Project - Time Series Forecasting

Balasubramaniyam Ravichandran 12/11/22

INDEX

Problem: Sparkling and Rose

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.

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

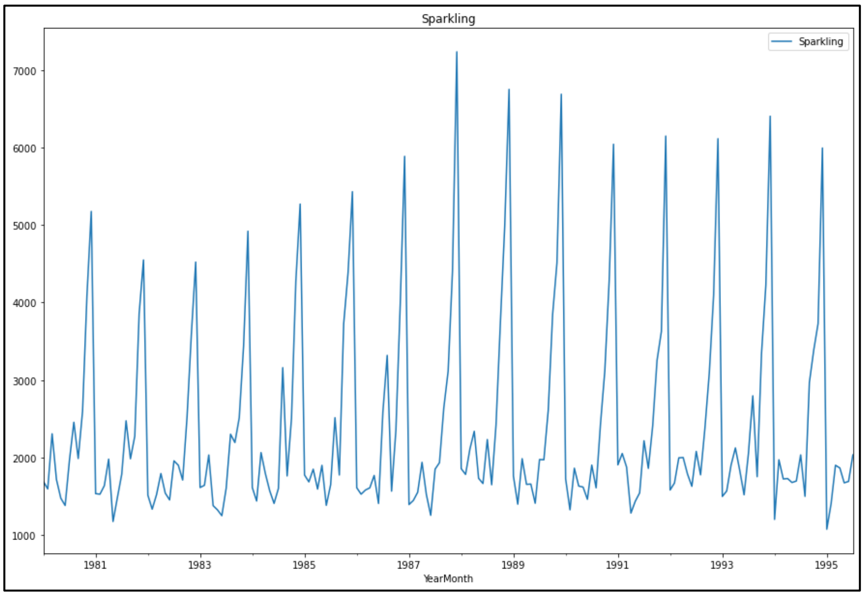
Sparkling:

```
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   YearMonth  187 non-null    datetime64[ns]
 1   Sparkling  187 non-null    int64
dtypes: datetime64[ns](1), int64(1)
```

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

1.1. Sparkling data info, Dataset

The dataset is read as a time series data. The dataset contains two columns Yearmonth with datetime datatype and the Sparkling column with integer datatype. The dataset is having 187 rows in it.



1.2. Sparkling data plot

The Sparkling dataset is plotted, and it is observed that there is some form of seasonality present in it.

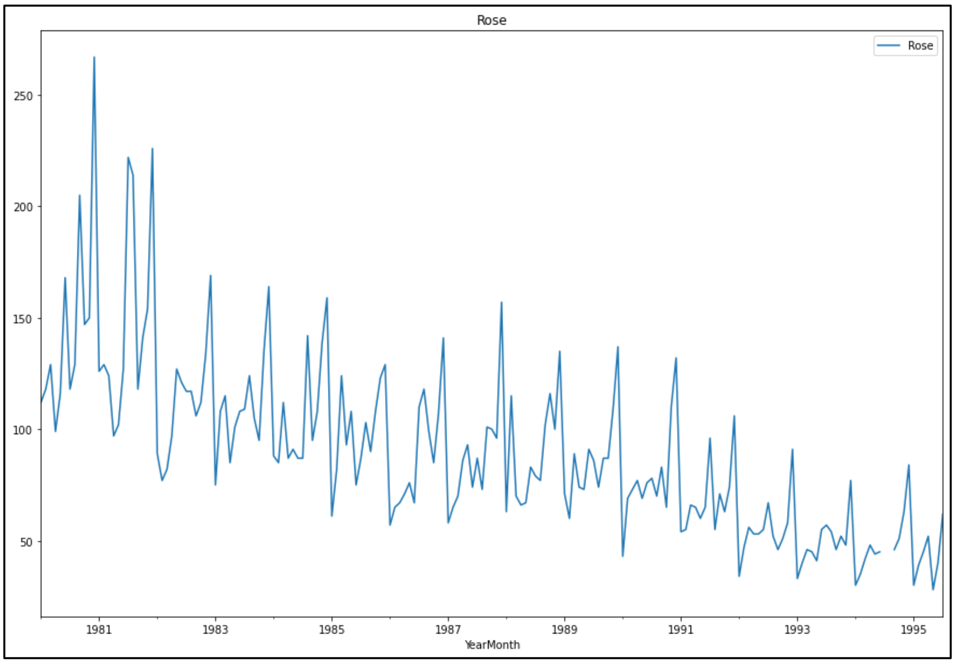
Rose:

```
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   YearMonth  187 non-null    datetime64[ns]
 1   Rose       185 non-null    float64
dtypes: datetime64[ns](1), float64(1)
```

	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

1.3. Rose Data info, Dataset

The dataset is read as a time series data. The dataset contains two columns Yearmonth with datetime datatype and the Rose column with float datatype. The dataset is having 187 rows in it.



1.4. Rose data plot

The Rose dataset is plotted, and it is observed that there is some form of seasonality present in it. The sales trend for Rose wine is negative.

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

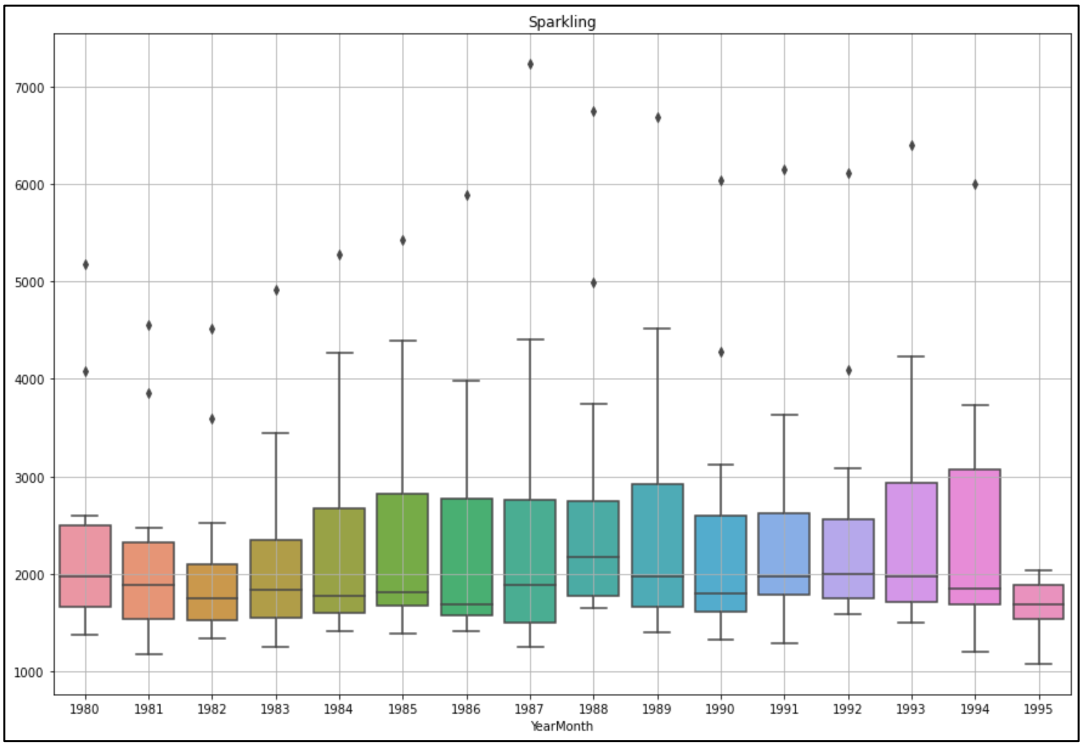
Sparkling wine:

	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

```
Sparkling    0
dtype: int64
```

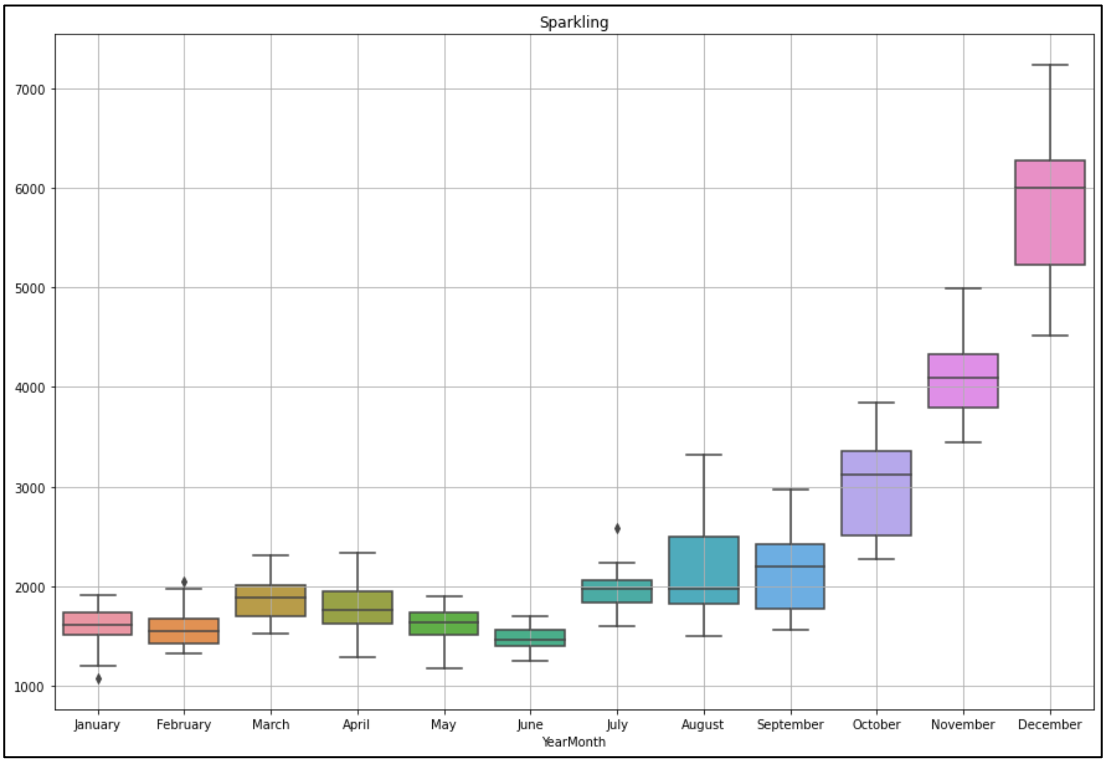
2.1. Sparkling data Summary, Null value check

From the data summary, it is observed that over the period of 1980 to 1995 the maximum sparkling units sold is 7242 and the minimum units sold is 1070. The average sale is about 2402 units with a standard deviation of 1295 units. The dataset is checked for null values and no null values are present in it.



2.2. Sparkling Yearly Boxplot

From the yearly boxplot, the yearly sales are good from year 1984 to 1989. Outliers are present in the dataset in all the sale years except the year 1995. These outliers are not treated, and further analysis is carried out with outliers present in the dataset.



2.3. Sparkling Monthly Boxplot

From the monthly boxplot, the monthly sales are good during the December, November, October months. The sales are low during the period from January to August. The indicates the sales is good during the holiday seasons and new year.

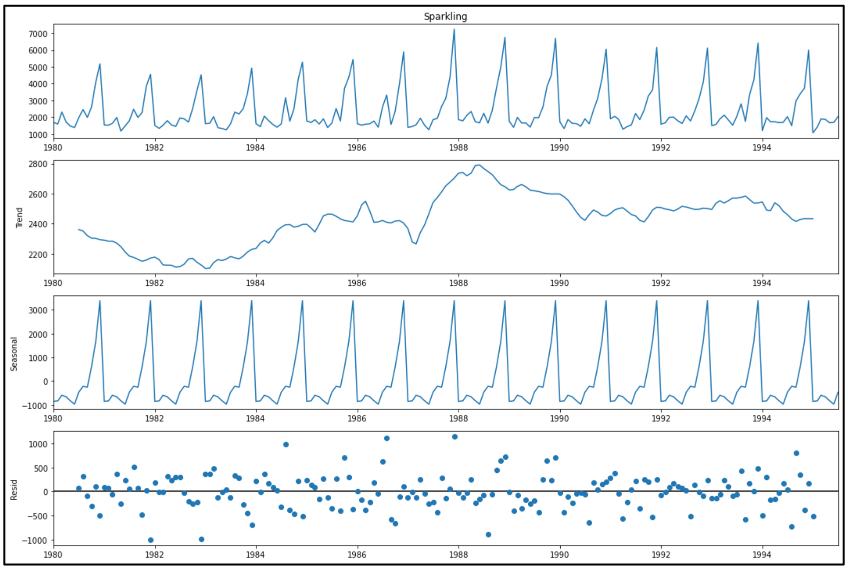
YearMonth	April	August	December	February	January	July	June	March	May	November	October	September
YearMonth												
1980	1712.0	2453.0	5179.0	1591.0	1686.0	1966.0	1377.0	2304.0	1471.0	4087.0	2596.0	1984.0
1981	1976.0	2472.0	4551.0	1523.0	1530.0	1781.0	1480.0	1633.0	1170.0	3857.0	2273.0	1981.0
1982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

2.4. Sparkling monthly sales across years

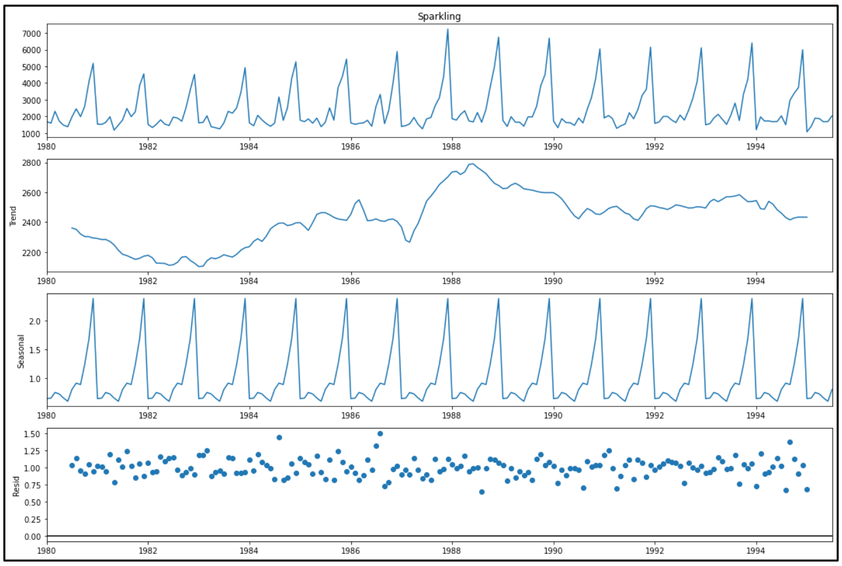
Decomposition of Sparkling Data

The decomposition of the dataset helps us to better understand the trend, seasonality, and residual present in it. Let's perform both the additive and multiplicative decomposition of our dataset.

Additive Decomposition



Multiplicative Decomposition



2.5. Sparkling Data Decomposition

From the decomposition it is observed a clear seasonality is present in the dataset with repetitive patterns. Trend however is positive till 1988 and becomes negative after that year. From the two methods multiplicative decomposition looks better after observing the residual pattern.

