ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

UNIVERSITY OF PIRAEUS

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

# ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

## "ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ & ΥΠΗΡΕΣΙΕΣ''

## ΠΡΟΒΛΕΠΤΙΚΗ ΑΝΑΛΥΤΙΚΗ ΜΕ ΤΗ ΧΡΗΣΗ ΧΡΟΝΟΣΕΙΡΩΝ


Από

ΒΑΓΓΈΛΗ ΚΑΡΑΣΑ


Υποβάλλεται

για την εκπλήρωση των προϋποθέσεων λήψης

Μεταπτυχιακού Διπλώματος

Στην ειδίκευση ΜΔΑ/ΠΠΣ/ΠΔ

του ΠΜΣ "Πληροφοριακά Συστήματα & Υπηρεσίες"

στο

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΦΕΒΡΟΥΑΡΙΟΣ 2024


Επιβλέπτων/Επιβλέπουσα: Μιχαήλ Φιλιππάκης

Ακαδημαϊκή Θέση: Καθήγητής Τμήματος Προηγμένων Πληροφοριακών Συστημάτων Πανεπιστήμιο Πειραιώς

Πανεπιστήμιο Πειραιώς.Κάτοχος όλων των δικαιωμάτων

**ΣΕΛΙΔΑ ΕΓΚΥΡΟΤΗΤΑΣ**

**Ονοματεπώνυμο Φοιτητή/Φοιτήτριας**: Καρασά Βαγγέλης

**Τίτλος Μεταπτυχιακής Διπλωματικής Εργασίας**: Προβλεπτική Αναλυτική με τη χρήση χρονοσειρών

Η παρούσα Μεταπτυχιακή Διπλωματική Εργασία υποβάλλεται ως μερική εκπλήρωση των απαιτήσεων του Προγράμματος Μεταπτυχιακών Σπουδών "Πληροφοριακά Συστήματα & Υπηρεσίες" του Τμήματος Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιώς και εγκρίθηκε στις 13-03-2024. από τα μέλη της Εξεταστικής Επιτροπής.

**Εξεταστική Επιτροπή**

**Κυριαζής Δημοσθένης** **Φιλιππάκης Μιχαήλ** **Χαλκίδη Μαρία**

**ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΑΥΘΕΝΤΙΚΟΤΗΤΑΣ**

Ο/Η Καρασάς Βαγγέλης , γνωρίζοντας τις συνέπειες της λογοκλοπής, δηλώνω υπεύθυνα ότι η παρούσα εργασία με τίτλο «Προβλεπτική Αναλυτική με τη χρήση Χρονοσειρών », αποτελεί προϊόν αυστηρά προσωπικής εργασίας και όλες οι πηγές που έχω χρησιμοποιήσει, έχουν δηλωθεί κατάλληλα στις βιβλιογραφικές παραπομπές και αναφορές. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή.

Επιπλέον δηλώνω υπεύθυνα ότι η συγκεκριμένη Μεταπτυχιακή Διπλωματική Εργασία έχει συγγραφεί από εμένα προσωπικά και δεν έχει υποβληθεί ούτε έχει αξιολογηθεί στο πλαίσιο κάποιου άλλου μεταπτυχιακού ή προπτυχιακού τίτλου σπουδών, στην Ελλάδα ή στο εξωτερικό.

Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου. Σε κάθε περίπτωση, αναληθούς ή ανακριβούς δηλώσεως, υπόκειμαι στις συνέπειες που προβλέπονται τις διατάξεις που προβλέπει η Ελληνική και Κοινοτική Νομοθεσία περί πνευματικής ιδιοκτησίας.

**Ο/Η ΔΗΛΩΝ/ΟΥΣΑ**

**Ονοματεπώνυμο: Καρασά Βαγγέλης**

**Αριθμός Μητρώου: me 2240**

**Υπογραφή**

## Ευχαριστίες

Με την ολοκλήρωση της μεταπτυχιακής διπλωματικής μου εργασίες θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες σε όλους όσους συνέβαλλαν στην εκπόνηση της.

Ευχαριστώ θερμά τον επιβλέπων καθηγητή μου κύριο Φιλλιπάκη Μιχαήλ, για την εμπιστοσύνη που μου έδειξε αναθέτοντάς μου το συγκεκριμένο θέμα , τις υποδείξεις του και την συνεχή του υποστήριξη που έδειξε από αρχή μέχρι το τέλος.

Τέλος, θα ήθελα να εκφράσω την ευγνωμοσύνη μου στην οικογένεια μου,τη συντροφό μου, τους φίλους μου για τη στήριξη, τη συμπαράσταση καθώς και τη κατανόησή τους καθ'όλη της διάρκεια των σπουδών μου.

# Περιεχόμενα

# Πίνακας περιεχομένων εικόνων

## Περίληψη

Οι προβλέψεις έχουν ένα μεγάλο ρόλο στη σημερινή κοινωνία καθώς και για τη λειτουργία της. Πολλές είναι οι έρευνες που έχουν διερευνήσει μεθόδους προβλέψεων καθώς και την εφαρμογή τους. Σκοπός της παρούσας εργασία είναι να διερευνηθεί η αποτελεσματικότητα των μοντέλων ARIMA σε ένα ευρύ φάσμα από δεδομένα που περιλαμβάνουν χρονοσειρές. Οι εφαρμογές ήταν σε τέσσερις περιπτώσεις από τις οποίες λάβαμε διάφορα συμπεράσματα σχετικά με την αποτελεσματικότητα τους καθώς και τις συνθήκες υπό τις οποίες είναι προτιμότερο να χρησιμοποιούνται.

Λέξεις κλειδιά: προβλέψεις , μοντέλα προβλέψεων , χρονοσειρές

## Abstract

Predictions have a big role in today's society as well as its functioning. There are many researches that have investigated forecasting methods as well as their application. The purpose of this paper is to investigate the effectiveness of ARIMA models on a wide range of data that include time series. The applications were in four cases from which we obtained various conclusions about their effectiveness as well as the conditions under which they are preferable to be used.

Keywords: forecasting, forecasting models, time series

# ΚΕΦΑΛΑΙΟ 1 ΕΙΣΑΓΩΓΗ

## ΤΙ ΕΙΝΑΙ Η ΠΡΟΒΛΕΨΗ

Συνολικά, οι δραστικές αλλαγές σε όλους τους τομείς της καθημερινής ζωής και η υφιστάμενη αβεβαιότητα έχουν προσελκύσει το ενδιαφέρον τόσο της ακαδημαϊκής κοινότητας όσο και του επιχειρηματικού κόσμου. Από τη σκοπιά των ακαδημαϊκών, δημιουργήθηκαν διάφορες μέθοδοι πρόβλεψης, κάποιες θεωρητικά θεμελιωμένες και άλλες επαναστατικές με βάση την τεχνολογία και τις διαθέσιμες δυνατότητες. Οι μέθοδοι επιλέγονται ανάλογα με παράγοντες όπως το αντικείμενο της πρόβλεψης, ο χρονικός ορίζοντας, τα υπάρχοντα δεδομένα, το κόστος και η ευκολία χρήσης. Η πραγματική αξία μιας μεθόδου αξιολογείται κατά την εφαρμογή της σε πραγματικές συνθήκες. Η συνεργασία μεταξύ ακαδημαϊκού και επιχειρηματικού κόσμου μπορεί να οδηγήσει σε βελτιωμένα μοντέλα πρόβλεψης που προσαρμόζονται στις ιδιαίτερες ανάγκες της κάθε επιχείρησης. Από την πλευρά του επιχειρησιακού περιβάλλοντος, η πρόβλεψη αποτελεί κρίσιμο στοιχείο για τη λήψη αποφάσεων, τον σχεδιασμό στρατηγικής, το χρονικό προγραμματισμό και την πολιτική λειτουργίας. Ακόμη και μια μικρή βελτίωση στην ακρίβεια των προβλέψεων μπορεί να έχει σημαντικές επιπτώσεις στα οικονομικά αποτελέσματα μιας επιχείρησης. Επομένως, οι υπηρεσίες πρόβλεψης και τα προϊόντα που παρέχουν είναι αναζητούμενα, με τη συνεχή ανάγκη για ακριβείς προβλέψεις να οδηγεί στην ανάπτυξη καινοτόμων διαδικασιών πρόβλεψης.

Για να χρησιμοποιήσει κάποιος προς όφελός του και με σωστό τρόπο τις προβλέψεις, πρέπει να μπορεί να κάνει τις πιο σοφές επιλογές (Μόλης Θ) . Και αυτό είναι απόρροια της πείρας. Στελέχη μεγάλων επιχειρήσεων ,στελέχη που εκλέγονται στα κράτη κ.α. πρέπει να διαθέτουν την κατάλληλη εμπειρία στις καταστάσεις, που έχουν να αντιμετωπίσουν και οι αποφάσεις τους να μην είναι αποτέλεσμα από προσωπικά ενδιαφέροντα, φιλοδοξίες και σκοπιμότητες. Για να επιτευχθεί ο εκάστοτε στόχος πρέπει οι αποφάσεις τους να είναι όσο το δυνατόν πιο αντικειμενικές και να μην βασίζονται μόνο στα “πειραματικά”, ή μόνο στα θεωρητικά αποτελέσματα, αλλά να έχουν την ικανότητα να συγκρίνουν και να συνδυάζουν τα αποτελέσματα και των δύο και να λαμβάνουν υπόψη τους τόσο τις οικονομικές καταστάσεις του κλάδου τους όσο και των χωρών που θα ενεργοποιηθούν.

## ΚΡΙΣΗ ΣΤΗ ΠΡΟΒΛΕΨΗ

Οι έντονες μεταβολές σε όλους τους τομείς της καθημερινής ζωής και η αβεβαιότητα σχετικά με την εξέλιξη των συνθηκών στο μέλλον έχουν επιστημονικό και επιχειρηματικό ενδιαφέρον. Ο ακαδημαϊκός κόσμος έχει αναπτύξει πολλές μεθόδους πρόβλεψης, ορισμένες θεωρητικές και άλλες πιο τεχνολογικά εξελιγμένες, προσαρμοσμένες στις ανάγκες της εποχής.

Η επιλογή της κατάλληλης μεθόδου πρόβλεψης εξαρτάται από παράγοντες όπως το αντικείμενο της πρόβλεψης, ο χρονικός ορίζοντας, τα διαθέσιμα δεδομένα, το κόστος και η ευκολία χρήσης. Επίσης, η ανθρώπινη κρίση παίζει καίριο ρόλο στη διαδικασία πρόβλεψης, καθώς η εμπειρία και η συνεισφορά της συνδυάζονται με τις ποσοτικές μεθόδους (Provost F, Fawcett T).

Σε μια εποχή που η τεχνολογία και οι υπολογιστές ευνοούν τις ποσοτικές προσεγγίσεις, η ανθρώπινη κρίση πρέπει να παραμένει κεντρική. Η σκέψη και η ανάλυση από πλευράς ανθρώπινης κρίσης είναι αναγκαίες για τη δημιουργία ισορροπημένων και αξιόπιστων προβλέψεων.

## ΤΥΠΟΙ ΠΡΟΒΛΕΨΗΣ

Είναι πολύ σημαντικό να καταλάβουμε ότι μία επιχείρηση πλέον μπορεί να συναγάγει χρήσιμες προβλέψεις αν ερμηνεύσει σωστά το παρελθόν. Τα δεδομένα της συμπεριφοράς των πελατών στο παρελθόν σκιαγραφούν τη μελλοντική συμπεριφορά τους. Η ζήτηση προκύπτει με βάση κάποιας συγκεκριμένης λογικής. Για τη διενέργεια των προβλέψεων έχουν αναπτυχθεί αρκετές μέθοδοι και η επιλογή τους εξαρτάται από τα εξής (Gebhard, 2006) :

Από το είδος των αποφάσεων που θα ληφθούν με βάση τις προβλέψει που θα προκύψουν

Από την περίοδο και τον ορίζοντα πρόβλεψης. Περίοδος πρόβλεψης είναι η χρονική μονάδα μέτρησης (μέρα εβδομάδα κ.α.) ενώ ορίζοντας ο αριθμός περιόδων για τον οποίο θα ληφθούν οι αποφάσεις. Για παράδειγμα, για στρατηγικές αποφάσεις ο ορίζοντας πρόβλεψης μπορεί να είναι δέκα έτη ενώ η περίοδος πρόβλεψης ένα έτος.

Από τη ζητούμενη ακρίβεια , το επίπεδο ακρίβειας των προβλέψεων είναι μια παράμετρος που καθορίζεται από τον χρήστη , ανάλογα με το είδος των προβλέψεων και την ασφάλεια που επιδιώκει να πετύχει. Η ακρίβεια μιας μεθόδου αυξάνει όσο περισσότερο στηρίζεται σε ποσοτικά στοιχεία, όσο μεγαλύτερο είναι το πλήθος των στοιχείων και όσο μικρότερος είναι ο χρονικός ορίζοντας των προβλέψεων.

Από τα διαθέσιμα στοιχεία. Το είδος και η ποσότητα των διαθεσίμων στοιχείων επηρεάζουν την επιλογή της μεθόδου αφού κάθε μέθοδος έχει διαφορετικές απαιτήσεις. Για παράδειγμα, οι μέθοδοι που βασίζονται σε χρονοσειρές απαιτούν ακριβή ποσοτικά στοιχεία για ένα συνήθως μεγάλο σύνολο περιόδων. Δεν συμβαίνει το ίδιο με τις ποιοτικές μεθόδους. Επιπλέον, τα διαθέσιμα στοιχεία συχνά μπορούν , μετά από κατάλληλη ανάλυση , να υποδείξουν τη συνάρτηση που χαρακτηρίζει τη μεταβλητή για την οποία ζητείται η πρόβλεψη. Έτσι για παράδειγμα , ίσως υποδεικνύουν ότι η μεταβλητή χαρακτηρίζεται από εποχικότητα και ότι επομένως πρέπει να επιλεγεί μέθοδος κατάλληλη για τέτοιες προβλέψεις.

Οι τεχνικές πρόβλεψης χωρίζονται σε μακροπρόθεσμες και βραχυπρόθεσμες, και και οι δύο είναι ουσιώδεις για την επιτυχημένη λειτουργία μιας επιχείρησης ή ενός κράτους. Ενώ οι ποσοτικές μέθοδοι προβλέψεων παρέχουν εξελιγμένες τεχνικές, η ανθρώπινη κρίση, η ανάλυση και η κοινή λογική πρέπει να καθοδηγούν τη διαδικασία. Η αποτελεσματική πρόβλεψη απαιτεί τον συνδυασμό ποσοτικών και ποιοτικών δεδομένων. Ένας αναλυτής πρέπει να είναι ικανός να σχηματίσει έναν ολοκληρωμένο συνδυασμό δεδομένων προκειμένου να αποφευχθούν σφάλματα και να επιτευχθεί η μέγιστη αξιοπιστία στις προβλέψεις. Είναι, ωστόσο, κρίσιμο να υπογραμμιστεί ότι η αποτελεσματική πρόβλεψη απαιτεί σκέψη, ανθρώπινη κρίση, και τη συνεισφορά καινοτόμων προσεγγίσεων. Ο συνδυασμός ποσοτικών και ποιοτικών δεδομένων είναι κλειδί για τη δημιουργία αξιόπιστων προβλέψεων που αντανακλούν τις πραγματικές ανάγκες και συνθήκες μιας επιχείρησης ή κοινωνίας. Συνοψίζοντας, ο πιο αποτελεσματικός αναλυτής πρέπει να είναι ικανός να ενσωματώνει και να αξιοποιεί τόσο τις ποσοτικές όσο και τις ποιοτικές πλευρές, δημιουργώντας έναν ισορροπημένο και ολοκληρωμένο τρόπο πρόβλεψης.

## ΤΑ ΒΗΜΑΤΑ ΤΗΣ ΠΡΟΒΛΕΨΗΣ

Το γεγονός ότι οι τεχνικές πρόβλεψης βασίζονται σε δεδομένα που προέρχονται από ιστορικά γεγονότα οδηγεί στον προσδιορισμό των παρακάτω πέντε βημάτων (Siegel, 2013):

1. Διατύπωση προβλήματος και συλλογή δεδομένων.
2. Διαχείριση δεδομένων και τελική επιλογή.
3. Κατασκευή υποδείγματος και αξιολόγηση του.
4. Εφαρμογή υποδείγματος (η πραγματική πρόβλεψη).
5. Αξιολόγηση πρόβλεψης.

**Στο πρώτο βήμα,** η διατύπωση του προβλήματος και η συλλογή των δεδομένων αντιμετωπίζονται ως ένα ενιαίο βήμα, διότι είναι αλληλένδετα. Το πρόβλημα καθορίζει τα κατάλληλα δεδομένα. Αν εξετάζεται μια ποσοτική μεθοδολογία πρόβλεψης, θα πρέπει να είναι διαθέσιμα και ακριβή τα κατάλληλα δεδομένα. Εάν τα κατάλληλα δεδομένα δεν είναι διαθέσιμα, τότε το πρόβλημα μπορεί να χρειαστεί να επαναπροσδιοριστεί. Επίσης μπορεί να εμφανιστούν προβλήματα συλλογής και ελέγχου ποιότητας, κάθε φορά που καθίσταται αναγκαία η λήψη κατάλληλων δεδομένων για να γίνει μια πρόβλεψη.

Στο δεύτερο βήμα, η διαχείριση των στοιχείων και η τελική επιλογή τους είναι απαραίτητη. Είναι δυνατόν να έχουμε, είτε πάρα πολλά δεδομένα, είτε πολύ λίγα στη διαδικασία πρόβλεψης. Ορισμένα δεδομένα μπορεί να μην είναι σχετικά με το πρόβλημα. Άλλα δεδομένα μπορεί να είναι κατάλληλα, αλλά μόνο σε συγκεκριμένες ιστορικές περιόδους. Συνήθως, απαιτείται κάποια προσπάθεια για να μετατραπούν τα δεδομένα στην κατάλληλη μορφή για τη χρήση συγκεκριμένων διαδικασιών πρόβλεψης.

Το τρίτο βήμα, η κατασκευή του μοντέλου και η αξιολόγηση του, περιλαμβάνει την τοποθέτηση των συλλεχθέντων στοιχείων σε ένα μοντέλο – υπόδειγμα πρόβλεψης, που είναι κατάλληλο για την ελαχιστοποίηση των σφαλμάτων προσαρμογής με τελικό στόχο την πρόβλεψη. Όσο πιο απλό είναι το μοντέλο, τόσο το καλύτερο για την αποδοχή της διαδικασίας πρόβλεψης από τους μάνατζερ. Προφανώς, η ανθρώπινη κρίση περιλαμβάνεται σε αυτό το βήμα.

Το τέταρτο βήμα, η εφαρμογή της μεθόδου, είναι η δημιουργία ενός πραγματικού μοντέλου με προβλέψεις, παράγουμε προβλέψεις, μόλις τα κατάλληλα δεδομένα έχουν συλλεχθεί και έχει γίνει η τελική επιλογή. Τα δεδομένα από τις πρόσφατες ιστορικές περιόδους χρησιμοποιούνται αργότερα για να γίνει ο έλεγχος για την ορθότητα της διαδικασίας.

Στο πέμπτο βήμα, η αξιολόγηση των προβλέψεων, περιλαμβάνεται η σύγκριση των τιμών με πραγματικές ιστορικές τιμές. Οι προβλέψεις αυτές συγκρίνονται στη συνέχεια με τις γνωστές ιστορικές τιμές και εξετάζονται τα τυχόν σφάλματα που προκύπτουν. Η

εξέταση των σφαλμάτων συχνά οδηγεί τον αναλυτή στην τροποποίηση του μοντέλου πρόβλεψης. Ειδικές μέθοδοι μέτρησης των σφαλμάτων αναλύονται στο τέλος του κεφαλαίου 3.

Οι διαδικασίες πρόβλεψης μπορούν επίσης να ταξινομηθούν ανάλογα με το αν είναι περισσότερο ποσοτικές, ή ποιοτικές. Από την μια μεριά, μία καθαρά ποιοτική τεχνική είναι μία, που δεν απαιτεί καμία εμφανή διαχείριση των δεδομένων. Μόνο η κρίση του αναλυτή απαιτείται. Από την άλλη μεριά, οι καθαρά ποσοτικές τεχνικές δεν απαιτούν ανθρώπινη κρίση. Είναι μηχανικές μαθηματικές διαδικασίες που καταλήγουν σε ποσοτικά αποτελέσματα. Ορισμένες ποσοτικές διαδικασίες, φυσικά, απαιτούν μια πολύ πιο εκλεπτυσμένη διαχείριση των στοιχείων από ότι άλλες.

Έχοντας αναφέρει κάποια βασικά στοιχεία της πρόβλεψης και τους τρόπους με τους οποίους τη διαχειριζόμαστε , θα αναφερθούμε σε ένα πιο συγκεκριμένο τύπο προβλέψεων. Τη προβλεπτική αναλυτική με τη ανάλυση των χρονοσειρών.

## ΤΑ ΣΦΑΛΜΑΤΑ ΤΗΣ ΠΡΟΒΛΕΨΗΣ

Ως αποτελεσματικότητα μίας πρόβλεψης ορίζουμε τη σχέση μεταξύ των αποτελεσμάτων ( προβλέψεων ) που παράγει η μέθοδος με τις τιμές που εμφανίζονται στη πράξη. Η αποτελεσματικότητα αξιολογείται με μια σειρά δεικτών οι οποίοι μετρούν τις αποκλίσεις ( σφάλματα ) πρόβλεψης . Έστω ότι για μια σειρά N περιόδων διατίθενται προβλέψεις και τις αντίστοιχες πραγματικές τιμές . Ορίζουμε ως Et το σφάλμα της πρόβλεψης , δηλαδή τη διαφορά ανάμεσα στη πραγματική τιμή Dt και την πρόβλεψη Ft της μεταβλητής για τη περίοδο τ, δηλαδή:

Et=Dt-Ft

Για την αξιολόγηση της αποτελεσματικότητας μιας μεθόδου , χρησιμοποιούνται οι παρακάτω δείκτες (Βιδάλης Μ.) :

Μεσο σφάλμα (Average error):

$$(\boldsymbol{M\Sigma}) = \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{\tau=1}}^{\boldsymbol{N}} \boldsymbol{e_t}$$

Η αποτελεσματικότητα μιας μεθόδου κρίνεται καλή όταν αυτός ο δείκτης τείνει στο μηδέν. Τυχόν υψηλή θετική τιμή δείχνει ότι η χρησιμοποιούμενη μέθοδος υποεκτιμά τη μεταβλητή για την οποία γίνεται η πρόβλεψη ενώ τυχόν υψηλή αρνητική τιμή δείχνει ότι η μέθοδος κάνει υπερεκτίμηση. Μειονέκτημα αυτού του δείκτη είναι το γεγονός ότι οι θετικές αποκλίσεις εξουδετερώνονται από τις αρνητικές με αποτέλεσμα να εμφανίζονται πλασματικά καλή αποτελεσματικότητα, δηλαδή μικρό μέσο σφάλμα αν και έχουν σημειωθεί στην πραγματικότητα πολύ μεγάλες αποκλίσεις.

Μέση απόλυτη απόκλιση(Mean absolute deviation , MAD):

$$(\boldsymbol{MAA}) = \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{t=1}}^{\boldsymbol{N}} |\boldsymbol{e_t}|$$

Για να αντισταθμίσουμε το προηγούμενο μειονέκτημα (εξουδετέρωση αρνητικών και θετικών αποκλίσεων) , εισάγουμε την απόλυτη τιμή της απόκλισης . Ο δείκτης αυτός δίνει ένα μέτρο μεγέθους των αποκλίσεων που παράγει η χρησιμοποιούμενη μέθοδος , αλλά δεν δίνει το πρόσημο τους . Δείχνει αν είναι καλή μια μέθοδος , αλλά σε περίπτωση που δεν είναι , δεν μας πληροφορεί αν έχουμε υπερεκτίμηση η υποεκτίμηση της μεταβλητής για την οποία γίνεται πρόβλεψη.

