

How to Check Stationary of a Time Series?

A TS is said to be stationary if its statistical properties such as mean, variance remain constant over time. But why is it important? Most of the TS models work on the assumption that the TS is stationary. Intuitively, we can say that if a TS has a particular behaviour over time, there is a very high probability that it will follow the same in the future. Also, the theories related to stationary series are more mature and easier to implement as compared to non-stationary series.

Stationary is defined using very strict criterion. However, for practical purposes we can assume the series to be stationary if it has constant statistical properties over time, ie. The following:

1. constant mean
2. constant variance
3. an auto covariance that does not depend on time

We can check stationary using the following:

- Plotting Rolling Statistics: We can plot the moving average or moving variance and see if it varies with time. By moving average/variance I mean that at any instant 't', we'll take the average/variance of the last year, i.e. last 12 months. But again this is more of a visual technique.
- Dickey-Fuller Test: This is one of the statistical tests for checking stationary. Here the null hypothesis is that the TS is non-stationary. The test results comprise of a Test Statistic and some Critical Values for difference confidence levels. If the 'Test Statistic' is less than the 'Critical Value', we can reject the null hypothesis and say that the series is stationary. Refer this article for details.

These concepts might not sound very intuitive at this point. I recommend going through the prequel article. If you're interested in some theoretical statistics, you can refer Introduction to Time Series and Forecasting by Brockwell and Davis. The book is a bit stats-heavy, but if you have the skill to read-between-lines, you can understand the concepts and tangentially touch the statistics.

Back to checking stationarity, we'll be using the rolling statistics plots along with Dickey-Fuller test results a lot so I have defined a function which takes a TS as input and generated them for us. Please note that I've plotted standard deviation instead of variance to keep the unit similar to mean.

The null hypothesis of the test is that the time series can be represented by a unit root, that it is not stationary (has some time-dependent structure). The alternate hypothesis (rejecting the null hypothesis) is that the time series is stationary.

- Null Hypothesis (H0): If accepted, it suggests the time series has a unit root, meaning it is non-stationary. It has some time dependent structure.
- Alternate Hypothesis (H1): The null hypothesis is rejected; it suggests the time series does not have a unit root, meaning it is stationary. It does not have time-dependent structure.

We interpret this result using the p-value from the test. A p-value below a threshold (such as 5% or 1%) suggests we reject the null hypothesis (stationary), otherwise a p-value above the threshold suggests we accept the null hypothesis (non-stationary).

- p-value > 0.05: Accept the null hypothesis (H0), the data has a unit root and is non-stationary.
- p-value <= 0.05: Reject the null hypothesis (H0), the data does not have a unit root and is stationary

Complementary if ADF Test is higher than Critical Value (5%) it is confirmed and there are some seasonality in the data

	Test Statistic	p-value	#Lags Used	Number of Observations Used	Critical Value (1%)	Critical Value (5%)	Critical Value (10%)	index	ESTATIONARY
0	-0,283389681	0,927824127	20	3504	-3,432217608	-2,862365203	-2,56720927	^GSPC	Non Stationary
1	0,375936496	0,980572404	20	3504	-3,432217608	-2,862365203	-2,56720927	SPY	Non Stationary
2	0,150255696	0,969329451	20	3504	-3,432217608	-2,862365203	-2,56720927	^IXIC	Non Stationary
3	-0,303551213	0,924999866	20	3504	-3,432217608	-2,862365203	-2,56720927	^DJI	Non Stationary
4	-0,780515119	0,824778711	5	3519	-3,432209641	-2,862361684	-2,567207396	^GDAXI	Non Stationary
5	-1,88149035	0,340818103	6	3518	-3,43221017	-2,862361918	-2,567207521	^FTSE	Non Stationary
6	-1,906067833	0,329212017	5	3519	-3,432209641	-2,862361684	-2,567207396	^FCHI	Non Stationary
7	-1,280201056	0,63818098	1	3523	-3,432207528	-2,862360751	-2,5672069	^N225	Non Stationary
8	-2,447901199	0,128670353	19	3505	-3,432217074	-2,862364968	-2,567209144	^HSI	Non Stationary
9	-1,957773436	0,30538918	11	3513	-3,432212819	-2,862363088	-2,567208144	^AXJO	Non Stationary
10	-2,165856047	0,218870229	28	3496	-3,432221884	-2,862367092	-2,567210276	ORB	Non Stationary
11	-1,921089503	0,322205902	23	3501	-3,432219209	-2,86236591	-2,567209646	EUR	Non Stationary
12	-2,715101553	0,071461298	30	3494	-3,432222957	-2,862367566	-2,567210528	AUD	Non Stationary
13	-0,513010473	0,889452486	30	3494	-3,432222957	-2,862367566	-2,567210528	GBP	Non Stationary
14	-1,51324426	0,526949716	1	3523	-3,432207528	-2,862360751	-2,5672069	JPY	Non Stationary
15	-1,689118516	0,436727486	30	3494	-3,432222957	-2,862367566	-2,567210528	SILVER	Non Stationary
16	-1,46335827	0,551553394	29	3495	-3,43222242	-2,862367329	-2,567210402	GOLD	Non Stationary
17	-2,233867048	0,194184028	16	3508	-3,432215476	-2,862364262	-2,567208769	PLAT	Non Stationary
18	-0,480699124	0,895728219	1	3523	-3,432207528	-2,862360751	-2,5672069	WT1010	Non Stationary

