Ανάλυση Δεδομένων

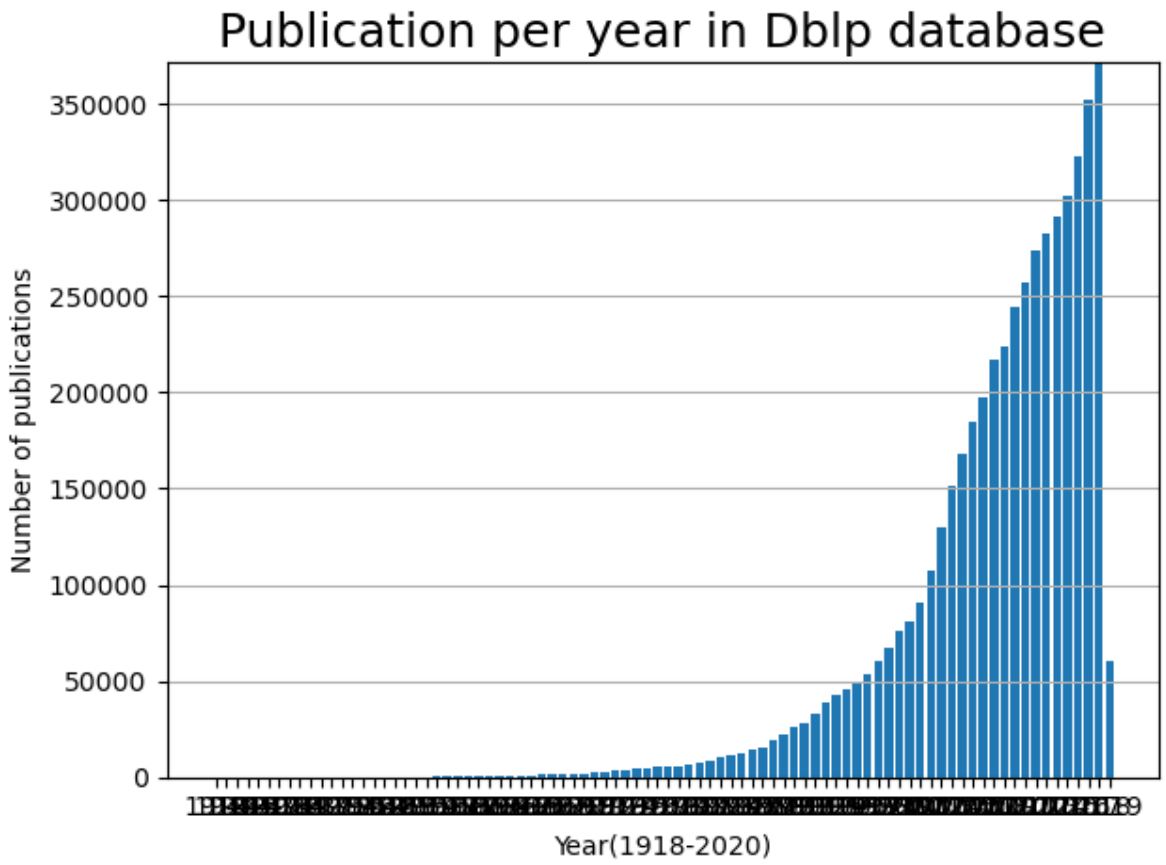
1η Εργασία

Βήμα 1ο:

Κατέβασα το αρχείο “dblp-2020-04-04.xml.gz” από τον σύνδεσμο https://dblp.org/xml/release/ .

Βήμα 2ο:

Χρησιμοποίησα κώδικα σε γλώσσα python, για να πάρω τα δεδομένα που χρειάζονται για την συγκεκριμένη ανάλυση.Χρησιμοποίησα ένα sax parser,για να πάρω τα στοιχεία που ήθελα απο το αρχείο xml.Όποτε διαβάζει κάποιον χρόνο τον προσθέτει σε ένα dictionary και θέτει έναν μετρητή γαι τον συγκεκριμένο χρόνο.Αν ο χρόνος υπάρχει ήδη αυξάνει τον μετρητή.Όταν τειειώσει αυτή η διαδικασία και ο μετρητής έχει φτάσει τον αριθμό 5015194(ο συνολιός αριθμός δημοσιεέσεων που βρήκα),εμφανίζει ταξινομημένα τα στοιχεία του dictionary.Δηλαδή το πλήθος δημοσιεύσεων ανά έτος.Στο ίδιο αρχείο κώδικα έγινε και η οπτικοποίηση του διαγράμματος των δημοσιεύσεων ανά έτος,με χρήση της βηβλιοθήκης matplotlib.

