PGPDSBA-Time Series Forecast Project

Sparkling and Rose wine sale analysis

Submitted by,

Deepa. K

Table of Contents

List of Figures:

List of Tables:

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

Solution:

Dataset 1: **Sparkling**

- Datatype: Object datatype with no null values present.
- Shape : (130,57)
- Missing values: No
- Year range: 1980-1995

Data description(Summary statistics):

	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

Figure 7:Summary

Plot:

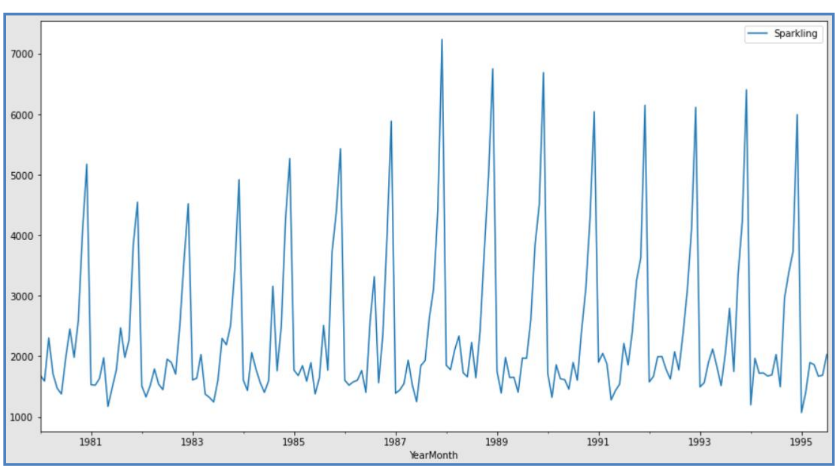

Figure 8:Sparkling-Plot

Dataset 2: **Rose**

- Datatype: Float datatype with 2 null values present.
- Shape : (130,57)
- Missing values: Interpolated with linear method
- Year range: 1980-1995

Data description(Summary statistics):

	Rose
count	185.000000
mean	90.394595
std	39.175344
min	28.000000
25%	63.000000
50%	86.000000
75%	112.000000
max	267.000000

Figure 9:Summary

Plot:

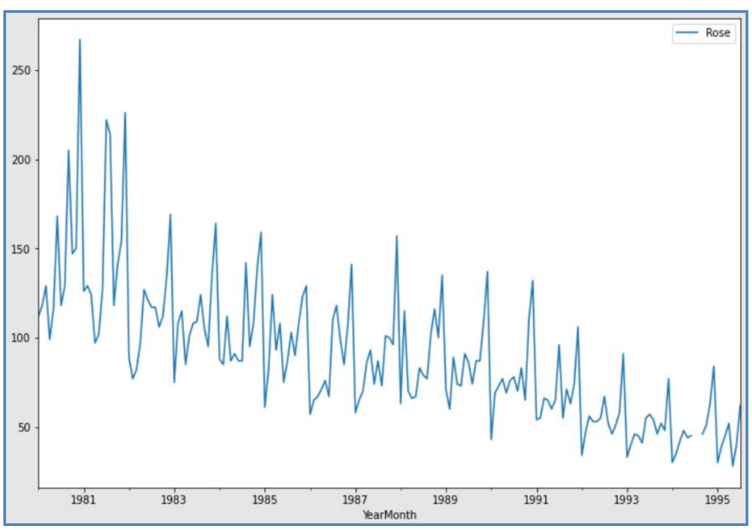

Figure 10:Rose-Plot

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

Solution:

Dataset 1: **Sparkling**

Univariate Time series: This time series data has single time stamped variable at time t and the year range is from January 1980 to July 1995.

Bivariate Time series: The below plots show year wise and month wise sales of Sparkling wine in 20th century.

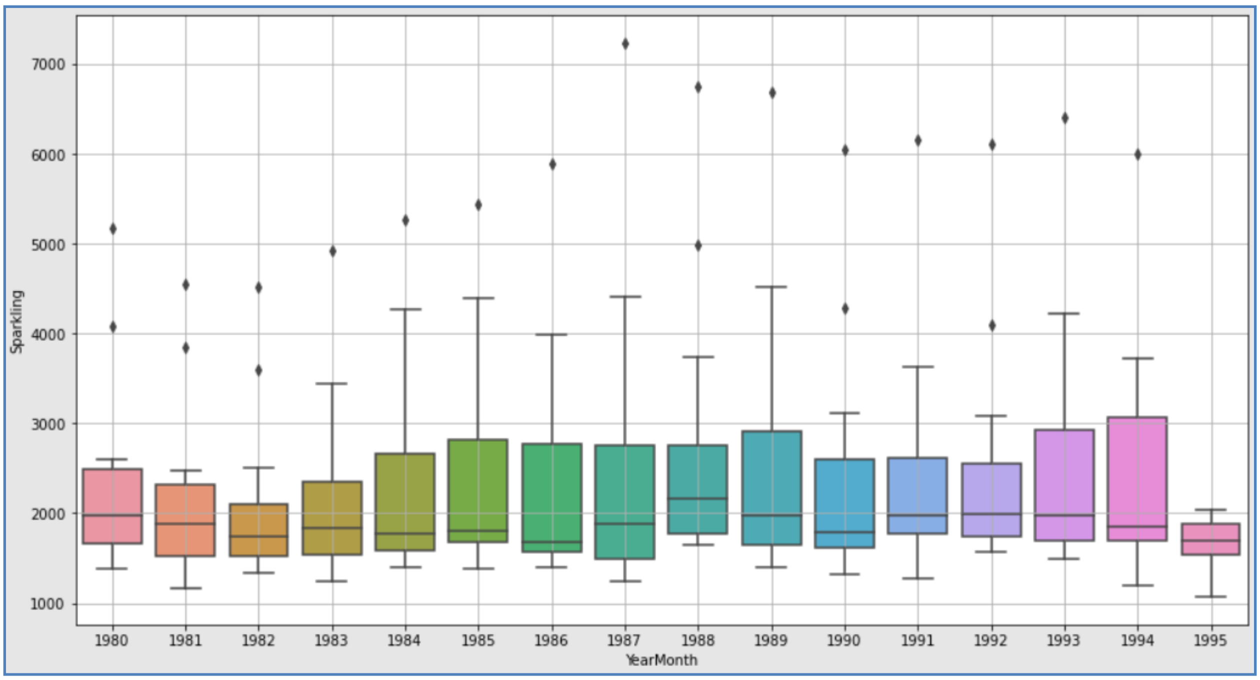

Figure 11:Year wise sale

- From the above plot,it is clear that 1985,1987 and 1989 years sales are comparatively higher as compared to other years.
- 1995-This year has very lesser sale.

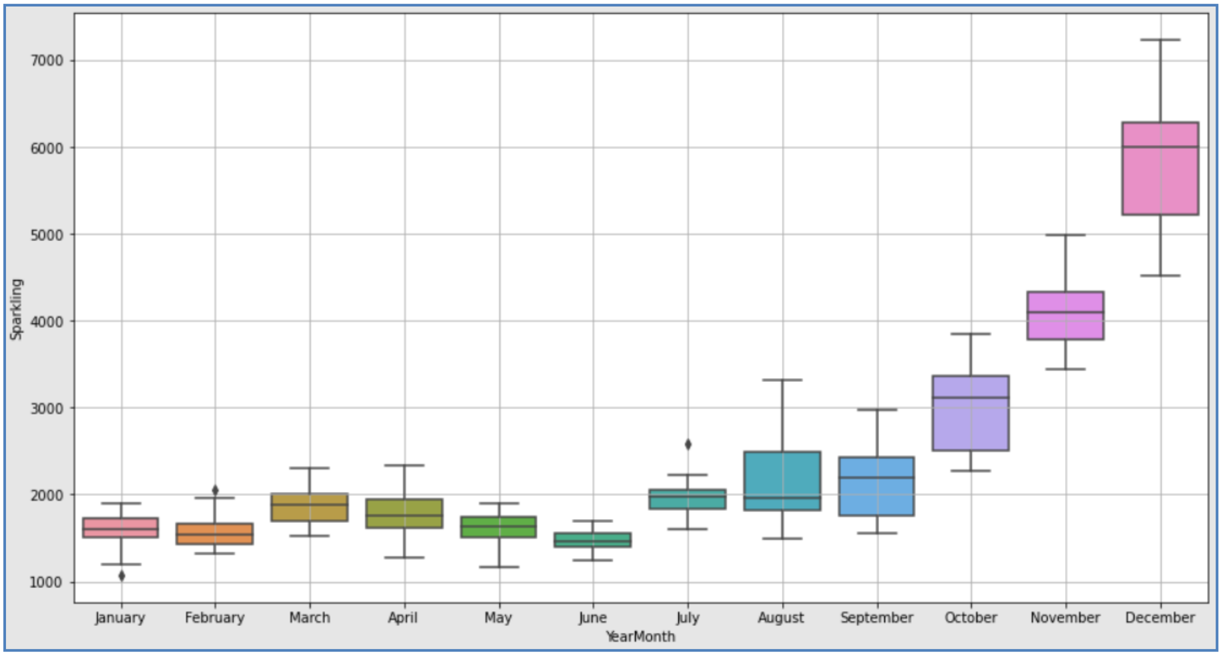

Figure 12:Month wise sale

- December month shows higher sale followed by November due to Christmas holidays and international vacation duration.
- June month has lesser sale than any other months.

Year-Month Table: No datapoints found after July-1995.

YearMonth	1	2	3	4	5	6	7	8	9	10	11	12
YearMonth												
1980	1686.0	1591.0	2304.0	1712.0	1471.0	1377.0	1966.0	2453.0	1984.0	2596.0	4087.0	5179.0
1981	1530.0	1523.0	1633.0	1976.0	1170.0	1480.0	1781.0	2472.0	1981.0	2273.0	3857.0	4551.0
1982	1510.0	1329.0	1518.0	1790.0	1537.0	1449.0	1954.0	1897.0	1706.0	2514.0	3593.0	4524.0
1983	1609.0	1638.0	2030.0	1375.0	1320.0	1245.0	1600.0	2298.0	2191.0	2511.0	3440.0	4923.0
1984	1609.0	1435.0	2061.0	1789.0	1567.0	1404.0	1597.0	3159.0	1759.0	2504.0	4273.0	5274.0
1985	1771.0	1682.0	1846.0	1589.0	1896.0	1379.0	1645.0	2512.0	1771.0	3727.0	4388.0	5434.0
1986	1606.0	1523.0	1577.0	1605.0	1765.0	1403.0	2584.0	3318.0	1562.0	2349.0	3987.0	5891.0
1987	1389.0	1442.0	1548.0	1935.0	1518.0	1250.0	1847.0	1930.0	2638.0	3114.0	4405.0	7242.0
1988	1853.0	1779.0	2108.0	2336.0	1728.0	1661.0	2230.0	1645.0	2421.0	3740.0	4988.0	6757.0
1989	1757.0	1394.0	1982.0	1650.0	1654.0	1406.0	1971.0	1968.0	2608.0	3845.0	4514.0	6694.0
1990	1720.0	1321.0	1859.0	1628.0	1615.0	1457.0	1899.0	1605.0	2424.0	3116.0	4286.0	6047.0
1991	1902.0	2049.0	1874.0	1279.0	1432.0	1540.0	2214.0	1857.0	2408.0	3252.0	3627.0	6153.0
1992	1577.0	1667.0	1993.0	1997.0	1783.0	1625.0	2076.0	1773.0	2377.0	3088.0	4096.0	6119.0
1993	1494.0	1564.0	1898.0	2121.0	1831.0	1515.0	2048.0	2795.0	1749.0	3339.0	4227.0	6410.0
1994	1197.0	1968.0	1720.0	1725.0	1674.0	1693.0	2031.0	1495.0	2968.0	3385.0	3729.0	5999.0
1995	1070.0	1402.0	1897.0	1862.0	1670.0	1688.0	2031.0	NaN	NaN	NaN	NaN	NaN

Figure 13:Year-month data

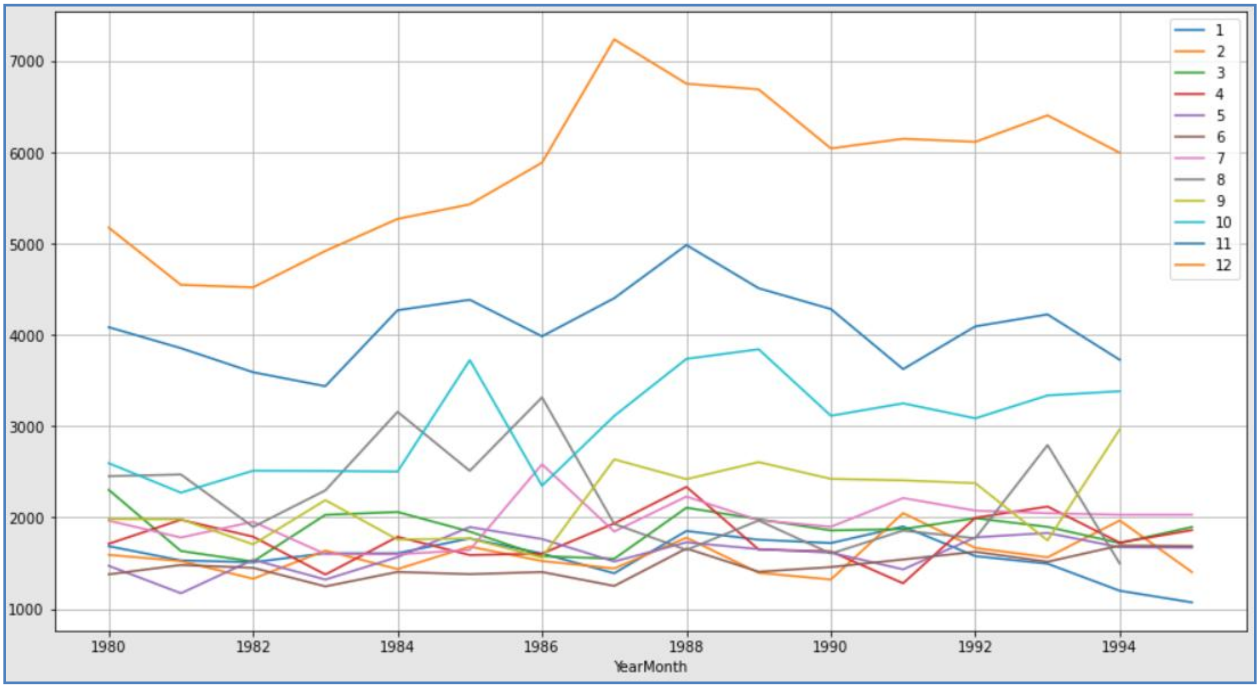

Figure 14:Line plot

- The above plot shows the line plot comparison of sparkling wine sale for 12 months ranging from 1980-1995.
- December month shows higher sale in almost all the years.

