PHAROS
THE GREEK AI FACTORY

Ανάλυση Χρονοσειρών και προβλεπτική αναλυτική (Time series analysis and forecasting)

Περιεχόμενα

- Εισαγωγή στις χρονοσειρές
- **Επιμέρους στοιχεία χρονοσειρών**
  - Τάση (Trend),
  - Εποχικότητα (Seasonality),
  - Θόρυβος (Noise)
- **Αναλυση χρονοσειρών**
  - Μορφή χρονοσειράς, Είδη χρονοσειράς
  - Διάγνωση Τάσης και Εποχικότητας και απαλοιφή, Εξομάλυνση (Smoothing)
  - Καμπύλες Αυτοσυσχέτισης και Μερικής Αυτοσυσχέτισης (ACF, PACF plots)
  - Απλός κινητός μέσος (Simple Moving Average), Σταθμισμένος κινητός μέσος (Weighted Moving Average)
  - Exponential Smoothing (Holt's Linear Model, Holt' Winters method)
  - Αυτοσυσχέτιση (autocorrelation)
  - Αυτοπαλίνδρομα μοντέλα: ARMA, ARIMA, SARIMA, SARIMAX

Εισαγωγικά στοιχεία (Ι)

Χρονοσειρά (*time-series*) ορίζεται μια συλλογή παρατηρήσεων που ευρετηριάζονται με βάση την ημερομηνία κάθε παρατήρησης [1].

Η **ανάλυση χρονοσειρών** *(time series analysis)* είναι η προσπάθεια εξαγωγής ουσιαστικών συνοπτικών και στατιστικών πληροφοριών από σημεία διατεταγμένα σε χρονολογική σειρά [2].

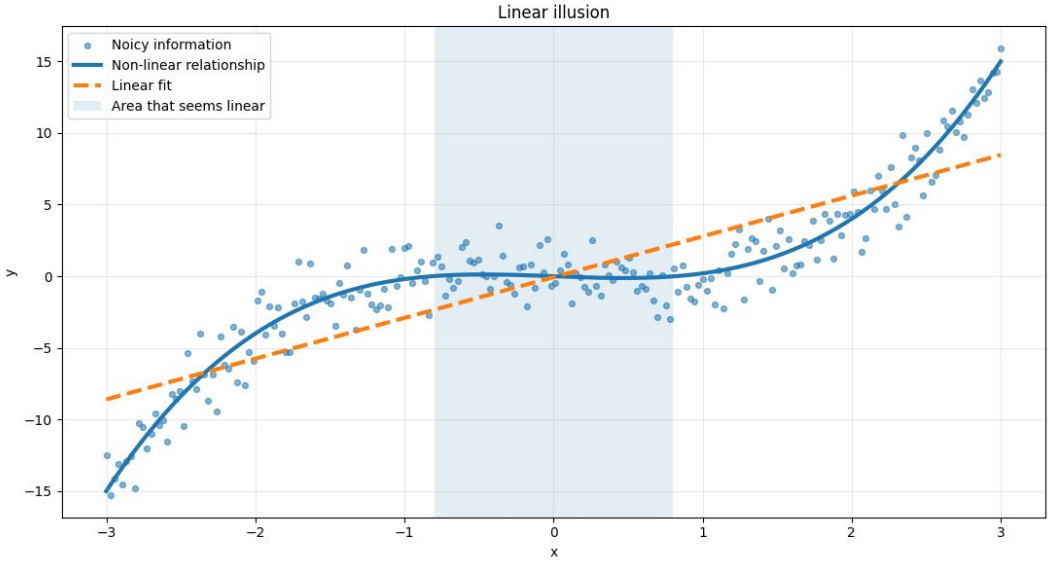
Η ανάλυση χρονοσειρών συνεισφέρει στη διάγνωση της συμπεριφοράς του παρελθόντος καθώς και στη πρόβλεψη της μελλοντικής συμπεριφοράς, τον εντοπισμό μοτίβων και φαινομένων.

Η συνεχής δημιουργία και η συλλογή δεδομένων αυξάνει την ανάγκη για κατάλληλη ανάλυση χρονοσειρών τόσο με τεχνικές στατιστικής όσο και με τεχνικές μηχανικής μάθησης.

[1] Hamilton, J. D. (2020). *Time series analysis*. Princeton university press.
[2] Nielsen, A. (2019). Practical time series analysis: Prediction with statistics and machine learning. O'Reilly Media.

Τα δεδομένα είναι πάντα έτσι όπως φαίνονται;

Η γραμμικότητα είναι συχνά μια γνωστική ψευδαίσθηση.



Εισαγωγικά στοιχεία (ΙΙ)

Παραδείγματα χρονοσειρών συναντάμε σε διάφορους τομείς (Οικονομία, Υγεία, Μετεωρολογία, Αθλητισμό κτλ.):

- **Τιμές μετοχών:** Ημερήσιες, ωριαίες ή ακόμη και κατά λεπτό τιμές μετοχών για μια χρονική περίοδο.
- **Δεδομένα καιρού:** Θερμοκρασία, βροχόπτωση, επίπεδα υγρασίας κ.λπ. για μια χρονική περίοδο.
- **Δεδομένα πωλήσεων:** Ημερήσια, μηνιαία ή ετήσια δεδομένα πωλήσεων για ένα προϊόν ή μια υπηρεσία.
- **Κατανάλωση ενέργειας:** Δεδομένα ωριαίας ή ημερήσιας κατανάλωσης ενέργειας για ένα κτίριο ή μια πόλη.

Εισαγωγικά στοιχεία (ΙΙ)

- **Οικονομικοί δείκτες:** Ποσοστά ανεργίας, ρυθμοί αύξησης του ΑΕΠ, ρυθμοί πληθωρισμού κ.λπ. σε μια χρονική περίοδο.

- **Ιατρικά δεδομένα:** Η αρτηριακή πίεση, ο καρδιακός ρυθμός, τα επίπεδα γλυκόζης κ.λπ. καταγράφηκαν με την πάροδο του χρόνου.

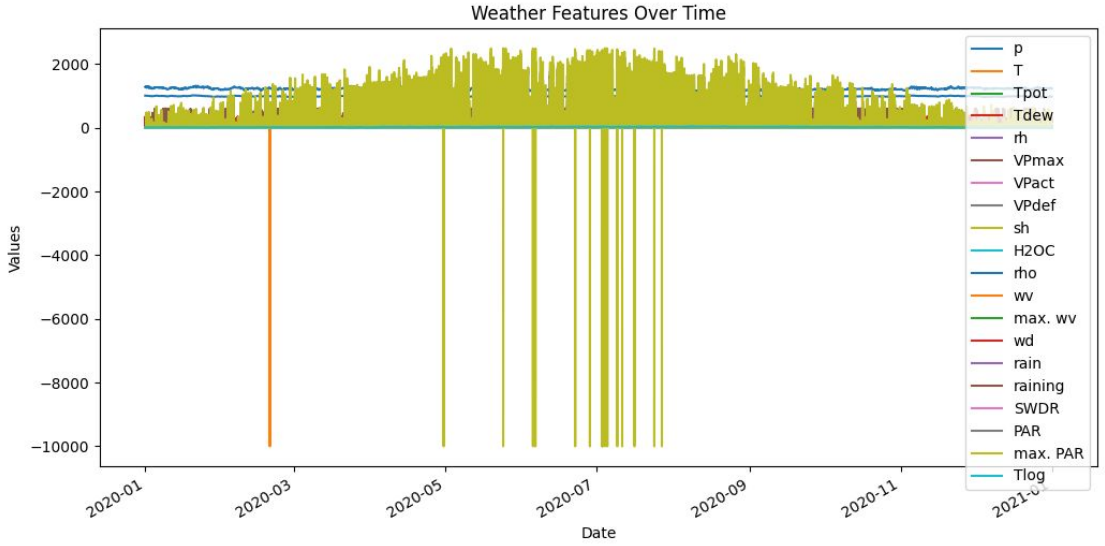
- **Επισκεψιμότητα ιστότοπου:** Ωριαία, ημερήσια ή εβδομαδιαία δεδομένα επισκεψιμότητας ιστότοπου.

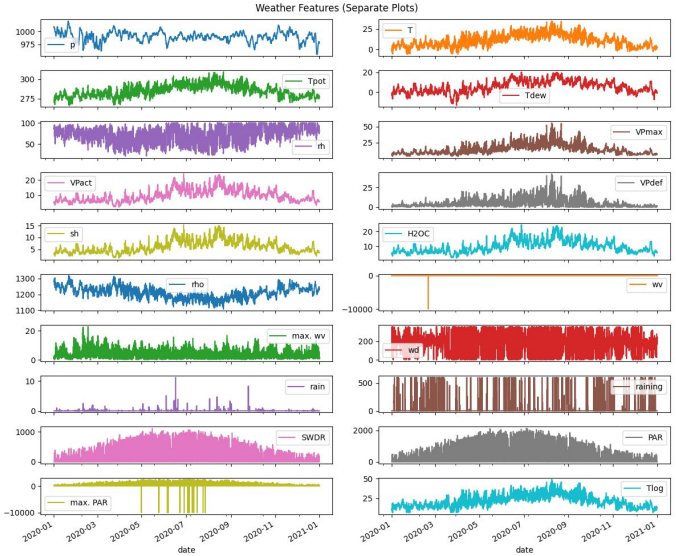
- **Δεδομένα αισθητήρα:** Θερμοκρασία, πίεση, επίπεδα υγρασίας κ.λπ. μετρημένα σε τακτά χρονικά διαστήματα.

- **Αθλητικά δεδομένα:** Οι βαθμολογίες μιας ομάδας ή ενός μεμονωμένου παίκτη για μια χρονική περίοδο.