Rose wine:

	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

```
Rose      2
dtype: int64
```

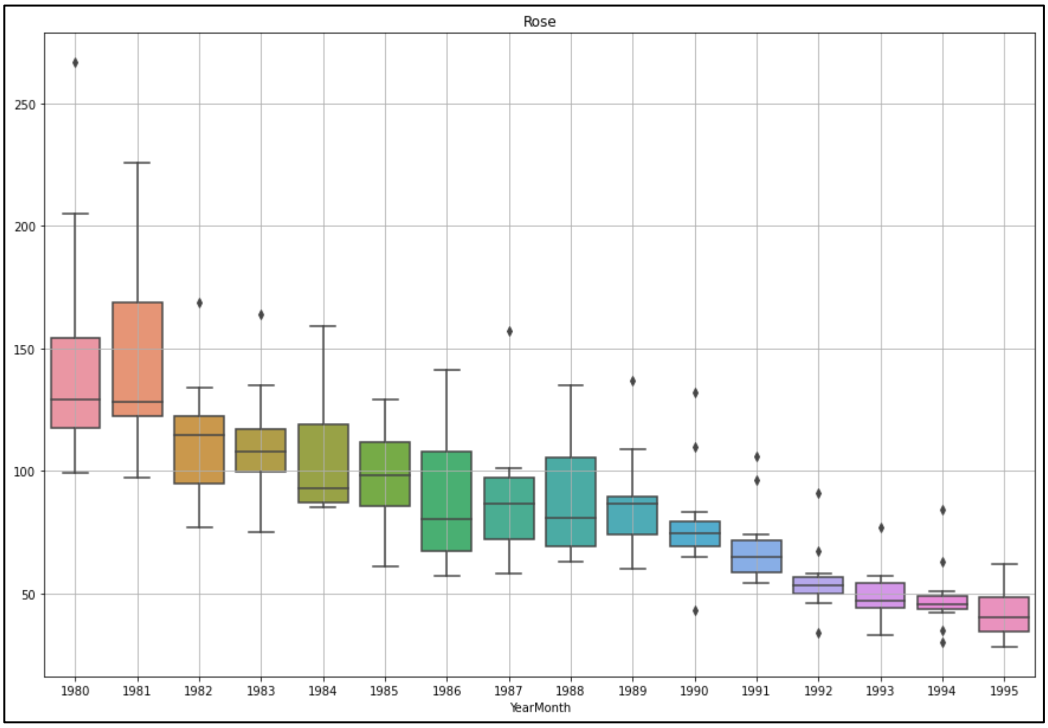
	Rose
YearMonth	
1994-01-01	30.000000
1994-02-01	35.000000
1994-03-01	42.000000
1994-04-01	48.000000
1994-05-01	44.000000
1994-06-01	45.000000
1994-07-01	45.333333
1994-08-01	45.666667
1994-09-01	46.000000
1994-10-01	51.000000
1994-11-01	63.000000
1994-12-01	84.000000

2.6. Sparkling data Summary, Null value check

From the data summary, it is observed that over the period of 1980 to 1995 the maximum Rose units sold is 267 and the minimum units sold is 28.

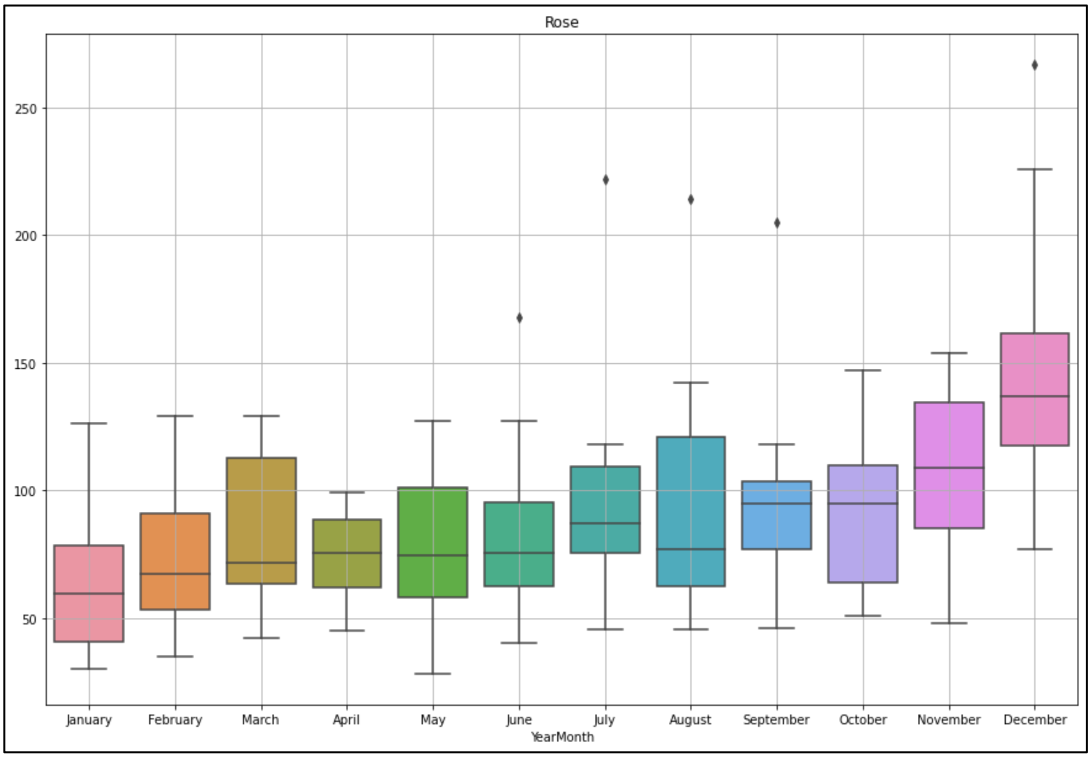
The average sale is about 90 units with a standard deviation of 39 units. The dataset is checked for null values and 2 null values are present in the year 1994.

The null values are filled with the value of 45.6 after using the interpolation method.



2.7. Rose Yearly Boxplot

From the yearly boxplot, the yearly sales of Rose wine are good from year 1980 to 1984. Outliers are present in the dataset. These outliers are not treated, and further analysis is carried out with outliers present in the dataset. We get an idea that the yearly sales are on a negative trend from the above plot.



2.8. Rose Monthly Boxplot

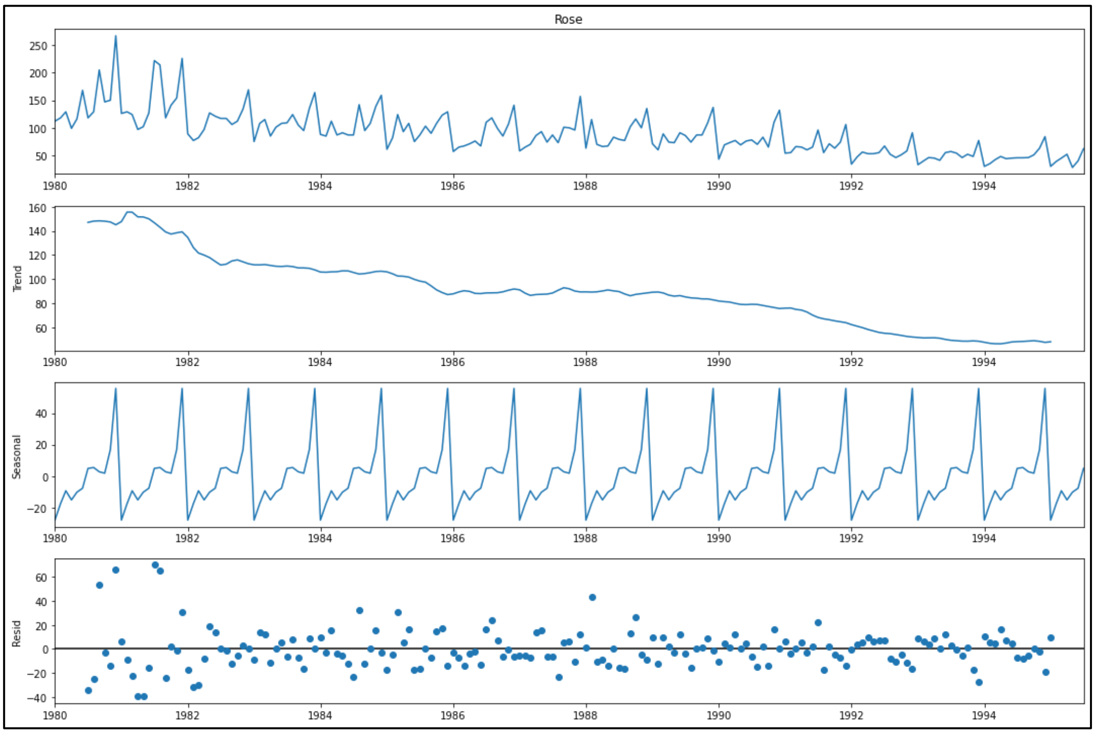
From the monthly boxplot, the monthly sales of rose are good during the December, November months. The sales are comparatively low during the period from January to September. The indicates the sales is good during the holiday seasons and new year time.

YearMonth	April	August	December	February	January	July	June	March	May	November	October	September
YearMonth												
1980	99.0	129.0	267.0	118.0	112.0	118.0	168.0	129.0	116.0	150.0	147.0	205.0
1981	97.0	214.0	226.0	129.0	126.0	222.0	127.0	124.0	102.0	154.0	141.0	118.0
1982	97.0	117.0	169.0	77.0	89.0	117.0	121.0	82.0	127.0	134.0	112.0	106.0
1983	85.0	124.0	164.0	108.0	75.0	109.0	108.0	115.0	101.0	135.0	95.0	105.0
1984	87.0	142.0	159.0	85.0	88.0	87.0	87.0	112.0	91.0	139.0	108.0	95.0
1985	93.0	103.0	129.0	82.0	61.0	87.0	75.0	124.0	108.0	123.0	108.0	90.0
1986	71.0	118.0	141.0	65.0	57.0	110.0	67.0	67.0	76.0	107.0	85.0	99.0
1987	86.0	73.0	157.0	65.0	58.0	87.0	74.0	70.0	93.0	96.0	100.0	101.0
1988	66.0	77.0	135.0	115.0	63.0	79.0	83.0	70.0	67.0	100.0	116.0	102.0
1989	74.0	74.0	137.0	60.0	71.0	86.0	91.0	89.0	73.0	109.0	87.0	87.0
1990	77.0	70.0	132.0	69.0	43.0	78.0	76.0	73.0	69.0	110.0	65.0	83.0
1991	65.0	55.0	106.0	55.0	54.0	96.0	65.0	66.0	60.0	74.0	63.0	71.0
1992	53.0	52.0	91.0	47.0	34.0	67.0	55.0	56.0	53.0	58.0	51.0	46.0
1993	45.0	54.0	77.0	40.0	33.0	57.0	55.0	46.0	41.0	48.0	52.0	46.0
1994	48.0	45.6	84.0	35.0	30.0	45.6	45.0	42.0	44.0	63.0	51.0	46.0
1995	52.0	NaN	NaN	39.0	30.0	62.0	40.0	45.0	28.0	NaN	NaN	NaN

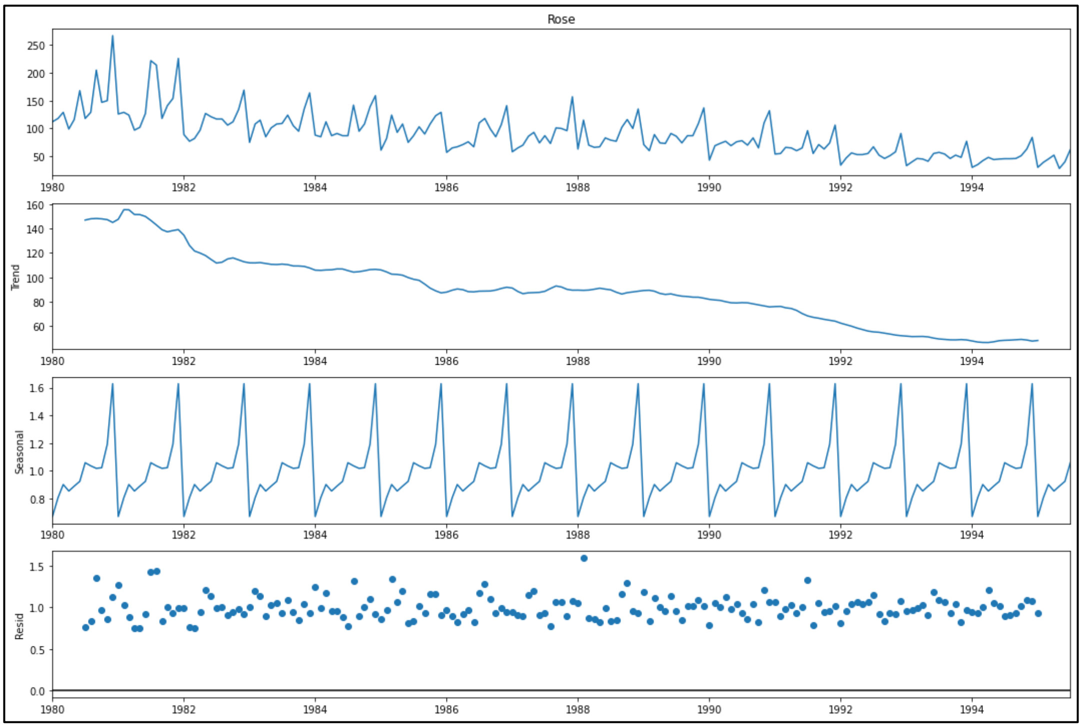
2.9. Rose monthly sales across years

Decomposition of Rose Data

Additive Decomposition



Multiplicative Decomposition



From the decomposition it is observed a clear seasonality is present in the dataset with repetitive patterns. Trend however is negative over the period of years, the sales has decreased significantly. From the two methods multiplicative decomposition looks better after observing the residual pattern.

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

Sparkling (Data split)

YearMonth	Sparkling
1980-01-01	1686
1980-02-01	1591
1980-03-01	2304
1980-04-01	1712
1980-05-01	1471
1980-06-01	1377
1980-07-01	1966
1980-08-01	2453
1980-09-01	1984
1980-10-01	2596
1980-11-01	4087
1980-12-01	5179

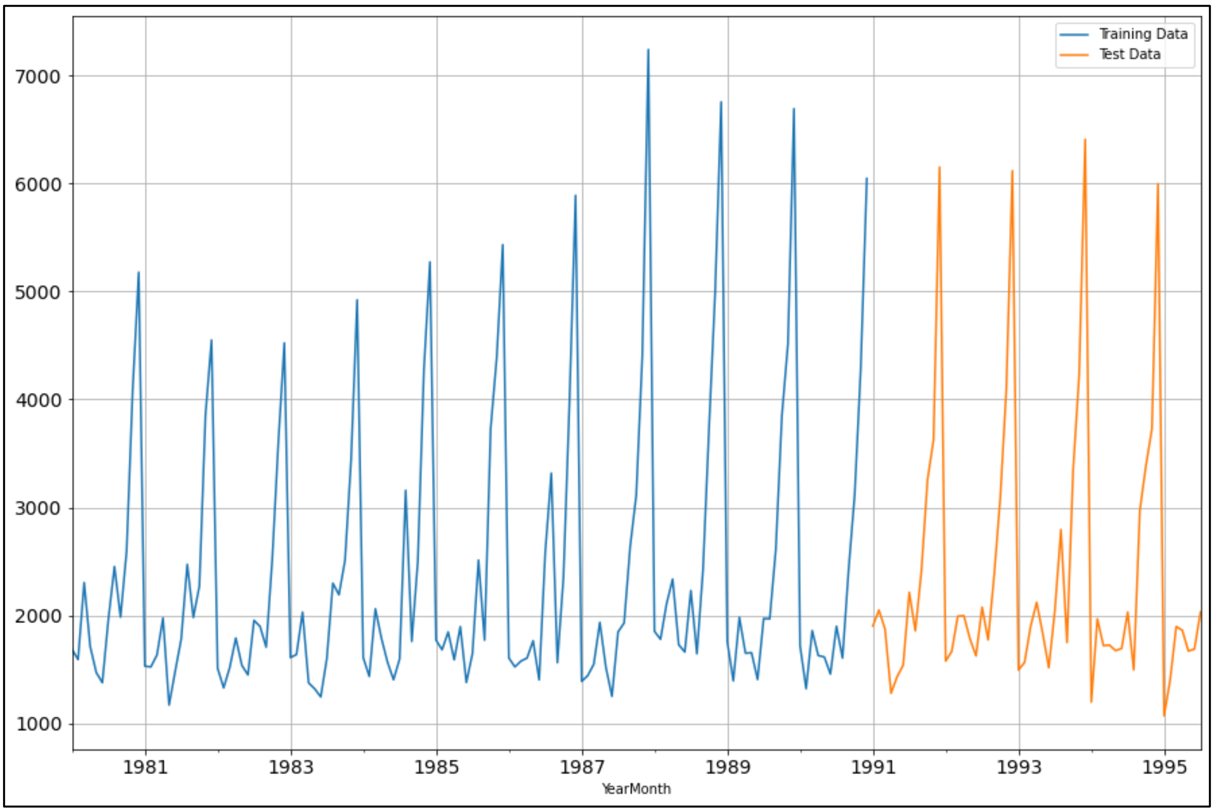
YearMonth	Sparkling
1990-01-01	1720
1990-02-01	1321
1990-03-01	1859
1990-04-01	1628
1990-05-01	1615
1990-06-01	1457
1990-07-01	1899
1990-08-01	1605
1990-09-01	2424
1990-10-01	3116
1990-11-01	4286
1990-12-01	6047