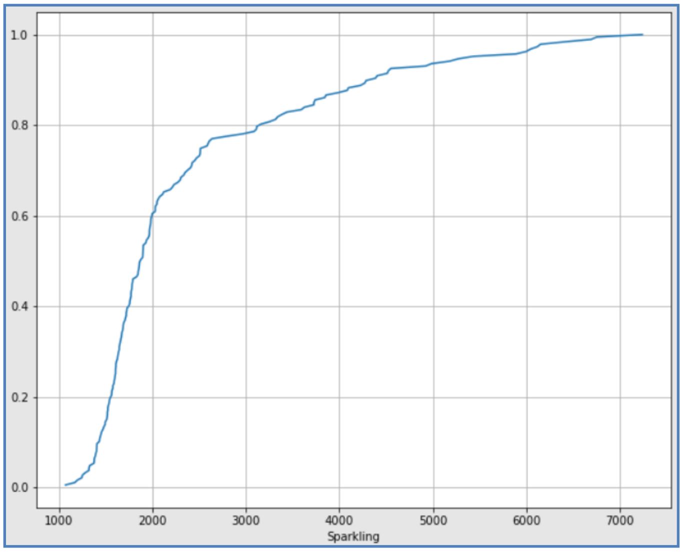

Figure 15:ECDF

- ECDF-Estimator of cumulative Distribution Function to plot the data points from lowest to highest value(min-1070 to max-7242).

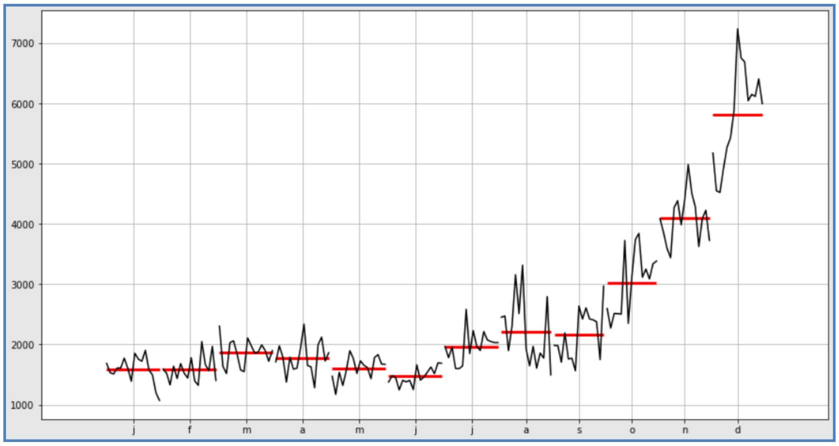

Figure 16:Month plot

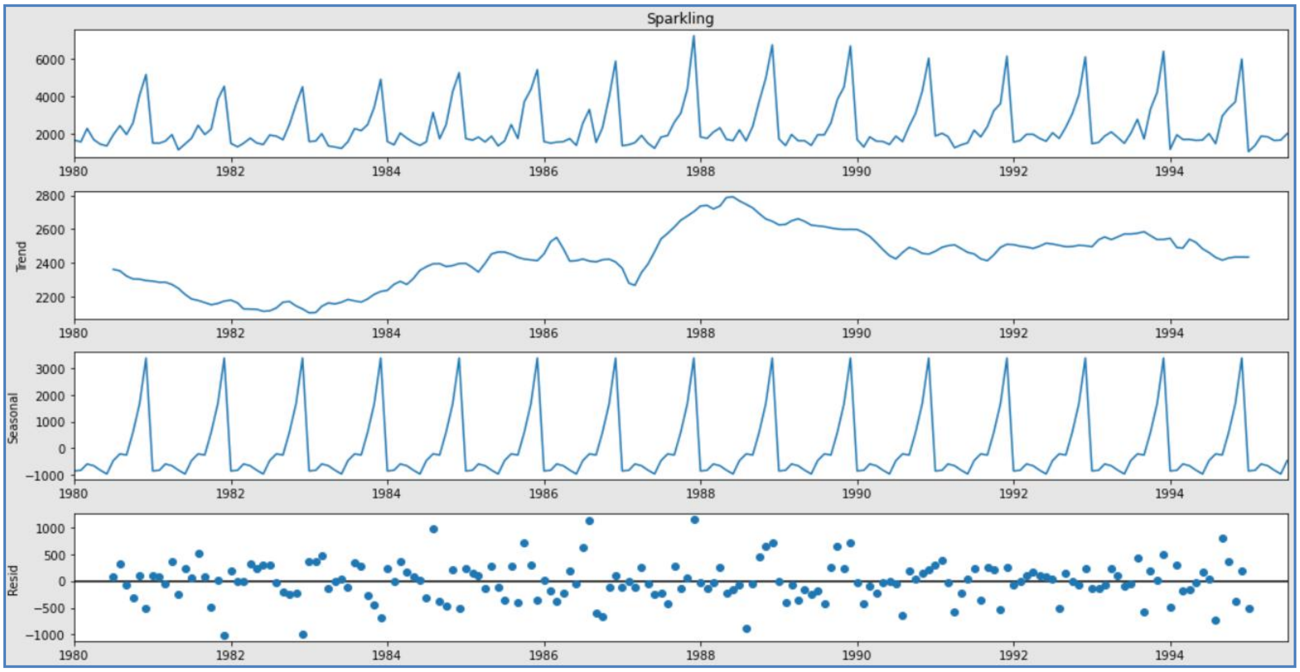

Figure 17:Additive decompose

- Additive decompose model adds the components together and the model shows both trend and seasonality.
- Here the linear seasonality is same.

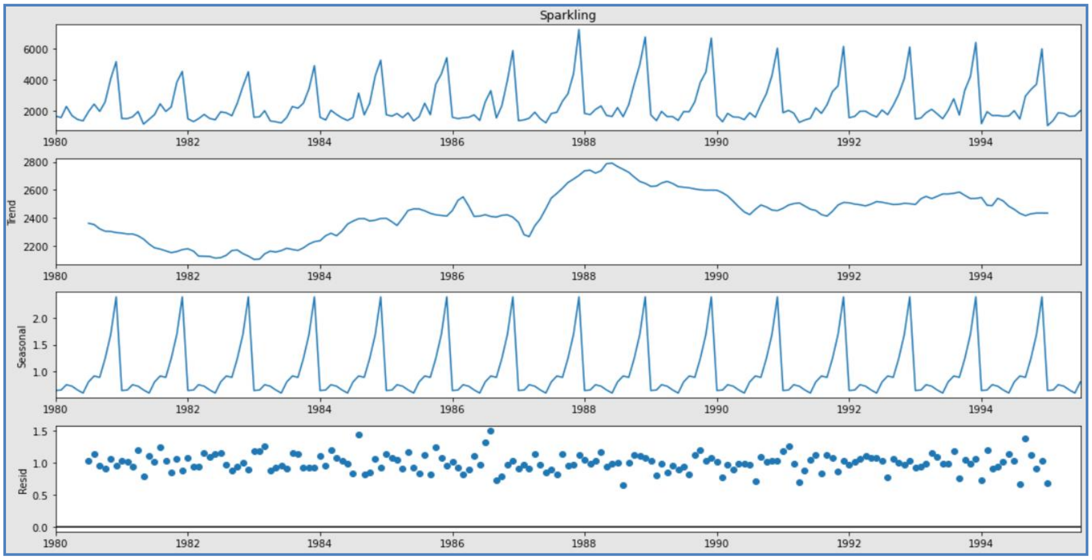

Figure 18:Multiplicative decompose

- Multiplicative decompose model multiplies the components together and it is non linear.
- Non linear seasonality is either increasing/decreasing frequency.

Dataset 2: **Rose**

Univariate Time series: This time series data has single time stamped variable at time t and the year range is from January 1980 to July 1995.

Bivariate Time series: The below plots show year wise and month wise sales of Rose wine in 20th century.

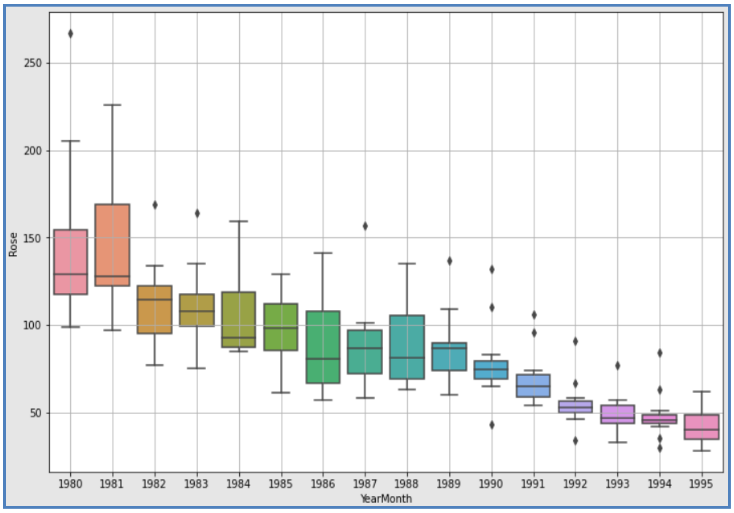

Figure 19:Year wise plot

- From the above plot,it is clear that 1980 and 1981 years sales are comparatively higher as compared to other years.
- 1995-This year has very lesser sale as only data is available until July month.

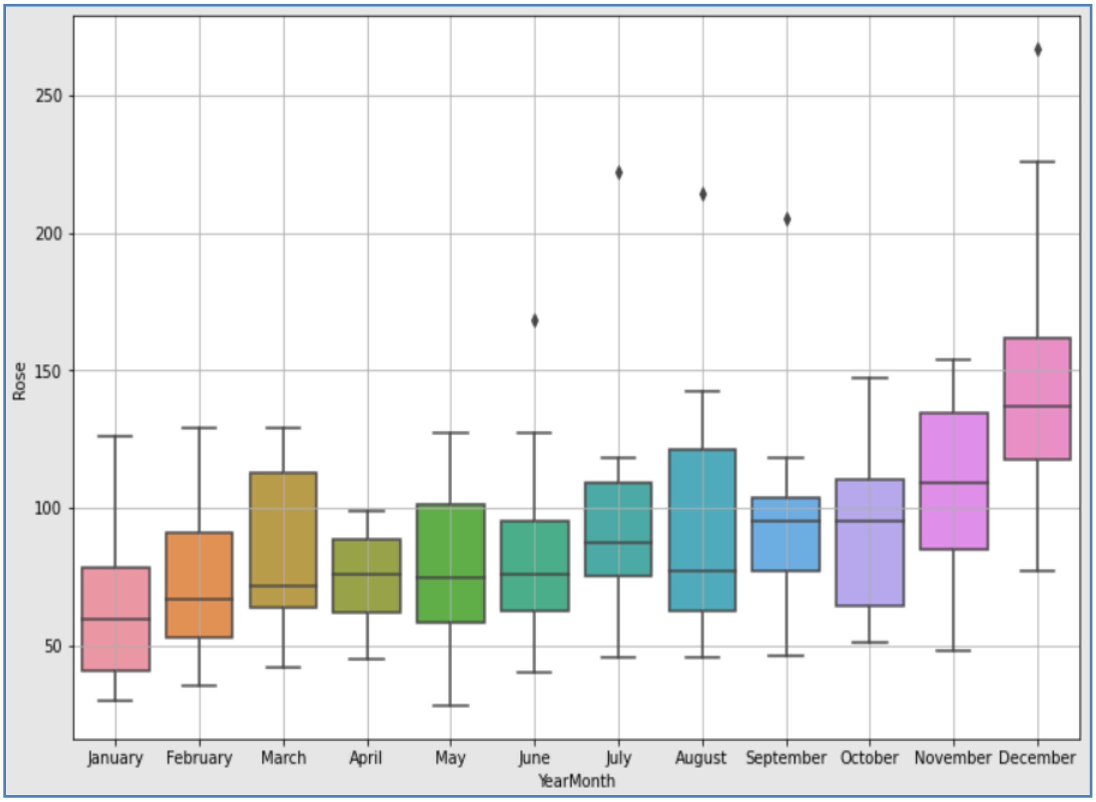

Figure 20:Month wise plot

- December month shows higher sale followed by November due to Christmas holidays and international vacation duration.
- April month shows lesser Rose wine sale.
- There are also few missing data points in the month of july.

