2/21/2022

Siddhi Padekar

[company name]

RST Soft Drink Company

Sales Forecast

Table of Contents

|  |
| --- |
| 1. Summary |
| 1. Introduction |
| 1. Reading data and plotting Data |
| * 1. Data Dictionary |
| * 1. Sample of the dataset |
| * 1. Count , mean , max , min |
| 1. Exploratory Data Analysis |
| 2.1 ECDF: Empirical Cumulative Distribution Function |
| 2.2 Histogram graph |
| 2.3 Time series plot diagram |
| 2.4 Month Plot showing monthly distribution |
| 2.5 Decomposition of Data – Additive Decomposition |
| 2.6 Decomposition of Data – Multiplicative Decomposition |
| 1. Split the data into t raining and test. The test data start in 1991. |
| 1. Build various exponential smoothing models on the t raining data and evaluate the model using RMSE on the test data. Other models such as regression, naïve forecast models, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE. |
| 4.1 Linear Regression |
| 4.2 Naiive Forecast Model |
| 4.3 Simple Average Model |
| 4.4 Moving Average |
| 1. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis f or the statistical t est. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05. |
| 1. Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE. |
| 1. Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the   training data and evaluate this model on the t est data using RMSE. |

Executive Summary:

This Data set is of JK Shoe Company, which deals in selling of shoes. The dataset contains details of pair of shoe sales from January 1980 to July 1995. We are expected to forecast the sales of the pairs of shoes for the upcoming 12 months from where the data ends.

Introduction:

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset; forecast the sales of the pairs of shoes for the upcoming 12 months from where the data ends.

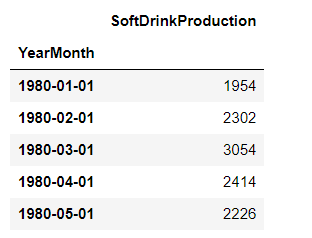
Reading data and plotting Data

Data Dictionary:

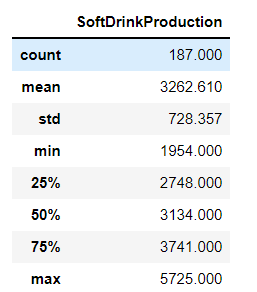
Below is the brief description of data set

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| YearMonth | Year and month |
| Shoe\_Sales | Pair of shoe sold |

Sample of the dataset :

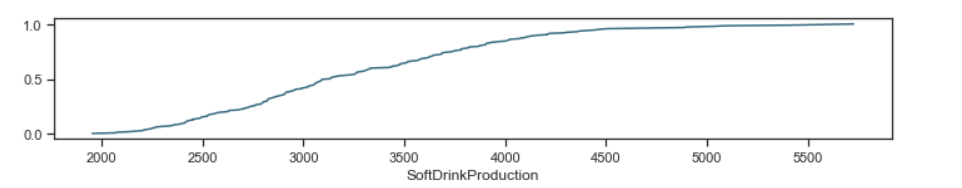


Count , mean , max , min :



Exploratory Data Analysis

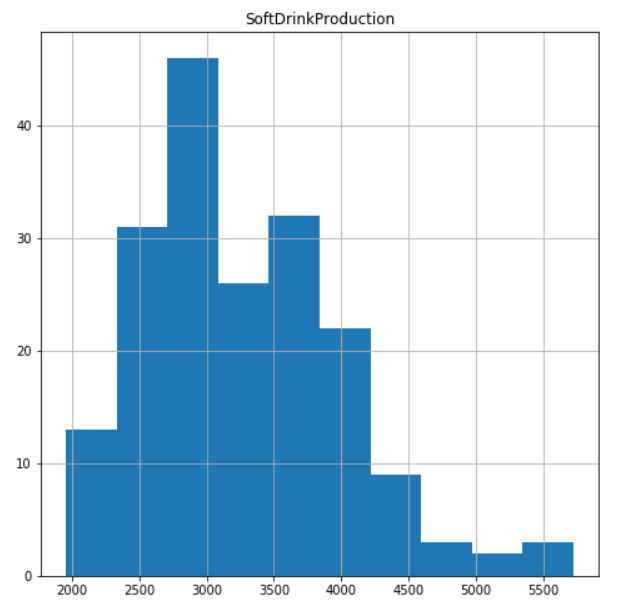
ECDF: Empirical Cumulative Distribution Function :



The empirical distribution function is an estimate of the cumulative distribution function that generated the points in the sample. The ECDF essentially allows you to plot a feature of your data in order from least to greatest and see the whole feature as if is distributed across the data set.

Dataset ranged from 1954 to 5725

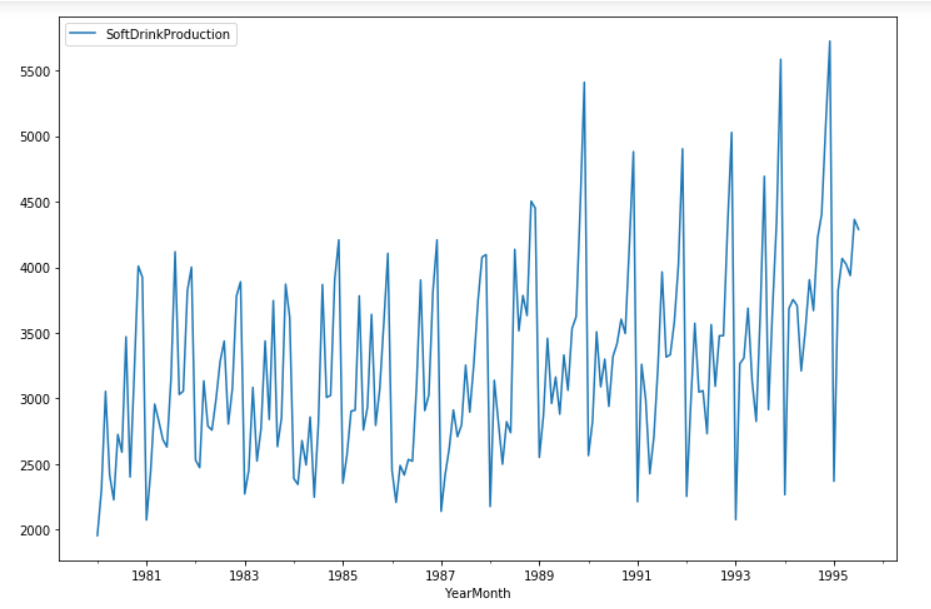
Histogram Graph



Histogram is **a chart that plots the distribution of a numeric variable's values as a series of bars**.

