**Project: Time Series Forecasting**

Pavan Kalyan B | PGP DSBA Oct\_C’21 | Jun 2022

**Table of contents**

1. Read the data as an appropriate Time Series data and plot the data.
2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.
3. Split the data into training and test. The test data should start in 1991.
4. Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other additional models such as regression, naïve forecast models, simple average models, moving average models should also be built on the training data and check the performance on the test data using RMSE.
5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. 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.
6. 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.
7. Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.
8. Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.
9. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.
10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

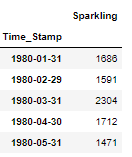
Problem:

For this particular assignment, the data of different types of wine sales in the 20th century is to be analysed. Both of these data are from the same company but of different wines. As an analyst in the ABC Estate Wines, you are tasked to analyse and forecast Wine Sales in the 20th century.

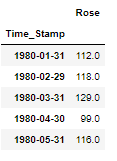
Data set for the Problem: [Sparkling.csv](https://olympus.mygreatlearning.com/courses/63577/files/5489615/download?verifier=LTuiihF1JHHna9N5Bszohhe3KJtr4wv88sbFPi2a&wrap=1) and [Rose.csv](https://olympus.mygreatlearning.com/courses/63577/files/5489614/download?verifier=rWyX1V5P6nPUtXUJHQTBquS3wadrC2qbHwZRDhds&wrap=1)

**1. Read the data as an appropriate Time Series data and plot the data.**

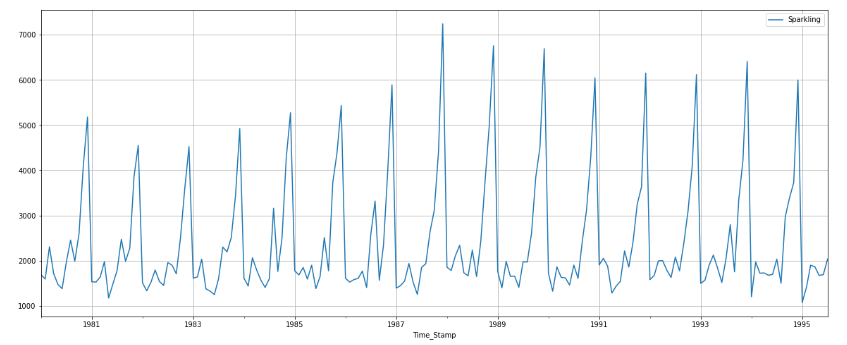
we are given with two types of data one is sparkling and other one is rose dataset. Each dataset contains different types of wine sales each month. We will be exploring the data



**Sparkling dataset sample**

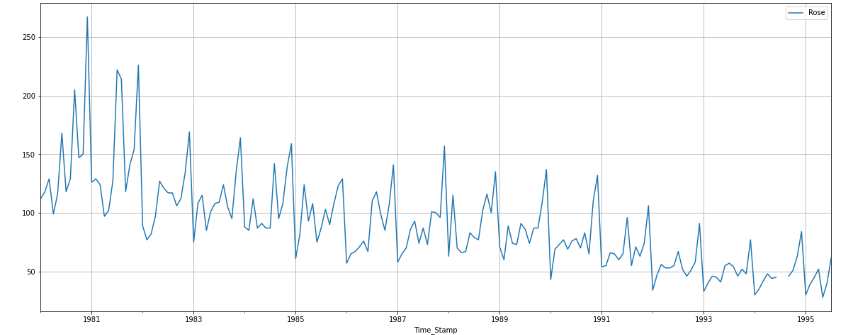
****

**Rose dataset sample**

****

**Figure 1.1 sparkling data graph**

We can observe that the graph is plotted as per month of the sales of sparkling company wines

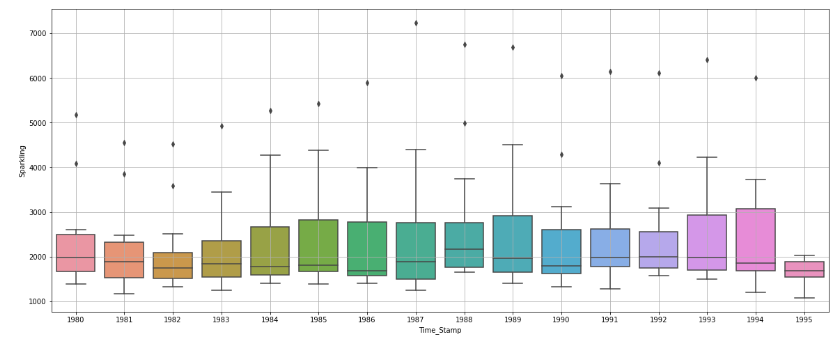


**Figure 1.2 Rose data graph**

We can observe that the graph is plotted as per month of the sales of rose company wines

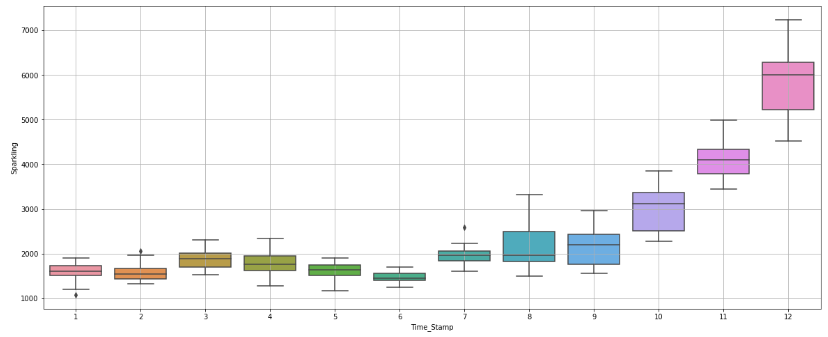
* Data contains of 187 rows and 1 column in both data set
* The data contains seasonality and trend as well

**2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.**

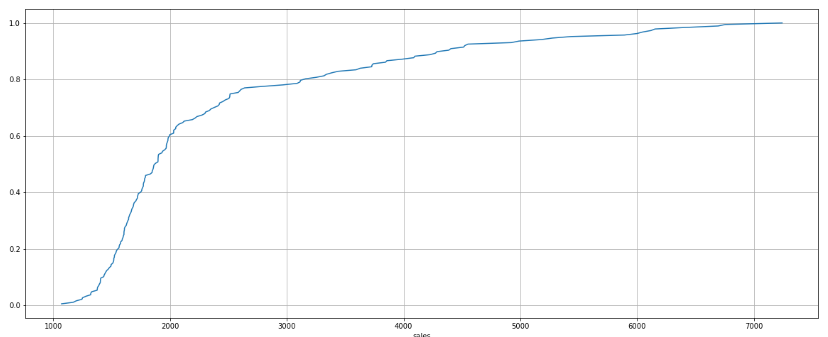


**Figure 2.1 Sparkling yearly plot**

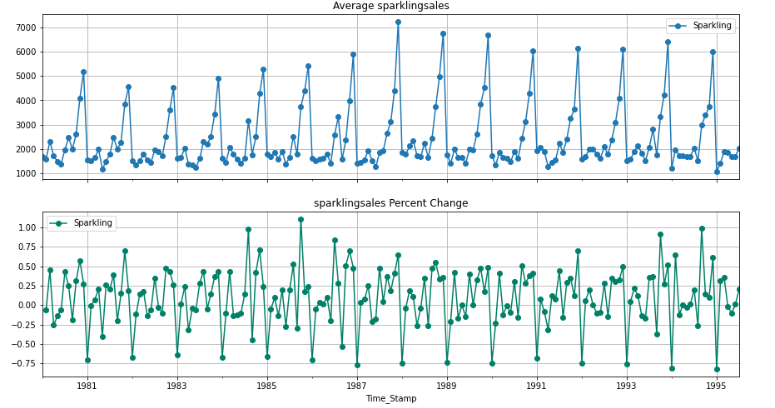
* We can observe that there are some outliers in the yearly plot
* Year 1985, 1985, 1993, 1994 the sales were high in sparkling
* In the year 1995 there were no outliers and it was a minimum sales

**  
Figure 2.2 Monthly plot of sparkling data**

* For each end of the year the sales were high i.e., December month has high sales
* We can observe that sales were increasing each month
* On the 6th month the sales are too low

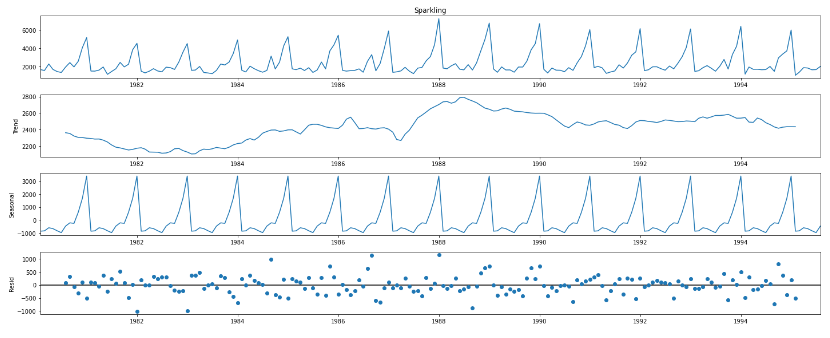
****

**Figure 2.3 Empirical Cumulative Distribution of sparkling**

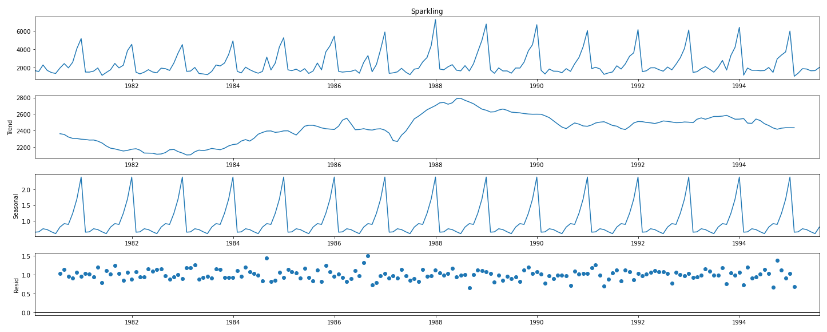
****

**Figure 2.4 Average sparkling sales per month**

* In between the year 1985 to 1987 the sales were more than 100%
* Below 1985 the sales were too low

