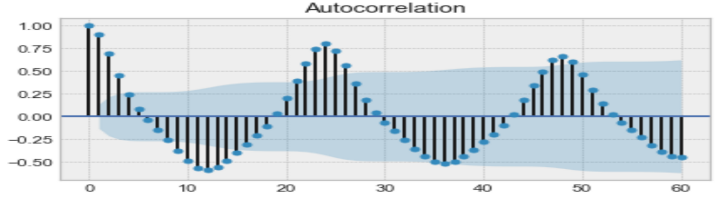
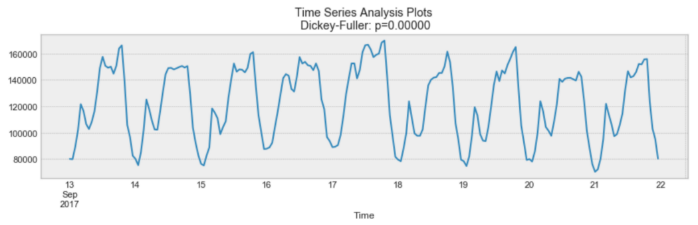
1. **Check the stationarity, trend and seasonality of the given data.**

A time series is simply a series of data points ordered in time. In a time series, time is often the independent variable and the goal is usually to make a forecast for the future. However, there are other aspects that come into play when dealing with time series.

* Is it stationary?
* Is there a seasonality?
* Is the target variable autocorrelated?

**Aim:** To introduce different characteristics of time series and how we can model them to obtain accurate (as much as possible) forecasts.

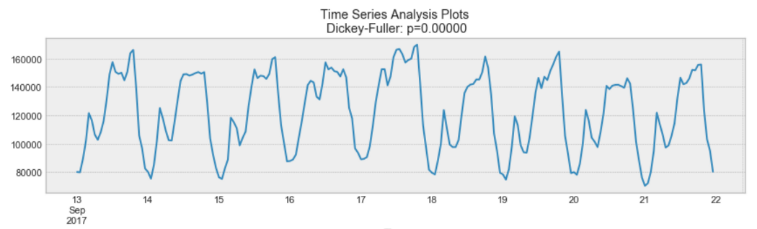
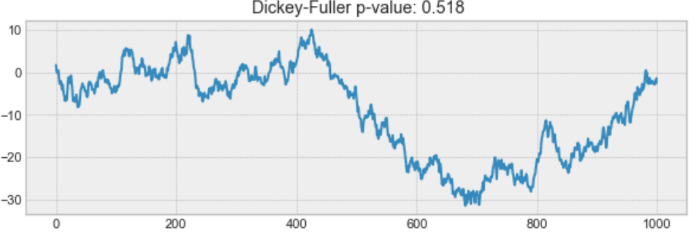
**Autocorrelation** Informally, is the similarity between observations as a function of the time lag between them.

Autocorrelation Seasonality

**Seasonality** - Seasonality refers to periodic fluctuations. Seasonality can also be derived from an autocorrelation plot if it has a sinusoidal shape.

**Stationarity** - Stationarity is an important characteristic of time series. A time series is said to be stationary if its statistical properties do not change over time. In other words, it has constant mean and variance, and covariance is independent of time. Want to have a stationary time series for modelling. Of course, not all of them are stationary, but we can make different transformations to make them stationary.

Stationary Non Stationary

***How to test if a process is stationary?***

Dickey-Fuller - is the statistical test that is run to determine if a time series is stationary or not. It tests the null hypothesis that a unit root is present. If it is, then p > 0, and the process is not stationary. Otherwise, p = 0, the null hypothesis is rejected, and the process is considered to be stationary.

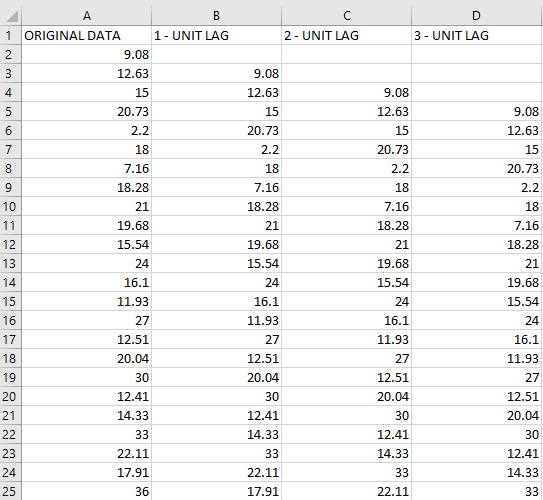
Below given properties influence the modeling process significantly.

