# 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**


# ΣΤΑΤΙΣΤΙΚΑ ΜΟΝΤΕΛΑ ΓΙΑ ΤΗΝ ΑΠΟΔΟΣΗ ΜΙΑΣ ΟΜΑΔΑΣ ΜΠΑΣΚΕΤ: ΠΟΙΑ ΣΤΑΤΙΣΤΙΚΑ ΣΤΟΙΧΕΙΑ ΕΙΝΑΙ ΚΑΘΟΡΙΣΤΙΚΑ ΓΙΑ ΤΗΝ ΑΠΟΔΟΣΗ ΤΗΣ ΟΜΑΔΑΣ, ΣΕ ΕΤΗΣΙΑ ΒΑΣΗ


Διονύσιος Γ. Καλλιακμάνης


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

Μάιος 2020

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 Πολίτης Κωνσταντίνος (Επιβλέπων)

- Επίκουρος Καθηγητής Ευαγγελάρας Χαράλαμπος

- Επίκουρος Καθηγητής Μπούτσικας Μιχαήλ

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**

# STATISTICAL MODELS FOR THE PERFORMANCE OF A BASKETBALL TEAM: WHICH STATISTICS IN THE BOXSCORE DETERMINE THE TEAM'S SEASON-LONG SUCCESS


By

Dionysios G. Kalliakmanis


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

May 2020

# Ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε στα πλαίσια του μεταπτυχιακού προγράμματος Εφαρμοσμένης Στατιστικής του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς. Ως την ελάχιστη δυνατή μνεία, με την παρούσα παράγραφο οφείλω να ευχαριστήσω όλους τους ανθρώπους που συνέβαλαν στην πραγματοποίηση και ολοκλήρωση αυτής της διπλωματικής εργασίας. Κατά κύριο λόγο θα ήθελα να ευχαριστήσω τον Επιβλέποντα Αναπληρωτή Καθηγητή κ. Πολίτη Κωνσταντίνο που μου έδωσε την ευκαιρία να ασχοληθώ με ένα τόσο ενδιαφέρον θέμα που συνδυάζει την στατιστική με τον αθλητισμό. Τον ευχαριστώ για την διαρκή καθοδήγηση που είχα από την αρχή μέχρι και το τέλος της συγγραφής αυτής της μελέτης καθώς και για την πολύτιμη υποστήριξή του, τις παραγωγικές υποδείξεις και το πολύ καλό κλίμα που διαμόρφωσε συμβάλλοντας τα μέγιστα για την κατάρτιση της διπλωματικής μου εργασίας. Τέλος, δεν θα μπορούσα να μην ευχαριστήσω τα αγαπημένα μου οικογενειακά και συγγενικά πρόσωπα, για την ανιδιοτελής υλική και ηθική υποστήριξη που μου παρείχαν, καθ' όλη την διάρκεια των μεταπτυχιακών μου σπουδών.

# Περίληψη

Στις μέρες μας παράγονται, επεξεργάζονται και αποθηκεύονται συνεχώς όλο και περισσότερα δεδομένα με ραγδαίους ρυθμούς. Αύτη η έκρηξη του όγκου των δεδομένων έχει επηρεάσει σαφώς και τον χώρο του αθλητισμού, και πιο συγκεκριμένα τον χώρο της καλαθοσφαίρισης. Η ανάλυση δεδομένων και ο αθλητισμός συμβαδίζουν εδώ και καιρό. Η στατιστική αναφέρεται σε ένα σύνολο εργαλείων για την μοντελοποίηση και την κατανόηση σύνθετων συνόλων δεδομένων. Οι αθλητικές αναλύσεις είναι μια συλλογή σχετικών, ιστορικών στατιστικών στοιχείων τα οποία, όταν εφαρμόζονται σωστά, μπορούν να προσφέρουν ένα ανταγωνιστικό πλεονέκτημα σε μία ομάδα ή και σε ατομικό επίπεδο. Μέσω της συλλογής και της ανάλυσης αυτών των στοιχείων, οι αθλητικές αναλύσεις ενημερώνουν τους παίχτες, τους προπονητές και το υπόλοιπο προσωπικό, για να διευκολύνουν τη λήψη αποφάσεων τόσο κατά την διάρκεια όσο και πριν από τα αθλητικά γεγονότα.

Σε αυτή την μελέτη, χρησιμοποιώντας πραγματικά δεδομένα από το πιο διάσημο Ευρωπαικό πρωτάθλημα καλαθοσφαίρισης, θα αναλύσουμε με τεχνικές στατιστικής και μηχανικής μάθησης, ποιες στατιστικά στοιχεία είναι καθοριστικά για την απόδοση της ομάδας, σε ετήσια βάση. Επιπλέον, θα παρουσιάσουμε μια περιγραφική ανάλυση και θα απεικονίσουμε τα αποτελέσματα μέσω γραφικών παραστάσεων, γραφημάτων και πινάκων. Στη συνέχεια, θα εφαρμόσουμε μοντέλα λογιστικής παλινδρόμησης με σκοπό να βρούμε τα βασικά χαρακτηριστικά που επηρεάζουν το τελικό αποτέλεσμα ενός παιχνιδιού. Τέλος, θα προβλέψουμε το ποσοστό ορθής ταξινόμησης και θα αξιολογήσουμε την απόδοση αυτών των μοντέλων μέσα από διαφορετικά κριτήρια, όπως πίνακες ταξινόμησης, την περιοχή κάτω από την καμπύλη και τον αλγόριθμο random forest.

# Abstract

Nowadays, more and more data is being produced, processed and stored at a rapid pace. This explosion in the volume of data has clearly affected the field of sports, and more specifically the field of basketball. Data analytics and sports are going hand-in-hand for some time now. Statistical learning refers to a set of tools for modeling and understanding complex datasets. **Sports analytics** are a collection of relevant, historical, statistics that when properly applied can provide a competitive advantage to a team or individual. Through the collection and analyzation of these data, sports analytics inform players, coaches and other staff in order to facilitate decision making both during and prior to sporting events.

In this study, using actual data from the most prestigious European Basketball Tournament, we will analyze with statistical and machine learning techniques, which statistics in the box-score determine the team's season-long success. Additionally, we will perform a descriptive analysis and visualize the results through graphs, charts and tables. Afterwards, we will employ logit, probit, cauchit and cloglog models aiming to found the key characteristics affecting the final outcome of a game. Finally, we will predict the accuracy and evaluate the performance of those models through different criteria, such as confusion matrix, area under curve (AUC) and random forest algorithm.

*" Torture data, and it will confess to anything"*

*Ronald Coase, Economist*

# Περιεχόμενα

# Κατάλογος πινάκων

# Κατάλογος σχημάτων

# ΚΕΦΑΛΑΙΟ 1

## Εισαγωγή

Σε αυτό το κεφάλαιο, θα αναφερθούν τεχνικές εξόρυξης γνώσης που έχουν εφαρμογή στο άθλημα του μπάσκετ. Στη συνέχεια θα γίνει μία ιστορική αναδρομή στην **Ευρωλίγκα,** την σπουδαιότερη διασυλλογική διοργάνωση καλαθοσφαίρισης για τους άντρες στην Ευρώπη. Τέλος, θα δοθεί μια συνοπτική περιγραφή των δεδομένων καθώς και των μεταβλητών που θα χρησιμοποιηθούν στην συγκεκριμένη ανάλυση.

### 1.1 Όγκος δεδομένων και εξόρυξη γνώσης

Στην σημερινή εποχή παράγονται και αποθηκεύονται συνεχώς όλο και περισσότερα δεδομένα με ραγδαίους ρυθμούς. Τέτοια δεδομένα μπορεί να είναι κυβερνητικά (εφορία), επιστημονικά (NASA), τραπεζικά, διαδικτυακά, αθλητικά κ.α. Χαρακτηριστικά αξίζει να αναφέρουμε πως στο διαδίκτυο υπάρχουν 50 δισεκατομμύρια σελίδες διασυνδεδεμένες, στο Facebook 400 εκατομμύρια χρήστες και 1 δισεκατομμύριο χρήστες χρησιμοποιούν το instant messenger, με τους αριθμούς αυτούς να συνεχίζουν να αυξάνονται σε καθημερινή βάση. Συνολικά το πλήθος των δεδομένων στις μέρες μας αγγίζει τα 44 zettabyte. Η πρόβλεψη για το 2025 είναι ότι το μέγεθος των δεδομένων θα φτάνει τα 175 zettabyte παγκοσμίως. Για να συνειδητοποιήσουμε αυτό το μέγεθος αρκεί να αναφέρουμε πως 1 zettabyte=1000 exabyte = 1 million petabyte = 1 billion terabyte = 1 trillion gigabyte .

(Πηγή : https://www.forbes.com/)

Ο άνθρωπος έχει από την φύση του περιορισμένες αναλυτικές δυνατότητες. Ακόμα και χωρίς την εκρηκτική αύξηση του όγκου των δεδομένων που παρατηρήθηκε ιδιαίτερα τα τελευταία χρόνια, τού είναι πολύ δύσκολο να επεξεργαστεί γρήγορα και αποτελεσματικά τα δεδομένα που έχει στην διάθεση του. Οι επιχειρήσεις για να είναι ανταγωνιστικές θα πρέπει να έχουν την δυνατότητα να αξιοποιήσουν όλες τις πληροφορίες και τα δεδομένα που συλλέγουν ώστε να οδηγηθούν στην μεγιστοποίηση του κέρδους τους. Για την επεξεργασία των στοιχείων αυτών, όμως, απαιτούνται εξειδικευμένα εργαλεία, ώστε να επιτευχθεί η ακριβέστερη και γρηγορότερη προσέγγισή τους. Η επιστήμη της Στατιστικής προσφέρει λύσεις ανάλυσης δεδομένων, δεν λαμβάνει όμως μέριμνα για το πρόβλημα του πολύ μεγάλου όγκου τους. Ο κλάδος των Βάσεων Δεδομένων είναι ο κατ' εξοχήν αρμόδιος για την τήρηση μεγάλου όγκου δεδομένων, όμως η σχεδιαστική φιλοσοφία του είναι προσανατολισμένη στην καταχώρηση, στη διαχείριση και στην ανάκτηση των δεδομένων, όχι όμως και στην ανάλυσή τους. Έτσι λοιπόν, το βασικό πρόβλημα που τίθεται είναι πώς θα μετατραπεί όλη αυτή η αυξανόμενη πληροφόρηση σε γνώση, με αποτελεσματικό τρόπο. Μια ευρέως διαδεδομένη μέθοδος επεξεργασίας και ανάλυσης δεδομένων είναι η **επιστήμη της Εξόρυξης Δεδομένων (Data Mining).** Με τον παραπάνω όρο εννοούμε την εξεύρεση μιας ενδιαφέρουσας, αυτόνομης, μη προφανούς και πιθανόν χρήσιμης πληροφορίας, με την χρήση αλγορίθμων ομαδοποίησης ή κατηγοριοποίησης και των

αρχών της στατιστικής, της τεχνητής νοημοσύνης, της μηχανικής μάθησης και των συστημάτων βάσεων δεδομένων. (Πηγή : https://el.wikipedia.org/ )

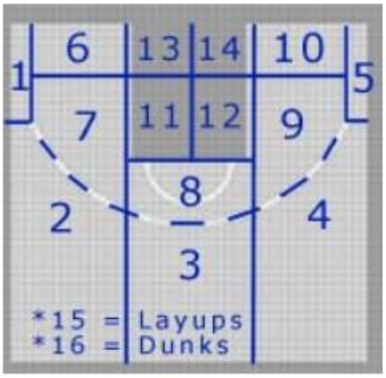
### 1.2 Τεχνικές εξόρυξης γνώσης για την πρόβλεψη αγώνων μπάσκετ

Αύτη η έκρηξη του όγκου των δεδομένων έχει επηρεάσει σαφώς και τον χώρο του αθλητισμού, και πιο συγκεκριμένα τον χώρο της καλαθοσφαίρισης. Παρακάτω αναφέρονται κάποιες μελέτες που έχουν προηγηθεί από ερευνητές σχετικές με το ποια στατιστικά στοιχεία είναι σημαντικά για την πρόβλεψη του τελικού αποτελέσματος σε έναν αγώνα μπάσκετ.

→Ο Oliver (1980) ήταν ο πρώτος ερευνητής που εφάρμοσε μία τεχνική εξόρυξης γνώσης σε έναν αγώνα μπάσκετ. Πιο συγκεκριμένα, δημιούργησε APBPmetrics ή με άλλα λόγια κάποια advanced statistics (Association of Professional Basketball Researchers) ώστε να χρησιμοποιούνται καλύτερα κριτήρια για στατιστικές αναλύσεις και συγκρίσεις. Η ανάλυσή του στηρίχθηκε σε ολόκληρη την ομάδα και όχι σε κάθε παίχτη ξεχωριστά.

→Ένας τρόπος για την αξιολόγηση της αποτελεσματικότητας του κάθε παίχτη ξεχωριστά είναι το player efficiency rating (PER), το οποίο βασίζεται στο rating ανά λεπτό. (Hollinger, 2002)

→Μία άλλη ενδιαφέρουσα τεχνική που έχει χρησιμοποιηθεί, κυρίως στο NBA, είναι αυτή των shot zones. Το γήπεδο χωρίζεται σε 16 περιοχές, όπου ο παίχτης που βρίσκεται στην επίθεση μπορεί να βρίσκεται σε πιθανή θέση για σουτ. Μέσα από την ανάλυση του ποσοστού ευστοχίας κάθε παίχτη σε κάθε ζώνη, μπορούν να εξαχθούν σημαντικά συμπεράσματα από τους εκάστοτε προπονητές για τον περιορισμό αυτών των σουτ. (Beech, 2008)




(Πηγή : https://www.82games.com/shotzones.htm)

Σχήμα 1.1 : Shot zones in NBA games

→ Οι Orendorff και Johnson χρησιμοποίησαν **Bayesian Logic (BLOG) and Markov Logic Networks (MLNs)** για να προβλέψουν το τελικό αποτέλεσμα για τους αγώνες

του NBA. Το project τους βασίστηκε πάνω σε ιστορικά δεδομένα. Αφού εφαρμόστηκε prediction accuracy για τα μοντέλα BLOG και MLN frameworks έγινε σύγκριση μεταξύ τους χρησιμοποιώντας cross validation για την περίοδο 2006-2007 . Η μέθοδος MLN έδωσε ποσοστό ορθής ταξινόμησης ίσο με 64% και 63% αντίστοιχα χρησιμοποιώντας το BLOG μοντέλο. (Orendorff and Johnson, 2007)

→ Οι Bernard Loeffelholz, Earl Bednarand και Kenneth W. Bauer έκαναν μία έρευνα με την οποία προέβλεπαν το τελικό αποτέλεσμα για τους αγώνες του NBA χρησιμοποιώντας neural networks. Οι ερευνητές διερεύνησαν τα υποσύνολα που λήφθηκαν από signal-to-noise ratios και γνώμες εμπειρογνωμόνων για να προσδιορίσουν ένα υποσύνολο χαρακτηριστικών που εισάγονται στα νευρωνικά δίκτυα. Τα αποτελέσματα που προέκυψαν από αυτά τα δίκτυα συγκρίθηκαν με τις προβλέψεις πολλών ειδικών στον τομέα του μπάσκετ. Μετά το πείραμα, το ποσοστό ορθής ταξινόμησης (accuracy) του μοντέλου ήταν 74.33% (Loeffelholz, B. et al, 2009)

→ Μια διαφορετική προσέγγιση έρχεται στην δημοσιότητα από τους Beckler, Wang και Papamichael οι οποίοι εφάρμοσαν applied machine learning τεχνικές για να προβλέψουν το τελικό αποτέλεσμα για αγώνες του NBA, χρησιμοποιώντας λογιστική παλινδρόμηση (Logistic Regression) με ποσοστό ορθής ταξινόμησης 68.1%, γραμμική παλινδρόμηση( linear regression) με ποσοστό ορθής ταξινόμησης 65.4% και Support Vector Machines (SVM) με τελικό ποσοστό ορθής ταξινόμησης ίσο με 66.9%.(Beckler, M. et al, 2008)

(Πηγή :
https://arrow.tudublin.ie/cgi/viewcontent.cgi?article=1040&context=scschcomdis )

Οι παραπάνω είναι ενδεικτικά κάποιες τεχνικές που χρησιμοποιήθηκαν για την πρόβλεψη του τελικού αποτελέσματος κυρίως στην Αμερική. Στο συγκεκριμένο project θα προσπαθήσουμε να διαπιστώσουμε ποια στατιστικά στοιχεία επηρεάζουν περισσότερο το τελικό αποτέλεσμα στην Ευρώπη χρησιμοποιώντας λογιστική παλινδρόμηση, με διαφορετικές συναρτήσεις σύνδεσης κάθε φορά ώστε να επιλέξουμε το μοντέλο που έχει τη μεγαλύτερη ακρίβεια.

### 1.3 Ιστορική αναδρομή του μπάσκετ στην Ευρώπη

Η **Ευρωλίγκα** είναι η σπουδαιότερη διασυλλογική διοργάνωση καλαθοσφαίρισης για τους άντρες στην Ευρώπη και αναδεικνύει τον Πρωταθλητή Ευρώπης στο άθλημα. Από την περίοδο 2000–01 μέχρι το 2004-05 η ευθύνη της διοργάνωσης ανήκει στην ιδιωτική εταιρεία Euroleague Basketball, καθώς και στις ομάδες που προέρχονται από μια πανευρωπαϊκή κοινοπραξία των κορυφαίων επαγγελματικών συλλόγων καλαθοσφαίρισης, που ονομάζεται Ένωση Ευρωπαϊκών Επαγγελματικών Ομάδων Καλαθοσφαίρισης (ULEB). Ακόμα από την περίοδο 2004-05 μέχρι το 2010-11 ο θεσμός διοργανώθηκε από την Fiba Europe. Από την περίοδο 2010-2011 ονομάζεται «Turkish Airlines Euroleague» (TAE) λόγω της υφιστάμενης χορηγίας της από την Turkish Airlines.

Η διοργάνωση ξεκίνησε την περίοδο 1957-58 (όλοι οι αγώνες διεξήχθησαν το ημερολογιακό έτος 1958) με την ονομασία *Κύπελλο Πρωταθλητριών*, κατά τα πρότυπα της αντίστοιχης ποδοσφαιρικής διοργάνωσης της UEFA. Μέχρι το 1990-91 μετείχαν

μόνο οι πρωταθλήτριες ομάδες της κάθε χώρας. Από την επόμενη σεζόν άρχισαν να προστίθενται δευτεραθλήτριες ή και τριταθλήτριες ομάδες από τις καλαθοσφαιρικά προηγμένες χώρες της και η διοργάνωση άλλαξε μορφή. Το 1996-97 καταργήθηκαν οι προκριματικοί νοκ - άουτ αγώνες και οι ομάδες ξεκινούν την πορεία τους από την αρχική φάση των ομίλων, ταυτόχρονα όμως έχασαν το δικαίωμα συμμετοχής οι χώρες όπου το άθλημα δε γνωρίζει ιδιαίτερη ανάπτυξη.

Ένα άλλο χαρακτηριστικό των πρόσφατων διοργανώσεων, έχει να κάνει με τη δέσμευση κλειστών συμβολαίων ανάμεσα στη Euroleague Basketball και κάποιους συλλόγους που θεωρούνται ιδιαίτερα «εμπορικοί». Αυτό πρακτικά σημαίνει πως κάποιες ομάδες έχουν δικαίωμα συμμετοχής ασχέτως των επιδόσεών τους στα εθνικά πρωταθλήματα. Κάτι τέτοιο έχει συμβεί και με ελληνικούς συλλόγους - Ολυμπιακός και ΑΕΚ έχουν στο παρελθόν αγωνισθεί στην Ευρωλίγκα, μολονότι δεν είχαν καταταγεί σε υψηλή θέση στο εθνικό πρωτάθλημα της προηγούμενης αγωνιστικής περιόδου.

Την αγωνιστική περίοδο 1965-66 δοκιμάστηκε για πρώτη φορά η διαδικασία Final Four ανάμεσα στις τέσσερις ομάδες που θα έφταναν στη φάση των ημιτελικών, κατά τα πρότυπα του αμερικανικού κολεγιακού πρωταθλήματος. Πολύ σύντομα η ιδέα εγκαταλείφθηκε, για να επανέλθει πολλά χρόνια αργότερα (1987-88) και να καθιερωθεί έως σήμερα. Μόνη εξαίρεση υπήρξε το 2001, όταν η διοργάνωση της FIBA περιλάμβανε final four, ενώ η αντίστοιχη της Euroleague Basketball όχι.

Τον Νοέμβριο του 2015 η Euroleague Basketball και η πολυεθνική εταιρεία IMG, ειδικευμένη σε αθλητικά, μόδα και μάρκετινγκ υπέγραψαν δεκαετές συμβόλαιο συνεργασίας, αλλάζοντας τη δομή της διοργάνωσης προσφέροντας παράλληλα στις ομάδες περισσότερα έσοδα. Στη νέα Ευρωλίγκα, λοιπόν, αποφασίστηκε να συμμετέχουν 16 ομάδες, και όλες να τεθούν αντιμέτωπες μεταξύ τους σε μία κανονική περίοδο που θα διαρκεί 30 αγωνιστικές. Στη συνέχεια, λήφθηκε η απόφαση οι οκτώ πρώτες να συμμετέχουν στην προημιτελική φάση (1-8, 2-7, 3-6, 4-5) και ύστερα από μία σειρά πέντε αγώνων, όποιος δηλαδή φτάσει τις τρεις νίκες, να προκριθεί στο Final 4. Από τις 16 ομάδες οι 11 έχουν συμβόλαιο τύπου Α, δηλαδή σταθερή παρουσία στο θεσμό για τα επόμενα δέκα χρόνια. Αυτές οι ομάδες είναι ο Παναθηναϊκός, ο Ολυμπιακός η Ρεάλ Μαδρίτης, η Μπαρτσελόνα, η Μπασκόνια, η ΤΣΣΚΑ Μόσχας, η Φενέρμπαχτσε, η Ανατολού Εφές, η Εμπόριο Αρμάνι Μιλάνο, η Μακάμπι Τελ Αβίβ και η Ζάλγκιρις Κάουνας. Οι υπόλοιπες πέντε άδειες που εξασφαλίζονται μέσω παρουσίας στις διοργανώσεις των χωρών, διανέμονται ως εξής: οι νικητές της Αδριατικής Λίγκα, και του πρωταθλήματος Γερμανίας, τη VTB League (ο φιναλίστ αν το πρωτάθλημα παίρνει η ΤΣΣΚΑ Μόσχας), τον κάτοχο του Eurocup της περασμένης περιόδου καθώς και μια ομάδα από την Ισπανία.

Από τη σεζόν 2019-20, οι ομάδες αυξήθηκαν σε 18, με το ισχύον καθεστώς των 11 ομάδων με συμβόλαιο τύπου Α αλλά και την προσθήκη ομάδων με δυο wild card που δόθηκαν σε Βιλερμπάν και Ζενίτ, για την συγκεκριμένη σεζόν.

(Πηγή : https://el.wikipedia.org/ )

### 1.4 Συλλογή δεδομένων

Για τις ανάγκες αυτής της μελέτης, συλλέξαμε στοιχεία από το επίσημο site της Euroleague. Τα δεδομένα μας αφορούν όλους τους αγώνες για την κανονική περίοδο τις χρονιές 2014-2015 και 2015-2016. Σε αυτή την διοργάνωση συμμετέχουν οι 24 καλύτερες ομάδες της Ευρώπης χωρισμένες σε 4 group. Το κάθε group αποτελείται

από 6 ομάδες, και από αυτές οι 4 πρώτες προκρίνονται στην επόμενη φάση (top-16), ενώ οι υπόλοιπες 2 δεν προκρίνονται στην επόμενη φάση. Σε κάθε όμιλο πραγματοποιούνται συνολικά 10 παιχνίδια από την κάθε ομάδα οπότε συνολικά έχουμε 30 αγώνες σε κάθε ένα από τα 4 group. Με άλλα λόγια τις χρονιές 2014-2015 και 2015-2016 έχουν διεξαχθεί συνολικά 240 παιχνίδια όσον αφορά την κανονική περίοδο. Για την ιστορία, αναφέρουμε ότι την περίοδο 2014-2015 κατέκτησε το τρόπαιο στην διοργάνωση η Real Madrid, ενώ την δεύτερη και τρίτη θέση κατέκτησαν ο Olympiacos Piraeus και η CSKA Moscow αντίστοιχα. Την επόμενη σεζόν την πρώτη θέση κατέκτησε η CSKA Moscow, ενώ την δεύτερη και τρίτη θέση κατέκτησαν η Fenerbahce Istanbul και η Lokomotiv Kuban Krasnodar αντίστοιχα. Επίσης αναφέρουμε πως όλα τα Ευρωπαικά παιχνίδια αποτελούνται από 4 περιόδους των 10 λεπτών. Στους αγώνες δεν υπάρχει ισοπαλία και σε αυτή την περίπτωση οδηγούμαστε σε 5λέπτη παράταση μέχρι μία από τις 2 ομάδες να αναδειχθεί νικήτρια.

### 1.5 Παρουσίαση μεταβλητών

Για να εξετάσουμε την εξέλιξη του σκορ σε κάθε αγώνα θα χρησιμοποιήσουμε τις παρακάτω μεταβλητές :

| Ranks | Κατάταξη των ομάδων στο Group |
|---|---|
| Qualified | Πρόκριση ή όχι της ομάδας στο top 16 |
| Pts.Total (points) | Συνολικοί πόντοι της ομάδας στα 10 παιχνίδια της regular season |
| Dif_Points | Συνολική διαφορά πόντων της ομάδας από την αντίπαλη στα 10 παιχνίδια της regular season |
| 2FG (2-point field goals) | Ποσοστά ευστοχίας για σουτ 2 πόντων στα 10 παιχνίδια της regular season |
| 3FG (3-point field goals) | Ποσοστά ευστοχίας για σουτ 3 πόντων στα 10 παιχνίδια της regular season |
| FT (free throw) | Ποσοστό ευστοχίας ελεύθερων βολών στα 10 παιχνίδια της regular season |
| O.TOTAL (offensive rebounds) | Συνολικά επιθετικά «ριμπάουντ» στα 10 παιχνίδια της regular season |
| D.TOTAL (defensive rebounds) | Συνολικά αμυντικά «ριμπάουντ» στα 10 παιχνίδια της regular season |

| | |
|---|---|
| T.TOTAL (rebounds) | Συνολικά «ριμπάουντ» στα 10 παιχνίδια της regular season |
| As.TOTAL (assists) | Συνολικές ασίστ στα 10 παιχνίδια της regular season |
| St.TOTAL (steals) | Συνολικά κλεψίματα στα 10 παιχνίδια της regular season |
| To.TOTAL (turnovers) | Συνολικά λάθη στα 10 παιχνίδια της regular season |
| Fv.TOTAL (blocks in favor) | Συνολικά κοψίματα υπέρ της ομάδας στα 10 παιχνίδια της regular season |
| Ag.TOTAL (blocks against) | Συνολικά κοψίματα κατά της ομάδας στα 10 παιχνίδια της regular season |
| Cm.TOTAL (fouls committed) | Συνολικά φάουλ που διέπραξε η ομάδα στα 10 παιχνίδια της regular season |
| Rv.TOTAL (fouls received) | Συνολικά φάουλ που δέχθηκε η ομάδα στα 10 παιχνίδια της regular season |
| PIR.TOTAL (performance index rating*) | Ειδικός δείκτης αξιολόγησης της euroleague στα 10 παιχνίδια της regular season |

Πίνακας 1.1 Βασικές μεταβλητές

Ο ειδικός δείκτης αξιολόγησης της διοργάνωσης λαμβάνει υπόψιν του αρκετούς παράγοντες και διαμορφώνεται ως εξής :

***PIR = (Points+Rebounds+Assists+Steals+Blocks+Fouls Drawn) – (Missed Field Goals + Missed Free Throws +Turnovers+Shots Rejected+Fouls Committed)**

Η μεταβλητή qualified είναι δίτιμη, παίρνοντας την τιμή 0 εάν η ομάδα δεν προκρίθηκε στην επόμενη φάση και την τιμή 1 εάν η ομάδα έχει προκριθεί στις 16 καλύτερες της διοργάνωσης.

Τα group που βρίσκεται η κάθε ομάδα και για τις 2 χρονιές που θα εξετάσουμε επισυνάπτονται στο Παράρτημα της εν λόγω εργασίας.

Μια πιο αναλυτική και αντιπροσωπευτική εικόνα για τις μεταβλητές μας μπορούμε να πάρουμε στο επόμενο κεφάλαιο.

Στο 5ο κεφάλαιο ,όπου θα δώσουμε βαρύτητα στο ερώτημα αν η έδρα της ομάδας επηρεάζει το τελικό αποτέλεσμα ενός αγώνα, θα χρησιμοποιηθούν επιπλέον οι μεταβλητές :

| | |
|---|---|
| Dif_Q1 | Συνολική διαφορά πόντων στο πρώτο δεκάλεπτο της γηπεδούχου ομάδας από την φιλοξενούμενη στα 120 παιχνίδια της regular season |
| Pts.Q1 | Συνολικοί πόντοι στο πρώτο δεκάλεπτο για κάθε ομάδα στα 120 παιχνίδια της regular season |
| Dif_Q2 | Συνολική διαφορά πόντων στο δεύτερο δεκάλεπτο της γηπεδούχου ομάδας από την φιλοξενούμενη στα 120 παιχνίδια της regular season |
| Pts.Q2 | Συνολικοί πόντοι στο δεύτερο δεκάλεπτο για κάθε ομάδα στα 120 παιχνίδια της regular season |
| Dif_Q3 | Συνολική διαφορά πόντων στο τρίτο δεκάλεπτο της γηπεδούχου ομάδας από την φιλοξενούμενη στα 120 παιχνίδια της regular season |
| Pts.Q3 | Συνολικοί πόντοι στο τρίτο δεκάλεπτο για κάθε ομάδα στα 120 παιχνίδια της regular season |
| Dif_Q4 | Συνολική διαφορά πόντων στο τέταρτο δεκάλεπτο της γηπεδούχου ομάδας από την φιλοξενούμενη στα 120 παιχνίδια της regular season |
| Pts.Q4 | Συνολικοί πόντοι στο τέταρτο δεκάλεπτο για κάθε ομάδα στα 120 παιχνίδια της regular season |

| | |
|---|---|
| Dif_ET | Συνολική διαφορά πόντων στην περίπτωση που είχαμε παράταση για τα 120 παιχνίδια της regular season |
| Pts.ET | Συνολικοί πόντοι στην περίπτωση που είχαμε παράταση για τα 120 παιχνίδια της regular season |
| HOME.AWAY | Γηπεδούχος/Φιλοξενουμενος |

Πίνακας 1.2 : Μεταβλητές που θα χρησιμοποιηθούν στο Κεφάλαιο 5

Τέλος, για όλους τους στατιστικούς ελέγχους και για την επεξεργασία των παραπάνω δεδομένων θα χρησιμοποιηθεί το λογισμικό ανοιχτού κώδικα (open source) «R». Αναλυτικά, ο κώδικας παρουσιάζεται στο επισυναπτόμενο παράρτημα.

# ΚΕΦΑΛΑΙΟ 2

## Περιγραφική ανάλυση

Σε αυτό το κεφάλαιο θα πάρουμε μια γενική εικόνα για τα δεδομένα μας ώστε στα επόμενα κεφάλαια να προχωρήσουμε σε περαιτέρω ανάλυση. Πιο συγκεκριμένα θα δούμε αναλυτικά κάποια περιγραφικά μέτρα για τις επεξηγηματικές μεταβλητές μας σε σχέση με την μεταβλητή απόκρισης qualified. Στη συνέχεια θα παρουσιάσουμε μια διαγραμματική απεικόνιση των δεδομένων και για τις 2 χρονιές ώστε τα αποτελέσματα να γίνουν ακόμη πιο κατανοητά και συγκρίσιμα από τον αναγνώστη.

### 2.1 Διερεύνηση για missing values

Αρχικά με την βοήθεια της βιβλιοθήκης Data Explorer που διαθέτει η R θα ελέγξουμε αν υπάρχουν ελλιπείς τιμές (missing values) στο δεδομένα μας. Τα αποτελέσματα δείχνουν ότι δεν υπάρχουν missing values. Παρατηρούμε δηλαδή ότι δεν υπάρχει κάποια τιμή NA. Αν υπήρχε θα έπρεπε να την αντικαταστήσουμε. Αυτό μπορεί να πραγματοποιηθεί με αρκετούς τρόπους. Μια λύση είναι να αγνοήσουμε την εγγραφή ή να συμπληρώσουμε την τιμή χειρωνακτικά, κάτι το οποίο δεν προτείνεται. Εναλλακτικά, μπορούμε να συμπληρώσουμε την τιμή αυτόματα με τη χρήση μιας γενικής σταθεράς, με την μέση τιμή του γνωρίσματος για όλα τα δείγματα που ανήκουν στην ίδια κατηγορία ή ακόμη και με την χρήση της πιο πιθανής τιμής. Στην τελευταία περίπτωση η εξαγωγή συμπερασμάτων γίνεται μέσω Bayesian Formula ή δέντρων απόφασης. Στην περίπτωσή μας δεν χρειάζεται να κάνουμε κάτι από τα παραπάνω, συνεπώς θα προχωρήσουμε στην περιγραφική ανάλυση. Με παρόμοιο τρόπο ελέγχουμε και για την επόμενη χρονιά, όπου ξανά δεν παρατηρούνται missing values.

### 2.2 Περιγραφικά μέτρα για την χρονιά 2014-2015

Κάποια βασικά στατιστικά περιγραφικά μέτρα παρουσιάζονται παρακάτω :

| $`0` ΓΙΑ ΤΙΣ ΟΜΑΔΕΣ ΠΟΥ ΔΕΝ ΠΡΟΚΡΙΘΗΚΑΝ | | | | | | |
|---|---|---|---|---|---|---|
| | **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** |
| Pts.Total..M.O | 70,2 | 74,82 | 76,05 | 75,85 | 77,35 | 80,56 |
| dif_Pts.Total..M.O | -12,4 | -10.175 | -7,7 | -6,138 | -3,4 | 3 |
| X2FG.M.O | 46,7 | 47,45 | 49,1 | 49,9 | 51,48 | 56,4 |
| X3FG.M.O | 32,4 | 33,05 | 35,55 | 34,94 | 36,2 | 37,4 |
| FT.M.O | 60,6 | 70,55 | 73,7 | 72,7 | 75,6 | 81 |
| O.TOTAL | 73 | 92,25 | 102,5 | 102,62 | 112,5 | 136 |
| D.TOTAL | 195 | 211 | 229,5 | 228,9 | 247,8 | 260 |
| T.TOTAL | 312 | 325 | 331 | 331,5 | 337,2 | 353 |
| As.Total | 156 | 159,5 | 161 | 163,2 | 166 | 176 |
| St.Total | 44 | 55,5 | 70 | 70,88 | 83,5 | 105 |
| To.Total | 103 | 121,8 | 129 | 126,8 | 135,2 | 147 |
| Fv.Total | 7 | 19 | 22,5 | 23,25 | 30 | 36 |
| Ag.Total | 22 | 23,75 | 24,5 | 28,25 | 33 | 40 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Cm.Total | 199 | 206,5 | 223 | 219,5 | 230,5 | 241 |
| Rv.Total | 194 | 197,5 | 203,5 | 205,1 | 213 | 220 |
| PIR.Total | 722 | 734,2 | 772 | 781,5 | 823 | 871 |
| $`1` ΓΙΑ ΤΙΣ ΟΜΑΔΕΣ ΠΟΥ ΠΡΟΚΡΙΘΗΚΑΝ | | | | | | |
| | **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** |
| Pts.Total..M.O | 71,4 | 76,28 | 77,95 | 79,13 | 81,3 | 88 |
| dif_Pts.Total..M.O | -5,9 | 0 | 1,95 | 3,069 | 5,6 | 16,2 |
| X2FG.M.O | 45,8 | 49,38 | 51,35 | 51,69 | 53,92 | 58,6 |
| X3FG.M.O | 29,8 | 33,75 | 35,45 | 35,55 | 37,48 | 42,5 |
| FT.M.O | 65,9 | 69,8 | 73,05 | 73,47 | 76,9 | 82,7 |
| O.TOTAL | 87 | 100,5 | 110,5 | 110,2 | 114 | 156 |
| D.TOTAL | 218 | 241,5 | 255,5 | 256 | 274,2 | 297 |
| T.TOTAL | 331 | 353,5 | 363 | 366,2 | 376,5 | 425 |
| As.Total | 130 | 162,2 | 169 | 174,5 | 189 | 224 |
| St.Total | 43 | 56 | 65,5 | 62,88 | 69 | 79 |
| To.Total | 100 | 116 | 128 | 129,4 | 140,2 | 165 |
| Fv.Total | 0 | 23 | 27,5 | 28,56 | 37,25 | 45 |
| Ag.Total | 19 | 24,75 | 28 | 28,44 | 31 | 44 |
| Cm.Total | 173 | 197,8 | 214,5 | 210,6 | 226,5 | 237 |
| Rv.Total | 187 | 207,5 | 210 | 210,6 | 217,2 | 231 |
| PIR.Total | 734 | 796,5 | 858,5 | 880,9 | 914,8 | 1107 |

Πίνακας 2.1 Βασικά περιγραφικά μέτρα για την χρονιά 2014-2015

Από τον παραπάνω πίνακα βλέπουμε τις ελάχιστες τιμές, τις μέγιστες τιμές, το μέσο όρο, την διάμεσο καθώς και το πρώτο και τρίτο τεταρτημόριο για τις επιλεγμένες μεταβλητές μας. Έτσι λοιπόν, έχοντας μια γενική εικόνα μπορούμε να παρατηρούμε ότι:

→ Οι ομάδες που προκρίθηκαν στην επόμενη φάση έχουν υψηλότερο μέσο όρο πόντων (79.13) και υψηλότερο δείκτη αξιολόγησης κατά μέσο όρο (880.9) σε σχέση με τις ομάδες που δεν προκρίθηκαν στην επόμενη φάση της διοργάνωσης (75.85 και 781.5 αντίστοιχα). Η ελάχιστη τιμή του ειδικού δείκτη αξιολόγησης της Euroleague διαπιστώθηκε ότι ισούται με 722 ενώ η μέγιστη τιμή του ήταν 1107.

→ Τα κλεψίματα για τις ομάδες που προκρίθηκαν ήταν κατά μέσο όρο 62.88, για το σύνολο όλων των αγώνων στην regular season, ενώ για τις ομάδες που δεν προκρίθηκαν ήταν 70.88. Σε αυτή την περίπτωση ενώ θα περιμέναμε το αντίθετο αποτέλεσμα, βλέπουμε ότι κάτι τέτοιο δεν ισχύει. Το γεγονός αυτό μας δίνει μία πρώτη εικόνα ότι η μεταβλητή αυτή ίσως δεν είναι στατιστικά σημαντική για την πρόβλεψη της ομάδας που προκρίθηκε ή όχι. Αντίθετα, οι assist υπερτερούν περίπου κατά 12 για τις ομάδες που προκρίθηκαν.

→ Ενδιαφέρον παρουσιάζουν και τα στατιστικά για τα σουτ των παιχτών. Όσον αφορά τα σουτ 2 πόντων οι ομάδες που προκρίθηκαν είχαν ποσοστό ευστοχίας ίσο με 51.69% ενώ οι 8 ομάδες που δεν προκρίθηκαν είχαν ποσοστό ευστοχίας 49.90%. Για τα σουτ

3 πόντων είχαμε ποσοστά ευστοχίας 35.55% και 34.94% αντίστοιχα για τις ομάδες που πέρασαν στην επόμενη φάση και για αυτές που δεν τα κατάφεραν. Σε αυτή την περίπτωση βλέπουμε ότι η διαφορά των 2 ποσοστών δεν είναι τόσο μεγάλη και ίσως και αυτή να μην επηρεάζει το τελικό αποτέλεσμα, με μία πρώτη εκτίμηση. Τέλος, αρκετά κοντά είναι και τα ποσοστά των ελεύθερων βολών για τις ομάδες.

→ Όσον αφορά τα συνολικά ριμπάουντ, 366 ήταν περίπου κατά μέσο όρο για τις ομάδες που προκρίθηκαν και 35 λιγότερα κατά μέσο όρο για τις ομάδες που δεν προκρίθηκαν.

→ Ακόμη αξίζει να αναφερθεί πως οι τιμές των μέσων όρων και των διαμέσων είναι αρκετά κοντά σε όλες τις μεταβλητές.

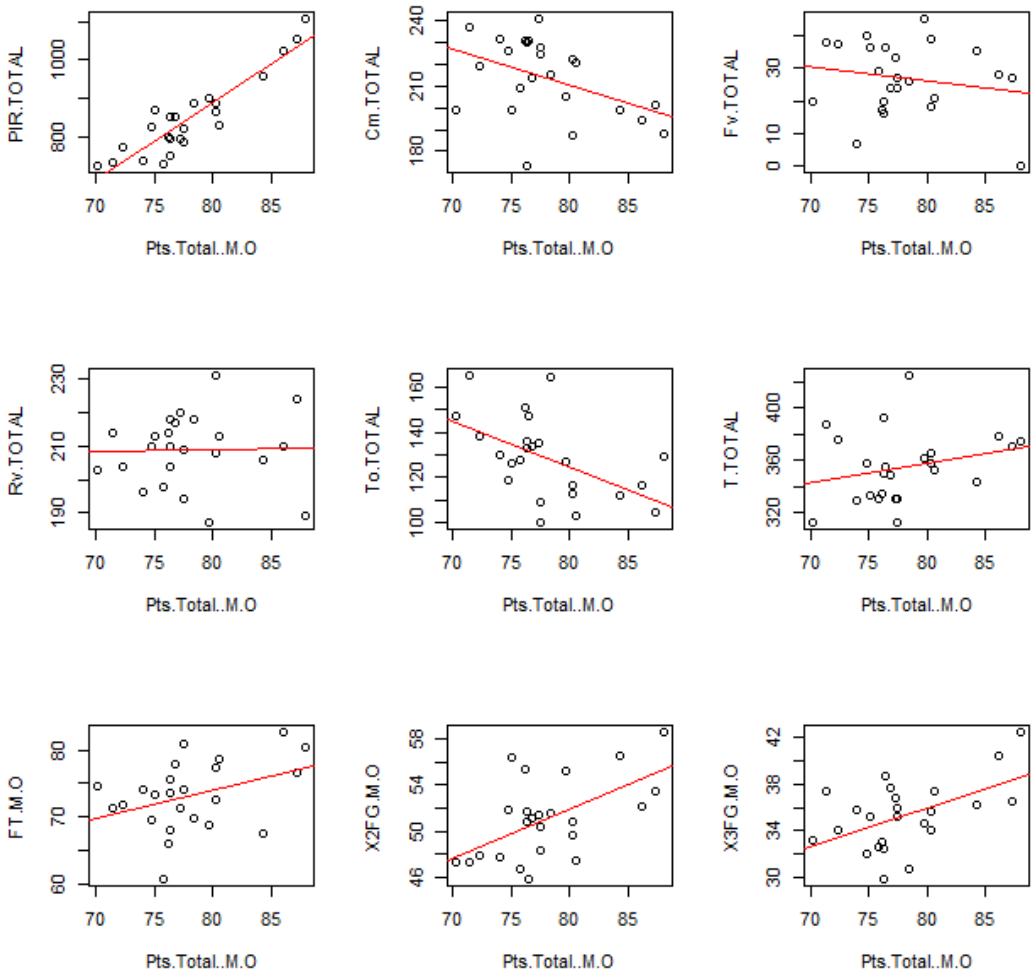
Θα εμβαθύνουμε περισσότερο στις σχέσεις μεταξύ αυτών των μεταβλητών στο επόμενο κεφάλαιο όπου θα εξετασθεί αναλυτικά ο βαθμός σύνδεσης τους, μέσω των συντελεστών συσχέτισης.

### 2.3 Γραφική ανάλυση για την χρονιά 2014-2015

Θα αναπαραστήσουμε ενδεικτικά τα ζεύγη των συνολικών πόντων με 9 από τις υπόλοιπες μεταβλητές μας στα παρακάτω διάγραμμα. Τα διαγράμματα αυτά λέγονται διαγράμματα διασποράς (scatter plot).

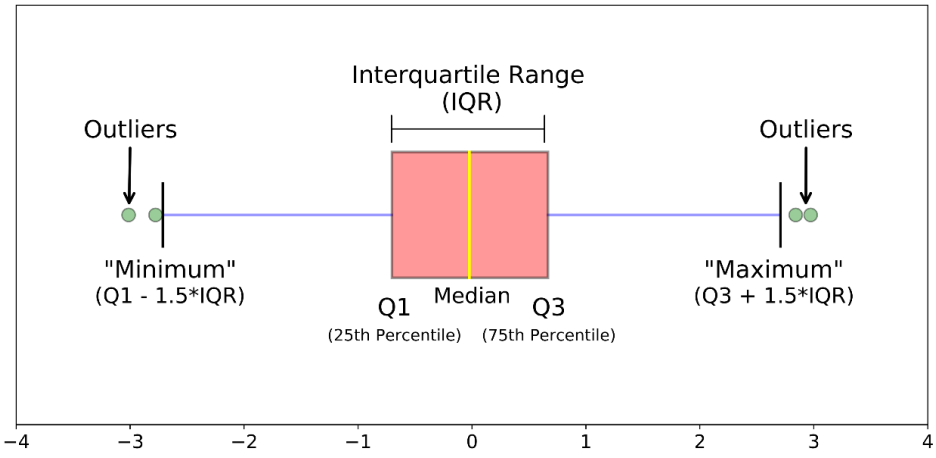
Πιο συγκεκριμένα, και για να γίνουμε περισσότερο κατανοητοί ας εξετάσουμε την σχέση μεταξύ του ειδικού δείκτη αξιολόγησης και των συνολικών πόντων. Από το παρακάτω σχήμα φαίνεται ότι η σχέση τους είναι γραμμική και η ευθεία παλινδρόμησης που προσαρμόστηκε έχει θετική κλίση. Ακόμη τα σημεία βρίσκονται διασκορπισμένα αρκετά κοντά στην ευθεία με αποτέλεσμα να μπορούμε να ισχυριστούμε πως η γραμμική σχέση είναι αρκετά ισχυρή. Έτσι λοιπόν οι πόντοι επηρεάζουν τους ειδικούς δείκτες αξιολόγησης. Αντίστοιχα, στο διάγραμμα όπου απεικονίζονται τα λάθη (turnovers) σε σχέση με τους συνολικούς πόντους βλέπουμε ότι η σχέση τους είναι γραμμική και η ευθεία παλινδρόμησης που προσαρμόστηκε έχει αρνητική κλίση.

Γενικά πάντως στις περισσότερες περιπτώσεις το μόνο που μπορούμε ίσως να πούμε είναι ότι τα σημεία βρίσκονται διασκορπισμένα στο χώρο (τυχαίο μοτίβο). Αξίζει να αναφέρουμε βέβαια πως τα συνολικά κοψίματα υπέρ της ομάδας (Fv) φαίνεται ότι είναι αρνητικά συσχετισμένα με τους πόντους.




Σχήμα 2.1 Scatterplot για 9 μεταβλητές σε σχέση με τους συνολικούς πόντους (2014-2015)

Μια άλλη μέθοδος γραφικής ανάλυσης που μπορούμε να χρησιμοποιήσουμε για συνεχείς μεταβλητές είναι το θηκόγραμμα (boxplot). Ένα παράδειγμα θηκογράμματος δίνεται στην συνέχεια :

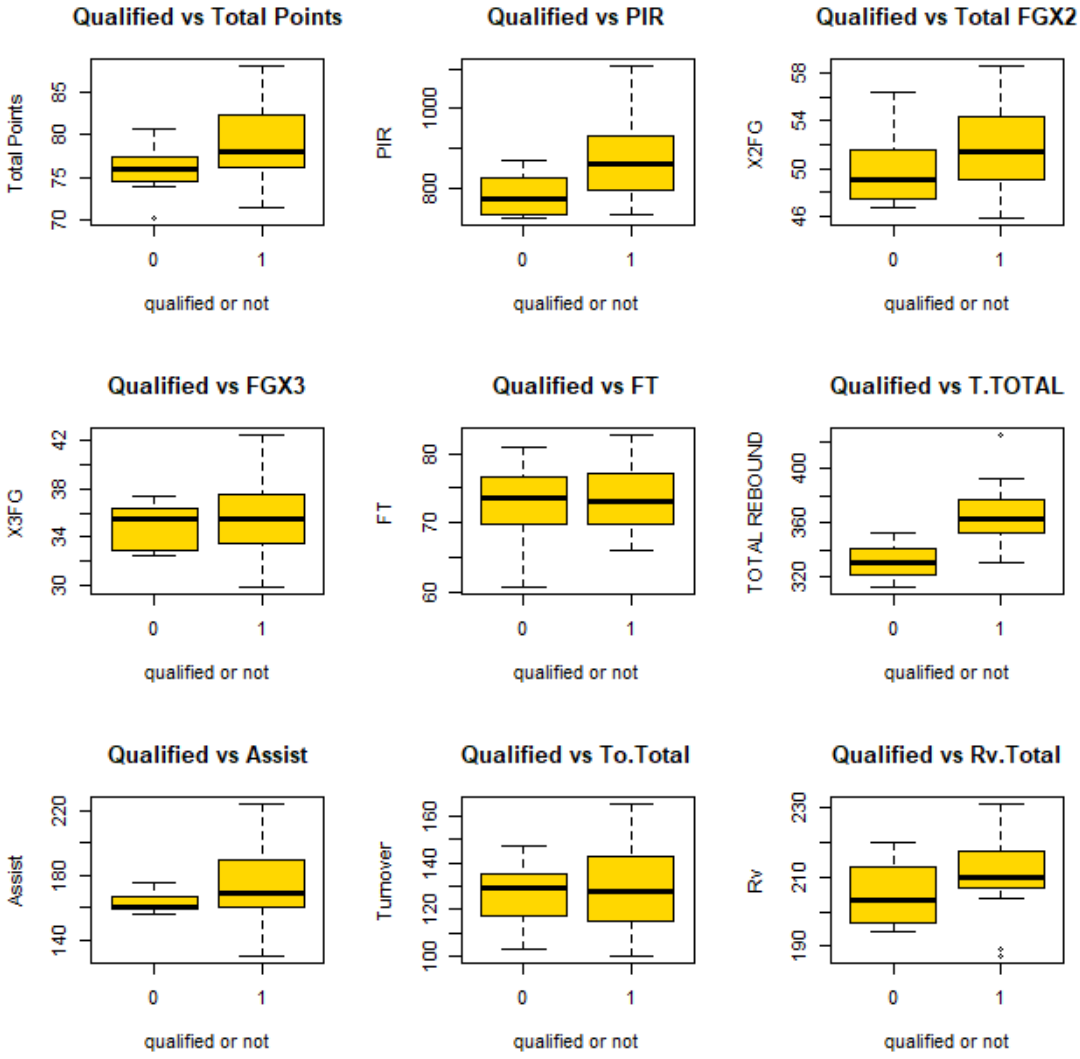


(Πηγή : https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51 )

Σχήμα 2.2 Basic boxplot

Στο παραπάνω σχήμα φαίνονται αναλυτικά σε ποιο ακριβώς σημείο απεικονίζονται τα outliers, η ελάχιστη και μέγιστη τιμή, το ενδοτεταρτημοριακό εύρος, το πρώτο και τρίτο τεταρτημόριο, καθώς και η διάμεσος. Η κατασκευή θηκογραμμάτων είναι ιδιαίτερα χρήσιμη στην περίπτωση που θέλουμε να ελέγξουμε αν υπάρχουν έκτροπες τιμές (outliers) στα δεδομένα που χρησιμοποιούμε ή ακόμη και ακραίες τιμές (extreme values), οι οποίες συνήθως εξαιρούνται από την ανάλυση.

Παρακάτω παρουσιάζουμε ενδεικτικά τα αντίστοιχα boxplots για 9 μεταβλητές σε σχέση με την μεταβλητή qualified :




Σχήμα 2.3 Boxplots για διάφορες μεταβλητές (2014-2015)

→ Είναι σαφές, πώς σχεδόν σε όλα τα σχήματα η διάμεσος για τις ομάδες που προκρίθηκαν (qualified = 1) βρίσκεται υψηλότερα από τις ομάδες που δεν προκρίθηκαν.

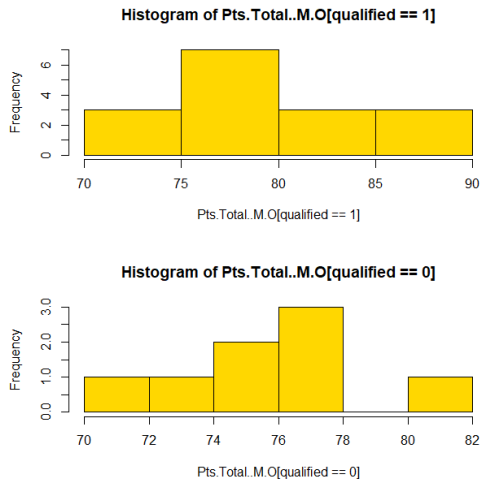
→ Φαίνεται πως παρατηρείται μία ακραία τιμή όσον αφορά τα συνολικά ριμπάουντ και 2 ακραίες τιμές όσον αφορά την μεταβλητή Rv για τις ομάδες που προκρίθηκαν στην επόμενη φάση. Επίσης, μία ακραία τιμή υπάρχει και στους πόντους για τις ομάδες που δεν προκρίθηκαν και ισούται περίπου με 70 όπως βλέπουμε στο πρώτο από τα 9 σχήματα.

→ Ακόμη βλέπουμε πως η διάμεσος του ειδικού δείκτη αξιολόγησης για τις ομάδες που προκρίθηκαν είναι πολύ υψηλότερα σε σχέση με τις ομάδες που δεν προκρίθηκαν, καθώς προσεγγίζει σχεδόν το τρίτο τεταρτημόριο τους ($2^{ο}$ σχήμα)

→ Επιπλέον, στις ελεύθερες βολές οι ομάδες που δεν έχουν προκριθεί στην επόμενη φάση φαίνεται πως έχουν μεγαλύτερο εύρος τιμών

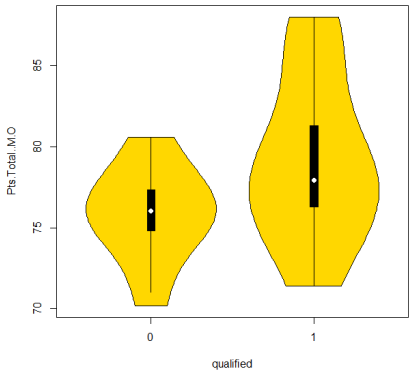
→ Δεν φαίνεται να υπάρχει κάποια έντονη ασσυμετρία, θετική ή αρνητική.

Στην συνέχεια παρουσιάζεται ένα ιστόγραμμα, χωρισμένο σε κλάσεις, που αντικατοπτρίζει το πώς κατανέμονται οι πόντοι των ομάδων. Η κανονικότητα των μεταβλητών ελέγχεται αναλυτικά, όπως θα δούμε στην συνέχεια, στο τρίτο κεφάλαιο.



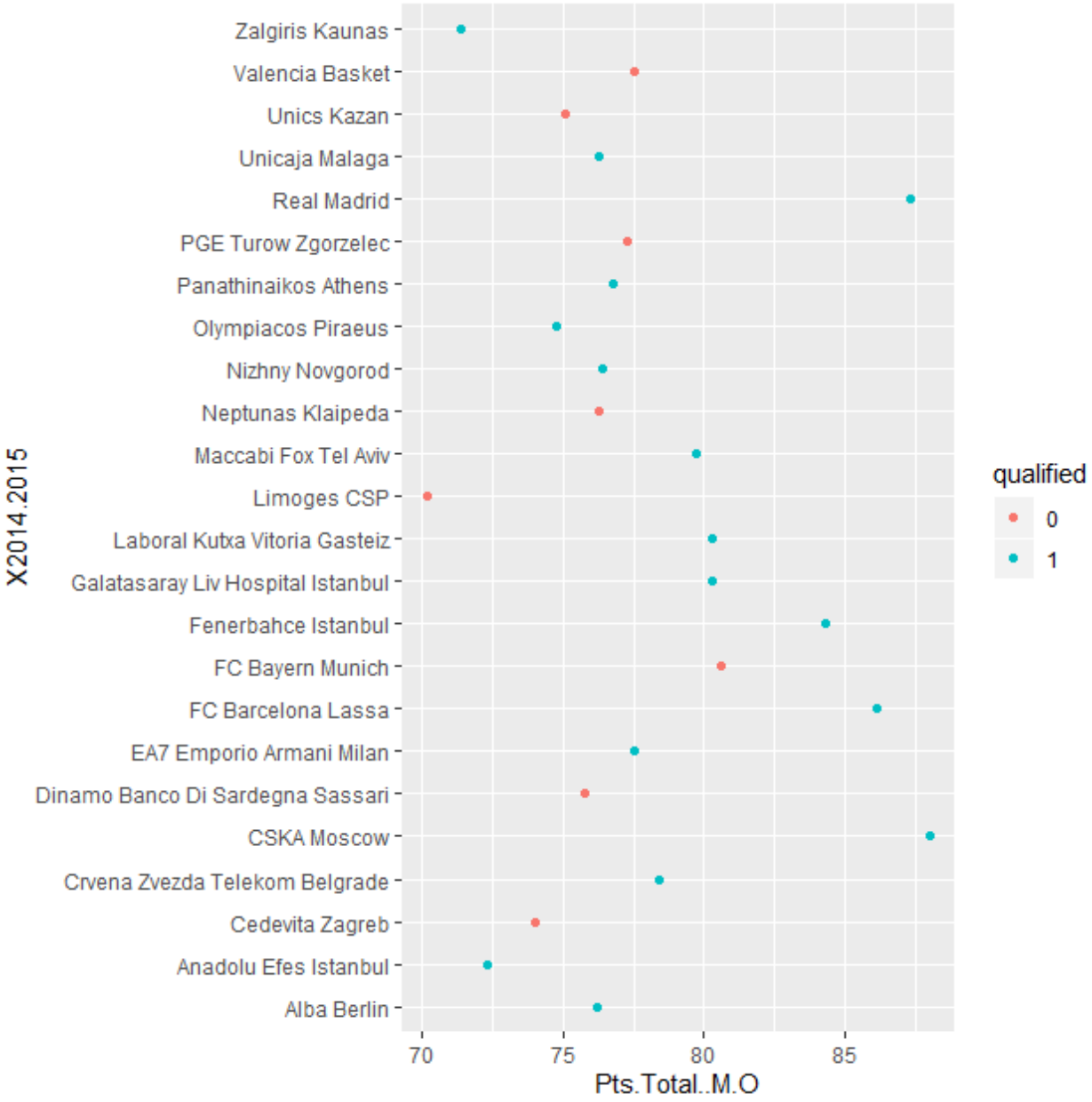

Σχήμα 2.4 Ιστόγραμμα συνολικών πόντων για την χρονιά 2014-2015

Μία παραλλαγή των θηκογραμμάτων αποτελούν τα violin plots, τα οποία κερδίζουν ολοένα και περισσότερο έδαφος όσο αφορά τον τομέα του exploratory data analysis (EDA). Στο παρακάτω σχήμα  πόντων – qualified βλέπουμε την μορφή τους. Η ερμηνεία είναι παρόμοια με αυτή του απλού boxplot. Η μόνη διαφορά τους είναι ότι παρουσιάζουν και μια εκτίμηση της συνάρτησης πυκνότητας για τα δεδομένα.