Μέσο τετραγωνικό σφάλμα(Mean squared error,MSE):

$$(\boldsymbol{MT\Sigma}) = \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{\tau=1}}^{\boldsymbol{N}} \boldsymbol{e_t^2}$$

Ο δείκτης παρέχει πληροφορίες παρόμοιες με τη ΜΑΑ, αλλά λόγω του εκθέτη στον οποίο υψώνεται κάθε απόκλιση ( υψώνεται στο τετράγωνο) παρουσιάζει μεγαλύτερη ευαισθησία στις αποκλίσεις. (Dorothy, 2018)

Τυπική απόκλιση σφαλμάτων(Standard deviation of error,σ):

Δίνει πληροφορίες για την απόκλιση των απόλυτων τιμών των σφαλμάτων από τη μέση απόκλιση σ των σφαλμάτων δίνεται από τη σχέση σ=1,25 ΜΑΑ.

Ποσοστιαίο σφάλμα(Percentage error, PE):

$$(\boldsymbol{\Pi\Sigma}) = \frac{\boldsymbol{e_t}}{\boldsymbol{D_t}} \mathbf{100}$$

Μετρά το ποσοστιαίο σφάλμα για μία μόνο περίοδο.

Μέσο ποσοστιαίο σφάλμα(Mean percentage error, MPE):

$$(\boldsymbol{M\Pi\Sigma}) = \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{\tau=1}}^{\boldsymbol{N}} \frac{\boldsymbol{e_t}}{\boldsymbol{D_t}} \mathbf{100}$$

Μετρά το μέσο ποσοστιαίο σφάλμα για N περιόδους.

Μέσο απόλυτο ποσοστιαίο σφάλμα(Mean absolute percentage error, MAPE):

$$(\boldsymbol{MA\Pi\Sigma}) = \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{t=1}}^{\boldsymbol{N}} \left| \frac{\boldsymbol{e_t}}{\boldsymbol{D_t}} \right| \mathbf{100}$$

Το ΜΑΠΣ μετρά πόσο αποκλίνουν οι προβλέψεις ως ποσοστά της πραγματικής τιμής της μεταβλητής.

Αμεροληψία(bias):

$$\boldsymbol{Bias} = \sum_{\boldsymbol{t=1}}^{\boldsymbol{N}} \boldsymbol{e_t}$$

Για να προσδιορίσουμε αν μια μέθοδος πρόβλεψης υπερεκτιμά η υποεκτιμά συστηματικά τη ζήτηση , πρέπει να χρησιμοποιήσουμε το άθροισμα των σφαλμάτων που παράγονται για να υπολογίσουμε την αμεροληψία της μεθόδου. Εφόσον τα σφάλματα είναι τελείως τυχαία και δεν ακολουθούν κάποια τάση ( υπερεκτίμησης , η υποεκτίμησης), η αμεροληψία πρέπει να είναι περίπου μηδενική.

Για τη σύγκριση των εναλλακτικών μεθόδων πρόβλεψης της ζήτησης μπορούν να χρησιμοποιηθούν ένας η (συνήθως) περισσότεροι δείκτες. Άλλωστε , για την ανάλυση της αξιοπιστίας και της ακρίβειας της εφαρμοζόμενης μεθόδου πρόβλεψης της ζήτησης , μπορεί να χρησιμοποιηθεί μια κατάλληλη μέθοδος ελέγχου η οποία θα δείχνει αν υπάρχει ανάγκη τροποποίηση η εγκατάλειψης της μεθόδου. Πιο συγκεκριμένα , χρησιμοποιείται το σήμα ανίχνευσης.

Σήμα ανίχνευσης(tracking signal):

$$\boldsymbol{\Sigma A = \frac{bias}{MAA}}$$

Όταν το ΣΑ βρίσκεται εκτός του εύρους των +- 6 , αυτό αποτελεί ένδειξη ότη η πρόβλεψη έχει τάση είτε υποεκτίμησης ( όταν TS<=-6) είτε υπερεκτίμησης ( TS>-6) , οπότε απαιτείται να διερευνηθεί η αιτία απόκλισης και τελικά, η καταληλλότητα της μεθόδου.

# ΚΕΦΑΛΑΙΟ 2

# ΠΡΟΒΛΕΠΤΙΚΗ ΑΝΑΛΥΤΙΚΗ ΜΕ ΤΗ ΑΝΑΛΥΣΗ ΧΡΟΝΟΣΕΙΡΩΝ

## 2.1 ΤΙ ΕΙΝΑΙ Η ΠΡΟΒΛΕΠΤΙΚΗ ΑΝΑΛΥΤΙΚΗ

Η προβλεπτική αναλυτική αναφέρεται στη χρήση στατιστικών και τεχνικών μοντέλων για να γίνουν προβλέψεις σχετικά με τα μελλοντικά αποτελέσματα και τις επιδόσεις έως τώρα . Η προβλεπτική αναλυτική εξετάζει τα τρέχοντα και ιστορικά μοτίβα δεδομένων για να καθορίσει εάν αυτά τα μοτίβα είναι πιθανό να εμφανιστούν ξανά. Αυτό επιτρέπει στις επιχειρήσεις και τους επενδυτές να προσαρμόσουν το πού χρησιμοποιούν τους πόρους τους για να επωφεληθούν από πιθανά μελλοντικά γεγονότα. Η προβλεπτική αναλυτική μπορεί επίσης να χρησιμοποιηθεί για τη βελτίωση της λειτουργικής αποτελεσματικότητας και τη μείωση του κινδύνου.

Βασίζεται σε μια σειρά τεχνικών για να κάνει αυτούς τους προσδιορισμούς, όπως η τεχνητή νοημοσύνη (AI), η εξόρυξη δεδομένων, η μηχανική μάθηση, η μοντελοποίηση και η στατιστική.Για παράδειγμα, η εξόρυξη δεδομένων περιλαμβάνει την ανάλυση μεγάλων συνόλων δεδομένων για τον εντοπισμό προτύπων από αυτήν. Η ανάλυση κειμένου κάνει το ίδιο, εκτός από μεγάλα μπλοκ κειμένου.

Τα μοντέλα πρόβλεψης χρησιμοποιούνται για όλα τα είδη εφαρμογών, συμπεριλαμβανομένων των προγνώσεων καιρού, της δημιουργίας βιντεοπαιχνιδιών, της μετάφρασης φωνής σε κείμενο, της εξυπηρέτησης πελατών και των στρατηγικών χαρτοφυλακίου επενδύσεων. Όλες αυτές οι εφαρμογές χρησιμοποιούν περιγραφικά στατιστικά μοντέλα υπαρχόντων δεδομένων για να κάνουν προβλέψεις σχετικά με μελλοντικά δεδομένα.

Τα προγνωστικά αναλυτικά στοιχεία είναι επίσης χρήσιμα για τις επιχειρήσεις για να τις

βοηθήσουν να διαχειριστούν το απόθεμα, να αναπτύξουν στρατηγικές μάρκετινγκ και να προβλέπουν πωλήσεις. Βοηθά επίσης τις επιχειρήσεις να επιβιώσουν, ειδικά εκείνες σε άκρως ανταγωνιστικούς κλάδους όπως η υγειονομική περίθαλψη και το λιανικό εμπόριο. Οι επενδυτές και οι επαγγελματίες του χρηματοοικονομικού τομέα μπορούν να αξιοποιήσουν αυτήν την τεχνολογία για να βοηθήσουν στη δημιουργία επενδυτικών χαρτοφυλακίων και στη μείωση της πιθανότητας κινδύνου.

Αυτά τα μοντέλα καθορίζουν σχέσεις, μοτίβα και δομές στα δεδομένα που μπορούν να χρησιμοποιηθούν για την εξαγωγή συμπερασμάτων σχετικά με το πώς οι αλλαγές στις υποκείμενες διαδικασίες που δημιουργούν τα δεδομένα θα αλλάξουν τα αποτελέσματα. Τα προγνωστικά μοντέλα βασίζονται σε αυτά τα περιγραφικά μοντέλα και εξετάζουν δεδομένα του παρελθόντος για να προσδιορίσουν την πιθανότητα ορισμένων μελλοντικών αποτελεσμάτων, δεδομένων των τρεχουσών συνθηκών ή ενός συνόλου αναμενόμενων μελλοντικών συνθηκών.

*Ένας βασικός τύπος προβλεπτικής αναλυτικής είναι η ανάλυση χρονοσειρών.*

## 2.2 ΤΙ ΕΙΝΑΙ Η ΧΡΟΝΟΣΕΙΡΑ

Μερικές φορές, τα δεδομένα σχετίζονται με το χρόνο και οι συγκεκριμένες προγνωστικές αναλύσεις βασίζονται στη σχέση μεταξύ του τι συμβαίνει πότε. Αυτοί οι τύποι μοντέλων αξιολογούν τις εισόδους σε συγκεκριμένες συχνότητες, όπως ημερήσιες, εβδομαδιαίες ή μηνιαίες επαναλήψεις. Στη συνέχεια, τα αναλυτικά μοντέλα αναζητούν εποχικότητα, τάσεις ή μοτίβα συμπεριφοράς με βάση το χρονοδιάγραμμα. Αυτός ο τύπος προγνωστικού μοντέλου μπορεί να είναι χρήσιμος για την πρόβλεψη πότε απαιτούνται περίοδοι αιχμής εξυπηρέτησης πελατών ή πότε θα πραγματοποιηθούν συγκεκριμένες

πωλήσεις.

Ειδικότερα , χρονοσειρά είναι μια ακολουθία { := 0,1,2,..}, t x t όπου κάθε t x εκφράζει την κατά την χρονική στιγμή κατάσταση συστήματος το οποίο εξελίσσεται στο χρόνο με ένα μη συγκεκριμένο τρόπο. Τέτοιες χρονοσειρές είναι I) Οι ημερήσιες αεροπορικές αφίξεις αφίξεις τουριστών στη χώρα μας t x t=1,2,…

II) Ο αριθμός t x πελατών μέσα σε ένα πολυκατάστημα κατά τη χρονική στιγμή t με [0 E ,T].

III) Η αγορά εμπορευμάτων για την επιχείρηση ανάλογα με τη ζήτηση των προϊόντων της μέσα σε ένα t , t=1,2,3,…

Σε κάποια προβλήματα, ο χρόνος στον οποίο συμπεριλαμβάνουμε από τη δειγματοληψία μπορεί να μην είναι σταθερός και τότε χρειάζεται επεξεργασία της χρονοσειράς για να γίνει η ανάλυση. Για παράδειγμα οι ημερήσιες τιμές ενός χρηματιστηριακού δείκτη (π.χ. του Χρηματιστηρίου Αξιών Αθηνών) συνιστούν μια χρονοσειρά με μεταβλητό φυσικό χρόνο δειγματοληψίας, αφού υπάρχουν Σαββατοκύριακα και αργίες, χρόνοι που είναι κλειστό το χρηματιστήριο. Ο τρόπος με τον οποίο αντιμετωπίζονται οι συγκεκριμένες περιπτώσεις είναι ορίζοντας τον οικονομικό χρόνο συναλλαγών ως χρόνο αναφοράς και σταθερό χρονικό βήμα μιας οικονομικής ημέρας από Παρασκευή σε Δευτέρα και κρατάμε τον χρόνο δειγματοληψίας σταθερό.

## 2.3 ΣΤΟΙΧΕΙΑ ΧΡΟΝΟΣΕΙΡΩΝ

Στις μεθόδους χρονοσειρών , για να προβλεφθούν οι μελλοντικές τιμές μίας μεταβλητής πχ ποσότητες της ζήτησης , χρησιμοποιούνται οι τιμές της μεταβλητής όπως διαμορφώθηκαν στο παρελθόν. Ειδικότερα , προσπαθούν να αναγνωρίσουν τα πρότυπα-χαρακτηριστικά με βάση τα οποία οι τιμές εξελίχθηκαν στο παρελθόν. Οι προβλέψεις στηρίζονται στην υπόθεση ότι αυτά τα πρότυπα θα διατηρηθούν και στο μέλλον. Ανεξάρτητα με το πώς διαμορφώνεται η πρόβλεψη της μεταβλητής (με ποια μέθοδο), αυτή θεωρείται καλή όταν είναι ακριβής, δηλαδή όταν πετυχαίνει τον στόχο της. Παράλληλα , η μέθοδος πρόβλεψης θα πρέπει να ξεχωρίζει τις μεταβολές της ζήτησης που οφείλονται σε αλλαγές των συνθηκών της αγοράς από εκείνες που προκαλούνται από τυχαία ( απρόβλεπτα ) γεγονότα.

Όταν λοιπόν μια επιχείρηση προσπαθεί να προβλέψει τη ζήτηση στηριζόμενη σε ιστορικά δεδομένα , γνωρίζει ότι η μελλοντική ζήτηση θα επηρεαστεί από την τρέχουσα ζήτηση , από ενδεχόμενα ιστορικά στοιχεία αύξησης η εποχικότητας.

Η μελλοντική ζήτηση που θα προκύψει θα περιέχει πάντα ένα στοιχείο τυχαιότητα που δεν μπορεί να εξηγηθεί ούτε από την τρέχουσα ζήτηση , ούτε από ενδεχόμενα ιστορικά στοιχεία αύξησης η εποχικότητας. Γι’αυτό κάθε παρατηρούμενη ζήτηση αναλύεται σε δύο συνιστώσες: στο συστηματικό (βασικό) στοιχείο και στο τυχαίο στοιχείο.

*Παρατηρούμενη ζήτηση = Συστηματικό στοιχείο + τυχαίο στοιχείο*

Το συστηματικό στοιχείο προσδιορίζει την αναμενόμενη τιμή της ζήτησης και αποτελείται από ένα η περισσότερα στοιχεία: από τη στάθμη η το επίπεδο (level) , την τάση , την εποχικότητα και την κυκλικότητα. Το τυχαίο στοιχείο είναι το μέρος της ζήτησης που αποκλίνει από το συστηματικό στοιχείο. Μια επιχείρηση δεν μπορεί (και δεν πρέπει) να προβλέπει το τυχαίο στοιχειο.

## 2.4 ΠΡΟΒΛΗΜΑ ΧΡΟΝΟΣΕΙΡΑΣ

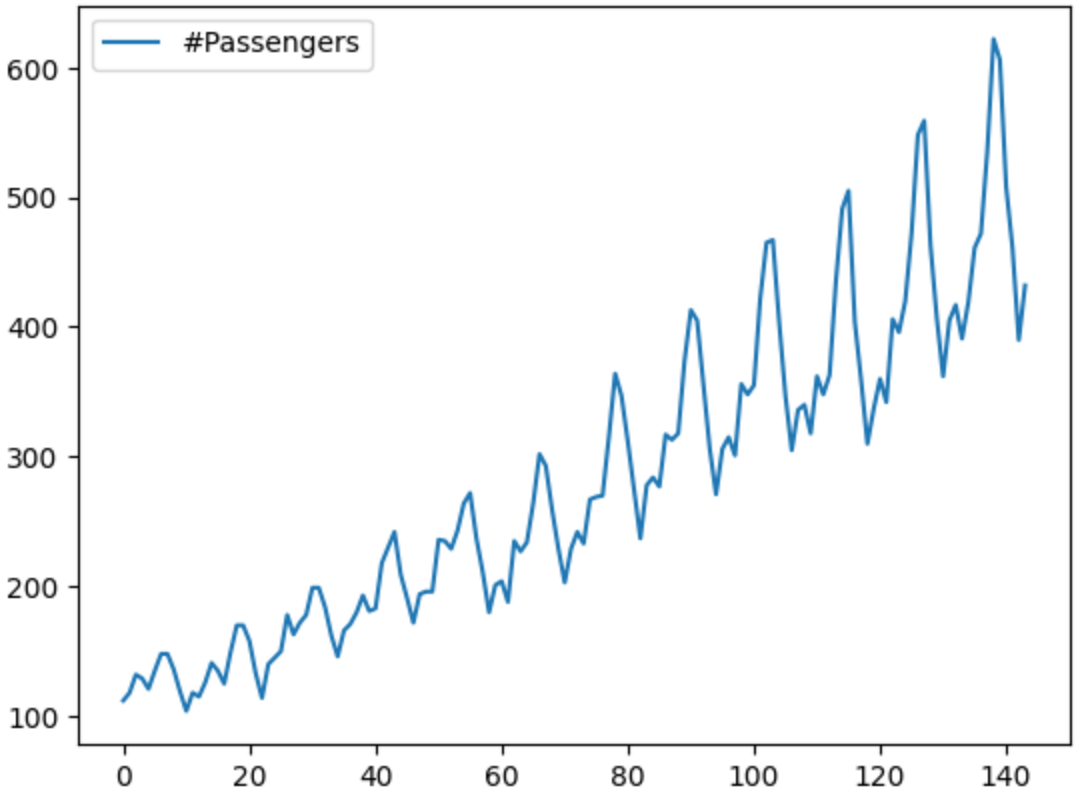
Το κύριο πρόβλημα κατά την ανάλυση με τη βοήθεια των χρονοσειρών είναι η εκτίμηση του μοντέλου που παράγει τη χρονοσειρά και ως αποτέλεσμα τις προβλέψεις μελλοντικών τιμών. Αρχικά , θα πρέπει να απορρίψουμε ότι η μεταβολή των τιμών που παρατηρούμε είναι τυχαία, δηλαδή ότι το σύστημα που έχουμε είναι λευκός θόρυβος.Αν οι τιμές της χρονοσειράς δεν είναι ανεξάρτητες , η πληροφορία που έχουμε στη χρονοσειρά μπορεί να είναι με διάφορες μορφές. Γνωρίζοντας τα παραπάνω , για να προχωρήσουμε και να προσαρμόσουμε κάποιο μοντέλο πρέπει να μελετήσουμε τα εξής:

I) Στασιμότητα : Σημαίνει ότι οι διακυμάνσεις των τιμών της χρονοσειράς δε διαφοροποιούνται με τη πέραση του χρόνου. (Brockwell, 2016)

Αντίθετα οι μη στάσιμες χρονοσειρές έχουν κάποια συγκεκριμένα χαρακτηριστικά όπως τάσεις (trends) , δηλαδή αλλαγές στη μέση τιμής με τη πέραση του χρόνου π.χ. η τιμή της βενζίνης ανάλογα με την αγορά και λόγο του πληθωρισμού.Επίσης, παρουσιάζεται η περιοδικότητα (periodicity)/εποχικότητα(seasonality) όπου γίνεται σε συγκεκριμένες περιόδους που αφορά και εποχές όπως μήνα,τρίμηνο κτλ. (Aptech.com., 2020)

Τέτοιο παράδειγμα είναι η τιμή του όζοντος , η κατανάλωση του πετρελαίου κ.α.

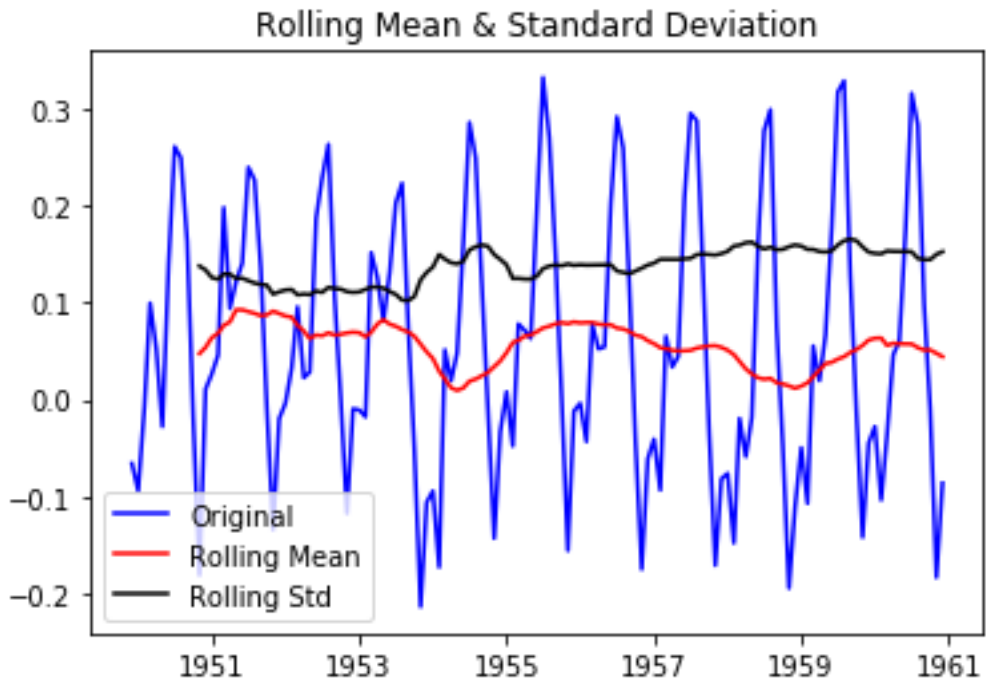
Πιο συγκεκριμένα, έχοντας πάρει ως παράδειγμα ένα πραγματικό data set το οποίο μας δείχνει τη χρονιά αλλά και τους ταξιδιώτες μέσω αεροπλάνου έχουμε το εξής :




**Εικόνα 2. 1 Ταξιδιώτες με αεροπλάνο**

Στο συγκεκριμένο διάγραμμα ο τρόπος η χρονοσειρά είναι ανοδική και είναι συνεχής , μας δείχνει ότι υπάρχει τάση για αύξηση των ταξιδιωτών κάτι το οποίο μας βοηθά να καταλάβουμε ότι δεν είναι στάσιμη.

Με διάφορους τρόπους που θα εξερευνήσουμε μεταγενέστερα , μετατρέψαμε την παραπάνω χρονοσειρά ώστε να έχει ένα μεγαλύτερο βαθμό στασιμότητας και ώς αποτέλεσμα :




**Εικόνα 2. 2 Κινούμενος μέσος και τυπική απόκλιση**

II) Αιτιοκρατία/Στοχαστικότητα : Οι χρονοσειρές περιέχουν «θόρυβο» και κατά συνέπεια όλες οι χρονοσειρές είναι στοχαστικές.

Η μεγαλύτερη πρόκληση στην ανάλυση πραγματικών χρονοσειρών είναι η διερεύνηση και ταύτιση ή εντοπισμός του αιτιοκρατικού μέρους του συστήματος που παράγει τη χρονοσειρά.

Αν για κάποιο λόγο μπορούμε να υποθέσουμε ότι το σύστημα που παράγει τη χρονοσειρά είναι κυρίως αιτιοκρατικό με κάποιες στοχαστικές διαταραχές που όμως δεν κυριαρχούν στην εξέλιξη του συστήματος (και της χρονοσειράς που μελετάμε), τότε μπορούμε να χρησιμοποιήσουμε διαφορετικές προσεγγίσεις που είναι κατάλληλες για αιτιοκρατικά δυναμικά συστήματα, π.χ. ανίχνευση κύριων περιόδων αν το σύστημα φαίνεται να είναι περιοδικό ή διερεύνηση της μη-γραμμικής δυναμικής αν το σύστημα φαίνεται να είναι χαοτικό.

Για παράδειγμα η μεταβολή τής στάθμης του όζοντος στην ατμόσφαιρα μπορεί να έχει διαφορετικές περιοδικότητες που θέλουμε να εντοπίσουμε με ακρίβεια (περίοδο έτους αλλά ίσως και άλλες περιόδους) και τότε καταφεύγουμε σε μεθόδους της φασματικής ανάλυσης. Μπορεί όμως απαλείφοντας την ετήσια εποχικότητα, να θεωρήσουμε ότι το σύστημα είναι στοχαστικό με ενδεχομένως κάποιες γραμμικές συσχετίσεις βραχείας διάρκειας που θα θέλαμε να εκτιμήσουμε και να προσαρμόσουμε κάποιο κατάλληλο στοχαστικό μοντέλο.

