The Task:

The purpose of the analysis is to forecast the number of customers for a Thai Tourism Company for next 12 months. Thai Tourism Company manages world-wide tour. Recently they were looking into their account of Africa Tour from the year 2010 to 2016. Now they want a prediction of customers for next 12 months i.e., for 2017. So, to solve this problem, the nature of the relationship between Tourist variable and Date must be understood. A statistical model is then adopted to further the analysis and arrive at the results and interpretation.

The Dataset:

The dataset contains the following variables:

**Date**

**Tourists**

The statistical model:

We have adopted the time series model analysis in this case. A time series is used to forecast the value of one variable with the help of the other dynamic nature of the other variable. So, to predict the tourists for next 12 months from the given data with the help of the other variable, the time series model is best suited to the purpose. In the following pages of documentation, the approach steps have been clearly outlined.

Setting up the R model by loading the required libraries:

library(forecast)

Data:

In the next step, the data is read into the R environment from the file.

C:\Users\ADMIN\Desktop\R Models\Time Series\Thai Tourism\data.PNG

Converting into time series data:

data <- ts(data[,2],start = c(2010,1),frequency = 12)

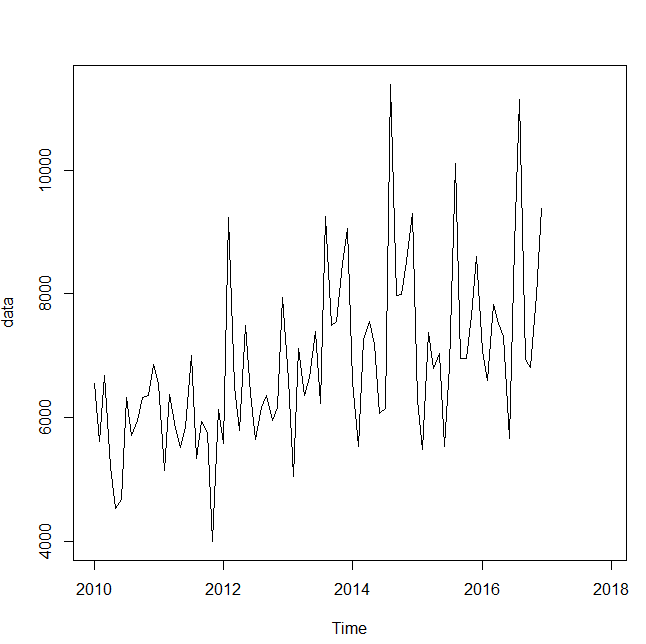
Here we are converting the data from the dataset into a time series data where we have taken the starting point to be first date of 2010 with a 12 months frequency so that the model could run yearly basis. Next we have to check whether the dataset is **Stationary** or not as we won’t be able to do analysis on a **Non-Stationary** dataset.

What is Stationary dataset?:

* A time series dataset is stationary if it doesnot have any trend.
* We should also ensure that there is no systematic change in variance.

Plotting the data to have a look at it:

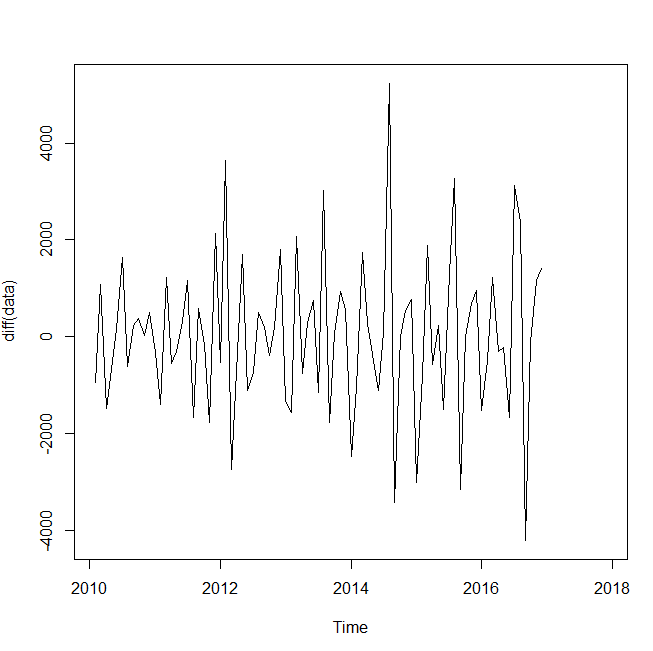
plot(data)



An upward rising trend has been found.