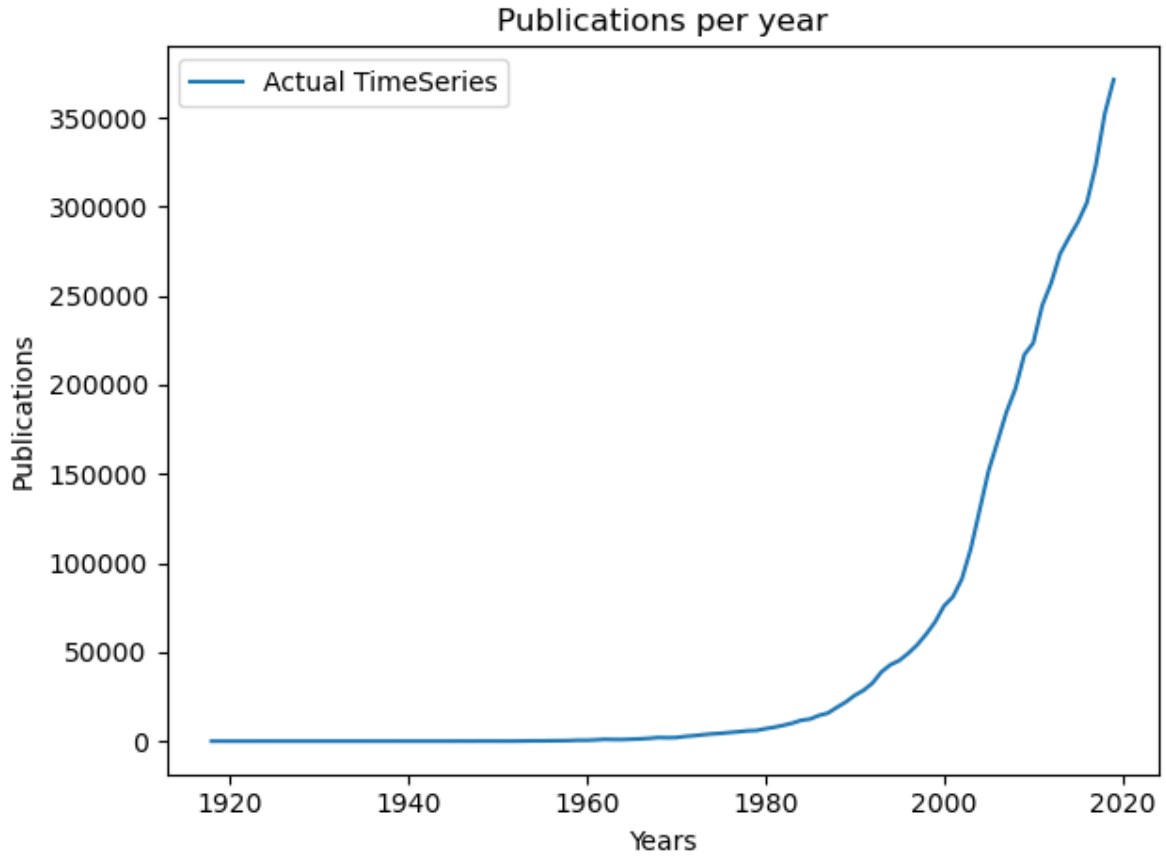


Βήμα 3ο:

Χρησιμοποιώ πάλι γλώσσα python για να κάνω αναπαράσταση της χρονοσειράς με τα δεδομένα που βρήκα πριν.(Χρησιμοποιούνται οι εξής βιβλιοθήκες:matplotlib,Pandas).

Απο την ανάλυση της χρονοσειράς εξήγαγα τα δεδομένα του 2020,γιατί τα δεδομένα του είναι μέχρι τον Απρίλιο.Επειδή λοιπόν δεν είναι ολοκληρωμένο έτος θεώρησα,οτι θα προστεθεί ένα λάθος στοιχείο στην χρονοσειρά δεδομένο οτι τα υπόλοιπα στοιχεία αναφέρονται ανά ολοκληρωμένο έτος.