Τέλος μπορεί να υποθέσουμε ότι η μη-κανονικότητα της χρονοσειράς (απαλλαγμένης από την ετήσια εποχικότητα) οφείλεται σε κάποιο μη-γραμμικό αιτιοκρατικό δυναμικό σύστημα, ενδεχομένως χαμηλής διάστασης και χαοτικό, που έχει τη δυνατότητα να παρουσιάζει φαινομενικά τυχαία συμπεριφορά.

III) Γραμμικότητα / Μη Γραμμικότητα : Σύμφωνα με τα παραπάνω φαίνεται αυτές οι δύο έννοιες να σχετίζονται με την αιτιοκρατία και στοχαστικότητα αλλά γενικά μπορούν να ορισθούν ανεξάρτητα από αυτές.

Η γραμμικότητα του συστήματος σημαίνει πως οι μεταβλητές του συστήματος (που μπορεί να μην έχουμε τη δυνατότητα να τις παρατηρήσουμε) αλληλο-επιδρούν γραμμικά, δηλαδή αν θα εκφράζαμε το σύστημα με αναλυτική μορφή όλοι οι όροι θα ήταν γραμμικοί ως προς τις μεταβλητές του συστήματος. Σε αντίθετη περίπτωση το σύστημα είναι μη-γραμμικό.

Για τη χρονοσειρά αυτό σημαίνει πως για ένα γραμμικό σύστημα ορίζουμε την εξέλιξη της χρονοσειράς ως γραμμικό συνδυασμό των προηγούμενων παρατηρήσεων της χρονοσειράς, ενώ για ένα μη-γραμμικό σύστημα μπορούμε να ορίσουμε την εξέλιξη της χρονοσειράς με μεγαλύτερη ακρίβεια αν θεωρήσουμε και τη συνδυασμένη επίδραση των προηγούμενων παρατηρήσεων σε διαφορετικές χρονικές στιγμές ή τις ίδιες. ΄Αρα λοιπόν ένα στοχαστικό σύστημα μπορεί να είναι γραμμικό ή μηγραμμικό και το ίδιο ισχύει για ένα αιτιοκρατικό σύστημα.

Βέβαια ένα αιτιοκρατικό γραμμικό σύστημα δεν παρουσιάζει ιδιαίτερο ενδιαφέρον γιατί τα γραμμικά αιτιοκρατικά δυναμικά συστήματα έχουν απλές λύσεις που στην απουσία θορύβου μπορούμε εύκολα να εντοπίσουμε (σταθερό σημείο, περιοδικά σημεία ή τροχιές). Εδώ σημειώνεται ότι κάποια δυσκολία μπορεί να παρουσιαστεί αν το σύστημα είναι πολλών διαστάσεων, υπάρχει κάποια τυχαία διαταραχή και το πλήθος των παρατηρήσεων είναι σχετικά μικρό. Από την άλλη μεριά, είναι ιδιαίτερα δύσκολο να εντοπίσουμε μη-γραμμικότητα σε ένα στοχαστικό σύστημα (ή διαδικασία όπως συνήθως λέγεται) αφού ο θόρυβος στο σύστημα δεν επιτρέπει τον εντοπισμό πολύπλοκων μη-γραμμικών σχέσεων. Σε μια τέτοια περίπτωση θα πρέπει να έχουμε ορίσει μια συγκεκριμένη μηγραμμική μορφή που θέλουμε να διερευνήσουμε. Συνήθως λοιπόν οι δύο κυρίαρχες κλάσεις συστημάτων που υποθέτουμε για στάσιμες χρονοσειρές είναι η

γραμμική στοχαστική διαδικασία (linear stochastic process) και το μη-γραμμικό δυναμικό (πιθανώς χαοτικό) σύστημα.

΄Οταν έχουμε τη δυνατότητα ταυτόχρονης παρατήρησης πολλών μεγεθών για το ίδιο σύστημα, όπως π.χ. καταγραφές σεισμικών κυμάτων από διαφορετικούς σταθμούς ή καταγραφή θερμοκρασίας και πίεσης, έχουμε πολλαπλές ταυτόχρονες χρονοσειρές ή αλλιώς έχουμε μια πολυ-διάστατη χρονοσειρά (multivariate time series).

## 2.4 ΣΧΕΤΙΚΕΣ ΕΝΝΟΙΕΣ ΚΑΙ ΜΟΝΤΕΛΑ ΣΤΗΝ ΑΝΑΛΥΣΗ ΧΡΟΝΟΣΕΙΡΩΝ

Για την κατανόηση και την ανάλυση των χρονοσειριακών δεδομένων είναι απαραίτητα κάποια μοντέλα και κάποιες έννοιες.

### ΑΠΟΣΥΝΘΕΣΗ ΧΡΟΝΟΣΕΙΡΩΝ (TIME SERIES DECOMPOSITION)

Είναι η διαδικασία διασπασμού μίας χρονοσειράς σε συνιστώσες και συνήθως τάση,εποχικότητα και υπολειπόμενο. Μέσω της διάσπασης μπορούμε να απομακρύνουμε κάποια δομικά στοιχεία και κατά συνέπεια μπορούμε να κατανοήσουμε και να προβλέψουμε πιο εύκολα με το μοντέλο μας.

Τα δεδομένα χρονοσειρών μπορούν να εμφανίσουν μια ποικιλία μοτίβων και συχνά είναι χρήσιμο να χωρίσουμε μια χρονοσειρά σε πολλά στοιχεία, καθένα από τα οποία αντιπροσωπεύει μια υποκείμενη κατηγορία προτύπων.

Όταν αποσυνθέτουμε μια χρονοσειρά σε στοιχεία, συνήθως συνδυάζουμε την τάση και τον κύκλο σε ένα ενιαίο στοιχείο κύκλου τάσης (συχνά ονομάζεται απλώς τάση για απλότητα). Επομένως, μπορούμε να σκεφτούμε μια χρονοσειρά που περιλαμβάνει τρία στοιχεία: μια συνιστώσα του κύκλου τάσης, μια εποχιακή συνιστώσα και μια συνιστώσα υπόλοιπο (που περιέχει οτιδήποτε άλλο στη χρονοσειρά). Για ορισμένες χρονολογικές σειρές (π.χ. αυτές που παρατηρούνται τουλάχιστον καθημερινά), μπορεί να υπάρχουν περισσότερες από μία εποχιακές συνιστώσες, που αντιστοιχούν στις διαφορετικές εποχιακές περιόδους.

Τέτοια παραδείγματα είναι η ρύθμιση του πληθυσμού, του πληθωρισμού κ.α. . Αντί να πάρουμε δεδομένα για τον συνολικό πληθυσμό, η για την αξία κάποιου χρηματικού ποσού θα μπορούσαμε να πάρουμε τα δεδομένα ανά άτομο/ανά χιλιάδες κ.α.

## ΠΑΡΑΜΕΤΡΙΚΆ ΜΟΝΤΕΛΑ ΧΡΟΝΟΣΕΙΡΩΝ

Αφορά τη χρήση μαθηματικών μοντέλων για τη περιγραφή τωμ χρονοσειριακών δεδομένω όπως τα μοντέλα ARIMA , AR , MA , SARIMA

κ.α.

Τα μοντέλα αυτά είναι στοχαστικά μαθηματικά μοντέλα τα οποία χρησιμοποιούνται και χρησιμεύουν για την πρόβλεψη και την εξέλιξη μεγεθών. Περιέχουν το τυχαίο παράγοντα σφάλμα πρόβλεψης , τιμές του μεγέθους που εμφανίστηκαν προηγουμένως και στοχαστικούς παράγοντες.

Για την εφαρμογή τους , η σειρά όπως έχουμε και αναφέρει προηγουμένως πρέπει να είναι διακριτή , στάσιμη οπότε τα χαρακτηριστηκά της δεν εξαρτώνται από τον χρόνο και μπορούμε να τη μελετήσουμε στοχαστικά.

Γενικά , τα συγκεκριμένα μοντέλα είναι χρήσιμα για την εξαγωγή βραχυπρόθεσμων προβλέψεων και όχι τόσο για μακροχρόνιες προβλέψεις.Αυτό συμβαίνει λόγο του ότι είναι ένας γραμμικός συνδυασμός των παλαιών τιμών της χρονοσειράς. Αν για παράδειγμα θελήσουμε να προβλέψουμε κάποια τιμή x την στιγμή τ θα πρέπει να γνωρίζουμε τις τιμές χ σε τ-1, χ σε τ-2 και ούτω καθεξής. Όταν όμως υπολογίσουμε την τιμή Χ την στιγμή τ, μια υποτιθέμενη μελλοντική πρόβλεψη της τιμής Χ σε ένα χρόνο τ+1 θα βασιστεί στην τιμή που έχουμε προβλέψει . Μία τιμή που μπορεί να έχει απόκλιση από την πραγματικότητα και κατά συνέπεια, μια τιμή που θα μεγαλώσει τις αποκλίσεις των πραγματικών τιμών με τις προβλεπόμενες τιμές.

Η 7-Eleven Ιαπωνίας έχει εκμεταλλευτε΄τις βραχυχρόνιεςπ ροβλέψεις για να βελτιώσει

την απόδοση της. Η επιχείρηση έχει καθιερώσει μια διαδικασία ανεφοδιασμού που της επιτρέπει να ανταποκρίνεται στις παράγγελίες σε διάστημα λίγων ωρών.Για παράδειγμα εάν ο διευθυντής του καταστήματος υποβάλει παραγγελία στις 10.π.μ , αυτή θα παραδοθεί στις 7 της ίδιας ημέρας. Έτσι ο διευθυντής πρέπει να προβλέψει τι θα πωληθει σε μία νύχτα , δηλαδή σε λιγότερο από 12 ώρες πριν την πραγματική πώληση. Σε αυτή τη περίπτωση , η πρόβλεψη ίσως να είναι ακριβέστερη από αυτήν που θα έκανε ο διευθυντής αν χρειαζόταν να προβλέψει τη ζήτηση ολόκληρης εβδομάδας εκ των προτέρων.

# ΚΕΦΑΛΑΙΟ 3

# ΜΟΝΤΕΛΑ ARIMA ΚΑΙ ΤΑ ΕΡΓΑΛΕΙΑ ΤΟΥΣ

## Τι είναι το μοντέλο ARIMA

Αυτοπαλινδρόμενος ολοκληρωμένος κινητός μέσος όρος η μοντέλο ARIMA είναι ένα μοντέλο στατιστικής ανάλυσης που με τη χρήση των χρονοσειρών μπορεί να κατανοήσει και να προβλέψει μελλοντικές τάσεις. Είναι αυτοπαλινδρομικό εάν μπορεί να προβλέψει μελλοντικές τιμές με βάση προηγούμενες τιμές . Ένα τέτοιο παράδειγμα είναι ένα μοντέλο ARIMA το οποίο μπορεί να προβλέψει τη μελλοντική τιμή μιας μετοχής μια επιχείρησης με βάση τις παλαιότερες τιμές η τα κέρδη μίας επιχείρησης βασισμένο σε προηγούμενες τιμές. (Contreras, 2003)

Μπορούμε να κατανοήσουμε τα συγκεκριμένα μοντέλα αναλύοντας τα συστατικά του μοντέλου και το τι σημαίνουν.

*Autoregression(AR):* Αναφέρεται σε ένα μοντέλο το οποίο εμφανίζει μια μεταβαλλόμενη μεταβλητή που παλινδρομεί ουσιαστικά στον εαυτό της (προηγούμενες τιμές της).Δηλαδή υποθέτει ότι η τιμή Yt βασίζεται στις προηγούμενες τιμές Yt-1,Yt-2,…

$$\widehat{y_t} = a_1\, y_{t-1} + \cdots + a_p y_{t-p}$$

Για την μελέτη ενός τέτοιου μοντέλου είναι απαραίτητη η χρήση της συνάρτησης μερικής αυτοσυσχέτισης (PACF). Δίνει τη μερική συσχέτιση μιας στάσιμης χρονοσειράς με τις δικές της τιμές καθυστέρησης, με παλινδρόμηση των τιμών της χρονοσειράς σε όλες τις μικρότερες καθυστερήσεις.

Intergrated(I):Εάν η χρονοσειρά μετατραπεί σε στάσιμη , χρησιμοποιώντας τις πρώτες διαφορές ονομάζεται ολοκληρώσιμη πρώτης τάξης και συμβολίζεται με I(1) . Αντίστοιχα σε άλλη περίπτωση, χρησιμοποιώντας τις δεύτερες διαφορές συμβολίζεται με I(20 είναι ολοκληρώσημη δεύτερης τάξης και ούτως καθεξής.

*Κινητός Μέσος Όρος(MA):*Ενσωματώνει την εξάρτηση μεταξύ μιας παρατήρησης και ενός υπολειπόμενου σφάλματος από ένα μοντέλο κινητού μέσου όρου που εφαρμόζεται σε παρατηρήσεις με καθυστέρηση.

$$\widehat{y_t} = \varepsilon_t + \beta_1\, \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q}$$

Το μοντέλο MA υποθέτει ότι η τρέχουσα τιμή Yt εξαρτάται από τους όρους σφάλματος, συμπεριλαμβανομένου του τρέχοντος σφάλματος ( Et , Et−1,Et−2,Et−3,...). Επειδή οι όροι σφάλματος είναι τυχαίοι, δεν υπάρχει γραμμική σχέση μεταξύ της τρέχουσας τιμής και των όρων σφάλματος.

Για την μελέτη του συγκεκριμένου στοιχείου είναι σημαντικός ο συντελεστής συσχέτισης. Ένας συντελεστής συσχέτισης 1 υποδεικνύει μια τέλεια θετική σχέση, ενώ ένας συντελεστής συσχέτισης -1 δείχνει μια τέλεια αρνητική σχέση. Ένας συντελεστής συσχέτισης 0 υποδεικνύει καμία σχέση μεταξύ των δύο μεταβλητών.

Το PCF και PACF υποθέτουν στασιμότητα. Η στασιμότητα μπορεί να ελεγχθεί με τον έλεγχο Dickey-Fuller.

## Test DICKEY-FULLER

Το τεστ Augmented Dickey-Fuller ανήκει σε μια κατηγορία δοκιμών που ονομάζεται «Unit Root Test», η οποία είναι η κατάλληλη μέθοδος για τον έλεγχο της σταθερότητας μιας χρονοσειράς.

Τι σημαίνει, λοιπόν, «Unit Root»;

Η ρίζα μονάδας είναι ένα χαρακτηριστικό μιας χρονοσειράς που την καθιστά μη στάσιμη. Από τεχνική άποψη, μια μοναδιαία ρίζα λέγεται ότι υπάρχει σε μια χρονοσειρά της τιμής του άλφα = 1 στην παρακάτω εξίσωση. (Dickey, 1997)

$$Y_t = aY_{t-1} + \beta X_e + \epsilon$$

όπου το Yt είναι η τιμή της χρονοσειράς τη χρονική στιγμή «t» και το Xe είναι μια εξωγενής μεταβλητή (μια ξεχωριστή επεξηγηματική μεταβλητή, η οποία είναι επίσης μια χρονοσειρά).

Τι σημαίνει αυτό για εμάς;

Η παρουσία μιας μοναδιαίας ρίζας σημαίνει ότι η χρονοσειρά είναι μη στάσιμη. Επιπλέον, ο αριθμός των μοναδιαίων ριζών που περιέχονται στη σειρά αντιστοιχεί στον

αριθμό των διαφοροποιημένων πράξεων που απαιτούνται για να γίνει η σειρά στάσιμη .

Το τεστ Dickey-Fuller (DF) αναπτύχθηκε και διαδόθηκε από τους Dickey και Fuller (1979). Η μηδενική υπόθεση του τεστ DF είναι ότι υπάρχει μια μοναδιαία ρίζα σε ένα μοντέλο AR που και σημαίνει ότι η σειρά δεδομένων δεν είναι σταθερή. Η εναλλακτική υπόθεση είναι γενικά η σταθερότητα ή η σταθερότητα τάσης, αλλά μπορεί να είναι διαφορετική ανάλογα με την έκδοση του τεστ που χρησιμοποιείται.

Συνδυάζοντας το με της μοναδιαίες ρίζες , είναι μια δοκιμή μοναδιαίας ρίζας που ελέγχει τη μηδενική υπόθεση ότι α=1 στην παρακάτω εξίσωση μοντέλου όπου άλφα είναι ο συντελεστής της πρώτης καθυστέρησης στο Y. (Kwiatkowski, 1992)

Μηδενική υπόθεση (H0): άλφα=1

$$y_t = c + \beta t + a y_{t-1} + \varphi \Delta Y_{t-1} + e_t$$

Όπου y(t-1) = υστέρηση 1 χρονοσειρών και

δέλτα Y(t-1) = πρώτη διαφορά της σειράς τη χρονική στιγμή (t-1)

Έχει παρόμοια μηδενική υπόθεση με τη δοκιμή μοναδιαίας ρίζας. Δηλαδή, ο συντελεστής του Y(t-1) είναι 1, υποδηλώνοντας την παρουσία μοναδιαίας ρίζας. Εάν δεν απορριφθεί, η σειρά θεωρείται ότι δεν είναι σταθερή.

Το τεστ Augmented Dickey-Fuller εξελίχθηκε με βάση την παραπάνω εξίσωση και είναι μια από τις πιο κοινές μορφές δοκιμής Unit Root.

Όπως υποδηλώνει το όνομα, η δοκιμή ADF είναι μια «επαυξημένη» έκδοση της δοκιμής Dickey Fuller.

Η δοκιμή ADF επεκτείνει την εξίσωση δοκιμής Dickey-Fuller για να συμπεριλάβει τη διαδικασία παλινδρόμησης υψηλής τάξης στο μοντέλο.

$$y_t = c + \beta t + \alpha y_{t-1} + \varphi_1 \Delta Y_{t-1} + \varphi_2 \Delta Y_{t-2} .. + \varphi_P \Delta Y_{\tau-\rho} + e_t$$

Αν παρατηρήσετε, προσθέσαμε μόνο περισσότερους διαφορετικούς όρους, ενώ η υπόλοιπη εξίσωση παραμένει η ίδια. Αυτό προσθέτει περισσότερη πληρότητα στη δοκιμή.

Ωστόσο, η μηδενική υπόθεση εξακολουθεί να είναι η ίδια με το τεστ Dickey Fuller.

Ένα βασικό σημείο που πρέπει να θυμάστε εδώ είναι: Εφόσον η μηδενική υπόθεση προϋποθέτει την παρουσία μοναδιαίας ρίζας, δηλαδή α=1, η τιμή p που λαμβάνεται θα πρέπει να είναι μικρότερη από το επίπεδο σημαντικότητας (ας πούμε 0,05) προκειμένου να απορριφθεί η μηδενική υπόθεση. Έτσι, συνάγοντας ότι η σειρά είναι ακίνητη.

Ωστόσο, αυτό είναι ένα πολύ συνηθισμένο λάθος που διαπράττουν οι αναλυτές με αυτό το τεστ. Δηλαδή, εάν η τιμή p είναι μικρότερη από το επίπεδο σημαντικότητας, οι άνθρωποι θεωρούν εσφαλμένα τη σειρά ως μη στάσιμη.

## Μοντέλο SARIMA

Μια σημαντική αναφορά που πρέπει να γίνει αφορά το μοντέλο SARIMA . Το συγκεκριμένο μοντέλο συνδυάζει τις έννοιες των αυτοσυσχετίσεων , κινούμενων μέσων και της εποχικής φύσης δεδομένων. Μοιάζει με το κλασικό μοντέλο ARIMA , προσθέτει όμως ένα βασικό στοιχείο το οποίο είναι σημαντικό σε αρκετές περιπτώσεις ,την εποχικότητα.Είναι κατάλληλο για τις χρονοσειρές που παρουσιάζουν κυκλικότητα στα δεδομένα τους. (Xinghua, 2012)

## Παράμετροι ARIMA

Για τα μοντέλα ARIMA, ένας τυπικός συμβολισμός θα ήταν ARIMA με p, d και q, όπου οι ακέραιες τιμές υποκαθιστούν τις παραμέτρους που υποδεικνύουν τον τύπο του μοντέλου ARIMA που χρησιμοποιείται. Οι παράμετροι μπορούν να οριστούν ως εξής:

I) p: Ο αριθμός παρατηρήσεων lag στο μοντέλο γνωστός και ώς σειρά υστέρησης.

II) d: Ο αριθμός των φορών που διαφοροποιούνται οι πρωτογενείς παρατηρήσεις γνωστό και ώς σειρά υστέρησης.

III) q: Το μέγεθος του κινητού μέσου γνωστό και ώς η σειρά του κινητού μέσου

## Επιπλέον παράμετροι SARIMA

Για τα μοντέλα SARIMA , πέρα από τις παραμέτρους ARIMA που υιοθετούν , υπάρχουν τέσσερις ακόμη παράμετροι οι οποιοί όμως δεν διαφοροποιούνται σε μέγαλο βαθμό από εκείνες των μοντέλων ARIMA.

I)m:Το στοιχείο που δείχνει την συχνότητα εμφάνισης της κυκλικότητας. Εάν παρατηρούμε μία κυκλικότητα για παράδειγμα ανά τρίμηνο του έτους , το m θα είναι 4.

II) P: Είναι το στοιχείο p προσαρμοσμένο στο m=4 , δηλαδή στην κυκλικότητα.

III) D: Είναι το στοιχείο d προσαρμοσμένο στο m=4 , δηλαδή στην κυκλικότητα.

III) Q:Είναι το στοιχείο q προσαρμοσμένο στο m=4 , δηλαδή στην κυκλικότητα

## Αξιολόγηση μοντέλου στη πρόβλεψη

Η αξιολόγηση του αποτελέσματος ενός μοντέλου (Κουγιουμτζή Μ) πρέπει να λάβει υπόψη και άλλα δεδομένα εκτός της ίδιας της πρόβλεψης. Η συγκέντρωση των σφαλμάτων πρόβλεψης για ένα βέλτιστο αριθμό προβλέψεων είναι σημαντικό. Υπάρχουν διάφορες τεχνικές για την αξιολόγηση. Ένας από αυτούς είναι η τεχνική αντεπικύρωσης (crossvalidation). Η μέθοδος όμως η οποία είναι σχετικά απλή για να χρησιμοποιηθεί στις εφαρμογές μας είναι η διαχώρηση της χρονοσειράς μήκους M , σε δύο υποσύνολα. Το ένα είναι το σύνολο εκμάθησης (training set) (X1,X2,…,Xm) στο οποίο κάνουμε εκτίμηση το μοντέλο και το δεύτερο είναι το σύνολο επικύρωσης (validation set)( $X_{M_{1+1}}, X_{M_{1+2}}, \dots\dots, X_M$) στο οποίο κάνουμε τις προβλέψεις και υπολογίζουμε το σφάλμα πρόβλεψης

### Εργαλεία Πρόβλεψης

Στο τεχνολογικό τομέα , οι γλώσσες προγραμματισμού έχουν βοηθήσει την διαδικασία της πρόβλεψης σε ένα πολύ μεγάλο βαθμό. Μηχανισμοί και αυτοματισμού , έχουν οδηγήσει την ανάκτηση και επεξεργασία των δεδομένων σε ένα μεγάλο βαθμό χωρίς να αποτελεί πλέον πρόβλημα στη διαδικασία της πρόβλεψης. Ο χειρισμός δεδομένων με γλώσσες προγραμματισμού όπως της R, Python έχουν βοηθήσει ένα μεγάλο μέρος ατόμων που ασχολούνται με την ανάλυση των δεδομένων να κάνουν τη διαδικασία της απόκτησης και της επεξεργασίας χωρίς να χρειάζονται να έχουν ιδιαίτερες γνώσεις επάνω στο αντικείμενο του προγραμματισμού.