3.1. Training data

YearMonth	Sparkling
1991-01-01	1902
1991-02-01	2049
1991-03-01	1874
1991-04-01	1279
1991-05-01	1432
1991-06-01	1540
1991-07-01	2214
1991-08-01	1857
1991-09-01	2408
1991-10-01	3252
1991-11-01	3627
1991-12-01	6153

YearMonth	Sparkling
1994-08-01	1495
1994-09-01	2968
1994-10-01	3385
1994-11-01	3729
1994-12-01	5999
1995-01-01	1070
1995-02-01	1402
1995-03-01	1897
1995-04-01	1862
1995-05-01	1670
1995-06-01	1688
1995-07-01	2031

3.1. Test data



3.1.1. Sparkling data split

Rose (Data split)

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
1980-06-01	168.0
1980-07-01	118.0
1980-08-01	129.0
1980-09-01	205.0
1980-10-01	147.0
1980-11-01	150.0
1980-12-01	267.0

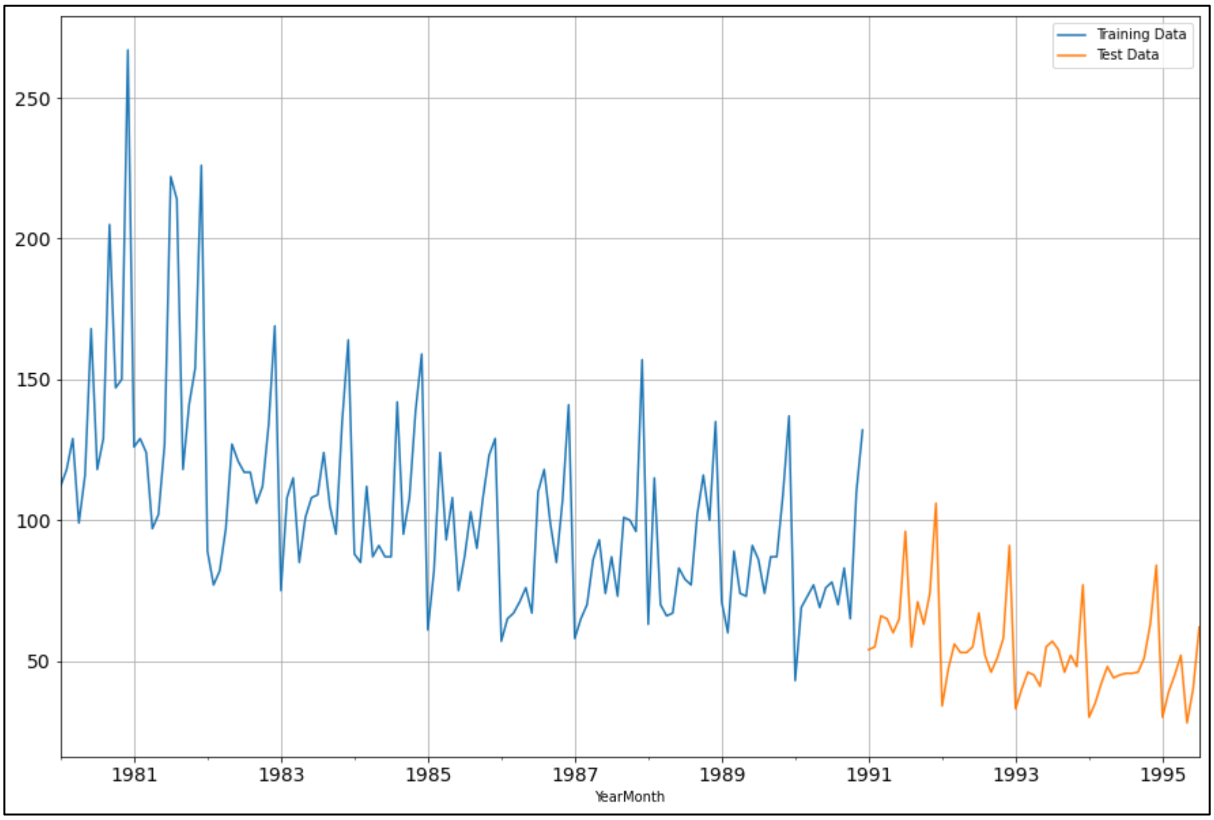
YearMonth	Rose
1990-01-01	43.0
1990-02-01	69.0
1990-03-01	73.0
1990-04-01	77.0
1990-05-01	69.0
1990-06-01	76.0
1990-07-01	78.0
1990-08-01	70.0
1990-09-01	83.0
1990-10-01	65.0
1990-11-01	110.0
1990-12-01	132.0

3.2. Training data

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
1991-06-01	65.0
1991-07-01	96.0
1991-08-01	55.0
1991-09-01	71.0
1991-10-01	63.0
1991-11-01	74.0
1991-12-01	106.0

YearMonth	Rose
1994-08-01	45.6
1994-09-01	46.0
1994-10-01	51.0
1994-11-01	63.0
1994-12-01	84.0
1995-01-01	30.0
1995-02-01	39.0
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

3.2. Test data



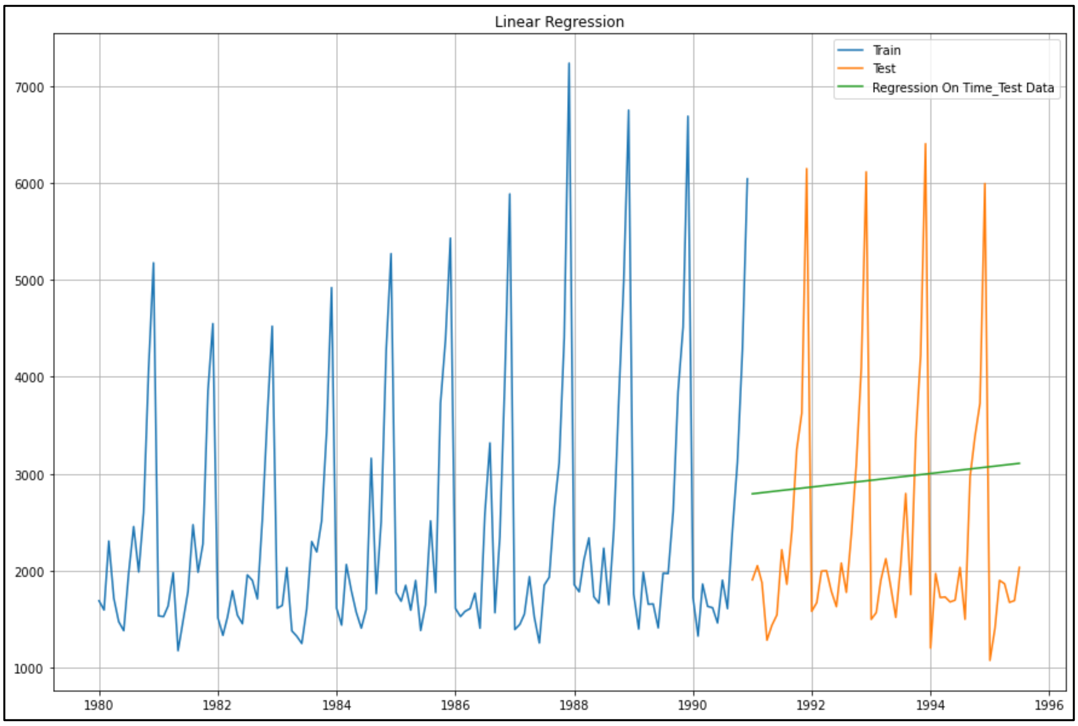
3.2.1. Rose data split

For building our forecasting models the dataset is split into train and test datasets. For both the Sparkling and Rose dataset the split is made for the training data from the year 1980 to 1990, the test data is from the year 1991 to 1995. The first and last 12 months of the training and test dataset is provided above for both the Sparkling and Rose datasets.

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.**

The train and test data of the sparkling and rose datasets are used to build the following forecasting models. Each model is built separately for both datasets,

Linear Regression



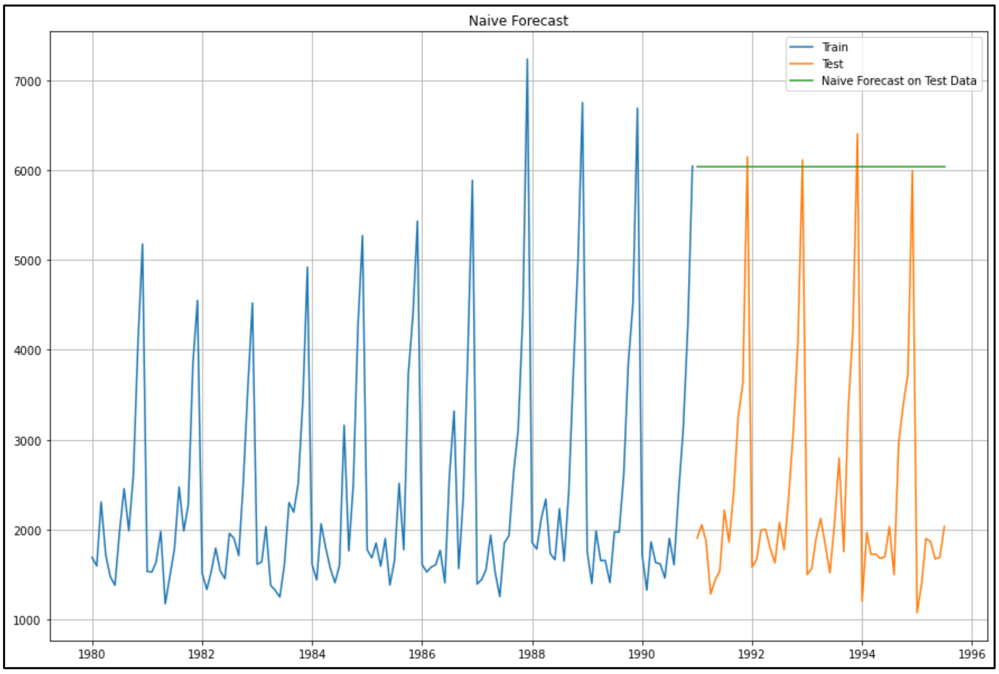
4.1. Linear Regression (Sparkling)

The forecast is a flat line, let's try to build other models and observe if we can a better forecast. The model is evaluated by using the RMSE value and it is shown below,

	Test RMSE
RegressionOnTime	1389.135175

4.2. RMSE Linear Regression

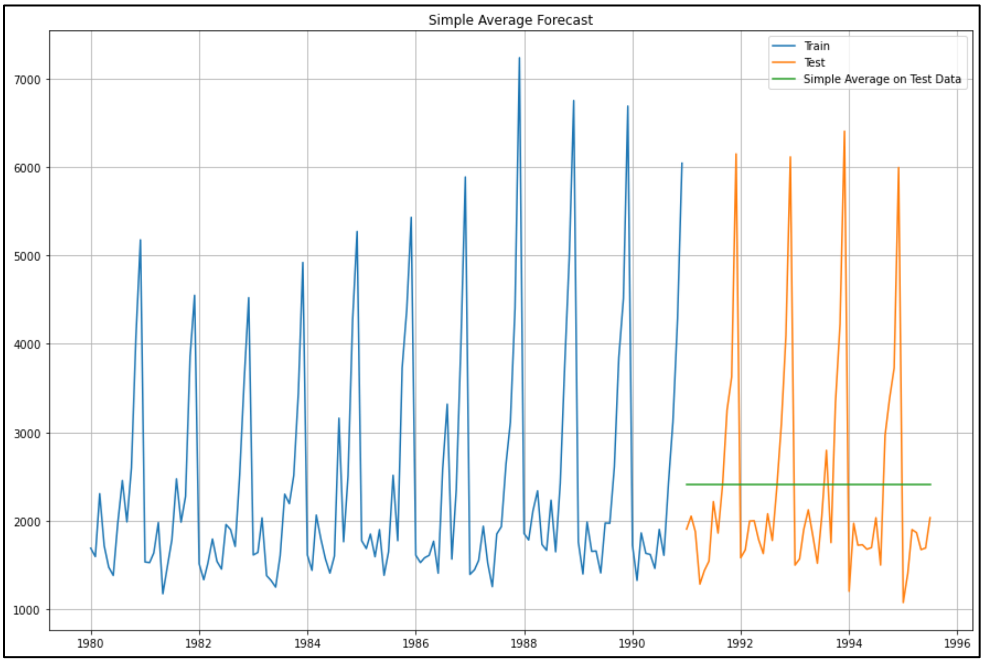
Naïve forecast



	Test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352

4.3. Naïve forecast (Sparkling), RMSE

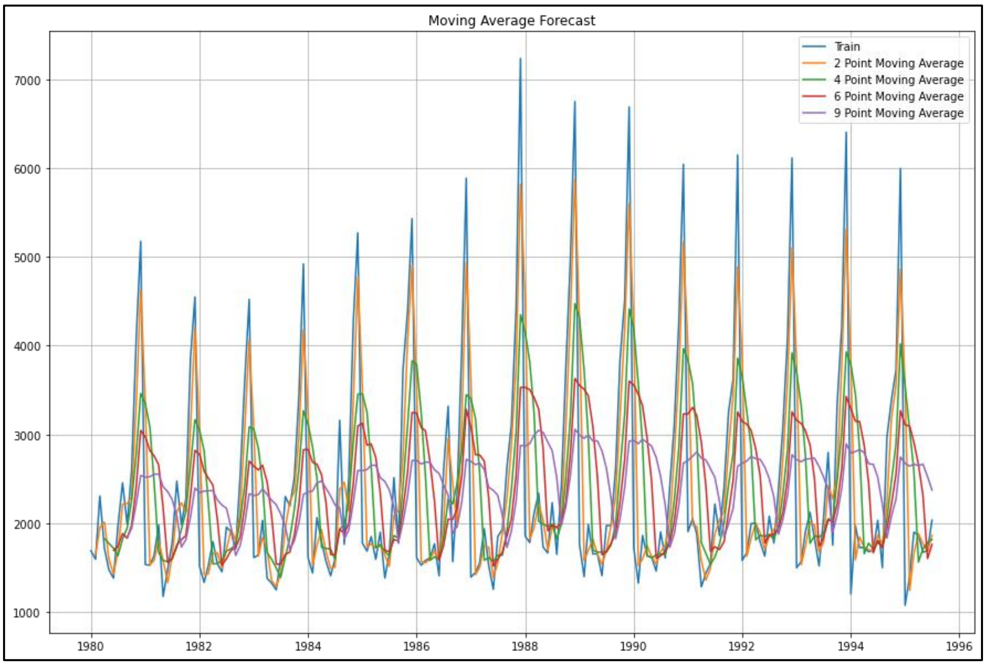
Simple Average forecast



	Test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804

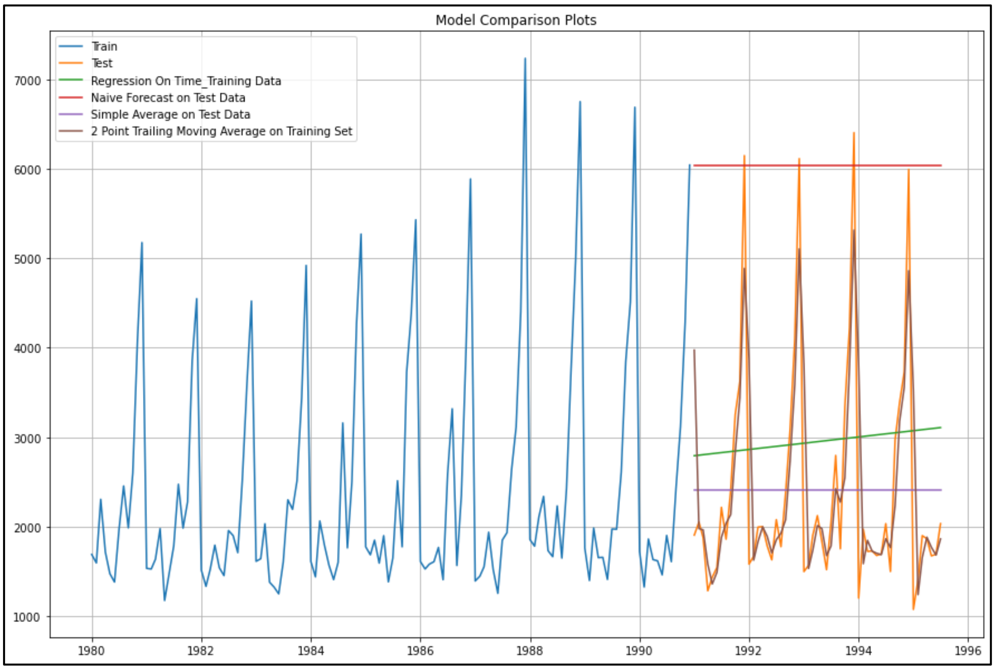
4.4. Simple average forecast (Sparkling), RMSE

Moving Average (MA)



	Test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315

4.5. Moving Average (MA-Sparkling), RMSE



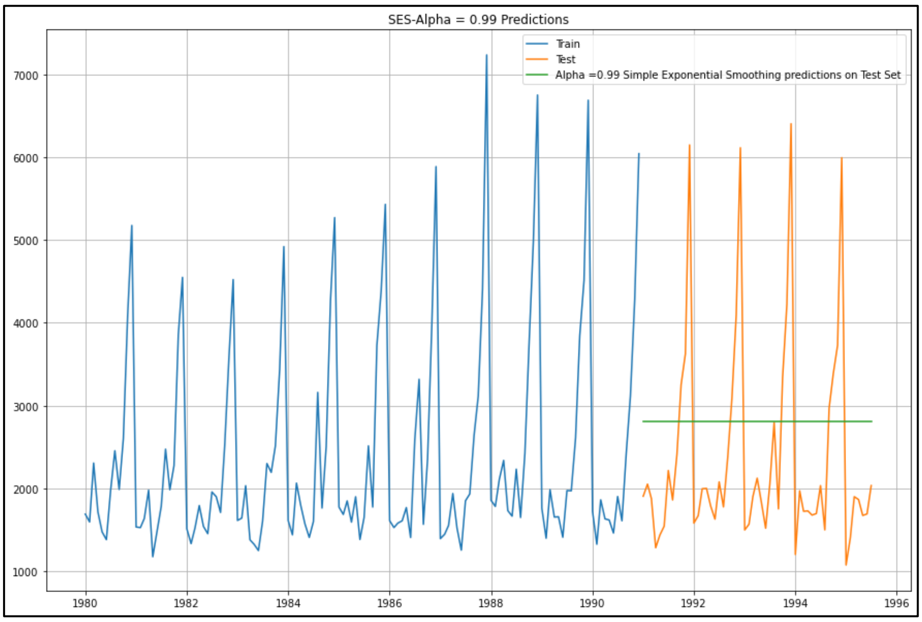
	Test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315

4.6. Model comparison, RMSE

From the model comparison plot and the RMSE model evaluation it is observed that the 2-point trailing MA model is having an optimum performance (RMSE-813.4) among the Regression, Naïve forecast, Simple average models.

