ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

Η ΚΑΝΟΝΙΚΗ ΠΕΡΙΟΔΟΣ ΚΑΙ ΤΑ ΠΛΕΙ-ΟΦΣ ΤΟΥ NBA: ΣΥΓΚΡΙΣΗ ΚΑΙ ΠΡΟΒΛΕΨΕΙΣ ΜΕ ΣΤΑΤΙΣΤΙΚΕΣ ΜΕΘΟΔΟΥΣ

Βασίλειος Χρυσής

Διπλωματική Εργασία
που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Μάρτιος 2024

Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

Η ΚΑΝΟΝΙΚΗ ΠΕΡΙΟΔΟΣ ΚΑΙ ΤΑ ΠΛΕΙ-ΟΦΣ ΤΟΥ ΝΒΑ: ΣΥΓΚΡΙΣΗ ΚΑΙ ΠΡΟΒΛΕΨΕΙΣ ΜΕ ΣΤΑΤΙΣΤΙΚΕΣ ΜΕΘΟΔΟΥΣ

Βασίλειος Χρυσής

Διπλωματική Εργασία
που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Μάρτιος 2024

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. …….. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Αναπληρωτής Καθηγητής Κωνσταντίνος Πολίτης (Επιβλέπων)
- Αναπληρωτής Καθηγητής Χαράλαμπος Ευαγγελάρας
- Επίκουρος Καθηγητής Ιωάννης Τριανταφύλλου

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

REGULAR SEASON AND PLAYOFFS AT THE NBA: COMPARISON AND PREDICTIONS USING STATISTICAL METHODS

By

Vasileios Chrysis

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
March 2024

Στους γονείς μου

Σταμάτη και Σταυρούλα

Ευχαριστίες

Η παρούσα διπλωματική εργασία με θέμα «Η κανονική περίοδος και τα πλειοφς του NBA: σύγκριση και προβλέψεις με στατιστικές μεθόδους» πραγματοποιήθηκε στο πλαίσιο του μεταπτυχιακού προγράμματος σπουδών «Εφαρμοσμένης Στατιστικής του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστήμιου Πειραιώς. Με την ολοκλήρωση της νιώθω την ανάγκη να ευχαριστήσω τους ανθρώπους που με στήριξαν κατά την διάρκεια της εκπόνησης αλλά και της πορείας μου σε όλο το μεταπτυχιακό πρόγραμμα. Αρχικά, ευχαριστώ τον επόπτη μου, κ. Κωσταντίνο Πολίτη, για την καθοριστική συμβολή στην ολοκλήρωση της παρούσας εργασίας και τον χρόνο που αφιέρωσε για την στενή παρακολούθηση της πορείας μου. Επιπροσθέτως, θα ήθελα να εκφράσω το μεγαλύτερο «ευχαριστώ» στην οικογένεια μου διότι ό,τι έχω καταφέρει έως σήμερα δεν θα ήταν εφικτό χωρίς την στήριξη και την κατανόηση τους.

Περίληψη

Τη σημερινή εποχή τα δεδομένα αποτελούν τη βάση της σύγχρονης κοινωνίας, αποτυπώνοντας τον ψηφιακό κόσμο και τις αλληλεπιδράσεις μας με αυτόν. Η συλλογή τους αποτελεί ένα σημαντικό βήμα στη διαδικασία ανάλυσης και εξαγωγής σημαντικών πληροφοριών. Από τη βιομηχανία μέχρι τον αθλητισμό, η τεχνολογία έχει δημιουργήσει μια μεγάλη συλλογή δεδομένων. Στον αθλητισμό και ειδικότερα στις διοργανώσεις της καλαθοσφαίρισης η ανάλυση δεδομένων έχει αλλάξει θεμελιωδώς τον τρόπο που οι ομάδες και οι προπονητές σχεδιάζουν τους αγώνες και αντιδρούν στον αγωνιστικό χώρο. Με τη συνεχή ροή στατιστικών δεδομένων, τόσο για τους παίκτες όσο και για τις ομάδες, δίνεται η δυνατότητα στους προπονητές να βελτιώσουν τα αδύναμα σημεία της ομάδας τους και να επαναπροσδιορίσουν τις στρατηγικές τους. Μέσω αυτών των δυνατοτήτων η ανάλυση δεδομένων αναδεικνύει ένα νέο επίπεδο προηγμένης προετοιμασίας και ανταγωνιστικότητας στα πρωταθλήματα.

Στην παρούσα διπλωματική εργασία εκτός από τις κλασικές μεταβλητές που συλλέγονται στους αγώνες καλαθοσφαίρισης της διοργάνωσης του National Basketball Association (NBA), όπως οι πόντοι ανά παιχνίδι και οι ασίστ, δημιουργήθηκαν και ορισμένες νέες στατιστικές μεταβλητές, όπως οι κατοχές και ο δείκτης DRt, ώστε να επιτευχθούν οι στόχοι που τέθηκαν εξαρχής. Αρχικά, παρουσιάζεται μέσω πινάκων και κατάλληλων γραφημάτων μια περιγραφική ανάλυση όλων των μεταβλητών για τα δεδομένα που συλλέχθηκαν από τις τελευταίες δεκαπέντε (15) σεζόν της διοργάνωσης. Έπειτα ακολουθούν ορισμένοι έλεγχοι, κανονικότητας και συσχέτισης, που εκτελέστηκαν προτού εφαρμοστούν τα κατάλληλα γενικευμένα γραμμικά μοντέλα, ώστε να βρεθούν οι σημαντικότερες μεταβλητές που επηρεάζουν την κατάκτηση του πρωταθλήματος και την πρόκριση (ή μη) μιας ομάδας στην φάση των Playoffs. Ακόμα, γίνεται μια ανάλυση των δεδομένων βάσει χρονοσειρών ώστε να παρουσιαστούν προβλέψεις για τις επόμενες δύο (2) σεζόν. Στη συνέχεια, χρησιμοποιούνται κατάλληλες τεχνικές μηχανικής μάθησης, ώστε να δημιουργηθούν ορισμένα μοντέλα ταξινόμησης (classification) και να διερευνηθεί αν υπάρχουν ανάμεσα στα δεδομένα κλάσεις με κοινά χαρακτηριστικά (clustering), για να εξεταστεί η πρόκριση των ομάδων στη φάση των Playoffs. Τέλος, παρουσιάζονται τα τελικά συμπεράσματα της εργασίας και τυχόν ομοιότητες με προηγούμενες αναλύσεις που έχουν γίνει.

Abstract

In modern times, data is considered an integral part of society, reflecting the digital world as well as our interactions within it. Data acquisition and, subsequently, its analysis leads to extraction of important information, shedding light to many "gray" aspects. From the pharmaceutical industry to sports, technology has contributed to the acquisition of a vast pool of data. More specifically in sports, particularly in basketball events, data analysis has fundamentally altered how teams and their coaches plan the games as well as how they behave on the field. With the continuous flow of statistical data for both players and teams, coaches get to have better insights and, consequently, improve the "Achilles Heel" of their team's and redefine their strategies. Thus, data analysis seems to have unlocked a whole new level of advanced preparation and competitiveness in championships.

This thesis focuses on sports and utilizes a combination of the performance indicators collected in National Basketball Association (NBA) games (e.g points per game and assists) and novel, recently created statistical variables, such as possessions and the DRt index. Initially, a descriptive analysis of all variables for the data collected from the last fifteen (15) seasons of the tournament is presented through tables and appropriate graphs. Then, certain tests of normality and correlation are conducted before applying suitable generalized linear models to find the most significant variables affecting championship conquest and a team's advancement (or lack thereof) to the Playoffs stage. Additionally, a time series data analysis is performed to provide forecasts for the next two (2) seasons. Subsequently, appropriate machine learning techniques are used to create classification models and explore if there are classes with common characteristics among the data (clustering) to examine team advancement to the Playoffs stage. The final conclusions of the study are presented along with comparisons to previous analyses conducted.

Πίνακας Περιεχομένων

Κατάλογος Σχημάτων και Πινάκων

ΚΕΦΑΛΑΙΟ 1ο

1. Εισαγωγή

Η παρούσα διπλωματική εργασία έχει ως σκοπό την ανάλυση των στατιστικών της διοργάνωσης National Basketball Association (NBA), με τη χρήση τόσο στατιστικών τεχνικών όσο και τεχνικών μηχανικής μάθησης. Η μελέτη στηρίζεται σε αριθμητικά δεδομένα από τη σεζόν 2009-2010 έως και την τελευταία σεζόν, που ολοκληρώθηκε πέρσι, 2022-2023. Τα δεδομένα που τίθενται προς ανάλυση αναφέρονται σε δύο διαφορετικές φάσεις της διοργάνωσης, την πρώτη που είναι η κανονική περίοδος (regular season) και τη δεύτερη που είναι τα Playoffs. Επίσης για την πραγματοποίηση της διερεύνησης το σύνολο των δεδομένων χωρίζεται σε δύο βασικές κατηγορίες, την ομάδα που κατέκτησε το πρωτάθλημα (Champions) και τις ομάδες που προκρίθηκαν ή όχι στα Playoffs (Playoffs).

Το δεύτερο κεφάλαιο της εργασίας αποσκοπεί στην περιγραφή της διοργάνωσης του NBA. Αρχικά, πραγματοποιείται μια ιστορική αναδρομή κατά την οποία παρουσιάζονται οι κυριότερες αλλαγές που συνέβησαν με το πέρασμα του χρόνου στην διοργάνωση. Έπειτα παρουσιάζονται οι ομάδες και οι μεταβλητές των δεδομένων που χρησιμοποιούνται στην ανάλυση της παρούσας εργασίας. Τέλος, γίνεται αναφορά σε διάφορα επιστημονικά άρθρα και τα αποτελέσματα τους, τα οποία αναλύουν στατιστικά της διοργάνωσης του NBA είτε παρόμοιων διοργανώσεων, όπως της αντίστοιχης διοργάνωσης της Ευρώπης (Euroleague).

Στο τρίτο κεφάλαιο, γίνεται αναλυτική παρουσίαση περιγραφικών μέτρων και διαφόρων διαγραμμάτων για τις επεξηγηματικές μεταβλητές του συνόλου δεδομένων για τις δύο φάσεις της διοργάνωσης, με σκοπό την καλύτερη κατανόηση από τον αναγνώστη. Στη συνέχεια ακολουθούν τρεις διαφορετικές αναλύσεις βάσει των κατηγορικών μεταβλητών Playoffs (πρόκριση ή όχι στα playoffs), Conference (η ομάδα συμμετέχει στην περιφέρεια East ή West) και Teams (όνομα ομάδας).

Στο τέταρτο κεφάλαιο, πραγματοποιούνται αρχικά οι έλεγχοι κανονικότητας για τις ποσοτικές μεταβλητές του συνόλου των δεδομένων για τις δύο φάσεις της διοργάνωσης, και σε ορισμένες επεξηγηματικές μεταβλητές (PPG, APG, SPG, BPG, RPG, TOV, DEFF) για τις δύο πρώτες αναλύσεις που έγιναν στο προηγούμενο κεφάλαιο. Στη συνέχεια παρουσιάζονται οι συντελεστές συσχέτισης των μεταβλητών και πραγματοποιούνται έλεγχοι για την ισότητα των μέσων τιμών. Οι έλεγχοι για τις μέσες τιμές γίνονται βάσει των δύο πρώτων αναλύσεων και για ζευγαρωτές παρατηρήσεις βάσει των στατιστικών που έχουν οι ομάδες που προκρίνονται και στις δύο φάσεις της διοργάνωσης. Για όλους τους προαναφερθέντες ελέγχους, το επίπεδο σημαντικότητας με το οποίο εργαστήκαμε, ισούται με 5%.

Το πέμπτο κεφάλαιο αποσκοπεί στην εύρεση των πιο σημαντικών μεταβλητών που επηρεάζουν την πρόκριση μιας ομάδας στη φάση των Playoffs και την κατάκτηση του πρωταθλήματος. Για την υλοποίηση αυτού του στόχου προσαρμόστηκαν κατάλληλα γενικευμένα γραμμικά μοντέλα με μεταβλητές απόκρισης τις κατηγορικές μεταβλητές

Playoffs (πρόκριση ή όχι) και Champions (πρωταθλήτρια ομάδα ή όχι). Πιο συγκεκριμένα αρχικά προσαρμόστηκαν μοντέλα όπου χρησιμοποιήθηκαν όλες οι επεξηγηματικές μεταβλητές και έπειτα χωρίς τις μεταβλητές ORt και DRt.

Το έκτο κεφάλαιο παρουσιάζει την ανάλυση ορισμένων μεταβλητών των δεδομένων με τη μορφή χρονοσειρών και έχει ως στόχο την εύρεση κατάλληλων μοντέλων ARIMA για την περιγραφή τους, τα οποία πληρούν τους ελέγχους στασιμότητας, κανονικότητας και ανεξαρτησίας των χρονοσειρών. Τέλος, πραγματοποιούνται προβλέψεις για τις τιμές των μεταβλητών στις δύο επόμενες σεζόν.

Στο έβδομο κεφάλαιο, γίνεται χρήση τεχνικών μηχανικής μάθησης. Αρχικά, εφόσον τα δεδομένα έχουν υποστεί την κατάλληλη επεξεργασία και έχουν επιλεχθεί οι σημαντικότερες μεταβλητές (feature selection), υλοποιούνται τεχνικές κατηγοριοποίησης (classification) προκειμένου να προσαρμοστούν κατάλληλα μοντέλα ταξινόμησης για το αν μια ομάδα μπορεί να προκριθεί στη δεύτερη φάση της διοργάνωσης (Playoffs). Στη συνέχεια χρησιμοποιούνται τεχνικές ομαδοποίησης (clustering) με στόχο την εύρεση μοντέλων και ομάδων παρατηρήσεων που έχουν παρόμοια χαρακτηριστικά ανάμεσα στα δεδομένα μας.

Στο όγδοο κεφάλαιο παρουσιάζονται συνοπτικά τα συμπεράσματα που προέκυψαν στα προηγούμενα κεφάλαια και γίνονται συγκρίσεις με άλλες παρόμοιες επιστημονικές έρευνες.

ΚΕΦΑΛΑΙΟ 2ο

2. Ιστορική εξέλιξη και περιγραφή δεδομένων

Το παρόν κεφάλαιο έχει ως στόχο να παρουσιάσει την ιστορική εξέλιξη του θεσμού του NBA, όπως και τον τρόπο διεξαγωγής του, και να εισάγει για πρώτη φορά τα δεδομένα που τέθηκαν προς επεξεργασία στα επόμενα κεφάλαια της εργασίας. Το θεωρητικό υπόβαθρο για την ιστορική αναδρομή της διοργάνωσης έχει αντληθεί από την επίσημη ιστοσελίδα Wikipedia.org.

2.1. Η διοργάνωση του NBA

2.1.1. Ιστορική αναδρομή

Το NBA (National Basketball Association) είναι η διοργανώτρια αρχή του πρωταθλήματος καλαθοσφαίρισης των Ηνωμένων Πολιτειών Αμερικής και αποτελεί επίσημο μέλος της USAB, η οποία είναι μια μη κερδοσκοπική οργάνωση που ιδρύθηκε το 1974 και αποτελεί τον επίσημο εκπρόσωπο των αμερικάνικων πρωταθλημάτων καλαθοσφαίρισης. Επιπλέον η USAB είναι αναγνωρισμένη ως υπεύθυνη για το άθλημα της καλαθοσφαίρισης στις Ηνωμένες Πολιτείες από την FIBA, η οποία αποτελεί την επίσημη αθλητική ομοσπονδία που είναι υπεύθυνη για τους κανονισμούς του μπάσκετ τόσο για το αγωνιστικό όσο και το διοικητικό επίπεδο σε παγκόσμια εμβέλεια.

Το πρωτάθλημα ονομάστηκε επίσημα για πρώτη φορά NBA στις 3 Αυγούστου του 1949 ως συνέπεια της συμφωνίας συγχώνευσης ανάμεσα στα πρωταθλήματα BAA και NBL. Το BAA ήταν ένα επίσημο πρωτάθλημα καλαθοσφαίρισης στις ΗΠΑ με ημερομηνία ίδρυσης τις 6 Ιουνίου του 1946 στην Νέα Υόρκη και αποτέλεσε το πρώτο πρωτάθλημα μπάσκετ που διεξαγόταν σε μεγάλες πόλεις και γήπεδα, ενώ το NBL είχε ιδρυθεί από το 1937 αποτελούμενο από μικρές ομάδες της περιοχής Great Lakes. Εκείνο το διάστημα το πρωτάθλημα αποτελούνταν από δεκαεπτά (17) ομάδες, εκ των οποίων μόνο οι έξι (6) προέρχονταν από την λίγκα του NBL, από μικρές και μεγάλες πόλεις, ενώ η συγχώνευση των δύο πρωταθλημάτων έμελλε να αλλάξει την ιστορία του μπάσκετ στις ΗΠΑ. Επίσημα η σεζόν 1949-1950 θεωρείται η τέταρτη σεζόν διεξαγωγής της διοργάνωσης, διότι διατηρήθηκαν τα στατιστικά και τα αρχεία που υπήρχαν από το πρωτάθλημα BAA σαν παρακαταθήκη της διοργάνωσης. Έπειτα από είκοσι επτά χρόνια, το 1976, συνέβη η τελευταία συγχώνευση πρωταθλημάτων, εφόσον το πρωτάθλημα ABA ενσωματώθηκε στο NBA και πρόσθεσε σε αυτό τέσσερις νέες ομάδες ώστε να πάρει και την τελική του μορφή. Το πρωτάθλημα ABA αποτελούσε και αυτό μια από τις μεγαλύτερες διοργανώσεις μπάσκετ της Αμερικής με έτος ίδρυσης το 1967. Σήμερα στο πρωτάθλημα συμμετέχουν πια τριάντα (30) ομάδες, εκ των οποίων οι είκοσι εννιά (29) προέρχονται από την Αμερική και μία (1) από τον Καναδά.

Οι δύο ομάδες με τους περισσότερους τίτλους της διοργάνωσης είναι οι Los Angeles Lakers και οι Boston Celtics με δεκαεπτά (17) τίτλους η κάθε μία. Έπειτα και από την ολοκλήρωση της σεζόν 2022-2023, όπου οι Denver Nuggets κατέκτησαν για πρώτη

φορά στην ιστορία τους το πρωτάθλημα, έχουν κατακτήσει τον τίτλο είκοσι (20) διαφορετικοί σύλλογοι.

Σχήμα 2.1: Οι εναλλαγές στο λογότυπο της διοργάνωσης έως και σήμερα.

(Πηγή: https://logos-world.net/nba-logo/)

2.1.2. Τα format της διοργάνωσης στο πέρασμα των χρόνων

Όπως αναφέρθηκε και σε προηγούμενη παράγραφο, μπορεί με κάθε επισημότητα η διοργάνωση να μετονομάστηκε σε NBA το 1949, όμως το πρώτο της πρωτάθλημα διεξάχθηκε την περίοδο 1946-1947 έχοντας τότε όνομα BAA. Από το πρώτο πρωτάθλημα έως και σήμερα η διαδικασία ανάδειξης του πρωταθλητή περνάει από δύο φάσεις. Η πρώτη φάση είναι η κανονική περίοδος (regular season). Η κανονική περίοδος τυπικά είναι ένας χρόνος αγώνων στον οποίο οι ομάδες διαιρούνται σε ομίλους και διεξάγεται ένας συγκεκριμένος αριθμός αγώνων ανάμεσα τους. Συγκεκριμένα, στο NBA κάθε ομάδα δίνει 82 παιχνίδια, ώστε να καταταχθεί στην αντίστοιχη θέση του πρωταθλήματος, και αφού βγει η τελική κατάταξη, οι πρωτοπόρες ομάδες κάθε ομίλου συμμετέχουν στην δεύτερη φάση, τα playoffs. Τα playoffs αποτελούν μία διοργάνωση που πραγματοποιείται μετά το πέρας της κανονικής περιόδου από τις κορυφαίες ομάδες με σκοπό την ανάδειξη του πρωταθλητή.

Στην πρώτη σεζόν, 1946-1947, συμμετείχαν συνολικά έντεκα (11) ομάδες σε δύο (2) ομίλους. Ο ένας όμιλος είχε την ονομασία Eastern Division με έξι (6) ομάδες που στην κανονική περίοδο έδωσαν εξήντα (60) παιχνίδια και οι τρεις (3) πρώτες πέρασαν στα playoffs. Αντίστοιχα, στον δεύτερο όμιλο με την ονομασία Western Division συμμετείχαν οι υπόλοιπες πέντε (5) ομάδες που έδωσαν εξήντα ένα (61) αγώνες. Τα playoffs χωρίζονταν σε τρεις (3) γύρους. Στον πρώτο γύρο υπήρχαν αγώνες μεταξύ των τρίτων κάθε ομίλου και των δεύτερων. Στον δεύτερο γύρο γίνονταν αγώνες μεταξύ των πρώτων κάθε ομίλου και των νικητών του πρώτου γύρου, απ' όπου και προέκυπταν οι συμμετέχοντες του μεγάλου τελικού.

Με το πέρασμα των χρόνων διάφορες ομάδες προστίθενταν στους ομίλους που υπήρχαν. Συγκεκριμένα την σεζόν 1949-1950 ήταν η πρώτη φορά που δημιουργήθηκαν τρεις (3) όμιλοι αντί για δύο (2). Έπειτα από την είσοδο και των ομάδων από το NBL έχουμε τις Eastern Division, Central Division και Western Division. Όπως ήταν αναμενόμενο, με την νέα ονομασία του πρωταθλήματος ήρθαν και οι αλλαγές στην μορφή των playoffs. Πλέον προκρίνονταν σε αυτά οι τέσσερις (4) πρώτοι από κάθε όμιλο και διεξάγονταν τρεις (3) τελικοί, ένας για κάθε όμιλο, και στην συνέχεια ένας (1) για την ανάδειξη του πρωταθλητή όλου του NBA. Αρχικά, οι πρώτοι της regular season έπαιζαν με τους αντίστοιχους τέταρτους του ομίλου τους και κατά συνέπεια οι δεύτεροι με τους τρίτους. Οι νικητές των αγώνων, οι οποίοι έπρεπε να φτάσουν σε δύο νίκες απέναντι στους αντιπάλους σε συνεχόμενα παιχνίδια, περνούσαν στον τελικό, όπου θα αναδεικνυόταν ο πρωταθλητής κάθε ομίλου. Στη συνέχεια υπήρχε ο αγώνας των ημιτελικών, όπου συμμετείχαν οι δύο (2) από τους τρεις (3) πρωταθλητές των ομίλων, και έπειτα διεξαγόταν ο τελικός. Έως και σήμερα το πρωτάθλημα έχει κρατήσει τον ίδιο τρόπο διεξαγωγής με μικρές διαφορές που γίνονταν ανά διαστήματα.

Ορόσημο στην διοργάνωση αποτέλεσε η σεζόν 1970-1971, όπου για πρώτη φορά υπήρξαν τέσσερις (4) όμιλοι, με τις ονομασίες Atlantic Division, Central Division, Midwest Division και Pacific Division, αποτελούμενοι από τέσσερις (4) ομάδες ο καθένας. Ακόμα, οι τέσσερις (4) όμιλοι για πρώτη φορά δημιουργούσαν δύο (2) μεγαλύτερες κατηγορίες γνωστές ως περιφέρειες με τα ονόματα Eastern Conference και Western Conference. Οι ομάδες έδιναν συνολικά ογδόντα δύο (82) παιχνίδια και οι τέσσερις (4) πρώτες από κάθε κατηγορία πέρναγαν στα playoffs. Τα playoffs διεξάγονταν με τον ίδιο τρόπο όπως την σεζόν 1949-1950, δηλαδή ο πρώτος έπαιζε με τον τέταρτο της περιφέρειας και ο δεύτερος με τον τρίτο και έπειτα οι νικητές συμμετείχαν στον τελικό της κατηγορίας τους. Οι πρωταθλητές των κατηγοριών προκρίνονταν στον τελικό του NBA για να στεφθεί ο πρωταθλητής.

Εφόσον τα δεδομένα της παρούσας εργασίας αναφέρονται από τη σεζόν 2008-2009 έως και την 2022-2023, θα περάσουμε στην ανάλυση του τρόπου διεξαγωγής καθεμίας απ’ αυτές αναλυτικότερα.

- <u>Σεζόν 2008-2009 έως και σεζόν 2010-2011</u>

Από τη σεζόν 2008-2009 έως και τη σεζόν 2010-2011 ο τρόπος διεξαγωγής της διοργάνωσης ήταν ο ίδιος κάθε χρονιά. Αρχικά, αρχές Οκτωβρίου ξεκινούσε πάντοτε η κανονική περίοδος που κρατούσε μέχρι τον Απρίλιο, όταν ξεκινούσαν τα playoffs. Στο πρωτάθλημα συμμετείχαν 30 ομάδες, οι οποίες χωρίζονταν σε πρώτη φάση σε δύο περιφέρειες τις Eastern Conference και Western Conference. Οι δύο περιφέρειες αποτελούνταν από δεκαπέντε (15) ομάδες η καθεμία και χωρίζονταν ξανά σε τρείς (3) ομίλους. Οι όμιλοι της πρώτης περιφέρειας (Eastern) ονομάζονταν Atlantic Division, Central Division και Southeast Division ενώ της δεύτερης περιφέρειας Northwest Division, Pacific Division και Southwest Division. Κάθε όμιλος περιείχε πέντε (5) ομάδες αντίστοιχα.

Κατά την διάρκεια της κανονικής περιόδου κάθε ομάδα πραγματοποιούσε ογδόντα δύο (82) αγώνες, σαράντα ένα (41) εντός έδρας και σαράντα ένα (41) εκτός έδρας. Συγκεκριμένα, κάθε ομάδα αντιμετώπιζε τέσσερις (4) φορές κάθε αντίπαλο από τον δικό της όμιλο (συνολικά δεκαέξι παιχνίδια). Έπειτα από τους άλλους δυο ομίλους της περιφέρειας της (δέκα ομάδες σύνολο και στους δυο ομίλους) αντιμετώπιζε τέσσερις (4) φορές τους έξι (6) αντιπάλους απ' αυτούς (είκοσι τέσσερα παιχνίδια) και τρείς (3) φορές τους υπόλοιπους τέσσερις (4) (δώδεκα παιχνίδια). Ακόμα, κάθε ομάδα αντιμετώπιζε δύο (2) φορές τις ομάδες από την αντίπαλη περιφέρεια (τριάντα αγώνες). Συνεπώς γίνεται αντιληπτό ότι το πρόγραμμα των ομάδων διαφοροποιούταν από ομάδα σε ομάδα, εφόσον υπήρχαν αυτές οι κατατάξεις των αγώνων. Στο τέλος της κανονικής περιόδου οι ομάδες κατατάσσονταν ανάλογα με το ρεκόρ τους σε σειρά βαθμολογίας στις περιφέρειες (Conference) που συμμετείχαν ώστε οι πρωτοπόροι να πάρουν μέρος στα playoffs.

Στα playoffs συμμετείχαν οι πρώτες οκτώ (8) ομάδες από κάθε περιφέρεια και έδιναν συγκεκριμένα παιχνίδια μεταξύ τους, ανάλογα με τη θέση που είχαν εξασφαλίσει κατά την regular season. Συγκεκριμένα, στον πρώτο γύρο ο πρώτος κάθε περιφέρειας έπαιζε με τον όγδοο της ίδιας περιφέρειας, ο δεύτερος με τον έβδομο, ο τρίτος με τον έκτο και ο τέταρτος με τον πέμπτο. Η ομάδα με την καλύτερη θέση στην βαθμολογία είχε το πλεονέκτημα έδρας για την σειρά των αγώνων. Για να προκριθεί μια ομάδα, έπρεπε να φτάσει στις τέσσερις (4) νίκες και έτσι γίνεται αντιληπτό πως ο μέγιστος αριθμός αγώνων σε κάθε γύρο ήταν επτά. Έπειτα ακολουθούσε ο δεύτερος γύρος όπου συμμετείχαν οι νικητές του προηγούμενου γύρου και ο νικητής από το ζευγάρι του πρώτου και του όγδοου προχωρούσε σε αναμέτρηση με τον νικητή του ζεύγους του τέταρτου και πέμπτου σε μια σειρά αγώνων όπως η προηγούμενη. Ο επόμενος γύρος ήταν οι τελικοί κάθε περιφέρειας ξεχωριστά, δηλαδή προέκυπτε ο πρωταθλητής του Eastern και Western Conference. Τέλος οι δύο πρωταθλητές έρχονταν αντιμέτωποι στους γενικούς τελικούς για την κατάκτηση του NBA, όπου και πάλι η νικήτρια ομάδα θα έπρεπε να νικήσει τέσσερις (4) φορές στα μεταξύ τους παιχνίδια.

- <u>Σεζόν 2011-2012</u>

Η σεζόν 2011-2012 ξεκίνησε με την υπογραφή μίας νέας συλλογικής σύμβασης εργασίας (CBA) ως αποτέλεσμα μιας διαπραγμάτευσης ανάμεσα στους ιδιοκτήτες των ομάδων και στους παίκτες του NBA για τα εργασιακά δικαιώματα των παικτών. Με την νέα σύμβαση η κανονική περίοδος θα αποτελούνταν από εξήντα έξι (66) παιχνίδια και όχι ογδόντα δύο (82) όπως τα προηγούμενα χρόνια. Συνεπώς υπήρχε νέος διακανονισμός για τους αγώνες της κανονικής περιόδου. Αυτή τη χρονιά κάθε ομάδα αντιμετώπισε σε σαράντα οκτώ (48) αγώνες ομάδες από την περιφέρεια στην οποία συμμετείχε και σε δεκαοκτώ (18) αγώνες ομάδες της αντίπαλης περιφέρειας. Όσον αφορά τα playoffs δεν άλλαξε κάτι στην διεξαγωγή τους.

- <u>Σεζόν 2012-2013 έως και 2018-2019</u>

Στις επόμενες σεζόν η διοργάνωση επέστρεψε στο πρότυπο που ακολουθούσε την σεζόν 2008-2009. Ακόμα, αυτή την περίοδο υπήρξαν δύο αλλαγές ονομάτων σε

ομάδες. Οι New Jersey Nets μετονομάστηκαν σε Brooklyn Nets (2012-2013) και οι Charlotte Bobcats σε Charlotte Hornets (2014-2015). Οι συγκεκριμένες αλλαγές ήταν ένα rebrand.

- Σεζόν 2019-2020

Η σεζόν 2019-2020 ξεκίνησε στις 22 Οκτωβρίου 2019, όμως διακόπηκε προσωρινά στις 11 Μαρτίου του 2020 λόγω της εμφάνισης του COVID-19[1]. Έτσι δεν ολοκληρώθηκαν ποτέ οι ογδόντα δύο (82) αγώνες που ήταν οριοθετημένοι για τη σεζόν και επομένως όλες οι ομάδες είχαν αγωνιστεί από εξήντα τρία έως εβδομήντα (63-70) παιχνίδια η καθεμία, αναλόγως το πρόγραμμα που είχε προκύψει στην αρχή της σεζόν έως την μέρα της διακοπής. Τελικά το πρωτάθλημα συνεχίστηκε έπειτα από συμφωνία της διοργάνωσης με τους παίκτες στις 4 Ιουνίου του 2020. Συγκεκριμένα, πήραν μέρος είκοσι δύο (22) ομάδες και έδωσαν ακόμα οκτώ (8) αγώνες ώστε να βγει η τελική κατάταξη. Λόγω της μη ολοκλήρωσης της κανονικής περιόδου υπήρξαν συγκεκριμένες αλλαγές για την εκπροσώπηση των ομάδων στα playoffs. Σε περίπτωση που οι θέσεις οκτώ (8) και εννέα (9) απείχαν τέσσερις (4) νίκες μεταξύ τους θα έπαιζαν στο play-in-tournament, δηλαδή θα έπαιζαν μία σειρά αγώνων μεταξύ τους. Αν οι όγδοοι νικούσαν το πρώτο παιχνίδι της σειράς τότε προκρίνονταν στα playoffs, αν όμως η ένατη θέση κέρδιζε το πρώτο παιχνίδι τότε διεξαγόταν ένα τελευταίο παιχνίδι και ο νικητής συνέχιζε στα playoffs. Συνεπώς, για να περάσει ο ένατος θα έπρεπε να κερδίσει δύο συνεχόμενους αγώνες. Αυτό το εσωτερικό τουρνουά δημιουργήθηκε, διότι οι ομάδες στην κανονική περίοδο δεν είχαν ολοκληρώσει, λόγω της ασθένειας, τους αγώνες που ήταν υποχρεωμένες να δώσουν και έτσι κρίθηκε δίκαιο να δοθεί κάτω από τις συγκεκριμένες συνθήκες μία ευκαιρία στον ένατο να διεκδικήσει τη θέση του στα playoffs. Η διεξαγωγή των playoffs ήταν ακριβώς η ίδια με τις προηγούμενες σεζόν.

- Σεζόν 2020-2021

Η επόμενη σεζόν έπειτα από την εμφάνιση του COVID-19 παρουσίασε κάποιες ιδιαιτερότητες. Αρχικά, η κανονική περίοδος ξεκίνησε στις 22 Δεκεμβρίου, εξαιτίας της αργοπορίας στο τέλος της διοργάνωσης της προηγούμενης χρονιάς, και περιλάμβανε εβδομήντα δύο (72) παιχνίδια για κάθε ομάδα. Ακόμα, οι ομάδες που κατατάσσονταν στις θέσεις επτά (7) έως και δέκα (10) σε κάθε περιφέρεια, έπειτα από το πέρας της κανονικής περιόδου, υποχρεούνταν να πάρουν μέρος σε ένα νέο πρωτάθλημα (play-in-tournament) για την συμμετοχή τους στα playoffs. Στον πρώτο γύρο ο ένατος αντιμετώπιζε τον δέκατο σε έναν αγώνα και ο νικητής περνούσε στον δεύτερο γύρο. Παράλληλα ο έβδομος αντιμετώπιζε τον όγδοο σε έναν αγώνα όπου όποιος κέρδιζε προκρινόταν στα playoffs. Τέλος, ο νικητής από το πρώτο γκρουπ ομάδων αντιμετώπιζε τον χαμένο από το δεύτερο για την πρόκριση στα playoffs. Έπειτα τα playoffs διεξάγονταν με τον ίδιο τρόπο που γίνονταν τα προηγούμενα χρόνια. Στην παρούσα εργασία δεν έχουν υπολογιστεί οι αγώνες των play-in-tournament· συνεπώς

[1] COVID-19 επίσης γνωστή ως οξεία αναπνευστική νόσος 2019-nCoV, είναι μία μολυσματική ασθένεια που προκαλείται από τον κορονοϊό SARS-CoV-2
Προσβάσιμο στο: https://el.wikipedia.org/wiki/COVID-19 [Ανακτήθηκε: 19-7-2023].

οι τελικές κατατάξεις για τις regular season υπολογίζονται αφού έχουν πραγματοποιηθεί οι αγώνες που αναφέρθηκαν στο play-in-tournament.

- Σεζόν 2021-2022 και 2022-2023

Σε αυτές τις δύο σεζόν κάθε ομάδα παίρνει μέρος σε ογδόντα δύο (82) αγώνες στη κανονική περίοδο για πρώτη φορά μετά την εμφάνιση του COVID-19. Ακόμα, το play-in-tournament, για τις θέσεις επτά (7) έως και δέκα (10), έχει γίνει πλέον επίσημο και πραγματοποιείται κάθε χρονιά για να ανακηρυχθούν οι συμμετέχοντες των playoffs.

2.2. Παρουσίαση των δεδομένων

Στην παρούσα διπλωματική εργασία τα δεδομένα που χρησιμοποιήθηκαν αφορούν στατιστικά δεδομένα των ομάδων του NBA τα τελευταία δεκαπέντε (15) χρόνια, ξεκινώντας από την αγωνιστική περίοδο 2008-2009 έως και την σεζόν 2022-2023.

Η προσαρμογή των δεδομένων έγινε με τη χρήση του προγράμματος επεξεργασίας υπολογιστικών φύλλων «Microsoft Excel». Τα δεδομένα αντλήθηκαν από την επίσημη σελίδα basketball-realgm.[2] Η ανάλυση και η επεξεργασία τους έγινε με τις γλώσσες προγραμματισμού «R» και «Python». Ο κώδικας παρουσιάζεται αναλυτικά στο επισυναπτόμενο παράρτημα στο τέλος της παρούσας εργασίας (Π1. Κώδικας υλοποίησης της παρούσας εργασίας).

Αρχής γενομένης από την σεζόν 2008-2009, πρώτη σεζόν που αναλύεται στα δεδομένα, το NBA απαρτίζεται από τριάντα (30) διαφορετικές ομάδες. Οι μόνες αλλαγές που πραγματοποιούνται έως και την σεζόν 2022-2023, όσον αφορά τις ομάδες της διοργάνωσης, είναι δύο rebrand. Το πρώτο, όπως ήδη αναφέρθηκε, με την ομάδα των New Jersey Nets να μετονομάζεται την σεζόν 2012-2013 σε Brooklyn Nets και αυτή των Charlotte Bobcats σε Charlotte Hornets το 2014-2015. Επειδή οι συγκεκριμένες αλλαγές έγιναν για χορηγικούς λόγους στην παρούσα εργασία παρουσιάζονται εξ αρχής και οι δύο ομάδες με τα τελικά τους ονόματα. Στον επόμενο πίνακα παρουσιάζονται και οι τριάντα (30) ομάδες που απαρτίζουν το πρωτάθλημα της διοργάνωσης αυτά τα χρόνια.

Team
Sacramento
Golden State
Atlanta
Boston
Oklahoma City
L.A. Lakers
Utah
Milwaukee

[2] Basketball RealGM αποτελεί έναν γρήγορα εξελίξιμο ιστότοπο, ο οποίος παρέχει στους χρήστες όλα τα δεδομένα και τα εργαλεία που απαιτούνται για την προσομοίωση της δουλείας ενός πραγματικού γενικού διευθυντή. Οι ομάδες του NBA χρησιμοποιούν το λογισμικό των υπηρεσιών εφαρμογών που παρέχεται στην σελίδα από το 2003.
Προσβάσιμο στο: https://basketball.realgm.com/info/about-us [Ανακτήθηκε: 22-7-2023].

Memphis
Indiana
New York
Denver
Minnesota
Philadelphia
New Orleans
Dallas
Phoenix
L.A. Clippers
Portland
Brooklyn
Washington
Chicago
San Antonio
Toronto
Cleveland
Orlando
Charlotte
Houston
Detroit
Miami

Πίνακας 2.1: Οι ομάδες που έχουν συμμετάσχει στην διοργάνωση του NBA

Οι μεταβλητές που χρησιμοποιήθηκαν στην ανάλυση των στατιστικών των ομάδων που προαναφέρθηκαν παρουσιάζονται στον επόμενο πίνακα.

Μεταβλητές	Επεξήγηση
Year	**Χρονιά διεξαγωγής πρωταθλήματος (2009:2008-2009, 2010:2009-2010 κ.ο.κ)**
Team	**Όνομα ομάδας**
Conference	**Πρωτάθλημα Περιφέρειας που συμμετέχει (1: West, 0: East)**
GP	**Αγώνες που συμμετείχε η ομάδα**
MPG	**Μέσος χρόνος διάρκεια αγώνα**
PPG	**Πόντοι ανά παιχνίδι**
PPGA	**Πόντοι που δέχεται ανά παιχνίδι**
FGM	**Εύστοχα δίποντα και τρίποντα ανά αγώνα**
FGA	**Συνολικές προσπάθειες δίποντων και τριπόντων ανά αγώνα**
MisFG	**Άστοχα δίποντα και τρίποντα ανά αγώνα**
FG%	**Ποσοστό εύστοχων δίποντων και τρίποντων ανά αγώνα**
FGAo	**Συνολικές προσπάθειες δίποντων και τριπόντων ανά αγώνα των αντιπάλων**
3PM	**Εύστοχα τρίποντα ανά αγώνα**
3PA	**Συνολικές προσπάθειες τριπόντων ανά αγώνα**
3P%	**Ποσοστό εύστοχων τριπόντων ανά αγώνα**
FTM	**Εύστοχες ελεύθερες βολές ανά αγώνα**

FTA	**Συνολικές προσπάθειες ελεύθερων βολών ανά αγώνα**
MisFT	**Άστοχες ελεύθερες βολές ανά αγώνα**
FT%	**Ποσοστό εύστοχων ελεύθερων βολών ανά αγώνα**
FTAo	**Συνολικές προσπάθειες ελεύθερων βολών ανά αγώνα των αντιπάλων**
ORB	**Επιθετικά "Ριμπάουντ" ανά αγώνα**
DRB	**Αμυντικά "Ριμπάουντ" ανά αγώνα**
RPG	**Συνολικά "Ριμπάουντ" ανά αγώνα**
ORBo	**Επιθετικά "Ριμπάουντ" ανά αγώνα των αντιπάλων**
APG	**"Ασίστ" (Τελικές πάσες που οδηγούν σε καλάθι) ανά αγώνα**
SPG	**Κλεψίματα ανά αγώνα**
BPG	**Κοψίματα ανά αγώνα**
TOV	**Λάθη ανά αγώνα**
TOVo	**Λάθη ανά αγώνα των αντιπάλων**
PF	**Φάουλ ανά αγώνα**
POSSt	**Κατοχές ανά αγώνα**
ORt	**Offensive Rating**
POSSo	**Κατοχές ανά αγώνα των αντιπάλων**
DRt	**Defensive Rating**
DEFF	**Διαφορά Offensive με Defensive Rating**
EFF	**Δείκτης Efficiency**
STATrs	**Κατάταξη της regular season (1:πρώτος, 2:δεύτερος κ.ο.κ, 9:ομάδες από την όγδοη θέση και έπειτα)**
Playoffs	**Ένδειξη για το αν η ομάδα πέρασε στα playoffs (1: ναι, 0:όχι)**
Champions	**Πρωταθλητής NBA (1:ναι, 0:όχι)**
Top Seed	**Ομάδα με τις περισσότερες νίκες στην σεζόν (1: ναι, 0: όχι)**

Πίνακας 2.2: Οι μεταβλητές

Όπως έχει ήδη αναφερθεί και παραπάνω, για τις περισσότερες μεταβλητές που χρησιμοποιήθηκαν προς ανάλυση στην παρούσα εργασία τα δεδομένα αντλήθηκαν από τον επίσημο ιστότοπο του basketball-realgm. Παρ' όλα αυτά κάποιες απ' αυτές δημιουργήθηκαν στη συνέχεια για τους λόγους της διπλωματικής.

Η μεταβλητή MisFG δημιουργήθηκε ώστε να είμαστε σε θέση να υπολογίσουμε τον δείκτη αποδοτικότητας (efficiency) κάθε ομάδας, και μετράει τις αποτυχημένες προσπάθειες των δίποντων και τρίποντων για μία ομάδα ανά αγώνα. Ο τύπος υπολογισμού της μεταβλητής είναι ο εξής:

$$\mathbf{MisFG = FGA - FGM}$$

Αντίστοιχα, η μεταβλητή MisFT, η οποία μετράει τις αποτυχημένες προσπάθειες ελεύθερων βολών της ομάδας ανά αγώνα, υπολογίζεται από τον τύπο:

$$\mathbf{MisFT = FTA - FTM}$$

Μία ακόμα μεταβλητή που υπολογίστηκε στην παρούσα εργασία είναι η POSSt, η οποία μετράει τις κατοχές που συγκέντρωνε κάθε ομάδα κατά μέσο όρο σε έναν αγώνα που έδινε. Πιο συγκεκριμένα η κατοχή για μια ομάδα ξεκινάει όταν οι παίκτες

κερδίζουν τον έλεγχο της μπάλας και τελειώνει όταν χάνει τον έλεγχο της. Κάποιοι τρόποι ώστε η ομάδα να χάσει την κατοχή στον αγώνα είναι:

- όταν καταφέρει να πετύχει ένα καλάθι είτε αυτό είναι ελεύθερη βολή είτε δίποντο είτε τρίποντο
- όταν χάσει ένα ριμπάουντ
- όταν υποπέσει σε κάποιο λάθος κατά την διάρκεια του αγώνα

Ο τύπος υπολογισμού της μεταβλητής είναι:

$$\mathbf{POSSt = 0.976 \times (FGA + 0.44 \times FTA - OREB + TOV)}$$

Με αντίστοιχο τρόπο υπολογίστηκαν και οι κατοχές ανά αγώνα των αντιπάλων ομάδων, με τον τροποποιημένο τύπο:

$$\mathbf{POSSo = 0.976 \times (FGAo + 0.44 \times FTAo - OREBo + TOVo)}$$

Επίσης, άλλες δύο μεταβλητές που δημιουργήθηκαν στην παρούσα εργασία είναι το offensive και defensive rating των ομάδων, κατά μέσο όρο με βάση τα στατιστικά δεδομένα των αγώνων που έχει δώσει τη συγκεκριμένη χρονιά. Οι δύο αυτές μεταβλητές θα χρησιμεύσουν ώστε να γίνει και ο υπολογισμός της διαφοράς των παραπάνω ratings.

$$\mathbf{ORt = (PTS / POSSt) \times 100}$$

$$\mathbf{DRt = (PTSo / POSSo) \times 100}$$

$$\mathbf{DEFF = ORt - DRt}$$

Τελευταία μεταβλητή που υπολογίστηκε στα δεδομένα είναι αυτή του δείκτη Efficiency (EFF). Ο συγκεκριμένος δείκτης δημιουργήθηκε από τον Martin Manley και έχει ως στόχο να υπολογίζει την αποδοτικότητα τόσο μίας ομάδας όσο και ατομικά κάθε παίχτη. Ο τύπος υπολογισμού του δείκτη είναι ο εξής:

$$\mathbf{(PTS + REB + AST + STL + BLK - Missed\ FG - Missed\ FT - TOV) / GP}$$

Εφόσον τα δεδομένα που επεξεργαζόμαστε στην παρούσα εργασία αναφέρονται όλα στα στατιστικά των ομάδων ανά αγώνα, θα χρησιμοποιηθεί ο παρών τύπος τροποποιημένος ως εξής:

$$\mathbf{PPG + RPG + APG + SPG + BPG - MisFG - MisFT - TOV}$$

Τέλος, παράλληλα με το excel για την κανονική περίοδο της διοργάνωσης δημιουργήθηκε ένα ακόμα για τα στατιστικά των αγώνων που έδωσαν οι ομάδες που προκρίθηκαν στα playoffs.

2.3. Βιβλιογραφική επισκόπηση

Σ' αυτή την ενότητα γίνεται αναφορά σε υπάρχουσες επιστημονικές έρευνες οι οποίες πραγματοποιήθηκαν παλαιότερα και είναι σχετικές με το θέμα της παρούσας εργασίας και των δεδομένων της.

Οι Casals και Martinez (2013) αποσκοπούσαν να ερευνήσουν τους κύριους παράγοντες που επηρεάζουν την απόδοση ενός παίκτη και χρησιμοποιώντας στατιστικά μοντέλα ήθελαν να ερμηνεύσουν την συνεισφορά τους προκειμένου να εξηγήσουν δύο αποτελέσματα: τους πόντους και τις νίκες που παράγουν (win score). Ο τύπος των νικών που παράγουν έχει δοθεί από τον David Berri και ισούται με Win Score = Points + Rebounds + Steals + 0.5 Assists + 0.5 Blocks – Turnovers – Field Goals Attempted – 0.5 Fouls – 0.5 Free Throws Attempted. Τα δεδομένα που χρησιμοποιήθηκαν ήταν 2187 παραδείγματα αποτελούμενα από είκοσι επτά (27) παίκτες και ογδόντα ένα (81) παιχνίδια, όπου χρησιμοποιώντας μικτά μοντέλα κατέληξαν στα παρακάτω συμπεράσματα. Για τους πόντους που πετύχαινε ο κάθε παίχτης σημαντικές μεταβλητές κρίθηκαν τα λεπτά συμμετοχής, το ποσοστό συμμετοχής του παίκτη (usage percentage[3]) και η διαφορά στην ποιότητα της ομάδας, ενώ για τις νίκες που παράγουν εκτός των παραπάνω χρήσιμη θεωρήθηκε και η αλληλεπίδραση μεταξύ της θέσης τους παίκτη και της ηλικίας του.

Οι Kubatko, Oliver, Pelton και Rosenbaum (2007) παρουσίασαν διάφορους τύπους για στατιστικά που δεν είναι ευρέως διαδεδομένα στον χώρο ώστε να είναι μελλοντικά διαθέσιμα προς ανάλυση. Αρχικά όρισαν τις κατοχές που κερδίζει η ομάδα στον αγώνα και υπέβαλαν διάφορους τύπους υπολογισμού διότι θεωρείται μια από τις πιο σημαντικές μεταβλητές που τίθεται πλέον προς ανάλυση. Παρόμοια διεργασία ακολούθησαν και με τις έννοιες offensive και defensive ratings, plays, effective field goals κ.ο.κ.

Οι Teramoto και Croos (2010) ερεύνησαν τον ισχυρισμό που αναφέρει πως η πορεία των ομάδων διαφέρει ανάμεσα στη κανονική περίοδο και τα playoffs. Συγκεκριμένα αναλύθηκαν τα στατιστικά δεδομένα, ως μεταβλητές χρησιμοποιήθηκαν τα offensive rating, defensive rating και Four Factors, από την σεζόν 1999-2000 έως και την σεζόν 2008-2009. Μέσω πολυωνυμικής γραμμικής παλινδρόμησης και λογιστικής παλινδρόμησης οι ερευνητές κατέληξαν στο συμπέρασμα πως παρόλο που και στις δύο φάσεις του πρωταθλήματος τα offensive και defensive ratings είναι απαραίτητα για την νίκη μίας ομάδας, στην φάση των playoffs φαίνεται πως ο δεύτερος παράγοντας παίζει σημαντικότερο ρόλο απ' ότι στη κανονική περίοδο. Ακόμα, παρατηρήθηκε πως τα επιτυχημένα ποσοστά σε σουτ καθώς και τα μικρά ποσοστά σε λάθη μπορούν να αποτελέσουν σημαντικό παράγοντα στην ανάδειξη του νικητή ενός αγώνα στη κανονική περίοδο. Τέλος, τα rebound παίζουν κυρίως σημαντικό ρόλο στην έκβαση των τελικών των περιφερειών, όταν οι δύο ομάδες έχουν περίπου κοινό ποσοστό σε επιτυχημένα ποσοστά σουτ και λάθη.

Ο Suumers (2013) αναζήτησε τον τρόπο που μπορεί κάποια ομάδα να νικήσει τους αγώνες των playoffs του NBA. Τα δεδομένα απαρτίζονταν από τα στατιστικά δεδομένα

[3] Usage percentage είναι ένα ποσοστό των φάσεων της ομάδας που συμμετέχει ο παίχτης όσο βρίσκεται στον αγωνιστικό χώρο και υπολογίζεται ως εξής: Usage percentage = 100 * ((Field goals attempted + 0.44 * Free throw attempted + Turnovers) * (Team minutes played/ 5)) / (Player minutes played * (Team Field goals attempted + 0.44 * Team Free throw attempted + Team Turnovers)). Προσβάσιμο στο: https://www.dailyfantasycafe.com/academy/nba/advanced-statistics-breakdown [Ανακτήθηκε: 26-7-2023].

που ήταν διαθέσιμα για κάθε ομάδα που συμμετείχε στα playoffs του 2012, όπου συνολικά πραγματοποιήθηκαν ογδόντα τέσσερις (84) αγώνες. Έπειτα από την έρευνα, με την μέθοδο πολυωνυμικής παλινδρόμησης που διεξάχθηκε κατέληξε στους τρεις σημαντικότερους παράγοντες που επηρεάζουν την νίκη της ομάδας στα playoffs, οι οποίοι ήταν η έδρα, τα ποσοστά στα δίποντα ή και τρίποντα (field goals) και τα αμυντικά ριμπάουντ. Επίσης, σε διάφορα επίπεδα σημαντικότητας βρήκε πως η νίκη επηρεάζεται από τις ασίστ, τα ριμπάουντ, τα κοψίματα, τις εύστοχες ελεύθερες βολές και τα προσωπικά φάουλ. Τέλος, κατέληξε στον ισχυρισμό πως οι ομάδες που αγωνίζονται στην έδρα τους χρεώνονται λιγότερα φάουλ απ' τις φιλοξενούμενες ομάδες.

Οι Puente, Coso, Salinero και Abián-Vicén (2015) ερεύνησαν στατιστικά από το πρωτάθλημα ACB, το πρωτάθλημα καλαθοσφαίρισης ανδρών της Ισπανίας γνωστό και ως Liga Endesa, τα οποία αναφέρονταν στις σεζόν 2003 έως 2013, με σκοπό να ανακαλύψουν ποιοι δείκτες επηρεάζουν περισσότερο την ομάδα ώστε να καταφέρει να φτάσει στην νίκη του αγώνα στη κανονική περίοδο. Μέσω πολυωνυμικής παλινδρόμησης κατέληξαν στο συμπέρασμα πως οι κύριοι δείκτες που επηρεάζουν τις νίκες είναι τα ποσοστά στις ελεύθερες βολές, στα δίποντα και στα τρίποντα, τα επιθετικά και αμυντικά ριμπάουντ, οι ασίστ, τα κλεψίματα, τα λάθη και τα κοψίματα που δέχεται μία ομάδα από τους αντιπάλους της. Πιο συγκεκριμένα τα ποσοστά των εκτελέσεων των σουτ που αναφέρθηκαν καταφέρνουν να εξηγήσουν 26% της ανερμήνευτης μεταβλητότητας. Στην συνέχεια τα ριμπάουντ εξηγούν το υπόλοιπο 23%, τα κλεψίματα το 9%, τα λάθη το 7%, τα κοψίματα που δέχεται η ομάδα και οι ασίστ από 6% έκαστο. Προφανώς το υπόλοιπο 24% είναι η ανερμήνευτη μεταβλητότητα. Τέλος, με το πέρας της σεζόν χωρίστηκαν οι ομάδες σε 3 κατηγορίες. Πρώτη ήταν οι οκτώ (8) πρώτες ομάδες, η δεύτερη αποτελούνταν από τις επόμενες οχτώ (8) και τέλος οι δύο (2) τελευταίοι της κατηγορίας, που την επόμενη σεζόν θα αγωνίζονταν στην κατώτερη κατηγορία. Ανάμεσα σε αυτές τις κατηγορίες οι ερευνητές παρατήρησαν πως η πρώτη είχε σημαντικά μεγαλύτερες τιμές ανάμεσα στις μεταβλητές πόντων, εύστοχων τριπόντων και των ποσοστών των εύστοχων τριπόντων, ποσοστών ευστοχίας δίποντων, συνολικών ριμπάουντ, συνολικών ασίστ, συνολικών κοψιμάτων και καρφωμάτων, ενώ λάμβανε μικρότερες τιμές στις μεταβλητές των συνολικών πόντων, των κοψιμάτων που δεχόταν και των λαθών. Επίσης σε σύγκριση ανάμεσα στην δεύτερη και τρίτη κατηγορία παρατηρήθηκε πως ο αριθμός των μεταβλητών των πόντων που δεχόταν μια ομάδα, των ποσοστών ευστοχίας στις ελεύθερες βολές, στα δίποντα, τα τρίποντα και οι ασίστ ήταν διαφορετικός.

Ο Jones (2007) επιχείρησε να αναλύσει τον τρόπο που επηρεάζει η έδρα την έκβαση του αγώνα. Τα στατιστικά δεδομένα που χρησιμοποιήθηκαν είναι από παιχνίδια δύο (2) σεζόν, τις 2002-2003 και 2003-2004, του NBA χωρισμένα ανά περιόδους και σε παρατάσεις σε περίπτωση ισοπαλίας στην κανονική διάρκεια του αγώνα. Αρχικά, παρατηρήθηκε πως τη σεζόν 2002-2003 η ομάδα που έπαιζε εντός έδρας σκόραρε 3,89 πόντους περισσότερους από τους φιλοξενούμενους και κέρδιζε το 62,9% των αγώνων, ενώ την σεζόν 2003-2004 σκόραρε 3,59 πόντους περισσότερο και κέρδιζε το 61,3% των αγώνων. Ακόμα, ο Jones κατέληξε πως η έδρα παίζει σημαντικό ρόλο στην πρώτη

περίοδο, εφόσον η διαφορά στο σκοράρισμα των δύο ομάδων είναι σημαντική, και με την πάροδο του χρόνου επηρεάζει όλο και λιγότερο τη διαφορά ανάμεσα στους πόντους που σκοράρουν οι δύο ομάδες. Επίσης, κατάφερε να αξιολογήσει τον τρόπο που αντιδρούν οι παίκτες κατά τη διάρκεια του αγώνα αναλόγως του σκορ του. Τα αποτελέσματα στα οποία κατέληξε είναι ξεκάθαρα και δείχνουν πως σε αγώνες που η ομάδα προηγείται των αντιπάλων αναμένεται στην περίοδο που θα ξεκινήσει να μειωθεί η διαφορά καθώς θα δεχθούν περισσότερους πόντους από τους αντιπάλους. Αν η ομάδα εντός έδρας έχανε από τους φιλοξενούμενους τότε αναμενόταν να αντιδράσει και να έχει μεγάλο πλεονέκτημα πόντων στη δεδομένη περίοδο που θα ξεκινούσε. Τέλος, σε περίπτωση που η περίοδος ξεκινούσε με τις δύο ομάδες ισόπαλες, τότε αναμενόταν ξανά να υπάρχει πλεονέκτημα πόντων της ομάδας που βρίσκεται εντός έδρας.

Ο Çene (2018) με την έρευνά του ήθελε να καθορίσει τους παράγοντες που είχαν τη μεγαλύτερη επιρροή στην έκβαση ενός αγώνα για τη σεζόν 2016-2017 της Euroleague, η οποία αποτελεί την κορυφαία διασυλλογική διοργάνωση καλαθοσφαίρισης στην Ευρώπη, γνωστή και ως «Turkish Airlines EuroLeague» για χορηγικούς λόγους. Τα δεδομένα αποτελούνταν από 259 αγώνες από τις φάσεις της κανονικής περιόδου, playoffs και των Final fours. Η φάση των Final four -που δεν υφίσταται στην διοργάνωση του NBA- αποτελεί ένα τουρνουά, συνήθως δύο αγώνων αποκλεισμού (ημιτελικοί και τελικοί), στο οποίο συμμετέχουν οι τέσσερις (4) κορυφαίες ομάδες που έχουν καταφέρει να διακριθούν στα playoffs. Σε ορισμένα τουρνουά οι δύο ηττημένες ομάδες των πρώτων αγώνων (ημιτελικοί) διαγωνίζονται σε ένα παιχνίδι παρηγοριάς, για να κριθεί ο τρίτος της κατηγορίας. Η ομάδα που θα κατορθώσει να φέρει στα δύο παιχνίδια ισάριθμες νίκες είναι και αυτή που χρίζεται πρωταθλήτρια,.

Αρχικά ο Çene μέσω της συσταδοποίησης, πιο συγκεκριμένα του αλγορίθμου kmeans, διαχώρισε τα παιχνίδια σε τρεις (3) κατηγορίες ανάλογα με τη διαφορά των πόντων ανάμεσα στις ομάδες. Η πρώτη κατηγορία αποτελούνταν από τα παιχνίδια που είχαν διαφορές ως -το πολύ- δέκα (10) πόντων, η δεύτερη τους αγώνες με διαφορές από δέκα (10) έως είκοσι ένα (21) πόντων και η τρίτη τα μη ισορροπημένα παιχνίδια, δηλαδή αυτά που η διαφορά πόντων ξεπέρναγε τους είκοσι ένα (21). Μέσω του correlation matrix βρέθηκε υψηλή συσχέτιση, $\rho>0,4$, ανάμεσα στην έκβαση των αγώνων και των μεταβλητών “FG%”, “FG” και “TS”. Επίσης, μέσω περιγραφικής στατιστικής, t-test και Cohen’s, ο Çene βρήκε πως τα ποσοστά στα σουτ, τα κλεψίματα και τα αμυντικά ριμπάουντ είναι στατιστικά σημαντικοί παράγοντες για όλες τις κατηγορίες αγώνων, ενώ τα κοψίματα βρέθηκαν πως είναι μη σημαντικά για τους αγώνες της πρώτης κατηγορίας και των μη ισορροπημένων αγώνων. Ακόμα, κατάφερε μέσω της μεθόδου BMA να καταλήξει στις μεταβλητές, αμυντικά ριμπάουντ, κλεψίματα, φάουλ που έκανε η ομάδα, φάουλ που κέρδισε η ομάδα, λάθη, πραγματικό ποσοστό εύστοχων σουτ και επιθετικά ριμπάουντ, που επηρεάζουν περισσότερο την έκβαση των αγώνων. Στη συνέχεια μέσω αλγορίθμων κατηγοριοποίησης (classification) κατάφερε να κατηγοριοποιήσει όλους τους αγώνες μαζί με ποσοστό ακρίβειας 80%. Συγκεντρωτικά, ομάδες με χαμηλό πραγματικό ποσοστό εύστοχων

σουτ, λίγα αμυντικά ριμπάουντ και κλεψίματα είναι πιθανότερο να είναι οι ηττημένες του αγώνα.

Οι Mandić, Erčulj, Jakovljevic και S˘trumbeljI (2019) επιχείρησαν να αναλύσουν και να συγκρίνουν τα στατιστικά δεδομένα του NBA και της Euroleague για τις σεζόν 2000 έως και 2017, τόσο της κανονικής περιόδου όσο και των playoffs. Αρχικά, μέσω περιγραφικής στατιστικής, διαπίστωσαν πως οι κατοχές με το πέρασμα του χρόνου όλο και αυξάνονταν, όπως αναφέρεται, δικαιολογημένα διότι στην Euroleague μειώθηκε ο χρόνος επίθεσης από 30 δευτερόλεπτα σε 24, ενώ και στο NBA υπήρχαν αλλαγές στα φάουλ και σε άλλα σημεία. Στη συνέχεια για το rotation των παικτών που χρησιμοποιούν οι ομάδες φάνηκε πως στη Euroleague είναι σημαντικότερο και συχνότερο, γεγονός που δικαιολογεί την μεγαλύτερη ένταση στον αγώνα των ομάδων. Επίσης, όσον αφορά τα ποσοστά ελεύθερων βολών, που το σημείο εκτέλεσης βρίσκεται στην ίδια απόσταση από το καλάθι και για τα δύο (2) πρωταθλήματα από το 1895, είναι υψηλότερα στο NBA με διαφορά της τάξης του 2,3%. Σε ό,τι έχει να κάνει με τα δίποντα τα ποσοστά των ομάδων στη Euroleague είναι υψηλότερα, ενώ οι προσπάθειες που επιχειρούν οι παίκτες στους αγώνες είναι χαμηλότερες απ' αυτές που επιχειρούν στο NBA. Αντίστοιχα, ποσοστά ευστοχίας για τα τρίποντα, έχουμε περισσότερες προσπάθειες στο ευρωπαϊκό πρωτάθλημα, ενώ και τα ριμπάουντ δεν διαφέρουν σημαντικά. Όσον αφορά τις ασίστ και τα κοψίματα είναι φανερό πως το NBA επικρατεί, ενώ στα κλεψίματα και τα λάθη δεν υπερτερεί έναντι του αντίστοιχου ευρωπαϊκού πρωταθλήματος. Τέλος, ερευνήθηκαν τα στατιστικά δεδομένα ανάμεσα στη regular season και τα playoffs. Για το NBA φάνηκε πως υπήρχε μείωση στις κατοχές ανά αγώνα, στις προσπάθειες σουτ δύο πόντων, στις ασίστ και στα λάθη ανά εκατό κατοχές από την κανονική περίοδο στα playoffs. Αντίθετα, υπήρξε αύξηση στις προσπάθειες σουτ τριών πόντων, ελεύθερων βολών και των φάουλ που έκανε κάθε ομάδα.

Οι Santos, Wang, Carlsson και Lambrix (2021) προσπάθησαν να προβλέψουν μέσω των στατιστικών δεδομένων κάθε ομάδας το αποτέλεσμα ενός αγώνα και μιας ολόκληρης σεζόν του NBA. Τα δεδομένα αποτελούνταν από τα στατιστικά των ομάδων από δέκα (10) season και μέσω της διαχείρισης τους κατέληξαν σε μία προσέγγιση με ακρίβεια της τάξης του 69,88%. Επίσης, αποδεικνύεται πως ανάμεσα στους πιο σημαντικούς παράγοντες για την εύρεση του αποτελέσματος είναι η απόδοση κατά την προηγούμενη σεζόν των παιχτών οι οποίοι παρέμειναν στην ομάδα, οι νίκες και ήττες της ομάδας τους τελευταίους δεκαπέντε (15) αγώνες και των προηγούμενων σεζόν, τα offensive και defensive performance των προηγούμενων σεζόν και οι αποδόσεις όλων των παιχτών ανεξαρτήτως αν αυτοί υπέγραψαν ή έφυγαν από την ομάδα.

Οι Nguyena, Nguyenb, Maa και Hua (2022) επιδίωξαν να προβλέψουν την μελλοντική απόδοση των παιχτών και την πιθανότητα επιλογής τους στο All-Star Game μέσω εφαρμογής machine learning και τους παράγοντες που επηρεάζουν αυτό το γεγονός. Το All-Star Game είναι μια διοργάνωση που διαδραματίζεται μια φορά κάθε σεζόν και συμμετέχουν οι αστέρες του NBA, οι οποίοι προκύπτουν έπειτα από ψηφοφορία ανάμεσα στους ίδιους τους παίκτες, τους δημοσιογράφους και στους

τηλεθεατές. Συνήθως διοργανώνεται κάθε χρόνο σε διαφορετική τοποθεσία και διεξάγεται σε τρεις μέρες (Παρασκευή-Σάββατο-Κυριακή). Όπως είναι οροθετημένο, την Παρασκευή γίνεται ο αγώνας ανάμεσα στους νεότερους αστέρες του NBA, το Σάββατο λαμβάνουν μέρος οι αγώνες δεξιοτήτων και την Κυριακή είναι ο αγώνας ανάμεσα στους αστέρες των περιφερειών του NBA, όπως αυτοί εκλέγονται έπειτα από την ψηφοφορία. Τελικώς, αποδείχθηκε, τόσο μέσω της ανάλυσης παλινδρόμησης όσο και της κατηγοριοποίησης, πως οι πόντοι είναι ο κύριος παράγοντας επιρροής και έπειτα ακολουθούν τα αμυντικά ριμπάουντ.

ΚΕΦΑΛΑΙΟ 3ο

3. Περιγραφική στατιστική

Στο παρόν κεφάλαιο της εργασίας παρουσιάζονται αναλυτικότερα κάποια περιγραφικά μέτρα και διάφορα διαγράμματα που σχετίζονται με τις μεταβλητές των δεδομένων μας. Αρχικά, παρουσιάζονται διαγραμματικές απεικονίσεις των δεδομένων και για τις δύο φάσεις του πρωταθλήματος. Έπειτα, ακολουθούν τρεις αναλύσεις που βασίζονται στις κατηγορικές μεταβλητές που έχουν συλλεχθεί στα δεδομένα. Πρώτος διαχωρισμός είναι ανάμεσα τις ομάδες που πέρασαν στα playoffs και αυτές που δεν τα κατάφεραν, ο δεύτερος συγκρίνει τις ομάδες σύμφωνα με το πρωτάθλημα περιφέρειας που συμμετέχουν, East και West, και τέλος γίνεται ένας γενικός απολογισμός όλων των ομάδων για τα τελευταία δεκαπέντε (15) χρόνια που αναλύουμε.

3.1. Διερεύνηση για ελλιπή δεδομένα

Για κάθε έρευνα που γίνεται σε μεγάλο όγκο δεδομένων θα πρέπει να γίνει και ο κατάλληλος έλεγχος για την ύπαρξη ελλιπών δεδομένων, γνωστών ως missing values. Τα missing values εμφανίζονται συνήθως όταν στις τιμές κάποιας μεταβλητής δεν έχει αποθηκευτεί καμία τιμή. Η επίδραση τους είναι αρκετά σημαντική διότι μπορεί να εξαχθούν λανθασμένα δεδομένα, να προκαλέσει σημαντική μεροληψία και να κάνει δυσκολότερη την ανάλυση των δεδομένων. Παρόλο που τα δεδομένα σχετίζονται και με την περίοδο του COVID-19, όπου και τα περισσότερα πρωταθλήματα παγκοσμίως διακόπηκαν οριστικά, δεν έχουμε ενδείξεις για ελλιπή δεδομένα. Όπως αναφέρθηκε και σε προηγούμενη παράγραφο, παρόλο που την σεζόν 2019-2020 αναβλήθηκαν ορισμένοι αγώνες, το πρωτάθλημα ολοκληρώθηκε κανονικά και τα δεδομένα διατηρήθηκαν και δόθηκαν προς ανάλυση.

Σε περίπτωση εύρεσης ελλιπών δεδομένων θα έπρεπε να υπάρχει και ανάλογη διαχείριση τους. Ένας τρόπος αντιμετώπισης, ο οποίος δεν προτείνεται, είναι να συμπληρωθούν οι τιμές χειρωνακτικά. Εναλλακτικά, επιτρέπεται να συμπληρωθούν οι τιμές αυτόματα με χρήση είτε της μέσης τιμής του γνωρίσματος για όλα τα όμοια δεδομένα είτε με χρήση της πιο πιθανής τιμής. (Πελέκης, 2022)

Στις ακόλουθες παραγράφους που ακολουθούν γίνεται παρουσίαση των περιγραφικών μέτρων και των διαγραμμάτων των ποσοτικών μεταβλητών, που περιέχονται στα δεδομένα, σύμφωνα με τις αναλύσεις που πραγματοποιήθηκαν.

3.2. Χαρακτηριστικά των μεταβλητών για την Κανονική Περίοδο

Με την ανάλυση των δεδομένων, που συλλέχτηκαν για την κανονική περίοδο όλων των τελευταίων δεκαπέντε (15) χρόνων, ερευνώνται τα περιγραφικά μέτρα, η μεταβολή των χαρακτηριστικών με το πέρασμα του χρόνου και η σχέση τους με την μεταβλητή EFF των ομάδων.

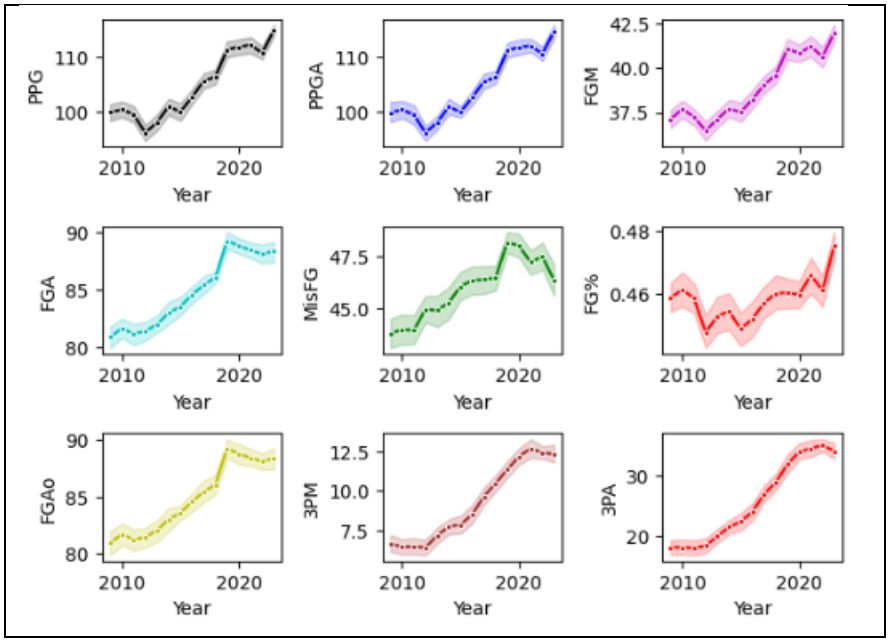
Κανονική περίοδος	Min	1st. Q	Median	Mean	3rd. Q	Max
PPG	87.00	98.92	104.05	104.68	110.60	120.70
PPGA	88.20	99.325	104.45	104.69	110.00	123.10
FGM	33.20	37.20	38.70	38.89	40.60	44.70
FGA	75.80	82.0	85.10	84.48	87.57	94.40
MisFG	38.20	44.49	46.00	45.94	47.69	51.40
FG%	0.40	0.44	0.45	0.45	0.46	0.50
FGAo	75.30	82.20	84.90	84.85	87.70	94.00
3PM	3.80	6.80	9.00	9.22	11.37	16.70
3PA	11.30	19.30	25.20	25.78	31.80	45.40
3P%	0.29	0.34	0.35	0.30	0.36	0.41
FTM	12.20	16.40	17.55	17.65	18.70	24.10
FTA	16.60	21.40	23.0	23.08	24.50	31.10
MisFt	3.09	4.70	5.39	5.42	5.90	9.00
FT%	0.66	0.74	0.76	0.76	0.78	0.83
FTAo	18.20	21.50	23.00	23.07	24.50	30.20
ORB	7.60	9.70	10.50	10.56	11.40	14.60
DRB	27.20	31.10	32.70	32.61	34.10	42.20
RPG	36.90	41.70	43.10	43.17	44.50	51.70
ORBo	8.00	9.90	10.60	10.57	11.10	14.20
APG	17.40	21.10	22.70	22.84	24.30	30.40
SPG	5.50	7.00	7.50	7.58	8.20	10.00
BPG	2.40	4.30	4.80	4.85	5.30	8.20
TOV	11.10	13.50	14.20	14.24	15.00	17.70
TOVo	11.30	13.42	41.20	14.24	15.00	18.40
PF	16.60	19.20	20.30	20.23	21.20	24.80
POSSt	87.78	93.40	96.36	96.31	98.96	104.99
ORt	94.65	105.32	108.18	108.61	111.76	119.78
POSSo	87.88	93.40	96.36	96.31	98.96	104.99
DRt	97.86	105.77	108.70	108.61	111.46	120.77
DEFF	-15.32	-3.28	0.35	-0.03	3.53	12.04
EFF	90.40	109.50	116.60	117.52	124.87	143.50

Πίνακας 3.1: Στατιστικά περιγραφικά μέτρα για την κανονική περίοδο

Στον Πίνακα 3.1 δίνονται η ελάχιστη τιμή, το πρώτο τεταρτημόριο, η διάμεσος, η μέση τιμή, το τρίτο τεταρτημόριο και η μέγιστη τιμή όλων των ποσοτικών μεταβλητών. Άξιο σχολιασμού είναι το γεγονός πώς η μέση τιμή των πόντων που δέχονται οι ομάδες κατά μέσο όρο στις σεζόν, 104.45, είναι μεγαλύτερη από την αντίστοιχη των πόντων που πετυχαίνουν, 104.05. Ακόμα, είναι φανερό πως το DEFF των ομάδων έχει

αρνητική μέση τιμή, -0.03, κάτι το οποίο μας δείχνει, έστω και σε μικρό βαθμό, πως τα defensive ratings είναι χειρότερα από τα αντίστοιχα offensive σε γενική εικόνα. Αυτό γίνεται αντιληπτό διότι από τον τύπο υπολογισμού μια ομάδα με καλή αμυντική λειτουργία θα έπρεπε να έχει μικρές τιμές του δείκτης DRt, εφόσον υπολογίζεται με βάση τους πόντους που δέχεται μια ομάδα.

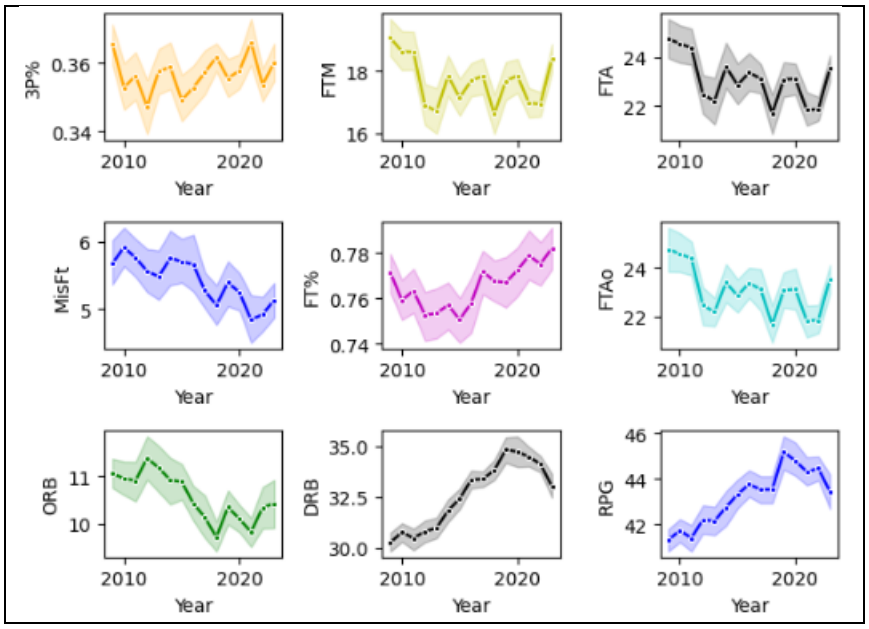
Για την παρουσίαση δεδομένων, ώστε να γίνει αντιληπτή η μεταβολή των τιμών των μεταβλητών στο πέρασμα του χρόνου ιδανικό στατιστικό γράφημα αποτελούν τα time series plots. Τα γραφήματα χρονοσειρών παρουσιάζουν τις τιμές των δεδομένων με την πάροδο του χρόνου. Είναι χρήσιμα διότι μπορούν να παρουσιάσουν την τάση των δεδομένων, πιθανά μοτίβα που ακολουθούν οι τιμές τους και να αποτελέσουν όργανο για τον έλεγχο της τυχαιότητας των σημείων. (Καμίτσης Α., 2023)



Σχήμα 3.1: Time Series Plots ορισμένων μεταβλητών της κανονικής περιόδου με την πάροδο των σεζόν

Στο Σχήμα 3.1 διακρίνονται τα γραφήματα των πρώτων εννέα (9) μεταβλητών του αρχείου σε σχέση με την πάροδο του χρόνου. Όπως φαίνεται στον οριζόντιο άξονα των γραφημάτων, οι τιμές ξεκινούν από το 2009, εφόσον είναι η κωδική ονομασία της πρώτης σεζόν 2008-2009, και τελειώνουν στην τιμή 2023. Όσον αφορά την ανάλυση των γραφημάτων είμαστε σε θέση να εξάγουμε ορισμένα αποτελέσματα για τον τρόπο που άλλαξαν οι μεταβλητές. Συγκεκριμένα, οι μεταβλητές των πόντων που σκοράρει κάθε ομάδα (PPG), των πόντων που δέχεται (PPGA), των διπόντων ή τριπόντων που επιτυγχάνει (FGM) και επιχειρεί τόσο η ίδια (FGA) όσο και η αντίπαλη ομάδα (FGAo) και των τριπόντων που επιχειρεί (3PA) και επιτυγχάνει (3PM), διακρίνεται πως έχουν μια αυξητική πορεία με την πάροδο του χρόνου. Τα συμπεράσματα πολύ πιθανόν να οφείλονται σε μία από τις μεγαλύτερες αλλαγές στους κανόνες του NBA έως και σήμερα, κατά την οποία δεν επιτρέπεται στους επιτιθέμενους παίχτες να βρίσκονται

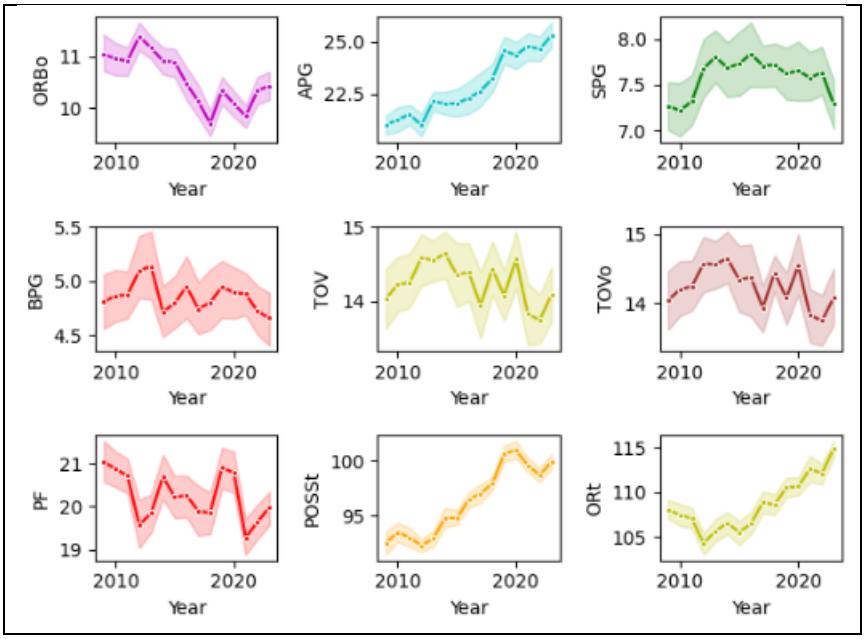
για πάνω από τρία (3) δευτερόλεπτα στη ρακέτα[4] της αντίπαλης ομάδας. Κατά συνέπεια, ο αμυντικός δεν χρειαζόταν να βρίσκεται σε επαφή με τον επιθετικό συνεχώς και θα μπορούσε να δίνει βοήθειες, πιέζοντας άλλους επιτιθέμενους, κοντά στο καλάθι. Κρίνοντας από το γράφημα των πόντων που σκοράρουν οι ομάδες είναι φανερό πως τις πρώτες έξι (6) σεζόν οι τιμές είναι αρκετά χαμηλές και έπειτα έχουμε την απρόσμενη αύξηση. Παρόλο, λοιπόν, που ο συγκεκριμένος κανόνας εισήχθη τη σεζόν 2001-2002, η πραγματική επιρροή φάνηκε από την σεζόν 2014-2015 και έπειτα, όπου οι προπονητές αναγνώρισαν τις αδυναμίες του κανόνα και άλλαξαν προσανατολισμό στη δημιουργία των ομάδων με παίχτες οι οποίοι ήταν ικανοί στο μακρινό παιχνίδι ώστε να απομακρύνουν πιθανούς αντιπάλους από το εσωτερικό της ρακέτας. Για την μεταβλητή των ποσοστών επιτυχημένων διπόντων ή τριπόντων (FG%) μια αυξομείωση στις τιμές της με την πάροδο του χρόνου.



Σχήμα 3.2: Time Series Plots των υπόλοιπων μεταβλητών για την κανονική περίοδο με την πάροδο των σεζόν

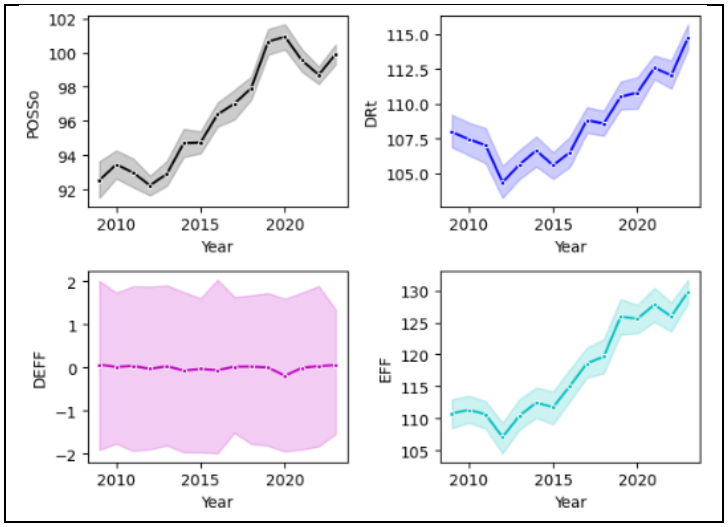
Στο Σχήμα 3.2, στο οποίο παρουσιάζονται τα γραφήματα χρονοσειρών των επόμενων εννέα (9) χαρακτηριστικών, παρατηρούμε πως υπάρχει αυξητική τάση για τις τιμές των μεταβλητών: αμυντικά ριμπάουντ (DRB) και συνολικά ριμπάουντ (RPG). Οι λόγοι που οδηγούν σε αυτή την τάση είναι οι ίδιοι που αναφέρθηκαν και προηγουμένως. Αντίθετα, για την μεταβλητή των επιθετικών ριμπάουντ (ORB), Σχήμα 3.2, παρατηρούμε μία απότομη μείωση με το πέρασμα του χρόνου, κάτι το οποίο είναι κατανοητό εφόσον οι επιθετικοί απομακρύνθηκαν από τη ρακέτα ώστε να δημιουργηθούν οι απαραίτητοι χώροι. Τέλος, για τις υπόλοιπες μεταβλητές παρατηρούμε μια αυξομείωση στις τιμές τους με την πάροδο του χρόνου.

[4] Ρακέτα, γνωστή και ως η λωρίδα ελεύθερων βολών, αποτελεί μία περιοχή στα γήπεδα μπάσκετ γύρω από το καλάθι. Συγκεκριμένα, ξεκινάει από την τελική γραμμή και φτάνει ως το ύψος των ελεύθερων βολών και μήκος ίσο με τις δύο πλάγιες γραμμές. Λόγω της κρισιμότητας της περιοχής και των κανόνων που ισχύουν σε αυτή βάφεται σε διαφορετικό χρώμα από το υπόλοιπο γήπεδο. Προσβάσιμο στο: https://en.wikipedia.org/wiki/Key_(basketball) [Ανακτήθηκε 1-8-2023].



