**Project - Time Series Forecasting**

Name: Varsha Srinivasan



Table of Contents

**Problem 1-Rose…………...………………………………………………………………………………………..7**

**Problem Statement**…………………………………………………………………………………………………………….**7**

**Introduction**……………………………………………………………………………………………………………………….**7**

**Data Description**…………………………………………………………………………………………………………………**7**

* 1. Read the data as an appropriate Time Series data and plot the data.………………….7
  2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition……………………………………………………………………………………..9
  3. Split the data into training and test. The test data should start in 1991.……………..20
  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……21
  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………………….…..…………………..36
  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……………….………………………37
  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…………..……..43
  8. Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data………………………….…………………………50
  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…………………………………………………………………………………………………51
  10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.……………...............................52

Problem 2-Sparkling…………………………………………………………………………………………….55

Problem Statement……………………………………………………………………………………………………..55

Introduction………………………………………………………………………………………………………………..55

**Data Description**…………………………………………………………………………………………………………….**55**

2.1. Read the data as an appropriate Time Series data and plot the data.………………….55

2.2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition …………………………………….……………………………………………..57

2.3. Split the data into training and test. The test data should start in 1991……………..66

2.4. 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, naïveforecast 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.…………………………………………………………………………………………………………………67

2.5. 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..………………………………..…………82

2.6. 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 ……………………………………….83

2.7. 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 ….……………….89

2.8. Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data………………………………………….………….95

2.9. 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.…………………………………………………………………………………………….….96

2.10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales ………….………………………….….98

**LIST OF TABLES**

[Table 1: Sample of the Dataset1 8](#_Toc105363594)

[Table 2:TimeStamped Dataset 8](#_Toc105363595)

[Table 3: Rose-Data Type and Missing Values 10](#_Toc105363596)

[Table 4: Imputed Values 10](#_Toc105363597)

[Table 5:Rose-Yearly sales 16](#_Toc105363598)

[Table 6:Rose -Train Data 21](#_Toc105363599)

[Table 7:Rose Test Data 21](#_Toc105363600)

[Table 8:Rose-MA-Train head 35](#_Toc105363601)

[Table 9:Rose-MA-train tail 35](#_Toc105363602)

[Table 10:Rose-Automated ARIMA AIC 38](#_Toc105363603)

[Table 11:Rose-Automated SARIMA AIC 41](#_Toc105363604)

[Table 12:Rose-Models-RMSE 51](#_Toc105363605)

[Table 13: Sample of the Dataset2 55](#_Toc105363606)

[Table 15:Sparkling Train Data 67](#_Toc105363607)

[Table 16:Sparkling Test Data 67](#_Toc105363608)

[Table 17:Sparkling LR Train Data 76](#_Toc105363609)

[Table 18:Naive Forecast Value Test 78](#_Toc105363610)

[Table 19:Sparkling-Model-RMSE 96](#_Toc105363611)

**LIST OF FIGURES**

[Figure 1: Rose Sales Plot 9](#_Toc105363612)

[Figure 2:Monthly Sales across years - Rose 11](#_Toc105363613)

[Figure 3:Rose-Barplot-Monthly 11](#_Toc105363614)

[Figure 4:Rose-Boxplot-Monthly 12](#_Toc105363615)

[Figure 5:Rose-Monthplot 13](#_Toc105363616)

[Figure 6:Rose-Quarterly Sales 13](#_Toc105363617)

[Figure 7:Rose-Quarterly-Barplot 14](#_Toc105363618)

[Figure 8:Rose-Quarterly-Boxplot 15](#_Toc105363619)

[Figure 9:Rose-Barplot-Yearly 17](#_Toc105363620)

[Figure 10:Rose-Boxplot-Yearly 17](#_Toc105363621)

[Figure 11:Rose Add Decompostion 18](#_Toc105363622)

[Figure 12:Rose -Multipl Decomp 19](#_Toc105363623)

[Figure 13:Rose SES 22](#_Toc105363624)

[Figure 14:Rose-DES 23](#_Toc105363625)

[Figure 15:Rose TES add seas 24](#_Toc105363626)

[Figure 16:Rose TES multipl seas 25](#_Toc105363627)

[Figure 17:Rose Iterative SES 26](#_Toc105363628)

[Figure 18:Rose Iter DES 28](#_Toc105363629)

[Figure 19:Rose Iter TES add seas 29](#_Toc105363630)

[Figure 20:Rose Iter multiplicative seas 30](#_Toc105363631)

[Figure 21:Rose-Linear Regr Plot 32](#_Toc105363632)

[Figure 22:Rose-Naive Forecast plot 33](#_Toc105363633)

[Figure 23:Rose-Simple Average Plot 34](#_Toc105363634)

[Figure 24:Rose-Moving Average plot 36](#_Toc105363635)

[Figure 25:Rose -Automated ARIMA diagnostics 40](#_Toc105363636)

[Figure 26:Rose-Auto SARIMA ACF 41](#_Toc105363637)

[Figure 27:Rose-Automated SARIMA diagnostics 43](#_Toc105363638)

[Figure 28:Rose-Manual ARIMA ACF 44](#_Toc105363639)

[Figure 29:Rose-Manual ARIMA PACF 44](#_Toc105363640)

[Figure 30:Rose-Manual ARIMA diagnostics 45](#_Toc105363641)

[Figure 31:Rose-Manual SARIMA plot 1 46](#_Toc105363642)

[Figure 32:Rose-Manual SARIMA plot 2 46](#_Toc105363643)

[Figure 33:Rose-Manual SARIMA plot 3 47](#_Toc105363644)

[Figure 34:Rose-Manual SARIMA ACF 48](#_Toc105363645)

[Figure 35:Rose-Manual SARIMA PACF 48](#_Toc105363646)

[Figure 36:Rose-Manual SARIMA diagnostics 50](#_Toc105363647)

[Figure 37:Rose-12 months Forecast 52](#_Toc105363648)

[Figure 38:Sparkling-Timestamped data 56](#_Toc105363649)

[Figure 39:Sales of Sparkling plot 56](#_Toc105363650)

[Figure 40:Sparkling description 57](#_Toc105363651)

[Figure 41:Sparkling Monthly sales over years 58](#_Toc105363652)

[Figure 42:Sparkling Barplot Monthly sales 58](#_Toc105363653)

[Figure 43:Sparkling Boxplot Monthly sales 59](#_Toc105363654)

[Figure 44:Sparkling Monthplot 60](#_Toc105363655)

[Figure 45:Sparkling Quarterly sales over years 60](#_Toc105363656)

[Figure 46:Sparkling Barplot Quarterly 61](#_Toc105363657)

[Figure 47:Sparkling Quarterly Boxplot 61](#_Toc105363658)

[Figure 48:Sparkling Yearly Sales 62](#_Toc105363659)

[Figure 49:Sparkling Barplot Yearly 63](#_Toc105363660)

[Figure 50:Sparkling Boxplot Yearly 63](#_Toc105363661)

[Figure 51:Sparkling Additive Decomp 64](#_Toc105363662)

[Figure 52:Sparkling Mulitpl Decomp 65](#_Toc105363663)

[Figure 53:Sparkling SES plot 68](#_Toc105363664)

[Figure 54:Sparkling DES plot 69](#_Toc105363665)

[Figure 55:Sparkling TES add seas plot 70](#_Toc105363666)

[Figure 56:Sparkling TES multipl seas plot 71](#_Toc105363667)

[Figure 57:Sparkling Iterative SES 72](#_Toc105363668)

[Figure 58:Sparkling Iter DES plot 73](#_Toc105363669)

[Figure 59:Sparkling-Iter TES add seas 74](#_Toc105363670)

[Figure 60:Sparkling Iterative TES multipl seas plot 75](#_Toc105363671)

[Figure 61:Sparkling-Linear Regr plot 77](#_Toc105363672)

[Figure 62:Sparkling-Naive Forecast plot 79](#_Toc105363673)

[Figure 63:Sparkling-SA plot 80](#_Toc105363674)

[Figure 64:Sparkling-MA-Train head 81](#_Toc105363675)

[Figure 65:Sparkling-MA-Train tail 81](#_Toc105363676)

[Figure 66:Sparkling-Mov Avg-plot 82](#_Toc105363677)

[Figure 67:Sparkling Automated ARIMA diagnostics 86](#_Toc105363678)

[Figure 68:Sparkling-Automated SARIMA-ACF 87](#_Toc105363679)

[Figure 69:Sparkling-Automated SARIMA diagnostics 89](#_Toc105363680)

[Figure 70:Sparkling-Manual ARIMA ACF 90](#_Toc105363681)

[Figure 71:Sparkling-Manual ARIMA-PACF 90](#_Toc105363682)

[Figure 72:Sparkling-Manual ARIMA diagnostics 91](#_Toc105363683)

[Figure 73:Sparkling-Manual SARIMA plot 1 92](#_Toc105363684)

[Figure 74:Sparkling-Manual SARIMA plot 2 92](#_Toc105363685)

[Figure 75:Sparkling-Manual SARIMA ACF 93](#_Toc105363686)

[Figure 76:Sparkling-Manual SARIMA PACF 93](#_Toc105363687)

[Figure 77:Sparkling-Manual SARIMA diagnostics 95](#_Toc105363688)

[Figure 78:Sparkling Optimum Model Plot 98](#_Toc105363689)

**PROBLEM 1 - Rose**

**Problem Statement**

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 Rose Wine Sales in the 20th century.

**Introduction**

The purpose of this whole exercise is to perform exploratory data analysis and perform Time Series Forecasting using Exponential Smoothing models, Regression, Naïve Forecast models, Simple Average models, Moving Average models, ARIMA and SARIMA models(using cut-off points of AIC,ACF and PACF plots) to forecast the sales of Rose wine.

**Data Description**

1. YearMonth: The Year and the Month on which its corresponding units of Rose Wine is sold.
2. Rose: Units of Rose Wine sold.
   1. **Read the data as an appropriate Time Series data and plot the data.**

**Sample of the dataset:**

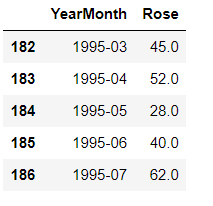
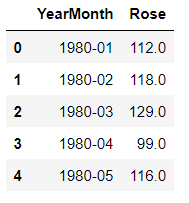


Table : Sample of the Dataset1

The data is read from the excel file and the above tables shows the first and last 5 rows of the dataset.There are 187 rows in the dataframe. The Rose is the variable to be forecasted . YearMonth denotes the year and month values ranging from Jan 1980 to July 1995.

There are no duplicates in the datatset.

There are 0 duplicates in the dataset

The initial datatype of the columns before indexing are:

YearMonth object

Rose float64

dtype: object

YearMonth column is converted into a Time Stamp index using to\_datetime

function and YearMonth is dropped.

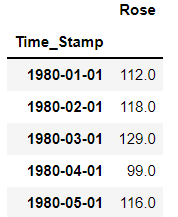


Table :TimeStamped Dataset

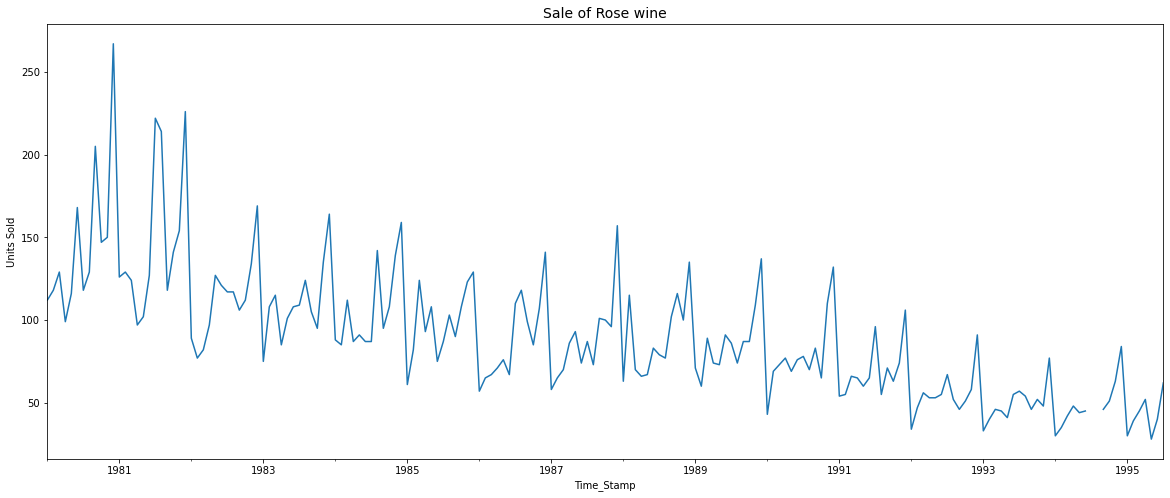


Figure : Rose Sales Plot

From the pot we can see a downwards trend and seasonality too. We can see a disconnect in the graph in 1994 . This might be because of missing values.

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

There are 0 duplicates in the dataset.

**Data Type and Missing Values**

The initial datatype of the columns after indexing are:

Rose float64

dtype: object

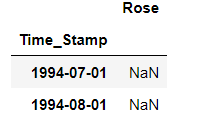
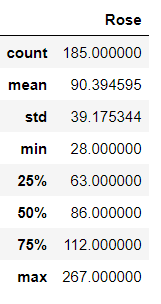


Table : Rose-Data Type and Missing Values

There are 2 null values which can also be inferred from the output of isnull function.

It is imputed with interpolation function with linear method. The values after interpolation are:

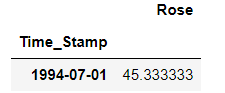
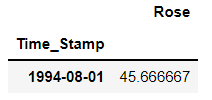
 

Table : Imputed Values

Both values are taken as 45 as the decimals doesn’t give any information when calculating units sold.

**Monthly Sales**

**Monthly Sales across years**

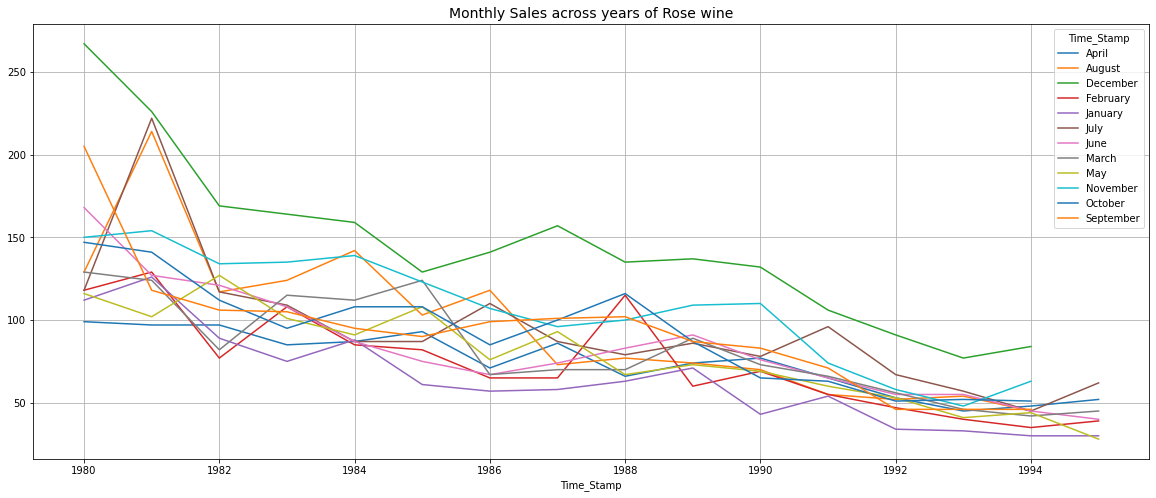
****

Figure :Monthly Sales across years - Rose

From the plot we can infer that December month has the highest sales across all years and January month has the lowest sales across most of the years.

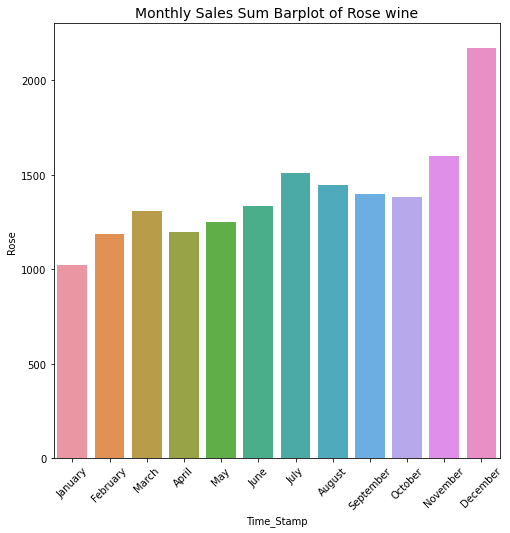
**Monthly Sales Sum Barplot of Rose wine**

Figure :Rose-Barplot-Monthly

December has highest sales combining all the years going above 2000 units sold while January has lowest sales combined with almost 1000 units sold.

Barplot is plotted using sum of the month values from all the years.

**Boxplot of Monthly Sales of Rose wine**

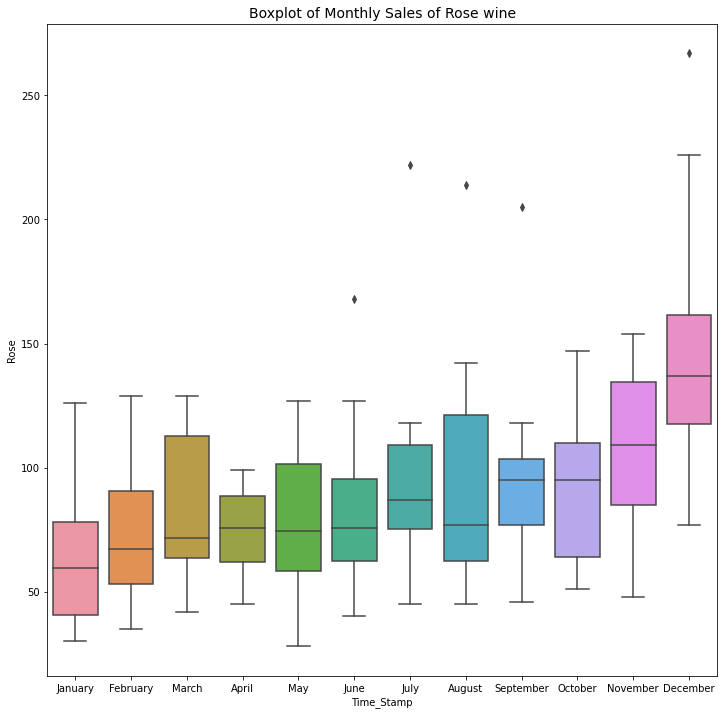


Figure :Rose-Boxplot-Monthly

We can see from the boxplot that there is an increasing trend in the sales from the increasing median in the subsequent months.

**Month Plot**

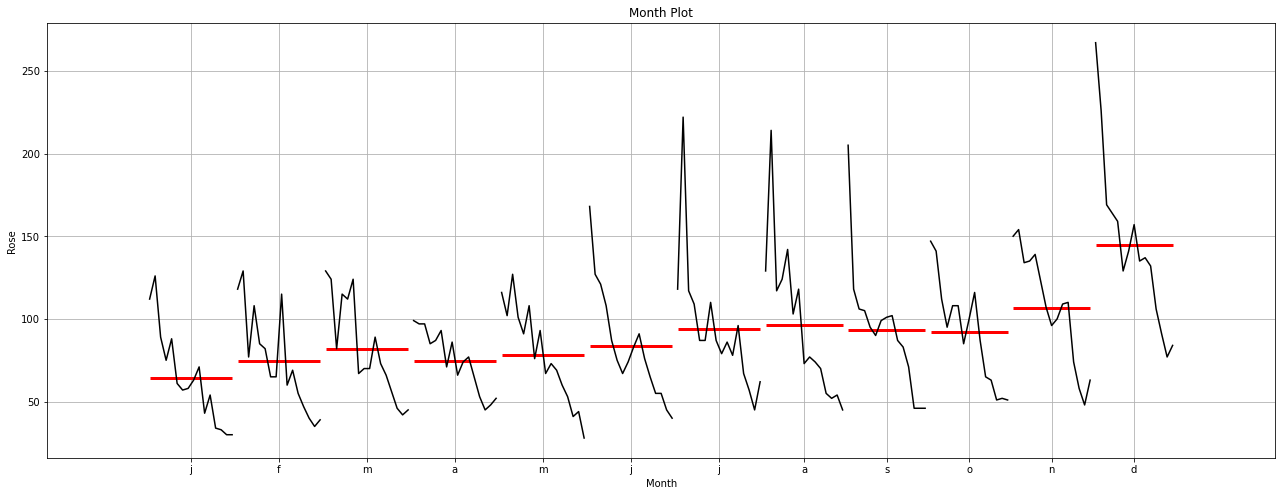


Figure :Rose-Monthplot

Month wise there is an increasing trend but year wise there is an decreasing trend in the units sold.

**Quarterly Sales**

**Quarterly Sales across years of Rose wine**

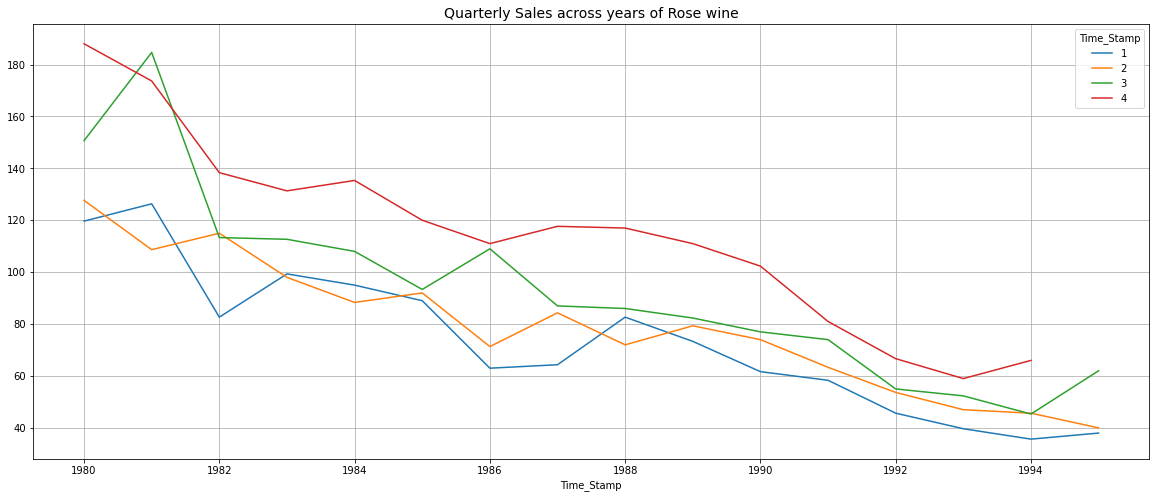
****

Figure :Rose-Quarterly Sales

Fourth quarter has the highest sales across all the years denoted by the red line. First quarter has the lowest sales across most of the years denoted by the blue line.

**Barplot of Quarterly Sales Sum of Rose wine**

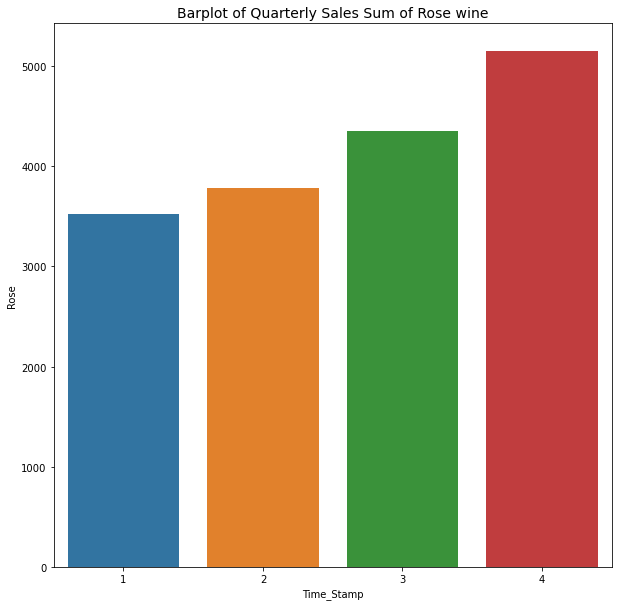
****

Figure :Rose-Quarterly-Barplot

Fourth quarter has the highest sum of sales across all the years with above 5000 units sold. First quarter has the lowest sum of sales across most of the years with almost 3500 units sold.

**Boxplot of Quarterly Sales of Rose wine**

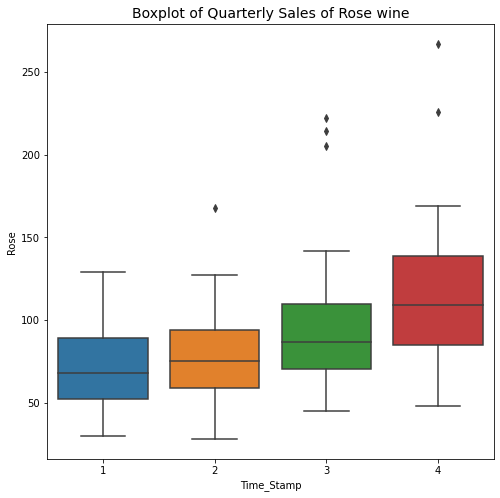
****

Figure :Rose-Quarterly-Boxplot

There is an increasing(upward) trend which we can see from the median values of individual boxplot with subsequent quarters.

**Yearly Sales**

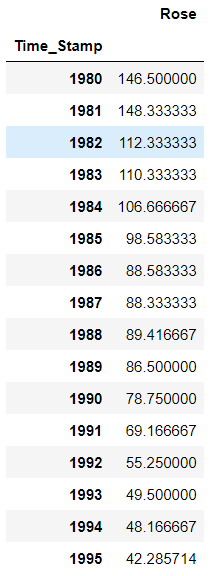


Table :Rose-Yearly sales

This table gives the mean value of the units of wine sold for each year. There is a decreasing trend in the sales yearwise.

**Barplot of Yearly Sales Sum of Rose wine**

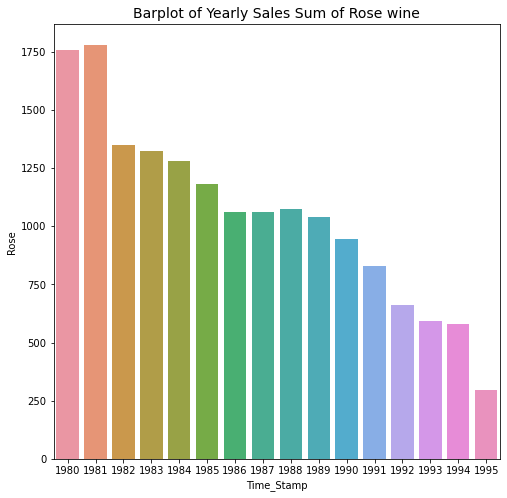
****

Figure :Rose-Barplot-Yearly

1981 has the highest sales amongst all years. 1995 has the lowest sales amongst all years.

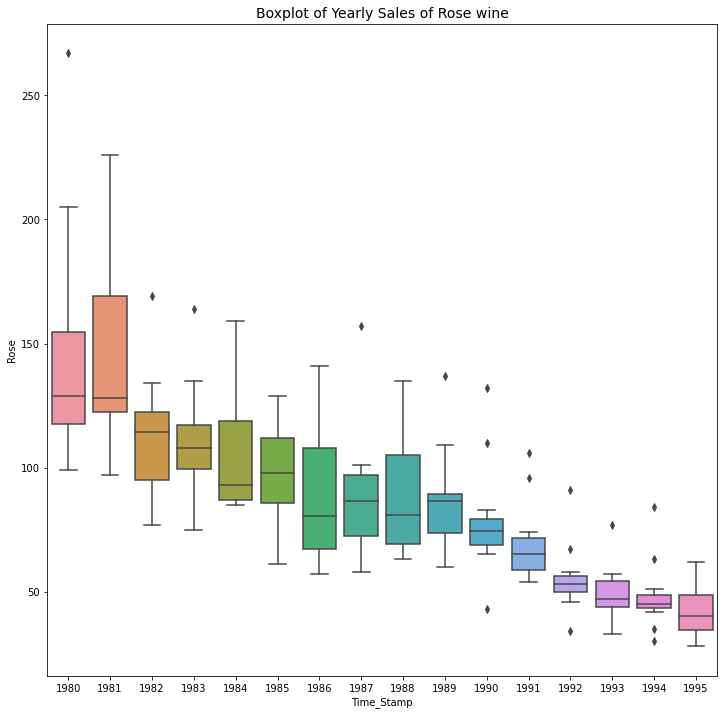
**Boxplot of Yearly Sales of Rose wine**

Figure :Rose-Boxplot-Yearly

There is a decreasing trend year-wise as shown by the medians of boxplots. The maximum sale ever happened is in 1980 which is above 250 units.