The Smoothening model for the sparkling dataset is also built and its performance is discussed below,

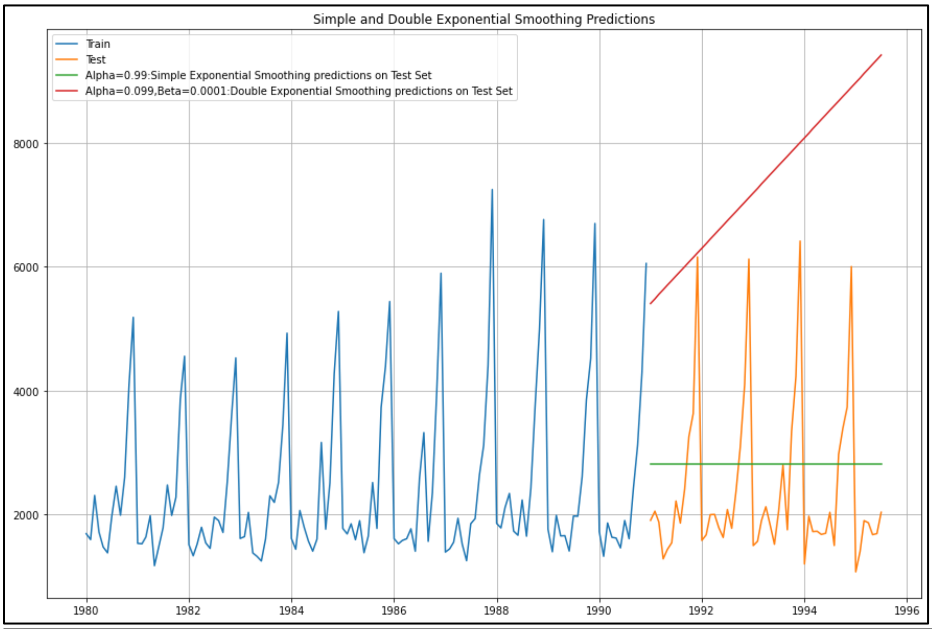
Simple Exponential Smoothing (Sparkling)




```
{'smoothing_level': 0.07028442075641193,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 1763.8402828521703,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

4.7. SES plot, optimum parameters

Double Exponential Smoothing (Sparkling)



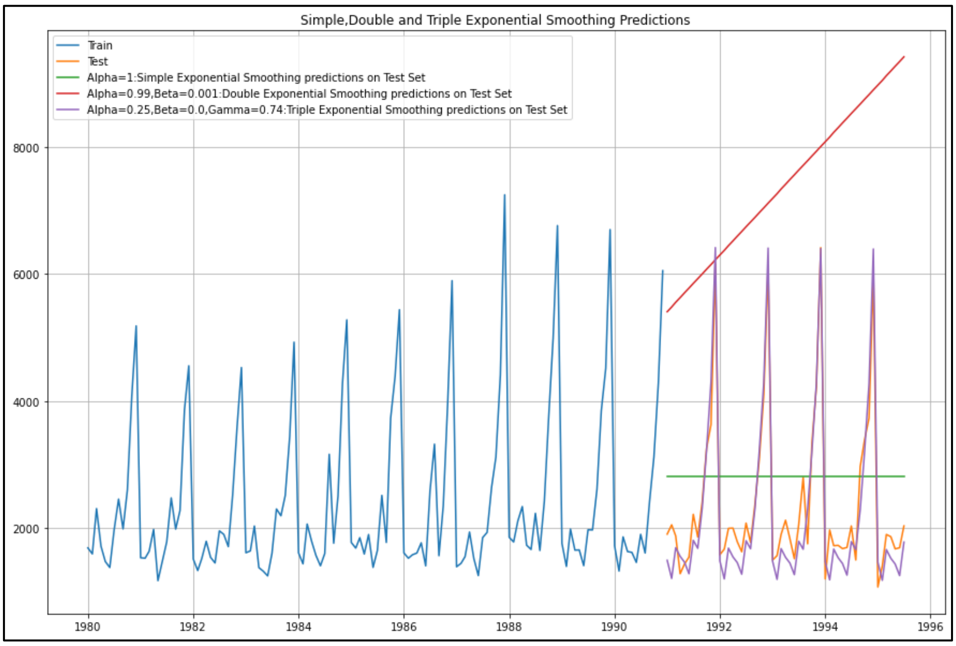

```
==Holt model Exponential Smoothing Estimated Parameters ==

{'smoothing_level': 0.6649999999999999, 'smoothing_trend': 0.0001, 'smoothing_seasonal': nan, 'damping_trend': nan, 'initial_le
vel': 1502.1999999999991, 'initial_trend': 74.87272727272739, 'initial_seasons': array([], dtype=float64), 'use_boxcox': False,
'lamda': None, 'remove_bias': False}
```

4.8. SES and DES plot-optimum parameters

The optimum parameters for the SES, DES models are found using the python and they are used to build the models.

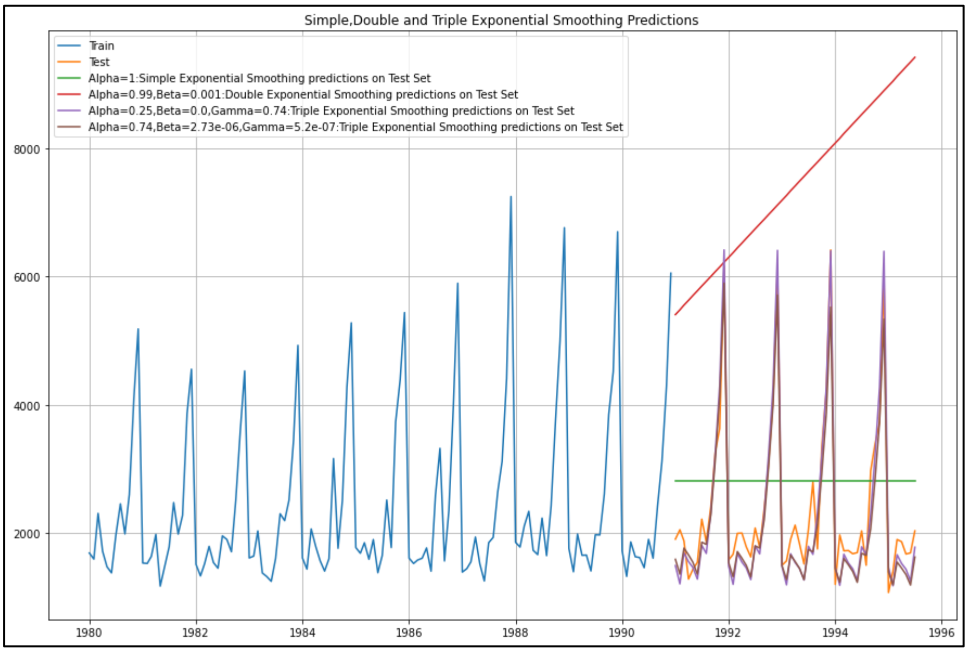
Triple Exponential Smoothing (Sparkling)




```
==Holt Winters model Exponential Smoothing Estimated Parameters ==

{'smoothing_level': 0.11127217693511166, 'smoothing_trend': 0.012360783126182025, 'smoothing_seasonal': 0.4607177659431463, 'da
mping_trend': nan, 'initial_level': 2356.5783078812697, 'initial_trend': -0.018442178724720648, 'initial_seasons': array([-636.
23349205, -722.98346399, -398.64349841, -473.43073157,
       -808.42502897, -815.35019273, -384.23061339,   72.99513671,
       -237.44278517,  272.32607144, 1541.37826596, 2590.07759442]), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}
```

4.9. SES and DES, TES plot-optimum parameters