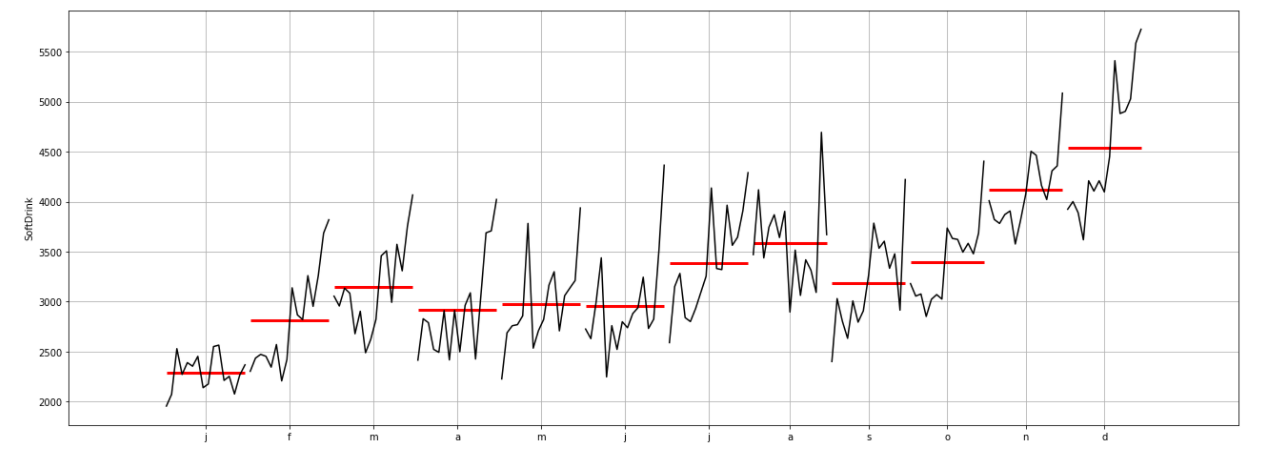
A distribution skewed to the left is said to be negatively skewed. This kind of distribution has a large number of occurrences in the upper value cells (right side) and few in the lower value cells (left side).

Time series plot diagram



The above diagram shows time series of sales data over years from 1980-1995. This is not a constant time series. However, there is a slight increasing trend. There is positive slope of trend. There is repetitive nature or repeatable patter every year. In addition, there is constant seasonality.

Month Plot showing monthly distribution



## Decomposition of data

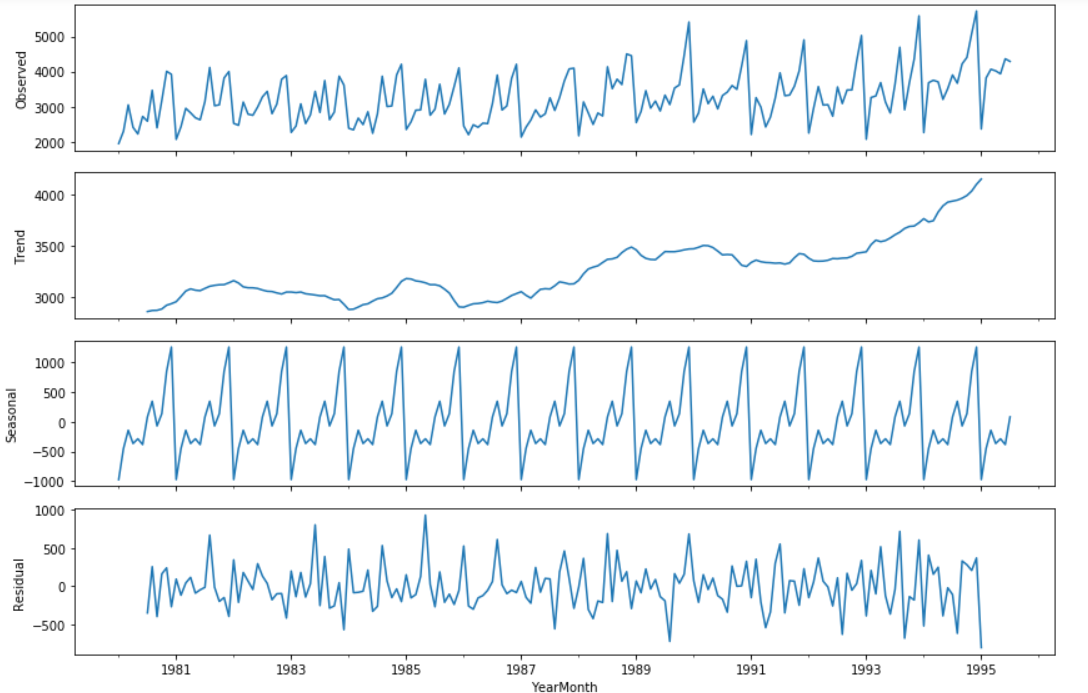
Time series decomposition involves thinking of a series as a combination of level, trend, seasonality, and noise components.

The additive decomposition is the most appropriate if the magnitude of the seasonal fluctuations, or the variation around the trend-cycle, does not vary with the level of the time series. When the variation in the seasonal pattern, or the variation around the trend-cycle, appears to be proportional to the level of the time series, then a multiplicative decomposition is more appropriate. Multiplicative decompositions are common with economic time series.

### **Additive Decomposition**

In Additive decomposition data is seprated into 4 different charts Residual , Seasonal , Trend , obsereved.

