Data Analysis and Decision Making

***RUTGERS BUSINESS SCHOOL - NEWARK***

*FORECASTING & TIME-SERIES*

***of the S&P 500 Stock Index***

MEMBERS:

1

Data Analysis and Decision Making

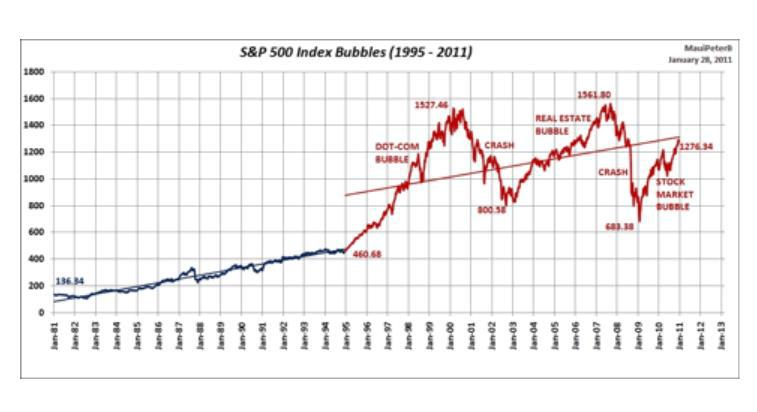
*FORECASTING AND TIME SERIES*

**Time-Series Analysis and Forecasting of the S&P 500 Stock Index:**

* Stock Market - Standard & Poor's 500 Index is one of the best representations of the U.S. stock market. S&P 500 index measures performance of 500 large companies in US.
* Data Source: Yahoo Finance

Data Fields: Open, close, Low, High and Adjusted Closing value

* We are using **Adjusted close** field for our analysis and predictions as it accurately reflects after adjustments of all splits and dividend distributions.



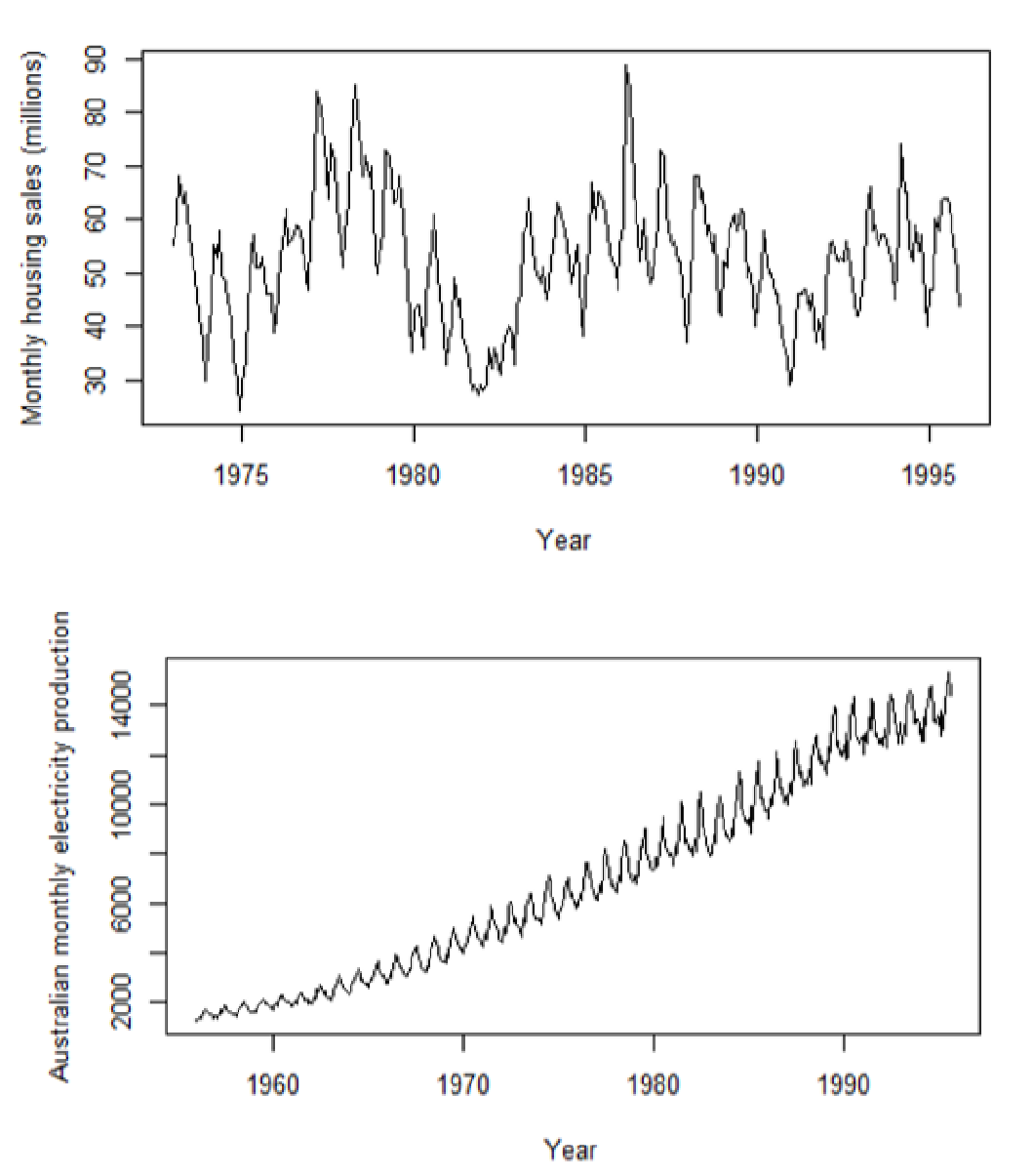
**Project Goal:**

* Time Series Analysis of S&P 500 Stock index
* Models: Mean, Naïve, Seasonal Naïve, ARIMA
* Which model fits better
* Forecasting future Stock index value

Data Analysis and Decision Making

**Time Series:**

* Data that exists over a continuous time interval with equal spacing between every two consecutive measurements.
* Can be used to understand the past as well as predict the future
* Examples: Monthly housing sales, Average Daily temperatures, Daily closing values of indexes



Data Analysis and Decision Making

**Time Series Components:**

* Trend:

A trend exists when there is a long-term increase or decrease in the data. It does not have to be linear. Sometimes we will refer to a trend "changing direction" when it might go from an increasing trend to a decreasing trend.

