**Business Report**

**Time Series Analysis and Forecast**

Contents

[**Problem 01:** 4](#_Toc121675095)

[Data Dictionary: 4](#_Toc121675096)

[Project Goal: 4](#_Toc121675097)

[Exploratory Data Analysis 4](#_Toc121675098)

[Converting Object variables to categorical variables 14](#_Toc121675099)

[Scaling of Data 15](#_Toc121675100)

[Train-Test Data Split 15](#_Toc121675101)

[Logistic Regression 15](#_Toc121675102)

[Linear Discriminant Analysis 18](#_Toc121675103)

[KNN Model Analysis 20](#_Toc121675104)

[Naïve Bayes Model Analysis 23](#_Toc121675105)

[Bagging with Random Forest Model: 25](#_Toc121675106)

[Gradient Boosting Model: 28](#_Toc121675107)

[Ada Boosting Model: 30](#_Toc121675108)

[Final Model: Comparing all the models 32](#_Toc121675109)

[Overall Best/Optimized model: 32](#_Toc121675110)

[Problem 02: 33](#_Toc121675111)

[2.1 Find the number of characters, words, and sentences for the mentioned documents. 33](#_Toc121675112)

[2.2 Remove all the stop words from all three speeches 33](#_Toc121675113)

[Libraries used: 33](#_Toc121675114)

[Stemming: 33](#_Toc121675115)

[Most common words in Roosevelt speech after removing stopwords 33](#_Toc121675116)

[Most common words in Kennedy speech after removing stopwords 34](#_Toc121675117)

[Most common words in Nixon speech after removing stopwords 34](#_Toc121675118)

[Top three words in Roosevelt's speech(after removing the stopwords) 34](#_Toc121675119)

[Top three words in Kennedy's speech(after removing the stopwords 34](#_Toc121675120)

[Top three words in Nixon's speech(after removing the stopwords 34](#_Toc121675121)

[Roosevelt speech 35](#_Toc121675122)

[Kennedy speech 36](#_Toc121675123)

[Nixon speech 37](#_Toc121675124)

[Conclusion: 37](#_Toc121675125)

List of all the tables and figures in the project:

|  |
| --- |
| List of all figures and tables: |
| Exploratory Data Analysis |
| Dataset (first 5) Sample |
| Dataset (last 5) Sample |
| Shape of the dataset |
| Datatypes in the dataset |
| Statistics of the dataset |
| Missing Values |
| Duplicated Values |
| Data Skewness Values |
| Unique values and their count in dataset |
| Boxplot and frequency distribution of other variables |
| Count plot of Gender |
| Count plot of vote |
| Hague versus political knowledge |
| Blair versus political knowledge |
| Vote versus age with gender |
| Boxplot and Strip plot of Hauge and Blair versus age |
| Before Outliers Treatment |
| After Outliers Treatment |
| Correlation Heatmap |
| Pair Plot |
| Logistic Regression |
| Accuracy for Train Data |
| Classification Report for Train Data |
| Confusion Matrix for Train Data |
| AUC Curve for Train Data |
| Accuracy for Test Data |
| Classification Report for Test Data |
| Confusion Matrix for Test Data |
| AUC Curve for Test Data |
| Linear Discriminant Analysis |
| Accuracy for Train Data |
| Classification Report for Train Data |
| Confusion Matrix for Train Data |
| AUC Curve for Train Data |
| Accuracy for Test Data |
| Classification Report for Test Data |
| Confusion Matrix for Test Data |
| AUC Curve for Test Data |
| KNN Model |
| Accuracy for Train Data |
| Classification Report for Train Data |
| Confusion Matrix for Train Data |
| AUC Curve for Train Data |
| Accuracy for Test Data |
| Classification Report for Test Data |
| Confusion Matrix for Test Data |
| AUC Curve for Test Data |
| Naïve Bayes Model |
| Accuracy for Train Data |
| Classification Report for Train Data |
| Confusion Matrix for Train Data |
| AUC Curve for Train Data |
| Accuracy for Test Data |
| Classification Report for Test Data |
| Confusion Matrix for Test Data |
| AUC Curve for Test Data |
| Bagging with Random Forest |
| Accuracy for Train Data |
| Classification Report for Train Data |
| Confusion Matrix for Train Data |
| AUC Curve for Train Data |
| Accuracy for Test Data |
| Classification Report for Test Data |
| Confusion Matrix for Test Data |
| AUC Curve for Test Data |
| Gradient Boosting Model |
| Accuracy for Train Data |
| Classification Report for Train Data |
| Confusion Matrix for Train Data |
| AUC Curve for Train Data |
| Accuracy for Test Data |
| Classification Report for Test Data |
| Confusion Matrix for Test Data |
| AUC Curve for Test Data |
| Ada Boosting |
| Accuracy for Train Data |
| Classification Report for Train Data |
| Confusion Matrix for Train Data |
| AUC Curve for Train Data |
| Accuracy for Test Data |
| Classification Report for Test Data |
| Confusion Matrix for Test Data |
| AUC Curve for Test Data |
| WordCloud |
| WordCloud with Roosevelt Speech |
| WordCloud with Kennedy Speech |
| WordCloud with Nixon Speech |

# **Project Goal:**

This report's goal is to use Python to analyse and forecast the datasets "Rose” and “Sparkling” and eventually produce insights about it. The core elements of this exploration summary report will be organized as follows:

* Using the dataset as input in a Jupyter notebook.
* Recognizing the dataset's structure.
* Exploratory Data Analysis.
* Graphical Interpretation; and
* Time Forecasting using different models
* Best Model for each dataset to forecast upcoming 12 months data
* Insights or Inferences from the data

## Rose Dataset

### Exploratory Data Analysis

#### Sample of the dataset:

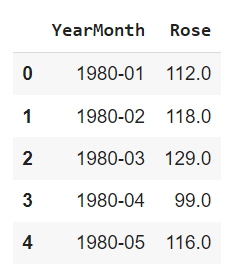


Fig 1. Sample of Rose Data

#### Data Types in the dataset before Data Modification:

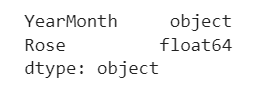


Fig 2. Datatype of Rose Data

#### Data Modification for easiness of Time Forecasting:

Since we have monthly data for years starting from 1980 to 1995 combined, let us draw a date range and incoporate as "TimeStamp" and drop "YearMonth" for ease.

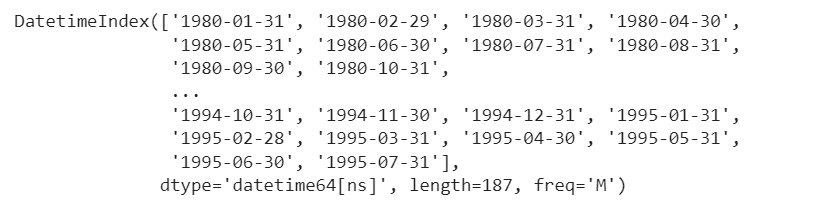


Fig 3. Date Index of Rose Data

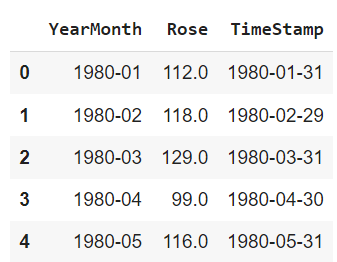


Fig 4. TimeStamp addition to Rose Data

After we have added the TimeStamp, the “YearMonth” is dropped, and TimeStamp is made the index. This would be the modified Dataset we use going forward.

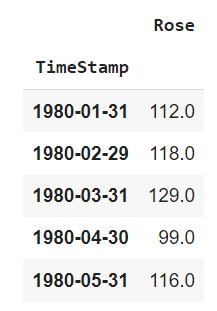


Fig 5. New Modified Rose Dataset

#### Basic details and Descriptive Statistics of Rose:



Fig 6. Length of Rose Dataset after modification

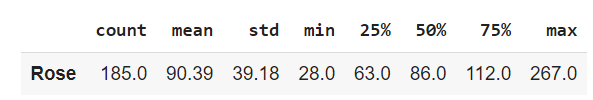


Fig 7. Descriptive Statistics of Rose Dataset after modification

#### Missing Values and Interpolation:

As we checked the dataset, there are 2 null or missing values. So, the values were interpolated. If this is not treated, it might cause a significant effect on the models we build.

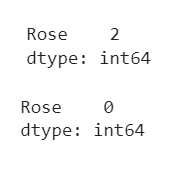


Fig 8. Null values before (above) and after (below) interpolation

#### Monthly Boxplot of Rose:

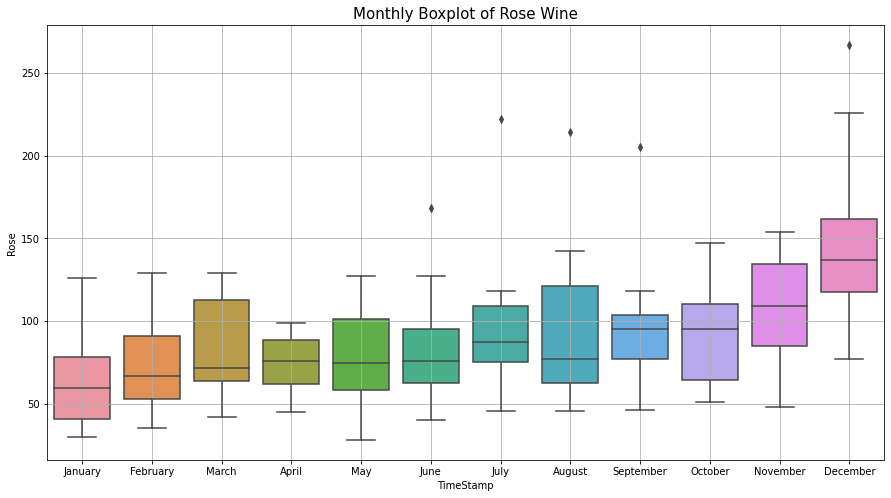


Fig 9. Monthly Boxplot of Rose

* The above plot shows a spike in the last quarter of October to December.

#### Monthly sales over years of Rose Wine:

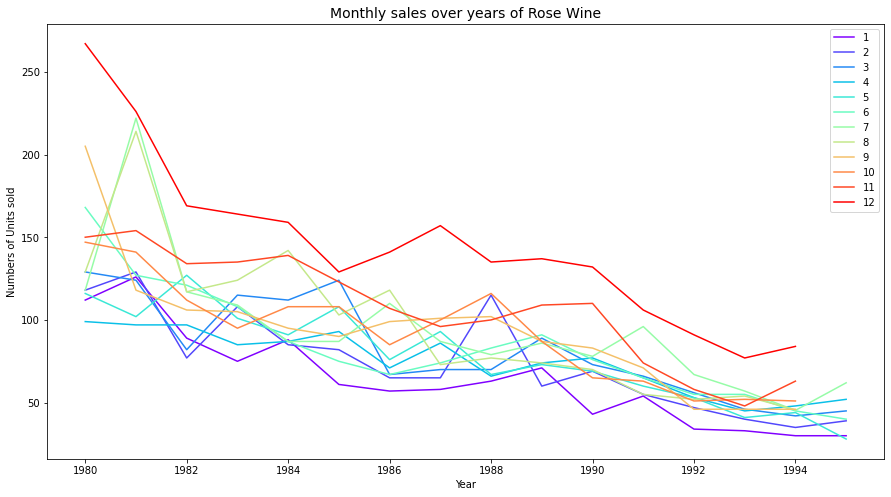


Fig 10. Monthly sales over years of Rose Wine

### Visualization of Time Series in Rose:

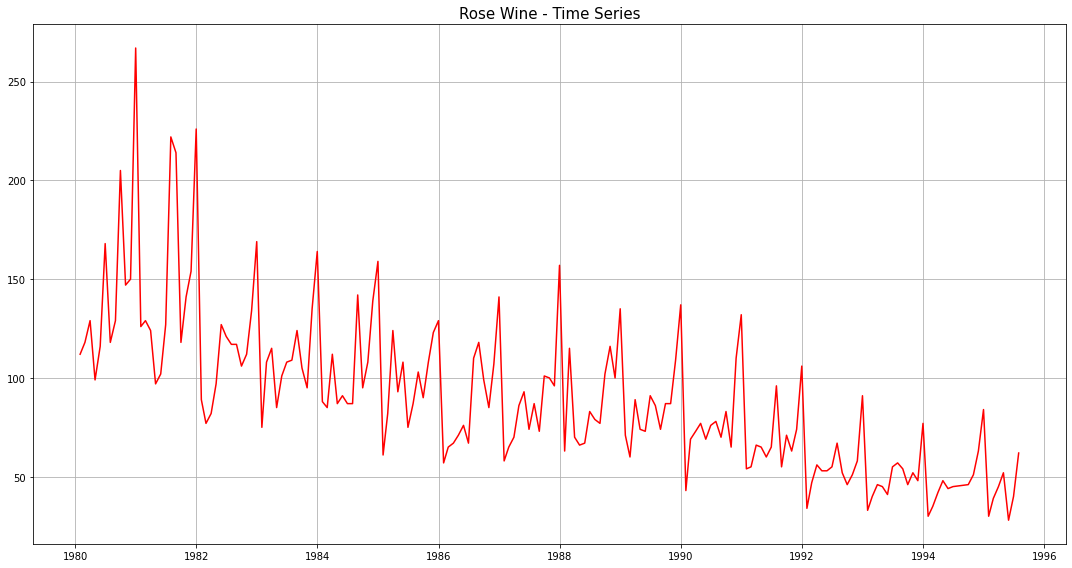


Fig 11. Visualization of Time Series in Rose

### Decomposition and plots of different component:

Additive Decomposition of Rose Dataset:

Once we decompose the Rose dataset using additive model, we obtain the following parameters and their respective plots.

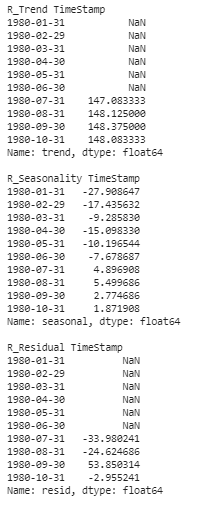


Fig 12. Trend, Seasonality and Residue of Rose

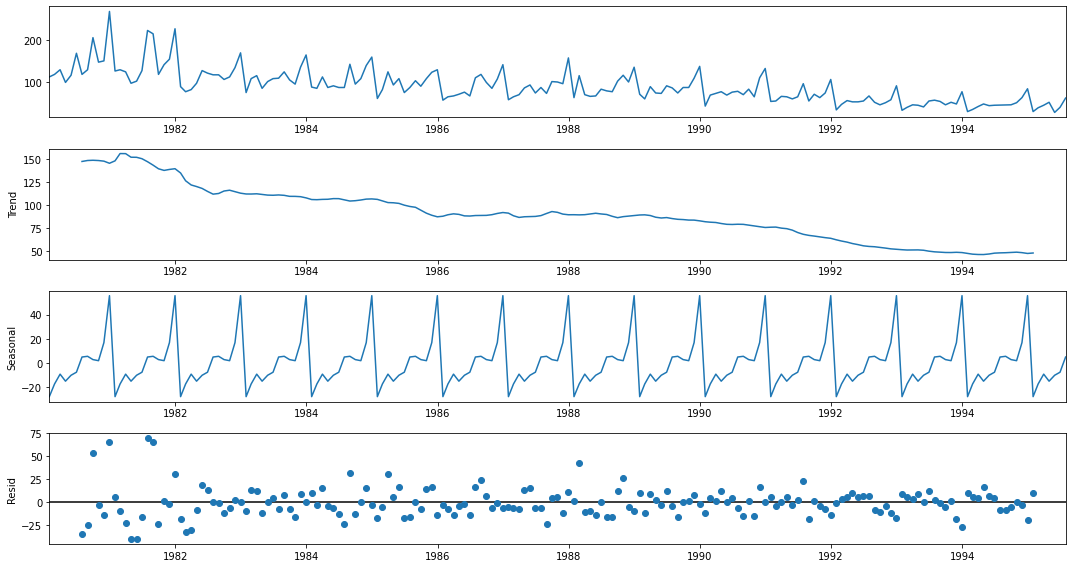


Fig 13. Trend, Seasonality and Residue plot for Rose

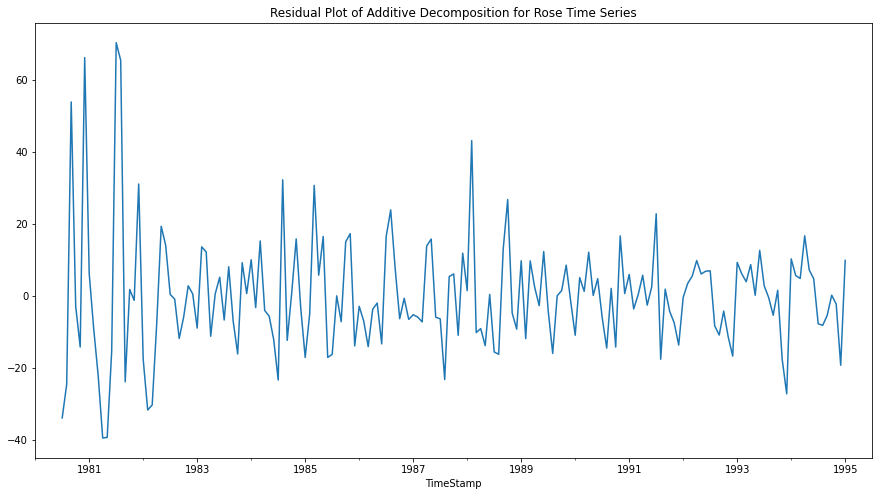


Fig 14. Residual Plot of Additive Decomposition for Rose Time Series

Multiplicative Decomposition of Rose Dataset:

Once we decompose the Rose dataset using additive model, we obtain the following parameters and their respective plots.