Οι βιβλιοθήκες των παραπάνω προγραμμάτων έχουν μία μεγάλη ανάπτυξη λόγω βελτιστοποίησης της υπολογιστικής ισχύς και στο μέλλον προβλέπεται μεγαλύτερη πρόοδος . Στην συνέχεια της διπλωματικής μας εργασίας , έχει γίνει η χρήση από βιβλιοθήκες της γλώσσας προγραμματισμού Python και ενδεικτικά κάποιες από αυτές είναι οι : statsmodels.tsa.statespace.sarimax ,statsmodels.graphics.tsaplots, numpy .

# ΚΕΦΑΛΑΙΟ 4

# ΕΦΑΡΜΟΓΕΣ ΤΩΝ ΜΟΝΤΕΛΩΝ ΣΤΗ ΠΡΑΞΗ

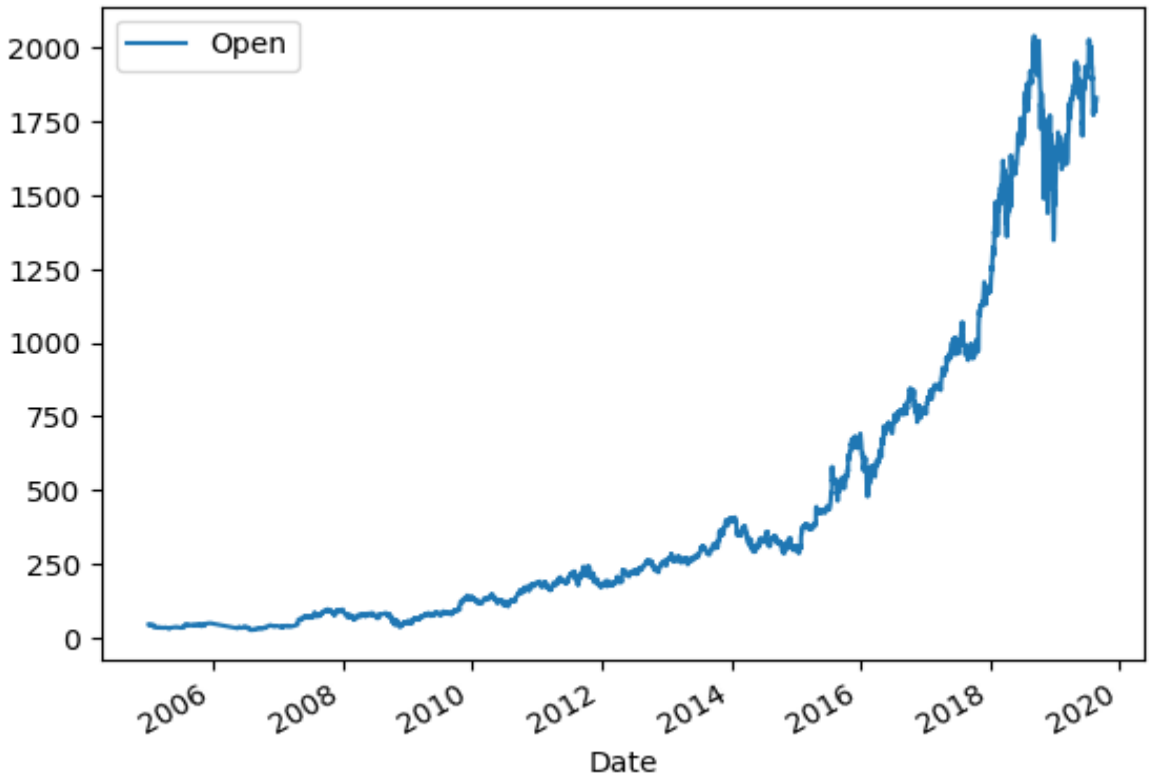
Για την διερεύνηση των συγκεκριμένων μοντέλων θα χρησιμοποιήσουμε κάποια Dataset τα οποία έχουν χειριστεί και επεξεργασθεί με το εργαλείο της Python.

## 4.1.1 Ανάλυση πρώτης χρονολογικής σειράς

Η πρώτη περίπτωση αφορά χρονολογική σειρά που αποτελείται από την τιμή ανοίγματος της μετοχής της Άμαζον τα τελευταία έτη.Παρουσιάζουμε αρχικά , κάποια βασικά στατιστικά στοιχεία από το συγκεκριμένο σύνολο δεδομένων καθώς και τη γραφική παράσταση τους.

**Πίνακας 4. 1 Περιγραφικά στατιστικά δεδομένα**

| | |
|---|---|
| Count | 3552.000000 |
| Mean | 451.961216 |
| std | 527.550229 |
| Min | 26.0900000 |
| 25% | 80.632500 |
| 50% | 230.9850000 |
| 75% | 618.1900000 |
| Max | 2038.110000 |




**Εικόνα 4. 1 Γραφική παράσταση τιμής μετοχής**

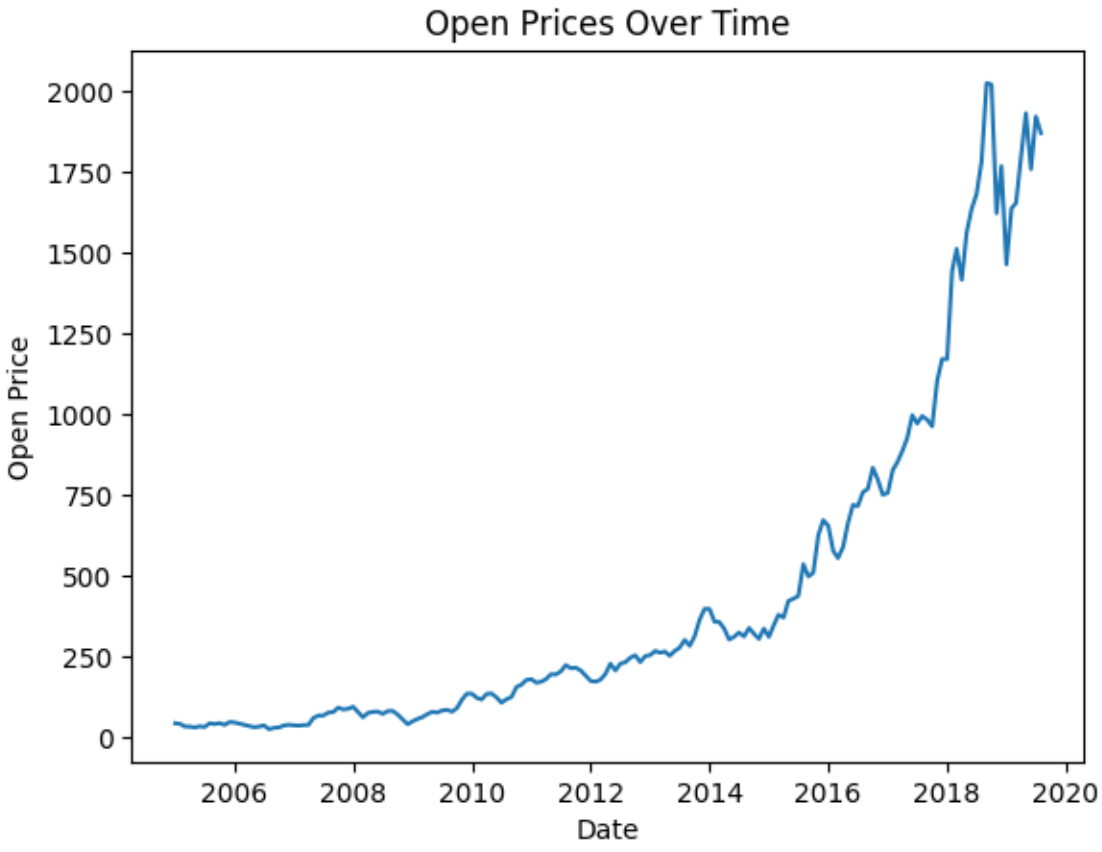
Από τη συγκεκριμένη γραφική παράσταση , υπάρχουν πολλά δεδομένα τα οποία θα μας δυσκολέψουν στη συνέχεια της εξέτασης της εφαρμογής και κάποιες μετατροπές είναι απαραίτητες. Μια από αυτές και ίσως η πιο σημαντική , είναι η μείωση των δεδομένων

μας κρατώντας μόνο την πρώτη τιμή από τον κάθε μήνα. Με αυτό το τρόπο , εμφανίζουμε αντιπροσωπευτικά δεδομένα τα οποία θα μπορούμε να διακρίνουμε και να εξετάσουμε με μεγαλύτερη ευκολία. Αυτό επιτυγχάνεται στη συγκεκριμένη περίπτωση με το εξής κομμάτι κώδικα της Python:

```
fom = df.groupby(pd.Grouper(key='Date',freq='MS')).first().
reset_index()
```

**Εικόνα 4. 2 Κομμάτι κώδικα**

Η γραφική παράσταση πλέον εμφανίζεται ως εξής:




**Εικόνα 4. 3 Διάγραμμα μετά τη παραμετροποίηση**

### 4.1.2 Χρήση Dickey-Fuller και Kwiatkowski-Phillips-Schmidt-Shin για τον έλεγχο στασιμότητας.

Αρχικά , όπως αναφέραμε και προηγουμένως , πριν την επεξεργασία και τον χειρισμό των δεδομένων θα πρέπει να εξεταστεί η στασιμότητα της εν προκειμένω χρονοσειράς. Κάτι το οποίο θα επιτευχθεί με τους ελέγχους Dickey-Fuller και Kwiatkowski-Phillips-Schmidt-Shin.

Για τους ελέγχους Dickey-Fuller/Augmented Dickey-Fuller , υπάρχει η μηδενική υπόθεση ότι η χρονοσειρά είναι μη στάσιμη. Για να απορριφθεί η μηδενική υπόθεση πρέπει το p-value να είναι μικρότερο του 0.05(p-value<0.005).

Για τον έλεγχο Kwiatkowski-Phillips-Schmidt-Shin(KPSS) , υπάρχει η μηδενική υπόθεση ότι τα δεδομένα είναι στάσιμα γύρω από μια συνιστώσα τάσης. Συνεπώς , η απόρριψη της μηδενικής υπόθεσης σε επίπεδο σημαντικότητας 5% , δείχνει ότι τα δεδομένα μας δεν παρουσιάζουν στασιμότητα.

*Τα αποτελέσματα μας είναι τα εξής:*

Πίνακας 4. 2 Αποτελέσματα Ελέγχου στασιμότητας Augmented Dickey-Fuller

| | |
|---|---|
| ADF Statistic: | 4.917598889809224 |
| P-value: | 1.0 |
| Lags Used: | 13 |
| Critical Values:<br>1%:<br>5%:<br>10%: | <br>-3.4322080557767<br>-2.8623609839255<br>-2.5672070234354 |

**Πίνακας 4. 3 Αποτελέσμα Ελέγχου στασιμότητας KPSS**

| | |
|---|---|
| KPSS Statistic: | 1.389335854535473 |
| P-value: | 0.01 |
| Lags Used: | 9 |
| Critical Values: | |

| | |
|---|---|
| 10%: | 0.347 |
| 5%: | 0.463 |
| 2.5%: | 0.574 |
| 1%: | 0.739 |

Τα συγκεκριμένα στοιχεία και ιδιαίτερα τα αποτελέσματα από τον έλεγχο Augmented Dickey-Fuller μας δείχνουν ότι δεν υπάρχουν αρκετές αποδείξεις για να απορρίψουμε την μηδενική υπόθεση. Αυτό απορρέει από το p-value το οποίο από τους υπολογισμούς μας είναι 1 , πολύ υψηλότερο από το συνηθισμένο 0.05 και κατά συνέπεια δεν είναι στάσιμη η χρονοσειρά μας.

Το επόμενο μας βήμα για να κάνουμε στάσιμη τη χρονοσειρά μας είναι η διαφοροποίηση πρώτου βαθμού.

Το αποτέλεσμα της διαφοροποίησης 1ου βαθμού είναι το εξής :


**Εικόνα 4. 4 Χρονοσειρά πριν και μετά τη διαφόριση πρώτου βαθμού**

**Εικόνα 4. 5 Χρονοσειρά πριν και μετά την διαφοροποίηση 1ου βαθμού**

Είναι προφανές πλέον ότι μετά τη διαφοροποίηση η χρονοσειρά είναι πλέον στάσιμη . Μένει μόνο πλέον να ξανατρέξουμε τον έλεγχο ADF .

**Πίνακας 4. 4 Αποτέλεσμα ελέγχου στασιμότητας Augmented-Dickey-Fuller με 1η διαφοροποίηση**

| | |
|---|---|
| ADF Statistic: | -3.5941866185322366 |
| P-value: | 0.0058731446691016615 |
| Lags Used: | 6 |
| Critical Values: | |
| 1%: | -3.4322080557767 |
| 5%: | -2.8623609839255 |
| 10%: | -2.5672070234354 |

Η τιμή του P στη συγκεκριμένη περίπτωση είναι μια εξαιρετικά μικρή τιμή, ουσιαστικά μηδενική. Αυτή η πολύ μικρή τιμή p υποδεικνύει ισχυρές ενδείξεις ενάντια στη μηδενική υπόθεση της ύπαρξης μοναδιαίας ρίζας, υποδηλώνοντας ότι η διαφοροποιημένη χρονική σειρά είναι πιθανόν ακίνητη.

Με μια τιμή p πολύ μικρότερη από το επίπεδο σημαντικότητας που χρησιμοποιείται συνήθως (π.χ. 0,05), θα απορρίψουμε συνήθως τη μηδενική υπόθεση και θα συμπεράνουμε ότι η χρονοσειρά είναι πιθανόν ακίνητη μετά τη διαφοροποίηση.

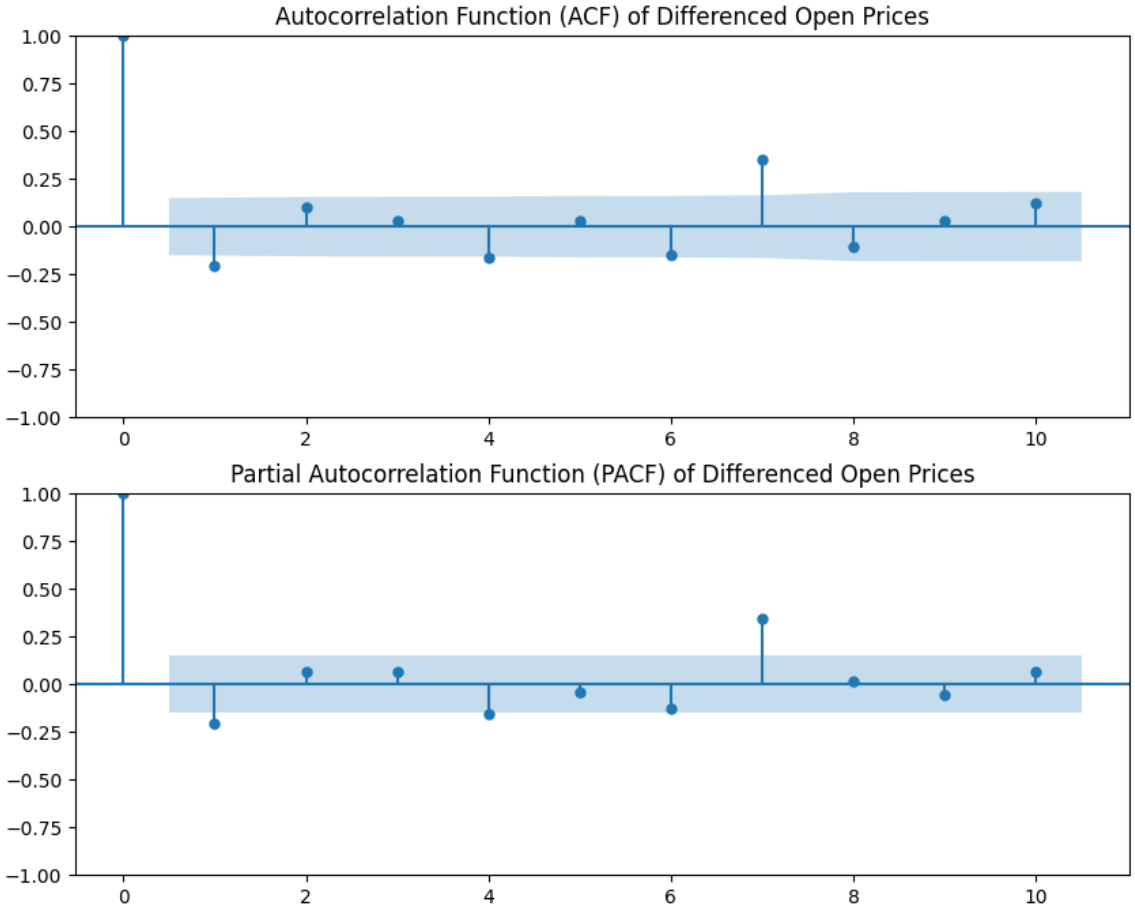
Επομένως, με βάση αυτήν την τιμή p, έχουμε πλέον ισχυρά στοιχεία που υποστηρίζουν τη σταθερότητα των διαφοροποιημένων χρονοσειρών , η οποία είναι επιθυμητή .

### 4.1.3 ΥΠΟΛΟΓΙΣΜΟΣ ΤΗΣ ΑΥΤΟΣΥΣΧΕΤΙΣΗΣ (ACF) ΚΑΙ ΤΗΣ ΜΕΡΙΚΗΣ ΑΥΤΟΣΥΣΧΕΤΙΣΗΣ (PACF)

Ο υπολογισμός των συντελεστών αυτοσυσχέτισης , για διάφορα lags μιας χρονοσειράς , βοηθά και καθορίζει την επιλογή των παραμέτρων των μοντέλων στο μέγιστο βαθμό.

Τυχαία χρονοσειρά θεωρείται η χρονοσειρά στην οποία κάθε παρατήρηση είναι ανεξάρτητη από οποιαδήποτε άλλη παρατήρηση. Σε μία τυχαία χρονοσειρά το 95% των συντελεστών αυτοσυσχέτισης βρίσκονται στο διάστημα που ορίζεται από τις τιμές

$\pm 1.96/\sqrt{n}$ όπου n είναι ο αριθμός των παρατηρήσεων. Εάν οι συντελεστές αυτοσυσχέτισης βρίσκονται εκτός των παραπάνω ορίων τότε υπάρχει συσχέτιση ανάμεσα στις παρατηρήσεις και άρα η χρονοσειρά δεν είναι τυχαία. Όσον αφορά τη στασιμότητα, για μία μη στάσιμη χρονοσειρά, οι συντελεστές αυτοσυσχέτισης είναι διάφοροι του μηδενός για αρκετές επαναλήψεις από τις πρώτες χρονικές υστερήσεις και αργά, προσεγγίζουν το μηδέν.




**Εικόνα 4. 6 Ολική και μερική αυτο συσχέτιση απο τις πρώτες διαφορές**

Στην εικόνα 4.3 , ο πρώτος όρος και ο έβδομος όρος της ACF φαίνεται να είναι οι μόνοι όροι που βρίσκονται εκτός των ορίων 5% ενώ όλοι οι υπόλοιποι φαίνεται να βρίσκονται εντώς. Λόγο του ότι ο όρος 7 είναι σχετικά μακρινός και ότι με τη κάθε υστέρηση χάνεται η γραμμική εξάρτηση των δεδομένων , δε θα θερωρηθεί στατιστικά σημαντικός.

Για τον ίδιο λόγο , θα επιλέξουμε το πρώτο όρος της PACF.

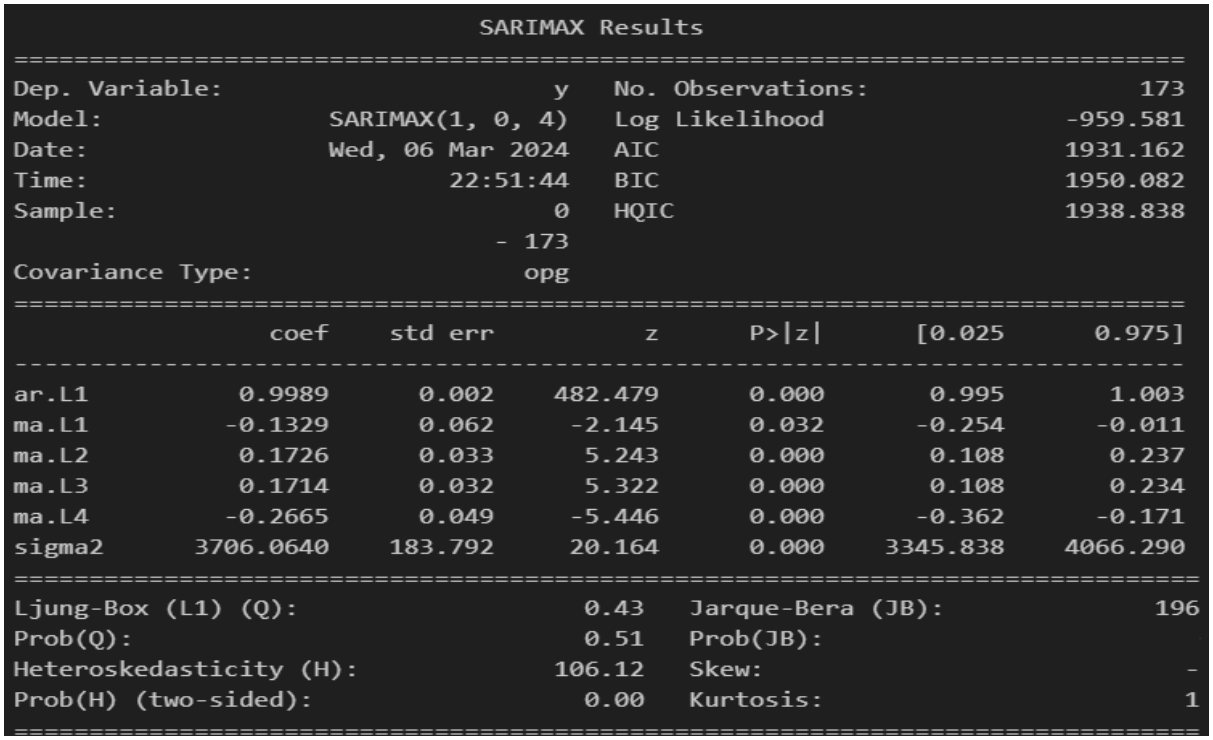
Γενικά , οι χρονοσειρές και οι προβλέψεις που αφορούν μετοχές και τιμές μετοχών είναι πολύ δύσκολο να προβλεφθούν και έχει μεγάλο ρόλο η μετοχή για την οποία πραγματοποιούμε

### 4.1.4 ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΟΥ ARIMA

Σύμφωνα με τα παραπάνω δεδομένα , δεν είναι ξεκάθαρο το ποιες παραμέτρους θα πρέπει να χρησιμοποιήσουμε ώστε να ρυθμίσουμε το μοντέλο μας με το καλύτερο δυνατό τρόπο. Ένας τρόπος για τη συγκεκριμένη περίπτωση ώστε να μπορέσουμε να βρούμε τις βέλτιστες παραμέτρους , είναι να χρησιμοποιήσουμε ένα εργαλείο της Python , τη βιβλιοθήκη pmdarima που περιέχει το auto-arima.

**Εικόνα 4. 7 Ολική και μερική συσχέτιση από τις πρώτες διαφορές.**

Το συγκεκριμένο εργαλείο θα μας βοηθήσει ώστε να επιλέξουμε τις βέλτιστες παραμέτρους αφού το ίδιο θα χρησιμοποιήσει όλους τους πιθανούς συνδυασμούς των p,d,q και θα μας δώσει το καλύτερο αποτέλεσμα.Το συγκεκριμένο αποτέλεσμα βασίζεται στον συντελεστή AIC και τη μικρότερη τιμή.



