**Forecasting Wine Sales**

**Report**

Submission By: **Raj Keswani**

Contents

List of Figures

List of Tables

Data and Data Dictionary

1. [Sparkling](https://olympus.mygreatlearning.com/courses/80068/files/7429593/download?verifier=PXXyRkHm5ssggA1vno2Avd4EonkVjjL2dqltdXXp&wrap=1)
   1. Read the data as an appropriate Time Series data and plot the data.
   2. Perform appropriate Exploratory Data Analysis to understand the data and perform decomposition.
   3. Split the data into training and test. The test data should start in 1991.
   4. Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other additional models such as regression, naïve forecast models, simple average models, moving average models should also be built on the training data and check the performance on the test data using RMSE.
   5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment.  
      Note: Stationarity should be checked at alpha = 0.05.
   6. Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.
   7. Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.
   8. Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.
   9. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.
   10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.
2. Rose
   1. Read the data as an appropriate Time Series data and plot the data.
   2. Perform appropriate Exploratory Data Analysis to understand the data and perform decomposition.
   3. Split the data into training and test. The test data should start in 1991.
   4. Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other additional models such as regression, naïve forecast models, simple average models, moving average models should also be built on the training data and check the performance on the test data using RMSE.
   5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment.  
      Note: Stationarity should be checked at alpha = 0.05.
   6. Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.
   7. Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.
   8. Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.
   9. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.
   10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

**List of Figures**

|  |  |  |
| --- | --- | --- |
| **S.No.** | **Figure** | **Page No.** |
| 1.1.A | Sales Data for Sparkling Wine | 7 |
| 1.2.A | Dataframe Information | 8 |
| 1.2.B | Dataframe Statistical Description | 8 |
| 1.2.C | Sales Distribution | 8 |
| 1.2.D | Additive Decomposition | 9 |
| 1.2.E | Multiplicative Decomposition | 9 |
| 1.4.A | Linear Regression Model | 10 |
| 1.4.B | Naïve Forecast Model | 11 |
| 1.4.C | Simple Average Model | 11 |
| 1.4.D | SES-A model | 12 |
| 1.4.E | DES-A model | 12 |
| 1.4.F | TES-A model | 13 |
| 1.4.G | TES-M model | 14 |
| 1.5.A | ADF Test Results for training data. | 15 |
| 1.5.B | ADF Test Results for differenced training data. | 15 |
| 1.5.c | Differenced Training Data. | 15 |
| 1.6.A | Automated ARIMA Model | 17 |
| 1.6.B | Automated SARIMA Model | 18 |
| 1.7.A | ACF and PACF Plots | 19 |
| 1.7.B | Manual ARIMA Model | 20 |
| 1.7.C | Manual SARIMA Model | 21 |
| 1.9.A | Predicted 12-month data. | 23 |
| 1.10.A | Last Available Vs Predicted 12-month data description. | 24 |
| 1.10.B | Predicted 12-month data trend | 24 |
| 1.10.C | Last available actual 12-month data | 24 |
| 2.1.A | Sales Data for Rose Wine | 25 |
| 2.2.A | Dataframe Information | 26 |
| 2.2.B | Null values. | 26 |
| 2.2.C | Dataframe Statistical Description | 26 |
| 2.2.D | Sales Distribution | 27 |
| 2.2.E | Additive Decomposition | 27 |
| 2.2.F | Multiplicative Decomposition | 28 |
| 2.4.A | Linear Regression Model | 29 |
| 2.4.B | Naïve Forecast Model | 30 |
| 2.4.C | Simple Average Model | 30 |
| 2.4.D | SES-A model | 31 |
| 2.4.E | DES-A model | 31 |
| 2.4.F | TES-A model | 32 |
| 2.4.G | TES-M model | 33 |
| 2.5.A | ADF Test Results for training data. | 34 |
| 2.5.B | ADF Test Results for differenced training data. | 34 |
| 2.5.c | Differenced Training Data. | 34 |
| 2.6.A | Automated ARIMA Model | 36 |
| 2.6.B | Automated SARIMA Model | 37 |
| 2.7.A | ACF and PACF Plots | 38 |
| 2.7.B | Manual ARIMA Model | 39 |
| 2.7.C | Manual SARIMA Model | 40 |
| 2.9.A | Predicted 12-month data. | 42 |
| 2.10.A | Last Available Vs Predicted 12-month data description. | 43 |
| 2.10.B | Predicted 12-month data trend | 43 |
| 2.10.C | Last available actual 12-month data | 43 |

**List of Tables­­­­­­­­­­­­­­**

|  |  |  |
| --- | --- | --- |
| **S.No.** | **Table** | **Page No.** |
| 1.6.1 | AIC Values for Various ARIMA Models. | 16 |
| 1.6.2 | ARIMA(2, 1, 2) Model | 16 |
| 1.6.3 | AIC Values for Various SARIMA Models. | 17 |
| 1.6.4 | SARIMA(1, 1, 1)(1, 0, 3, 12) Model | 18 |
| 1.7.1 | ARIMA(3, 1, 2) Model | 20 |
| 1.7.2 | SARIMA(3, 1, 2)(3, 0, 1, 12) Model | 21 |
| 1.8.1 | Test RMSE values for various models. | 22 |
| 2.6.1 | AIC Values for Various ARIMA Models. | 35 |
| 2.6.2 | ARIMA(2, 1, 3) Model | 35 |
| 2.6.3 | AIC Values for Various SARIMA Models. | 36 |
| 2.6.4 | SARIMA(3, 1, 3)(3, 0, 1, 12) Model | 37 |
| 2.7.1 | ARIMA(2, 1, 2) Model | 39 |
| 2.7.2 | SARIMA(2, 1, 2)(1, 0, 1, 12) Model | 40 |
| 2.8.1 | Test RMSE values for various models. | 41 |

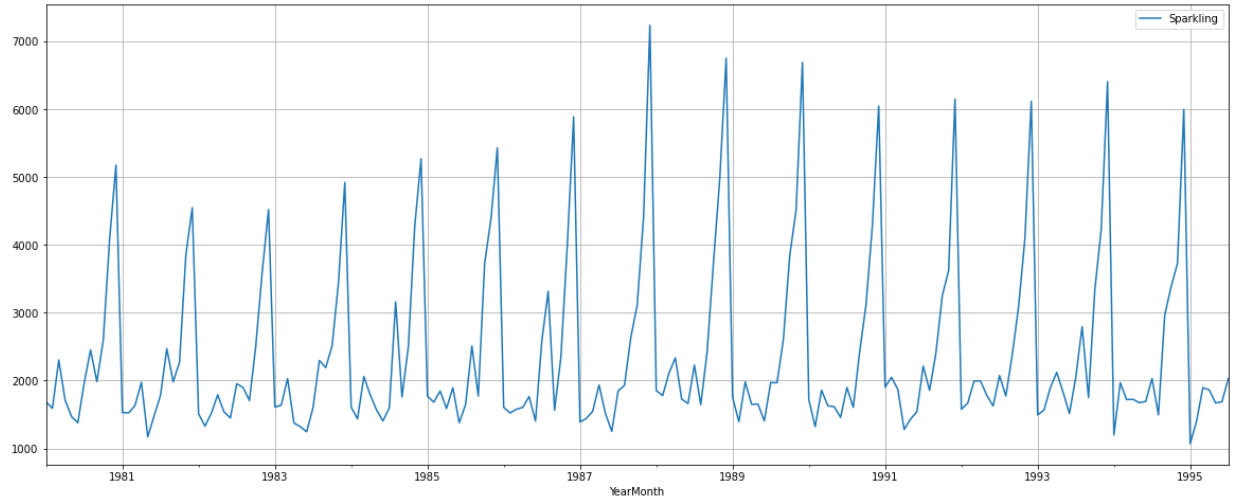
**­­­­­­­Data and Data Dictionary.**

* There are two data sets provided.
* One data set has sales records for sparkling wine.
* One dataset has sales records for rose wine.
* Both the wines belong to a single company.
* Both the datasets have 2 columns: YearMonth and Sales (with name Sparkling/Rose).
* It is a time series data with periodicity of 1 month.
* YearMonth column contains the year and month details.
* Sales column contains the cumulative sales for the corresponding month.

**1. Sparkling**

1.1 Read the data as an appropriate Time Series data and plot the data.

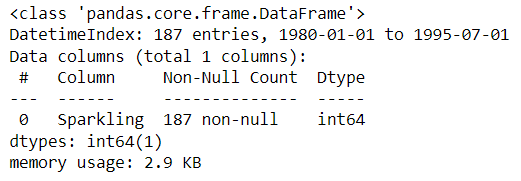
* The data was read using read\_csv function from Pandas library.
* The YearMonth column was parsed to be of Date format and was set as index column.
* The data was plotted and can be visualized in the following plot.



**Figure 1.1.A** Sales Data for Sparkling Wine

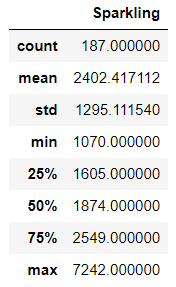
1.2 Perform appropriate Exploratory Data Analysis to understand the data and perform decomposition.

* There were 187 unique values in the dataset. No null values were found.



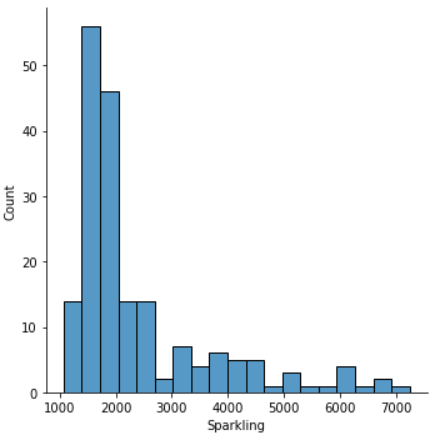
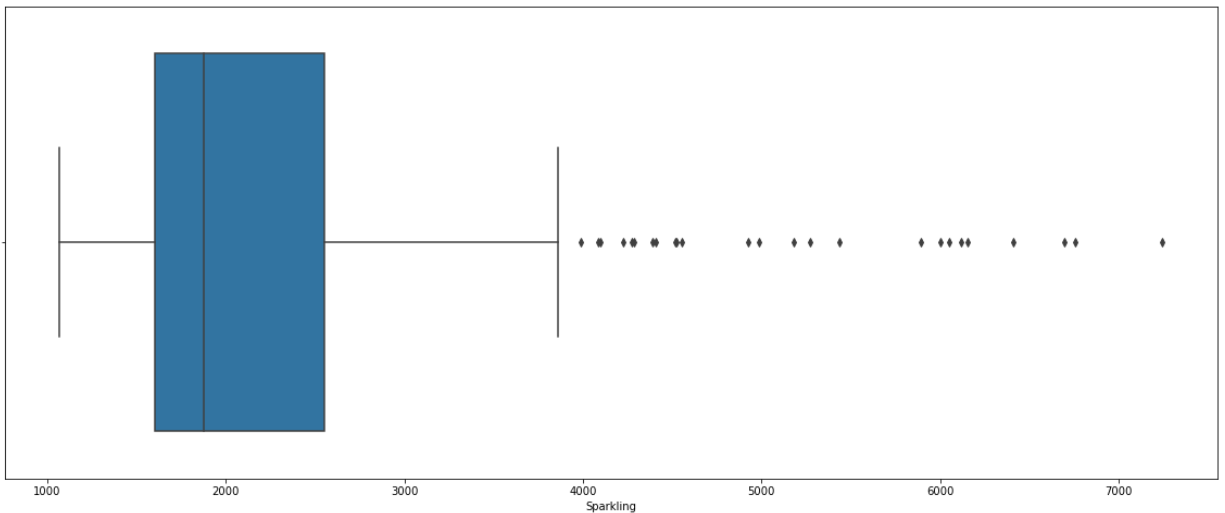
**Figure 1.2.A** Dataframe Information

* The monthly sales were averaged at 2402.
* Maximum sales recorded were 7242.
* Minimum sales recorded were 1070.



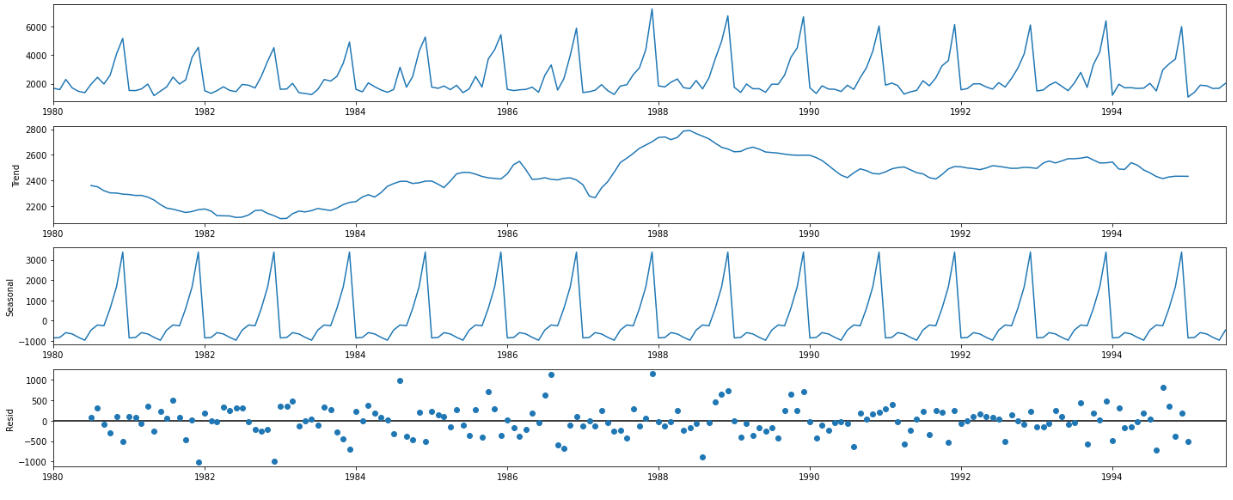
**Figure 1.2.B** Dataframe Statistical Description

* Most of the months, the sales are in between 1000-2700. Followed by fewer months with higher sales from 2700-4000 and not many months with exceptional sales >4000.