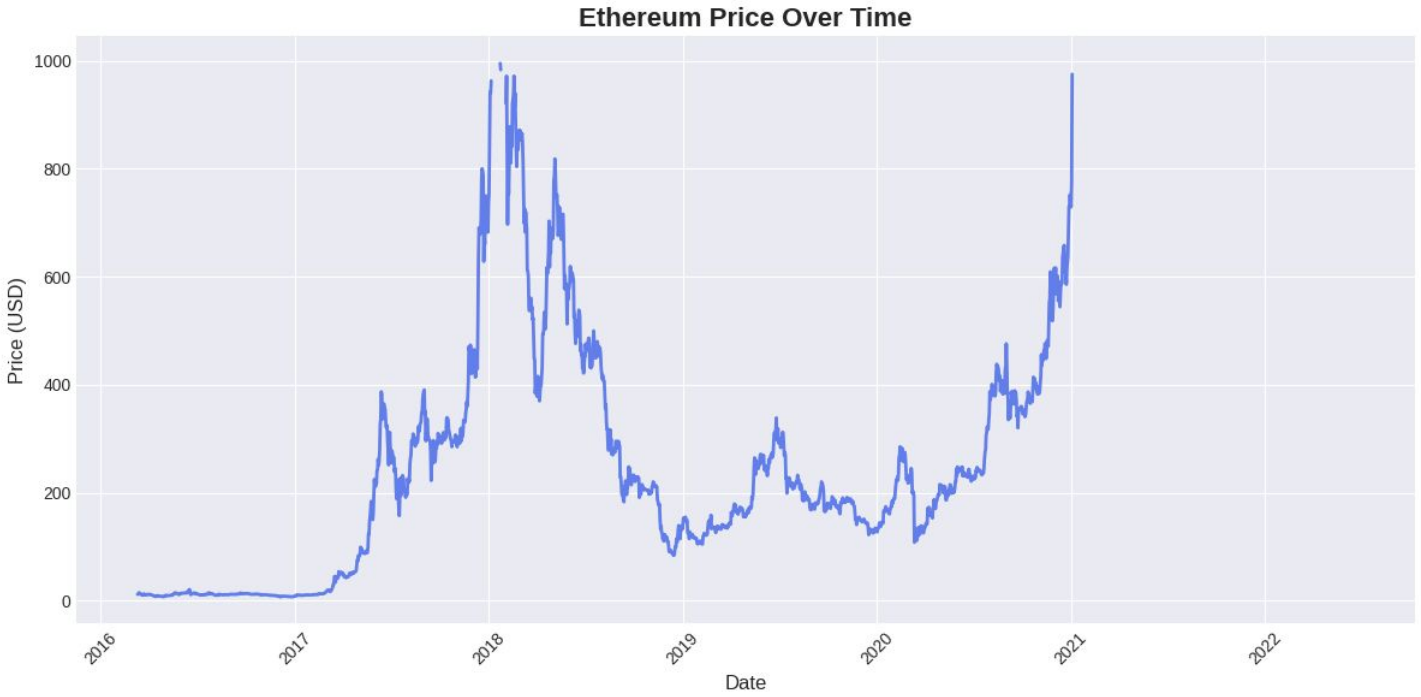
Παράδειγμα 1/5

Παράδειγμα 2/5

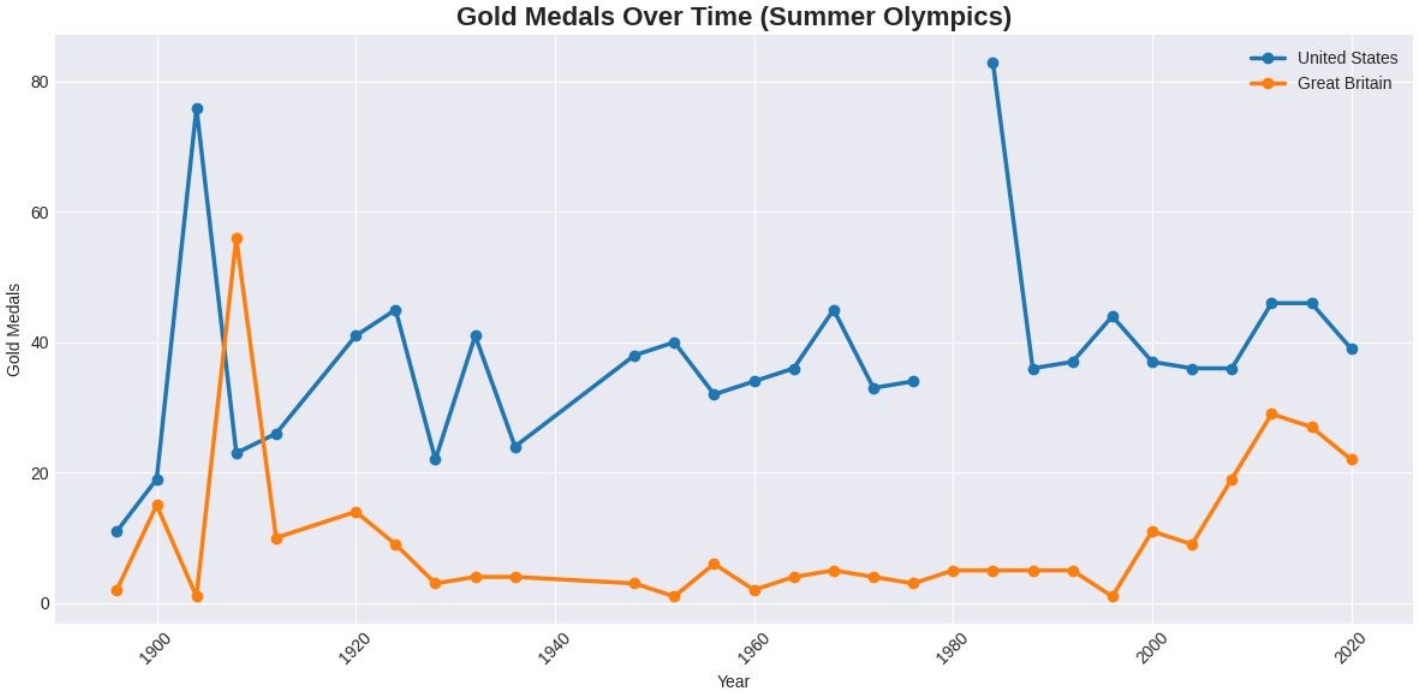




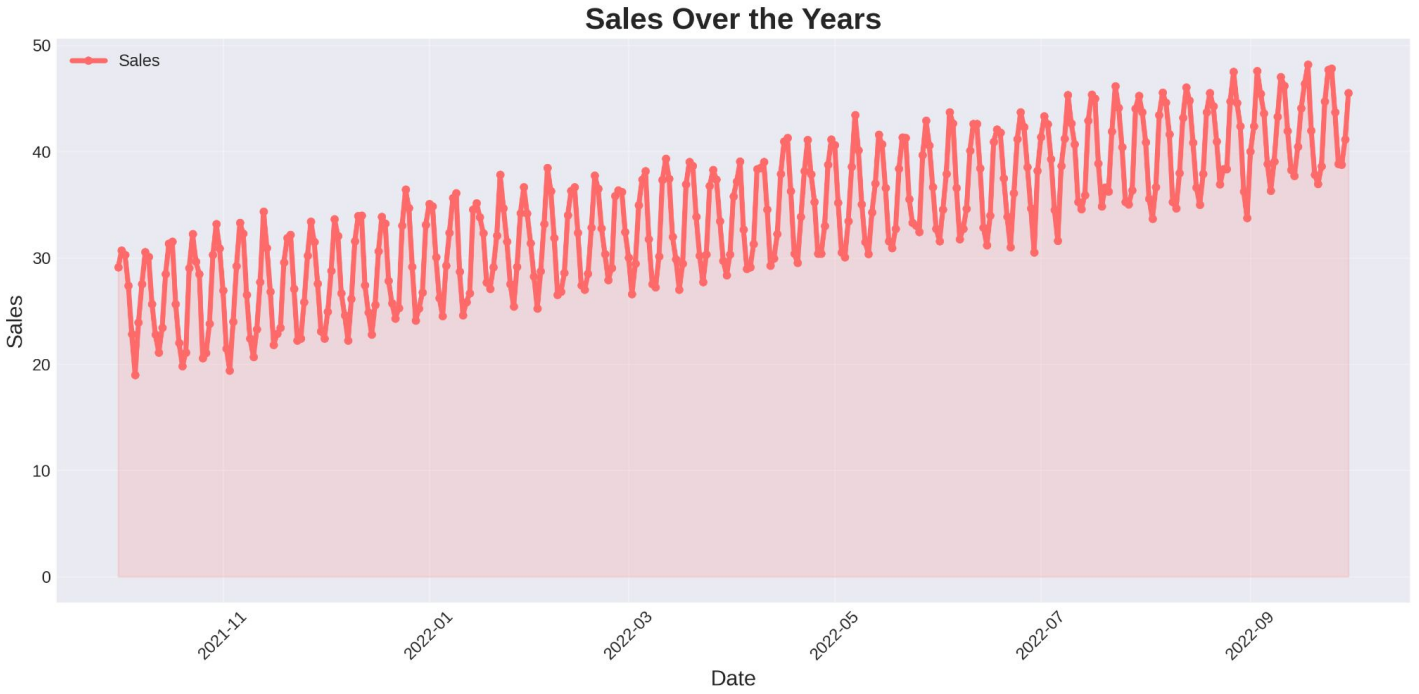
Παράδειγμα 3/5



Παράδειγμα 4/5



Παράδειγμα 5/5



Μορφή Χρονοσειράς ως σύνολο δεδομένων

- Τα σύνολα δεδομένων που απεικονίζουν μια χρονοσειρά έχουν δομημένη μορφή (structured data).

- Συνήθεις μορφές δεδομένων (.csv, .xlxs, .json, .txt)

Date	Daily minimum temperatures
1/1/81	20.7
1/2/81	17.9
1/3/81	18.8
1/4/81	14.6
1/5/81	15.8
1/6/81	15.8
1/7/81	15.8
1/8/81	17.4
1/9/81	21.8
1/10/81	20
1/11/81	16.2
1/12/81	13.3
1/13/1981	16.7
1/14/1981	21.5
1/15/1981	25
1/16/1981	20.7
1/17/1981	20.6
1/18/1981	24.8
1/19/1981	17.7
1/20/1981	15.5
1/21/1981	18.2
1/22/1981	12.1
1/23/1981	14.4
1/24/1981	16
1/25/1981	16.5
1/26/1981	18.7
1/27/1981	19.4
1/28/1981	17.2
1/29/1981	15.5
1/30/1981	15.1
1/31/1981	15.4

Month	Monthly beer production
1956-01	93.2
1956-02	96
1956-03	95.2
1956-04	77.1
1956-05	70.9
1956-06	64.8
1956-07	70.1
1956-08	77.3
1956-09	79.5
1956-10	100.6
1956-11	100.7
1956-12	107.1
1957-01	95.9
1957-02	82.8
1957-03	83.3
1957-04	80
1957-05	80.4
1957-06	67.5
1957-07	75.7
1957-08	71.1


```
{
  {
    "date":"1-1-2016",
    "shampoo_price":"2.9$"
  },
  {
    "date":"1-3-2017",
    "shampoo_price":"2.8$"
  },
  {
    "date":"1-3-2018",
    "shampoo_price":"2.85$"
  },
  {
    "date":"1-3-2019",
    "shampoo_price":"2.9$"
  },
  {
    "date":"1-1-2020",
    "shampoo_price":"3$"
  },
  {
    "date":"1-3-2021",
    "shampoo_price":"3.2$"
  },
  {
    "date":"1-3-2022",
    "shampoo_price":"4.12$"
  },
  {
    "date":"1-3-2023",
    "shampoo_price":"4.15$"
  }
}
```


data-01.txt


```
04-21-1991    9:09    58    100
04-21-1991    9:09    33    009
04-21-1991    9:09    34    013
04-21-1991    17:08   62    119
04-21-1991    17:08   33    007
04-21-1991    22:51   48    123
04-22-1991    7:35    58    216
04-22-1991    7:35    33    010
04-22-1991    7:35    34    013
04-22-1991    13:40   33    002
04-22-1991    16:56   62    211
04-22-1991    16:56   33    007
04-23-1991    7:25    58    257
04-23-1991    7:25    33    011
04-23-1991    7:25    34    013
04-23-1991    17:25   62    129
04-23-1991    17:25   33    007
04-24-1991    7:52    58    239
04-24-1991    7:52    33    010
04-24-1991    7:52    34    014
04-24-1991    12:00   33    004
04-24-1991    17:10   62    129
04-24-1991    22:09   48    340
04-24-1991    22:09   33    005
04-25-1991    7:29    58    067
04-25-1991    7:29    33    009
04-25-1991    7:29    34    014
04-25-1991    12:49   33    004
04-25-1991    17:24   62    206
04-25-1991    17:24   33    007
```

Είδη Χρονοσειράς (Ι)

1. **Μονομεταβλητά Υποδείγματα Χρονοσειρών (Univariate Time Series)**
   Μια μονομεταβλητή χρονοσειρά αποτελείται από παρατηρήσεις μίας μόνο μεταβλητής ενδιαφέροντος καταγεγραμμένες διαχρονικά. Δηλαδή, παρακολουθούμε την εξέλιξη μιας συγκεκριμένης ποσότητας με την πάροδο του χρόνου.

   Παράδειγμα: **Ημερήσιες τιμές κλεισίματος ενός χρηματιστηριακού δείκτη.**

2. **Πολυμεταβλητά Υποδείγματα Χρονοσειρών (Multivariate Time Series)**
   Μια πολυμεταβλητή χρονοσειρά αποτελείται από παρατηρήσεις δύο ή περισσότερων μεταβλητών που εξελίσσονται ταυτόχρονα στον χρόνο. Επιτρέπει τη μελέτη σχέσεων και αλληλεπιδράσεων μεταξύ διαφορετικών μεταβλητών.

   Παράδειγμα: Παρακολούθηση της πορείας μιας εταιρείας στον χρόνο λαμβάνοντας υπόψη μεταβλητές όπως έσοδα, έξοδα, κέρδος ή αριθμό εργαζομένων.