```
==Holt Winters model Exponential Smoothing Estimated Parameters ==

{'smoothing_level': 0.11101471561088701, 'smoothing_trend': 0.0493145907614654, 'smoothing_seasonal': 0.36244934537370843, 'dam
ping_trend': nan, 'initial_level': 2356.496908624238, 'initial_trend': -9.809526161838415, 'initial_seasons': array([0.713711
, 0.68278724, 0.90458411, 0.8053878 , 0.65571739,
       0.65388935, 0.88616088, 1.13350811, 0.91894498, 1.21186447,
       1.87099202, 2.37505867]), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}
```

4.10. SES and DES, TES plot-Alternative parameters

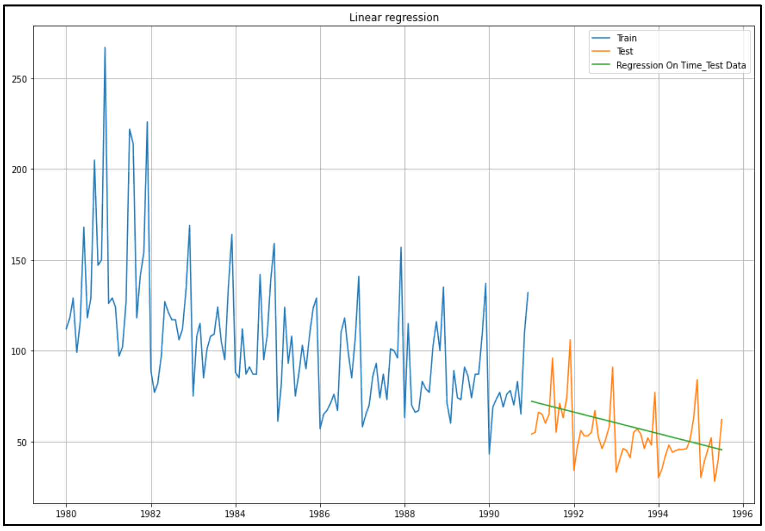
RMSE of smoothening models (Sparkling)

	Test RMSE
Alpha=0.99,SES	1338.000861
Alpha=1,Beta=0.0189:DES	5291.879833
Alpha=0.25,Beta=0.0,Gamma=0.74:TES	378.625883
Alpha=0.74,Beta=2.73e-06,Gamma=5.2e-07,Gamma=0:TES	402.936179

From the RMSE values it is observed that the TES model with RMSE (378.6) is the optimum smoothening model for forecasting. This TES model is even better that the 2-point trailing MA model (RMSE-813.4).

Forecast models for Rose wine:

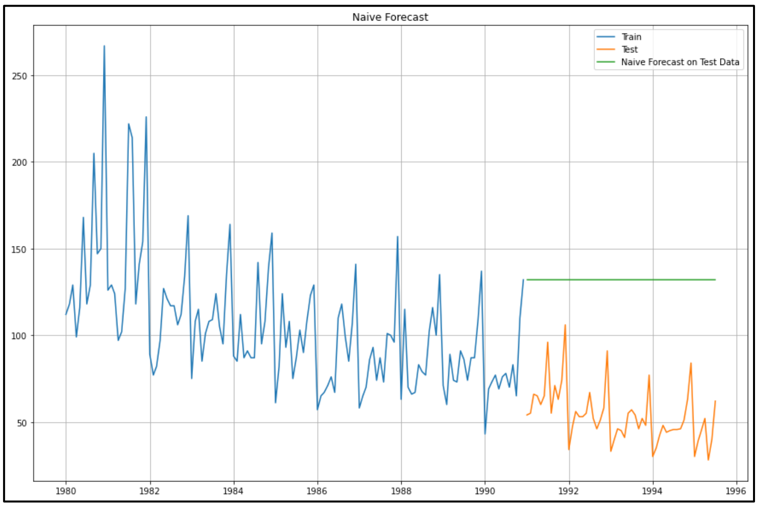
Linear Regression



	Test RMSE
RegressionOnTime	15.267514

4.11. Linear Regression (Rose), RMSE

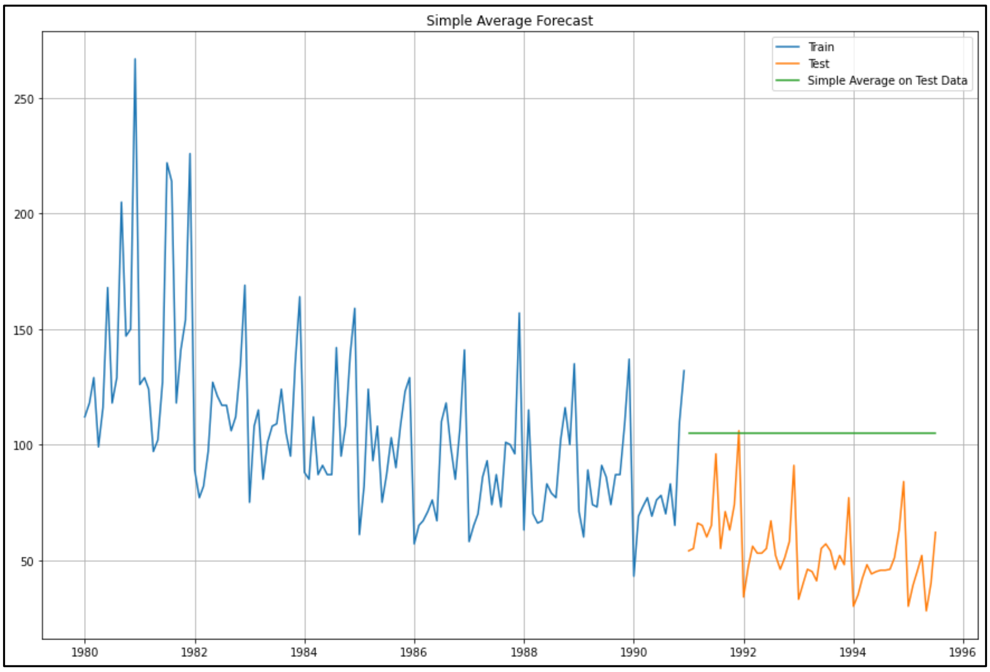
Naïve forecast



	Test RMSE
RegressionOnTime	15.267514
NaiveModel	79.714824

4.12. Naïve Forecast (Rose), RMSE

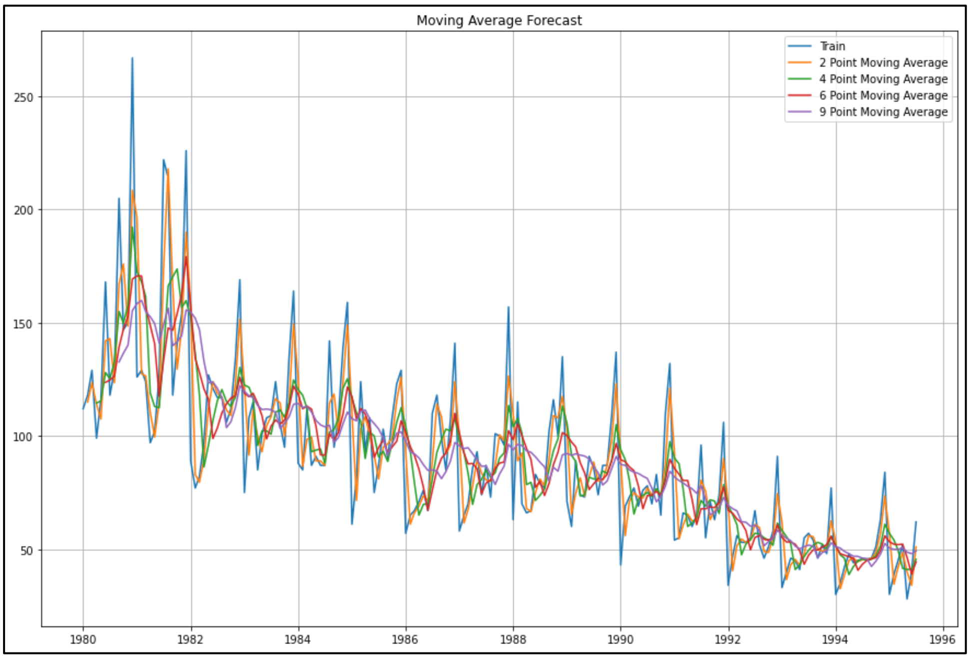
Simple Average



	Test RMSE
RegressionOnTime	15.267514
NaiveModel	79.714824
SimpleAverageModel	53.456520

4.13. Simple Average (Rose), RMSE

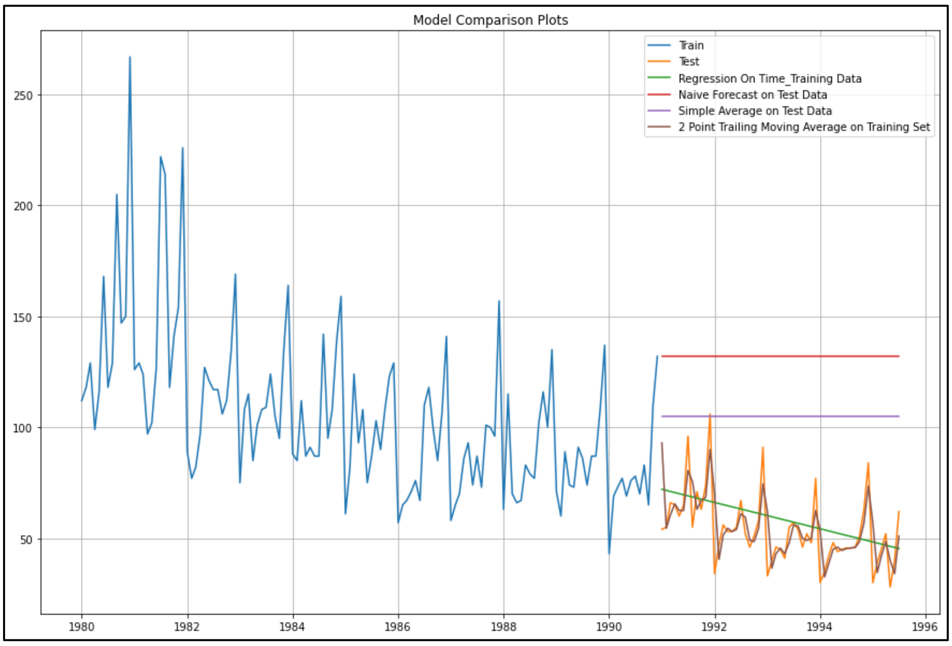
Moving Average



	Test RMSE
RegressionOnTime	15.267514
NaiveModel	79.714824
SimpleAverageModel	53.456520
2pointTrailingMovingAverage	11.529314
4pointTrailingMovingAverage	14.451239
6pointTrailingMovingAverage	14.564591
9pointTrailingMovingAverage	14.726926

4.14. Moving Average (Rose), RMSE

The Models are built for the Rose wine dataset and the respective model evaluations are made using the RMSE values. Let us make a comparison of the above models using a comparison plot.



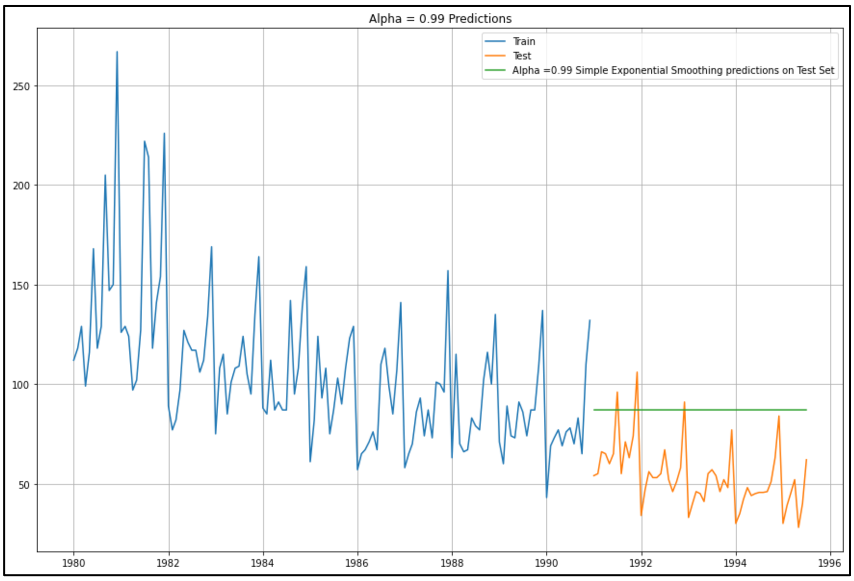
	Test RMSE
RegressionOnTime	15.267514
NaiveModel	79.714824
SimpleAverageModel	53.456520
2pointTrailingMovingAverage	11.529314
4pointTrailingMovingAverage	14.451239
6pointTrailingMovingAverage	14.564591
9pointTrailingMovingAverage	14.726926

4.15. Model comparison (Rose)

From the model comparison plot and the RMSE model evaluation it is observed that the 2-point trailing MA model is having an optimum performance (RMSE-11.5) among the Regression, Naïve forecast, Simple average models.

The Smoothening model for the sparkling dataset is also built and its performance is discussed below,

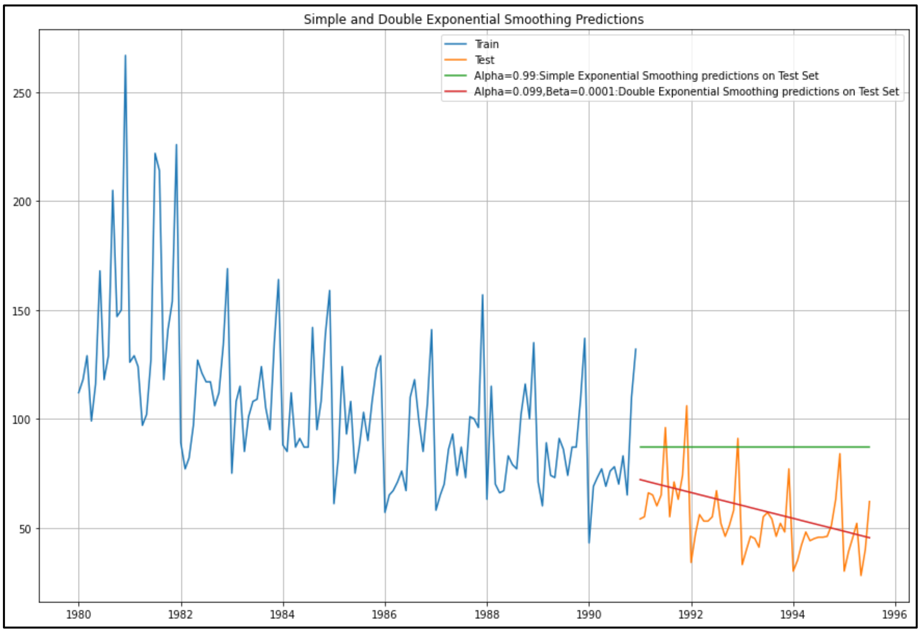
Simple Exponential Smoothing (Rose)




```
{'smoothing_level': 0.09874963957110783,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 134.38708961485827,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

4.16. SES plot, optimum parameters

Double Exponential Smoothing

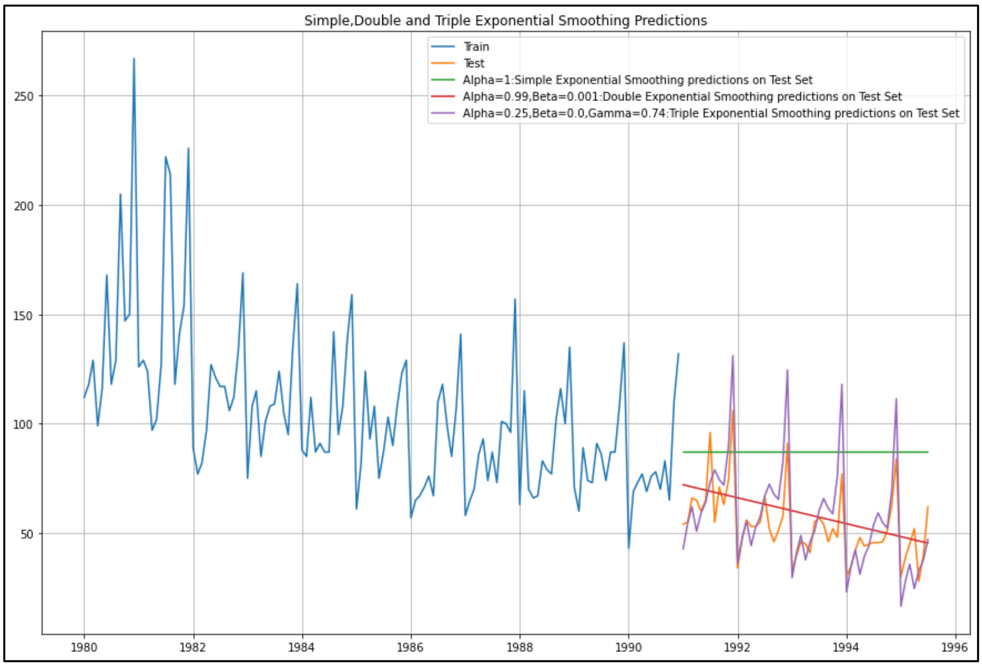



```
==Holt model Exponential Smoothing Estimated Parameters ==

{'smoothing_level': 1.4901247095597348e-08, 'smoothing_trend': 7.3896641488640725e-09, 'smoothing_seasonal': nan, 'damping_tren
d': nan, 'initial_level': 137.81551313502814, 'initial_trend': -0.4943777717865305, 'initial_seasons': array([], dtype=float6
4), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}
```

4.16. SES and DES plot, optimum parameters

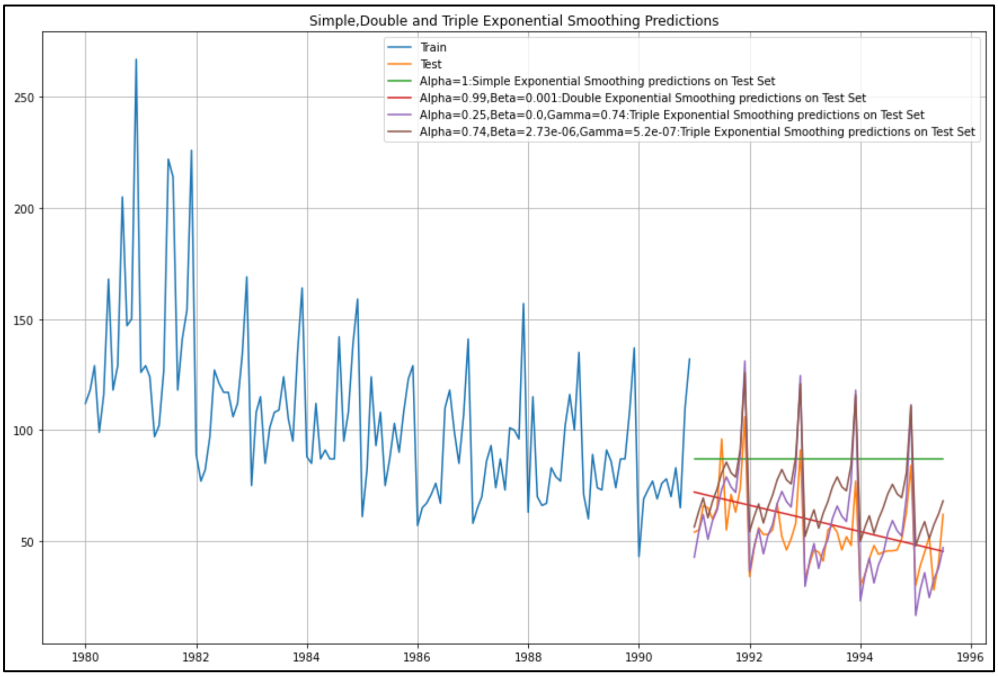
Triple Exponential Smoothing




```
==Holt Winters model Exponential Smoothing Estimated Parameters ==

{'smoothing_level': 0.09467987567540882, 'smoothing_trend': 2.31999683285252e-05, 'smoothing_seasonal': 0.0004175285691922314,
'damping_trend': nan, 'initial_level': 146.40142527639352, 'initial_trend': -0.5464913833622084, 'initial_seasons': array([-31.
19268548, -18.83344765, -10.84745053, -21.48718886,
       -12.67654312,  -7.19154248,   2.65454402,   8.80233514,
         4.79913097,   2.91389547,  21.00157004,  63.18716583]), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}
```

4.17. SES and DES, TES plot-optimum parameters




```
==Holt Winters model Exponential Smoothing Estimated Parameters ==

{'smoothing_level': 0.07130285749243212, 'smoothing_trend': 0.04550837652110988, 'smoothing_seasonal': 8.385716703273524e-05,
'damping_trend': nan, 'initial_level': 163.60092654560762, 'initial_trend': -0.9804841883026134, 'initial_seasons': array([0.68
714163, 0.77936108, 0.85184662, 0.74446365, 0.8372947 ,
       0.91182237, 1.00282327, 1.06745268, 1.01025249, 0.98957378,
       1.1535151 , 1.59037115]), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}
```

4.18.SES and DES, TES plot-Alternative parameters

RMSE of smoothening models (Rose)

	Test RMSE
Alpha=0.99,SES	36.792115
Alpha=1,Beta=0.0189:DES	15.267515
Alpha=0.25,Beta=0.0,Gamma=0.74:TES	14.276827
Alpha=0.74,Beta=2.73e-06,Gamma=5.2e-07,Gamma=0:TES	20.185370

From the RMSE values it is observed that the TES model with RMSE (14.27) is the optimum smoothening model for forecasting. But comparatively the 2-point trailing MA model is having a better optimum overall performance (RMSE-11.5) than the smoothening model.

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.

The Augmented Dickey-Fuller test is a unit root test which determines whether there is a unit root and subsequently whether the series is non-stationary. The hypothesis in a simple form for the ADF test is:

H_0 : The Time Series has a unit root and is thus non-stationary.

H_1 : The Time Series does not have a unit root and is thus stationary.

We would want the series to be stationary for building ARIMA models and thus we would want the p-value of this test to be less than the α value (0.05).

Stationarity check (Sparkling dataset)

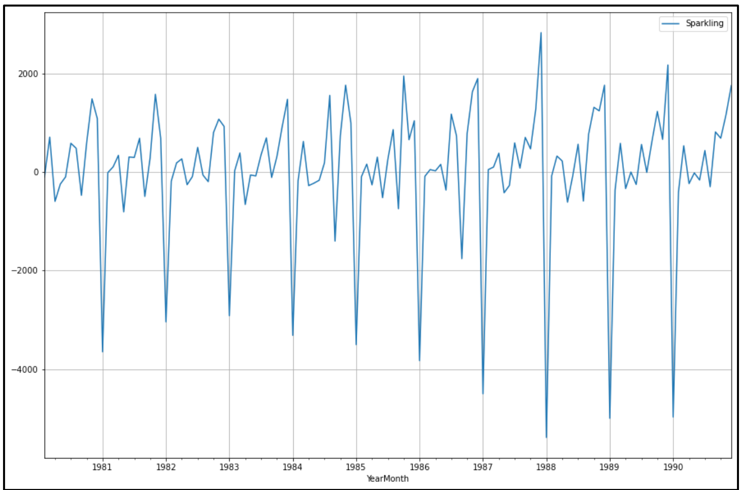
```
DF test statistic is -1.798
DF test p-value is 0.7055958459932692
Number of lags used 12
```

We see that at 5% significant level the Sparkling data series is non-stationary, (p>α).