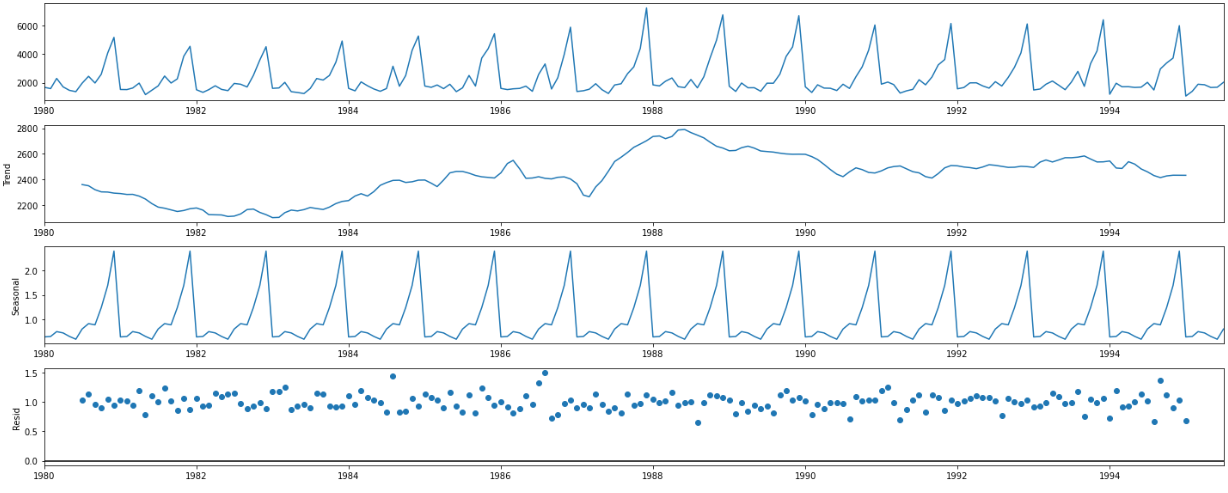
**Figure 1.2.C** Sales Distribution

* Additive decomposition was performed for the time series and the results are shared below.



**Figure 1.2.D** Additive Decomposition

* Both the trend and seasonality components were observed in the series.
* The residuals seemed to have a pattern suggesting additive decomposition was unable to completely decompose the series. Hence, we moved to multiplicative decomposition.
* Multiplicative decomposition was performed for the time series and the results are shared below.



**Figure 1.2.E** Multiplicative Decomposition

* Multiplicative decomposition was able to perform better suggesting the nature of seasonality is multiplicative.

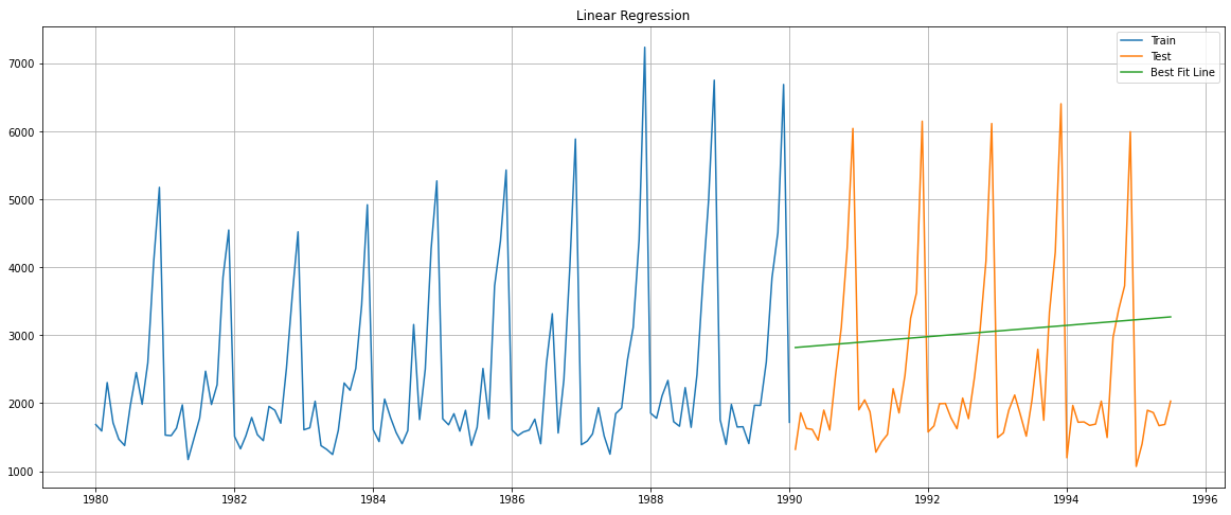
1.3. Split the data into training and test. The test data should start in 1991.

* The data frame was split into training and testing data using python slicing feature.
* The data until 1990 was used as a training data and the data from 1991 and beyond was used as testing 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.

**Regression Model**

* Linear Regression model was built and validated with the test data.

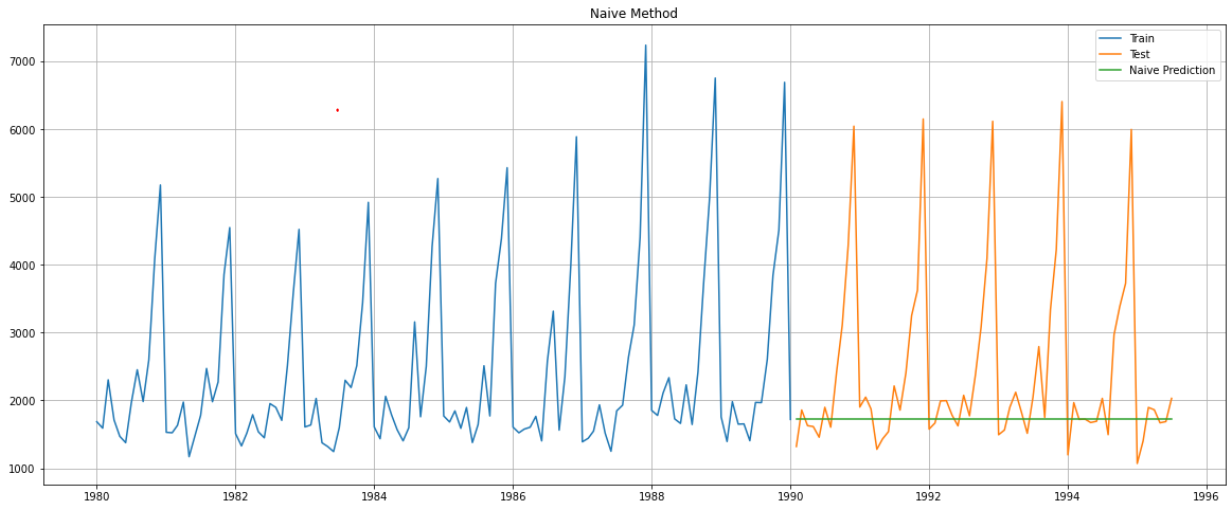


**Figure 1.4.A** Linear Regression Model

* As it can be clearly observed, the Linear regression model failed to explain the nature of test series.
* Even though it might give an idea about the trend but fails miserably to explain the seasonality of the series.
* RMSE vale of 1446.9 was observed.

**Naïve Forecast Model**

* Naïve forecast model was built and validated across the test data.

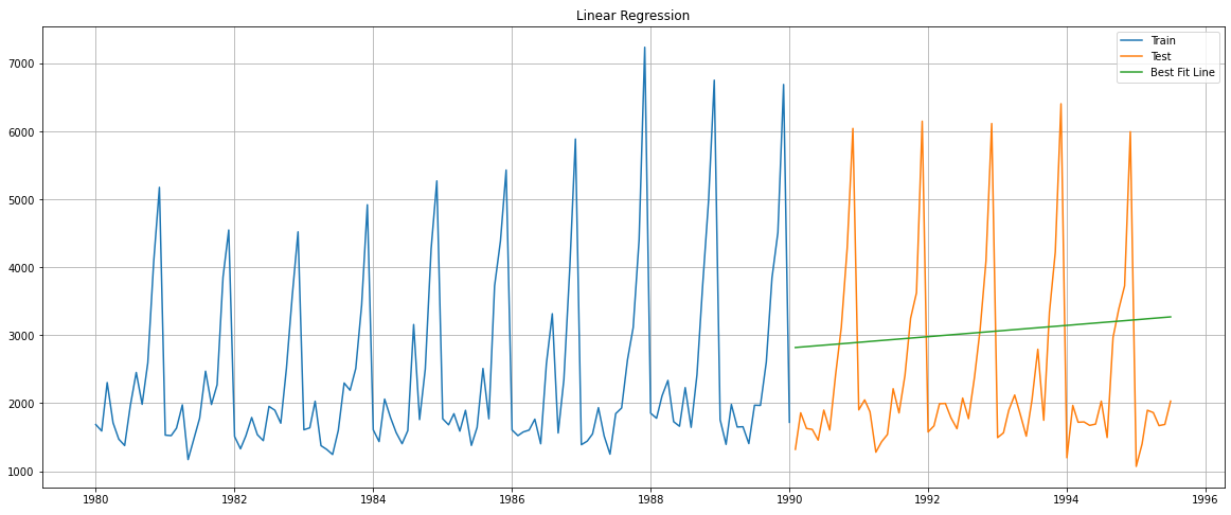


**Figure 1.4.B** Naïve Forecast Model

* As it can be clearly observed, the Naïve forecast model failed to explain the nature of test series.
* It fails to explain both the trend and the seasonality component of the time series.
* RMSE vale of 1471.2 was observed which was even higher than the regression model.

**Simple Average Model**

* Simple Average model was built and validated with the test data.

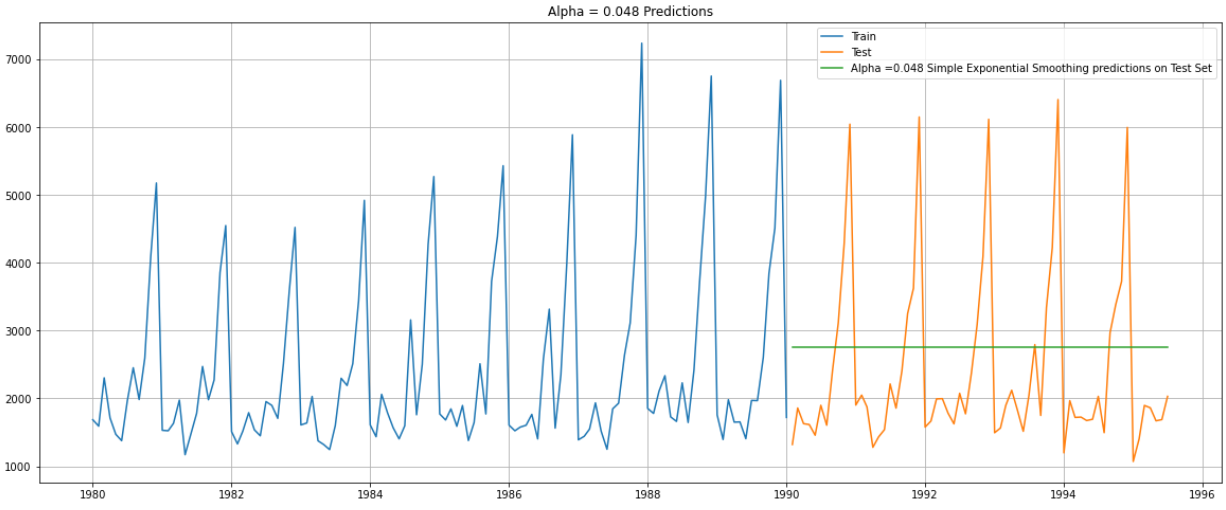


**Figure 1.4.C** Simple Average Model

* As it can be clearly observed, the Linear regression model failed to explain the nature of test series.
* Even though it might give an idea about the trend but fails miserably to explain the seasonality of the series.
* RMSE vale of 1298.2 was observed.
* It performed better than both regression and naïve forecasting models for this dataset.

### **Simple Exponential Smoothing Model with Additive Errors**

* SES-A model was built and validated with test data.

