Time Series Forecasting

Wine Sales

ANIL KUMAR PGBPDSBA AUG'2023

Contents

Time Series and Forecasting Project

1.0 Problem Statement – Wine Sales

Context:

As an analyst at ABC Estate Wines, we are presented with historical data encompassing the sales of different types of wines throughout the 20th century. These datasets originate from the same company but represent sales figures for distinct wine varieties. Our objective is to delve into the data, analyze trends, patterns, and factors influencing wine sales over the course of the century. By leveraging data analytics and forecasting techniques, we aim to gain actionable insights that can inform strategic decision-making and optimize sales strategies for the future.

Objective:

The primary objective of this project is to analyze and forecast wine sales trends for the 20th century based on historical data provided by ABC Estate Wines. We aim to equip ABC Estate Wines with the necessary insights and foresight to enhance sales performance, capitalize on emerging market opportunities, and maintain a competitive edge in the wine industry.

Data Dictionary

Rose.CSV

YearMonth: Combination of Year and month
Rose: Sales of Rose (Brand of wine)
Sparkling.csv

YearMonth: Combination of Year and month
Rose: Sales of Sparkling(Brand of wine)

1.1 Read Data and Plot Data

```
dfr = pd.read_csv('/content/drive/MyDrive/Python Course/Rose.csv') ##  Fill the blank to read the data
```

```
[4] dfs = pd.read_csv('/content/drive/MyDrive/Python Course/Sparkling.csv') ##  Fill the blank to read the data
```

- Displaying few rows of the dataset
  - Rose Wine

	YearMonth	Rose
0	1980-01	112.0
1	1980-02	118.0
2	1980-03	129.0
3	1980-04	99.0
4	1980-05	116.0

  - Sparkling Wine

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

- Cl

- Checking the data types of the columns for the dataset(Rose)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 187 entries, 0 to 186
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   YearMonth  187 non-null    object 
 1   Rose       185 non-null    float64
dtypes: float64(1), object(1)
memory usage: 3.0+ KB
```

- Checking the data types of the columns for the dataset(Sparkling)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 187 entries, 0 to 186
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   YearMonth  187 non-null    object
 1   Sparkling  187 non-null    int64 
dtypes: int64(1), object(1)
memory usage: 3.0+ KB
```

- Statistical summary of the Rose dataset

	count	mean	std	min	25%	50%	75%	max
Rose	185.0	90.394595	39.175344	28.0	63.0	86.0	112.0	267.0

- Statistical summary of the Sparkling dataset

	count	mean	std	min	25%	50%	75%	max
Sparkling	187.0	2402.417112	1295.11154	1070.0	1605.0	1874.0	2549.0	7242.0

- Checking the Null Values for Rose Dataset

We observed that there are two missing values present

```
YearMonth    0
Rose         2
dtype: int64
```

- Checking the Null Values for Sparkling Dataset

There are no null values for sparkling

```
YearMonth    0
Sparkling    0
dtype: int64
```

- Treatment of Null Values in Rose Dataset

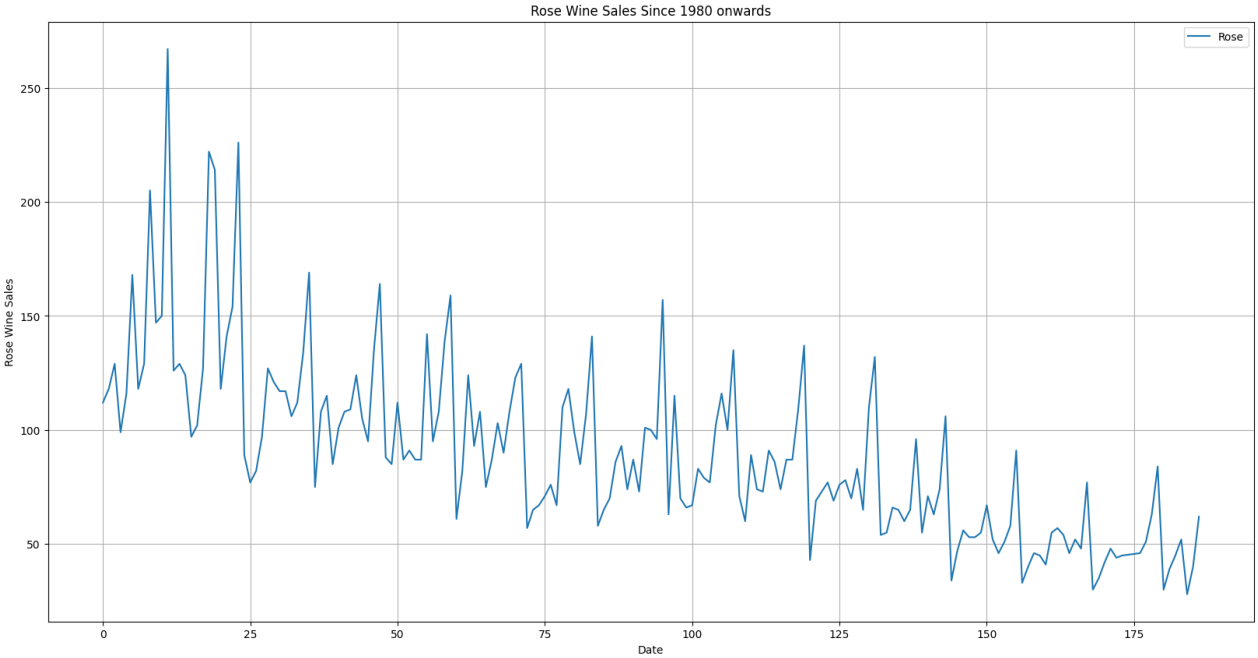
```
[18] # Impute Missing Values
     dfr=dfr.interpolate()
```

- Check after imputing missing values

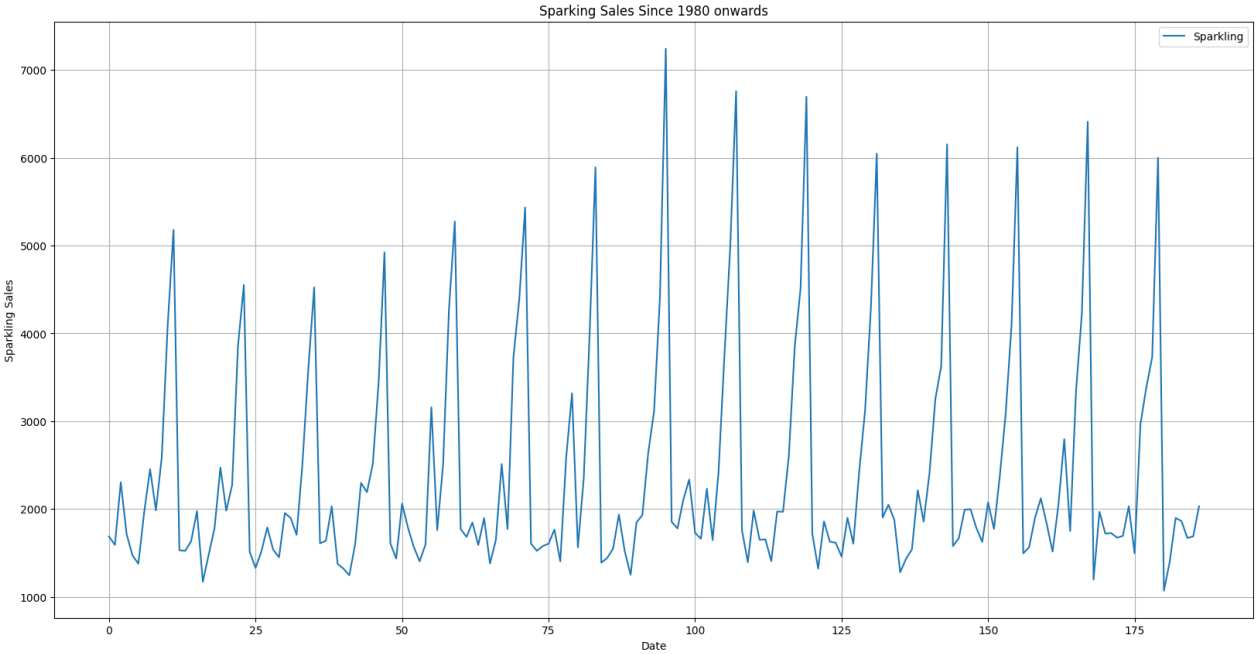
```
YearMonth    0
Rose         0
dtype: int64
```

No Missing values Found

Plot Data – Rose Dataset



Plot Data – Sparkling Dataset



Making Timestamp data by setting to index column

```
DatetimeIndex(['1980-01-31', '1980-02-29', '1980-03-31', '1980-04-30',
               '1980-05-31', '1980-06-30', '1980-07-31', '1980-08-31',
               '1980-09-30', '1980-10-31',
               ...
               '1994-10-31', '1994-11-30', '1994-12-31', '1995-01-31',
               '1995-02-28', '1995-03-31', '1995-04-30', '1995-05-31',
               '1995-06-30', '1995-07-31'],
              dtype='datetime64[ns]', length=187, freq='M')
```

Replace yearmonth by date

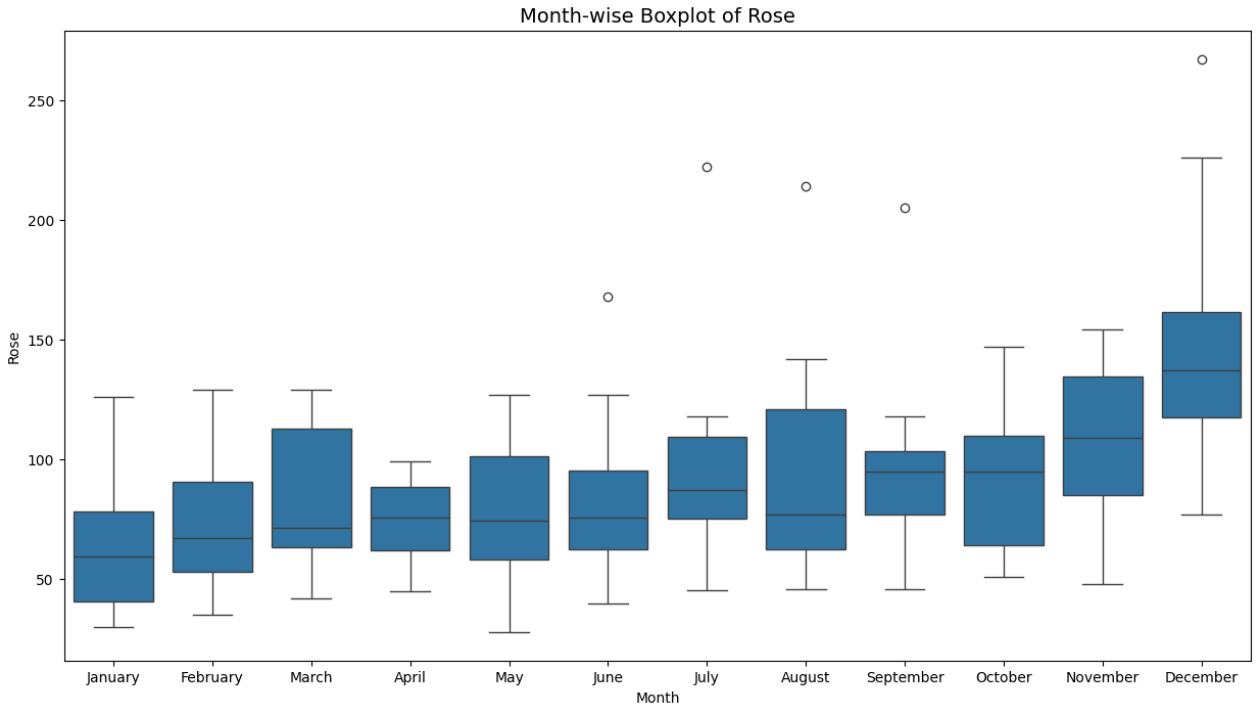
```
dfr['Month']=date
dfr.drop("YearMonth",axis=1,inplace=True)
dfr=dfr.set_index('Month')
dfr.head()
```

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

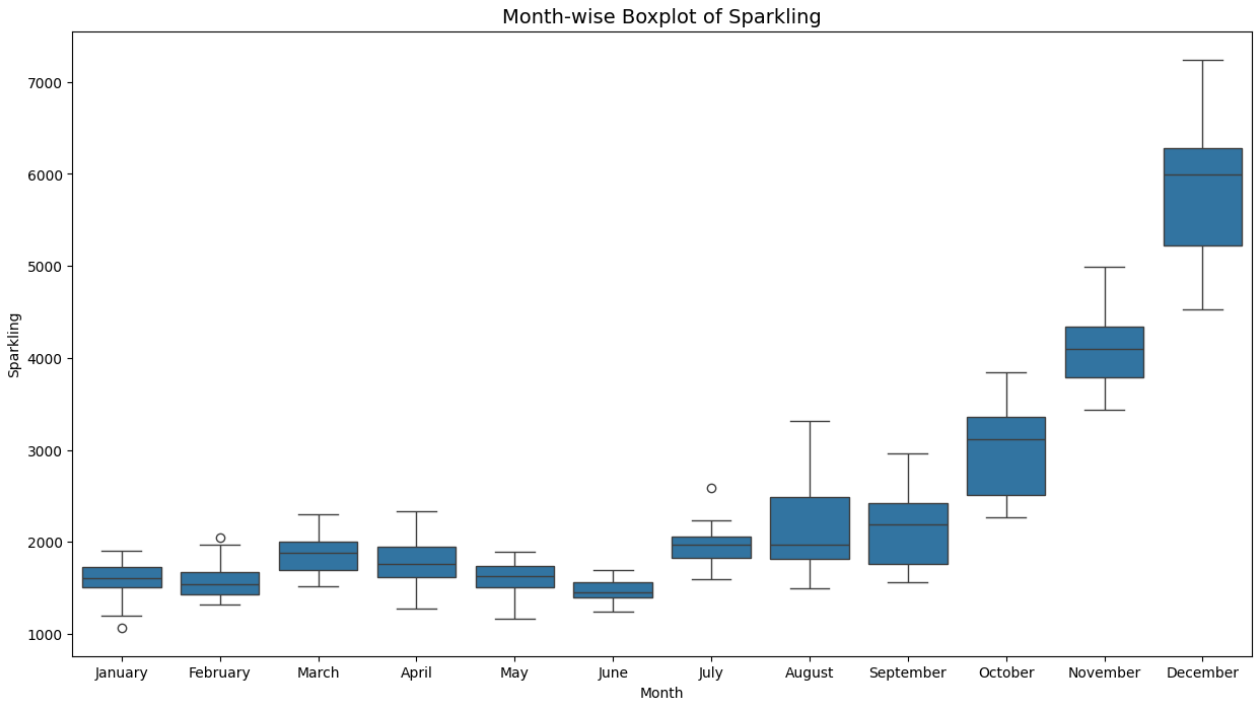
	Sparkling
Month	
1980-01-31	1686
1980-02-29	1591
1980-03-31	2304
1980-04-30	1712
1980-05-31	1471