**Decomposition**

**Additive Decomposition of Rose wine**

Additive Decomposition of Rose wine

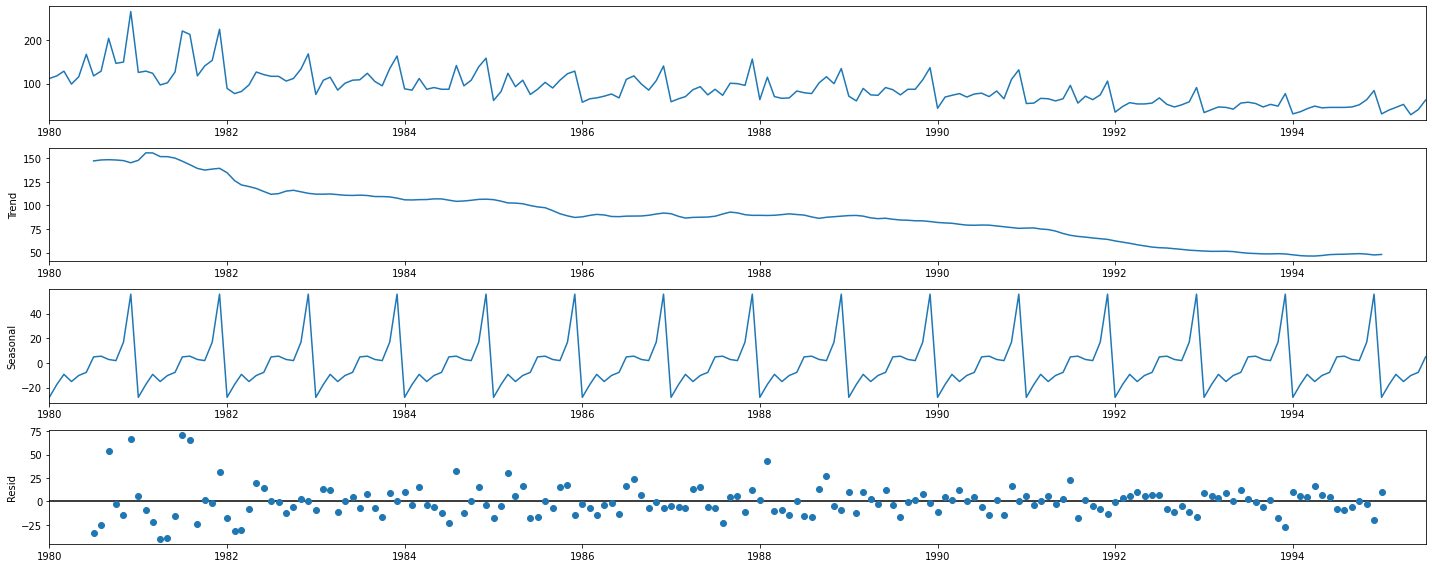
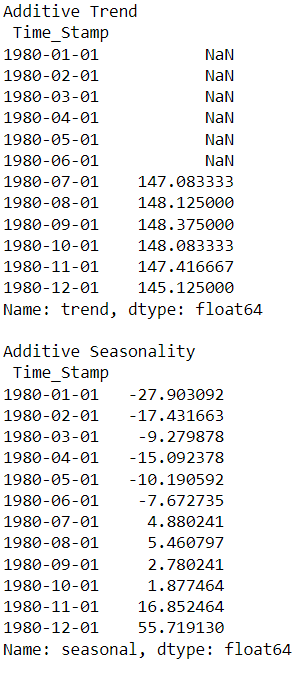
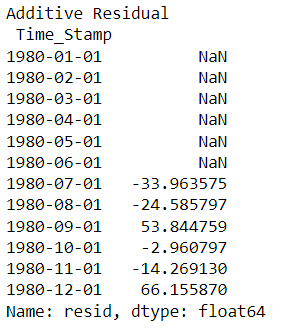
****

Figure :Rose Add Decompostion

Residuals have a pattern and it is around 0. So additive decomposition is suitable for the data.

The trend,seasonality and residual for 12 months

**Multiplicative Decomposition of Rose wine**

Multiplicative Decomposition of Rose wine

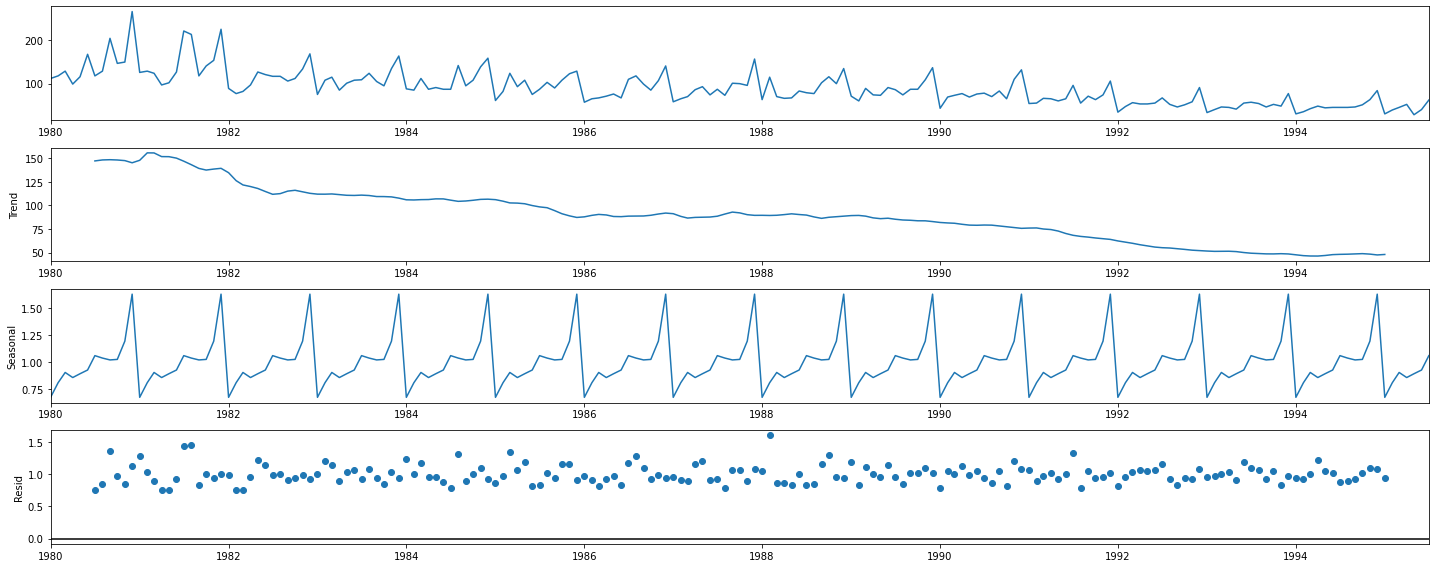
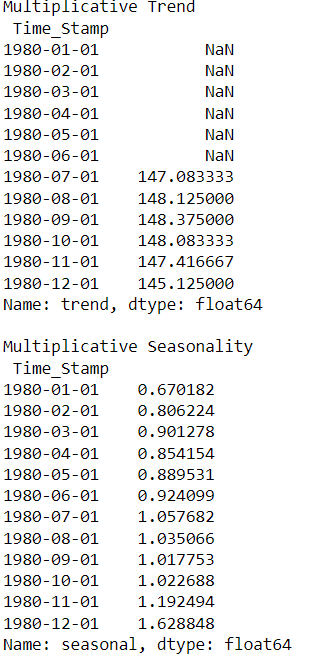
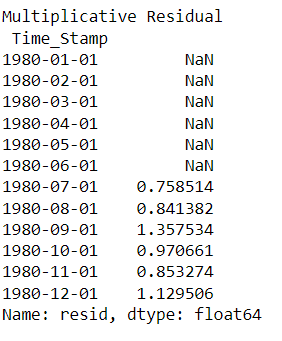
****

Figure :Rose -Multipl Decomp

Residuals have a pattern and it is around 1. So additive decomposition is not suitable for the data.

The trend ,seasonality and residuals for first 12 months is.

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

The first 5 and last 5 rows of train data shows that it starts from January 1980 and ends at December 1990.

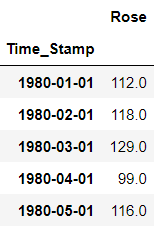
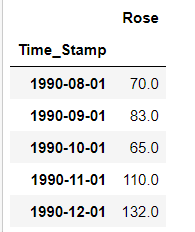
 

Table :Rose -Train Data

The first 5 and last 5 rows of test data shows that it starts from January 1991 and ends at July 1995.

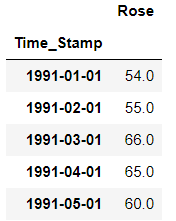
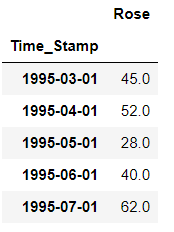
 

Table :Rose Test Data

**1.4. 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.**

**Simple Exponential Smoothing Auto Fit Model**

Simple Exponential Smoothing(SES) is a time series forecasting method for univariate data without a trend or seasonality. It requires a single parameter, called alpha (a), also called the smoothing factor or smoothing coefficient.

A SES model is built on train data with initialization\_method value as estimated and the following parameters with values.

initialization\_method - Method for initialize the recursions.

Use\_brute=True -> Search for good starting values using a brute force (grid) optimizer

**Optimized**=True -> Estimate model parameters by maximizing the log-likelihood.

{'smoothing\_level': 0.09874983698117956,

'smoothing\_trend': nan,

'smoothing\_seasonal': nan,

'damping\_trend': nan,

'initial\_level': 134.38702481818487,

'initial\_trend': nan,

'initial\_seasons': array([], dtype=float64),

'use\_boxcox': False,

'lamda': None,

'remove\_bias': False}

The model built is used to forecast for next 55 months which is the length of

test data.

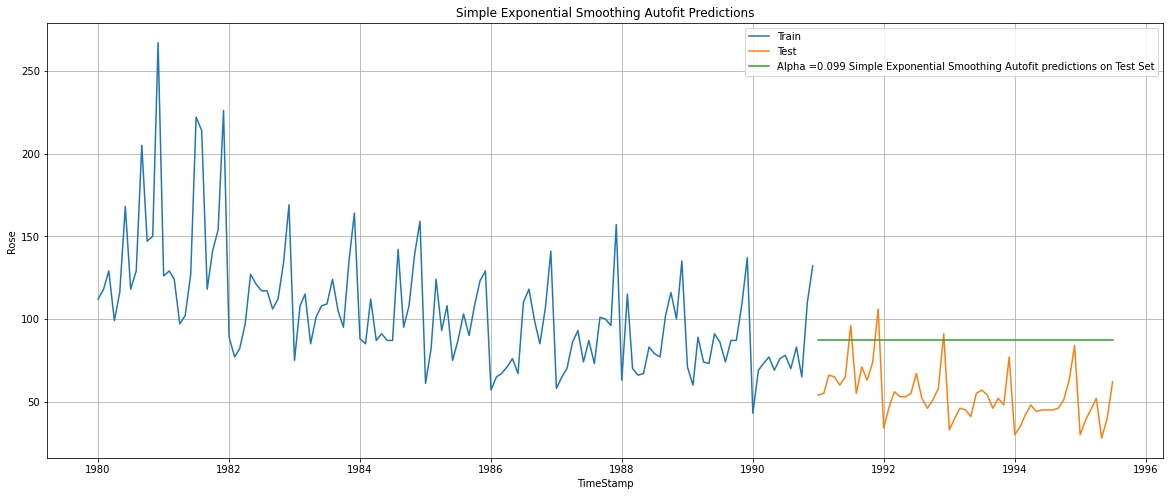


Figure :Rose SES

RMSE is calculated on test data.

SES Autofit RMSE: 36.82

**Double Exponential Smoothing - Holt Autofit Model**

Double exponential smoothing(DES) employs a level component and a trend component at each period

A DES model is built on train data with initialization\_method value as estimated , use\_brute=True and Optimized=True. The other parameters and the values are:

{'smoothing\_level': 1.4901161193847656e-08, 'smoothing\_trend': 1.6610391146660035e-10, 'smoothing\_seasonal': nan, 'damping\_trend': nan, 'initial\_level': 137.81553690867275, 'initial\_trend': -0.4943781897068274, 'initial\_seasons': array([], dtype=float64), 'use\_boxcox': False, 'lamda': None, 'remove\_bias': False}

The model built is used to forecast for next 55 months which is the length of

test data.

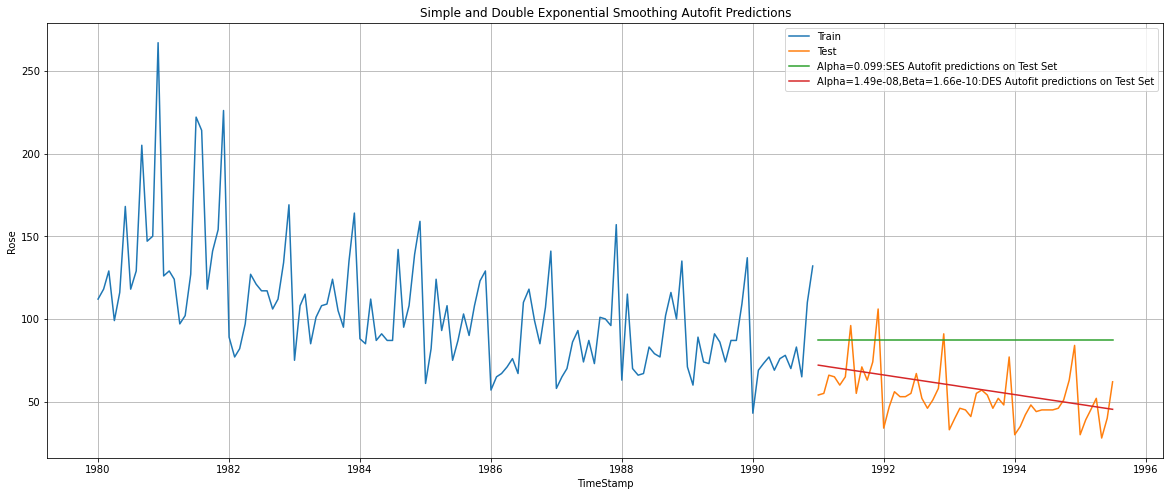
****

Figure :Rose-DES

We see that the double exponential smoothing is picking up the trend component along with the level component as well

RMSE is calculated on test data.

DES Autofit RMSE: 15.28

**Triple Exponential Smoothing - ETS(A, A, A) - Holt Winter's linear method with additive errors Autofit Model**

Triple exponential smoothing(TES) is used to handle the time series data containing a seasonal component. This method is based on three smoothing equations: stationary component, trend, and seasonal. Both seasonal and trend can be additive or multiplicative

In this model both trend and seasonality is chosen as additive. A DES model is built on train data with initialization\_method value as estimated , use\_brute=True and Optimized=True. The other parameters and the values are:

{'smoothing\_level': 0.08954054664605082, 'smoothing\_trend': 0.0002400108693915795, 'smoothing\_seasonal': 0.003466872515750747, 'damping\_trend': nan, 'initial\_level': 146.5570157826235, 'initial\_trend': -0.547196983509005, 'initial\_seasons': array([-31.17478463, -18.74839869, -10.76961776, -21.36741017,

-12.63775539, -7.27430333, 2.61279801, 8.69603625,

4.79381122, 2.96110122, 21.05738849, 63.18279918]), 'use\_boxcox': False, 'lamda': None, 'remove\_bias': False}

The model built is used to forecast for next 55 months which is the length of

test data.

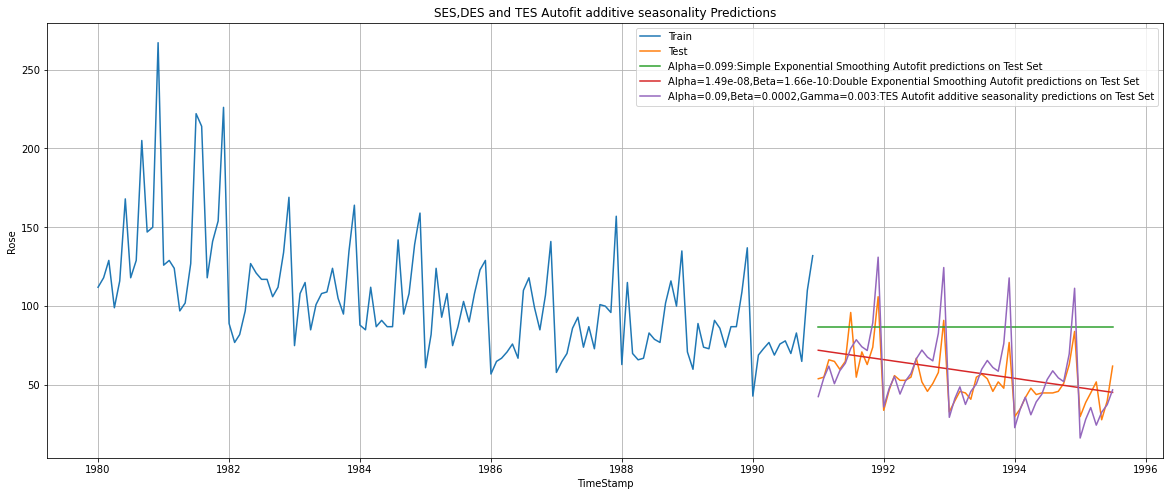


Figure :Rose TES add seas

RMSE is calculated on test data.

TES\_A Autofit RMSE: 14.26

**Triple Exponential Smoothing - ETS(A, A, M) - Holt Winter's linear method with multiplicative errors**

In this model trend is additive and seasonality is chosen as multiplicative. A DES model is built on train data with initialization\_method value as estimated , use\_brute=True and Optimized=True. The other parameters and the values are:

The model built is used to forecast for next 55 months which is the length of

test data.

{'smoothing\_level': 0.0715106306609405, 'smoothing\_trend': 0.04529179757535142, 'smoothing\_seasonal': 7.244325029450242e-05, 'damping\_trend': nan, 'initial\_level': 130.40839142502193, 'initial\_trend': -0.77985743179386, 'initial\_seasons': array([0.86218996, 0.977675 , 1.0687727 , 0.93403881, 1.050625 ,

1.14410977, 1.25836944, 1.33937772, 1.26778766, 1.24131254,

1.44724625, 1.99553681]), 'use\_boxcox': False, 'lamda': None, 'remove\_bias': False}

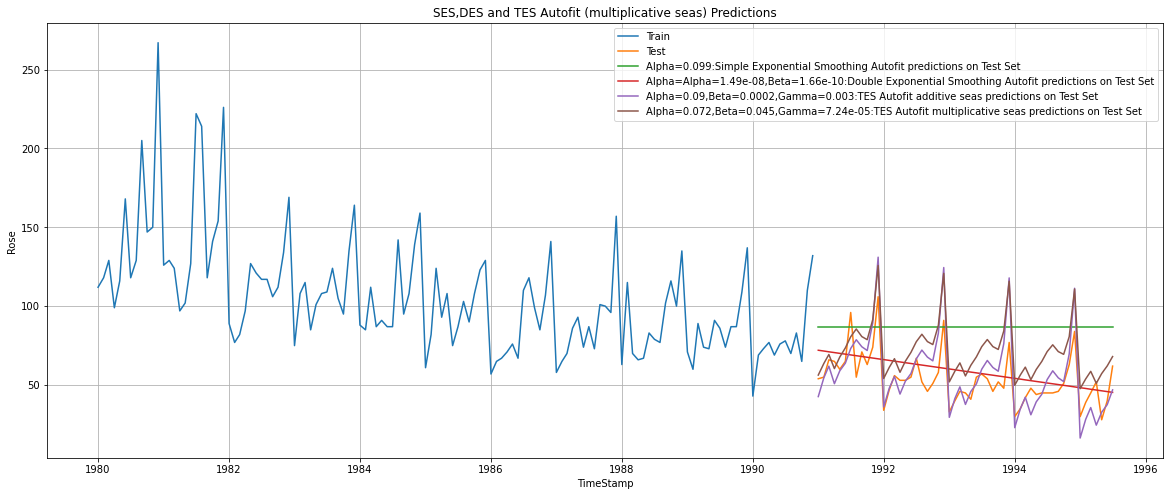
****

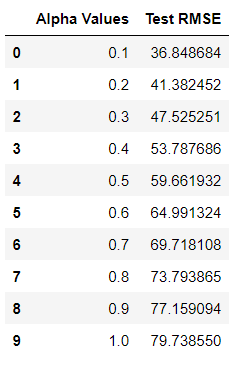
Figure :Rose TES multipl seas

RMSE is calculated on test data.

TES\_m Autofit RMSE: 20.18

## **Iterative Method for Simple Exponential Smoothing**

A SES model is built on train data with initialization\_method value as estimated , use\_brute=True and Optimized=False. The value of smoothing\_level(alpha) is taken from 0.1 to 1 and its corresponding RMSE value is calculated as shown below.



We can see that alpha as 0.1 as the least RMSE.

Plot the prediction of the model with smoothing level value as 0.1.

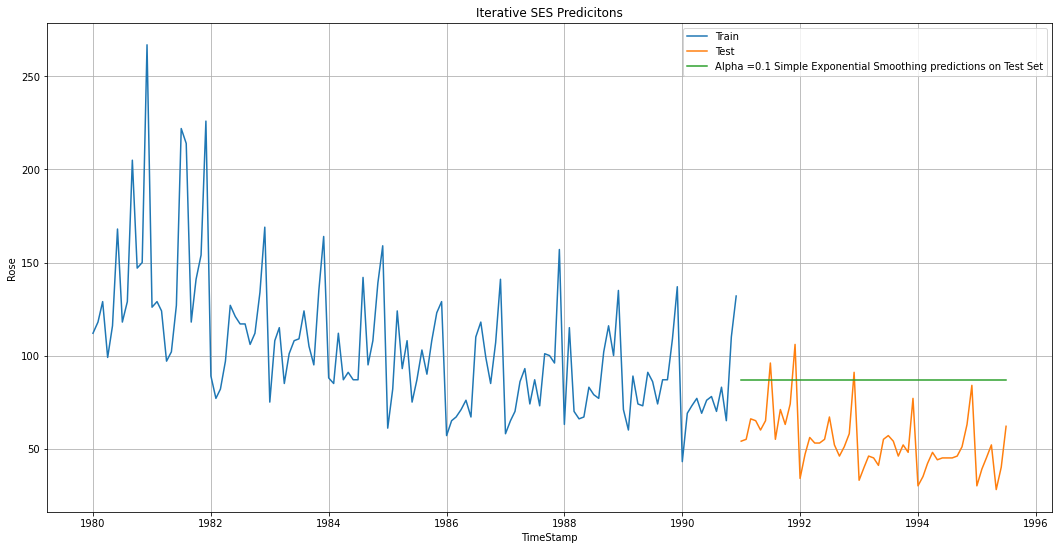
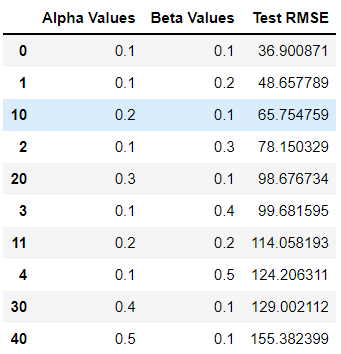


Figure :Rose Iterative SES

The RMSE of the iterative SES is 36.85

**Iterative Method for Double Exponential Smoothing**

A DES model is built on train data with initialization\_method value as estimated , use\_brute=True and Optimized=False. The values of smoothing\_level(alpha) and smoothing\_trend(beta) is taken from 0.1 to 1 and its corresponding RMSE value is calculated .The first 10 rows with least RMSE is shown below.



The model with smoothing level as 0.1 and smoothing trend as 0.1 has the least RMSE value 36.9.

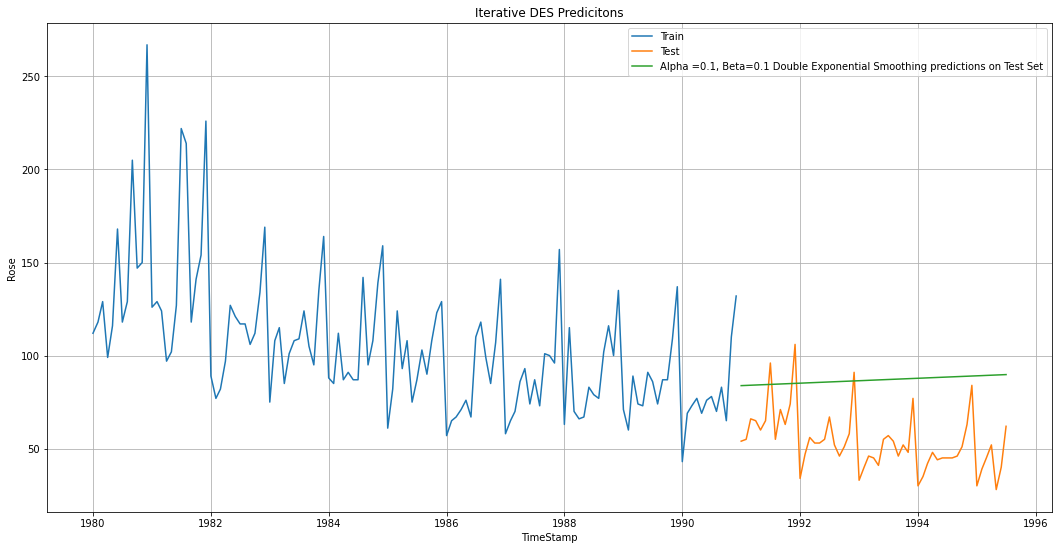
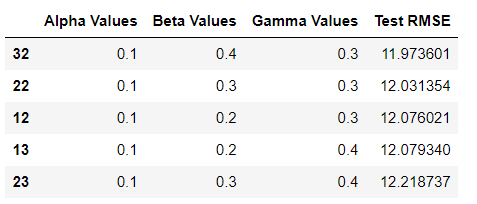
****

Figure :Rose Iter DES

The RMSE of Iterative DES is 36.9.

### **Iterative Method - Triple Exponential Smoothing - ETS(A, A, A)**

A TES model with trend and seasonality as additive is built on train data with initialization\_method value as estimated , use\_brute=True and Optimized=False. The values of smoothing\_level(alpha) ,smoothing\_trend(beta) and smoothing\_seasonal(gamma)is taken from 0.1 to 1 and its corresponding RMSE value is calculated. First 5 rows of the data which shows the 5 least RMSE ones are shown below.



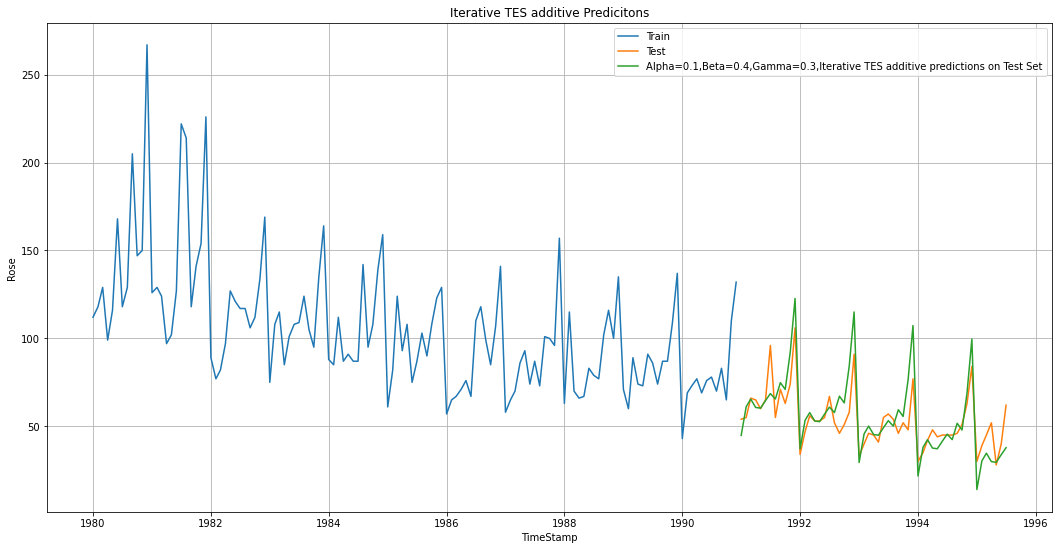
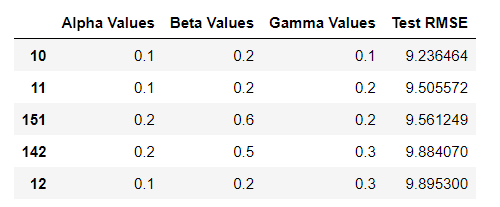


Figure :Rose Iter TES add seas

The RMSE of iterative TES additive is 11.97.

