Cijith, Jose (J.)



**TIME SERIES FORECASTING PROJECT**

Prepared by: Cijith Jose

Contents

[PROBLEM 1 – SPARKLING DATASET 3](#_Toc93861918)

[1. Read the data as an appropriate Time Series data and plot the data . 3](#_Toc93861919)

[2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition. 4](#_Toc93861920)

[3. Split the data into training and test. The test data should start in 1991. 9](#_Toc93861921)

[4. Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression,naïve forecast models, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE. 11](#_Toc93861922)

[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. 19](#_Toc93861923)

[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. 21](#_Toc93861924)

[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. 24](#_Toc93861925)

[8. Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data. 30](#_Toc93861926)

[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. 30](#_Toc93861927)

[10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales. 31](#_Toc93861928)

[PROBLEM 2 – ROSE DATASET 32](#_Toc93861929)

[1. Read the data as an appropriate Time Series data and plot the data. 32](#_Toc93861930)

[2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition. 33](#_Toc93861931)

[3. Split the data into training and test. The test data should start in 1991. 38](#_Toc93861932)

[4. Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression,naïve forecast models, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE. 39](#_Toc93861933)

[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. 48](#_Toc93861934)

[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. 50](#_Toc93861935)

[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. 53](#_Toc93861936)

[8. Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data. 58](#_Toc93861937)

[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. 59](#_Toc93861938)

[10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales. 60](#_Toc93861939)

[Table 1:Dataset of Sparkling Wine 4](#_Toc93862005)

[Table 2:Dataset after converting into Timeseries 4](#_Toc93862006)

[Table 3:Data Description 5](#_Toc93862007)

[Table 4:Pivot Table of Monthly sales across years 7](#_Toc93862008)

[Table 5: Summarized performance of the model 12](#_Toc93862009)

[Table 6: Summarized performance of the model 13](#_Toc93862010)

[Table 7:Summarized performance of the model 14](#_Toc93862011)

[Table 8:Summarised performance of the models 15](#_Toc93862012)

[Table 9:Summarised performance of the models 16](#_Toc93862013)

[Table 10:Summarised performance of the models 17](#_Toc93862014)

[Table 11:Summaruized performances of all the models 19](#_Toc93862015)

[Table 12:AIC values of first 5 Params of SARIMA model in descending order 23](#_Toc93862016)

[Table 13:Summarized performance of all the models 25](#_Toc93862017)

[Table 14:Predicted sales for next 12 months 32](#_Toc93862018)

[Table 15:Dataset of Rose 33](#_Toc93862019)

[Table 16:Dataset after converting into time series 33](#_Toc93862020)

[Table 17:Data Description of Rose Wine Dataset 34](#_Toc93862021)

[Table 18:Pivot Table of Monthly Sales across years 35](#_Toc93862022)

[Table 19:Summarised performance of models 41](#_Toc93862023)

[Table 20:Summarised performance of models 41](#_Toc93862024)

[Table 21:Summarized performance of models 42](#_Toc93862025)

[Table 22:Summarized performance of the model 43](#_Toc93862026)

[Table 23:Summarised performance of the models 44](#_Toc93862027)

[Table 24:Summarised performance of the models 45](#_Toc93862028)

[Table 25:AIC Values of the first 5 params of SARIMA model in descending order 53](#_Toc93862029)

[Table 26:Summarized performance of all the models 54](#_Toc93862030)

[Table 27:Predicted Sale of Rose 60](#_Toc93862031)

[Figure 1: of Sparkling Time series plot 8](#_Toc93862035)

[Figure 2:Yearly Box Plot 9](#_Toc93862036)

[Figure 3:Monthly Box Plot 9](#_Toc93862037)

[Figure 4:Monthly Sales across year Plot 10](#_Toc93862038)

[Figure 5:Time series Month Plot 11](file:///C:\Users\jcijith\Documents\Time%20Series%20Forecasting%20Project%20-CIJITH%20JOSE.docx#_Toc93862039)

[Figure 6:Emprical Cumulative Distribution Plot 11](#_Toc93862040)

[Figure 7:average retails sales per month and month on month percentage change of Sparkling sales. 12](#_Toc93862041)

[Figure 8:Additive model of the decomposition 12](#_Toc93862042)

[Figure 9:Multiplicative model decomposition 13](#_Toc93862043)

[Figure 10: Plot for Training and Test Dataframes 14](#_Toc93862044)

[Figure 11:Linear Regression Outcome of Sparkling wine Time series 15](#_Toc93862045)

[Figure 12:Naive Forecast outcome of Sparkling wine Time series 16](#_Toc93862046)

[Figure 13:Simple Average Forecast outcome of Sparkling wine Time series 16](#_Toc93862047)

[Figure 14:Moving Average Model outcome on the sparkling wine time series 17](#_Toc93862048)

[Figure 15:SES parameters for Sparkling wine dataset 18](#_Toc93862049)

[Figure 16:Simple Exponential Smoothing Outcome on the Sparkling Wine Time Series 18](#_Toc93862050)

[Figure 17:DES parameters for sparkling wine dataset 19](#_Toc93862051)

[Figure 18:Double Exponential Smoothing Outcome on the Sparkling Wine Time Series 19](#_Toc93862052)

[Figure 19:TES parameters for sparkling wine dataset 20](#_Toc93862053)

[Figure 20:TES Outcome on Sparkling wine Timeseries 20](#_Toc93862054)

[Figure 21:TES Parameters for Sparkling wine dataset 21](#_Toc93862055)

[Figure 22:TES outcome of sparkling wine timeseries 21](#_Toc93862056)

[Figure 23:TES parameters for sparkling wine dataset 22](#_Toc93862057)

[Figure 24:TES outcome on sparkling wine timeseries 22](#_Toc93862058)

[Figure 25:Summary of all prediction on a single graph 23](#_Toc93862059)

[Figure 26:Stationarity graph 24](#_Toc93862060)

[Figure 27:Stationarity graph after differencing 24](#_Toc93862061)

[Figure 28:Statistical analysis of ARIMA model 25](#_Toc93862062)

[Figure 29:ARIMA Plot 26](#_Toc93862063)

[Figure 30:Auto ARIMA Plot on the test data and forecasted values 26](#_Toc93862064)

[Figure 31:Statistical Analysis of SARIMA Model 27](#_Toc93862065)

[Figure 32:SARIMA Plot 27](#_Toc93862066)

[Figure 33:Auto SARIMA Plot 28](#_Toc93862067)

[Figure 34:ACF Plot of ARIMA Model 29](#_Toc93862068)

[Figure 35:PACF Plot of ARIMA Model 29](#_Toc93862069)

[Figure 36:Statistical Analysis of Manual ARIMA model 30](#_Toc93862070)

[Figure 37:Manual ARIMA Plot 30](#_Toc93862071)

[Figure 38:Manual ARIMA plot prediction on training and test data 31](#_Toc93862072)

[Figure 39:ACF Plot of SARIMA Model 31](#_Toc93862073)

[Figure 40:PACF Model of SARIMA Model 32](#_Toc93862074)

[Figure 41:Statistical Analysis of the Manual SARIMA Model 33](#_Toc93862075)

[Figure 42:Manual SARIMA Plot 33](#_Toc93862076)

[Figure 43:Plot of Trianing , Test and forecasted values of Manual SARIMA Model 33](#_Toc93862077)

[Figure 44:Summarized Performance of all the models 34](#_Toc93862078)

[Figure 45:Model performnace using full data 35](#_Toc93862079)

[Figure 46:Graph of Predicted sales of Sparkling Wine 35](#_Toc93862080)

[Figure 47:Rose Timeseries Plot 36](#_Toc93862081)

[Figure 48:Yearly Box Plot of Rose Wine dataset 37](#_Toc93862082)

[Figure 49:Monthly Box plot of Rose Wine Data 38](#_Toc93862083)

[Figure 50:Monthly Sales across year Plot 39](#_Toc93862084)

[Figure 51:Time Series Month Plot 39](#_Toc93862085)

[Figure 52:Empirical Cumulative Distribution Plot 40](#_Toc93862086)

[Figure 53:average retails sales per month and month on month percentage change of Rose sales. 40](#_Toc93862087)

[Figure 54: Additive Model of Decomposition 41](#_Toc93862088)

[Figure 55:Multiplicative Model of Decomposition 41](#_Toc93862089)

[Figure 56:Plot for training and test dataframes 43](#_Toc93862090)

[Figure 57:Linear Regression Outcome of Rose wine Time series 43](#_Toc93862091)

[Figure 58:Naive Forecast outcome of Rose WIne Time Series 44](#_Toc93862092)

[Figure 59:Simple Average Forecast Outcome of Rose wine Time series 45](#_Toc93862093)

[Figure 60:Moving Average Forecast outcome on Rose wine Time series 46](#_Toc93862094)

[Figure 61:SES parameters for Rose wine Forecast 46](#_Toc93862095)

[Figure 62:Simple Exponential smoothing outcome on Rose wine Timeseries 47](#_Toc93862096)

[Figure 63:DES parameters for Rose wine Forecast 47](#_Toc93862097)

[Figure 64:Double Exponential Smoothing Outcome on Rose Wine time series 48](#_Toc93862098)

[Figure 65:TES Parameters for Rose wine sales forecast 48](#_Toc93862099)

[Figure 66:TES Add\_Add outcome on Rose Wine sales Time series 49](#_Toc93862100)

[Figure 67:TES Parameters for Rose wine sales forecast 49](#_Toc93862101)

[Figure 68:TES Add\_Mul outcome on Rose wine sales time series 50](#_Toc93862102)

[Figure 69:TES parameters for Rose Wine sales forecast 50](#_Toc93862103)

[Figure 70:TES Mul\_Mul outcome of Rose wine sale time series 51](#_Toc93862104)

[Figure 71:Consolidated prediction Plot of Rose Wine Sales Data 52](#_Toc93862105)

[Figure 72:Stationarity Graph 53](#_Toc93862106)

[Figure 73:Stationarity Graph after differencing 53](#_Toc93862107)

[Figure 74:Statistical Analysis of ARIMA model 54](#_Toc93862108)

[Figure 75:Diagnosis of ARIMA Model 55](#_Toc93862109)

[Figure 76:Auto ARIMA Plot of Test and forecasted values 55](#_Toc93862110)

[Figure 77:Statistical Analysis of SARIMA Model 56](#_Toc93862111)

[Figure 78:Diagnosis of SARIMA Model 56](#_Toc93862112)

[Figure 79:Auto SARIMA plot 57](#_Toc93862113)

[Figure 80:ACF Plot of ARIMA Model 58](#_Toc93862114)

[Figure 81:PACF Plot of ARIMA Model 58](#_Toc93862115)

[Figure 82:Statistical Analysis of Manual ARIMA Model 59](#_Toc93862116)

[Figure 83:Diagnosis of Manual ARIMA model 59](#_Toc93862117)

[Figure 84:Manual ARIMA Predictions Plot 59](#_Toc93862118)

[Figure 85:ACF plot of SARIMA Model 60](#_Toc93862119)

[Figure 86:PACF Plot of SARIMA Model 60](#_Toc93862120)

[Figure 87:Statistical Analysis of Manual SARIMA Model 61](#_Toc93862121)

[Figure 88:Diagnosis of manual SARIMA plot 61](#_Toc93862122)

[Figure 89:Manual SARIMA predictions plot 62](#_Toc93862123)

[Figure 90:Summarized performance of all the models 62](#_Toc93862124)

[Figure 91:Model Performance using full data 63](#_Toc93862125)

# PROBLEM 1 – SPARKLING DATASET

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

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

I have imported the data series and observed that there are 2 columns , one which has an YearMonth value with it , which needs to converted to appropriate time series data for analysis and the other column contains the monthly sales data of Sparkling wine from 1980 to 1995 . The table below shows the data imported from the dataset .

|  |  |  |
| --- | --- | --- |
|  | YearMonth | Sparkling |
| 0 | 1980-01 | 1686 |
| 1 | 1980-02 | 1591 |
| 2 | 1980-03 | 2304 |
| 3 | 1980-04 | 1712 |
| 4 | 1980-05 | 1471 |

Table 1:Dataset of Sparkling Wine

Below is the table after converting the YearMonth column into the time series data and making our time series reference as the index, since we can conveniently do the slicing i.e obtain data for a specific time period, and check values for corresponding time point .

|  |  |
| --- | --- |
| **Time\_Stamp** | **Sparkling** |
| 1/31/1980 | 1686 |
| 2/29/1980 | 1591 |
| 3/31/1980 | 2304 |
| 4/30/1980 | 1712 |
| 5/31/1980 | 1471 |

Table 2:Dataset after converting into Timeseries

Lets go ahead and Plot the graph of Sparkling data set:

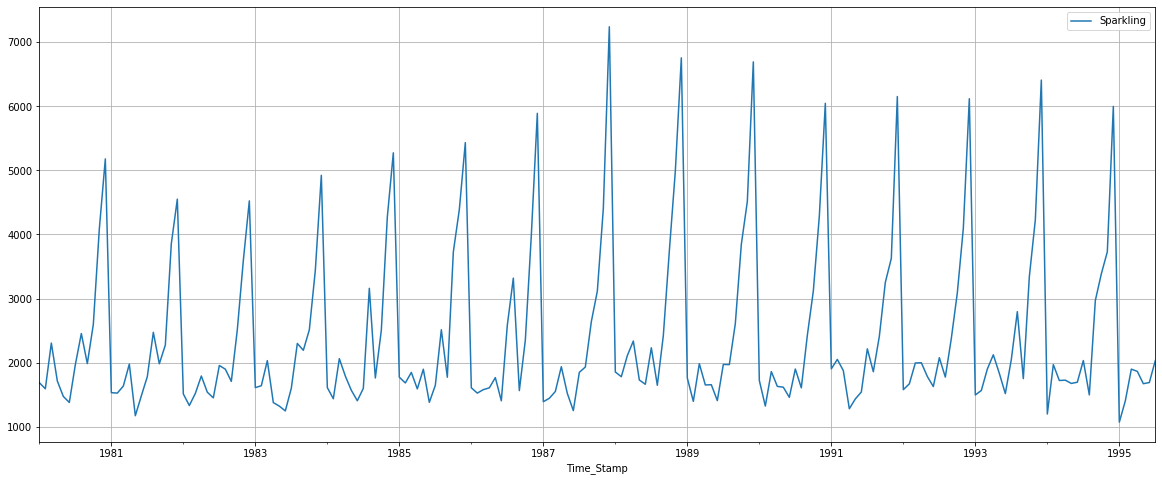


Figure 1: of Sparkling Time series plot

From the above graph we can observe that the sales of sparkling wines are showing a mixed trend of upwards and downwards for the last 15 years . We can also observe that there is an evident seasonality element that is present, where in it can be understood that the sales are more towards the end of the year rather than the early months. Further in our analysis we will explore more about trends and seasonality when we perform decomposition .

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

Now we go ahead with the Exploratory Data Analysis , we saw that there are 187 entries for sales column and 2 columns including time series column which is used as index for referencing . Further we saw that there are no null values, but 11 duplicate entries which can be ignored since each value correspond to a different time index.

Below is the data description :

|  |  |
| --- | --- |
|  | **Sparkling** |
| count | 187 |
| mean | 2402.417 |
| std | 1295.112 |
| min | 1070 |
| 25% | 1605 |
| 50% | 1874 |
| 75% | 2549 |
| max | 7242 |

Table 3:Data Description

We can see form the above data that the maximum sales is 7242 nad the average sale is 2402 and the minimum sale that has happened is 1070 . Hence we can also understand that there is skewness in the data. As mentioned earlier there are 187 records in the dataset .

Let’s explore the data more by plotting both Yearly and Monthly Box Plots:

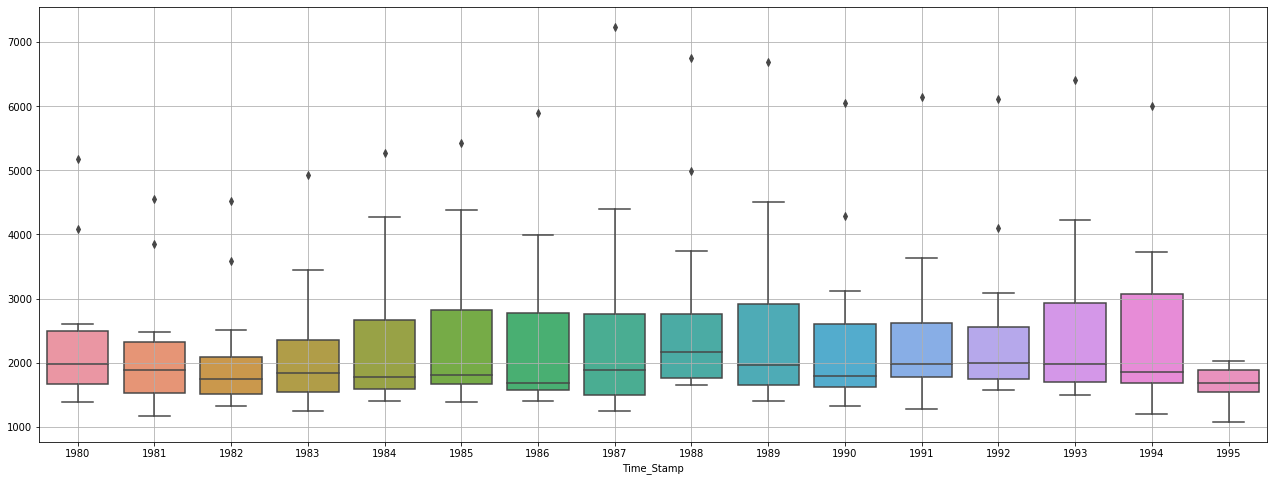


Figure 2:Yearly Box Plot

From the above Box Plot we can see that there are variations each year, apart from the first two years and the period from 1985 to 1988. The highest sales has seem to come in 1994, and the lowest sales recorded is of 1982 with an exception of 1995 which has only 7 months sales data. There is a declining trend from 1980 onwards to 1982 and an increasing trend from 1982 to 1989, with a fluctuation again in the trends from 1989 onwards that can be observed. We can also notice there are outliers every year but that can be ignored as it is a Time Series .

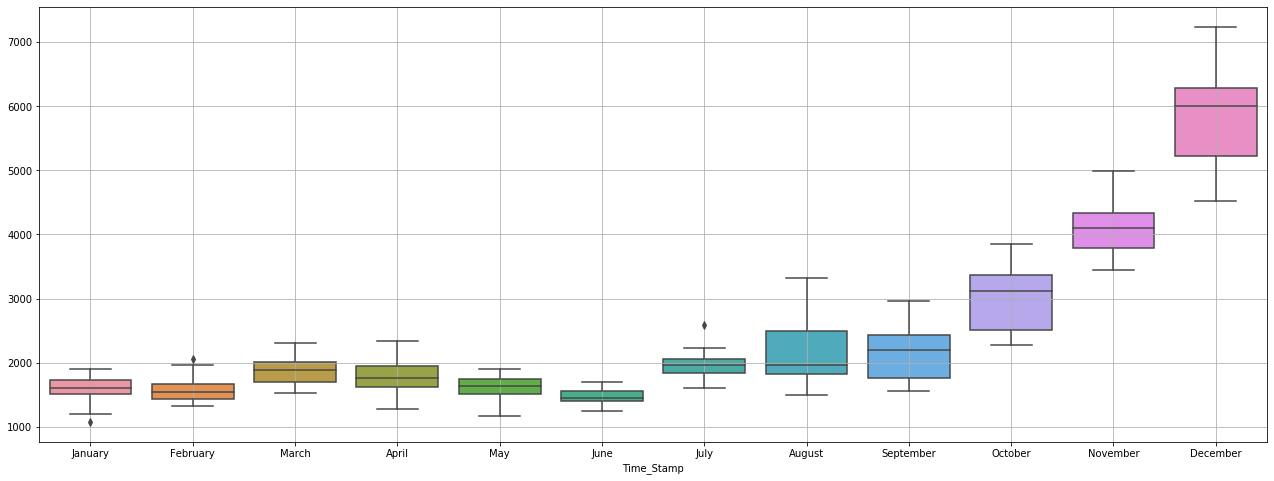


Figure 3:Monthly Box Plot

From the above Monthly Box Plot we can observe that there is an high element of seasonality as discussed earlier where December is a high performing month in terms of Sparkling Wine Sales with highest sales figures followed by November, October, August and September. This also tells us that the Sales increase towards the second half of the year and are comparatively lower during the first half of the year. The lowest sale recorded is in the month of June .

