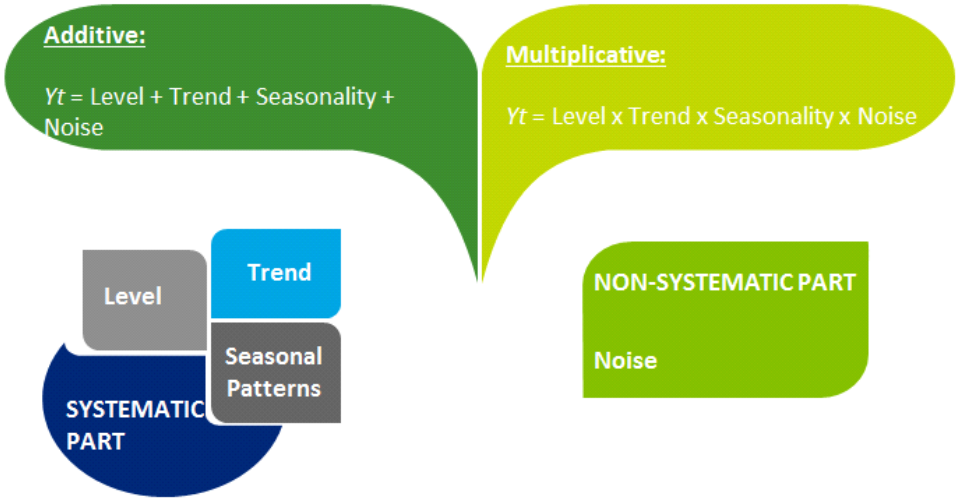
Excel R

10 April 2022 10:02

Generally, Two types of data we have

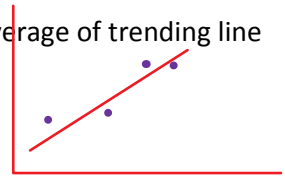
1. Cross sectional data: our Regular data that we used for our machine learning models.
2. Time series data: A series of data which is following with Time as an index.

Components of TSA



What is Level:

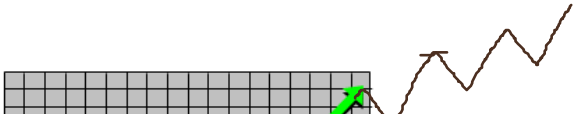
Level is nothing but Average of trending line

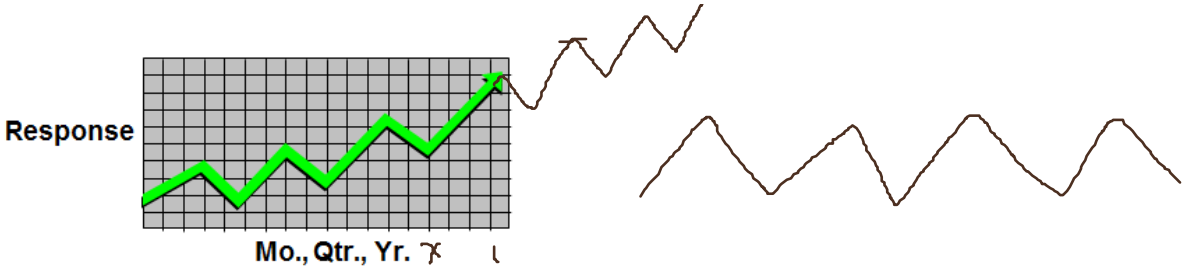


d1	24
d2	25
d3	26
d4	25
avg	25

What is Trend?

The trend is **the component of a time series that represents variations of low frequency in a time series, the high and medium frequency fluctuations having been filtered out**



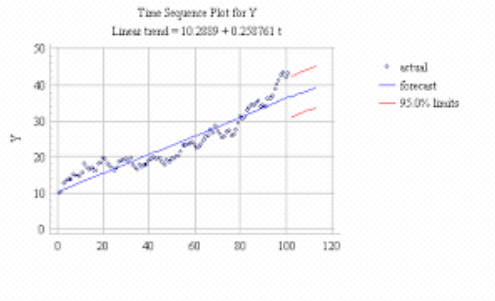


Trend can be divided in two again.