Σχήμα 3.3: Time Series Plots επόμενων μεταβλητών για την κανονική περίοδο με την πάροδο των σεζόν

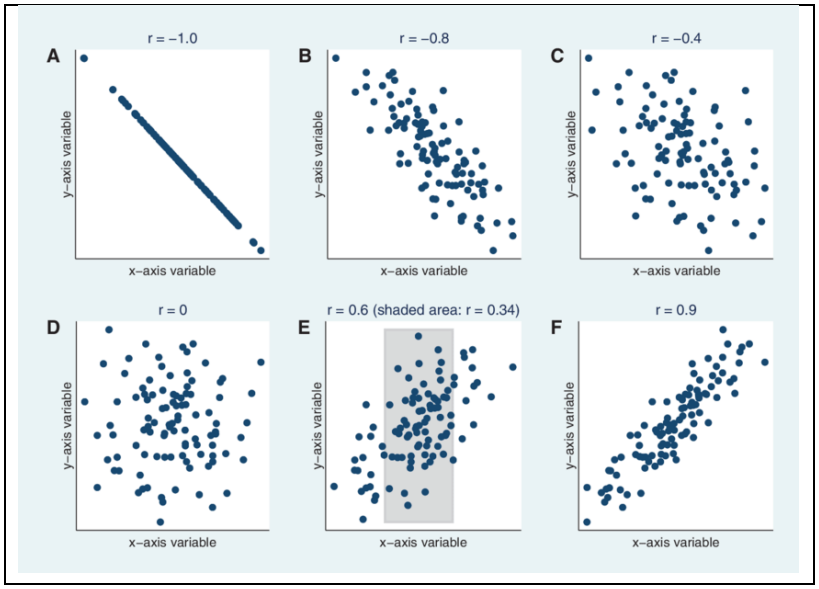
Στο Σχήμα 3.3, στο οποίο παρουσιάζονται τα γραφήματα χρονοσειρών των επόμενων εννέα (9) χαρακτηριστικών, παρατηρούμε πως υπάρχει αυξητική τάση για τις τιμές των μεταβλητών: ασίστ (APG) και κατοχές των ομάδων (POSSt). Αναφορικά με τις κατοχές ανά αγώνα στο ίδιο αποτέλεσμα είχαν καταλήξει στην έρευνά τους και οι Mandić (et al. 2005) όπου παρατήρησαν την αύξηση των κατοχών ανά χρονιά από την σεζόν 2000 έως 2017 τόσο στην διοργάνωση του NBA όσο και στην Euroleague. Ακόμα, για τις κατοχές των ομάδων παρατηρούμε πως τη σεζόν 2017-2018 υπάρχει ένα απότομο αυξητικό άλμα το οποίο ενδεχομένως να οφείλεται στον νέο κανόνα που εισήχθη εκείνη τη σεζόν. Σύμφωνα με αυτόν τον κανόνα έπειτα από επιθετικό ριμπάουντ ο χρόνος ανανέωσης της επίθεσης ορίζεται στα δεκατέσσερα (14) δευτερόλεπτα έναντι των είκοσι τεσσάρων (24) που ήταν μέχρι τότε γεγονός που επέφερε περισσότερες κατοχές σε κάθε ομάδα. Από την άλλη πλευρά, οι τιμές των επιθετικών ριμπάουντ των αντίπαλων ομάδων (ORBo) ακολουθούν πτωτική πορεία, όπως και τα επιθετικά των ομάδων που είδαμε στο προηγούμενο σχήμα.



Σχήμα 3.4: Time Series Plots τελευταίων τεσσάρων μεταβλητών

Επίσης, στο Σχήμα 3.4 παρουσιάζονται τα γραφήματα των τεσσάρων τελευταίων μεταβλητών. Απ' αυτά παρατηρούμε πως το γράφημα της μεταβλητής DEFF φαίνεται να μην έχει τάση και να αποτελεί μια στάσιμη χρονοσειρά με τιμές γύρω από το μηδέν (0). Τέλος, οι υπόλοιπες τρεις μεταβλητές του σχήματος φαίνεται πως έχουν αυξητική τάση με το πέρας του χρόνου.

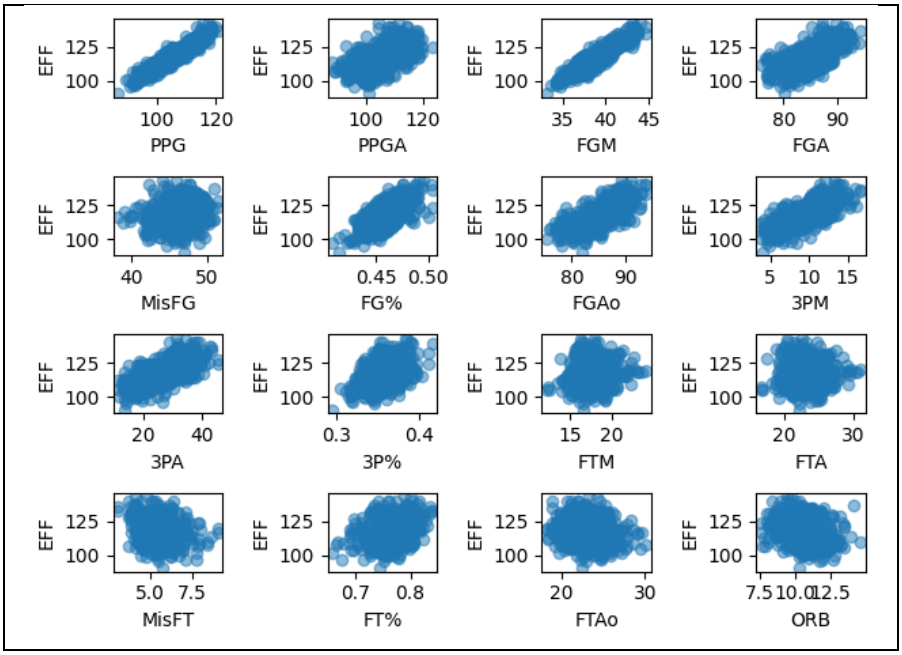
Στη συνέχεια της μελέτης θα αναπαρασταθούν τα διαγράμματα διασποράς, γνωστά και ως scatter plot, όλων των μεταβλητών σε σύγκριση με την μεταβλητή της αποδοτικότητας μίας ομάδας (EFF). Στα διαγράμματα διασποράς απεικονίζονται οι τιμές των παρατηρούμενων μεταβλητών που μας ενδιαφέρει να παρατηρήσουμε ώστε να μελετηθεί η συσχέτιση ανάμεσα τους. Οι σχέσεις που παρατηρούνται ανάμεσα στις μεταβλητές συνήθως χωρίζονται σε τρείς (3) κατηγορίες. Στην πρώτη κατηγορία ανήκουν οι μεταβλητές που έχουν θετική συσχέτιση και στην περίπτωση αυτή αναμένουμε να έχουμε αυξητική σχέση ανάμεσα στις τιμές των μεταβλητών. Όσο οι τιμές της μεταβλητής του άξονα x αυξάνονται, τόσο αυξάνονται και οι τιμές της μεταβλητής y. Αντίθετα, στη δεύτερη κατηγορία έχουμε αρνητική συσχέτιση ανάμεσα στις μεταβλητές και συνεπώς περιμένουμε μειωτική συμπεριφορά των τιμών. Τέλος, στην τελευταία κατηγορία υπάγονται οι ασυσχέτιστες μεταβλητές, όπου η αύξηση στις τιμές της μίας μεταβλητής δεν συνεπάγεται κάποια δεδομένη κίνηση στις τιμές της άλλης. Στο Σχήμα 3.5 δίνονται διάφορες τιμές του συντελεστή συσχέτισης και τα αντίστοιχα γραφήματα. Στα A, B, C γραφήματα παρατηρείται αρνητική συσχέτιση ανάμεσα στις μεταβλητές, στο D δύο ασυσχέτιστες μεταβλητές και στα E, F περιέχονται δύο θετικά συσχετισμένες μεταβλητές.



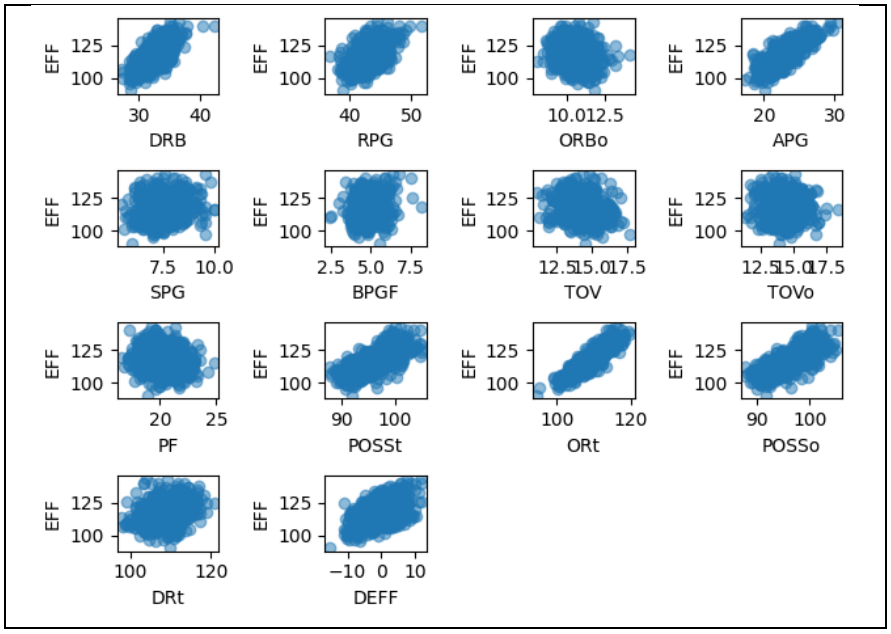
Σχήμα 3.5: Χαρακτηριστικά διαγράμματα διασποράς

(Πηγή: Schober et al., 2018)

Στο Σχήμα 3.6, όπου παρουσιάζονται οι συσχετίσεις των πρώτων δεκαέξι (16) μεταβλητών με τον δείκτη αποδοτικότητας (EFF), παρατηρούμε πως οι μεταβλητές των πόντων που βάζει μία ομάδα (PPG) και των εύστοχων δίποντων ή τρίποντων (FGM) έχουν ισχυρή θετική συσχέτιση με την μεταβλητή EFF. Στη συνέχεια, οι μεταβλητές πόντων που δέχεται η ομάδα (PPGA), ποσοστού εύστοχων (FG%) και προσπαθειών δίποντων ή τρίποντων (FGA) τόσο της ομάδας όσο και των αντιπάλων (FGAo), εύστοχων (3PM), προσπαθειών που επιχειρήθηκαν (3PA) και ποσοστών εύστοχων τρίποντων (3P%) έχουν θετική συσχέτιση με την μεταβλητή, όμως δεν είναι τόσο ισχυρή όπως προηγουμένως. Αντιθέτως, οι μεταβλητές χαμένων ελεύθερων βολών (MisFT) και επιτυχημένων ελεύθερων βολών (FTM) έχουν αρνητική συσχέτιση με τον δείκτη. Ακόμα οι υπόλοιπες μεταβλητές έχουν οριακές τιμές συντελεστών, επομένως δεν μπορούν να χαρακτηριστούν ως αρνητικές-θετικές.



Σχήμα 3.6: Scatter Plots ορισμένων μεταβλητών της κανονικής περιόδου σε σχέση με EFF



Σχήμα 3.7: Scatter Plots των υπόλοιπων μεταβλητών της κανονικής περιόδου σε σχέση με EFF

Αντίστοιχα, στο Σχήμα 3.7 οι μεταβλητές των αμυντικών ριμπάουντ, ασίστ και offensive rating έχουν ισχυρή θετική συσχέτιση με την ζητούμενη μεταβλητή. Οι μεταβλητές των συνολικών ριμπάουντ (RPG), των κατοχών των ομάδων (POSSt) είτε των αντιπάλων (POSSo) και ο δείκτης DEFF έχουν μικρότερη αλλά παραμένει θετική η συσχέτιση με τον συντελεστή. Επίσης, τα επιθετικά ριμπάουντ των αντιπάλων ομάδων (ORBo) και τα λάθη (TOV) φαίνεται πως έχουν αρνητική συσχέτιση με τον συντελεστή. Όλες οι υπόλοιπες μεταβλητές έχουν μικρή συσχέτιση με αυτόν.

3.3. Χαρακτηριστικά των μεταβλητών για τα Playoffs

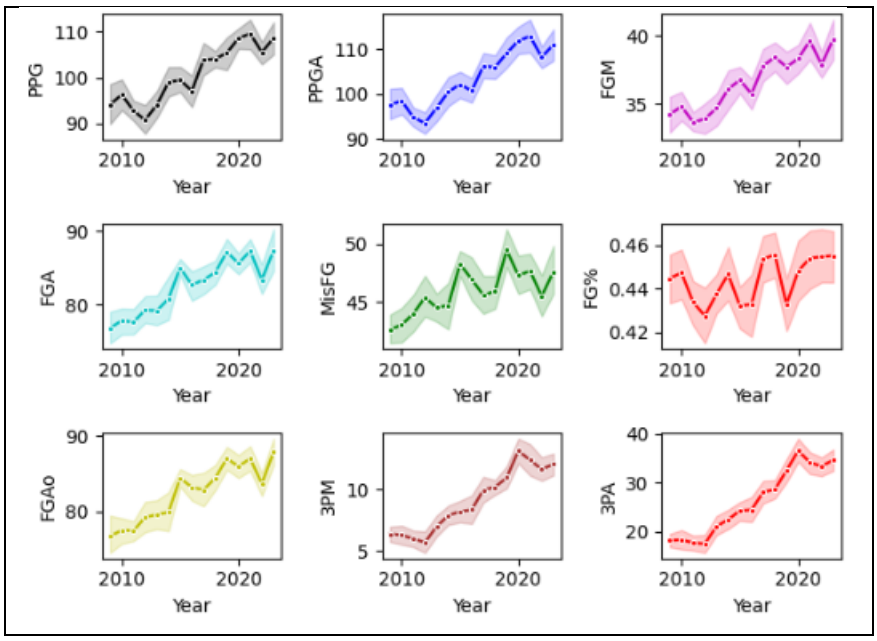
Όπως έχει αναφερθεί και σε προηγούμενη παράγραφο, για την παρούσα εργασία έχει δημιουργηθεί και ένα δεύτερο αρχείο excel το οποίο περιέχει τα στατιστικά των ομάδων που συμμετείχαν στους αγώνες των playoffs. Συνεπώς θα ακολουθήσει, με παρόμοια σειρά, παρουσίαση των περιγραφικών μέτρων και διαγραμματικές απεικονίσεις γι' αυτά τα δεδομένα. Ξεκινώντας, λοιπόν, από τον Πίνακα 3.2. παρουσιάζονται οι τιμές περιγραφικών μέτρων όπως αυτές υπολογίστηκαν. Σύμφωνα με τον πίνακα, είναι φανερό πως τα στατιστικά των μεταβλητών έχουν μειωθεί αρκετά σε σχέση με τους αγώνες που έχουν πραγματοποιηθεί την κανονική περίοδο. Παραδείγματος χάρη οι πόντοι των ομάδων έχουν μειωθεί, καθώς στα playoffs η μέση τιμή είναι 100.59 ενώ στον Πίνακα 3.1 η μέση τιμή που αντιστοιχεί σε εκείνη την περίοδο είναι ίση με 104.68. Η μεγαλύτερη διαφορά ωστόσο φαίνεται στο efficiency των ομάδων. Στη κανονική περίοδο, όπως είδαμε σε προηγούμενη παράγραφο, η μέση τιμή ήταν 117.52. Στα playoffs όμως η τιμή έχει πτωτική συμπεριφορά καθώς φτάνει να έχει μέση τιμή ίση με 106.43. Αυτή η συμπεριφορά ήταν αναμενόμενη διότι τα παιχνίδια των playoffs διακρίνονται για τις πιο δυνατές άμυνες που υπάρχουν, κάτι που οδηγεί σε λιγότερο θεαματικό μπάσκετ με αποτέλεσμα τη μείωση των τιμών των μεταβλητών που συμμετέχουν στον υπολογισμό του efficiency.

Playoffs	Min	1st. Q	Median	Mean	3rd. Q	Max
PPG	78.00	93.80	101.1	100.59	107.70	119.50
PPGA	85.90	96.47	103.30	103.34	109.32	126.70
FGM	27.80	34.30	36.70	36.62	38.80	44.20
FGA	69.00	78.47	82.85	82.55	86.02	96.80
MisFG	36.80	43.40	45.99	45.92	48.19	59.30
FG%	0.38	0.42	0.44	0.44	0.46	0.50
FGAo	68.80	78.27	82.80	82.47	86.60	94.50
3PM	2.30	6.60	8.80	9.06	11.30	18.00
3PA	10.80	19.30	25.54	26.16	32.72	46.80
3P%	0.20	0.32	0.34	0.34	0.37	0.46
FTM	11.30	16.37	18.30	18.29	20.02	28.20
FTA	14.40	21.30	23.70	23.94	26.22	36.80
MisFt	2.20	4.67	5.40	5.65	6.59	12.70
FT%	0.60	0.73	0.76	0.76	0.79	0.88
FTAo	15.30	21.97	24.30	4.48	26.60	38.00
ORB	5.20	8.90	10.00	10.18	11.60	16.70
DRB	24.20	30.00	31.70	31.82	33.80	41.80
RPG	33.50	39.77	42.05	42.00	44.30	51.90
ORBo	5.30	8.80	10.00	10.12	11.30	16.10
APG	12.60	18.30	20.45	20.55	22.52	28.40
SPG	3.40	6.20	6.90	6.97	7.80	10.80
BPG	2.10	3.80	4.60	4.68	5.50	8.10
TOV	8.40	12.37	13.30	13.47	14.52	21.00
TOVo	8.00	12.20	13.25	13.23	14.20	18.30
PF	15.50	20.37	21.55	21.71	22.90	30.80
POSSt	82.02	90.45	94.02	94.06	98.05	104.70
ORt	91.38	102.92	106.67	106.86	111.03	122.49
POSSo	82.26	90.43	93.94	94.04	97.47	105.24
DRt	96.12	105.56	109.57	109.80	113.20	130.00
DEFF	-26.60	-7.52	-2.48	-2.93	2.20	13.38
EFF	77.40	98.97	107.00	106.43	114.00	130.90

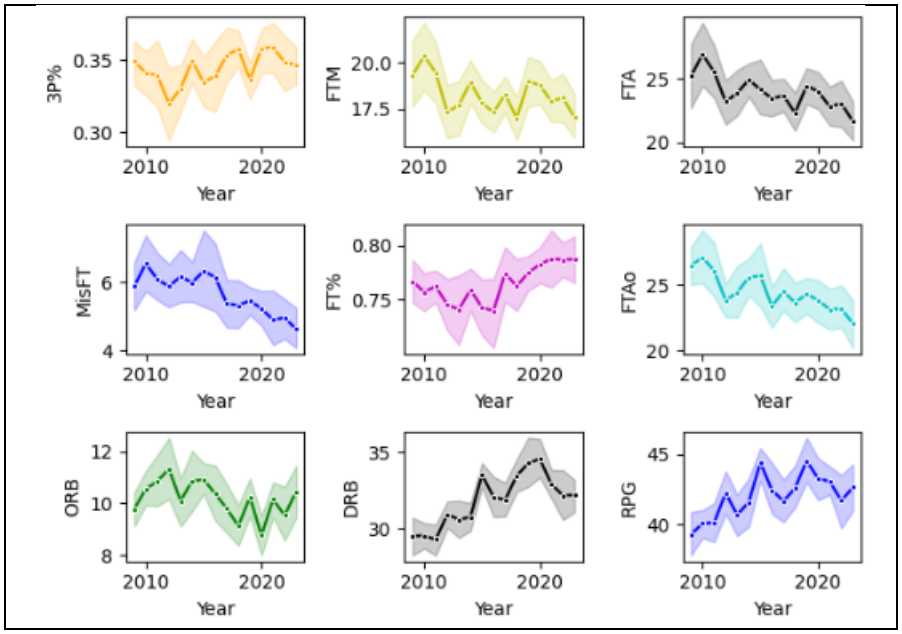
Πίνακας 3.2: Στατιστικά περιγραφικά μέτρα για τα Playoffs

Στη συνέχεια παρουσιάζονται τα διαγράμματα χρονοσειρών ανά χρονική περίοδο για όλες τις μεταβλητές. Στα Σχήματα 3.8 και 3.9 παρουσιάζονται τα time series plots των πρώτων δεκαοκτώ (18) μεταβλητών του αρχείου. Όπως είναι φανερό, οι περισσότερες μεταβλητές ακολουθούν μία αυξητική πορεία, όπως άλλωστε συνέβη και στη κανονική περίοδο με την πάροδο του χρόνου. Πιο αναλυτικά, οι μεταβλητές των πόντων που βάζουν (PPG) και δέχονται οι ομάδες (PPGA), επιτυχημένων (FGM) και καταβληθεισών προσπαθειών για δίποντα ή τρίποντα τόσο των ομάδων (FGA) όσο και των αντιπάλων (FGAo), των επιτυχημένων (3PM), καταβληθεισών προσπαθειών σουτ τριών πόντων (3PA) αμυντικά (DRB) και συνολικά ριμπάουντ (RPG) έχουν αυξητική πορεία με βάση την πάροδο του χρόνου. Αντίθετα, οι μεταβλητές των επιτυχημένων (FTM), αποτυχημένων (MisFT) και των καταβληθεισών προσπαθειών για ελεύθερες βολές των ομάδων (FTA) και των αντιπάλων τους (FTAo) έχουν πτωτική πορεία με την

πάροδο του χρόνου. Τα ποσοστά επιτυχίας των ελεύθερων βολών (FT%) σε γενικές γραμμές στις περισσότερες σεζόν έχουν αυξομειώσεις, ενώ από την σεζόν 2017-2018, και έπειτα, έχουν μόνο αυξητική τάση.



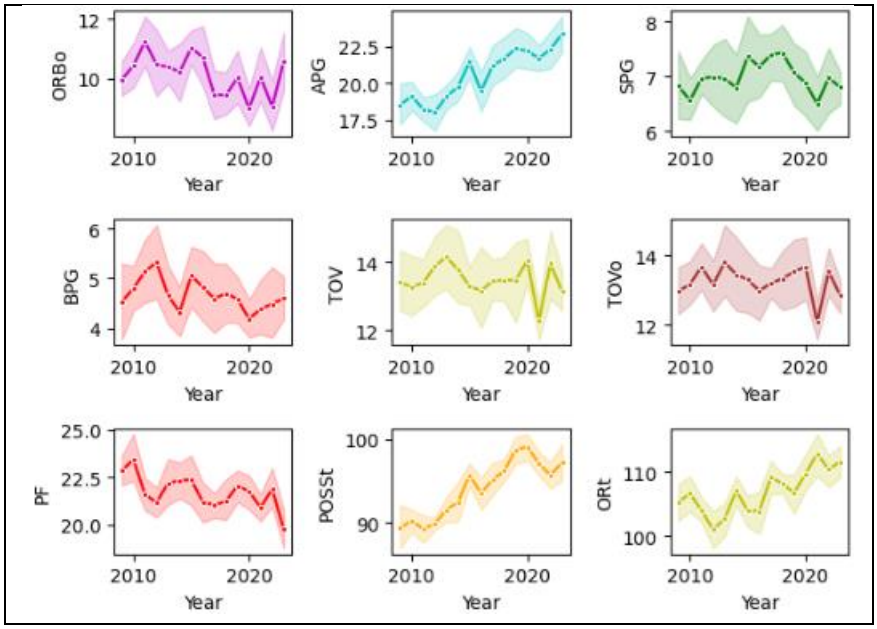
Σχήμα 3.8: Time Series Plots ορισμένων μεταβλητών των Playoffs με την πάροδο των σεζόν



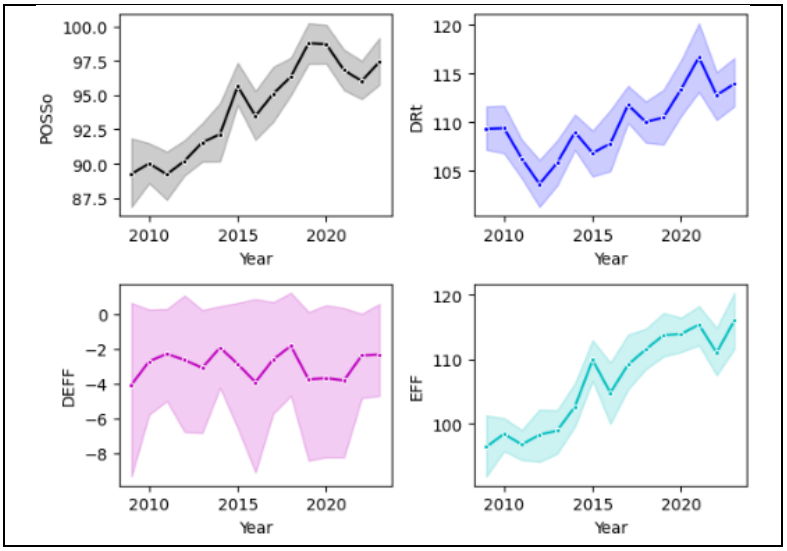
Σχήμα 3.9: Time Series Plots των υπόλοιπων μεταβλητών των Playoffs με την πάροδο των σεζόν

Στα Σχήμα 3.10 και 3.11 υπάρχουν τα αντίστοιχα διαγράμματα για τις υπόλοιπες μεταβλητές που αναλύονται στην παρούσα εργασία. Αυξητική συμπεριφορά διαγραμμάτων κατά μέσο όρο έχουν οι παρακάτω μεταβλητές: ασίστ (APG), κατοχές της ομάδας (POSSt) και των αντιπάλων (POSSo), offensive (ORt) και defensive rating (DRt) και ο δείκτης efficiency (EFF). Επίσης, διακρίνεται μείωση στις τιμές της

μεταβλητής των προσωπικών φάουλ που πραγματοποιεί μία ομάδα (PF). Η συμπεριφορά αυτή ίσως να οφείλεται στις αλλαγές του τρόπου παιχνιδιού των ομάδων, που, όπως αναφέρθηκε προηγουμένως, προτίμησαν να απομακρύνουν τους παίχτες από τη ρακέτα.



Σχήμα 3.10: Time Series Plots επόμενων μεταβλητών των Playoffs με την πάροδο των σεζόν

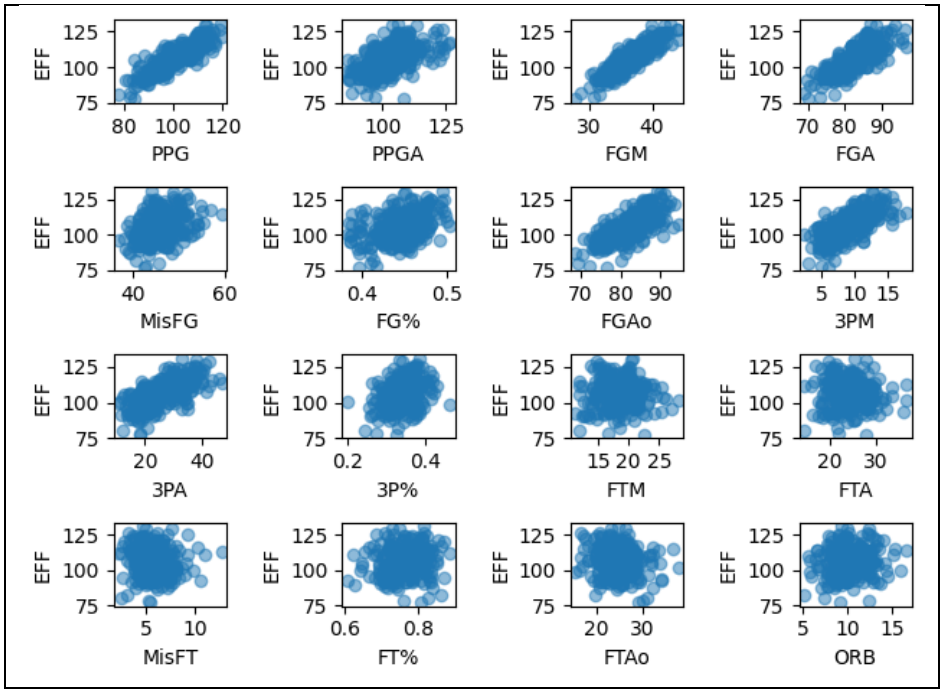


Σχήμα 3.11: Time Series Plots τελευταίων τεσσάρων μεταβλητών των Playoffs με την πάροδο των σεζόν

Μια διαφορά ανάμεσα στις δύο φάσεις της διοργάνωσης, δηλαδή στην κανονική περίοδο και τα playoffs, που παρατηρείται μέσα από τα time series plot που δημιουργήθηκαν είναι πως στις μεταβλητές των επιτυχημένων δίποντων ή τρίποντων (FGM), επιτυχημένων τρίποντων (3PM) και του δείκτη EFF οι τιμές των ομάδων έχουν πτωτική συμπεριφορά. Παραδείγματος χάρη, για τη μεταβλητή επιτυχημένων τρίποντων παρατηρούμε πως στην κανονική περίοδο οι τιμές κυμαίνονται ανάμεσα στο 7.5 με 12.5, ενώ στην επόμενη φάση της διοργάνωσης έχουμε τιμές που ξεκινάνε από

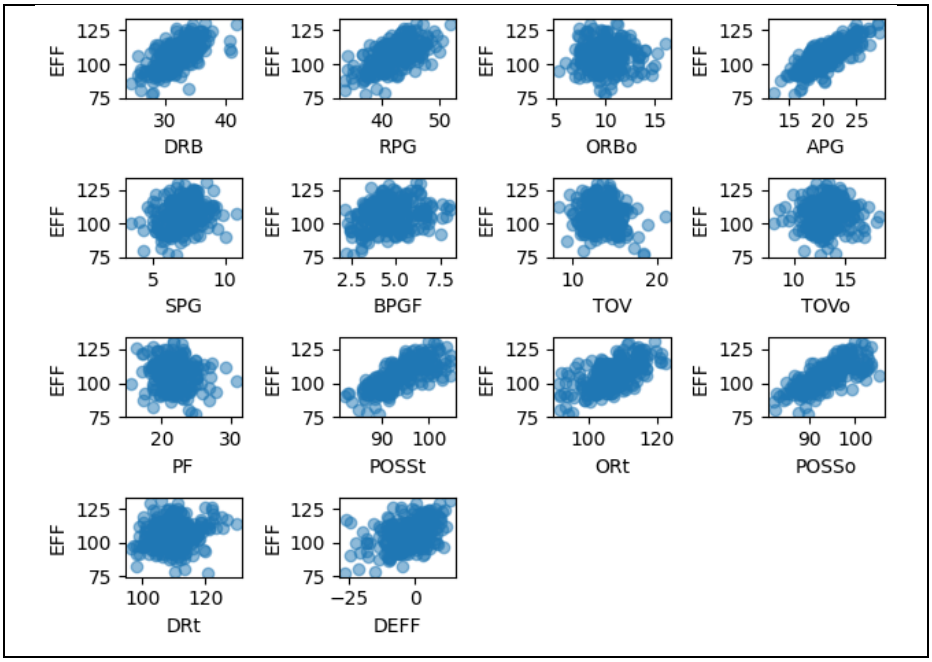
το 5. Για τη μεταβλητή των πόντων στα playoffs οι τιμές κυμαίνονται από 90 έως 110, ενώ στην κανονική περίοδο από 100 έως 120. Αυτό το φαινόμενο εξηγείται από τον πιο συντηρητικό τρόπο παιχνιδιού στη φάση των playoffs, καθώς τα παιχνίδια πλέον είναι σημαντικότερα. Ακόμα, για την μεταβλητή των ασίστ στη φάση των playoffs οι τιμές κυμαίνονται από 17.5 έως 22.5, ενώ στην κανονική περίοδο από 21 έως 25.2. Αυτό φυσικά είναι μία αναμενομένη εξέλιξη διότι στην φάση των playoffs, όπως αναφέρθηκε και παραπάνω, οι ομάδες ασκούν μεγαλύτερη πίεση στο αντίπαλο με αποτέλεσμα οι επιχειρήσεις για κάποιο πετυχημένο καλάθι να γίνονται δυσκολότερες. Αντιθέτως, για την μεταβλητή των προσπαθειών τριπόντων (3PA) παρατηρούμε πως υπάρχει αύξηση σε σχέση με την κανονική περίοδο. Λόγω των δυνατών αμυνών που έχει αυτή η φάση της διοργάνωσης, πολλές φορές οι ομάδες οδηγούνται σε προσπάθειες για να πετύχουν κάποιο καλάθι από πιο μακρινή απόσταση, όπου θα βρεθεί κάποιος ελεύθερος χώρος. Για τον δείκτη DEFF παρατηρούμε πως στη φάση των playoffs παίρνει μόνο αρνητικές τιμές και επομένως με βάση τον τύπο υπολογισμού του, που αναφέρθηκε στο δεύτερο κεφάλαιο, συμπεραίνουμε πως η μεταβλητή DRt παίρνει μεγαλύτερες τιμές από τον δείκτη ORt. Σε ανάλογα συμπεράσματα έχουν καταλήξει στην έρευνά τους οι Cabarkapa et al. (2022), οι οποίοι ανέφεραν αυτή την αναμενόμενη μείωση στις τιμές ορισμένων μεταβλητών όταν οι ομάδες συμμετέχουν στη φάση των Playoffs.

Στο Σχήμα 3.12, όπου παρουσιάζονται οι συσχετίσεις των πρώτων δεκαέξι (16) μεταβλητών με τον δείκτη αποδοτικότητας (EFF), γίνεται αντιληπτό πως οι μεταβλητές των πόντων που βάζει μία ομάδα (PPG), των εύστοχων (FGM) και οι προσπάθειες (FGA) δίποντων ή τρίποντων έχουν ισχυρή θετική συσχέτιση με την μεταβλητή EFF. Στη συνέχεια, οι παρακάτω μεταβλητές: πόντοι που δέχεται η ομάδα (PPGA), ποσοστό εύστοχων (FG%) δίποντων ή τρίποντων, οι προσπάθειες των αντιπάλων για την επίτευξη ίδιων σουτ (FGAo), εύστοχα (3PM), προσπάθειες που επιχειρήθηκαν (3PA) και τα ποσοστά (3P%) τρίποντων έχουν θετική συσχέτιση με την μεταβλητή, όμως δεν είναι τόσο ισχυρή όπως προηγουμένως. Αντιθέτως, η μεταβλητή χαμένων ελεύθερων βολών (MisFT) έχει αρνητική συσχέτιση με τον δείκτη. Ακόμα οι υπόλοιπες μεταβλητές έχουν οριακές τιμές συντελεστών, που μπορούν ενδεχομένως να θεωρηθούν και ασυσχέτιστες.



Σχήμα 3.12: Scatter Plots ορισμένων μεταβλητών των Playoff σε σχέση με EFF

Αντίστοιχα, στο Σχήμα 3.13, για τις υπόλοιπες μεταβλητές παρατηρούμε πως οι μόνες που έχουν ισχυρή συσχέτιση με τον δείκτη είναι τα αμυντικά ριμπάουντ (DRB) και οι ασίστ (APG). Τα συνολικά ριμπάουντ (RPG), οι κατοχές των ομάδων (POSSt) και των αντιπάλων τους (POSSo) και το offensive rating (ORt) έχουν θετική συσχέτιση με τον δείκτη αλλά μικρότερου βαθμού. Από την άλλη πλευρά, τα λάθη (TOV) έχουν αρνητική συσχέτιση με τον δείκτη.



Σχήμα 3.13: Scatter Plots των υπόλοιπων μεταβλητών των Playoffs σε σχέση με EFF

Στις επόμενες παραγράφους θα δοθούν τρείς (3) διαφορετικές αναλύσεις βάσει των κατηγορικών μεταβλητών που περιέχονται στα δεδομένα που συλλέχτηκαν για την εργασία. Αναλυτικότερα, στην πρώτη ενότητα θα δοθούν τα περιγραφικά μέτρα και διαγράμματα για να παρουσιαστούν οι διαφορές των ομάδων που συμμετέχουν στα Playoffs και αυτών που δεν καταφέρνουν να προκριθούν. Στη δεύτερη θα γίνουν συγκρίσεις ανάμεσα στις ομάδες των δύο περιφερειών και στην τελευταία θα γίνει ανάλυση για τις ομάδες που έχουν διακριθεί στις τελευταίες δεκαπέντε (15) σεζόν σε ορισμένες στατιστικές κατηγορίες. Σε όλες τις αναλύσεις θα χρησιμοποιηθούν οι μεταβλητές των πόντων που πετυχαίνει κάθε ομάδα (PPG), των ασίστ (APG), των κλεψιμάτων (SPG), των συνολικών ριμπάουντ (RPG), των λαθών (TOV) και του δείκτη DEFF.

3.4. Ανάλυση με βάση την πρόκριση στα Playoffs

Μια από τις σημαντικότερες αναλύσεις που μπορούν να γίνουν στα δεδομένα που έχουν συλλεχθεί είναι ανάμεσα στις ομάδες οι οποίες προκρίθηκαν στα playoffs και σε αυτές που δεν κατάφεραν να συνεχίσουν στην επόμενη φάση. Αρχικά δίνονται τα στατιστικά περιγραφικά μέτρα των ομάδων. Όπως είναι φανερό, στους δύο πίνακες, 3.3 και 3.4, οι ομάδες που συμμετείχαν στα playoffs διατηρούν σε όλες τις μεταβλητές, εκτός των λαθών και των κατοχών που κερδίζουν ανά αγώνα, υψηλότερες τιμές από τις αντίστοιχες ομάδες που δεν προκρίθηκαν.

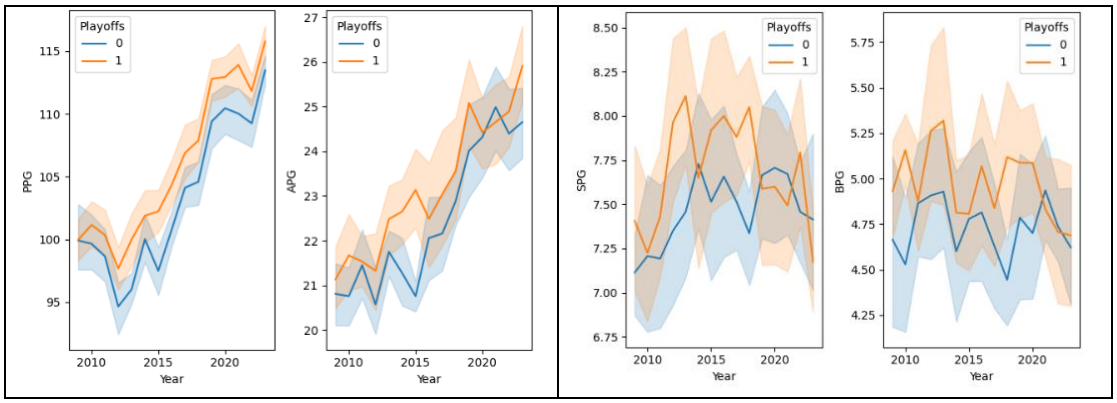
No Playoffs	Min.	1st Q.	Median	Mean	3rd Q.	Max.
PPG	87.00	97.90	102.90	103.26	108.80	117.50
APG	17.40	20.90	22.30	22.46	23.80	28.10
SPG	5.60	7.00	7.50	7.46	7.90	10.00
BPG	2.40	4.25	4.70	4.72	5.10	6.40
RPG	38.40	41.35	42.70	42.76	44.10	47.50
TOV	11.40	13.30	14.50	14.51	15.20	17.70
DEFF	-15.32	-6.97	-3.62	-3.48	-1.49	3.16
POSSt	88.94	93.76	97.05	96.81	99.68	104.99

Πίνακας 3.3: Στατιστικά περιγραφικά μέτρα των ομάδων που δεν προκρίθηκαν στα Playoffs

Playoffs	Min.	1st Q.	Median	Mean	3rd Q.	Max.
PPG	91.80	100.70	105.30	105.93	111.75	120.70
APG	18.50	21.40	23.00	23.19	24.80	30.40
SPG	5.50	7.10	7.60	7.68	8.30	10.00
BPG	3.00	4.40	5.00	4.97	5.40	8.20
RPG	36.90	42.10	43.50	43.54	44.85	51.70
TOV	11.10	13.30	14.00	14.01	14.70	16.80
DEFF	-3.64	1.38	3.36	3.62	5.59	12.04
POSSt	87.78	93.21	96.06	95.87	98.51	104.79

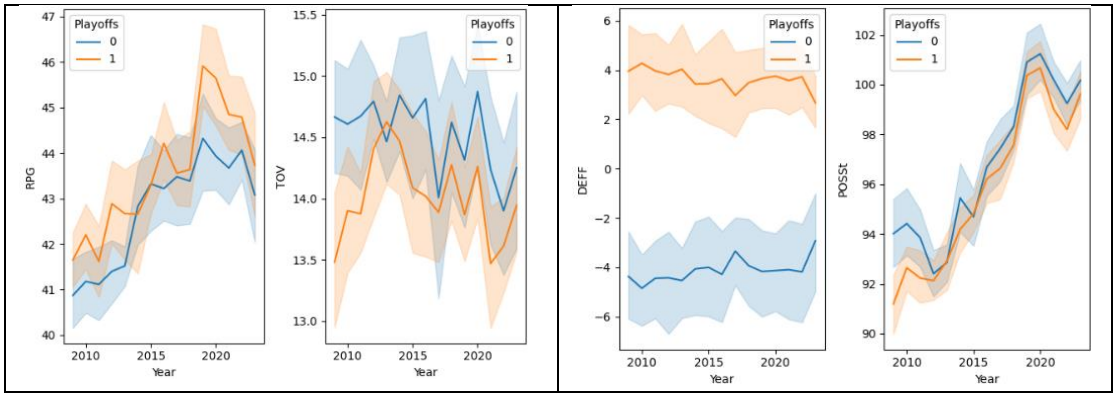
Πίνακας 3.4: Στατιστικά περιγραφικά μέτρα των ομάδων που προκρίθηκαν στα Playoffs

Στη συνέχεια δίνονται τα αντίστοιχα γραφήματα χρονοσειρών για τις μεταβλητές με την πάροδο του χρόνου. Στο Σχήμα 3.14 παρατηρούμε πως και για τους πόντους που πετυχαίνουν οι ομάδες, τις ασίστ, τα κλεψίματα και τα κοψίματα ανά αγώνα οι τιμές των ομάδων που προκρίνονται στην φάση των playoffs είναι υψηλότερες από τις αντίστοιχες των ομάδων που δεν προκρίνονται. Φυσικά, κάποιες χρονιές στις μεταβλητές των ασίστ και των κλεψιμάτων ανά αγώνα η μέση τιμή των ομάδων που δεν καταφέρνουν να προκριθούν τυγχάνει να είναι μεγαλύτερη από τις άλλες.



Σχήμα 3.14: Διαγράμματα χρονοσειρών για τις προκρίσεις στα Playoffs

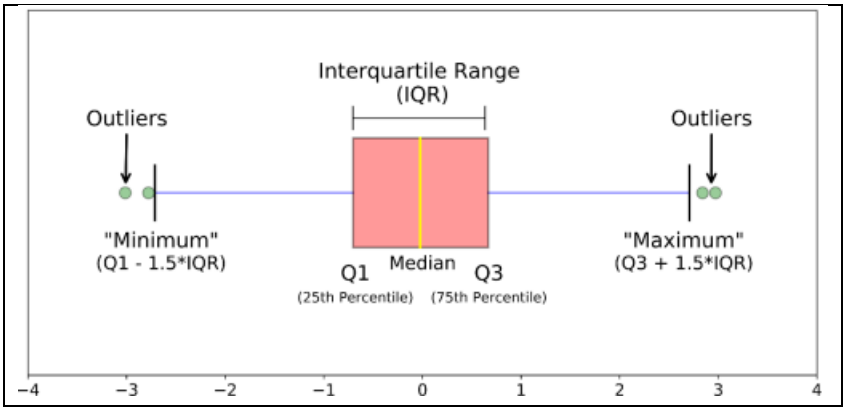
Στο Σχήμα 3.15 παρατηρούμε πως στα συνολικά ριμπάουντ ανά αγώνα και τον δείκτη DEFF οι ομάδες που προκρίνονται στα playoffs διατηρούν υψηλότερες τιμές, ενώ για τα λάθη και τις κατοχές έχουν μικρότερες τιμές. Άξια σχολιασμού είναι η ξεκάθαρη εικόνα που επικρατεί στις μέσες τιμές του δείκτη DEFF. Οι ομάδες που προκρίνονται στα playoffs λαμβάνουν θετικές τιμές για τον συγκεκριμένο δείκτη. Αντιθέτως οι ομάδες που δεν τα καταφέρνουν τείνουν να έχουν αρνητικές τιμές.



Σχήμα 3.15: Διαγράμματα χρονοσειρών για τις προκρίσεις στα Playoffs

Μια ακόμα μέθοδος γραφικής ανάλυσης που χρησιμοποιήθηκε στην παρούσα εργασία είναι το boxplot, γνωστό και ως θηκόγραμμα. Στα θηκογράμματα παρουσιάζεται η ελάχιστη και η μέγιστη τιμή, το ενδοτεταρτημοριακό εύρος, το πρώτο και το τρίτο τεταρτημόριο, η διάμεσος και οι ακραίες τιμές που λαμβάνουν οι μεταβλητές. Τα συγκεκριμένα γραφήματα είναι σημαντικά για την εύρεση των outliers τα οποία δεν χρησιμοποιούνται στην ανάλυση διότι μπορεί να οδηγήσουν σε λάθος

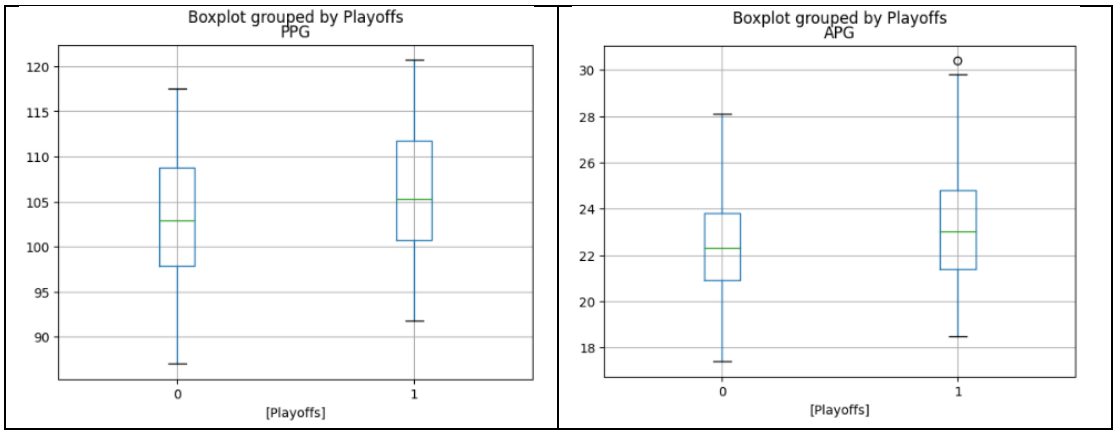
αποτελέσματα. Στο Σχήμα 3.16 δίνεται ένα παράδειγμα ενός boxplot για να γίνει πιο κατανοητό τι παρουσιάζεται σε ένα τέτοιο γράφημα. (Καλλιακμάνης, 2020)



Σχήμα 3.16: Χαρακτηριστικό παράδειγμα θηκογράμματος

(Πηγή: https://builtin.com/data-science/boxplot)

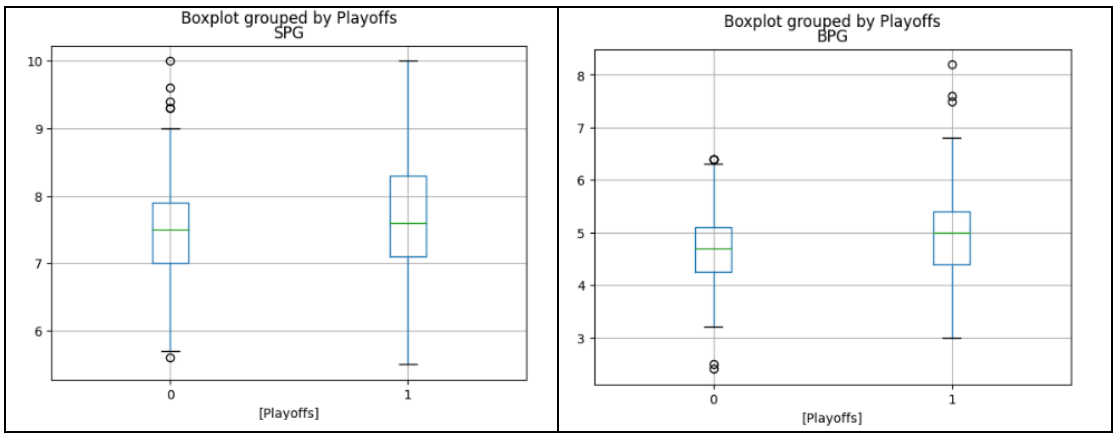
Στο Σχήμα 3.17 παρουσιάζονται τα θηκογράμματα των πόντων ανά αγώνα στα αριστερά και των ασίστ ανά αγώνα στα δεξιά. Όπως είναι φανερό, για τους πόντους ανά αγώνα και τις ασίστ οι ομάδες που προκρίθηκαν στα playoffs λαμβάνουν υψηλότερες τιμές στη διάμεσο τους. Συγκεκριμένα, για τους πόντους ξεπερνάει το 105 και τις ασίστ είναι κοντά στο 24. Ακόμα, για τις ασίστ έχουμε και μία ακραία τιμή που λαμβάνει η απόδοση της ομάδας Golden State Warriors την σεζόν 2016-2017, όπου και μέτρησαν συνολικά 30.40 ασίστ ανά αγώνα.



Σχήμα 3.17: Boxplots για PPG και APG για τις προκρίσεις στα Playoffs

Αντίστοιχα, στο επόμενο σχήμα, 3.18, παρουσιάζονται τα διαγράμματα για τα κλεψίματα ανά αγώνα, στα αριστερά, και για τα κοψίματα ανά αγώνα, στα δεξιά. Στις δύο αυτές περιπτώσεις οι τιμές των διαμέσων για τις ομάδες που προκρίθηκαν στα playoffs έχουν μεγαλύτερες τιμές. Ακόμα, και στα δύο γραφήματα παρουσιάζονται έκτροπες τιμές. Για τα κλεψίματα ανά αγώνα έχουμε αρκετές τιμές υψηλότερες του άνω ορίου στο διάγραμμα. Πιο συγκεκριμένα: οι αποδόσεις των ομάδων Golden State Warriors τη σεζόν 2009-2010 με απόδοση 9.30, με την ίδια απόδοση την σεζόν 2013-2014 οι Philadelphia 76ers, η ίδια ομάδα την επόμενη σεζόν 2014-2015 με απόδοση 9.60, οι Chicago Bulls με απόδοση 10.00 τη σεζόν 2019-2020 και τέλος οι Toronto Raptors με απόδοση 9.40 τη σεζόν 2022-2023. Για τιμές κάτω από το κάτω όριο στο

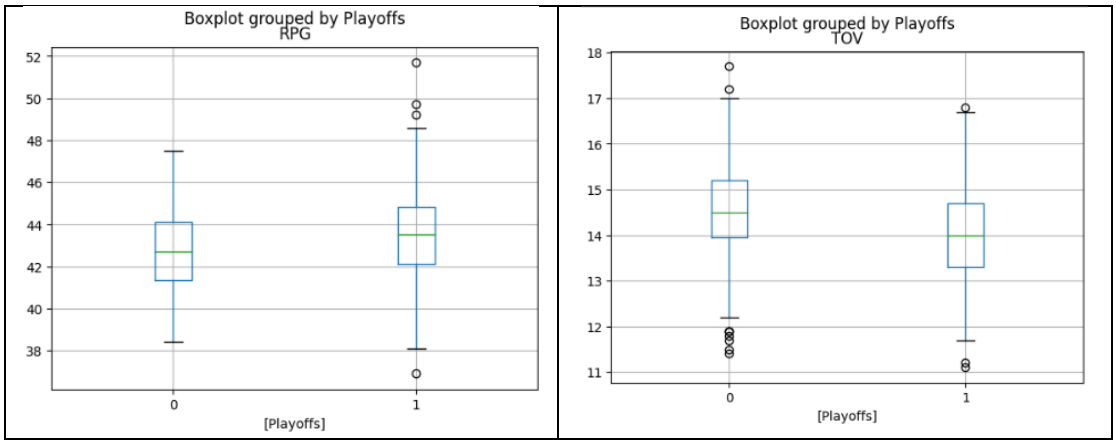
διάγραμμα οι Brooklyn Nets τη σεζόν 2010-2011 με απόδοση 5.60 κλεψίματα ανά αγώνα. Όλες αυτές οι τιμές αντιστοιχούσαν σε ομάδες που δεν προκρίθηκαν στα playoffs. Έπειτα, για τα κοψίματα ανά αγώνα έχουμε ακραίες τιμές και για ομάδες που προκρίθηκαν και για αυτές που δεν προκρίθηκαν. Για τις ομάδες που προκρίθηκαν στα playoffs υπάρχουν τρεις (3) έκτροπες τιμές, οι οποίες ξεπερνούν την τιμή του άνω ορίου. Αυτές αφορούν στις ομάδες Oklahoma City Thunder για τη σεζόν 2011-2012 και 2012-2013 με απόδοση 8.20 και 7.60 αντίστοιχα και στους Golden State Warriors τη σεζόν 2017-2018 με απόδοση 7.50. Για τις ομάδες που δεν προκρίθηκαν στα playoffs έχουμε τιμές που ξεπερνούν και το άνω αλλά και το κάτω όριο του διαγράμματος. Για τις τιμές για το κάτω όριο υπάρχουν δύο (2) έκτροπες τιμές που σχετίζονται με τις αποδόσεις των ομάδων: New York Knicks τη σεζόν 2008-2009 με τιμή 2.40 και Cavaliers Cleveland τη σεζόν 2018-2019 με απόδοση 2.50. Τέλος, υπάρχει και μία (1) τιμή που ξεπερνάει το άνω όριο του διαγράμματος και είναι ίση με 6.40 που έχουν σημειώσει τρεις (3) ομάδες. Συγκεκριμένα είναι οι: Golden State Warriors τη σεζόν 2008-2009, New Orleans Pelicans τη σεζόν 2013-2014 και οι Indiana Pacers τη σεζόν 2020-2021.



Σχήμα 3.18: Boxplots για SPG και BPG για τις προκρίσεις στα Playoffs

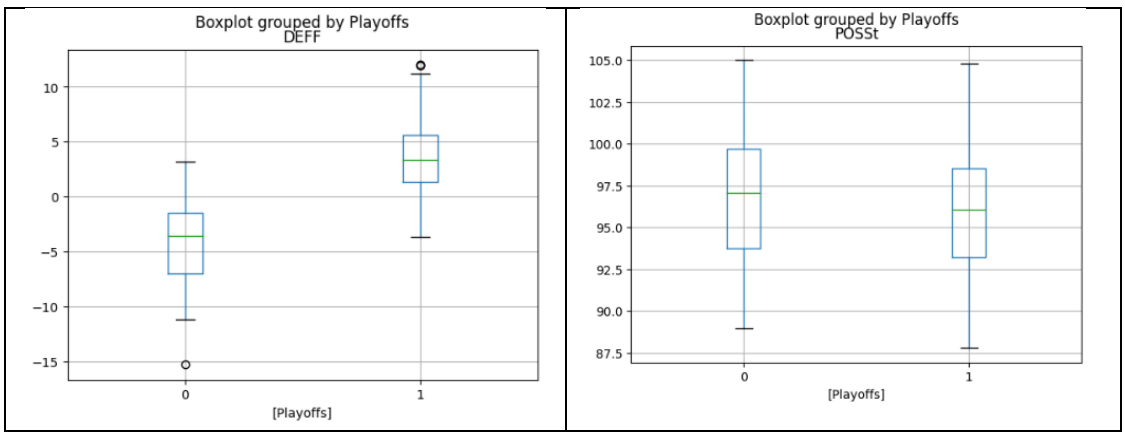
Στο επόμενο σχήμα, 3.19, διακρίνονται τα boxplots των μεταβλητών ριμπάουντ ανά αγώνα και των λαθών στα οποία υποπίπτουν οι ομάδες. Όπως είδαμε και προηγουμένως, στην περίπτωση των λαθών οι ομάδες που δεν αγωνίζονται στα playoffs λαμβάνουν μεγαλύτερη τιμή διαμέσου. Οι τιμές που ξεπερνούν τη maximum τιμή έχουν σχέση με τις ομάδες: Oklahoma City Thunder τη σεζόν 2015-2016 με απόδοση 48.60, Milwaukee Bucks τη σεζόν 2018-2019 και 2019-2020 ίση με 49.70 και 51.70 αντίστοιχα, Memphis Grizzlies τη σεζόν 2021-2022 με απόδοση 49.20 και Milwaukee Bucks τη σεζόν 2022-2023 με απόδοση 48.60. Ακόμα υπάρχει και μια απόδοση που περνάει την τιμή του κάτω ορίου και σχετίζεται με την επίδοση των Miami Heat τη σεζόν 2013-2014 που ήταν ίση με 36.90. Σε ό,τι έχει να κάνει με τα λάθη ανά αγώνα παρατηρούμε περισσότερες έκτροπες τιμές, οι οποίες σχετίζονται κυρίως με τις ομάδες που δεν προκρίθηκαν στην φάση των playoffs. Για τις αποδόσεις που είναι μικρότερες από το κάτω όριο του διαγράμματος σε σχέση με τις ομάδες που δεν προκρίθηκαν στα playoffs έχουμε: τους Charlotte Hornets τη σεζόν 2014-2015, τους Dallas Mavericks τη σεζόν 2016-2017, τους Detroit Pistons τη σεζόν 2016-2017,

όλους με 11.90, τους Charlotte Hornets ξανά τη σεζόν 2016-2017 με 11.50, τους Chicago Bulls τη σεζόν 2020-2021 με 11.40, τους Atlanta Hawks τη σεζόν 2021-2022 με 11.80 και τους Toronto Raptors τη σεζόν 2022-2023 με 11.70. Για τις ομάδες που προκρίθηκαν στα playoffs έχουμε τους: Philadelphia 76ers τη σεζόν 2011-2012 με 11.20 και τους Portland Trail Blazers σεζόν 2020-2021 με 11.10. Έπειτα από έρευνα βρέθηκαν και τιμές οι οποίες ξεπερνάνε την τιμή του άνω ορίου και για τις δύο κατηγορίες ομάδων που ψάχνουμε. Σε πρώτη φάση για αυτές που δεν προκρίθηκαν στα playoffs βρέθηκαν δύο (2) αποδόσεις που ξεπέρασαν την αντίστοιχη τιμή. Η πρώτη είναι τη σεζόν 2014-2015 με τους New York Knicks να μετράνε 17.70 λάθη και η δεύτερη τη σεζόν 2015-2016 με τους Phoenix Suns να έχουν 17.20. Για τις ομάδες που προκρίθηκαν βρέθηκε μόνο μία (1) τιμή που έχει σχέση με την απόδοση των Philadelphia 76ers τη σεζόν 2017-2018 ίση με 16.80.



Σχήμα 3.19: Boxplots για RPG και TOV για τις προκρίσεις στα Playoffs

Σε ό,τι έχει να κάνει σχέση με τις δύο τελευταίες μεταβλητές παρατηρούμε πως βάσει του δείκτη DEFF η διάμεσος βρίσκεται υψηλότερα για τις ομάδες που βρέθηκαν στα playoffs, ενώ για τις κατοχές ανά αγώνα γίνεται ακριβώς το αντίθετο. Έκτροπες τιμές των μεταβλητών παρουσιάζονται μόνο στην περίπτωση του δείκτη DEFF. Η τιμή που ξεπερνάει το άνω όριο του διαγράμματος έχει σχέση με ομάδες που προκρίθηκαν στα playoffs και συγκεκριμένα με τους Golden State Warriors τη σεζόν 2015-2016 όπου και είχαν επίδοση ίση με 12.04. Με τη τιμή που ξεπερνούν το κάτω όριο του διαγράμματος σχετίζεται η ακραία τιμή των New Orleans Pelicans τη σεζόν 2011-2012, οι οποίοι δεν προκρίθηκαν στη φάση των playoffs, ίση με -15.32.

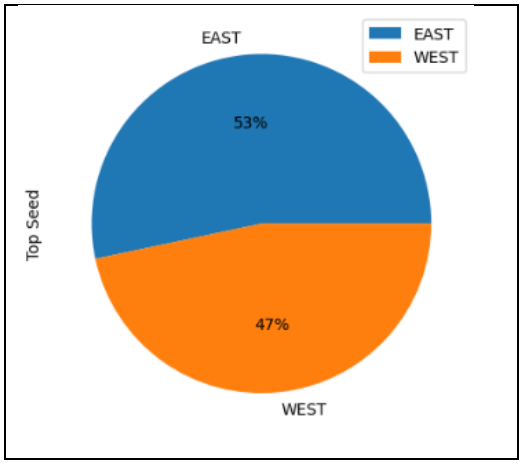


Σχήμα 3.20: Boxplots για DEFF και POSSt για τις προκρίσεις στα Playoffs

3.5. Ανάλυση με βάση τον διαχωρισμό σε Περιφέρειες

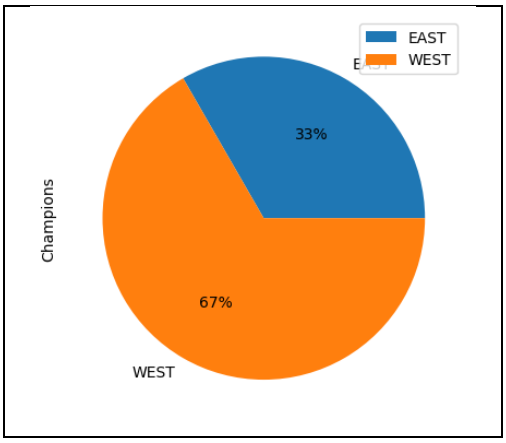
Σε αυτή την ενότητα θα γίνει σύγκριση ανάμεσα στα στατιστικά των ομάδων για κάθε περιφέρεια, με σκοπό να συγκρίνουμε τις ομάδες που συμμετέχουν σε κάθε διοργάνωση. Αρχικά θα γίνει ανάλυση για τα στατιστικά δεδομένα που έχουν συλλεχθεί για τη φάση της κανονικής περιόδου και έπειτα θα διερευνηθούν οι ίδιες μεταβλητές για τη φάση των playoffs. Στην ανατολική περιφέρεια (East Conference) συμμετέχουν οι ομάδες Atlanta Hawks, Boston Celtics, Brooklyn Nets, Charlotte Hornets, Chicago Bulls, Cleveland Cavaliers, Detroit Pistons, Indiana Pacers, Miami Heat, Milwaukee Bucks, New York Knicks, Orlando Magic, Philadelphia 76ers, Toronto Raptors και Washington Wizards. Προφανώς σε κάθε περιφέρεια συμμετέχει το ίδιο πλήθος αριθμός ομάδων.

Πριν από την ανάλυση των ομάδων που ανήκουν στις δύο περιφέρειες κρίνεται ορθό να διερευνήσουμε σε ποια περιφέρεια συμμετείχε κάθε αγωνιστική περίοδο η ομάδα που κατάφερε να έχει το καλύτερο ποσοστό νικών έπειτα από τα ογδόντα δύο (82) παιχνίδια που διαδραματίζονται σε αυτή τη φάση. Για τη συγκεκριμένη ανάλυση θα χρησιμοποιηθεί η μεταβλητή «Top Seed», η οποία έχει δημιουργηθεί βάσει αυτού του χαρακτηριστικού. Το ιδανικότερο στατιστικό διάγραμμα για αυτή την ανάλυση είναι τα pie charts. Πιο αναλυτικά, το γνωστό και ως γράφημα «πίτα», pie chart, είναι ένα κυκλικό στατιστικό διάγραμμα που διαμελίζεται σε κομμάτια για να διακρίνουμε τα ποσοστά που κατέχει το κάθε κομμάτι στις κατηγορίες που επιθυμούμε. Όπως φαίνεται στο Σχήμα 3.21, δεν υπάρχει μεγάλη διαφορά στα ποσοστά των περιφερειών που ανήκουν οι ομάδες που καταφέρνουν να πετύχουν το καλύτερο ποσοστό νικών ανά αγωνιστική περίοδο. Συγκεκριμένα, η ανατολική περιφέρεια κατέχει το ποσοστό του 53%, δηλαδή οι ομάδες που συμμετέχουν σε αυτή είχαν οκτώ (8) χρονιές από τις δεκαπέντε (15) που αναλύουμε συνολικά το καλύτερο ποσοστό νικών. Εφόσον οι ομάδες της ανατολικής περιφέρειας είχαν τις περισσότερες χρονιές, έστω και για ένα μικρό ποσοστό, το καλύτερο ποσοστό νικών τους δινόταν το πλεονέκτημα έδρας στους τελικούς της διοργάνωσης. Με λίγα λόγια, όποια ομάδα κατάφερνε να στεφθεί πρωταθλήτρια της ανατολικής περιφέρειας και προκρινόταν στους τελικούς όλης της διοργάνωσης θα ξεκίναγε με το πρώτο παιχνίδι στο γήπεδο της και αν έφταναν οι τελικοί στο έβδομο (7) παιχνίδι θα πραγματοποιούνταν τα τέσσερα (4) παιχνίδια στην έδρα της.



Σχήμα 3.21: Pie chart για την κατάκτηση του καλύτερου ποσοστού νικών βάση των περιφερειών

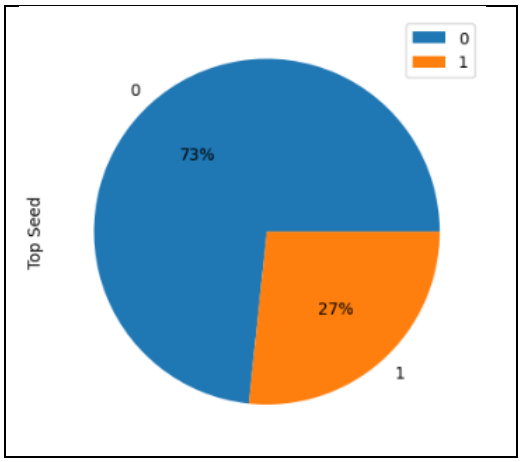
Δεδομένου του γεγονότος πως οι ομάδες της ανατολικής περιφέρειας είχαν το πλεονέκτημα έδρας τις περισσότερες χρονιές της διοργάνωσης θα περιμέναμε αυτή η υπεροχή να τους δώσει το πλεονέκτημα νίκης και στους τελικούς όλης της διοργάνωσης. Για να διερευνήσουμε αυτόν τον ισχυρισμό θα χρησιμοποιήσουμε αυτή την φορά σαν μεταβλητή την «Champions», ώστε να δούμε ποια περιφέρεια κατάφερε να κερδίσει τα περισσότερα πρωταθλήματα τα τελευταία δεκαπέντε (15) χρόνια. Όπως φαίνεται στο Σχήμα 3.22, παρόλο που οι ομάδες που είχαν το καλύτερο ποσοστό νικών ανήκουν στην περιφέρεια EAST, οι πρωταθλήτριες ομάδες της διοργάνωσης ανήκουν στην περιφέρεια WEST έχοντας κατακτήσει το 67% των πρωταθλημάτων, δηλαδή τα εννέα (9) από τα δεκαπέντε (15) πρωταθλήματα. Το συγκεκριμένο αποτέλεσμα ερμηνεύεται και από την ανάλυση που ακολουθεί, στην οποία θα φανεί πως οι συγκεκριμένες ομάδες κατείχαν υψηλότερες τιμές σε όλες σχεδόν τις κατηγορίες μεταβλητών που χρησιμοποιήθηκαν προς ανάλυση.



Σχήμα 3.22: Pie chart για την κατάκτηση περισσότερων πρωταθλημάτων βάση των περιφερειών

Προφανώς μέσα από αυτές τις αναλύσεις που είδαμε προηγουμένως δημιουργείται η απορία αν τελικά οι ομάδες που καταφέρνουν να έχουν το καλύτερο ποσοστό νικών στην κανονική περίοδο της διοργάνωσης είναι και αυτές που θα κατακτήσουν στο τέλος το πρωτάθλημα. Χρησιμοποιώντας τις μεταβλητές «Champions» και «Top Seed»

καταφέρνουμε να δημιουργήσουμε ένα διάγραμμα πίτας που θα μας δώσει αυτή την πληροφορία. Σύμφωνα με το Σχήμα 3.23, συμπεραίνουμε πως οι ομάδες που έχουν το καλύτερο ποσοστό νικών έχουν καταφέρει να κατακτήσουν μόνο τέσσερα (4) από τα δεκαπέντε (15) πρωταθλήματα της διοργάνωσης που αναλύουμε. Αυτή είναι άλλη μια ένδειξη πως στην φάση των playoffs οι ομάδες δίνουν μεγαλύτερη βαρύτητα στον τρόπο που ανταποκρίνονται στα παιχνίδια.



Σχήμα 3.23: Pie chart για την κατάκτηση περισσότερων πρωταθλημάτων βάση της μεταβλητής Top Seed

3.5.1. Κανονική περίοδος

Ξεκινώντας την ανάλυση παρουσιάζουμε τα περιγραφικά στατιστικά των ομάδων που συμμετέχουν σε κάθε περιφέρεια με απώτερο σκοπό τη σύγκριση των δύο περιφερειών ανάλογα με τις τιμές που παίρνουν οι μεταβλητές που χρησιμοποιούνται.

EAST	Min.	1st Q.	Median	Mean	3rd Q.	Max.
PPG	87.00	97.88	102.90	103.39	109.20	120.10
APG	18.50	21.00	22.70	22.64	24.10	28.10
SPG	5.60	7.00	7.40	7.48	8.00	10.00
BPG	2.40	4.30	4.80	4.81	5.30	6.70
RPG	36.90	41.50	42.80	42.94	44.40	51.70
TOV	11.20	13.50	14.10	14.16	14.90	17.70
DEFF	-15.32	-4.06	-0.09	-0.49	3.09	10.50
POSSt	88.03	92.76	96.15	95.85	98.67	104.91

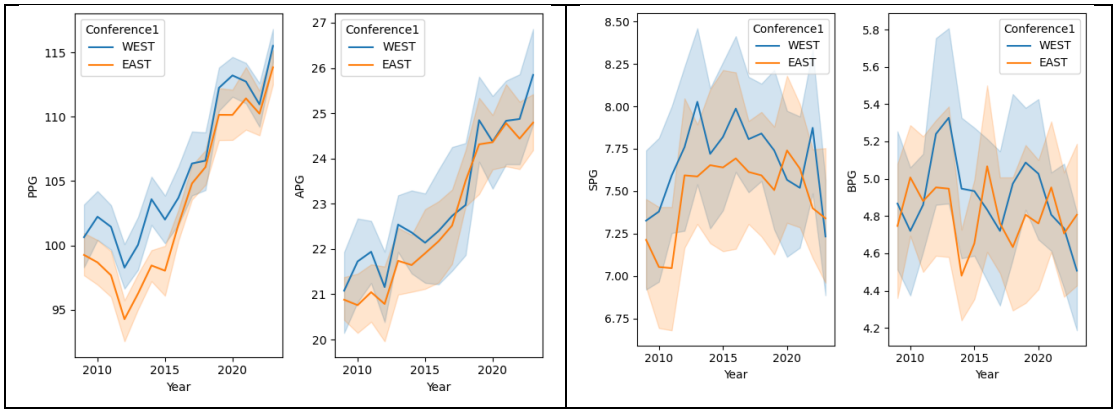
Πίνακας 3.5: Στατιστικά περιγραφικά μέτρα περιφέρειας EAST (κανονική περίοδος)

WEST	Min.	1st Q.	Median	Mean	3rd Q.	Max.
PPG	89.60	100.70	105.30	105.97	111.70	120.70
APG	17.40	21.30	22.70	23.05	24.50	30.40
SPG	5.50	7.10	7.60	7.67	8.30	10.00
BPG	3.40	4.40	4.80	4.90	5.30	8.20
RPG	38.40	42.00	43.40	43.41	44.70	49.20
TOV	11.10	13.50	14.40	14.33	15.10	17.20

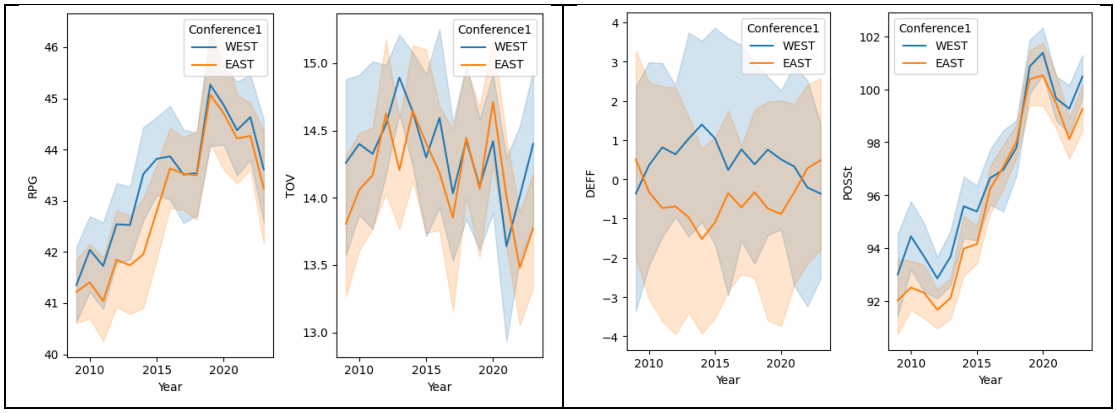
DEFF	-11.03	-2.73	1.04	0.48	4.14	12.04
POSSt	87.78	94.40	96.6	96.77	99.50	104.99

Πίνακας 3.6: Στατιστικά περιγραφικά μέτρα περιφέρειας WEST (κανονική περίοδος)

Στη συνέχεια παρουσιάζονται διαγράμματα των μεταβλητών, time series plot, με την πάροδο του χρόνου και για τις δύο περιφέρειες. Όπως είναι φανερό στα σχήματα 3.24 και 3.25, στις διαγραμματικές απεικονίσεις οι ομάδες που κατατάσσονται στην περιφέρεια WEST λαμβάνουν μεγαλύτερες τιμές στις μεταβλητές με την πάροδο του χρόνου σε σχέση με εκείνες της EAST. Στα συγκεκριμένα γραφήματα φαίνεται η πορεία των μέσων τιμών ανά χρονιά (έντονη γραμμή) αλλά και το εύρος των τιμών με πιο απαλό χρώμα γύρω από την γραμμή των μέσων. Από τη σεζόν 2019-2020 και έπειτα παρατηρούμε πως αρχίζουν οι τιμές των ομάδων της περιφέρειας EAST να πλησιάζουν και σε κάποιες μεταβλητές τελικά να ξεπερνάνε τις τιμές της αντίπαλης περιφέρειας.



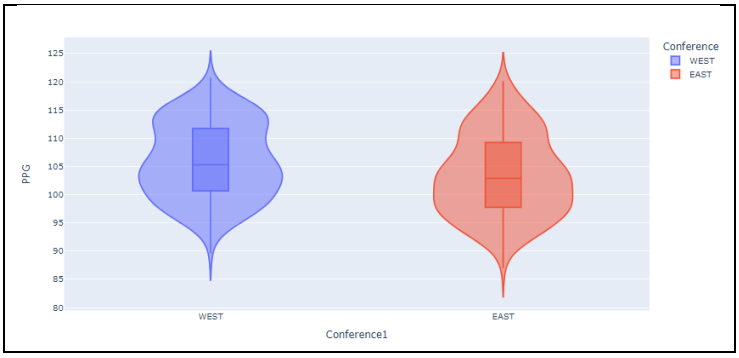
Σχήμα 3.24: Διαγράμματα χρονοσειρών των περιφερειών (κανονική περίοδος)



Σχήμα 3.25: Διαγράμματα χρονοσειρών των περιφερειών (κανονική περίοδος)

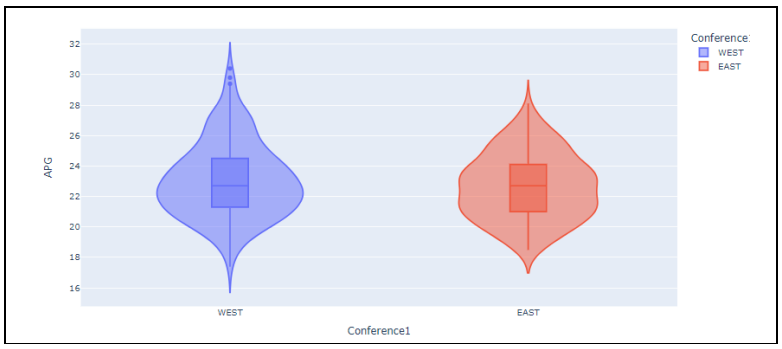
Μία ακόμη μέθοδος γραφικής ανάλυσης, η οποία αποτελεί μεγάλο εφόδιο για την οπτικοποίηση των δεδομένων, είναι η δημιουργία των violin plots. Τα violin plots αποτελούν μια αναβάθμιση των boxplots, εφόσον εσωτερικά περιέχουν και αυτά, και μας δίνουν περαιτέρω πληροφορίες. Πιο συγκεκριμένα η διαφορά τους αναγνωρίζεται όταν η κατανομή των δεδομένων είναι πολυκόρυφη (Καμίτσης, 2023). Στον Σχήμα 3.26 απεικονίζονται τα violin plots των πόντων που σκοράρουν οι ομάδες διαχωρισμένες στις δύο περιφέρειες. Όπως γίνεται αντιληπτό, η τιμή της διαμέσου για τις ομάδες της

WEST περιφέρειας είναι υψηλότερη από της αντίπαλης περιφέρειας, κάτι το οποίο ήταν αναμενόμενο βάση των προηγούμενων διαγραμμάτων χρονοσειρών. Ακόμα, δεν διακρίνονται ακραίες τιμές στα δεδομένα των πόντων για καμία περιφέρεια.



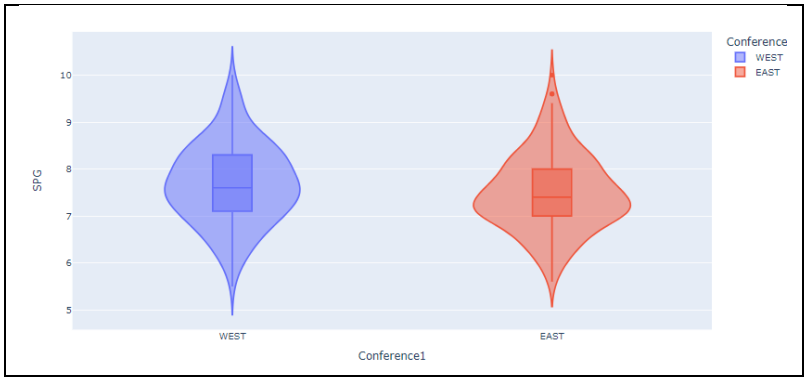
Σχήμα 3.26: Violin plot PPG με βάση τις περιφέρειες (κανονική περίοδος)

Για το αντίστοιχο διάγραμμα που παριστάνει τις ασίστ, Σχήμα 3.27, φαίνεται πως οι ομάδες των περιφερειών έχουν ίσες διαμέσους και παρατηρούμε πως για τις ομάδες της WEST περιφέρειας υπάρχουν τρείς έκτροπες τιμές. Όλες οι έκτροπες τιμές, που διαφαίνονται στο γράφημα εκτός των ορίων του boxplots με έντονες κουκίδες, έχουν να κάνουν με την απόδοση των Golden State Warriors τις σεζόν 2022-2023, 2018-2019 και 2017-2018, όπου μετρούσαν 29.80, 29.40 και 30.40 ασίστ ανά αγώνα αντίστοιχα.



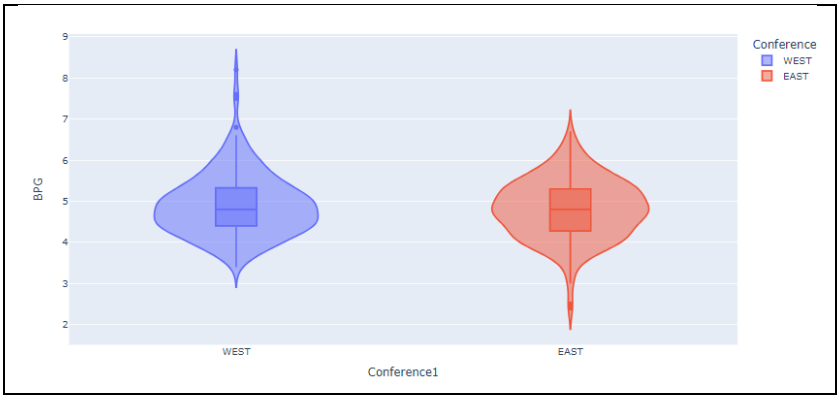
Σχήμα 3.27: Violin plot APG με βάση τις περιφέρειες (κανονική περίοδος)

Όσον αφορά το violin plot για τα κλεψίματα των ομάδων παρατηρείται μεγαλύτερη τιμή στη διάμεσο των ομάδων της περιφέρειας WEST, καθώς και τρεις έκτροπες τιμές στην άλλη περιφέρεια. Οι τιμές αυτές έχουν σχέση με την απόδοση 9.60 κλεψιμάτων ανά αγώνα από τις ομάδες Memphis Grizzlies τη σεζόν 2012-2013, Los Angeles Clippers 2013-2014, Milwaukee Bucks 2015-2016 και Golden State Warriors τη σεζόν 2017-2018. Έπειτα με την απόδοση 9.80 κλεψιμάτων από την ομάδα Memphis Grizzlies τη σεζόν 2021-2022. Τέλος, η μεγαλύτερη τιμή που λαμβάνεται είναι ίση με 10.00 κλεψίματα ανά αγώνα από την ομάδα Chicago Bulls τη σεζόν 2020-2021.



Σχήμα 3.28: Violin plot SPG με βάση τις περιφέρειες (κανονική περίοδος)

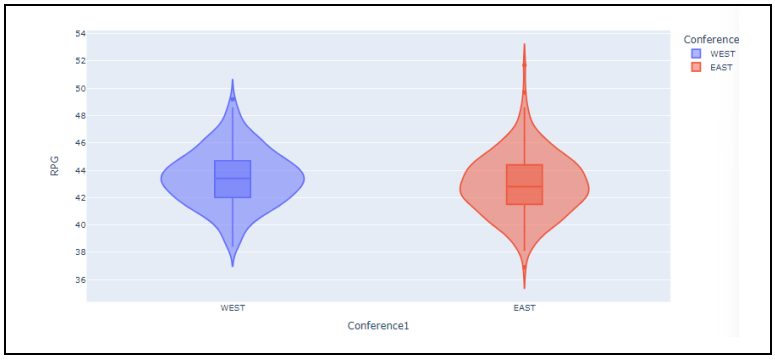
Στη συνέχεια δημιουργώντας το ίδιο γράφημα για τα κοψίματα της ομάδας ανά αγώνα διακρίνουμε πως η διάμεσος των τιμών για τις περιφέρειες είναι ίσες. Επίσης, παρατηρώντας το σχήμα έξω από τα boxplot είμαστε σε θέση να αντιληφθούμε την σαφή υπεροχή των ομάδων της WEST περιφέρειας παρόλο που έχουν ίση διάμεσο. Αυτό το καταλαβαίνουμε διότι οι τιμές που παίρνει είναι από το τρία (3) σαν ελάχιστη και φτάνουν έως το οκτώ (8) σαν μέγιστη, με πολλές τιμές κοντά στη μέγιστη. Από την άλλη πλευρά στην EAST περιφέρεια δεν έχουμε τόσο ακραίες τιμές μέγιστης επίδοσης αλλά αρκετές κοντά στη διάμεσο. Όσον αφορά τις ακραίες τιμές, όπως φαίνεται στο Σχήμα 3.29, υπάρχουν και στις δύο περιφέρειες. Σε ό,τι έχει να κάνει με την περιφέρεια WEST οι ακραίες τιμές εμφανίζονται στις τιμές που είναι μεγαλύτερες του άνω ορίου του διαγράμματος. Πιο συγκεκριμένα, η πρώτη ακραία τιμή αντιστοιχεί στην απόδοση της ομάδας της Oklahoma City Thunder τη σεζόν 2011-2012, όπου μετρούσε κατά μέσο όρο 8.20 κλεψίματα. Έπειτα, πάλι η ίδια ομάδα τη σεζόν 2012-2013 μέτρησε συνολικά 7.60 κλεψίματα ανά αγώνα. Η τρίτη ακραία τιμή ανήκει στους Golden State Warriors τη σεζόν 2016-2017 με 6.80 και, τέλος, την επόμενη σεζόν οι ίδιοι με 7.50. Στην άλλη περιφέρεια οι ακραίες τιμές έχουν να κάνουν με τις ελάχιστες αποδόσεις των New York Knicks τη σεζόν 2008-2009 που μέτραγαν 2.50 και των Cleveland Cavaliers τη σεζόν 2018-2019 με 2.40 κοψίματα ανά αγώνα.



Σχήμα 3.29: Violin plot BPG με βάση τις περιφέρειες (κανονική περίοδος)

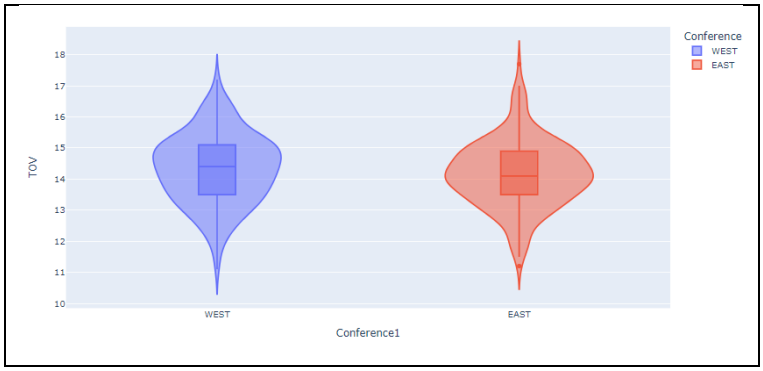
Σε σχέση με τα ριμπάουντ ανά αγώνα είναι φανερό από το Σχήμα 3.30 πως η τιμή της διαμέσου των ομάδων που συμμετέχουν στην περιφέρεια WEST είναι ελαφρώς υψηλότερη. Για τις ακραίες τιμές έχουμε τους Milwaukee Bucks τις σεζόν 2018-2019, 2019-2020 με 49.70 και 51.70 απόδοση αντίστοιχα, τιμές υψηλότερες από το

συνηθισμένο. Αντίθετα, μια αρκετά μικρή απόδοση στα συνολικά ριμπάουντ, που θεωρείται και ακραία τιμή, είχαν οι Miami Heat τη σεζόν 2013-2014 με τιμή 36.90. Τέλος, στην άλλη περιφέρεια έχουμε τη σεζόν 2021-2022 τους Memphis Grizzlies με απόδοση 49.20. Από το συγκεκριμένο σχήμα μπορούμε ακόμα να συμπεράνουμε πως η μεταβλητότητα των τιμών της EAST περιφέρειας είναι μεγαλύτερη από της αντίπαλης, διότι οι τιμές που παίρνει η συγκεκριμένη μεταβλητή είναι σε ένα αρκετά μεγάλο διάστημα, από τριάντα πέντε (35) έως και πενήντα τρία (53).