Lets go ahead and see the table of monthly sales of wine across years.



Table 4:Pivot Table of Monthly sales across years

Lets us now visually see the same table to understand the sales performance better :



Figure 4:Monthly Sales across year Plot

The graph again talks about the seasonality element which is present as we can see December seems to be the month that drives the highest sales for Sparkling Wine, with the second highest being in November and so on .

Lets go ahead and plot a time series month plot to understand the spread accidents across different years and within different months across years .

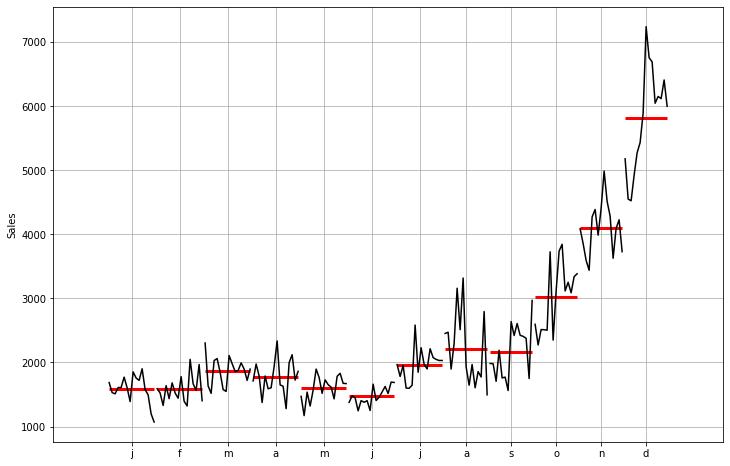
  
This plot shows us the behaviour of time series (Sparkling Sales) across various months. The red line is the median Value.

Figure 5:Time series Month Plot

The below figure shows the Emprical Cumulative Dsitribution:

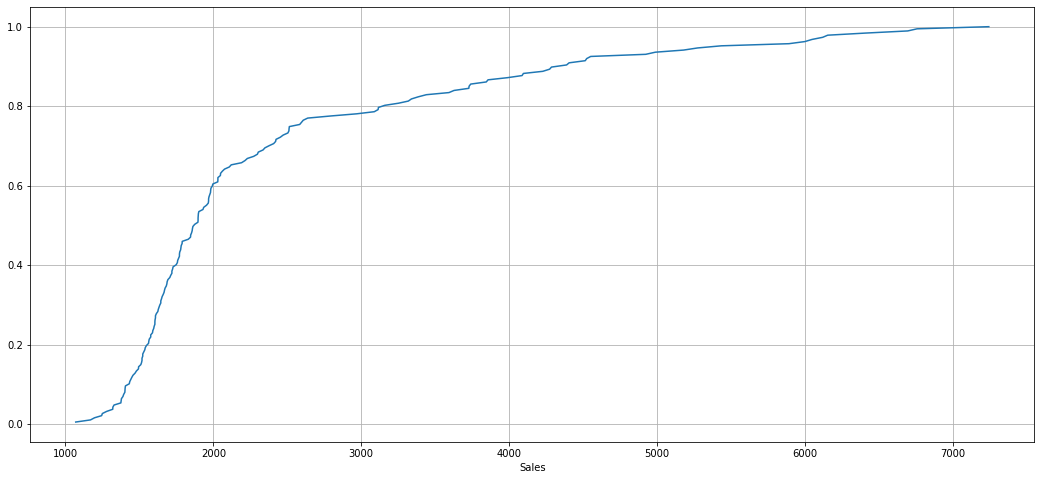


Figure 6:Emprical Cumulative Distribution Plot

This particular graph tells us what percentage of data points refer to what number of Sales .

Lets now plot the average retails sales per month and month on month percentage change of Sparkling sales .

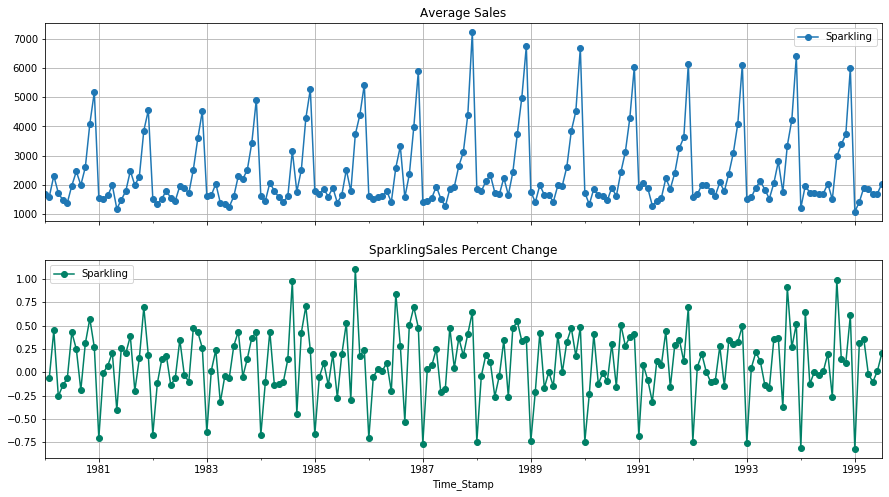


Figure 7:average retails sales per month and month on month percentage change of Sparkling sales.

The above two graphs tells us the Average 'Sparkling Sales' and the Percentage change of 'Sparkling sales’ with respect to the time.

Now we will move to perform the Decomposition:

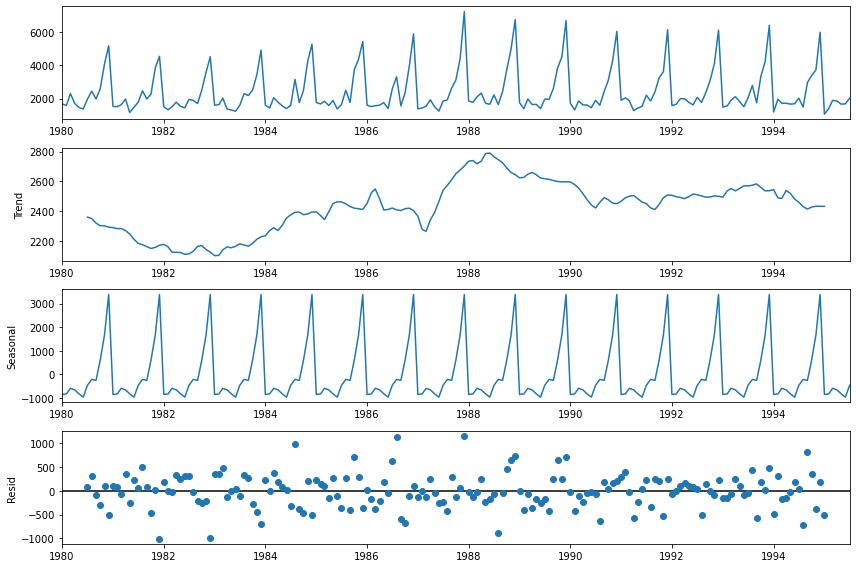


Figure 8:Additive model of the decomposition

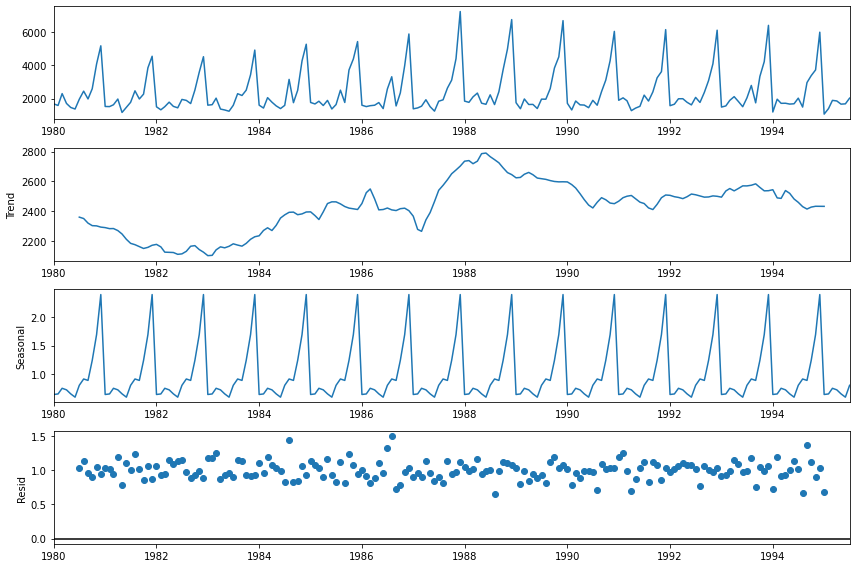


Figure 9:Multiplicative model decomposition

From the above figure we can clearly understand there is an upward and downward trend which indicate that the sales are not stable and not uniform. They have a strong factor of seasonality present in the data. The residual in the additive model shows a random pattern hence we can choose to go ahead with additive model .

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

Now we go ahead and split the time series datasets into Train and Test dataset. As given in the question that the Test data should start in 1991, so I have used 71% of the datasets for the Training Set and the rest of the dataset for Test datasets.

Below table shows the head and tail of training and test dataset .

First Few rows of the Training Dataset

|  |  |
| --- | --- |
|  | **Sparkling** |
| **Time\_Stamp** |  |
| **1/31/1980** | 1686 |
| **2/29/1980** | 1591 |
| **3/31/1980** | 2304 |
| **4/30/1980** | 1712 |
| **5/31/1980** | 1471 |

Last few rows of the Training Dataset

|  |  |
| --- | --- |
| **Time\_Stamp** | **Sparkling** |
| **8/31/1990** | 1605 |
| **9/30/1990** | 2424 |
| **10/31/1990** | 3116 |
| **11/30/1990** | 4286 |
| **12/31/1990** | 6047 |

First few rows of the Test Dataset

|  |  |
| --- | --- |
| **Time\_Stamp** | **Sparkling** |
| **1/31/1991** | 1902 |
| **2/28/1991** | 2049 |
| **3/31/1991** | 1874 |
| **4/30/1991** | 1279 |
| **5/31/1991** | 1432 |

Last few rows of the Test Dataset

|  |  |
| --- | --- |
| **Time\_Stamp** | **Sparkling** |
| **3/31/1995** | 1897 |
| **4/30/1995** | 1862 |
| **5/31/1995** | 1670 |
| **6/30/1995** | 1688 |
| **7/31/1995** | 2031 |

There are a total of 132 records in train data and 55 records in test dataset .



Figure 10: Plot for Training and Test Dataframes

The above figure shows the train dataset from 1981 to 1990 in blue color and Test dataset from 1991 onwards in orange color.

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

Lets us build various time series forecasting models to check the performance using RMSE scores .

Linear Regression model:

The below graph is the extract of the result from Linear Regression model :

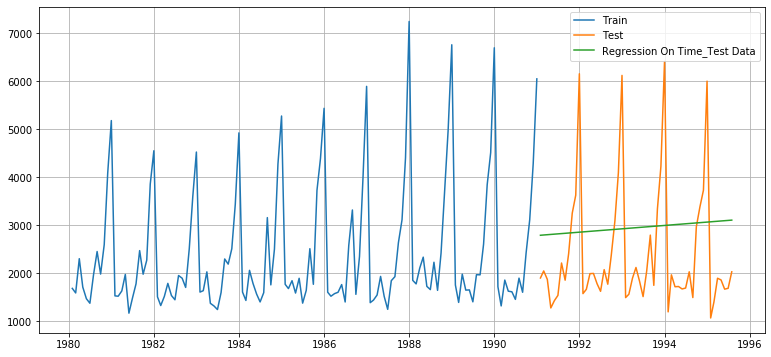


Figure 11:Linear Regression Outcome of Sparkling wine Time series

We can see from the plot that sparkling wine sales shows an upward trend. The RMSE score for this model is 1389.135 .

The summarized performance of the model run are :

|  |  |
| --- | --- |
|  | Test RMSE |
| RegressionOnTime | 1389.135175 |

Table 5: Summarized performance of the model

Naive Forecast model:

The below graph is the extract of the result from Naïve Forecast model

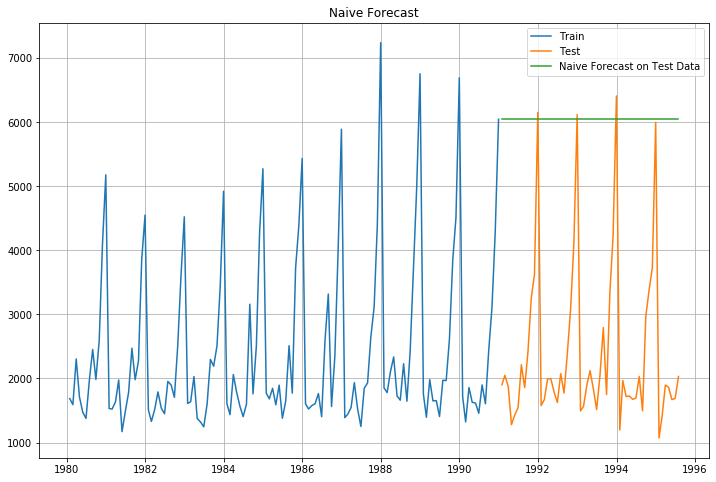


Figure 12:Naive Forecast outcome of Sparkling wine Time series

The RSMSE score of Naïve model is 3864.279, the naïve model is not suitable for the wine dataset since the forecast depends on the previous last observation .

The summarized performance of the models run are :

|  |  |
| --- | --- |
|  | Test RMSE |
| RegressionOnTime | 1389.135175 |
| NaiveModel | 3864.279352 |

Table 6: Summarized performance of the model

Simple Average

The below graph is the extract of the result from Simple Average Forecast model

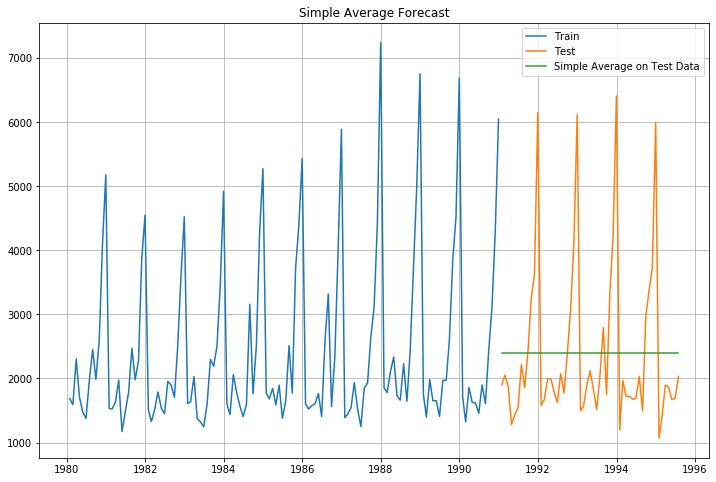


Figure 13:Simple Average Forecast outcome of Sparkling wine Time series

The RMSE score of Simple Average Forecast model is 1275.08

The summarized performance of the model are:

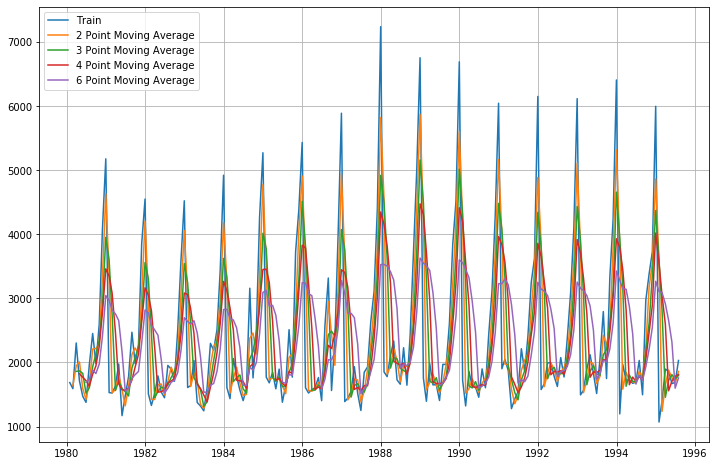
|  |  |
| --- | --- |
|  | Test RMSE |
| RegressionOnTime | 1389.135175 |
| NaiveModel | 3864.279352 |
| SimpleAverageModel | 1275.081804 |

Table 7:Summarized performance of the model

As we can see the Simple Average Model is the best performing model from all the models run till now.

Moving Average

The below graph is the extract of the result after running Moving Average Forecast model



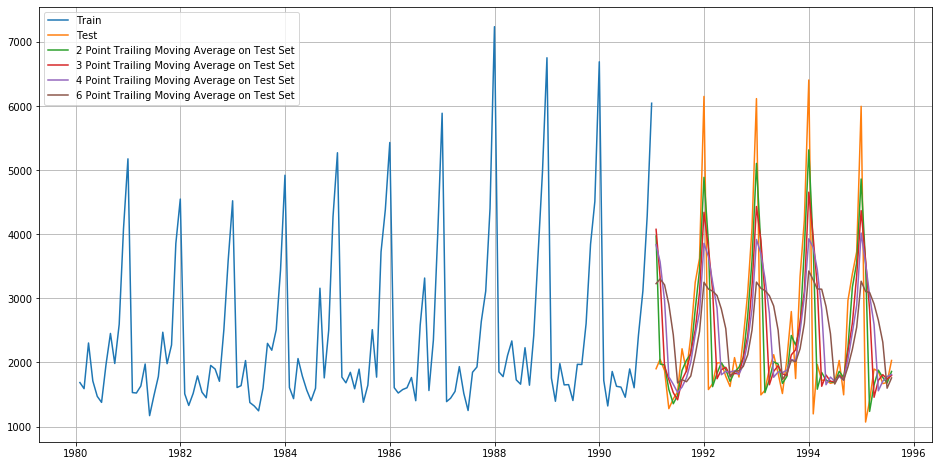


Figure 14:Moving Average Model outcome on the sparkling wine time series

For 2 point Moving Average Model forecast on the Training Data, RMSE is 813.401

For 3 point Moving Average Model forecast on the Training Data, RMSE is 1028.606

For 4 point Moving Average Model forecast on the Training Data, RMSE is 1156.590

For 6 point Moving Average Model forecast on the Training Data, RMSE is 1283.927

The summarized performance of the model are:

|  |  |
| --- | --- |
| **Test RMSE** |  |
| RegressionOnTime | 1389.135175 |
| NaiveModel | 3864.279352 |
| SimpleAverageModel | 1275.081804 |
| 2pointTrailingMovingAverage | 813.400684 |
| 3pointTrailingMovingAverage | 1028.605756 |
| 4pointTrailingMovingAverage | 1156.589694 |
| 6pointTrailingMovingAverage | 1283.927428 |

Table 8:Summarised performance of the models

From the above table we can see 2 point Trailing Moving Average is the best performing model till now.