EDA- Exploratory Data Analysis and Decomposition

Month-wise sales of Rose

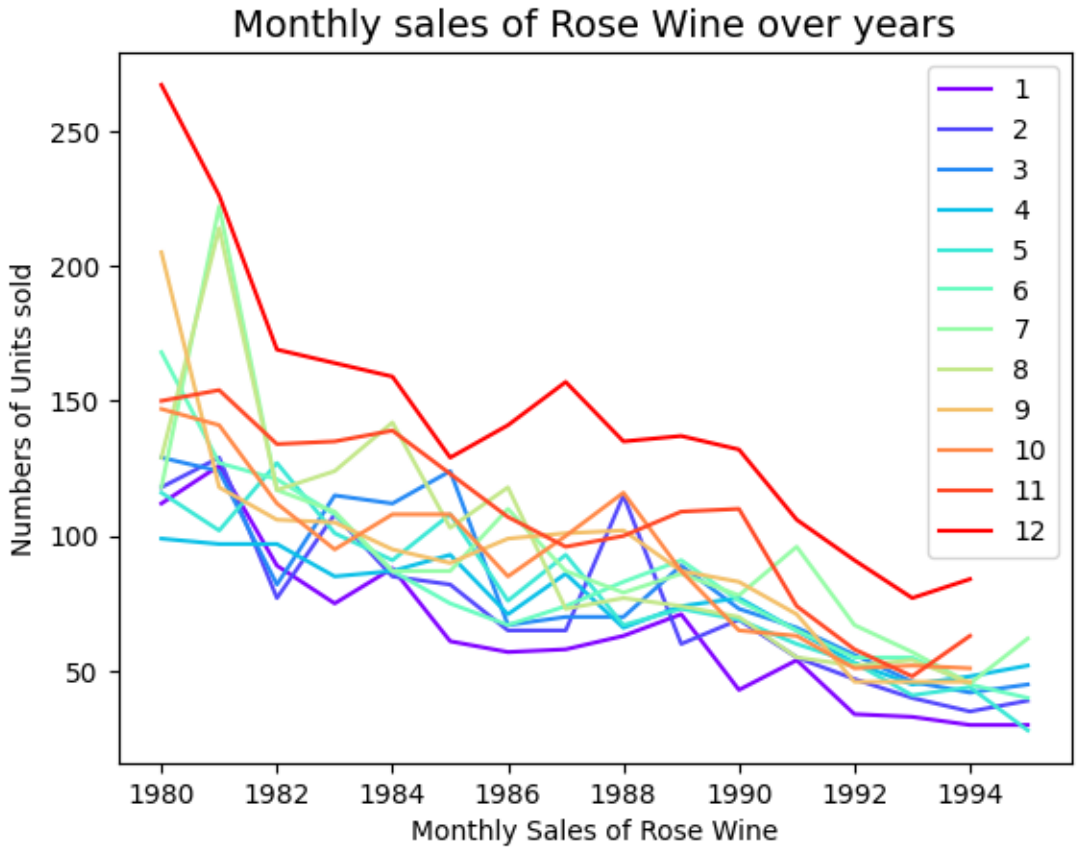


Month-wise sales of Sparkling



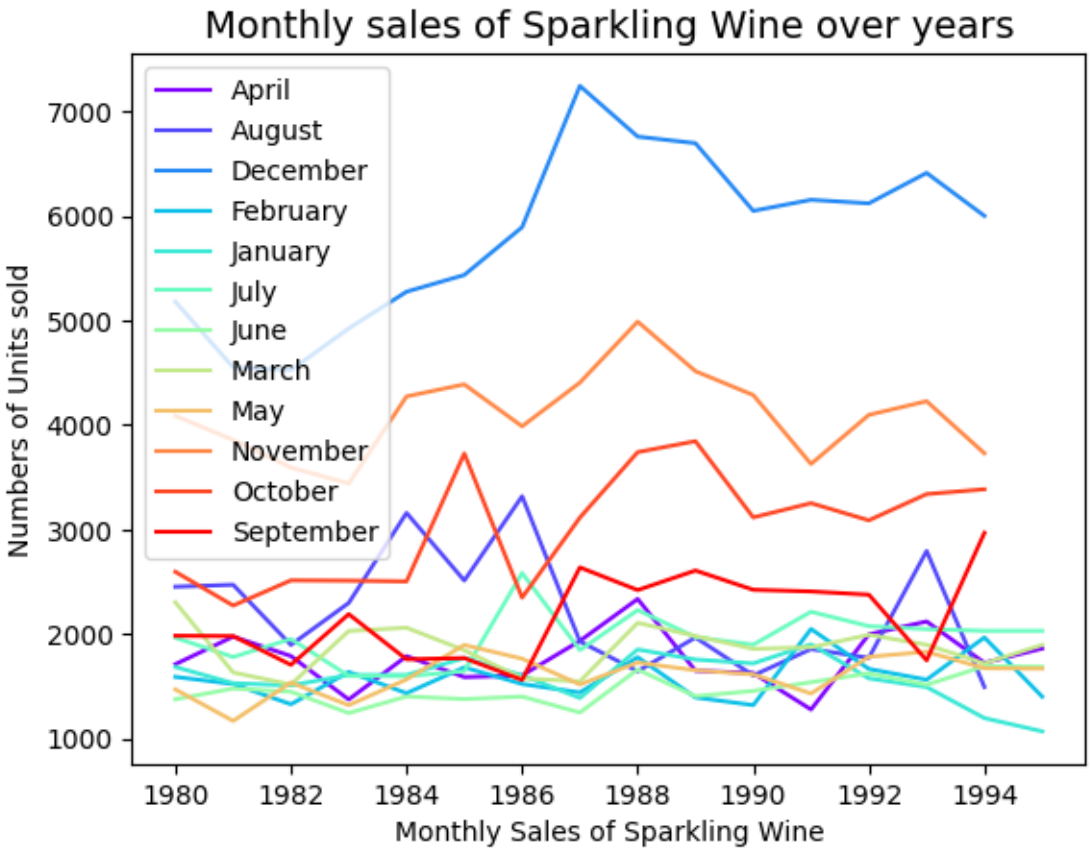
Month-Year wise sales of Rose

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



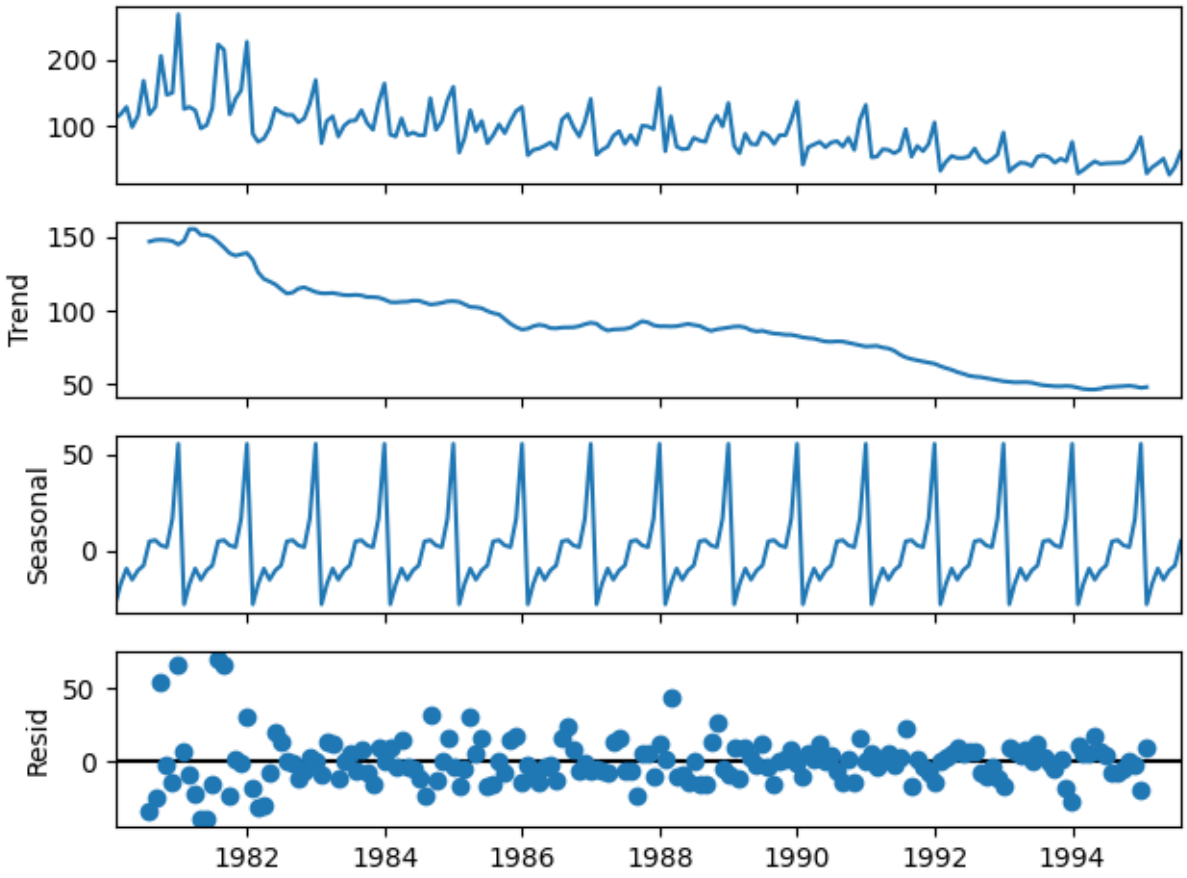
Month-Year wise sales of Sparkling

Month	April	August	December	February	January	July	June	March	May	November	October	September
Month												
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



Decomposition

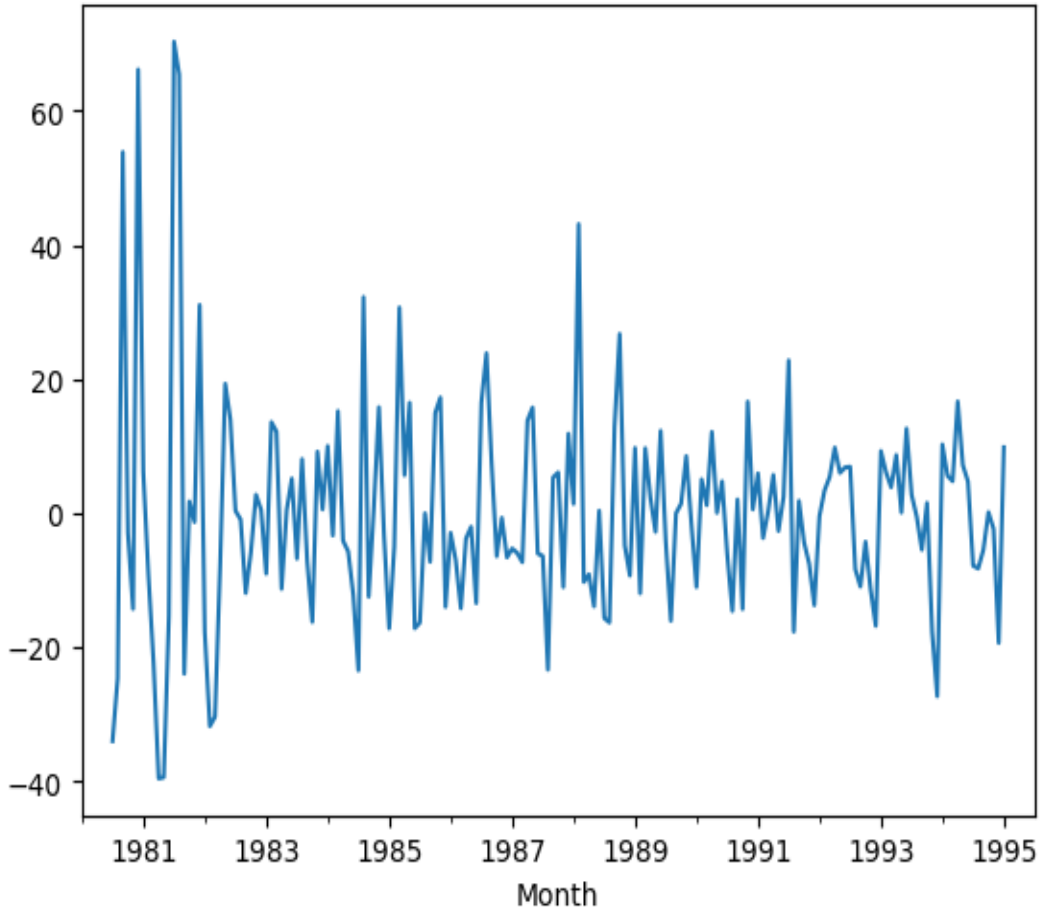
➢ Additive Decomposition of Rose



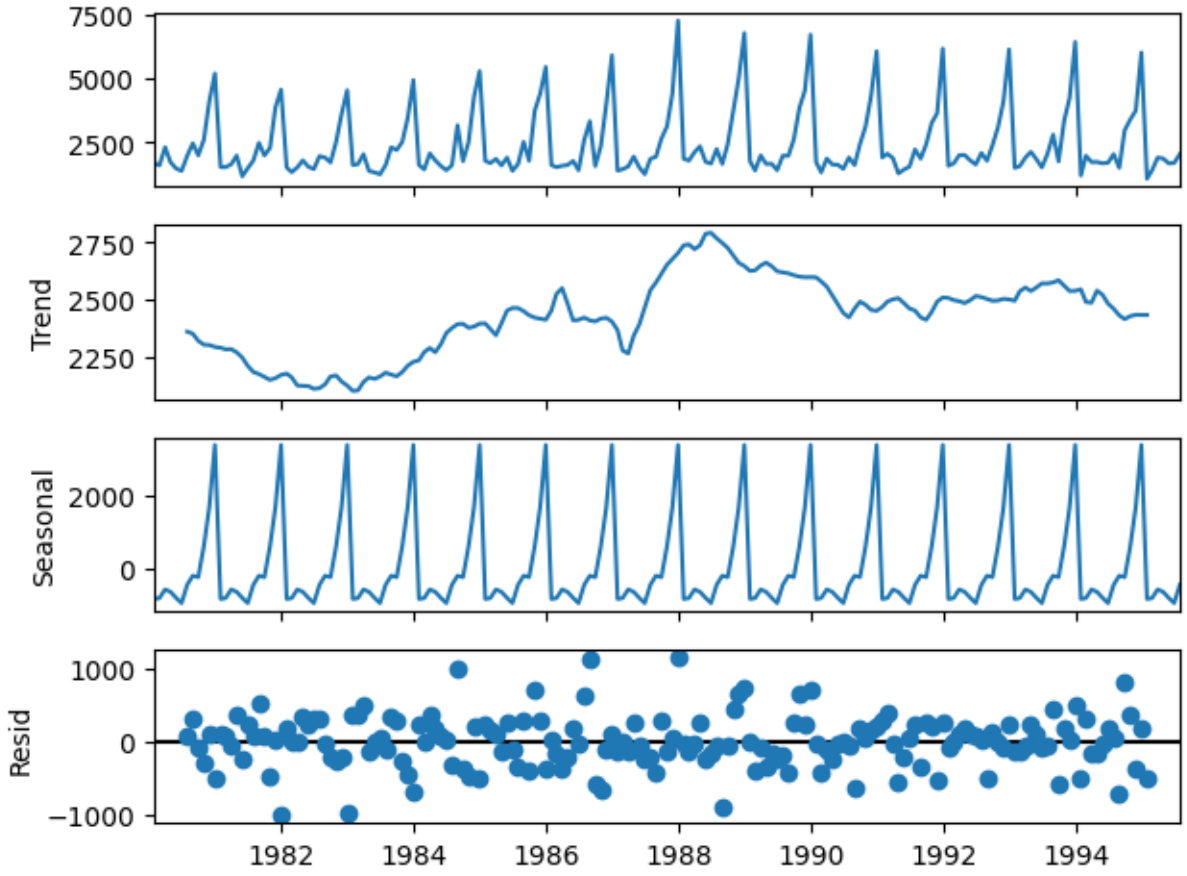

```
Trend
 Month
1980-01-31           NaN
1980-02-29           NaN
1980-03-31           NaN
1980-04-30           NaN
1980-05-31           NaN
1980-06-30           NaN
1980-07-31    147.083333
1980-08-31    148.125000
1980-09-30    148.375000
1980-10-31    148.083333
1980-11-30    147.416667
1980-12-31    145.125000
```

```
Name: trend, dtype: float64


Seasonality
 Month
1980-01-31   -27.908647
1980-02-29   -17.435632
1980-03-31    -9.285830
1980-04-30   -15.098330
1980-05-31   -10.196544
1980-06-30    -7.678687
1980-07-31     4.896908
1980-08-31     5.499686
1980-09-30     2.774686
1980-10-31     1.871908
1980-11-30    16.846908
1980-12-31    55.713575
Name: seasonal, dtype: float64


Residual
 Month
1980-01-31          NaN
1980-02-29          NaN
1980-03-31          NaN
1980-04-30          NaN
1980-05-31          NaN
1980-06-30          NaN
1980-07-31   -33.980241
1980-08-31   -24.624686
1980-09-30    53.850314
1980-10-31    -2.955241
1980-11-30   -14.263575
1980-12-31    66.161425
    Name: resid, dtype: float64
```

