TIME SERIES FORECASTING PROJECT

BUSINESS REPORT

BY SUDEEP KUMAR DAS

PGP-DSBA

INDEX:

List Of Figures:

List of Tables:

ABOUT THE DATA

PROBLEM:

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

Data set for the Problem: Sparkling.csv and Rose.csv

SPARKLING WINE DATASET

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

After importing the necessary libraries, we will load the Sparkling wine data set and below are the first few rows of the dataset. We have also converted the Year Month column into a timestamp and index of the data frame using 'Parse Dates' function.

	Sparkling
YearMonth	
1980-01-01	1686
1980-02-01	1591
1980-03-01	2304
1980-04-01	1712
1980-05-01	1471

Table 1: Sparkling Wine sales Data

Next we will check the info of the data frame:

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
Data columns (total 1 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Sparkling  187 non-null    int64
dtypes: int64(1)
memory usage: 2.9 KB
```

From the above output we can derive the following

i. The data frame consists of 187 observations
ii. The data frame has no missing values

iii. The date time index ranges from 01-01-1980 to 09-07-1995
iv. The 'Sparkling' columns has wine sales values and it is of integer datatype.

Next we will plot the data. The below data shows the time plot of the dataset;

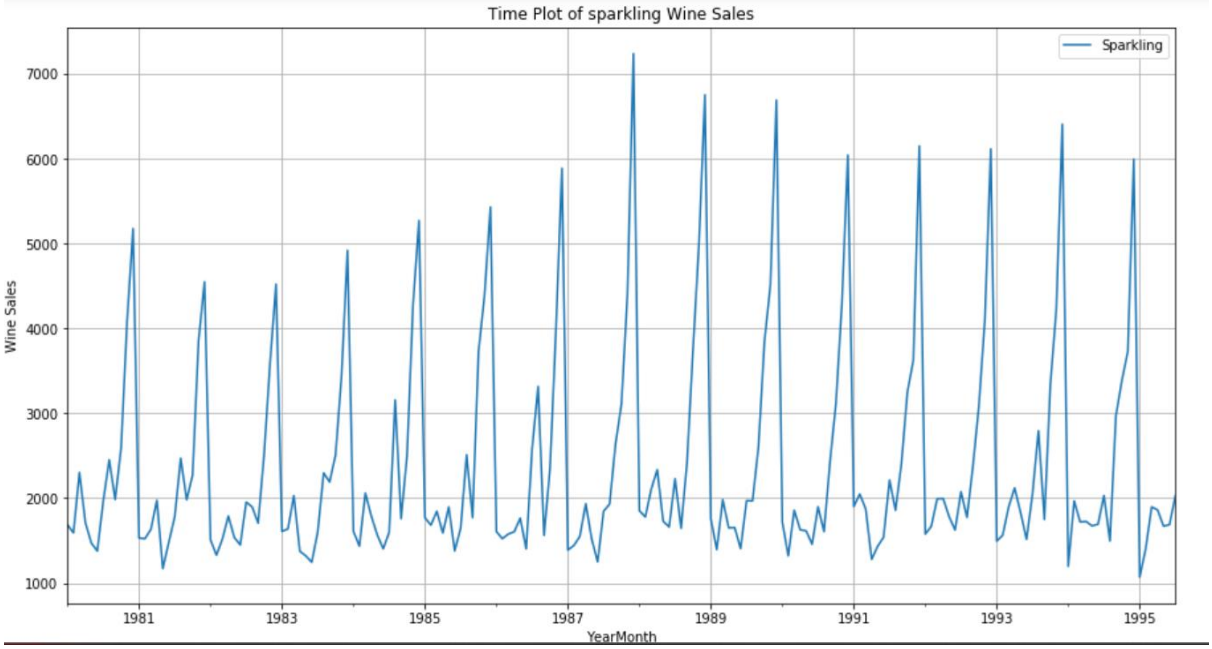

Figure 1: Time plot of Sparkling wine Sales

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

Below shows the 5 point summary of the dataset:

	Sparkling
count	187.000000
mean	2402.417112
std	1295.111540
min	1070.000000
25%	1605.000000
50%	1874.000000
75%	2549.000000
max	7242.000000

Table 2: 5 point summary of the data

The mean of the dataset is 2402.417 with a standard deviation of 1295.11. The dataset is ranged from 1070 to 7242.

Below is the Histplot distribution of the dataset:

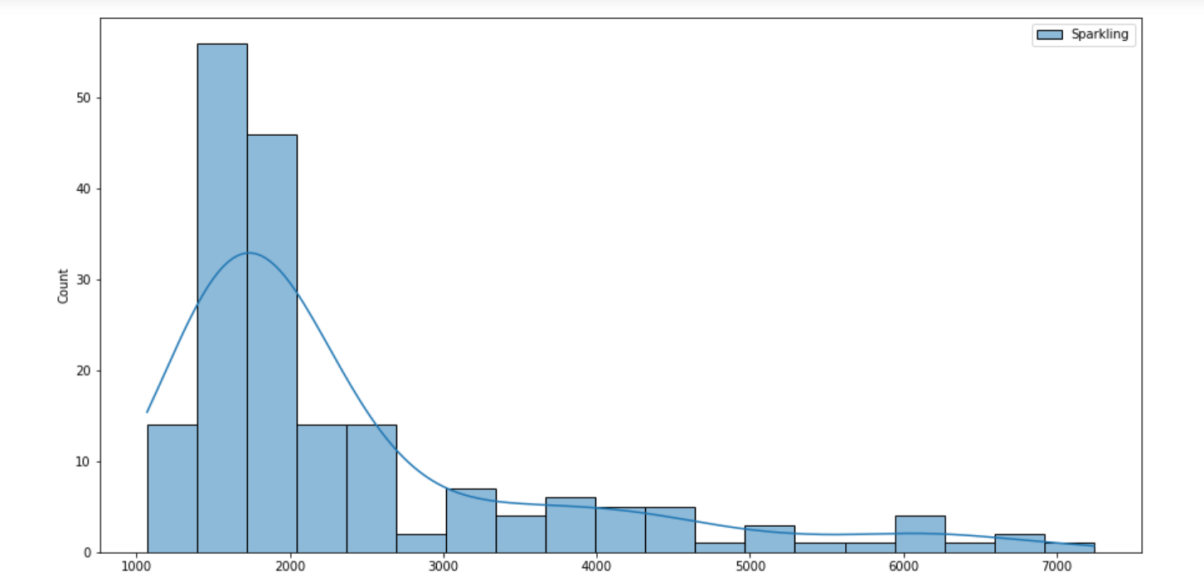

Figure 2: Histplot Distribution of the dataset

Below is the yearly plot of the dataset:

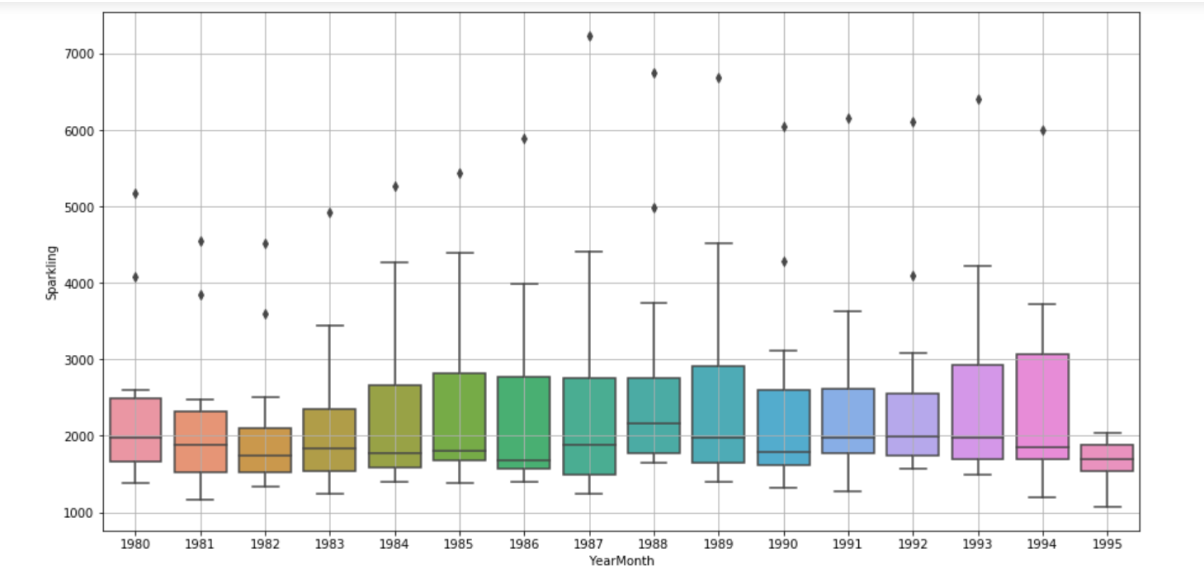

Figure 3: Yearly plot of the dataset

From the above plot we can see that the wine sales has decrease from 1980 to 1982 and the there is an increase in the sales from 1983 to 1989. The wine sales remains at a constant from 1990 to 1992 then again increases and then decreases till 1995.

Below is the monthly plot of the dataset:

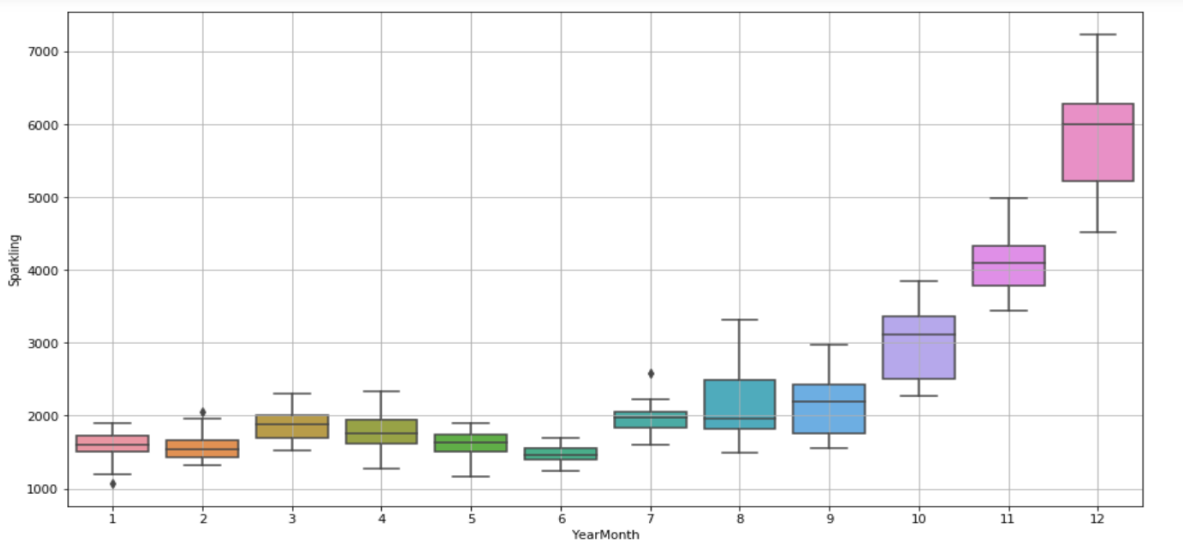

Figure 4: Monthly Plot of the dataset

In the month wise dataset we see that wine sales are at highest in the month of December and the lowest in the month of June. Wine sales are particularly low in the first half of the year and high in the last half.

Monthly plot of the data:

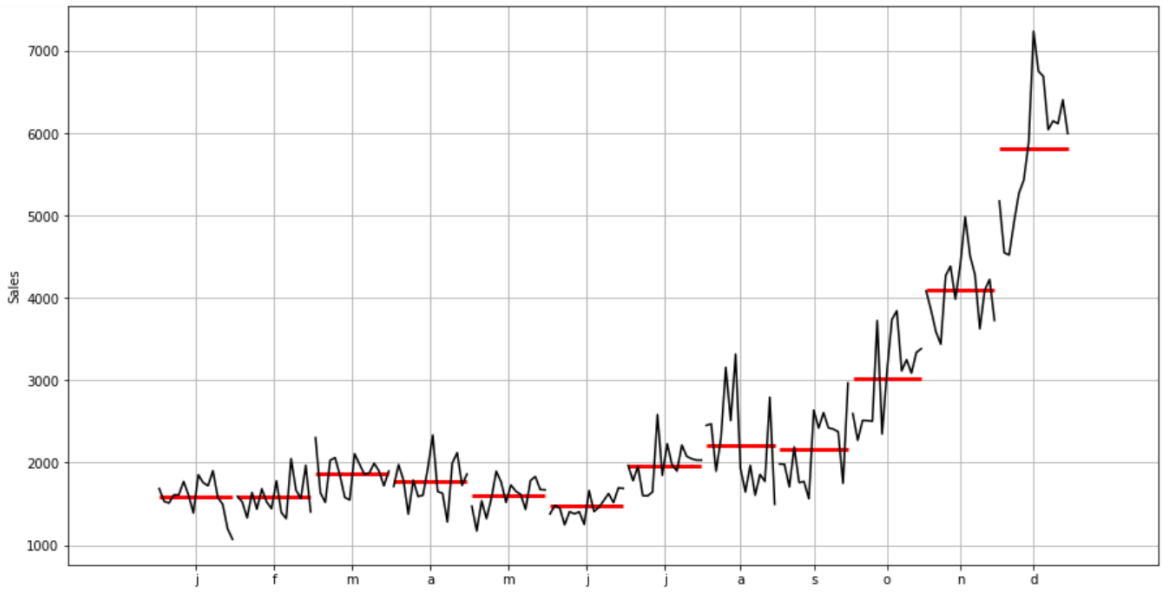

Figure 5: Month wise plot of the dataset

Monthly wise sales across years:

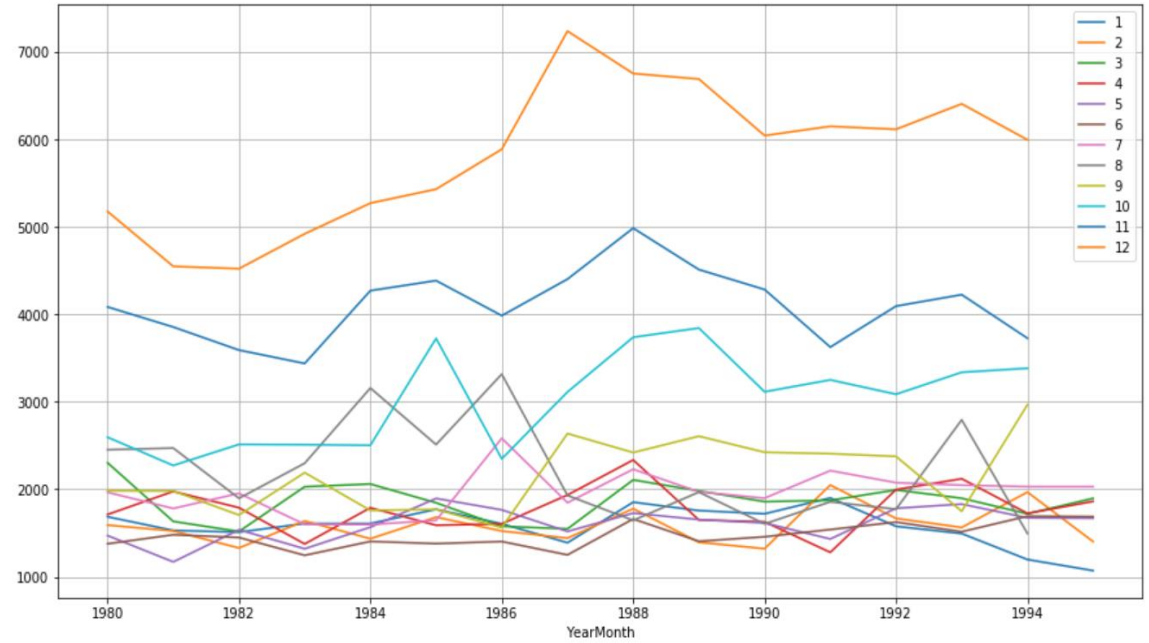

Figure 6: Monthly sales plot across years

Yearly average sales and percentage change of wine sales:

Figure 7: Average sales and percentage change

Decomposition:
The Data set is decomposed using additive and multiplicative model:

i. MULTIPLICATIVE:

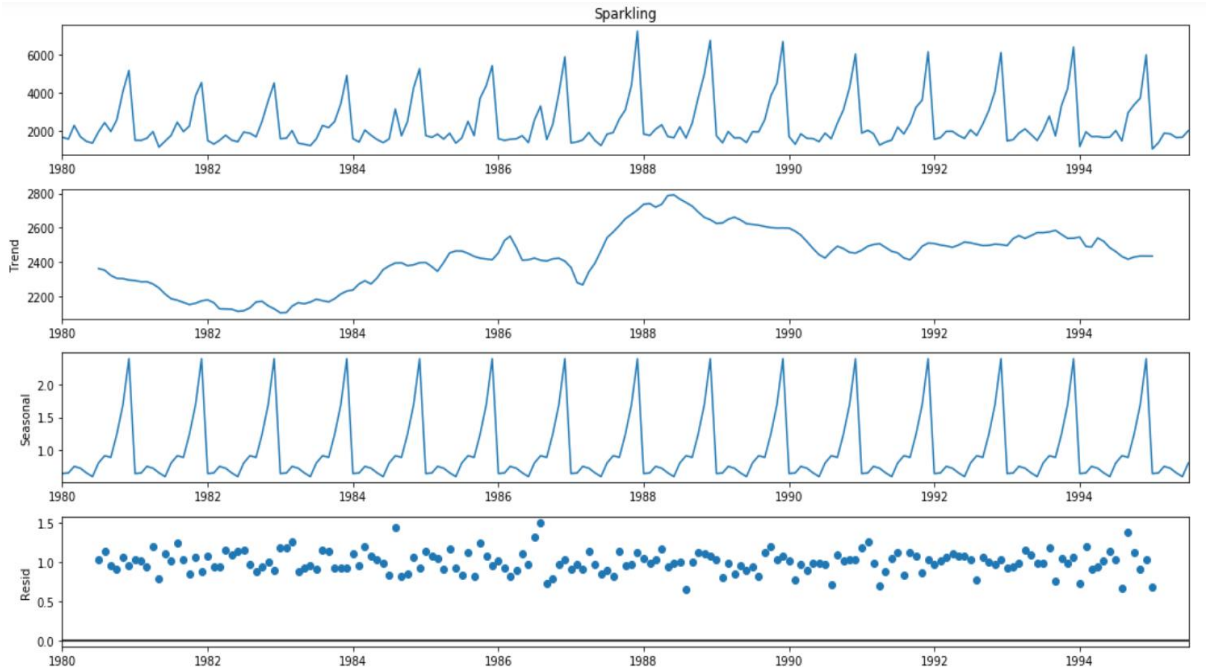

Figure 8: Multiplicative Decomposition of the data

ii. ADDITIVE:

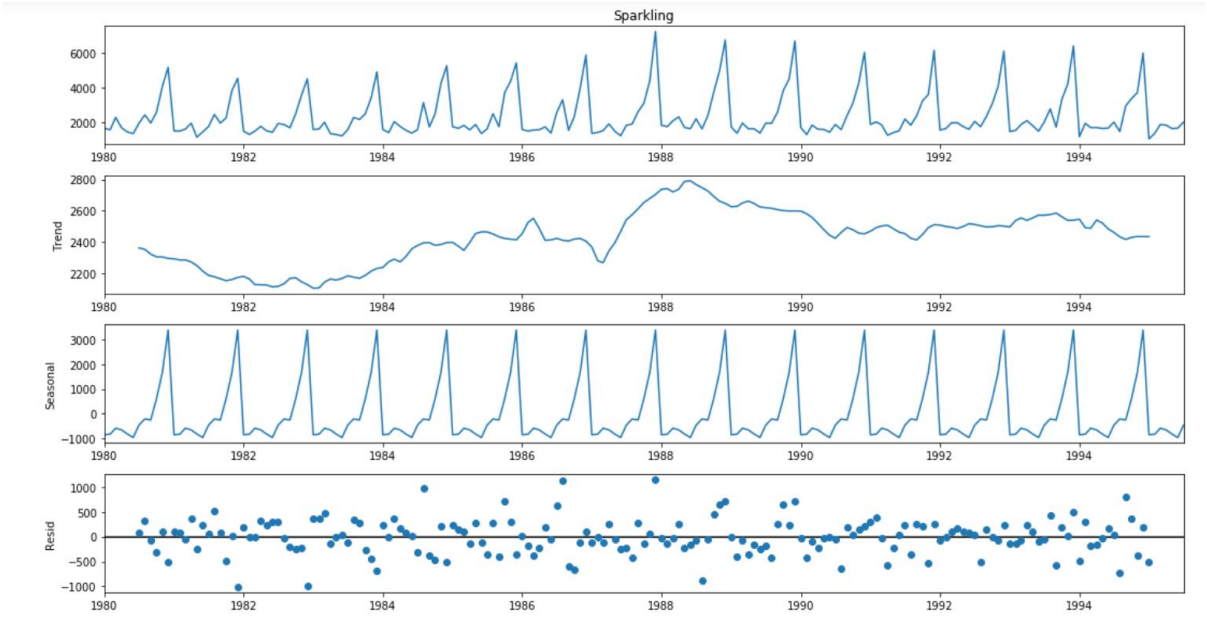

Figure 9: Additive Decomposition of the data

From the decomposition we see that the dataset has an irregular dataset. Also we can see that there is seasonality in the dataset where sales are low at the starting of the year and peaks near the end of the year. Also there is a certain pattern in the residual plot indicating that not all seasonal functions are captured by the model.

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

The test data should contain the recent years and hence, for the split we will consider the data from 1991 as test data and data before that as train data,

First few rows of the training data is

YearMonth	Sparkling
1980-01-01	1686
1980-02-01	1591
1980-03-01	2304
1980-04-01	1712
1980-05-01	1471

Last few rows of the training data is

YearMonth	Sparkling
1990-08-01	1605
1990-09-01	2424
1990-10-01	3116
1990-11-01	4286
1990-12-01	6047

First Few rows of the test data is

YearMonth	Sparkling
1991-01-01	1902
1991-02-01	2049
1991-03-01	1874
1991-04-01	1279
1991-05-01	1432

Last Few rows of the test data is

YearMonth	Sparkling
1995-03-01	1897
1995-04-01	1862
1995-05-01	1670
1995-06-01	1688
1995-07-01	2031

Table 3: First and last few rows of train and test data

Plot of Train and Test Data:

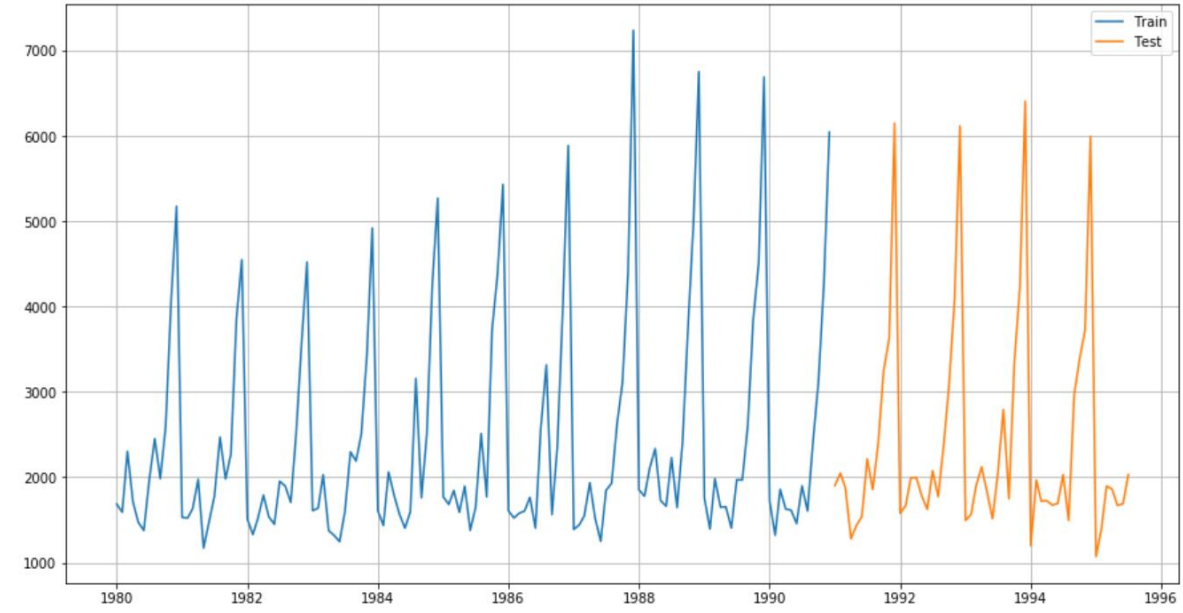

Figure 10: train and Test data Plot

Shape of the train and test data:

```
Shape of the training data is
(132, 1)
Shape of the test data is
(55, 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**

i. Naïve Forecast:

Naïve Forecast is an Estimating technique in which the last period's actuals are used as this period's forecast, without adjusting them or attempting to establish causal factors.