Simple Exponential Smoothing

After running the SES model we can see the performance as below,

{'smoothing\_level': 0.0,

'smoothing\_slope': nan,

'smoothing\_seasonal': nan,

'damping\_slope': nan,

'initial\_level': 2403.7856210776245,

'initial\_slope': nan,

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

'use\_boxcox': False,

'lamda': None,

'remove\_bias': False}

Figure 15:SES parameters for Sparkling wine dataset

The below graph is the extract of the result after running Simple Exponential Smoothing Forecast model

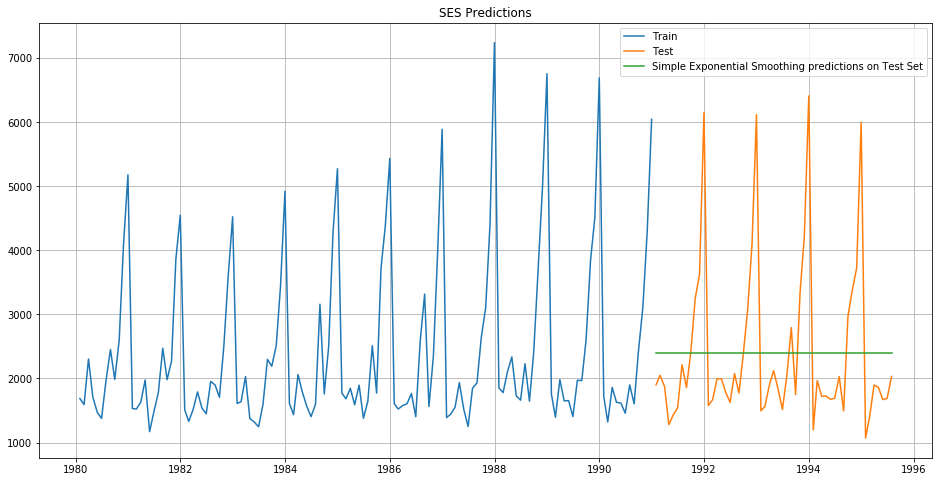


Figure 16:Simple Exponential Smoothing Outcome on the Sparkling Wine Time Series

The RMSE for the SES model is 1275.082

The summarized performance of the model are:

|  |  |
| --- | --- |
|  | Test RMSE |
| RegressionOnTime | 1389.135175 |
| NaiveModel | 3864.279352 |
| SimpleAverageModel | 1275.081804 |
| 2pointTrailingMovingAverage | 813.400684 |
| 3pointTrailingMovingAverage | 1028.605756 |
| 4pointTrailingMovingAverage | 1156.589694 |
| 6pointTrailingMovingAverage | 1283.927428 |
| SimpleExponentialSmoothing | 1275.081823 |

Table 9:Summarised performance of the models

As we can see above the best model till now is 2 point Trailing Moving Average.

Double Exponential Smoothing

After running the SES model we can see the performance as below,

{'smoothing\_level': 0.6478137983008563,

'smoothing\_slope': 0.0,

'smoothing\_seasonal': nan,

'damping\_slope': nan,

'initial\_level': 1686.0837778875616,

'initial\_slope': 27.06051184255422,

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

'use\_boxcox': False,

'lamda': None,

'remove\_bias': False}

Figure 17:DES parameters for sparkling wine dataset

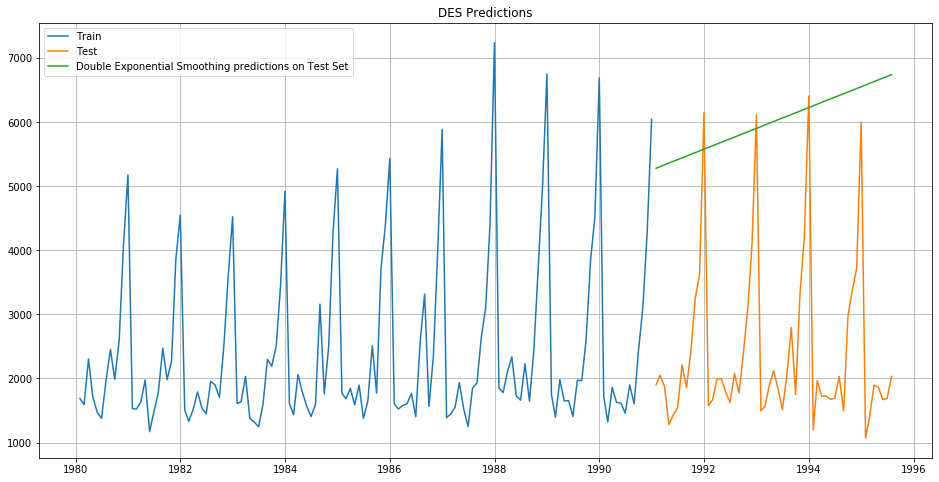


Figure 18:Double Exponential Smoothing Outcome on the Sparkling Wine Time Series

The RMSE of DES model is 3851.053

The summarized performance of the model are

|  |  |
| --- | --- |
|  | Test RMSE |
| RegressionOnTime | 1389.135175 |
| NaiveModel | 3864.279352 |
| SimpleAverageModel | 1275.081804 |
| 2pointTrailingMovingAverage | 813.400684 |
| 3pointTrailingMovingAverage | 1028.605756 |
| 4pointTrailingMovingAverage | 1156.589694 |
| 6pointTrailingMovingAverage | 1283.927428 |
| SimpleExponentialSmoothing | 1275.081823 |
| DoubleExponentialSmoothing | 3851.072597 |

Table 10:Summarised performance of the models

Triple Exponential Smoothing:

After running the TES model with Trend as Additive and Seasonality as Multiplicative , we can see the parameters as below

{'smoothing\_level': 0.15422043724083787,

'smoothing\_slope': 2.7016118310306043e-21,

'smoothing\_seasonal': 0.37132361273463527,

'damping\_slope': nan,

'initial\_level': 1639.999343205301,

'initial\_slope': 4.848648986644784,

'initial\_seasons': array([1.00841815, 0.96898257, 1.24179222, 1.13205604, 0.93980962,

0.93811109, 1.22458123, 1.54428823, 1.27336092, 1.63198272,

2.48293081, 3.11862208]),

'use\_boxcox': False,

'lamda': None,

'remove\_bias': False}

Figure 19:TES parameters for sparkling wine dataset

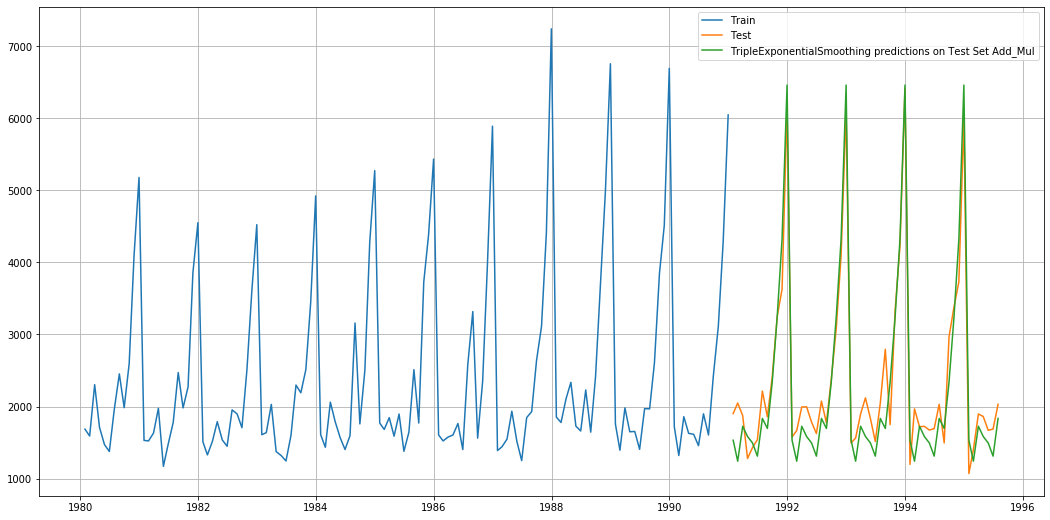


Figure 20:TES Outcome on Sparkling wine Timeseries

The RMSE of the above TES model is 383.182

After running the TES model with Trend as Additive and Seasonality as Additive, we can see the parameters as below

{'smoothing\_level': 0.08621947613454733,

'smoothing\_slope': 2.6874330863382154e-08,

'smoothing\_seasonal': 0.4763612183448062,

'damping\_slope': nan,

'initial\_level': 1684.809720734794,

'initial\_slope': 0.006601124491771914,

'initial\_seasons': array([ 39.19059509, -37.24835927, 464.88056614, 205.99095389,

-140.66424075, -156.79570166, 338.0811185 , 856.82160873,

403.52711408, 971.26615796, 2401.64073231, 3426.75506275]),

'use\_boxcox': False,

'lamda': None,

'remove\_bias': False}

Figure 21:TES Parameters for Sparkling wine dataset

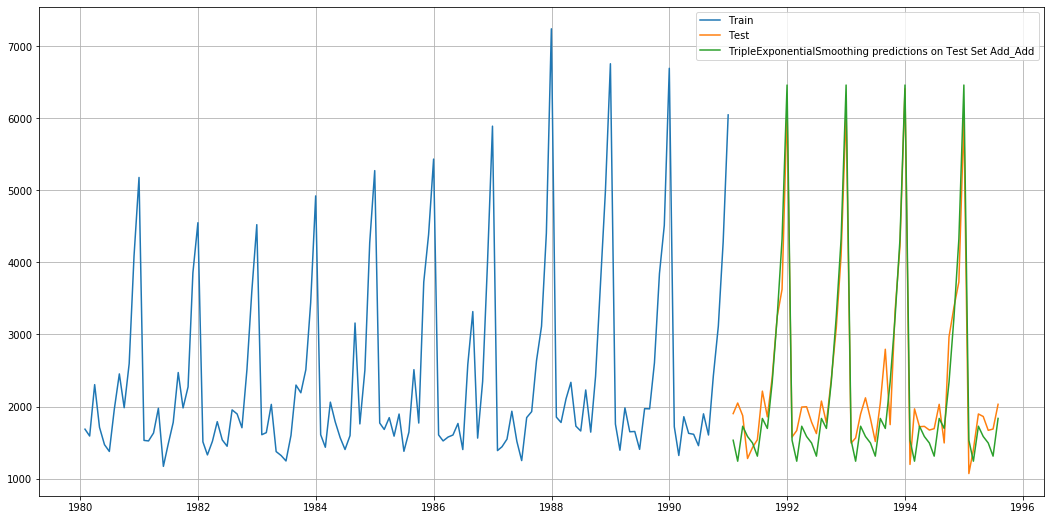


Figure 22:TES outcome of sparkling wine timeseries

The RMSE for above TES model is 362.731

After running the TES model with Trend as Multiplicative and Seasonality as Multiplicative , we can see the parameters as below

{'smoothing\_level': 0.15340799299595567,

'smoothing\_slope': 1.4578242858544946e-16,

'smoothing\_seasonal': 0.36918563259948295,

'damping\_slope': nan,

'initial\_level': 1640.00008048211,

'initial\_slope': 1.002822086250115,

'initial\_seasons': array([1.00890155, 0.96940434, 1.2425392 , 1.13281286, 0.94019572,

0.93836077, 1.225938 , 1.54668454, 1.27513351, 1.63558138,

2.48868762, 3.12682145]),

'use\_boxcox': False,

'lamda': None,

'remove\_bias': False}

Figure 23:TES parameters for sparkling wine dataset

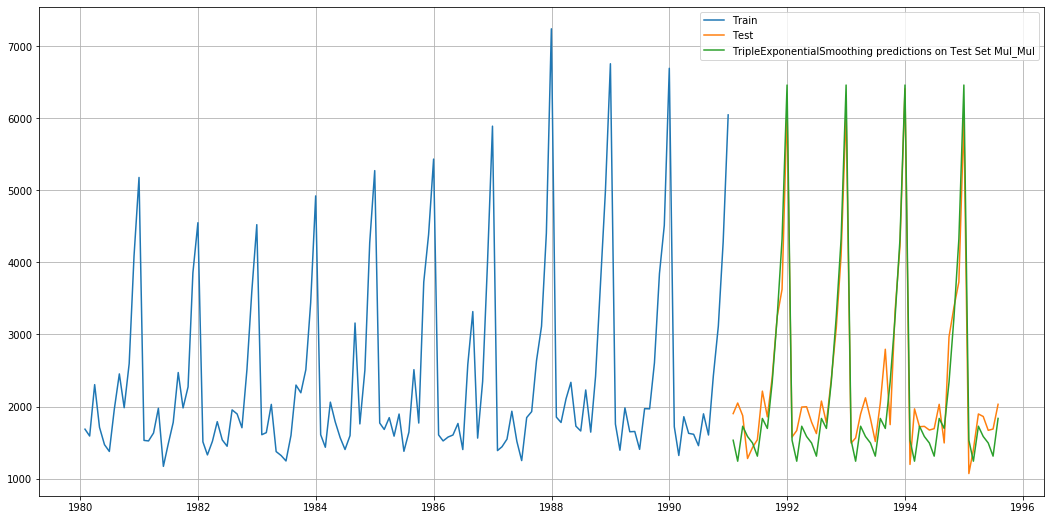


Figure 24:TES outcome on sparkling wine timeseries

The RMSE of above model is 392.869

The summarized performance of all the model are as below :

|  |  |
| --- | --- |
|  | Test RMSE |
| RegressionOnTime | 1389.135175 |
| NaiveModel | 3864.279352 |
| SimpleAverageModel | 1275.081804 |
| 2pointTrailingMovingAverage | 813.400684 |
| 3pointTrailingMovingAverage | 1028.605756 |
| 4pointTrailingMovingAverage | 1156.589694 |
| 6pointTrailingMovingAverage | 1283.927428 |
| SimpleExponentialSmoothing | 1275.081823 |
| DoubleExponentialSmoothing | 3851.072597 |
| TES\_Add\_Mul | 383.1827454 |
| TES\_MUL\_MUL | 392.8691267 |
| TES\_ADD\_ADD | 362.7316572 |

Table 11:Summaruized performances of all the models

From the above table we can see that out of all the models the Triple Exponential smoothening with additive trend and seasonality is the best performing model with lowest RMSE score of 362.73165

Now that we have run all the models , lets now see the summary of all the models on a single graph as below :

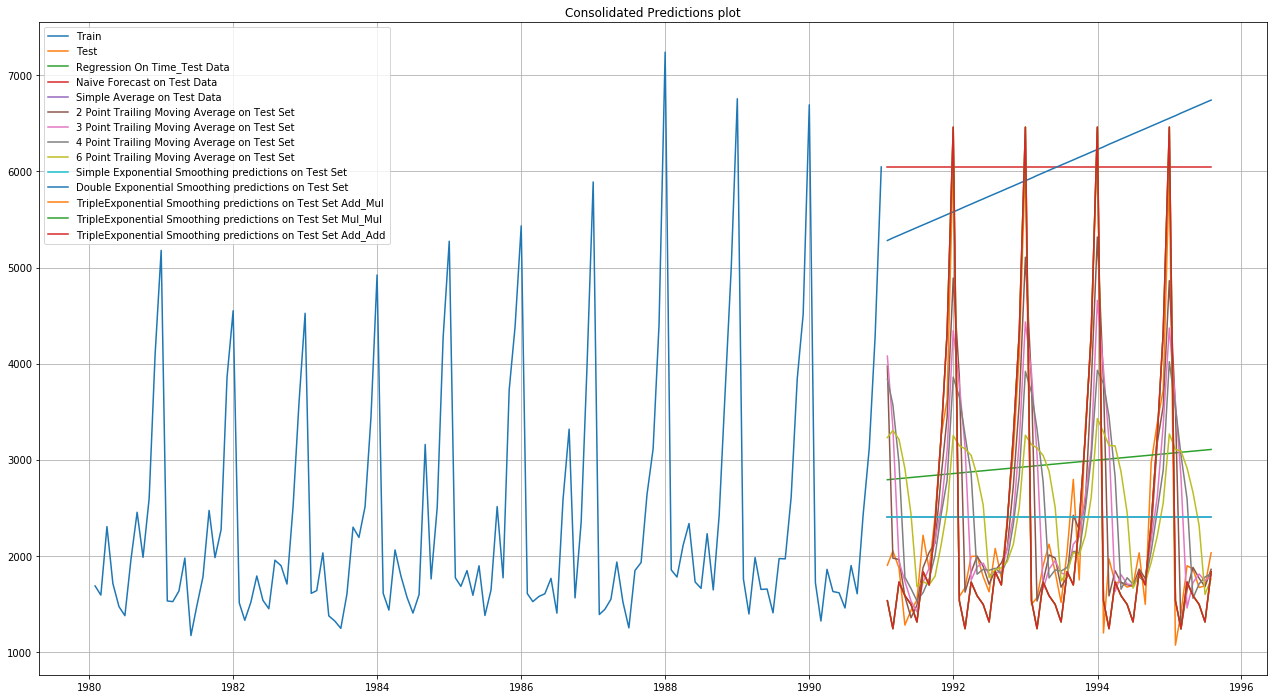


Figure 25:Summary of all prediction on a single graph

## 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.

To check whether the series is stationary , we use the Augmented Dickey Fuller (ADF) test. 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:

* 𝐻0 : The Time Series has a unit root and is thus non-stationary.
* 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 𝛼α value of 0.05 as mentioned in the question .

After performing the Stationarity check we get the below result :

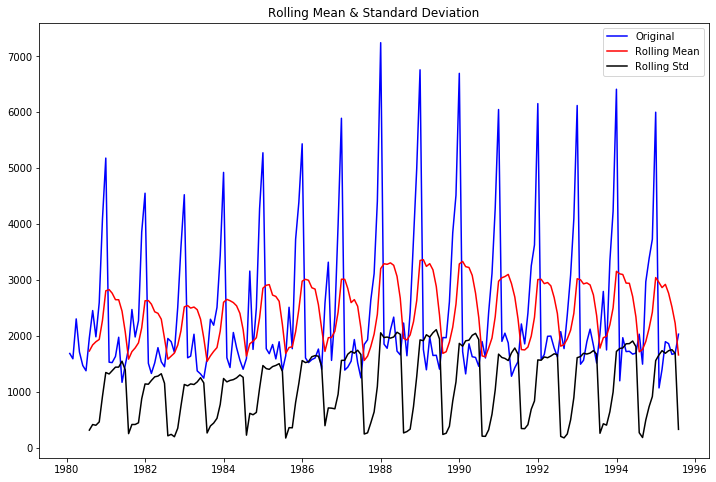


Figure 26:Stationarity graph

Results of Dickey-Fuller Test:

Test Statistic -1.360497

p-value 0.601061

#Lags Used 11.000000

Number of Observations Used 175.000000

Critical Value (1%) -3.468280

Critical Value (5%) -2.878202

Critical Value (10%) -2.575653

Since the p value is greater than alpha 0.05, time series is not stationary. We can take the next levels of differencing to make the time series stationary . Let us take a difference of order 1 and check whether the Time Series is stationary or not.

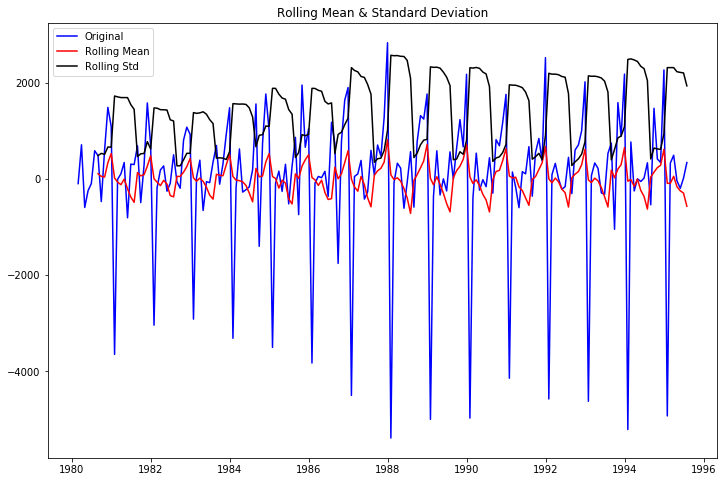


Figure 27:Stationarity graph after differencing

Results of Dickey-Fuller Test:

Test Statistic -45.050301

p-value 0.000000

#Lags Used 10.000000

Number of Observations Used 175.000000

Critical Value (1%) -3.468280

Critical Value (5%) -2.878202

Critical Value (10%) -2.575653

We can see that the alpha value is less than 0.05 and the Time Series now is indeed stationary .

## 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 MODEL:**

ARIMA model AIC values of the parameter combinations:

ARIMA(0, 1, 0) - AIC:2267.6630357855465

ARIMA(0, 1, 1) - AIC:2263.060015592425

ARIMA(0, 1, 2) - AIC:2234.4083231342192

ARIMA(1, 1, 0) - AIC:2266.608539319009

ARIMA(1, 1, 1) - AIC:2235.75509467321

ARIMA(1, 1, 2) - AIC:2234.527200452548

ARIMA(2, 1, 0) - AIC:2260.3657439680865

ARIMA(2, 1, 1) - AIC:2233.7776263299156

ARIMA(2, 1, 2) - AIC:2213.5092123639834

We can see above the lowest AIC is 2213.509 with (2,1,2) param .