Looking at the outcomes of the Dickey-Fuller Test in our 19 Time series, we can conclude that HO is accepted in all of them, meaning that we have seasonality, and in some of them, it is really evident (in red colour).

Autocorrelation PLOTS

Statistical correlation summarizes the strength of the relationship between two variables.

We can assume the distribution of each variable fits a Gaussian (bell curve) distribution. If this is the case, we can use the Pearson's correlation coefficient to summarize the correlation between the variables.

The Pearson's correlation coefficient is a number between -1 and 1 that describes a negative or positive correlation respectively. A value of zero indicates no correlation.

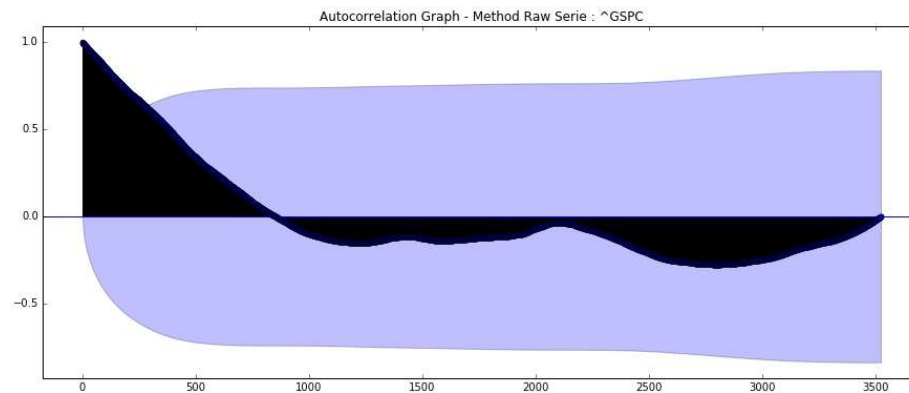
We can calculate the correlation for time series observations with observations with previous time steps, called lags. Because the correlation of the time series observations is calculated with values of the same series at previous times, this is called a serial correlation, or an autocorrelation.

A plot of the autocorrelation of a time series by lag is called the **Auto-Correlation Function**, or the acronym **ACF**. This plot is sometimes called a correlogram or an autocorrelation plot.

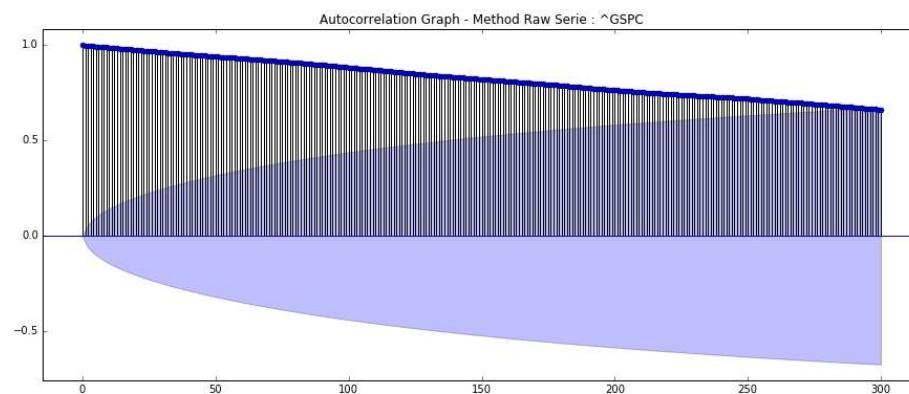
Autocorrelation plots are a commonly-used tool for checking randomness in a data set. This randomness is ascertained by computing autocorrelations for data values at varying time lags. If random, such autocorrelations should be near zero for any and all time-lag separations. If non-random, then one or more of the autocorrelations will be significantly non-zero.