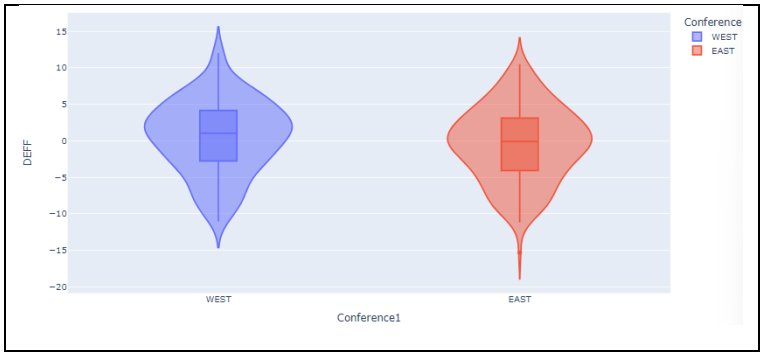
Σχήμα 3.30: Violin plot RPG με βάση τις περιφέρειες (κανονική περίοδος)

Ακόμα και στα λάθη διακρίνεται η διάμεσος των τιμών των ομάδων της περιφέρειας WEST να είναι υψηλότερη. Οι ακραίες τιμές που αφορούν χαμηλότερες τιμές αντιστοιχούν στους Portland Trail Blazers για τη σεζόν 2020-2021 με 11.10 και τους Philadelphia 76ers τη σεζόν 2011-2012 με 11.20. Ακόμα για τους Philadelphia 76ers αντιστοιχεί και μία ακραία τιμή για τη σεζόν 2014-2015 με τιμή 17.70.



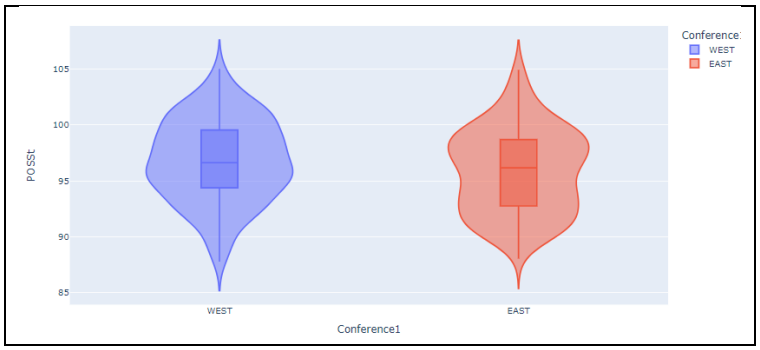
Σχήμα 3.31: Violin plot TOV με βάση τις περιφέρειες (κανονική περίοδος)

Τέλος, για τα violin plots του δείκτη DEFF με βάση τις περιφέρειες παρατηρούμε πως η διάμεσος των ομάδων της WEST κατηγορίας είναι υψηλότερη και πως υπάρχει μία μόνο ακραία τιμή που αντιστοιχεί στην απόδοση της ομάδας Charlotte Hornets τη σεζόν 2010-2012 με απόδοση -15.32.



Σχήμα 3.32: Violin plot DEFF με βάση τις περιφέρειες (κανονική περίοδος)

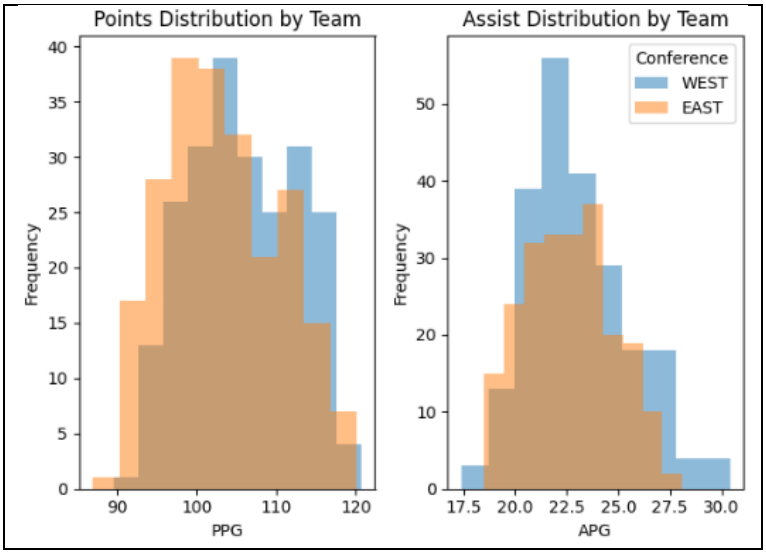
Για την τελευταία μεταβλητή που χρησιμοποιούμε προς ανάλυση διακρίνουμε πως δεν υπάρχουν ακραίες τιμές για καμία περιφέρεια και πως η διάμεσος των τιμών της WEST είναι υψηλότερη.



Σχήμα 3.33: Violin plot POSSt με βάση τις περιφέρειες (κανονική περίοδος)

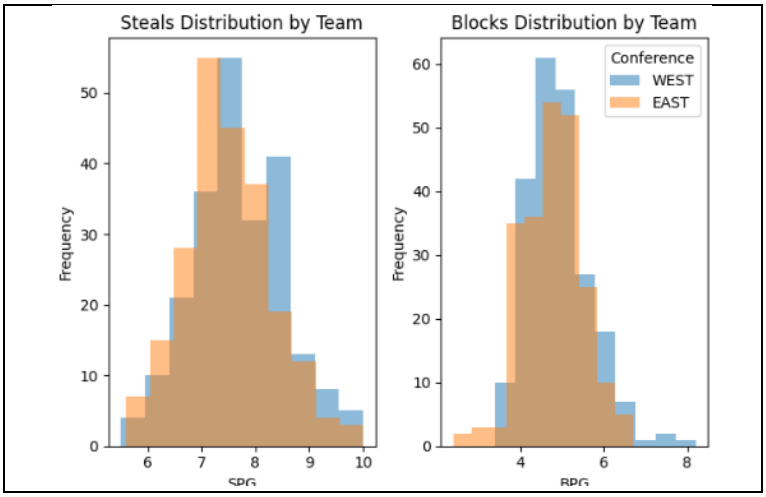
Ακόμα μια μέθοδος που χρησιμοποιείται για τη γραφική απεικόνιση στατιστικών συχνοτήτων είναι τα ιστογράμματα. Πιο συγκεκριμένα, παρουσιάζεται με κατακόρυφα ορθογώνια που η επιφάνεια κάθε ενός απ' αυτά είναι μέτρο της συχνότητας εμφάνισής του και το ύψος του ισούται με τον λόγο συχνότητας προς το εύρος των τιμών που αντιπροσωπεύει το εκάστοτε ορθογώνιο.

Στα σχήματα που ακολουθούν για τις μεταβλητές που συμμετέχουν στην ανάλυση με πορτοκαλί χρώμα διακρίνονται οι τιμές των μεταβλητών και η συχνότητα παρουσίασης αυτών. Με πορτοκαλί χρώμα απεικονίζονται οι τιμές της περιφέρειας EAST και με μπλε χρώμα τα ιστογράμματα των τιμών της αντίπαλης περιφέρειας. Στο Σχήμα 3.34 παρουσιάζεται αριστερά το ιστόγραμμα για τους πόντους που πετυχαίνουν οι ομάδες σε κάθε περιφέρεια και δεξιά για τις ασίστ. Στον κατακόρυφο άξονα παρουσιάζεται η συχνότητα παρουσίασης των τιμών που αναφέρονται στους οριζόντιους άξονες. Όσον αφορά τους πόντους που πετυχαίνουν οι ομάδες φαίνεται πως οι ομάδες της WEST περιφέρειας πετυχαίνουν συχνά περίπου 105 πόντους ανά αγώνα, ενώ οι ομάδες της αντίπαλης περιφέρειας γύρω στους 102. Για τις ασίστ είναι ορατό πως οι περιφέρειες λαμβάνουν σχεδόν τις ίδιες μέσες τιμές.



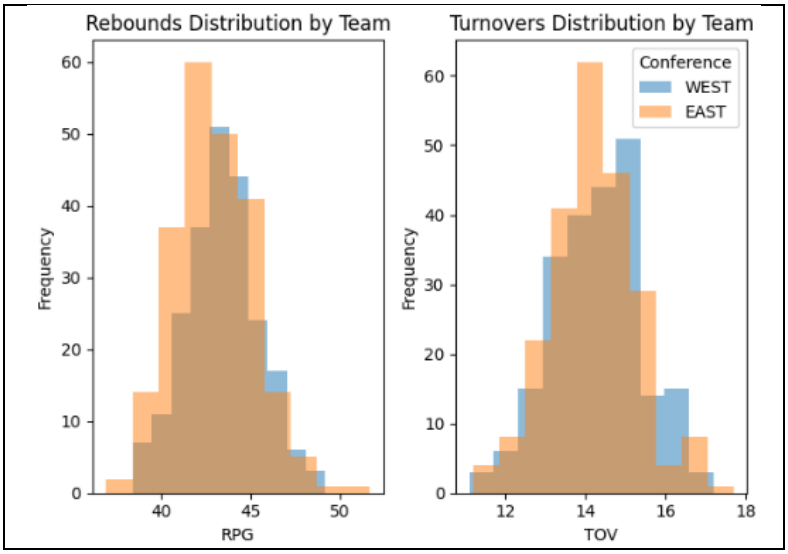
Σχήμα 3.34: Ιστογράμματα των περιφερειών (κανονική περίοδος)

Στο Σχήμα 3.35, που παρουσιάζονται τα ιστογράμματα των κλεψιμάτων και των κοψιμάτων, παρατηρούμε πως στην πρώτη περίπτωση η περιφέρεια των WEST λαμβάνει τιμές γύρω στα 7.5-7.8 κλεψίματα ενώ στην άλλη περιφέρεια τιμές από 7.2-7.4. Για τα κοψίματα που επιτυγχάνουν οι ομάδες παρατηρούμε πως η περιφέρεια των EAST λαμβάνει μικρότερη μέση τιμή.



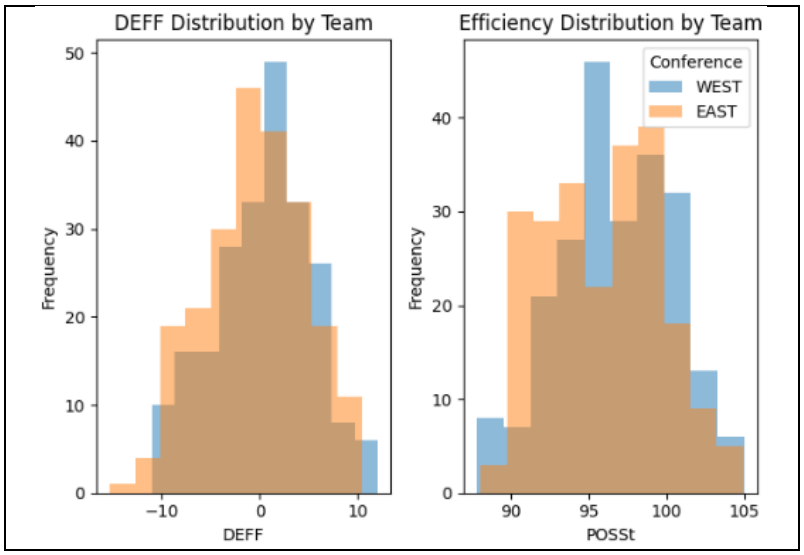
Σχήμα 3.35: Ιστογράμματα των περιφερειών (κανονική περίοδος)

Στη συνέχεια για τις μεταβλητές ριμπάουντ ανά αγώνα και λαθών ανά αγώνα, στο Σχήμα 3.36, παρατηρούμε ξανά πως η περιφέρεια WEST υπερτερεί σε τιμές από την αντίπαλη περιφέρεια.



Σχήμα 3.36: Ιστογράμματα των περιφερειών (κανονική περίοδος)

Για τις δύο τελευταίες μεταβλητές που έχουν σχέση με τον δείκτη DEFF και τις κατοχές που κερδίζουν οι ομάδες ανά αγώνα παρατηρούμε πως και στις δύο υψηλότερες τιμές λαμβάνουν οι ομάδες που συμμετέχουν στην WEST περιφέρεια. Πιο συγκεκριμένα, για τις τιμές του δείκτη DEFF φαίνεται πως η περιφέρεια EAST λαμβάνει αρνητική τιμή κατά μέσο όρο, γύρω στο -0.5 έως και 0, ενώ η άλλη περιφέρεια κινείται στα θετικά κομμάτια του άξονα.



Σχήμα 3.37: Ιστογράμματα των περιφερειών (κανονική περίοδος)

Εν κατακλείδι, φαίνεται πως εκτός από τα κλεψίματα, τα συνολικά ριμπάουντ και τα λάθη οι μέσοι όροι των μεταβλητών κατανέμονται μη-κανονικά και για τις δύο περιφέρειες.

3.5.2. Playoffs

Όπως αναφέρθηκε και προηγουμένως, θα ακολουθήσουμε την ίδια στατιστική ανάλυση και για την φάση των playoffs. Αρχικά θα παρουσιαστούν τα περιγραφικά μέτρα των μεταβλητών που επιλέχθηκαν προς ανάλυση.

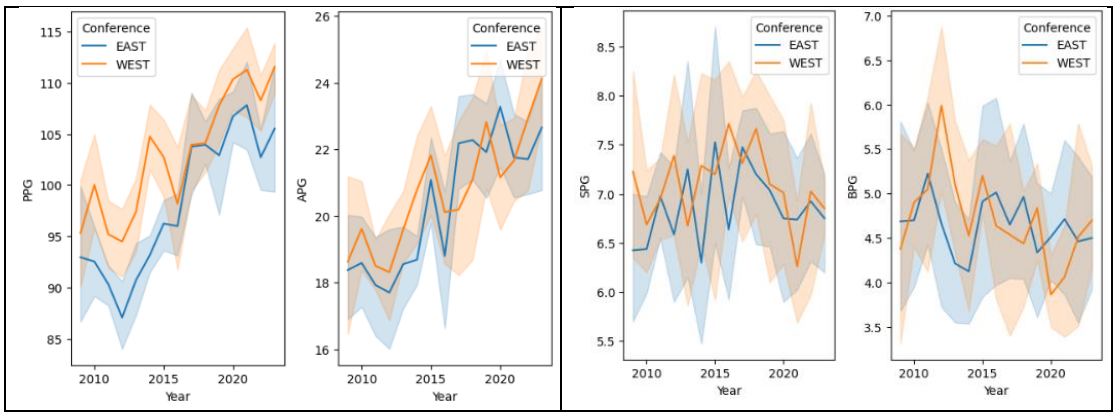
EAST	Min.	1st Q.	Median	Mean	3rd Q.	Max.
PPG	78.00	91.80	97.05	98.16	105.27	118.80
APG	12.60	18.27	20.35	20.36	22.50	26.90
SPG	4.30	6.17	6.85	6.86	7.52	10.80
BPG	2.20	3.80	4.50	4.64	5.50	8.10
RPG	33.50	39.10	41.20	41.54	43.85	49.90
TOV	9.40	12.47	13.40	13.59	14.70	18.40
DEFF	-25.87	-8.27	-2.58	-3.23	2.31	10.40
POSSt	82.02	89.48	92.83	93.29	96.45	104.70

Πίνακας 3.7: Στατιστικά περιγραφικά μέτρα περιφέρειας EAST (playoffs)

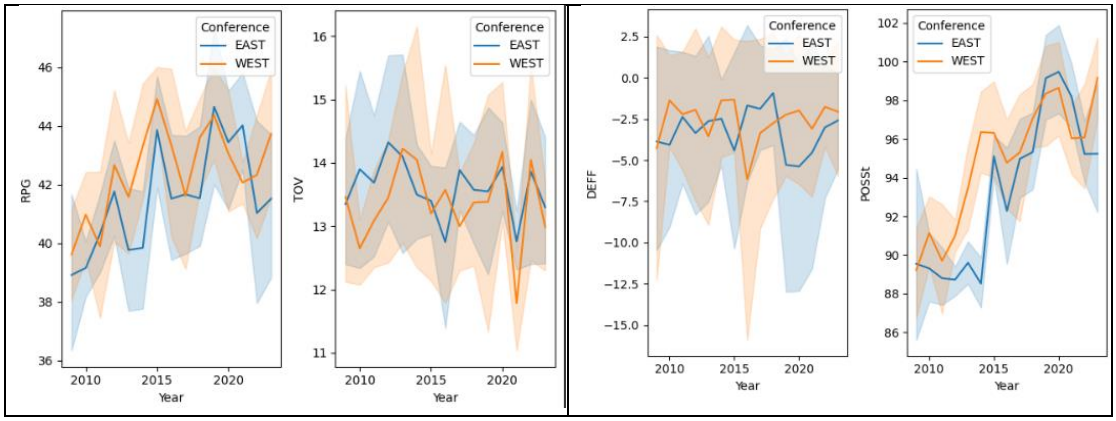
WEST	Min.	1st Q.	Median	Mean	3rd Q.	Max.
PPG	81.00	97.40	103.35	103.03	109.02	119.50
APG	14.70	18.37	20.55	20.75	22.65	28.40
SPG	3.40	6.27	7.00	7.09	8.00	9.80
BPG	2.10	3.80	4.60	4.71	5.50	8.00
RPG	33.50	40.20	42.85	42.47	44.80	49.90
TOV	8.40	12.10	13.10	13.36	14.50	21.00
DEFF	-26.60	-6.86	-2.25	-2.63	1.90	13.38
POSSt	82.63	91.86	94.67	94.84	98.29	104.43

Πίνακας 3.8: Στατιστικά περιγραφικά μέτρα περιφέρειας WEST (playoffs)

Στη συνέχεια παρουσιάζονται διαγράμματα time series με την πάροδο των χρόνων. Όπως είναι φανερό στα σχήματα 3.38 και 3.39, στις διαγραμματικές απεικονίσεις οι ομάδες που κατατάσσονται στην περιφέρεια WEST λαμβάνουν μεγαλύτερες τιμές στις μεταβλητές με την πάροδο του χρόνου σε σχέση με εκείνες της EAST. Στα συγκεκριμένα γραφήματα φαίνεται η πορεία των μέσων τιμών ανά χρονιά (έντονη γραμμή) αλλά και το εύρος των τιμών με πιο απαλό χρώμα γύρω από την γραμμή που αναφέρθηκε προηγουμένως. Από τη σεζόν 2019-2020 και έπειτα παρατηρούμε πως αρχίζουν οι τιμές των ομάδων της περιφέρειας EAST να πλησιάζουν και σε κάποιες μεταβλητές, τελικώς, να ξεπερνάνε τις τιμές της αντίπαλης περιφέρειας.



Σχήμα 3.38: Διαγράμματα χρονοσειρών των περιφερειών (playoffs)



Σχήμα 3.39: Διαγράμματα χρονοσειρών των περιφερειών (playoffs)

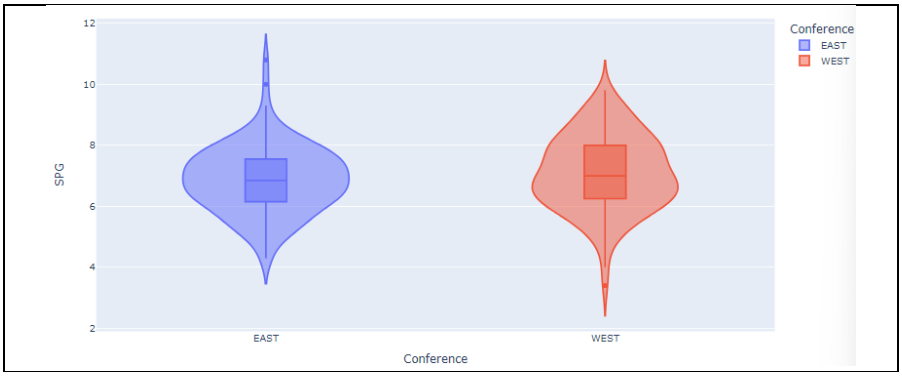
Στη συνέχεια παρουσιάζονται τα violin plots για τις μεταβλητές και τις τιμές τους στη φάση των playoffs. Για τους πόντους ανά αγώνα παρατηρούμε πως η διάμεσος των τιμών της περιφέρειας των WEST λαμβάνει αρκετά μεγαλύτερη τιμή. Ενδεικτικά η τιμή της είναι κοντά στους 105 πόντους, όταν η αντίστοιχη τιμή της άλλης περιφέρειας δεν ξεπερνάει την τιμή των 100 πόντων ανά αγώνα. Επιπλέον στο γράφημα δεν παρατηρούνται ακραίες τιμές.

Σχήμα 3.40: Violin plot PPG με βάση τις περιφέρειες (playoffs)

Για τις ασίστ που επιτυγχάνουν οι ομάδες των περιφερειών δεν παρατηρούμε μεγάλη διαφορά στη διάμεσο των τιμών, ενώ ενδεικτικά φαίνεται πως η διάμεσος των ομάδων WEST λαμβάνει οριακά μεγαλύτερη τιμή. Όπως και προηγουμένως, βάσει του γραφήματος δεν παρατηρούμε να εμφανίζονται ακραίες τιμές.

Σχήμα 3.41: Violin plot APG με βάση τις περιφέρειες (playoffs)

Όσον αφορά τα κλεψίματα ανά αγώνα των περιφερειών παρατηρούμε πως η περιφέρεια WEST λαμβάνει οριακά μεγαλύτερη τιμή διαμέσου. Επίσης, σε αυτό το γράφημα παρατηρούνται ακραίες τιμές που έχουν σχέση και με τις δύο περιφέρειες. Αρχικά, για την EAST περιφέρεια οι ακραίες τιμές έχουν σχέση με υψηλότερες τιμές. Τη σεζόν 2012-2013 και την 2014-2015 οι Milwaukee Bucks μέτρησαν κατά μέσο όρο 10.00 και 10.80 κλεψίματα ανά αγώνα αντίστοιχα. Για την περιφέρεια των WEST ακραία τιμή αποτέλεσε η απόδοση των Portland Trail Blazers τη σεζόν 2014-2015, που μέτρησαν κατά μέσο όρο 3.40 κλεψίματα ανά αγώνα.

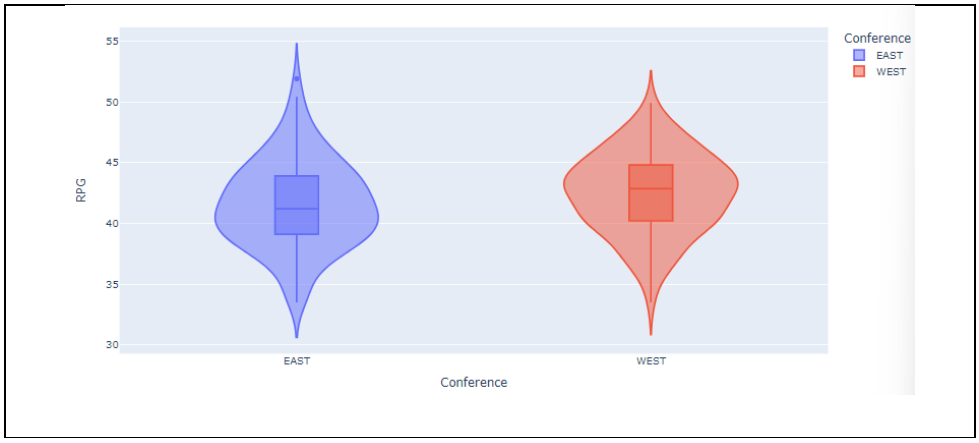


Σχήμα 3.42: Violin plot SPG με βάση τις περιφέρειες (playoffs)

Για τα κοψίματα ανά αγώνα που πετυχαίνουν οι ομάδες για κάθε περιφέρεια η τιμή της διαμέσου είναι οριακά, ξανά, υψηλότερη για την περιφέρεια των WEST. Ακόμα, στην περιφέρεια EAST υπάρχει μία ακραία τιμή, που λαμβάνεται τη σεζόν 2008-2009, όπου η ομάδα των Chicago Bulls μέτρησε κατά μέσο όρο 8.10 κοψίματα ανά αγώνα.

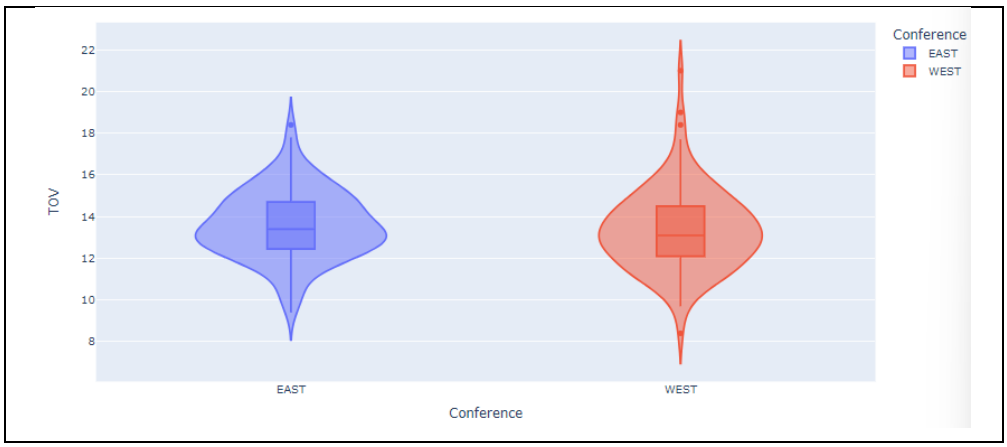
Σχήμα 3.43: Violin plot BPG με βάση τις περιφέρειες (playoffs)

Όπως και προηγουμένως, στα ριμπάουντ ανά αγώνα η περιφέρεια WEST λαμβάνει υψηλότερη τιμή από την αντίστοιχη αντίπαλη περιφέρεια, η οποία αγγίζει το 42.85. Για τις ακραίες τιμές έχουμε μόνο στην περιφέρεια EAST, όπου τη σεζόν 2018-2019 οι Milwaukee Bucks μέτρησαν κατά μέσο όρο 51.90 ριμπάουντ ανά αγώνα.



Σχήμα 3.44: Violin plot RPG με βάση τις περιφέρειες (playoffs)

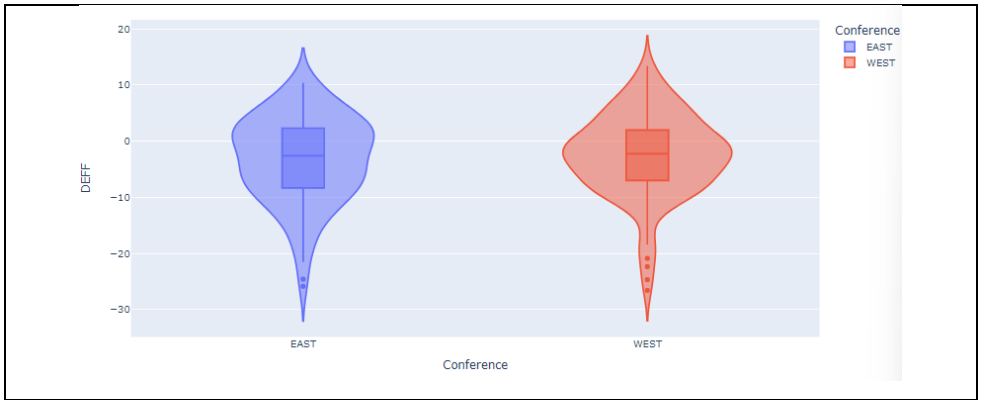
Για τα λάθη στα οποία υποπίπτουν οι ομάδες των περιφερειών στα playoffs παρατηρούμε για πρώτη φορά την περιφέρεια EAST να λαμβάνει υψηλότερη τιμή διαμέσου, περίπου ίση με 13.40. Στις τιμές αυτής της μεταβλητής διακρίνονται οι περισσότερες ακραίες τιμές μέχρι στιγμής στη φάση των playoffs. Αρχικά, τη σεζόν 2008-2009 οι New Orleans Pelicans έχουν απόδοση ίση με 18.40 για τα λάθη ανά αγώνα και ακολουθούν με την ίδια απόδοση οι New York Knicks τη σεζόν 2011-2012. Σε υψηλότερη τιμή απόδοσης, 21.00, φτάνουν οι Golden State Warriors τη σεζόν 2014-2015 και οι Houston Rockets τη σεζόν 2015-2016 με απόδοση 19.00. Τέλος, έχουμε και στην περιφέρεια WEST μία κατώτερη ακραία τιμή ίση με 8.40 τη σεζόν 2018-2019 από τους San Antonio Spurs.



Σχήμα 3.45: Violin plot TOV με βάση τις περιφέρειες (playoffs)

Για τον δείκτη DEFF έχουμε και για τις δύο περιφέρειες αρνητική τιμή στη διάμεσο που αντιστοιχεί στις τιμές των ομάδων στην φάση των playoffs. Παρ' όλα αυτά έστω και οριακά η περιφέρεια των WEST έχει μεγαλύτερη τιμή. Από ακραίες τιμές υπάρχουν επτά (7), οι οποίες αντιστοιχούν σε τιμές χαμηλότερες από το σύνηθες. Αρχικά, τη σεζόν 2008-2009 οι ομάδες Detroit Pistons και New Orleans Pelicans μέτρησαν τιμές ίσες με -21.51 και -26.60 αντίστοιχα. Τη σεζόν 2012-2013 οι Los Angeles Lakers είχαν απόδοση ίση με -20.89, ενώ τη σεζόν 2015-2016 οι ομάδες των Memphis Grizzlies και Houston Rockets είχαν -24.86 και -22.36 αντίστοιχα. Τέλος, τη σεζόν 2018-2019 οι

Detroit Pistons είχαν απόδοση -24.60 και τη σεζόν 2019-2020 οι Brooklyn Nets σημείωσαν απόδοση ίση με -25.87.



Σχήμα 3.46: Violin plot DEFF με βάση τις περιφέρειες (playoffs)

Τελευταία μεταβλητή που συμμετέχει στην ανάλυση είναι αυτή των κατοχών που κερδίζουν οι ομάδες. Όπως είναι φανερό στο Σχήμα 3.13, οι ομάδες που απαρτίζουν της περιφέρεια WEST έχουν και την υψηλότερη τιμή διαμέσου, η οποία έχει οριακά μικρότερη τιμή από 95 κατοχές. Ακραίες τιμές δεν εμφανίζονται σε αυτή την μεταβλητή σε καμία περιφέρεια.

Σχήμα 3.47: Violin plot POSSt με βάση τις περιφέρειες (playoffs)

3.6. Ανάλυση με βάση τον διαχωρισμό σε Ομάδες

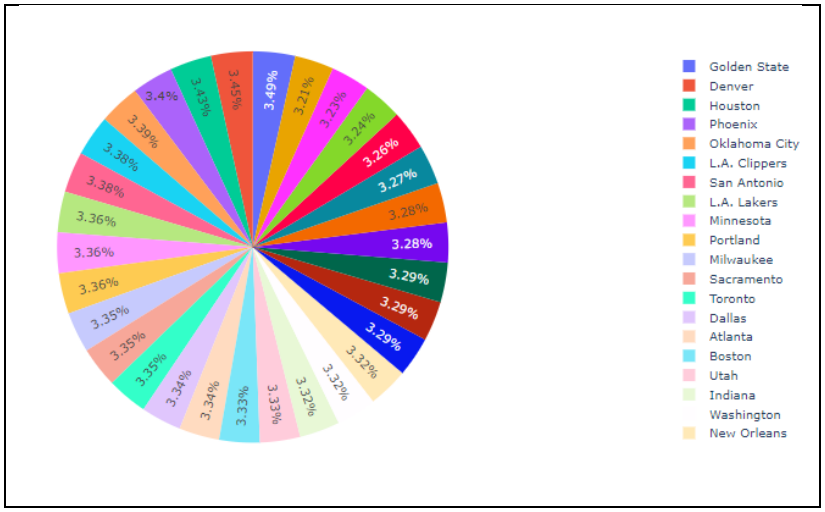
Στη συνέχεια της περιγραφικής ανάλυσης διερευνάται η ανάλυση των μεταβλητών ως προς την κατηγορική μεταβλητή των ομάδων. Ακολουθούν κάποια pie charts για τα στατιστικά των ομάδων τα τελευταία χρόνια, με σκοπό να βρεθεί η ομάδα που ξεχώρισε σε κάθε κατηγορία και για τις δύο φάσεις του πρωταθλήματος. Ο λόγος για τον οποίο χρησιμοποιούνται μόνο pie charts σε αυτή την ανάλυση είναι ο μεγάλος όγκος τιμών που λαμβάνει αυτή η μεταβλητή.

Συγκεκριμένα, στο πρωτάθλημα συμμετέχουν τριάντα (30) ομάδες και επομένως είναι δύσκολο να δημιουργηθούν άλλα γραφήματα για όλες τις ομάδες ξεχωριστά.

3.6.1. Κανονική περίοδος

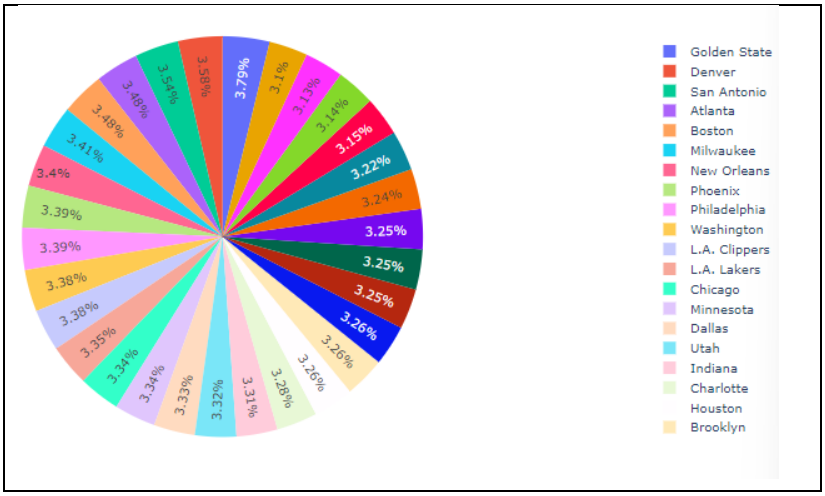
Ξεκινώντας με τη φάση της κανονικής περιόδου στο Σχήμα 3.48 διακρίνουμε την κατανομή των πόντων που έχουν επιτύχει οι ομάδες τα χρόνια αυτά. Σύμφωνα με το διάγραμμα η ομάδα των Golden State Warriors έχει το μεγαλύτερο ποσοστό πόντων

ίσο με το 3.49% από τους συνολικούς πόντους που έχουν επιτύχει όλες οι ομάδες. Το αποτέλεσμα αυτό είναι αναμενόμενο εφόσον είναι η ομάδα που ξεκίνησε πρώτη να εκμεταλλεύεται το πλεονέκτημα που έδινε ο κανόνας των τριών (3) δευτερολέπτων στην ρακέτα, δημιουργώντας το καλύτερο δίδυμο παικτών το οποίο απαρτίζεται από δύο παίκτες με τα καλύτερα ποσοστά τριπόντων στο NBA, τον Stephen Curry και τον Klay Thompson, και συμπληρώνοντας τις υπόλοιπες θέσεις της ομάδας με παίκτες εξίσου καλούς στο μακρινό σουτ (Sanjeev Lingam-Nattamai, 2022). Τελευταία ομάδα στην επιτυχία πόντων είναι οι Detroit Pistons με ποσοστό που αγγίζει το 3.20%.



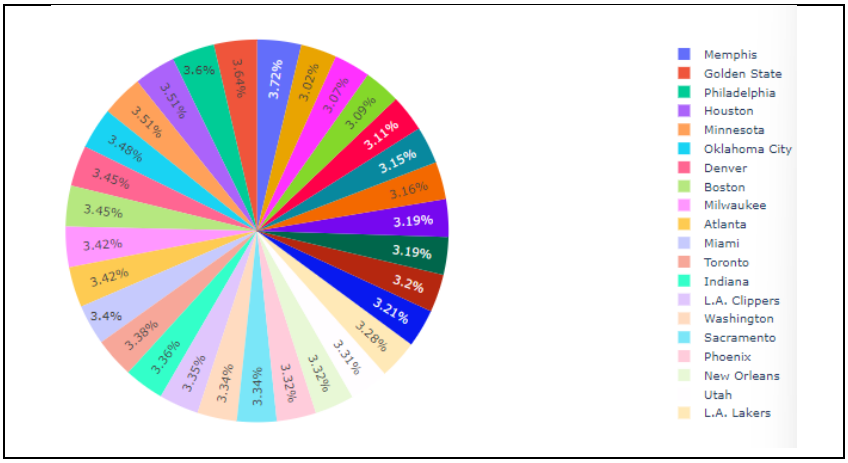
Σχήμα 3.48: Pie chart των PPG με βάση τις ομάδες (κανονική περίοδος)

Εν συνεχεία και στο Σχήμα 3.49 βρίσκονται τα ποσοστά κατανομής των ασίστ που έχουν μοιράσει οι ομάδες. Οι Golden State Warriors είναι πρώτοι με 3.79%, με μεγάλη διαφορά από τον δεύτερο που έχει ποσοστό 3.58%. Τελευταίοι στην κατηγορία των ασίστ βρίσκονται οι New York Nicks με το ποσοστό του 3.10%.



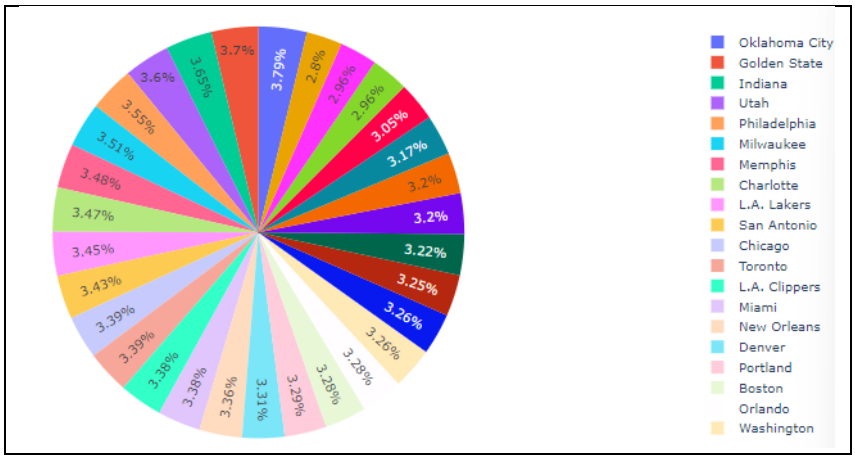
Σχήμα 3.49: Pie chart των APG με βάση τις ομάδες (κανονική περίοδος)

Όσον αφορά τα αμυντικά χαρακτηριστικά μίας ομάδας, κλεψίματα και κοψίματα,, στο Σχήμα 3.50 διακρίνουμε τους Memphis Grizzlies να είναι πρώτοι στην κατηγορία με το ποσοστό του 3.72% και τελευταίοι οι Portland.



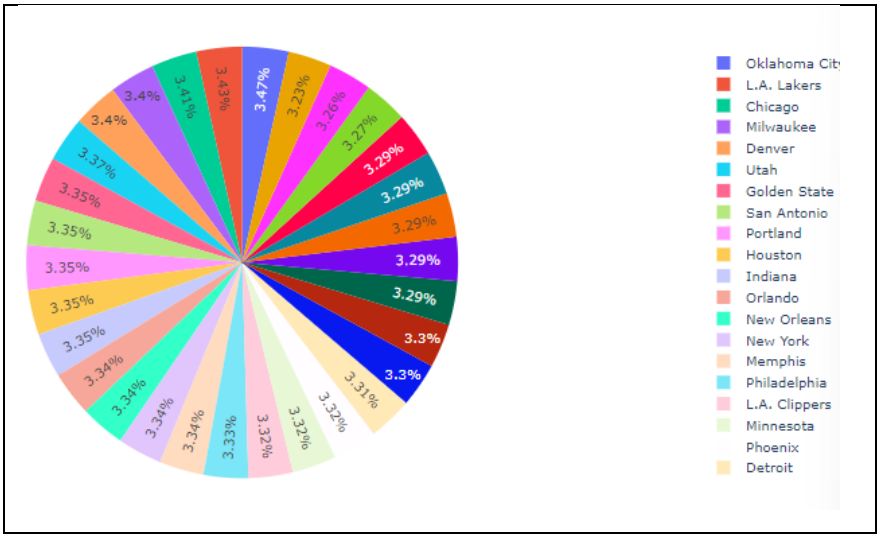
Σχήμα 3.50: Pie chart των SPG με βάση τις ομάδες (κανονική περίοδος)

Για τα κοψίματα, Σχήμα 3.51, πρώτοι της κατηγορίας είναι οι Oklahoma City Thunder με ποσοστό 3.79%. Αυτό ίσως αποτελεί και ένα αναμενόμενο αποτέλεσμα, εφόσον τα τελευταία χρόνια οι κινήσεις τους στην αγορά παιχτών έχουν στόχο ψηλούς guard με έφεση στο σουτ. Τελευταίοι της κατηγορίας είναι οι Cleveland Cavaliers με 2.8%.



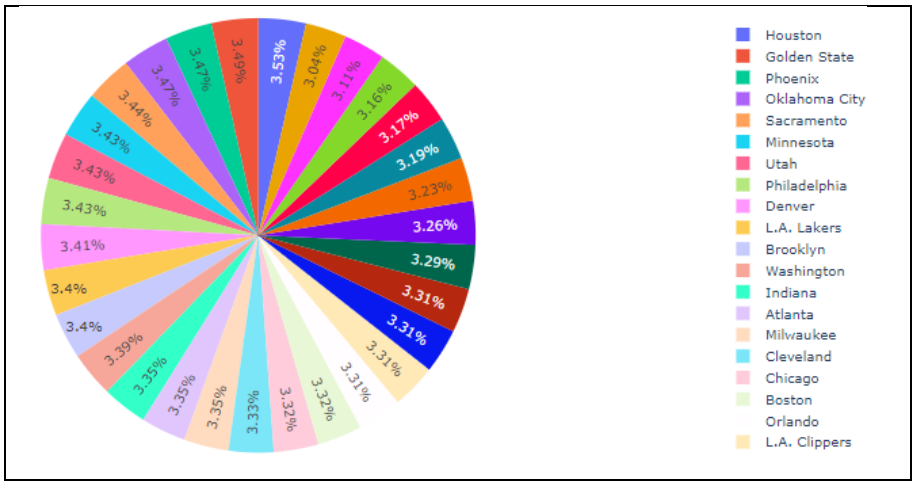
Σχήμα 3.51: Pie chart των BPG με βάση τις ομάδες (κανονική περίοδος)

Προηγουμένως αναλύθηκε ο τρόπος που επιλέγονται οι παίχτες των Oklahoma City Thunder οπότε για το Σχήμα 3.52 και τον πρώτο στην κατηγορία των ριμπάουντ βρίσκουμε τους ίδιους με ποσοστό της τάξης του 3.47%. Αντίθετα τελευταίοι σε αυτή την στατιστική κατηγορία βρίσκονται οι Miami Heat με ποσοστό 3.23%.



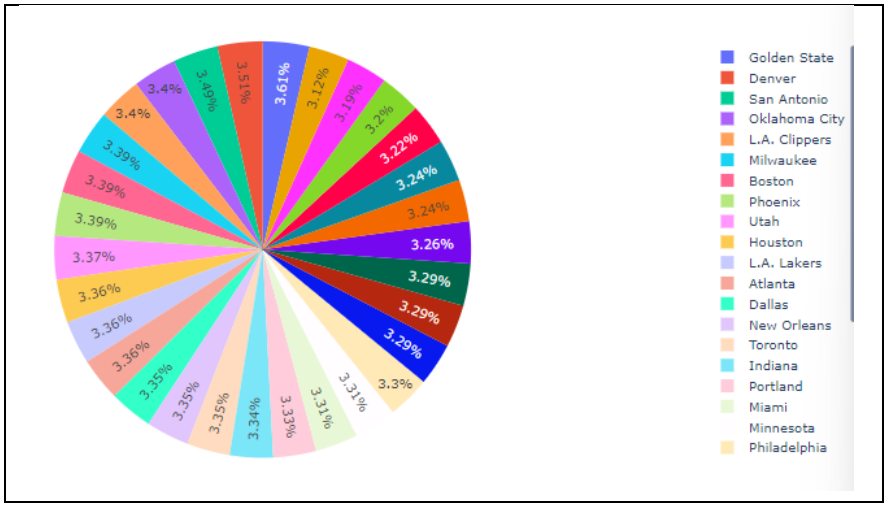
Σχήμα 3.52: Pie chart των RPG με βάση τις ομάδες (κανονική περίοδος)

Για τις τελευταίες μεταβλητές, στο Σχήμα 3.53 παρουσιάζονται τα ποσοστά των ομάδων σε σχέση με τα λάθη στα οποία υποπίπτουν κατά τη διάρκεια των αγώνων. Πρώτοι σε αυτοί την κατηγορία είναι οι Houston Rockets με ποσοστό 3.53%.

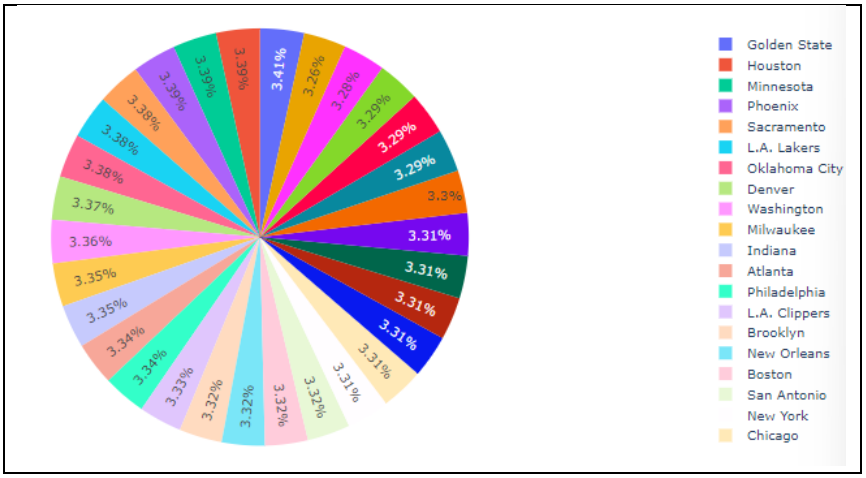


Σχήμα 3.53: Pie chart των TOV με βάση τις ομάδες (κανονική περίοδος)

Επίσης, σε αυτή την ανάλυση αντί για τον δείκτη του DEFF χρησιμοποιήθηκε ο δείκτης Efficiency των ομάδων, για τον οποίο μπορούσαμε να δημιουργήσουμε το συγκεκριμένο γράφημα, Σχήμα 3.54. Το μεγαλύτερο ποσοστό το κατέχει η ομάδα των Golden State Warriors με 3.41%. Τέλος, για τη μεταβλητή των κατοχών που κερδίζουν οι ομάδες, σχήμα 3.55, πρώτοι έρχονται ξανά οι Golden State Warriors με το ποσοστό 3.41% και τελευταίοι οι Miami Heat με ποσοστό ίσο με 3.26%.



Σχήμα 3.54: Pie chart των EFF με βάση τις ομάδες (κανονική περίοδος)

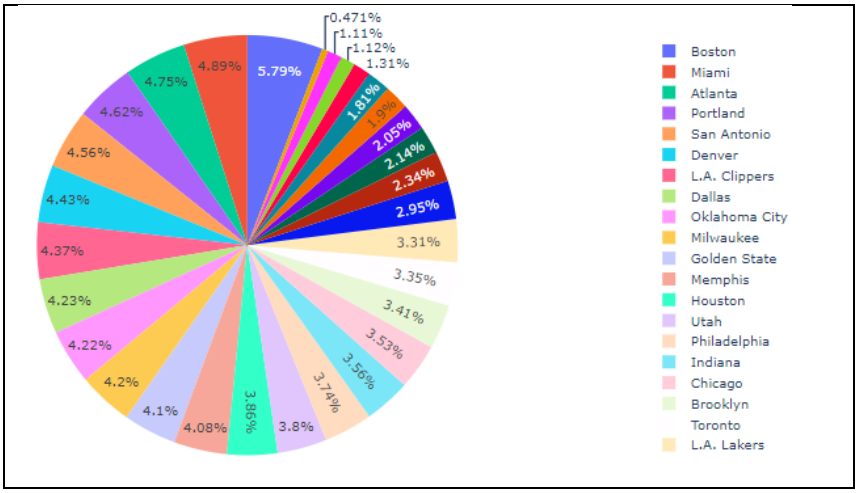


Σχήμα 3.55: Pie chart των POSSt με βάση τις ομάδες (κανονική περίοδος)

3.6.2. Playoffs

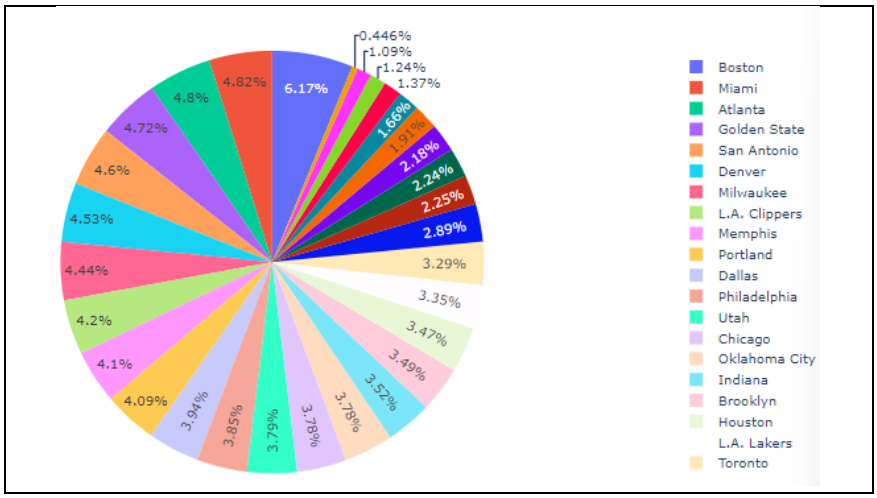
Αφού αναλύσαμε όλες τις μεταβλητές για τη κανονική περίοδο, θα συνεχίσουμε την ίδια ανάλυση και για την φάση των playoffs. Στο σχήμα 3.56 για τους πόντους που πετυχαίνουν οι ομάδες, πρώτοι έρχονται οι Boston Celtics με ποσοστό 5.79%. Παρατηρώντας το διάγραμμα αυτό αντιλαμβανόμαστε πως σε σύγκριση με την κανονική περίοδο στα playoffs διακρίνονται άλλες ομάδες στην στατιστική κατηγορία ως καλύτερες. Το συγκεκριμένο αποτέλεσμα είναι αναμενόμενο, διότι οι ομάδες οι οποίες καταφέρνουν να φτάσουν στους τελικούς της διοργάνωσης έχουν συμμετάσχει σε περισσότερα παιχνίδια από τις ομάδες που σταμάτησαν σε κάποια προηγούμενη φάση των playoffs. Παραδείγματος χάρη, αν μία ομάδα αποκλειστεί στον πρώτο γύρω των playoffs θα συμμετάσχει σε τέσσερα (4) έως επτά (7) παιχνίδια. Επομένως οι πόντοι θα μετρηθούν σε αυτά τα παιχνίδια μόνο, ενώ μια ομάδα που φτάνει στους τελικούς θα έχει παίξει πολλά περισσότερα παιχνίδια μέχρι το πέρας των φάσεων των playoffs και επομένως θα γεμίσει την στατιστική της. Ακόμα, είναι λογικό να βλέπουμε και μεγάλες αποκλίσεις στα ποσοστά του πρώτου και του τελευταίου, όπως γίνεται και στην κατανομή των πόντων όπου ο τελευταίος έχει ποσοστό της τάξης του 0.45%. Αυτό

συμβαίνει διότι η διαχρονική πορεία των ομάδων σε αυτή τη φάση του πρωταθλήματος είναι διαφορετική, αφού υπάρχουν χρονιές που πολλές ομάδες έμειναν εκτός των playoffs και κατά συνέπεια δεν πρόσθεσαν νέα στατιστικά στις κατηγορίες που ερευνώνται ώστε να βρίσκονται ψηλότερα στην λίστα.

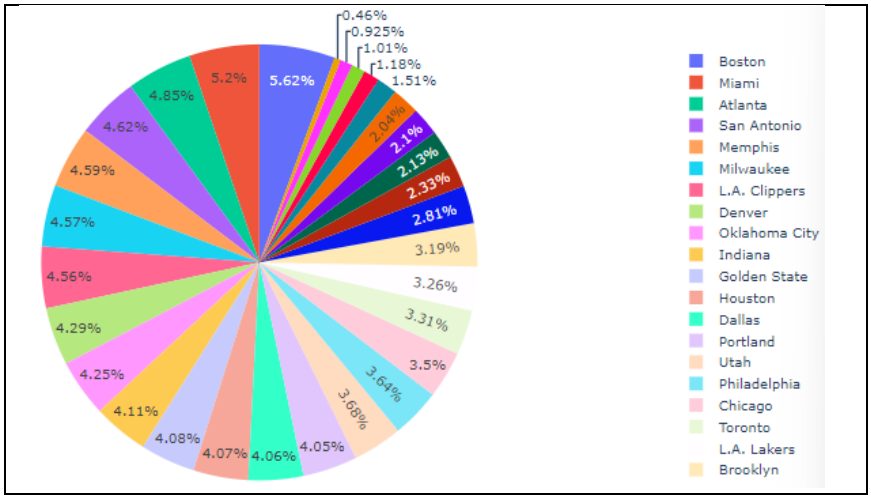


Σχήμα 3.56: Pie chart των PPG με βάση τις ομάδες (playoffs)

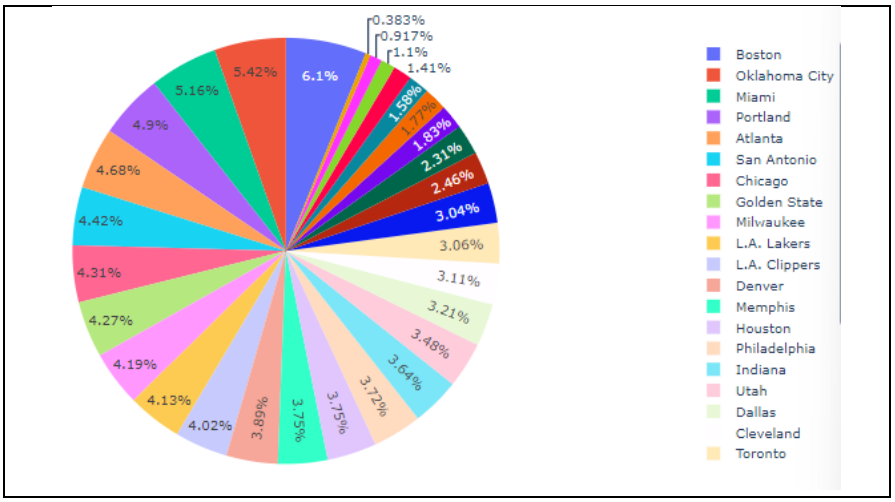
Στη συνέχεια ακολουθούν τα γραφήματα για όλες τις μεταβλητές που χρησιμοποιήθηκαν στην ανάλυση. Όπως φαίνεται σε αυτά, οι Boston Celtics είναι πρώτοι σε όλες τις κατηγορίες με αρκετά μεγαλύτερα ποσοστά από τους επομένους. Το συμπέρασμα αυτό πολύ πιθανόν να έχει προκύψει διότι είναι μία ομάδα που παραδοσιακά συμμετέχει σε όλα τα χρόνια στις φάσεις των playoffs, γεγονός που αποδεικνύεται στη συνέχεια της εργασίας. Αντίθετα, τελευταίοι σε όλες τις κατηγορίες βρίσκονται οι Sacramento με ποσοστά τα οποία δεν ξεπερνούν το 1%.



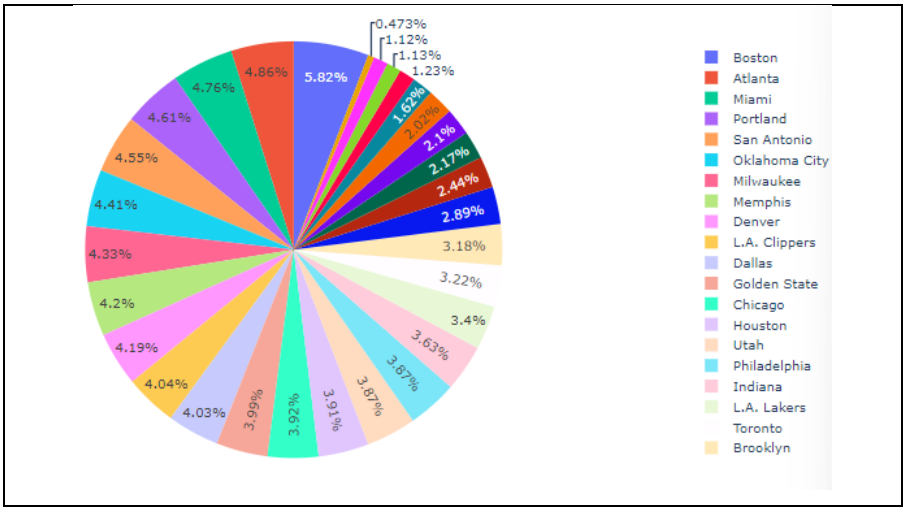
Σχήμα: 3.57: Pie chart των APG με βάση τις ομάδες (playoffs)



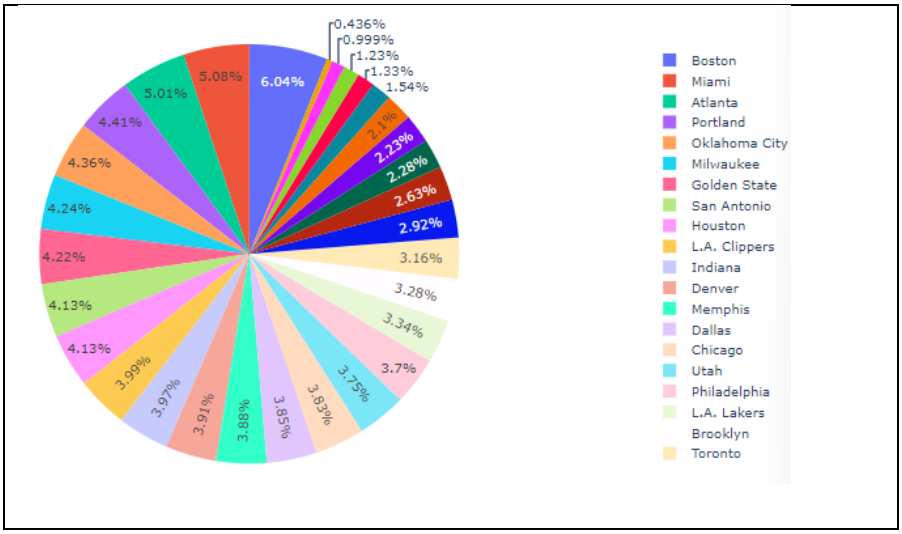
Σχήμα 3.58: Pie chart των SPG με βάση τις ομάδες (playoffs)



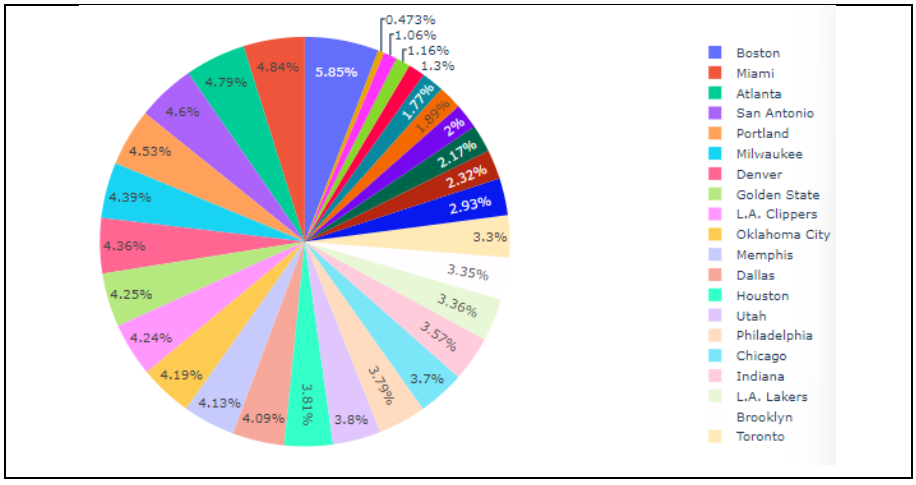
Σχήμα 3.59: Pie chart των BPG με βάση τις ομάδες (playoffs)



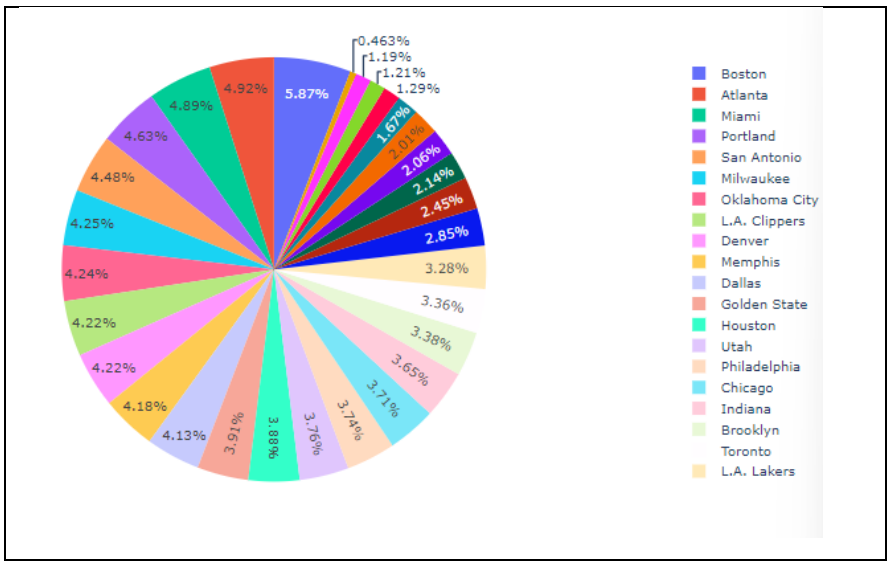
Σχήμα 3.60: Pie chart των RPG με βάση τις ομάδες (playoffs)



Σχήμα 3.61: Pie chart των TOV με βάση τις ομάδες (playoffs)



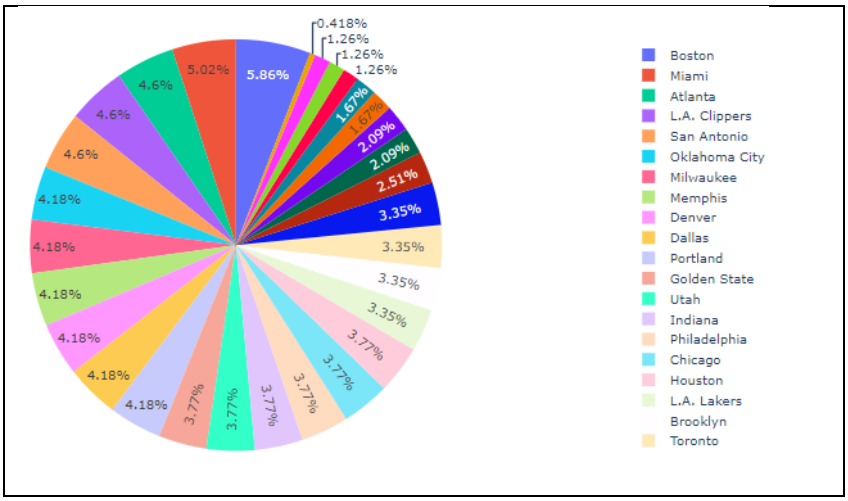
Σχήμα 3.62: Pie chart του δείκτη EFF με βάση τις ομάδες (playoffs)



Σχήμα 3.63: Pie chart των POSSt με βάση τις ομάδες (playoffs)

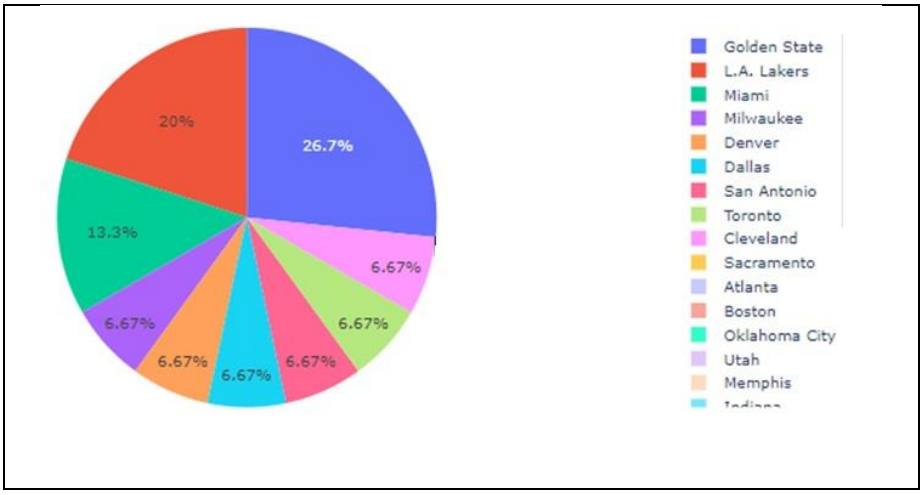
Όπως έχει αναφερθεί ήδη, στη φάση των playoffs δεν έχουν συμμετάσχει όλες οι ομάδες με την ίδια συχνότητα και, αν έχουν υπάρξει ομάδες που συμμετείχαν με την

ίδια συχνότητα, σίγουρα δεν έχουν μετρήσει τα ίδια στατιστικά, εφόσον μπορεί κάποιες να έχουν αποκλειστεί σε διαφορετικά στάδια στη φάση αυτή. Ο καθένας θα μπορούσε να αντιληφθεί πως οι Boston Celtics, οι οποίοι φάνηκε από την προηγούμενη ανάλυση πως ήταν πρώτοι σε όλες τις κατηγορίες που μελετήθηκαν, θα είναι μία από τις ομάδες με τις περισσότερες συμμετοχές στη φάση των playoffs. Στο Σχήμα 3.64 αναφέρονται τα ποσοστά συμμετοχών στα playoffs για όλες τις ομάδες διαχρονικά. Όπως είναι φανερό, οι Boston Celtics έχουν προκριθεί δεκατέσσερις (14) φορές στη συγκεκριμένη φάση, με ποσοστό που αντιστοιχεί στο 5.86%, ενώ οι δεύτεροι, Miami Heat, έχουν δώδεκα (12) συμμετοχές. Επίσης, στην ανάλυση της κανονικής περιόδου είδαμε πως οι Golden State Warriors βρίσκονταν σε πολλές κατηγορίες πρώτοι, ενώ στα playoffs δεν παρατηρήσαμε κάτι παρόμοιο. Αυτό το γεγονός συμβαίνει διότι έχουν προκριθεί στην εν λόγω διοργάνωση του πρωταθλήματος μόνο εννέα (9) φορές από τις τελευταίες δεκαπέντε (15) περιόδους που αναλύουμε.



Σχήμα 3.64: Pie chart για τις συμμετοχές των ομάδων στα playoffs

Τέλος, δημιουργήθηκε και το γράφημα που παρουσιάζει τα ποσοστά των ομάδων που έχουν αναδειχθεί πρωταθλήτριες τα τελευταία χρόνια. Πρώτοι στα πρωταθλήματα με ποσοστό 26.7% είναι οι Golden State Warriors, δηλαδή έχουν στεφθεί πρωταθλητές τέσσερις (4) φορές. Έπειτα ακολουθούν οι L.A. Lakers με το 20% των πρωταθλημάτων, ποσοστό που αντιστοιχεί σε τρείς (3) στέψεις και οι Miami Heat με δύο (2), 13.30% ποσοστό. Τέλος υπάρχουν έξι (6) διαφορετικές ομάδες οι Milwaukee Bucks, Denver Nuggets, Dallas Mavericks, San Antonio Spurs, Toronto Raptors και Cleveland Cavaliers που έχουν κατακτήσει από ένα (1) πρωτάθλημα τα τελευταία χρόνια. Επομένως, τα τελευταία δεκαπέντε (15) χρόνια έχουν στεφθεί συνολικά εννέα (9) διαφορετικοί πρωταθλητές εκ των οποίων, όπως αναλύθηκε και προηγουμένως, τα δέκα (10) πρωταθλήματα έχουν καταλήξει στα χέρια της WEST περιφέρειας.



Σχήμα 3.65: Pie chart για τα πρωταθλήματα των ομάδων

ΚΕΦΑΛΑΙΟ 4ο

4. Στατιστικοί έλεγχοι και συσχετίσεις μεταξύ των μεταβλητών

Στο παρόν κεφάλαιο της εργασίας θα εκτελεστούν οι έλεγχοι κανονικότητας για τις ποσοτικές μεταβλητές που συμμετέχουν στην ανάλυσή μας και έπειτα λόγω του μεγάλου πλήθους μεταβλητών και της ανάγκης να βρεθούν αυτές που καθορίζουν την απόδοση μιας ομάδας γεννάται η ανάγκη να εξεταστούν οι συσχετίσεις μεταξύ αυτών, ώστε να έχουμε μια πρώτη εικόνα, όχι καθοριστική καθώς δεν διαθέτει την ακρίβεια που έχει η κατάρτιση ενός μοντέλου. Τέλος, θα γίνουν και κάποιοι έλεγχοι για τις μέσες τιμές ορισμένων μεταβλητών βάσει των αναλύσεων που έχουν προηγηθεί.

ΣΧΟΛΙΟ: Για όλους τους ελέγχους, το επίπεδο σημαντικότητας με το οποίο εργαστήκαμε ισούται με 5%.

4.1 Έλεγχοι κανονικότητας

Ένας από τους βασικότερους στατιστικούς ελέγχους υποθέσεων είναι ο έλεγχος της κανονικότητας που θα αναλυθεί στην παρούσα παράγραφο. Με τον όρο κανονικότητα των δεδομένων εννοούμε πως το δείγμα που έχουμε προς ανάλυση προέρχεται από ένα σύνολο δεδομένων το οποίο ακολουθεί την κανονική κατανομή. Ο παρών έλεγχος θεωρείται από τους σπουδαιότερους στη Στατιστική, διότι μας επιτρέπει αναλόγως του αποτελέσματός του να αποφασίσουμε αν θα κάνουμε χρήση παραμετρικού ή μη παραμετρικού ελέγχου στην εξέταση επόμενων μηδενικών υποθέσεων για τον έλεγχο μέσω τιμών, διαμέσων κ.ά.. Η μορφή των στατιστικών ελέγχων αυτού του είδους είναι:

Μηδενική υπόθεση H_0 : η κατανομή από την οποία προέρχεται το δείγμα είναι η κανονική

vs

Εναλλακτική υπόθεση H_1 : η κατανομή από την οποία προέρχεται το δείγμα δεν είναι η κανονική

Για τη διεξαγωγή του ελέγχου υπάρχουν αρκετές διαφορετικές μέθοδοι, είτε γραφικά είτε με κάποιον συγκεκριμένο έλεγχο. Οι περισσότεροι έλεγχοι που υπάρχουν δεν μπορούν να ανταπεξέλθουν σε όλα τα μεγέθη των δειγμάτων και γι' αυτό παρακάτω θα αναφέρουμε διάφορους ελέγχους και τις προϋποθέσεις που πρέπει να ισχύουν για να τους χρησιμοποιήσουμε. (Ευαγγελάρας, 2022)

Kolmogorov-Smirnov (K-S)

Το κριτήριο K-S χρησιμοποιείται για τον έλεγχο καλής προσαρμογής ενός δείγματος σε μία δεδομένη συνεχή κατανομή. Συγκεκριμένα, βασίζεται στην διαφορά της εμπειρικής συνάρτησης κατανομής, που προέρχεται από το δείγμα, και της αναμενόμενης συνάρτησης κατανομής υπό την μηδενική υπόθεση. Το βασικό μειονέκτημα αυτού του ελέγχου σχετίζεται με το γεγονός πως πρέπει να είναι γνωστές οι παράμετροι της κατανομής πριν εκτελεστεί. (Μπούτσικας, 2004)

Shapiro-Wilk

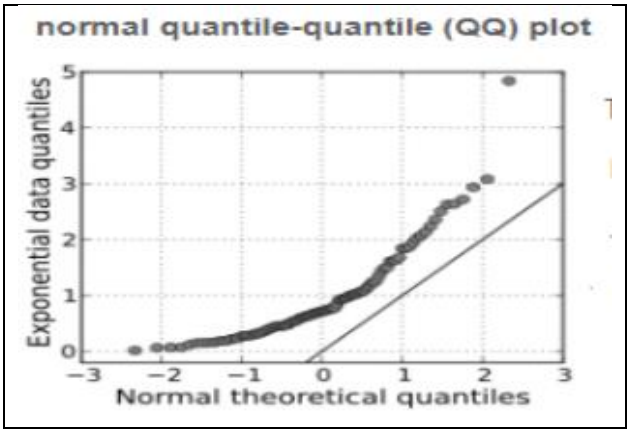
Ένας έλεγχος ο οποίος προτιμάται για μικρά δείγματα (μέγεθος < 50) έναντι του ελέγχου K-S είναι ο έλεγχος Shapiro-Wilk και γενικά έχει μεγάλη ισχύ. Ειδικότερα, η στατιστική συνάρτηση ελέγχου αποτιμά πόσο κοντά είναι τα εμπειρικά ποσοστιαία σημεία του δείγματος από τα αντίστοιχα θεωρητικά ποσοστιαία σημεία της αντίστοιχης κανονικής κατανομής. (Αντζουλάκος, 2021)

Έλεγχος Lilliefors

Όπως προαναφέρθηκε, για να εφαρμοστεί ο έλεγχος Kolmogorov – Smirnov, πρέπει να είναι εκ των προτέρων γνωστές οι παράμετροι της κατανομής του πληθυσμού. Επομένως γίνεται αντιληπτό πως αν δεν είναι γνωστές οι παράμετροι και συμπληρωθούν με εκτιμήσεις το αποτέλεσμα κρίνεται αναξιόπιστο. Σε αυτή τη δεδομένη περίπτωση είναι προτιμότερο να χρησιμοποιήσουμε μια παραλλαγή του ελέγχου K-S, τον έλεγχο Lilliefors. (Αντζουλάκος, 2021)

Ακόμα, υπάρχουν και περαιτέρω έλεγχοι που δεν είναι τόσο διαδεδομένοι όπως ο Anderson-Darling, Shapiro-Francia, Cramer-von Mises και Perason chi-square.

Εκτός από τους παραπάνω ελέγχους, όπως αναφέρθηκε και προηγουμένως, μπορεί να γίνει ο έλεγχος κανονικότητας και μέσω γραφήματος. Ένας τέτοιος έλεγχος είναι το γράφημα QQ-plot. Το γράφημα χρησιμοποιείται για να συγκριθούν κατά πόσο δύο δείγματα προέρχονται από την ίδια κατανομή. Πιο συγκεκριμένα, απεικονίζονται τα ποσοστημόρια της μίας κατανομής σε σχέση με της άλλης.



Σχήμα 4.1: Παράδειγμα ενός Q-Q plot.

(Πηγή: https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot)

Στο συγκεκριμένο παράδειγμα που παρουσιάστηκε συγκρίνεται η θεωρητική κανονική κατανομή (άξονας x) με ένα σετ δεδομένων (άξονας y). Όπως είναι φανερό, τα σημεία δεν συμβαδίζουν με την ευθεία των 45^{o} μοιρών και επομένως το δείγμα φαίνεται πως δεν ακολουθεί την κανονική κατανομή. (Κάτρης, 2021)

4.1.1. Έλεγχοι κανονικότητας για όλες τις μεταβλητές

Στην παρούσα παράγραφο θα γίνει διερεύνηση για το αν υπάρχουν μεταβλητές ανάμεσα στα δεδομένα μας που ακολουθούν την κανονική κατανομή για τις δύο φάσεις του πρωταθλήματος. Ο έλεγχος που θα χρησιμοποιηθεί είναι ο Shapiro-Wilk, που αναφέρθηκε και παραπάνω. Εφόσον γίνει ο έλεγχος κανονικότητας, για να

ισχυριστούμε πως απορρίπτουμε την μηδενική υπόθεση, δηλαδή πως τα δεδομένα μας δεν προέρχονται από κανονική κατανομή, θα πρέπει η τιμή του p_{value} του ελέγχου να είναι μικρότερη του 0.05, εφόσον σαν επίπεδο σημαντικότητας χρησιμοποιούμε το 5%.

4.1.1.1. Έλεγχοι για την κανονική περίοδο

Όπως φαίνεται στον Πίνακα 4.1, για τις μεταβλητές: ποσοστά ευστοχίας δίποντων ή τρίποντων (FG%), ποσοστό εύστοχων τρίποντων (3P%), επιθετικά ριμπάουντ (ORB), συνολικά ριμπάουντ (RPG), επιθετικά ριμπάουντ της αντίπαλης ομάδας (ORBo), λάθη της ομάδας (TOV) και της αντιπάλου (TOVo), φάουλ (PF) και defensive rating (DRt) τα p_{value} του ελέγχου είναι μεγαλύτερα από το 0.05 και επομένως δεν υπάρχουν ενδείξεις απόρριψης της μηδενικής υπόθεσης, δηλαδή τα δείγματα ακολουθούν την κανονική κατανομή. Για τις υπόλοιπες μεταβλητές τα p_{value} είναι μικρότερα και συνεπώς θα πρέπει να απορρίψουμε τη μηδενική υπόθεση του ελέγχου.

Μεταβλητή	p_{value} Shapiro-Wilk
PPG	**2.79959e-06**
PPGA	**0.00063**
FGM	**0.00036**
FGA	**0.01596**
MisFG	**0.02392**
FG%	**0.34156**
FGAo	**0.00591**
3PM	**3.94893e-06**
3PA	**2.04131e-07**
3P%	**0.67951**
FTM	**0.00043**
FTA	**0.00179**
MisFT	**6.84685e-08**
FT%	**0.00103**
FTAo	**0.01312**
ORB	**0.39897**
DRB	**0.00061**
RPG	**0.23202**
ORBo	**0.07283**
APG	**0.00012**
SPG	**0.02554**
BPG	**2.40231e-05**
TOV	**0.37460**
TOVo	**0.34047**
PF	**0.33963**
POSSt	**0.00182**
ORt	**0.04500**

POSSo	**0.00121**
DRt	**0.68410**
DEFF	**0.00571**
EFF	**0.00063**

Πίνακας 4.1: Αποτελέσματα ελέγχου κανονικότητας των μεταβλητών

4.1.1.1. Έλεγχοι για τα playoffs

Σύμφωνα με τον Πίνακα 4.2, για τις μεταβλητές: εύστοχα δίποντα και τρίποντα (FGM), συνολικές προσπάθειες δίποντων και τριπόντων της ομάδας (FGA) και των αντιπάλων τους (FGAo), άστοχα δίποντα και τρίποντα (MisFG), ποσοστό εύστοχων τριπόντων (3P%), εύστοχες ελεύθερες βολές (FTM), ποσοστό εύστοχων ελεύθερων βολών (FT%), επιθετικά ριμπάουντ (ORB), συνολικά ριμπάουντ (RPG), κλεψίματα (SPG), λάθη της αντίπαλης ομάδας (TOVo), κατοχές της ομάδας (POSSt) και της αντιπάλου (POSSo), offensive rating (ORt) και efficiency (EFF) συμπεραίνουμε πως η υπόθεση της κανονικότητας δεν απορρίπτεται, εφόσον οι τιμές του p_{value} του ελέγχου είναι μεγαλύτερες του επιπέδου σημαντικότητας. Για τις υπόλοιπες μεταβλητές τα p_{value} είναι μικρότερα και συνεπώς θα πρέπει να απορρίψουμε την μηδενική υπόθεση του ελέγχου.

Μεταβλητή	p_{value} Shapiro-Wilk
PPG	**0.04127**
PPGA	**0.01177**
FGM	**0.48011**
FGA	**0.64615**
MisFG	**0.10962**
FG%	**0.01846**
FGAo	**0.05648**
3PM	**0.00962**
3PA	**0.00037**
3P%	**0.13966**
FTM	**0.17687**
FTA	**0.00888**
MisFT	**1.40631e-05**
FT%	**0.13210**
FTAo	**0.00591**
ORB	**0.06868**
DRB	**0.04585**
RPG	**0.93303**
ORBo	**0.03033**
APG	**0.40653**
SPG	**0.90782**
BPG	**0.01451**
TOV	**0.00643**
TOVo	**0.30060**

PF	**0.00191**
POSSt	**0.08399**
Ort	**0.13519**
POSSo	**0.14360**
DRt	**0.02876**
DEFF	**0.00048**
EFF	**0.24027**

Πίνακας 4.2: Αποτελέσματα ελέγχου κανονικότητας των μεταβλητών

4.1.2. Έλεγχοι κανονικότητας με βάση την πρόκριση στα playoffs

Στη συνέχεια θα γίνουν οι έλεγχοι κανονικότητας για τις μεταβλητές: πόντοι (PPG), ασίστ (APG), κλεψίματα (SPG), κοψίματα (BPG), ριμπάουντ (RPG), λάθη της ομάδας (TOV) και του δείκτη DEFF για την ανάλυση των ομάδων με βάση την πρόκρισή τους ή μη στα playoffs.

Για τις ομάδες που προκρίθηκαν τα τελευταία χρόνια στα playoffs, όπως φαίνεται στον Πίνακα 4.3, οι μεταβλητές κλεψιμάτων (SPG), ριμπάουντ (RPG) και λαθών (TOV) ανά αγώνα σύμφωνα με τον έλεγχο Shapiro-Wilk πως η υπόθεση της κανονικότητας δεν απορρίπτεται, εφόσον με βάση το p_{value} δεν μπορεί να απορριφθεί η μηδενική υπόθεση.

Μεταβλητή	p_{value} Shapiro-Wilk
PPG	**0.00012**
APG	**0.00742**
SPG	**0.20369**
BPG	**0.00015**
RPG	**0.33434**
TOV	**0.53241**
DEFF	**0.01286**

Πίνακας 4.3: Αποτελέσματα ελέγχου κανονικότητας των ομάδων που προκρίθηκαν στα playoffs

Για την άλλη κατηγορία της μεταβλητής των playoffs, ομάδες που δεν προκρίθηκαν σε αυτά, όπως φαίνεται στον Πίνακα 4.4 για τις μόνες μεταβλητές που έχουμε ενδείξεις ότι ακολουθούν κανονική κατανομή, σύμφωνα με τον έλεγχο και το p_{value}, είναι τα κλεψίματα (SPG), ριμπάουντ (RPG) και λάθη (TOV) ανά αγώνα.

Μεταβλητή	p_{value} Shapiro-Wilk
PPG	**0.00077**
APG	**0.01314**
SPG	**0.19335**
BPG	**0.02891**
RPG	**0.20620**
TOV	**0.09034**
DEFF	**0.00487**

Πίνακας 4.4: Αποτελέσματα ελέγχου κανονικότητας των ομάδων που δεν προκρίθηκαν στα playoffs

4.1.3. Έλεγχοι κανονικότητας για την ανάλυση των Περιφερειών

Η επόμενη ανάλυση, που έγινε και στο προηγούμενο κεφάλαιο, είναι αυτή που χώριζε τις ομάδες στις περιφέρειές τους. Προφανώς θα γίνουν οι έλεγχοι κανονικότητας και για τα δεδομένα που έχουν συλλεχθεί τόσο για την κανονική περίοδο όσο και για τα playoffs.

4.1.3.1. Κανονική περίοδος

Για τα στατιστικά της κανονικής περιόδου για την West περιφέρεια, όπως φαίνεται και στον Πίνακα 4.5, σύμφωνα με τα p_{value} του ελέγχου Shapiro-Wilk οι μόνες μεταβλητές που ακολουθούν την κανονική κατανομή είναι και πάλι τα κλεψίματα (SPG), ριμπάουντ (RPG) και λάθη (TOV).

Μεταβλητή	p_{value} Shapiro-Wilk
PPG	**0.00026**
APG	**0.00108**
SPG	**0.70659**
BPG	**1.13647e-06**
RPG	**0.78493**
TOV	**0.67888**
DEFF	**0.02054**

Πίνακας 4.5:Αποτελέσματα ελέγχου κανονικότητας της West περιφέρειας για την Κανονική περίοδο

Αντίστοιχα, για την East περιφέρεια στον Πίνακα 4.6 διακρίνεται πως υπάρχουν ενδείξεις μη απόρριψης της μηδενικής υπόθεσης του ελέγχου της κανονικότητας για τις μεταβλητές κοψιμάτων (BPG), ριμπάουντ (RPG), λαθών (TOV) ανά αγώνα και του δείκτη DEFF.

Μεταβλητή	p_{value} Shapiro-Wilk
PPG	**0.00066**
APG	**0.01032**
SPG	**0.04100**
BPG	**0.17260**
RPG	**0.13235**
TOV	**0.09911**
DEFF	**0.15193**

Πίνακας 4.6:Αποτελέσματα ελέγχου κανονικότητας της East περιφέρειας για την Κανονική περίοδο

4.1.3.2. Playoffs

Για τα στατιστικά των playoffs για την περιφέρεια της West περιφέρειας, σύμφωνα με τον Πίνακα 4.7 και τον έλεγχο Shapiro-Wilk, οι μόνες μεταβλητές που ακολουθούν

την κανονική κατανομή είναι οι πόντοι (SPG), τα κλεψίματα (SPG), τα κοψίματα (BPG) και τα ριμπάουντ (RPG).

Μεταβλητή	p_{value} Shapiro-Wilk
PPG	**0.24850**
APG	**0.04479**
SPG	**0.47703**
BPG	**0.15651**
RPG	**0.71072**
TOV	**0.00857**
DEFF	**0.00619**

Πίνακας 4.7:Αποτελέσματα ελέγχου κανονικότητας της West περιφέρειας για τα Playoffs

Για την αντίπαλη περιφέρεια και βάσει του Πίνακα 4.8 έχουμε την απόρριψη της μηδενικής υπόθεσης μόνο για την μεταβλητή του δείκτη DEFF. Επομένως για τις υπόλοιπες δεν υπάρχουν αρκετές ενδείξεις απόρριψης της μηδενικής υπόθεσης.