Είδη Χρονοσειράς (II)

Month	Sales of shampoo over a three year period
1-Jan	266
1-Feb	145.9
1-Mar	183.1
1-Apr	119.3
1-May	180.3
1-Jun	168.5
1-Jul	231.8
1-Aug	224.5
1-Sep	192.8
1-Oct	122.9
1-Nov	336.5
1-Dec	185.9
2-Jan	194.3
2-Feb	149.5
2-Mar	210.1
2-Apr	273.3
2-May	191.4
2-Jun	287
2-Jul	226

Univariate time series

No	year	month	day	hour	pm2.5	DEWP	TEMP	PRES	wind direction	wind speed	snow	rain
1	2010	1	1	0	NA	-21	-11	1021	NW	1.79	0	0
2	2010	1	1	1	NA	-21	-12	1020	NW	4.92	0	0
3	2010	1	1	2	NA	-21	-11	1019	NW	6.71	0	0
4	2010	1	1	3	NA	-21	-14	1019	NW	9.84	0	0
5	2010	1	1	4	NA	-20	-12	1018	NW	12.97	0	0
6	2010	1	1	5	NA	-19	-10	1017	NW	16.1	0	0
7	2010	1	1	6	NA	-19	-9	1017	NW	19.23	0	0
8	2010	1	1	7	NA	-19	-9	1017	NW	21.02	0	0
9	2010	1	1	8	NA	-19	-9	1017	NW	24.15	0	0
10	2010	1	1	9	NA	-20	-8	1017	NW	27.28	0	0
11	2010	1	1	10	NA	-19	-7	1017	NW	31.3	0	0
12	2010	1	1	11	NA	-18	-5	1017	NW	34.43	0	0
13	2010	1	1	12	NA	-19	-5	1015	NW	37.56	0	0
14	2010	1	1	13	NA	-18	-3	1015	NW	40.69	0	0
15	2010	1	1	14	NA	-18	-2	1014	NW	43.82	0	0
16	2010	1	1	15	NA	-18	-1	1014	cv	0.89	0	0
17	2010	1	1	16	NA	-19	-2	1015	NW	1.79	0	0
18	2010	1	1	17	NA	-18	-3	1015	NW	2.68	0	0
19	2010	1	1	18	NA	-18	-5	1016	NE	1.79	0	0

Multivariate time series

Hands-on

Hands-on 1 — Εισαγωγή και Βασική Διερευνητική Ανάλυση Δεδομένων

Να πραγματοποιηθεί βασική διερευνητική ανάλυση σε **μονομεταβλητό** και **πολυμεταβλητό** σύνολο δεδομένων.

Συγκεκριμένα, ζητείται:

- Εισαγωγή των συνόλων δεδομένων στο περιβάλλον ανάλυσης.
- Εκτύπωση τιμών για αρχική κατανόηση της δομής των δεδομένων.
- Προβολή επιλεγμένων υποσυνόλων τιμών (π.χ. αρχικές και τελικές εγγραφές).
- Υπολογισμός και παρουσίαση περιγραφικών στατιστικών (descriptive statistics).
- Οπτικοποίηση των δεδομένων με κατάλληλες γραφικές παραστάσεις όπου απαιτείται.

Βασικά Στοιχεία Χρονοσειρών και Δομική Ανάλυση

- **Τάση (Trend)**

Η τάση περιγράφει τη μακροχρόνια κατεύθυνση της χρονοσειράς, δηλαδή αν οι τιμές αυξάνονται, μειώνονται ή παραμένουν σταθερές με την πάροδο του χρόνου.

- **Εποχικότητα (Seasonality)**

Η εποχικότητα αναφέρεται σε επαναλαμβανόμενες μεταβολές που εμφανίζονται σε σταθερά χρονικά διαστήματα (π.χ. ανά ημέρα, μήνα ή έτος). Συνήθως σχετίζεται με εξωτερικούς παράγοντες, όπως ο καιρός, οι αργίες ή οι καταναλωτικές συνήθειες.

- **Περιοδικότητα (Periodicity / Cyclicality)**

Η περιοδικότητα αποτελεί γενικότερο όρο και περιγράφει κάθε μορφή επαναλαμβανόμενου μοτίβου σε μια χρονοσειρά, ανεξάρτητα από το αν σχετίζεται με εποχικούς παράγοντες ή όχι.

Επιμέρους στοιχεία χρονοσειρών

- **Θόρυβος (Noise)**

Ο θόρυβος αναφέρεται στις τυχαίες διακυμάνσεις ή αποκλίσεις που εμφανίζονται στα δεδομένα και δεν μπορούν να εξηγηθούν από την τάση ή την εποχικότητα. Η κατανόηση και, όπου είναι δυνατό, η μείωση του θορύβου αποτελεί σημαντικό στάδιο στη διαδικασία ανάλυσης χρονοσειρών.

- **Άλλα Φαινόμενα**

Σε μια χρονοσειρά μπορεί να εμφανίζονται και άλλα σημαντικά χαρακτηριστικά, όπως ξαφνικές αλλαγές στη συνολική συμπεριφορά, έντονες ακραίες τιμές (outliers), ελλιπή δεδομένα ή άλλες ανωμαλίες που αξίζει να διερευνηθούν.

Hands-on 2 — Οπτική Ανάδειξη Φαινομένων Χρονοσειρών

Να πραγματοποιηθεί διερευνητική οπτικοποίηση χρονοσειρών με στόχο την ανάδειξη βασικών χαρακτηριστικών που περιγράφουν τη συμπεριφορά των δεδομένων στον χρόνο.

Συγκεκριμένα, να αναδειχθούν φαινόμενα όπως η τάση, η εποχικότητα και οι ακραίες τιμές, τόσο σε μονομεταβλητό όσο και σε πολυμεταβλητό σύνολο δεδομένων, μέσω κατάλληλων γραφικών αναπαραστάσεων και τεχνικών επεξεργασίας δεδομένων.

Ανάλυση Χρονοσειρών

- **Περιγραφή (Description) – Εστίαση στο Παρελθόν**
  Κατανόηση της συμπεριφοράς των δεδομένων και εντοπισμός τάσης, εποχικότητας ή σημαντικών αλλαγών.
  Παράδειγμα: Ανάλυση ιστορικών πωλήσεων, μελέτη θερμοκρασιών προηγούμενων ετών.

- **Πρόβλεψη (Prediction) – Εστίαση στο Μέλλον**
  Εκτίμηση μελλοντικών τιμών με βάση ιστορικά δεδομένα.
  Παράδειγμα: Πρόβλεψη τιμών μετοχών ή οικονομικών δεικτών.

- **Έλεγχος (Control) – Εστίαση στο Παρόν**
  Παρακολούθηση διαδικασιών σε πραγματικό χρόνο ώστε να παραμένουν εντός επιθυμητών ορίων.
  Παράδειγμα: Παρακολούθηση θερμοκρασίας παραγωγής, έλεγχος ποιότητας προϊόντων.

Απαιτήσεις και ο πραγματικός κόσμος των χρονοσειρών

Οι περισσότερες κλασικές μέθοδοι ανάλυσης χρονοσειρών βασίζονται σε συγκεκριμένες υποθέσεις σχετικά με τα δεδομένα που αναλύονται.

Συνήθως θεωρείται ότι:

- Οι παρατηρήσεις έχουν συλλεχθεί σε ίσα χρονικά διαστήματα, χωρίς ελλείποντα δεδομένα.
  Η χρονοσειρά έχει επαρκές μήκος ώστε να μπορεί να υποστηρίξει αξιόπιστη ανάλυση.
  Η χρονοσειρά είναι στάσιμη (stationary), δηλαδή δεν παρουσιάζει μακροχρόνια τάση ή εποχικότητα και τα στατιστικά χαρακτηριστικά της (όπως μέση τιμή και διακύμανση) παραμένουν σταθερά στο χρόνο.

Στην πράξη, όμως, τα πραγματικά δεδομένα σπάνια ικανοποιούν πλήρως αυτές τις υποθέσεις. Συχνά εμφανίζονται ελλιπή δεδομένα, θόρυβος, τάσεις ή εποχικά μοτίβα.

Για τον λόγο αυτό, πριν την ανάλυση ή τη μοντελοποίηση, είναι συνήθως απαραίτητη η **διαδικασία καθαρισμού και προεπεξεργασίας δεδομένων (data cleaning).**

Ανάλυση Χρονοσειράς: Κατάργηση Τάσης και Εποχικότητας (II)

Μια χρονοσειρά μπορεί να θεωρηθεί ως συνδυασμός διαφορετικών συνιστωσών που περιγράφουν τη συμπεριφορά της στον χρόνο. Συγκεκριμένα, μια παρατήρηση της χρονοσειράς σε χρονική στιγμή *t* μπορεί να εκφραστεί ως άθροισμα της **τάσης**, της **εποχικότητας** και του **τυχαίου υπολοίπου**:

$$X_t = m_t + s_t + y_t$$

- **m□ (Τάση):** Εκφράζει τη μακροχρόνια πορεία της χρονοσειράς, δηλαδή αν τα δεδομένα αυξάνονται, μειώνονται ή παραμένουν σταθερά στον χρόνο.
- **s□ (Εποχικότητα / Περιοδικότητα):** Περιγράφει τα επαναλαμβανόμενα μοτίβα που εμφανίζονται σε συγκεκριμένα χρονικά διαστήματα (π.χ. ανά έτος, μήνα ή εβδομάδα).
- **y□ (Υπόλοιπο / Θόρυβος):** Αντιπροσωπεύει την τυχαία μεταβλητότητα που παραμένει αφού αφαιρεθούν η τάση και η εποχικότητα από τη χρονοσειρά.

Ανάλυση Χρονοσειράς: Κατάργηση Τάσης και Εποχικότητας

- Η κατάργηση της τάσης και της εποχικότητας μπορεί να μας βοηθήσει να κατανοήσουμε καλύτερα και να αναλύσουμε τα δεδομένα χρονοσειρών και μπορεί να οδηγήσει σε ακριβέστερες προβλέψεις και πληροφορίες.
- Καταργώντας την τάση και την εποχικότητα, μπορούμε πιο εύκολα να εντοπίσουμε άλλα μοτίβα στις χρονοσειρές, (κύκλους ή ακανόνιστες διακυμάνσεις), που μπορεί να είναι πιο σημαντικά για την κατανόηση της συμπεριφοράς του της μεταβλητής που μελετάται.
- Συμβάλει σε ακριβέστερες προβλέψεις. Η τάση και η εποχικότητα μπορεί να παραμορφωσουν τα μοτίβα στα δεδομένα και να κάνουν πιο δύσκολη την πρόβλεψη μελλοντικών τιμών. Αφαιρώντας αυτά τα στοιχεία, μπορούμε να κάνουμε πιο ακριβείς προβλέψεις για τις μελλοντικές τιμές.

Ανάλυση Χρονοσειράς: Διάγνωση Τάσης

Η διάγνωση της τάσης μπορεί να γίνει μέσω οπτικοποίησης (Visualization).
Ωστόσο, υπάρχουν και στατιστικά τεστ που χρησιμοποιούνται για τον έλεγχο της στασιμότητας μιας χρονοσειράς, όπως:

Augmented Dickey–Fuller Test (ADF)

Το Augmented Dickey–Fuller είναι ένα στατιστικό τεστ τύπου unit root test, το οποίο χρησιμοποιείται για να ελέγξει αν μια χρονοσειρά είναι στάσιμη.

$$y_t = c + \beta t + \alpha y_{t-1} + \sum_{i=1}^{p} \phi_i \Delta y_{t-i} + e_t$$

Υποθέσεις Τεστ

H_0 (Null Hypothesis): Η χρονοσειρά έχει unit root → είναι μη στάσιμη
H_1 (Alternative Hypothesis): Η χρονοσειρά είναι στάσιμη

Κριτήριο Απόφασης

Αν το p-value < επίπεδο σημαντικότητας (π.χ. 0.05)
Απορρίπτουμε την H_0 → Η χρονοσειρά θεωρείται στάσιμη.

Ανάλυση Χρονοσειράς: Διάγνωση Τάσης (II)

Το **Mann-Kendall Test** είναι ένα μη παραμετρικό στατιστικό τεστ (δεν απαιτεί υπόθεση κανονικότητας των δεδομένων), το οποίο χρησιμοποιείται για τον εντοπισμό ύπαρξης τάσης σε μια χρονοσειρά.
Βασίζεται στη σύγκριση της σχετικής τάξης των παρατηρήσεων και μπορεί να ανιχνεύσει μονοτονικές τάσεις, είτε αυξητικές είτε φθίνουσες.

Υποθέσεις Τεστ

H_0 (Null Hypothesis): Δεν υπάρχει τάση στα δεδομένα.
H_1 (Alternative Hypothesis): Υπάρχει τάση στα δεδομένα (θετική ή αρνητική).