In addition, autocorrelation plots are used in the model identification stage for Box-Jenkins autoregressive, moving average time series models.

Time Series ^GSPC



Confidence intervals are drawn as a cone. By default, this is set to a 95% confidence interval, suggesting that correlation values outside of this cone are very likely a correlation and not a statistical fluke. Looking at the graph we see that approximately lag=300 when the correlation line crosses the confidence interval, which is confirmed plotting PAFC graph with lag =300



Removing seasonality

Though stationary assumption is taken in many TS models, almost none of practical time series are stationary. So statisticians have figured out ways to make series stationary, which we'll discuss now. Actually, it's almost impossible to make a series perfectly stationary, but we try to take it as close as possible.

Let's understand what is making a TS non-stationary. There are 2 major reasons behind non-stationary of a TS:

1. **Trend** – varying mean over time. For example, in this case we saw that on average, the number of passengers was growing over time.
2. **Seasonality** – variations at specific time-frames. Example, people might have a tendency to buy cars in a particular month because of pay increment or festivals.

The underlying principle is to model or estimate the trend and seasonality in the series and remove those from the series to get a stationary series. Then statistical forecasting techniques can be implemented on this series. The final step would be to convert the forecasted values into the original scale by applying trend and seasonality constraints back.

Estimating & Eliminating Trend

One of the first tricks to reduce trend can be transformation. For example, in this case we can clearly see that there is a significant positive trend. So we can apply transformation which penalize higher values more than smaller values. These can be taking a log, square root, cube root, etc.

We can use some techniques to estimate or model the trend and then remove it from the series. There can be many ways of doing it and some of most commonly used are:

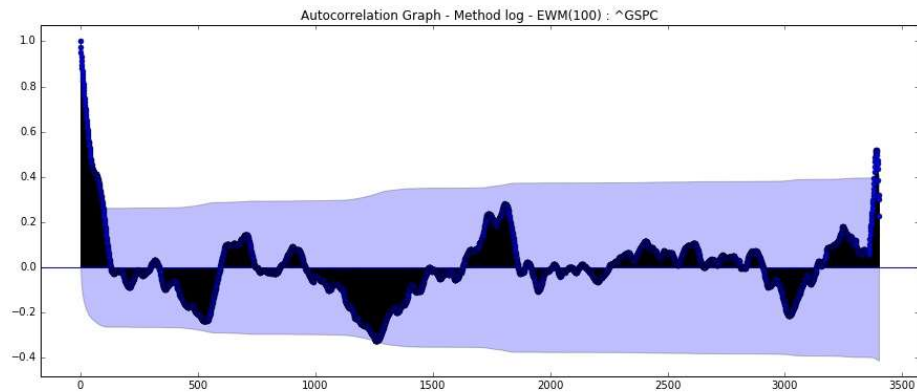
1. **Aggregation** – taking average for a time period like monthly/weekly averages
2. **Smoothing** – taking rolling averages
3. **Polynomial Fitting** – fit a regression model

Moving average

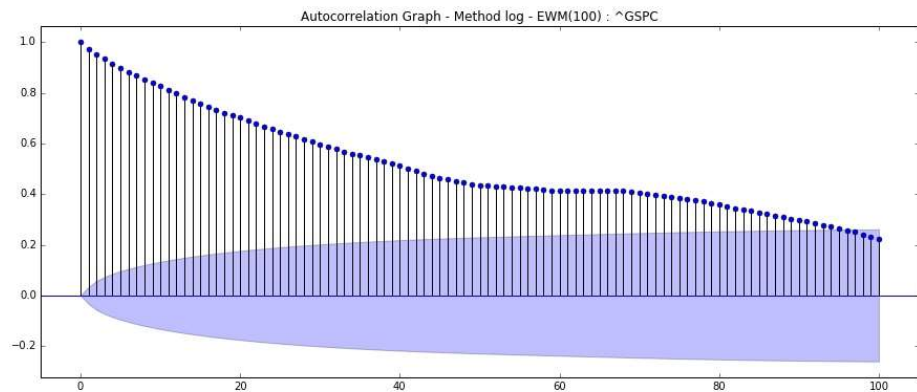
In this approach, we take average of 'k' consecutive values depending on the frequency of time series, then subtract this from the original series. A drawback in this particular approach is that the time-period has to be strictly defined. In this case we can take yearly averages but in complex situations like forecasting a stock price, it's difficult to come up with a number. So we take a 'weighted moving average' where more recent values are given a higher weight. There can be many technique for assigning weights. A popular one is exponentially weighted moving average where weights are assigned to all the previous values with a decay factor

