Executive Summary:-

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.

Introduction:-

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset and analysis of using method time series ABC Estate Wines, you are tasked to analyse and forecast Wine Sales in the 20th century. We are also going to analyse results from the different models and build a final model to predict the future sales using the model that provides the best results.

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

Solution:-

The two datasets with Sales data of two types of wine “Rose” and “Sparkling” are imported. We also parse the Year and month information to Date time datatype using parse_dates function.

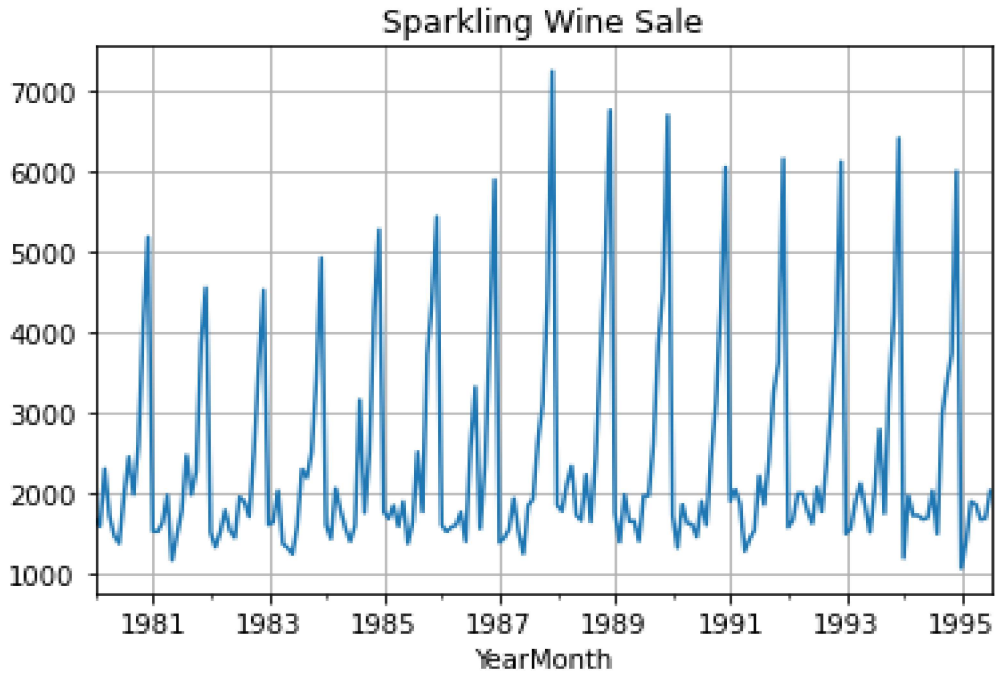


Fig-1.1

Conclusion | insight:

Observations on Sparkling Year wise sales data –

- We notice that there is not much trend in the plot
- The seasonality seems to have a pattern on yearly basis

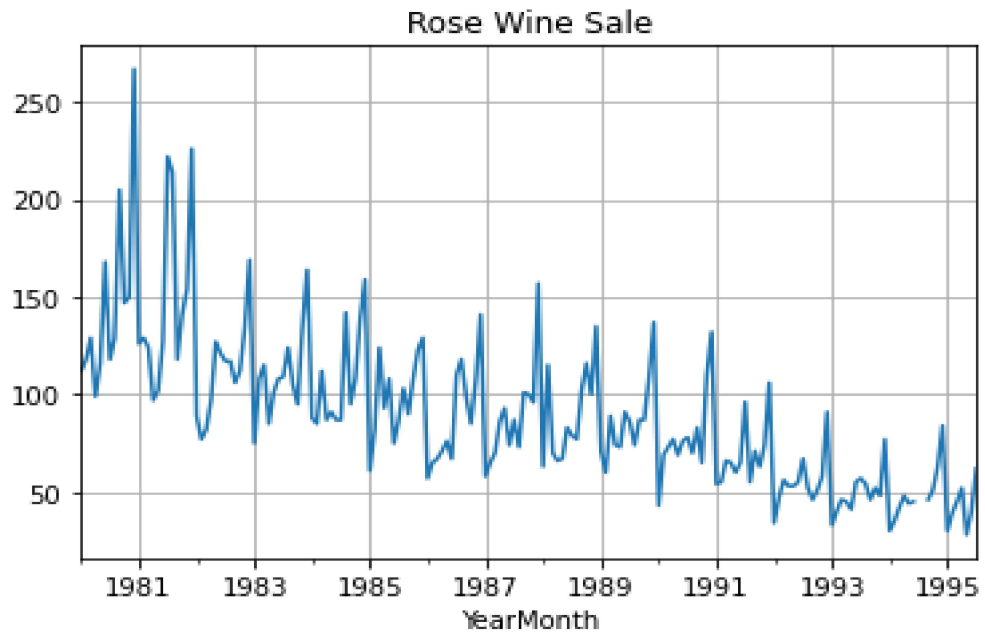


Fig-1.2

Conclusion | insight:

Observations on Rose Year wise sales data –

- We notice that there is an decreasing trend in the initial years which stabilizes after few years and again shows a decreasing trend
- We also observe seasonality in the data trend and pattern seem to repeat on yearly basis

Q2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

Solution:-

We have 187 data points in Rose wine data with two missing values and 187 data points in Sparkling wine data. The basic measures of descriptive statistics tell us how the Sales have varied across years. However, for this measure of descriptive statistics we have averaged over the whole data without taking the time component into account hence should look at the box plots year wise and month wise.

Sparkling wine data – Descriptive statistics:-

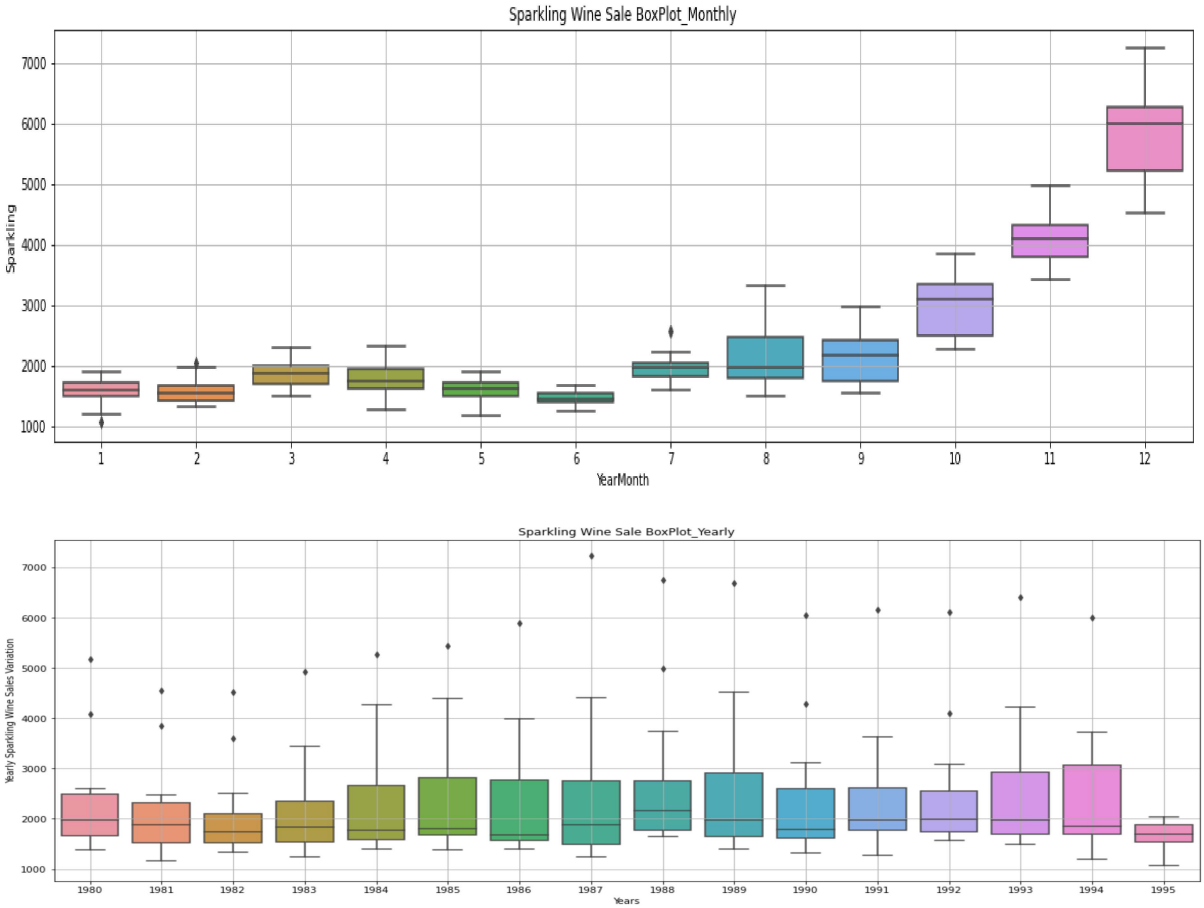
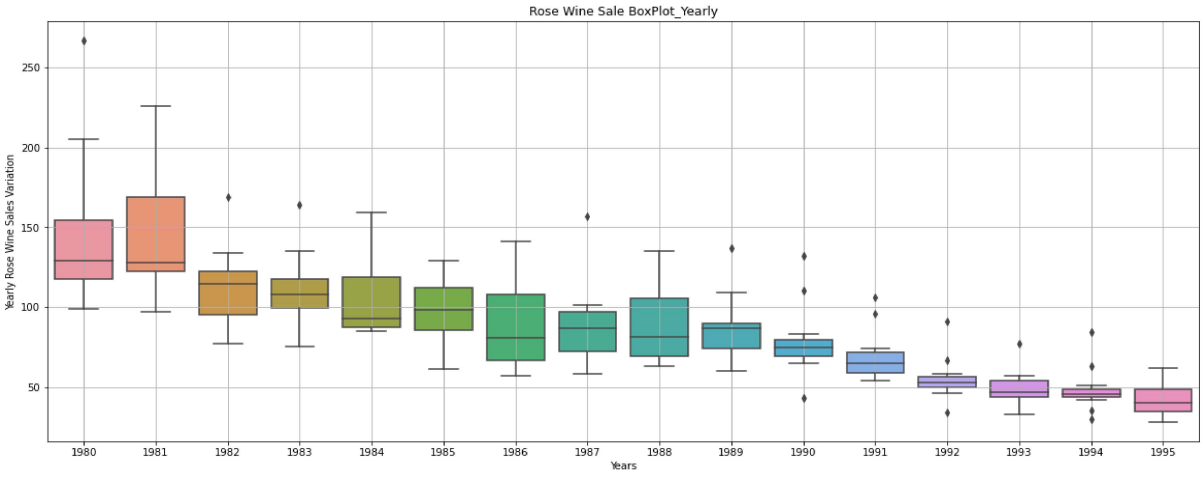


Fig-1.3

- As observed in the Time Series plot, the boxplots over here also do not indicate trend
- Also, we see that the sales of Sparkling wine has some outliers for almost all years except 1995
- We also observe December month has the highest sales value for Sparkling wine.

Rose wine data – Descriptive statistics:-



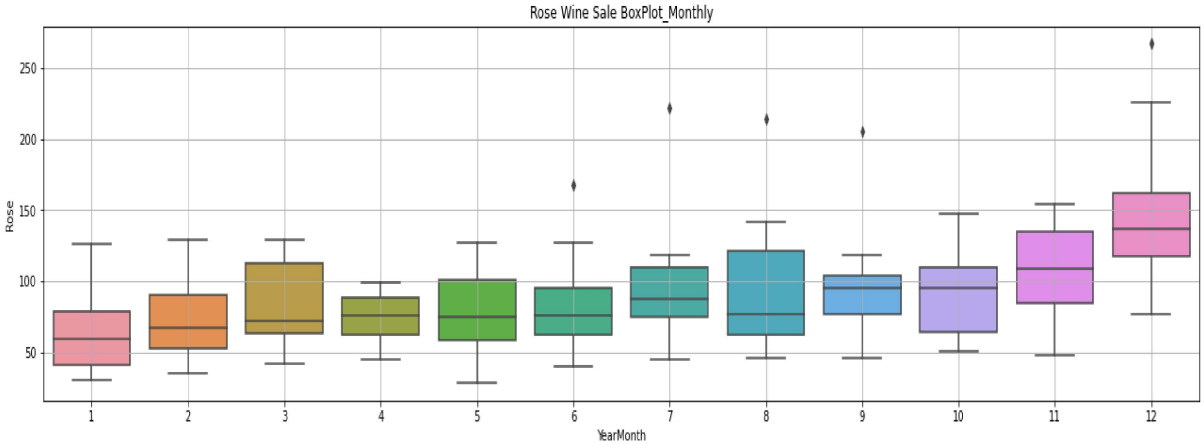


Fig-1.4

- As observed in the Time Series plot, the year wise boxplots over here also indicate a measure of downward trend.
- Also, we see that the sales of Rose wine has some outliers for certain years.
- December seems to have the highest sales of Rose wine and there are also outlier in June, July, August and September months.