Κριτήριο Απόφασης

Αν το p-value < επίπεδο σημαντικότητας (π.χ. 0.05)
Απορρίπτουμε την H_0 → Υπάρχει στατιστικά σημαντική τάση στη χρονοσειρά.

Hands-on 3: Έλεγχος Στασιμότητας σε Μονομεταβλητό και Πολυμεταβλητό Σύνολο Δεδομένων

Να εφαρμοστεί το στατιστικό τεστ Augmented Dickey–Fuller (ADF) με σκοπό τον έλεγχο της στασιμότητας χρονοσειρών τόσο σε μονομεταβλητό όσο και σε πολυμεταβλητό σύνολο δεδομένων. Ειδικότερα,

- Να εφαρμοστεί το τεστ **ADF** στη χρονοσειρά του **μονομεταβλητού συνόλου δεδομένων.**
- Να εφαρμοστεί το τεστ ADF σε κατάλληλα επιλεγμένη μεταβλητή του **πολυμεταβλητού συνόλου δεδομένων**.
- Να πραγματοποιηθεί ερμηνεία των αποτελεσμάτων και να εξαχθεί συμπέρασμα σχετικά με τη στασιμότητα των χρονοσειρών, με **βάση επίπεδο σημαντικότητας 5%.**

Ανάλυση Χρονοσειράς: Διάγνωση Εποχικότητας

Υπάρχουν διάφορες μέθοδοι για να ελέγξουμε αν μια χρονοσειρά παρουσιάζει εποχικότητα. Μία απλή και ιδιαίτερα χρήσιμη προσέγγιση είναι η οπτική διερεύνηση με ειδικά διαγράμματα.

Seasonal Subseries Plot (Οπτικοποίηση εποχικών υποσειρών)

Σε αυτήν τη μέθοδο, η χρονοσειρά χωρίζεται σε εποχικά τμήματα (π.χ. ανά μήνα ή ανά εβδομάδα) και οι τιμές του ίδιου "εποχικού σημείου" συγκρίνονται διαχρονικά (π.χ. όλοι οι Ιανουάριοι μεταξύ τους).

- Αν παρατηρείται σταθερό επαναλαμβανόμενο μοτίβο στις αντίστοιχες εποχικές υπο-σειρές, τότε υπάρχει ένδειξη εποχικότητας.
- Αν δεν υπάρχει σταθερότητα στο μοτίβο, η εποχικότητα είναι ασθενής ή ανύπαρκτη.

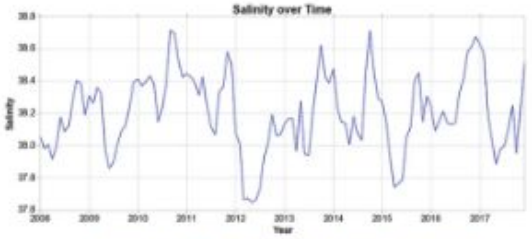


Figure 10: Time series data per year

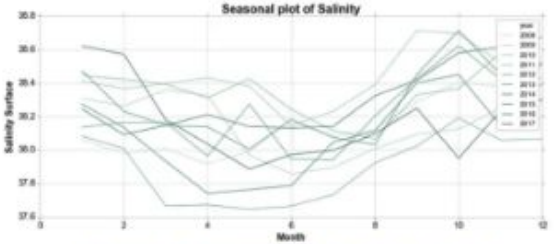


Figure 10: Seasonal subseries plot to detect seasonality per year

Ανάλυση Χρονοσειράς: Διάγνωση Εποχικότητας (II)

Συνάρτηση Αυτοσυσχέτισης (ACF)

Η ACF μετρά τη συσχέτιση μιας χρονοσειράς με καθυστερημένες εκδόσεις του εαυτού της (**lags**).

Αν παρατηρούνται σημαντικές κορυφές σε συγκεκριμένες καθυστερήσεις (ιδιαίτερα σε εποχικά lags, π.χ. 12 για μηνιαία δεδομένα ή 52 για εβδομαδιαία), τότε υπάρχει ένδειξη εποχικότητας.

Kruskal–Wallis Test

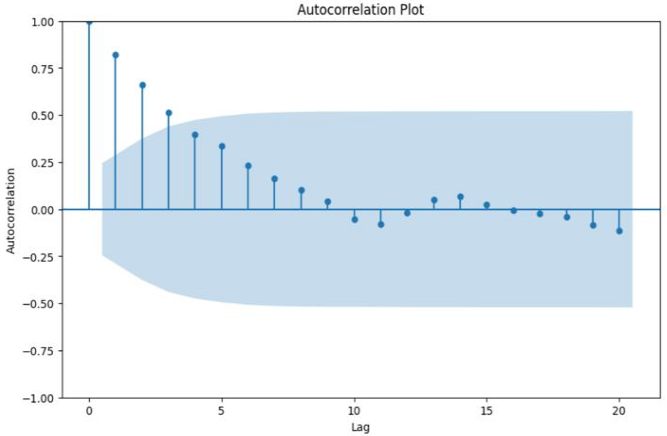
Το Kruskal–Wallis είναι ένα μη παραμετρικό στατιστικό τεστ που χρησιμοποιείται για τη σύγκριση πολλαπλών ομάδων δεδομένων.

Μπορεί να χρησιμοποιηθεί για τον έλεγχο εποχικότητας συγκρίνοντας τις τιμές της χρονοσειράς μεταξύ διαφορετικών εποχικών περιόδων (π.χ. μήνες ή εβδομάδες του έτους).

- Αν **p-value < 0.05**, απορρίπτουμε την υπόθεση ότι δεν υπάρχει εποχικότητα.
- Αν **p-value ≥ 0.05**, δεν υπάρχουν στατιστικά σημαντικές ενδείξεις εποχικότητας.

Ανάλυση Χρονοσειράς: Αυτοσυσχέτιση (Autocorrelation) Καμπυλη ACF

Η αυτοσυσχέτιση μπορεί να χρησιμοποιηθεί για τον εντοπισμό προτύπων σε μια χρονοσειρά, όπως τάση, κύκλοι και εποχικότητα, εξετάζοντας πόσο σχετίζονται οι τιμές της σειράς με προηγούμενες τιμές της.



Για τον οπτικό έλεγχο της αυτοσυσχέτισης χρησιμοποιούμε το διάγραμμα ACF (Autocorrelation Function).

Στο ACF plot:

- x-axis: η καθυστέρηση (*lag*)
- y-axis: ο συντελεστής αυτοσυσχέτισης

Ο συντελεστής αυτοσυσχέτισης κυμαίνεται από -1 έως 1:

- τιμές κοντά στο **1** υποδηλώνουν ισχυρή θετική αυτοσυσχέτιση
- τιμές κοντά στο **-1** υποδηλώνουν ισχυρή αρνητική αυτοσυσχέτιση
- τιμές κοντά στο **0** υποδηλώνουν ασθενή ή μηδενική αυτοσυσχέτιση

Ανάλυση Χρονοσειράς: Αυτοσυσχέτιση (Autocorrelation) Καμπυλη ACF (II)

Τι μας δείχνει μια καμπύλη ACF?

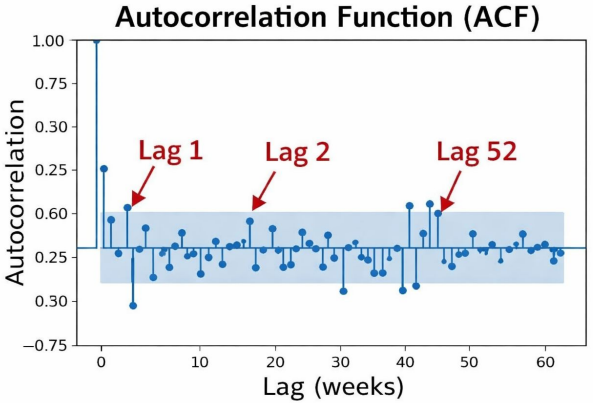
Δείχνει πόσο σχετίζονται οι τωρινές τιμές μιας χρονοσειράς με προηγούμενες τιμές της.

Βασικά στοιχεία:

Lag 1: σχέση με προηγούμενη περίοδο
Lag k: σχέση με k περιόδους πριν
Bars έξω από confidence band: **σημαντική** συσχέτιση

Βοηθάει να διαπιστωθεί αν:

(α) Υπάρχει pattern ή όχι
(β) Υπάρχει εποχικότητα (seasonality)



Significant autocorrelations indicate temoral dependence

Ανάλυση Χρονοσειράς: Αυτοσυσχέτιση (Autocorrelation) Καμπυλη Partial-ACF (PACF)

Τι μας δείχνει μια καμπύλη PACF?

Άμεση (καθαρή) σχέση παρόντος με παρελθόν

Βασικά στοιχεία:

Lag 1: **άμεση** σχέση με προηγούμενη περίοδο

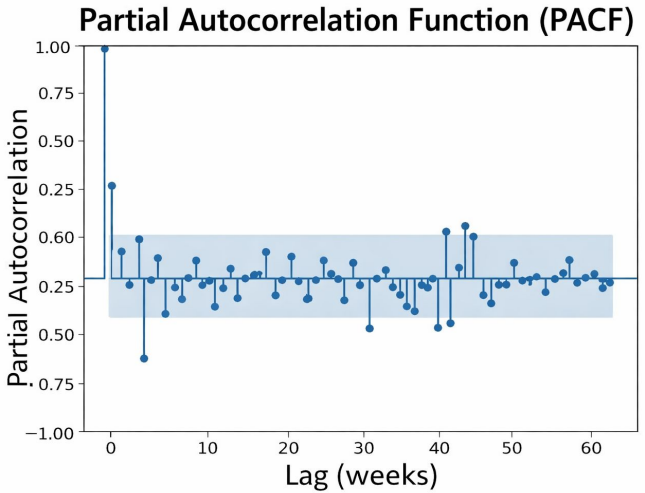
Lag k: μόνο άμεση σχέση με k lag

Bars έξω από band: σημαντική άμεση συσχέτιση

Βοηθάει να διαπιστωθεί αν:

(α) PACF βοηθά να εντοπιστεί ο αριθμός των Autoregressive (AR) terms που χρειάζεται το μοντέλο.

(β) Δείχνει ποιες προηγούμενες περίοδοι επηρεάζουν άμεσα το παρόν.



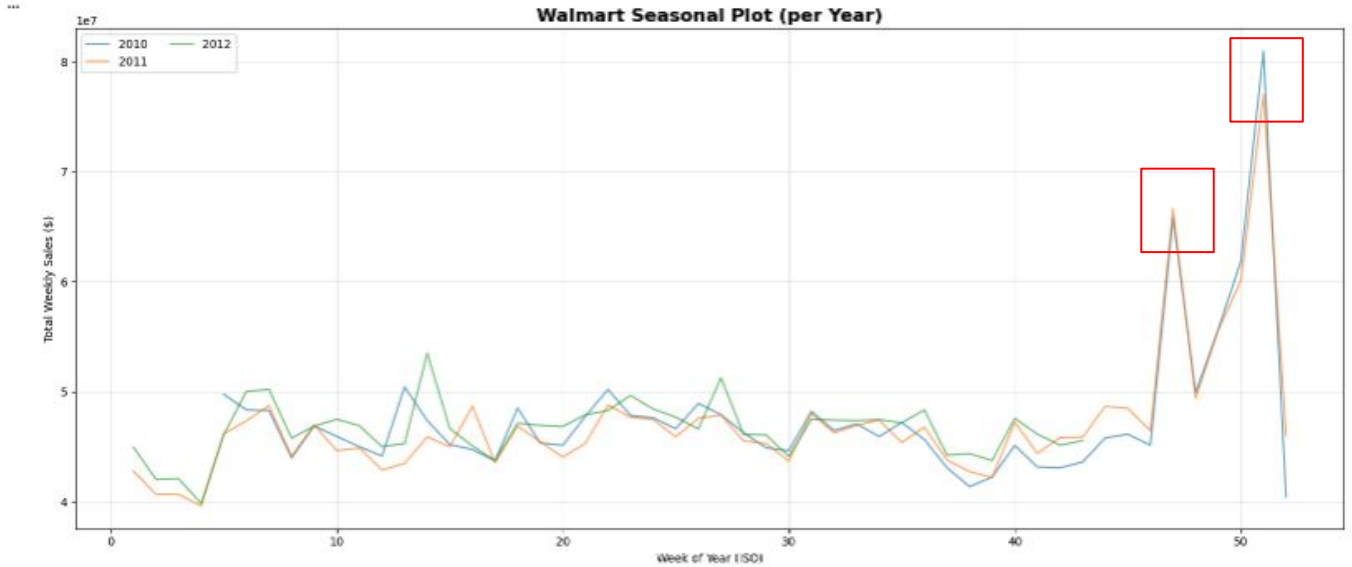
Hands-on 4 — Διερεύνηση Εποχικότητας με Οπτικοποίηση, ACF και Στατιστικό Τεστ στο πολυμεταβλητό σύνολο δεδομένων χρονοσειράς.