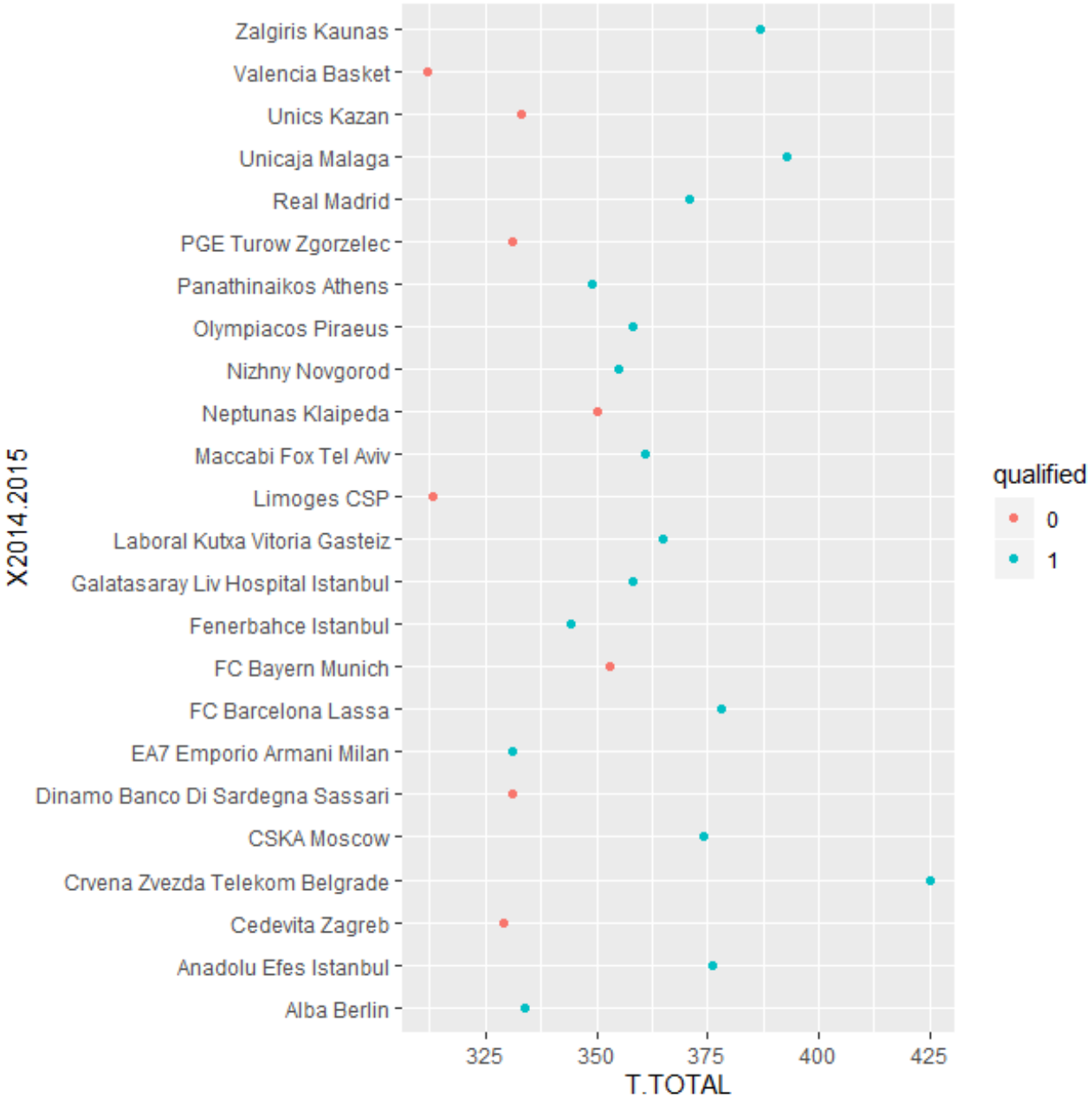

Σχήμα 2.5 Violin plot συνολικών πόντων για την χρονιά 2014-2015

Επιπλέον, χρησιμοποιώντας την βιβλιοθήκη ggplot2 που προσφέρεται από την R και χρησιμοποιείται ευρέως στην οπτικοποίηση των δεδομένων (visualization), παίρνουμε τα παρακάτω διαγράμματα για τους συνολικούς πόντους και τα συνολικά ριμπάουντ αναλυτικά για κάθε μια από τις 24 ομάδες συγκριτικά με το αν προκρίθηκαν ή όχι.




Σχήμα 2.6 Συνολικοί πόντοι για κάθε μία ομάδα ξεχωριστά (2014-2015)

Είναι σαφές ότι τους περισσότερους πόντους κατά μέσο όρο στην regular season 2014-2015 σκόραρε η CSKA Moscow, η Real Madrid και η FC Barcelona Lassa κατά σειρά. Αξίζει να σημειώσουμε ότι η Real Madrid κατέκτησε τελικά το τρόπαιο, λαμβάνοντας την πρώτη θέση στο θεσμό. Την χειρότερη απόδοση σε πόντους φαίνεται πως παρουσιάζει η Limoges CSP με μέσο σκοράρισμα γύρω στους 70 πόντους ανά αγώνα. Σε χαμηλά ποσοστά κυμαίνεται και η Zalgiris Kaunas με περίπου 72 πόντους κατά μέσο όρο. Παρόλα αυτά βλέπουμε πως παρουσιάζεται με πράσινο χρώμα, το οποίο υποδεικνύει ότι προκρίθηκε στο top-16 της διοργάνωσης.




Σχήμα 2.7 Συνολικά ριμπάουντ για κάθε μία ομάδα ξεχωριστά(2014-2015)

Όπως και παραπάνω και εδώ φαίνεται ότι οι ομάδες που είχαν περισσότερα συνολικά ριμπάουντ πέρασαν στην επόμενη φάση του θεσμού. Χαρακτηριστικά αναφέρουμε ότι ο Ερυθρός Αστέρας είχε τα περισσότερα συνολικά ριμπάουντ (425) στην regular season 2014-2015. Η Real Madrid, η οποία κατέκτησε και την διοργάνωση είχε λίγο λιγότερα από 375 ριμπάουντ στους 10 αγώνες που έδωσε στην κανονική περίοδο.

### 2.4 Περιγραφικά μέτρα για την χρονιά 2015-2016

Κάποια βασικά στατιστικά περιγραφικά μέτρα παρουσιάζονται παρακάτω :

| $`0` ΓΙΑ ΤΙΣ ΟΜΑΔΕΣ ΠΟΥ ΔΕΝ ΠΡΟΚΡΙΘΗΚΑΝ | | | | | | |
|---|---|---|---|---|---|---|
| | **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** |
| Pts.Total..M.O | 66,4 | 69,6 | 70,45 | 71,39 | 74,03 | 76,03 |
| dif_Pts.Total..M.O | -14,9 | -11 | -7,9 | -8,175 | -5,55 | -1,7 |
| X2FG.M.O | 44,8 | 47,08 | 48,2 | 47,99 | 49,15 | 50 |
| X3FG.M.O | 30,3 | 31,27 | 33,85 | 34,44 | 36,98 | 39,5 |
| FT.M.O | 67 | 72,53 | 75,7 | 75,15 | 77,7 | 82,3 |
| O.TOTAL | 75 | 92,25 | 99,5 | 98,38 | 103,25 | 123 |
| D.TOTAL | 197 | 215 | 221,5 | 223,5 | 231 | 252 |
| T.TOTAL | 272 | 311 | 323,5 | 321,9 | 334,5 | 360 |
| As.Total | 143 | 161,8 | 168 | 167,1 | 178 | 179 |
| St.Total | 43 | 57,25 | 66 | 66,75 | 76,5 | 84 |
| To.Total | 119 | 127,2 | 133 | 137,6 | 138,8 | 176 |
| Fv.Total | 15 | 18 | 24,5 | 24,38 | 31 | 35 |
| Ag.Total | 17 | 23,5 | 27,5 | 29,12 | 36,75 | 40 |
| Cm.Total | 194 | 203,8 | 207 | 214,5 | 232 | 235 |
| Rv.Total | 187 | 188,5 | 200 | 197,8 | 201,5 | 215 |
| PIR.Total | 639 | 693 | 732 | 725,6 | 744 | 834 |
| $`1` ΓΙΑ ΤΙΣ ΟΜΑΔΕΣ ΠΟΥ ΠΡΟΚΡΙΘΗΚΑΝ | | | | | | |
| | **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** |
| Pts.Total..M.O | 69,7 | 75,97 | 76,8 | 78,84 | 83 | 91,1 |
| dif_Pts.Total..M.O | -4,7 | 2 | 5,35 | 3,925 | 6,45 | 12,7 |
| X2FG.M.O | 47,3 | 50 | 51,65 | 52,15 | 53,75 | 58 |
| X3FG.M.O | 32,9 | 33,98 | 35,5 | 36,56 | 37,33 | 46,6 |
| FT.M.O | 68,4 | 71,42 | 74,5 | 74,21 | 77,75 | 79,7 |
| O.TOTAL | 82 | 99,75 | 108 | 107,31 | 117 | 127 |
| D.TOTAL | 213 | 237 | 250 | 248,3 | 257,5 | 278 |
| T.TOTAL | 329 | 340 | 353 | 355,6 | 370,8 | 383 |
| As.Total | 154 | 166 | 176 | 180,2 | 191,2 | 210 |
| St.Total | 52 | 58,75 | 65,5 | 67,44 | 73,75 | 87 |
| To.Total | 105 | 129,8 | 136,5 | 136,2 | 147,2 | 155 |
| Fv.Total | 16 | 27,5 | 29,5 | 31 | 35,25 | 55 |
| Ag.Total | 18 | 24 | 25,5 | 28,31 | 30,75 | 42 |
| Cm.Total | 186 | 203 | 209,5 | 208,7 | 219,2 | 225 |
| Rv.Total | 190 | 202,5 | 213,5 | 213,7 | 219,2 | 243 |
| PIR.Total | 729 | 832,2 | 865,5 | 887,4 | 970,5 | 1101 |

Πίνακας 2.2 Βασικά περιγραφικά μέτρα για την χρονιά 2015-2016

Έτσι λοιπόν, έχοντας μια γενική εικόνα μπορούμε να παρατηρούμε ότι:

→ Οι ομάδες που προκρίθηκαν στην επόμενη φάση έχουν υψηλότερο μέσο όρο πόντων (78.84) και υψηλότερο δείκτη αξιολόγησης κατά μέσο όρο (887.4) σε σχέση

με τις ομάδες που δεν προκρίθηκαν στην επόμενη φάση της διοργάνωσης (71.39 και 725.6 αντίστοιχα). Συγκριτικά με την προηγούμενη χρονιά, αξίζει να αναφέρουμε πώς ο μέσος όρος των πόντων καθώς και ο ειδικός δείκτης αξιολόγησης για τις ομάδες που δεν προκρίθηκαν στην επόμενη φάση μειώθηκε αισθητά. Πιο συγκεκριμένα από 75.85 σε 71.39 και από 781.5 σε 725.6 αντίστοιχα. Αυτό το γεγονός αποτελεί ίσως ένα πρώτο σημάδι ότι οι ομάδες ήταν περισσότερο άστοχες σε αυτή την χρονιά ή πήραν λιγότερες προσπάθειες για σουτ.

→ Τα κλεψίματα για τις ομάδες που προκρίθηκαν ήταν κατά μέσο όρο 67.44 (αρκετά περισσότερα από την προηγούμενη χρονιά που ήταν 62.88), για το σύνολο όλων των αγώνων στην regular season, ενώ για τις ομάδες που δεν προκρίθηκαν ήταν 65.75. Έτσι λοιπόν εδώ φαίνεται ότι η ομάδα με τα περισσότερα κλεψίματα πέρασε στην επόμενη φάση. Οπότε η μεταβλητή που δηλώνει τα κλεψίματα ίσως είναι στατιστικά σημαντική και θα πρέπει να συμπεριληφθεί στο μοντέλο που θα κατασκευάσουμε αργότερα. Η διαφορά στις assist για τις ομάδες που προκρίθηκαν ήταν περίπου 13 μονάδες.

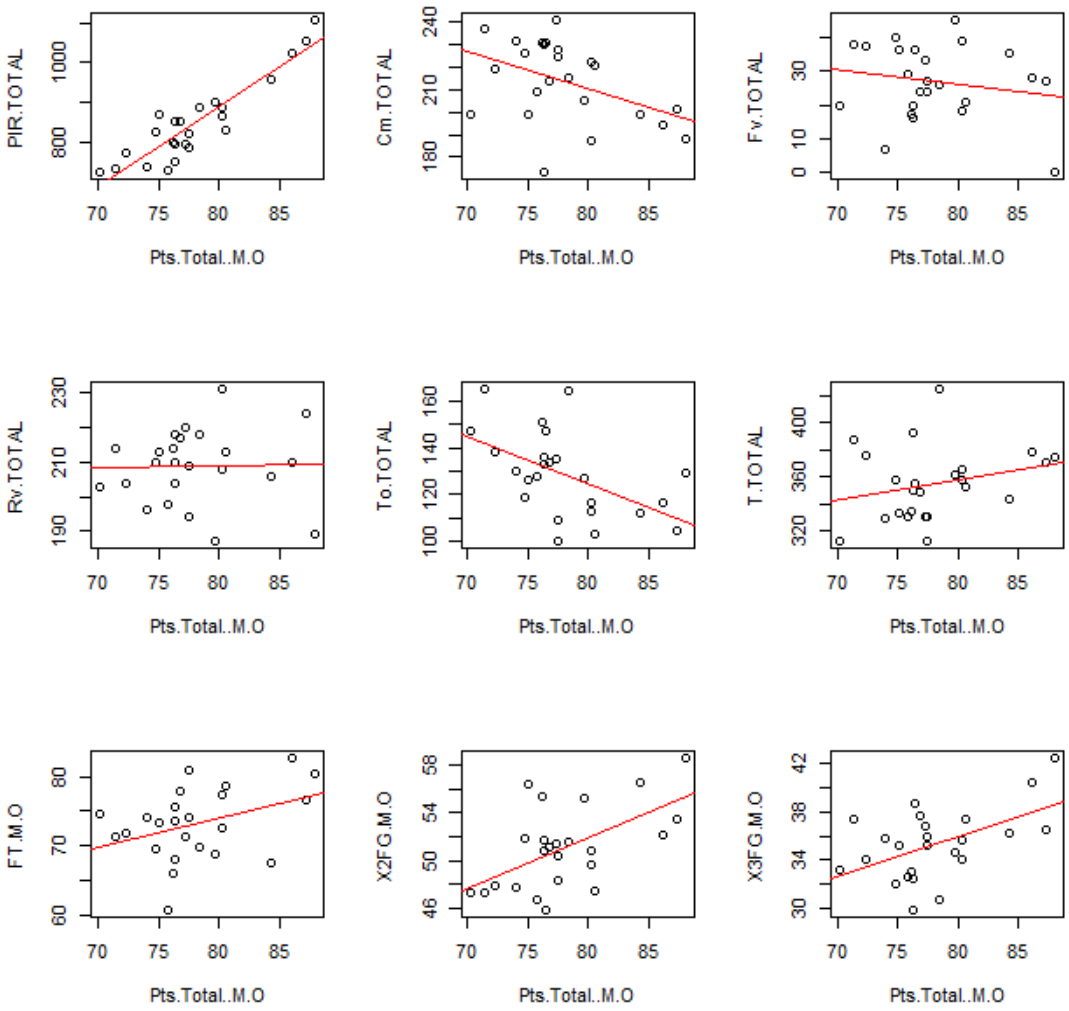
→ Ενδιαφέρον παρουσιάζουν και τα στατιστικά για τα σουτ των παιχτών. Όσον αφορά τα σουτ 2 πόντων οι ομάδες που προκρίθηκαν είχαν ποσοστό ευστοχίας ίσο με 52.15% , ελαφρώς μεγαλύτερο σε σύγκριση με την προηγούμενη χρονιά, ενώ οι 8 ομάδες που δεν προκρίθηκαν είχαν ποσοστό ευστοχίας 47.99%, δηλαδή περίπου 2 ποσοστιαίες μονάδες λιγότερες σε σχέση με την προηγούμενη χρονιά. Για τα σουτ 3 πόντων είχαμε ποσοστά ευστοχίας 36.56% και 33.85% αντίστοιχα για τις ομάδες που πέρασαν στην επόμενη φάση και για αυτές που δεν τα κατάφεραν. Σε αυτή την περίπτωση βλέπουμε ότι η διαφορά των 2 ποσοστών δεν είναι τόσο μεγάλη και ίσως και αυτή να μην επηρεάζει το τελικό αποτέλεσμα, με μία πρώτη εκτίμηση. Τέλος, αρκετά κοντά είναι και τα ποσοστά των ελεύθερων βολών για τις ομάδες.

→ Όσον αφορά τα συνολικά ριμπάουντ 355 ήταν περίπου κατά μέσο όρο για τις ομάδες που προκρίθηκαν (11 λιγότερα από την προηγούμενη χρονιά) και περίπου 321 για τις ομάδες που δεν προκρίθηκαν.

→ Ακόμη αξίζει να αναφερθεί πως οι τιμές των μέσων όρων και των διαμέσων είναι αρκετά κοντά σε όλες τις μεταβλητές, όπως και την προηγούμενη χρονιά.

### 2.5 Γραφική ανάλυση για την χρονιά 2015-2016

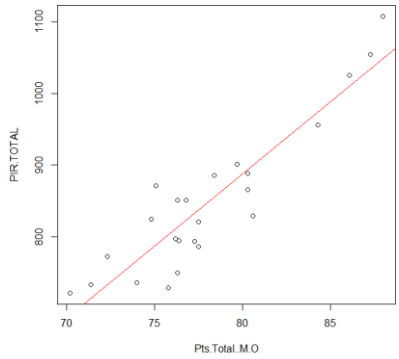
Θα αναπαραστήσουμε ενδεικτικά τα ζεύγη των συνολικών πόντων με 9 από τις υπόλοιπες μεταβλητές μας στα παρακάτω διάγραμμα.




Σχήμα 2.8 Scatterplot για 9 μεταβλητές σε σχέση με τους συνολικούς πόντους (2015-2016)

Το μόνο που μπορούμε ίσως να πούμε είναι ότι στις περισσότερες περιπτώσεις τα σημεία βρίσκονται διασκορπισμένα στο χώρο, όπως και την χρονιά 2014-2015. Εξαίρεση αποτελούν και εδώ τα συνολικά κοψίματα υπέρ της ομάδας (Fv) που φαίνεται ότι είναι αρνητικά συσχετισμένα με τους πόντους.

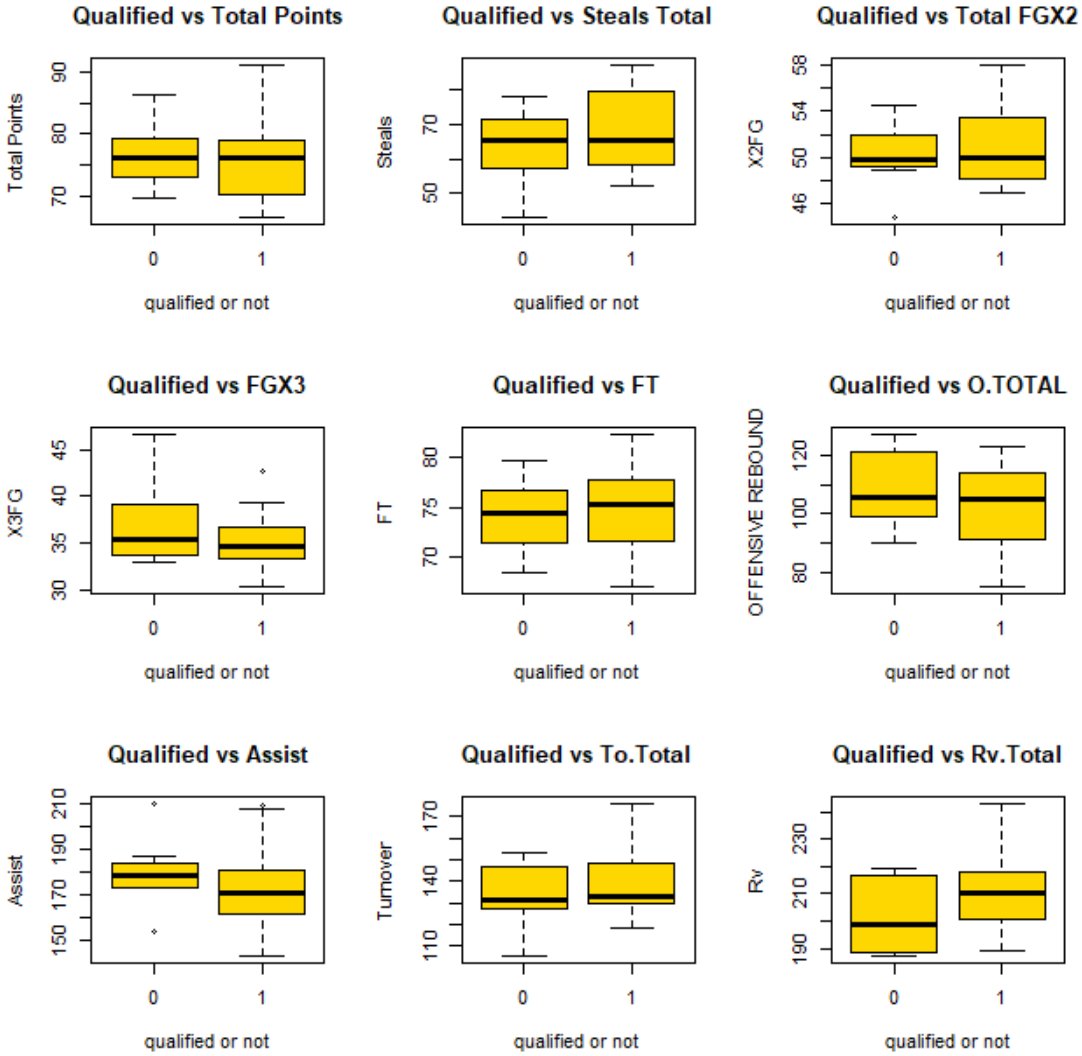
Μια πιο κατανοητή αναπαράσταση δίνεται εάν εξετάσουμε την συσχέτιση των συνολικών πόντων και του ειδικού δείκτη αξιολόγησης.




Σχήμα 2.9 Scatterplot για πόντους και PIR (2015-2016)

Από το παραπάνω σχήμα φαίνεται ότι η σχέση τους είναι γραμμική και η ευθεία παλινδρόμησης που προσαρμόστηκε έχει θετική κλίση. Ακόμη τα σημεία βρίσκονται διασκορπισμένα αρκετά κοντά στην ευθεία με αποτέλεσμα να μπορούμε να ισχυριστούμε πως η γραμμική σχέση είναι αρκετά ισχυρή. Έτσι λοιπόν οι πόντοι επηρεάζουν και αυτή την χρόνια τον δείκτη PIR, κάτι το οποίο είναι προφανές καθώς είναι μία μεταβλητή που αποτελείται από έναν συνδυασμό άλλων μεταβλητών.

Παρακάτω παρουσιάζουμε ενδεικτικά τα αντίστοιχα boxplots για 9 μεταβλητές σε σχέση με την μεταβλητή qualified :




Σχήμα 2.10 Boxplots για διάφορες μεταβλητές (2015-2016)

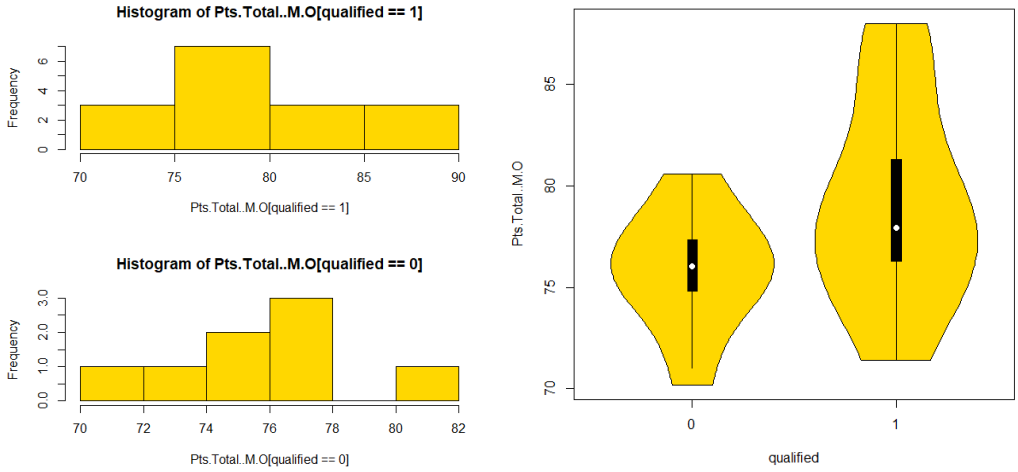
→ Βλέπουμε πώς δεν είναι σαφές ότι σε όλα τα θηκογράμματα η διάμεσος για τις ομάδες που προκρίθηκαν (qualified = 1) βρίσκεται υψηλότερα από τις ομάδες που δεν προκρίθηκαν, έστω και οριακά.

→ Φαίνεται πως παρατηρούνται δύο ακραίες τιμές όσον αφορά τις assist για τις ομάδες που δεν προκρίθηκαν και μία ελαφριά ασυμμετρία προς τα δεξιά. Επίσης, μία ακραία τιμή υπάρχει και στα σουτ τριών πόντων για τις ομάδες που προκρίθηκαν και ισούται περίπου με 42% όπως βλέπουμε στο σχήμα. Βέβαια δεν πρέπει να βγει από το dataset καθώς δεν αποτελεί extreme value. Ένα μέτρο για να το επαληθεύσουμε αυτό είναι ο υπολογισμός $\mu + 3\sigma$ (όπου μ ο μέσος όρος και σ η τυπική απόκλιση). Εάν η τιμή είναι μεγαλύτερη από το προηγούμενο αποτέλεσμα ίσως θα έπρεπε να αφαιρεθεί από το dataset. Τέλος, οι assist φαίνεται πως είναι περισσότερες για τις ομάδες που δεν προκρίθηκαν, όπως και τα επιθετικά ριμπάουντ.

→ Επιπλέον, τα σουτ 3 πόντων καθώς και τα λάθη παρουσιάζουν μεγάλη ασσυμετρία στις ομάδες που δεν προκρίθηκαν στην επόμενη φάση της διοργάνωσης.

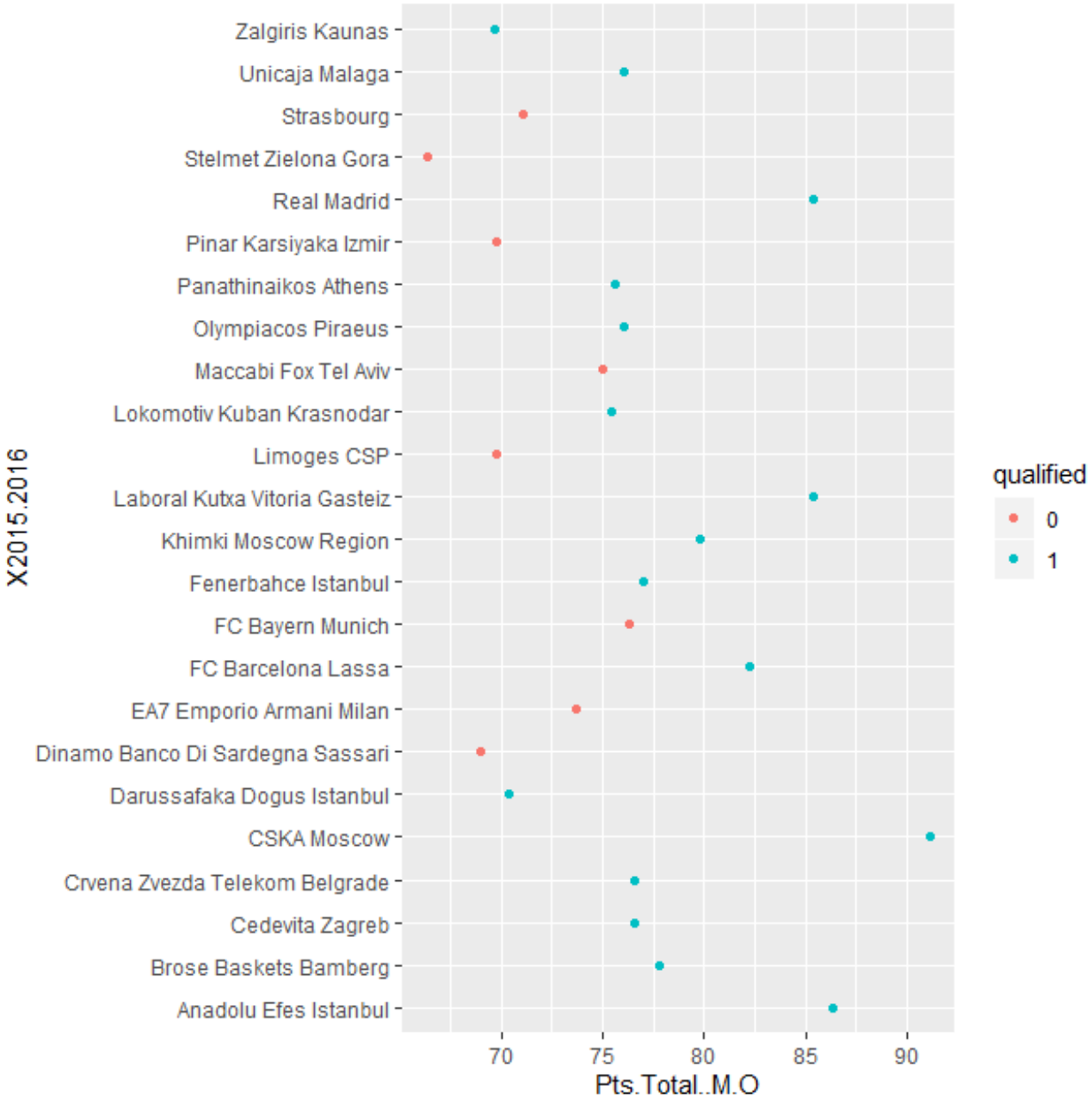
→ Δεν φαίνεται να υπάρχει κάποια έντονη ασσυμετρία, θετική ή αρνητική.

Στην συνέχεια παρουσιάζεται ένα ιστόγραμμα, χωρισμένο σε κλάσεις, που αντικατοπτρίζει το πώς κατανέμονται οι πόντοι των ομάδων. Βλέπουμε ότι για τις ομάδες που προκρίθηκαν οι περισσότερες παρατηρήσεις βρίσκονται στην κλάση [75,80). Συμπληρωματικά, το violin plot μας δίνει την επιπλέον πληροφορία για την κατανομή (συνάρτηση πυκνότητας) των μεταβλητών.




Σχήμα 2.11 Ιστόγραμμα και violin plot για το σύνολο των πόντων (2015-2016)

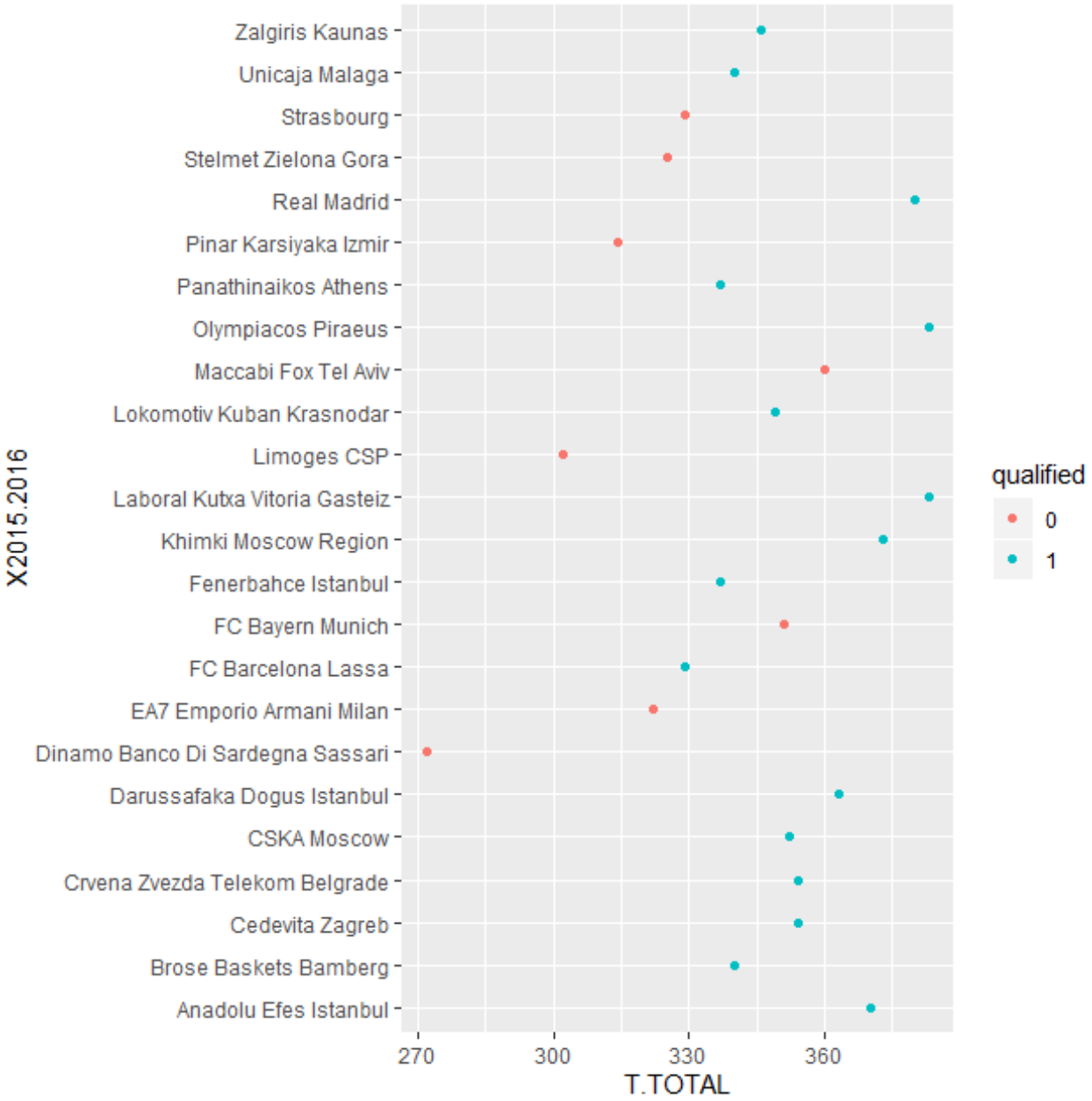
Τέλος, χρησιμοποιώντας την βιβλιοθήκη ggplot2 έχουμε ενδεικτικά τα 2 παρακάτω διαγράμματα 2.12 και 2.13 για τους πόντους κατά μέσο όρο και τα συνολικά ριμπάουντ που πέτυχαν οι 24 ομάδες στα 10 παιχνίδια που αγωνίστηκαν στην κανονική περίοδο.




Σχήμα 2.12 Συνολικοί πόντοι για κάθε μία ομάδα ξεχωριστά (2015-2016)

Είναι σαφές ότι τους περισσότερους πόντους κατά μέσο όρο στην regular season 2015-2016 σκόραρε ξανά η CSKA Moscow.Στην δεύτερη θέση έρχεται αυτή την χρονιά η τουρκική Anadolu Efes Istanbul. Αξίζει να σημειώσουμε ότι η CSKA Moscow κατέκτησε τελικά το τρόπαιο, λαμβάνοντας την πρώτη θέση στο θεσμό. Την χειρότερη απόδοση σε πόντους φαίνεται πως παρουσιάζει η Stelmet Zielona Gora με μέσο σκοράρισμα γύρω στους 65 πόντους ανά αγώνα. Σε χαμηλά ποσοστά κυμαίνεται και η Zalgiris Kaunas με περίπου 70 πόντους κατά μέσο όρο. Παρόλα αυτά βλέπουμε πως παρουσιάζεται με πράσινο χρώμα, το οποίο υποδεικνύει ότι προκρίθηκε στο top-16 της διοργάνωσης και αυτή την χρόνια κάτι το οποίο είναι εντυπωσιακό καθώς κατάφερε να

προκριθεί στην επόμενη φάση 2 συνεχόμενες χρονιές με 72 και 70 πόντους κατά μέσο όρο σε κάθε παιχνίδι.




Σχήμα 2.13 Συνολικά ριμπάουντ για κάθε μία ομάδα ξεχωριστά (2015-2016)

Όσον αφορά τα συνολικά ριμπάουντ παρατηρούμε ότι μία ομάδα για να προκριθεί στην επόμενη φάση πρέπει να είναι τουλάχιστον 330 και πάνω. Εξαίρεση αποτελεί η FC Bayern Munich με περίπου 350 και η Maccabi Fox Tel Aviv με 360. Τέλος, τα περισσότερα ριμπάουντ είχαν ο Olympiacos Piraeus και η Laboral Kutxa. Πληροφοριακά αναφέρουμε πως η Laboral τερμάτισε στην τέταρτη θέση του θεσμού.

# ΚΕΦΑΛΑΙΟ 3

## Συσχετίσεις μεταξύ των μεταβλητών

Στο παρόν κεφάλαιο θα εξετάσουμε τις σχέσεις που έχουν οι μεταβλητές μεταξύ τους. Η συγκεκριμένη ανάλυση θα γίνει μέσω των συντελεστών συσχέτισης ανά group και ανά χρονιά. Με την μέθοδο της συσχέτισης μπορούμε να μετρήσουμε τον βαθμό της αλληλεξάρτησης ανάμεσα σε δύο ή και περισσότερες μεταβλητές και να πάρουμε μια πρώτη εικόνα για τα δεδομένα μας. Επιπλέον, θα εφαρμόσουμε διαχωριστική ανάλυση με δύο groups (qualified και non-qualified) ως προς τις υπόλοιπες μεταβλητές.

### 3.1 Βασικοί συντελεστές συσχέτισης

Στην στατιστική, η εξάρτηση είναι οποιαδήποτε στατιστική σχέση μεταξύ δύο τυχαίων μεταβλητών ή δύο συνόλων δεδομένων. Η συσχέτιση αναφέρεται σε μια ευρεία κατηγορία στατιστικών σχέσεων με τη συμμετοχή της εξάρτησης, αν και σε κοινή χρήση συχνότερα αναφέρεται στο βαθμό με τον οποίο δύο μεταβλητές έχουν μια γραμμική σχέση η μία με την άλλη.

#### 3.1.1 Συντελεστής συσχέτισης του Pearson (r)

Το πιο γνωστό μέτρο της εξάρτησης μεταξύ δύο ποσοτήτων είναι ο συντελεστής συσχέτισης Pearson, ή " Pearson συντελεστής συσχέτισης ", που συνήθως ονομάζεται απλά "ο συντελεστής συσχέτισης". Είναι το πηλίκο της διαίρεσης της συνδιακύμανσης των δύο μεταβλητών με το γινόμενο των τυπικών αποκλίσεων.

Ο γνωστός συντελεστής συσχέτισης $P_{X,Y}$ μεταξύ δύο τυχαίων μεταβλητών $X$ και $Y$ με τις αναμενόμενες τιμές $\mu_X$ και $\mu_Y$ και τυπική απόκλιση $\sigma_X$ και $\sigma_Y$ ορίζεται ως:

$$P_{X,Y} = \text{corr}\,(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

όπου $E$ είναι η αναμενόμενη τιμή, *cov* σημαίνει συνδιακύμανση, και *corr* είναι μια ευρέως χρησιμοποιούμενη εναλλακτική συντομογραφία για το συντελεστή συσχέτισης.

Ο συντελεστής συσχέτισης Pearson ορίζεται μόνο αν και οι δύο τυπικές αποκλίσεις είναι πεπερασμένες και μη μηδενικές. Είναι απόρροια της Caushy-Schwarz ανισότητας ότι η συσχέτιση δεν μπορεί να υπερβαίνει το 1, σε απόλυτη τιμή. Ο συντελεστής συσχέτισης είναι συμμετρικός: $\text{corr}(X,Y) = \text{corr}(Y,X)$.

Η συσχέτιση Pearson είναι +1 σε περίπτωση μίας τέλειας (αύξουσας) γραμμικής σχέσης (θετική συσχέτιση), -1 σε περίπτωση μίας τέλειας φθίνουσας (αντίστροφης) γραμμικής σχέσης (αρνητική συσχέτιση), και κάποια τιμή μεταξύ -1 και 1 σε όλες τις άλλες περιπτώσεις, που δείχνει το βαθμό της γραμμικής εξάρτησης μεταξύ των μεταβλητών. Καθώς πλησιάζει το μηδέν υπάρχει λιγότερη σχέση (πιο κοντά σε ασυσχέτιστα). Όσο πιο κοντά είναι ο συντελεστής είτε στο -1 ή στο 1, τόσο ισχυρότερη είναι η συσχέτιση μεταξύ των μεταβλητών.

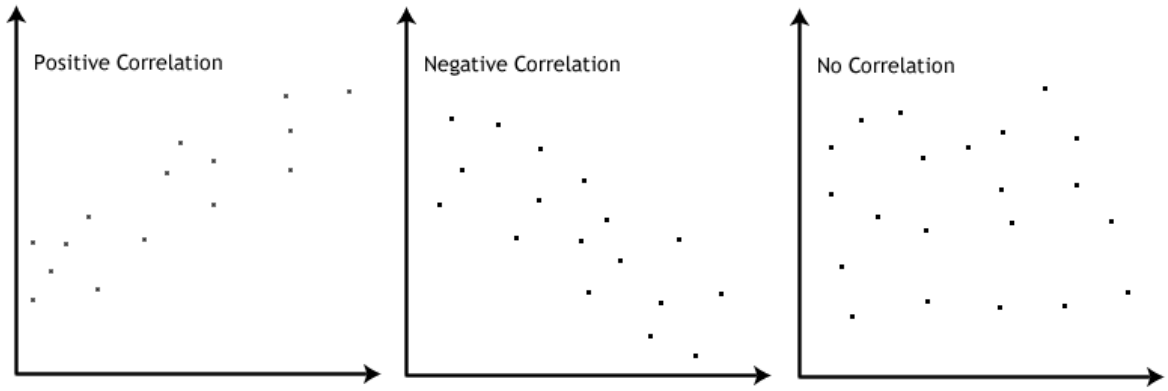
Αν οι μεταβλητές είναι ανεξάρτητες, ο συντελεστής συσχέτισης Pearson είναι 0, αλλά το αντίστροφο δεν είναι αληθές, διότι ο συντελεστής συσχέτισης ανιχνεύει μόνο γραμμική εξάρτηση μεταξύ των δύο μεταβλητών. Για παράδειγμα, ας υποθέσουμε ότι η τυχαία μεταβλητή $X$ είναι συμμετρικά κατανεμημένη στο μηδέν, και $Y = X^2$. Τότε

το *Y* καθορίζεται εντελώς από το *X*, ώστε *οι X* και *Y* είναι απόλυτα εξαρτημένες, αλλά η συσχέτιση είναι μηδενική (είναι ασυσχέτιστες). Ωστόσο, στην ειδική περίπτωση, όταν *X* και *Y* είναι από κοινού κανονικές, το ότι δε συσχετίζονται είναι ισοδύναμο με την ανεξαρτησία.

Αν έχουμε μια σειρά από *n* μετρήσεις των *X* και *Y* γραμμένες ως $x_i$ και $y_i$ για $i = 1, 2, ..., n$, τότε ο δειγματικός συντελεστής συσχέτισης μπορεί να χρησιμοποιηθεί για την εκτίμηση του πληθυσμιακού συντελεστή συσχέτισης Pearson *r* μεταξύ *X* και *Y*. Ο δειγματικός συντελεστής συσχέτισης γράφεται :

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

Παρακάτω μπορούμε να πάρουμε μια πρώτη εικόνα για την διαγραμματική απεικόνιση του συντελεστή συσχέτισης r , όταν υπάρχει θετική, αρνητική ή και καμία συσχέτιση αντίστοιχα :




Σχήμα 3.1 : Διάγραμμα Διασποράς απεικόνισης συσχετίσεων

(Πηγή: http://www.statstutor.ac.uk/resources/uploaded/pearsons.pdf )

Χρησιμοποιείται για τον έλεγχο της σχέσης μεταξύ 2 μεταβλητών όταν :

→ ακολουθούν και οι 2 την κανονική κατανομή.

### 3.1.2 Συντελεστής συσχέτισης του Spearman (ρ)

Ο συντελεστής συσχέτισης Spearman, που πήρε το όνομά του από τον Charles Spearman και συχνά συμβολίζεται με το ελληνικό γράμμα ρ ή ως $r_s$, είναι ένα μη-παραμετρικό μέτρο της στατιστικής εξάρτησης μεταξύ δύο μεταβλητών. Αξιολογεί το πόσο καλά μπορεί να περιγραφεί η σχέση μεταξύ των δύο μεταβλητών χρησιμοποιώντας μια μονότονη συνάρτηση. Εάν δεν υπάρχουν επαναλαμβανόμενες τιμές των δεδομένων, μια τέλεια συσχέτιση Spearman κατά +1 ή -1 εμφανίζεται όταν κάθε μία από τις μεταβλητές είναι μια τέλεια μονότονη συνάρτηση της άλλης. Ο συντελεστής Spearman είναι κατάλληλος και για συνεχείς και για διακριτές μεταβλητές, συμπεριλαμβανομένων των διατακτικών διακριτών μεταβλητών. Ορίζεται με παρόμοιο τρόπο όπως και ο συντελεστής συσχέτισης Pearson. Το πρόσημο της συσχέτισης Spearman δείχνει την κατεύθυνση της σχέσης μεταξύ της *X* (ανεξάρτητη μεταβλητή) και της *Y* (εξαρτημένη μεταβλητή). Εάν

η *Y* τείνει να αυξάνεται όταν η *X* αυξάνει, ο συντελεστής συσχέτισης Spearman είναι θετικός.

Εάν η *Y* τείνει να μειώνεται όταν η *X* αυξάνει, ο συντελεστής συσχέτισης Spearman είναι αρνητικός. Μια μηδενική συσχέτιση Spearman δείχνει ότι δεν υπάρχει τάση για την *Y* είτε να αυξηθεί ή να μειωθεί, όταν η *X* αυξάνει. Η συσχέτιση Spearman αυξάνει σε μέγεθος, όταν η *X* και η *Y* είναι πιο κοντά στο να είναι τέλειες μονότονες συναρτήσεις η μία της άλλης.. Όταν η *X* και η *Y* έχουν απόλυτη μονοτονική σχέση, ο συντελεστής συσχέτισης Spearman γίνεται 1.

Ο συντελεστής συσχέτισης Spearman συχνά περιγράφεται ως "μη παραμετρικός." Αυτό μπορεί να έχει δύο έννοιες. Πρώτον, το γεγονός ότι μια τέλεια συσχέτιση Spearman προκύπτει όταν *X* και *Y* σχετίζονται με οποιαδήποτε μονότονη συνάρτηση, που μπορεί να αντιπαραβληθεί με τη συσχέτιση Pearson, η οποία δίνει μόνο μια τέλεια τιμή όταν *X* και *Y* σχετίζονται με μια γραμμική συνάρτηση. Η άλλη έννοια με την οποία ο συντελεστής Spearman είναι μη παραμετρικός είναι ότι η ακριβής κατανομή της δειγματοληψίας του μπορεί να ληφθεί χωρίς να απαιτείται γνώση της από κοινού κατανομής πιθανότητας της *X* και *Y*.

Όπως και για τον συντελεστή του Pearson, ο συντελεστής του Spearman δίνεται από τον τύπο :

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

(Πηγή : https://el.wikipedia.org/ )

Χρησιμοποιείται για τον έλεγχο της σχέσης μεταξύ 2 μεταβλητών όταν :

→ η μία είναι συνεχής με κανονική κατανομή και η άλλη δεν ακολουθεί την κανονική κατανομή ή είναι κατηγορική

→ δεν ακολουθούν και οι 2 την κανονική κατανομή

### 3.1.3 Συντελεστής συσχέτισης του Kendall (τ)

Ο συντελεστής συσχέτισης του Kendall, γνωστός και ως συντελεστής συμφωνίας του Kendall, μοιάζει με τον συντελεστή ρ του Spearman ως προς το ότι υπολογίζεται με βάση την τάξη (rank) μεγέθους των παρατηρήσεων και όχι με βάση τις παρατηρήσεις αυτές καθεαυτές. Επιπλέον, η κατανομή του δεν εξαρτάται από την κατανομή των μεταβλητών X και Y, όταν αυτές είναι ανεξάρτητες και συνεχείς. Το κύριο πλεονέκτημα του μέτρου αυτού σε σχέση με το μέτρο ρ του Spearman είναι ότι τείνει πιο γρήγορα στην κανονική κατανομή .(Πηγή : Croux, C. and Dehon, C. (2010))

Ένα άλλο πλεονέκτημα βρίσκεται στο γεγονός ότι μπορεί άμεσα να ερμηνευθεί μέσω των πιθανοτήτων με τις οποίες παρατηρούμε εναρμονισμένα (concordant) ζεύγη τιμών και μη εναρμονισμένα (discordant) ζεύγη τιμών. Σημειώνουμε πως δύο παρατηρήσεις ονομάζονται εναρμονισμένες, αν και τα δύο μέλη της μίας παρατήρησης είναι

μεγαλύτερα (ή μικρότερα) από τα αντίστοιχα μέλη της άλλης παρατήρησης. Αν η διάταξη των πρώτων μελών τους είναι αντίθετη από την διάταξη των δεύτερων μελών τους οι παρατηρήσεις ονομάζονται μη εναρμονισμένες. Τα ζεύγη των παρατηρήσεων ($X_i$ , $Y_j$) και ($X_k$ , $Y_{k)}$ για τα οποία ισχύει $X_j = X_k$ ή/και $Y_j = Y_k$ , δεν είναι ούτε εναρμονισμένα ούτε μη εναρμονισμένα και ονομάζονται ισοβαθμούντα (tied).

Έτσι λοιπόν ο συντελεστής συσχέτισης του Kendall ανάμεσα σε 2 τυχαίες μεταβλητές με n παρατηρήσεις ορίζεται ως εξής :

$$\tau = \frac{(\#concordant\ pairs) - (\#discordant\ pairs)}{n(n-1)/2}$$

Χρησιμοποιείται για τον έλεγχο της σχέσης μεταξύ 2 μεταβλητών όταν :

→ είναι κατηγορικές διατάξιμες

### 3.2 Σχέσεις των μεταβλητών

Στην περίπτωση μας θα υπολογίσουμε, σύμφωνα με τον συντελεστή συσχέτισης tau (τ) του Kendall, αν υπάρχει ισχυρή εξάρτηση ανάμεσα στην βαθμολογία κατάταξης στο group (1,2,3,4,5,6) και στα χαρακτηριστικά που επηρεάζουν ή όχι το τελικό αποτέλεσμα για κάθε μια ομάδα ξεχωριστά. Επιλέγουμε τον συγκεκριμένο συντελεστή καθώς αναζητούμε την σχέση μιας διατάξιμης και μιας ποσοτικής μεταβλητής. Πιο συγκεκριμένα, θα υπολογίσουμε τους συντελεστές συσχέτισης για όλα τα group (A,B,C,D) ανά χρονιά (2014-2015 και 2015-2016). Οι τιμές αυτών των συντελεστών, όπως είπαμε και παραπάνω, θα βρίσκονται ανάμεσα στο -1 και στο +1. Όσο μεγαλύτερες κατ' απόλυτη τιμή είναι οι τιμές των συντελεστών, τόσο πιο ισχυρή θετική ή αντίστοιχα αρνητική συσχέτιση έχουμε για τις μεταβλητές που μας απασχολούν. Αν η τιμή ίση ή πολύ κοντά στο 0, υποδηλώνει ότι δεν υπάρχει καμία σχέση ανάμεσα στις μεταβλητές, ενώ μια τιμή ίση ή πολύ κοντά στο -1 ή το +1, υποδηλώνει ότι η μια μεταβλητή μπορεί να προβλέψει με υψηλή ακρίβεια την άλλη. Το πρόσημο μας δίνει πληροφορίες για το εάν μια αύξηση της μίας μεταβλητής, οδηγεί σε αύξηση ή μείωση της άλλης. Ο βαθμός έντασης της σχέσης που έχουν μεταξύ τους είναι υποκειμενικός, αλλά σε γενικές γραμμές ισχύει ο παρακάτω κανόνας :

$|r| < 0.1$ → καμία ή αμυδρή σχέση

$0.1 \leq |r| < 0.3$ → αδύναμη σχέση

$0.3 \leq |r| < 0.6$ → σχέση μέτριας έντασης

$|r| \geq 0.6$ → ισχυρή σχέση

Για την σχέση μιας δίτιμης (qualified/not-qualified) και μιας διατάξιμης μεταβλητής προτείνεται η χρήση του συντελεστή rank biserial που είναι μια ειδική περίπτωση του Somers'D.

Ο Somers το 1962 καθιέρωσε τον συντελεστή D, ένα μέτρο για την πρόβλεψη της εξαρτημένης μεταβλητής γνωρίζοντας την ανεξάρτητη. Είναι παραλλαγή του μέτρου Gamma, λαμβάνοντας υπόψη τυχόν ισοβαθμίες (ties) των ζευγών των παρατηρήσεων. Λαμβάνει τιμές στο διάστημα [-1, 1], με D = 1 όταν δεν υπάρχουν «ασύμφωνα» ζεύγη μεταξύ των μεταβλητών X και Y, και η διάταξη των X είναι σε πλήρη συμφωνία με τη διάταξη των τιμών της Y (ισχυρή συσχέτιση μεταξύ των δύο μεταβλητών), ενώ D = -1 όταν δεν υπάρχουν «σύμφωνα» ζεύγη μεταξύ των μεταβλητών X και Y, και η διάταξη των τιμών της X είναι σε πλήρη ασυμφωνία με τη διάταξη των τιμών της Y.

Τέλος, θα επιβεβαιώσουμε την σημαντικότητα της σχέσης 2 ποιοτικών μεταβλητών με έναν chi-square ( $\chi^2$ ) έλεγχο και θα γίνει διαχωριστική ανάλυση με δύο groups(qualified και not-qualified) ως προς τις άλλες μεταβλητές.

### 3.2.1 Σχέσεις των μεταβλητών ανά χρονιά (2014-2015)

Αρχικά, θα υπολογιστούν όλες οι ανά 2 συσχετίσεις των μεταβλητών μεταξύ τους, τόσο των ποσοτικών όσο και των ποιοτικών. Τα πλήρη αποτελέσματα της ανάλυσης είναι διαθέσιμα στο επισυναπτόμενο Παράρτημα.

Συνολικά για την χρονιά 2014-2015 προκύπτουν τα παρακάτω αποτελέσματα :

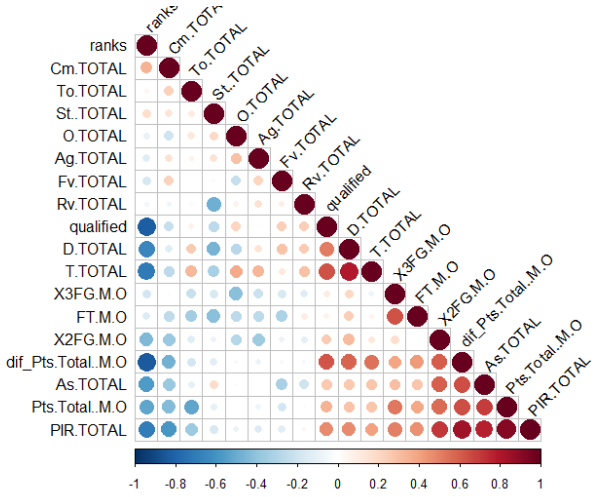
- ➔ Η μεταβλητή ranks είναι έντονα αρνητικά συσχετισμένη κατά φθίνουσα σειρά με τις μεταβλητές Pts.Total(-0.75), T.Total(-0.61) και qualified(-0.73). Η τιμή -0.73 προκύπτει λόγω των δεσμών (ties). Η παραπάνω διαπίστωση θα λέγαμε πως είναι προφανής καθώς οι ομάδες που προκρίνονται από κάθε group(που αποτελείται από 6 ομάδες) είναι οι 4 πρώτες στην αντίστοιχη κατάταξη.
- ➔ Το αν προκρίθηκε μια ομάδα ή όχι φαίνεται να είναι έντονα αρνητικά συσχετισμένο με τα συνολικά ριμπάουντ της κάθε ομάδας(T.Total).
- ➔ Οι συνολικοί πόντοι φαίνεται ότι είναι έντονα θετικά συσχετισμένοι με τον δείκτη αξιολόγησης (PIR) της Euroleague ,αφού ο αντίστοιχος συντελεστής συσχέτισης είναι 66%.
- ➔ Επιπλέον, η διαφορά των συνολικών πόντων και το ποσοστό ευστοχίας στα δίποντα είναι έντονα θετικά συσχετισμένα με τον δείκτη αξιολόγησης, με συντελεστές συσχέτισης 0.60 και 0.55 αντίστοιχα. Επίσης, έντονη αρνητική συσχέτιση παρατηρείται μεταξύ της συνολικής διαφοράς των πόντων και της κατάταξης των ομάδων μέσα στο group. (-0.75).
- ➔ Αξίζει να αναφέρουμε πως το ποσοστό ευστοχίας για τα σουτ 3 πόντων, το ποσοστό ευστοχίας ελευθέρων βολών, τα συνολικά επιθετικά ριμπάουντ, οι ασιστ, τα κλεψίματα, τα λάθη, τα κοψίματα κατά της ομάδας και τα φάουλ δεν φαίνεται να σχετίζονται έντονα θετικά ή αρνητικά με κάποια από τις υπόλοιπες μεταβλητές. Αυτό βέβαια δεν σημαίνει ότι δεν υπάρχει σχέση μεταξύ τους, απλά αυτή η σχέση δεν είναι δεν είναι τόσο έντονη όσο οι προηγούμενες που αναφέρθηκαν. ( $|r| < 0.5$ ).

