**TIME SERIES FORECASTING**

**INDIVIDUAL ASSIGNEMENT**

**SANJEEVI RS**

**DSBA FEB2020**

**DATE: 20/10/2021**

**PROBLEM: -**

The ABC Estate company wants to analyse and forecast the wine sales in the 20th century. Data given for 2 different wine sales in the 20th century is to be analysed.

**Results- of SPARKLING DATA SET**

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

* As we need to forecast the wine sales, we will look at the past values and try to gauge and extract a pattern.
* Here we observe a pattern within each year indicating a seasonal effect. Such observations will help us in predicting future values.
* We have used only one variable, sales (the sales of the past 15 years) between 1980 t0 1995.
* This is a Univariate Time Series Analysis/Forecasting.

**Reading Sparkling data:**

* Sparkling data set has 2 variables.
  + Date time index from 1980 till 1995 around 187 entries.
  + Wine sales for each month in integer around 187 entries.
  + No null values observed in this dataset.
  + Below fig 1 shows the head and tail of the sparkling data set.

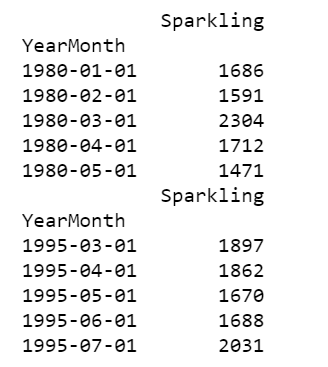
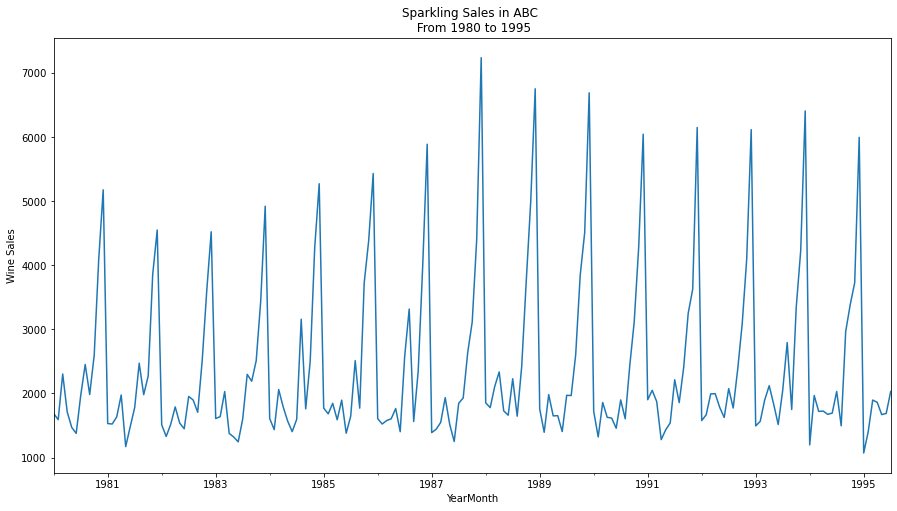
****

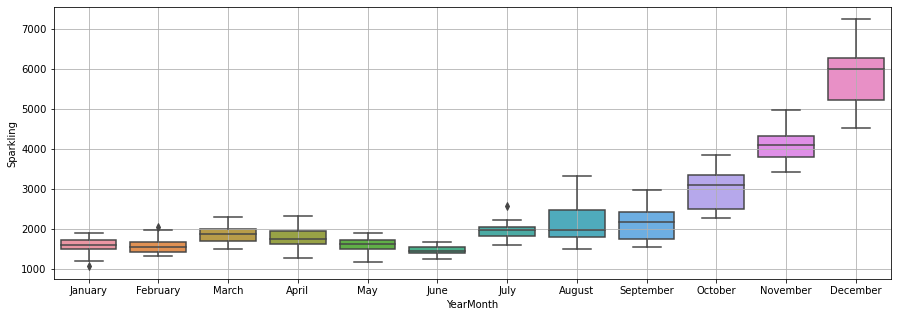
Fig 1

 Graph 1

* Above Graph 1 shows the seasonality and trend of the sparkling wine sales from 1980 till 1995.
* Sales in the end of each year is peak especially in Oct, Nov & Dec.

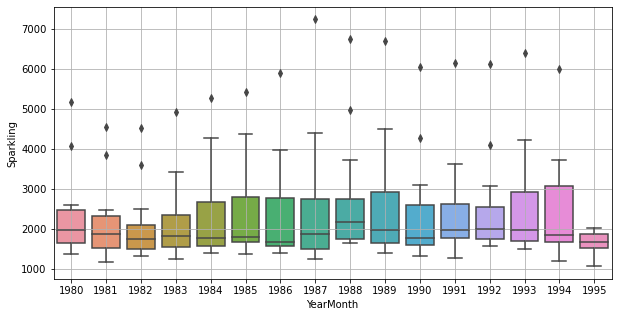
**2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.**

Monthly BOXPLOT



Graph 2

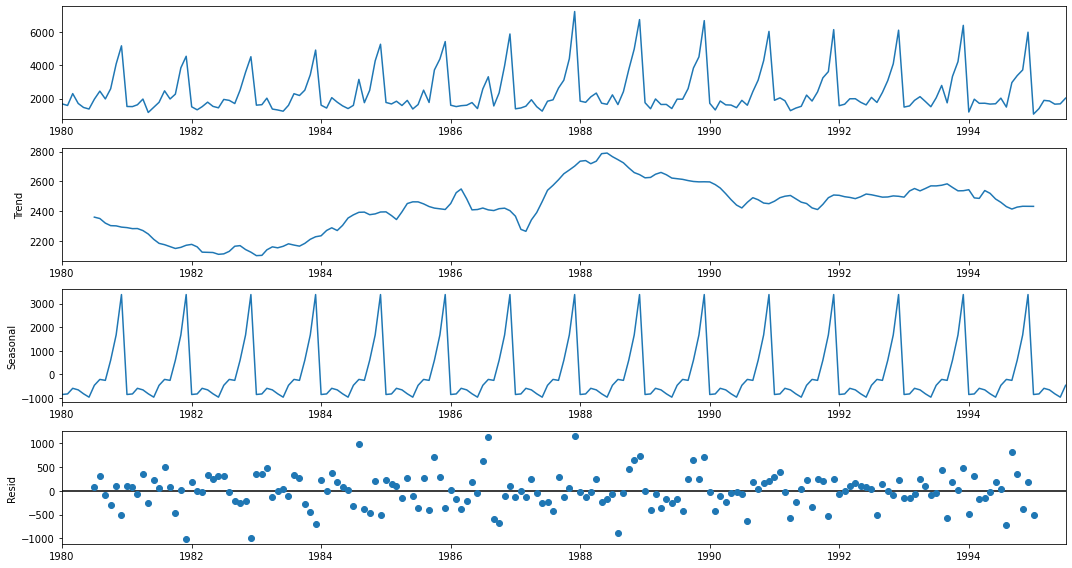
Yearly BOXPLOT



Graph 3

* + Graph 2 shows the sparkling sales of different month and Graph 3 shows the sales of different years.
  + The average sales of 1986 is the lowest and 1988 is the highest, other years average sales are almost same.

Decompose The Time Series into Various Components



Graph 4

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

* Sparkling dataset has 187 rows and 2 columns.
* Train data is >1991 and it has 132 rows.
* Test data is <=1991 and it has 55 rows.
* Below fig 2 shows the first and last few rows of Train and Test data.
* Test and Train data plot and shown the trend in Graph 5 below.

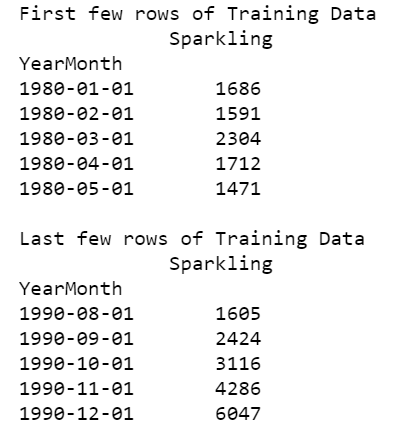
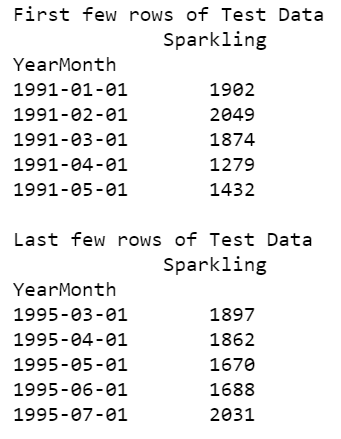
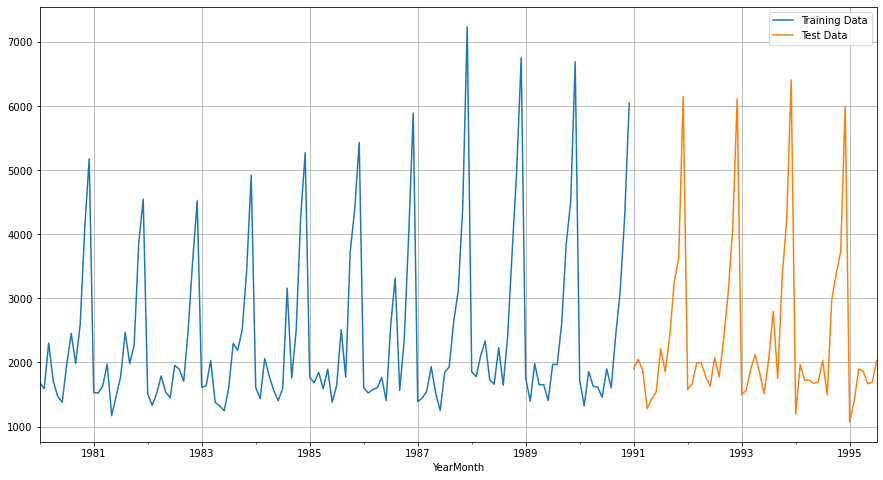
 

Fig 2



Graph 5

**4. Build various 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, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE.**

1. **Linear Regression: -**

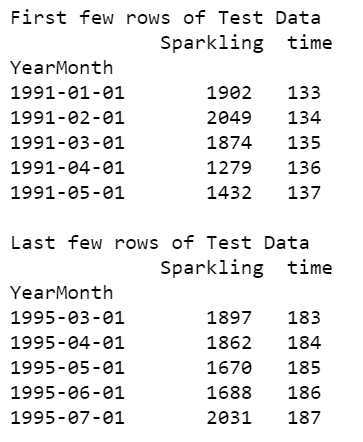
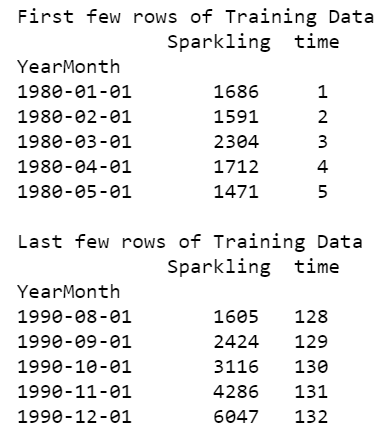


Fig 3

* Time instance added in train and test dataset
* Length of Train Data is 1-132.
* Length of Test Data is 133-187.
* Above Fig 3 shows the first and last few rows for Train and Test Data along with Time instance.
* Below Graph 6 shows the Regression on Time Test data and on Train Data.
* The graph 6 clearly shows the line across the train and test data.
* Below Fig 4 shows the Test RMSE and MAPE for Regression Model.
* For Regression on time forecast on the Training Data

RMSE is 1279.322

MAPE is 40.70

* For Regression on time forecast on the Test Data

RMSE is 1389.135

MAPE is 59.35

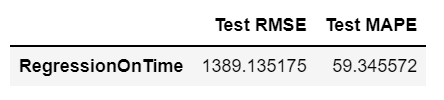
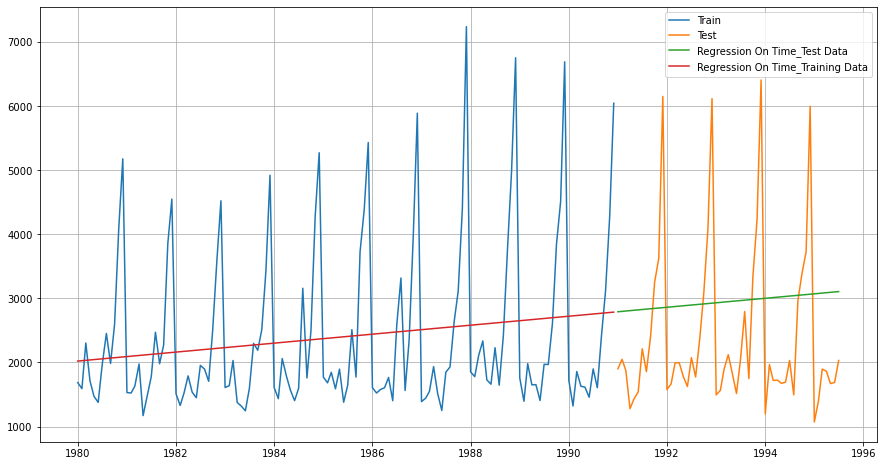


Fig 4

****

Graph 6

1. **Naive: -**

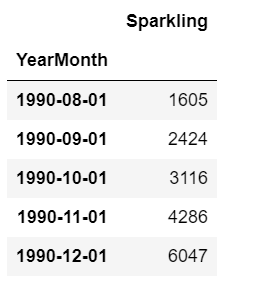
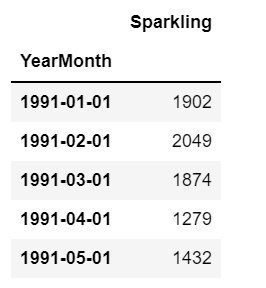
 

Fig 5

* Above Fig 5 shows the Train and Test of sparkling data for Naive model.
* Below Fig 6 shows the First Few rows of Naïve Test data.
* Length of Test Data is 133-187.
* Below graph 7 shows the Naïve forecast on test data.
* For Regression on time forecast on the Test Data

RMSE is 3864.279

MAPE is 201.33

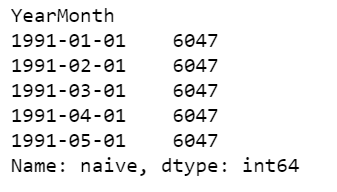
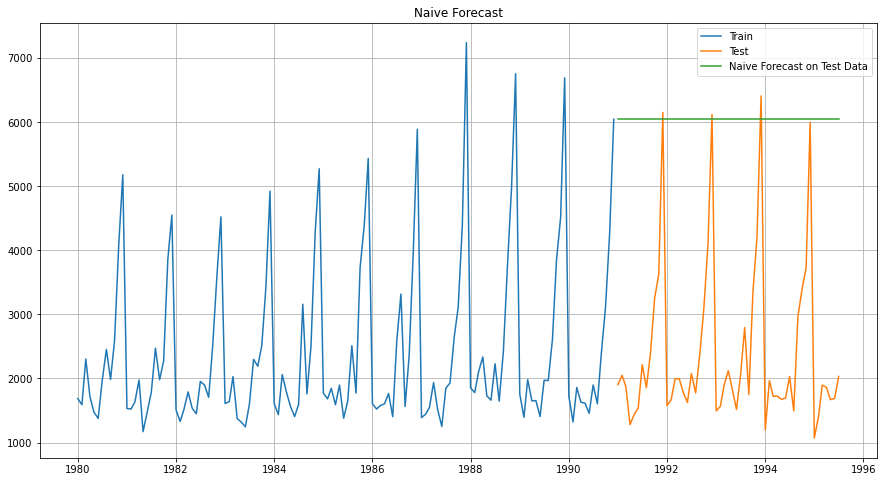