Year/month table: No data points found after July-1995.

YearMonth	1	2	3	4	5	6	7	8	9	10	11	12
YearMonth												
1980	112.0	118.0	129.0	99.0	116.0	168.0	118.000000	129.000000	205.0	147.0	150.0	267.0
1981	126.0	129.0	124.0	97.0	102.0	127.0	222.000000	214.000000	118.0	141.0	154.0	226.0
1982	89.0	77.0	82.0	97.0	127.0	121.0	117.000000	117.000000	106.0	112.0	134.0	169.0
1983	75.0	108.0	115.0	85.0	101.0	108.0	109.000000	124.000000	105.0	95.0	135.0	164.0
1984	88.0	85.0	112.0	87.0	91.0	87.0	87.000000	142.000000	95.0	108.0	139.0	159.0
1985	61.0	82.0	124.0	93.0	108.0	75.0	87.000000	103.000000	90.0	108.0	123.0	129.0
1986	57.0	65.0	67.0	71.0	76.0	67.0	110.000000	118.000000	99.0	85.0	107.0	141.0
1987	58.0	65.0	70.0	86.0	93.0	74.0	87.000000	73.000000	101.0	100.0	96.0	157.0
1988	63.0	115.0	70.0	66.0	67.0	83.0	79.000000	77.000000	102.0	116.0	100.0	135.0
1989	71.0	60.0	89.0	74.0	73.0	91.0	86.000000	74.000000	87.0	87.0	109.0	137.0
1990	43.0	69.0	73.0	77.0	69.0	76.0	78.000000	70.000000	83.0	65.0	110.0	132.0
1991	54.0	55.0	66.0	65.0	60.0	65.0	96.000000	55.000000	71.0	63.0	74.0	106.0
1992	34.0	47.0	56.0	53.0	53.0	55.0	67.000000	52.000000	46.0	51.0	58.0	91.0
1993	33.0	40.0	46.0	45.0	41.0	55.0	57.000000	54.000000	46.0	52.0	48.0	77.0
1994	30.0	35.0	42.0	48.0	44.0	45.0	45.333333	45.666667	46.0	51.0	63.0	84.0
1995	30.0	39.0	45.0	52.0	28.0	40.0	62.000000	NaN	NaN	NaN	NaN	NaN

Figure 21:Year month data

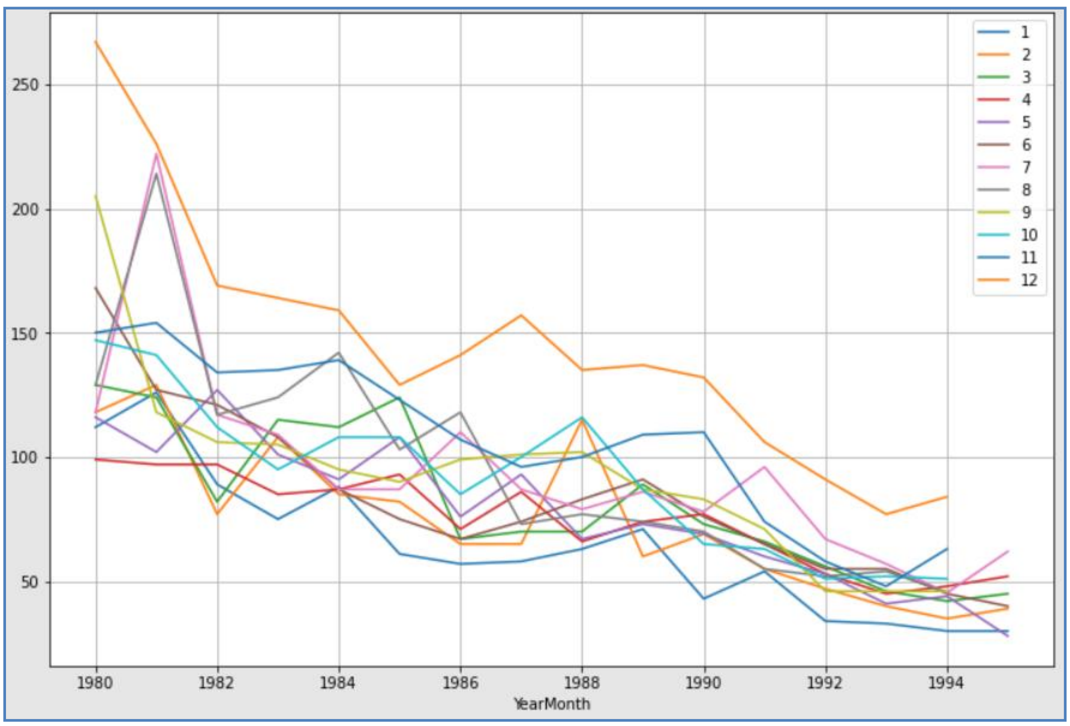

Figure 22:Line plot

- The above plot shows the line plot comparison of Rose wine sale for 12 months ranging from 1980-1995.
- December month shows higher sale in almost all the years.

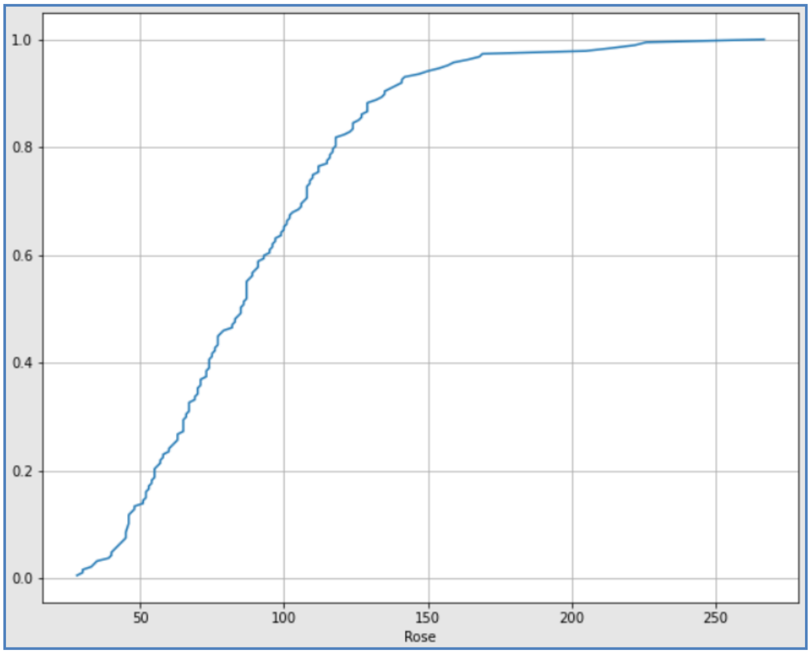

Figure 23:ECDF

- ECDF-Estimator of cumulative Distribution Function to plot the data points from lowest to highest value(min-28 to max-267).

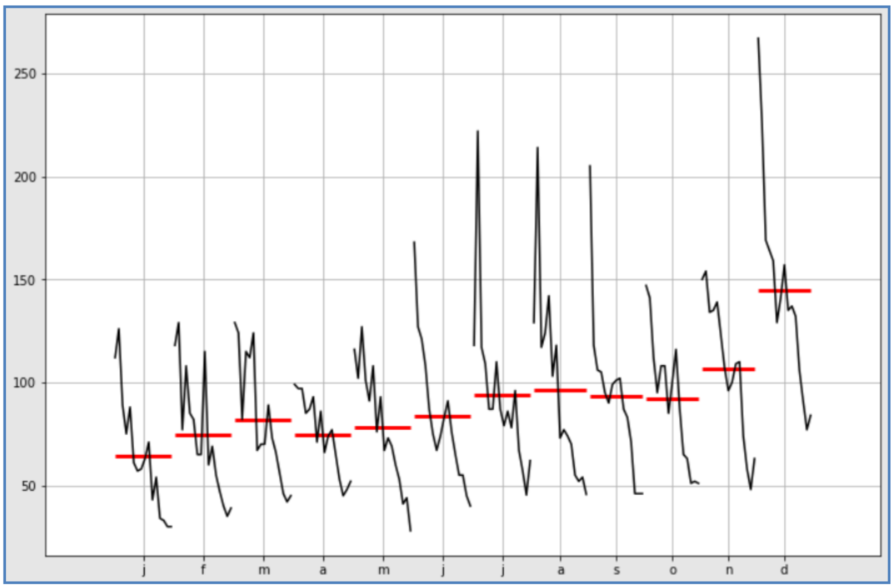

Figure 24:Month plot

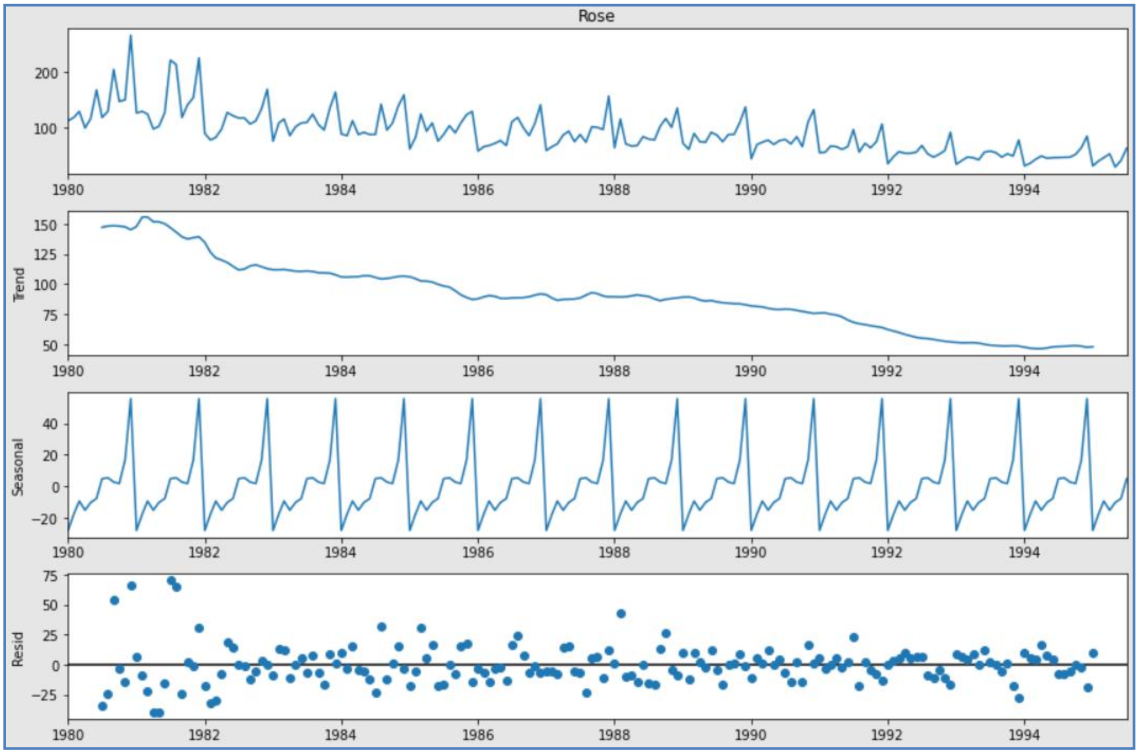

Figure 25: Additive decompose

- Additive decompose model adds the components together and the model shows both trend and seasonality.
- Here the linear seasonality is same.

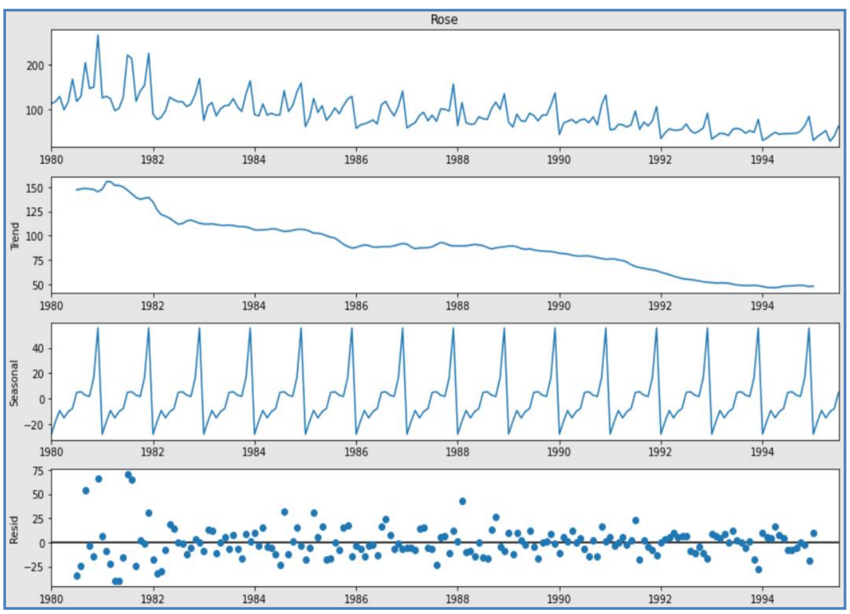

Figure 26: Multiplicative decompose

- Multiplicative decompose model multiplies the components together and it is non linear.
- Non linear seasonality is either increasing/decreasing frequency.