➢ Residual



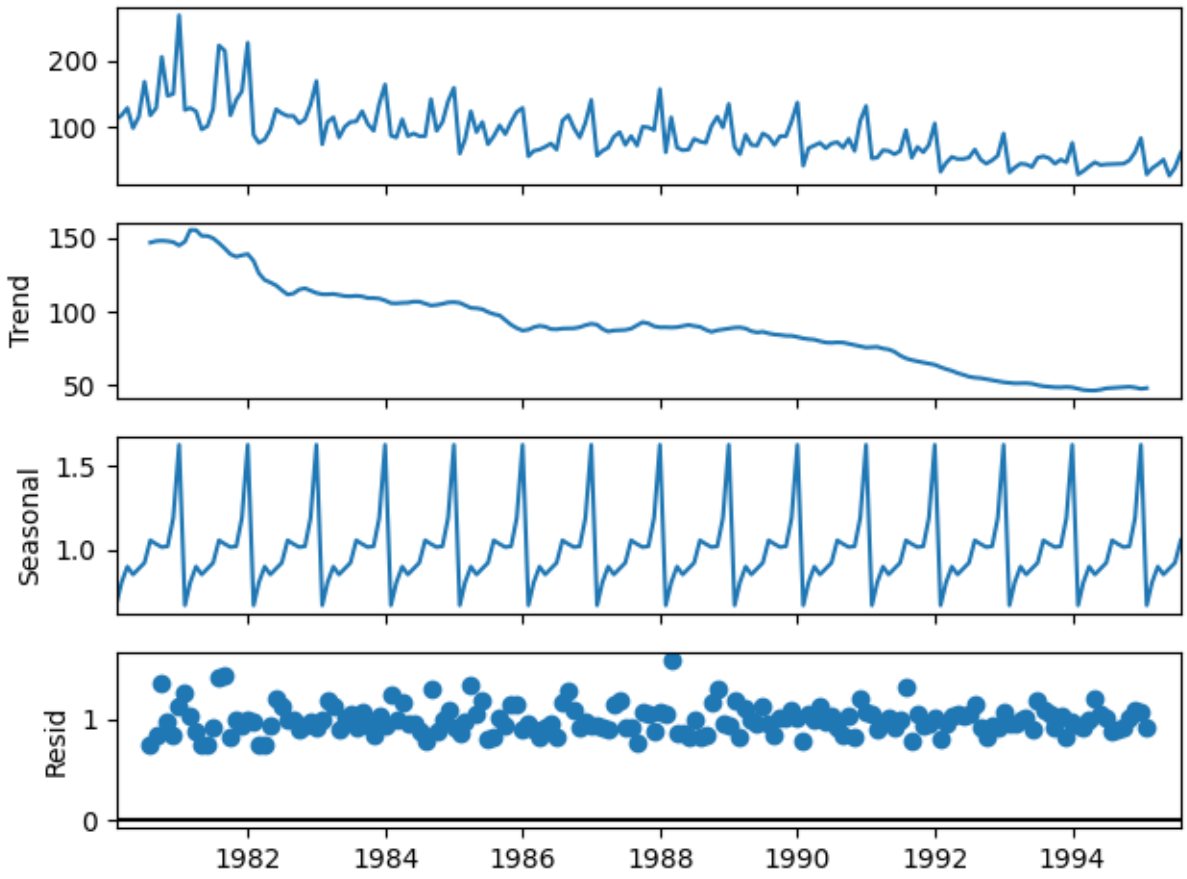
➢ Additive Decomposition of Sparkling




```
Trend
 Month
1980-01-31             NaN
1980-02-29             NaN
1980-03-31             NaN
1980-04-30             NaN
1980-05-31             NaN
1980-06-30             NaN
1980-07-31     2360.666667
1980-08-31     2351.333333
1980-09-30     2320.541667
1980-10-31     2303.583333
1980-11-30     2302.041667
1980-12-31     2293.791667
Name: trend, dtype: float64
```

```
Seasonality
 Month
1980-01-31    -854.260599
1980-02-29    -830.350678
1980-03-31    -592.356630
1980-04-30    -658.490559
1980-05-31    -824.416154
1980-06-30    -967.434011
1980-07-31    -465.502265
1980-08-31    -214.332821
1980-09-30    -254.677265
1980-10-31     599.769957
1980-11-30    1675.067179
1980-12-31    3386.983846
Name: seasonal, dtype: float64

Residual
 Month
1980-01-31           NaN
1980-02-29           NaN
1980-03-31           NaN
1980-04-30           NaN
1980-05-31           NaN
1980-06-30           NaN
1980-07-31     70.835599
1980-08-31    315.999487
1980-09-30    -81.864401
1980-10-31   -307.353290
1980-11-30    109.891154
1980-12-31   -501.775513
Name: resid, dtype: float64
```

➢ Multiplicative Decomposition of Rose

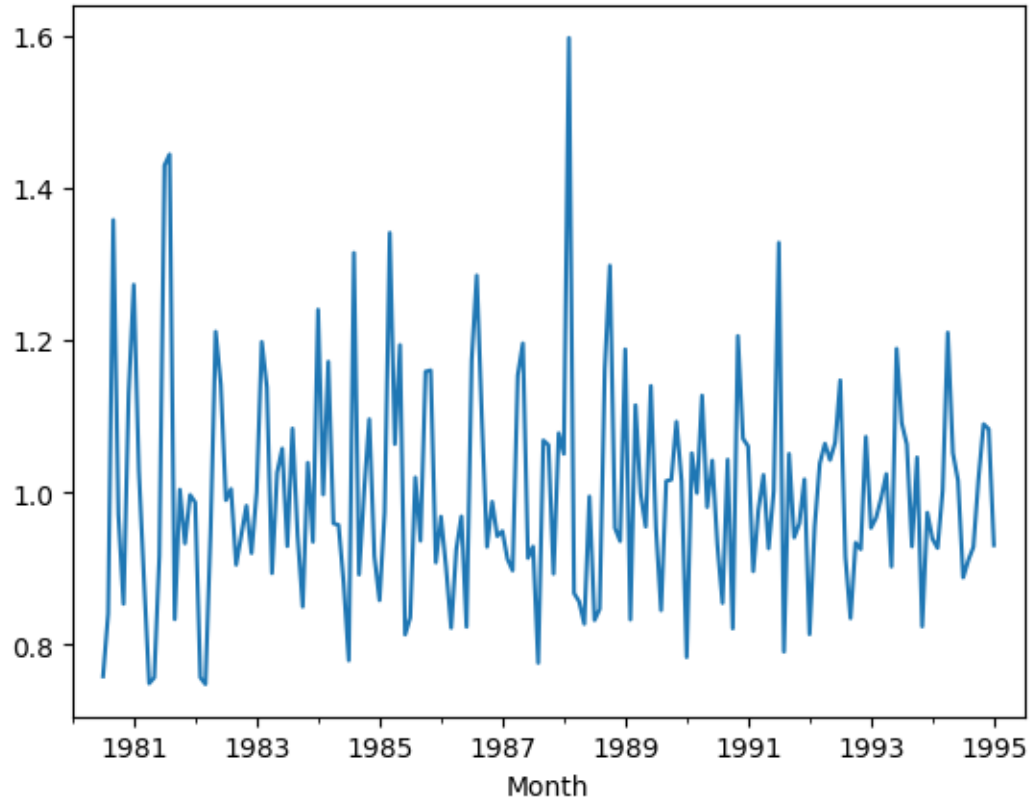


Trend & Seasonality

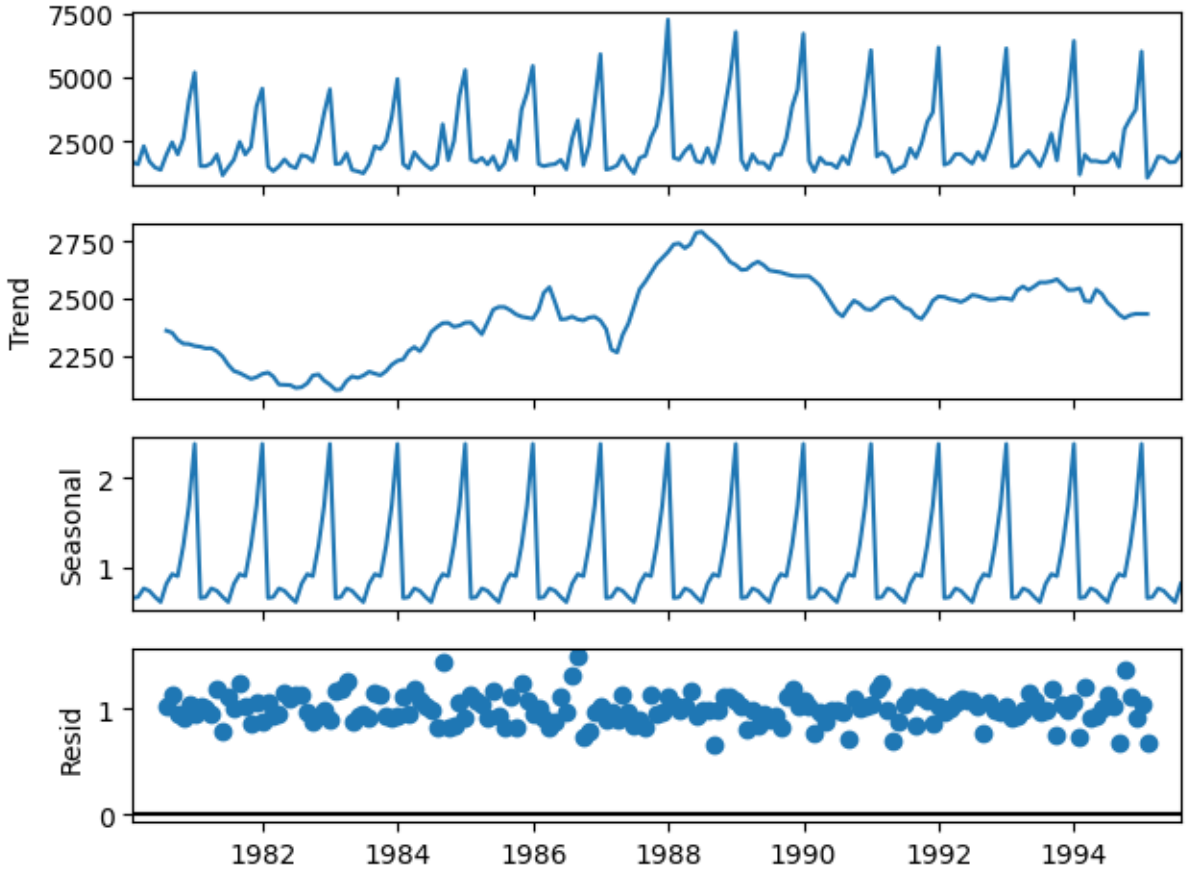
```
Trend
 Month
1980-01-31           NaN
1980-02-29           NaN
1980-03-31           NaN
1980-04-30           NaN
1980-05-31           NaN
1980-06-30           NaN
1980-07-31    147.083333
1980-08-31    148.125000
1980-09-30    148.375000
1980-10-31    148.083333
1980-11-30    147.416667
```

```
1980-12-31    145.125000
Name: trend, dtype: float64

Seasonality
 Month
1980-01-31    0.670111
1980-02-29    0.806163
1980-03-31    0.901164
1980-04-30    0.854024
1980-05-31    0.889415
1980-06-30    0.923985
1980-07-31    1.058038
1980-08-31    1.035881
1980-09-30    1.017648
1980-10-31    1.022573
1980-11-30    1.192349
1980-12-31    1.628646
Name: seasonal, dtype: float64
Residual
 Month
1980-01-31         NaN
1980-02-29         NaN
1980-03-31         NaN
1980-04-30         NaN
1980-05-31         NaN
1980-06-30         NaN
1980-07-31    0.758258
1980-08-31    0.840720
1980-09-30    1.357674
1980-10-31    0.970771
1980-11-30    0.853378
1980-12-31    1.129646
      Name: resid, dtype: float64
```

Residual



➢ Multiplicative Decomposition of Sparkling



Trend, seasonality and Residual

```
Trend

 Month
1980-01-31           NaN
1980-02-29           NaN
1980-03-31           NaN
1980-04-30           NaN
1980-05-31           NaN
1980-06-30           NaN
1980-07-31    2360.666667
1980-08-31    2351.333333
1980-09-30    2320.541667
1980-10-31    2303.583333
1980-11-30    2302.041667
1980-12-31    2293.791667
```

```
Name: trend, dtype: float64


Seasonality
 Month
1980-01-31    0.649843
1980-02-29    0.659214
1980-03-31    0.757440
1980-04-30    0.730351
1980-05-31    0.660609
1980-06-30    0.603468
1980-07-31    0.809164
1980-08-31    0.918822
1980-09-30    0.894367
1980-10-31    1.241789
1980-11-30    1.690158
1980-12-31    2.384776
Name: seasonal, dtype: float64


Residual
 Month
1980-01-31         NaN
1980-02-29         NaN
1980-03-31         NaN
1980-04-30         NaN
1980-05-31         NaN
1980-06-30         NaN
1980-07-31    1.029230
1980-08-31    1.135407
1980-09-30    0.955954
1980-10-31    0.907513
1980-11-30    1.050423
1980-12-31    0.946770
      Name: resid, dtype: float64
```

1.2 Data Preprocessing

- Train-test split

```
# ROSE DATA SPLIT
rtrain = dfr[dfr.index<'1991']
rtest = dfr[dfr.index>='1991']
```

```
# SPARKLING DATA SPLIT
strain = dfs[dfs.index<'1991']
stest = dfs[dfs.index>='1991']
```

```
print(rtrain.shape)
print(rtest.shape)
```

```
(132, 1)
(55, 1)
```

```
print(strain.shape)
print(stest.shape)
```

```
(132, 1)
(55, 1)
```