Στη συνέχεια παρουσιάζουμε ενδεικτικά έναν πίνακα με τις μεταβλητές που έχουν την σημαντικότερη επιρροή, σύμφωνα με τον συντελεστή συσχέτισης του Kendall :

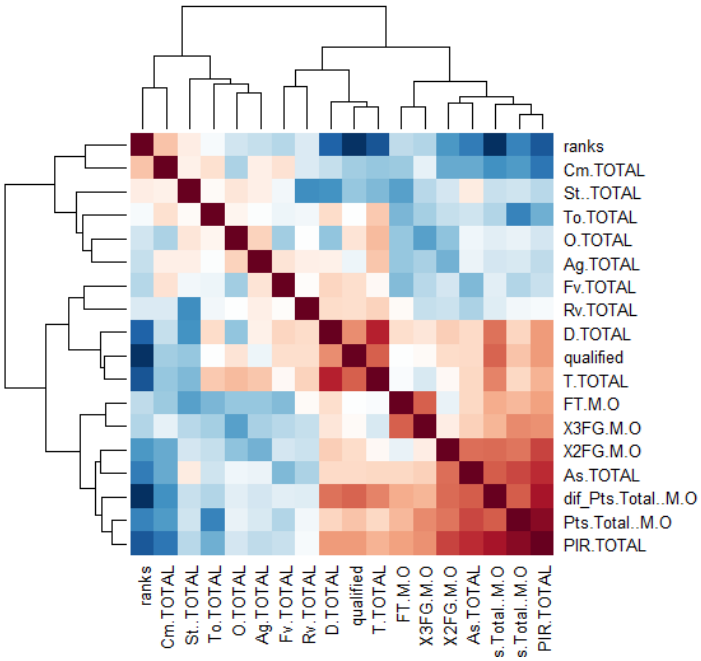
| | Ranks | Qualified | Pts | Dif.Pts | X2FG | D | T | PIR |
|---|---|---|---|---|---|---|---|---|
| Ranks | 1.00 | **-0.73** | -0.33 | **-0.75** | -0.39 | -0.50 | **-0.61** | -0.54 |
| Qualified | **-0.73** | 1.00 | 0.27 | 0.49 | 0.23 | 0.42 | **-0.58** | 0.42 |
| Pts | -0.33 | 0.27 | 1.00 | 0.33 | 0.31 | 0.17 | 0.21 | **0.66** |
| Dif.Pts | **-0.75** | 0.49 | 0.33 | 1.00 | 0.50 | 0.43 | 0.42 | **0.60** |
| X2FG | -0.39 | 0.23 | 0.31 | 0.50 | 1.00 | 0.29 | 0.12 | 0.55 |
| D | -0.50 | 0.42 | 0.17 | 0.43 | 0.29 | 1.00 | **0.64** | 0.39 |
| T | **-0.61** | 0.58 | 0.21 | 0.42 | 0.12 | **0.64** | 1.00 | 0.36 |
| PIR | -0.54 | 0.42 | **0.66** | **0.60** | 0.55 | 0.39 | 0.36 | 1.00 |

Πίνακας 3.1: Συντελεστές συσχέτισης για 2014-2015

Η R μας δίνει την δυνατότητα να απεικονίσουμε και γραφικά τα παραπάνω αποτελέσματα για τους συντελεστές συσχέτισης με αρκετές επιλογές. Εμείς θα παρουσιάσουμε το correlogram, το οποίο είναι χρήσιμο ώστε να φανούν ποιες μεταβλητές είναι συσχετισμένες από τα δεδομένα που έχουμε και το heatmap. Το correlogram μπορεί να παρουσιαστεί σε 3 μορφές : full, upper και lower. Εμείς θα παρουσιάσουμε το κάτω τρίγωνο του correlation matrix, καθώς οι τιμές των συσχετίσεων είναι συμμετρικές ως προς τη διαγώνιο από το άνω δεξιό μέχρι και το κάτω αριστερό κελί.




Σχήμα 3.2: Correlogram για 2014-2015




Σχήμα 3.3: Heatmap για 2014-2015

Αυτή η τεχνική οπτικοποίησης δεδομένων μας δείχνει το μέγεθος των μεταβλητών ως χρώμα σε 2 διαστάσεις. Εδώ με έντονο κόκκινο χρώμα παρουσιάζονται οι θετικά συσχετισμένες μεταβλητές και με μπλε οι αρνητικά συσχετισμένες μεταβλητές για την χρονιά 2014-2015 (όπως φαίνεται και στο υπόμνημα κάτω από το γράφημα οι τιμές κυμαίνονται από -1 έως +1 ). Ακόμη παρατηρούμε ότι οι ισχυρότερες θετικές συσχετίσεις συγκεντρώνονται κυρίως στην κάτω δεξιά γωνία του heatmap. Όσο πιο ανοιχτός είναι ο χρωματισμός τόσο ποιο αδύναμη είναι και η σχέση των 2 μεταβλητών. Η παραπάνω οπτικοποίηση επιβεβαιώνει τον πίνακα των συντελεστών συσχέτισης του Kendall.

Στη συνέχεια θα πρέπει να ελέγξουμε αν οι παραπάνω μεταβλητές είναι στατιστικά σημαντικές συγκριτικά με την μεταβλητή qualified. Η στατιστική σημαντικότητα είναι η δήλωση της πιθανότητας να προκύψει ένας συγκεκριμένος συντελεστής συσχέτισης ή μία πιο ακραία τιμή αυτού για ένα δείγμα δεδομένων αν δεν υπάρχει συσχέτιση στον πληθυσμό από τον λήφθηκε το δείγμα. Τα πιο συνήθη επίπεδα σημαντικότητας που χρησιμοποιούμε είναι 10%, 5% και 1%. Για το σκοπό αυτό θα δημιουργήσουμε έναν πίνακα συνάφειας 2x2, αποτελούμενο από τις συχνότητες για τις 24 ομάδες.

**Pts vs qualified**

Σε αυτό το σημείο θα θεωρήσουμε ως επιτυχία αν ο μέσος όρος των πόντων που πέτυχε η κάθε ομάδα είναι μεγαλύτερος από 80 και ως αποτυχία αν ο μέσος όρος είναι μικρότερος ή ίσος από 80. Έτσι λοιπόν δημιουργούμε μια δίτιμη μεταβλητή με τιμές 1 (επιτυχία) και 0 (όχι επιτυχία) αντίστοιχα.

Έχουμε ότι :

$H_0$ : οι μεταβλητές points και qualified είναι ασυσχέτιστες

έναντι της

$H_1$ : οι μεταβλητές points και qualified είναι συσχετισμένες.

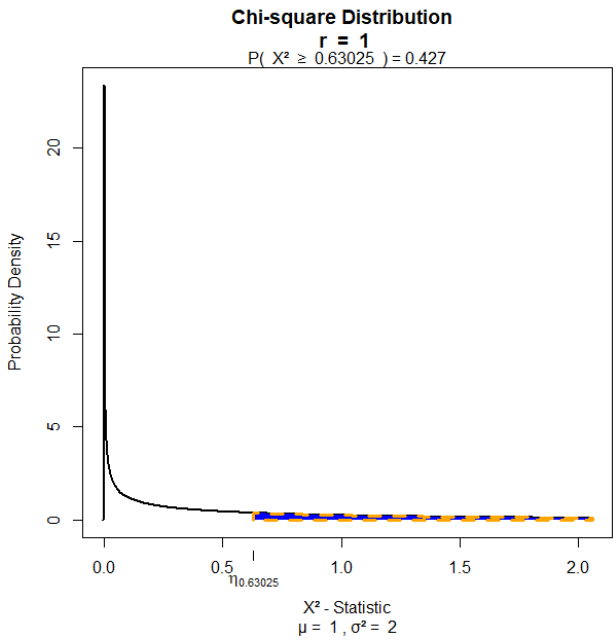
Σε αυτή την περίπτωση ο πίνακας συνάφειας θα είναι ο παρακάτω :

| | | Non-qualified | Qualified | Total |
|---|---|---|---|---|
| **Points** | ≤80 | 7 | 10 | 17 |
| | >80 | 1 | 6 | 7 |
| **Total** | | 8 | 16 | 24 |

Πίνακας 3.2: Πίνακας συνάφειας points-qualified για 2014-2015

Πραγματοποιώντας τον έλεγχο χ-τετράγωνο η R μας δίνει τα παρακάτω αποτελέσματα :

| **Pearson's Chi-squared test with Yates' continuity correction** |
|---|
| **data: tb1** |
| **X-squared = 0.63025, df = 1, p-value = 0.4273** |




Σχήμα 3.4: Chi-square Distribution points-qualified για 2014-2015

Παρατηρούμε πως η τιμή της στατιστικής συνάρτησης ισούται με 0.63025 και το αντίστοιχο p-value είναι ίσο με 0.4273 > 0.05 (όπως φαίνεται και από την γραφική απεικόνιση). Συνεπώς σε επίπεδο στατιστικής σημαντικότητας 5% δεν μπορούμε να απορρίψουμε την μηδενική μας υπόθεση, κάτι το οποίο σημαίνει πως το αν θα προκριθεί μια ομάδα ή όχι στην επόμενη φάση δεν σχετίζεται με το αν θα έχει περισσότερους από 80 πόντους κατά μέσο όρο.

Για τις υπόλοιπες μεταβλητές θα θεωρήσουμε ότι υπάρχουν επιτυχίες με βάση τα παρακάτω κριτήρια :

Dif_points > |5|, 2FG > 50%, 3FG > 35 %, FT > 75%, O > 110, D >250, T > 360, As > 170,St >70,To >130, Fv >30, Ag >30, Cm >210, Rv >210, PIR > 800

Έτσι λοιπόν συγκεντρωτικά προκύπτει ο παρακάτω πίνακας με τις τιμές των p-values μετά τον έλεγχο χ-τετράγωνο για κάθε δίτιμη μεταβλητή σε σύγκριση με την μεταβλητή qualified :

| **Μεταβλητές** | **p-value** |
|---|---|
| Dif_points vs qualified | 0.4689388 |
| 2FG vs qualiied | 0.6547208 |
| 3FG vs qualified | 1 |
| FT vs qualified | 0.8783249 |
| O vs qualified | 0.8848361 |
| D vs qualified | 0.05970881 |
| **T vs qualified** | **0.02534732*** |
| As vs qualified | 0.1797125 |
| St vs qualified | 0.2663799 |
| To vs qualified | 1 |
| Fv vs qualified | 0.6547208 |
| Ag vs qualified | 1 |
| Cm vs qualified | 1 |
| Rv vs qualified | 1 |

| | |
|---|---|
| PIR vs qualified | 0.3055071 |

Πίνακας 3.3: Στατιστικά σημαντικές μεταβλητές ως προς την πρόκριση στα play-offs για 2014-2015

Παρατηρούμε πως σύμφωνα με τα κριτήρια που έχουμε ορίσει για το έτος 2014-2015 αν τα συνολικά rebound είναι περισσότερα από 360 τότε υπάρχει ισχυρή ένδειξη ότι η ομάδα θα προκριθεί στο top-16 (p-value = 0.025). Οι υπόλοιπες δίτιμες μεταβλητές δεν φαίνεται να σχετίζονται στατιστικά σημαντικά με την μεταβλητή qualified (p-value > 5%). Οριακή φαίνεται πως ίσως και τα αμυντικά ριμπάουντ να σχετίζονται στατιστικά σημαντικά με την μεταβλητή qualified (p - value = 0.059)

Συγκεντρωτικά όλα τα αποτελέσματα παρουσιάζονται στο επισυναπτόμενο παράρτημα.

Στην περίπτωση μας η τιμή του συντελεστή Somer's D που ισούται με -0.53 υποδηλώνει μια μέτρια αρνητική συσχέτιση μεταξύ των ποιοτικών μεταβλητών qualified και κατάταξης στο group (ranks) . Αυτή η διαπίστωση μπορεί να επιβεβαιωθεί και με τον αντίστοιχο έλεγχο $\chi^2$.

### 3.2.2 Σχέσεις των μεταβλητών ανά χρονιά (2015-2016)

Συνολικά για την χρονιά 2015-2016 προκύπτουν τα παρακάτω αποτελέσματα :

- ➔ Η κατάταξη των ομάδων είναι έντονα αρνητικά συσχετισμένη κατά φθίνουσα σειρά με τους συνολικούς πόντους (-0.74), με το αν η ομάδα προκρίθηκε ή όχι (-0.73) καθώς και με τον ειδικό δείκτη αξιολόγησης (-0.58). Αντίστοιχα η μεταβλητή qualified είναι ισχυρά θετικά συσχετισμένη με τους συνολικούς πόντους, το ποσοστό ευστοχίας στα σουτ 2 πόντων (0.60) και τον δείκτη PIR (0.60).
- ➔ Ακόμη φαίνεται πως οι συνολικοί πόντοι κάθε ομάδας στους αγώνες που έδωσε κατά την διάρκεια της χρονιάς, όπως και η συνολική διαφορά των πόντων, είναι πολύ έντονα θετικά συσχετισμένοι με τον δείκτη αξιολόγησης ( r =0.80 και r=0.70 αντίστοιχα). Επιπλέον η μεταβλητή Dif.Pts παρουσιάζει έντονη αρνητική συσχέτιση με την μεταβλητή ranks (-0.74) και ισχυρή θετική συσχέτιση με τη μεταβλητή qualified (0.63).
- ➔ Είναι προφανές πώς τα αμυντικά και επιθετικά ριμπάουντ επηρεάζουν τα συνολικά ($r_{D,T}$ = 0.66 και $r_{O,T}$ =0.50). Αυτό που αξίζει να σημειώσουμε είναι ότι τα αμυντικά ριμπάουντ σχετίζονται σε μεγάλο βαθμό με την συνολική διαφορά των πόντων (0.54).
- ➔ Η μεταβλητή Performance Index Rating (PIR) παρουσιάζει έντονη θετική συσχέτιση κατά φθίνουσα σειρά με τους συνολικούς πόντους (0.80), την διαφορά των συνολικών πόντων (0.70) και τα επιθετικά φάουλ (0.60).
- ➔ Τέλος, το ποσοστό ευστοχίας στα τρίποντα, το ποσοστό ευστοχίας στις ελεύθερες βολές, οι ασιστ, τα κλεψίματα, τα λάθη και τα κοψίματα δεν σχετίζονται σε μεγάλο βαθμό με κάποια από τις υπόλοιπες μεταβλητές.

Εξαίρεση αποτελεί η μέτριας έντασης σχέση που έχουν οι μεταβλητές total points και PIR με την μεταβλητή blocks in favor.

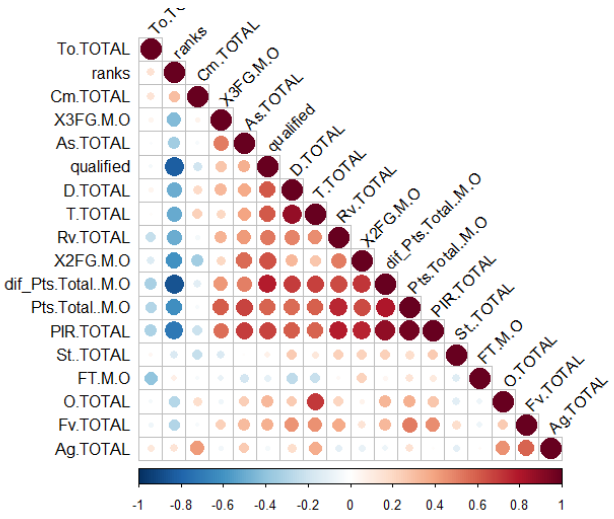
Στη συνέχεια θα παρουσιάσουμε έναν πίνακα με τις μεταβλητές που έχουν την σημαντικότερη επιρροή, σύμφωνα με τον συντελεστή συσχέτισης του Kendall, όπως έγινε και παραπάνω για να εντοπίσουμε αν υπάρχει κάποια σημαντική διαφορά ανάμεσα στους συντελεστές για τα 2 έτη :

| | Ranks | Qualified | Pts | Dif.Pts | X2FG | D | T | PIR |
|---|---|---|---|---|---|---|---|---|
| Ranks | 1.00 | **-0.73** | -0.50 | **-0.74** | -0.50 | -0.37 | -0.33 | -0.58 |
| Qualified | **-0.73** | 1.00 | 0.53 | **0.63** | **0.60** | 0.50 | 0.50 | **0.60** |
| Pts | -0.50 | 0.53 | 1.00 | 0.56 | 0.47 | 0.39 | 0.43 | **0.80** |
| Dif.Pts | **-0.74** | **0.63** | 0.56 | 1.00 | 0.48 | 0.54 | 0.44 | **0.70** |
| X2FG | -0.50 | **0.60** | 0.47 | 0.48 | 1.00 | 0.27 | 0.19 | 0.54 |
| D | -0.37 | 0.50 | 0.39 | 0.54 | 0.27 | 1.00 | **0.66** | 0.45 |
| T | -0.33 | 0.50 | 0.43 | 0.44 | 0.19 | **0.66** | 1.00 | 0.37 |
| PIR | -0.58 | **0.60** | **0.80** | **0.70** | 0.54 | 0.45 | 0.37 | 1.00 |

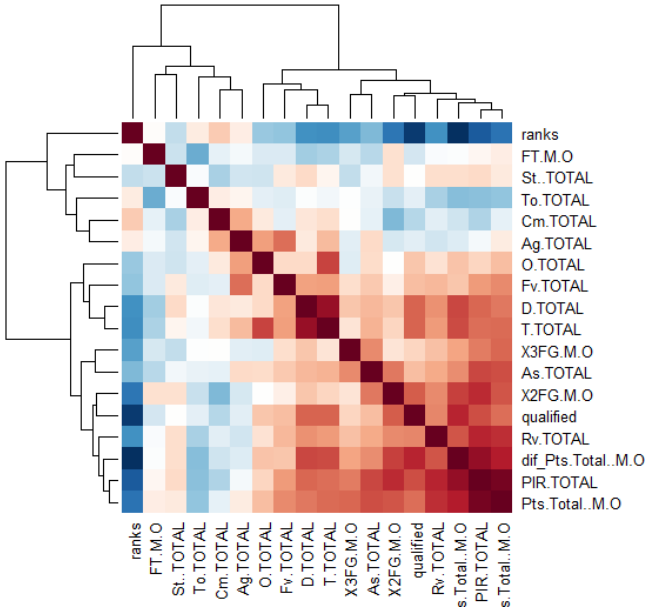
Πίνακας 3.4: Συντελεστές συσχέτισης για 2015-2016

Συγκριτικά με το προηγούμενο έτος παρατηρείται μια οριακή αύξηση των περισσότερων συντελεστών συσχέτισης. Ενδεικτικά αναφέρουμε πως την περίοδο 2014-2015 ο συντελεστής συσχέτισης μεταξύ των συνολικών πόντων και του PIR ήταν 0.66 και την περίοδο 2015-2016 είναι 0.80.

Στη συνέχεια, παρουσιάζονται τα παραπάνω αποτελέσματα γραφικά με το αντίστοιχο correlogram και heatmap :




Σχήμα 3.5: Correlogram για 2015-2016




Σχήμα 3.6: Heatmap για 2015-2016

Σε αυτά τα 2 σχήματα είναι εύκολα κατανοητή η σχέση των μεταβλητών μεταξύ τους ανάλογα με τα χρώματα, όπως αναφέραμε και παραπάνω. Ενδεικτικά αναφέρουμε πώς παρουσιάζεται έντονα αρνητική συσχέτιση (μπλε χρώμα)μεταξύ της συνολικής διαφοράς των πόντων και της μεταβλητής ranks (περίπου – 0.8).

Στη συνέχεια θα πρέπει να ελέγξουμε αν οι παραπάνω μεταβλητές είναι στατιστικά σημαντικές συγκριτικά με την μεταβλητή qualified. Για το σκοπό αυτό, όπως και παραπάνω, θα δημιουργήσουμε έναν πίνακα συνάφειας 2x2, αποτελούμενο από τις συχνότητες για τις 24 ομάδες.

**Pts vs qualified**

Σε αυτό το σημείο θα θεωρήσουμε ξανά ως επιτυχία αν ο μέσος όρος των πόντων που πέτυχε η κάθε ομάδα την χρονιά 2015-2016 είναι μεγαλύτερος από 80 και ως αποτυχία αν ο μέσος όρος είναι μικρότερος ή ίσος από 80. Έτσι λοιπόν δημιουργούμε μια δίτιμη μεταβλητή με τιμές 1 (επιτυχία) και 0 (όχι επιτυχία) αντίστοιχα.

Έχουμε ότι :

$H_0$ : οι μεταβλητές points και qualified είναι ασυσχέτιστες

έναντι της

$H_1$ : οι μεταβλητές points και qualified είναι συσχετισμένες.

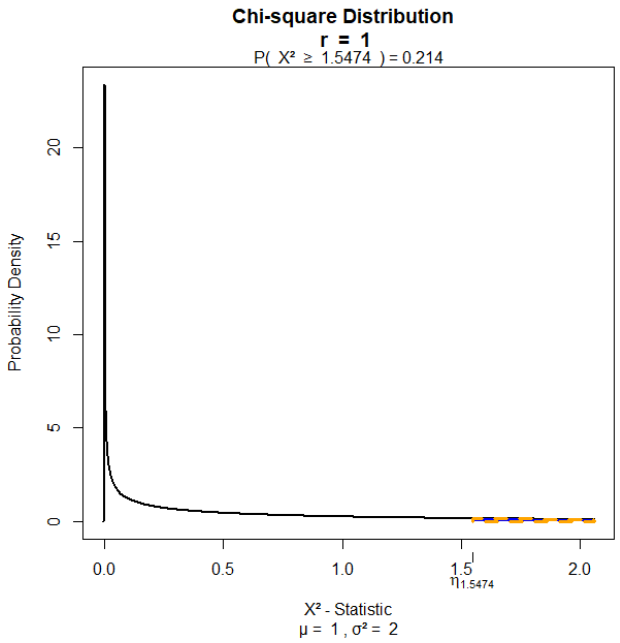
Σε αυτή την περίπτωση ο πίνακας συνάφειας θα είναι ο παρακάτω :

| | | Non-qualified | Qualified | Total |
|---|---|---|---|---|
| **Points** | ≤80 | 8 | 11 | 19 |
| | >80 | 0 | 5 | 5 |
| **Total** | | 8 | 16 | 24 |

Πίνακας 3.5: Πίνακας συνάφειας points-qualified για 2015-2016

Πραγματοποιώντας τον έλεγχο χ-τετράγωνο η R μας δίνει τα παρακάτω αποτελέσματα :

| **Pearson's Chi-squared test with Yates' continuity correction** |
|---|
| **data: tb1** |
| **X-squared = 1.5474, df = 1, p-value = 0.2135** |




Σχήμα 3.7: Chi-square Distribution points-qualified για 2015-2016

Παρατηρούμε πως η τιμή της στατιστικής συνάρτησης ισούται με 1.5474 και το αντίστοιχο p-value είναι ίσο με 0.214 > 0.05 (όπως φαίνεται και από την γραφική απεικόνιση). Συνεπώς σε επίπεδο στατιστικής σημαντικότητας 5% δεν μπορούμε να απορρίψουμε την μηδενική μας υπόθεση, κάτι το οποίο σημαίνει πως το αν θα προκριθεί μια ομάδα ή όχι στην επόμενη φάση δεν σχετίζεται με το αν θα έχει περισσότερους από 80 πόντους κατά μέσο όρο.

Για τις υπόλοιπες μεταβλητές θα θεωρήσουμε ξανά όπως και την προηγούμενη χρονιά, ότι υπάρχουν επιτυχίες με βάση τα παρακάτω κριτήρια :

Dif_points > |5|, 2FG > 50%, 3FG > 35 %, FT > 75%, O > 110, D >250, T > 360, As > 170,St >70,To >130, Fv >30, Ag >30, Cm >210, Rv >210, PIR > 800

Έτσι λοιπόν συγκεντρωτικά προκύπτει ο παρακάτω πίνακας με τις τιμές των p-values μετά τον έλεγχο χ-τετράγωνο για κάθε δίτιμη μεταβλητή σε σύγκριση με την μεταβλητή qualified για την χρονια 2015-2016 :

| **Μεταβλητές** | **p-value** |
|---|---|
| Dif_points vs qualified | 0.4642143 |
| **2FG vs qualiied** | **0.01282669*** |
| 3FG vs qualified | 0.8848361 |
| FT vs qualified | 0.8836175 |

| | |
|---|---|
| O vs qualified | 0.1797125 |
| D vs qualified | 0.1797125 |
| T vs qualified | 0.1336144 |
| As vs qualified | 0.6547208 |
| St vs qualified | 0.6547208 |
| To vs qualified | 1 |
| Fv vs qualified | 1 |
| Ag vs qualified | 0.8738447 |
| Cm vs qualified | 1 |
| **Rv vs qualified** | **0.05970881*** |
| **PIR vs qualified** | **0.01380479*** |

Πίνακας 3.6: Στατιστικά σημαντικές μεταβλητές ως προς την πρόκριση στα play-offs για 2015-2016

Παρατηρούμε πως σύμφωνα με τα κριτήρια που έχουμε ορίσει για το έτος 2015-2016 οι μεταβλητές που είναι στατιστικά σημαντικές και μπορούν να καθορίσουν ποιες ομάδες θα προκριθούν ή όχι διαφοροποιούνται από την προηγούμενη χρονιά.

- ➔ Πιο συγκεκριμένα, την συγκεκριμένη χρόνια βλέπουμε πως αν το ποσοστό ευστοχίας στα σουτ 2 πόντων είναι μεγαλύτερο από 50% τότε υπάρχει ισχυρή ένδειξη ότι η ομάδα θα προκριθεί στο top-16. (p-value = 0.012).
- ➔ Επιπλέον, μπορούμε να ισχυριστούμε πως εάν κάποια από τις 24 ομάδες έχει ειδικό δείκτη αξιολόγησης της Euroleague μεγαλύτερο από 800 μονάδες τότε υπάρχει ξανά ισχυρή ένδειξη πρόκρισης στην επόμενη φάση της διοργάνωσης. (p-value = 0.013)
- ➔ Τέλος αξίζει να αναφέρουμε πως και τα συνολικά φάουλ που δέχθηκε μια ομάδα μπορεί να αποτελέσει έναν σημαντικό παράγοντα πρόκρισης σε επίπεδο σημαντικότητας α = 10% (p-value = 0.059).
- ➔ Οι υπόλοιπες δίτιμες μεταβλητές δεν φαίνεται να σχετίζονται στατιστικά σημαντικά με την μεταβλητή qualified (p-value > 5%).

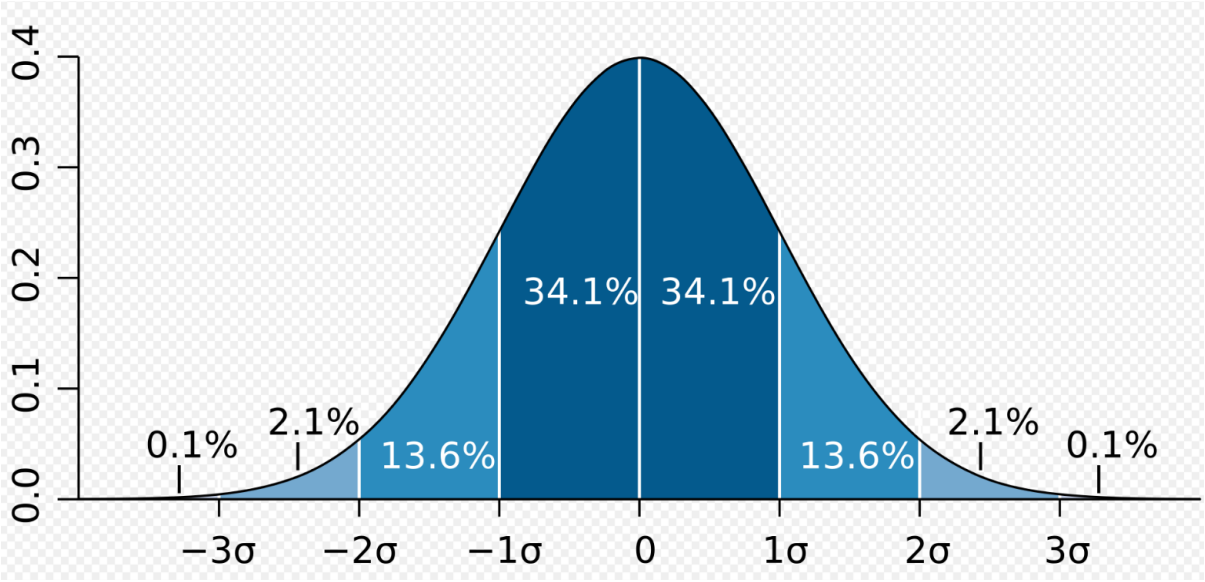
Συγκεντρωτικά όλα τα αποτελέσματα και οι εντολές που χρησιμοποιήθηκαν στην R παρουσιάζονται στο επισυναπτόμενο παράρτημα.

Τέλος, για την χρονιά 2015-2016 η τιμή του συντελεστή Somer's D ισούται ξανά με -0.53 και υποδηλώνει μια μέτρια αρνητική συσχέτιση μεταξύ των ποιοτικών

μεταβλητών qualified και κατάταξης στο group (ranks) . Αυτή η διαπίστωση μπορεί να επιβεβαιωθεί και με τον αντίστοιχο έλεγχο $\chi^2$.

### 3.3 Κανονικότητα κατανομών των ανεξάρτητων μεταβλητών

Σε πολλές στατιστικές αναλύσεις είναι αναγκαίος ο έλεγχος κανονικότητας κατανομών των ανεξάρτητων μεταβλητών. Ο λόγος που χρειαζόμαστε την κανονικότητα των δεδομένων είναι για να έχουν ισχύ κάποιες στατιστικές τεχνικές που θα χρησιμοποιήσουμε όπως έλεγχοι μέσων, διαμέσων κ.ά. Ένας απλός και εύκολος διαγραμματικός τρόπος εξακρίβωσης της κανονικότητας είναι μέσω του ιστογράμματος. Απαραίτητη προϋπόθεση για το ιστόγραμμα όπως και για όλα τα διαγράμματα είναι η μεταβλητή μας να είναι ποσοτική.




Σχήμα 3.8 : Τυποποιημένη κανονική κατανομή (μ=0,σ=1). Το σκούρο μπλε είναι λιγότερο από μία τυπική απόκλιση από το μέσο. Στην κανονική κατανομή, αυτό αφορά στο 68% των παρατηρήσεων, ενώ δύο τυπικές αποκλίσεις από τον μέσο (μπλε και σκούρο μπλε) αφορούν στο 95%, και τρεις τυπικές αποκλίσεις (ανοιχτό μπλε, μπλε και σκούρο μπλε) αφορούν το 99,7%.

Πηγή : https://www.wikiwand.com/el/

Το ιστόγραμμα όμως δεν είναι ικανό να μας απαντήσει αν τα δεδομένα προέρχονται από μια κανονική κατανομή με έναν συγκεκριμένο μέσο και μια διακύμανση. Η οπτική διερεύνηση δεν είναι πάντα σωστή για αυτό τον λόγο καταφεύγουμε σε τεστ κανονικότητας για να απαντήσουμε στο ερώτημα της κανονικότητας. Οι υποθέσεις στον έλεγχο κανονικότητας διαμορφώνονται ως εξής:

$H_0$ : Η κατανομή των δεδομένων δεν διαφέρει από την κανονική, έναντι της

$H_1$ : Η κατανομή των δεδομένων διαφέρει από την κανονική.

Οι πιο διαδεδομένοι στατιστικοί έλεγχοι που μας προσφέρουν την πληροφορία της κανονικής κατανομής είναι ο έλεγχος των Kolmogorov-Smirnov και Shapiro Wilk. Ο έλεγχος των Shapiro-Wilk έχει μεγαλύτερη στατιστική ισχύ από τον έλεγχο των Kolmogorov-Smirnov στην αναγνώριση μεταβλητών που δεν ακολουθούν κανονική κατανομή.

### 3.3.1 Έλεγχος Shapiro-Wilk για την χρονιά 2014-2015

Στην περίπτωσή μας εφαρμόζοντας τον μη-παραμετρικό έλεγχο κανονικότητας Shapiro-Wilk για κάθε μία από τις ανεξάρτητες ποσοτικές μας μεταβλητές την χρονιά 2014-2015 παρουσιάζουμε τον παρακάτω πίνακα με τις αντίστοιχες τιμές p-value του ελέγχου :

| **Μεταβλητές** | **p-value** |
|---|---|
| Pts.Total | 0.1266 |
| Dif_points | 0.7364 |
| 2FG | 0.2582 |
| 3FG | 0.9541 |
| FT | 0.9878 |
| O | 0.2147 |
| D | 0.9580 |
| T | 0.4855 |
| **As** | **0.02512** |
| St | 0.2847 |
| To | 0.5675 |
| Fv | 0.5235 |
| Ag | 0.1331 |
| Cm | 0.3311 |
| Rv | 0.8861 |
| **PIR** | **0.02626** |

Πίνακας 3.7 : Αποτελέσματα ελέγχου Shapiro-Wilk για την χρονιά 2014-2015

Έτσι λοιπόν από τον έλεγχο Shapiro-Wilk παρατηρούμε πως δεν μπορούμε να απορρίψουμε ότι σχεδόν όλες οι μεταβλητές μας ακολουθούν την κανονική κατανομή σε επίπεδο στατιστικής σημαντικότητας 5%. Εξαίρεση αποτελούν οι assist (p-value = 0.02512) και ο ειδικός δείκτης αξιολόγησης της διοργάνωσης (p-value = 0.02626) που όπως φαίνεται η μηδενική μας υπόθεση απορρίπτεται για α = 0.05 και συνεπώς δεν μπορούμε να ισχυριστούμε πως ακολουθούν την κανονική κατανομή.

### 3.3.2 Έλεγχος Shapiro-Wilk για την χρονιά 2015-2016

Στην περίπτωσή μας εφαρμόζοντας τον μη-παραμετρικό έλεγχο κανονικότητας Shapiro-Wilk για κάθε μία από τις ανεξάρτητες ποσοτικές μας μεταβλητές την χρονιά 2015-2016 παρουσιάζουμε τον παρακάτω πίνακα με τις αντίστοιχες τιμές p-value του ελέγχου :

| **Μεταβλητές** | **p-value** |
|---|---|
| Pts.Total | 0.1607 |
| Dif_points | 0.1454 |
| 2FG | 0.4507 |
| 3FG | 0.06571 |
| FT | 0.8041 |
| O | 0.8202 |
| D | 0.7847 |
| T | 0.3478 |
| As | 0.1066 |
| St | 0.5993 |
| To | 0.4522 |
| Fv | 0.1051 |
| **Ag** | **0.02401** |
| Cm | 0.5557 |
| Rv | 0.3165 |
| PIR | 0.6051 |

Πίνακας 3.8 : Αποτελέσματα ελέγχου Shapiro-Wilk για την χρονιά 2015-2016

Έτσι λοιπόν από τον έλεγχο Shapiro-Wilk παρατηρούμε πως δεν μπορούμε να απορρίψουμε ότι σχεδόν όλες οι μεταβλητές μας ακολουθούν την κανονική κατανομή σε επίπεδο στατιστικής σημαντικότητας 5%. Εξαίρεση αποτελεί η μεταβλητή Ag (p-value = 0.02401) που όπως φαίνεται η μηδενική μας υπόθεση απορρίπτεται για α = 0.05 και συνεπώς δεν μπορούμε να ισχυριστούμε πως ακολουθεί την κανονική κατανομή.

### 3.4 Διαχωριστική Ανάλυση

Η διαχωριστική ανάλυση είναι μια χρήσιμη στατιστική τεχνική που σκοπό έχει την διάκριση των διαφορών μεταξύ δύο ή περισσότερων αντικειμένων σε σχέση με πολλές ανεξάρτητες μεταβλητές (μεταβλητές πρόβλεψης) ταυτοχρόνως (Klecka, 1980). Οι εξαρτημένες μεταβλητές πρέπει να είναι κατηγορικές και οι κατηγορίες τους πρέπει να είναι διακεκριμένες. Η αρχή στην οποία στηρίζεται είναι ο γραμμικός συνδυασμός των ανεξάρτητων μεταβλητών έτσι ώστε να επιτευχθεί η άριστη ένταξη μιας μεταβλητής σε κάποια από τις διακριτές ομάδες. Η διαχωριστική ανάλυση, προτάθηκε από τον Ronald Fisher το 1936 για την επίλυση προβλημάτων ταξινόμησης των φυτών με τεχνικές που περιγράφονται σε πολλές στατιστικές αναλύσεις όπως του Cooley and Lohnes (1971), Klecka (1980) κ.ά.

Η βασική ιδέα της διαχωριστικής ανάλυσης (discriminant analysis) είναι να κατατάξει παρατηρήσεις (συνήθως πολυδιάστατες) σε γνωστούς πληθυσμούς με γνωστές κατανομές για κάθε πληθυσμό. Η διαχωριστική ανάλυση αποτελεί μια μέθοδο με πλήθος εφαρμογών σε πολλές επιστήμες. Έστω ότι υπάρχουν k υπο-πληθυσμοί (ομάδες), Π1, Π2, . . . , $\Pi_k$ με $k \geq 2$. Για τον κάθε υπο-πλυθυσμό $\Pi_k$ υπάρχει και μία κατανομή, $f_k$. Η διαχωριστική ανάλυση έχει 2 στόχους :

1. Τη διαχώριση ενός πληθυσμού σε ευδιάκριτα σύνολα (υπο-πληθυσμούς) και

2. Την ταξινόμηση παρατηρήσεων στους προηγούμενους γνωστούς πληθυσμούς με γνωστές κατανομές για κάθε πληθυσμό, με τη βοήθεια ενός κανόνα.

Αυτό που θα εξεταστεί εδώ είναι το πώς οι επεξηγηματικές μεταβλητές συνεισφέρουν στην σωστή ταξινόμηση των ομάδων, των οποίων η ιδιότητα είναι ήδη γνωστή (supervised classification). Για τη διαχώριση σε k ομάδες χρειάζονται $k - 1$ διαχωριστές (descriminators). Οι συναρτήσεις που χρειάζονται για τη διαχωριστική ανάλυση βρίσκονται στη βιβλιοθήκη MASS.

Οι μέθοδοι με τις οποίες μπορεί να πραγματοποιηθεί διαχωριστική ανάλυση ποικίλλουν:

- **Linear discriminant analysis** (**LDA**)
- **Quadratic discriminant analysis** (**QDA**)
- **Mixture discriminant analysis** (**MDA**)
- **Flexible Discriminant Analysis** (**FDA**)
- **Regularized discriminant anlysis** (**RDA**)

(Πηγή : https://el.wikipedia.org/)

Στην περίπτωσή μας θα εφαρμόσουμε linear discriminant analysis για την μεταβλητή qualified σε σχέση με όλες τις υπόλοιπες μεταβλητές μας.

### 3.4.1 Προυποθέσεις χρήσης διαχωριστικής ανάλυσης

Όπως σε όλες τις πολυμεταβλητές μεθόδους έτσι και στην διαχωριστική ανάλυση υπάρχουν κάποιες προϋποθέσεις που πρέπει να ικανοποιούνται. Αρχικά το μέγεθος του δείγματος θα πρέπει να είναι όσο το δυνατόν πιο μεγάλο έτσι ώστε να εξασφαλιστεί η αποτελεσματικότητα της διαδικασίας και η ορθότητα των συμπερασμάτων που θα εξάγουμε. Οι μεταβλητές θα πρέπει να ακολουθούν **κανονική κατανομή** και να είναι ανεξάρτητες και ασυσχέτιστες μεταξύ τους. Ενδέχεται κάποιες μεταβλητές να περιέχουν ακραίες παρατηρήσεις οι οποίες είναι πιθανό να διαστρεβλώσουν και να επηρεάσουν σε σημαντικό βαθμό τα συμπεράσματα. Για τον λόγο αυτό ο ερευνητής οφείλει να είναι προσεκτικός στην χρήση των δεδομένων του. Τέλος η διαχωριστική ανάλυση προϋποθέτει την ομοιογένεια των πινάκων διασποράς-συνδιασποράς.

### 3.4.2 Εφαρμογή διαχωριστικής ανάλυσης για την χρονιά 2014-2015

Αρχικά, θα πρέπει να φορτώσουμε στην R τα πακέτα tidyverse και MASS τα οποία μας είναι ιδιαίτερα χρήσιμα για την ανάλυσή μας. Στη συνέχεια θα χωρίσουμε το dataset μας σε 2 subsets : ένα training set και ένα testing set αντίστοιχα. Θα χρησιμοποιήσουμε το training test για να κατασκευάσουμε ένα προβλεπτικό μοντέλο και το testing set για να αξιολογήσουμε την ακρίβεια του μοντέλου. Πιο συγκεκριμένα θα χρησιμοποιήσουμε το 60% των δεδομένων μας σαν training set και το υπόλοιπο 40% σαν testing set. Ενδεικτικά παρουσιάζεται ο κώδικας που θα χρησιμοποιήσουμε στην R :

```
training_sample <- sample(c(TRUE, FALSE), nrow(mydata), replace = T, prob = c(0.6,0.4))
```

```
train <- mydata[training_sample,3:19 ];train
```

```
test <- mydata[!training_sample,3:19 ];test
```

```
lda.mydata<-
lda(qualified~Pts.Total..M.O+dif_Pts.Total..M.O+X2FG.M.O+X3FG.M.O+FT.M.O+T.TOTAL+St..TOTAL+To.TOTAL+Fv.TOTAL+Ag.TOTAL+Cm.TOTAL+Rv.TOTAL, train)
```

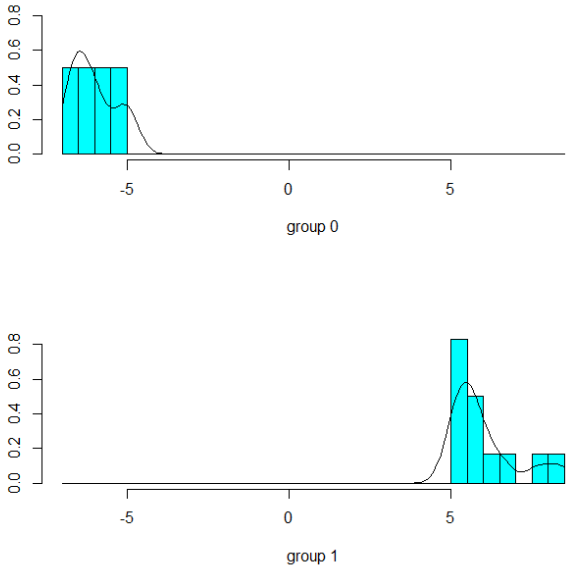
```
lda.mydata #show results
```

Έτσι λοιπόν προκύπτουν τα παρακάτω αποτελέσματα:

- ➔ Οι prior probabilities των 2 groups είναι 0.67 και 0.33 αντίστοιχα για το αν θα προκριθεί η ομάδα στο top-16 ή όχι, κάτι το οποίο περιμέναμε καθώς η διοργάνωση αποτελείται από 24 ομάδες και από αυτές οι 16 προκρίνονται.

➔ Δημιουργείται ένας linear discriminant(LD1) ο οποίος είναι ο παρακάτω γραμμικός συνδυασμός :
(0.56*Pts)+(1.19*dif_Points)+(-0.79*X2)+(-1.40*X3)+(-0.50*FT)+(-0.07*T)+(-0.21*St)+(0.20*To)+(0.08*Fv)+(-0.20*Ag)+(-0.05*Cm)+(0.09*Rv) και εξηγεί το 100% της μεταβλητότητας.

Οπτικοποιώντας το παραπάνω αποτέλεσμα μπορούμε να πάρουμε μία πιο σαφή εικόνα για τα group που σχηματίζονται καθώς και για το αν υπάρχουν αλληλοκαλυπτόμενες περιοχές ανάμεσα στα 2 group (qualified και non_qualified)




Σχήμα 3.9 : Διαχωρισμός μεταξύ των 2 groups για την χρονιά 2014-2015

Στη συνέχεια θα αξιολογήσουμε την προβλεπτική ικανότητα του μοντέλου δημιουργώντας ένα confusion matrix με τις actual και predicted τιμές για την μεταβλητή qualified. Αρχικά για το training set παρατηρούμε ότι ο συνολικός αριθμός σωστός προβλέψεων που είναι το άθροισμα της διαγώνιου είναι 16, κάτι το οποί σημαίνει ότι το μοντέλο μας προσαρμόζεται σωστά για όλες τις ομάδες που προκρίθηκαν.

| | Non-Qualified | Qualified |
|---|---|---|
| Non-qualified | 4 | 0 |
| Qualified | 0 | 12 |

Πίνακας 3.9 : confusion matrix για το training set την χρονιά 2014-2015

Για το testing set παρατηρούμε ότι 3 από τις 8 ομάδες έχουν ταξινομηθεί λάθος οπότε θα χρειαστεί περαιτέρω ανάλυση καθώς το accuracy είναι σχετικά μικρό και η

διαχωριστική ανάλυση ίσως δεν είναι η κατάλληλη μέθοδος για να χρησιμοποιήσουμε στην περίπτωση αυτή.

| | Non-Qualified | Qualified |
|---|---|---|
| Non-qualified | 3 | 2 |
| Qualified | 1 | 2 |

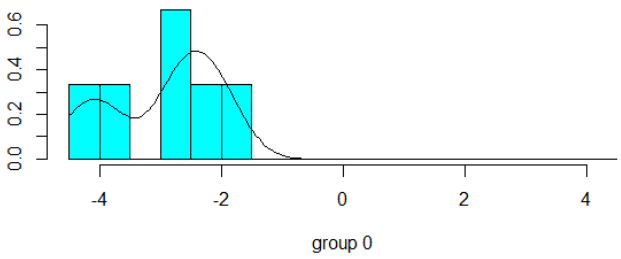
Πίνακας 3.10 : confusion matrix για το testing set την χρονιά 2014-2015

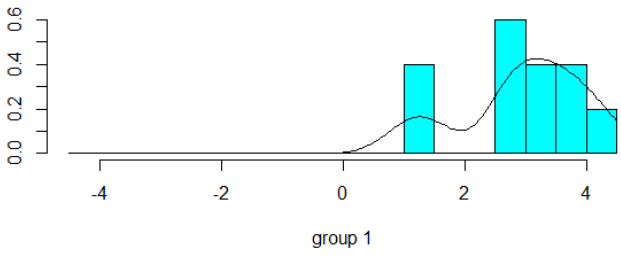
### 3.4.3 Εφαρμογή διαχωριστικής ανάλυσης για την χρονιά 2015-2016

Με παρόμοιο τρόπο χωρίζοντας το dataset μας για την χρονιά 2015-2016 σε training και testing set προκύπτουν τα παρακάτω αποτέσματα

- ➔ Οι prior probabilities των 2 groups είναι 0.33 και 0.67 αντίστοιχα για το αν θα προκριθεί η ομάδα στο top-16 ή όχι
- ➔ Δημιουργείται ένας linear discriminant(LD1) ο οποίος είναι ο παρακάτω γραμμικός συνδυασμός :
  (0.03*Pts)+(-0.01*dif_Points)+(0.65*X2)+(-0.01*X3)+(0.16*FT)+(0.07*T)+(-0.001*St)+(0.03*To)+(0.17*Fv)+(-0.24*Ag)+(0.07*Cm)+(-0.11*Rv) και εξηγεί το 100% της μεταβλητότητας.

Οπτικοποιώντας το παραπάνω αποτέλεσμα μπορούμε να πάρουμε μία πιο σαφή εικόνα για τα group που σχηματίζονται καθώς και για το αν υπάρχουν αλληλοκαλυπτόμενες περιοχές ανάμεσα στα 2 group (qualified και non_qualified)



Σχήμα 3.10 : Διαχωρισμός μεταξύ των 2 groups για την χρονιά 2015-2016

Στη συνέχεια θα αξιολογήσουμε την προβλεπτική ικανότητα του μοντέλου δημιουργώντας ένα confusion matrix με τις actual και predicted τιμές για την μεταβλητή qualified. Αρχικά για το training set παρατηρούμε ότι ο συνολικός αριθμός σωστός προβλέψεων που είναι το άθροισμα της διαγώνιου είναι 16, κάτι το οποίο σημαίνει ότι το μοντέλο μας προσαρμόζεται σωστά για όλες τις ομάδες που προκρίθηκαν.

| | Non-Qualified | Qualified |
|---|---|---|
| Non-qualified | 6 | 0 |
| Qualified | 0 | 10 |

Πίνακας 3.11 : confusion matrix για το training set την χρονιά 2015-2016

Για το testing set παρατηρούμε ότι 4 από τις 8 ομάδες έχουν ταξινομηθεί λάθος οπότε θα χρειαστεί περαιτέρω ανάλυση καθώς το accuracy είναι 50% και η διαχωριστική ανάλυση ίσως δεν είναι η κατάλληλη μέθοδος για να χρησιμοποιήσουμε στην περίπτωση αυτή.

| | Non-Qualified | Qualified |
|---|---|---|
| Non-qualified | 1 | 3 |
| Qualified | 1 | 3 |

Πίνακας 3.12 : confusion matrix για το testing set την χρονιά 2014-2015

# ΚΕΦΑΛΑΙΟ 4

## Προσαρμογή και μέθοδοι αξιολόγησης μοντέλων

Σε αυτό το κεφάλαιο θα αναφερθούμε στα μοντέλα λογιστικής παλινδρόμησης logit, probit, cauchit, cloglog και στην συνέχεια για το μοντέλο που θα επιλέξουμε θα αξιολογήσουμε την προβλεπτική του ισχύ. Τέλος, θα προχωρήσουμε σε σύγκριση αυτών των μοντέλων για να δούμε ποια χαρακτηριστικά επηρεάζουν περισσότερο την εξέλιξη των 24 ομάδων μέσα στην διοργάνωση.

### 4.1 Εισαγωγή στη Λογιστική Παλινδρόμηση

Η **παλινδρόμηση** είναι μια ευρέως χρησιμοποιούμενη στατιστική τεχνική μοντελοποίησης για την έρευνα της συσχέτισης μεταξύ μίας εξαρτώμενης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών (Κόκκινος, 2011).

Το λογιστικό μοντέλο είναι ένα μοντέλο, τα σφάλματα, του οποίου δεν υπακούν στην κανονική κατανομή και η μεταβλητή απόκρισης είναι διακριτή.

Η λογιστική παλινδρόμηση χρησιμοποιείται όταν επιθυμούμε να προβλέψουμε την απουσία ή την παρουσία ενός χαρακτηριστικού, ή ενός συμβάντος. Είναι μια γενίκευση της απλής γραμμικής παλινδρόμησης για την περίπτωση όπου η εξαρτημένη μεταβλητή (Y) είναι δίτιμη ( δηλαδή παίρνει την τιμή 0 όταν απουσιάζει το χαρακτηριστικό ή την τιμή 1 όταν υπάρχει το χαρακτηριστικό).

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

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

Παρόλο που στα δύο πρώτα προβλήματα είναι δυνατό σε κάποιες περιπτώσεις να τα παραλείψουμε και να χρησιμοποιήσουμε την 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 στην 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 $\beta_i$ 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 υπολείμματα 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:

1. Δίτιμη ή δυαδική ή διχοτομική (binary) ή διμερής εξαρτημένη μεταβλητή. Συνίσταται από δύο κατηγορίες, όπως π.χ. είναι οι εκβάσεις επιτυχία/αποτυχία, ΝΑΙ/ΟΧΙ, γεγονός απόν/παρόν.

2. Διατάξιμη (ordinal) μεταβλητή. Η εξαρτημένη μεταβλητή συνίσταται από τρεις ή περισσότερες κατηγορίες μεταξύ των οποίων ισχύει η έννοια της διάταξης, όπως π.χ. σε μια ερώτηση της κλίμακας διαφωνώ καθόλου, λίγο, μέτρια, αρκετά, πολύ, στην κατάταξη ενός στρώματος υλικού ως λεπτού, μεσαίου, παχέος.

3. 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, όπως π.χ. ο χαρακτηρισμός ενός τροφίμου ως τραγανού, μαλακού, εύθρυπτου ή του χρώματος αντικειμένων ως ερυθρού, πράσινου, κίτρινου κτλ.

(Πηγή : https://el.wikipedia.org/)

### 4.2 Το μοντέλο logit

Η λογιστική παλινδρόμηση είναι το γενικευμένο γραμμικό μοντέλο για δίτιμες αποκρίσεις με συνάρτηση σύνδεσης την logit. Αν Y είναι μία δίτιμη απόκριση με P(Y = 1) = π = E(Y ) το μοντέλο της λογιστικής παλινδρόμησης εκφράζεται ως εξής :

$$\text{logit}(\pi) \equiv \log\left(\frac{\pi}{1-\pi}\right) = x'\beta \equiv \beta_0 + \beta_1 x_1 + \ldots + \beta_\rho x_\rho$$

όπου $\chi = (1, x_1, \ldots, x_\rho)'$ το διάνυσμα των επεξηγηματικών μεταβλητών.

Αντιστρέφοντας την logit βλέπουμε ότι το μοντέλο της λογιστικής παλινδρόμησης εκφράζει την πιθανότητα :

$$\pi = \frac{\exp(x'\beta)}{1+\exp(x'\beta)}$$ ή ισοδύναμα την σχετική πιθανότητα (odds) ως $\frac{\pi}{1-\pi} = \exp(x'\beta)$

Με άλλα λόγια, η logit συνάρτηση είναι ο λογάριθμος της σχετικής πιθανότητας για ένα γεγονός (odds), δηλαδή ο λογάριθμος της πιθανότητας να συμβεί ένα γεγονός προς την πιθανότητα να μην συμβεί. Οι συντελεστές β δείχνουν πόσο αλλάζει το logit βασισμένο στις τιμές των επεξηγηματικών μεταβλητών.

Η logit είναι γνησίως αύξουσα συνάρτηση και ισχύουν τα εξής :

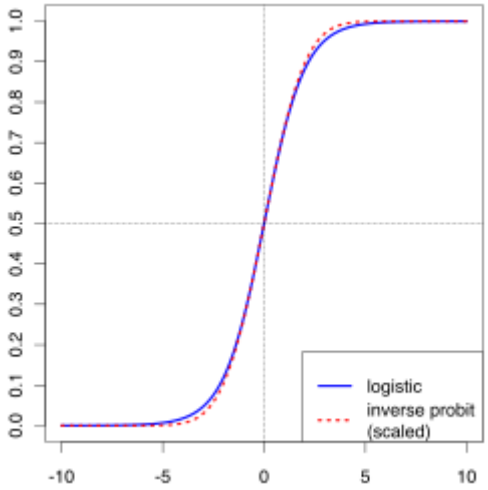
$\text{logit}(\pi) < 0 \Leftrightarrow \pi < \frac{1}{2}$

logit (π) = 0 ⇔ π = $\frac{1}{2}$

logit (π) > 0 ⇔ π > $\frac{1}{2}$

Η αντίστροφη συνάρτηση της logit δίνεται από τον τύπο :

logit$^{-1}$ (a) = $\frac{e^a}{1+e^a}$ , όπου a = log ($\frac{\pi}{1-\pi}$ ) ⇔ π = $\frac{e^a}{1+e^a}$




Σχήμα 4.1 : Λογιστική Παλινδρόμηση

(Πηγή : https://el.wikipedia.org/)

### 4.3 Το μοντέλο probit

Το μοντέλο probit εμφανίσθηκε στη βιβλιογραφία το 1934 οταν ο Bliss το χρησιμοποίησε για να αναλύσει δεδομένα τα οποία σχετίζονταν με σκαθάρια. Ο ίδιος το ονόμασε έτσι ενώνοντας κομμάτια των λέξεων probability unit. Η προσαρμογή των μοντέλων probit δεν διαφέρει πολύ από την προσαρμογή των μοντέλων logit. Οι οικονομολόγοι προτιμούν συνήθως το μοντέλο probit για τη μοντελοποίηση δίτιμων αποκρίσεων. Αυτά τα μοντέλα χρησιμοποιούν την αθροιστική Gaussian κατανομή σε αντίθεση με τα μοντέλα logit που χρησιμοποιούν την λογιστική συνάρτηση. Αν υποθέσουμε ότι Y είναι μια δίτιμη μεταβλητή απόκρισης που μπορεί να έχει μόνο 2 αποτελέσματα (0 και 1) και X μια επεξηγηματική μεταβλητή που επηρεάζει το αποτέλεσμα της Y, τότε η μορφή που έχει το μοντέλο είναι :

P (Y = 1 | X) = Φ ($X^T$ β ) , όπου Φ η αθροιστική συνάρτηση κατανομής (CDF) της τυπικής κανονικής κατανομής.