Μεταβλητή	p_{value} Shapiro-Wilk
PPG	**0.23119**
APG	**0.87447**
SPG	**0.17981**
BPG	**0.09011**
RPG	**0.69751**
TOV	**0.056076**
DEFF	**0.01772**

Πίνακας 4.8:Αποτελέσματα ελέγχου κανονικότητας της East περιφέρειας για τα Playoffs

4.2. Συντελεστές συσχέτισης των μεταβλητών

Ένα ακόμα βασικό εργαλείο της στατιστικής είναι ο συντελεστής συσχέτισης ανάμεσα στις μεταβλητές. Υπολογίζοντας τις συσχετίσεις των μεταβλητών είμαστε σε θέση να κρίνουμε τον βαθμό αλληλεξάρτησης ανάμεσά τους και να πάρουμε μια πρώτη εικόνα για τα δεδομένα μας. Στην παρούσα ανάλυση θα αναφερθούμε στους δύο σημαντικότερους συντελεστές συσχέτισης, του Pearson και του Spearman.

Συντελεστής συσχέτισης του Pearson

Ο πιο γνωστός συντελεστής συσχέτισης μεταξύ δύο τυχαίων μεταβλητών είναι ο συντελεστής του Pearson, γνωστός και ως συντελεστής θεωρητικής συσχέτισης (correlation coefficient). Χρησιμοποιείται στην περίπτωση που έχουμε δύο τυχαίες μεταβλητές που αλληλοεπηρεάζονται.

Ο συγκεκριμένος συντελεστής συμβολίζεται με το γράμμα ρ(X,Y) ή r και υπολογίζεται από τον τύπο:

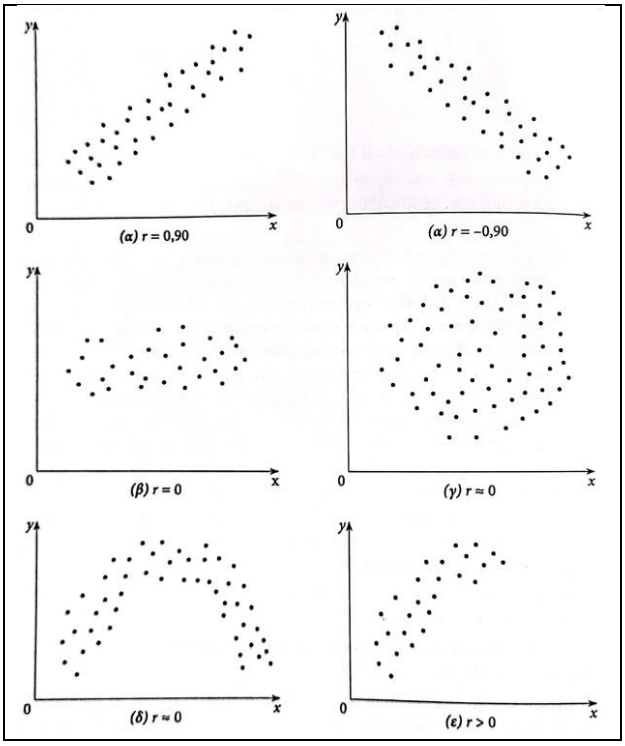
$$\rho(X,Y) = \frac{cov(X,Y)}{\sqrt{VarX * VarY}} \text{ με } cov(X,Y) = E[(X-EX)(Y-EY)] = EXY - EXEY$$

Ακόμα, αποδεικνύεται πως ο συντελεστής μπορεί να πάρει τιμές $|\rho| \leq 1$ και ανάλογα με την τιμή του χαρακτηρίζουμε τη δυναμικότητα της συσχέτισης. Δηλαδή, αν οι τιμές του δείκτη είναι ανάμεσα $0 < \rho \leq 1$, τότε οι μεταβλητές είναι θετικά συσχετισμένες, ενώ αν $-1 \leq \rho < 0$, τότε οι μεταβλητές έχουν αρνητική συσχέτιση. Προφανώς όσο το αποτέλεσμα του συντελεστή πλησιάζει την τιμή 0 τόσο φθίνει η δυναμικότητα της συσχέτισης ανάμεσα στις μεταβλητές. Τέλος, αν οι μεταβλητές έχουν $\rho = 0$, τότε θεωρούνται ασυσχέτιστες ή μη γραμμικά συσχετισμένες. (Κολύβα-Μαχαίρα και Μπόρα-Σέντα, 2013)

Ακόμα, ο δειγματικός συντελεστής συσχέτισης, ο οποίος συμβολίζεται με το γράμμα r, υπολογίζεται με βάση τις εκτιμήτριες των παραπάνω ποσοτήτων, με συνέπεια να προκύπτει ο τύπος:

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Γενικότερα, τόσο για τον δειγματικό συντελεστή όσο και για τον Pearson ισχύει πως οι τιμές που παίρνουν είναι ανάμεσα στο -1 και +1. Όσο η τιμή πλησιάζει στο -1, τόσο πιο ισχυρή αρνητική συσχέτιση υπάρχει ανάμεσα στις μεταβλητές, κάτι που αποδεικνύει πως μία αύξηση της Y σημαίνει μείωση της X και το αντίστροφο. Από την άλλη, όσο η τιμή πλησιάζει το +1, μία αύξηση της Y σημαίνει αύξηση της X και το ανάποδο. Αν ο συντελεστής δώσει μηδενική συσχέτιση, τότε ο τρόπος με τον οποίο η Y λαμβάνει τιμές δεν σχετίζεται με τον τρόπο που η X λαμβάνει τιμές. Στο επόμενο Σχήμα, 4.2, παρουσιάζονται οι διαγραμματικές απεικονίσεις για διάφορες τιμές του συντελεστή συσχέτισης, όπως περιγράψαμε σε αυτή την παράγραφο.



Σχήμα 4.2: Διαγράμματα διασποράς για διάφορες τιμές συσχετίσεων

(Πηγή: Κολύβα-Μαχαίρα και Μπόρα-Σέντα, 2013)

Συντελεστής συσχέτισης του Spearman

Ο δεύτερος συντελεστής που παρουσιάζεται είναι ο Spearman, ο οποίος πήρε το όνομα του από τον Charles Spearman και συμβολίζεται με τα γράμματα ρ ή r, όπως και ο προηγούμενος. Χρησιμοποιείται σε ένα μη-παραμετρικό πλαίσιο που μας δίνει τη συσχέτιση ανάμεσα στην βαθμολογία/τάξη (rank) που παίρνουν οι τιμές δύο μεταβλητών. (Σπυριδάκης, 2022)

Αυτός ο συντελεστής είναι μια εξειδίκευση του συντελεστή Pearson στην κατάταξη των τιμών των παρατηρήσεων των δύο μεταβλητών και υπολογίζεται από τον τύπο:

$$r_s = \rho_{rg_x, rg_y} = \frac{cov(rg_x, rg_y)}{\sigma_{rg_x}, \sigma_{rg_y}},$$

Αξιολογεί το πόσο καλά μπορεί να περιγραφεί η σχέση των δύο μεταβλητών με την χρήση μιας μονότονης συνάρτησης. Το πρόσημο του συντελεστή Spearman υποδεικνύει την κατεύθυνση της σχέσης μεταξύ των δύο μεταβλητών. Εάν η μεταβλητή Y τείνει να αυξάνεται όταν η X αυξάνει, ο συντελεστής συσχέτισης είναι θετικός. Από την άλλη, αν η Y τείνει να μειώνεται όταν η X αυξάνει, τότε ο συντελεστής είναι αρνητικός. Σε περίπτωση μηδενικής συσχέτισης ο συντελεστής Spearman δείχνει πως δεν υπάρχει τάση για την Y, όταν η X αυξάνει. Γενικότερα αυτός ο συντελεστής παίρνει τιμές μεταξύ -1 και του +1. (Τριανταφύλλου, 2023)

Αντίστοιχα, στην περίπτωση που υπάρχει τουλάχιστον μία περίπτωση ισοβαθμίας στην κατάταξη των παρατηρήσεων μίας εκ των δύο μεταβλητών, ο δειγματικός συντελεστής συσχέτισης υπολογίζεται με τον ίδιο ακριβώς τρόπο που υπολογίζεται και ο δειγματικός συντελεστής συσχέτισης του Pearson:

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Όπως είναι αναμενόμενο, και γι' αυτόν τον συντελεστή ισχύει πως οι τιμές που λαμβάνει ανάμεσα στο -1 και το +1. Για τιμές κοντά στο +1 συνεπάγεται ότι οι μεταβλητές σχετίζονται άριστα μέσω μίας μονότονης συνάρτησης όπου η αύξηση της τιμής του Y σημαίνει αύξηση της τιμής του X, ενώ για τιμές κοντά στο -1 ισχύει ότι οι μεταβλητές σχετίζονται άριστα μέσω μιας γραμμικής συνάρτησης, αλλά η αύξηση της τιμής του Y σημαίνει μείωση της τιμής του X.

4.2.1. Υπολογισμός συσχετίσεων για όλες τις μεταβλητές

Στην παρούσα παράγραφο θα υπολογιστεί ο συντελεστής συσχέτισης του Pearson για όλες τις μεταβλητές των δεδομένων. Αρχικά θα γίνει για την κανονική περίοδο και έπειτα για τα playoffs. Όπως αναφέρθηκε περιληπτικά και παραπάνω, τα αποτελέσματα των τιμών του συντελεστή συσχέτισης χαρακτηρίζονται ως προς την δυναμικότητα του με τον παρακάτω τρόπο.

- Τέλεια Συσχέτιση: εάν η τιμή του συντελεστή είναι πολύ κοντά στο ± 1.

- Πολύ Υψηλή Συσχέτιση: εάν η τιμή του συντελεστή κυμαίνεται μεταξύ ± 0,80 και ± 1 .
- Υψηλή Συσχέτιση: εάν η τιμή κυμαίνεται μεταξύ ± 0,60 και ± 0,79 .
- Μέτρια Συσχέτιση: εάν η τιμή κυμαίνεται μεταξύ ± 0,40 και ± 0,59 .
- Ασθενής Συσχέτιση: όταν η τιμή κυμαίνεται μεταξύ ± 0,20 και ± 0,39
- Πολύ Ασθενής Συσχέτιση: εάν η τιμή του συντελεστή είναι κάτω από ± 0,19 .
- Χωρίς Συσχέτιση: όταν η τιμή είναι μηδέν. (Evans, 1996)

4.2.1.1. Συσχετίσεις για την κανονική περίοδο

Αρχικά, στο δεύτερο μέρος του Παραρτήματος της παρούσας εργασίας παρουσιάζονται οι τιμές του δείκτη του Pearson και ανάλογα με τη δυναμικότητα της συσχέτισης που υπάρχει ανάμεσα στις μεταβλητές το τετράγωνο παίρνει το κατάλληλο χρώμα. Για συσχετίσεις κοντά στο -1 το τετράγωνο έχει οριακά λευκό χρώμα και όσο αυξάνεται η τιμή του αποτελέσματος τόσο πιο σκούρο κόκκινο γίνεται.

Λόγω του όγκου των μεταβλητών που διαθέτουν τα δεδομένα μας δεν θα σχολιαστούν όλες οι συσχετίσεις μεταξύ τους, παρόλα αυτά θα γίνει ανάλυση για τις συσχετίσεις όλων των μεταβλητών με τις κατοχές των ομάδων (POSSt). Αρχικά, στον παραπάνω πίνακα παρατηρούμε πως οι μεταβλητές που δημιουργήθηκαν για την παρούσα εργασία έχουν αρκετά υψηλές τιμές συσχέτισης με την μεταβλητή των πόντων (PPG). Παραδείγματος χάρη οι μεταβλητές των κατοχών των ομάδων ανά αγώνα (POSSt, POSSo) έχουν συντελεστή συσχέτισης με τους πόντους 0.84 και 0.85 αντίστοιχα. Τις συγκεκριμένες τιμές τις αναμέναμε αφού στους τύπους των μεταβλητών, οι οποίοι παρουσιάστηκαν στο κεφάλαιο 2, η μεταβλητή των πόντων συμμετέχει στην υπολογισμό τους. Στη συνέχεια θα δούμε αναλυτικά τις συσχετίσεις που έχουν δημιουργηθεί ανάμεσα σε όλες τις μεταβλητές με την μεταβλητή κατοχών ανά παιχνίδι (POSSt). Όπως είναι φανερό και στο Σχήμα 4.3, πολύ ισχυρή θετική συσχέτιση έχει με τις μεταβλητές: πόντοι (PPG), πόντοι των αντιπάλων (PPGA), επιτυχημένα δίποντα και τρίποντα (FGM), συνολικές προσπάθειες δίποντων και τριπόντων (FGA), συνολικές προσπάθειες δίποντων και τριπόντων των αντιπάλων (FGAo) και τις κατοχές των αντίπαλων ομάδων ανά αγώνα (POSSo). Με την μεταβλητή των κατοχών των αντιπάλων ανά αγώνα παρατηρούμε πως η τιμή του συντελεστή συσχέτισης είναι της τάξης 0.99 (≈1), κάτι που προειδοποιεί πως οι δύο μεταβλητές δεν θα μπορούσαν να χρησιμοποιηθούν ταυτόχρονα σε κάποιο (γενικευμένο) γραμμικό μοντέλο. Το ίδιο συμβαίνει και με την μεταβλητή FGAo η οποία έχει συντελεστή συσχέτισης με τις κατοχές της ομάδας ανά αγώνα ίσο με 0.9. Υψηλή θετική συσχέτιση υπάρχει με τις μεταβλητές: άστοχα δίποντα και τρίποντα (MisFG), επιτυχημένα τρίποντα (3PM), συνολικές προσπάθειες τριπόντων (3PA), αμυντικά ριμπάουντ (DRB), ασίστ (APG) και ο δείκτης efficiency (EFF). Μέτρια θετική συσχέτιση υπάρχει με τις μεταβλητές συνολικά ριμπάουντ (RPG) και τους δείκτες offensive rating (ORt) και defensive rating (DRt). Ασθενής θετική συσχέτιση υπάρχει με τα ποσοστά ευστοχίας των δίποντων και τρίποντων (FG%). Με τις

υπόλοιπες μεταβλητές υπάρχει πολύ ασθενής συσχέτιση και συγκεκριμένα με τις άστοχες ελεύθερες βολές (MisFT), επιθετικά ριμπάουντ της ομάδας (ORB) και της αντιπάλου (ORBo) και τον δείκτη DEFF έχει αρνητική ασθενή συσχέτιση. Με παρόμοια λογική μπορεί να γίνει και η ανάλυση για τις υπόλοιπες μεταβλητές που είναι διαθέσιμες στα δεδομένα που συλλέξαμε.

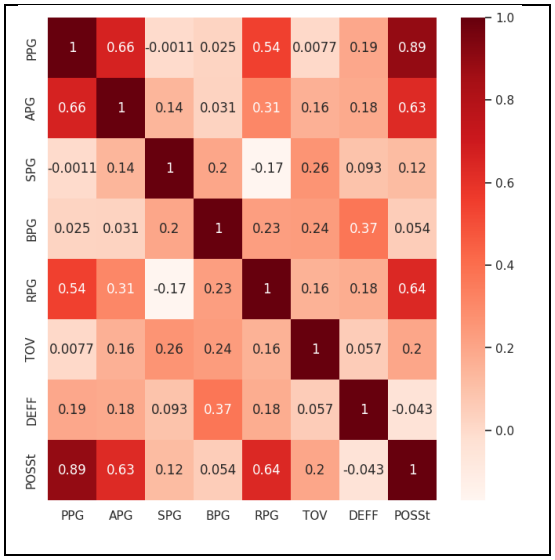
4.2.1.1. Συσχετίσεις για τα playoffs

Όπως και στην κανονική περίοδο, στο Παράρτημα της εργασίας παρουσιάζονται οι τιμές του συντελεστή Pearson για την φάση των playoffs και το ανάλογο heatmap. Σε αυτή την παράγραφο θα αναλυθεί η τιμή του συντελεστή συσχέτισης όλων των μεταβλητών με την μεταβλητή κατοχών των ομάδων, ώστε να γίνει κατανοητό αν υπάρχουν διαφοροποιήσεις με την κανονική φάση της διοργάνωσης. Αρχικά, παρατηρούμε πως πολύ υψηλή συσχέτιση με τη μεταβλητή αυτή έχουν οι μεταβλητές πόντοι που σκοράρει η αντίπαλη ομάδα (PPGA) και συνολικές προσπάθειες δίποντων και τρίποντων της ομάδας (FGA) και της αντιπάλου (FGAo) και οι κατοχές που έχουν οι αντίπαλοι ανά αγώνα (POSSo). Υψηλή συσχέτιση έχουν οι πόντοι που πετυχαίνει η ομάδα (PPG), εύστοχα δίποντα και τρίποντα (FGM), άστοχα δίποντα και τρίποντα (FGA), επιτυχημένα τρίποντα (3PM), συνολικές προσπάθειες τριπόντων (3PA) και ο δείκτης efficiency (EFF). Μέτρια συσχέτιση έχουν οι μεταβλητές: αμυντικά ριμπάουντ (DRB), συνολικά ριμπάουντ (RPG) και οι ασίστ ανά αγώνα (APG). Ασθενή συσχέτιση έχουν οι μεταβλητές: κλεψίματα, λάθη της ομάδας και της αντιπάλου και οι δείκτες ORt και DRt. Οι υπόλοιπες μεταβλητές έχουν πολύ ασθενή συσχέτιση με τη μεταβλητή των κατοχών της ομάδας ανά αγώνα. Η μεγαλύτερη διαφορά που παρατηρούμε σε σχέση με την κανονική περίοδο είναι πως υπάρχει μόνο μια μεταβλητή που έχει αρνητική συσχέτιση με τις κατοχές και αυτή είναι ο δείκτης DEFF, ενώ προηγουμένως ήταν περισσότερες.

4.2.2. Υπολογισμός συσχετίσεων με βάση την πρόκριση στα playoffs

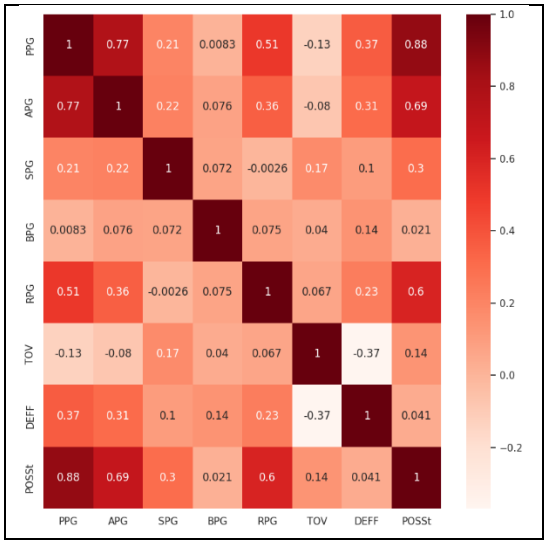
Στη συνέχεια θα υπολογιστεί ο θεωρητικός συντελεστής συσχέτισης για τις μεταβλητές: πόντοι (PPG), ασίστ (APG), κλεψίματα (SPG), κοψίματα (BPG), ριμπάουντ (RPG), λάθη της ομάδας (TOV), DEFF, κατοχές της ομάδας ανά αγώνα (POSSt) για την ανάλυση των ομάδων με βάση την πρόκρισή τους ή μη στα playoffs.

Για τις ομάδες που προκρίθηκαν τα τελευταία χρόνια στα playoffs, όπως φαίνεται στο Σχήμα 4.3, δίνονται όλες οι συσχετίσεις ανάμεσα στις μεταβλητές που χρησιμοποιήθηκαν προς την συγκεκριμένη ανάλυση. Πιο αναλυτικά, η μεταβλητή των κατοχών ανά αγώνα της ομάδας (POSSt) έχει πολύ ισχυρή θετική συσχέτιση με την μεταβλητή των πόντων της ομάδας ανά αγώνα (PPG). Υψηλή συσχέτιση υπάρχει με τις μεταβλητές των ασίστ (APG) και των συνολικών ριμπάουντ ανά αγώνα (RPG). Με όλες τις υπόλοιπες έχει ασθενή συσχέτιση και με τον δείκτη DEFF ασθενή αρνητική.



Σχήμα 4.3: Heatmap του συντελεστή Pearson για τις ομάδες που προκρίθηκαν στα playoffs

Αντίστοιχα, στο Σχήμα 4.4, που αναφέρεται στους συντελεστές συσχέτισης για τις ομάδες που δεν προκρίθηκαν στα playoffs παρατηρούμε μια παρόμοια κατάσταση για τη μεταβλητή των κατοχών ανά αγώνα με τη διαφορά πως πλέον δεν υπάρχει μεταβλητή η οποία έχει αρνητική συσχέτιση μαζί της.

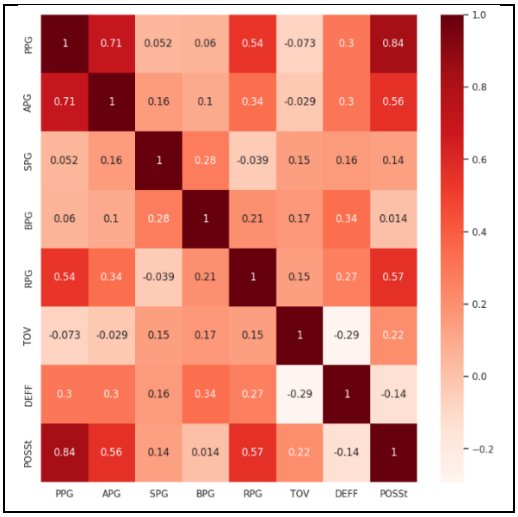


Σχήμα 4.4: Heatmap του συντελεστή Pearson για τις ομάδες που δεν προκρίθηκαν στα playoffs

4.2.3. Υπολογισμός συσχετίσεων για την ανάλυση των Περιφερειών

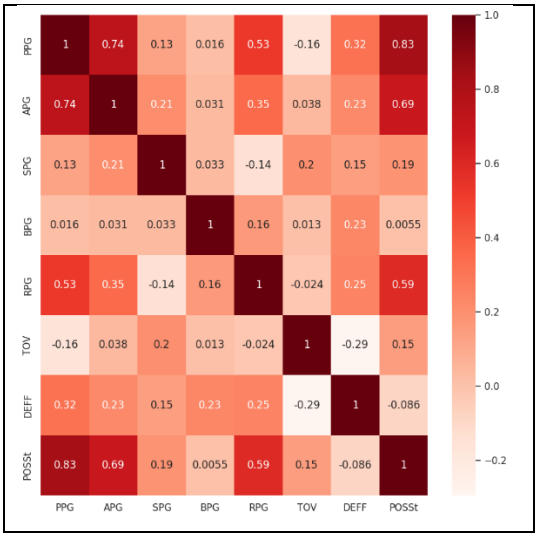
Τελευταία ανάλυση για την οποία θα υπολογιστεί ο συντελεστής συσχέτισης ανάμεσα στις μεταβλητές της είναι ο διαχωρισμός των ομάδων σε περιφέρειες.

Για την περιφέρεια West, όπως φαίνεται στο Σχήμα 4.5, οι κατοχές της ομάδας ανά αγώνα (POSSt) έχουν πολύ ισχυρή θετική συσχέτιση μόνο με την μεταβλητή των πόντων ανά αγώνα (PPG). Μέτρια θετική συσχέτιση έχει με τις μεταβλητές των ασίστ (APG) και των συνολικών ριμπάουντ ανά αγώνα (RPG), ενώ με τις υπόλοιπες έχει ασθενή συσχέτιση.



Σχήμα 4.5: Heatmap του συντελεστή Pearson για την περιφέρεια West

Αντίστοιχα, για την περιφέρεια East δίνονται τα αποτελέσματα του συντελεστή συσχέτισης Pearson στο Σχήμα 4.6.

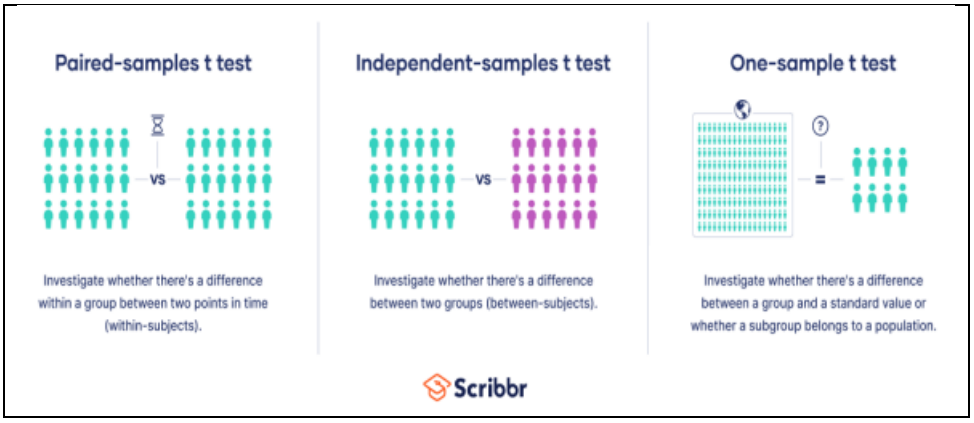


Σχήμα 4.6: Heatmap του συντελεστή Pearson για την περιφέρεια East

4.3. Έλεγχοι υποθέσεων για την ισότητα μέσων τιμών δύο δειγμάτων

Όπως είναι γνωστό, στους στατιστικούς ελέγχους υποθέσεων επιθυμούμε να ελέγξουμε αν μία ή περισσότερες παράμετροι ενός πληθυσμού ή πληθυσμών ικανοποιούν μια βασική υπόθεση έναντι μίας δεύτερης (εναλλακτική). Αυτοί οι έλεγχοι είναι από τις σημαντικότερες στατιστικές μεθόδους για τη σύγκριση των μέσων δύο ομάδων. Οι έλεγχοι χωρίζονται σε δύο κατηγορίες, τους παραμετρικούς και τους μη-παραμετρικούς. Οι παραμετρικοί έλεγχοι προϋποθέτουν τα δεδομένα μας να είναι ανεξάρτητα, να ακολουθούν την κανονική κατανομή και να έχουν σταθερή διακύμανση. Σε περίπτωση που κάποια από τις συνθήκες αυτές δεν ισχύει, τότε αναγκαστικά θα πρέπει να χρησιμοποιηθούν οι μη-παραμετρικοί έλεγχοι.

Τα είδη των t-test που υπάρχουν είναι διάφορα, ανάλογα με το εάν οι ομάδες που χρησιμοποιούνται προέρχονται από έναν μόνο πληθυσμό ή δύο διαφορετικούς. Για να γίνει πιο κατανοητό, δίνεται το Σχήμα 4.7 που περιέχει τα τεστ και τις ονομασίες τους.



Σχήμα 4.7: Διαφορετικές κατηγορίες t-test

(Πηγή: https://www.scribbr.com/)

Η πρώτη κατηγορία ελέγχων (αριστερό μέρος του σχήματος) αναφέρεται ως paired-sample t-test και χρησιμοποιείται όταν οι ομάδες προέρχονται από έναν πληθυσμό και μελετάται η επίδραση κάποιου παράγονται σε αυτόν, τα δεδομένα προέρχονται από μία μέτρηση πριν και μία αφότου χρησιμοποιήθηκε ο παράγοντας τον οποίο αναλύουμε. Δεύτερη κατηγορία είναι τα Independent-samples t-test χρησιμοποιούνται όταν τα δείγματα προέρχονται από δύο διαφορετικούς πληθυσμούς και χρειάζεται να ελεγχθεί αν οι μέσες τιμές του χαρακτηριστικού είναι ίσες. Ακόμα, εάν υπάρχει μία ομάδα και χρειάζεται να γίνει έλεγχος για την υπόθεση ότι η μέση τιμή ενός χαρακτηριστικού του πληθυσμού είναι ίση με κάποια δεδομένη τιμή, χρησιμοποιείται ο έλεγχος One-sample t-test. Τέλος, αν μας ενδιαφέρει να εξετάσουμε αν η μέση τιμή ενός πληθυσμού διαφέρει από κάποιου άλλου, τότε γίνεται χρήση ενός αμφίπλευρου t-test.

Στην παρούσα εργασία, παρόλο που τα δεδομένα μας στις περισσότερες μεταβλητές δεν ακολουθούν την κανονική κατανομή, λόγω του Κεντρικού Οριακού Θεωρήματος για μεγάλο πλήθος δεδομένων (n>50) δίνεται η δυνατότητα να πραγματοποιηθούν οι έλεγχοι independent-samples t test και paired-samples t-test για παραμετρικούς ελέγχους. Σε περίπτωση που διαθέταμε μικρότερο δείγμα δεδομένων θα έπρεπε να χρησιμοποιηθούν οι αντίστοιχοι μη παραμετρικοί έλεγχοι.

4.3.1. Έλεγχος για την ισότητα των μέσων τιμών δύο δειγμάτων

Σε περίπτωση που οι ομάδες έχουν ίση διακύμανση μεταξύ τους, τότε ο τύπος που στηρίζεται ο έλεγχος μετατρέπεται σε:

$$t = \frac{\overline{x_1} - \overline{x_2}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

όπου το s είναι το ομαδοποιημένο τυπικό σφάλμα (pooled standard error) των δύο ομάδων. (Kim, 2015) Η μηδενική και η ενναλακτική υπόθεση του ελέγχου είναι:

Μηδενική υπόθεση H_0 : $\mu_1 = \mu_2$

vs

Εναλλακτική υπόθεση H_1 : $\mu_1 \neq \mu_2$

Σε περίπτωση μη απόρριψης της μηδενικής υπόθεσης του ελέγχου, οι τιμές των δύο μεταβλητών δεν διαφέρουν σημαντικά. (Ευαγγελάρας, 2022) Ακόμα, εφόσον δεν έχει γίνει κάποια υπόθεση για την ισότητα των διακυμάνσεων των δειγμάτων θα χρησιμοποιηθεί το "Welch Two Sample t-test" για τις αναλύσεις που ακολουθούν.

4.3.1.1. Έλεγχος μέσων βάσει της πρόκρισης στα Playoffs

Εφόσον εκτελέστηκαν οι έλεγχοι για τις μεταβλητές: πόντοι (PPG), ασίστ (APG), κλεψίματα (SPG), κοψίματα (BPG), ριμπάουντ (RPG), λάθη της ομάδας (TOV), βρέθηκαν τα p_{value} των ελέγχων, τα οποία είναι όλα μικρότερα από το επίπεδο σημαντικότητας 5% και επομένως έχουμε σημαντικές ενδείξεις απόρριψης της μηδενικής υπόθεσης. Άρα οι μέσες τιμές για τις μεταβλητές των ομάδων που προκρίθηκαν στα playoffs διαφέρουν σημαντικά από τις μέσες τιμές των ομάδων που δεν τα κατάφεραν. Οι ομάδες που προκρίθηκαν σημειώνουν υψηλότερες τιμές σε όλες τις μεταβλητές εκτός των λαθών (TOV), όπου μεγαλύτερες τιμές λαμβάνουν οι ομάδες που δεν προκρίθηκαν.

Μεταβλητές ελέγχου	t-statistic	p_{value}
PPG	**4.12468**	**4.42592e-05**
APG	**3.40010**	**0.00073**
SPG	**2.81026**	**0.00516**
BPG	**3.54885**	**0.00042**
RPG	**3.89639**	**0.00011**
TOV	**-5.02337**	**7.34669e-07**

Πίνακας 4.9: Αποτελέσματα ελέγχου ισότητας μέσων με βάση την πρόκριση στα Playoffs

4.3.1.1. Έλεγχος μέσων βάσει των Περιφερειών

Όπως και προηγουμένως, αφού εκτελέστηκαν οι έλεγχοι για τις μεταβλητές: πόντοι (PPG), ασίστ (APG), κλεψίματα (SPG), κοψίματα (BPG), ριμπάουντ (RPG), λάθη της ομάδας (TOV), βρέθηκαν τα p_{value} των ελέγχων. Για τις μεταβλητές: ασίστ (APG), κοψίματα (BPG) και λάθη (TOV) ανά αγώνα, τα p_{value} των ελέγχων βρέθηκαν μεγαλύτερα του επιπέδου σημαντικότητας 5% και συνεπώς έχουμε αρκετές ενδείξεις για τη μη απόρριψη της μηδενικής υπόθεσης. Για τις υπόλοιπες μεταβλητές απορρίπτεται η μηδενική υπόθεση του ελέγχου και επομένως δεν ικανοποιείται η ισότητα των μέσων των τιμών για τα στατιστικά των ομάδων των δύο περιφερειών. Οι τιμές των ομάδων της περιφέρειας West είναι υψηλότερες απ' αυτές της East.

Μεταβλητές ελέγχου	t-statistic	p_{value}
PPG	**3.97619**	**8.16248e-05**
APG	**1.90958**	**0.05682**
SPG	**2.46520**	**0.01406**
BPG	**1.36869**	**0.17178**
RPG	**2.32555**	**0.02048**
TOV	**1.63756**	**0.10221**

Πίνακας 4.10: Αποτελέσματα ελέγχου ισότητας μέσων βάσει των Περιφερειών

4.3.2. Έλεγχος ισότητας μέσου για ζευγαρωτές παρατηρήσεις (Paired- T test)

Στον επόμενο έλεγχο προσήμου δεν ελέγχουμε την ισότητα των διαμέσων δύο πληθυσμών, αλλά αν η διάμεσος των διαφορών είναι ίση με μηδέν (0). Η συγκεκριμένη μέθοδος υπολογίζει το πρόσημο των διαφορών. Π.χ. αν $x_i - y_i > 0$ τότε το πρόσημο είναι + και η διαδικασία υλοποιείται ξεχωριστά για κάθε παρατήρηση. Τέλος, υπολογίζεται το p_{value} του ελέγχου σύμφωνα με τον τύπο:

$$p = 2 \sum_{i=1}^{Q} \binom{n^*}{i} 2^{-n^*} \text{ , όπου } Q = \min\{d_+, n^* - d_+\}$$

Μηδενική υπόθεση H_0 : $\mu_1 - \mu_2 = 0$

vs

Εναλλακτική υπόθεση H_1 : $\mu_1 - \mu_2 \neq 0$

Συνεπώς σε περίπτωση μη απόρριψης της μηδενικής υπόθεσης του ελέγχου οι δύο μεταβλητές μας δεν έχουν κάποια στατιστικά σημαντική διαφορά. (Ευαγγελάρας, 2022)

Κλείνοντας, αναφέρουμε ότι σ' αυτή την ενότητα πραγματοποιήθηκαν t-test για τις μεταβλητές: πόντοι (PPG), ασίστ (APG), κλεψίματα (SPG), κοψίματα (BPG), ριμπάουντ (RPG), λάθη της ομάδας (TOV) για τα δεδομένα των ομάδων -για την κανονική περίοδο και τα playoffs- που προκρίθηκαν στην φάση των playoffs. Πιο συγκεκριμένα, ο έλεγχος είχε σχέση με τα στατιστικά των ομάδων για τις δύο φάσεις που χωρίζεται η οργάνωση του NBA. Αρχικά, εκτελέστηκε ο έλεγχος για τους πόντους (PPG) που σκοράρουν οι ομάδες ανά αγώνα στην κανονική περίοδο και τα playoffs. Όπως φαίνεται στον Πίνακα 4.8, το p_{value} του ελέγχου για τους πόντους είναι μικρότερο του επιπέδου σημαντικότητας 5% και επομένως υπάρχει στατιστικά σημαντική διαφορά στους πόντους που σκοράρουν οι ομάδες. Οι ομάδες που συμμετείχαν και στις δύο φάσεις του πρωταθλήματος φαίνεται πως σκοράρουν λιγότερους πόντους στην φάση των playoffs. Το ίδιο συμπέρασμα προκύπτει και για τις υπόλοιπες μεταβλητές, εφόσον τα p_{value} τους συνεχίζουν να είναι μικρότερα από τα υπόλοιπα.

Μεταβλητές ελέγχου	Statistic	p_{value}
PPG	**-15.96490**	**1.53194e-39**
APG	**-20.85359**	**1.02313e-55**
SPG	**-10.09386**	**3.48440e-20**
BPG	**-4.00111**	**8.41210e-05**
RPG	**-8.39068**	**4.25197e-15**
TOV	**-4.69604**	**4.47396e-06**

Πίνακας 4.11: Αποτελέσματα ελέγχου ισότητας μέσου για ζευγαρωτές παρατηρήσεις

ΚΕΦΑΛΑΙΟ 5ο

5. Γενικευμένα γραμμικά μοντέλα

Στόχος του παρόντος κεφαλαίου είναι να εξεταστεί η σημαντικότητα στον ρόλο των μεταβλητών όσον αφορά την πρόκριση των ομάδων στην φάση των playoffs και την πρωταθλήτρια ομάδα. Για να καταφέρουμε να φέρουμε εις πέρας το συγκεκριμένο ερώτημα, θα πρέπει να προσαρμόσουμε κατάλληλα γενικευμένα γραμμικά μοντέλα που θα έχουν για μεταβλητές απόκρισης την κατηγορική μεταβλητή Playoffs και την Champions.

5.1. Ανάλυση Παλινδρόμησης

Αρκετά συχνά εμφανίζεται η ανάγκη ταυτόχρονης μελέτης δύο ή περισσότερων χαρακτηριστικών μεταβλητών με στόχο τον προσδιορισμό του τρόπου με τον οποίο οι μεταβλητές αυτές σχετίζονται μεταξύ τους. Η πλέον συνήθης μορφή σχέσης δύο μεταβλητών, Χ και Υ με απώτερο σκοπό την πρόβλεψη της μίας απ' αυτές μέσω της άλλης λέγεται ανάλυση απλής παλινδρόμησης (simple regression analysis). Η μεταβλητή Χ ονομάζεται ανεξάρτητη μεταβλητή, εφόσον οι τιμές της (x_i) καθορίζονται από τον αναλυτή, ενώ η Υ ονομάζεται εξαρτημένη μεταβλητή αφού οι τιμές της (y_i) εξαρτώνται από τις τιμές της Χ. Σε περίπτωση που η παλινδρόμηση που θέλουμε να εκτελέσουμε έχει περισσότερες από μία ανεξάρτητες μεταβλητές, δηλαδή X_1 , X_2 , … , X_p, τότε το μοντέλο που εφαρμόζουμε ονομάζεται πολλαπλή παλινδρόμηση και αποτελεί μια πιο γενική μορφή της απλής παλινδρόμησης. Ο γενικός τύπος αυτού του μοντέλου είναι ο εξής:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i$$

όπου Y_i , i=1,..ν είναι η i-οστή τιμή της εξαρτημένης μεταβλητής Υ, αντίστοιχα τα x_i είναι οι i-οστές τιμές των p ανεξάρτητων X_i μεταβλητών, ενώ β_0 , β_1, …, β_p είναι οι τιμές των παραμέτρων του μοντέλου και τα σφάλματα συμβολίζονται με ε_i.

Για να χρησιμοποιηθεί το παραπάνω μοντέλο πολλαπλής παλινδρόμησης, εκτός της κανονικότητας των σφαλμάτων και κατ' επέκταση της εξαρτημένης μεταβλητής, θα πρέπει να ισχύουν οι εξής προϋποθέσεις:

- Οι ποσότητες β_0 , β_1, …, β_p είναι άγνωστες παράμετροι.

- Τα x_{i1},…, x_{ip} είναι γνωστοί αριθμοί. Πιο συγκεκριμένα, είναι οι τιμές των ανεξάρτητων (ή προσβλεπουσών) μεταβλητών κατά την i-οστή επανάληψη του πειράματος. Οι τιμές αυτές καθορίζονται από τον ερευνητή που εκτελεί το πείραμα.

- Το Y_i είναι η τιμή της εξαρτημένης μεταβλητής (ή μεταβλητής απόκρισης) κατά την i-οστή επανάληψη του πειράματος. Το Y_i είναι τυχαία μεταβλητή.

- Τα ε_i , i=1,…,ν είναι τυχαία σφάλματα με μέση τιμή 0 και διασπορά σ^2 .

- Τα σφάλματα εi και εj που αντιστοιχούν σε διαφορετικές επαναλήψεις του πειράματος (j ≠ i) θεωρούνται ασυσχέτιστα, δηλαδή ισχύει $Cov(\varepsilon_i, \varepsilon_j) = 0$ για j ≠ i. (Κούτρας, 2020)

Όπως αναφέραμε και προηγουμένως, η ανάλυση που θα γίνει στο παρόν κεφάλαιο σχετίζεται με τη δίτιμη μεταβλητή που ερμηνεύει την προαγωγή μιας ομάδας στη φάση των playoffs. Δεδομένης της μορφής της, δίτιμη και κατηγορική μεταβλητή, δεν συνάδει να εξετάσουμε την κανονικότητα της μεταβλητής και επομένως δεν θα ήταν ορθό να εκτελεστεί ένα μοντέλο παλινδρόμησης. Για τον παραπάνω λόγο θα αναλύσουμε τα δεδομένα βάσει γενικευμένων γραμμικών μοντέλων, τα οποία αποτελούν γενίκευση του μοντέλου παλινδρόμησης.

5.2. Γενικευμένα Γραμμικά Μοντέλα

Ένας απλός τρόπος για να γίνει κατανοητό ότι τα γενικευμένα γραμμικά μοντέλα αποτελούν γενίκευση του μοντέλου της παλινδρόμησης είναι συνειδητοποιώντας πως αντί της μέσης τιμής της Y χρησιμοποιείται μια συνάρτηση αυτής της μέσης τιμής, η οποία συμβολίζεται με g. Με λίγα λόγια, μια συνάρτηση σύνδεσης g συνδέει το στοχαστικό μέρος του μοντέλου, μέση τιμή της Y, με το μη στοχαστικό μέρος του, το οποίο είναι γραμμικός συνδυασμός των ερμηνευτικών μεταβλητών X_i. Συνεπώς η συνάρτηση πρόβλεψης διατυπώνεται ως εξής:

$$\eta_i = g(\mu_i) = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij}$$

όπου το X_{ij} είναι η τιμή της μεταβλητής X_j για την παρατήρηση i.

Ο τύπος που αναφέρθηκε για τη συνάρτηση $g(\mu_i)$ αφορά κατανομές οι οποίες ανήκουν στην εκθετική οικογένεια κατανομών. Ένα από τα βασικότερα πλεονεκτήματα των γενικευμένων γραμμικών μοντέλων είναι το γεγονός πως δεν γίνεται καμία υπόθεση για την κατανομή των σφαλμάτων, σε σχέση με την παλινδρόμηση που χρειάζονται κανονικά σφάλματα. Ακόμα, γίνεται αντιληπτό πως τα συγκεκριμένα μοντέλα έχουν μεγαλύτερο φάσμα εφαρμογών και πως οι εκτιμήσεις των παραμέτρων προκύπτουν με τη μέθοδο μέγιστης πιθανοφάνειας και κατά συνέπεια έχουν μια σειρά από επιθυμητές ιδιότητες. Τέλος, στο μεγαλύτερο μέρος των περιπτώσεων δε χρειάζεται να υποθέσουμε σταθερή διακύμανση για τις τιμές της Y και δε χρειάζεται να χρησιμοποιούνται διαφορετικά μοντέλα ανάλογα με το αν οι ερμηνευτικές μεταβλητές είναι ποσοτικές ή ποιοτικές, σημειώνεται ωστόσο πως στην ανάλυση παλινδρόμησης χρησιμοποιούνται ποσοτικές μεταβλητές. (Πολίτης, 2021)

5.2.1. Μέθοδοι για δίτιμα δεδομένα

Στο παρόν κεφάλαιο θα γίνει χρήση γενικευμένων γραμμικών μοντέλων για δίτιμες μεταβλητές (binary data). Για τα δίτιμα δεδομένα, η κατανομή πιθανότητας των Y_i είναι:

$$P[\, Y_i = y_i \,] = p_i^{y_i} \, (1 - p_i)^{1-y_i} \text{ , } y_i = 0,1$$

η οποία αποτελεί την συνάρτηση πιθανότητας Bernoulli για υπαρκτές τιμές 0, 1.

Οι τρείς συναρτήσεις σύνδεσης που χρησιμοποιούνται για τέτοιου είδους δεδομένα είναι:

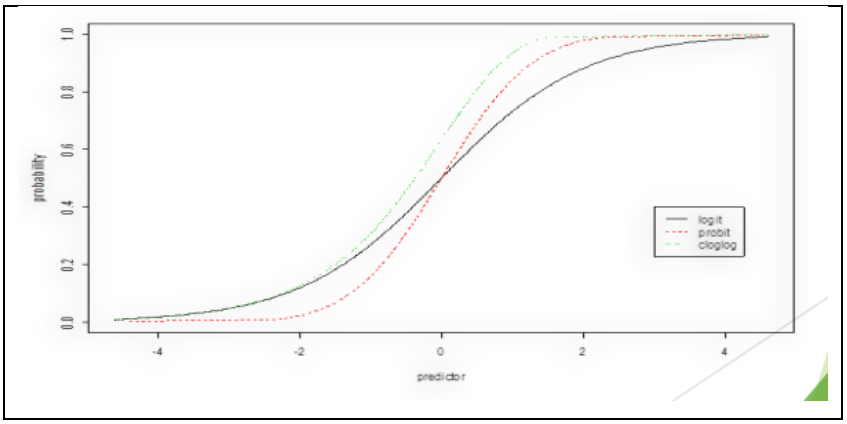
- **Logit:** $\eta_i = \text{logit}\,(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$.

- **Probit:** $\eta_i = \text{Probit}(p_i) = \Phi^{-1}(p_i)$,

όπου η Φ είναι η αθροιστική συνάρτηση της τυποποιημένης κανονικής.

- **Complementary log-log:** $\eta_i = \log[-[\log(1-p_i)]$.

Στο Σχήμα 5.1 δίνονται οι γραφικές αναπαραστάσεις των συναρτήσεων σύνδεσης. Όταν χρησιμοποιείται ως συνάρτηση σύνδεσης η συνάρτηση logit, όπως θα αναλύσουμε στην επόμενη παράγραφο, τότε το μοντέλο μας ονομάζεται μοντέλο λογιστικής παλινδρόμησης.



Σχήμα 5.1: Γραφικές αναπαραστάσεις των συναρτήσεων σύνδεσης

(Πηγή: Πολίτης, 2021)

5.3. Λογιστική Παλινδρόμηση

Η λογιστική παλινδρόμηση χρησιμοποιείται όταν επιθυμία μας είναι να προβλέψουμε την απουσία ή την παρουσία ενός χαρακτηριστικού, ή ενός συμβάντος. Αξιοποιείται όταν η εξαρτημένη μεταβλητή Υ είναι δίτιμη.

Το κανονικό γραμμικό μοντέλο είναι αδύνατο να χρησιμοποιηθεί όταν η μεταβλητή Υ είναι δυαδική και έχουμε τα εξής τρία προβλήματα:

- Τα σφάλματα δεν είναι κανονικά.
- Τα σφάλματα έχουν άνισες διασπορές
- Περιορισμός στη συνάρτηση απόκρισης (η προβλεπόμενη πιθανότητα θα πρέπει να ανήκει στα διάστημα (0,1))

Παρόλο που τα δύο πρώτα προβλήματα είναι δυνατό σε κάποιες περιπτώσεις να τα παραλείψουμε και να χρησιμοποιήσουμε τη γραμμική παλινδρόμηση εφαρμόζοντας κάποιες άλλες τεχνικές, το τρίτο πρόβλημα μας το απαγορεύει ρητά, γιατί δεν μπορεί να αντιμετωπιστεί με διαφορετικό τρόπο.

Η σημαντικότερη διαφοροποίηση μεταξύ λογιστικής και γραμμικής παλινδρόμησης βασίζεται στη φύση της επιλεγμένης μεταβλητής απόκρισης, η οποία στη μεν πρώτη μπορεί να είναι κατηγορική, (τακτική ή ονομαστική), στη δε δεύτερη αποκλειστικά ποσοτική και συνεχής. Ενώ κατά την κλασική γραμμική παλινδρόμηση η εκτίμηση των παραμέτρων α και β_i γίνεται με τη μέθοδο των ελάχιστων τετραγώνων, κατά τη

λογιστική παλινδρόμηση η εκτίμηση των παραμέτρων γίνεται με τη μέθοδο της μέγιστης πιθανοφάνειας (μέθοδος συνήθως εφαρμοζόμενη στα γενικευμένα γραμμικά υποδείγματα), δηλαδή επιλέγονται οι πιο πιθανοφανείς εκτιμήσεις των παραμέτρων, προκειμένου να οδηγήσουν στα παρατηρούμενα αποτελέσματα. Ως επακόλουθο, η πρώτη παραδέχεται την ύπαρξη ομοιογένειας (ομοσκεδαστικότητας) στα υπόλοιπα των αποκρίσεων, ενώ στη δεύτερη αναπτύσσεται πάντα ετεροσκεδαστικότητα σε κάθε προβλεπόμενη τιμή εξαιτίας του μεταβαλλόμενου ποσοστού διακύμανσης που αναλογεί σε αυτήν.

Διακρίνονται τρεις τύποι λογιστικής παλινδρόμησης ανάλογα με την ιδιαίτερη φύση της εξαρτημένης κατηγορικής μεταβλητής, η οποία μπορεί να είναι:

- Δίτιμη ή δυαδική ή διχοτομική (binary) ή διμερής εξαρτημένη μεταβλητή. Συνίσταται από δύο κατηγορίες.
- Διατάξιμη (ordinal) μεταβλητή. Η εξαρτημένη μεταβλητή συνίσταται από τρεις ή περισσότερες κατηγορίες μεταξύ των οποίων ισχύει η έννοια της διάταξης.
- Ονομαστική (Nominal) ή πολυωνυμική (polynomial) ή πολυτομική (polychotomus) ή κατηγορική αδιαβάθμητη (non-ordered categorical) ή πολυμερής μεταβλητή απόκρισης. Περιέχει τρεις ή περισσότερες κατηγορίες χωρίς κάποια φυσική διαβάθμιση (Καλλιακμάνης, 2020).

5.3.1. Μοντέλα λογιστικής παλινδρόμησης

Πριν εφαρμοστεί η λογιστική παλινδρόμηση θα πρέπει να επιλεχτεί και το κατάλληλο μοντέλο που θα χρησιμοποιήσει ο αναλυτής. Παρακάτω παρουσιάζονται τα τρία επικρατέστερα μοντέλα.

Το μοντέλο probit

Το μοντέλο που στηρίζεται στην συνάρτηση σύνδεσης probit εισήχθη για πρώτη φορά από τον Chester Ittner Bliss το 1934. Η προσαρμογή του μοντέλου έχει ως σκοπό την εκτίμηση της πιθανότητας μιας παρατήρησης με συγκεκριμένα χαρακτηριστικά ώστε να εμπίπτει σε μια συγκεκριμένη κατηγορία. Η μορφή του μοντέλου που προκύπτει είναι η εξής:

$$P(Y = 1| X) = \Phi^{-1}(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$$

όπου η Y αποτελεί μια δίτιμη εξαρτημένη μεταβλητή, $X = (X_1, X_2, \ldots, X_p)$ επεξηγηματικές μεταβλητές και η Φ είναι η αθροιστική συνάρτηση κατανομής της τυπικής κανονικής κατανομής.

Το μοντέλο logit

Όταν το γενικευμένο γραμμικό μοντέλο έχει ως συνάρτηση σύνδεσης την logit, η μορφή του μετατρέπεται στην:

$$\text{logit}(p_i) = \log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Λύνοντας την παραπάνω σχέση ως προς p_i επαληθεύεται αρχικά η σχετική πιθανότητα (odds) του μοντέλου της λογιστικής παλινδρόμησης:

$$\frac{p_i}{1-p_i} = e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}$$

και συνεχίζοντας προκύπτει και ο τύπος που εκφράζει την πιθανότητα:

$$p_i = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+ e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}$$

Ο λόγος για τον οποίο προτιμάται η συνάρτηση logit, σε σχέση με τις υπόλοιπες, είναι η εύκολη διαισθητική ερμηνεία των αποτελεσμάτων με βάση τη σχετική πιθανότητα. Η συνάρτηση αναφέρεται στον λογάριθμο της σχετικής πιθανότητας του ενδεχομένου που μας ενδιαφέρει (odds).

Το μοντέλο clog-log

Το complementary log log μοντέλο ταιριάζει σε περιπτώσεις που η P(Y=1) πλησιάζει γρήγορα τη μονάδα, αλλά αργά το 0. Το μοντέλο αυτό έχει τη μορφή:

$$\text{cloglog}(\pi) \equiv \log(-\log(1-\pi_x)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

όπου $\pi_x = P(Y = 1|X_i = x_i)$.

5.3.2. Μέτρα προσαρμογής

Εφόσον προσαρμόσουμε το καταλληλότερο μοντέλο για τα δεδομένα μας, θα χρειαστεί να επιλεγεί και αυτό που περιγράφει καλύτερα την μεταβλητή απόκρισης. Στη συνήθη παλινδρόμηση ένα αριθμητικό μέτρο που αξιολογεί την προσαρμογή ενός μοντέλου είναι ο συντελεστής προσδιορισμού R^2. Για την αξιολόγηση ενός λογιστικού μοντέλου (ή, γενικότερα, ενός ΓΓΜ) δεν υπάρχει ένα γενικά αποδεκτό μέτρο, αντίστοιχο του R^2. Υπάρχουν ωστόσο διάφορα μέτρα που έχουν προταθεί (συνήθως είναι γνωστά ως pseudo-R^2). Στη συνέχεια, L_M είναι η μέγιστη πιθανοφάνεια υπό το μοντέλο M και L_0 είναι η μέγιστη πιθανοφάνεια του μοντέλου μόνο με το σταθερό όρο. Παρακάτω παρουσιάζονται τα σημαντικότερα μέτρα προσαρμογής.

AIC

Ένα από τα πιο δημοφιλή κριτήρια για την εξής επιλογή είναι το κριτήριο AIC. Για ένα μοντέλο με k παραμέτρους το κριτήριο ορίζεται ως:

$$\text{AIC} = -2\ log\ L_M + 2k$$

Ανάμεσα σε διαφορετικά μοντέλα, επιλέγεται ως βέλτιστο αυτό που έχει την μικρότερη τιμή του δείκτη AIC.

BIC

Αυτό το κριτήριο προτάθηκε από τον Schwarz το 1978 και χρησιμοποιείται όταν έχουμε να επιλέξουμε ανάμεσα σε δύο ή περισσότερα μοντέλα. Το κριτήριο BIC είναι πιο αυστηρό σε σχέση με το AIC, διότι επιφέρει μεγαλύτερη ποινή για τον αριθμό παραμέτρων στο μοντέλο που χρησιμοποιείται. Ο τύπος ορίζεται ως εξής:

$$BIC = \log(n)\, k - 2\log(L)$$

McFadden

Το κριτήριο McFadden είναι το πιο δημοφιλές κριτήριο μέτρου προσαρμογής που έχει να κάνει με τα ψευδό – R^2. Τιμές ανάμεσα στο 0.2 και 0.4 δηλώνουν καλή προσαρμογή. Ο τύπος του είναι:

$$1 - \frac{logL_M}{logL_0}$$

Cox and Snell

Ο τύπος είναι:

$$1 - (\frac{L_0}{L_M})^{\frac{2}{n}}$$

Nagelkerke /Cragg and Uhler

Τροποποίηση του παραπάνω, όταν διαιρεθεί με τη μέγιστη τιμή του. Η τιμή του είναι:

$$\frac{1 - (L_0 - L_M)^{\frac{2}{n}}}{1 - L_0^{\frac{2}{n}}}$$

5.3.3. Έλεγχοι στατιστικής σημαντικότητας μεταβλητών

Έπειτα από την κατάλληλη επιλογή του μοντέλου που θα χρησιμοποιηθεί για την ανάλυση, θα πρέπει να γίνει έλεγχος για την σημαντικότητα των μεταβλητών που είναι χρήσιμες για να εξηγηθεί η μεταβλητότητα. Μια μέθοδος για να ελέγξουμε την σημαντικότητα κάθε μεταβλητής σε ένα μοντέλο λογιστικής παλινδρόμησης είναι ο έλεγχος του Wald. Από την θεωρία εκτιμητικής, οι ε.μ.π. για τις παραμέτρους ενός γενικευμένου γραμμικού μοντέλου έχουν ασυμπτωτικά την κανονική κατανομή. Συνεπώς, για μεγάλο αριθμό 30 παρατηρήσεων, μπορούμε να χρησιμοποιήσουμε την κανονική κατανομή για να εξετάσουμε αν μία παράμετρος β διαφέρει σημαντικά από το μηδέν. Η μηδενική και η εναλλακτική υπόθεση που ελέγχει ο έλεγχος είναι:

Μηδενική υπόθεση H_0: $\beta = 0$

VS

Εναλλακτική υπόθεση H_1: $\beta \neq 0$

Η στατιστική συνάρτηση που χρησιμοποιείται για τον έλεγχο της μηδενικής υπόθεσης, στατιστικής σημαντικότητας της μεταβλητής του μοντέλου είναι:

$$z = \frac{\beta}{s.e(\hat{\beta})}$$

που, όταν ισχύει η μηδενική υπόθεση H_0: $\beta = 0$, ακολουθεί ασυμπτωτικά τη N (0,1) κατανομή.

Παρόλο που ο έλεγχος Wald αποτελεί σημαντικό έλεγχο για την στατιστική σημαντικότητα των μεταβλητών σε ένα μοντέλο λογιστικής παλινδρόμησης, δεν είναι ο μοναδικός που μπορεί να πραγματοποιηθεί. Η βασική έννοια, η οποία χρησιμοποιείται για να ελέγξουμε την καλή προσαρμογή ενός μοντέλου είναι αυτή της απόκλισης (Πολίτης, 2021). Η έννοια αυτή είναι παρόμοια με αυτή του αθροίσματος τετραγώνων των καταλοίπων που υπάρχει στα κανονικά γραμμικά μοντέλα, καθώς αποτελεί μέτρο της ανερμήνευτης μεταβλητότητας του μοντέλου επί μια σταθερά. Ως απόκλιση θεωρούμε τον λογάριθμο της πιθανοφάνειας του προσαρμοσμένου μοντέλου. Διαισθητικά, όσο πιο μικρή είναι η απόκλιση ενός μοντέλου, τόσο πιο κοντά είναι στο κορεσμένο μοντέλο, και αυτό παρέχει ένδειξη καλής προσαρμογής.

Στη γενική περίπτωση, η κατανομή της απόκλισης δεν είναι γνωστή. Λόγω αυτού, αξιοποιείται ο έλεγχος του λόγου πιθανοφανειών, που στην ουσία αποτελεί την διαφορά της απόκλισης του κορεσμένου μοντέλου (αυτού που αποτελείται από τόσες μεταβλητές, όσες και οι παρατηρήσεις μας) και του προσαρμοσμένου μοντέλου, η οποία εκφράζεται από τον παρακάτω τύπο:

$$D_1 - D_2 = -2\left(\frac{logL(reduced\ model)}{logL(saturated\ model)}\right)$$

Υπό τη μηδενική υπόθεση, ότι το μοντέλο μας δεν διαφέρει από το κορεσμένο μοντέλο, η συγκεκριμένη ποσότητα γνωρίζουμε ότι ακολουθεί την κατανομή X_p^2, όπου ισχύει $p = \mathrm{d}f_1 - \mathrm{d}f_2$, ενώ $L()$ είναι η συνάρτηση πιθανοφάνειας.

5.3.4. Έλεγχος Πολυσυγγραμικότητας (VIF)

Ένας ακόμα έλεγχος που αποφέρει σημαντικά συμπεράσματα για το μοντέλο που έχουμε δημιουργήσει είναι ο έλεγχος της πολυσυγγραμικότητας των μεταβλητών. Η πολυσυγγραμικότητα (multicollinearity) δεν επηρεάζει την πρόβλεψη της μεταβλητής απόκρισης, προκαλεί όμως σύγχυση στην εκτίμηση των συντελεστών του μοντέλου (προκύπτουν μεγάλα τυπικά σφάλματα), δηλαδή δε μπορούν να καθοριστούν οι επιδράσεις των επεξηγηματικών μεταβλητών στη μεταβλητή απόκρισης. Η πολυσυγγραμικότητα μπορεί να ανιχνευθεί κυρίως με τη βοήθεια της ανοχής (tolerance) και του συντελεστή πληθωρισμού διακύμανσης (VIF).

Η ανοχή (Senaviratna & A. Cooray, 2019) είναι το ποσοστό της διακύμανσης μιας επεξηγηματικής μεταβλητής, που δεν μπορεί να εξηγηθεί από τις άλλες ανεξάρτητες μεταβλητές. Εξ ορισμού η ανοχή οποιασδήποτε συγκεκριμένης επεξηγηματικής μεταβλητής ισούται με $1 - R_i^2$, όπου R_i^2 είναι ο συντελεστής προσδιορισμού που προκύπτει από την παλινδρόμηση των άλλων μεταβλητών πάνω στην i-οστή μεταβλητή. Τιμές ανοχής κοντά στο 1 δείχνουν ότι υπάρχει μικρή πολυσυγγραμικότητα, ενώ μια τιμή κοντά στο μηδέν υποδηλώνει ότι η πολυσυγγραμικότητα μπορεί να αποτελεί πρόβλημα. Ο συντελεστής πληθωρισμού διακύμανσης (VIF) έχει τύπο:

$$VIF = \frac{1}{1-R^2}$$

και δείχνει πόσο διογκώνεται η διακύμανση της εκτίμησης του συντελεστή από την ύπαρξη πολυσυγγραμικότητας. Η τετραγωνική ρίζα του VIF υποδηλώνει πόσο μεγαλύτερο είναι το τυπικό σφάλμα, σε σύγκριση με το πόσο θα ήταν, εάν αυτή η μεταβλητή δεν ήταν συσχετισμένη με τις άλλες επεξηγηματικές μεταβλητές. Τιμές VIF που υπερβαίνουν το 10 συχνά θεωρείται ότι υποδεικνύουν πολυσυγγραμικότητα, αλλά σε ασθενέστερα μοντέλα (κάτι που συμβαίνει συχνά στην λογιστική παλινδρόμηση) τιμές πάνω από το 5 μπορεί να είναι αιτία ανησυχίας.

5.3.5. Έλεγχος ολικής επάρκειας μοντέλου

Τελευταίος έλεγχος που χρειάζεται να υλοποιηθεί είναι ο έλεγχος της ολικής επάρκειας του μοντέλου, όπου για δίτιμα δεδομένα προτείνεται ο έλεγχος των Hosmer-Lemeshow. Οι υποθέσεις του συγκεκριμένου ελέγχου είναι:

Μηδενική υπόθεση H_0: οι παρατηρηθείσες τιμές της Y δε διαφέρουν σημαντικά, από τις εκτιμώμενες τιμές.

VS

Εναλλακτική υπόθεση H_1: όχι H_0

Για την εφαρμογή του ελέγχου θα πρέπει πρώτα να διαταχθούν οι παρατηρήσεις ανάλογα με την προβλεπόμενη πιθανότητα επιτυχίας. Έπειτα, χωρίζονται οι διατεταγμένες παρατηρήσεις σε g ομάδες και για καθεμία από αυτές καταγράφουμε τον αριθμό επιτυχιών και αποτυχιών, σχηματίζοντας έτσι έναν πίνακα διαστάσεων g×2. Η στατιστική συνάρτηση του ελέγχου ορίζεται ως εξής:

$$X_{HL} = \sum_{g=1}^{G} \frac{(O_g - E_g)^2}{E_g(1-\frac{E_g}{n_g})}$$

όπου O_g είναι οι παρατηρηθείσες τιμές, E_g είναι οι εκτιμώμενες τιμές και n_g είναι ο αριθμός των παρατηρήσεων για την g-οστή ομάδα, και G είναι ο αριθμός των ομάδων. Η στατιστική συνάρτηση, υπό την H_0, ακολουθεί την κατανομή X^2_{g-2}.

5.4. Προσαρμογή λογιστικής παλινδρόμησης

Στην παρούσα παράγραφο γίνεται η εφαρμογή της θεωρίας που παρουσιάστηκε στις προηγούμενες ενότητες για τις κατηγορικές μεταβλητές Playoffs και Champions στις δύο φάσεις της διοργάνωσης του NBA. Επειδή αμφότερες οι μεταβλητές είναι δίτιμες, θα χρησιμοποιηθεί η δίτιμη λογιστική παλινδρόμηση σε όλες τις αναλύσεις.

5.4.1. Λογιστική παλινδρόμηση με χρήση όλων των μεταβλητών

Στο πρώτο μέρος της λογιστικής παλινδρόμησης θα παρουσιαστούν γραμμικά μοντέλα τα οποία περιέχουν όλες τις μεταβλητές που έχουν συλλεχθεί στα δεδομένα μας. Πιο συγκεκριμένα, για την φάση της κανονικής περιόδου της διοργάνωσης εκτελέστηκαν δύο μοντέλα λογιστικής παλινδρόμησης. Το πρώτο χρησιμοποιεί σαν κατηγορική μεταβλητή την Playoffs το δεύτερο την μεταβλητή Champions.

5.4.1.1. Για την κανονική περίοδο βάσει της κατηγορικής μεταβλητής Playoffs

Το μοντέλο που θα προσαρμοστεί έχει ως μεταβλητή απόκρισης την κατηγορική μεταβλητή των Playoffs και ως ανεξάρτητες μεταβλητές όλες τις ποσοτικές μεταβλητές που περιέχονται στα δεδομένα που έχουμε συλλέξει για τις τελευταίες δεκαπέντε (15) σεζόν της διοργάνωσης. Λόγω του όγκου των μεταβλητών έγινε χρήση της συνάρτησης step του πακέτου 'MASS' στο στατιστικό πακέτο της R, εφόσον παρόμοια συνάρτηση δεν υπάρχει στην Python. Όσον αφορά την κατεύθυνση που θα χρησιμοποιηθεί, είναι αυτή που συνδυάζει και τις δύο περιπτώσεις (both), δηλαδή και τον προς τα πίσω αποκλεισμό και τον προς τα εμπρός.

Για την σωστή εκτέλεση της συνάρτησης step πρέπει να οριστεί το απλούστερο μοντέλο λογιστικής παλινδρόμησης που προκύπτει μέσα από τα δεδομένα και το πολυπλοκότερο. Χρησιμοποιώντας μόνο τις μεταβλητές χωρίς τις αλληλεπιδράσεις τους, παρουσιάστηκε πρόβλημα στο πολυπλοκότερο μοντέλο λόγω των συσχετίσεων ανάμεσα στις μεταβλητές. Πιο συγκεκριμένα, οι μεταβλητές MisFG, MisFT, POSSt, POSSo, DEFF και EFF έχουν τιμές συσχέτισης, με άλλες μεταβλητές των δεδομένων, που οριακά πλησιάζουν το ένα (τέλεια συσχέτιση) και το στατιστικό πρόγραμμα εμφανίζει NA τιμές στους συντελεστές της παλινδρόμησης, η συγκεκριμένη ένδειξη είναι φανερή στο Παράρτημα της εργασίας όπου και δίνεται το αποτέλεσμα της παλινδρόμησης από το στατιστικό πακέτο της R (Π3. Συσχετίσεις στην λογιστική πσλινδρόμηση). Μια παρόμοια περίπτωση σχολιάστηκε και στο κεφάλαιο 4 της παρούσας εργασίας. Εν ολίγοις, το στατιστικό πρόγραμμα προτείνει να αφαιρεθούν από την ανάλυση οι παρούσες μεταβλητές λόγω της μεγάλης συσχέτισης ανάμεσα τους. Συνεπώς, το πολυπλοκότερο πλέον μοντέλο περιέχει τις υπόλοιπες ποσοτικές μεταβλητές. Έπειτα από την εκτέλεση της συνάρτησης step καταλήξαμε πως το καταλληλότερο μοντέλο είναι αυτό που χρησιμοποιεί σαν ανεξάρτητες μεταβλητές τις ORt και DRt, με δείκτη AIC ίσο με 212,10.

Εφόσον βρέθηκε το καταλληλότερο μοντέλο από την χρήση του στατιστικού πακέτου της R θα εφαρμόσουμε γι' αυτές τις μεταβλητές την λογιστική παλινδρόμηση για να υλοποιήσουμε και τον έλεγχο στατιστικής σημαντικότητας των μεταβλητών. Όπως φαίνεται στο Σχήμα 5.3, και οι δύο μεταβλητές έχουν p_{value} για τον έλεγχο Wald μικρότερο του επιπέδου σημαντικότητας που είναι ίσο με 5%. Ακόμα είναι φανερό πως ο αλγόριθμος μεγιστοποίησε το λογάριθμο της πιθανοφάνειας έπειτα από επτά (7) επαναλήψεις, το οποίο αποτελεί θετικό σημάδι, αφού χρειαζόμαστε όσες το λιγότερο επαναλήψεις για να είναι πιο ευσταθές το μοντέλο.

Generalized Linear Model Regression Results

Dep. Variable:	Playoffs	No. Observations:	450
Model:	GLM	Df Residuals:	447
Model Family:	Binomial	Df Model:	2
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-103.05
Date:	Mon, 18 Sep 2023	Deviance:	206.10
Time:	08:57:09	Pearson chi2:	255.
No. Iterations:	7	Pseudo R-squ. (CS):	0.6032
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	13.8982	5.570	2.495	0.013	2.982	24.815
ORt	0.9699	0.107	9.099	0.000	0.761	1.179
DRt	-1.0964	0.120	-9.121	0.000	-1.332	-0.861

Σχήμα: 5.3: Output για το προσαρμοσμένο μοντέλο της μεταβλητής playoffs

Για τον έλεγχο καλής προσαρμογής του μοντέλου βάσει της απόκλισης χρησιμοποιήθηκε η εντολή anova(), που περιέχεται στο στατιστικό πακέτο της R και παρόμοια δεν υπάρχει στην Python. Σύμφωνα με τον Πίνακα 5.1, τα p_{value} είναι μικρότερα του επιπέδου σημαντικότητας. Επομένως και οι δύο μεταβλητές προσφέρουν κάτι σημαντικό αναφορικά με την απόκλιση στο προσαρμοσμένο μοντέλο.

Μεταβλητές	p_{value}
ORt	**<2.2e-16**
DRt	**<2.2e-16**

Πίνακας 5.1: Τιμές των p_{value} για τον έλεγχο σημαντικότητας μεταβλητών βάσει της απόκλισης για το προσαρμοσμένο μοντέλο της μεταβλητής playoffs

Στη συνέχεια γίνεται ο έλεγχος της πολυσυγγραμικότητας των μεταβλητών που συμμετείχαν στην λογιστική παλινδρόμηση. Όπως φαίνεται στον Πίνακα 5.2, δεν υπάρχει θέμα με την πολυσυγγραμικότητα των μεταβλητών.

Μεταβλητές	Τιμές VIF
ORt	**1.09408**
DRt	**1.09408**

Πίνακας 5.2: Τιμές του δείκτη VIF για το προσαρμοσμένο μοντέλο της μεταβλητής playoffs

Τέλος, εκτελέστηκε ο έλεγχος για την ολική επάρκεια του μοντέλου των Hosmer – Lemeshow, όπου σύμφωνα με τον Πίνακα 5.3, καθώς το p_{value} ήταν ίσο με 0.9225, δηλαδή μεγαλύτερο του επιπέδου σημαντικότητας 5%, δεν έχουμε ενδείξεις απόρριψης της μηδενικής υπόθεσης. Συνεπώς, σε αυτή την περίπτωση έχουμε μία πολύ ισχυρή ένδειξη καλής προσαρμογής του μοντέλου μας.

Έλεγχος	p_{value}
Hosmer-Lemeshow	**0.9225**

Πίνακας 5.3: Output για τον έλεγχο Hosmer – Lemeshow για το προσαρμοσμένο μοντέλο της μεταβλητής playoffs

Συνεπώς καταλήγουμε στο συμπέρασμα, μέσω λογιστικής παλινδρόμησης, πως οι μεταβλητές offensive rating και defensive rating είναι απαραίτητες για την προαγωγή ή μη των ομάδων στην φάση των playoffs. Σε ανάλογο συμπέρασμα για τις νίκες των ομάδων στη φάση της κανονικής περιόδου είχαν καταλήξει οι Teramoto και Croos (2010). Αυτό το αποτέλεσμα ήταν αναμενόμενο, εφόσον οι δύο μεταβλητές που είναι σημαντικότερες είχαν δημιουργηθεί βάσει κάποιων μεταβλητών από τις υπόλοιπες. Οι μεταβλητές που χρησιμοποιήθηκαν για τη δημιουργία τους παρουσιάστηκαν αναλυτικά στο κεφάλαιο 2, όπου και έχει δοθεί ο τύπος τους. Τέλος, το μοντέλο έχει τύπο:

$$\log\left(\frac{p_i}{1-p_i}\right) = 13.8982 + 0.9699\ \text{ORt} - 1.0964\ \text{DRt}$$

Για την ερμηνεία των συντελεστών αυτού του μοντέλου προκύπτουν τα παρακάτω συμπεράσματα:

- Αν αυξήσουμε την τιμή της μεταβλητής ORt κατά μία μονάδα, ενώ διατηρούμε σταθερή την τιμή της άλλης μεταβλητής, ο λογάριθμος της σχετικής πιθανότητας επιτυχίας θα αυξηθεί κατά 0.9699 μονάδες.
- Αν αυξήσουμε την τιμή της μεταβλητής DRt κατά μία μονάδα, ενώ διατηρούμε σταθερή την τιμή της μεταβλητής ORt, ο λογάριθμος της σχετικής πιθανότητας επιτυχίας θα μειωθεί τόσες μονάδες όσες είναι η εκτίμηση του συντελεστή, -1.0964.

Συνεπώς, οι μεταβλητές που επηρεάζουν περισσότερο την είσοδο μίας ομάδας στην φάση των playoffs είναι οι δείκτες ORt και DRt. Πιο συγκεκριμένα, όσο αυξάνεται ο δείκτης ORt, τόσο πιο πιθανό είναι η ομάδα να περάσει στη δεύτερη φάση της διοργάνωσης. Το ακριβώς αντίθετο συμβαίνει με τον DRt δείκτη.

5.4.1.2. Για την κανονική περίοδο βάσει της κατηγορικής μεταβλητής Champions

Για την ανάλυση βάσει της κατηγορικής μεταβλητής Champions για την κανονική περίοδο ακολουθήθηκε η ίδια μέθοδος με αυτή που χρησιμοποιήθηκε στην προηγουμένη παράγραφο. Με τη χρήση της συνάρτησης step στο στατιστικό πακέτο της R, χρησιμοποιώντας ως απλούστερο και πολυπλοκότερο μοντέλο τα ίδια με την προηγούμενη ανάλυση, καταλήξαμε πως το καταλληλότερο μοντέλο είναι αυτό που χρησιμοποιεί σαν ανεξάρτητες μεταβλητές τις: ποσοστά επιτυχίας δίποντων και τρίποντων (FG%), defensive rating (DRt) και επιτυχημένες προσπάθειες δίποντων και τρίποντων (FGM). Όπως φαίνεται στο Σχήμα 5.4 και βάσει των p_{value} και οι τρείς μεταβλητές, σύμφωνα με τον έλεγχο Wald, είναι στατιστικά σημαντικές. Για τον έλεγχο βάσει της απόκλισης του μοντέλου, σύμφωνα με τον Πίνακα 5.4, και οι τρείς μεταβλητές προσφέρουν κάτι σημαντικό στο μοντέλο. Για τον έλεγχο της πολυσυγγραμικότητας, όπως φαίνεται στον Πίνακα 5.5, οι τιμές του δείκτη VIF είναι αρκετά μικρές και επομένως δεν προκύπτει κανένα πρόβλημα.

Generalized Linear Model Regression Results

Dep. Variable:	Champions	No. Observations:	450
Model:	GLM	Df Residuals:	446
Model Family:	Binomial	Df Model:	3
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-41.473
Date:	Mon, 18 Sep 2023	Deviance:	82.946
Time:	15:18:28	Pearson chi2:	199.
No. Iterations:	8	Pseudo R-squ. (CS):	0.1023
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-21.6891	12.984	-1.670	0.095	-47.138	3.760
df['FG%']	82.9683	25.506	3.253	0.001	32.977	132.959
DRt	-0.3525	0.093	-3.790	0.000	-0.535	-0.170
FGM	0.4284	0.176	2.428	0.015	0.083	0.774

Σχήμα: 5.4: Output για το προσαρμοσμένο μοντέλο της μεταβλητής champions για την κανονική περίοδο

Μεταβλητές	p_{value}
FG%	**2.945e-8**

DRt	**0.00063**
FGM	**0.01296**

Πίνακας 5.4: Τιμές των p_{value} για τον έλεγχο σημαντικότητας μεταβλητών βάσει της απόκλισης για το προσαρμοσμένο μοντέλο της μεταβλητής champions για την κανονική περίοδο

Μεταβλητές	Τιμές VIF
FG%	**2.04375**
DRt	**1.49464**
FGM	**2.6224**

Πίνακας 5.5: Τιμές του δείκτη VIF για το προσαρμοσμένο μοντέλο της μεταβλητής champions για την κανονική περίοδο

Τέλος, για τον έλεγχο ολικής επάρκειας του μοντέλου, Πίνακας 5.6, δεν έχουμε ενδείξεις απόρριψης της μηδενικής υπόθεσης εφόσον το p_{value} είναι ίσο με 0.9742. Επομένως, για την πρόβλεψη της ομάδας που θα στεφθεί πρωταθλήτρια στην διοργάνωση του NBA σημαντικότερες μεταβλητές για την κανονική περίοδο κρίνονται τα ποσοστά επιτυχίας δίποντων και τρίποντων (FG%), defensive rating (DRt) και επιτυχημένες προσπάθειες δίποντων και τρίποντων (FGM).

Έλεγχος	p_{value}
Hosmer-Lemeshow	**0.9742**

Πίνακας 5.6: Output για τον έλεγχο Hosmer – Lemeshow για το προσαρμοσμένο μοντέλο της μεταβλητής champions για την κανονική περίοδο

Τέλος, το μοντέλο έχει τύπο:

$$\log(\frac{p_i}{1-p_i}) = -21.6841 + 82.9683\ FG\% - 0.3525\ DRt + 0.4284\ FGM$$

Για την ερμηνεία των συντελεστών αυτού του μοντέλου προκύπτουν τα παρακάτω συμπεράσματα:

- Για κάθε αύξηση του ποσοστού της μεταβλητής ποσοστών επιτυχίας δίποντων ή τριπόντων (FG%) κατά μία μονάδα, ενώ διατηρούμε σταθερή τις τιμές των άλλων μεταβλητών, ο λογάριθμος της σχετικής πιθανότητας επιτυχίας θα αυξηθεί κατά 82.9683 μονάδες.
- Αν αυξήσουμε την τιμή της μεταβλητής DRt κατά μία μονάδα, ενώ διατηρούμε σταθερές τις τιμές των άλλων μεταβλητών, ο λογάριθμος της σχετικής πιθανότητας επιτυχίας θα μειωθεί τόσες μονάδες όσες είναι η εκτίμηση του συντελεστή, -0.3525.
- Αν αυξήσουμε την τιμή της μεταβλητής επιτυχημένων δίποντων ή τριπόντων κατά μία μονάδα, ενώ διατηρούμε σταθερές τις τιμές των άλλων μεταβλητών, ο λογάριθμος της σχετικής πιθανότητας επιτυχίας θα αυξηθεί κατά 0.4284 μονάδες.

Από την ανάλυση που προηγήθηκε συμπεραίνουμε πως οι μεταβλητές που κρίνουν τον πρωταθλητή της διοργάνωσης του NBA είναι οι δείκτες DRt, τα ποσοστά επιτυχίας

στα δίποντα ή τρίποντα και οι επιτυχημένες προσπάθειες τους. Ακόμα, όπως γίνεται αντιληπτό, όσο αυξάνονται οι τιμές των ποσοστών επιτυχίας των δίποντων ή τρίποντων, αυτό εξελίσσεται σε θετικό παράγοντα για να κατακτήσει η ομάδα τον τίτλο του πρωταθλητή. Το ίδιο ισχύει και για την αύξηση των τιμών των επιτυχημένων προσπαθειών δίποντων και τρίποντων.

5.4.1.3. Για τα Playoffs βάσει της κατηγορικής μεταβλητής Champions

Σε αυτήν την φάση της διοργάνωσης η μόνη ανάλυση που μπορεί να γίνει είναι βάσει της κατηγορικής μεταβλητής Champions. Προφανώς, ακολουθήθηκε η ίδια διεργασία που εφαρμόστηκε παραπάνω. Μέσω της συνάρτησης step βρέθηκε το καταλληλότερο μοντέλο, το οποίο περιέχει τις μεταβλητές offensive rating, defensive rating, συνολικές προσπάθειες δίποντων και τρίποντων των αντιπάλων (FGAo), συνολικές προσπάθειες ελεύθερων βολών της ομάδας (FTA) και τα επιθετικά ριμπάουντ της ομάδας (ORB). Έπειτα από την εκτέλεση της λογιστικής παλινδρόμησης, Σχήμα 5.5, οι μεταβλητές βάσει του στατιστικού ελέγχου Wald κρίνονται όλες στατιστικά σημαντικές.

Generalized Linear Model Regression Results

Dep. Variable:	Champions	No. Observations:	240
Model:	GLM	Df Residuals:	234
Model Family:	Binomial	Df Model:	5
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-16.570
Date:	Tue, 19 Sep 2023	Deviance:	33.140
Time:	12:12:52	Pearson chi2:	68.7
No. Iterations:	10	Pseudo R-squ. (CS):	0.2807
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-15.5546	12.441	-1.250	0.211	-39.939	8.830
ORt	1.2488	0.356	3.509	0.000	0.551	1.946
DRt	-0.9882	0.302	-3.267	0.001	-1.581	-0.395
FTA	-0.5760	0.208	-2.773	0.006	-0.983	-0.169
FGAo	-0.2597	0.131	-1.979	0.048	-0.517	-0.003
ORB	1.3529	0.475	2.851	0.004	0.423	2.283

Σχήμα: 5.5: Output για το προσαρμοσμένο μοντέλο της μεταβλητής champions για την φάση των playoffs

Για τον έλεγχο της απόκλισης του μοντέλου, Πίνακας 5.7, διακρίνουμε πως ορισμένες μεταβλητές έχουν p_{value} μεγαλύτερο του επιπέδου σημαντικότητας και επομένως δεν προσφέρουν κάποια σημαντική πληροφορία στο μοντέλο. Εφόσον αφαιρέθηκαν καθεμία ξεχωριστά, παρατηρήθηκε ότι καταλληλότερο μοντέλο είναι αυτό με μεταβλητές τις offensive rating, defensive rating και τα επιθετικά ριμπάουντ της ομάδας (ORB), με δείκτη AIC ίσο με 54,596. Ακόμα, ο αλγόριθμος μεγιστοποίησε το λογάριθμο της πιθανοφάνειας μετά από 9 επαναλήψεις σε αντίθεση με το προηγούμενο μοντέλο που χρειαζόταν 10.

Μεταβλητές	p_{value}
Ort	**9.786e-6**
DRt	**2.119e-10**
FTA	**0.07345**
FGAo	**0.10482**
ORB	**0.00026**

Πίνακας 5.7: Τιμές των p_{value} για τον έλεγχο σημαντικότητας μεταβλητών βάσει της απόκλισης για το προσαρμοσμένο μοντέλο της μεταβλητής champions για την φάση των playoffs

Όπως είναι φανερό στο Σχήμα 5.5, οι τιμές του ελέγχου Wald για τις μεταβλητές που χρησιμοποιήθηκαν είναι μικρότερες του επιπέδου σημαντικότητας. Επίσης, για τον έλεγχο της απόκλισης στον Πίνακα 5.8 διακρίνουμε πως οι τιμές του p_{value} είναι όλες αρκετά μικρότερες του 5% και επομένως κρίνονται ως στατιστικά σημαντικές.

Generalized Linear Model Regression Results

Dep. Variable:	Champions	No. Observations:	240
Model:	GLM	Df Residuals:	236
Model Family:	Binomial	Df Model:	3
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-23.298
Date:	Tue, 19 Sep 2023	Deviance:	46.596
Time:	13:36:17	Pearson chi2:	52.6
No. Iterations:	9	Pseudo R-squ. (CS):	0.2392
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-18.8238	11.508	-1.636	0.102	-41.379	3.732
ORt	0.7385	0.172	4.299	0.000	0.402	1.075
DRt	-0.6810	0.168	-4.047	0.000	-1.011	-0.351
ORB	0.6553	0.288	2.276	0.023	0.091	1.220

Σχήμα: 5.6: Output για το νέο προσαρμοσμένο μοντέλο της μεταβλητής champions για την φάση των playoffs

Μεταβλητές	p_{value}
Ort	**9.786e-6**
DRt	**2.119e-10**
ORB	**0.0168**

Πίνακας 5.8: Τιμές των p_{value} για τον έλεγχο σημαντικότητας μεταβλητών βάσει της απόκλισης για το νέο προσαρμοσμένο μοντέλο της μεταβλητής champions για την φάση των playoffs

Για τον έλεγχο της πολυσυγγραμικότητας, όπως φαίνεται στον Πίνακα 5.9, οι τιμές του δείκτη VIF είναι αρκετά μικρές και επομένως δεν προκύπτει κανένα πρόβλημα, και για τον έλεγχο ολικής επάρκειας του μοντέλου, σύμφωνα με τον Πίνακα 5.1, δεν έχουμε ενδείξεις απόρριψης της μηδενικής υπόθεσης εφόσον το p_{value} είναι ίσο με 0.9974. Επομένως, για την πρόβλεψη της ομάδας που θα στεφθεί πρωταθλήτρια στην διοργάνωση του NBA απαραίτητες μεταβλητές για τη φάση των playoffs κρίνονται οι offensive rating, defensive rating και τα επιθετικά ριμπάουντ της ομάδας (ORB).