Let us take one level of differencing to make the series stationary.

```
DF test statistic is -44.912
DF test p-value is 0.0
Number of lags used 10
```

p-value is less than 0.05, the sparkling dataset is now stationary.




```
DF test statistic is -2.062
DF test p-value is 0.5674110388593684
Number of lags used 12
```

```
DF test statistic is -7.968
DF test p-value is 8.47921065551504e-11
Number of lags used 11
```

5.1. Stationarity of Training data (before & after differencing)

Stationarity check (Rose dataset)

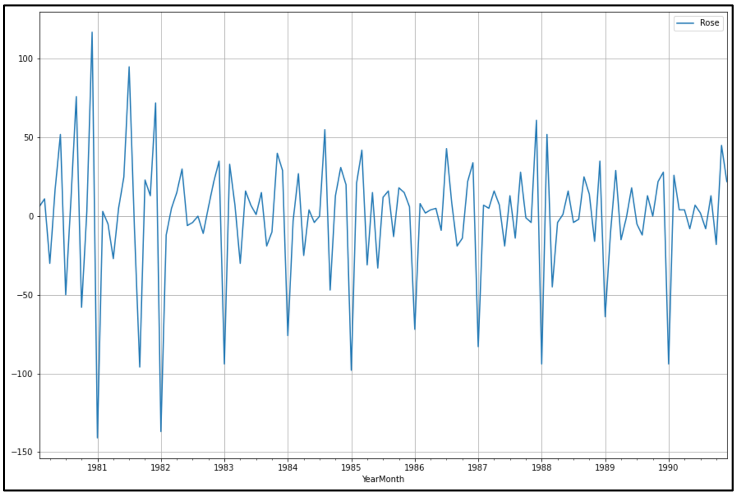
```
DF test statistic is -2.241
DF test p-value is 0.4669327030226468
Number of lags used 13
```

We see that at 5% significant level the Rose data series is non-stationary, (p>α).

Let us take one level of differencing to make the series stationary.

```
DF test statistic is -8.162
DF test p-value is 3.015793288348449e-11
Number of lags used 12
```

p-value is less than 0.05, the Rose dataset is now stationary.




```
DF test statistic is -1.686
DF test p-value is 0.7569093051047106
Number of lags used 13
```

```
DF test statistic is -6.804
DF test p-value is 3.894831356781761e-08
Number of lags used 12
```

5.2. Stationarity of Training data (before & after differencing)

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 (Sparkling Dataset)

The combination of different parameters of p and q in the range of 0 and 2 are used. The value of d is kept as 1 as we need to take a difference of the series to make it stationary.

```
Examples of the 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)
```

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1101.755
Date:                Sun, 11 Dec 2022   AIC                           2213.509
Time:                        18:36:10   BIC                           2227.885
Sample:                    01-01-1980   HQIC                          2219.351
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.3121      0.046     28.782      0.000       1.223       1.401
ar.L2         -0.5593      0.072     -7.740      0.000      -0.701      -0.418
ma.L1         -1.9917      0.109    -18.216      0.000      -2.206      -1.777
ma.L2          0.9999      0.110      9.109      0.000       0.785       1.215
sigma2      1.099e+06   1.99e-07   5.51e+12      0.000     1.1e+06     1.1e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.19   Jarque-Bera (JB):                14.46
Prob(Q):                              0.67   Prob(JB):                         0.00
Heteroskedasticity (H):               2.43   Skew:                             0.61
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.08
===================================================================================
```

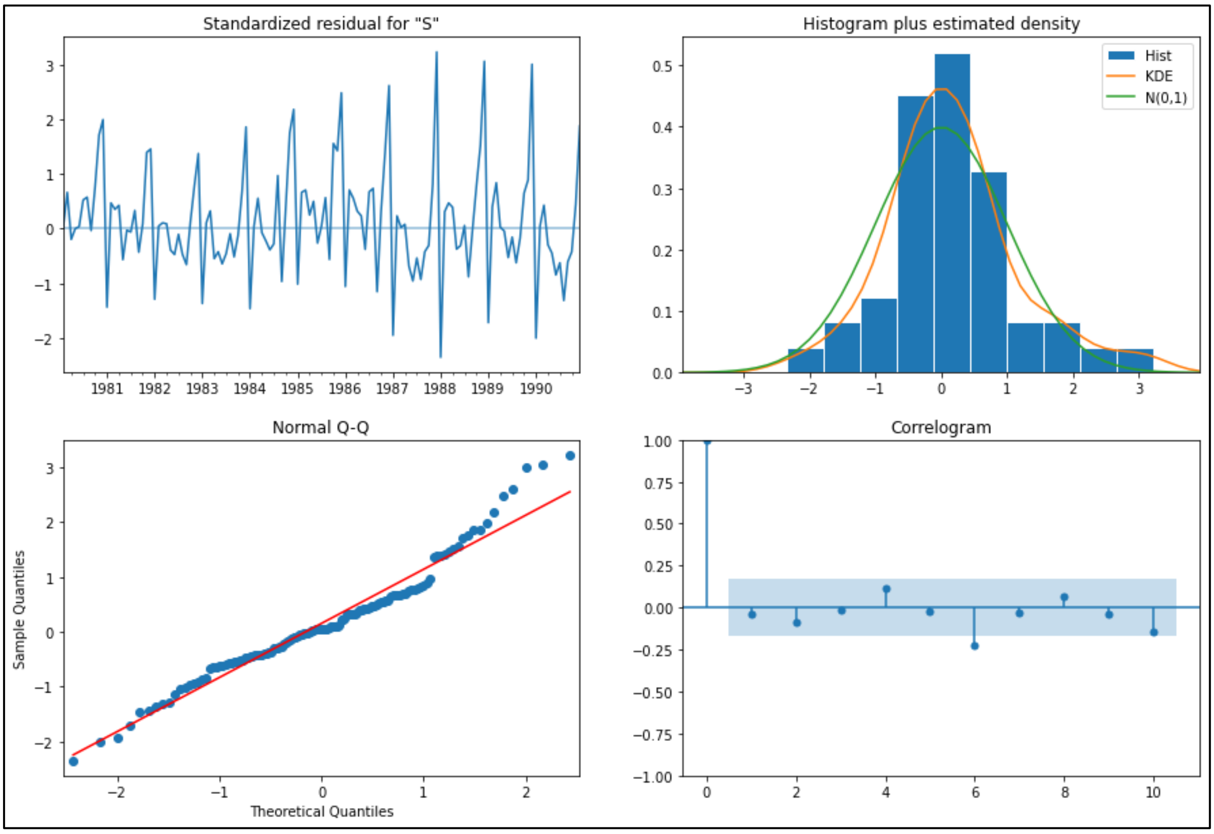
	param	AIC
10	(2, 1, 2)	2213.509212
15	(3, 1, 3)	2221.461689
14	(3, 1, 2)	2230.825009
11	(2, 1, 3)	2232.811211
9	(2, 1, 1)	2233.777626

6.1. Auto ARIMA summary, AIC values

The parameters (2, 1, 2) with AIC value 2213.51 is selected for the model.

	RMSE	MAPE
ARIMA(2,1,2)	1299.979821	47.099974

6.2. Auto ARIMA RMSE, MAPE



6.3. Auto ARIMA diagnostic plot

Automated SARIMA (Sparkling Dataset)

Like the Automated ARIMA model, the parameter combinations are generated for the automated SARIMA model. Here the order of seasonality is taken as 4 based on the observations from the ACF plot.

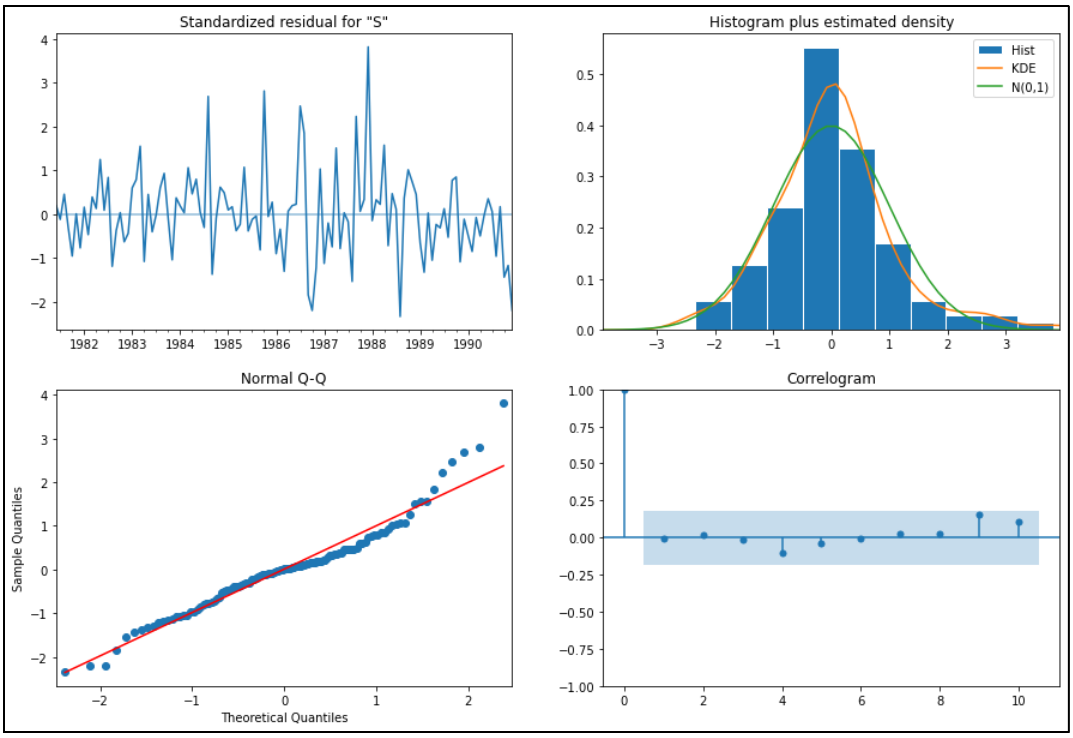
```
Examples of the parameter combinations for the Model are
Model: (0, 1, 1)(0, 0, 1, 4)
Model: (0, 1, 2)(0, 0, 2, 4)
Model: (0, 1, 3)(0, 0, 3, 4)
Model: (1, 1, 0)(1, 0, 0, 4)
Model: (1, 1, 1)(1, 0, 1, 4)
Model: (1, 1, 2)(1, 0, 2, 4)
Model: (1, 1, 3)(1, 0, 3, 4)
Model: (2, 1, 0)(2, 0, 0, 4)
Model: (2, 1, 1)(2, 0, 1, 4)
Model: (2, 1, 2)(2, 0, 2, 4)
Model: (2, 1, 3)(2, 0, 3, 4)
Model: (3, 1, 0)(3, 0, 0, 4)
Model: (3, 1, 1)(3, 0, 1, 4)
Model: (3, 1, 2)(3, 0, 2, 4)
Model: (3, 1, 3)(3, 0, 3, 4)
```

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                          Sparkling   No. Observations:                  132
Model:             SARIMAX(1, 1, 3)x(3, 0, 3, 4)   Log Likelihood                -844.750
Date:                            Sat, 10 Dec 2022   AIC                           1711.501
Time:                                    22:01:37   BIC                           1741.695
Sample:                                01-01-1980   HQIC                          1723.757
                                     - 12-01-1990
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5053      0.438     -1.154      0.248      -1.363       0.353
ma.L1         -0.2776      0.444     -0.625      0.532      -1.149       0.593
ma.L2         -0.5892      0.313     -1.884      0.060      -1.202       0.024
ma.L3          0.0808      0.151      0.536      0.592      -0.215       0.376
ar.S.L4       -0.0072      0.010     -0.747      0.455      -0.026       0.012
ar.S.L8       -0.0245      0.009     -2.782      0.005      -0.042      -0.007
ar.S.L12       1.0436      0.008    136.816      0.000       1.029       1.059
ma.S.L4        0.1131      0.253      0.448      0.654      -0.382       0.608
ma.S.L8        0.0758      0.260      0.291      0.771      -0.434       0.586
ma.S.L12      -1.1879      0.145     -8.217      0.000      -1.471      -0.905
sigma2      8.785e+04   7.45e-06   1.18e+10      0.000    8.79e+04    8.79e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                27.06
Prob(Q):                              0.93   Prob(JB):                         0.00
Heteroskedasticity (H):               2.61   Skew:                             0.69
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.93
===================================================================================
```

	param	seasonal	AIC
63	(0, 1, 3)	(3, 0, 3, 4)	1710.552843
127	(1, 1, 3)	(3, 0, 3, 4)	1711.500926
191	(2, 1, 3)	(3, 0, 3, 4)	1712.736903
255	(3, 1, 3)	(3, 0, 3, 4)	1714.608407
251	(3, 1, 3)	(2, 0, 3, 4)	1714.755979

6.4. Auto SARIMA summary, AIC values

The parameters (1, 1, 3), seasonal (3, 0, 3, 4) with AIC value 1711.5 is selected for the auto SARIMA model.



	RMSE	MAPE
ARIMA(2,1,2)	1299.979821	47.099974
SARIMA(1, 1, 3)(3, 0, 3, 4)	596.585216	22.389227

6.5. Auto SARIMA diagnostic plot, RMSE, MAPE

Automated ARIMA (Rose Dataset)

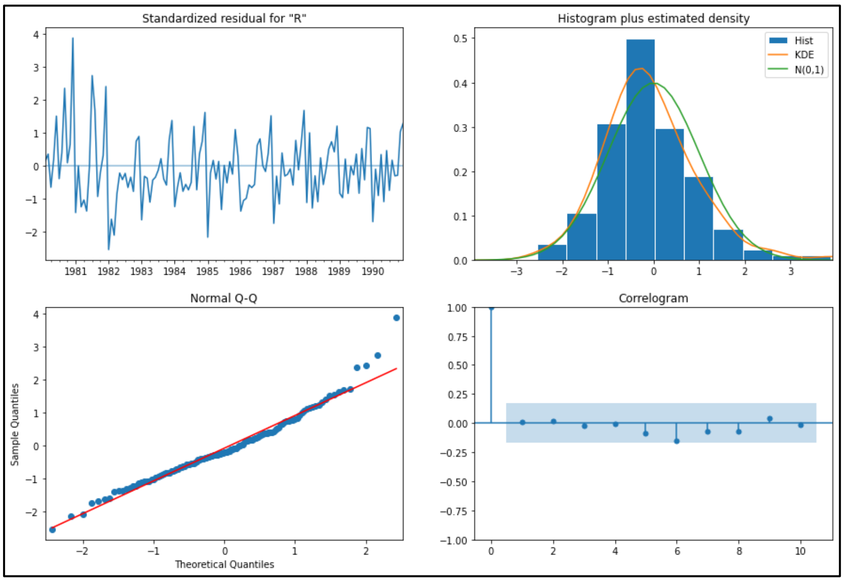
```
Examples of the parameter combinations for the Model
Model: (0, 1, 0)
Model: (0, 1, 1)
Model: (0, 1, 2)
Model: (0, 1, 3)
Model: (1, 1, 0)
Model: (1, 1, 1)
Model: (1, 1, 2)
Model: (1, 1, 3)
Model: (2, 1, 0)
Model: (2, 1, 1)
Model: (2, 1, 2)
Model: (2, 1, 3)
Model: (3, 1, 0)
Model: (3, 1, 1)
Model: (3, 1, 2)
Model: (3, 1, 3)
```