Fig 6

****

Graph 7

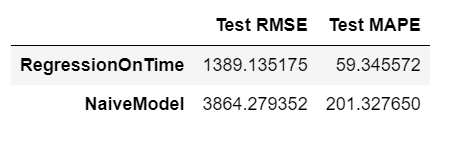


Fig 7

1. **Simple Average: -**

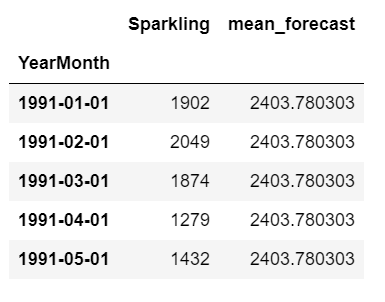


Fig 8

* Above Fig 8 shows the first few rows of Simple Average Test Data.
* In below Fig 9 the green line denotes the simple average on test data
* For Simple average forecast on the Test Data

RMSE is 1275.082

MAPE is 39.16

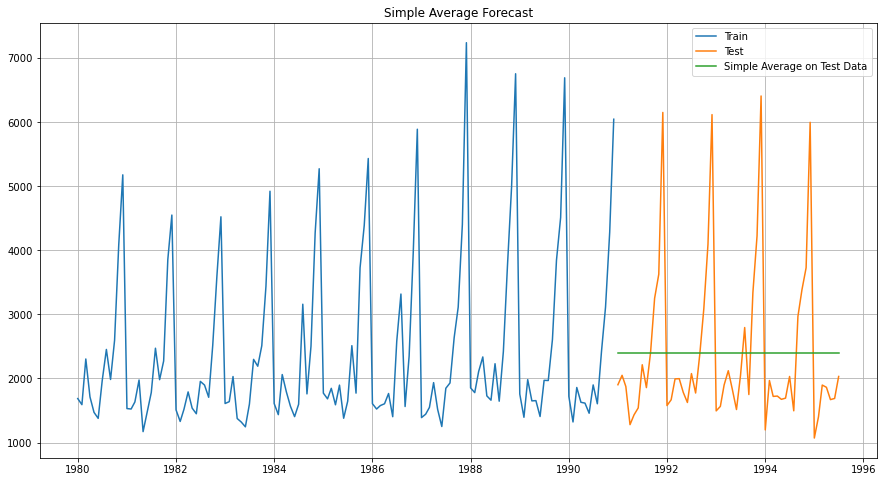


Fig 9

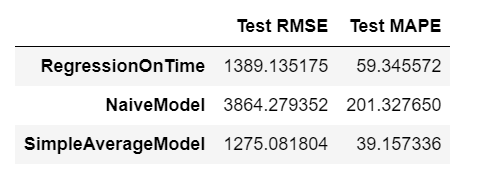


Fig 10

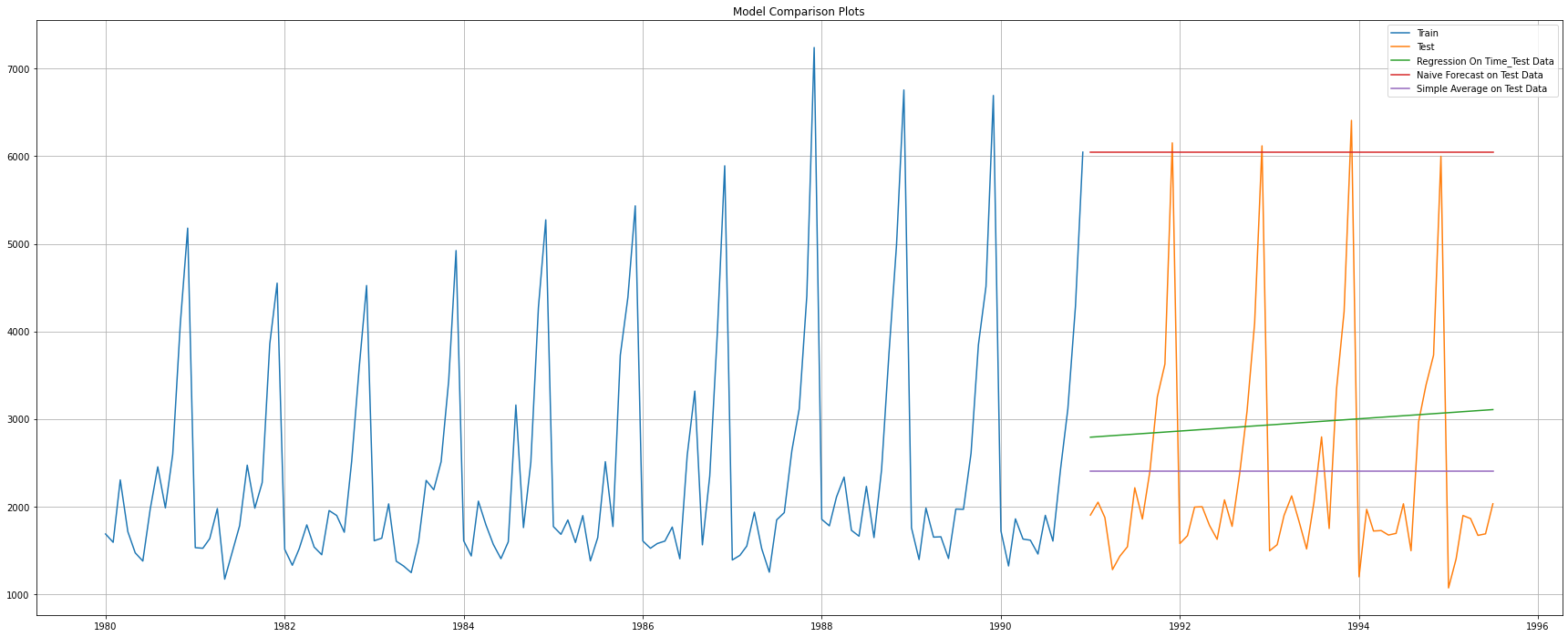


Fig 11

* All model RMSE and MAPE shown in Fig 10 to compare.
* In Fig 11 all the models are plotted to compare the time series on Test Data.

1. **Simple Exponential Smoothing: -**

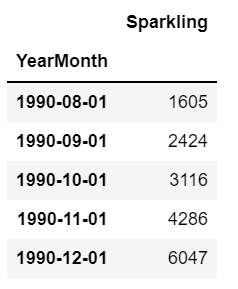
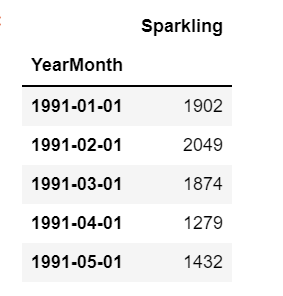
 

Fig 12

* Above Fig 12 shows the Train and Test Data for Simple Exponential Smoothing.
* The Alpha Value SES is 0.0496.
* In Fig 13 the first few rows and last few rows of Train Predict values are shown.
* Fig 14 shows the first few rows of test predict.

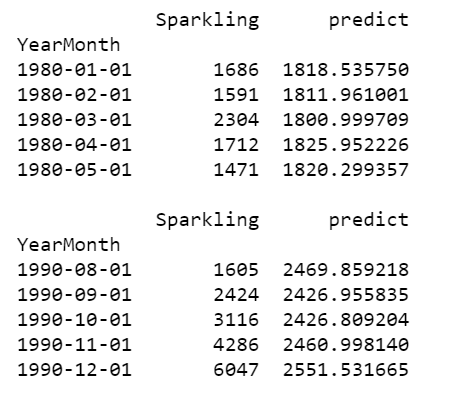


Fig 13

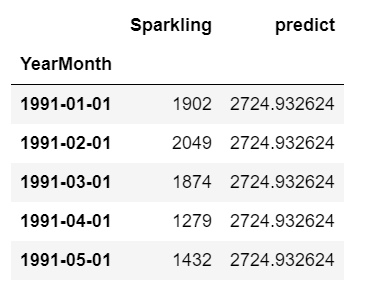


Fig 14

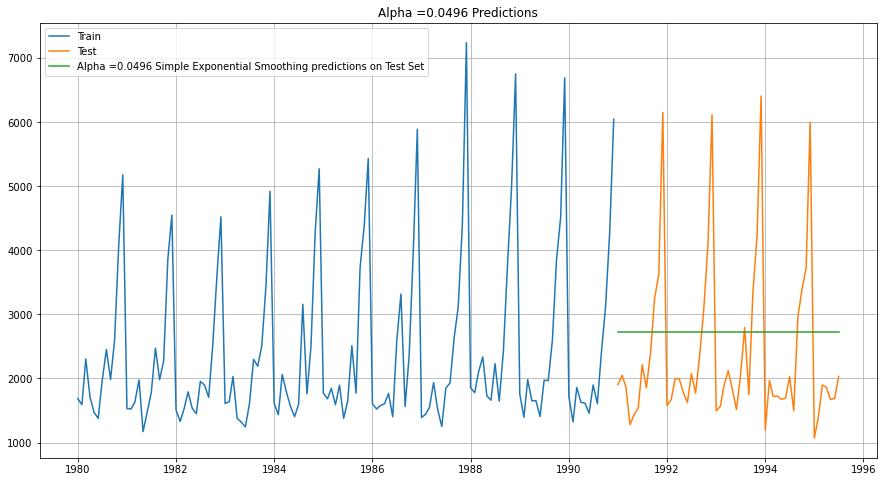
****

Fig 15

* Above Fig 15 Simple Exponential Smoothing predictions on Test Set for Alpha = 0.0496.
* We can see the forecast model and the respective RMSE and MAPE in Fig 16.
* SES forecast done for four Apha Values.
* All the values plotted in both train and test data shown in Fig 17.

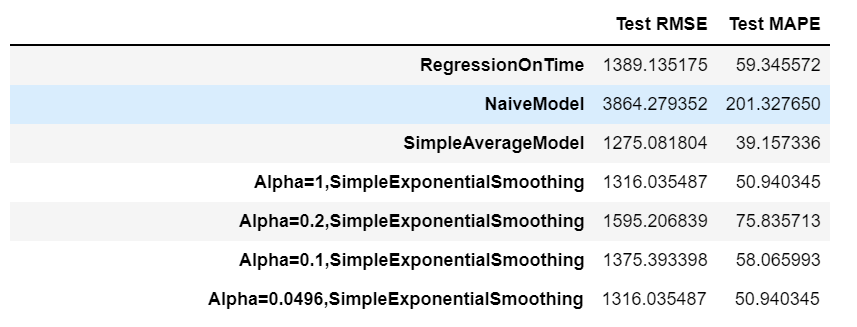
****

Fig 16

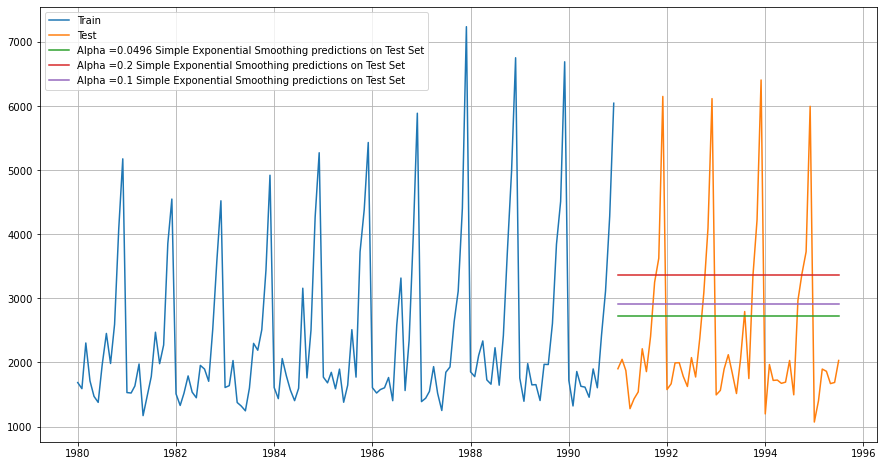
****

Fig 17

1. **Double Exponential Smoothing: -**

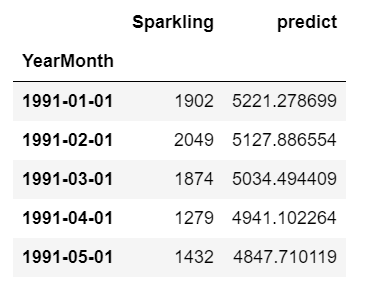
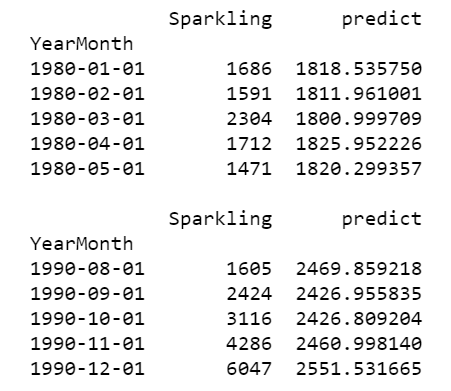
 

Fig 18 Fig 19

* DES Test Predict First Few rows shown in Fig 18.
* DES Train Predict First and Last few rows are shows in Fig 19.
* DES prediction on test set for Alpha =0.6885, Beta = 9.9999 is shown in Fig 20.
* RMSE and MAPE for multiple Alpha and Beta values are compared in Fig 21 and sorted low to high.
* The two low values are plotted in Fig 22.

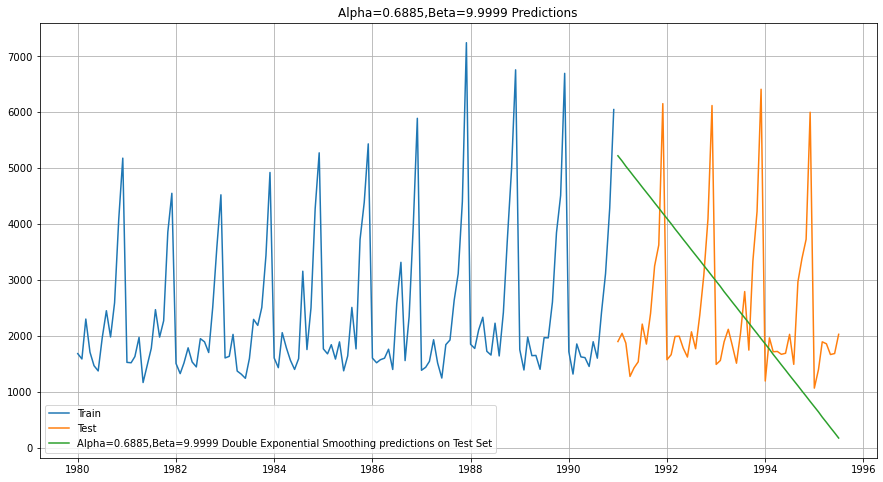


Fig 20

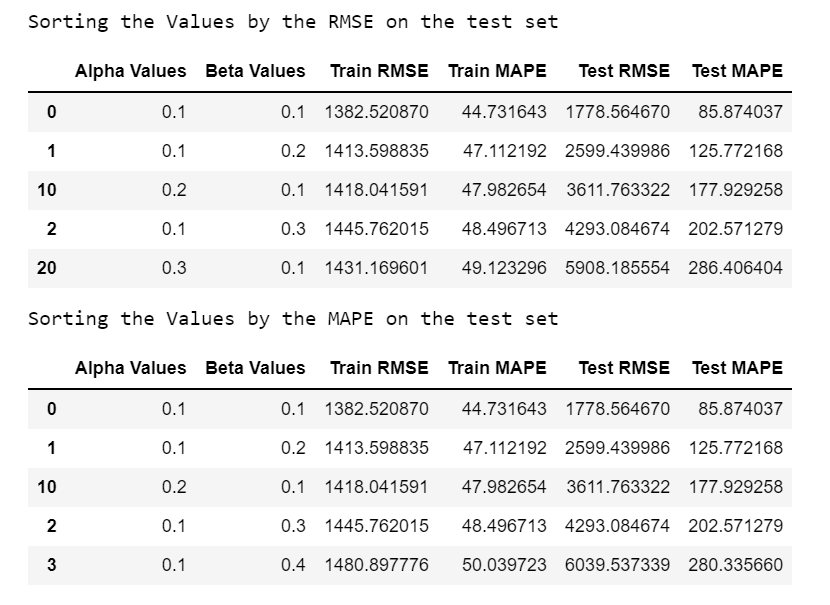


Fig 21

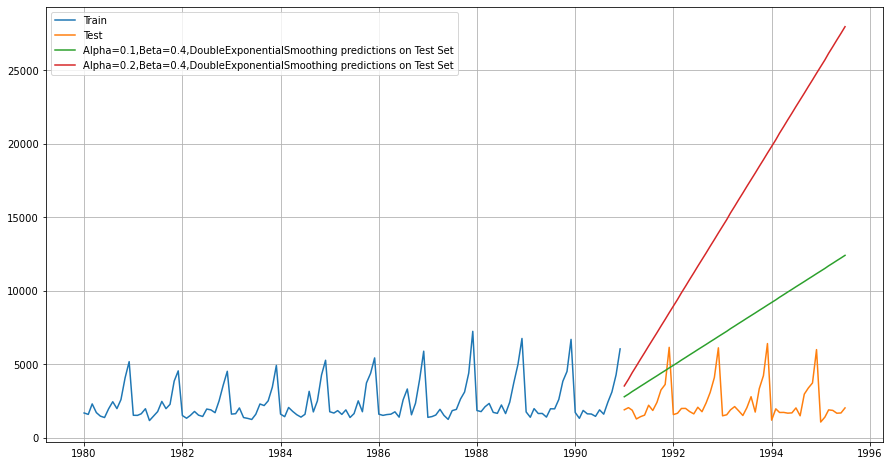
****

Fig 22

1. **Triple Exponential Smoothing: -**

* Below Fig 23 shows the Train and Test set of TES.
* TES prediction for Alpha = 0.1464, Beta = 0, Gama = 0.4377 on test set is shown in Fig 24.
* TES prediction for three different Alpha, Beta, Gama on test set is plotted in Fig 25.
* Fig 26 shows the sorted RMSE value of all the Model completed so far and TES has the least RMSE.
* MAPE sorted in Fig 27 and TES is the lowest again.