Μεταβλητές	Τιμές VIF
Ort	**1.09496**
DRt	**1.09236**
ORB	**1.00339**

Πίνακας 5.9: Τιμές του δείκτη VIF για το νέο προσαρμοσμένο μοντέλο της μεταβλητής champions για την φάση των playoffs

Έλεγχος	p_{value}
Hosmer-Lemeshow	**0.9974**

Πίνακας 5.10: Output για τον έλεγχο Hosmer – Lemeshow για το νέο προσαρμοσμένο μοντέλο της μεταβλητής champions για την φάση των playoffs

Όπως και στην φάση της κανονικής περιόδου, έτσι και στα playoffs καταλήξαμε σε ανάλογο συμπέρασμα για τις νίκες των ομάδων με τους Teramoto και Croos (2010). Τέλος, το μοντέλο έχει τύπο:

$$\log(\frac{p_i}{1-p_i}) = -18.8238 + 0.7385\ ORt - 0.6810\ DRt + 0.6553\ ORB$$

Για την ερμηνεία των συντελεστών αυτού του μοντέλου προκύπτουν τα παρακάτω συμπεράσματα:

- Αν αυξήσουμε την τιμή της μεταβλητής ORt, ενώ διατηρούμε σταθερές τις τιμές των άλλων μεταβλητών, ο λογάριθμος της σχετικής πιθανότητας επιτυχίας θα αυξηθεί κατά 0.7385 μονάδες.
- Αν αυξήσουμε την τιμή της μεταβλητής DRt κατά μία μονάδα, ενώ διατηρούμε σταθερές τις τιμές των άλλων μεταβλητών, ο λογάριθμος της σχετικής πιθανότητας επιτυχίας θα μειωθεί τόσες μονάδες όσες είναι η εκτίμηση του συντελεστή, -0.6810.
- Αν αυξήσουμε την τιμή της μεταβλητής επιθετικών ριμπάουντ (ORB), ενώ διατηρούμε σταθερές τις τιμές των άλλων μεταβλητών, ο λογάριθμος της σχετικής πιθανότητας επιτυχίας θα αυξηθεί κατά 0.6553 μονάδες.

Όπως γίνεται αντιληπτό, όσο αυξάνονται οι τιμές του δείκτη ORt και των επιθετικών ριμπάουντ για μία ομάδα, τόσο πιο πιθανό είναι αυτή να κατακτήσει τον τίτλο, εφόσον έχει προκριθεί στην φάση των playoffs.

5.4.2. Λογιστική παλινδρόμηση χωρίς χρήση των δεικτών DRt και ORt

Όπως αναφέρθηκε και παραπάνω, οι τύποι των μεταβλητών DRt και ORt χρησιμοποιούν πληροφορίες από άλλες μεταβλητές που βρίσκονται στην παρούσα εργασία. Επομένως είναι λογικό πως όταν βρεθούν σε ένα γραμμικό μοντέλο θα είναι αυτές που θα είναι σημαντικότερες και δεν θα χρησιμοποιούνται οι υπόλοιπες. Γι' αυτόν τον λόγο κρίθηκε ορθό να παρουσιαστούν και γραμμικά μοντέλα βάσει της μεταβλητής Playoffs για την κανονική περίοδο και βάσει της μεταβλητής Champions για τις δύο φάσεις της διοργάνωσης, στα οποία οι δύο μεταβλητές δεν θα χρησιμοποιούνται εξ αρχής.

5.4.2.1. Για την κανονική περίοδο βάσει της μεταβλητής Playoffs

Σε αυτή την υποενότητα θα εκτελεστεί η λογιστική παλινδρόμηση για την κανονική περίοδο βάσει της μεταβλητής Playoffs, χωρίς την χρήση των δύο δεικτών που αναφέρθηκαν προηγουμένως. Ακολουθώντας την ίδια διεργασία που εφαρμόστηκε παραπάνω και μέσω της συνάρτησης step βρέθηκε το καταλληλότερο μοντέλο, το οποίο περιέχει τις μεταβλητές: πόντοι που πετυχαίνει η ομάδα (PPG), πόντοι αντιπάλων ομάδων (PPGA), λάθη της ομάδας (TOV) και των αντιπάλων (TOVo), επιτυχημένες προσπάθειες δίποντων ή τρίποντων (FGM), ποσοστά ευστοχίας στα δίποντα ή τρίποντα (FG%) και αμυντικά ριμπάουντ (DRB). Όπως φαίνεται στον Πίνακα 5.11 οι τιμές του δείκτη VIF, για τις μεταβλητές PPG και PPGA, ξεπερνάνε το χαμηλότερο όριο (10

μονάδες) και επομένως αποφασίζεται να αφαιρεθεί από την ανάλυση η μεταβλητή των πόντων των αντίπαλων ομάδων (PPGA). Η ίδια λογική χρησιμοποιήθηκε και στις επόμενες αναλύσεις που ακολουθούν επειδή και σε αυτές οι μεταβλητές των πόντων και των δύο ομάδων συμμετείχαν στη δημιουργία του καλύτερου μοντέλου.

Μεταβλητές	Τιμές VIF
PPG	**28.461**
PPGA	**22.078**

Πίνακας 5.11: Τιμές του δείκτη VIF των πόντων της ομάδας και των αντιπάλων (χωρίς DRt,ORt)

Εκτελώντας την παλινδρόμηση για τις υπόλοιπες μεταβλητές παρατηρούμε στο Σχήμα 5.7 πως όλες οι μεταβλητές είναι στατιστικά σημαντικές σύμφωνα με τον έλεγχο Wald.

Generalized Linear Model Regression Results

Dep. Variable:	Playoffs	No. Observations:	450
Model:	GLM	Df Residuals:	443
Model Family:	Binomial	Df Model:	6
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-181.12
Date:	Sun, 12 Nov 2023	Deviance:	362.24
Time:	09:22:51	Pearson chi2:	420.
No. Iterations:	6	Pseudo R-squ. (CS):	0.4387
Covariance Type:	nonrobust		

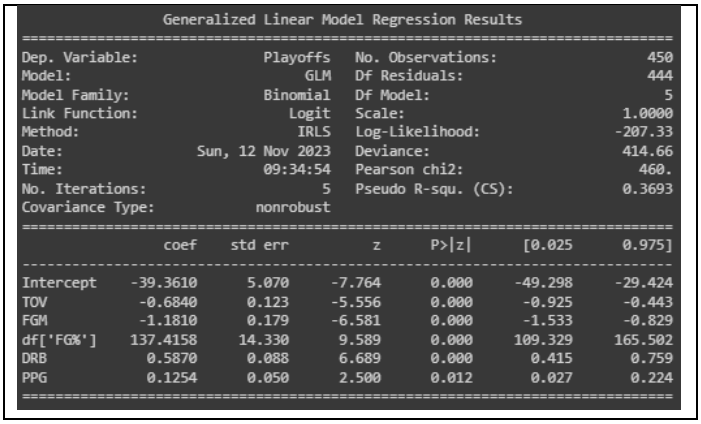
	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-61.7489	6.865	-8.995	0.000	-75.204	-48.294
TOV	-1.0852	0.153	-7.100	0.000	-1.385	-0.786
TOVo	1.0000	0.151	6.613	0.000	0.704	1.296
FGM	-1.5633	0.208	-7.507	0.000	-1.971	-1.155
df['FG%']	169.8566	17.001	9.991	0.000	136.536	203.178
DRB	1.0332	0.125	8.247	0.000	0.788	1.279
PPG	0.1187	0.054	2.214	0.027	0.014	0.224

Σχήμα: 5.7: Output για το νέο προσαρμοσμένο μοντέλο της μεταβλητής playoffs για την κανονική περίοδο (χωρίς DRt,ORt)

Για τον έλεγχο της απόκλισης του μοντέλου, Πίνακας 5.12, διακρίνουμε πως η μεταβλητή TOVo έχει p_{value} μεγαλύτερο του επιπέδου σημαντικότητας και επομένως δεν προσφέρει κάποια σημαντική πληροφορία στο μοντέλο. Εφόσον αφαιρέθηκε, παρατηρήθηκε ότι καταλληλότερο μοντέλο είναι αυτό με όλες τις προηγούμενες μεταβλητές πλην της συγκεκριμένης. Ακόμα, ο αλγόριθμος μεγιστοποίησε το λογάριθμο της πιθανοφάνειας μετά από πέντε (5) επαναλήψεις σε αντίθεση με το προηγούμενο μοντέλο που χρειαζόταν έξι (6).

Μεταβλητές	p_{value}
TOV	**6.734e-07**
TOVo	**0.091738**
FGM	**0.006766**
FG%	**< 2.2e-16**
DRB	**< 2.2e-16**
PPG	**0.024669**

Πίνακας 5.12: Τιμές των p_{value} για τον έλεγχο σημαντικότητας μεταβλητών βάσει της απόκλισης του προσαρμοσμένο μοντέλο της μεταβλητής playoffs για την κανονική περίοδο (χωρίς DRt,ORt)



Generalized Linear Model Regression Results

Dep. Variable:	Playoffs	No. Observations:	450
Model:	GLM	Df Residuals:	444
Model Family:	Binomial	Df Model:	5
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-207.33
Date:	Sun, 12 Nov 2023	Deviance:	414.66
Time:	09:34:54	Pearson chi2:	460.
No. Iterations:	5	Pseudo R-squ. (CS):	0.3693
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-39.3610	5.070	-7.764	0.000	-49.298	-29.424
TOV	-0.6840	0.123	-5.556	0.000	-0.925	-0.443
FGM	-1.1810	0.179	-6.581	0.000	-1.533	-0.829
df['FG%']	137.4158	14.330	9.589	0.000	109.329	165.502
DRB	0.5870	0.088	6.689	0.000	0.415	0.759
PPG	0.1254	0.050	2.500	0.012	0.027	0.224

Σχήμα: 5.8: Output για το νέο προσαρμοσμένο μοντέλο της μεταβλητής playoffs για την κανονική περίοδο (χωρίς DRt,ORt)

Όπως είναι φανερό στο Σχήμα 5.8, τα p_{value} του ελέγχου Wald για τις μεταβλητές που χρησιμοποιήθηκαν είναι μικρότερα του επιπέδου σημαντικότητας. Επίσης, για τον έλεγχο της απόκλισης, Πίνακας 5.13, διακρίνουμε πως οι τιμές είναι όλες μικρότερες του 5%.

Μεταβλητές	p_{value}
TOV	**6.734e-07**
FGM	**0.006908**
FG%	**< 2.2e-16**
DRB	**2.722e-16**
PPG	**0.011464**

Πίνακας 5.13: Τιμές των p_{value} για τον έλεγχο σημαντικότητας μεταβλητών βάσει της απόκλισης του προσαρμοσμένο μοντέλο της μεταβλητής playoffs για την κανονική περίοδο (χωρίς DRt,ORt)

Για τον έλεγχο της πολυσυγγραμικότητας, όπως φαίνεται στον Πίνακα 5.14, οι τιμές του δείκτη VIF είναι αρκετά μικρές εκτός από δύο τιμές που είναι κοντά στο δέκα (10), το οποίο είναι και το όριο του δείκτη. Για τον έλεγχο ολικής επάρκειας του μοντέλου, σύμφωνα με τον Πίνακα 5.15, έχουμε ενδείξεις απόρριψης της μηδενικής υπόθεσης, εφόσον το p_{value} είναι ίσο με 0.03269. Επομένως, για να βρεθεί το καταλληλότερο μοντέλο θα αφαιρέσουμε από το μοντέλο μας τη μεταβλητή FGM, που είναι η μια από τις δύο που έχουν αρκετά μεγάλη τιμή VIF, όπως φάνηκε στον Πίνακα 5.12.

Μεταβλητές	Τιμές VIF
TOV	**1.105**
FGM	**11.087**
FG%	**2.550**
DRB	**2.161**
PPG	**7.889**

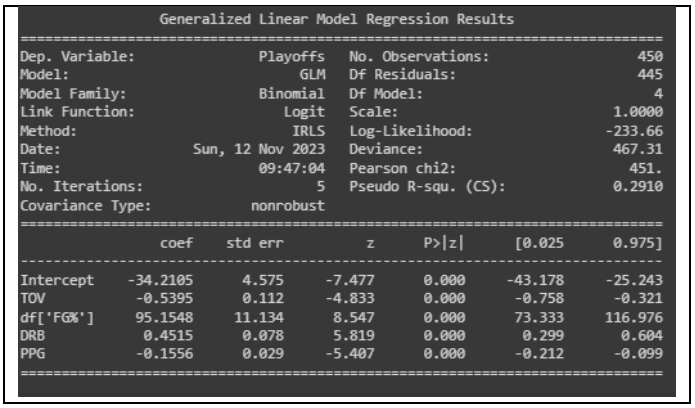
Πίνακας 5.14: Τιμές του δείκτη VIF για το νέο προσαρμοσμένο μοντέλο της μεταβλητής playoffs για την κανονική περίοδο (χωρίς DRt,ORt)

Έλεγχος	p_{value}

Hosmer-Lemeshow	0.03269

Πίνακας 5.15: Output για τον έλεγχο Hosmer – Lemeshow για το νέο προσαρμοσμένο μοντέλο της μεταβλητής champions για την φάση των playoffs

Εφόσον αφαιρέθηκε η μεταβλητή FGM, στο σχήμα 5.9 παρουσιάζονται τα αποτελέσματα του νέου γραμμικού μοντέλου.



Generalized Linear Model Regression Results

Dep. Variable:	Playoffs	No. Observations:	450
Model:	GLM	Df Residuals:	445
Model Family:	Binomial	Df Model:	4
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-233.66
Date:	Sun, 12 Nov 2023	Deviance:	467.31
Time:	09:47:04	Pearson chi2:	451.
No. Iterations:	5	Pseudo R-squ. (CS):	0.2910
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-34.2105	4.575	-7.477	0.000	-43.178	-25.243
TOV	-0.5395	0.112	-4.833	0.000	-0.758	-0.321
df['FG%']	95.1548	11.134	8.547	0.000	73.333	116.976
DRB	0.4515	0.078	5.819	0.000	0.299	0.604
PPG	-0.1556	0.029	-5.407	0.000	-0.212	-0.099

Σχήμα: 5.9: Output για το δεύτερο νέο προσαρμοσμένο μοντέλο της μεταβλητής playoffs για την κανονική περίοδο (χωρίς DRt,ORt)

Όπως είναι φανερό στο Σχήμα 5.9, οι τιμές του ελέγχου Wald για τις μεταβλητές που χρησιμοποιήθηκαν είναι μικρότερες του επιπέδου σημαντικότητας, όπως και για τον έλεγχο της απόκλισης, Πίνακας 5.16.

Μεταβλητές	p_{value}
TOV	**7.58e-14**
FG%	**1.34e-06**
DRB	**5.91e-09**
PPG	**6.39e-08**

Πίνακας 5.16: Τιμές των p_{value} για τον έλεγχο σημαντικότητας μεταβλητών βάσει της απόκλισης του δεύτερου νέου προσαρμοσμένου μοντέλου της μεταβλητής playoffs για την κανονική περίοδο (χωρίς DRt,ORt)

Για τον έλεγχο της πολυσυγγραμικότητας, όπως φαίνεται στον Πίνακα 5.16, οι τιμές του δείκτη VIF είναι αρκετά μικρές. Για τον έλεγχο ολικής επάρκειας του μοντέλου, σύμφωνα με τον Πίνακα 5.17, έχουμε ενδείξεις μη απόρριψης της μηδενικής υπόθεσης, εφόσον το p_{value} είναι ίσο με 0.07679. Επομένως, το μοντέλο που χρησιμοποιεί τις μεταβλητές των λαθών της ομάδας (TOV), των ποσοστών δίποντων ή τριπόντων (FG%), των αμυντικών ριμπάουντ (DRB) και των πόντων της ομάδας (PPG) κρίνεται ως καταλληλότερο.

Μεταβλητές	Τιμές VIF
TOV	**1.044**
FG%	**1.728**
DRB	**1.986**
PPG	**2.930**

Πίνακας 5.16: Τιμές του δείκτη VIF για το νέο προσαρμοσμένο μοντέλο της μεταβλητής playoffs για την κανονική περίοδο (χωρίς DRt,ORt)

Έλεγχος	p_{value}
Hosmer-Lemeshow	**0.07679**

Πίνακας 5.17: Output για τον έλεγχο Hosmer – Lemeshow για το νέο προσαρμοσμένο μοντέλο της μεταβλητής playoffs για την κανονική περίοδο (χωρίς DRt,ORt)

Τέλος, το μοντέλο έχει τύπο:

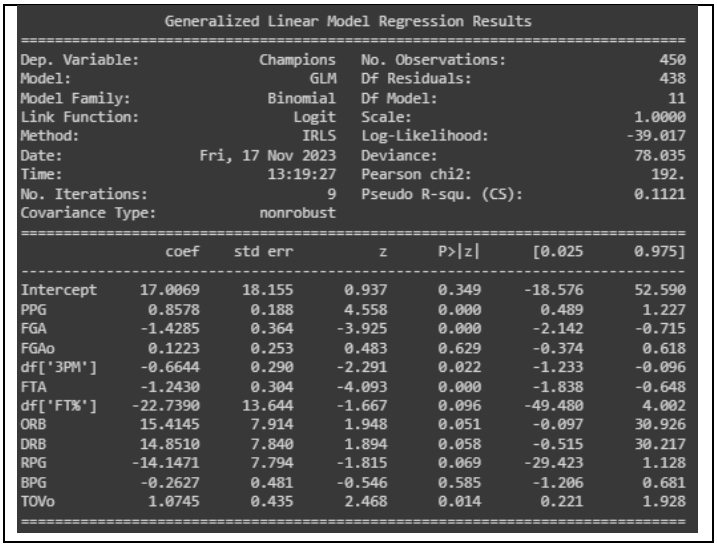
$$\log(\frac{p_i}{1-p_i}) = -34.2105 - 0.5395\ TOV + 95.1548\ FG\% + 0.4515\ DRB - 0.1556\ PPG$$

Η ερμηνεία των συντελεστών αυτού του μοντέλου γίνεται με την ίδια λογική που ακολουθήθηκε και προηγουμένως.

Όπως είναι αντιληπτό από την ανάλυση που έγινε, θετικό ρόλο στην εισαγωγή μίας ομάδας στην φάση των playoffs παίζουν οι αυξήσεις των τιμών των μεταβλητών FG% και DRB. Ακόμα η αύξηση στους δείκτες των λαθών και των πόντων της ομάδας θα έχει αρνητικό αντίκτυπο στην προσπάθεια τους να περάσουν στην δεύτερη φάση της διοργάνωσης.

5.4.2.2. Για την κανονική περίοδο βάσει της μεταβλητής Champions

Έπειτα από την εκτέλεση της διαδικασίας step, στο στατιστικό πρόγραμμα της R, καταλήγουμε πως οι σημαντικότερες μεταβλητές για την περιγραφή αυτού του γραμμικού μοντέλου είναι: πόντοι που πετυχαίνει η ομάδα (PPG), προσπάθειες δίποντων ή τρίποντων (FGA) και των αντιπάλων (FGAo), επιτυχημένες προσπάθειες τρίποντων (3PM), συνολικές προσπάθειες ελεύθερων βολών (FTA) και ποσοστό επιτυχημένων (FT%), επιθετικά (ORB) και αμυντικά ριμπάουντ (DRB), συνολικά ριμπάουντ (RPG), συνολικά κοψίματα (BPG) και λάθη των αντιπάλων (TOVo). Αφού εκτελέστηκε η λογιστική παλινδρόμηση, Σχήμα 5.10, οι μεταβλητές: προσπάθειες δίποντων ή τρίποντων των αντιπάλων (FGAo), ποσοστό επιτυχημένων ελεύθερων βολών (FT%), επιθετικά (ORB) και αμυντικά ριμπάουντ (DRB), συνολικά ριμπάουντ (RPG) και συνολικά κοψίματα (BPG) κρίνονται στατιστικά μη σημαντικές, σύμφωνα με τον έλεγχο του Wald, και γι' αυτόν τον λόγο θα πρέπει να αφαιρεθούν καθεμιά ξεχωριστά για να βρούμε το καταλληλότερο μοντέλο που αναζητούμε.



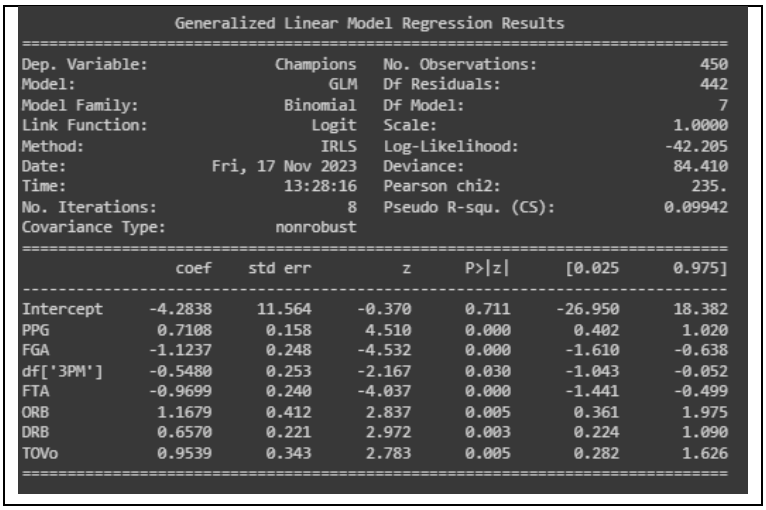
Generalized Linear Model Regression Results

Dep. Variable:	Champions	No. Observations:	450
Model:	GLM	Df Residuals:	438
Model Family:	Binomial	Df Model:	11
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-39.017
Date:	Fri, 17 Nov 2023	Deviance:	78.035
Time:	13:19:27	Pearson chi2:	192.
No. Iterations:	9	Pseudo R-squ. (CS):	0.1121
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	17.0069	18.155	0.937	0.349	-18.576	52.590
PPG	0.8578	0.188	4.558	0.000	0.489	1.227
FGA	-1.4285	0.364	-3.925	0.000	-2.142	-0.715
FGAo	0.1223	0.253	0.483	0.629	-0.374	0.618
df['3PM']	-0.6644	0.290	-2.291	0.022	-1.233	-0.096
FTA	-1.2430	0.304	-4.093	0.000	-1.838	-0.648
df['FT%']	-22.7390	13.644	-1.667	0.096	-49.480	4.002
ORB	15.4145	7.914	1.948	0.051	-0.097	30.926
DRB	14.8510	7.840	1.894	0.058	-0.515	30.217
RPG	-14.1471	7.794	-1.815	0.069	-29.423	1.128
BPG	-0.2627	0.481	-0.546	0.585	-1.206	0.681
TOVo	1.0745	0.435	2.468	0.014	0.221	1.928

Σχήμα: 5.10: Output για το προσαρμοσμένο μοντέλο της μεταβλητής champions για την κανονική περίοδο (χωρίς DRt,ORt)

Αφαιρώντας καθεμιά ξεχωριστά, ξεκινώντας απ' αυτή που είχε το μεγαλύτερο p_{value}, συμπεραίνουμε πως το καταλληλότερο μοντέλο βάσει του ελέγχου του Wald είναι αυτό που περιέχει τις μεταβλητές: πόντοι που πετυχαίνει η ομάδα (PPG), προσπάθειες δίποντων ή τρίποντων (FGA), επιτυχημένες προσπάθειες τρίποντων (3PM), συνολικές προσπάθειες ελεύθερων βολών (FTA), επιθετικά (ORB) και αμυντικά ριμπάουντ (DRB) και λάθη των αντιπάλων (TOVo), όπως φαίνεται στο Σχήμα 5.11.



Generalized Linear Model Regression Results

Dep. Variable:	Champions	No. Observations:	450
Model:	GLM	Df Residuals:	442
Model Family:	Binomial	Df Model:	7
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-42.205
Date:	Fri, 17 Nov 2023	Deviance:	84.410
Time:	13:28:16	Pearson chi2:	235.
No. Iterations:	8	Pseudo R-squ. (CS):	0.09942
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-4.2838	11.564	-0.370	0.711	-26.950	18.382
PPG	0.7108	0.158	4.510	0.000	0.402	1.020
FGA	-1.1237	0.248	-4.532	0.000	-1.610	-0.638
df['3PM']	-0.5480	0.253	-2.167	0.030	-1.043	-0.052
FTA	-0.9699	0.240	-4.037	0.000	-1.441	-0.499
ORB	1.1679	0.412	2.837	0.005	0.361	1.975
DRB	0.6570	0.221	2.972	0.003	0.224	1.090
TOVo	0.9539	0.343	2.783	0.005	0.282	1.626

Σχήμα: 5.10: Output για το νέο προσαρμοσμένο μοντέλο της μεταβλητής champions για την κανονική περίοδο (χωρίς DRt,ORt)

Σύμφωνα όμως με τον έλεγχο απόκλισης θα πρέπει να αφαιρεθεί από το μοντέλο μας η μεταβλητή των επιτυχημένων τρίποντων εφόσον έχει $p_{value} = 0.355$, όπως φαίνεται στον Πίνακα 5.18.

Μεταβλητές	p_{value}
PPG	**0.01601**
FGA	**0.00295**
3PM	**0.35551**

FTA	**0.00011**
ORB	**0.04987**
DRB	**0.038187**
TOVo	**0.003376**

Πίνακας 5.18: Τιμές των p_{value} για τον έλεγχο σημαντικότητας μεταβλητών βάσει της απόκλισης του νέου προσαρμοσμένο μοντέλο της μεταβλητής champions για την κανονική περίοδο (χωρίς DRt,ORt)

Εφόσον αφαιρέθηκε η μεταβλητή των επιτυχημένων τρίποντων (3PM) και υπολογίσαμε το νέο προσαρμοσμένο μοντέλο, ο έλεγχος της απόκλισης έδειξε πως θα έπρεπε να αφαιρεθεί και η μεταβλητή των αμυντικών ριμπάουντ (DRB), καθώς είχε p_{value} = 0.115. Έτσι καταλήξαμε πως το μοντέλο που περιγράφει καλύτερα την μεταβλητή Champions είναι αυτό που περιέχει τις μεταβλητές: πόντοι που πετυχαίνει η ομάδα (PPG), προσπάθειες δίποντων ή τρίποντων (FGA), συνολικές προσπάθειες ελεύθερων βολών (FTA), επιθετικά (ORB) και λάθη των αντιπάλων (TOVo), Σχήμα 5.11.

Generalized Linear Model Regression Results

Dep. Variable:	Champions	No. Observations:	450
Model:	GLM	Df Residuals:	444
Model Family:	Binomial	Df Model:	5
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-49.026
Date:	Fri, 17 Nov 2023	Deviance:	98.052
Time:	13:42:10	Pearson chi2:	314.
No. Iterations:	8	Pseudo R-squ. (CS):	0.07170
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	12.5078	9.803	1.276	0.202	-6.706	31.722
PPG	0.5147	0.112	4.584	0.000	0.295	0.735
FGA	-0.8610	0.203	-4.235	0.000	-1.260	-0.463
FTA	-0.7022	0.196	-3.579	0.000	-1.087	-0.318
ORB	0.8836	0.363	2.436	0.015	0.173	1.595
TOVo	0.6271	0.273	2.293	0.022	0.091	1.163

Σχήμα: 5.11: Output για το τελικό προσαρμοσμένο μοντέλο της μεταβλητής champions για την κανονική περίοδο (χωρίς DRt,ORt)

Για τον έλεγχο της πολυσυγγραμικότητας, όπως φαίνεται στον Πίνακα 5.19, οι τιμές του δείκτη VIF είναι αποδεκτές. Για τον έλεγχο ολικής επάρκειας του μοντέλου, σύμφωνα με τον Πίνακα 5.20, έχουμε ενδείξεις μη απόρριψης της μηδενικής υπόθεσης, εφόσον το p_{value} είναι ίσο με 0.6248. Επομένως, το μοντέλο που αναλύουμε κρίνεται ως καταλληλότερο.

Μεταβλητές	Τιμές VIF
PPG	**7.451**
FGA	**8.131**
FTA	**2.154**
ORB	**2.445**
TOVo	**1.235**

Πίνακας 5.19: Τιμές του δείκτη VIF για το τελικό προσαρμοσμένο μοντέλο της μεταβλητής champions για την κανονική περίοδο (χωρίς DRt,ORt)

Έλεγχος	p_{value}
Hosmer-Lemeshow	**0.6248**

Πίνακας 5.20: Output για τον έλεγχο Hosmer – Lemeshow για το νέο προσαρμοσμένο μοντέλο της μεταβλητής champions για την κανονική περίοδο (χωρίς DRt,ORt)

Τέλος, το μοντέλο έχει τύπο:

$$\log(\frac{p_i}{1-p_i}) = - 34.2105 + 0.5147\ PPG - 0.8610\ FGA - 0.7022\ FTA + 0.8836\ ORB + 0.6271\ TOVo$$

Η ανάλυση που μόλις πραγματοποιήθηκε έχει ως σκοπό να ανακαλύψει τις μεταβλητές που επηρεάζουν περισσότερο μία ομάδα που συμμετέχει στην κανονική περίοδο ώστε να στεφθεί πρωταθλήτρια. Όπως φάνηκε, η αύξηση στις τιμές των δεικτών: πόντοι της ομάδας, επιθετικά ριμπάουντ και λάθη των αντιπάλων έχει θετική επιρροή στο να αναδειχθεί η ομάδα πρωταθλήτρια. Αντίθετα για τις τιμές των μεταβλητών: συνολικές προσπάθειες δίποντων ή τρίποντων και ελεύθερων βολών παρατηρούμε ότι όσο αυξάνονται αυτές η πιθανότητα της ομάδας να στεφθεί πρωταθλήτρια μειώνεται.

5.4.2.2. Για τα Playoffs βάσει της μεταβλητής Champions

Ακολουθώντας την ίδια διεργασία που εφαρμόστηκε παραπάνω και μέσω της συνάρτησης step συμπεραίνουμε πως το καταλληλότερο μοντέλο είναι αυτό που περιέχει τις μεταβλητές: πόντοι που πετυχαίνει η ομάδα (PPG), επιθετικά ριμπάουντ (ORB), προσπάθειες ελεύθερων βολών (FTA), προσπάθειες δίποντων ή τρίποντων (FGA), λάθη αντίπαλης ομάδας (TOVo) και ποσοστά επιτυχημένων τρίποντων (3P%). Έπειτα από την εκτέλεση της λογιστικής παλινδρόμησης, Σχήμα 5.12, η μεταβλητή των ποσοστών επιτυχημένων τρίποντων κρίνεται στατιστικά μη σημαντική, σύμφωνα με τον έλεγχο του Wald, και γι' αυτόν τον λόγο θα πρέπει να αφαιρεθεί από το γραμμικό μοντέλο.

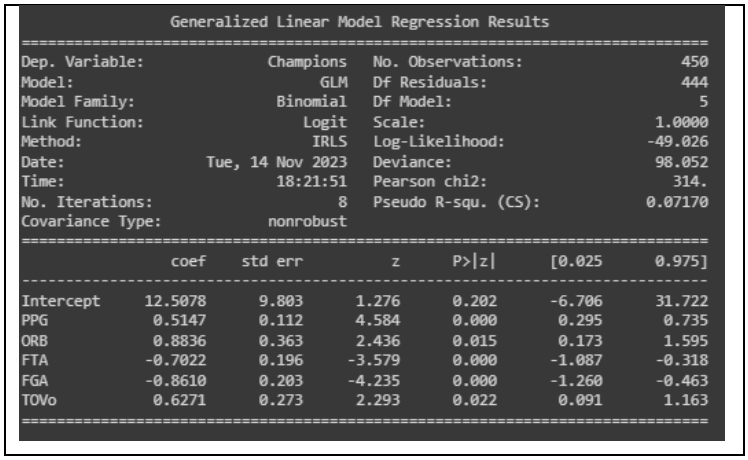
Generalized Linear Model Regression Results

Dep. Variable:	Champions	No. Observations:	450
Model:	GLM	Df Residuals:	443
Model Family:	Binomial	Df Model:	6
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-48.987
Date:	Tue, 14 Nov 2023	Deviance:	97.975
Time:	18:19:51	Pearson chi2:	325.
No. Iterations:	8	Pseudo R-squ. (CS):	0.07186
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	8.9875	15.976	0.563	0.574	-22.326	40.301
PPG	0.4943	0.133	3.703	0.000	0.233	0.756
ORB	0.8595	0.373	2.306	0.021	0.129	1.590
FTA	-0.6726	0.222	-3.030	0.002	-1.108	-0.238
FGA	-0.8252	0.239	-3.449	0.001	-1.294	-0.356
TOVo	0.6323	0.274	2.304	0.021	0.094	1.170
df['3P%']	5.9362	21.267	0.279	0.780	-35.746	47.618

Σχήμα: 5.12: Output για το νέο προσαρμοσμένο μοντέλο της μεταβλητής champions για τα playoffs (χωρίς DRt,ORt)

Έπειτα από την εκτέλεση του νέου προσαρμοσμένου μοντέλου όλες οι μεταβλητές κρίνονται στατιστικά σημαντικές σύμφωνα με τον έλεγχο του Wald, εφόσον όλες οι τιμές λαμβάνουν τιμές μικρότερες του 5%, Σχήμα 5.13.



Generalized Linear Model Regression Results

Dep. Variable:	Champions	No. Observations:	450
Model:	GLM	Df Residuals:	444
Model Family:	Binomial	Df Model:	5
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-49.026
Date:	Tue, 14 Nov 2023	Deviance:	98.052
Time:	18:21:51	Pearson chi2:	314.
No. Iterations:	8	Pseudo R-squ. (CS):	0.07170
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	12.5078	9.803	1.276	0.202	-6.706	31.722
PPG	0.5147	0.112	4.584	0.000	0.295	0.735
ORB	0.8836	0.363	2.436	0.015	0.173	1.595
FTA	-0.7022	0.196	-3.579	0.000	-1.087	-0.318
FGA	-0.8610	0.203	-4.235	0.000	-1.260	-0.463
TOVo	0.6271	0.273	2.293	0.022	0.091	1.163

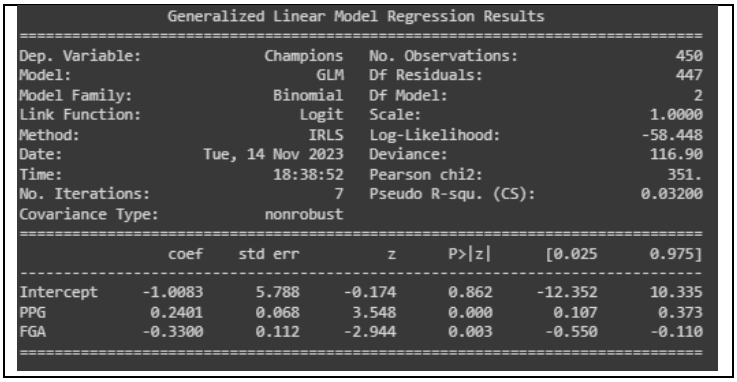
Σχήμα: 5.13: Output για το δεύτερο προσαρμοσμένο μοντέλο της μεταβλητής champions για τα playoffs (χωρίς DRt,ORt)

Για τον έλεγχο της απόκλισης του μοντέλου, Πίνακας 5.21, διακρίνουμε πως οι μεταβλητή ORB και FTA έχουν p_{value} μεγαλύτερο του επιπέδου σημαντικότητας και επομένως θα πρέπει να αφαιρεθούν κάθε μια ξεχωριστά από την ανάλυση. Εφόσον αφαιρέθηκαν, παρατηρήθηκε ότι καταλληλότερο μοντέλο είναι αυτό που περιέχει μόνο τις μεταβλητές PPG και FGA, καθώς ο έλεγχος απόκλισης έπειτα από την αφαίρεση των δύο πρώτων μεταβλητών έδειξε πως και η μεταβλητή TOVo είναι στατιστικά μη σημαντική (p_{value} = 0.099163).

Μεταβλητές	p_{value}
PPG	**0.007196**
ORB	**0.795325**
FTA	**0.361828**
FGA	**3.247e-05**
TOVo	**0.006124**

Πίνακας 5.21: Τιμές των p_{value} για τον έλεγχο σημαντικότητας μεταβλητών βάσει της απόκλισης του προσαρμοσμένο μοντέλο της μεταβλητής champions για τα playoffs (χωρίς DRt,ORt)

Ακόμα, ο αλγόριθμος μεγιστοποίησε το λογάριθμο της πιθανοφάνειας μετά από επτά (7) επαναλήψεις σε αντίθεση με το προηγούμενο μοντέλο που χρειαζόταν οκτώ (8).



Generalized Linear Model Regression Results

Dep. Variable:	Champions	No. Observations:	450
Model:	GLM	Df Residuals:	447
Model Family:	Binomial	Df Model:	2
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-58.448
Date:	Tue, 14 Nov 2023	Deviance:	116.90
Time:	18:38:52	Pearson chi2:	351.
No. Iterations:	7	Pseudo R-squ. (CS):	0.03200
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.0083	5.788	-0.174	0.862	-12.352	10.335
PPG	0.2401	0.068	3.548	0.000	0.107	0.373
FGA	-0.3300	0.112	-2.944	0.003	-0.550	-0.110

Σχήμα: 5.14: Output για το τρίτο προσαρμοσμένο μοντέλο της μεταβλητής champions για τα playoffs (χωρίς DRt,ORt)

Όπως είναι φανερό στο Σχήμα 5.8, τα p_{value} του ελέγχου Wald για τις μεταβλητές που χρησιμοποιήθηκαν είναι μικρότερα του επιπέδου σημαντικότητας. Επίσης, για τον έλεγχο της απόκλισης, Πίνακας 5.22, διακρίνουμε πως οι τιμές είναι όλες μικρότερες του 5%.

Μεταβλητές	p_{value}
PPG	**0.007196**
FGA	**0.020850**

Πίνακας 5.22: Τιμές των p_{value} για τον έλεγχο σημαντικότητας μεταβλητών βάσει της απόκλισης του προσαρμοσμένο μοντέλο της μεταβλητής champions για τα playoffs (χωρίς DRt,ORt)

Για τον έλεγχο της πολυσυγγραμικότητας, όπως φαίνεται στον Πίνακα 5.23, οι τιμές του δείκτη VIF είναι αρκετά μικρές και επομένως δεν διακρίνεται κάποιο πρόβλημα όπως και για τον δείκτη Hosmer-Lemeshow που έχει τιμή ίση με 0.4844.

Μεταβλητές	Τιμές VIF
PPG	**2.051**
FGA	**2.051**

Πίνακας 5.23: Τιμές του δείκτη VIF για το νέο προσαρμοσμένο μοντέλο της μεταβλητής champions για τα playoffs (χωρίς DRt,ORt)

Έλεγχος	p_{value}
Hosmer-Lemeshow	**0.4844**

Πίνακας 5.24: Output για τον έλεγχο Hosmer – Lemeshow για το νέο προσαρμοσμένο μοντέλο της μεταβλητής champions για τα playoffs (χωρίς DRt,ORt)

Τέλος, το μοντέλο έχει τύπο:

$$\log(\frac{p_i}{1-p_i}) = 1.0083 + 0.4401\ \text{PPG} - 0.3300\ \text{FGA}$$

Για την ερμηνεία των συντελεστών αυτού του μοντέλου προκύπτουν τα παρακάτω συμπεράσματα:

- Αν αυξήσουμε την τιμή της μεταβλητής PPG, ενώ διατηρούμε σταθερές τις τιμές των άλλων μεταβλητών, ο λογάριθμος της σχετικής πιθανότητας επιτυχίας θα αυξηθεί κατά 0.4401 μονάδες.
- Αν αυξήσουμε την τιμή της μεταβλητής FGA, ενώ διατηρούμε σταθερές τις τιμές των άλλων μεταβλητών, ο λογάριθμος της σχετικής πιθανότητας επιτυχίας θα μειωθεί κατά 0.3300 μονάδες.

Όπως φαίνεται στην ανάλυση που έγινε, αν η ομάδα συμμετέχει στην φάση των playoffs, τότε οι μεταβλητές που παίζουν σημαντικότερο ρόλο για να αναδειχθεί πρωταθλήτρια είναι οι πόντοι της ομάδας και οι συνολικές προσπάθειες δίποντων ή τρίποντων. Η αύξηση στις τιμές των δεικτών των πόντων της ομάδας έχει θετική επιρροή στο να στεφθεί η ομάδα πρωταθλήτρια. Αντίθετα για τις τιμές της μεταβλητής των συνολικών προσπαθειών δίποντων ή τρίποντων η πιθανότητα κατάκτησης του πρωταθλήματος μειώνεται.

ΚΕΦΑΛΑΙΟ 6ο

6. Χρονοσειρές

Στο συγκεκριμένο κεφάλαιο της εργασίας θα γίνει ανάλυση ορισμένων μεταβλητών των δεδομένων με τη μορφή χρονοσειρών. Θα ελεγχθούν οι χρονοσειρές ως προς τη στασιμότητά τους, την κανονικότητα και την ανεξαρτησία των δεδομένων μέσω των χρονοσειρών ARIMA. Ακόμα, θα γίνουν οι προβλέψεις για τις επόμενες δύο (2) σεζόν των μεταβλητών και θα παρουσιαστούν και τα ανάλογα γραφήματα.

6.1. Θεωρητικό Υπόβαθρο

Ένα σύνολο παρατηρήσεων που εκφράζουν την χρονική εξέλιξη κάποιου χαρακτηριστικού ενός στοχαστικού φαινομένου αποτελεί μια χρονοσειρά. Οι παρατηρήσεις του συνόλου καταγράφονται ακολουθιακά στη διάρκεια του χρόνου και συνήθως ισαπέχουν χρονικά. Επειδή η χρονική εξέλιξη της τιμής του χαρακτηριστικού που αναλύεται γίνεται με τυχαίο τρόπο, μπορεί να περιγράφει από μια οικογένεια τυχαίων μεταβλητών X_t , t = 1, 2, …T. Στόχος της ανάλυσης χρονοσειρών είναι να χρησιμοποιήσει τις παρατηρηθείσες τιμές ώστε να κατασκευάσει ένα στοχαστικό πρότυπο μέσω του οποίου θα μπορέσει να κάνει προβλέψεις για μελλοντικές τιμές του χαρακτηριστικού που αναλύεται κάθε φορά. Αρχικά, για να είναι εφικτό να αναλυθεί μια χρονοσειρά, θα πρέπει να ισχύουν για αυτή οι προϋποθέσεις της στασιμότητας (Τριανταφύλλου, 2023).

6.1.1. Στασιμότητα χρονοσειράς

Η μη-στασιμότητα αποτελεί σοβαρό πρόβλημα στην ανάλυση χρονοσειρών -και ιδιαίτερα όταν προσπαθούμε να κάνουμε προβλέψεις- εφόσον οδηγούμαστε σε λανθασμένα αποτελέσματα. Μια χρονοσειρά $\{X_t, t = 1, 2, \ldots\}$ ονομάζεται στάσιμη, αν ισχύουν τα ακόλουθα τρία ζητούμενα:

- $E(X_t) = \mu$, για κάθε t = 1, 2, …
- $V(X_t) = \sigma^2$, για κάθε t = 1, 2, …
- $Cov(X_{t_s}, X_{t_u}) = \gamma_{t_s - t_u}$

όπου η συνάρτηση $\gamma_j = \gamma_{t_s - t_u}$ είναι γνωστή ως συνάρτηση αυτοσυνδιακύμανσης με υστέρηση ίση με j. Η αυτοσυνδιακύμανση της χρονοσειράς εκφράζει την συνδιακύμανση μεταξύ των τιμών της χρονοσειράς που αντιστοιχούν σε δύο χρονικές στιγμές οι οποίες απέχουν μεταξύ τους απόσταση j. Η συνάρτηση αυτή λαμβάνει τιμές στο διάστημα $(-\infty, +\infty)$, ενώ επηρεάζεται και από τη μονάδα μέτρησης του υπό μελέτη χαρακτηριστικού.

Όπως έχει ήδη αναφερθεί, βασική προϋπόθεση για την ανάλυση μια χρονοσειράς είναι η ύπαρξη στασιμότητας. Ωστόσο στην πράξη οι χρονοσειρές δεν είναι πάντοτε στάσιμες και σε τέτοιες περιπτώσεις απαιτείται ο μετασχηματισμός τους σε στάσιμες εξομαλύνοντας τις παραβιάσεις. Οι παραβιάσεις της στασιμότητας των χρονοσειρών

χωρίζονται σε δύο μεγάλες κατηγορίες, τις χρονοσειρές με τάση και τις χρονοσειρές που παρουσιάζουν τάση και εποχικότητα.

6.1.1.1. Χρονοσειρές με τάση

Στην πρώτη κατηγορία παραβιάσεων μιας χρονοσειράς είναι οι χρονοσειρές που παρουσιάζουν τάση. Οι συγκεκριμένες χρονοσειρές περιγράφονται από τον τύπο:

$$X_t = m_t + Y_t \text{ , } t = 1, 2, \ldots, T$$

όπου m_t είναι μια συνάρτηση που εκφράζει την τάση της χρονοσειράς, η οποία αποτελεί την παραβίαση και θα πρέπει να αφαιρεθεί, και Y_t είναι μια στάσιμη χρονοσειρά με μέση τιμή 0 και σταθερή διασπορά σ^2. Στόχος των μεθόδων που παρουσιάζονται παρακάτω είναι η εύρεση της συνάρτησης της τάσης και η επιτυχής αφαίρεση της, ώστε να δημιουργηθεί μία στάσιμη χρονοσειρά που θα είναι εφικτό να αναλυθεί. Για την αντιμετώπιση της τάσης μίας χρονοσειράς παρουσιάζονται τρείς διαφορετικοί τρόποι.

Μέθοδος γραμμικής τάσης

Στην πρώτη μέθοδο εύρεσης και αφαίρεσης της τάσης βασική υπόθεση είναι πως η τάση που περιέχεται στη χρονοσειρά που αναλύουμε είναι γραμμική, δηλαδή είναι της μορφής:

$$m_t = \beta_0 + \beta_1 t, \; t = 1, 2, \ldots, T$$

όπου β_0, β_1 είναι άγνωστες παράμετροι. Για να βρεθεί το καλύτερο μοντέλο που περιγράφει τη χρονοσειρά, πρέπει να ελαχιστοποιηθεί το άθροισμα των τετραγώνων των αποστάσεων των παρατηρηθέντων σημείων της χρονοσειράς από την προκύπτουσα ευθεία $\widehat{m}_t = \hat{\beta}_0 + \hat{\beta}_1 t$. Δηλαδή οι εκτιμήτριες ελαχίστων τετραγώνων, $\hat{\beta}_0$ και $\hat{\beta}_1$, να ελαχιστοποιούν το άθροισμα $SS = \sum_{t=1}^{T}(X_t - \beta_0 - \beta_1 t)^2$. Χρησιμοποιώντας τον τύπο αυτόν αποδεικνύεται πως οι εκτιμήτριες βρίσκονται από τους παρακάτω τύπους:

$$\hat{\beta}_1 = \frac{\sum_{t=1}^{T}(t-\bar{t})(X_t-\bar{X})}{\sum_{t=1}^{T}(t-\bar{t})^2} \text{ και } \hat{\beta}_0 = \bar{X} - \hat{\beta}_1 t$$

όπου $\bar{X}$, $\bar{t}$ είναι οι μέσες τιμές των παρατηρήσεων της χρονοσειράς και των τιμών t = 1, 2, … , T αντίστοιχα. Τέλος, πραγματοποιείται η εξάλειψη της τάσης αφαιρώντας της εκτιμώμενη τάση και λαμβάνοντας μια καινούργια στάσιμη χρονοσειρά, $\widehat{Y_t}$, με τύπο:

$$\widehat{Y_t} = X_t - (\hat{\beta}_0 + \hat{\beta}_1 t).$$

Μέθοδος πολυωνυμικής τάσης

Σε αυτή τη μέθοδο βασική υπόθεση αποτελεί το γεγονός ότι η τάση είναι πολυωνυμικής μορφής, δηλαδή βασίζεται στον τύπο:

$$m_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \ldots + \beta_k t^k \text{ , } t = 1, 2, \ldots, T$$

όπου β_0, β_1, β_2 είναι οι άγνωστες παράμετροι που είναι αναγκαίο να εκτιμηθούν ώστε να γίνει η αφαίρεση της τάσης. Ακολουθώντας τη διεργασία που αναλύθηκε και

προηγουμένως (μέθοδος γραμμικής τάσης) προσδιορίζονται και εδώ οι εκτιμήτριες ελαχίστων τετραγώνων $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$. Στόχος είναι η εύρεση αυτών των εκτιμητριών ελάχιστων τετραγώνων και έπειτα η εξάλειψη της τάσης, λαμβάνοντας την ακόλουθη στάσιμη χρονοσειρά:

$$\hat{Y}_t = X_t - (\hat{\beta}_0 + \hat{\beta}_1 t + \hat{\beta}_2 t^2 + \dots + \hat{\beta}_k t^k),\ t = 1, 2, \dots, T$$

Μέθοδος διαφορών

Σε αντίθεση με τις προηγούμενες μεθόδους αυτή η μέθοδος δεν εκτιμάει την τάση αλλά την αφαιρεί από την υπό μελέτη χρονοσειρά και τελικώς προκύπτει μία νέα στάσιμη χρονοσειρά. Για να προχωρήσουμε στην αφαίρεση της τάσης από την χρονοσειρά θα πρέπει να υποθέσουμε την μορφή της τάσης που παρουσιάζει.

- Αν η τάση είναι γραμμική ($m_t = \beta_0 + \beta_1 t$), τότε η στάσιμη χρονοσειρά που προκύπτει από την εφαρμογή του τελεστή διαφορών, $\nabla X_t = X_t - X_{t-1}$, ορίζεται ως εξής:
$$\nabla X_t = \nabla(m_t + Y_t) = \beta_1 + \nabla Y_t$$
- Αν η τάση είναι πολυωνυμική με βαθμό k ($m_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \dots + \beta_k t^k$), τότε η νέα στάσιμη χρονοσειρά που θα προκύψει από την αφαίρεση της τάσης και την εφαρμογή του τελεστή διαφορών δίνεται από τον τύπο:
$$\nabla^k X_t = \nabla^k(\beta_0 + \beta_1 t + \beta_2 t^2 + \dots + \beta_k t^k + Y_t) = k!\beta_k + \nabla^k Y_t$$

6.1.1.2. Χρονοσειρές με τάση και εποχικότητα

Σε αυτή την κατηγορία η υπό μελέτη χρονοσειρά X_t περιγράφεται από τον τύπο:

$$X_t = m_t + s_t + Y_t,\ t = 1, 2, \dots, T$$

όπου Y_t είναι μια στάσιμη χρονοσειρά με μέση τιμή 0 και σταθερή διασπορά σ^2, m_t είναι η συνάρτηση της τάσης και s_t αποτελεί την συνάρτηση που εκφράζει την εποχικότητα της χρονοσειράς. Σκοπός της μεθόδου που θα αναλυθεί παρακάτω είναι η εξάλειψη της τάσης και της εποχικότητας χωρίς την εκτίμηση τους.

Μέθοδος διαφορών

Έστω πως η περίοδος της εποχικότητας της χρονοσειράς είναι τ, δηλαδή $s_t = s_{t+\tau}$, τότε προσδιορίζουμε τις διαφορές τάξης τ ως εξής:

$$\nabla_\tau X_t = X_t - X_{t-\tau} = m_t + s_t + Y_t - m_{t+\tau} - s_{t+\tau} - Y_{t+\tau} = (m_t - m_{t+\tau}) + (Y_t - Y_{t+\tau}).$$

Συνεπώς δημιουργήθηκε μια χρονοσειρά η οποία περιλαμβάνει την τάση, με μορφή $m_t - m_{t+\tau}$, και μία στάσιμη χρονοσειρά $\nabla_\tau Y_t = Y_t - Y_{t+\tau}$. Επομένως σε αυτή τη φάση αφαιρέθηκε η εποχικότητα που παρουσίαζε η χρονοσειρά. Στη συνέχεια εφαρμόζουμε μια από τις μεθόδους αφαίρεσης της τάσης που αναφέρθηκαν στην προηγούμενη παράγραφο με στόχο να φτάσουμε σε μία στάσιμη χρονοσειρά.

Όταν μία χρονοσειρά εμφανίζει και τάση και εποχικότητα, προτιμάται να διενεργείται η μέθοδος διαφορών που παρουσιάστηκε, ώστε να εξαλειφθεί η εποχικότητα της χρονοσειράς, και έπειτα εκτελείται η μέθοδος διαφορών για την

εξάλειψη της τάσης. Προφανώς για να εξασφαλιστεί η στασιμότητα μίας χρονοσειράς θα πρέπει να γίνει και ο κατάλληλος έλεγχος μοναδιαίας ρίζας.

6.1.1.3. Έλεγχος μοναδιαίας ρίζας

Ο πιο γνωστός έλεγχος στασιμότητας για μια χρονοσειρά είναι ο έλεγχος που αναπτύχθηκε από τους Dickey και Fuller. Ο συγκεκριμένος έλεγχος εξετάζει την παρουσία μοναδιαίας ρίζας και ονομάζεται επαυξημένος Dickey-Fuller έλεγχος (Augmented Dickey-Fuller test). Πιο αναλυτικά υπάρχουν τρείς (3) διαφορετικές εκδοχές που αντιστοιχούν σε τρία διαφορετικά υποδείγματα.

- $\nabla X_t = \delta\, X_{t-1} + \sum_{i=1}^{k} \nabla X_{t-i} + W_t$
- $\nabla X_t = \mu + \delta\, X_{t-1} + \sum_{i=1}^{k} \gamma_i \nabla X_{t-i} + W_t$
- $\nabla X_t = \mu + \beta\, t + \delta\, X_{t-1} + \sum_{i=1}^{k} \gamma_i \nabla X_{t-i} + W_t$

Η μηδενική και η εναλλακτική υπόθεση γι' αυτόν τον έλεγχο ορίζεται ως εξής:

Μηδενική υπόθεση H_0: $\delta = 0$

vs

Εναλλακτική υπόθεση H_1 : $\delta < 0$

Προφανώς όταν υπάρχουν ενδείξεις μη απόρριψης της μηδενικής υπόθεσης ($p_{value} > 0.05$) τότε υπάρχει μοναδιαία ρίζα στην υπό μελέτη χρονοσειρά και συνεπώς δεν χαρακτηρίζεται στάσιμη, οπότε θα πρέπει να ακολουθηθούν άλλες μέθοδοι εξομάλυνσης των παραβιάσεων της.

6.1.2. Επιλογή μοντέλου ερμηνείας και χρήσιμοι έλεγχοι

Εφόσον έχουν υλοποιηθεί τα προηγούμενα βήματα και πλέον υπάρχει μία στάσιμη χρονοσειρά είναι πλέον εφικτό να γίνει η ανάλυση της. Αρχικά, μέσω των διαγραμμάτων αυτοσυσχέτισης (ACF) και μερικής αυτοσυσχέτισης (PACF) υπάρχει η δυνατότητα επιλογής του καταλληλότερου ARIMA(p,d,q)$(P,D,Q)_s$. Κάθε αναλυτής μπορεί να επιλέξει διαφορετικό μοντέλο ερμηνείας για την εκάστοτε χρονοσειρά, επομένως δημιουργείται η ανάγκη ελέγχου της επιλογής. Αν υπάρχουν περισσότερα από ένα υποψήφια υποδείγματα ARIMA που μπορούν να περιγράψουν την χρονοσειρά, τότε χρησιμοποιούνται οι δείκτες AIC, BIC και AICc για την επιλογή του βέλτιστού. Η επιλογή του βέλτιστου γίνεται με βάση τη μικρότερη τιμή των δεικτών. Εφόσον επιλεγεί το μοντέλο με την μικρότερη τιμή των δεικτών, στη συνέχεια ελέγχεται και η καλή προσαρμογή του μέσω διαφόρων ελέγχων για τα κατάλοιπα, ώστε να ελεγχθεί η εγκυρότητα της επιλογής που έγινε.

Τα μοντέλα ARIMA(p,d,q)$(P,D,Q)_s$ είναι αυτά που χρησιμοποιούνται συνήθως για παρόμοιες αναλύσεις. Το μέρος $(P,D,Q)_s$ αναφέρεται στο εποχικό μέρος της χρονοσειράς. Προφανώς είναι πιθανό να υπάρξουν και χρονοσειρές που δεν έχουν εποχικό μέρος, οπότε αυτές οι παράμετροι αντιστοιχούν όλες σε μηδέν (0). Στην περίπτωση ύπαρξης εποχικού μέρους η εξίσωση που περιγράφει το μοντέλο είναι:

$$\Phi_P(B^s)X_t = \Theta_Q(B^s)W_t$$

όπου $\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \ldots - \Phi_P B^{Ps}$, $\Theta_Q(B^s) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} - \ldots - \Theta_Q B^{Qs}$.

Στην περίπτωση μη ύπαρξης εποχικού μέρους στην χρονοσειρά το μοντέλο παίρνει την μορφή ARIMA(p,d,q)$(0,0,0)_0$ και περιγράφεται από την εξίσωση:

$$X_t = \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \ldots + \varphi_p X_{t-p} + W_t + \theta_1 W_{t-1} + \ldots + \theta_q W_{t-q}$$

όπου οι παράμετροι $\varphi_1, \varphi_2, \ldots, \varphi_p \in$ R και $\theta_1, \theta_2, \ldots, \theta_p \in$ R σταθμίζουν τη συμβολή των όρων της υπό μελέτη χρονοσειράς.

Εφόσον επιλεχθούν τα μοντέλα που θεωρείται πως περιγράφουν καλύτερα την χρονοσειρά, θα πρέπει να επιλεχθεί το βέλτιστο μοντέλο και έπειτα να γίνουν οι απαραίτητοι έλεγχοι. Όπως αναφέρθηκε και προηγουμένως, η επιλογή του βέλτιστου μοντέλου γίνεται με τη χρήση των δεικτών AIC, BIC και AICc.

Δείκτης AIC

Το κριτήριο AIC (Akaike's Information Criterion) ορίζεται ως:

$$\text{AIC} = 2k - 2\ln L(\widehat{\varphi_1}, \widehat{\varphi_2}, \ldots, \widehat{\varphi_p}, \widehat{\theta_1}, \ldots, \widehat{\theta_q}, \hat{\mu}, \widehat{\sigma^2})$$

όπου k = p + q + 1 ή k = p + q + 2.

Δείκτης BIC

Το κριτήριο BIC (Bayesian Information Criterion) ορίζεται ως:

$$\text{BIC} = 2\ln(n) - 2\ln L(\widehat{\varphi_1}, \widehat{\varphi_2}, \ldots, \widehat{\varphi_p}, \widehat{\theta_1}, \ldots, \widehat{\theta_q}, \hat{\mu}, \widehat{\sigma^2}).$$

Συνήθως το μοντέλο που επιλέγεται με αυτό το κριτήριο είναι είτε το ίδιο που έχει επιλεχθεί με το κριτήριο AIC είτε με ένα λιγότερο παράγοντα, διότι ο συντελεστής 2ln(n) μειώνει περισσότερο το πλήθος των παραμέτρων που εισέρχονται στο μοντέλο.

Δείκτης AICc

Για μικρά δείγματα, όπως στην παρούσα εργασία που αναφέρεται σε δεκαπέντε (15) σεζόν, προτιμότερο κριτήριο για την εύρεση του βέλτιστου μοντέλου είναι το διορθωμένο κριτήριο AIC (AIC corrected ή AICc). Ο υπολογισμός του κριτηρίου γίνεται μέσω του τύπου:

$$\text{AICc} = \text{AIC} + \frac{2k^2+2k}{n-k-1}$$

καθώς το $n \to \infty$, το AICc $\to$ AIC.

Έλεγχος κανονικότητας

Ένας βασικός έλεγχος που πρέπει να γίνεται σε κάθε στάσιμη χρονοσειρά, όταν έχουμε καταλήξει σε κάποιο μοντέλο, είναι ο έλεγχος κανονικότητας των καταλοίπων. Οι έλεγχοι κανονικότητας, όπως έχει αναφερθεί ήδη σε προηγούμενο κεφάλαιο, μπορούν να γίνουν είτε γραφικά είτε με ένα στατιστικό τεστ. Στο παρόν κεφάλαιο θα

χρησιμοποιηθεί ο έλεγχος Anderson-Darling, ενώ ενδεικτικά μόνο υπολογίζεται και ο Shapiro-Wilk, διότι δεν θεωρείται έμπιστος για τα παραδείγματα χρονοσειρών.

Έλεγχος ανεξαρτησίας

Ο έλεγχος που σχετίζεται με την ανεξαρτησία των καταλοίπων, δηλαδή για μηδενική αυτοσυσχέτιση, είναι ο έλεγχος Box-Ljung. Ο συγκεκριμένος έλεγχος πήρε το όνομα του από τους Greta M. Ljung και George E.P Box. Η μηδενική και η εναλλακτική υπόθεση του ελέγχου είναι:

Μηδενική υπόθεση H_0: Τα δεδομένα κατανέμονται ανεξάρτητα

vs

Εναλλακτική υπόθεση H_1 : Τα δεδομένα δεν κατανέμονται ανεξάρτητα

Ο έλεγχος ανεξαρτησίας των καταλοίπων έχει το προτέρημα πως ελέγχει όλες τις αυτοσυσχετίσεις αν είναι μηδέν και όχι κάθε μία ξεχωριστά. (Ρακιτζής, 2023)

Σε περίπτωση που το μοντέλο που έχει επιλεγεί απέχει από την κανονικότητα και την ανεξαρτησία των καταλοίπων, τότε αυξάνεται κατά μία μονάδα η τάξη του μοντέλου μέχρι να ικανοποιηθούν και οι δύο συνθήκες.

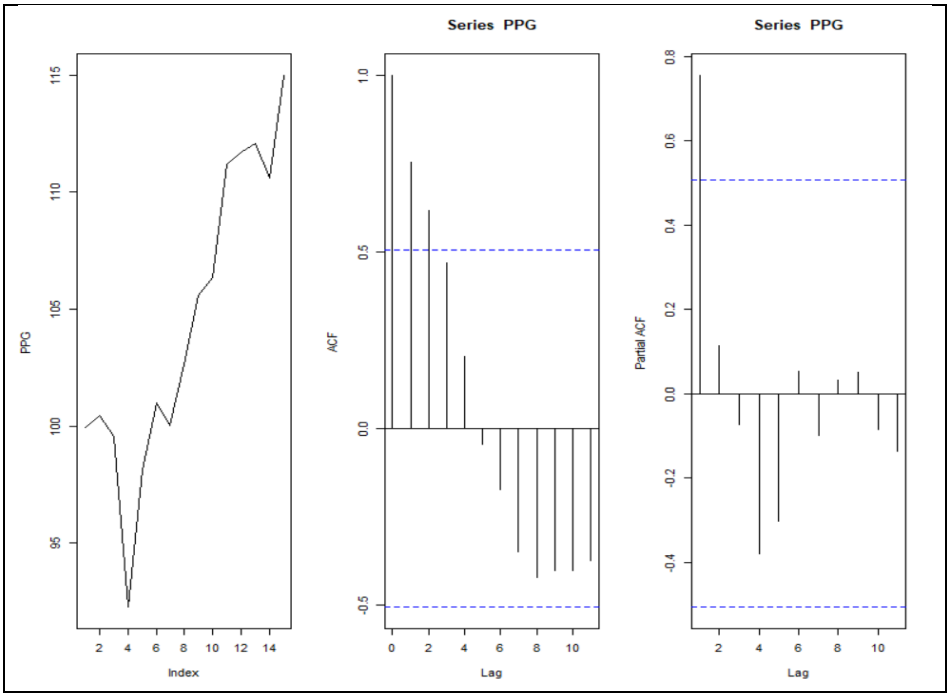
6.2. Προσαρμογή μεθόδων χρονοσειρών

Σε αυτή την ενότητα θα εκτελεστούν όλες οι μέθοδοι που έχουν αναφερθεί παραπάνω, ώστε να γίνει η ανάλυση διαφόρων μεταβλητών, σαν χρονοσειρές, και τελικώς να γίνουν και οι προβλέψεις των επομένων πέντε (5) σεζόν. Οι μεταβλητές που θα χρησιμοποιηθούν προς ανάλυση είναι οι πόντοι ανά αγώνα (PPG), οι ασίστ (APG), τα ριμπάουντ (RPG), τα κλεψίματα (SPG) και τα κοψίματα της ομάδας ανά αγώνα (BPG). Η ανάλυση των χρονοσειρών θα γίνει για τους μέσους όρους ανά χρονιά για τα δεδομένα που συλλέχθηκαν για την κανονική περίοδο της διοργάνωσης. Γενικότερα το πλήθος των παρατηρήσεων των χρονοσειρών μπορεί να ποικίλει, αλλά προτείνεται να είναι τουλάχιστον είκοσι (20) παρατηρήσεις και πολλά μοντέλα απαιτούν το λιγότερο πενήντα (50) παρατηρήσεις για ακριβή εκτίμηση (McCleary et al., 1980, σελ. 20). Προφανώς περισσότερα δεδομένα είναι πάντα προτιμότερα αλλά μια χρονοσειρά θα πρέπει να είναι αρκετά μεγάλη ώστε να καταγράφει τα φαινόμενα που μας ενδιαφέρουν. Στην παρούσα εργασία έχουν καταγραφεί τα δεδομένα των τελευταίων δεκαπέντε (15) σεζόν της διοργάνωσης, δηλαδή έχουμε στην διάθεση μας δεκαπέντε παρατηρήσεις για κάθε μεταβλητή, και συνεπώς ίσως δημιουργηθεί πρόβλημα στην ακρίβεια των προβλέψεων. Παρόλα αυτά θα προχωρήσουμε στην ανάλυση ορισμένων μεταβλητών ώστε να παρουσιαστούν οι τεχνικές που χρησιμοποιούνται σε μοντέλα χρονοσειρών.

6.2.1. Ανάλυση της χρονοσειράς πόντοι ανά αγώνα (PPG)

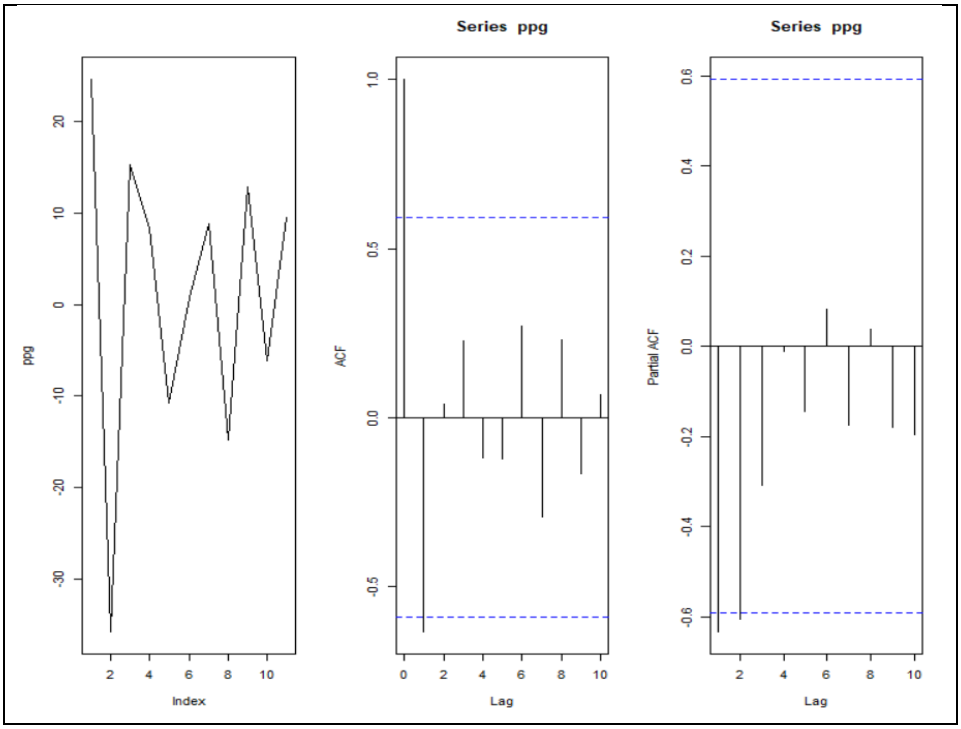
Σε πρώτη φάση, ακολουθώντας το θεωρητικό υπόβαθρο, πρέπει να δημιουργήσουμε τα γραφήματα της χρονοσειράς, ώστε να μπορέσουμε να ελέγξουμε την στασιμότητα της. Στο Σχήμα 6.1 παρουσιάζονται το γράφημα της χρονοσειράς στα αριστερά, το γράφημα της αυτοσυσχέτισης στο κέντρο και το γράφημα της μερικής αυτοσυσχέτισης

στα δεξιά του σχήματος. Όπως φαίνεται στο Σχήμα 6.1, η χρονοσειρά παρουσιάζει τάση η οποία είναι αυξητική, και επομένως δεν μπορεί να θεωρηθεί στάσιμη και έτοιμη προς ανάλυση.



Σχήμα 6.1: Γραφήματα χρονοσειράς πόντων ανά αγώνα

Χρησιμοποιώντας τη μέθοδο των διαφορών για τάξη που ισούται με τέσσερα (4), καταλήξαμε στη νέα στάσιμη χρονοσειρά, η οποία έλαβε το όνομα ppg, που απεικονίζεται στο Σχήμα 6.2. Προφανώς, κάνοντας και τον έλεγχο της μοναδιαίας ρίζας, Πίνακας 6.1 στήλη p_{value} adf test, το p_{value} είναι αρκετά μικρό και επομένως υπάρχουν ενδείξεις απόρριψης της μηδενικής υπόθεσης, δηλαδή μη ύπαρξη μοναδιαίας ρίζας.



Σχήμα 6.2: Γραφήματα της νέας στάσιμης χρονοσειράς πόντων ανά αγώνα (ppg)

Χρονοσειρά	p_{value} adf test	p_{value} Shapiro-Wilk	p_{value} Anderson-Darling	p_{value} Box-Ljung
Ppg	**<0.01**	**0.9964**	**0.9596**	**0.2222**

Πίνακας 6.1: Έλεγχοι της στάσιμης χρονοσειράς ppg

Αφού βρέθηκε η στάσιμη χρονοσειρά που προκύπτει μέσω της μεταβλητής των πόντων ανά αγώνα, θα προχωρήσουμε στην ανάλυσή της. Σε πρώτη φάση θα υπολογίσουμε τις τιμές των δεικτών για κάποια μοντέλα ARIMA(p,d,q), εφόσον δεν υπήρχε εποχικότητα στην χρονοσειρά που έχουμε υπό μελέτη. Όπως φαίνεται στα γραφήματα τα οποία περιέχονται στον Πίνακα 6.2, αυτοσυσχέτισης και μερικής αυτοσυσχέτισης, πιθανά μοντέλα που ερμηνεύουν την χρονοσειρά είναι τα ARIMA(0,0,1), ARIMA(0,0,2), ARIMA(1,0,1) και ARIMA(1,0,2). Αυτές οι επιλογές έγιναν διότι στο γράφημα της αυτοσυσχέτισης παρατηρούμε πως οι δύο πρώτες είναι στατιστικά σημαντικές και στο γράφημα της μερικής αυτοσυσχέτισης συμβαίνει το ίδιο. Σύμφωνα με τους δείκτες που υπολογίστηκαν μικρότερη τιμή, σε όλους τους δείκτες, έχει το μοντέλο ARIMA(0,0,2), και επομένως θα προχωρήσουμε στην ανάλυση με αυτό. Επόμενοι έλεγχοι που θα πρέπει να γίνουν για να εξασφαλιστεί η ορθή επιλογή του μοντέλου είναι ο έλεγχος κανονικότητας και ανεξαρτησίας των καταλοίπων του. Στον Πίνακα 6.3 υπολογίστηκαν οι παράμετροι του μοντέλου που επιλέχτηκε, ενώ στον Πίνακα 6.1 παρουσιάζονται τα p_{value} των ελέγχων κανονικότητας και ανεξαρτησίας. Από τις τιμές των p_{value} έχουμε ενδείξεις μη απόρριψης των μηδενικών υποθέσεων και για τους δύο ελέγχους και επομένως το μοντέλο που επιλέχθηκε είναι κατάλληλο για την περιγραφή της χρονοσειράς.

Χρονοσειρά	ppg	ppg	ppg	ppg

Μοντέλο	ARIMA(0,0,1)	ARIMA(0,0,2)	ARIMA(1,0,1)	ARIMA(1,0,2)
AIC	88.49468	83.6731	86.49722	84.40914
BIC	89.68836	85.26468	88.0888	86.3982
AICc	91.92325	90.33977	93.16389	96.40914

Πίνακας 6.2: Δείκτες AIC, BIC και AICc για την χρονοσειρά ppg

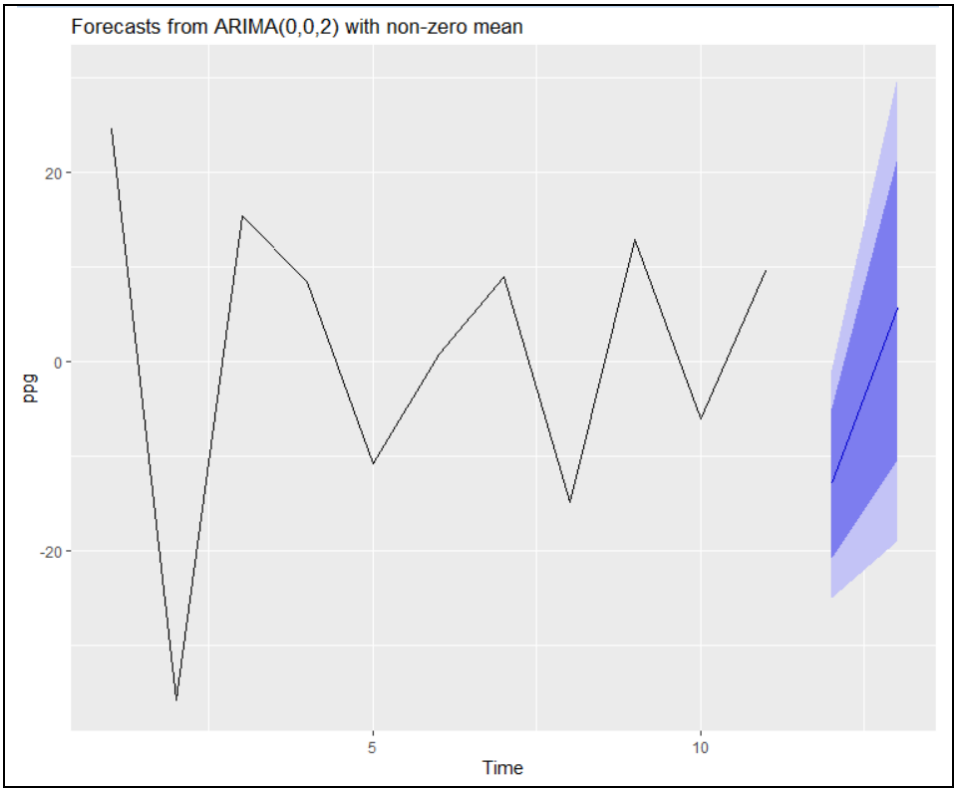
Χρονοσειρά	Μέθοδος	Εκτίμηση Παραμέτρων
ppg	**mle**	$\theta_1 = -1.902, \theta_2 = 0.999$

Πίνακας 6.3: Εκτίμηση των παραμέτρων του μοντέλου ARIMA(0,0,2) για την χρονοσειρά ppg

Εφόσον βρέθηκε το κατάλληλο μοντέλο για την περιγραφή της χρονοσειράς, θα γίνουν οι προβλέψεις για τις επόμενες δύο (2) σεζόν. Στον Πίνακα 6.4 φαίνονται οι τιμές που λαμβάνει η χρονοσειρά και στο Σχήμα 6.3 υπάρχει το γράφημα των τιμών μαζί με τα διαστήματα εμπιστοσύνης των τιμών. Πιο συγκεκριμένα στην σκουρόχρωμη περιοχή του μπλε χρώματος φαίνεται το 20% διάστημα εμπιστοσύνης των τιμών, ενώ στο πιο ανοιχτόχρωμο φαίνεται το 5% διάστημα εμπιστοσύνης.

Χρονοσειρά	ppg	Ppg
Προβλέψεις	**1η**	**2η**
Τιμές	**-12.9329**	**5.56598**

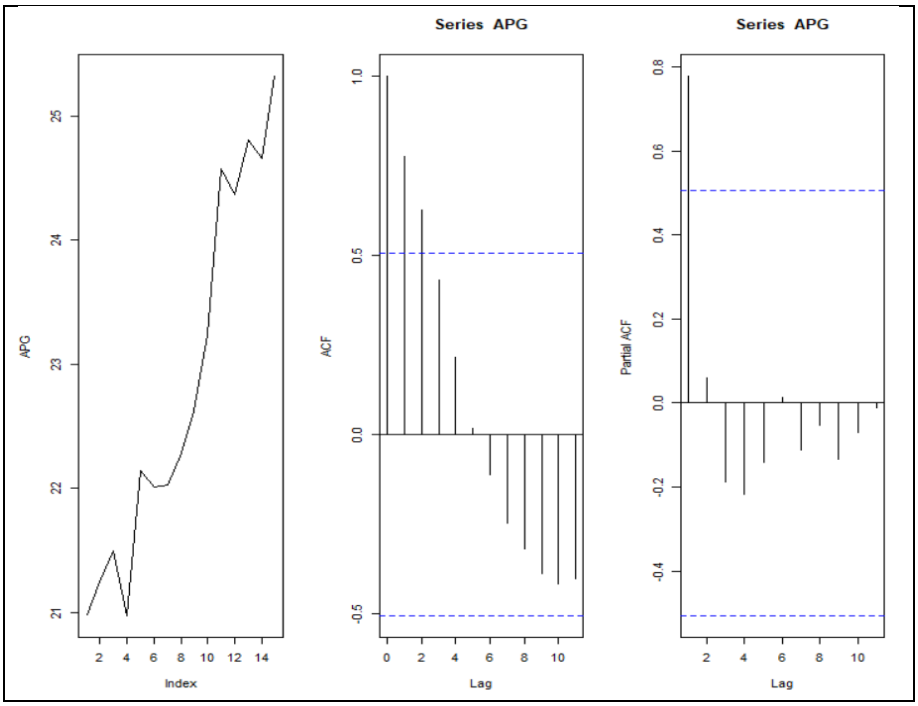
Πίνακας 6.4: Πρόβλεψη μελλοντικών τιμών της χρονοσειράς ppg



Σχήμα 6.3: Γράφημα προβλέψεων των μελλοντικών τιμών της χρονοσειράς ppg

6.2.2. Ανάλυση της χρονοσειράς ασίστ ανά αγώνα (APG)

Όσον αφορά τη χρονοσειρά των ασίστ ανά αγώνα, όπως φαίνεται στο Σχήμα 6.4 στο αριστερό γράφημα, δεν είναι στάσιμη διότι παρουσιάζει αυξητική τάση. Επομένως ακολουθώντας τη μέθοδο των διαφορών, αυτή τη φορά με τάξη ίση με πέντε (5), δημιουργούμε μία νέα στάσιμη χρονοσειρά, την apg.

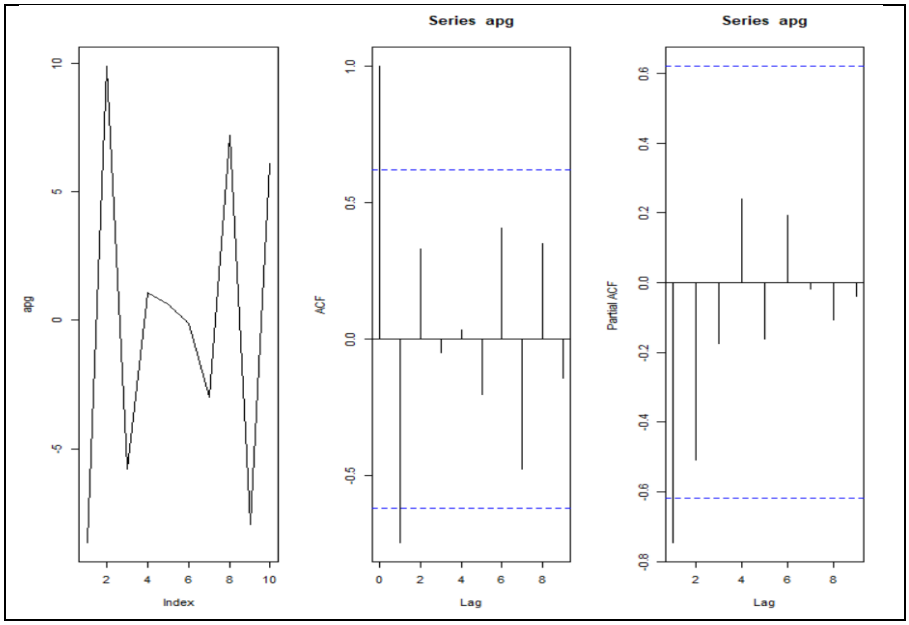


Σχήμα 6.4: Γραφήματα χρονοσειράς ασίστ ανά αγώνα

Στο Σχήμα 6.5 παρουσιάζονται τα γραφήματα της νέας χρονοσειράς apg, ενώ στον Πίνακα 6.5 οι έλεγχοι που διενεργούνται καθ' όλη την διάρκεια των χρονοσειρών. Αρχικά, στον Πίνακα 6.5 παρατηρούμε τη στήλη με το p_{value} του adf test, έλεγχος μοναδιαίας ρίζας, με τιμή 0.00 απ' όπου συμπεραίνουμε πως η νέα χρονοσειρά είναι στάσιμη.

Χρονοσειρά	p_{value} adf test	p_{value} Shapiro-Wilk	p_{value} Anderson-Darling	p_{value} Box-Ljung
apg	**<0.01**	**0.7751**	**0.8961**	**0.1983**

Πίνακας 6.5: Έλεγχοι της στάσιμης χρονοσειράς apg



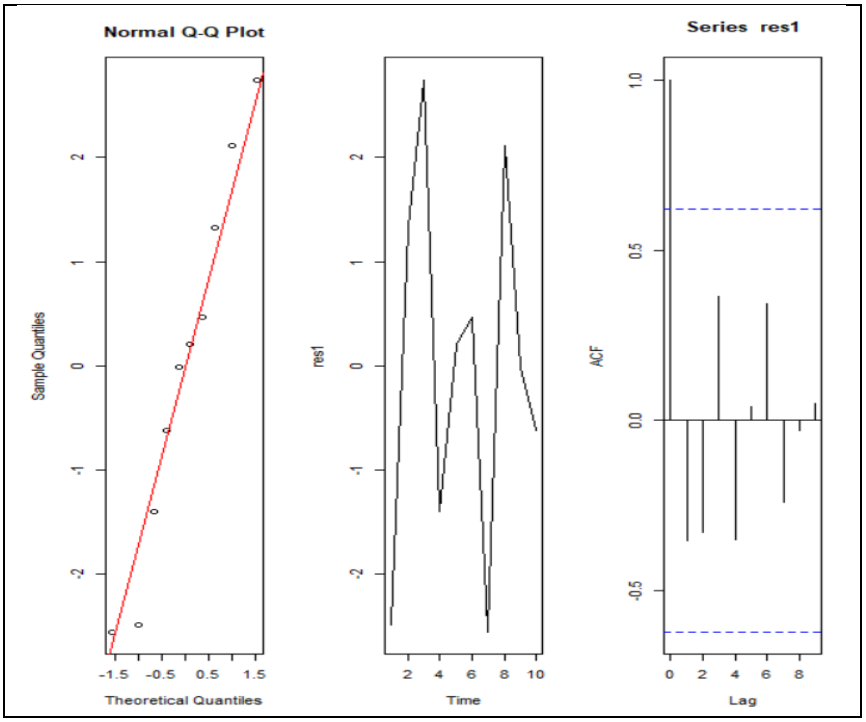
Σχήμα 6.5: Γραφήματα της νέας στάσιμης χρονοσειράς ασίστ ανά αγώνα (apg)

Σύμφωνα με τα γραφήματα, τα οποία περιέχονται στον Πίνακα 6.5, αυτοσυσχέτισης και μερικής αυτοσυσχέτισης, πιθανά μοντέλα που ερμηνεύουν την χρονοσειρά είναι τα: ARIMA(0,0,0), ARIMA(0,0,1), ARIMA(0,0,2) και ARIMA(1,0,1).

Αυτές οι επιλογές έγιναν διότι στο γράφημα της αυτοσυσχέτισης παρατηρούμε πως οι δύο πρώτες είναι στατιστικά σημαντικές, ενώ στο γράφημα της μερικής αυτοσυσχέτισης παρατηρούμε πως η μόνη στατικά σημαντική είναι η πρώτη. Σύμφωνα με τους δείκτες που υπολογίστηκαν, Πίνακας 6.6, μικρότερη τιμή έχει το μοντέλο ARIMA(1,0,1). Στη συνέχεια εκτελώντας τους ελέγχους κανονικότητας και ανεξαρτησίας των καταλοίπων του, στον Πίνακα 6.5 παρουσιάζονται τα p_{value}, παρατηρούμε πως δεν υπάρχουν ενδείξεις απόρριψης των μηδενικών υποθέσεων και επομένως το μοντέλο είναι κατάλληλο για την περιγραφή της χρονοσειράς. Επίσης στο Σχήμα 6.6 παρουσιάζονται γραφικά οι έλεγχοι της κανονικότητας στο αριστερό γράφημα, και της ανεξαρτησίας στο δεξιό γράφημα, καθώς και η χρονοσειρά των καταλοίπων στο κέντρο τους σχήματος. Ακόμα στον Πίνακα 6.7 παρουσιάζονται οι εκτιμήτριες των παραμέτρων του μοντέλου.

Χρονοσειρά	apg	Apg	apg	apg
Μοντέλο	**ARIMA(0,0,0)**	**ARIMA(0,0,1)**	**ARIMA(0,0,2)**	**ARIMA(1,0,1)**
AIC	**68.4564**	**60.54731**	**54.12943**	**51.75933**
BIC	**69.0615**	**61.45506**	**55.33977**	**52.96967**
AICc	**70.1707**	**64.54731**	**62.12943**	**59.75933**

Πίνακας 6.6: Δείκτες AIC, BIC και AICc για την χρονοσειρά apg



Σχήμα 6.6: Γραφικές παραστάσεις καταλοίπων της χρονοσειράς apg

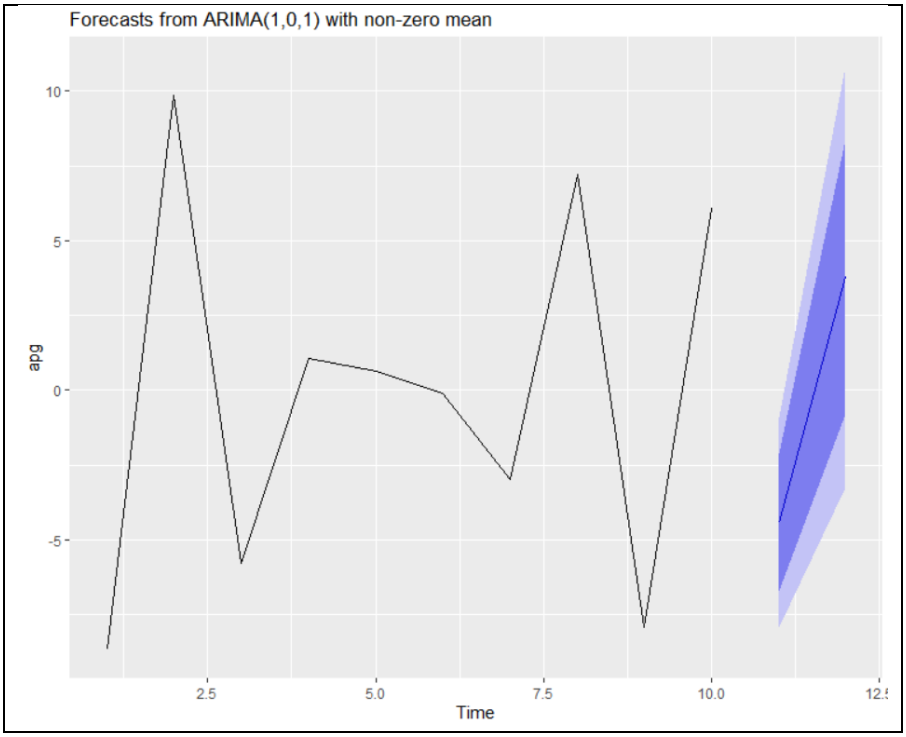
Χρονοσειρά	Μέθοδος	Εκτίμηση Παραμέτρου
rpg	**Mle**	**φ_1 = -0.83702, θ_1 = -0.9998**

Πίνακας 6.7: Εκτίμηση των παραμέτρων του μοντέλου ARIMA(1,0,1) για την χρονοσειρά apg

Αφού καταλήξαμε στο βέλτιστο μοντέλο για την περιγραφή της νέας χρονοσειράς, θα προχωρήσουμε στις προβλέψεις των επόμενων πέντε (5) σεζόν. Στον Πίνακα 6.8 φαίνονται οι τιμές τους και στο Σχήμα 6.7 παρουσιάζεται το αντίστοιχο γράφημα μελλοντικών τιμών.