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(2, 1, 3)   Log Likelihood                -631.348
Date:                Sun, 11 Dec 2022   AIC                           1274.695
Time:                        20:01:01   BIC                           1291.947
Sample:                    01-01-1980   HQIC                          1281.705
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.6781      0.084    -19.992      0.000      -1.843      -1.514
ar.L2         -0.7289      0.084     -8.684      0.000      -0.893      -0.564
ma.L1          1.0446      0.628      1.665      0.096      -0.185       2.275
ma.L2         -0.7720      0.133     -5.824      0.000      -1.032      -0.512
ma.L3         -0.9046      0.569     -1.590      0.112      -2.020       0.210
sigma2       860.6996    528.714      1.628      0.104    -175.560    1896.959
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                24.48
Prob(Q):                              0.88   Prob(JB):                         0.00
Heteroskedasticity (H):               0.40   Skew:                             0.71
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.57
===================================================================================
```

	param	AIC
11	(2, 1, 3)	1274.695356
15	(3, 1, 3)	1278.661965
2	(0, 1, 2)	1279.671529
6	(1, 1, 2)	1279.870723
3	(0, 1, 3)	1280.545376

6.6. Auto ARIMA summary, AIC values

The parameters (2, 1, 3) with AIC value 1274.7 is selected for the auto ARIMA model.



	RMSE	MAPE
ARIMA(2,1,3)	36.809369	75.824713

6.7. Auto ARIMA diagnostic plot, RMSE, MAPE

Automated SARIMA (Rose Dataset)

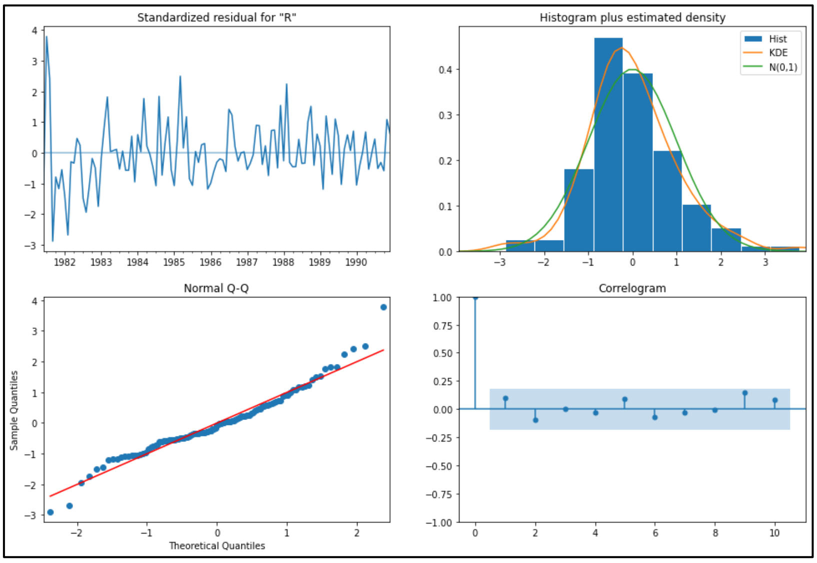
```
Examples of the parameter combinations for the Model are
Model: (0, 1, 1)(0, 0, 1, 4)
Model: (0, 1, 2)(0, 0, 2, 4)
Model: (0, 1, 3)(0, 0, 3, 4)
Model: (1, 1, 0)(1, 0, 0, 4)
Model: (1, 1, 1)(1, 0, 1, 4)
Model: (1, 1, 2)(1, 0, 2, 4)
Model: (1, 1, 3)(1, 0, 3, 4)
Model: (2, 1, 0)(2, 0, 0, 4)
Model: (2, 1, 1)(2, 0, 1, 4)
Model: (2, 1, 2)(2, 0, 2, 4)
Model: (2, 1, 3)(2, 0, 3, 4)
Model: (3, 1, 0)(3, 0, 0, 4)
Model: (3, 1, 1)(3, 0, 1, 4)
Model: (3, 1, 2)(3, 0, 2, 4)
Model: (3, 1, 3)(3, 0, 3, 4)
```

```
                                      SARIMAX Results
=========================================================================================
Dep. Variable:                                Rose   No. Observations:                132
Model:             SARIMAX(2, 1, 3)x(3, 0, 3, 4)   Log Likelihood                -501.204
Date:                             Sun, 11 Dec 2022   AIC                           1026.409
Time:                                     20:30:29   BIC                           1059.348
Sample:                                 01-01-1980   HQIC                          1039.779
                                      - 12-01-1990
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.6781      0.035    -19.370      0.000      -0.747      -0.609
ar.L2         -0.8521      0.031    -27.149      0.000      -0.914      -0.791
ma.L1         -0.1524     19.864     -0.008      0.994     -39.084      38.780
ma.L2          0.2678     32.164      0.008      0.993     -62.772      63.308
ma.L3         -0.9352     47.446     -0.020      0.984     -93.927      92.057
ar.S.L4        0.0106      0.032      0.336      0.737      -0.051       0.073
ar.S.L8       -0.0420      0.026     -1.591      0.112      -0.094       0.010
ar.S.L12       0.8998      0.025     35.727      0.000       0.850       0.949
ma.S.L4       -0.0370    211.473     -0.000      1.000    -414.517     414.443
ma.S.L8        0.0370    204.619      0.000      1.000    -401.009     401.083
ma.S.L12      -1.0002    247.684     -0.004      0.997    -486.451     484.451
sigma2       261.5209   6.51e+04      0.004      0.997   -1.27e+05    1.28e+05
===================================================================================
Ljung-Box (L1) (Q):                   1.11   Jarque-Bera (JB):                18.96
Prob(Q):                              0.29   Prob(JB):                         0.00
Heteroskedasticity (H):               0.38   Skew:                             0.50
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.72
===================================================================================
```

	param	seasonal	AIC
191	(2, 1, 3)	(3, 0, 3, 4)	1026.408988
254	(3, 1, 3)	(3, 0, 2, 4)	1034.712742
63	(0, 1, 3)	(3, 0, 3, 4)	1040.705448
127	(1, 1, 3)	(3, 0, 3, 4)	1041.926683
255	(3, 1, 3)	(3, 0, 3, 4)	1046.087758

6.8. Auto SARIMA summary, AIC values

The parameters (2, 1, 3), seasonal (3, 0, 3, 4) with AIC value 1026.41 is selected for the auto SARIMA model.

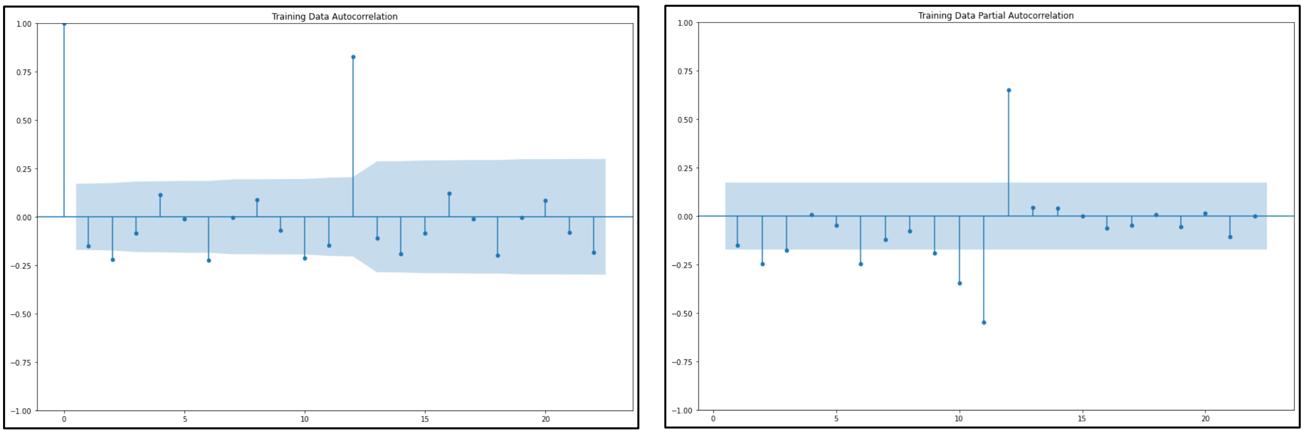


	RMSE	MAPE
ARIMA(2,1,3)	36.809369	75.824713
SARIMA(2, 1, 3)(3, 0, 3, 4)	21.501007	43.054333

6.9. Auto SARIMA diagnostic plot, RMSE, MAPE

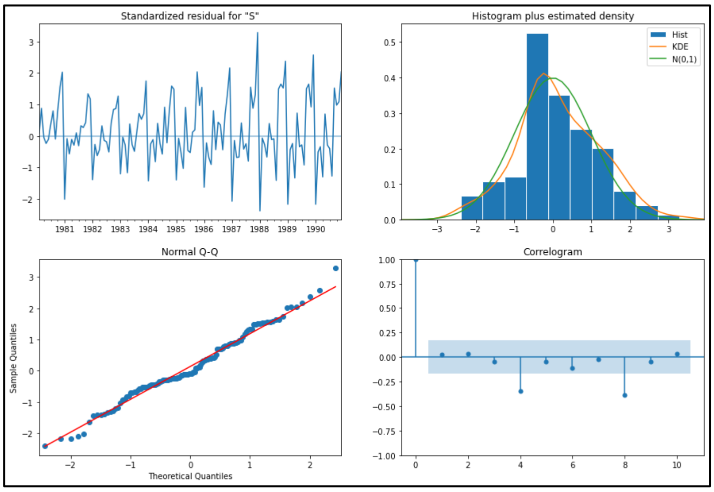
7. **Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.**

Manual ARIMA (Sparkling Dataset)



7.1. ACF and PACF plots

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(4, 1, 4)   Log Likelihood               -1097.624
Date:                Sun, 11 Dec 2022   AIC                           2213.248
Time:                        21:46:02   BIC                           2239.125
Sample:                    01-01-1980   HQIC                          2223.763
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.4452      0.109     -4.087      0.000      -0.659      -0.232
ar.L2         -0.4492      0.076     -5.926      0.000      -0.598      -0.301
ar.L3         -0.4463      0.088     -5.091      0.000      -0.618      -0.275
ar.L4          0.5500      0.068      8.126      0.000       0.417       0.683
ma.L1         -0.0044      7.181     -0.001      1.000     -14.079      14.070
ma.L2          0.0181     14.247      0.001      0.999     -27.905      27.942
ma.L3         -0.0328      6.920     -0.005      0.996     -13.595      13.529
ma.L4         -0.9809      0.156     -6.287      0.000      -1.287      -0.675
sigma2      9.083e+05   3.05e-05   2.98e+10      0.000    9.08e+05    9.08e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.11   Jarque-Bera (JB):                 0.68
Prob(Q):                              0.74   Prob(JB):                         0.71
Heteroskedasticity (H):               2.83   Skew:                             0.17
Prob(H) (two-sided):                  0.00   Kurtosis:                         3.06
===================================================================================
```

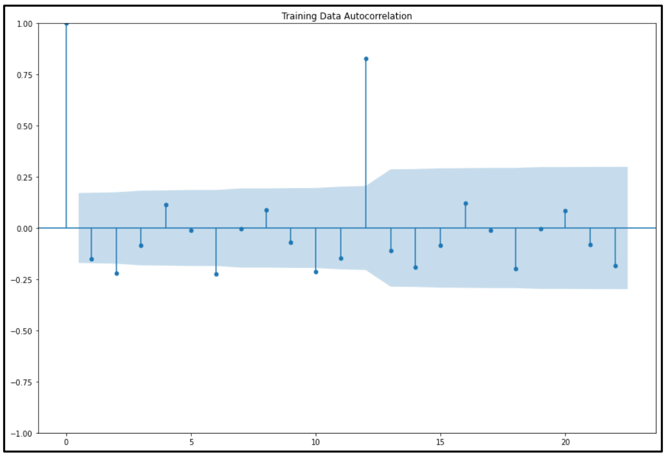


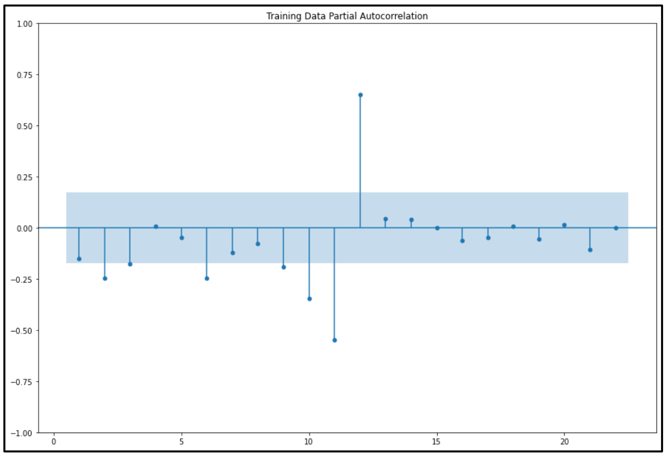
7.2. MANUAL ARIMA summary and diagnostic plot

	RMSE	MAPE
ARIMA(2,1,2)	1299.979821	47.099974
SARIMA(1, 1, 3)(3, 0, 3, 4)	596.585216	22.389227
ARIMA(4,1,4)	1212.918076	40.214639

7.3. MANUAL ARIMA RMSE, MAPE

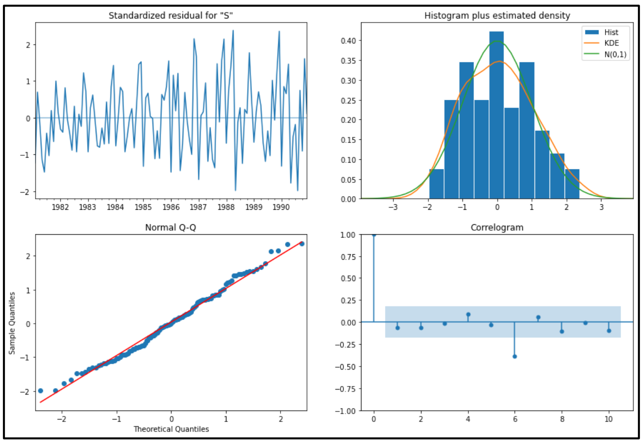
Manual SARIMA (Sparkling Dataset)