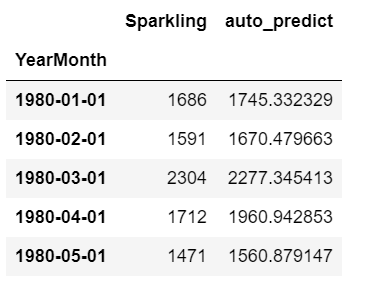
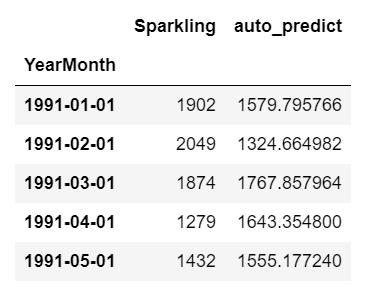
 

Fig 23

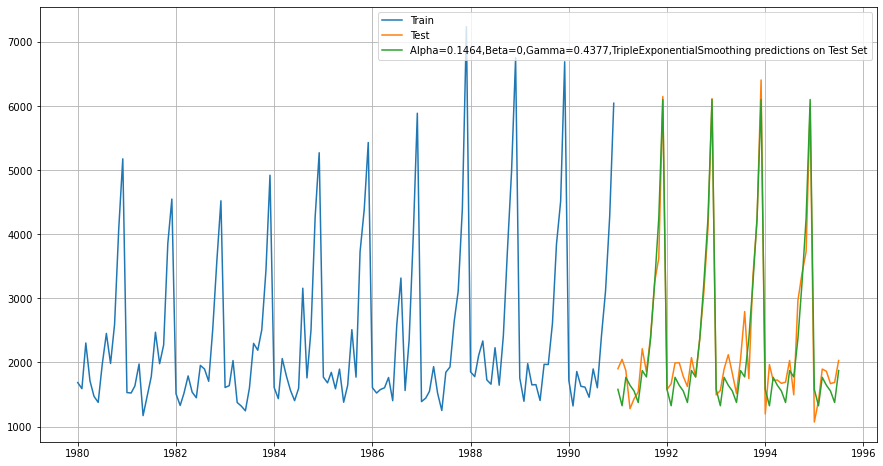
****

Fig 24

****

Fig 25

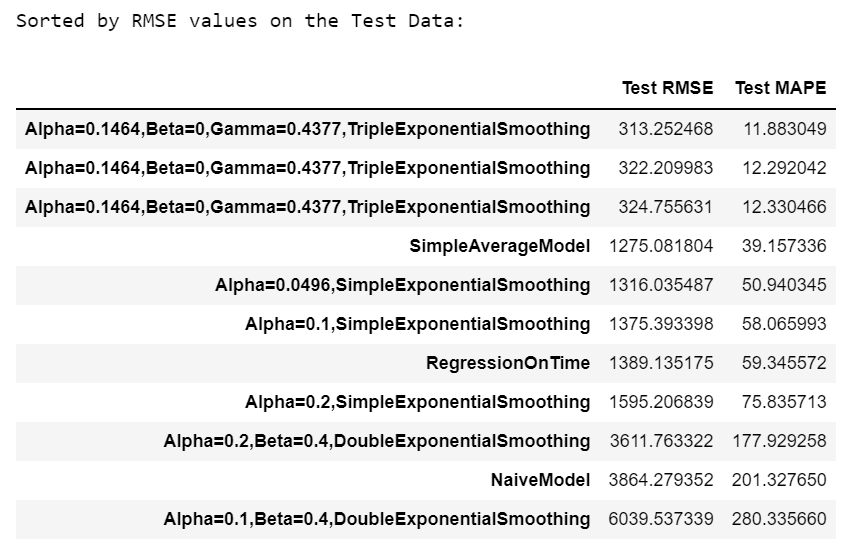


Fig 26

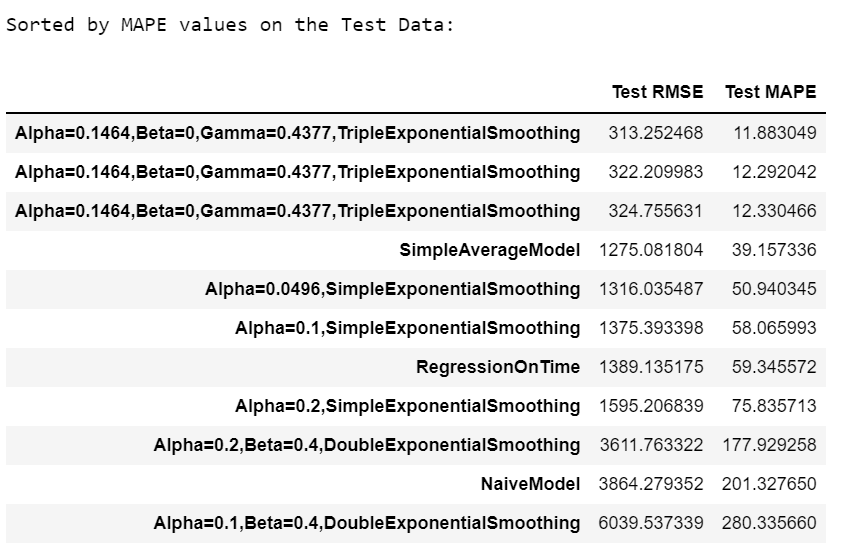


Fig 27

**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.**

* The p Value for test set is 0.5674.
* TES prediction for Alpha = 0.1464, Beta = 0, Gama = 0.4377 on test set is shown in Fig 24.
* The differenced data auto correlation shown in Fig 30 and partial auto correlation data shown in Fig 31.

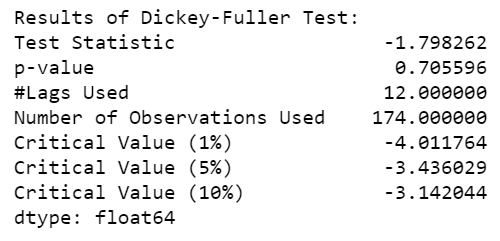


Fig 28

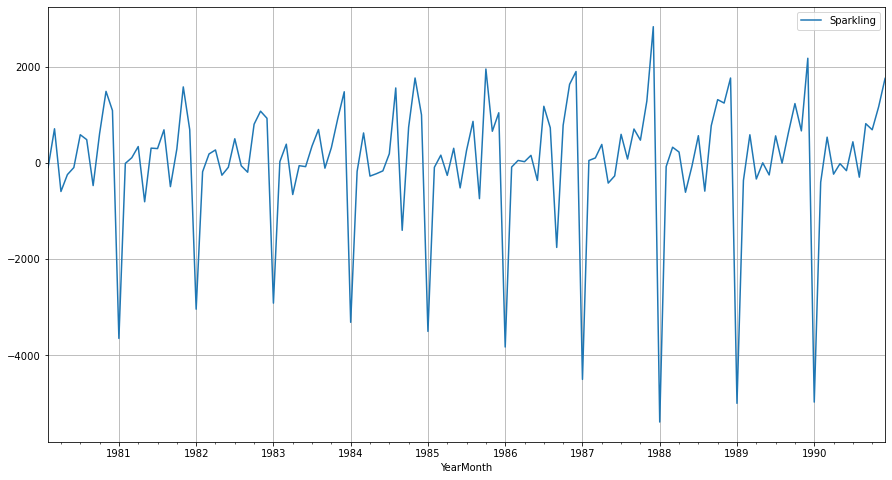


Fig 29

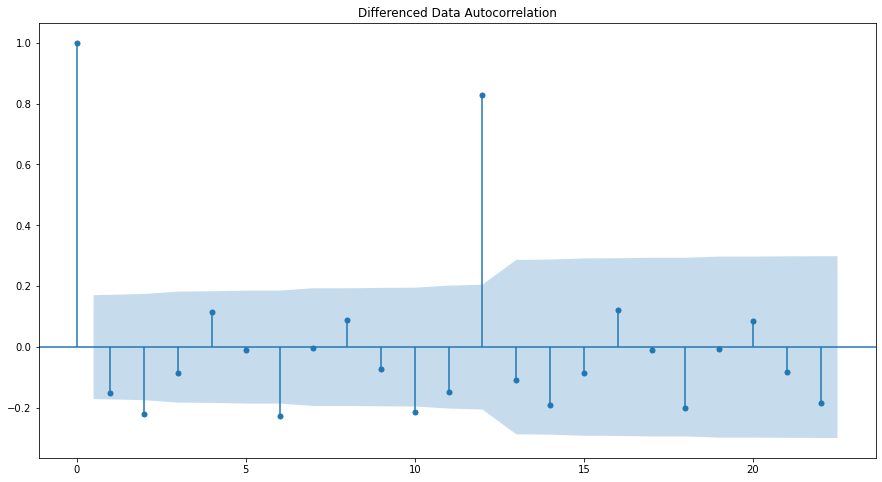
****

Fig 30

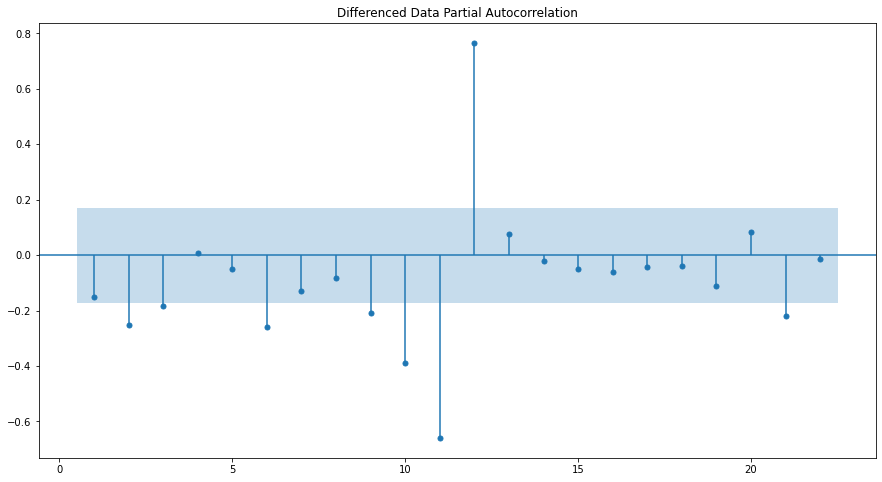
****

Fig 31

**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.**

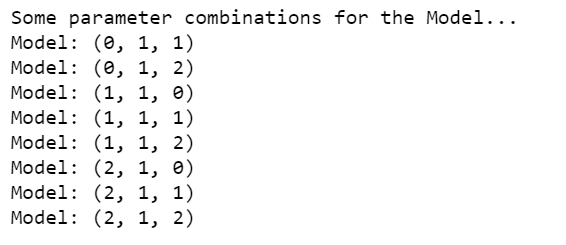
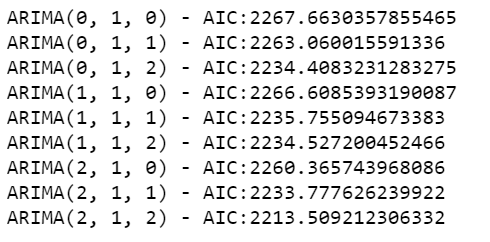
 

Fig 32 Fig 33

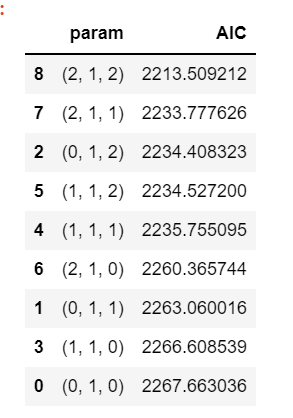


Fig 34

SARIMAX RESULT- (Auto ARIMA)

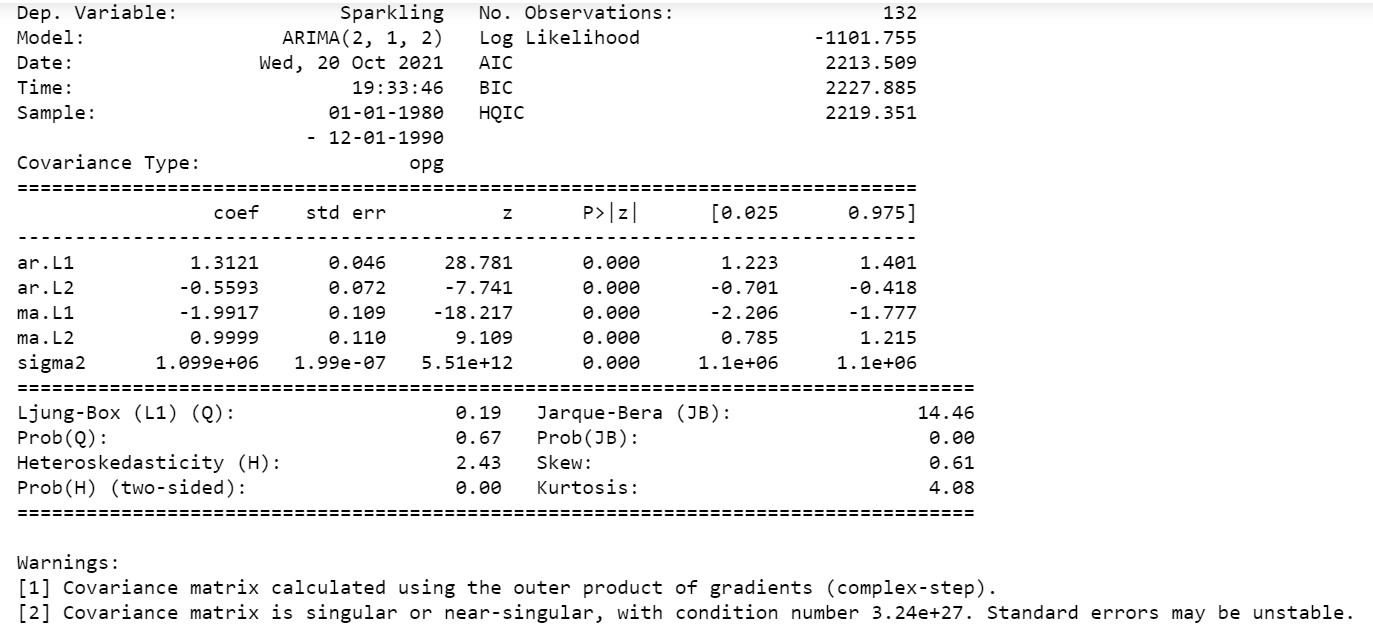


Fig 35

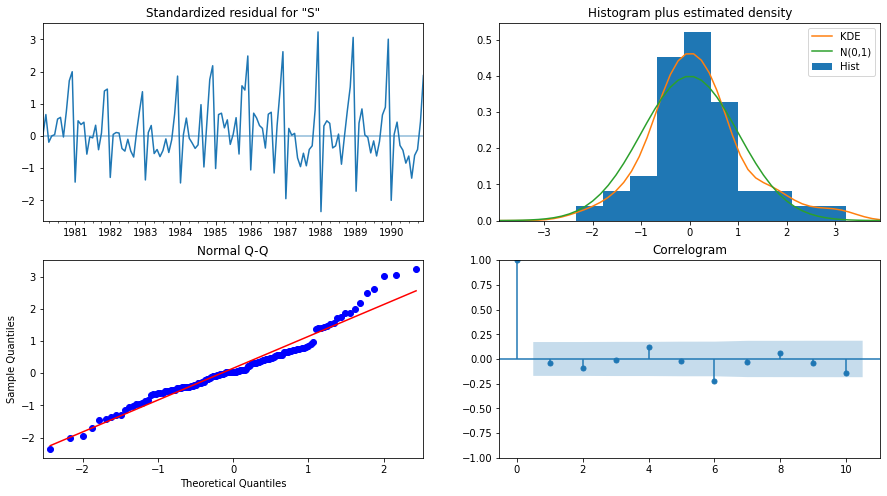
****

Fig 36

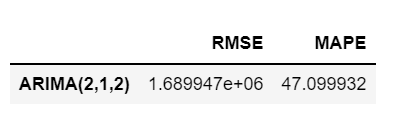


Fig 37

SARIMAX RESULT- (Manual ARIMA)