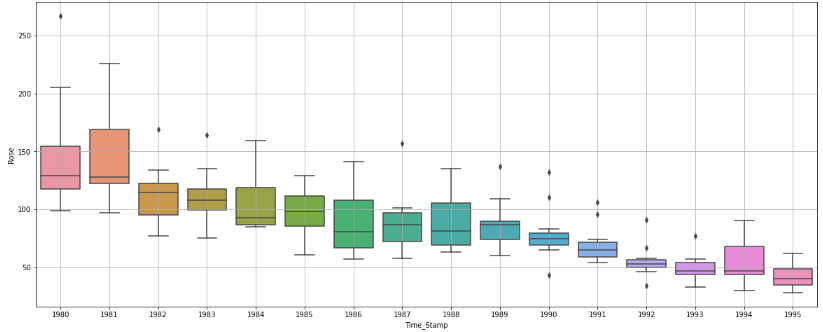
****

**Figure 2.5 Decomposition additive of sparkling**

****

**Figure 2.6 Decomposition multiplicative of sparkling**

* From the above graphs of decomposition, we can observe that there is a seasonality in the data

****

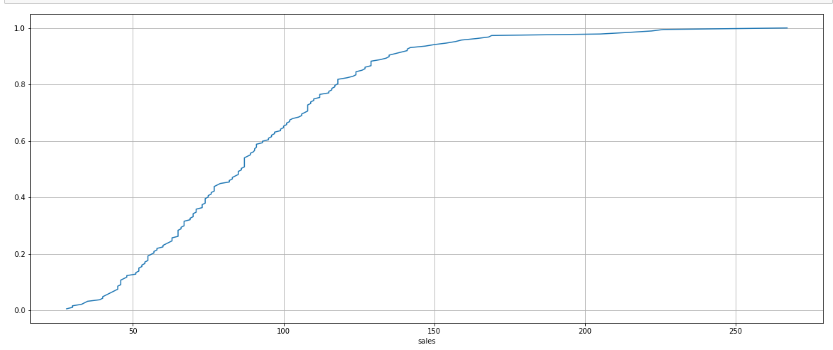
**Figure 2.7 Yearly plot of rose data**

* We can observe that there are some outliers in the data
* There was a high sale in year 1991 and 1990
* Yearly plot contains outliers in the years 1980, 1982, 1983, 1987, 1989, 1990, 1991, 1992, 1993

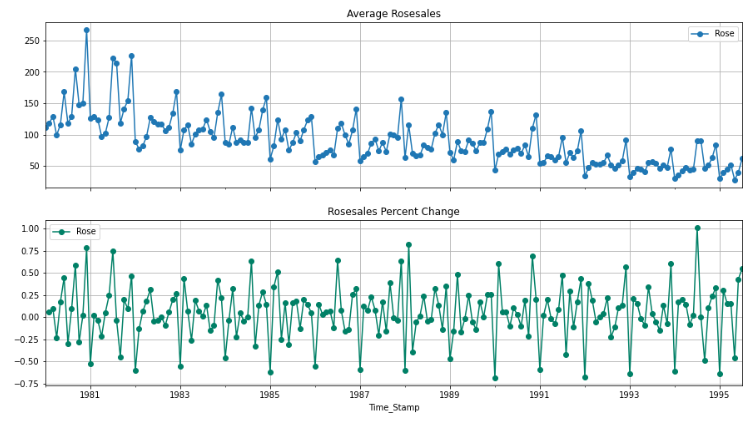
****

**Figure 2.8 Monthly plot of rose data**

* The sales were increasing each month but has an outlier
* The sales are high after 7th month which seems to be a seasonal
* From 7th to 12th there is a rapid sale but not as expected

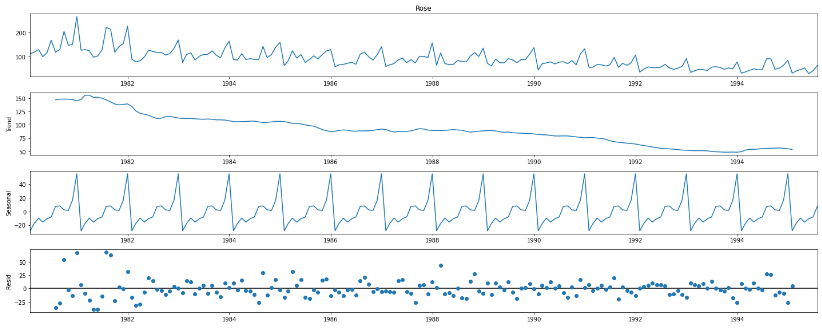
****

**Figure 2.9 Empirical Cumulative Distribution of rose**

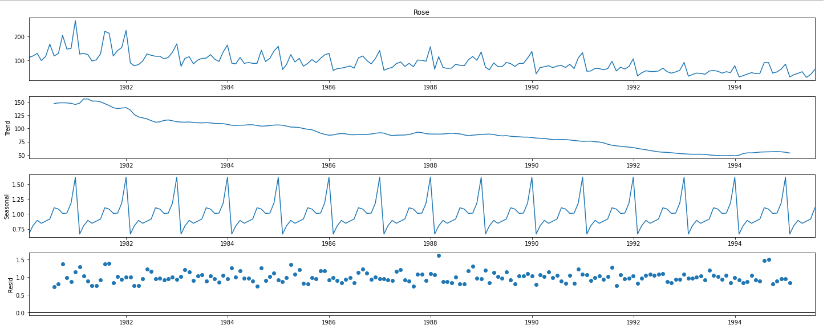
****

**Figure 2.10 Average rose sales per month**

* The sales were decreasing each year
* After the year 1981, the sale started decreasing

****

**Figure 2.11 Decomposition additive of rose data**

****

**Figure 2.12 Decomposition Multiplicative of rose**

* This data has seasonality in the data
* The trend seems to be getting low

**3. Split the data into training and test. The test data should start in 1991.**

****

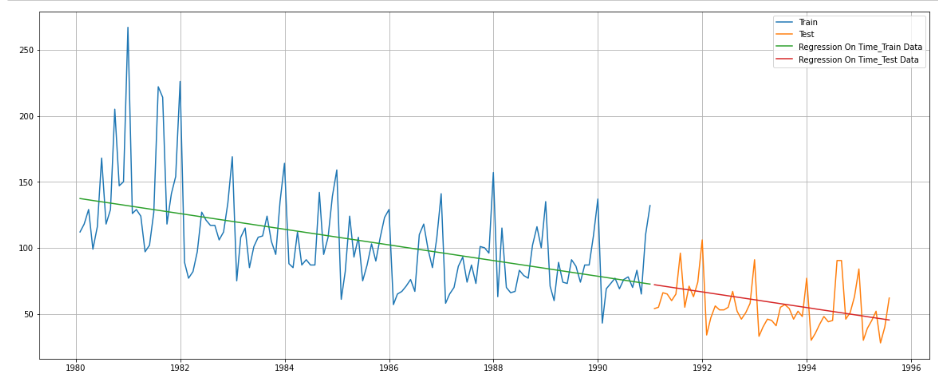
* As the split it using 1991, we get 132 rows and 1 column in the training part and 55 rows and 1 column in the testing part
* We are using the past data to get the forecast

**4. Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression,naïve forecast models and simple average models. should also be built on the training data and check the performance on the test data using RMSE.**

****

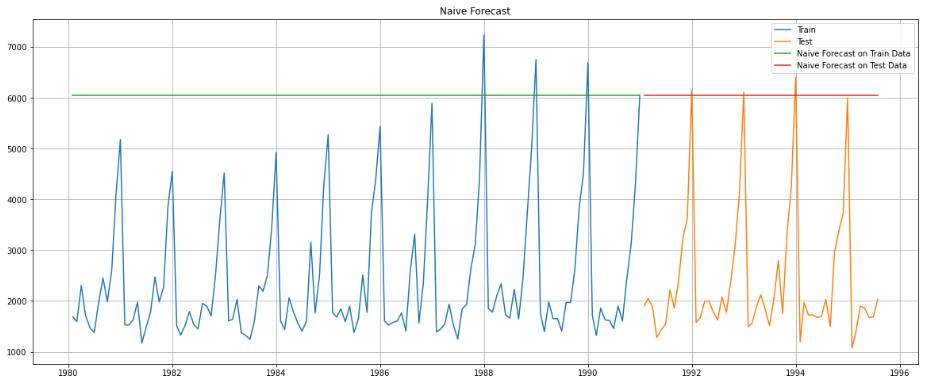
**Figure 4.1 Linear regression on sparkling data**

****

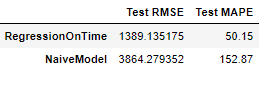
****

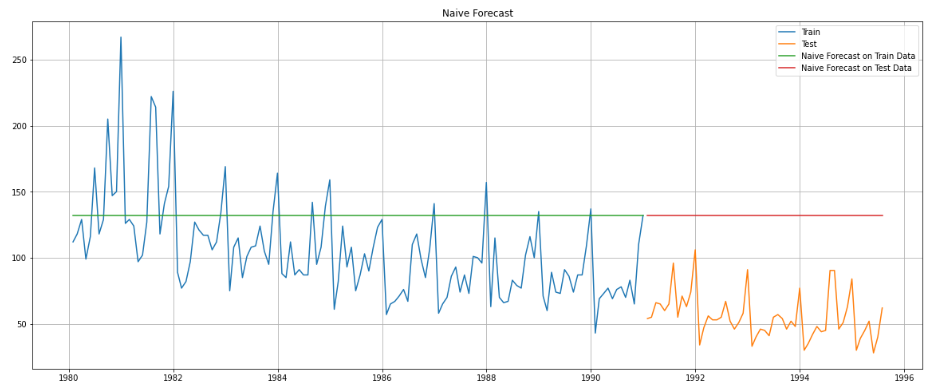
**Figure 4.2 Linear regression on rose data**

****

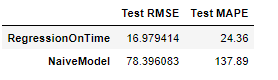
****

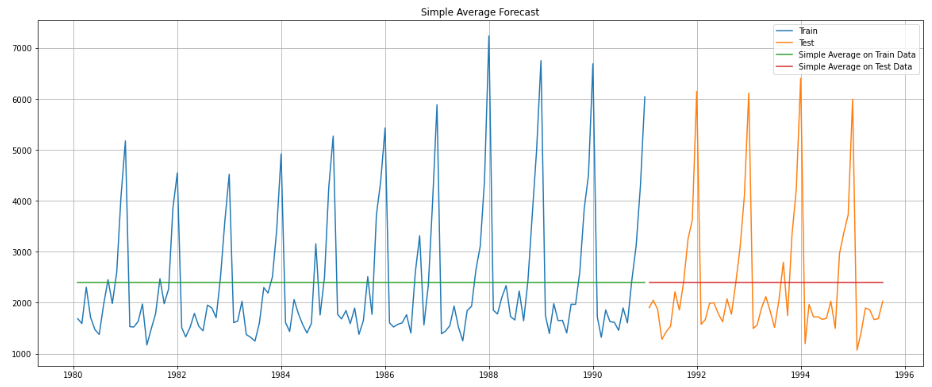
**Figure 4.3 Naïve bayes on sparkling data**