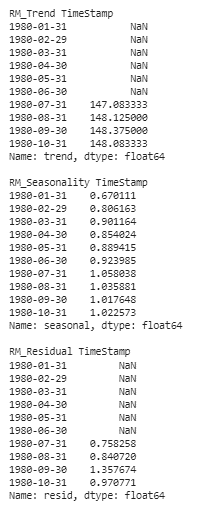


Fig 13. Trend, Seasonality and Residue of Rose (Multiplicative)

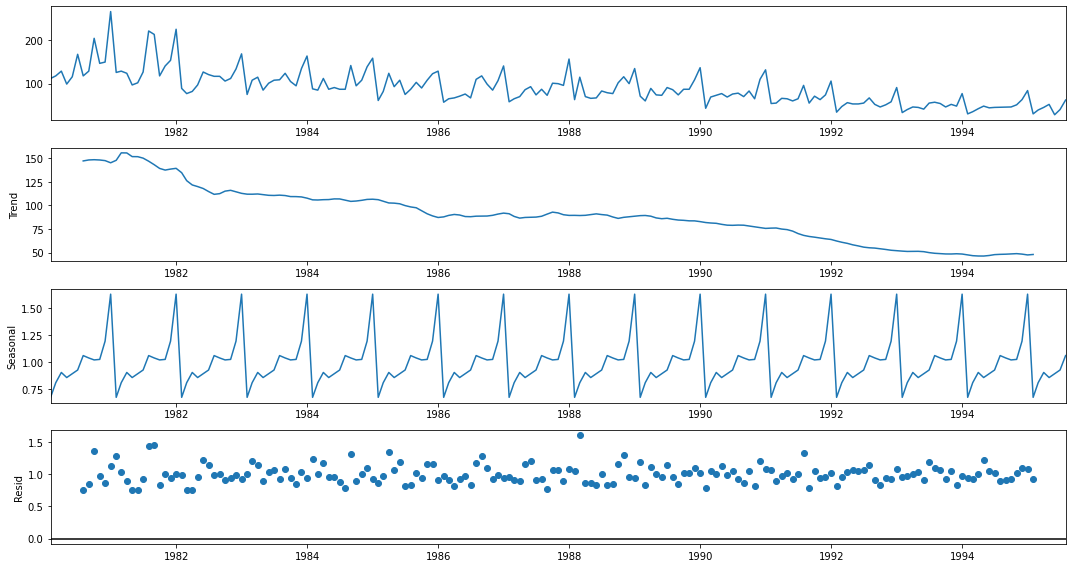


Fig 14. Trend, Seasonality and Residue plot for Rose (Multiplicative)

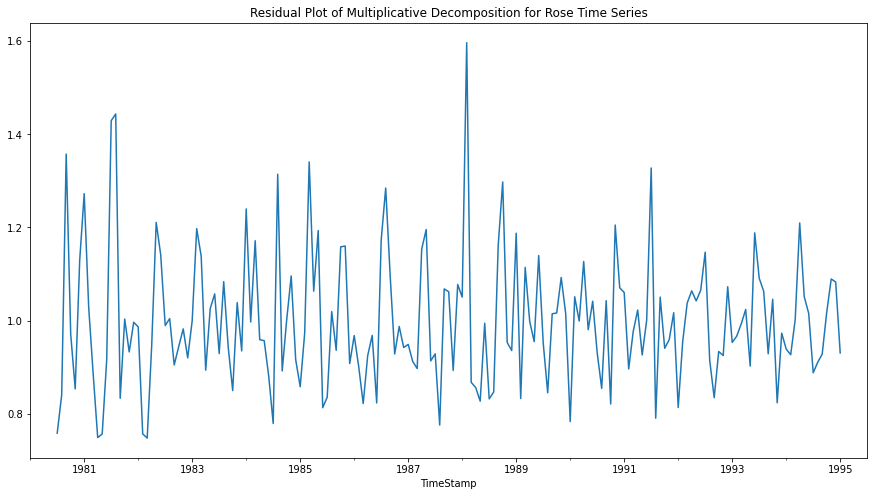


Fig 15. Residual Plot of Additive Decomposition for Rose Time Series (Multiplicative)

Conclusion:

Multiplicative Models

The seasonality increases or decreases over time. It is proportionate to the trend

Yt= Trend \* Seasonality \* Residual

Here by just observing the Residual patterns of Additive and Multiplicative models of Rose, we can conclude that Rose follows multiplicative model.

### Test and Train Split:

The whole Rose data was split into test and training dataset such that the test data starts in 1991.