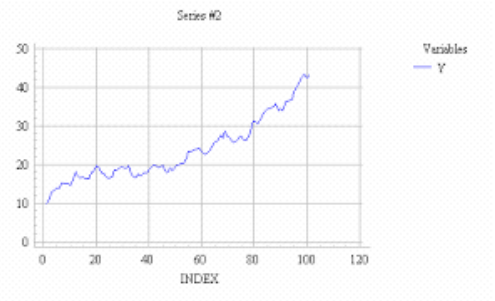
1. Linear Trend
2. Non-Linear Trend

What is a Linear Trend?

The linear trend is **the steady increase or decrease of the variables over the period of time**. The model observes the previous data and predicts the future growth or pattern. On the graph, the model is shown as a straight line towards upwards or downwards direction.



Linear trend model
faculty.fuqua.duke.edu

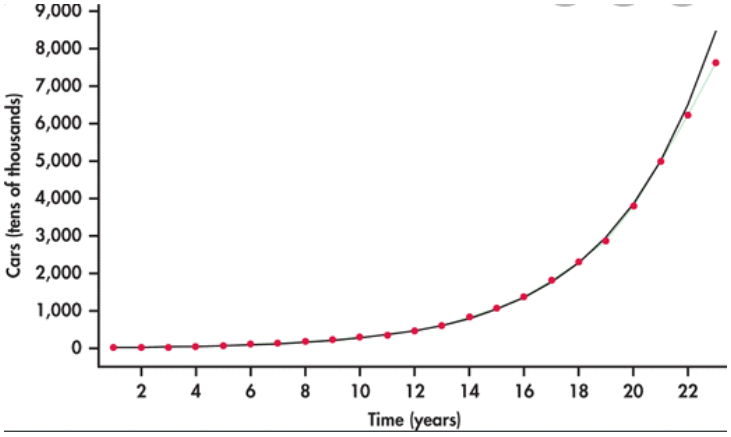


Linear trend model
faculty.fuqua.duke.edu

What is a Non - Linear Trend?

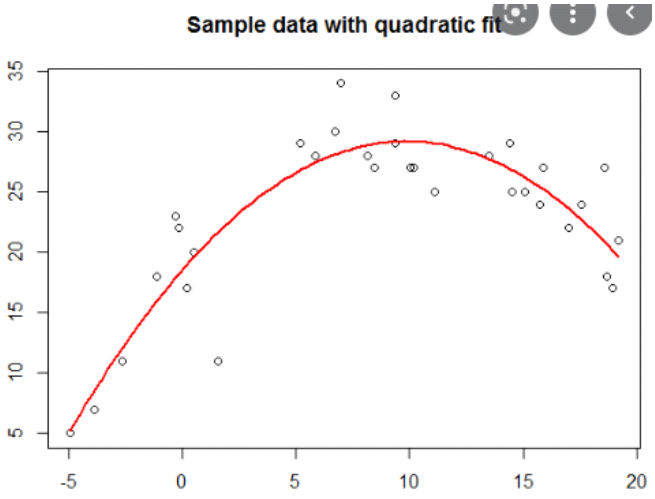
Non-Linear Trend can be divided in to again