Lets go ahead and evaluate the model

SARIMAX Results

==============================================================================

Dep. Variable: Sparkling No. Observations: 132

Model: ARIMA(2, 1, 2) Log Likelihood -1101.755

Date: Sun, 23 Jan 2022 AIC 2213.509

Time: 08:25:08 BIC 2227.885

Sample: 01-31-1980 HQIC 2219.351

- 12-31-1990

Covariance Type: opg

==============================================================================

coef std err z P>|z| [0.025 0.975]

------------------------------------------------------------------------------

ar.L1 1.3121 0.046 28.782 0.000 1.223 1.401

ar.L2 -0.5593 0.072 -7.741 0.000 -0.701 -0.418

ma.L1 -1.9917 0.109 -18.217 0.000 -2.206 -1.777

ma.L2 0.9999 0.110 9.109 0.000 0.785 1.215

sigma2 1.099e+06 1.99e-07 5.51e+12 0.000 1.1e+06 1.1e+06

===================================================================================

Ljung-Box (Q): 293.72 Jarque-Bera (JB): 14.46

Prob(Q): 0.00 Prob(JB): 0.00

Heteroskedasticity (H): 2.43 Skew: 0.61

Prob(H) (two-sided): 0.00 Kurtosis: 4.08

===================================================================================

Figure 28:Statistical analysis of ARIMA model

The RMSE and MAPE values are:1299 .979 & 47.099 .

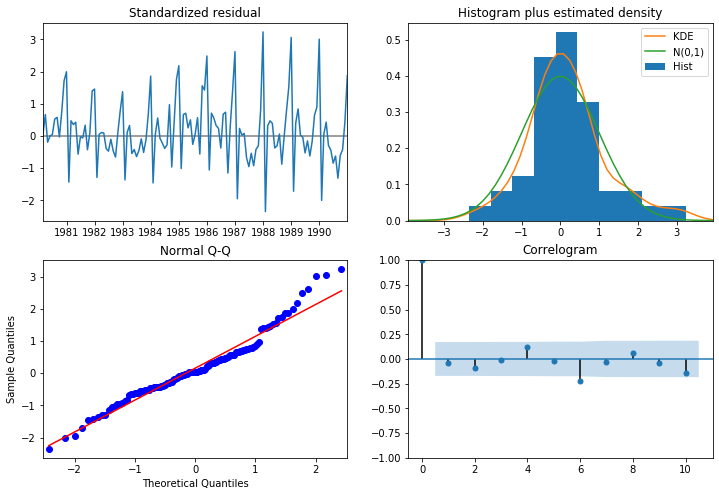


Figure 29:ARIMA Plot

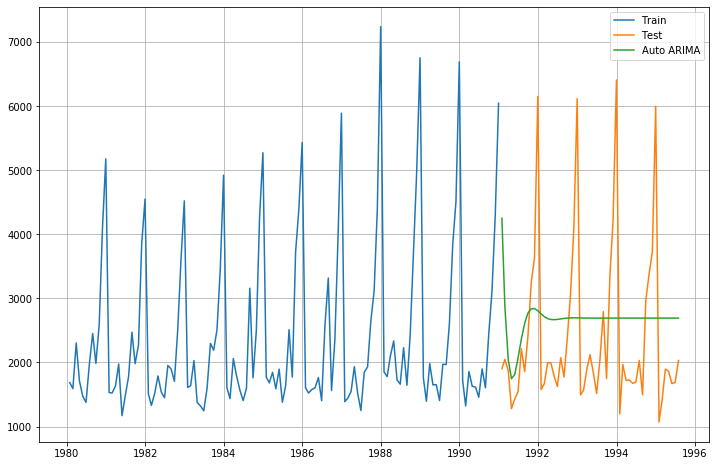


Figure 30:Auto ARIMA Plot on the test data and forecasted values

**SARIMA MODEL :** SARIMA model AIC values of the parameter combinations in descending order of the first records are as below :

|  |  |  |  |
| --- | --- | --- | --- |
|  | **param** | **seasonal** | **AIC** |
| 47 | (1, 1, 2) | (0, 1, 2, 12) | 1382.348 |
| 20 | (0, 1, 2) | (0, 1, 2, 12) | 1382.484 |
| 50 | (1, 1, 2) | (1, 1, 2, 12) | 1384.138 |
| 74 | (2, 1, 2) | (0, 1, 2, 12) | 1384.318 |
| 23 | (0, 1, 2) | (1, 1, 2, 12) | 1384.399 |

Table 12:AIC values of first 5 Params of SARIMA model in descending order

We can see the lowest AIC value is 1382.348 with param (1,1,2) and seasonal (0,1,2,12) .

Lets go ahead and evaluate the model using statistical values :

SARIMAX Results

==========================================================================================

Dep. Variable: Sparkling No. Observations: 132

Model: SARIMAX(1, 1, 2)x(0, 1, 2, 12) Log Likelihood -685.174

Date: Sun, 23 Jan 2022 AIC 1382.348

Time: 09:02:28 BIC 1397.479

Sample: 01-31-1980 HQIC 1388.455

- 12-31-1990

Covariance Type: opg

==============================================================================

coef std err z P>|z| [0.025 0.975]

------------------------------------------------------------------------------

ar.L1 -0.5507 0.287 -1.922 0.055 -1.112 0.011

ma.L1 -0.1612 0.235 -0.687 0.492 -0.621 0.299

ma.L2 -0.7218 0.175 -4.132 0.000 -1.064 -0.379

ma.S.L12 -0.4062 0.092 -4.401 0.000 -0.587 -0.225

ma.S.L24 -0.0274 0.138 -0.198 0.843 -0.298 0.243

sigma2 1.705e+05 2.45e+04 6.956 0.000 1.22e+05 2.19e+05

===================================================================================

Ljung-Box (Q): 20.51 Jarque-Bera (JB): 13.48

Prob(Q): 1.00 Prob(JB): 0.00

Heteroskedasticity (H): 0.89 Skew: 0.60

Prob(H) (two-sided): 0.75 Kurtosis: 4.44

===================================================================================

Figure 31:Statistical Analysis of SARIMA Model

The RMSE and MAPE values are 382.576 & 15.332 .

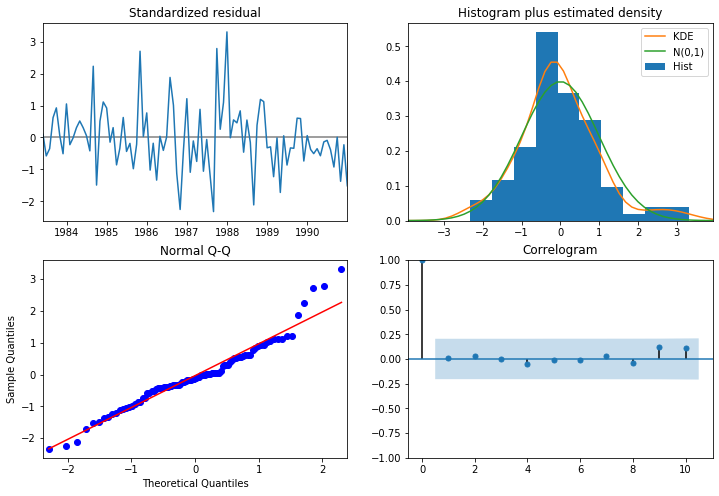


Figure 32:SARIMA Plot

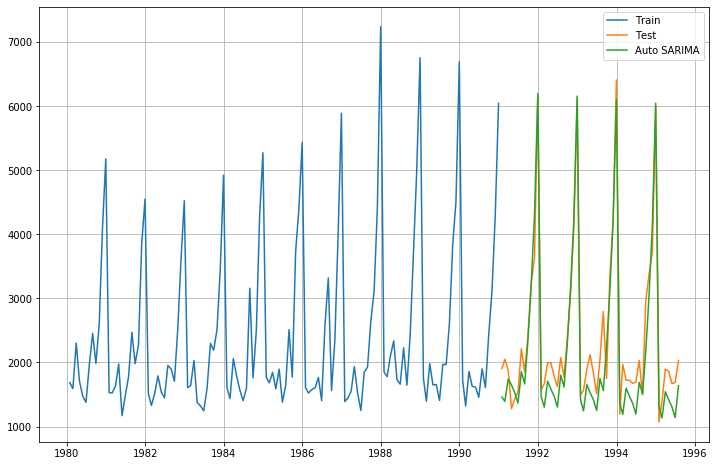


Figure 33:Auto SARIMA Plot

The summarized performances of the all the models built are as below :

|  |  |
| --- | --- |
|  | Test RMSE |
| RegressionOnTime | 1389.135175 |
| NaiveModel | 3864.279352 |
| SimpleAverageModel | 1275.081804 |
| 2pointTrailingMovingAverage | 813.400684 |
| 3pointTrailingMovingAverage | 1028.605756 |
| 4pointTrailingMovingAverage | 1156.589694 |
| 6pointTrailingMovingAverage | 1283.927428 |
| SimpleExponentialSmoothing | 1275.081823 |
| DoubleExponentialSmoothing | 3851.072597 |
| TES\_Add\_Mul | 383.1827454 |
| TES\_MUL\_MUL | 392.8691267 |
| TES\_ADD\_ADD | 362.7316572 |
| AIC\_ARIMA(2,1,2) | 1299.97947 |
| AIC\_SARIMA(1,1,2)(0,1,2,12) | 382.5767219 |

Table 13:Summarized performance of all the models

From the above summary of the performance of models we can see that the Triple exponential smoothing of additive models is still the best performing model till now .

## 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.

**ARIMA model based on the cut off points**

Lets first have a look at the ACF and PACF Plot:

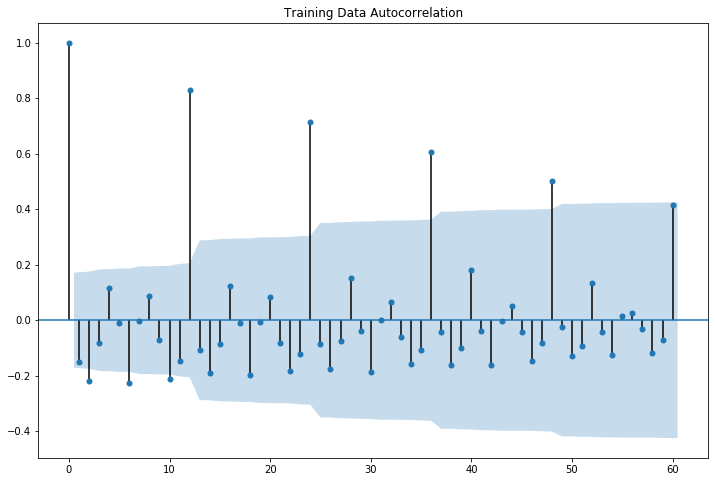


Figure 34:ACF Plot of ARIMA Model

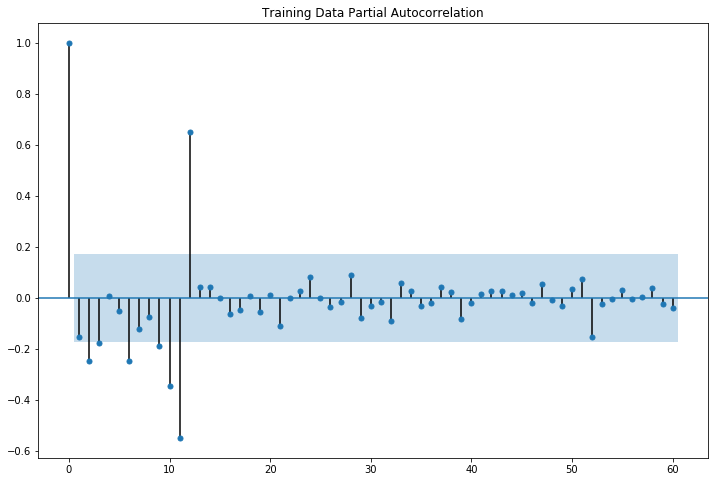


Figure 35:PACF Plot of ARIMA Model

Here, we have taken alpha=0.05.

* 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 can say that both the PACF and ACF plot cuts-off at lag 0.

Lets go ahead and evaluate the model using statistical values by taking the parameters as (0,1,0):

SARIMAX Results

==============================================================================

Dep. Variable: Sparkling No. Observations: 132

Model: ARIMA(0, 1, 0) Log Likelihood -1132.832

Date: Sun, 23 Jan 2022 AIC 2267.663

Time: 09:00:03 BIC 2270.538

Sample: 01-31-1980 HQIC 2268.831

- 12-31-1990

Covariance Type: opg

==============================================================================

coef std err z P>|z| [0.025 0.975]

------------------------------------------------------------------------------

sigma2 1.885e+06 1.29e+05 14.658 0.000 1.63e+06 2.14e+06

===================================================================================

Ljung-Box (Q): 347.60 Jarque-Bera (JB): 198.83

Prob(Q): 0.00 Prob(JB): 0.00

Heteroskedasticity (H): 2.46 Skew: -1.92

Prob(H) (two-sided): 0.00 Kurtosis: 7.65

Figure 36:Statistical Analysis of Manual ARIMA model

Below is the diagnosis for Manual ARIMA Plot:

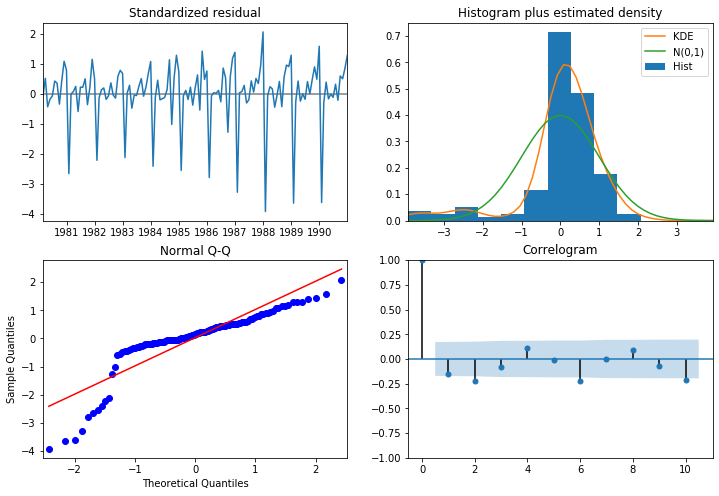


Figure 37:Manual ARIMA Plot

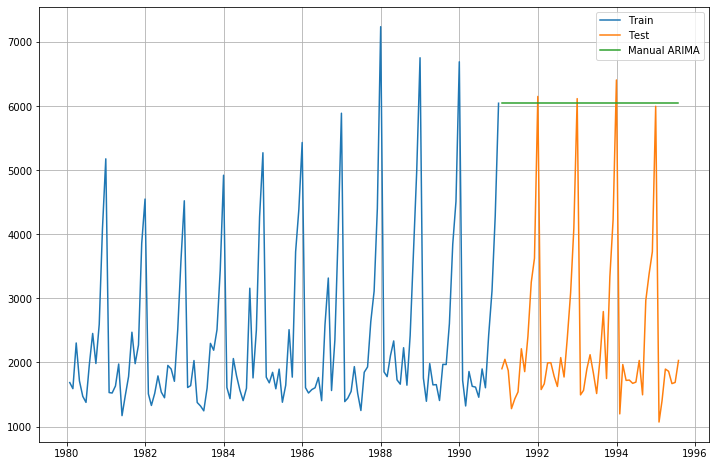


Figure 38:Manual ARIMA plot prediction on training and test data

The RSMSE AND MAPE values of Manual ARIMA model are 3864.279 & 201.327

**SARIMA model based on the cut off points**

Lets first have a look at the ACF and PACF Plot:

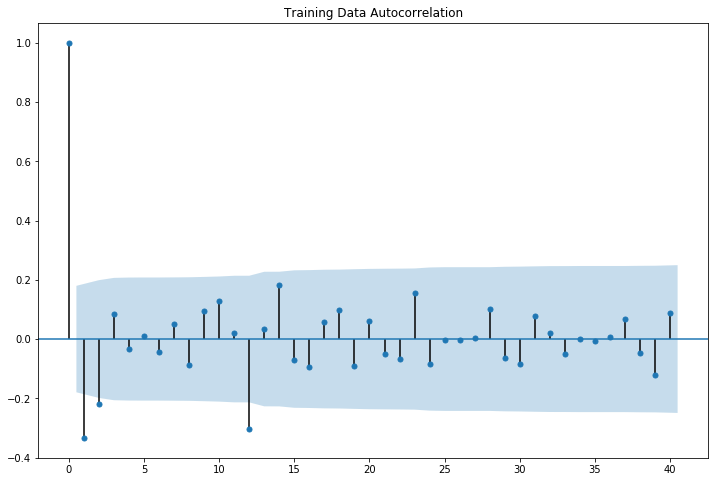
****

Figure 39:ACF Plot of SARIMA Model

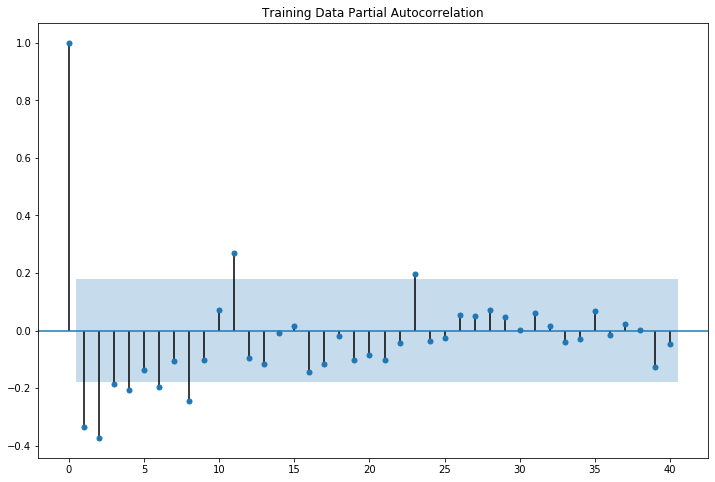


Figure 40:PACF Model of SARIMA Model

Here, we have taken alpha=0.05.

We are going to take the seasonal period as 12 or its multiple e.g. 24. We are taking the p value to be 4 and the q value to be 2 as the 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 1.

We check the ACF and the PACF plots only at multiples of 12 (since 12 is the seasonal period).