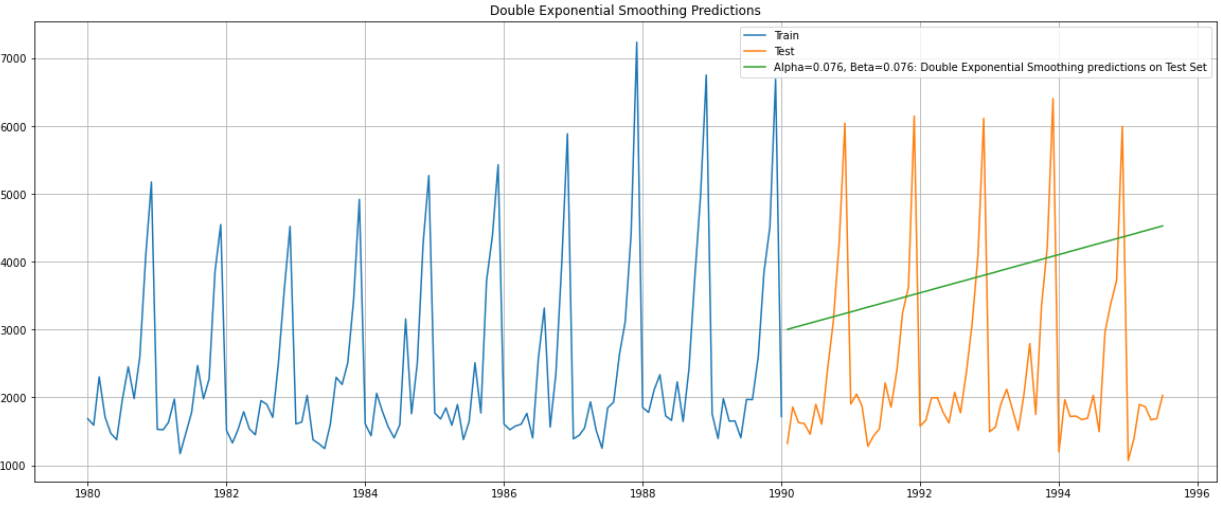


**Figure 1.4.D** SES-A model

* Alpha value was found to be 0.048.
* As it can be clearly observed, the SES model failed to explain the nature of test series.
* It fails to explain both the trend and the seasonality component of the time series.
* RMSE vale of 1344.4 was observed. It performed worse than simply average model.

### **Double Exponential Smoothing Model with Additive Errors**

* DES-A model was built and validated with test data.



**Figure 1.4.E** DES-A model

* Alpha and Beta, both the values were observed to be 0.76.
* As it can be clearly observed, the DES model failed to explain the nature of test series.
* Even though it might give an idea about the trend but fails miserably to explain the seasonality of the series.
* RMSE vale of 1919.8 was observed. It performed worse than SES model.

### **Triple Exponential Smoothing Model with Additive Errors**

* TES-A model was built and validated with test data.

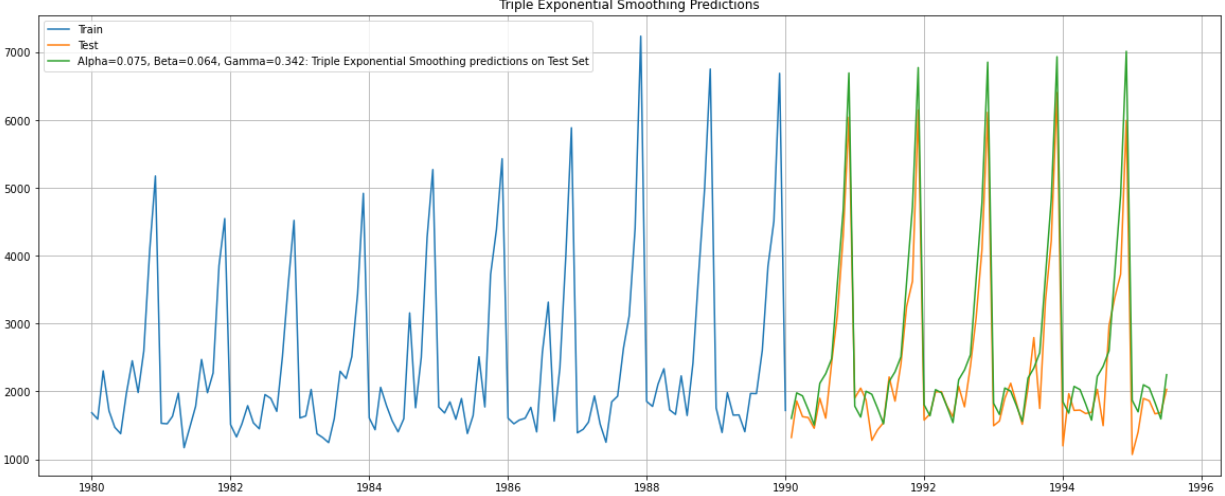


**Figure 1.4.F** TES-A model

* Alpha, Beta and Gama values were found to be 0.068, 0.035 and 0.434.
* TES-A model explains both the trend and seasonality of the time series.
* RMSE vale of 481.5 was observed. It performed much better than all other models tried before.

### **Triple Exponential Smoothing Model with Multiplicative Errors**

* TES-M model was built and validated with test data.



**Figure 1.4.G** TES-M model

* Alpha, Beta and Gama values were found to be 0.075, 0.064 and 0.342.
* TES-M model explains both the trend and seasonality of the time series.
* RMSE vale of 441.7 was observed. It performed much better than all other models tried before.

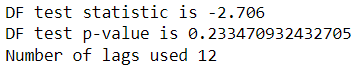
1.5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05.

The Augmented Dickey-Fuller test is a 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:

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

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

* ADF test was performed on the training data and following observations were made.



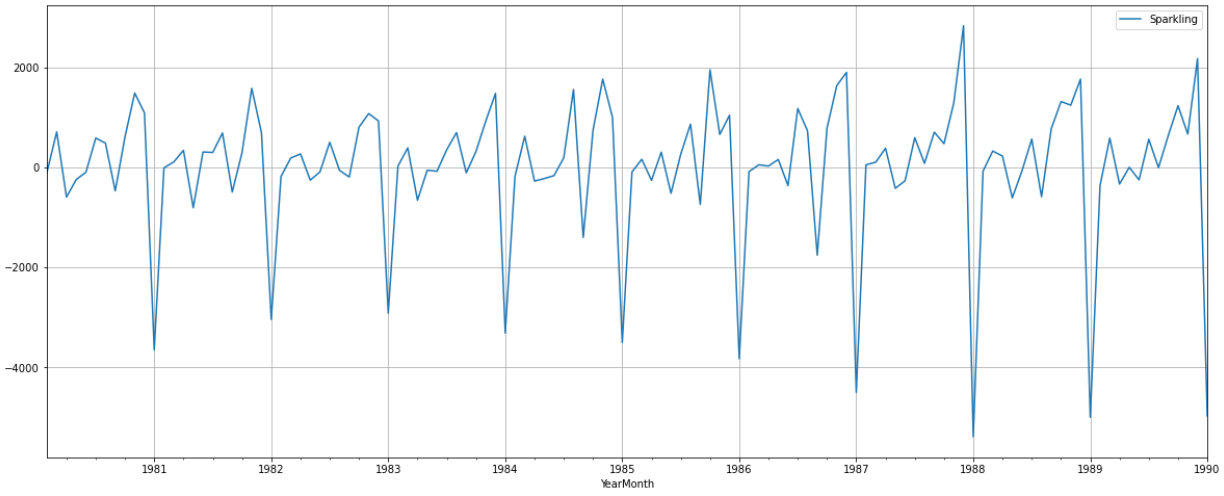
**Figure 1.5.A** ADF Test Results for training data.

* P-value is > 0.05, thus, null hypothesis cannot be rejected, implying the training data is not stationary.
* Since, the training data was found to be non-stationary, differencing was applied to the training data and differenced training data was again tested using ADF.



**Figure 1.5.B** ADF Test Results for differenced training data.

* P-value is much less than 0.05, thus, null hypothesis can be rejected, implying the differenced training data is stationary.

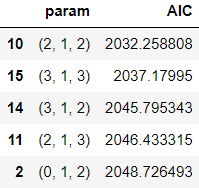


**Figure 1.5.C** Differenced Training Data.

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.

**Automated ARIMA Model**

* Multiple values of p, d and q were used and ARIMA models were built for various permutations and combinations.
* AIC values were calculated for each model.
* Minimum AIC value of 2032.25 was observed for a value of 2, 1 and 2 for p, d and q respectively.

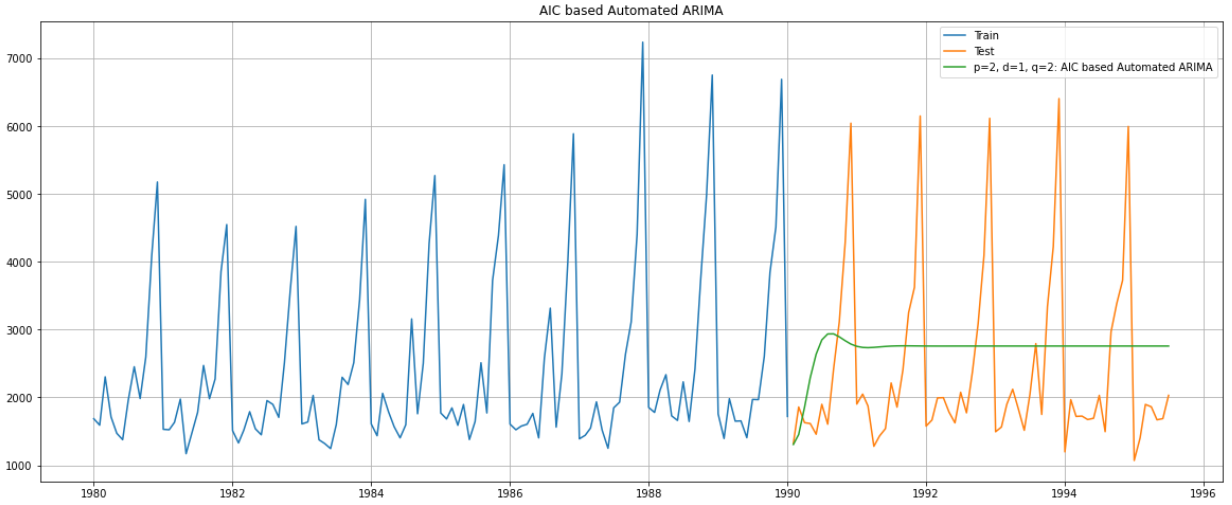


**Table 1.6.1** AIC Values for Various ARIMA Models.

* ARIMA(2, 1, 2) Model was picked and built and is shared as follows.



**Table 1.6.2** ARIMA(2, 1, 2) Model



**Figure 1.6.A** Automated ARIMA Model

* RMSE value was found to be 1314.91.
* The model was unable to explain the trend and the seasonality.

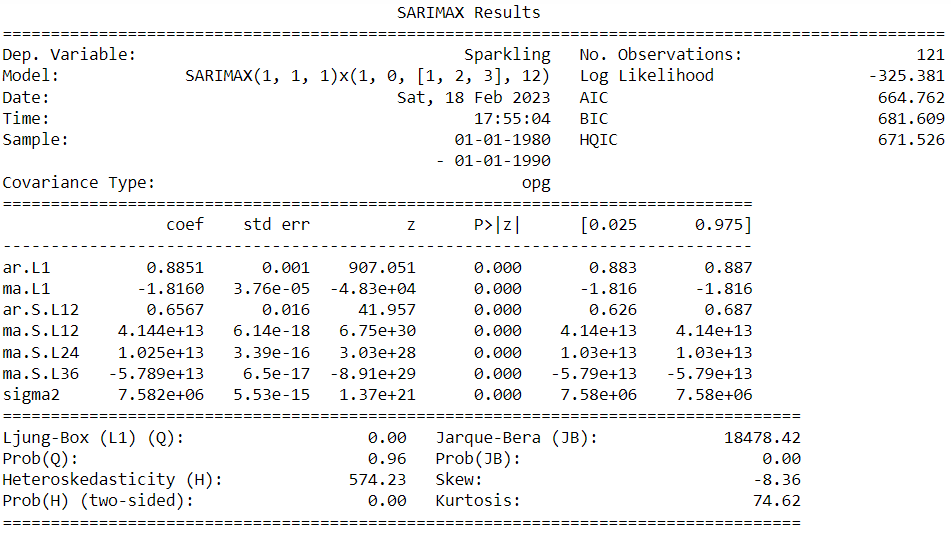
**Automated SARIMA Model**

* Multiple values of p, d, q, P, D, Q and F=12 were used and SARIMA models were built for various permutations and combinations.
* AIC values were calculated for each model.
* Minimum AIC value of 664.76 was observed for a value of 1, 1, 1, 1, 0 , 3 and 12 for p, d, q, P, D, Q and F respectively.

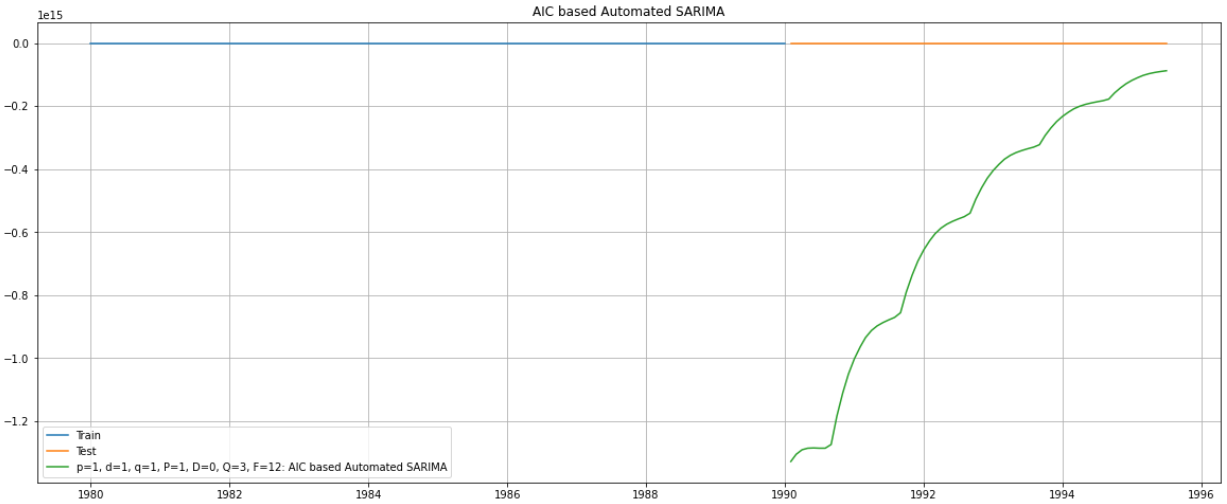


**Table 1.6.3** AIC Values for Various SARIMA Models.

* SARIMA(1, 1, 1)(1, 0, 3, 12) Model was picked and built and is shared as follows.