* Autocorrelation
* Seasonality
* Stationarity

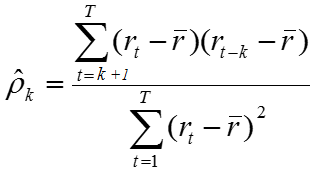
## Autocorrelation: -

***Autocorrelation*** is a mathematical representation of the degree of similarity between a given time series and the lagged version of itself over successive time intervals. In other words, instead of calculating the correlation between two different series, we calculate the correlation of the series with an “x” unit lagged version (x**∈N**) of itself.

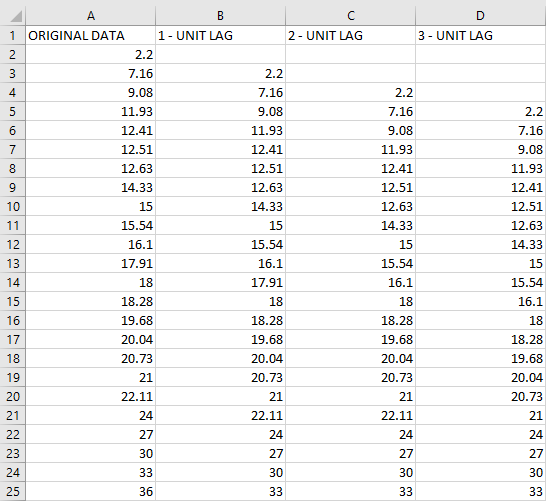
Autocorrelation is also known as **lagged correlation** or **serial correlation**. The value of autocorrelation varies between +1 & -1. If the autocorrelation of series is a very small value that does not mean, there is no correlation. The correlation could be non-linear. Let us understand by a hand calculated example.



**Eg**.: Consider the above dataset. The data represents the monthly sales for 2 years and the lagged versions of itself. For the sake of this example, we have a 3 unit lagged series. We are going to use the below given formula to calculate the autocorrelation for the time series.



**Step 1: Sort the data into ascending order.**



Step 2: Calculate the average of the original data (first column) which is r(t) = 18.1933.

Step 3: Now calculate the denominator of the formula which is the ***sum of the squared deviations*** of the original data = 1455.431.

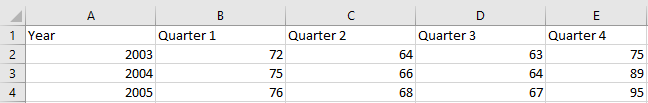
* Now have the values of the denominator and the average of the original dataset (r(t)).

Step 4: Let’s calculate the numerator:

* Subtract the average of the original data from the first column starting from the fourth row and store it in a separate column.
* Do the same with the fourth column (3 – unit lag column) and store the output in a separate column.
* Once you have both the columns in place***, multiply them, find the total of it and divide it by the sum of squared deviations of the original dataset*** which will be 647.8286/1455.431 = 0.46 (approx.).
* This is the value of your autocorrelation which could indicate ***seasonality*** for a time interval of 3 months.

**Seasonality**:

Seasonality in time series data means ***periodic fluctuations.*** It is often considered when the graph of the time series resembles a sinusoidal shape, which means that the graph looks like a sine function or shows repetitions after every fixed interval of time. This ***repetition*** interval is known as your ***period***. Seasonality means the data shows a repetitive structure every one-year. There is a difference between ***cyclic*** and ***seasonal*** data. When talking about cyclic data, the ***period can be of variable lengths*** like 2 days, 2 months, 2 years etc. but if you are meant to find seasonality then it has to be ***over a period of*** one calendar year or financial year. It is important for companies that have been in the market for a long time. Let’s understand seasonality by a small handwritten example.



Consider the above example which shows three-year sales of a company per quarter. Let’s calculate seasonality index for this data.

Step 1: Calculate the average sales for 2003, 2004 and 2005.

Average sales for year 2003 = 68.5

Average sales for year 2004 = 73.5

Average sales for year 2005 = 76.5

Step 2: Now divide the values of quarters by their respective year averages.

|  |  |  |  |
| --- | --- | --- | --- |
| 1.051 | 0.934 | 0.920 | 1.095 |
| 1.020 | 0.898 | 0.871 | 1.211 |
| 0.993 | 0.888 | 0.876 | 1.242 |

If we had data only for the year 2003 then the values in the first row would be our seasonal indexes, but we have data for 3 consecutive years so we’ll have to take the average of all of them which will result in the ***seasonal indexes***.

Step 3: Calculate the seasonal index by taking averages of the columns in the above table.

|  |  |  |  |
| --- | --- | --- | --- |
| 1.021 | 0.97 | 0.889 | 1.183 |

These are our seasonal indexes. To check whether or not we have arrived to the correct values, add all of them and the sum should be equal to 4. In our example the value should be equal to 4 but if the data had sales for 12 months then the value of the indexes should sum up to 12.