- First/last few rows of Train dataset after split - Rose

First few rows of Rose Training Data

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

Last few rows of Rose Training Data

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

- First/last few rows of Test dataset after split - Rose

First few rows of Rose Test Data

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

Last few rows of Rose Test Data

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

➢ First/last few rows of Train dataset after split – Sparkling

First few rows of Sparkling Training Data

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

Last few rows of Sparkling Training Data

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

➢ First/last few rows of Test dataset after split – Sparkling

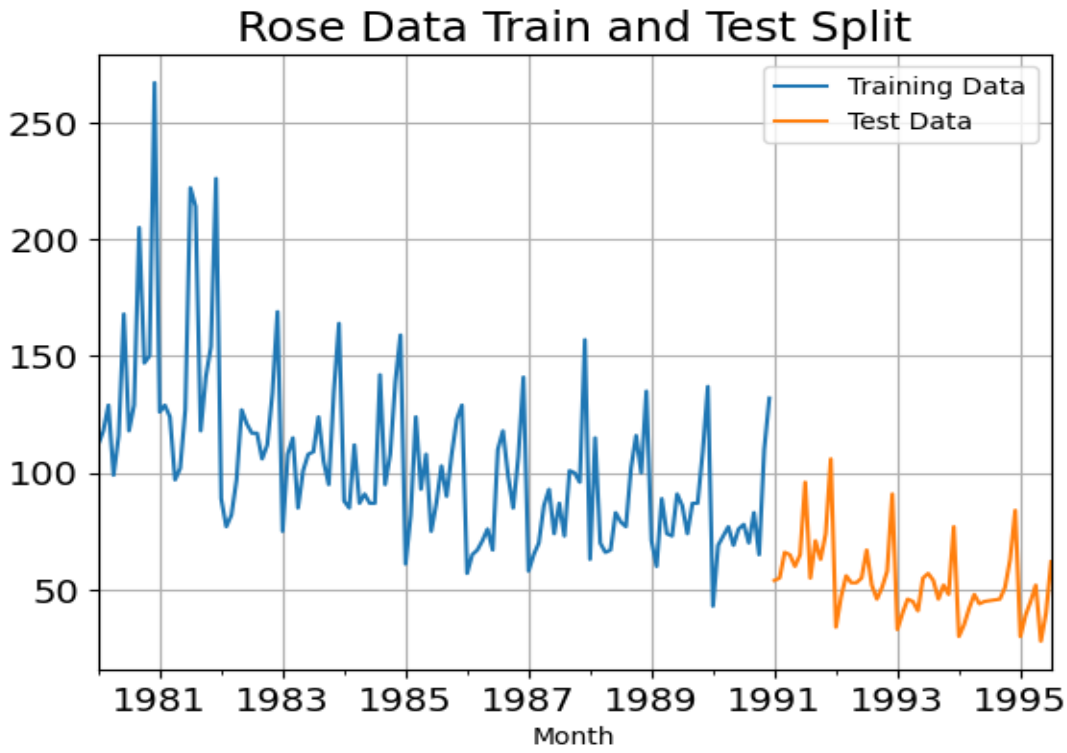
First few rows of Sparkling Test Data

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

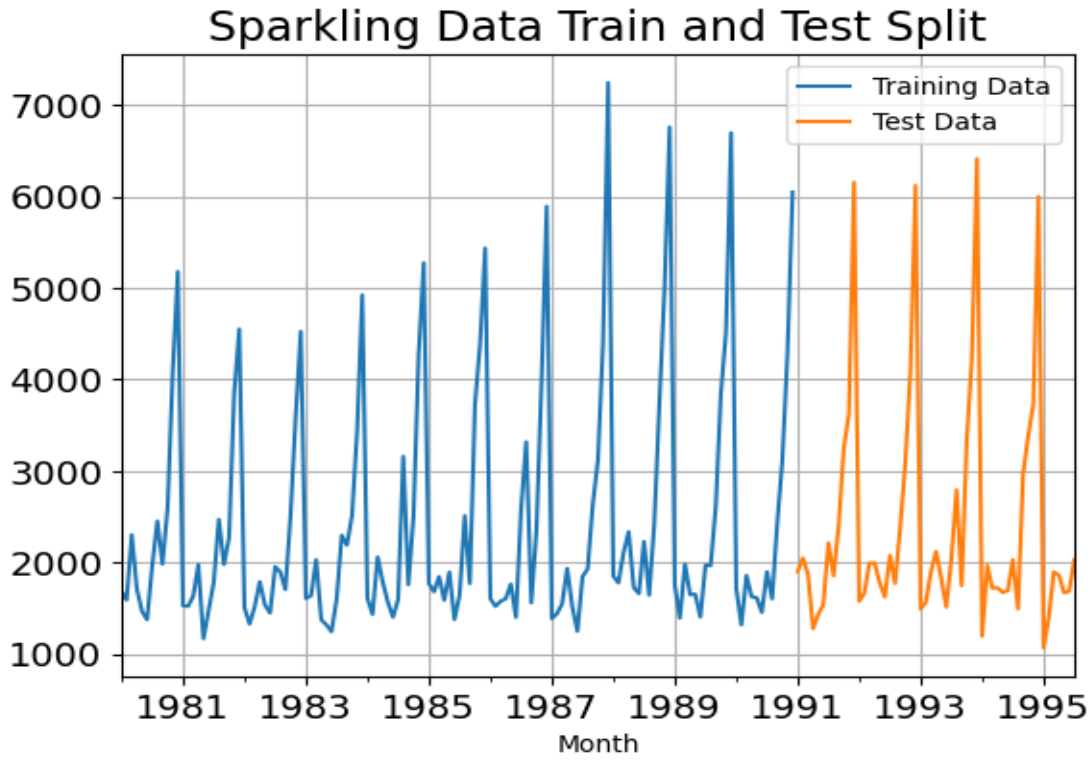
Last few rows of Sparkling Test Data

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

- Rose Data Train and Test Split



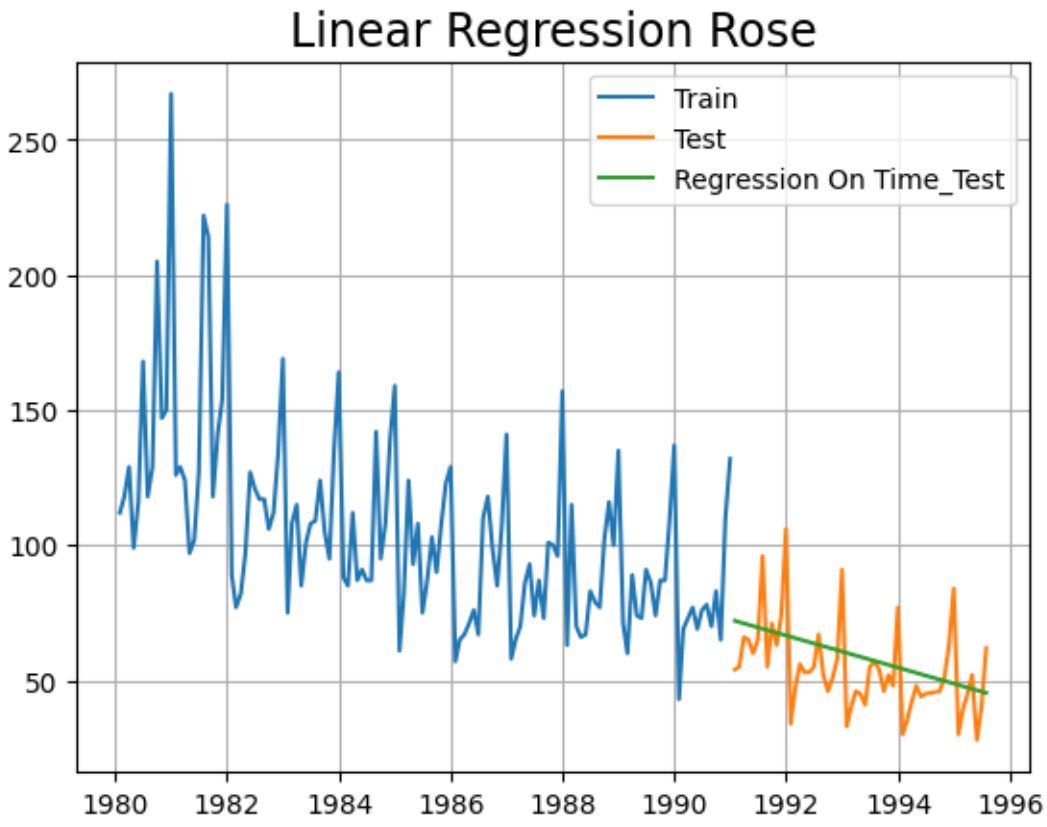
- Sparkling Data Train and Test Split



1.3 Model Building

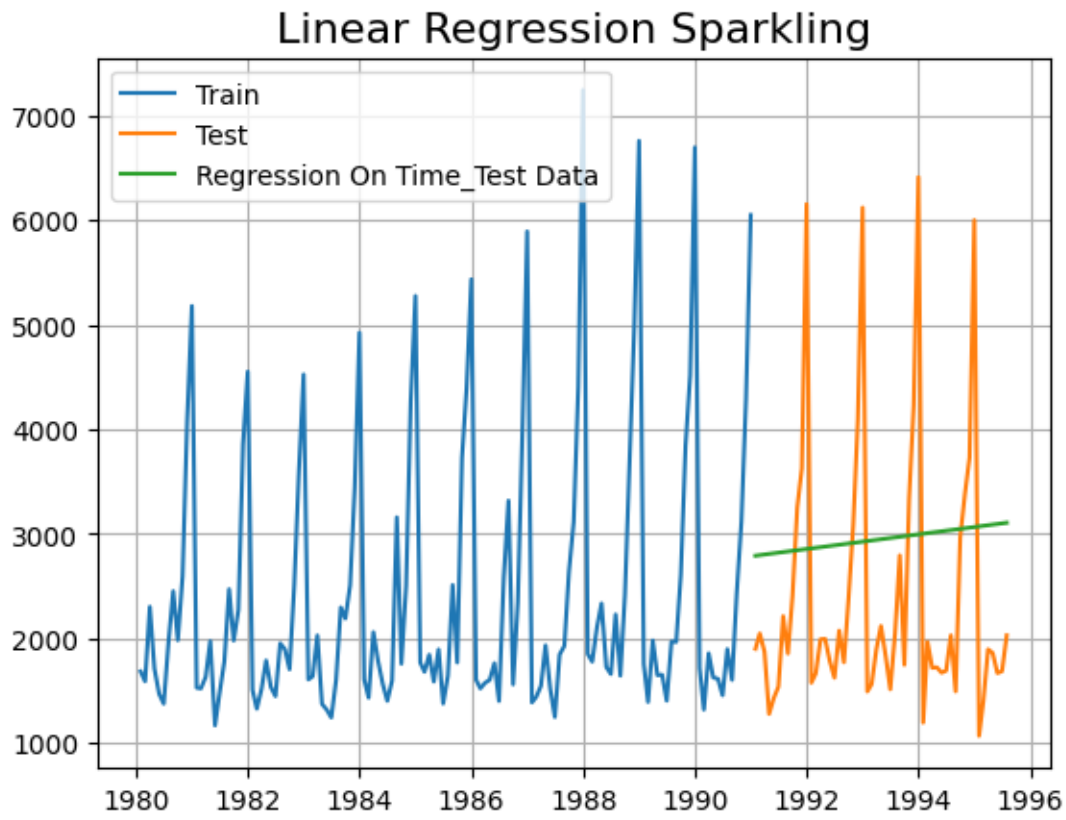
➢ 1.3.1 Model 1 Linear Regression on Rose

```
LinearRegression
LinearRegression()
```



	Test RMSE Rose
RegressionOnTime	15.268955

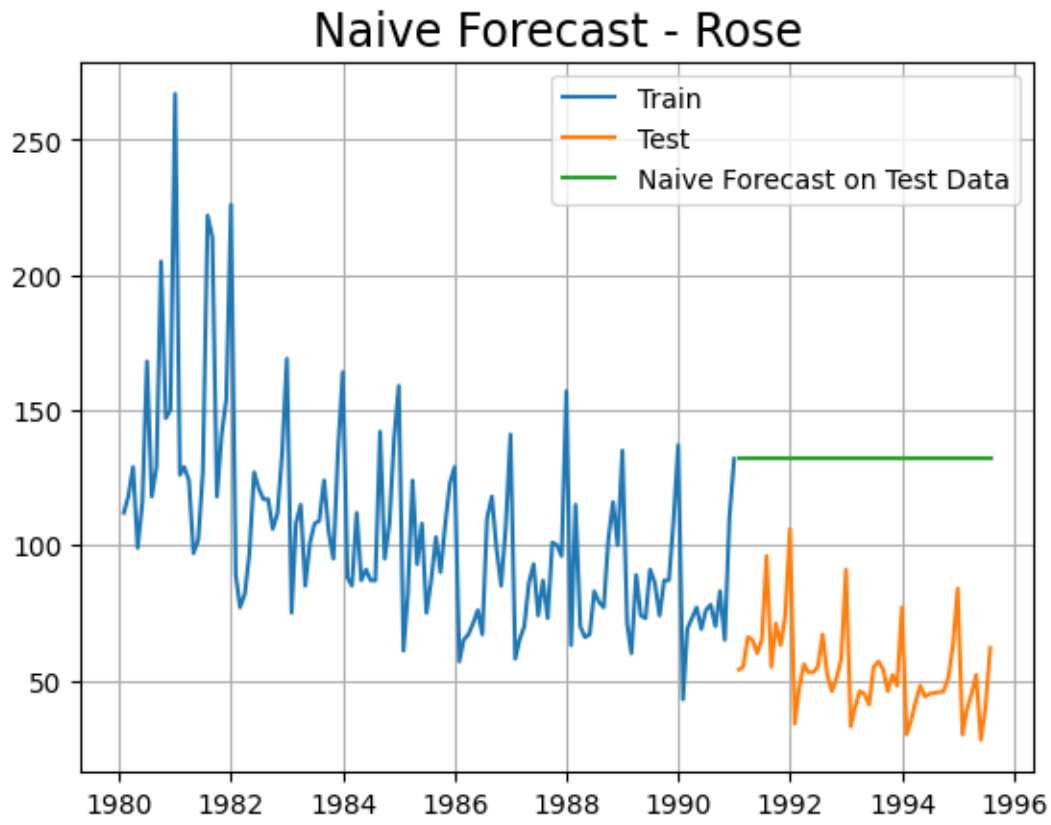
1.3.2 Model 1 Linear Regression on Sparkling



[1]

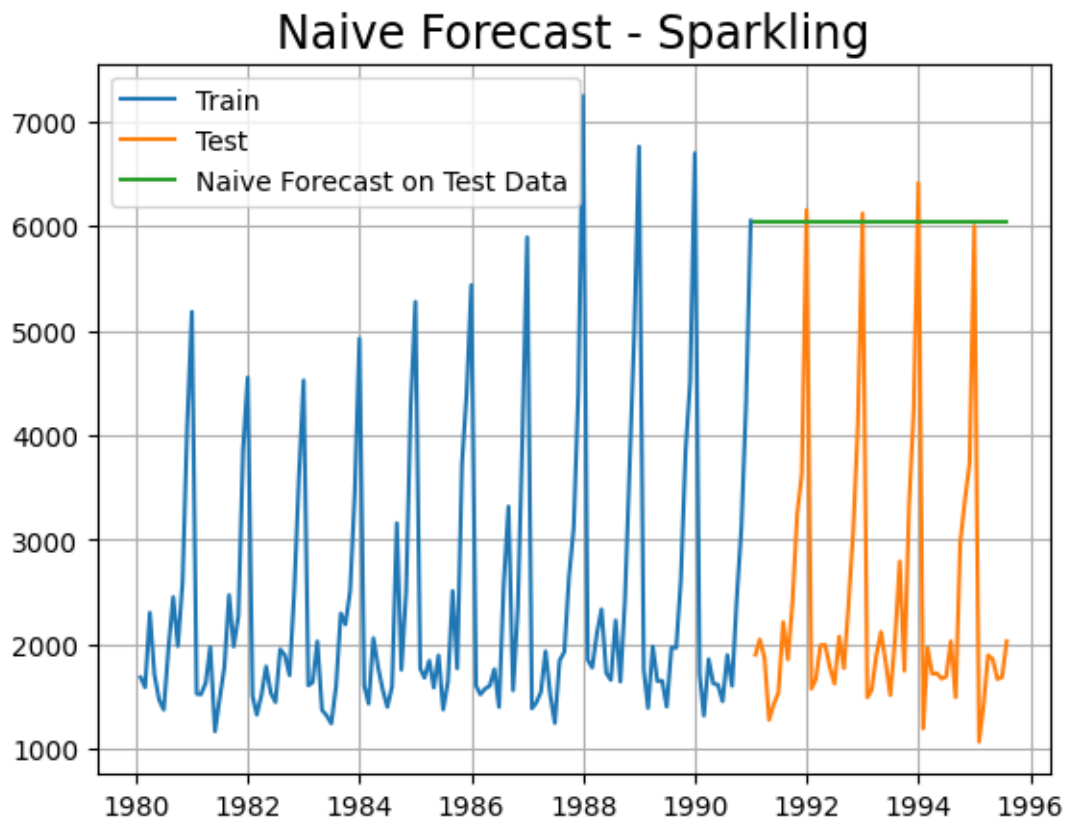
	Test RMSE Rose	Test RMSE Sparkling
RegressionOnTime	15.268955	1389.135175