**Table 1.6.4** SARIMA(1, 1, 1)(1, 0, 3, 12) Model



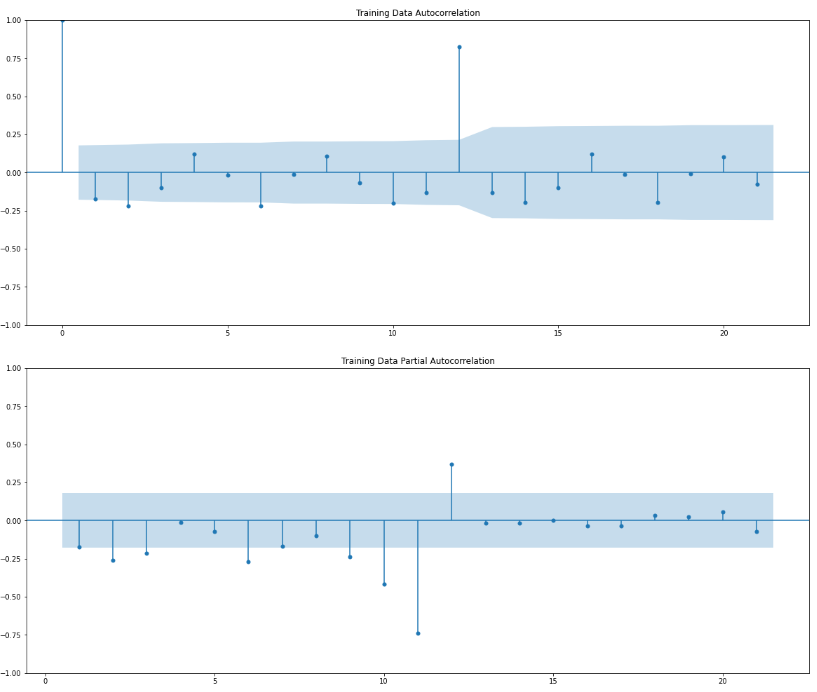
**Figure 1.6.B** Automated SARIMA Model

* There is a discrepancy in the model, here Q value picked up to be greater than q that is not expected.
* As per our expectation, the model didn’t perform very well and a very large RMSE value of 694733001166717.2 was observed.
* There was a requirement to pick up the values manually and tune the model which has been done in section 1.7.

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.

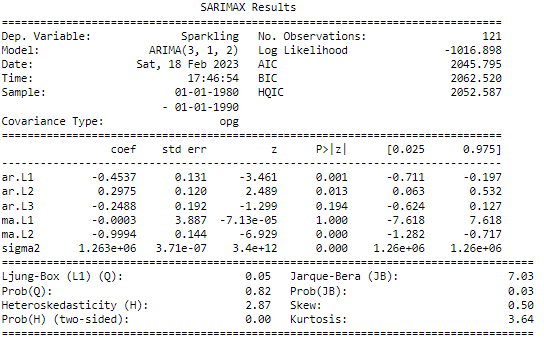
**Manual ARIMA Model**

* Value of p and q was selected to be 3 and 2 respectively looking at the PACF and ACF plots as shown below.
* Value of d was picked to be 1 as after one level of differencing the training data became stationary.

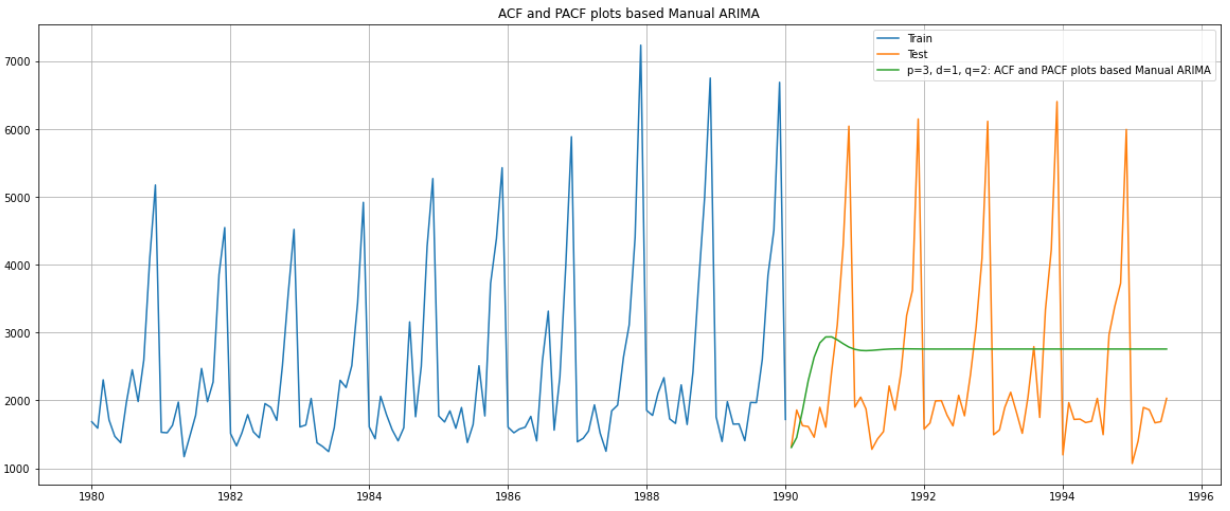


**Figure 1.7.A** ACF and PACF Plots

* ARIMA(3, 1, 2) Model was picked and built and is shared as follows.



**Table 1.7.1** ARIMA(3, 1, 2) Model

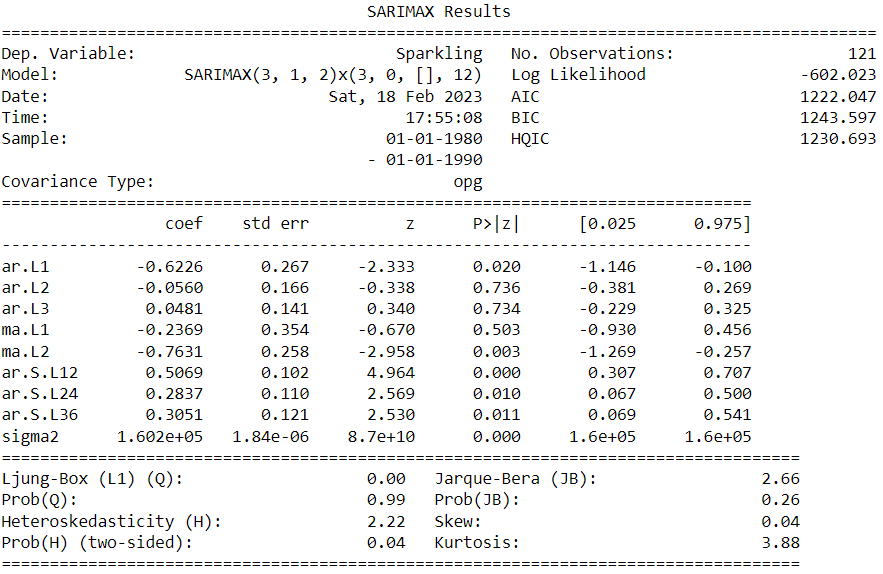


**Figure 1.7.B** Manual ARIMA Model

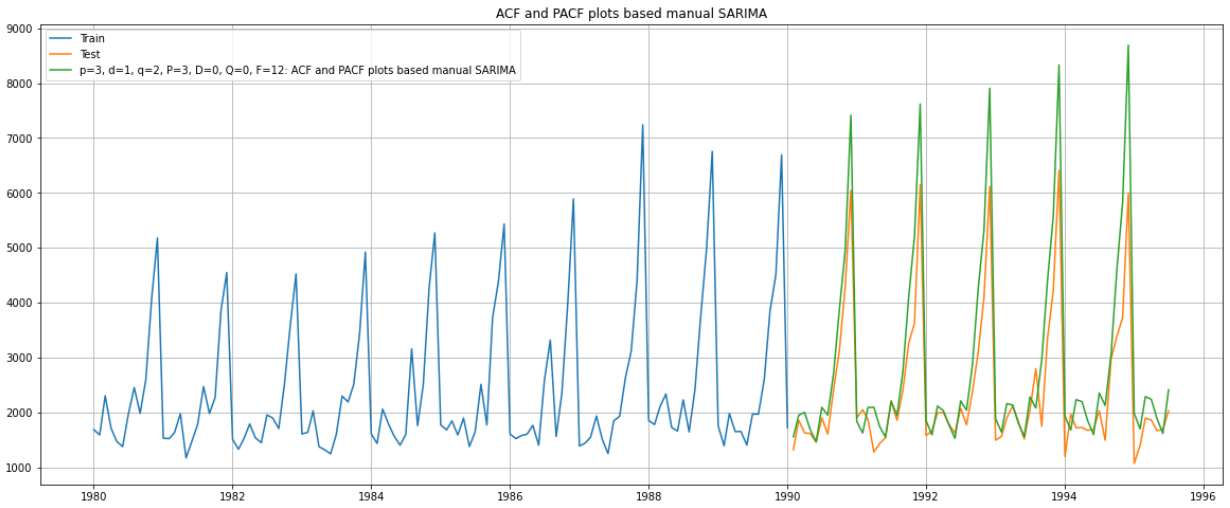
* RMSE value was found to be 1275.
* The model performed better than automated ARIMA model.
* The model was unable to explain the trend and the seasonality.

**Manual SARIMA Model**

* Value of p and q was selected to be 3 and 2 respectively looking at the PACF and ACF plots as shown below.
* Value of d was picked to be 1 as after one level of differencing the training data became stationary.
* There was no indication from the ACF and PACF plot on P and Q values and since picking both to be 0 would imply only an ARIMA model only, values of P and Q were picked iteratively with conditions 0 < P <= p and 0 <= Q <= q and the permutation which yielded the least RMSE value was picked i.e P=3 and Q=0.
* D was picked to be 0.
* SARIMA(3, 1, 2)(3, 0, 1, 12) Model was picked and built and is shared as follows.



**Table 1.7.2** SARIMA(3, 1, 2)(3, 0, 1, 12) Model

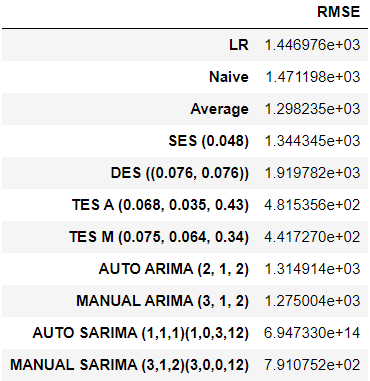


**Figure 1.7.C** Manual SARIMA Model

* Manual SARIMA model perform well in explaining the trend and seasonality.
* RMSE value of 791.07 was observed.

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.

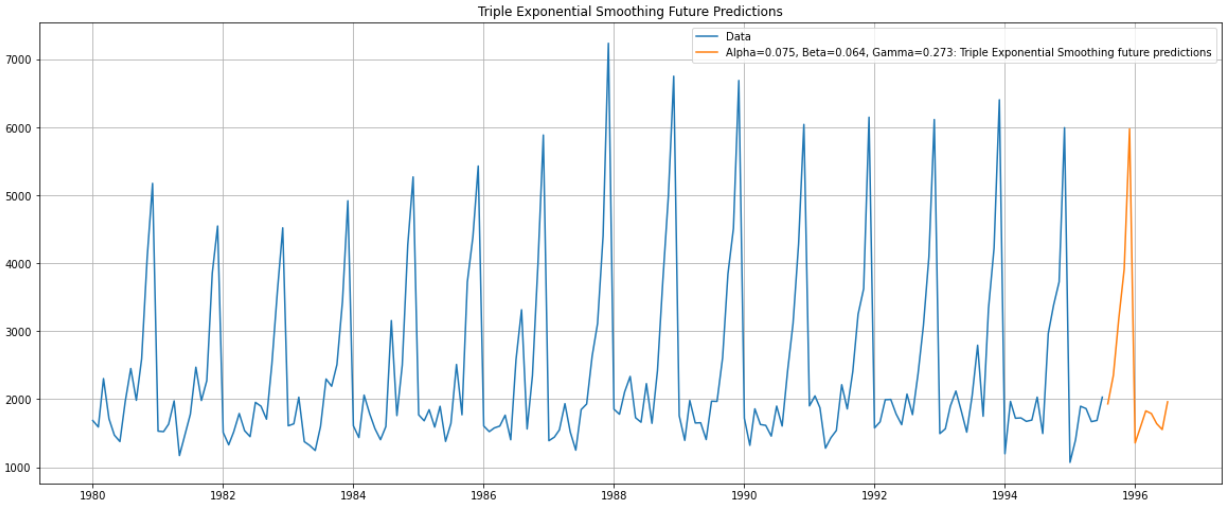
* Looking at the RMSE scores of various models, it was observed that TES-M model performed the best and was successfully able to explain both the trend and seasonality component of the time series.



**Table 1.8.1** Test RMSE values for various models.

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.

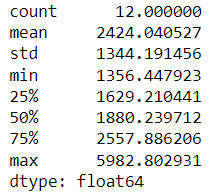
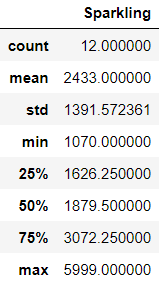
* TES-M was picked and built for the entire data available.
* Alpha, Beta and Gamma values were observed to be 0.075, 0.064 and 0.273 respectively.
* 12 months future data was predicted.



**Figure 1.9.A** Predicted 12-month data.

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.