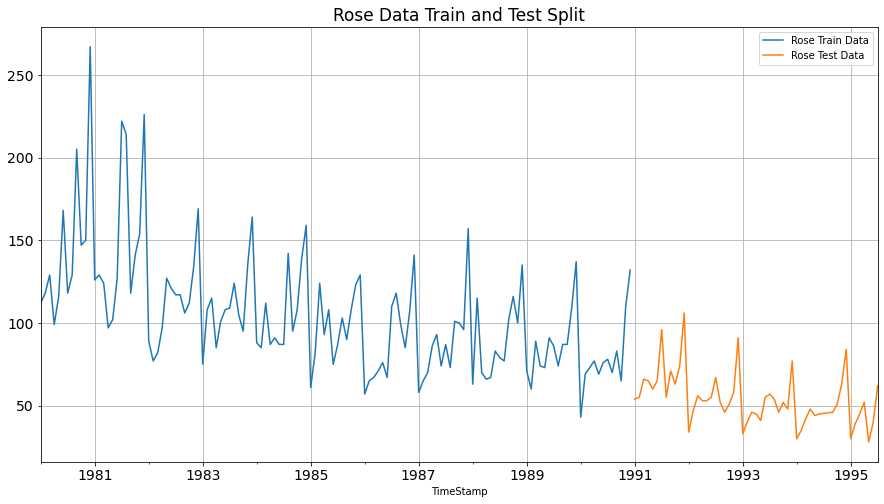


Fig 16. Test and training data for Rose

### Various Models on the Rose Dataset to forecast the time series:

#### Model 1 – Linear Regression

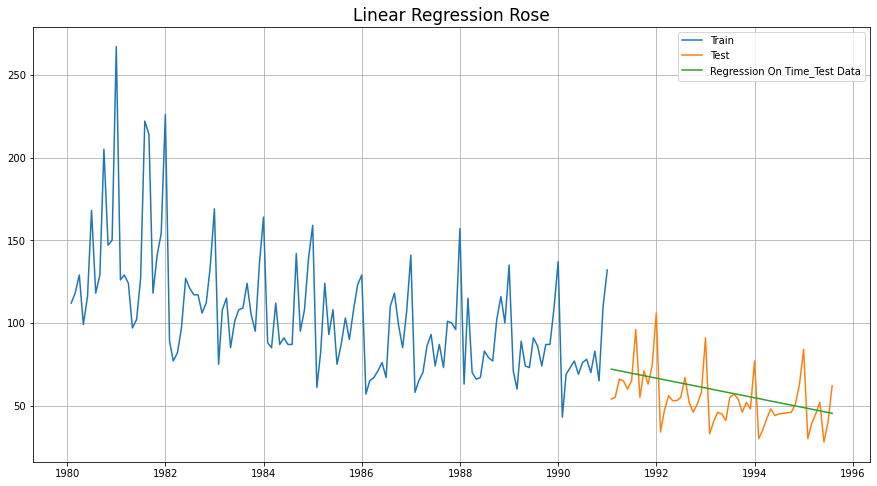


Fig 17. Linear Regression for Rose

#### Model 2 - Naive Bayes

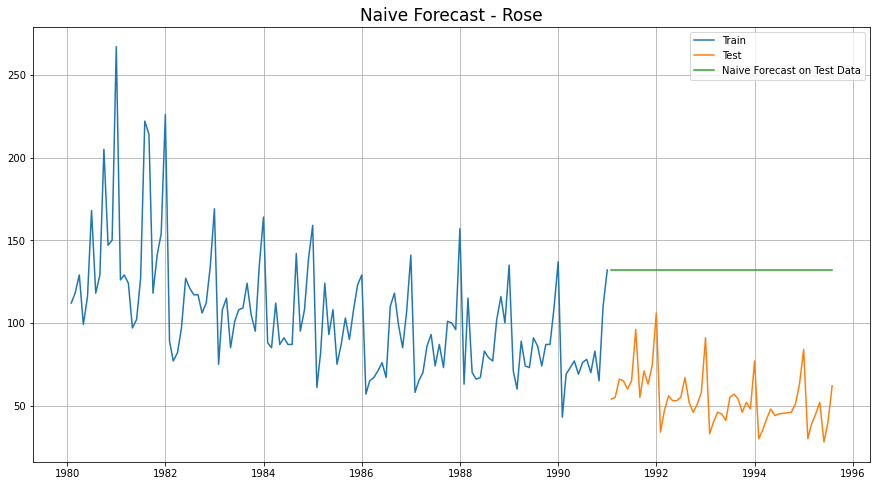


Fig 18. Naïve’s Approach for Rose

#### Model 3 - Simple Average

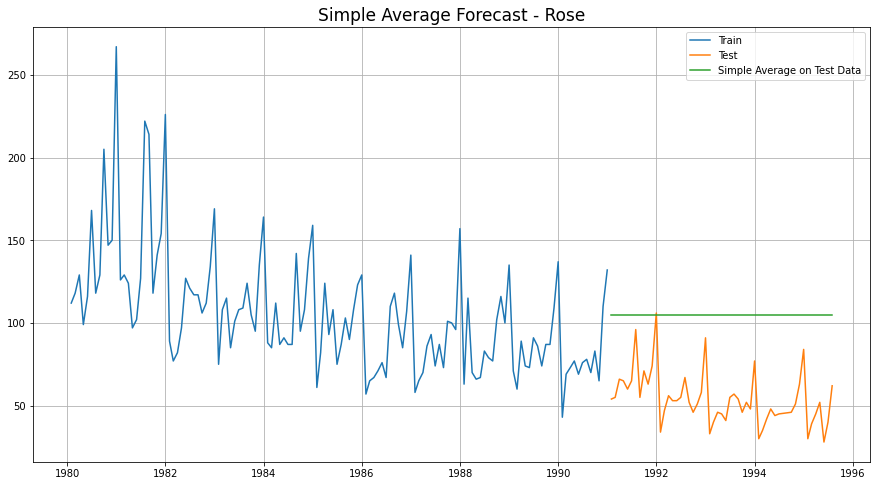


Fig 19. Simple Average Model for Rose

#### Model 4- Moving Average (Rose)

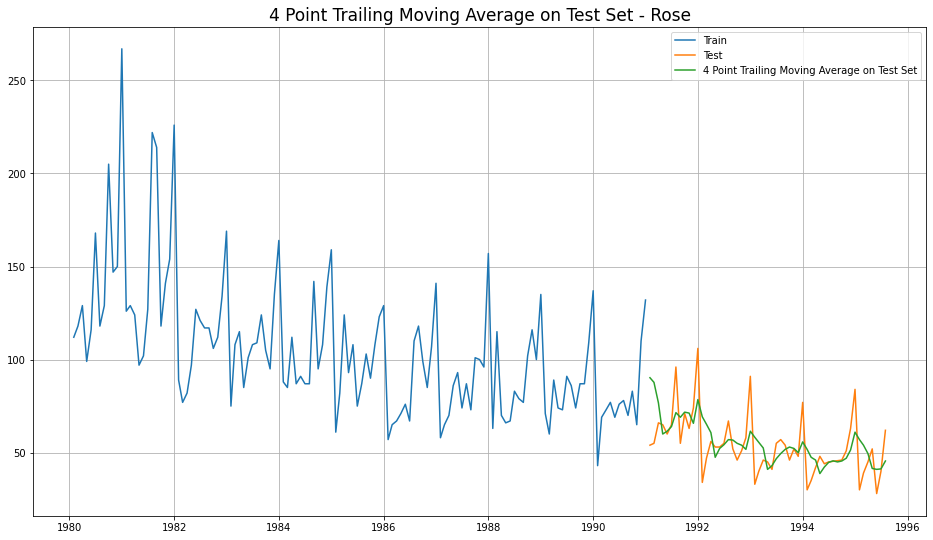


Fig 20. 4 Point Trailing Moving Average for Rose

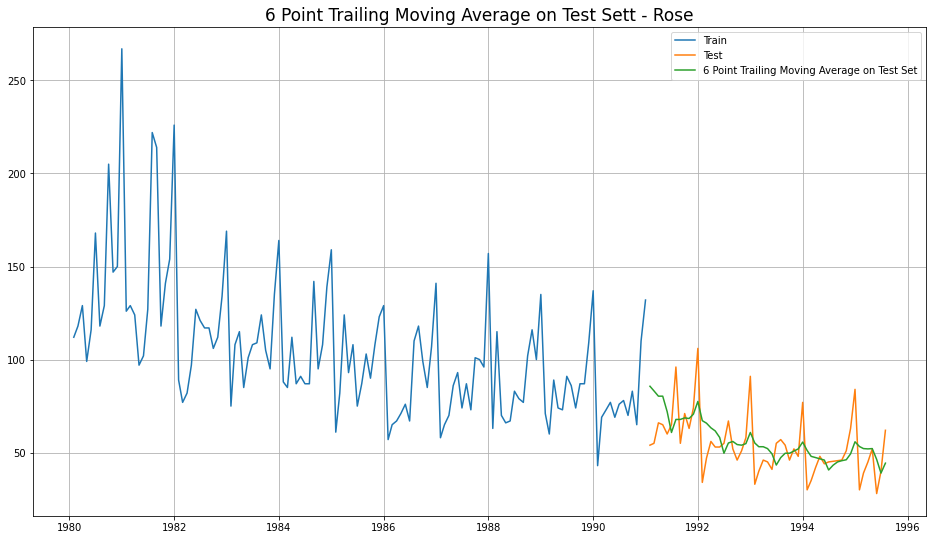


Fig 21. 6 Point Trailing Moving Average for Rose

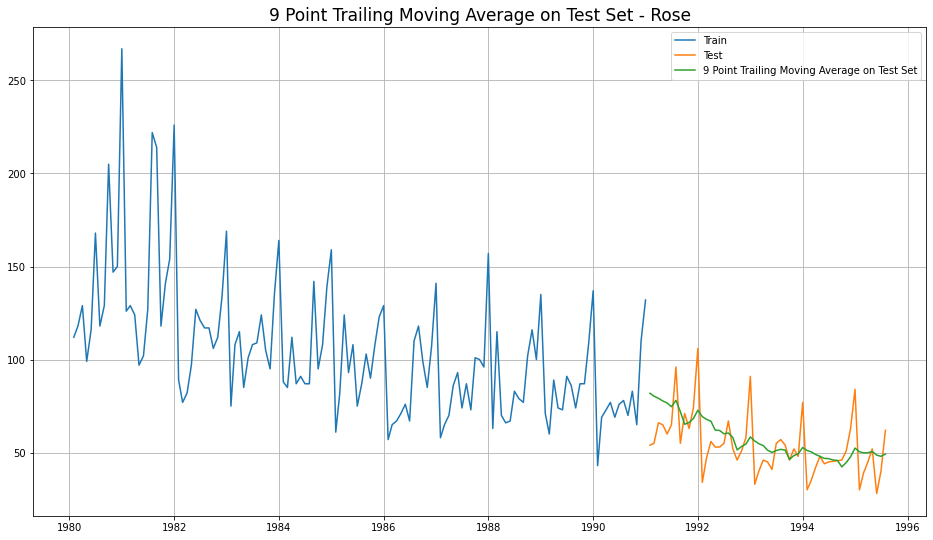


Fig 22. 9 Point Trailing Moving Average for Rose

#### Consolidated Moving Average Forecasts (Rose)

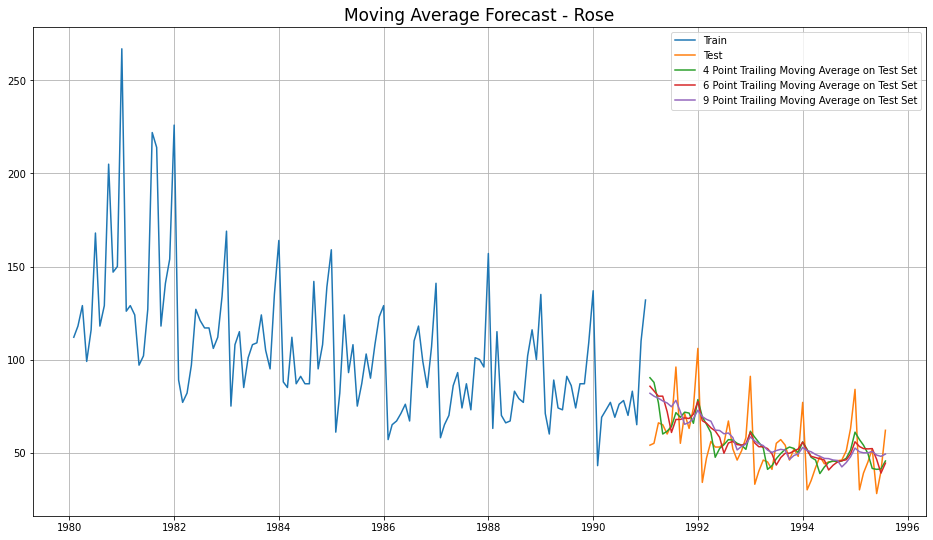


Fig 22. 4 Consolidated Trailing Moving Average for Rose

Let us see the best of all models built until now, before we go forward with Exponential Smoothing Models.

|  |  |
| --- | --- |
| **Model Name** | **Test RMSE Rose** |
| Regression On Time | 15.268955 |
| Naive Model | 79.718773 |
| Simple Average Model | 53.46057 |
| 4point Trailing MovingAverage | 14.451403 |
| 6point Trailing MovingAverage | 14.566327 |
| 9point Trailing MovingAverage | 14.72763 |

Table 1 . Consolidated Scores of Regression, Naive, Simple Average & Moving Average

Till now, the best model built is **4point Trailing MovingAverage**

#### Model 5 - Single Exponential Smoothing

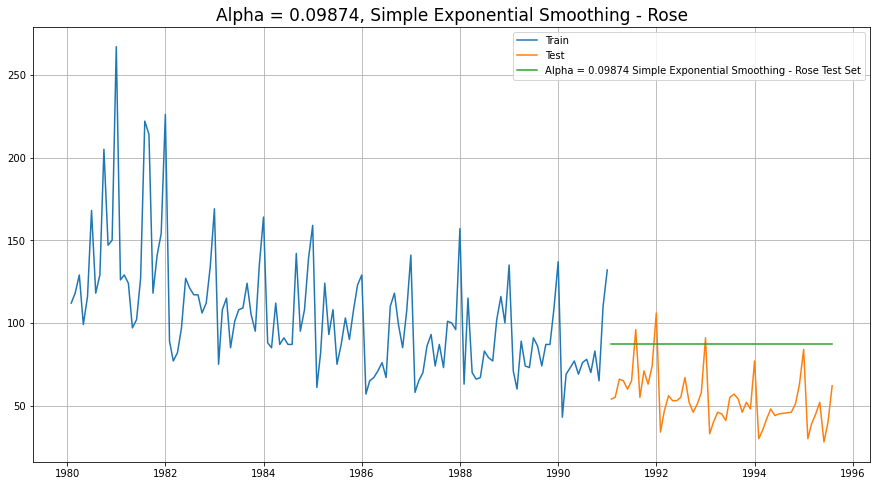


Fig 22. Single Exponential Smoothing

#### Model 6 - Double Exponential Smoothing

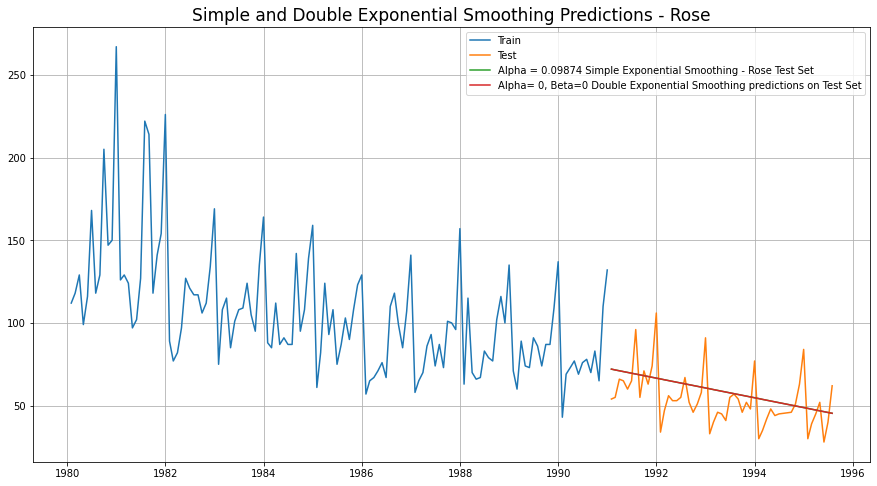


Fig 23. Single and Double Exponential Smoothing

#### Model 7 - Holt-Winter's Model (Exponential Smoothing) with additive errors

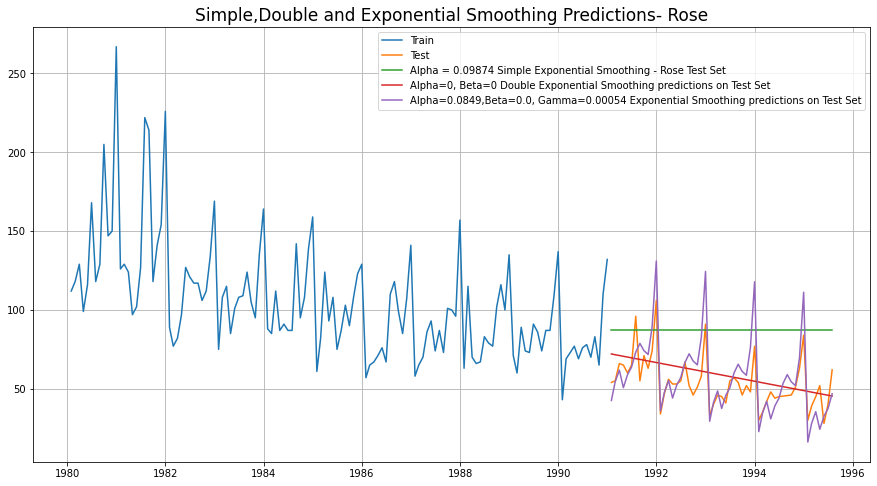


Fig 24. Simple, Double and Exponential Smoothing Predictions

#### Model 8 - Holt-Winter's Model (Exponential Smoothing) with multiplicative seasonality

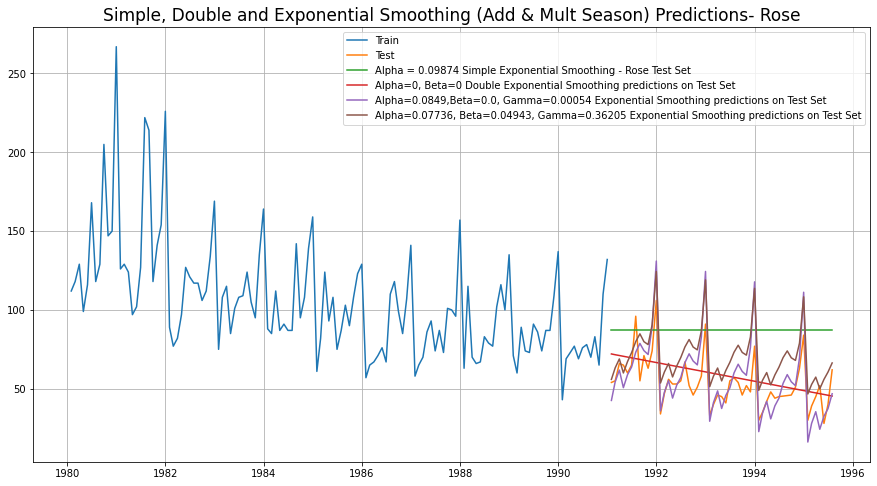


Fig 25. Simple, Double and Exponential Smoothing (Add & Mult.) Season

Conclusion at the end of Exponential Modelling, we see that the Exponential Smoothing with Additive Seasonality is the best model (least RMSE)

|  |  |
| --- | --- |
| **Model Name** | **Test RMSE Rose** |
| Regression On Time | 15.268955 |
| Naive Model | 79.718773 |
| Simple Average Model | 53.46057 |
| 4point Trailing MovingAverage | 14.451403 |
| 6point Trailing MovingAverage | 14.566327 |
| 9point Trailing MovingAverage | 14.72763 |
| Simple Exponential Smoothing | 36.796225 |
| Double Exponential Smoothing | 15.270968 |
| Exponential Smoothing (Additive Seasonality) | 14.24324 |
| Exponential Smoothing (Multi Seasonality) | 19.11311 |

Table 2 -Consolidated Scores of Regression, Naive, Simple Average & Moving Average, Exponential Smoothing

#### Check for stationarity of the whole Time Series data

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

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

1. H0: The Time Series has a unit root and is thus non-stationary.
2. H1:The Time Series does not have a unit root and is thus stationary

We would want the series to be stationary for building ARIMA models and thus we would want the p-value of this test to be less than the α (0.05) value.

Since the p-value is greater than 0.05, the Rose data set is non-stationary.

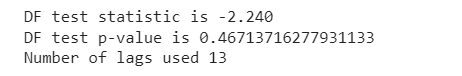


Fig 26. ADF test for Rose

To make the data stationary, lets difference and repeat the ADF for Rose

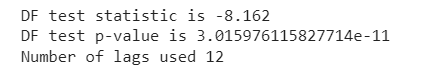
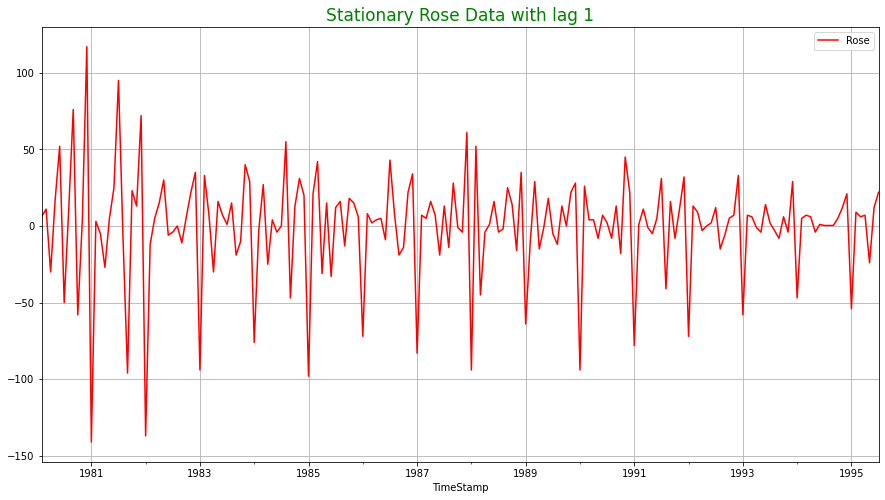
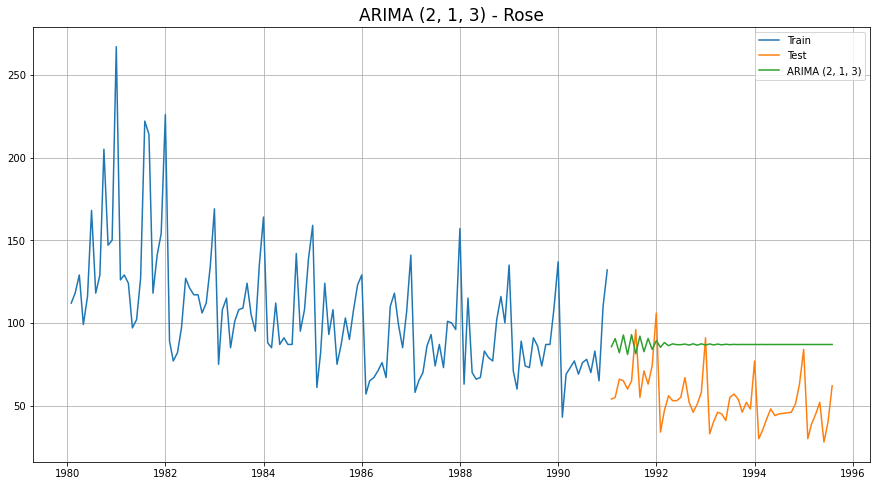


Fig 27. ADF test for Rose (diff)

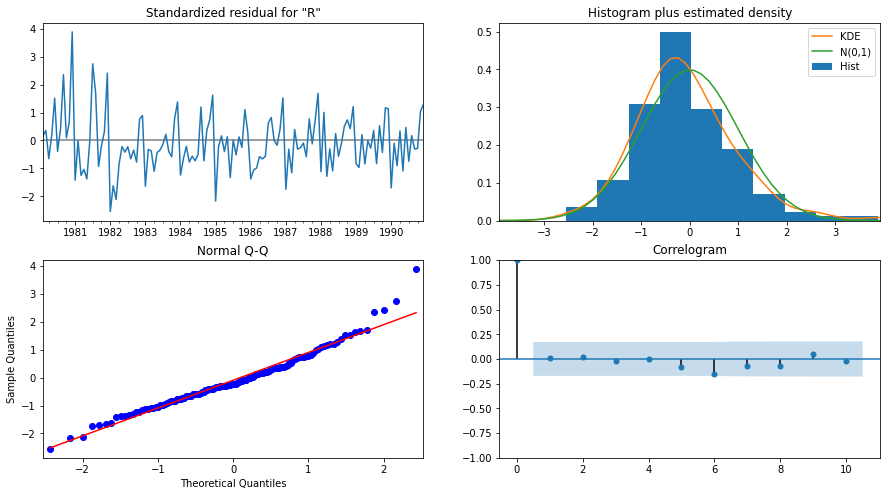


#### Model 9 – Automated ARIMA Model

* We create a grid of all possible combinations of (p, d, q)
* Range of p = Range of q = 0 to 3, Constant d = 1
* Few Examples of the grid -
  + - Model: (0, 1, 0)
    - Model: (0, 1, 1)
    - Model: (0, 1, 2)
    - Model: (0, 1, 3)
    - Model: (1, 1, 0)
    - Model: (1, 1, 1)
    - Model: (1, 1, 2)
    - Model: (1, 1, 3)
    - Model: (2, 1, 0)
    - Model: (2, 1, 1)
    - Model: (2, 1, 2)
    - Model: (2, 1, 3)
    - Model: (3, 1, 0)
    - Model: (3, 1, 1)
    - Model: (3, 1, 2)
    - Model: (3, 1, 3)
* We fit ARIMA models to each of these combinations for both datasets
* We choose the combination with the least Akaike Information Criteria (AIC) - ARIMA (2, 1, 3)
* We fit ARIMA to this combination of (p, d, q) to the Train set and forecast on the Test set
* Finally, we check the accuracy of this model by checking RMSE of Test set