### 4.4 Το μοντέλο cauchit

Το μοντέλο cauchit χρησιμοποιείται σπανιότερα από τα υπόλοιπα. Λόγω του ότι η κατανομή Cauchy έχει εξαιρετικά βαριές ουρές, το μοντέλο προσαρμόζεται καλύτερα σε περιπτώσεις που η πιθανότητα π «αργεί» πολύ να πλησιάσει το 0 και το 1 ως συνάρτηση των επεξηγηματικών μεταβλητών. Η συνάρτηση σύνδεσης σε αυτή την περίπτωση είναι η αντίστροφη αθροιστική συνάρτηση κατανομής για μια τυπική

Cauchy κατανομή. Η συνάρτηση πυκνότητας πιθανότητας και η συνάρτηση κατανομής για μια τυπική Cauchy κατανομή είναι :

PDF : f(x) = $\frac{1}{\pi\,(1+\,x^2)}$ , όπου x $\epsilon R$ και

CDF : F(x) = 0.5 + $\frac{tan^{-1}x}{\pi}$ ,όπου x $\epsilon R$

### 4.5 Το μοντέλο cloglog

Το complementary log log μοντέλο ταιριάζει σε περιπτώσεις που η P(Y =1 ) πλησιάζει γρήγορα τη μονάδα αλλά αργά το 0. Η μορφή αυτού του μοντέλου είναι η παρακάτω :

cloglog(π) ≡ log(- log (1 - $\pi_x$)) = xβ , όπου $\pi_x$ = P (Y =1 | X=x)

Σημειώνουμε πως εκτός από αυτές τις συναρτήσεις σύνδεσης υπάρχουν αρκετές ακόμη για την περίπτωση των γενικευμένων γραμμικών μοντέλων.

| Family | Link | Mean Function | $\Psi(\mathbf{x}_i'\boldsymbol{\beta})$ |
|---|---|---|---|
| gaussian | identity | $\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$ | $1/\sigma^2$ |
| binomial | logit | $\mu_i = \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta})}$ | $\mu_i(1-\mu_i)$ |
| binomial | probit | $\mu_i = \Phi(\mathbf{x}_i'\boldsymbol{\beta})$ | $\frac{\phi(\mathbf{x}_i'\boldsymbol{\beta})^2}{\Phi(\mathbf{x}_i'\boldsymbol{\beta})(1-\Phi(\mathbf{x}_i'\boldsymbol{\beta}))}$ |
| binomial | cloglog | $\mu_i = 1 - \exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))$ | $\frac{1-\mu_i}{\mu_i}[\log(1-\mu_i)]^2$ |
| poisson | log | $\mu_i - \exp(\mathbf{x}_i'\boldsymbol{\beta})$ | $\mu_i$ |
| poisson | identity | $\mu_i - \mathbf{x}_i'\boldsymbol{\beta}$ | $1/\mu_i$ |
| poisson | sqrt | $\mu_i = (\mathbf{x}_i'\boldsymbol{\beta})^2$ | 4 |
| gamma | inverse | $\mu_i = (\mathbf{x}_i'\boldsymbol{\beta})^{-1}$ | $a\mu_i^2$ |
| gamma | identity | $\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$ | $a/\mu_i^2$ |
| gamma | log | $\mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$ | $a$ |
| inverse gaussian | inverse squared | $\mu_i - (\mathbf{x}_i'\boldsymbol{\beta})^{-1/2}$ | $\lambda\mu_i^3/4$ |

Πηγή : Arthur Charpentier (2014), Computational Actuarial Science with R

Πίνακας 4.1 : Συναρτήσεις σύνδεσης για γενικευμένα γραμμικά μοντέλα

### 4.6 Εφαρμογή για την χρονιά 2014-2015

Στην συνέχεια θα προσαρμόσουμε και τα 4 μοντέλα που αναφέραμε ώστε να δούμε ποιο δίνει ακριβέστερα αποτελέσματα για την ανάλυσή μας.

#### 4.6.1 Ποιες μεταβλητές είναι στατιστικά σημαντικές για την χρονιά 2014-2015;

Σε αυτό το σημείο πριν προχωρήσουμε στην εύρεση ενός ολοκληρωμένου μοντέλου λογιστικής παλινδρόμησης με συνάρτηση σύνδεσης την logit, θα δοκιμάσουμε όλα εκείνα τα μοντέλα που έχουν μόνο μια ερμηνευτική μεταβλητή για να διαπιστώσουμε πώς επιδρά η κάθε μεταβλητή ξεχωριστά στην μεταβλητή απόκρισης. Ως μεταβλητή απόκρισης θεωρούμε και σε αυτή την περίπτωση αν μια ομάδα προκρίθηκε ή όχι στην επόμενη φάση της διοργάνωσης. Για τον σκοπό αυτό θα παρουσιάσουμε παρακάτω έναν συγκεντρωτικό πίνακα με τα αντίστοιχα p-value, καθώς και τις τιμές των AIC για το κάθε μοντέλο που τρέξαμε στην R. Πιο συγκεκριμένα θα παραθέσουμε τα

αποτελέσματα 16 διαφορετικών μοντέλων με τις αντίστοιχες ανεξάρτητες μεταβλητές :

| Qualified~ | AIC | p-value |
|---|---|---|
| **Pts.Total** | 31.432 | 0.124 |
| **Dif_Points** | 22.39 | **0.0134*** |
| **X2** | 33 | 0.239 |
| **X3** | 34.31 | 0.626 |
| **FT** | 34.43 | 0.727 |
| **O** | 33.39 | 0.311 |
| **D** | 27.68 | **0.0276*** |
| **T** | 20.567 | **0.0139*** |
| **As** | 32.766 | 0.233 |
| **St** | 32.783 | 0.209 |
| **To** | 34.425 | 0.722 |
| **Fv** | 33.22 | 0.263 |
| **Ag** | 34.548 | 0.946 |
| **Cm** | 33.127 | 0.256 |
| **Rv** | 33.095 | 0.244 |
| **PIR** | 27.343 | **0.0438*** |

Πίνακας 4.2 : Επιρροή της κάθε μεταβλητής ξεχωριστά

Όπως είναι λογικό η συνολική διαφορά των πόντων για την κάθε ομάδα, που ουσιαστικά είναι η διαφορά του τελικού σκορ κάθε αγώνα, είναι στατιστικά σημαντική αφού προσδιορίζει άριστα τον τελικό νικητή. Για τον συγκεκριμένο λόγο δεν μπορεί να συμπεριληφθεί στο τελικό μας μοντέλο. Επίσης η μεταβλητή PIR παρέχει μια συνδυαστική πληροφορία για την κάθε ομάδα και θα ήταν καλύτερο να μην την συμπεριλάβουμε στο μοντέλο μας, αφού μας ενδιαφέρει να εξετάσουμε ποια συγκεκριμένη μεταβλητή επηρεάζει το τελικό αποτέλεσμα. Από όλα τα υπόλοιπα χαρακτηριστικά, τα συνολικά ριμπάουντ φαίνεται να έχουν καθοριστικό ρόλο ( p-value

= 0.0139) στην έκβαση του αποτελέσματος καθώς και τα αμυντικά ριμπάουντ (p-value = 0.0276).

### 4.6.2 Επιλογή χαρακτηριστικών (feature selection) για 2014-2015

Για να προχωρήσουμε σε διαδικασίες εξόρυξης γνώσης θα πρέπει να καταλήξουμε σε εκείνες τις μεταβλητές οι οποίες ουσιαστικά συνεισφέρουν σημαντικά και συντελούν στην αυξημένη προβλεπτική ικανότητα του μοντέλου. Υπάρχουν πολλές μέθοδοι και τεχνικές για την επιλογή των κατάλληλων χαρακτηριστικών (στηλών), ώστε το τελικό μοντέλο που θα χρησιμοποιήσουμε να είναι πιο απλοποιημένο και κατανοητό για τον αναγνώστη.

**α. Random forest**

Η πρώτη μέθοδος που μπορούμε να χρησιμοποιήσουμε για την επιλογή των κατάλληλων μεταβλητών που θα εισάγουμε στο μοντέλο είναι η random forest.

Είναι από τους δημοφιλέστερους αλγόριθμους στην κατηγορία του, κυρίως για την ταχύτητα αλλά και την ακρίβεια που προσφέρει. Σύμφωνα με τον δημιουργό του (Tin Kan Ho, 1995) αλλά και από μεταγενέστερες βελτιώσεις του αλγορίθμου (Breiman, 2001 και Cutler, 2008) :

→Προσφέρει την καλύτερη ακρίβεια μεταξύ των υπαρχόντων αλγορίθμων

→Η ταχύτητά του είναι πολύ καλή ακόμα και σε πολύ μεγάλα σύνολα δεδομένων

→Μπορεί να χειριστεί αποδοτικά πάρα πολύ μεγάλο αριθμό χαρακτηριστικών

→Δίνει μια εκτίμηση για το ποια χαρακτηριστικά είναι τα πιο σημαντικά στην κατηγοριοποίηση

→Δεν χρειάζεται την χρήση διαφορετικού συνόλου δεδομένων για τον έλεγχο ακρίβειας (δεν είναι δηλαδή απαραίτητο το cross-validation), καθώς η εκτίμηση του λάθους γενίκευσης γίνεται από τον ίδιο τον αλγόριθμο κατά την εκτέλεσή του.

→Μπορεί να χειριστεί αποδοτικά ελλιπή δεδομένα

Ο αλγόριθμος random forest αποτελεί μια supervised τεχνική κυρίως για προβλήματα ταξινόμησης και λειτουργεί με τον παρακάτω τρόπο :

Βήμα 1 : Δημιουργούμε ένα bootstrapped dataset. Επιλέγουμε τυχαία δείγματα από τα αρχικά δεδομένα μας. Ένα σημαντικό σημείο που πρέπει να τονίσουμε είναι ότι μπορούμε να επιλέξουμε ένα συγκεκριμένο δείγμα και περισσότερες από μία φορές.

Βήμα 2 : Δημιουργούμε decision trees (δέντρα αποφάσεων).

Βήμα 3 : Γυρίζουμε ξανά στο Βήμα 1 και επαναλαμβάνουμε την διαδικασία.

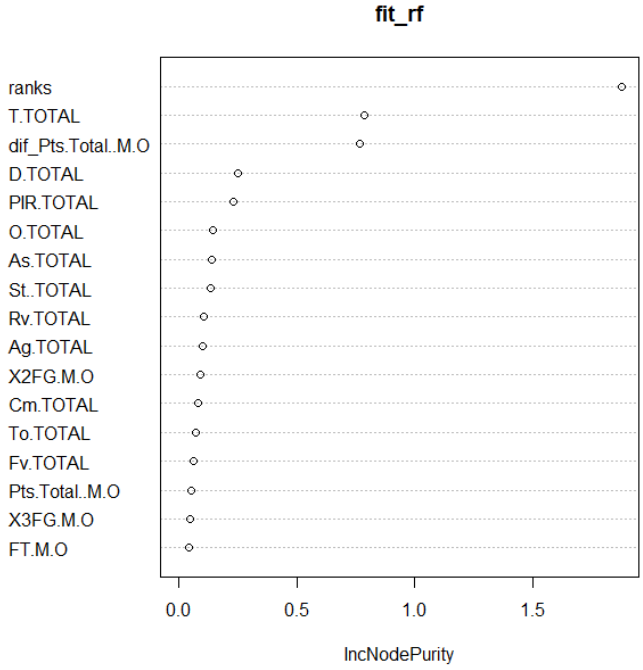
Βήμα 4 : Προβλέπουμε το αποτέλεσμα από ένα καινούργιο data point, χρησιμοποιώντας το άθροισμα από όλα τα δέντρα μαζί. Η συγκεκριμένη διαδικασία είναι ευρέως γνωστή ως bagging.

Βήμα 5 : Τέλος, γίνεται η αξιολόγηση του μοντέλου μας. Στην πραγματικότητα το 1/3 του original dataset δεν περιλαμβάνεται στο bootstrapped dataset. Αυτό το dataset λέγεται out-of-bag (OOB) και χρησιμοποιείται για να ελέγξει την ακρίβεια του μοντέλου, δηλαδή εάν είναι αποτελεσματικό ή όχι.

(Πηγή:https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/)

Έτσι λοιπόν εκτελώντας τις παρακάτω εντολές στην R , διαπιστώνουμε πως με εξαίρεση την κατηγορική μεταβλητή ranks, οι 2 πιο σημαντικές μεταβλητές που πρέπει να εισαχθούν στο μοντέλο μας είναι τα συνολικά ριμπάουντ καθώς και η συνολική διαφορά των πόντων.

```
library(randomForest)

library(Metrics)

fit_rf = randomForest(qualified~.,data=mydata[,-1])

# Create an importance based on mean decreasing gini

importance(fit_rf)

varImpPlot(fit_rf)
```




Σχήμα 4.2 : Γραφική αναπαράσταση της σημαντικότητας των μεταβλητών με τη μέθοδο random forest.

### β. Μέθοδος Stepwise

Μια εναλλακτική μέθοδος για την επιλογή μεταβλητών είναι η μέθοδος stepwise. Σύμφωνα με τη μέθοδο αυτή, για κάθε διαδοχικό βήμα η υπόθεση $H_0 : \beta_j = 0$ ελέγχεται για όλες τις δυνατές ανεξάρτητες μεταβλητές ώστε να αποκλείονται εκείνες για τις οποίες οι τιμές της στατιστικής συνάρτησης $|T_j|$ είναι μικρότερες από ένα προκαθορισμένο κρίσιμο επίπεδο. Η επόμενη μεταβλητή προστίθεται στο υποσύνολο με την ίδια διαδικασία χρησιμοποίησης του κριτηρίου του μεγίστου συντελεστή συσχέτισης όπως στη μέθοδο της προοδευτικής προσθήκης μεταβλητών (forward selection). Αυτή η επιλογή συνεχίζεται μέχρι να φτάσουμε σε ένα υποσύνολο μεταβλητών για το οποίο καμμιά από τις μεταβλητές που περιέχει το υποσύνολο αυτό δεν έχουν τιμή για τη στατιστική συνάρτηση $|T_j|$ μικρότερη από κάποια συγκεκριμένη κρίσιμη τιμή της μεταβλητής t και δεν υπάρχουν άλλες μεταβλητές που θα πρέπει να αξιολογηθούν για να περιληφθούν στο μοντέλο. Εφαρμόζοντας την παραπάνω διαδικασία στην R, προκύπτει ότι η πιο σημαντική μεταβλητή για να εισάγουμε στο μοντέλο μας, είναι η μεταβλητή ranks κάτι το οποίο δεν επιθυμούμε στην περίπτωση μας.

```
base.mod <- glm(qualified ~ 1 ,family=binomial, data= mydata[,-1])  # base intercept only model

all.mod <- glm(qualified ~ . ,family=binomial, data= mydata[,-1]) # full model with all predictors
```

```
stepMod <- step(base.mod, scope = list(lower = base.mod, upper = all.mod), direction = "both", trace = 0, steps = 1000)  # perform step-wise algorithm

shortlistedVars <- names(unlist(stepMod[[1]])) # get the shortlisted variable.

shortlistedVars <- shortlistedVars[!shortlistedVars %in% "(Intercept)"]  # remove intercept

print(shortlistedVars)
```

**γ. Μέθοδος Boruta**

Ο συγκεκριμένος αλγόριθμος παρουσιάστηκε από τους Πολωνούς ερευνητές Kursa και Rudnicki το 2010. Το όνομα του προέρχεται από μια σλαβική μυθολογική φιγούρα που ενσαρκώνει το πνεύμα του δάσους. Ουσιαστικά ο αλγόριθμος Boruta αποτελεί μια πιο γρήγορη εφαρμογή της μεθόδου random forest και αναλυτικά παρουσιάζεται παρακάτω ο τρόπος λειτουργείας του :

Βήμα 1 : Εισάγουμε τυχαιότητα ή θόρυβο στα δεδομένα που μας έχουν δωθεί δημιουργώντας «ανακατεμένα αντίγραφα» (shuffled copies), τα οποία ονομάζονται σκιώδη χαρακτηριστικά (shadow features).

Βήμα 2 : Κάνουμε train έναν random forest classifier στα διευρυμένα πλέον δεδομένα εφαρμόζοντας ένα μέτρο σημαντικότητας για την επιλογή των χαρακτηριστικών ( η προεπιλογή είναι το mean decrease accuracy).

Βήμα 3 : Σε κάθε επανάληψη, ελέγχεται πότε ένα χαρακτηριστικό έχει υψηλότερη σημαντικότητα από τα αντίστοιχα καλύτερα σκιώδη χαρακτηριστικά.

Βήμα 4 : Ο αλγόριθμος σταματάει είτε όταν όλα τα χαρακτηριστικά επιβεβαιωθούν είτε απορριφθούν είτε φτάσουμε σε ένα συγκεκριμένο όριο.

(Πηγή :https://www.andreaperlato.com/mlpost/feature-selection-using-boruta-algorithm/)

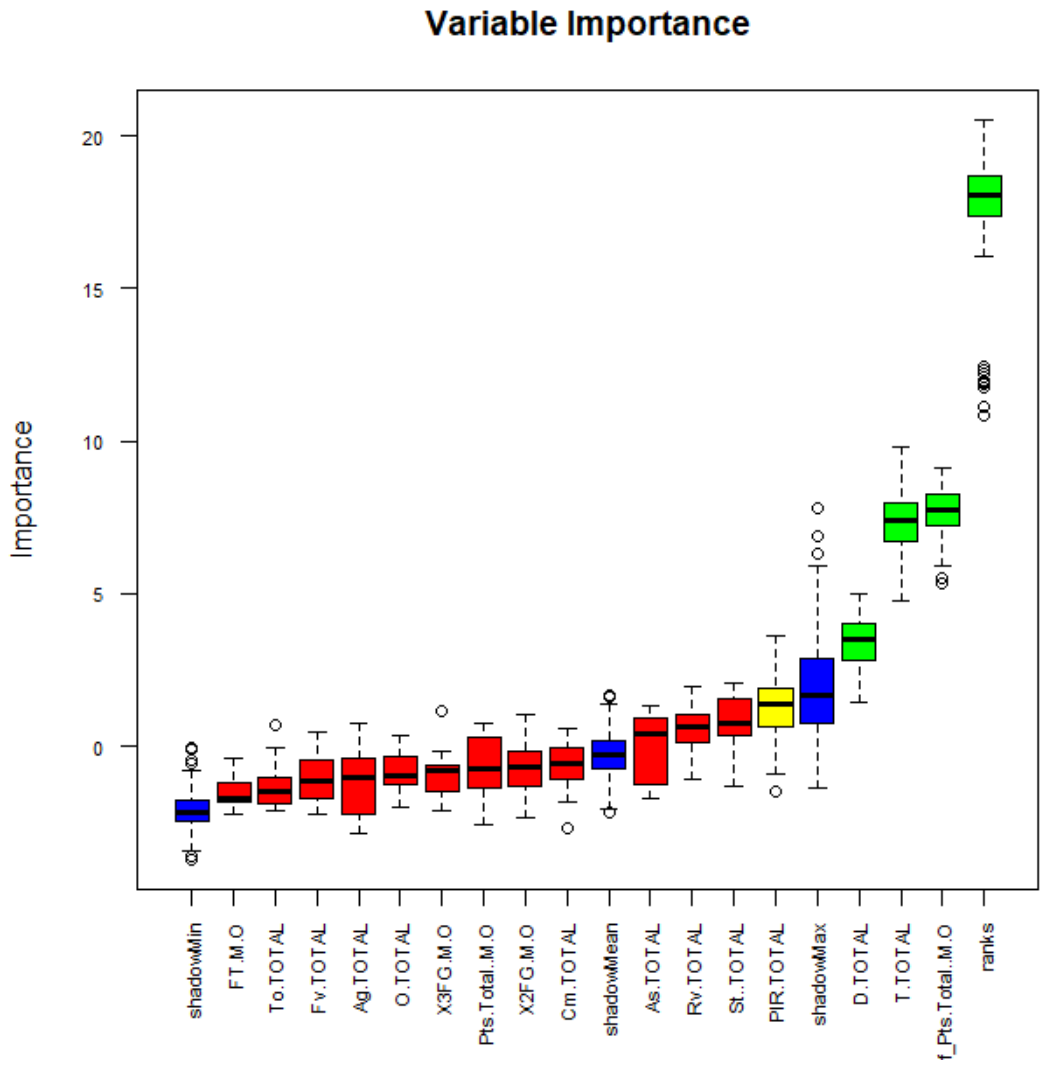
(Πηγή : https://www.datacamp.com/community/tutorials/feature-selection-R-boruta)

Ο αλγόριθμος Boruta, δεν ενδείκνυται να χρησιμοποιείται όταν τα δεδομένα μας περιέχουν ελλιπείς τιμές.

Χρησιμοποιώντας στην R τις παρακάτω εντολές έχουμε ότι :

```
library(Boruta)

# Decide if a variable is important or not using Boruta
```

```
boruta_output <- Boruta(qualified~.,data=na.omit(mydata[,-1]),doTrace=2)  # perform Boruta search

boruta_signif <- names(boruta_output$finalDecision[boruta_output$finalDecision %in% c("Confirmed", "Tentative")])  # collect Confirmed and Tentative variables

print(boruta_signif)  # significant variables

plot(boruta_output, cex.axis=.7, las=2, xlab="", main="Variable Importance")
```




Σχήμα 4.3 : Γραφική αναπαράσταση της σημαντικότητας των μεταβλητών με τη μέθοδο Boruta.

Παρατηρούμε ότι έγινε μια ταξινόμηση των μεταβλητών σε 3 επίπεδα, τα οποία στο σχήμα δηλώνονται με τα αντίστοιχα χρώματα: πράσινο, κόκκινο, κίτρινο.

Με πράσινο χρώμα είναι οι μεταβλητές που έχουν επιβεβαιωθεί (confirmed) από τον random forest (RF) classifier. Έτσι λοιπόν σύμφωνα με το σχήμα **και σε σύγκριση με το δεύτερο κεφάλαιο**, παρατηρούμε πως οι πιο σημαντικές μεταβλητές, αν εξαιρέσουμε την μεταβλητή ranks, είναι η διαφορά των πόντων και τα συνολικά ριμπάουντ.

Με κόκκινο χρώμα παρουσιάζονται τα χαρακτηριστικά που δεν έχουν εγκριθεί (p-value > 0.05 ) από τον RF. Χαρακτηριστικά, βλέπουμε ότι έχουν ¨αρνητική σημαντικότητα¨ στο σχήμα.

Με κίτρινο χρώμα απεικονίζονται οι λεγόμενες tentative μεταβλητές, δηλαδή μεταβλητές που είναι στην κρίση του εκάστοτε αναλυτή αν θα τις συμπεριλάβει στην μελέτη του και η συνεισφορά τους στο μοντέλο είναι υπό διερεύνηση.

Τέλος, με μπλε χρώμα δηλώνονται τα shadow features που όπως είπαμε στα βήματα περιγραφής του αλγορίθμου, δεν είναι πραγματικές μεταβλητές, απλά χρησιμοποιούνται για να αποφασίσουμε αν μια μεταβλητή είναι σημαντική ή όχι. Οι τιμές τους αντιστοιχούν στο min, average και max z-score.

### 4.6.3 Εφαρμογή logit μοντέλου για 2014-2015

Αφού ελέγξαμε την σημαντικότητα των μεταβλητών μας και μετά από τη εφαρμογή πολλαπλών μοντέλων ώστε να διαπιστώσουμε ποιες μεταβλητές προβλέπουν καλύτερα αν μια ομάδα προκρίθηκε στην επόμενη φάση ή όχι, είμαστε πλέον σε θέση να δημιουργήσουμε το τελικό μοντέλο που θα χρησιμοποιήσουμε. Η κατηγορική μεταβλητή ranks δεν μπορεί να χρησιμοποιηθεί καθώς, αν και στις περισσότερες μεθόδους φαίνεται να είναι στατιστικά σημαντική, περιέχει μεγάλο αριθμό κατηγοριών (6) σε σύγκριση με το μέγεθος του δείγματος. Ακόμη πρέπει να αποφευχθεί η συνύπαρξη μεταβλητών με ισχυρές συσχετίσεις. Έτσι λοιπόν συνδυάζοντας όλα τα παραπάνω επιλέξαμε να φτιάξουμε ένα λογιστικό μοντέλο με μόνη ερμηνευτική τα συνολικά ριμπάουντ, ώστε να διερευνήσουμε με μεγαλύτερη ακρίβεια αν όντως επιδρούν σημαντικά στην απόδοση μιας ομάδας.

Στην ερώτηση πόσες επεξηγηματικές μεταβλητές μπορούμε να χρησιμοποιήσουμε οι απαντήσεις ποικίλλουν. Αν τα δεδομένα είναι μη-ισορροπημένα, δηλαδή αν η απόκριση Y =1 εμφανίζεται περισσότερες ή λιγότερες φορές σε σχέση με την Y = 0, τότε δεν είναι σωστό να χρησιμοποιήσουμε στο μοντέλο μας πολλές επεξηγηματικές μεταβλητές. Ένας χονδρικός κανόνας είναι να χρησιμοποιήσουμε 1 επεξηγηματική μεταβλητή για τουλάχιστον 10 αποκρίσεις ανά κατηγορία. Στην περίπτωση που εξετάζουμε, από κάθε group 4 ομάδες προκρίνονται στην επόμενη φάση ( Y =1) και 2 δεν προκρίνονται (Y = 0). Αυτός είναι ένας από τους λόγους που επιλέγουμε να χρησιμοποιήσουμε μια επεξηγηματική μεταβλητή στο μοντέλο μας. Επιπλέον, αν οι επεξηγηματικές μεταβλητές εμφανίζουν πολυσυγγραμμικότητα(multicollinearity) τότε δημιουργούνται προβλήματα καθώς μπορεί να φαίνεται ότι καμία από τις μεταβλητές δεν είναι στατιστικά σημαντική όταν όλες μπαίνουν στο μοντέλο. Αυτό συμβαίνει

διότι, υπό πολυσυγγραμμικότητα, οι εκτιμήσεις των τυπικών σφαλμάτων μπορεί να προκύψουν πολύ μεγάλες.

```
fit.logit<-glm(qualified~T.TOTAL,family=binomial,data=mydata)
summary(fit.logit)
```

```
Call:
glm(formula = qualified ~ T.TOTAL, family = binomial, data = mydata)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.8210  -0.4028   0.1767   0.5012   1.6368

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -38.38803   15.71802  -2.442   0.0146 *
T.TOTAL       0.11285    0.04587   2.460   0.0139 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 30.553  on 23  degrees of freedom
Residual deviance: 16.567  on 22  degrees of freedom
AIC: 20.567

Number of Fisher Scoring iterations: 6
```

Προκύπτει λοιπόν ότι το εκτιμώμενο μοντέλο είναι :

$\text{logit}(\hat{\pi}) \equiv \log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -38.38 + 0.11*\text{T.TOTAL}$

που σημαίνει ότι η εκτιμώμενη σχετική πιθανότητα είναι :

$\frac{\hat{\pi}}{1-\hat{\pi}} = e^{-38.38 + 0.11*\text{T.TOTAL}}$ και $\hat{\pi} = \frac{e^{-38.38 + 0.11*\text{T.TOTAL}}}{1+ e^{-38.38 + 0.11*\text{T.TOTAL}}}$

Δηλαδή για κάθε μοναδιαία αύξηση των συνολικών ριμπάουντ, ο λογάριθμος της σχετικής πιθανότητας για το αν η ομάδα θα προκριθεί ή όχι αυξάνεται κατά περίπου 0.11 μονάδες.

Ο έλεγχος Wald για την επεξηγηματική μεταβλητή ριμπάουντ μας λέει ότι είναι στατιστικά σημαντική :

$\frac{\hat{b}}{\hat{\sigma}(\hat{b})} = \frac{0.11285}{0.04587} = 2.46$ , και p-value = 0.0139 < 5%

Και επειδή $\hat{b} > 0$ συμπεραίνουμε ότι όσα περισσότερα είναι τα συνολικά ριμπάουντ τόσο αυξάνεται η πιθανότητα η ομάδα να προκριθεί στο top-16.

Η τελική απόκριση του μοντέλου 16.567 κάτι το οποίο είναι ικανοποιητικό καθώς ένα καλό μοντέλο θα πρέπει να έχει μικρή τιμή. Τέλεια προσαρμογή υπάρχει όταν η απόκλιση είναι μηδενική προφανώς.

Τέλος, ο αλγόριθμος μεγιστοποίησε το λογάριθμο της πιθανοφάνειας μετά από 6 επαναλήψεις. Γενικά, όσο λιγότερες είναι οι επαναλήψεις τόσο πιο ευσταθές είναι και το μοντέλο που έχουμε προσαρμόσει.

### 4.6.4 Εφαρμογή probit μοντέλου για 2014-2015

```
fit.probit<-glm(qualified~T.TOTAL,family=binomial(link="probit"),data=mydata)
summary(fit.probit)
```

```
Call:
glm(formula = qualified ~ T.TOTAL, family = binomial(link = "probit"),
    data = mydata)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
```

```
-1.8046  -0.3660   0.1147   0.4970   1.6369


Coefficients:

            Estimate Std. Error z value Pr(>|z|)

(Intercept) -23.09328    8.65923  -2.667  0.00766 **

T.TOTAL       0.06784    0.02521   2.691  0.00712 **

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


(Dispersion parameter for binomial family taken to be 1)


    Null deviance: 30.553  on 23  degrees of freedom

Residual deviance: 16.311  on 22  degrees of freedom

AIC: 20.311


Number of Fisher Scoring iterations: 7
```

Προκύπτει λοιπόν ότι το εκτιμώμενο μοντέλο είναι :

probit($\hat{\pi}$) ≡ $\Phi^{-1}$ ($\hat{\pi}$) = -23.09 + 0.06*T.TOTAL

που σημαίνει ότι η εκτιμώμενη πιθανότητα είναι :

και $\hat{\pi}$ = Φ (-23.09 + 0.06*T.TOTAL)

Ο έλεγχος Wald για την επεξηγηματική μεταβλητή ριμπάουντ μας λέει ότι είναι στατιστικά σημαντική :

$\frac{\hat{b}}{\hat{\sigma}(\hat{b})} = \frac{0.06784}{0.02521}$ = 2.691 , και p-value = 0.007 < 5%

Και επειδή $\hat{b} > 0$ συμπεραίνουμε ότι όσα περισσότερα είναι τα συνολικά ριμπάουντ τόσο αυξάνεται η πιθανότητα η ομάδα να προκριθεί στο top-16.

Η τελική απόκριση του μοντέλου 16.311 κάτι το οποίο είναι ικανοποιητικό καθώς ένα καλό μοντέλο θα πρέπει να έχει μικρή τιμή. Τέλεια προσαρμογή υπάρχει όταν η απόκλιση είναι μηδενική προφανώς.

Τέλος, ο αλγόριθμος μεγιστοποίησε το λογάριθμο της πιθανοφάνειας μετά από 7 επαναλήψεις. Γενικά, όσο λιγότερες είναι οι επαναλήψεις τόσο πιο ευσταθές είναι και το μοντέλο που έχουμε προσαρμόσει.

### 4.6.5 Εφαρμογή cauchit μοντέλου για 2014-2015

```
fit.cauchit<-
glm(qualified~T.TOTAL,family=binomial(link="cauchit"),data=mydata)

summary(fit.cauchit)
```

```
Call:

glm(formula = qualified ~ T.TOTAL, family = binomial(link = "cauchit"),

    data = mydata)


Deviance Residuals:

    Min       1Q   Median       3Q      Max

-1.9234  -0.5356   0.3693   0.5119   1.5872


Coefficients:

             Estimate Std. Error z value Pr(>|z|)

(Intercept) -40.88668   26.84313  -1.523    0.128

T.TOTAL       0.12109    0.07969   1.519    0.129


(Dispersion parameter for binomial family taken to be 1)


    Null deviance: 30.553  on 23  degrees of freedom

Residual deviance: 17.943  on 22  degrees of freedom
```

```
AIC: 21.943

Number of Fisher Scoring iterations: 9
```

Προκύπτει λοιπόν ότι το εκτιμώμενο μοντέλο είναι :

cauchit($\hat{\pi}$) ≡ tan($\hat{\pi}$ ($\hat{\pi}$ − 0.5) = -40.88 + 0.12*T.TOTAL

που σημαίνει ότι η εκτιμώμενη πιθανότητα είναι :

και $\hat{\pi}$ = 0.5 + $\pi^{-1}$arctan (-40.88 + 0.12*T.TOTAL)

Ο έλεγχος Wald για την επεξηγηματική μεταβλητή ριμπάουντ μας λέει ότι δεν είναι στατιστικά σημαντική :

$\frac{\hat{b}}{\hat{\sigma}(\hat{b})} = \frac{0.12109}{0.07969}$ = 1.519, και p-value = 0.129 > 5%

Η τελική απόκριση του μοντέλου 17.943 κάτι το οποίο είναι ικανοποιητικό καθώς ένα καλό μοντέλο θα πρέπει να έχει μικρή τιμή.

Τέλος, ο αλγόριθμος μεγιστοποίησε το λογάριθμο της πιθανοφάνειας μετά από 9 επαναλήψεις.

**4.6.6 Εφαρμογή cloglog μοντέλου για 2014-2015**

```
fit.cloglog<-
glm(qualified~T.TOTAL,family=binomial(link="cloglog"),data=mydata)

summary(fit.cloglog)
```

```
Call:

glm(formula = qualified ~ T.TOTAL, family = binomial(link = "cloglog"),

    data = mydata)


Deviance Residuals:

     Min       1Q   Median       3Q      Max

-1.74559  -0.47717   0.01631   0.48303   1.64546
```

Coefficients:

| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | -25.71933 | 10.63973 | -2.417 | 0.0156 | * |
| T.TOTAL | 0.07405 | 0.03049 | 2.429 | 0.0152 | * |

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 30.553 on 23 degrees of freedom

Residual deviance: 16.266 on 22 degrees of freedom

AIC: 20.266

Number of Fisher Scoring iterations: 8

Προκύπτει λοιπόν ότι το εκτιμώμενο μοντέλο είναι :

cloglog($\hat{\pi}$) ≡ log(- log (1 -$\hat{\pi}$)) = -25.71 + 0.07*T.TOTAL

που σημαίνει ότι η εκτιμώμενη πιθανότητα είναι :

και $\hat{\pi}$ = 1 – exp(- exp (-25.71 + 0.07*T.TOTAL))

Ο έλεγχος Wald για την επεξηγηματική μεταβλητή ριμπάουντ μας λέει ότι είναι στατιστικά σημαντική :

$\frac{\hat{b}}{\hat{\sigma}(\hat{b})} = \frac{0.07404}{0.03049}$ = 2.42 , και p-value = 0.0152 < 5%

Η τελική απόκριση του μοντέλου 16.266 και ο αλγόριθμος μεγιστοποίησε το λογάριθμο της πιθανοφάνειας μετά από 8 επαναλήψεις.

Συγκεντρωτικά, βλέπουμε πως τα μοντέλα logit, probit και cloglog μας οδηγούν στο ίδιο συμπέρασμα, καθώς η μηδενική μας υπόθεση απορρίπτεται και η μεταβλητή που εξετάζουμε είναι στατιστικά σημαντική. Αντίθετα το μοντέλο cauchit δεν απορρίπτει

την μηδενική μας υπόθεση και δεν μπορούμε να θεωρήσουμε ότι η μεταβλητή «συνολικά ριμπάουντ» είναι στατιστικά σημαντική για το μοντέλο μας.

Στη συνέχεια για να πάρουμε μια πρώτη εικόνα για το ποιο από τα 4 μοντέλα προσαρμόζεται καλύτερα στα δεδομένα μας θα ελέγξουμε τις αντίστοιχες στατιστικές συναρτήσεις χ-τετράγωνο του pearson εκτελώντας τις παρακάτω εντολές στην R:

```
X2.logit=sum(residuals(fit.logit,type="pearson")^2)

X2.probit=sum(residuals(fit.probit,type="pearson")^2)

X2.cauchit=sum(residuals(fit.cauchit,type="pearson")^2)

X2.cloglog=sum(residuals(fit.cloglog,type="pearson")^2)

X2.all=c(X2.logit,X2.probit,X2.cauchit,X2.cloglog)

X2.table=rbind(X2.all,1-pchisq(X2.all,df=fit.logit$df.residual))

colnames(X2.table)=c("logit","probit","cauchit","cloglog")

rownames(X2.table)=c("X2","p")

round(X2.table,4)
```

```
      logit    probit   cauchit  cloglog
X2  15.3534  15.0087  17.5442  14.4900
p    0.8466   0.8619   0.7326   0.8832
```

Έτσι λοιπόν βλέπουμε ότι το μοντέλο complementary log log δείχνει να ταιριάζει καλύτερα από τα υπόλοιπα 3 στα δεδομένα μας.

Το ότι το cloglog μοντέλο προσαρμόζεται καλύτερα μπορούμε να το επιβεβαιώσουμε και από τις αποκλίσεις( Deviance =16.266).

Το μοντέλο με την ελάχιστη απόκλιση μεταξύ των 4 έχει και τη μέγιστη πιθανοφάνεια. Μια και τα 4 μοντέλα έχουν ουσιαστικά 2 παραμέτρους, την $\beta_0$ και $\beta_1$, θα μπορούσαμε να πούμε ότι το cloglog μεγιστοποιεί την πιθανοφάνεια της παραμέτρου $(M,\beta_0,\beta_1)$ στον παραμετρικό χώρο {logit, probit, cauchit, cloglog}x R x R.

Αξίζει να συμπληρώσουμε ότι το complementary log log μοντέλο ταιριάζει σε περιπτώσεις που η P ( Y =1) πλησιάζει γρήγορα τη μονάδα αλλά αργά το μηδέν.

### 4.6.7 Μέτρα Προσαρμογής

Στο κλασσικό γραμμικό μοντέλο το κυριότερο μέτρο προσαρμογής είναι ο συντελεστής προσδιορισμού $R^2$ ,ο οποίος παίρνει τιμές στο διάστημα (0,1). Μεγάλες τιμές του συντελεστή προσδιορισμού δείχνουν καλή προσαρμογή ενώ μικρές τιμές του όχι. Τα κριτήρια που χρησιμοποιούμε για να ορίσουμε αυτές τις τιμές ποικίλλουν αλλά συνήθως μεγάλες τιμές θεωρούμε αυτές που είναι πάνω από 0.70-0.80. Ερμηνεύεται ως το ποσοστό της συνολικής μεταβλητότητας της μεταβλητής απόκρισης που εξηγείται από το μοντέλο.

Όταν σε ένα μοντέλο περιλαμβάνουμε περισσότερες μεταβλητές συνήθως ο συντελεστής προσδιορισμού αυξάνεται για αυτό χρησιμοποιούμε τον προσαρμοσμένο (ή τροποποιημένο) συντελεστή προσδιορισμού ο οποίος ορίζεται ως εξής :

$R^2_{adj}$ = 1 - $\frac{SSE/(n-m)}{SST/(n-1)}$

Στα γενικευμένα γραμμικά μοντέλα για δίτιμα δεδομένα δεν υπάρχει ένα και μοναδικό τέτοιο μέτρο αλλά πολλά. Θεωρούμε ότι $L_M$ είναι η μέγιστη πιθανοφάνεια υπό το μοντέλο M και $L_0$ η μέγιστη πιθανοφάνεια του μοντέλου που δεν χρησιμοποιεί καμία επεξηγηματική μεταβλητή. Έτσι λοιπόν, κάποια άλλα μέτρα καλής προσαρμογής που χρησιμοποιούνται αρκετά συχνά είναι :

Πηγή: (Γ. Ηλιόπουλος, Πειραιάς 2019. Γενικευμένα γραμμικά μοντέλα, Πανεπιστημιακές Σημειώσεις.)

**1. McFadden**

Είναι το πιο δημοφιλές ψευδο-$R^2$. Τιμές μεταξύ 0.2 και 0.4 δείχνουν καλή προσαρμογή ή με άλλα λόγια ότι το συγκεκριμένο μοντέλο έχει πολύ καλύτερη προσαρμογή από το αντίστοιχο που έχει μόνο τη σταθερά.

Δίνεται από τον τύπο $R^2 = 1 - \frac{logL_M}{logL_0}$

**2.McFadden adjusted**

Δίνεται από τον τύπο $R^2 = 1 - \frac{logL_M - m}{logL_0}$

**3.Cox & Snell**

Ο λόγος των πιθανοφανειών μας δείχνει τη βελτίωση του μοντέλου μας σε σχέση με αυτό που έχει μόνο τη σταθερά. Όσο μικρότερος είναι ο λόγος, τόσο μεγαλύτερη βελτίωση θα υπάρχει.

Δίνεται από τον τύπο $R^2 = 1 - (L_0/L_M)^{2/n}$

**4.Nagelkerke/Cragg & Uhler**

Ουσιαστικά είναι το $R^2$ των Cox & Snell διαιρεμένο με την μέγιστη τιμή του και δίνεται από τον τύπο

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

**5.McKelvey & Zavoina**

Δίνεται από τον τύπο $R^2 = \frac{\sigma^2(\hat{z}_k)}{\sigma^2(\hat{z}_k) + \sigma^2(\hat{U}_k))}$

**6.Efron Pseudo $R^2$**

Δίνεται από τον τύπο $R^2_{efron}$ = 1 - $\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ και αποτελεί μία προέκταση για την μεταβλητότητα που εξηγείται σε ένα μοντέλο παλινδρόμησης.

**7.AIC (Akaike Information Criterion)**

Το κριτήριο πληροφορίας του Akaike είναι αρκετά δημοφιλές κριτήριο επιλογής μοντέλου και όχι μόνο για δίτιμες αποκρίσεις. Επιβάλλει ποινή για το πλήθος m των παραμέτρων. Όσο μικρότερο τόσο καλύτερα. Δίνεται από τον τύπο

$AIC = -2\log L_M + 2m$

**8.$AIC_C$ (Corrected AIC)**

Η συγκεκριμένη διόρθωση έχει νόημα μόνο όταν τα σφάλματα είναι κανονικά κατανεμημένα και δίνεται από τον τύπο

$AIC_C = AIC + \frac{2m(m+1)}{n-m-1}$

**9.BIC(Bayesian Information Criterion)**

Είναι γνωστό και ως κριτήριο πληροφορίας του Schawrz. Όπως και στο AIC επιλέγουμε την min τιμή του. Δίνεται από τον τύπο

$BIC = -2\log L_M + m\log n$ όπου n είναι το πλήθος των παρατηρήσεων

Το AIC αντανακλά τον κίνδυνο ένα μοντέλο να υπερπροσαρμοστεί (overfitting), ενώ το BIC να υποπροσαρμοστεί (underfitting). Οι απόψεις των ερευνητών για το ποιο πληροφοριακό κριτήριο δίνει καλύτερα και πιο αξιόπιστα αποτελέσματα διαφέρουν.

Γενικά τα κριτήρια δεν επιλέγουν πάντα τα ίδια μοντέλα. Αν και είναι ιδιαίτερα χρήσιμα σε περιπτώσεις όπου έχουμε περισσότερες από μια επεξηγηματικές μεταβλητές κρίνεται σημαντικό να τα εξετάσουμε και στην περίπτωσή μας έχοντας σαν ερμηνευτική μεταβλητή τα συνολικά ριμπάουντ. Έτσι λοιπόν παρακάτω παρουσιάζονται τα μέτρα προσαρμογής για μοντέλα logit, probit, cauchit και cloglog που προσαρμόσαμε :

```
library(BaylorEdPsych)
```

```
PseudoR2(fit.logit)
PseudoR2(fit.probit)
PseudoR2(fit.cauchit)
PseudoR2(fit.cloglog)
```

➔ Για το μοντέλο logit :

```
 McFadden     Adj.McFadden       Cox.Snell       Nagelkerke
    0.4577588       0.2613767       0.4416336       0.6133651
McKelvey.Zavoina          Effron           Count       Adj.Count
    0.7236189       0.4771114       0.8333333       0.5000000
         AIC    Corrected.AIC
   20.5669216       21.1383501
```

➔ Για το μοντέλο probit :

```
     McFadden     Adj.McFadden       Cox.Snell       Nagelkerke
    0.4661295       0.2697474       0.4475521       0.6215850
McKelvey.Zavoina          Effron           Count       Adj.Count
    0.7568798       0.4799487       0.8333333       0.5000000
         AIC    Corrected.AIC
   20.3111735       20.8826021
```

➔ Για το μοντέλο cauchit :

```
     McFadden     Adj.McFadden       Cox.Snell       Nagelkerke
    0.4127180       0.2163359       0.4086822       0.5676003
McKelvey.Zavoina          Effron           Count       Adj.Count
          NA       0.4640551       0.8333333       0.5000000
         AIC    Corrected.AIC
```

| |
|---|
| 21.9430400 22.5144685 |

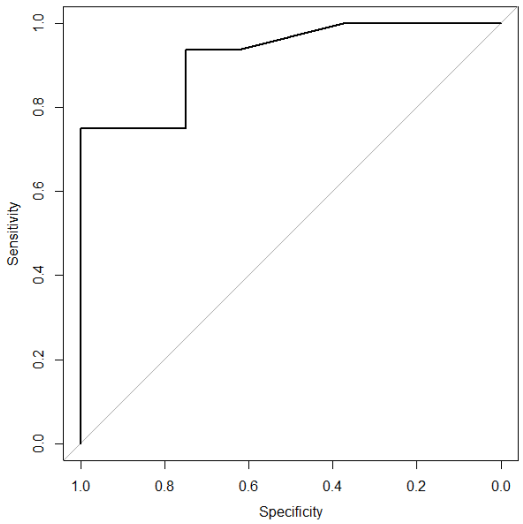
➔ Για το μοντέλο cloglog :

| McFadden | Adj.McFadden | Cox.Snell | Nagelkerke |
|---|---|---|---|
| 0.4676048 | 0.2712227 | 0.4485886 | 0.6230246 |
| McKelvey.Zavoina | Effron | Count | Adj.Count |
| NA | 0.4777902 | 0.8333333 | 0.5000000 |
| AIC | Corrected.AIC | | |
| 20.2661003 | 20.8375289 | | |

Γενικά παρατηρούμε ότι το ψευδο-$R^2$ του McFadden είναι μεγαλύτερο του 0.4 κάτι το οποίο δείχνει πολύ καλή προσαρμογή και για τα 4 μοντέλα. Φαίνεται ότι το μοντέλο cloglog προσαρμόζεται λίγο καλύτερα συγκριτικά με τα υπόλοιπα.

### 4.6.8 Προβλεπτική ισχύς του μοντέλου

Ένα άλλο μέτρο που δείχνει την προβλεπτική ικανότητα του μοντέλου είναι η καμπύλη ROC. Πιο συγκεκριμένα για το μοντέλο cloglog που επιλέξαμε έχουμε :




Σχήμα 4.4 : Καμπύλη ROC για το cloglog μοντέλο

Η διαγώνιος είναι η καμπύλη ROC για το μοντέλο που προβλέπει Y =1 και Y=0 με πιθανότητες 50-50, ανεξάρτητα από τις τιμές των επεξηγηματικών μεταβλητών.

Προφανώς, όσο πιο ψηλά είναι η καμπύλη ενός μοντέλου (δηλαδή όσο περισσότερο απέχει από την διαγώνιο) τόσο καλύτερη προβλεπτική ικανότητα θα έχει το μοντέλο που έχουμε κατασκευάσει. Στην περίπτωση μας το εμβαδόν κάτω από την καμπύλη (area under the curve, AUC) ισούται με 0.9219 το οποίο κρίνεται αρκετά ικανοποιητικό για την προβλεπτική ικανότητα του complementary log log μοντέλου. Το συγκεκριμένο μέτρο αναφέρεται και ως δείκτης συμφωνίας (concordance index). Μια δημοφιλής μέθοδος για την επιλογή του βέλτιστου κατωφλιού $\pi_0$ είναι η μέθοδος Youden. Ο δείκτης J του Youden είναι η ποσότητα :

$J = J(\pi_0) = \text{ευαισθησία}(\pi_0) - \text{ειδικότητα}(\pi_0) - 1,$

και η μέθοδος επιλέγει το $\pi_0$ που τον μεγιστοποιεί.

Έτσι λοιπόν το μοντέλο που εφαρμόσαμε είναι αρκετά καλύτερο από ένα τυχαίο μοντέλο που έχει AUC = 0.5 και συνεπώς οι προβλέψεις με τις πραγματικές παρατηρήσεις θα είναι πολύ κοντά. Το cutoff σημείο ισούται με 0.8055, είναι πιο κοντά στην πάνω αριστερή γωνία του σχήματος και θεωρείται βέλτιστο, όσον αφορά το αν η ομάδα προκρίθηκε ή όχι στην επόμενη φάση.

### 4.6.9 Αξιολόγηση της προβλεπτικής ικανότητας του μοντέλου

Οι πίνακες ταξινόμησης (classification tables) είναι ένας από τους πιο διαδεδομένους τρόπους για να αξιολογήσουμε την προβλεπτική ικανότητα του μοντέλου που χρησιμοποιούμε. Αν η εκτιμώμενη πιθανότητα είναι ίση ή μεγαλύτερη συνήθως του 0.50 (cutoff) τότε η ομάδα έχει προκριθεί (Y = 1) αλλιώς δεν έχει προκριθεί ( Y = 0 ). Προφανώς το ποσοστό ορθής ταξινόμησης θα πρέπει να είναι αρκετά μεγαλύτερο από μία πρόβλεψη που θα επιτύχουμε από τύχη.

```
qualified.fit=1*(pred.fit>0.5)

classification.matrix=matrix(nrow=2,byrow=T,

+c(sum(qualified*qualified.fit),sum(qualified*(1-qualified.fit)),

+sum((1-qualified)*qualified.fit),sum((1-qualified)*(1-qualified.fit))))

rownames(classification.matrix)=c("y=1","y=0")

colnames(classification.matrix)=c("y.hat=1","y.hat=0")

classification.matrix

sum(qualified*qualified.fit)/sum(qualified) #sensitivity

sum((1-qualified)*(1-qualified.fit))/sum(1-qualified) #specificity

mean(qualified==qualified.fit)
```

| CLASSIFICATION TABLE | | | | | |
|---|---|---|---|---|---|
| | | PREDICTED | | | |
| | QUALIFIED | 0 | 1 | | %CORRECT |
| OBSERVED | 0 | 6 | 2 | 8 | 75% |
| | 1 | 2 | 14 | 16 | 87.5% |
| OVERALL | | 8 | 16 | 24 | 83.33% |

Πίνακας 4.3 : classification table με cutoff = 0.5

Έτσι λοιπόν βλέπουμε ότι περίπου το 83% των αποτελεσμάτων προβλέφθηκε σωστά από το μοντέλο, δηλαδή 4 ομάδες από τις 24 δεν κατηγοριοποιήθηκαν σωστά. Το ποσοστό ορθής ταξινόμησης προκύπτει και από τους παραπάνω πίνακες καλής προσαρμογής και πιο συγκεκριμένα είναι το μέτρο count που επιστρέφει η συνάρτηση PseudoR2. Πιο συγκεκριμένα από τις 8 ομάδες που δεν προκρίθηκαν στο top-16 εκτιμήθηκαν σωστά οι 6, ποσοστό 75% (specificity) και από τις 16 ομάδες που προκρίθηκαν εκτιμήθηκαν σωστά οι 14, ποσοστό 87,5% (sensitivity). Σύμφωνα με αρκετές έρευνες το ποσοστό της ορθής ταξινόμησης μπορεί να θεωρηθεί ικανοποιητικό αν είναι κατά 25% μεγαλύτερο από την κατά τύχη πιθανότητα κάποιος να βρει τις ομάδες που θα προκριθούν ή όχι στην επόμενη φάση.

(Πηγή : Srivastava T N, Rego S. (2011). Business research methodology.)

Η κατά τύχη πιθανότητα στην περίπτωσή μας μπορεί να υπολογιστεί σταθμίζοντας τις 2 κατηγορίες της μεταβλητής qualified. Θεωρώντας ότι το ποσοστό πρόκρισης είναι 66% (4 από τις 6 ομάδες) και το ποσοστό αποκλεισμού είναι 34% (2 από τις 6 ομάδες) έχουμε ότι :

**P(random choice) = $0.66^2$ + $0.34^2$ = 0.4356 + 0.1156 = 0.5512 ή 55%**

Με άλλα λόγια κάποιος φίλαθλος μπορεί, χωρίς να χρησιμοποιήσει κάποιο μοντέλο λογιστικής παλινδρόμησης να προβλέψει αν η ομάδα πέρασε από την regular season, με πιθανότητα περίπου 55 %. Έτσι λοιπόν, σύμφωνα με τον κανόνα που θέσαμε παραπάνω το accuracy του μοντέλου που κατασκευάσαμε θα πρέπει να είναι μεγαλύτερο του 55.12 % + 25% = 80.12 %. Αφού στην περίπτωση μας βρέθηκε ότι το ποσοστό σωστής ταξινόμησης ισούται περίπου με 83% δεν προκύπτει κάποιο πρόβλημα για το μοντέλο μας.

Αξίζει να σημειώσουμε πως επιλέξαμε ότι το κατώφλι ισούται με 0.50. Μία άλλη συνηθισμένη επιλογή είναι το ποσοστό των Y =1 στα δεδομένα μας, δηλαδή το ποσοστό πρόκρισης (66%). Έτσι λοιπόν ορίζοντας $\pi_0 = 0.66$ έχουμε :

| CLASSIFICATION TABLE | | | | | |
|---|---|---|---|---|---|
| | | PREDICTED | | | |
| | QUALIFIED | 0 | 1 | | %CORRECT |
| OBSERVED | 0 | 6 | 2 | 8 | 75% |
| | 1 | 3 | 13 | 16 | 81.25% |
| OVERALL | | 9 | 15 | 24 | 79.16% |

Πίνακας 4.4 : classification table με cutoff = 0.66

Παρατηρούμε πως δεν παίρνουμε τα επιθυμητά αποτελέσματα καθώς το accuracy του μοντέλου είναι μικρότερο του 80.12% και πιο συγκεκριμένα ισούται με 79.16%.

## 4.7 Εφαρμογή για την χρονιά 2015-2016

### 4.7.1 Ποιες μεταβλητές είναι στατιστικά σημαντικές για την χρονιά 2015-2016?

| Qualified~ | AIC | p-value |
|---|---|---|
| **Pts.Total** | 23.156 | **0.0171*** |
| **Dif_Points** | 14.444 | 0.0836 |
| **X2** | 20.325 | **0.0333*** |
| **X3** | 32.541 | 0.204 |
| **FT** | 34.247 | 0.584 |
| **O** | 32.114 | 0.140 |
| **D** | 23.952 | **0.0138*** |
| **T** | 23.354 | **0.0240*** |
| **As** | 31.2 | 0.111 |
| **St** | 34.429 | 0.726 |
| **To** | 34.501 | 0.820 |
| **Fv** | 31.149 | 0.0977 |
| **Ag** | 34.492 | 0.805 |

| | | |
|---|---|---|
| **Cm** | 33.605 | 0.342 |
| **Rv** | 26.563 | **0.027*** |
| **PIR** | 18.309 | **0.0201*** |

Πίνακας 4.5 : Επιρροή της κάθε μεταβλητής ξεχωριστά

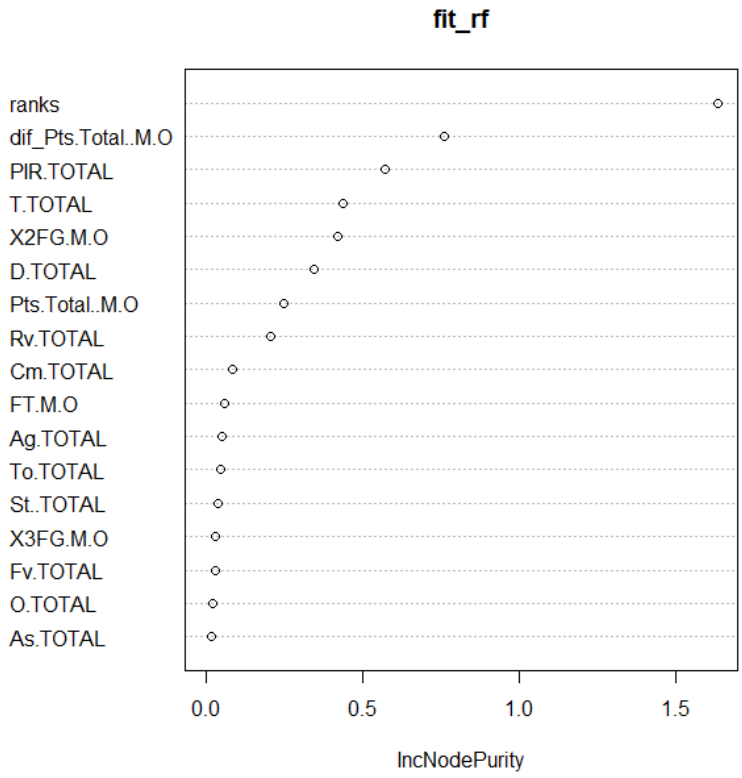
Διαπιστώνουμε ότι όπως και την προηγούμενη χρονιά τα συνολικά και τα αμυντικά ριμπάουντ έχουν καθοριστικό ρόλο στην έκβαση του τελικού αποτελέσματος ( p-value = 0.0240 και 0.0138 αντίστοιχα). Επιπλέον παρατηρούμε πως για την χρονιά 2015-2016 σημαντικό ρολό για το αν η ομάδα προκρίθηκε ή όχι στην επόμενη φάση φαίνεται πως έχει και το ποσοστό ευστοχίας στα σουτ 2 πόντων που πραγματοποιήθηκαν.

### 4.7.2 Επιλογή χαρακτηριστικών για 2015-2016

Όμοια, όπως και για την προηγούμενη χρονιά θα ελέγξουμε με τις μεθόδους random forest, stepwise regression, Boruta και anova ποιες μεταβλητές φαίνονται να είναι σημαντικές ώστε να τις συμπεριλάβουμε στα μοντέλα που θα κατασκευάσουμε. Σημειώνουμε ξανά πως οι μεταβλητές ranks, points, dif_points και PIR δεν μπορούν να συμπεριληφθούν στο μοντέλο μας καθώς είτε προβλέπουν σχεδόν απόλυτα αν η ομάδα προκρίθηκε στην επόμενη φάση είτε παρέχουν συνδυαστική πληροφορία για παραπάνω από μία μεταβλητή. Παρόλο αυτά στις παρακάτω μεθόδους παρουσιάζονται τα πλήρη αποτελέσματα καθώς είναι ενσωματωμένες και αυτές οι μεταβλητές.

**α.Random Forest**

Οι 2 πιο σημαντικές μεταβλητές που πρέπει να εισαχθούν στο μοντέλο μας όπως φαίνεται από το παρακάτω σχήμα είναι τα συνολικά ριμπάουντ καθώς και το ποσοστό ευστοχίας στα σουτ δύο πόντων, εξαιρώντας τις μεταβλητές που αναφέραμε προηγουμένως.




Σχήμα 4.5 : Γραφική αναπαράσταση της σημαντικότητας των μεταβλητών με τη μέθοδο random forest.