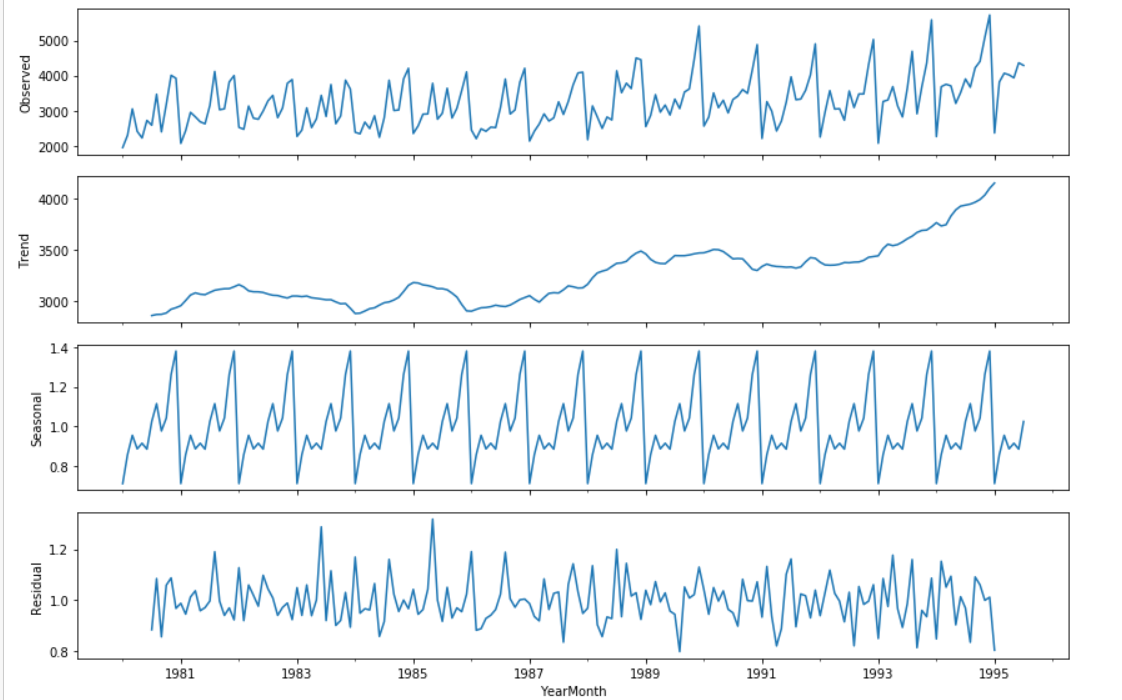
We can see there is an increasing trend with a seasonality.



**Multiplicative Decomposition**

In Additive decomposition data is seprated into 4 different charts Residual , Seasonal , Trend , obsereved.

We can see there is an increasing trend with a seasonality.



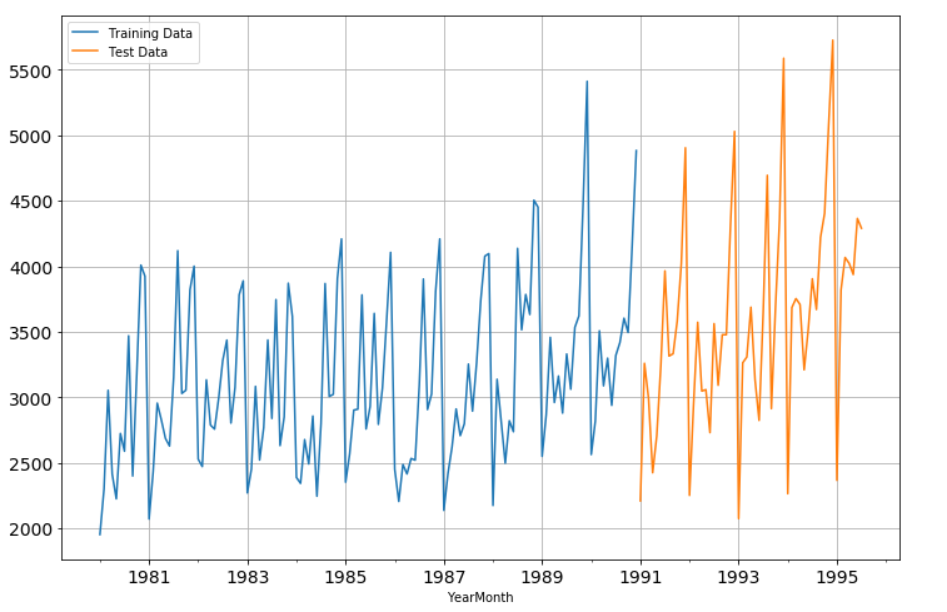
Split the data into t raining and test. The test data should start in 1991.

Data is split into train and test to predict the forecast of sales.

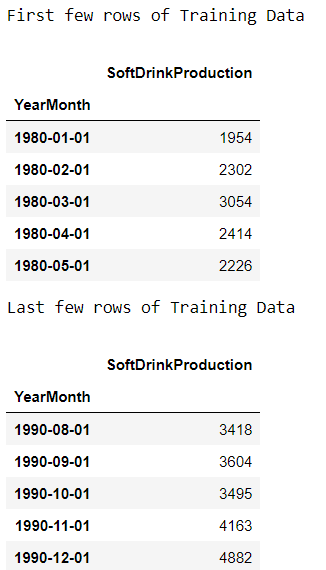
Train data (132 ,1)

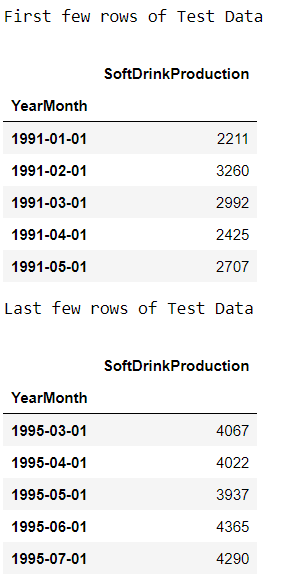
Test data (55, 1)