Να πραγματοποιηθεί διερεύνηση της εποχικότητας σε χρονοσειρά δεδομένων πωλήσεων μιας εταιρείας με δίκτυο καταστημάτων, χρησιμοποιώντας συνδυασμό οπτικών και στατιστικών μεθόδων ανάλυσης.

Συγκεκριμένα, ζητείται:

- Να προετοιμαστεί η χρονοσειρά συνολικών εβδομαδιαίων πωλήσεων μέσω κατάλληλης συγκέντρωσης (aggregation) των δεδομένων ανά χρονική στιγμή.
- Να πραγματοποιηθεί οπτική διερεύνηση της εποχικότητας μέσω κατάλληλων διαγραμμάτων (π.χ. seasonal plot ή seasonal subseries plot), με στόχο τον εντοπισμό επαναλαμβανόμενων μοτίβων μέσα στο έτος.
- Να εφαρμοστεί η Συνάρτηση Αυτοσυσχέτισης (ACF) για τον εντοπισμό πιθανών εποχικών καθυστερήσεων (lags), με έμφαση στον εντοπισμό επαναλαμβανόμενων μοτίβων σε εποχικές περιόδους.
- Να εφαρμοστεί κατάλληλο στατιστικό τεστ εποχικότητας (π.χ. Kruskal–Wallis test), συγκρίνοντας τις τιμές της χρονοσειράς μεταξύ διαφορετικών εποχικών περιόδων (π.χ. εβδομάδες του έτους).
- Να εφαρμοστει η Συνάρτηση μερικής αυτοσυσχέτισης και να εξαχθούν συμπεράσματα. Να παρουσιαστούν και να ερμηνευθούν τα αποτελέσματα κάθε μεθόδου και να εξαχθεί συνολικό συμπέρασμα σχετικά με την ύπαρξη ή μη εποχικότητας στη χρονοσειρά.

Ανάλυση Χρονοσειράς: Συμπεράσματα (I)



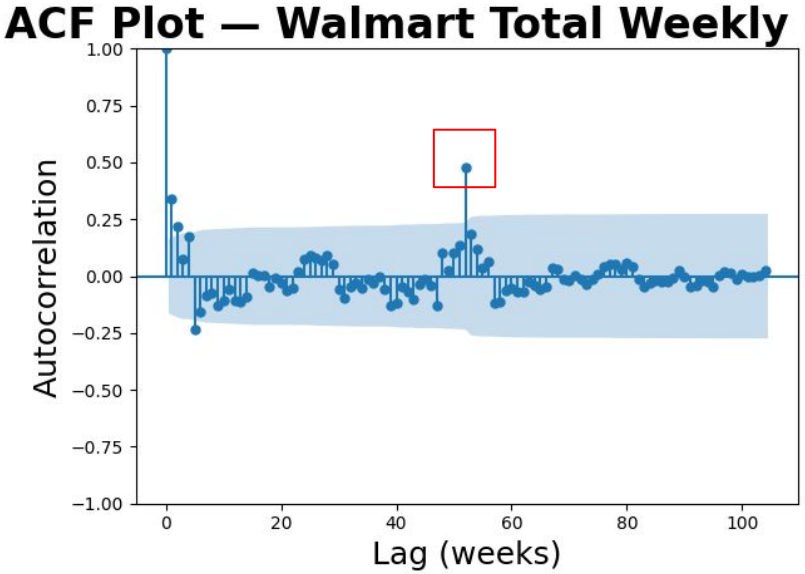
Παρόμοιο μοτίβο πωλήσεων εμφανίζεται κάθε χρόνο == ένδειξη σταθερής εποχικότητας.

Πολύ έντονη αύξηση πωλήσεων στις τελευταίες εβδομάδες του έτους (holiday season).

Σημαντική πτώση πωλήσεων αμέσως μετά το τέλος του έτους.

Σχετικά σταθερό επίπεδο πωλήσεων στο μεγαλύτερο μέρος του έτους.

Ανάλυση Χρονοσειράς: Συμπεράσματα (II)



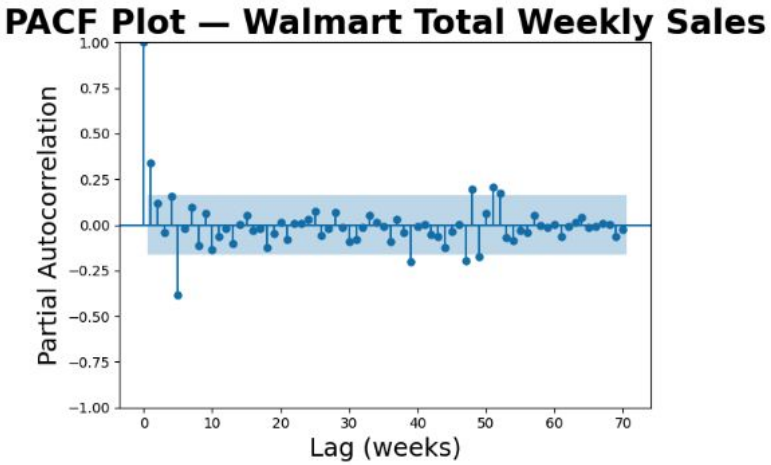
Υπάρχει ισχυρή κορυφή στο lag ≈ 52, που δείχνει ετήσια εποχικότητα στις εβδομαδιαίες πωλήσεις.

Οι περισσότερες άλλες καθυστερήσεις (lags) έχουν χαμηλή συσχέτιση, άρα δεν υπάρχει ισχυρή μακροχρόνια εξάρτηση.

Το spike στο lag 52 βρίσκεται εκτός του confidence band, άρα είναι στατιστικά σημαντικό.

Το μοτίβο υποδηλώνει επαναλαμβανόμενη ετήσια συμπεριφορά πωλήσεων (π.χ. περίοδος εορτών).

Ανάλυση Χρονοσειράς: Συμπεράσματα (III)



- Lag 1: Ισχυρή θετική συσχέτιση
  Strong dependence από την προηγούμενη εβδομάδα
  Ένδειξη για πιθανό AR(1)
- Lag ~5: Αρνητική συσχέτιση (πιθανώς σημαντική)
  Πιθανό short-term correction effect
- Lag ~50–52: Μικρά spikes
  Πιθανή ετήσια εποχικότητα (seasonality) για weekly data

Ανάλυση Χρονοσειράς: Κατάργηση Τάσης-Εποχικότητας

Υπάρχουν διάφορες μέθοδοι για την αφαίρεση της τάσης και της εποχικότητας από δεδομένα χρονοσειρών, ανάλογα με τη φύση των δεδομένων και τους στόχους της ανάλυσης. Μερικές από τις πιο συχνά χρησιμοποιούμενες τεχνικές είναι οι εξής:

Κατάργηση Τάσης μέσω Διαφοροποίησης (Detrending by Differencing)

Μία από τις απλούστερες μεθόδους αφαίρεσης γραμμικής τάσης είναι ο υπολογισμός της πρώτης διαφοράς της χρονοσειράς.

Η μέθοδος αυτή μειώνει ή εξαλείφει τη μακροχρόνια τάση και βοηθά στη μετατροπή της σειράς σε πιο στάσιμη μορφή.

Κινούμενοι Μέσοι Όροι (Moving Average)

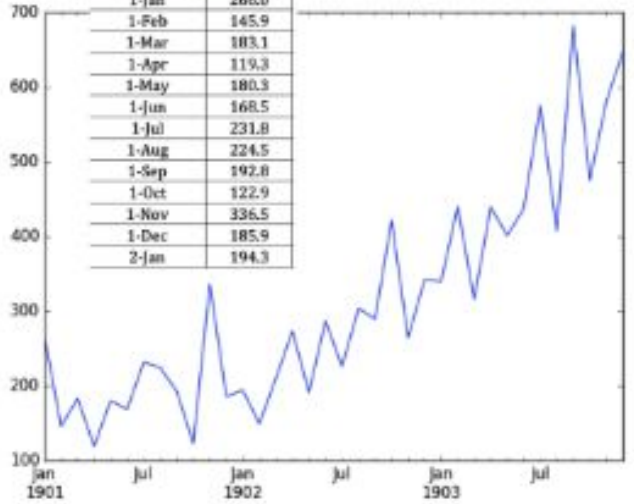
Η μέθοδος βασίζεται στον υπολογισμό του μέσου όρου ενός συγκεκριμένου αριθμού διαδοχικών παρατηρήσεων (time window) γύρω από κάθε χρονικό σημείο.

Οι κινούμενοι μέσοι όροι βοηθούν στο φιλτράρισμα των βραχυπρόθεσμων διακυμάνσεων και στην ανάδειξη της μακροχρόνιας τάσης της χρονοσειράς.

Ανάλυση Χρονοσειράς: Κατάργηση Τάσης – Παράδειγμα (1/2)

- Το σύνολο δεδομένων περιγράφει τον μηνιαίο αριθμό πωλήσεων σαμπουάν για περίοδο 3 ετών.
- Η υπό μελέτη χρονοσειρά παρουσιάζει αυξητική τάση.
- Η τάση θα αφαιρεθεί χρησιμοποιώντας τη μέθοδο της διαφοροποίησης (differencing).
- Η αφαίρεση της τάσης αποτελεί βασικό στάδιο στη διαδικασία προ-επεξεργασίας και καθαρισμού δεδομένων.

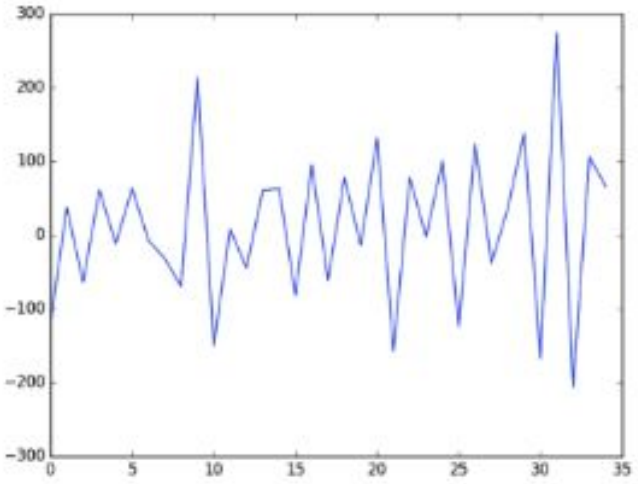
Month	Sales
1-Jan	266.0
1-Feb	145.9
1-Mar	183.1
1-Apr	119.3
1-May	180.3
1-Jun	168.5
1-Jul	231.8
1-Aug	224.5
1-Sep	192.8
1-Oct	122.9
1-Nov	336.5
1-Dec	185.9
2-Jan	194.3



Ανάλυση Χρονοσειράς: Κατάργηση Τάσης – Παράδειγμα (2/2)

Υπολογισμός Πρώτης Διαφοράς:
Η πρώτη διαφορά υπολογίζεται αφαιρώντας κάθε παρατήρηση της χρονοσειράς από την αμέσως προηγούμενη. Η διαδικασία αυτή χρησιμοποιείται συχνά για την αφαίρεση της τάσης.

Έλεγχος Στασιμότητας:
Μετά τον υπολογισμό της πρώτης διαφοράς, ελέγχεται αν η χρονοσειρά είναι στάσιμη. Μια στάσιμη χρονοσειρά χαρακτηρίζεται από σταθερή μέση τιμή και σταθερή διακύμανση με την πάροδο του χρόνου.