* Seasonal:

A seasonal pattern exists when a series is influenced by seasonal factors (e.g., the quarter of the year, the month, or day of the week). Seasonality is always of a fixed and known period.

* Autocorrelation:

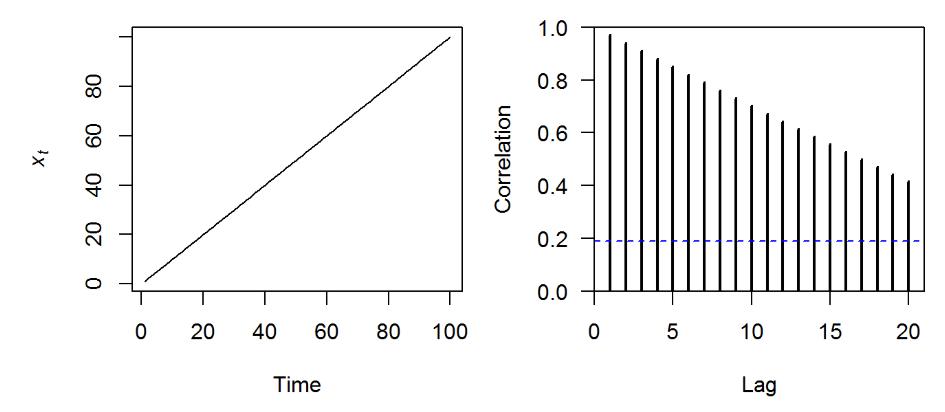
The similarity between observations as a function of the time lag between them.

* If a series is,

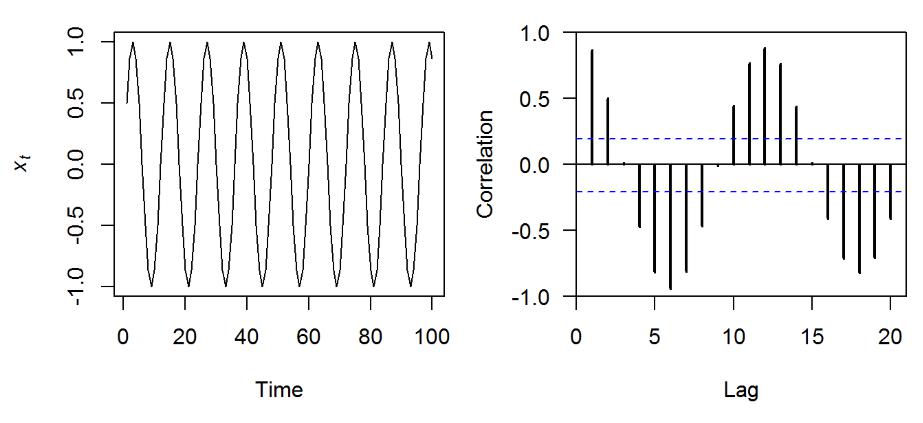
Random = autocorrelation coefficients (ACs) are close to zero