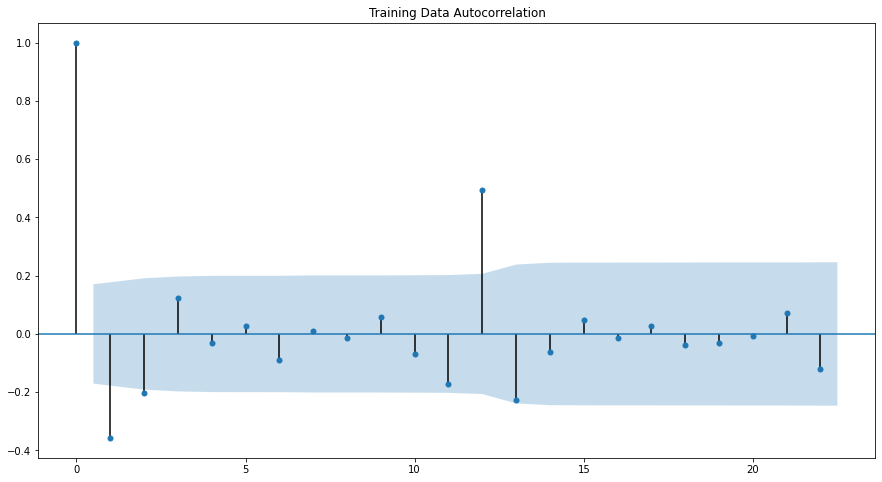
ARIMA (2, 1, 3) – Automated ARIMA



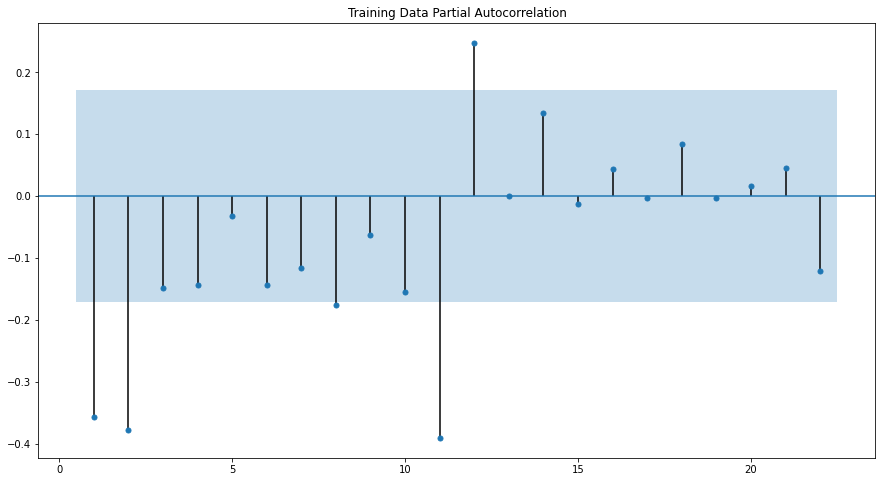
ARIMA (2, 1, 3) Diagnostic Plot – Rose

#### Model 10 – Manual ARIMA Model

* The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off to 2.
* The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 2.
* By looking at the above plots, we will take the value of p and q to be 2 and 2 respectively.

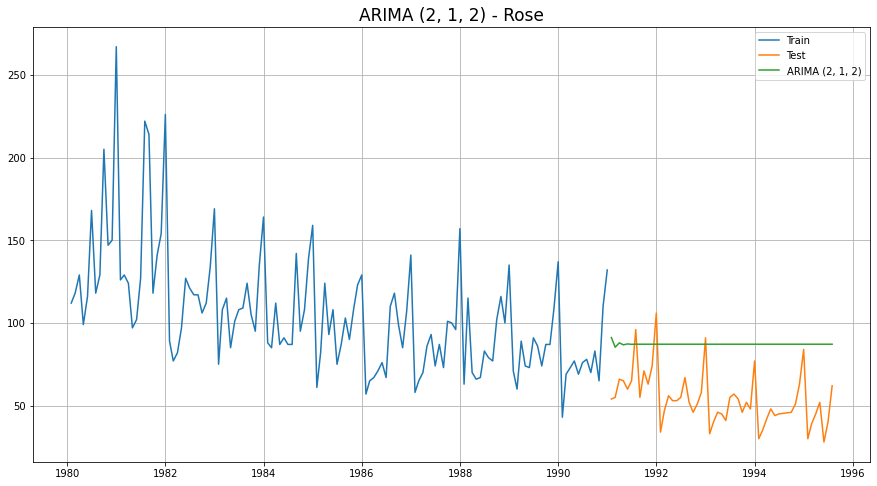


Training Data Autocorrelation

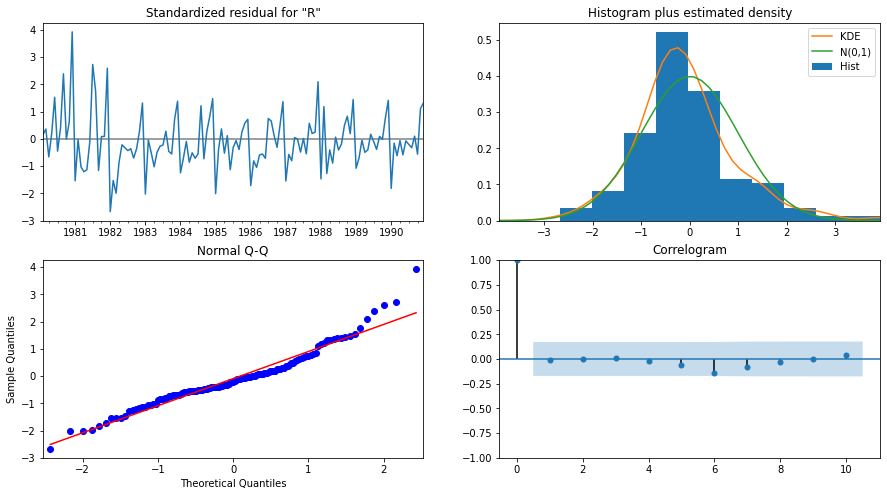


Training Data Partial Autocorrelation

ARIMA Manual - Rose - (2, 1, 2)



ARIMA (2, 1, 2) – Manual ARIMA



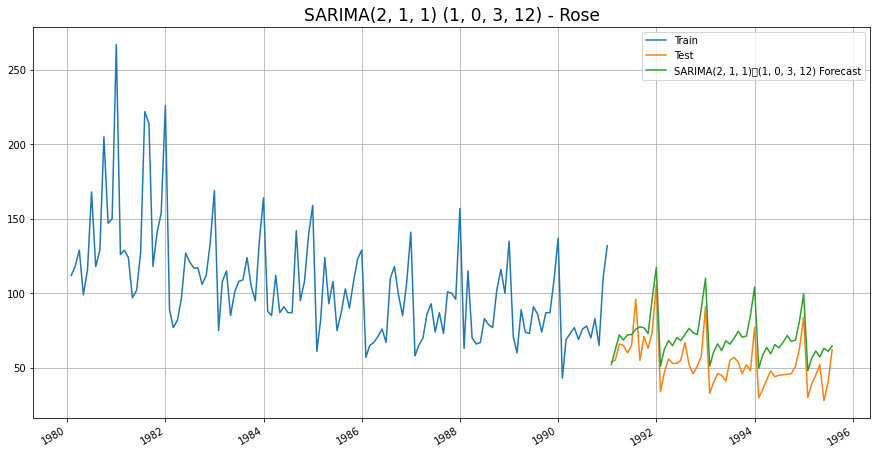
ARIMA (2, 1, 2) Diagnostic Plot - Rose

#### Model 11 – Automated SARIMA Model

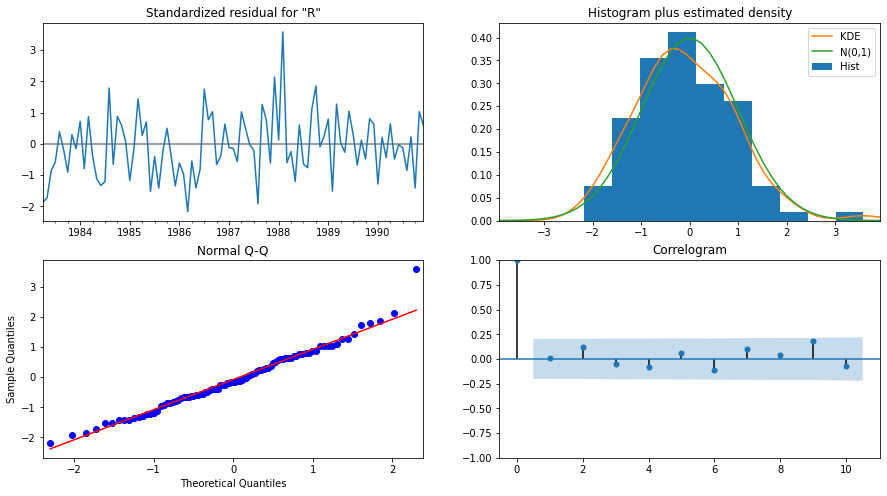
* We create a grid of all possible combinations of (p, d, q) along with Seasonal

(P, D, Q) & Seasonality of 12 (for both datasets)

* Range of p = Range of q = 0 to 3, Constant d = 1
* Range of Seasonal P = Range of Seasonal Q = 0 to 3, Constant D = 1, Seasonality m = 12
* Few Examples of the grid (p, d, q) (P, D, Q, m) -
  + Model: (0, 1, 1)(0, 0, 1, 12)
  + Model: (0, 1, 2)(0, 0, 2, 12)
  + Model: (0, 1, 3)(0, 0, 3, 12)
  + Model: (1, 1, 0)(1, 0, 0, 12)
  + Model: (1, 1, 1)(1, 0, 1, 12)
  + Model: (1, 1, 2)(1, 0, 2, 12)
  + Model: (1, 1, 3)(1, 0, 3, 12)
  + Model: (2, 1, 0)(2, 0, 0, 12)
  + Model: (2, 1, 1)(2, 0, 1, 12)
  + Model: (2, 1, 2)(2, 0, 2, 12)
  + Model: (2, 1, 3)(2, 0, 3, 12)
  + Model: (3, 1, 0)(3, 0, 0, 12)
  + Model: (3, 1, 1)(3, 0, 1, 12)
  + Model: (3, 1, 2)(3, 0, 2, 12)
  + Model: (3, 1, 3)(3, 0, 3, 12)
* We fit SARIMA models to each of these combinations and select with least AIC
* We fit SARIMA to this best combination of (p, d, q) (P, D, Q, m) to the Train set and forecast on the Test set. Then, we check accuracy using RMSE on Test set
* For Rose, Best Combination with Least AIC is - (2, 1, 1) (1, 0, 3, 12)



SARIMA (2, 1, 1) (1, 0, 3, 12) - Automated SARIMA

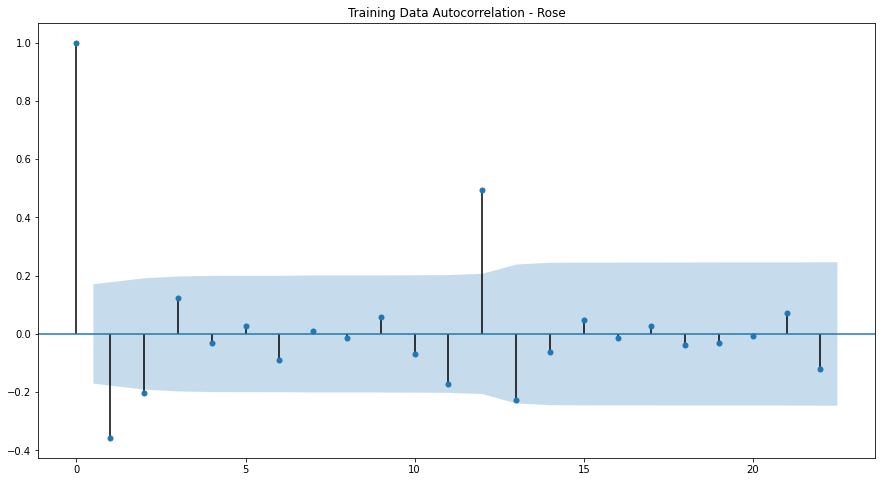


SARIMA (2, 1, 1) (1, 0, 3, 12) Diagnostic Plot – Rose

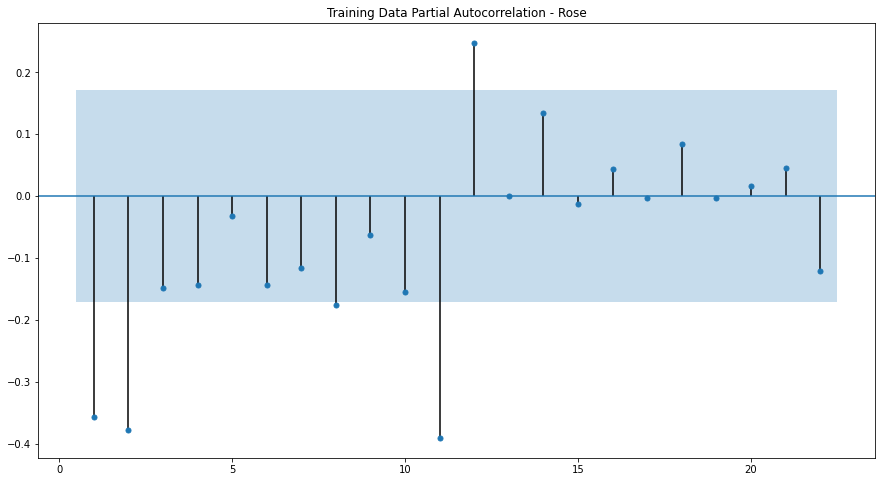
#### Model 12 – Manual SARIMA Model

We are going to take the seasonal period as 12 We are taking the p value to be 2 and the q value also to be 2 as the parameters same as the ARIMA model.

The Auto-Regressive parameter in a SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 2. The Moving-Average parameter in a SARIMA model is 'Q' which comes from the significant lag after which the ACF plot cuts-off to 3.