**Iterative Method - Triple Exponential Smoothing - ETS(A, A, M)**

A TES model with trend as additive and seasonality as multiplicative is built on train data with initialization\_method value as estimated , use\_brute=True and Optimized=False. The values of smoothing\_level(alpha) ,smoothing\_trend(beta) and smoothing\_seasonal(gamma)is taken from 0.1 to 1 and its corresponding RMSE value is calculated. First 5 rows of the data which shows the 5 least RMSE ones are shown below.



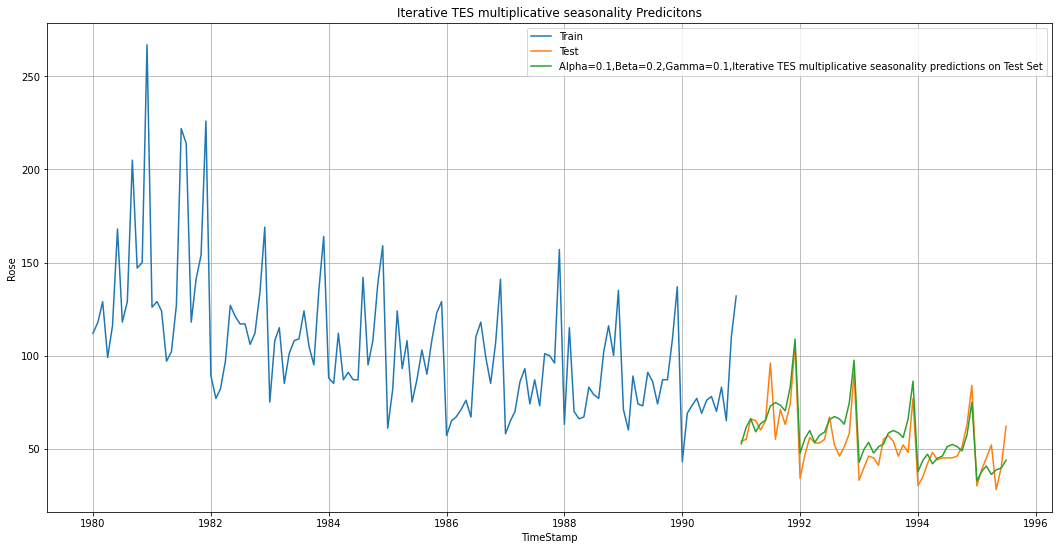
****

Figure :Rose Iter multiplicative seas

The RMSE of iterative TES multiplicative is 9.24.

### **Linear Regression Model**

Linear regression uses the relationship between the data-points to draw a straight line through all them. This line can be used to predict future values.

The training and testing time instance is created.

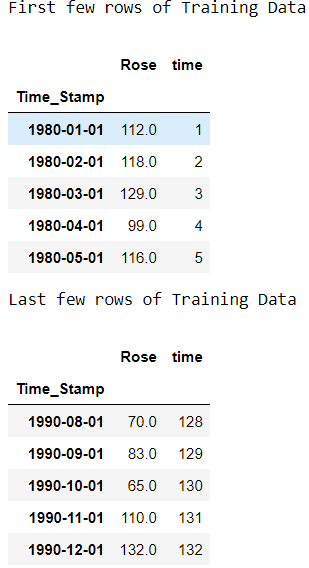
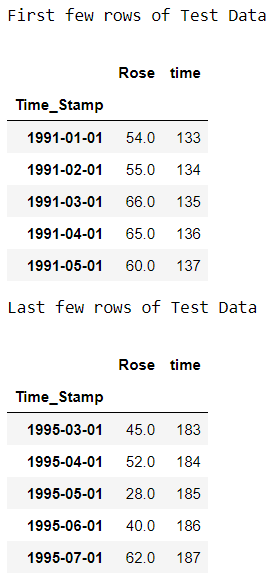
Training Time instance

[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132]

Test Time instance

[133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187]

The instances are added to the train and test data as column ‘time’ and treated as an independent variable.

Rose column is treated as a dependent variable. Both the train independent and dependent variable are fitted on the Linear Regression model from sklearn library.

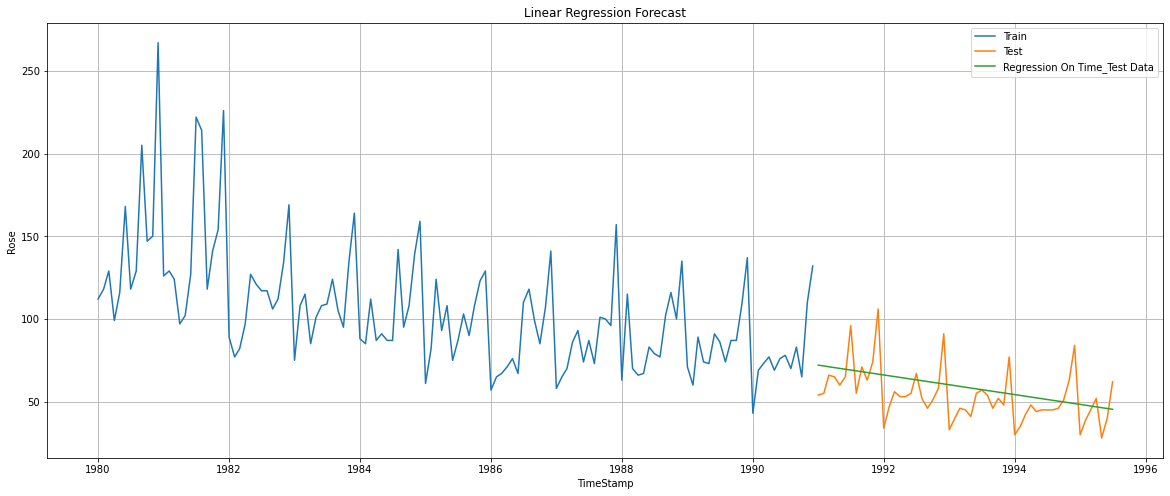


Figure :Rose-Linear Regr Plot

The fitted model is sued to predict on test data and the output of Rose wine

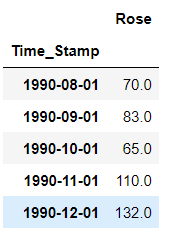
sales is used to calculate RMSE along with actual Rose wine sales.

The RMSE of Linear Regression forecast on the Test Data is 15.280

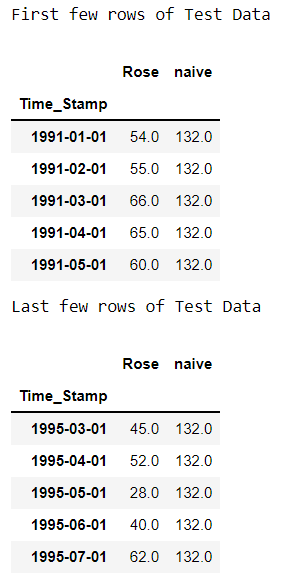
### **Naive Forecast Model:**

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.

The last 5 rows of the train data are:



The last value is 132 which is treated as the forecasted value for the future time period as shown below.



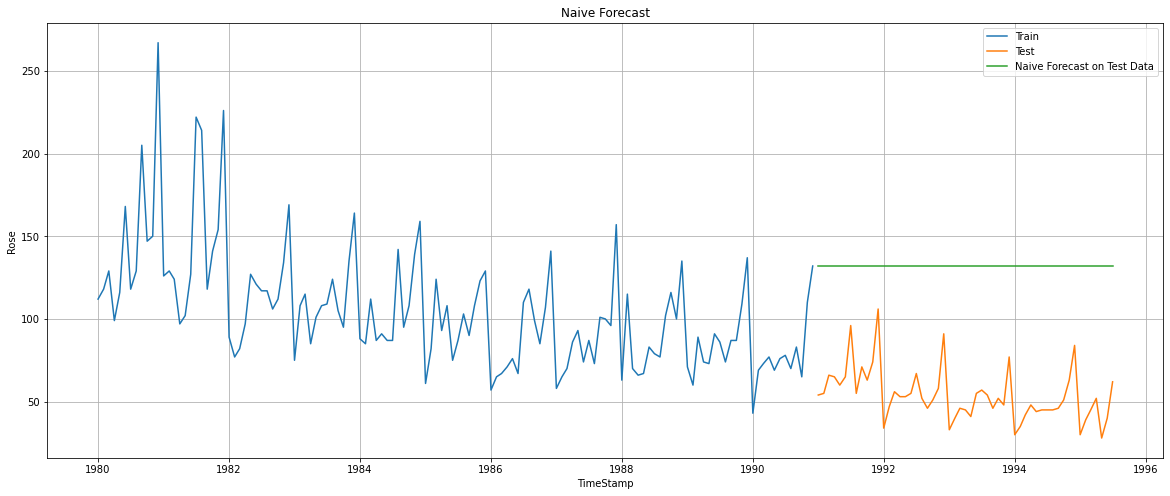


Figure :Rose-Naive Forecast plot

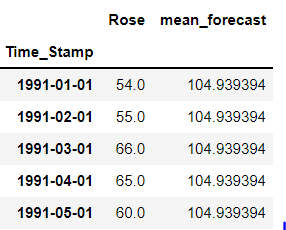
The predicted test output along with the actual test output is used for calculating RMSE.

The RMSE for NaiveModel forecast on the Test Data is 79.74

### **Simple Average Model**

### In the Simple Average model ,forecast is equal to the average of historical data.

The average of the train data is taken and added as the forecast of Test data as shown below.



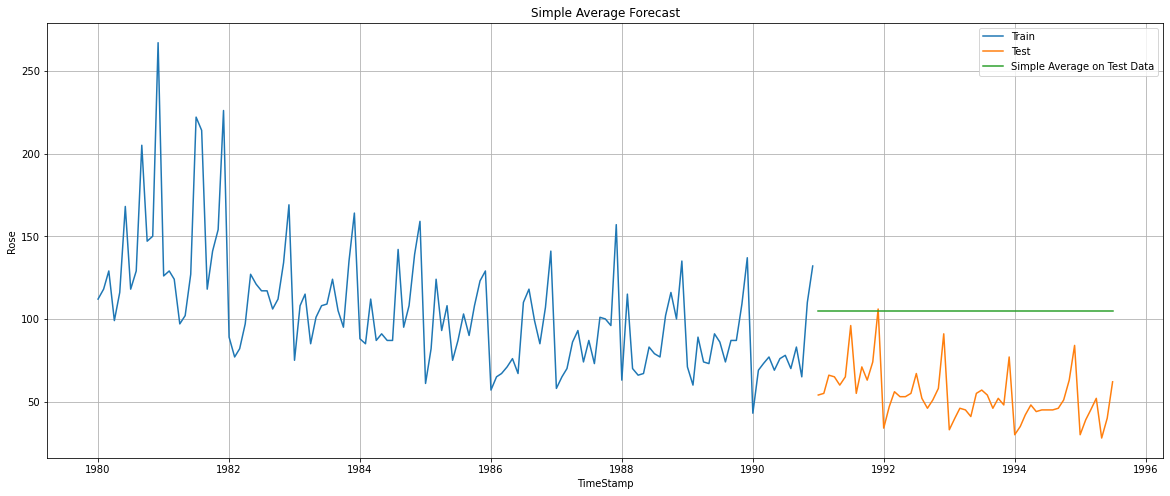


Figure :Rose-Simple Average Plot

The predicted test output along with the actual test output is used for calculating RMSE.

The RMSE for Simple Average forecast on the Test Data is 53.48

**Moving Average Forecast Model**

Moving Average Forecast Model takes an average of a set of numbers in a given range while moving the range.

Moving Average is calculated on train data for which the following values are given to rolling function.

2-takes moving average of 2 months of data

3-takes moving average of 3 months of data

6-takes moving average of 6 months of data

12-takes moving average of 12 months of data

The first 5 rows of training data is the following:

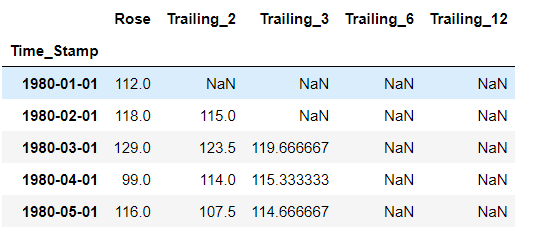


Table :Rose-MA-Train head

The last 5 rows of training data is the following:

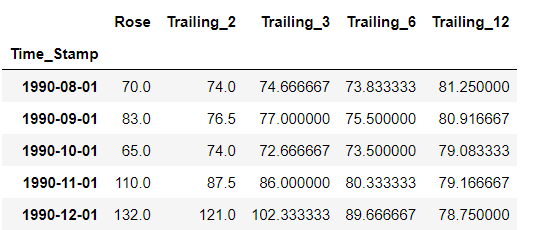


Table :Rose-MA-train tail

The value 121 is taken as the forecasted value for the test for 2 point Moving Average.

The value 102 is taken as the forecasted value for the test for 3 point Moving Average.( 102.333333 rounded off)

The value 90 is taken as the forecasted value for the test for 6 point Moving Average. ( 89.666667 rounded off)

The value 79 is taken as the forecasted value for the test for 12 point Moving Average. (78.750000 rounded off)

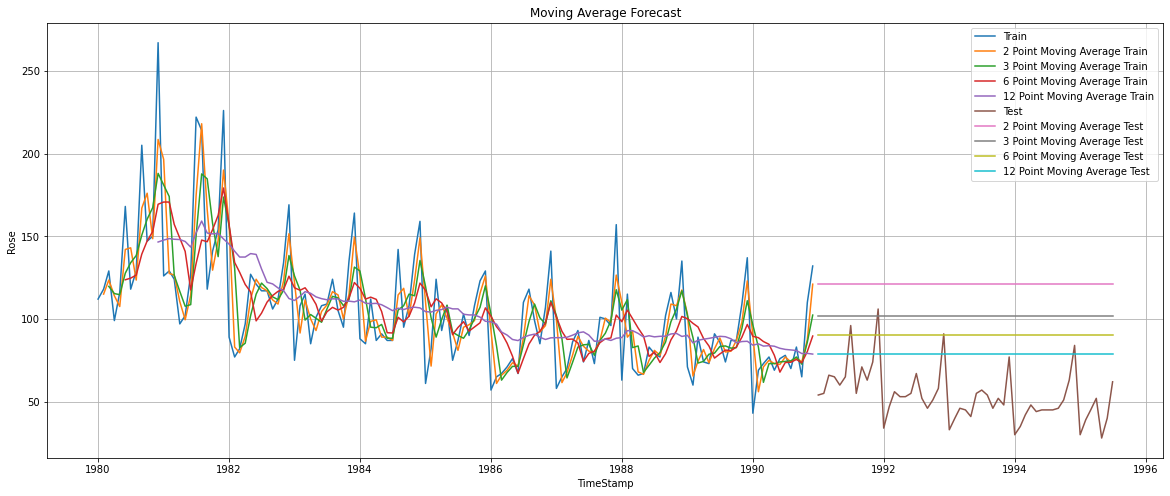


Figure :Rose-Moving Average plot

The RMSE for 2 point Moving Average forecast on the Test Data is 68.99

The RMSE for 3 point Moving Average forecast on the Test Data is 50.68

The RMSE for 6 point Moving Average forecast on the Test Data is 39.45

The RMSE for 12 point Moving Average forecast on the Test Data is 29.7

**1.5. 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.**

Augmented Dickey Fuller test (ADF Test) is a common statistical test used to

test whether a given Time series is stationary or not. It is one of the most

commonly used statistical test when it comes to analyzing the stationary of a

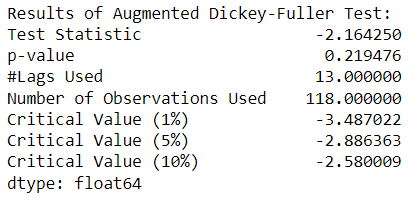
series.

The hypothesis for the statistical test is:

H0-Null Hypothesis: Time series is non-stationary

H1-Alternate Hypothesis: Time series is stationary

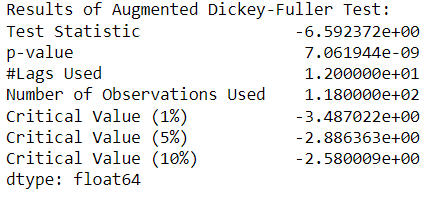
The ADF Test is conducted on the train data



The p-value obtained by the test should be less than the significance level (say 0.05) to reject the Null hypothesis or it fails to reject the Null hypothesis.

p value obtained from the ADF test is 0.219 which is greater than 0.05 . Hence we fail to reject the Null Hypothesis and so we can say that data is non-stationary.

To convert the data into a stationary one, the difference of a Dataframe value with the value in the previous row is taken and remove missing values. The ADF Test is taken again on the modified train data.



After modifying, the p-value 7.061943750942e-09 obtained by the test is less than 0.05. Now the data has been converted into a stationary one.

**1.6. 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.**

**ARIMA Automated Train Model**

Auto Regressive Integrated Moving Average (ARIMA) models are applied on time series data when the current value is assumed to be correlated to past values and past prediction errors. Therefore, these models are used in defining current value as a linear combination of past values and past prediction errors. Here, we have defined a few terms that would be useful in understanding ARIMA models in detail. ARIMA models can only be applied only on stationary time series data.

The Akaike information criterion (AIC) is an estimator of out-of-sample prediction error and thereby relative quality of statistical models for a given set of data. The least the AIC the better the model is .

The p,q value is taken from 0 to 3.‘d’ is taken as 1.

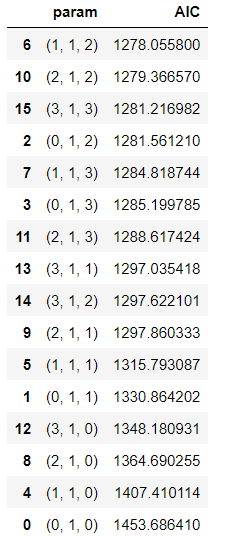


Table :Rose-Automated ARIMA AIC

The model has the parameter ‘order’ which has its values in the form of (p,d,q) where

p: Trend autoregression order.

d: Trend difference order.

q: Trend moving average order.

The p value is taken as 1, q as 2 and d as 1.(1,1,2) as it has the least AIC value 1278.06.

ARIMA is built using stationary data after dropping its NA values since it reduces the AIC value .

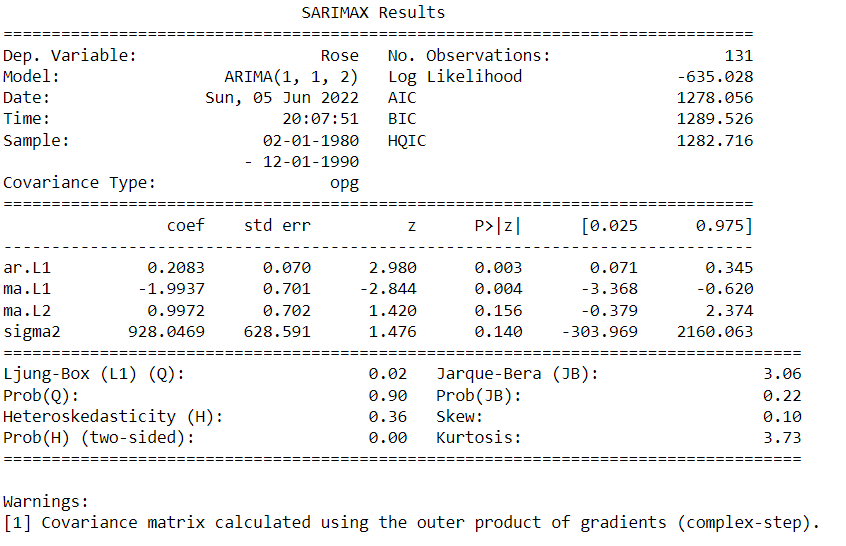
enforce\_stationarity🡪Whether or not to transform the AR parameters to enforce stationarity in the autoregressive component of the model.

enforce\_invertibility🡪Whether or not to transform the MA parameters to enforce invertibility in the moving average component of the model.

enforce\_stationarity and enforce\_ invertibility is given as false.

The ARIMA model is built with those values and RMSE is calculated on test data.

The Summary of ARIMA model is



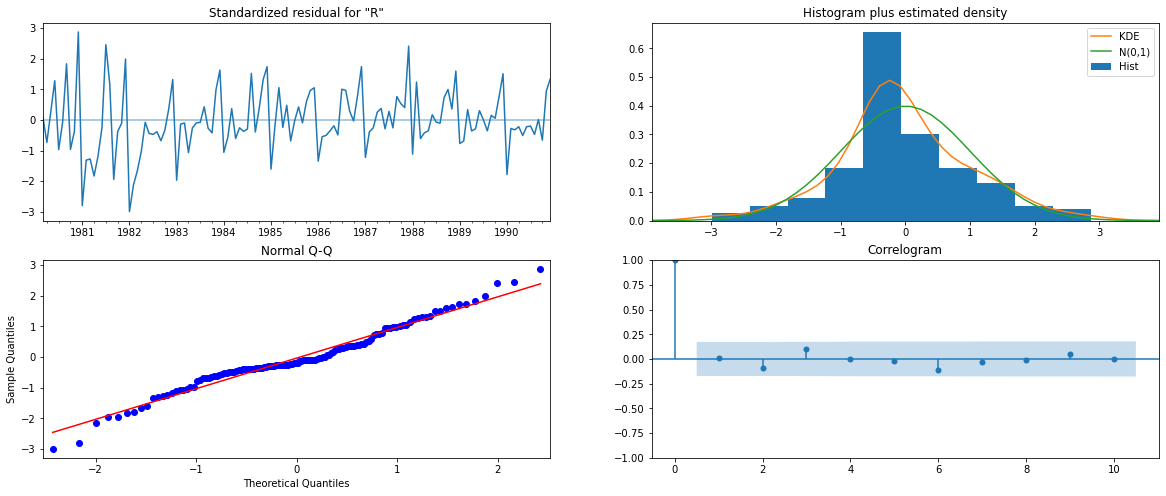


Figure :Rose -Automated ARIMA diagnostics

The diagnostics look good here.

RMSE of Automated ARIMA model on Test data is: 56.93

**SARIMA Automated Train Model**

Seasonal Autoregressive Integrated Moving Average, SARIMA or Seasonal ARIMA, is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component.

It adds three new hyperparameters to specify the autoregression (AR), differencing (I) and moving average (MA) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality.

**Autocorrelation Function (ACF)**

A plot of auto-correlation of different lags is called ACF.The plot summarizes the correlation of an observation with lag values. The x-axis shows the lag and the y-axis shows the correlation coeﬃcient between -1 and 1 for negative and positive correlation.

**Partial Autocorrelation Function (PACF)**

A plot of partial auto-correlation for different values of lags is called PACF.

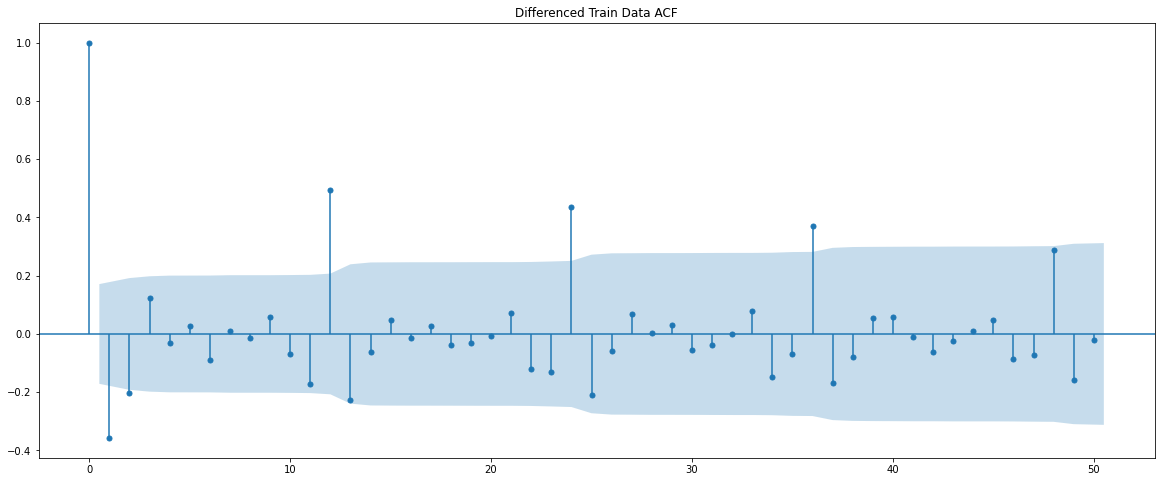


Figure :Rose-Auto SARIMA ACF

The model has the parameter ‘seasonal order’ which has its values in the form of (P, D, Q, s):

P: Seasonal autoregressive order.

D: Seasonal difference order.

Q: Seasonal moving average order.

s: The number of time steps for a single seasonal period.

’s’ is determined from acf plot

The P,Q value is taken from 0 to 2. D is as 0 and 1. ‘d’ is taken as 1. From the above ACF plot we can say that ,Seasonality after every 12th lag is visible. We will run our auto SARIMA models by setting seasonality as 12. SARIMA is built using stationary data after dropping its NA values since it reduces the AIC value.



Table :Rose-Automated SARIMA AIC

The parameters in the first row has the least AIC so its taken to build the SARIMA model. ‘p’ is taken as 0, d as 1, q as 2, P as 2, D as 1, Q as 2 and s as 12.

SARIMA is built using stationary data after dropping its NA values since it reduces the AIC value to 775.38.

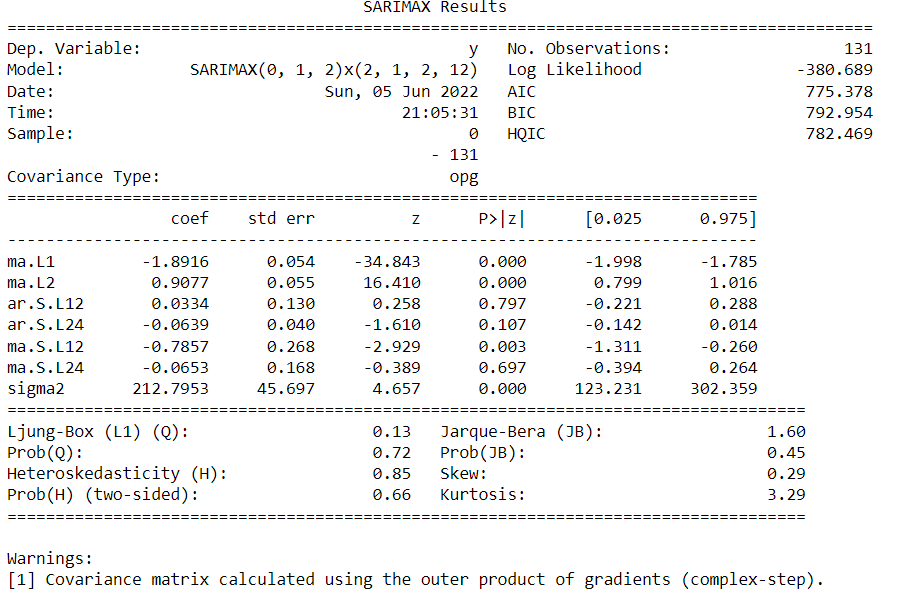
enforce\_stationarity🡪Whether or not to transform the AR parameters to enforce stationarity in the autoregressive component of the model.

enforce\_invertibility🡪Whether or not to transform the MA parameters to enforce invertibility in the moving average component of the model.

enforce\_stationarity and enforce\_ invertibility is given as false.

The SARIMA model is built with those values and RMSE is calculated on test data.