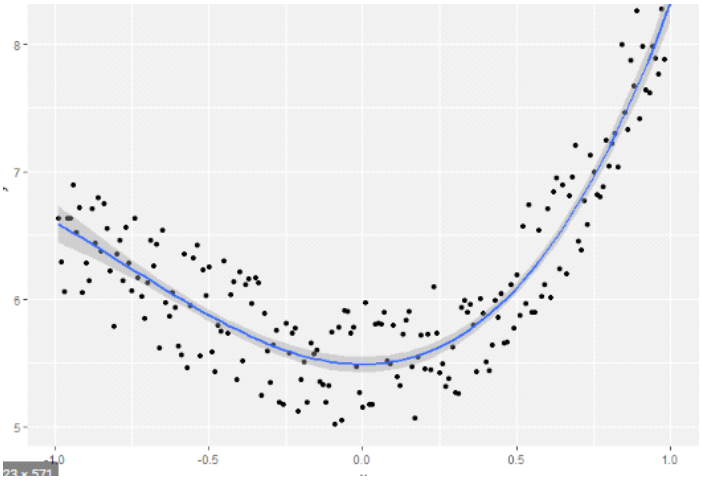
a. Exponential



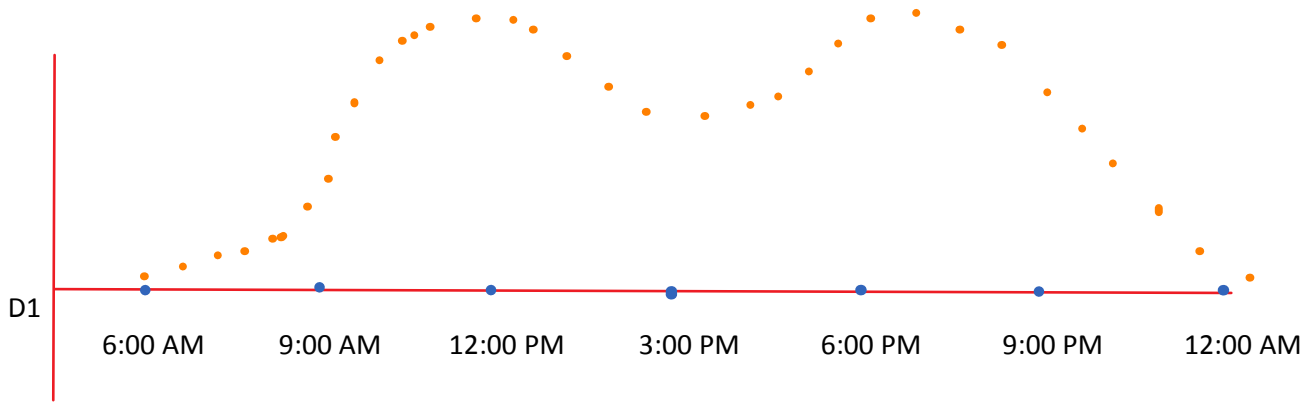
b. Polynomial

It will be again 2nd , 3rd order, 4th order….