SARIMAX Results

| | | | |
|---|---|---|---|
| Dep. Variable: | y | No. Observations: | 173 |
| Model: | SARIMAX(1, 0, 4) | Log Likelihood | -959.581 |
| Date: | Wed, 06 Mar 2024 | AIC | 1931.162 |
| Time: | 22:51:44 | BIC | 1950.082 |
| Sample: | 0 | HQIC | 1938.838 |
| | - 173 | | |
| Covariance Type: | opg | | |

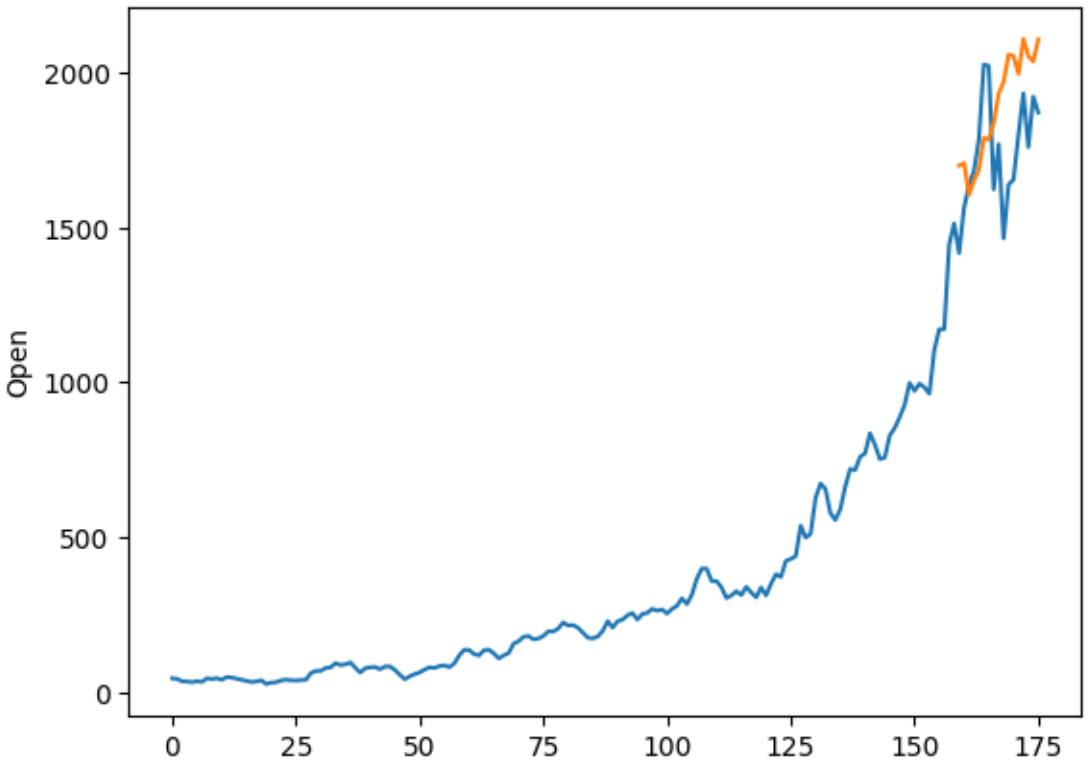
| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| ar.L1 | 0.9989 | 0.002 | 482.479 | 0.000 | 0.995 | 1.003 |
| ma.L1 | -0.1329 | 0.062 | -2.145 | 0.032 | -0.254 | -0.011 |
| ma.L2 | 0.1726 | 0.033 | 5.243 | 0.000 | 0.108 | 0.237 |
| ma.L3 | 0.1714 | 0.032 | 5.322 | 0.000 | 0.108 | 0.234 |
| ma.L4 | -0.2665 | 0.049 | -5.446 | 0.000 | -0.362 | -0.171 |
| sigma2 | 3706.0640 | 183.792 | 20.164 | 0.000 | 3345.838 | 4066.290 |

| | | | |
|---|---|---|---|
| Ljung-Box (L1) (Q): | 0.43 | Jarque-Bera (JB): | 196 |
| Prob(Q): | 0.51 | Prob(JB): | |
| Heteroskedasticity (H): | 106.12 | Skew: | - |
| Prob(H) (two-sided): | 0.00 | Kurtosis: | 1 |

**Εικόνα 4. 8 Αποτελέσματα Auto-arima**

Τα συγκεκριμένα αποτελέσματα υποδεικνύουν την χρησιμοποίηση των συντελεστών 1,0,4.

Θα χρησιμοποιήσουμε το 90% των δεδομένων και θα προσπαθήσουμε να προβλέψουμε το απομένων 10%.




**Εικόνα 4. 9 Πρόβλεψη μοντέλου ARIMA σε δεδομένα τεστ**

Γίνεται αντιληπτό ότι το συγκεκριμένο μοντέλο ARIMA που προτείνεται από το εργαλείο auto-arima έχει κάνει μια πρόβλεψη της στιγμιαίας αύξησης της τιμής. Παρόλο αυτά , μετά από το πρώτο τρίμηνο του 2018 έχει καταφέρει να αποτύχει να προσεγγίσει τις πραγματικές τιμές συνεχίζοντας την ανοδική πορεία της χρονοσειράς.

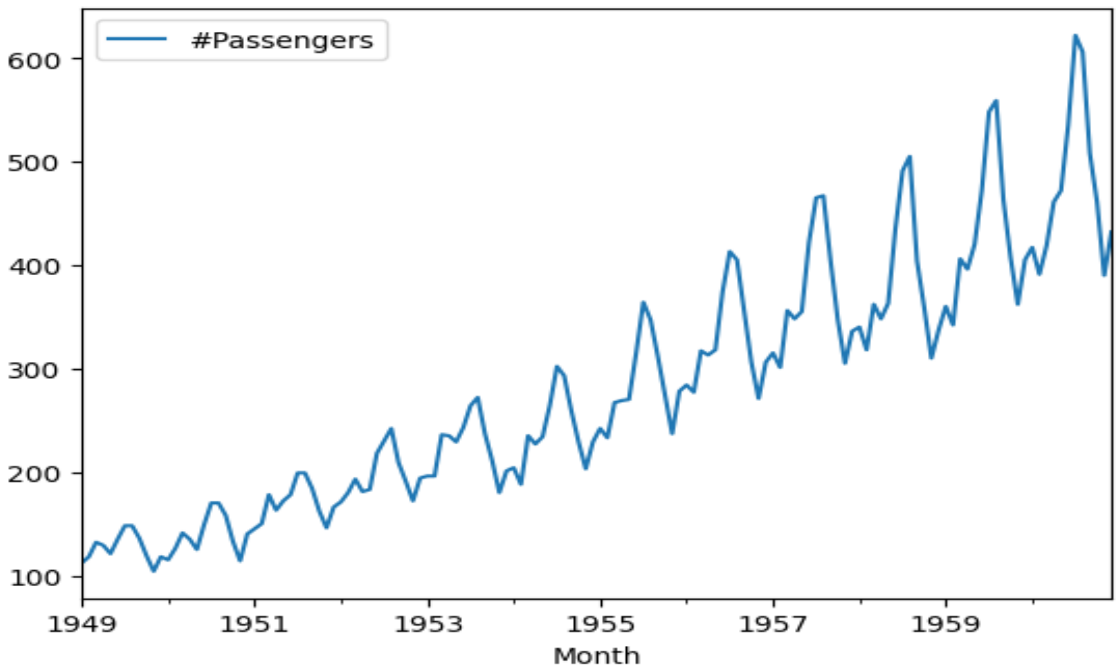
### 4.2.1 Ανάλυση δεύτερης χρονολογικής σειράς

Η δεύτερη χρονοσειρά εξέτασης αφορά δεδομένα σχετικά με το πλήθος ανθρώπων που χρησιμοποίησαν κάποιο εναέριο μεταφορικό μέσο κάθε έτος από το 1949 μέχρι το 1960. Η συγκεκριμένη χρονοσειρά  δείχνει τα συγκεκριμένα στατιστικά στοιχεία

Πίνακας 4. 5 Στατιστικά δεδομένα των δεδομένων

| | |
|---|---|
| Count | 144.000000 |
| Mean | 280.298611 |
| std | 119.966317 |
| Min | 104.000000 |
| 25% | 180.000000 |
| 50% | 265.500000 |
| 75% | 360.500000 |
| Max | 622.000000 |

Με το διάγραμμα του συνόλου των δεδομένων να απεικονίζεται ώς εξής




Εικόνα 4. 10 Διάγραμμα με ταξιδιώτες αεροπλάνων

Στο συγκεκριμένο διάγραμμα παρατηρείται άμμεσα ότι πρόκειται για μία χρονοσειρά η οποία δεν είναι στάσιμη. Είναι προφανές ότι υπάρχει μια αυξητική τάση καθώς και μία εποχικότητα ανάμεσα στα δεδομένα.Αυτή η εποχικότητα είναι κατανοητή και χωρίς την ανάγκη για τη χρησιμοποίηση κάποιου μοντέλου η κάποιου όρου. Είναι λογικό εκείνα τα χρόνια, που η τεχνολογία των εναέριων μέσων ήταν ακόμη σε αρχικό στάδιο , να μην πραγματοποιούνταν πτήσεις προς το τέλος του έτους που επικρατούν άσχημες συνθήκες.

### 4.2.2 Μετασχηματισμός και υπολογισμός πρώτης διαφόρισης.

Πραγματοποιήσαμε τη πρώτη διαφοροποίηση μέσω της Python κάνοντας την επακόλουθη διαδικασία στο Dataset (Perktold, 2009-2023).

Από το αρχικό Dataset , φτιάχνουμε ένα νέο dataframe μέσω της συνάρτησης diff() όπου:

```
NewPass= Passengers[['#Passengers']].copy(deep=True)


NewPass.head()

NewPass['firstDiff']=NewPass['#Passengers'].diff()
NewPass['Diff12']=NewPass['#Passengers'].diff(12)

NewPass.head()
```

**Εικόνα 4. 11 Διαδικασία Διαφοροποίησης μέσω της Python**

*Τα αποτελέσματα από τη συγκεκριμένη διαδικασία είναι τα εξής*

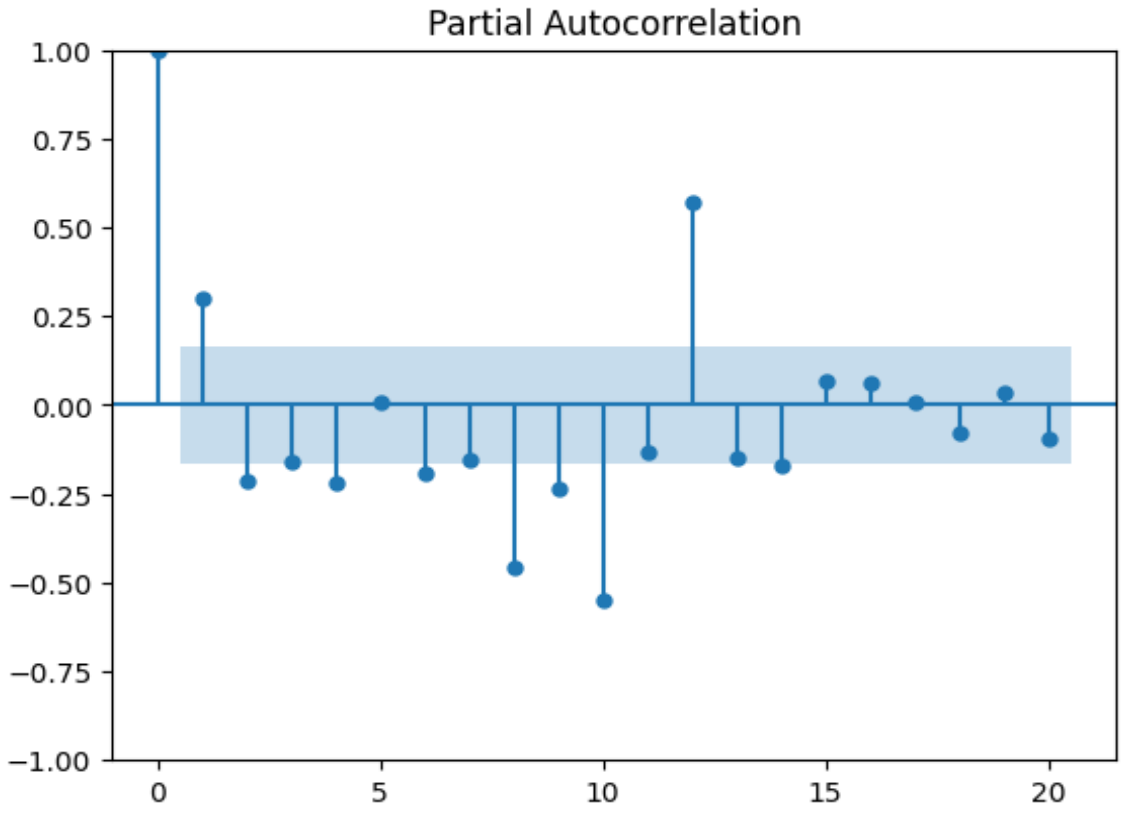
Πίνακας 4. 6 Πρώτης Διαφοροποίησης

| *Month* | *#Passengers* | *firstDiff* |
|---|---|---|
| *1949-01-01* | *112* | *NaN* |
| *1949-02-01* | *118* | *6.0* |
| *1949-03-01* | *132* | *14.0* |
| *1949-04-01* | *129* | *-3.0* |
| *1049-05-01* | *121* | *-8.0* |

Επόμενο βήμα είναι και πάλι να προχωρήσουμε στη δημιουργία διαγραμμάτων σχετικά με την ολική και μερική αυτόσυσχέτιση.

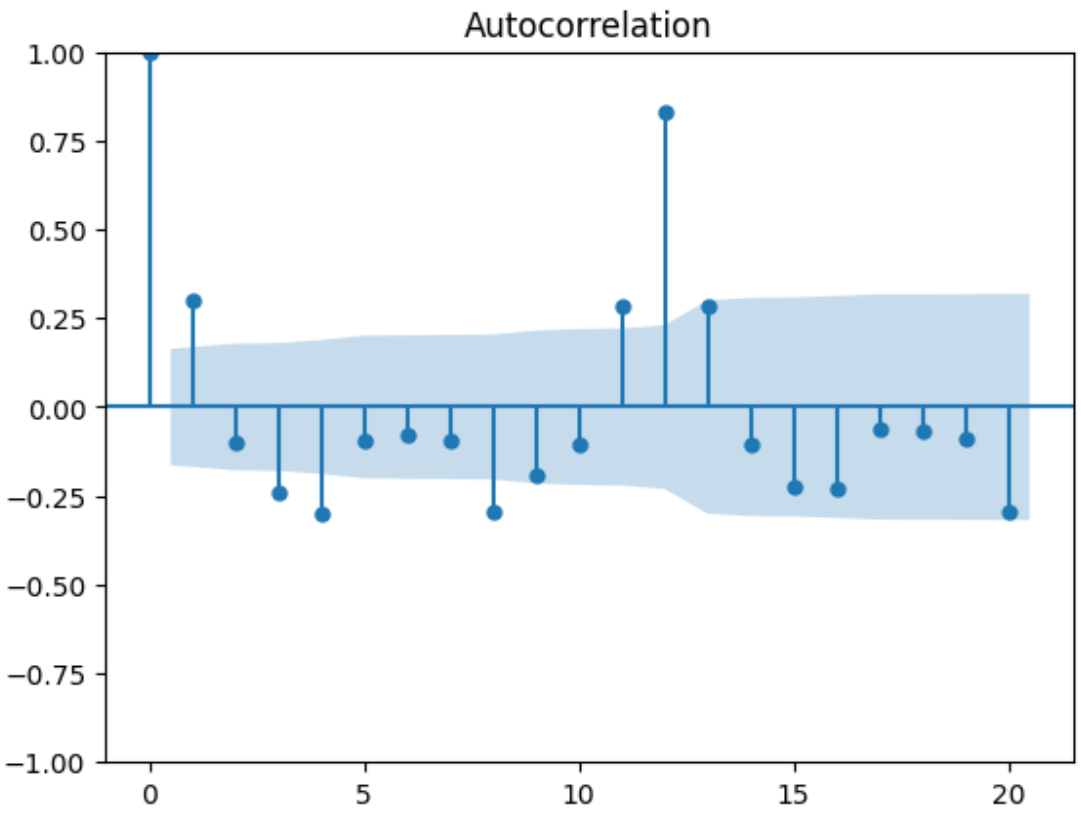
### 4.2.3 Υπολογισμός συνάρτησης ολικής και μερικής αυτοσυσχέτισης (ACF/PACF)

Για τον υπολογισμό της συνάρτησης ACF/PACF χρησιμοποιούμε τη βιβλιοθήκη της Python: statsmodels.graphics.tsaplots




**Εικόνα 4. 12 Μερικής αυτοσυσχέτιση**

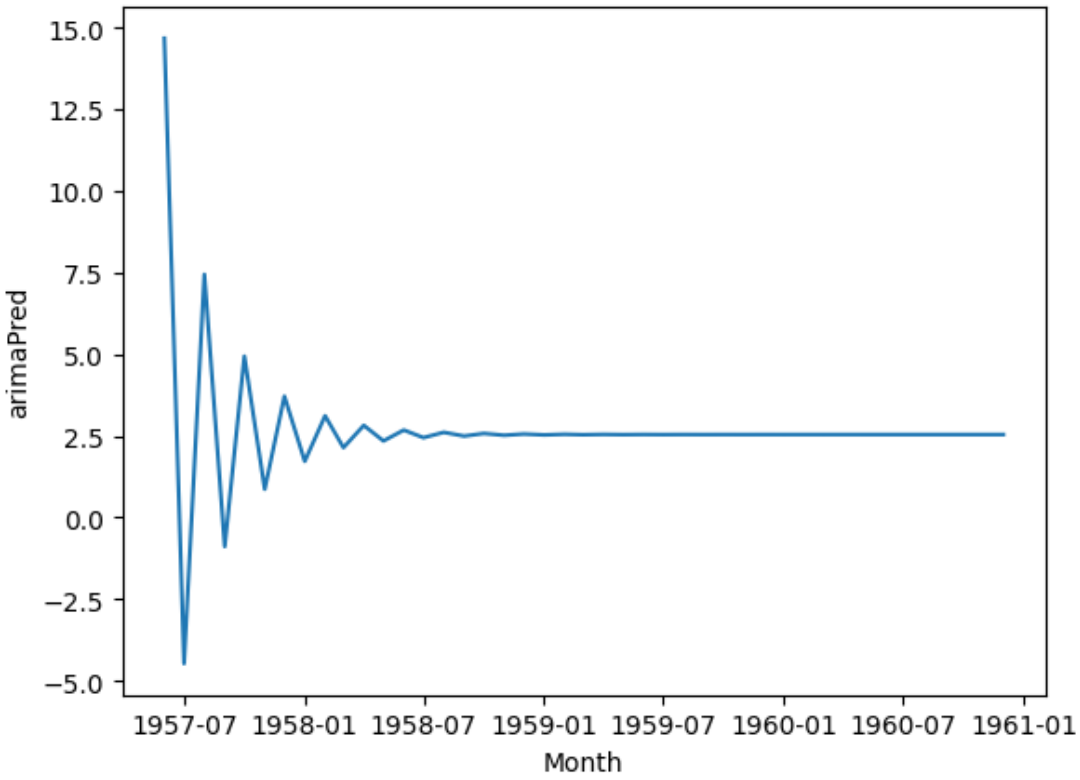
Βλέπουμε και είναι εμφανές ότι οι όροι που θα μπορούσαμε να χρησιμοποιήσουμε είναι ο 1ος ,2ος και τέταρτος όρος καθώς οι υπόλοιποι σημαντικού όροι βρίσκονται μετά το 7ο lag.



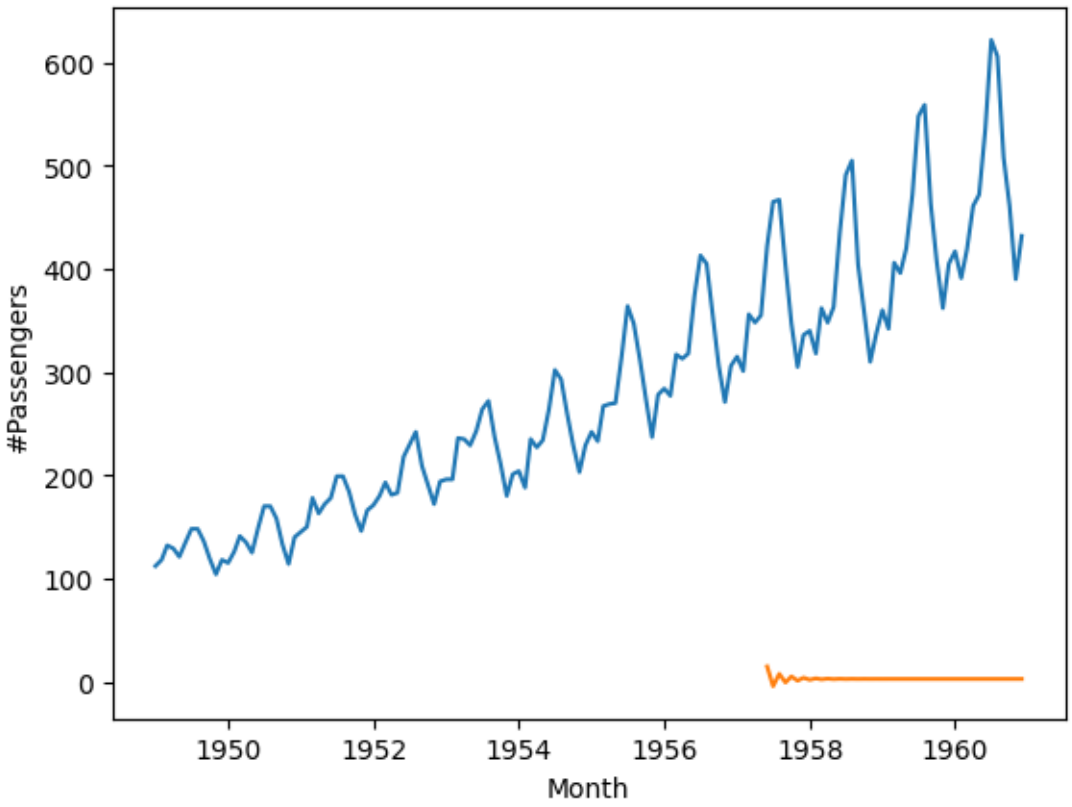

**Εικόνα 4. 13 Διάγραμμα ολικής αυτοσυσχέτισης**

Όπως είναι ξεκάθαρο και στη συγκεκριμένη περίπτωση οι μεταβλητές τις οποιές θα δοκίμασουμε να χρησιμοποιήσουμε είναι οι p=1 d=1 και q= 1 η 3 καθώς στο μέρος του AR λάβαμε από την μερική αυτοσυσχέτιση τη σημαντικότητα στο 1ο lag. Στην ολική συσχέτιση από την άλλη βλέπουμε ότι οι μετάβλητές 1 και 3 είναι οι σημαντικότερες.

Χρησιμοποιήσαμε το 70% των δεδομένων ως train δεδομένα ενώ το υπόλοιπο 30% ως τεστ για το μοντέλο ARIMA(1,1,3).