Plotting a month plot of the given Time Series:-

Sparkling Sales – Month, Cumulative % and Month on Month % Sales plots

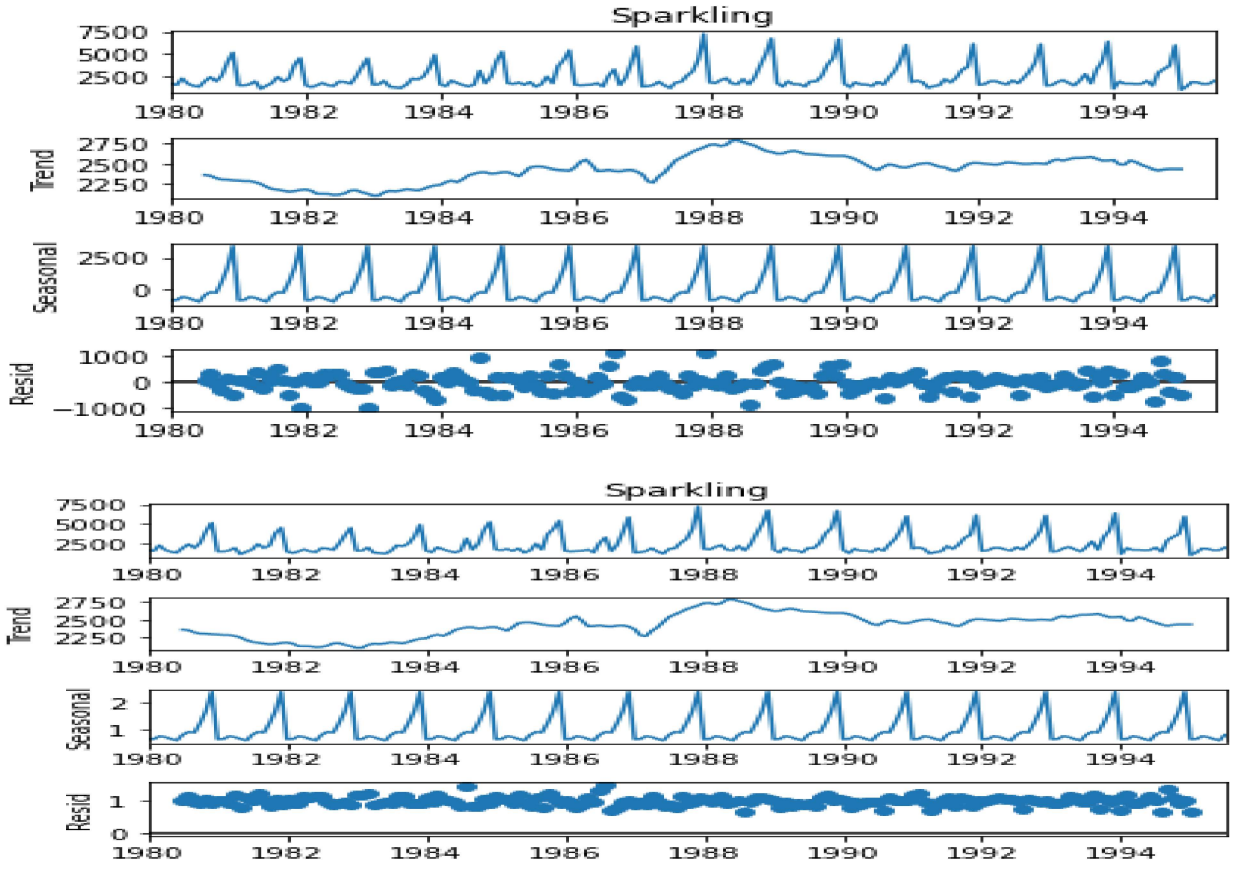


Fig-1.5

Rose Sales – Month, Cumulative % and Month on Month % Sales plots:-

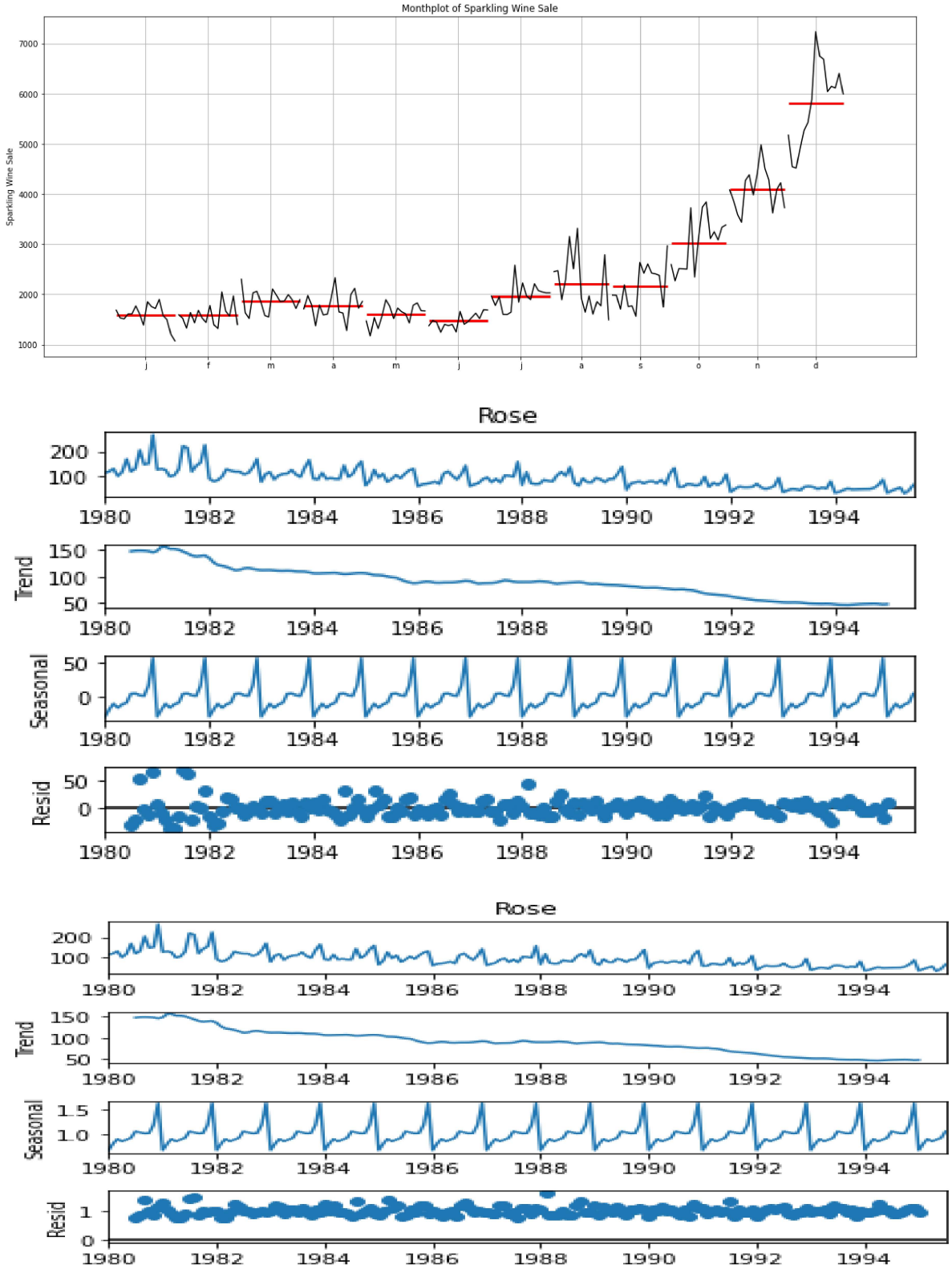


Fig-1.6

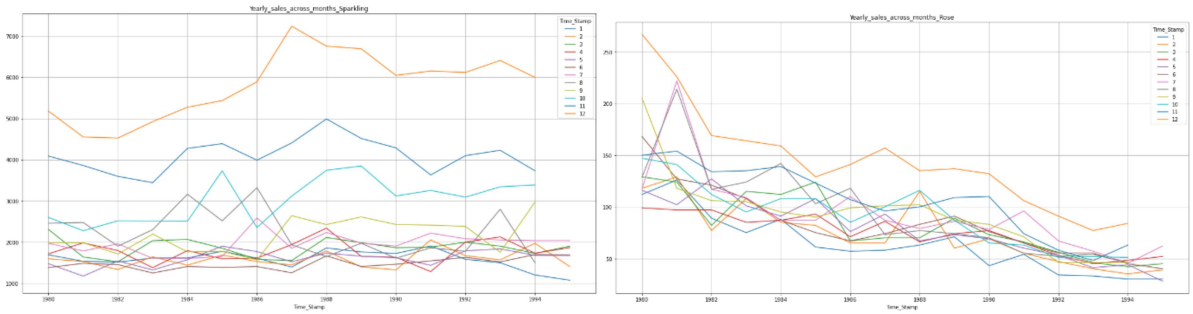


Fig-1.7

- We observe from the above Line plots of Year/month wise sales data of Rose and Sparkling wine that December month has the highest Sales and January, February and March months show lower sales values.

Q3. Split the data into training and test. The test data should start in 1991.

Solution:-

First few rows of Train Data:

First few rows of Training Data
Sparkling

Time_Stamp	
1980-01-31	1686
1980-02-29	1591
1980-03-31	2304
1980-04-30	1712
1980-05-31	1471

Last few rows of Train data:

Last few rows of Training Data
Sparkling

Time_Stamp	
1990-08-31	1605
1990-09-30	2424
1990-10-31	3116
1990-11-30	4286
1990-12-31	6047

First few rows of Test data:

First few rows of Test Data
Sparkling

Time_Stamp	
1991-01-31	1902
1991-02-28	2049
1991-03-31	1874
1991-04-30	1279
1991-05-31	1432

Last few rows of Test data:

Last few rows of Test Data
Sparkling

Time_Stamp	
1995-03-31	1897
1995-04-30	1862
1995-05-31	1670
1995-06-30	1688
1995-07-31	2031

Table-1.8

Sparkling Sales Train and Test data:-

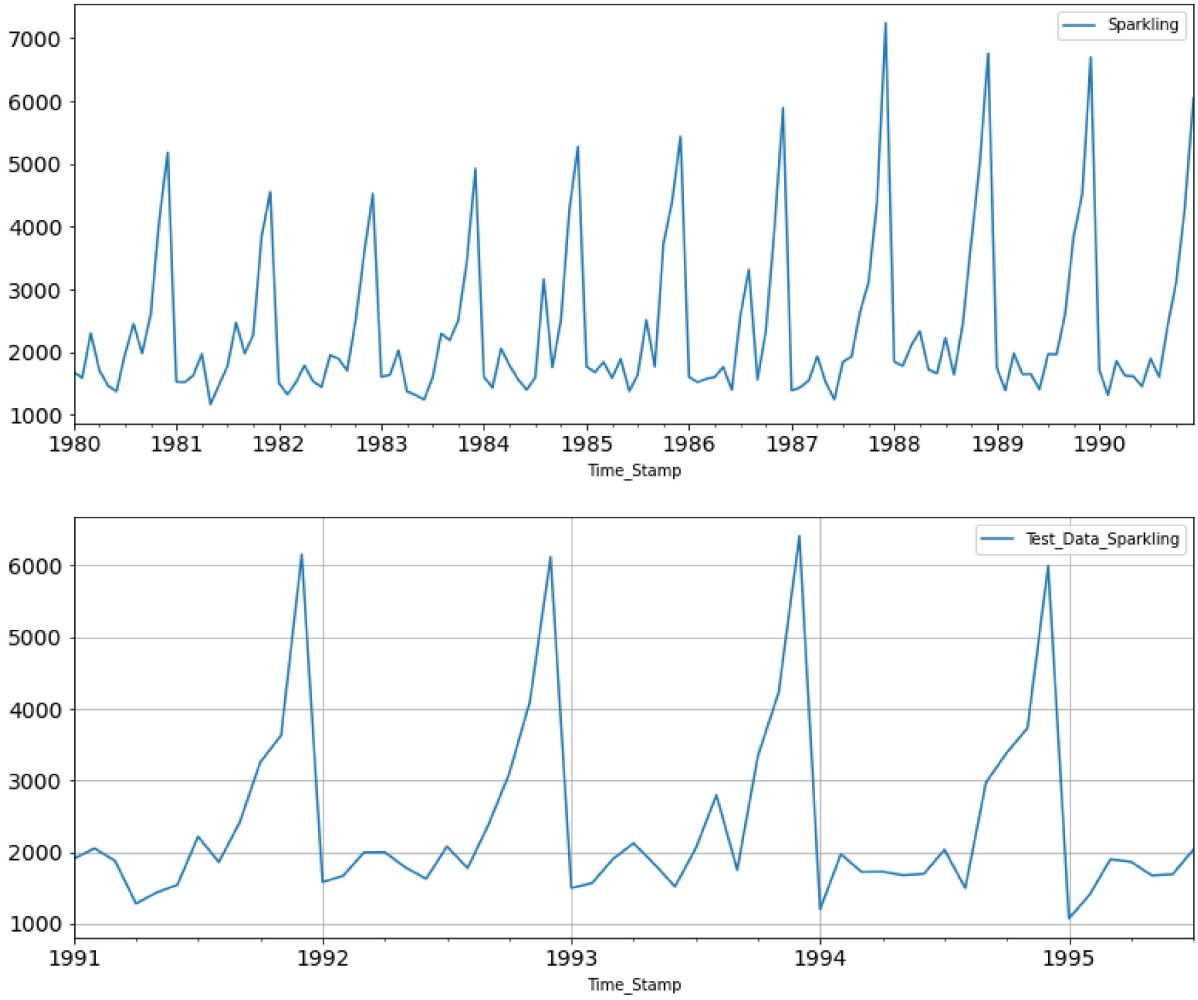


Fig-1.9

- The Train data of Sparkling wine sales has been split for data up to 1990 and has 132 data points
- The Test data of Sparkling wine sales has been split for data from 1991 and has 55 data points
- From our train-test split we are predicting the future sales as compared to the past years.

Rose Sales Train and Test data:-

First few rows of Train data:

First few rows of Training Data

Time_Stamp	Rose
1980-01-31	112.0
1980-02-29	118.0
1980-03-31	129.0
1980-04-30	99.0
1980-05-31	116.0

Last few rows of Train data:

Last few rows of Training Data

Time_Stamp	Rose
1990-08-31	70.0
1990-09-30	83.0
1990-10-31	65.0
1990-11-30	110.0
1990-12-31	132.0

First few rows of Test data:

First few rows of Test Data

Time_Stamp	Rose
1991-01-31	54.0
1991-02-28	55.0
1991-03-31	66.0
1991-04-30	65.0
1991-05-31	60.0

Last few rows of Test data:

Last few rows of Test Data

Time_Stamp	Rose
1995-03-31	45.0
1995-04-30	52.0
1995-05-31	28.0
1995-06-30	40.0
1995-07-31	62.0

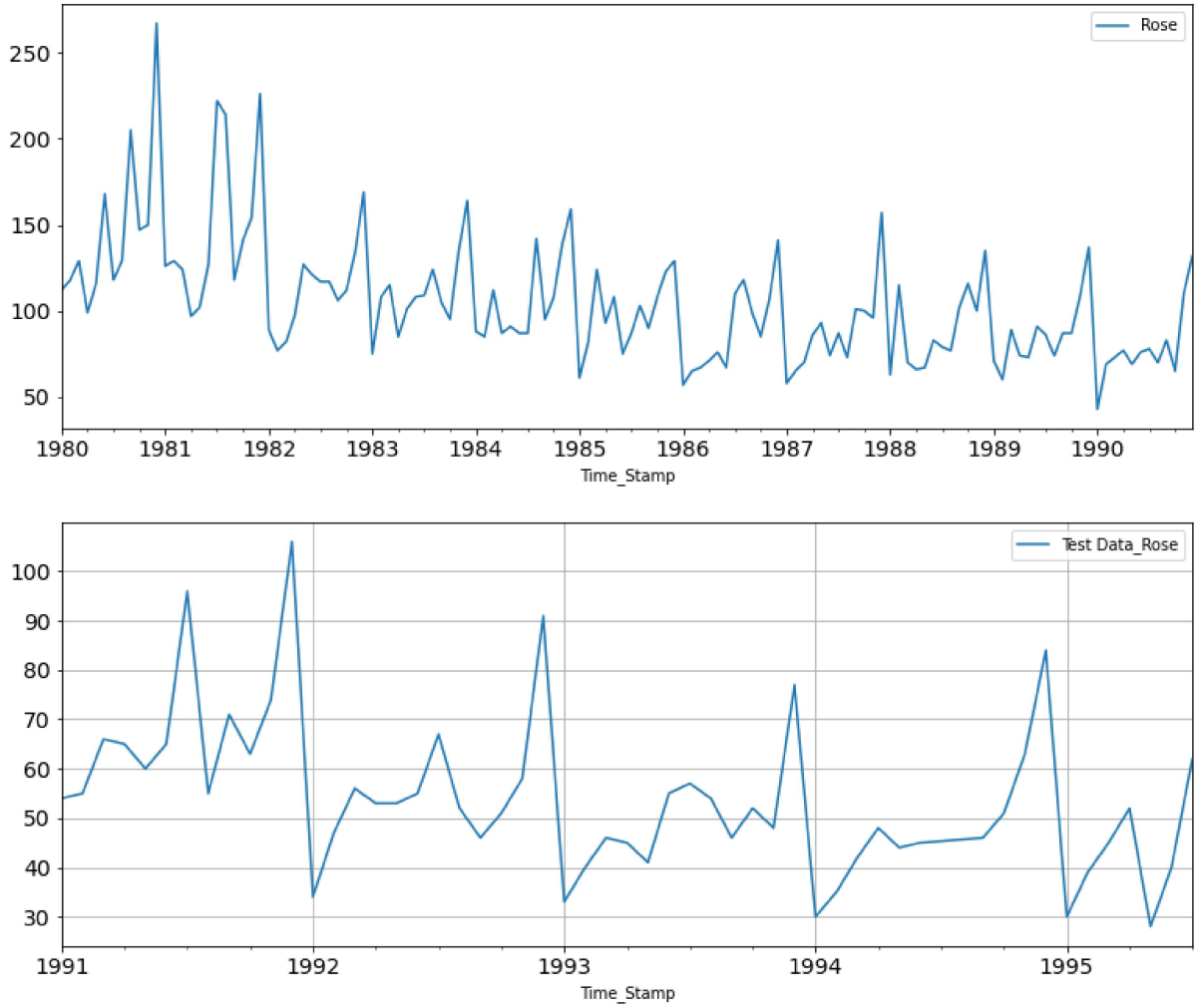


Fig-1.10

- The Train data of Rose wine sales has been split for data up to 1990 and has 132 data points
- The Test data of Rose wine sales has been split for data from 1991 and has 55 data points
- From our train-test split we are predicting the future sales as compared to the past years.