The Summary of SARIMA model built is:



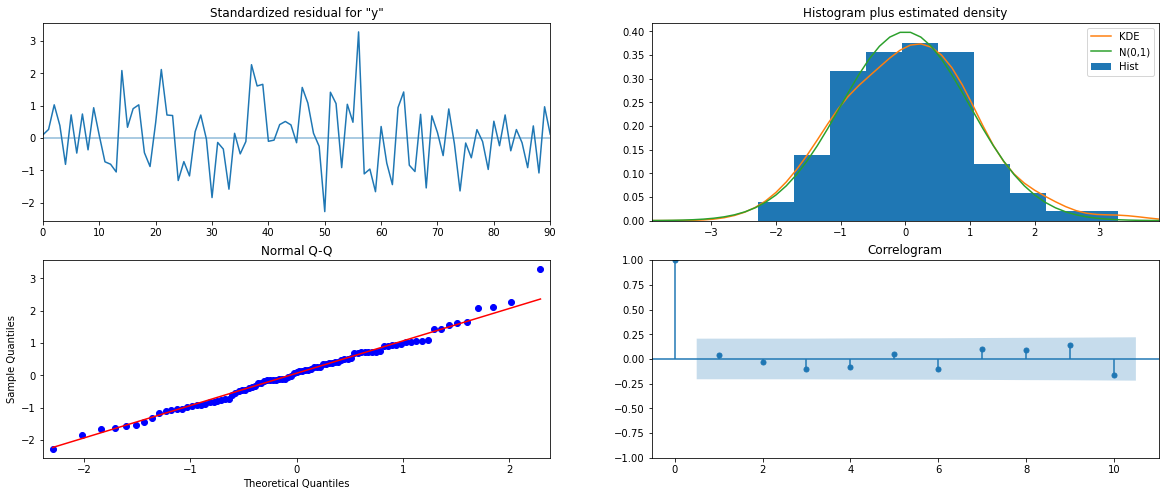


Figure :Rose-Automated SARIMA diagnostics

RMSE of Automated SARIMA model on Test data is: 61.44

**1.7. 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.**

**Manual ARIMA Model**

ACF is used for identifying the value of q and PACF is used for identifying the value of 𝑝.

The p value of the Augmented Dickey-Fuller Test on Train data: 0.219476 .‘p’ is not less than 0.05 so the data isnt stationary.

The p value of the Augmented Dickey-Fuller Test on Train data: 7.061943750942e-09. The data after first difference is stationary as the p value is less than 0.05. Thus we can consider the value of d as 1.

ACF and PACF plotted with differenced data is below:

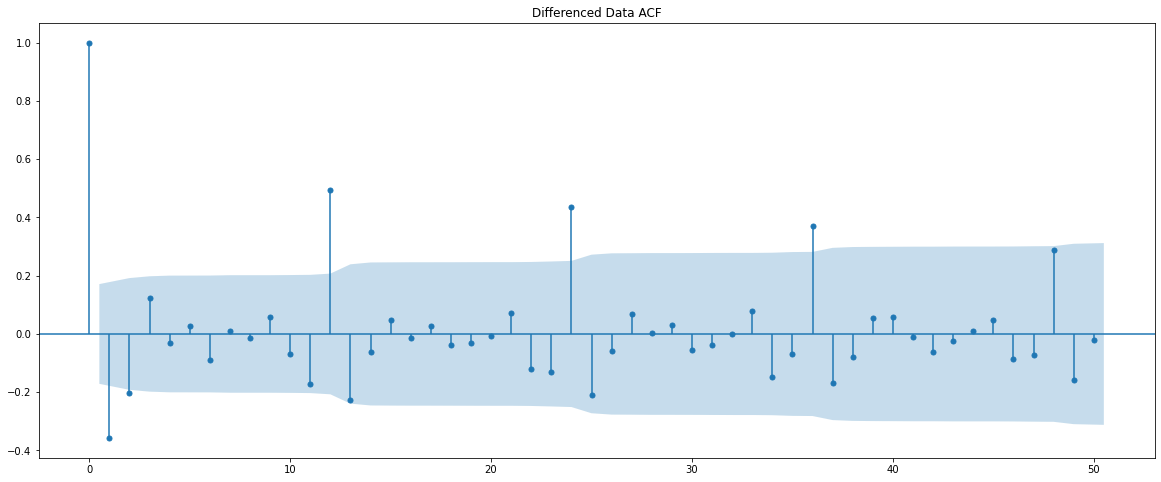
****

Figure :Rose-Manual ARIMA ACF

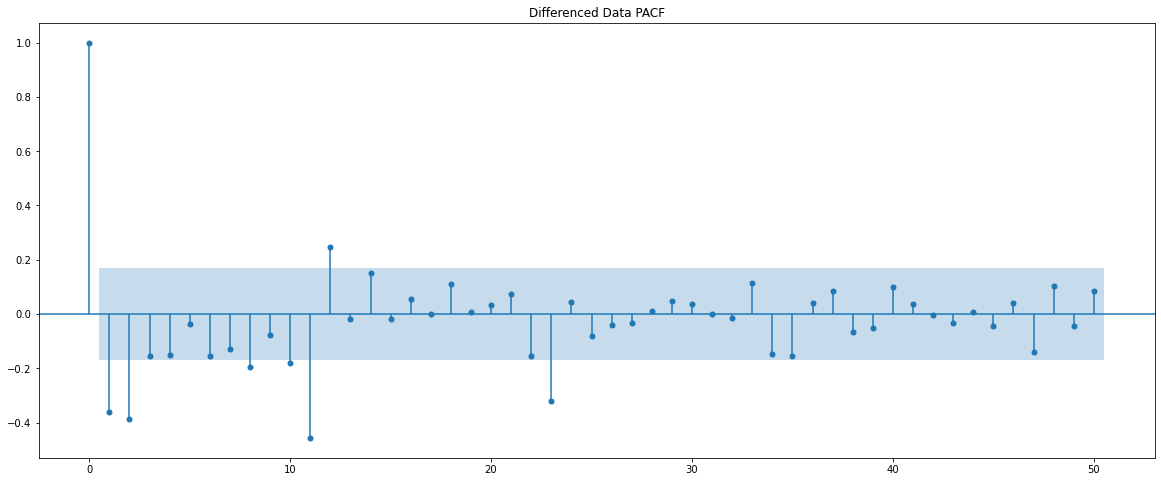


Figure :Rose-Manual ARIMA PACF

From the ACF and PACF models we can take the value of p and q as 2. Since difference of order 1 has been taken on the data to make it stationary ,d=1.

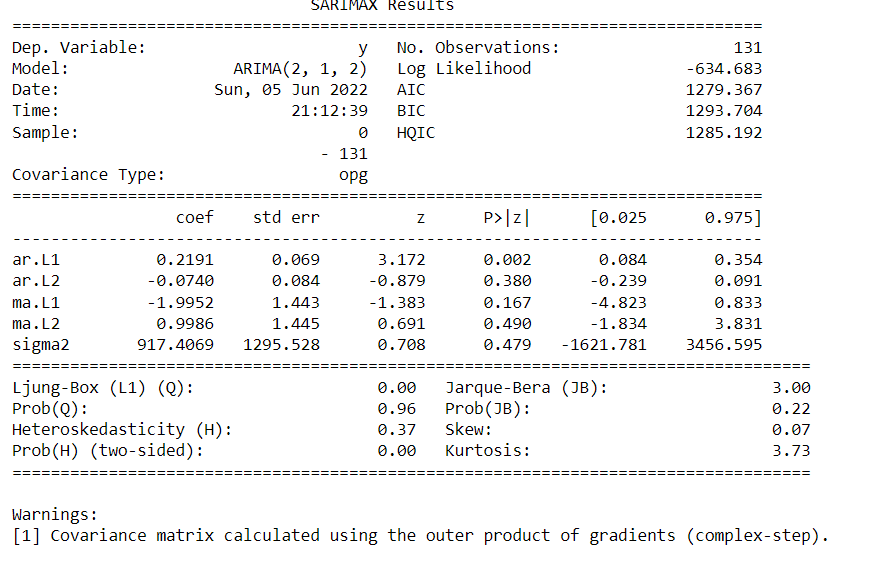
enforce\_stationarity🡪Whether or not to transform the AR parameters to enforce stationarity in the autoregressive component of the model.

enforce\_invertibility🡪Whether or not to transform the MA parameters to enforce invertibility in the moving average component of the model.

enforce\_stationarity and enforce\_ invertibility is given as false.

ARIMA is built using stationary data after dropping its NA values .The ARIMA model is built with those values and RMSE is calculated on test data.

The summary of Manual ARIMA model is:



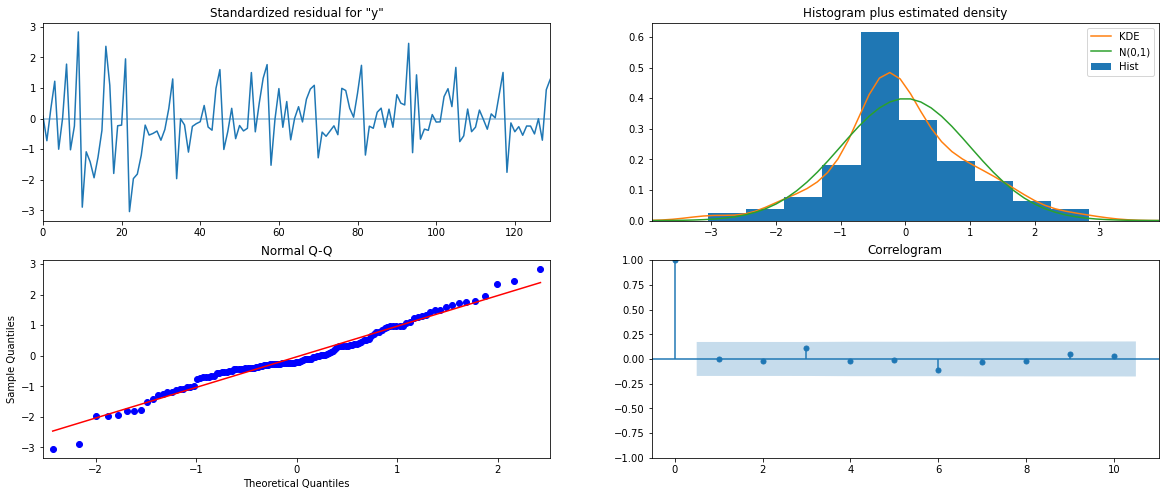


Figure :Rose-Manual ARIMA diagnostics

The diagnostics are good here.

RMSE of Manual ARIMA model on Test data is: 56.97

**Manual SARIMA Model:**

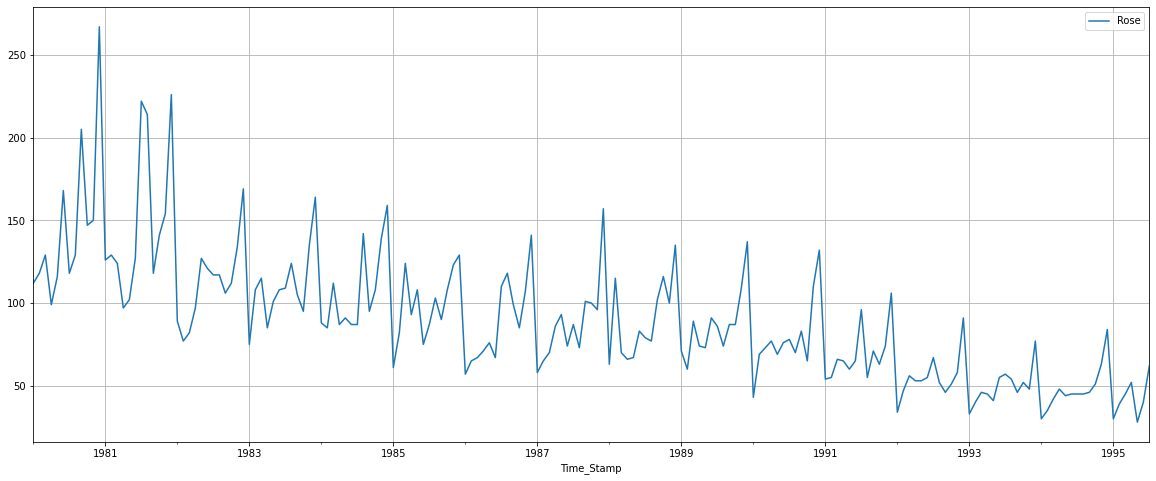
****

Figure :Rose-Manual SARIMA plot 1

We see that there is a trend and a seasonality. So, now we take a seasonal differencing and check the series.

Since the seasonality parameter is 12 we can plot the graph for the difference of order 12 with NA values dropped.

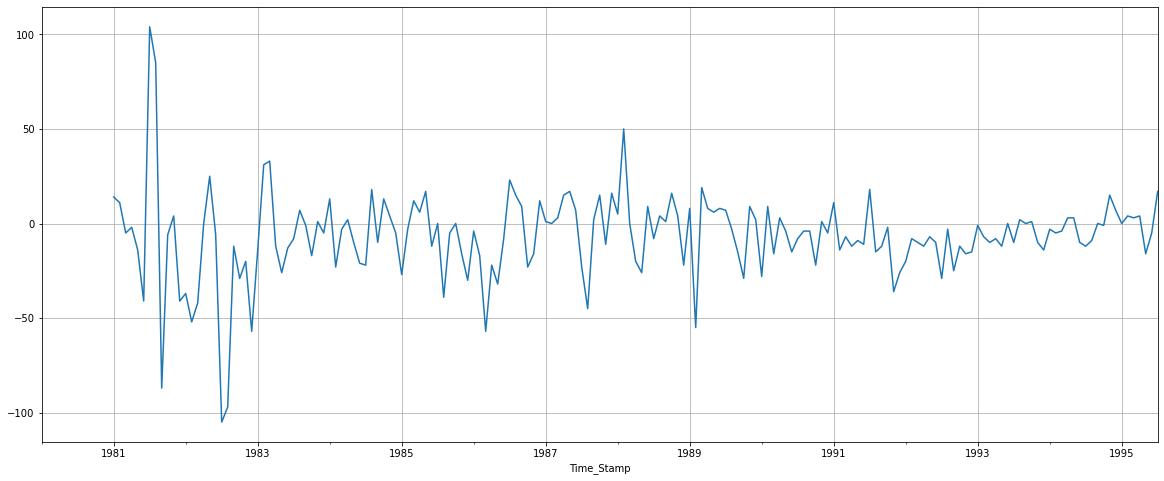
****

Figure :Rose-Manual SARIMA plot 2

As there is a slight trend in the graph we can take a differencing of first order on the seasonally differenced series .

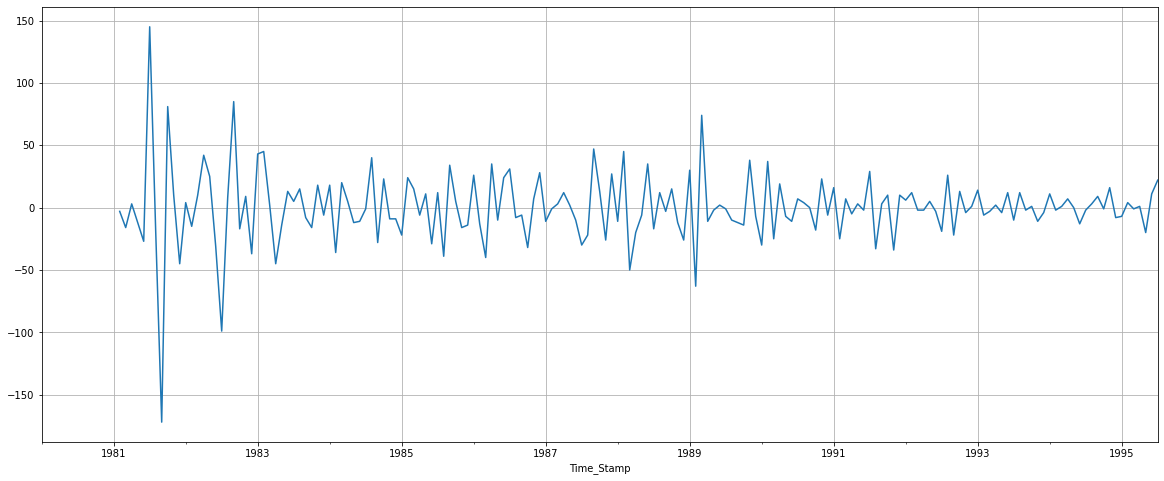
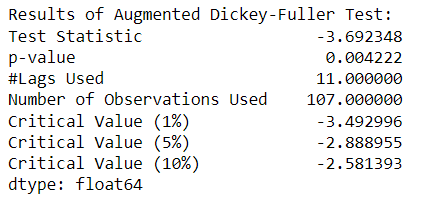


Figure :Rose-Manual SARIMA plot 3

Now we see that there is almost no trend present in the data. Seasonality is only present in the data.

Let us go ahead and check the stationarity of the above series before fitting the SARIMA model.



The series is stationary.

The first difference of seasonal differenced train data is used to plot ACF and PACF as below:

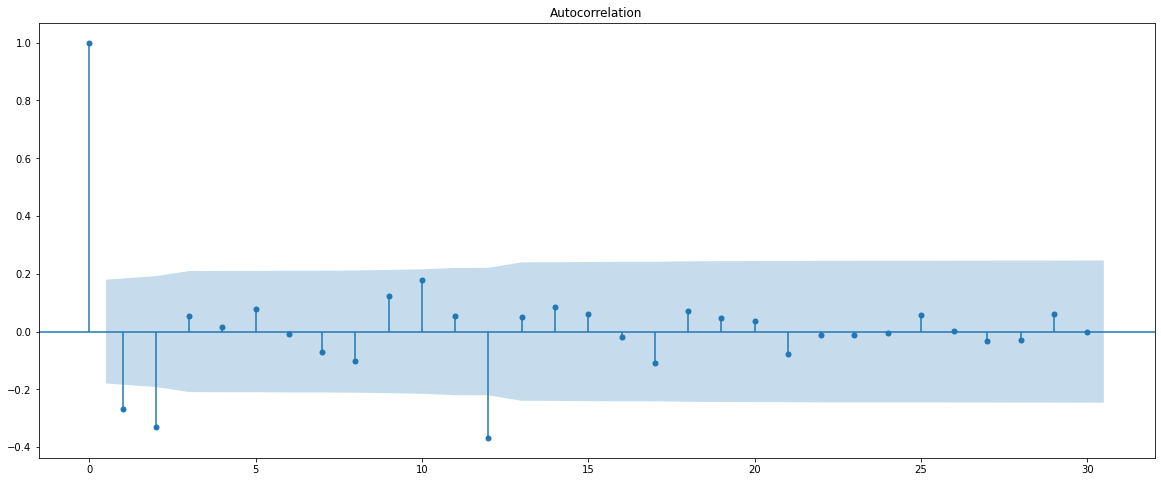


Figure :Rose-Manual SARIMA ACF

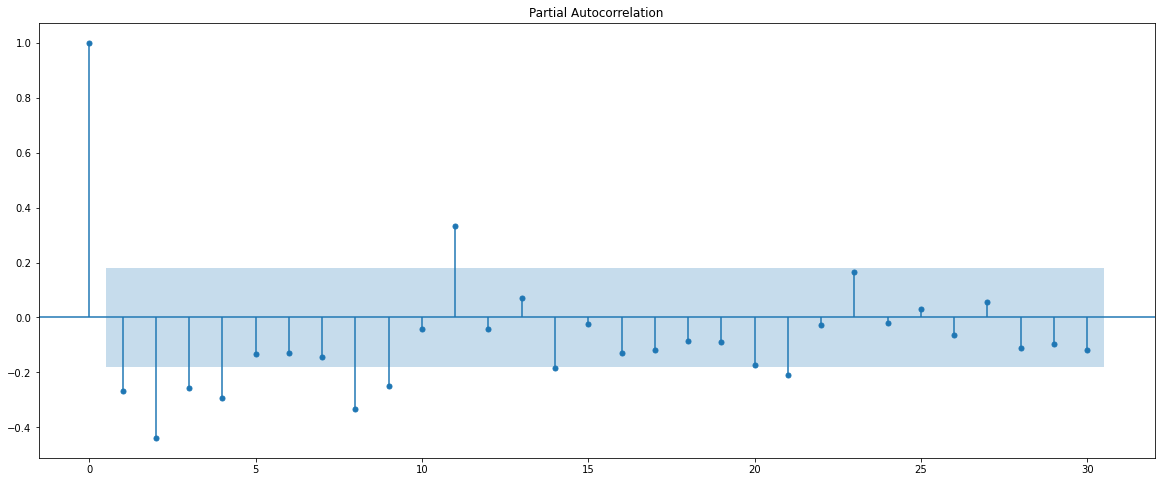


Figure :Rose-Manual SARIMA PACF

Here, we have taken alpha=0.05.

We are going to take the seasonal period as 12. We will keep the p(1) and q(1) parameters same as the ARIMA model.

The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 0.

The Moving-Average parameter in an SARIMA model is 'Q' which comes from the significant lag after which the ACF plot cuts-off to 0.

By looking at the plots we see that the ACF and the PACF do not directly cut-off to 0.

Hence P=4 and Q=2. As we have taken a differencing of first order on the seasonally differenced series D is taken as 1.

‘p’,’q’,’d’ has the same value as the one calculated in the ARIMA model.

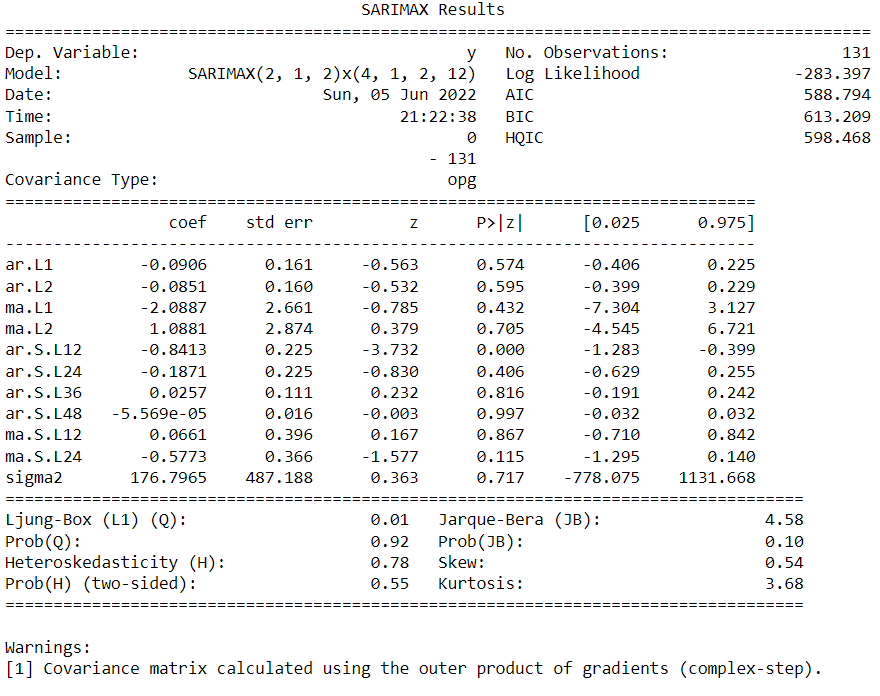
Order=(2,1,2) and Seasonal\_order=(4,1,2,12) are the parameters to be passed to the SARIMA model along with the difference data after dropping NA values.

enforce\_stationarity🡪Whether or not to transform the AR parameters to enforce stationarity in the autoregressive component of the model.

enforce\_invertibility🡪Whether or not to transform the MA parameters to enforce invertibility in the moving average component of the model.

enforce\_stationarity and enforce\_ invertibility is given as false.

The model is built with these values and RMSE is calculated.



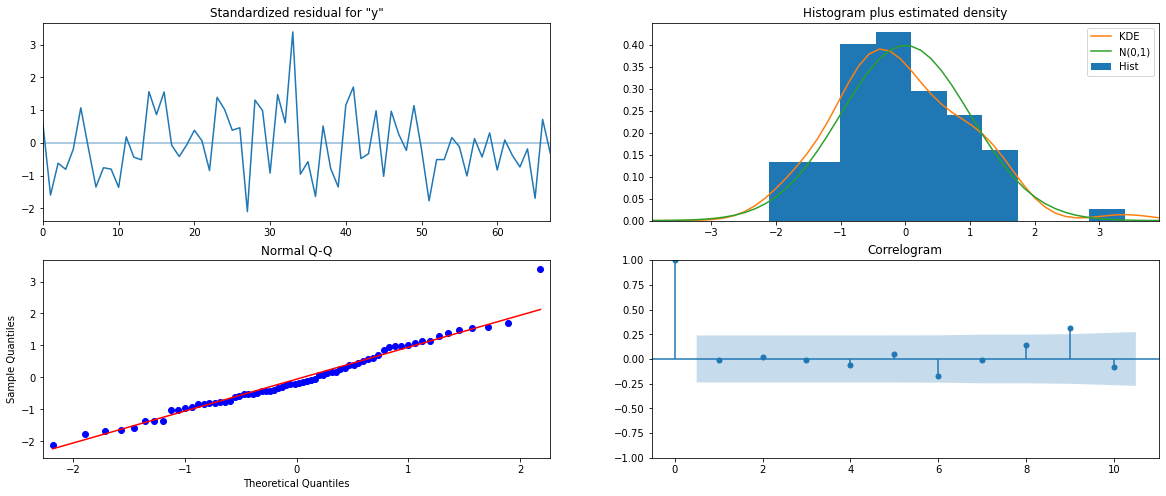


Figure :Rose-Manual SARIMA diagnostics

RMSE of Manual SARIMA model on Test data is: 60.24

**1.8.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.**

The model with its parameter and its corresponding RMSE value is shown below in ascending order of RMSE value from which we can see that Iterative Triple Exponential Smoothing with additive trend and Multiplicative Seasonality has the least RMSE value of 9.24. Hence it’s the optimum model.

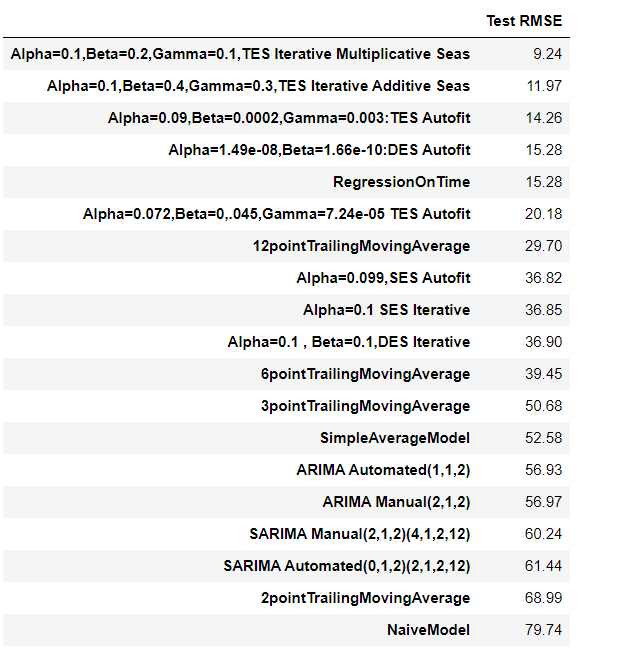


Table :Rose-Models-RMSE

**1.9.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**

The optimum model Iterative Triple Exponential Smoothing with additive trend and Multiplicative Seasonality is built and forecasted for 12 months into the future.

The paramters are as below:

{'smoothing\_level': 0.1, 'smoothing\_trend': 0.2, 'smoothing\_seasonal': 0.1, 'damping\_trend': nan, 'initial\_level': 145.27499999999992, 'initial\_trend': 0.7643939393939481, 'initial\_seasons': array([0.75572235, 0.80417408, 0.89051255, 0.75964172, 0.88067767,

0.92553586, 1.08905952, 1.13538639, 1.03030222, 0.96585715,

1.13788391, 1.62524659]), 'use\_boxcox': False, 'lamda': None, 'remove\_bias': False}

The forecast is shown with 95% confidence interval band.