Χρονοσειρά:



Χρήση διπλής εκθετικής εξομάλυνσης:

Βλέπουμε οτι η συγκεκριμένη χρονοσειρά έχει τάση,και φαίνεται οτι δεν έχει εποχικότητα.

Ο κώδικας που χρησιμοποίησα βρίσκεται στο αρχείο(Diplh_ekthetiki_exomalunsh.py)

Έστω οτι έχουμε τα μαθηματικά μοντέλα :

$P_t = a * x_t + (1-a) * (P_{t-1} + b_{t-1})$,a$\in(0,1)$

b_t =γ$* (P_t - P_{t-1}) + (1-\gamma) * b_{t-1}, \gamma \in (0,1)$

Όπου το άλφα και το γάμμα είναι οι συντελεστές εξομάλυνσης και παίρνουν τιμές απο το 0 έως το 1.Πρέπει να βρούμε τις τιμές του άλφα και του γάμμα.Αυτό θα γίνει δοκιμάζοντας

όλες τις τιμές σε αυτό το πεδίο και στις δύο μεταβλητές και θα χρησιμοποιήσουμε αυτές που δίνουν το μικρότερο μέσο τετραγωνικό σφλάμα,με τον παρακάτω τύπο.

$$P_{t+m} = P_{t+m} * b_t$$

Αυτό που παρατηρώ είναι οτι αν χρησιμοποιήσω την χρονοσειρά απο την αρχή,η πρόβλεψη είναι μικρότερη απο την πραγματική.Φαίνεται λογικό δεδομένο οτι μόλις το 1983 οι δημοσιεύσεις ξεπέρασαν τις 10 χιλιάδες.Θα μπορούσε να χρησιμοποιηθεί ο μέσος αλλά όπως θα φανεί και σε επόμενο διάγραμμα,λόγω της ραγδαίας αύξησης με την πάροδο τον χρόνων,τήνει στην ίδια πορεία με παρόμοιες τιμές.

Σε μία περίπτωση που θα πάρουμε δεδομένα απο το 2004 μέχρι το 2012 έχουμε το εξής αποτέλεσμα:

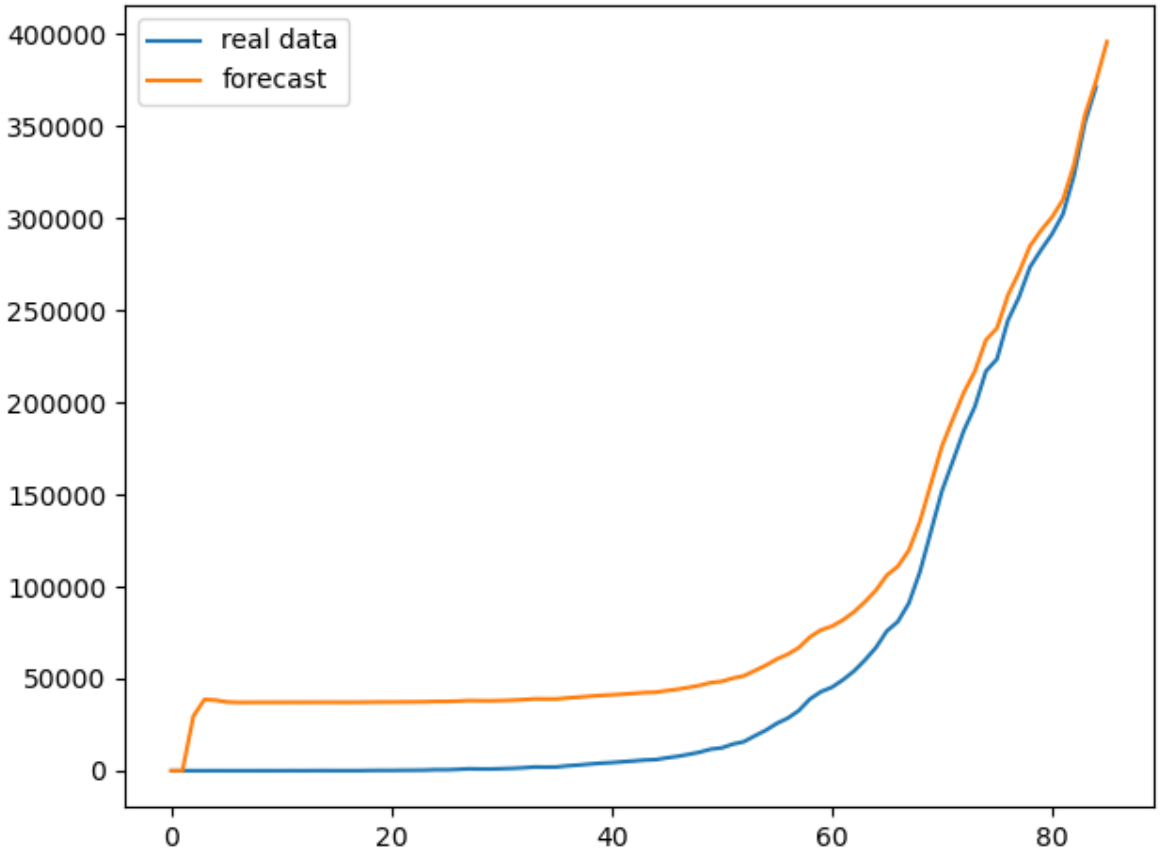
```
Best MSE = 173195.76427250472
Optimal alpha = 0.9
Optimal gamma = 0.1
P_t = 257510.9327773599
b_t= 13415.852463813297
Next observation = 270926.78524117323
```