7.4. ACF and PACF plots

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                             Sparkling   No. Observations:               132
Model:             SARIMAX(1, 1, 3)x(0, 0, [1, 2], 4)   Log Likelihood                -985.398
Date:                               Sun, 11 Dec 2022   AIC                           1984.797
Time:                                       21:46:03   BIC                           2004.251
Sample:                                   01-01-1980   HQIC                          1992.697
                                        - 12-01-1990
Covariance Type:                                 opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.9980      0.005   -216.319      0.000      -1.007      -0.989
ma.L1          0.5828      0.761      0.766      0.444      -0.909       2.075
ma.L2         -1.0017      0.396     -2.529      0.011      -1.778      -0.225
ma.L3         -0.5846      0.120     -4.869      0.000      -0.820      -0.349
ma.S.L4       -0.6477      0.077     -8.404      0.000      -0.799      -0.497
ma.S.L8        0.7239      0.075      9.678      0.000       0.577       0.870
sigma2      8.494e+05   7.72e-07    1.1e+12      0.000    8.49e+05    8.49e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.42   Jarque-Bera (JB):                 2.56
Prob(Q):                              0.52   Prob(JB):                         0.28
Heteroskedasticity (H):               2.49   Skew:                             0.20
Prob(H) (two-sided):                  0.00   Kurtosis:                         2.41
===================================================================================
```

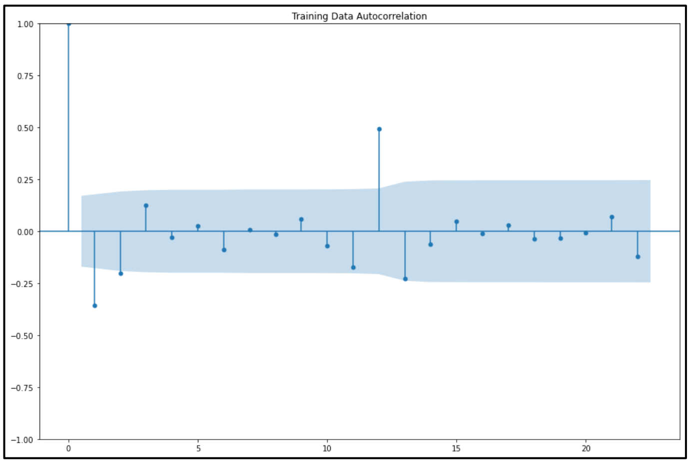


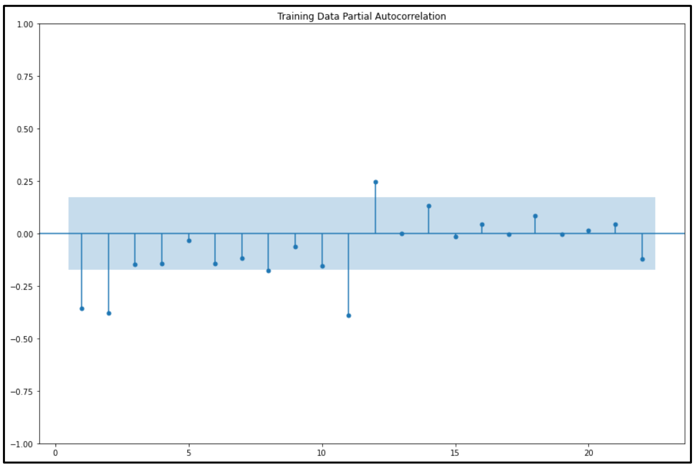
7.5. MANUAL SARIMA summary and diagnostic plot

	RMSE	MAPE
ARIMA(2,1,2)	1299.979821	47.099974
SARIMA(1, 1, 3)(3, 0, 3, 4)	596.585216	22.389227
ARIMA(4,1,4)	1212.918076	40.214639
SARIMA(1,1,3)(0,0,2,4)	1252.716993	38.122371

7.6. MANUAL SARIMA RMSE, MAPE

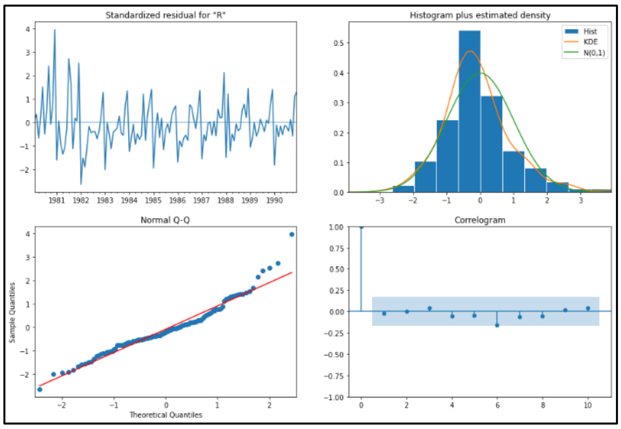
Manual ARIMA (Rose Dataset)





7.7. ACF and PACF plots

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(0, 1, 3)   Log Likelihood                -636.273
Date:                Sun, 11 Dec 2022   AIC                           1280.545
Time:                        22:12:30   BIC                           1292.046
Sample:                    01-01-1980   HQIC                          1285.219
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.7035      0.070    -10.117      0.000      -0.840      -0.567
ma.L2         -0.2887      0.103     -2.807      0.005      -0.490      -0.087
ma.L3          0.0939      0.084      1.122      0.262      -0.070       0.258
sigma2       957.0093     90.801     10.540      0.000     779.043    1134.975
===================================================================================
Ljung-Box (L1) (Q):                   0.05   Jarque-Bera (JB):                35.24
Prob(Q):                              0.82   Prob(JB):                         0.00
Heteroskedasticity (H):               0.37   Skew:                             0.81
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.96
===================================================================================
```

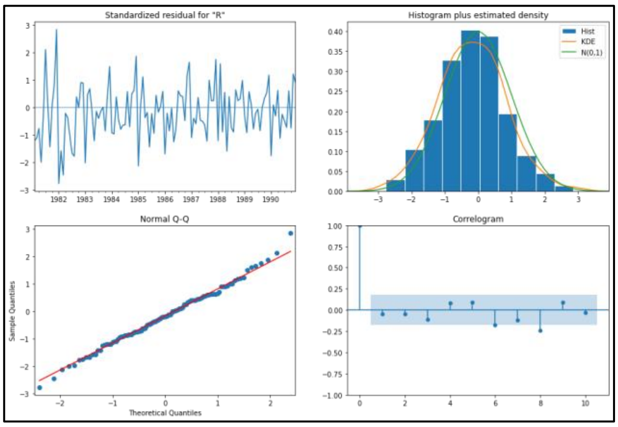


ARIMA(0,1,3) 36.715834 75.712739

7.8. MANUAL ARIMA summary and diagnostic plot

Manual SARIMA (Rose Dataset)

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                               Rose   No. Observations:                  132
Model:             SARIMAX(0, 1, 3)x(0, 0, [1, 2], 4)   Log Likelihood                -567.422
Date:                            Sun, 11 Dec 2022   AIC                           1146.844
Time:                                    22:14:00   BIC                           1163.519
Sample:                                01-01-1980   HQIC                          1153.615
                                     - 12-01-1990
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.6577      0.085     -7.708      0.000      -0.825      -0.490
ma.L2         -0.2539      0.098     -2.584      0.010      -0.447      -0.061
ma.L3          0.0285      0.084      0.341      0.733      -0.135       0.192
ma.S.L4       -0.3374      0.083     -4.047      0.000      -0.501      -0.174
ma.S.L8        0.4610      0.116      3.960      0.000       0.233       0.689
sigma2       791.7344    106.423      7.439      0.000     583.148    1000.320
===================================================================================
Ljung-Box (L1) (Q):                   0.30   Jarque-Bera (JB):                 0.45
Prob(Q):                              0.59   Prob(JB):                         0.80
Heteroskedasticity (H):               0.47   Skew:                             0.10
Prob(H) (two-sided):                  0.02   Kurtosis:                         3.22
===================================================================================
```



SARIMA(0,1,3)(0,0,2,4) 34.496181 70.201699

7.9. MANUAL SARIMA summary and diagnostic plot

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

RMSE values of models on test data as follows,

Sparking dataset (RMSE values)

Regression, Naïve, Simple Avg , MA models		**Smoothing models (SES, DES, TES)**	
RegressionOnTime	1389.135175	**Alpha=0.99, SES**	1338.000861
NaiveModel	3864.279352	**Alpha=1, Beta=0.0189: DES**	5291.879833
SimpleAverageModel	1275.081804	**Alpha=0.25, Beta=0.0, Gamma=0.74: TES**	378.625883
2pointTrailingMovingAvg	813.400684	**Alpha=0.74, Beta=2.73e-06, Gamma=5.2e-07, Gamma=0:TES**	402.936179
4pointTrailingMovingAvg	1156.589694		
6pointTrailingMovingAvg	1283.927428		
9pointTrailingMovingAvg	1346.278315		

ARIMA, SARIMA models (RMSE, MAPE)			**Parameters**
AUTO ARIMA (2,1,2)	1299.979821	47.099974	The optimal parameters for each model are shown in the respective model sections.
AUTO SARIMA (1, 1, 3) (3, 0, 3, 4)	596.585216	22.389227	
MANUAL ARIMA (4,1,4)	1212.918076	40.214639	
MANUAL SARIMA (1,1,3) (0,0,2,4)	1252.716993	38.122371	

Rose dataset (RMSE values)

Regression, Naïve, Simple Avg , MA models		Smoothing models (SES, DES, TES)	
RegressionOnTime	15.267514	**Alpha=0.99, SES**	36.792115
NaiveModel	79.714824	**Alpha=1, Beta=0.0189: DES**	15.267515
SimpleAverageModel	53.456520	**Alpha=0.25, Beta=0.0, Gamma=0.74: TES**	14.276827
2pointTrailingMovingAvg	11.529314	**Alpha=0.74, Beta=2.73e-06, Gamma=5.2e-07, Gamma=0: TES**	20.185370
4pointTrailingMovingAvg	14.451239		
6pointTrailingMovingAvg	14.564591		
9pointTrailingMovingAvg	14.726926		

ARIMA, SARIMA models (RMSE, MAPE)			Parameters
AUTO ARIMA (2,1,3)	36.80937	75.82471	The optimal parameters for each model are shown in the respective model sections.
AUTO SARIMA (2, 1, 3) (3, 0, 3, 4)	21.50101	43.05433	
MANUAL ARIMA (0,1,3)	36.71583	75.71274	
MANUAL SARIMA (0,1,3) (0,0,2,4)	34.49618	70.2017	

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.**

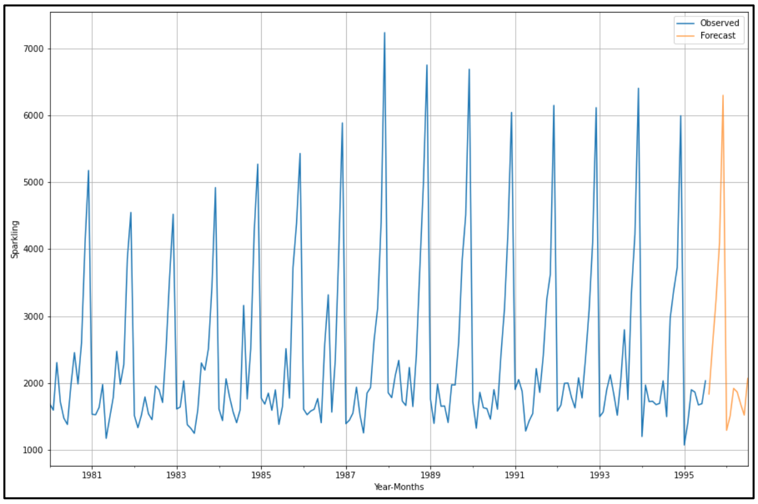
Sparkling dataset (Optimum model)

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                          Sparkling   No. Observations:                  187
Model:             SARIMAX(1, 1, 3)x(3, 0, 3, 4)   Log Likelihood               -1247.401
Date:                           Sun, 11 Dec 2022   AIC                           2516.803
Time:                                   22:56:15   BIC                           2551.297
Sample:                               01-01-1980   HQIC                          2530.800
                                    - 07-01-1995
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.3005      0.844     -0.356      0.722      -1.956       1.355
ma.L1          1.0418      2.655      0.392      0.695      -4.163       6.246
ma.L2         -2.5023      3.640     -0.688      0.492      -9.636       4.631
ma.L3          0.1619      1.046      0.155      0.877      -1.888       2.212
ar.S.L4        0.0002      0.011      0.017      0.986      -0.021       0.022
ar.S.L8       -0.0093      0.008     -1.207      0.227      -0.025       0.006
ar.S.L12       1.0119      0.009    111.612      0.000       0.994       1.030
ma.S.L4       -0.1467      0.427     -0.344      0.731      -0.983       0.690
ma.S.L8       -0.1550      0.352     -0.441      0.659      -0.844       0.534
ma.S.L12      -0.6849      0.298     -2.296      0.022      -1.269      -0.100
sigma2       2.17e+04   5.67e+04      0.383      0.702   -8.95e+04    1.33e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                60.63
Prob(Q):                              0.98   Prob(JB):                         0.00
Heteroskedasticity (H):               1.39   Skew:                             0.79
Prob(H) (two-sided):                  0.21   Kurtosis:                         5.46
===================================================================================
```

9.1. Summary

RMSE of the Full Model 525.16

Sparkling	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-01	1833.267312	360.917909	1125.881208	2540.653415
1995-09-01	2551.740052	366.467319	1833.477305	3270.002798
1995-10-01	3234.758932	366.780249	2515.882853	3953.635011
1995-11-01	4102.430636	368.956037	3379.290092	4825.571181
1995-12-01	6304.313240	369.225281	5580.644987	7027.981492
1996-01-01	1289.950498	370.131586	564.505919	2015.395076
1996-02-01	1503.641886	371.517527	775.480913	2231.802860
1996-03-01	1917.365597	372.667274	1186.951162	2647.780031
1996-04-01	1863.795064	373.437475	1131.871063	2595.719066
1996-05-01	1681.442981	373.946373	948.521558	2414.364404
1996-06-01	1518.082227	374.953827	783.186230	2252.978223
1996-07-01	2066.054172	375.716258	1329.663839	2802.444506



9.2. Model predictions

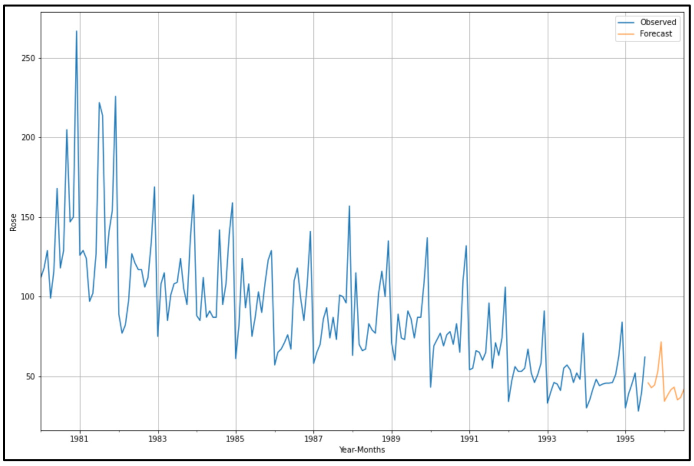
Rose dataset (Optimum model)

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                Rose   No. Observations:                 187
Model:             SARIMAX(2, 1, 3)x(3, 0, 3, 4)   Log Likelihood                -719.510
Date:                             Sun, 11 Dec 2022   AIC                           1463.020
Time:                                     23:02:27   BIC                           1500.649
Sample:                                 01-01-1980   HQIC                          1478.289
                                      - 07-01-1995
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.3153      0.035    -37.567      0.000      -1.384      -1.247
ar.L2         -0.8418      0.030    -27.991      0.000      -0.901      -0.783
ma.L1          0.5383      0.067      8.024      0.000       0.407       0.670
ma.L2         -0.3194      0.077     -4.129      0.000      -0.471      -0.168
ma.L3         -0.8715      0.062    -13.996      0.000      -0.994      -0.749
ar.S.L4        0.0308      0.020      1.539      0.124      -0.008       0.070
ar.S.L8       -0.0123      0.017     -0.714      0.475      -0.046       0.021
ar.S.L12       0.9212      0.016     56.539      0.000       0.889       0.953
ma.S.L4       -0.0434    220.681     -0.000      1.000    -432.571     432.484
ma.S.L8        0.0434    218.350      0.000      1.000    -427.914     428.001
ma.S.L12      -1.0001    315.548     -0.003      0.997    -619.463     617.463
sigma2       225.7910   7.13e+04      0.003      0.997   -1.39e+05     1.4e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.92   Jarque-Bera (JB):                61.91
Prob(Q):                              0.34   Prob(JB):                         0.00
Heteroskedasticity (H):               0.14   Skew:                             0.56
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.74
===================================================================================
```

9.3. Summary

RMSE of the Full Model 26.8

Rose	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-01	45.756424	15.524812	15.328351	76.184497
1995-09-01	42.689788	15.907865	11.510947	73.868630
1995-10-01	44.472443	15.951391	13.208290	75.736595
1995-11-01	53.323918	15.968752	22.025740	84.622096
1995-12-01	71.570373	16.268907	39.683901	103.456844
1996-01-01	34.172715	16.272239	2.279714	66.065717
1996-02-01	38.021251	16.401244	5.875403	70.167100
1996-03-01	41.433995	16.564480	8.968210	73.899780
1996-04-01	43.085396	16.593262	10.563200	75.607592
1996-05-01	35.028295	16.818313	2.065007	67.991584
1996-06-01	36.616906	16.874588	3.543321	69.690491
1996-07-01	41.675068	16.925436	8.501823	74.848313



9.4. Model predictions

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

Steps involved:

1. The dataset is read as a proper time series data and EDA is done in it.
2. Visualization of the data is performed for understanding the data distribution, sales over the years and monthly sales.
3. The data is decomposed to understand the seasonality and trend in it.
4. Data is split into train and test datasets; this train data is used to build different forecasting models and the model with the optimum RMSE value is selected for forecasting.
5. From the RMSE values the optimum model for forecasting the sparkling dataset are Alpha=0.25, Beta=0.0, Gamma=0.74: TES and Auto SARIMA models. For the Rose dataset it is 2pointTrailingMovingAvg and Alpha=0.25, Beta=0.0, Gamma=0.74: TES.
6. The optimum model is selected, and the predictions are made for the next 12 months using it.

INSIGHTS:

SPARKLING WINE: The sparkling wine sales is high during the December and November months; the company can provide offers during these months and increase the production of this type of wine for sale during this period.

- The sale of the sparkling wine will follow the same seasonality pattern as previous years.
- The sparkling wine is forecasted to sell over 6000 units.
- Sparkling wine sales is comparatively higher than rose wine.

ROSE WINE: The rose wine sales is high during the December and November months; the company can provide offers during these months. The sale of this type of wine is declining over the years and the company have to perform marker surveys to decide upon the demand for this wine for next year production.

- The sale of the rose wine will follow the same seasonality pattern as previous years.
- The sparkling wine is forecasted to sell between 50 to 100 units.
- The rose wine is losing its popularity among people. The company may look for producing a different type of wine.
- The introduction of a new wine type may boost sales.