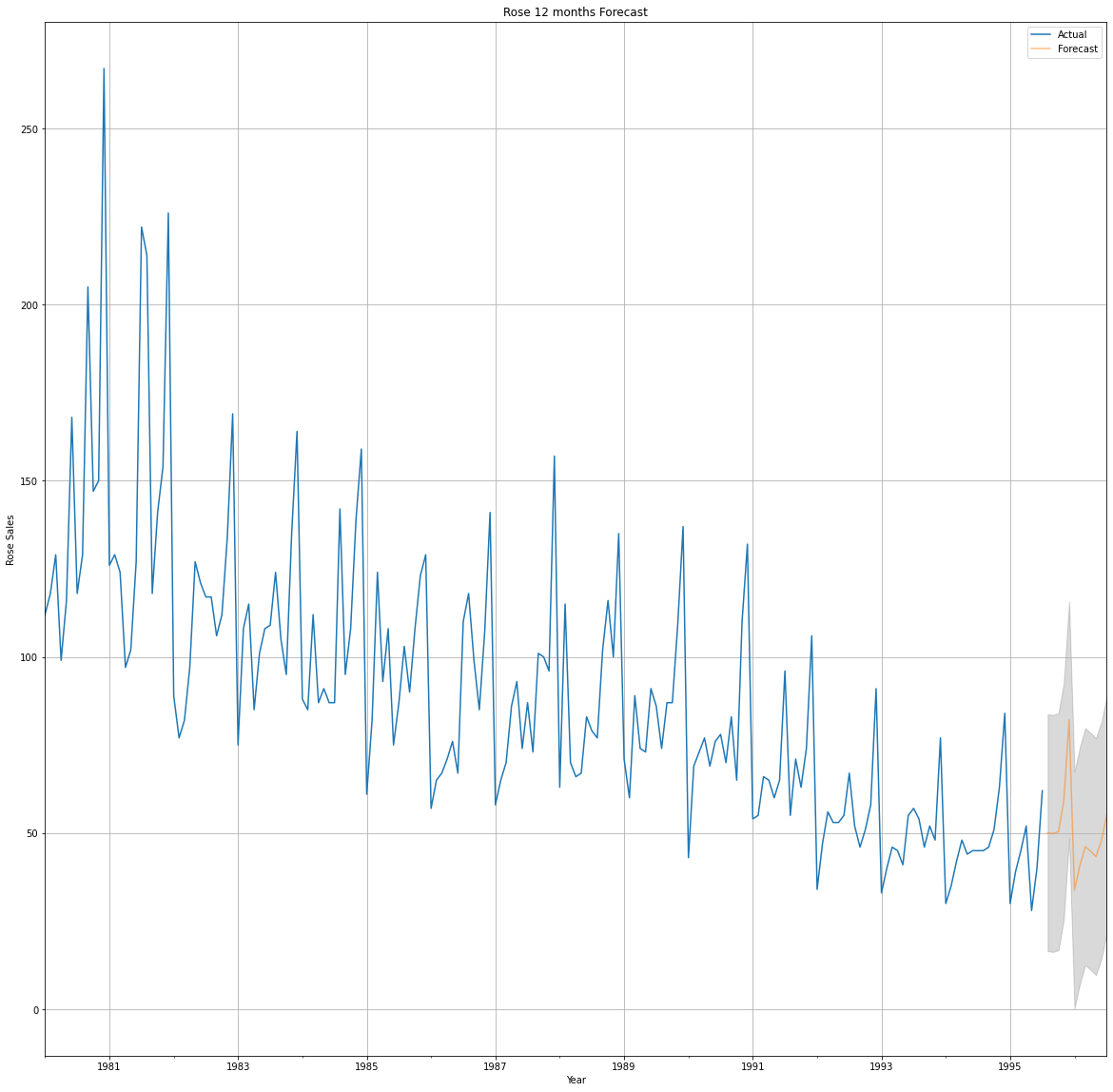


Figure :Rose-12 months Forecast

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

1. First read the data as a time series and plotted it on a graph to show how sales for Sparkling wines over the years.

2. Then performed some exploratory data analyses on the data sets, creating various types of charts for analyze the sales. The missing values are also imputed.

3. I split the data into test(data from the year 1991) and train(data before the year 1991).

4. Next I built the following models :

• Simple Exponential Smoothing Model

• Iterative Simple Exponential Smoothing Model

• Double Exponential Smoothing Model

• Iterative Double Exponential Smoothing Model

• Triple Exponential Smoothing Model

• Iterative Triple Exponential Smoothing Model.

• Linear Regression Model

• Naïve Approach

• Simple Average Model

• Moving Average Model

For all the above models RMSE value was calculated to understand the performance.

5. The stationarity of the data was checked by stating hypothesis for statistical testing and using ADF Test.

6. From here, we build ARIMA and SARIMA models, but first we examine the dataset. If the series is not stationary, we take the first difference of the series and converted into a stationary series.

7. The ARIMA/SARIMA models are built using AIC scores, we select the parameter with the least AIC and the model is built with it. RMSE is also calculated to check the performance.

8. The ARIMA/SARIMA models are built manually by calculating value of p,q,P,Q,s,d,D from ACF , PACF graph. RMSE is also calculated to check the performance.

9. Finally, we take the model with minimum RMSE value and build the most optimum model on the complete data .The sales for the next 12 months in future with 95% confidence intervals is predicted.

**Recommendations**

--> Fourth quarter has the highest sales among other quarter. So the company can stock up the wines in the second quarter itself to prepare themselves to supply the high demand in the fourth quarter.

--> Proper branding advertising in leading newspaper and magazines can be done. Social Media Advertising can also be done to improve the sales in the first 3 quarters.

--> First quarter has the lowest sales . So coupons and differs can be offered to boost up the sales.

--> Over the years sales are decreasing this can be due to more competition with new companies. As time progresses the wine company must bring in unique tastes.

--> Further information like age group,location of the customers can be analyzed to improve the model performance and get a better understanding of the Rose wine sales.

**PROBLEM 2 - Sparkling**

**Problem Statement**

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 Sparkling Wine Sales in the 20th century.

**Introduction**

The purpose of this whole exercise is to perform exploratory data analysis and perform Time Series Forecasting using Exponential Smoothing models, Regression, Naïve Forecast models, Simple Average models, Moving Average models, ARIMA and SARIMA models(using cut-off points of AIC,ACF and PACF plots) to forecast the sales of Sparkling wine.

**Data Description**

1. YearMonth: The Year and the Month on which its corresponding units of Rose Wine is sold.
2. Sparkling: Units of Sparkling Wine sold.
   1. **Read the data as an appropriate Time Series data and plot the data.**

**Sample of the dataset:**

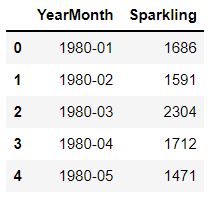
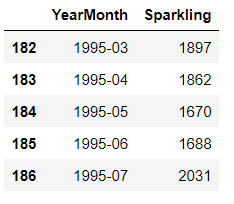
 

Table : Sample of the Dataset2

The data is read from the excel file and the above tables shows the first and last 5 rows of the dataset.There are 187 rows in the dataframe. The Sparkling is the variable to be forecasted . YearMonth denotes the year and month values ranging from Jan 1980 to July 1995.

There are no duplicates in the datatset.

There are 0 duplicates in the dataset

The initial datatype of the columns before indexing are:

YearMonth object

Sparkling int64

dtype: object

YearMonth column is converted into a Time Stamp index using

to\_datetime function and YearMonth is dropped.

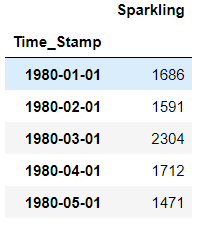


Figure :Sparkling-Timestamped data

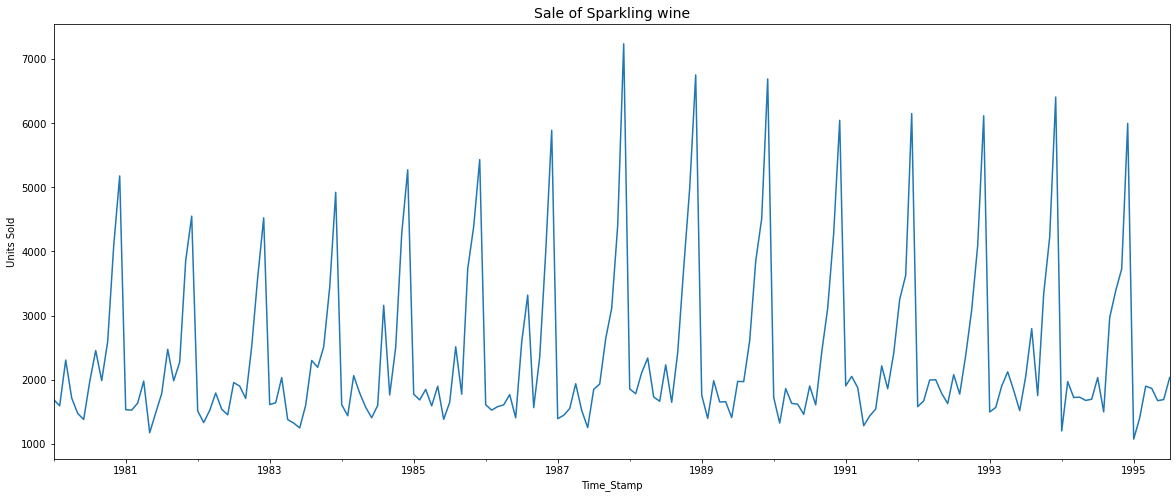


Figure :Sales of Sparkling plot

From the pot we can see there is no trend but seasonality is present. There is no missing value

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

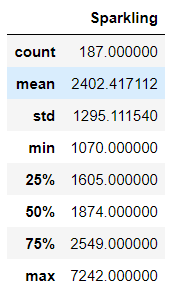


Figure :Sparkling description

The minimum sales was 1070 units of wine and maximum was 7242 units of wine through all the years.

There are no null values in the dataset.

Sparkling 0

dtype: int64

**Monthly Sales**

**Monthly Sales across years**

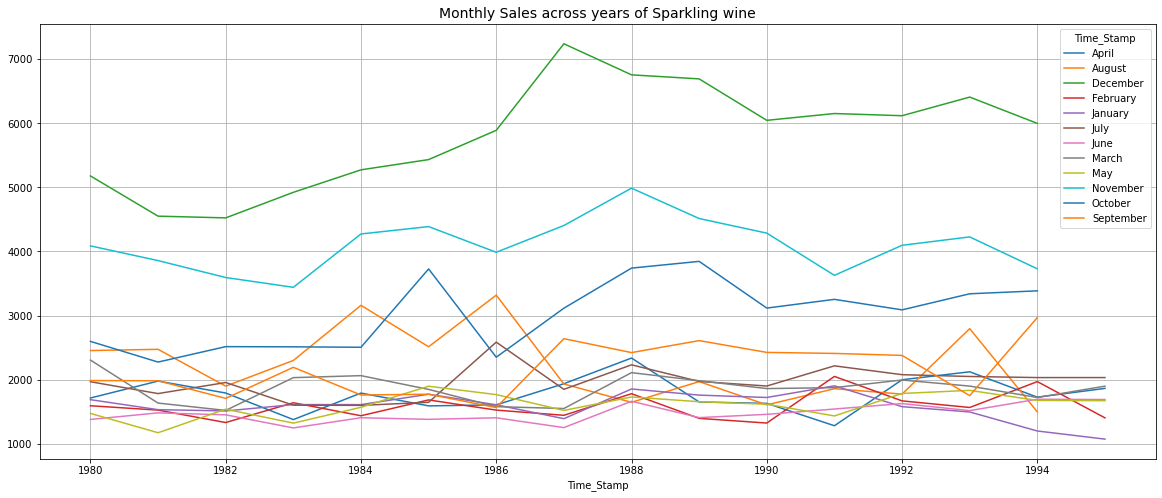
****

Figure :Sparkling Monthly sales over years

From the plot we can infer that December month has the highest sales across all years and November month has the second highest sales across most of the years.

**Monthly Sales Sum Barplot of Sparkling wine**

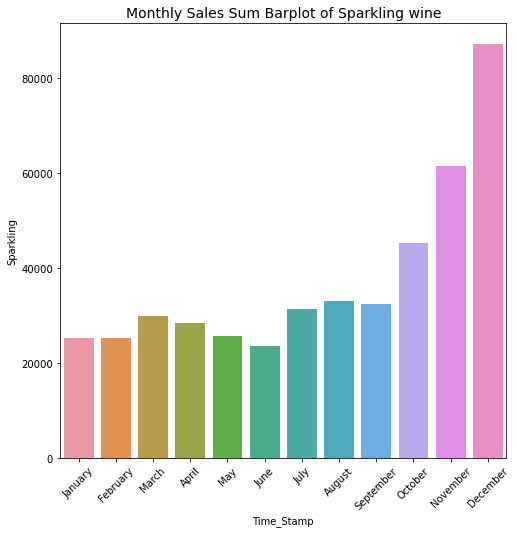


Figure :Sparkling Barplot Monthly sales

December has the highest sales of Sparkling wine with above 80000 units combining all the years. June has the lowest sales of Sparkling wine with slightly above 20000 units combining all the years.

**Boxplot of Monthly Sales of Sparkling wine**

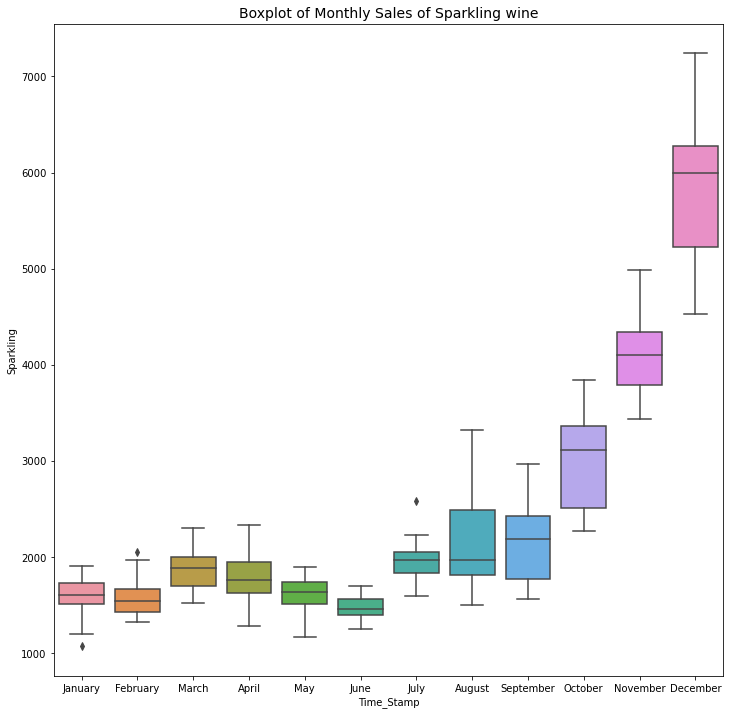
****

Figure :Sparkling Boxplot Monthly sales

**There is an increasing trend month wise with January has the sale with the least value of almost 1000.**

**Month Plot**

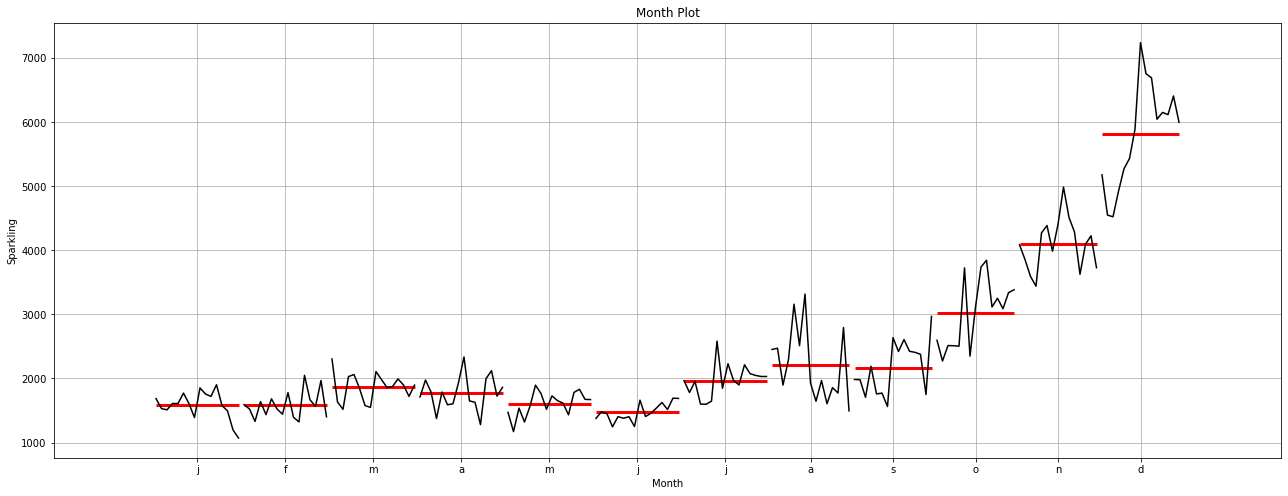
****

Figure :Sparkling Monthplot

The months are marked orderly in x-axis. There is an increasing trend with the months.

**Quarterly Sales**

**Quarterly Sales across years of Sparkling wine**

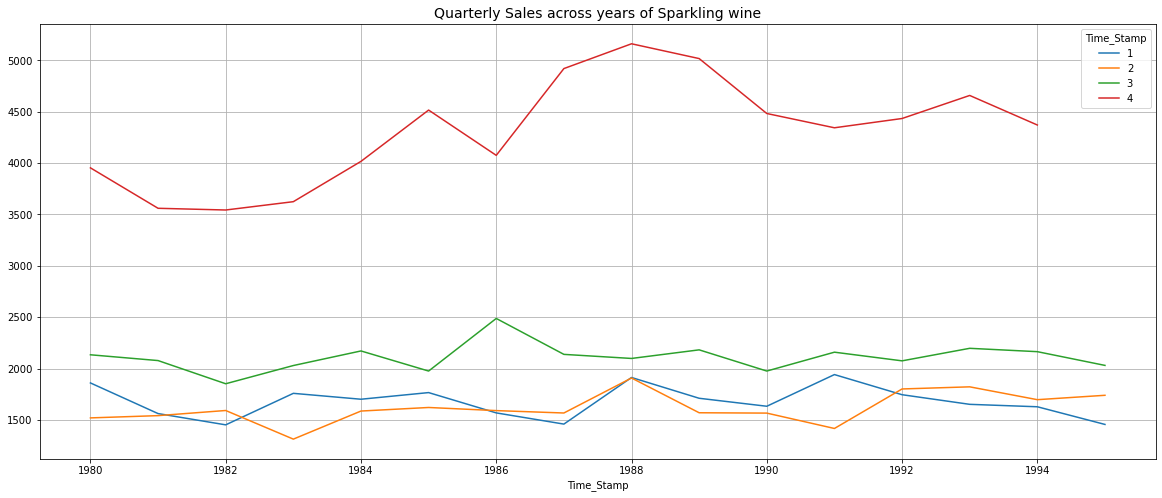
****

Figure :Sparkling Quarterly sales over years

**The fourth quarter has the most sales of wine throughout all the years. Third quarter has the second highest sales of wine throughout all the years**

**Barplot of Quarterly Sales Sum of Sparkling wine**

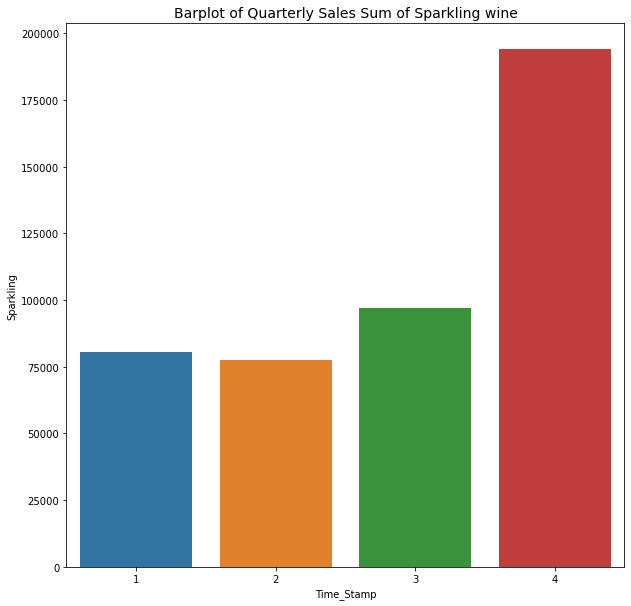
****

Figure :Sparkling Barplot Quarterly

4th quarter has the highest sales with nearly 190000 units of wine combining all the years. 2nd quarter has lesser sales than the first quarter combining all the years.

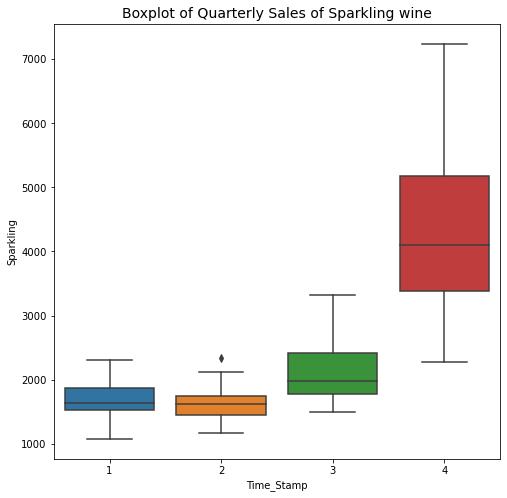
**Boxplot of Quarterly Sales of Sparkling wine**

Figure :Sparkling Quarterly Boxplot

**The sales are increasing with each subsequent quarter throughout all the years.**

### **Yearly Sales**

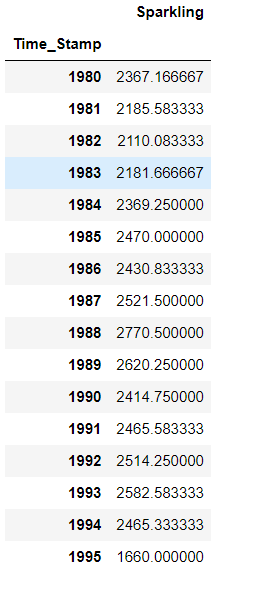


Figure :Sparkling Yearly Sales

Above are the average sales each year.

**Barplot of Yearly Sales Sum of Sparkling wine**

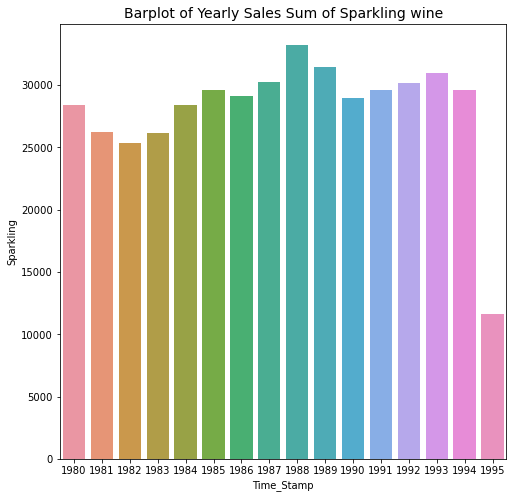
****

Figure :Sparkling Barplot Yearly

1988 has the highest sale of all the years.

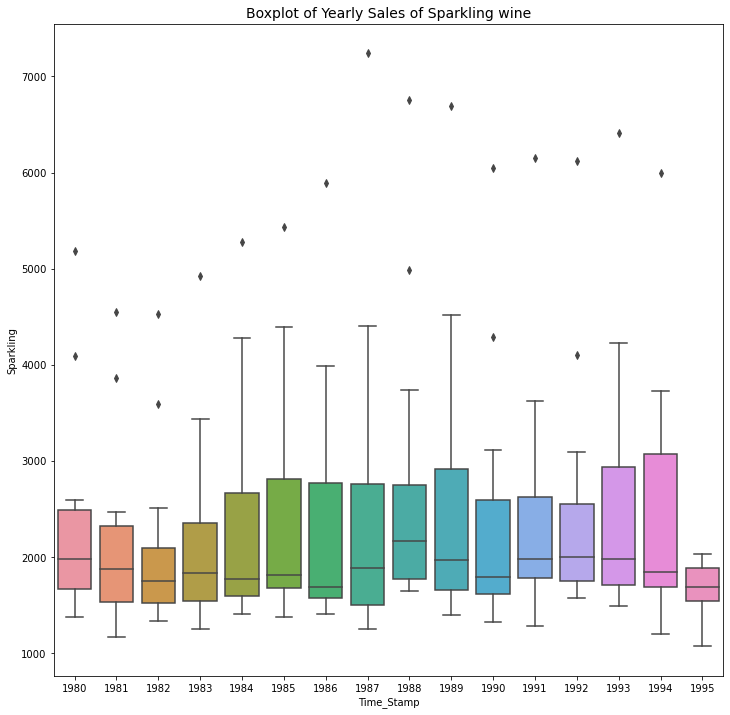
**Boxplot of Yearly Sales of Sparkling wine**

Figure :Sparkling Boxplot Yearly

**No trend is visible is seen in the data. 1995 has the lowest sale of wine.**

**Decomposition**

Additive Decomposition of Sparkling wine

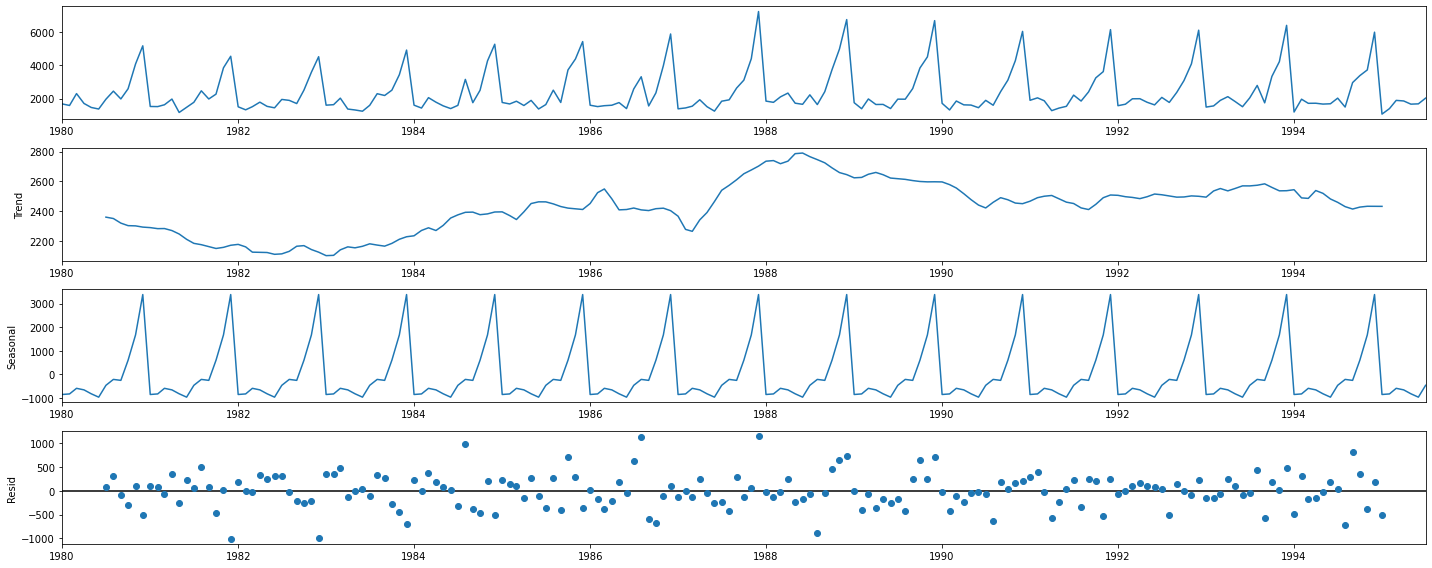
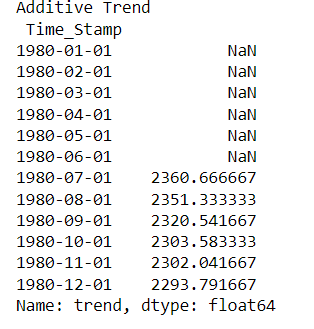
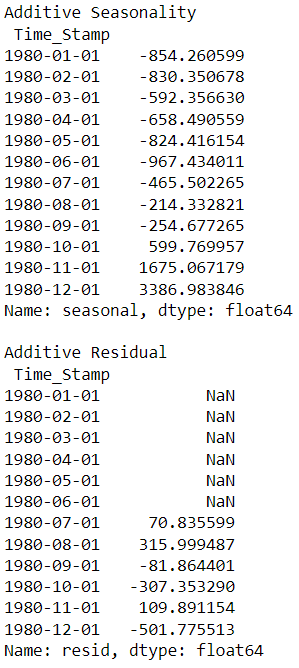
****

Figure :Sparkling Additive Decomp

We see that the residuals are located around 0 from the residual and patter is present. The residuals are ranging from -1000 to +1000.

so further decomposing to multiplicative model to minimize the residuals. There is seasonality and we don't observe pronounced trend.