Plot Diagram of data split into Training and testing



**First 5 rows sample of train & test split data**





Various exponential smoothing models on the t raining data

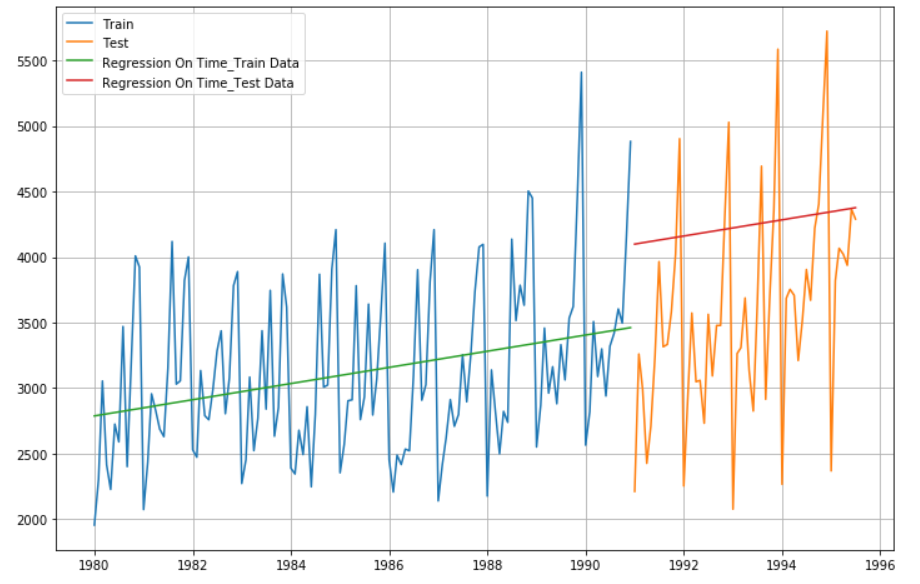
Evaluation of model using RMSE on the test data.

**Model 1: Linear Regression**

Linear Regression  **in which the model finds the best fit linear line between the independent and dependent variable** i.e it finds the linear relationship between the dependent and independent variable.

Here the dependent variable is sales of production of soft drinks over a period of time.

Diagram



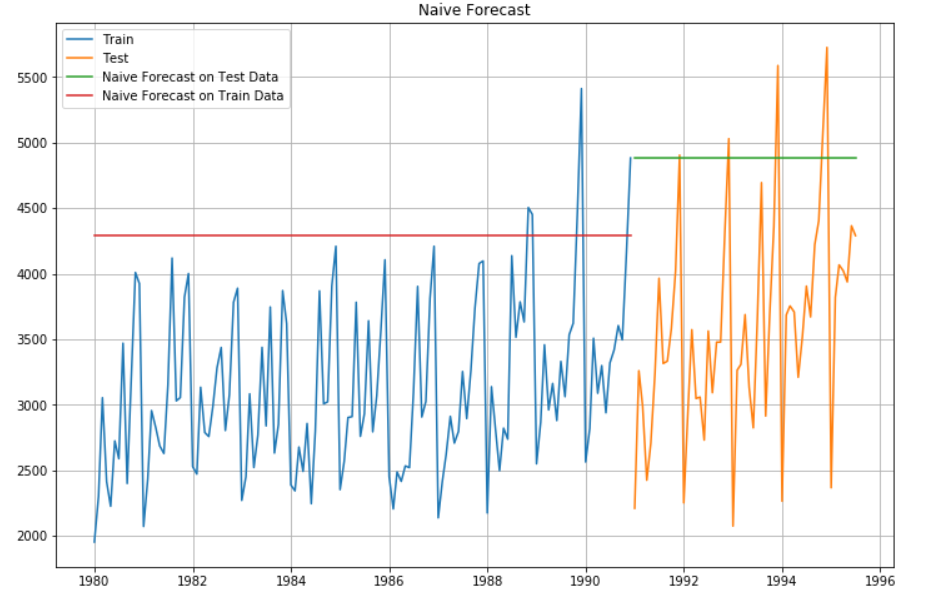
**For RegressionOnTime forecast on the Test Data, RMSE is 1015063.096**

**For RegressionOnTime forecast on the Train Data, RMSE is 372512.417**

**Model 2: Naive Approach: y^t+1=yt**

Naïve forecasting is the technique in which the last period's sales are used for the next period's forecast without predictions or adjusting the factors. Forecasts produced using a naïve approach are equal to the final observed value.