Method used Log (Series) - exponential weighted Series (Span =100)

As we can see on the graph using a window of 100 days the correlation line cross the 95% confidence interval



Here we confirm our hypothesis plotting PAFC with lags=100



Eliminating Trend and Seasonality

The simple trend reduction techniques discussed before don't work in all cases, particularly the ones with high seasonality. Let's discuss two ways of removing trend and seasonality:

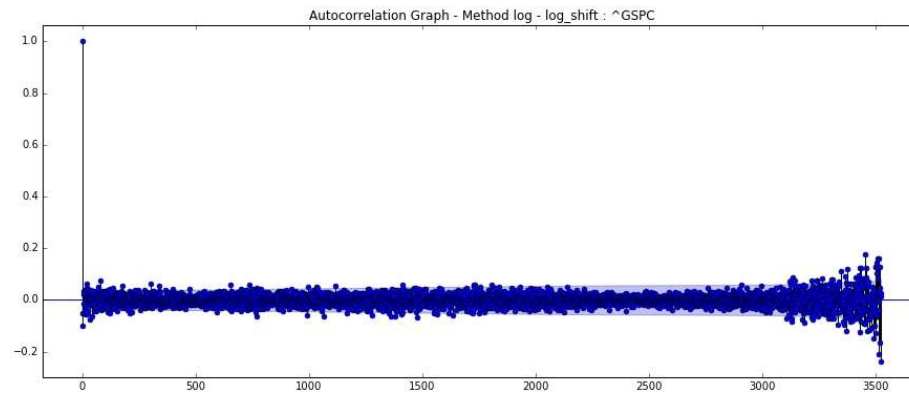
1. **Differencing** – taking the difference with a particular time lag
2. **Decomposition** – modelling both trend and seasonality and removing them from the model.

Differencing

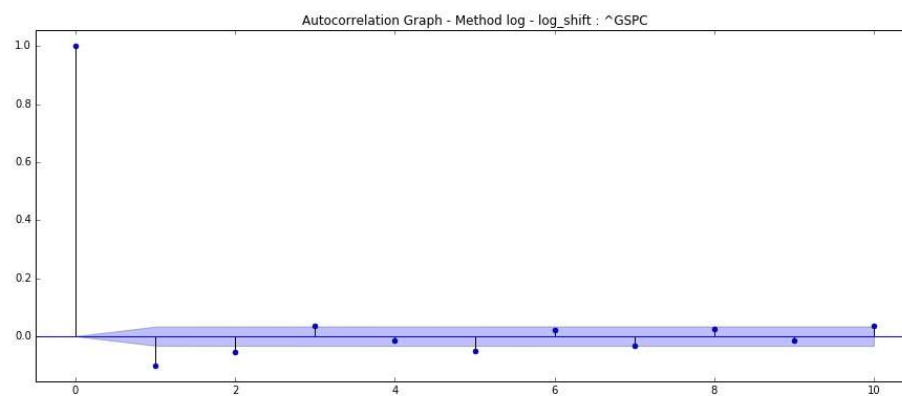
One of the most common methods of dealing with both trend and seasonality is differencing. In this technique, we take the difference of the observation at a particular instant with that at the previous instant