Χρονοσειρά	apg	apg
Προβλέψεις	**1η**	**2η**
Τιμές	**-4.413671**	**3.791528**

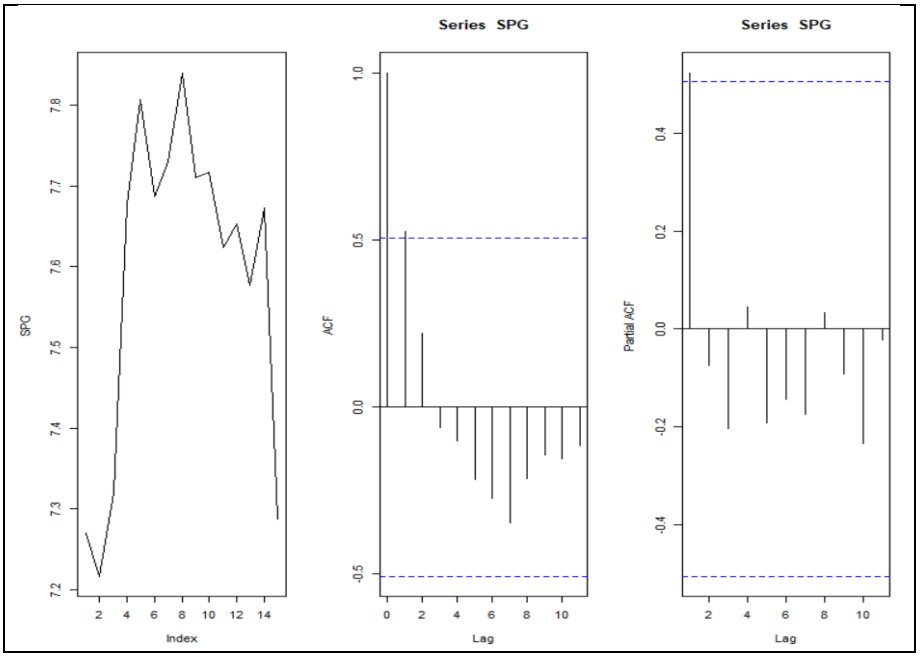
Πίνακας 6.8: Πρόβλεψη μελλοντικών τιμών της χρονοσειράς apg



Σχήμα 6.7: Γράφημα προβλέψεων των μελλοντικών τιμών της χρονοσειράς apg

6.2.3. Ανάλυση της χρονοσειράς κλεψιμάτων ανά αγώνα (SPG)

Σύμφωνα με το Σχήμα 6.8 η χρονοσειρά των κλεψιμάτων ανά αγώνα που τίθεται προς ανάλυση δεν είναι στάσιμη. Επομένως, όπως και στις προηγούμενες παραγράφους, θα πρέπει να αφαιρεθεί η τάση που συμπεριλαμβάνεται σε αυτή, ώστε να δημιουργήσουμε μία νέα στάσιμη χρονοσειρά.

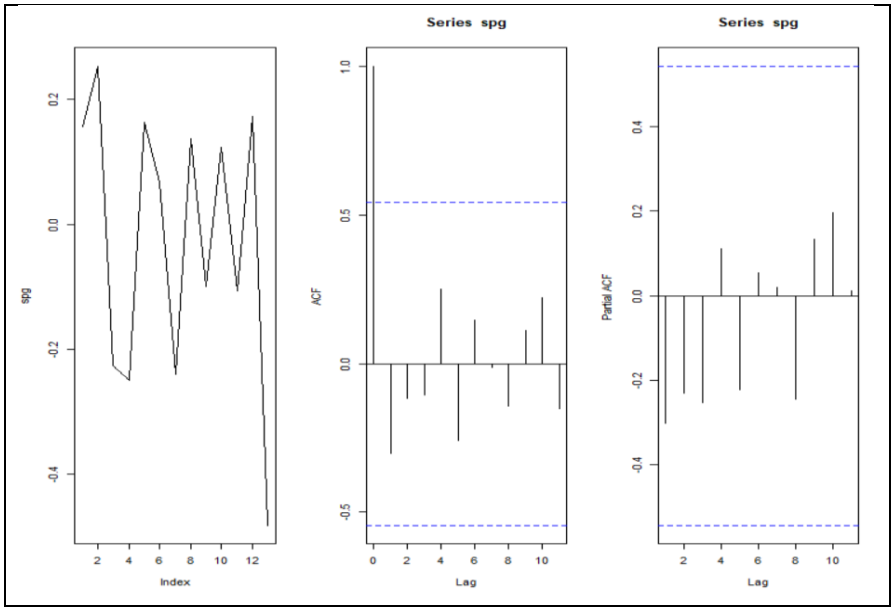


Σχήμα 6.8: Γραφήματα χρονοσειράς κλεψιμάτων ανά αγώνα

Χρησιμοποιώντας τις δεύτερης τάξης διαφορές δημιουργούμε μια νέα στάσιμη χρονοσειρά, ο επαυξημένος έλεγχος Dickey-Fuller παρουσιάζεται στον Πίνακα 6.9 στη στήλη p_{value} adf test όπου και παρατηρούμε πως δεν υπάρχουν ενδείξεις μη απόρριψης της μηδενικής υπόθεσης (p_{value}<0.01). Ακόμα, στο Σχήμα 6.9 παρουσιάζονται γραφήματα της νέας στάσιμης χρονοσειράς, spg.

Χρονοσειρά	p_{value} adf test	p_{value} Shapiro-Wilk	p_{value} Anderson-Darling	p_{value} Box-Ljung
spg	**<0.01**	**0.974**	**0.8158**	**0.5288**

Πίνακας 6.9: Έλεγχοι της στάσιμης χρονοσειράς spg



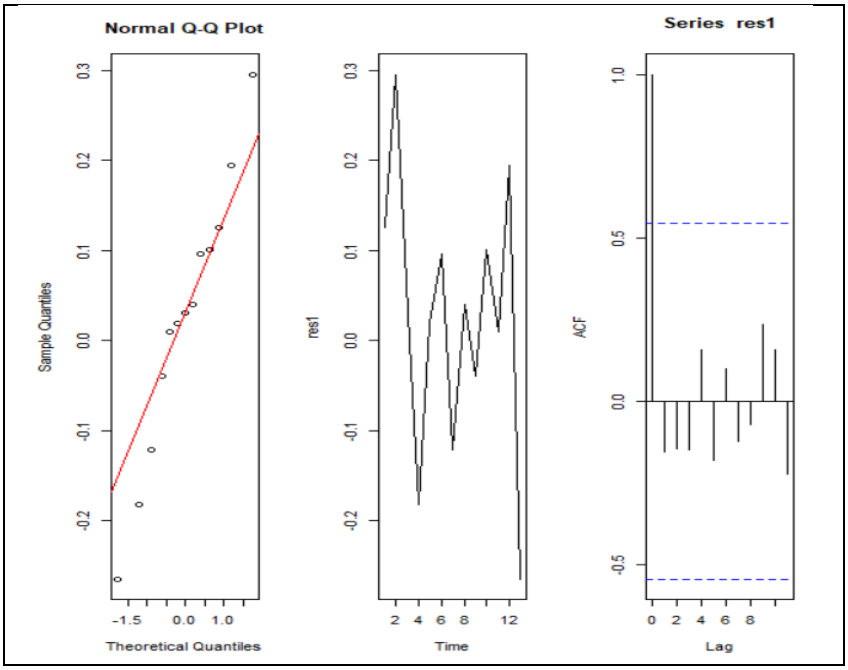
Σχήμα 6.9: Γραφήματα της νέας στάσιμης χρονοσειράς κλεψιμάτων ανά αγώνα (spg)

Από τα γραφήματα τα οποία περιέχονται στον Πίνακα 6.9, αυτοσυσχέτισης και μερικής αυτοσυσχέτισης, τα πιθανότερα μοντέλα που ερμηνεύουν σωστά τη νέα στάσιμη χρονοσειρά είναι τα ARIMA(0,0,0), ARIMA(0,0,1), ARIMA(0,0,2) και ARIMA(0,0,3). Οι επιλογές αυτές έγιναν διότι μόνο η πρώτη αυτοσυσχέτιση στα δύο γραφήματα είναι στατιστικά σημαντική. Στον Πίνακα 6.10 παρουσιάζονται οι τιμές των δεικτών για τη νέα στάσιμη χρονοσειρά. Σύμφωνα λοιπόν με τους υπολογισμούς τα κριτήρια AIC και BIC παίρνουν μικρότερες τιμές για το μοντέλο ARIMA(0,0,2), όμως το κριτήριο AICc λαμβάνει τη μικρότερη τιμή του για το μοντέλο ARIMA(0,0,1). Όπως αναφέρθηκε και στην παράγραφο του θεωρητικού υποβάθρου, επειδή τα δεδομένα μας είναι λίγα, θα ακολουθήσουμε την απόφαση του κριτηρίου AICc και θα συνεχίσουμε την ανάλυση με το μοντέλο ARIMA(0,0,1). Έπειτα εκτελώντας τους ελέγχους κανονικότητας και ανεξαρτησίας των καταλοίπων του, στον Πίνακα 6.9 παρουσιάζονται τα p_{value}, παρατηρούμε πως δεν υπάρχουν ενδείξεις απόρριψης των μηδενικών υποθέσεων. Συνεπώς το μοντέλο που επιλέχθηκε θεωρείται κατάλληλο για την περιγραφή της χρονοσειράς, στο Σχήμα 6.10 δίνονται και τα γραφήματα των

ελέγχων κανονικότητας και ανεξαρτησίας. Ο Πίνακας 6.11 περιέχει τις εκτιμήτριες των παραμέτρων του μοντέλου.

Χρονοσειρά	Spg	Spg	spg	spg
Μοντέλο	**ARIMA(0,0,0)**	**ARIMA(0,0,1)**	**ARIMA(0,0,2)**	**ARIMA(0,0,3)**
AIC	**0.923907**	**-4.20728**	**-5.76555**	**-4.211204**
BIC	**2.05380**	**-2.51243**	**-3.505755**	**-1.41645**
AICc	**2.12390**	**-1.54061**	**-0.76555**	**4.33022**

Πίνακας 6.10: Δείκτες AIC, BIC και AICc για την χρονοσειρά spg



Σχήμα 6.10: Γραφικές παραστάσεις καταλοίπων της χρονοσειράς spg

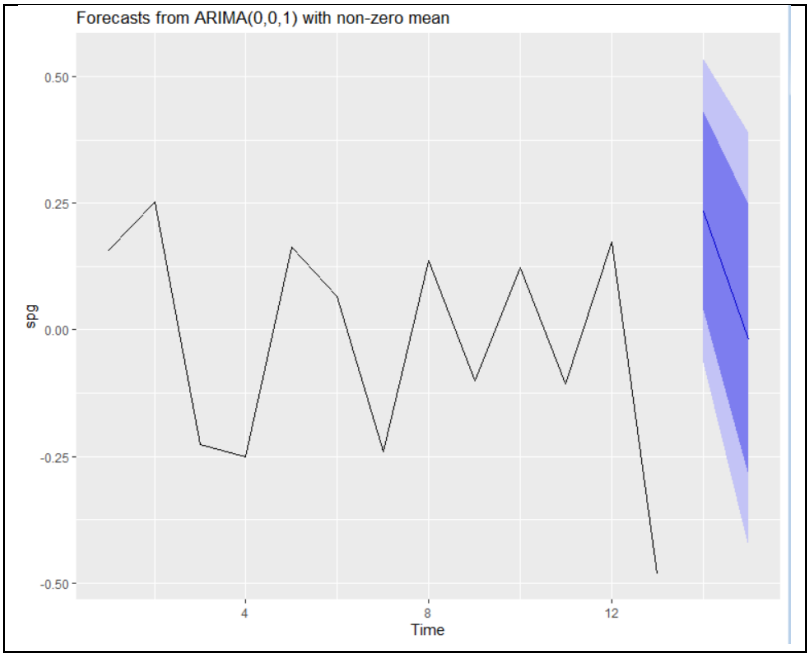
Χρονοσειρά	Μέθοδος	Εκτίμηση Παραμέτρων
spg	**mle**	**θ_1= -0.9999957**

Πίνακας 6.11: Εκτίμηση των παραμέτρων του μοντέλου ARIMA(0,0,1) για την χρονοσειρά spg

Βασικός στόχος της ανάλυσης είναι να γίνουν οι προβλέψεις, οι οποίες παρουσιάζονται στον Πίνακα 6.12 και στο Σχήμα 6.11.

Χρονοσειρά	spg	spg
Προβλέψεις	**1η**	**2η**
Τιμές	**0.23526**	**-0.02025**

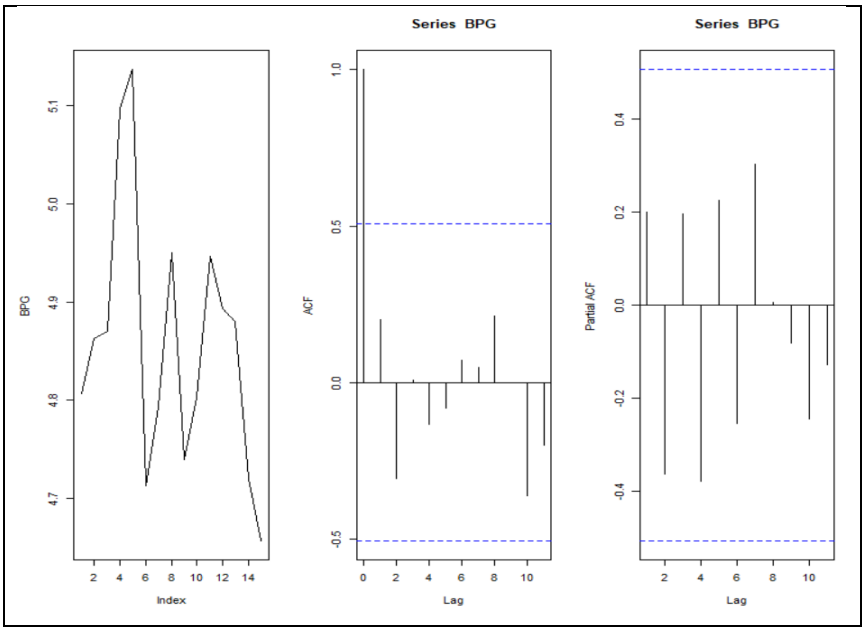
Πίνακας 6.12: Πρόβλεψη μελλοντικών τιμών της χρονοσειράς spg



Σχήμα 6.11: Γράφημα προβλέψεων των μελλοντικών τιμών της χρονοσειράς spg

6.2.4. Ανάλυση της χρονοσειράς κοψιμάτων ανά αγώνα (BPG)

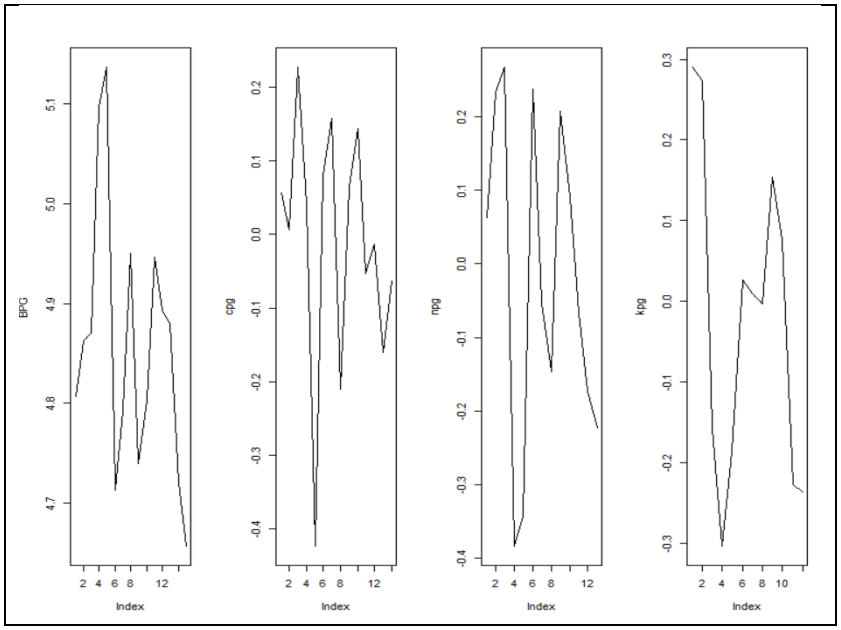
Στο Σχήμα 6.12 παρουσιάζονται τα γραφήματα της υπό μελέτη χρονοσειράς, κοψιμάτων ανά αγώνα, όπου και παρατηρούμε πως υπάρχει εποχικότητα στις τιμές της. Επομένως θα πρέπει να χρησιμοποιήσουμε τη μέθοδο των διαφορών για την εξάλειψη της εποχικότητας.



Σχήμα 6.12: Γραφήματα χρονοσειράς κοψιμάτων ανά αγώνα

Στο Σχήμα 6.13 παρουσιάζονται από αριστερά προς δεξιά το γράφημα της υπό μελέτη χρονοσειράς, καθώς και των νέων χρονοσειρών με εποχικές διαφορές πρώτης, δεύτερης και τρίτης τάξης. Προφανώς το γράφημα με τις διαφορές τρίτης τάξης δείχνει πως έχει εξαλείψει την εποχικότητα και είναι η χρονοσειρά που επιλέγεται για τη συνέχεια της ανάλυσης, με όνομα kpg. Στη συνέχεια της ανάλυσης χρησιμοποιούμε τις

διαφορές για την εξάλειψη της τάσης, διότι η χρονοσειρά που δημιουργήθηκε, kpg, δεν είναι στάσιμη σύμφωνα με τον επαυξημένο έλεγχο Dickey-Fuller που λαμβάνει τιμή p_{value} ίση με 0.8197, μη απόρριψη της μηδενικής υπόθεσης.



Σχήμα 6.13: Γραφήματα χρονοσειρών κοψιμάτων ανά αγώνα με αφαίρεση της εποχικότητας

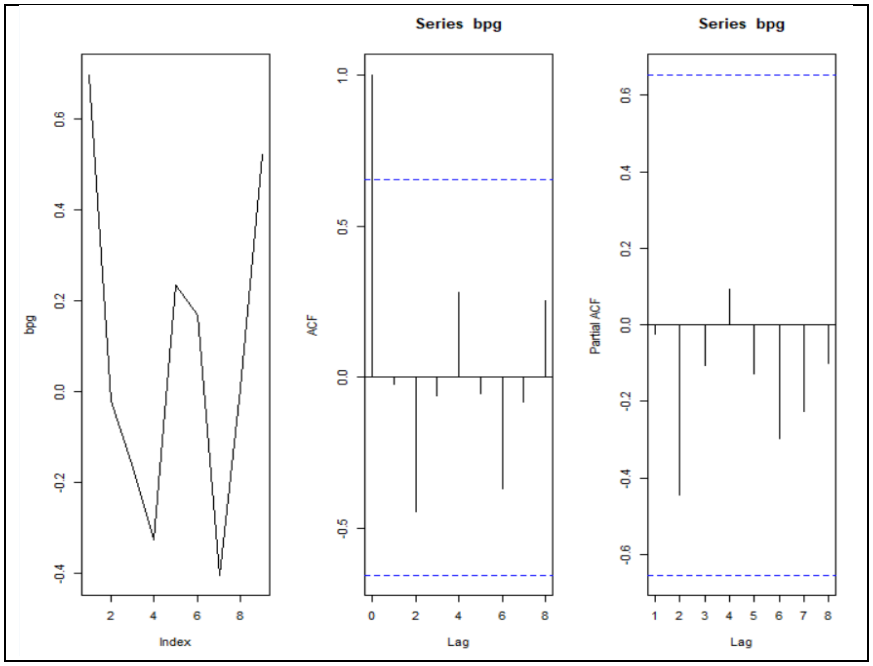
Στον Πίνακα 6.13 παρουσιάζεται η τιμή του επαυξημένου ελέγχου για τη μοναδιαία ρίζα της χρονοσειράς, αφού χρησιμοποιήσαμε τις διαφορές τρίτης τάξης για την εξάλειψη της τάση. Σύμφωνα με το p_{value} του ελέγχου δεν υπάρχουν αρκετές ενδείξεις για τη μη απόρριψη της μηδενικής υπόθεσης και επομένως η χρονοσειρά είναι στάσιμη. Τα νέα γραφήματα της στάσιμης χρονοσειράς, bpg, παρουσιάζονται στο Σχήμα 6.14. Για την εύρεση του βέλτιστου μοντέλου περιγραφής της χρονοσειράς κοψιμάτων ανά αγώνα τα πιθανότερα μοντέλα είναι: ARIMA$(0,3,0)(0,1,0)_3$, ARIMA$(0,3,1)(0,1,0)_3$, ARIMA$(0,3,2)(0,1,0)_3$ και ARIMA$(1,3,1)(0,1,0)_3$. Στον Πίνακα 6.14 παρουσιάζονται οι τιμές των δεικτών AIC και BIC για τα μοντέλα που επιλέχθηκαν προς ανάλυση και είναι φανερό πως το καλύτερο είναι το μοντέλο ARIMA$(0,3,1)(0,1,0)_3$.

Χρονοσειρά	p_{value} adf test	p_{value} Shapiro-Wilk	p_{value} Anderson-Darling	p_{value} Box-Ljung
bpg	**<0.01**	**0.1164**	**0.04633**	**0.5288**

Πίνακας 6.13: Έλεγχοι της στάσιμης χρονοσειράς BPG

Χρονοσειρά	BPG	BPG	BPG	BPG
Μοντέλο	**ARIMA$(0,3,0)(0,1,0)_3$**	**ARIMA$(0,3,1)(0,1,0)_3$**	**ARIMA$(0,3,2)(0,1,0)_3$**	**ARIMA$(1,3,1)(0,1,0)_3$**
AIC	**14.0084**	**12.45438**	**14.52353**	**14.3952**
BIC	**14.20562**	**12.84883**	**15.1152**	**14.98687**

Πίνακας 6.14: Δείκτες AIC, BIC και AICc για την χρονοσειρά BPG

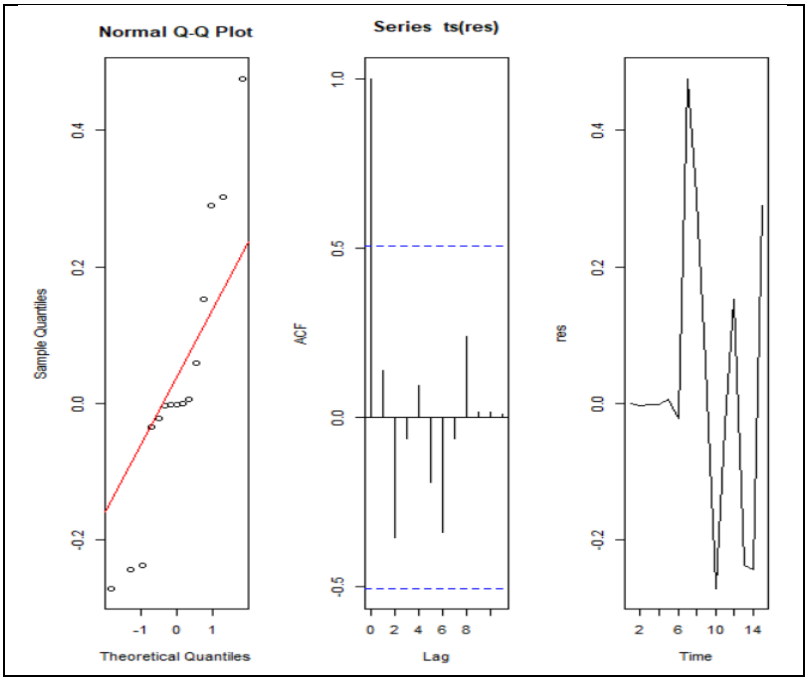


Σχήμα 6.14: Γραφήματα της νέας στάσιμης χρονοσειράς κοψιμάτων ανά αγώνα (bpg)

Για τους ελέγχους κανονικότητας και ανεξαρτησίας των καταλοίπων της χρονοσειράς παρατηρούμε στον Πίνακα 6.13 πως και το p_{value} του ελέγχου Shapiro-Wilk και του Box-Ljung παίρνουν αρκετά μεγάλες τιμές και επομένως δεν έχουμε ενδείξεις για την απόρριψη της μηδενικής υπόθεσης. Παρ' όλα αυτά στον έλεγχο του Anderson-Darling το p_{value} για την κανονικότητα των καταλοίπων απορρίπτει την μηδενική υπόθεση, ίσο με 0.04633. Συνεπώς θα δοκιμάσουμε τα υπόλοιπα μοντέλα με πρώτο αυτό με την μικρότερη τιμή στους δείκτες. Για το μοντέλο ARIMA$(0,3,2)(0,1,0)_3$, σύμφωνα με τον Πίνακα 6.15, όλοι οι έλεγχοι των καταλοίπων λαμβάνουν υψηλές τιμές και κρίνεται το καταλληλότερο για την περιγραφή της χρονοσειράς. Στο Σχήμα 6.15 παρουσιάζονται και τα γραφήματα των καταλοίπων του μοντέλου που επιλέχθηκε.

Χρονοσειρά	p_{value} Shapiro-Wilk	p_{value} Anderson-Darling	p_{value} Box-Ljung
BPG	**0.160**	**0.07539**	**0.4079**

Πίνακας 6.15: Έλεγχοι των καταλοίπων του μοντέλου ARIMA$(0,3,2)(0,1,0)_3$

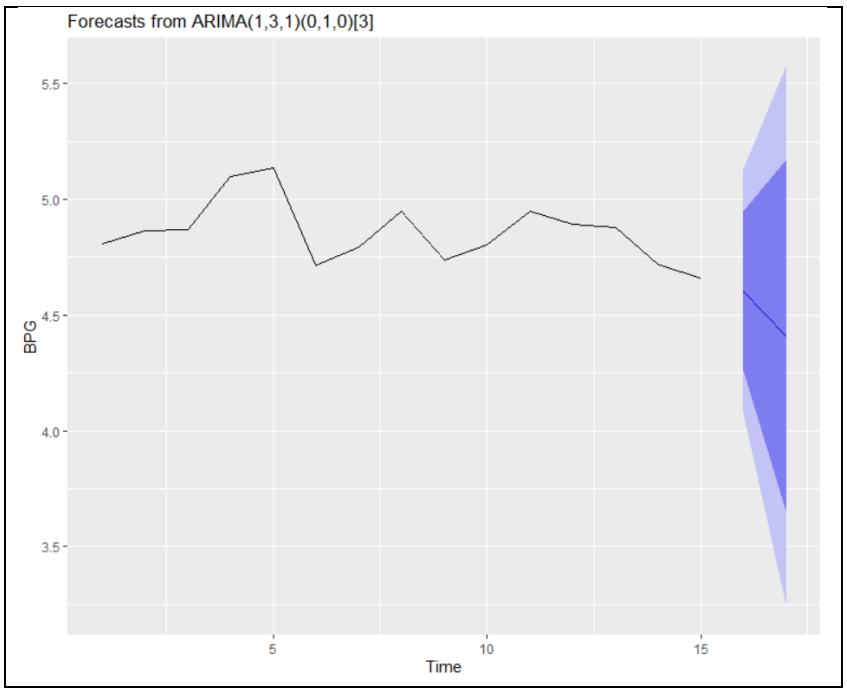


Σχήμα 6.15: Γραφικές παραστάσεις καταλοίπων της χρονοσειράς bpg

Τέλος, στον Πίνακα 6.16 και στο Σχήμα 6.16 δίνονται οι προβλέψεις της χρονοσειράς των κοψιμάτων ανά αγώνα για τις επόμενες πέντε (5) σεζόν που ακολουθούν.

Χρονοσειρά	BPG	BPG
Προβλέψεις	**1η**	**2η**
Τιμές	**4.60279**	**4.40753**

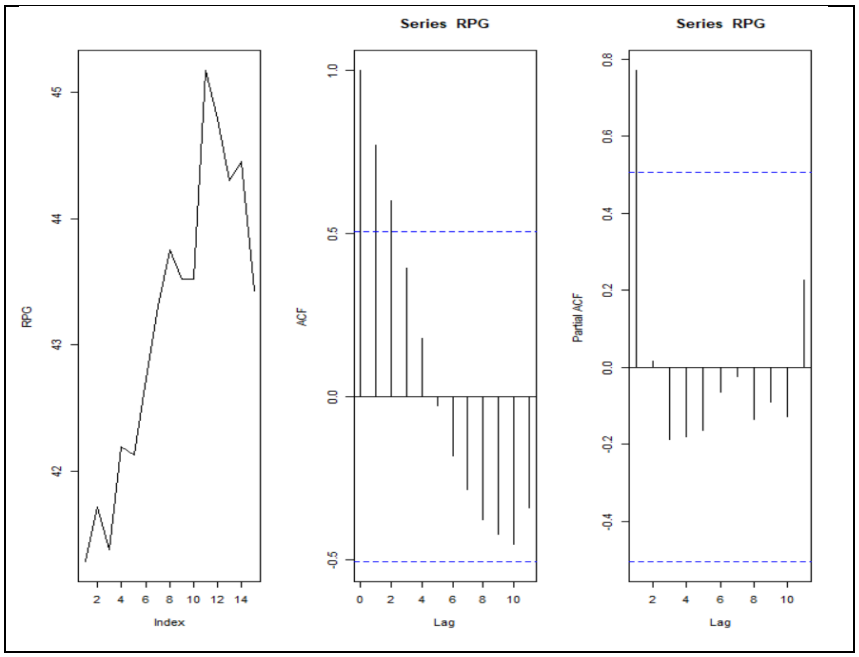
Πίνακας 6.16: Πρόβλεψη μελλοντικών τιμών της χρονοσειράς BPG



Σχήμα 6.16: Γράφημα προβλέψεων των μελλοντικών τιμών της χρονοσειράς BPG

6.2.5. Ανάλυση της χρονοσειράς ριμπάουντ ανά αγώνα (RPG)

Στην παρούσα παράγραφο θα γίνει η ανάλυση της χρονοσειράς των ριμπάουντ ανά αγώνα. Στο Σχήμα 6.17 παρουσιάζονται τα γραφήματα της χρονοσειράς. Όπως γίνεται κατανοητό απ' αυτό το σχήμα, η υπό μελέτη χρονοσειρά παρουσιάζει τάση και επομένως δεν θεωρείται στάσιμη. Γι' αυτό τον λόγο εκτελείται η μέθοδος διαφορών τρίτης τάξης για εξάλειψη της τάσης.

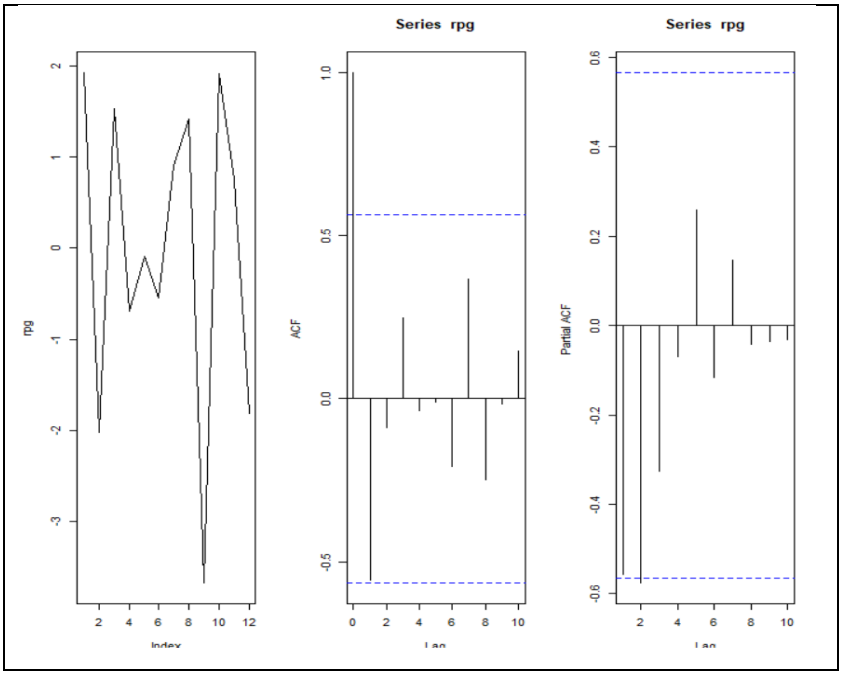


Σχήμα 6.17: Γραφήματα χρονοσειράς ριμπάουντ ανά αγώνα

Εφόσον δημιουργήσαμε την νέα χρονοσειρά rpg, στον Πίνακα 6.17 παρουσιάζεται το p_{value} adf test του επαυξημένου ελέγχου Dickey-Fuller με τιμή μικρότερη του 0.05 απ' όπου συμπεραίνουμε πως η αυτή είναι στάσιμη. Ακόμα, στο Σχήμα 6.18 παρουσιάζονται τα γραφήματα της ώστε να είμαστε σε θέση να ανακαλύψουμε το καταλληλότερο μοντέλο για την περιγραφή της.

Χρονοσειρά	p_{value} adf test	p_{value} Shapiro-Wilk	p_{value} Anderson-Darling	p_{value} Box-Ljung
Rpg	**0.04362**	**0.0672**	**0.1289**	**0.1884**

Πίνακας 6.17: Έλεγχοι της στάσιμης χρονοσειράς rpg

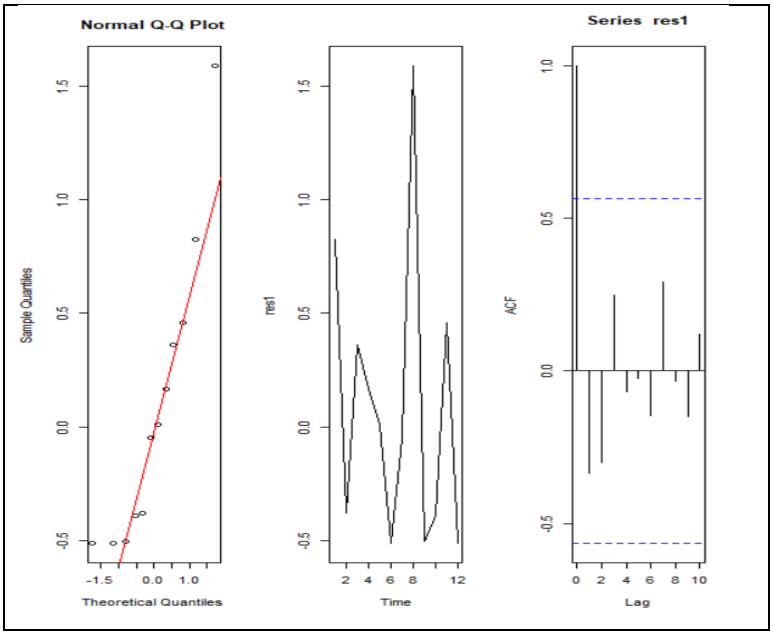


Σχήμα 6.18: Γραφήματα της νέας στάσιμης χρονοσειράς ριμπάουντ ανά αγώνα (rpg)

Από τα γραφήματα της νέας στάσιμης χρονοσειράς τα πιθανότερα μοντέλα που θα δοκιμαστούν για την περιγραφή της είναι τα ARIMA(0,0,0), ARIMA(0,0,1), ARIMA(0,0,2) και ARIMA(0,0,3). Ο λόγος για τον οποίο επιλέχθηκαν τα συγκεκριμένα μοντέλα είναι διότι η μόνη στατιστικά σημαντική αυτοσυσχέτιση στο αντίστοιχο γράφημα είναι η πρώτη. Οι τιμές των τριών δεικτών για τα μοντέλα δίνονται στον Πίνακα 6.18, όπου και αντιλαμβανόμαστε πως σύμφωνα με το κριτήριο AICc το καταλληλότερο μοντέλο είναι το ARIMA(0,0,2). Αντίθετα για τα κριτήρια AIC και BIC καλύτερο μοντέλο είναι το ARIMA(0,0,3). Συνεχίζοντας την ανάλυση για το μοντέλο που δόθηκε βάσει του κριτηρίου AICc, όπως έχουμε αναφέρει προηγουμένως προτιμάται για μικρά δείγματα όπως αυτό της εργασίας, εκτελέστηκαν οι έλεγχοι κανονικότητας και ανεξαρτησίας όπου και στους δύο δεν είχαμε ενδείξεις μη απόρριψης της μηδενικής υπόθεσης εφόσον οι τιμές του p_{value} ήταν μεγαλύτερες του 0.05 (Πίνακας 6.17). Επομένως το μοντέλο κρίνεται ορθό για την περιγραφή της χρονοσειράς. Επίσης στο Σχήμα 6.19 δίνονται τα γραφήματα των καταλοίπων και των ελέγχων τους. Στον Πίνακα 6.19 παρουσιάζονται οι εκτιμήτριες των συντελεστών του μοντέλου.

Χρονοσειρά	rpg	Rpg	rpg	rpg
Μοντέλο	**ARIMA(0,0,0)**	**ARIMA(0,0,1)**	**ARIMA(0,0,2)**	**ARIMA(0,0,3)**
AIC	**50.8428**	**42.5281**	**36.9427**	**35.6615**
BIC	**51.8126**	**43.9828**	**38.8823**	**38.0861**
AICc	**52.1761**	**45.5281**	**42.6563**	**45.6615**

Πίνακας 6.18: Δείκτες AIC, BIC και AICc για την χρονοσειρά rpg



Σχήμα 6.19: Γραφικές παραστάσεις καταλοίπων της χρονοσειράς rpg

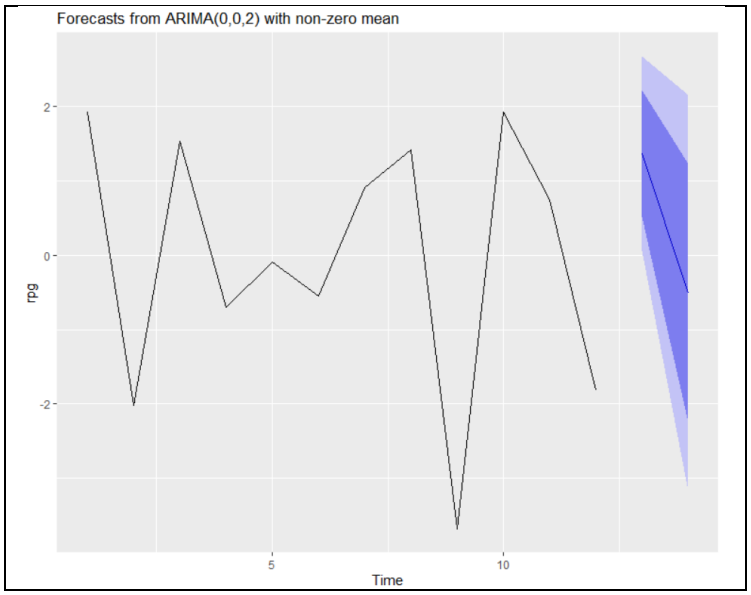
Χρονοσειρά	Μέθοδος	Εκτίμηση Παραμέτρων
spg	**Mle**	**$\theta_1 = -1.91138$, $\theta_2 = 0.9999$**

Πίνακας 6.19: Εκτίμηση των παραμέτρων του μοντέλου ARIMA(0,0,2) για την χρονοσειρά rpg

Εφόσον βρέθηκε το μοντέλο που περιγράφει καλύτερα την χρονοσειρά rpg, θα γίνουν οι προβλέψεις για τις τιμές των επόμενων πέντε (5) σεζόν. Στον Πίνακα 6.20 δίνονται οι τιμές τους και στο Σχήμα 6.19 το γράφημα τον μελλοντικών τιμών.

Χρονοσειρά	rpg	Rpg
Προβλέψεις	**1η**	**2η**
Τιμές	**1.3673**	**-0.51000**

Πίνακας 6.20: Πρόβλεψη μελλοντικών τιμών της χρονοσειράς rpg



Σχήμα 6.20: Γράφημα προβλέψεων των μελλοντικών τιμών της χρονοσειράς rpg

ΚΕΦΑΛΑΙΟ 7ο

7. Μηχανική Μάθηση

Στον παρόν κεφάλαιο της εργασίας, μέσω της εκτέλεσης τεχνικών μηχανικής μάθησης και της κατάλληλης επεξεργασίας των δεδομένων μας, αποσκοπούμε στην εξαγωγή χρήσιμης πληροφορίας. Πιο συγκεκριμένα, έπειτα από την επιλογή των καταλληλότερων μεταβλητών (feature selection) από τα δεδομένα μας, έγινε χρήση αλγορίθμων κατηγοριοποίησης (classification), προκειμένου να προσαρμοστούν διάφορα μοντέλα ταξινόμησης για το αν μία ομάδα θα προκρινόταν στην φάση των Playoffs της διοργάνωσης του NBA, καθώς και αλγορίθμων ομαδοποίησης (clustering), ώστε να εντοπιστούν ομάδες παρατηρήσεων με παρόμοια χαρακτηριστικά ανάμεσα στα δεδομένα μας. Προφανώς, ως μεταβλητή απόκρισης θα χρησιμοποιηθεί η μεταβλητή «Playoffs», η οποία παρέχει την πληροφορία για το αν η ομάδα κατάφερε να προκριθεί στην φάση των playoffs ή όχι.

7.1. Θεωρητικό υπόβαθρο

Στην παρούσα ενότητα θα παρουσιαστούν οι θεωρητικές γνώσεις που χρειάζονται ώστε να υλοποιηθούν οι τεχνικές της μηχανικής μάθησης.

7.1.1. Εξόρυξη δεδομένων

Με τον όρο εξόρυξη δεδομένων (Data Mining) αναφερόμαστε σε μια διαδικασία ανακάλυψης ενδιαφερόντων προτύπων και γνώσεων από μεγάλες βάσεις δεδομένων, με χρήση αλγορίθμων ομαδοποίησης ή κατηγοριοποίησης και των αρχών της στατιστικής, της τεχνητής νοημοσύνης, της μάθησης και των συστημάτων βάσεων δεδομένων. Οι πηγές των δεδομένων μπορούν να περιλαμβάνουν βάσεις δεδομένων, αποθήκες δεδομένων, το διαδίκτυο και άλλα αποθετήρια πληροφοριών ή δεδομένα που μεταδίδονται δυναμικά στο σύστημα. Ο στόχος της εξόρυξης δεδομένων είναι η πληροφορία που θα εξαχθεί και τα πρότυπα που θα προκύψουν να έχουν δομή κατανοητή προς τον άνθρωπο, ώστε να τον βοηθήσουν να πάρει τις κατάλληλες αποφάσεις για περαιτέρω ανάλυση ή για παράδειγμα στην εκμάθηση μηχανής και στην προγνωστική ανάλυση. (Han et al., 2012)

7.1.1.1. Βήματα διαδικασίας εξόρυξης δεδομένων

Τα δεδομένα που χρησιμοποιούνται στους αλγορίθμους πρέπει να έχουν ποιότητα ώστε να ικανοποιούν τον επιδιωκόμενο σκοπό. Συνεπώς η προ-επεξεργασία είναι μία κρίσιμη διαδικασία της εξόρυξης δεδομένων. Τα βασικότερα βήματα που εμπλέκονται στην διαδικασία αυτή παρουσιάζονται αναλυτικά παρακάτω και γραφικά στο Σχήμα 7.1.

- Καθαρισμός δεδομένων

Ο καθαρισμός δεδομένων είναι το πρώτο βήμα στην διαδικασία εξόρυξης δεδομένων. Είναι ιδιαίτερης σημασία διότι αν τα δεδομένα δεν καθαριστούν και χρησιμοποιηθούν απευθείας, μπορούν να προκαλέσουν σύγχυση στις διαδικασίες που θα ακολουθήσουν και να παράγουν ανακριβή αποτελέσματα. Εν ολίγοις, το συγκεκριμένο βήμα περιλαμβάνει την αφαίρεση του θορύβου ή των ελλιπών δεδομένων από την συλλογή.

- Ενσωμάτωση δεδομένων

Σε αυτό το βήμα πολλαπλές ετερογενείς πηγές δεδομένων, όπως βάσεις δεδομένων, κύβοι δεδομένων ή αρχεία, συνδυάζονται για την ανάλυση. Αυτή η διαδικασία ονομάζεται ολοκλήρωση δεδομένων.

- Μείωση/επιλογή δεδομένων

Αυτή η τεχνική χρησιμοποιείται για τη λήψη καταλληλότερων δεδομένων για την ανάλυση τους. Το μέγεθος της αναπαράστασης είναι πολύ μικρότερο σε όγκο διατηρώντας όμως ταυτόχρονα την ακεραιότητα τους. Πιο συγκεκριμένα, στόχος αυτής της διαδικασίας είναι η μείωση των διαστάσεων και της πληθικότητας, όπως και η συμπίεση των δεδομένων.

- Μετασχηματισμός δεδομένων

Στην παρούσα διαδικασία τα δεδομένα μετατρέπονται ώστε να λάβουν κατάλληλη μορφή για την εφαρμογή της εξόρυξης δεδομένων. Οι στρατηγικές που χρησιμοποιούνται είναι η εξομάλυνση, η συσσωμάτωση, η ομαλοποίηση και η διακριτική ευχέρεια.

- Εξόρυξη δεδομένων

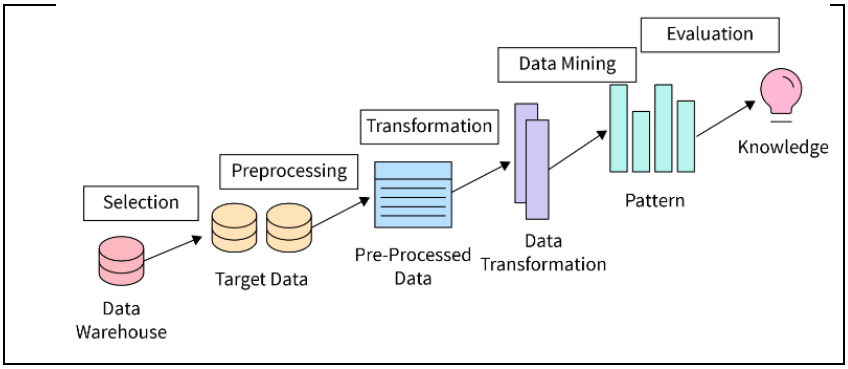
Σ' αυτό το βήμα εφαρμόζονται τεχνικές ταξινόμησης και ομαδοποίησης για τον εντοπισμό ενδιαφερόντων προτύπων και γνώσεων από τα δεδομένα μας.

- Αξιολόγηση προτύπων

Σε αυτό το βήμα γίνεται ο εντοπισμός των πραγματικά ενδιαφερόντων μοτίβων που αντιπροσωπεύουν τη γνώση, με βάση τα μέτρα ενδιαφέροντος.

- Αναπαράσταση γνώσης

Χρησιμοποιώντας τις τεχνικές οπτικοποίησης και αναπαράστασης δεδομένων παρουσιάζεται η γνώση που προέκυψε από τα παραπάνω βήματα. (Han et al., 2012)



Σχήμα 7.1: Βήματα διαδικασίας εξόρυξης δεδομένων

(Πηγή: https://www.scaler.com/topics/kdd-in-data-mining/)

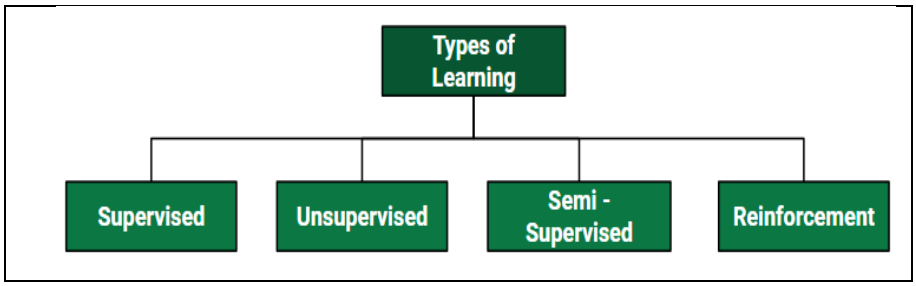
7.2.1. Μηχανική μάθηση

Ως επιστημονικός κλάδος, η μηχανική μάθηση διερευνά την ανάλυση και την κατασκευή αλγορίθμων που μπορούν να εκπαιδεύονται από δεδομένα και να κάνουν

προβλέψεις σε αυτά. Διακρίνεται σε τέσσερις κατηγορίες, οι οποίες αναλύονται στην συνέχεια της ενότητας εκτενέστερα: την Εποπτευόμενη Μηχανική Μάθηση (Supervised Machine Learning), τη Μη Εποπτευόμενη Μηχανική Μάθηση (Unsupervised Machine Learning), τη Μηχανική Μάθηση με Ημι-επίβλεψη (Semi-sypervised Machine Learning) και τέλος την Ενισχυτική Μάθηση (Reinforcement).

7.2.1.1. Είδη μηχανικής μάθησης

Στην παρούσα ενότητα θα αναλυθούν και οι τέσσερις κατηγορίες μηχανικής μάθησης, όπως αυτές έχουν αναλυθεί από τον Bishop (2006). Συνοπτικά στο Σχήμα 7.2 έχουμε τα είδη των τεσσάρων κατηγοριών.



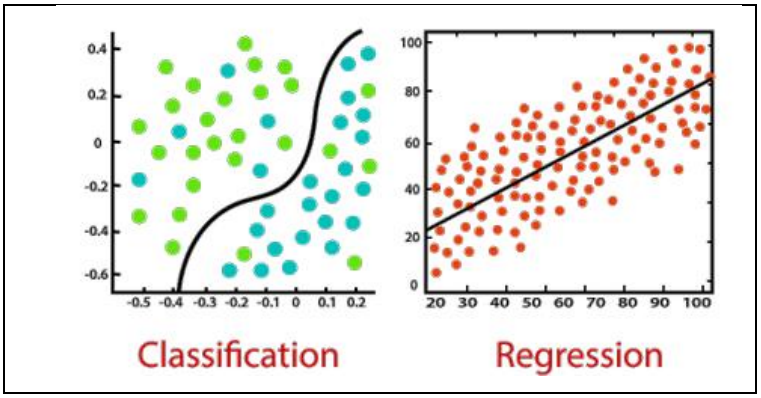
Σχήμα 7.2: Είδη μηχανικής μάθησης

(Πηγή: https://www.geeksforgeeks.org/supervised-machine-learning/)

7.2.1.1.1. Εποπτευόμενη μάθηση (Supervised learning)

Η βασική διαφορά της εποπτευόμενης μάθησης είναι η δυνατότητα εκπαίδευσης σε σχολιασμένα δεδομένα, ή αλλιώς χαρακτηρισμένα (labeled) δεδομένα εκπαίδευσης (training data), δηλαδή δεδομένα όπου η κατηγορία/έξοδός τους είναι γνωστή. Ο αλγόριθμος μαθαίνει να αντιστοιχίζει τις εισόδους στις εξόδους και μπορεί να χρησιμοποιηθεί για να κάνει προβλέψεις σε νέα δεδομένα. Η διαδικασία αυτή χρησιμοποιείται σε δύο διαφορετικούς τύπους προβλημάτων:

- σε προβλήματα ταξινόμησης, που γίνεται χρήση αλγορίθμων με στόχο την ανάθεση δεδομένων ελέγχου (test data) σε συγκεκριμένες κατηγορίες. Στην ουσία μέσω των χαρακτηριστικών κάθε οντότητας ο αλγόριθμος προσπαθεί να καταλήξει σε συμπεράσματα σχετικά με την ταξινόμηση αυτών σε διάφορες κατηγορίες. Οι πιο γνωστοί αλγόριθμοι ταξινόμησης είναι οι γραμμικοί ταξινομητές (linear classifiers), οι μηχανές διανυσμάτων (SVM), ο k-κοντινότερος γείτονας (k-nearest neighbour), το τυχαίο δάσος (random forest) και τα δέντρα απόφασης (decision trees).
- σε προβλήματα παλινδρόμησης (regression), που χρησιμοποιούνται για την κατανόηση της σχέσης μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών.



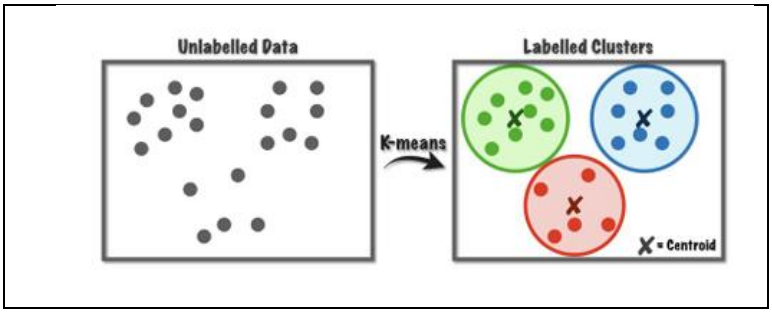
Σχήμα 7.3: Προβλήματα ομαδοποίησης και παλινδρόμησης

(Πηγή: Belkacem, 2021)

7.2.1.1.2. Μη εποπτευόμενη μάθηση (Unsupervised learning)

Η μη εποπτευόμενη μάθηση, χρησιμοποιεί αλγορίθμους με στόχο την ανακάλυψη κρυμμένων μοτίβων ή ομαδοποιήσεων ανάμεσα στα δεδομένα μας χωρίς χρήση χαρακτηρισμένων δεδομένων (unlabeled). Όπως και η προηγούμενη κατηγορία, έτσι και αυτή χρησιμοποιείται σε δύο περιπτώσεις προβλημάτων.

- σε προβλήματα ομαδοποίησης (clustering), που στόχος των διαδικασιών είναι η δημιουργία ομάδων για μη ταξινομημένα δεδομένα που αντιπροσωπεύονται από δομές ή μοτίβα (patterns). Ο δημοφιλέστερος αλγόριθμος ομαδοποίησης, ο οποίος χρησιμοποιείται και στην παρούσα εργασία όπως θα δούμε σε επόμενη παράγραφο, είναι η ομαδοποίηση k μέσων (k-means clustering).
- σε προβλήματα μείωσης των διαστάσεων (dimensionality reduction). Η συγκεκριμένη διαδικασία έχει ως σκοπό τη μείωση του αριθμού των εισερχόμενων (in put data) δεδομένων διατηρώντας παράλληλα στο δυνατότερο επίπεδο την ακεραιότητα τους. Μια από τις πιο γνωστές μεθόδους είναι η ανάλυση των κύριων συνιστωσών (principal components analysis).



Σχήμα 7.4: Η διαδικασίας ομαδοποίησης με k-means

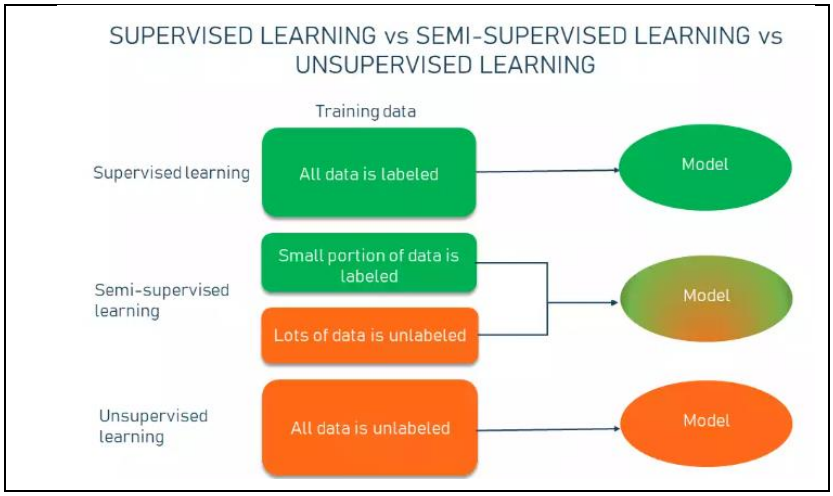
(Πηγή:Jeffares, 2019)

7.2.1.1.3. Ημι-εποπτευόμενη μάθηση (Semi-supervised learning)

Η συγκεκριμένη κατηγορία μηχανικής μάθησης περιλαμβάνει ένα μικρό αριθμό επισημασμένων δεδομένων (labeled) και ένα μεγάλο αριθμό μη επισημασμένων

δεδομένων (unlabeled). Η μέθοδος χρησιμοποιείται σε περιπτώσεις όπου η επισήμανση δεδομένων είναι δαπανηρή ή χρονοβόρα.

Οι διαφορές που δημιουργούνται ανάμεσα στις τρείς πρώτες κατηγορίες μηχανικής μάθησης είναι στα επισημασμένα δεδομένα (training data). Όπως φαίνεται και στο Σχήμα 7.5, στην πρώτη κατηγορία γνωρίζουμε την έξοδο όλων των δεδομένων, στην ημι-εποπτευόμενη μάθηση γνωρίζουμε την έξοδο ενός μικρού συνόλου των δεδομένων και τέλος στην μη εποπτευόμενη δεν γνωρίζουμε την έξοδο κανενός δεδομένου.

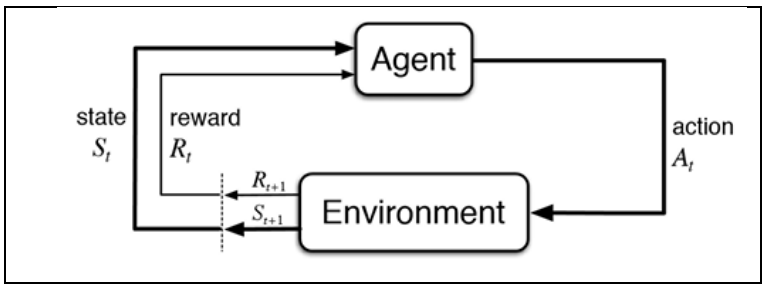


Σχήμα 7.5: Διαφορές ανάμεσα στις κατηγορίες της μηχανικής μάθησης

(Πηγή: https://www.altexsoft.com/blog/semi-supervised-learning/)

7.2.1.1.4. Ενισχυτική μάθηση (Reinforcement learning)

Η ενισχυτική μάθηση σχετίζεται με τη λήψη κατάλληλων αποφάσεων για τη μεγιστοποίηση της ανταμοιβής σε μια συγκεκριμένη περίπτωση. Σ' αυτή την κατηγορία μηχανικής μάθησης δεν υπάρχει απάντηση, αλλά ο ενισχυτικός παράγοντας αποφασίζει τι χρειάζεται να ακολουθήσει ώστε να εκτελέσει τη συγκεκριμένη εργασία. Λόγω της έλλειψης του συνόλου δεδομένων εκπαίδευσης, ο αλγόριθμος είναι αναγκασμένος να μάθει από την εμπειρία του.



Σχήμα 7.6: Η διαδικασίας της ενισχυτικής μάθησης

(Πηγή: De Castro et al., 2020)

7.1.3. Μείωση/επιλογή δεδομένων (Feature selection)

Όπως αναφέραμε και σε προηγούμενη παράγραφο του κεφαλαίου, προτού ξεκινήσουμε να εφαρμόζουμε τους αλγορίθμους της εξόρυξης δεδομένων, θα πρέπει να μειώσουμε το πλήθος των μεταβλητών, σε περίπτωση που αυτό είναι μεγάλο. Στην

παρούσα εργασία έχουμε καταγράψει αρκετές μεταβλητές και επομένως κρίνεται αναγκαίο να γίνει η επιλογή των καταλληλότερων μεταβλητών. Σ' αυτή τη διαδικασία επιλέγεται ένα υποσύνολο μεταβλητών/χαρακτηριστικών από ένα ευρύτερο σύνολο για την δημιουργία ενός προγνωστικού μοντέλου. Αυτό συνήθως γίνεται για να βελτιωθεί η ακρίβεια του μοντέλου, να αποφευχθεί οποιαδήποτε μορφή υπερ-προσαρμογής (over-fitting) και να επιταχυνθεί η διαδικασία εκπαίδευσης του μοντέλου. Για την επιλογή των καταλληλότερων μεταβλητών υπάρχουν δύο βασικές κατηγορίες μεθόδων. Η πρώτη κατηγορία είναι οι μέθοδοι χωρίς επίβλεψη (Unsupervised) και η δεύτερη με επίβλεψη (Supervised). Αυτές οι κατηγορίες αναφέρθηκαν και προηγουμένως ως κατηγορίες μηχανικής μάθησης και στην ουσία έχουν τα ίδια χαρακτηριστικά. Δηλαδή, στις μεθόδους με επίβλεψη για την επιλογή των καταλληλότερων μεταβλητών χρησιμοποιείται ως γνωστή η κατηγορία εξόδου και στόχος είναι οι μεταβλητές που θα επιλεχθούν να αυξάνουν την αποτελεσματικότητα του μοντέλου. Αντίθετα, στις μεθόδους χωρίς επίβλεψη δεν χρησιμοποιείται η κατηγορία εξόδου για την επιλογή των χαρακτηριστικών. Πιο συγκεκριμένα, οι μέθοδοι με επίβλεψη μπορούν να χωριστούν σε ακόμη τρείς υπό κατηγορίες:

- Μέθοδος φίλτρου (Filter)

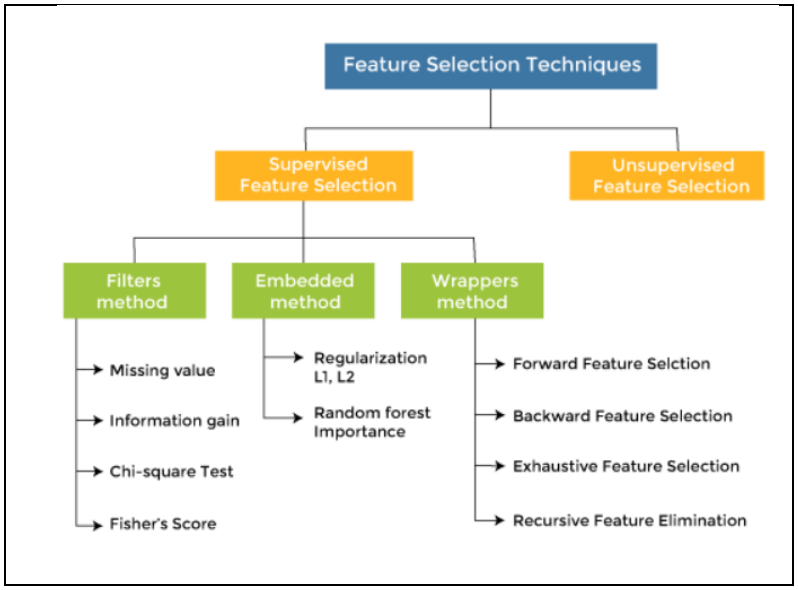
Σ' αυτή τη μέθοδο, οι μεταβλητές διαλέγονται βάσει της σχέσης τους με την κατηγορία εξόδου, ή ανάλογα με τη συσχέτιση που έχουν με αυτή. Η συσχέτιση χρησιμοποιείται για να ελεγχθεί αν οι μεταβλητές συσχετίζονται θετικά ή αρνητικά με τη μεταβλητή εξόδου. Για τον έλεγχο συσχέτισης χρησιμοποιούνται παραδείγματος χάρη το Chi-Square Test, το Fisher's Score κ.ά

- Μέθοδος περιτύλιξης (Wrapper)

Σ' αυτή την κατηγορία μεθόδων τα δεδομένα πρώτα διαχωρίζονται σε υποσύνολα και έπειτα εκπαιδεύεται ένα μοντέλο βάσει αυτών. Με βάση την έξοδο του μοντέλου, προαφαιρούνται χαρακτηριστικά και στη συνέχεια εκπαιδεύεται ξανά το μοντέλο. Διαμορφώνει τα υποσύνολα χρησιμοποιώντας μια «άπληστη» προσέγγιση και αξιολογεί την ακρίβεια όλων των πιθανών συνδυασμών των χαρακτηριστικών. Τέτοιες μέθοδοι είναι η επιλογή προς τα εμπρός (Forward Selection), η εξάλειψη προς τα πίσω (Backwards Elimination) κ.λπ.

- Ενσωματωμένη μέθοδος (Embedded)

Η συγκεκριμένη μέθοδος είναι ένας συνδυασμός των δύο προηγούμενων για τη δημιουργία του καλύτερου υποσυνόλου. Τέτοιες μέθοδοι είναι οι Lasso και Ridge παλινδρομήσεις. (Καμίτσης, 2023)



Σχήμα 7.7: Μέθοδοι για την μείωση δεδομένων

(Πηγή:https://www.javatpoint.com/feature-selection-techniques-in-machine-learning)

Στην παρούσα εργασία θα χρησιμοποιηθούν δύο μέθοδοι για την επιλογή χαρακτηριστικών ανάμεσα στις διαθέσιμες μεταβλητές των δεδομένων: η Boruta και η k-means.

Μέθοδος Boruta

Η πρώτη απ' αυτές είναι η μέθοδος Boruta, που κατατάσσεται στην κατηγορία wrapper. Ο αλγόριθμος που στηρίζεται αυτή η μέθοδος παρουσιάστηκε από τους Πολωνούς ερευνητές Kursa και Rudnicki το 2010. Οι ερευνητές βασίστηκαν στον αλγόριθμο του τυχαίου δάσους (Random Forest Classifier), ο οποίος αναλύεται στην παράγραφο 7.4.1.2. Πιο συγκεκριμένα προσπαθεί να κατονομάσει τις πιο σημαντικές μεταβλητές που μπορεί να υπάρχουν στο σύνολο δεδομένων αναφορικά πάντα με την μεταβλητή απόκρισης. Στο πρώτο του βήμα ο αλγόριθμος, προσθέτει τυχαιότητα στο σύνολο δεδομένων που έχουμε διαθέσιμο, δημιουργώντας ανακατεμένα αντίγραφα όλων των χαρακτηριστικών που ονομάζονται σκιώδη χαρακτηριστικά (shadow features). Έπειτα, αφού εκπαιδεύσει έναν ταξινομητή τυχαίου δάσους, εφαρμόζει ένα μέτρο σημαντικότητας χαρακτηριστικών, όπως το Mean Decrease Accuracy, ώστε να αξιολογήσει τη σημασία της κάθε μεταβλητής. Σε κάθε επανάληψη, ο αλγόριθμος συγκρίνει τα Z-scores των σκιωδών χαρακτηριστικών και των αρχικών χαρακτηριστικών, για να διαπιστώσει αν τα δεύτερα είχαν καλύτερη απόδοση από τα πρώτα. Εάν ναι, ο αλγόριθμος θα χαρακτηρίσει το χαρακτηριστικό ως σημαντικό. Στην ουσία, ο αλγόριθμος προσπαθεί να επικυρώσει τη σημασία του χαρακτηριστικού συγκρίνοντάς το με τα τυχαία ανακατεμένα αντίγραφα, γεγονός που αυξάνει την ανθεκτικότητα. Ο αλγόριθμος σταματάει μετά από έναν προκαθορισμένο αριθμό επαναλήψεων, ή αν όλα τα χαρακτηριστικά έχουν είτε επιβεβαιωθεί είτε απορριφθεί. (Kursa & Rudnicki, 2010)

Μέθοδος επιλογής k καλύτερων χαρακτηριστικών (SelectKBest)

Η μέθοδος SelectKBest είναι άλλη μια μέθοδος επιλογής των καλύτερων δεδομένων ανάμεσα στις μεταβλητές που έχουμε προς επεξεργασία. Πιο συγκεκριμένα, η μέθοδος λαμβάνει ως παράμετρο μια συνάρτηση, η οποία πρέπει να εφαρμοστεί σε ένα ζεύγος (X, y). Αυτή η συνάρτηση πρέπει να επιστρέψει έναν πίνακα με βαθμολογίες, μία για κάθε χαρακτηριστικό X[:,i] του X. Εφόσον γίνουν οι μετρήσεις, η μέθοδος διατηρεί τις μεταβλητές με τις υψηλότερες επιδόσεις. Ως συνάρτηση στην παρούσα εργασία έχει επιλεγεί η chi2 επομένως το SelectKBest θα υπολογίσει το στατιστικό chi2 μεταξύ κάθε χαρακτηριστικού του X και y και αναλόγως την τιμή θα γίνει η επιλογή των καλύτερων k μεταβλητών. Μια μικρή τιμή θα σημαίνει ότι το χαρακτηριστικό είναι ανεξάρτητο από το y. Μια μεγάλη τιμή θα σημαίνει ότι το χαρακτηριστικό δεν σχετίζεται τυχαία με το y, και έτσι είναι πιθανό να παρέχει σημαντικές πληροφορίες.

7.1.4. Κατηγοριοποίηση/Ταξινόμηση (Classification)

Η ταξινόμηση είναι μία θεμελιώδης τεχνική στη μηχανική μάθηση και περιλαμβάνει την πρόβλεψη της κλάσης ή της κατηγορίας των δεδομένων. Οι μέθοδοι ταξινόμησης είναι αλγόριθμοι που χρησιμοποιούν επισημασμένα δεδομένα (labelled data) για να ανακαλύψουν ένα όριο απόφασης που μπορεί να χρησιμοποιηθεί για την ταξινόμηση νέων, μη επισημασμένων δεδομένων (unlabelled data) σε μία ή περισσότερες προκαθορισμένες κλάσεις. Για τον λόγο αυτό στις συγκεκριμένες μεθόδους θα πρέπει να γίνει διαχωρισμός των δεδομένων σε σύνολα εκπαίδευσης (training set) και ελέγχου (testing test). Με αυτόν τον τρόπο ένα μέρος των δεδομένων θα χρησιμοποιείται ώστε να εκπαιδεύεται το μοντέλο και στη συνέχεια να δοκιμάζεται στο άλλο μέρος. Στην παρούσα εργασία τα δεδομένα χωρίστηκαν σε αναλογία 70% – 30% ανάμεσα στα δεδομένα εκπαίδευσης και ελέγχου. Στόχος είναι η εκμάθηση ενός μοντέλου που μπορεί να προβλέψει με ακρίβεια την κατηγορία των νέων παρατηρήσεων. Συνολικά, οι αλγόριθμοι ταξινόμησης είναι ισχυρά εργαλεία για την επίλυση ποικίλων προβλημάτων. Τέλος, στην παρούσα παράγραφο θα αναλυθούν και ορισμένες τεχνικές αξιολόγησης των μοντέλων ταξινόμησης, ώστε να είμαστε σε θέση να τις συγκρίνουμε.

7.1.4.1. Μέθοδοι κατηγοριοποίησης

Αρχικά δίνεται το θεωρητικό υπόβαθρο των μεθόδων και στη συνέχεια γίνεται χρήση των αλγορίθμων και παρουσίαση των αποτελεσμάτων τους επάνω στα δεδομένα. Στην παρούσα εργασία θα αναλυθούν οι τρείς παρακάτω μέθοδοι:

- Κ Κοντινότεροι Γείτονες (K-Nearest Neighbors, KNN)
- Τυχαίο Δάσος (Random Forest)
- Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines, SVM)

7.1.4.1.1. Κ Κοντινότεροι Γείτονες (K-Nearest Neighbors, KNN)

Η μέθοδος των k κοντινότερων γειτόνων είναι ένας αλγόριθμος κατηγοριοποίησης που προβλέπει την κλάση ενός σημείου βάσει της κλάσης των k κοντινότερων γειτόνων του στα δεδομένα εκπαίδευσης, ενώ το k είναι μια παράμετρος που ορίζει ο εκάστοτε χρήστης. Ο συγκεκριμένος αλγόριθμος είναι μια μη παραμετρική μέθοδος, που σημαίνει ότι δεν κάνει υποθέσεις σχετικά με την κατανομή των δεδομένων ή τη σχέση μεταξύ των χαρακτηριστικών και της μεταβλητής-στόχου.

Ο τρόπος λειτουργίας του αλγορίθμου εξαρτάται από τις αποστάσεις του νέου σημείου των δεδομένων και κάθε σημείου που βρίσκεται στα δεδομένα εκπαίδευσης. Αρχικά, υπολογίζεται η συγκεκριμένη απόσταση με χρήση κάποιου από τα συνήθη μέτρα απόστασης. Σ' αυτά τα μέτρα ανήκουν η Ευκλείδεια απόσταση, η απόσταση Manhattan και η απόσταση Minkowski. Πιο συγκεκριμένα τα είδη των αποστάσεων ορίζονται παρακάτω:

Ευκλείδεια απόσταση

Υπολογίζει το μήκος της ευθείας γραμμής μεταξύ δύο σημείων με χρήση του τύπου

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Απόσταση Manhattan

Η συγκεκριμένη απόσταση μετρά την απόλυτη τιμή μεταξύ δύο σημείων και συνήθως δίνει ίδια αποτελέσματα με την ευκλείδεια απόσταση, εκτός από την περίπτωση που υπάρχουν στα δεδομένα έκτροπες παρατηρήσεις. Ο τύπος υπολογισμού της είναι:

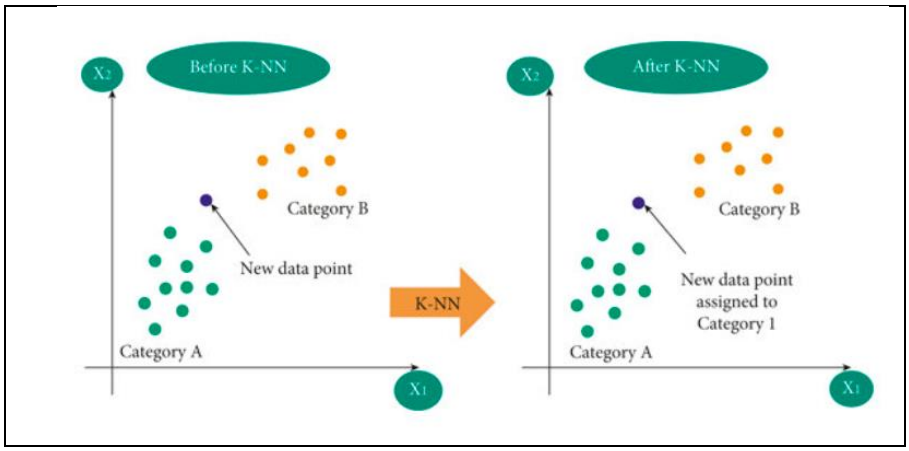
$$d(x,y) = \sum_{i=1}^{m} |x_i - y_i|$$

Απόσταση Minkowski

Η απόσταση Minkowski αποτελεί μια γενικευμένη μορφή των προηγούμενων δύο αποστάσεως που αναλύσαμε. Η παράμετρος λ (λ≥1) επιτρέπει τη δημιουργία άλλων μετρικών απόστασης. Για λ=2, προκύπτει η ευκλείδεια απόσταση, ενώ για λ=1 έχουμε την απόσταση Manhattan. Ο τύπος της είναι:

$$d(x,y) = (\sum_{i=1}^{n} |x_i - y_i|^{\lambda})^{1/\lambda}$$

Εφόσον υπολογιστούν οι αποστάσεις, ο αλγόριθμος προσδιορίζει τους k πλησιέστερους γείτονες του νέου σημείου με βάσει αυτές τις αποστάσεις. Στη συνέχεια, η κλάση του νέου σημείου δεδομένων καθορίζεται με ψηφοφορία πλειοψηφίας μεταξύ των κλάσεων των k πλησιέστερων γειτόνων του. Εάν k = 1, το νέο σημείο δεδομένων λαμβάνει απλώς την κλάση του πλησιέστερου γείτονά του. (Harrison, 2018)

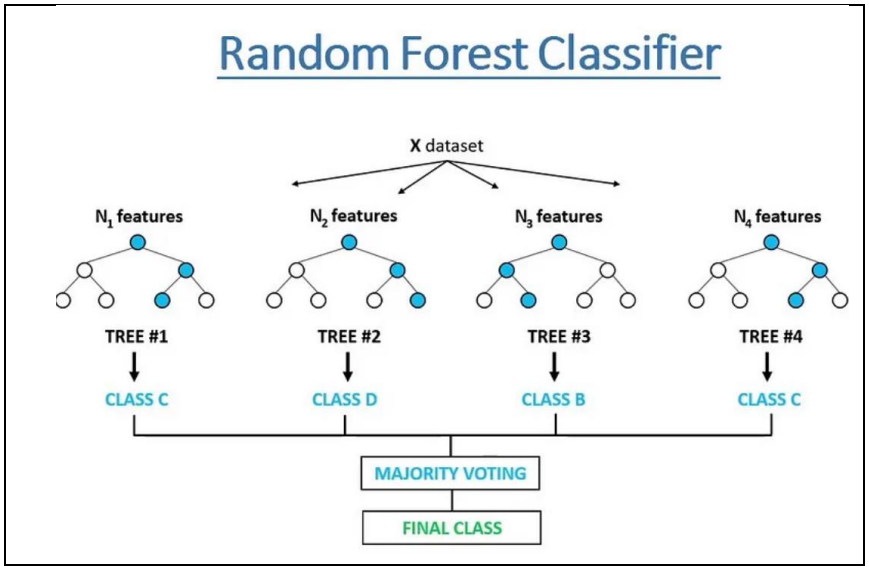


Σχήμα 7.8: Λειτουργία μεθόδου KNN Classification

(Πηγή:Khalid et al., 2022)

7.1.4.1.2. Τυχαία Δάση (Random Forests)

Ένας ακόμα αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται ως διαδικασία ταξινόμησης είναι αυτός του Τυχαίου Δάσους. Αυτή τη διαδικασία την παρουσίασαν πρώτη φορά οι Leo Breiman και Adele Culter το 2001. Πρόκειται για μια τεχνική που σκοπό έχει να βελτιώσει την ακρίβεια των προβλέψεων συνδυάζοντας πολλαπλά δέντρα αποφάσεων και τη μείωση της υπερ-προσαρμογής του. Στο Σχήμα 7.9 φαίνεται η διαδικασία του αλγορίθμου για την κατηγοριοποίηση ενός συνόλου δεδομένων.



Σχήμα 7.9: Λειτουργία μεθόδου Random Forest Classification

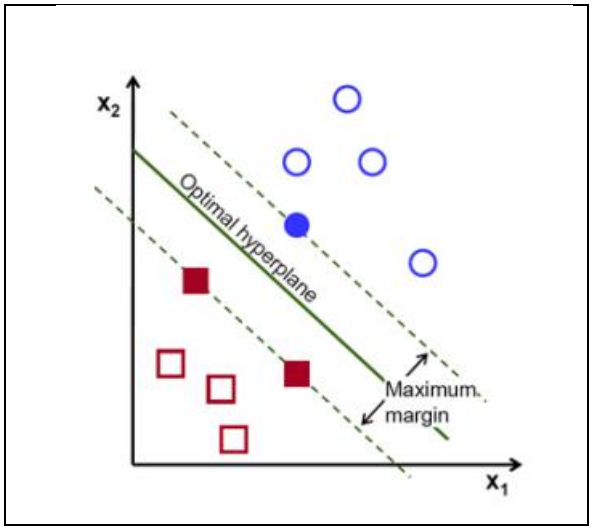
(Πηγή:https://medium.com/@mrmaster907/introduction-random-forest-classification-by-example-6983d95c7b91)

Σε ένα τυχαίο δάσος, κάθε δέντρο απόφασης εκπαιδεύεται σε ένα τυχαίο υποσύνολο των δεδομένων εισόδου. Αρχικά η διαδικασία δημιουργεί πολλά νέα σύνολα εκπαίδευσης και έστω ότι ο αριθμός των νέων συνόλων είναι K. Με βάση κάθε σύνολο του K_i, από τα K, κατασκευάζεται ένα δέντρο με την χρήση ενός ταξινομητή βάσης. Η διαφορά από άλλες μεθόδους είναι πως σε κάθε νέα προσθήκη κόμβου δεν χρησιμοποιούνται όλα τα χαρακτηριστικά που είναι διαθέσιμα αλλά μόνο ένα μέρος αυτών. Ο αριθμός των χαρακτηριστικών που χρησιμοποιείται είναι σταθερός κατά την εκτέλεση του αλγορίθμου. Η τελική πρόβλεψη καθορίζεται κατά κύριο λόγο βάσει μιας ψηφοφορίας πλειοψηφίας, όπου η πιο συχνά προβλεπόμενη κλάση σε όλα τα δέντρα απόφασης επιλέγεται ως τελική έξοδος. (Breiman, 2001)

7.1.4.1.3. Μηχανές Διανυσματικής Υποστήριξης (Support Vector Machines)

Ο τελευταίος αλγόριθμος που θα χρησιμοποιηθεί για τη διαδικασία της κατηγοριοποίησης στην παρούσα εργασία είναι οι μηχανές διανυσματικής υποστήριξης, ενώ ο παρών αλγόριθμος μπορεί να χρησιμοποιηθεί και για εργασίες παλινδρόμησης. Ο συγκεκριμένος αλγόριθμος έχει δημιουργηθεί γύρω από την ιδέα της εύρεσης ενός υπερ-επιπέδου που διαχωρίζει στοιχεία δύο κλάσεων σε έναν χώρο μεγαλύτερων διαστάσεων. Στόχος του είναι να καθορίσει το υπερ-επίπεδο που μεγιστοποιεί το περιθώριο, δηλαδή την απόσταση ανάμεσα στο υπερ-επίπεδο και στα

πλησιέστερα σημεία από κάθε κλάση, μεταξύ των δύο κλάσεων. Το υπερ-επίπεδο το ορίζουν τα σημεία των δεδομένων που βρίσκονται πλησιέστερα σε αυτό, τα οποία ονομάζονται διανύσματα υποστήριξης. (Hastie et al., 2009)



Σχήμα 7.10: Γραφική απεικόνιση της επιλογής του μέγιστου περιθωρίου μεταξύ των σημείων δεδομένων των κατηγοριών

(Πηγή: Fidan et al., 2020)

Αρχικά, αυτός ο αλγόριθμος μετασχηματίζει τα δεδομένα σε έναν χώρο υψηλότερων διαστάσεων, όπου είναι πιθανό τα δεδομένα να είναι γραμμικά διαχωρίσιμα. Για να υλοποιηθεί αυτός ο μετασχηματισμός χρησιμοποιείται μια συνάρτηση πυρήνας. Η συνάρτηση πυρήνας απεικονίζει τα δεδομένα στο χώρο υψηλότερων διαστάσεων, δίχως να υπολογίζει τις συντεταγμένες των δεδομένων στον καινούργιο χώρο. Η επιλογή αυτής της συνάρτησης είναι κρίσιμη, εφόσον καθορίζει το σχήμα του ορίου απόφασης. Παρακάτω θα αναλυθούν κάποιες συναρτήσεις πυρήνα. Έπειτα το SVM βρίσκει το υπερ-επίπεδο που μεγιστοποιεί το περιθώριο μεταξύ των κλάσεων επιλύοντας ένα πρόβλημα βελτιστοποίησης με περιορισμούς, όπου επιθυμία είναι η μεγιστοποίηση του περιθωρίου με τον περιορισμό ότι τα σημεία των δεδομένων ταξινομούνται σωστά.

Ορισμένες από τις πιο σύνηθες συναρτήσεις πυρήνα είναι οι:

- **<u>Πολυωνυμική συνάρτηση πυρήνα (Polynomial Kernel Function)</u>**

Αυτή μετασχηματίζει τα σημεία δεδομένων χρησιμοποιώντας το γινόμενο τελείας και μετασχηματίζοντας τα δεδομένα σε μια "n-διάσταση", δηλαδή ο μετασχηματισμός θα είναι είτε τετραγωνικό γινόμενο είτε υψηλότερο. Συνεπώς, τα δεδομένα αναπαρίστανται σε χώρο υψηλότερων διαστάσεων χρησιμοποιώντας τα νέα μετασχηματισμένα σημεία. Ο τύπος που χρησιμοποιεί αυτή η συνάρτηση πυρήνα είναι:

$$\mathrm{k}(x, y) = (\alpha x^T y + c)^d$$

όπου ο βαθμός της πολυωνυμικής συνάρτησης συμβολίζεται με d και είναι μια παράμετρος που πρέπει να ρυθμιστεί, ενώ οι παράμετροι α και c ελέγχουν την επιρροή των όρων υψηλότερης τάξης.