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

Solution:

Dataset 1: **Sparkling**

- Training data shape:(130,1)
- Test data shape:(57,1)
- The test data starts from 1991 and the previous years are into the training set.

First few rows of Training Data	Last few rows of training Data
1686	1605
1591	2424
2304	3116
1712	4286
1471	6047

Table 1:Training data points

First few rows of Test Data	Last few rows of test Data
1902	1897
2049	1862
1874	1670
1279	1688
1432	2031

Table 2:Test data points

Dataset 2: **Rose**

- Training data shape:(130,1)
- Test data shape:(57,1)
- The test data starts from 1991 and the previous years are into the training set.

First few rows of Training Data	Last few rows of training Data
112	70
118	83
129	65
99	110
116	132

Table 3:Training data points

First few rows of Test Data	Last few rows of test Data
54	45
55	52
66	28
65	40
60	62

Table 4:Test data points

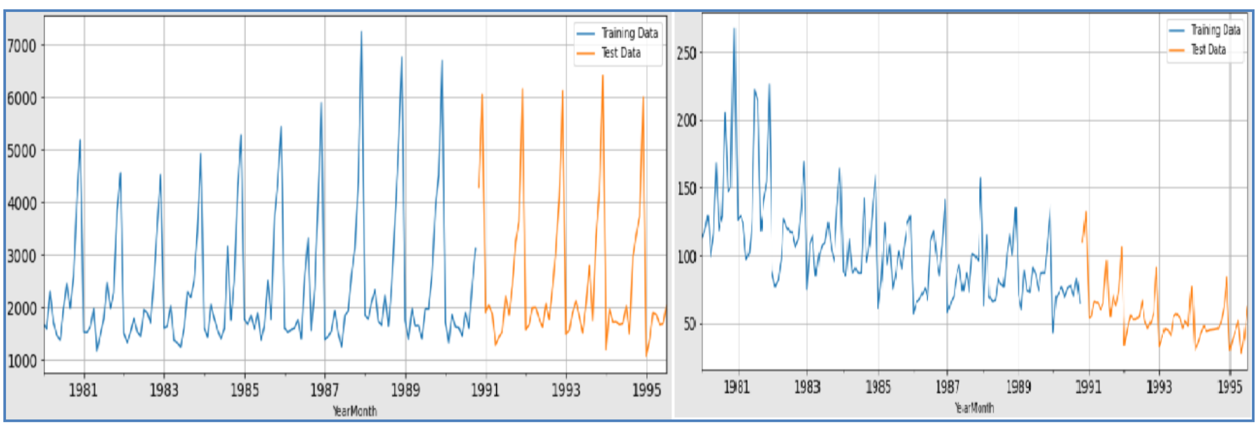

Figure 27:Joint plot-Sparkling&Rose

4. Build all the 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 and simple average models. should also be built on the training data and check the performance on the test data using RMSE.

Solution:

Linear Regression plot for Sparkling and Rose wine:

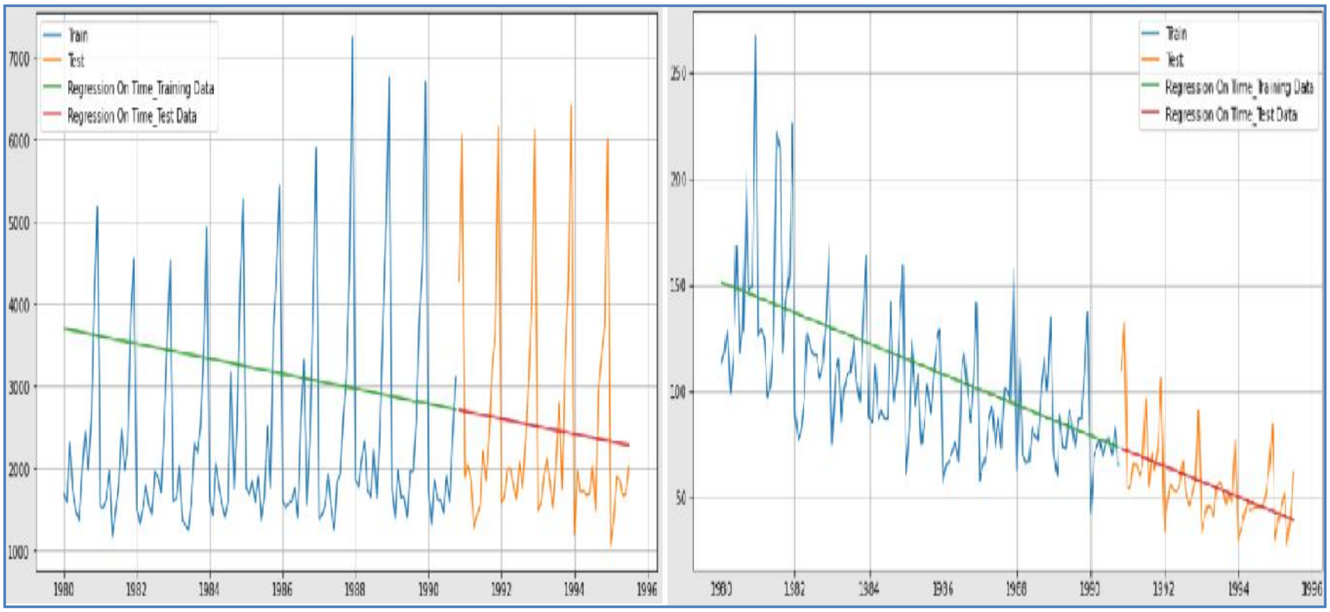

Figure 28: Linear Regression

- Test RMSE for Sparkling: 1356.3
- Test RMSE for Rose: 17.2

Naïve forecast plot for sparkling and Rose wine:

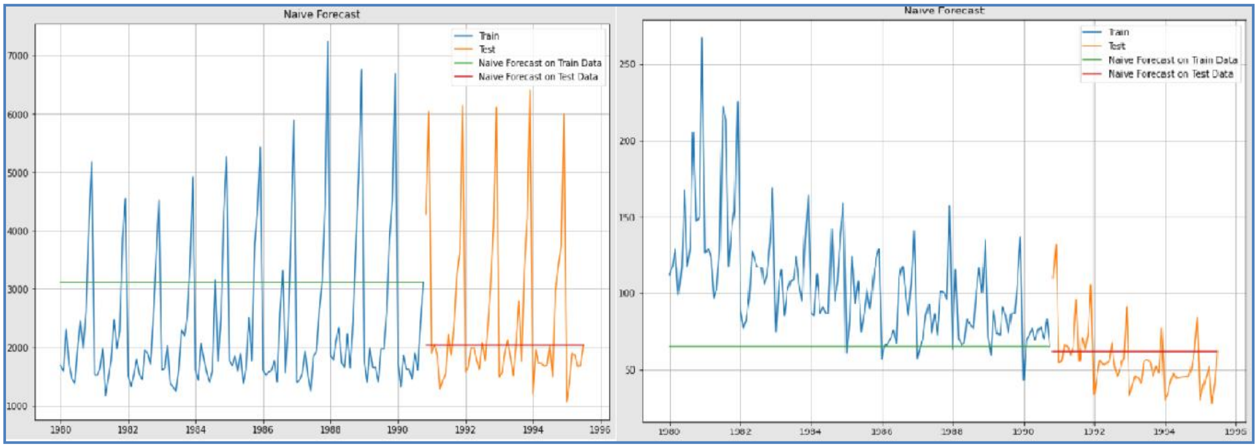

Figure 29: Naive forecast

- Test RMSE for Sparkling: 1439.3
- Test RMSE for Rose: 20.7

Simple Average forecast plot for Sparkling and Rose wine:

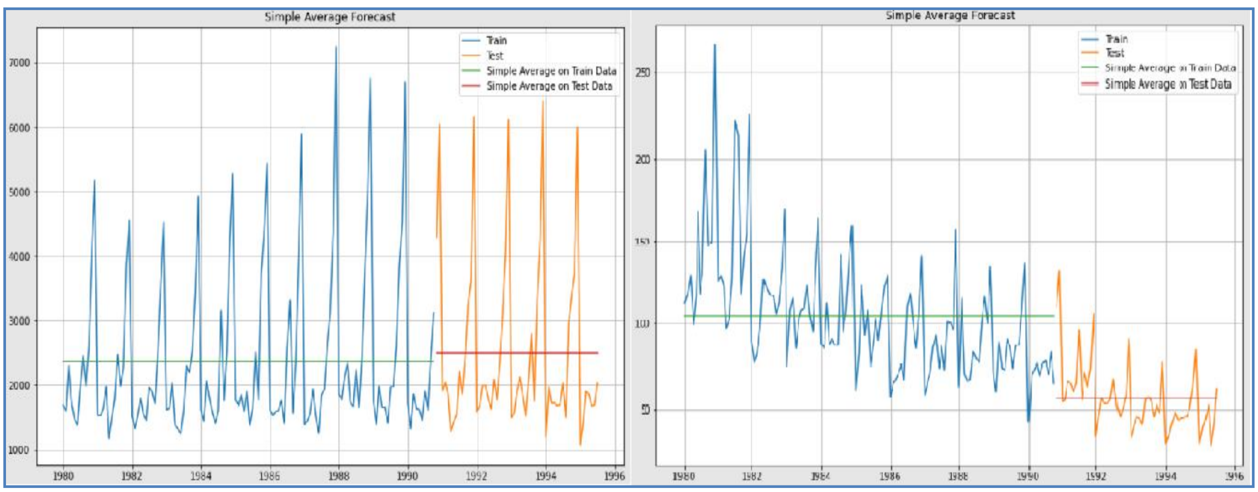

Figure 30: Simple average

- Test RMSE for Sparkling: 1362.07
- Test RMSE for Rose: 19.9

Moving Average plot for Sparkling and Rose wine:

- Rolling means for different intervals are taken into consideration and best interval is determined by the model with minimum error.Average of entire data is considered.