Trendy = ACs for the first several time lags are high, and then decrease

Seasonal = significant ACs will occur at the seasonal time lag

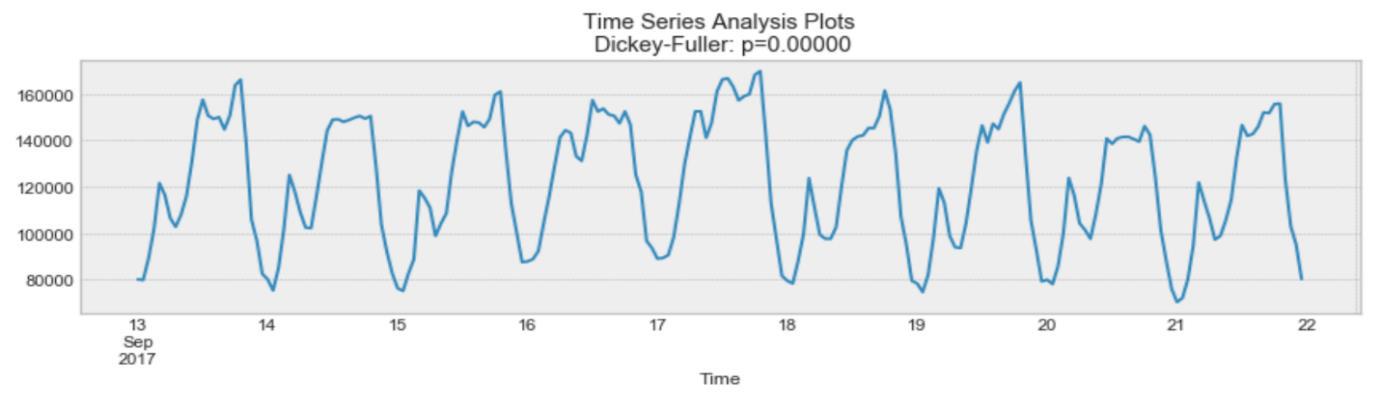


Data Analysis and Decision Making



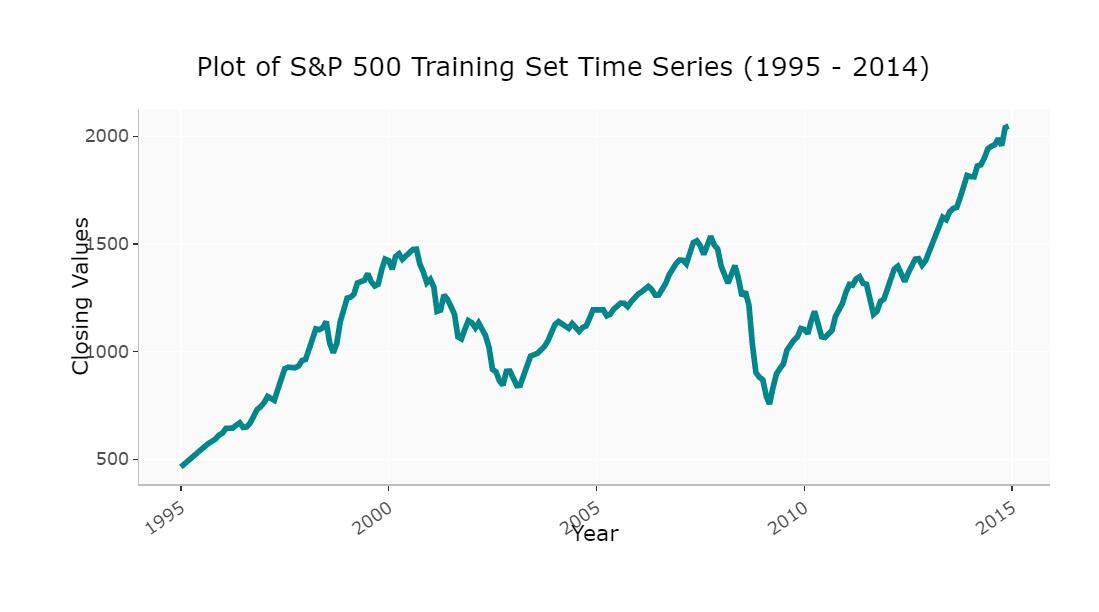
* Stationarity:

Time series is stationary if its statistical properties do not change over time i.e. it has constant mean and variance.



Above, process looks Stationary.

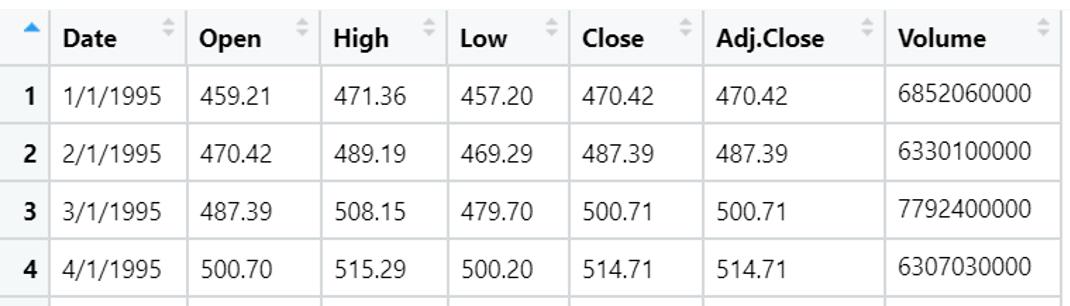
Data Analysis and Decision Making



* Growing Trend, Variance increases
* Volatility increases over time
* Mean is not constant over time So, **S&P500 is NON-Stationary**
* Tests like Augmented Dickey-Fuller Test are implemented to check if the data is stationary or not.

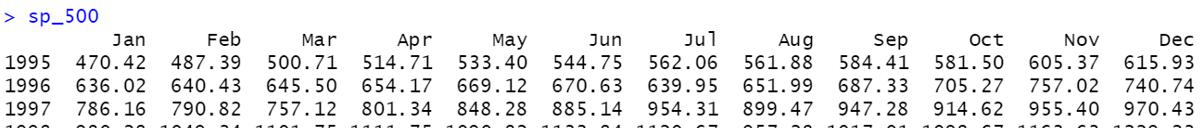
**Creating a Time series from S&P 500 index data:**

**Data95tooct2019.csv**



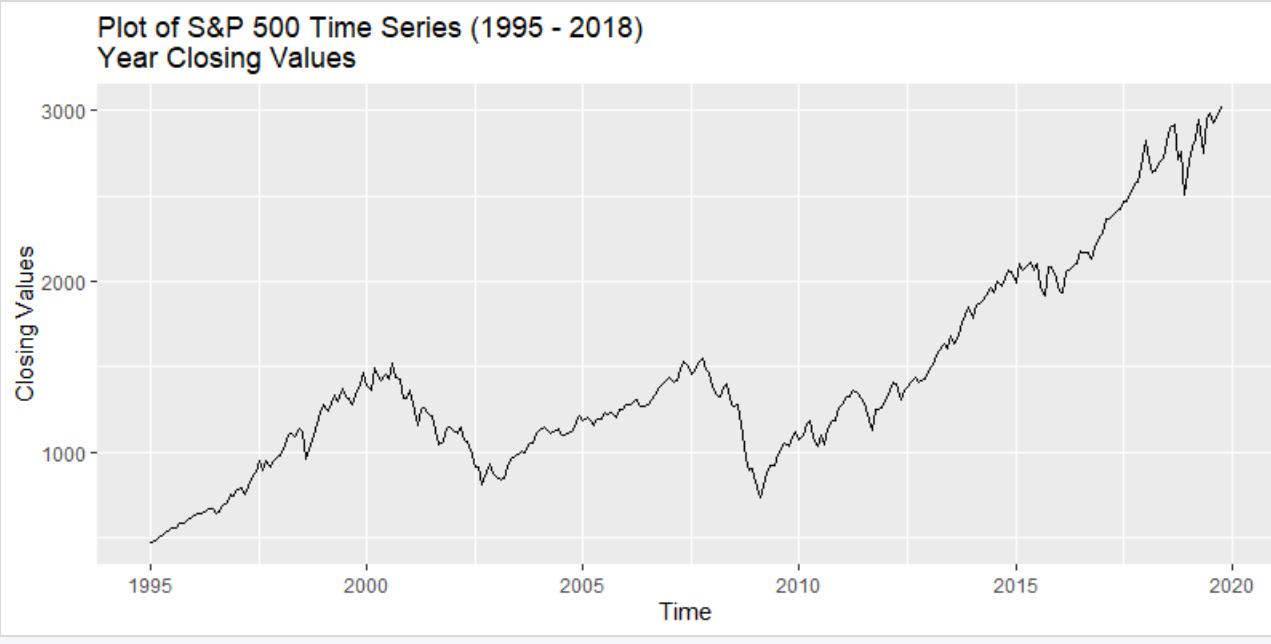
**data\_master <- read.csv("C:/Semester1/DADM\_Project/Data95tooct2019.csv")**