Method Used Log (Series) - Log (Series).shift

This method is much efficient removing seasonality



Here we confirm our hypothesis plotting PAFC with lags=10



Decomposing

In this approach, both trend and seasonality are modelled separately and the remaining part of the series is returned

indice	Test Statistic	p-value	#Lags Used	Critical Value (1%)	Critical Value (10%)	Critical Value (5%)	Number of Observations Used	Test	stationary
^AXJO	-1,957773436	0,30538918	11	-3,432212819	-2,862363088	-2,567208144	3513	Raw Series	No Stationary
^AXJO	-18,83928944	0	10	-2,862362854	-3,432212289	-2,567208019	3514	log - log_shift	Stationary
^AXJO	-4,954768764	2,72525E-05	11	-2,862393217	-3,43228103	-2,567224184	3389	log - EWM(100)	Stationary
^DJI	-0,303551213	0,924999866	20	-3,432217608	-2,862365203	-2,56720927	3504	Raw Series	No Stationary
^DJI	-13,00932507	2,58764E-24	21	-2,862365439	-3,432218141	-2,567209395	3503	log - log_shift	Stationary
^DJI	-4,913748109	3,2788E-05	21	-2,862395743	-3,432286749	-2,567225529	3379	log - EWM(100)	Stationary
^FCHI	-1,906067833	0,329212017	5	-3,432209641	-2,862361684	-2,567207396	3519	Raw Series	No Stationary
^FCHI	-29,44283574	0	4	-2,862361451	-3,432209112	-2,567207272	3520	log - log_shift	Stationary
^FCHI	-5,729138398	6,66298E-07	5	-2,862391709	-3,432277614	-2,567223381	3395	log - EWM(100)	Stationary
^FTSE	-1,88149035	0,340818103	6	-3,43221017	-2,862361918	-2,567207521	3518	Raw Series	No Stationary
^FTSE	-26,67831436	0	5	-2,862361684	-3,432209641	-2,567207396	3519	log - log_shift	Stationary
^FTSE	-6,084790173	1,07083E-07	6	-2,86239196	-3,432278183	-2,567223514	3394	log - EWM(100)	Stationary
^GDAXI	-0,780515119	0,824778711	5	-3,432209641	-2,862361684	-2,567207396	3519	Raw Series	No Stationary
^GDAXI	-28,53113581	0	4	-2,862361451	-3,432209112	-2,567207272	3520	log - log_shift	Stationary
^GDAXI	-5,825648375	4,08537E-07	5	-2,862391709	-3,432277614	-2,567223381	3395	log - EWM(100)	Stationary
^GSPC	-0,283389681	0,927824127	20	-3,432217608	-2,862365203	-2,56720927	3504	Raw Series	No Stationary
^GSPC	-12,58396353	1,88366E-23	21	-2,862365439	-3,432218141	-2,567209395	3503	log - log_shift	Stationary
^GSPC	-4,780865157	5,91601E-05	21	-2,862395743	-3,432286749	-2,567225529	3379	log - EWM(100)	Stationary
^HSI	-2,447901199	0,128670353	19	-3,432217074	-2,862364968	-2,567209144	3505	Raw Series	No Stationary
^HSI	-13,09924802	1,72352E-24	18	-2,862364732	-3,432216541	-2,567209019	3506	log - log_shift	Stationary
^HSI	-4,991433952	2,3076E-05	19	-2,862395237	-3,432285602	-2,567225259	3381	log - EWM(100)	Stationary
^IXIC	0,150255696	0,969329451	20	-3,432217608	-2,862365203	-2,56720927	3504	Raw Series	No Stationary
^IXIC	-13,18804831	1,15925E-24	19	-2,862364968	-3,432217074	-2,567209144	3505	log - log_shift	Stationary
^IXIC	-5,304916348	5,3487E-06	20	-2,86239549	-3,432286176	-2,567225394	3380	log - EWM(100)	Stationary
^N225	-1,280201056	0,63818098	1	-3,432207528	-2,862360751	-2,5672069	3523	Raw Series	No Stationary
^N225	-36,26219537	0	2	-2,862360984	-3,432208056	-2,567207024	3522	log - log_shift	Stationary
^N225	-5,640588281	1,03873E-06	3	-2,862391207	-3,432276479	-2,567223114	3397	log - EWM(100)	Stationary
AUD	-2,715101553	0,071461298	30	-3,432222957	-2,862367566	-2,567210528	3494	Raw Series	Stationary
AUD	-13,31397333	6,65998E-25	18	-2,862364732	-3,432216541	-2,567209019	3506	log - log_shift	Stationary
AUD	-4,744081127	6,94902E-05	19	-2,862395237	-3,432285602	-2,567225259	3381	log - EWM(100)	Stationary
EUR	-1,921089503	0,322205902	23	-3,432219209	-2,86236591	-2,567209646	3501	Raw Series	No Stationary
EUR	-12,91040307	4,06839E-24	20	-2,862365203	-3,432217608	-2,56720927	3504	log - log_shift	Stationary
EUR	-5,158927368	1,06583E-05	21	-2,862395743	-3,432286749	-2,567225529	3379	log - EWM(100)	Stationary
GBP	-0,513010473	0,889452486	30	-3,432222957	-2,862367566	-2,567210528	3494	Raw Series	No Stationary
GBP	-12,96465647	3,17204E-24	22	-2,862365675	-3,432218675	-2,567209521	3502	log - log_shift	Stationary
GBP	-4,389696097	0,000309547	23	-2,86239625	-3,432287897	-2,567225799	3377	log - EWM(100)	Stationary
GOLD	-1,46335827	0,551553394	29	-3,43222242	-2,862367329	-2,567210402	3495	Raw Series	No Stationary