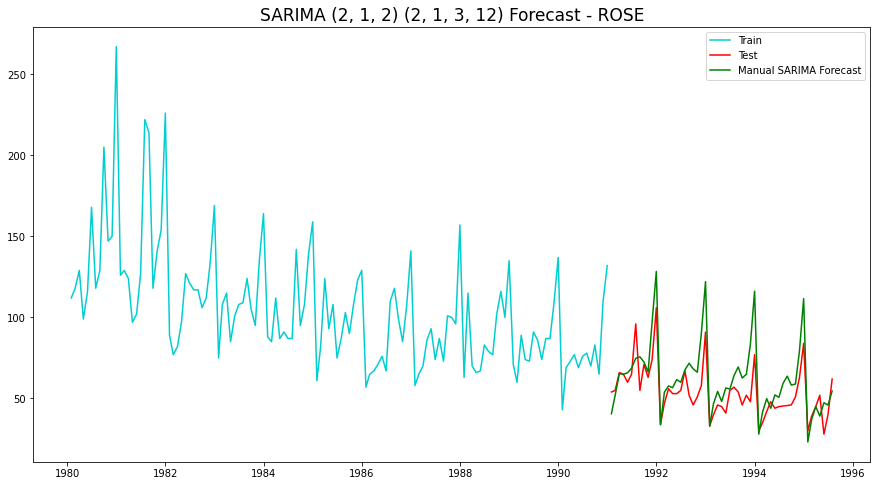


Training Data Autocorrelation

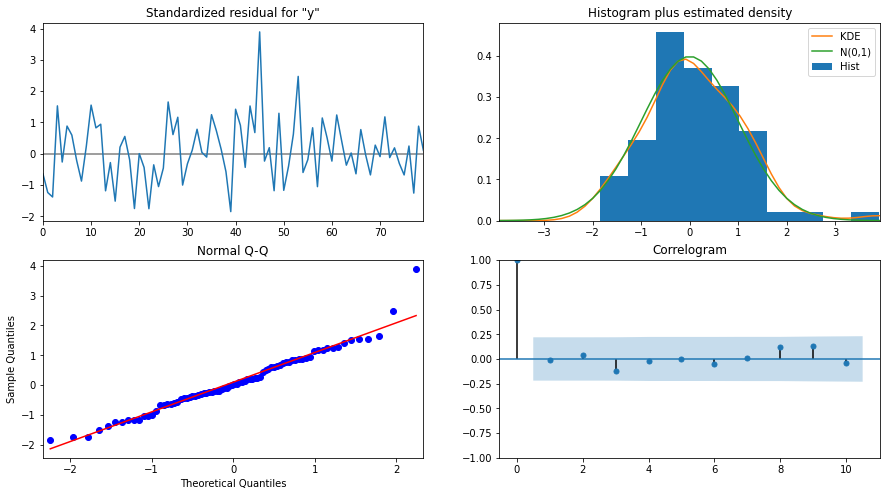


Training Data Partial Autocorrelation

SARIMA Manual - Rose - (2, 1, 2) (2, 1, 3, 12)



SARIMA (2, 1, 2) (2, 1, 3, 12) – Manual SARIMA



SARIMA (2, 1, 2) (2, 1, 3, 12) - Diagnostic Plot - Rose

### Data Frame with all the RMSE values for models built:

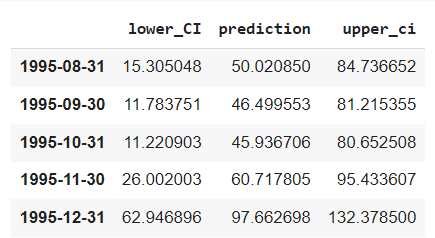
|  |  |
| --- | --- |
| **Model Name** | **Test RMSE Rose** |
| Regression On Time | 15.268955 |
| Naive Model | 79.718773 |
| Simple Average Model | 53.46057 |
| 4point Trailing MovingAverage | 14.451403 |
| 6point Trailing MovingAverage | 14.566327 |
| 9point Trailing MovingAverage | 14.72763 |
| Simple Exponential Smoothing | 36.796225 |
| Double Exponential Smoothing | 15.270968 |
| Exponential Smoothing (Additive Seasonality) | 14.24324 |
| Exponential Smoothing (Multi Seasonality) | 19.11311 |
| ARIMA (2, 1, 3) | 36.81575 |
| ARIMA (2, 1, 2) | 36.871197 |
| SARIMA (2, 1, 1) (1, 0, 3, 12) | 19.003214 |
| SARIMA (2, 1, 2) (2, 1, 3, 12) | 14.799237 |

Conclusion:

As we can see the above table, the lease RMSE out of all models built is Exponential Smoothing (Additive Seasonality).

Let us now build the model on full data and predict the upcoming 12 months.

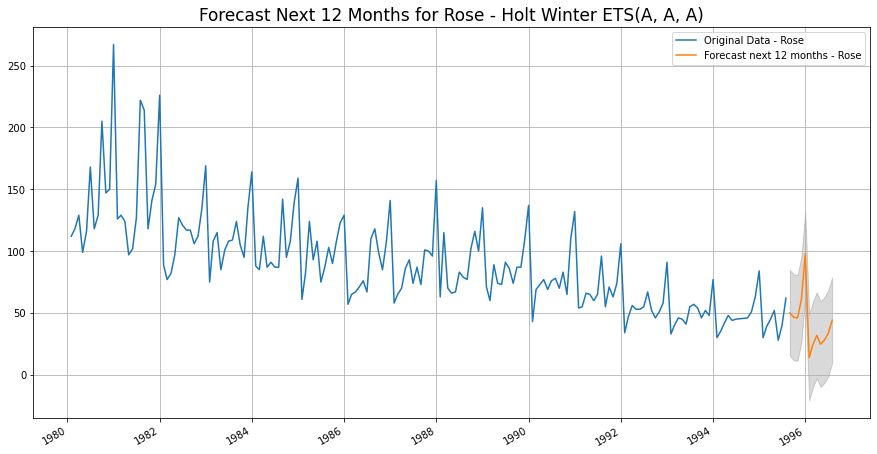
### Final Model – Exponential Smoothing with additive seasonality



Final Model Forecast: Exponential Smoothing (Additive Seasonality - Rose

### Rose Wine Sales - Suggestions :

* First and foremost, the holiday season is rapidly approaching, with sales expected to increase and peak in December. As a result, the company should stock up. However, the long-term decline in Rose Wine sales should be investigated with more data crunching.
* The company can rebrand its Rose variant as well as hire a new Wine-master.
* The company should capitalise on the impending spike from August to October by launching aggressive offers and advertising campaigns.
* This will entice first-time wine drinkers and fence sitters (who do not have specific loyalties to any particular brand).
* However, if there is no significant increase in sales by December, the company has two options: invest in R&D or consider discontinuing this variant and coming up with something completely new.

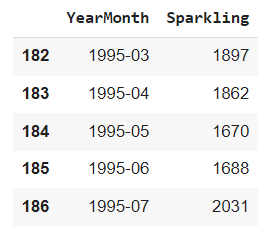


Final Model: Exponential Smoothing (Additive Seasonality - Rose

## Sparkling Dataset

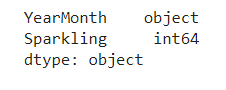
### Exploratory Data Analysis

#### Sample of the dataset:



Sample of Sparkling Data

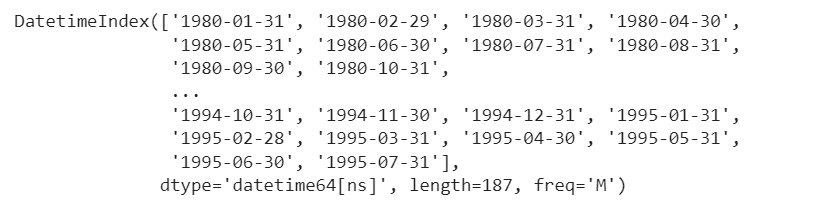
#### Data Types in the dataset before Data Modification:



Datatype of Sparkling Data

#### Data Modification for easiness of Time Forecasting:

Since we have monthly data for years starting from 1980 to 1995 combined, let us draw a date range and incorporate as "TimeStamp" and drop "YearMonth" for ease.

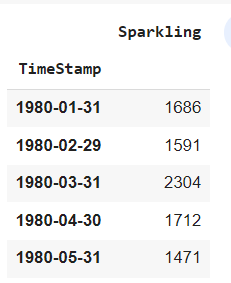


Date Index of Sparkling Data



TimeStamp addition to Sparkling Data

After we have added the TimeStamp, the “YearMonth” is dropped, and TimeStamp is made the index. This would be the modified Dataset we use going forward.

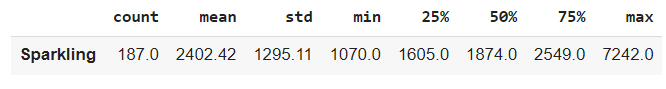


New Modified Sparkling Dataset

#### Basic details and Descriptive Statistics of Rose:



Length of Sparkling Dataset after modification



Descriptive Statistics of Sparkling Dataset after modification

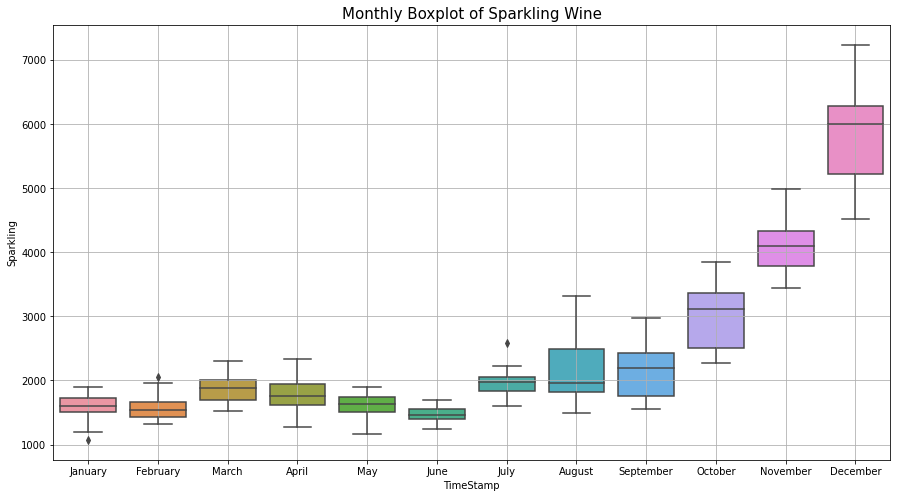
#### Missing Values and Interpolation:

As we checked the dataset, there are no null values. Let us go forward with a couple of analysis before building the models



No Null Values

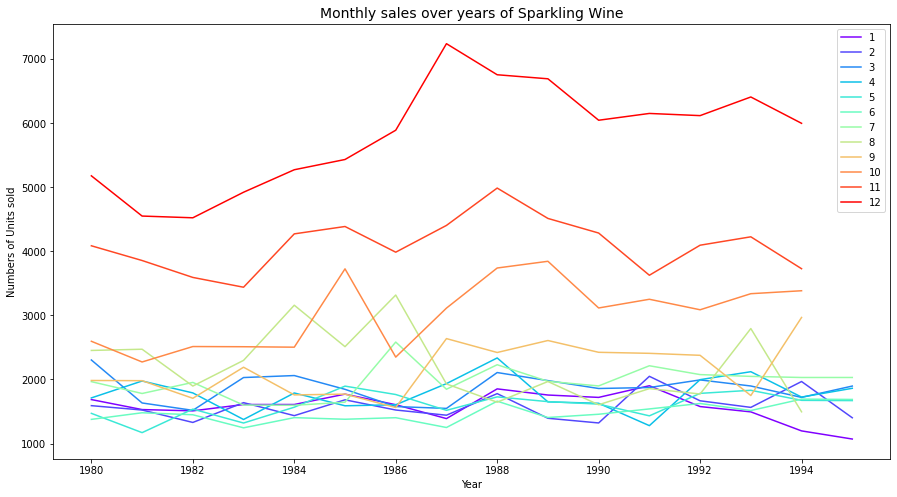
#### Monthly Boxplot of Sparkling:



Monthly Boxplot of Sparkling

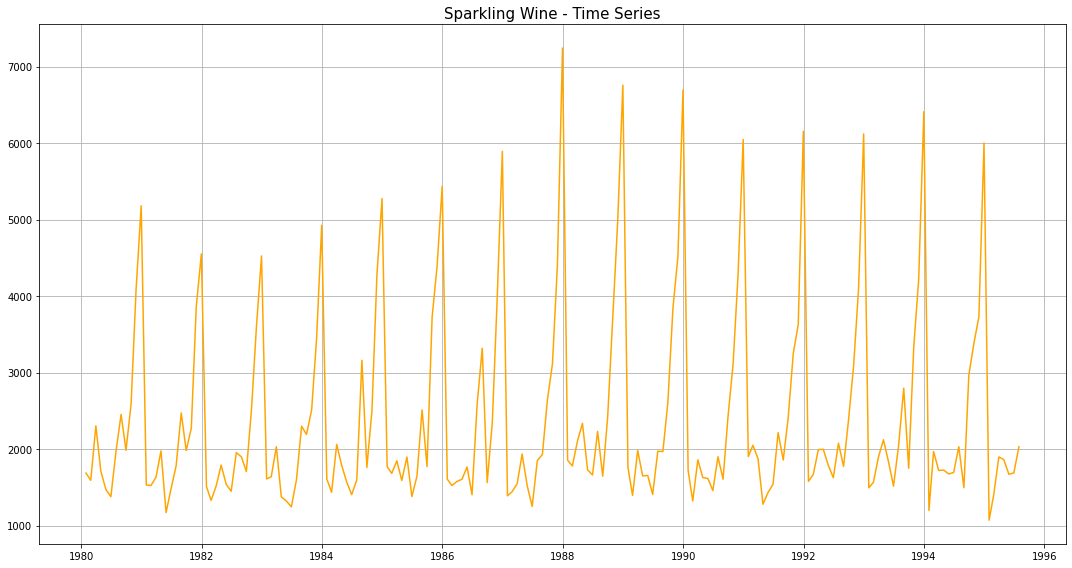
* Sales of both - Rose and Sparkling, show a spike in the last quarter of Oct to Dec
* Spike is much more accentuated in Sparkling sales
* This spike may be due to the Holiday season starting in Oct

#### Monthly sales over years of Sparkling Wine:



Monthly sales over years of Sparkling Wine

### Visualization of Time Series in Sparkling:

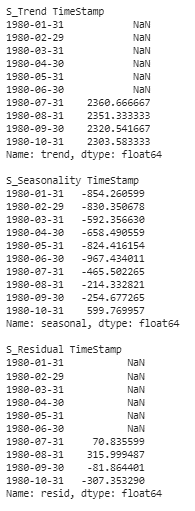


Visualization of Time Series in Sparkling

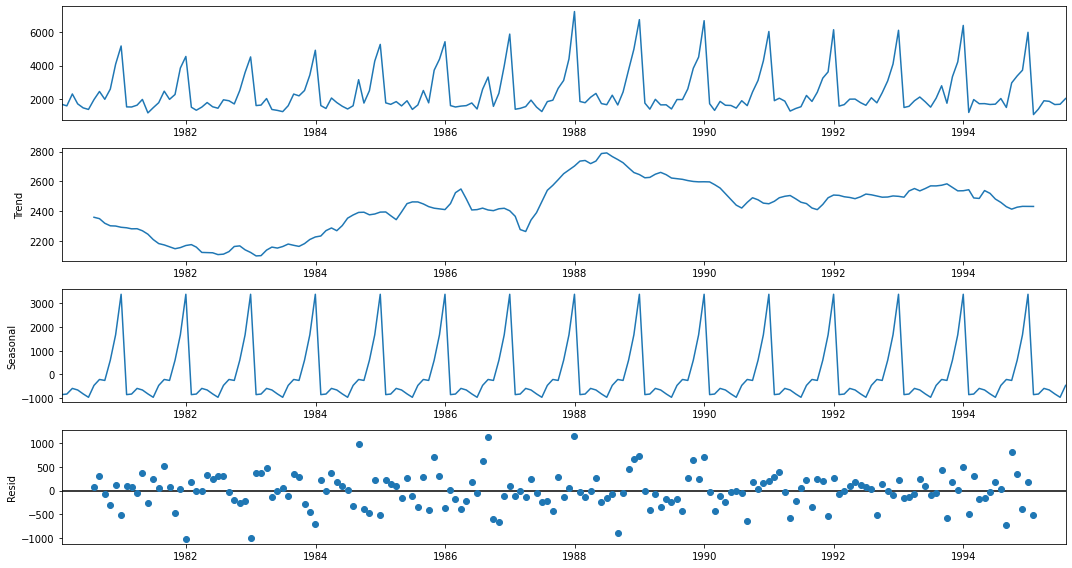
### Decomposition and plots of different component:

Additive Decomposition of Sparkling Dataset:

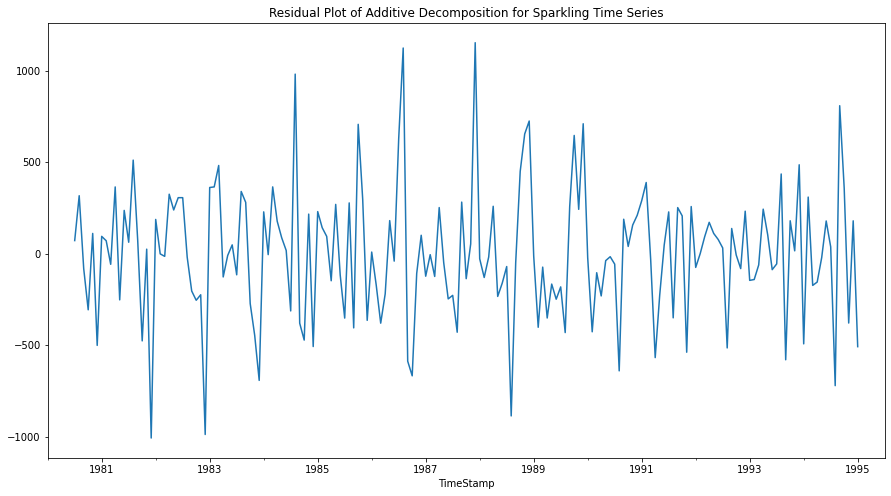
Once we decompose the Rose dataset using additive model, we obtain the following parameters and their respective plots.