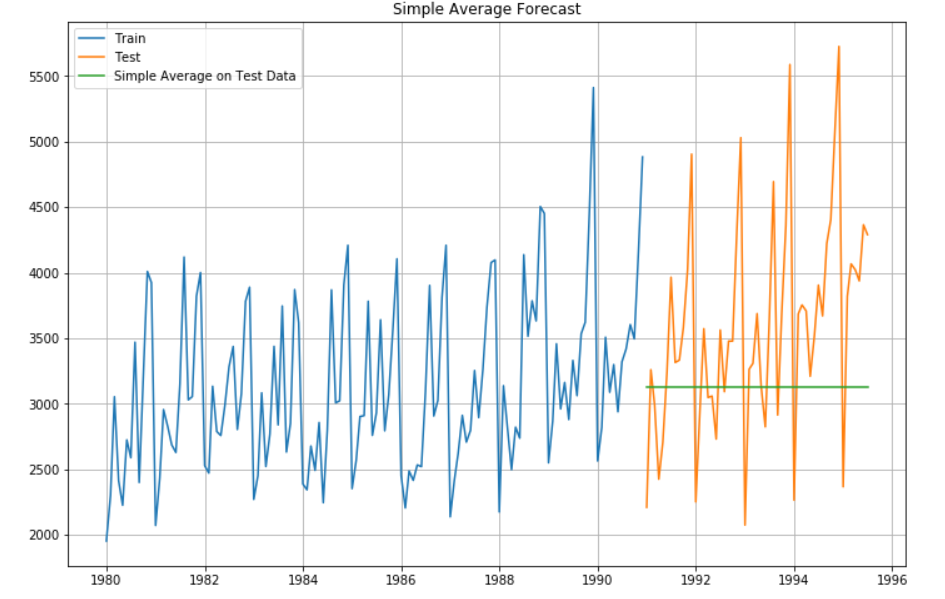
Diagram



For RegressionOnTime forecast on the Test Data, RMSE is 2308148.618

For RegressionOnTime forecast on the Train Data, RMSE is 1770092.091

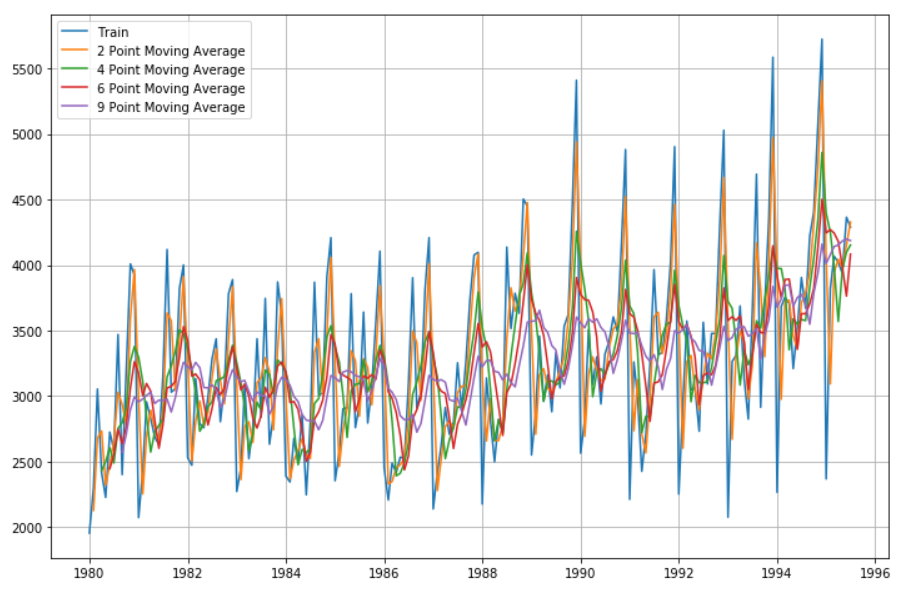
### Method 3: Simple Average

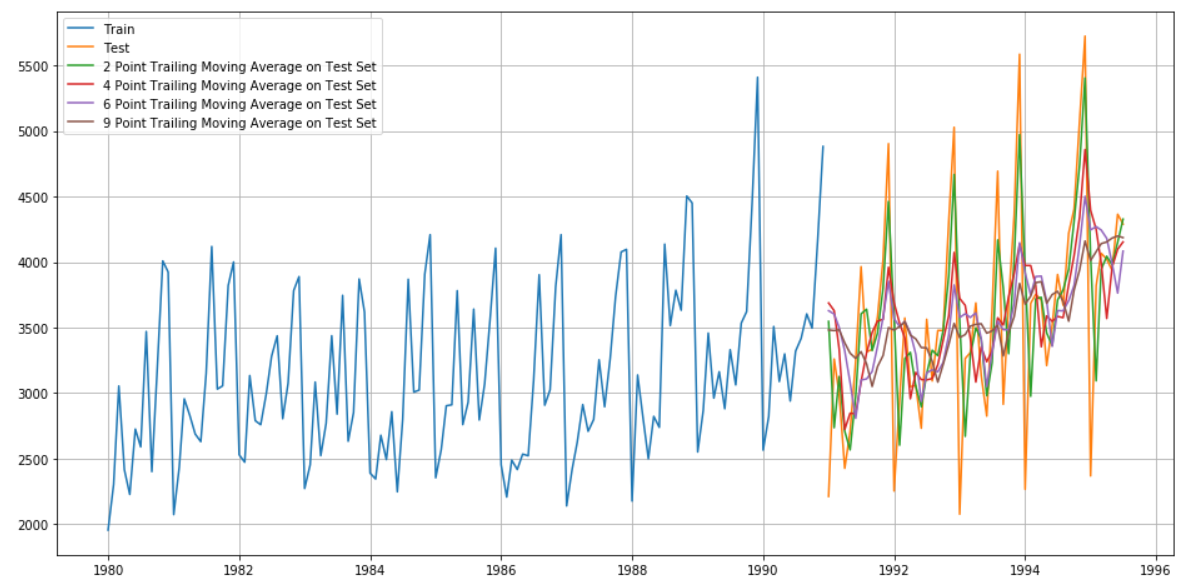


**For Simple Average forecast on the Test Data, RMSE is 873016.197**

### **Method 4: Moving Average(MA)**

A moving average of order mm can be written as^Tt=1mk∑j=−kyt+j,(6.1)(6.1)T^t=1m∑j=−kkyt+j,where m=2k+1m=2k+1. That is, the estimate of the trend-cycle at time tt is obtained by averaging values of the time series within kk periods of tt. Observations that are nearby in time are also likely to be close in value. Therefore, the average eliminates some of the randomness in the data, leaving a smooth trend-cycle component. We call this an mm**-MA**, meaning a moving average of order mm.





**For 2 point Moving Average Model forecast on the Training Data, RMSE is 309943.191**

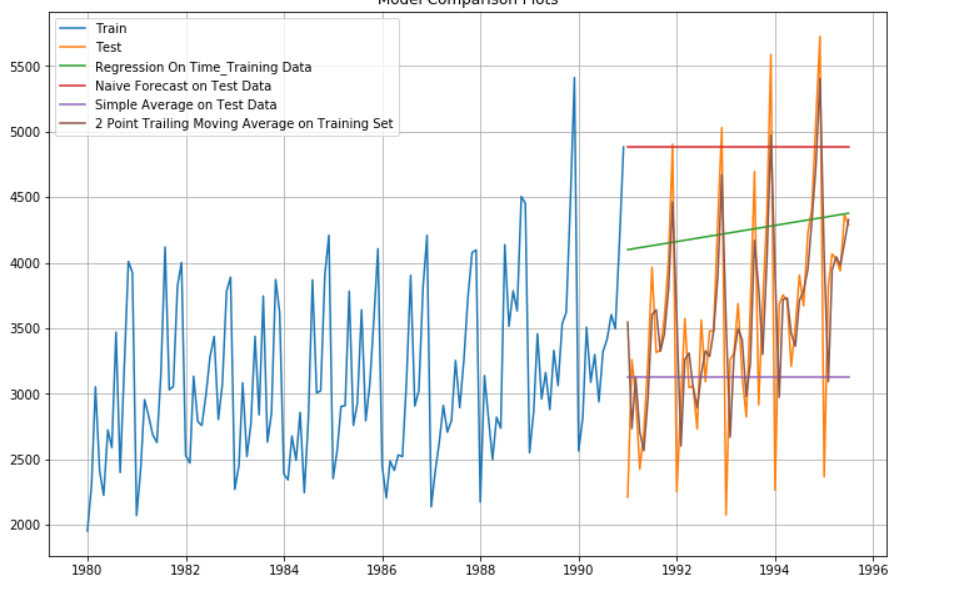
**For 4 point Moving Average Model forecast on the Training Data, RMSE is 472218.725**