The below table shows the Naïve forecast on the test data :

YearMonth	Sparkling	Naive
1991-01-01	1902	6047
1991-02-01	2049	6047
1991-03-01	1874	6047
1991-04-01	1279	6047
1991-05-01	1432	6047

Table 4: Naïve Forecast

Plot of naïve forecast on test data:

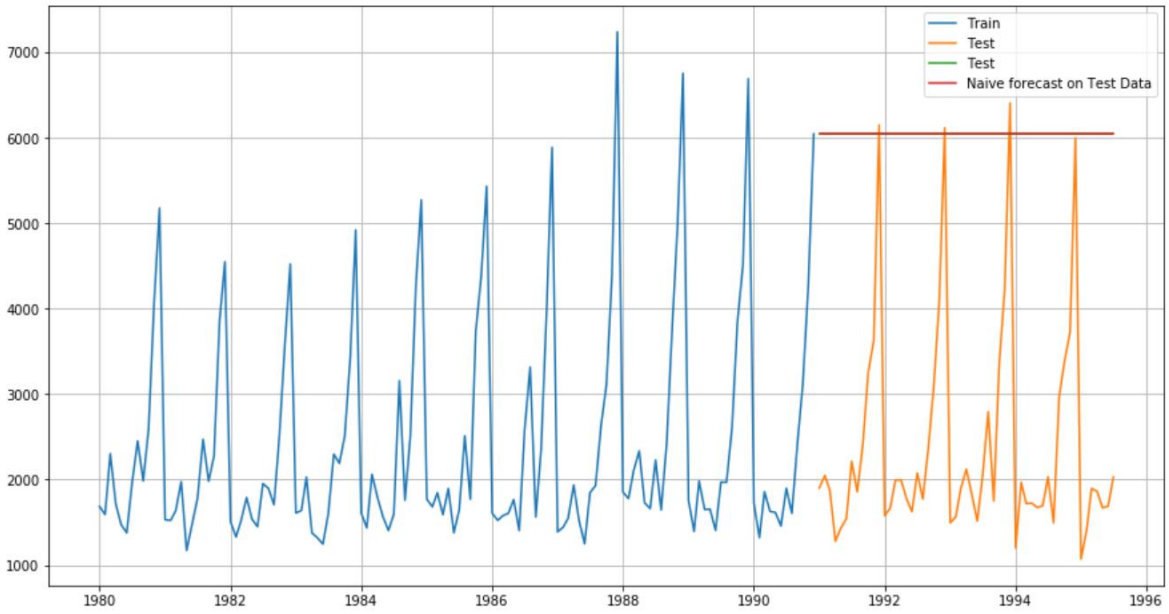

Figure 11: Naïve Forecast plot on test data

The Naïve forecast is a very basic model and produces a very straight line. It does not takes into account the trend or the seasonality. We calculate the RMSE for the model which is as follows:

```
RMSE for Naive forecast model on Sparkling wine data is 3864.279
```

ii. Linear Regression

For a linear regression model, the data must contain at least one predictor variable. The Sparkling data only contains a target variable which is the sales of the Sparkling wine. Therefore, we will have to create a new predictor variable to build the model. Hence, we will regress the Sparkling column with the order of occurrence of the values. The below tables show the training and test set after adding the new predictor variables:

Training data:

YearMonth	Sparkling	Time
1980-01-01	1686	1
1980-02-01	1591	2
1980-03-01	2304	3
1980-04-01	1712	4
1980-05-01	1471	5

YearMonth	Sparkling	Time
1990-08-01	1605	128
1990-09-01	2424	129
1990-10-01	3116	130
1990-11-01	4286	131
1990-12-01	6047	132

Test Data:

YearMonth	Sparkling	Time
1991-01-01	1902	133
1991-02-01	2049	134
1991-03-01	1874	135
1991-04-01	1279	136
1991-05-01	1432	137

YearMonth	Sparkling	Time
1995-03-01	1897	183
1995-04-01	1862	184
1995-05-01	1670	185
1995-06-01	1688	186
1995-07-01	2031	187

Plot of Linear Regression on test Data:

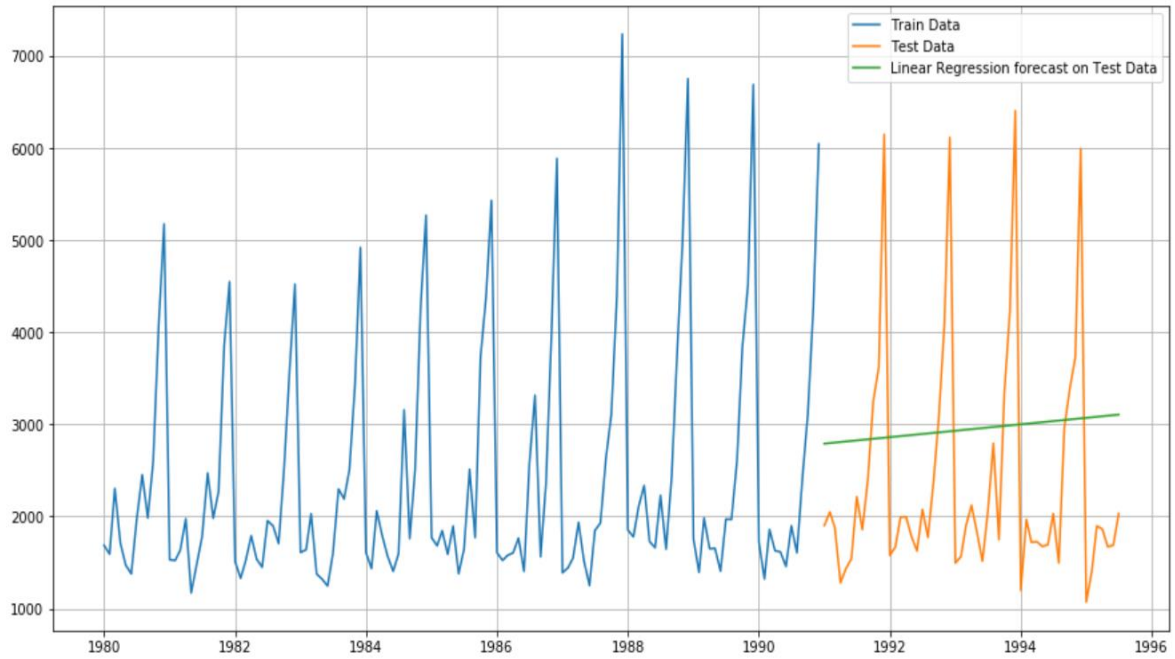

Figure 12: Linear regression plot on test data

The RMSE for the model is as follows:

```
RMSE for Linear Regression forecast model on Sparkling wine data is 1389.135
```

iii. Simple Average Forecast:

In a Simple Average forecasting technique, the mean of the training data is used as the forecast for all the periods in the test data. Like Naïve forecast this is also a very basic forecasting technique and it doesnot take into account the trend or seasonality. The below table shows the forecast for the Simple Average model:

	Sparkling	Average
YearMonth		
1991-01-01	1902	2403.780303
1991-02-01	2049	2403.780303
1991-03-01	1874	2403.780303
1991-04-01	1279	2403.780303
1991-05-01	1432	2403.780303

Table 5 Simple Average forecast on Test Data

Plot of Simple Average forecast on Test Data:

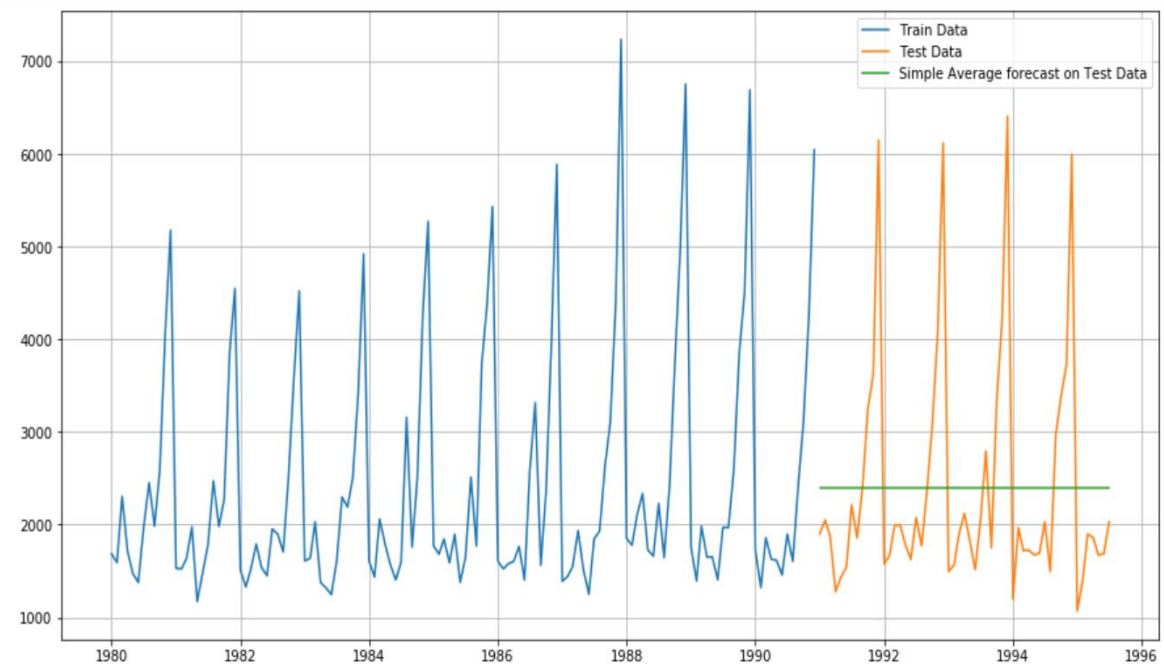

Figure 13: Simple average forecast on test data

The RMSE of the model is as follows:

```
RMSE for Simple Average forecast model on Sparkling wine data is 1275.082
```

iv. Moving Average Forecast:

A moving average is a technique that calculates the overall trend in a data set. This technique is very useful for forecasting short-term trends. It is simply the average of a select set of time periods. The moving averages smoothens the seasonal fluctuations in the time series data. The forecasts for first k-1 periods are null values as k number of data points are required to find the averages. Therefore, we will find various moving averages for the complete data and then split the data into training and test sets so that the test set does not contain any null values. We will calculate 2, 4, 6, and 9 point moving averages for the entire data.

Moving Average in entire data:

YearMonth	Sparkling	2 point moving average	4 point moving average	6 point moving average	9 point moving average
1980-01-01	1686	NaN	NaN	NaN	NaN
1980-02-01	1591	1638.5	NaN	NaN	NaN
1980-03-01	2304	1947.5	NaN	NaN	NaN
1980-04-01	1712	2008.0	1823.25	NaN	NaN
1980-05-01	1471	1591.5	1769.50	NaN	NaN

Table 6: 2point, 4point, 6point & 9point moving average on whole data

Moving Average on Test Data:

YearMonth	Sparkling	2 point moving average	4 point moving average	6 point moving average	9 point moving average
1991-01-01	1902	3974.5	3837.75	3230.000000	2705.666667
1991-02-01	2049	1975.5	3571.00	3304.000000	2753.888889
1991-03-01	1874	1961.5	2968.00	3212.333333	2800.222222
1991-04-01	1279	1576.5	1776.00	2906.166667	2731.333333
1991-05-01	1432	1355.5	1658.50	2430.500000	2712.111111

Table 7: 2point, 4point, 6point & 9point moving average on whole data

The smoothening effect produced by the moving averages depend on the window of rolling mean. Larger the window more is the smoothening effect. This can be confirmed from the below figure

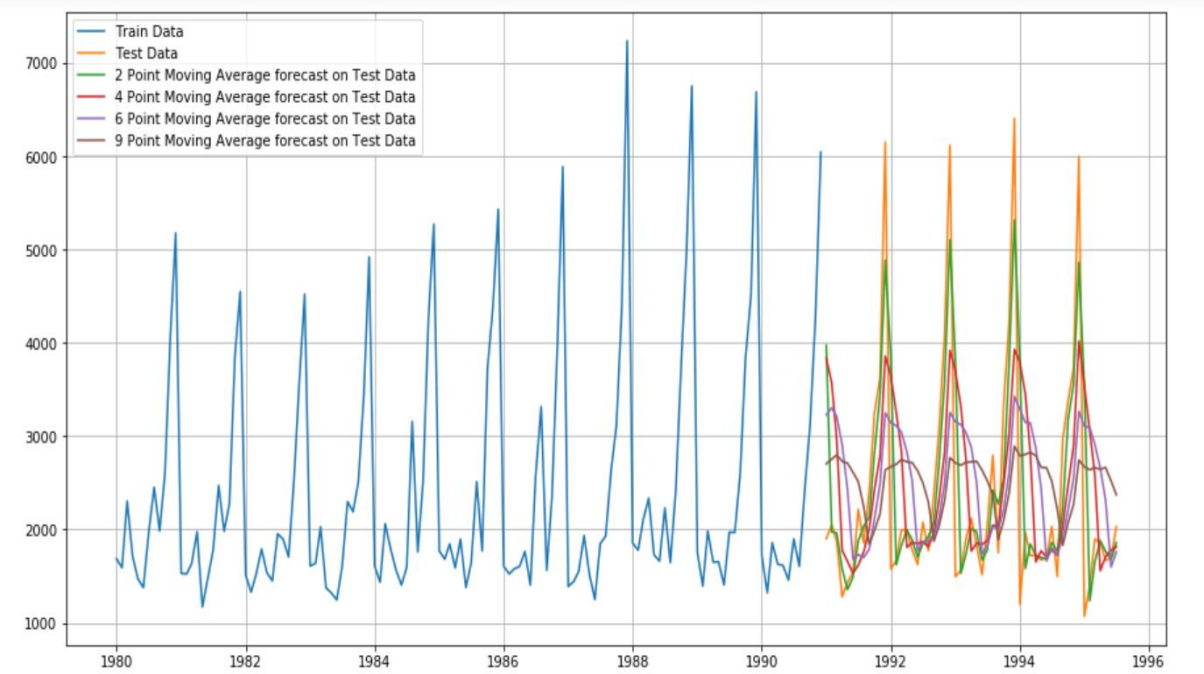

Figure 14: 2point, 4point, 6point & 9point moving average plot on test data

The RMSE for Moving Average forecast is as follows:

```
RMSE for 2 point Moving Average forecast model on Sparkling wine data is 813.401
RMSE for 4 point Moving Average forecast model on Sparkling wine data is 1156.590
RMSE for 6 point Moving Average forecast model on Sparkling wine data is 1283.927
RMSE for 9 point Moving Average forecast model on Sparkling wine data is 1346.278
```

Since the 2 Point moving average is the closest to the test data, it has the lowest RMSE value.

v. Simple Exponential Smoothing

A simple exponential smoothing is one of the simplest ways to forecast a time series. The basic idea of this model is to assume that the future will be more or less the same as the (recent) past. Thus, the only pattern that this model will learn from demand history is its level.
It is a time series forecasting method for univariate data without a trend or seasonality. It requires a single parameter, called alpha, also called the smoothing factor or smoothing coefficient. The following are the parameters of the SES Model:

```
{'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}
```

Here, the smoothing trend and smoothing seasonal parameters are NaN because we are using Simple Exponential Smoothing model. The below table shows the predictions done:

YearMonth	Sparkling	Prediction
1991-01-01	1902	2403.785621
1991-02-01	2049	2403.785621
1991-03-01	1874	2403.785621
1991-04-01	1279	2403.785621
1991-05-01	1432	2403.785621

Table 8: Simple Exponential Smoothing Prediction on test data

The plot of the Simple exponential smoothing at alpha=0 is as follows:

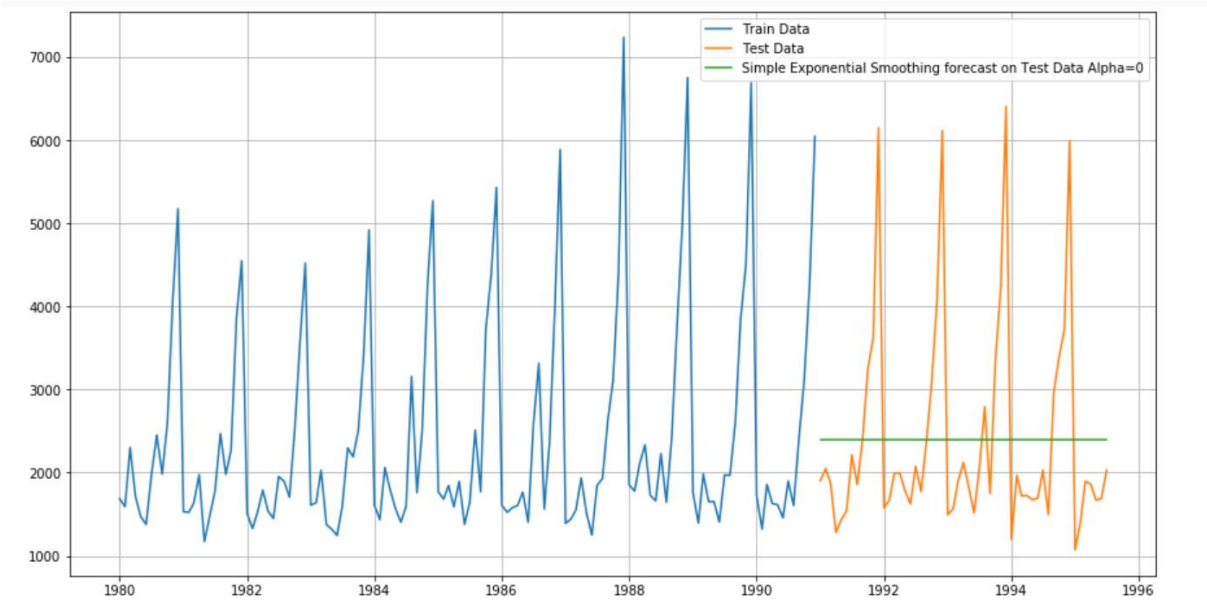

Figure 15: Simple Exponential Smoothing Forecast on test data

From the above figure we see that the forecast done by the Simple Exponential Smoothing model is a straight line. This is due the fact that SES model is only used for data which does not have any trend or seasonality.
The RMSE of the model is as follows:

```
RMSE for SES forecast model on Sparkling wine data is 1275.082
```

vi. Double Exponential Smoothing:

Double exponential smoothing employs a level component and a trend component at each period. Double exponential smoothing uses two weights, (also called smoothing parameters), to update the components at each period. This method is also called as Holt's trend corrected or second-order exponential smoothing. This method is used for forecasting the time series when the data has a linear trend and no seasonal pattern.