Trend, Seasonality and Residue of Sparkling



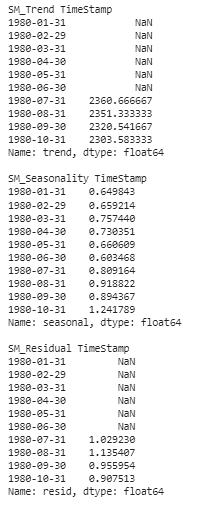
Trend, Seasonality and Residue plot for Sparkling



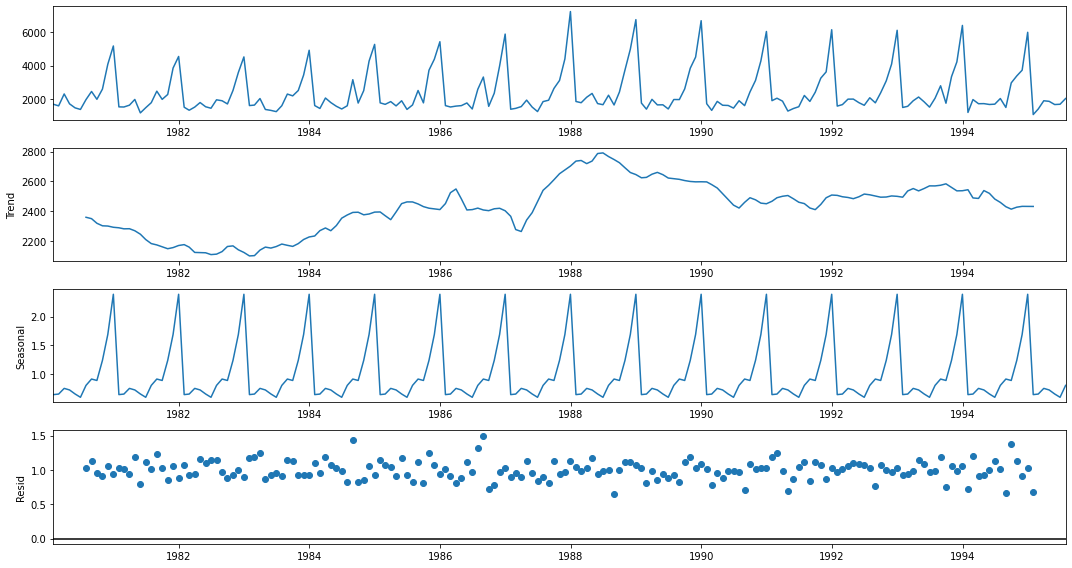
Residual Plot of Additive Decomposition for Sparkling Time Series

Multiplicative Decomposition of Sparkling Dataset:

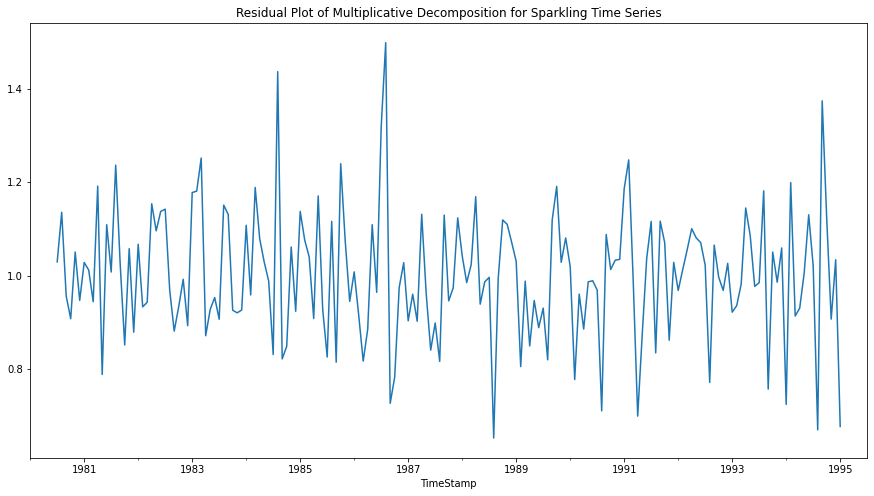
Once we decompose the Rose dataset using additive model, we obtain the following parameters and their respective plots.



Trend, Seasonality and Residue of Sparkling (Multiplicative)



Trend, Seasonality and Residue plot for Sparkling (Multiplicative)



Residual Plot of Additive Decomposition for Sparkling Time Series (Multiplicative)

Conclusion:

Additive Model-

The seasonality is relatively constant over time

Yt= Trend \* Seasonality \* Residual

Here by just observing the Residual patterns of Additive and Multiplicative models of Sparkling, we can conclude that Sparkling follows additive model.

### Test and Train Split:

The whole Rose data was split into test and training dataset such that the test data starts in 1991.

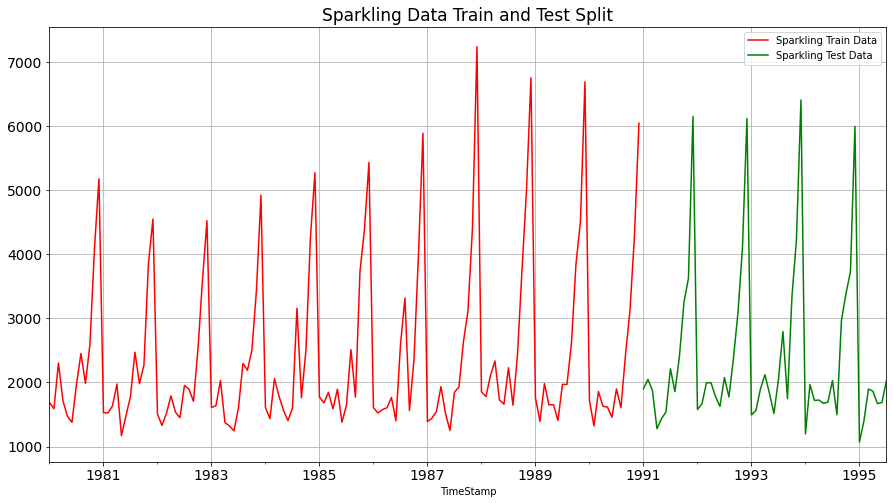


Fig 16. Test and training data for Sparkling

### Various Models on the Sparkling Dataset to forecast the time series:

#### Model 1 – Linear Regression

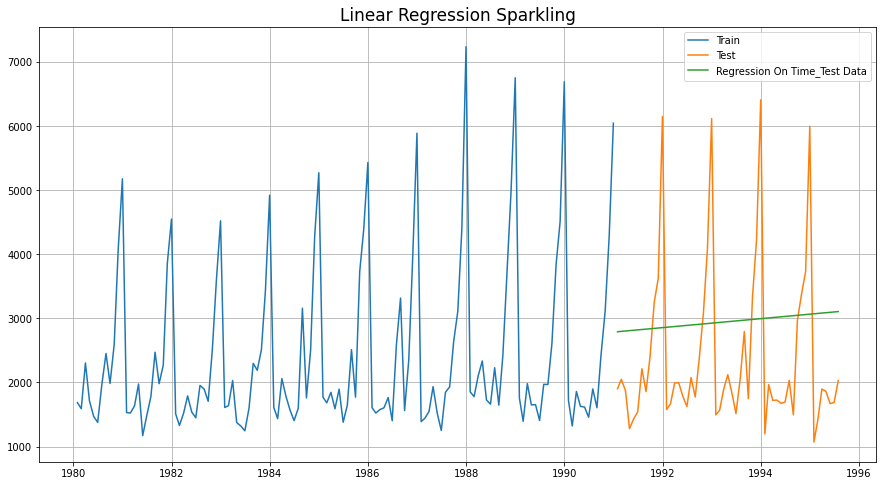


Fig 17. Linear Regression for Sparkling

#### Model 2 - Naive Bayes

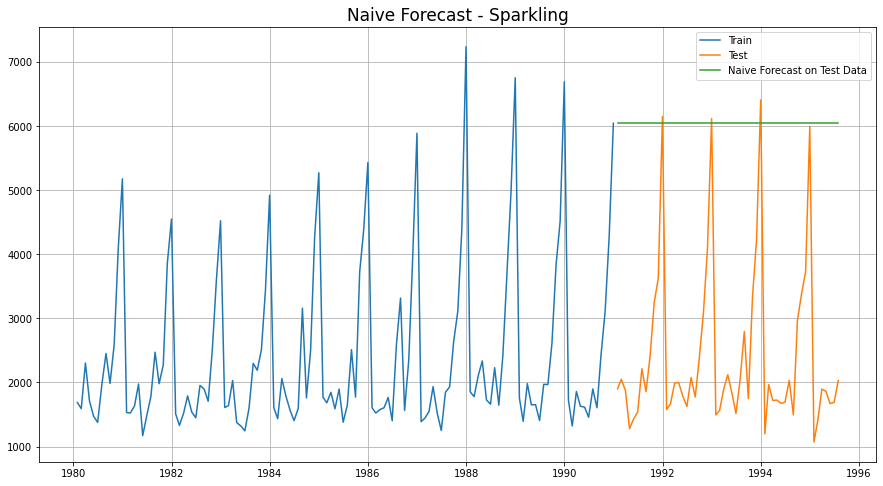


Fig 18. Naïve’s Approach for Sparkling

#### Model 3 - Simple Average

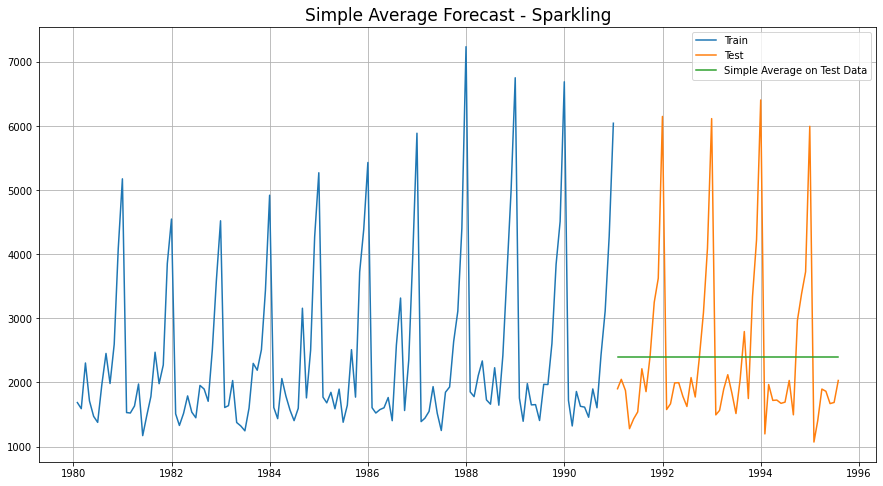
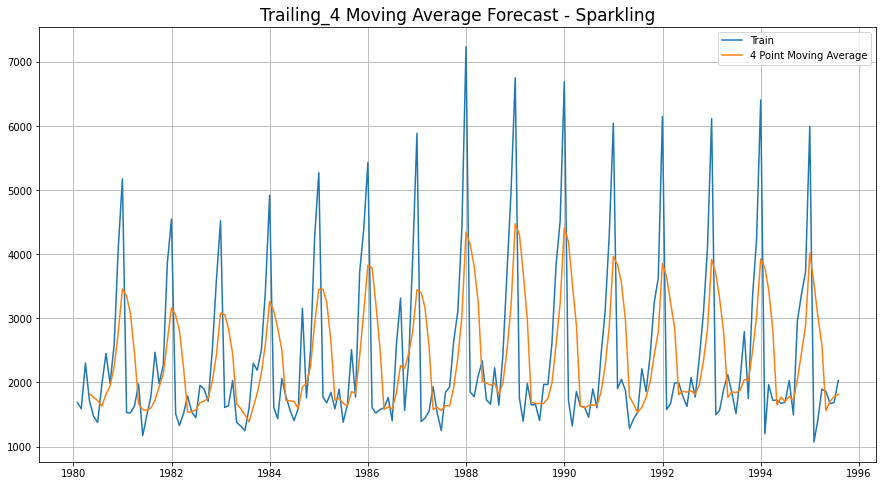
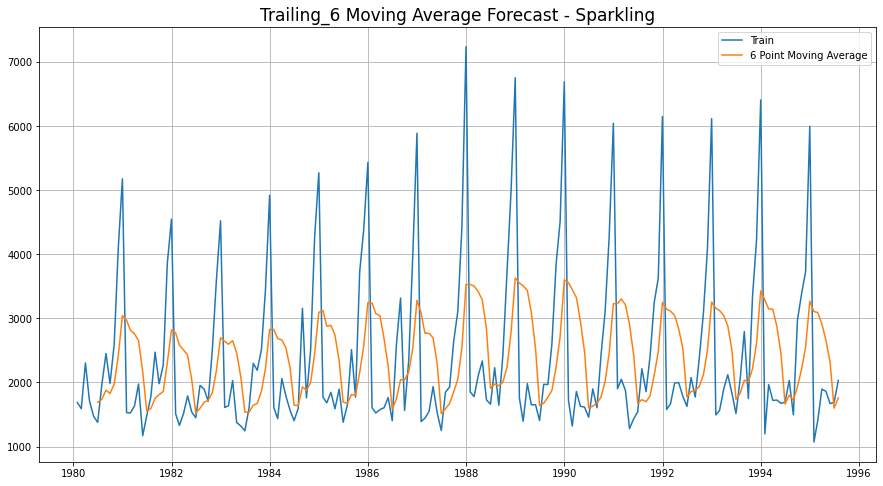


Fig 19. Simple Average Model for Sparkling

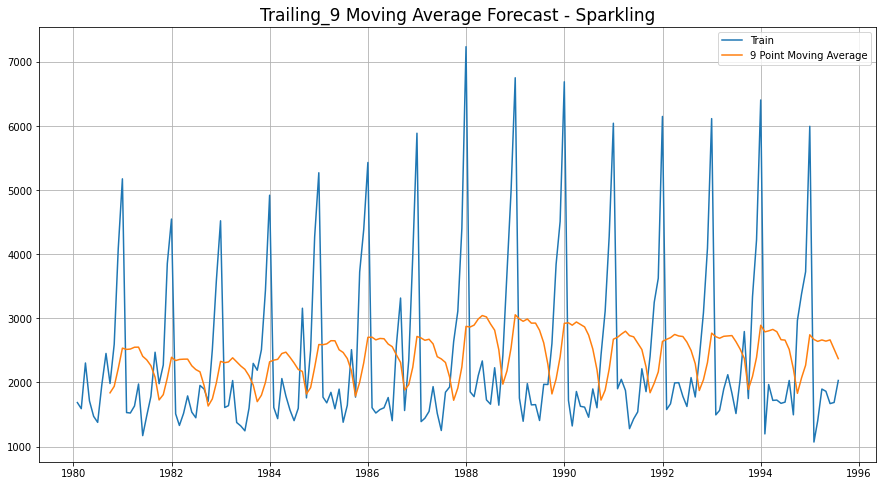
#### Model 4- Moving Average (Sparkling)