The first 12 months trend, seasonality and residual values of Sparkling wine sales dataset are shown below:

Multiplicative Decomposition of Sparkling wine

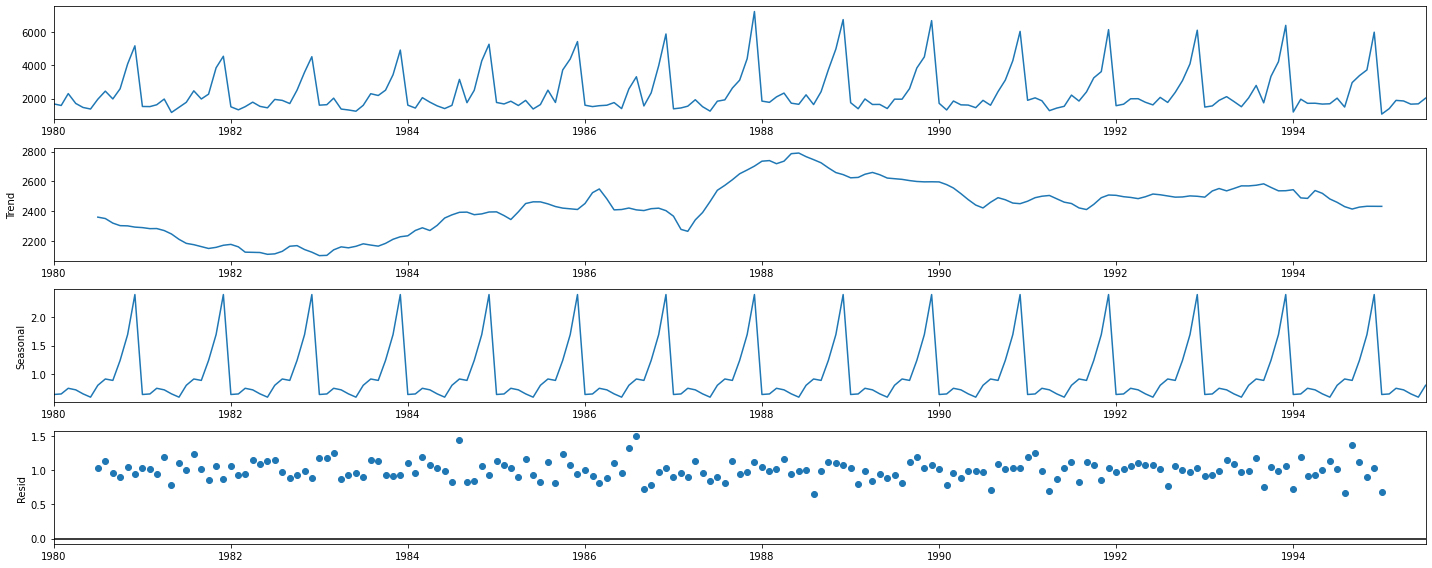
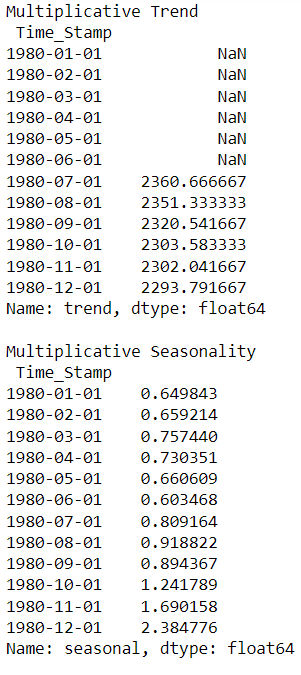
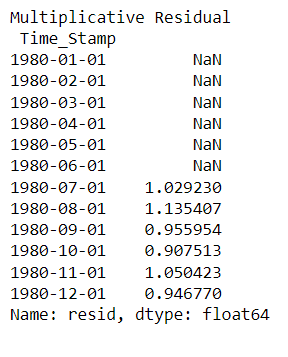


Figure :Sparkling Mulitpl Decomp

In the multiplicative decomposition a lot of residuals are located around 1 and it has pattern too. There is seasonality and we don't observe pronounced trend.

Since both additive and multiplicative has pattern in its residuals we can choose additive has the simpler and suitable decomposition method.

The first 12 months multiplicative trend, seasonality and residual values of Sparkling wine sales dataset are shown below:

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

The first 5 and last 5 rows of train data shows that it starts from January 1980 and ends at December 1990.

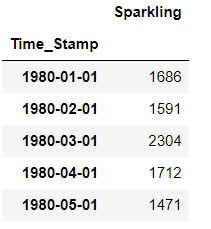
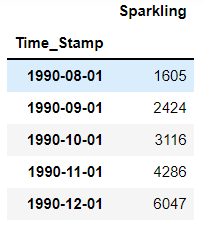
 

Table :Sparkling Train Data

The first 5 and last 5 rows of test data shows that it starts from January 1991 and ends at July 1995.

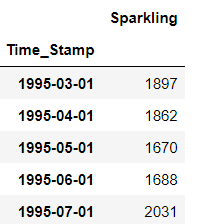
 

Table :Sparkling Test Data

**2.4. 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.**

**Simple Exponential Smoothing Auto Fit Model**

Simple Exponential Smoothing(SES) is a time series forecasting method for univariate data without a trend or seasonality. It requires a single parameter, called alpha (a), also called the smoothing factor or smoothing coefficient.

A SES model is built on train data with initialization\_method value as estimated and the following parameters with values.

initialization\_method - Method for initialize the recursions.

Use\_brute=True -> Search for good starting values using a brute force (grid) optimizer

**Optimized**=True -> Estimate model parameters by maximizing the log-likelihood.

{'smoothing\_level': 0.07029120765764557,

'smoothing\_trend': nan,

'smoothing\_seasonal': nan,

'damping\_trend': nan,

'initial\_level': 1764.0137060346985,

'initial\_trend': nan,

'initial\_seasons': array([], dtype=float64),

'use\_boxcox': False,

'lamda': None,

'remove\_bias': False}

The model built is used to forecast for next 55 months which is the length of

test data.

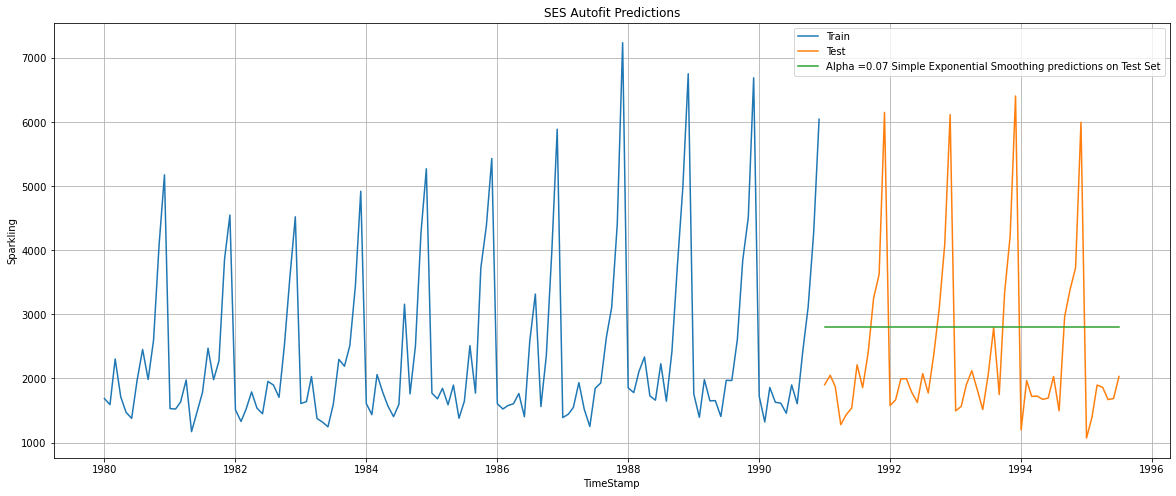


Figure :Sparkling SES plot

RMSE is calculated on test data.

SES Autofit RMSE on Test data: 1338.01

**Double Exponential Smoothing - Holt Autofit Model**

Double exponential smoothing(DES) employs a level component and a trend

component at each period.

A DES model is built on train data with initialization\_method value as estimated , use\_brute=True and Optimized=True. The other parameters and the values are:

{'smoothing\_level': 0.6649999999999999, 'smoothing\_trend': 0.0001, 'smoothing\_seasonal': nan, 'damping\_trend': nan, 'initial\_level': 1502.1999999999991, 'initial\_trend': 74.87272727272739, 'initial\_seasons': array([], dtype=float64), 'use\_boxcox': False, 'lamda': None, 'remove\_bias': False}

In [381]:

The model built is used to forecast for next 55 months which is the length of

test data.

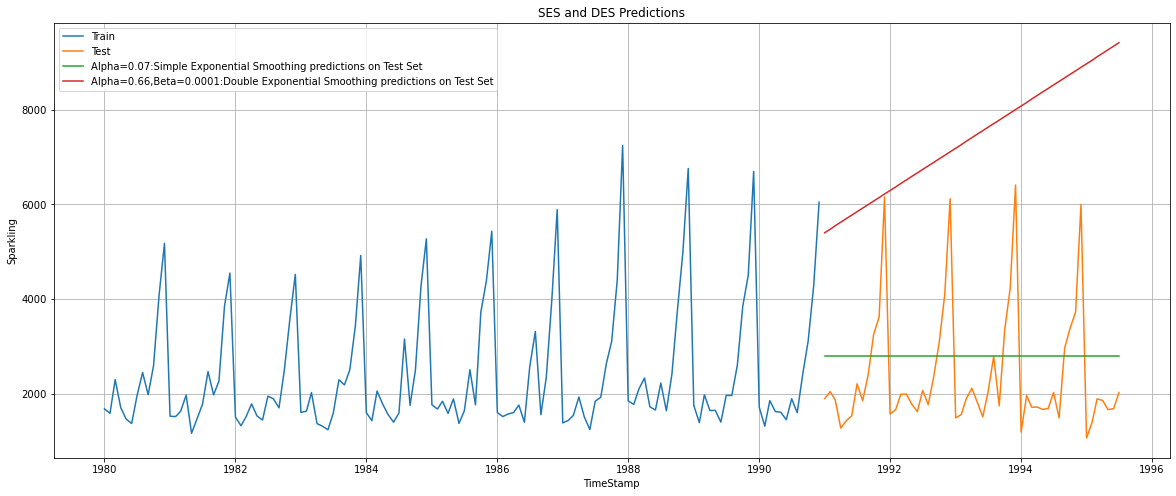
****

Figure :Sparkling DES plot

RMSE is calculated on test data.

DES Autofit RMSE on Test Data: 5291.88

**Triple Exponential Smoothing - ETS(A, A, A) - Holt Winter's linear method with additive errors Autofit Model**

Triple exponential smoothing(TES) is used to handle the time series data containing a seasonal component. This method is based on three smoothing equations: stationary component, trend, and seasonal. Both seasonal and trend can be additive or multiplicative

In this model both trend and seasonality is chosen as additive. A DES model is built on train data with initialization\_method value as estimated , use\_brute=True and Optimized=True. The other parameters and the values are:

{'smoothing\_level': 0.11127227248079453, 'smoothing\_trend': 0.012360804305088534, 'smoothing\_seasonal': 0.46071766688111543, 'damping\_trend': nan, 'initial\_level': 2356.577980956387, 'initial\_trend': -0.10243675533021725, 'initial\_seasons': array([-636.23319334, -722.9832009 , -398.64410813, -473.43045416,

-808.42473284, -815.34991402, -384.23065038, 72.99484403,

-237.44226045, 272.32608272, 1541.37737052, 2590.07692296]), 'use\_boxcox': False, 'lamda': None, 'remove\_bias': False}

In [386]:

The model built is used to forecast for next 55 months which is the length of

test data.

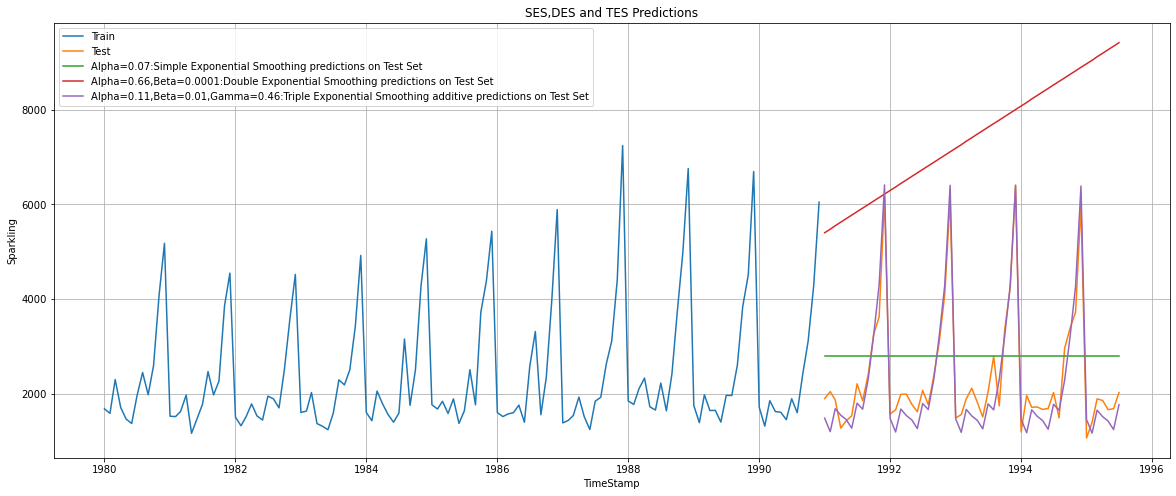


Figure :Sparkling TES add seas plot

RMSE is calculated on test data.

TES Additive Autofit RMSE on Test data: 378.95

**Triple Exponential Smoothing - ETS(A, A, M) - Holt Winter's linear method with multiplicative errors**

In this model trend is additive and seasonality is chosen as multiplicative. A DES model is built on train data with initialization\_method value as estimated , use\_brute=True and Optimized=True. The other parameters and the values are:

The model built is used to forecast for next 55 months which is the length of

test data.

{'smoothing\_level': 0.11133818361298699, 'smoothing\_trend': 0.049505131019509915, 'smoothing\_seasonal': 0.3620795793580111, 'damping\_trend': nan, 'initial\_level': 2356.4967888704355, 'initial\_trend': -10.187944726007238, 'initial\_seasons': array([0.71296382, 0.68242226, 0.90755008, 0.80515228, 0.65597218,

0.65414505, 0.88617935, 1.13345121, 0.92046306, 1.21337874,

1.87340336, 2.37811768]), 'use\_boxcox': False, 'lamda': None, 'remove\_bias': False}

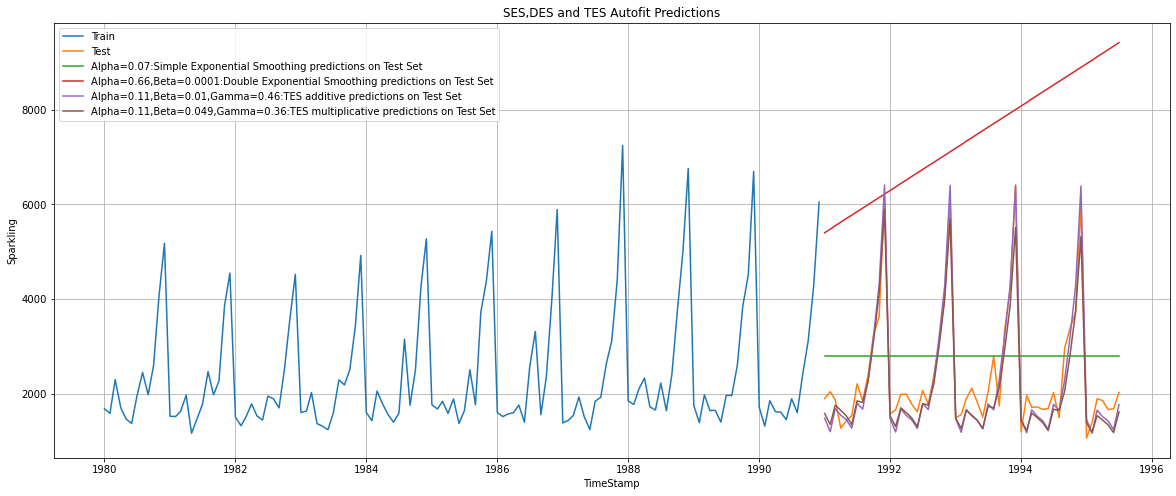
****

Figure :Sparkling TES multipl seas plot

RMSE is calculated on test data.

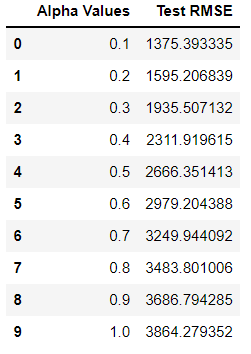
TES multiplicative autofit RMSE on Test data : 404.29

Both the Triple Exponential Smoothing models are picking up the seasonal component as well

which can be inferred from the graph.

## **Iterative Method for Simple Exponential Smoothing**

A SES model is built on train data with initialization\_method value as estimated , use\_brute=True and Optimized=False. The value of smoothing\_level(alpha) is taken from 0.1 to 1 and its corresponding RMSE value is calculated as shown below.



We can see that alpha as 0.1 as the least RMSE.

Plot the prediction of the model with smoothing level value as 0.1.

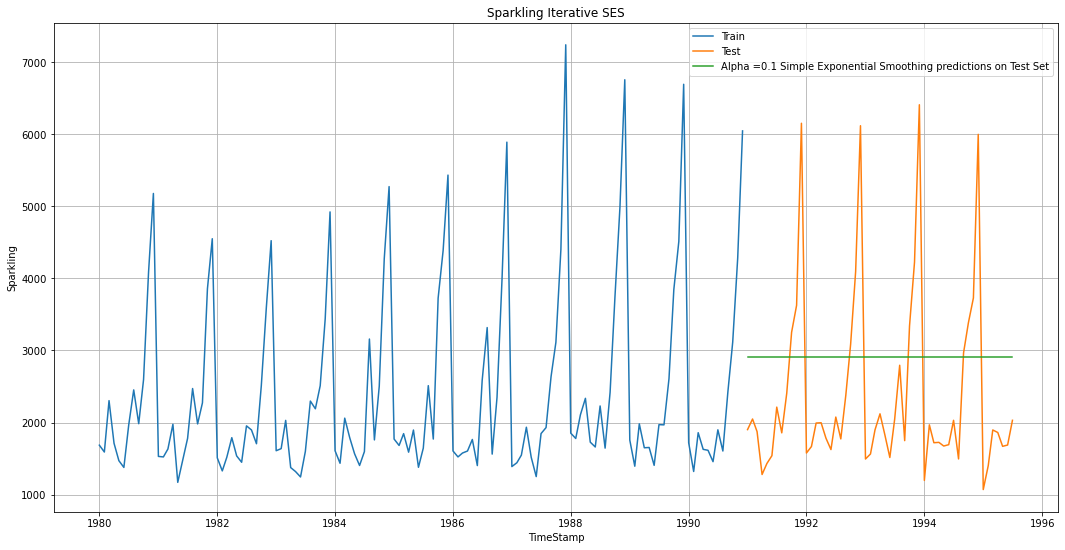
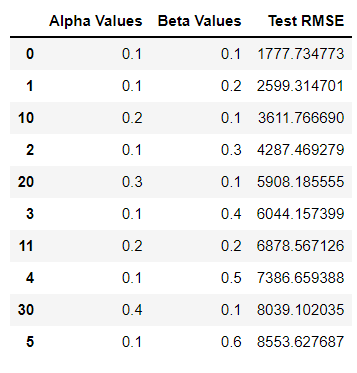


Figure :Sparkling Iterative SES

Iterative SES RMSE on Test data : 1375.39

**Iterative Method for Double Exponential Smoothing**

A DES model is built on train data with initialization\_method value as estimated , use\_brute=True and Optimized=False. The values of smoothing\_level(alpha) and smoothing\_trend(beta) is taken from 0.1 to 1 and its corresponding RMSE value is calculated .The first 10 rows with least RMSE is shown below.



The model with smoothing level as 0.1 and smoothing trend as 0.1 has the least RMSE value 1777.73

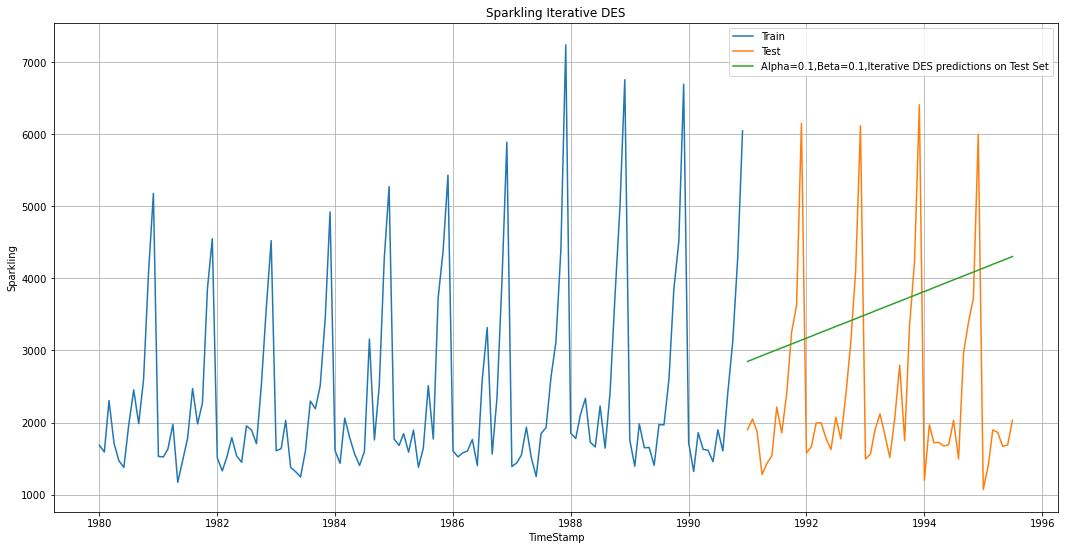
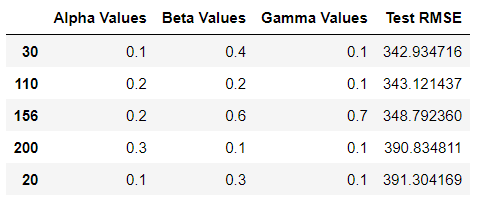
****

Figure :Sparkling Iter DES plot

### Iterative SES RMSE on Test data : 1777.73

### **Iterative Method - Triple Exponential Smoothing - ETS(A, A, A)**

A TES model with trend and seasonality as additive is built on train data with initialization\_method value as estimated , use\_brute=True and Optimized=False. The values of smoothing\_level(alpha) ,smoothing\_trend(beta) and smoothing\_seasonal(gamma)is taken from 0.1 to 1 and its corresponding RMSE value is calculated. First 5 rows of the data which shows the 5 least RMSE ones are shown below.



The model with smoothing level as 0.1,smoothing trend as 0.4 and smoothing seasonal 0.1 has the least RMSE value 342.93

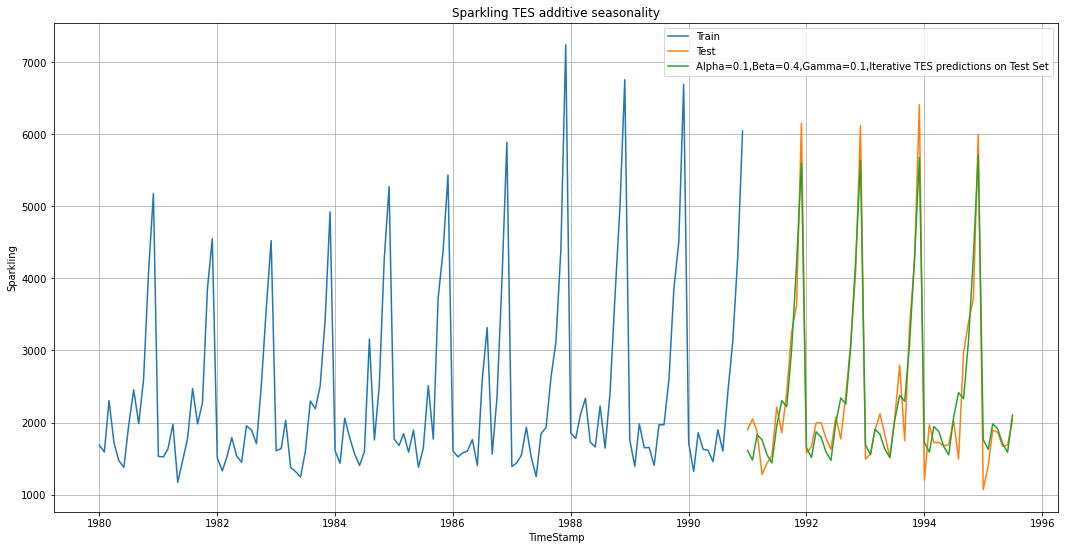
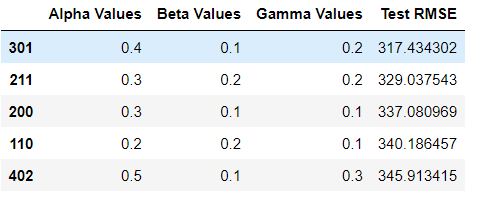


Figure :Sparkling-Iter TES add seas

Iterative TES Additive RMSE on Test data : 342.93

**Iterative Method - Triple Exponential Smoothing - ETS(A, A, M)**

A TES model with trend as additive and seasonality as multiplicative is built on train data with initialization\_method value as estimated , use\_brute=True and Optimized=False. The values of smoothing\_level(alpha) ,smoothing\_trend(beta) and smoothing\_seasonal(gamma)is taken from 0.1 to 1 and its corresponding RMSE value is calculated. First 5 rows of the data which shows the 5 least RMSE ones are shown below.



The model with smoothing level as 0.4,smoothing trend as 0.1 and smoothing seasonal 0.2 has the least RMSE value 317.43

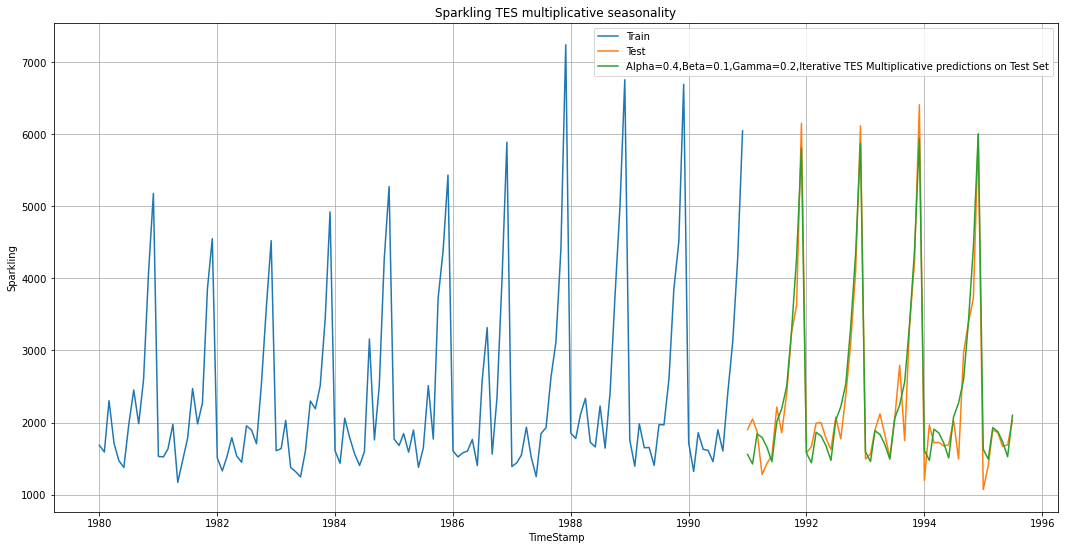
****

Figure :Sparkling Iterative TES multipl seas plot

### The RMSE of iterative TES multiplicative is 317.43

### **Linear Regression Model**

Linear regression uses the relationship between the data-points to draw a straight line through all them. This line can be used to predict future values.

The training and testing time instance is created.

Training Time instance

[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132]

Test Time instance

[133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187]

The instances are added to the train and test data as column ‘time’ and treated as an independent variable.