Από την εξαγωγή των δεδομένων που κάναμε μπορούμε να παρατηρήσουμε ότι η πρόβλεψη είναι αρκετά κοντά στον πραγματικό αριθμό δημοσιεύσεων για το 2013 που είναι 273452.

Με την ίδια λογική μπορούμε να κάνουμε πρόβλεψη για το 2020.Ο αριθμός που προβλέπεται είναι 381398.

Αν τρέξουμε τον αλγόριθμο για όλες τις δημοσιεύσεις(απο το 1918-2019) ή πρόβλεψει που παίρνουμε είναι 372069.Έχουμε και το αντίστοιχο διάγραμμα.

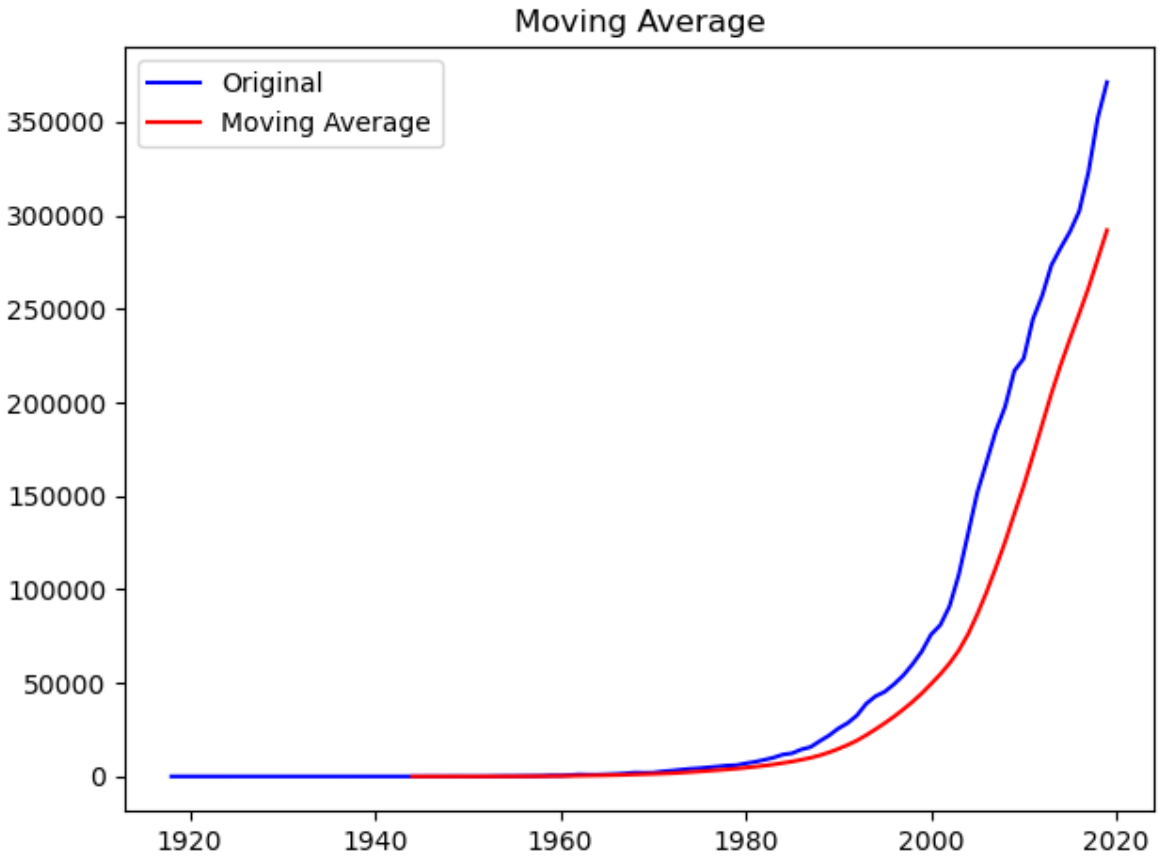
Στον άξονα Χ αναγράφεται ο αριθμός των ετών και όχι το έτος.