4 Point Trailing Moving Average for Sparkling



6 Point Trailing Moving Average for Sparkling



9 Point Trailing Moving Average for Sparkling

#### Consolidated Moving Average Forecasts (Sparkling)



4 Consolidated Trailing Moving Average for Sparkling

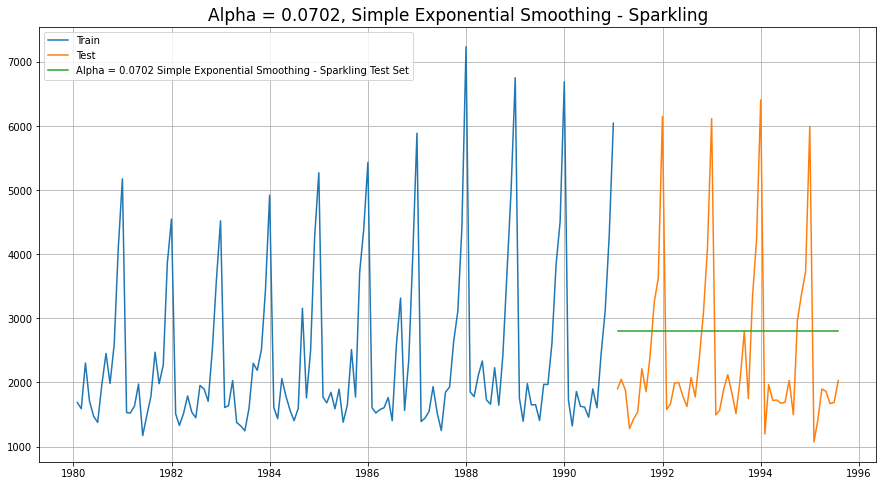
Let us see the best of all models built until now, before we go forward with Exponential Smoothing Models.

|  |  |
| --- | --- |
| **Model Name** | **Test RMSE Sparkling** |
| Regression On Time | 1389.135175 |
| Naive Model | 3864.279352 |
| Simple Average Model | 1275.081804 |
| 4point Trailing MovingAverage | 1156.589694 |
| 6point Trailing MovingAverage | 1283.927428 |
| 9point Trailing MovingAverage | 1346.278315 |

Consolidated Scores of Regression, Naive, Simple Average & Moving Average

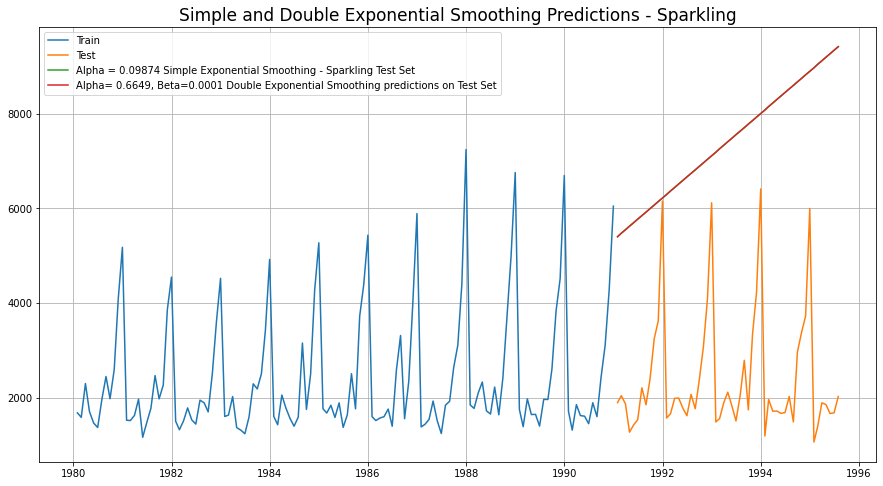
Till now, the best model built is **4point Trailing MovingAverage**

#### Model 5 - Single Exponential Smoothing



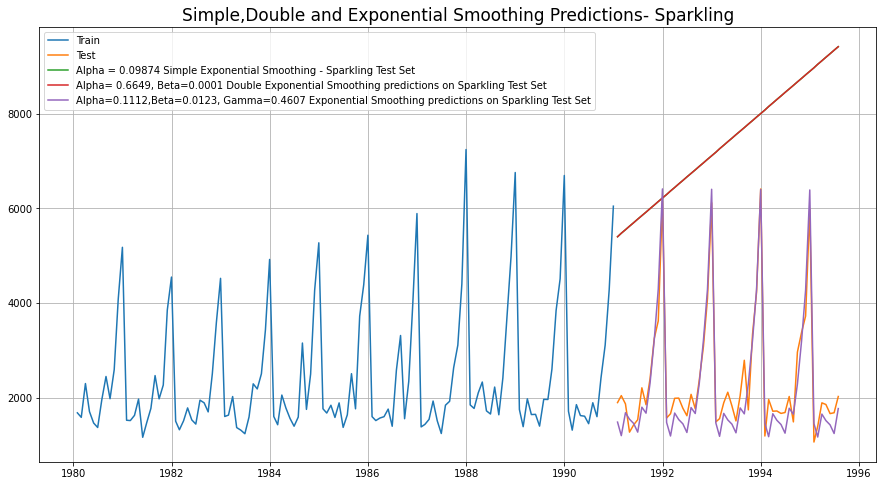
Single Exponential Smoothing

#### Model 6 - Double Exponential Smoothing



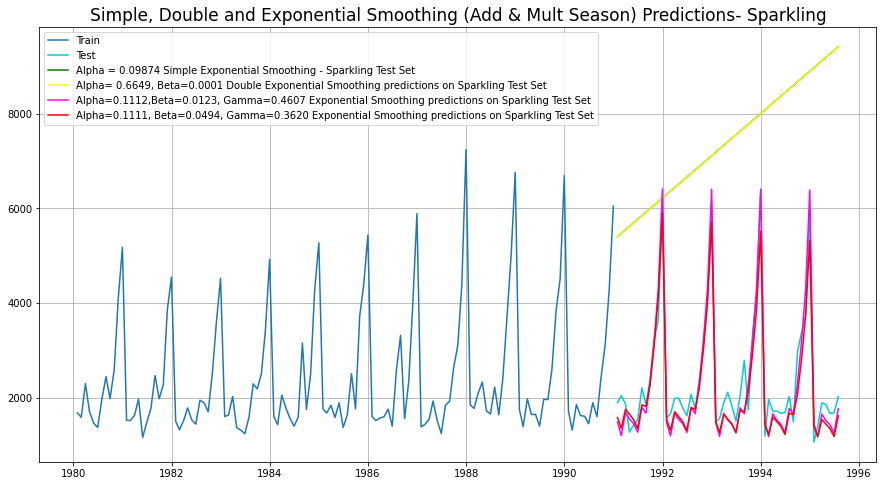
Single and Double Exponential Smoothing

#### Model 7 - Holt-Winter's Model (Exponential Smoothing) with additive errors



Simple, Double and Exponential Smoothing Predictions

#### Model 8 - Holt-Winter's Model (Exponential Smoothing) with multiplicative seasonality



Simple, Double and Exponential Smoothing (Add & Mult.) Season

Conclusion at the end of Exponential Modelling, we see that the Exponential Smoothing with Additive Seasonality is the best model (least RMSE)

|  |  |
| --- | --- |
| **Model Name** | **Test RMSE Rose** |
| Regression On Time | 1389.135175 |
| Naive Model | 3864.279352 |
| Simple Average Model | 1275.081804 |
| 4point Trailing MovingAverage | 1156.589694 |
| 6point Trailing MovingAverage | 1283.927428 |
| 9point Trailing MovingAverage | 1346.278315 |
| Simple Exponential Smoothing | 1338.004623 |
| Double Exponential Smoothing | 5291.879833 |
| Exponential Smoothing (Additive Seasonality) | 378.626241 |
| Exponential Smoothing (Multi Seasonality) | 403.706228 |

Consolidated Scores of Regression, Naive, Simple Average & Moving Average, Exponential Smoothing

#### Check for stationarity of the whole Time Series data

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

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

1. H0: The Time Series has a unit root and is thus non-stationary.
2. H1:The Time Series does not have a unit root and is thus stationary

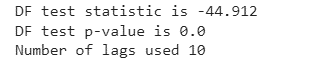
We would want the series to be stationary for building ARIMA models and thus we would want the p-value of this test to be less than the α (0.05) value.

Since the p-value is greater than 0.05, the Rose data set is non-stationary.

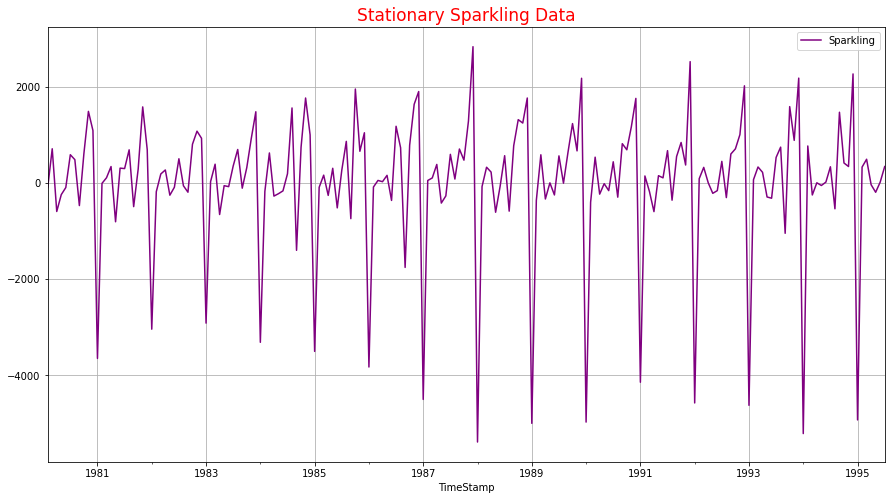


ADF test for Sparkling

To make the data stationary, lets difference and repeat the ADF for Rose

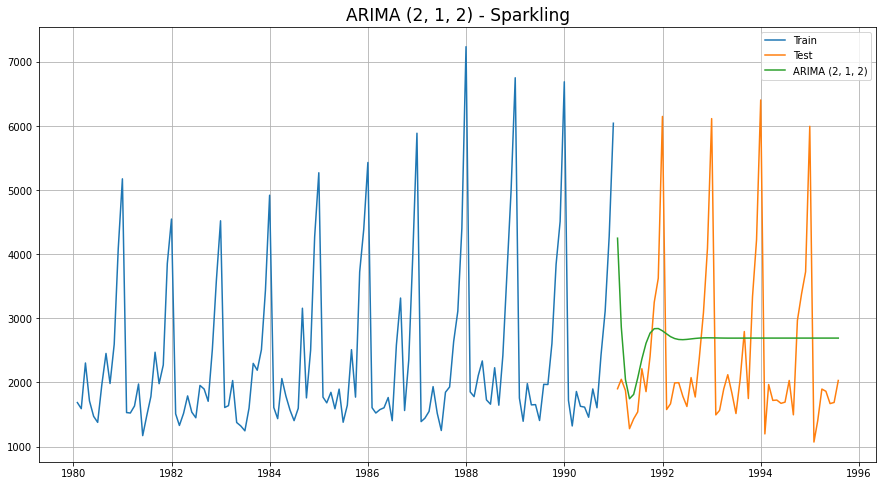


ADF test for Sparkling (diff)

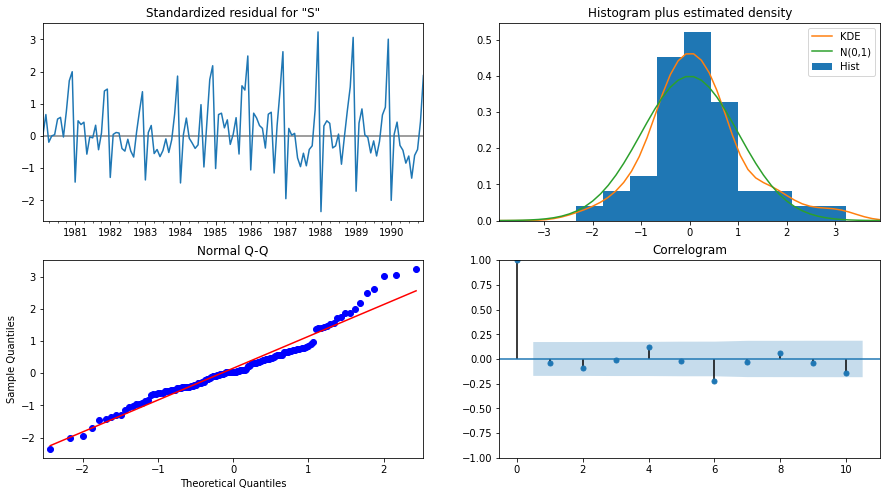


#### Model 9 – Automated ARIMA Model

* We create a grid of all possible combinations of (p, d, q)
* Range of p = Range of q = 0 to 3, Constant d = 1
* Few Examples of the grid -
  + - Model: (0, 1, 0)
    - Model: (0, 1, 1)
    - Model: (0, 1, 2)
    - Model: (0, 1, 3)
    - Model: (1, 1, 0)
    - Model: (1, 1, 1)
    - Model: (1, 1, 2)
    - Model: (1, 1, 3)
    - Model: (2, 1, 0)
    - Model: (2, 1, 1)
    - Model: (2, 1, 2)
    - Model: (2, 1, 3)
    - Model: (3, 1, 0)
    - Model: (3, 1, 1)
    - Model: (3, 1, 2)
    - Model: (3, 1, 3)
* We fit ARIMA models to each of these combinations for both datasets
* We choose the combination with the least Akaike Information Criteria (AIC) - ARIMA (2, 1, 2)
* We fit ARIMA to this combination of (p, d, q) to the Train set and forecast on the Test set
* Finally, we check the accuracy of this model by checking RMSE of Test set



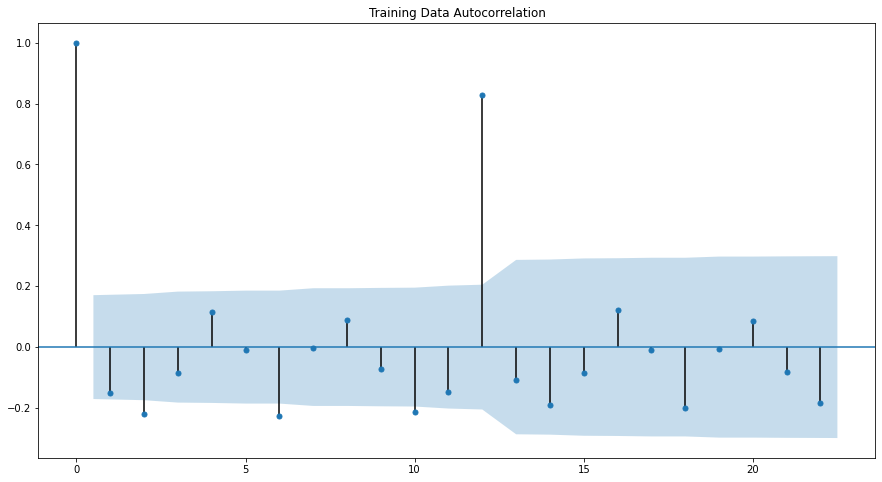
ARIMA (2, 1, 2) – Automated ARIMA



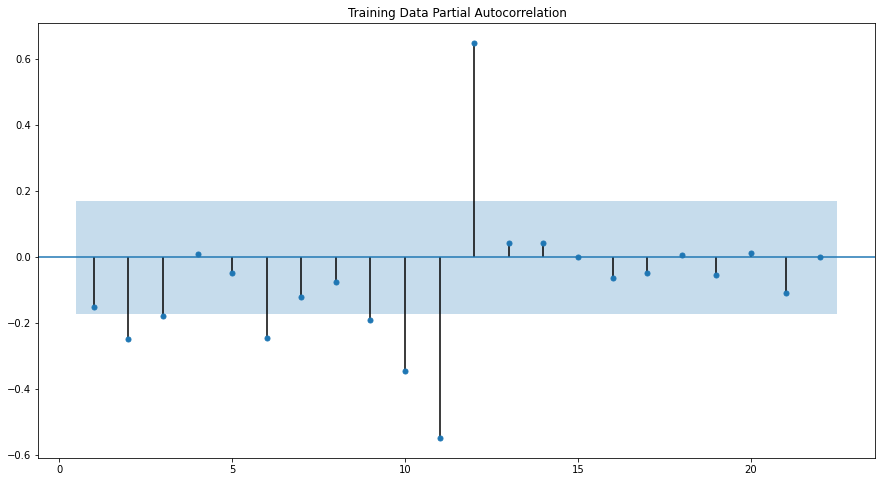
ARIMA (2, 1, 2) Diagnostic Plot – Sparkling

#### Model 10 – Manual ARIMA Model

* The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off to 0.
* The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 0.
* By looking at the above plots, we will take the value of p and q to be 0 and 0 respectively.