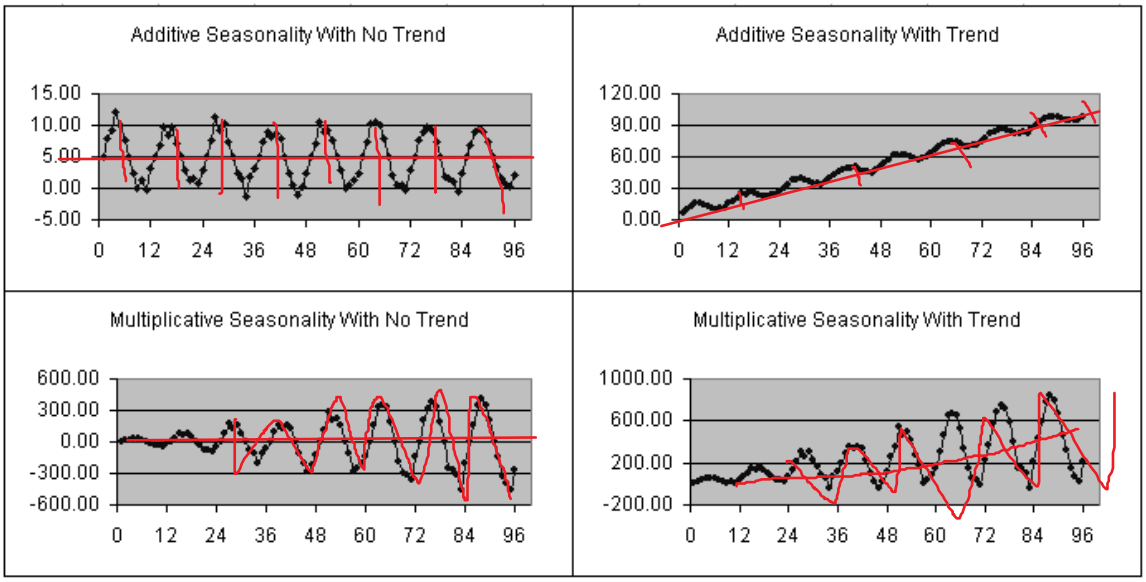


What is Seasonal component

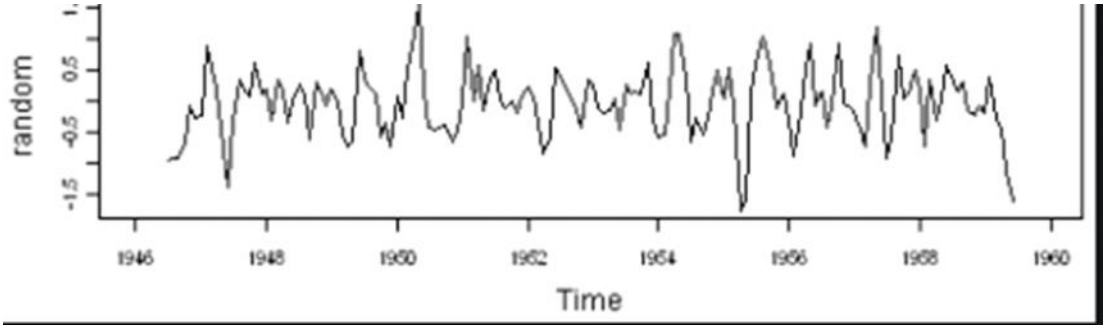


Seasonal components two things again.

i. Additive Seasonality
ii. Multiplicative Seasonality



Irregular/Random/Noise Component



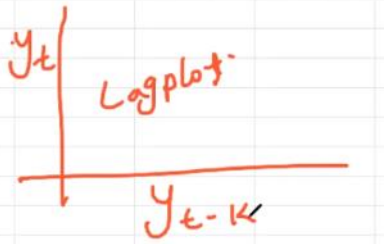
Time series Data Visualization

Time series lends itself naturally to visualization. Line plots of observations over time are popular, but there is a suite of other plots that you can use to learn more about your problem. The more you learn about your data, the more likely you are to develop a better forecasting model.

1. Line Plots.
2. Histograms and Density Plots.
3. Box and Whisker Plots.
4. Heat Maps.

What is meant by Lag plot?

Month	Sales (Yt)	Yt-1	Yt-2
Jan	272		
Feb	522	272	
Mar	114	522	272
Apr	424	114	522
May	486	424	114
Jun	667	486	424
Jul	782	667	486
Aug	906	782	667
Sep	137	906	782
Oct	290	137	906
Nov	408	290	137
Dec	212	408	290



If you see a good relationship between Y_t and Y_{t-1} that means , we can use these two variables for model forecasting.

We can confirm that **Auto correlation** existence in the data

So we will try with

$'r_0$ ---------> Y_t and Y_t ---------> 1

$'r_1$ ---------> Y_t and Y_{t-1} ---------> 0.2

$'r_2$ ---------> Y_t and Y_{t-2} ---------> 0.4

$'r_3$ ---------> Y_t and Y_{t-3} ---------> 0.9

$'r_4$ ---------> Y_t and Y_{t-4} ---------> 0.6

$'r_5$ ---------> Y_t and Y_{t-5} ---------> 0.75

$'r_6$ ---------> Y_t and Y_{t-6} ---------> 0.3

Wherever we could see high autocorrelation we will take as best and also we need to look how much statistical significance it is near to the actual population data.

For that purpose we have to look out for Standard Error for each $'r_k$

- The standard error is

$$SE(r_k) = \sqrt{\frac{1 + 2\sum_{i=1}^{k=1} r_i^2}{n}}$$

- Increases progressively with k, but eventually reaches a maximum value
- If the 'true' autocorrelation is 0, then the estimate r_k should be in the interval $(-2SE(r_k), 2SE(r_k))$ 95% of the time.
- Sometimes $SE(r_k)$ is approximated by $\sqrt{1/n}$

+- 2 Standard Error Range

'r_0 ---------> Y_t and Y_t ---------> 1

'r_1 ---------> Y_t and Y_{t-1} ---------> 0.2 ----------> [-0.2 , 0.7]