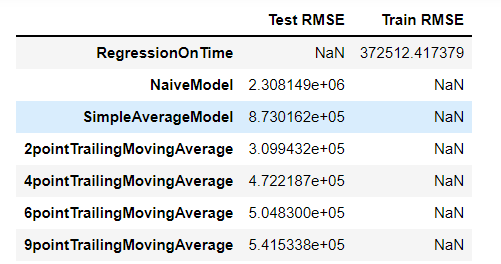
**For 6 point Moving Average Model forecast on the Training Data, RMSE is 504829.969**

**For 9 point Moving Average Model forecast on the Training Data, RMSE is 541533.838**

**Model comparison Plot**

Comparison of RMSE for all the models.





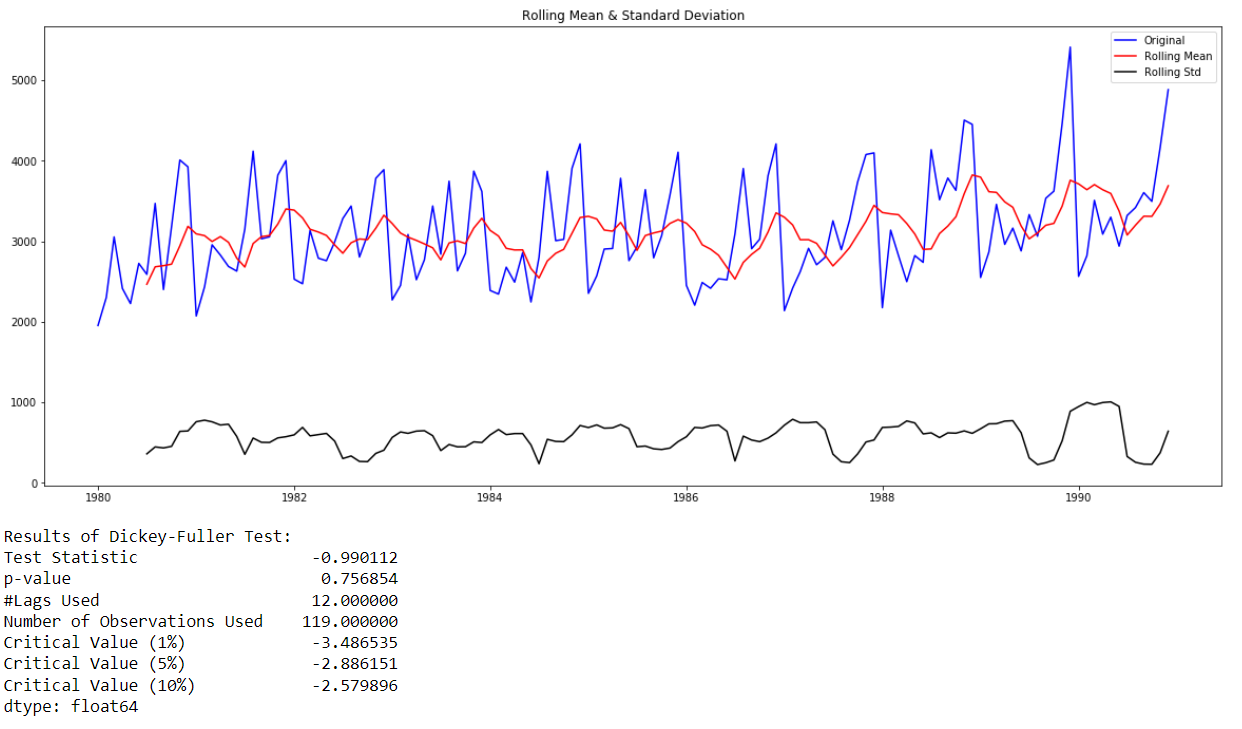
Stationarity of the data

#### The Augmented Dickey-Fuller test is an unit root test which determines whether there is a unit root and subsequently whether the series is non-stationary.

#### The hypothesis in a simple form for the ADF test is:

#### H0 : The Time Series has a unit root and is thus non-stationary.

#### H1 : The Time Series does not have a unit root and is thus stationary.



#### **Since the p-value is more than original form at alpha = 0.05 .**We would want the series to be stationary for building ARIMA models and thus we would want the p-value of this test to be less than the α value.¶

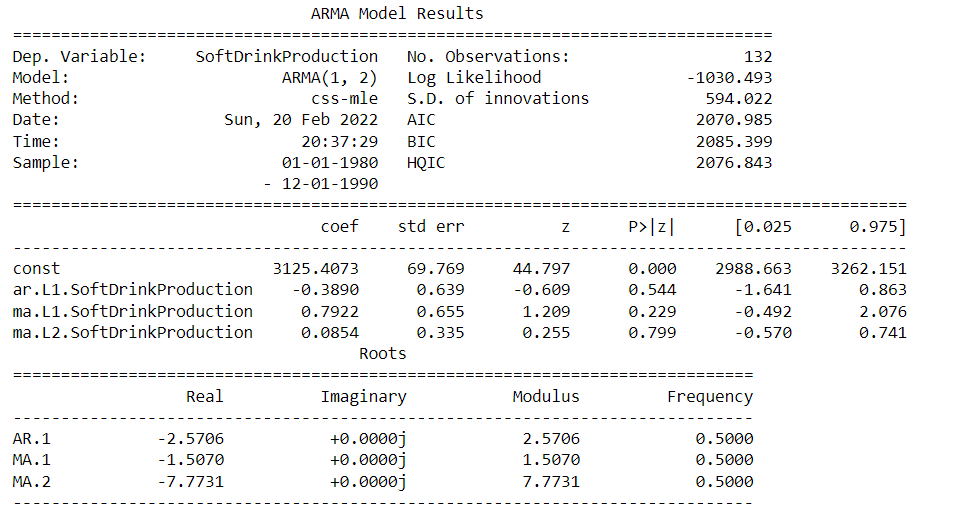
ARIMA/SARIMA model

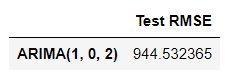
**Autoregressive Integrated Moving Average (ARIMA)** is a popular time series forecasting model. It is used in forecasting time series variable such as price, sales, production, demand etc.