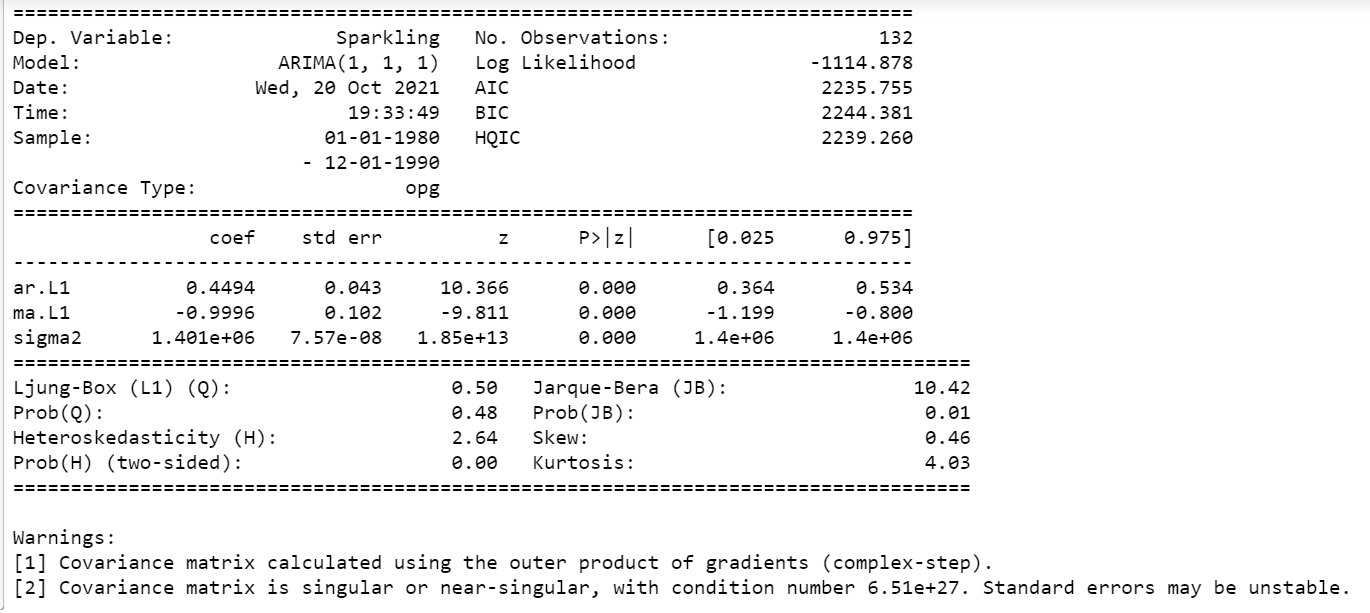


Fig 38

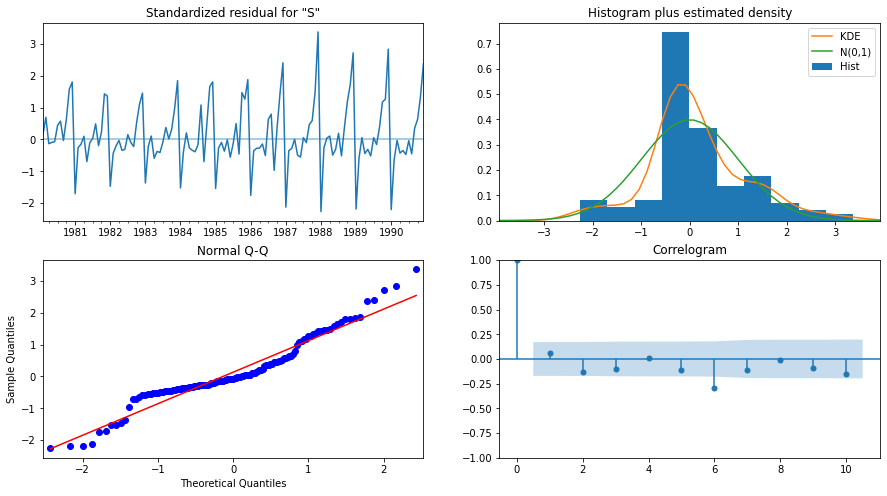


Fig 39

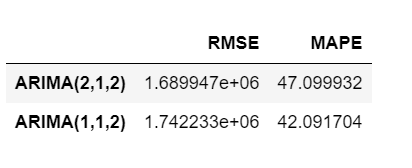


Fig 40

**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.**

**a) ARIMA**

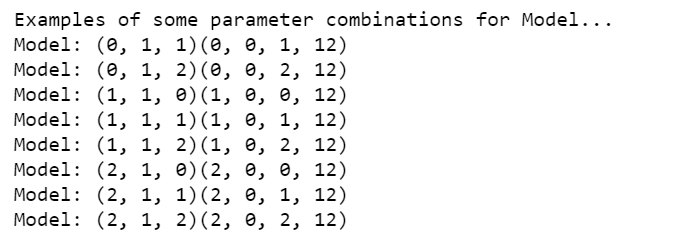
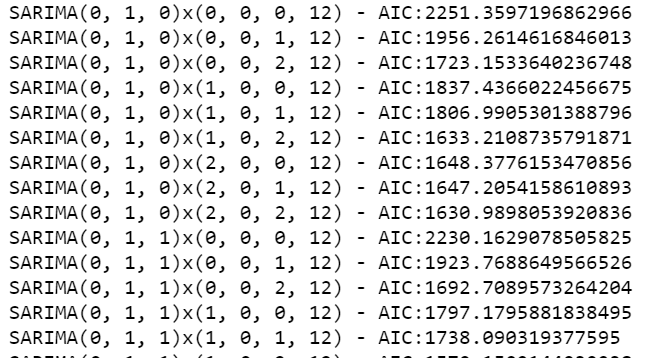
 

Fig 41

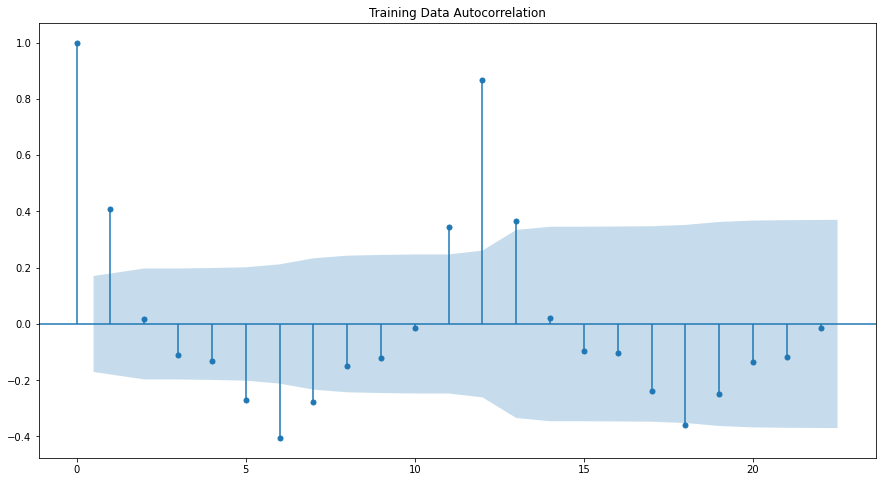
****

Fig 42

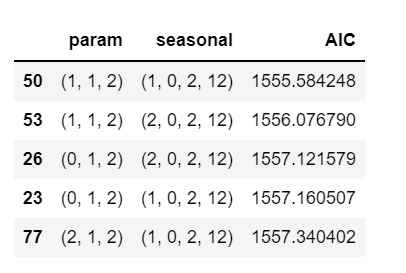


Fig 43

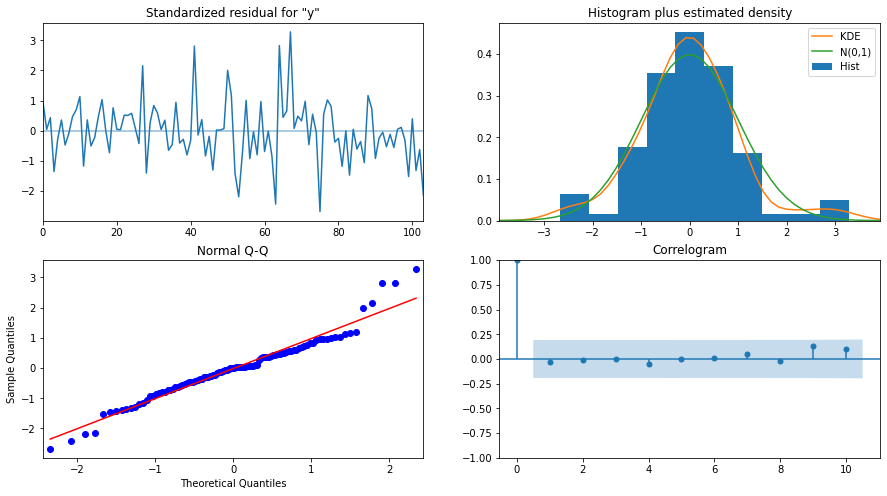


Fig 44

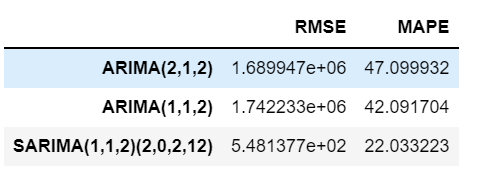


Fig 45

b) **SARIMA**

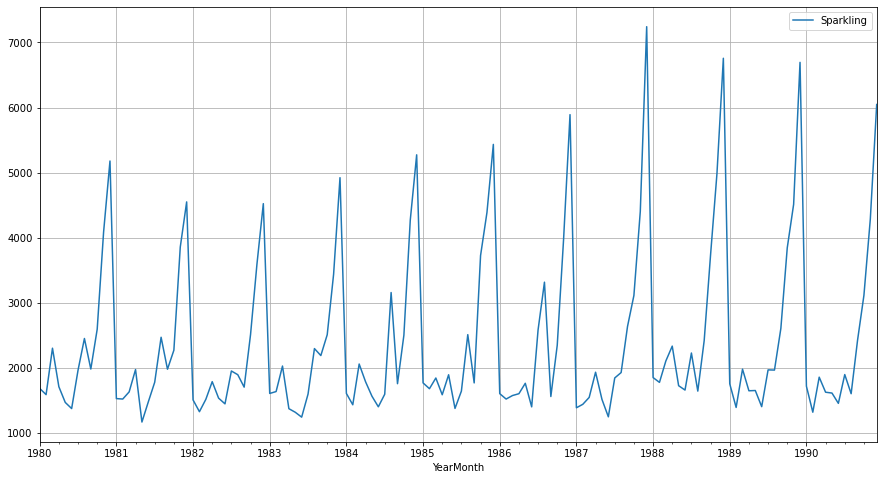


Fig 46

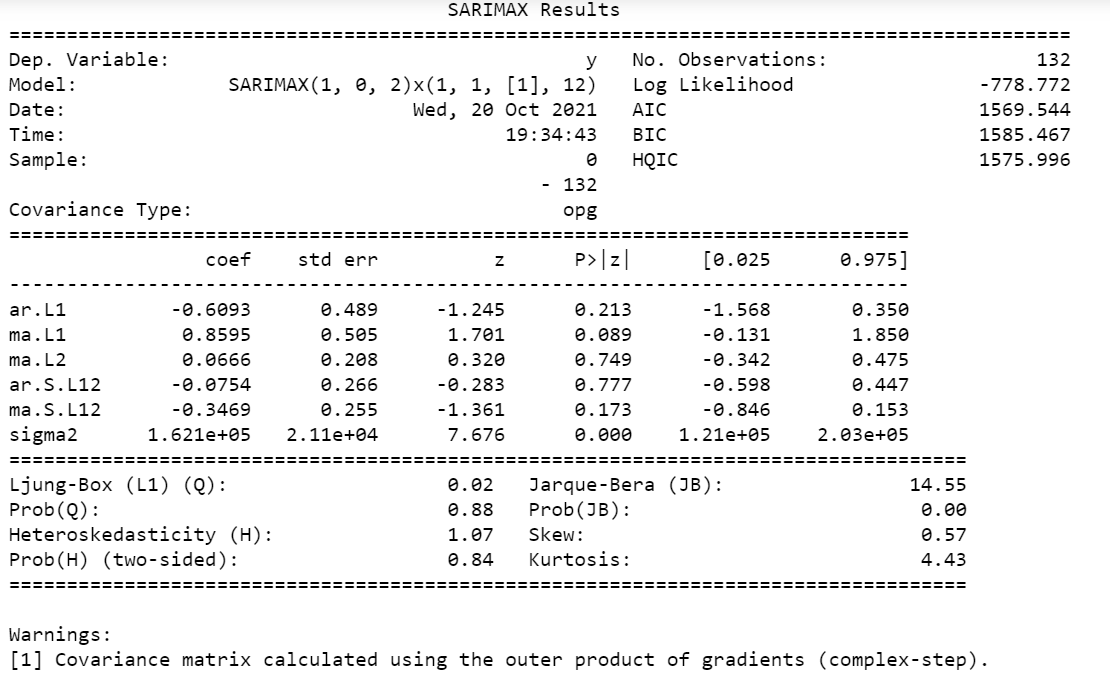


Fig 47

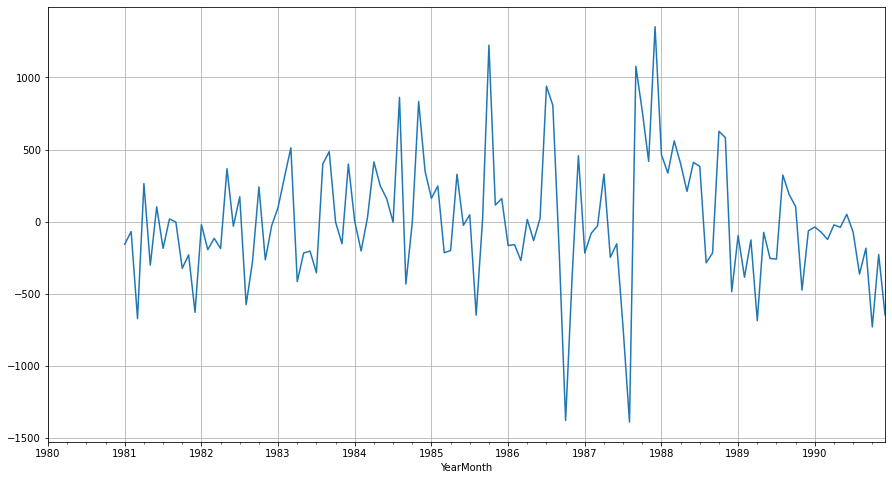
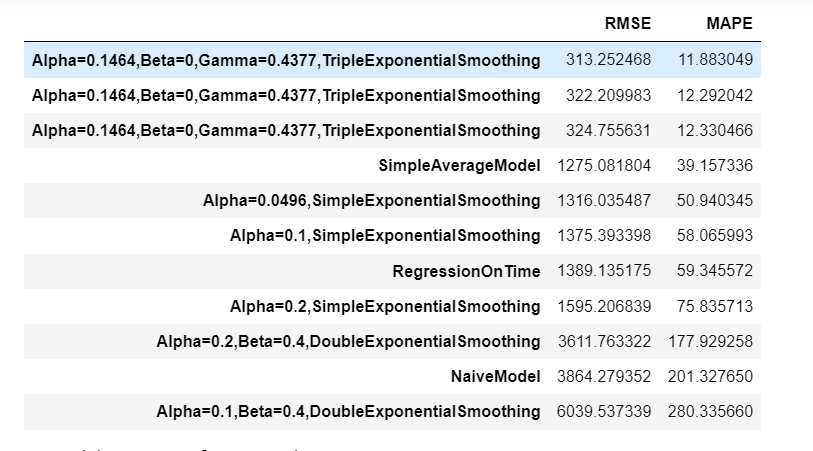
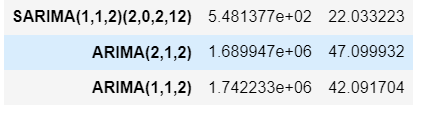