1.3.3 Model 2 Naïve - Rose



	Test RMSE Rose	Test RMSE Sparkling
RegressionOnTime	15.268955	1389.135175

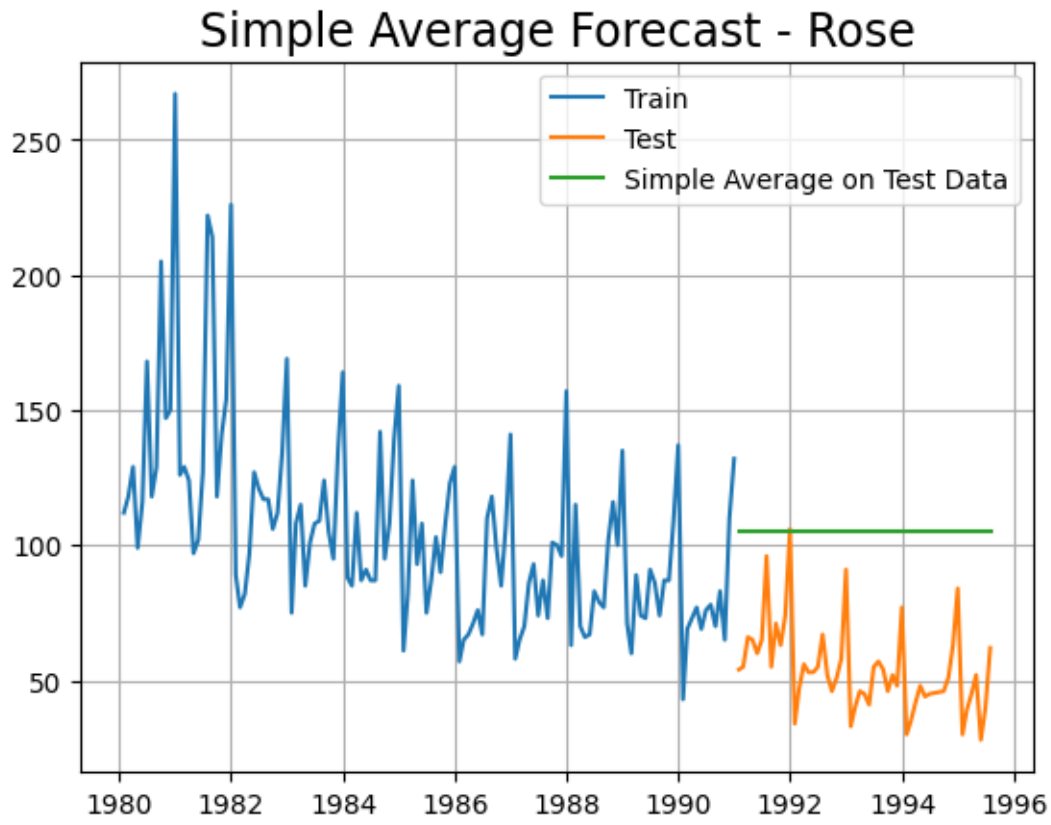
1.3.4 Model 2 Naïve – Sparkling



	Test RMSE Rose	Test RMSE Sparkling
NaiveModel	79.718773	3864.279352

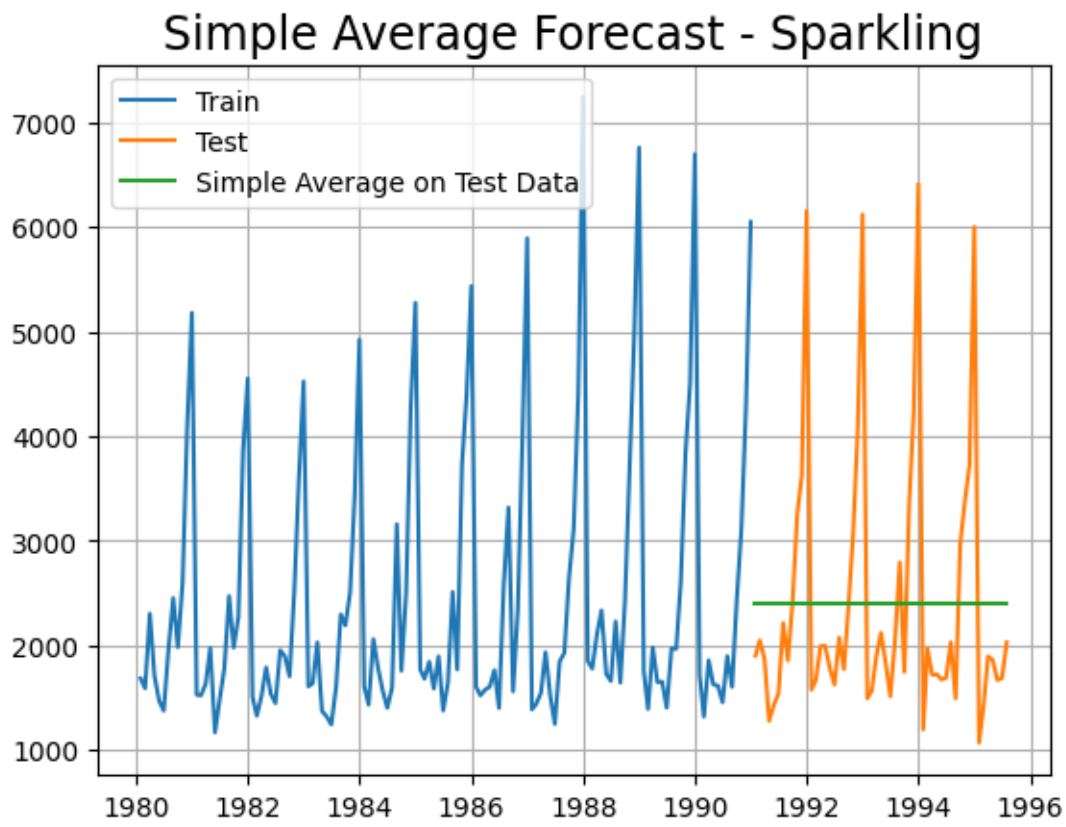
	Test RMSE Rose	Test RMSE Sparkling
RegressionOnTime	15.268955	1389.135175
NaiveModel	79.718773	3864.279352

1.3.5 Model 3 Simple Average – Rose



	Test RMSE Rose
SimpleAverageModel	53.46057

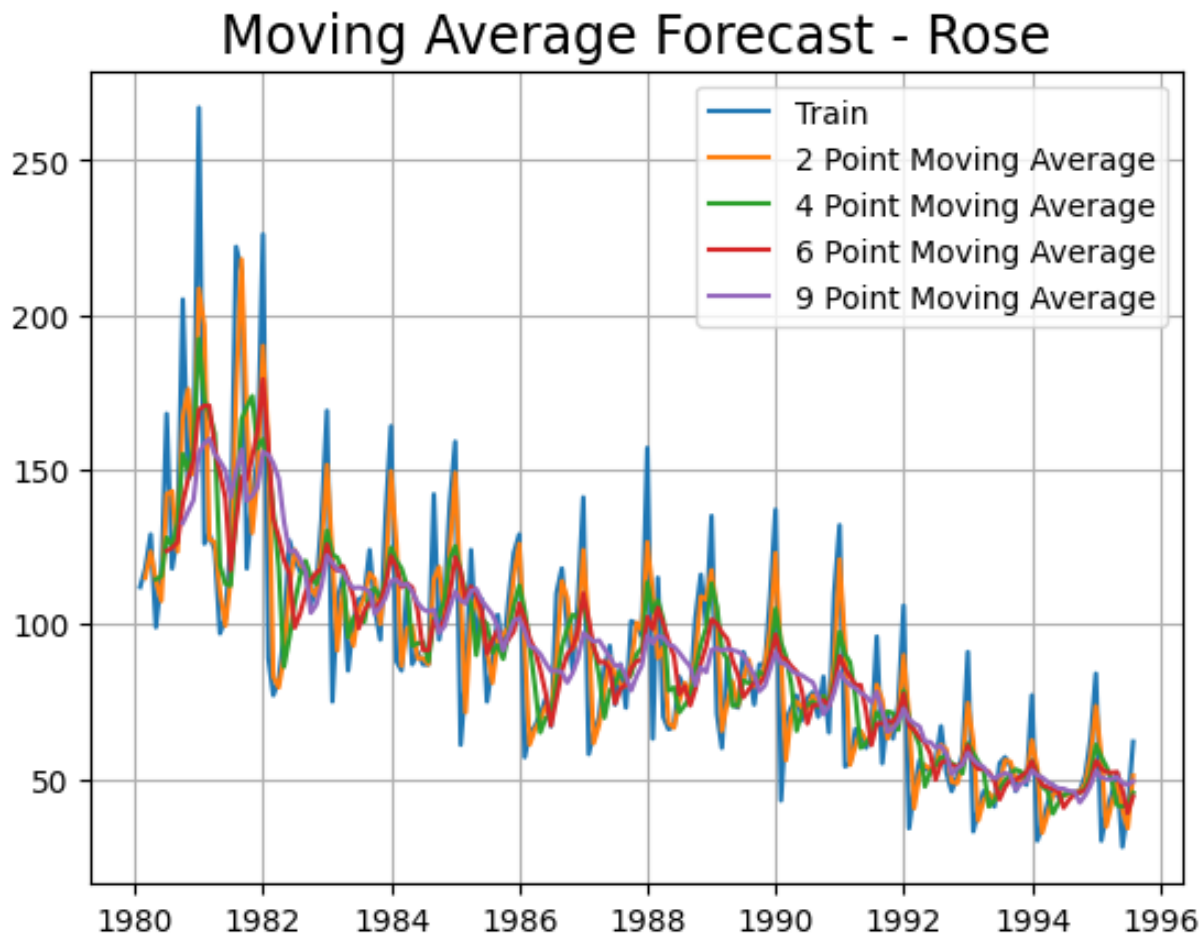
1.3.6 Model 3 Simple Average – Sparkling



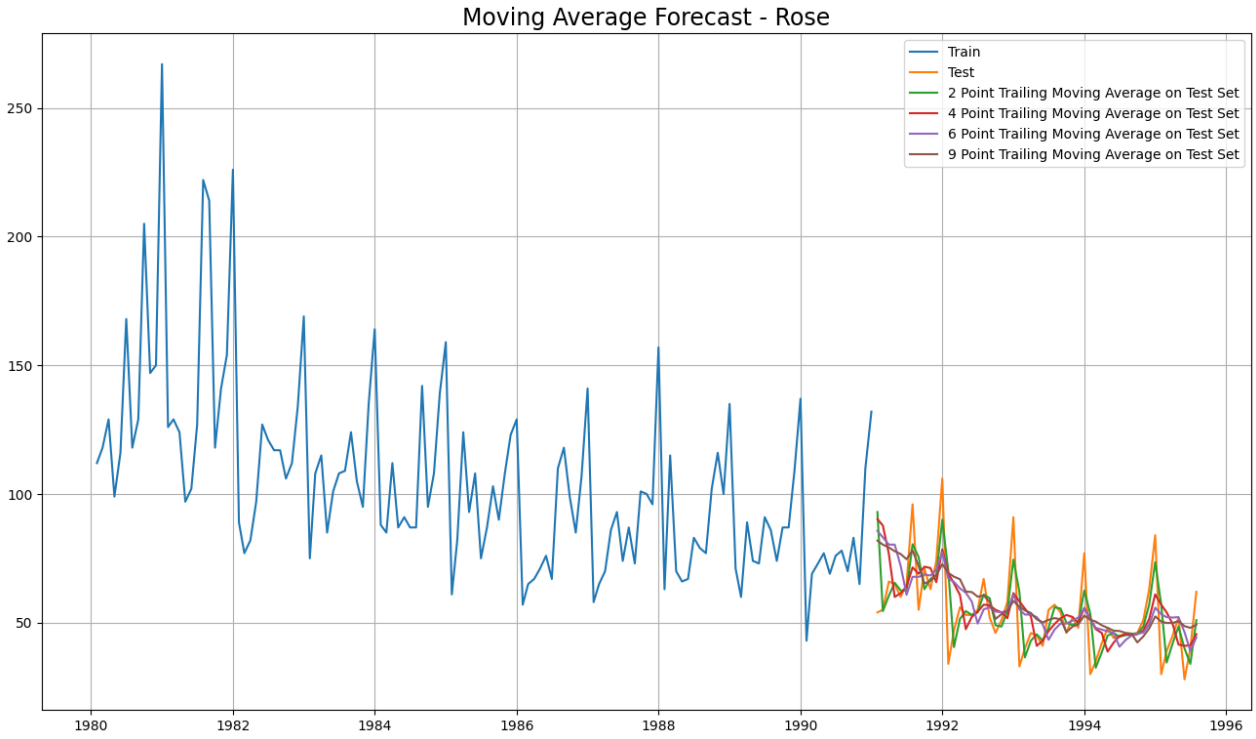
	Test RMSE Sparkling
SimpleAverageModel	1275.081804

	Test RMSE Rose	Test RMSE Sparkling
RegressionOnTime	15.268955	1389.135175
NaiveModel	79.718773	3864.279352
SimpleAverageModel	53.460570	1275.081804