* Plot the data

Now let’s deseasonalize (reducing noise) our data.

Step 4: Take the original data and divide the data point in each column by their respective seasonal indexes.

|  |  |  |  |
| --- | --- | --- | --- |
| 71 | 71 | 71 | 63 |
| 73 | 73 | 72 | 75 |
| 74 | 75 | 75 | 80 |

This is our deseasonalized data.

Step 5 : Compares to the original data by plotting them on the same graph.

We can see that the data looks much smoother and this transformation makes it easy to identify trends in the data.

* If you wish you to go back to the original data then, original\_data = deseasonalized\_data \* index

There are 3 ways to spot seasonality:

* Lookup operation
* 3 median method
* Least squares method
  + If we were to predict the deseasonalized sales for the first quarter of 2006 using least squares method, we’ll take all the values, put them in a calculator and find the regression equation.

y = 0.78x + 67.68

x takes quarters as input and while doing regression the quarters are represented by 1,2,3,…… so the first quarter of 2006 means x=13. This will give us 77.82.

The entire process can be replicated in python by following the steps given in implementation.

Importantly, a time series where the seasonal component has been removed is called seasonal stationary. A time series with a clear seasonal component is referred to as non-stationary.

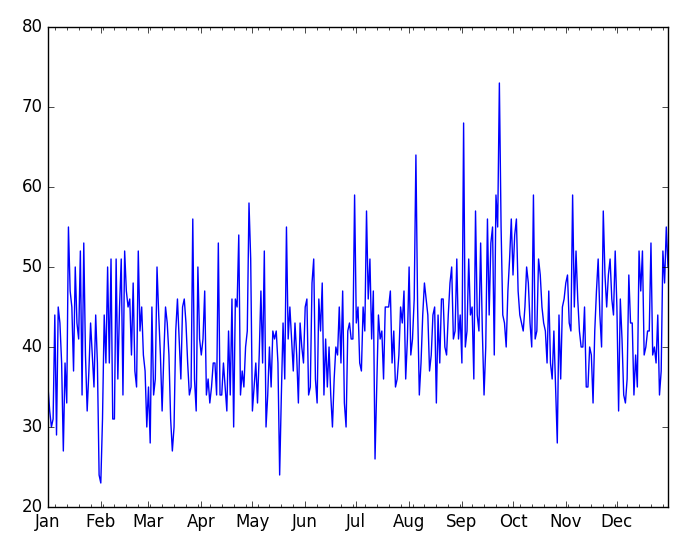
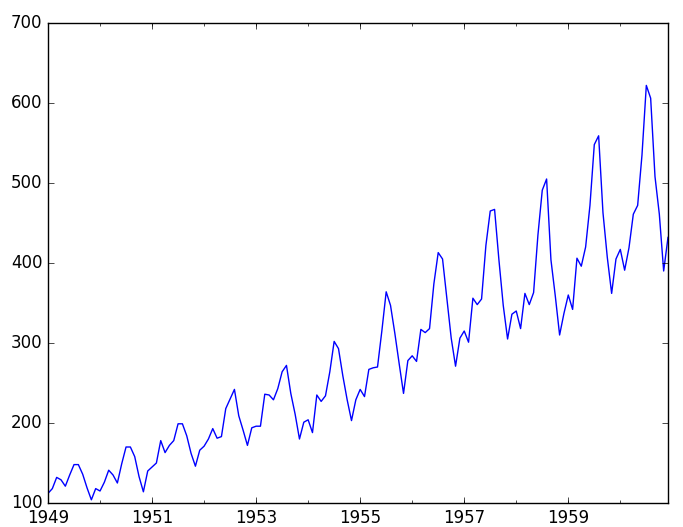
**Stationarity**:

This is one of the most important characteristics of time series data. A time series is said to be ***stationary if it has constant mean, variance and the covariance is independent of time.*** In ideal situations we would prefer a stationary series, but in real world, that’s not the case. There are different types of stationary time series as follows:

* Stationary process: A process that generates a stationary series of observations
* Stationary model: A model that describes a stationary series of observations
* Trend stationary: A time series does not show a trend
* Seasonal stationery: A time series does not depict seasonality
* Strictly stationary: A mathematical definition of a stationary process, specifically that the joint distribution of observations is invariant to time shift.

Identifying stationarity in the time series can be tricky at times. There are multiple ways to deal with it.

* Looking at the plots: By far the easiest and most straightforward method to decide whether the series is stationary or non-stationary.
* Summary statistics: This could be a cheat way to identify stationarity in the series.
  + To put it into simple words, break down the series into two or more than two parts and compare the mean and variance among all the parts.
  + If they happen to be similar, then the series is stationary.