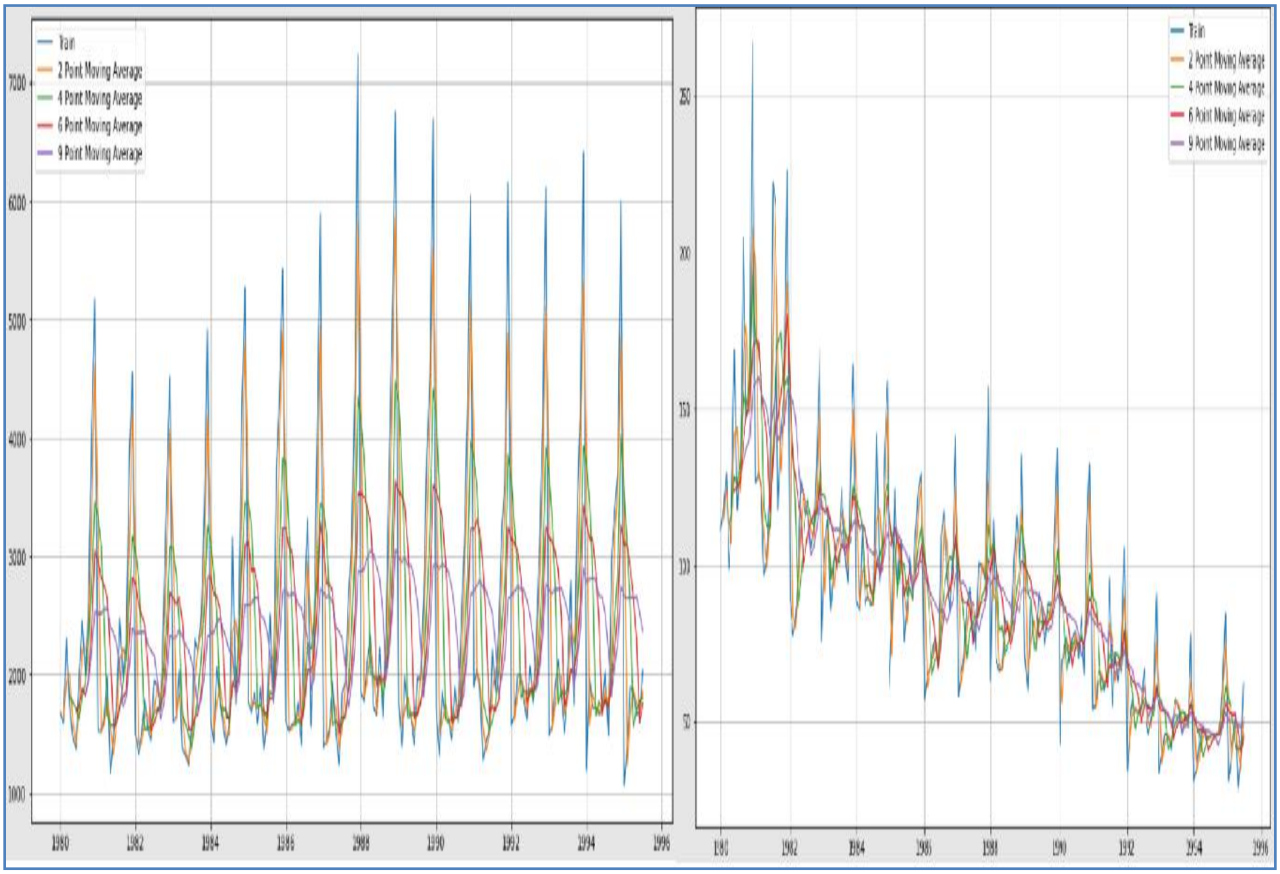

Figure 31:Moving average

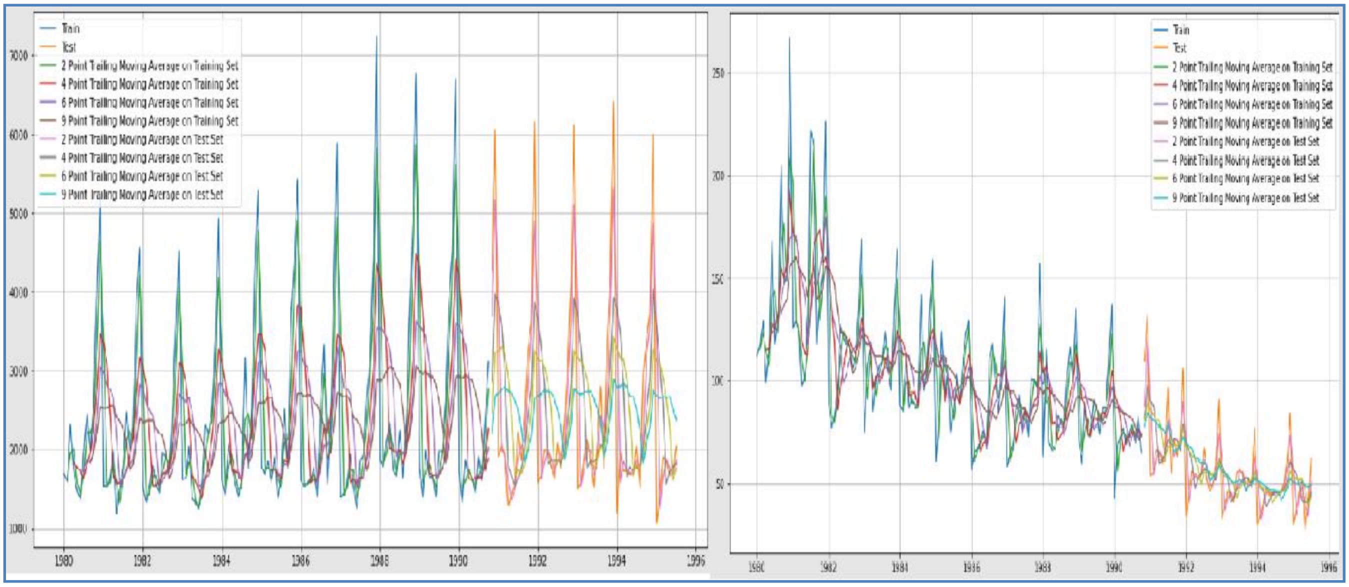

Figure 32: Rolling means

Sparkling:

- 2 point moving average on test data, RMSE is 811.17
- 4 point moving average on test data, RMSE is 1184.21
- 6 point moving average on test data, RMSE is 1337.20
- 9 point moving average on test data, RMSE is 1422.65

Rose:

- 2 point Moving Average Model on the Test Data, RMSE is 11.801
- 4 point Moving Average Model on the Test Data, RMSE is 15.367
- 6 point Moving Average Model on the Test Data, RMSE is 15.862
- 9 point Moving Average Model on the Test Data, RMSE is 16.342

Simple Exponential Smoothing(SES-Sparkling & Rose):

- The time series neither have pronounced trend nor seasonality(almost non available)
- The below plot(test set) is for sparkling dataset at alpha = 0.06

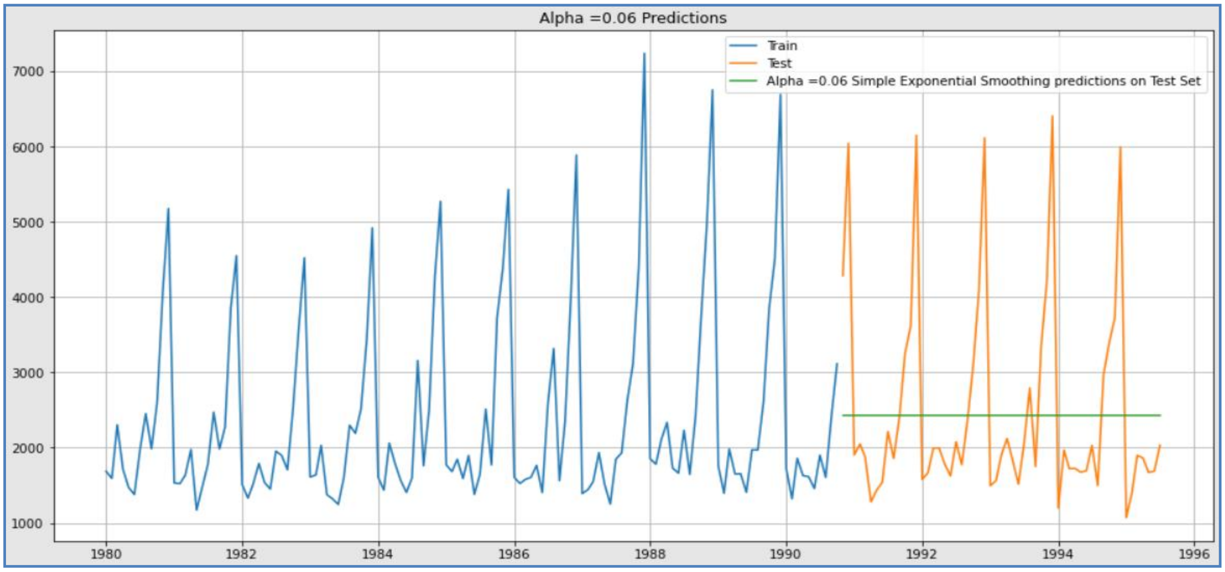

Figure 33: SES-Sparkling

- Test RMSE for SES Sparkling dataset at alpha=0.06 is 1363.702

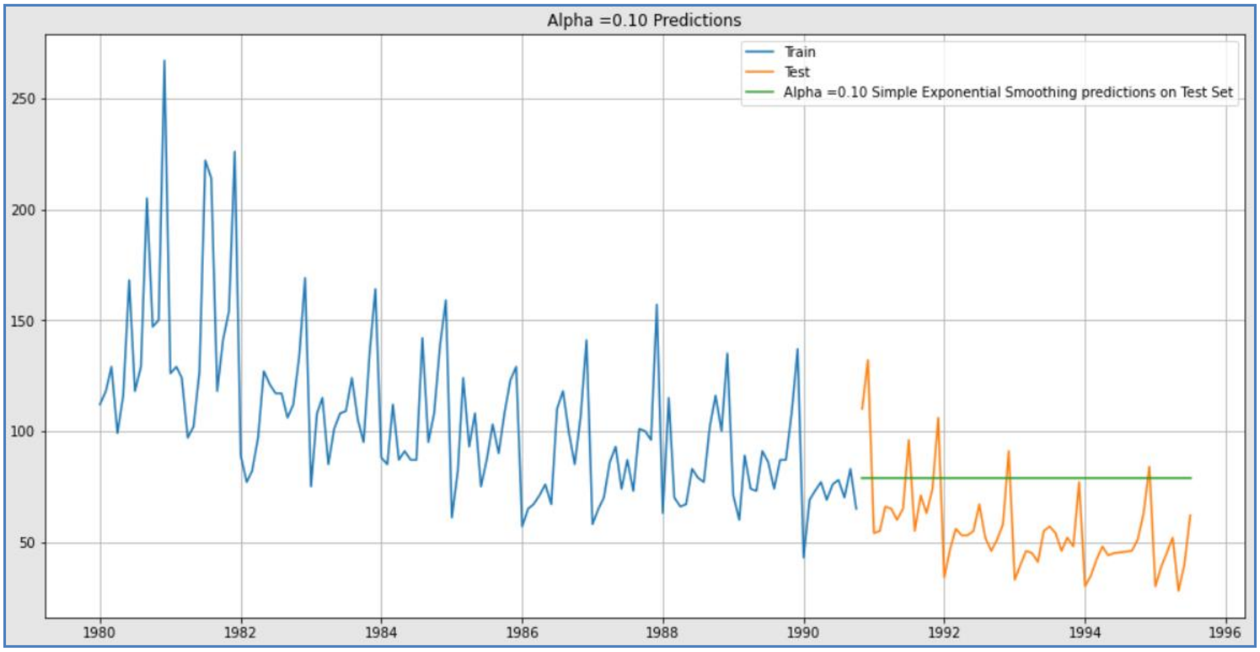

Figure 34: SES-Rose

- Test RMSE for SES Rose dataset at alpha=0.10 is 30.18

Double Exponential Smoothing(DES-Sparkling & Rose):

- Applicable when data has trend but no seasonality.
- Also known as Holt’s model.

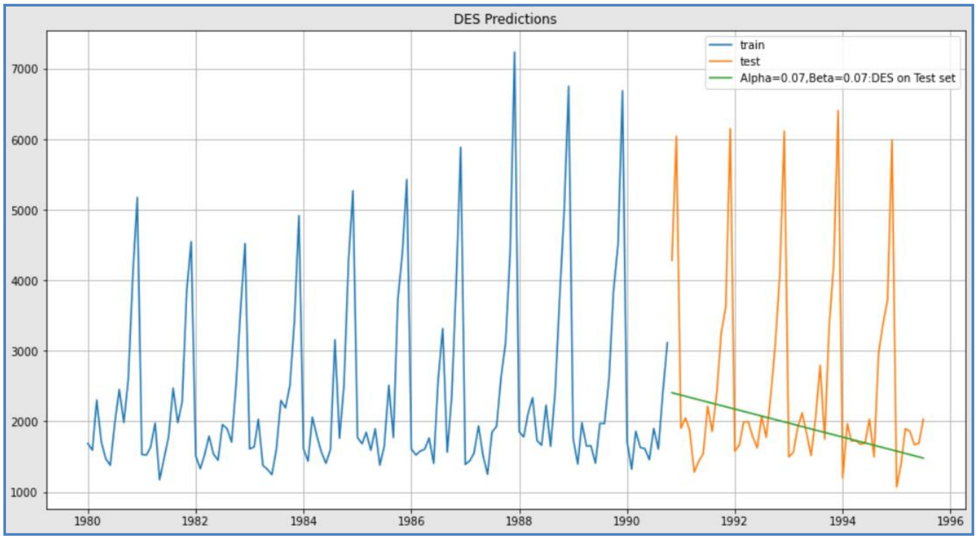

Figure 35:DES-Sparkling

- For Alpha=0.07,Beta=0.07 Double Exponential Smoothing Model forecast on the Test Data, RMSE is 1472.254

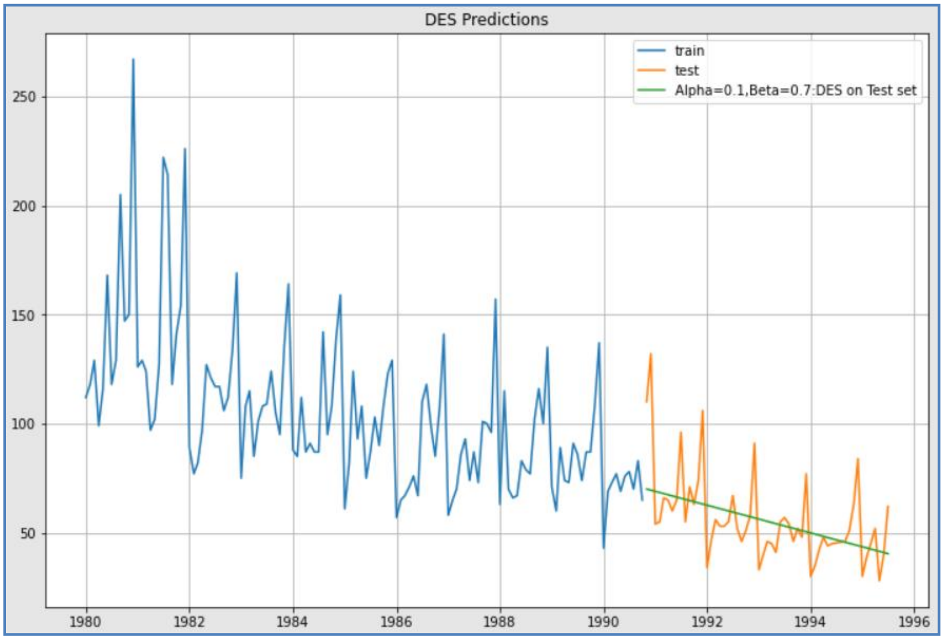

Figure 36:DES-Rose

- For Alpha=0.1,Beta=0.7 Double Exponential Smoothing Model forecast on the Test Data, RMSE is 17.356

Triple Exponential Smoothing(TES-Sparkling & Rose):