GOLD	-13,52644612	2,6728E-25	21	-2,862365439	-3,432218141	-2,567209395	3503	log - log_shift	Stationary
GOLD	-5,494319798	2,14035E-06	22	-2,862395997	-3,432287323	-2,567225664	3378	log - EWM(100)	Stationary
JPY	-1,51324426	0,526949716	1	-3,432207528	-2,862360751	-2,5672069	3523	Raw Series	No Stationary
JPY	-52,1344132	0	0	-2,862360518	-3,432207	-2,567206776	3524	log - log_shift	Stationary
JPY	-5,423681445	3,02011E-06	1	-2,862390706	-3,432275344	-2,567222847	3399	log - EWM(100)	Stationary
ORB	-2,165856047	0,218870229	28	-3,432221884	-2,862367092	-2,567210276	3496	Raw Series	No Stationary
ORB	-9,719225577	9,64378E-17	25	-2,862366383	-3,432220278	-2,567209898	3499	log - log_shift	Stationary
ORB	-4,760751617	6,46107E-05	24	-2,862396504	-3,432288471	-2,567225934	3376	log - EWM(100)	Stationary
PLAT	-2,233867048	0,194184028	16	-3,432215476	-2,862364262	-2,567208769	3508	Raw Series	No Stationary
PLAT	-9,535261453	2,81991E-16	30	-2,862367566	-3,432222957	-2,567210528	3494	log - log_shift	Stationary
PLAT	-4,182252056	0,000705993	24	-2,862396504	-3,432288471	-2,567225934	3376	log - EWM(100)	Stationary
SILVER	-1,689118516	0,436727486	30	-3,432222957	-2,862367566	-2,567210528	3494	Raw Series	No Stationary
SILVER	-64,12735734	0	0	-2,862360518	-3,432207	-2,567206776	3524	log - log_shift	Stationary
SILVER	-6,087909533	1,05348E-07	1	-2,862390706	-3,432275344	-2,567222847	3399	log - EWM(100)	Stationary
SPY	0,375936496	0,980572404	20	-3,432217608	-2,862365203	-2,56720927	3504	Raw Series	No Stationary
SPY	-12,62526031	1,54664E-23	21	-2,862365439	-3,432218141	-2,567209395	3503	log - log_shift	Stationary
SPY	-4,797423876	5,50064E-05	21	-2,862395743	-3,432286749	-2,567225529	3379	log - EWM(100)	Stationary

Looking to the outcome on the table after running the Dickey-Fuller Test with two differentiating methods, 1) Log (Series) - exponential weighted Series (Span =100)
And 2) Log (Series) - Log (Series).shift

We interpret this result using the p-value from the test. A p-value below a threshold (such as 5% or 1%) suggests we reject the null hypothesis (stationary), otherwise a p-value above the threshold suggests we accept the null hypothesis (non-stationary).