**β. Μέθοδος Stepwise**

Εκτελώντας τις ακόλουθες εντολές στην R, προκύπτει ότι τα αμυντικά ριμπάουντ είναι σημαντικά για το μοντέλο που θα προσαρμόσουμε.

```
varImpPlot(fit_rf)

base.mod <- glm(qualified ~ 1 ,family=binomial, data= mydata2[,-1])  # base intercept only model

all.mod <- glm(qualified ~ . ,family=binomial, data= mydata2[,-1]) # full model with all predictors

stepMod <- step(base.mod, scope = list(lower = base.mod, upper = all.mod), direction = "both", trace = 0, steps = 1000)  # perform step-wise algorithm

shortlistedVars <- names(unlist(stepMod[[1]])) # get the shortlisted variable.

shortlistedVars <- shortlistedVars[!shortlistedVars %in% "(Intercept)"]  # remove intercept
```

```
print(shortlistedVars)

[1] "ranks"   "D.TOTAL"
```

### γ. Μέθοδος Boruta

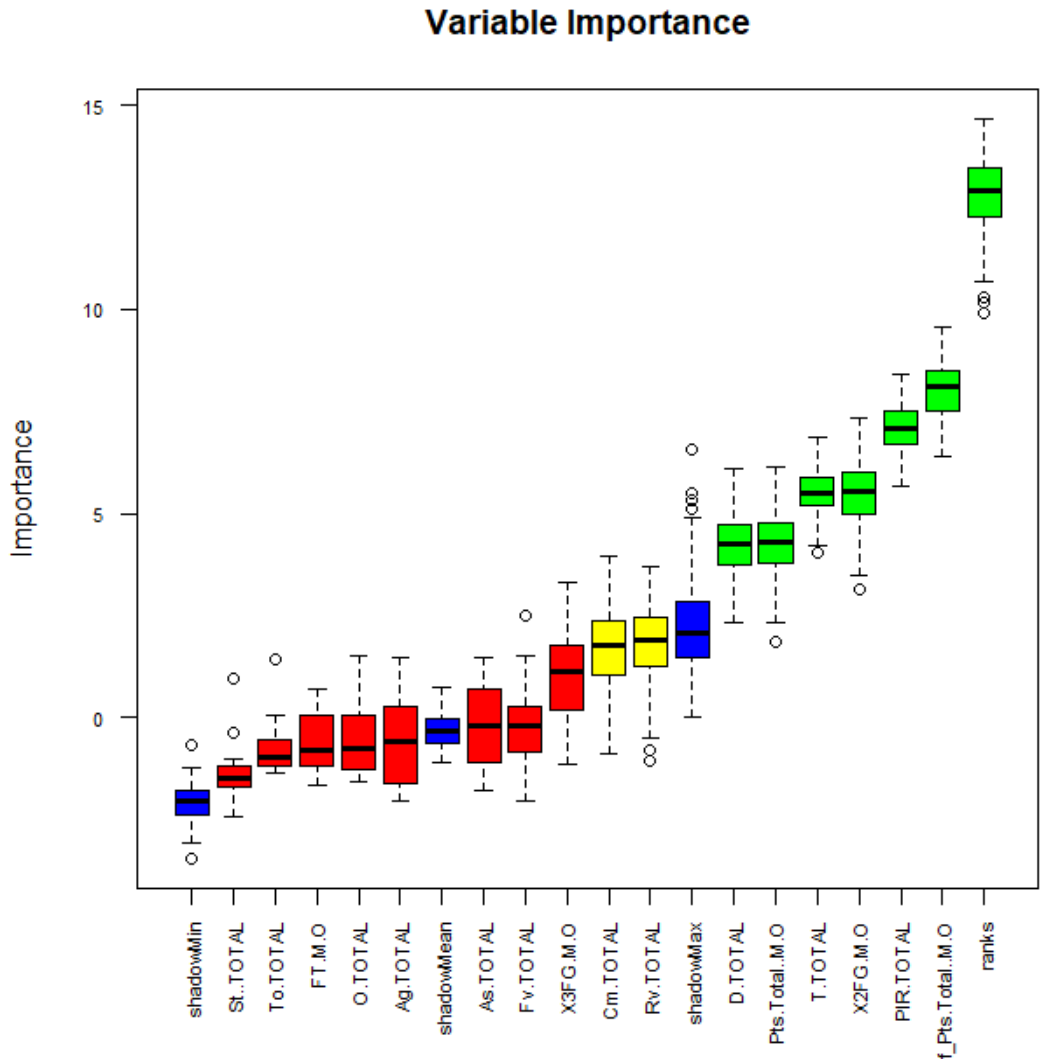
Παρατηρούμε ότι έγινε μια ταξινόμηση των μεταβλητών σε 3 επίπεδα, τα οποία στο σχήμα δηλώνονται με τα αντίστοιχα χρώματα:  πράσινο, κόκκινο, κίτρινο.

Με πράσινο χρώμα είναι οι μεταβλητές που έχουν γίνει confirmed από τον random forest (RF) classifier. Έτσι λοιπόν σύμφωνα με το σχήμα **και σε σύγκριση με το δεύτερο κεφάλαιο**, παρατηρούμε πως οι πιο σημαντικές μεταβλητές, αν εξαιρέσουμε την μεταβλητή ranks και το PIR, είναι η διαφορά των πόντων, τα συνολικά ριμπάουντ και το ποσοστό ευστοχίας στα σουτ 2 πόντων.

Με κόκκινο χρώμα παρουσιάζονται τα features που δεν έχουν εγκριθεί (p-value > 0.05 ) από τον RF. Χαρακτηριστικά, βλέπουμε ότι έχουν ¨αρνητική σημαντικότητα¨ στο σχήμα.

Με κίτρινο χρώμα απεικονίζονται οι λεγόμενες tentative μεταβλητές, που στην περίπτωσή μας είναι η Rv και η Cm.

Τέλος, με μπλε χρώμα δηλώνονται τα shadow features, όπως αναφέραμε και για την προηγούμενη χρονιά.




Σχήμα 4.6 : Γραφική αναπαράσταση της σημαντικότητας των μεταβλητών με τη μέθοδο Boruta.

### 4.7.3 Εφαρμογή logit μοντέλου για 2015-2016

Αφού ελέγξαμε την σημαντικότητα των μεταβλητών μας και μετά από τη εφαρμογή πολλαπλών μεθόδων ώστε να διαπιστώσουμε ποιες μεταβλητές προβλέπουν καλύτερα αν μια ομάδα προκρίθηκε στην επόμενη φάση ή όχι είμαστε πλέον σε θέση να δημιουργήσουμε το τελικό μοντέλο που θα χρησιμοποιήσουμε. Έτσι λοιπόν συνδυάζοντας όλα τα παραπάνω επιλέξαμε να φτιάξουμε ένα λογιστικό μοντέλο με ερμηνευτικές τα συνολικά ριμπάουντ και το ποσοστό ευστοχίας στα σουτ 2 πόντων, ώστε να διερευνήσουμε με μεγαλύτερη ακρίβεια αν όντως επιδρούν σημαντικά στην εξέλιξη της διοργάνωσης.

```
Call:
glm(formula = qualified ~ X2FG.M.O + T.TOTAL, family = binomial,
    data = mydata2)
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-1.99313  -0.05733   0.02910   0.17475   1.09599
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -117.21074   59.63744  -1.965   0.0494 *
X2FG.M.O       1.30521    0.71963   1.814   0.0697 .
T.TOTAL        0.15473    0.09112   1.698   0.0895 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 30.5527  on 23  degrees of freedom
Residual deviance:  6.8739  on 21  degrees of freedom
AIC: 12.874
Number of Fisher Scoring iterations: 8
```

Προκύπτει λοιπόν ότι το εκτιμώμενο μοντέλο είναι :

$\text{logit}(\hat{\pi}) \equiv \log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -117.21 + 0.15*\text{T.TOTAL} + 1.3*\text{X2FG}$

που σημαίνει ότι η εκτιμώμενη σχετική πιθανότητα είναι :

$\frac{\hat{\pi}}{1-\hat{\pi}} = e^{-117.21 + 0.15*\text{T.TOTAL} + 1.3*\text{X2FG}}$ και $\hat{\pi} = \frac{e^{-117.21 + 0.15*\text{T.TOTAL} + 1.3*\text{X2FG}}}{1 + e^{-117.21 + 0.15*\text{T.TOTAL} + 1.3*\text{X2FG}}}$

Δηλαδή για κάθε μοναδιαία αύξηση των συνολικών ριμπάουντ, ο λογάριθμος της σχετικής πιθανότητας για το αν η ομάδα θα προκριθεί ή όχι αυξάνεται κατά περίπου 0.15 μονάδες και για κάθε μοναδιαία αύξηση των ποσοστών ευστοχίας στα σούτ 2 πόντων ο λογάριθμος της σχετικής πιθανότητας για το αν η ομάδα θα προκριθεί ή όχι αυξάνεται κατά περίπου 1.3 μονάδες

Ο έλεγχος Wald για την επεξηγηματική μεταβλητή ριμπάουντ μας λέει ότι είναι στατιστικά σημαντική σε επίπεδο σημαντικότητας 10% αφού p-value = 0.0895 < 10%

Ο έλεγχος Wald για την επεξηγηματική μεταβλητή X2FG μας λέει ότι είναι στατιστικά σημαντική σε επίπεδο σημαντικότητας 10% αφού p-value = 0.0697 < 10%

Και επειδή $\hat{b} > 0$ συμπεραίνουμε ότι όσα περισσότερα είναι τα συνολικά ριμπάουντ και όσο μεγαλύτερα είναι τα ποσοστά ευστοχίας στα σουτ 2 πόντων τόσο αυξάνεται η πιθανότητα η ομάδα να προκριθεί στο top-16.

Η τελική απόκριση του μοντέλου 6.8736 κάτι το οποίο είναι ικανοποιητικό καθώς ένα καλό μοντέλο θα πρέπει να έχει μικρή τιμή. Τέλεια προσαρμογή υπάρχει όταν η απόκλιση είναι μηδενική προφανώς.

Τέλος, ο αλγόριθμος μεγιστοποίησε το λογάριθμο της πιθανοφάνειας μετά από 8 επαναλήψεις.

Ακόμη για την χρονιά 2015-2016 αξίζει να ελέγξουμε και αν είναι στατιστικά σημαντική η αλληλεπίδραση αυτών των 2 μεταβλητών μέσα στο μοντέλο που έχουμε κατασκευάσει.

```
Call:

glm(formula = qualified ~ X2FG.M.O * T.TOTAL, family = binomial,

    data = mydata2)

Deviance Residuals:

     Min       1Q   Median       3Q      Max

-1.83035  -0.00037   0.00094   0.14385   1.05508

Coefficients:

                    Estimate      Std. Error    z value    Pr(>|z|)

(Intercept)        -2.024e+03    1.825e+03     -1.109      0.267

X2FG.M.O            3.944e+01    3.597e+01      1.096      0.273

T.TOTAL             5.420e+00    4.917e+00      1.102      0.270

X2FG.M.O:T.TOTAL -1.053e-01  9.678e-02  -1.088    0.277

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 30.5527  on 23  degrees of freedom

Residual deviance:  5.5548  on 20  degrees of freedom
```

AIC: 13.555

Number of Fisher Scoring iterations: 10

Όπως βλέπουμε η αλληλεπίδραση μεταξύ των 2 μεταβλητών δεν είναι στατιστικά σημαντική καθώς το αντίστοιχο p-value ισούται με 0.277 > 10% άρα δεν κρίνεται σημαντική η είσοδος της αλληλεπίδρασης στο μοντέλο μας όσον αφορά τα συνολικά ριμπάουντ και το ποσοστό ευστοχίας στα σουτ 2 πόντων.

### 4.7.4 Εφαρμογή probit μοντέλου για 2015-2016

Call:

glm(formula = qualified ~ X2FG.M.O + T.TOTAL, family = binomial(link = "probit"),

data = mydata2)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.93656 -0.01436 0.00324 0.13392 1.13108

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -67.97647 31.27669 -2.173 0.0298 *

X2FG.M.O 0.75631 0.37278 2.029 0.0425 *

T.TOTAL 0.08970 0.04908 1.828 0.0676 .

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 30.5527 on 23 degrees of freedom

Residual deviance: 6.7043 on 21 degrees of freedom

AIC: 12.704

Number of Fisher Scoring iterations: 10

Προκύπτει λοιπόν ότι το εκτιμώμενο μοντέλο είναι :

probit($\hat{\pi}$) ≡ $\Phi^{-1}$ ($\hat{\pi}$) = -67.97 + 0.08*T.TOTAL + 0.75*X2FG

που σημαίνει ότι η εκτιμώμενη πιθανότητα είναι :

και $\hat{\pi}$ = Φ (-67.97 + 0.08*T.TOTAL + 0.75*X2FG)

Ο έλεγχος Wald για την επεξηγηματική μεταβλητή ριμπάουντ μας λέει ότι είναι στατιστικά σημαντική αφού p-value = 0.0676 < 10% και αντίστοιχα για τα ποσοστά ευστοχίας στα σουτ 2 πόντων αφού το p-value (= 0.0425) είναι μικρότερο και σε 5% και σε 10% επίπεδο στατιστικής σημαντικότητας.

Και επειδή $\hat{b} > 0$ συμπεραίνουμε ότι όσα περισσότερα είναι τα συνολικά ριμπάουντ και τα σουτ 2 πόντων αντίστοιχα τόσο αυξάνεται η πιθανότητα η ομάδα να προκριθεί στο top-16.

Η τελική απόκριση του μοντέλου 6.7043 κάτι το οποίο είναι αρκετά ικανοποιητικό.

### 4.7.5 Εφαρμογή cauchit μοντέλου για 2015-2016

Ο έλεγχος Wald και για τις 2 επεξηγηματικές μεταβλητές μας λέει ότι δεν είναι στατιστικά σημαντικές ( p-value = 0.412 και 0.434 αντίστοιχα)

Συνεπώς, και στις 2 χρονιές το μοντέλο cauchit δεν δίνει ικανοποιητικά αποτελέσματα και δεν μπορεί να χρησιμοποιηθεί για την εξαγωγή ασφαλών συμπερασμάτων.

### 4.7.6 Εφαρμογή cloglog μοντέλου για 2015-2016

Call:

glm(formula = qualified ~ X2FG.M.O + T.TOTAL, family = binomial(link = "cloglog"),

data = mydata2)

Deviance Residuals:

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -1.83824 | -0.12428 | 0.00000 | 0.05369 | 1.29644 |

Coefficients:

| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | -75.89144 | 37.46108 | -2.026 | 0.0428 | * |
| X2FG.M.O | 0.82084 | 0.41200 | 1.992 | 0.0463 | * |

T.TOTAL 0.10172 0.06019 1.690 0.0910 .

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 30.5527 on 23 degrees of freedom

Residual deviance: 7.1314 on 21 degrees of freedom

AIC: 13.131

Number of Fisher Scoring iterations: 11

Προκύπτει λοιπόν ότι το εκτιμώμενο μοντέλο είναι :

cloglog($\hat{\pi}$) ≡ log(- log (1 - $\hat{\pi}$)) = - 75.89 + 0.10*T.TOTAL + 0.82*X2FG

που σημαίνει ότι η εκτιμώμενη πιθανότητα είναι :

και $\hat{\pi}$ = 1 – exp(- exp (-75.89 + 0.10*T.TOTAL +0.82*X2FG))

Ο έλεγχος Wald και για τις 2 επεξηγηματικές μεταβλητές μας λέει ότι είναι στατιστικά σημαντικές (p-value = 0.0910 και 0.0463) σε επίπεδο στατιστικής σημαντικότητας 10% και 5% αντίστοιχα.

Η τελική απόκριση του μοντέλου 7.1314 κάτι το οποίο είναι ικανοποιητικό καθώς ένα καλό μοντέλο θα πρέπει να έχει μικρή τιμή. Τέλεια προσαρμογή υπάρχει όταν η απόκλιση είναι μηδενική προφανώς.

Τέλος, ο αλγόριθμος μεγιστοποίησε το λογάριθμο της πιθανοφάνειας μετά από 11 επαναλήψεις.

Συγκεντρωτικά, βλέπουμε πως τα μοντέλα logit, probit και cloglog μας οδηγούν στο ίδιο συμπέρασμα, καθώς η μηδενική μας υπόθεση απορρίπτεται και οι 2 μεταβλητές που εξετάζουμε είναι στατιστικά σημαντικές.

Στη συνέχεια για να πάρουμε μια πρώτη εικόνα για το ποιο από τα 4 μοντέλα προσαρμόζεται καλύτερα στα δεδομένα μας θα ελέγξουμε τις αντίστοιχες στατιστικές συναρτήσεις χ-τετράγωνο του pearson.

| | Logit | probit | cauchit | cloglog |
|---|---|---|---|---|
| X2 | 8.0703 | 7.3839 | 15.7536 | **6.9690** |
| p | 0.9948 | 0.9973 | 0.7833 | 0.9982 |

Έτσι λοιπόν βλέπουμε ότι το μοντέλο complementary log log δείχνει να ταιριάζει καλύτερα από τα υπόλοιπα 3 στα δεδομένα μας. Αρκετά κοντά πάντως είναι κατά σειρά και τα μοντέλα probit και logit.

### 4.7.7 Μέτρα Προσαρμογής

➔ Για το μοντέλο logit :

| McFadden | Adj.McFadden | Cox.Snell | Nagelkerke |
|---|---|---|---|
| 0.7750144 | 0.5131716 | 0.6271635 | 0.8710392 |
| McKelvey.Zavoina | Effron | Count | Adj.Count |
| 0.9269598 | 0.7965558 | 0.9583333 | 0.8750000 |
| AIC | Corrected.AIC | | |
| 12.8739129 | 14.0739129 | | |

➔ Για το μοντέλο probit :

| McFadden | Adj.McFadden | Cox.Snell | Nagelkerke |
|---|---|---|---|
| 0.7805663 | 0.5187234 | 0.6297893 | 0.8746860 |
| McKelvey.Zavoina | Effron | Count | Adj.Count |
| 0.9334467 | 0.7950415 | 0.9583333 | 0.8750000 |
| AIC | Corrected.AIC | | |
| 12.7042889 | 13.9042889 | | |

➔ Για το μοντέλο cauchit :

| McFadden | Adj.McFadden | Cox.Snell | Nagelkerke |
|---|---|---|---|
| 0.7537822 | 0.4919394 | 0.6169486 | 0.8568522 |
| McKelvey.Zavoina | Effron | Count | Adj.Count |
| NA | 0.8201211 | 0.9583333 | 0.8750000 |
| AIC | Corrected.AIC | | |

| |
|---|
| 13.5226135 14.7226135 |

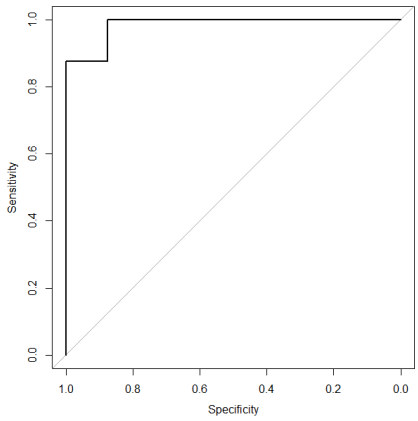
➔ Για το μοντέλο cloglog

| | | | |
|---|---|---|---|
| McFadden | Adj.McFadden | Cox.Snell | Nagelkerke |
| 0.7665872 | 0.5047444 | 0.6231422 | 0.8654541 |
| McKelvey.Zavoina | Effron | Count | Adj.Count |
| NA | 0.7752519 | 0.9166667 | 0.7500000 |
| AIC | Corrected.AIC | | |
| 13.1313859 | 14.3313859 | | |

Γενικά παρατηρούμε ότι το Pseudo-$R^2$ του McFadden είναι μεγαλύτερο του 0.4 κάτι το οποίο δείχνει πολύ καλή προσαρμογή και για τα 4 μοντέλα. Φαίνεται ότι το μοντέλο probit υποστηρίζεται λίγο καλύτερα συγκριτικά με τα υπόλοιπα αν λάβουμε υπόψιν μας το πληροφοριακό κριτήριο του Akaike (AIC = 12.7) καθώς και τα υπόλοιπα Pseudo-$R^2$. Πάντως και το logit μοντέλο προσαρμόζεται αρκετά ικανοποιητικά.

### 4.7.8 Προβλεπτική ισχύς του μοντέλου

Για το μοντέλο probit που επιλέξαμε έχουμε :




Σχήμα 4.7 : Καμπύλη ROC για το probit μοντέλο

Η διαγώνιος είναι η καμπύλη ROC για το μοντέλο που προβλέπει Y =1 και Y=0 με πιθανότητες 50-50, ανεξάρτητα από τις τιμές των επεξηγηματικών μεταβλητών. Προφανώς, όσο πιο ψηλά είναι η καμπύλη ενός μοντέλου (δηλαδή όσο περισσότερο απέχει από την διαγώνιο) τόσο καλύτερη προβλεπτική ικανότητα θα έχει το μοντέλο που έχουμε κατασκευάσει. Στην περίπτωση μας το εμβαδόν κάτω από την καμπύλη

(area under the curve, AUC) ισούται με 0.9843 το οποίο κρίνεται αρκετά ικανοποιητικό για την προβλεπτική ικανότητα του probit μοντέλου.

Έτσι λοιπόν το μοντέλο που εφαρμόσαμε είναι αρκετά καλύτερο από ένα τυχαίο μοντέλο που έχει AUC = 0.5 και συνεπώς οι προβλέψεις με τις πραγματικές παρατηρήσεις θα είναι πολύ κοντά. Το cutoff σημείο ισούται με 0.877, είναι πιο κοντά στην πάνω αριστερή γωνία του σχήματος και θεωρείται βέλτιστο, όσον αφορά το αν η ομάδα προκρίθηκε ή όχι στην επόμενη φάση.

### 4.7.9 Αξιολόγηση της προβλεπτικής ικανότητας του μοντέλου

Ακολουθώντας παρόμοια λογική όπως και για την προηγούμενη χρονιά θα κατασκευάσουμε τον πίνακα ταξινόμησης για το probit μοντέλο που επιλέξαμε ως κατάλληλο. Έτσι λοιπόν έχουμε :

| **CLASSIFICATION TABLE** | | | | | |
|---|---|---|---|---|---|
| | | **PREDICTED** | | | |
| | **QUALIFIED** | 0 | 1 | | **%CORRECT** |
| **OBSERVED** | 0 | 7 | 1 | 8 | 87.5% |
| | 1 | 0 | 16 | 16 | 100% |
| **OVERALL** | | 7 | 17 | 24 | **95.83%** |

Πίνακας 4.6 : classification table με cutoff = 0.50

Παρατηρούμε ότι περίπου το 95.83% των αποτελεσμάτων προβλέφθηκε σωστά από το μοντέλο, δηλαδή μόνο 1 ομάδα από τις 24 δεν κατηγοριοποιήθηκε σωστά, κάτι το οποίο είναι αρκετά εντυπωσιακό. Το ποσοστό ορθής ταξινόμησης προκύπτει και από τους παραπάνω πίνακες καλής προσαρμογής και πιο συγκεκριμένα είναι το μέτρο count που επιστρέφει η συνάρτηση PseudoR2. Πιο συγκεκριμένα από τις 8 ομάδες που δεν προκρίθηκαν στο top-16 εκτιμήθηκαν σωστά οι 7, ποσοστό 87.5% (specificity) και από τις 16 ομάδες που προκρίθηκαν εκτιμήθηκαν σωστά και οι 16, ποσοστό 100% (sensitivity).

Έτσι λοιπόν, σύμφωνα με τον κανόνα που θέσαμε παραπάνω το accuracy του μοντέλου που κατασκευάσαμε θα πρέπει να είναι μεγαλύτερο του 55.12 % + 25% = 80.12 %. Αφού στην περίπτωση μας βρέθηκε ότι το ποσοστό σωστής ταξινόμησης ισούται περίπου με 95.83% είμαστε απόλυτα ευχαριστημένοι.

Αξίζει να σημειώσουμε πως επιλέξαμε ότι το κατώφλι ισούται με 0.50. Μία άλλη συνηθισμένη επιλογή είναι το ποσοστό των Y =1 στα δεδομένα μας, δηλαδή το ποσοστό πρόκρισης (66%). Έτσι λοιπόν ορίζοντας $\pi_0 = 0.66$ έχουμε :

| CLASSIFICATION TABLE | | | | | |
|---|---|---|---|---|---|
| | | PREDICTED | | | |
| | **QUALIFIED** | 0 | 1 | | **%CORRECT** |
| **OBSERVED** | 0 | 7 | 1 | 8 | 87.5% |
| | 1 | 2 | 14 | 16 | 87,5% |
| **OVERALL** | | 9 | 15 | 24 | **87.5%** |

Πίνακας 4.7 : classification table με cutoff = 0.66

Βλέπουμε ότι το ποσοστό ορθής ταξινόμησης σε αυτή την περίπτωση είναι μικρότερο από όταν $\pi_0 = 0.5$ , συνεπώς θα παραμείνουμε στην πρώτη μας επιλογή.

# ΚΕΦΑΛΑΙΟ 5

## Στατιστική ανάλυση με βάση την έδρα της ομάδας

Σε αυτό το κεφάλαιο θα διαπιστώσουμε μέσα από την εφαρμογή στατιστικών ελέγχων και με την μέθοδο random forest ποια στατιστικά στοιχεία είναι σημαντικά για την εξέλιξη του σκορ σε έναν αγώνα μπάσκετ. Αυτή την φορά θα ελέγξουμε κατά πόσο ο παράγοντας έδρα επιδρά στην εξέλιξη του κάθε παιχνιδιού ξεχωριστά για την regular season 2014-2015 και 2015-2016. Στην συνέχεια για το μοντέλο που θα επιλέξουμε θα αξιολογήσουμε την προβλεπτική του ισχύ. Τέλος, θα προχωρήσουμε σε σύγκριση αυτών των μοντέλων για να δούμε ποια χαρακτηριστικά επηρεάζουν περισσότερο την εξέλιξη των 24 ομάδων μέσα στην διοργάνωση με βάση την έδρα.

### 5.1 Συλλογή δεδομένων

Όπως σε όλα τα αθλητικά γεγονότα, η έδρα της ομάδας θεωρείται ότι παίζει αρκετά σημαντικό ρόλο στην εξέλιξη του αγώνα και στον καθορισμό του τελικού νικητή. Το ζητούμενο είναι να εντοπίσουμε τις μεταβλητές εκείνες που επηρεάζουν την γηπεδούχο ομάδα, στο να φτάσει στην νίκη ή στην ήττα αντίστοιχα. Έτσι λοιπόν, σε αντίθεση με τα προηγούμενα κεφάλαια, θα χρησιμοποιήσουμε στατιστικά στοιχεία για τον κάθε αγώνα ξεχωριστά. Συνολικά και τις 2 χρονιές που θα εξετάσουμε πραγματοποιήθηκαν 240 αγώνες οι οποίοι φαίνονται αναλυτικά στο επισυναπτόμενο παράρτημα. Αξίζει να σημειώσουμε πώς την χρονιά 2015-2016 από τα 120 παιχνίδια τα 3 κρίθηκαν στην παράταση και από την χρονιά 2014-2015 τα 4 κρίθηκαν στην παράταση και μάλιστα στα 3 από αυτά τελική νικήτρια ήταν η φιλοξενούμενη ομάδα. Σε αυτή την ανάλυση αντί για την μεταβλητή qualified θα χρησιμοποιήσουμε την δίτιμη μεταβλητή HOME.AWAY με επιτρεπτές τιμές 1 (γηπεδούχος) και 0 (φιλοξενούμενος).

### 5.2 Exploratory data analysis για την χρονιά 2014-2015

Αρχικά με την βοήθεια της βιβλιοθήκης Data Explorer που διαθέτει η R θα ελέγξουμε αν υπάρχουν ελλιπείς τιμές (missing values) στο dataset μας.

Από τα αποτελέσματα δεν παρατηρούνται ελλιπείς τιμές στα δεδομένα μας. Αν εργαστούμε με τον ίδιο τρόπο θα διαπιστώσουμε ότι ούτε για την επόμενη χρονιά υπάρχουν missing values, συνεπώς συνεχίζουμε κανονικά στην ανάλυσή μας. Στην συνέχεια με την εντολή str(df) μπορούμε να πάρουμε μία γενική εικόνα για τα δεδομένα μας.

```
'data.frame':   240 obs. of  27 variables:
 $ GROUP.A      : Factor w/ 35 levels "Alba Berlin",..: 2 33 24 8 29 35 33 29 35 24
...
 $ Pts.Q1.      : num  20 19 15 26 13 14 16 19 23 17 ...
 $ Pts.Q2.      : num  25 15 28 21 19 21 15 16 25 7 ...
```

```
$ Pts.Q3.          : num  19 18 23 14 24 14 25 20 27 18 ...
$ Pts.Q4.          : num  18 24 22 25 24 22 19 21 22 21 ...
$ Pts.ET.          : num  0 0 0 0 0 0 0 0 0 0 ...
$ Pts.Total.       : num  82 76 88 86 80 71 75 76 97 63 ...
$ dif_Q1           : num  1 -1 -11 11 -1 1 -3 3 6 -6 ...
$ dif_Q2           : num  10 -10 7 -7 -2 2 -1 1 18 -18 ...
$ dif_Q3           : num  1 -1 9 -9 10 -10 5 -5 9 -9 ...
$ dif_Q4           : num  -6 6 -3 3 2 -2 -2 2 1 -1 ...
$ dif_Pts.Total.   : num  6 -6 2 -2 9 -9 -1 1 34 -34 ...
$ X2FG             : num  46.9 54.3 45.8 51.2 46.8 50 58.8 41.3 72.9 38.3 ...
$ X3FG             : num  35.3 47.1 52.6 39.3 27.3 50 29.6 24 36.4 23.5 ...
$ FT               : num  85.7 66.7 70 64.3 66.7 57.7 84.6 83.3 75 51.7 ...
$ O                : num  12 4 19 15 17 7 9 16 6 18 ...
$ D                : num  24 26 25 21 22 26 29 24 33 18 ...
$ T                : num  36 30 44 36 39 33 38 40 39 36 ...
$ As               : num  13 12 16 20 21 11 14 18 22 14 ...
$ St.              : num  8 5 9 11 12 4 3 6 6 8 ...
$ To               : num  7 14 16 14 8 16 14 5 14 15 ...
$ Fv               : num  1 3 7 4 1 3 3 1 2 2 ...
$ Ag               : num  3 1 4 7 3 1 1 3 2 2 ...
$ Cm               : num  25 24 18 21 28 25 25 23 27 24 ...
$ Rv               : num  24 25 20 17 25 27 21 24 23 27 ...
$ PIR              : num  89 80 105 89 89 71 79 84 121 53 ...
$ HOME.AWAY        : num  1 0 1 0 1 0 1 0 1 0 ...
```

Πίνακας 5.1 Δομή των δεδομένων για την regular season 2014-2015

Παρατηρούμε λοιπόν ότι το dataset μας αποτελείται από 240 σειρές και 27 μεταβλητές. Σημειώνουμε πως η μεταβλητή HOME.AWAY εμφανίζεται ως numeric και πρέπει να μετατραπεί σε factor.

Κάποια βασικά στατιστικά περιγραφικά μέτρα παρουσιάζονται παρακάτω :

| $`0` ΕΚΤΟΣ ΕΔΡΑΣ | | | | | | |
|---|---|---|---|---|---|---|
| | **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** |
| Pts.Q1 | 10 | 16 | 19 | 18,69 | 21 | 30 |
| Pts.Q2 | 7 | 16 | 19 | 19,32 | 23 | 38 |
| Pts.Q3 | 8 | 16 | 18 | 18,42 | 21 | 36 |
| Pts.Q4 | 8 | 16 | 19 | 19.14 | 22 | 33 |
| Pts.Total | 56 | 68.75 | 76 | 76,01 | 82 | 104 |
| dif_Q1 | -17 | -6 | -1 | -1,258 | 3,25 | 15 |
| dif_Q2 | -19 | -5 | -1 | -0,425 | 3 | 16 |
| dif_Q3 | -19 | -5 | -1 | -1,658 | 2 | 20 |
| dif_Q4 | -18 | -4 | -1 | -0,717 | 3 | 22 |
| dif_Pts.Total | -38 | -11,25 | -3,5 | -4,058 | 5 | 39 |
| X2FG | 28,6 | 43,6 | 50 | 50,12 | 55,38 | 69,2 |
| X3FG | 10 | 27,6 | 34,7 | 34,41 | 42,1 | 60 |
| FT | 37,5 | 66,7 | 73,95 | 78,18 | 81,42 | 84 |
| O.TOTAL | 3 | 8 | 11 | 10,93 | 13 | 19 |
| D.TOTAL | 13 | 20 | 24 | 24.11 | 28 | 35 |
| T.TOTAL | 19 | 31 | 36 | 34,89 | 39 | 47 |
| As.Total | 1 | 13 | 16 | 16,27 | 20 | 30 |
| St.Total | 1 | 5 | 6 | 6,392 | 8 | 13 |
| To.Total | 5 | 10 | 13 | 13,14 | 16 | 31 |
| Fv.Total | 0 | 1 | 2,5 | 2,833 | 4 | 9 |
| Ag.Total | 0 | 1,75 | 2,5 | 2,858 | 4 | 9 |
| Cm.Total | 1 | 19 | 22 | 21,45 | 24 | 33 |
| Rv.Total | 12 | 18 | 20,5 | 20,73 | 23 | 33 |
| PIR.Total | 43 | 68 | 78,5 | 79,84 | 93,25 | 133 |
| $`1` ΕΝΤΟΣ ΕΔΡΑΣ | | | | | | |
| | **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** |
| Pts.Q1 | 2 | 17 | 20 | 19,95 | 23 | 32 |
| Pts.Q2 | 9 | 15 | 20 | 19,75 | 23 | 32 |
| Pts.Q3 | 7 | 17 | 20 | 20,07 | 24 | 32 |
| Pts.Q4 | 5 | 16 | 20 | 19,86 | 24 | 32 |
| Pts.Total | 55 | 71 | 79,5 | 80,07 | 89 | 115 |
| dif_Q1 | -15 | -3 | 1 | 0,425 | 5 | 19 |
| dif_Q2 | -16 | -3 | 1 | 0,425 | 5 | 19 |
| dif_Q3 | -20 | -2 | 1 | 1,658 | 5 | 19 |
| dif_Q4 | -22 | -3 | 1 | 0,7167 | 4 | 18 |
| dif_Pts.Total | -39 | -5 | 3,5 | 4,058 | 11,25 | 38 |
| X2FG | 24,3 | 46,12 | 52,85 | 52,31 | 57,62 | 72,9 |
| X3FG | 9,1 | 29,2 | 35,3 | 36,17 | 43 | 69,2 |
| FT | 33,3 | 67,83 | 74,55 | 73,72 | 81,8 | 100 |

| | | | | | | |
|---|---|---|---|---|---|---|
| O.TOTAL | 3 | 8 | 10 | 10,6 | 12 | 25 |
| D.TOTAL | 14 | 22 | 25 | 25,28 | 28,25 | 36 |
| T.TOTAL | 25 | 32 | 36 | 35,88 | 39 | 52 |
| As.Total | 7 | 14 | 17,5 | 17,8 | 21 | 33 |
| St.Total | 1 | 4,75 | 6 | 6,717 | 8,25 | 22 |
| To.Total | 4 | 10 | 13 | 12,57 | 15 | 23 |
| Fv.Total | 0 | 1,75 | 2,5 | 2,858 | 4 | 9 |
| Ag.Total | 0 | 1 | 2,5 | 2,817 | 4 | 9 |
| Cm.Total | 13 | 18 | 21 | 21,12 | 24 | 31 |
| Rv.Total | 13 | 19 | 21 | 21,11 | 23 | 33 |
| PIR.Total | 44 | 75 | 88,5 | 89,72 | 104,25 | 158 |

Πίνακας 5.2 Βασικά περιγραφικά μέτρα (2014-2015)

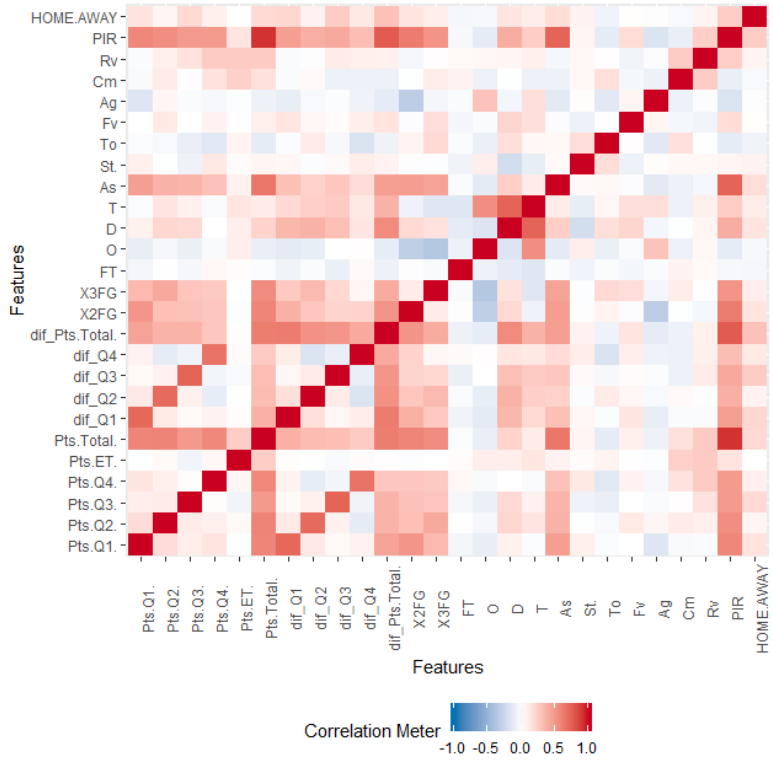
→Βλέπουμε ότι όσον αφορά κάθε δεκάλεπτο ξεχωριστά η φιλοξενούμενη ομάδα είχε κατά μέσο όρο περισσότερους πόντους στο δεύτερο (19.32) και η γηπεδούχος στο τρίτο δεκάλεπτο (20.07). Για τους συνολικούς πόντους βλέπουμε πώς την regular season 2014-2015 η γηπεδούχος ομάδα σκόραρε κατά μέσο όρο περίπου 80 πόντους σε κάθε παιχνίδι ενώ η φιλοξενούμενη περίπου 76 πόντους. Επιπλέον στον παραπάνω πίνακα παρουσιάζονται αναλυτικά και οι διαφορές μεταξύ των δεκαλέπτων, όπου για την φιλοξενούμε ομάδα ο μέσος όρος είναι αρνητικός. Δηλαδή κατά μέσο όρο όλες οι φιλοξενούμενες ομάδες βρίσκονταν πίσω στο σκορ. Αυτό το γεγονός αποτελεί μια πρώτη ένδειξη ότι ο παράγοντας έδρα όντως παίζει σημαντικό αποτέλεσμα στον καθορισμό του τελικού νικητή.

→Ενδιαφέρον παρουσιάζουν και τα ποσοστά ευστοχίας στις ελεύθερες βολές, στα σουτ 2 και 3 πόντων. Για την γηπεδούχο ομάδα τα ποσοστά αυτά κατά μέσο όρο ήταν αντίστοιχα 73.72% , 52.31% και 36.17%. Για την φιλοξενούμενη ομάδα τα ποσοστά αυτά ήταν αντίστοιχα 73.95%, 50.12% και 34.41%. Παρατηρούμε ότι οι φιλοξενούμενες ομάδες είχαν έστω και οριακά καλύτερο ποσοστό ευστοχίας κατά μέσο όρο στις ελεύθερες βολές.

→Η γηπεδούχος ομάδα είχε κατά μέσο όρο 17,8 assist και 6,7 κλεψίματα σε κάθε παιχνίδι, ενώ η φιλοξενούμενη ομάδα 16,3 και 6,4 αντίστοιχα.

→Ο ειδικός δείκτης αξιολόγησης της διοργάνωσης, όπως θα περίμενε κανείς είναι μεγαλύτερος για τις ομάδες που αγωνίζονταν εντός έδρας κατά μέσο όρο. (89.72 έναντι 79.84). Ο μεγαλύτερος δείκτης για τις ομάδες που έπαιζαν ως γηπεδούχοι ήταν 158.

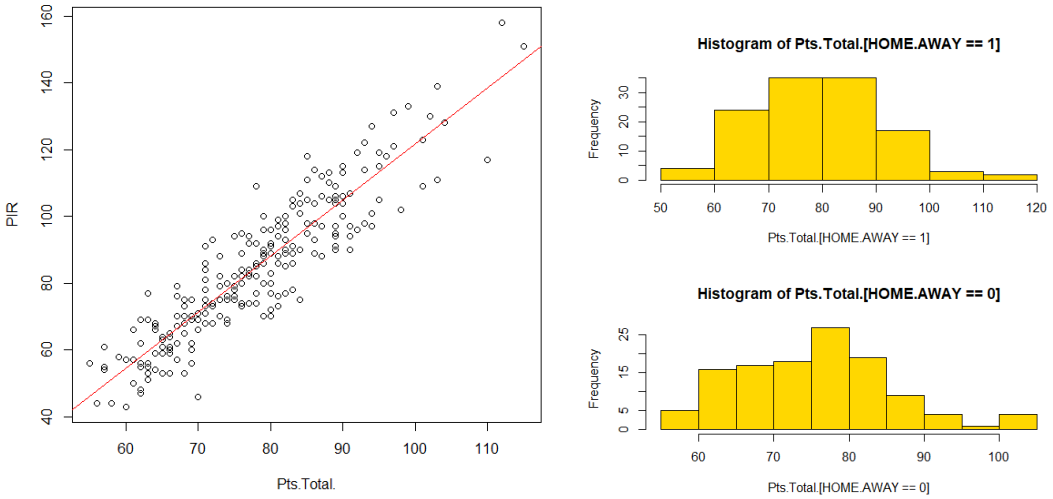
Για να εμβαθύνουμε περισσότερο στην σχέση όλων αυτών των μεταβλητών μεταξύ τους αρκεί να παρουσιαστεί ο πίνακας των συντελεστών συσχέτισης (Pearson):




Σχήμα 5.1 Correlogram 2014-2015

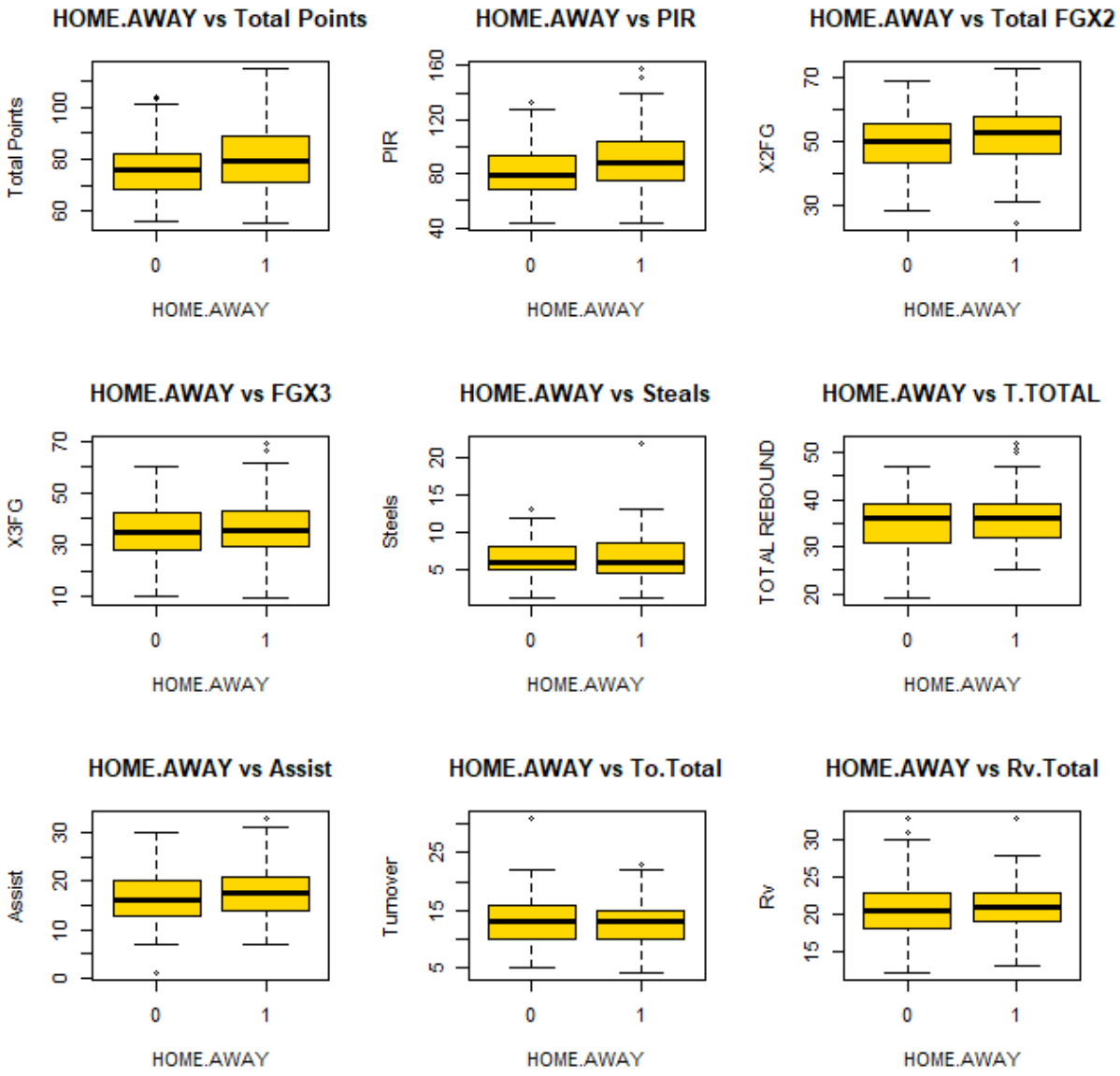
Εδώ με έντονο κόκκινο χρώμα παρουσιάζονται οι θετικά συσχετισμένες μεταβλητές και με μπλε οι αρνητικά συσχετισμένες μεταβλητές για την χρονιά 2014-2015 (όπως φαίνεται και στο υπόμνημα κάτω από το γράφημα οι τιμές κυμαίνονται από -1 έως +1 ). Ακόμη παρατηρούμε ότι οι ισχυρότερες θετικές συσχετίσεις συγκεντρώνονται κυρίως στις μεταβλητές PIR και Pts_Total. Όσο πιο ανοιχτός είναι ο χρωματισμός στο heatmap τόσο ποιο αδύναμη είναι και η σχέση των 2 μεταβλητών. Ενδεικτικά αναφέρουμε ότι τα επιθετικά ριμπάουντ έχουν αρνητική συσχέτιση με τα ποσοστά ευστοχίας στα σουτ 2 και 3 πόντων, κάτι το οποίο είναι και λογικό.

Στην συνέχεια για να γίνουν ακόμη πιο κατανοητές από τον αναγνώστη οι μεταβλητές μας και η επιρροή τους θα τις οπτικοποιήσουμε χρησιμοποιώντας, όπως και στο δεύτερο κεφάλαιο, διαγράμματα.




Σχήμα 5.2 Scatterplot (PIR vs Pts.Total) και ιστόγραμμα πόντων (2014-2015)

Από το παραπάνω σχήματα φαίνεται ότι η σχέση των πόντων και του δείκτη PIR είναι γραμμική και η ευθεία παλινδρόμησης που προσαρμόστηκε έχει θετική κλίση. Ακόμη τα σημεία βρίσκονται διασκορπισμένα αρκετά κοντά στην ευθεία με αποτέλεσμα να μπορούμε να ισχυριστούμε πως η γραμμική σχέση είναι αρκετά ισχυρή. Έτσι λοιπόν οι πόντοι επηρεάζουν τους ειδικούς δείκτες αξιολόγησης. Στα δεξιά παρουσιάζεται ένα ιστόγραμμα, χωρισμένο σε κλάσεις, που αντικατοπτρίζει το πώς κατανέμονται οι πόντοι των ομάδων. Θα μπορούσαμε να ισχυριστούμε ότι παρατηρείται μια κανονική κατανομή, τόσο στο σύνολο των παιχνιδιών για τις ομάδες που έπαιζαν ενός έδρας αλλά και για αυτές που ήταν φιλοξενούμενες. Το ιστόγραμμα όμως δεν είναι ικανό να μας απαντήσει αν τα δεδομένα προέρχονται από μια κανονική κατανομή με έναν συγκεκριμένο μέσο και μια διακύμανση. Η οπτική διερεύνηση δεν είναι πάντα σωστή για αυτό τον λόγο θα καταφύγουμε σε τεστ κανονικότητας για να απαντήσουμε στο ερώτημα της κανονικότητας.




Σχήμα 5.3 boxplots (2014-2015)

Είναι σαφές, πώς σχεδόν σε όλα τα σχήματα η διάμεσος για τις ομάδες που έπαιζαν εντός έδρας(HOME.AWAY = 1) βρίσκεται υψηλότερα από τις ομάδες που δεν έπαιζαν εντός έδρας (HOME.AWAY = 0). Ακόμη, φαίνεται πως παρατηρούνται κάποιες ακραίες τιμές όσον αφορά τον ειδικό δείκτη αξιολόγησης στους γηπεδούχους, στα ποσοστά ευστοχίας για σουτ 3 πόντων, στα συνολικά ριμπάουντ. Για τους φιλοξενούμενους ακραίες τιμές ίσως παρουσιάζονται στα λάθη και στην μεταβλητή Rv. Σημειώνουμε πώς αυτές οι παρατηρήσεις θα συμπεριληφθούν στην ανάλυση καθώς δεν αποτελούν extreme values.

Επιπλέον, χρησιμοποιώντας την βιβλιοθήκη ggplot2 που προσφέρεται από την R και χρησιμοποιείται ευρέως στην εξόρυξη γνώσης, παίρνουμε τα παρακάτω διαγράμματα για τους συνολικούς πόντους και τα συνολικά ριμπάουντ αναλυτικά για κάθε μια από τις 24 ομάδες συγκριτικά με την έδρα.


Σχήμα 5.4 : home.away vs points graph(2014—2015)

Από το παραπάνω σχήμα βλέπουμε ότι η Real Madrid στα 10 παιχνίδια που έδωσε κατά την διάρκεια της regular season 2014-2015 αγωνίστηκε 5 φορές σαν γηπεδούχος και 5 φορές σαν φιλοξενούμενη. Παρατηρούμε ότι εντός έδρας (πράσινο χρώμα) σημείωσε σε 2 από τα 5 παιχνίδια πάνω από 100 πόντους. Χαρακτηριστικά παραδείγματα φιλοξενούμενων που σημείωσαν πάνω από 100 πόντους είναι ο Ερυθρός Αστέρας, η Μπαρτσελόνα και η Αρμάνι. Σε σύγκριση και με το αντίστοιχο διάγραμμα που έγινε

στο δεύτερο κεφάλαιο μπορούμε να ισχυριστούμε ότι οι περισσότερες ομάδες που προκρίθηκαν είχαν υψηλή απόδοση στο σκορ είτε έπαιζαν εντός είτε εκτός έδρας.


Σχήμα 5.5 home.away vs rebounds graph (2014-2015)

Τέλος, όσον αφορά τα ριμπάουντ δεν υπάρχει μία ξεκάθαρη εικόνα. Αξίζει να σημειώσουμε πώς η Alba Berlin σε ένα παιχνίδι που έπαιζε ως φιλοξενούμενη είχε μόλις 19 συνολικά ριμπάουντ. Η Anatolu Efes Istanbul, αγωνιζόμενη ως γηπεδούχος είχε τα περισσότερα ριμπάουντ στην regular season 2014-2015, ξεπερνώντας τα 50 σε ένα παιχνίδι. Σε 2 από τα 5 παιχνίδια που έδωσε εντός έδρας η Unicaja Malaga είχε και αυτή 50 ή περισσότερα συνολικά ομαδικά ριμπάουντ.

### 5.3 Κανονικότητα της κατανομής των ανεξάρτητων μεταβλητών

Οι υποθέσεις στον έλεγχο κανονικότητας διαμορφώνονται ως εξής:

$H_0$ : Η κατανομή των δεδομένων δεν διαφέρει από την κανονική, έναντι της

$H_1$ : Η κατανομή των δεδομένων διαφέρει από την κανονική.

Στην περίπτωσή μας εφαρμόζοντας τον μη-παραμετρικό έλεγχο κανονικότητας Shapiro-Wilk για κάθε μία από τις ανεξάρτητες ποσοτικές μας μεταβλητές την χρονιά 2014-2015 παρουσιάζουμε τον παρακάτω πίνακα με τις αντίστοιχες τιμές p-value του ελέγχου :

| **Μεταβλητές** | **p-value** |
|---|---|
| Pts.Q1 | 0.1064 |
| Pts.Q2 | 0.07434 |
| Pts.Q3 | 0.04755 |
| Pts.Q4 | 0.3825 |
| Pts.Total | 0.0115 |
| Dif_Q1 | 0.5766 |
| Dif_Q2 | 0.4993 |
| Dif_Q3 | 0.02193 |
| Dif_Q4 | 0.012 |
| Dif_Pts.Total | 0.5193 |
| X2FG | 0.7834 |
| X3FG | 0.5666 |
| FT | 0 |
| O | 0.0023 |
| D | 0.05662 |
| T | 0 |
| As | 0.0011 |
| St | 0 |

| | |
|---|---|
| To | 0.0002 |
| Fv | 0 |
| Ag | 0 |
| Cm | 0 |
| Rv | 0.02 |
| PIR | 0.0088 |

Πίνακας 5.3 Κανονικότητα μεταβλητών (2014-2015)

Σε αυτό το σημείο αξίζει να σημειώσουμε πως τα αποτελέσματα δεν διαφοροποιούνται όταν ο έλεγχος εφαρμόζεται ξεχωριστά για τους 2 ανεξάρτητους υπό-πληθυσμούς (γηπεδούχος – φιλοξενούμενος) και όχι συνολικά για την μεταβλητή Home.Away, η οποία περιλαμβάνει αυτούς τους 2 πληθυσμούς.

Έτσι λοιπόν από τον έλεγχο Shapiro-Wilk παρατηρούμε πως δεν μπορούμε να απορρίψουμε ότι σχεδόν όλες οι μεταβλητές μας ακολουθούν την κανονική κατανομή σε επίπεδο στατιστικής σημαντικότητας 5%. Εξαίρεση αποτελούν οι Pts.Q3, Pts.ET, Pts.Total, Dif.Q3, Dif.Q4, FT, O,T, assists, steals, To, Fv, Ag, Cm, Rv και ο ειδικός δείκτης αξιολόγησης της διοργάνωσης (p-value = 0.0088) που όπως φαίνεται η μηδενική μας υπόθεση απορρίπτεται για $\alpha = 0.05$ και συνεπώς δεν μπορούμε να ισχυριστούμε πως ακολουθούν την κανονική κατανομή.

### 5.4 Έλεγχοι t-test & Mann-Whitney για regular season 2014-2015

Αφού διαπιστώσαμε ποιες από τις μεταβλητές μας ακολουθούν την κανονική κατανομή και ποιες όχι, στην συνέχεια θα προχωρήσουμε στους αντίστοιχους στατιστικούς ελέγχους για κάθε μία περίπτωση. Αρχικά, στην περίπτωση της κανονικότητας ο έλεγχος που θα εφαρμόσουμε είναι one sample t-test.

Ο έλεγχος t ενός δείγματος (one sample t-test) χρησιμοποιείται σε περιπτώσεις προβλημάτων στα οποία θέλουμε να ελέγξουμε αν ένα δείγμα προέρχεται από κάποιο πληθυσμό με γνωστό μέσο όρο ή να ελέγξουμε αν ο μέσος όρος ενός δείγματος είναι ίσος με τον μέσο όρο του γενικού πληθυσμού που θεωρούμε ότι είναι γνωστός (Daniel, 2005).

Για την πραγματοποίηση του sample t-test, απαιτείται για τα δεδομένα μας να ισχύουν οι ακόλουθες παραδοχές:

- Η εξαρτημένη μεταβλητή ελέγχου θα πρέπει να προσεγγίζει την κανονική κατανομή. Ο έλεγχος κανονικότητας μπορεί να γίνει ποιοτικά με τη μελέτη των Q-Q plots και P-P plots ή ποσοτικά με χρήση του Kolmogorov-Smirnov test ή του Shapiro-Wilk test (όπως ελέγχθηκε στην παράγραφο 5.3).

• Θα πρέπει στις τιμές να μην υπάρχουν σημαντικά ακραίες τιμές (outliers). Το πρόβλημα με τις ακραίες τιμές είναι ότι μπορούν να έχουν αρνητική επίδραση στο t-test, μειώνοντας την ακρίβεια των αποτελεσμάτων που προκύπτουν. Ο έλεγχος για ακραίες τιμές μπορεί να πραγματοποιηθεί με τη δημιουργία και παρατήρηση του θηκογράμματος, όπως περιγράφηκε στα προηγούμενα κεφάλαια.

• Τα στοιχεία του δείγματος θα πρέπει να είναι ανεξάρτητα (μη συσχετισμένα), το οποίο σημαίνει ότι δεν υπάρχει σχέση μεταξύ των παρατηρήσεων. Η παραβίαση αυτής της παραδοχής συνήθως έχει να κάνει με λανθασμένη σχεδίαση της μελέτης.

• Η εξαρτημένη μεταβλητή ελέγχου πρέπει να είναι ποσοτική μεταβλητή, είτε διαστήματος (interval) είτε αναλογίας (ratio). Παράδειγμα ποσοτικής μεταβλητής διαστήματος θα μπορούσε να θεωρηθεί η θερμοκρασία, όπου η έννοια του μηδέν δεν σχετίζεται με τη μη ύπαρξη μετρήσιμης ποσότητας (π.χ. θερμοκρασία 0 δεν σημαίνει ότι δεν υπάρχει θερμοκρασία, αφού σε άλλη κλίμακα η τιμή θα ήταν διαφορετική). Παράδειγμα ποσοτικής μεταβλητής αναλογίας (ratio) θα μπορούσε να θεωρηθεί η ταχύτητα, όπου το μηδέν δηλώνει ανυπαρξία ταχύτητας ανεξάρτητα από την κλίμακα μέτρησης.