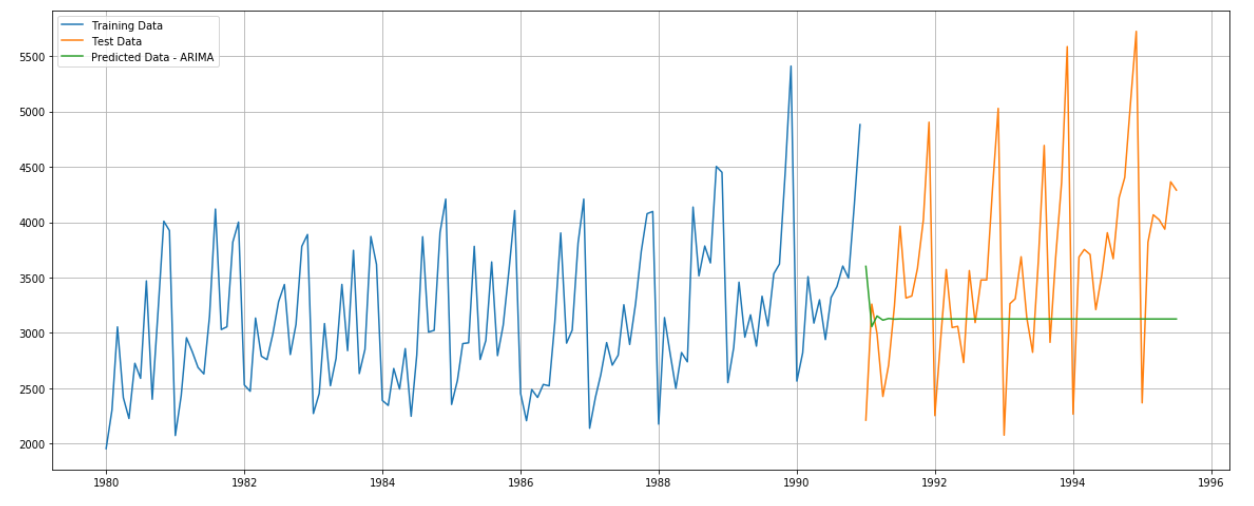
A **seasonal autoregressive integrated moving average** (SARIMA) model is one step different from an ARIMA model based on the concept of seasonal trends. In many time series data, frequent seasonal effects come into play.