'r_2 ---------> Y_t and Y_{t-2} ---------> 0.4 ----------> [-0.1 , 0.6]

'r_3 ---------> Y_t and Y_{t-3} ---------> 0.9 ----------> [-0.2 , 0.7]

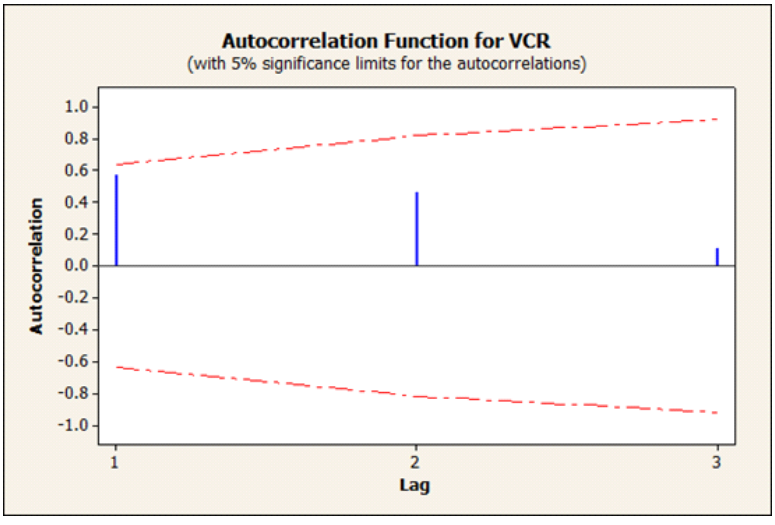
'r_4 ---------> Y_t and Y_{t-4} ---------> 0.6 ----------> [-0.1 , 0.9]

'r_5 ---------> Y_t and Y_{t-5} ---------> 0.75 ----------> [0.2 , 0.8]

'r_6 ---------> Y_t and Y_{t-6} ---------> 0.3 ----------> [0.1 , 0.6]

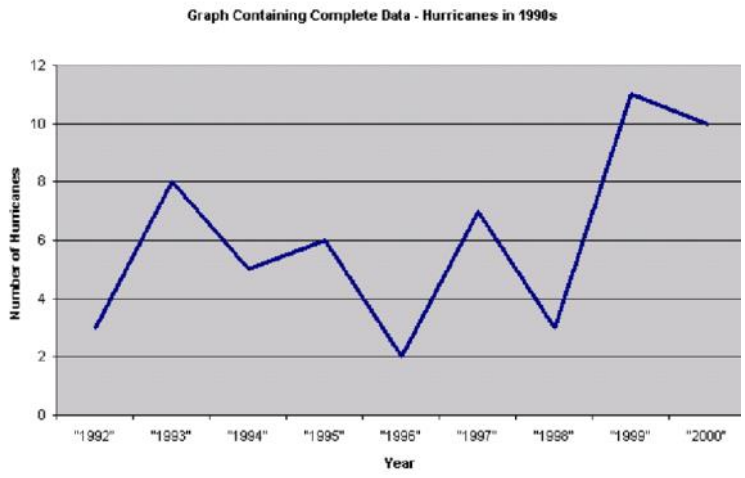
Which Auto correlation value is highly significant and not following with in this range that specific lag variable you are going to choose and will start for model fitting.

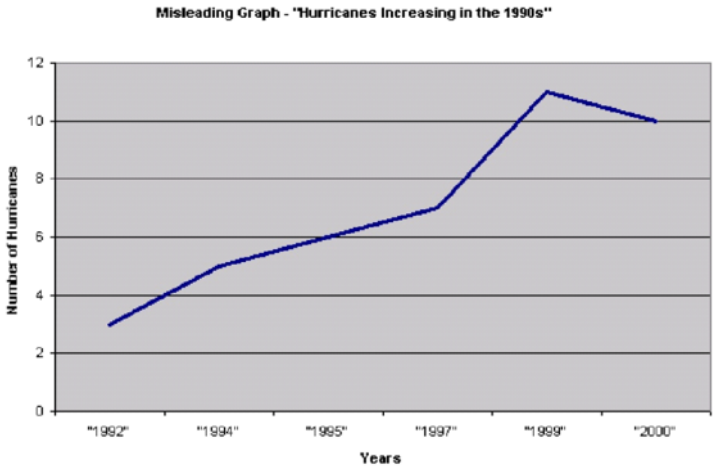
The above all checks that we can covered in the one of the plot will called as " **ACF plot** "



Red color line is the Standard Error line, we have to verify that which auto correlation is outside of this range, that is statistically high significance farthest from zero, centre.