Lets go ahead and evaluate the model using statistical values by taking the parameters as (4,1,2):

SARIMAX Results

==========================================================================================

Dep. Variable: Sparkling No. Observations: 132

Model: SARIMAX(4, 1, 2)x(0, 1, 2, 12) Log Likelihood -684.793

Date: Sun, 23 Jan 2022 AIC 1387.586

Time: 09:04:47 BIC 1410.283

Sample: 01-31-1980 HQIC 1396.747

- 12-31-1990

Covariance Type: opg

==============================================================================

coef std err z P>|z| [0.025 0.975]

------------------------------------------------------------------------------

ar.L1 -0.4852 0.368 -1.320 0.187 -1.206 0.235

ar.L2 0.0471 0.178 0.265 0.791 -0.301 0.395

ar.L3 0.0413 0.118 0.349 0.727 -0.191 0.273

ar.L4 -0.0669 0.201 -0.332 0.740 -0.462 0.328

ma.L1 -0.2212 0.353 -0.627 0.530 -0.912 0.470

ma.L2 -0.6791 0.349 -1.945 0.052 -1.363 0.005

ma.S.L12 -0.4262 0.105 -4.077 0.000 -0.631 -0.221

ma.S.L24 -0.0233 0.144 -0.162 0.871 -0.305 0.259

sigma2 1.688e+05 2.84e+04 5.939 0.000 1.13e+05 2.24e+05

===================================================================================

Ljung-Box (Q): 19.72 Jarque-Bera (JB): 10.42

Prob(Q): 1.00 Prob(JB): 0.01

Heteroskedasticity (H): 0.90 Skew: 0.57

Prob(H) (two-sided): 0.78 Kurtosis: 4.18

Figure 41:Statistical Analysis of the Manual SARIMA Model

Below is the diagnosis of the Manual SARIMA Model:

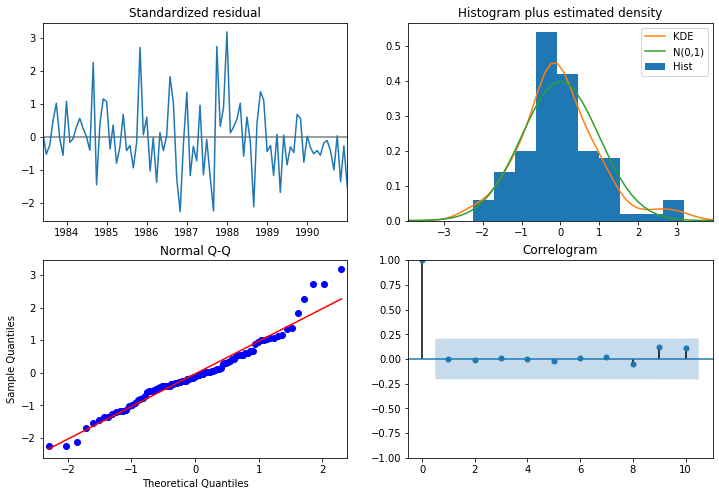


Figure 42:Manual SARIMA Plot

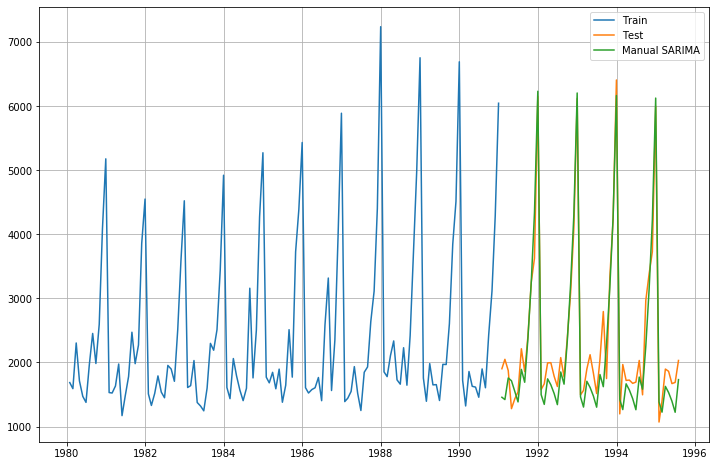


Figure 43:Plot of Trianing , Test and forecasted values of Manual SARIMA Model

The RSME and MAPE value of SARIMA model are 352.106 and 13.807

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

The summarized performances of the all the models built are as below :

|  |  |
| --- | --- |
|  | Test RMSE |
| RegressionOnTime | 1389.135175 |
| NaiveModel | 3864.279352 |
| SimpleAverageModel | 1275.081804 |
| 2pointTrailingMovingAverage | 813.400684 |
| 3pointTrailingMovingAverage | 1028.605756 |
| 4pointTrailingMovingAverage | 1156.589694 |
| 6pointTrailingMovingAverage | 1283.927428 |
| SimpleExponentialSmoothing | 1275.081823 |
| DoubleExponentialSmoothing | 3851.072597 |
| TES\_Add\_Mul | 383.1827454 |
| TES\_MUL\_MUL | 392.8691267 |
| TES\_ADD\_ADD | 362.7316572 |
| AIC\_ARIMA(2,1,2) | 1299.97947 |
| Manual\_ARIMA(0,1,0) | 3864.279352 |
| AIC\_SARIMA(1,1,2)(0,1,2,12) | 382.5767219 |
| Manual\_SARIMA(4,1,2)(0,1,2,12) | 352.1061891 |

Figure 44:Summarized Performance of all the models

From all the above models built we can see that the lowest RMSE value is for Manual SARIMA model of 352.106 with (4,1,2) params, hence concluding this as the best among all the model built .

## 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.

SARIMA model with parameters (1,1,2) x (1,0,2,12) is proposed to use for forecast of next 12 months using full data. Details of the model are as under :

SARIMAX Results

==========================================================================================

Dep. Variable: Sparkling No. Observations: 187

Model: SARIMAX(1, 1, 2)x(1, 0, 2, 12) Log Likelihood -1173.413

Date: Sun, 23 Jan 2022 AIC 2360.826

Time: 11:25:53 BIC 2382.308

Sample: 01-31-1980 HQIC 2369.549

- 07-31-1995

Covariance Type: opg

==============================================================================

coef std err z P>|z| [0.025 0.975]

------------------------------------------------------------------------------

ar.L1 -0.6610 0.242 -2.734 0.006 -1.135 -0.187

ma.L1 -0.1801 0.219 -0.822 0.411 -0.609 0.249

ma.L2 -0.7387 0.192 -3.847 0.000 -1.115 -0.362

ar.S.L12 1.0157 0.012 84.464 0.000 0.992 1.039

ma.S.L12 -1.3874 0.338 -4.101 0.000 -2.050 -0.724

ma.S.L24 -0.1460 0.146 -1.001 0.317 -0.432 0.140

sigma2 6.532e+04 2.08e+04 3.136 0.002 2.45e+04 1.06e+05

===================================================================================

Ljung-Box (Q): 18.00 Jarque-Bera (JB): 27.47

Prob(Q): 1.00 Prob(JB): 0.00

Heteroskedasticity (H): 1.03 Skew: 0.52

Prob(H) (two-sided): 0.93 Kurtosis: 4.76

===================================================================================

Figure 45:Model performnace using full data

The RMSE of the full model is 539.923722

Below is the graphical representation of the predicted sales for sparkling wine for next 12 months :

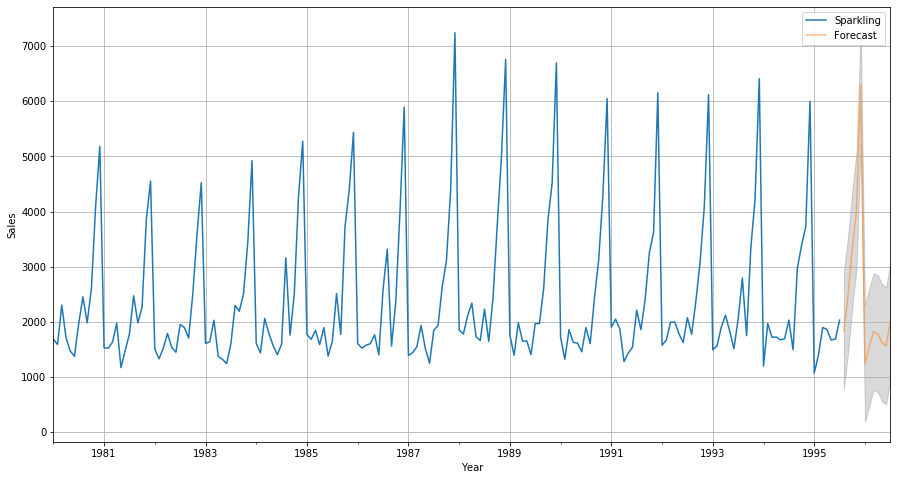


Figure 46:Graph of Predicted sales of Sparkling Wine

The predicted sales of Sparkling Wine for next 12 months is as below :



Table 14:Predicted sales for next 12 months

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

Busines study may be conducted to find why sales are not increasing and what are the contributing factors. Study can also include to see which which other wine product has substituted higher sales in the years of low sales of sparkling. With promotion and focused effort with micro detailing it may be feasible to increase the sales. Sales of sparkling wine is higher in the later part of the year, which may be due to the the climatic conditions or festive seasons towards the end of the year .

# PROBLEM 2 – ROSE DATASET

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

I have imported the data series and observed that there are 2 columns , one which has an YearMonth value with it , which needs to converted to appropriate time series data for analysis and the other column contains the monthly sales data of Rose wine from 1980 to 1995 . The table below shows the data imported from the dataset .

|  |  |  |
| --- | --- | --- |
|  | YearMonth | Rose |
| 0 | 1980-01 | 112 |
| 1 | 1980-02 | 118 |
| 2 | 1980-03 | 129 |
| 3 | 1980-04 | 99 |
| 4 | 1980-05 | 116 |
|  |  |  |

Table 15:Dataset of Rose

Below is the table after converting the YearMonth column into the time series data and making our time series reference as the index, since we can conveniently do the slicing i.e obtain data for a specific time period, and check values for corresponding time point .

|  |  |
| --- | --- |
| YearMonth | Rose |
| 1/1/1980 | 112 |
| 2/1/1980 | 118 |
| 3/1/1980 | 129 |
| 4/1/1980 | 99 |
| 5/1/1980 | 116 |

Table 16:Dataset after converting into time series

Lets go ahead and Plot the graph of Rose data set:

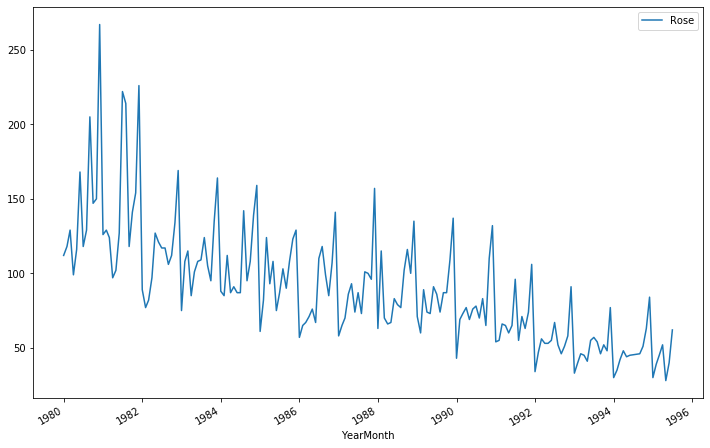


Figure 47:Rose Timeseries Plot

From the above graph we can observe that the sales of Rose wines are showing a downward trend for the last 15 years . We can also observe that there is an seasonality element that is present, where in it can be understood that the sales are more towards the beginning and end of the year in comparison with other months. Further in our analysis we will explore more about trends and seasonality when we perform decomposition .

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

Now we go ahead with the Exploratory Data Analysis , we saw that there are 187 entries for sales column and 2 columns including time series column which is used as index for referencing . Further we saw that there are 2 null values which have been dropped, but 90 duplicate entries which can be ignored since each value correspond to a different time index.

Below is the data description :

|  |  |
| --- | --- |
|  | Rose |
| count | 185 |
| mean | 90.39459 |
| std | 39.17534 |
| min | 28 |
| 25% | 63 |
| 50% | 86 |
| 75% | 112 |
| max | 267 |

Table 17:Data Description of Rose Wine Dataset

We can see form the above data that the maximum sales is 267 , the average sale is 90.39 and the minimum sale that has happened is 28 . Hence we can also understand that there is slight skewness in the data. As mentioned earlier there are 187 records in the dataset, but after dropping the null values the available records are 185 .

Let’s explore the data more by plotting both Yearly and Monthly Box Plots:



Figure 48:Yearly Box Plot of Rose Wine dataset

From the above Box Plot we can see that the highest sale observed was in 1981 and lowest in 1994, considering 1995 data is only till July and sales have already reached to the level of 1994 . Similarly the highest and the lowest variation seem to be in 1981 and 1994 respectively.

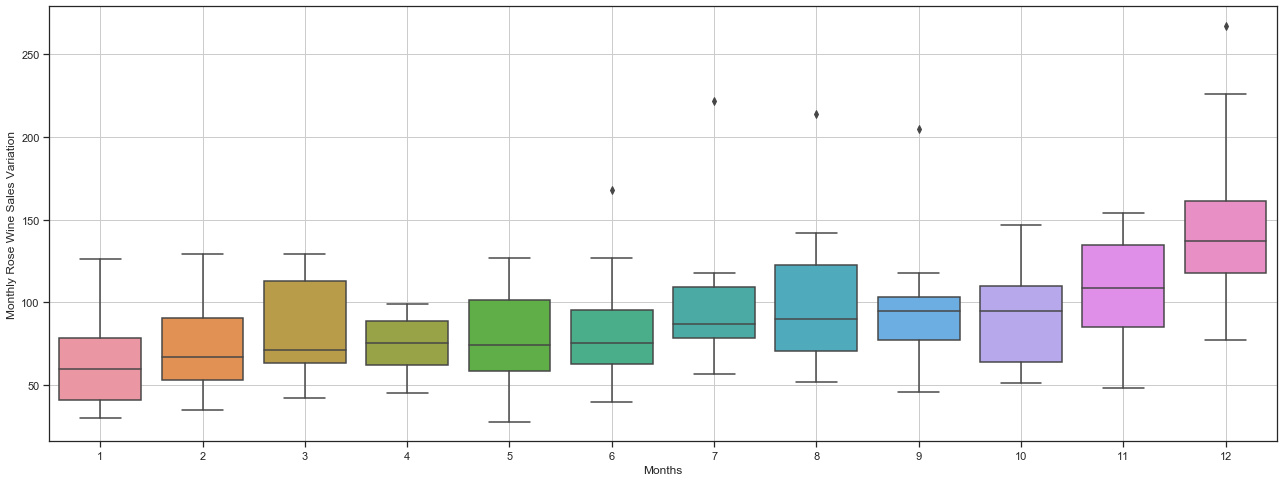


Figure 49:Monthly Box plot of Rose Wine Data

We can observe from the Monthly Box Plot that there is a slight seasonality element visible in the data. The sales for Rose wine seem to pick up from January month till the end of first quarter and is less consistent and stagnant through the second quarter onwards till fourth quarter. Similarly there is again an upward trend in sales through the last quarter .

Lets go ahead and see the table of monthly sales of wine across years.



Table 18:Pivot Table of Monthly Sales across years

Lets us now visually see the same table to understand the sales performance better :

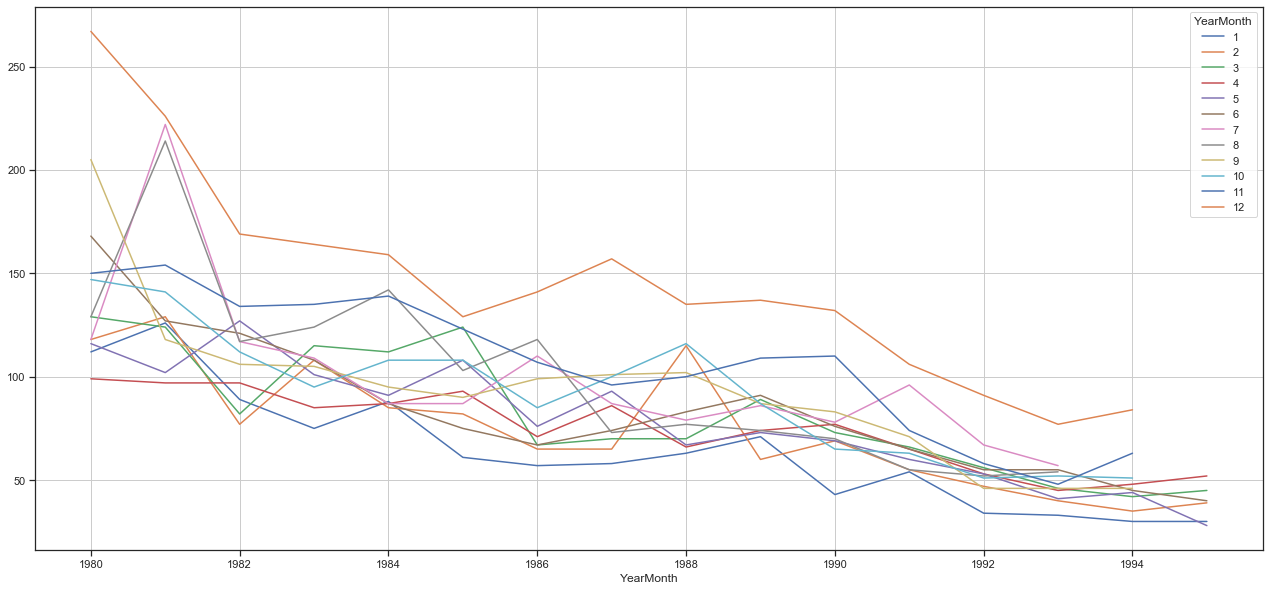


Figure 50:Monthly Sales across year Plot

The graph again talks about the seasonality element which is present as we can see December seems to be the month that drives the highest sales for Rose Wine, with the second highest being in November and so on .

Lets go ahead and plot a time series month plot to understand the spread accidents across different years and within different months across years .

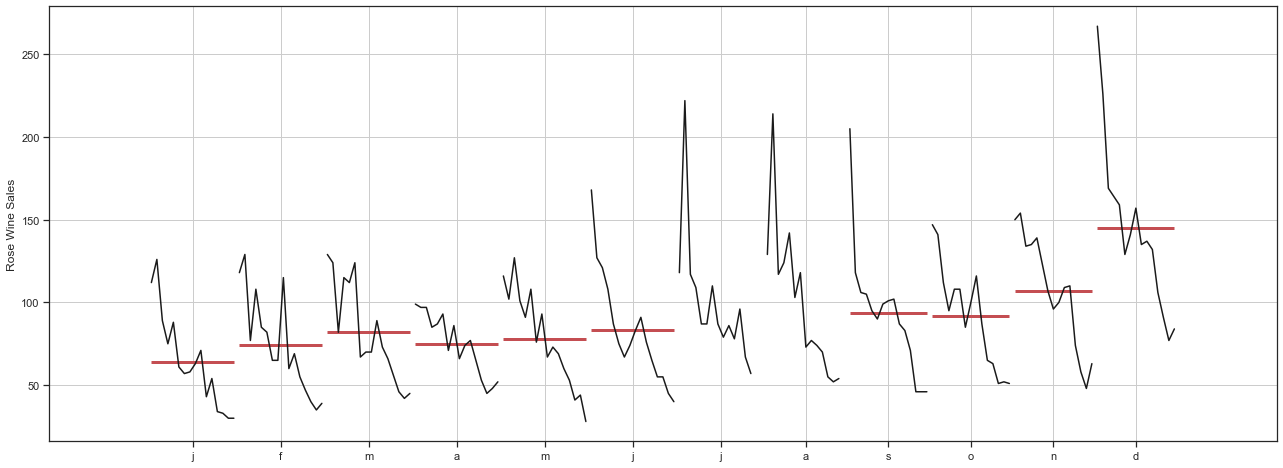


Figure 51:Time Series Month Plot

This plot shows us the behaviour of time series (Rose wine Sales) across various months. The red line is the median Value.

The below figure shows the Emprical Cumulative Dsitribution:

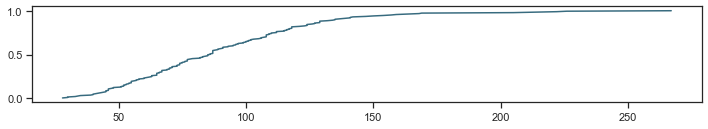


Figure 52:Empirical Cumulative Distribution Plot

This particular graph tells us what percentage of data points refer to what number of Sales .

Lets now plot the average retails sales per month and month on month percentage change of Rose wine sales :



Figure 53:average retails sales per month and month on month percentage change of Rose sales.

The above two graphs tells us the Average 'Sparkling Sales' and the Percentage change of 'Rose sales’ with respect to the time.