****

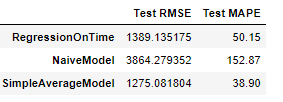
****

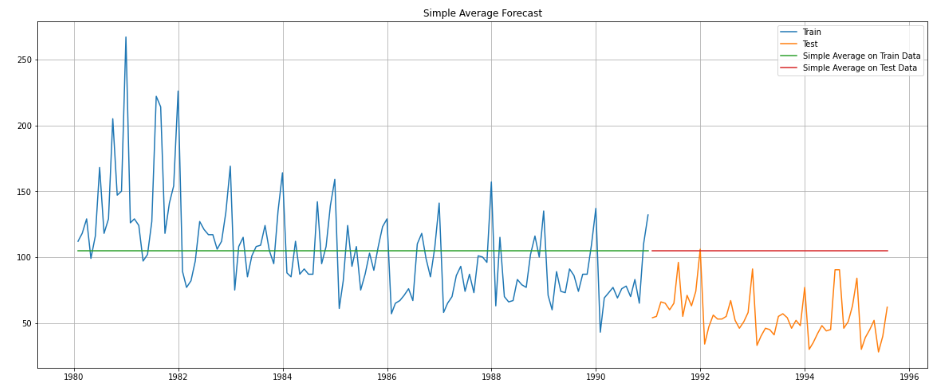
**Figure 4.4 Naïve bayes on rose data**

****

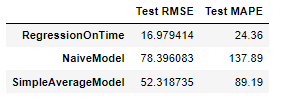
****

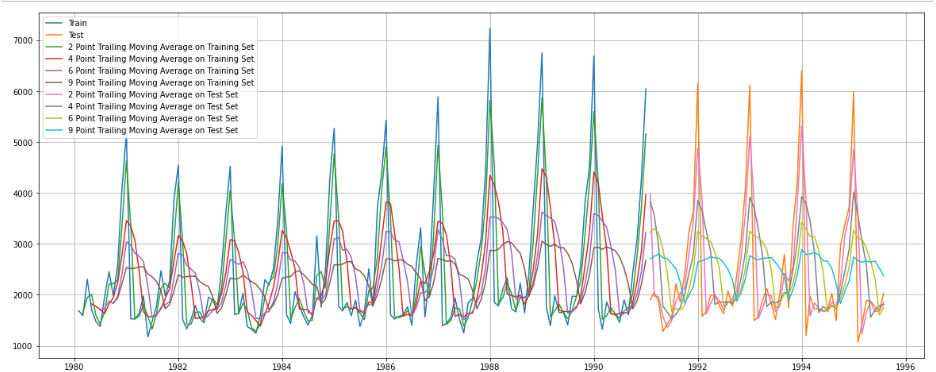
**Figure 4.5 simple average on sparkling data**

****

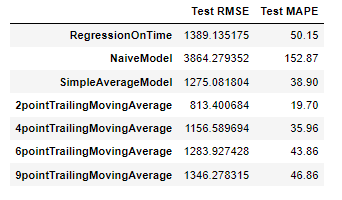
****

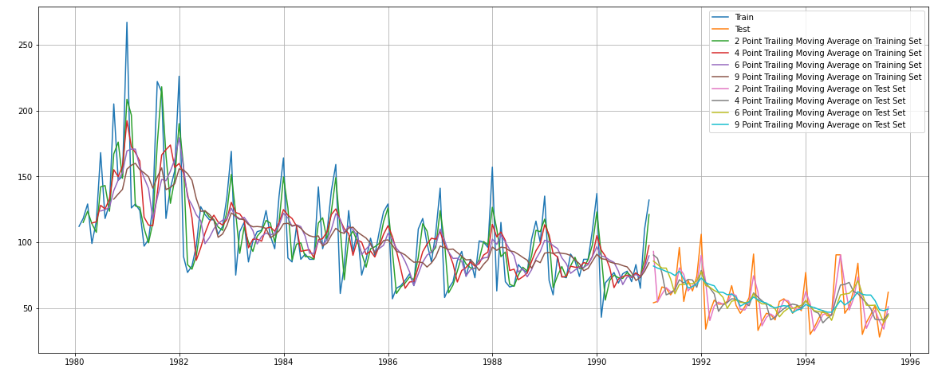
**Figure 4.6 simple average on rose data**

****

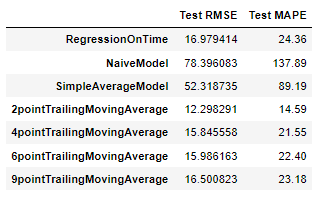
****

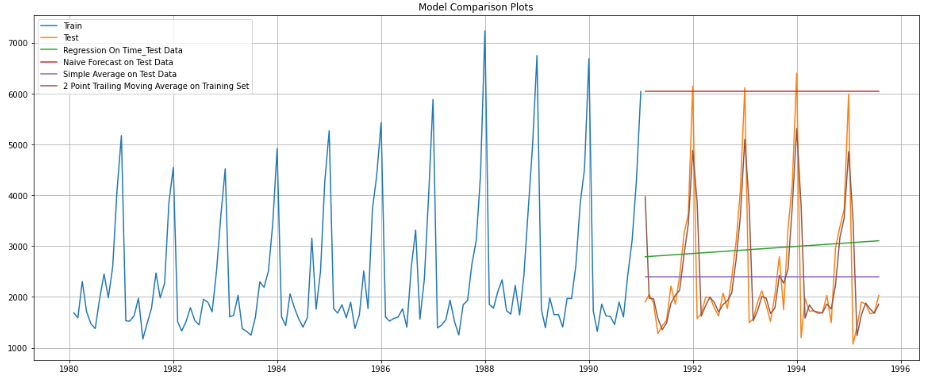
**Figure 4.7 Moving average on training and testing for sparkling data**

****

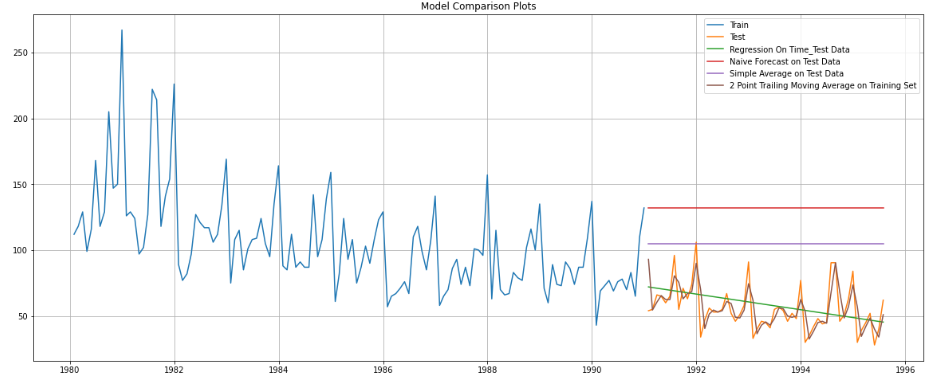
****

**Figure 4.8 moving average on training and testing for rose data**

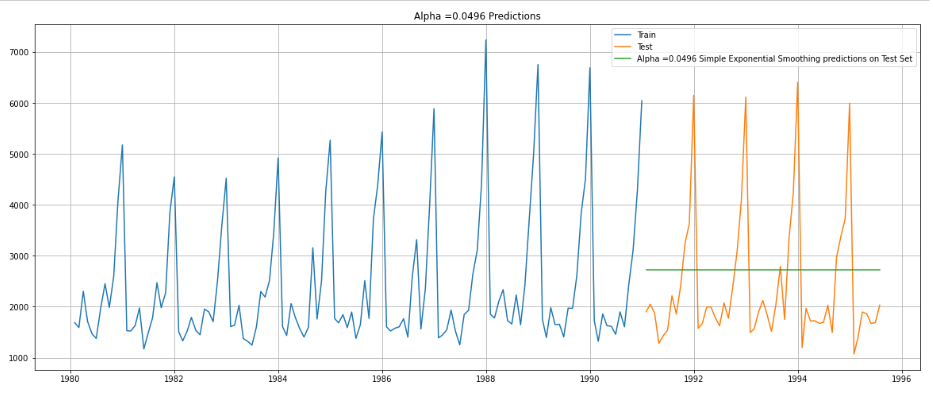
****

****

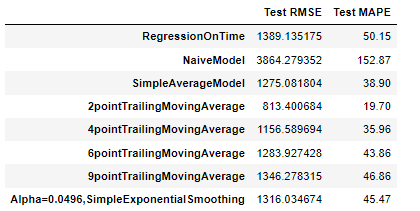
**Figure 4.9 plotting all models for sparkling data**

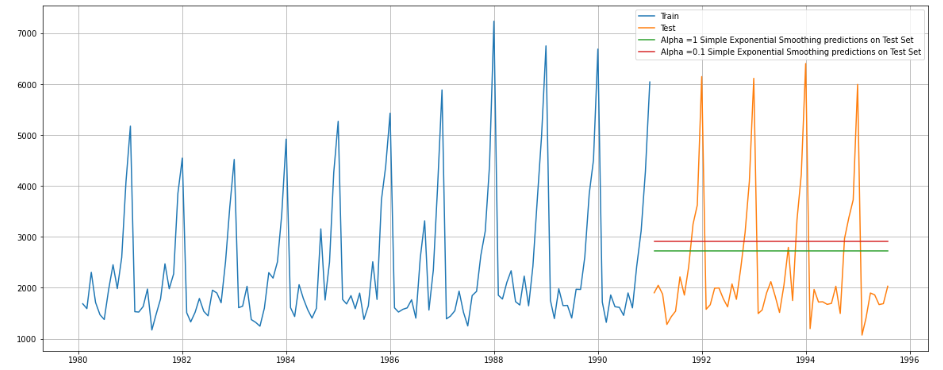
****

**Figure 4.10 plotting all models for rose data**

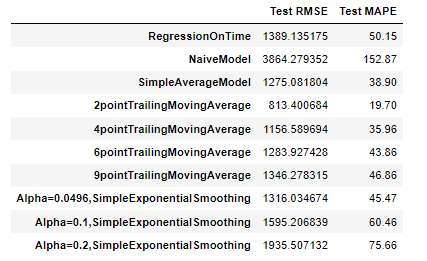
****

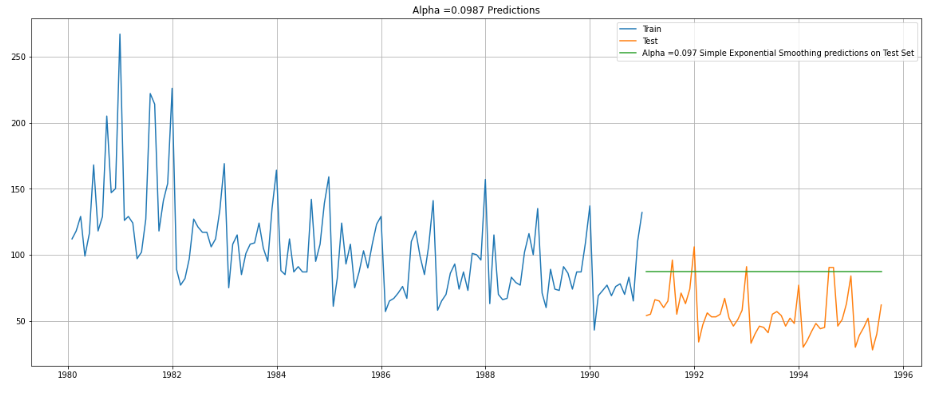
**Figure 4.11 simple exponential model for sparkling data**