**sp\_500 <- ts(data\_master$Adj.Close, start=c(1995, 1), freq=12)**



View(sp\_500)

Data Analysis and Decision Making

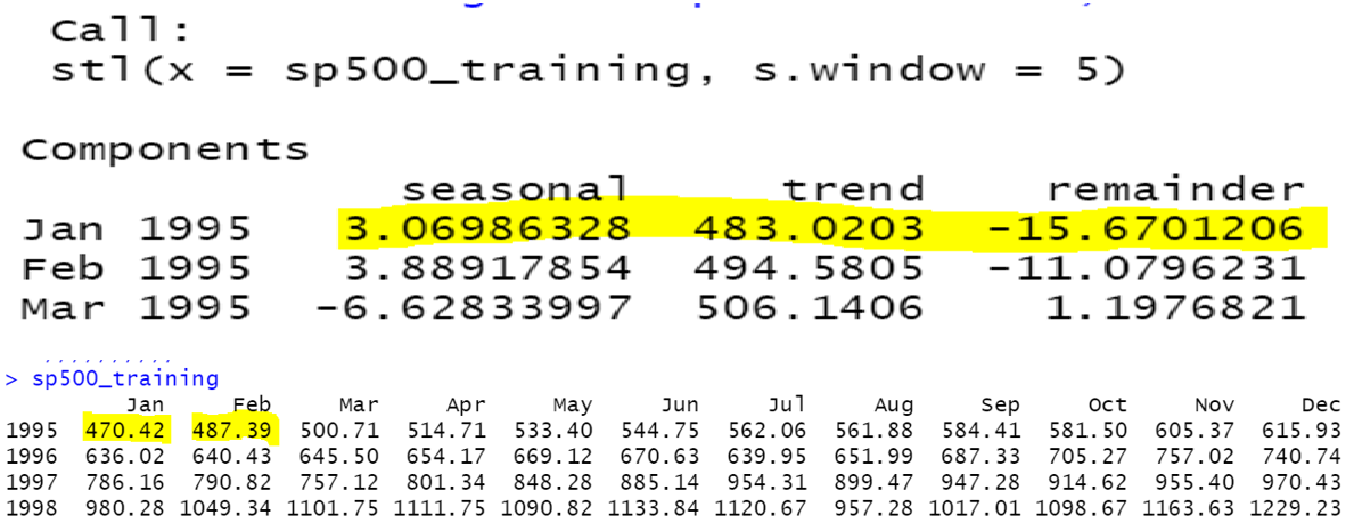


**Time Series Decomposition:**

* 3 Main Components of Time series: Seasonality, Trend and remainder/Error

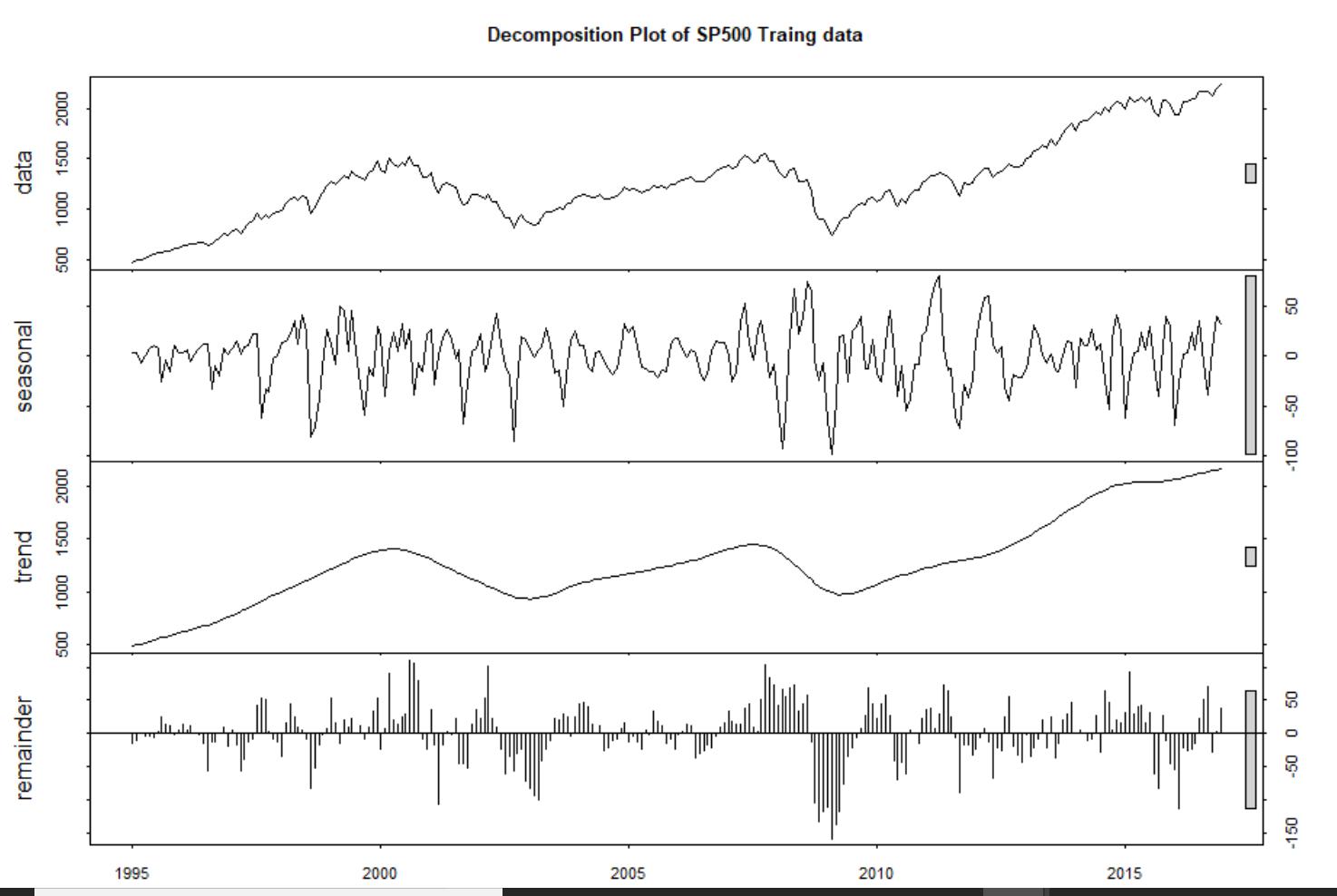
**= St + Tt + Et**

* Time series components Separated by **STL Decomposition** in R.



**Plotting STL for S&P500 index:**

Data Analysis and Decision Making



Data Analysis and Decision Making

**TIME SERIES FORECASTING METHODS:**

1. SIMPLE AVERAGE METHOD:

• Uses the mean of all the relevant historical observations as the forecast of the next period.

* If we let the historical data be denoted by 1, …, , then we can write the forecasts as

Example - Forecasts for time t5 -> t8 given t1->t4 using simple average method.



|  |  |
| --- | --- |
| Time | Yt |
| 1 | 52 |
| 2 | 56 |
| 3 | 45 |
| 4 | 50 |
| 5 | 50.75 |
| 6 | 50.75 |
| 7 | 50.75 |
| 8 | 50.75 |

1. NAÏVE METHOD:

* This method is only appropriate for time series data.
* All forecasts are simply set to be the value of the last observation.
* That is, the forecasts of all future values are set to be, whereis the last observed value.
* This method works remarkably well for many economic and financial time series.

Example - Forecasts for time t5 -> t8 given t1->t4 using Naïve method.

|  |  |
| --- | --- |
| Time | Yt |
| 1 | 52 |
| 2 | 56 |
| 3 | 45 |
| 4 | 50 |
| 5 | 50 |
| 6 | 50 |
| 7 | 50 |
| 8 | 50 |

Data Analysis and Decision Making

1. Seasonal Naïve Forecasting Method:

A similar method is useful for highly seasonal data. In this case, we set each forecast to be equal to the last observed value from the same season of the year (e.g., the same month of the previous year).

Formally, the forecast for time **T+h**

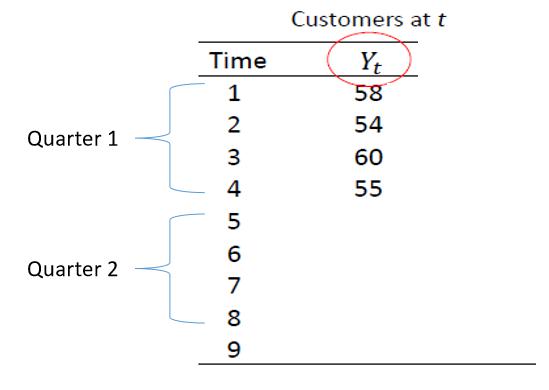
is written as



where **m=** the seasonal period, and **k** is the integer part of (h−1)/m (i.e., the number of complete

years in the forecast period prior to time T+h). This looks more complicated than it really is. For example, with monthly data, the forecast for all future February values is equal to the last observed February value. With quarterly data, the forecast of all future Q2 values is equal to the last observed Q2 value (where Q2 means the second quarter). Similar rules apply for other months and quarters, and for other seasonal periods.