Επίσης προσπάθησα να χρησιμοποιήσω το μοντέλο ARIMA για να κάνω πρόβλεψη.

Χρησιμοπόιησα το ADF(Augmented Dickey – Fuller Test) για να καθορίσω αν η χρονοσειρά είναι στάσιμη.΄Εγινε χρήση της βιβλιοθήκης statsmodels.tsa.statools.Ο κώδικας για το μοντέλο αυτό βρίσκεται στο αρχείο Arima.py.

Αρχίκα βρήκα τον κινητό μέσο όρο 10 παρατηρήσεων.



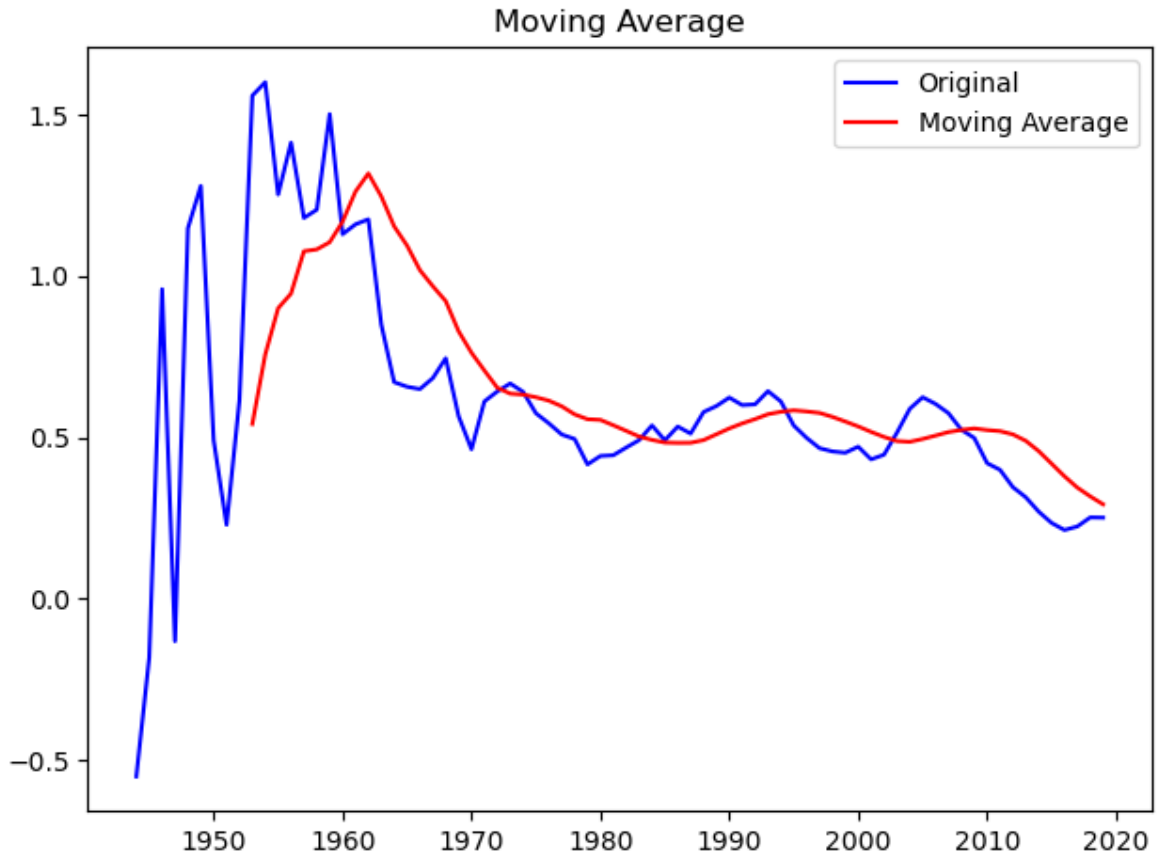

```
Results of Dickey-Fuller Test:
Test Statistic                   2.490244
p-value                          0.999047
#Lags Used                      11.000000
Number of Observations Used     73.000000
Critical Value (1%)             -3.523284
Critical Value (5%)             -2.902031
Critical Value (10%)            -2.588371
dtype: float64
```