**Differencing the data to get rid off the trend:**

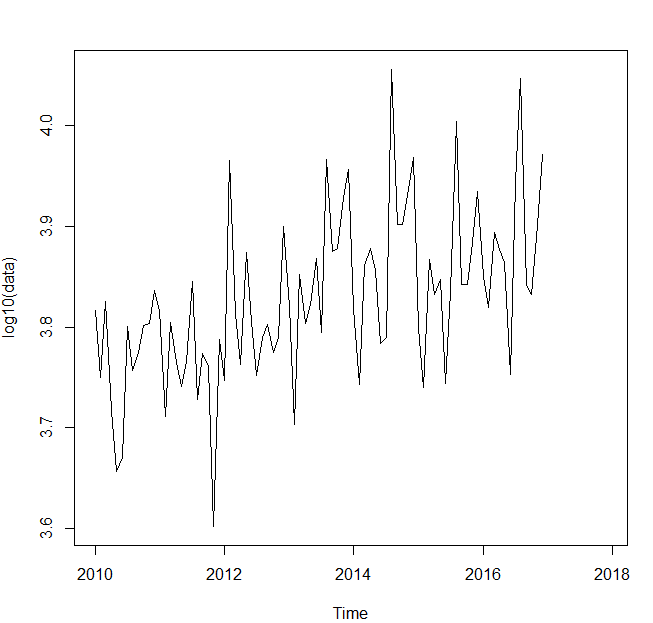
plot(diff(data))



After removing the trend we can see there is uneven variance in the dataset which is also known as **Seasonality**.

Log transforming to get rid off Seasonality:

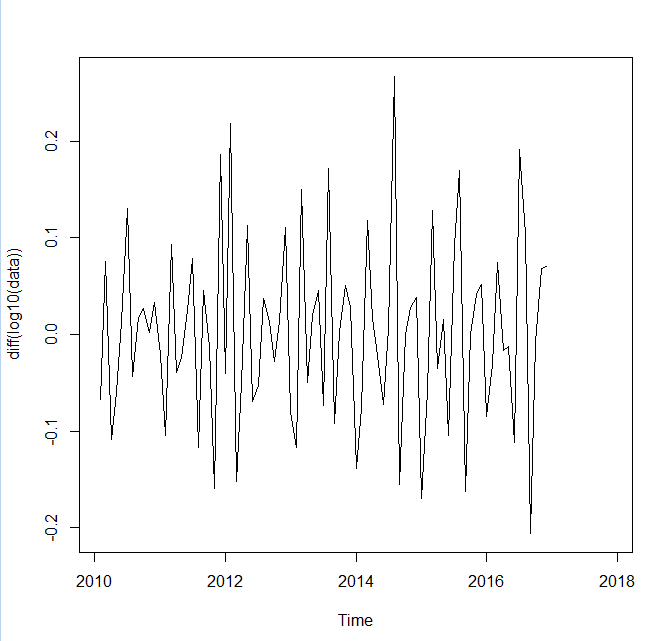
plot(log10(data))



By log transforming we can see there is no uneven variance in the dataset but we can clearly get the picture that even by using the functions diff and log one at time, either of the one will perform its action.

**Combining both log and differencing to get rid off trend and uneven variance:**

plot(diff(log10(data)))

****

Now we can see there is neither trend nor seasonality. Thus the dataset has become a stationary data. Now we can proceed further with the analysis.

**Augmented Dickey-Fuller (ADF) Test:**

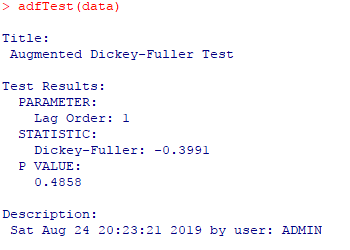
library(fUnitRoots)

adfTest(data)

adfTest(diff(data))

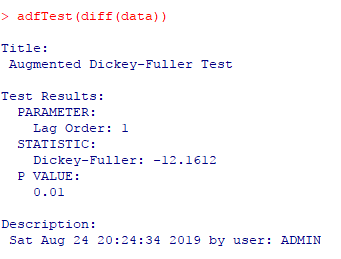
* H0: The series is not stationary.
* H1: The series is stationary.
* Low P-Value indicates the dataset is stationary.

adfTest(data)



Here P-Value is more than 0.05. Thus the dataset is non-stationary.

adfTest(diff(data))



Now after running daftest on the differenced data we can see the P-Value is 0.01 which is lower the 0.05. Thus the dataset is stationary.