**Εικόνα 4. 14 Αποτέλεσμα ARIMA σε test δεδομένα.**



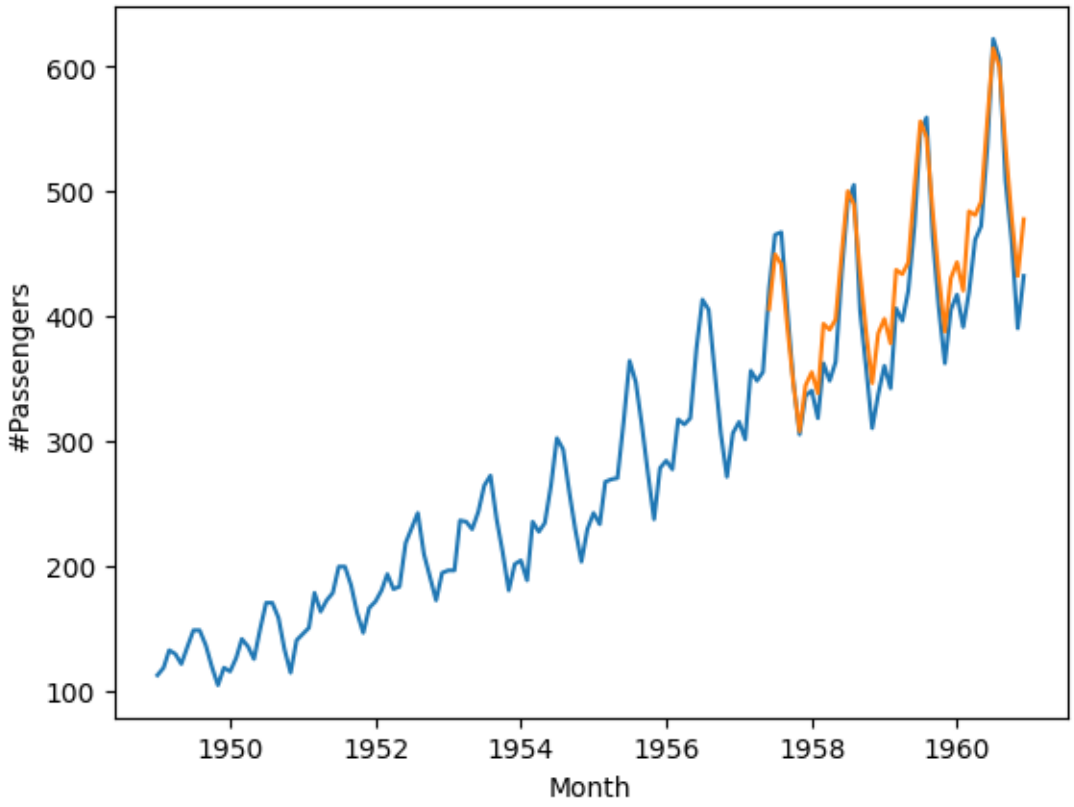

**Εικόνα 4. 15 Πρόβλεψη ARIMA σε αρχικά δεδομένα**

Όπως εμφανίζεται, η απόκλιση με τα πραγματικά δεδομένα είναι πολύ μεγάλη και το συγκεκριμένο μοντέλο ARIMA δεν μας βοήθησε αρκετά .

Υπάρχει ένα στοιχείο που θα πρέπει να λάβουμε υπόψη και κατά συνέπεια θα μας δείξει ότι η επιλογή μοντέλου εξαρτάται στο μέγιστο βαθμό στα χαρακτηριστηκά της χρονοσειράς. Αυτό είναι η εποχικότητα.

Είδαμε γενικά και στο διάγραμμα της συγκεκριμένης χρονοσειράς ότι είναι εμφανής η εποχικότητα και ότι θα χρειαστεί να τη συμπεριλάβουμε στα στοιχεία των προβλέψεων μας.

Αυτό επιτυγχάνεται με το μοντέλο SARIMA το οποίο λαμβάνει υπόψιν και το εποχιακό στοιχείο στη συνάρτηση.Στο δικό μας παράδειγμα χρησιμοποιήσαμε τη βιβλιοθήκη της Python statsmodels.tsa.statespace.sarimax και τα δεδομένα p,q,d που διαλέξαμε στο αρχικό μοντέλο.




**Εικόνα 4. 16 Αποτελέσματα μοντέλου SARIMA**

Παρατηρούμε ότι το συγκεκριμένο μοντέλο έχει πετύχει και έχει καταφέρει να κάνει μια αρκετά καλή πρόβλεψη των αρχικών τιμών που έιχε η χρονοσειρά. Είναι ασφαλές να συμπεράνουμε ότι, η εποχικότητα είναι ένα σημαντικό στοιχείο το οποίο θα πρέπει να λάβουμε υπόψη στο μοντέλο το οποίο χειριζόμαστε.

## 4.3.1 Ανάλυση τρίτης χρονολογικής σειράς

Για το συγκεκριμένο υπόδειγμα , χρησιμοποιήσαμε πραγματικά δεδομένα από Ελληνική επιχείρηση τα οποία μας δείχνουν τις αγορές της ανά ημέρα.Θα προσπαθήσουμε να προβλέψουμε τις αγορές που πρέπει να κάνει με την χρήση των χρονοσειρών.

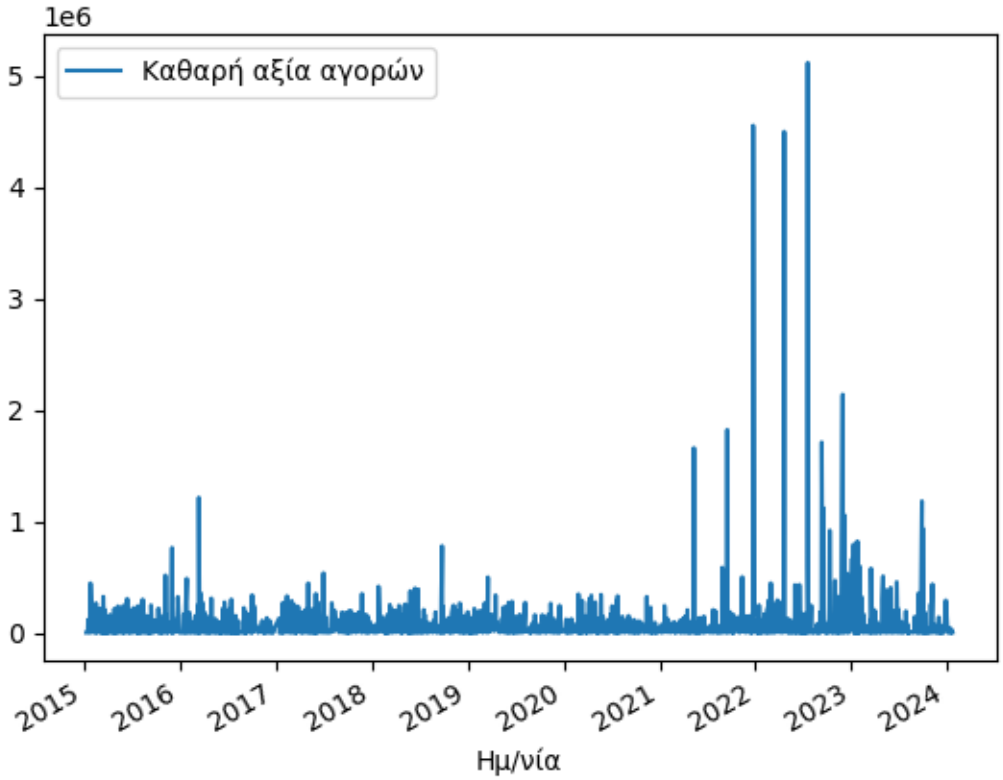
### 4.3.2 Μετασχηματισμός δεδομένων

Ένα σημαντικό κομμάτι στη προβλεπτική αναλυτική είναι ο χειρισμός των δεδομένων μας. Κατά κύριο λόγο , τα δεδομένα δεν έρχονται ακριβώς όπως τα επιθυμούμε και χρειάζεται να γίνουν κάποιες διορθώσεις σε εκείνα για να μπορέσουμε να τα διαχειριστούμε και να βγάλουμε συμπεράσματα. Στην συγκεκριμένη περίπτωση , είναι δύσκολο να χειριστούμε και να εφαρμόσουμε το μοντέλο ARIMA σε ημερήσια δεδομένα.

Στην περίπτωση μας , τα δεδομένα μας ήταν στη συγκεκριμένη μορφή σε πρώτο στάδιο:

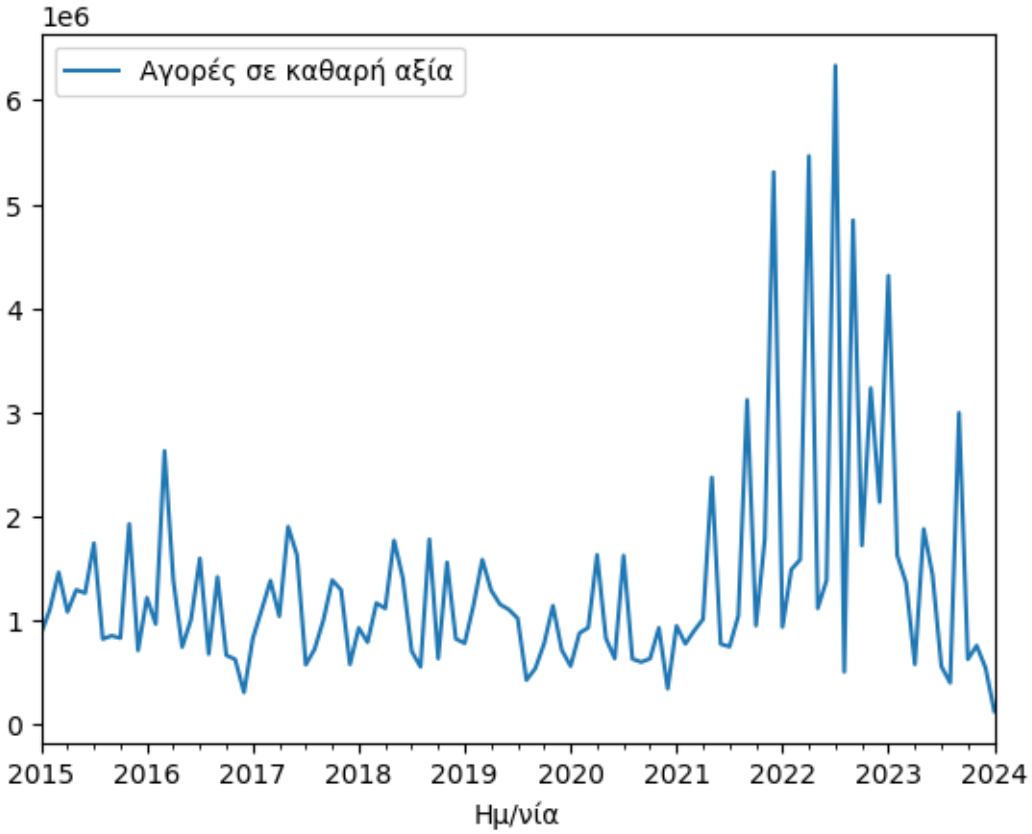
**Πίνακας 4. 7 Δεδομένα πριν την επεξεργασία**

| Ημ/νία | Καθαρή αξία σε ευρώ |
|---|---|
| 7/1/2015 | 2516,85 |
| 12/1/2015 | 4067,5 |
| 12/1/2015 | 300 |
| 13/1/2015 | 600 |
| 13/1/2015 | 1044,2 |
| 13/1/2015 | 29737,83 |
| 14/1/2015 | 118240 |
| 15/1/2015 | 206 |
| 15/1/2015 | 36260,89 |
| 15/1/2015 | 3042 |
| 16/1/2015 | 60534 |
| 19/1/2015 | 139,52 |
| 19/1/2015 | 4252,5 |
| 21/1/2015 | 1044,2 |
| 21/1/2015 | 3725,1 |
| 22/1/2015 | 1260 |
| 23/1/2015 | 141962,85 |
| 23/1/2015 | 217855,1 |
| 23/1/2015 | 36000 |




**Εικόνα 4. 17 Διάγραμμα δεδομένων πριν την επεξεργασία**

Όπως είναι φανερό δεν μας βοηθάει το συγκεκριμένο διάγραμμα. Έπειτα από την επεξεργασία των δεδομένων ώστε να εμφανίζονται σε μηνιαία βάση, έχουμε το εξής αποτέλεσμα:




**Εικόνα 4. 18 Διάγραμμα δεδομένων μετά την επεξεργασία**

Με μία πρώτη ματιά , δεν είναι ξεκάθαρο το τι συμβαίνει και αν αφορά μια στάσιμη σειρά. Επόμενο βήμα , είναι οι έλεγχοι στασιμότητας που χρησιμοποιήσαμε και στο πρώτο υπόδειγμα.

### 4.3.3 Χρήση Dickey-Fuller και Kwiatkowski-Phillips-Schmidt-Shin για τον έλεγχο στασιμότητας.

Για τους ελέγχους Dickey-Fuller/Augmented Dickey-Fuller , υπάρχει η μηδενική υπόθεση ότι η χρονοσειρά είναι μη στάσιμη. Για να απορριφθεί η μηδενική υπόθεση πρέπει το p-value να είναι μικρότερο του 0.05(p-value<0.005).

Για τον έλεγχο Kwiatkowski-Phillips-Schmidt-Shin(KPSS) , υπάρχει η μηδενική υπόθεση ότι τα δεδομένα είναι στάσιμα γύρω από μια συνιστώσα τάσης. Συνεπώς , η απόρριψη της μηδενικής υπόθεσης σε επίπεδο σημαντικότητας 5% , δείχνει ότι τα δεδομένα μας δεν παρουσιάζουν στασιμότητα.

Πίνακας 4. 8 Αποτελέσματα δεδομένων test Augmented-Dickey-Fuller

| | |
|---|---|
| ADF Statistic: | -1.9807566853210401 87 |
| P-value: | 0.295073833775119 |
| Lags Used: | 3 |
| Critical Values: | |
| 1%: | -3.4922080557767 |
| 5%: | -2.8893609839255 |
| 10%: | -2.58162070234354 |

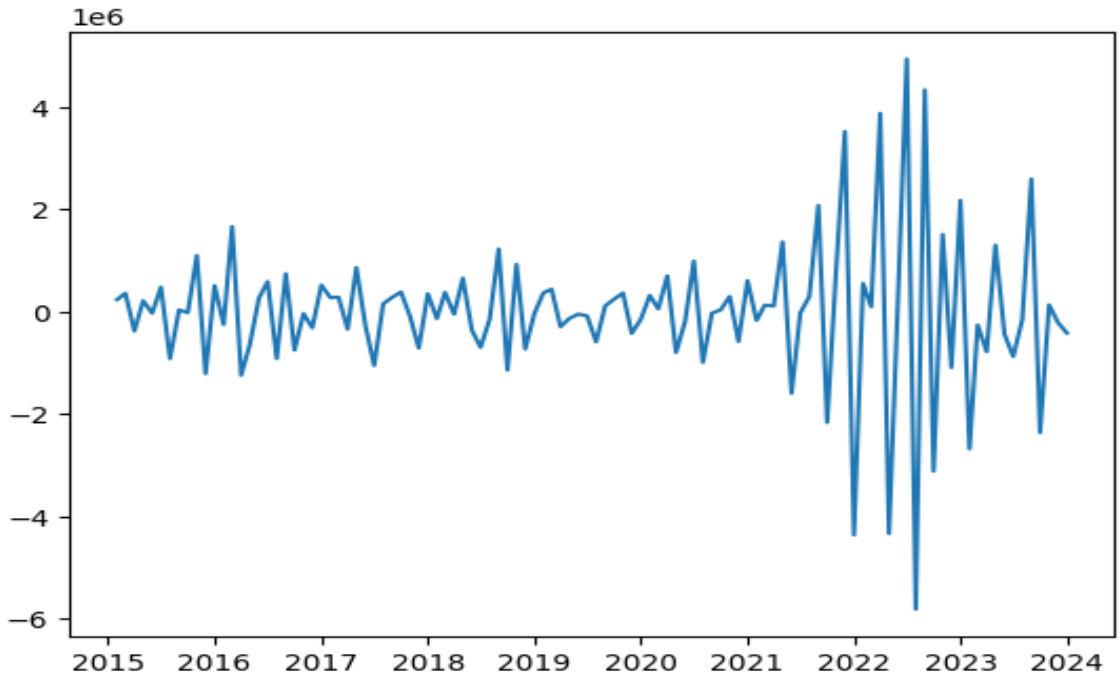
Στη συγκεκριμένη περίπτωση η τιμή P είναι πολύ μεγαλύτερη από το 0.05 και συνεπώς δεν μπορούμε να απορρίψουμε το null hypothesis και να υπιοθέσουμε στασιμότητα. Είναι εμφανές ότι θα πρέπει να προχωρήσουμε στην διαφόριση πρώτου βαθμού.

### 4.4.4 Υπολογισμός πρώτης διαφόρισης

Η διαδικασία της πρώτης διαφόρισης παραμένει η ίδια με το αρχικό υπόδειγμα. Χρησιμοποιήσαμε το εργαλείο της Python και τα αποτελέσματα είναι τα εξής:

### 4.4.5 Μέρος κώδικα για την πρώτη διαφόριση.

```
NewDF = df.diff(periods=1)
NewDF = NewDF[1:]
NewDF.head()
```




**Εικόνα 4. 19 Διάγραμμα πρώτης διαφόρισης**

Φαίνεται πλέον ότι η χρονοσειρά είναι στάσιμη , είναι όμως απαραίτητο να επιβεβαιωθεί με τη μέθοδο Augmented-Dickey-Fuller.

**Πίνακας 4. 9 Πίνακας δεδομένων Augmented-Dickey-Fuller**

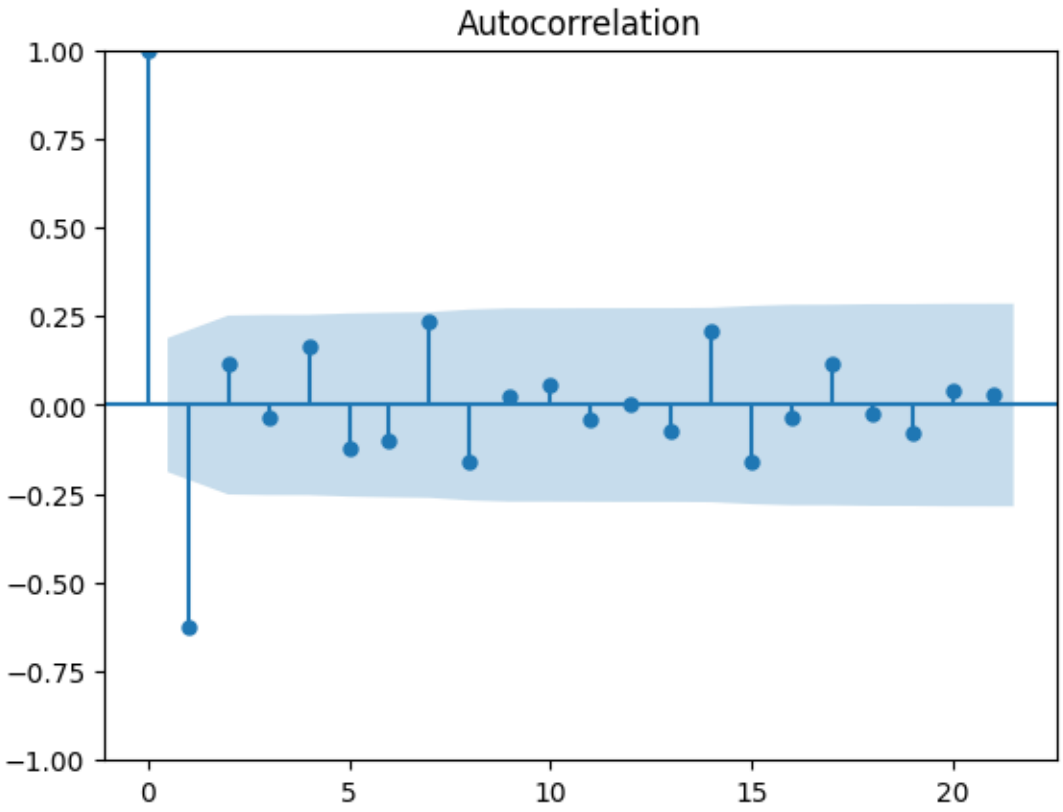
| | |
|---|---|
| ADF Statistic: | -13.20883110480247 |
| P-value: | 1.057215937690581e-24 |
| Lags Used: | 2 |
| Critical Values:<br>1%:<br>5%:<br>10%: | <br>-3.4922080557767<br>-2.8893609839255<br>-2.58162070234354 |

Πλέον , το p-value είναι πολύ κοντά στο μηδέν κάτι που φανερώνει ότι

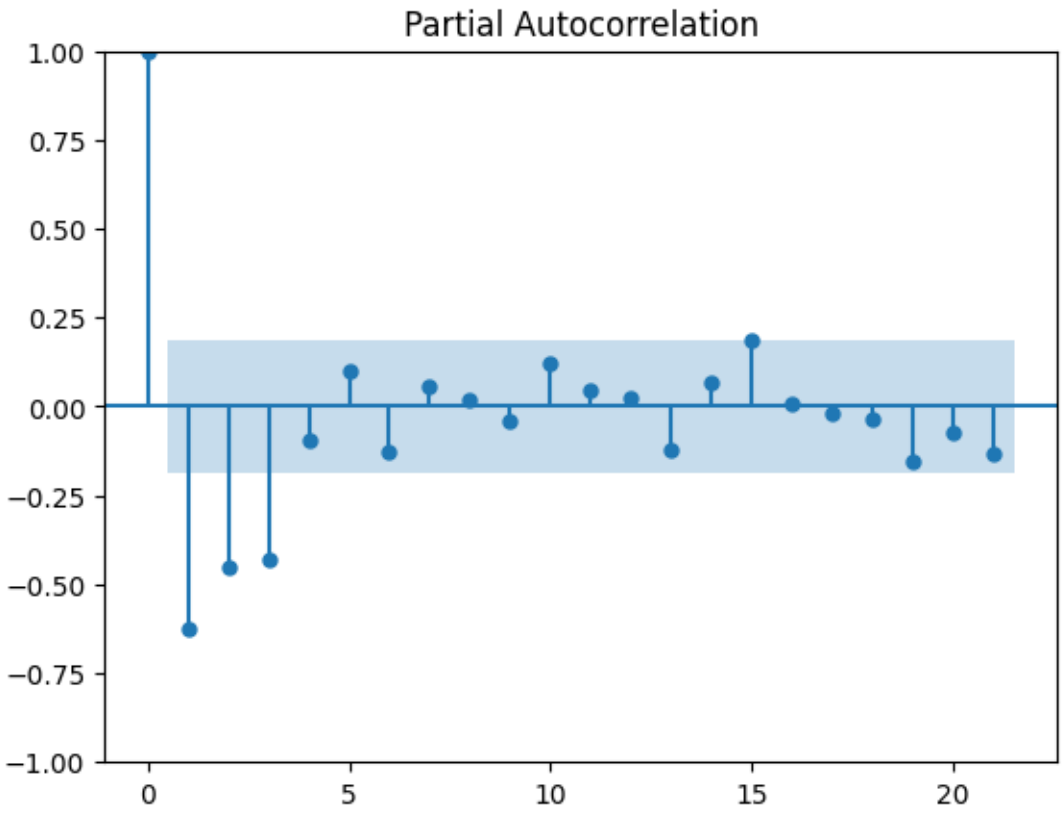
υπάρχουν στοιχεία ενάντια στην υπόθεση μηδενικής ρίζας και μπορούμε να την απορρίψουμε. Μπορούμε να υποθέσουμε πλέον ότι η χρονοσειρά είναι σταθερή.