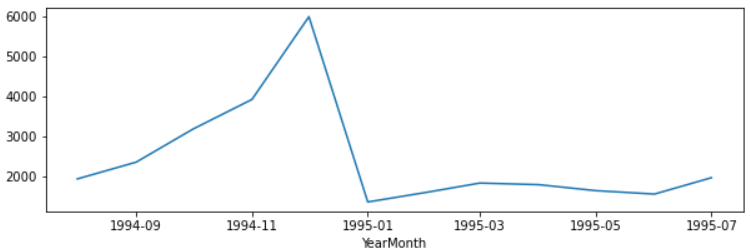
* Comparison between the last available 12-month data and the new predicted 12 months data is done and the results are shown below.



**Figure 1.10.A** Last Available Vs Predicted 12-month data description.



**Figure 1.10.B** Predicted 12-month data trend



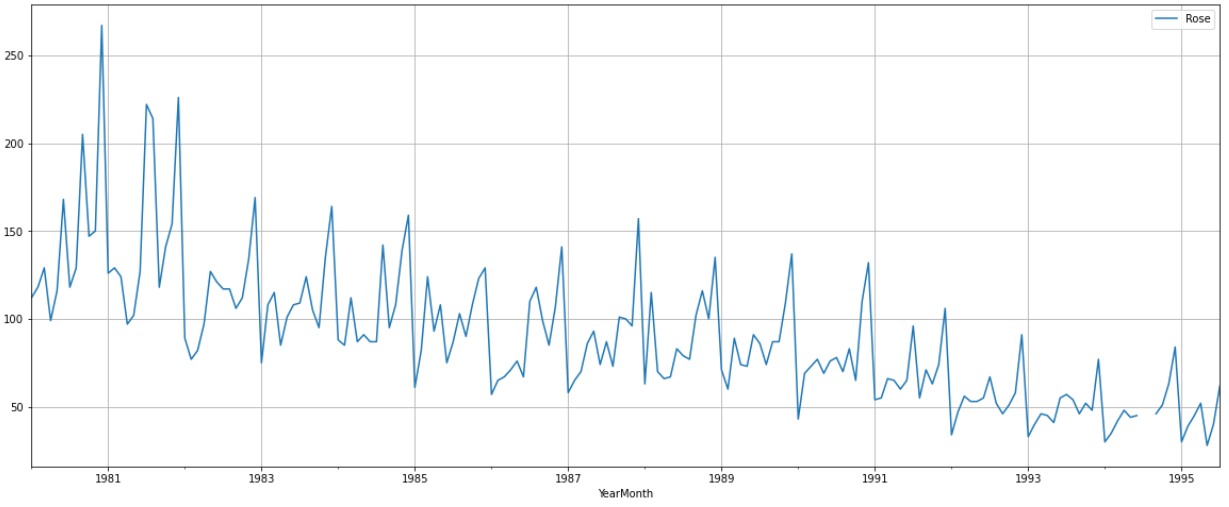
**Figure 1.10.C** Last available actual 12-month data

* Predicted 12-month data resembles the last available 12-month data.
* The model is beautifully able to capture the seasonality and trend.
* There is a slight dip forecasted in the average Y-o-Y sales of sparkling wine bottles.
* It also highlights that the first 6 month of the year, people do not seem to buy much sparkling wine.
* The sales of sparkling wine continue to increase significantly from the 7th month and hit the peak in last month of the year and then there is a significant downfall.

**2. Rose**

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

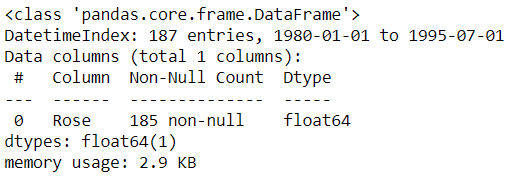
* The data was read using read\_csv function from pandas library.
* The YearMonth column was parsed to be of Date format and was set as index column.
* The data was plotted and can be visualized in the following plot.



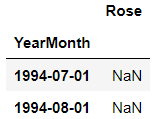
**Figure 2.1.A** Sales Data for Rose Wine

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

* There were 187 unique entries in the dataset. 2 null values were found.

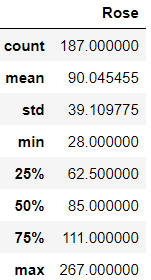


**Figure 2.2.A** Dataframe Information



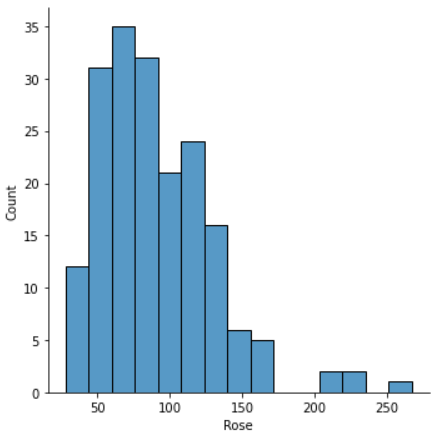
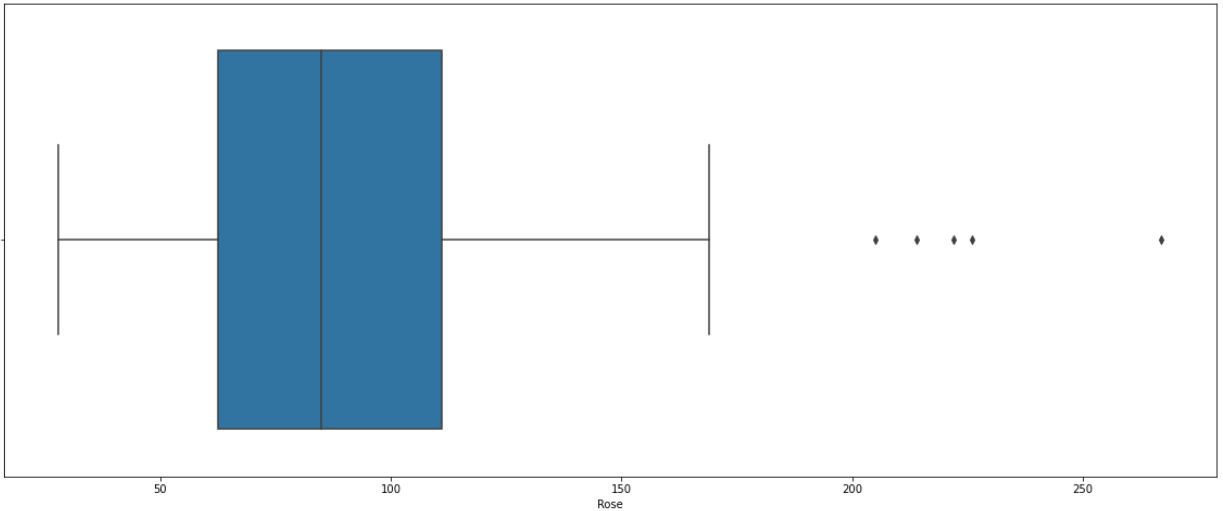
**Figure 2.2.B** Null values.

* Null Values were imputed using suitable means.
* The monthly sales were averaged at 90.
* Maximum sales recorded were 267.
* Minimum sales recorded were 28.



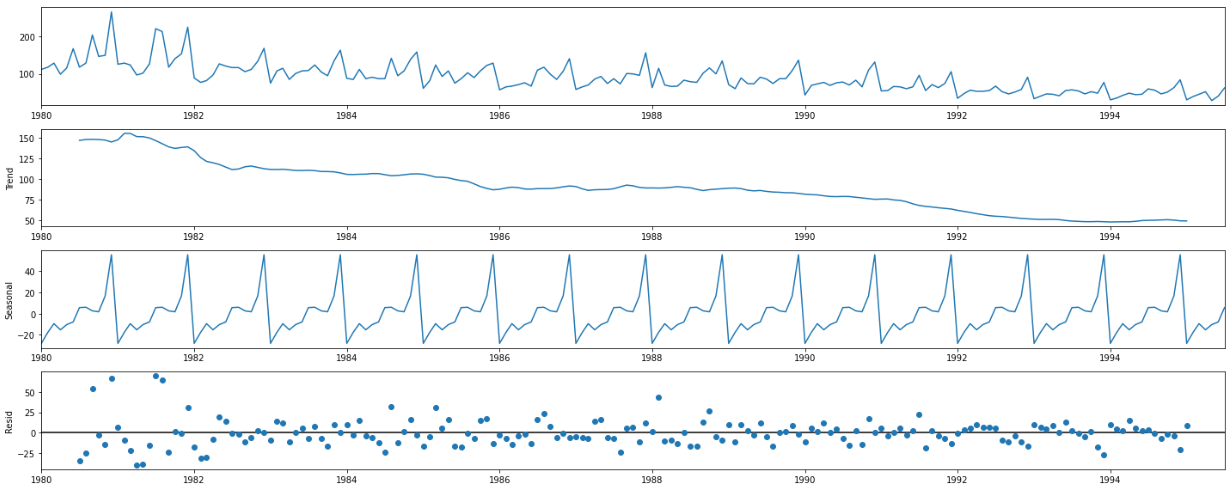
**Figure 2.2.C** Dataframe Statistical Description

* Most of the months, the sales are in between 40-140. Followed by fewer months with higher sales from 140-170 and not many months with exceptional sales >200.

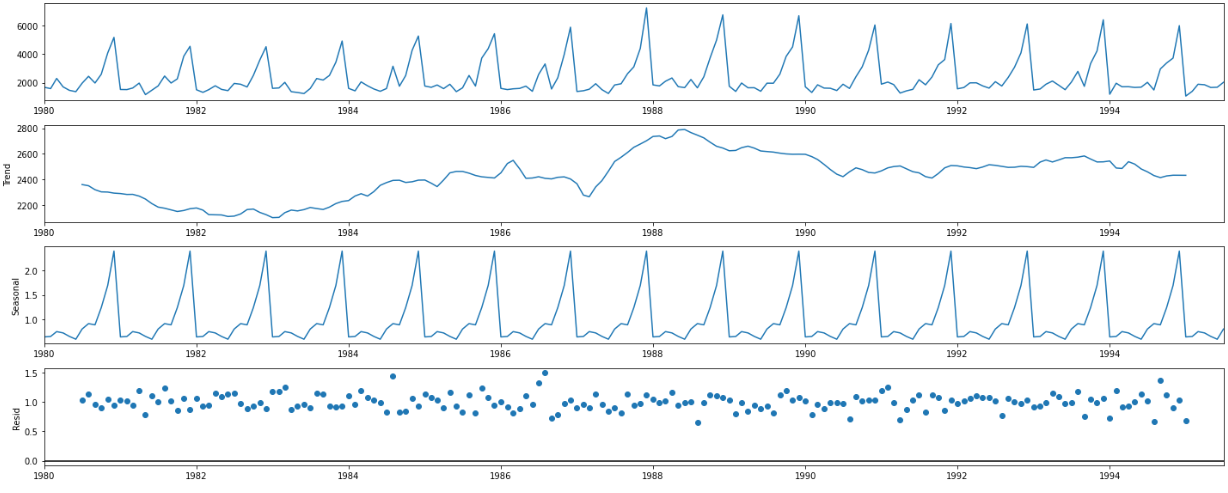
**Figure 2.2.D** Sales Distribution

* Additive decomposition was performed for the time series and the results are shared below.



**Figure 2.2.E** Additive Decomposition

* Both the trend and seasonality components were observed in the series.
* The residuals seemed to have a pattern suggesting additive decomposition was unable to completely decompose the series. Hence, we moved to multiplicative decomposition.
* Multiplicative decomposition was performed for the time series and the results are shared below.



**Figure 2.2.F** Multiplicative Decomposition

* Multiplicative decomposition was also unable to perform much better as similar trends are observed in residuals as were observed for the case for additive decomposition.

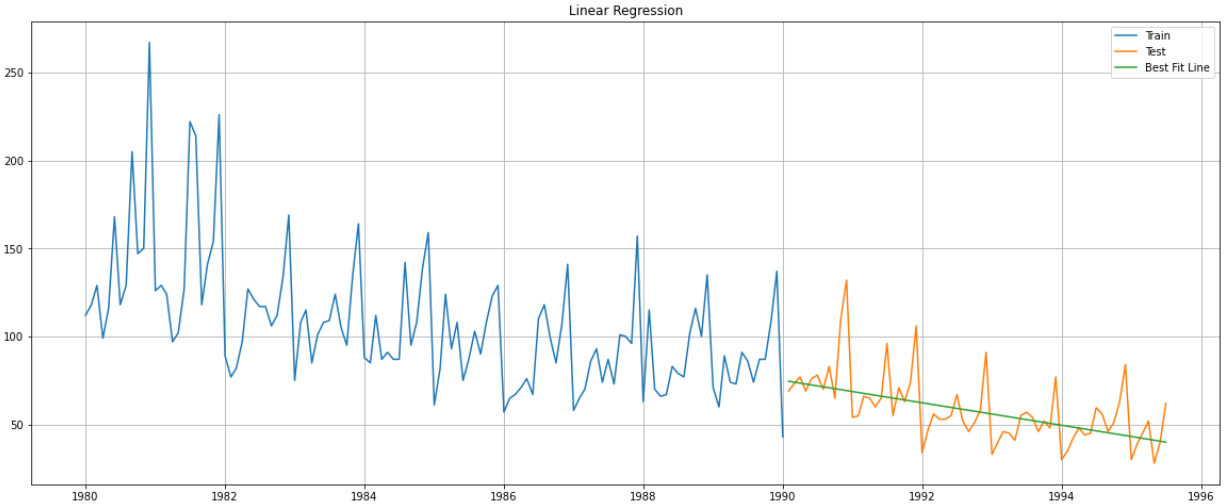
2.3. Split the data into training and test. The test data should start in 1991.