Βλέπουμε οτι απο το ADF τεστ η χρονοσειρά δεν είναι σταθερή γιατί το δεδομένο test statistic είναι πολυ μεγαλύτερο απο τις κρίσιμες τιμές.

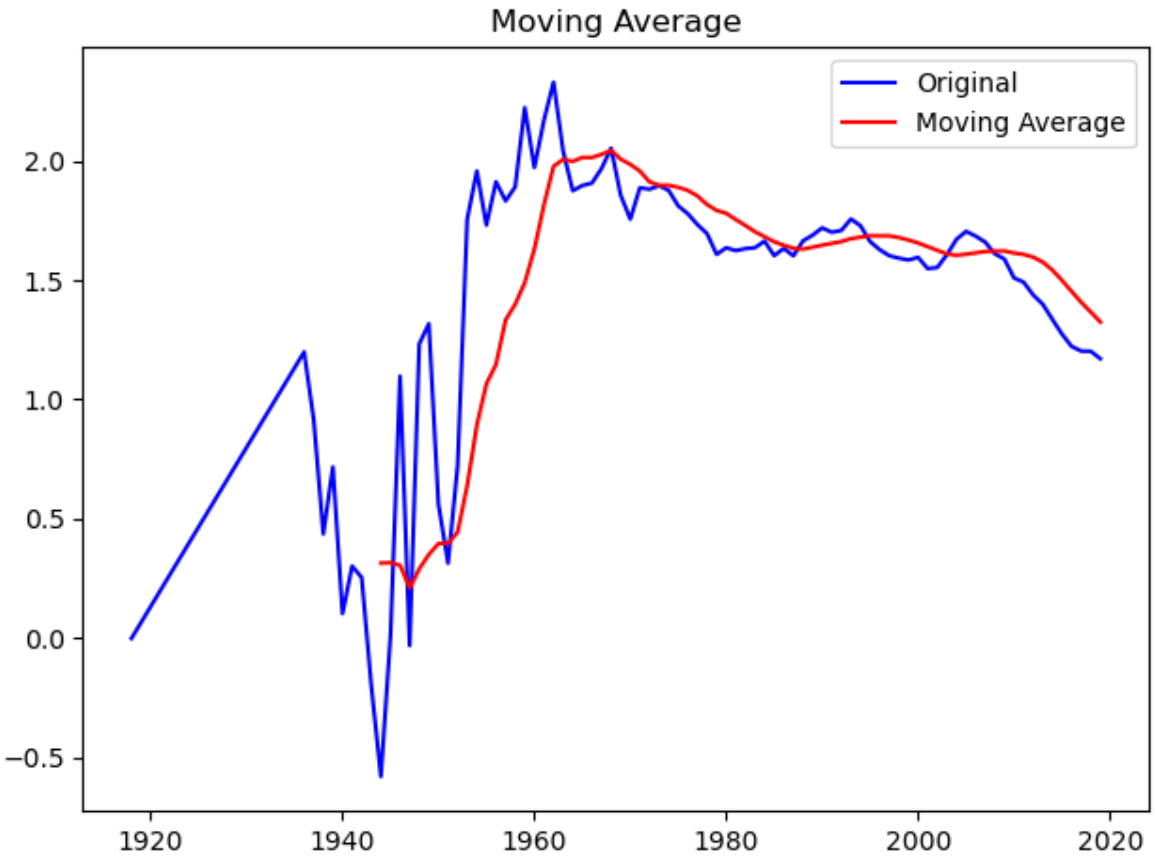
Για να γίνει σταθερή η χρονοσειρά πρέπει να αφαιρεθεί η τάση και/ή εποχικότητα.