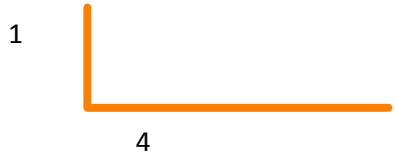
Problems with Time series plots





1. If our data contains missing information we may get wrong interpretation

2. if our data doesn't have a proper scale on y - axis you may not clear what exact relationship. Same with Y and X -axis a well.
3. However, for long time series data aspect ratio should be around 0.25. To understand the impact of aspect ratio see the following two plots. i.e

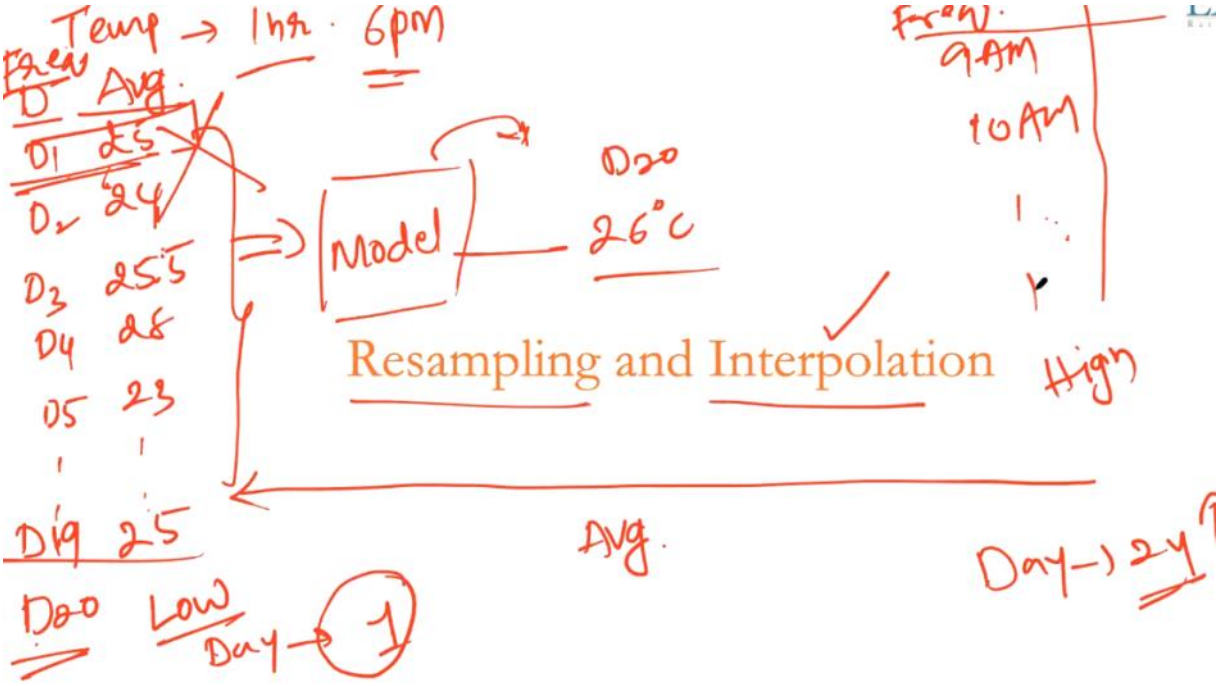


What happens if my data is not in our expectations

We expect the data in hours , but data receives us in Days

Using Re-sampling and interpolation method it is possible.