Fig 48

**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.**





**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.**

* Triple Exponential smoothing is the most optimal model as per the RMSE value. The prediction for 12 months not the future is mentioned below with the appropriate confidence

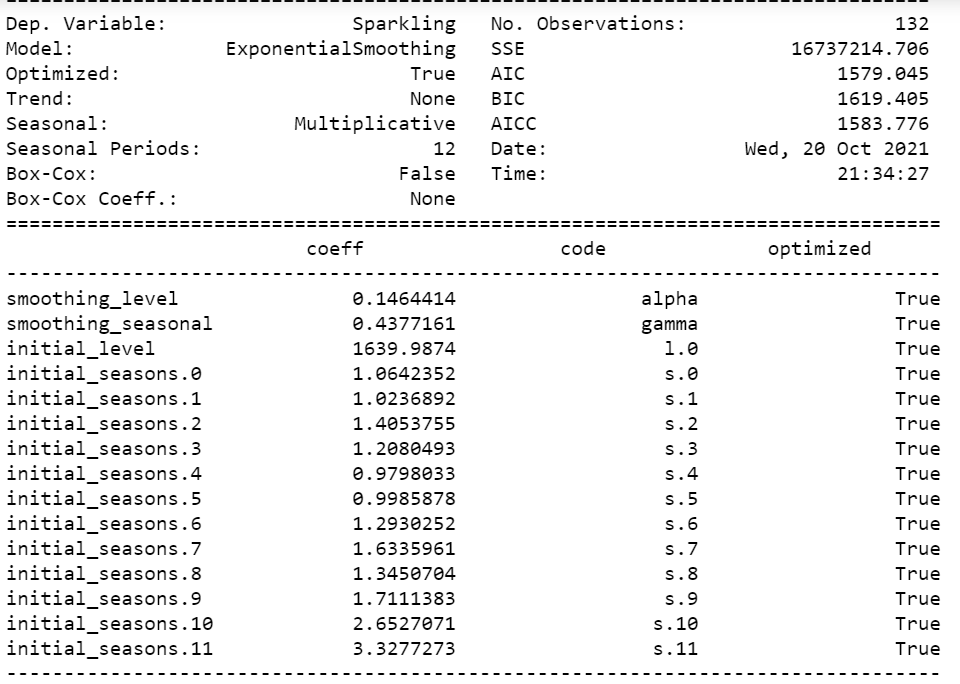


Fig 49

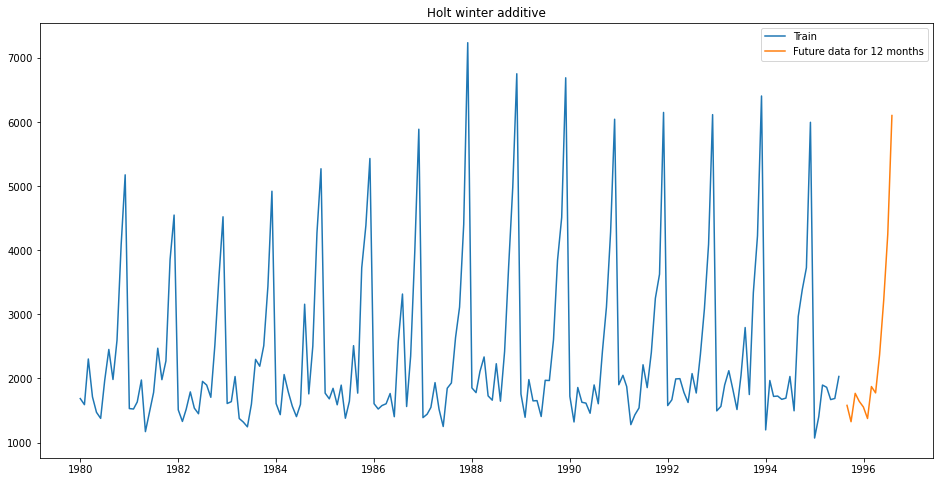
****

Fig 50

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

**Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.**

* As the sales is seasonal and spike at the end of the year, promotions and action required in the first half of the year.
* This is to be followed every year, the sale is increasing year on year.
* The average sales of most the years are similar, only 1988 is above average sales.
* Prediction for upcoming years are mostly the same average sales can be expected.

**Results- of ROSE DATA SET**

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

* As we need to forecast the wine sales, we will look at the past values and try to gauge and extract a pattern.
* Here we observe a pattern within each year indicating a seasonal effect. Such observations will help us in predicting future values.
* We have used only one variable, sales (the sales of the past 15 years) between 1980 t0 1995.
* This is a Univariate Time Series Analysis/Forecasting.

**Reading Rose data:**

* Rose data set has 2 variables.
  + Date time index from 1980 till 1995 around 187 entries.
  + Wine sales for each month in integer around 185 entries.
  + 2 null values observed in this dataset.
  + Below fig 51 shows the head and tail of the sparkling data set.

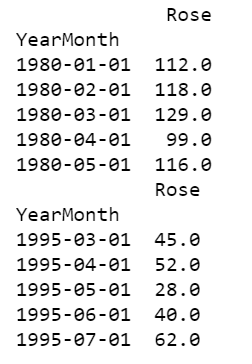
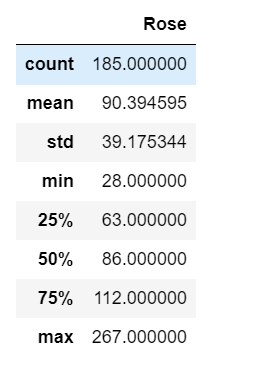
 

Fig 51

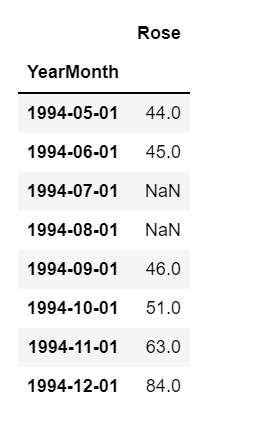
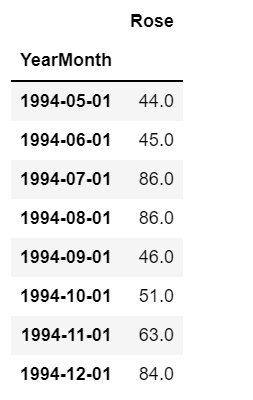
 

Fig 52

* + Null values imputed with the median of the overall Rose sales.
  + Fig 52 shows the missing 07-1994 and 08-1994. Imputed values is also shown.

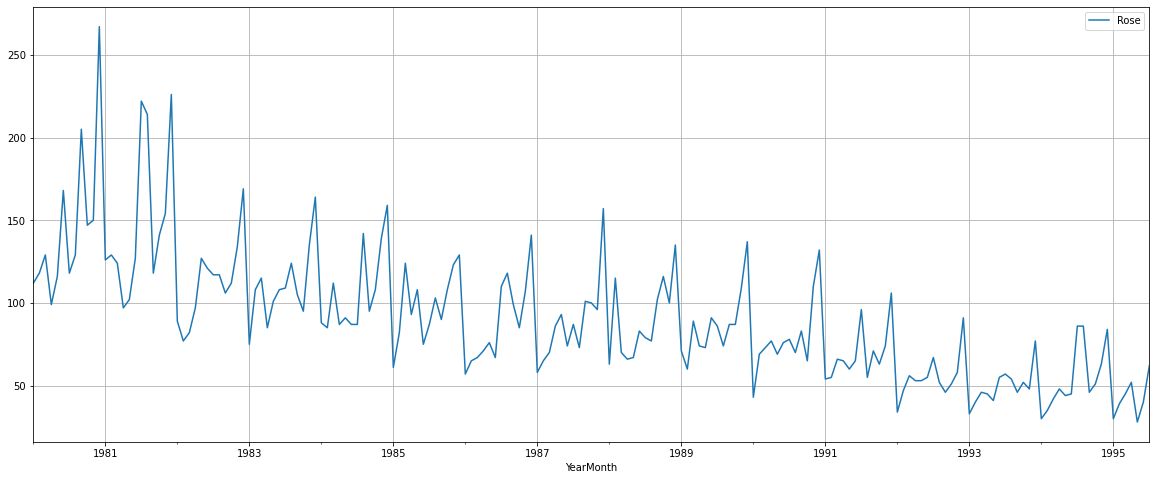
****

Fig 53

1. **Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.**
   * Fig 54 shows the rose sales in different month and Fig 55 shows the sales in different years.
   * The average rose sales decreasing year on year is observed in the above Fig 55.

Monthly Boxplot

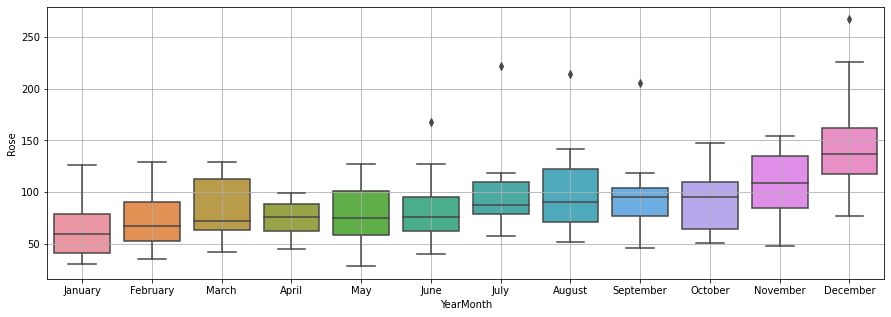


Fig 54

Yearly Boxplot

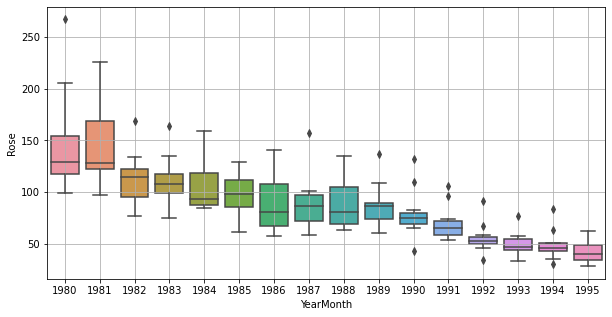


Fig 55

Decompose The Time Series into Various Components

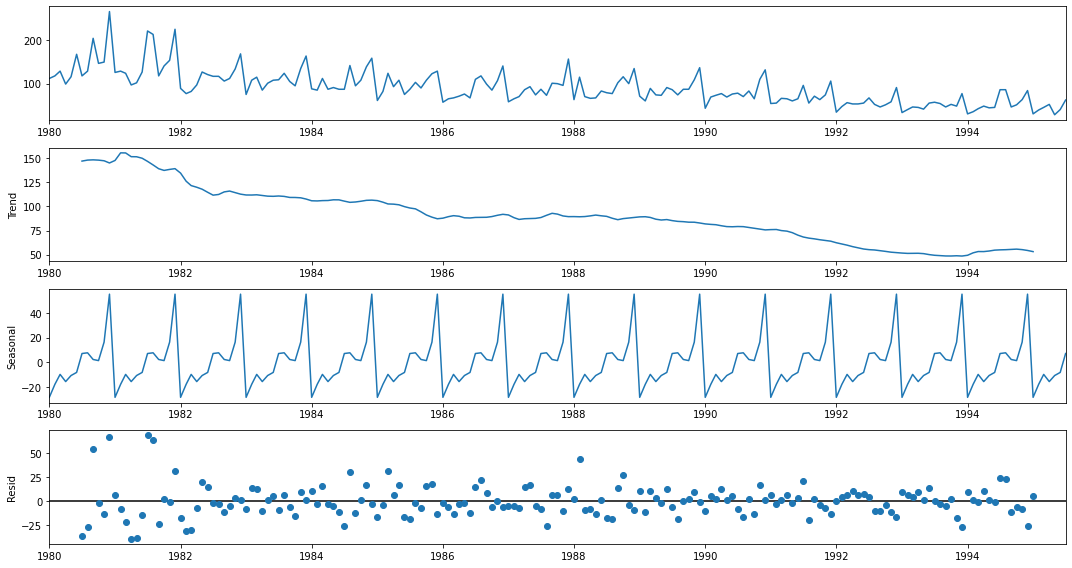


Fig 56

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

* Sparkling dataset has 187 rows and 2 columns.
* Train data is >1991 and it has 132 rows.
* Test data is <=1991 and it has 55 rows.
* Below fig 57 shows the first and last few rows of Train and Test data.
* Test and Train data plot and shown the trend in Fig 58 below.