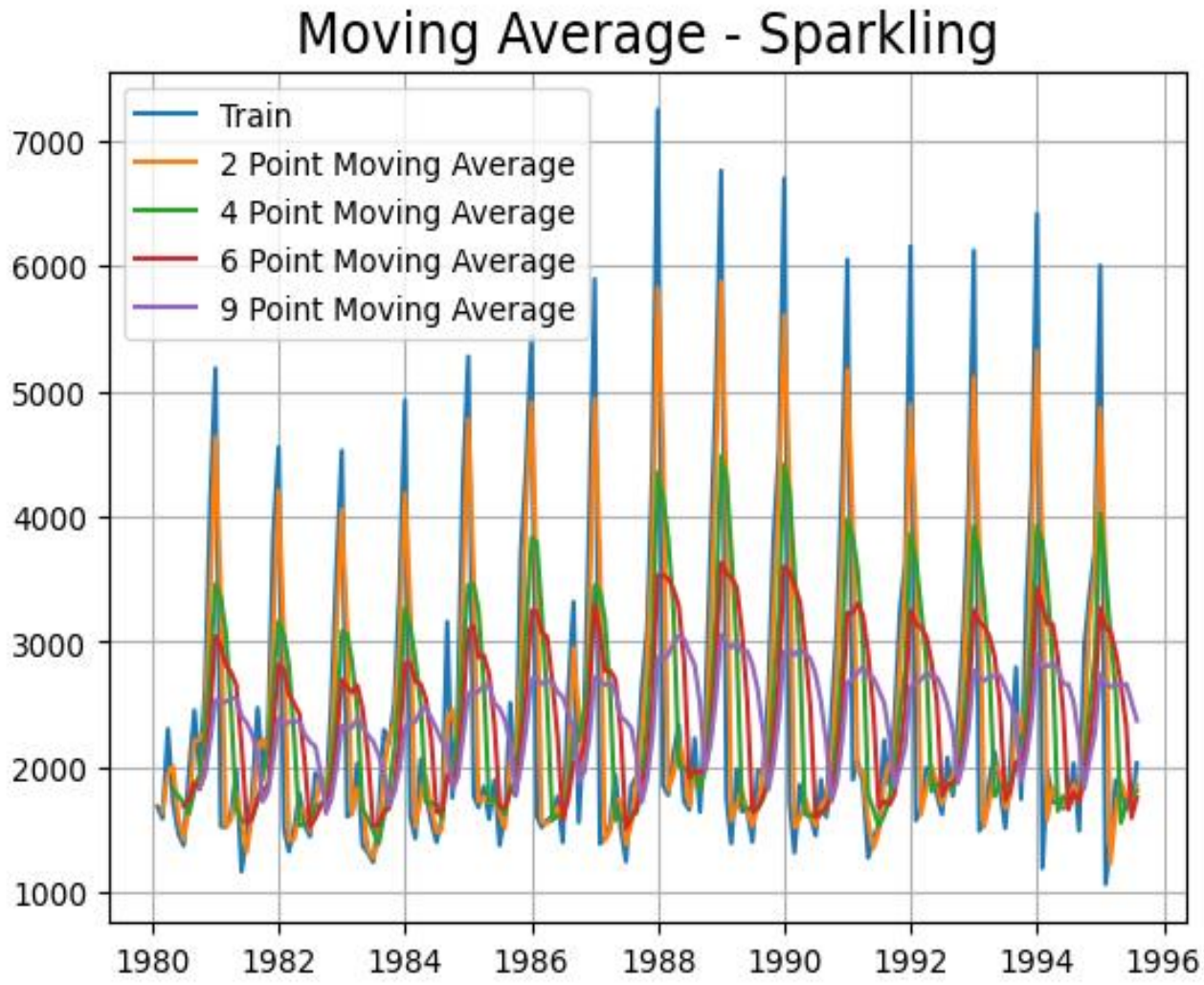
1.3.7 Model 4 Moving Average – Rose



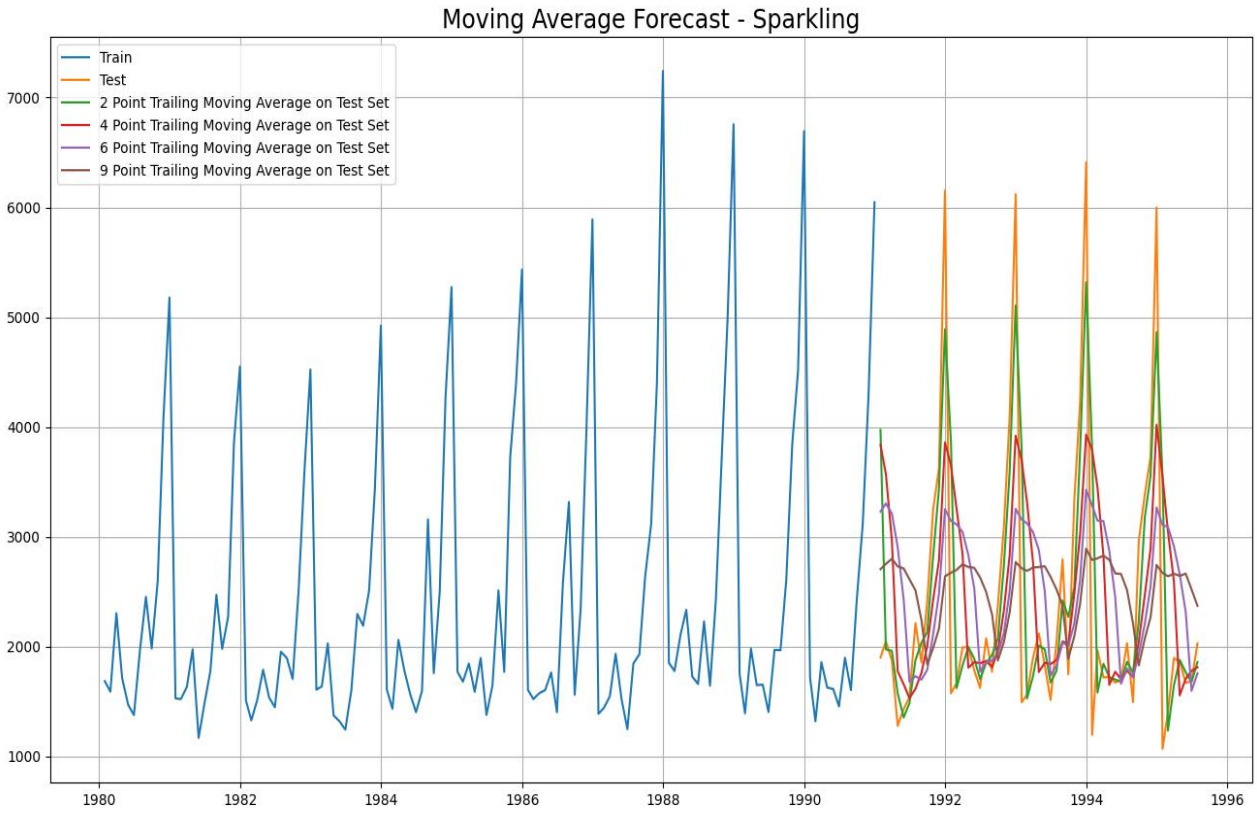
Moving Average on Train and Test (Both)- Rose Data set



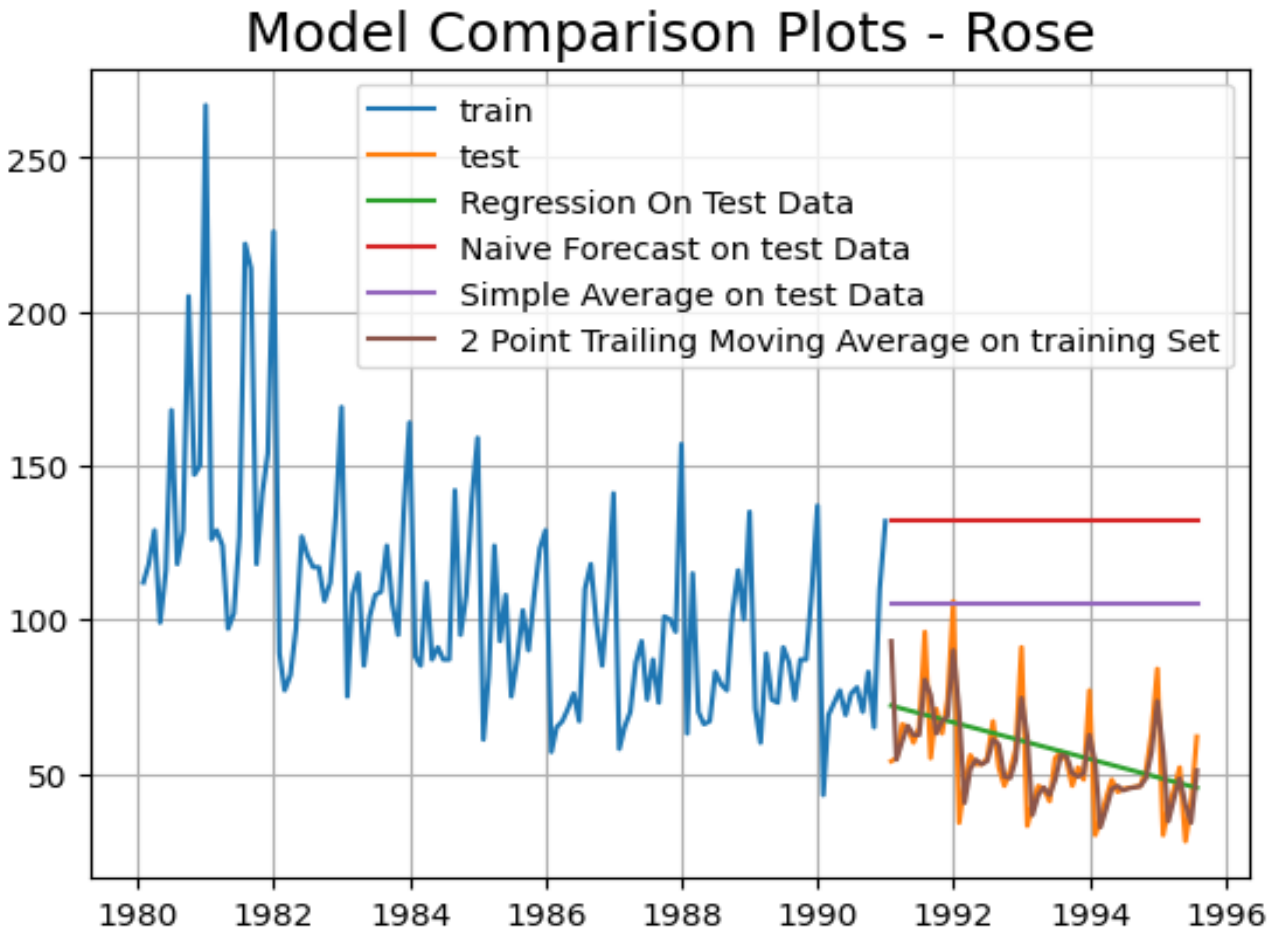
1.3.7 Model 4 Moving Average – Sparkling



Moving Average on Train and Test (Both)- Sparkling Data set

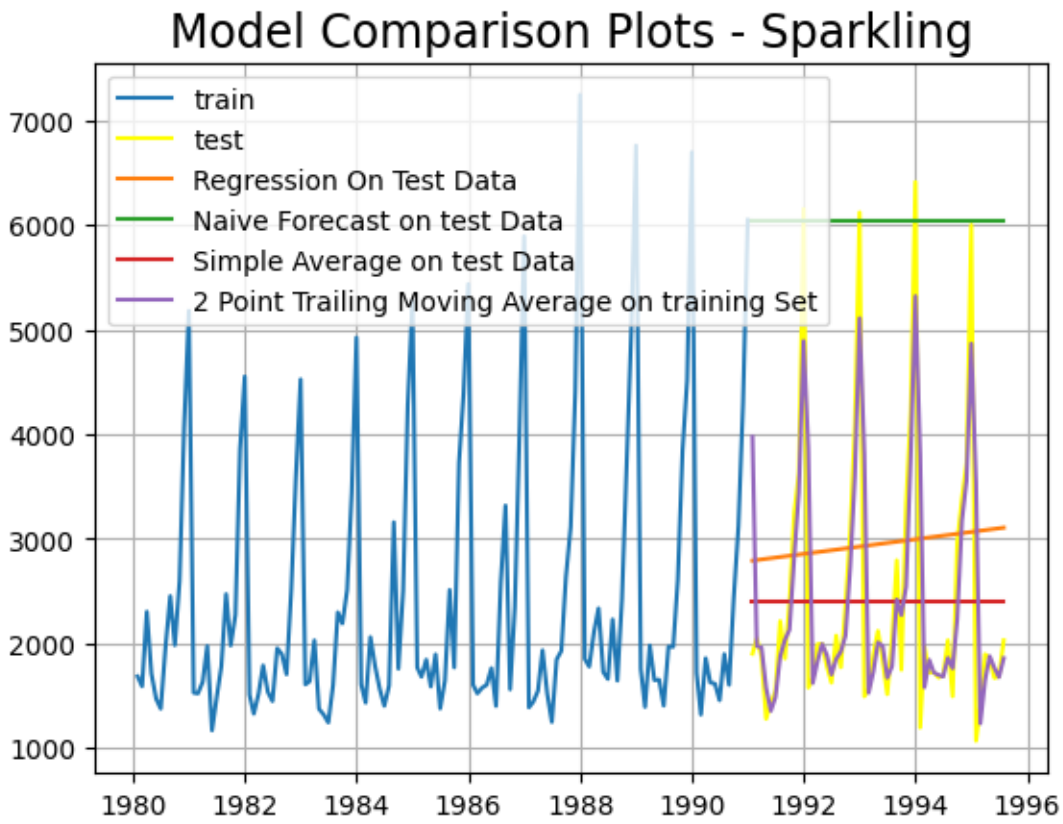


➢ Consolidated plots of All Models- Rose



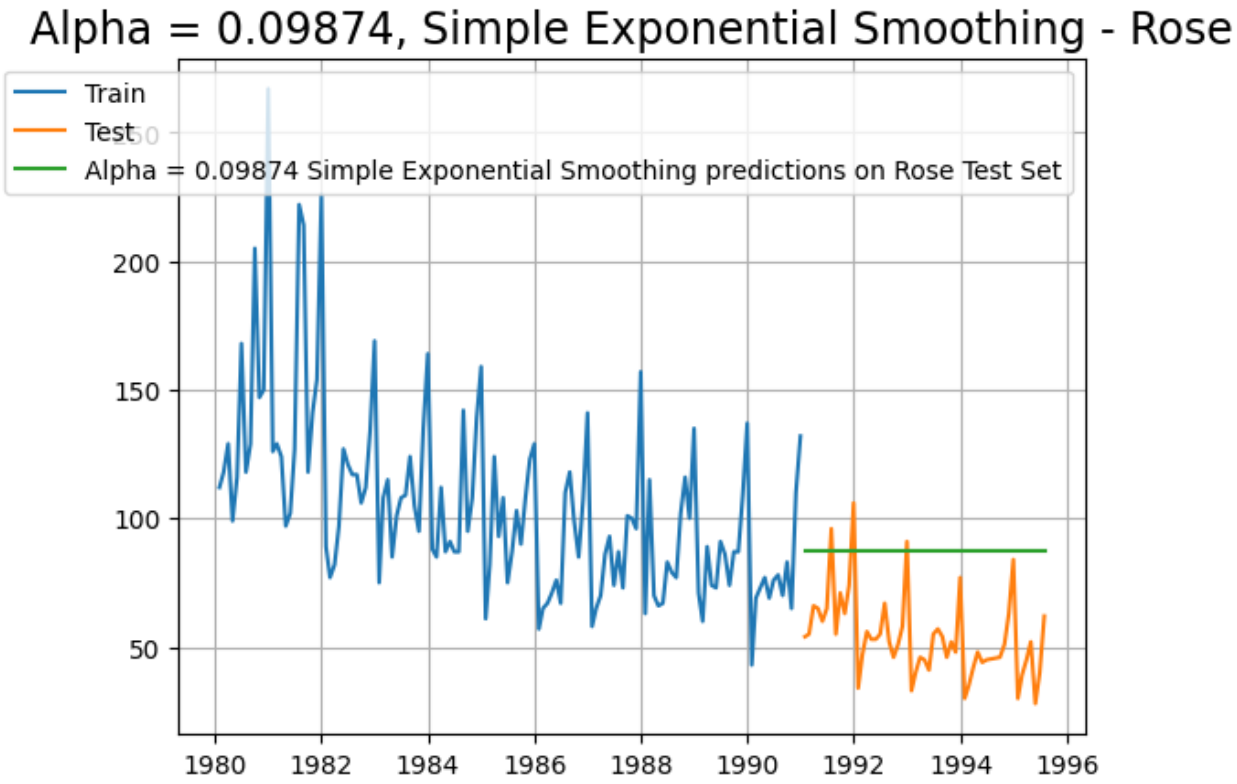
	Test RMSE Rose	Test RMSE Sparkling
RegressionOnTime	15.268955	1389.135175
NaiveModel	79.718773	3864.279352
SimpleAverageModel	53.460570	1275.081804
2pointTrailingMovingAverage	11.529278	813.400684
4pointTrailingMovingAverage	14.451403	1156.589694
6pointTrailingMovingAverage	14.566327	1283.927428
9pointTrailingMovingAverage	14.727630	1346.278315

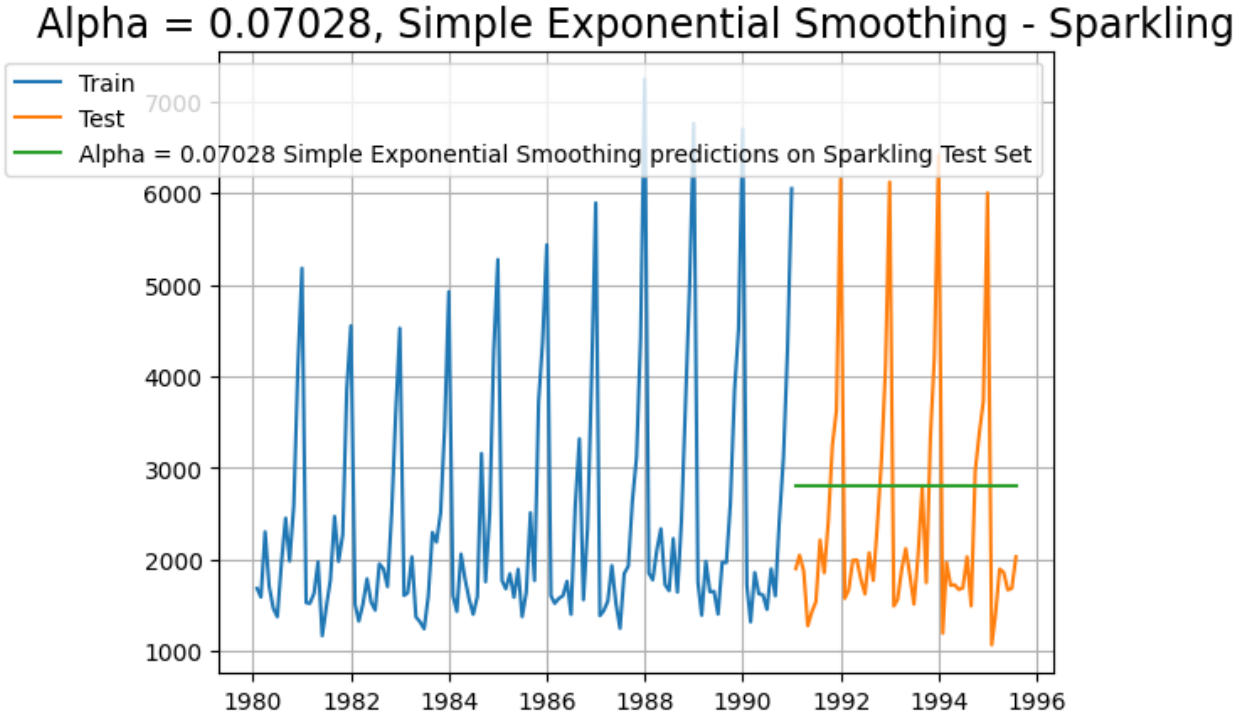
➢ Consolidated plots of All Models- Sparkling