ARIMA

**Parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data**

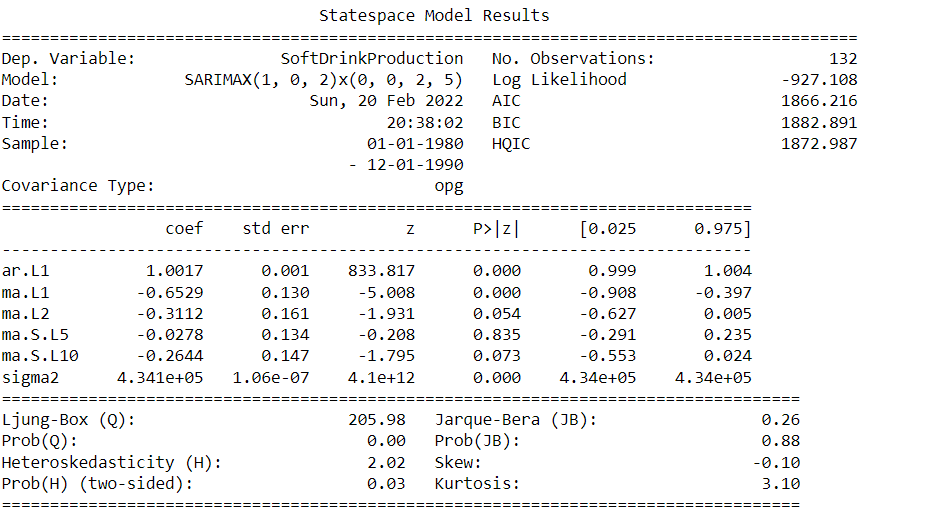


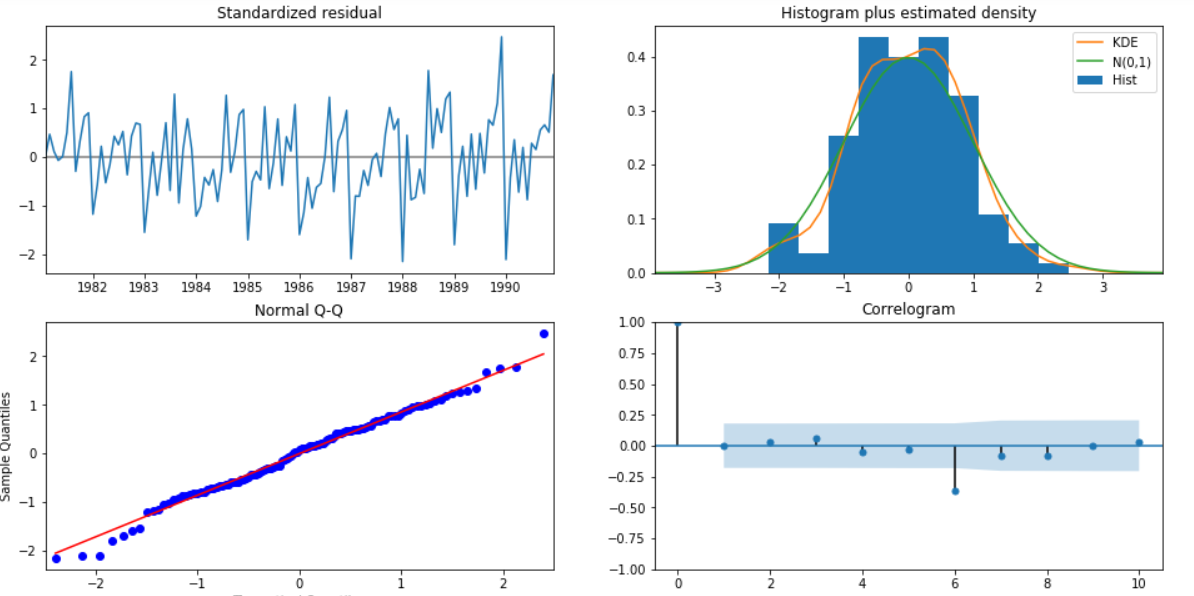




SARIMA

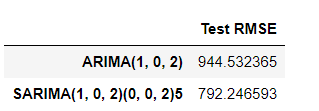
**Parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data**



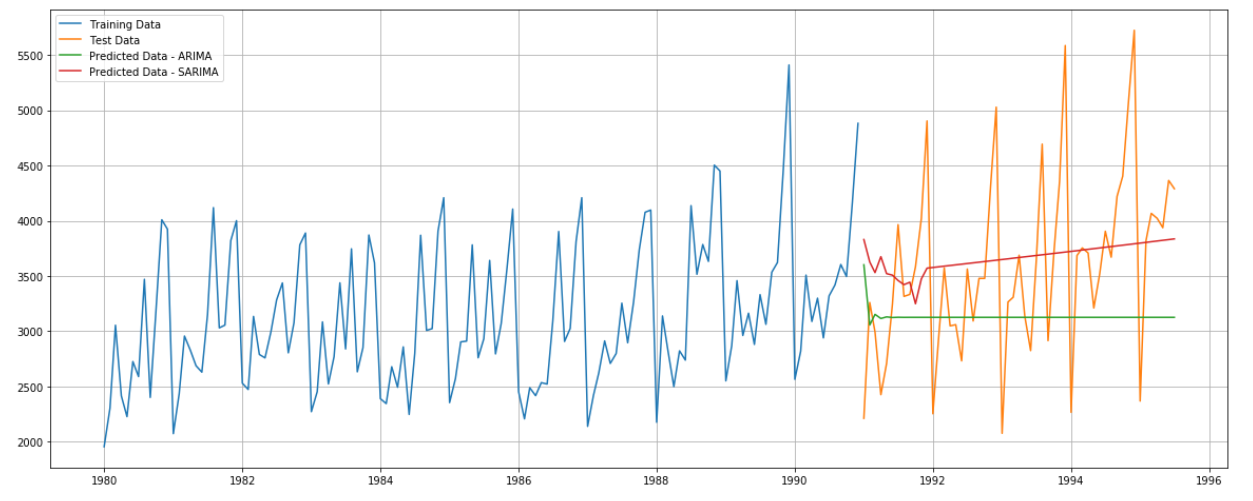


RMSE - 792.246593472974

RMSE Comparison of ARIMA /SARIMA

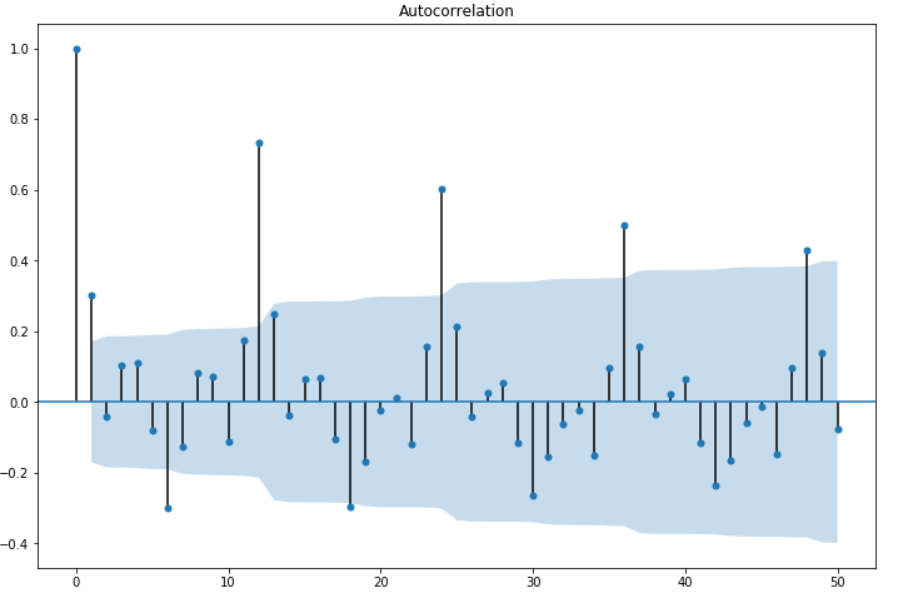


#### **RMSE OF SARIMA has reduced in comparison to ARIMA when seasonality was introduced**



ACF and PACF on train data

ACF - **ACF** is an (complete) auto-correlation function which gives us values of auto-correlation of any series with its lagged values. We plot these values along with the confidence band and tada! We have an ACF plot. In simple terms, it describes how well the present value of the series is related with its past values. A time series can have components like trend, seasonality, cyclic and residual. ACF considers all these components while finding correlations hence it’s a ‘complete auto-correlation plot’.



PACF – **PACF** is a partial auto-correlation function. Basically instead of finding correlations of present with lags like ACF, it finds correlation of the residuals (which remains after removing the effects which are already explained by the earlier lag(s)) with the next lag value hence ‘partial’ and not ‘complete’ as we remove already found variations before we find the next correlation. So if there is any hidden information in the residual which can be modeled by the next lag, we might get a good correlation and we will keep that next lag as a feature while modeling. Remember while modeling we don’t want to keep too many features which are correlated as that can create multicollinearity issues. Hence we need to retain only the relevant features.