Q4. 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.

Solution:-

Model 1- Linear Regression-Sparkling:-

- In Linear regression, we regress the 'Sparkling' sales against the order of the occurrence. Hence we modified our training data before fitting it into a linear regression by generating numerical time instance order for train and test data. We then ran the Linear regression model and got Test RMSE score 1275.87

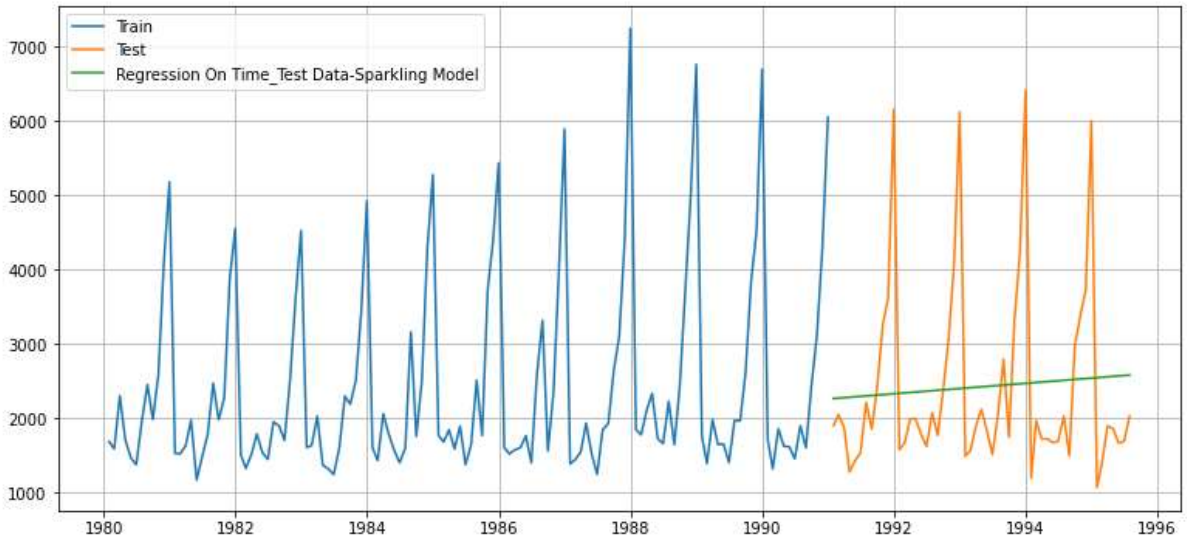


Fig-1.11

Model Evaluation:-

	Test RMSE
RegressionOnTime	1275.867052

Model-1 – Linear Regression –Rose:-

In Linear regression, we regress the 'Rose' sales against the order of the occurrence. Hence we modified our training data before fitting it into a linear regression by generating numerical time instance order for train and test data. We then ran the Linear regression model and got Test RMSE score 2372.45.

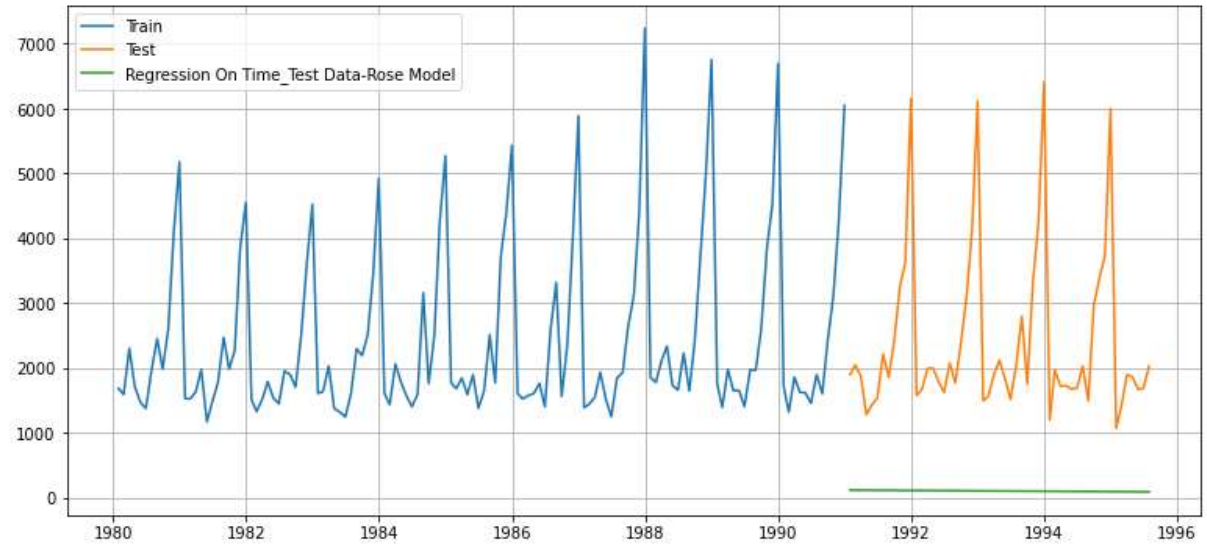


Fig-1.13

Model Evaluation –Rose

	Test RMSE
RegressionOnTime	2372.449884

Model -2 Naïve Approach – Sparkling

- We ran the Naïve Model for Sparkling Sales and got the test RMSE score – 3864.28. We observe that the green line in the Chart which shows the Naïve forecast plotting is a straight line given the Naïve models approach where Sales for tomorrow is the same as today and it applies to all the future periods.

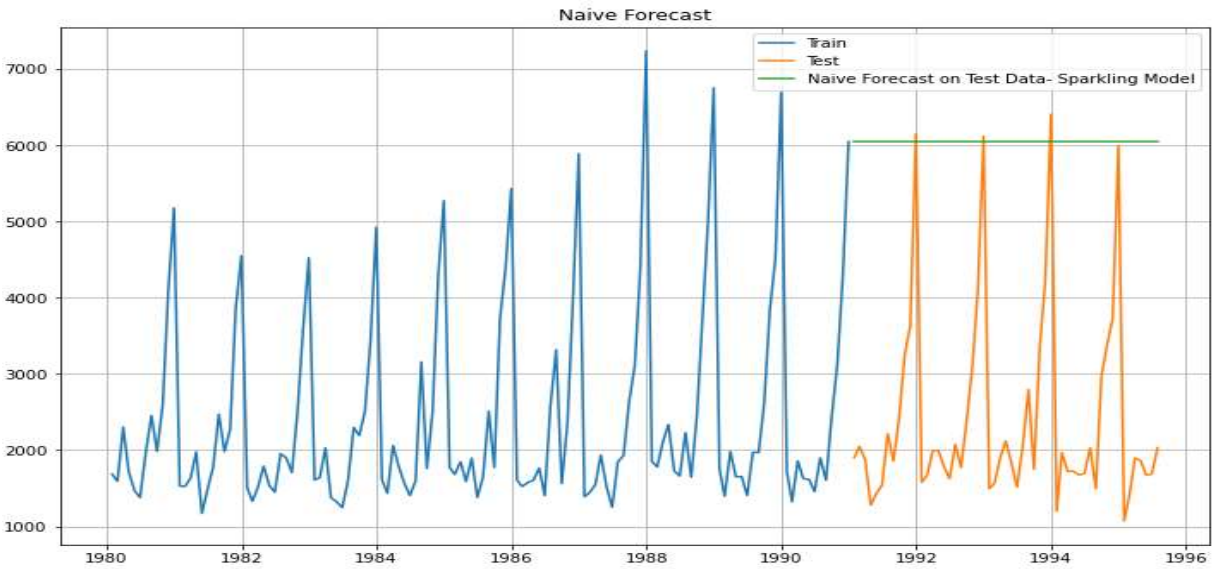


Fig-1.14

Model Evaluation – Sparkling

	Test RMSE
RegressionOnTime	1275.867052
NaiveModel	3864.279352

Model -2 Naïve Approach – Rose Sales:-

- We ran the Naïve Model for Rose Sales and got the test RMSE score – 79.72. We observe that the green line in the Chart which shows the Naïve forecast plotting is a straight line given the Naïve models approach where Sales for tomorrow is the same as today and it applies to all the future periods.

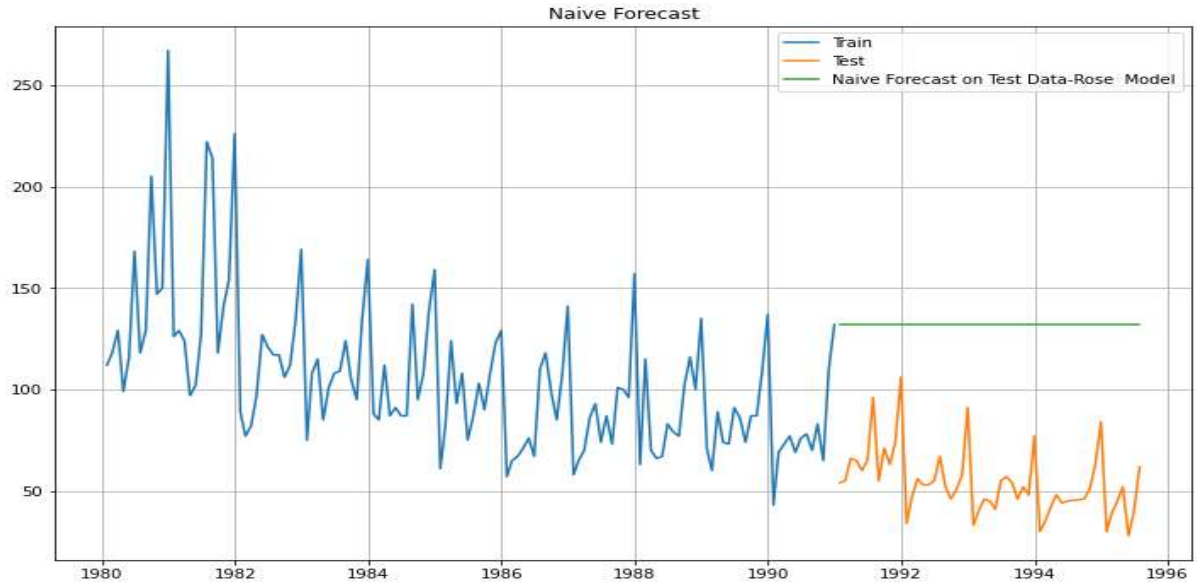


Fig-1.15

Model Evaluation:-

	Test RMSE
RegressionOnTime	2372.449884
NaiveModel	79.718773

Model -3 Simple Average Model – Sparkling Sales:-

- We ran the Simple Average model for Sparkling sales and got the Test RMSE score – 1275.08. We observe that the green line in the Chart which shows the Simple Average forecast plotting is

a straight line given the Simple Average models approach where we use the average Sales value to forecast future Sales.

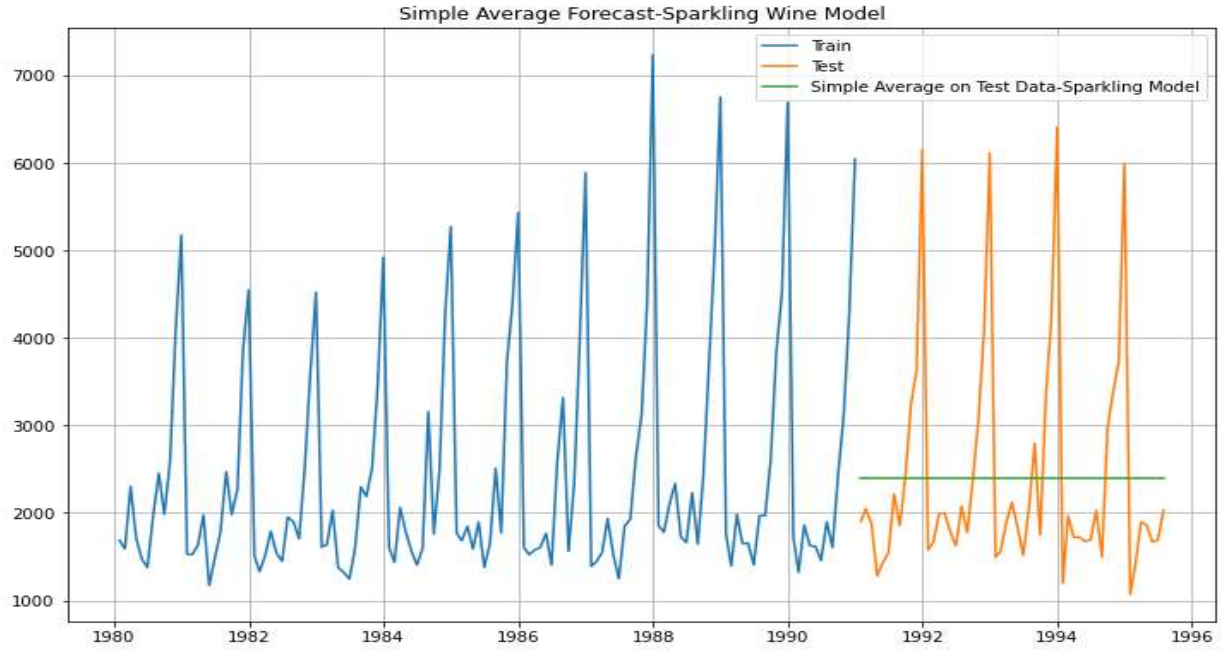


Fig-1.17

Model Evaluation:-

	Test RMSE
RegressionOnTime	1275.867052
NaiveModel	3864.279352
SimpleAverageModel	1275.081804

Model -3 Simple Average Model – Rose Sales

We ran the Simple Average model for Rose sales and got the Test RMSE score – 53.46. We observe that the green line in the Chart which shows the Simple Average forecast plotting is a straight line given the Simple Average models approach where we use the average Sales value to forecast future Sales.

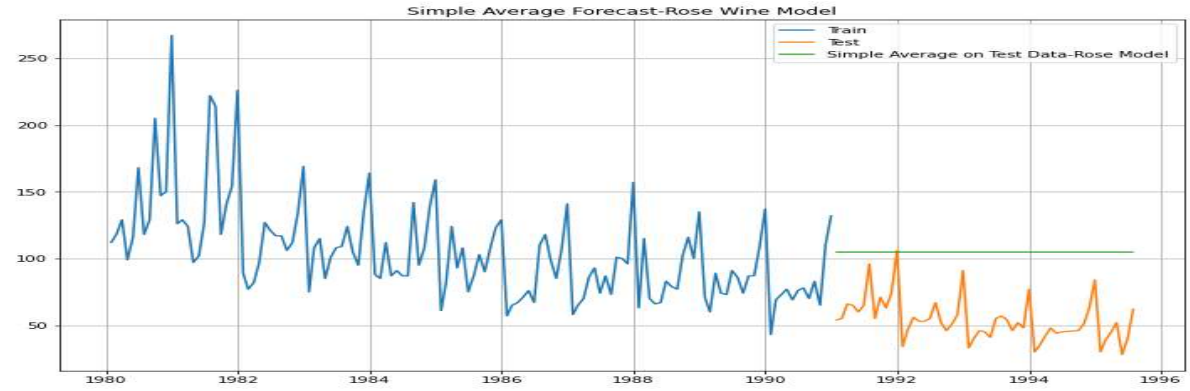


Fig-1.18

Model Evaluation:-

	Test RMSE
RegressionOnTime	2372.449884
NaiveModel	79.718773
SimpleAverageModel	53.460570

Model -4 Simple Exponential Smoothing –Sparkling Sales:-

While trying to forecast Model Evaluation for alpha = 0.995: Simple Exponential Smoothing, we get an RMSE Score of 1316.13 for the test data.