Εφόσον διαπιστώσαμε τη μη παραβίαση των παραπάνω κριτηρίων, είμαστε έτοιμοι για την πραγματοποίηση των στατιστικών ελέγχων με την χρήση της R.

```
t.test(Pts.Q1.~HOME.AWAY,var.equal=TRUE)

        Two Sample t-test

data:  Pts.Q1. by HOME.AWAY

t = -2.0935, df = 238, p-value = 0.03737

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 -2.44244617 -0.07422049

sample estimates:

mean in group 0 mean in group 1

      18.69167        19.95000
```

Ο έλεγχος t-test μας έδειξε ότι απορρίπτουμε την μηδενική μας υπόθεση, δηλαδή υπάρχει σημαντική διαφορά ( t = -2.0935, p-value = 0.03737) μεταξύ των πόντων που σημειώθηκαν στο πρώτο δεκάλεπτο για τις ομάδες που αγωνίστηκαν εντός έδρας σε σύγκριση με τις ομάδες που αγωνίστηκαν εκτός έδρας. Μάλιστα ο μέσος για τις ομάδες

που αγωνίστηκαν εκτός έδρας ήταν ίσος με 18.69 πόντους ενώ για τους γηπεδούχους 19.95.

```
        Two Sample t-test

data:  Pts.Q2. by HOME.AWAY

t = -0.66763, df = 238, p-value = 0.505

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 -1.6790597  0.8290597

sample estimates:

mean in group 0 mean in group 1

        19.325          19.750
```

Ο έλεγχος t-test μας έδειξε ότι δεν απορρίπτουμε την μηδενική μας υπόθεση, δηλαδή δεν υπάρχει σημαντική διαφορά ( t = -0.66763, p-value = 0.505) μεταξύ των πόντων που σημειώθηκαν στο δεύτερο δεκάλεπτο για τις ομάδες που αγωνίστηκαν εντός έδρας σε σύγκριση με τις ομάδες που αγωνίστηκαν εκτός έδρας. Μάλιστα ο μέσος για τις ομάδες που αγωνίστηκαν εκτός έδρας ήταν ίσος με 19.325 πόντους ενώ για τους γηπεδούχους 19.750 και πράγματι δεν παρατηρείται μεγάλη διαφορά.

```
t.test(Pts.Q4.~HOME.AWAY,var.equal=TRUE)

        Two Sample t-test

data:  Pts.Q4. by HOME.AWAY

t = -1.0666, df = 238, p-value = 0.2872

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 -2.0403468  0.6070135

sample estimates:

mean in group 0 mean in group 1

       19.14167        19.85833
```

Ο έλεγχος t-test μας έδειξε ότι δεν απορρίπτουμε την μηδενική μας υπόθεση, δηλαδή δεν υπάρχει σημαντική διαφορά ( t = -1.0666, p-value = 0.2872) μεταξύ των πόντων που σημειώθηκαν στο τέταρτο δεκάλεπτο για τις ομάδες που αγωνίστηκαν εντός έδρας σε σύγκριση με τις ομάδες που αγωνίστηκαν εκτός έδρας. Μάλιστα ο μέσος για τις ομάδες που αγωνίστηκαν εκτός έδρας ήταν ίσος με 19.14167 πόντους ενώ για τους γηπεδούχους 19.85833 και πράγματι δεν παρατηρείται μεγάλη διαφορά.

Με παρόμοιο τρόπο για τις υπόλοιπες μεταβλητές που ακολουθούν την κανονική κατανομή διαπιστώνουμε ότι :

Ο έλεγχος t-test μας έδειξε ότι υπάρχει σημαντική διαφορά ( t = -2.9922, p-value = 0.003062) μεταξύ της διαφοράς των πόντων που σημειώθηκαν στο πρώτο δεκάλεπτο για τις ομάδες που αγωνίστηκαν εντός έδρας σε σύγκριση με τις ομάδες που αγωνίστηκαν εκτός έδρας, ενώ δεν υπάρχει σημαντική διαφορά (t = -0.96538, p-value =0.3353) μεταξύ της διαφοράς των πόντων που σημειώθηκαν στο δεύτερο δεκάλεπτο για τις ομάδες που αγωνίστηκαν εντός έδρας σε σύγκριση με τις ομάδες που αγωνίστηκαν εκτός έδρας. Η διαφορά των συνολικών πόντων, όπως είναι λογικό, είναι στατιστικά σημαντική ανάμεσα στην γηπεδούχο και την φιλοξενούμενη ομάδα (t = -4.7463, p-value = $3.577*10^{-6}$). Τέλος, δεν υπάρχει σημαντική διαφορά μεταξύ των σουτ 2 πόντων (t = -1.931 , p-value =0.05467), των σουτ 3 πόντων (t = -1.2643 , p-value = 0.2074) και των αμυντικών ριμπάουντ (t = -1.9157 , p-value = 0.0566) συγκριτικά με την έδρα.

Για τις υπόλοιπες μεταβλητές που δεν ακολουθούν την κανονική κατανομή θα προχωρήσουμε στην ανάλυση με τον έλεγχο Mann-Whitney U.

Ο έλεγχος Mann-Whitney U αποτελεί έναν μη παραμετρικό έλεγχο ισοδύναμο με το independent samples t-test και χρησιμοποιείται για να συγκρίνει πότε η διαφορά μεταξύ της εξαρτημένης μεταβλητής για 2 ανεξάρτητα γκρουπ είναι στατιστικά σημαντική. Ουσιαστικά εξετάζει πότε η κατανομή της εξαρτημένης μεταβλητής είναι η ίδια για τα 2 γκρουπ (γηπεδούχος – φιλοξενούμενος) και συνεπώς αν προέρχονται από τον ίδιο πληθυσμό. Απορρίπτει την μηδενική υπόθεση ( $H_0$ : Δεν υπάρχει διαφορά μεταξύ των αγώνων που πραγματοποιούνται εντός έδρας και εκτός έδρας όσον αφορά την εκάστοτε μεταβλητή που μας ενδιαφέρει, π.χ. πόντοι που επιτεύχθηκαν κατά την διάρκεια του τρίτου δεκαλέπτου) αν το p-value για το τεστ είναι μικρότερο από το 0.05. Η R χρησιμοποιεί την εντολή wilcox.test(dependent~independent). Στην αριστερή πλευρά του τύπου τοποθετούμε την συνεχή μεταβλητή και το group (γηπεδούχος ή φιλοξενούμενη) στην δεξιά πλευρά. Η τιμή που θα μας δώσει ο έλεγχος, Wilcoxon W, είναι απλά το χαμηλότερο άθροισμα των ranks. Για να υπολογίσει το p-value η R χρησιμοποιεί μια προσέγγιση της κανονικής κατανομής μαζί με μία διόρθωση συνέχειας. Η προσέγγιση είναι λιγότερο αξιόπιστη για μικρά μεγέθη δείγματος.

Στην περίπτωση μας εμείς θέλουμε να ελέγξουμε, για τις μεταβλητές που δεν βρέθηκαν ότι προσεγγίζουν την κανονική κατανομή, αν υπάρχει διαφορά μεταξύ του μέσου

αριθμού της κάθε μεταβλητής για τις γηπεδούχους και φιλοξενούμενες ομάδες αντίστοιχα.

Έτσι λοιπόν θα εξετάσουμε αναλυτικά κάθε μια από τις περιπτώσεις :

```
wilcox.test(Pts.Q3.~HOME.AWAY)

        Wilcoxon rank sum test with continuity correction

data:  Pts.Q3. by HOME.AWAY

W = 5651.5, p-value = 0.003911

alternative hypothesis: true location shift is not equal to 0
```

Ο έλεγχος Mann-Whitney U έδειξε ότι απορρίπτουμε την μηδενική μας υπόθεση, δηλαδή υπάρχει σημαντική διαφορά ( W = 5651.5, p-value = 0.003911) μεταξύ των πόντων που σημειώθηκαν στο τρίτο δεκάλεπτο για τις ομάδες που αγωνίστηκαν εντός έδρας σε σύγκριση με τις ομάδες που αγωνίστηκαν εκτός έδρας.

```
wilcox.test(Pts.Total.~HOME.AWAY)

        Wilcoxon rank sum test with continuity correction

data:  Pts.Total. by HOME.AWAY

W = 5747, p-value = 0.00689

alternative hypothesis: true location shift is not equal to 0
```

Ο έλεγχος Mann-Whitney U έδειξε ότι απορρίπτουμε την μηδενική μας υπόθεση, δηλαδή υπάρχει σημαντική διαφορά ( W = 5747, p-value = 0.00689) μεταξύ των συνολικών πόντων που σημειώθηκαν κατά την διάρκεια του αγώνα για τις ομάδες που αγωνίστηκαν εντός έδρας σε σύγκριση με τις ομάδες που αγωνίστηκαν εκτός έδρας.

Με παρόμοιο τρόπο για τις υπόλοιπες μεταβλητές που δεν ακολουθούν την κανονική κατανομή διαπιστώνουμε ότι :

Ο έλεγχος Mann-Whitney U έδειξε ότι υπάρχει σημαντική διαφορά ( W = 5174, p-value = 0.0001611) μεταξύ της διαφοράς των πόντων που σημειώθηκαν κατά την διάρκεια του τρίτου δεκαλέπτου, ενώ δεν υπάρχει διαφορά ( W = 6291, p-value = 0.9063) μεταξύ της διαφοράς των πόντων που σημειώθηκαν κατά την διάρκεια του τέταρτου δεκαλέπτου για τις ομάδες που αγωνίστηκαν εντός έδρας σε σύγκριση με τις ομάδες που αγωνίστηκαν εκτός έδρας. Ακόμη, δεν υπάρχει διαφορά μεταξύ των ελεύθερων βολών (W = 6782, p-value = 0.4373), των επιθετικών ριμπάουντ (W = 7587, p-value = 0.4706), των συνολικών ριμπάουντ ( W = 6827.5, p-value = 0.4883),των

κλεψιμάτων ( W = 6875, p-value = 0.5436), των λαθών (W = 7726 , p-value = 0.3268),των συνολικών κοψιμάτων υπέρ ( W = 7726, p-value = 0.3268) και κατά (W = 7263.5 , p-value = 0.9053), των συνολικών φάουλ που διέπραξε (W = 7827, p-value = 0.2424) και που δέχθηκε(W= 6561.5, p-value = 0.2338) συγκριτικά με τον παράγοντα έδρα.

Τέλος, από τα αποτελέσματα του ελέγχου διαπιστώνουμε ότι υπάρχει σημαντική διαφορά μεταξύ των ασσιστ ( W = 6101.5 , p-value = 0.04067) καθώς και του ειδικού δείκτη αξιολόγησης ( W = 5381 , p-value = 0.0007193) για τις ομάδες που αγωνίστηκαν εντός έδρας σε σύγκριση με τις ομάδες που αγωνίστηκαν εκτός έδρας.

### 5.5 Random Forest για regular season 2014-2015

Ο αλγόριθμος random forest αποτελεί μια supervised τεχνική κυρίως για classification προβλήματα και ο τρόπος με τον οποίο λειτουργεί έχει αναλυθεί λεπτομερώς στο προηγούμενο κεφάλαιο. Σε αυτό το σημείο θα εφαρμόσουμε την συγκεκριμένη μέθοδο για να διαπιστώσουμε ποιες μεταβλητές διαφέρουν ως προς την έδρα της ομάδας. Για να γίνει αυτό αρχικά θα πρέπει να χωρίσουμε τα δεδομένα μας σε train και validation set. Η αναλογία επιλέγουμε να είναι 70 % για το training sample και 30% για το validation sample.Θα μπορούσαμε να δημιουργήσουμε και ένα test set αλλά για τα συγκεκριμένα δεδομένα δεν κρίνεται απαραίτητο.

```
# Split into Train and Validation sets

# Training Set : Validation Set = 70 : 30 (random)

set.seed(100)

train <- sample(nrow(df), 0.7*nrow(df), replace = FALSE)

TrainSet <- df[train,]

ValidSet <- df[-train,]

summary(TrainSet)

summary(ValidSet)
```

Στη συνέχεια θα δημιουργήσουμε το random forest μοντέλο με προκαθορισμένο αριθμό παραμέτρων. Φυσικά μπορούμε να τροποποιήσουμε το μοντέλο μας αλλάζοντας την μεταβλητή ntree (αριθμός των δέντρων που θα χρησιμοποιηθούν), καθώς και των αριθμό των μεταβλητών που επιλέχθηκαν με τυχαίο τρόπο σε κάθε στάδιο (mtry) .

Πιο συγκεκριμένα η μεταβλητή ntree δηλώνει των αριθμό των δέντρων που θα χρησιμοποιηθούν στο μοντέλο. Αυτός ο αριθμός δεν πρέπει να οριστεί με πολύ χαμηλή τιμή για να σιγουρευτούμε ότι κάθε παρατήρηση προβλέπεται από το μοντέλο αρκετές

φορές. Τέλος, η μεταβλητή mtry, δηλώνει τον αριθμό των μεταβλητών που θα χρησιμοποιηθούν με τυχαία δειγματοληψία κάθε φορά που γίνεται η διάσπαση των δέντρων.

Έτσι λοιπόν έχουμε,

```
# Create a Random Forest model with default parameters

model1 <- randomForest(HOME.AWAY ~ ., data = TrainSet, importance = TRUE)

model1

Call:

 randomForest(formula = HOME.AWAY ~ ., data = TrainSet, importance = TRUE)

               Type of random forest: classification

                     Number of trees: 500

No. of variables tried at each split: 5

        OOB estimate of  error rate: 40.48%

Confusion matrix:

   0  1 class.error

0 45 35      0.4375

1 33 55      0.3750
```

Όπως βλέπουμε ο αριθμός των δέντρων είναι 500 και οι μεταβλητές που δοκιμάστηκαν σε κάθε διάσπαση είναι 5

Ο αριθμός των out of bag (OOB) error είναι 40.68% συνεπώς το ποσοστό ορθής ταξινόμησης (accuracy) του συγκεκριμένου μοντέλου ισούται με 1 – OOB = **59.32%**

```
# Fine tuning parameters of Random Forest model

model2 <- randomForest(HOME.AWAY ~ ., data = TrainSet, ntree = 500, mtry = 3, importance = TRUE)

model2

Call:

 randomForest(formula = HOME.AWAY ~ ., data = TrainSet, ntree = 500,      mtry = 6, importance = TRUE)

               Type of random forest: classification
```

Number of trees: 500

No. of variables tried at each split: 6

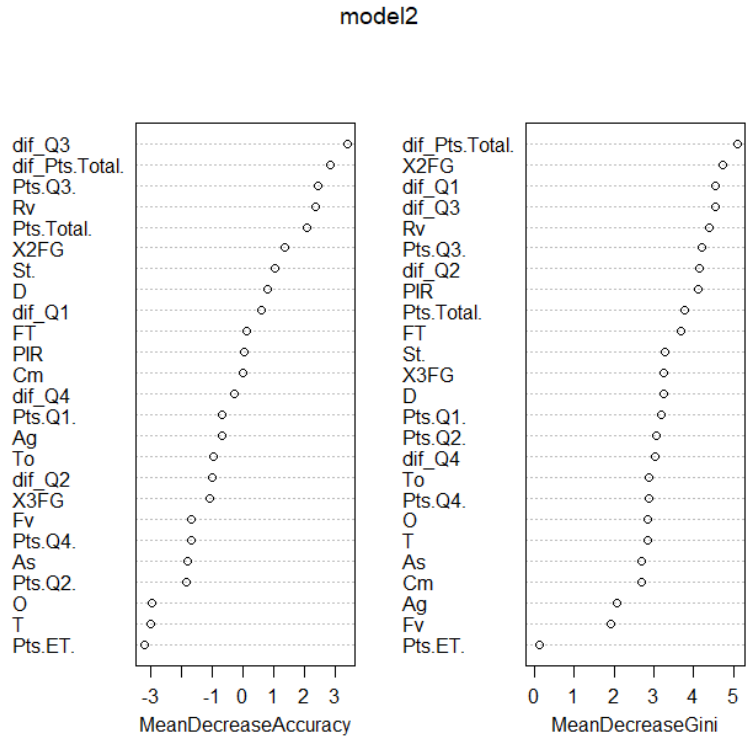
OOB estimate of error rate: 44.64%

Confusion matrix:

| | 0 | 1 | class.error |
|---|---|---|---|
| 0 | 41 | 39 | 0.4875000 |
| 1 | 36 | 52 | 0.4090909 |

Όπως βλέπουμε τώρα που αυξήσαμε τον αριθμό των παραμέτρων από τους 5 στους 6 , το out of bag error αυξήθηκε από 40.68% σε 44.64%. έτσι λοιπόν σε αυτή την περίπτωση το ποσοστό ορθής ταξινόμησης μειώθηκε σε **55.36%**

Τέλος, από τα παρακάτω διαγράμματα βλέπουμε την σημαντικότητα των μεταβλητών κατά σειρά. Αρκετά ψηλά βρίσκονται, τα σουτ 2 πόντων και τα φάουλ που δέχθηκε η ομάδα.



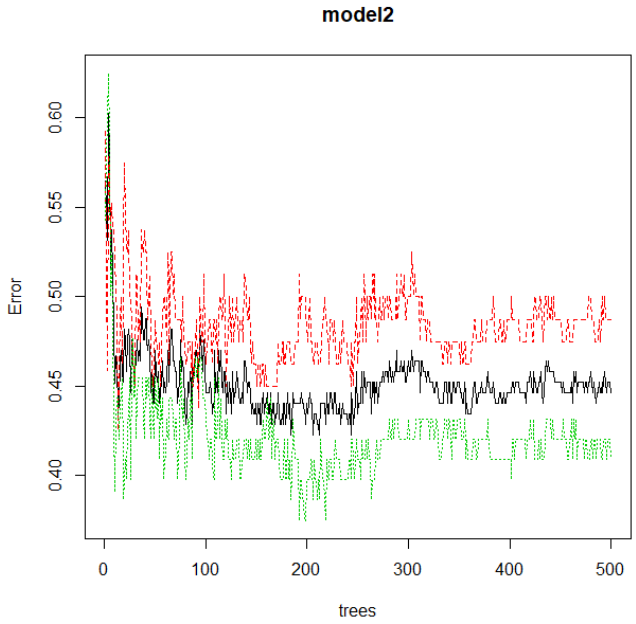

Σχήμα 5.6 Σημαντικότητα των μεταβλητών με την μέθοδο RF (2014-2015)

Στη συνέχεια θα δοκιμάσουμε διάφορα πλήθη μεταβλητών και πιο συγκεκριμένα από 1 μέχρι 10 για να οδηγηθούμε σε αυτό που δίνει το μεγαλύτερο επίπεδο ορθής ταξινόμησης :

| Number of variables tried at each split (mtry) | Out of bag error (OOB) | Accuracy of the model |
|---|---|---|
| 1 | 45.83% | 54.17% |
| 2 | 46.43% | 53.57% |
| 3 | 45.24% | 54.76% |
| 4 | 44.05% | 55.95% |
| 5 | 40.68% | 59.32% |
| 6 | 44.64% | 55.36% |
| 7 | 45.24% | 54.76% |
| 8 | 42.86% | 57.14% |
| 9 | 41.67% | 58.33% |
| 10 | 46.43% | 53,57% |

Έτσι λοιπόν το πλήθος των μεταβλητών που δίνουν το μεγαλύτερο ποσοστό ορθής ταξινόμησης είναι 5 ή 9. Συνήθως θέλουμε όσο λιγότερες μεταβλητές για αυτό θα μείνουμε με την επιλογή mtry = 5

Από το παρακάτω διάγραμμα μπορούμε να πάρουμε μια αναλυτική εικόνα και για τον κατάλληλο αριθμό των δέντρων σε σύγκριση με OOB error :




Σχήμα 5.7 Error vs trees με την μέθοδο RF (2014-2015)

### 5.6 Exploratory data analysis για την χρονιά 2015-2016

Κάποια βασικά στατιστικά περιγραφικά μέτρα για την regular season 2015-2016 παρουσιάζονται παρακάτω :

| $`0` ΕΚΤΟΣ ΕΔΡΑΣ | | | | | | |
|---|---|---|---|---|---|---|
| | **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** |
| Pts.Q1 | 6 | 16 | 20 | 19,48 | 22 | 30 |
| Pts.Q2 | 3 | 15 | 18 | 18,34 | 22 | 35 |
| Pts.Q3 | 7 | 14 | 18 | 17,94 | 21 | 30 |
| Pts.Q4 | 4 | 14 | 19 | 18,17 | 22 | 30 |
| Pts.Total | 49 | 66 | 74 | 74,27 | 84 | 107 |
| dif_Q1 | -16 | -5 | -1 | -0,3167 | 4 | 20 |
| dif_Q2 | -22 | -6 | -1 | -1,075 | 4 | 20 |
| dif_Q3 | -22 | -6 | -1 | -1,458 | 3 | 16 |
| dif_Q4 | -18 | -6 | -1,5 | -1,175 | 4 | 14 |
| dif_Pts.Total | -38 | -14 | -5 | -4,142 | 7,25 | 36 |
| X2FG | 29,3 | 43,58 | 48,7 | 50,37 | 58,65 | 75,8 |
| X3FG | 4,4 | 25 | 35,85 | 34,75 | 44,17 | 65,4 |
| FT | 40,9 | 63,9 | 73,3 | 73,14 | 84,05 | 100 |
| O.TOTAL | 3 | 8 | 10 | 10,4 | 12,25 | 22 |
| D.TOTAL | 15 | 21 | 23 | 23,69 | 27 | 36 |
| T.TOTAL | 22 | 30 | 34 | 34,09 | 38 | 49 |
| As.Total | 6 | 13,75 | 16 | 16,74 | 20 | 33 |
| St.Total | 1 | 4 | 6 | 6,25 | 8 | 17 |
| To.Total | 6 | 12 | 14 | 14,16 | 16 | 26 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Fv.Total | 0 | 1 | 2 | 2,583 | 4 | 7 |
| Ag.Total | 0 | 1 | 3 | 2,992 | 4 | 10 |
| Cm.Total | 2 | 19 | 22 | 21,35 | 24 | 29 |
| Rv.Total | 13 | 18 | 20 | 20,38 | 23 | 30 |
| PIR.Total | 23 | 62 | 77 | 77,62 | 92,25 | 149 |
| $`1` ΕΝΤΟΣ ΕΔΡΑΣ | | | | | | |
| | **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** |
| Pts.Q1 | 9 | 16 | 19 | 19,82 | 23 | 33 |
| Pts.Q2 | 9 | 16 | 19 | 19,33 | 23 | 34 |
| Pts.Q3 | 9 | 15,75 | 20 | 19,42 | 23 | 31 |
| Pts.Q4 | 9 | 15 | 19 | 19,28 | 22 | 29 |
| Pts.Total | 51 | 71 | 78 | 78,32 | 85 | 107 |
| dif_Q1 | -20 | -4 | 1 | 0,3167 | 6 | 16 |
| dif_Q2 | -20 | -4 | 1 | 1,075 | 6 | 22 |
| dif_Q3 | -16 | -3 | 1 | 1,458 | 6 | 22 |
| dif_Q4 | -14 | -4 | 1 | 1,175 | 6 | 18 |
| dif_Pts.Total | -36 | -7,25 | 5 | 4,142 | 12,5 | 38 |
| X2FG | 30,2 | 46 | 51,75 | 51,94 | 57,52 | 76,3 |
| X3FG | 13,6 | 29,9 | 36 | 36,21 | 43,85 | 64 |
| FT | 38,5 | 68,3 | 77,8 | 82,51 | 83,47 | 95 |
| O.TOTAL | 3 | 8 | 11 | 10,33 | 13 | 22 |
| D.TOTAL | 15 | 20 | 24 | 24,42 | 28 | 38 |
| T.TOTAL | 20 | 30 | 34 | 34,75 | 40 | 57 |
| As.Total | 8 | 15 | 18 | 18,35 | 21 | 36 |
| St.Total | 2 | 5 | 6 | 6,917 | 8,25 | 18 |
| To.Total | 4 | 11 | 13 | 13,07 | 15 | 24 |
| Fv.Total | 0 | 1 | 3 | 3,008 | 4 | 10 |
| Ag.Total | 0 | 1 | 2 | 2,6 | 4 | 7 |
| Cm.Total | 2 | 18 | 20 | 20,36 | 23 | 30 |
| Rv.Total | 2 | 19 | 22 | 21,2 | 24 | 29 |
| PIR.Total | 45 | 74 | 89 | 89,08 | 102 | 143 |

Πίνακας 5.4 Βασικά περιγραφικά μέτρα (2015-2016)

→Βλέπουμε ότι όσον αφορά κάθε δεκάλεπτο ξεχωριστά η φιλοξενούμενη ομάδα είχε κατά μέσο όρο περισσότερους πόντους στο πρώτο δεκάλεπτο(19.48) σε σχέση με την προηγούμενη χρονιά όπως και η γηπεδούχος (19.82). Ενδιαφέρον παρουσιάζει το γεγονός ότι μία φιλοξενούμενη ομάδα σκόραρε μόνο 3 πόντους στην δεύτερη περίοδο. Η αντίστοιχη ελάχιστη τιμή για μια γηπεδούχο ομάδα ήταν 9 πόντοι. Για τους συνολικούς πόντους βλέπουμε πώς την regular season 2015-2016 η γηπεδούχος ομάδα σκόραρε κατά μέσο όρο περίπου 78 πόντους, δηλαδή 2 πόντους λιγότερους σε σχέση με την περσινή χρονιά, ενώ η φιλοξενούμενη περίπου 74 πόντους. Επιπλέον στον παραπάνω πίνακα παρουσιάζονται αναλυτικά και οι διαφορές μεταξύ των δεκαλέπτων, όπου για την φιλοξενούμε ομάδα ο μέσος όρος είναι αρνητικός. Δηλαδή κατά μέσο όρο όλες οι φιλοξενούμενες ομάδες βρίσκονταν πίσω στο σκορ. Αυτό το γεγονός αποτελεί

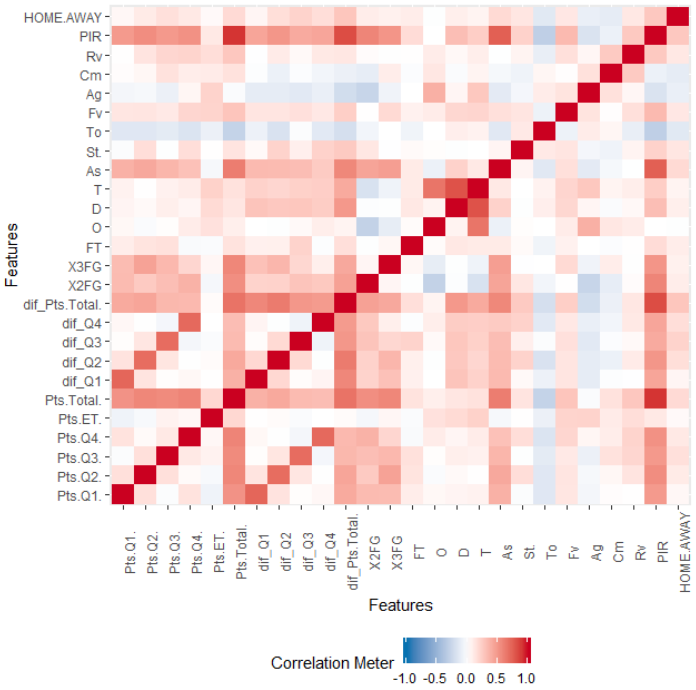
μια πρώτη ένδειξη ότι ο παράγοντας έδρα όντως παίζει σημαντικό αποτέλεσμα στον καθορισμό του τελικού νικητή. Η μικρότερη διαφορά πάντως παρατηρείται κατά τη διάρκεια της πρώτης περιόδου.

→Ενδιαφέρον παρουσιάζουν και τα ποσοστά ευστοχίας στις ελεύθερες βολές, στα σουτ 2 και 3 πόντων. Για την γηπεδούχο ομάδα τα ποσοστά αυτά κατά μέσο όρο ήταν αντίστοιχα 82.51% (εντυπωσιακή αύξηση 9 ποσοστιαίων μονάδων σε σχέση με την περσινή χρονιά), 51.94% και 36.21%. Για την φιλοξενούμενη ομάδα τα ποσοστά αυτά ήταν αντίστοιχα 73.14%, 50.37% και 34.75%. Παρατηρούμε λοιπόν ότι το ποσοστό ευστοχίας στις ελεύθερες βολές των γηπεδούχων ομάδων αυξήθηκε σημαντικά σε σχέση με την regular season 2014-2015

→Η γηπεδούχος ομάδα είχε κατά μέσο όρο 18,3 assist και 6,9 κλεψίματα σε κάθε παιχνίδι, ενώ η φιλοξενούμενη ομάδα 16,7 και 6,2 αντίστοιχα.

→Ο ειδικός δείκτης αξιολόγησης της διοργάνωσης, όπως θα περίμενε κανείς είναι μεγαλύτερος για τις ομάδες που αγωνίζονταν εντός έδρας κατά μέσο όρο. (89.1 έναντι 77.6). Αξιοσημείωτο είναι το γεγονός ότι ο μεγαλύτερος δείκτης αφορά ομάδα που αγωνίστηκε ως φιλοξενούμενη, και πιο συγκεκριμένα ισούται με 149. Αντίστοιχα την τιμή του μικρότερου PIR την έχει και πάλι φιλοξενούμενη ομάδα, και ισούται μόλις με 23(!).

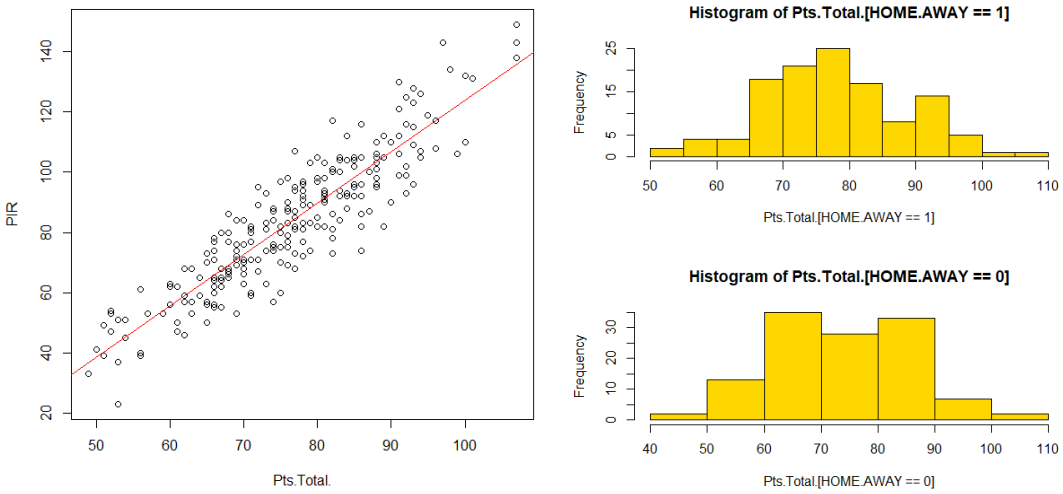
Για να εμβαθύνουμε περισσότερο στην σχέση όλων αυτών των μεταβλητών μεταξύ τους αρκεί να παρουσιαστεί ο πίνακας των συντελεστών συσχέτισης (Pearson):



Σχήμα 5.8 Correlogram 2015-2016

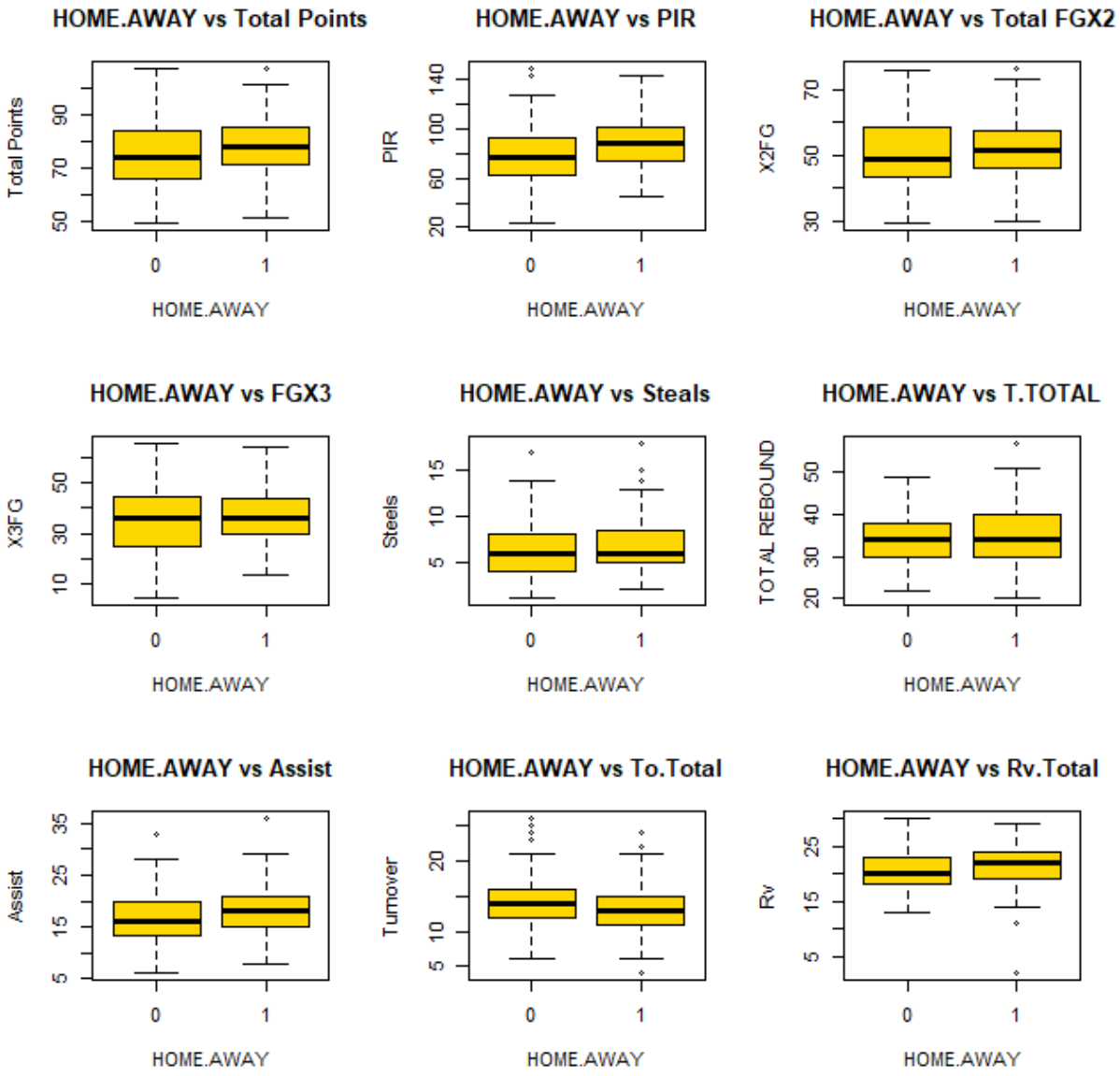
Εδώ με έντονο κόκκινο χρώμα παρουσιάζονται ξανά οι θετικά συσχετισμένες μεταβλητές και με μπλε οι αρνητικά συσχετισμένες μεταβλητές για την χρονιά 2015-2016 (όπως φαίνεται και στο υπόμνημα κάτω από το γράφημα οι τιμές κυμαίνονται από -1 έως +1 ). Ακόμη παρατηρούμε ότι οι ισχυρότερες θετικές συσχετίσεις συγκεντρώνονται κυρίως στις μεταβλητές PIR και Pts_Total, όπως και την προηγούμενη χρονιά. Ισχυρή θετική συσχέτιση φαίνεται πως υπάρχει και μεταξύ των assist και των συνολικών πόντων. Όσο πιο ανοιχτός είναι ο χρωματισμός στο heatmap τόσο ποιο αδύναμη είναι και η σχέση των 2 μεταβλητών. Ενδεικτικά αναφέρουμε ότι τα συνολικά ριμπάουντ έχουν αρνητική συσχέτιση με τα ποσοστά ευστοχίας στα σουτ 2 και 3 πόντων, όπως και τα συνολικά λάθη σε σχέση με των ειδικό δείκτη αξιολόγησης.

Στην συνέχεια για να γίνουν ακόμη πιο κατανοητές από τον αναγνώστη οι μεταβλητές μας και η επιρροή τους θα τις οπτικοποιήσουμε χρησιμοποιώντας, όπως και στο δεύτερο κεφάλαιο, γραφική ανάλυση.




Σχήμα 5.9 Scatterplot (PIR vs Pts.Total) και ιστόγραμμα πόντων (2015-2016)

Από το παραπάνω σχήματα φαίνεται ότι η σχέση των πόντων και του δείκτη PIR είναι ξανά γραμμική και η ευθεία παλινδρόμησης που προσαρμόστηκε έχει θετική κλίση. Ακόμη τα σημεία βρίσκονται διασκορπισμένα αρκετά κοντά στην ευθεία με αποτέλεσμα να μπορούμε να ισχυριστούμε πως η γραμμική σχέση είναι αρκετά ισχυρή. Έτσι λοιπόν οι πόντοι επηρεάζουν τους ειδικούς δείκτες αξιολόγησης. Στην συνέχεια παρουσιάζεται ένα ιστόγραμμα, χωρισμένο σε κλάσεις, που αντικατοπτρίζει το πώς κατανέμονται οι πόντοι των ομάδων. Θα μπορούσαμε να ισχυριστούμε ότι παρατηρείται μια κανονική κατανομή, τόσο στο σύνολο των παιχνιδιών για τις ομάδες που έπαιζαν ενός έδρας αλλά και για αυτές που ήταν φιλοξενούμενες. Συγκριτικά με την προηγούμενη χρονιά πάντως, φαίνεται πως οι πόντοι κατανέμονται πιο ομαλά, αν προσαρμόσουμε την καμπύλη της κανονικής κατανομής πάνω στο σχήμα μας.




Σχήμα 5.10 boxplots (2015-2016)

Είναι σαφές, πώς σχεδόν σε όλα τα σχήματα η διάμεσος για τις ομάδες που έπαιζαν εντός έδρας(HOME.AWAY = 1) βρίσκεται υψηλότερα από τις ομάδες που δεν έπαιζαν εντός έδρας (HOME.AWAY = 0), εκτός από τα λάθη. Ακόμη, φαίνεται πως παρατηρούνται κάποιες ακραίες τιμές, με περισσότερες στην μεταβλητή turnovers για τους φιλοξενούμενους. Σημειώνουμε πώς αυτές οι παρατηρήσεις θα συμπεριληφθούν στην ανάλυση καθώς δεν αποτελούν extreme values.

Επιπλέον, χρησιμοποιώντας την βιβλιοθήκη ggplot2 που προσφέρεται από την R, παίρνουμε τα παρακάτω διαγράμματα για τους συνολικούς πόντους και τα συνολικά ριμπάουντ αναλυτικά για κάθε μια από τις 24 ομάδες συγκριτικά με την έδρα.


Σχήμα 5.11 : home.away vs points graph(2015 - 2016)

Από το παραπάνω σχήμα βλέπουμε ότι οι 24 ομάδες στα 10 παιχνίδια που έδωσαν κατά την διάρκεια της regular season 2015-2016 αγωνίστηκαν 5 φορές σαν γηπεδούχοι και 5 φορές σαν φιλοξενούμενοι. Για παράδειγμα παρατηρούμε ότι εντός έδρας (πράσινο χρώμα) η CSKA Moscow σημείωσε σε 2 από τα 5 παιχνίδια κοντά στους 95 πόντους, ενώ σε 2 από τα 5 παιχνίδια που αγωνίστηκε ως φιλοξενούμενος πέτυχε πάνω από 100 πόντους, κάτι το οποίο είναι εντυπωσιακό. Χαρακτηριστικά παραδείγματα φιλοξενούμενης που σημείωσε πάνω από 100 πόντους είναι η Laboral Kutxa. Σε σύγκριση και με το αντίστοιχο διάγραμμα που έγινε στο δεύτερο κεφάλαιο μπορούμε να ισχυριστούμε ότι οι περισσότερες ομάδες που προκρίθηκαν είχαν υψηλή απόδοση στο σκορ είτε έπαιζαν εντός είτε εκτός έδρας. Θυμίζουμε ότι την season 2015-2016 την πρώτη θέση κατέκτησε η CSKA Moscow και την δεύτερη η Fenerbahce.


Σχήμα 5.12 home.away vs rebounds graph (2015-2016)

Τέλος, όσον αφορά τα ριμπάουντ δεν υπάρχει μία ξεκάθαρη εικόνα. Αξίζει να σημειώσουμε πώς τα λιγότερα rebound αγωνιζόμενη σαν γηπεδούχος είχε η Dinamo Banco Di Sardegna Sassari, η οποία όπως είδαμε και στην ανάλυση που προηγήθηκε στα προηγούμενα κεφάλαια στην προκρίθηκε στην επόμενη φάση του θεσμού. Η Real Madrid και η Darussafaka Dogus Istanbul, αγωνιζόμενες ως γηπεδούχοι είχαν τα περισσότερα ριμπάουντ στην regular season 2014-2015, ξεπερνώντας τα 50. Η CSKA Moscow αν και κινήθηκε σε χαμηλότερα επίπεδα, όχι μόνο προκρίθηκε στην επόμενη φάση αλλά πρόσθεσε ένα ακόμη τρόπαιο στην συλλογή της.

### 5.7 Κανονικότητα της κατανομής των ανεξάρτητων μεταβλητών (2015-2016)

Οι υποθέσεις στον έλεγχο κανονικότητας διαμορφώνονται ως εξής:

$H_0$ : Η κατανομή των δεδομένων δεν διαφέρει από την κανονική, έναντι της

$H_1$ : Η κατανομή των δεδομένων διαφέρει από την κανονική.

Στην περίπτωσή μας εφαρμόζοντας τον μη-παραμετρικό έλεγχο κανονικότητας Shapiro-Wilk για κάθε μία από τις ανεξάρτητες ποσοτικές μας μεταβλητές την χρονιά 2015-2016 παρουσιάζουμε τον παρακάτω πίνακα με τις αντίστοιχες τιμές p-value του ελέγχου :

| **Μεταβλητές** | **p-value** |
|---|---|
| Pts.Q1 | 0.2562 |
| Pts.Q2 | 0.0235 |
| Pts.Q3 | 0.09352 |
| Pts.Q4 | 0.03664 |
| Pts.Total | 0.3943 |
| Dif_Q1 | 0.7557 |
| Dif_Q2 | 0.2256 |
| Dif_Q3 | 0.9071 |
| Dif_Q4 | 0.1106 |
| Dif_Pts.Total | 0.1362 |
| X2FG | 0.3548 |
| X3FG | 0.246 |
| FT | 0 |
| O | 0.0014 |
| D | 0.0038 |
| T | 0.0235 |
| As | 0 |

| | |
|---|---|
| St | 0 |
| To | 0.0001 |
| Fv | 0 |
| Ag | 0 |
| Cm | 0 |
| Rv | 0.0001 |
| PIR | 0.5859 |

Πίνακας 5.5 Κανονικότητα μεταβλητών (2015-2016)

Έτσι λοιπόν από τον έλεγχο Shapiro-Wilk παρατηρούμε πως δεν μπορούμε να απορρίψουμε ότι σχεδόν όλες οι μεταβλητές μας ακολουθούν την κανονική κατανομή σε επίπεδο στατιστικής σημαντικότητας 5%. Εξαίρεση αποτελούν οι Pts.Q2, Pts.Q4, Pts.ET, FT, O, D, T, assists, steels, To, Fv, Ag, Cm, Rv) που όπως φαίνεται η μηδενική μας υπόθεση απορρίπτεται για $\alpha = 0.05$ και συνεπώς δεν μπορούμε να ισχυριστούμε πως ακολουθούν την κανονική κατανομή.

### 5.8 Έλεγχοι t-test & Mann-Whitney για regular season 2014-2015

Αφού διαπιστώσαμε ποιες από τις μεταβλητές μας ακολουθούν την κανονική κατανομή και ποιες όχι, στην συνέχεια θα προχωρήσουμε στους αντίστοιχους στατιστικούς ελέγχους για κάθε μία περίπτωση. Αρχικά, στην περίπτωση της κανονικότητας ο έλεγχος που θα εφαρμόσουμε είναι one sample t-test.

Ο έλεγχος t ενός δείγματος (one sample t-test) χρησιμοποιείται σε περιπτώσεις προβλημάτων στα οποία θέλουμε να ελέγξουμε αν ένα δείγμα προέρχεται από κάποιο πληθυσμό με γνωστό μέσο όρο ή να ελέγξουμε αν ο μέσος όρος ενός δείγματος είναι ίσος με τον μέσο όρο του γενικού πληθυσμού που θεωρούμε ότι είναι γνωστός (Daniel, 2005).

Για την πραγματοποίηση του sample t-test, απαιτείται για τα δεδομένα μας να ισχύουν ξανά οι παραδοχές που αναφέραμε και προηγουμένως. Εφόσον διαπιστώσαμε τη μη παραβίαση των παραπάνω κριτηρίων, είμαστε έτοιμοι για την πραγματοποίηση των στατιστικών ελέγχων με την χρήση της R.

```
t.test(Pts.Q1.~HOME.AWAY,var.equal=TRUE)

        Two Sample t-test

data:  Pts.Q1. by HOME.AWAY
```

```
t = -0.56257, df = 238, p-value = 0.5743

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 -1.5756109  0.8756109

sample estimates:

mean in group 0 mean in group 1

        19.475          19.825
```

Ο έλεγχος t-test μας έδειξε ότι δεν απορρίπτουμε την μηδενική μας υπόθεση, δηλαδή δεν υπάρχει σημαντική διαφορά ( t = -0.56257, p-value = 0.5743) μεταξύ των πόντων που σημειώθηκαν στο πρώτο δεκάλεπτο για τις ομάδες που αγωνίστηκαν εντός έδρας σε σύγκριση με τις ομάδες που αγωνίστηκαν εκτός έδρας. Μάλιστα ο μέσος για τις ομάδες που αγωνίστηκαν εκτός έδρας ήταν ίσος με 19.475 πόντους ενώ για τους γηπεδούχους 19.825.

```
t.test(Pts.Q3.~HOME.AWAY,var.equal=TRUE)

        Two Sample t-test

data:  Pts.Q3. by HOME.AWAY

t = -2.2669, df = 238, p-value = 0.02429

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 -2.7568094 -0.1931906

sample estimates:

mean in group 0 mean in group 1

       17.94167        19.41667
```

Αντίθετα, ο έλεγχος t-test μας έδειξε ότι απορρίπτουμε την μηδενική μας υπόθεση, δηλαδή υπάρχει σημαντική διαφορά ( t = -2.2669, p-value = 0.02429) μεταξύ των πόντων που σημειώθηκαν στο τρίτο δεκάλεπτο για τις ομάδες που αγωνίστηκαν εντός έδρας σε σύγκριση με τις ομάδες που αγωνίστηκαν εκτός έδρας. Μάλιστα ο μέσος για τις ομάδες που αγωνίστηκαν εκτός έδρας ήταν ίσος με 17.94167 πόντους ενώ για τους γηπεδούχους 19.41667.

```
t.test(Pts.Total.~HOME.AWAY,var.equal=TRUE)

        Two Sample t-test

data:  Pts.Total. by HOME.AWAY

t = -2.7084, df = 238, p-value = 0.007252

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 -6.995827 -1.104173

sample estimates:

mean in group 0 mean in group 1

      74.26667        78.31667
```

Ο έλεγχος t-test μας έδειξε ότι απορρίπτουμε την μηδενική μας υπόθεση, δηλαδή υπάρχει σημαντική διαφορά ( t = -2.7084, p-value = 0.007252) μεταξύ των πόντων που σημειώθηκαν σε όλη την κανονική διάρκεια του αγώνα για τις ομάδες που αγωνίστηκαν εντός έδρας σε σύγκριση με τις ομάδες που αγωνίστηκαν εκτός έδρας. Μάλιστα ο μέσος για τις ομάδες που αγωνίστηκαν εκτός έδρας ήταν ίσος με 72.26 πόντους ενώ για τους γηπεδούχους 78.16.

Με παρόμοιο τρόπο για τις υπόλοιπες μεταβλητές που ακολουθούν την κανονική κατανομή διαπιστώνουμε ότι :

Ο έλεγχος t-test μας έδειξε ότι δεν υπάρχει σημαντική διαφορά ( t = -0.71234, p-value = 0.477) μεταξύ της διαφοράς των πόντων που σημειώθηκαν στο πρώτο δεκάλεπτο για τις ομάδες που αγωνίστηκαν εντός έδρας σε σύγκριση με τις ομάδες που αγωνίστηκαν εκτός έδρας, ενώ υπάρχει σημαντική διαφορά μεταξύ της διαφοράς των πόντων που σημειώθηκαν στο δεύτερο δεκάλεπτο (t = -2.3821, p-value =0.018) , στο τρίτο δεκάλεπτο ( t = -3.3076 , p-value = 0.001086) και στο τέταρτο δεκάλεπτο (t = -2.4255, p-value =0.01603) για τις ομάδες που αγωνίστηκαν εντός έδρας σε σύγκριση με τις ομάδες που αγωνίστηκαν εκτός έδρας.

Η διαφορά των συνολικών πόντων, όπως είναι λογικό, είναι στατιστικά σημαντική ανάμεσα στην γηπεδούχο και την φιλοξενούμενη ομάδα (t = -4.3212, p-value = $2.281*10^{-5}$).

Τέλος, δεν υπάρχει σημαντική διαφορά μεταξύ των σουτ 2 πόντων (t = -1.3281 , p-value =0.1854), των σουτ 3 πόντων (t = -1.0489 , p-value = 0.2953) ενώ υπάρχει σημαντική διαφορά όσον αφορά τον ειδικό δείκτη αξιολόγησης (t = -4.1248 , p-value = $5.131*10^{-5}$ ) συγκριτικά με την έδρα.

Για τις υπόλοιπες μεταβλητές που δεν ακολουθούν την κανονική κατανομή θα προχωρήσουμε στην ανάλυση με τον έλεγχο Mann-Whitney U.

```
wilcox.test(Pts.Q2.~HOME.AWAY)

        Wilcoxon rank sum test with continuity correction

data:  Pts.Q2. by HOME.AWAY

W = 6295.5, p-value = 0.09214

alternative hypothesis: true location shift is not equal to 0
```

Ο έλεγχος Mann-Whitney U έδειξε ότι δεν απορρίπτουμε την μηδενική μας υπόθεση, δηλαδή δεν υπάρχει σημαντική διαφορά ( W = 6295.5, p-value = 0.09214) μεταξύ των πόντων που σημειώθηκαν στο δεύτερο δεκάλεπτο για τις ομάδες που αγωνίστηκαν εντός έδρας σε σύγκριση με τις ομάδες που αγωνίστηκαν εκτός έδρας.

```
wilcox.test(Pts.Q4.~HOME.AWAY)

        Wilcoxon rank sum test with continuity correction

data:  Pts.Q4. by HOME.AWAY

W = 6417.5, p-value = 0.1451

alternative hypothesis: true location shift is not equal to 0
```

Ο έλεγχος Mann-Whitney U έδειξε ότι δεν απορρίπτουμε την μηδενική μας υπόθεση, δηλαδή δεν υπάρχει σημαντική διαφορά ( W = 6417.5, p-value = 0.1451) μεταξύ των πόντων που σημειώθηκαν στο τέταρτο δεκάλεπτο για τις ομάδες που αγωνίστηκαν εντός έδρας σε σύγκριση με τις ομάδες που αγωνίστηκαν εκτός έδρας.

Με παρόμοιο τρόπο για τις υπόλοιπες μεταβλητές που δεν ακολουθούν την κανονική κατανομή διαπιστώνουμε ότι :

Ο έλεγχος Mann-Whitney U έδειξε ότι δεν υπάρχει διαφορά μεταξύ των πόντων που σημειώθηκαν στην παράταση (W = 7194, p-value = 0.9738), των ελεύθερων βολών (W = 6346, p-value = 0.1124), των επιθετικών (W = 7169.5, p-value = 0.9554),των αμυντικών (W = 6629.5, p-value = 0.2881) και των συνολικών ριμπάουντ ( W = 6831.5, p-value = 0.4932),των κλεψιμάτων ( W = 6322, p-value = 0.1001),των συνολικών κοψιμάτων υπέρ ( W = 6455, p-value = 0.1601) και κατά (W = 7861, p-value = 0.2127), συγκριτικά με τον παράγοντα έδρα.

Τέλος, από τα αποτελέσματα του ελέγχου διαπιστώνουμε ότι υπάρχει σημαντική διαφορά μεταξύ των ασσιστ ( W = 5656 , p-value = 0.003989), των λαθών ( W = 8448.5

, p-value = 0.0198) καθώς και των συνολικών φάουλ που διέπραξε (W = 8438.5, p-value = 0.02088) και που δέχθηκε (W= 6039.5, p-value = 0.03037) για τις ομάδες που αγωνίστηκαν εντός έδρας σε σύγκριση με τις ομάδες που αγωνίστηκαν εκτός έδρας.

### 5.9 Random Forest για regular season 2015-2016

Θα εφαρμόσουμε ξανά την συγκεκριμένη μέθοδο για να διαπιστώσουμε ποιες μεταβλητές επηρεάζουν σημαντικά την έδρα της ομάδας. Για να γίνει αυτό αρχικά θα πρέπει να χωρήσουμε τα δεδομένα μας σε train και validation set. Η αναλογία επιλέγουμε να είναι 70 % για το training sample και 30% για το validation sample, όπως και για την προηγούμενη σεζον.

Στη συνέχεια θα δημιουργήσουμε το random forest μοντέλο με προκαθορισμένο αριθμό παραμέτρων. Φυσικά μπορούμε να τροποποιήσουμε το μοντέλο μας αλλάζοντας την μεταβλητή ntree (αριθμός των δέντρων που θα χρησιμοποιηθούν), καθώς και των αριθμό των μεταβλητών που επιλέχθηκαν με τυχαίο τρόπο σε κάθε στάδιο (mtry) .

Έτσι λοιπόν έχουμε,

```
# Create a Random Forest model with default parameters

model1 <- randomForest(HOME.AWAY ~ ., data = TrainSet, importance = TRUE)

model1

Call:

 randomForest(formula = HOME.AWAY ~ ., data = TrainSet, importance = TRUE)

               Type of random forest: classification

                     Number of trees: 500

No. of variables tried at each split: 5

        OOB estimate of  error rate: 42.26%

Confusion matrix:

   0  1 class.error

0 45 35   0.4375000

1 36 52   0.4090909
```

Όπως βλέπουμε ο αριθμός των δέντρων είναι 500 και οι μεταβλητές που δοκιμάστηκαν σε κάθε διάσπαση είναι 5

Το out of bag (OOB) error ισούται με 42.26% συνεπώς το ποσοστό ορθής ταξινόμησης (accuracy) του συγκεκριμένου μοντέλου ισούται με 1 – OOB = **57.74%**

# Fine tuning parameters of Random Forest model

model2 <- randomForest(HOME.AWAY ~ ., data = TrainSet, ntree = 500, mtry = 5, importance = TRUE)

model2

Call:

randomForest(formula = HOME.AWAY ~ ., data = TrainSet, ntree = 500, mtry = 6, importance = TRUE)

Type of random forest: classification

Number of trees: 500

No. of variables tried at each split: 6

OOB estimate of error rate: 44.64%

Confusion matrix:

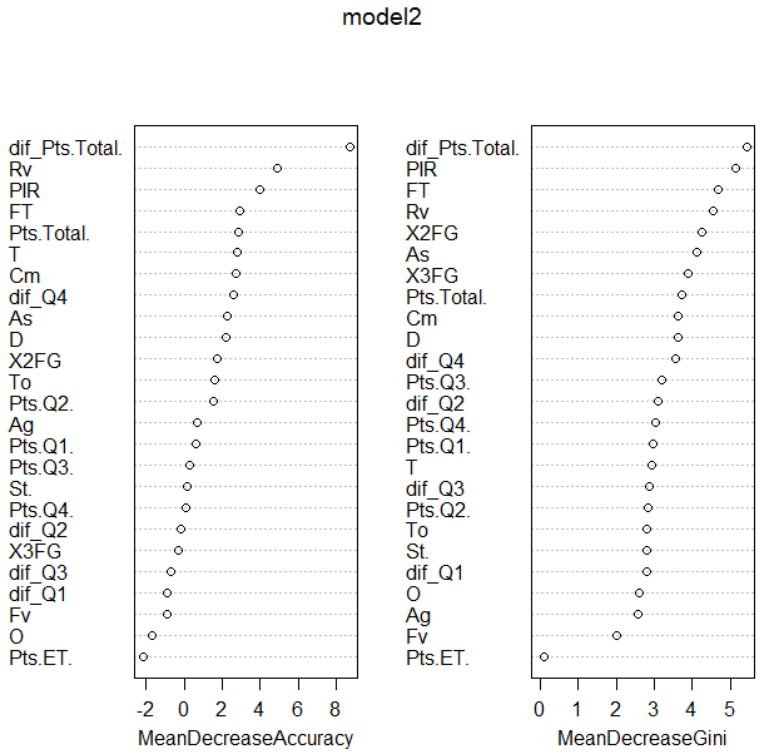
0 1 class.error

0 43 37 0.4625000

1 38 50 0.4318182

Όπως βλέπουμε τώρα που αυξήσαμε τον αριθμό των παραμέτρων από τους 5 στους 6 , το out of bag error αυξήθηκε από 42.26% σε 44.64%. έτσι λοιπόν σε αυτή την περίπτωση το ποσοστό ορθής ταξινόμησης μειώθηκε σε **55.36%**

Τέλος, από τα παρακάτω διαγράμματα βλέπουμε την σημαντικότητα των μεταβλητών κατά σειρά :




Σχήμα 5.13 Σημαντικότητα των μεταβλητών με την μέθοδο RF (2015-2016)

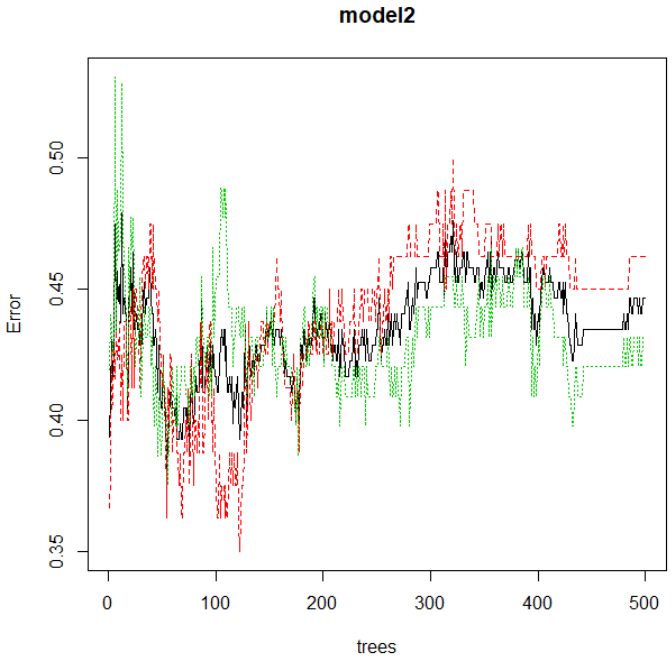
Στη συνέχεια θα δοκιμάσουμε διάφορα πλήθη μεταβλητών και πιο συγκεκριμένα από 1 μέχρι 10 για να οδηγηθούμε σε αυτό που δίνει το μεγαλύτερο επίπεδο ορθής ταξινόμησης :

| Number of variables tried at each split (mtry) | Out of bag error (OOB) | Accuracy of the model |
|---|---|---|
| 1 | 42.26% | 57.74% |
| 2 | 41.07% | 58.93% |
| 3 | 41.67% | 58.33% |
| 4 | 43.45% | 56.55% |
| 5 | 42.26% | 57.74% |
| 6 | 44.64% | 55.36% |
| 7 | 44.05% | 55.95% |
| 8 | 41.67% | 58.33% |
| 9 | 45.83% | 54.17% |

| | | |
|---|---|---|
| 10 | 44.64% | 55.36% |

Έτσι λοιπόν το πλήθος των μεταβλητών που δίνουν το μεγαλύτερο ποσοστό ορθής ταξινόμησης είναι 2 ή 3 ή 8. Συνήθως θέλουμε όσο λιγότερες μεταβλητές για αυτό θα μείνουμε με την επιλογή mtry = 2. Έτσι λοιπόν, ο αλγόριθμος random forest, προσαρμόστηκε καλύτερα όταν χρησιμοποιήσαμε 2 μεταβλητές και πιο συγκεκριμένα κατά σειρά σημαντικότητας, την συνολική διαφορά των πόντων και τα φάουλ που δέχθηκε η ομάδα (Σχήμα 5.10).