Example:



Here the forecast values for quarter 2 is:

|  |  |
| --- | --- |
| Time | Yt |
|  |  |
| 5 | 58 |
|  |  |
| 6 | 54 |
|  |  |
| 7 | 60 |
|  |  |
| 8 | 55 |
|  |  |

Data Analysis and Decision Making

1. ARIMA:

ARIMA, short for ‘Auto Regressive Integrated Moving Average’ is actually a class of models that ‘explains’ a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values.

Any ‘non-seasonal’ time series that exhibits patterns and is not a random white noise can be modelled with ARIMA models.

An ARIMA model is characterized by 3 terms: p, d, q where,

p is the order of the AR term

q is the order of the MA term

d is the number of differencing required to make the time series stationary

If a time series, has seasonal patterns, then you need to add seasonal terms a nd it becomes SARIMA, short for ‘Seasonal ARIMA’.

**p, d and q in ARIMA model.**

The first step to build an ARIMA model is to [make the time series stationary.](https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/www.machinelearningplus.com/stationary-time-series)

Because, term ‘Auto Regressive’ in ARIMA means it is a [linear regression model](https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/) that uses its own lags as predictors. Linear regression models, as we know, work best when the predictors are not correlated and are independent of each other.

To make a series stationary-the most common approach is to difference it. That is, subtract the previous value from the current value. Sometimes, depending on the complexity of the series, more than one differencing may be needed.

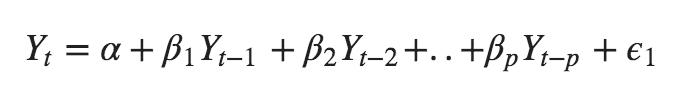
The value of d, therefore, is the minimum number of differencing needed to make the series stationary. And if the time series is already stationary, then d = 0.

Now, ‘p’ is the order of the ‘Auto Regressive’ (AR) term. It refers to the number of lags o f Y to be used as predictors. And ‘q’ is the order of the ‘Moving Average’ (MA) term. It refers to the number of lagged forecast errors that should go into the ARIMA Model.

**AR and MA models**

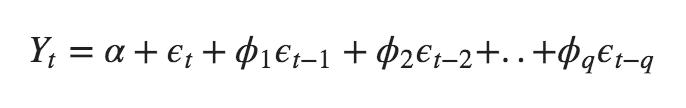
A pure **Auto Regressive (AR only) model** is one where Yt depends only on its own lags.