- **Η συνάρτηση ακτινωτής βάσης (Radial Basis Function, RBF)**

Η συνάρτηση αυτή αρχικά μετασχηματίζει τα δεδομένα αναπαριστώντας τα σε άπειρες διαστάσεις και έπειτα τα ταξινομεί χρησιμοποιώντας τον σταθμισμένο κοντινότερο γείτονα. Η ακτινωτή συνάρτηση μπορεί να είναι είτε Gaussian είτε Laplace. Αυτό εξαρτάται από μια υπερπαράμετρο γνωστή ως γάμμα, η οποία ελέγχει το πλάτος της «Γκαουσιανής» συνάρτησης και πρέπει να ρυθμιστεί. Ο τύπος της είναι:

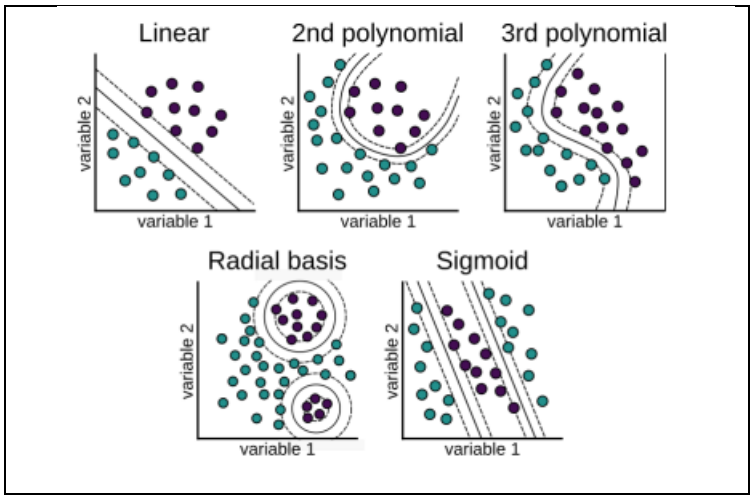
$$\mathrm{k}(x, y) = exp(\gamma \|x - y\|^2)$$

- **Ο γραμμικός πυρήνας (Linear Kernel)**

Αυτή συνάρτηση απλά αναπαριστά τα σημεία δεδομένων χρησιμοποιώντας μια γραμμική σχέση. Πιο συγκεκριμένα υπολογίζει το τετραγωνικό γινόμενο μεταξύ δύο διανυσμάτων εισόδου στον αρχικό χώρο τους. Έχει τύπο:

$$k(x, y) = x^T y$$

(Schölkopf & Smola, 2002)



Σχήμα 7.11: Γραφική απεικόνιση των τεχνασμάτων των πυρήνων του SVM

(Πηγή: Awais et al., 2022)

7.1.4.2. Τεχνικές αξιολόγησης των μεθόδων κατηγοριοποίησης

Ένα σημαντικό κομμάτι της ταξινόμησης είναι και η αξιολόγηση του μοντέλου που δημιουργείται από τη χρήση των μεθόδων. Οι συνήθεις μετρικές περιλαμβάνουν: την ορθότητα (accuracy), την ακρίβεια (precision), την ανάκληση (recall), το F1 score και τον πίνακα σύγχυσης (confusion matrix).

Σκοπός του πίνακα σύγχυσης είναι η οπτικοποίηση των ποσοστών αληθώς θετικών, αληθώς αρνητικών, ψευδώς θετικών και ψευδώς αρνητικών του μοντέλου. Πιο συγκεκριμένα, με τον όρο αληθώς θετικό (TP) αναφέρουμε τον αριθμό των παρατηρήσεων που προβλέφθηκαν σωστά ως θετικές από το μοντέλο. Με τον όρο αληθώς αρνητικό (TN) εννοούμε τον αριθμό των παρατηρήσεων που το μοντέλο προβλέπει σωστά ως αρνητικές. Ο όρος ψευδώς θετικό (FP) αναφέρεται στον αριθμό

των περιπτώσεων που προβλέφθηκαν ως θετικές από το μοντέλο, αλλά στην πραγματικότητα ήταν αρνητικές. Τέλος, ο τέταρτος όρος ψευδώς αρνητικό (FN) αναφέρεται στον αριθμό των περιπτώσεων που προβλέφθηκαν ως αρνητικές από το μοντέλο, αλλά ήταν στην πραγματικότητα θετικές. Ο όρος «θετικές» αναφέρεται στις τιμές της κατηγορικής μεταβλητής που ισούνται με 1, ενώ ο όρος «αρνητικές» όταν έχουμε τιμή ίση με 0.

Με βάση τις τιμές του πίνακα σύγχυσης μπορούν να υπολογιστούν οι μετρικές που αναφέραμε και προηγουμένως.

Ορθότητα (Accuracy)

Αυτή μετρά τη συνολική ορθότητα του μοντέλου και υπολογίζεται από τον τύπο:

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Ακρίβεια (Precision)

Αυτή η μετρική μετρά το ποσοστό των αληθώς θετικών μεταξύ όλων των θετικών προβλέψεων και υπολογίζεται από τον παρακάτω τύπο:

$$\frac{TP}{TP + FP}$$

Ανάκληση (Recall)

Η συγκεκριμένη τιμή μετρά την αναλογία των αληθώς θετικών μεταξύ όλων των πραγματικών θετικών προβλέψεων και υπολογίζεται από τον τύπο:

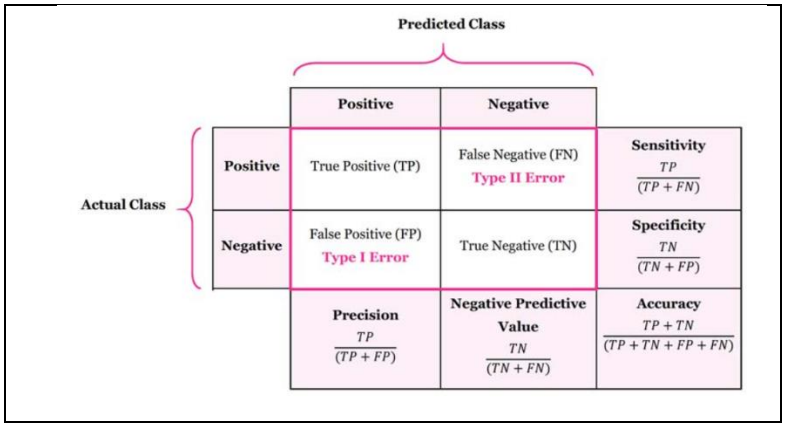
$$\frac{TP}{TP + FN}$$

Αυτή η μετρική αναφέρεται και ως sensitivity, Σχήμα 7.12.

Βαθμολογία F1 (F-score)

Πρόκειται για τον αρμονικό μέσο όρο της ακρίβειας και της ανάκλησης και υπολογίζεται ως:

$$2 * \frac{Precision * Recall}{Precision + Recall}$$



Σχήμα 7.12: Παράδειγμα ενός πίνακα σύγχυσης

(Πηγή: https://manisha-sirsat.blogspot.com/2019/04/confusion-matrix.html)

7.1.5. Ανάλυση κατά συστάδες (Clustering)

Η ανάλυση κατά συστάδες (συσταδοποίηση) είναι μια από τις βασικότερες τεχνικές στην εξόρυξη δεδομένων και τη μηχανική μάθηση, η οποία περιλαμβάνει την κατανομή ενός συνόλου αντικειμένων σε ομάδες ή συστάδες, έτσι ώστε τα αντικείμενα εντός της κάθε ομάδας να μοιάζουν περισσότερο μεταξύ τους, να είναι δηλαδή πιο ομοιογενή, σε σχέση με εκείνα των υπολοίπων ομάδων. Η συσταδοποίηση χρησιμοποιείται για διάφορους λόγους, όπως ο εντοπισμός φυσικών ομαδοποιήσεων παρόμοιων αντικειμένων, η ανακάλυψη ακραίων τιμών ή ανωμαλιών σε δεδομένα (outlier detection) και η μείωση της διαστατικότητας σε μεγάλα σύνολα δεδομένων. Η επιλογή του αλγορίθμου εξαρτάται από τη φύση των δεδομένων και τους στόχους της ανάλυσης. (Tan et al., 2018)

Για να γίνει η υλοποίηση των μεθόδων συσταδοποίησης, θα πρέπει να οριστεί ένα μέτρο απόστασης που θα μετράει την ομοιότητα ή τη διαφορετικότητα μεταξύ των παρατηρήσεων. Πιο συγκεκριμένα, αυτό καθορίζει τον τρόπο υπολογισμού της απόστασης μεταξύ δύο οποιωνδήποτε παρατηρήσεων στο σύνολο δεδομένων. Προφανώς υπάρχουν διάφορα μέτρα απόστασης, ορισμένα εκ των οποίων έχουν ήδη αναλυθεί στην παράγραφο 7.4.1.1.

7.1.5.1. Μέθοδοι συσταδοποίησης

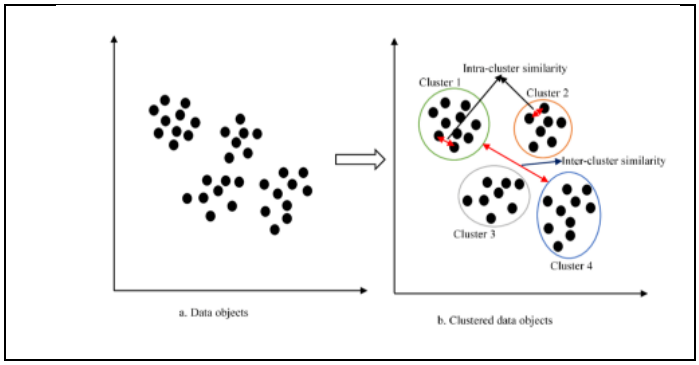
Στην παρούσα εργασία θα χρησιμοποιηθούν οι δύο παρακάτω αλγόριθμοι συσταδοποίησης:

- K-Means
- Birch

7.1.5.1.1. Αλγόριθμος K-Means για συσταδοποίηση

Ο αλγόριθμος K-means είναι ένας από τους πιο δημοφιλείς αλγορίθμους ομαδοποίησης. Ο αλγόριθμος αποσκοπεί στην κατάτμηση ενός δεδομένου συνόλου δεδομένων σε k διακριτές, μη επικαλυπτόμενες συστάδες με βάση την ομοιότητα των σημείων δεδομένων.

Για την υλοποίησή του αρχικά ο χρήστης πρέπει να προσδιορίσει τα k αρχικά κέντρα τυχαία από το σύνολο δεδομένων. Έπειτα, κάθε σημείο δεδομένων ανατίθεται στη συστάδα της οποίας το κέντρο είναι πλησιέστερα σε αυτό, χρησιμοποιώντας ένα μέτρο απόστασης απ' αυτά που αναφέραμε σε προηγούμενη παράγραφο. Τα κέντρα των k συστάδων ενημερώνονται με τον υπολογισμό του μέσου όρου όλων των σημείων που έχουν ανατεθεί στη συγκεκριμένη συστάδα. Τέλος, οι δύο προηγούμενες κινήσεις επαναλαμβάνονται μέχρι σύγκλισης, δηλαδή μέχρι η ανάθεση των σημείων στις συστάδες να μην αλλάζει πλέον. Ο συγκεκριμένος έλεγχος εκτελείται πολλές φορές, διότι τα αποτελέσματα του σχετίζονται σημαντικά με την αρχική επιλογή των κέντρων και αναλόγως την επιλογή μπορεί να αυξηθούν οι πιθανότητες εύρεσης μιας καλής λύσης.



Σχήμα 7.13: Διαγραμματική απεικόνιση της ομαδοποίησης με k-means

(Πηγή: Ezugwu et al., 2021)

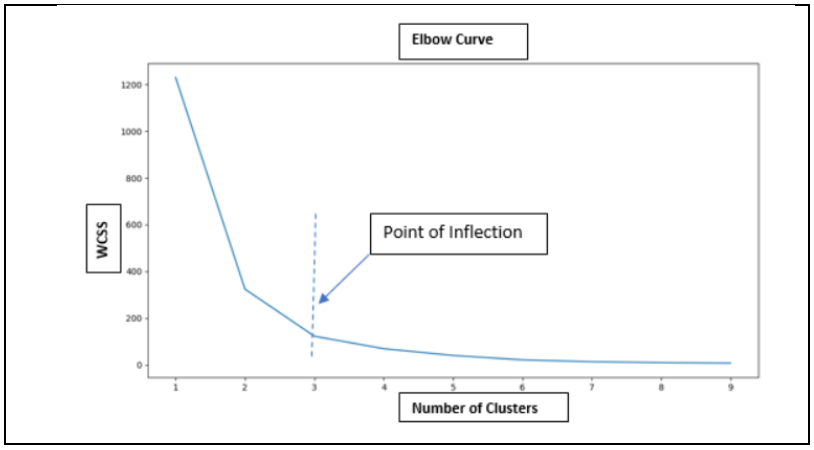
Όπως είναι φανερό, ένα σημαντικό μειονέκτημα της μεθόδου είναι ο προσδιορισμός του αριθμού των κέντρων, ο οποίος δεν είναι γνωστός από την αρχή. Ωστόσο για την επιλογή του k μπορεί να χρησιμοποιηθεί η μέθοδος του αγκώνα.

Μέθοδος αγκώνα (Elbow Method)

Ειδικότερα, η συγκεκριμένη μέθοδος στηρίζεται στη σχεδίαση του αθροίσματος των τετραγώνων εντός συστάδας (within-cluster sum of squares - WCSS) ως συνάρτηση του αριθμού συστάδων k ώστε να βρεθεί ένας αγκώνας στο διάγραμμα. Το άθροισμα αυτό υπολογίζεται από τον τύπο:

$$\text{WCSS} = \sum_i \sum_j ||x_i - c_j||^2$$

όπου xi είναι η i-οστή παρατήρηση, cj είναι το κέντρο της j-οστής συστάδας και ||.|| δηλώνει την ευκλείδεια απόσταση. Στο σχήμα 7.13 παρουσιάζεται ένα παράδειγμα της επιλογής του k μέσω της διαδικασίας του αγκώνα. Όπως φαίνεται σ' αυτό, το k που επιλέγεται είναι ίσο με 3 καθώς βλέπουμε ότι μετά από αυτή τη τιμή, η γραμμή που αντιστοιχεί στο WCSS έχει όλο και μικρότερη κλήση, καθώς μεγαλώνουν οι τιμές του k.



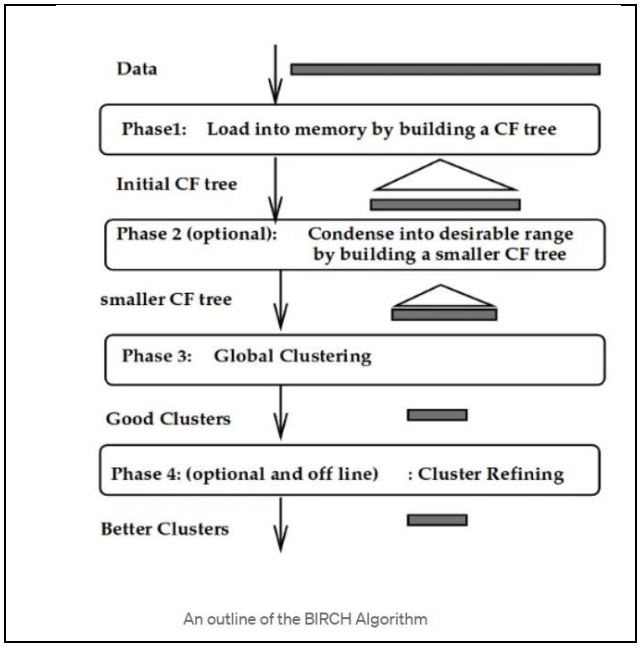
Σχήμα 7.14: Παράδειγμα μεθόδου αγκώνα

(Πηγή: Khan, 2021)

7.1.5.1.2. Αλγόριθμος Birch (Balanced Iterative Reducing and Clustering using Hierarchies) για συσταδοποίηση

Ο δεύτερος αλγόριθμος που θα χρησιμοποιηθεί στην παρούσα εργασία για την διαδικασία της συσταδοποίησης είναι ο Birch. Ο αλγόριθμος αυτός είναι γνωστός και ως αλγόριθμος δύο βημάτων διότι χωρίζεται σε δύο μικρότερες διαδικασίες.

Η πρώτη διαδικασία του αλγορίθμου περιέχει την δημιουργία του δέντρου της συσταδοποίησης (CF tree), στο οποίο αποθηκεύονται οι σχετικές πληροφορίες και στόχος της είναι να ελαχιστοποιήσει την χωρητικότητα των πληροφοριών ενός μεγάλου συνόλου δεδομένων (large dataset). Για την ελαχιστοποίηση της μνήμης ο αλγόριθμος συνοψίζει τις πληροφορίες των δεδομένων και τις καταχωρεί ως Clustering Feature (CF). Αυτό το σύμπλεγμα περιέχει μία τριπλή εισαγωγή CF = (N, LS, SS), όπου το N είναι ο αριθμός των σημείων που περιέχονται στο σύμπλεγμα αναφοράς, το LS είναι το γραμμικό άθροισμα των N σημείων και το SS είναι το άθροισμα τετραγώνων των σημείων που περιέχονται στο σύμπλεγμα. Κάθε σύμπλεγμα που δημιουργείται στη συνέχεια προστίθεται ως κόμβος στο δέντρο συσταδοποίησης. Επομένως κάθε φύλλο του δέντρου δεν αποτελεί ένα απλό σημείο του συνόλου δεδομένων αλλά μια συστάδα αυτού και περιέχει τις πληροφορίες όλης της συστάδας. Κάθε καταχώρηση σε ένα δέντρο περιέχει έναν δείκτη σ' έναν θυγατρικό κόμβο (υποσυστάδα του παιδιού του). Υπάρχει ένας μέγιστος αριθμός καταχωρήσεων σε κάθε κόμβο φύλλου, όπου αυτός ο αριθμός ονομάζεται κατώφλι. Το μέγεθος του δέντρου είναι συνάρτηση του κατωφλίου, όπου όσο μεγαλύτερος είναι ο αριθμός του κατωφλίου, τόσο μικρότερο είναι το δέντρο. Κάθε καινούργιος κόμβος εισάγεται στο πλησιέστερο στοιχείο του δέντρου. Σημαντική παρατήρηση σε αυτή την διαδικασία είναι πως ο αλγόριθμος την εκτελεί μόνο μια φορά.



Σχήμα 7.15: Παράδειγμα μεθόδου Birch

(Πηγή: Benzer, 2022)

Εφόσον έχει δημιουργηθεί το δέντρο συσταδοποίησης, ο αλγόριθμος εκτελεί τη δεύτερη διαδικασία, η οποία ονομάζεται Global Clustering. Σ' αυτή τη διαδικασία ο αλγόριθμος απλά προσπαθεί να κάνει καλύτερη την ποιότητα της συσταδοποίησης στην οποία έχει καταλήξει κατά τη διάρκεια δημιουργίας του δέντρου, επαναλαμβάνοντας τα προηγούμενα βήματα (Zhang et al., 1996).

7.5.2. Τεχνικές αξιολόγησης των μεθόδων συσταδοποίησης

Υπάρχουν διάφοροι μέθοδοι για την αξιολόγηση της ομαδοποίησης, συμπεριλαμβανομένων των εσωτερικών, των εξωτερικών και των σχετικών μέτρων. Η επιλογή του μέτρου αξιολόγησης εξαρτάται από την εκάστοτε εφαρμογή και τους στόχους της συσταδοποίησης.

- Τα εσωτερικά μέτρα αξιολογούν την ποιότητα της ομαδοποίησης με βάση αποκλειστικά τα δεδομένα και τον αλγόριθμο ομαδοποίησης, χωρίς να χρησιμοποιούν εξωτερικές πληροφορίες ή τις ετικέτες/κλάσεις που συνοδεύουν κάθε σημείο/παρατήρηση.

- Τα εξωτερικά μέτρα αξιολογούν την ποιότητα της συσταδοποίησης συγκρίνοντας τα αποτελέσματα της συσταδοποίησης με κάποιο εξωτερικό κριτήριο, όπως οι ετικέτες των κλάσεων στις οποίες ανήκουν τα σημεία.

- Τα σχετικά μέτρα αξιολογούν την ποιότητα της συσταδοποίησης συγκρίνοντας τα αποτελέσματα διαφορετικών αλγορίθμων συσταδοποίησης ή διαφορετικών ρυθμίσεων για τις παραμέτρους του ίδιου αλγορίθμου.

Στην παρούσα εργασία χρησιμοποιούνται για την αξιολόγηση των αποτελεσμάτων μας ο συντελεστής σιλουέτας και ο δείκτης Calinski - Harabasz.

Συντελεστής σιλουέτας (The silhouette Coefficient)

Ένας τρόπος για να αξιολογήσουμε το μοντέλο συσταδοποίησης που έχουμε δημιουργήσει είναι ο συντελεστής σιλουέτας (The Silhouette Coefficient). Ο συντελεστής αυτός ορίζεται για κάθε συστάδα που έχουμε δημιουργήσει και υπολογίζει τις παρακάτω δύο διαφορετικές ποσότητες.

a: Η μέση απόσταση ανάμεσα στο σημείο i και σε όλα τα υπόλοιπα σημεία που βρίσκονται στην ίδια συστάδα με αυτό.

b: Η μέση απόσταση μεταξύ του σημείου i και σε όλα τα υπόλοιπα σημεία που βρίσκονται στην πλησιέστερη συστάδα.

Για τον υπολογισμό του συντελεστή σιλουέτας χρησιμοποιείται ο τύπος:

$$s = \frac{b-a}{\max(a,b)}$$

Ο μέσος συντελεστής σιλουέτας σε όλα τα σημεία δεδομένων υπολογίζεται στη συνέχεια ως ο μέσος όρος των συντελεστών σιλουέτας για κάθε σημείο δεδομένων. Ο συντελεστής αυτός παίρνει τιμές ανάμεσα στο διάστημα -1 έως και 1. Τιμές κοντά στο -1 παίρνουν οι συσταδοποιήσεις που δεν έχουν καλή προσαρμογή, ενώ τιμές θετικές και κοντά στο 1 αποδεικνύουν καλές ομαδοποιήσεις. Τιμές του δείκτη γύρω από το 0 δείχνουν δεδομένα τα οποία μπορούν να ανήκουν σε παραπάνω από μία συστάδα και δεν διαχωρίζονται ξεκάθαρα. (Tan et al., 2018)

Δείκτης Calinski - Harabasz

Άλλη μια τεχνική αξιολόγησης μεθόδων ομαδοποίησης είναι ο δείκτης Calinski - Harabasz. Ο συγκεκριμένος δείκτης συνήθως χρησιμοποιείται για τη σύγκριση ανάμεσα σε δύο μεθόδους συσταδοποίησης στο ίδιο σύνολο δεδομένων. Η συνοχή των συστάδων υπολογίζεται βάσει των αποστάσεων των σημείων από το κέντρο της συστάδας. Γενικότερα, όσο υψηλότερες τιμές λαμβάνει ο δείκτης Calinski – Harabasz, τόσο καλύτερο είναι το αποτέλεσμα της ομαδοποίησης που παρουσιάστηκε. Ο τύπος υπολογισμού του είναι:

$$CH = \left[\frac{\sum_{k=1}^{K} n_k ||c_k - c||^2}{k-1}\right] / \left[\frac{\sum_{k=1}^{K} \sum_{i=1}^{n_k} ||d_i - c_k||^2}{N-K}\right]$$

όπου n_k είναι το πλήθος των σημείων που βρίσκονται στον k-οστή συστάδα, c_k είναι το κέντρο της k-οστή συστάδας, c είναι το κέντρο όλων των σημείων, d_i είναι η στήλη i από το σύνολο δεδομένων που χρησιμοποιήθηκε για την ανάλυση και N είναι το πλήθος των παρατηρήσεων που χρησιμοποιήθηκαν συνολικά.

7.2 Προσαρμογή θεωρητικού υπόβαθρου σε δεδομένα του NBA

Σ' αυτή την ενότητα θα εκτελεστούν όλες οι μέθοδοι που έχουν αναφερθεί προηγουμένως ώστε να καταλήξουμε στα αποτελέσματα της μηχανικής μάθησης στα δεδομένα μας. Σκοπός της ενότητας είναι να εκτελεστούν οι διαδικασίες που

παρουσιάστηκαν, ώστε να κατηγοριοποιηθούν τα δεδομένα μας ως προς την μεταβλητή των Playoffs και να δημιουργηθούν και οι κατάλληλες συστάδες. Στο πρώτο μέρος της ενότητας θα παρουσιαστούν οι διαδικασίες κατηγοριοποίησης και συσταδοποίησης βάσει της μεθόδου μείωσης δεδομένων Boruta. Στο δεύτερο μέρος θα εκτελεστεί η ίδια διαδικασία για τη μέθοδο μείωσης δεδομένων Κ καλύτερων χαρακτηριστικών. Τέλος, θα γίνει και η σύγκριση ανάμεσα στα αποτελέσματα των δύο αυτών μεθόδων επιλογής χαρακτηριστικών.

7.2.1. Μέθοδος Boruta

Εκτελώντας την μέθοδο Boruta στη γλώσσα προγραμματισμού της Python ο αλγόριθμος παρουσίασε δεκαέξι (16) χρήσιμες μεταβλητές. Οι μεταβλητές που θα χρησιμοποιηθούν στην συνέχεια για την μέθοδο παρουσιάζονται στον Πί0,νακα 7.1.

Χρήσιμες μεταβλητές
PPG
ORt
POSSt
TOV
SPG
RPG
DRB
3P%
FTAo
3PM
PPGA
FGM
FGA
MisFG
FG%
DRt

Πίνακας 7.1: Χρήσιμες μεταβλητές που υπέδειξε η μέθοδος Boruta

Επομένως στη συνέχεια θα χρησιμοποιηθούν μόνο αυτές οι μεταβλητές, του Πίνακα 7.1, για την εκτέλεση των μεθόδων της κατηγοριοποίησης και της συσταδοποίησης.

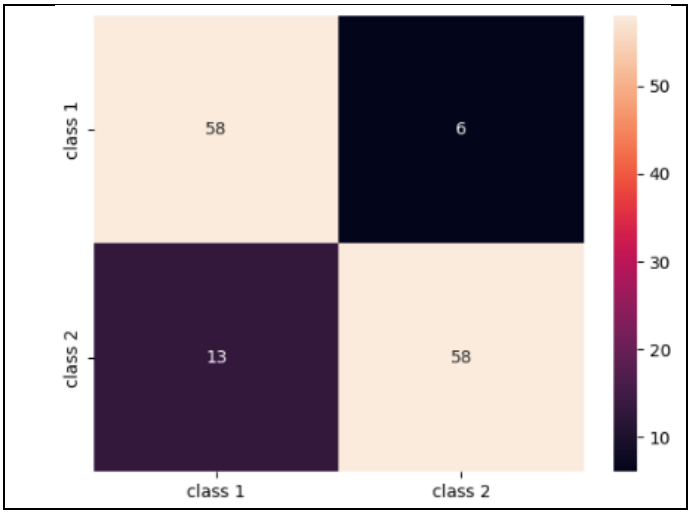
7.2.1.1. Κατηγοριοποίηση

Εφόσον βρέθηκαν οι μεταβλητές που θα χρησιμοποιηθούν, προχωράμε με την κατηγοριοποίηση των δεδομένων βάσει των μεθόδων που παρουσιάστηκαν στο θεωρητικό μέρος του κεφαλαίου.

7.2.1.1.1. Μέθοδος Κ κοντινότερων γειτόνων

Εκτελώντας την μέθοδο των k κοντινότερων γειτόνων παρατηρούμε πως σύμφωνα με το Σχήμα 7.17 ταξινομήθηκαν σωστά πενήντα οκτώ (58) παρατηρήσεις στις ομάδες που δεν προκρίθηκαν στη φάση των Playoffs και πενήντα οκτώ (58) στις ομάδες που προκρίθηκαν. Ακόμα, το μοντέλο μας ταξινόμησε λανθασμένα δεκατρείς (13) ομάδες που ενώ προκρίθηκαν σε ομάδες που δεν κατάφεραν να προκριθούν, και έξι (6) ομάδες

που δεν έχουν καταφέρει να προκριθούν στην κατηγορία των ομάδων που κατάφεραν να προκριθούν. Συνολικά, ταξινόμησε ορθά εκατόν δεκαέξι (116) παρατηρήσεις και δεκαεννιά (19) λανθασμένα με βάση το σύνολο δοκιμών.



Σχήμα 7.16: Πίνακας σύγχυσης της μεθόδου Κ κοντινότερων γειτόνων

Στη συνέχεια ακολουθεί ο πίνακας με τις τιμές των μέτρων αξιολόγησης για την κατηγοριοποίηση που πραγματοποιήθηκε. Γενικά, φαίνεται να έχει πραγματοποιηθεί μια καλή ταξινόμηση από το μοντέλο μας.

Μέτρα αξιολόγησης	Τιμές
Accuracy	**0.859**
Precision	**0.906**
Recall	**0.816**
F1 score	**0.861**

Πίνακας 7.2: Μέτρα αξιολόγησης της μεθόδου Κ κοντινότερων γειτόνων

7.2.1.1.2. Μέθοδος τυχαίου δάσους

Εκτελώντας τη μέθοδο του τυχαίου δάσους, παρατηρούμε πως τα δεδομένα ταξινομήθηκαν σωστά πενήντα οκτώ (58) παρατηρήσεις στις ομάδες που δεν προκρίθηκαν στη φάση των Playoffs και πενήντα εννέα (59) στις ομάδες που προκρίθηκαν. Ακόμα, το μοντέλο μας ταξινόμησε λανθασμένα τρείς (3) ομάδες που ενώ προκρίθηκαν σε ομάδες που δεν κατάφεραν να προκριθούν, και δεκαπέντε (15) ομάδες που δεν προκρίθηκαν στην κατηγορία των ομάδων που προκρίθηκαν. Συνολικά, ταξινόμησε ορθά εκατόν δεκαεπτά (117) παρατηρήσεις και δεκαοκτώ (18) λανθασμένα με βάση το σύνολο δοκιμών.

Σχήμα 7.17: Πίνακας σύγχυσης της μεθόδου τυχαίου δάσους

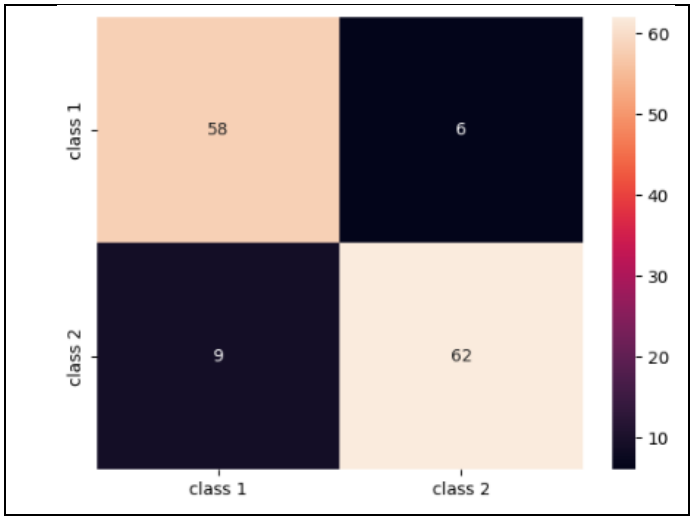
Στη συνέχεια ακολουθεί ο πίνακας με τις τιμές των μέτρων αξιολόγησης για την κατηγοριοποίηση που πραγματοποιήθηκε. Γενικά, φαίνεται να έχει πραγματοποιηθεί μια καλή ταξινόμηση από το μοντέλο μας και φανερά καλύτερη από την προηγούμενη.

Μέτρα αξιολόγησης	Τιμές
Accuracy	**0.866**
Precision	**0.950**
Recall	**0.794**
F1 score	**0.865**

Πίνακας 7.3: Μέτρα αξιολόγησης της μεθόδου τυχαίου δάσους

7.2.1.1.3. Μέθοδος SVM

Ακολουθώντας τη διαδικασία SVM με συνάρτηση πυρήνα την πολυωνυμική και τιμή για τη παράμετρο γ ίση με 1 προέκυψε ο παρακάτω πίνακας σύγχυσης, Σχήμα 7.18.



Σχήμα 7.18: Πίνακας σύγχυσης της μεθόδου SVM

Παρατηρούμε ότι μ' αυτή τη μέθοδο ταξινομήθηκαν σωστά πενήντα οκτώ (58) παρατηρήσεις στις ομάδες που δεν προκρίθηκαν στη φάση των Playoffs και εξήντα δύο

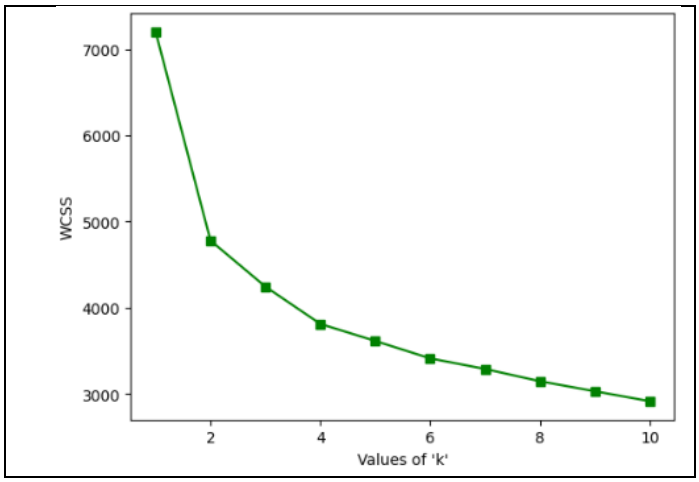
(62) στις ομάδες που προκρίθηκαν. Παράλληλα, το μοντέλο μας ταξινόμησε λανθασμένα εννέα (9) ομάδες που ενώ κατάφεραν να προκριθούν σε ομάδες που δεν προκρίθηκαν και έξι (6) ομάδες που δεν προκρίθηκαν στην κατηγορία των ομάδων που πέρασαν στην επόμενη φάση. Συνολικά, το μοντέλο ταξινόμησε ορθά εκατό είκοσι (120) παρατηρήσεις και μόλις δεκαπέντε (15) λανθασμένα με βάση το σύνολο δοκιμών. Ακολουθεί ο πίνακας με τις τιμές των μέτρων αξιολόγησης για την κατηγοριοποίηση. Γενικά, φαίνεται να έχει πραγματοποιηθεί μια αρκετά καλή ταξινόμηση από το μοντέλο μας με την παρούσα μέθοδο και προφανώς είναι καλύτερη από τις προηγούμενες δύο μεθόδους που παρουσιάστηκαν.

Μέτρα αξιολόγησης	Τιμές
Accuracy	**0.889**
Precision	**0.906**
Recall	**0.865**
F1 score	**0.885**

Πίνακας 7.4: Μέτρα αξιολόγησης της μεθόδου SVM

7.2.1.2. Συσταδοποίηση

Στη συνέχεια θα εκτελεστούν οι κώδικες συσταδοποίησης για τα δεδομένα που έχει επιλέξει η μέθοδος μείωσης χαρακτηριστικών Boruta. Ως πρώτο βήμα εκτελέστηκε η μέθοδος του αγκώνα ώστε να γίνει ξεκάθαρο πόσες συστάδες θα ζητήσουμε από τις μεθόδους συσταδοποίησης να δημιουργήσουν. Όπως φαίνεται στο Σχήμα 7.19, μια ορθή επιλογή για το πλήθος των αρχικών συστάδων είναι δύο (2), δηλαδή k = 2.



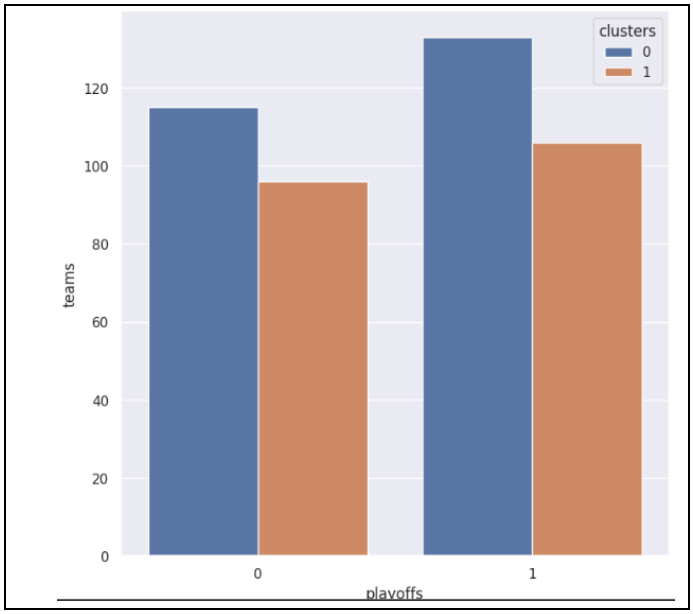
Σχήμα 7.19: Γράφημα για τη μέθοδο του αγκώνα για τα Playoffs

7.2.1.2.1. Μέθοδος K-Means

Εφόσον επιλέχθηκε το πλήθος των συστάδων, προχωράμε στην εφαρμογή του αλγόριθμου K-Means και καταλήγουμε στο παρακάτω συμπέρασμα σχετικά με την κατανομή των δεδομένων στις δύο συστάδες:

- από τις διακόσες έντεκα (210) ομάδες που δεν προκρίθηκαν στα Playoffs της διοργάνωσης, οι εκατό δεκατέσερις (114) ομάδες βρέθηκαν στην 1η συστάδα (cluster 0) και οι ενενήντα έξι (96) ομάδες βρέθηκαν στην 2η συστάδα (cluster 1).

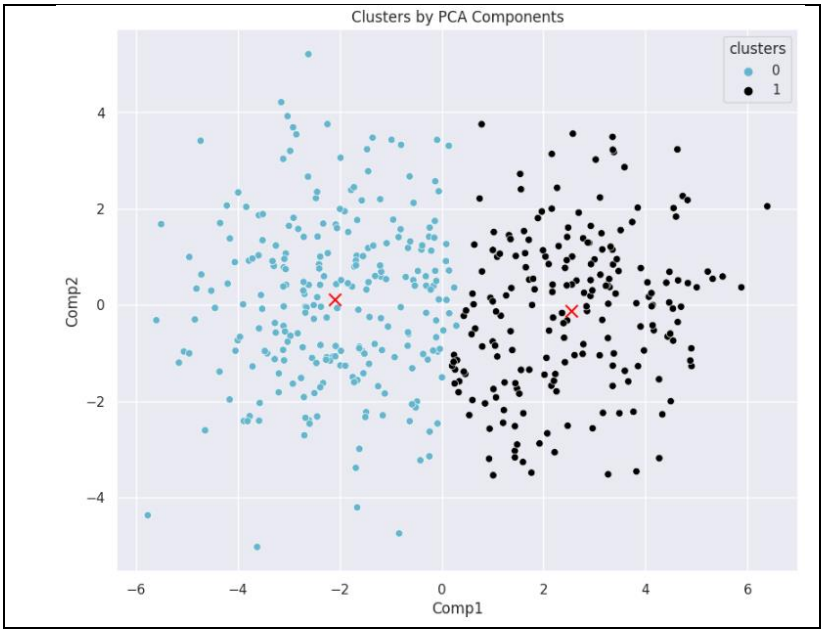
- από τις διακόσες τριάντα εννέα (240) ομάδες που προκρίθηκαν στα Playoffs της διοργάνωσης, οι εκατό τριάντα τέσσερις (134) ομάδες βρέθηκαν στην 1η συστάδα (cluster 0) και οι εκατό έξι (106) ομάδες βρέθηκαν στην 2η συστάδα (cluster 1).



Σχήμα 7.20: Bar plot για την μέθοδο K-Means

Από τα συμπεράσματα που αναλύσαμε και γραφικά, Σχήμα 7.20, φαίνεται πως η συσταδοποίηση που έχει πραγματοποιηθεί δεν είναι αρκετά καλή διότι το μοντέλο ταξινομεί αρκετές παρατηρήσεις λάθος. Στην τελευταία υπο-ενότητα του κεφαλαίου θα γίνει μια διερεύνηση για τον λόγο που ταξινομούνται οι παρατηρήσεις τόσο λανθασμένα.

Στη συνέχεια της ανάλυσης κατά συστάδες με τη μέθοδο K-Means θα απεικονίσουμε τα δεδομένα μας ώστε να φαίνονται οι δύο συστάδες που δημιούργησε ο αλγόριθμός μας. Προφανώς στην συγκεκριμένη περίπτωση δεν μπορεί να χρησιμοποιηθεί ένα κλασικό διάγραμμα διασποράς, διότι το σύνολο δεδομένων μας περιέχει περισσότερες από δύο μεταβλητές. Αυτό το πρόβλημα λύνεται με τη χρήση ενός αλγορίθμου μείωσης διαστασιμότητας, όπως η ανάλυση κύριων συνιστωσών (PCA), που δημιουργεί νέες μεταβλητές, οι οποίες είναι γραμμικοί συνδυασμοί των αρχικών μας μεταβλητών και προσπαθούν να ερμηνεύσουν όσο το δυνατόν περισσότερη πληροφορία από το αρχικό σύνολο δεδομένων. Η πρώτη κύρια συνιστώσα ερμηνεύει το μεγαλύτερο ποσοστό της μεταβλητότητας, η δεύτερη το αμέσως επόμενο κ.ο.κ. Στο Σχήμα 7.21 παρουσιάζονται οι δύο συστάδες που δημιούργησε το μοντέλο μας.



Σχήμα 7.21: Cluster plot για τη συσταδοποίηση μέσω του K-means

Τέλος, στον Πίνακα 7.5 παρουσιάζονται τα μέτρα αξιολόγησης της συσταδοποίησης που έγινε με την χρήση της μεθόδου K-Means. Από τις τιμές που λαμβάνουν οι δείκτες της συσταδοποίησης συμπεραίνουμε πως η συσταδοποίηση που πραγματοποιήθηκε δεν είναι αρκετά αποτελεσματική.

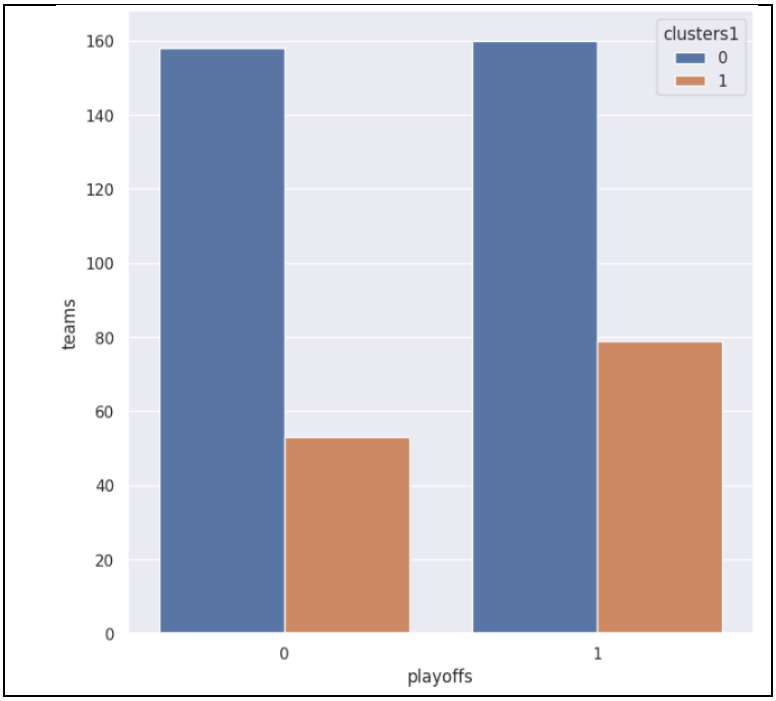
Μέτρα αξιολόγησης	Τιμή μέτρου αξιολόγησης
Silhouette	**0.284**
Calinski – Harabasz	**227.136**

Πίνακας 7.5: Μέτρα αξιολόγησης της μεθόδου K-Means συσταδοποίησης

7.2.1.2.2. Μέθοδος Birch

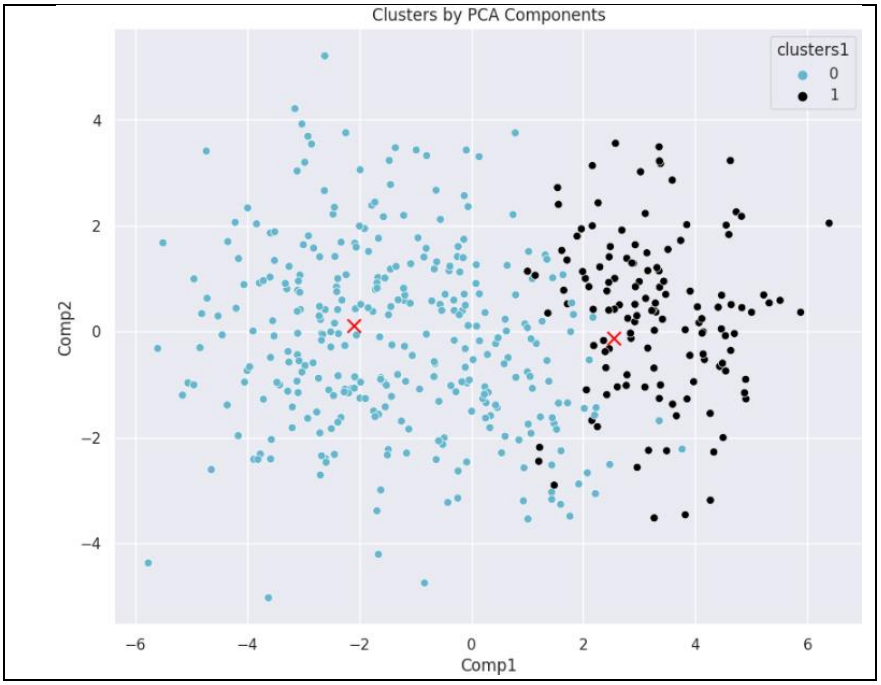
Εκτελώντας τη μέθοδο συσταδοποίησης Birch παρατηρούμε πως:

- από τις διακόσες έντεκα (210) ομάδες που δε προκρίθηκαν στα Playoffs της διοργάνωσης, οι εκατό πενήντα επτά (157) ομάδες βρέθηκαν στην 1η συστάδα (cluster 0) και οι πενήντα τρία (53) ομάδες βρέθηκαν στην 2η συστάδα (cluster 1).
- • Από τις διακόσες τριάντα εννέα (240) ομάδες που προκρίθηκαν στα Playoffs της διοργάνωσης, οι εκατό εξήντα ένα (161) ομάδες βρέθηκαν στην 1η συστάδα (cluster 0) και οι εβδομήντα εννέα (79) ομάδες βρέθηκαν στην 2η συστάδα (cluster 1).



Σχήμα 7.22: Bar plot για την μέθοδο Birch

Από τα συμπεράσματα που αναλύσαμε και γραφικά, Σχήμα 7.22, φαίνεται πως η συσταδοποίηση που έχει πραγματοποιηθεί είναι διαφορετική απ' αυτή που έδωσε η προηγούμενη μέθοδος. Στο Σχήμα 7.23 παρουσιάζεται το Cluster plot βάσει των δύο πρώτων μεταβλητών της μεθόδου pca, απ' όπου φαίνεται ότι στο κέντρο του σχήματος μερικά σημεία έχουν ταξινομηθεί διαφορετικά απ' ό,τι με την μέθοδο K-Means. Η πραγματική διαφορά τους αναμένεται να είναι φανερή μέσω των αποτελεσμάτων των δεικτών.



Σχήμα 7.23: Cluster plot για τη συσταδοποίηση μέσω Birch συσταδοποίησης

Τέλος, στον Πίνακα 7.6 παρουσιάζονται τα μέτρα αξιολόγησης της συσταδοποίησης που έγινε με την χρήση της μεθόδου K-Means. Από τις τιμές που λαμβάνουν οι δείκτες της συσταδοποίησης συμπεραίνουμε πως η συσταδοποίηση που πραγματοποιήθηκε δεν είναι αρκετά αποτελεσματική και βάσει του δείκτη Calinski – Harabasz δεν είναι σίγουρα προτιμότερη από την προηγούμενη, εφόσον σ' αυτή την περίπτωση λαμβάνει μικρότερη τιμή.

Μέτρα αξιολόγησης	Τιμή μέτρου αξιολόγησης
Silhouette	**0.256**
Calinski – Harabasz	**176.217**

Πίνακας 7.6: Μέτρα αξιολόγησης της μεθόδου Birch συσταδοποίησης

7.2.2. Μέθοδος Κ καλύτερων χαρακτηριστικών

Σ' αυτή την ενότητα θα παρουσιαστούν τα αποτελέσματα για τη μέθοδο των k καλύτερων χαρακτηριστικών. Για τη συγκεκριμένη εργασία επιλέχτηκε να τεθεί το k ίσο με οκτώ (8). Οι σημαντικότερες μεταβλητές δίνονται στον Πίνακα 7.7.

Χρήσιμες μεταβλητές
PPG
PPGA
MisFG
3PM
3PA
FTAo
ORt
DRt

Πίνακας 7.7: Χρήσιμες μεταβλητές που υπέδειξε η μέθοδος Κ καλύτερων χαρακτηριστικών

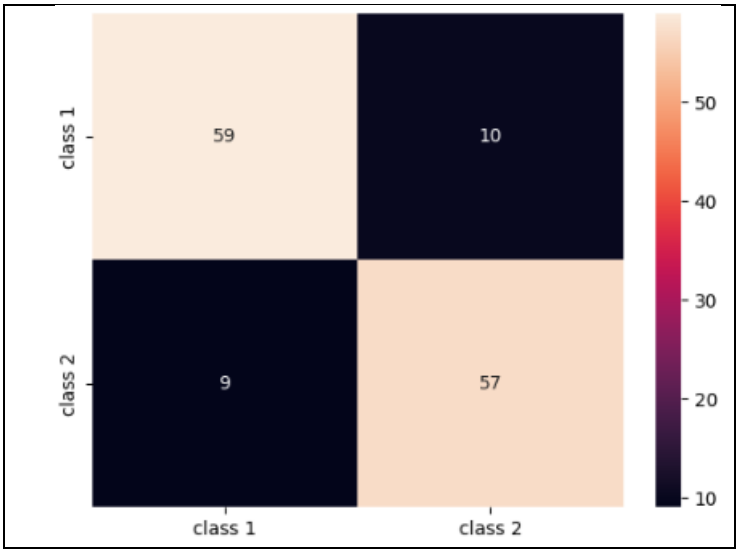
Επομένως στη συνέχεια θα χρησιμοποιηθούν μόνο αυτές οι μεταβλητές, του Πίνακα 7.7.

7.2.2.1. Κατηγοριοποίηση

Για τη διαδικασία κατηγοριοποίησης θα χρησιμοποιηθούν και σ' αυτήν την ενότητα οι τρείς (3) μέθοδοι που χρησιμοποιήθηκαν και προηγουμένως με την μόνη διαφορά να βρίσκεται στις μεταβλητές που χρησιμοποιήθηκαν.

7.2.2.1.1. Μέθοδος Κ κοντινότερων γειτόνων

Εκτελώντας τη μέθοδο των k κοντινότερων γειτόνων παρατηρούμε πως σύμφωνα με το Σχήμα 7.24 ταξινομήθηκαν σωστά πενήντα εννιά (59) παρατηρήσεις στις ομάδες που δεν προκρίθηκαν στη φάση των Playoffs και πενήντα επτά (57) στις ομάδες που προκρίθηκαν. Ακόμα, το μοντέλο μας ταξινόμησε λανθασμένα εννιά (9) ομάδες που προκρίθηκαν σε ομάδες που δεν κατάφεραν να προκριθούν, και δέκα (10) ομάδες που δεν προκρίθηκανσε ομάδες που προκρίθηκαν. Συνολικά, ταξινόμησε ορθά εκατό δεκαέξι (116) παρατηρήσεις και δεκαεννιά (19) λανθασμένα, με βάση το σύνολο δοκιμών.



Σχήμα 7.24: Πίνακας σύγχυσης της μεθόδου Κ κοντινότερων γειτόνων

Έπειτα παρουσιάζονται τα μέτρα αξιολόγησης της μεθόδου από τα οποία συμπεραίνουμε πως έχει γίνει μια καλή ταξινόμηση.

Μέτρα αξιολόγησης	Τιμές
Accuracy	**0.851**
Precision	**0.855**
Recall	**0.867**
F1 score	**0.861**

Πίνακας 7.8: Μέτρα αξιολόγησης της μεθόδου Κ κοντινότερων γειτόνων

7.2.2.1.2. Μέθοδος τυχαίου δάσους

Εκτελώντας τη μέθοδο του τυχαίου δάσους παρατηρούμε πως τα δεδομένα ταξινομήθηκαν με διαφορετικό τρόπο απ' ότι στην προηγούμενη μέθοδο, παρόλο που στη μέθοδο Boruta είχαν ακριβώς ίδιο τρόπο ταξινόμησης. Όπως φαίνεται στο Σχήμα 7.25, ο αλγόριθμος ταξινομεί σωστά πενήντα επτά (57) παρατηρήσεις στις ομάδες που δεν προκρίθηκαν στη φάση των Playoffs και πενήντα έξι (56) στις ομάδες που προκρίθηκαν. Αντίθετα, το μοντέλο μας ταξινόμησε λανθασμένα δέκα (10) ομάδες που προκρίθηκαν σε ομάδες που δεν πήραν την πρόκριση, και δώδεκα (12) ομάδες που ενώ δεν προκρίθηκαν στην κατηγορία των ομάδων που πέρασαν στην επόμενη φάση.

Σχήμα 7.25: Πίνακας σύγχυσης της μεθόδου Κ κοντινότερων γειτόνων

Στον Πίνακα 7.9 παρουσιάζονται τα μέτρα αξιολόγησης της μεθόδου, από τα οποία συμπεραίνουμε πως έχει γίνει μια καλή ταξινόμηση και σ' αυτή την περίπτωση αλλά λιγότερο καλή από την προηγούμενη μέθοδο.

Μέτρα αξιολόγησης	Τιμές
Accuracy	**0.837**
Precision	**0.826**
Recall	**0.850**
F1 score	**0.838**

Πίνακας 7.9: Μέτρα αξιολόγησης της μεθόδου τυχαίου δάσους

7.2.2.1.3. Μέθοδος SVM

Ακολουθώντας τη διαδικασία SVM με συνάρτηση πυρήνα την πολυωνυμική και τιμή για τη παράμετρο γ ίση με 1 προέκυψε ο παρακάτω πίνακας σύγχυσης, Σχήμα 7.26.

Σχήμα 7.26: Πίνακας σύγχυσης της μεθόδου SVM

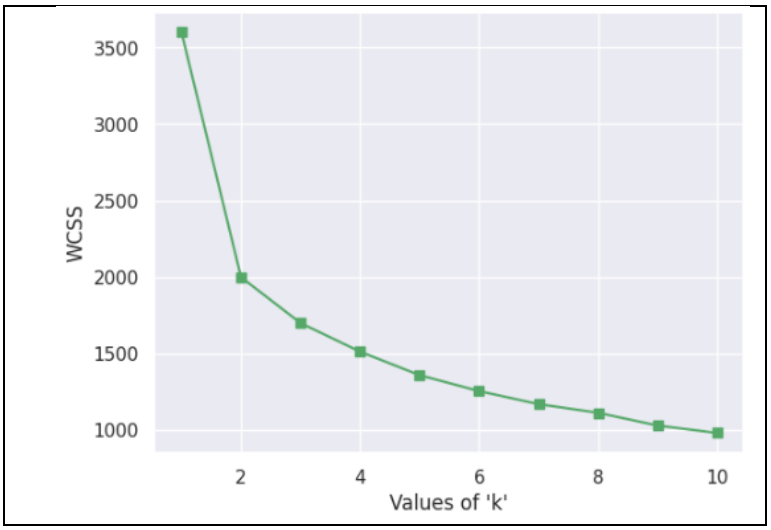
Παρατηρούμε ότι μ' αυτή τη μέθοδο ταξινομήθηκαν σωστά πενήντα οκτώ (58) παρατηρήσεις στις ομάδες που δεν προκρίθηκαν στη φάση που τίθεται προς μελέτη και πενήντα εννιά (59) στις ομάδες που προκρίθηκαν, ενώ λανθασμένα ταξινομήθηκαν επτά (7) ομάδες που προκρίθηκαν σε ομάδες που δεν πήραν την πρόκριση, και έντεκα (11) ομάδες που δεν προκρίθηκαν στην κατηγορία των ομάδων που προκρίθηκαν. Στη συνέχεια παρουσιάζεται ο πίνακας με τις τιμές των μέτρων αξιολόγησης του μοντέλου, απ' όπου και συμπεραίνουμε πως έχει πραγματοποιηθεί μια αρκετά καλή ταξινόμηση από το μοντέλο μας με την παρούσα μέθοδο. Ανάμεσα στις τρείς μεθόδους που εκτελέστηκαν για το συγκεκριμένο σύνολο δεδομένων καλύτερη μέθοδος κατηγοριοποίησης ήταν η SVM, εφόσον είχε τα υψηλότερα μέτρα αξιολόγησης.

Μέτρα αξιολόγησης	Τιμές
Accuracy	**0.867**
Precision	**0.840**
Recall	**0.892**
F1 score	**0.865**

Πίνακας 7.10: Μέτρα αξιολόγησης της μεθόδου SVM

7.2.2.2. Συσταδοποίηση

Για να εκτελέσουμε τις μεθόδους συσταδοποίησης θα πρέπει και για αυτές τις μεταβλητές να εκτελεστεί η μέθοδος αγκώνα ώστε να δούμε το ιδανικό πλήθος k για τις συστάδες. Όπως είναι φανερό στο Σχήμα 7.27, το k που κρίνεται ορθό να επιλέξουμε είναι ίσο με 2.



Σχήμα 7.27: Γράφημα για τη μέθοδο του αγκώνα για τα Playoffs

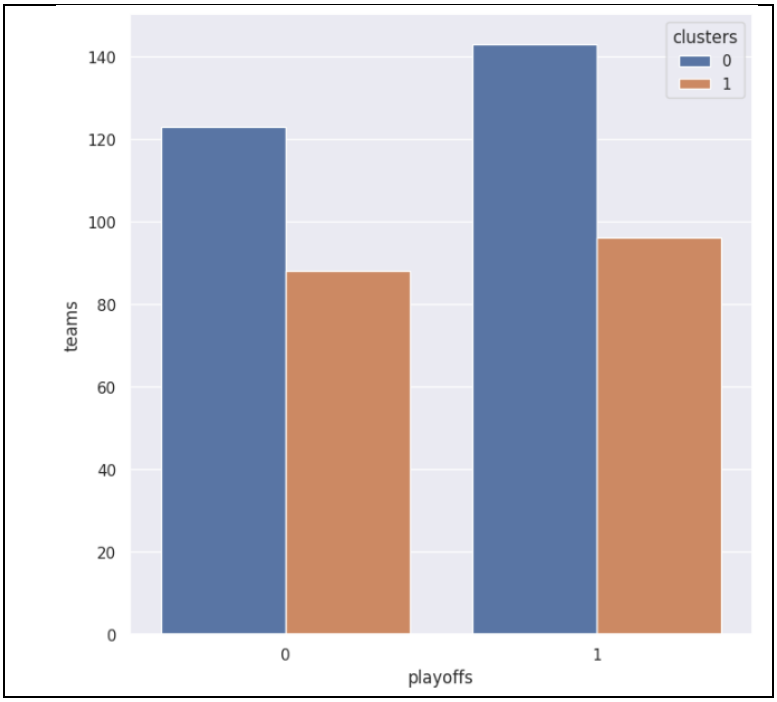
7.2.1.2.1. Μέθοδος K-Means

Εφόσον επιλέχθηκε το πλήθος των συστάδων, προχωράμε στην εφαρμογή του αλγόριθμου K-Means και καταλήγουμε στο παρακάτω συμπέρασμα σχετικά με την κατανομή των δεδομένων στις δύο συστάδες:

- από τις διακόσες έντεκα (210) ομάδες που δεν προκρίθηκαν στα Playoffs της διοργάνωσης, οι εκατό είκοσι δύο (122) ομάδες βρέθηκαν στην 1η συστάδα

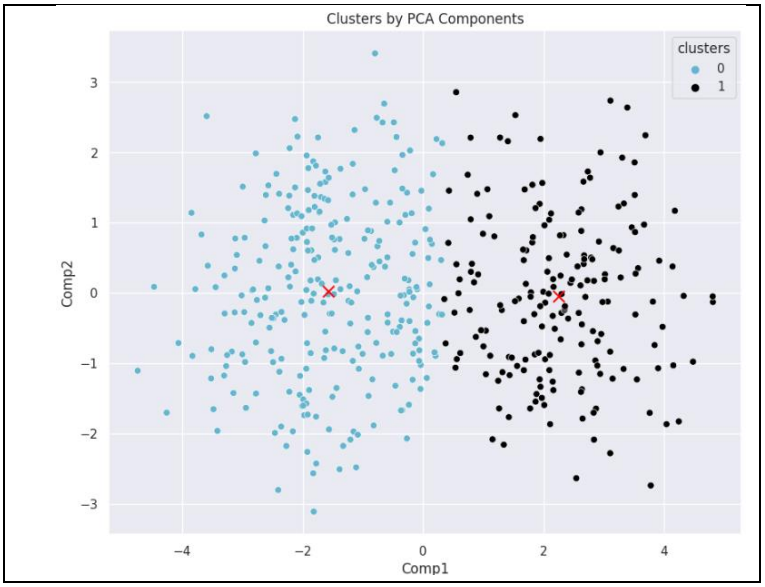
(cluster 0) και οι ογδόντα οκτώ (88) ομάδες βρέθηκαν στην 2η συστάδα (cluster 1).

- από τις διακόσες τριάντα εννέα (240) ομάδες που προκρίθηκαν στα Playoffs της διοργάνωσης, οι εκατό σαράντα τρείς (143) ομάδες βρέθηκαν στην 1η συστάδα (cluster 0) και οι ενενήντα επτά (97) ομάδες βρέθηκαν στην 2η συστάδα (cluster 1).



Σχήμα 7.28: Bar plot για την μέθοδο K-Means

Από τα συμπεράσματα που αναλύσαμε και γραφικά, Σχήμα 7.28, φαίνεται πως η συσταδοποίηση που έχει πραγματοποιηθεί δεν είναι αρκετά καλή, διότι το μοντέλο ταξινομεί αρκετές παρατηρήσεις λανθασμένα. Όπως φαίνεται, το μοντέλο ταξινομεί και αυτό όλες τις παρατηρήσεις των πρώτων σεζόν στην κατηγορία των ομάδων που δεν προκρίνονται στα Playoffs, όπως έγινε και προηγουμένως. Παρ' όλα αυτά στο Σχήμα 7.29 παρουσιάζονται οι παρατηρήσεις κάθε συστάδας.



Σχήμα 7.29: Cluster plot για τη συσταδοποίηση μέσω του K-means

Τέλος, στον Πίνακα 7.11 παρουσιάζονται τα μέτρα αξιολόγησης της συσταδοποίησης που έγινε με τη χρήση της μεθόδου K-Means, απ' όπου αντιλαμβανόμαστε πως η συσταδοποίηση που πραγματοποιήθηκε δεν είναι αρκετά αποτελεσματική.

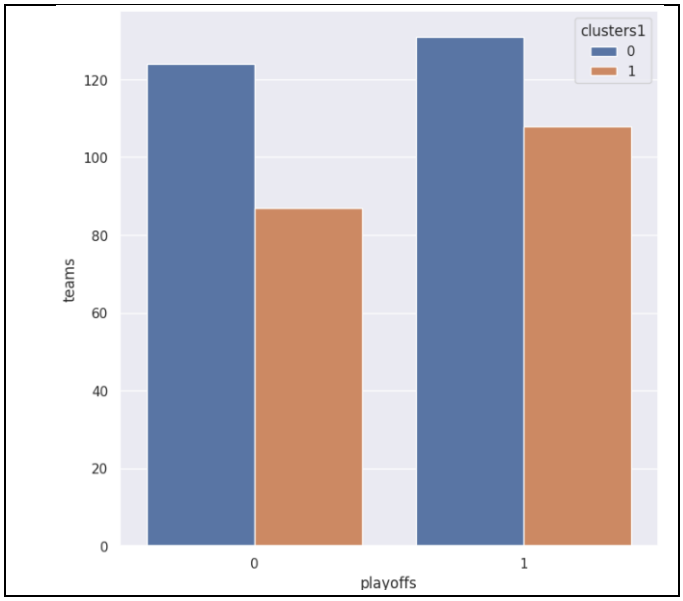
Μέτρα αξιολόγησης	Τιμή μέτρου αξιολόγησης
Silhouette	**0.373**
Calinski - Harabasz	**358.968**

Πίνακας 7.11: Μέτρα αξιολόγησης της μεθόδου K-Means συσταδοποίησης

7.2.1.2.2. Μέθοδος Birch

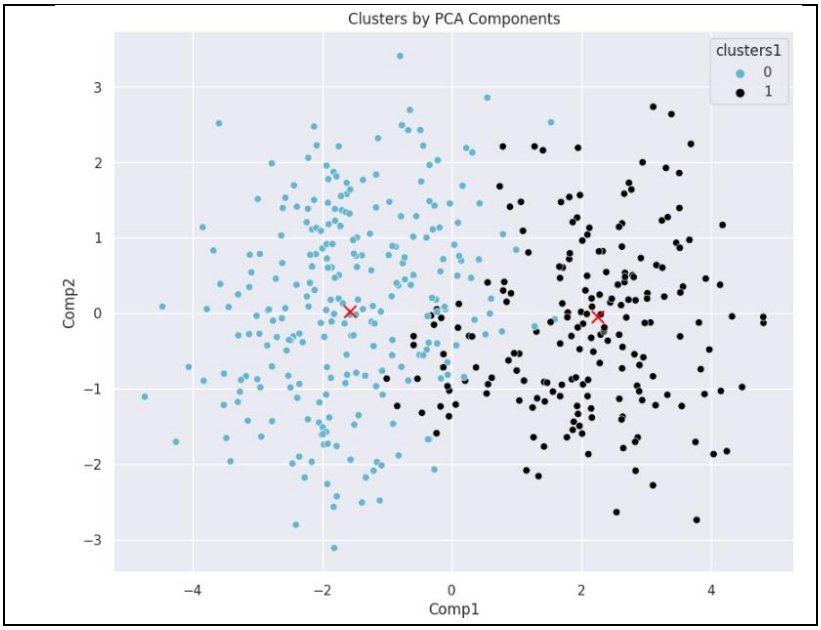
Εκτελώντας τη μέθοδο συσταδοποίησης Birch παρατηρούμε πως:

- από τις διακόσες έντεκα (210) ομάδες που δεν προκρίθηκαν στα Playoffs της διοργάνωσης, οι εκατό είκοσι τρείς (123) ομάδες βρέθηκαν στην 1η συστάδα (cluster 0) και οι ογδόντα επτά (87) ομάδες βρέθηκαν στην 2η συστάδα (cluster 1).
- από τις διακόσες τριάντα εννέα (240) ομάδες που προκρίθηκαν στα Playoffs της διοργάνωσης, οι εκατό τριάντα δύο (132) ομάδες βρέθηκαν στην 1η συστάδα (cluster 0) και οι εκατό οκτώ (108) ομάδες βρέθηκαν στην 2η συστάδα (cluster 1).



Σχήμα 7.30: Bar plot για την μέθοδο Birch

Από τα συμπεράσματα που αναλύσαμε και γραφικά, Σχήμα 7.30, φαίνεται πως η συσταδοποίηση που έχει πραγματοποιηθεί είναι ελάχιστα διαφορετική απ' αυτή που έδωσε η προηγούμενη μέθοδος. Στο Σχήμα 7.31 παρουσιάζεται το Cluster plot βάσει των δύο πρώτων μεταβλητών της μεθόδου pca.



Σχήμα 7.31: Cluster plot για τη συσταδοποίηση μέσω του K-means

Τέλος, στον Πίνακα 7.12 παρουσιάζονται τα μέτρα αξιολόγησης της συσταδοποίησης που έγινε με την χρήση της μεθόδου Birch. Από τις τιμές που λαμβάνουν οι δείκτες της συσταδοποίησης συμπεραίνουμε πως η συσταδοποίηση που πραγματοποιήθηκε δεν είναι αποτελεσματική και βάσει του δείκτη Calinski – Harabasz δεν είναι αποτελεσματικότερη από την προηγούμενη.

Μέτρα αξιολόγησης	Τιμή μέτρου αξιολόγησης
Silhouette	**0.254**

Calinski - Harabasz	194.139

Πίνακας 7.12: Μέτρα αξιολόγησης της μεθόδου Birch συσταδοποίησης

7.2.3. Σύγκριση μεθόδων μείωσης χαρακτηριστικών

Έπειτα από την εκτέλεση όλων των διαδικασιών που θέλαμε να υλοποιήσουμε στην παρούσα εργασία θα ήταν ορθό να γίνει μία σύγκριση ανάμεσα στις μεθόδους μείωσης χαρακτηριστικών. Όπως είδαμε παραπάνω, πρώτα εκτελέστηκε η μέθοδος Boruta, στην οποία για τη διαδικασία κατηγοριοποίησης καλύτερη τεχνική ήταν οι μηχανές διανυσματικής υποστήριξης. Για τη συσταδοποίηση αυτής της διαδικασίας μείωσης η καλύτερη μέθοδος ήταν η διαδικασία K-Means. Ακριβώς οι ίδιες μέθοδοι φάνηκε πως είναι καλύτερες και για τη μέθοδο μείωσης χαρακτηριστικών βάσει k καλύτερων χαρακτηριστικών.

Για την κατηγοριοποίηση βάσει των μέτρων αξιολόγησης, Πίνακες 7.4 και 7.10, είναι φανερό πως η μέθοδος μείωσης χαρακτηριστικών Boruta δίνει ελάχιστα καλύτερα αποτελέσματα από την άλλη. Για τη συσταδοποίηση η μέθοδος Boruta δίνει για τον δείκτη Calinski - Harabasz τιμή 227.136, Πίνακας 7.5, ενώ η μέθοδος k καλύτερων χαρακτηριστικών τιμή 358.968, Πίνακας 7.11. Επομένως με αυτό το κριτήριο καλύτερη μέθοδος κρίνεται αυτή που προέκυψε βάσει της επιλογής των k καλύτερων χαρακτηριστικών. Στην πραγματικότητα οι συσταδοποιήσεις που εκτελέστηκαν είχαν όλες αρκετά χαμηλούς δείκτες, κάτι που έδειξε πως δεν ήταν ικανοποιητικές. Έτσι δημιουργήθηκε ένα καινούργιο αρχείο δεδομένων για να αναδειχθούν οι λόγοι που οδήγησαν σε αυτά τα αποτελέσματα. Μετά από την εκτέλεση της συσταδοποίησης με την μέθοδο K-means, χρησιμοποιώντας τις μεταβλητές που ήταν σημαντικές με την μέθοδο Boruta, εκτυπώθηκαν τα labels ώστε να είναι φανερό πώς χωρίστηκαν οι παρατηρήσεις. Όπως έγινε κατανοητό η πρώτη συστάδα αποτελούνταν από τις παρατηρήσεις των πρώτων επτά (7) σεζόν και οι υπόλοιπες παρατηρήσεις αποτελούν την δεύτερη συστάδα. Έπειτα από την παρατήρηση αυτού του γεγονότους δημιουργήθηκε το νέο αρχείο δεδομένων το οποίο αποτελείται από τα στατιστικά δεδομένα των ομάδων. Σ' αυτό το αρχείο οι μεταβλητές χωρίστηκαν σε δύο υποκατηγορίες ώστε να είναι εφικτό να εκτελεστούν τα απαραίτητα paired samples t-test, που έχουν παρουσιαστεί στο κεφάλαιο 4. Η μεταβλητή με το όνομα PPG αναφέρεται πλέον στους πόντους των ομάδων που καταγράφηκαν τις πρώτες επτά (7) σεζόν, δηλαδή από τη σεζόν 2022-2023 έως και τη σεζόν 2016-2017, ενώ δημιουργήθηκε μια ακόμα μεταβλητή, η PPG1, που αναφέρεται στους πόντους που σκόραραν οι ίδιες ομάδες από τη σεζόν 2014-2015 έως και τη σεζόν 2008-2009. Ο παράγοντας που ελέγχεται με τη χρήση των ελέγχων ισότητας για τη μέση τιμή είναι η αλλαγή στις τιμές των μεταβλητών μετά από το πέρασμα οχτώ (8) σεζόν της διοργάνωσης. Με αυτόν τον τρόπο χωρίστηκαν όλες οι μεταβλητές που χρησιμοποιήθηκαν στις μεθόδους συσταδοποίησης, για να γίνουν τα t-test και να ελεγχθεί εάν οι μέσοι των μεταβλητών διαφέρουν σημαντικά ώστε να αποδειχθεί πως οι χαμηλές τιμές επίδοσης των συσταδοποιήσεων οφείλονται στην διαφορά επιπέδου των ομάδων και των στατιστικών που καταγράφουν. Στον Πίνακα 7.13 παρουσιάζονται τα p-values των paired t-test όπου, όπως φαίνεται, οι μέσες τιμές έχουν στατιστικά σημαντικές διαφορές.

Μεταβλητές ελέγχου	t-statistic	p_{value}
PPG	**5.4118**	**1.702e-07**
ORt	**62.6995**	**<2.2e-16**
POSSt	**974.3530**	**<2.2e-16**
SPG	**--110.1583**	**<2.2e-16**
RPG	**70.9429**	**<2.2e-16**
TOV	**105.3432**	**<2.2e-16**
DRB	**163.9706**	**<2.2e-16**
3P%	**2.0487**	**0.04173**
FTAo	**152.8128**	**<2.2e-16**
3PM	**30.3084**	**<2.2e-16**
PPGA	**24.5909**	**<2.2e-16**
FGM	**21.2940**	**<2.2e-16**
FGA	**23.3593**	**<2.2e-16**
MisFG	**12.3974**	**<2.2e-16**
FG%	**5.4603**	**1.705e-07**
DRt	**71.1035**	**<2.2e-16**

Πίνακας 7.13: Αποτελέσματα ελέγχου ισότητας μέσων βάσει του νέου αρχείο δεδομένων

Συνεπώς, οι αλγόριθμοι συσταδοποίησης δεν έχουν καλές επιδόσεις εφόσον διακρίνουν τις διαφορές στις τιμές των μεταβλητών με το πέρασμα οχτώ (8) σεζόν. Αυτές οι διαφορές είχαν αναγνωριστεί και στο τρίτο (3) κεφάλαιο, όπου είχαν γίνει τα time series plot των μεταβλητών και ήταν φανερή η άνοδος – κάθοδος των τιμών στις τελευταίες σεζόν, μόνο που τώρα αποδείχθηκαν και με τους κατάλληλους ελέγχους.

ΚΕΦΑΛΑΙΟ 8ο

8. Συμπεράσματα

Στην παρούσα διπλωματική εργασία συλλέχθηκαν στατιστικά δεδομένα της διοργάνωσης National Basketball Association (NBA) αναφορικά με τις τελευταίες δεκαπέντε (15) σεζόν, δηλαδή από τη σεζόν 2008-2009 έως και την περσινή 2022-2023. Τα στατιστικά προήλθαν από την επίσημη σελίδα basketball-realgm με βασικό σκοπό της διπλωματικής εργασίας την εύρεση των σημαντικότερων μεταβλητών που οδηγούν μία ομάδα στην πρόκρισή της στη φάση των playoffs και στην κατάκτηση του τίτλου. Στο σύνολο των παραγόντων που χρησιμοποιήθηκαν ανήκουν προφανώς τα βασικά χαρακτηριστικά που μετριούνται σε έναν αγώνα καλαθοσφαίρισης, όπως οι πόντοι της ομάδας, τα επιτυχημένα δίποντα και τρίποντα, οι ασίστ, τα αμυντικά, τα επιθετικά και τα συνολικά ριμπάουντ της ομάδας, τα κλεψίματά της, τα κοψίματα που καταφέρνει να κάνει στους αντιπάλους και τα φάουλ στα οποία υποπίπτει, αλλά και παράγοντες που περιέχουν συνδυαστική πληροφορία, όπως οι κατοχές μιας ομάδας στον αγώνα και οι δείκτες defensive rating και offensive rating, οι οποίοι υπολογίστηκαν βάσει των τύπων που παρουσίασαν οι Kubatko et al. (2007) στην έρευνά τους.

Στο δεύτερο κεφάλαιο πραγματοποιήθηκε η ιστορική αναδρομή της διοργάνωσης του NBA και δόθηκε μεγάλη σημασία στο format της με το πέρασμα του χρόνου. Αμέσως μετά παρουσιάστηκαν οι ομάδες που συμμετείχαν στην διοργάνωση τα χρόνια στα οποία βασίζεται η ανάλυση της διπλωματικής εργασίας. Στο τέλος αυτού του κεφαλαίου παρουσιάστηκαν υπάρχοντα επιστημονικά άρθρα και μελέτες που είναι άμεσα σχετικά με το θέμα της παρούσας εργασίας.

Στο τρίτο κεφάλαιο, παρουσιάστηκαν περιγραφικά μέτρα και διάφορα διαγράμματα που σχετίζονται με τις μεταβλητές που υπάρχουν στα δεδομένα μας, για να γίνουν καλύτερα αντιληπτές από τον αναγνώστη. Οι αναλύσεις που ακολούθησαν σε αυτό το κεφάλαιο βασίστηκαν στις κατηγορικές μεταβλητές που περιέχονται στο σύνολο δεδομένων, Playoffs (πρόκριση ή όχι στα playoffs), Champions (κατάκτηση ή όχι του πρωταθλήματος), Conference (αν η ομάδα συμμετέχει στο πρωτάθλημα της East περιφέρειας ή West) και Teams (οι τριάντα διαφορετικές ομάδες). Πριν την ανάλυση εκτελέστηκε ο έλεγχος για ελλιπή δεδομένα, ο οποίος έδειξε πως στο σύνολο δεδομένων μας δεν υπάρχουν τέτοιες τιμές.

Στη συνέχεια, εξετάστηκαν τα στατιστικά των ομάδων για τη φάση της κανονικής περιόδου (regular season). Πιο συγκεκριμένα, δόθηκε ο πίνακας με τα περιγραφικά μέτρα και μετά παρουσιάστηκαν τα γραφήματα χρονοσειρών για να αποκτήσουμε μια γενική οπτική για το πως μεταβάλλονται τα δεδομένα με το πέρασμα των δεκαπέντε σεζόν. Βάσει των γραφημάτων τα συμπεράσματα που εξάγουμε δείχνουν πως οι περισσότερες μεταβλητές έχουν ανοδική πορεία έπειτα από τη σεζόν 2014-2015. Τέτοιες μεταβλητές είναι οι πόντοι ανά αγώνα, τα επιτυχημένα δίποντα ή τρίποντα, τα αμυντικά ριμπάουντ και τα συνολικά, οι κατοχές των ομάδων και των αντιπάλων τους. Για τις κατοχές ανά αγώνα έχουν καταλήξει στο ίδιο συμπέρασμα στις έρευνές τους και οι Mandić et al. (2019). Σχολιάστηκε, αναλυτικά στην παράγραφο 3.2, ακόμα πως

οι παραπάνω αλλαγές μπορεί να οφείλονται στις τροποποιήσεις στους κανονισμούς της διοργάνωσης που οδήγησαν σε μικρότερο χρόνο επίθεσης έπειτα από κάποιο επιθετικό ριμπάουντ και στον κανονισμό των τριών δευτερολέπτων εντός ρακέτας. Αντίθετα, ορισμένες μεταβλητές ακολούθησαν πτωτική πορεία, όπως τα επιθετικά ριμπάουντ της ομάδας και των αντιπάλων τους. Σ' αυτή την ανάλυση δημιουργήθηκαν ακόμα κατάλληλα scatter plot για να παρουσιαστεί η σχέση του δείκτη Efficiency με τις υπόλοιπες μεταβλητές. Όπως φάνηκε, αναλυτικά στην παράγραφο 3.2, υπήρξαν συσχετίσεις ποικίλης μορφής ανάμεσα στις μεταβλητές και τον δείκτη.

Η ίδια διαδικασία με την κανονική περίοδο ακολουθήθηκε και στα στατιστικά δεδομένα που συλλέχθηκαν για τις ομάδες που πέρασαν στην φάση των playoffs. Στα γραφήματα χρονοσειρών σε αυτή την φάση της διοργάνωσης παρατηρήθηκαν τα ίδια συμπεράσματα με την κανονική περίοδο. Όμως συγκριτικά για τις δύο φάσεις της διοργάνωσης παρατηρήθηκαν διαφορές στις τιμές, μείωση, που λαμβάνουν με το πέρασμα των χρόνων ορισμένες μεταβλητές. Σε ανάλογα συμπεράσματα έχουν καταλήξει στην έρευνά τους οι Cabarkapa et al. (2022) οι οποίοι ανέφεραν αυτή την αναμενόμενη μείωση στις τιμές ορισμένων μεταβλητών, όταν οι ομάδες συμμετέχουν στη φάση των Playoffs. Όσον αφορά τα scatter plot για τις μεταβλητές αποδείχθηκε πως μόνο οι μεταβλητές των λαθών και των χαμένων ελεύθερων βολών είχαν αρνητική σχέση με τον δείκτη ενώ οι υπόλοιπες είχαν θετική συσχέτιση.

Η πρώτη ανάλυση που έγινε ήταν βάσει της μεταβλητής που δείχνει αν μια ομάδα προκρίθηκε στα playoffs ή όχι. Αρχικά, παρουσιάστηκαν τα περιγραφικά μέτρα και διάφορα γραφήματα χρονοσειρών, απ' όπου και συμπεράναμε πως σε όλες τις μεταβλητές που χρησιμοποιήθηκαν σ' αυτή την ανάλυση πλην των λαθών και των κατοχών (οι πόντοι, οι ασίστ, τα κλεψίματα, τα κοψίματα, τα ριμπάουντ, τα λάθη ανά αγώνα, ο δείκτης DEFF και οι κατοχές των ομάδων) οι ομάδες που προηγούνται στα playoffs λαμβάνουν μεγαλύτερες τιμές σε όλα τα χρόνια. Στη συνέχεια δημιουργήθηκαν και τα κατάλληλα boxplot των μεταβλητών, ώστε να γίνει η έρευνα για τις έκτροπες τιμές (outliers) που παρουσιάστηκαν αναλυτικά στην παράγραφο 3.2.3.1. Διερευνώντας τα boxplot παρατηρήθηκε πως η διάμεσος για τις ομάδες που προκρίθηκαν, βρισκόταν υψηλότερα απ' ό,τι για τις ομάδες που δεν προκρίθηκαν σε όλες τις μεταβλητές εκτός των λαθών ανά αγώνα και των κατοχών.

Στη δεύτερη ανάλυση τέθηκε προς επεξεργασία η μεταβλητή Conference που χωρίζει τις ομάδες σ' αυτές που συμμετέχουν στην περιφέρεια East και αυτές που συμμετέχουν στην West. Η συγκεκριμένη ανάλυση χωρίστηκε σε δύο κατηγορίες: πρώτα έγινε για τα στατιστικά δεδομένα της κανονικής περιόδου και έπειτα για τα playoffs. Πριν την ανάλυση που περιγράφηκε έγιναν τρείς διαφορετικές έρευνες για να δοθεί μια αρχική εικόνα στον αναγνώστη για τις περιφέρειες. Πρώτα αποδείχθηκε πως η ομάδα με το καλύτερο ποσοστό νικών ανήκε για οκτώ (8) σεζόν στην East περιφέρεια, με αποτέλεσμα να έχει και το πλεονέκτημα έδρας. Επίσης, αποδείχθηκε πως η West περιφέρεια έχει κατακτήσει εννέα (9) φορές τα τελευταία δεκαπέντε (15) χρόνια το πρωτάθλημα, γεγονός που δείχνει πως λογικά είναι η περιφέρεια με τα καλύτερα στατιστικά. Τέλος, φάνηκε πως μόνο τέσσερις (4) φορές έχει καταφέρει να

πάρει το πρωτάθλημα η ομάδα στην οποία ανήκε το καλύτερο ποσοστό νικών στην κανονική περίοδο.

Για τη φάση της κανονικής περιόδου παρουσιάστηκαν αρχικά τα περιγραφικά μέτρα και τα διαγράμματα χρονοσειρών, όπου και φάνηκε πως χρονικά οι ομάδες της East περιφέρειας λαμβάνουν μικρότερες τιμές στις μεταβλητές από την αντίπαλη περιφέρεια. Ακόμα σε αυτή την ανάλυση έγιναν τα violin plots, στα οποία φάνηκε πως οι διάμεσοι της West περιφέρειας λαμβάνουν μεγαλύτερες τιμές σε όλες τις μεταβλητές ετκός των ασίστ και των κοψιμάτων. Επίσης, μέσω των violin plots παρουσιάστηκαν οι διάφορες έκτροπες τιμές που υπήρχαν στα δεδομένα. Γι' αυτή την φάση της διοργάνωσης δόθηκαν και τα ιστογράμματα των μεταβλητών, όπου και φάνηκε πως εκτός από τα κλεψίματα, τα συνολικά ριμπάουντ και τα λάθη οι μέσοι όροι των μεταβλητών κατανέμονται μη-κανονικά και για τις δύο περιφέρειες.

Για τη φάση των playoffs ακολουθήθηκε η ίδια ανάλυση με τη φάση της κανονικής περιόδου. Στα διαγράμματα των χρονοσειρών σ' αυτή τη φάση της διοργάνωσης δεν υπάρχει ξεκάθαρη εικόνα, καθώς υπάρχουν χρονιές, όπου η East περιφέρεια λαμβάνει υψηλότερες τιμές από τη West και το αντίστροφο. Βάσει των violin plots τα αποτελέσματα που βρέθηκαν είναι τα ίδια με την προηγούμενη ανάλυση, με την διαφορά πως η διάμεσος στα λάθη που κάνουν οι ομάδες είναι υψηλότερη για την περιφέρεια East.

Στο τέταρτο κεφάλαιο, αρχικά εκτελέστηκε ο έλεγχος κανονικότητας για τις ποσοτικές μεταβλητές του συνόλου δεδομένων, για κάθε ανάλυση που πραγματοποιήθηκε στο προηγούμενο κεφάλαιο, και υπολογίστηκαν οι συσχετίσεις ανάμεσα τους. Έπειτα υλοποιήθηκαν δύο είδη ελέγχων υποθέσεων για την ισότητα των μέσων, δύο διαφορετικών δειγμάτων και ζευγαρωτών παρατηρήσεων, για τις μεταβλητές των πόντων που σκοράρει μια ομάδα, των ασίστ, των κλεψιμάτων, των κοψιμάτων, των συνολικών ριμπάουντ και των λαθών ανά αγώνα.

Αναφορικά με τους ελέγχους κανονικότητας, αυτοί υλοποιήθηκαν σε πρώτη φάση στο σύνολο δεδομένων της κανονικής περιόδου. Σύμφωνα με τον έλεγχο Shapiro-Wilk δεν υπήρξαν ενδείξεις απόρριψης της μηδενικής υπόθεσης, σε επίπεδο σημαντικότητας 5%, για τις μεταβλητές: ποσοστό ευστοχίας δίποντων ή τρίποντων, ποσοστό εύστοχων τρίποντων, επιθετικά ριμπάουντ, συνολικά ριμπάουντ, επιθετικά ριμπάουντ της αντίπαλης ομάδας, λάθη της ομάδας και της αντιπάλου, φάουλ και defensive rating. Αντίθετα, για όλες τις υπόλοιπες μεταβλητές απορρίφθηκε η υπόθεση της κανονικότητας. Ακόμα, για αυτή την ανάλυση παρουσιάστηκαν οι συσχετίσεις των μεταβλητών και πιο συγκεκριμένα αναλύθηκαν, στην παράγραφο 4.2.1.1, μόνο αυτές με τη μεταβλητή των κατοχών ανά αγώνα λόγω του όγκου των μεταβλητών.