- TES(Holt winters) model can be additive or multiplicative that simultaneously smooths the level,trend and seasonality.
- Alpha=0.07,Beta=0.03,Gamma=0.47(Additive)

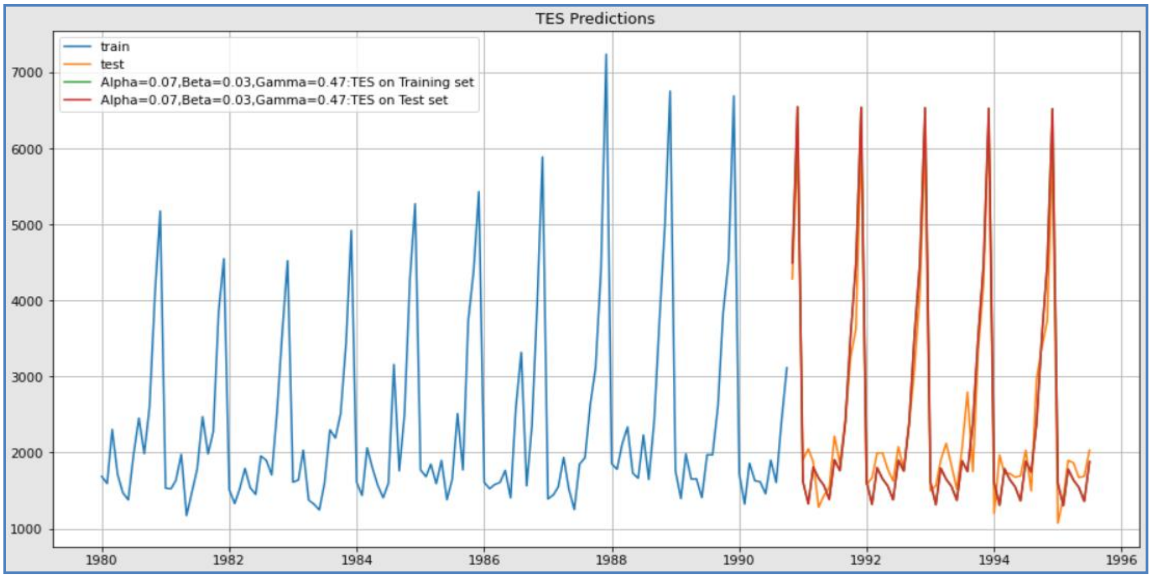

Figure 37:TES Additive-Sparkling

- For Alpha=0.07,Beta=0.03,Gamma=0.47 TES additive Model forecast on the Test Data, RMSE is 366.859

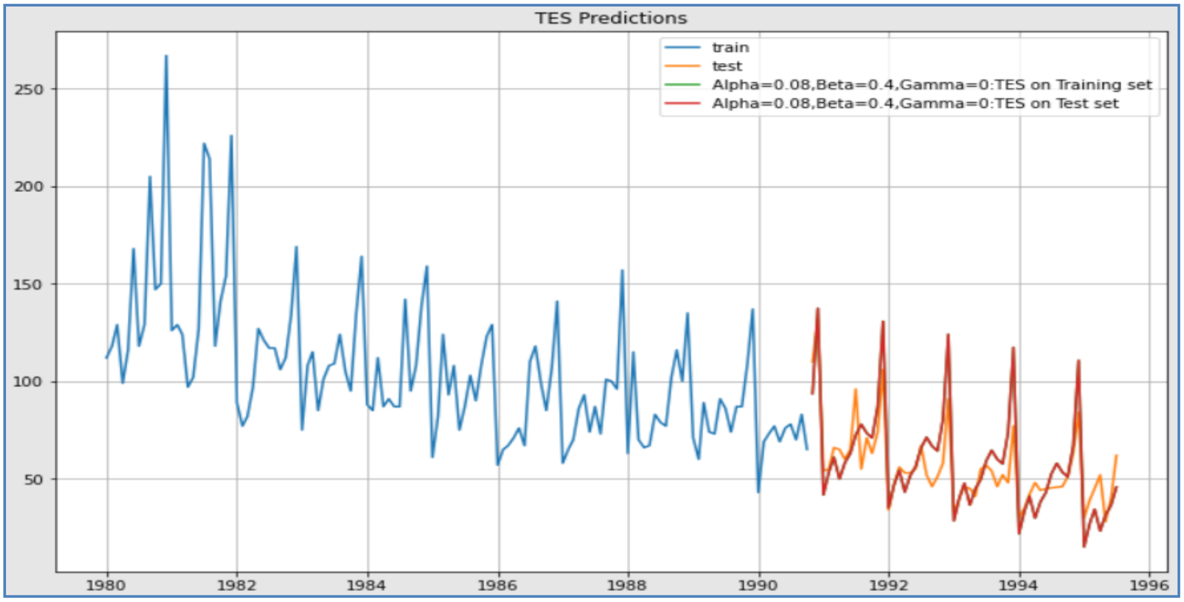

Figure 38:TES Additive-Rose

- For Alpha=0.08,Beta=0.4,Gamma=0 TES additive Model forecast on the Test Data, RMSE is 13.96

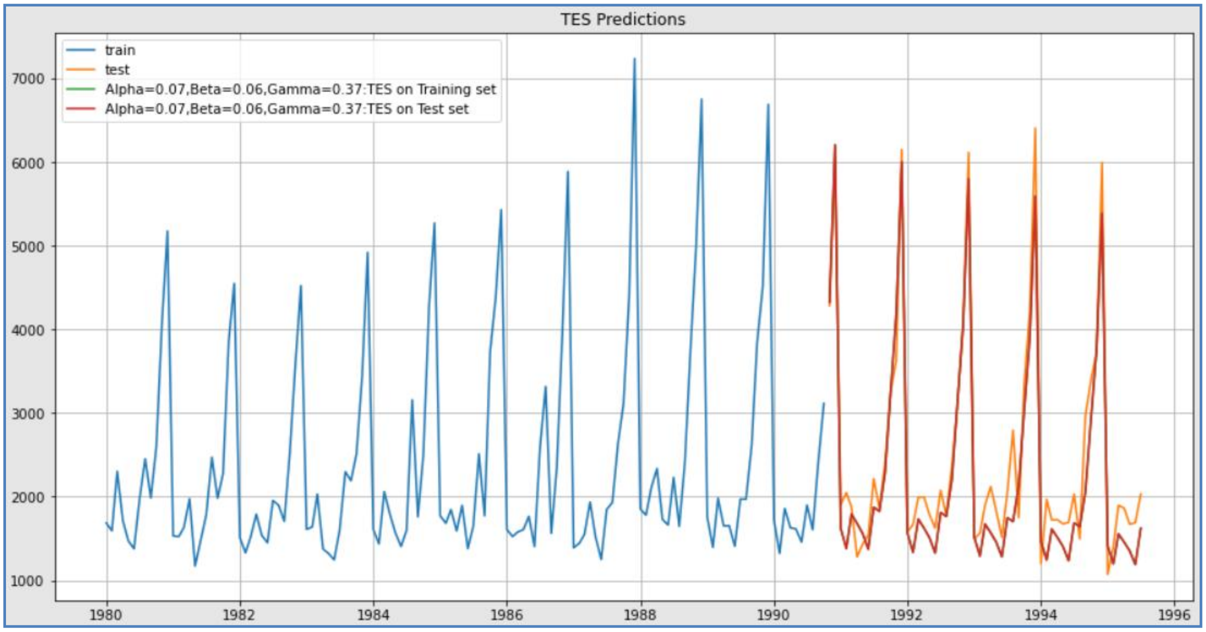

Figure 39:TES Multiplicative-Sparkling

- For Alpha=0.07,Beta=0.03,Gamma=0.47 TES Multiplicative Model forecast on the Test Data, RMSE is 381.655

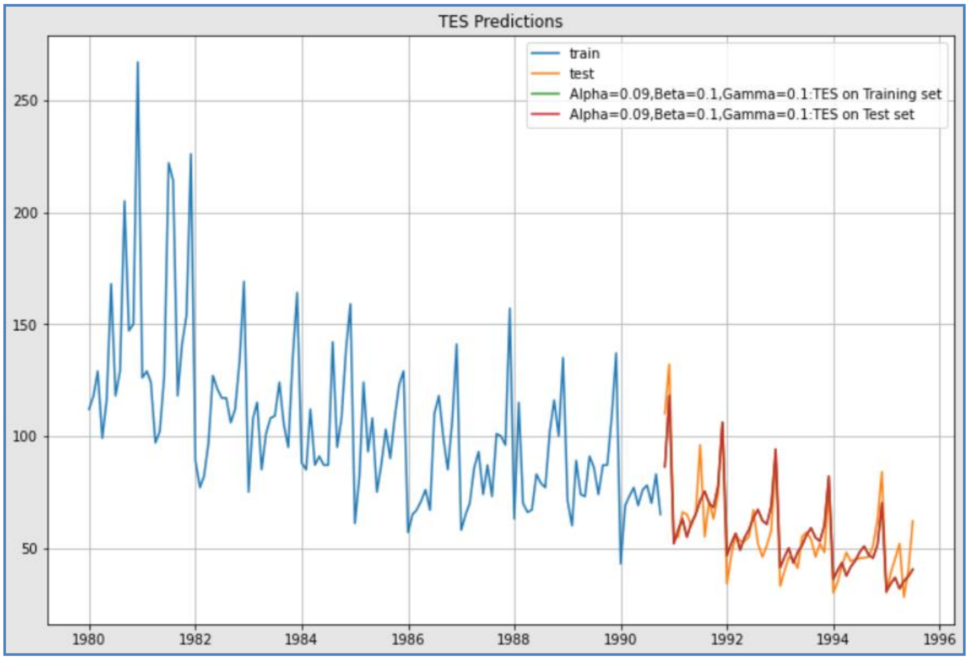

Figure 40:TES Multiplicative-Rose

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:

Null and Alternate hypothesis for ADF(Sparkling and Rose dataset)

Null hypothesis(H_0):The series is not stationary

Alternate Hypothesis(Ha):The series is stationary.

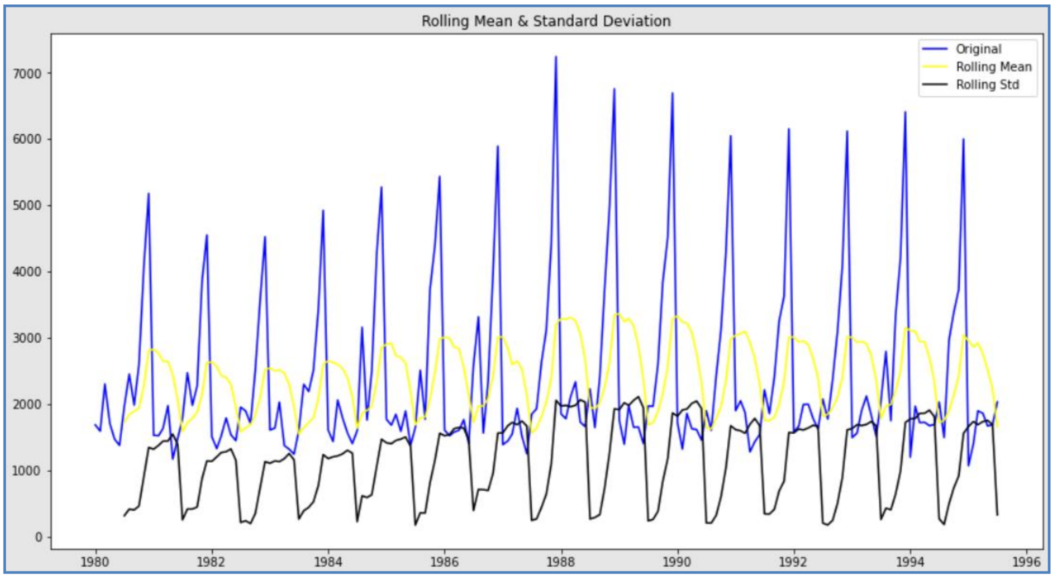

Figure 41:Sparkling-Rolling mean & Std.Dev

Results of Dickey-Fuller Test: Sparkling Dataset	
Test statistic	-1.360
P value	0.60
Lags used	11.00
No.of.Observations	175.00
Critical value(1%)	-3.468
Critical value(5%)	-2.878
Critical value(10%)	-2.575

Table 5:ADF-Sparkling

- The series is not stationary with original form at alpha =0.05.