* The data frame was split into training and testing data using python slicing feature.
* The data until 1990 was used as a training data and the data from 1991 and beyond was used as testing 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.

**Regression Model**

* Linear Regression model was built and validated with the test data.

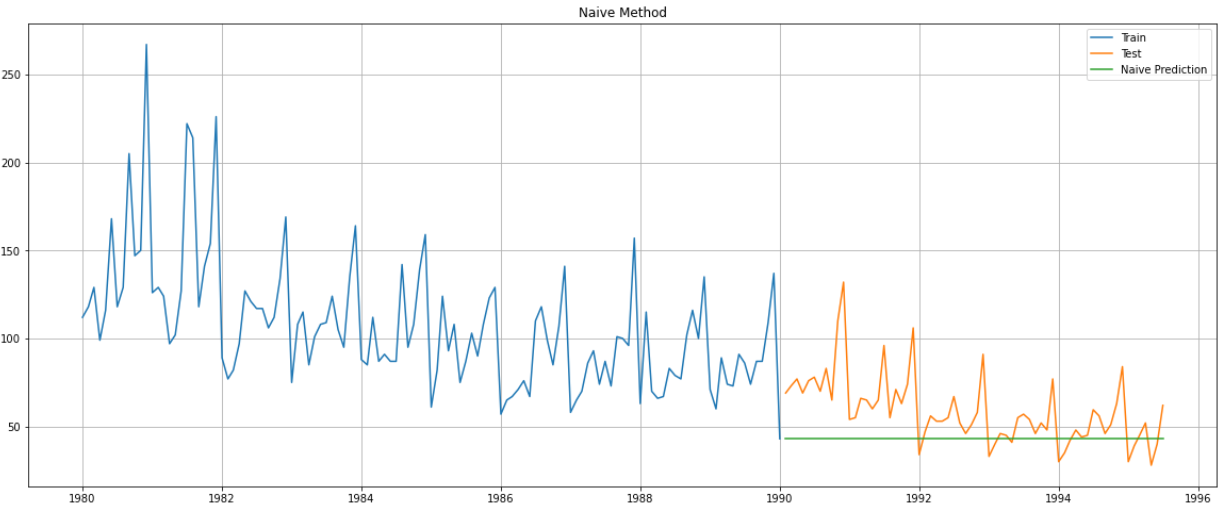


**Figure 2.4.A** Linear Regression Model

* As it can be clearly observed, the Linear regression model was able to explain the nature of the time series to some extent
* It gave an idea about the trend but fails to explain the seasonality of the series.
* RMSE vale of 16.4 was observed.

**Naïve Forecast Model**

* Naïve forecast model was built and validated across the test data.

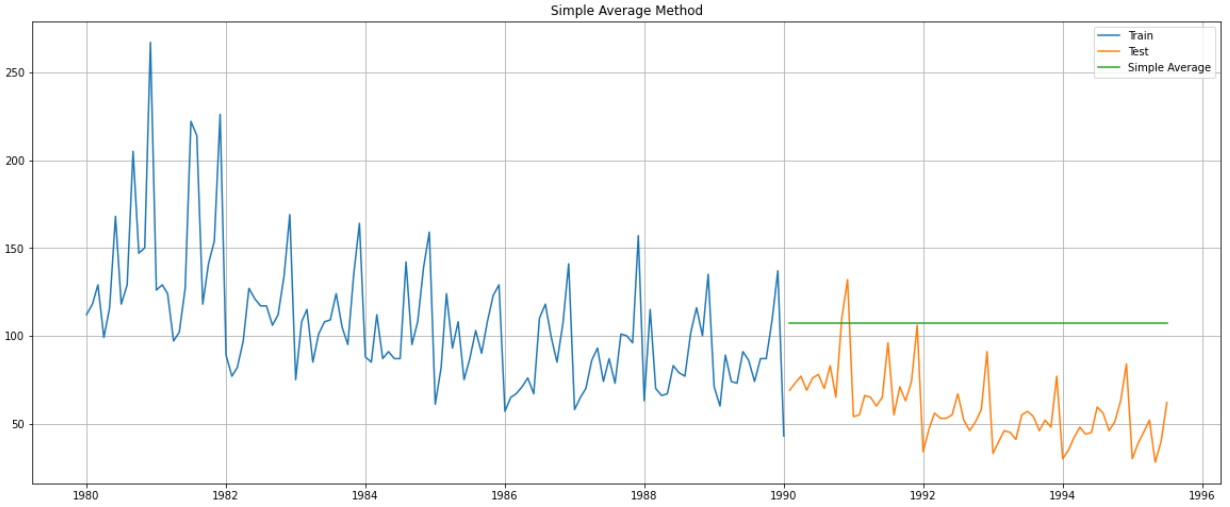


**Figure 2.4.B** Naïve Forecast Model

* As it can be clearly observed, the Naïve forecast model failed to explain the nature of test series.
* It fails to explain both the trend and the seasonality component of the time series.
* RMSE vale of 25.08 was observed which was even higher than the regression model.

**Simple Average Model**

* Simple Average model was built and validated with the test data.

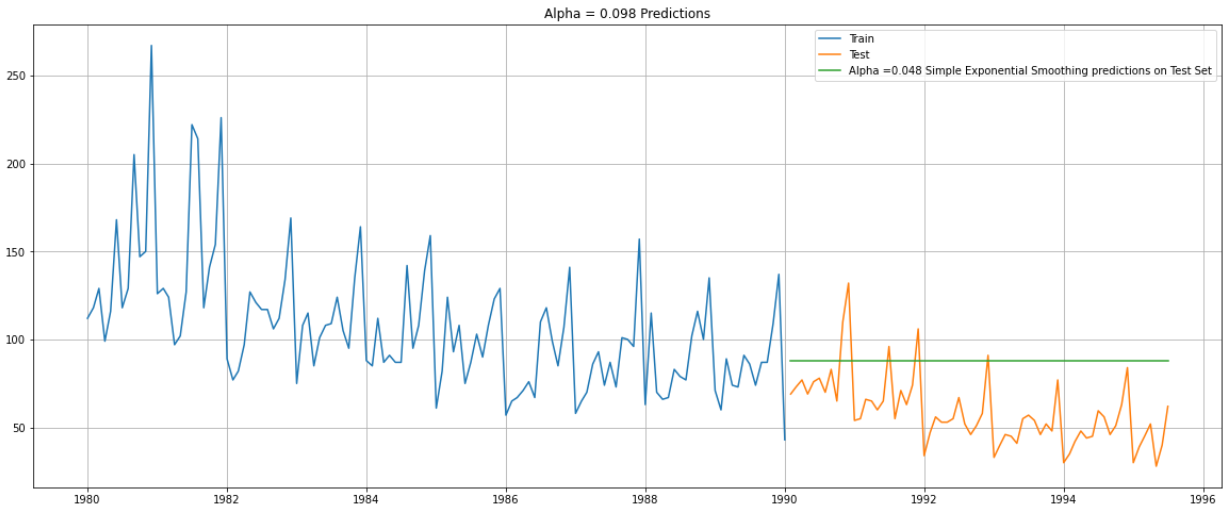


**Figure 2.4.C** Simple Average Model

* As it can be clearly observed, the Simple Average Model failed to explain the nature of test series.
* It fails to explain both the trend and seasonality component of the time series.
* RMSE vale of 51.86 was observed.
* It performed worse than both regression and naïve forecasting models for this dataset.

### **Simple Exponential Smoothing Model with Additive Errors**

* SES-A model was built and validated with test data.

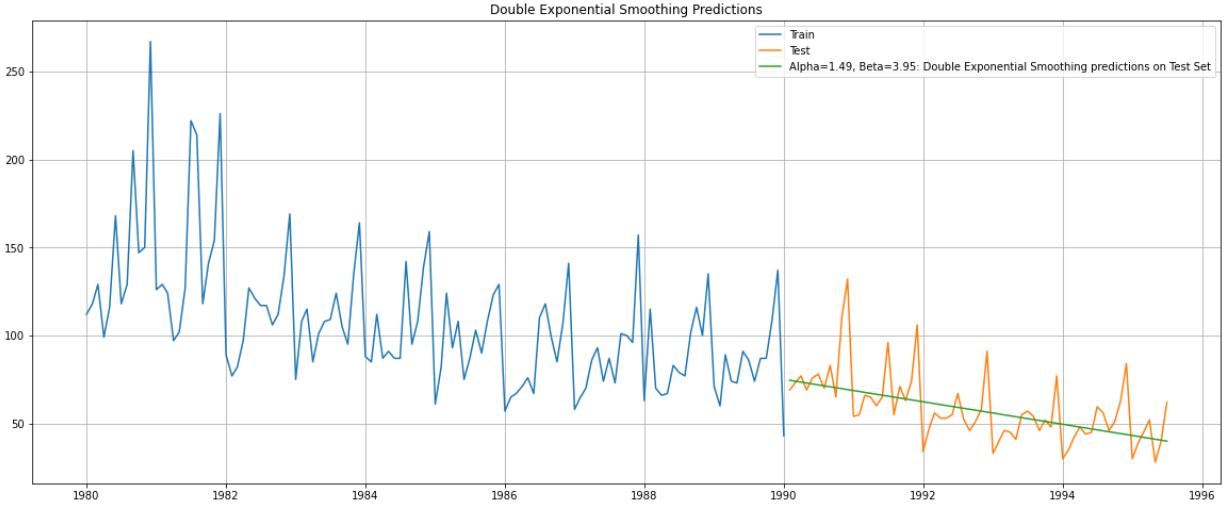


**Figure 2.4.D** SES-A model

* Alpha value was found to be 0.098.
* As it can be clearly observed, the SES model failed to explain the nature of test series.
* It fails to explain both the trend and the seasonality component of the time series.
* RMSE vale of 35.02 was observed. It performed better than simply average model.

### **Double Exponential Smoothing Model with Additive Errors**

* DES-A model was built and validated with test data.

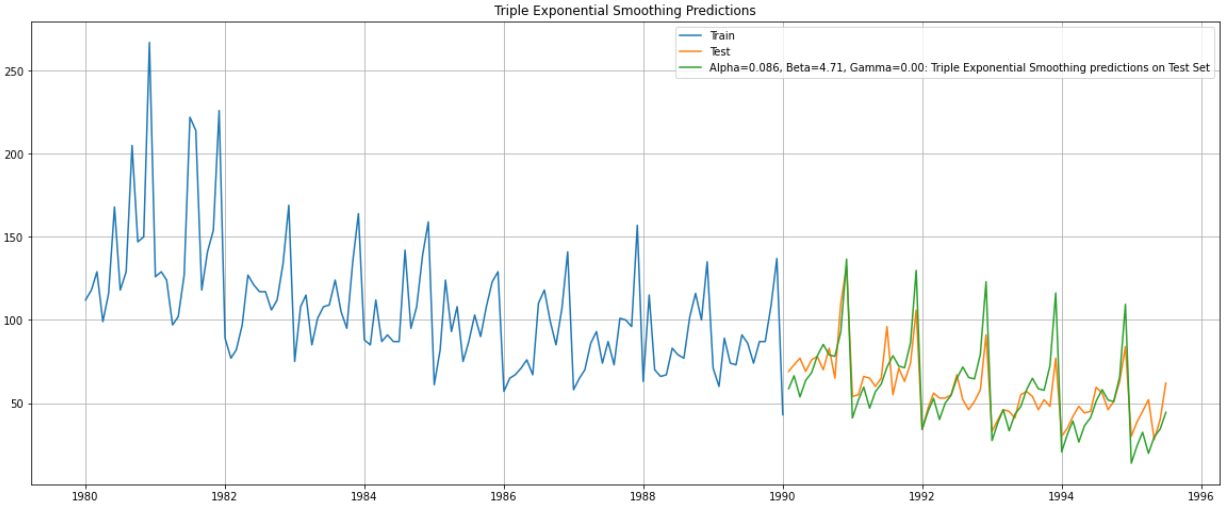


**Figure 2.4.E** DES-A model

* Alpha and Beta vales were observed to be 1.49 and 3.95 respectively.
* The DES model explains nature of test series to some extent.
* Even though it might give an idea about the trend but fails to explain the seasonality of the series.
* RMSE vale of 16.41 was observed. It performed better than SES model.

### **Triple Exponential Smoothing Model with Additive Errors**

* TES-A model was built and validated with test data.

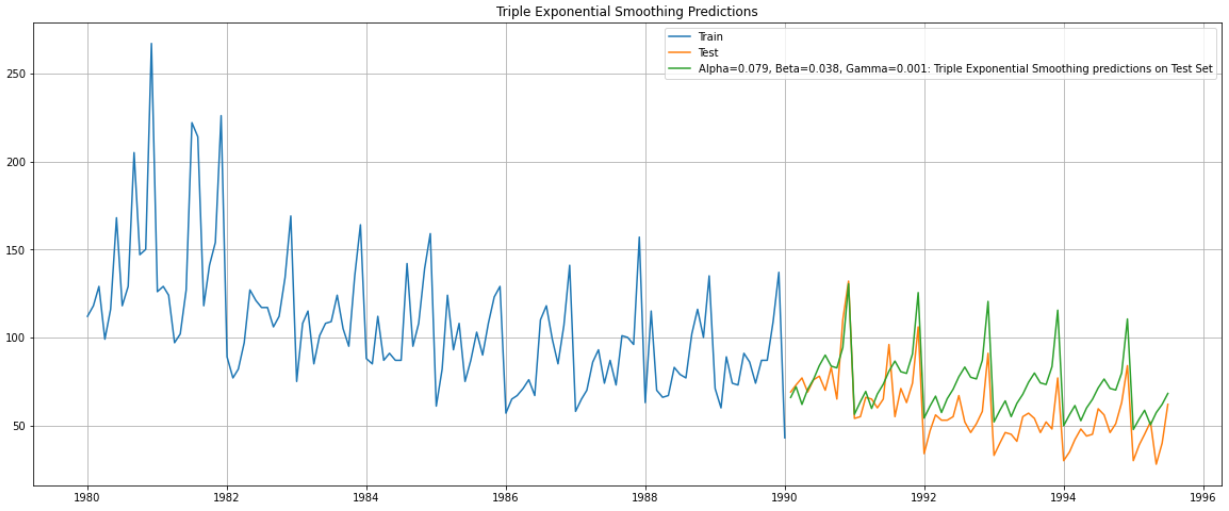


**Figure 2.4.F** TES-A model

* Alpha, Beta and Gama values were found to be 0.086, 4.71 and 0.0005.
* TES-A model explains both the trend and seasonality of the time series.
* RMSE vale of 13.88 was observed. It performed much better than all other models tried before.