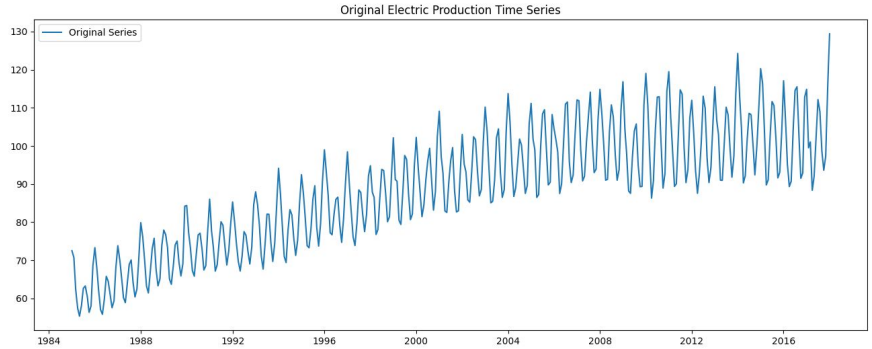
Hands-on 5: Αφαίρεση Τάσης με Πρώτη Διαφοροποίηση σε Σύνολο Δεδομένων Ενέργειας

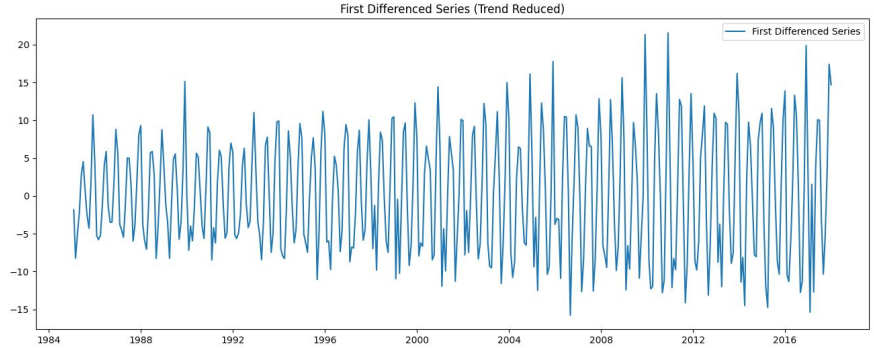
Να πραγματοποιηθεί **αφαίρεση της τάσης από μονομεταβλητό σύνολο δεδομένων** που αφορά την παραγωγή ηλεκτρικής ενέργειας, χρησιμοποιώντας τη μέθοδο της πρώτης διαφοροποίησης (first differencing).

Συγκεκριμένα, Ζητείται:

- Να προετοιμαστεί η χρονοσειρά παραγωγής ενέργειας.
- Να υπολογιστεί η **πρώτη διαφορά της χρονοσειράς.**
- Να πραγματοποιηθεί οπτική σύγκριση της αρχικής και της διαφοροποιημένης χρονοσειράς.
- Να ελεγχθεί αν η διαφοροποιημένη χρονοσειρά παρουσιάζει βελτιωμένη στασιμότητα.
- Να εξαχθούν συμπεράσματα σχετικά με την επίδραση της διαφοροποίησης στην τάση της χρονοσειράς.

Ανάλυση Χρονοσειράς: Κατάργηση Τάσης - Συμπεράσματα





Αρχική χρονοσειρά

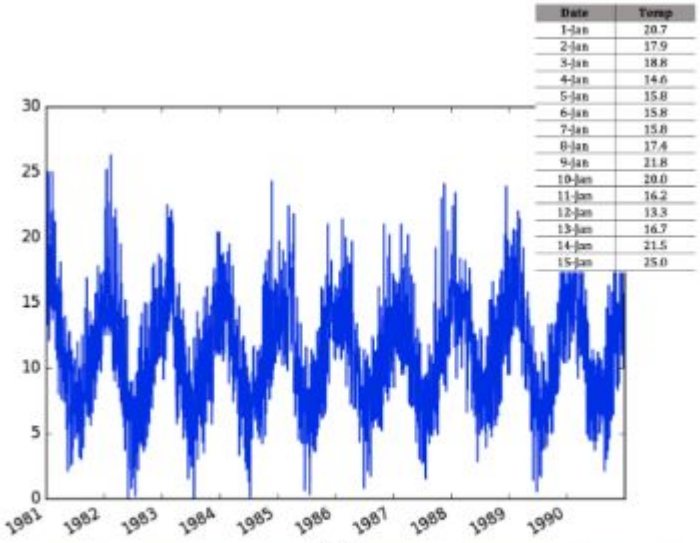
- Παρουσιάζει σαφή γραμμική τάση (ανοδική).
  Ο μέσος όρος μεταβάλλεται με τον χρόνο.
- Η σειρά δεν μπορεί να θεωρηθεί στάσιμη.

Μετά το 1st differencing

- Η τάση εξαλείφεται αποτελεσματικά.
- Η νέα χρονοσειρά ταλαντώνεται γύρω από σταθερό μέσο όρο.
- Η διακύμανση παραμένει περίπου σταθερή στον χρόνο.

Ανάλυση Χρονοσειράς: Κατάργηση Εποχικότητας – Παράδειγμα (1/2)

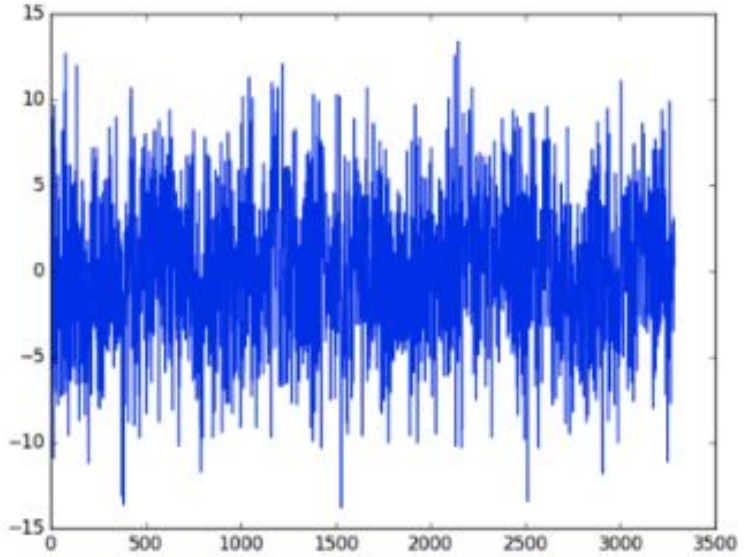
- Το σύνολο δεδομένων περιγράφει τις ελάχιστες ημερήσιες θερμοκρασίες για περίοδο 10 ετών (1981–1990) σε μια πόλη.
- Η υπό μελέτη χρονοσειρά παρουσιάζει εποχικότητα.
- Η εποχικότητα θα αφαιρεθεί χρησιμοποιώντας τη μέθοδο της διαφοροποίησης (seasonal differencing).
- Η αφαίρεση της εποχικότητας αποτελεί βασικό στάδιο στη διαδικασία προ-επεξεργασίας και καθαρισμού δεδομένων.



Ανάλυση Χρονοσειράς: Κατάργηση Εποχικότητας – Παράδειγμα (2/2)

- **Υπολογισμός πρώτης διαφοράς** και στη συνέχεια έλεγχος στασιμότητας της χρονοσειράς (όπως και στην περίπτωση της τάσης).

- Αν η εναπομείνουσα χρονοσειρά δεν παρουσιάζει πλέον τάση ή εποχικότητα, τότε η διαδικασία ανάλυσης μπορεί να προχωρήσει στο επόμενο στάδιο.



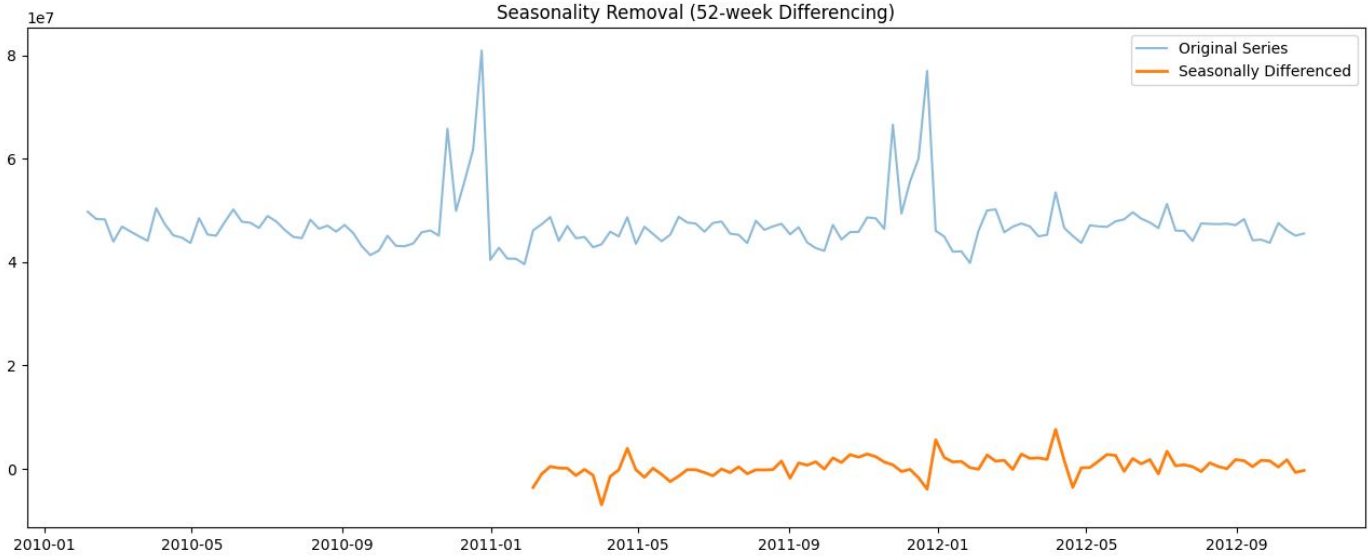
Hands-on 6: Αφαίρεση Εποχικότητας στο πολυμεταβλητό σύνολο δεδομένων

Να πραγματοποιηθεί αφαίρεση της εποχικότητας από τη χρονοσειρά των συνολικών εβδομαδιαίων πωλήσεων της εταιρείας με δίκτυο καταστημάτων.

Συγκεκριμένα, ζητείται:

- Να προετοιμαστεί η χρονοσειρά πωλήσεων
- Να εντοπιστεί η εποχική περίοδος της χρονοσειράς (π.χ. ετήσια εποχικότητα σε εβδομαδιαία δεδομένα).
- Να εφαρμοστεί κατάλληλη μέθοδος αφαίρεσης εποχικότητας, όπως η εποχική διαφοροποίηση (seasonal differencing).
- Να πραγματοποιηθεί οπτική σύγκριση της αρχικής και της εποχικά διαφοροποιημένης χρονοσειράς.
- Να εξαχθούν συμπεράσματα σχετικά με την επίδραση της αφαίρεσης εποχικότητας στη συμπεριφορά της χρονοσειράς.

Ανάλυση Χρονοσειράς: Κατάργηση Εποχικότητας - Συμπεράσματα



- Η εφαρμογή **52-week seasonal differencing αφαίρεσε αποτελεσματικά την ετήσια εποχικότητα** από τη χρονοσειρά.
- Η διαφοροποιημένη σειρά κινείται γύρω από το μηδέν, ένδειξη μείωσης συστηματικών εποχικών μοτίβων.

Ανάλυση Χρονοσειράς: Εξομάλυνση (Smoothing)

Η αφαίρεση του θορύβου, ή τουλάχιστον η μείωση της επίδρασής του, αποτελεί σημαντικό στάδιο στην ανάλυση χρονοσειρών.