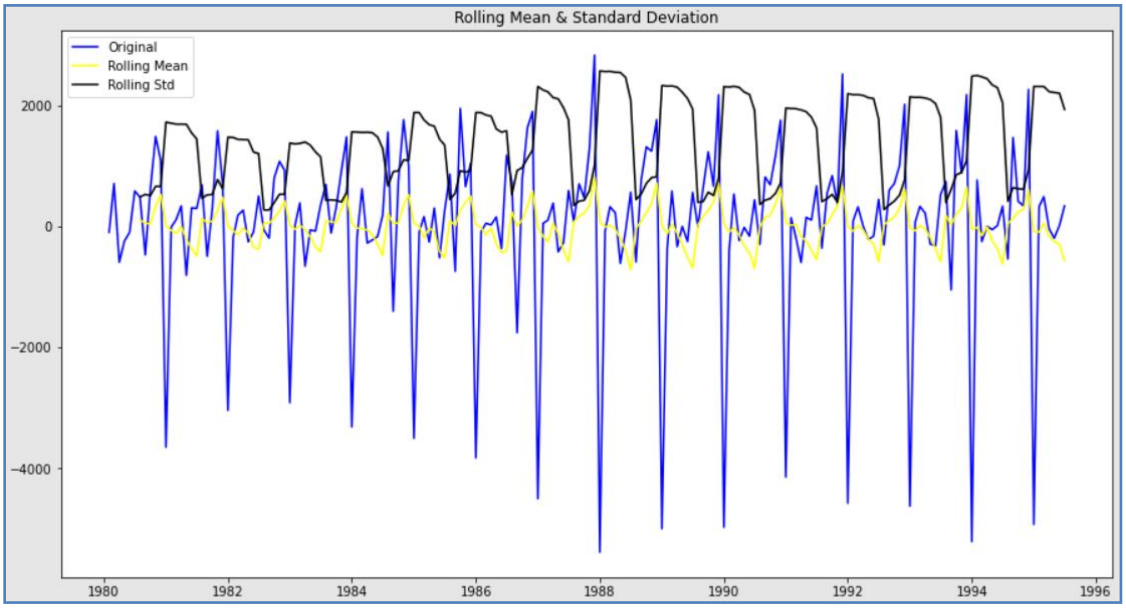

Figure 42: After differencing

Results of Dickey-Fuller Test: Sparkling Dataset(Stationary)	
Test statistic	-45.05
P value	0.00
Lags used	10.00
No.of.Observations	175.00
Critical value(1%)	-3.468
Critical value(5%)	-2.878
Critical value(10%)	-2.575

Table 6:ADF-Stationary

- The series is stationary with original form at alpha =0.05.

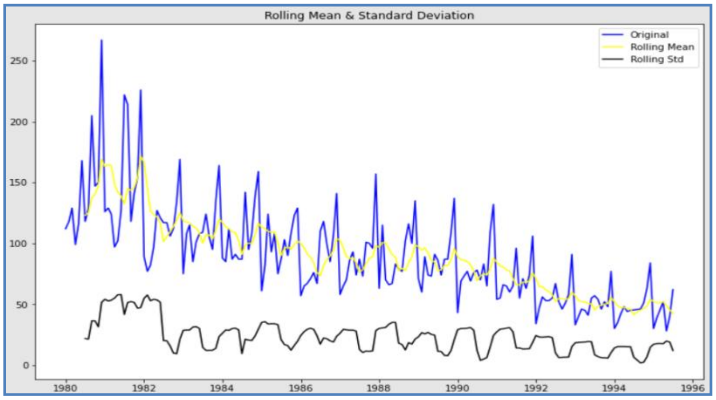

Figure 43:Rose-Rolling mean

Results of Dickey-Fuller Test: Rose Dataset	
Test statistic	-1.879
P value	0.34
Lags used	13.00
No.of.Observations	173.00
Critical value(1%)	-3.468
Critical value(5%)	-2.878
Critical value(10%)	-2.575

Table 7:ADF-Rose

- The series is not stationary with original form at alpha =0.05.

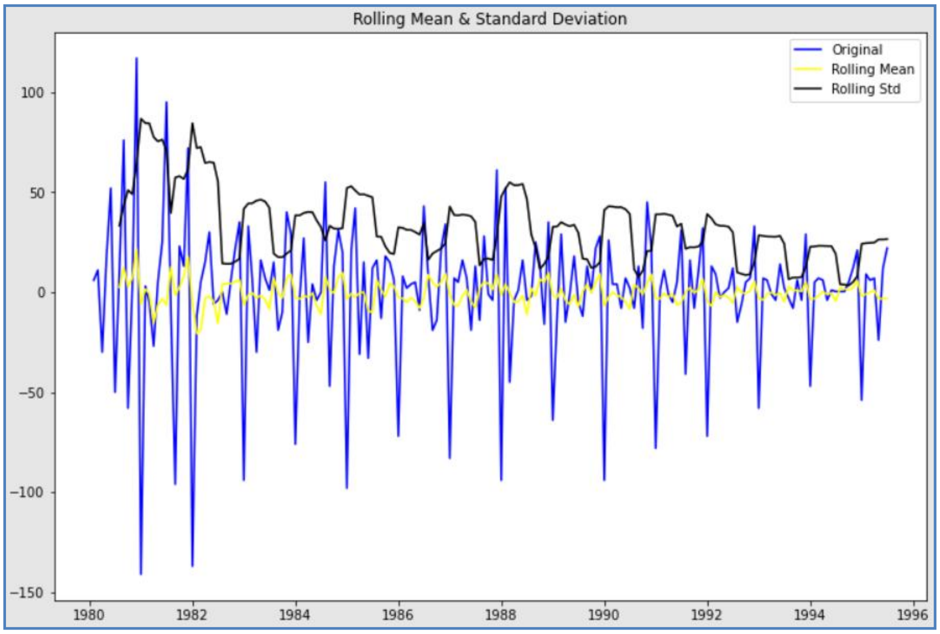

Figure 44:After Differencing

Results of Dickey-Fuller Test: Rose Dataset	
Test statistic	-8.044392e+00
P value	1.810895e-12
Lags used	1.200000e+01
No.of.Observations	1.730000e+02
Critical value(1%)	-3.468726e+00
Critical value(5%)	-2.878396e+00
Critical value(10%)	-2.575756e+00

Table 8:ADF-Stationary

- The series is stationary with original form at alpha =0.05.

6. Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.

Solution:

ARIMA(Sparkling)-Auto Regressive Integrated Moving Average.

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:           D.Sparkling   No. Observations:                  129
Model:                ARIMA(2, 1, 2)   Log Likelihood               -1081.784
Method:                      css-mle   S.D. of innovations           1008.048
Date:               Sat, 12 Nov 2022   AIC                           2175.569
Time:                       20:20:22   BIC                           2192.728
Sample:                   02-01-1980   HQIC                          2182.541
                        - 10-01-1990
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 5.3750      0.602      8.927      0.000       4.195       6.555
ar.L1.D.Sparkling     1.2595      0.075     16.837      0.000       1.113       1.406
ar.L2.D.Sparkling    -0.5364      0.074     -7.210      0.000      -0.682      -0.391
ma.L1.D.Sparkling    -1.9960      0.046    -43.120      0.000      -2.087      -1.905
ma.L2.D.Sparkling     0.9960      0.046     21.453      0.000       0.905       1.087
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.1740           -0.6971j            1.3654           -0.0853
AR.2            1.1740           +0.6971j            1.3654            0.0853
MA.1            1.0001           +0.0000j            1.0001            0.0000
MA.2            1.0039           +0.0000j            1.0039            0.0000
-----------------------------------------------------------------------------
```

Figure 45:ARIMA-Sparkling

- ARIMA-Means regression of a variable on itself.
- Lowest AIC=2175.56 with parameter (p,d,q)(2,1,2)
- Test RMSE: 1365.48

SARIMAX(Sparkling)-Seasonal ARIMA with Exogenous factor.

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                  130
Model:             SARIMAX(1, 1, 2)x(2, 0, 2, 12)   Log Likelihood                -752.865
Date:                             Sat, 12 Nov 2022   AIC                           1521.730
Time:                                     20:21:35   BIC                           1542.730
Sample:                                          0   HQIC                          1530.234
                                           - 130
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.6467      0.268     -2.413      0.016      -1.172      -0.121
ma.L1          0.1770      0.344      0.515      0.607      -0.497       0.851
ma.L2         -1.2994      0.314     -4.139      0.000      -1.915      -0.684
ar.S.L12       0.7526      0.509      1.477      0.140      -0.246       1.751
ar.S.L24       0.3257      0.542      0.601      0.548      -0.737       1.389
ma.S.L12      -0.9786      0.491     -1.994      0.046      -1.941      -0.016
ma.S.L24      -0.5636      0.671     -0.840      0.401      -1.879       0.752
sigma2      4.452e+04   2.07e+04      2.147      0.032    3883.248    8.52e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.16   Jarque-Bera (JB):                 8.08
Prob(Q):                              0.69   Prob(JB):                         0.02
Heteroskedasticity (H):               1.46   Skew:                             0.21
Prob(H) (two-sided):                  0.27   Kurtosis:                         4.31
===================================================================================
```

Figure 46: SARIMAX-Sparkling

- SARIMAX-considers both trend and seasonality.
- AIC Value:1521.73
- SARIMAX (1,1,2) X (2,0,2,12)
- Test RMSE:712.39

The better model is built by the one with lower AIC Value.Hence SARIMAX is the best model to analyse the Sparkling wine sale.

```
                              ARIMA Model Results
==============================================================================
Dep. Variable:                 D.Rose   No. Observations:                  129
Model:                 ARIMA(0, 1, 2)   Log Likelihood                -623.393
Method:                       css-mle   S.D. of innovations             29.847
Date:                Sun, 13 Nov 2022   AIC                           1254.787
Time:                        11:53:08   BIC                           1266.226
Sample:                    02-01-1980   HQIC                          1259.435
                         - 10-01-1990
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const           -0.5281      0.085     -6.222      0.000      -0.694      -0.362
ma.L1.D.Rose    -0.7780      0.100     -7.805      0.000      -0.973      -0.583
ma.L2.D.Rose    -0.2219      0.096     -2.318      0.020      -0.410      -0.034
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
MA.1            1.0001           +0.0000j            1.0001            0.0000
MA.2           -4.5060           +0.0000j            4.5060            0.5000
-----------------------------------------------------------------------------
```

Figure 47:ARIMA-Rose

- ARIMA-Means regression of a variable on itself.
- Lowest AIC=1254.78 with parameter (p,d,q)(0,1,2)
- Test RMSE: 17.42

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  130
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 12)   Log Likelihood                -428.538
Date:                            Sun, 13 Nov 2022   AIC                            871.075
Time:                                    11:53:46   BIC                            889.450
Sample:                                         0   HQIC                           878.516
                                            - 130
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.8367    239.175     -0.003      0.997    -469.611     467.937
ma.L2         -0.1633     39.038     -0.004      0.997     -76.676      76.349
ar.S.L12       0.3494      0.079      4.408      0.000       0.194       0.505
ar.S.L24       0.3067      0.075      4.103      0.000       0.160       0.453
ma.S.L12       0.0454      0.134      0.338      0.735      -0.218       0.309
ma.S.L24      -0.0912      0.145     -0.628      0.530      -0.376       0.193
sigma2       250.7786      6e+04      0.004      0.997   -1.17e+05    1.18e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.09   Jarque-Bera (JB):                 3.10
Prob(Q):                              0.76   Prob(JB):                         0.21
Heteroskedasticity (H):               0.88   Skew:                             0.43
Prob(H) (two-sided):                  0.71   Kurtosis:                         3.05
===================================================================================
```

Figure 48:SARIMAX-Rose

- SARIMAX-considers both trend and seasonality.
- AIC Value:871.07
- SARIMAX (0,1,2) X (2,0,2,12)
- Test RMSE:25.34

The better model is built by the one with lower AIC Value.Hence ARIMA is the best model to analyse the Rose wine sale.

7. Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.

Solution:

ACF/PACF-Auto/Partial correlation Factor(Sparkling):

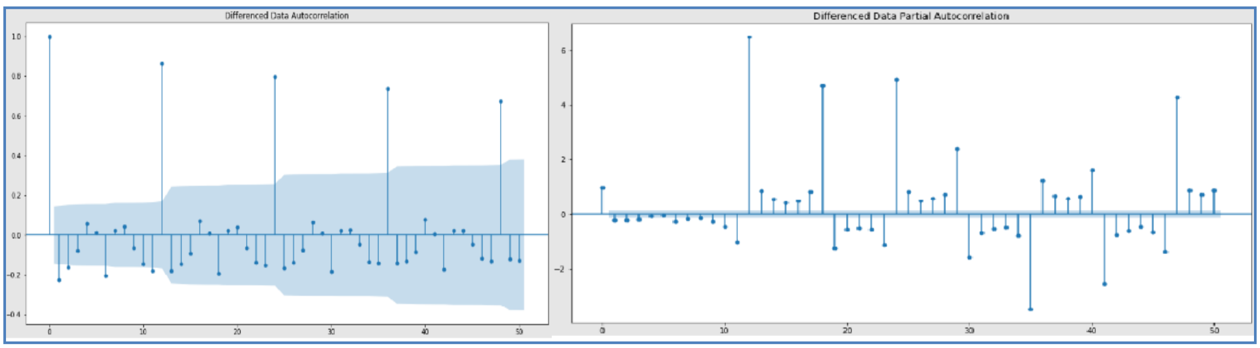

Figure 49:Manual ARIMA

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:             D.Sparkling   No. Observations:                  129
Model:                  ARIMA(0, 1, 0)   Log Likelihood               -1115.354
Method:                            css   S.D. of innovations           1376.382
Date:                 Sat, 12 Nov 2022   AIC                           2234.707
Time:                         20:21:36   BIC                           2240.427
Sample:                     02-01-1980   HQIC                          2237.031
                          - 10-01-1990
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         11.0853    121.184      0.091      0.927    -226.430     248.601
==============================================================================
```