Now we will move to perform the Decomposition:

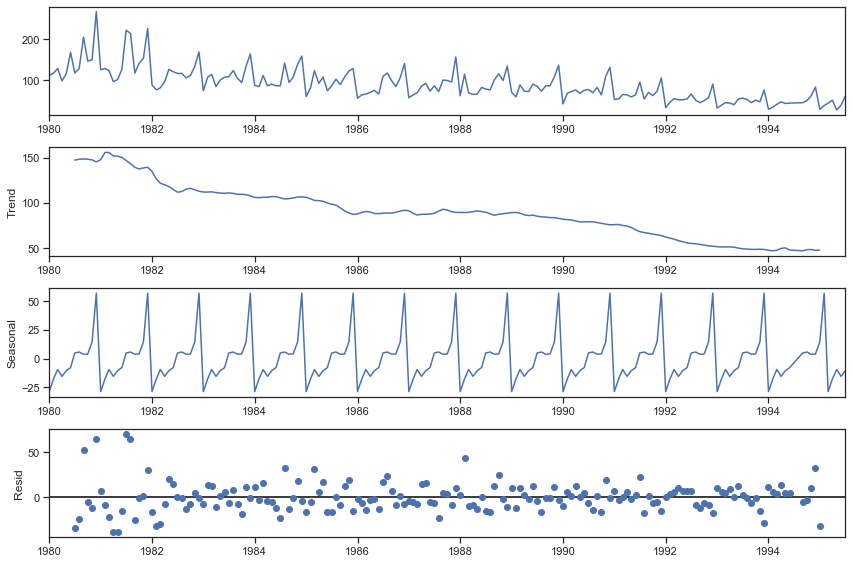


Figure 54: Additive Model of Decomposition

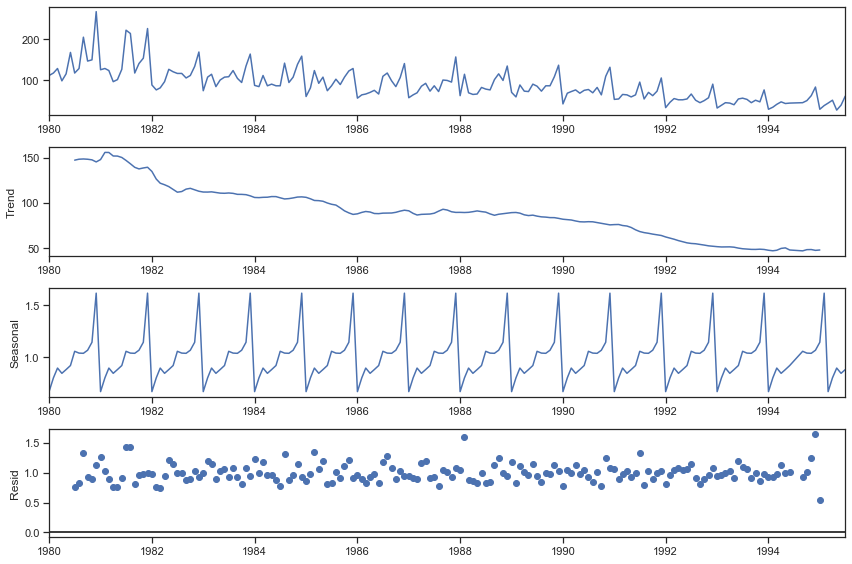


Figure 55:Multiplicative Model of Decomposition

From the above figure we can clearly understand there is an downward trend which indicate that the sales are not performing well. They have a factor of seasonality present in the data. The residual in the additive model shows a random pattern hence we can choose to go ahead with additive model.

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

Now we go ahead and split the time series datasets into Train and Test dataset. As given in the question that the Test data should start in 1991, so I have used 71% of the datasets for the Training Set and the rest of the dataset for Test datasets.

Below table shows the head and tail of training and test dataset .

First Few rows of the Training Dataset

|  |  |
| --- | --- |
| **YearMonth** | **Rose** |
| **1/1/1980** | 112 |
| **2/1/1980** | 118 |
| **3/1/1980** | 129 |
| **4/1/1980** | 99 |
| **5/1/1980** | 116 |

Last few rows of the Training Dataset

|  |  |
| --- | --- |
| **YearMonth** | **Rose** |
| **8/1/1990** | 70 |
| **9/1/1990** | 83 |
| **10/1/1990** | 65 |
| **11/1/1990** | 110 |
| **12/1/1990** | 132 |

First few rows of the Test Dataset

|  |  |
| --- | --- |
| **YearMonth** | **Rose** |
| **1/1/1991** | 54 |
| **2/1/1991** | 55 |
| **3/1/1991** | 66 |
| **4/1/1991** | 65 |
| **5/1/1991** | 60 |

Last few rows of the Test Dataset

|  |  |
| --- | --- |
| **YearMonth** | **Rose** |
| **3/1/1995** | 45 |
| **4/1/1995** | 52 |
| **5/1/1995** | 28 |
| **6/1/1995** | 40 |
| **7/1/1995** | 62 |

There are a total of 132 records in train data and 53 records in test dataset .

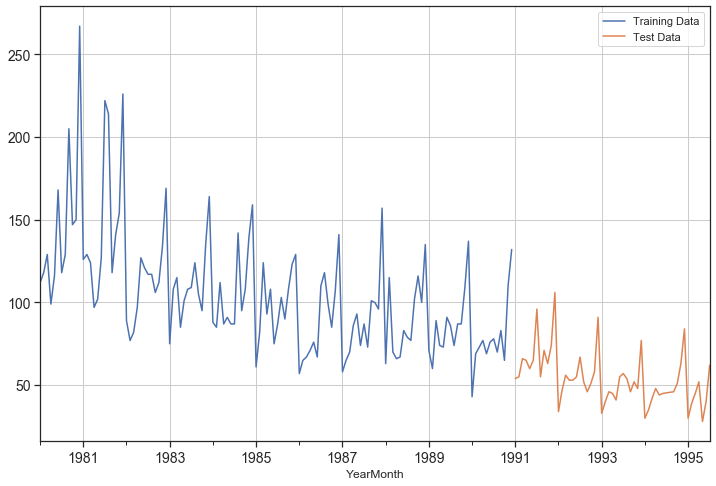


Figure 56:Plot for training and test dataframes

The above figure shows the train dataset from 1981 to 1990 in blue color and Test dataset from 1991 onwards in orange color.

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

Lets us build various time series forecasting models to check the performance using RMSE scores .

Linear Regression model: The below graph is the extract of the result from Linear Regression model :

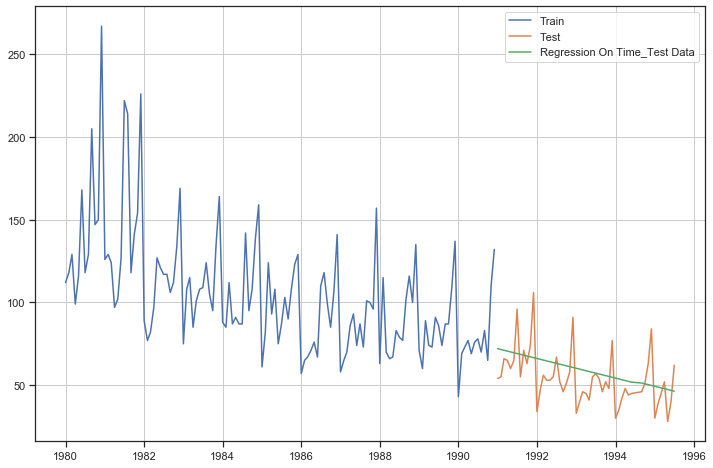


Figure 57:Linear Regression Outcome of Rose wine Time series

We can see from the plot that Rose wine sales shows an downward trend. The RMSE score for this model is 15.507 .

The summarized performance of the model run are :

|  |  |
| --- | --- |
|  | Test RMSE |
| RegressionOnTime | 15.507 |

Table 19:Summarised performance of models

Naive Forecast model:

The below graph is the extract of the result from Naïve Forecast model :

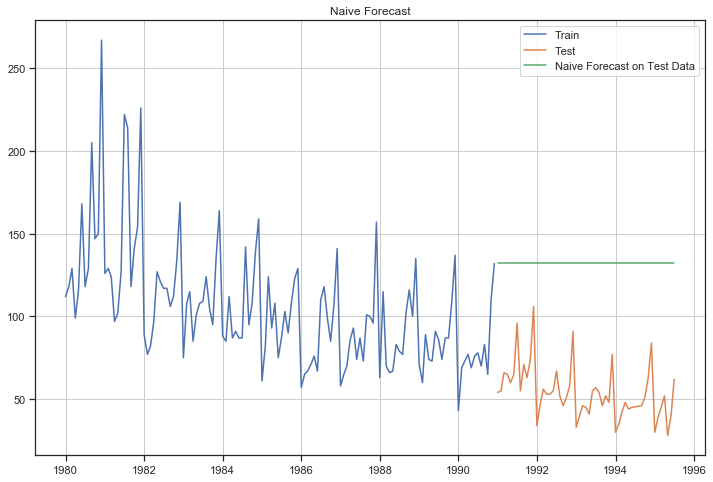


Figure 58:Naive Forecast outcome of Rose WIne Time Series

The RSMSE score of Naïve model is 3864.279, the naïve model is not suitable for the wine dataset since the forecast depends on the previous last observation .

The summarized performance of the models run are :

|  |  |
| --- | --- |
|  | Test RMSE |
| RegressionOnTime | 15.5067 |
| NaiveModel | 79.4515 |

Table 20:Summarised performance of models

Simple Average

The below graph is the extract of the result from Simple Average Forecast model

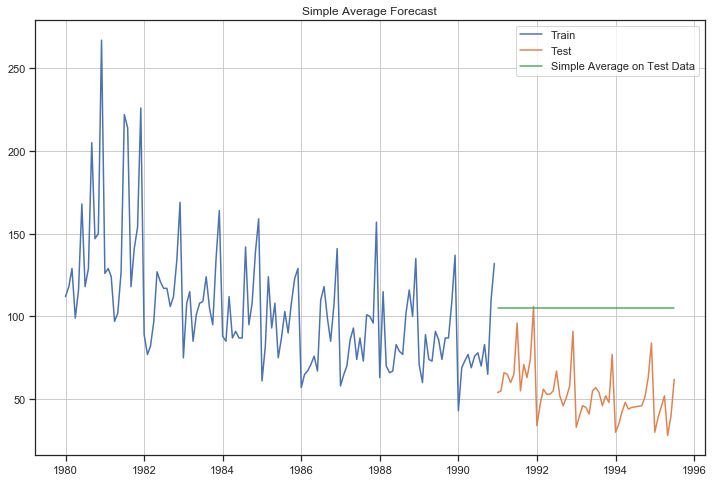


Figure 59:Simple Average Forecast Outcome of Rose wine Time series

The RMSE score of Simple Average Forecast model is 53.222

The summarized performance of the model are:

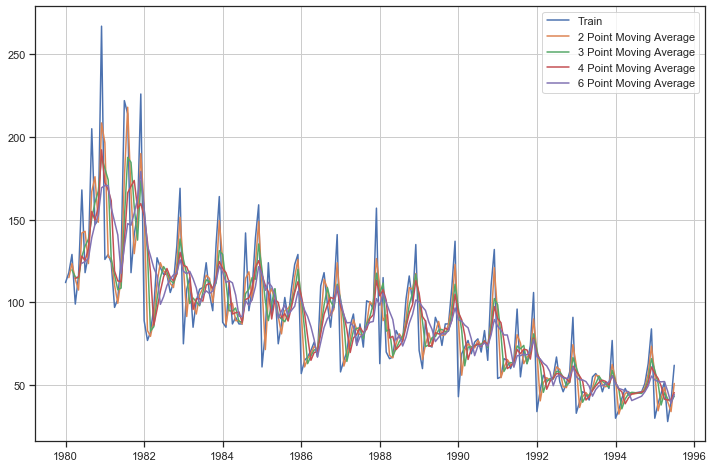
|  |  |
| --- | --- |
|  | Test RMSE |
| RegressionOnTime | 15.5067 |
| NaiveModel | 79.4515 |
| SimpleAverageModel | 53.2217 |

Table 21:Summarized performance of models

As we can see the Linear Regression Model is the best performing model from all the models run till now.

Moving Average

The below graph is the extract of the result after running Moving Average Forecast model



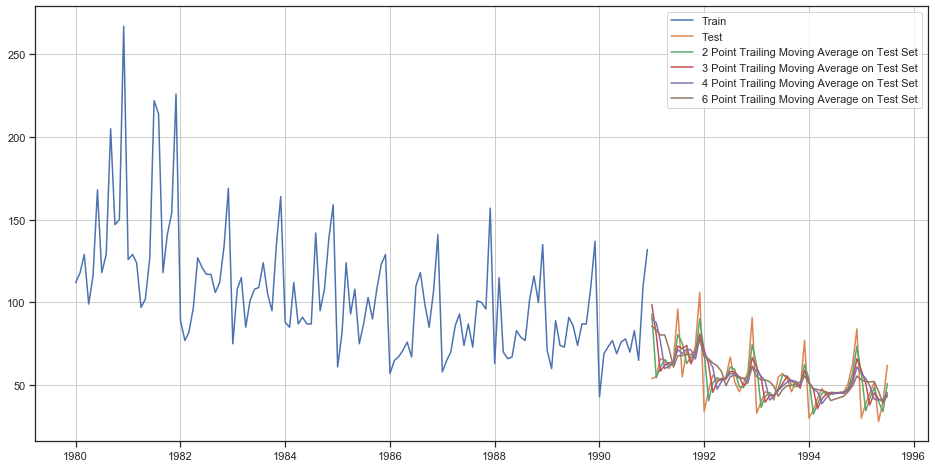


Figure 60:Moving Average Forecast outcome on Rose wine Time series

For 2 point Moving Average Model forecast on the Training Data, RMSE is 11.745

For 3 point Moving Average Model forecast on the Training Data, RMSE is 14.392

For 4 point Moving Average Model forecast on the Training Data, RMSE is 14.726

For 6 point Moving Average Model forecast on the Training Data, RMSE is 14.847

The summarized performance of the model are:

|  |  |
| --- | --- |
|  | **Test RMSE** |
| RegressionOnTime | 15.5067 |
| NaiveModel | 79.4515 |
| SimpleAverageModel | 53.2217 |
| 2pointTrailingMovingAverage | 11.745 |
| 3pointTrailingMovingAverage | 14.392 |
| 4pointTrailingMovingAverage | 14.726 |
| 6pointTrailingMovingAverage | 14.847 |

Table 22:Summarized performance of the model

From the above table we can see 2 point Trailing Moving Average is the best performing model till now.

**Simple Exponential Smoothing** : After running the SES model we can see the performance as below,

{'smoothing\_level': 0.09874976263905368,

'smoothing\_slope': nan,

'smoothing\_seasonal': nan,

'damping\_slope': nan,

'initial\_level': 134.38751258560546,

'initial\_slope': nan,

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

'use\_boxcox': False,

'lamda': None,

'remove\_bias': False}

Figure 61:SES parameters for Rose wine Forecast

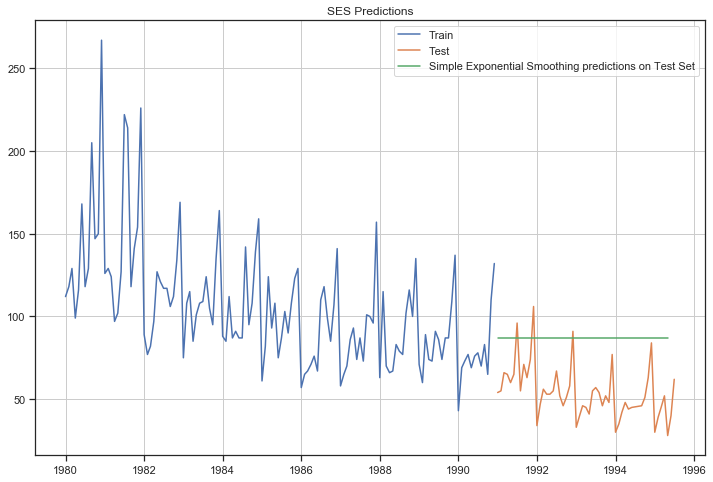


Figure 62:Simple Exponential smoothing outcome on Rose wine Timeseries

The RMSE for the SES model is 36.602394

The summarized performance of the model are:

|  |  |
| --- | --- |
|  | Test RMSE |
| RegressionOnTime | 15.5067 |
| NaiveModel | 79.4515 |
| SimpleAverageModel | 53.2217 |
| 2pointTrailingMovingAverage | 11.745 |
| 3pointTrailingMovingAverage | 14.392 |
| 4pointTrailingMovingAverage | 14.726 |
| 6pointTrailingMovingAverage | 14.847 |
| SimpleExponentialSmoothing | 36.602394 |

Table 23:Summarised performance of the models

As we can see above the best model till now is 2 point Trailing Moving Average.

Double Exponential Smoothing

After running the SES model we can see the performance as below,

{'smoothing\_level': 0.15789473684210525,

'smoothing\_slope': 0.15789473684210525,

'smoothing\_seasonal': nan,

'damping\_slope': nan,

'initial\_level': 112.0,

'initial\_slope': 6.0,

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

'use\_boxcox': False,

'lamda': None,

'remove\_bias': False}

Figure 63:DES parameters for Rose wine Forecast

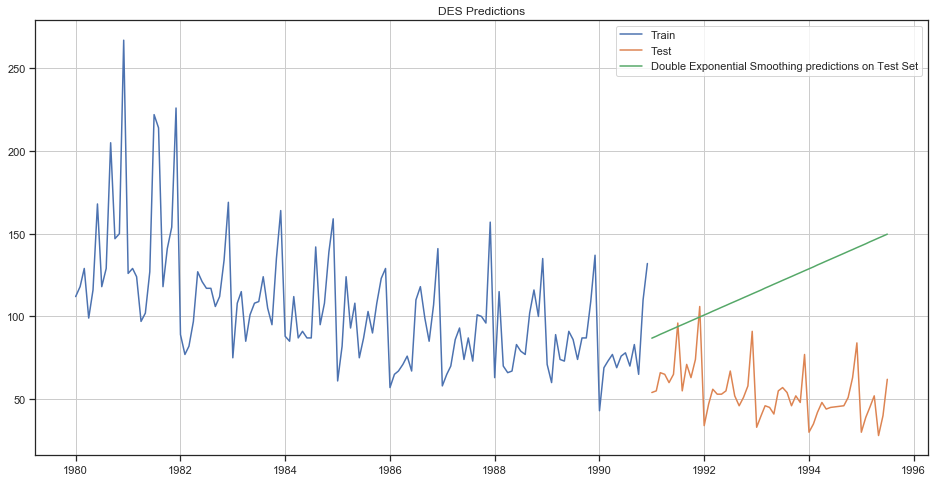


Figure 64:Double Exponential Smoothing Outcome on Rose Wine time series

The RMSE for this model is 69.03888

The summarized performance of the model are

|  |  |
| --- | --- |
|  | Test RMSE |
| RegressionOnTime | 15.5067 |
| NaiveModel | 79.4515 |
| SimpleAverageModel | 53.2217 |
| 2pointTrailingMovingAverage | 11.745 |
| 3pointTrailingMovingAverage | 14.392 |
| 4pointTrailingMovingAverage | 14.726 |
| 6pointTrailingMovingAverage | 14.847 |
| SimpleExponentialSmoothing | 36.602394 |
| DoubleExponentialSmoothing | 69.03888 |

Table 24:Summarised performance of the models

Triple Exponential Smoothing:

After running the TES model with Trend as Additive and Seasonality as Additive , we can see the parameters as below

{'smoothing\_level': 0.13346419155496156,

'smoothing\_slope': 0.013799003011050462,

'smoothing\_seasonal': 0.0,

'damping\_slope': nan,

'initial\_level': 77.88760840203805,

'initial\_slope': 0.0,

'initial\_seasons': array([ 37.22512664, 49.55914984, 57.49063909, 46.84824177,

55.60691281, 61.03952791, 70.96491019, 76.99009393,

73.00798464, 71.11386606, 89.18174761, 131.39710677]),

'use\_boxcox': False,

'lamda': None,

'remove\_bias': False}

Figure 65:TES Parameters for Rose wine sales forecast

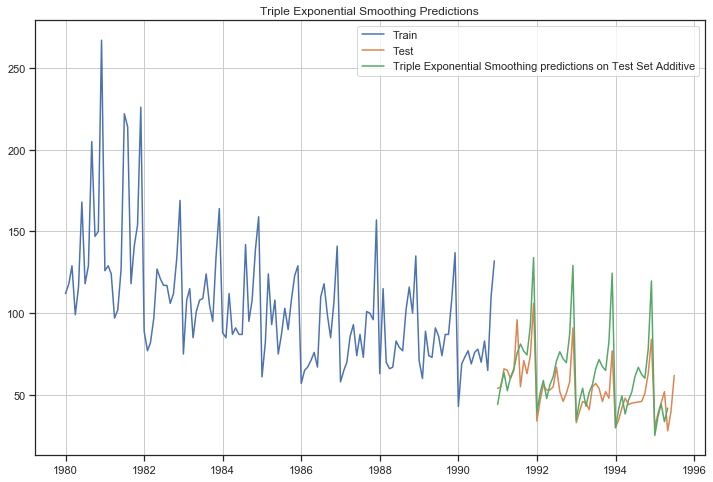


Figure 66:TES Add\_Add outcome on Rose Wine sales Time series

The RMSE for above TES Add\_Add model is 20.7717

After running the TES model with Trend as Additive and Seasonality as multiplicative , we can see the parameters as below