- P-value > 0.05: Accept the null hypothesis (H0), the data has a unit root and is non-stationary.
- p-value <= 0.05: Reject the null hypothesis (H0), the data does not have a unit root and is stationary

Complementary if ADF Test is higher than Critical Value (5%) it is confirmed and there are some seasonality in the data

While in both cases we get stationary Series according to the DF test, I prefer to use the method 2 as we get with all series a confidence higher of 99% that the resultant Series don't contains Seasonality

Forecasting a Time Series (ARIMA Forecasting)

We saw different techniques and all of them worked reasonably well for making the TS stationary. Let's make model on the TS after differencing as it is a very popular technique. Also, it's relatively easier to add noise and seasonality back into predicted residuals in this case. Having performed the trend and seasonality estimation techniques, there can be two situations:

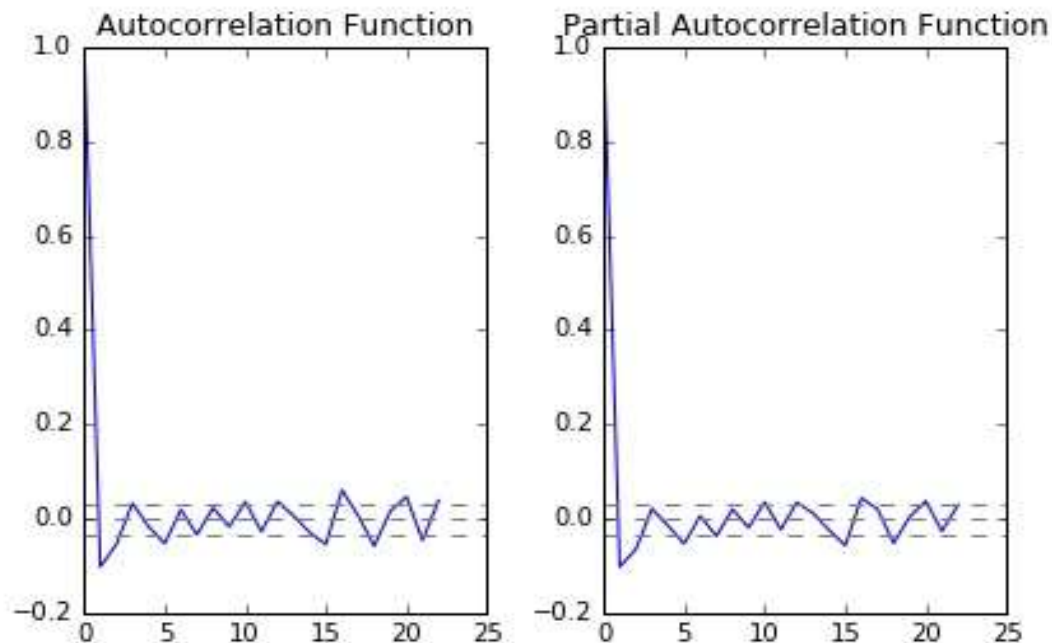
1. A **strictly stationary series** with no dependence among the values. This is the easy case wherein we can model the residuals as white noise. But this is very rare.
2. A series with significant **dependence among values**. In this case we need to use some statistical models like ARIMA to forecast the data.

Let me give you a brief introduction to ARIMA. I won't go into the technical details but you should understand these concepts in detail if you wish to apply them more effectively. ARIMA stands for Auto-Regressive Integrated Moving Averages. The ARIMA forecasting for a stationary time series is nothing but a linear (like a linear regression) equation. The predictors depend on the parameters (p,d,q) of the ARIMA model:

1. **Number of AR (Auto-Regressive) terms (p):** AR terms are just lags of dependent variable. For instance if p is 5, the predictors for $x(t)$ will be $x(t-1) \dots x(t-5)$.
2. **Number of MA (Moving Average) terms (q):** MA terms are lagged forecast errors in prediction equation. For instance if q is 5, the predictors for $x(t)$ will be $e(t-1) \dots e(t-5)$ where $e(i)$ is the difference between the moving average at i^{th} instant and actual value.
3. **Number of Differences (d):** These are the number of no seasonal differences, i.e. in this case we took the first order difference. So either we can pass that variable and put $d=0$ or pass the original variable and put $d=1$. Both will generate same results.

An importance concern here is how to determine the value of 'p' and 'q'. We use two plots to determine these numbers. Let's discuss them first.