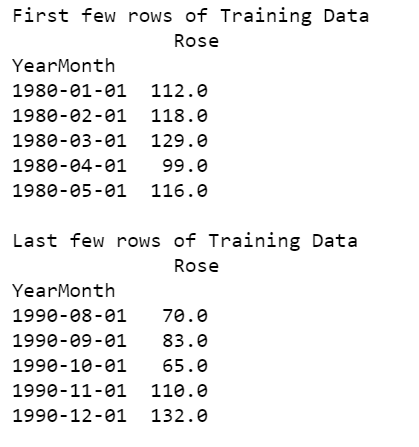
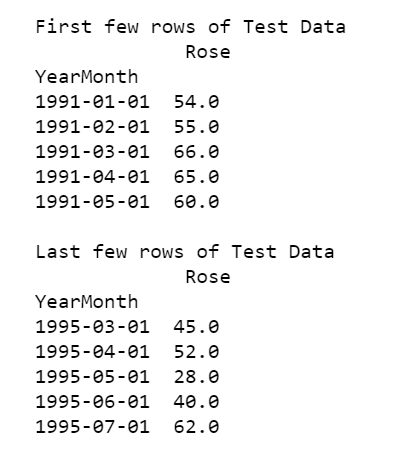
 

Fig 57

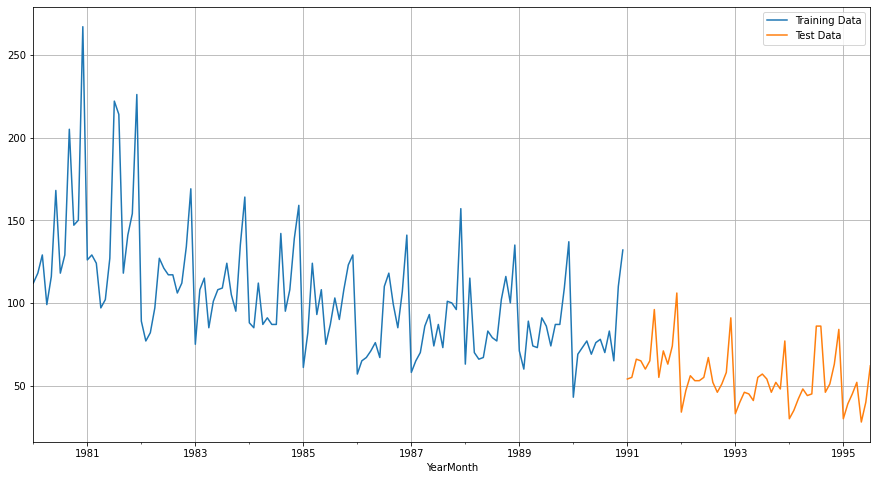


Fig 58

1. **Build various 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, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE.**
2. **Linear Regression: -**