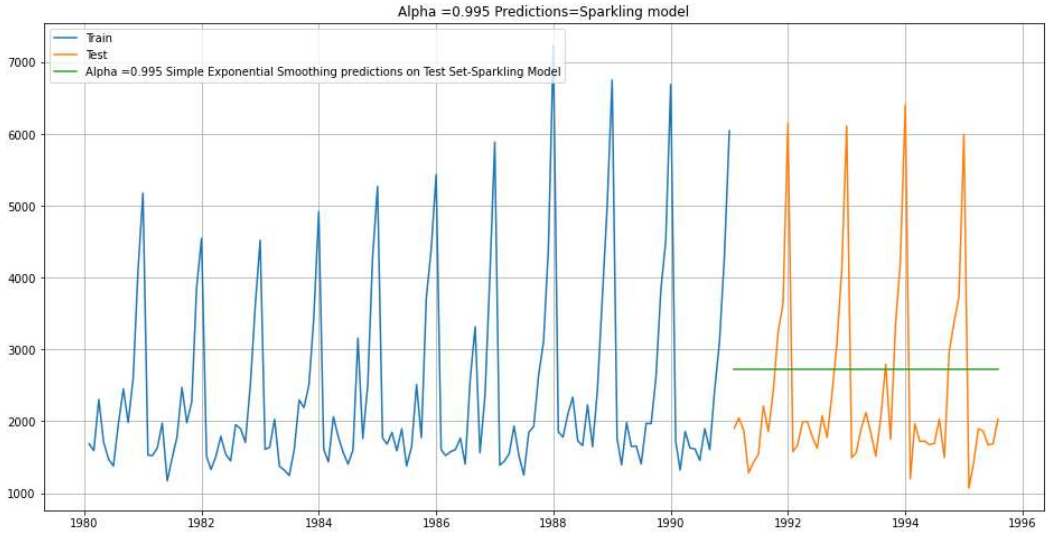


Fig-1.20

Model Evaluation

	Test RMSE
RegressionOnTime	1275.867052
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
Alpha=0.995,SimpleExponentialSmoothing	1316.135411

Model -4 Simple Exponential Smoothing – Rose Sales:-

While trying to forecast Model Evaluation for alpha = 0.995: Simple Exponential Smoothing, we get an RMSE Score of 36.80 for the test data.

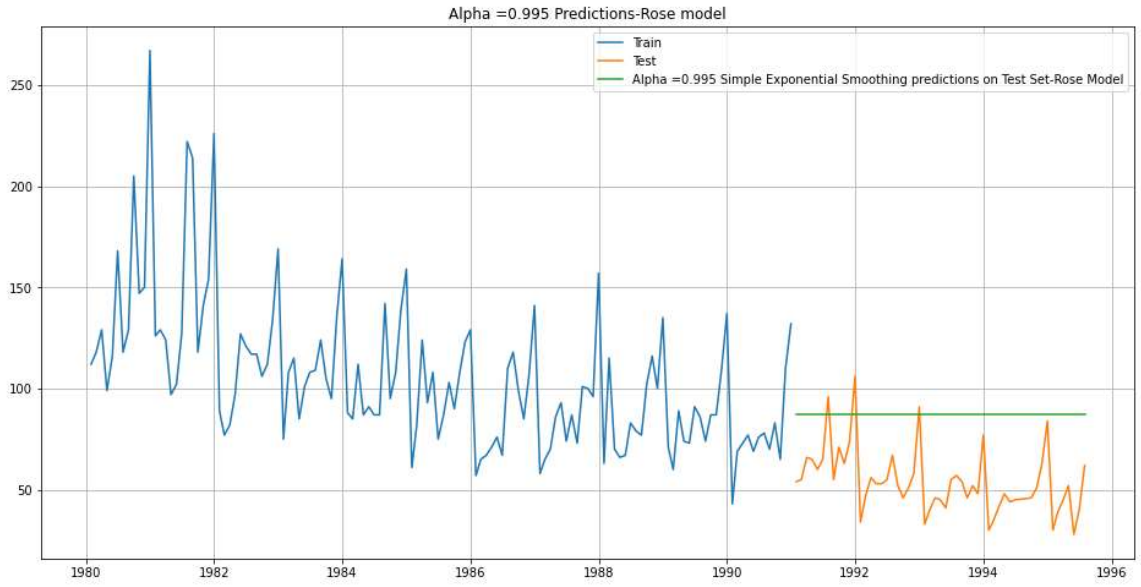


Fig-1.21

Model Evaluation:-

	Test RMSE
RegressionOnTime	2372.449884
NaiveModel	79.718773
SimpleAverageModel	53.460570
Alpha=0.995,SimpleExponentialSmoothing	36.796242

Model - 5 Double Exponential Smoothing (Level and Trend) –Sparkling Sales:-

We ran the Double Exponential Smoothing for Sparkling Sales data, at alpha =0.3 and Beta = 0.3 and the RMSE obtained for test data is 2007.23. The result is shown below.

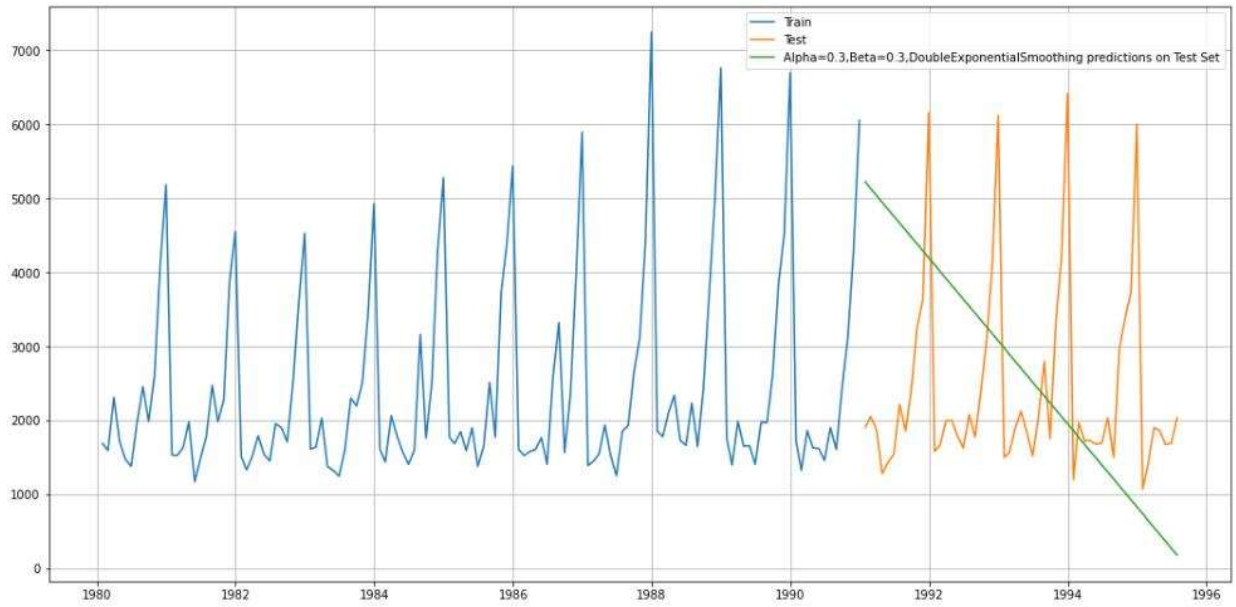


Fig-1.22

Model Evaluation:-

	Test RMSE
RegressionOnTime	1275.867052
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
Alpha=0.995,SimpleExponentialSmoothing	1316.135411
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	2007.238526

Model – 5 Double Exponential Smoothing (Level and Trend) –Rose Sales:-

We ran the Double Exponential Smoothing for Rose Sales data at alpha =0.3 and Beta = 0.3 and RMSE score for test data obtained is 15.569.

The result is shown below.

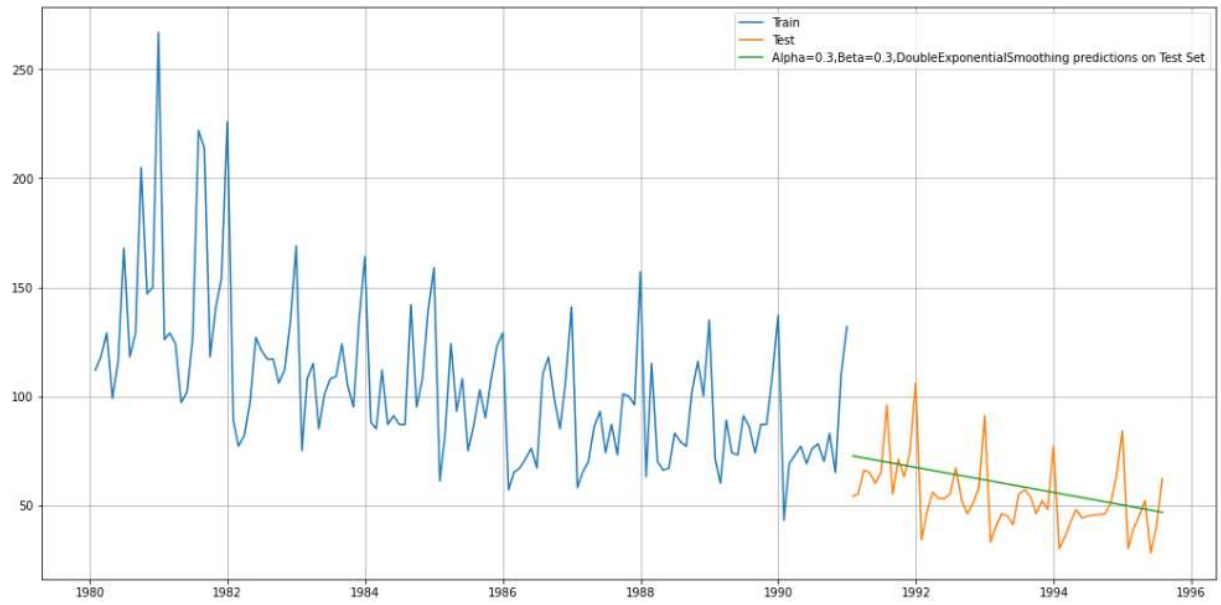


Fig-1.24

Model Evaluation:-

	Test RMSE
RegressionOnTime	2372.449884
NaiveModel	79.718773
SimpleAverageModel	53.460570
Alpha=0.995,SimpleExponentialSmoothing	36.796242
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	15.569001

Model – 6 Triple Exponential Smoothing (Level, Trend and Seasonality) –Sparkling Sales:-

We ran the Triple Exponential Smoothing for Sparkling sales with auto fit parameters which resulted in Smoothing level(Alpha) 0.676, Smoothing slope/trend(Beta) 0.088 and Smoothing seasonality(Gamma) 0.323 and we got Test RMSE score 469.659.

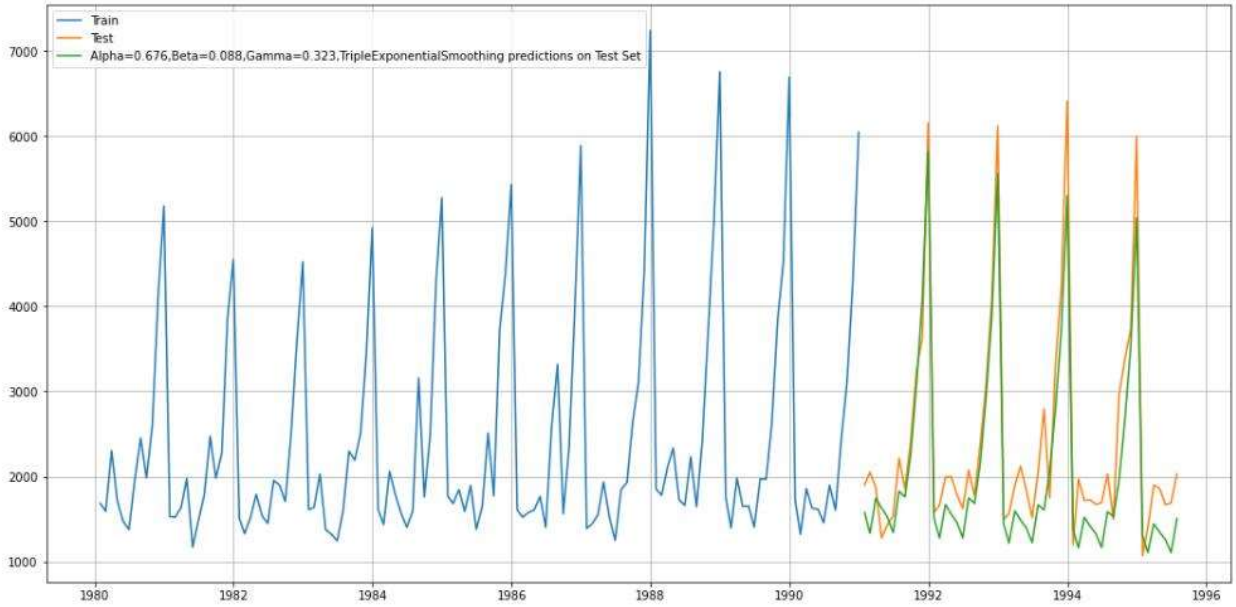


Fig-1.27

Model Evaluation:-

	Test RMSE
RegressionOnTime	1275.867052
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
Alpha=0.995,SimpleExponentialSmoothing	1316.135411
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	2007.238526
Alpha=0.676,Beta=0.088,Gamma=0.323,TripleExponentialSmoothing	469.659106

Model – 6 Triple Exponential Smoothing (Level, Trend and Seasonality) Rose Sales :-

We ran the Triple Exponential Smoothing for Sparkling sales with auto fit parameters which resulted in Smoothing level(Alpha) 0.676, Smoothing slope/trend(Beta) 0.088 and Smoothing seasonality(Gamma) 0.323 and we got Test RMSE score 21.154.

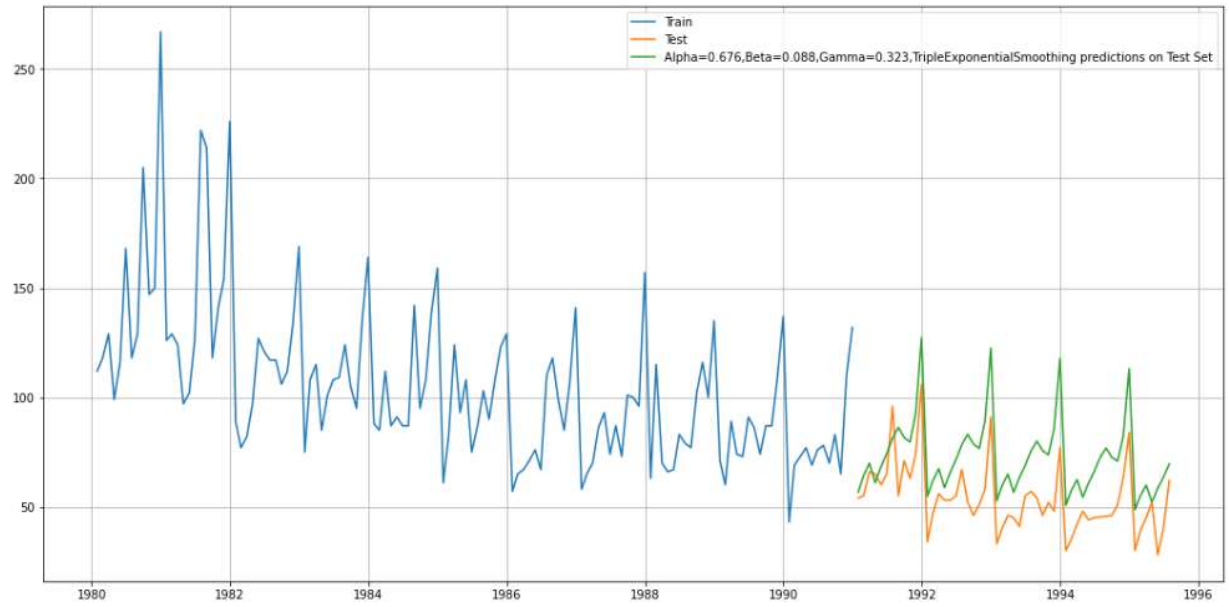


Fig-1.28

Model Evaluation:-

	Test RMSE
RegressionOnTime	2372.449884
NaiveModel	79.718773
SimpleAverageModel	53.460570
Alpha=0.995,SimpleExponentialSmoothing	36.796242
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	15.569001
Alpha=0.676,Beta=0.088,Gamma=0.323,TripleExponentialSmoothing	21.154772

Conclusion | insight: All Modal are mention in file.