****

****

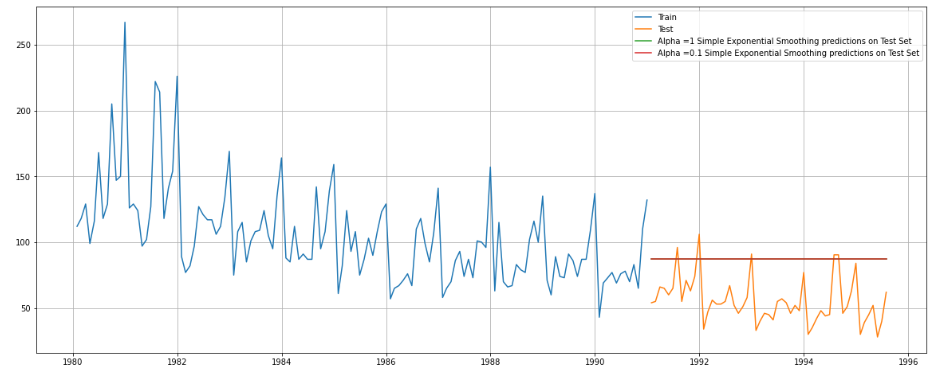
**Figure 4.12 simple exponential model iterative for sparkling data**

****

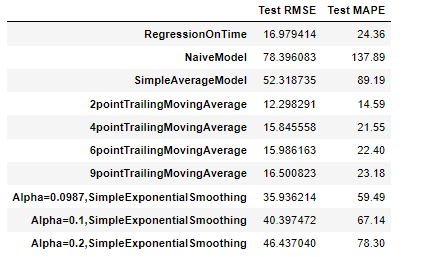
****

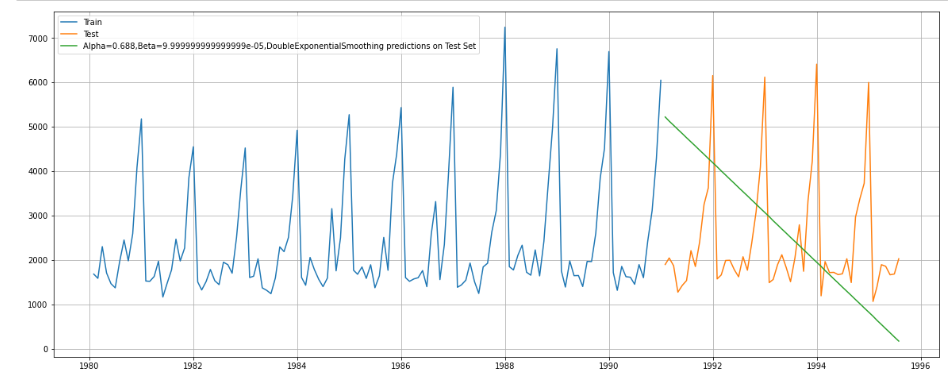
**Figure 4.13 simple exponential model for rose data**

****

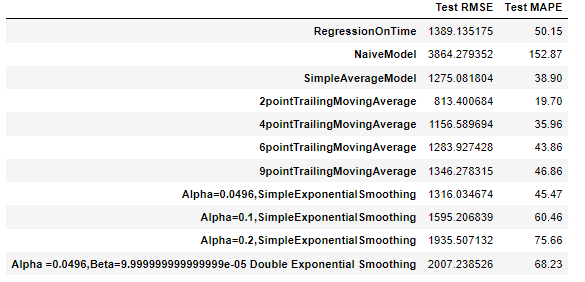
****

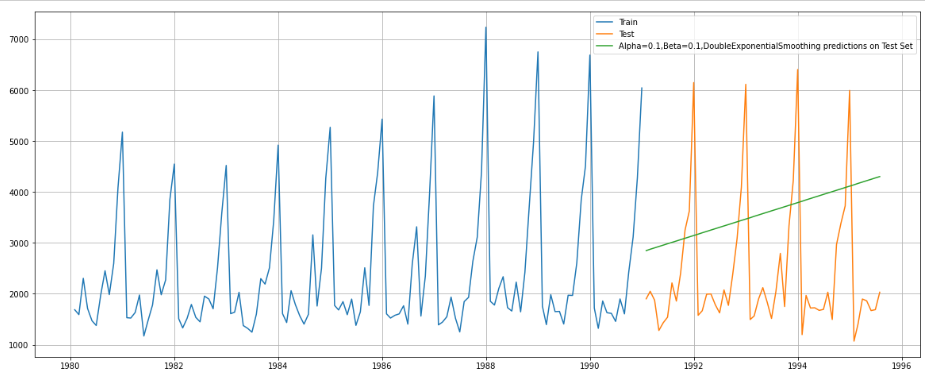
**Figure 4.14 simple exponential model iterative for rose data**

****

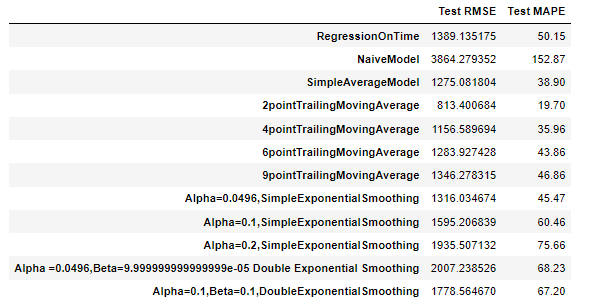
****

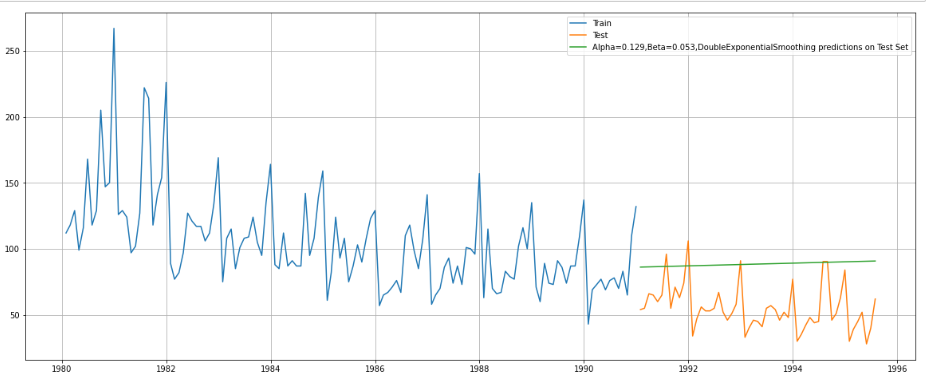
**Figure 4.15 double exponential model on sparkling data**

****

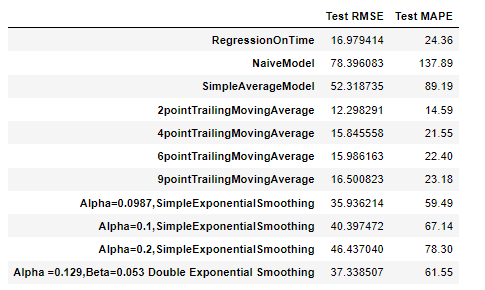
****

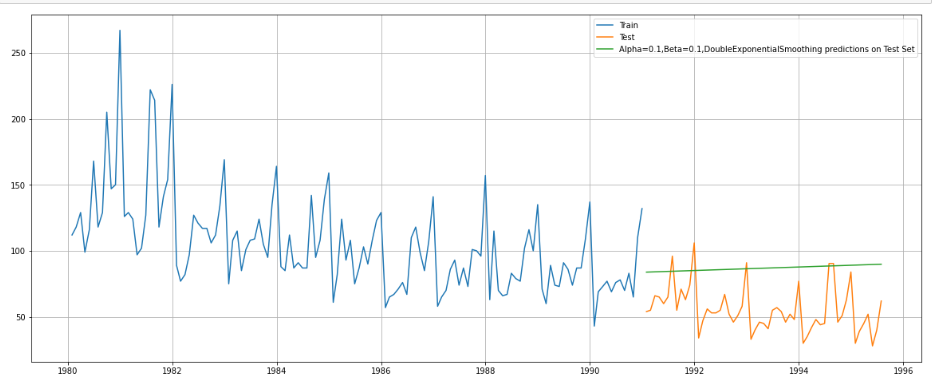
**Figure 4.16 double exponential iterative model on sparkling data**

****

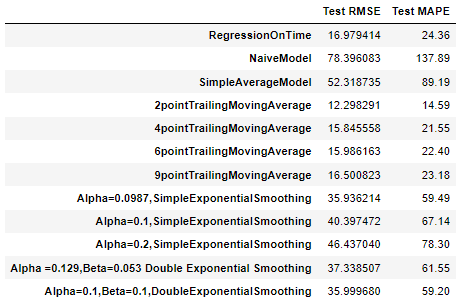
****

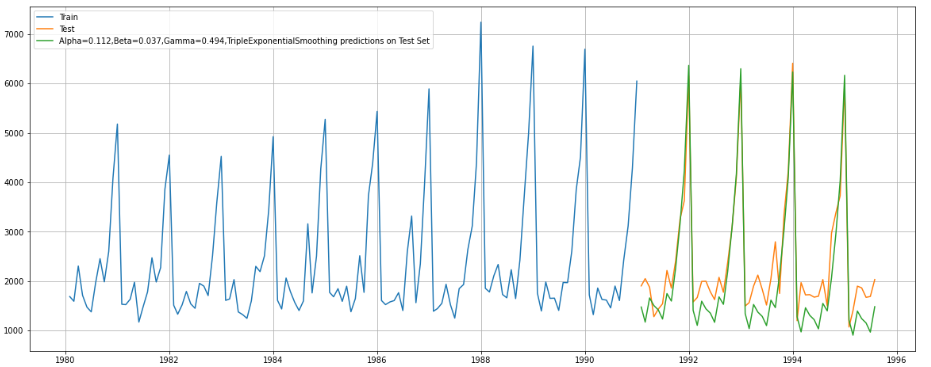
**Figure 4.17 double exponential model on rose data**