Στις επόμενες αναλύσεις που παρουσιάστηκαν χρησιμοποιήθηκαν σαν μεταβλητές ανάλυσης οι πόντοι, οι ασίστ, τα κλεψίματα, τα κοψίματα, τα συνολικά ριμπάουντ, τα λάθη ανά αγώνα, ο δείκτης DEFF και οι κατοχές της ομάδας. Η πρώτη ανάλυση απ' αυτές έγινε βάσει της κατηγορικής μεταβλητής Playoffs, όπου οι ομάδες χωρίστηκαν σε δύο κατηγορίες. Για τις ομάδες που κατάφεραν να προκριθούν στη φάση των

playoffs η υπόθεση της κανονικότητας απορρίφθηκε για τις μεταβλητές: πόντοι που σκοράρει η ομάδα, ασίστ, κοψίματα και ο δείκτης DEFF, ενώ για τις μεταβλητές των κλεψιμάτων και των ριμπάουντ ανά αγώνα υπήρχαν ενδείξεις αποδοχής της μηδενικής υπόθεσης. Από τον πίνακα συσχετίσεων φάνηκε πως η μεταβλητή των πόντων ανά αγώνα έχει πολύ ισχυρή θετική συσχέτιση με τις κατοχές της ομάδας. Υψηλή συσχέτιση είχε με τις μεταβλητές των ασίστ και των συνολικών ριμπάουντ. Ασθενή συσχέτιση είχε με τις μεταβλητές των κλεψιμάτων, των κοψιμάτων και των λαθών ανά αγώνα, ενώ ασθενή αρνητική είχε με τον δείκτη DEFF. Για τις ομάδες που δεν κατάφεραν να προκριθούν στη δεύτερη φάση των playoffs η μόνη διαφορά για τους ελέγχους κανονικότητας είναι στη μεταβλητή των λαθών, όπου φάνηκε πως η υπόθεση του ελέγχου δεν απορρίφθηκε. Αντίστοιχα, σε ό,τι έχει να κάνει με τους συντελεστές συσχέτισης παρατηρούμε μια παρόμοια κατάσταση για τη μεταβλητή των κατοχών ανά αγώνα με τη διαφορά πως πλέον δεν υπάρχει μεταβλητή η οποία έχει αρνητική συσχέτιση μαζί της.

Τελευταία ανάλυση που έγινε είναι αυτή προς τη μεταβλητή Conference. Για τις ομάδες της East περιφέρειας οι έλεγχοι κανονικότητας έδειξαν πως για τις μεταβλητές των πόντων, των ασίστ, των κοψιμάτων ανά αγώνα και τον δείκτη DEFF απορρίπτεται η υπόθεση της κανονικότητας. Για τις συσχετίσεις των μεταβλητών με τις κατοχές ανά αγώνα ισχυρή θετική συσχέτιση είχε μόνο η μεταβλητή των πόντων ανά αγώνα. Μέτρια θετική συσχέτιση είχε με τις μεταβλητές των ασίστ και των συνολικών ριμπάουντ ανά αγώνα, ενώ με τις υπόλοιπες είχε ασθενή συσχέτιση. Για τις ομάδες που συμμετέχουν στην West περιφέρεια σε ό,τι αφορά τους ελέγχους κανονικότητας υπήρξε μία απόρριψη της μηδενικής υπόθεσης για τη μεταβλητή του δείκτη DEFF, ενώ για τις συσχετίσεις ισχύουν τα ίδια επίπεδα με την περιφέρεια East.

Τέλος, στο κλείσιμο του τέταρτου κεφαλαίου εκτελέστηκαν δύο ειδών έλεγχοι υποθέσεων για την ισότητα μέσω τιμών δύο δειγμάτων. Οι πρώτοι έλεγχοι που έγιναν ήταν για την ισότητα των μέσων τιμών δύο δειγμάτων βάσει των κατηγορικών μεταβλητών Playoffs και Conference. Βάσει της κατηγορικής μεταβλητής Playoffs οι μέσες τιμές των πόντων, των ασίστ, των κλεψιμάτων, των κοψιμάτων, των συνολικών ριμπάουντ και των λαθών διέφεραν σημαντικά. Οι ομάδες που προκρίνονταν κάθε φορά στη φάση των playoffs σημείωναν υψηλότερες τιμές για όλες τις μεταβλητές εκτός των λαθών ανά αγώνα. Για την κατηγορική μεταβλητή Conference βάσει του ίδιου ελέγχου φάνηκε πως έχουμε απόρριψη της μηδενικής υπόθεσης για τις μεταβλητές των πόντων, των κλεψιμάτων και των συνολικών ριμπάουντ, ενώ για τις υπόλοιπες δεν υπήρξαν αρκετές ενδείξεις απόρριψής της. Ο τελευταίος έλεγχος που διενεργήθηκε ήταν για την ισότητα μέσων τιμών για ζευγαρωτές παρατηρήσεις. Οι έλεγχοι έδειξαν πως οι μέσες τιμές των πόντων, των ασίστ, των κλεψιμάτων, των κοψιμάτων, των συνολικών ριμπάουντ και των λαθών διέφεραν σημαντικά ανάμεσα στις δύο φάσεις. Ειδικότερα, οι ομάδες φαίνεται πως στη φάση των playoffs κατάφερναν να λαμβάνουν μικρότερες τιμές στις μεταβλητές απ' ό,τι στην κανονική περίοδο. Τα συγκεκριμένα αποτελέσματα συμφωνούν, στον μεγαλύτερο βαθμό, με την έρευνα των Cabarkapa et al. (2022), οι οποίοι έχουν ερευνήσει τη στατιστική διαφορά ανάμεσα στα δεδομένα της κανονικής περιόδου και της φάσης των playoffs. Πιο

συγκεκριμένα, στην έρευνα τους αναφέρεται πως για όλες τις μεταβλητές, εκτός των κοψιμάτων, υπάρχουν στατιστικά σημαντικές διαφορές ανάμεσα στις δύο φάσεις, καθώς οι ομάδες έχουν μικρότερες τιμές στατιστικών στη φάση των playoffs. Επίσης και οι Mandić et al. (2019) κατέληξαν στο συμπέρασμα πως οι ομάδες αγωνιζόμενες στη φάση των playoffs ανά εκατό κατοχές πετυχαίνουν λιγότερες ασίστ και λάθη σε σχέση με την κανονική περίοδο. Φυσικά στην παρούσα εργασία, παρόλο που τα δεδομένα μας στις περισσότερες μεταβλητές δεν ακολουθούσαν την κανονική κατανομή, λόγω του Κεντρικού Οριακού Θεωρήματος για μεγάλο πλήθος δεδομένων (n>50), στις στατιστικές μεθόδους που αναφέρθηκαν θεωρήθηκε πως οι μεταβλητές ακολουθούν την κανονική κατανομή.

Στο πέμπτο κεφάλαιο ο στόχος ήταν να εξεταστούν ποιες μεταβλητές από το σύνολο δεδομένων παίζουν καθοριστικό ρόλο στο αν μία ομάδα καταφέρνει να περάσει στη φάση των playoffs και στην κατάκτηση του πρωταθλήματος της διοργάνωσης. Ακριβώς γι' αυτόν τον λόγο προσαρμόστηκαν τα κατάλληλα γενικευμένα γραμμικά μοντέλα, ενώ ως μεταβλητές απόκρισης χρησιμοποιήθηκαν οι Playoffs και Champions. Συνολικά προσαρμόστηκαν έξι (6) διαφορετικά μοντέλα, από τα οποία τα δύο πρώτα χρησιμοποίησαν όλες τις μεταβλητές που υπήρχαν στο σύνολο δεδομένων της κανονικής περιόδου και προσπάθησαν να ερευνήσουν τις σημαντικότερες μεταβλητές και για τις δύο μεταβλητές απόκρισης. Ακολούθως, το επόμενο μοντέλο στηρίχτηκε στο σύνολο δεδομένων των playoffs και κατά συνέπεια χρησιμοποιήθηκε ως μεταβλητή απόκρισης μόνο η μεταβλητή Champions, αφού έγινε χρήση όλων των στατιστικών για τη φάση των playoffs. Στα τελευταία τρία μοντέλα έγιναν οι ίδιες αναλύσεις με τη διαφορά πως δεν χρησιμοποιήθηκαν ως μεταβλητές αυτές των δεικτών offensive rating και defensive rating.

Για το σύνολο δεδομένων της κανονικής περιόδου και με μεταβλητή απόκρισης την Playoffs, το τελικό μοντέλο που προσαρμόστηκε είχε ως στατιστικά σημαντικές μεταβλητές μόνο τους δείκτες offensive rating και defensive rating. Παρόμοια αποτελέσματα παρουσίασαν οι Teramoto et al. (2010), που στην έρευνά τους διαπίστωσαν πως οι μεταβλητές defensive rating και offensive είναι απαραίτητες για την νίκη μίας ομάδας στην κανονική περίοδο.

Η ίδια διαδικασία με μεταβλητή απόκρισης την Champions είχε διαφορετικά αποτελέσματα. Το καταλληλότερο μοντέλο φάνηκε πως ήταν αυτό που είχε ως μεταβλητές το ποσοστό των εύστοχων δίποντων και τρίποντων, τον δείκτη defensive rating και τα εύστοχα δίποντα και τρίποντα.

Για το σύνολο δεδομένων που αφορά τη φάση των Playoffs, όπως προαναφέραμε, χρησιμοποιήθηκε ως μεταβλητή απόκρισης μόνο η μεταβλητή Champions. Το μοντέλο που είχε την καλύτερη προσαρμογή είναι αυτό με τις μεταβλητές των δεικτών ORt και DRt και της μεταβλητής των επιθετικών ριμπάουντ.

Στις επόμενες αναλύσεις που παρουσιάστηκαν, όπως αναφέρθηκε παραπάνω, δεν χρησιμοποιήθηκαν εξαρχής οι δείκτες offensive rating και defensive rating. Η πρώτη ανάλυση για την κανονική περίοδο είχε ως μεταβλητές στο καταλληλότερο μοντέλο τα

λάθη της ομάδας, τα ποσοστά εύστοχων δίποντων και τρίποντων, τα επιθετικά ριμπάουντ και τους πόντους που πετυχαίνει η ομάδα.

Στην επόμενη ανάλυση χρησιμοποιήθηκε ως μεταβλητή απόκρισης η Champions στο σύνολο δεδομένων της κανονικής περιόδου. Σημαντικές μεταβλητές γι' αυτό το μοντέλο βρέθηκαν οι πόντοι ανά αγώνα, οι συνολικές προσπάθειες δίποντων και τρίποντων, όσο και των ελεύθερων βολών, τα επιθετικά ριμπάουντ και τα λάθη της αντίπαλης ομάδας.

Τέλος, η ανάλυση που έγινε για το σύνολο δεδομένων των playoffs έδειξε πως οι σημαντικότερες μεταβλητές, ως προς την μεταβλητή Champions, ήταν οι πόντοι της ομάδας και οι συνολικές προσπάθειες για δίποντα και τρίποντα.

Στο έκτο κεφάλαιο στόχος ήταν να γίνει μία ανάλυση ορισμένων μεταβλητών που επιλέχθηκαν από το σύνολο δεδομένων της κανονικής περιόδου με τη μορφή χρονοσειρών. Οι μεταβλητές που επιλέχθηκαν είναι: οι πόντοι, οι ασίστ, τα ριμπάουντ, τα κλεψίματα και τα κοψίματα της ομάδας ανά αγώνα. Επίσης, μέσω του κατάλληλου μοντέλου έγιναν και οι προβλέψεις για τις δύο (2) επόμενες σεζόν.

Για την ανάλυση των πόντων ανά αγώνα δημιουργήθηκαν τα κατάλληλα γραφήματα και φάνηκε πως η χρονοσειρά είχε αυξητική τάση και δεν ήταν στάσιμη. Εφόσον αφαιρέθηκε η τάση και δημιουργήθηκε η νέα στάσιμη χρονοσειρά, ppg, βρέθηκε ένα κατάλληλο μοντέλο ARIMA που μπορεί να την περιγράψει, το οποίο ήταν το ARIMA(0,0,2). Αντίστοιχα για τις ασίστ ανά αγώνα παρουσιάστηκε πάλι αυξητική τάση στην αρχική χρονοσειρά και αφού δημιουργήθηκε η στάσιμη χρονοσειρά, apg, βρέθηκε ένα κατάλληλο μοντέλο περιγραφής ώστε να γίνουν οι προβλέψεις της, ARIMA(1,0,1). Για τη χρονοσειρά των κλεψιμάτων δημιουργήθηκε μια νέα στάσιμη χρονοσειρά, spg, βρέθηκε ένα μοντέλο που την περιγράφει κατάλληλα, το ARIMA(0,0,1). Για τα κοψίματα αφού αφαιρέθηκε η εποχικότητα της χρονοσειράς και στην συνέχει έγινε στάσιμη αποδείχθηκε πως ένα κατάλληλο μοντέλο περιγραφής ήταν το $ARIMA(0,3,2)(0,1,0)_3$. Τέλος, για τη χρονοσειρά των ριμπάουντ ανά αγώνα σαν καταλληλότερο μοντέλο περιγραφής της νέας στάσιμης χρονοσειράς, rpg, βρέθηκε το ARIMA(0,0,2).

Στο έβδομο κεφάλαιο με τη βοήθεια τεχνικών μηχανικής μάθησης είχαμε ως στόχο την εξαγωγή χρήσιμων πληροφοριών. Πιο συγκεκριμένα, αφού εκτελέσαμε δύο διαφορετικές μεθόδους για την επιλογή των καταλληλότερων μεταβλητών (feature selection) που ήταν χρήσιμες για την διεξαγωγή των αλγορίθμων, χρησιμοποιήσαμε τεχνικές κατηγοριοποίησης (K-κοντινότερων γειτόνων, τυχαίου δάσους και Support Vector Machines) προκειμένου να προσαρμόσουμε κατάλληλα μοντέλα ταξινόμησης βάσει της μεταβλητής Playoffs. Έπειτα χρησιμοποιήσαμε και μεθόδους ομαδοποίησης για να εντοπίσουμε μοτίβα παρατηρήσεων με παρόμοια χαρακτηριστικά μέσα στα δεδομένα μας, με την βοήθεια των αλγορίθμων K-Means και Birch.

Η πρώτη μέθοδος επιλογής των καταλληλότερων μεταβλητών ήταν η μέθοδος Boruta, η οποία ανέδειξε ως σημαντικότερες τις μεταβλητές των πόντων, του δείκτη offensive rating, των κατοχών της ομάδας, των λαθών, των κλεψιμάτων, των

συνολικών ριμπάουντ, των αμυντικών ριμπάουντ, του ποσοστού εύστοχων τριπόντων, των προσπαθειών δίποντων και τρίποντων της αντίπαλης ομάδας, των επιτυχημένων τρίποντων, των πόντων που σκοράρει η αντίπαλη ομάδα, των επιτυχημένων δίποντων και τρίποντων, των συνολικών προσπαθειών δίποντων και τρίποντων, των χαμένων δίποντων και τρίποντων, του ποσοστού δίποντων και τρίποντων και του δείκτη defensive rating.

Αναφορικά με την κατηγοριοποίηση που εκτελέστηκε ως προς τη μεταβλητή Playoffs, και οι τρεις μέθοδοι δημιούργησαν αρκετά αποτελεσματικά μοντέλα. Πιο συγκεκριμένα, η μέθοδος K-κοντινότερων γειτόνων μας έδωσε accuracy 0.859, η μέθοδος του τυχαίου δάσους 0.866 και η μέθοδος SVM 0.889. Συνεπώς, επιλέχθηκε ως καλύτερη μέθοδος η SVM.

Στη συνέχεια, για τις ίδιες μεταβλητές εκτελέστηκαν και οι δύο μέθοδοι συσταδοποίησης, οι οποίες δεν έδωσαν καλά αποτελέσματα. Αρχικά, με τη μέθοδο του αγκώνα βρέθηκε πως το κατάλληλο πλήθος συστάδων που θα έπρεπε να δημιουργηθεί είναι δύο (2). Επομένως, οι αλγόριθμοι αναγκάστηκαν να χωρίσουν τα δεδομένα σε δύο (2) διαφορετικές συστάδες ανάλογα τις επιδόσεις στις μεταβλητές που επιλέχθηκαν. Η μέθοδος K-Means είχε για τον δείκτη silhouette τιμή ίση με 0.284 ενώ για τον Calinski Harabasz 227.136. Αντίστοιχα, η μέθοδος Birch είχε τιμές 0.256 και 176.217. Όπως προαναφέρθηκε, οι συσταδοποιήσεις δεν είχαν καλά αποτελέσματα και γι' αυτόν τον λόγο έγινε περαιτέρω διερεύνηση. Παρατηρώντας την ταξινόμηση των παρατηρήσεων με βάση την μέθοδο K-Means φάνηκε πως ο αλγόριθμος ταξινομούσε σχεδόν όλες τις παρατηρήσεις των τελευταίων επτά (7) σεζόν σε μία συστάδα και τις υπόλοιπες σε μία άλλη. Επομένως, δημιουργήθηκε ένα νέο σετ δεδομένων, σύμφωνα με το οποίο χωρίστηκαν οι ομάδες με τις αποδόσεις τους σε δύο κατηγορίες, από τη σεζόν 2008-2009 έως και τη σεζόν 2014-2015 η πρώτη, ενώ η δεύτερη από 2016-2017 έως την περσινή. Στην ουσία εξετάσαμε αν το επίπεδο της διοργάνωσης διαφέρει σε απόκλιση οκτώ (8) ετών. Όπως φάνηκε από τους ελέγχους για την ισότητα μέσων τιμών, paired t-test, υπάρχει στατιστικά σημαντική διαφορά στις τιμές που συλλέχθηκαν. Επομένως, θεωρήθηκε πως αυτός είναι και ο λόγος που δεν επιτεύχθηκαν καλές ταξινομήσεις.

Η δεύτερη μέθοδος επιλογής των σημαντικότερων μεταβλητών ήταν η μέθοδος K-καλύτερων χαρακτηριστικών, στην οποία επιλέξαμε το πλήθος των σημαντικότερων μεταβλητών να είναι ίσο με οκτώ (8). Οι μεταβλητές ήταν: οι πόντοι της ομάδας και της αντιπάλου, τα χαμένα δίποντα και τρίποντα, τα επιτυχημένα και οι συνολικές προσπάθειες για τρίποντα, οι συνολικές προσπάθειες δίποντων και τρίποντων της αντίπαλής ομάδας και οι δείκτες offensive και defensive rating.

Για την κατηγοριοποίηση των δεδομένων βάσει των νέων μεταβλητών υπήρξαν μικρές μειώσεις στον δείκτη accuracy σε σχέση με τη μέθοδο Boruta. Για τη μέθοδο των K-κοντινότερων γειτόνων ο δείκτης έλαβε την τιμή 0.851, για τη μέθοδο του τυχαίου δάσους την τιμή 0.837 και για τη μέθοδο SVM την τιμή 0.866, που και πάλι ήταν η υψηλότερη ανάμεσα στις τρεις μεθόδους που υλοποιήθηκαν. Συνολικά για τις κατηγοριοποιήσεις καταλήξαμε πως η προτιμότερη είναι αυτή με τη μέθοδο SVM και

με τις μεταβλητές που υποδείχτηκαν από τη μέθοδο επιλογής καλύτερων μεταβλητών Boruta.

Τέλος, για την συσταδοποίηση με τις νέες μεταβλητές οι αλγόριθμοι εξακολούθησαν να μην έχουν καλά αποτελέσματα. Καλύτερη συσταδοποίηση φάνηκε πως έγινε με τη μέθοδο K-Means και τη μέθοδο επιλογής καλύτερων μεταβλητών Boruta.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική

Αντζουλάκος, Δ., (2021). Σημειώσεις μαθήματος «Ανάλυση δεδομένων με τη χρήση στατιστικών πακέτων: Εισαγωγή στην R», Τμήμα Στατιστικής & Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς, ΠΜΣ «Εφαρμοσμένη Στατιστική», Κ. 6 σ. 27.

Ευγγελάρας, Χ., (2022). Σημειώσεις μαθήματος «Ανάλυση δεδομένων με τη χρήση στατιστικών πακέτων: Σημειώσεις για το IBM SPSS Statistics», Τμήμα Στατιστικής & Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς, ΠΜΣ «Εφαρμοσμένη Στατιστική», σ. 46.

Καλλιακμάνης, Δ.Γ., (2020). Στατιστικά μοντέλα για την απόδοση μιας ομάδας μπάσκετ: ποια στατιστικά στοιχεία είναι καθοριστικά για την απόδοση της ομάδας, σε ετήσια βάση. Προσβάσιμο στο: https://dione.lib.unipi.gr/xmlui/handle/unipi/12717

Καμίτσης, Α., (2023). Δείκτες για την αξιολόγηση της απόδοσης παικτών και ομάδων σε αγώνες μπάσκετ και παράγοντες που τους επηρεάζουν. https://dione.lib.unipi.gr/xmlui/handle/unipi/15593

Κάτρης, Χ. (2021). Σημειώσεις μαθήματος «Εισαγωγή στην ανάλυση δεδομένων», Τμήμα Μηχανολόγων και Αεροναυπηγών, Πανεπιστήμιο Πατρών, Δ.1 σ. 25. https://eclass.upatras.gr/modules/document/file.php/MATH1244/ BALE KAI TO ALLO

Κολύβα-Μαχαίρα, Φ., Μπόρα-Σέντα, Ε., (2013). Στατιστική, Εκδόσεις Ζήτη, 331-334.

Κούτρας, Μ., (2020). Σημειώσεις μαθήματος «Ανάλυση παλινδρόμησης και ανάλυση διακύμανσης», Τμήμα Στατιστικής & Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς, ΠΜΣ «Εφαρμοσμένη Στατιστική», σ. 114.

Μπούτσικας, Μ., (2004). Σημειώσεις μαθήματος «Στατιστικά Προγράμματα», Τμήμα Στατιστικής & Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς, σ. 32. https://www.samos.aegean.gr/actuar/zimste/notes/SPSS_lesson5-6.pdf

Πελέκης Ν., (2022). Σημειώσεις μαθήματος «Στατιστικές Μέθοδοι Εξόρυξης Δεδομένων», Τμήμα Στατιστικής & Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς, ΠΜΣ «Εφαρμοσμένη Στατιστική», Κ. 1 σ. 7.

Πολίτης, Κ., (2021). Σημειώσεις μαθήματος «Γενικευμένα Γραμμικά Μοντέλα», Τμήμα Στατιστικής & Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς, ΠΜΣ «Εφαρμοσμένη Στατιστική», σ. 13.

Ρακιτζής, Α., (2023). Σημειώσεις μαθήματος «Πρόβλεψη - Χρονοσειρές», Τμήμα Στατιστικής & Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς, ΠΜΣ «Εφαρμοσμένη Στατιστική».

Σπυριδάκης, Α.Ε., (2022). Στατιστική ανάλυση για τους παράγοντες που επηρεάζουν την απόδοση των ομάδων ποδοσφαίρου στις ευρωπαϊκές διοργανώσεις. https://dione.lib.unipi.gr/xmlui/handle/unipi/14180

Τριανταφύλλου, Ι.Σ (2023). Σημειώσεις μαθήματος «Πρόβλεψη - Χρονοσειρές», Τμήμα Στατιστικής & Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς, ΠΜΣ «Εφαρμοσμένη Στατιστική».

Ξένη

Adele C., Cutler D. R., Stevens R. J. (2021). *Random Forest*. https://www.researchgate.net/publication/236952762_Random_Forests

Awais M., Iqbal S., Rasool Q., Kousar T., (2022). *Optical Character Recognition of Urdu Text using Histogram of Oriented Gradient Features*, Research Square. v1_covered.pdf (researchsquare.com)

Belkacem, S (2021) *Machine learning approaches to rank news feed updates on social media*. PhD Thesis, University of Science and Technology Houari Boumediene, Algeria 4: Classification vs. Regression | Download Scientific Diagram (researchgate.net)

Benzer N. J., (2022). *Balanced iterative reducing and clustering using hierarchies(BIRCH)*, Medium. The BIRCH clustering algorithm explained | Medium

Bishop, C.M. (2006) *Pattern recognition and machine learning*, New York: Springer.

Breiman, L. (2001) *Random Forests*, Machine Learning, 45, pp. 5–32 https://doi.org/10.1023/A:1010933404324

Casals, M., Martinez, J.A., (2013). *Modelling player performance in basketball through mixed models*. International Journal of Performance Analysis in Sport, 64-82. https://www.researchgate.net/publication/235992657_Modelling_player_performance_in_basketball_through_mixed_models

Çene, E., (2018). *What is the difference between a winning and a losing team: insights from Euroleague basketball*. International Journal of Performance Analysis in Sport. 18:1, 55-68. https://www.tandfonline.com/doi/abs/10.1080/24748668.2018.1446234?journalCode=rpan20

De Castro M., Zona U., Bocci F. (2020). *Dalle Teaching Machines al Machine Learning*, Padova University Press, pp 27-33. (PDF) "L'apprendimento macchinico tra Skinner box e Deep Reinforcement Learning. Rischi e opportunità. Machine Learning between Skinner box and Deep Reinforcement Learning. Risks and opportunities", in "Dalle Teaching Machines al Machine Learning" a cura di Graziano Cecchinato, Valentina Grion - PREPRINT (researchgate.net)

Evans, R. H. (1996). *An Analysis of Criterion Variable Reliability in Conjoint Analysis*, Perceptual and Motor Skills, 82(3), pp. 988–990. https://doi.org/10.2466/pms.1996.82.3.988

Ezugwu A.E.S., Shukla A.K., Agbaje M., Jose-Garcia A., (2021). *Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature*, Neural Computing and Applications, pp 14-15. (PDF) Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature (researchgate.net)

Fidan U., Uzunhisarcikli E., Clikusu I., (2020). *Classification of Dermatological Data with Self Organizing Maps and Support Vector Machine*. (PDF) Classification of Dermatological Data with Self Organizing Maps and Support Vector Machine Dermatolojik Verilerin Öz Düzenleyici Harita ve Destek Vektör Makinaları ile Sınıflandırılması Classification of Dermatological Data with Self Organizing Maps and Support Vector Machine, Fidan et al. 895 (researchgate.net)

González, T., Santos, D., Wang, C., Carlsson, N., Lambrix, P., (2021). *Predicting Season Outcomes for the NBA*. https://www.researchgate.net/publication/353193640_Predicting_Season_Outcomes_for_the_NBA

Han J., Kamber M, Pei J., (2012). *Data Mining Concepts and Techniques*, Morgan Kaufmann, pp 6-8. https://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

Harrison O., (2018). *Machine Learning Basics with the K-Nearest Neighbors Algorithm*. https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

Jeffares A. (2019) *K-Means: A complete Introduction*, Towards Data Science. K-means: A Complete Introduction. K-means is an unsupervised clustering… | by Alan Jeffares | Towards Data Science

Jones, M.B., (2007). *Home Advantage in the NBA as a Game-Long Process*. Journal of Quantitative Analysis in Sports, Vol. 3: Iss.4, Article 2. https://www.degruyter.com/document/doi/10.2202/1559-0410.1081/pdf

Khalid S., Khan M.A., Mazliham M.S., (2022). *Predicting Risk through Artificial Intelligence Based on Machine Learning Algorithms: A Case of Pakistani Nonfinancial Firms*. Illustration of KNN technique (adapted from JavaTpoint). | Download Scientific Diagram (researchgate.net)

Khan M., (2021). *Discuss the elbow method*. Discuss the elbow method. (csias.in)

Kim, T.K, (2015). *T test as a parametric statistic*. Korean Journal of Anesthesiology. 540-546. https://synapse.koreamed.org/articles/1156170

Kubatko, J., Oliver, D., Pelton, K., Rosenbaum, D.T., (2007). *A starting Point for Analyzing Basketball Statistics.* Journal of Quantitative Analysis in Sports, Vol. 3: Iss. 3, Article 1. http://vishub.org/officedocs/18024.pdf

Mandić, R., Jakovljević, S., Erčulj, F., Sˇtrumbelj, E., (2019). *Trends in NBA and Euroleague basketball: Analysis and comparison of statistical data from 2000 to 2017*. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0223524

McCleary, R., Hay, R. A., Meidinger, E. E., and McDowall, D. (1980). *Applied time series analysis for the social sciences*. Beverly Hills, CA: Sage.

Nguyena, N.H., Nguyenb, D.T., Maa, B., Hua, J., (2021). *The application of machine learning and deep learning in sport: predicting NBA players' performance and popularity.* Journal of Information and Telecommunication, 6:2, 217-235. The application of machine learning and deep learning in sport: predicting NBA players' performance and popularity (tandfonline.com)

Puente, C., Del Coso, J., Salinero, J.J., Abián-Vicén, J., (2015). *Basketball performance indicators during the ACB regular season from 2003 to 2013.* International Journal of Performance Analysis in Sport, 935-948. https://www.tandfonline.com/doi/abs/10.1080/24748668.2015.11868842

Sanjeev Lingam-Nattamai, (2022). *How the three-point shot is changing the NBA, Deep Dives with Data*. https://medium.com/deep-dives-with-data/how-the-three-point-shot-is-changing-the-nba-3fa312e9dc21

Schober, Patrick & Boer, Christa & Schwarte, Lothar. (2018). *Correlation Coefficients: Appropriate Use and Interpretation*. Anesthesia & Analgesia. 126. 1. 10.1213/ANE.0000000000002864

Schölkopf, B. and Smola, A.J. (2002) *Learning with kernels: Support vector machines, regularization, optimization, and beyond.* Cambridge, MA: MIT Press.

Suumers, M.R., (2013). *How to Win in the NBA Playoffs: A Statistical Analysis.* http://www.na-businesspress.com/AJM/SummersMR_Web13_3_.pdf

Tan, P.-N. et al. (2018) *Introduction to data mining, 2nd edn*. Boston, MA: Pearson Education.

Teramoto, M., Croos, (2010). *Relative Importance of Performance Factors in Winning NBA Games in Regular Season versus Playoffs.* Journal of Quantitative Analysis in Sports, Vol. 6: Iss. 3, 1-19. https://www.degruyter.com/document/doi/10.2202/1559-0410.1260/html

Zhang, R. and L., *BIRCH: an efficient data clustering method for very large databases* in ACM Sigmod Record, ACM, vol. 25, pp. 103–114

Σύνδεσμοι

https://basketball.realgm.com/

https://logos-world.net/nba-logo/

https://builtin.com/data-science/boxplot

https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot

https://www.scribbr.com/

https://www.scaler.com/topics/kdd-in-data-mining/

https://www.geeksforgeeks.org/supervised-machine-learning/

https://www.altexsoft.com/blog/semi-supervised-learning/

https://www.javatpoint.com/feature-selection-techniques-in-machine-learning

https://medium.com/@mrmaster907/introduction-random-forest-classification-by-example-6983d95c7b91

https://manisha-sirsat.blogspot.com/2019/04/confusion-matrix.html

https://www.researchgate.net/figure/A-F-Scatter-plots-with-data-sampled-from-simulated-bivariate-normal-distributions-with_fig1_323388613

https://en.wikipedia.org/wiki/Season_(sports)

https://en.wikipedia.org/wiki/Playoffs

https://en.wikipedia.org/wiki/1949%E2%80%9350_NBA_season
https://en.wikipedia.org/wiki/Collective_bargaining
https://el.wikipedia.org/wiki/COVID-19

https://en.wikipedia.org/wiki/2019%E2%80%9320_NBA_season

https://en.wikipedia.org/wiki/Efficiency_(basketball)

https://www.nbastuffer.com/analytics101/win-score/

https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

https://www.scribbr.com/

https://builtin.com/data-science/boxplot

ΠΑΡΑΡΤΗΜΑ

Π1. Κώδικας υλοποίησης της παρούσας εργασίας

Παρακάτω επισυνάπτεται ο κώδικας της εργασίας για το σύνολο δεδομένων που αναφέρεται στα δεδομένα από την κανονική περίοδο διότι οι εντολές που χρησιμοποιήθηκαν για τα playoffs είναι ίδιες.

```
#Εισαγωγή βιβλιοθηκών
import pandas as pd
import numpy as np
from numpy import array
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.express as px
from sklearn import linear_model
from scipy import stats
from statsmodels.stats.diagnostic import Lilliefors
import statsmodels.api as sm
import statsmodels.formula.api as smf
from patsy import dmatrices
from statsmodels.stats.outliers_influence import variance_inflation_factor
from scipy.stats import shapiro
from scipy.stats import wilcoxon
from boruta import BorutaPy
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
from sklearn.feature_selection import f_regression
from tqdm.notebook import tqdm
from sklearn.decomposition import PCA
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import accuracy_score
from sklearn.metrics import, classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import f1_score
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from sklearn.metrics import calinski_harabasz_score
from sklearn.cluster import Birch

#ΚΕΦΑΛΑΙΟ 3Ο

#Εισαγωγή δεδομένων
df = pd.read_excel('regularseason1.xlsx')
```

```
df=pd.DataFrame(df)

#Χαρακτηριστικά των μεταβλητών
df.describe()

#Time Series Plot
plt.subplot(3,3,1)
sns.lineplot(x='Year',y='PPG',data=df,marker ='.',color="k")
plt.subplot(3,3,2)
sns.lineplot(x='Year',y='PPGA',data=df,marker ='.',color="b")
plt.subplot(3,3,3)
sns.lineplot(x='Year',y='FGM',data=df,marker ='.',color="m")
plt.subplot(3,3,4)
sns.lineplot(x='Year',y='FGA',data=df,marker ='.',color="c")
plt.subplot(3,3,5)
sns.lineplot(x='Year',y='MisFG',data=df,marker ='.',color="g")
plt.subplot(3,3,6)
sns.lineplot(x='Year',y='FG%',data=df,marker ='.',color="r")
plt.subplot(3,3,7)
sns.lineplot(x='Year',y='FGAo',data=df,marker ='.',color="y")
plt.subplot(3,3,8)
sns.lineplot(x='Year',y='3PM',data=df,marker ='., color="brown")
plt.subplot(3,3,9)
sns.lineplot(x='Year',y='3PA',data=df,marker ='.',color="r")
plt.tight_layout()
plt.figure(figsize=(10000,6000))
plt.show()
plt.subplot(3,3,1)
sns.lineplot(x='Year',y='3P%',data=df,marker ='.',color="orange")
plt.subplot(3,3,2)
sns.lineplot(x='Year',y='FTM',data=df,marker ='.',color="y")
plt.subplot(3,3,3)
sns.lineplot(x='Year',y='FTA',data=df,marker ='.',color="k")
plt.subplot(3,3,4)
sns.lineplot(x='Year',y='MisFt',data=df,marker ='.',color="b")
plt.subplot(3,3,5)
sns.lineplot(x='Year',y='FT%',data=df,marker ='.',color="m")
plt.subplot(3,3,6)
sns.lineplot(x='Year',y='FTAo',data=df,marker ='.',color="c")
plt.subplot(3,3,7)
sns.lineplot(x='Year',y='ORB',data=df,marker ='.',color="g")
plt.subplot(3,3,8)
sns.lineplot(x='Year',y='DRB',data=df,marker ='.',color="k")
plt.subplot(3,3,9)
sns.lineplot(x='Year',y='RPG',data=df,marker ='.',color="b")
plt.tight_layout()
plt.figure(figsize=(10000,6000))
plt.show()
plt.subplot(3,3,1)
sns.lineplot(x='Year',y='ORBo',data=df,marker ='.',color="m")
```

```
plt.subplot(3,3,2)
sns.lineplot(x='Year',y='APG',data=df,marker ='.',color="c")
plt.subplot(3,3,3)
sns.lineplot(x='Year',y='SPG',data=df,marker ='.',color="g")
plt.subplot(3,3,4)
sns.lineplot(x='Year',y='BPG',data=df,marker ='.',color="r")
plt.subplot(3,3,5)
sns.lineplot(x='Year',y='TOV',data=df,marker ='.',color="y")
plt.subplot(3,3,6)
sns.lineplot(x='Year',y='TOVo',data=df,marker ='.', color="brown")
plt.subplot(3,3,7)
sns.lineplot(x='Year',y='PF',data=df,marker ='.',color="r")
plt.subplot(3,3,8)
sns.lineplot(x='Year',y='POSSt',data=df,marker ='.',color="orange")
plt.subplot(3,3,9)
sns.lineplot(x='Year',y='ORt',data=df,marker ='.',color="y")
plt.tight_layout()
plt.figure(figsize=(10000,6000))
plt.show()
plt.subplot(2,2,1)
sns.lineplot(x='Year',y='POSSo',data=df,marker ='.',color="k")
plt.subplot(2,2,2)
sns.lineplot(x='Year',y='DRt',data=df,marker ='.',color="b")
plt.subplot(2,2,3)
sns.lineplot(x='Year',y='DEFF',data=df,marker ='.',color="m")
plt.subplot(2,2,4)
sns.lineplot(x='Year',y='EFF',data=df,marker ='.',color="c")
plt.tight_layout()
plt.figure(figsize=(10000,6000))
plt.show()
```

#Scatter Plots των μεταβλητών με τον δείκτη EFF

```
plt.subplot(4,4,1)
plt.scatter(df['PPG'],df['EFF'],alpha=0.5)
plt.xlabel("PPG")
plt.ylabel("EFF")
plt.subplot(4,4,2)
plt.scatter(df['PPGA'],df['EFF'],alpha=0.5)
plt.xlabel("PPGA")
plt.ylabel("EFF")
plt.subplot(4,4,3)
plt.scatter(df['FGM'],df['EFF'],alpha=0.5)
plt.xlabel("FGM")
plt.ylabel("EFF")
plt.subplot(4,4,4)
plt.scatter(df['FGA'],df['EFF'],alpha=0.5)
plt.xlabel("FGA")
plt.ylabel("EFF")
plt.subplot(4,4,5)
plt.scatter(df['MisFG'],df['EFF'],alpha=0.5)
```

```
plt.xlabel("MisFG")
plt.ylabel("EFF")
plt.subplot(4,4,6)
plt.scatter(df['FG%'],df['EFF'],alpha=0.5)
plt.xlabel("FG%")
plt.ylabel("EFF")
plt.subplot(4,4,7)
plt.scatter(df['FGAo'],df['EFF'],alpha=0.5)
plt.xlabel("FGAo")
plt.ylabel("EFF")
plt.subplot(4,4,8)
plt.scatter(df['3PM'],df['EFF'],alpha=0.5)
plt.xlabel("3PM")
plt.ylabel("EFF")
plt.subplot(4,4,9)
plt.scatter(df['3PA'],df['EFF'],alpha=0.5)
plt.xlabel("3PA")
plt.ylabel("EFF")
plt.subplot(4,4,10)
plt.scatter(df['3P%'],df['EFF'],alpha=0.5)
plt.xlabel("3P%")
plt.ylabel("EFF")
plt.subplot(4,4,11)
plt.scatter(df['FTM'],df['EFF'],alpha=0.5)
plt.xlabel("FTM")
plt.ylabel("EFF")
plt.subplot(4,4,12)
plt.scatter(df['FTA'],df['EFF'],alpha=0.5)
plt.xlabel("FTA")
plt.ylabel("EFF")
plt.subplot(4,4,13)
plt.scatter(df['MisFt'],df['EFF'],alpha=0.5)
plt.xlabel("MisFT")
plt.ylabel("EFF")
plt.subplot(4,4,14)
plt.scatter(df['FT%'],df['EFF'],alpha=0.5)
plt.xlabel("FT%")
plt.ylabel("EFF")
plt.subplot(4,4,15)
plt.scatter(df['FTAo'],df['EFF'],alpha=0.5)
plt.xlabel("FTAo")
plt.ylabel("EFF")
plt.subplot(4,4,16)
plt.scatter(df['ORB'],df['EFF'],alpha=0.5)
plt.xlabel("ORB")
plt.ylabel("EFF")
plt.tight_layout()
plt.figure(figsize=(10000,6000))
plt.show()
plt.subplot(4,4,1)
```

```
plt.scatter(df['DRB'],df['EFF'],alpha=0.5)
plt.xlabel("DRB")
plt.ylabel("EFF")
plt.subplot(4,4,2)
plt.scatter(df['RPG'],df['EFF'],alpha=0.5)
plt.xlabel("RPG")
plt.ylabel("EFF")
plt.subplot(4,4,3)
plt.scatter(df['ORBo'],df['EFF'],alpha=0.5)
plt.xlabel("ORBo")
plt.ylabel("EFF")
plt.subplot(4,4,4)
plt.scatter(df['APG'],df['EFF'],alpha=0.5)
plt.xlabel("APG")
plt.ylabel("EFF")
plt.subplot(4,4,5)
plt.scatter(df['SPG'],df['EFF'],alpha=0.5)
plt.xlabel("SPG")
plt.ylabel("EFF")
plt.subplot(4,4,6)
plt.scatter(df['BPG'],df['EFF'],alpha=0.5)
plt.xlabel("BPGF")
plt.ylabel("EFF")
plt.subplot(4,4,7)
plt.scatter(df['TOV'],df['EFF'],alpha=0.5)
plt.xlabel("TOV")
plt.ylabel("EFF")
plt.subplot(4,4,8)
plt.scatter(df['TOVo'],df['EFF'],alpha=0.5)
plt.xlabel("TOVo")
plt.ylabel("EFF")
plt.subplot(4,4,9)
plt.scatter(df['PF'],df['EFF'],alpha=0.5)
plt.xlabel("PF")
plt.ylabel("EFF")
plt.subplot(4,4,10)
plt.scatter(df['POSSt'],df['EFF'],alpha=0.5)
plt.xlabel("POSSt")
plt.ylabel("EFF")
plt.subplot(4,4,11)
plt.scatter(df['ORt'],df['EFF'],alpha=0.5)
plt.xlabel("ORt")
plt.ylabel("EFF")
plt.subplot(4,4,12)
plt.scatter(df['POSSo'],df['EFF'],alpha=0.5)
plt.xlabel("POSSo")
plt.ylabel("EFF")
plt.subplot(4,4,13)
plt.scatter(df['DRt'],df['EFF'],alpha=0.5)
plt.xlabel("DRt")
```

```
plt.ylabel("EFF")
plt.subplot(4,4,14)
plt.scatter(df['DEFF'],df['EFF'],alpha=0.5)
plt.xlabel("DEFF")
plt.ylabel("EFF")
plt.tight_layout()
plt.figure(figsize=(10000,6000))
plt.show()

#Ανάλυση με βάση την πρόκριση των ομάδων στα Playoffs
#Χαρακτηριστικά των μεταβλητών που επιλέχθηκαν γι' αυτή την ανάλυση
df2=df.groupby('Playoffs')
df2['PPG','APG','SPG','BPG','RPG','TOV','EFF','DEFF','POSSt'].describe()

#Time Series Plot
plt.subplot(121)
sns.lineplot(x='Year',y='PPG',data=df,hue='Playoffs')
plt.subplot(122)
sns.lineplot(x='Year',y='APG',data=df,hue='Playoffs')
plt.tight_layout()
plt.show()
plt.subplot(121)
sns.lineplot(x='Year',y='SPG',data=df,hue='Playoffs')
plt.subplot(122)
sns.lineplot(x='Year',y='BPG',data=df,hue='Playoffs')
plt.tight_layout()
plt.show()
plt.subplot(121)
sns.lineplot(x='Year',y='RPG',data=df,hue='Playoffs')
plt.subplot(122)
sns.lineplot(x='Year',y='TOV',data=df,hue='Playoffs')
plt.tight_layout()
plt.show()
plt.subplot(121)
sns.lineplot(x='Year',y='DEFF',data=df,hue='Playoffs')
plt.subplot(122)
sns.lineplot(x='Year',y='POSSt',data=df,hue='Playoffs')
plt.tight_layout()
plt.show()

#Boxplots δίνονται δύο παραδείγματα καθώς στα υπόλοιπα αλλάζουμε απλώς τις μεταβλητές
boxplot=df.boxplot(column=['PPG'],by=['Playoffs'])
boxplot=df.boxplot(column=['APG'],by=['Playoffs'])

#Ανάλυση με βάση τον διαχωρισμό των ομάδων σε Περιφέρειες
#Pie Charts
df.groupby(['Conference1']).sum().plot(kind='pie', y='Top Seed',
autopct='%1.0f%%')
```

```
df.groupby(['Conference1']).sum().plot(kind='pie', y='Champions',
autopct='%1.0f%%')
df.groupby(['Champions']).sum().plot(kind='pie', y='Top Seed',
autopct='%1.0f%%')

#Χαρακτηριστικά των μεταβλητών που επιλέχθηκαν γι' αυτή την ανάλυση
df3=df.groupby('Conference1')
df3['PPG','APG','SPG','BPG','RPG','TOV','EFF','DEFF','POSSt'].describe()

#Time Series Plots
plt.subplot(121)
sns.lineplot(x='Year',y='PPG',data=df,hue='Conference1')
plt.subplot(122)
sns.lineplot(x='Year',y='APG',data=df,hue='Conference1')
plt.tight_layout()
plt.show()
plt.subplot(121)
sns.lineplot(x='Year',y='SPG',data=df,hue='Conference1')
plt.subplot(122)
sns.lineplot(x='Year',y='BPG',data=df,hue='Conference1')
plt.tight_layout()
plt.show()
plt.subplot(121)
sns.lineplot(x='Year',y='RPG',data=df,hue='Conference1')
plt.subplot(122)
sns.lineplot(x='Year',y='TOV',data=df,hue='Conference1')
plt.tight_layout()
plt.show()
plt.subplot(121)
sns.lineplot(x='Year',y='DEFF',data=df,hue='Conference1')
plt.subplot(122)
sns.lineplot(x='Year',y='POSSt',data=df,hue='Conference1')
plt.tight_layout()
plt.show()

#Violin Plots δίνονται δύο παραδείγματα καθώς στα υπόλοιπα αλλάζουμε
απλώς τις μεταβλητές
fig=px.violin(df,y='PPG',x='Conference1',color='Conference1', box=True)
fig.show()
fig=px.violin(df,y='APG',x='Conference1',color='Conference1', box=True)
fig.show()

#Histograms δίνονται δύο παραδείγματα καθώς στα υπόλοιπα αλλάζουμε
απλώς τις μεταβλητές
plt.subplot(121)
A=df.loc[df['Conference1']=='WEST','PPG']
plt.hist(A,alpha=0.5,label='WEST')
B=df.loc[df['Conference1']=='EAST','PPG']
plt.hist(B,alpha=0.5,label='EAST')
plt.title('Points Distribution by Team')
```

```
plt.xlabel('PPG')
plt.ylabel('Frequency')
plt.subplot(122)
A=df.loc[df['Conference1']=='WEST','APG']
plt.hist(A,alpha=0.5,label='WEST')
B=df.loc[df['Conference1']=='EAST','APG']
plt.hist(B,alpha=0.5,label='EAST')
plt.title('Assist Distribution by Team')
plt.xlabel('APG')
plt.ylabel('Frequency')
plt.legend(title='Conference')
plt.tight_layout()
plt.figure(figsize=(10000,6000))
plt.show()
plt.subplot(121)

#Ανάλυση με βάση τον διαχωρισμό των ομάδων σε Περιφέρειες
#Pie Charts δίνονται δύο παραδείγματα καθώς στα υπόλοιπα αλλάζουμε απλώς τις μεταβλητές
fig = px.pie(df, values='PPG', names='Team')
fig.show()
fig = px.pie(df, values='APG', names='Team')
fig.show()

#ΚΕΦΑΛΑΙΟ 4Ο

#Ελεγχοι κανονικότητας (Shapiro – Wilk) δίνονται δύο παραδείγματα καθώς στα υπόλοιπα αλλάζουμε απλώς τις μεταβλητές
print(shapiro(df_group1['PPG']))
print(shapiro(df_group1['PPGA']))

#Ελεγχοι κανονικότητας (Shapiro – Wilk) με βάση την πρόκριση στα Playoffs
#ομάδες που κατάφεραν να προκριθούν δίνονται δύο παραδείγματα καθώς στα υπόλοιπα αλλάζουμε απλώς τις μεταβλητές
df_playoffs=df.query('Playoffs == 1')
print(shapiro(df_playoffs['PPG']))
print(shapiro(df_playoffs['APG']))

#ομάδες που δεν κατάφεραν να προκριθούν δίνονται δύο παραδείγματα καθώς στα υπόλοιπα αλλάζουμε απλώς τις μεταβλητές
df_nonplayoffs=df.query('Playoffs == 0')
print(shapiro(df_nonplayoffs['PPG']))
print(shapiro(df_nonplayoffs['APG']))

#Ελεγχοι κανονικότητας (Shapiro – Wilk) βάσει του διαχωρισμού των ομάδων σε Περιφέρειες
#ομάδες που ανήκουν στην WEST περιφέρεια δίνονται δύο παραδείγματα καθώς στα υπόλοιπα αλλάζουμε απλώς τις μεταβλητές
df_west=df.query('Conference1 == "WEST"')
print(shapiro(df_west['PPG']))
```

```
print(shapiro(df_west['APG']))

#ομάδες που ανήκουν στην EAST περιφέρεια δίνονται δύο παραδείγματα
καθώς στα υπόλοιπα αλλάζουμε απλώς τις μεταβλητές
df_east=df.query('Conference1 == "EAST"')
print(shapiro(df_east['PPG']))
print(shapiro(df_east['APG']))

#Heatmaps
data=df.drop(['Year','Team','Conference1','Conference','GP','MPG','STATrs','Playoff
s','Champions'],axis=1)
cor = data.corr()
sns.heatmap(cor, annot=True, cmap=plt.cm.Reds)
sns.set(rc={'figure.figsize':(20,20)})
plt.show()

#Heatmaps με βάση την πρόκριση στα Playoffs
data_playoffs=df_playoffs[['PPG','APG','SPG','BPG','RPG','TOV','DEFF','POSSt']]
cor = data_playoffs.corr()
sns.heatmap(cor, annot=True, cmap=plt.cm.Reds)
sns.set(rc={'figure.figsize':(10,10)})
plt.show()
data_nonplayoffs=df_nonplayoffs[['PPG','APG','SPG','BPG','RPG','TOV','DEFF','P
OSSt']]
cor = data_nonplayoffs.corr()
sns.heatmap(cor, annot=True, cmap=plt.cm.Reds)
sns.set(rc={'figure.figsize':(10,10)})
plt.show()

#Heatmaps βάσει του διαχωρισμού των ομάδων σε Περιφέρειες
data_west=df_west[['PPG','APG','SPG','BPG','RPG','TOV','DEFF','POSSt']]
cor = data_west.corr()
sns.heatmap(cor, annot=True, cmap=plt.cm.Reds)
sns.set(rc={'figure.figsize':(10,10)})
plt.show()
data_east=df_east[['PPG','APG','SPG','BPG','RPG','TOV','DEFF','POSSt']]
cor = data_east.corr()
sns.heatmap(cor, annot=True, cmap=plt.cm.Reds)
sns.set(rc={'figure.figsize':(10,10)})
plt.show()

#Έλεγχος για την ισότητα μέσων τιμών (t-tests) δίνονται δύο παραδείγματα
καθώς στα υπόλοιπα αλλάζουμε απλώς τις μεταβλητές
#με βάση την πρόκριση στα Playoffs
t_stat, p_val = stats.ttest_ind(df_playoffs['PPG'], df_nonplayoffs['PPG'])
print("t-statistic = " + str(t_stat))
print("p-value = " + str(p_val))
t_stat, p_val = stats.ttest_ind(df_playoffs['APG'], df_nonplayoffs['APG'])
print("t-statistic = " + str(t_stat))
print("p-value = " + str(p_val))
```

#βάσει του διαχωρισμού των ομάδων σε Περιφέρειες δίνονται δύο παραδείγματα καθώς στα υπόλοιπα αλλάζουμε απλώς τις μεταβλητές

```
t_stat, p_val = stats.ttest_ind(df_west['PPG'], df_east['PPG'])
print("t-statistic = " + str(t_stat))
print("p-value = " + str(p_val))
t_stat, p_val = stats.ttest_ind(df_west['SPG'], df_east['SPG'])
print("t-statistic = " + str(t_stat))
print("p-value = " + str(p_val))
```

#Ελεγχοι για ζευγαρωτές παρατηρήσεις δίνονται δύο παραδείγματα καθώς στα υπόλοιπα αλλάζουμε απλώς τις μεταβλητές

```
t_stat, p_val = stats.ttest_rel(df['PPGp'],df['PPG'])
print("t-statistic = " + str(t_stat))
print("p-value = " + str(p_val))
t_stat, p_val = stats.ttest_rel(df['APGp'],df['APG'])
print("t-statistic = " + str(t_stat))
print("p-value = " + str(p_val))
```

#ΚΕΦΑΛΑΙΟ 5^{Ο}

#Στο παρόν κεφάλαιο χρησιμοποιήθηκαν και ορισμένες βιβλιοθήκες που διατίθενται μόνο στην R· επομένως ορισμένες εντολές είναι από αυτό το στατιστικό πακέτο. Ακόμα, στο παράρτημα θα δοθούν τα παραδείγματα για την κανονική περίοδο με όλες τις μεταβλητές ως προς τις δύο κατηγορικές μεταβλητές Playoffs και Champions. Για τις υπόλοιπες αναλύσεις που παρουσιάστηκαν στην εργασία έχουν χρησιμοποιηθεί οι ίδιες διαδικασίες.

#Διαδικασία step βάσει της κατηγορικής μεταβλητής Playoffs (στην R)

```
library("readxl")
b=read_excel("regularseason.xlsx")
attach(b)
Playoffs<-as.factor(Playoffs)
model1<-glm(Playoffs~1,family=binomial(link=logit))
stepMod1 <- step(model1, scope = list(lower = model1, upper = model2),direction
= "both")
model10<-glm(Playoffs~DRt+ORt,family=binomial)
anova(model10,test="Chisq")
```

#τελικό μοντέλο (Python)

```
model = smf.glm(formula = "Playoffs ~ ORt+DRt",
                data = df,
                family = sm.families.Binomial())
result = model.fit()
print(result.summary())
```

#Δείκτης VIF (Python)

```
y,X = dmatrices('Playoffs ~ ORt+DRt', data=df, return_type='dataframe')
vif_df = pd.DataFrame()
vif_df['variable'] = X.columns
vif_df['VIF'] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
```

```
print(vif_df)

#Hosmer-Lemeshow (R)
y<-Playoffs
install.packages("GOF")
library(ResourceSelection)
hoslem.test<-hoslem.test(model10$y,fitted(model10))
hoslem.test

#Διαδικασία step βάσει της κατηγορικής μεταβλητής Champions (στην R)
Champions<-as.factor(Champions)
model1<-glm(Champions~1,family=binomial(link=logit))
stepMod1 <- step(model1, scope = list(lower = model1, upper = model2),direction
= "both")
model10<-glm(Champions~XFG+DRt+FGM,family=binomial)
anova(model10,test="Chisq")

#τελικό μοντέλο (Python)
model = smf.glm(formula = "Champions ~ df['FG%']+DRt+FGM",
            data = df,
            family = sm.families.Binomial())
result = model.fit()
print(result.summary())

#Δείκτης VIF (Python)
y,X1 = dmatrices("Champions ~ df['FG%']+DRt+FGM", data=df,
return_type='dataframe')
vif_df = pd.DataFrame()
vif_df['variable'] = X1.columns
vif_df['VIF'] = [variance_inflation_factor(X1.values, i) for i in range(X1.shape[1])]
print(vif_df)

#Hosmer-Lemeshow (R)
y<-Playoffs
hoslem.test<-hoslem.test(model10$y,fitted(model10))
hoslem.test

#ΚΕΦΑΛΑΙΟ 6Ο

#Ανάλυση χρονοσειράς της μεταβλητής PPG

#Γραφήματα χρονοσειράς και έλεγχος στασιμότητας
par(mfrow=c(1,3))
plot(PPG,type="l")
acf(PPG)
pacf(PPG)
ppg=diff(PPG,d=4)
T=length(PPG);T
plot(ppg,type="l")
acf(ppg)
```

```
pacf(ppg)
library(tseries)
adf.test(ppg)

#Υπολογισμός AICc
n=length(ppg)
k=3;AICc1=AIC(arima(ppg,order=c(0,0,1)))+2*k*(k+1)/(n-k-1)
k=4;AICc2=AIC(arima(ppg,order=c(0,0,2)))+2*k*(k+1)/(n-k-1)
k=4;AICc3=AIC(arima(ppg,order=c(1,0,1)))+2*k*(k+1)/(n-k-1)
k=5;AICc4=AIC(arima(ppg,order=c(1,0,2)))+2*k*(k+1)/(n-k-1)

#Υπολογσιμός AIC και BIC
AIC(arima(ppg,order=c(0,0,1)));BIC(arima(ppg,order=c(0,0,1)));AICc1
AIC(arima(ppg,order=c(0,0,2)));BIC(arima(ppg,order=c(0,0,2)));AICc2
AIC(arima(ppg,order=c(1,0,1)));BIC(arima(ppg,order=c(1,0,1)));AICc3
AIC(arima(ppg,order=c(1,0,2)));BIC(arima(ppg,order=c(1,0,2)));AICc4

#Ελεγχοι για την επιλεγμένη χρονοσειρά
library(tseries)
arima(ppg,order=c(0,0,2),method="ML")$coef
res1=residuals(arima(ppg,order=c(0,0,2),method="ML"))
library(nortest)
res1=residuals(arima(ppg,order=c(0,0,2)))
shapiro.test(res1);ad.test(res1)
Box.test(res1,lag=1,type = "Ljung-Box")
par(mfrow=c(1,4));hist(res1,30,prob=TRUE);xd2=seq(0,50,0.5);
lines(xd2,dnorm(xd2,mean(res1),sd(res1)),col="blue");qqnorm(res1);
qqline(res1,col="red");plot(res1);acf(res1)

#Γραφική παράσταση
m1=arima(ppg,order=c(0,0,2))
library(forecast)
forecast(m1,2)
autoplot(forecast(m1,2))

#Ανάλυση χρονοσειράς της μεταβλητής APG

#Γραφήματα χρονοσειράς και έλεγχος στασιμότητας
par(mfrow=c(1,3))
plot(APG,type="l")
acf(APG)
pacf(APG)
apg=diff(APG,d=5)
T=length(APG);T
plot(apg,type="l")
acf(apg)
pacf(apg)
library(tseries)
adf.test(apg)
```

```
#Υπολογισμός AICc
n=length(apg)
k=2;AICc1=AIC(arima(apg,order=c(0,0,0)))+2*k*(k+1)/(n-k-1)
k=3;AICc2=AIC(arima(apg,order=c(0,0,1)))+2*k*(k+1)/(n-k-1)
k=4;AICc3=AIC(arima(apg,order=c(0,0,2)))+2*k*(k+1)/(n-k-1)
k=4;AICc4=AIC(arima(apg,order=c(1,0,1)))+2*k*(k+1)/(n-k-1)

#Υπολογισμός AIC και BIC
AIC(arima(apg,order=c(0,0,0)));BIC(arima(apg,order=c(0,0,0)));AICc1
AIC(arima(apg,order=c(0,0,1)));BIC(arima(apg,order=c(0,0,1)));AICc2
AIC(arima(apg,order=c(0,0,2)));BIC(arima(apg,order=c(0,0,2)));AICc3
AIC(arima(apg,order=c(1,0,1)));BIC(arima(apg,order=c(1,0,1)));AICc4

#Ελεγχοι για την επιλεγμένη χρονοσειρά
library(tseries)
arima(apg,order=c(1,0,1),method="ML")$coef
res1=residuals(arima(apg,order=c(1,0,1),method="ML"))
library(nortest)
res1=residuals(arima(apg,order=c(1,0,1)))
shapiro.test(res1);ad.test(res1)
Box.test(res1,lag=1,type = "Ljung-Box")
par(mfrow=c(1,4));hist(res1,30,prob=TRUE);xd2=seq(0,50,0.5);
lines(xd2,dnorm(xd2,mean(res1),sd(res1)),col="blue");qqnorm(res1);
qqline(res1,col="red");plot(res1);acf(res1)

#Γραφική παράσταση
m2=arima(apg,order=c(1,0,1))
library(forecast)
forecast(m2,2)
autoplot(forecast(m2,2))
auto.arima(PPG,stepwise=FALSE)

#Ανάλυση χρονοσειράς της μεταβλητής SPG

#Γραφήματα χρονοσειράς και έλεγχος στασιμότητας
par(mfrow=c(1,3))
plot(SPG,type="l")
acf(SPG)
pacf(SPG)
spg=diff(SPG,d=2)
T=length(SPG);T
plot(spg,type="l")
acf(spg)
pacf(spg)
library(tseries)
adf.test(spg)

#Υπολογισμός AICc
n=length(spg)
k=2;AICc1=AIC(arima(spg,order=c(0,0,0)))+2*k*(k+1)/(n-k-1)
```

```
k=3;AICc2=AIC(arima(spg,order=c(0,0,1)))+2*k*(k+1)/(n-k-1)
k=4;AICc3=AIC(arima(spg,order=c(0,0,2)))+2*k*(k+1)/(n-k-1)
k=5;AICc4=AIC(arima(spg,order=c(0,0,3)))+2*k*(k+1)/(n-k-1)

#Υπολογισμός AIC και BIC
AIC(arima(spg,order=c(0,0,0)));BIC(arima(spg,order=c(0,0,0)));AICc1
AIC(arima(spg,order=c(0,0,1)));BIC(arima(spg,order=c(0,0,1)));AICc2
AIC(arima(spg,order=c(0,0,2)));BIC(arima(spg,order=c(0,0,2)));AICc3
AIC(arima(spg,order=c(0,0,3)));BIC(arima(spg,order=c(0,0,3)));AICc4

#Ελεγχοι για την επιλεγμένη χρονοσειρά
library(tseries)
arima(spg,order=c(0,0,1),method="ML")$coef
res1=residuals(arima(spg,order=c(0,0,1),method="ML"))
library(nortest)
res1=residuals(arima(spg,order=c(0,0,1)))
shapiro.test(res1);ad.test(res1)
Box.test(res1,lag=1,type = "Ljung-Box")
par(mfrow=c(1,4));hist(res1,30,prob=TRUE);xd2=seq(0,50,0.5);
lines(xd2,dnorm(xd2,mean(res1),sd(res1)),col="blue");qqnorm(res1);
qqline(res1,col="red");plot(res1);acf(res1)4

#Γραφική παράσταση
m3=arima(spg,order=c(0,0,1))
library(forecast)
forecast(m3,2)
autoplot(forecast(m3,2))

#Ανάλυση χρονοσειράς της μεταβλητής BPG

#Γραφήματα χρονοσειράς και έλεγχος στασιμότητας
par(mfrow=c(1,3))
plot(BPG,type="l")
acf(BPG)
pacf(BPG)
par(mfrow=c(1,4))
plot(BPG,type="l")
cpg=diff(BPG,lag=1)
npg=diff(BPG,lag=2)
kpg=diff(BPG,lag=3)
plot(cpg,type="l")
plot(npg,type="l")
plot(kpg,type="l")
par(mfrow=c(1,3))
bpg=diff(kpg,d=3)
T=length(BPG);T
plot(bpg,type="l")
acf(bpg)
pacf(bpg)
adf.test(bpg)
```

```
#Υπολογισμός AIC και BIC
m1=arima(BPG,order=c(0,3,0),seasonal=list(order=c(0,1,0),period=3),method="M
L")
m2=arima(BPG,order=c(0,3,1),seasonal=list(order=c(0,1,0),period=3),method="M
L")
m3=arima(BPG,order=c(0,3,2),seasonal=list(order=c(0,1,0),period=3),method="M
L")
m4=arima(BPG,order=c(1,3,1),seasonal=list(order=c(0,1,0),period=3),method="M
L")
AIC(m1);AIC(m2);AIC(m3);AIC(m4);
BIC(m1);BIC(m2);BIC(m3);BIC(m4);

#Ελεγχοι για την επιλεγμένη χρονοσειρά
res=residuals(m4)
par(mfrow=c(1,4));
hist(res,prob=TRUE);x=seq(-
1.5,2,0.01);lines(x,dnorm(x,mean(res),sd(res)),col="blue");
qqnorm(res);qqline(res,col="red");acf(ts(res));plot(res,type="l")
shapiro.test(res);Box.test(res,lag=3,type = "Ljung-Box");ad.test(res)

#Γραφική παράσταση
forecast(m4,2);
autoplot(forecast(m4,2))

#Ανάλυση χρονοσειράς της μεταβλητής RPG

#Γραφήματα χρονοσειράς και έλεγχος στασιμότητας
par(mfrow=c(1,3))
plot(RPG,type="l")
acf(RPG)
pacf(RPG)
rpg=diff(RPG,d=3)
library(tseries)
adf.test(rpg)
T=length(RPG);T
plot(rpg,type="l")
acf(rpg)
pacf(rpg)
#Υπολογισμός AICc
n=length(rpg)
k=2;AICc1=AIC(arima(rpg,order=c(0,0,0)))+2*k*(k+1)/(n-k-1)
k=3;AICc2=AIC(arima(rpg,order=c(0,0,1)))+2*k*(k+1)/(n-k-1)
k=4;AICc3=AIC(arima(rpg,order=c(0,0,2)))+2*k*(k+1)/(n-k-1)
k=5;AICc4=AIC(arima(rpg,order=c(0,0,3)))+2*k*(k+1)/(n-k-1)

#Υπολογισμός AIC και BIC
AIC(arima(rpg,order=c(0,0,0)));BIC(arima(rpg,order=c(0,0,0)));AICc1
AIC(arima(rpg,order=c(0,0,1)));BIC(arima(rpg,order=c(0,0,1)));AICc2
AIC(arima(rpg,order=c(0,0,2)));BIC(arima(rpg,order=c(0,0,2)));AICc3
```

```
AIC(arima(rpg,order=c(0,0,3)));BIC(arima(rpg,order=c(0,0,3)));AICc4

#Ελεγχοι για την επιλεγμένη χρονοσειρά
library(tseries)
arima(rpg,order=c(0,0,2),method="ML")$coef
res1=residuals(arima(rpg,order=c(0,0,2),method="ML"))
library(nortest)
res1=residuals(arima(rpg,order=c(0,0,2)))
shapiro.test(res1);ad.test(res1)
Box.test(res1,lag=1,type = "Ljung-Box")
par(mfrow=c(1,4));hist(res1,30,prob=TRUE);xd2=seq(0,50,0.5);
lines(xd2,dnorm(xd2,mean(res1),sd(res1)),col="blue");qqnorm(res1);
qqline(res1,col="red");plot(res1);acf(res1)

#Γραφική παράσταση
m1=arima(rpg,order=c(0,0,2))
library(forecast)
forecast(m1,2)
autoplot(forecast(m1,2))

#ΚΕΦΑΛΑΙΟ 7Ο

#Σε αυτό το κεφάλαιο θα δοθούν αναλυτικά οι αλγόριθμοι που
χρησιμοποιήθηκαν για την μέθοδο Boruta και η μείωση επιλογών της μεθόδου
KMEANS, διότι οι τρόποι υλοποίησης των διαδικασιών, αφού βρεθούν οι
μεταβλητές που θα χρησιμοποιηθούν, είναι οι ίδιοι.

#Μέθοδος Boruta
X=
df[['PPG','PPGA','FGM','FGA','MisFG','FG%','FGAo','3PM','3PA','3P%','FTM','FT
A','MisFt','FT%','FTAo','ORB','DRB','RPG','ORBo','APG','SPG','BPG','TOV','TOVo
','PF','POSSt','ORt','POSSo','DRt']].values
Y= df['Playoffs'].values
rf = RandomForestClassifier(n_estimators=200,n_jobs=4, class_weight='balanced',
max_depth=6)
feat_selector = BorutaPy(rf, n_estimators='auto', verbose=2)
feat_selector.fit(X, Y)
df_selected=df[['PPG','ORt','POSSt','TOV','SPG','RPG','DRB','3P%','FTAo','3PM','P
PGA','FGM','FGA','MisFG','FG%','DRt']]

#Κανονικοποίηση δεδομένων και δημιουργία δεδομένων εκπαίδευσης και μη
scaler = StandardScaler()
input_data = scaler.fit_transform(df_selected)
output_data = df['Playoffs']
X_train, X_test, y_train, y_test = train_test_split(input_data,output_data,
test_size=0.30)

#Κατηγοριοποίηση K-κοντινότερων γειτόνων – υπολογισμός μέτρων
αξιολόγησης
model1 = KNeighborsClassifier()
```

```
model1.fit(X_train, y_train)
y_pred1 = model1.predict(X_test)
print(classification_report(y_test, y_pred1, digits=4))
print('AUC score:', roc_auc_score(y_test, y_pred1))
scorers = {'accuracy': make_scorer(accuracy_score),
 'precision': make_scorer(precision_score, pos_label=1),
 'recall': make_scorer(recall_score, pos_label=1),
 'f1': make_scorer(f1_score, pos_label=1),
 'auc': make_scorer(roc_auc_score)}
cm = confusion_matrix(y_test, y_pred1)
print("Confusion matrix:")
print(cm)
print("Total samples:", np.sum(cm))
print("True positives:", cm[1, 1])
print("False positives:", cm[0, 1])
print("True negatives:", cm[0, 0])
print("False negatives:", cm[1, 0])

#Κατηγοριοποίηση τυχαίου δάσους
model4 = RandomForestClassifier(n_estimators=300, max_depth=10,
random_state=40)
model4.fit(X_train, y_train)
y_pred1 = model4.predict(X_test)
print(classification_report(y_test, y_pred1, digits=4))
print('AUC score:', roc_auc_score(y_test, y_pred1))
scorers = {'accuracy': make_scorer(accuracy_score),
 'precision': make_scorer(precision_score, pos_label=1),
 'recall': make_scorer(recall_score, pos_label=1),
 'f1': make_scorer(f1_score, pos_label=1),
 'auc': make_scorer(roc_auc_score)}
cm = confusion_matrix(y_test, y_pred1)
print("Confusion matrix:")
print(cm)
print("Total samples:", np.sum(cm))
print("True positives:", cm[1, 1])
print("False positives:", cm[0, 1])
print("True negatives:", cm[0, 0])
print("False negatives:", cm[1, 0])

#Κατηγοριοποίηση SVM
model3 = SVC(kernel='linear', C=1, random_state=42)
model3.fit(X_train, y_train)
y_pred = model3.predict(X_test)
print(classification_report(y_test, y_pred, digits=4))
print('AUC score:', roc_auc_score(y_test, y_pred))
scorers = {'accuracy': make_scorer(accuracy_score),
 'precision': make_scorer(precision_score, pos_label=1),
 'recall': make_scorer(recall_score, pos_label=1),
 'f1': make_scorer(f1_score, pos_label=1),
 'auc': make_scorer(roc_auc_score)}
```

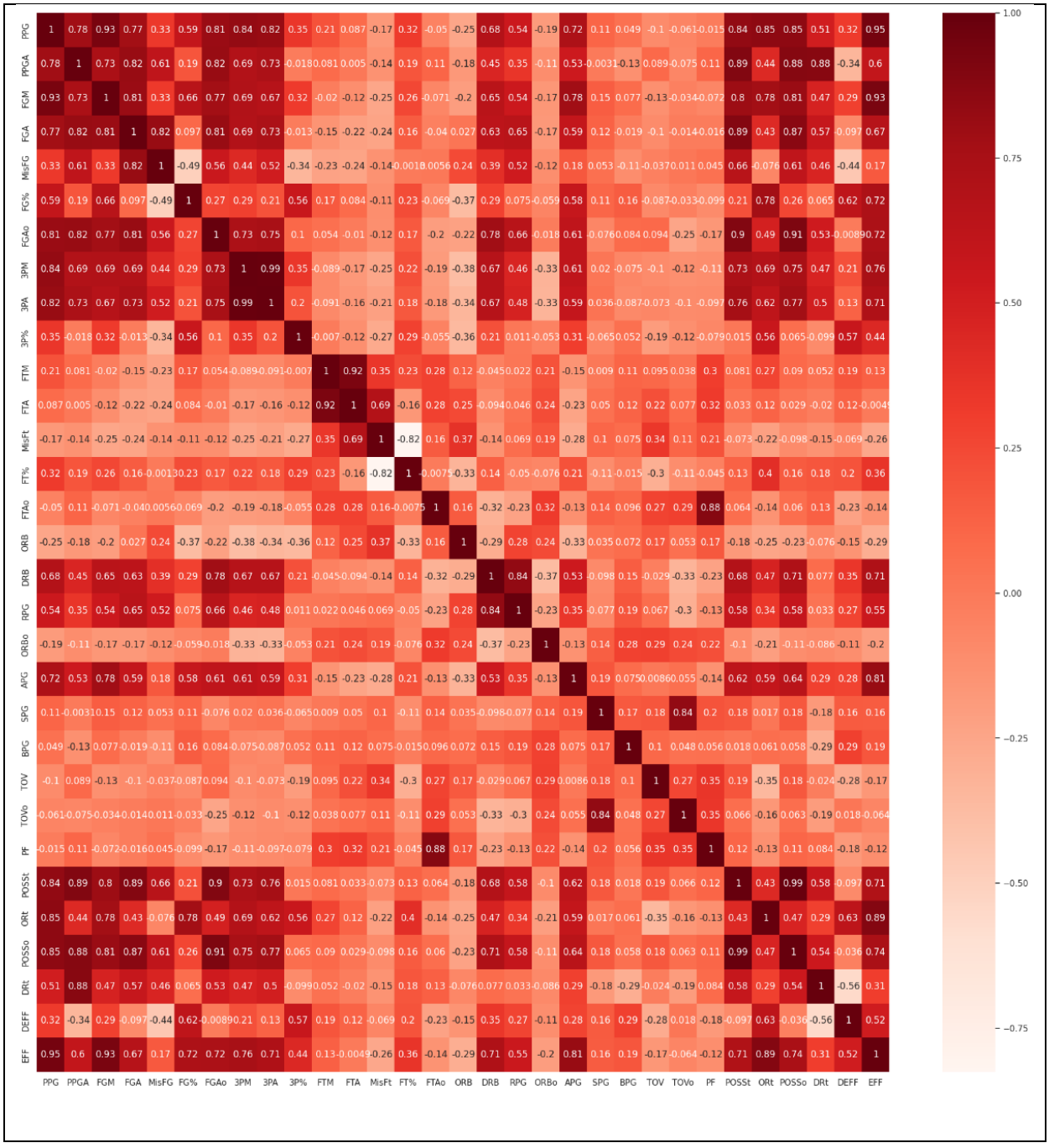
```
cm = confusion_matrix(y_test, y_pred)
print("Confusion matrix:")
print(cm)
print("Total samples:", np.sum(cm))
print("True positives:", cm[1, 1])
print("False positives:", cm[0, 1])
print("True negatives:", cm[0, 0])
print("False negatives:", cm[1, 0])

#Ορισμός συνόλου δεδομένων και εκτέλεση της μεθόδου αγκώνα για την συσταδοποίηση
df_selected=df[['PPG','ORt','POSSt','TOV','SPG','RPG','DRB','3P%','FTAo','3PM','PPGA','FGM','FGA','MisFG','FG%','DRt']]
data = df_selected.values
scaler = StandardScaler()
X = StandardScaler().fit_transform(data)
wcss = {}
for I in range(1, 11):
   kmeans = Kmeans(n_clusters = I, init = 'k-means++', random_state = 42)
   kmeans.fit(X)
   wcss[i] = kmeans.inertia_

plt.plot(wcss.keys(), wcss.values(), 'gs-')
plt.xlabel("Values of 'k'")
plt.ylabel('WCSS')
plt.show()

#KMEANS συσταδοποίηση
kmeans=Kmeans(n_clusters=2)
kmeans=kmeans.fit(X)
labels=kmeans.labels_
kmeans.cluster_centers_

#Εύρεση σωστών και λανθασμένων τοποθετήσεων
df_selected['playoffs']=df['Playoffs']
df2 = df_selected.pivot_table(index = ['playoffs', 'clusters'], aggfunc ='size')
print(df2)
my_np1=np.array([[0,0,1,1],[0,1,0,1]])
my_pd=pd.DataFrame(data=[my_np1[0],my_np1[1]]).T
my_pd.columns=['playoffs','clusters']
my_pd['teams']=[115,96,133,106]
print(my_pd)
#Boxplot
sns.set(style="darkgrid")
plt.figure(figsize=(8, 8))
sns.barplot(x="playoffs", y="teams", hue="clusters", data=my_pd, ci=None);

#Υπολογισμός μέτρων αξιολόγησης
score = silhouette_score(X,labels, metric='euclidean')
print('Silhouetter Score: %.3f' % score)
```

```
calinski_harabasz_score(X, labels)

#PCA
pca=PCA(n_components=2)
df_component=pd.DataFrame(data=pca.fit_transform(X),columns=['Comp1','Comp
2'])
df_component.head()
centers=pca.transform(kmeans.cluster_centers_)

#Cluster plot
x_axis=df_component['Comp1']
y_axis=df_component['Comp2']
plt.figure(figsize=(10,8))
sns.scatterplot(x=x_axis,y=y_axis,hue=df_component['clusters'],palette=['c','black']
)
plt.scatter(centers[:,0],centers[:,1],marker='x',s=100,c='red')
plt.title('Clusters by PCA Components')
plt.show

#BIRCH συσταδοποίηση
df_selected=df[['PPG','ORt','POSSt','TOV','SPG','RPG','DRB','3P%','FTAo','3PM','P
PGA','FGM','FGA','MisFG','FG%','DRt']]
data = df_selected
scaler = StandardScaler()
X1 = StandardScaler().fit_transform(data)
model_clus=Birch(n_clusters=2)
model_clus=model_clus.fit(X1)
labels1=model_clus.labels_

#Εύρεση σωστών και λανθασμένων τοποθετήσεων
df_selected['playoffs']=df['Playoffs']
df3 = df_selected.pivot_table(index = ['playoffs', 'clusters1'], aggfunc ='size')
print(df3)
my_np1=np.array([[0,0,1,1],[0,1,0,1]])
my_pd1=pd.DataFrame(data=[my_np1[0],my_np1[1]]).T
my_pd1.columns=['playoffs','clusters1']
my_pd1['teams']=[158,53,160,79]
print(my_pd1)

#Box plot
sns.set(style="darkgrid")

plt.figure(figsize=(8, 8))
sns.barplot(x="playoffs", y="teams", hue="clusters1", data=my_pd1, ci=None);

#Υπολογισμός μέτρων αξιολόγησης
score = silhouette_score(X1,labels1, metric='euclidean')
print('Silhouetter Score: %.3f' % score)
calinski_harabasz_score(X1, labels1)
```

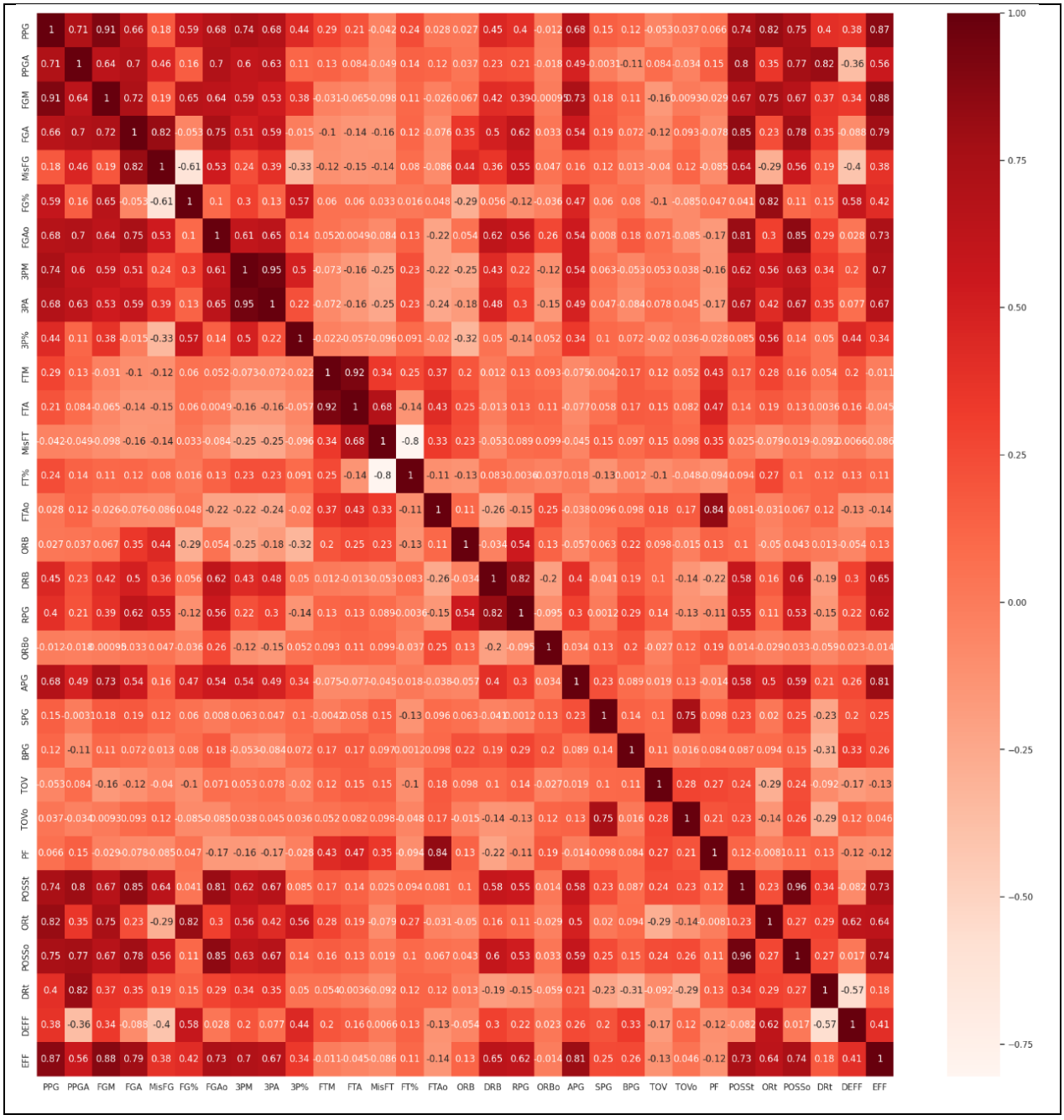
```
#PCA
pca=PCA(n_components=2)
df_component=pd.DataFrame(data=pca.fit_transform(X1),columns=['Comp1','Com
p2'])
df_component.head()

#Cluster plot
x_axis=df_component['Comp1']
y_axis=df_component['Comp2']
plt.figure(figsize=(10,8))
sns.scatterplot(x=x_axis,y=y_axis,hue=df_component['clusters1'],palette=['c','black'
])
plt.scatter(centers[:,0],centers[:,1],marker='x',s=100,c='red')
plt.title('Clusters by PCA Components')
plt.show

#Μέθοδος KMEANS
X=
df[['PPG','PPGA','FGM','FGA','MisFG','FG%','FGAo','3PM','3PA','3P%','FTM','FT
A','MisFt','FT%','FTAo','ORB','DRB','RPG','ORBo','APG','SPG','BPG','TOV','TOVo
','PF','POSSt','ORt','POSSo','DRt']]
Y= df['Playoffs']
from sklearn.feature_selection import SelectKBest, chi2
selector=SelectKBest(chi2,k=8)
selector.fit(X,Y)
X.columns[selector.get_support()]
new_df=df['PPG','PPGA','MisFG','3PM','3PA','FTAo','ORt','DRt']

#Εισαγωγή των δεδομένων από το νέο αρχείο δεδομένων που δημιουργήθηκε
για τον έλεγχο διαφορών ανάμεσα στις επιδόσεις των ομάδων των πρώτων και
των τελευταίων σεζόν της διοργάνωσης
df = pd.read_excel('Book3.xlsx')
df=pd.DataFrame(df)

#Paired T-test δίνονται δύο παραδείγματα καθώς στα υπόλοιπα αλλάζουμε
απλώς τις μεταβλητές
stats.ttest_rel(df['PPG'],df['PPG2'])
stats.ttest_rel(df[‘ORt’], df[‘ORt2’])
```

Π2. Heatmaps

Heatmap του συντελεστή Pearson για την κανονική περίοδο



Heatmap του συντελεστή Pearson για τα playoffs



Π3. Συσχετίσεις στην λογιστική παλινδρόμηση

Προβλήματα συσχετίσεων στην λογιστική παλινδρόμηση

```
    ORB + DRB + RPG + ORBo + APG + SPG + BPG + TOV + TOVo + PF +
    POSSt + ORt + POSSo + DRt + DEFF + EFF, family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.49181  -0.15344   0.00729   0.20958   2.25313

Coefficients: (6 not defined because of singularities)
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 29.877965 217.989802   0.137    0.891
PPG          1.037975   4.232876   0.245    0.806
PPGA         1.787337   3.596127   0.497    0.619
FGM         -6.437801   5.466620  -1.178    0.239
FGA          2.423066   3.668002   0.661    0.509
MisFG              NA         NA      NA       NA
XFG        460.992694 354.233102   1.301    0.193
FGAo        -2.583899   3.849268  -0.671    0.502
X3PM         0.344589   3.370475   0.102    0.919
X3PA        -0.319698   0.619766  -0.516    0.606
X3P        -44.404254  44.212085  -1.004    0.315
FTM         -2.210929   3.756648  -0.589    0.556
FTA          1.309533   2.439729   0.537    0.591
MisFt              NA         NA      NA       NA
XFT         40.636692  58.616117   0.693    0.488
FTAo        -1.000737   1.704807  -0.587    0.557
ORB          2.181408   4.919146   0.443    0.657
DRB          2.865527   3.478215   0.824    0.410
RPG         -2.525953   3.531172  -0.715    0.474
ORBo         2.893425   3.852339   0.751    0.453
APG         -0.009721   0.157351  -0.062    0.951
SPG          0.103964   0.514449   0.202    0.840
BPG         -0.263543   0.335538  -0.785    0.432
TOV         -0.100491   3.560233  -0.028    0.977
TOVo        -2.567460   3.892135  -0.660    0.509
PF          -0.085742   0.336573  -0.255    0.799
POSSt              NA         NA      NA       NA
ORt          0.601964   3.198384   0.188    0.851
POSSo              NA         NA      NA       NA
DRt         -2.890493   3.477614  -0.831    0.406
DEFF               NA         NA      NA       NA
EFF                NA         NA      NA       NA
```