That is, Yt is a function of the ‘lags of Yt’.



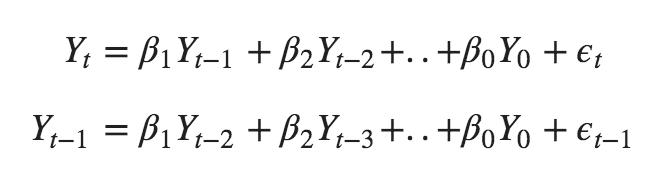
where, $Y*{t-1}$ is the lag1 of the series, $\beta*1$ is the coefficient of lag1 that the model estimates and $\alpha$ is the intercept term, also estimated by the model.

Likewise, a pure **Moving Average (MA only) model** is one where Yt depends only on the lagged forecast errors.



where the error terms are the errors of the autoregressive models of the respective lags. The errors Et and E(t-1) are the errors from the following equations:

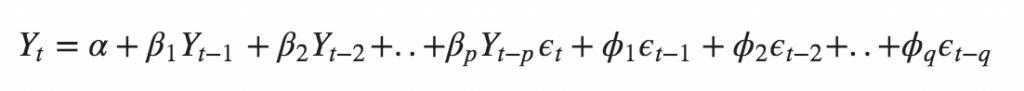
Data Analysis and Decision Making



That was AR and MA models respectively.

Also,

An ARIMA model is one where the time series was differenced at least once to make it stationary and you combine the AR and the MA terms. So, the equation becomes:



**ARIMA model in words:**

**Predicted Yt = Constant + Linear combination Lags of Y (up to p lags) + Linear Combination of Lagged forecast errors (up to q lags)**

The objective, therefore, is to identify the values of p, d and q.

**Finding the order of differencing (d) in ARIMA model**

The purpose of differencing it to make the time series stationary.

But we need to be careful to not over-difference the series. Because, an over differenced series may still be stationary, which in turn will affect the model parameters.

The right order of differencing is the minimum differencing required to get a near -stationary series which roams around a defined mean and the ACF plot reaches to zero fairly quick. If the autocorrelations are positive for many numbers of lags (10 or more), then the series needs further differencing. On the other hand, if the lag 1 autocorrelation itself is too negative, then the series is probably over-differenced.

In the event, you can’t really decide between two orders of differencing, then go with the order that gives the least standard deviation in the differenced series.

Let’s see how to do it with an example.

First, we are going to check if the series is stationary

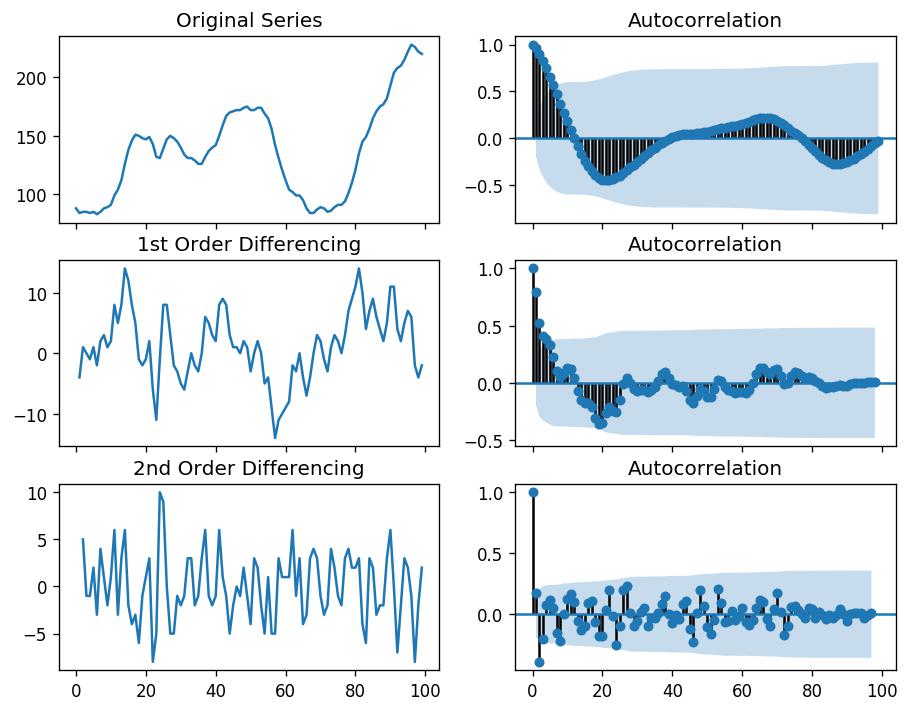
This check is because, we need differencing only if the series is non -stationary. Else, no differencing is needed, that is, d=0.

The null hypothesis of the ADF test is that the time series is non -stationary. So, if the p-value of the test is less than the significance level (0.05) then you reject the null hypothesis and infer that the time series is indeed stationary.

So, in our case, if P Value > 0.05 we go ahead with finding the order of differencing.

If P-value is greater than the significance level, then we difference the series.

Data Analysis and Decision Making



ORDER OF DIFFERENCING

For the above series, the time series reaches stationarity with two orders of differencing. But on looking at the autocorrelation plot for the 2nd differencing the lag goes into the far negative zone fairly quick, which indicates, the series might have been over differenced.

So, we are going to tentatively fix the order of differencing as 1 even though the series is not perfectly stationary (weak stationarity).

**Find the order of the AR term (p)**

The next step is to identify if the model needs any AR terms. You can find out the required number of AR terms by inspecting the Partial Autocorrelation (PACF) plot.