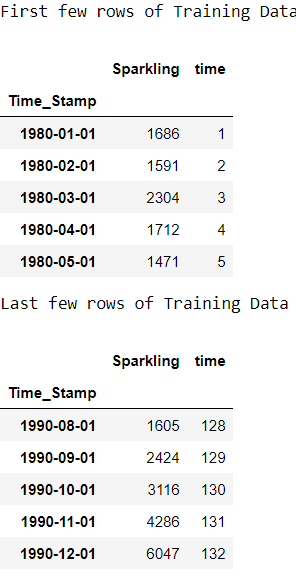
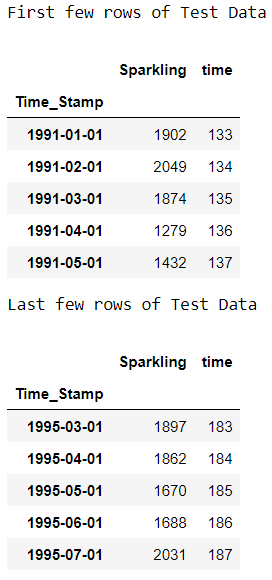
 

Table :Sparkling LR Train Data

Sparkling column is treated as a dependent variable. Both the train independent and dependent variable are fitted on the Linear Regression model from sklearn library.

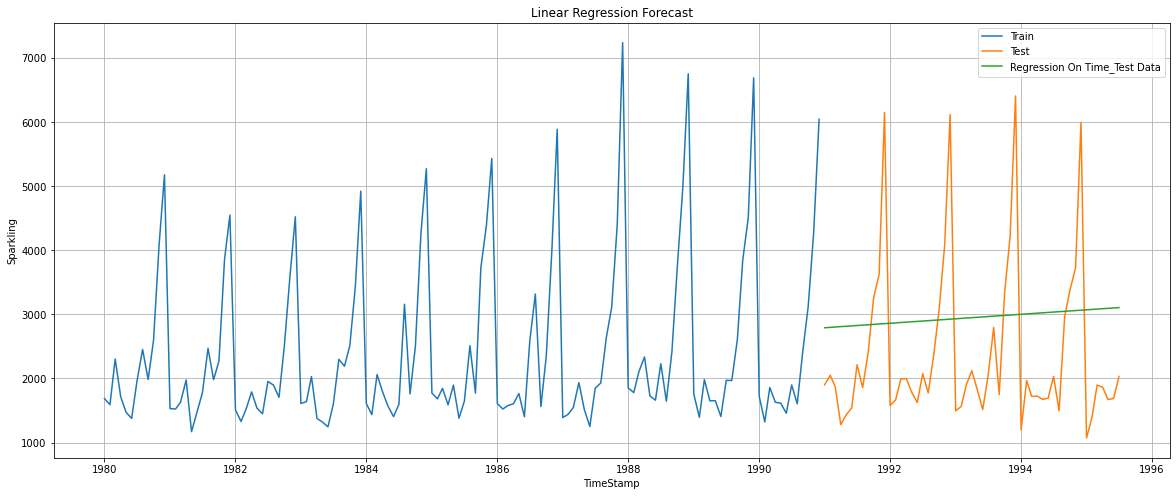


Figure :Sparkling-Linear Regr plot

The fitted model is sued to predict on test data and the output of Rose wine

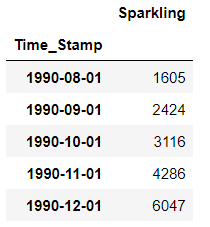
sales is used to calculate RMSE along with actual Rose wine sales.

### The RMSE of Linear Regression forecast on the Test Data is 1389.14

### **Naive Forecast Model:**

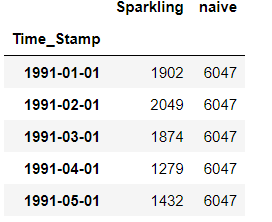
Naïve forecasting is the technique in which the last period's sales are used for the next period's forecast without predictions or adjusting the factors. Forecasts produced using a naïve approach are equal to the final observed value.

The last 5 rows of the train data are:



The last value is 6047 which is treated as the forecasted value for the future time period as shown below.

The first 5 rows of test data are:



The last 5 rows of test data are:

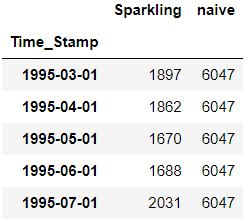


Table :Naive Forecast Value Test

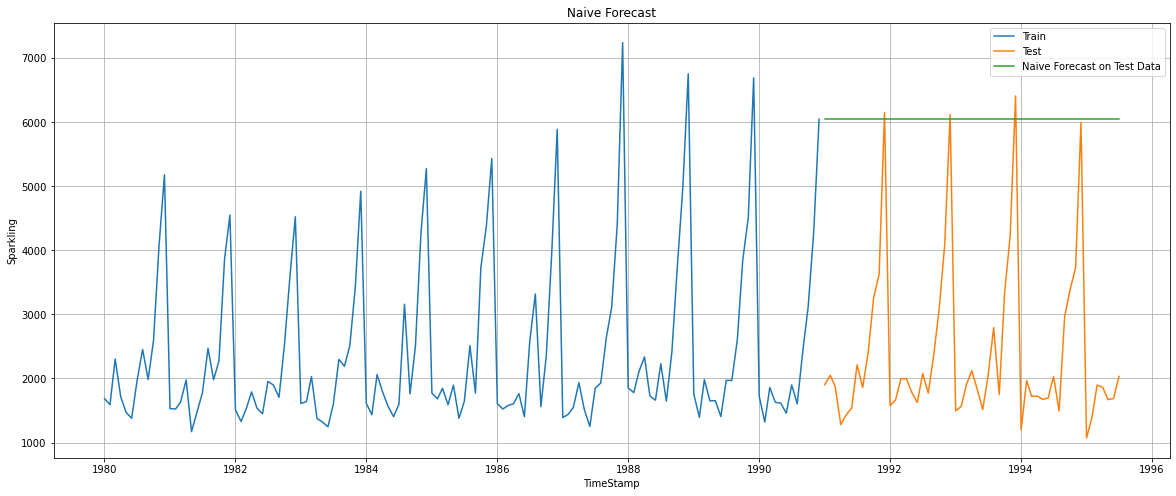


Figure :Sparkling-Naive Forecast plot

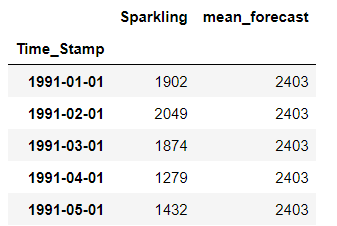
The predicted test output along with the actual test output is used for calculating RMSE.

### The RMSE for NaiveModel forecast on the Test Data 3864.28

### **Simple Average Model**

### In the Simple Average model ,forecast is equal to the average of historical data.

The average of the train data is taken and added as the forecast of Test data after converting into an integer number(mean -> 2403.780303030303 ) as shown below.



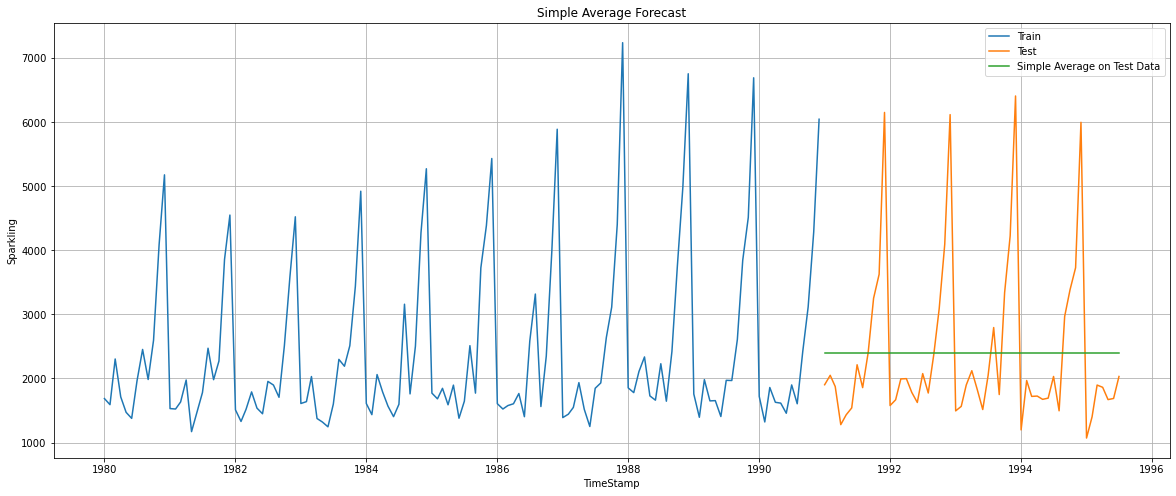


Figure :Sparkling-SA plot

The predicted test output along with the actual test output is used for calculating RMSE.

The RMSE for Simple Average forecast on the Test Data 1275.08.

**Moving Average Forecast Model**

Moving Average Forecast Model takes an average of a set of numbers in a given range while moving the range.

Moving Average is calculated on train data for which the following values are given to rolling function.

2-takes moving average of 2 months of data

3-takes moving average of 3 months of data

6-takes moving average of 6 months of data

12-takes moving average of 12 months of data

The first 5 rows of training data is the following:

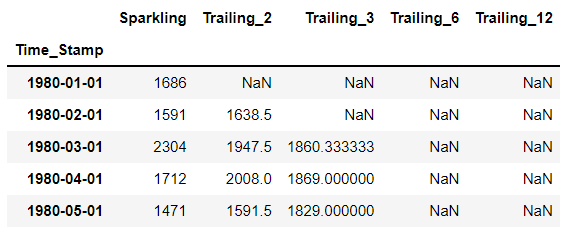


Figure :Sparkling-MA-Train head

The last 5 rows of training data is the following:

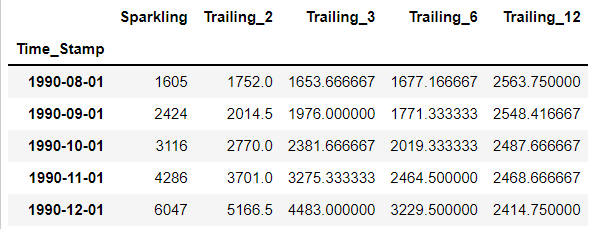


Figure :Sparkling-MA-Train tail

The value 5166 is taken as the forecasted value for the test for 2 point Moving Average. ( 5166.5 rounded off)

The value 4483 is taken as the forecasted value for the test for 3 point Moving Average.

The value 3229 is taken as the forecasted value for the test for 6 point Moving Average. ( 3229.500000 rounded off)

The value 2414 is taken as the forecasted value for the test for 12 point Moving Average. (2414.750000 rounded off)

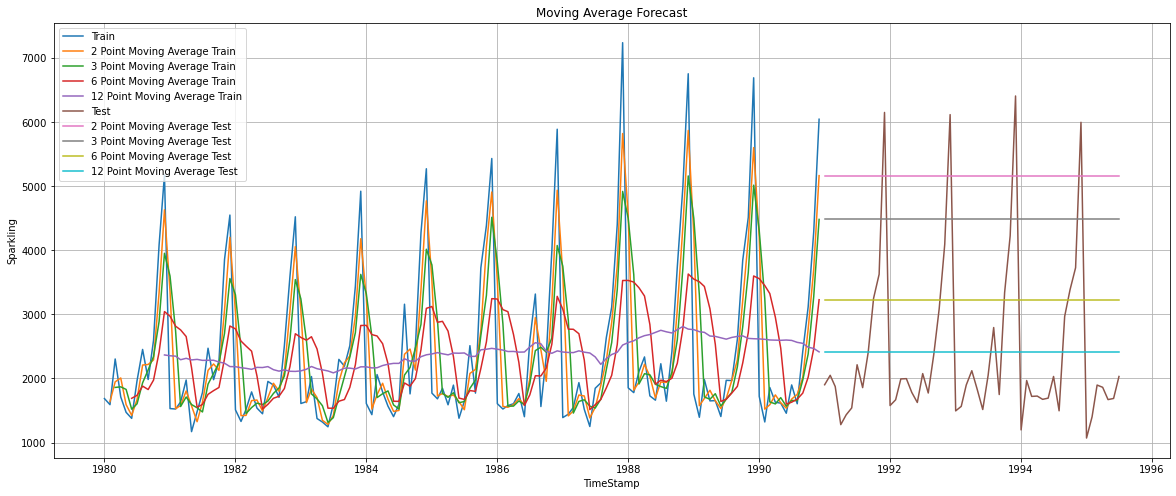


Figure :Sparkling-Mov Avg-plot

The RMSE for 2 point Moving Average Model forecast on the Test Data is

3046.52

The RMSE for 3 point Moving Average Model forecast on the Test Data is

2443.0

The RMSE for 6 point Moving Average Model forecast on the Test Data is

1521.34

The RMSE for 12 point Moving Average Model forecast on the Test Data is

1275.16

**2.5. 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.**

Augmented Dickey Fuller test (ADF Test) is a common statistical test used to

test whether a given Time series is stationary or not. It is one of the most

commonly used statistical test when it comes to analyzing the stationary of a

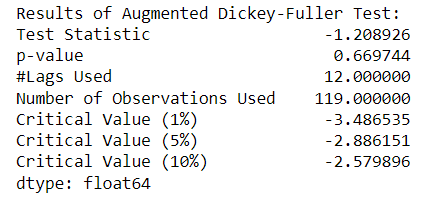
series.

The hypothesis for the statistical test is:

H0-Null Hypothesis: Time series is non-stationary

H1-Alternate Hypothesis: Time series is stationary

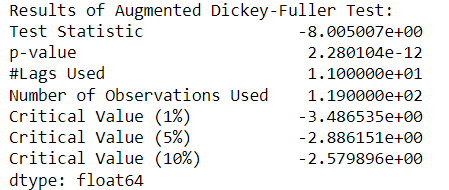
The ADF Test is conducted on the train data



The p-value obtained by the test should be less than the significance level (say 0.05) to reject the Null hypothesis or it fails to reject the Null hypothesis.

p value obtained from the ADF test is 0.6697 which is greater than 0.05 . Hence we fail to reject the Null Hypothesis and so we can say that data is non-stationary.

To convert the data into a stationary one, the difference of a Dataframe value with the value in the previous row is taken and remove missing values. The ADF Test is taken again on the modified train data.



After modifying, the p-value 2.280104e-12 obtained by the test is less than 0.05. Now the data has been converted into a stationary one.

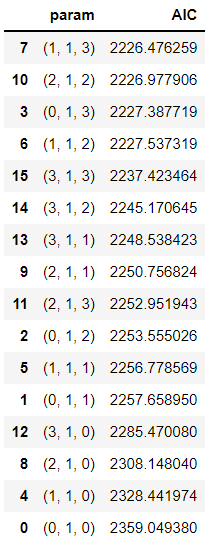
**2.6. Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.**

**ARIMA Automated Train Model**

Auto Regressive Integrated Moving Average (ARIMA) models are applied on time series data when the current value is assumed to be correlated to past values and past prediction errors. Therefore, these models are used in defining current value as a linear combination of past values and past prediction errors. Here, we have defined a few terms that would be useful in understanding ARIMA models in detail. ARIMA models can only be applied only on stationary time series data.

The Akaike information criterion (AIC) is an estimator of out-of-sample prediction error and thereby relative quality of statistical models for a given set of data. The least the AIC the better the model is .

The p,q value is taken from 0 to 3.‘d’ is taken as 1.



The model has the parameter ‘order’ which has its values in the form of (p,d,q) where

p: Trend autoregression order.

d: Trend difference order.

q: Trend moving average order.

The p value is taken as 1, q as 2 and d as 1.(1,1,2) as it has the least AIC value.

ARIMA is built using stationary data after dropping its NA values since it reduces the AIC value .

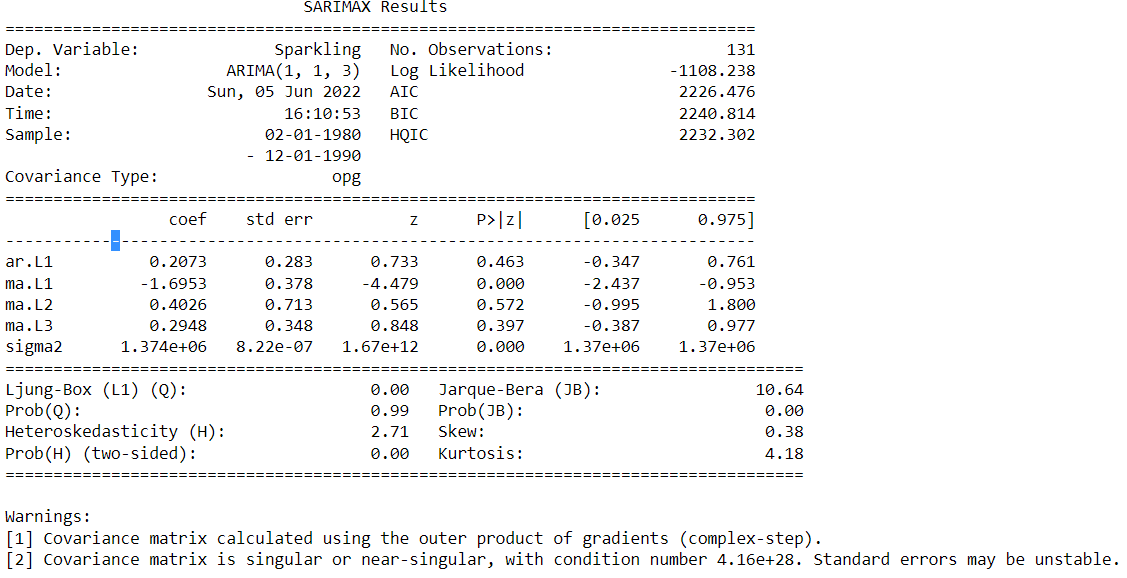
enforce\_stationarity🡪Whether or not to transform the AR parameters to enforce stationarity in the autoregressive component of the model.

enforce\_invertibility🡪Whether or not to transform the MA parameters to enforce invertibility in the moving average component of the model.

enforce\_stationarity and enforce\_ invertibility is given as false.

The ARIMA model is built with those values and RMSE is calculated on test data.

The Summary of ARIMA model is



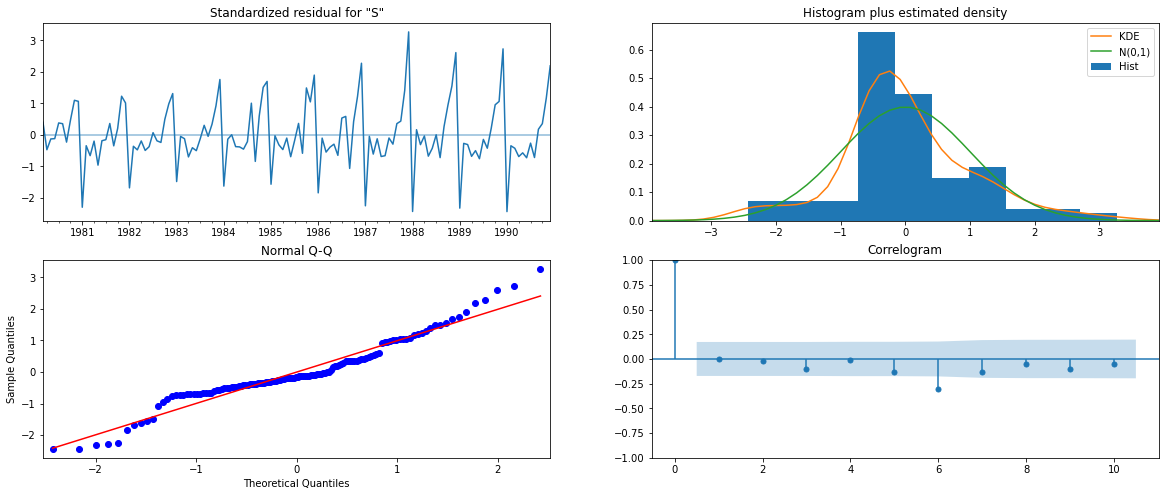


Figure :Sparkling Automated ARIMA diagnostics

The diagnostics look good here.

RMSE of Automated ARIMA model on Test data is: 2763.74

**SARIMA Automated Train Model**

Seasonal Autoregressive Integrated Moving Average, SARIMA or Seasonal ARIMA, is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component.

It adds three new hyperparameters to specify the autoregression (AR), differencing (I) and moving average (MA) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality.

**Autocorrelation Function (ACF)**

A plot of auto-correlation of different lags is called ACF.The plot summarizes the correlation of an observation with lag values. The x-axis shows the lag and the y-axis shows the correlation coeﬃcient between -1 and 1 for negative and positive correlation.

**Partial Autocorrelation Function (PACF)**

A plot of partial auto-correlation for different values of lags is called PACF.

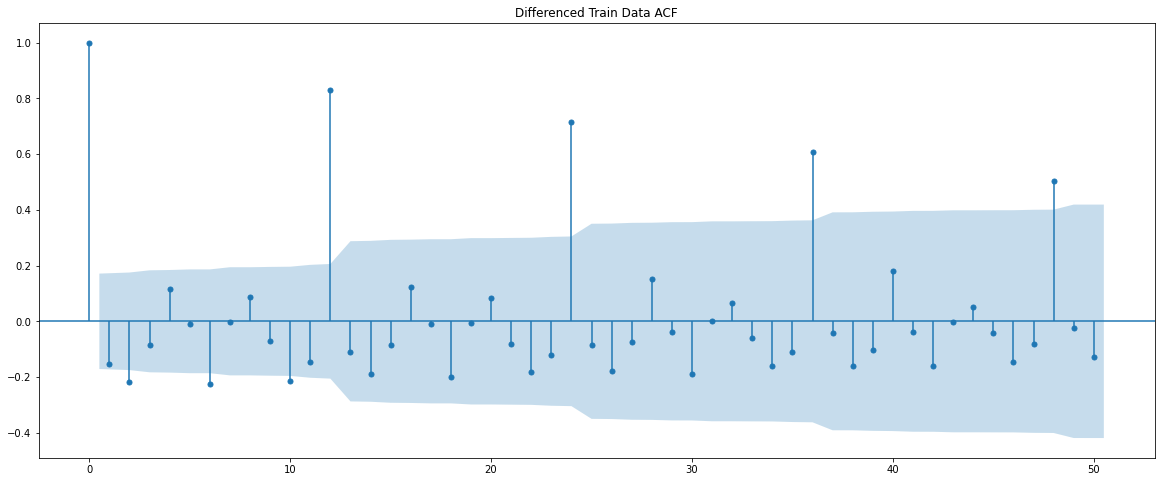


Figure :Sparkling-Automated SARIMA-ACF

The model has the parameter ‘seasonal order’ which has its values in the form of (P, D, Q, s):

P: Seasonal autoregressive order.

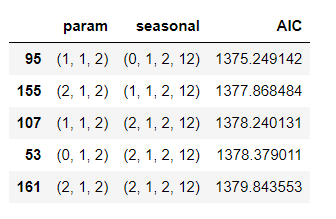
D: Seasonal difference order.

Q: Seasonal moving average order.

s: The number of time steps for a single seasonal period.

’s’ is determined from acf plot

The P,Q value is taken from 0 to 2. D is as 0 and 1. ‘d’ is taken as 1. From the above ACF plot we can say that ,Seasonality after every 12th lag is visible. We will run our auto SARIMA models by setting seasonality as 12. SARIMA is built using stationary data after dropping its NA values since it reduces the AIC value.



The parameters in the first row has the least AIC so its taken to build the SARIMA model. P is taken as 1, d as 1, q as 1, P as 0, D as 1, Q as 2 and s as 12.

SARIMA is built using stationary data after dropping its NA values since it reduces the AIC value.

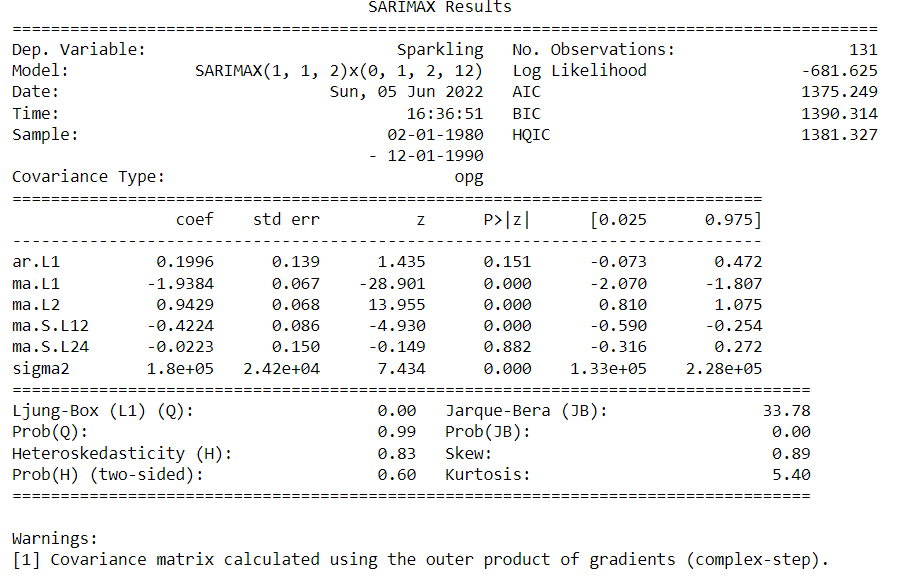
enforce\_stationarity🡪Whether or not to transform the AR parameters to enforce stationarity in the autoregressive component of the model.

enforce\_invertibility🡪Whether or not to transform the MA parameters to enforce invertibility in the moving average component of the model.

enforce\_stationarity and enforce\_ invertibility is given as false.

The SARIMA model is built with those values and RMSE is calculated on test data.

The Summary of SARIMA model built is:



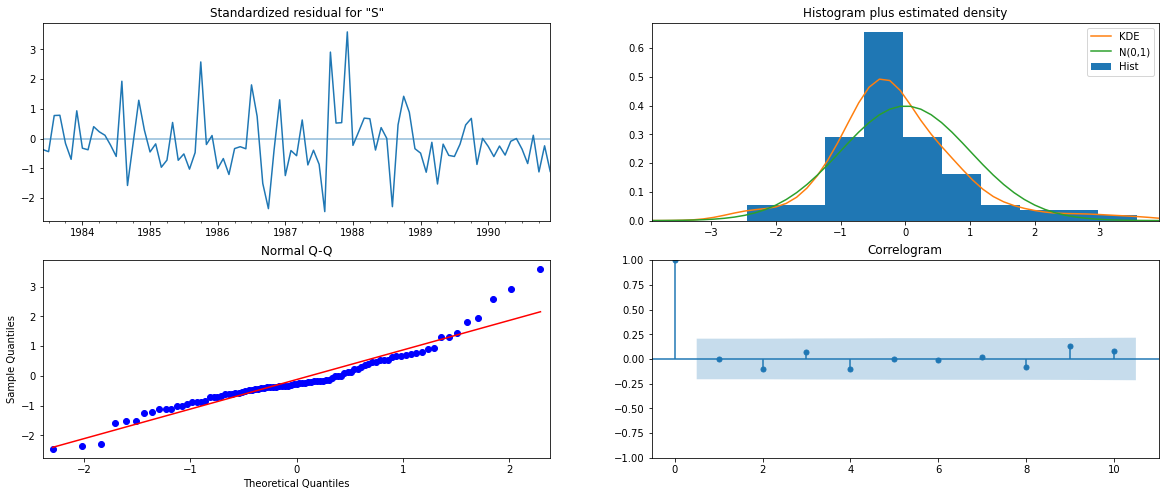


Figure :Sparkling-Automated SARIMA diagnostics

The diagnostics look good here.

RMSE of Automated SARIMA model on Test data is: 2866.01

**2.7. 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.**

**Manual ARIMA Model**

ACF is used for identifying the value of q and PACF is used for identifying the value of 𝑝.

The p value of the Augmented Dickey-Fuller Test on Train data: 0.669744

‘p’ is not less than 0.05 so the data isnt stationary.

The p value of the Augmented Dickey-Fuller Test on Train first difference data: 2.280104355826159e-12.

The data after first difference is stationary as the p value is less than 0.05. Thus we can consider the value of d as 1.