The frequency of Days would be lower but when we convert to hours, our frequency could be increased.
That means, moving from low to high frequency or high to low is called interpolation.
So, using interpolation technique t is possible.



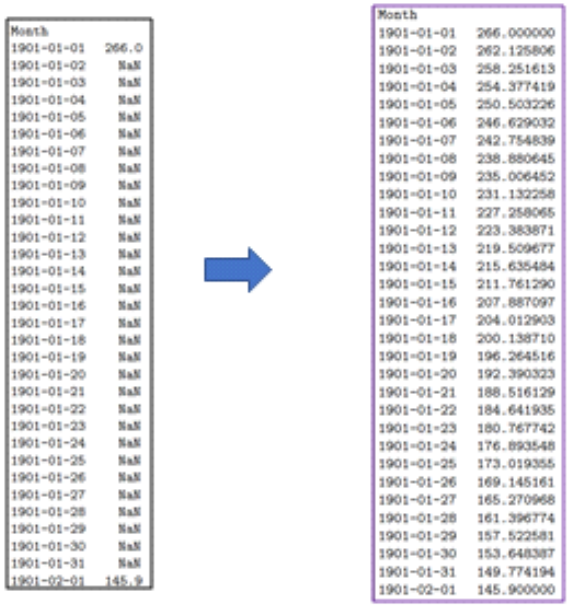
Using the interpolation technique we will do resampling according to our requirements.

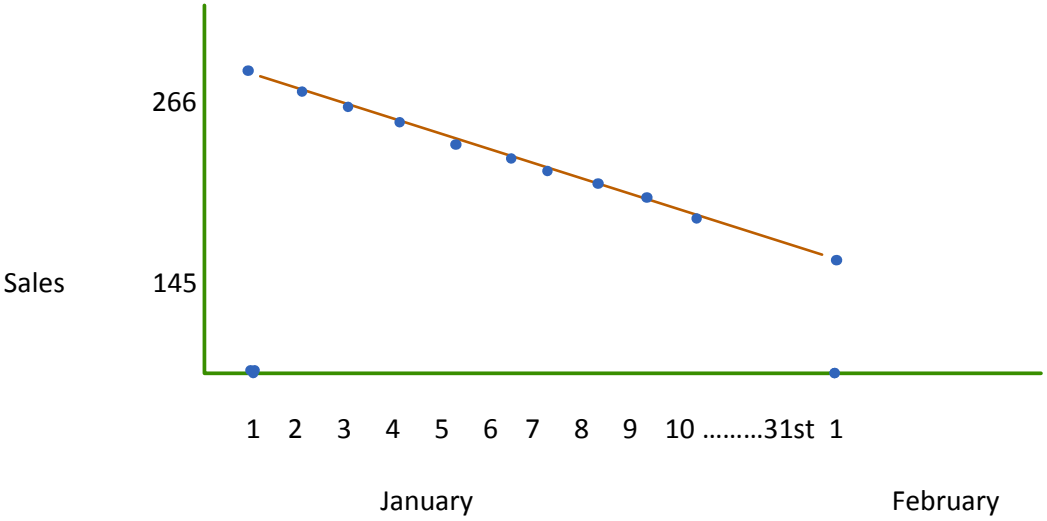
Resampling involves changing the frequency of your time series observations. Two types of resampling are:

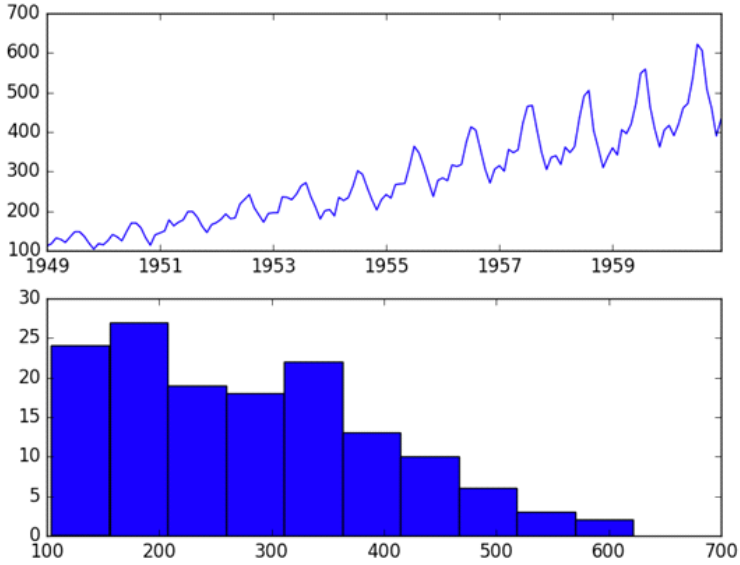
- Up sampling : Where you increase the frequency of the samples, such as from minutes to seconds.
- Down sampling : Where you decrease the frequency of the samples, such as from days to months.