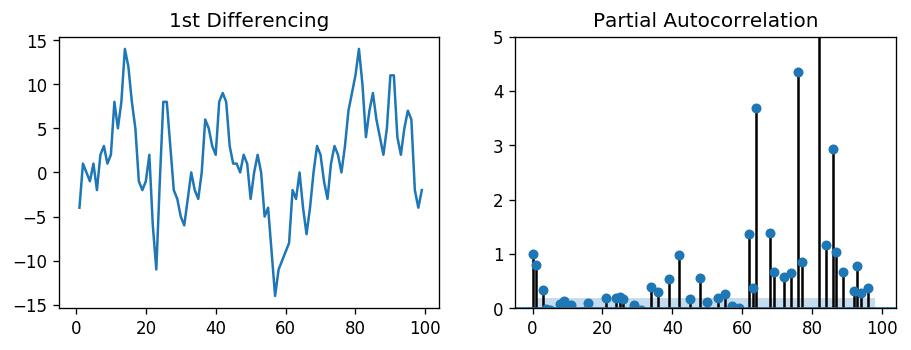
Partial autocorrelation can be imagined as the correlation between the series and its lag, after excluding the contributions from the intermediate lags. So, PACF sort of conveys the pure correlation between a lag and the series. That way, you will know if that lag is needed in the AR term or not.

Partial autocorrelation of lag (k) of a series is the coefficient of that lag in the autoregression equation of Y.

Data Analysis and Decision Making

Finding the number of AR terms-

Any autocorrelation in a stationarized series can be rectified by adding enough AR terms. So, we initially take the order of AR term to be equal to as many lags that crosses the significance limit in the PACF plot.



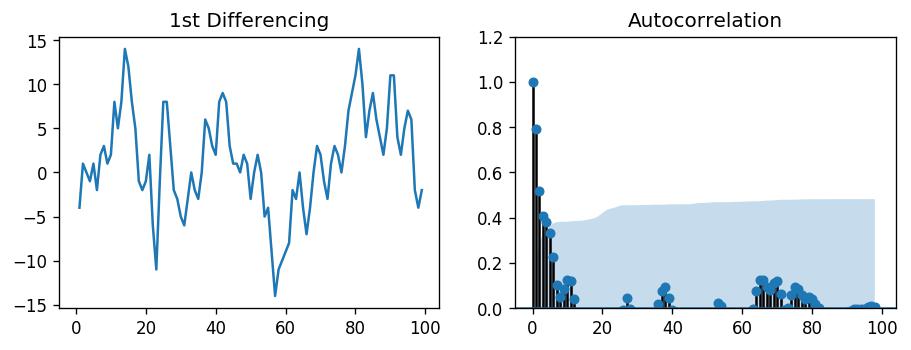
ORDER OF AR TERM

We can observe that the PACF lag 1 is quite significant since is well above the significance line. Lag 2 turns out to be significant as well, slightly managing to cross the significance limit (blue region). But I am going to be conservative and tentatively fix the p as 1.

**Finding the order of the MA term (q)**

Just like how we looked at the PACF plot for the number of AR terms, we can look at the ACF plot for the number of MA terms. An MA term is technically, the error of the lagged forecast. The ACF tells how many MA terms are required to remove any autocorrelation in the stationaries series.

Let’s see the autocorrelation plot of the differenced series.



ORDER OF MA TERM

Couple of lags are well above the significance line. So, let’s tentatively fix q as 2. When in doubt, go with the simpler model that sufficiently explains the Y.

Data Analysis and Decision Making

**Handling time series if it is slightly under or over differenced**

It may so happen that your series is slightly under differenced, that differencing it one more time makes it slightly over-differenced. In such case, if our series is slightly under differenced, adding one or more additional AR terms usually makes it up. Likewise, if it is slightly over - differenced, try adding an additional MA term.

**Building the ARIMA Model**

Now that we have determined the values of p, d and q, we have everything needed to fit the ARIMA model by using the ARIMA () implementation.



**Forecast KPI (Key Performance Index)**

Mean Forecast Error



Mean forecast error shows the deviation of a forecast from actual demand. This is the mean of the differences per period between a number of period forecasts and the actual demand for the corresponding periods.

Mean forecast error is used as a basis for following up and adjusting forecasts. When it is positive, the forecasts have been low in relation to actual demand and when it is negative, the forecasts have been too high.

Let yi denote the h observation and ̂i denote a forecast of yi

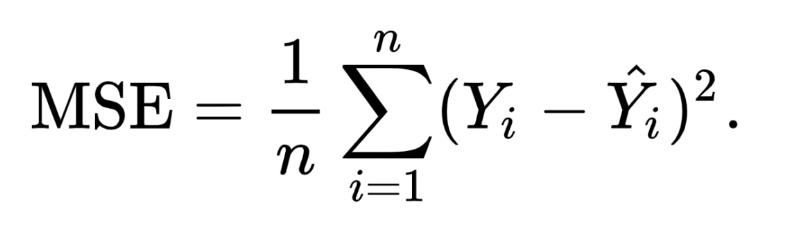
The forecast error is simply ei = yi − ̂i, which is on the same scale as the data. **(This is also called**

**RESIDUAL)**

• Accuracy measures that are based on are therefore scale-dependent and

cannot be used to make comparisons between series that are on different scales.

MEAN SQUARED ERROR:



Data Analysis and Decision Making



When comparing forecast methods on a single data set, the MAE is popular as it is easy to understand and compute.

MSE prefers many small errors to one large error.

MAPE is a good measure to explain and interpret the model accuracy. Use MSE if we do not want large errors. Use MAPE if concerned with interpretation.

The smaller value, better it is.

Percentage Errors

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| The | percentage | error | is | given | by |  | = | 100 / |
| Percentage | errors | have | the | advantage | of | being | scale-independent | |

Frequently used to compare forecast performance between different data sets ▪ The most commonly used measure is:



Akaike’s Information Criterion (AIC) and Bayesian information criterion (BIC):

The Akaike Information Criteria (AIC) is a widely used measure of a statistical model. It basically quantifies 1) the goodness of fit, and 2) the simplicity/parsimony, of the model into a single statistic.

When comparing two models, the one with the lower AIC is generally “better”.