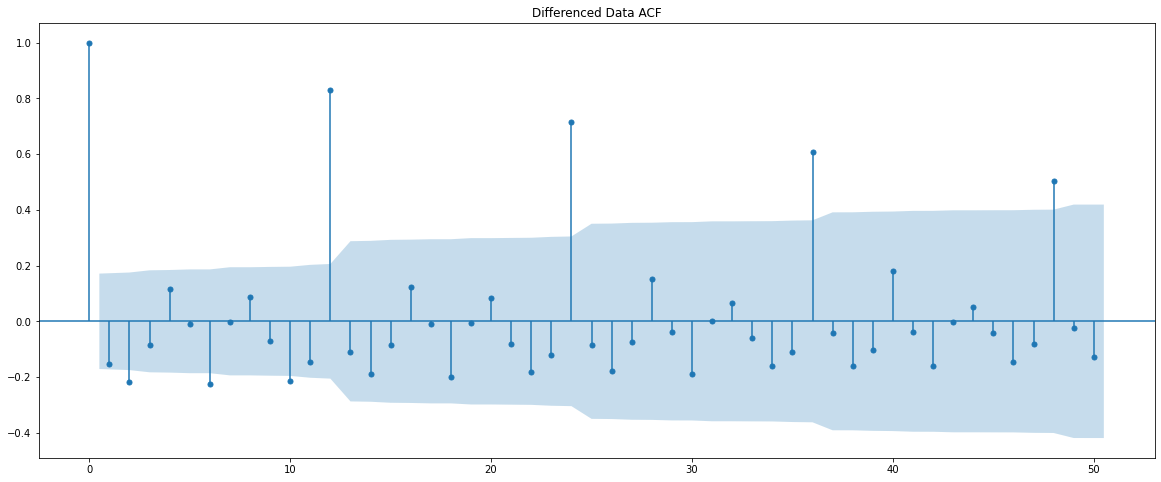
****

Figure :Sparkling-Manual ARIMA ACF

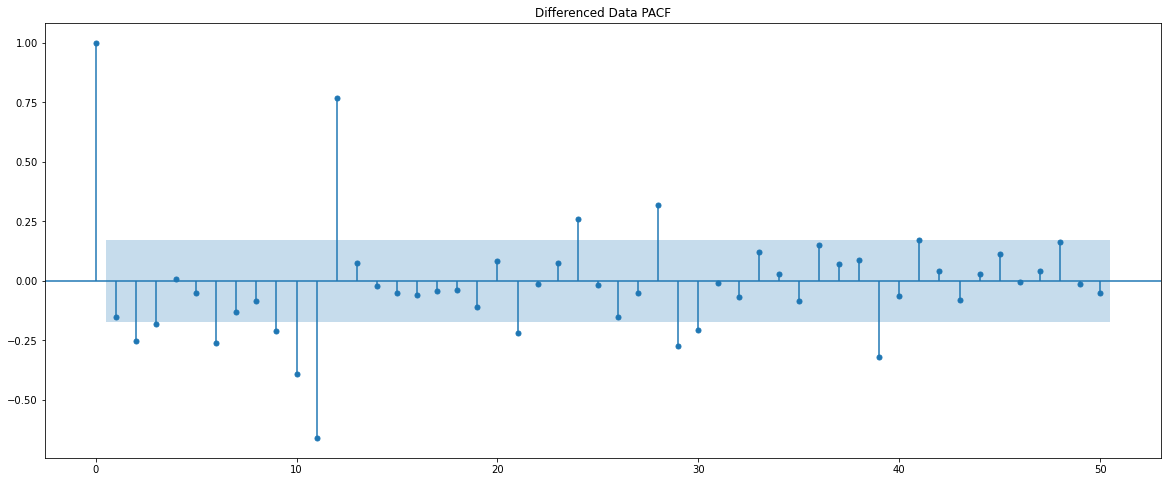


Figure :Sparkling-Manual ARIMA-PACF

From the ACF and PACF models we can take the value of p and q as 0. Since difference of order 1 has been taken on the data to make it stationary ,d=1.

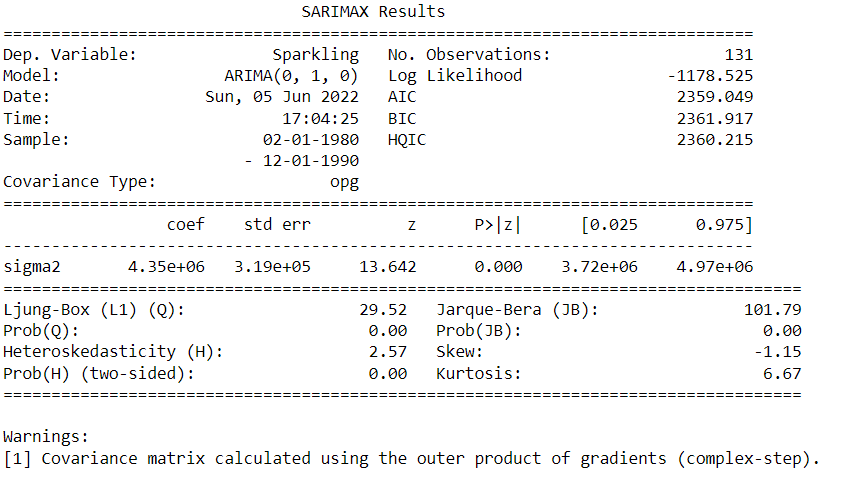
enforce\_stationarity🡪Whether or not to transform the AR parameters to enforce stationarity in the autoregressive component of the model.

enforce\_invertibility🡪Whether or not to transform the MA parameters to enforce invertibility in the moving average component of the model.

enforce\_stationarity and enforce\_ invertibility is given as false.

ARIMA is built using stationary data after dropping its NA values .The ARIMA model is built with those values and RMSE is calculated on test data.

The summary of Manual ARIMA model is:



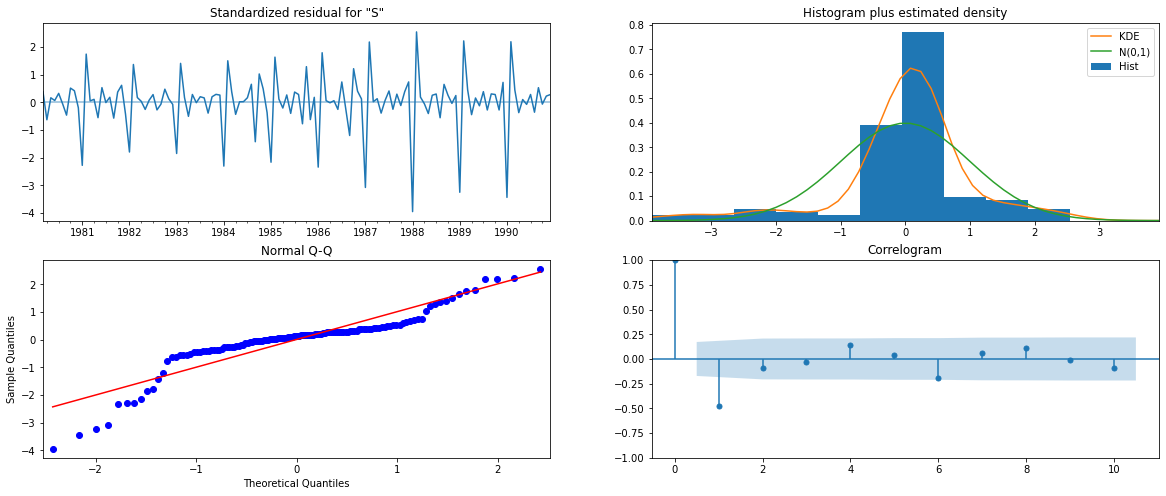


Figure :Sparkling-Manual ARIMA diagnostics

The diagnostics is okay here.

RMSE of Manual ARIMA model on Test data is: 1425.85

**Manual SARIMA Model:**

Since the seasonality parameter is 12 we can plot the graph for the difference of order 12 with NA values dropped.

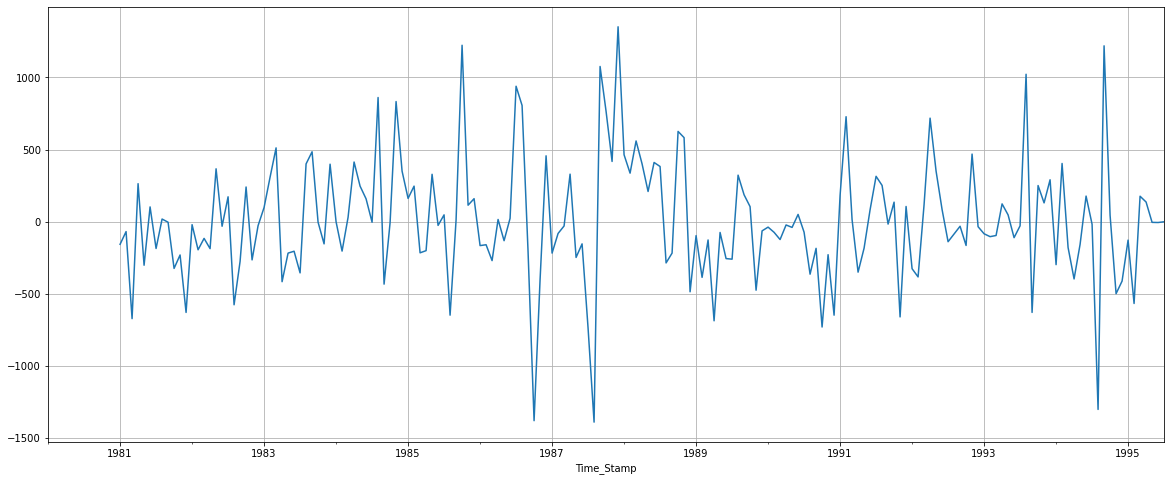
****

Figure :Sparkling-Manual SARIMA plot 1

As there is a slight trend in the graph we can take a differencing of first order on the seasonally differenced series .

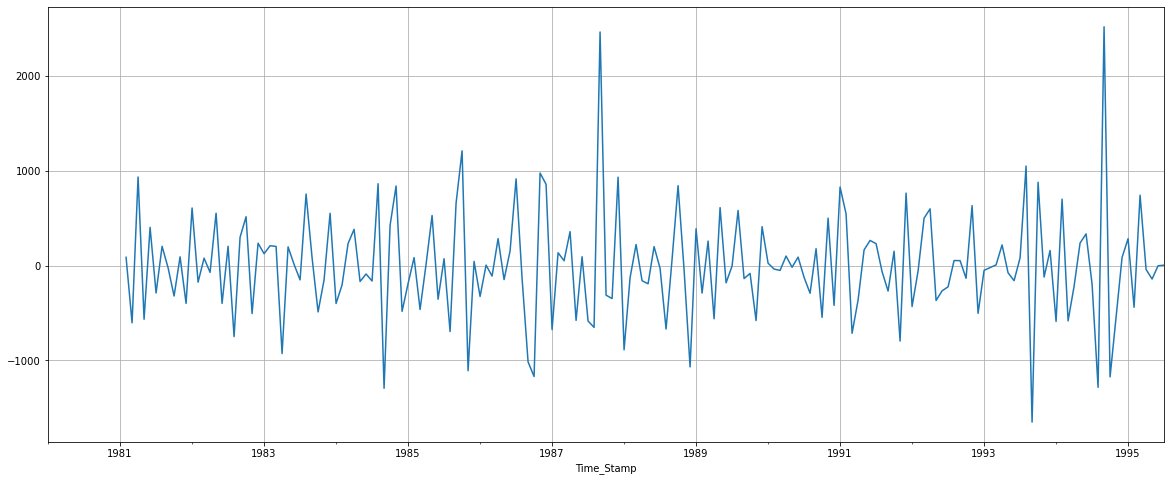
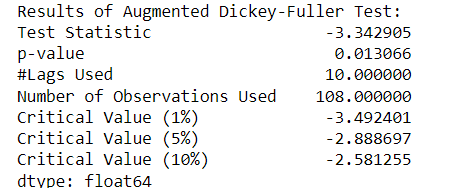


Figure :Sparkling-Manual SARIMA plot 2

Now we see that there is almost no trend present in the data. Seasonality is only present in the data.

Let us go ahead and check the stationarity of the above series before fitting the SARIMA model.



The first difference of seasonal differenced train data is used to plot ACF and PACF as below:

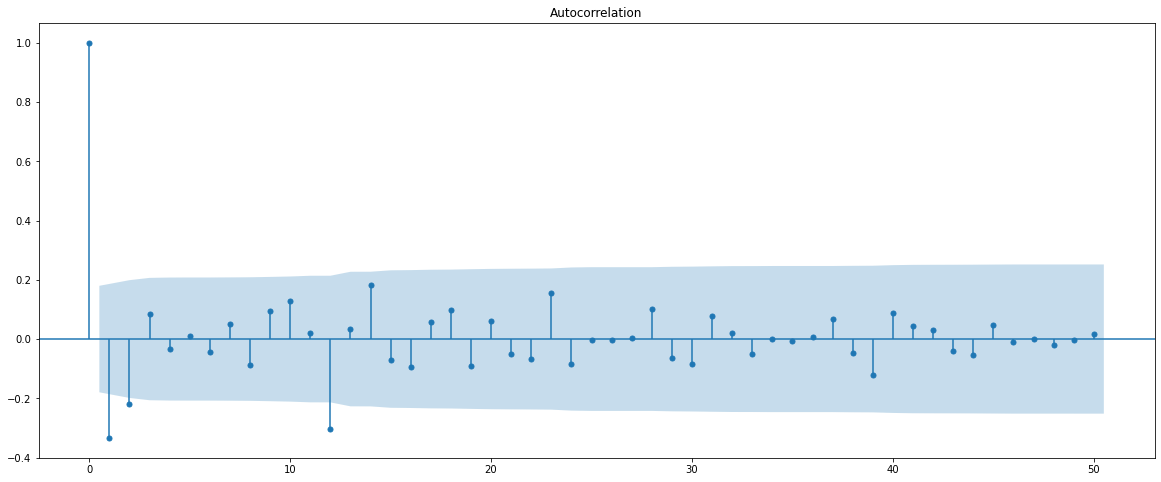


Figure :Sparkling-Manual SARIMA ACF

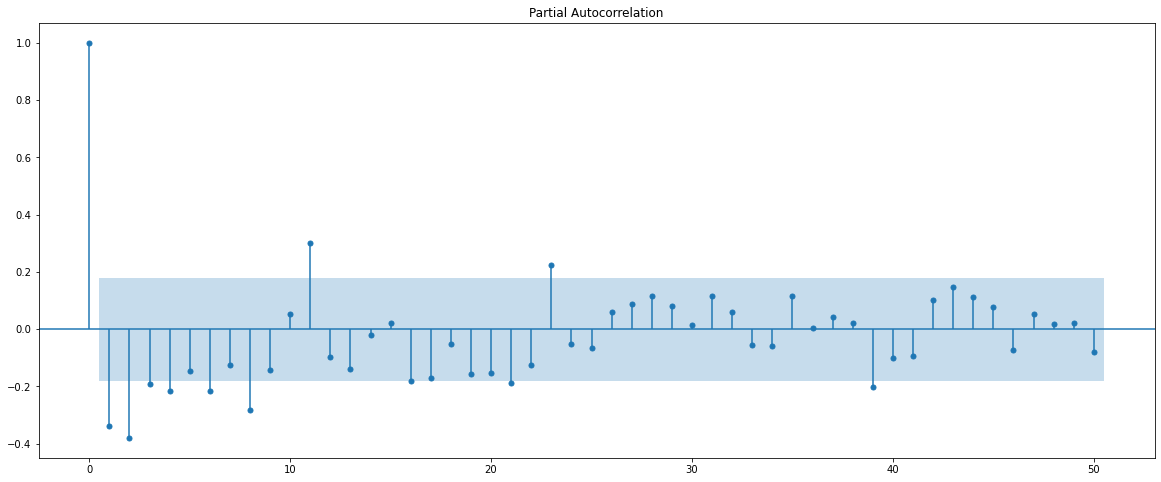


Figure :Sparkling-Manual SARIMA PACF

Here, we have taken alpha=0.05.

We are going to take the seasonal period as 12. We will keep the p(1) and q(1) parameters same as the ARIMA model.

The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 0.

The Moving-Average parameter in an SARIMA model is 'Q' which comes from the significant lag after which the ACF plot cuts-off to 0.

By looking at the plots we see that the ACF and the PACF do not directly cut-off to 0.

Hence P=4 and Q=2. As we have taken a differencing of first order on the seasonally differenced series D is taken as 1.

‘p’,’q’,’d’ has the same value as the one calculated in the ARIMA model.

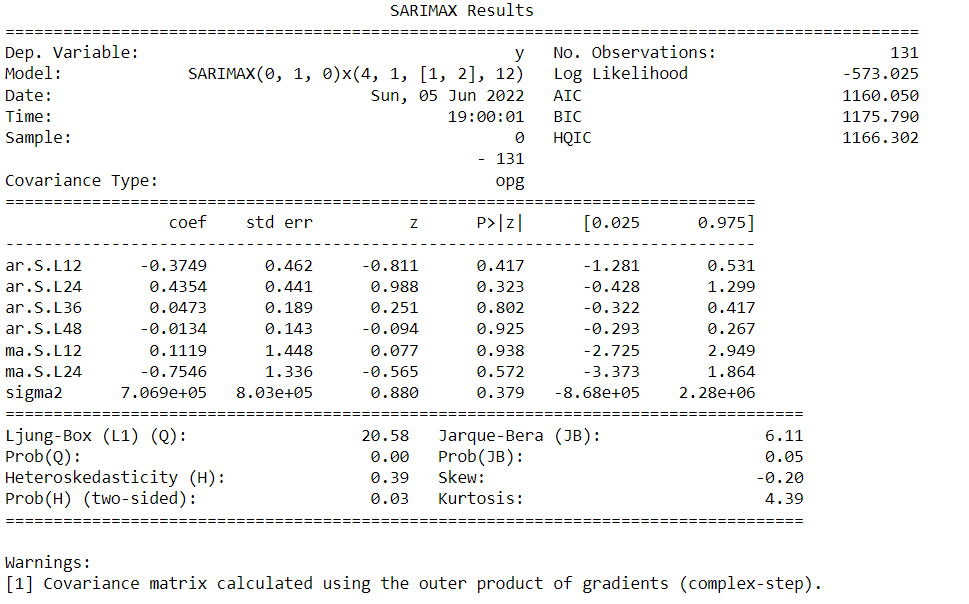
Order=(0,1,0) and Seasonal\_order=(4,1,2,12) are the parameters to be passed to the SARIMA model along with the difference data after dropping NA values.

enforce\_stationarity🡪Whether or not to transform the AR parameters to enforce stationarity in the autoregressive component of the model.

enforce\_invertibility🡪Whether or not to transform the MA parameters to enforce invertibility in the moving average component of the model.

enforce\_stationarity and enforce\_ invertibility is given as false.

The model is built with these values and RMSE is calculated.



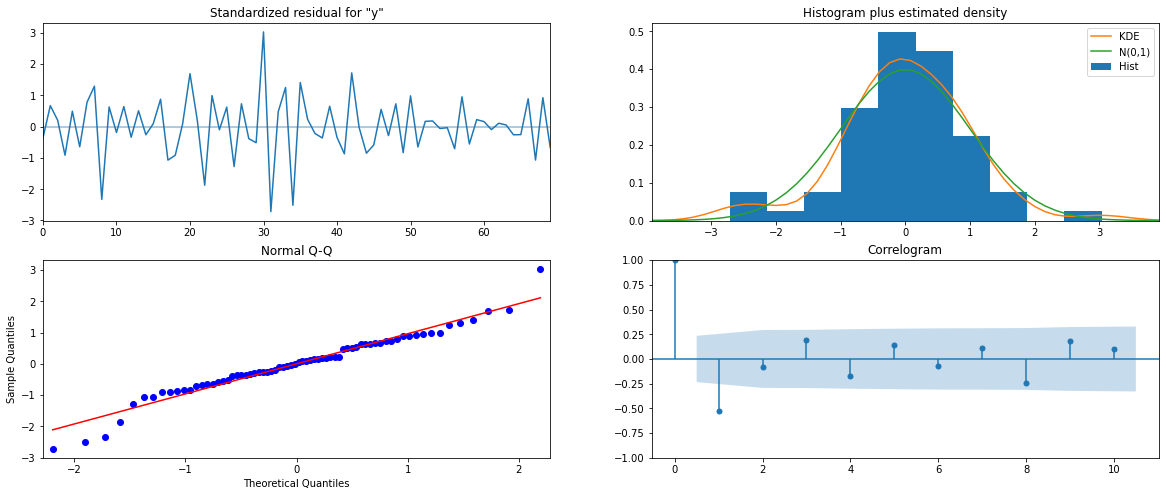


Figure :Sparkling-Manual SARIMA diagnostics

The diagnostics look good here.

RMSE of Manual SARIMA model on Test data is: 3437.38

**2.8.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.**

The model with its parameter and its corresponding RMSE value is shown below in ascending order of RMSE value from which we can see that Iterative Triple Exponential Smoothing with additive trend and Multiplicative Seasonality has the least RMSE value. Hence it’s the optimum model.

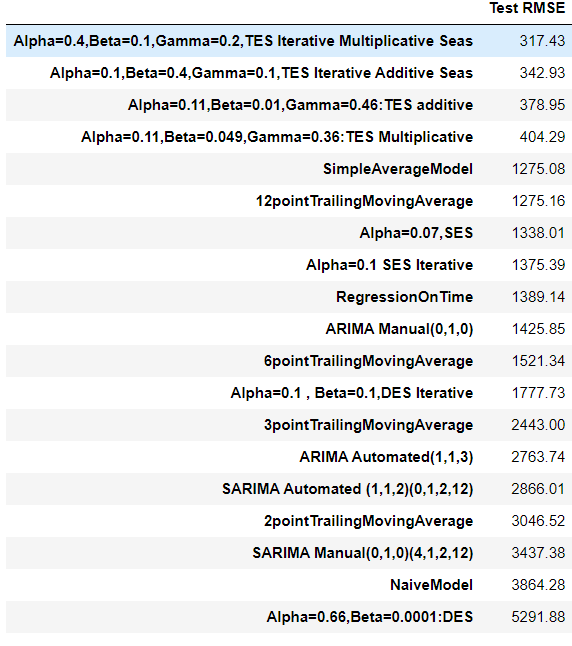


Table :Sparkling-Model-RMSE

**2.9.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**

The optimum model Iterative Triple Exponential Smoothing with additive trend and Multiplicative Seasonality is built and forecasted for 12 months into the future.

The paramters are as below:

{'smoothing\_level': 0.4, 'smoothing\_trend': 0.1, 'smoothing\_seasonal': 0.2, 'damping\_trend': nan, 'initial\_level': 2356.541666666665, 'initial\_trend': -9.181060606060463, 'initial\_seasons': array([0.71166877, 0.67309316, 0.81943184, 0.78429538, 0.63424785,

0.63175794, 0.82647725, 1.0318111 , 0.89263071, 1.1231428 ,

1.69872589, 2.17271729]), 'use\_boxcox': False, 'lamda': None, 'remove\_bias': False}

The forecast is shown with 95% confidence interval band.

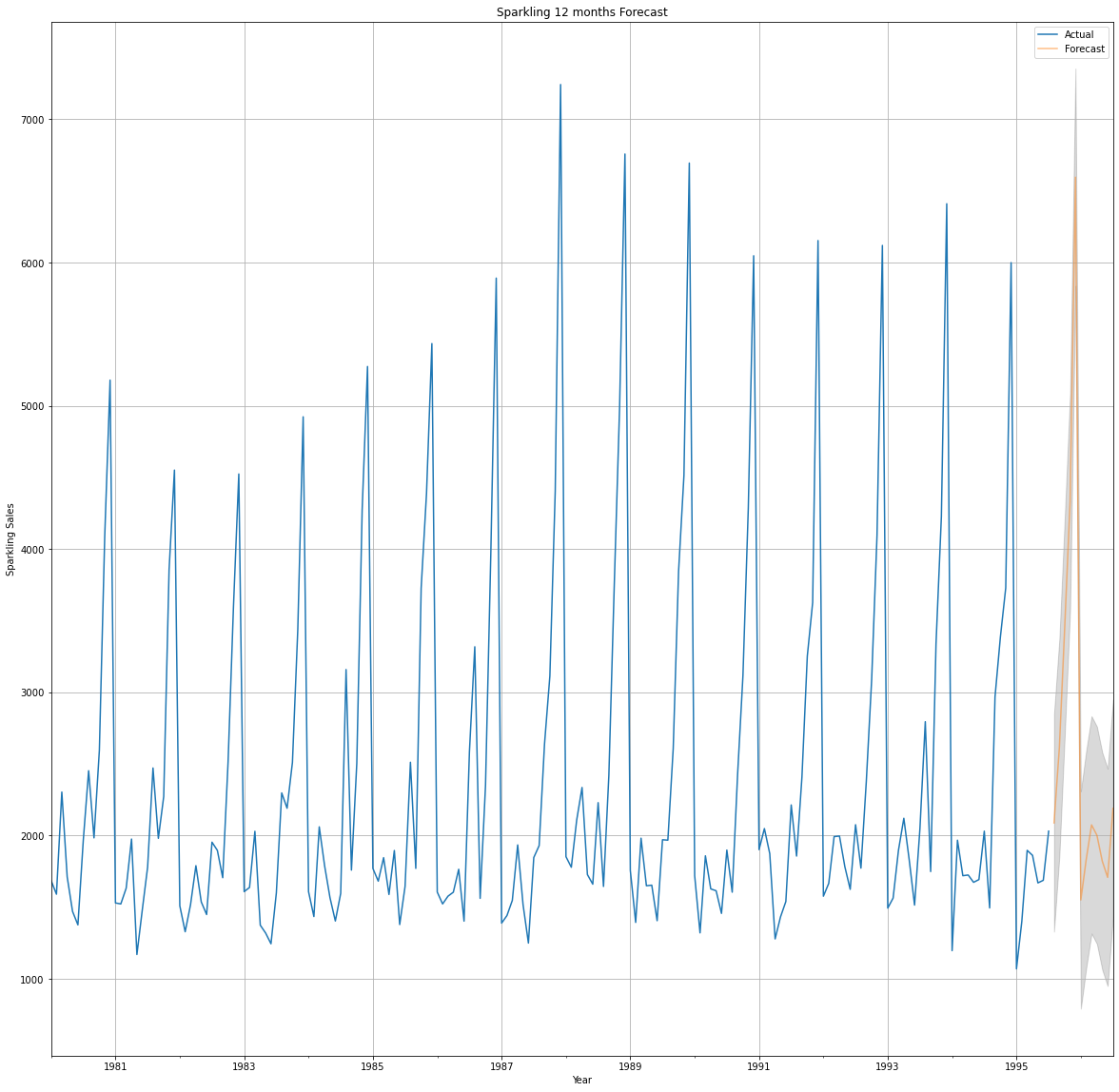


Figure :Sparkling Optimum Model Plot

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

1. First ,read the data as a time series and plotted it on a graph to show how sales for Sparkling wines over the years.

2. Then performed some exploratory data analyses on the data sets, creating various types of charts for analyze the sales.

3. I split the data into test(data from the year 1991) and train(data before the year 1991).

4. Next I built the following models :

• Simple Exponential Smoothing Model

• Iterative Simple Exponential Smoothing Model

• Double Exponential Smoothing Model

• Iterative Double Exponential Smoothing Model

• Triple Exponential Smoothing Model

• Iterative Triple Exponential Smoothing Model.

• Linear Regression Model

• Naïve Approach

• Simple Average Model

• Moving Average Model

For all the above models RMSE value was calculated to understand the performance.

5. The stationarity of the data was checked by stating hypothesis for statistical testing and using ADF Test.

6. From here, we build ARIMA and SARIMA models, but first we examine the dataset. If the series is not stationary, we take the first difference of the series and converted into a stationary series.

7. The ARIMA/SARIMA models are built using AIC scores, we select the parameter with the least AIC and the model is built with it. RMSE is also calculated to check the performance.

8. The ARIMA/SARIMA models are built manually by calculating value of p,q,P,Q,s,d,D from ACF , PACF graph. RMSE is also calculated to check the performance.

9. Finally, we take the model with minimum RMSE value and build the most optimum model on the complete data .The sales for the next 12 months in future with 95% confidence intervals is predicted.

**Recommendations**

--> Fourth quarter has the highes sales among other quarter. So the company can stock up the wines in the second quarter itself to prepare themselves to supply the high demand in the fourth quarter.

--> Proper branding advertising in leading newspaper and magazines can be done. Social Media Advertising can also be done to improve the sales in the first 3 quarters.

--> Second quarter has the lowest sales . So coupons and differs can be offered to boost up the sales.

--> The quality and the taste of the wines can be improved.

--> Further information like age group,location of the customers can be analyzed to improve the model performance and get a better understanding of the Sparkling wine sales.

--------------------------X----------X---------X--------------------------------------------