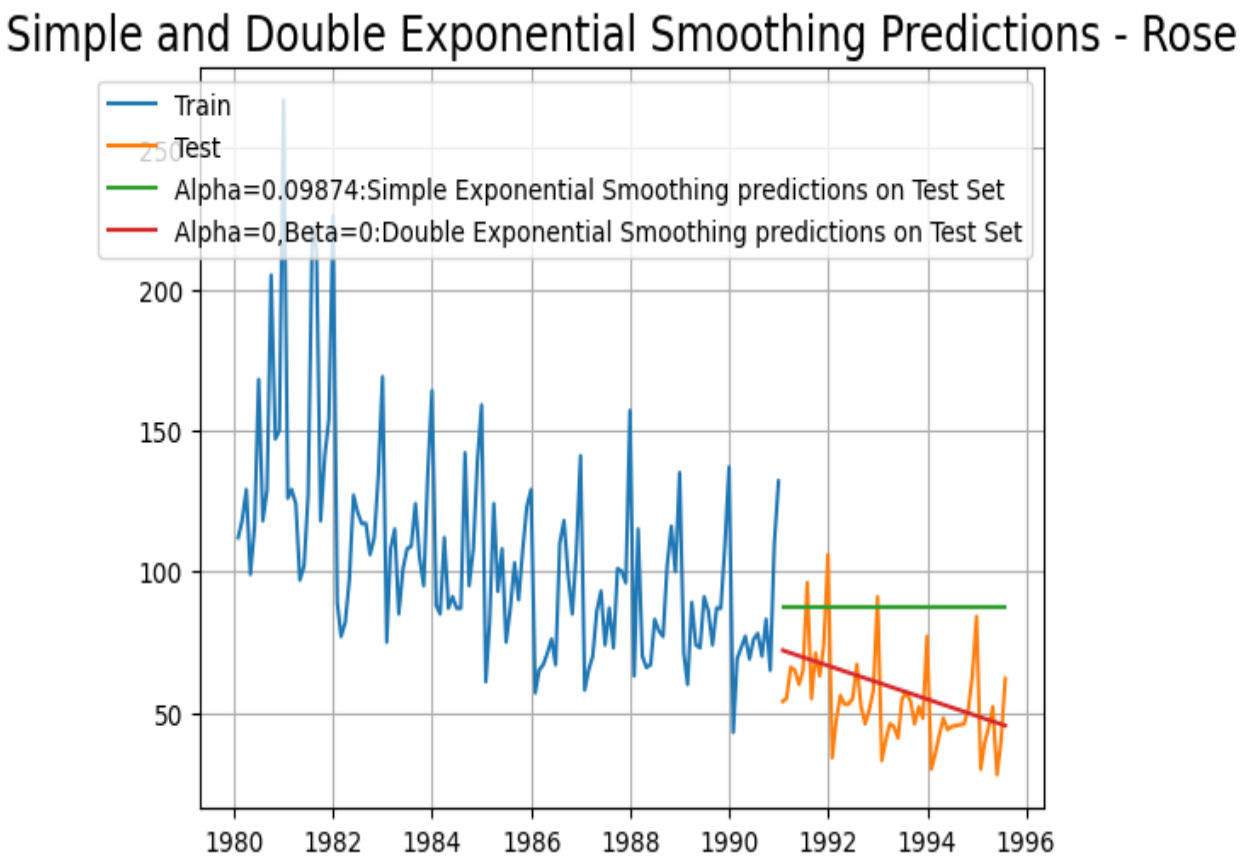
	Test RMSE Rose	Test RMSE Sparkling
RegressionOnTime	15.268955	1389.135175
NaiveModel	79.718773	3864.279352
SimpleAverageModel	53.460570	1275.081804
2pointTrailingMovingAverage	11.529278	813.400684
4pointTrailingMovingAverage	14.451403	1156.589694
6pointTrailingMovingAverage	14.566327	1283.927428
9pointTrailingMovingAverage	14.727630	1346.278315

- Simple Exponential Smoothing, Holt's Model (Double Exponential Smoothing) & Holt-Winter's
  - Model (Triple Exponential Smoothing)

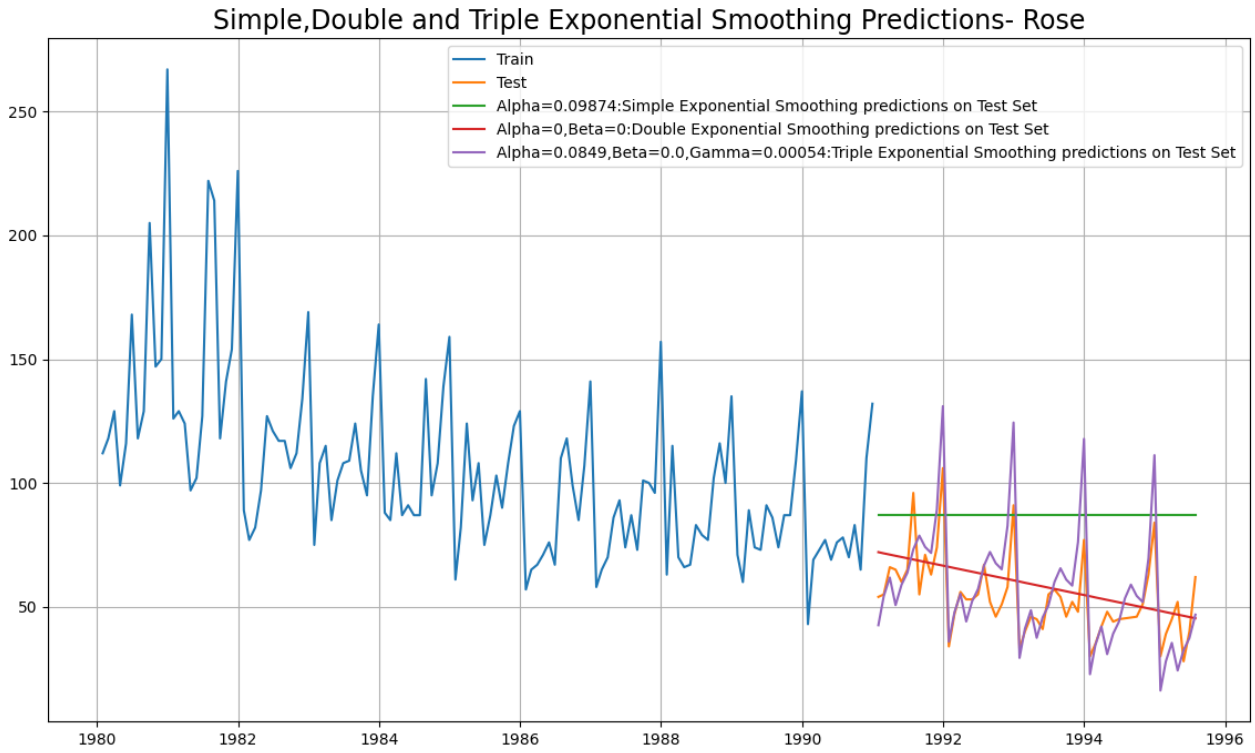




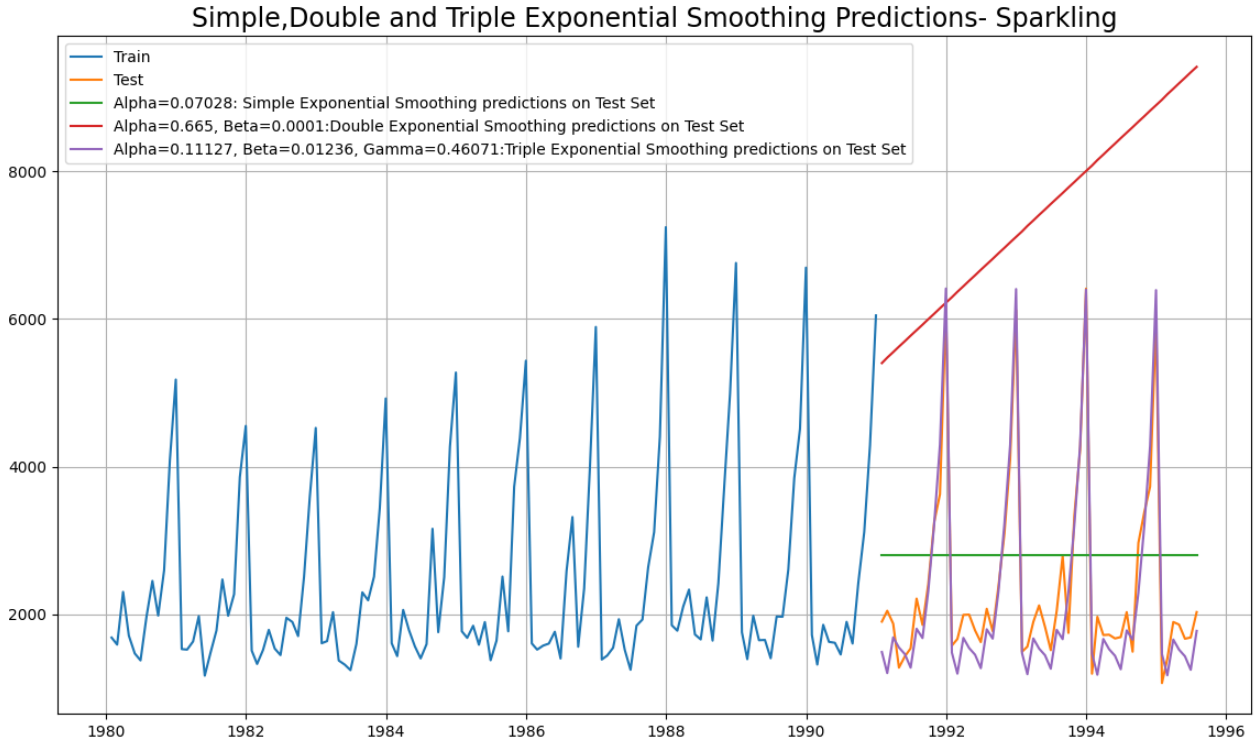
- Holt - ETS(A, A, N) - Holt's linear method with additive errors - Rose
  - Double Exponential Smoothing - Rose



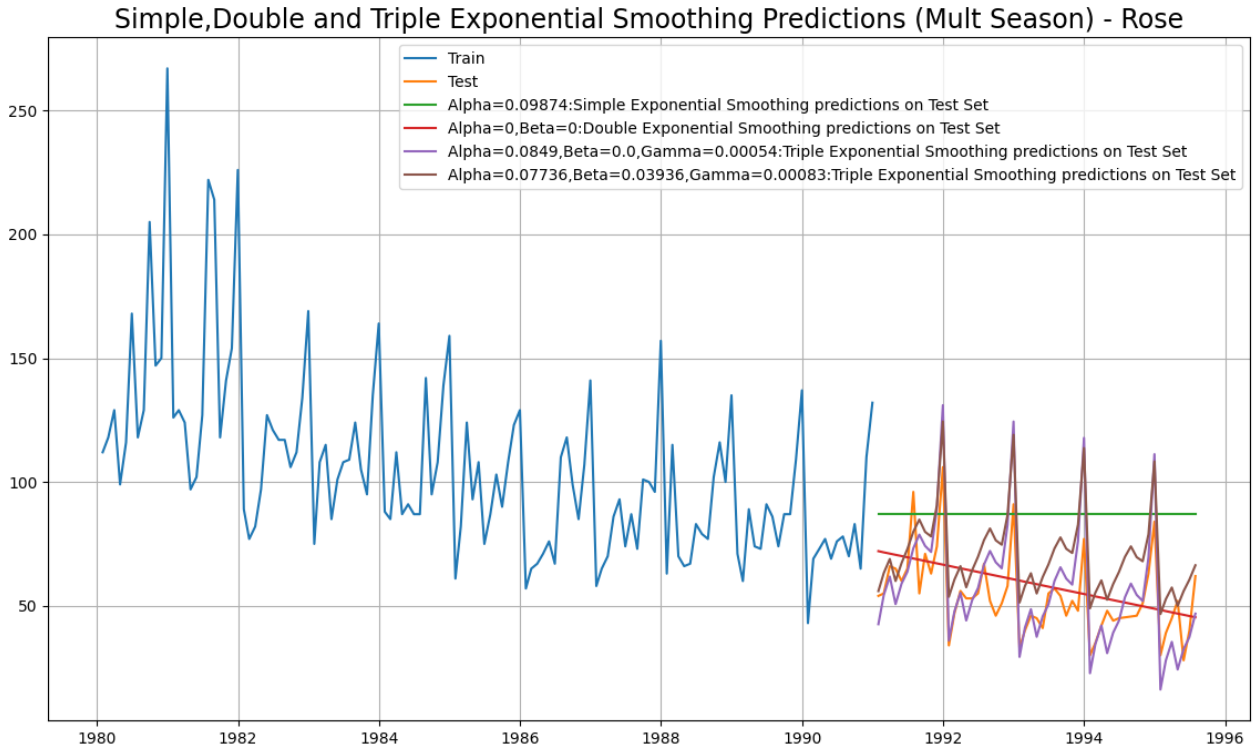
- Holt-Winters - ETS(A, A, A) - Holt Winter's linear method with additive errors
  - Rose



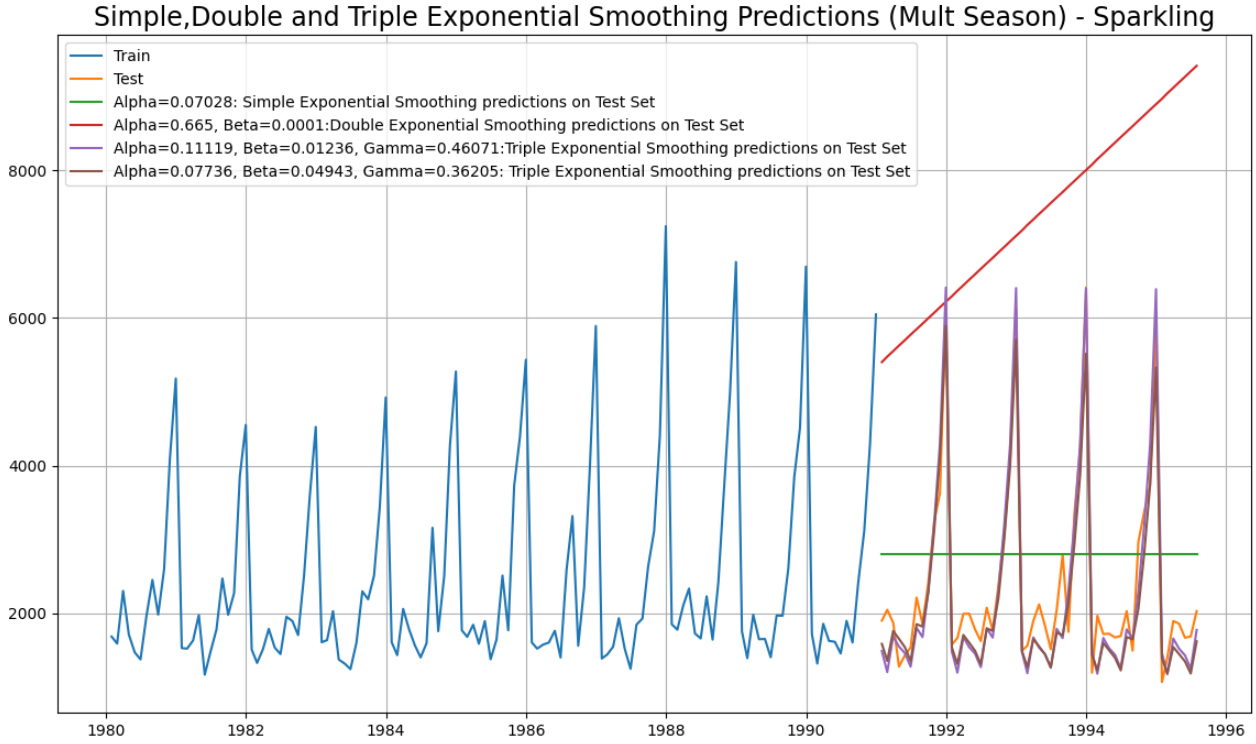
- Holt-Winters - ETS(A, A, A) - Holt Winter's linear method with additive errors
  - Sparkling