- Η **εξομάλυνση (smoothing) βοηθά στη μείωση του θορύβου** και των τυχαίων διακυμάνσεων στα δεδομένα, διευκολύνοντας τον εντοπισμό βασικών μοτίβων, όπως τάση και εποχικότητα.
- **Μπορεί να μειώσει την επίδραση ακραίων τιμών (outliers),** οι οποίες ενδέχεται να επηρεάσουν αρνητικά τα αποτελέσματα της στατιστικής ανάλυσης και των μοντέλων πρόβλεψης.
- **Βελτιώνει την πρόβλεψη**, καθώς δημιουργεί ένα πιο καθαρό και λιγότερο ασταθές σήμα, οδηγώντας σε πιο αξιόπιστες εκτιμήσεις μελλοντικών τιμών.
- **Βοηθά στον εντοπισμό ανωμαλιών και ακραίων τιμών (outliers)**, οι οποίες μπορεί να παραμορφωσουν τα αποτελέσματα της ανάλυσης.

Ανάλυση Χρονοσειράς: Βασικές Μέθοδοι Εξομάλυνσης σε Χρονοσειρές

Μέθοδος	Περιγραφή	Πότε Χρησιμοποιείται
Κινούμενος Μέσος Όρος (Moving Average)	Υπολογίζει τον μέσο όρο διαδοχικών παρατηρήσεων για μείωση θορύβου.	Όταν θέλουμε να εξομαλύνουμε βραχυχρόνιες διακυμάνσεις και να δούμε τη γενική τάση.
Εκθετική Εξομάλυνση (Simple Exponential Smoothing)	Δίνει μεγαλύτερη βαρύτητα στις πιο πρόσφατες παρατηρήσεις.	Για δεδομένα χωρίς έντονη τάση ή εποχικότητα.
Holt (Double Exponential Smoothing)	Επεκτείνει την εκθετική εξομάλυνση λαμβάνοντας υπόψη και την τάση.	Για δεδομένα με τάση αλλά χωρίς έντονη εποχικότητα.
Holt–Winters (Triple Exponential Smoothing)	Λαμβάνει υπόψη τάση και εποχικότητα.	Για δεδομένα με επαναλαμβανόμενα εποχικά μοτίβα.

Ανάλυση Χρονοσειράς: Απλός Κινητός Μέσος όρος (Simple Moving Average)

Η **μέθοδος του κινούμενου μέσου όρου** βασίζεται στον υπολογισμό του μέσου όρου ενός συγκεκριμένου αριθμού διαδοχικών παρατηρήσεων μιας χρονοσειράς και στη χρήση αυτού του μέσου όρου ως εξομαλυμένης τιμής.

Με τον τρόπο αυτό **μειώνονται οι βραχυπρόθεσμες διακυμάνσεις (θόρυβος)** και γίνεται πιο εμφανής η υποκείμενη τάση.

Για την εφαρμογή της μεθόδου, απαιτείται ο καθορισμός του μεγέθους του παραθύρου (window size), δηλαδή του πλήθους των διαδοχικών παρατηρήσεων που θα συμμετέχουν σε κάθε υπολογισμό.

$$SMA_t = \frac{A_t + A_{t-1} + \cdots + A_{t-n+1}}{n}$$

n : μέγεθος παραθύρου (αριθμός χρονικών περιόδων)

A_t : τιμή της χρονοσειράς στη χρονική στιγμή t

Ανάλυση Χρονοσειράς: Απλός Κινητός Μέσος όρος (Simple Moving Average) (II)

- Αφού καθοριστεί το μέγεθος του παραθύρου, υπολογίζεται ο μέσος όρος των παρατηρήσεων που περιλαμβάνονται σε αυτό και η τιμή της χρονοσειράς αντικαθίσταται από τον αντίστοιχο μέσο όρο.

- Η διαδικασία αυτή επαναλαμβάνεται για κάθε χρονικό σημείο της χρονοσειράς, με αποτέλεσμα την παραγωγή μιας εξομαλυμένης καμπύλης.

- Ένα πιθανό μειονέκτημα της μεθόδου του απλού κινούμενου μέσου όρου είναι ότι εισάγει υστέρηση (lag), καθώς οι εξομαλυμένες τιμές βασίζονται σε προηγούμενες παρατηρήσεις.

Supplier	$	MA
1	9	
2	8	
3	9	8.667
4	12	9.667
5	9	10.000
6	12	11.000
7	11	10.667
8	7	10.000
9	13	10.333
10	9	9.667
11	11	11.000
12	10	10.000

Month	Temp. (°F)	Moving average
Jan	39	
Feb	42	
Mar	50	44
Apr	60	51
May	71	60
Jun	79	70
Jul	85	78
Aug	81	82
Sep	76	81

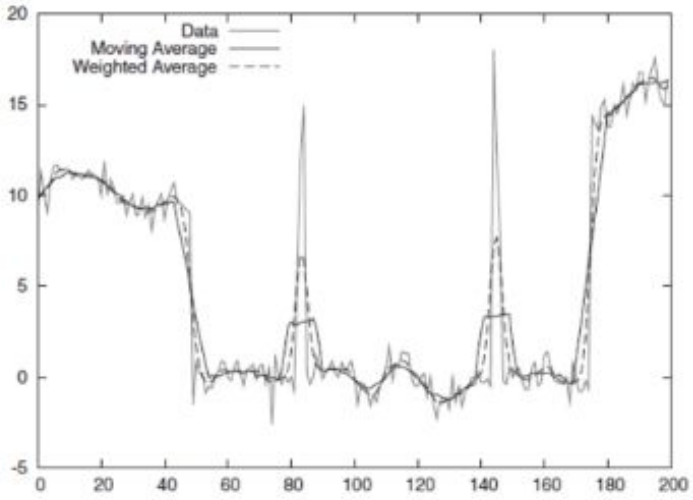
Ανάλυση Χρονοσειράς:Εξομάλυνση – Τρέχοντες μέσοι όροι

Ο **απλός κινούμενος μέσος όρος (Simple Moving Average – SMA)** αποδίδει την ίδια βαρύτητα σε όλες τις παρατηρήσεις που περιλαμβάνονται στο παράθυρο υπολογισμού.

Πρόβλημα:

Όταν μια **ακραία τιμή (outlier) εισέρχεται στο τρέχον παράθυρο** (π.χ. παράθυρο 11 παρατηρήσεων), ο κινούμενος μέσος όρος μπορεί να μεταβληθεί απότομα. Η επίδραση της ακραίας τιμής παραμένει μέχρι η τιμή αυτή να εξέλθει από το παράθυρο υπολογισμού.

Αντίθετα, **σε σταθμισμένους κινούμενους μέσους όρους (Weighted Moving Average),** οι πρόσφατες τιμές λαμβάνουν μεγαλύτερη βαρύτητα, με αποτέλεσμα η εξομαλυμένη σειρά να επηρεάζεται λιγότερο από απότομες μεταβολές ή ακραίες τιμές στα δεδομένα.



Hands-on 7: Εφαρμογή Απλού και Σταθμισμένου Κινούμενου Μέσου Όρου στο μονομεταβλητο σύνολο δεδομένων (χρονοσειρά Ενέργειας)

Να πραγματοποιηθεί εξομάλυνση της χρονοσειράς παραγωγής ηλεκτρικής ενέργειας χρησιμοποιώντας τις μεθόδους του Απλού Κινούμενου Μέσου Όρου (**Simple Moving Average — SMA**) και του Σταθμισμένου Κινούμενου Μέσου Όρου (**Weighted Moving Average — WMA)**.

Συγκεκριμένα, ζητείται:

- Να προετοιμαστεί η χρονοσειρά παραγωγής ενέργειας.
- Να εφαρμοστεί ο απλός κινούμενος μέσος όρος με κατάλληλο μέγεθος παραθύρου.
- Να εφαρμοστεί ο σταθμισμένος κινούμενος μέσος όρος, δίνοντας μεγαλύτερη βαρύτητα στις πιο πρόσφατες παρατηρήσεις.
- Να πραγματοποιηθεί οπτική σύγκριση της αρχικής χρονοσειράς με τις εξομαλυμένες σειρές.
- Να εξαχθούν συμπεράσματα σχετικά με τη διαφορά συμπεριφοράς μεταξύ SMA και WMA ως προς την εξομάλυνση και την απόκριση στις μεταβολές των δεδομένων.

Ανάλυση Χρονοσειράς:Εξομάλυνση – Σταθμισμένος Κινητός Μέσος όρος (Gaussian Weighted Moving Average)

Ο **Gaussian Weighted Moving Average** είναι μια μορφή **σταθμισμένου κινούμενου μέσου όρου (WMA)**, όπου τα βάρη προκύπτουν από Gaussian (κανονική) κατανομή.

Βασική Ιδέα

- Τα σημεία κοντά στο κέντρο του παραθύρου έχουν μεγαλύτερο βάρος.
- Τα σημεία στις άκρες έχουν μικρότερο βάρος.
- Τα βάρη είναι συμμετρικά και αθροίζουν σε 1.

$$s_i = \sum_{j=-k}^{k} w_j x_{i+j} \sum_{j=-k}^{k} w_j = 1$$

Γιατί το χρησιμοποιούμε;

- Μειώνει τον θόρυβο χωρίς απότομες αλλαγές (no sharp window edges).
- Είναι λιγότερο ευαίσθητο σε ακραίες τιμές σε σχέση με τον SMA.
- Παρέχει πιο “ομαλή” και φυσική εξομάλυνση.

Ανάλυση Χρονοσειράς: Εξομάλυνση – Η εκθετική εξομάλυνση (Exponential Smoothing - Holt Winters Method)

Όλες οι μέθοδοι κινούμενου μέσου όρου **παρουσιάζουν ορισμένους περιορισμούς:**

- Δυσκολία στην αντικειμενική αξιολόγηση της καταλληλότητας του παραθύρου.
- Ευαισθησία στην παρουσία ακραίων τιμών (outliers).
- Περιορισμένη ικανότητα ανίχνευσης απότομων ή δομικών αλλαγών στη χρονοσειρά.
- Πιθανή απώλεια πληροφορίας λόγω υπερβολικής εξομάλυνσης.

Η εκθετική εξομάλυνση αποτελεί εναλλακτική προσέγγιση, η οποία δίνει μεγαλύτερη βαρύτητα στις πιο πρόσφατες παρατηρήσεις.

- **Απλή Εκθετική Εξομάλυνση (Single Exponential Smoothing):** Χρησιμοποιείται για χρονοσειρές χωρίς τάση και χωρίς εποχικότητα.
- **Διπλή Εκθετική Εξομάλυνση (Double Exponential Smoothing — Holt's Linear Model):** Χρησιμοποιείται για χρονοσειρές που εμφανίζουν τάση αλλά όχι εποχικότητα.
  Τριπλή Εκθετική Εξομάλυνση (Triple Exponential Smoothing — Holt–Winters Model): Χρησιμοποιείται για χρονοσειρές που εμφανίζουν τόσο τάση όσο και εποχικότητα.

Ανάλυση Χρονοσειράς: Εισαγωγή στα αυτοπαλίνδρομα μοντέλα (Autoregressive models)

Τα αυτοπαλίνδρομα μοντέλα είναι μοντέλα χρονοσειρών όπου η τωρινή τιμή εξαρτάται από προηγούμενες τιμές της ίδιας μεταβλητής. (*π.χ, Πωλήσεις σήμερα σχετίζονται με τις πωλήσεις προηγούμενων εβδομάδων)*

Σημαντικό στοιχείο: Η παρούσα τιμή εξηγείται από προηγούμενες τιμές της ίδιας σειράς.

Βασική μορφή:

$$\mathbf{AR(1)\ Model}\ Y_t = c + \phi Y_{t-1} + \epsilon_t$$

Y□: τιμή στη χρονική στιγμή t
c : σταθερός όρος
φ: πόσο επηρεάζει η προηγούμενη τιμή την τωρινή
ε□: τυχαίο σφάλμα (white noise)

Τι σημαίνει AR(p)

AR(1): εξάρτηση από 1 προηγούμενη περίοδο
AR(p): εξάρτηση από p προηγούμενες περιόδους

Ανάλυση Χρονοσειράς: Εισαγωγή στα αυτοπαλίνδρομα μοντέλα (Autoregressive models) (II)

- Τα στατιστικά μοντέλα χρονοσειρών χρησιμοποιούνται για την πραγματοποίηση προβλέψεων βασισμένων σε ιστορικά δεδομένα.

- Συνήθως είναι κατάλληλα για σχετικά απλές χρονοσειρές ή για σύνολα δεδομένων με περιορισμένο αριθμό παρατηρήσεων.