{'smoothing\_level': 0.1060963216313948,

'smoothing\_slope': 0.04843849811887955,

'smoothing\_seasonal': 0.0,

'damping\_slope': nan,

'initial\_level': 76.65565108581343,

'initial\_slope': 0.0,

'initial\_seasons': array([1.47550255, 1.65927136, 1.80572635, 1.58888816, 1.77822697,

1.92604355, 2.11649447, 2.25135196, 2.11690577, 2.08112826,

2.40927271, 3.30448121]),

'use\_boxcox': False,

'lamda': None,

'remove\_bias': False}

Figure 67:TES Parameters for Rose wine sales forecast

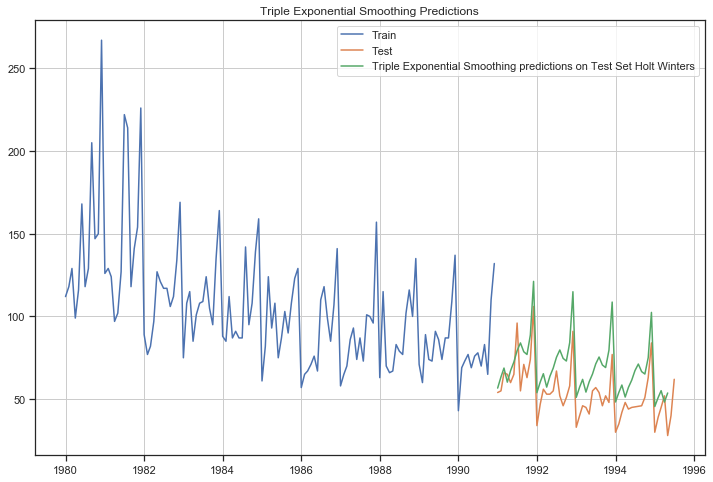


Figure 68:TES Add\_Mul outcome on Rose wine sales time series

The RMSE of this model is 19.9165

After running the TES model with Trend as Multiplicative and Seasonality as multiplicative , we can see the parameters as below

{'smoothing\_level': 0.0699773504936336,

'smoothing\_slope': 3.4486598091050817e-18,

'smoothing\_seasonal': 0.0,

'damping\_slope': nan,

'initial\_level': 76.65475181082856,

'initial\_slope': 0.9939059043685701,

'initial\_seasons': array([1.4524715 , 1.64401816, 1.79897871, 1.57420499, 1.76942144,

1.90969487, 2.10120743, 2.24519705, 2.1142589 , 2.07332592,

2.4162356 , 3.31040422]),

'use\_boxcox': False,

'lamda': None,

'remove\_bias': False}

Figure 69:TES parameters for Rose Wine sales forecast

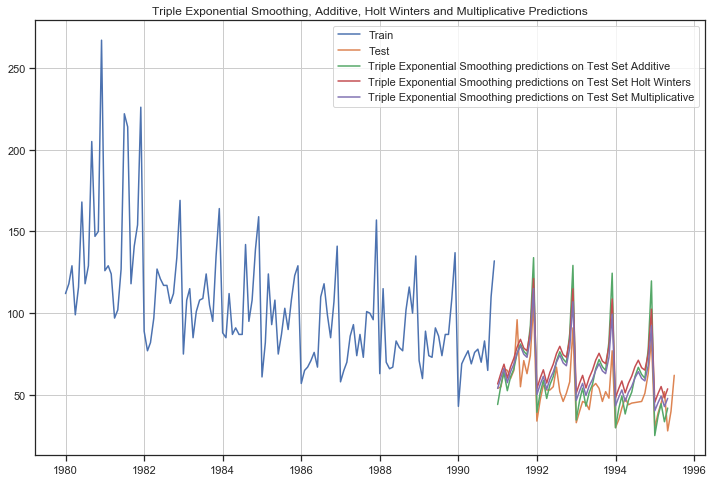


Figure 70:TES Mul\_Mul outcome of Rose wine sale time series

The RMSE of this model is 15.8578

The summarized performance of all the models built are as below :

|  |  |
| --- | --- |
|  | Test RMSE |
| RegressionOnTime | 15.50670843 |
| NaiveModel | 79.45153976 |
| NaiveModel | 79.45153976 |
| SimpleAverageModel | 53.22179459 |
| 2pointTrailingMovingAverage | 11.74493066 |
| 3pointTrailingMovingAverage | 14.39193811 |
| 4pointTrailingMovingAverage | 14.72631613 |
| 6pointTrailingMovingAverage | 14.84684213 |
| SES | 36.60239402 |
| DES | 69.03888163 |
| TES\_Add | 20.77173073 |
| TES\_Add\_Mul | 19.9165808 |
| TES\_Mul\_Mul | 15.85789014 |

From the above table we can see that out of all the models the Two point trailing moving average is the best performing model with lowest RMSE score of 11.7449

Now that we have run all the models , lets now see the summary of all the models on a single graph as below :

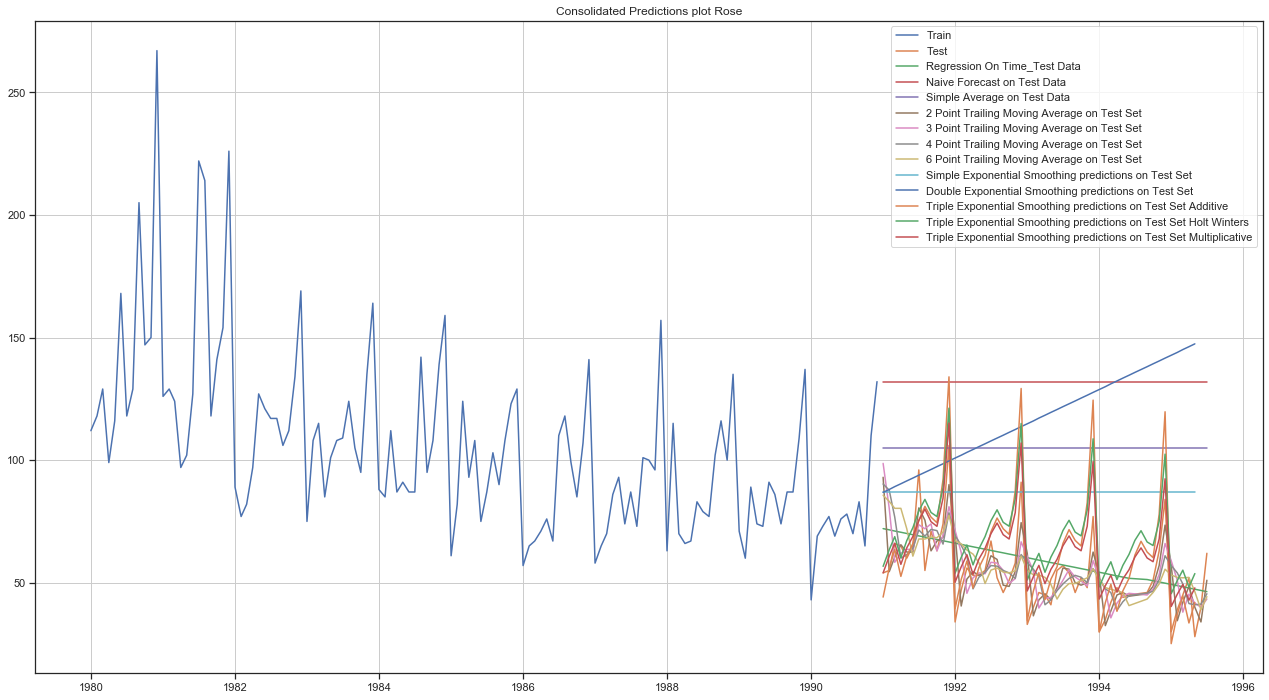


Figure 71:Consolidated prediction Plot of Rose Wine Sales Data

## 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.

To check whether the series is stationary , we use the Augmented Dickey Fuller (ADF) test. 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:

* 𝐻0 : The Time Series has a unit root and is thus non-stationary.
* 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 𝛼α value of 0.05 as mentioned in the question .

After performing the Stationarity check we get the below result :

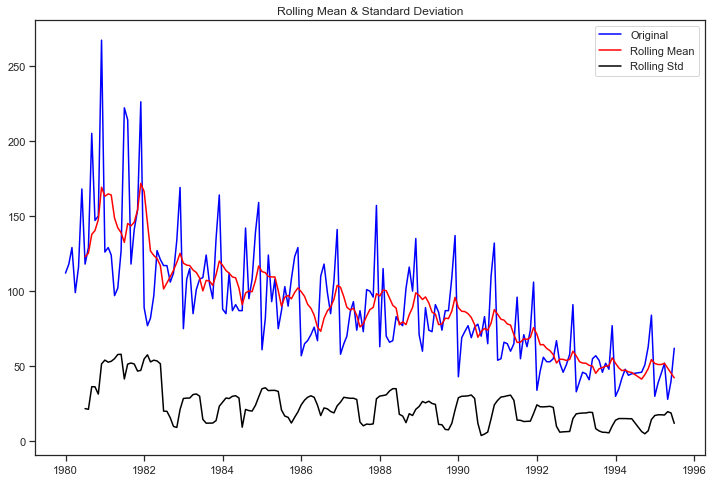


Figure 72:Stationarity Graph

Results of Dickey-Fuller Test:

Test Statistic -1.838033

p-value 0.361750

#Lags Used 13.000000

Number of Observations Used 171.000000

Critical Value (1%) -3.469181

Critical Value (5%) -2.878595

Critical Value (10%) -2.575863

Since the p value is greater than alpha 0.05, time series is not stationary. We can take the next levels of differencing to make the time series stationary . Let us take a difference of order 1 and check whether the Time Series is stationary or not.

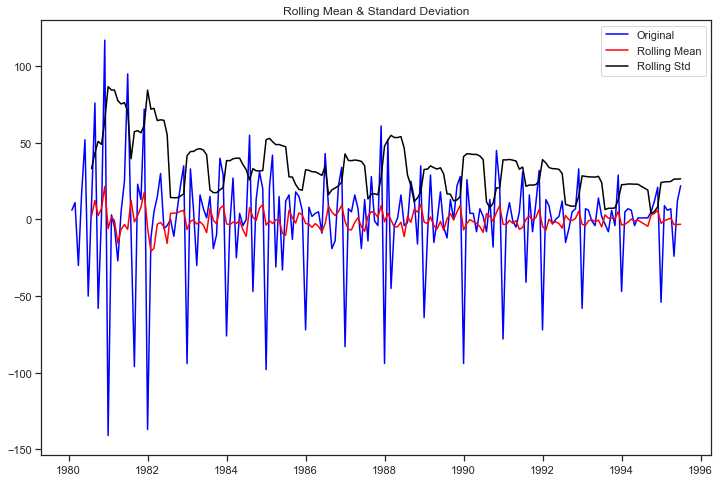


Figure 73:Stationarity Graph after differencing

Results of Dickey-Fuller Test:

Test Statistic -8.167161e+00

p-value 8.819858e-13

#Lags Used 1.200000e+01

Number of Observations Used 1.710000e+02

Critical Value (1%) -3.469181e+00

Critical Value (5%) -2.878595e+00

Critical Value (10%) -2.575863e+00

We can see that the alpha value is less than 0.05 and the Time Series now is indeed stationary .

## 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 MODEL:**

ARIMA model AIC values of the parameter combinations:

ARIMA(0, 1, 0) - AIC:1333.1546729124348

ARIMA(0, 1, 1) - AIC:1282.3098319748315

ARIMA(0, 1, 2) - AIC:1279.671528853576

ARIMA(1, 1, 0) - AIC:1317.3503105381535

ARIMA(1, 1, 1) - AIC:1280.5742295380041

ARIMA(1, 1, 2) - AIC:1279.8707234231897

ARIMA(2, 1, 0) - AIC:1298.6110341604985

ARIMA(2, 1, 1) - AIC:1281.507862186858

ARIMA(2, 1, 2) - AIC:1281.8707222264297

We can see above the lowest AIC is 1279.671 with (0,1,2) param .

Lets go ahead and evaluate the model

SARIMAX Results

==============================================================================

Dep. Variable: Rose No. Observations: 132

Model: ARIMA(0, 1, 2) Log Likelihood -636.836

Date: Sun, 23 Jan 2022 AIC 1279.672

Time: 18:15:51 BIC 1288.297

Sample: 01-01-1980 HQIC 1283.176

- 12-01-1990

Covariance Type: opg

==============================================================================

coef std err z P>|z| [0.025 0.975]

------------------------------------------------------------------------------

ma.L1 -0.6970 0.072 -9.689 0.000 -0.838 -0.556

ma.L2 -0.2042 0.073 -2.794 0.005 -0.347 -0.061

sigma2 965.8407 88.305 10.938 0.000 792.766 1138.915

===================================================================================

Ljung-Box (Q): 112.54 Jarque-Bera (JB): 39.24

Prob(Q): 0.00 Prob(JB): 0.00

Heteroskedasticity (H): 0.36 Skew: 0.82

Prob(H) (two-sided): 0.00 Kurtosis: 5.13

===================================================================================

Figure 74:Statistical Analysis of ARIMA model

The RMSE and MAPE values are 37.1179 & 76.2155

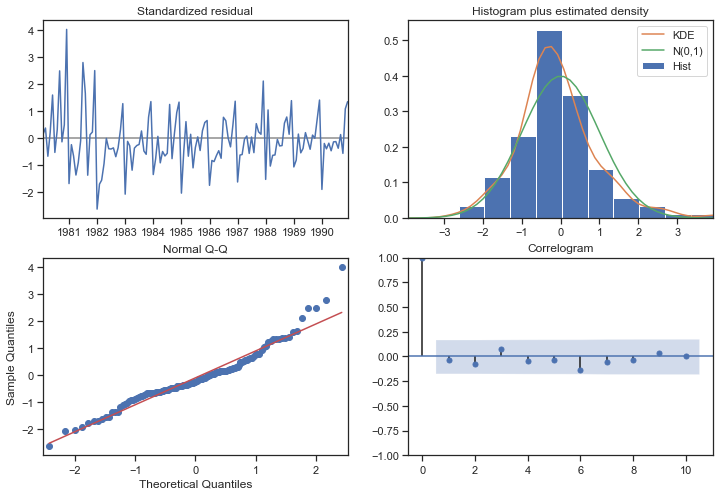


Figure 75:Diagnosis of ARIMA Model

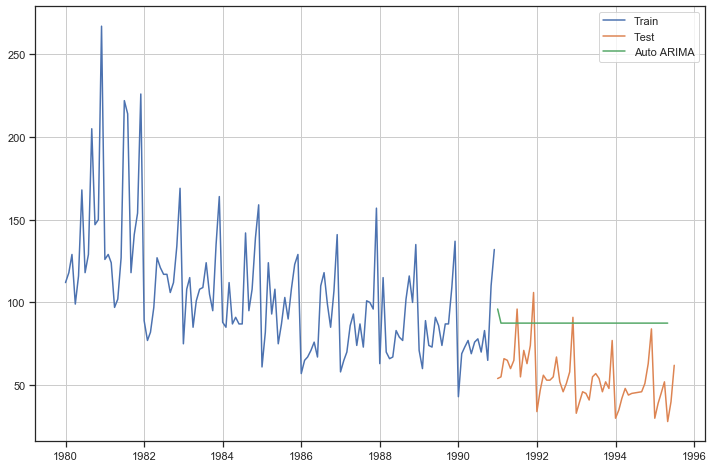


Figure 76:Auto ARIMA Plot of Test and forecasted values

**SARIMA MODEL :** SARIMA model AIC values of the parameter combinations in descending order of the first records are as below :

|  |  |  |  |
| --- | --- | --- | --- |
|  | param | seasonal | AIC |
| 26 | (0, 1, 2) | (2, 1, 2, 12) | 774.9691 |
| 53 | (1, 1, 2) | (2, 1, 2, 12) | 776.9401 |
| 80 | (2, 1, 2) | (2, 1, 2, 12) | 776.9961 |
| 17 | (0, 1, 1) | (2, 1, 2, 12) | 782.1539 |
| 79 | (2, 1, 2) | (2, 1, 1, 12) | 783.7037 |

Table 25:AIC Values of the first 5 params of SARIMA model in descending order

We can see the lowest AIC value is 774.9691 with param (0,1,2) and seasonal (2,1,2,12) .

Lets go ahead and evaluate the model using statistical values :

SARIMAX Results

==========================================================================================

Dep. Variable: Rose No. Observations: 132

Model: SARIMAX(0, 1, 2)x(2, 1, 2, 12) Log Likelihood -380.485

Date: Sun, 23 Jan 2022 AIC 774.969

Time: 18:20:47 BIC 792.622

Sample: 01-01-1980 HQIC 782.094

- 12-01-1990

Covariance Type: opg

==============================================================================

coef std err z P>|z| [0.025 0.975]

------------------------------------------------------------------------------

ma.L1 -0.9524 0.184 -5.166 0.000 -1.314 -0.591

ma.L2 -0.0764 0.126 -0.605 0.545 -0.324 0.171

ar.S.L12 0.0480 0.177 0.271 0.786 -0.299 0.395

ar.S.L24 -0.0419 0.028 -1.513 0.130 -0.096 0.012

ma.S.L12 -0.7525 0.301 -2.503 0.012 -1.342 -0.163

ma.S.L24 -0.0721 0.204 -0.354 0.723 -0.472 0.327

sigma2 187.8667 45.275 4.149 0.000 99.130 276.603

===================================================================================

Ljung-Box (Q): 31.31 Jarque-Bera (JB): 4.86

Prob(Q): 0.84 Prob(JB): 0.09

Heteroskedasticity (H): 0.91 Skew: 0.41

Prob(H) (two-sided): 0.79 Kurtosis: 3.77

===================================================================================

Figure 77:Statistical Analysis of SARIMA Model

The RMSE and MAPE value are 20.5049 & 25.2693

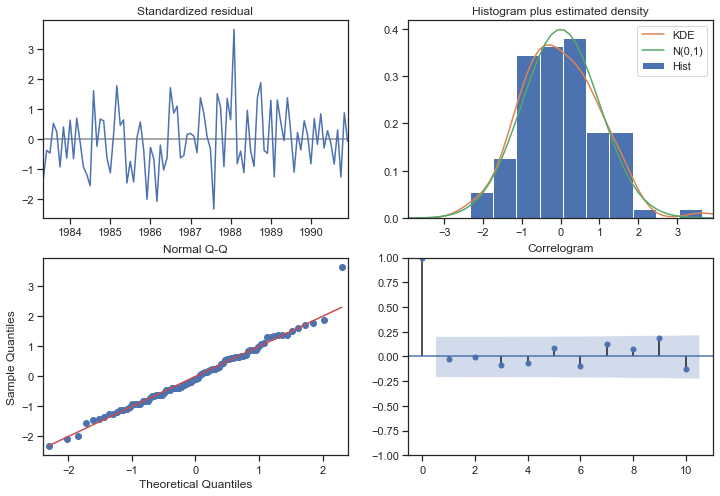


Figure 78:Diagnosis of SARIMA Model

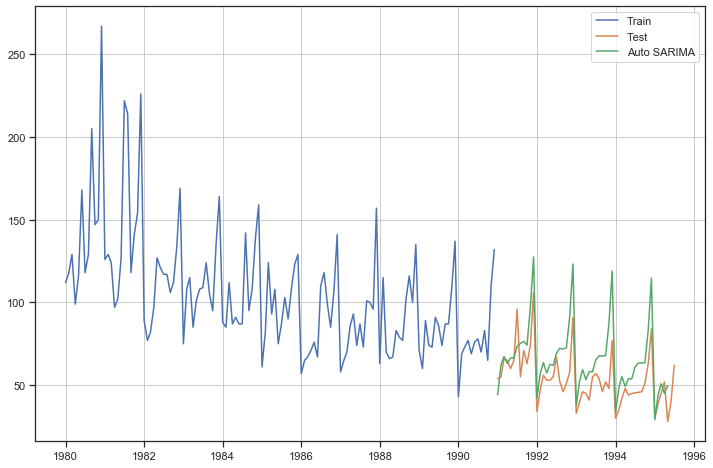


Figure 79:Auto SARIMA plot

The summarized model of all the models built are as below :

|  |  |
| --- | --- |
|  | Test RMSE |
| RegressionOnTime | 15.506708 |
| NaiveModel | 79.45154 |
| NaiveModel | 79.45154 |
| SimpleAverageModel | 53.221795 |
| 2pointTrailingMovingAverage | 11.744931 |
| 3pointTrailingMovingAverage | 14.391938 |
| 4pointTrailingMovingAverage | 14.726316 |
| 6pointTrailingMovingAverage | 14.846842 |
| SES | 36.602394 |
| DES | 69.038882 |
| TES\_Add | 20.771731 |
| TES\_Add\_Mul | 19.916581 |
| TES\_Mul\_Mul | 15.85789 |
| TES\_Mul\_Mul | 15.85789 |
| AIC\_ARIMA(0,1,2) | 37.117942 |
| Manual\_ARIMA(0,1,0) | 79.45154 |
| AIC\_SARIMA(0,1,2)(2,1,2,12) | 20.504966 |
| Manual\_SARIMA(4,1,2)(0,1,2,12) | 19.926729 |

Table 26:Summarized performance of all the models

From the above summary of the performance of models we can see that the Two point Trailing Moving Average models is still the best performing model till now .