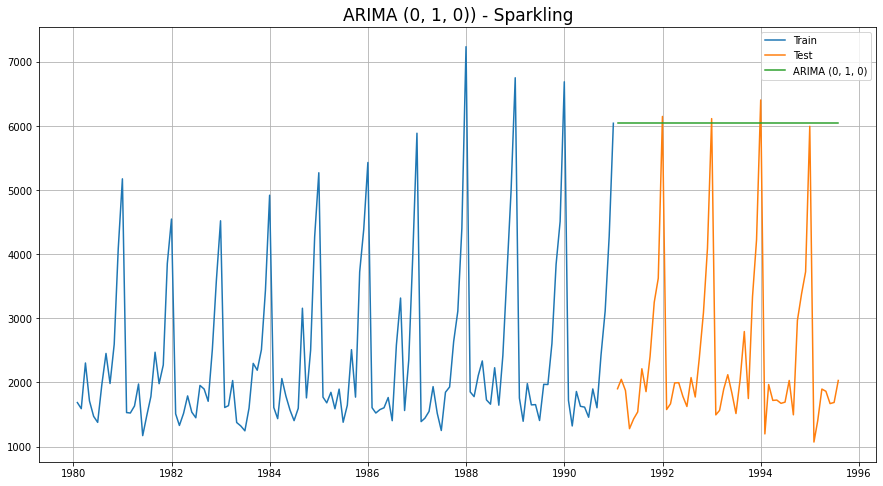


Training Data Autocorrelation

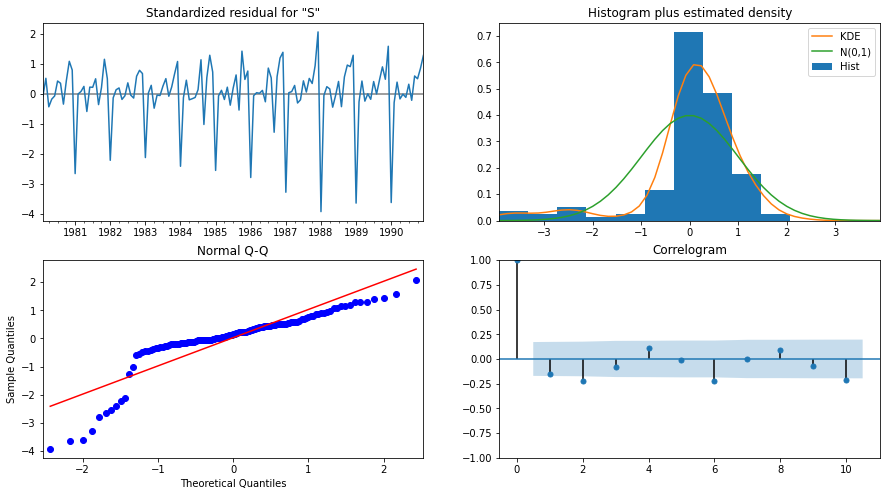


Training Data Partial Autocorrelation

ARIMA Manual - Rose - (0,1,0)



ARIMA (0,1,0) – Manual ARIMA



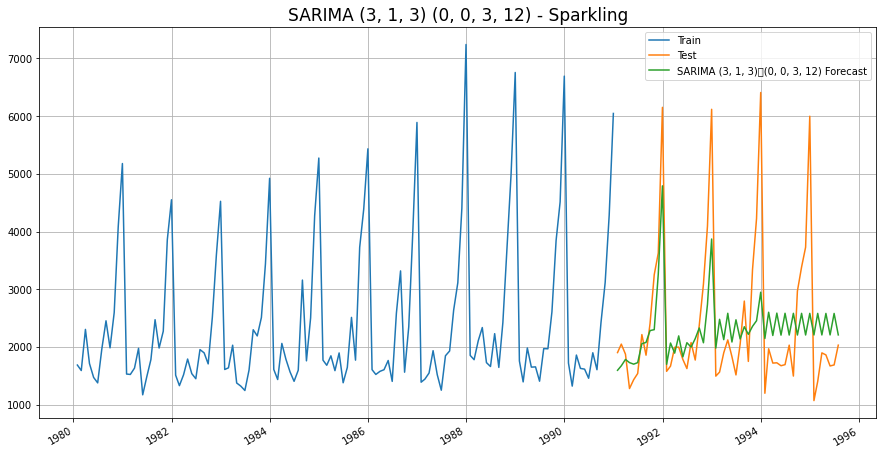
ARIMA (0,1,0) Diagnostic Plot - Rose

#### Model 11 – Automated SARIMA Model

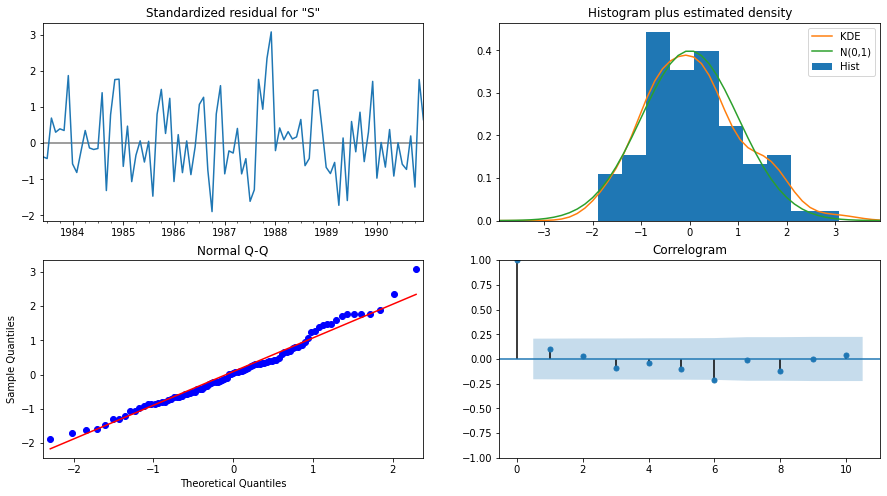
* We create a grid of all possible combinations of (p, d, q) along with Seasonal

(P, D, Q) & Seasonality of 12 (for both datasets)

* Range of p = Range of q = 0 to 3, Constant d = 1
* Range of Seasonal P = Range of Seasonal Q = 0 to 3, Constant D = 1, Seasonality m = 12
* Few Examples of the grid (p, d, q) (P, D, Q, m) -
  + Model: (0, 1, 1)(0, 0, 1, 12)
  + Model: (0, 1, 2)(0, 0, 2, 12)
  + Model: (0, 1, 3)(0, 0, 3, 12)
  + Model: (1, 1, 0)(1, 0, 0, 12)
  + Model: (1, 1, 1)(1, 0, 1, 12)
  + Model: (1, 1, 2)(1, 0, 2, 12)
  + Model: (1, 1, 3)(1, 0, 3, 12)
  + Model: (2, 1, 0)(2, 0, 0, 12)
  + Model: (2, 1, 1)(2, 0, 1, 12)
  + Model: (2, 1, 2)(2, 0, 2, 12)
  + Model: (2, 1, 3)(2, 0, 3, 12)
  + Model: (3, 1, 0)(3, 0, 0, 12)
  + Model: (3, 1, 1)(3, 0, 1, 12)
  + Model: (3, 1, 2)(3, 0, 2, 12)
  + Model: (3, 1, 3)(3, 0, 3, 12)
* We fit SARIMA models to each of these combinations and select with least AIC
* We fit SARIMA to this best combination of (p, d, q) (P, D, Q, m) to the Train set and forecast on the Test set. Then, we check accuracy using RMSE on Test set
* For Rose, Best Combination with Least AIC is - (3, 1, 3) (0, 0, 3, 12)



SARIMA (3, 1, 3)(0, 0, 3, 12)- Automated SARIMA

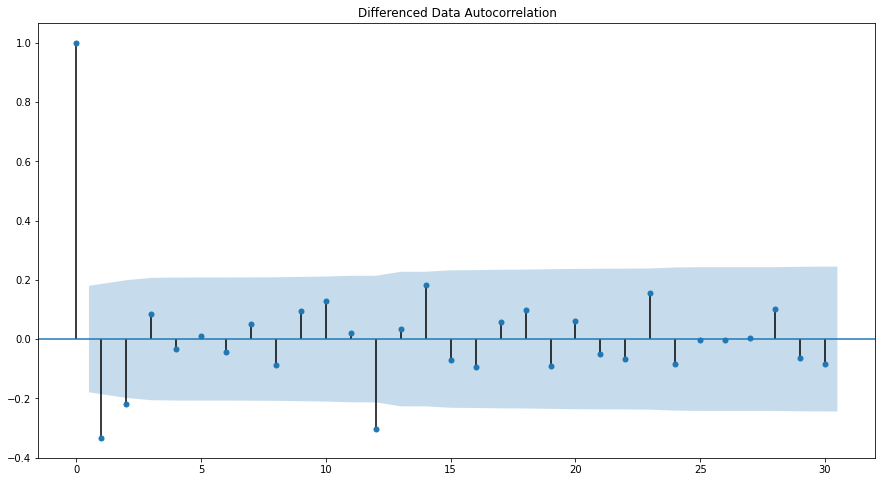


SARIMA (3, 1, 3)(0, 0, 3, 12)Diagnostic Plot – Rose

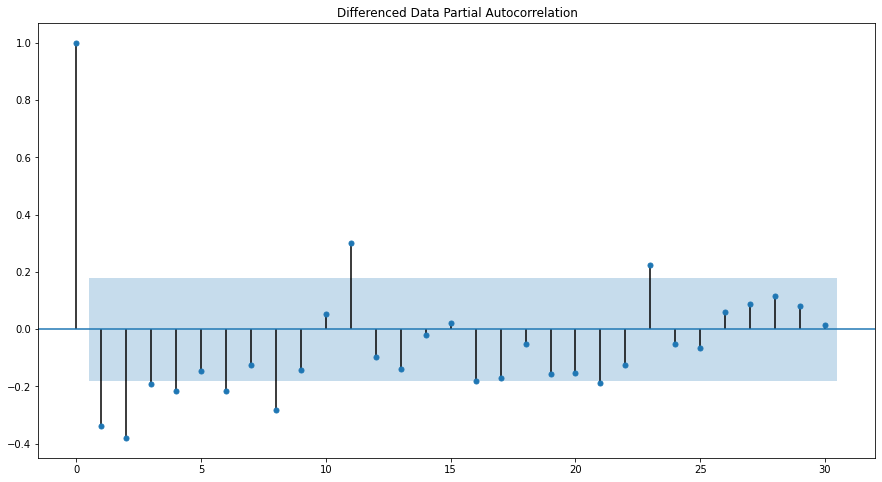
#### Model 12 – Manual SARIMA Model

We are going to take the seasonal period as 12 We are taking the p value to be 2 and the q value also to be 0 as the parameters same as the ARIMA model.

The Auto-Regressive parameter in a SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 4. The Moving-Average parameter in a SARIMA model is 'Q' which comes from the significant lag after which the ACF plot cuts-off to 2.

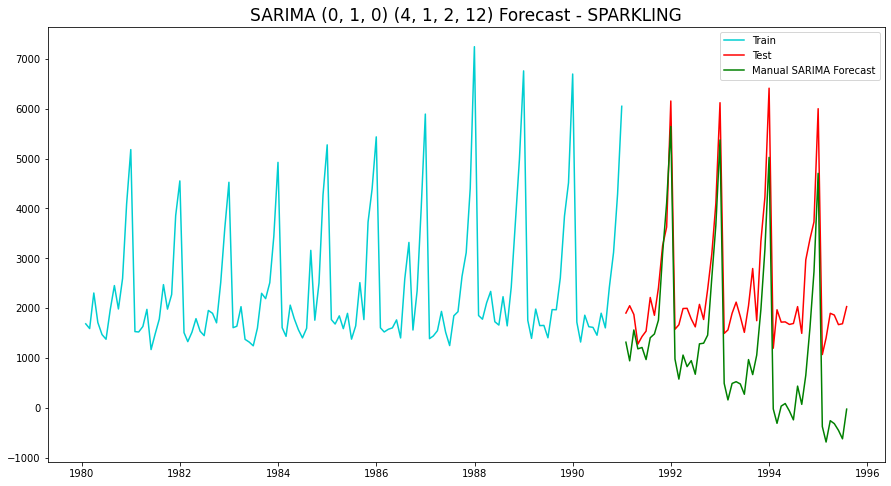


Training Data Autocorrelation



Training Data Partial Autocorrelation

SARIMA Manual - Sparkling - (0, 1, 0) (4, 1, 2, 12)



SARIMA (0, 1, 0) (4, 1, 2, 12)– Manual SARIMA



SARIMA (0, 1, 0) (4, 1, 2, 12) - Diagnostic Plot – Sparkling

### Data Frame with all the RMSE values for models built:

|  |  |
| --- | --- |
| **Model Name** | **Test RMSE Sparkling** |
| Regression On Time | 1389.135175 |
| Naive Model | 3864.279352 |
| Simple Average Model | 1275.081804 |
| 4point Trailing MovingAverage | 1156.589694 |
| 6point Trailing MovingAverage | 1283.927428 |
| 9point Trailing MovingAverage | 1346.278315 |
| Simple Exponential Smoothing | 1338.004623 |
| Double Exponential Smoothing | 5291.879833 |
| Exponential Smoothing (Additive Seasonality) | 378.626241 |
| Exponential Smoothing (Multi Seasonality) | 403.706228 |
| ARIMA (2, 1, 3) | 1299.98066 |
| ARIMA (2, 1, 2) | 3864.279352 |
| SARIMA (2, 1, 1) (1, 0, 3, 12) | 1000.643795 |
| SARIMA (2, 1, 2) (2, 1, 3, 12) | 1336.560606 |

Conclusion:

As we can see the above table, the lease RMSE out of all models built is Exponential Smoothing (Additive Seasonality).

Let us now build the model on full data and predict the upcoming 12 months.

### Final Model – Exponential Smoothing with additive seasonality



Final Model Forecast: Exponential Smoothing (Additive Seasonality) - Sparkling



Final Model: Exponential Smoothing (Additive Seasonality) - Sparkling

### Sparkling Wine Sales - Comments :

* Sparkling wine sales show no upward or downward trend - this indicates flat sales over the long term range.
* There is also a very high spike in sales seen in the last quarter of every year from October to December.
* The highest peak in sales is seen every year in December
* Dec sales are almost three times that of Sep sales
* Similar to Rose Wine, even in Sparkling sales, an instant crashing slump is seen in the first quarter of every year beginning in January
* This could be due to the after effect or hangover of Holidays