- Για την εφαρμογή κάθε μοντέλου απαιτείται η ικανοποίηση συγκεκριμένων προϋποθέσεων που σχετίζονται με τα χαρακτηριστικά της χρονοσειράς, όπως η στασιμότητα, η παρουσία τάσης και η ύπαρξη εποχικότητας.

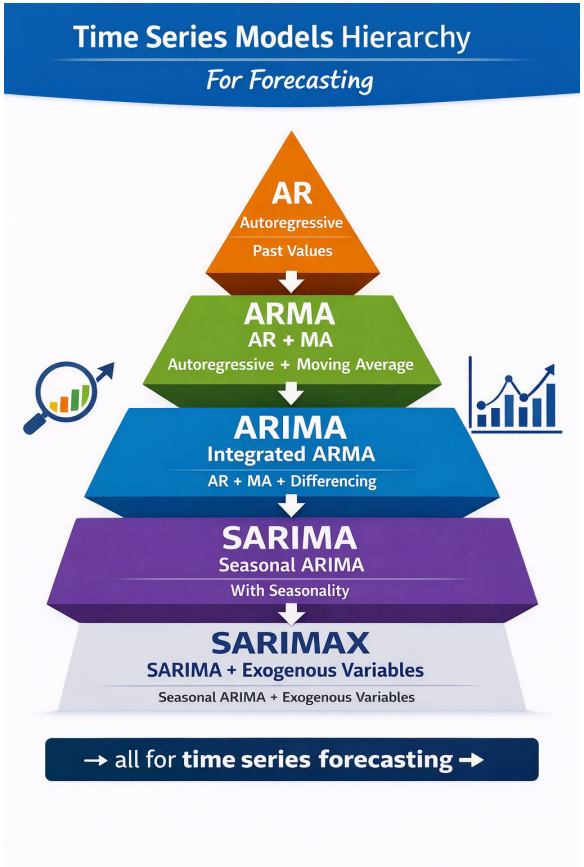
- Ο καθορισμός των παραμέτρων των μοντέλων βασίζεται σε αποτελέσματα οπτικοποίησης και στατιστικής ανάλυσης, όπως τα διαγράμματα **ACF** (Autocorrelation Function), **PACF** (Partial Autocorrelation Function) και στατιστικά τεστ όπως το **ADF** (Augmented Dickey-Fuller test).

Ανάλυση Χρονοσειράς: Αυτοπαλινδρομικά Μοντέλα Απαιτήσεις

- **Καθαρά δεδομένα** (Χωρίς ελλείπουσες τιμές και σε σωστή χρονική σειρά με σταθερη συχνότητα (π.χ. weekly))
- **Επαρκές μέγεθος δείγματος** (π.χ το μοντέλο ARIMA χρειάζεται συνήθως ~50–100+ observations)
- **Exploratory Analysis (EDA)** (Αρχική οπτικοποίηση Time series, ACF plot, PACF plot)
- **Διαχωρισμός δεδομένων** Train – Test Split (χρονικό, όχι random) Train = παρελθόν και Test = future
- **Αξιολόγηση προβλεψης:** Δείκτες μέτρησης: MAE, RMSE, MAPE

Ανάλυση Χρονοσειράς: Κύρια Στατιστικά Μοντέλα Χρονοσειρών

- **AR (Autoregressive)** Η τωρινή τιμή εξαρτάται από προηγούμενες τιμές της ίδιας σειράς.
- **MA (Moving Average)** Η τωρινή τιμή εξαρτάται από προηγούμενα σφάλματα (errors).
- **ARMA (Autoregressive Moving Average):** Συνδυάζει AR + MA.
- **ARIMA (Autoregressive Integrated Moving Average):** AR + MA + Differencing (για stationarity).
- **SARIMA (Seasonal ARIMA):** ARIMA + Seasonality component.
- **SARIMAX:** SARIMA + εξωτερικές μεταβλητές (exogenous variables - external factors επηρεάζουν τη σειρά (π.χ. promotions, holidays)



Πίνακας Σύγκρισης Μοντέλων Χρονοσειρών

Μοντέλο	Χαρακτηριστικά	Παράμετροι	Πότε το χρησιμοποιούμε
ARMA	- Συνδυασμός Αυτοπαλίνδρομου μοντέλου (AR) & Κινητού Μέσου Όρου (MA)	p, q	Όταν η σειρά είναι στασιμή, χωρίς τάση ή εποχικότητα
ARIMA	- Όπως το ARMA, αλλά με διαφοροποίηση (I) για αφαίρεση τάσης	p, d, q	Όταν υπάρχει τάση, αλλά όχι εποχικότητα
SARIMA	- Επέκταση του ARIMA με εποχικά στοιχεία	p, d, q, P, D, Q, s	Όταν υπάρχει τάση και εποχικότητα
SARIMAX	- Όπως το SARIMA, αλλά με εξωτερικές μεταβλητές (exog)	p, d, q, P, D, Q, s, exog	Όταν υπάρχει εποχικότητα και επηρεάζεται από άλλες μεταβλητές

Ανάλυση Χρονοσειράς: Αυτοπαλινδρομικά Μοντέλα Απαιτήσεις - Διαχείριση ελλειπουσών τιμών (Ι)

Ένα συχνό φαινόμενο σε πραγματικά σύνολα δεδομένων χρονοσειρών ειναι οι ε**λλειπούσες τιμές.**

Πώς προκύπτουν:

- **MCAR (Missing Completely at Random):** Τα δεδομένα λείπουν τελείως τυχαία, Δεν σχετίζονται ούτε με observed ούτε με missing values
- **MAR (Missing at Random):** Τα missing values σχετίζονται με observed δεδομένα, Δεν σχετίζονται με την ίδια την τιμή που λείπει
- **MNAR (Missing Not at Random):** Τα missing values σχετίζονται με την ίδια την missing τιμή (Δεν είναι τυχαίο ότι λείπει, λείπει επειδή είναι μεγάλη / μικρή / ευαίσθητη τιμή.)

Ανάλυση Χρονοσειράς: Αυτοπαλινδρομικά Μοντέλα Απαιτήσεις - Διαχείριση ελλειπουσών τιμών (II)

Γιατί είναι πρόβλημα:

- Μπορούν να αλλοιώσουν τα στατιστικά αποτελέσματα
- Μπορούν να επηρεάσουν training μοντέλων
- Πολλά μοντέλα δεν δέχονται NaN values

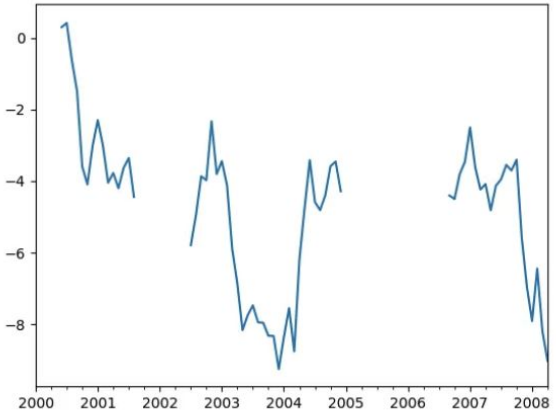
Μέθοδοι Διαχείρισης

- **Διαγραφή (Deletion)** - Drop rows με missing values (λίγες παρατηρήσεις)
- **Αντικατάσταση (Imputation)**
  - με περιγραφική στατιστικη: (Mean / Median / Mode)
  - έτερες μέθοδοι: Forward fill (*ffill*) Backward fill (*bfill*), Interpolation, Machine Learning (knn algorithm)

Ανάλυση Χρονοσειράς: Αυτοπαλινδρομικά Μοντέλα Απαιτήσεις - Διαχείριση ελλειπουσών τιμών

	Site Num	Latitude	Longitude	Sample Measurement	Date	Humidity	Pressure	Temperature	Weather_descr	Wind_dir	Wind_speed
0	1	41.670992	-87.732457	17.7	2013-01-01 00:00:00	NaN	1024.0	-0.19	16.0	200.0	4.0
1	1	41.670992	-87.732457	14.6	2013-01-01 01:00:00	64.0	1022.0	0.28	0.0	180.0	3.0
2	1	41.670992	-87.732457	13.5	2013-01-01 02:00:00	69.0	1022.0	0.33	16.0	190.0	6.0
3	1	41.670992	-87.732457	11.9	2013-01-01 03:00:00	NaN	1021.0	0.12	16.0	190.0	7.0
4	1	41.670992	-87.732457	10.3	2013-01-01 04:00:00	68.0	1021.0	0.04	0.0	210.0	7.0
5	1	41.670992	-87.732457	8.4	2013-01-01 05:00:00	68.0	1020.0	-0.04	16.0	200.0	7.0
6	1	41.670992	-87.732457	4.9	2013-01-01 06:00:00	NaN	NaN	0.00	22.0	63.0	3.0
7	1	41.670992	-87.732457	3.7	2013-01-01 07:00:00	68.0	1018.0	-0.15	16.0	200.0	4.0
8	1	41.670992	-87.732457	5.7	2013-01-01 08:00:00	NaN	1018.0	0.60	0.0	210.0	6.0
9	1	41.670992	-87.732457	7.2	2013-01-01 09:00:00	NaN	1018.0	0.52	0.0	210.0	5.0
10	1	41.670992	-87.732457	6.9	2013-01-01 10:00:00	69.0	1017.0	0.83	16.0	220.0	6.0
11	1	41.670992	-87.732457	5.7	2013-01-01 11:00:00	64.0	1017.0	1.01	16.0	230.0	6.0
12	1	41.670992	-87.732457	5.7	2013-01-01 12:00:00	55.0	1017.0	1.29	16.0	240.0	7.0
13	1	41.670992	-87.732457	6.8	2013-01-01 13:00:00	59.0	1017.0	1.20	0.0	240.0	5.0
14	1	41.670992	-87.732457	6.9	2013-01-01 14:00:00	59.0	1017.0	1.34	16.0	230.0	5.0

Outliers in time-series dataset



Outliers in time-series dataset with plot

Ανάλυση Χρονοσειράς: Μετρικές Αξιολόγησης Προβλέψεων σε Μοντέλα Χρονοσειρών

Γιατι χρειάζονται:
Η αξιολόγηση της ακρίβειας των προβλέψεων που παράγουν τα μοντέλα χρονοσειρών, μέσω ποσοτικών μετρικών σφάλματος.

- Mean Absolute Error (**MAE**)
  Υπολογίζει τον μέσο όρο των απόλυτων σφαλμάτων.
- Root Mean Squared Error (**RMSE**)
  Δίνει μεγαλύτερο βάρος στα μεγάλα σφάλματα.
- Mean Absolute Percentage Error (MAPE)
  Μετρά το σφάλμα ως ποσοστό της **πραγματικής** τιμής.
- Akaike Information Criterion (**AIC**) / Bayesian Information Criterion (**BIC**)
  Δείχνουν πόσο καλά προσαρμόζεται το μοντέλο στα δεδομένα

Hands-on 8: Εφαρμογή αυτοπαλίνδρομων στατιστικών μοντέλων

Στο μονομεταβλητό σύνολο δεδομένων, να εφαρμοστεί το καταλληλότερο αυτοπαλίνδρομο στατιστικό μοντέλο για πρόβλεψη.

Ζητούμενα:

1. Να φορτωθεί η χρονοσειρά και να χρησιμοποιηθεί **η stationary εκδοχή της** (μετά την αφαίρεση τάσης).
2. Να πραγματοποιηθεί **train/test split** με βάση τη χρονική σειρά.
3. Να εκπαιδευτούν και να συγκριθούν τα ακόλουθα μοντέλα:
   - **AR(p)**
   - **MA(q)**
   - **ARMA(p, q)**
4. Η σύγκριση των μοντέλων να γίνει με βάση:
   - **AIC / BIC** (in-sample αξιολόγηση)
   - **MAE / RMSE** στο test set (out-of-sample αξιολόγηση)
5. Να επιλεγεί το βέλτιστο μοντέλο και να παρουσιαστεί η **πρόβλεψη (forecast)** στο test set.

Ερωτήσεις

EuroHPC Joint Undertaking
Co-funded by the European Union
AI Factories PHAROS GREECE
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Ψηφιακής Διακυβέρνησης και Τεχνητής Νοημοσύνης