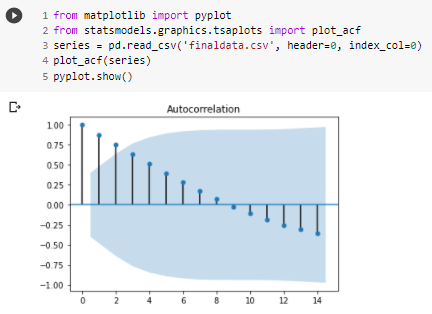
 

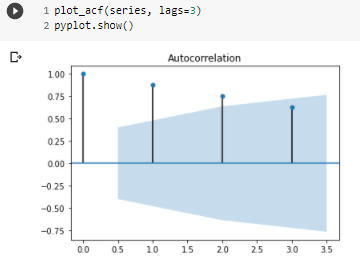
Stationary Series             Non-stationary series

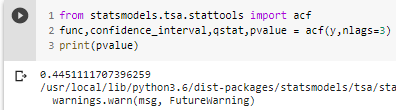
**Autocorrelation**

**Implementation**

Having done that, let’s see the easier way to do all the calculation in python. First plot the data and see how our data looks. Now let’s see how to visualize autocorrelation with unit lag = 3 in python.

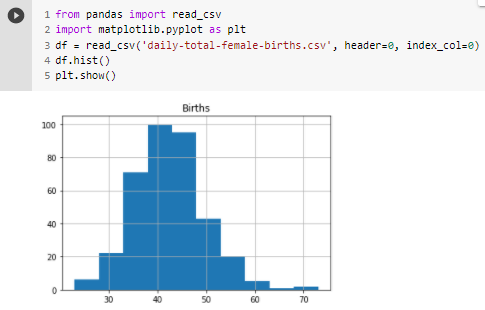




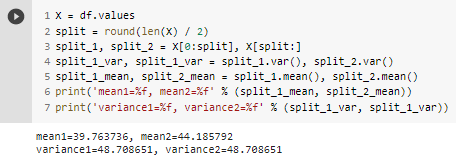


**Stationarity**:

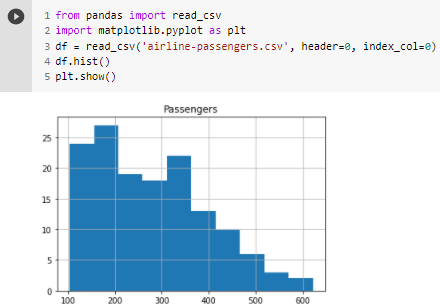
Let’s take the classic example of the birth and airline passenger csv dataset.



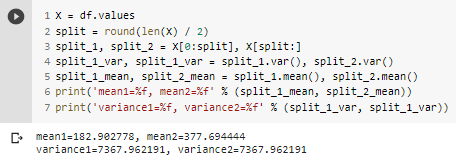
Looking at the histogram we can observe that the data forms a bell curve resulting in a normal distribution. Now following our summary statistics approach, let’s divide the dataset into two parts. Since the length of the dataset is an odd number, we’ll round the result of division by 2, to make two parts of the series.



Looking at the values we could say that the series is stationary as the values are quite close to each other, but if we take the example of the classic airline data then you’ll see that the values are quite far from each other.

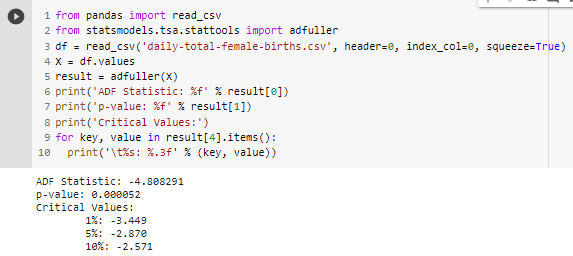


Here one thing is quite evident, that the graphs don’t show stationarity, but if we visit the time series plot for airline passengers again, we can see exponential seasonal growth. To flatten out this growth we could do a log transform on all the values, plot the transformed values and find mean and variance. Now the values will be pretty close to each other.



* Statistical tests:

The most famous one is the ***Augmented Dickey-Fuller test*** (ADF). It is also called a ***unit root test***. ADF uses an ***autoregressive*** model and optimizes an information criterion across multiple different lag values. The simple idea behind this is looking at p-value. If the p-value is <=0.05 then we reject the null hypothesis as the data does not have a unit root and is stationary.



* Higher negative value of ADF Statistic means we can reject the null hypothesis.
* The p-value is also less than 0.05 which means the series is stationary and we can reject the null hypothesis.
* ADF statistic value is less than value of -3.499 at 1%. This means we can reject the null hypothesis with a significance level of less than 1%.