- Test RMSE(Manual_ARIMA-Sparkling):1679.32

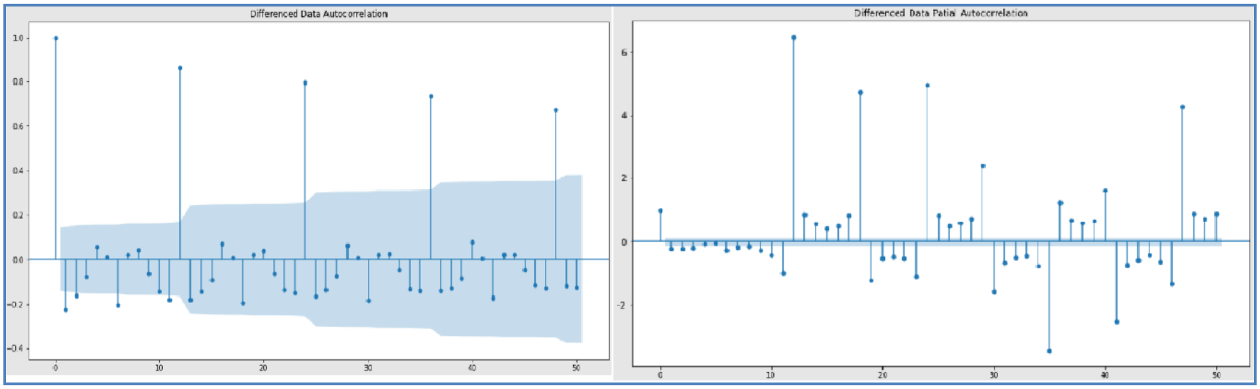

Figure 50:Manual-SARIMA-Sparkling

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  130
Model:           SARIMAX(0, 1, 0)x(1, 1, [1, 2, 3], 6)   Log Likelihood            -797.057
Date:                           Sat, 12 Nov 2022   AIC                           1604.113
Time:                                   20:21:38   BIC                           1617.335
Sample:                                        0   HQIC                          1609.470
                                           - 130
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.S.L6       -1.0181      0.015    -66.917      0.000      -1.048      -0.988
ma.S.L6        0.0330      0.178      0.186      0.852      -0.315       0.381
ma.S.L12      -0.4594      0.082     -5.590      0.000      -0.621      -0.298
ma.S.L18       0.0732      0.166      0.442      0.659      -0.252       0.398
sigma2      2.642e+05   2.94e+04      9.000      0.000    2.07e+05    3.22e+05
===================================================================================
Ljung-Box (L1) (Q):                  14.96   Jarque-Bera (JB):                31.98
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               1.15   Skew:                             0.67
Prob(H) (two-sided):                  0.69   Kurtosis:                         5.36
===================================================================================
```

- Test RMSE(Manual SARIMAX –Sparkling):1359.341

ACF/PACF-Auto/Partial correlation Factor(Rose):

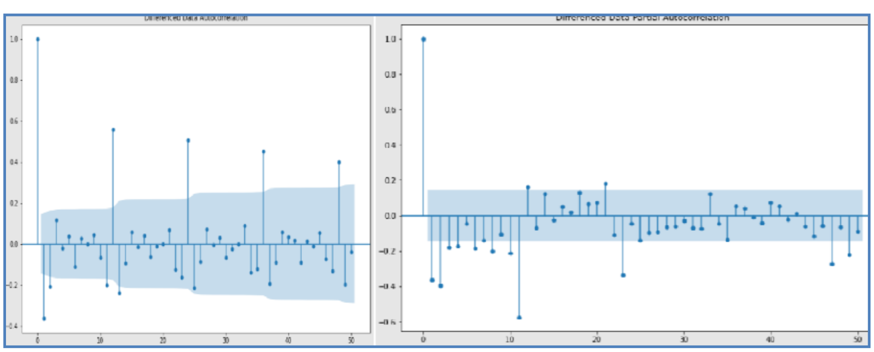

Figure 51:Manual ARIMA

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:                 D.Rose   No. Observations:                  129
Model:                 ARIMA(0, 1, 0)   Log Likelihood                -655.582
Method:                           css   S.D. of innovations             38.982
Date:                Sun, 13 Nov 2022   AIC                           1315.165
Time:                        11:53:47   BIC                           1320.884
Sample:                    02-01-1980   HQIC                          1317.489
                         - 10-01-1990
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.3643      3.432     -0.106      0.915      -7.091       6.363
==============================================================================
```

- Test RMSE(Manual_ARIMA-Rose):17.80

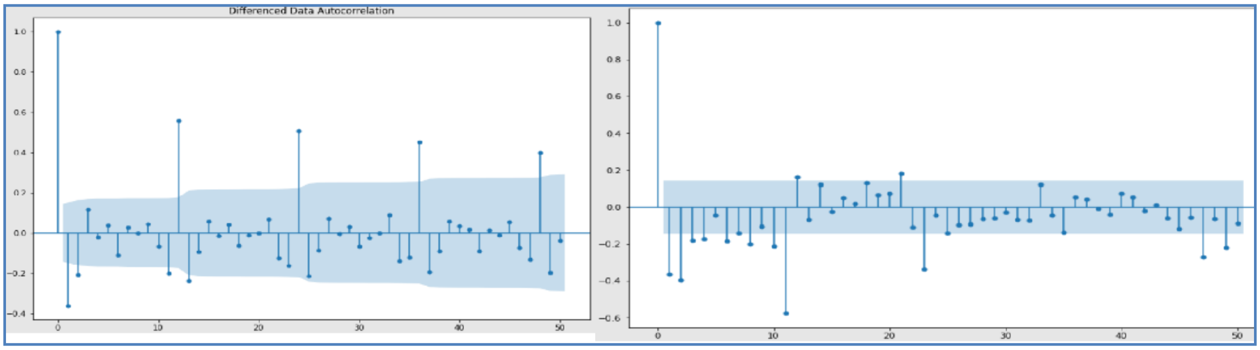

Figure 52:Manual-SARIMAX

```
                                       SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                 130
Model:             SARIMAX(0, 1, 0)x(1, 1, [1, 2, 3], 6)   Log Likelihood            -469.371
Date:                             Sun, 13 Nov 2022   AIC                           948.741
Time:                                     11:53:49   BIC                           961.963
Sample:                                          0   HQIC                          954.098
                                             - 130
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.S.L6       -0.8592      0.036    -23.819      0.000      -0.930      -0.789
ma.S.L6       -0.1913      0.112     -1.715      0.086      -0.410       0.027
ma.S.L12      -0.4801      0.114     -4.198      0.000      -0.704      -0.256
ma.S.L18      -0.0685      0.100     -0.685      0.493      -0.265       0.128
sigma2       472.7674     67.750      6.978      0.000     339.979     605.556
===================================================================================
Ljung-Box (L1) (Q):                   8.37   Jarque-Bera (JB):                 0.00
Prob(Q):                              0.00   Prob(JB):                         1.00
Heteroskedasticity (H):               0.80   Skew:                             0.01
Prob(H) (two-sided):                  0.52   Kurtosis:                         3.03
===================================================================================
```

- Test RMSE(Manual_SARIMAX-Rose):15.80

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

Solution:

Sparkling Dataset:

	Test RMSE
RegressionOnTime	1356.301492
NaiveModel	1439.341693
Simple Average	1362.075999
2pointTrailingMovingAverage	811.178937
4pointTrailingMovingAverage	1184.213295
6pointTrailingMovingAverage	1337.200524
9pointTrailingMovingAverage	1422.653281
Alpha=0.06,SimpleExponentialSmoothing	1363.702251
Alpha=0.07,Beta=0.07,DoubleExponentialSmoothing	1472.253632
Alpha=0.07,Beta=0.03,Gamma=0.47,TES Additive	366.859156
Alpha=0.07,Beta=0.03,Gamma=0.47,TES Multiplicative	381.655272
ARIMA_AIC[2,1,2]	1365.489906
SARIMAX(1, 1, 2)x(2, 0, 2, 12)	712.394763
ARIMA_ACF/PACF	1679.321024
SARIMA_ACF/PACF	1359.341878

Figure 53:Test RMSE:Sparkling

Rose Dataset:

	Test RMSE
RegressionOnTime	17.286999
NaiveModel	20.737945
Simple Average	19.913422
2pointTrailingMovingAverage	11.801043
4pointTrailingMovingAverage	15.367212
6pointTrailingMovingAverage	15.862350
9pointTrailingMovingAverage	16.341919
Alpha=0.10,SimpleExponentialSmoothing	30.188326
Alpha=0.1,Beta=0.7,DoubleExponentialSmoothing	17.355728
Alpha=0.08,Beta=0.04,Gamma=0,TES Additive	13.963361
Alpha=0.09,Beta=0.1,Gamma=0.1,TES Multiplicative	9.325439
ARIMA_AIC[0,1,2]	17.428095
SARIMAX(0, 1, 2)x(2, 0, 2, 12)	25.343325
ARIMA_ACF/PACF	17.808181
SARIMA_ACF/PACF	15.808075

Figure 54:Test RMSE-Rose

9. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

Solution:

SARIMAX Diagnostic plot(Sparkling & Rose):

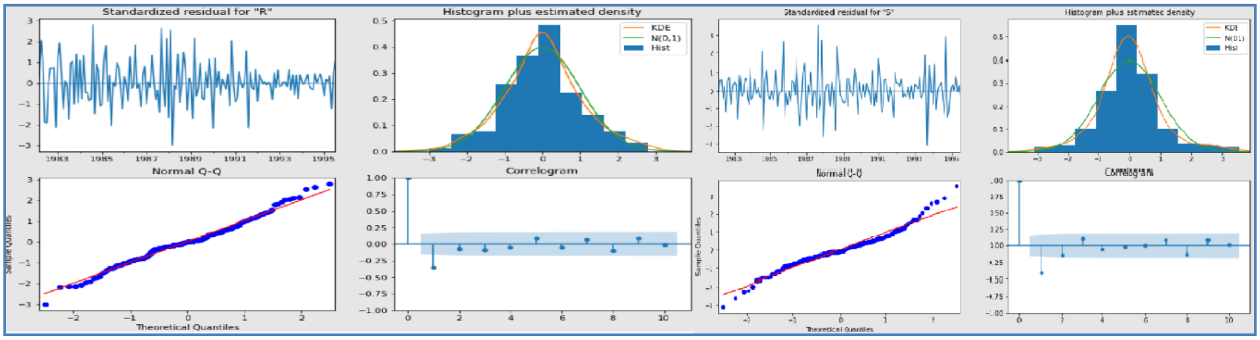

Figure 55: SARIMAX plot

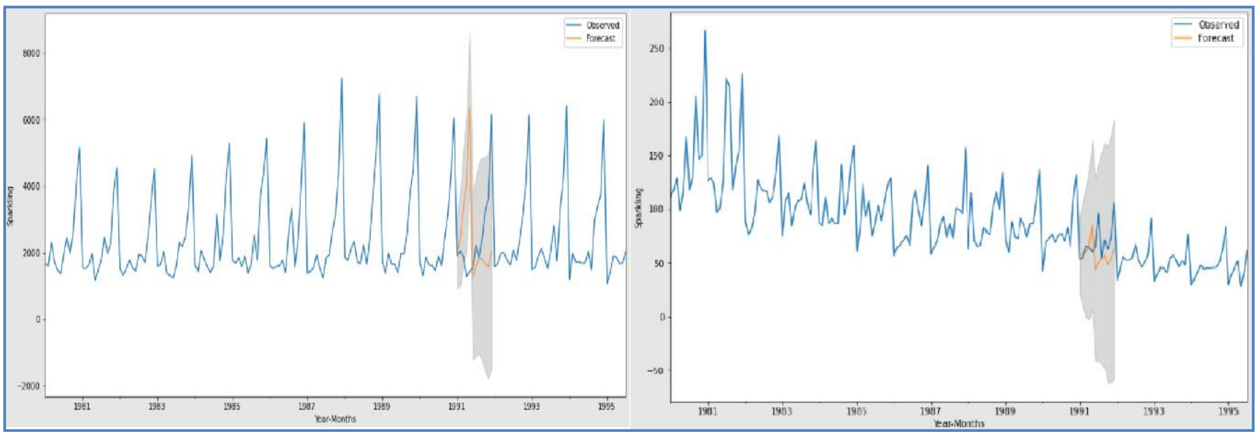

Figure 56: Forecast plot

- Forecasted data is incorporated and shown in above plot.
- RMSE of full data(12 months):608,29.

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

Solution:

1.Sparkling:

- KDE plot is similar with the normal distribution.
- QQ Plot shows distribution of residuals(normal distribution of residuals)
- ACF/PACF plot residuals are random.
- Overall-ARIMA is best fit model for analyzing Sparkling wine dataset.

2.Rose:

- Triple exponential smoothing model is best with optimum alpha,beta and gamma values.
- Lowest AIC-ARIMA Model is also best.

Steps used:

- Time series data has been split into train and test set.
- Univariate/Bivariate analysis are carried out.
- Decomposition of dataset is done to identify trend,level and seasonality.
- Model built using Linear regression,Naïve,Simple average,Moving average is carried out.
- Exponential smoothing on Simple,Double and Triple(Additive and Multiplicative models)
- ACF/PACF Plots ARIMA,SARIMA Plots are built.
- Lowest AIC score is identified.
- Best model is built and predicted 12 months into future.
- Corresponding RMSE Values are stored in results.

******Thank You******