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.

Solution:-

Stationarity Test:- Sparkling Sales -On training Data

- We check the stationarity of the Sparkling sales data at alpha 0.05 and observe from the following result table that P value (0.66) is greater than alpha value. Hence we fail to reject the null hypothesis that the data is not stationary.

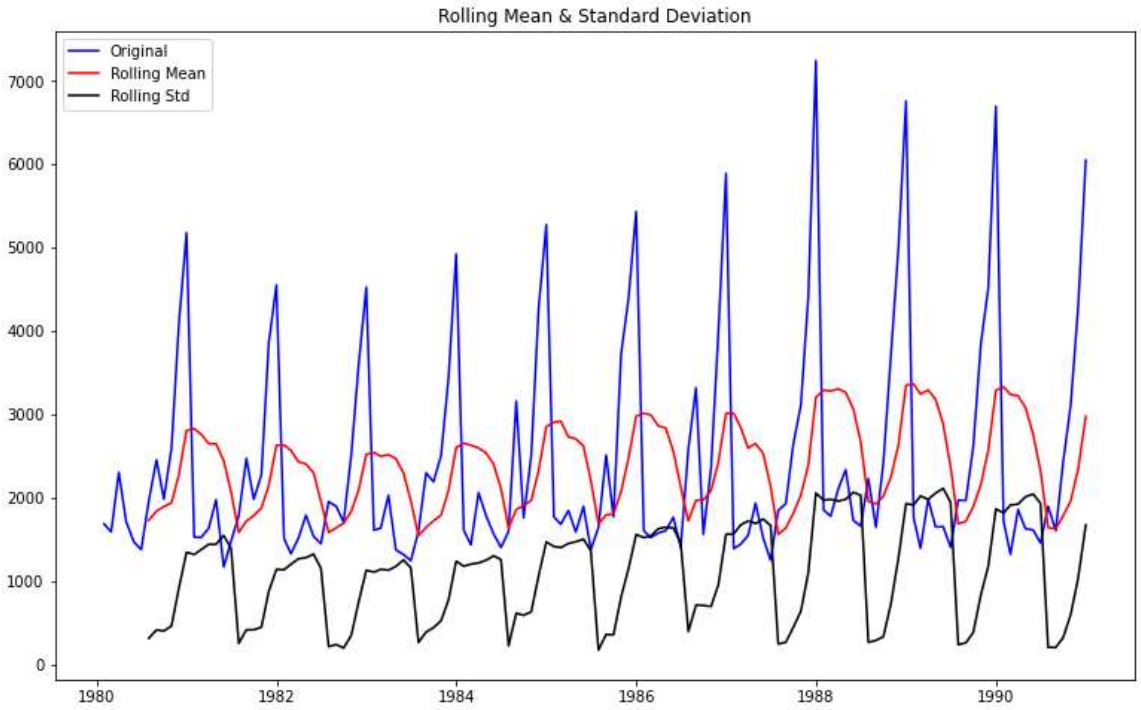


Fig-1.29

Results of Dickey fuller test:-

```
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
```

- We see that at 5% significant level the Time Series is non-stationary. Let us take a difference of order 1 and check whether the Time Series is stationary or not.

Stationary test –Sparkling after 1 Level order of Differencing:-

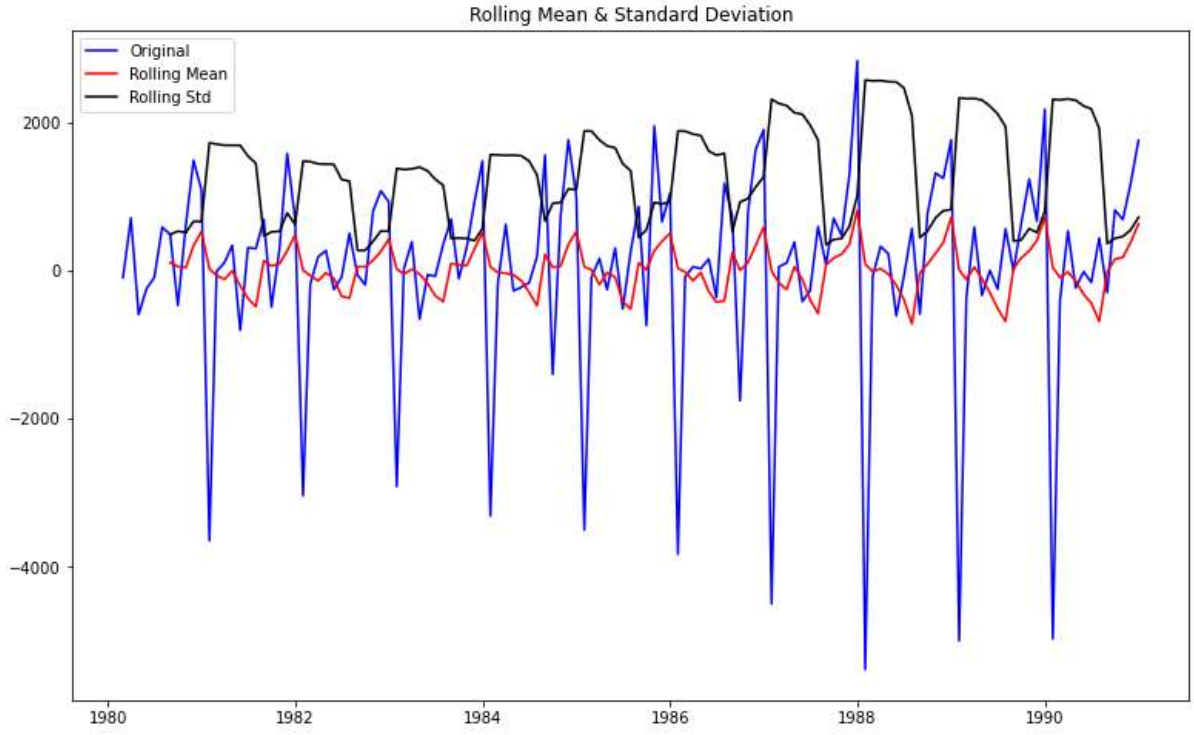


Fig-1.29

Results of Dickey fuller test:-

```
Results of Dickey-Fuller Test:
Test Statistic                -8.005007e+00
p-value                        2.280104e-12
#Lags Used                     1.100000e+01
Number of Observations Used    1.190000e+02
Critical Value (1%)           -3.486535e+00
Critical Value (5%)           -2.886151e+00
Critical Value (10%)          -2.579896e+00
dtype: float64
```

Observation:

- Now we can see that after 1 level of differencing, p-value is much less than alpha=0.05, and it rejects the null hypothesis and proves that data is stationary now.

Stationarity Test:- Rose Sales -On training Data:-

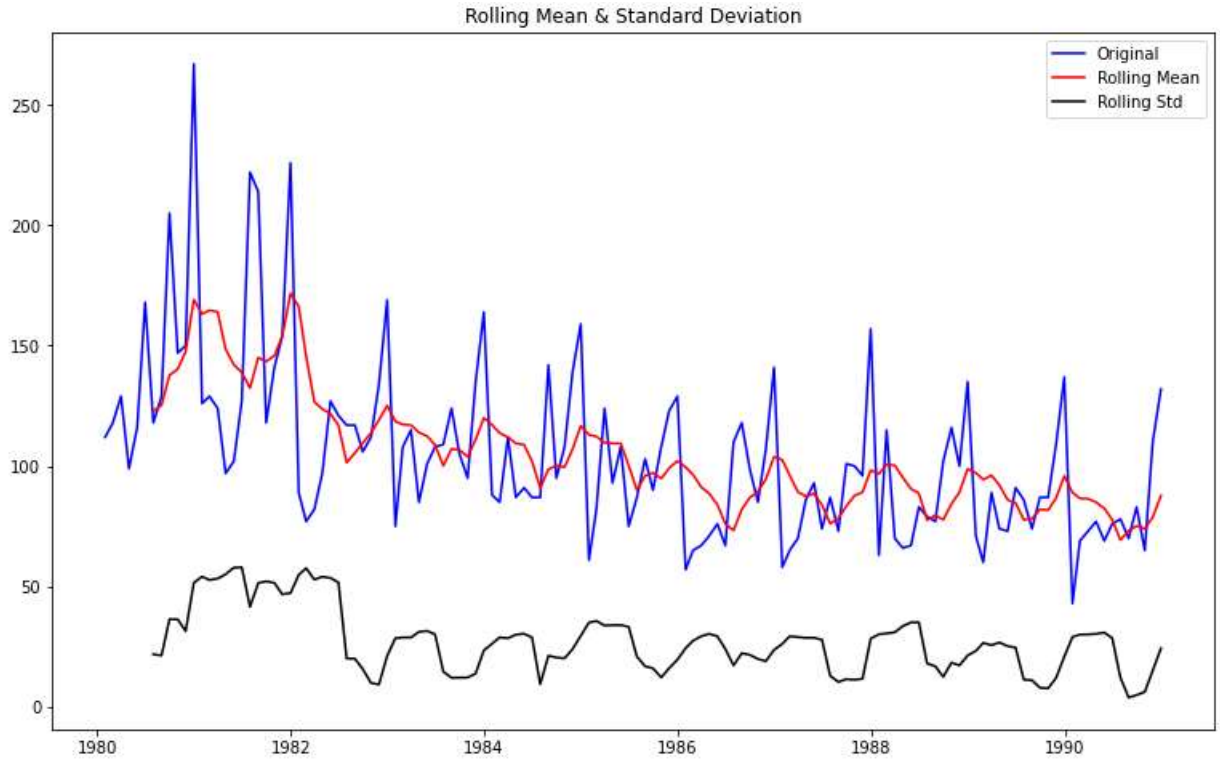


Fig-1.30

Results of Dickey-Fuller Test:-

```
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
```

Observation:

- As we see here, p-value is 0.21 which is greater than alpha = 0.05, and we are unable to reject the null hypothesis: Time Series is non-stationary.
- Let us take a difference of order 1 and check whether the Time Series is stationary or not. Stationary test –Rose Data after 1 Level order of Differencing:

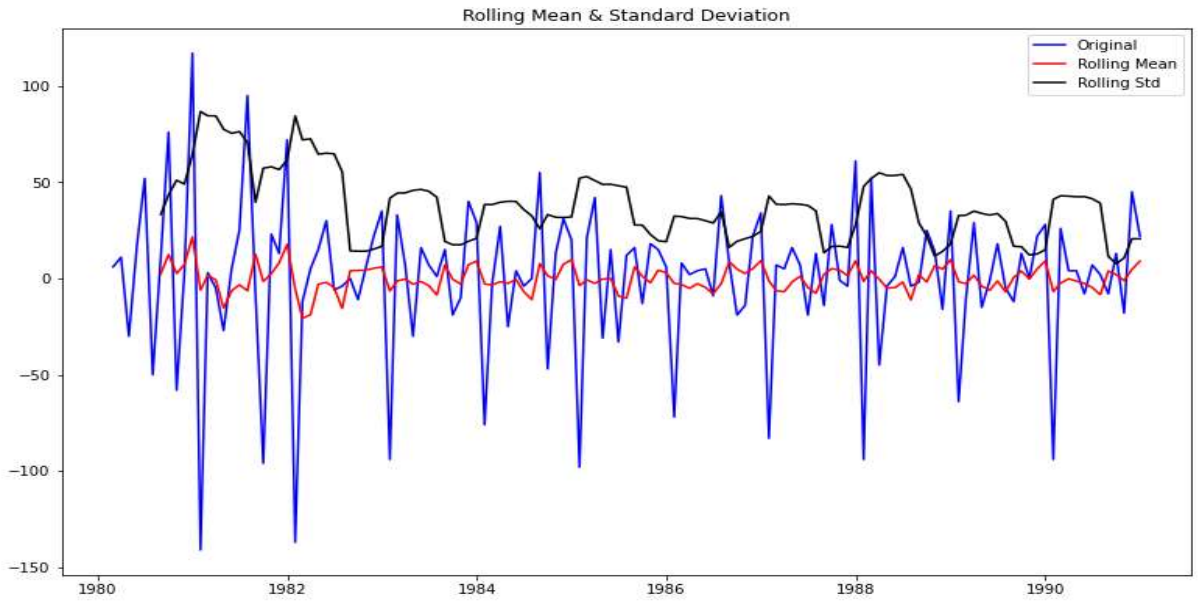


Fig-1.32

Results of Dickey-Fuller Test:-

```
Results of Dickey-Fuller Test:
Test Statistic                -6.592372e+00
p-value                        7.061944e-09
#Lags Used                     1.200000e+01
Number of Observations Used    1.180000e+02
Critical Value (1%)           -3.487022e+00
Critical Value (5%)           -2.886363e+00
Critical Value (10%)          -2.580009e+00
dtype: float64
```

Observation:

- As we see here, p-value is much less than alpha = 0.05, and we are able to reject the null hypothesis: Time Series is non-stationary. So Time Series is stationary now.

Q6. 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.

Solution:-

Automated ARIMA – Sparkling:

- We ran the automated ARIMA model for Sparkling Sales and sorted the AIC values output from lowest to highest. We then proceeded to build the ARIMA model with the lowest Akaike Information Criteria and got the test RMSE Score 1374.75.

Sorted AIC Values in ascending order:-

	param	AIC
8	(2, 1, 2)	2210.617939
7	(2, 1, 1)	2232.360490
2	(0, 1, 2)	2232.783098
5	(1, 1, 2)	2233.597647
4	(1, 1, 1)	2235.013945
6	(2, 1, 0)	2262.035600
1	(0, 1, 1)	2264.906439
3	(1, 1, 0)	2268.528061
0	(0, 1, 0)	2269.582796

ARIMA Model Results:- Sparkling

ARIMA Model Results

Dep. Variable:	D.Sparkling	No. Observations:	131
Model:	ARIMA(2, 1, 2)	Log Likelihood	-1099.309
Method:	css-mle	S.D. of innovations	1012.457
Date:	Sat, 09 Oct 2021	AIC	2210.618
Time:	18:32:36	BIC	2227.869
Sample:	02-29-1980 - 12-31-1990	HQIC	2217.628

	coef	std err	z	P>|z|	[0.025	0.975]
const	5.5857	0.517	10.814	0.000	4.573	6.598
ar.L1.D.Sparkling	1.2699	0.074	17.046	0.000	1.124	1.416
ar.L2.D.Sparkling	-0.5602	0.074	-7.618	0.000	-0.704	-0.416
ma.L1.D.Sparkling	-1.9983	0.042	-47.170	0.000	-2.081	-1.915
ma.L2.D.Sparkling	0.9983	0.042	23.560	0.000	0.915	1.081

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.1335	-0.7073j	1.3361	-0.0888
AR.2	1.1335	+0.7073j	1.3361	0.0888
MA.1	1.0002	+0.0000j	1.0002	0.0000
MA.2	1.0015	+0.0000j	1.0015	0.0000

Fig-1.33

Predicting on the Test Set using this model and evaluating the model.

	RMSE
ARIMA(2,1,2)	1374.759476

Automated ARIMA – Rose

- We ran the automated ARIMA model for Sparkling Sales and sorted the AIC values output from lowest to highest. We then proceeded to build the ARIMA model with the lowest Akaike Information Criteria and got the test RMSE Score 15.617.