Example for up sampling:

Month	Sales
1-01	266
1-02	145.9
1-03	183.1
1-04	119.3
1-05	180.3
1-06	168.5
1-07	231.8
1-08	224.5
1-09	192.8







The dataset is non-stationary, meaning that the mean and the variance of the observations change over time. This makes it difficult to model for both classical statistical methods, like ARIMA, and more sophisticated machine learning methods, like neural networks.

Using transformation techniques we can solve some of the problems.

Methods:

Forecast method: Last sample $\hat{Y}_{t+1} = Y_t$

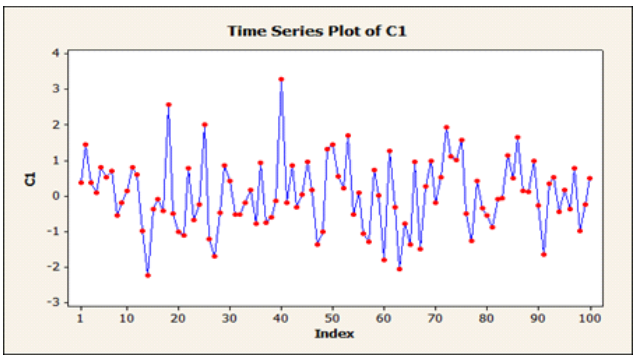
The first basic model to start up
When we don't have any past data we will ensure that yesterday's data as current predicted data

k-step ahead $F_{t+k} = Y_t$

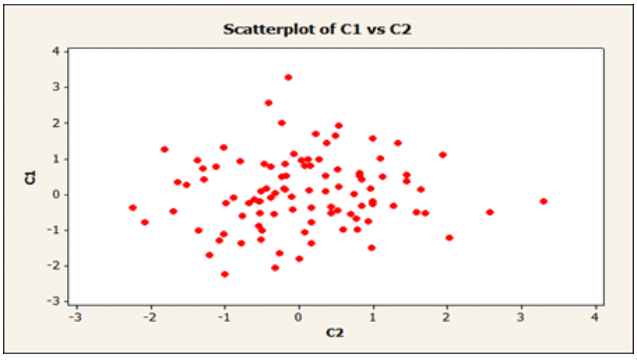
K = number of days you are adding in to predicted with future data.

However once you fitted the model we will have errors

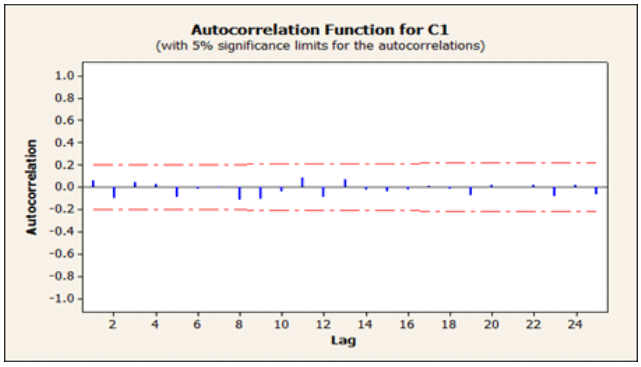
1. If you plot error on y-axis. We should not see any patterns then we can understand that our model is good.



2. If you don't see any relationship between et and et-1 then it is a good model.



3. if our data points are out of red color lines. Then the model is not so good.



4. Errors on Histogram should be normally distributed