****

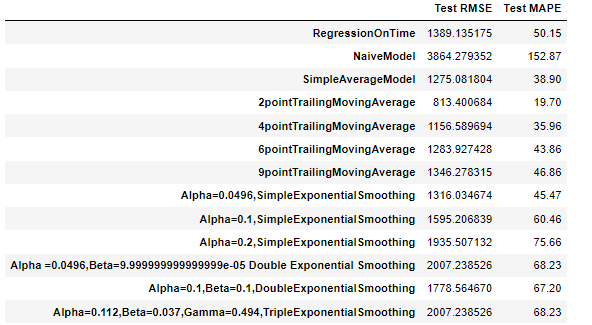
****

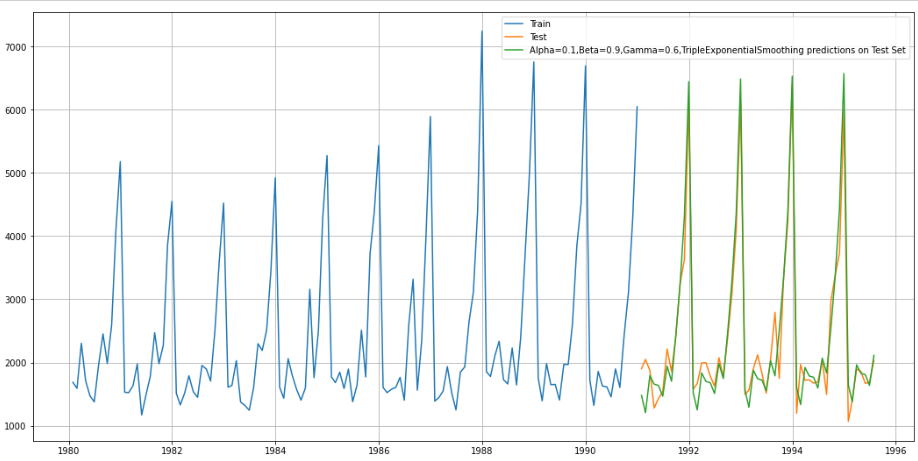
**Figure 4.18 double exponential iterative model on rose data**

****

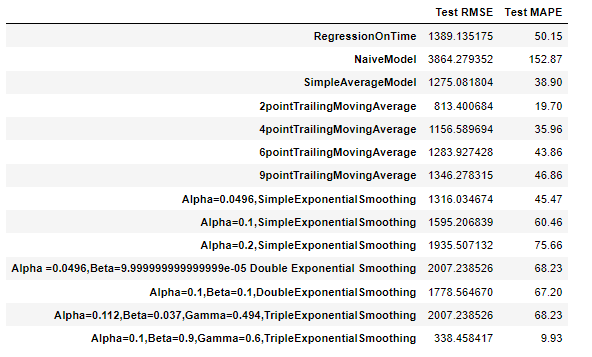
****

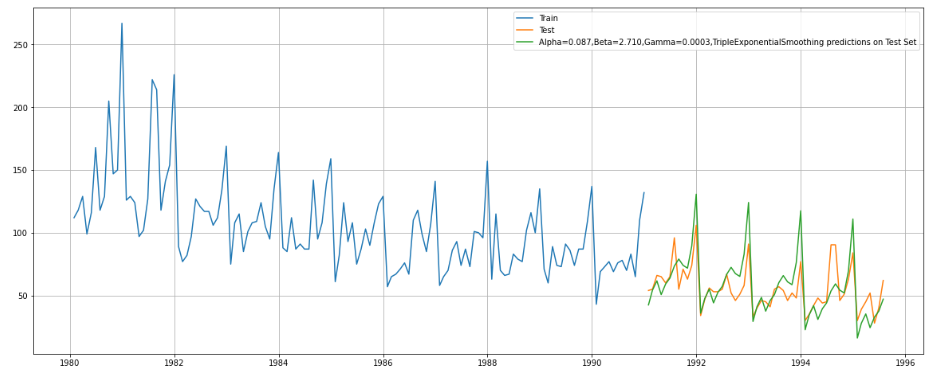
**Figure 4.19 triple exponential model on sparkling**

****

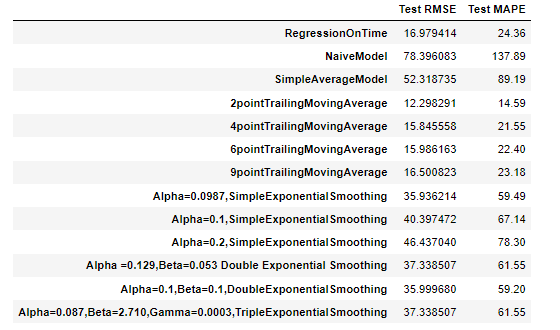
****

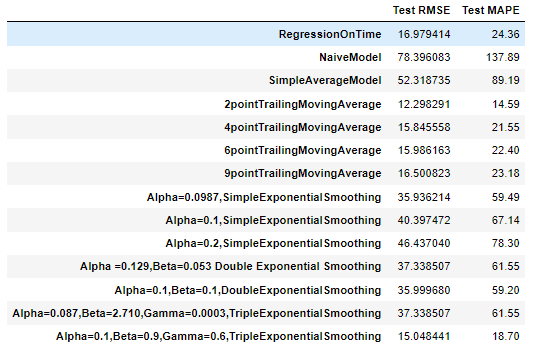
**Figure 4.20 triple exponential iterative model on sparkling data**

****

****

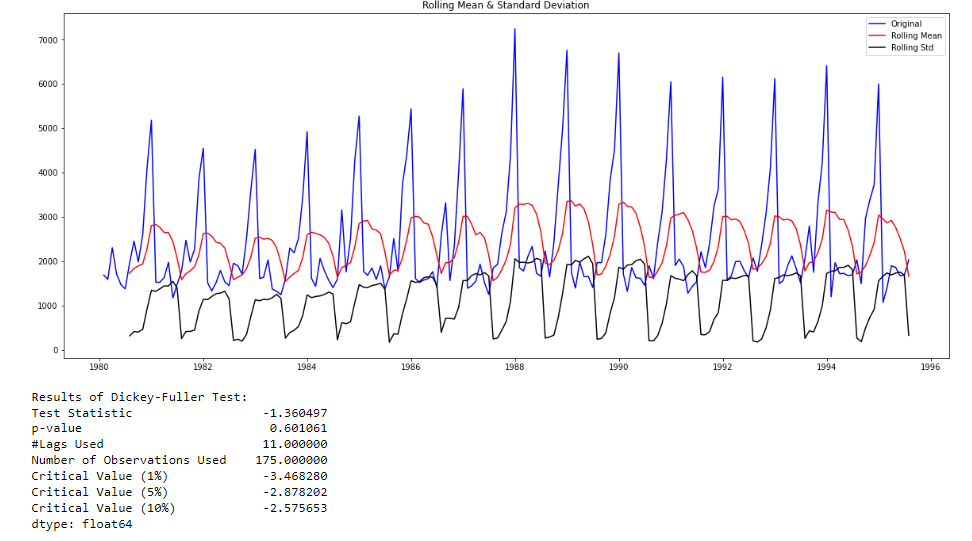
**Figure 4.21 triple exponential model on rose data**

****

****

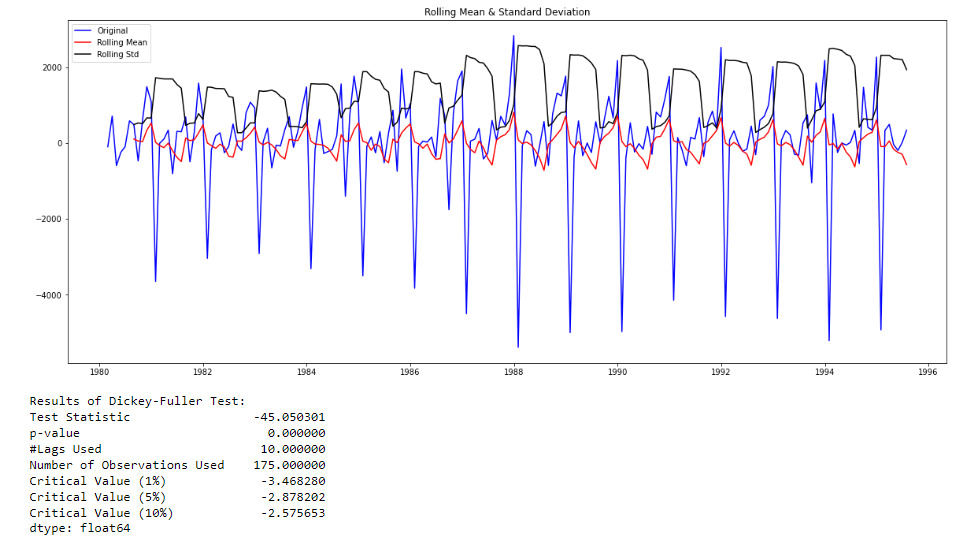
**Triple exponential iterative model on rose data table**

**5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. 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.**

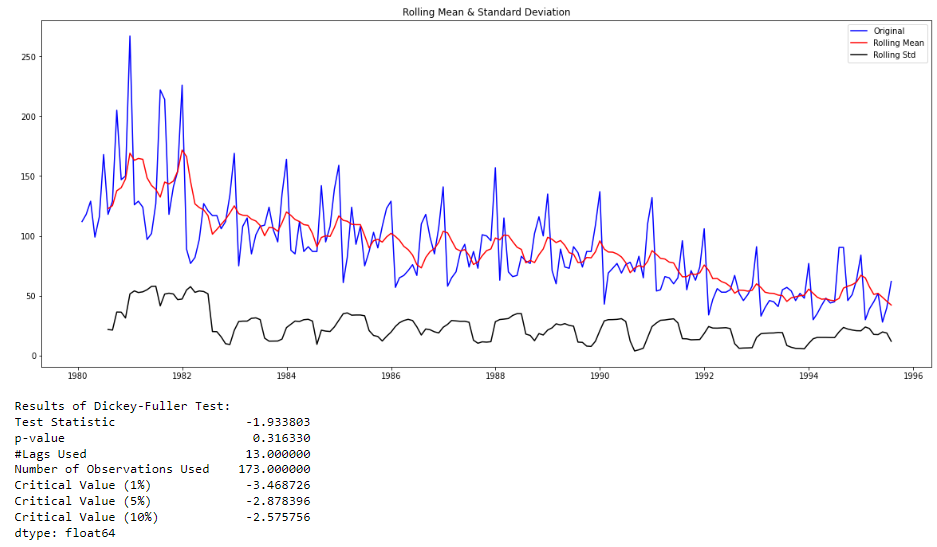
****

**Figure 5.1 non stationary on sparkling data**

As the pvalue is greater than 5% we fail to reject null hypothesis, the data exists as non-stationary, we have to make the data stationary using diff().