Τέλος, από το παρακάτω διάγραμμα μπορούμε να πάρουμε μια αναλυτική εικόνα και για τον κατάλληλο αριθμό των δέντρων σε σύγκριση με OOB error :




Σχήμα 5.14 Error vs trees με την μέθοδο RF (2015-2016)

# ΚΕΦΑΛΑΙΟ 6

## Συμπεράσματα

Για τις ανάγκες αυτής της ανάλυσης, συλλέξαμε στοιχεία από το επίσημο site της Euroleague. Τα δεδομένα μας αφορούν όλους τους αγώνες για την κανονική περίοδο τις χρονιές 2014-2015 και 2015-2016 και πιο συγκεκριμένα τα συνολικά παιχνίδια που εξετάσθηκαν ήταν 240. Η μελέτη επικεντρώθηκε στην εύρεση αυτών των στατιστικών στοιχείων που ήταν καθοριστικά για την απόδοση της κάθε ομάδας κατά την διάρκεια αυτών των παιχνιδιών. Για να εξετάσουμε την εξέλιξη του σκορ σε κάθε αγώνα χρησιμοποιήσαμε μεταβλητές όπως, οι συνολικοί πόντοι, τα ποσοστά ευστοχίας στις ελεύθερες βολές, στα σουτ 2 και 3 πόντων, οι ασσιστ, τα λάθη, τα κλεψίματα, τα ριμπάουντ αλλά και μεταβλητές που παρείχαν πιο συνδυαστική πληροφορία, όπως ο ειδικός δείκτης αξιολόγησης της Euroleague. Στην πρώτη προσέγγιση της ανάλυσης η μεταβλητή απόκρισης που χρησιμοποιήσαμε ήταν η πρόκριση ή όχι της ομάδας στην επόμενη φάση του θεσμού, ενώ στην συνέχεια δώσαμε βαρύτητα στο ερώτημα αν η έδρα της ομάδας επηρεάζει το τελικό αποτέλεσμα ενός αγώνα και χρησιμοποιήσαμε κάποιες επιπλέον μεταβλητές, όπως οι πόντοι ανά περίοδο και η διαφορά των πόντων.

Συγκεκριμένα, στο 2$^{ο}$ κεφάλαιο παρουσιάσαμε αναλυτικά κάποια περιγραφικά μέτρα για τις επεξηγηματικές μεταβλητές μας σε σχέση με την μεταβλητή απόκρισης qualified. Αρχικά έγινε μία διερεύνηση αν υπάρχουν ελλιπείς τιμές στα δεδομένα που θα χρησιμοποιήσουμε για την ανάλυση και διαπιστώσαμε πώς δεν υπάρχουν σε καμία από τις 2 χρονιές που εξετάζουμε.

Στη συνέχεια για την χρονιά 2014-2015, είδαμε ότι οι ομάδες που προκρίθηκαν στην επόμενη φάση έχουν υψηλότερο μέσο όρο πόντων (79.13) και υψηλότερο δείκτη αξιολόγησης κατά μέσο όρο (880.9) σε σχέση με τις ομάδες που δεν προκρίθηκαν στην επόμενη φάση της διοργάνωσης (75.85 και 781.5 αντίστοιχα). Οι ομάδες που δεν προκρίθηκαν είχαν περισσότερα (70.88) κλεψίματα κατά μέσο όρο από τις ομάδες που δεν συνέχισαν στον θεσμό. Ενδιαφέρον παρουσιάζουν και τα στατιστικά για τα σουτ των παιχτών των 24 ομάδων. Όσον αφορά τα σουτ 2 πόντων οι ομάδες που προκρίθηκαν είχαν ποσοστό ευστοχίας ίσο με 51.69% ενώ οι 8 ομάδες που δεν προκρίθηκαν είχαν ποσοστό ευστοχίας 49.90%. Για τα σουτ 3 πόντων είχαμε ποσοστά ευστοχίας 35.55% και 34.94% αντίστοιχα για τις ομάδες που πέρασαν στην επόμενη φάση και για αυτές που δεν τα κατάφεραν. Επιπλέον, έγινε γραφική ανάλυση για αρκετές μεταβλητές ώστε να γίνει περισσότερο κατανοητό ποια στατιστικά στοιχεία είναι κρίσιμα για το αν η ομάδα προκρίθηκε ή όχι. Αρχικά έγιναν κάποια διαγράμματα διασποράς από τα οποία διαπιστώσαμε ότι οι πόντοι επηρεάζουν τους ειδικούς δείκτες αξιολόγησης αλλά γενικότερα δεν βγήκε κάποια ξεκάθαρη πληροφορία καθώς τα περισσότερα σημεία βρίσκονταν διασκορπισμένα στον χώρο. Μια δεύτερη γραφική ανάλυση έγινε χρησιμοποιώντας θηκογράμματα, ιστογράμματα καθώς και violin plots. Αυτό που παρατηρήσαμε ήταν ότι στις ελεύθερες βολές οι ομάδες που δεν έχουν προκριθεί στην επόμενη φάση φαίνεται πως έχουν μεγαλύτερο εύρος τιμών, δεν φάνηκε κάποια ασυμμετρία θετική ή αρνητική και σχεδόν σε όλες τις περιπτώσεις οι

ομάδες που προκρίθηκαν παρουσίασαν υψηλότερη απόδοση στις μεταβλητές που εξετάστηκαν. Χαρακτηριστικό είναι το παράδειγμα των ριμπάουντ, καθώς οι ομάδες που είχαν περισσότερα συνολικά ριμπάουντ πέρασαν στην επόμενη φάση του θεσμού. Αντίστοιχα και για την επόμενη χρονιά παρουσιάστηκαν τόσο τα περιγραφικά μέτρα των μεταβλητών όσο και η γραφική ανάλυση τους. Πιο συγκεκριμένα, οι ομάδες που προκρίθηκαν στην επόμενη φάση έχουν υψηλότερο μέσο όρο πόντων (78.84) και υψηλότερο δείκτη αξιολόγησης κατά μέσο όρο (887.4) σε σχέση με τις ομάδες που δεν προκρίθηκαν στην επόμενη φάση της διοργάνωσης (71.39 και 725.6 αντίστοιχα). Συγκριτικά με την προηγούμενη χρονιά, αξίζει να αναφέρουμε πώς ο μέσος όρος των πόντων καθώς και ο ειδικός δείκτης αξιολόγησης για τις ομάδες που δεν προκρίθηκαν στην επόμενη φάση μειώθηκε αισθητά. Όσον αφορά τα συνολικά ριμπάουντ 355 ήταν περίπου κατά μέσο όρο για τις ομάδες που προκρίθηκαν (11 λιγότερα από την προηγούμενη χρονιά) και περίπου 321 για τις ομάδες που δεν προκρίθηκαν. Ειδικότερα, μία ομάδα για να προκριθεί στην επόμενη φάση πρέπει να έχει 330 ριμπάουντ και πάνω. Ακόμη, την σεζόν 2015-2016, τα σουτ 3 πόντων καθώς και τα λάθη παρουσιάζουν μεγάλη ασσυμετρία στις ομάδες που δεν προκρίθηκαν στην επόμενη φάση της διοργάνωσης. Στις προσπάθειες για σουτ 2 πόντων οι ομάδες που προκρίθηκαν είχαν ποσοστό ευστοχίας ίσο με 52.15% , ελαφρώς μεγαλύτερο σε σύγκριση με την προηγούμενη χρονιά, ενώ οι 8 ομάδες που δεν προκρίθηκαν είχαν ποσοστό ευστοχίας 47.99%, δηλαδή περίπου 2 ποσοστιαίες μονάδες λιγότερες σε σχέση με την προηγούμενη χρονιά. Για τα σουτ 3 πόντων είχαμε ποσοστά ευστοχίας 36.56% και 33.85% αντίστοιχα για τις ομάδες που πέρασαν στην επόμενη φάση και για αυτές που δεν τα κατάφεραν.

Στο 3ο κεφάλαιο, εξετάσαμε τις σχέσεις των μεταβλητών μεταξύ τους, μέσω των συντελεστών συσχέτισης του Kendall. Για την χρονιά 2014-2015 διαπιστώθηκε ότι τα ποσοστά ευστοχίας στα σουτ 2 πόντων είχαν έντονη θετική συσχέτιση με την διαφορά των συνολικών πόντων. Αξίζει ακόμη να αναφέρουμε πως το ποσοστό ευστοχίας για τα σουτ 3 πόντων, το ποσοστό ευστοχίας ελευθέρων βολών, τα συνολικά επιθετικά ριμπάουντ, οι ασσιστ, τα κλεψίματα, τα λάθη, τα κοψίματα κατά της ομάδας και τα φάουλ δεν φαίνεται να σχετίζονται έντονα θετικά ή αρνητικά με κάποια από τις υπόλοιπες μεταβλητές. Στην συνέχεια, ελέγξαμε αν οι παραπάνω μεταβλητές είναι στατιστικά σημαντικές συγκριτικά με την μεταβλητή qualified, και βρέθηκε ότι τα συνολικά ριμπάουντ έχουν σημαντικό ρόλο στην εξέλιξη του σκορ. Για την επόμενη χρονιά, παρατηρείται μια οριακή αύξηση των περισσότερων συντελεστών συσχέτισης. Επιπλέον, βλέπουμε πως αν το ποσοστό ευστοχίας στα σουτ 2 πόντων είναι μεγαλύτερο από 50% ή ο ειδικός δείκτης αξιολόγησης είναι μεγαλύτερος του 800 τότε υπάρχει ισχυρή ένδειξη ότι η ομάδα θα προκριθεί στο top-16. Αξίζει να αναφέρουμε πως και τα συνολικά φάουλ που δέχθηκε μια ομάδα μπορεί να αποτελέσει έναν σημαντικό παράγοντα πρόκρισης για την συγκεκριμένη περίοδο. Εφαρμόζοντας τον μη-παραμετρικό έλεγχο κανονικότητας Shapiro-Wilk για κάθε μία από τις ανεξάρτητες ποσοτικές μας μεταβλητές την χρονιά 2014-2015 είδαμε ότι μόνο οι ασσιστ και ο δείκτης PIR δεν ακολουθούν την κανονική κατανομή και την χρονιά 2015-2016 εξαίρεση αποτέλεσαν μόνο τα συνολικά κοψίματα κατά της ομάδας. Τέλος, η

διαχωριστική ανάλυση παρουσίασε χαμηλό ποσοστό ορθής ταξινόμησης (περίπου 50 %) οπότε αποφασίστηκε ότι θα χρειαστεί περαιτέρω ανάλυση αφού ίσως δεν είναι η κατάλληλη μέθοδος για να χρησιμοποιήσουμε στην περίπτωση αυτή.

Στο 4ο κεφάλαιο, προχωρήσαμε στην σύγκριση των μοντέλων λογιστικής παλινδρόμησης logit, probit, cauchit, cloglog για να δούμε ποια χαρακτηριστικά επηρεάζουν περισσότερο την εξέλιξη των 24 ομάδων μέσα στην διοργάνωση. Για την περίοδο 2014-2015 διαπιστώσαμε λοιπόν ότι, τα συνολικά ριμπάουντ φαίνεται να έχουν καθοριστικό ρόλο ( p-value = 0.0139) στην έκβαση του αποτελέσματος . Επιπλέον εφαρμόστηκαν διάφοροι μέθοδοι επιλογής των πιο σημαντικών μεταβλητών (feature selection). Ο αλγόριθμος random forest έδειξε ότι οι 2 πιο σημαντικές μεταβλητές που πρέπει να εισαχθούν στο μοντέλο μας είναι τα συνολικά ριμπάουντ καθώς και η συνολική διαφορά των πόντων. Μια παραλλαγή του αλγορίθμου random forest, η μέθοδος Boruta, επιβεβαίωσε το παραπάνω αποτέλεσμα. Η τελική μας επιλογή ήταν να φτιάξουμε ένα λογιστικό μοντέλο με μόνη ερμηνευτική μεταβλητή τα συνολικά ριμπάουντ. Από τα 4 μοντέλα, το μοντέλο cloglog φαίνεται πως υποστηρίζεται καλύτερα συγκριτικά με τα υπόλοιπα, έχοντας υψηλή προβλεπτική ικανότητα (AUC=0.9219) όπως και ποσοστό ορθής ταξινόμησης (83.33%) με την επιλογή του βέλτιστου κατωφλιού $\pi_0$, χρησιμοποιώντας την μέθοδο του Youden. Με παρόμοια λογική την επόμενη χρονιά είδαμε πως σημαντικό ρολό για το αν η ομάδα προκρίθηκε ή όχι στην επόμενη φάση φαίνεται πως έχει και το ποσοστό ευστοχίας στα σουτ 2 πόντων που πραγματοποιήθηκαν. Συνδυάζοντας ξανά διαφορετικές μεθόδους επιλέξαμε να φτιάξουμε ένα λογιστικό μοντέλο με ερμηνευτικές τα συνολικά ριμπάουντ και το ποσοστό ευστοχίας στα σουτ 2 πόντων, ώστε να διερευνήσουμε με μεγαλύτερη ακρίβεια αν όντως επιδρούν σημαντικά στην εξέλιξη της διοργάνωσης. Από τα 4 μοντέλα λογιστικής παλινδρόμησης, το μοντέλο probit φαίνεται πως υποστηρίζεται καλύτερα συγκριτικά με τα υπόλοιπα, αν λάβουμε υπόψιν μας το πληροφοριακό κριτήριο του Akaike (AIC = 12.7) καθώς και τα υπόλοιπα Pseudo-$R^2$ για αυτή την χρονιά, έχοντας υψηλή προβλεπτική ικανότητα (AUC=0.9843) όπως και ποσοστό ορθής ταξινόμησης (95.83%) με την επιλογή του βέλτιστου κατωφλιού $\pi_0$, χρησιμοποιώντας την μέθοδος του Youden. Η έρευνα από τους Beckler, Wang και Papamichael (2008-09) οι οποίοι εφάρμοσαν applied machine learning τεχνικές για να προβλέψουν το τελικό αποτέλεσμα για αγώνες του NBA, χρησιμοποιώντας λογιστική παλινδρόμηση (Logistic Regression) είχε αντίστοιχο ποσοστό ορθής ταξινόμησης 68.1%, Οι υψηλές τιμές των ελέγχων καλής προσαρμογής και η μέθοδος cross-validation μας διασφαλίζουν ότι το μοντέλο μας είναι αρκετά αξιόπιστο και μπορεί να γενικευτεί και σε ένα ευρύτερο σετ δεδομένων που αφορούν αγώνες μπάσκετ.

Τέλος, στο 5ο κεφαλαίο έγινε μια διαφορετική προσέγγιση σε σχέση με τα προηγούμενα εστιάζοντας στα χαρακτηριστικά που επηρεάζουν περισσότερο την εξέλιξη των 24 ομάδων μέσα στην διοργάνωση με βάση τον παράγοντα έδρα.

Από την περιγραφική ανάλυση για την χρονιά 2014-2015 διαπιστώσαμε πως κατά μέσο όρο όλες οι φιλοξενούμενες ομάδες βρίσκονταν πίσω στο σκορ και είχαν έστω και οριακά καλύτερο ποσοστό ευστοχίας κατά μέσο όρο στις ελεύθερες βολές από ότι οι

ομάδες που αγωνίστηκαν εντός έδρας. Ο ειδικός δείκτης αξιολόγησης της διοργάνωσης, όπως θα περίμενε κανείς είναι μεγαλύτερος για τις ομάδες που αγωνίζονταν εντός έδρας κατά μέσο όρο. (89.72 έναντι 79.84). Η γραφική ανάλυση επιβεβαίωσε τα παραπάνω συμπεράσματα. Στην συνέχεια εφαρμόστηκαν έλεγχοι στατιστικής σημαντικότητας για τις μεταβλητές που ακολουθούν την κανονική κατανομή (t-test) ή όχι (Mann-Whitney). Ο έλεγχος t-test έδειξε ότι είναι σημαντική η διαφορά μόνο μεταξύ των πόντων που σημειώθηκαν στο πρώτο δεκάλεπτο για τις ομάδες που αγωνίστηκαν εντός έδρας σε σύγκριση με τις ομάδες που αγωνίστηκαν εκτός έδρας, αλλά όχι και στο δεύτερο και στο τέταρτο. Το τρίτο δεκάλεπτο δεν φαίνεται πως ακολουθεί την κανονική κατανομή. Ακόμη, δεν υπάρχει σημαντική διαφορά μεταξύ των σουτ 2 και 3 πόντων καθώς και των αμυντικών ριμπάουντ συγκριτικά με την έδρα. Αντίστοιχα, ο μη παραμετρικός έλεγχος Mann Whitney έδειξε ότι οι μεταβλητές που διαφοροποιούνται περισσότερο ως προς τον παράγοντα έδρα, είναι, οι πόντοι που σημειώθηκαν στο τρίτο δεκάλεπτο, οι ασσιστ καθώς και ο ειδικός δείκτης αξιολόγησης. Ο αλγόριθμος random forest, προσαρμόστηκε καλύτερα όταν χρησιμοποιήσαμε 5 μεταβλητές και πιο συγκεκριμένα κατά σειρά σημαντικότητας, την διαφορά των πόντων στο τρίτο δεκάλεπτο, την συνολική διαφορά των πόντων, τα φάουλ που δέχθηκε η ομάδα, τα σουτ 2 πόντων και τα κλεψίματα.

Από την περιγραφική ανάλυση για την χρονιά 2015-2016 διαπιστώσαμε πως κατά μέσο όρο όλες οι φιλοξενούμενες ομάδες βρίσκονταν πίσω στο σκορ και η μικρότερη διαφορά πόντων παρατηρείται κατά τη διάρκεια της πρώτης περιόδου. Το ποσοστό ευστοχίας στις ελεύθερες βολές των γηπεδούχων ομάδων αυξήθηκε σημαντικά σε σχέση με την regular season 2014-2015. Όσον αφορά τα ριμπάουντ δεν υπάρχει μία ξεκάθαρη εικόνα. Ο έλεγχος t-test έδειξε ότι είναι σημαντική η διαφορά μόνο μεταξύ των πόντων που σημειώθηκαν στο τρίτο δεκάλεπτο για τις ομάδες που αγωνίστηκαν εντός έδρας σε σύγκριση με τις ομάδες που αγωνίστηκαν εκτός έδρας, αλλά όχι και στο πρώτο. Δεν υπάρχει σημαντική διαφορά μεταξύ των σουτ 2 και 3 πόντων. . Αντίστοιχα, ο μη παραμετρικός έλεγχος Mann Whitney έδειξε ότι οι μεταβλητές που διαφοροποιούνται περισσότερο ως προς τον παράγοντα έδρα είναι, οι ασσιστ, τα λάθη, καθώς και τα συνολικά φάουλ που διέπραξε ή δέχθηκε η ομάδα. Ο αλγόριθμος random forest, προσαρμόστηκε καλύτερα όταν χρησιμοποιήσαμε 2 μεταβλητές και πιο συγκεκριμένα κατά σειρά σημαντικότητας, την συνολική διαφορά των πόντων και τα φάουλ που δέχθηκε η ομάδα.

## ΠΑΡΑΡΤΗΜΑ

Παρακάτω επισυνάπτεται ο κώδικας που χρησιμοποιήθηκε για την πραγματοποίηση της μελέτης :

### Π1. R script

```
install.packages("xlsx")
library("xlsx")
mydata<-read.xlsx("C:/Users/Dionysis/Desktop/thesis.xlsx",5,header=TRUE)
mydata
mydata<-mydata[,-20]
mydata
names(mydata)
attach(mydata)

##missing values
install.packages("DataExplorer")
library(DataExplorer)
plot_missing(mydata)

#descriptive ksexwrista
library(purrr)
mydata[,c(4:19)]%>%split(mydata$qualified)%>%map(summary)

#descriptive total
summary(mydata[,c(4:19)])

#scatterplot
pairs(mydata[,c(4:19)])

par(mfrow=c(3,3))
plot(Pts.Total..M.O,PIR.TOTAL)
abline(lm(PIR.TOTAL~Pts.Total..M.O),col="red")
plot(Pts.Total..M.O,Cm.TOTAL)
abline(lm(Cm.TOTAL~Pts.Total..M.O),col="red")
plot(Pts.Total..M.O,Fv.TOTAL)
abline(lm(Fv.TOTAL~Pts.Total..M.O),col="red")
plot(Pts.Total..M.O,Rv.TOTAL)
abline(lm(Rv.TOTAL~Pts.Total..M.O),col="red")
plot(Pts.Total..M.O,To.TOTAL)
abline(lm(To.TOTAL~Pts.Total..M.O),col="red")
plot(Pts.Total..M.O,T.TOTAL)
abline(lm(T.TOTAL~Pts.Total..M.O),col="red")
plot(Pts.Total..M.O,FT.M.O)
abline(lm(FT.M.O~Pts.Total..M.O),col="red")
plot(Pts.Total..M.O,X2FG.M.O)
abline(lm(X2FG.M.O~Pts.Total..M.O),col="red")
plot(Pts.Total..M.O,X3FG.M.O)
abline(lm(X3FG.M.O~Pts.Total..M.O),col="red")

#boxplots
par(mfrow=c(3,3))
b1<-boxplot(Pts.Total..M.O~qualified,data=mydata[,c(4:19)],
   main="Qualified vs Total Points",xlab="qualified or not",ylab="Total Points",col="gold")
b2<-boxplot(PIR.TOTAL~qualified,data=mydata[,c(4:19)],
   main="Qualified vs PIR",xlab="qualified or not",ylab="PIR",col="gold")
```

```
b3<-boxplot(X2FG.M.O~qualified,data=mydata[,c(4:19)],
   main="Qualified vs Total FGX2",xlab="qualified or not",ylab="X2FG",col="gold")
b4<-boxplot(X3FG.M.O~qualified,data=mydata[,c(4:19)],
   main="Qualified vs FGX3",xlab="qualified or not",ylab="X3FG",col="gold")
b5<-boxplot(FT.M.O~qualified,data=mydata[,c(4:19)],
   main="Qualified vs FT",xlab="qualified or not",ylab="FT",col="gold")
b6<-boxplot(T.TOTAL~qualified,data=mydata[,c(4:19)],
   main="Qualified vs T.TOTAL",xlab="qualified or not",ylab="TOTAL REBOUND",col="gold")
b7<-boxplot(As.TOTAL~qualified,data=mydata[,c(4:19)],
   main="Qualified vs Assist",xlab="qualified or not",ylab="Assist",col="gold")
b8<-boxplot(To.TOTAL~qualified,data=mydata[,c(4:19)],
   main="Qualified vs To.Total",xlab="qualified or not",ylab="Turnover",col="gold")
b9<-boxplot(Rv.TOTAL~qualified,data=mydata[,c(4:19)],
   main="Qualified vs Rv.Total",xlab="qualified or not",ylab="Rv",col="gold")

#histogram
par(mfrow=c(2,1))
hist(Pts.Total..M.O[qualified==1],col="gold")
hist(Pts.Total..M.O[qualified==0],col="gold")

#violin plots
install.packages("vioplot")
library(vioplot)
vioplot(Pts.Total..M.O~qualified,col="gold")

###new_visual
install.packages("ggplot2")
library(ggplot2)
mydata$qualified<-as.factor(mydata$qualified)
qplot(Pts.Total..M.O,X2014.2015,data=mydata,color=qualified)
qplot(T.TOTAL,X2014.2015,data=mydata,color=qualified)

#cor ana group(year 2014-2015)
group_a<-mydata[1:6,-1];group_a
group_b<-mydata[7:12,-1];group_b
group_c<-mydata[13:18,-1];group_c
group_d<-mydata[19:24,-1];group_d
round(cor(group_a,method="kendall"),2)
round(cor(group_b,method="kendall"),2)
round(cor(group_c,method="kendall"),2)
round(cor(group_d,method="kendall"),2)

#cor tin xronia 2014-2015
year14_15<-mydata[1:24,-1];year14_15
round(cor(year14_15,method="kendall"),2)
source("http://www.sthda.com/upload/rquery_cormat.r")
require("corrplot")
rquery.cormat(year14_15)
cormat<-rquery.cormat(year14_15, graphType="heatmap")
rquery.cormat(year14_15, type="flatten", graph=FALSE)

#chi-square
chis<-read.xlsx("C:/Users/Dionysis/Desktop/thesis.xlsx",7,header=TRUE)
chis
names(chis)
library(MASS)
tb1<-table(chis$Pts.Total..M.O,chis$qualified);tb1
chisq1<-chisq.test(tb1);chisq1

install.packages("visualize")
```

```
library(visualize)
visualize.chisq(stat = 0.63025, df = 1, section = "upper")

tb2<-table(chis$dif_Pts.Total..M.O,chis$qualified);tb2
chisq2<-chisq.test(tb2);chisq2
tb3<-table(chis$X2FG.M.O,chis$qualified);tb3
chisq3<-chisq.test(tb3);chisq3
tb4<-table(chis$X3FG.M.O,chis$qualified);tb4
chisq4<-chisq.test(tb4);chisq4
tb5<-table(chis$FT.M.O,chis$qualified);tb5
chisq5<-chisq.test(tb5);chisq5
tb6<-table(chis$O.TOTAL,chis$qualified);tb6
chisq6<-chisq.test(tb6);chisq6
tb7<-table(chis$D.TOTAL,chis$qualified);tb7
chisq7<-chisq.test(tb7);chisq7
tb8<-table(chis$T.TOTAL,chis$qualified);tb8
chisq8<-chisq.test(tb8);chisq8
tb9<-table(chis$As.TOTAL,chis$qualified);tb9
chisq9<-chisq.test(tb9);chisq9
tb10<-table(chis$St..TOTAL,chis$qualified);tb10
chisq10<-chisq.test(tb10);chisq10
tb11<-table(chis$To.TOTAL,chis$qualified);tb11
chisq11<-chisq.test(tb11);chisq11
tb12<-table(chis$Fv.TOTAL,chis$qualified);tb12
chisq12<-chisq.test(tb12);chisq12
tb13<-table(chis$Ag.TOTAL,chis$qualified);tb13
chisq13<-chisq.test(tb13);chisq13
tb14<-table(chis$Cm.TOTAL,chis$qualified);tb14
chisq14<-chisq.test(tb14);chisq14
tb15<-table(chis$Rv.TOTAL,chis$qualified);tb15
chisq15<-chisq.test(tb15);chisq15
tb16<-table(chis$PIR.TOTAL,chis$qualified);tb16
chisq16<-chisq.test(tb16);chisq16
df<-
data.frame(chisq2$p.value,chisq3$p.value,chisq4$p.value,chisq5$p.value,chisq6$p.value,chisq7$p.v
alue,chisq8$p.value,chisq9$p.value,chisq10$p.value,chisq11$p.value,chisq12$p.value,chisq13$p.va
lue,chisq14$p.value,chisq15$p.value,chisq16$p.value);df

names(chis)
install.packages("DescTools")
library(DescTools)
SomersDelta(chis$ranks,chis$qualified,direction  = "column",conf.level = 0.95)

#normality test for independent variables
shapiro.test(mydata$Pts.Total..M.O)
shapiro.test(mydata$dif_Pts.Total..M.O)
shapiro.test(mydata$X2FG.M.O)
shapiro.test(mydata$X3FG.M.O)
shapiro.test(mydata$FT.M.O)
shapiro.test(mydata$O.TOTAL)
shapiro.test(mydata$D.TOTAL)
shapiro.test(mydata$T.TOTAL)
shapiro.test(mydata$As.TOTAL)
shapiro.test(mydata$St..TOTAL)
shapiro.test(mydata$To.TOTAL)
shapiro.test(mydata$Fv.TOTAL)
shapiro.test(mydata$Ag.TOTAL)
shapiro.test(mydata$Cm.TOTAL)
shapiro.test(mydata$Rv.TOTAL)
shapiro.test(mydata$PIR.TOTAL)
```

```
library(tidyverse)
library(MASS)
set.seed(101)
sample_n(mydata[,3:19], 10)
training_sample <- sample(c(TRUE, FALSE), nrow(mydata), replace = T, prob = c(0.6,0.4))
train <- mydata[training_sample,3:19 ];train
test <- mydata[!training_sample,3:19 ];test
lda.mydata <- lda(qualified ~
Pts.Total..M.O+dif_Pts.Total..M.O+X2FG.M.O+X3FG.M.O+FT.M.O+T.TOTAL+St..TOTAL+To.
TOTAL+Fv.TOTAL+Ag.TOTAL+Cm.TOTAL+Rv.TOTAL, train)
lda.mydata #show results
plot(lda.mydata, col = as.integer(train$qualified))
plot(lda.mydata, dimen = 1, type = "b")
lda.train <- predict(lda.mydata)
train$lda <- lda.train$class
table(train$lda,train$qualified)
lda.test <- predict(lda.mydata,test)
test$lda <- lda.test$class
table(test$lda,test$qualified)

#logit gia kathe metabliti gia na dw an einai statistika shmantiki
names(mydata)
fit1<-glm(qualified~Pts.Total..M.O,family=binomial,data=mydata);summary(fit1)
fit2<-glm(qualified~dif_Pts.Total..M.O,family=binomial,data=mydata);summary(fit2)
fit3<-glm(qualified~X2FG.M.O,family=binomial,data=mydata);summary(fit3)
fit4<-glm(qualified~X3FG.M.O,family=binomial,data=mydata);summary(fit4)
fit5<-glm(qualified~FT.M.O,family=binomial,data=mydata);summary(fit5)
fit6<-glm(qualified~O.TOTAL,family=binomial,data=mydata);summary(fit6)
fit7<-glm(qualified~D.TOTAL,family=binomial,data=mydata);summary(fit7)
fit8<-glm(qualified~T.TOTAL,family=binomial,data=mydata);summary(fit8)
fit9<-glm(qualified~As.TOTAL,family=binomial,data=mydata);summary(fit9)
fit10<-glm(qualified~St..TOTAL,family=binomial,data=mydata);summary(fit10)
fit11<-glm(qualified~To.TOTAL,family=binomial,data=mydata);summary(fit11)
fit12<-glm(qualified~Fv.TOTAL,family=binomial,data=mydata);summary(fit12)
fit13<-glm(qualified~Ag.TOTAL,family=binomial,data=mydata);summary(fit13)
fit14<-glm(qualified~Cm.TOTAL,family=binomial,data=mydata);summary(fit14)
fit15<-glm(qualified~Rv.TOTAL,family=binomial,data=mydata);summary(fit15)
fit16<-glm(qualified~PIR.TOTAL,family=binomial,data=mydata);summary(fit16)

fit17<-
glm(qualified~X2FG.M.O+T.TOTAL+As.TOTAL+St..TOTAL+Fv.TOTAL+Cm.TOTAL+Rv.TOT
AL,family=binomial,data=mydata)
summary(fit17)
step(fit17)

#feature selection gia 2014-2015

1. Random forest method

library(randomForest)
install.packages("Metrics")
library(Metrics)
fit_rf = randomForest(qualified~.,data=mydata[,-1])
# Create an importance based on mean decreasing gini
importance(fit_rf)
varImpPlot(fit_rf)

2.Stepwise Method
```

```
base.mod <- glm(qualified ~ 1 ,family=binomial, data= mydata[,-1])  # base intercept only model
all.mod <- glm(qualified ~ . ,family=binomial, data= mydata[,-1]) # full model with all predictors
stepMod <- step(base.mod, scope = list(lower = base.mod, upper = all.mod), direction = "both",
trace = 0, steps = 1000)  # perform step-wise algorithm
shortlistedVars <- names(unlist(stepMod[[1]])) # get the shortlisted variable.
shortlistedVars <- shortlistedVars[!shortlistedVars %in% "(Intercept)"]  # remove intercept
print(shortlistedVars)
```

3.Boruta Method

```
install.packages("Boruta")
library(Boruta)
# Decide if a variable is important or not using Boruta
boruta_output <- Boruta(qualified~.,data=na.omit(mydata[,-1]),doTrace=2) # perform Boruta search
boruta_signif <- names(boruta_output$finalDecision[boruta_output$finalDecision %in%
c("Confirmed", "Tentative")])  # collect Confirmed and Tentative variables
print(boruta_signif)  # significant variables
plot(boruta_output, cex.axis=.7, las=2, xlab="", main="Variable Importance")
```

4. Anova

```
anova<-aov(qualified~.,data=mydata[,-1])
summary(anova)

#GLM gia 2014-2015

fit.logit<-glm(qualified~T.TOTAL,family=binomial,data=mydata)
summary(fit.logit)
fit.probit<-glm(qualified~T.TOTAL,family=binomial(link="probit"),data=mydata)
summary(fit.probit)
fit.cauchit<-glm(qualified~T.TOTAL,family=binomial(link="cauchit"),data=mydata)
summary(fit.cauchit)
fit.cloglog<-glm(qualified~T.TOTAL,family=binomial(link="cloglog"),data=mydata)
summary(fit.cloglog)
library(BaylorEdPsych)
PseudoR2(fit.logit)
PseudoR2(fit.probit)
PseudoR2(fit.cauchit)
PseudoR2(fit.cloglog)
X2.logit=sum(residuals(fit.logit,type="pearson")^2)
X2.probit=sum(residuals(fit.probit,type="pearson")^2)
X2.cauchit=sum(residuals(fit.cauchit,type="pearson")^2)
X2.cloglog=sum(residuals(fit.cloglog,type="pearson")^2)
X2.all=c(X2.logit,X2.probit,X2.cauchit,X2.cloglog)
X2.table=rbind(X2.all,1-pchisq(X2.all,df=fit.logit$df.residual))
colnames(X2.table)=c("logit","probit","cauchit","cloglog")
rownames(X2.table)=c("X2","p")
round(X2.table,4)
D.all=cbind(fit.logit$deviance,fit.probit$deviance,fit.cauchit$deviance,fit.cloglog$deviance)
round(D.all,3)

install.packages("ResourceSelection")
library(ResourceSelection)
phat=predict(fit.logit,data.frame(T.TOTAL),type="response")
hoslem.test(qualified,phat)

pred.fit=predict(fit.cloglog,type="response")
library(pROC)
plot.roc(mydata$qualified,pred.fit)
auc(mydata$qualified,pred.fit)
```

```
coords(roc(mydata$qualified,pred.fit),"best",best.method="youden")

#accuracy 2014-2015
attach(mydata)
qualified.fit=1*(pred.fit>0.66)
classification.matrix=matrix(nrow=2,byrow=T,
+c(sum(qualified*qualified.fit),sum(qualified*(1-qualified.fit)),
+sum((1-qualified)*qualified.fit),sum((1-qualified)*(1-qualified.fit))))
rownames(classification.matrix)=c("y=1","y=0")
colnames(classification.matrix)=c("y.hat=1","y.hat=0")
classification.matrix
mean(qualified==qualified.fit)
sum(qualified*qualified.fit)/sum(qualified) #sensitivity
sum((1-qualified)*(1-qualified.fit))/sum(1-qualified) #specificity

#outliers

outliers<-boxplot.stats(T.TOTAL)$out
boxplot(T.TOTAL,main="total rebound",boxwex=0.1,horizontal=T)
mtext(paste("Outliers:",paste(outliers,collapse=",")),cex=0.6)

rebound_new<-T.TOTAL[T.TOTAL>275]
outliers<-boxplot.stats(rebound_new)$out
boxplot(total_rebound_new,main="total rebound",boxwex=0.1,horizontal=T)
mtext(paste("Outliers:",paste(outliers,collapse=",")),cex=0.6)

#YEAR 2015-2016

mydata2<-read.xlsx("C:/Users/Dionysis/Desktop/thesis.xlsx",6,header=TRUE)
mydata2
names(mydata2)

#descriptive ksexwrista
library(purrr)
mydata2[,c(4:19)]%>%split(mydata2$qualified)%>%map(summary)

#descriptive total
summary(mydata2[,c(4:19)])

#scatterplot

par(mfrow=c(3,3))
plot(Pts.Total..M.O,PIR.TOTAL)
abline(lm(PIR.TOTAL~Pts.Total..M.O),col="red")
plot(Pts.Total..M.O,Cm.TOTAL)
abline(lm(Cm.TOTAL~Pts.Total..M.O),col="red")
plot(Pts.Total..M.O,Fv.TOTAL)
abline(lm(Fv.TOTAL~Pts.Total..M.O),col="red")
plot(Pts.Total..M.O,Rv.TOTAL)
abline(lm(Rv.TOTAL~Pts.Total..M.O),col="red")
plot(Pts.Total..M.O,To.TOTAL)
abline(lm(To.TOTAL~Pts.Total..M.O),col="red")
plot(Pts.Total..M.O,T.TOTAL)
abline(lm(T.TOTAL~Pts.Total..M.O),col="red")
plot(Pts.Total..M.O,FT.M.O)
abline(lm(FT.M.O~Pts.Total..M.O),col="red")
plot(Pts.Total..M.O,X2FG.M.O)
abline(lm(X2FG.M.O~Pts.Total..M.O),col="red")
plot(Pts.Total..M.O,X3FG.M.O)
abline(lm(X3FG.M.O~Pts.Total..M.O),col="red")
```

```
#boxplots
par(mfrow=c(3,3))
b1<-boxplot(Pts.Total..M.O~qualified,data=mydata2[,c(4:19)],
   main="Qualified vs Total Points",xlab="qualified or not",ylab="Total Points",col="gold")
b2<-boxplot(St.TOTAL~qualified,data=mydata2[,c(4:19)],
   main="Qualified vs Steals Total",xlab="qualified or not",ylab="St..Total",col="gold")
b3<-boxplot(X2FG.M.O~qualified,data=mydata2[,c(4:19)],
   main="Qualified vs Total FGX2",xlab="qualified or not",ylab="X2FG",col="gold")
b4<-boxplot(X3FG.M.O~qualified,data=mydata2[,c(4:19)],
   main="Qualified vs FGX3",xlab="qualified or not",ylab="X3FG",col="gold")
b5<-boxplot(FT.M.O~qualified,data=mydata2[,c(4:19)],
   main="Qualified vs FT",xlab="qualified or not",ylab="FT",col="gold")
b6<-boxplot(O.TOTAL~qualified,data=mydata2[,c(4:19)],
   main="Qualified vs O.TOTAL",xlab="qualified or not",ylab="OFFENSIVE
REBOUND",col="gold")
b7<-boxplot(As.TOTAL~qualified,data=mydata2[,c(4:19)],
   main="Qualified vs Assist",xlab="qualified or not",ylab="Assist",col="gold")
b8<-boxplot(To.TOTAL~qualified,data=mydata2[,c(4:19)],
   main="Qualified vs To.Total",xlab="qualified or not",ylab="Turnover",col="gold")
b9<-boxplot(Rv.TOTAL~qualified,data=mydata2[,c(4:19)],
   main="Qualified vs Rv.Total",xlab="qualified or not",ylab="Rv",col="gold")

#histogram
par(mfrow=c(2,1))
hist(Pts.Total..M.O[qualified==1],col="gold")
hist(Pts.Total..M.O[qualified==0],col="gold")

#violin plots
install.packages("vioplot")
library(vioplot)
vioplot(Pts.Total..M.O~qualified,col="gold")

###new_visual
install.packages("ggplot2")
library(ggplot2)
mydata2$qualified<-as.factor(mydata2$qualified)
qplot(Pts.Total..M.O,X2015.2016,data=mydata2,color=qualified)
qplot(T.TOTAL,X2015.2016,data=mydata2,color=qualified)

#cor ana group(year 2015-2016)
group_a<-mydata2[1:6,-1];group_a
group_b<-mydata2[7:12,-1];group_b
group_c<-mydata2[13:18,-1];group_c
group_d<-mydata2[19:24,-1];group_d
round(cor(group_a,method="kendall"),2)
round(cor(group_b,method="kendall"),2)
round(cor(group_c,method="kendall"),2)
round(cor(group_d,method="kendall"),2)

#cor tin xronia 2015-2016
year15_16<-mydata2[1:24,-1];year15_16
round(cor(year15_16,method="kendall"),2)
source("http://www.sthda.com/upload/rquery_cormat.r")
require("corrplot")
rquery.cormat(year15_16)
cormat<-rquery.cormat(year15_16, graphType="heatmap")
rquery.cormat(year15_16, type="flatten", graph=FALSE)

install.packages("PerformanceAnalytics")
```

```
library("PerformanceAnalytics")
chart.Correlation(group_a, histogram=TRUE, pch=19)

#chi-square
chis<-read.xlsx("C:/Users/Dionysis/Desktop/thesis.xlsx",8,header=TRUE)
chis
names(chis)
library(MASS)
tb1<-table(chis$Pts.Total..M.O,chis$qualified);tb1
chisq1<-chisq.test(tb1);chisq1
visualize.chisq(stat = 1.5474, df = 1, section = "upper")

tb2<-table(chis$dif_Pts.Total..M.O,chis$qualified);tb2
chisq2<-chisq.test(tb2);chisq2
tb3<-table(chis$X2FG.M.O,chis$qualified);tb3
chisq3<-chisq.test(tb3);chisq3
tb4<-table(chis$X3FG.M.O,chis$qualified);tb4
chisq4<-chisq.test(tb4);chisq4
tb5<-table(chis$FT.M.O,chis$qualified);tb5
chisq5<-chisq.test(tb5);chisq5
tb6<-table(chis$O.TOTAL,chis$qualified);tb6
chisq6<-chisq.test(tb6);chisq6
tb7<-table(chis$D.TOTAL,chis$qualified);tb7
chisq7<-chisq.test(tb7);chisq7
tb8<-table(chis$T.TOTAL,chis$qualified);tb8
chisq8<-chisq.test(tb8);chisq8
tb9<-table(chis$As.TOTAL,chis$qualified);tb9
chisq9<-chisq.test(tb9);chisq9
tb10<-table(chis$St..TOTAL,chis$qualified);tb10
chisq10<-chisq.test(tb10);chisq10
tb11<-table(chis$To.TOTAL,chis$qualified);tb11
chisq11<-chisq.test(tb11);chisq11
tb12<-table(chis$Fv.TOTAL,chis$qualified);tb12
chisq12<-chisq.test(tb12);chisq12
tb13<-table(chis$Ag.TOTAL,chis$qualified);tb13
chisq13<-chisq.test(tb13);chisq13
tb14<-table(chis$Cm.TOTAL,chis$qualified);tb14
chisq14<-chisq.test(tb14);chisq14
tb15<-table(chis$Rv.TOTAL,chis$qualified);tb15
chisq15<-chisq.test(tb15);chisq15
tb16<-table(chis$PIR.TOTAL,chis$qualified);tb16
chisq16<-chisq.test(tb16);chisq16
df<-
data.frame(chisq2$p.value,chisq3$p.value,chisq4$p.value,chisq5$p.value,chisq6$p.value,chisq7$p.v
alue,chisq8$p.value,chisq9$p.value,chisq10$p.value,chisq11$p.value,chisq12$p.value,chisq13$p.va
lue,chisq14$p.value,chisq15$p.value,chisq16$p.value);df

names(chis)
install.packages("DescTools")
library(DescTools)
SomersDelta(chis$ranks,chis$qualified,direction  = "column",conf.level = 0.95)

#normality test for independent variables
shapiro.test(mydata2$Pts.Total..M.O)
shapiro.test(mydata2$dif_Pts.Total..M.O)
shapiro.test(mydata2$X2FG.M.O)
shapiro.test(mydata2$X3FG.M.O)
shapiro.test(mydata2$FT.M.O)
shapiro.test(mydata2$O.TOTAL)
shapiro.test(mydata2$D.TOTAL)
```

```
shapiro.test(mydata2$T.TOTAL)
shapiro.test(mydata2$As.TOTAL)
shapiro.test(mydata2$St..TOTAL)
shapiro.test(mydata2$To.TOTAL)
shapiro.test(mydata2$Fv.TOTAL)
shapiro.test(mydata2$Ag.TOTAL)
shapiro.test(mydata2$Cm.TOTAL)
shapiro.test(mydata2$Rv.TOTAL)
shapiro.test(mydata2$PIR.TOTAL)

library(tidyverse)
library(MASS)
set.seed(101)
sample_n(mydata2[,3:19], 10)
training_sample <- sample(c(TRUE, FALSE), nrow(mydata2), replace = T, prob = c(0.6,0.4))
train <- mydata2[training_sample,3:19 ];train
test <- mydata2[!training_sample,3:19 ];test
lda.mydata2<- lda(qualified ~
Pts.Total..M.O+dif_Pts.Total..M.O+X2FG.M.O+X3FG.M.O+FT.M.O+T.TOTAL+St..TOTAL+To.
TOTAL+Fv.TOTAL+Ag.TOTAL+Cm.TOTAL+Rv.TOTAL, train)
lda.mydata2 #show results
plot(lda.mydata2, col = as.integer(train$qualified))
plot(lda.mydata2, dimen = 1, type = "b")
lda.train <- predict(lda.mydata2)
train$lda <- lda.train$class
table(train$lda,train$qualified)
lda.test <- predict(lda.mydata2,test)
test$lda <- lda.test$class
table(test$lda,test$qualified)

#logit gia kathe metabliti gia na dw an einai statistika shmantiki
names(mydata2)
fit1<-glm(qualified~Pts.Total..M.O,family=binomial,data=mydata2);summary(fit1)
fit2<-glm(qualified~dif_Pts.Total..M.O,family=binomial,data=mydata2);summary(fit2)
fit3<-glm(qualified~X2FG.M.O,family=binomial,data=mydata2);summary(fit3)
fit4<-glm(qualified~X3FG.M.O,family=binomial,data=mydata2);summary(fit4)
fit5<-glm(qualified~FT.M.O,family=binomial,data=mydata2);summary(fit5)
fit6<-glm(qualified~O.TOTAL,family=binomial,data=mydata2);summary(fit6)
fit7<-glm(qualified~D.TOTAL,family=binomial,data=mydata2);summary(fit7)
fit8<-glm(qualified~T.TOTAL,family=binomial,data=mydata2);summary(fit8)
fit9<-glm(qualified~As.TOTAL,family=binomial,data=mydata2);summary(fit9)
fit10<-glm(qualified~St..TOTAL,family=binomial,data=mydata2);summary(fit10)
fit11<-glm(qualified~To.TOTAL,family=binomial,data=mydata2);summary(fit11)
fit12<-glm(qualified~Fv.TOTAL,family=binomial,data=mydata2);summary(fit12)
fit13<-glm(qualified~Ag.TOTAL,family=binomial,data=mydata2);summary(fit13)
fit14<-glm(qualified~Cm.TOTAL,family=binomial,data=mydata2);summary(fit14)
fit15<-glm(qualified~Rv.TOTAL,family=binomial,data=mydata2);summary(fit15)
fit16<-glm(qualified~PIR.TOTAL,family=binomial,data=mydata2);summary(fit16)
fit17<-glm(qualified~T.TOTAL+Rv.TOTAL,family=binomial,data=mydata2)
summary(fit17)

#feature selection gia 2015-2016

1. Random forest method

library(randomForest)
install.packages("Metrics")
library(Metrics)
fit_rf = randomForest(qualified~.,data=mydata2[,-1])
# Create an importance based on mean decreasing gini
```

```
importance(fit_rf)
varImpPlot(fit_rf)

2.Stepwise Method

base.mod <- glm(qualified ~ 1 ,family=binomial, data= mydata2[,-1])  # base intercept only model
all.mod <- glm(qualified ~ . ,family=binomial, data= mydata2[,-1]) # full model with all predictors
stepMod <- step(base.mod, scope = list(lower = base.mod, upper = all.mod), direction = "both",
trace = 0, steps = 1000)  # perform step-wise algorithm
shortlistedVars <- names(unlist(stepMod[[1]])) # get the shortlisted variable.
shortlistedVars <- shortlistedVars[!shortlistedVars %in% "(Intercept)"]  # remove intercept
print(shortlistedVars)

3.Boruta Method

install.packages("Boruta")
library(Boruta)
# Decide if a variable is important or not using Boruta
boruta_output <- Boruta(qualified~.,data=na.omit(mydata2[,-1]),doTrace=2) # perform Boruta
search
boruta_signif <- names(boruta_output$finalDecision[boruta_output$finalDecision %in%
c("Confirmed", "Tentative")])  # collect Confirmed and Tentative variables
print(boruta_signif)  # significant variables
plot(boruta_output, cex.axis=.7, las=2, xlab="", main="Variable Importance")

4. Anova

anova<-aov(qualified~.,data=mydata2[,-1])
summary(anova)

#GLM gia 2015-2016

names(mydata2)

fit.logit<-glm(qualified~X2FG.M.O+T.TOTAL,family=binomial,data=mydata2)
summary(fit.logit)

fit2.logit<-glm(qualified~X2FG.M.O*T.TOTAL,family=binomial,data=mydata2)
summary(fit2.logit)

fit.probit<-glm(qualified~X2FG.M.O+T.TOTAL,family=binomial(link="probit"),data=mydata2)
summary(fit.probit)
fit.cauchit<-glm(qualified~X2FG.M.O+T.TOTAL,family=binomial(link="cauchit"),data=mydata2)
summary(fit.cauchit)
fit.cloglog<-glm(qualified~X2FG.M.O+T.TOTAL,family=binomial(link="cloglog"),data=mydata2)
summary(fit.cloglog)
library(BaylorEdPsych)
PseudoR2(fit.logit)
PseudoR2(fit.probit)
PseudoR2(fit.cauchit)
PseudoR2(fit.cloglog)
X2.logit=sum(residuals(fit.logit,type="pearson")^2)
X2.probit=sum(residuals(fit.probit,type="pearson")^2)
X2.cauchit=sum(residuals(fit.cauchit,type="pearson")^2)
X2.cloglog=sum(residuals(fit.cloglog,type="pearson")^2)
X2.all=c(X2.logit,X2.probit,X2.cauchit,X2.cloglog)
X2.table=rbind(X2.all,1-pchisq(X2.all,df=fit.logit$df.residual))
colnames(X2.table)=c("logit","probit","cauchit","cloglog")
rownames(X2.table)=c("X2","p")
round(X2.table,4)
```

```
D.all=cbind(fit.logit$deviance,fit.probit$deviance,fit.cauchit$deviance,fit.cloglog$deviance)
round(D.all,3)

install.packages("ResourceSelection")
library(ResourceSelection)
phat=predict(fit.logit,data.frame(X2FG.M.O+T.TOTAL),type="response")
hoslem.test(qualified,phat)

pred.fit=predict(fit.probit,type="response")
library(pROC)
plot.roc(mydata2$qualified,pred.fit)
auc(mydata2$qualified,pred.fit)
coords(roc(mydata2$qualified,pred.fit),"best",best.method="youden")

#accuracy 2015-2016
attach(mydata2)
qualified.fit=1*(pred.fit>0.66)
classification.matrix=matrix(nrow=2,byrow=T,
+c(sum(qualified*qualified.fit),sum(qualified*(1-qualified.fit)),
+sum((1-qualified)*qualified.fit),sum((1-qualified)*(1-qualified.fit))))
rownames(classification.matrix)=c("y=1","y=0")
colnames(classification.matrix)=c("y.hat=1","y.hat=0")
classification.matrix
sum(qualified*qualified.fit)/sum(qualified) #sensitivity
sum((1-qualified)*(1-qualified.fit))/sum(1-qualified) #specificity
mean(qualified==qualified.fit)

#outliers

outliers<-boxplot.stats(T.TOTAL)$out
boxplot(T.TOTAL,main="total rebound",boxwex=0.1,horizontal=T)
mtext(paste("Outliers:",paste(outliers,collapse=",")),cex=0.6)

rebound_new<-T.TOTAL[T.TOTAL>275]
outliers<-boxplot.stats(rebound_new)$out
boxplot(total_rebound_new,main="total rebound",boxwex=0.1,horizontal=T)
mtext(paste("Outliers:",paste(outliers,collapse=",")),cex=0.6)

outliers<-boxplot.stats(X2FG.M.O)$out;outliers
boxplot(X2FG.M.O,main="X2FG",boxwex=0.1,horizontal=T)
mtext(paste("Outliers:",paste(outliers,collapse=",")),cex=0.6)

#---------- Chapter 5 ----------

##2015-2016
a<-read.xlsx("C:/Users/Dionysis/Desktop/thesis.xlsx",1,header=TRUE);a
b<-a[,-1];b

#group seperation

group_a<-b[c(1:60),];group_a
group_b<-b[c(62:121),];group_b
group_c<-b[c(123:182),];group_c
group_d<-b[c(184:243),];group_d

#gia oli tin xronia

df1<-rbind(group_a,group_b,group_c,group_d);df1
names(df1)
df<-df1[,-27] #vgazw to win
```

```
attach(df)
str(df)

library(DataExplorer)
plot_correlation(df[,-1])

#descriptive ksexwrista
library(purrr)
df[,-1]%>%split(df$HOME.AWAY)%>%map(summary)

#scatterplot
plot(Pts.Total.,PIR)
abline(lm(PIR~Pts.Total.),col="red")

#boxplots
par(mfrow=c(3,3))
b1<-boxplot(Pts.Total.~HOME.AWAY,data=df,
   main="HOME.AWAY vs Total Points",xlab="HOME.AWAY",ylab="Total Points",col="gold")
b2<-boxplot(PIR~HOME.AWAY,data=df,
   main="HOME.AWAY vs PIR",xlab="HOME.AWAY",ylab="PIR",col="gold")
b3<-boxplot(X2FG~HOME.AWAY,data=df,
   main="HOME.AWAY vs Total FGX2",xlab="HOME.AWAY",ylab="X2FG",col="gold")
b4<-boxplot(X3FG~HOME.AWAY,data=df,
   main="HOME.AWAY vs FGX3",xlab="HOME.AWAY",ylab="X3FG",col="gold")
b5<-boxplot(St.~HOME.AWAY,data=df,
   main="HOME.AWAY vs Steals",xlab="HOME.AWAY",ylab="Steels",col="gold")
b6<-boxplot(T~HOME.AWAY,data=df,
   main="HOME.AWAY vs T.TOTAL",xlab="HOME.AWAY",ylab="TOTAL
REBOUND",col="gold")
b7<-boxplot(As~HOME.AWAY,data=df,
   main="HOME.AWAY vs Assist",xlab="HOME.AWAY",ylab="Assist",col="gold")
b8<-boxplot(To~HOME.AWAY,data=df,
   main="HOME.AWAY vs To.Total",xlab="HOME.AWAY",ylab="Turnover",col="gold")
b9<-boxplot(Rv~HOME.AWAY,data=df,
   main="HOME.AWAY vs Rv.Total",xlab="HOME.AWAY",ylab="Rv",col="gold")

#histogram
par(mfrow=c(2,1))
hist(Pts.Total.[HOME.AWAY==1],col="gold")
hist(Pts.Total.[HOME.AWAY==0],col="gold")

#violin plots
install.packages("vioplot")
library(vioplot)
vioplot(Pts.Total.~HOME.AWAY,col="gold")

###new_visual
install.packages("ggplot2")
library(ggplot2)
df$HOME.AWAY<-as.factor(df$HOME.AWAY)
qplot(Pts.Total.,GROUP.A,data=df,color=HOME.AWAY)
qplot(T,GROUP.A,data=df,color=HOME.AWAY)

#normality test for independent variables
shapiro.test(df$Pts.Q1.)
shapiro.test(df$Pts.Q2.)
shapiro.test(df$Pts.Q3.)
shapiro.test(df$Pts.Q4.)
shapiro.test(df$Pts.ET.)
shapiro.test(df$Pts.Total.)
```

```
shapiro.test(df$dif_Q1)
shapiro.test(df$dif_Q2)
shapiro.test(df$dif_Q3)
shapiro.test(df$dif_Q4)
shapiro.test(df$dif_Pts.Total.)
shapiro.test(df$X2FG)
shapiro.test(df$X3FG)
shapiro.test(df$FT)
shapiro.test(df$O)
shapiro.test(df$D)
shapiro.test(df$T)
shapiro.test(df$As)
shapiro.test(df$St.)
shapiro.test(df$To)
shapiro.test(df$Fv)
shapiro.test(df$Ag)
shapiro.test(df$Cm)
shapiro.test(df$Rv)
shapiro.test(df$PIR)

#feature selection

library(Boruta)
boruta_output <- Boruta(HOME.AWAY~.,data=na.omit(df),doTrace=2)
boruta_signif <- names(boruta_output$finalDecision[boruta_output$finalDecision %in%
c("Confirmed", "Tentative")])  # collect Confirmed and Tentative variables
print(boruta_signif)  # significant variables
plot(boruta_output, cex.axis=.7, las=2, xlab="", main="Variable Importance")

#t-test

t.test(Pts.Q1.~HOME.AWAY,var.equal=TRUE)
t.test(Pts.Q3.~HOME.AWAY,var.equal=TRUE)
t.test(Pts.Total.~HOME.AWAY,var.equal=TRUE)
t.test(dif_Q1~HOME.AWAY,var.equal=TRUE)
t.test(dif_Q2~HOME.AWAY,var.equal=TRUE)
t.test(dif_Q3~HOME.AWAY,var.equal=TRUE)
t.test(dif_Q4~HOME.AWAY,var.equal=TRUE)
t.test(dif_Pts.Total.~HOME.AWAY,var.equal=TRUE)
t.test(X2FG~HOME.AWAY,var.equal=TRUE)
t.test(X3FG~HOME.AWAY,var.equal=TRUE)
t.test(PIR~HOME.AWAY,var.equal=TRUE)
var.test(PIR,HOME.AWAY)
t.test(PIR,HOME.AWAY,var.equal=FALSE)

#Mann-Whitney U
wilcox.test(Pts.Q2.~HOME.AWAY)
wilcox.test(Pts.Q4.~HOME.AWAY)
wilcox.test(Pts.ET.~HOME.AWAY)
wilcox.test(FT~HOME.AWAY)
wilcox.test(O~HOME.AWAY)
wilcox.test(D~HOME.AWAY)
wilcox.test(T~HOME.AWAY)
wilcox.test(As~HOME.AWAY)
wilcox.test(St.~HOME.AWAY)
wilcox.test(To~HOME.AWAY)
wilcox.test(Fv~HOME.AWAY)
wilcox.test(Ag~HOME.AWAY)
wilcox.test(Cm~HOME.AWAY)
wilcox.test(Rv~HOME.AWAY)
```

```
names(df)
#random forest

df$HOME.AWAY<-as.factor(df$HOME.AWAY)
######### Random forest method

library(randomForest)
install.packages("Metrics")
library(Metrics)
fit=randomForest(HOME.AWAY~., data=df[,-1],importance=TRUE, proximity=TRUE)
fit
varImpPlot(fit)
plot(fit)

df<-df[,-1]

# Split into Train and Validation sets
# Training Set : Validation Set = 70 : 30 (random)
set.seed(100)
train <- sample(nrow(df), 0.7*nrow(df), replace = FALSE)
TrainSet <- df[train,]
ValidSet <- df[-train,]
summary(TrainSet)
summary(ValidSet)

# Create a Random Forest model with default parameters
model1 <- randomForest(HOME.AWAY ~ ., data = TrainSet, importance = TRUE)
model1

# Fine tuning parameters of Random Forest model
model2 <- randomForest(HOME.AWAY ~ ., data = TrainSet, ntree = 500, mtry = 7, importance =
TRUE)
model2

# To check important variables
varImpPlot(model2)
plot(model2)

##########Boruta Method

install.packages("Boruta")
library(Boruta)
# Decide if a variable is important or not using Boruta
boruta_output <- Boruta(HOME.AWAY~.,data=na.omit(df),doTrace=2) # perform Boruta search
boruta_signif <- names(boruta_output$finalDecision[boruta_output$finalDecision %in%
c("Confirmed", "Tentative")])  # collect Confirmed and Tentative variables
print(boruta_signif)  # significant variables
plot(boruta_output, cex.axis=.7, las=2, xlab="", main="Variable Importance")

###2014-2015

c<-read.xlsx("C:/Users/Dionysis/Desktop/thesis.xlsx",1,header=TRUE);c
d<-c[,-1];d

#group seperation

group_a<-d[c(245:304),];group_a
group_b<-d[c(306:365),];group_b
group_c<-d[c(367:426),];group_c
```

```
group_d<-d[c(428:487),];group_d

#gia oli tin xronia

df2<-rbind(group_a,group_b,group_c,group_d);df2
names(df2)
df<-df2[,-27] #vgazw to win
attach(df)
str(df)
dim(df)

library(DataExplorer)
plot_missing(df)
plot_correlation(df)

#descriptive ksexwrista
library(purrr)
df[,-1]%>%split(df$HOME.AWAY)%>%map(summary)

#scatterplot
plot(Pts.Total.,PIR)
abline(lm(PIR~Pts.Total.),col="red")

#boxplots
par(mfrow=c(3,3))
b1<-boxplot(Pts.Total.~HOME.AWAY,data=df,
   main="HOME.AWAY vs Total Points",xlab="HOME.AWAY",ylab="Total Points",col="gold")
b2<-boxplot(PIR~HOME.AWAY,data=df,
   main="HOME.AWAY vs PIR",xlab="HOME.AWAY",ylab="PIR",col="gold")
b3<-boxplot(X2FG~HOME.AWAY,data=df,
   main="HOME.AWAY vs Total FGX2",xlab="HOME.AWAY",ylab="X2FG",col="gold")
b4<-boxplot(X3FG~HOME.AWAY,data=df,
   main="HOME.AWAY vs FGX3",xlab="HOME.AWAY",ylab="X3FG",col="gold")
b5<-boxplot(St.~HOME.AWAY,data=df,
   main="HOME.AWAY vs Steals",xlab="HOME.AWAY",ylab="Steels",col="gold")
b6<-boxplot(T~HOME.AWAY,data=df,
   main="HOME.AWAY vs T.TOTAL",xlab="HOME.AWAY",ylab="TOTAL
REBOUND",col="gold")
b7<-boxplot(As~HOME.AWAY,data=df,
   main="HOME.AWAY vs Assist",xlab="HOME.AWAY",ylab="Assist",col="gold")
b8<-boxplot(To~HOME.AWAY,data=df,
   main="HOME.AWAY vs To.Total",xlab="HOME.AWAY",ylab="Turnover",col="gold")
b9<-boxplot(Rv~HOME.AWAY,data=df,
   main="HOME.AWAY vs Rv.Total",xlab="HOME.AWAY",ylab="Rv",col="gold")

#histogram
par(mfrow=c(2,1))
hist(Pts.Total.[HOME.AWAY==1],col="gold")
hist(Pts.Total.[HOME.AWAY==0],col="gold")

#violin plots
install.packages("vioplot")
library(vioplot)
vioplot(Pts.Total.~HOME.AWAY,col="gold")

###new_visual
install.packages("ggplot2")
library(ggplot2)
df$HOME.AWAY<-as.factor(df$HOME.AWAY)
qplot(Pts.Total.,GROUP.A,data=df,color=HOME.AWAY)
```

```
qplot(T,GROUP.A,data=df,color=HOME.AWAY)

#normality test for independent variables
shapiro.test(df$Pts.Q1.)
shapiro.test(df$Pts.Q2.)
shapiro.test(df$Pts.Q3.)
shapiro.test(df$Pts.Q4.)
shapiro.test(df$Pts.ET.)
shapiro.test(df$Pts.Total.)
shapiro.test(df$dif_Q1)
shapiro.test(df$dif_Q2)
shapiro.test(df$dif_Q3)
shapiro.test(df$dif_Q4)
shapiro.test(df$dif_Pts.Total.)
shapiro.test(df$X2FG)
shapiro.test(df$X3FG)
shapiro.test(df$FT)
shapiro.test(df$O)
shapiro.test(df$D)
shapiro.test(df$T)
shapiro.test(df$As)
shapiro.test(df$St.)
shapiro.test(df$To)
shapiro.test(df$Fv)
shapiro.test(df$Ag)
shapiro.test(df$Cm)
shapiro.test(df$Rv)
shapiro.test(df$PIR)

#feature selection

library(Boruta)
boruta_output <- Boruta(HOME.AWAY~.,data=na.omit(df),doTrace=2)
boruta_signif <- names(boruta_output$finalDecision[boruta_output$finalDecision %in%
c("Confirmed", "Tentative")])  # collect Confirmed and Tentative variables
print(boruta_signif)  # significant variables
plot(boruta_output, cex.axis=.7, las=2, xlab="", main="Variable Importance")

#t-test

t.test(Pts.Q1.~HOME.AWAY,var.equal=TRUE)
t.test(Pts.Q2.~HOME.AWAY,var.equal=TRUE)
t.test(Pts.Q4.~HOME.AWAY,var.equal=TRUE)
t.test(dif_Q1~HOME.AWAY,var.equal=TRUE)
t.test(dif_Q2~HOME.AWAY,var.equal=TRUE)
t.test(dif_Pts.Total.~HOME.AWAY,var.equal=TRUE)
t.test(X2FG~HOME.AWAY,var.equal=TRUE)
t.test(X3FG~HOME.AWAY,var.equal=TRUE)
t.test(D~HOME.AWAY,var.equal=TRUE)
var.test(D,HOME.AWAY)
t.test(D,HOME.AWAY,var.equal=FALSE)

#Mann-Whitney U
wilcox.test(Pts.Q3.~HOME.AWAY)
wilcox.test(Pts.ET.~HOME.AWAY)
wilcox.test(Pts.Total.~HOME.AWAY)
wilcox.test(dif_Q3~HOME.AWAY)
wilcox.test(dif_Q4~HOME.AWAY)
wilcox.test(FT~HOME.AWAY)
wilcox.test(O~HOME.AWAY)
```

```
wilcox.test(T~HOME.AWAY)
wilcox.test(As~HOME.AWAY)
wilcox.test(St.~HOME.AWAY)
wilcox.test(To~HOME.AWAY)
wilcox.test(Fv~HOME.AWAY)
wilcox.test(Ag~HOME.AWAY)
wilcox.test(Cm~HOME.AWAY)
wilcox.test(Rv~HOME.AWAY)
wilcox.test(PIR~HOME.AWAY)

#random forest

df$HOME.AWAY<-as.factor(df$HOME.AWAY)
######### Random forest method

library(randomForest)
install.packages("Metrics")
library(Metrics)
fit=randomForest(HOME.AWAY~., data=df[,-1],importance=TRUE, proximity=TRUE)
fit
varImpPlot(fit)
plot(fit)

df<-df[,-1]
df
# Split into Train and Validation sets
# Training Set : Validation Set = 70 : 30 (random)
set.seed(100)
train <- sample(nrow(df), 0.7*nrow(df), replace = FALSE)
TrainSet <- df[train,]
ValidSet <- df[-train,]
summary(TrainSet)
summary(ValidSet)

# Create a Random Forest model with default parameters
model1 <- randomForest(HOME.AWAY ~ ., data = TrainSet, importance = TRUE)
model1

# Fine tuning parameters of Random Forest model
model2 <- randomForest(HOME.AWAY ~ ., data = TrainSet, ntree = 500, mtry = 6, importance =
TRUE)
model2
plot(model2)

# To check important variables
varImpPlot(model2)

##########Boruta Method

install.packages("Boruta")
library(Boruta)
# Decide if a variable is important or not using Boruta
boruta_output <- Boruta(HOME.AWAY~.,data=na.omit(df),doTrace=2) # perform Boruta search
boruta_signif <- names(boruta_output$finalDecision[boruta_output$finalDecision %in%
c("Confirmed", "Tentative")])  # collect Confirmed and Tentative variables
print(boruta_signif)  # significant variables
plot(boruta_output, cex.axis=.7, las=2, xlab="", main="Variable Importance")
```