Ένας τρόπος για να γίνει αυτό είναι με λογαριθμικό μετασχηματισμό.

Μπορούμε να αφαιρέσουμε το νέο moving average του μετασχηματισμού απο αυτόν και να πάρουμε το εξης διάγραμμα.



Στη συνέχεια παίρνουμε το exponentially weighted moving average και το αφαιρούμε απο την χρονοσειρά για να παρουμε το εξης διάγραμμα.



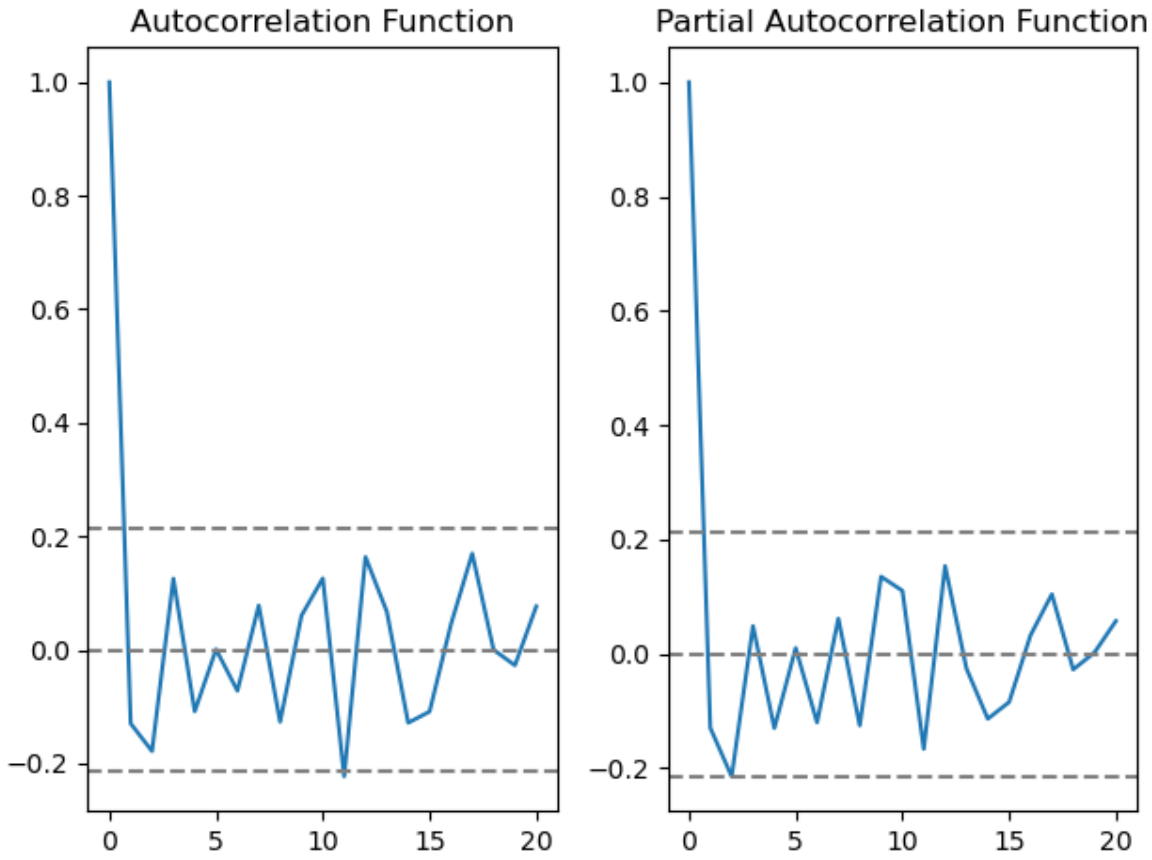
Και απο το ADF Test έχουμε το εξής:

```
Results of Dickey-Fuller Test:
Test Statistic                  -5.449295
p-value                          0.000003
#Lags Used                       9.000000
Number of Observations Used     75.000000
Critical Value (1%)             -3.520713
Critical Value (5%)             -2.900925
Critical Value (10%)            -2.587781
dtype: float64
```

Autocorrelation Function: Είναι μέτρηση της συνάρτησης συσχέτισης μεταξύ αυτής και με τον εαυτό της σε τιμή υστέρησης.