Sorted AIC Values in ascending order:-

	param	AIC
2	(0, 1, 2)	1276.835372
5	(1, 1, 2)	1277.359222
4	(1, 1, 1)	1277.775752
7	(2, 1, 1)	1279.045689
8	(2, 1, 2)	1279.298694
1	(0, 1, 1)	1280.726183
6	(2, 1, 0)	1300.609261
3	(1, 1, 0)	1319.348311
0	(0, 1, 0)	1335.152658

Predicting on the Test Set using this model and evaluating the model

	RMSE
ARIMA(0,1,2)	15.617949

ARIMA Model Results: - Rose

ARIMA Model Results

Dep. Variable:	D.Rose	No. Observations:	131
Model:	ARIMA(0, 1, 2)	Log Likelihood	-634.418
Method:	css-mle	S.D. of innovations	30.167
Date:	Sat, 09 Oct 2021	AIC	1276.835
Time:	18:32:39	BIC	1288.336
Sample:	02-29-1980 - 12-31-1990	HQIC	1281.509

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.4885	0.085	-5.742	0.000	-0.655	-0.322
ma.L1.D.Rose	-0.7601	0.101	-7.499	0.000	-0.959	-0.561
ma.L2.D.Rose	-0.2398	0.095	-2.518	0.012	-0.427	-0.053

Roots

	Real	Imaginary	Modulus	Frequency
MA.1	1.0000	+0.0000j	1.0000	0.0000
MA.2	-4.1695	+0.0000j	4.1695	0.5000

Fig-1.34

Automated Version of SARIMA Model – Sparkling Sales:-

We build the automated version of SARIMA model based on the right order obtained from lowest AIC Values.

Sorted AIC Values in ascending order:-

	param	seasonal	AIC
53	(1, 1, 2)	(2, 0, 2, 6)	1727.678701
26	(0, 1, 2)	(2, 0, 2, 6)	1727.888804
80	(2, 1, 2)	(2, 0, 2, 6)	1729.363556
17	(0, 1, 1)	(2, 0, 2, 6)	1741.696451
44	(1, 1, 1)	(2, 0, 2, 6)	1743.379778

SARIMA MODEL Result:-

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1, 1, 2)x(2, 0, 2, 6)	Log Likelihood	-855.839
Date:	Sat, 09 Oct 2021	AIC	1727.679
Time:	18:33:42	BIC	1749.707
Sample:	0	HQIC	1736.621
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.6448	0.286	-2.256	0.024	-1.205	-0.085
ma.L1	-0.1069	0.250	-0.428	0.669	-0.596	0.383
ma.L2	-0.7006	0.202	-3.470	0.001	-1.096	-0.305
ar.S.L6	-0.0045	0.027	-0.165	0.869	-0.058	0.049
ar.S.L12	1.0361	0.018	56.061	0.000	1.000	1.072
ma.S.L6	0.0676	0.152	0.444	0.657	-0.231	0.366
ma.S.L12	-0.6122	0.093	-6.588	0.000	-0.794	-0.430
sigma2	1.448e+05	1.71e+04	8.464	0.000	1.11e+05	1.78e+05

Ljung-Box (L1) (Q):	0.09	Jarque-Bera (JB):	25.23
Prob(Q):	0.77	Prob(JB):	0.00
Heteroskedasticity (H):	2.63	Skew:	0.47
Prob(H) (two-sided):	0.00	Kurtosis:	5.09

Warnings:

Fig-1.35

Model Diagnostic Plot: Sparkling:-

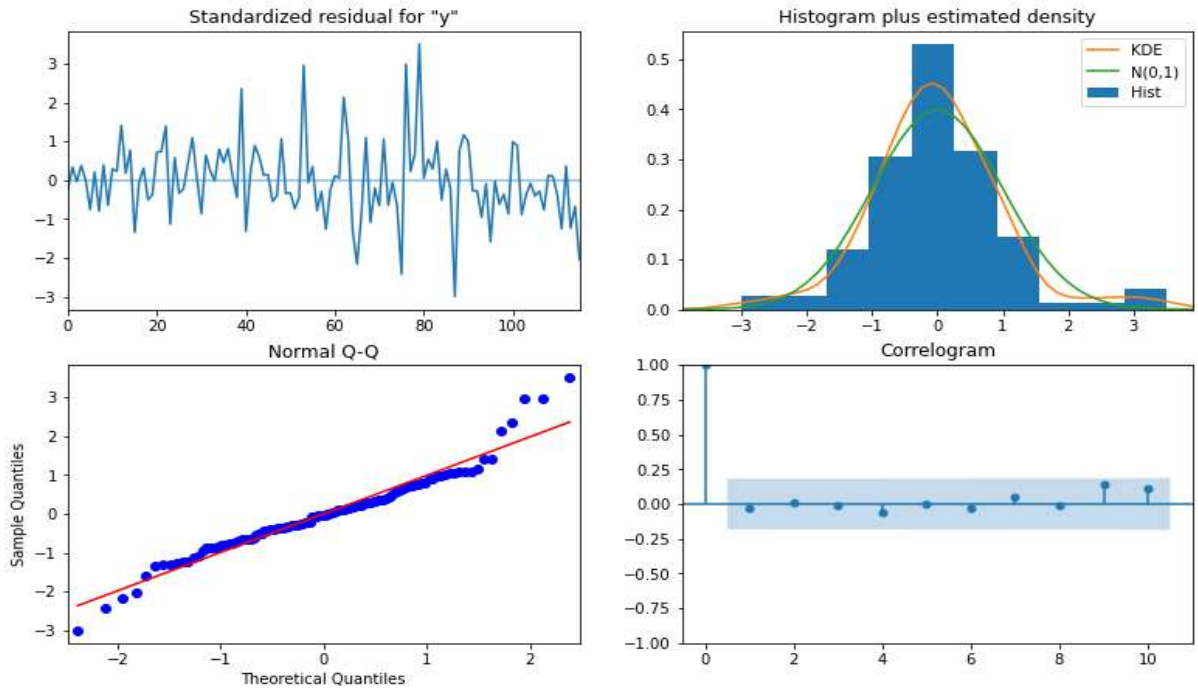


Fig-1.37

Observation:-

- Inference from model diagnostic confirms that the model residuals are normally distributed.
- Standardized Residual – Do not display any obvious seasonality.
- Histogram plus estimated density – The KDE plot of the residuals is similar with the normal distribution.
- Normal Q-Q Plot: There is an ordered distribution of residuals (blue dots) following the linear trend of the samples taken from a standard normal distribution with N (0, 1).
- Correlogram – The time series residuals have low correlation with lagged versions of itself.

Prediction on Test set:-

	RMSE
ARIMA(2,1,2)	1374.759476
SARIMA(1,1,2)(2,0,2,6)	626.894467

Automated Version of SARIMA Model – Rose Sales:-

- We build the automated version of SARIMA model for Rose Sales data based on the right order obtained from lowest AIC Values. Sorted AIC Values in ascending order.

	param	seasonal	AIC
53	(1, 1, 2)	(2, 0, 2, 6)	1041.655818
26	(0, 1, 2)	(2, 0, 2, 6)	1043.600261
80	(2, 1, 2)	(2, 0, 2, 6)	1045.231328
71	(2, 1, 1)	(2, 0, 2, 6)	1051.673461
44	(1, 1, 1)	(2, 0, 2, 6)	1052.778470

SARIMA Model Result: Rose Data:-

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1, 1, 2)x(2, 0, 2, 6)	Log Likelihood	-512.828
Date:	Sat, 09 Oct 2021	AIC	1041.656
Time:	18:34:22	BIC	1063.685
Sample:	0	HQIC	1050.598
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5940	0.152	-3.900	0.000	-0.892	-0.295
ma.L1	-0.1954	277.656	-0.001	0.999	-544.392	544.001
ma.L2	-0.8046	223.454	-0.004	0.997	-438.766	437.157
ar.S.L6	-0.0625	0.035	-1.764	0.078	-0.132	0.007
ar.S.L12	0.8451	0.039	21.887	0.000	0.769	0.921
ma.S.L6	0.2225	431.320	0.001	1.000	-845.148	845.593
ma.S.L12	-0.7774	335.378	-0.002	0.998	-658.106	656.551
sigma2	335.1816	1.79e+05	0.002	0.999	-3.5e+05	3.51e+05

Ljung-Box (L1) (Q):	0.07	Jarque-Bera (JB):	56.67
Prob(Q):	0.78	Prob(JB):	0.00
Heteroskedasticity (H):	0.47	Skew:	0.52
Prob(H) (two-sided):	0.02	Kurtosis:	6.26

Fig-1.40

Model Diagnostic Plot: Rose Sales Data:-

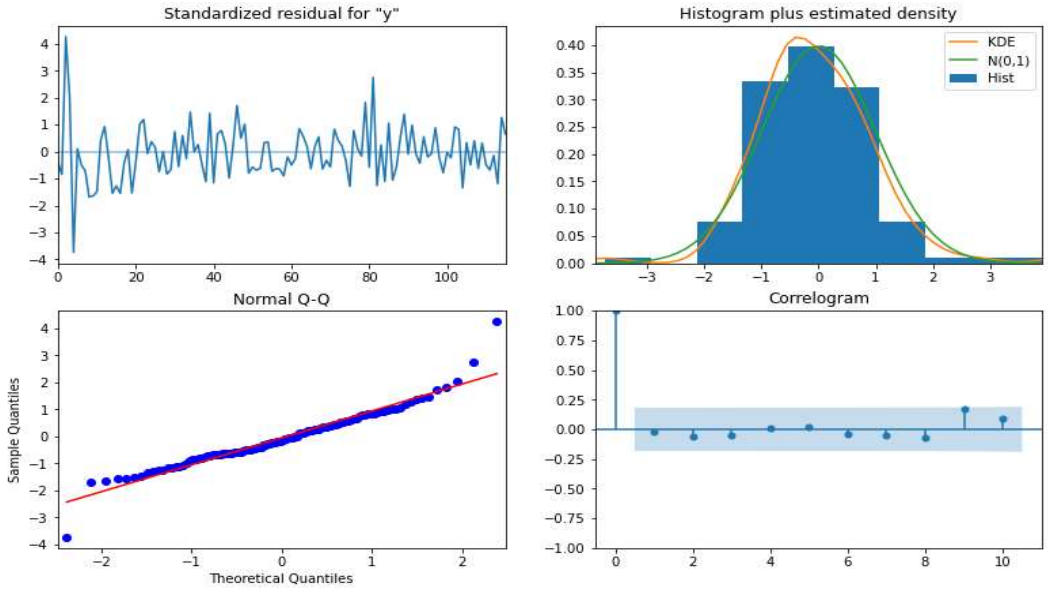


Fig-1.41

Observation:

- Inference from model diagnostic confirms that the model residuals are normally distributed.
- Standardized Residual – Do not display any obvious seasonality.
- Histogram plus estimated density – The KDE plot of the residuals is similar with the normal distribution.

- Normal Q-Q Plot: There is an ordered distribution of residuals (blue dots) following the linear trend of the samples taken from a standard normal distribution with N (0, 1).

- Correlogram – The time series residuals have low correlation with lagged versions of itself.

Prediction on Test set:

	RMSE
ARIMA(0,1,2)	15.617949
SARIMA(1,1,2)(2,0,2,6)	26.133705

- **Conclusion | insight**: All The time series residuals have low correlation with lagged versions of itself.

Q7. 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.

Solution:-

Manual ARIMA Model – Sparkling Data:-

- We now built manual ARIMA model for Sparkling Sales based on the ACF and PACF plots. Hence we chose the AR parameter p value 1, Moving average parameter q value 2 and d =1 based on below plots.

ACF Plot:-

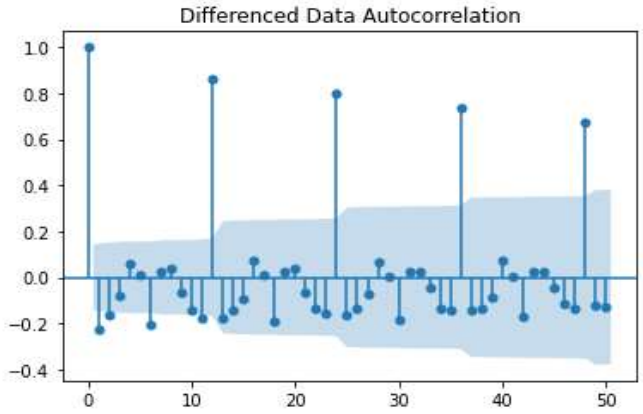


Fig-1.41

PACF Plot:-

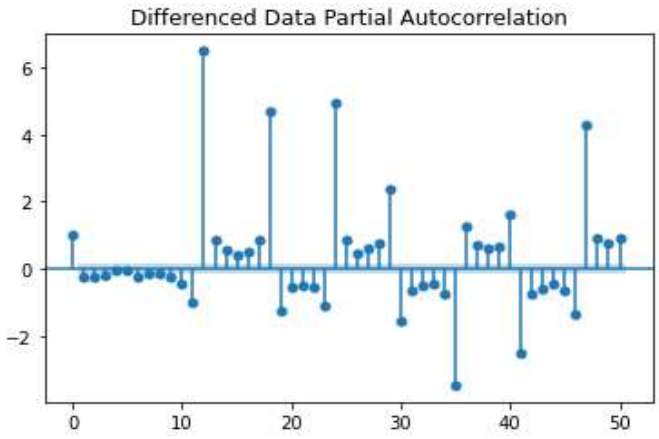


Fig-1.42

Observation: -

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 1.
- The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 2.

ARIMA Model Results:-

ARIMA Model Results

Dep. Variable:	D.Sparkling	No. Observations:	131
Model:	ARIMA(1, 1, 2)	Log Likelihood	-1111.799
Method:	css-mle	S.D. of innovations	1155.290
Date:	Sat, 09 Oct 2021	AIC	2233.598
Time:	18:34:25	BIC	2247.974
Sample:	02-29-1980 - 12-31-1990	HQIC	2239.439

	coef	std err	z	P>|z|	[0.025	0.975]
const	6.4581	4.211	1.534	0.125	-1.795	14.711
ar.L1.D.Sparkling	0.1896	0.166	1.143	0.253	-0.135	0.515
ma.L1.D.Sparkling	-0.6951	0.153	-4.548	0.000	-0.995	-0.396
ma.L2.D.Sparkling	-0.3049	0.152	-2.009	0.045	-0.602	-0.007

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	5.2740	+0.0000j	5.2740	0.0000
MA.1	1.0000	+0.0000j	1.0000	0.0000
MA.2	-3.2802	+0.0000j	3.2802	0.5000

Fig-1.43

Prediction on Test Set: -

	RMSE
ARIMA(2,1,2)	1374.759476
SARIMA(1,1,2)(2,0,2,6)	626.894467
ARIMA(1,1,2)	1436.731556

Manual ARIMA Model – Rose Data:-

- We now built manual ARIMA model for Rose Sales based on the ACF and PACF plots. Hence we chose the AR parameter p value 5, Moving average parameter q value 2 and d =1 based on below plots.

ACF Plot:-

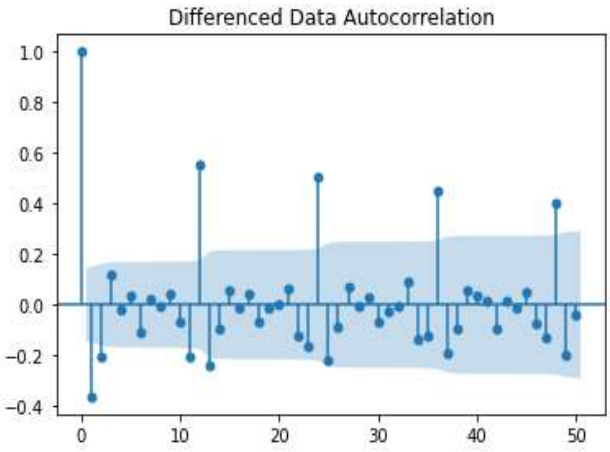


Fig-1.43

PACF Plot:-

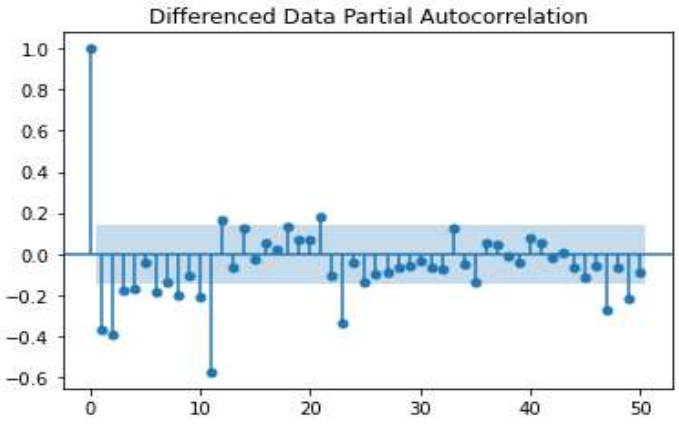


Fig-1.44

Observation:

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 5.
- 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 can say that both the PACF and ACF plot cuts-off at lag 5 & 2 respectively.