1. **Autocorrelation Function (ACF):** It is a measure of the correlation between the TS with a lagged version of itself. For instance at lag 5, ACF would compare series at time instant 't1'...'t2' with series at instant 't1-5'...'t2-5' (t1-5 and t2-5 being end points).
2. **Partial Autocorrelation Function (PACF):** This measures the correlation between the TS with a lagged version of itself but after eliminating the variations already explained by the intervening comparisons. Eg at lag 5, it will check the correlation but remove the effects already explained by lags 1 to 4



In this plot, the two dotted lines on either sides of 0 are the confidence intervals. These can be used to determine the 'p' and 'q' values as:

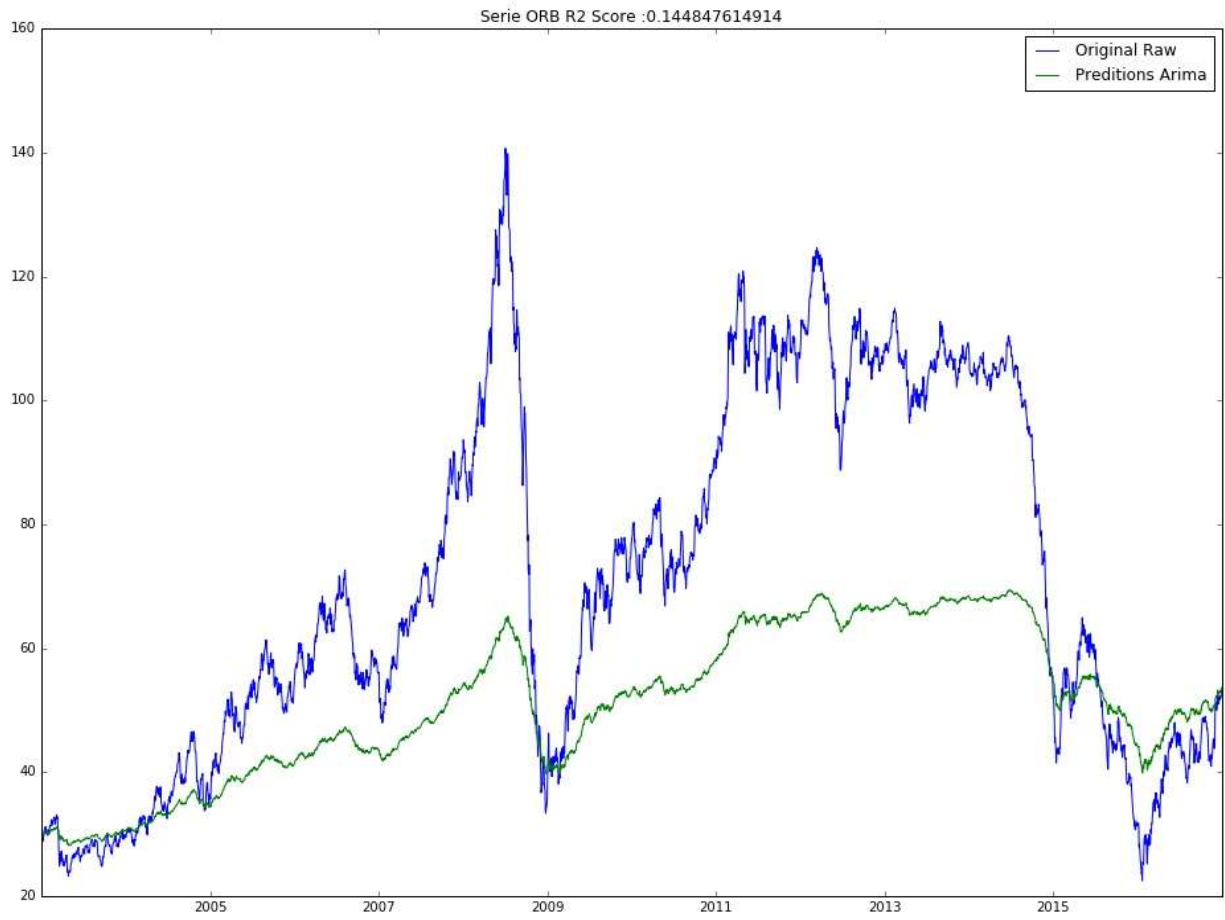
1. **p** – The lag value where the **PACF** chart crosses the upper confidence interval for the first time. If you notice closely, in this case $p=1$.
2. **q** – The lag value where the **ACF** chart crosses the upper confidence interval for the first time. If you notice closely, in this case $q=1$.

ARIMA Prediction

R^2 (coefficient of determination) regression score function.

Best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse). A constant model that always predicts the expected value of y , disregarding the input features, would get a R^2 score of 0.0.





ACF- PACF Rest of Index