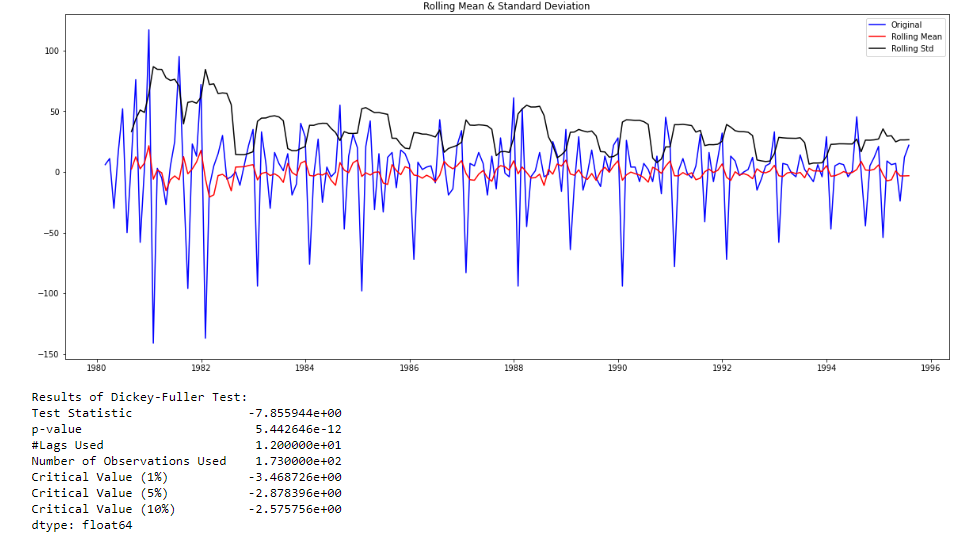


**Figure 5.2 stationary data on sparkling data**

****

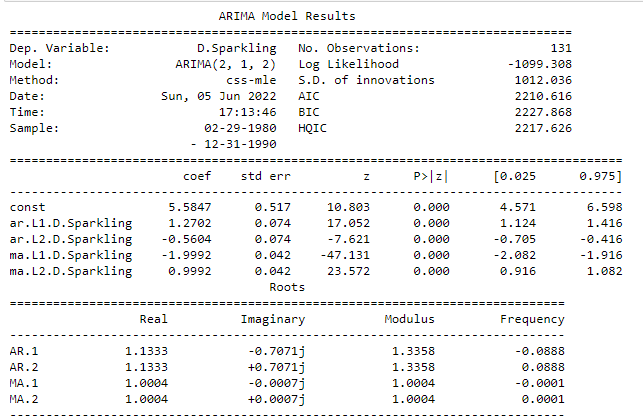
**Figure 5.3 non stationary on rose data**

As the pvalue is greater than 5% we fail to reject null hypothesis, the data exists as non-stationary, we have to make the data stationary using diff().

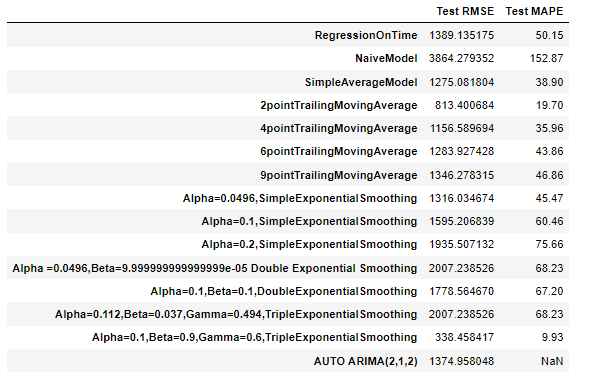


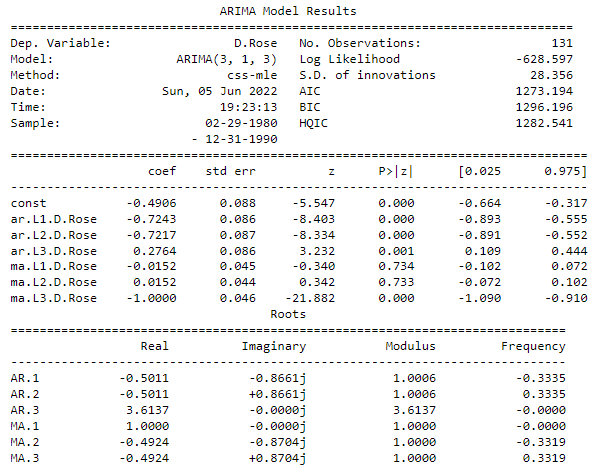
**Figure 5.4 stationary on rose data**

**6. 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.**

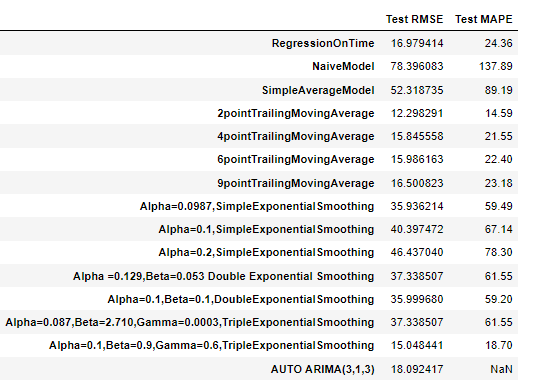
****

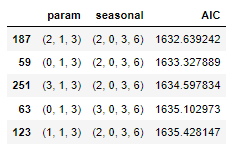
**Figure 6.1 Automated arima on sparkling data**

****

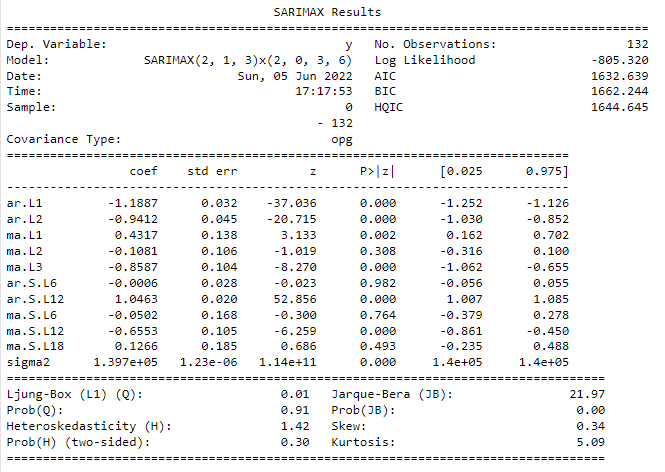
****

**Figure 6.2 Automated arima on rose data**

****

****

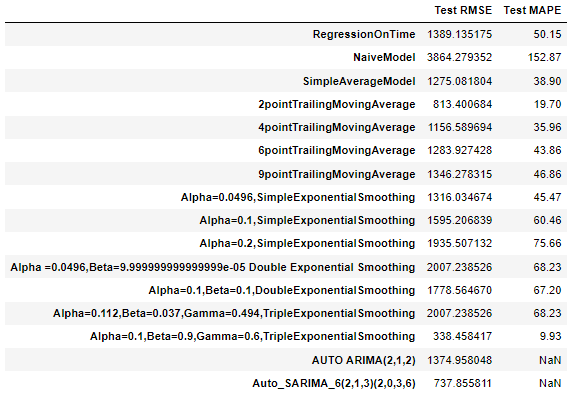
* The above table has the parameters for the normal method on sparkling data