ARIMA Model Result – Rose Data: -

ARIMA Model Results

Dep. Variable:	D.Rose	No. Observations:	131
Model:	ARIMA(5, 1, 2)	Log Likelihood	-633.565
Method:	css-mle	S.D. of innovations	29.960
Date:	Sat, 09 Oct 2021	AIC	1285.129
Time:	18:34:29	BIC	1311.006
Sample:	02-29-1980 - 12-31-1990	HQIC	1295.644

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.4908	0.084	-5.843	0.000	-0.655	-0.326
ar.L1.D.Rose	-0.7739	0.129	-6.013	0.000	-1.026	-0.522
ar.L2.D.Rose	0.1214	0.181	0.670	0.503	-0.234	0.477
ar.L3.D.Rose	0.0130	0.300	0.043	0.965	-0.575	0.601
ar.L4.D.Rose	0.0532	0.096	0.552	0.581	-0.136	0.242
ar.L5.D.Rose	-0.0644	0.101	-0.639	0.523	-0.262	0.133
ma.L1.D.Rose	-9.946e-08	0.047	-2.1e-06	1.000	-0.093	0.093
ma.L2.D.Rose	-1.0000	0.047	-21.065	0.000	-1.093	-0.907

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-1.0000	-0.0000j	1.0000	-0.5000
AR.2	-0.8867	-1.5841j	1.8154	-0.3312
AR.3	-0.8867	+1.5841j	1.8154	0.3312
AR.4	1.7993	-1.2132j	2.1702	-0.0944
AR.5	1.7993	+1.2132j	2.1702	0.0944
MA.1	1.0000	+0.0000j	1.0000	0.0000
MA.2	-1.0000	+0.0000j	1.0000	0.5000

Fig-1.45

Prediction on Test set:-

	RMSE
ARIMA(0,1,2)	15.617949
SARIMA(1,1,2)(2,0,2,6)	26.133705
ARIMA(5,1,2)	15.400620

Manual SARIMA Model – Sparkling Data:-

- We now built manual SARIMA model for Sparking Sales based on the ACF and PACF plots. Hence we chose the AR parameter p value 1, Moving average parameter q value 2 and d value 1 based on the below plots. We then derive the Seasonal parameters based on the seasonal cut-offs on the 6 months series plot and chose Seasonal AR parameter P value 2.

ACF Plot:

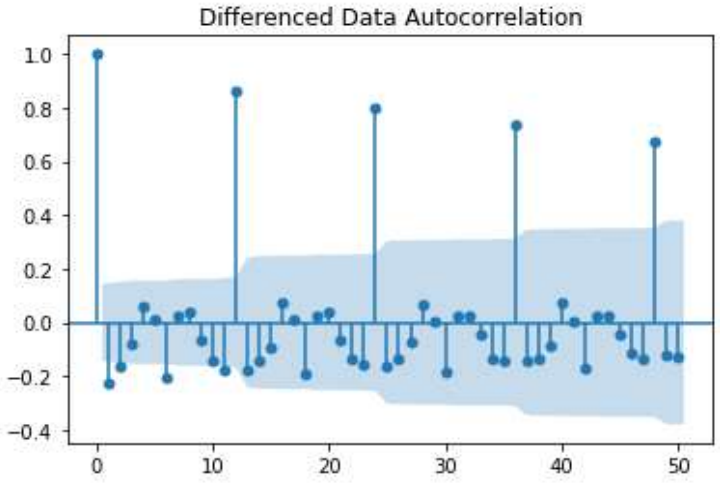


Fig-1.44

PACF Plot:

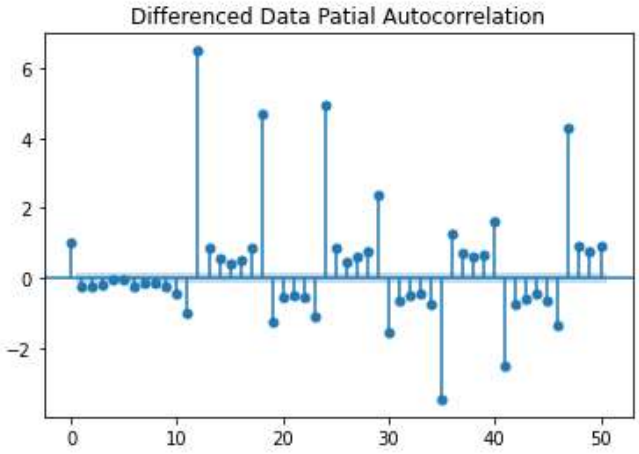


Fig-1.45

SARIMA Model Results:-

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1, 1, 2)x(2, 0, 2, 6)	Log Likelihood	-855.839
Date:	Sat, 09 Oct 2021	AIC	1727.679
Time:	18:34:32	BIC	1749.707
Sample:	0	HQIC	1736.621
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.6448	0.286	-2.256	0.024	-1.205	-0.085
ma.L1	-0.1069	0.250	-0.428	0.669	-0.596	0.383
ma.L2	-0.7006	0.202	-3.470	0.001	-1.096	-0.305
ar.S.L6	-0.0045	0.027	-0.165	0.869	-0.058	0.049
ar.S.L12	1.0361	0.018	56.061	0.000	1.000	1.072
ma.S.L6	0.0676	0.152	0.444	0.657	-0.231	0.366
ma.S.L12	-0.6122	0.093	-6.588	0.000	-0.794	-0.430
sigma2	1.448e+05	1.71e+04	8.464	0.000	1.11e+05	1.78e+05

Ljung-Box (L1) (Q):	0.09	Jarque-Bera (JB):	25.23
Prob(Q):	0.77	Prob(JB):	0.00
Heteroskedasticity (H):	2.63	Skew:	0.47
Prob(H) (two-sided):	0.00	Kurtosis:	5.09

Fig-1.47

Model Diagnostic Plot:

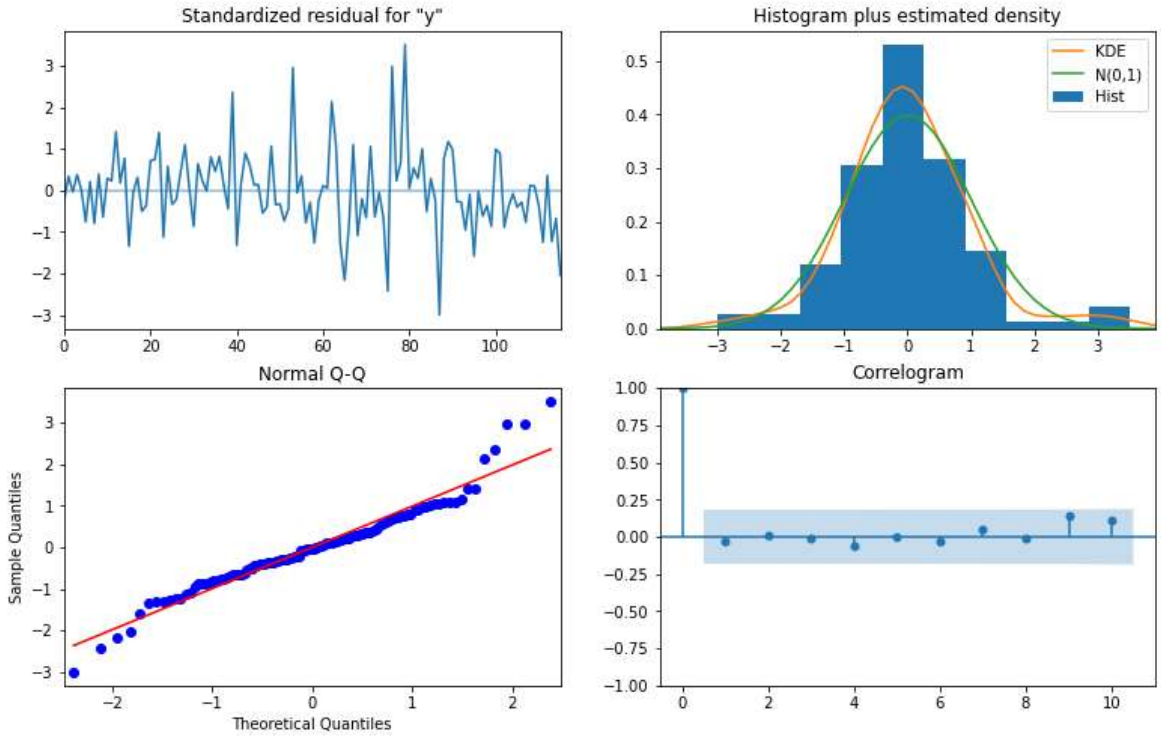


Fig-1.46

Observation:

- Model diagnostics confirms that the model residuals are normally distributed. Standardized residual do not display any obvious seasonality, Histogram plus estimated density – The KDE plot has normally distribution. The time series residuals have low correlation with lagged version of itself.

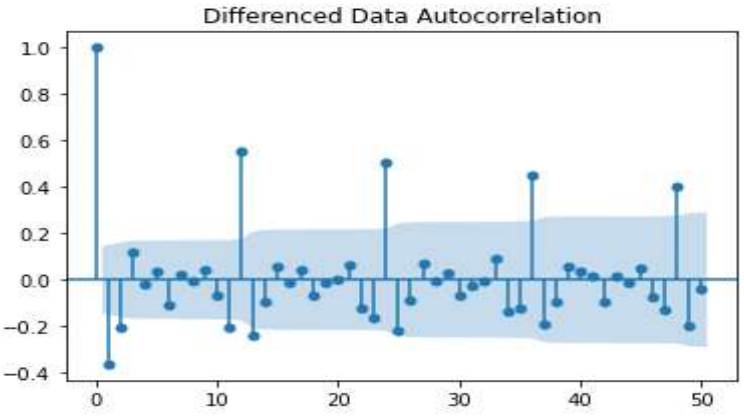
RMSE Score:

	RMSE
ARIMA(2,1,2)	1374.759476
SARIMA(1,1,2)(2,0,2,6)	626.894467
ARIMA(1,1,2)	1436.731556
SARIMA(1,1,2)(2,0,2,6)	626.894467

Manual SARIMA Model – Rose Data:-

- We now built manual SARIMA model for Sparking Sales based on the ACF and PACF plots. Hence we chose the AR parameter p value 4, Moving average parameter q value 2 and d value 1 based on the below plots. We then derive the Seasonal parameters based on the seasonal cut-offs on the 6 months series plot and chose Seasonal AR parameter P value 2.

ACF Plot: -



PACF plot:-

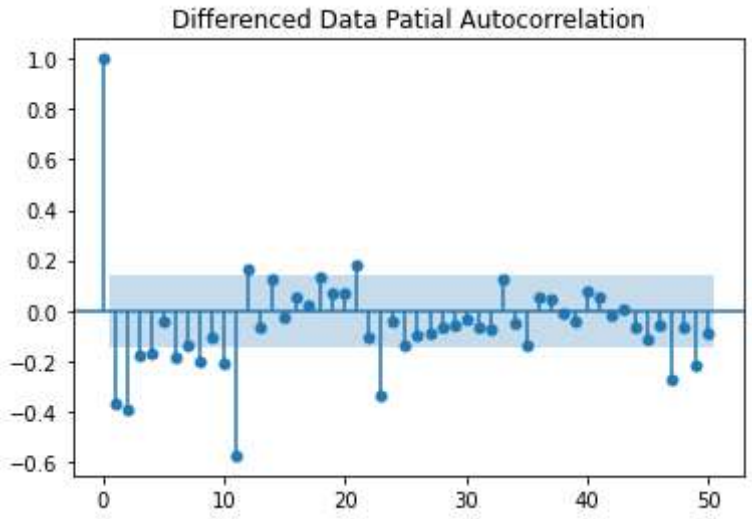


Fig-1.52

SARIMAX Model Result-Rose Data:-

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(4, 1, 2)x(2, 0, 2, 6)	Log Likelihood	-508.444
Date:	Sat, 09 Oct 2021	AIC	1038.887
Time:	18:34:37	BIC	1069.082
Sample:	0	HQIC	1051.143
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5877	0.243	-2.414	0.016	-1.065	-0.111
ar.L2	-0.0798	0.117	-0.684	0.494	-0.308	0.149
ar.L3	-0.0695	0.147	-0.471	0.637	-0.358	0.219
ar.L4	-0.0182	0.067	-0.274	0.784	-0.149	0.112
ma.L1	-0.2266	147.516	-0.002	0.999	-289.352	288.899
ma.L2	-0.7734	114.025	-0.007	0.995	-224.257	222.711
ar.S.L6	-0.0645	0.036	-1.815	0.070	-0.134	0.005
ar.S.L12	0.8427	0.034	24.490	0.000	0.775	0.910
ma.S.L6	0.2181	147.485	0.001	0.999	-288.847	289.283
ma.S.L12	-0.7820	115.279	-0.007	0.995	-226.724	225.160
sigma2	334.7658	0.704	475.523	0.000	333.386	336.146

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	43.27
Prob(Q):	0.99	Prob(JB):	0.00
Heteroskedasticity (H):	0.47	Skew:	0.53
Prob(H) (two-sided):	0.02	Kurtosis:	5.81

Fig-1.55

Model Diagnostic Plot:-

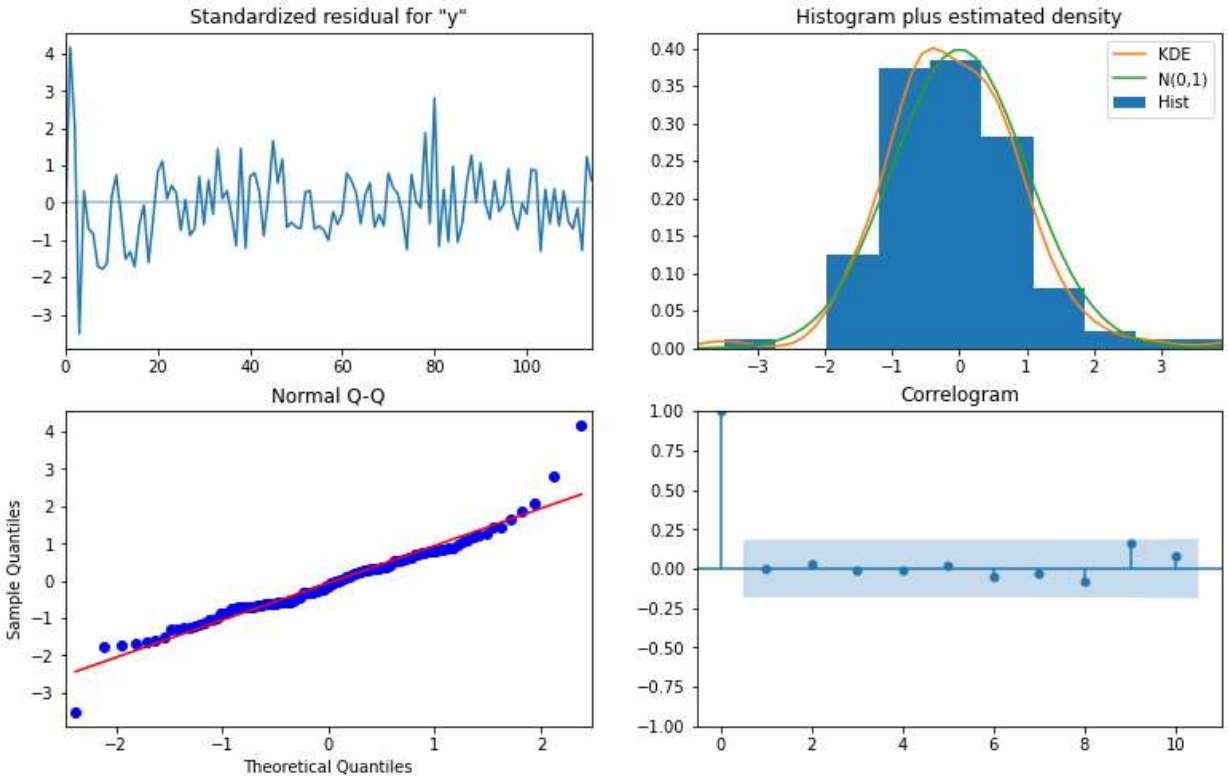


Fig-1.56

Observation:-

- Model diagnostics confirms that the model residuals are normally distributed. Standardized residual do not display any obvious seasonality, Histogram plus estimated density – The KDE plot has normally distribution. The time series residuals have low correlation with lagged version of itself. RMSE Score:

	RMSE
ARIMA(0,1,2)	15.617949
SARIMA(1,1,2)(2,0,2,6)	26.133705
ARIMA(5,1,2)	15.400620
SARIMA(4,1,2)(2,0,2,6)	26.187831

Fig-1.58

Q8. Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

Solution:-

Q9. 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.