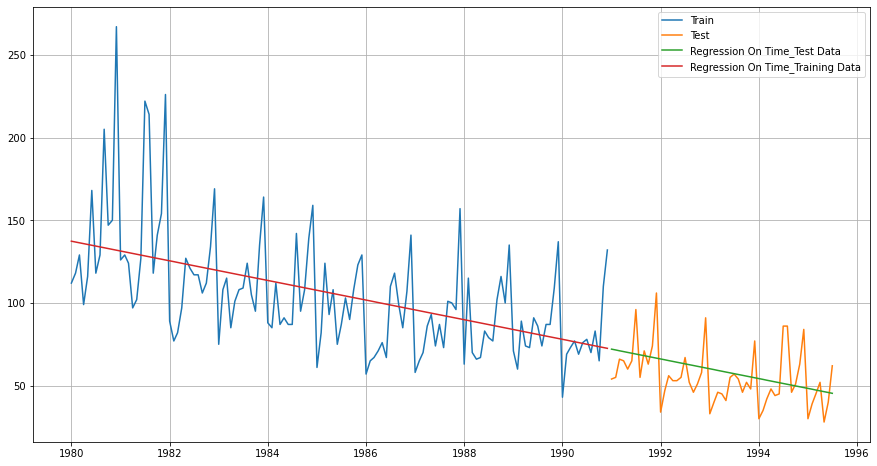


Fig 59

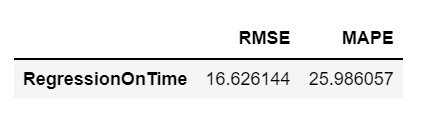


Fig 60

1. **Naive: -**

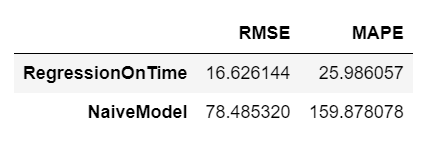


Fig 61

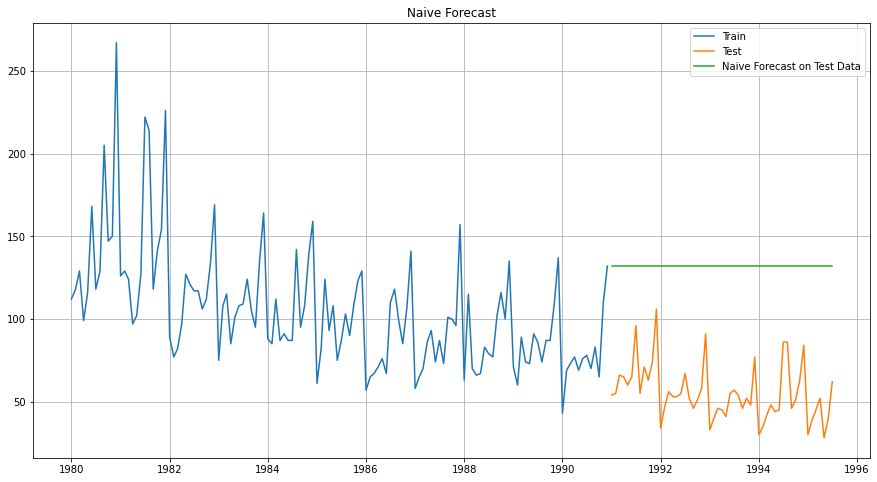


Fig 62

1. **Simple Average: -**

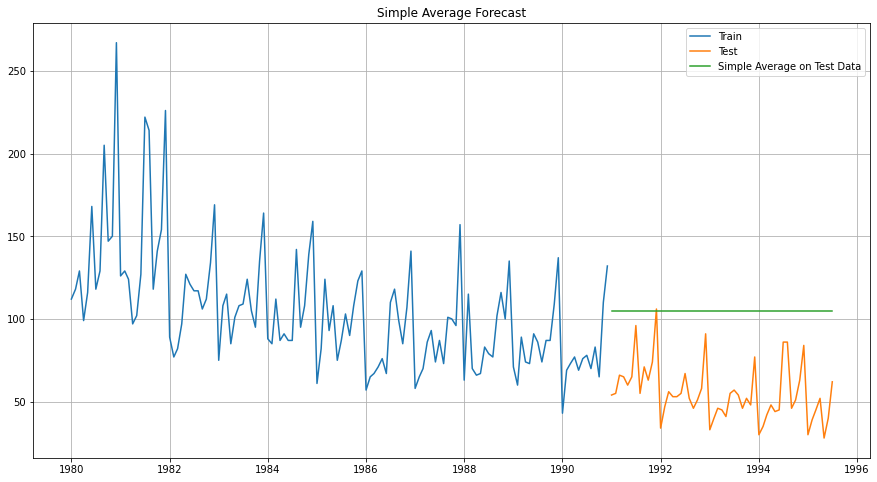


Fig 63

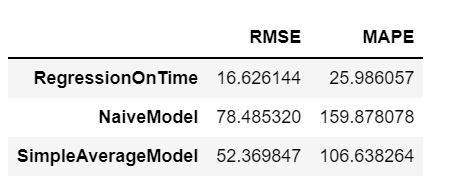


Fig 64

1. **Simple Exponential Smoothing: -**

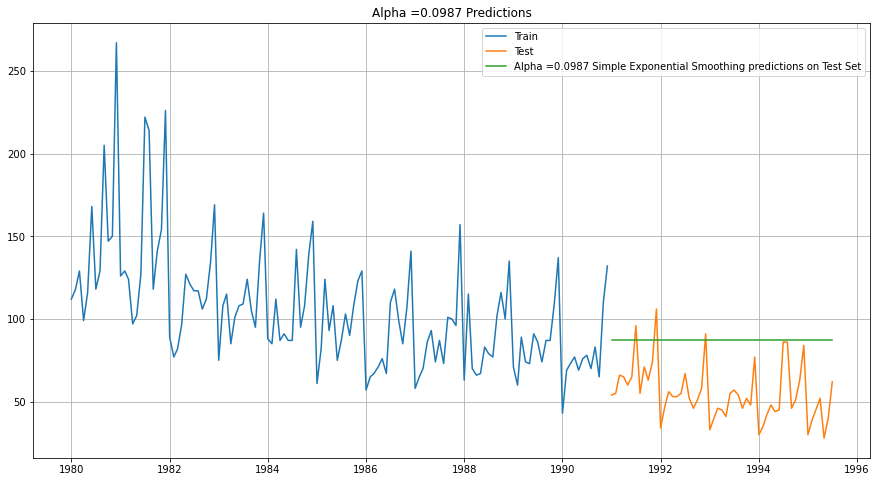


Fig 65

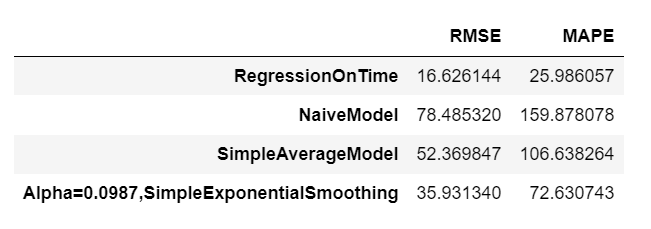


Fig 66

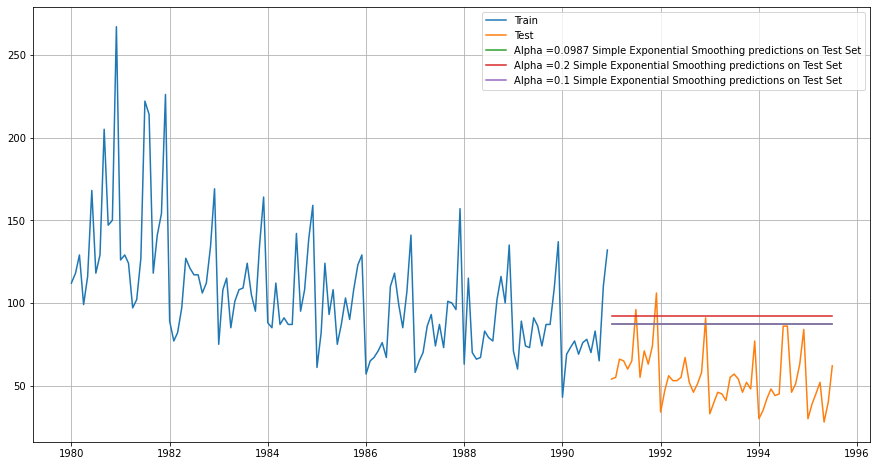


Fig 67



Fig 68

1. **Double Exponential Smoothing: -**

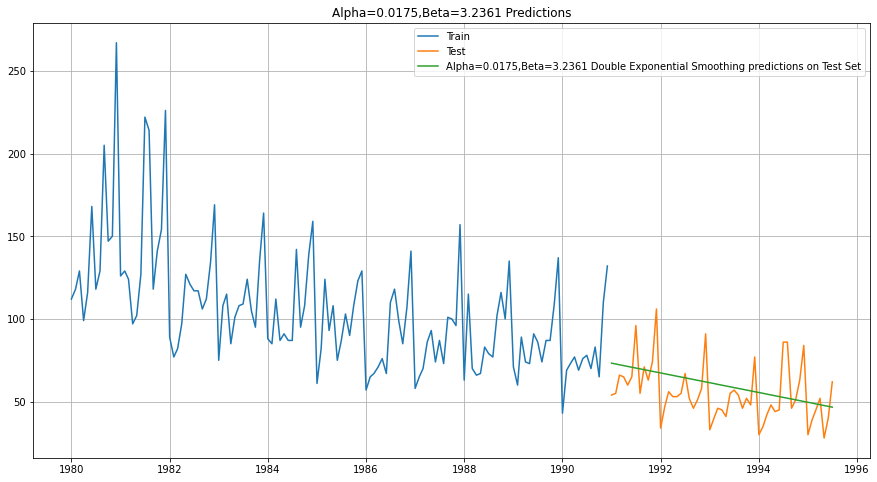


Fig 69

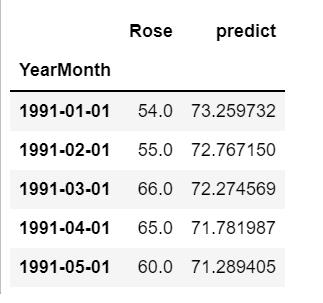


Fig 70

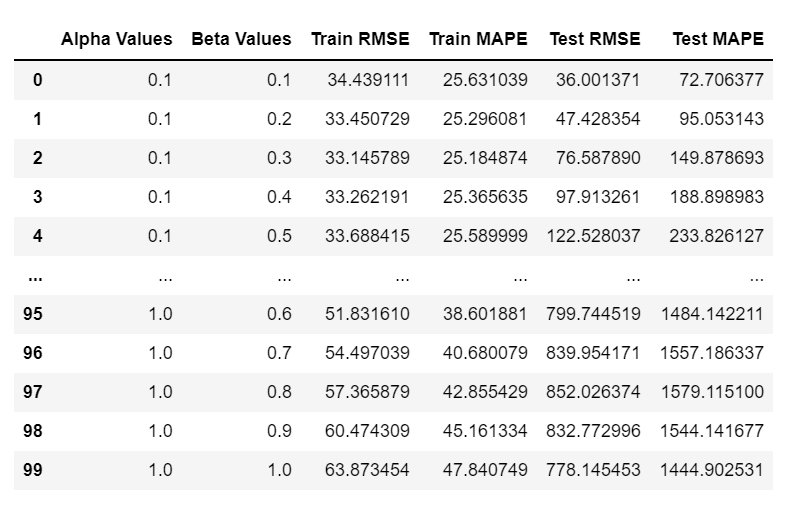


Fig 71

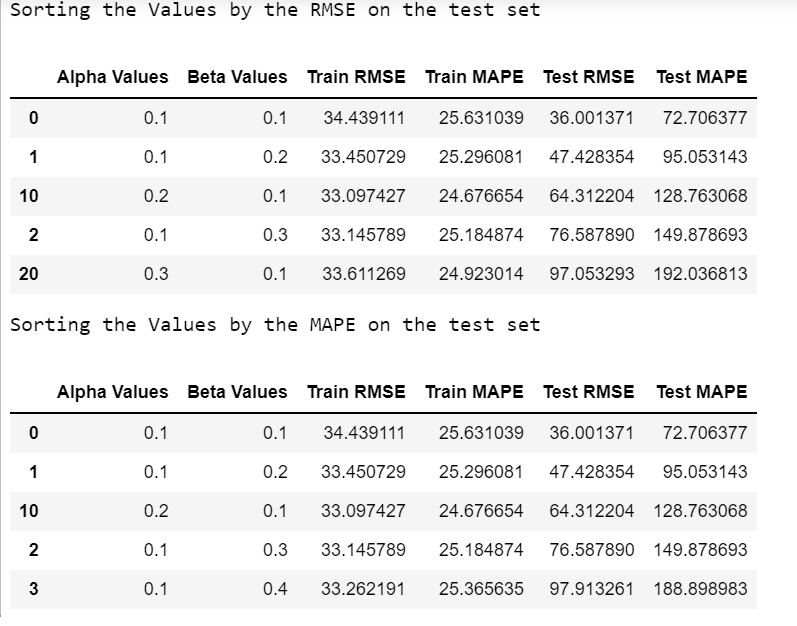


Fig 72



Fig 73

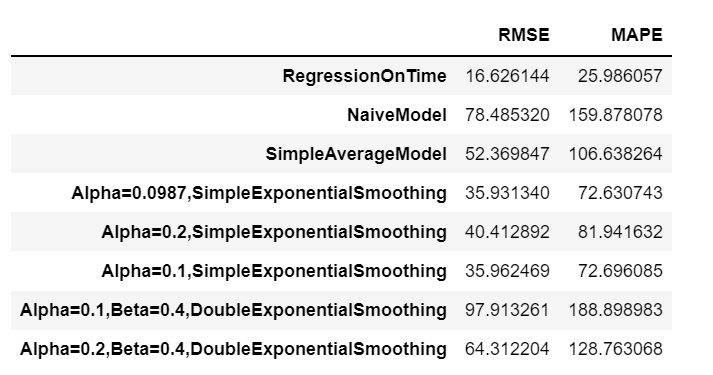


Fig 74

1. **Triple Exponential Smoothing: -**

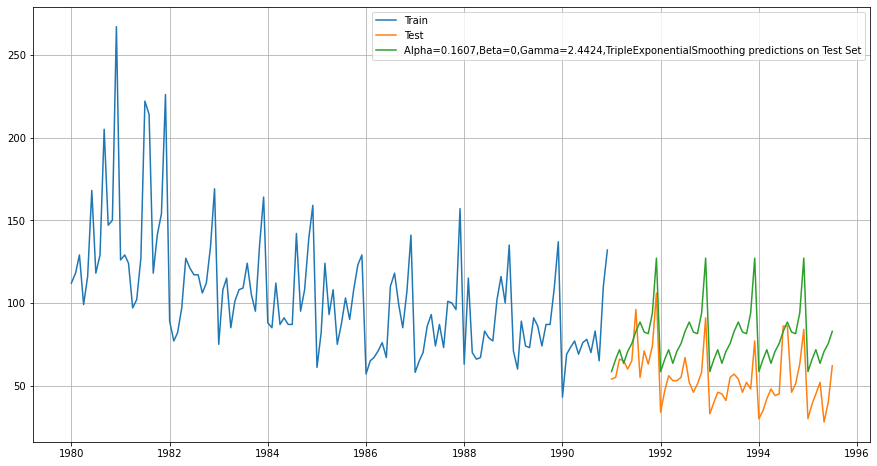


Fig 75

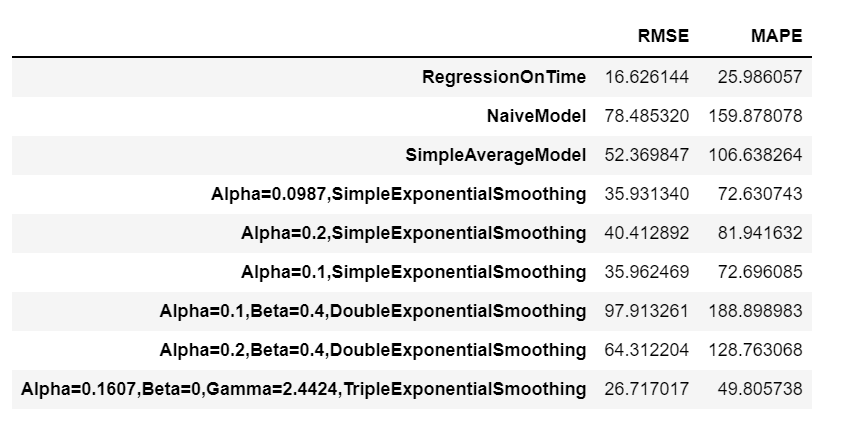


Fig 74

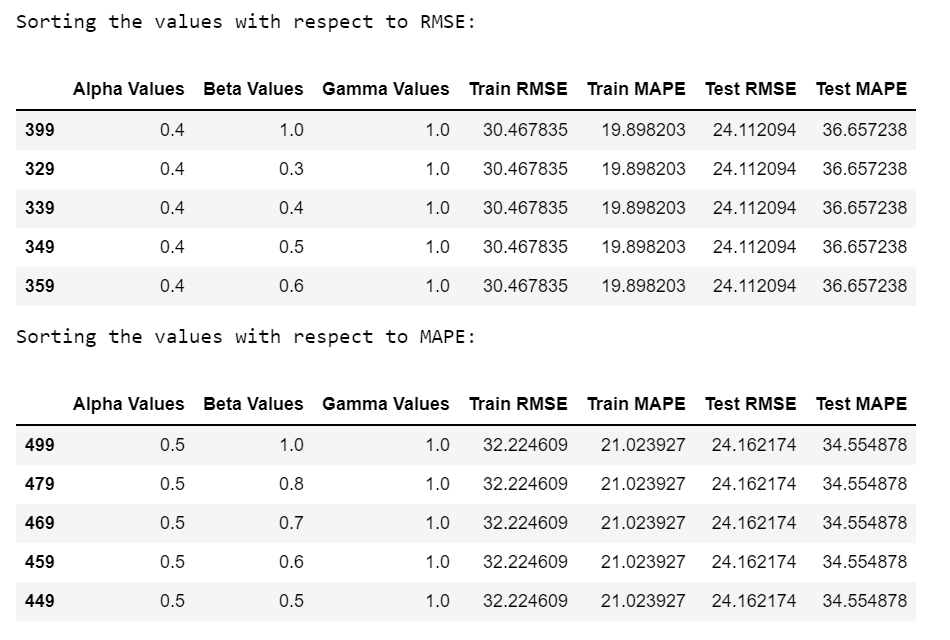


Fig 75

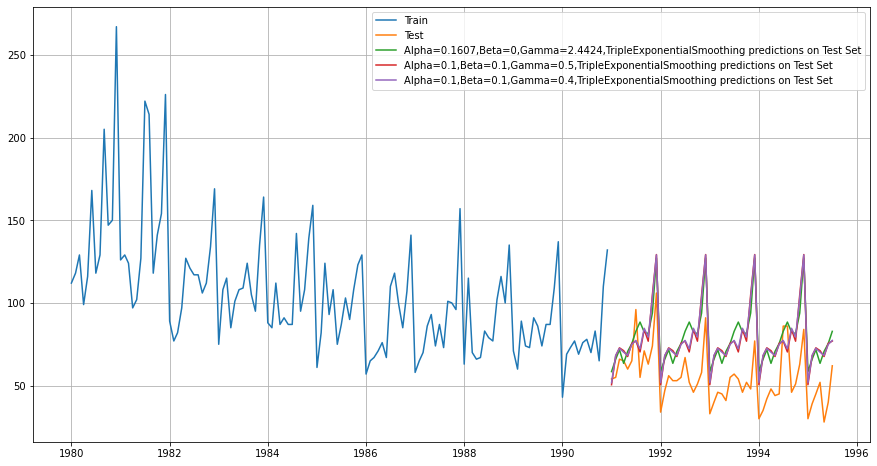


Fig 76

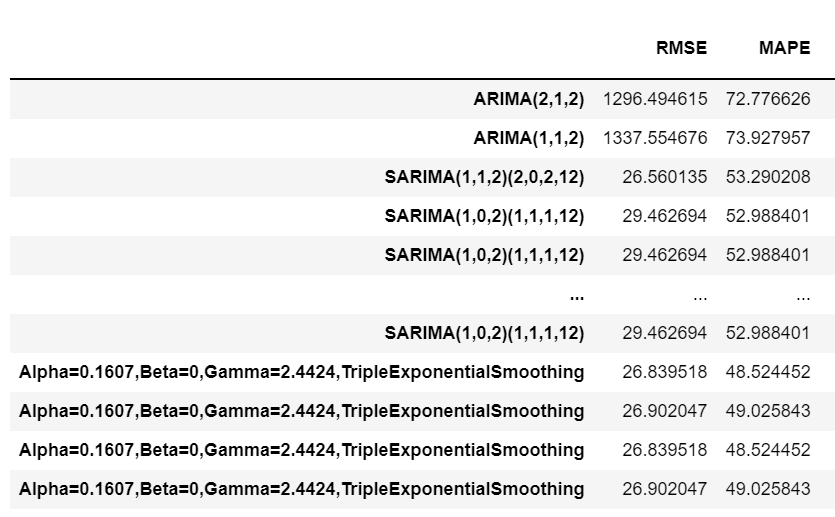


Fig 77

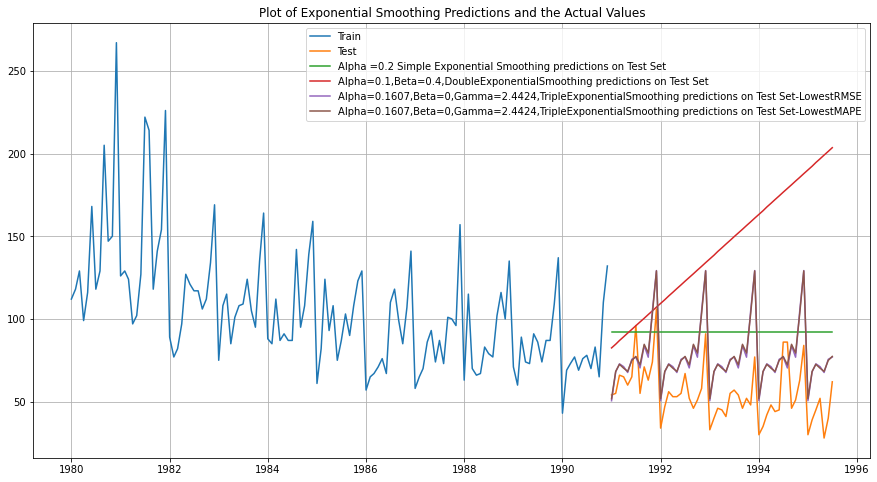


Fig 78

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 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.**

* The p Value for test set is 0.3166.

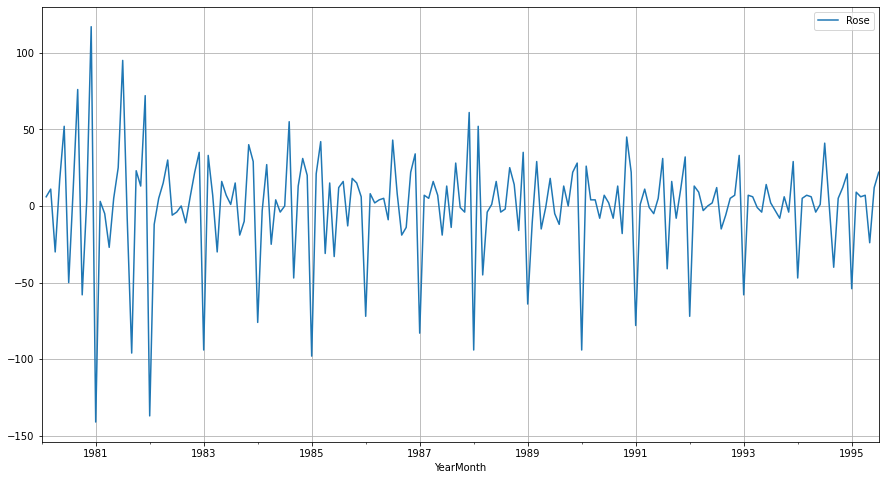
****

Fig 79

**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.**

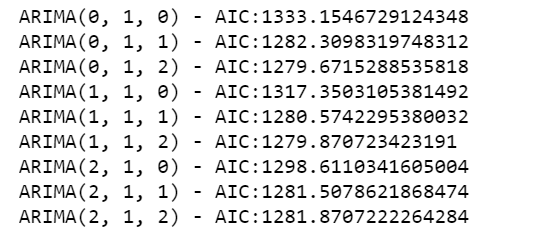
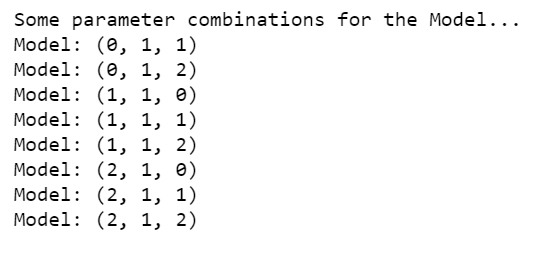
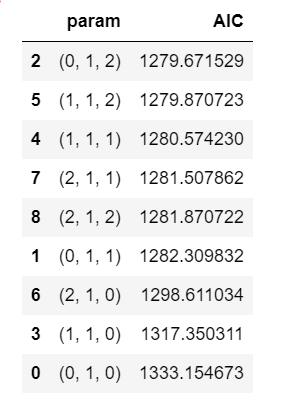


Fig 80

Fig 81

SARIMAX RESULT- (Auto ARIMA)

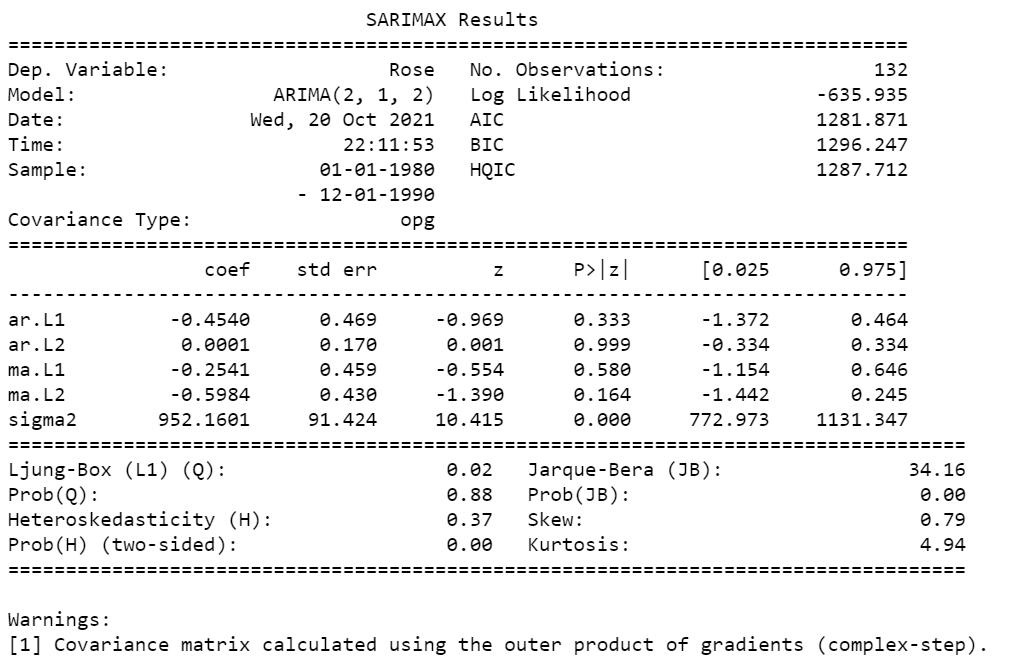


Fig 82

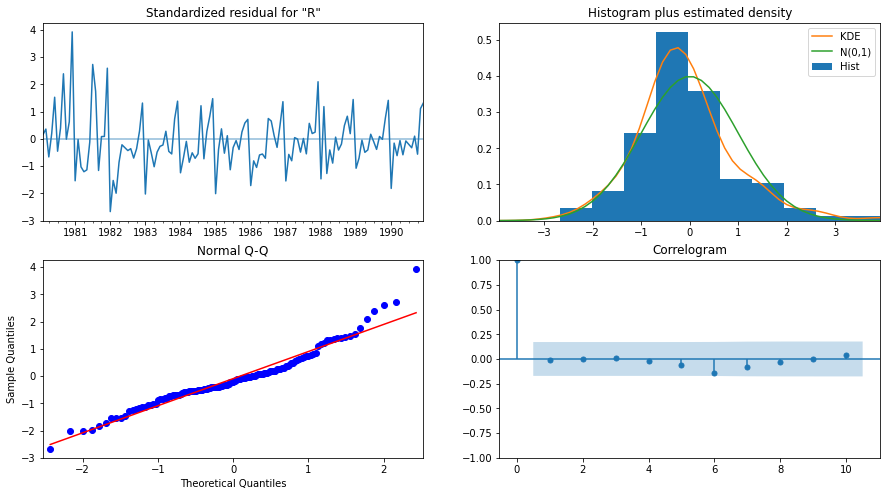


Fig 83

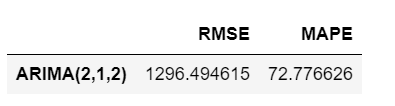


Fig 84

SARIMAX RESULT- (Manual ARIMA)