### **Triple Exponential Smoothing Model with Multiplicative Errors**

* TES-M model was built and validated with test data.



**Figure 2.4.G** TES-M model

* Alpha, Beta and Gama values were found to be 0.079, 0.038 and 0.001.
* TES-M model explains both the trend and seasonality of the time series.
* RMSE vale of 18.48 was observed. It performed worse than TES-A model.

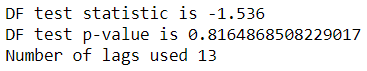
2.5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05.

The Augmented Dickey-Fuller test is a 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:

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

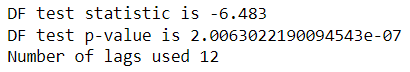
𝐻a : The Time Series does not have a unit root and is thus stationary.

* ADF test was performed on the training data and following observations were made.



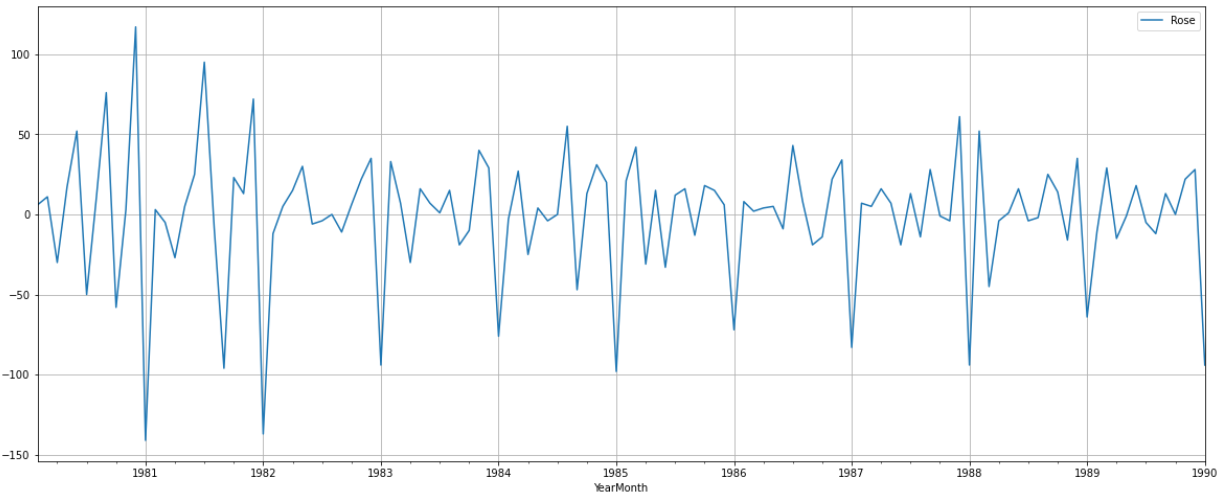
**Figure 2.5.A** ADF Test Results for training data.

* P-value is > 0.05, thus, null hypothesis cannot be rejected, implying the training data is not stationary.
* Since, the training data was found to be non-stationary, differencing was applied to the training data and differenced training data was again tested using ADF.



**Figure 2.5.B** ADF Test Results for differenced training data.

* P-value is much less than 0.05, thus, null hypothesis can be rejected, implying the differenced training data is stationary.

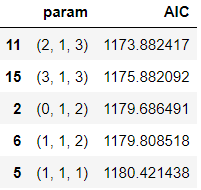


**Figure 2.5.C** Differenced Training Data.

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.

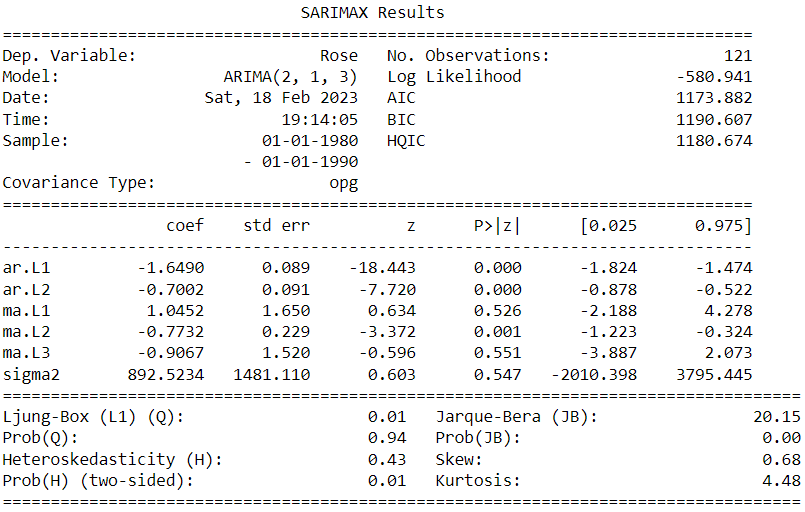
**Automated ARIMA Model**

* Multiple values of p, d and q were used and ARIMA models were built for various permutations and combinations.
* AIC values were calculated for each model.
* Minimum AIC value of 1173.88 was observed for a value of 2, 1 and 3 for p, d and q respectively.

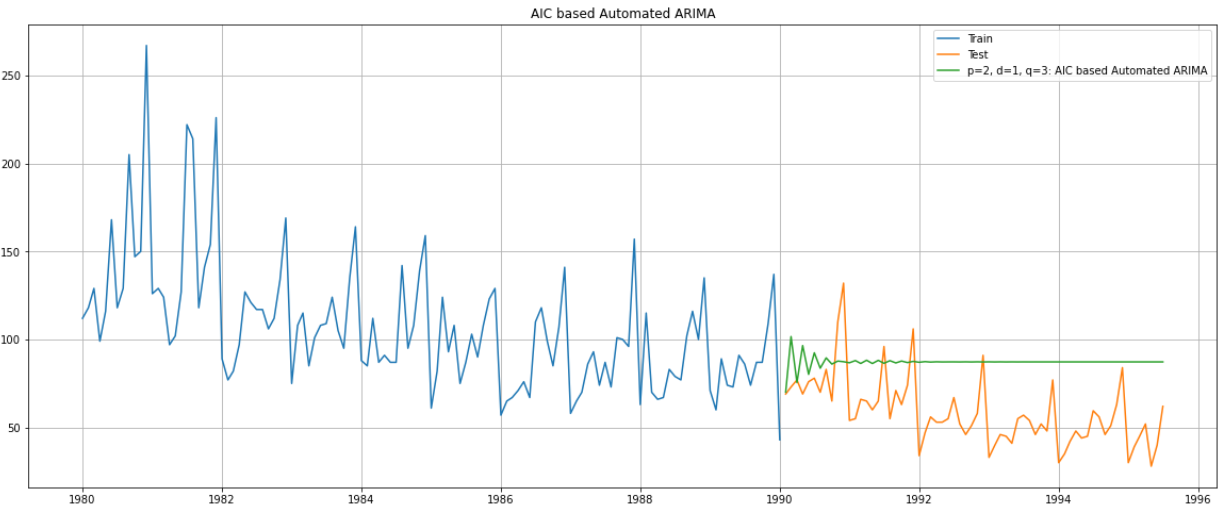


**Table 2.6.1** AIC Values for Various ARIMA Models.

* ARIMA(2, 1, 3) Model was picked and built and is shared as follows.



**Table 2.6.2** ARIMA(2, 1, 3) Model

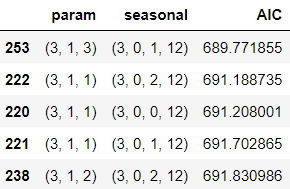


**Figure 2.6.A** Automated ARIMA Model

* RMSE value was found to be 34.43.
* The model was unable to explain the trend and the seasonality component of the series.

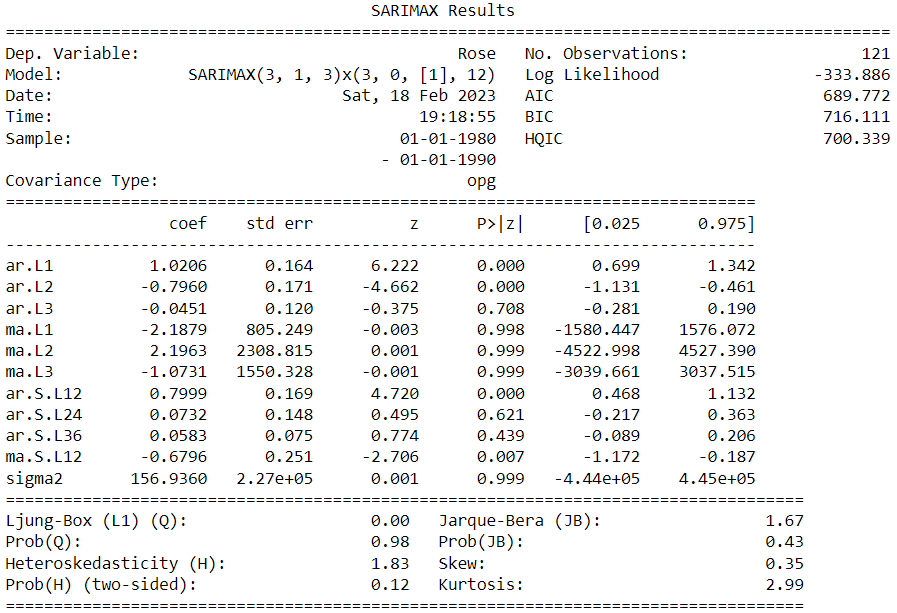
**Automated SARIMA Model**

* Multiple values of p, d, q, P, D, Q and F=12 were used and SARIMA models were built for various permutations and combinations.
* AIC values were calculated for each model.
* Minimum AIC value of 689.77was observed for a value of 3, 1, 3, 3, 0 , 1 and 12 for p, d, q, P, D, Q and F respectively.

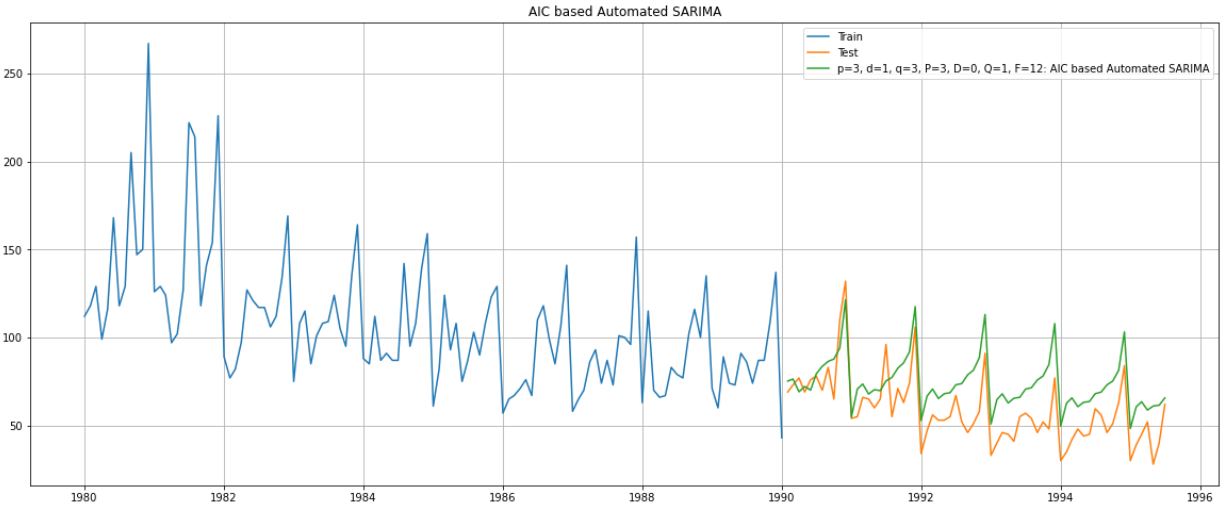


**Table 2.6.3** AIC Values for Various SARIMA Models.

* SARIMA(3, 1, 3)(3, 0, 1, 12) Model was picked and built and is shared as follows.