The following are the parameters of the Double Exponential Smoothing:

```
{'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}
```

The smoothing level parameter is the alpha value while the smoothing trend parameter is the beta value. The model found the optimal alpha value to be 0.647 and beta value as zero. The below table shows the predictions done by the model:

YearMonth	Sparkling	Prediction
1991-01-01	1902	5281.503187
1991-02-01	2049	5308.563699
1991-03-01	1874	5335.624211
1991-04-01	1279	5362.684723
1991-05-01	1432	5389.745234

Table 9: Double Exponential Smoothing prediction on test data

The Plot of the forecast is as follows:

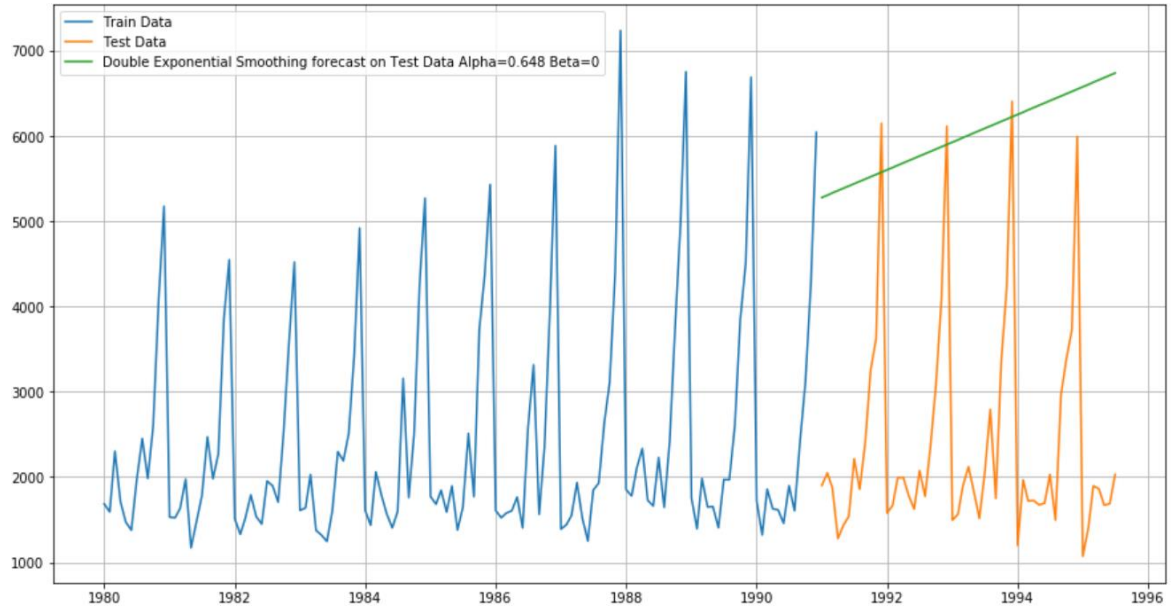

Figure 16: Double Exponential Smoothing forecast on test data

The RMSE of the model is:

```
RMSE for DES forecast model on Sparkling wine data is 3851.073
```

vii. Triple Exponential Smoothing(Holt-Winters Model)

Triple exponential smoothing (Holt-Winters model) employs level, trend and seasonal components at each period. Holt-Winters model is used for data that has both trend and seasonality components. Triple exponential smoothing uses three smoothing parameters, alpha, beta and gamma, to update the components at each period. Alpha is the level smoothing parameter, beta is the trend smoothing parameter and gamma is the seasonal smoothing parameter. All of these parameters range between 0 and 1. Large values means that the model pays attention mainly to the most recent past observations, whereas smaller values mean more of the history is taken into account when making a prediction.
Following are the parameters of Triple Exponential smoothing:

```
{'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}
```

The smoothing level parameter is the alpha value, the smoothing trend parameter is the beta value and the smoothing seasonal parameter is the gamma value. The model found the optimal alpha value to be 0.086, beta value as 2.68 and gamma value to be 0.476. The below table shows the predictions done by the model:

YearMonth	Sparkling	Prediction
1991-01-01	1902	1532.422837
1991-02-01	2049	1241.383411
1991-03-01	1874	1726.781799
1991-04-01	1279	1584.323098
1991-05-01	1432	1494.024475

Table 10: Triple Exponential Smoothing prediction on test data

The plot of the forecast is as follows:

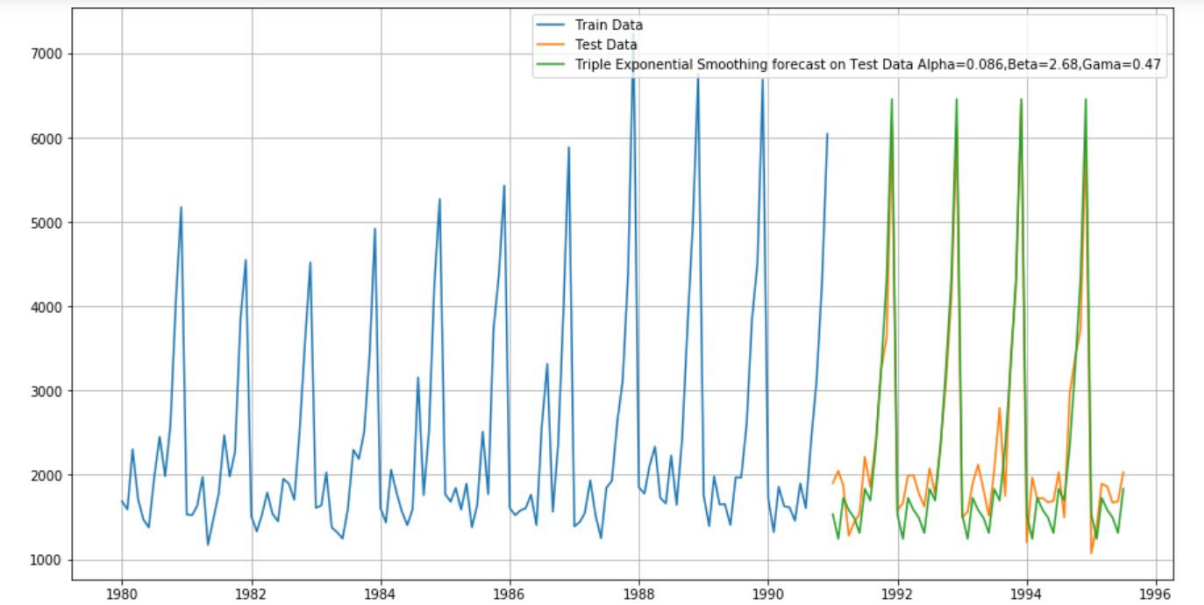

Figure 17: Triple Exponential Smoothing forecast on sales data

The RMSE of the model is as follows:

```
RMSE for TES forecast model on Sparkling wine data is 362.732
```

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.**

A stationary time series is one whose properties do not depend on the time at which the series is observed. Thus, time series with trends, or with seasonality, are not stationary — the trend and seasonality will affect the value of the time series at different times.
Time series are stationary if they do not have trend or seasonal effects. Summary statistics calculated on the time series are consistent over time, like the mean or the variance of the observations.

There are various statistical tests to check stationarity, including the Augmented Dickey-Fuller (ADF) test. The ADF test is a widely used test for checking the stationarity of a time series, and it checks for the presence of a unit root in the data. It is a hypothesis test with the null and alternate hypothesis as follows:

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.

The following is the output of the ADF Test:

```
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
dtype: float64
```

The following is the plot of the ADF Test:

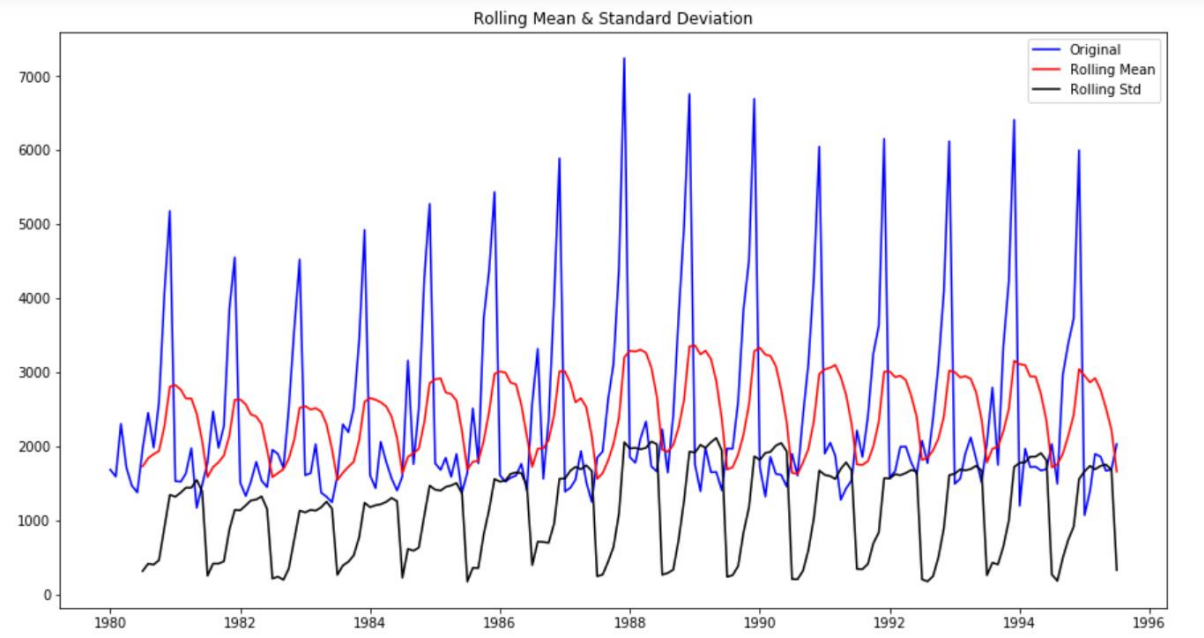

Figure 18: Check of stationarity on whole data

The p-value from the ADF test is greater than 0.05 and therefore we fail to reject the Null hypothesis. Hence, the Sparkling data is non-stationary.

Differencing and checking for stationarity:

Differencing can help stabilise the mean of a time series by removing changes in the level of a time series, and therefore eliminating (or reducing) trend and seasonality thereby making the series stationary. As our data is a monthly data, one seasonal period is of 12 months.

The result of the ADF test is as follows:

```
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
dtype: float64
```

The plot of the test is as follows:

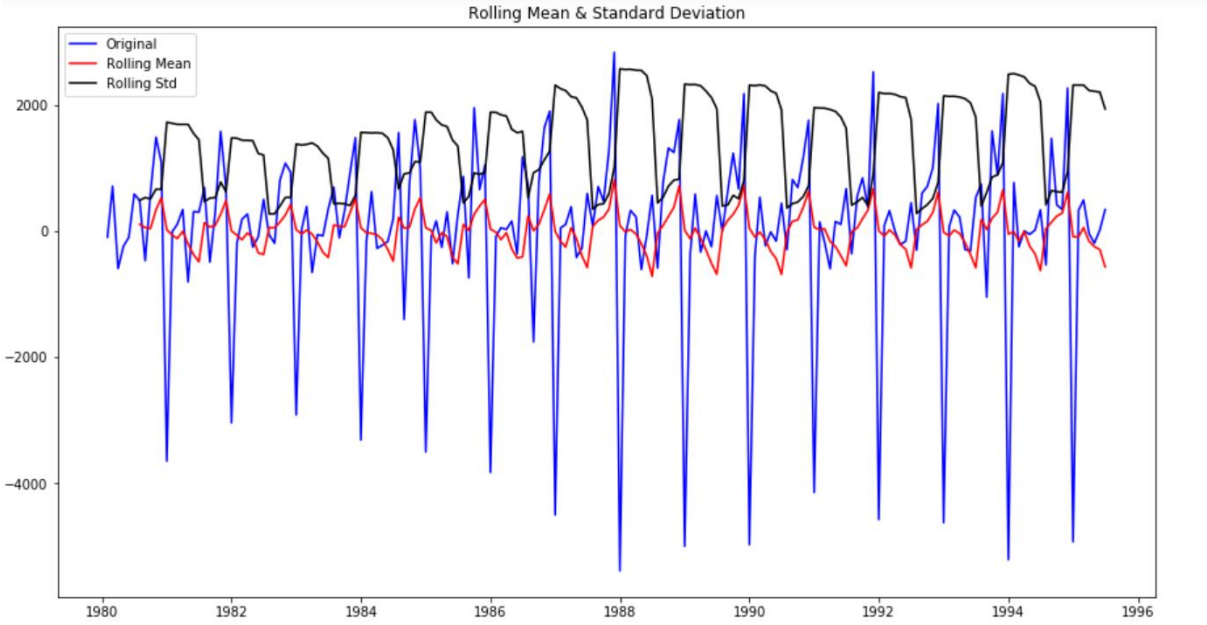

Figure 19: Differencing and check for stationarity on whole data

The p-value from the ADF test is less than 0.05 and therefore we reject the Null hypothesis. Therefore, taking the difference of the data has made the data stationary.

Checking for stationarity on the training data:

The following is the output of ADF test on training data:

```
Results of Dickey-Fuller Test:
Test Statistic                -1.208926
p-value                        0.669744
#Lags Used                    12.000000
Number of Observations Used   119.000000
Critical Value (1%)            -3.486535
Critical Value (5%)            -2.886151
Critical Value (10%)           -2.579896
dtype: float64
```

The plot of the test is as follows:

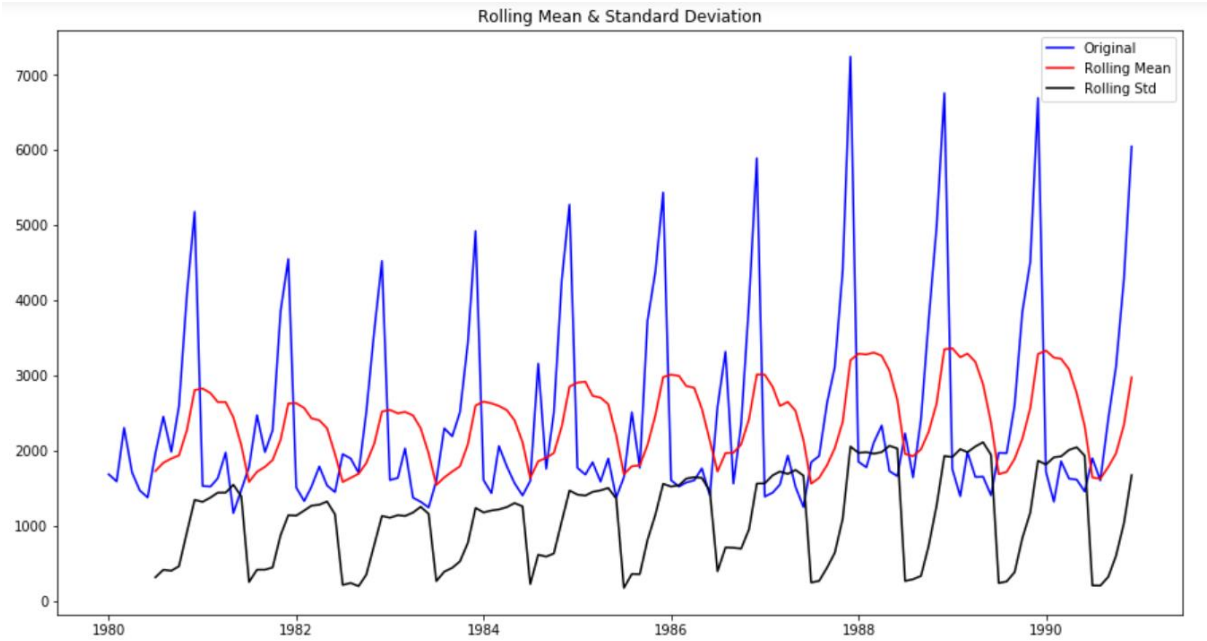

Figure 20: Check for stationarity on training data

The p-value from the ADF test is greater than 0.05 and therefore we fail to reject the Null hypothesis. Hence, the training data is non-stationary. Therefore we have to do differencing.

The following is the output of the ADF after differencing:

```
Results of Dickey-Fuller Test:
Test Statistic                  -3.342905
p-value                          0.013066
#Lags Used                      10.000000
Number of Observations Used    108.000000
Critical Value (1%)             -3.492401
Critical Value (5%)             -2.888697
Critical Value (10%)            -2.581255
dtype: float64
```

The following is the plot of the test:

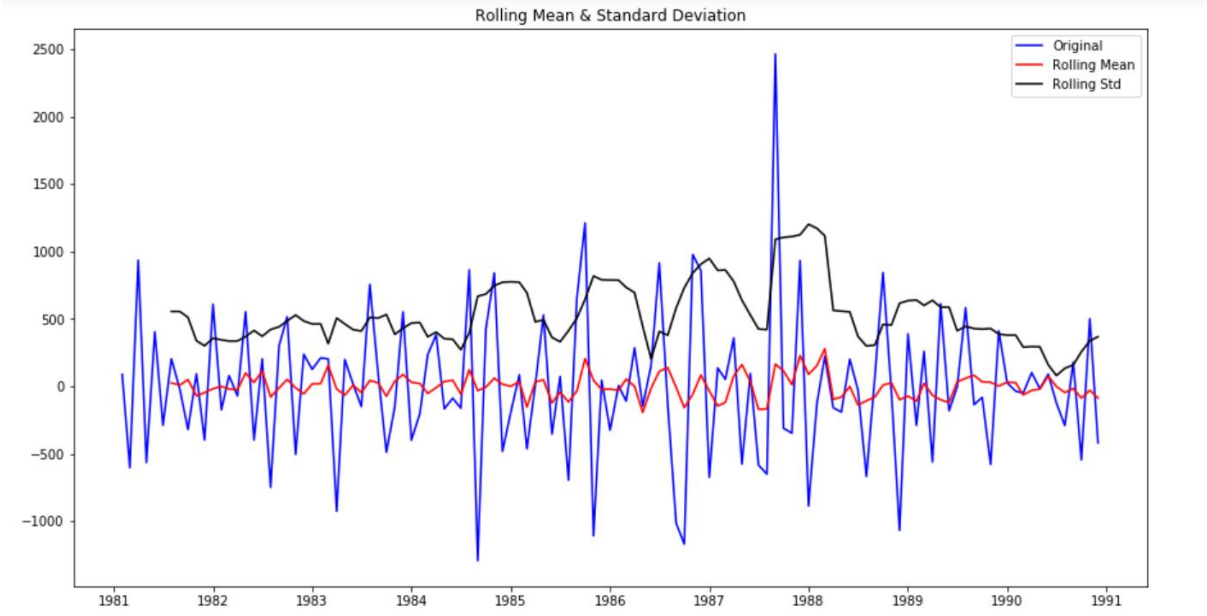

Figure 21: Differencing and check for stationarity on training data

The p-value from the ADF test is less than 0.05 and therefore we reject the Null hypothesis. Therefore, taking the difference of the training data has made the data stationary.

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

ARIMA stands for Auto Regressive Integrated Moving Average model. There are total three parameters related to three different components of the model. The p parameter indicates the lag used in the Auto Regressive component. The d parameter is the order of level differencing applied to make the data stationary. The q parameter is the lag used in the Moving Average component. Here, we are using a value of d = 1. Various combinations of the p, d and q parameter are used to build various ARIMA models and the combination that give the lowest AIC value is used to evaluate the model on the test data.

Some parameter combinations for the Model...
Model: (0, 1, 0)
Model: (0, 1, 1)
Model: (0, 1, 2)
Model: (0, 1, 3)
Model: (1, 1, 0)
Model: (1, 1, 1)
Model: (1, 1, 2)
Model: (1, 1, 3)
Model: (2, 1, 0)
Model: (2, 1, 1)
Model: (2, 1, 2)
Model: (2, 1, 3)
Model: (3, 1, 0)
Model: (3, 1, 1)
Model: (3, 1, 2)
Model: (3, 1, 3)

The Below table shows the AIC values in ascending orders:

	param	AIC
15	(3, 1, 3)	2155.774954
3	(0, 1, 3)	2168.092541
7	(1, 1, 3)	2171.026404
11	(2, 1, 3)	2171.039589
10	(2, 1, 2)	2176.868115
14	(3, 1, 2)	2187.314727
2	(0, 1, 2)	2187.441010
13	(3, 1, 1)	2188.222098
6	(1, 1, 2)	2188.463345
9	(2, 1, 1)	2199.858613
5	(1, 1, 1)	2204.934049
12	(3, 1, 0)	2208.402501
8	(2, 1, 0)	2227.302762
1	(0, 1, 1)	2230.162908
4	(1, 1, 0)	2250.318127
0	(0, 1, 0)	2251.359720

Table 11: ARIMA AIC in ascending order

We will take the p,d,q as 0,1,3. The below is the output which shows the model summary of ARIMA model:

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	132
Model:	ARIMA(0, 1, 3)	Log Likelihood	-1112.997
Date:	Fri, 26 May 2023	AIC	2233.995
Time:	20:01:36	BIC	2245.496
Sample:	01-01-1980	HQIC	2238.668
	- 12-01-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.4677	0.804	-0.582	0.561	-2.044	1.108
ma.L2	-0.3782	0.537	-0.704	0.481	-1.431	0.674
ma.L3	-0.1532	0.217	-0.706	0.480	-0.578	0.272
sigma2	1.359e+06	1.14e+06	1.197	0.231	-8.66e+05	3.58e+06

Ljung-Box (Q):	316.81	Jarque-Bera (JB):	10.48
Prob(Q):	0.00	Prob(JB):	0.01
Heteroskedasticity (H):	2.76	Skew:	0.36
Prob(H) (two-sided):	0.00	Kurtosis:	4.18

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

The diagnostic plot is as follows:

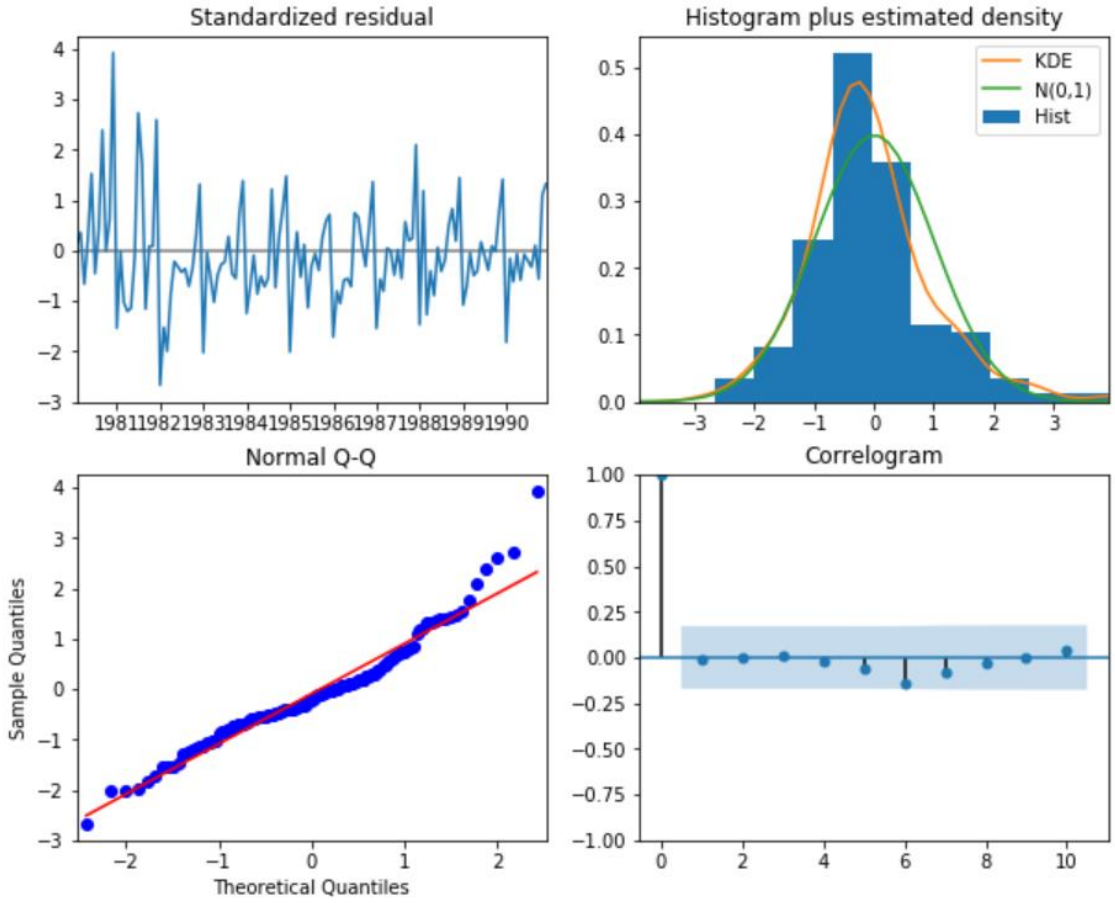

Figure 22: Diagnostic plot ARIMA (0,1,3)