### 4.4.5 Υπολογισμός συνάρτησης ολικής και μερικής αυτοσυσχέτισης (ACF/PACF)

Ο υπολογισμός της ολικής και μερικής αυτό συσχέστισης έχουν τα συγκεκριμένα αποτελέσματα:



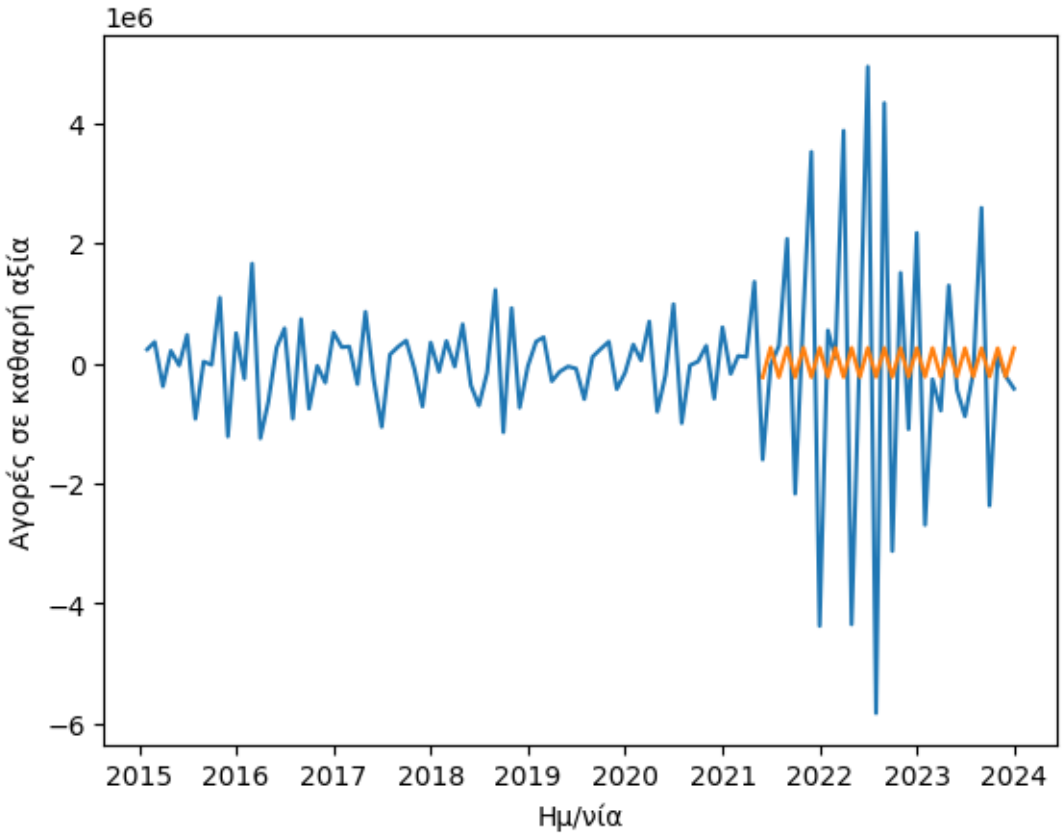

**Εικόνα 4. 20 Διάγραμμα μερικής αυτό-συσχέτισης**



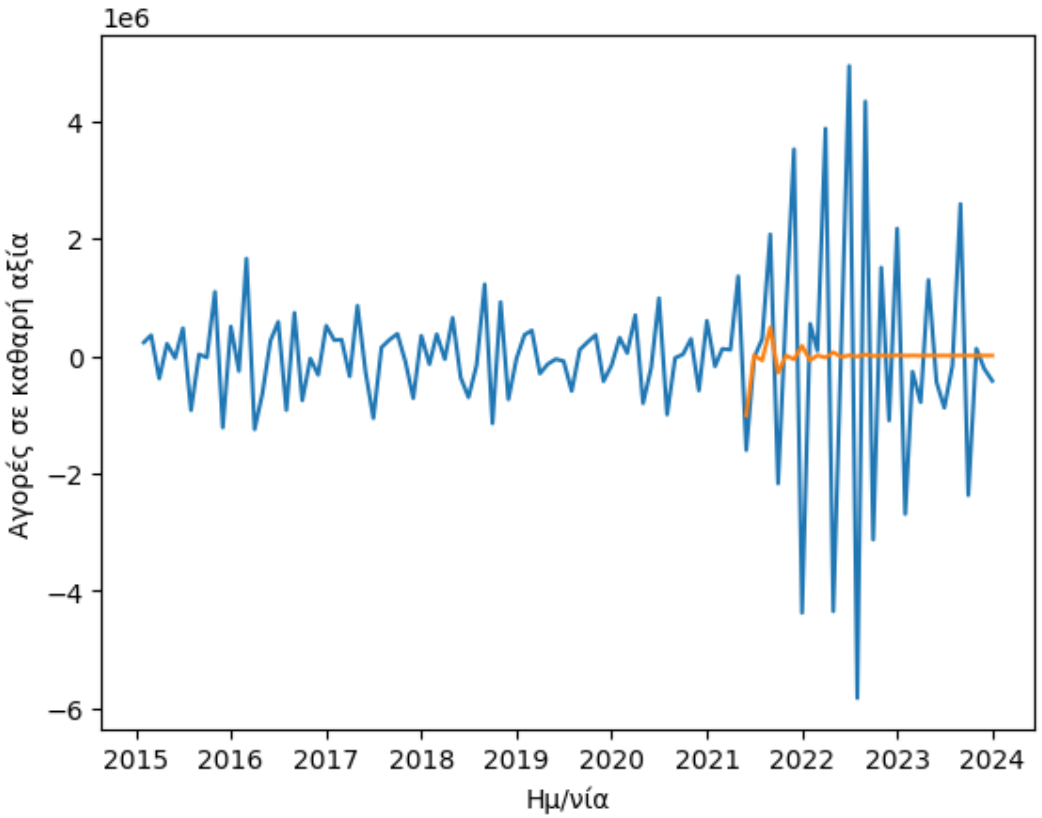

**Εικόνα 4. 21 Διάγραμμα ολικής αυτό-συσχέτισης**

Συμπεραίνουμε ότι το στη μερική αυτό συσχέτιση ο σημαντικός όρος είναι στο 1ο lag ενώ για τη μερική αυτό συσχέτιση έχουμε το 1ο,2ο και 3ο.

Χωρίζοντας τα δεδομένα σε δύο κατηγορίες (test 70%,train 30%) δοκιμάζουμε το μοντέλο ARIMA με τις παραπάνω παραμέτρους τα αποτελέσματα είναι τα εξής:

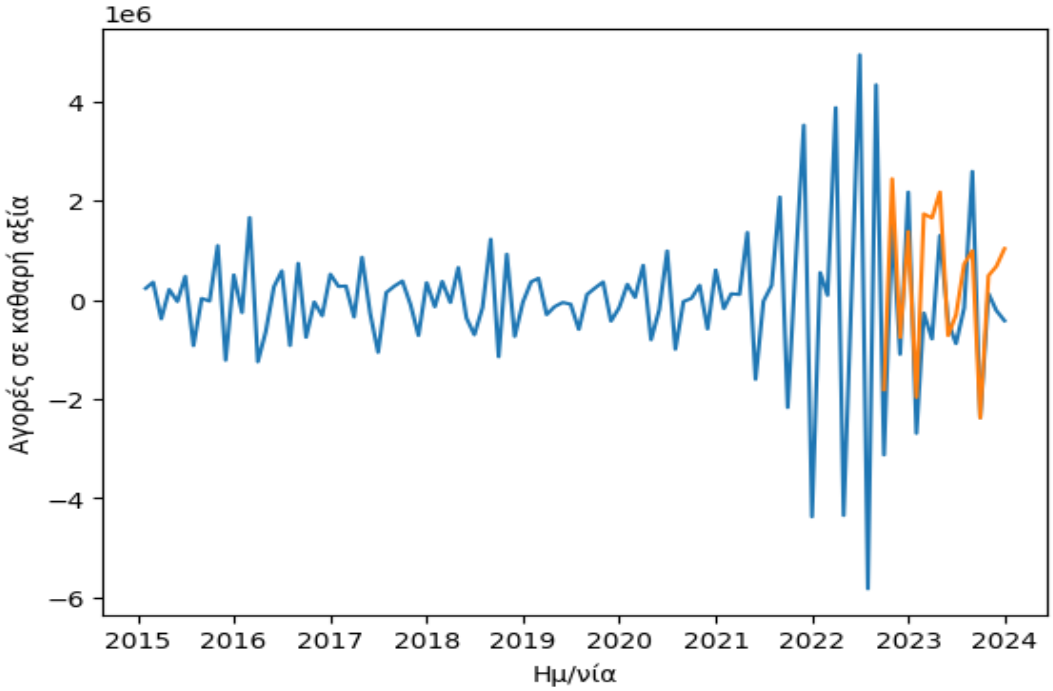



**Εικόνα 4. 22 Διάγραμμα ARIMA(1,1,1)**

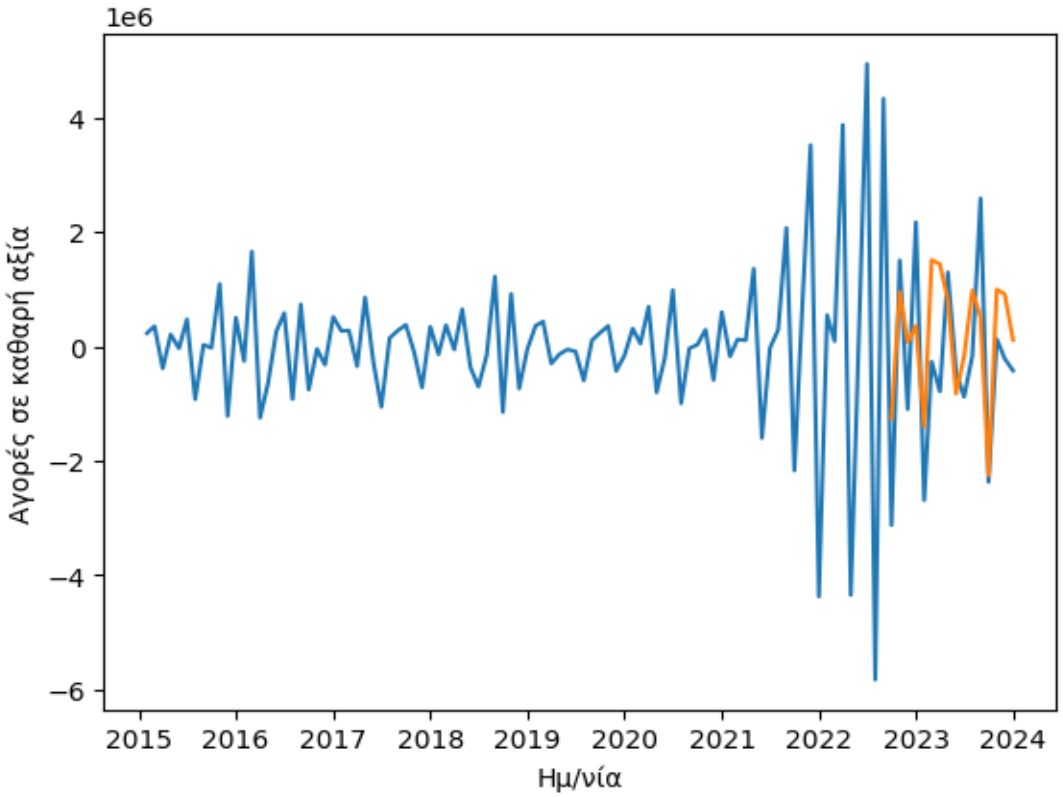



**Εικόνα 4. 23 Διάγραμμα ARIMA(2,1,1)**

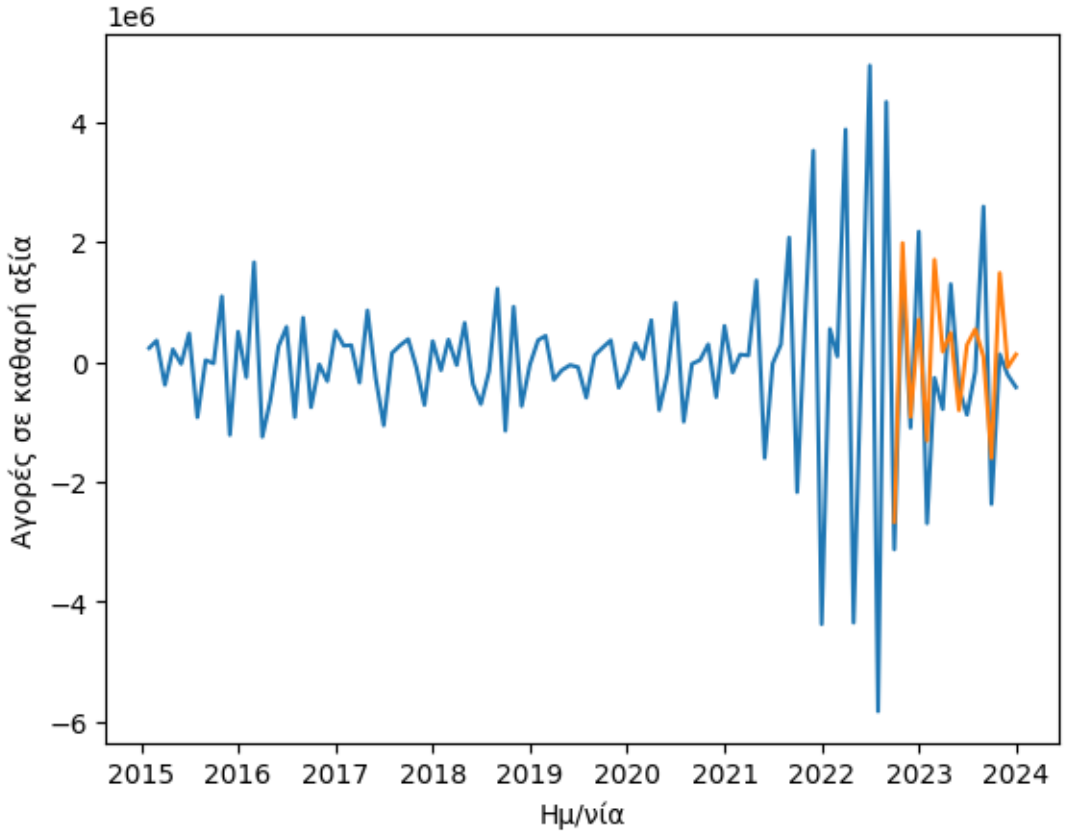
Τα αποτελέσματα του μοντέλου με τη συγκεκριμένη μορφή δεν μας βοηθάνε να κάνουμε την οποιαδήποτε πρόβλεψη. Είναι σημαντικό το γεγονός της επιλογής του σωστού test/train data. Γνωρίζουμε ότι τα μοντέλα χρονοσειρών δεν λειτουργούν σωστά σε μακροχρόνια δεδομένα και φαίνεται ξεκάθαρα στα επάνω διαγράμματα. Επόμενο βήμα είναι η ανακατανομή των δεδομένων σε 85% test / 15% train data. Το αποτέλεσμα είναι το εξής:




**Εικόνα 4. 24 Διάγραμμα ARIMA(3,1,1) με 85% test data/15% train data**




**Εικόνα 4. 25 Διάγραμμα ARIMA(2,1,1) με 85% test data/15% train data**




**Εικόνα 4. 26 Διάγραμμα ARIMA(1,1,1) με 85% test data/15% train data**

Είναι εμφανές το πόσο σημαντικό είναι η σωστή ρύθμιση των δεδομένων που λαμβάνουμε υπόψη , καθώς και ότι ο διαχωρισμός τους με βραχυπρόθεσμο ορίζοντα είναι μια μεγάλη προτεραιότητα. Όλα τα παραπάνω μοντέλα και ιδιαίτερα το ARIMA(1,1,1) και ARIMA(2,1,1) είναι επιτυχημένα καθώς επιτυγχάνουν τη πορεία της πρόβλεψης της τιμής σε ένα μεγάλο βαθμό.

## 4.4.1 Ανάλυση τέταρτης χρονολογικής σειράς

Τα στοιχεία της συγκεκριμένης χρονοσειράς αφορούν τις εβδομαδιαίες πώλησεις της Amazon σε διάφορα υποκαταστήματα της. Εμείς θα αφοσιωθούμε στο πρώτο υποκατάστημα μιας και είναι εκείνο με τις περισσότερες πωλήσεις και δεν θα

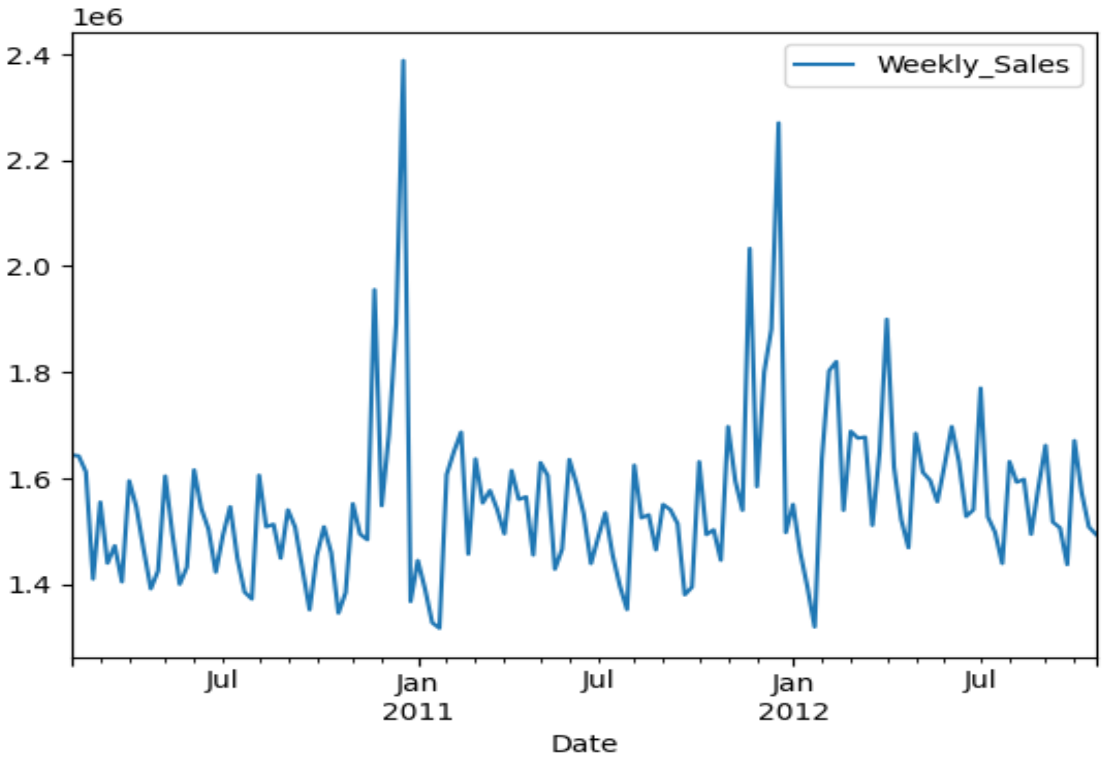
χρειαστεί να περιπλέξουμε άσκοπα το υπόδειγμά μας προσθέτοντας περισσότερα υποκαταστήματα.

### 4.4.2 Μετασχηματισμός δεδομένων

Τα δεδομένα χρειάστηκαν μετασχηματισμό . Κρατήσαμε μόνο τις στήλες τις οποίες θα χρειαστούμε για τη συγκεκριμένη ανάλυση οι οποίες και είναι οι εβδομαδιαίες πωλήσεις καθώς και η ημερομηνία. Κρατήσαμε τα δεδομένα έχοντας ώς κριτήριο το υποκατάστημα 1 και επίσης κάναμε μετασχηματισμό στην ημερομηνία σε τύπου Datetime κρατώντας τη σε μορφή yyyyMMdd.

**Πίνακας 4. 10 Δεδομένα πωλήσεων ανά εβδομάδα**

| Ημερομηνία | Εβδομαδιαίες Πωλήσεις |
|---|---|
| 2010-02-05 | 1643690.90 |
| 2010-02-12 | 1641957.44 |
| 2010-02-19 | 1611968.17 |
| 2010-02-26 | 1409727.59 |
| 2010-03-05 | 1554806.68 |




**Εικόνα 4. 27 Διάγραμμα πωλήσεων ανά εβδομάδα**

### 4.4.3 Έλεγχος KPSS/DICKEY-FULLER

Πραγματοποιώντας τον έλεγχο ADF μέσω της βιβλιοθήκης της Python (adfuller) το αποτέλεσμα είναι το εξής:

**Πίνακας 4. 11 Έλεγχος test Augmented-Dickey-Fuller**

| | |
|---|---|
| ADF Statistic: | -5.102186145192285 |
| P-value: | 1.3877788330759535e-05 |
| Lags Used: | 4 |
| Critical Values:<br>1%:<br>5%: | <br>-3.4786080557767<br>-2.8827609839255 |

| | |
|---|---|
| ADF Statistic: | -5.102186145192285 |
| 10%: | -2.57862070234354 |

Στη συγκεκριμένη περίπτωση βλέπουμε ότι ο στατιστικός ADF είναι -5.1021 και η τιμή p είναι 1.39e-05. Στο συγκεκριμένο τεστ όπως έχουμε αναφέρει και προηγουμένως , η μηδενική υπόθεση είναι ότι η χρονοσειρά έχει μοναδιαία ρίζα και η εναλλακτική υπόθεση είναι ότι η χρονοσειρά είναι στάσιμη.

Δεδομένου ότι η τιμή p είναι σημαντικά χαμηλότερη από τα συνηθή επίπεδα σημαντικότητας , απορρίπτουμε τη μηδενική υπόθεση. Συνεπώς, έχουμε ενδείξεις να υποθέσουμε ότι η χρονοσειρά είναι στάσιμη.

Οι κρίσιμες τιμές που παρέχονται είναι κατώτατα όρια για διάφορα επίπεδα σημαντικότητας (1%, 5% και 10%). Αυτές οι τιμές συγκρίνονται με το δοκιμαστικό στατιστικό ADF για να καθοριστεί εάν θα απορριφθεί η μηδενική υπόθεση. Σε αυτήν την περίπτωση, ο δοκιμαστικός στατιστικός είναι πιο ακραίος (δηλαδή, μικρότερος σε απόλυτη τιμή) από ακόμα και την κρίσιμη τιμή του 1%, υποστηρίζοντας περαιτέρω την απόρριψη της μηδενικής υπόθεσης.

Συνολικά τα στοιχεία από το τέστ ADF, υποδηλώνουν ότι η χρονοσειρά είναι στάσιμη.

**Πίνακας 4. 12 Έλεγχος KPSS**

| | |
|---|---|
| KPSS Statistic: | 0.4758942566191388 |
| P-value: | 0.047095888148842614 |
| Lags Used: | 4 |
| Critical Values: | |
| 10%: | 0.346 |
| 5%: | 0.463 |
| 2.5%: | 0.574 |
| 1%: | 0.739 |

Η τιμή του στατιστικού KPSS είναι 0.4758942566191388 και η αντίστοιχη p-τιμή είναι περίπου 0.0471.

Στο τεστ KPSS, η μηδενική υπόθεση (H0) είναι ότι η χρονοσειρά είναι στάσιμη, ενώ η εναλλακτική υπόθεση (H1) είναι ότι δεν είναι στάσιμη.