**Table 2.6.4** SARIMA(3, 1, 3)(3, 0, 1, 12) Model



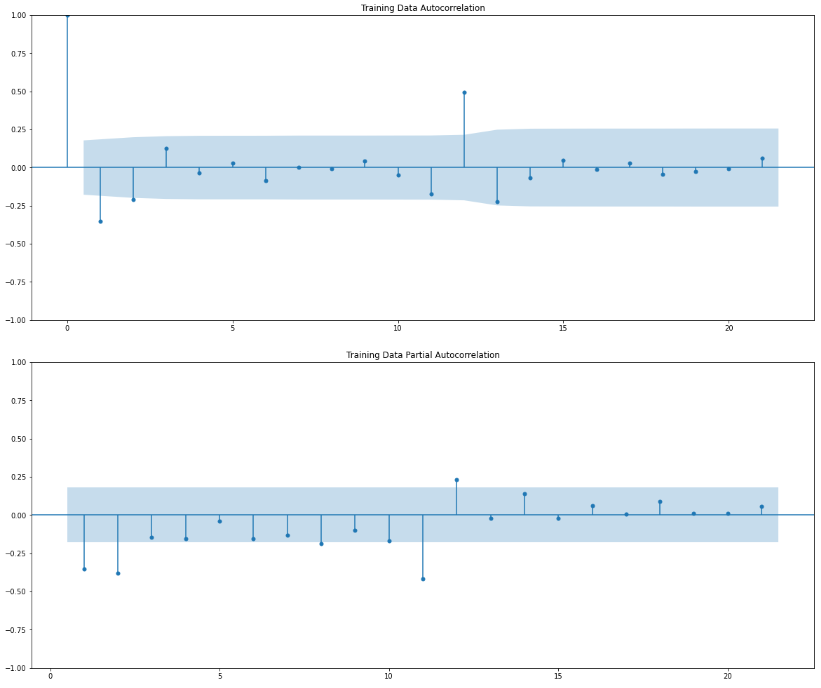
**Figure 2.6.B** Automated SARIMA Model

* The model was able to explain both the trend and seasonality component.
* RMSE value was observed to be 18.9. It performed better than AUTO ARIMA model.

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**

* Value of p and q was selected to be 2 and 2 respectively looking at the PACF and ACF plots as shown below.
* Value of d was picked to be 1 as after one level of differencing the training data became stationary.

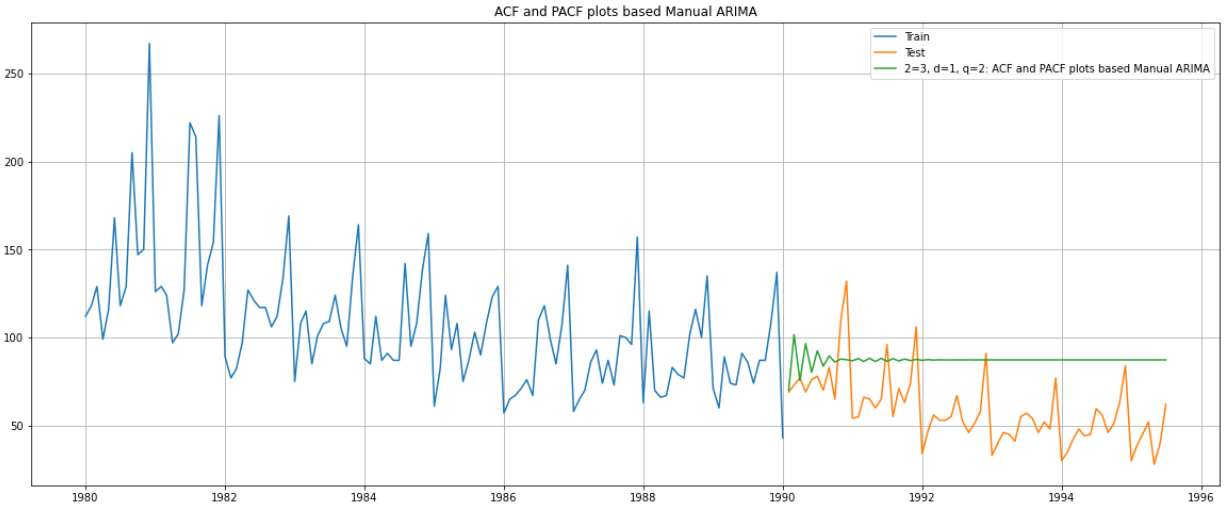


**Figure 2.7.A** ACF and PACF Plots

* ARIMA(2, 1, 2) Model was picked and built and is shared as follows.



**Table 2.7.1** ARIMA(2, 1, 2) Model

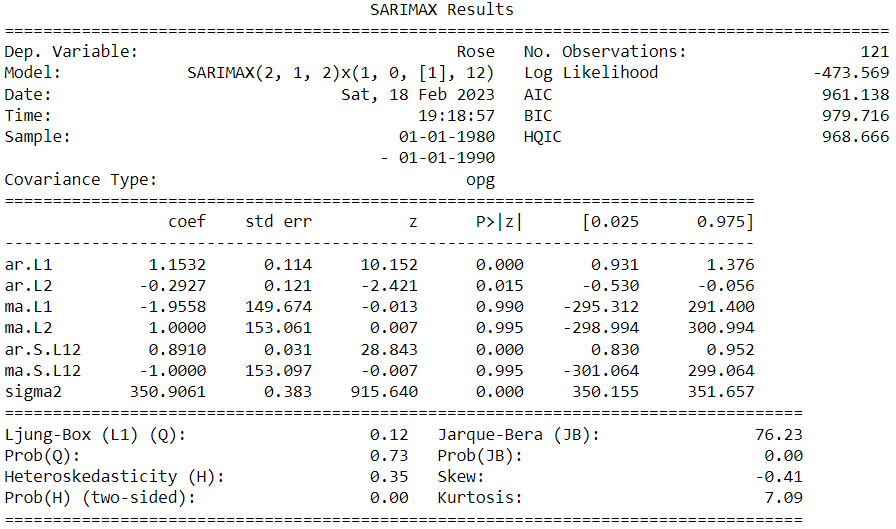


**Figure 2.7.B** Manual ARIMA Model

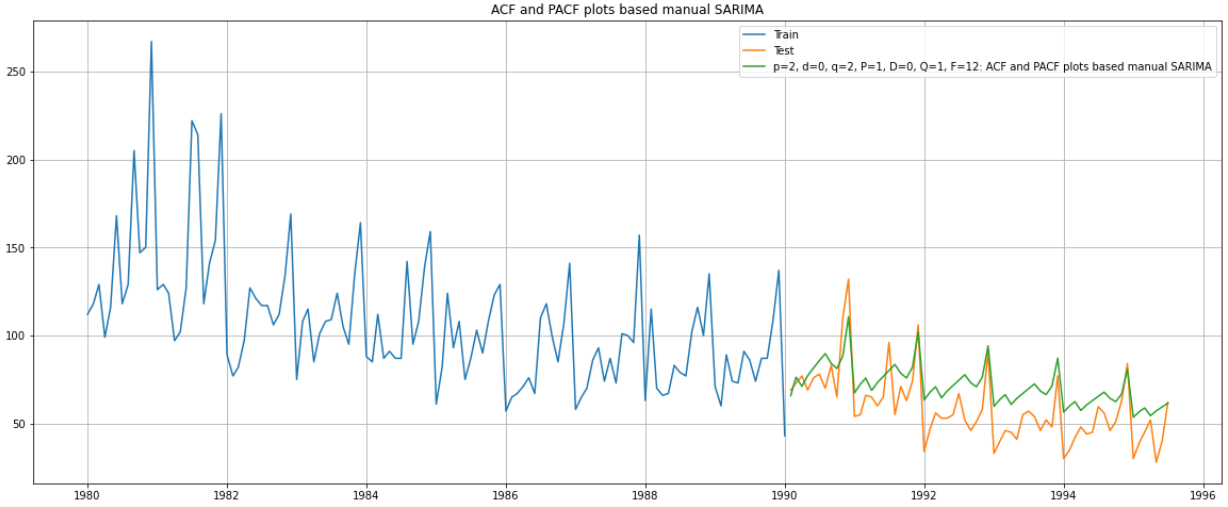
* RMSE value was found to be 34.38.
* The model performed better than automated ARIMA model.
* The model was unable to explain the trend and the seasonality.

**Manual SARIMA Model**

* Value of p and q was selected to be 2 and 2 respectively looking at the PACF and ACF plots as shown below.
* Value of d was picked to be 1 as after one level of differencing the training data became stationary.
* There was no indication from the ACF and PACF plot on P and Q values and since picking both to be 0 would imply only an ARIMA model only, values of P and Q were picked iteratively with conditions 0 < P <= p and 0 <= Q <= q and the permutation which yielded the least RMSE value was picked i.e P=1 and Q=1.
* D was picked to be 0.
* SARIMA(2, 1, 2)(1, 0, 1, 12) Model was picked and built and is shared as follows.



**Table 2.7.2** SARIMA(2, 1, 2)(1, 0, 1, 12) Model

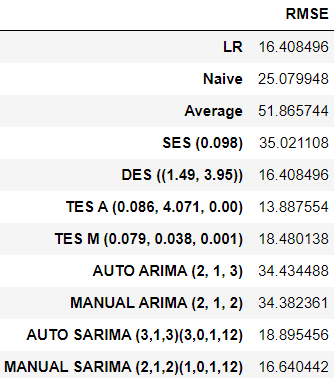


**Figure 2.7.C** Manual SARIMA Model

* Manual SARIMA model perform well in explaining the trend and seasonality.
* RMSE value of 16.64 was observed.

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.

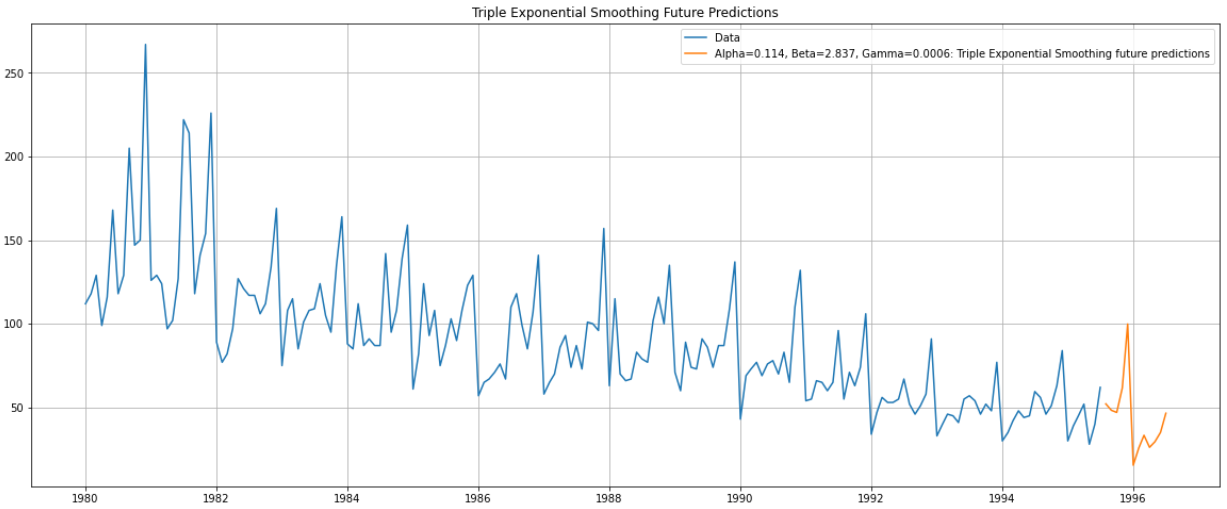
* Looking at the RMSE scores of various models, it was observed that TES-A model performed the best and was successfully able to explain both the trend and seasonality component of the time series.



**Table 2.8.1** Test RMSE values for various models.

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.

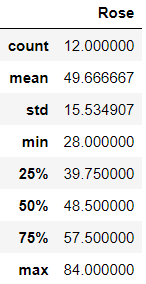
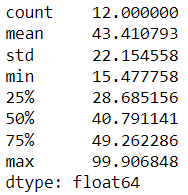
* TES-A was picked and built for the entire data available.
* Alpha, Beta and Gamma values were observed to be 0.114, 2.837 and 0.0006 respectively.
* 12 months future data was predicted.



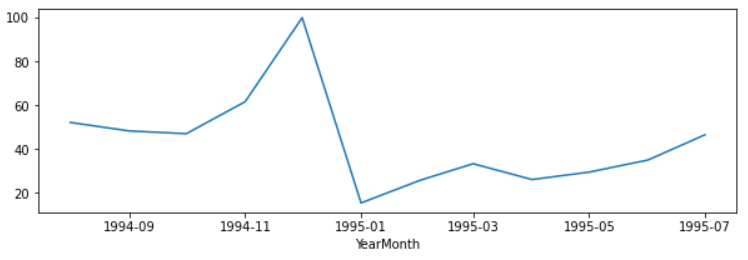
**Figure 2.9.A** Predicted 12-month data.

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.

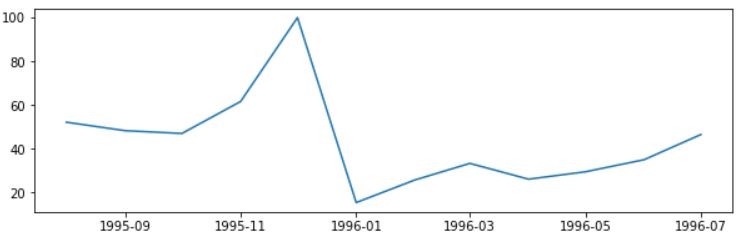
* Comparison between the last available 12-month data and the new predicted 12 months data is done and the results are shown below.

**Figure 2.10.A** Last Available Vs Predicted 12-month data description.



**Figure 2.10.B** Actual 12-month data trend



**Figure 2.10.C** Predicted 12-month data trend

* Predicted 12-month data resembles the last available 12-month data.
* The model is beautifully able to capture the seasonality and trend.
* There is a dip forecasted in the average Y-o-Y sales of rose wine bottles.
* During the peak season time, the sale is expected to be slightly better than last year but during the off season the sales are expected to be worse than last year.
* The sales as compared to sparkling wine are quite less. If the wine has proportionally higher profit margin than it is alright but otherwise it seems like the product is less liked by the people and might be discontinued.