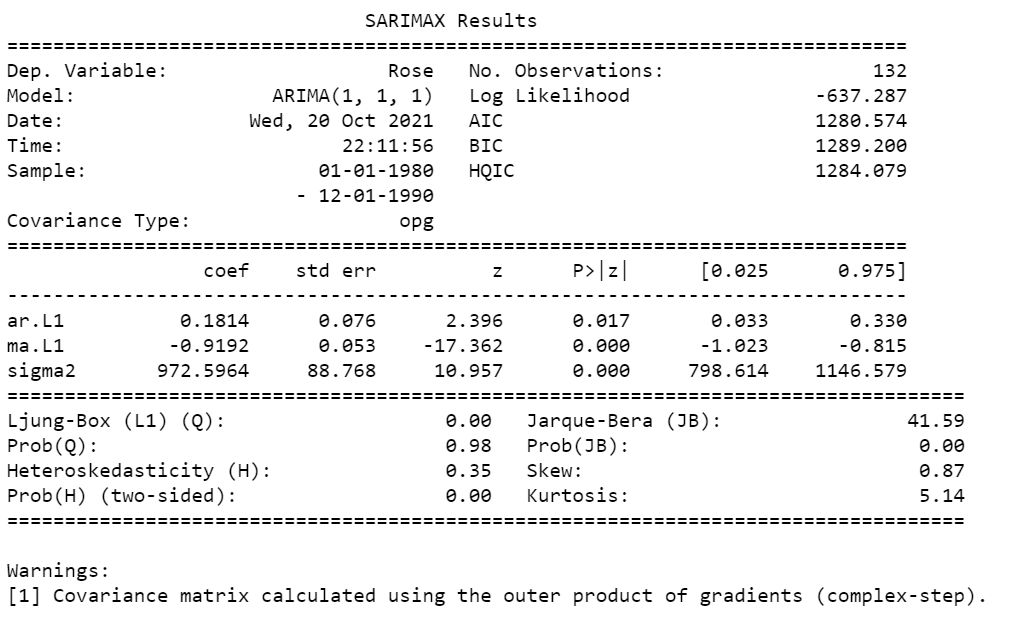
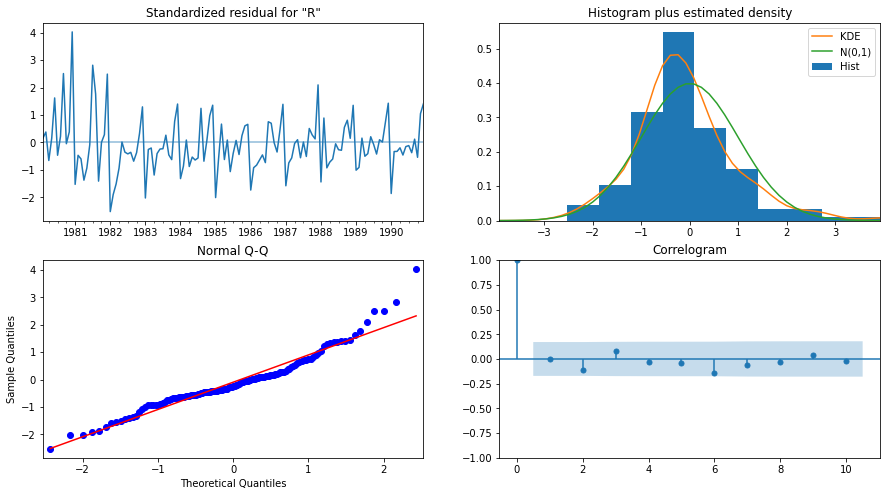


Fig 85

 Fig 86

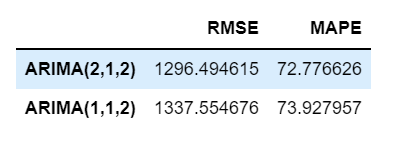


Fig 87

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 test data using RMSE.**
2. **ARIMA**

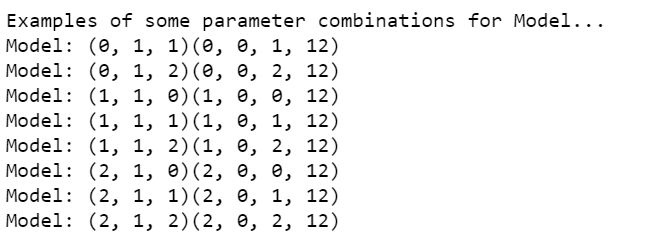
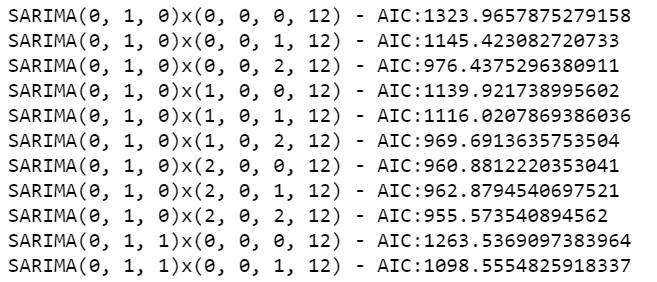
 

Fig 88

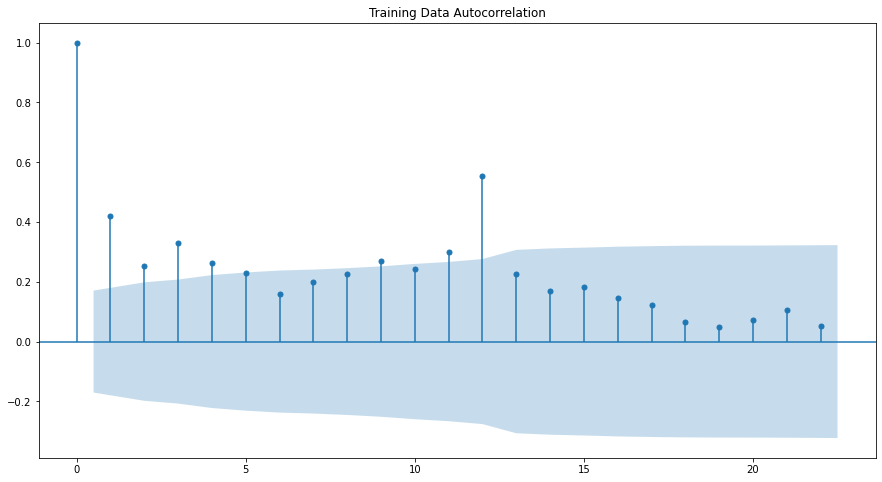


Fig 89

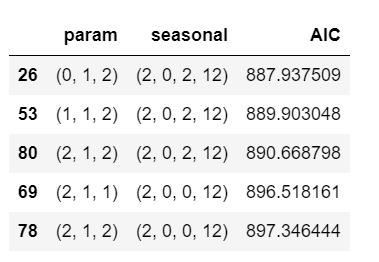


Fig 90

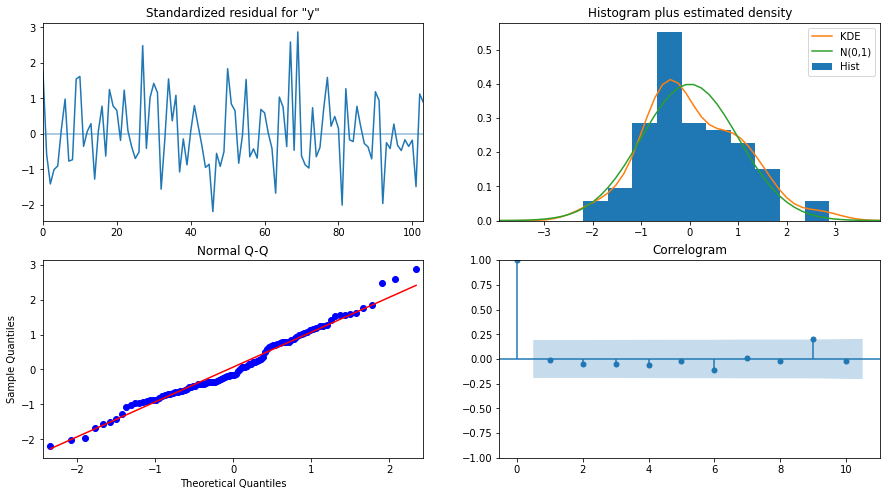


Fig 91

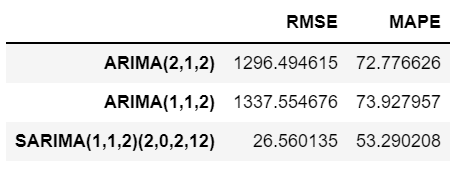
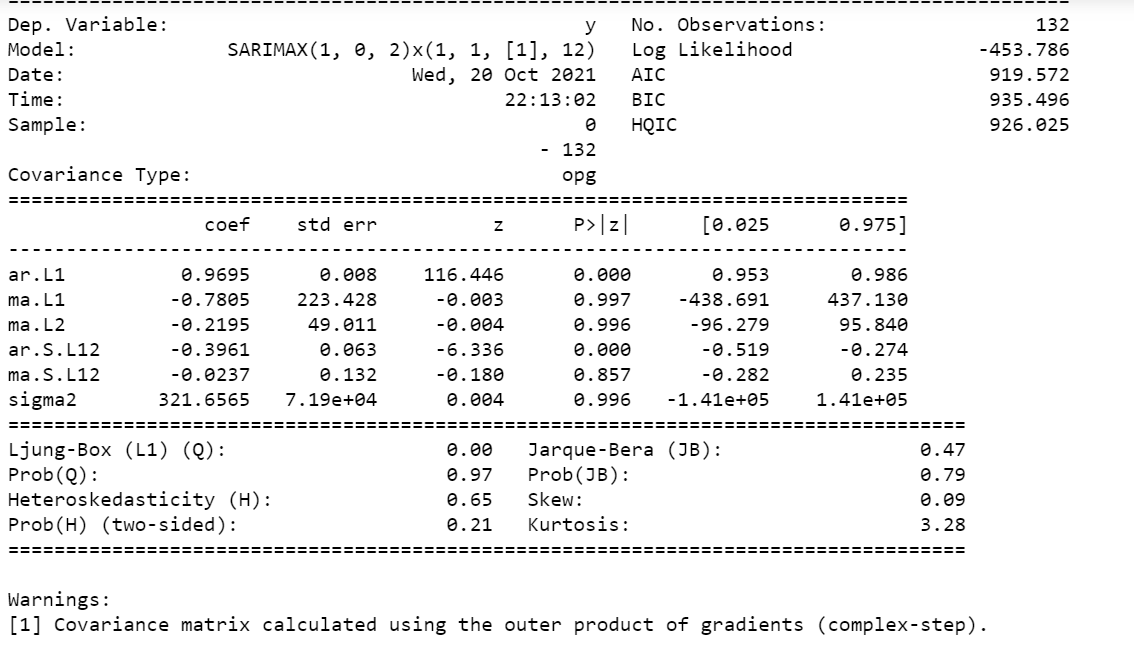


Fig 92

1. **SARIMA**



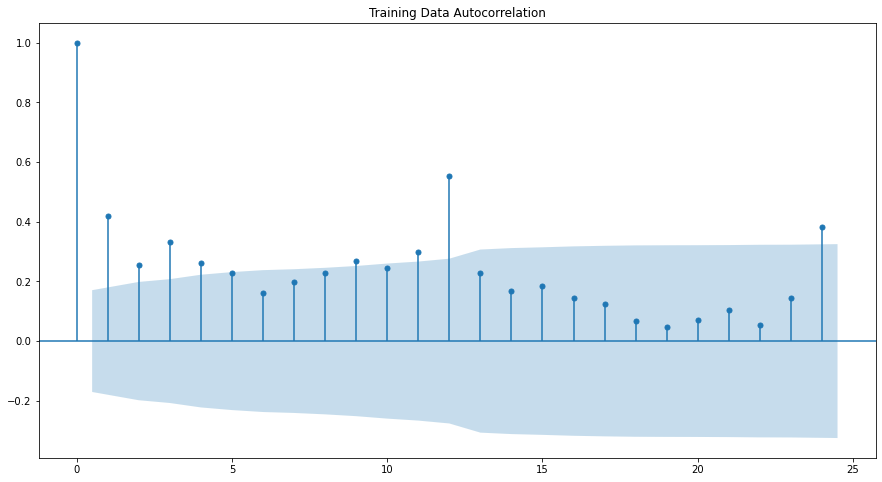


Fig 93



Fig 94

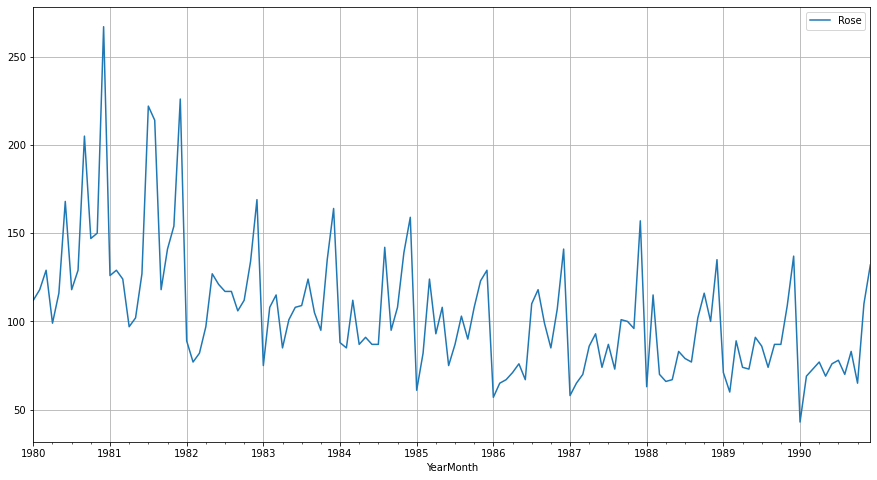


Fig 95

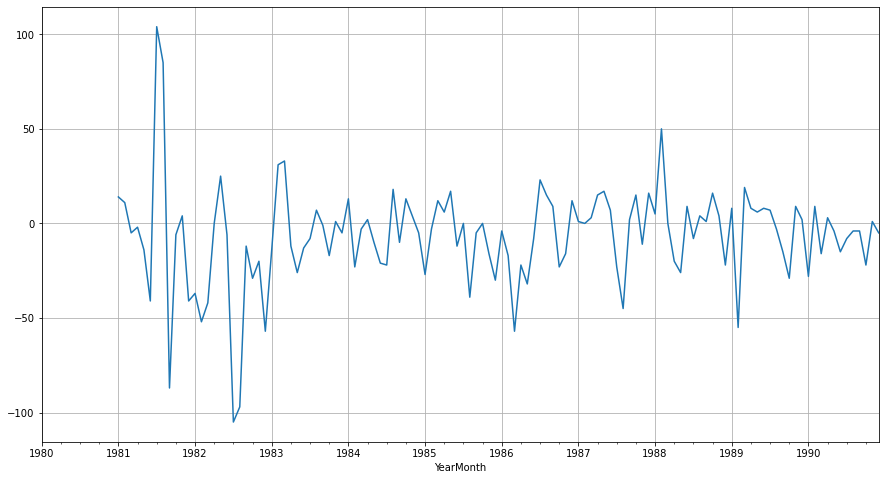
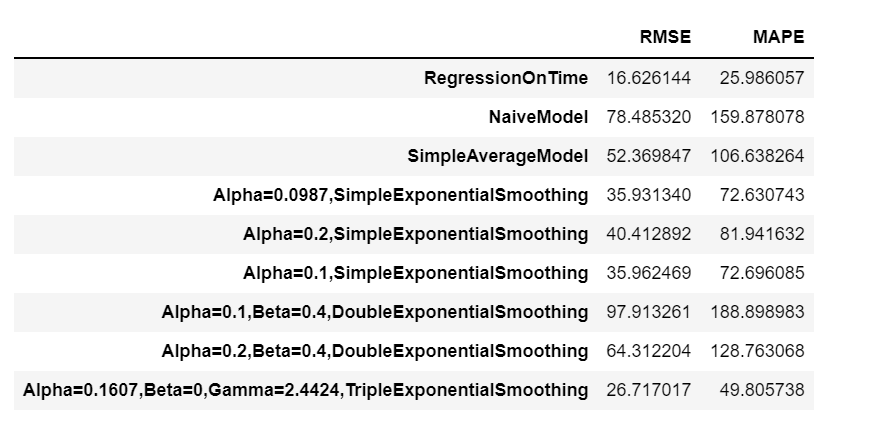


Fig 96

**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.**



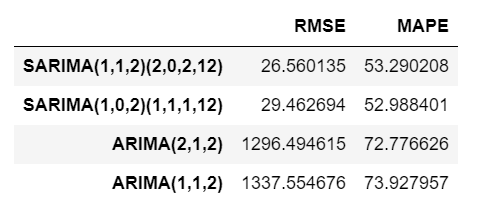


Fig 97

**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.**

* Regression ontime is the most optimal model as per the RMSE value. The prediction for 12 months not the future is mentioned below with the appropriate confidence

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

**Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.**

The sales is gradually decreasing year on year, although the spikes observed end of every year.

It is clearly visible, the prediction is average sales decreasing the proper action and promotions required to improve the sales in upcoming years.

**END**