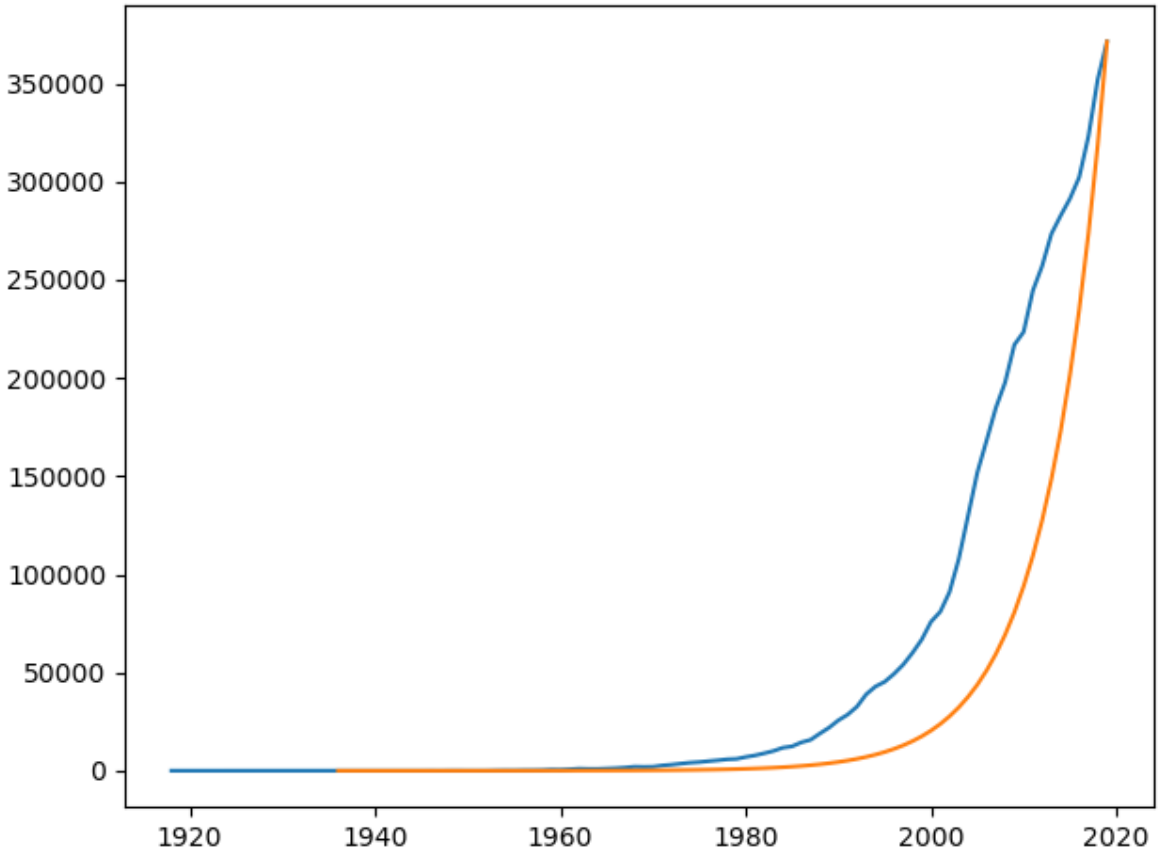
Partial Autocorrelation Function: Κάνει το ίδιο αλλα αφαιρεί τις αλλαγές απο τις προηγούμενες υστερήσεις που έχει δεχτεί.

Βρίσκουμε λοιπόν τα εξής διαγράμματα:



Απο τα δύο αυτά διαγράμματα παίρνουμε τις τιμές που έχουν οταν ακουμπάν πρώτη φορά το πάνω όριο και τις χρησιμοποιούμε για να κάνουμε πρόβλεψη με το ARIMA.

Δεν κατάφερα να βγάλω κάποιο έγκυρο αποτέλεσμα αλλά αυτό είναι το τελικό διάγραμμα.


350000
300000
250000
200000
150000
100000
50000
0
1920
1940
1960
1980
2000
2020