## II2. GROUPS

## YEAR 2014-2015

### GROUP A

| | |
|---|---|
| 1 | Real Madrid |
| 2 | Anatolou Efes Istanbul |
| 3 | Zalgiris Kaunas |
| 4 | Nizhny Novgorov |
| 5 | Unics Kazan |
| 6 | Dinamo Banco Di Sardegna Sassari |

### GROUP B

| | |
|---|---|
| 1 | CSKA Moscow |
| 2 | Maccabi Fox Tel Aviv |
| 3 | Unicaja Malaga |
| 4 | Alba Berlin |
| 5 | Cedevita Zagreb |
| 6 | Limoges CSP |

### GROUP C

| | |
|---|---|
| 1 | FC Barcelona Lass |
| 2 | Fenerbahce Istanbul |

| | |
|---|---|
| 3 | Panathinaikos Athens |
| 4 | EA7 Emporio Armani Milan |
| 5 | FC Bayern Munich |
| 6 | PGE Turow Zgorzelec |

**GROUP D**

| | |
|---|---|
| 1 | Olympiacos Piraeus |
| 2 | Crvena Zvezda Telekom Belgrade |
| 3 | Laboral Kutxa Vitoria Gasteiz |
| 4 | Galatasaray Liv Hospital Istanbul |
| 5 | Neptunas Klaipeda |
| 6 | Valencia Basket |

**YEAR 2015-2016**

**GROUP A**

| | |
|---|---|
| 1 | Fenerbache Istanbul |
| 2 | Khimki Moscow Region |
| 3 | Crvena Zvezda Telekom Belgrade |
| 4 | Real Madrid |
| 5 | FC Byern Munich |
| 6 | Strasbourg |

**GROUP B**

| | |
|---|---|
| 1 | Olympiacos Piraeus |
| 2 | Anatolou Efes Istanbul |
| 3 | Laboral Kutxa Vitoria Gasteiz |
| 4 | Cedevita Zagreb |
| 5 | Limoges CSP |
| 6 | EA7 Emporio Armani Milan |

**GROUP C**

| | |
|---|---|
| 1 | Lokomotiv Kuban Krasnodar |
| 2 | FC Barcelona Lassa |
| 3 | Panathinaikos Athens |
| 4 | Zalgiris Kaunas |
| 5 | Pinar Karsiyaka Izmir |
| 6 | Stelmet Zielona Gora |

**GROUP D**

| | |
|---|---|
| 1 | CSKA Moscow |
| 2 | Unicaja Malaga |
| 3 | Brose Baskets Bamberg |
| 4 | Darussafaka Dogus Istanbul |
| 5 | Maccabi Fox Tel Aviv |
| 6 | Dinamo Banco Di Sardegna Sassari |

## Π3. ΤΑ 240 ΠΑΙΧΝΙΔΙΑ ΤΗΣ ΔΙΟΡΓΑΝΩΣΗΣ ΓΙΑ ΤΙΣ ΠΕΡΙΟΔΟΥΣ 2014-2015 ΚΑΙ 2015-2016

| 2015-16 | GROUP A |
|---|---|
| 1 | Crvena Zvezda Telekom Belgrade |
| | Strasbourg |
| 2 | Fenerbahce Istanbul |
| | FC Bayern Munich |
| 3 | Khimki Moscow Region |
| | Real Madrid |
| 4 | FC Bayern Munich |
| | Khimki Moscow Region |
| 5 | Real Madrid |
| | Crvena Zvezda Telekom Belgrade |
| 6 | Strasbourg |
| | Fenerbahce Istanbul |
| 7 | Khimki Moscow Region |
| | Crvena Zvezda Telekom Belgrade |
| 8 | Fenerbahce Istanbul |
| | Real Madrid |
| 9 | FC Bayern Munich |
| | Strasbourg |
| 10 | Khimki Moscow Region |
| | Strasbourg |
| 11 | Real Madrid |
| | FC Bayern Munich |
| 12 | Crvena Zvezda Telekom Belgrade |
| | Fenerbahce Istanbul |
| 13 | FC Bayern Munich |
| | Crvena Zvezda Telekom Belgrade |
| 14 | Strasbourg |
| | Real Madrid |
| 15 | Fenerbahce Istanbul |
| | Khimki Moscow Region |
| 16 | Real Madrid |
| | Khimki Moscow Region |
| 17 | FC Bayern Munich |
| | Fenerbahce Istanbul |
| 18 | Strasbourg |
| | Crvena Zvezda Telekom Belgrade |
| 19 | Khimki Moscow Region |
| | FC Bayern Munich |
| 20 | Fenerbahce Istanbul |
| | Strasbourg |
| 21 | Crvena Zvezda Telekom Belgrade |

| | |
|---|---|
| | Real Madrid |
| 22 | Real Madrid |
| | Fenerbahce Istanbul |
| 23 | Crvena Zvezda Telekom Belgrade |
| | Khimki Moscow Region |
| 24 | Strasbourg |
| | FC Bayern Munich |
| 25 | FC Bayern Munich |
| | Real Madrid |
| 26 | Strasbourg |
| | Khimki Moscow Region |
| 27 | Fenerbahce Istanbul |
| | Crvena Zvezda Telekom Belgrade |
| 28 | Khimki Moscow Region |
| | Fenerbahce Istanbul |
| 29 | Real Madrid |
| | Strasbourg |
| 30 | Crvena Zvezda Telekom Belgrade |
| | FC Bayern Munich |
| | GROUP B |
| 31 | Olympiacos Piraeus |
| | Cedevita Zagreb |
| 32 | Limoges CSP |
| | Anadolu Efes Istanbul |
| 33 | EA7 Emporio Armani Milan |
| | Laboral Kutxa Vitoria Gasteiz |
| 34 | Anadolu Efes Istanbul |
| | EA7 Emporio Armani Milan |
| 35 | Cedevita Zagreb |
| | Limoges CSP |
| 36 | Laboral Kutxa Vitoria Gasteiz |
| | Olympiacos Piraeus |
| 37 | Anadolu Efes Istanbul |
| | Cedevita Zagreb |
| 38 | Limoges CSP |
| | Laboral Kutxa Vitoria Gasteiz |
| 39 | EA7 Emporio Armani Milan |
| | Olympiacos Piraeus |
| 40 | Laboral Kutxa Vitoria Gasteiz |
| | Anadolu Efes Istanbul |
| 41 | EA7 Emporio Armani Milan |
| | Cedevita Zagreb |
| 42 | Olympiacos Piraeus |
| | Limoges CSP |
| 43 | Anadolu Efes Istanbul |
| | Olympiacos Piraeus |

| | |
|---|---|
| 44 | Limoges CSP |
| | EA7 Emporio Armani Milan |
| 45 | Cedevita Zagreb |
| | Laboral Kutxa Vitoria Gasteiz |
| 46 | Cedevita Zagreb |
| | Olympiacos Piraeus |
| 47 | Laboral Kutxa Vitoria Gasteiz |
| | EA7 Emporio Armani Milan |
| 48 | Anadolu Efes Istanbul |
| | Limoges CSP |
| 49 | Limoges CSP |
| | Cedevita Zagreb |
| 50 | EA7 Emporio Armani Milan |
| | Anadolu Efes Istanbul |
| 51 | Olympiacos Piraeus |
| | Laboral Kutxa Vitoria Gasteiz |
| 52 | Cedevita Zagreb |
| | Anadolu Efes Istanbul |
| 53 | Laboral Kutxa Vitoria Gasteiz |
| | Limoges CSP |
| 54 | Olympiacos Piraeus |
| | EA7 Emporio Armani Milan |
| 55 | Anadolu Efes Istanbul |
| | Laboral Kutxa Vitoria Gasteiz |
| 56 | Limoges CSP |
| | Olympiacos Piraeus |
| 57 | Cedevita Zagreb |
| | EA7 Emporio Armani Milan |
| 58 | Laboral Kutxa Vitoria Gasteiz |
| | Cedevita Zagreb |
| 59 | EA7 Emporio Armani Milan |
| | Limoges CSP |
| 60 | Olympiacos Piraeus |
| | Anadolu Efes Istanbul |
| | GROUP C |
| 61 | Pinar Karsiyaka Izmir |
| | FC Barcelona Lassa |
| 62 | Lokomotiv Kuban Krasnodar |
| | Panathinaikos Athens |
| 63 | Stelmet Zielona Gora |
| | Zalgiris Kaunas |
| 64 | Panathinaikos Athens |
| | Pinar Karsiyaka Izmir |
| 65 | Zalgiris Kaunas |
| | Lokomotiv Kuban Krasnodar |
| 66 | FC Barcelona Lassa |

| | |
|---|---|
| | Stelmet Zielona Gora |
| 67 | Pinar Karsiyaka Izmir |
| | Zalgiris Kaunas |
| 68 | FC Barcelona Lassa |
| | Panathinaikos Athens |
| 69 | Stelmet Zielona Gora |
| | Lokomotiv Kuban Krasnodar |
| 70 | Lokomotiv Kuban Krasnodar |
| | Pinar Karsiyaka Izmir |
| 71 | Zalgiris Kaunas |
| | FC Barcelona Lassa |
| 72 | Stelmet Zielona Gora |
| | Panathinaikos Athens |
| 73 | FC Barcelona Lassa |
| | Lokomotiv Kuban Krasnodar |
| 74 | Pinar Karsiyaka Izmir |
| | Stelmet Zielona Gora |
| 75 | Panathinaikos Athens |
| | Zalgiris Kaunas |
| 76 | Zalgiris Kaunas |
| | Stelmet Zielona Gora |
| 77 | FC Barcelona Lassa |
| | Pinar Karsiyaka Izmir |
| 78 | Panathinaikos Athens |
| | Lokomotiv Kuban Krasnodar |
| 79 | Stelmet Zielona Gora |
| | FC Barcelona Lassa |
| 80 | Lokomotiv Kuban Krasnodar |
| | Zalgiris Kaunas |
| 81 | Pinar Karsiyaka Izmir |
| | Panathinaikos Athens |
| 82 | Zalgiris Kaunas |
| | Pinar Karsiyaka Izmir |
| 83 | Lokomotiv Kuban Krasnodar |
| | Stelmet Zielona Gora |
| 84 | Panathinaikos Athens |
| | FC Barcelona Lassa |
| 85 | Pinar Karsiyaka Izmir |
| | Lokomotiv Kuban Krasnodar |
| 86 | FC Barcelona Lassa |
| | Zalgiris Kaunas |
| 87 | Panathinaikos Athens |
| | Stelmet Zielona Gora |
| 88 | Zalgiris Kaunas |
| | Panathinaikos Athens |
| 89 | Lokomotiv Kuban Krasnodar |

| | |
|---|---|
| | FC Barcelona Lassa |
| 90 | Stelmet Zielona Gora |
| | Pinar Karsiyaka Izmir |
| | GROUP D |
| 91 | CSKA Moscow |
| | Maccabi Fox Tel Aviv |
| 92 | Darussafaka Dogus Istanbul |
| | Dinamo Banco Di Sardegna Sassari |
| 93 | Unicaja Malaga |
| | Brose Baskets Bamberg |
| 94 | Maccabi Fox Tel Aviv |
| | Unicaja Malaga |
| 95 | Brose Baskets Bamberg |
| | Darussafaka Dogus Istanbul |
| 96 | Dinamo Banco Di Sardegna Sassari |
| | CSKA Moscow |
| 97 | CSKA Moscow |
| | Brose Baskets Bamberg |
| 98 | Maccabi Fox Tel Aviv |
| | Dinamo Banco Di Sardegna Sassari |
| 99 | Unicaja Malaga |
| | Darussafaka Dogus Istanbul |
| 100 | Brose Baskets Bamberg |
| | Maccabi Fox Tel Aviv |
| 101 | Darussafaka Dogus Istanbul |
| | CSKA Moscow |
| 102 | Unicaja Malaga |
| | Dinamo Banco Di Sardegna Sassari |
| 103 | Maccabi Fox Tel Aviv |
| | Darussafaka Dogus Istanbul |
| 104 | CSKA Moscow |
| | Unicaja Malaga |
| 105 | Dinamo Banco Di Sardegna Sassari |
| | Brose Baskets Bamberg |
| 106 | Maccabi Fox Tel Aviv |
| | CSKA Moscow |
| 107 | Brose Baskets Bamberg |
| | Unicaja Malaga |
| 108 | Dinamo Banco Di Sardegna Sassari |
| | Darussafaka Dogus Istanbul |
| 109 | Unicaja Malaga |
| | Maccabi Fox Tel Aviv |
| 110 | CSKA Moscow |
| | Dinamo Banco Di Sardegna Sassari |
| 111 | Darussafaka Dogus Istanbul |
| | Brose Baskets Bamberg |

| | |
|---|---|
| 112 | Dinamo Banco Di Sardegna Sassari |
| | Maccabi Fox Tel Aviv |
| 113 | Brose Baskets Bamberg |
| | CSKA Moscow |
| 114 | Darussafaka Dogus Istanbul |
| | Unicaja Malaga |
| 115 | Maccabi Fox Tel Aviv |
| | Brose Baskets Bamberg |
| 116 | Dinamo Banco Di Sardegna Sassari |
| | Unicaja Malaga |
| 117 | CSKA Moscow |
| | Darussafaka Dogus Istanbul |
| 118 | Darussafaka Dogus Istanbul |
| | Maccabi Fox Tel Aviv |
| 119 | Brose Baskets Bamberg |
| | Dinamo Banco Di Sardegna Sassari |
| 120 | Unicaja Malaga |
| | CSKA Moscow |
| 2014-15 | GROUP A |
| 121 | Anadolu Efes Istanbul |
| | Unics Kazan |
| 122 | Nizhny Novgorod |
| | Dinamo Banco Di Sardegna Sassari |
| 123 | Real Madrid |
| | Zalgiris Kaunas |
| 124 | Unics Kazan |
| | Real Madrid |
| 125 | Zalgiris Kaunas |
| | Nizhny Novgorod |
| 126 | Dinamo Banco Di Sardegna Sassari |
| | Anadolu Efes Istanbul |
| 127 | Unics Kazan |
| | Dinamo Banco Di Sardegna Sassari |
| 128 | Anadolu Efes Istanbul |
| | Zalgiris Kaunas |
| 129 | Real Madrid |
| | Nizhny Novgorod |
| 130 | Real Madrid |
| | Dinamo Banco Di Sardegna Sassari |
| 131 | Nizhny Novgorod |
| | Anadolu Efes Istanbul |
| 132 | Zalgiris Kaunas |
| | Unics Kazan |
| 133 | Unics Kazan |
| | Nizhny Novgorod |
| 134 | Anadolu Efes Istanbul |

| | |
|---|---|
| | Real Madrid |
| 135 | Dinamo Banco Di Sardegna Sassari |
| | Zalgiris Kaunas |
| 136 | Unics Kazan |
| | Anadolu Efes Istanbul |
| 137 | Zalgiris Kaunas |
| | Real Madrid |
| 138 | Dinamo Banco Di Sardegna Sassari |
| | Nizhny Novgorod |
| 139 | Real Madrid |
| | Unics Kazan |
| 140 | Nizhny Novgorod |
| | Zalgiris Kaunas |
| 141 | Anadolu Efes Istanbul |
| | Dinamo Banco Di Sardegna Sassari |
| 142 | Nizhny Novgorod |
| | Real Madrid |
| 143 | Dinamo Banco Di Sardegna Sassari |
| | Unics Kazan |
| 144 | Zalgiris Kaunas |
| | Anadolu Efes Istanbul |
| 145 | Unics Kazan |
| | Zalgiris Kaunas |
| 146 | Anadolu Efes Istanbul |
| | Nizhny Novgorod |
| 147 | Dinamo Banco Di Sardegna Sassari |
| | Real Madrid |
| 148 | Real Madrid |
| | Anadolu Efes Istanbul |
| 149 | Nizhny Novgorod |
| | Unics Kazan |
| 150 | Zalgiris Kaunas |
| | Dinamo Banco Di Sardegna Sassari |
| | GROUP B |
| 151 | Cedevita Zagreb |
| | Unicaja Malaga |
| 152 | Maccabi Fox Tel Aviv |
| | Limoges CSP |
| 153 | Alba Berlin |
| | CSKA Moscow |
| 154 | CSKA Moscow |
| | Maccabi Fox Tel Aviv |
| 155 | Limoges CSP |
| | Cedevita Zagreb |
| 156 | Unicaja Malaga |
| | Alba Berlin |

| | |
|---|---|
| 157 | Alba Berlin |
| | Maccabi Fox Tel Aviv |
| 158 | Cedevita Zagreb |
| | CSKA Moscow |
| 159 | Unicaja Malaga |
| | Limoges CSP |
| 160 | Maccabi Fox Tel Aviv |
| | Cedevita Zagreb |
| 161 | CSKA Moscow |
| | Unicaja Malaga |
| 162 | Alba Berlin |
| | Limoges CSP |
| 163 | Unicaja Malaga |
| | Maccabi Fox Tel Aviv |
| 164 | Cedevita Zagreb |
| | Alba Berlin |
| 165 | Limoges CSP |
| | CSKA Moscow |
| 166 | CSKA Moscow |
| | Alba Berlin |
| 167 | Limoges CSP |
| | Maccabi Fox Tel Aviv |
| 168 | Unicaja Malaga |
| | Cedevita Zagreb |
| 169 | Maccabi Fox Tel Aviv |
| | CSKA Moscow |
| 170 | Alba Berlin |
| | Unicaja Malaga |
| 171 | Cedevita Zagreb |
| | Limoges CSP |
| 172 | Limoges CSP |
| | Unicaja Malaga |
| 173 | Maccabi Fox Tel Aviv |
| | Alba Berlin |
| 174 | CSKA Moscow |
| | Cedevita Zagreb |
| 175 | Cedevita Zagreb |
| | Maccabi Fox Tel Aviv |
| 176 | Unicaja Malaga |
| | CSKA Moscow |
| 177 | Limoges CSP |
| | Alba Berlin |
| 178 | CSKA Moscow |
| | Limoges CSP |
| 179 | Maccabi Fox Tel Aviv |
| | Unicaja Malaga |

| | |
|---|---|
| 180 | Alba Berlin |
| | Cedevita Zagreb |
| | GROUP C |
| 181 | Panathinaikos Athens |
| | PGE Turow Zgorzelec |
| 182 | Fenerbahce Istanbul |
| | EA7 Emporio Armani Milan |
| 183 | FC Barcelona Lassa |
| | FC Bayern Munich |
| 184 | FC Bayern Munich |
| | Panathinaikos Athens |
| 185 | EA7 Emporio Armani Milan |
| | FC Barcelona Lassa |
| 186 | PGE Turow Zgorzelec |
| | Fenerbahce Istanbul |
| 187 | Panathinaikos Athens |
| | Fenerbahce Istanbul |
| 188 | FC Bayern Munich |
| | EA7 Emporio Armani Milan |
| 189 | FC Barcelona Lassa |
| | PGE Turow Zgorzelec |
| 190 | PGE Turow Zgorzelec |
| | FC Bayern Munich |
| 191 | Fenerbahce Istanbul |
| | FC Barcelona Lassa |
| 192 | Panathinaikos Athens |
| | EA7 Emporio Armani Milan |
| 193 | FC Bayern Munich |
| | Fenerbahce Istanbul |
| 194 | EA7 Emporio Armani Milan |
| | PGE Turow Zgorzelec |
| 195 | FC Barcelona Lassa |
| | Panathinaikos Athens |
| 196 | FC Bayern Munich |
| | FC Barcelona Lassa |
| 197 | PGE Turow Zgorzelec |
| | Panathinaikos Athens |
| 198 | EA7 Emporio Armani Milan |
| | Fenerbahce Istanbul |
| 199 | Fenerbahce Istanbul |
| | PGE Turow Zgorzelec |
| 200 | Panathinaikos Athens |
| | FC Bayern Munich |
| 201 | FC Barcelona Lassa |
| | EA7 Emporio Armani Milan |
| 202 | EA7 Emporio Armani Milan |

| | |
|---|---|
| | FC Bayern Munich |
| 203 | Fenerbahce Istanbul |
| | Panathinaikos Athens |
| 204 | PGE Turow Zgorzelec |
| | FC Barcelona Lassa |
| 205 | FC Bayern Munich |
| | PGE Turow Zgorzelec |
| 206 | FC Barcelona Lassa |
| | Fenerbahce Istanbul |
| 207 | EA7 Emporio Armani Milan |
| | Panathinaikos Athens |
| 208 | PGE Turow Zgorzelec |
| | EA7 Emporio Armani Milan |
| 209 | Panathinaikos Athens |
| | FC Barcelona Lassa |
| 210 | Fenerbahce Istanbul |
| | FC Bayern Munich |
| | GROUP D |
| 211 | Laboral Kutxa Vitoria Gasteiz |
| | Neptunas Klaipeda |
| 212 | Crvena Zvezda Telekom Belgrade |
| | Galatasaray Liv Hospital Istanbul |
| 213 | Valencia Basket |
| | Olympiacos Piraeus |
| 214 | Neptunas Klaipeda |
| | Crvena Zvezda Telekom Belgrade |
| 215 | Galatasaray Liv Hospital Istanbul |
| | Valencia Basket |
| 216 | Olympiacos Piraeus |
| | Laboral Kutxa Vitoria Gasteiz |
| 217 | Crvena Zvezda Telekom Belgrade |
| | Valencia Basket |
| 218 | Neptunas Klaipeda |
| | Olympiacos Piraeus |
| 219 | Laboral Kutxa Vitoria Gasteiz |
| | Galatasaray Liv Hospital Istanbul |
| 220 | Crvena Zvezda Telekom Belgrade |
| | Olympiacos Piraeus |
| 221 | Galatasaray Liv Hospital Istanbul |
| | Neptunas Klaipeda |
| 222 | Valencia Basket |
| | Laboral Kutxa Vitoria Gasteiz |
| 223 | Neptunas Klaipeda |
| | Valencia Basket |
| 224 | Laboral Kutxa Vitoria Gasteiz |
| | Crvena Zvezda Telekom Belgrade |

| | |
|---|---|
| 225 | Olympiacos Piraeus |
| | Galatasaray Liv Hospital Istanbul |
| 226 | Neptunas Klaipeda |
| | Laboral Kutxa Vitoria Gasteiz |
| 227 | Olympiacos Piraeus |
| | Valencia Basket |
| 228 | Galatasaray Liv Hospital Istanbul |
| | Crvena Zvezda Telekom Belgrade |
| 229 | Crvena Zvezda Telekom Belgrade |
| | Neptunas Klaipeda |
| 230 | Laboral Kutxa Vitoria Gasteiz |
| | Olympiacos Piraeus |
| 231 | Valencia Basket |
| | Galatasaray Liv Hospital Istanbul |
| 232 | Galatasaray Liv Hospital Istanbul |
| | Laboral Kutxa Vitoria Gasteiz |
| 233 | Olympiacos Piraeus |
| | Neptunas Klaipeda |
| 234 | Valencia Basket |
| | Crvena Zvezda Telekom Belgrade |
| 235 | Neptunas Klaipeda |
| | Galatasaray Liv Hospital Istanbul |
| 236 | Laboral Kutxa Vitoria Gasteiz |
| | Valencia Basket |
| 237 | Olympiacos Piraeus |
| | Crvena Zvezda Telekom Belgrade |
| 238 | Galatasaray Liv Hospital Istanbul |
| | Olympiacos Piraeus |
| 239 | Valencia Basket |
| | Neptunas Klaipeda |
| 240 | Crvena Zvezda Telekom Belgrade |
| | Laboral Kutxa Vitoria Gasteiz |

## Π4. Συντελεστές συσχέτισης την χρονιά 2014-2015

| | ranks | qualified | Pts.Total..M.O | dif_Pts.Total..M.O | X2FG.M.O |
|---|---|---|---|---|---|
| ranks | 1.00 | -0.73 | -0.33 | -0.75 | -0.39 |
| qualified | -0.73 | 1.00 | 0.27 | 0.49 | 0.23 |
| Pts.Total..M.O | -0.33 | 0.27 | 1.00 | 0.33 | 0.31 |
| dif_Pts.Total..M.O | -0.75 | 0.49 | 0.33 | 1.00 | 0.50 |
| X2FG.M.O | -0.39 | 0.23 | 0.31 | 0.50 | 1.00 |
| X3FG.M.O | -0.09 | 0.06 | 0.28 | 0.14 | 0.02 |
| FT.M.O | 0.00 | -0.01 | 0.26 | 0.12 | -0.15 |
| O.TOTAL | -0.07 | 0.16 | 0.00 | -0.04 | -0.23 |
| D.TOTAL | -0.50 | 0.42 | 0.17 | 0.43 | 0.29 |
| T.TOTAL | -0.61 | 0.58 | 0.21 | 0.42 | 0.12 |
| As.TOTAL | -0.39 | 0.23 | 0.27 | 0.45 | 0.40 |
| St..TOTAL | 0.05 | -0.17 | -0.14 | 0.00 | -0.01 |
| To.TOTAL | 0.06 | 0.03 | -0.40 | -0.15 | -0.09 |
| Fv.TOTAL | -0.22 | 0.23 | -0.08 | 0.15 | 0.01 |
| Ag.TOTAL | -0.12 | 0.05 | -0.09 | -0.06 | -0.18 |
| Cm.TOTAL | 0.26 | -0.18 | -0.31 | -0.37 | -0.24 |
| Rv.TOTAL | -0.11 | 0.23 | 0.10 | 0.03 | 0.03 |
| PIR.TOTAL | -0.54 | 0.42 | 0.66 | 0.60 | 0.55 |

| | X3FG.M.O | FT.M.O | O.TOTAL | D.TOTAL | T.TOTAL | As.TOTAL | St..TOTAL |
|---|---|---|---|---|---|---|---|
| ranks | -0.09 | 0.00 | -0.07 | -0.50 | -0.61 | -0.39 | 0.05 |
| qualified | 0.06 | -0.01 | 0.16 | 0.42 | 0.58 | 0.23 | -0.17 |
| Pts.Total..M.O | 0.28 | 0.26 | 0.00 | 0.17 | 0.21 | 0.27 | -0.14 |
| dif_Pts.Total..M.O | 0.14 | 0.12 | -0.04 | 0.43 | 0.42 | 0.45 | 0.00 |
| X2FG.M.O | 0.02 | -0.15 | -0.23 | 0.29 | 0.12 | 0.40 | -0.01 |
| X3FG.M.O | 1.00 | 0.45 | -0.25 | 0.14 | 0.00 | 0.09 | -0.18 |
| FT.M.O | 0.45 | 1.00 | -0.19 | 0.07 | 0.03 | 0.13 | -0.22 |
| O.TOTAL | -0.25 | -0.19 | 1.00 | -0.24 | 0.13 | 0.03 | 0.23 |
| D.TOTAL | 0.14 | 0.07 | -0.24 | 1.00 | 0.64 | 0.11 | -0.31 |
| T.TOTAL | 0.00 | 0.03 | 0.13 | 0.64 | 1.00 | 0.15 | -0.26 |
| As.TOTAL | 0.09 | 0.13 | 0.03 | 0.11 | 0.15 | 1.00 | 0.17 |
| St..TOTAL | -0.18 | -0.22 | 0.23 | -0.31 | -0.26 | 0.17 | 1.00 |
| To.TOTAL | -0.18 | -0.24 | 0.07 | 0.13 | 0.13 | 0.04 | 0.05 |
| Fv.TOTAL | -0.01 | -0.24 | -0.11 | 0.23 | 0.15 | -0.10 | 0.06 |
| Ag.TOTAL | -0.16 | -0.24 | 0.25 | 0.08 | 0.22 | 0.03 | 0.13 |
| Cm.TOTAL | 0.00 | -0.19 | 0.02 | -0.16 | -0.22 | -0.19 | 0.12 |
| Rv.TOTAL | 0.00 | 0.06 | 0.01 | 0.22 | 0.25 | -0.10 | -0.37 |
| PIR.TOTAL | 0.22 | 0.18 | -0.10 | 0.39 | 0.36 | 0.40 | -0.18 |

| | To.TOTAL | Fv.TOTAL | Ag.TOTAL | Cm.TOTAL | Rv.TOTAL | PIR.TOTAL |
|---|---|---|---|---|---|---|
| ranks | 0.06 | -0.22 | -0.12 | 0.26 | -0.11 | -0.54 |
| qualified | 0.03 | 0.23 | 0.05 | -0.18 | 0.23 | 0.42 |
| Pts.Total..M.O | -0.40 | -0.08 | -0.09 | -0.31 | 0.10 | 0.66 |
| dif_Pts.Total..M.O | -0.15 | 0.15 | -0.06 | -0.37 | 0.03 | 0.60 |
| X2FG.M.O | -0.09 | 0.01 | -0.18 | -0.24 | 0.03 | 0.55 |
| X3FG.M.O | -0.18 | -0.01 | -0.16 | 0.00 | 0.00 | 0.22 |
| FT.M.O | -0.24 | -0.24 | -0.24 | -0.19 | 0.06 | 0.18 |
| O.TOTAL | 0.07 | -0.11 | 0.25 | 0.02 | 0.01 | -0.10 |
| D.TOTAL | 0.13 | 0.23 | 0.08 | -0.16 | 0.22 | 0.39 |
| T.TOTAL | 0.13 | 0.15 | 0.22 | -0.22 | 0.25 | 0.36 |
| As.TOTAL | 0.04 | -0.10 | 0.03 | -0.19 | -0.10 | 0.40 |
| St..TOTAL | 0.05 | 0.06 | 0.13 | 0.12 | -0.37 | -0.18 |
| To.TOTAL | 1.00 | 0.01 | 0.09 | 0.21 | 0.07 | -0.29 |
| Fv.TOTAL | 0.01 | 1.00 | 0.11 | 0.10 | -0.01 | 0.03 |
| Ag.TOTAL | 0.09 | 0.11 | 1.00 | 0.12 | 0.10 | -0.15 |
| Cm.TOTAL | 0.21 | 0.10 | 0.12 | 1.00 | 0.00 | -0.48 |
| Rv.TOTAL | 0.07 | -0.01 | 0.10 | 0.00 | 1.00 | 0.20 |
| PIR.TOTAL | -0.29 | 0.03 | -0.15 | -0.48 | 0.20 | 1.00 |

## Π.5 Συντελεστές συσχέτισης την χρονιά 2015-2016

| | ranks | qualified | Pts.Total..M.O | dif_Pts.Total..M.O | X2FG.M.O |
|---|---|---|---|---|---|
| ranks | 1.00 | -0.73 | -0.50 | -0.74 | -0.50 |
| qualified | -0.73 | 1.00 | 0.53 | 0.63 | 0.60 |
| Pts.Total..M.O | -0.50 | 0.53 | 1.00 | 0.56 | 0.47 |
| dif_Pts.Total..M.O | -0.74 | 0.63 | 0.56 | 1.00 | 0.48 |
| X2FG.M.O | -0.50 | 0.60 | 0.47 | 0.48 | 1.00 |
| X3FG.M.O | -0.33 | 0.24 | 0.37 | 0.38 | 0.13 |
| FT.M.O | 0.02 | -0.04 | 0.12 | 0.11 | 0.19 |
| O.TOTAL | -0.17 | 0.26 | 0.23 | 0.18 | 0.02 |
| D.TOTAL | -0.37 | 0.50 | 0.39 | 0.54 | 0.27 |
| T.TOTAL | -0.33 | 0.50 | 0.43 | 0.44 | 0.19 |
| As.TOTAL | -0.27 | 0.27 | 0.49 | 0.24 | 0.30 |
| St..TOTAL | -0.12 | 0.05 | 0.11 | 0.19 | 0.13 |
| To.TOTAL | 0.09 | 0.03 | -0.15 | -0.16 | -0.09 |
| Fv.TOTAL | -0.18 | 0.29 | 0.48 | 0.20 | 0.08 |
| Ag.TOTAL | 0.09 | -0.02 | 0.21 | -0.04 | -0.10 |
| Cm.TOTAL | 0.18 | -0.08 | 0.03 | -0.07 | -0.19 |
| Rv.TOTAL | -0.45 | 0.47 | 0.49 | 0.53 | 0.34 |
| PIR.TOTAL | -0.58 | 0.60 | 0.80 | 0.70 | 0.54 |

| | X3FG.M.O | FT.M.O | O.TOTAL | D.TOTAL | T.TOTAL | As.TOTAL | St..TOTAL |
|---|---|---|---|---|---|---|---|
| ranks | -0.33 | 0.02 | -0.17 | -0.37 | -0.33 | -0.27 | -0.12 |
| qualified | 0.24 | -0.04 | 0.26 | 0.50 | 0.50 | 0.27 | 0.05 |
| Pts.Total..M.O | 0.37 | 0.12 | 0.23 | 0.39 | 0.43 | 0.49 | 0.11 |
| dif_Pts.Total..M.O | 0.38 | 0.11 | 0.18 | 0.54 | 0.44 | 0.24 | 0.19 |
| X2FG.M.O | 0.13 | 0.19 | 0.02 | 0.27 | 0.19 | 0.30 | 0.13 |
| X3FG.M.O | 1.00 | -0.04 | -0.10 | 0.22 | 0.15 | 0.26 | -0.05 |
| FT.M.O | -0.04 | 1.00 | 0.00 | -0.17 | -0.14 | -0.06 | -0.12 |
| O.TOTAL | -0.10 | 0.00 | 1.00 | 0.14 | 0.50 | 0.14 | -0.15 |
| D.TOTAL | 0.22 | -0.17 | 0.14 | 1.00 | 0.66 | 0.22 | 0.17 |
| T.TOTAL | 0.15 | -0.14 | 0.50 | 0.66 | 1.00 | 0.19 | 0.04 |
| As.TOTAL | 0.26 | -0.06 | 0.14 | 0.22 | 0.19 | 1.00 | -0.02 |
| St..TOTAL | -0.05 | -0.12 | -0.15 | 0.17 | 0.04 | -0.02 | 1.00 |
| To.TOTAL | 0.04 | -0.33 | 0.02 | 0.06 | 0.07 | 0.00 | 0.04 |
| Fv.TOTAL | 0.21 | -0.04 | 0.22 | 0.35 | 0.46 | 0.19 | 0.05 |
| Ag.TOTAL | 0.00 | -0.02 | 0.29 | 0.14 | 0.28 | 0.25 | -0.08 |
| Cm.TOTAL | 0.07 | -0.08 | 0.20 | 0.20 | 0.23 | -0.05 | -0.15 |
| Rv.TOTAL | 0.29 | 0.01 | 0.11 | 0.35 | 0.33 | 0.15 | 0.22 |
| PIR.TOTAL | 0.38 | 0.10 | 0.22 | 0.45 | 0.37 | 0.51 | 0.17 |

| | To.TOTAL | Fv.TOTAL | Ag.TOTAL | Cm.TOTAL | Rv.TOTAL | PIR.TOTAL |
|---|---|---|---|---|---|---|
| ranks | 0.09 | -0.18 | 0.09 | 0.18 | -0.45 | -0.58 |
| qualified | 0.03 | 0.29 | -0.02 | -0.08 | 0.47 | 0.60 |
| Pts.Total..M.O | -0.15 | 0.48 | 0.21 | 0.03 | 0.49 | 0.80 |
| dif_Pts.Total..M.O | -0.16 | 0.20 | -0.04 | -0.07 | 0.53 | 0.70 |
| X2FG.M.O | -0.09 | 0.08 | -0.10 | -0.19 | 0.34 | 0.54 |
| X3FG.M.O | 0.04 | 0.21 | 0.00 | 0.07 | 0.29 | 0.38 |
| FT.M.O | -0.33 | -0.04 | -0.02 | -0.08 | 0.01 | 0.10 |
| O.TOTAL | 0.02 | 0.22 | 0.29 | 0.20 | 0.11 | 0.22 |
| D.TOTAL | 0.06 | 0.35 | 0.14 | 0.20 | 0.35 | 0.45 |
| T.TOTAL | 0.07 | 0.46 | 0.28 | 0.23 | 0.33 | 0.37 |
| As.TOTAL | 0.00 | 0.19 | 0.25 | -0.05 | 0.15 | 0.51 |
| St..TOTAL | 0.04 | 0.05 | -0.08 | -0.15 | 0.22 | 0.17 |
| To.TOTAL | 1.00 | -0.02 | 0.14 | 0.11 | -0.15 | -0.16 |
| Fv.TOTAL | -0.02 | 1.00 | 0.38 | 0.04 | 0.28 | 0.40 |
| Ag.TOTAL | 0.14 | 0.38 | 1.00 | 0.32 | -0.10 | 0.06 |
| Cm.TOTAL | 0.11 | 0.04 | 0.32 | 1.00 | -0.02 | -0.09 |
| Rv.TOTAL | -0.15 | 0.28 | -0.10 | -0.02 | 1.00 | 0.60 |
| PIR.TOTAL | -0.16 | 0.40 | 0.06 | -0.09 | 0.60 | 1.00 |

## Π.6 Πίνακες συνάφειας για την χρονιά 2014-2015

Έλεγχος για το αν οι παραπάνω μεταβλητές είναι στατιστικά σημαντικές συγκριτικά με την μεταβλητή qualified.

**Οι επιτυχίες ορίζονται με βάση τα κριτήρια που υπάρχουν στους παρακάτω πίνακες :**

| | | Non-qualified | Qualified | Total |
|---|---|---|---|---|
| **Points** | ≤80 | 7 | 10 | 17 |
| | >80 | 1 | 6 | 7 |
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |
| **Dif_Points** | ≤ \|5\| | 3 | 10 | 13 |
| | >\|5\| | 5 | 6 | 11 |
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |
| **2FG** | ≤ 50% | 4 | 5 | 9 |
| | >50% | 4 | 11 | 15 |
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |
| **3FG** | ≤ 35% | 3 | 7 | 10 |
| | >35% | 5 | 9 | 14 |
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |
| **FT** | ≤ 75% | 6 | 10 | 16 |
| | >75% | 2 | 6 | 8 |
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |
| **O** | ≤ 110 | 5 | 8 | 13 |
| | >110 | 3 | 8 | 11 |

| | | | | |
|---|---|---|---|---|
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |
| **D** | ≤ 250 | 7 | 6 | 13 |
| | >250 | 1 | 10 | 11 |
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |
| **T** | ≤ 360 | 8 | 7 | 15 |
| | >360 | 0 | 9 | 9 |
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |
| **As** | ≤ 170 | 7 | 8 | 15 |
| | >170 | 1 | 8 | 9 |
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |
| **St** | ≤ 70 | 4 | 13 | 17 |
| | >70 | 4 | 3 | 7 |
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |
| **To** | ≤ 130 | 5 | 9 | 14 |
| | >130 | 3 | 7 | 10 |
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |
| **Fv** | ≤ 30 | 6 | 9 | 15 |
| | >30 | 2 | 7 | 9 |
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |

| | | | | |
|---|---|---|---|---|
| **Ag** | ≤ 30 | 5 | 11 | 16 |
| | >30 | 3 | 5 | 8 |
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |
| **Cm** | ≤ 210 | 3 | 7 | 10 |
| | >210 | 5 | 9 | 14 |
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |
| **Rv** | ≤ 210 | 5 | 9 | 14 |
| | >210 | 3 | 7 | 10 |
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |
| **PIR** | ≤800 | 5 | 5 | 10 |
| | >800 | 3 | 11 | 14 |
| **Total** | | 8 | 16 | 24 |

**Π.7 Πίνακες συνάφειας για την χρονιά 2015-2016**

| | | Non-qualified | Qualified | Total |
|---|---|---|---|---|
| **Points** | ≤80 | 8 | 11 | 19 |
| | >80 | 0 | 5 | 5 |
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |
| **Dif_Points** | ≤ \|5\| | 2 | 8 | 10 |
| | >\|5\| | 6 | 8 | 14 |
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |

| | | | | |
|---|---|---|---|---|
| **2FG** | ≤ 50% | 8 | 6 | 14 |
| | >50% | 0 | 10 | 10 |
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |
| **3FG** | ≤ 35% | 5 | 8 | 13 |
| | >35% | 3 | 8 | 11 |
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |
| **FT** | ≤ 75% | 4 | 10 | 14 |
| | >75% | 4 | 6 | 10 |
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |
| **O** | ≤ 110 | 7 | 8 | 15 |
| | >110 | 1 | 8 | 9 |
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |
| **D** | ≤ 250 | 7 | 8 | 15 |
| | >250 | 1 | 8 | 9 |
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |
| **T** | ≤ 360 | 8 | 10 | 18 |
| | >360 | 0 | 6 | 6 |
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |
| **As** | ≤ 170 | 4 | 5 | 9 |
| | >170 | 4 | 11 | 15 |

| | | | | |
|---|---|---|---|---|
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |
| **St** | ≤ 70 | 4 | 11 | 15 |
| | >70 | 4 | 5 | 9 |
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |
| **To** | ≤ 130 | 3 | 6 | 9 |
| | >130 | 5 | 10 | 15 |
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |
| **Fv** | ≤ 30 | 5 | 9 | 14 |
| | >30 | 3 | 7 | 10 |
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |
| **Ag** | ≤ 30 | 5 | 12 | 17 |
| | >30 | 3 | 4 | 7 |
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |
| **Cm** | ≤ 210 | 5 | 9 | 14 |
| | >210 | 3 | 7 | 10 |
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |
| **Rv** | ≤ 210 | 7 | 6 | 13 |
| | >210 | 1 | 10 | 11 |
| **Total** | | 8 | 16 | 24 |
| | | Non-qualified | Qualified | Total |

| | | | | |
|---|---|---|---|---|
| **PIR** | ≤800 | 7 | 4 | 11 |
| | >800 | 1 | 12 | 13 |
| **Total** | | 8 | 16 | 24 |

# ΒΙΒΛΙΟΓΡΑΦΙΑ

## ΕΛΛΗΝΙΚΗ


Γ. Ηλιόπουλος, Αθήνα 2013. Βασικές Μέθοδοι Εκτίμησης Παραμέτρων με σημείο και με διάστημα, Β΄ Έκδοση, Εκδόσεις Σταμούλη Α.Ε.

Γ. Ηλιόπουλος, Πειραιάς 2019. Γενικευμένα γραμμικά μοντέλα, Πανεπιστημιακές Σημειώσεις.

Γ. Κόκκινος, 2011. Παράλληλοι αλγόριθμοι εξόρυξης γνώσης από βάσεις δεδομένων με τεχνητά νευρωνικά δίκτυα και μηχανές υποστήριξης.

Δ. Ανζουλάκος, Πειραιάς 2017. Ανάλυση Δεδομένων με την Χρήση Στατιστικών Πακέτων : Εισαγωγή στην R, Σημειώσεις Παραδόσεων.

Ε. Ταβλάκης, Διπλωματική Εργασία, Πειραιάς 2013. Στατιστικά μοντέλα για την εξέλιξη του σκορ και το τελικό αποτέλεσμα σε έναν αγώνα μπάσκετ.

Μ. Κούτρας, Πειραιάς 2019. Εφαρμοσμένη Πολυμεταβλητή Ανάλυση, Εργαστηριακές Σημειώσεις.

Μ. Κούτρας, Χ. Ευαγγελάρας, Αθήνα 2010. Ανάλυση Παλινδρόμησης, Θεωρία και Εφαρμογές, Εκδόσεις Σταμούλη Α.Ε.

Μ. Κούτρας, Αθήνα 2004. Εισαγωγή στις Πιθανότητες : Θεωρία και Εφαρμογές, Μέρος Ι, Εκδόσεις Σταμούλη Α.Ε.

Μ. Κούτρας, Αθήνα 2005. Εισαγωγή στις Πιθανότητες : Θεωρία και Εφαρμογές, Μέρος ΙΙ, Εκδόσεις Σταμούλη Α.Ε.

Ν. Πελέκης, Πειραιάς 2019. Στατιστικές Μέθοδοι Εξόρυξης Δεδομένων, Σημειώσεις.

Τζαβελάς Γ. (2007). Γενικευμένα Γραμμικά Μοντέλα, Μέρος Α΄ - Λογιστική Παλινδρόμηση, Πανεπιστήμιο Πειραιώς.

# ΞΕΝΗ


Agresti, A. (2007). Building and Applying Logistic Regression Models, An Introduction to Categorical Data Analysis 2nd Edition. Wiley-Interscience.

Beckler M., Wang H., Papamichael M. (2008-09). NBA Oracle

Cene E. (2019). What is the difference between a winning and a losing team : insights from Euroleague Basketball.

Charpentier A. (2014). Computational Actuarial Science with R.

Croux, C. and Dehon, C. (2010). Influence functions of the Spearman and Kendall correlation measures. Statistical Methods and Applications, 19, 497-515.

Dunham. (2003). Data Mining, Introductory and Advanced Topics, Prentice Hall.

Egidi L. (2019). Comparing Statistical Models and Machine Learning Algorithms in Predicting Football Outcomes. 4th Athens University of Economics & Business Sports Analytics Workshop.

Kubatko J., Oliver D., Pelton K., Dan T. Rosenbaum. (2007). Journal of Quantitative Analysis in Sports, A Starting Point for Analyzing Basketball Statistics

Mandic R., Jakovljevic S., Erculj F., Strumbelj E. (2019). Trends in NBA and Euroleague basketball : Analysis and comparison of statistical data from 2000 to 2017.

Manisera M. (2019). Measuring spatial performance in Basketball by CART. 4th Athens University of Economics & Business Sports Analytics Workshop.

Marmarinos C. (2019). Basic Basketball Analytics : Meaning, Formulas, Interpretation and Forms of Visualization of Data. 4th Athens University of Economics & Business Sports Analytics Workshop.

Ricci V. (2005). Fitting Distributions in R.

Schwab A. J. (2003). Logistic Regression – Complete Problems, University of Texas.

Shahzeb, F. (2017). The Evolution and Future of Analytics in Sport. *Proem Sports | Sports Analytics | Singapore & India*.

Smith T. & Schwertman C. Neil (1999). Can the NCAA Basketball Tournament Seeding be Used to Predict Margin of Victory? , The American Statistician, Vol. 53, No. 2., pp. 94-98.

Srivastava T N, Rego S. (2011). Business research methodology.

Stekler H.O. and Klein A. (2011). Predicting the Outcomes of NCAA Basketball Championship Game, Research Program on Forecasting, Department of Economics, George Washington University.

Tan, Steinbach, Kumar, (2006). Introduction to Data Mining, Addison Wesley.

William R. Klecka (1980). Discriminant Analysis.

Witkos R. (2010). Determing the Success of NCAA Basketball Teams through Team Characteristics.

WorkInSports. (2017). "How Data Analytics Helps Coaches in Planning".

Zuccolotto P. and Manisera M. (2020). Basketball Data Science – With Applications in R, Chapman and Hall/CRC.

# ΣΥΝΔΕΣΜΟΙ


https://www.euroleague.net/competition/teams/showteam?clubcode=RED&seasoncode=E2015#!E2015_RS

https://el.wikipedia.org/wiki/%CE%A3%CF%85%CF%83%CF%87%CE%AD%CF%84%CE%B9%CF%83%CE%B7_%CE%BA%CE%B1%CE%B9_%CE%B5%CE%BE%CE%AC%CF%81%CF%84%CE%B7%CF%83%CE%B7

http://www.r-tutor.com/gpu-computing/correlation/kendall-rank-coefficient

http://www.sthda.com/english/wiki/correlation-analyses-in-r

http://rcompanion.org/handbook/

https://repository.kallipos.gr/bitstream/11419/5081/1/07_chapter6.pdf

http://www.mas.ucy.ac.cy/~fokianos/GreekRbook/pca&discriminant.pdf

https://rstudio-pubs-static.s3.amazonaws.com/298913_9bd76dd24a9241cfa112d19a5e50610e.html

https://repository.kallipos.gr/bitstream/11419/2128/1/04_chapter03.pdf

https://www.r-bloggers.com/feature-selection-with-the-boruta-algorithm/

https://www.r-bloggers.com/how-to-implement-random-forests-in-r/

https://www.sheffield.ac.uk/polopoly_fs/1.714563!/file/stcp-karadimitriou-MannWhitR.pdf

https://www.wikiwand.com/el/

https://www.forbes.com/

https://www.82games.com/shotzones.html

https://arrow.tudublin.ie/cgi/viewcontent.cgi?article=1040&context=scschcomdis

https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

http://www.statstutor.ac.uk/resources/uploaded/pearsons.pdf

https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/

https://www.andreaperlato.com/mlpost/feature-selection-using-boruta-algorithm/

https://www.datacamp.com/community/tutorials/feature-selection-R-boruta