Time Series Modeling Steps

1. Plot the time series data
2. Check If the data are stationary with the help of Dickey fuller test: if found non stationary then take first differences of the data until the data becomes stationary.
3. Check whether data contains seasonality. If yes, two options - either take seasonal differencing or fit seasonal arima model.
4. Identify orders of p,d and q by examining the ACF/PACF.
5. Try your chosen models, and use the AICC/BIC to search for a better model.
6. Evaluate the model with the help of MSE and RMSE.
7. Once above step are completed, calculate forecasts.

Many of the simple time series models are special cases of ARIMA Model

1. Simple Exponential Smoothing ARIMA(0,1,1)
2. Holt's Exponential Smoothing ARIMA(0,2,2)
3. Autoregession ARIMA(p,0,0)
4. Moving average ARIMA(0,0,q)
5. ARMA Model (p,d,q)
6. ARIMA Model (p,d,q)
7. SARIMAX Model (p,d,q,P,D,Q)
8. SARIMAX with Extragenous Variable (p,d,q,P,D,Q)

- ## What is Time Series ?

A time series is a data series consisting of several values over a time interval. e.g. daily BSE Sensex closing point, weekly sales and monthly profit of a company etc.

Typically, in a time series it is assumed that value at any given point of time is a result of its historical values. This assumption is the basis of performing a time series analysis.

So talking mathematically, Vt = p(Vt-n) + e

It means Value (V) at time "t" is a function of value at time "n" instance ago with an error (e). Value at time "t" can depend on one or various lags of various order.

Example :

Suppose Mr. X starts his job in year 2010 and his starting salary was $5,000 per month. Every years he is appraised and salary reached to a level of $20,000 per month in year 2014. His annual salary can be considered a time series and it is clear that every year's salary is function of previous year's salary (here function is appraisal rating).

- ## Components of a Time Series :

1. **Trend:-** Series could be **constantly increasing or decreasing** or **first decreasing for a considerable time period and then decreasing**. This trend is identified and then removed from the time series in ARIMA forecasting process.

2. **Seasonality:-** Repeating pattern with fixed period.

Example - Sales in festive seasons. Sales of Candies and sales of Chocolates peaks in every October Month and December month respectively every year in US. It is because of Halloween and Christmas falling in those months. The time-series should be de-seasonalized in ARIMA forecasting process.

3. **Random Variation (Irregular Component):-** This is the unexplained variation in the time-series which is totally random. Erratic movements that are not predictable because they do not follow a pattern.
Example - Earthquake

- **Stationary Series**

A stationary series is one whose mean and variance of the series is constant over time.

The series has to be stationary before building a time series with ARIMA. Most of the time series are non-stationary. If series is non-stationary, we need to make it stationary with detrending, differencing etc.

- **White Noise**

A white noise process is one with a constant mean of zero, a constant variance and no correlation between its values at different times. White noise series exhibit a very erratic, jumpy, unpredictable behavior. Since values are uncorrelated, previous values do not help us to forecast future values.
Note:- White noise series themselves are quite uninteresting from a forecasting standpoint (they are no linearly forecastable).

- **Autocorrelation**

Autocorrelation refers to the correlation of a time series with its own past and future values. Autocorrelation is also sometimes called “lagged correlation” or “serial correlation”.

- **Random Walk**

It means past data provides no information about the direction of future movements.

$$Yt = Yt-1$$

It is called **random-walk-without-drift model**: it assumes that, at each point in time, the series merely takes a random step away from its last recorded position, with steps whose mean value is zero.

If the mean step size is some nonzero value α, the process is said to be a **random-walk-with -drift** (slow steady change) whose prediction equation is $\hat{Y}t = Yt\text{-}1 + \alpha$

A random walk process is non-stationary as its mean and variance increases with t.

- **What are the Data Preparation Steps For ARIMA Modeling?**

Check if there is variance that changes with time – Volatility. For ARIMA, the volatility should not be very high.

If the volatility is very high, we need to make it non-volatile.

Check for Stationary – a series should be stationary before performing ARIMA.