Solution:-

Most Optimum Model on full data- Sparkling:-

- We observed from the RMSE scores that Triple Exponential Smoothing would work better for the Sparkling Sales data where Seasonality and trend are present in the data.

SARIMAX Model Result on Full data:-

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	187
Model:	SARIMAX(1, 1, 2)x(2, 0, 2, 6)	Log Likelihood	-1257.650
Date:	Sat, 09 Oct 2021	AIC	2531.301
Time:	18:34:42	BIC	2556.434
Sample:	01-31-1980 - 07-31-1995	HQIC	2541.499
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5767	0.331	-1.741	0.082	-1.226	0.073
ma.L1	-0.3873	0.301	-1.287	0.198	-0.977	0.202
ma.L2	-0.7391	0.330	-2.243	0.025	-1.385	-0.093
ar.S.L6	0.0073	0.017	0.435	0.664	-0.026	0.040
ar.S.L12	1.0178	0.011	93.233	0.000	0.996	1.039
ma.S.L6	-0.4310	0.105	-4.094	0.000	-0.637	-0.225
ma.S.L12	-0.9295	0.103	-9.014	0.000	-1.132	-0.727
sigma2	8.252e+04	1.34e+04	6.146	0.000	5.62e+04	1.09e+05

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	41.30
Prob(Q):	0.96	Prob(JB):	0.00
Heteroskedasticity (H):	1.26	Skew:	0.54
Prob(H) (two-sided):	0.38	Kurtosis:	5.15

Fig-1.61

Model Diagnostic Plot:-

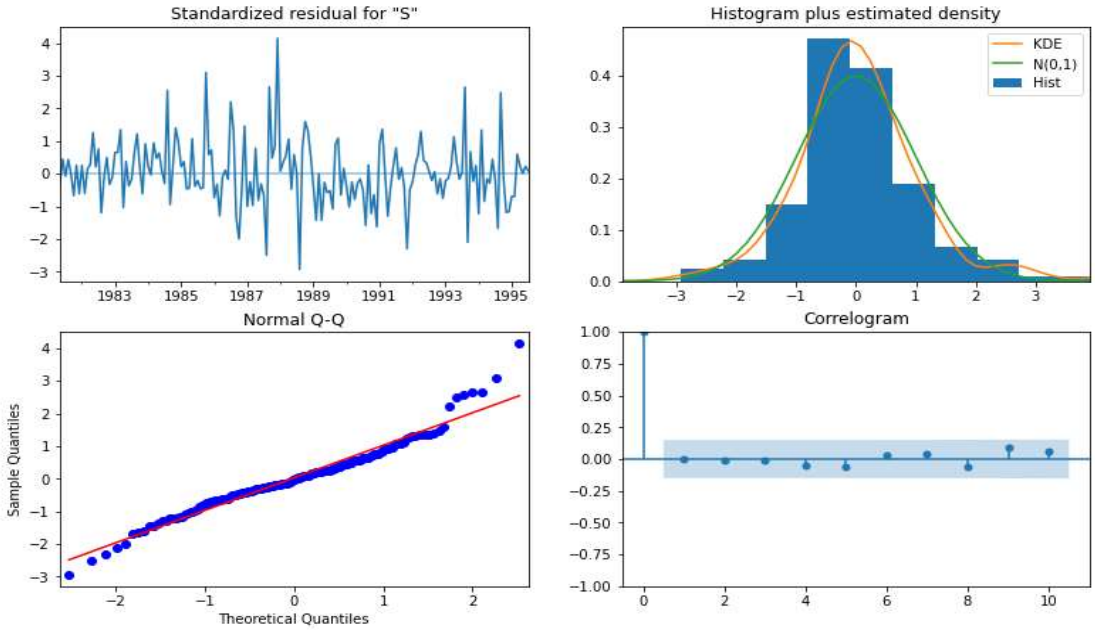


Fig-1.56

Observation:

- Model diagnostics confirms that the model residuals are normally distributed. Standardized residual do not display any obvious seasonality, Histogram plus estimated density – The KDE plot has normally distribution. The time series residuals have low correlation with lagged version of itself.

RMSE Score of Full Model

```
RMSE of the Full Model 531.216033517846
```

Prediction Plot- Sparkling Sales at 95% confidence Interval for next 12 months

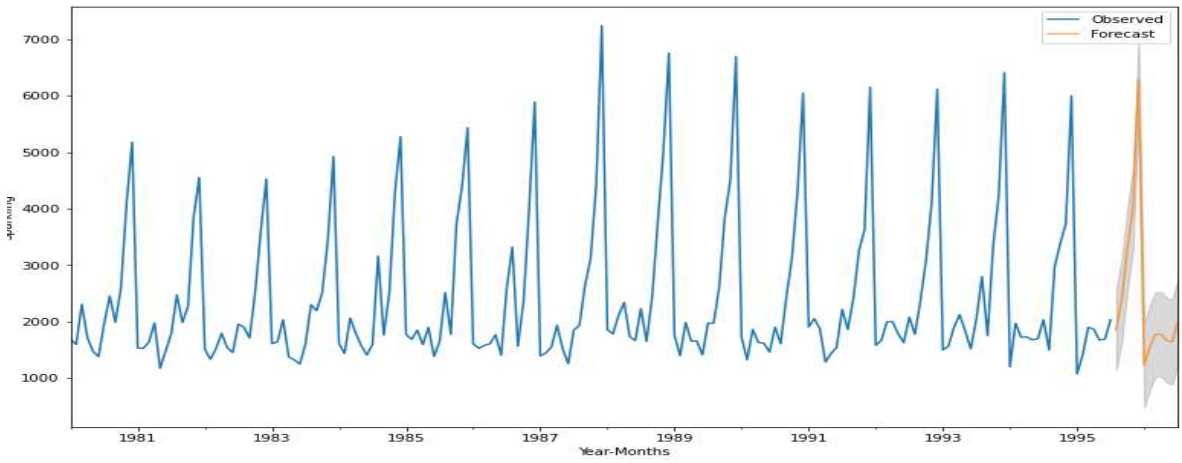


Fig-1.60

Observation:-

The prediction for next 12 months in future is in sync with the plot for previous year's data.

Most Optimum Model on full data- Rose

We observed from the RMSE scores that Triple Exponential Smoothing would work better for the Rose Sales data where Seasonality and trend are present in the data.

SARIMAX Model Result on Full data:-

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	187
Model:	SARIMAX(4, 1, 2)x(2, 0, 2, 6)	Log Likelihood	-727.094
Date:	Sat, 09 Oct 2021	AIC	1476.188
Time:	18:34:49	BIC	1510.682
Sample:	01-31-1980	HQIC	1490.185
	- 07-31-1995		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.6148	0.186	-3.313	0.001	-0.978	-0.251
ar.L2	-0.1217	0.126	-0.969	0.332	-0.368	0.124
ar.L3	-0.0884	0.118	-0.750	0.453	-0.319	0.143
ar.L4	-0.0274	0.054	-0.511	0.609	-0.132	0.078
ma.L1	-0.1355	0.240	-0.565	0.572	-0.606	0.334
ma.L2	-0.6701	0.245	-2.730	0.006	-1.151	-0.189
ar.S.L6	-0.0288	0.027	-1.071	0.284	-0.082	0.024
ar.S.L12	0.8912	0.024	36.482	0.000	0.843	0.939
ma.S.L6	0.1805	0.206	0.877	0.381	-0.223	0.584
ma.S.L12	-0.8785	0.191	-4.611	0.000	-1.252	-0.505
sigma2	257.8606	63.739	4.046	0.000	132.934	382.787

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	130.11
Prob(Q):	0.96	Prob(JB):	0.00
Heteroskedasticity (H):	0.18	Skew:	0.57
Prob(H) (two-sided):	0.00	Kurtosis:	7.13

Fig-1.63

Model Diagnostic Plot: Rose

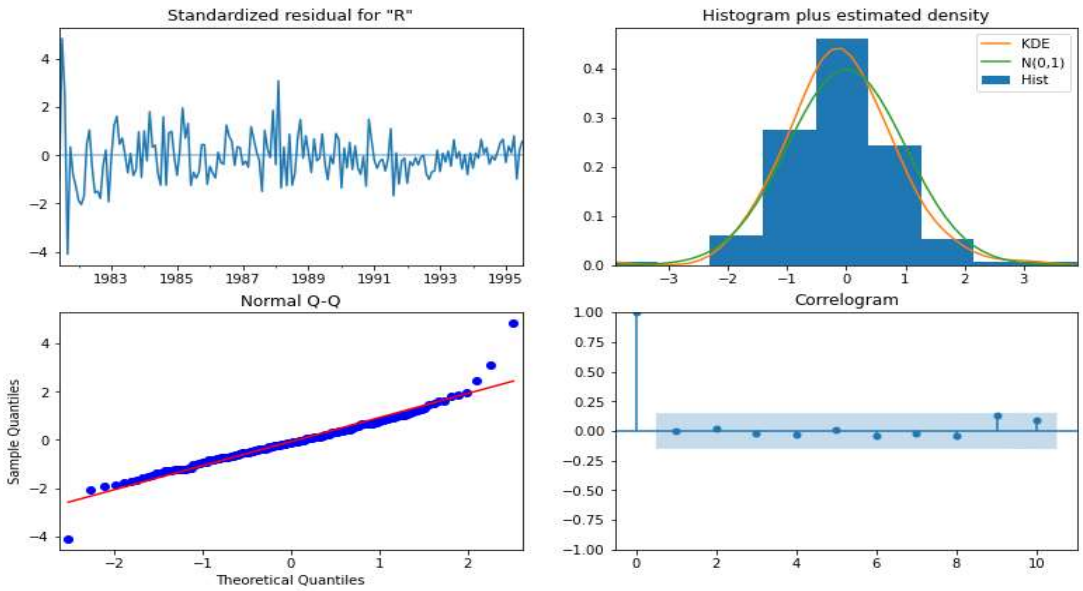


Fig-1.64

Observation:

- Model diagnostics confirms that the model residuals are normally distributed. Standardized residual do not display any obvious seasonality, Histogram plus estimated density – The KDE plot has normally distribution. The time series residuals have low correlation with lagged version of itself.

RMSE Score of Full Model: Rose

```
RMSE of the Full Model_Rose 26.061959085934813
```

Prediction Plot- Rose Sales at 95% confidence Interval for next 12 months

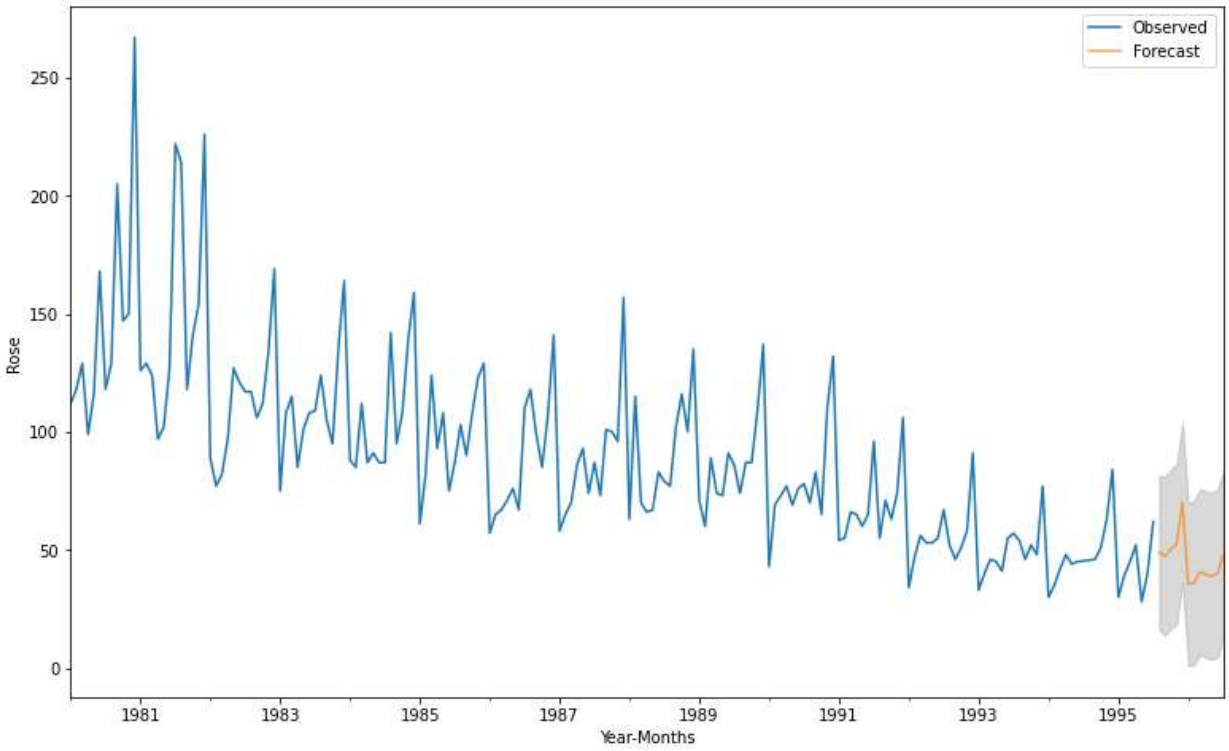


Fig-1.62

Observation:

The prediction for next 12 months in future is in sync with the plot for previous year's data. Most of observation are mention in all report.

Q10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

Solution:-

Inference and Recommendations

Sparkling Sales Model:-

- Sparkling sales shows stabilized values and not much trend compared to previous years.
- December month shows the highest Sales across the years from 1980-1994.
- The models are built considering the Trend and Seasonality in to account and we see from the output plot that the future prediction is in line with the trend and seasonality in the previous years.
- The Sales of Sparkling wine is seasonal, hence the company cannot have the same stock through the year. The predictions would help here to plan the Stock need basis the forecasted sales.
- The company should use the prediction results and capitalize on the high demand seasons and ensure to source and supply the high demand.
- The company should use the prediction results to plan the low demand seasons to stock as per the demand.

Inference and Recommendations

Rose Sales Model:-

- Rose sales shows a decrease in trend compared to the previous years.
- December month shows the highest Sales across the year while the value has come down through the years from 1980-1994.
- The models are built considering the Trend and Seasonality in to account and we see from the output plot that the future prediction is in line with the trend and seasonality in the previous years.
- The Sales of Rose wine is seasonal and also has trend, hence the company cannot have the same stock through the year. The predictions would help here to plan the Stock need basis the forecasted sales.
- The company should use the prediction results and capitalize on the high demand seasons and ensure to source and supply the high demand.
- The company should use the prediction results to plan the low demand seasons to stock as per the demand.