The RMSE of the model is :

```
RMSE for Auto ARIMA forecast model on Sparkling wine data is 1310.517
```

The forecast plot is as follows:

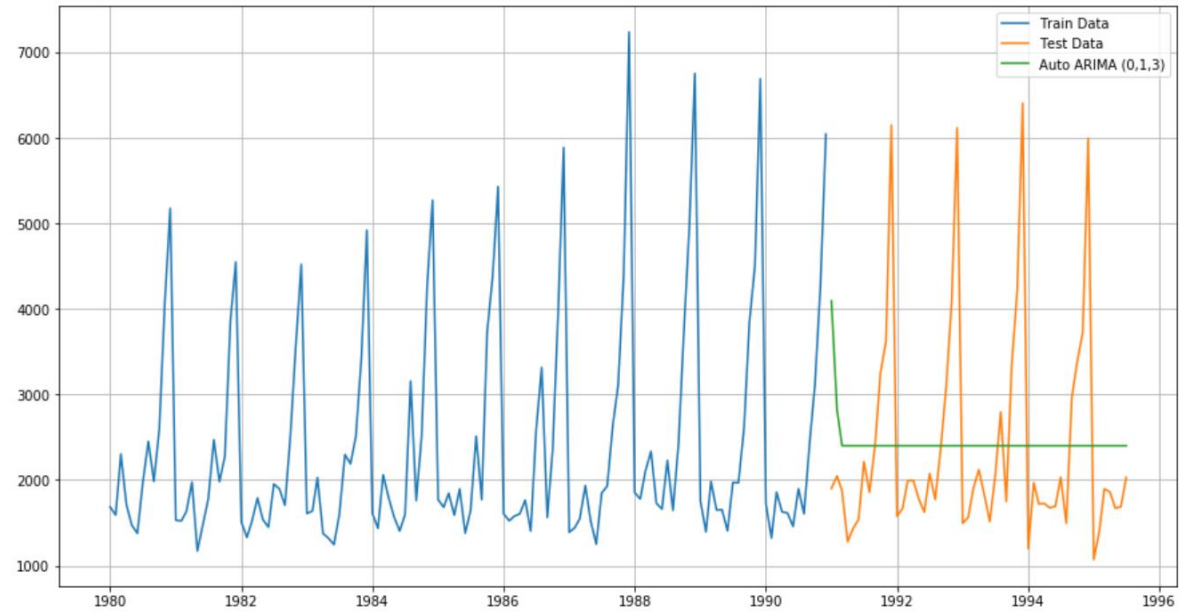

Figure 23: ARIMA (0,1,3) forecast on test data

We will also try p,d,q as 3,1,3. The following is the summary results:

```
                               SARIMAX Results
==========================================================================================
Dep. Variable:                      Sparkling   No. Observations:                  132
Model:                 ARIMA(3, 1, 3)   Log Likelihood               -1103.722
Date:                Fri, 26 May 2023   AIC                           2221.444
Time:                        20:01:38   BIC                           2241.570
Sample:                    01-01-1980   HQIC                          2229.622
                         - 12-01-1990
Covariance Type:                  opg
==========================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.4753      0.143      3.327      0.001       0.195       0.755
ar.L2         -0.9849      0.056    -17.733      0.000      -1.094      -0.876
ar.L3          0.4820      0.118      4.074      0.000       0.250       0.714
ma.L1         -0.9275      0.324     -2.867      0.004      -1.562      -0.293
ma.L2          0.9137      0.239      3.824      0.000       0.445       1.382
ma.L3         -0.9853      0.208     -4.746      0.000      -1.392      -0.578
sigma2      1.517e+06   2.84e-07   5.34e+12      0.000    1.52e+06    1.52e+06
==========================================================================================
Ljung-Box (Q):                      317.29   Jarque-Bera (JB):                 3.20
Prob(Q):                              0.00   Prob(JB):                         0.20
Heteroskedasticity (H):               2.75   Skew:                             0.29
Prob(H) (two-sided):                  0.00   Kurtosis:                         3.51
==========================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 1.83e+28. Standard errors may be unstable.
```

The diagnostic plot as follows:

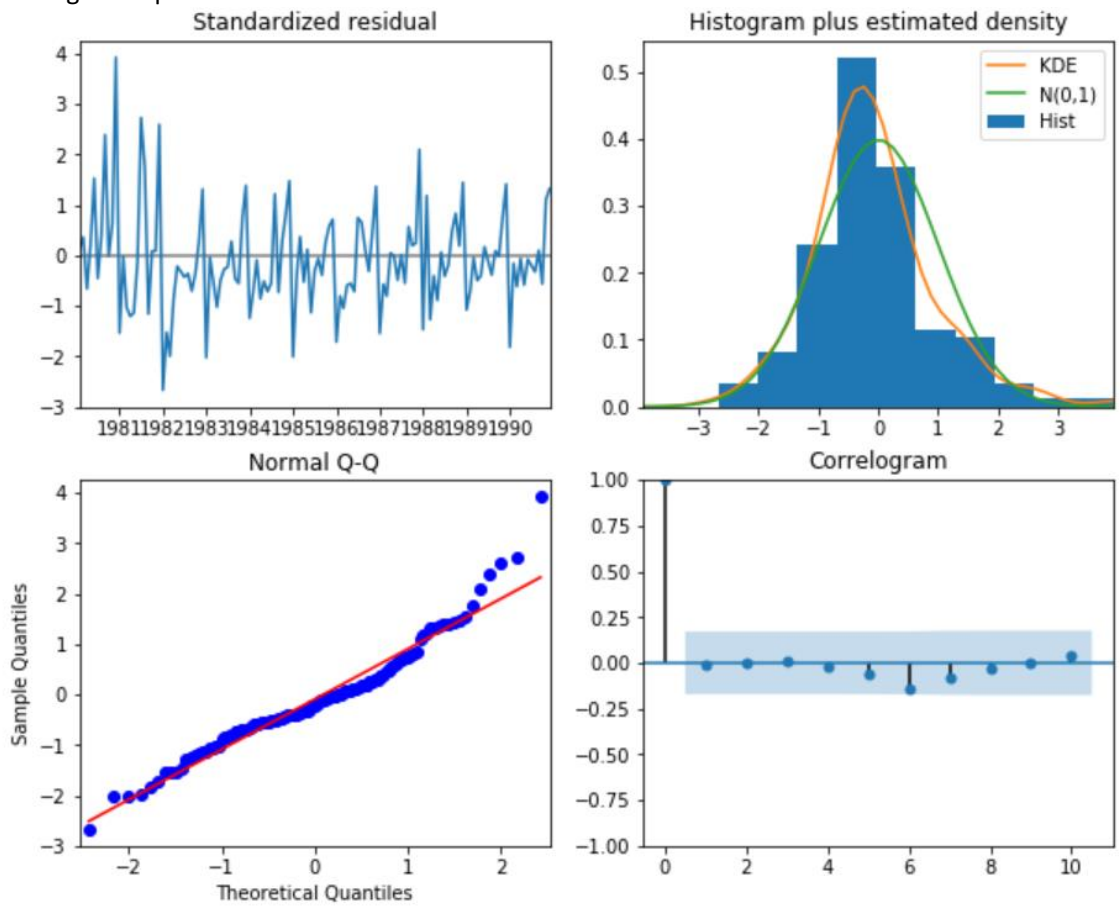

Figure 24: Diagnostic plot ARIMA (3,1,3)

The RMSE of the model is as follows:

```
RMSE for Auto ARIMA forecast model on Sparkling wine data is 1229.645
```

The forecast plot is as follows:

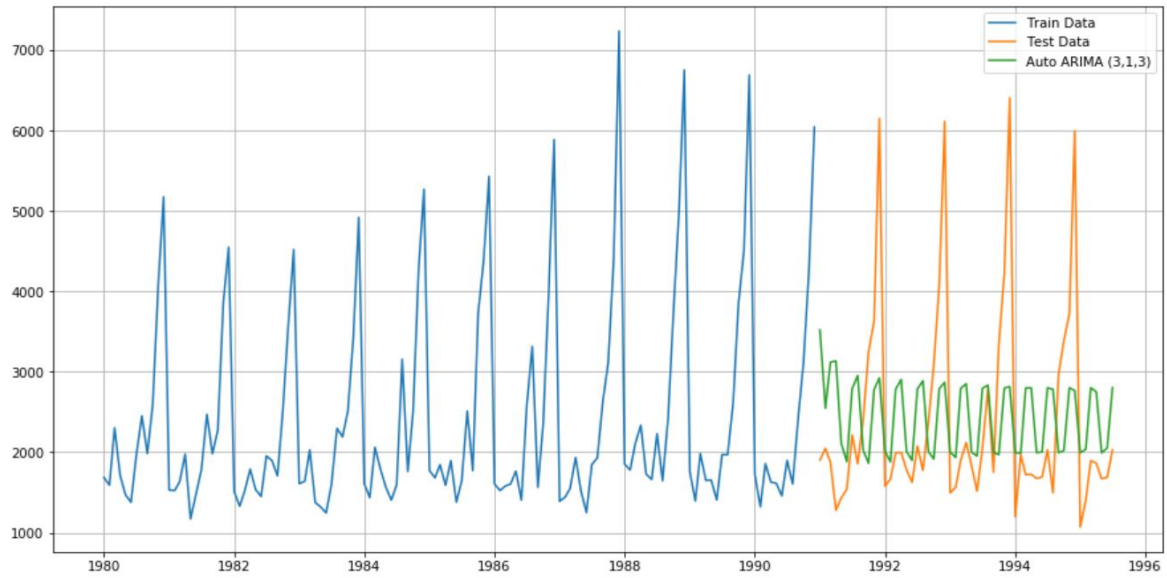

Figure 25: ARIMA (3,1,3) forecast on test data

This is the ideal model compared to the earlier.

Automated SARIMA

SARIMA stands for Seasonal Auto Regressive Integrated Moving Average model. It is an extension of the ARIMA model. There are total 7 parameters used to define the model. The parameters (p, d, q) are the same as the ARIMA model. The parameters (P, D, Q) are the seasonal counterparts of (p, d, q). The parameter F is the seasonality of the data which is 12 in our case. Here, we are using a value of d = 1 and D = 1. Various combinations of the p, d, q, P, D, and Q parameter are used to build various SARIMA models and the combination that give the lowest AIC value is used to evaluate the model on the test data.

To confirm the seasonality we can also look at the ACF graph. The ACF graph is as follows:

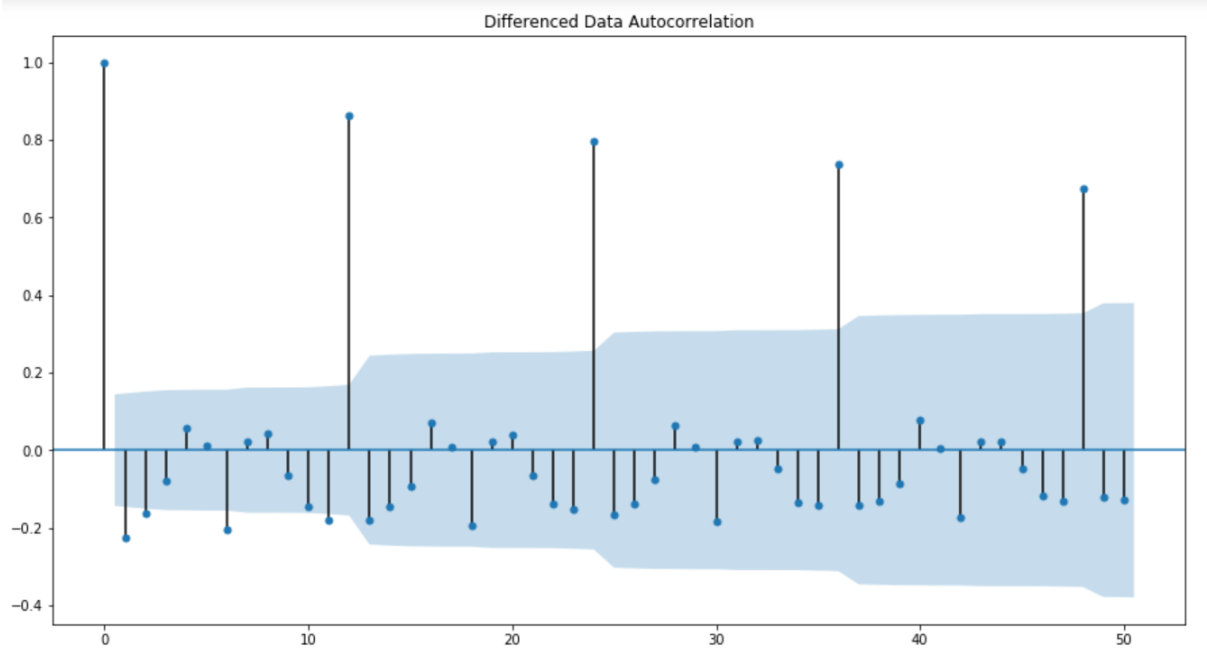

Figure 26: ACF graph

The ACF graph confirms the seasonality being 12. So, we will take seasonality 12.

```
Examples of some parameter combinations for Model...
Model: (0, 1, 1)(0, 0, 1, 12)
Model: (0, 1, 2)(0, 0, 2, 12)
Model: (1, 1, 0)(1, 0, 0, 12)
Model: (1, 1, 1)(1, 0, 1, 12)
Model: (1, 1, 2)(1, 0, 2, 12)
Model: (2, 1, 0)(2, 0, 0, 12)
Model: (2, 1, 1)(2, 0, 1, 12)
Model: (2, 1, 2)(2, 0, 2, 12)
```

The below table shows the AIC values of different parameter combinations in ascending order:

	param	seasonal	AIC
50	(1, 1, 2)	(1, 0, 2, 12)	1555.584247
53	(1, 1, 2)	(2, 0, 2, 12)	1555.929669
26	(0, 1, 2)	(2, 0, 2, 12)	1557.121563
23	(0, 1, 2)	(1, 0, 2, 12)	1557.160507
77	(2, 1, 2)	(1, 0, 2, 12)	1557.340402

Table 12: SARIMA AIC in ascending order

The parameter combination of p=1, d=1, q=2, P=1, D=0, Q=2, and F=12 gives the lowest AIC value. We will use this model for making predictions on the test data. The below output shows the model summary of SARIMA(1, 1, 2)(1, 0, 2, 12) model:

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1, 1, 2)x(1, 0, 2, 12)	Log Likelihood	-770.792
Date:	Fri, 26 May 2023	AIC	1555.584
Time:	20:22:13	BIC	1574.095
Sample:	0 - 132	HQIC	1563.083
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.6282	0.255	-2.463	0.014	-1.128	-0.128
ma.L1	-0.1041	0.225	-0.463	0.643	-0.545	0.337
ma.L2	-0.7276	0.154	-4.735	0.000	-1.029	-0.426
ar.S.L12	1.0439	0.014	72.840	0.000	1.016	1.072
ma.S.L12	-0.5550	0.098	-5.663	0.000	-0.747	-0.363
ma.S.L24	-0.1355	0.120	-1.133	0.257	-0.370	0.099
sigma2	1.506e+05	2.03e+04	7.401	0.000	1.11e+05	1.9e+05

Ljung-Box (Q):	23.02	Jarque-Bera (JB):	11.72
Prob(Q):	0.99	Prob(JB):	0.00
Heteroskedasticity (H):	1.47	Skew:	0.36
Prob(H) (two-sided):	0.26	Kurtosis:	4.48

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

The Diagnostic plot is as follows:

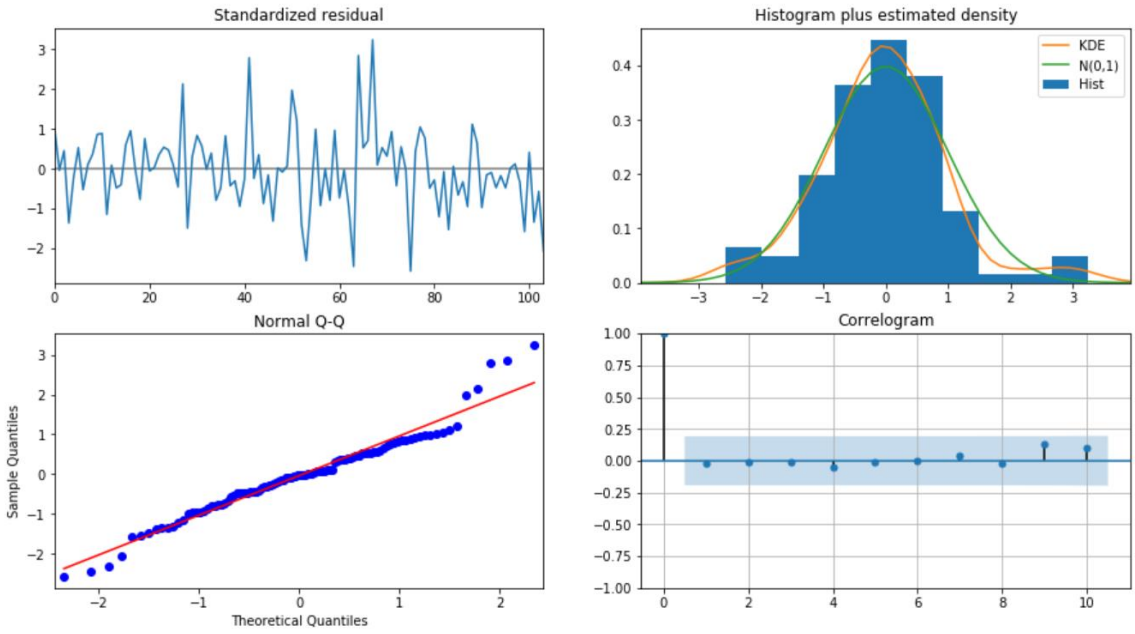

Figure 27: Diagnostic Plot SARIMA (1,1,2) (1,0,2,12)

The RMSE for the SARIMA model is:

RMSE for Auto SARIMA forecast model on Sparkling wine data is 528.636

The auto SARIMA summary frame is as follows:

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1327.364146	388.345557	566.220841	2088.507451
1	1315.100345	402.009781	527.175652	2103.025037
2	1621.570193	402.003388	833.658031	2409.482354
3	1598.854330	407.241640	800.675383	2397.033278
4	1392.671880	407.971816	593.061814	2192.281946

Table 13: Auto SARIMA summary frame

The Plot for the SARIMA model is as follows:

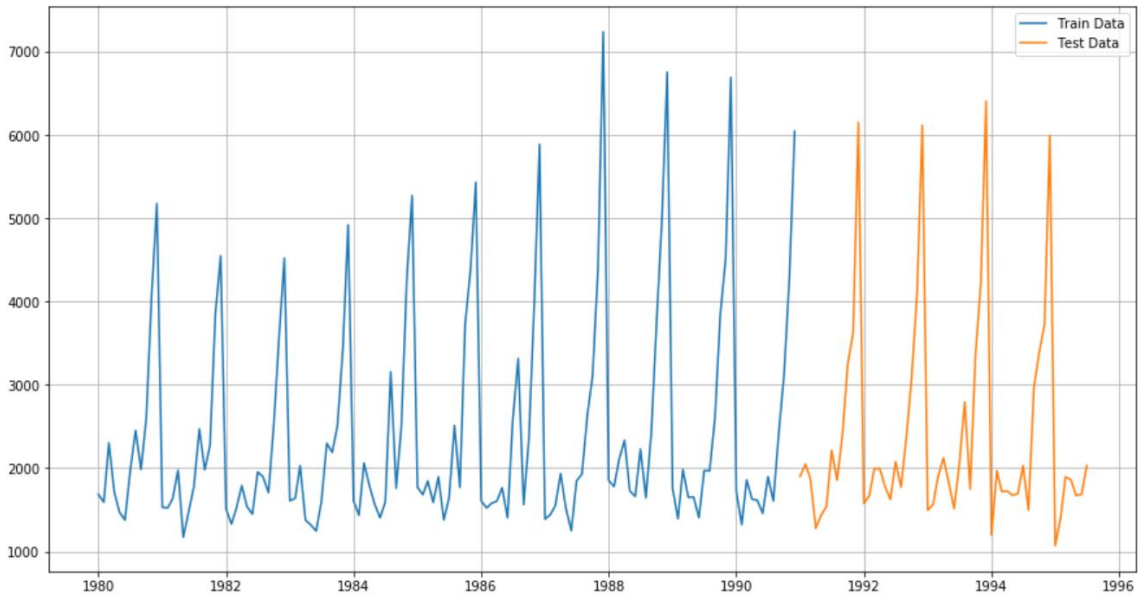

Figure 28: SARIMA (1,1,2) (1,0,2,12) plot

7. **Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.**

The below table shows the summary of RMSE values of the all the models built above in ascending order:

	Test RMSE
Alpha=0.086, Beta=2.68, Gamma=0.47, TES	362.731657
Auto SARIMAX(1,1,2)(1,0,2,12)	528.635883
2 Point Moving Average	813.400684
4 Point Moving Average	1156.589694
Auto ARIMA(3,1,3)	1229.644724
Simple Average	1275.081804
Alpha=0, SES	1275.081823
6 Point Moving Average	1283.927428
Auto ARIMA(0,1,3)	1310.517315
9 Point Moving Average	1346.278315
Linear Regression	1389.135175
Alpha=0.648, Beta=0, DES	3851.072597
Naive Model	3864.279352

Table 14: RMSE values of all models in ascending order

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

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

We had the dataset of Sparkling Wine sales for ABC Estate Wines which contains the wine sales from 1980 to 1995 and we were tasked to analyse and predict the sales for the next 12 months. The dataset was read into a dataframe and proper date time index was created. We saw that there was no missing data so we advanced to do the exploratory data analysis. We plotted the data on a histplot and then observed that wine sales peaked during the last quarter of the year. The highest sales recorded in the month of December. Then we went ahead and performed additive and multiplicative decomposition od the dataset. We observed that there was no clear trend in the dataset but had a clear seasonality. Then we divided the data into train and test to perform the forecast models. We performed the simple Smoothing models for forecast like Naïve, Linear Regression, Simple Exponential Smoothing, Holt(Double Exponential Smoothing), Holt-Winters(Triple Exponential Smoothing) and Moving Average. Then we performed the differencing and ARIMA and SARIMA forecasting and calculated all the RMSE for comparison and selecting the best forecasting Models.

Inferences:

i. Sales Peaked every year at the last quarter specially in December. The least sales recorded in the month of June.
ii. Sales was lowest in the year of 1995 among all the years. This could be due to the fact that data of 1995 was collected till July.
iii. The data follows no clear trend in sales although it had a clear seasonality to the data.
iv. Highest sales of Sparkling wine was recorded in the month of December 1987 at 7242.