If data is non-stationary, we need to make it stationary.

Check for Seasonality in the data

- Augmented Dickey-Fuller Test

The Dickey-Fuller test is the statistical test used to find whether the time series is stationary or not.

Null Hypothesis : Non-Stationary
Alternative Hypothesis : Stationary

If the p value is less than 0.05 than the series is Stationary at 5% level of significance.

There are three types by which you can calculate test statistics of dickey-fuller test.

- Zero Mean - No Intercept. Series is a random walk without drift.
- Single Mean - Includes Intercept. Series is a random walk with drift.
- Trend - Includes Intercept and Trend. Series is a random walk with linear trend.

How to Make Non-Stationary Data to Stationary

Differencing process is used for making the series stationary.

Differencing *:* Transformation of the series to a new time series where the values are the differences between consecutive values

Differencing Procedure may be applied consecutively more than once, giving rise to the **"first differences"**, **"second differences"**, etc.

Differencing Orders :

1st order : ⍰xt = xt – xt-1. **For eg. Sales - lag1(Sales)**

2nd order : ⍰2xt = (⍰xt - ⍰xt-1)=xt – 2xt-1 + xt-2

Note:- It is unlikely that more than two differencing orders would ever be required.

Split Data into Training and Validation

Spliting data into **Training and Validation** samples requires recent data for validation and remaining data be used to train the model. We would develop ARIMA model and forecast on Testing part and would check the results on Validation part.

Note:- We cannot use random sampling like we do in regression models to split the data.

- **Explain Cyclic Variation?**

The variation of observations in a time series occurring generally in business and economics where the rises and falls in the data are not of fixed period is known as Cyclic Variation.

The duration of these cycles is more than a year.

Example:- Sensex Price

- **What are the conditions are satisfied then a time series is stationary?**
- Mean is constant and does not depend on time
- Autocovariance function depends on s and t only through their difference |s-t| (where t and s are moments in time)
- The time series under considerations is a finite variance process

These conditions are essential prerequisites for mathematically representing a time series to be used for analysis and forecasting. Thus stationarity is a desirable property.

- Why Stationary?

To calculate the expected value, we generally take a mean across time intervals. The mean across many time intervals makes sense only when the expected value is the same across those time periods. If the mean and population variance can vary, there is no point estimating by taking an average across time.

- Lag

Lag is basically value at a previous point of time.

In simpler terms, shifting series to 1 timeframe downward means 1 lag and shifting series to 2 timeframe downward means 2 lag.

- **Auto-regressive Model**

It implies relationship of a value of a series at a point of time with its own previous values. Such relationship can exist with any order of lag.

Autoregressive models are based on the idea that the current value of the series, xt, can be explained as a function of p past values.

- **Autocorrelation Function (ACF)**

Autocorrelation is a correlation coefficient. However, instead of correlation between two different variables, the correlation is between two values of the same variable at times Xt and Xt-h. Correlation between two or more lags.

- **Partial Autocorrelation Function (PACF)**

For a time series, the partial autocorrelation between xt and xt-h is defined as the conditional correlation between xt and xt-h, conditional on xt-h+1, ... , xt-1, the set of observations that come between the time points t and t−h.

In simpler terms, PACF is same as ACF just that the intermediate lags between t and t-p is removed i.e. correlation between Y(t) and Y(t-p) with p-1 lags excluded.

- **Order of ARIMA**

The order of an ARIMA (autoregressive integrated moving-average) model is usually denoted by the notation ARIMA(p,d,q) or it can be read as AR(p) , I(d), MA(q), where

p = Order of Autoregression (Individual values of time series can be described by linear models based on preceding observations.

For instance: x(t) = 3 x(t-1) - 4 x(t-2))

d = Order of differencing (No. of times data to be differenced to become stationary)

q = Order of Moving Average (Number of lagged forecast errors in the prediction equation. Past estimation or forecasting errors are taken into account when estimating the next time series value. The difference between the estimation x(t) and the actually observed value x(t) is denoted ε(t).

For instance: x(t) = 3 ε(t-1) - 4 ε(t-2).)