****

**Figure 6.3 Automated sarima 6 on sparkling data**

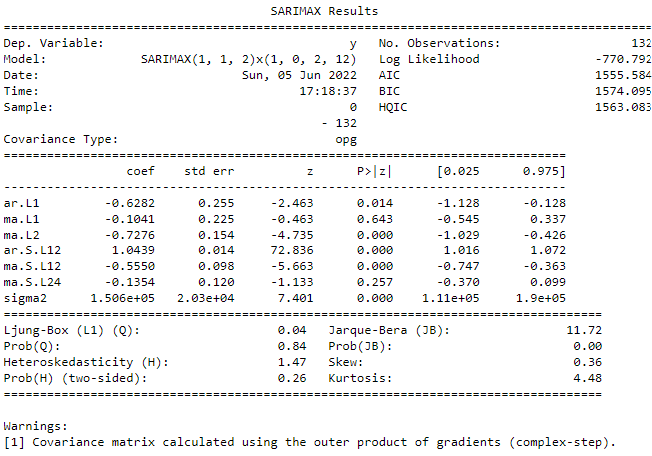
****

**Figure 6.4 Automated sarima 6 on sparkling data**

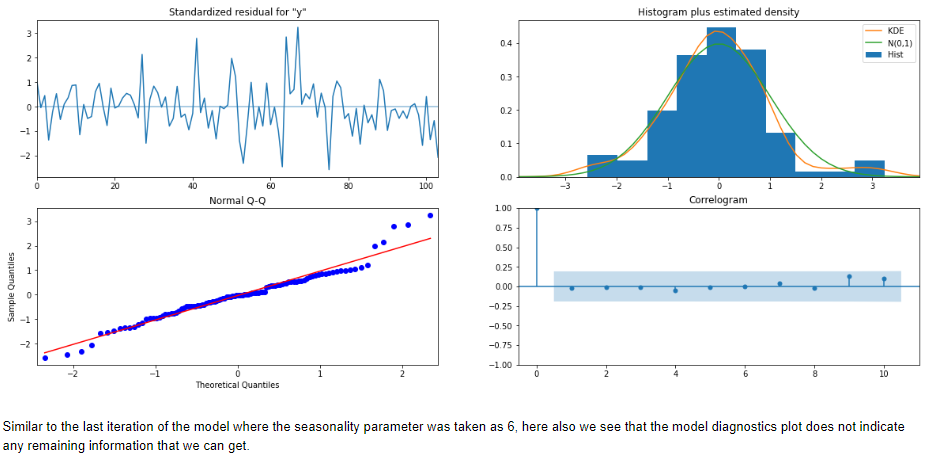
****

****

* The above table has the parameters for the normal method on sparkling data

****

**Figure 6.5 Automated sarima 12 on sparkling data**

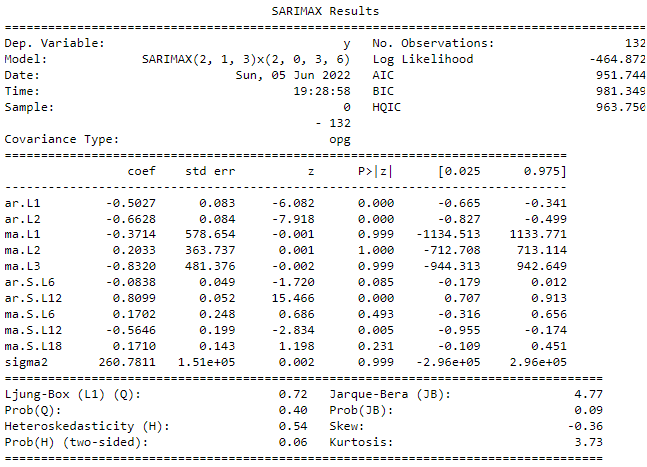
****

**Figure 6.6 Automated sarima 12 on sparkling data**

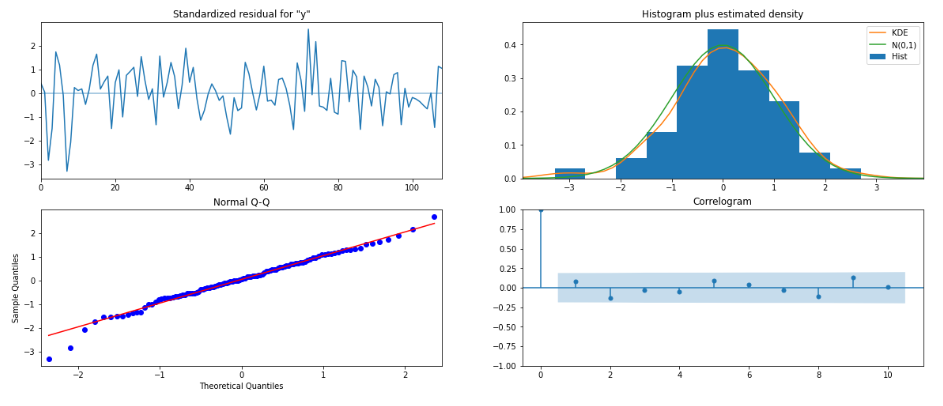
****

****

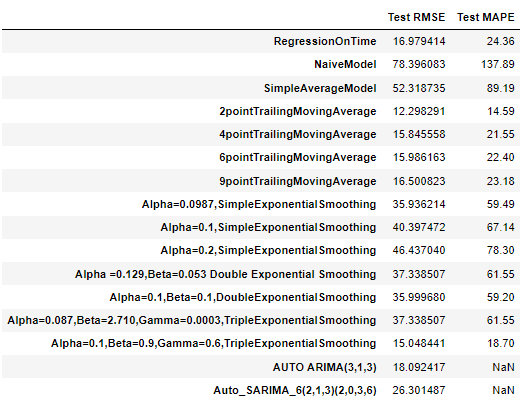
* The above table has the parameters for the normal method on rose data

****

**Figure 6.7 Automated sarima 6 on rose data**

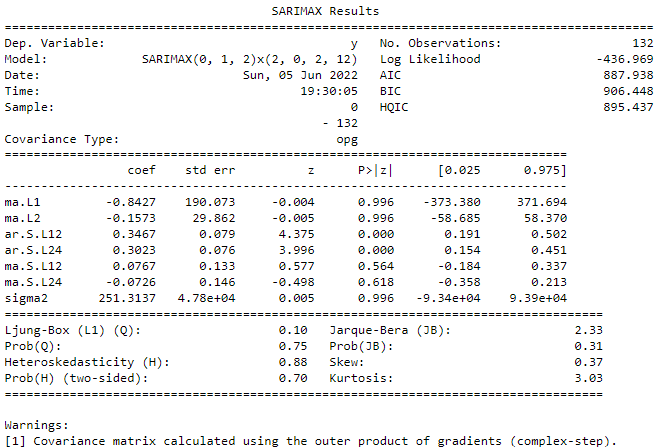
****

**Figure 6.8 Automated sarima 6 on rose data**

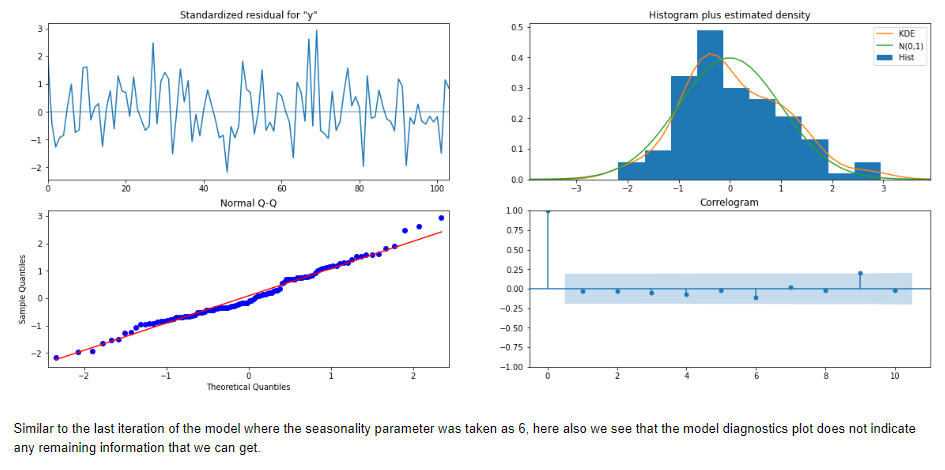
****

****

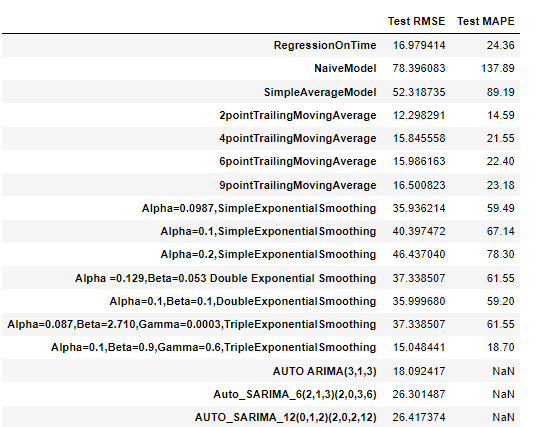
* The above table has the parameters for the normal method on rose data

****

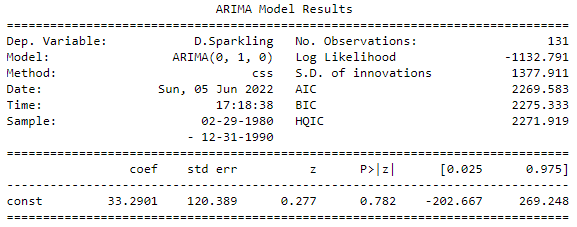
**Figure 6.9 Automated sarima 12 on rose data**

****

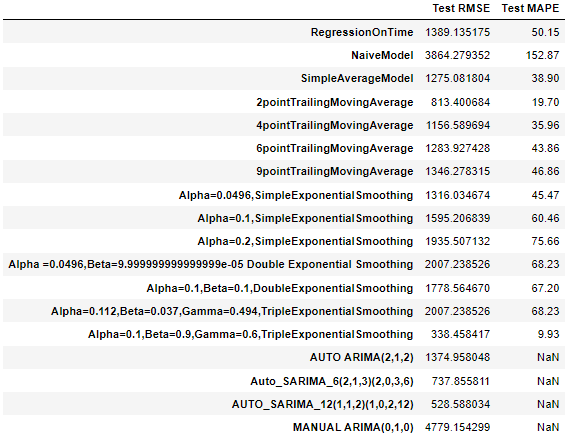
**Figure 6.10 Automated sarima 12 on rose data**

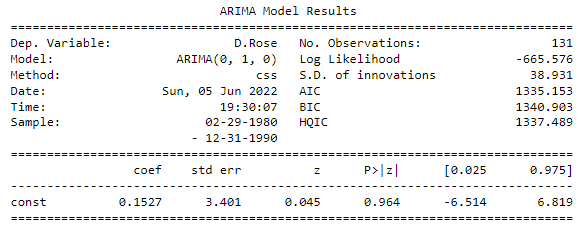
****

**7. Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.**

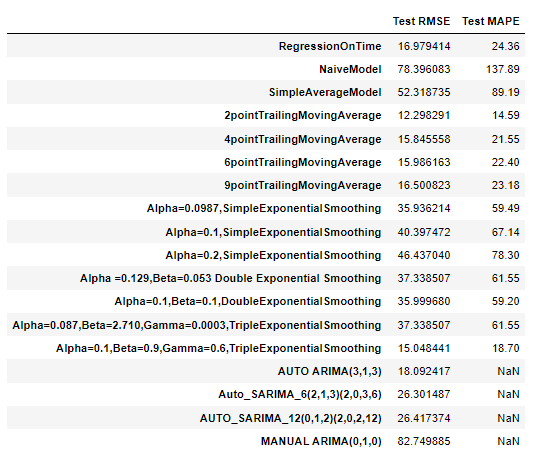
****

**Figure 7.1 Manual arima on sparkling data**

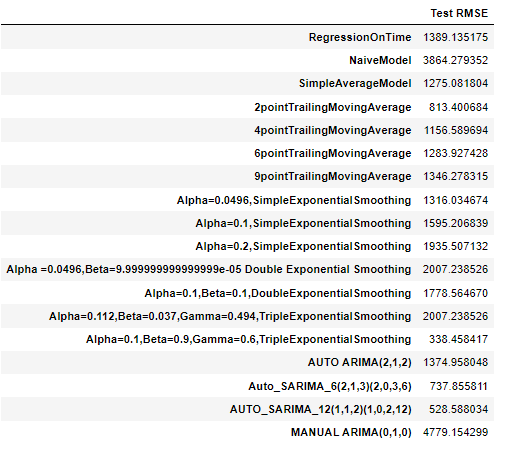
****

****

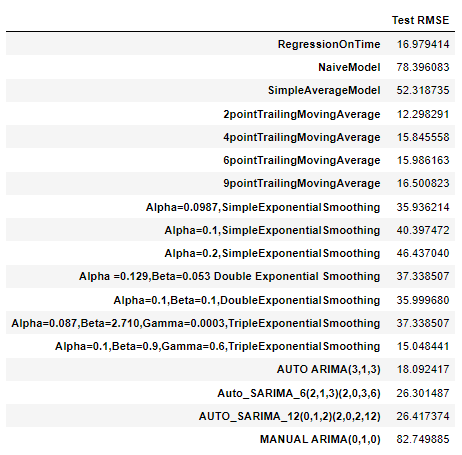
**Figure 7.2 Manual arima on rose data**