Suggestions:

i. The sales have evidently decreased of the years and the company should try and investigate the reason behind this.
ii. The company should come up with promotional offers to increase the sales in the first half of the year when the sales are the lowest.
iii. Historically the sales are highest in the last half of the year so the company should prepare for the upcoming sales season in 1995.
iv. The company should focus on sales more in the month of June when the sales are the lowest in a year.

ROSE WINE DATASET

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

After importing the necessary libraries, we will load the Rose wine data set and below are the first few rows of the dataset. We have also converted the Year Month column into a timestamp and index of the data frame using 'Parse Dates' function.

	Rose
YearMonth	
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0

Table 15: Rose wine sales data

Next we will check the info of the data frame:

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Rose    185 non-null    float64
dtypes: float64(1)
memory usage: 2.9 KB
```

From the above output we can derive the following

i. The data frame consists of 185 observations
ii. The data frame has 2 missing values
iii. The date time index ranges from 01-01-1980 to 09-07-1995
iv. The 'Rose' columns has wine sales values and it is of integer datatype.

Next, we will perform treatment of missing values.

Next, we will plot the data. The below data shows the time plot of the dataset;

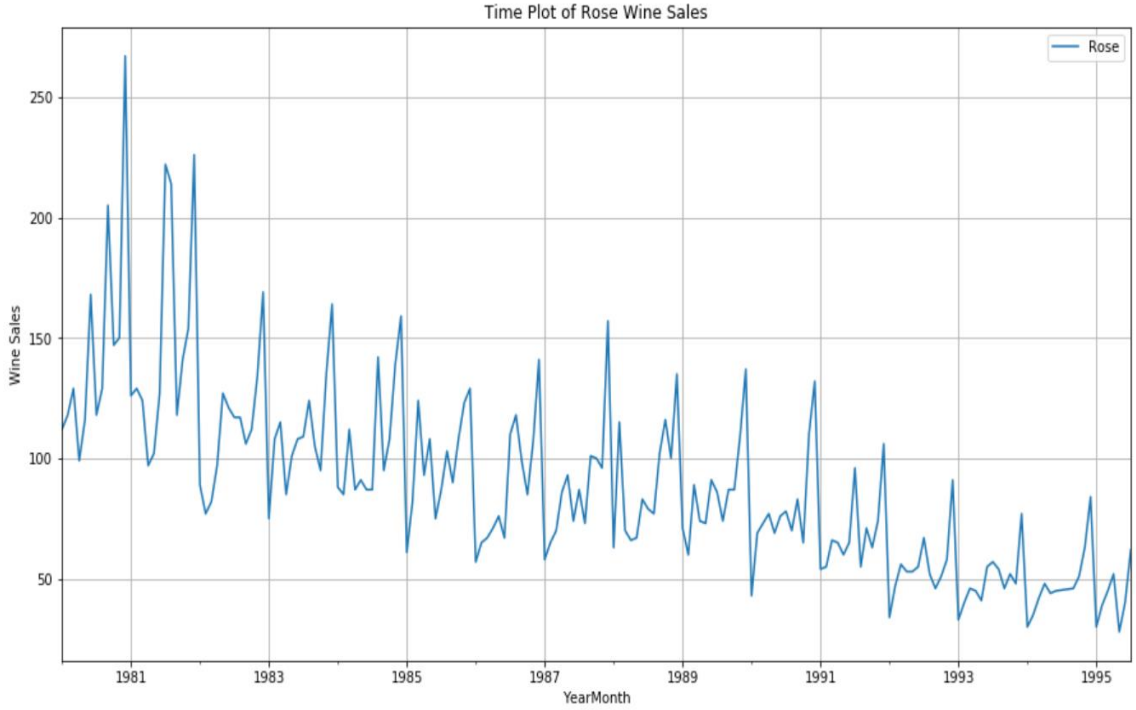

Figure 29: Time plot of Rose wine sales

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

Below shows the 5 point summary of the dataset:

	Rose
count	187.000000
mean	89.914439
std	39.238325
min	28.000000
25%	62.500000
50%	85.000000
75%	111.000000
max	267.000000

Table 16: 5point summary of the data

The mean of the dataset is 89.914 with a standard deviation of 39.238. The dataset is ranged from 28 to 267. There are no missing values.

Below is the Histplot distribution of the dataset:

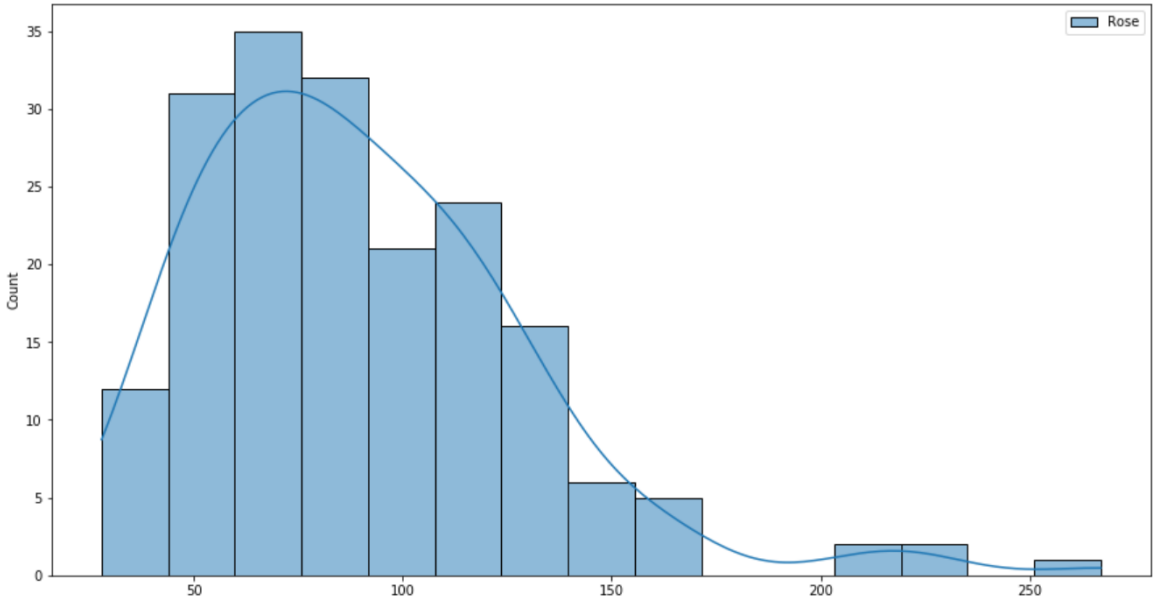

Figure 30: Histplot Distribution of the dataset

Below is the yearly plot of the dataset:

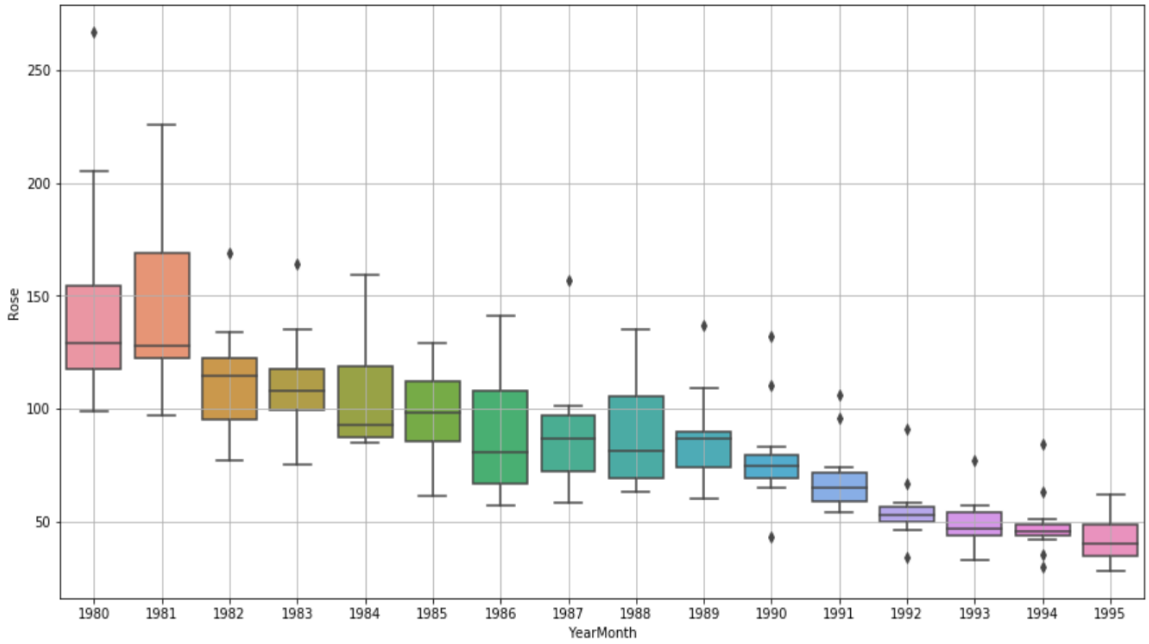

Figure 31: Yearly plot of the dataset

From the above plot we can see that the wine sales has increased in 1980 and 1981 and the there is a decrease in the sales from 1983 to 1995. The sales has clearly decreased over the years. There are some years with good sales like 1986 and 1988 but the rest have decreased in sales.

Below is the monthly plot of the dataset:

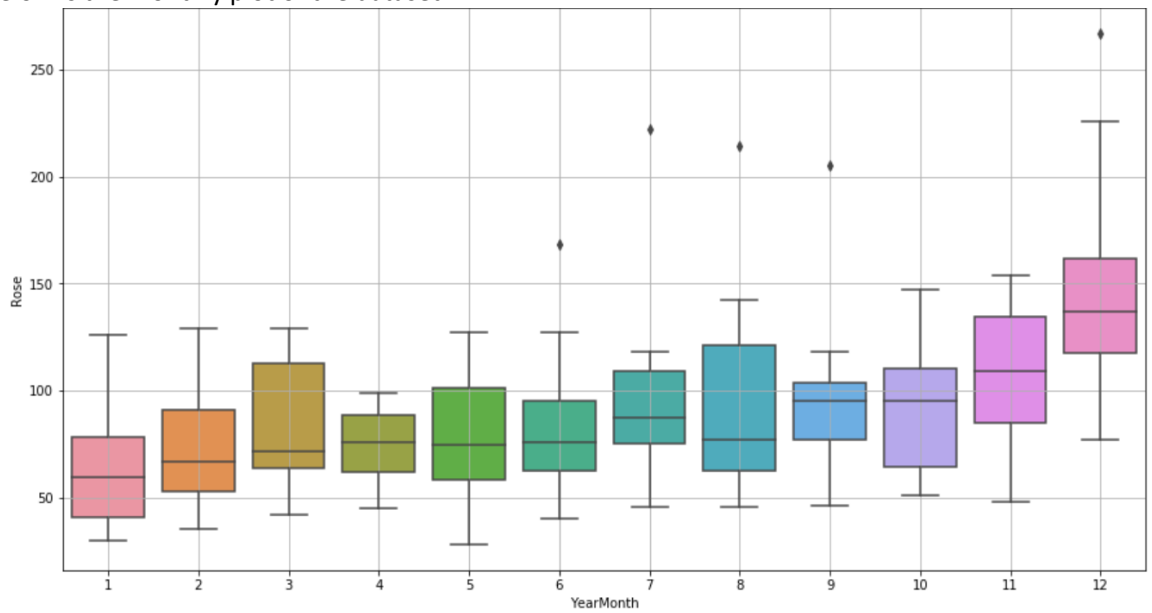

Figure 32: Monthly plot of the dataset

In the month wise dataset we see that wine sales are at highest in the month of December and the lowest in the month of April. Wine sales are particularly low in the first half of the year and high in the last half.

Monthly plot of the data:

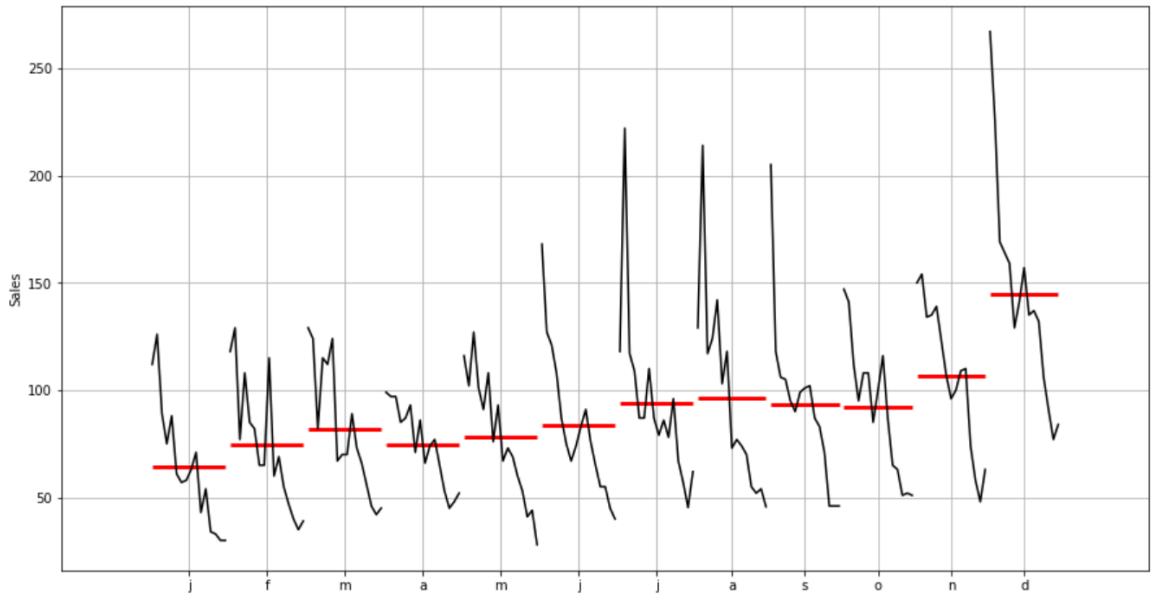

Figure 33: Month wise plot of the dataset

Monthly wise sales across years:

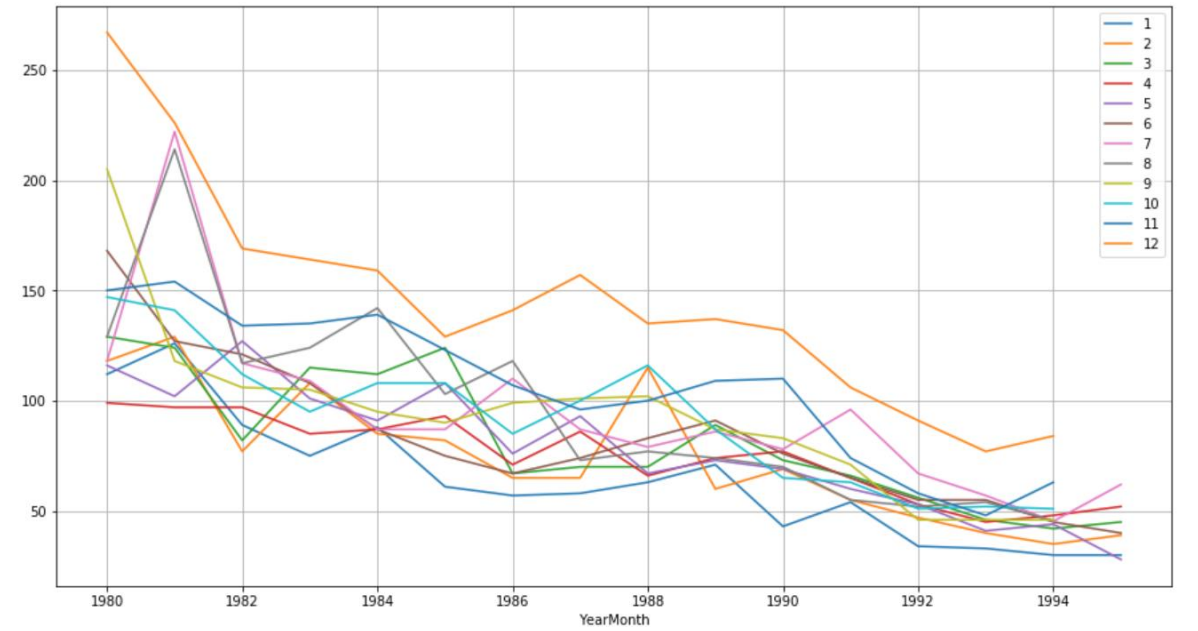

Figure 34: Monthly sales across years

The month wise sales plot shows a clear decrease in sales over the years.

Yearly average sales and percentage change of wine sales:

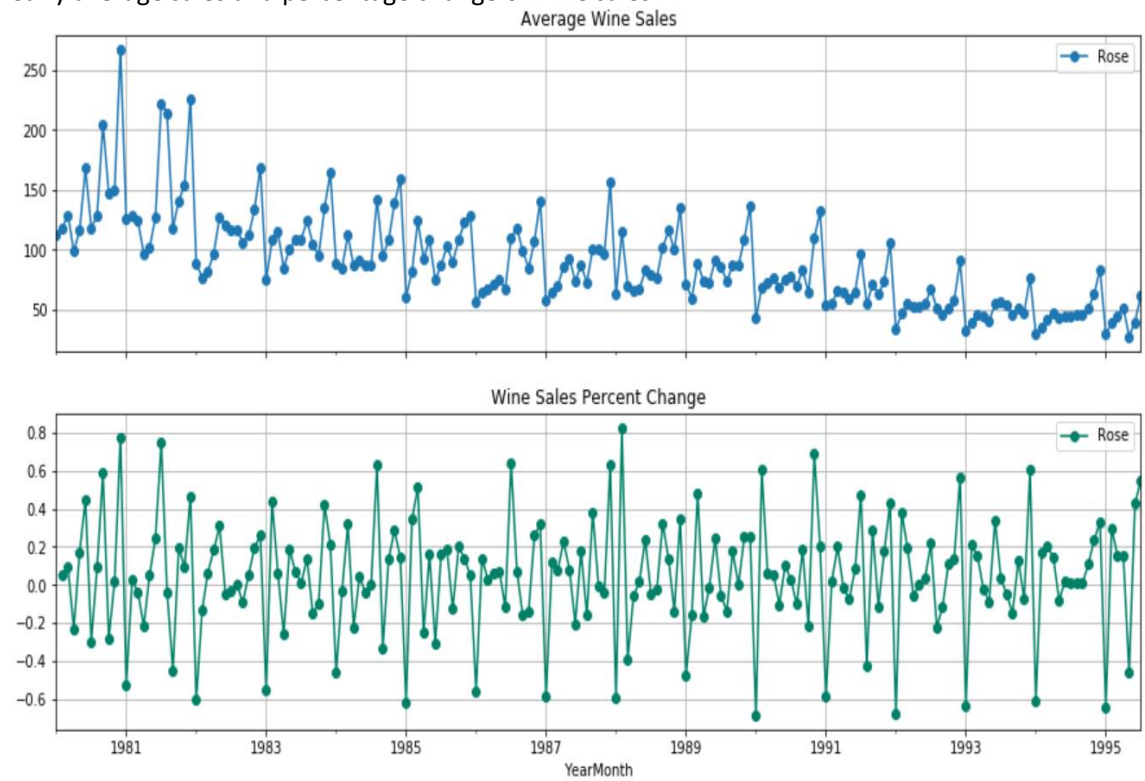

Figure 35: Average Sales and Percentage change

Decomposition:

The Data set is decomposed using additive and multiplicative model:

i. MULTIPLICATIVE:

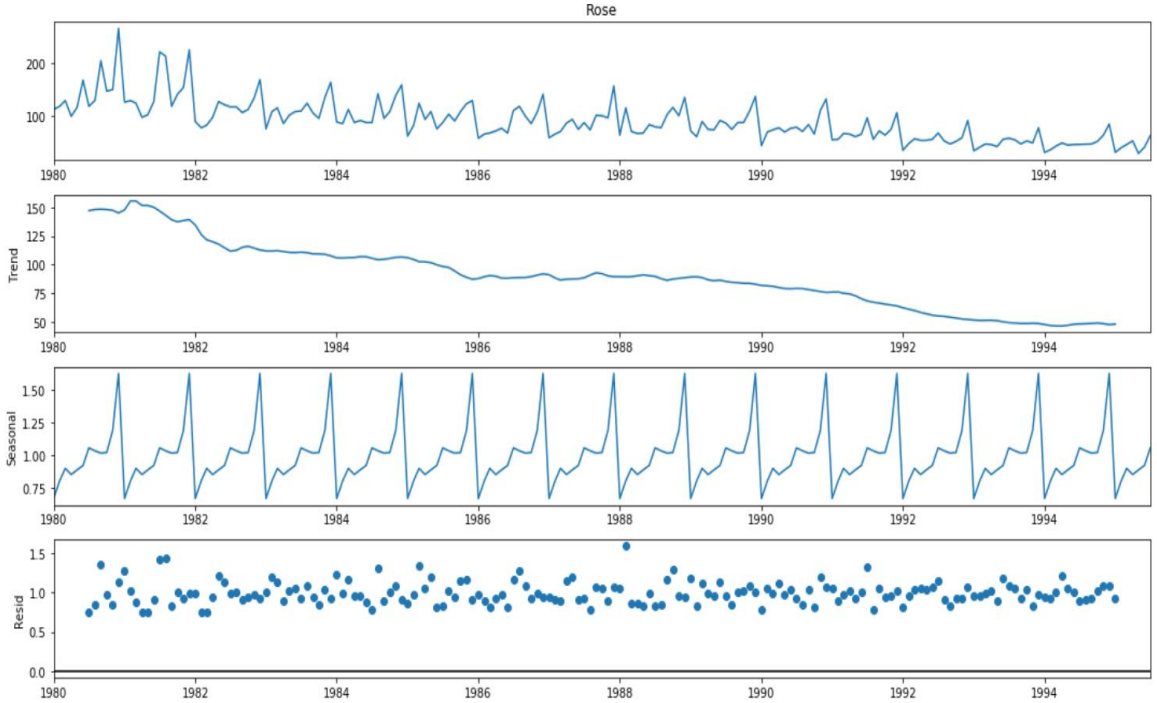

Figure 36: Multiplicative Distribution of the data

ii. ADDITIVE

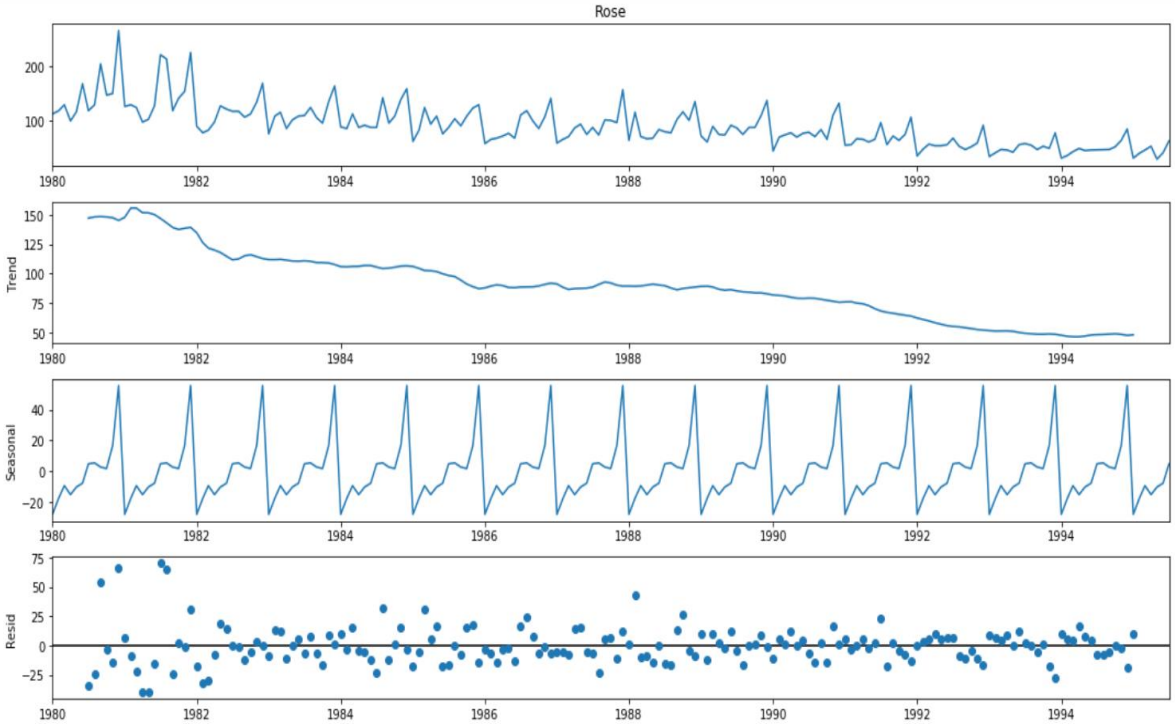

Figure 37: Additive decomposition of the data

From the decomposition plot we can see that there is a decreasing trend in the sales of Rose wine. Also a clear seasonality is visible in the dataset indicating that sales are high in the last half of the year and sales are low in the first half of the year. Also, there is a certain pattern in the residual plot indicating that not all seasonal functions are captured by the model.

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

The test data should contain the recent years and hence, for the split we will consider the data from 1991 as test data and data before that as train data,

The shape of the train and test dataset is as follows:

```
Shape of the training data is

(132, 1)

Shape of the test data is

(55, 1)
```

The first and last few rows of the train and test dataset are as follows:

First few rows of the training data is

YearMonth	Rose
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0

Last few rows of the training data is

YearMonth	Rose
1990-08-01	70.0
1990-09-01	83.0
1990-10-01	65.0
1990-11-01	110.0
1990-12-01	132.0

First Few rows of the test data is

YearMonth	Rose
1991-01-01	54.0
1991-02-01	55.0
1991-03-01	66.0
1991-04-01	65.0
1991-05-01	60.0

Last Few rows of the test data is

YearMonth	Rose
1995-03-01	45.0
1995-04-01	52.0
1995-05-01	28.0
1995-06-01	40.0
1995-07-01	62.0

Table 17: First and last few rows of train and test data

The plot of the training and test data are as follows:

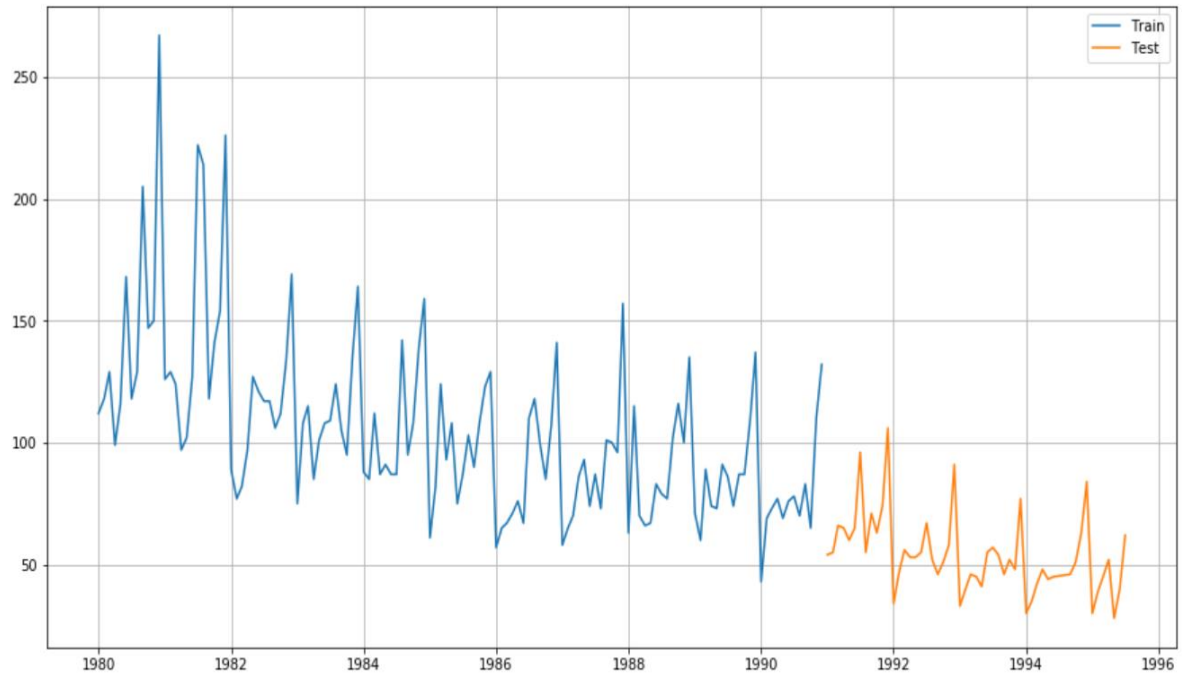

Figure 38: Train and Test data plot

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

i. Naïve Forecast:

Naïve Forecast is an Estimating technique in which the last period's actuals are used as this period's forecast, without adjusting them or attempting to establish causal factors.

The below table shows the Naïve forecast on the test data:

YearMonth	Rose	Naive
1991-01-01	54.0	132.0
1991-02-01	55.0	132.0
1991-03-01	66.0	132.0
1991-04-01	65.0	132.0
1991-05-01	60.0	132.0

Table 18: Naïve Forecast

The forecast of Naïve on Test data is:

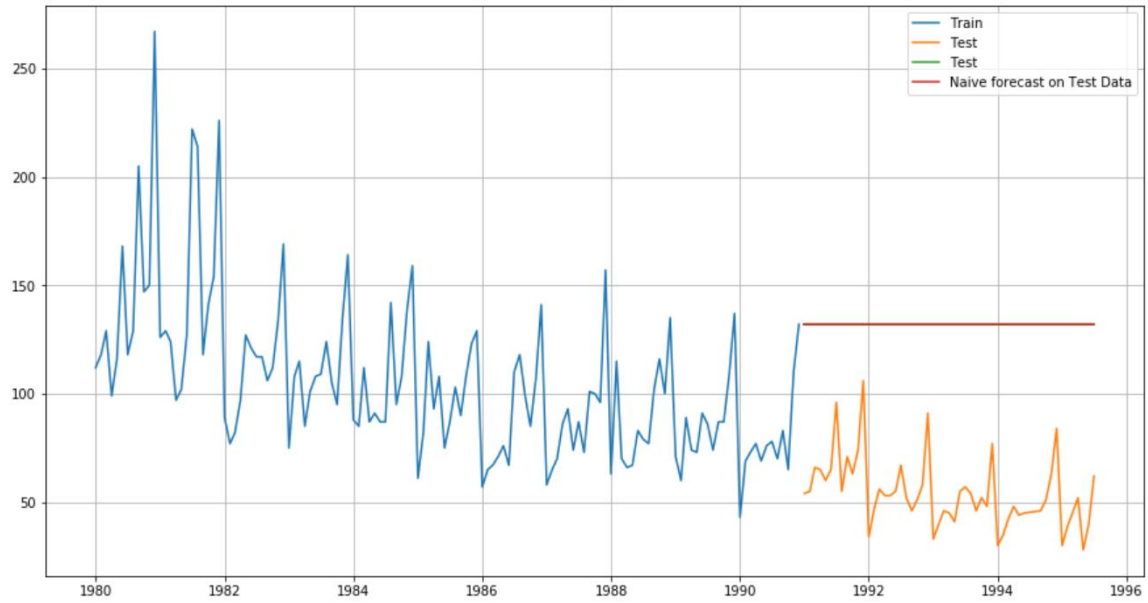

Figure 39: Naïve forecast plot on test data

The Naïve forecast is a very basic model and produces a very straight line. It does not takes into account the trend or the seasonality. We calculate the RMSE for the model which is as follows:

```
RMSE for Naive forecast model on Rose wine data is 79.719
```

ii. Linear Regression

For a linear regression model, the data must contain at least one predictor variable. The Rose data only contains a target variable which is the sales of the Rose wine. Therefore, we will have to create a new predictor variable to build the model. Hence, we will regress the Rose column with the order of occurrence of the values. The below tables show the training and test set after adding the new predictor variables:

Training data:

YearMonth	Rose	Time
1980-01-01	112.0	1
1980-02-01	118.0	2
1980-03-01	129.0	3
1980-04-01	99.0	4
1980-05-01	116.0	5

YearMonth	Rose	Time
1990-08-01	70.0	128
1990-09-01	83.0	129
1990-10-01	65.0	130
1990-11-01	110.0	131
1990-12-01	132.0	132

Test Data:

YearMonth	Rose	Time
1991-01-01	54.0	133
1991-02-01	55.0	134
1991-03-01	66.0	135
1991-04-01	65.0	136
1991-05-01	60.0	137

YearMonth	Rose	Time
1995-03-01	45.0	183
1995-04-01	52.0	184
1995-05-01	28.0	185
1995-06-01	40.0	186
1995-07-01	62.0	187

Plot of the Linear regression on test data:

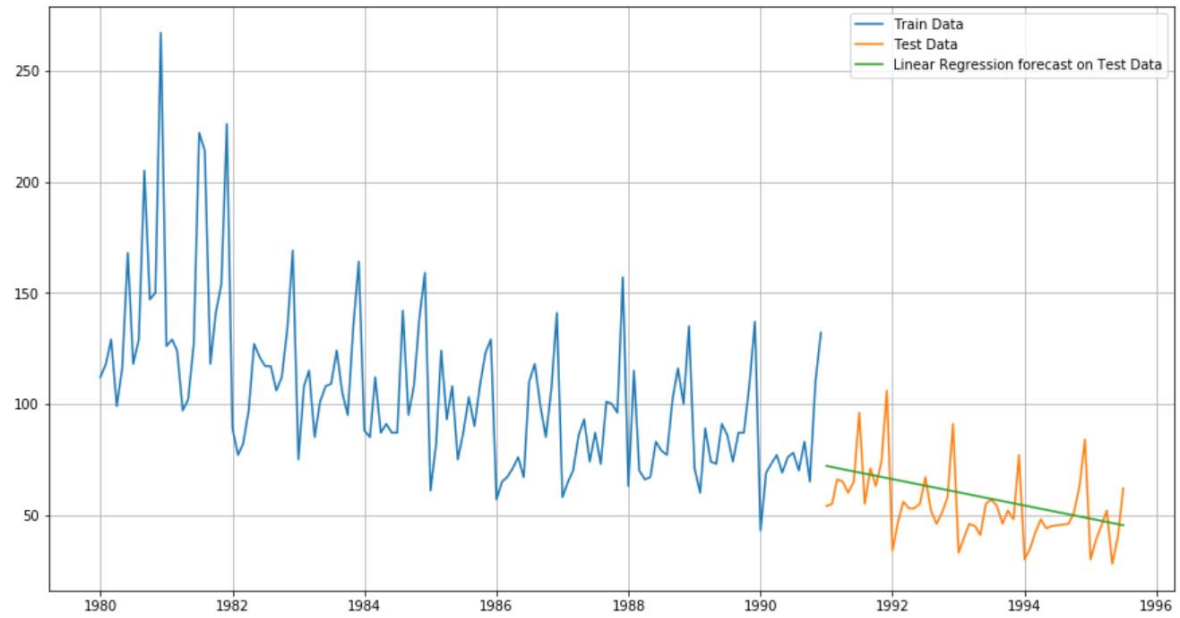

Figure 40: Linear Regression plot on test data

The RMSE of the model is as follows:

```
RMSE for Linear Regression forecast model on Rose wine data is 15.269
```

iii. Simple Average Forecast:

In a Simple Average forecasting technique, the mean of the training data is used as the forecast for all the periods in the test data. Like Naïve forecast this is also a very basic forecasting technique and it doesnot take into account the trend or seasonality. The below table shows the forecast for the Simple Average model:

YearMonth	Rose	Average
1991-01-01	54.0	104.939394
1991-02-01	55.0	104.939394
1991-03-01	66.0	104.939394
1991-04-01	65.0	104.939394
1991-05-01	60.0	104.939394

Table 19: Simple Average forecast on test data

Plot of Simple Average forecast on test data:

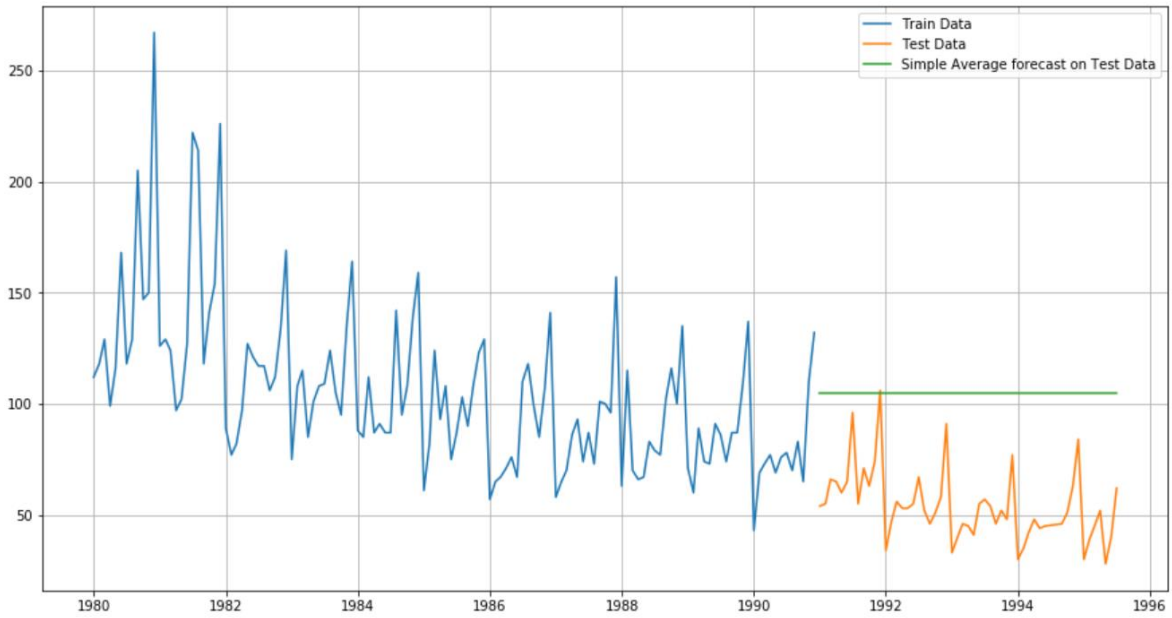

Figure 41: Simple average forecast on test data

The RMSE of the model is as follows:

```
RMSE for Simple Average forecast model on Rose wine data is 53.461
```

iv. Moving Average Forecast:

A moving average is a technique that calculates the overall trend in a data set. This technique is very useful for forecasting short-term trends. It is simply the average of a select set of time periods. The moving averages smoothens the seasonal fluctuations in the time series data. The forecasts for first k-1 periods are null values as k number of data points are required to find the averages. Therefore, we will find various moving averages for the complete data and then split the data into training and test sets so that the test set does not contain any null values. We will calculate 2, 4, 6, and 9 point moving averages for the entire data.

Moving Average in entire data:

YearMonth	Rose	2 point moving average	4 point moving average	6 point moving average	9 point moving average
1980-01-01	112.0	NaN	NaN	NaN	NaN
1980-02-01	118.0	115.0	NaN	NaN	NaN
1980-03-01	129.0	123.5	NaN	NaN	NaN
1980-04-01	99.0	114.0	114.5	NaN	NaN
1980-05-01	116.0	107.5	115.5	NaN	NaN

Table 20: 2point, 4point, 6point & 9point moving average on whole data

Moving Average on Test Data:

YearMonth	Rose	2 point moving average	4 point moving average	6 point moving average	9 point moving average
1991-01-01	54.0	93.0	90.25	85.666667	81.888889
1991-02-01	55.0	54.5	87.75	83.166667	80.333333
1991-03-01	66.0	60.5	76.75	80.333333	79.222222
1991-04-01	65.0	65.5	60.00	80.333333	77.777778
1991-05-01	60.0	62.5	61.50	72.000000	76.666667

Table 21: 2point, 4point, 6point & 9point moving average on test data

The smoothening effect produced by the moving averages depend on the window of rolling mean. Larger the window more is the smoothening effect. This can be confirmed from the below figure

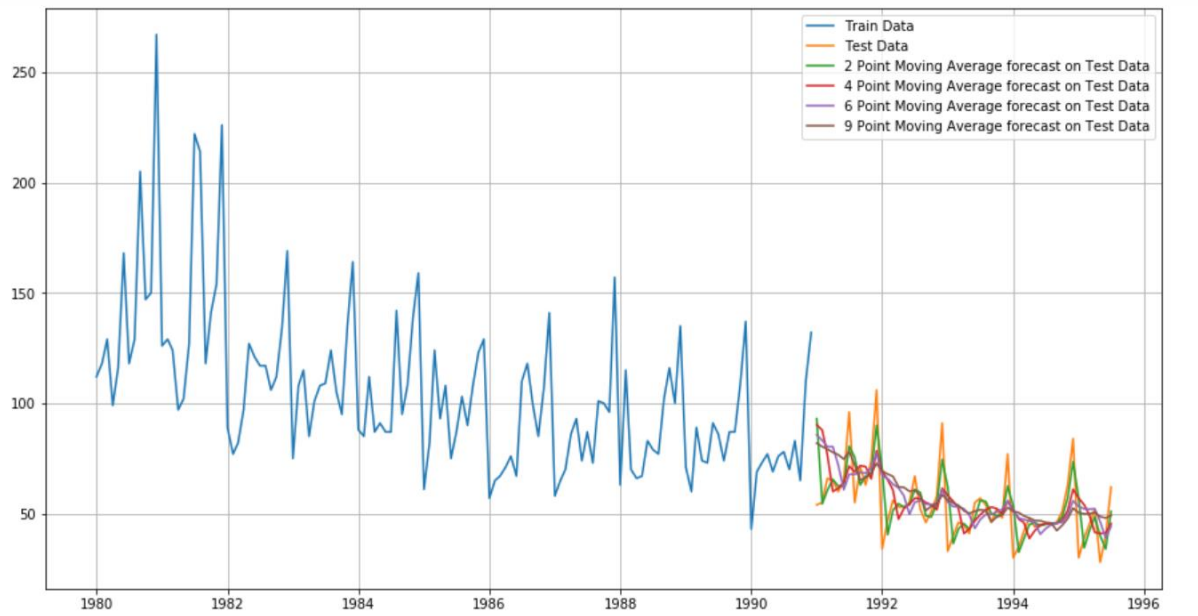

Figure 42: 2point, 4point, 6point, 9point moving average plot on test data

The RMSE for Moving Average forecast is as follows:

```
RMSE for 2 point Moving Average forecast model on Rose wine data is 11.529
RMSE for 4 point Moving Average forecast model on Rose wine data is 14.451
RMSE for 6 point Moving Average forecast model on Rose wine data is 14.566
RMSE for 9 point Moving Average forecast model on Rose wine data is 14.728
```

Since the 2 Point moving average is the closest to the test data, it has the lowest RMSE value.

v. Simple Exponential Smoothing

A simple exponential smoothing is one of the simplest ways to forecast a time series. The basic idea of this model is to assume that the future will be more or less the same as the (recent) past. Thus, the only pattern that this model will learn from demand history is its level.
It is a time series forecasting method for univariate data without a trend or seasonality. It requires a single parameter, called alpha, also called the smoothing factor or smoothing coefficient. The following are the parameters of the SES Model:

```
{'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}
```

Here, the smoothing trend and smoothing seasonal parameters are NaN because we are using Simple Exponential Smoothing model. The below table shows the predictions done:

YearMonth	Rose	Prediction
1991-01-01	54.0	87.104995
1991-02-01	55.0	87.104995
1991-03-01	66.0	87.104995
1991-04-01	65.0	87.104995
1991-05-01	60.0	87.104995

Table 22: Simple Exponential Smoothing prediction on test data

The plot of the Simple exponential smoothing at alpha=0.09 is as follows:

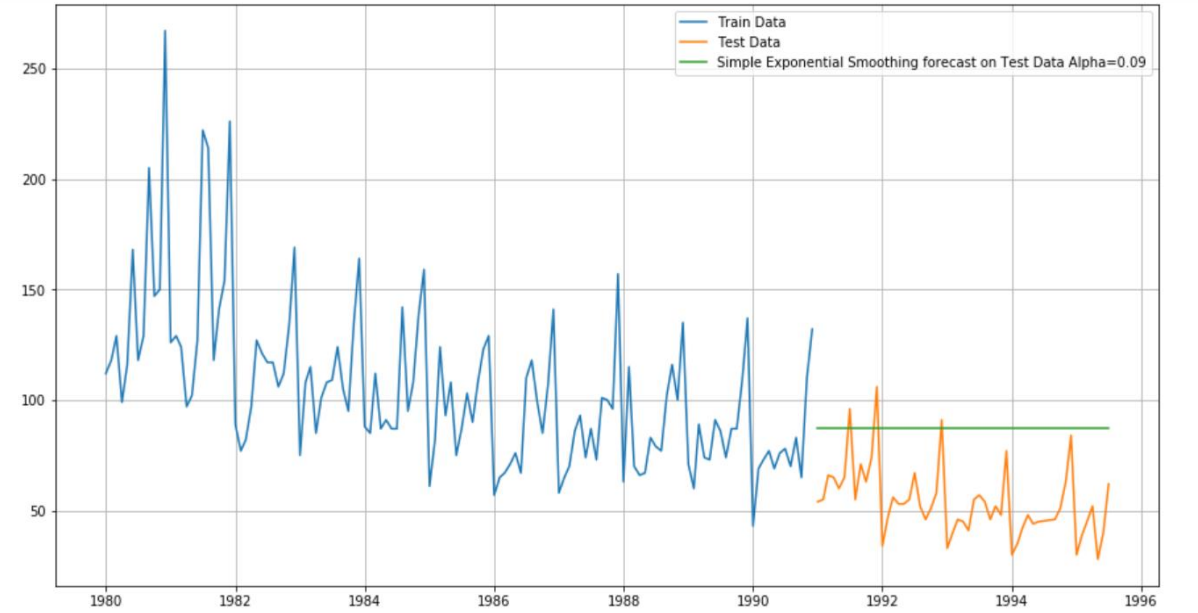

Figure 43: Simple Exponential Smoothing forecast on test data

From the above figure we see that the forecast done by the Simple Exponential Smoothing model is a straight line. This is due the fact that SES model is only used for data which does not have any trend or seasonality.
The RMSE of the model is as follows:

```
RMSE for SES forecast model on Rose wine data is 36.796
```

vi. Double Exponential Smoothing:

Double exponential smoothing employs a level component and a trend component at each period. Double exponential smoothing uses two weights, (also called smoothing parameters), to update the components at each period. This method is also called as Holt's trend corrected or second-order exponential smoothing. This method is used for forecasting the time series when the data has a linear trend and no seasonal pattern.

The following are the parameters of the Double Exponential Smoothing:

```
{'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}
```

The smoothing level parameter is the alpha value while the smoothing trend parameter is the beta value. The model found the optimal alpha value to be 0.15 and beta value as 0.15. The below table shows the predictions done by the model:

YearMonth	Rose	Prediction
1991-01-01	54.0	86.863579
1991-02-01	55.0	88.028056
1991-03-01	66.0	89.192534
1991-04-01	65.0	90.357011
1991-05-01	60.0	91.521488

Table 23: Double Exponential Smoothing prediction on test data

The Plot of the forecast is as follows:

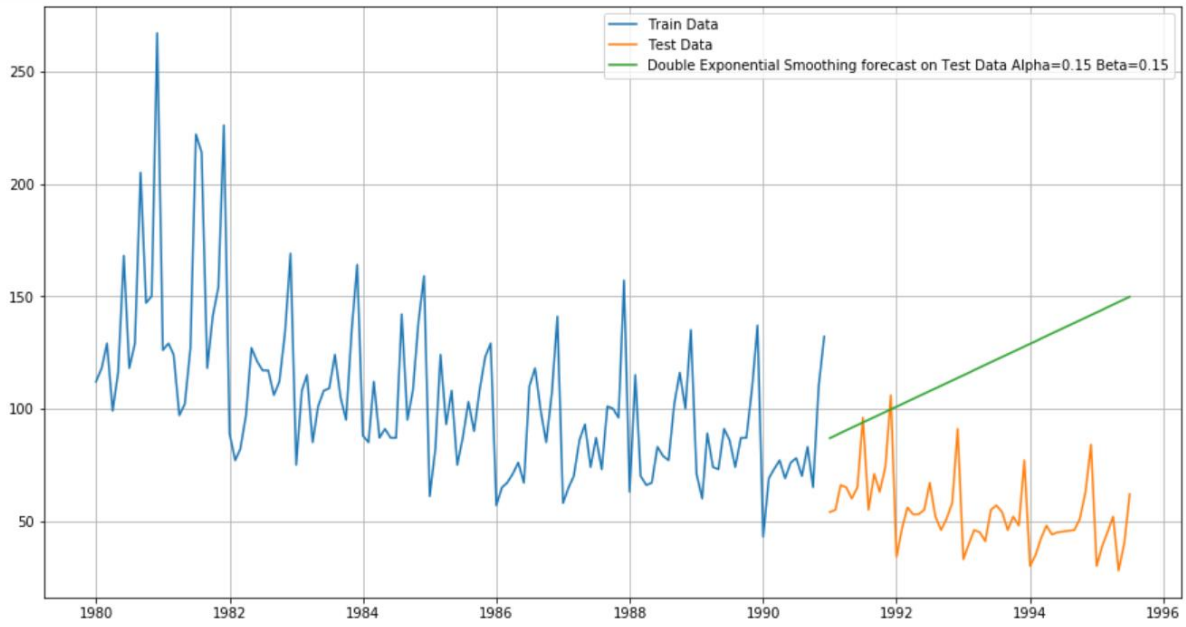

Figure 44: Double Exponential Smoothing Forecast on test data

The RMSE of the model is:

```
RMSE for DES forecast model on Rose wine data is 70.572
```

vii. Triple Exponential Smoothing(Holt-Winters Model)

Triple exponential smoothing (Holt-Winters model) employs level, trend and seasonal components at each period. Holt-Winters model is used for data that has both trend and seasonality components. Triple exponential smoothing uses three smoothing parameters, alpha, beta and gamma, to update the components at each period. Alpha is the level smoothing parameter, beta is the trend smoothing parameter and gamma is the seasonal smoothing parameter. All of these parameters range between 0 and 1. Large values means that the model pays attention mainly to the most recent past

observations, whereas smaller values mean more of the history is taken into account when making a prediction.
Following are the parameters of Triple Exponential smoothing:

```
{'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}
```

The smoothing level parameter is the alpha value, the smoothing trend parameter is the beta value and the smoothing seasonal parameter is the gamma value. The model found the optimal alpha value to be 0.13, beta value as 0.013 and gamma value to be 0. The below table shows the predictions done by the model:

YearMonth	Rose	Prediction
1991-01-01	54.0	44.139564
1991-02-01	55.0	56.077991
1991-03-01	66.0	63.613884
1991-04-01	65.0	52.575891
1991-05-01	60.0	60.938966

Table 24: Triple Exponential Smoothing prediction on test data

The plot of the forecast is as follows:

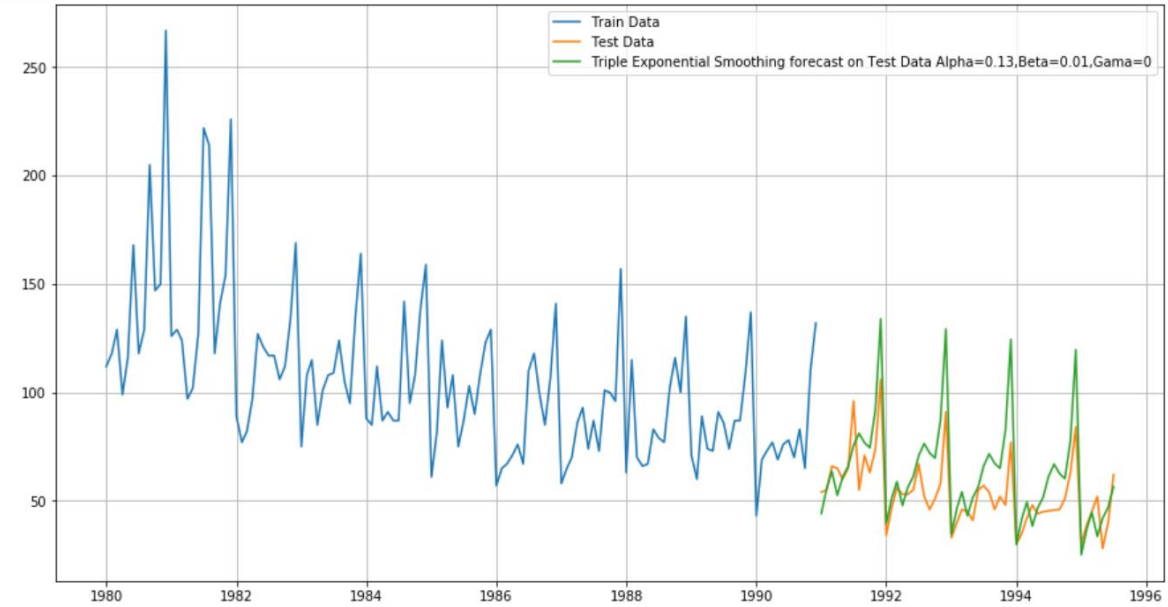

Figure 45: Triple Exponential Smoothing Forecast on test data

The RMSE of the model is as follows:

```
RMSE for TES forecast model on Rose wine data is 16.443
```

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.

A stationary time series is one whose properties do not depend on the time at which the series is observed. Thus, time series with trends, or with seasonality, are not stationary — the trend and seasonality will affect the value of the time series at different times.
Time series are stationary if they do not have trend or seasonal effects. Summary statistics calculated on the time series are consistent over time, like the mean or the variance of the observations.

There are various statistical tests to check stationarity, including the Augmented Dickey-Fuller (ADF) test. The ADF test is a widely used test for checking the stationarity of a time series, and it checks for the presence of a unit root in the data. It is a hypothesis test with the null and alternate hypothesis as follows:

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.

The following is the output of the ADF Test:

```
Results of Dickey-Fuller Test:
Test Statistic                  -1.876699
p-value                          0.343101
#Lags Used                      13.000000
Number of Observations Used    173.000000
Critical Value (1%)             -3.468726
Critical Value (5%)             -2.878396
Critical Value (10%)            -2.575756
dtype: float64
```

The following is the plot of the ADF Test

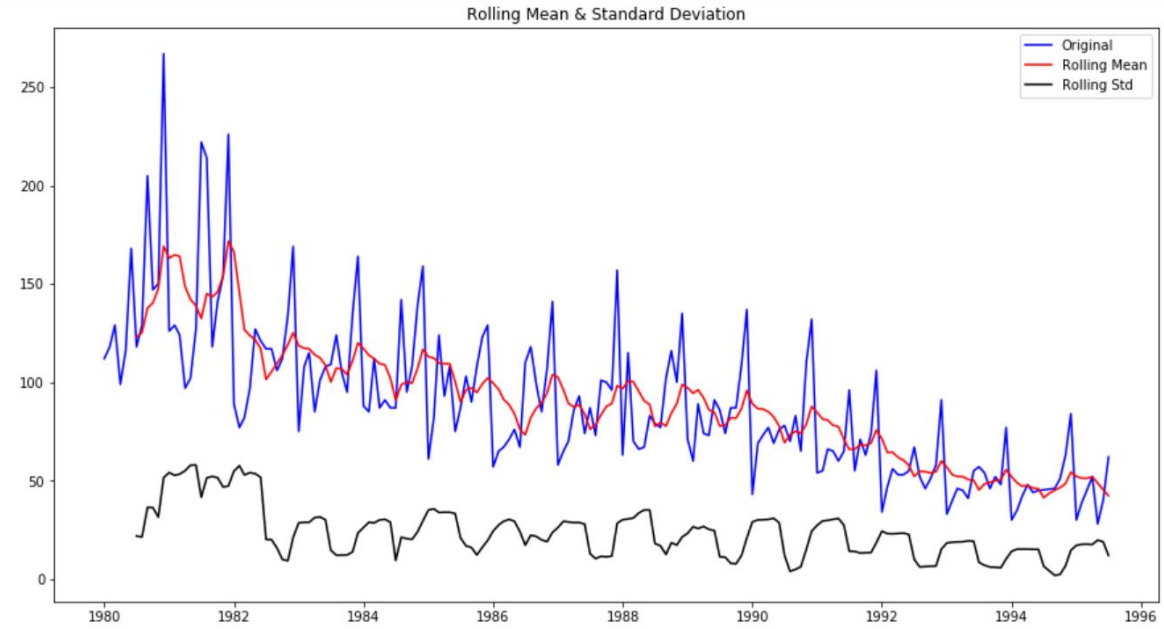

Figure 46: Check for Stationarity on whole data

The p-value from the ADF test is greater than 0.05 and therefore we fail to reject the Null hypothesis. Hence, the Rose data is non-stationary.

Checking for stationarity on the training data:
The following is the output of ADF test on training data:

```
Results of Dickey-Fuller Test:
Test Statistic                  -2.164250
p-value                          0.219476
#Lags Used                      13.000000
Number of Observations Used    118.000000
Critical Value (1%)             -3.487022
Critical Value (5%)             -2.886363
Critical Value (10%)            -2.580009
dtype: float64
```

The plot of is as follows:

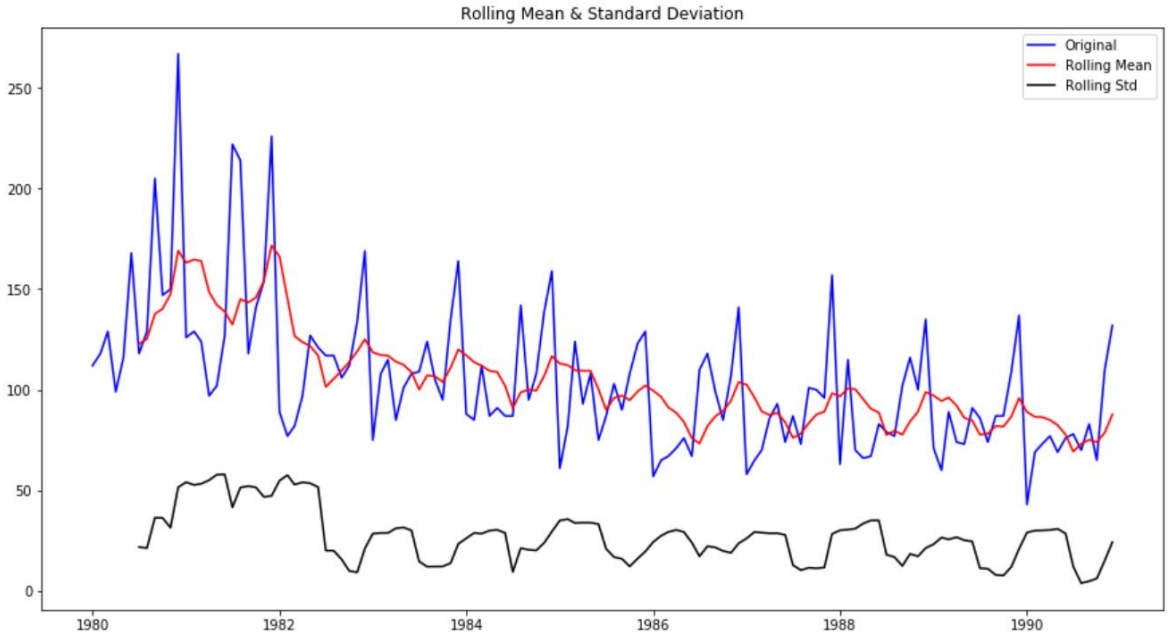

Figure 47: Check for stationarity on training data

The p-value from the ADF test is greater than 0.05 and therefore we fail to reject the Null hypothesis. Hence, the training data is non-stationary. Therefore we have to do differencing.

Differencing and checking for stationarity:

Differencing can help stabilise the mean of a time series by removing changes in the level of a time series, and therefore eliminating (or reducing) trend and seasonality thereby making the series stationary. As our data is a monthly data, one seasonal period is of 12 months.
The result of the ADF test is as follows:

```
Results of Dickey-Fuller Test:
Test Statistic                  -3.692348
p-value                          0.004222
#Lags Used                      11.000000
Number of Observations Used    107.000000
Critical Value (1%)             -3.492996
Critical Value (5%)             -2.888955
Critical Value (10%)            -2.581393
dtype: float64
```

The plot of the test is as follows:

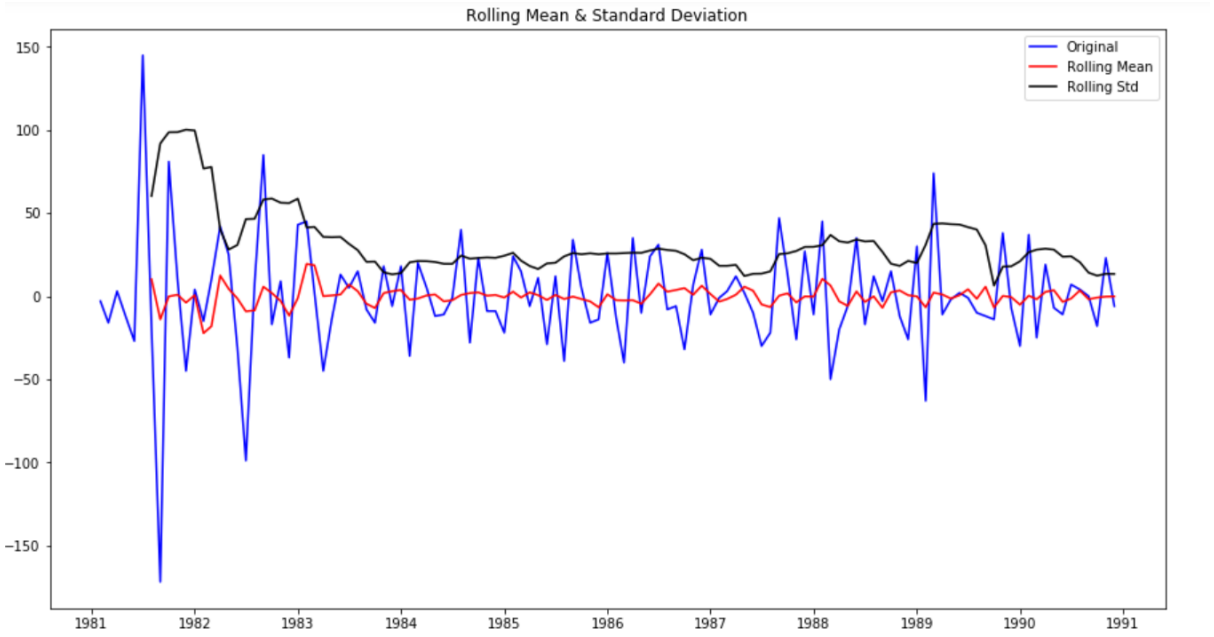

Figure 48: Differencing and check for stationarity on training data

The p-value from the ADF test is less than 0.05 and therefore we reject the Null hypothesis. Therefore, taking the difference of the training data has made the data stationary.

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

ARIMA stands for Auto Regressive Integrated Moving Average model. There are total three parameters related to three different components of the model. The p parameter indicates the lag used in the Auto Regressive component. The d parameter is the order of level differencing applied to make the data stationary. The q parameter is the lag used in the Moving Average component. Here, we are using a value of d = 1. Various combinations of the p, d and q parameter are used to build various ARIMA models and the combination that give the lowest AIC value is used to evaluate the model on the test data.

```
Some parameter combinations for the Model...
Model: (0, 1, 0)
Model: (0, 1, 1)
Model: (0, 1, 2)
Model: (1, 1, 0)
Model: (1, 1, 1)
Model: (1, 1, 2)
Model: (2, 1, 0)
Model: (2, 1, 1)
Model: (2, 1, 2)
```

The Below table shows the AIC values in ascending orders:

	param	AIC
2	(0, 1, 2)	1251.667543
5	(1, 1, 2)	1251.949504
8	(2, 1, 2)	1253.910212
4	(1, 1, 1)	1262.184006
7	(2, 1, 1)	1263.231523
1	(0, 1, 1)	1263.536910
6	(2, 1, 0)	1280.253756
3	(1, 1, 0)	1308.161871
0	(0, 1, 0)	1323.965788

Table 25: ARIMA AIC in ascending order

We will take the p,d,q as 0,1,2. The below is the output which shows the model summary of ARIMA model:

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(0, 1, 2)   Log Likelihood                -636.836
Date:                Sat, 27 May 2023   AIC                           1279.672
Time:                        07:29:28   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
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

The diagnostic plot is as follows:

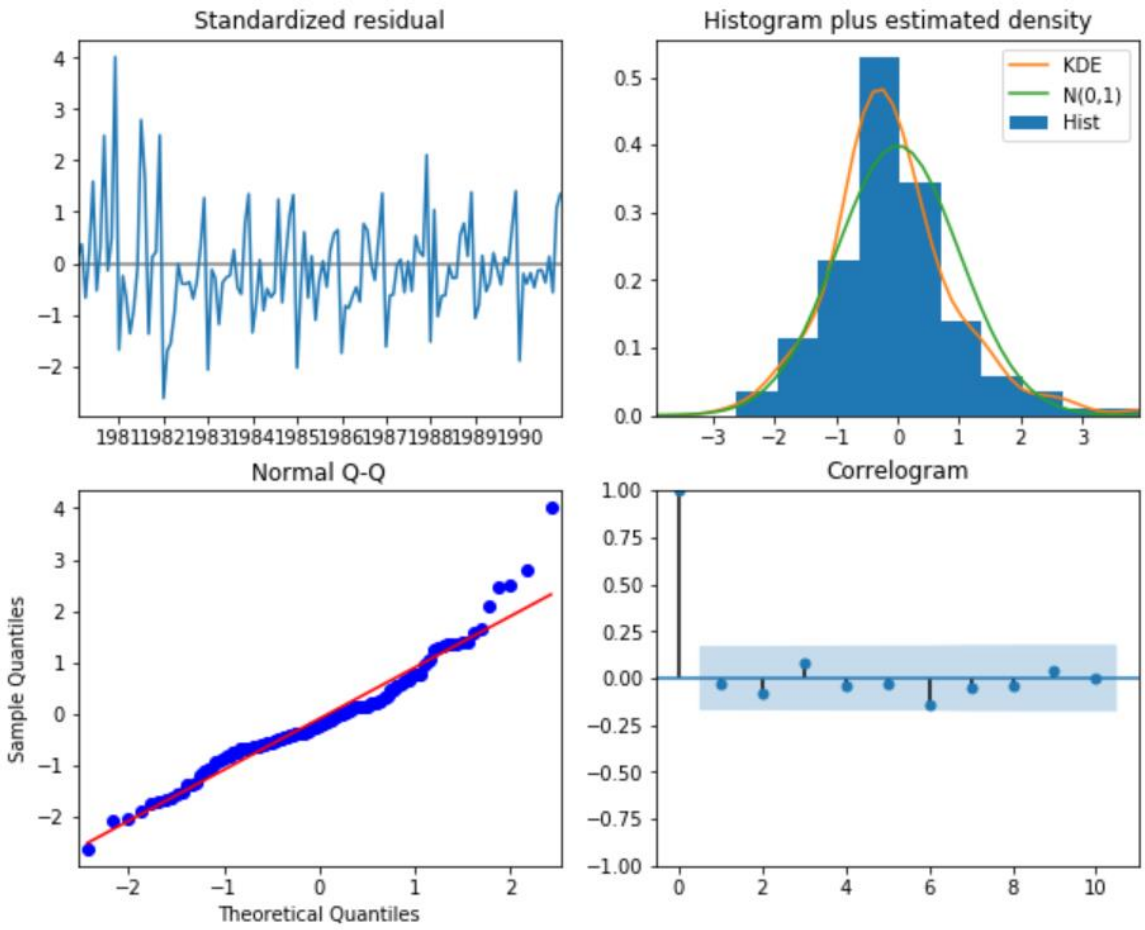

Figure 49: Diagnostic plot ARIMA (0,1,2)

The RMSE for the model is:

```
RMSE for Auto ARIMA forecast model on Rose wine data is 37.306
```

The forecast plot is as follows:

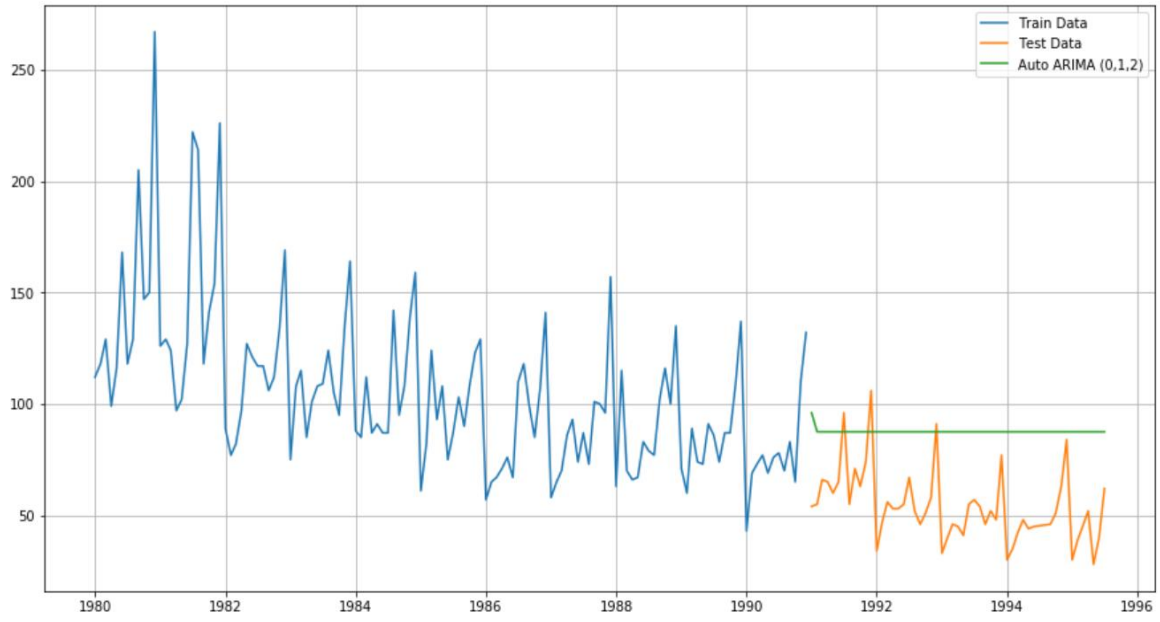

Figure 50: ARIMA(0,1,2) forecast on test data

We will also try p,d,q as (1,1,2). The following is the summary result:

```
                               SARIMAX Results
==========================================================================================
Dep. Variable:                       Rose   No. Observations:                  132
Model:                 ARIMA(1, 1, 2)   Log Likelihood                -635.935
Date:                Sun, 28 May 2023   AIC                           1279.871
Time:                        06:46:34   BIC                           1291.372
Sample:                    01-01-1980   HQIC                          1284.544
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.4540      0.263     -1.723      0.085      -0.970       0.062
ma.L1         -0.2542      0.242     -1.049      0.294      -0.729       0.221
ma.L2         -0.5983      0.201     -2.970      0.003      -0.993      -0.203
sigma2       952.1489     89.279     10.665      0.000     777.165    1127.133
===================================================================================
Ljung-Box (Q):                      112.12   Jarque-Bera (JB):                34.15
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.37   Skew:                             0.79
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.94
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

The diagnostic plot is as follows:

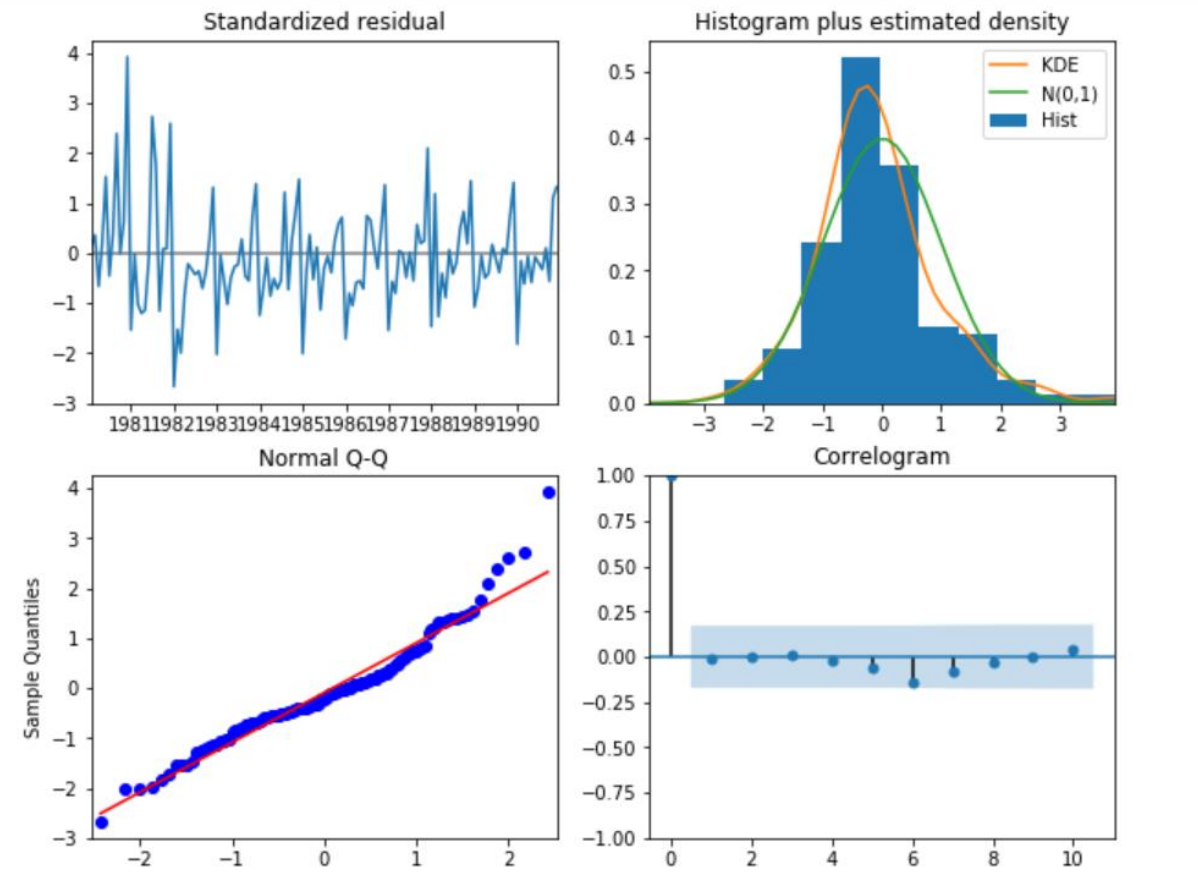

Figure 51: Diagnostic plot ARIMA (1,1,2)

The RMSE of the model is as follows:

```
RMSE for Auto ARIMA forecast model on Rose wine data is 36.871
```

The forecast plot on the test data is as follows:

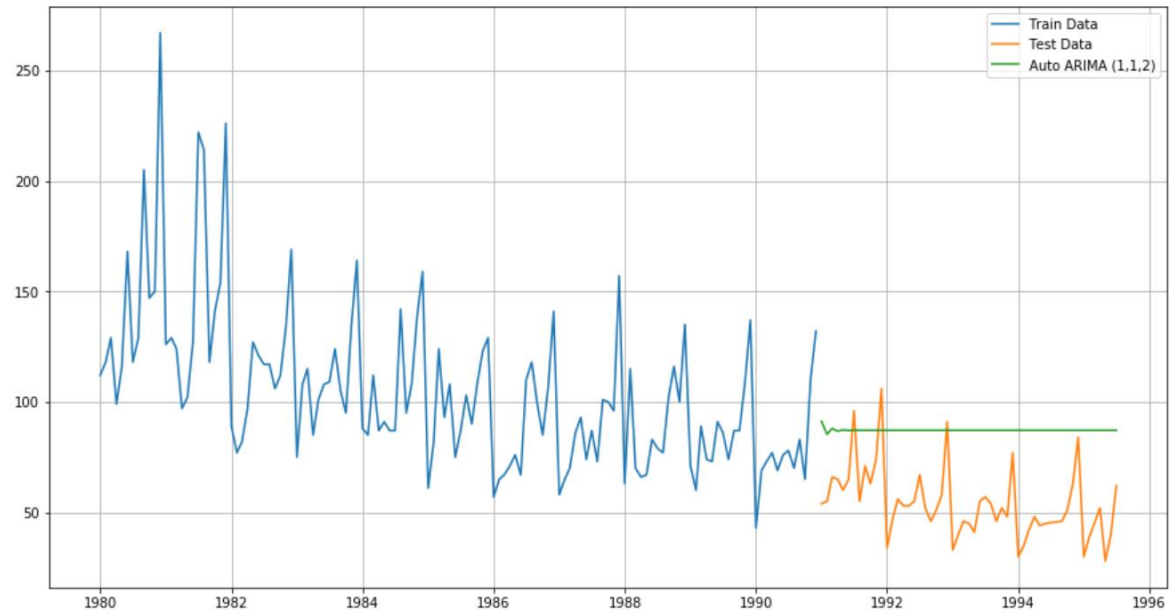

Figure 52: ARIMA (1,1,2) forecast on test data

This seams to be the ideal model as the RMSE is the lowest among the two.

Automated SARIMA

SARIMA stands for Seasonal Auto Regressive Integrated Moving Average model. It is an extension of the ARIMA model. There are total 7 parameters used to define the model. The parameters (p, d, q) are the same as the ARIMA model. The parameters (P, D, Q) are the seasonal counterparts of (p, d, q). The parameter F is the seasonality of the data which is 12 in our case. Here, we are using a value of d = 1 and D = 1. Various combinations of the p, d, q, P, D, and Q parameter are used to build various SARIMA models and the combination that give the lowest AIC value is used to evaluate the model on the test data.
To confirm the seasonality we can also look at the ACF graph. The ACF graph is as follows:

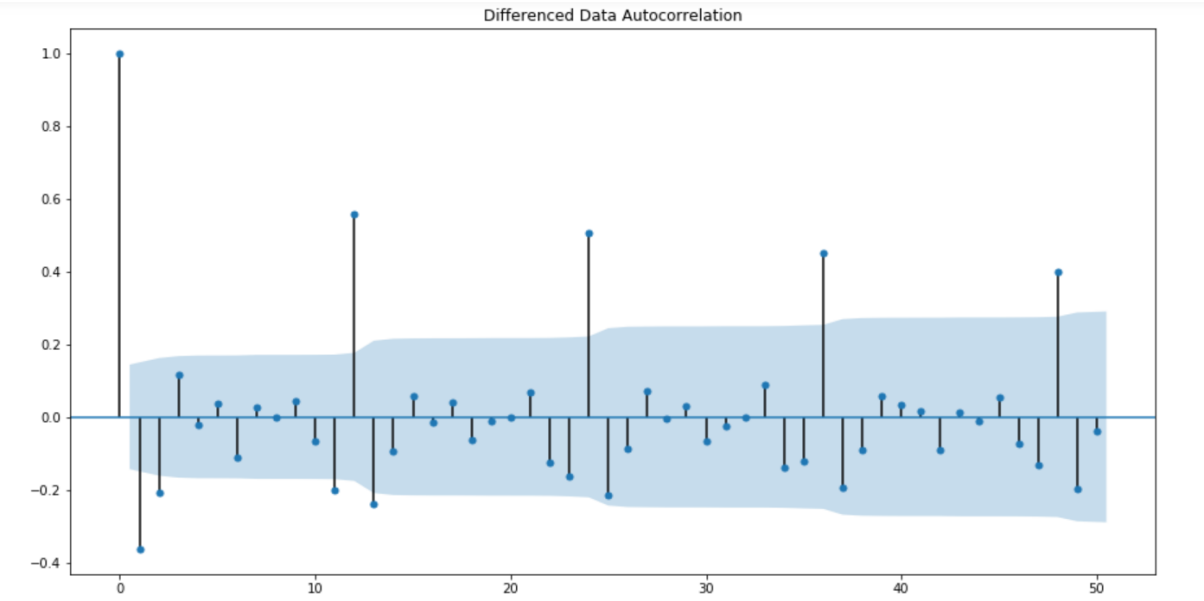

Figure 53: ACF Graph

The Rose wine sales is following a seasonality of 12.

```
Examples of some parameter combinations for Model...
Model: (0, 1, 1)(0, 0, 1, 12)
Model: (0, 1, 2)(0, 0, 2, 12)
Model: (1, 1, 0)(1, 0, 0, 12)
Model: (1, 1, 1)(1, 0, 1, 12)
Model: (1, 1, 2)(1, 0, 2, 12)
Model: (2, 1, 0)(2, 0, 0, 12)
Model: (2, 1, 1)(2, 0, 1, 12)
Model: (2, 1, 2)(2, 0, 2, 12)
```

The below table shows the AIC values of different parameter combinations in ascending order:

	param	seasonal	AIC
26	(0, 1, 2)	(2, 0, 2, 12)	887.937509
53	(1, 1, 2)	(2, 0, 2, 12)	889.903151
80	(2, 1, 2)	(2, 0, 2, 12)	890.668798
69	(2, 1, 1)	(2, 0, 0, 12)	896.518161
78	(2, 1, 2)	(2, 0, 0, 12)	897.346444

Table 26: SARIMA AIC in ascending order

The parameter combination of p=0, d=1, q=2, P=2, D=0, Q=2, and F=12 gives the lowest AIC value. We will use this model for making predictions on the test data. The below output shows the model summary of SARIMA(0, 1, 2)(2, 0, 2, 12) model:

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(0, 1, 2)x(2, 0, 2, 12)	Log Likelihood	-436.969
Date:	Sun, 28 May 2023	AIC	887.938
Time:	06:47:46	BIC	906.448
Sample:	0	HQIC	895.437
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.8427	189.863	-0.004	0.996	-372.967	371.281
ma.L2	-0.1573	29.828	-0.005	0.996	-58.620	58.305
ar.S.L12	0.3467	0.079	4.375	0.000	0.191	0.502
ar.S.L24	0.3023	0.076	3.996	0.000	0.154	0.451
ma.S.L12	0.0767	0.133	0.577	0.564	-0.184	0.337
ma.S.L24	-0.0726	0.146	-0.498	0.618	-0.358	0.213
sigma2	251.3137	4.77e+04	0.005	0.996	-9.33e+04	9.38e+04

Ljung-Box (Q):	24.56	Jarque-Bera (JB):	2.33
Prob(Q):	0.97	Prob(JB):	0.31
Heteroskedasticity (H):	0.88	Skew:	0.37
Prob(H) (two-sided):	0.70	Kurtosis:	3.03

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

The Diagnostic plot is as follows:

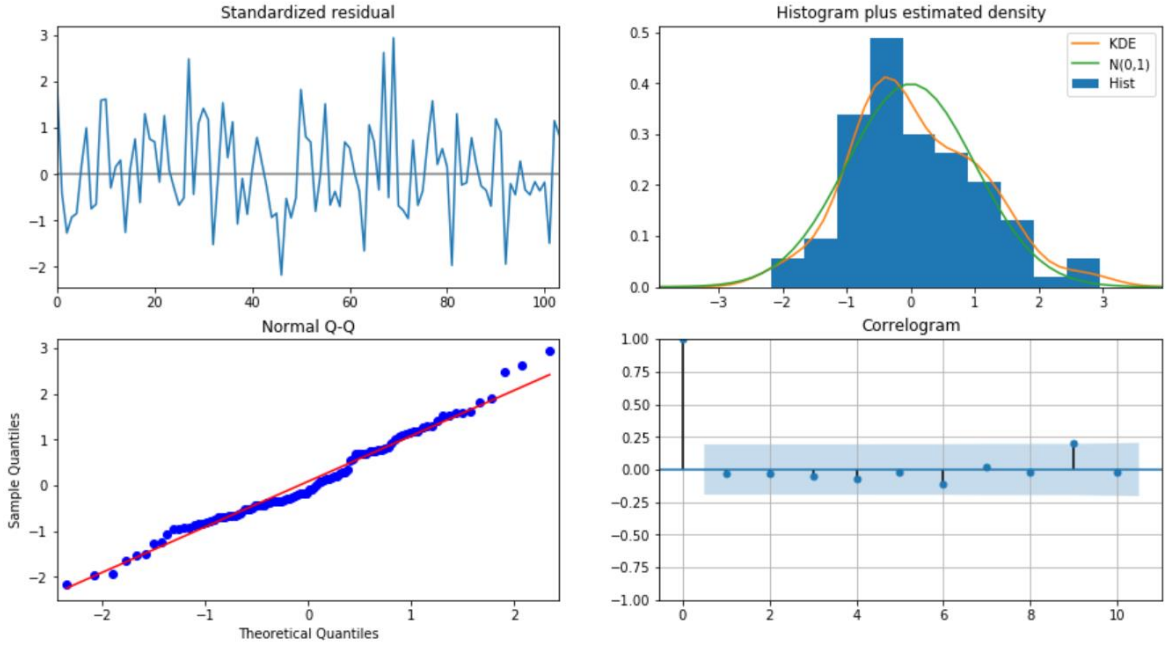

Figure 54: Diagnostic plot SARIMA (0,1,2) (2,0,2,12)

The Predicted Auto SARIMA summary frame is as follows:

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	62.867263	15.928501	31.647976	94.086551
1	70.541190	16.147659	38.892360	102.190020
2	77.356411	16.147656	45.707586	109.005236
3	76.208814	16.147656	44.559989	107.857639
4	72.747398	16.147656	41.098573	104.396223

Table 27: Auto SARIMA summary frame

The RMSE of the SARIMA(0, 1, 2)(2, 0, 2, 12) model is as follows:

```
RMSE for Auto SARIMA forecast model on Rose wine data is 26.928
```

The Plot for the SARIMA model is as follows:

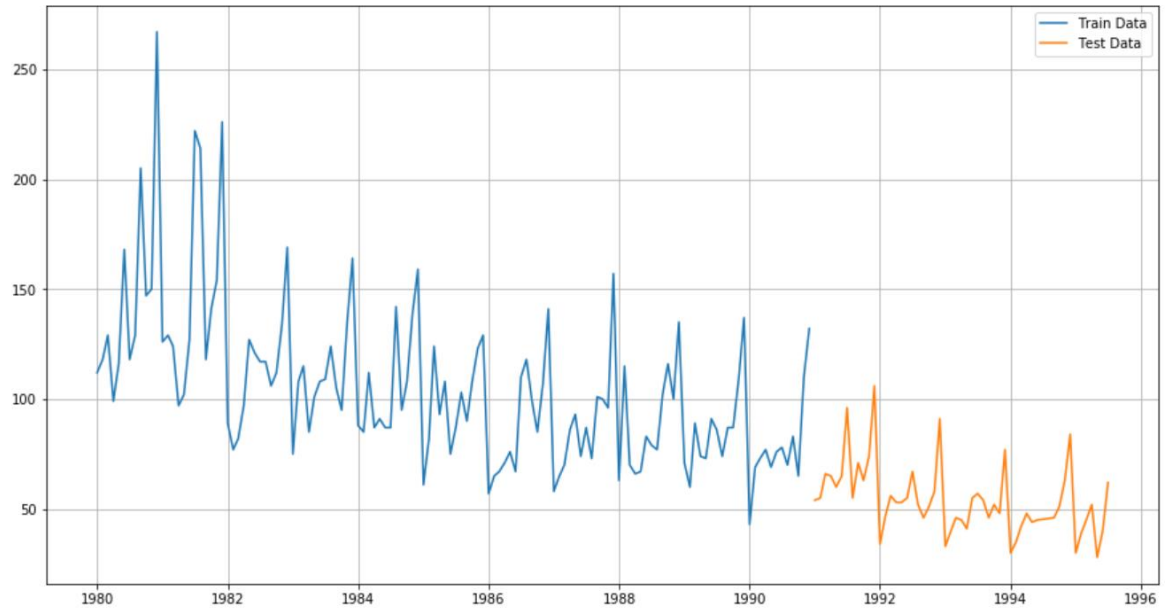

Figure 55: SARIMA(0,1,2)(2,0,2,12)

7. Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

The below table shows the summary of RMSE values of the all the models built above in ascending order:

	Test RMSE
2 Point Moving Average	11.529278
4 Point Moving Average	14.451403
6 Point Moving Average	14.566327
9 Point Moving Average	14.727630
Linear Regression	15.268955
Alpha=0.13, Beta=0.01, Gamma=0, TES	16.443252
Auto SARIMAX(0,1,2)(2,0,2,12)	26.928362
Alpha=0.09, SES	36.796239
Auto ARIMA(1,1,2)	36.870660
Auto ARIMA(0,1,2)	37.306480
Simple Average	53.460570
Alpha=0.15, Beta=0.15, DES	70.572452
Naive Model	79.718773

Table 28: RMSE value of all models in ascending order

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

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

We had the dataset of Rose Wine sales for ABC Estate Wines which contains the wine sales from 1980 to 1995 and we were tasked to analyse and predict the sales for the next 12 months. The dataset was read into a dataframe and proper date time index was created. We saw that there was 2 missing data so we imputed the missing values and advanced to do the exploratory data analysis. We plotted the data on a histplot and then observed that wine sales peaked during the last quarter of the year. The highest sales recorded in the month of December. Then we went ahead and performed additive and multiplicative decomposition of the dataset. We observed that there was a decreasing trend in the dataset and had a clear seasonality. Then we divided the data into train and test to perform the forecast models. We performed the simple Smoothing models for forecast like Naïve, Linear Regression, Simple Exponential Smoothing, Holt(Double Exponential Smoothing), Holt-Winters(Triple Exponential Smoothing) and Moving Average. Then we performed the differencing and ARIMA and SARIMA forecasting and calculated all the RMSE for comparison and selecting the best forecasting Models.

Inferences:

v. Sales Peaked every year at the last quarter specially in December. The least sales recorded in the month of April.
vi. Sales was lowest in the year of 1994 among all the years. Also the sales for the year 1995 is recorded till the month of July.
vii. The data follows a clear decreasing trend in sales and it has a clear seasonality to the data.
viii. Highest sales of Rose wine was recorded in the month of December 1980 at 267.

Suggestions:

v. The sales have evidently decreased over the years and the company should try and investigate the reason behind this.
vi. The company should come up with promotional offers to increase the sales in the first half of the year when the sales are the lowest.
vii. Historically the sales are highest in the last half of the year so the company should prepare for the upcoming sales season in 1995.
viii. The company should focus on sales more in the month of April when the sales are the lowest in a year.