## 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.

**ARIMA model based on the cut off points**

Lets first have a look at the ACF and PACF Plot:

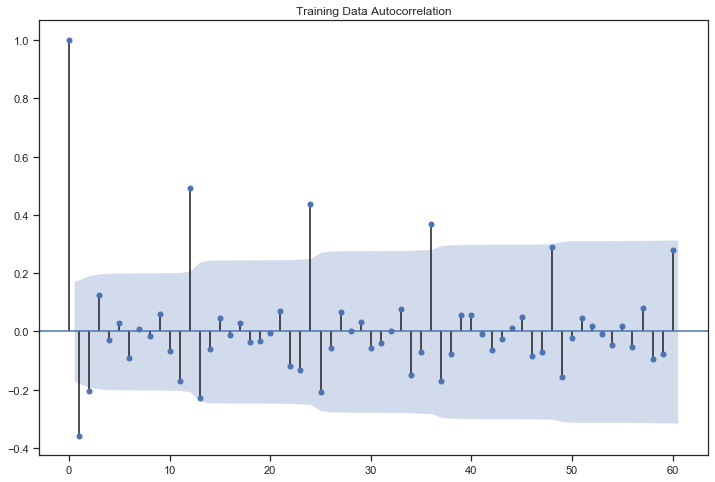


Figure 80:ACF Plot of ARIMA Model

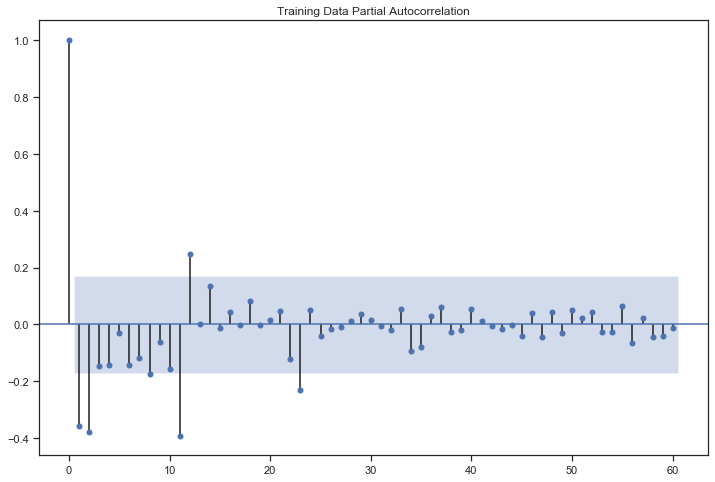


Figure 81:PACF Plot of ARIMA Model

Here, we have taken alpha=0.05.

* 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.

Lets go ahead and evaluate the model using statistical values by taking the parameters as (2,1,2):

SARIMAX Results

==============================================================================

Dep. Variable: Rose No. Observations: 132

Model: ARIMA(2, 1, 2) Log Likelihood -635.935

Date: Sun, 23 Jan 2022 AIC 1281.871

Time: 19:18:12 BIC 1296.247

Sample: 01-01-1980 HQIC 1287.712

- 12-01-1990

Covariance Type: opg

==============================================================================

coef std err z P>|z| [0.025 0.975]

------------------------------------------------------------------------------

ar.L1 -0.4540 0.469 -0.969 0.333 -1.372 0.464

ar.L2 0.0001 0.170 0.001 0.999 -0.334 0.334

ma.L1 -0.2541 0.459 -0.554 0.580 -1.154 0.646

ma.L2 -0.5984 0.430 -1.390 0.164 -1.442 0.245

sigma2 952.1601 91.424 10.415 0.000 772.973 1131.347

===================================================================================

Ljung-Box (Q): 112.12 Jarque-Bera (JB): 34.16

Prob(Q): 0.00 Prob(JB): 0.00

Heteroskedasticity (H): 0.37 Skew: 0.79

Prob(H) (two-sided): 0.00 Kurtosis: 4.94

===================================================================================

Figure 82:Statistical Analysis of Manual ARIMA Model

Below is the diagnosis of Manual ARIMA plot:

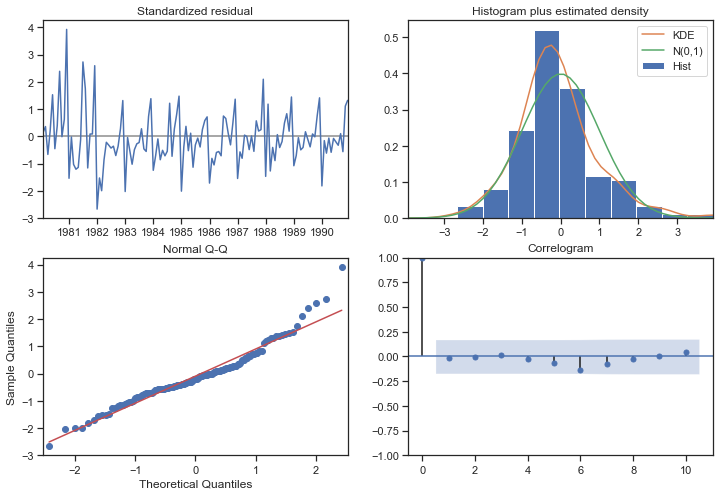


Figure 83:Diagnosis of Manual ARIMA model

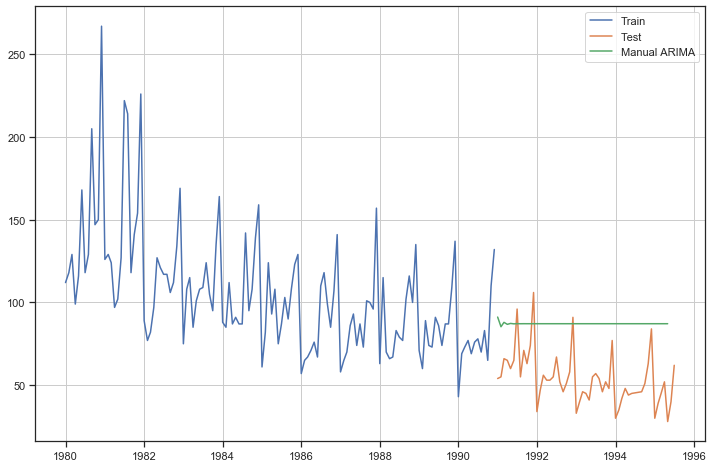


Figure 84:Manual ARIMA Predictions Plot

The RSMSE AND MAPE values of Manual ARIMA model are 36.6792 & 75.3273

**SARIMA model based on the cut off points**

Lets first have a look at the ACF and PACF Plot:

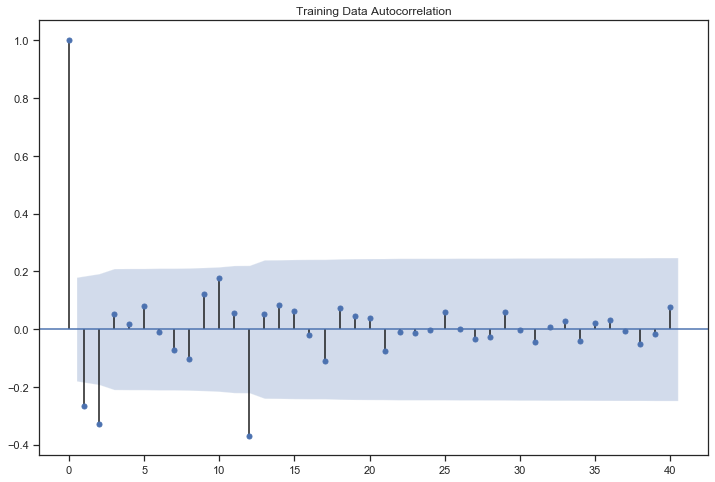


Figure 85:ACF plot of SARIMA Model

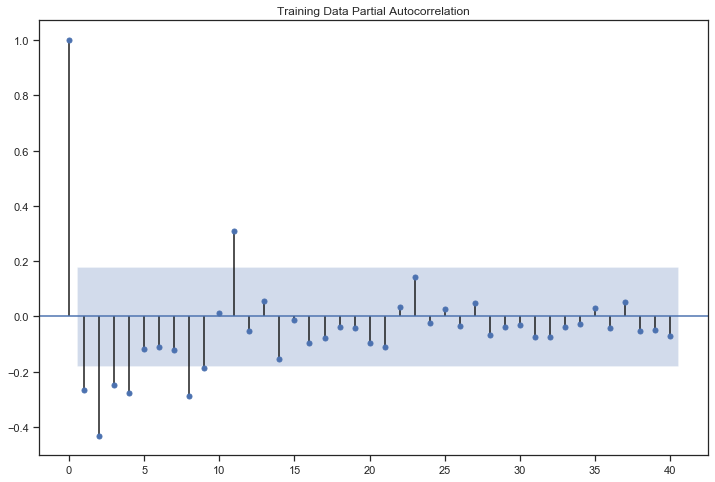


Figure 86:PACF Plot of SARIMA Model

Here, we have taken alpha=0.05.

We are going to take the seasonal period as 12 or its multiple e.g. 24

* The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off to 4.
* 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 4 and 2 respectively.

Lets go ahead and evaluate the model using statistical values by taking the parameters as (4,1,2):

SARIMAX Results

==========================================================================================

Dep. Variable: Rose No. Observations: 132

Model: SARIMAX(4, 1, 2)x(0, 1, 2, 12) Log Likelihood -384.369

Date: Sun, 23 Jan 2022 AIC 786.737

Time: 18:23:14 BIC 809.433

Sample: 01-01-1980 HQIC 795.898

- 12-01-1990

Covariance Type: opg

==============================================================================

coef std err z P>|z| [0.025 0.975]

------------------------------------------------------------------------------

ar.L1 -0.8967 0.132 -6.814 0.000 -1.155 -0.639

ar.L2 0.0165 0.171 0.097 0.923 -0.319 0.352

ar.L3 -0.1132 0.174 -0.650 0.515 -0.454 0.228

ar.L4 -0.1598 0.116 -1.380 0.168 -0.387 0.067

ma.L1 0.1508 0.174 0.866 0.387 -0.191 0.492

ma.L2 -0.8492 0.164 -5.166 0.000 -1.171 -0.527

ma.S.L12 -0.3907 0.102 -3.848 0.000 -0.590 -0.192

ma.S.L24 -0.0887 0.091 -0.977 0.329 -0.267 0.089

sigma2 238.9649 0.001 2.02e+05 0.000 238.963 238.967

===================================================================================

Ljung-Box (Q): 27.59 Jarque-Bera (JB): 0.01

Prob(Q): 0.93 Prob(JB): 0.99

Heteroskedasticity (H): 0.76 Skew: -0.01

Prob(H) (two-sided): 0.46 Kurtosis: 3.06

Figure 87:Statistical Analysis of Manual SARIMA Model

Below is the Diagnosis of Manual SARIMA plot

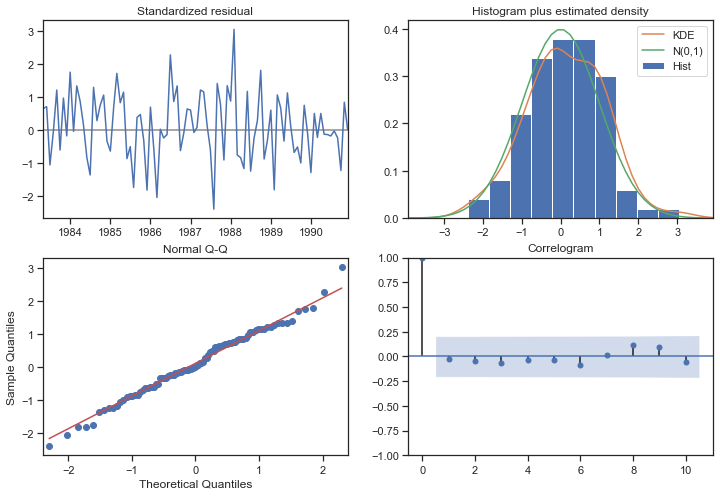


Figure 88:Diagnosis of manual SARIMA plot

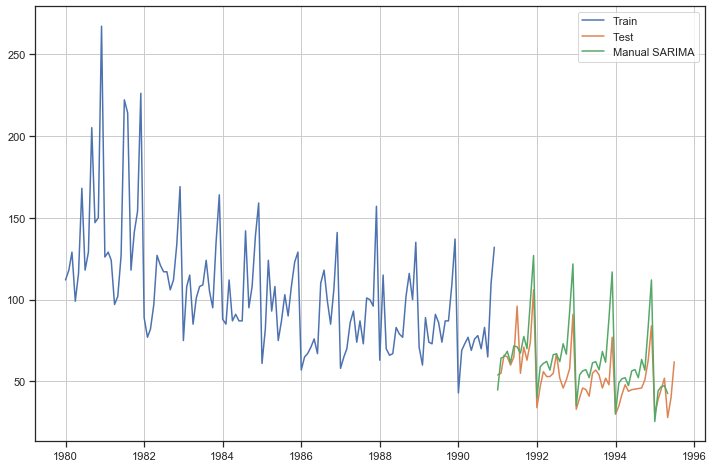


Figure 89:Manual SARIMA predictions plot

The RMSE and MAPE values for this model is 19.9267 & 22.3583

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

The summarized performances of the all the models built are as below :

|  |  |
| --- | --- |
|  | Test RMSE |
| RegressionOnTime | 15.506708 |
| NaiveModel | 79.45154 |
| NaiveModel | 79.45154 |
| SimpleAverageModel | 53.221795 |
| 2pointTrailingMovingAverage | 11.744931 |
| 3pointTrailingMovingAverage | 14.391938 |
| 4pointTrailingMovingAverage | 14.726316 |
| 6pointTrailingMovingAverage | 14.846842 |
| SES | 36.602394 |
| DES | 69.038882 |
| TES\_Add | 20.771731 |
| TES\_Add\_Mul | 19.916581 |
| TES\_Mul\_Mul | 15.85789 |
| AIC\_ARIMA(0,1,2) | 37.117942 |
| Manual\_ARIMA(0,1,0) | 36.67929 |
| AIC\_SARIMA(0,1,2)(2,1,2,12) | 20.504966 |
| Manual\_SARIMA(4,1,2)(0,1,2,12) | 19.926729 |

Figure 90:Summarized performance of all the models

From all the above models built we can see that the lowest RMSE value is for 2 point Trailing Moving Average model of 11.7449 hence concluding this as the best among all the models built .

## 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.

SARIMA model with parameters (4,1,2) x (1,0,2,12) is proposed to use for forecast of next 12 months using full data. Details of the model are as under :

SARIMAX Results

==========================================================================================

Dep. Variable: Rose No. Observations: 185

Model: SARIMAX(4, 1, 2)x(1, 0, 2, 12) Log Likelihood -655.578

Date: Sun, 23 Jan 2022 AIC 1331.156

Time: 20:02:16 BIC 1361.719

Sample: 0 HQIC 1343.569

- 185

Covariance Type: opg

==============================================================================

coef std err z P>|z| [0.025 0.975]

------------------------------------------------------------------------------

ar.L1 0.3793 0.530 0.716 0.474 -0.660 1.418

ar.L2 -0.1996 0.113 -1.765 0.077 -0.421 0.022

ar.L3 -0.0516 0.129 -0.401 0.688 -0.303 0.200

ar.L4 -0.0109 0.121 -0.090 0.928 -0.248 0.226

ma.L1 -1.1680 0.533 -2.190 0.029 -2.214 -0.122

ma.L2 0.2527 0.476 0.531 0.595 -0.679 1.185

ar.S.L12 0.4447 0.063 7.084 0.000 0.322 0.568

ma.S.L12 0.1309 0.089 1.464 0.143 -0.044 0.306

ma.S.L24 0.2711 0.108 2.513 0.012 0.060 0.482

sigma2 243.1937 25.836 9.413 0.000 192.556 293.831

===================================================================================

Ljung-Box (Q): 30.36 Jarque-Bera (JB): 5.64

Prob(Q): 0.86 Prob(JB): 0.06

Heteroskedasticity (H): 0.40 Skew: 0.30

Prob(H) (two-sided): 0.00 Kurtosis: 3.71

===================================================================================

Figure 91:Model Performance using full data

The RMSE for the full Model is 31.56823

The predicted sale of Rose wine for next 12 months is as below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | mean | mean\_se | mean\_ci\_lower | mean\_ci\_upper |
| 8/1/1995 | 52.2935 | 15.594668 | 21.72851447 | 82.85848865 |
| 9/1/1995 | 45.12412 | 15.938882 | 13.88448755 | 76.36375708 |
| 10/1/1995 | 49.97538 | 15.948158 | 18.71756106 | 81.23319361 |
| 11/1/1995 | 57.4091 | 15.951966 | 26.14382334 | 88.67438081 |
| 12/1/1995 | 69.76605 | 15.98062 | 38.44461413 | 101.0874917 |
| 1/1/1996 | 39.27217 | 16.075546 | 7.764684558 | 70.77966531 |
| 2/1/1996 | 43.95084 | 16.177615 | 12.24330222 | 75.6583862 |
| 3/1/1996 | 47.0288 | 16.257949 | 15.16380918 | 78.89379987 |
| 4/1/1996 | 50.65175 | 16.324224 | 18.65685669 | 82.64663868 |
| 5/1/1996 | 36.95568 | 16.387821 | 4.836144556 | 69.07522318 |
| 6/1/1996 | 45.5642 | 16.453667 | 13.31560149 | 77.81278926 |
| 7/1/1996 | 57.93478 | 16.521561 | 25.55311168 | 90.31644126 |

Table 27:Predicted Sale of Rose

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

Trend in sale of Rose is continuously decreasing over the period. Detailed Study may be required to see whether decreasing trend is due to change in customer preference or due to substitution. Seasonality of sales is observed , and higher sales is maintained in the end of the year. Some promotion schemes and improvement/ quality enhancers in the product can be examined so as to attract new young generation customers .

-----------------------------------------------------xxxxxxxxxxxxxxxxxxxxxxxxxxx-----------------------------------------------