Akaike’s Information Criterion, which was useful in selecting predictors for regression, is also useful for determining the order of an ARIMA model. It can be written as

*AIC* = −2log(*L*) + 2(p + *q* + *k* +1),

where *L* is the likelihood of the data, *k*=1 if *c*≠0 and *k*=0 if *c*=0. Note that the last term in parentheses is the number of parameters in the model (including *σ2*, the variance of the residuals).

* For ARIMA models, the corrected AIC can be written as 2(*p*+*q*+*k* +1) (*p*+*q*+*k* +2)

*AICc* =*AIC*+*T*−*p*−*p*−*k*−2.

Data Analysis and Decision Making

the Bayesian Information Criterion can be written as



*BIC* = *AIC* + (log (*T)* − 2) (p+ *q* + *k* + 1).

* Good models are obtained by minimizing either the *AIC, AICc* or *BIC*.

**AIC should be used for selecting a model intended for *prediction* while BIC should be used for selecting a model for *explanation*.**

R-Squared:

R-Squared (R² or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by

the [independent variable.](https://corporatefinanceinstitute.com/resources/knowledge/modeling/independent-variable/) In other words, r-squared tells how well the data fit the regression model (the goodness of fit).

R-squared can take any values between 0 to 1.

The most common interpretation of r-squared is how well the regression model fits the observed data. For example, an r-squared of 60% reveals that 60% of the data fit the regression model. Generally, a higher r-squared indicates a better fit for the model.

The formula for calculating R-squared is:



Where:

SSregression is the sum of squares due to regression (explained sum of squares)

SStotal is the total sum of squares.

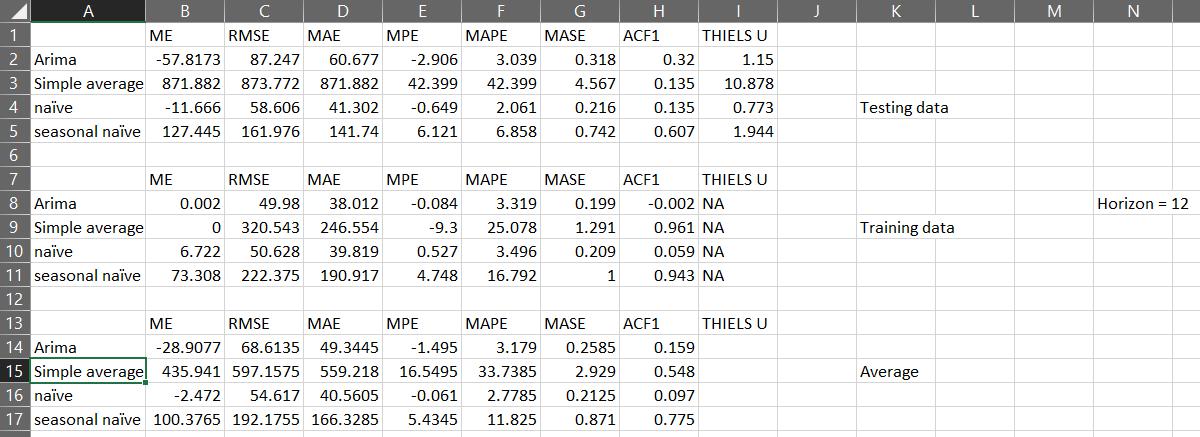
The sum of squares due to regression measures how well the regression model represents the data that were used for modelling. The total sum of squares measures the variation in the observed data (data used in regression modelling).

Out of Simple Average, Naïve, Seasonal Naïve and ARIMA Modelling, In general if we consider MSE as accuracy metric, Simple Average is better. ARIMA is better model for all other metrics.

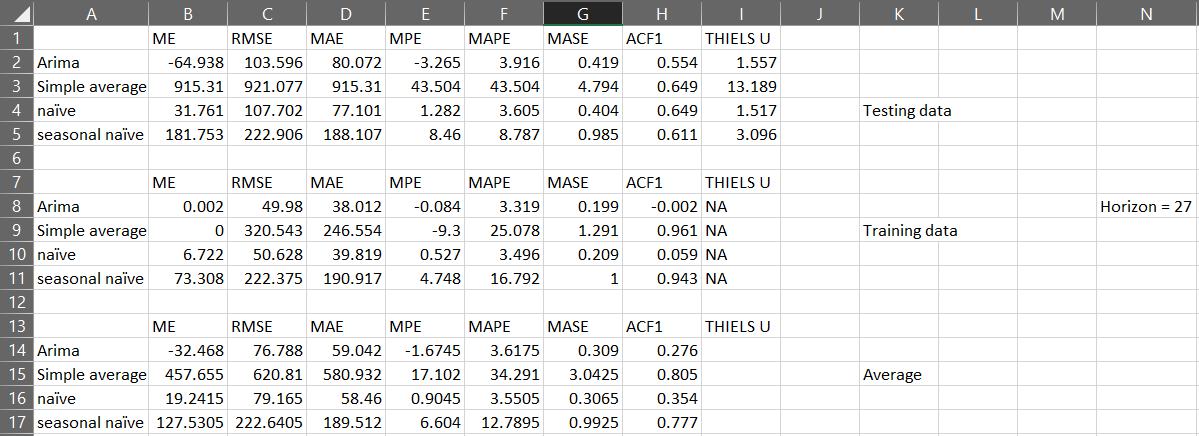
For 2nd part we have taken testing data of 24 months. When we take the average of both training and testing data, we take RMSE as the KPI. ARIMA has the lowest value of 77 and Naïve has value of 79.

**So, ARIMA is the best model for now**.

Comparing models of January 2015 – December 2015 & January 2015 – December 2017



January 2015 – December 2015



January 2015 – December 2017

**Conclusion:**

The forecasting method we used to find the best model receives the lowest MAE and MAPE. The above tables are the results for the test set (the function will return both training and test set metrics, but we're only concerned with the test set metrics).

Surprisingly, other models performed better on the test set than the ARIMA model. This could be a result of overfitting, since our test set was really small. Hence, if we increase the testing data set, we can conclude that ARIMA performs better than other models.

We do, however, encourage that you explore the models with a larger test set.