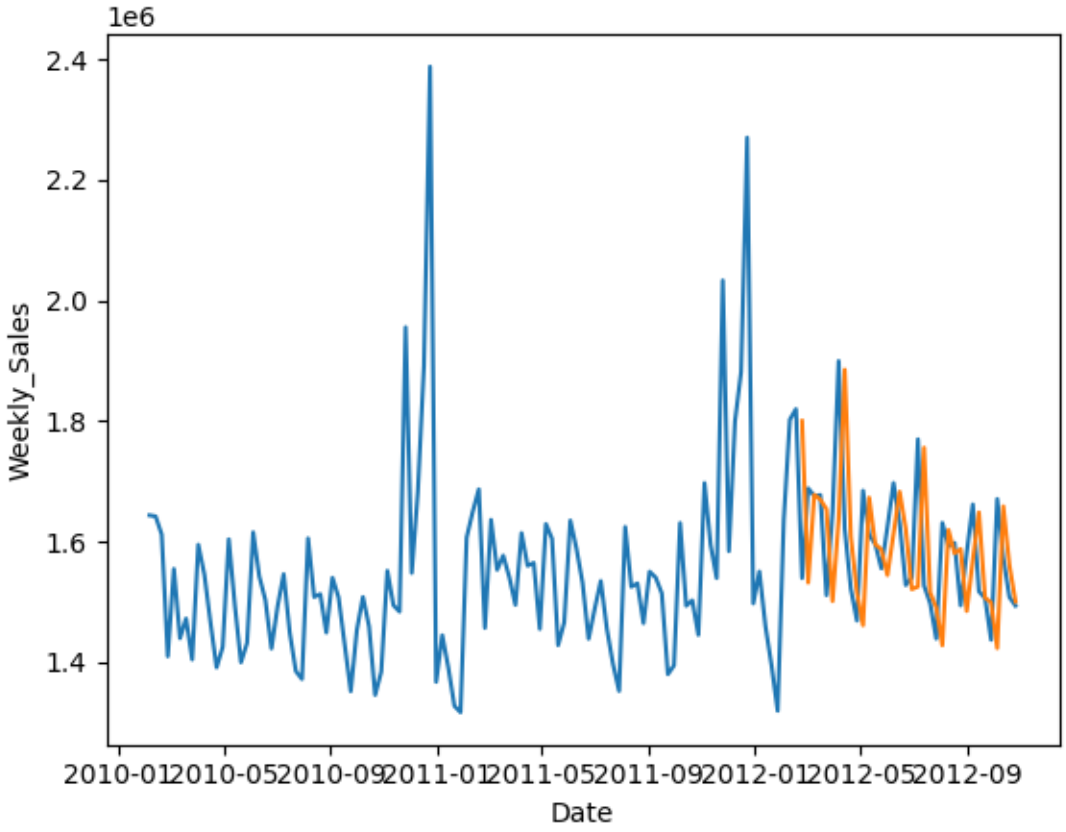
Εδώ, η τιμή p (0.0471) είναι μικρότερη από το επίπεδο σημαντικότητας 0.05, οπότε απορρίπτουμε τη μηδενική υπόθεση και δεν έχουμε αρκετά αποδεικτικά στοιχεία υπέρ της στάσιμης φύσης της χρονοσειράς.

Οι κρίσιμες τιμές που παρέχονται είναι οι κατώτατες τιμές για διάφορα επίπεδα σημαντικότητας (10%, 5%, 2.5%, 1%). Αν η τιμή του στατιστικού είναι μεγαλύτερη από αυτές τις κρίσιμες τιμές, τότε απορρίπτεται η μηδενική υπόθεση. Σε αυτήν την περίπτωση, η τιμή του στατιστικού δεν υπερβαίνει την κρίσιμη τιμή σε κανένα από τα επίπεδα σημαντικότητας, οπότε απορρίπτουμε τη μηδενική υπόθεση και συμπεραίνουμε ότι η χρονοσειρά μας δεν είναι στάσιμη.

Τα συμπεράσματα από τους παραπάνω ελέγχους δεν είναι ξεκάθαρα και οι τρόποι συνέχισης της συγκεκριμένης ανάλυσης είναι αρκετοί. Στη συγκεκριμένη περίπτωση , με τη χρήση της συνάρτησης και βιβλιοθήκης auto-arima (pmdarima), θα βρούμε το καλύτερο δυνατό μοντέλο και έπειτα θα το συγκρίνουμε με τα αποτελέσματα διαγράμματα μερικής και ολικής αυτό-συσχέτισης. Έχει σημασία να δούμε , τη διαφορά που υπάρχει ανάμεσα στο καλύτερο μοντέλο και τις επιλογές που μας δείχνουν ως προτιμότερες τα παραπάνω διαγράμματα.

### 4.4.4 Εφαρμογή Auto-ARIMA

Με την εφαρμογή του μοντέλου Auto-ARIMA , το αποτέλεσμα που εμφανίζεται είναι ότι ο καλύτερος συνδυασμός μεταβλητών-μοντέλου είναι το SARIMA(1,0,0). Με τα συγκεκριμένα δεδομένα , μπορούμε πλέον στο διάγραμμα των δεδομένων να εντοπίσουμε αυτή τη κυκλικότητα που εμφανίζεται ότι υπάρχει και που μας προτρέπει στην χρήση του μοντέλου SARIMA με μεταβλητές (1,0,0). Τα αποτελέσματα της χρήσης του συγκεκριμένου συνδυασμού σε δεδομένα train(75%) και test(25%) είναι τα εξής.

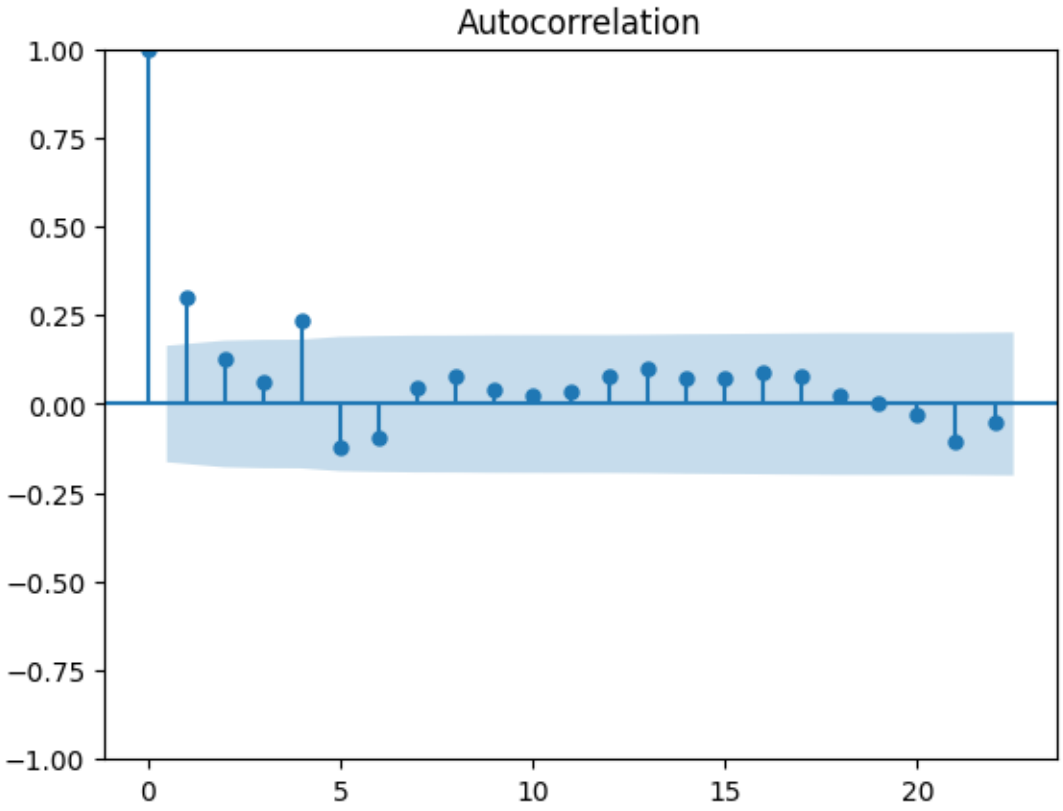



**Εικόνα 4. 28 Γράφημα αποτελεσμάτων μοντέλο SARIMA(1,0,0)**

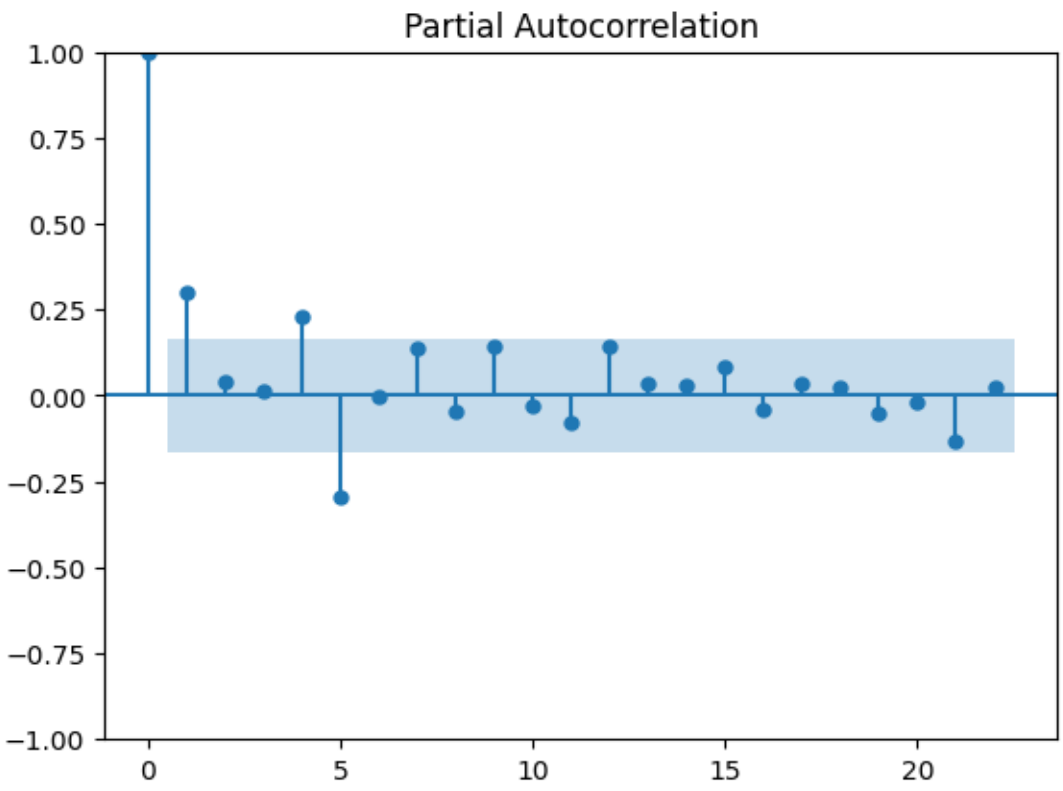
Είναι εμφανές ότι τα συγκεκριμένα αποτελέσματα είναι απόρροια του ότι έχουμε λάβει υπόψη στα αποτελέσματα μας και τα στοιχεία της κυκλικότητας.

### 4.4.5 Σύγκριση με ACF/PACF

Σε αντίθεση με το μοντέλο SARIMA που επέλεξε ως το βέλτιστο για εμάς το auto-ARIMA, μπορούμε να ελέγξουμε τα διαγράμματα μερικής και ολικής αυτό-συσχέτισης.

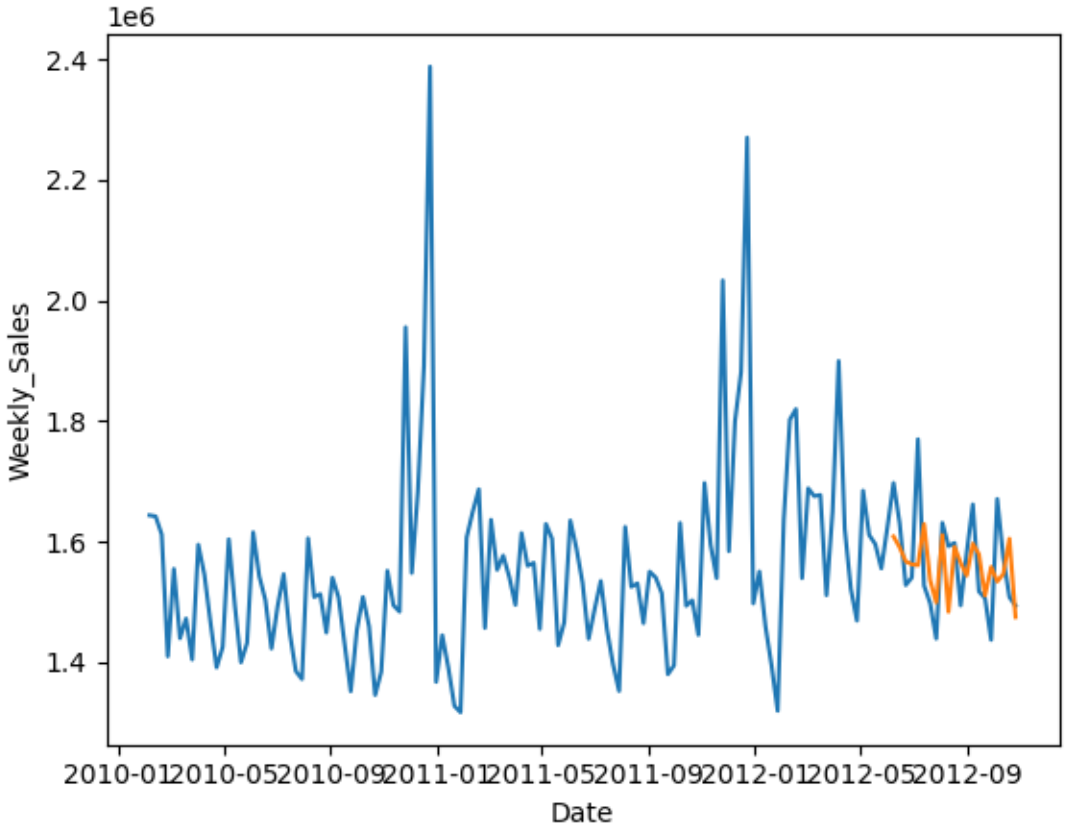



**Εικόνα 4. 29 Γράφημα ολικής αυτό-συσχέτισης**




**Εικόνα 4. 30 Γράφημα μερικής αυτό-συσχέτισης**

Από τα συγκεκριμένα διαγράμματα και χωρίς την δυνατότητα χρησιμοποίησης του auto-ARIMA η επιλογή των παραμέτρων ARIMA θα ήταν για εμάς η (1,0,5) αφού από τα διαγράμματα βλέπουμε ότι οι σημαντικότερες παρατηρήσεις για το καθένα βρίσκεται στις συγκεκριμένες τιμές. Το μοντέλο ARIMA με τις συγκεκριμένες παραμέτρους μας δίνει το εξής αποτέλεσμα.




**Εικόνα 4. 31 Γράφημα αποτελεσμάτων μοντέλου ARIMA(1,0,5)**

Το συγκεκριμένο μοντέλο , παρόλο που δεν δίνει σε καμία περίπτωση την ίδια ακρίβεια με εκείνο που λειτουργήσαμε από το auto-ARIMA, έχει καταφέρει να δείξει την τροχιά και να πρόβλεψει τις ακριβής τιμές κατά ένα μεγάλο βαθμό.

Γενικά, ένα μεγάλο συμπέρασμα από τις παραπάνω εφαρμογές είναι η χρησιμότητα της χρησιμοποίησης του auto-Arima καθώς δεν βασίζεται σε υποκειμενική άποψη και είναι ίσως το πιο χρήσιμο εργαλείο για την επιλογή του μοντέλου μας.

# Συμπεράσματα

Στην παρούσα εργασία μελετήθηκαν χρονοσειρές ζήτησης αλλά και προσφοράς επιχειρήσεων , η τιμή της μετοχής της AMAZON στο χρηματιστήριο καθώς και μία γενικότερη εισαγωγική χρονοσειρά με τους επιβάτες αεροπλάνων . Για την ανάλυση των παραπάνω χρονοσειρών , χρησιμοποιήσαμε μοντέλα ARIMA καθώς και τις εναλλακτικές μορφές που μπορεί να πάρει , προσθέτοντας η αφαιρώντας ανάλογα μεταβλητές. Η επιλογή των μεταβλητών έγινε με διάφορους τρόπους σε κάθε εφαρμογή , σύμφωνα με τα δεδομένα της κάθε χρονοσειράς .

Τα αποτελέσματα των προβλέψεων είχαν μεγάλη επιτυχία και ειδικότερα στις περιπτώσεις των πωλήσεων και των αγορών όπου τα δεδομένα είχαν ένα στοιχείο εποχικότητας. Σε αυτές τις περιπτώσεις , εντάχθηκε το στοιχείο της εποχικότητας στο μοντέλο ARIMA μετατρέποντας το σε ένα υβριδικό μοντέλο SARIMA. Το συγκεκριμένο μοντέλο είχε μεγάλο βαθμό επιτυχίας. Από την άλλη μεριά, η εφαρμογή του μοντέλου ARIMA στη τιμή της μετοχής της AMAZON δεν έφερε κάποιο ουσιαστικό αποτέλεσμα. Η πρόβλεψη σε τιμές μετοχών είναι ένα δύσκολο αντικείμενο το οποίο λόγο της μεγαλύτερης πολυπλοκότητας που έχει , δεν έφερε κάποιο ουσιαστικό αποτέλεσμα στην εφαρμογή .

Θα είχε ενδιαφέρον, η περαιτέρω μελέτη χρονοσειρών πωλήσεων και αγορών καθώς φαίνεται πως μπορεί να κάνει εύστοχες προβλέψεις σε επιχειρήσεις.

# Βιβλιογραφία


Aptech.com. (2020). *Introduction To The Fundamentals Of Time Series Data and Analysis.* Ανάκτηση από https://www.aptech.com/blog/introduction-tothe-fundamentals-of-time-series-data-and-analysis/.

Brockwell, R. (2016). *Time Series and forecasting.* Brockwell.

Contreras, J. (2003). ARIMA models to predict next-day electricity prices.

Dickey, D. (1997). *Distribution of the estimators for autoregressive time series with a unit root.* Journal of the American Statistical Association.

Dorothy, J. (2018). *The SAGE Encyclopedia of Educational Research , Measurement , and Evaluation.* SAGE Publiccation , Inc.

Gebhard, K. (2006). *Introduction to Modern Time Series Analysis.* Springer.

Hamilton, J. (1994). *Time Series Analysis.* Princeton: Princeton University Press.

Henrik, M. (1955). *Time Series Analysis.* Texts in Statistical Science.

Kwiatkowski, D. ,. (1992). *Testing the null hypothesis of stationarity against the alternartive of a unit root.* Journal of Econometrics.

Perktold, J. (2009-2023). Ανάκτηση από www.statsmodels.org: https://www.statsmodels.org/stable/generated/statsmodels.tsa.arima.model.ARIMA.html#statsmodels.tsa.arima.model.ARIMA-attributes

Provost F., &. F. (2019). *Η επιστήμη των δεδομένων για επιχειρήσεις.* Κλειδάριθμος.

Siegel, E. (2013). *Predictive Analytics : The Power to Predict Who Will Click , Buy , Lie or Die.*

Xinghua, C. (2012). Seasonal Autoregressive Integrated Moving Average Model for Precipitaion Time Series.

Βιδάλης, Μ. (2007). *Εφοδιαστική(Logistics) Μία ποσοτική προσέγγιση.* Αθήνα: Κλειδάριθμος.

Δημέλη, Σ. (2003). *Σύγχρονές Μέθοδοι Ανάλυσης Χρονολογικών Σειρών.* Εκδόσεις Κριτική.

Καρακασίδης, Θ. (2017). *Ανάλυση χρονοσειρών και δεδομένων περιβαλλοντικών κινδύνων.*

Κοκολάκης, Γ. (2009). *Στατιστική Θεωρία & Εφαρμογές.* Αθήνα.

Κουγιουμτζής, Σ. (2005). *Γραμμική ανάλυση χρονοσειρών.* Πανεπιστημιακές σημειώσεις.

Μόλης, Θ. (1995). *Προβλέψεις.* Πανεπιστημιακές εκδόσεις Κρήτης.

# Εργαλεία και κώδικας

## Βιβλιοθήκες

- statsmodels.tsa.statespace.sarimax
- statsmodels.graphics.tsaplots
- pmdarima
- datetime
- matplotlib
- pandas.plotting
- statsmodels.tsa.arima.model
- sklearn.metrics mean_squared_error
- math
- seaborn
- statsmodels.tsa.stattools
- numpy
- pandas

## Κώδικας

- Επεξεργασία δεδομένων:

```
#Διάβασμα αρχείου csv
df = pandas.DataFrame(pd.read_csv(
   r'C:\Users\vagge\Downloads\archive (3)\mydata.csv'))


#Μετατροπή ημερομηνίας απο string σε Date
df['Date'] = pandas.to_datetime(df['Date'])
# %%
#Κράταμε μόνο τις κολώνες Date και Open price που χρειάζομαστε
df.drop(df.columns.difference(['Date', 'Open']), 1, inplace=True)
```

```
#Ομαδοποίηση δεδομένων ανά εβδομάδα
fom = df.groupby(pd.Grouper(key='Date',freq='MS')).first().reset_index()

# %%
#Εμφάνιση τις τιμές των τελευταίων 5 δεδομένων
fom.tail()

#Για την αφαίρεση των μηδενικών τιμών
fom=fom.dropna()

#Χωρισμός δεδομένων σε train/test Data
train = fom[:round(len(fom)*90/100)]
test = fom[round(len(fom)*90/100):]
```

- Οπτικοποίηση δεδομένων:

```
# Εμφάνιση της πρώτης και της διαφοροποιημένης χρονοσειράς
plt.figure(figsize=(10, 5))
plt.subplot(2, 1, 1)
plt.plot(fom.index, fom['Open'])
plt.title('Original Open Prices')
plt.xlabel('Date')
plt.ylabel('Open Price')

plt.subplot(2, 1, 2)
plt.plot(fom.index, fom['Open_diff'], color='orange')
plt.title('First Difference of Open Prices')
plt.xlabel('Date')
plt.ylabel('Open Price Difference')

plt.tight_layout()
```

```
plt.show()

#Εμφάνιση της ολικής και της μερικής αυτοσυσχέτισης
fig, (ax1, ax2) = plt.subplots(2, 1, figsize=(10, 8))

# ACF plot
plot_acf(fom['Open'].dropna(), ax=ax1, lags=10)
ax1.set_title('Autocorrelation Function (ACF) of Differenced Open Prices')

# PACF plot
plot_pacf(fom['Open'].dropna(), ax=ax2, lags=10)
ax2.set_title(
   'Partial Autocorrelation Function (PACF) of Differenced Open Prices')

plt.show()

#Εμφάνιση αποτελεσμάτων πρόβλεψεις και πραγματικότητας
sns.lineplot(data=fom, x=fom.index, y='Open')
sns.lineplot(data=fom, x=fom.index, y='arimaPred')
```

- Χρήση στατιστικών μοντέλων ARIMA

```
#Χρήση μοντέλου auto-Arima
model = pm.auto_arima(train['Open'], start_p=1, start_q=1,
                      test='adf',      # use adftest to find optimal 'd'
                      max_p=10, max_q=10,  # maximum p and q
                      m=1,             # frequency of series
                      d=None,          # let model determine 'd'
                      seasonal=False,   # No Seasonality
                      start_P=0,
                      D=0,
                      trace=True,
```

```
                error_action='ignore',
                suppress_warnings=True,
                stepwise=True)
#Χρήση μοντέλου Arima μοντέλου στα train δεδομένα
model = SARIMAX(train['Open'], order=(1, 1, 1),seasonal_order=(1,1,1,12)).fit()
pred = model.predict(start=len(train), end=(len(fom)-1))
#Υπολογισμός test Dickey-Fuller
 result = adfuller(fom['Open'])
print(f'ADF Statistic: {result[0]}')
print(f'P-value: {result[1]}')
print(f'Lags Used: {result[2]}')
print('Critical Values:')
for key, value in result[4].items():
   print(f'   {key}: {value}')
```