****

**8. Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.**

****

**RMSE results for all models on sparkling data**

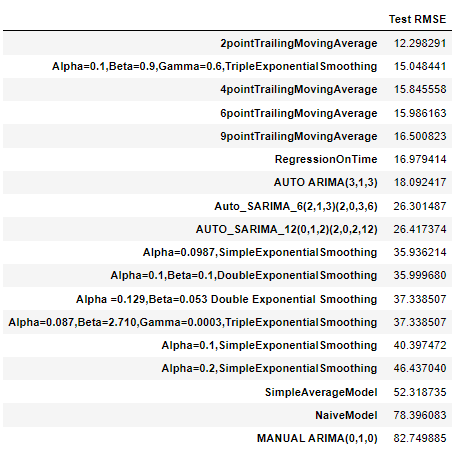
****

**RMSE results for all models on rose data**

**9. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.**

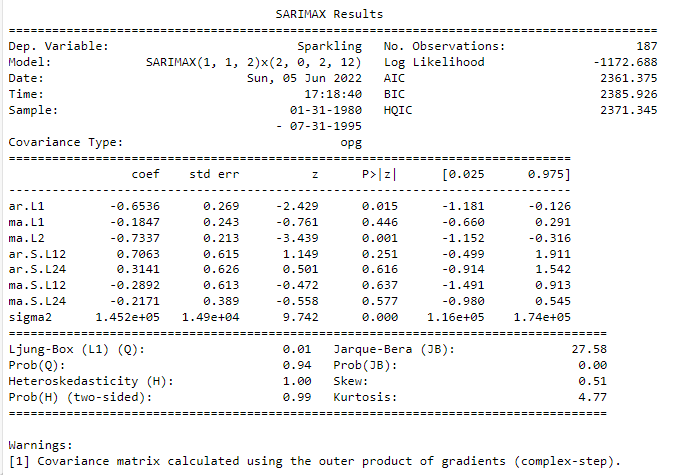
****

**RMSE arranged on sparkling data**

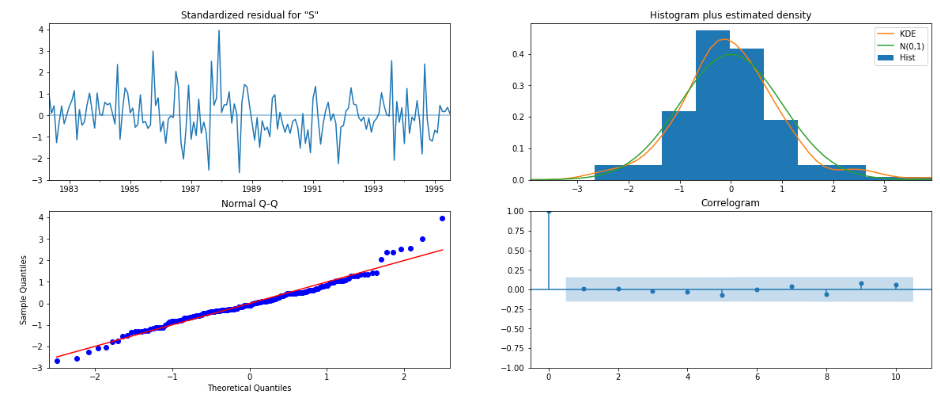
****

**RMSE arranged on rose data**

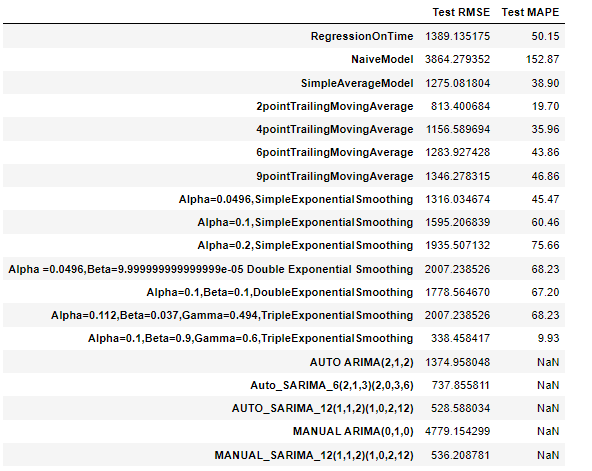
* As we arrange all RMSE values, model with lowest RMSE will be the best model

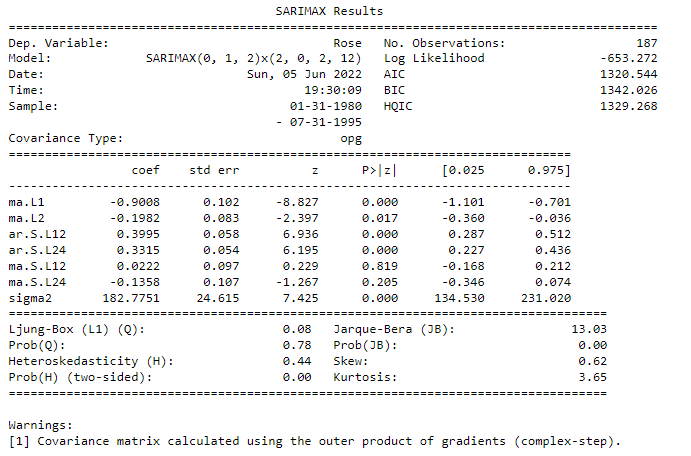
****

**Figure 9.1 Manual sarima on full data for sparkling data**

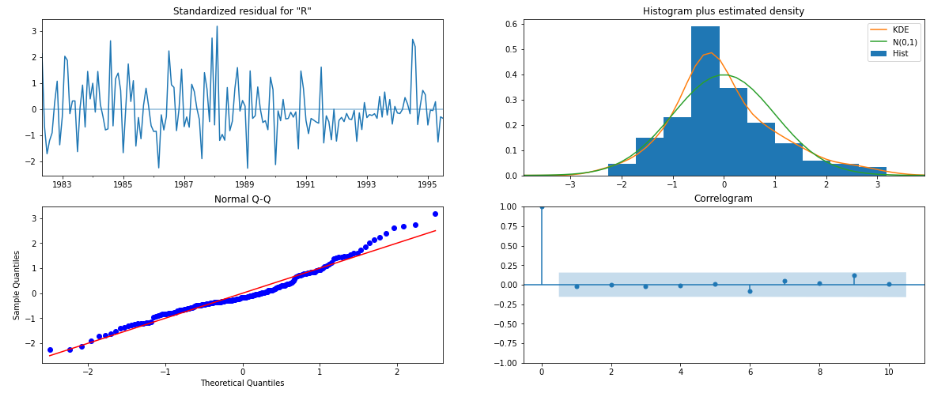
****

**Figure 9.2 Manual sarima on full data plot of sparkling data**

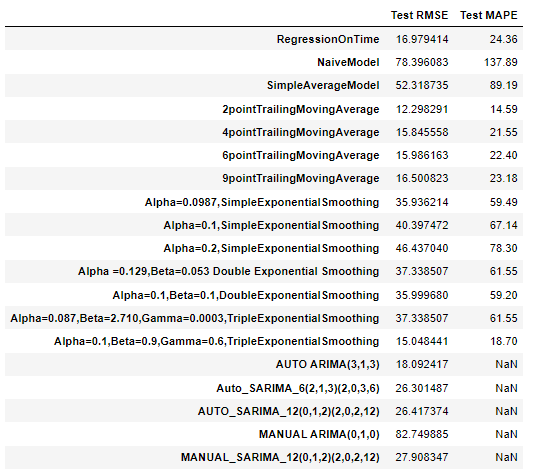
****

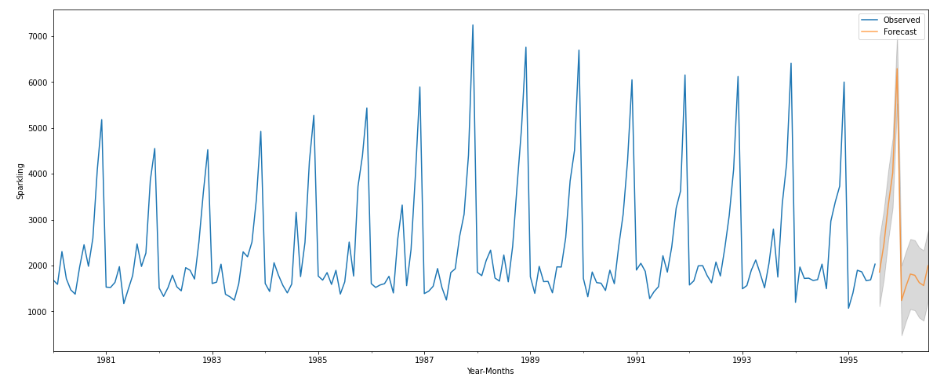
****

**Figure 9.3 Manual sarima on full data for rose data**

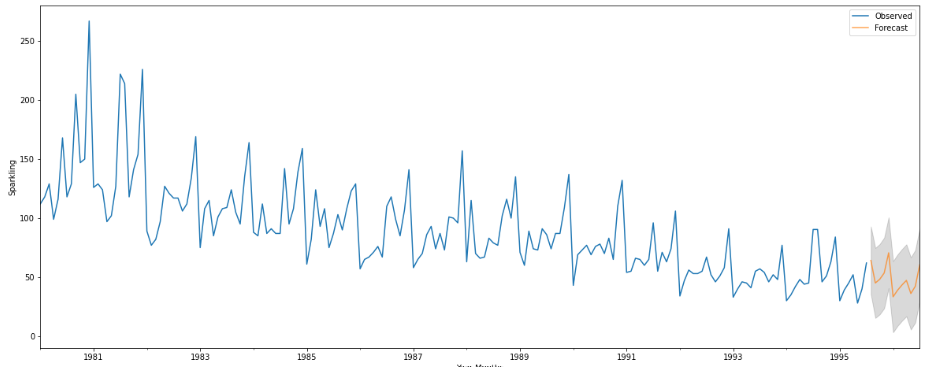
****

**Figure 9.4 Manual sarima on full data plot for rose data**

****

****

**Figure 9.5 Forecast for 12 months on sparkling data**

****

**Figure 9.5 Forecast for 12 months on rose data**

**10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.**

**For sparkling data:**

* From all the above models, we take less RMSE model as the most optimum model
* We have predicted the next 12 months for next year
* From the prediction we can observe that next year sales are going to be high and profitable

**For Rose data:**

* From all the above models, we take less RMSE model as the most optimum model
* We have predicted the next 12 months for next year
* From the prediction we can observe that next year sales are going an average sales for the rose company