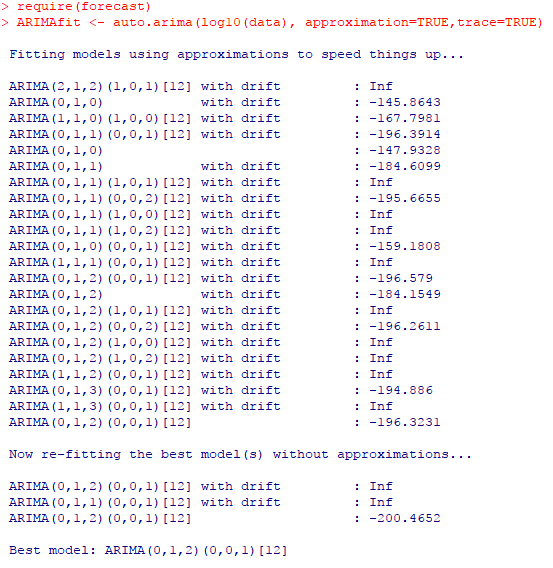
So, here with one time differencing we transformed this non-stationary dataset to a stationary dataset.

Running ARIMA model:

The acronym ARIMA stands for **Auto-Regressive Integrated Moving Average**. Lags of the stationarized series in the forecasting equation are called "autoregressive" terms, lags of the forecast errors are called "moving average" terms, and a time series which needs to be differenced to be made stationary is said to be an "integrated" version of a stationary series.

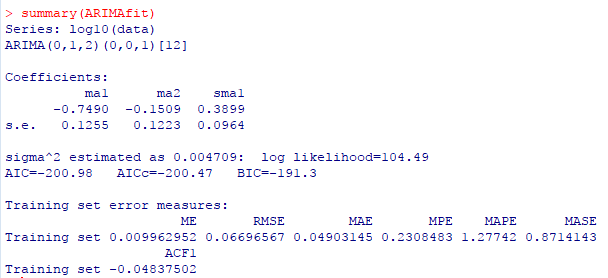
A non-seasonal ARIMA model is classified as an "ARIMA(p,d,q)" model, where:

* **p** is the number of autoregressive terms,
* **d** is the number of nonseasonal differences needed for stationarity, and
* **q** is the number of lagged forecast errors in the prediction equation.



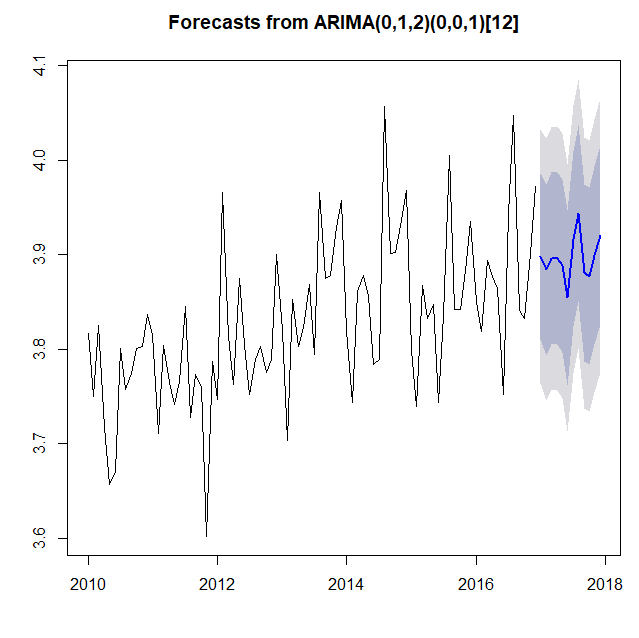
The Arimafit is showing the best model to be Arima(p,d,q) and Seasonality(p,d,q) of the model calculated to be (0,1,2) and (0,0,1) with frequency = 12.

Summary:



The model has calculated minimum mape, rmse, mae, mpe mase at best AIC and BIC.

Plotting the forecasted values:



The blue line indicates that we have successfully integrated the forecasted value with the original dataset.

Showing values:

10^(pred$pred)

pred <- 10^(pred$pred)

write.csv(pred,"Afrofothers$result.csv")

Converting and storing the log values to exponential value to present the forecasted values as general numeric values and then exporting it. We have also merged it with the original dataset to make it easy for the client to observe.

The Business Interpretation

For the Thai Tourism company, as per the above forecasted graph for the year 2017, there will be a slight fall and again it will rise which means there is a lesser chance for number of tourists for the company in the beginning of the year and again there will be another huge decrease in the number of tourists in the beginning of the middle of the year. The number of tourists will increase with an increasing rate at the end of the middle of the year but it is again going to fall but by the end of the year it will rise seemingly. Therefore, as per our analogy the autumn of the year will be prosperous for the company.