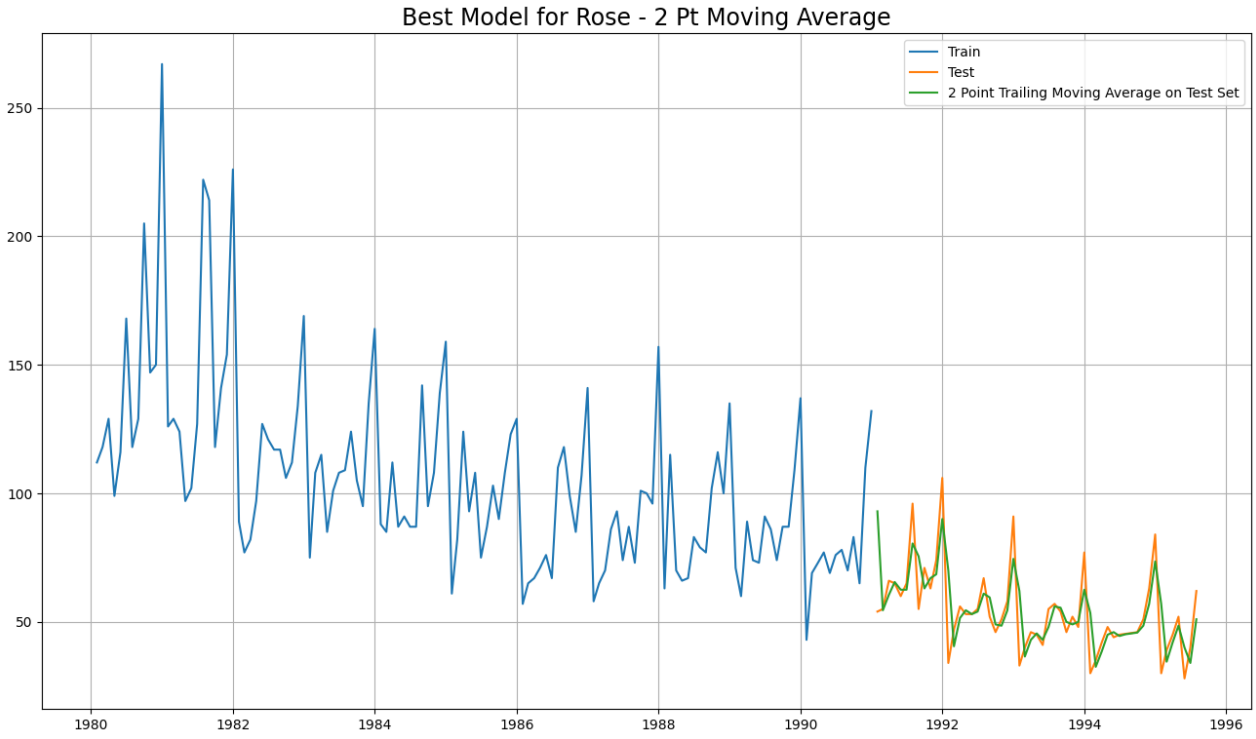


- Holt-Winters - ETS(A, A, M) - Holt Winter's linear method – ROSE
  - Multi seasonality



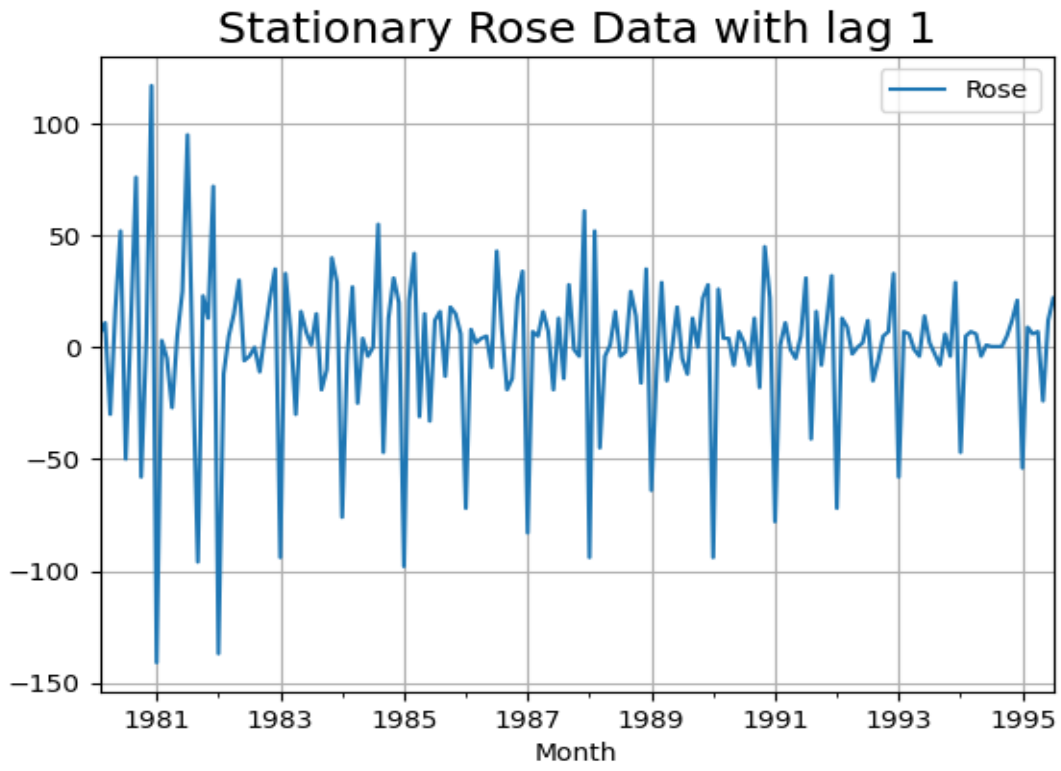
- Holt-Winters - ETS(A, A, M) - Holt Winter's linear method – Sparkling
  - Multi seasonality



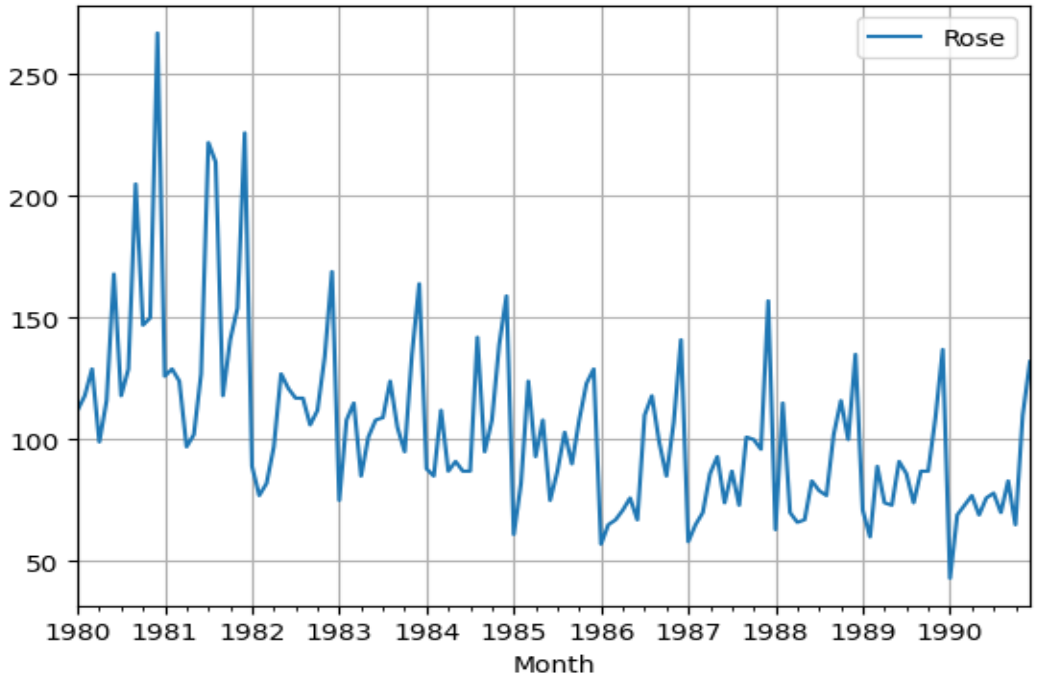

Best Model for Rose - 2 Pt Moving Average
Train
Test
2 Point Trailing Moving Average on Test Set
250
200
150
100
50
1980
1982
1984
1986
1988
1990
1992
1994
1996

1.4 Check Stationary

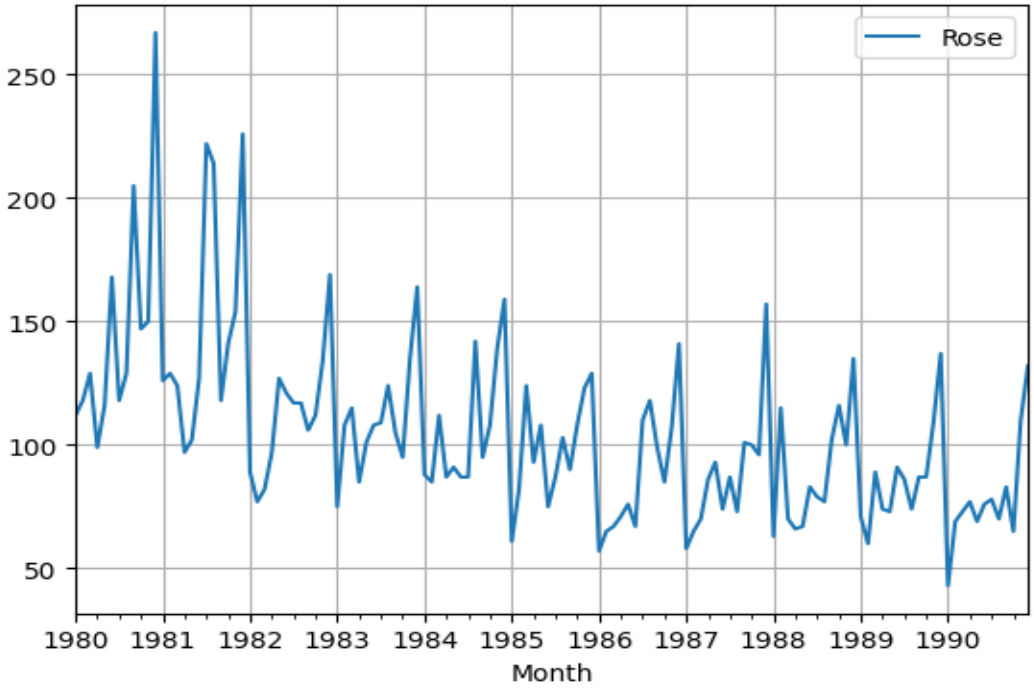
- Stationary Rose Data with lag 1



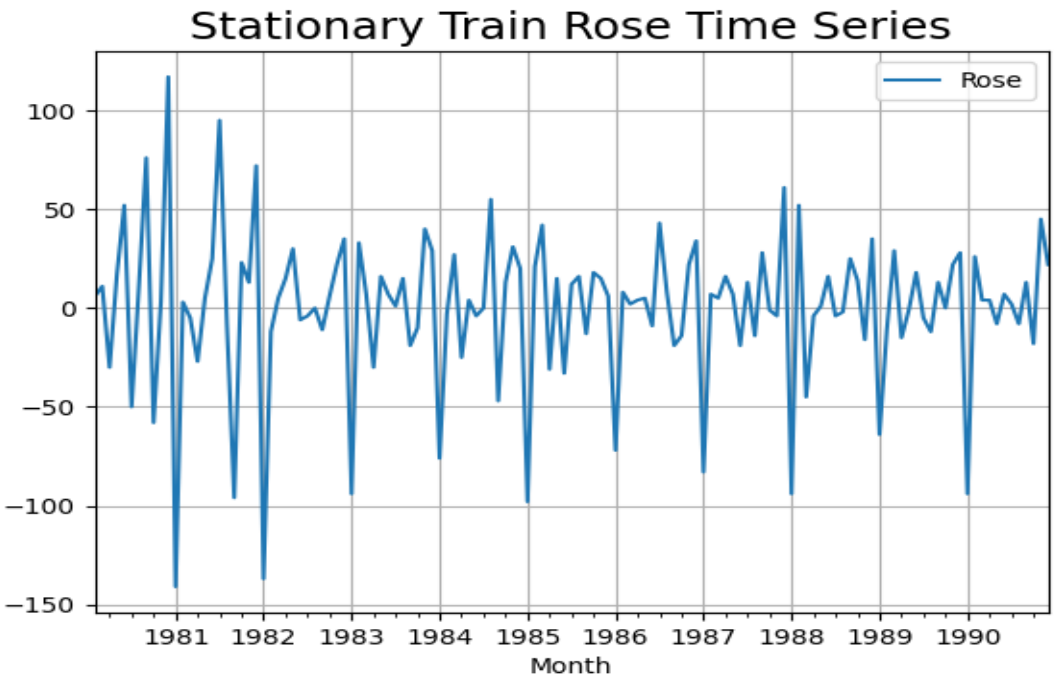
- Stationary Sparkling Data with lag 1



1.5 Model Building Stationary Data



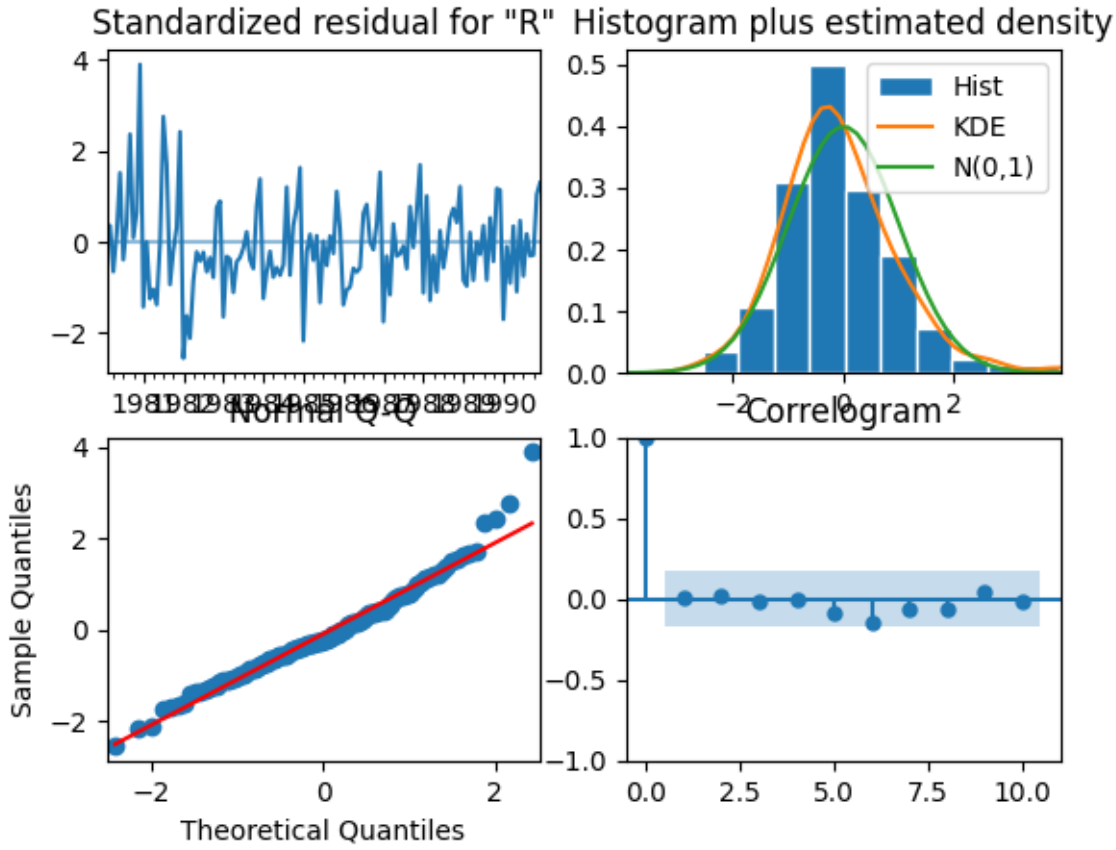
- Check for Stationarity of Rose Train Data



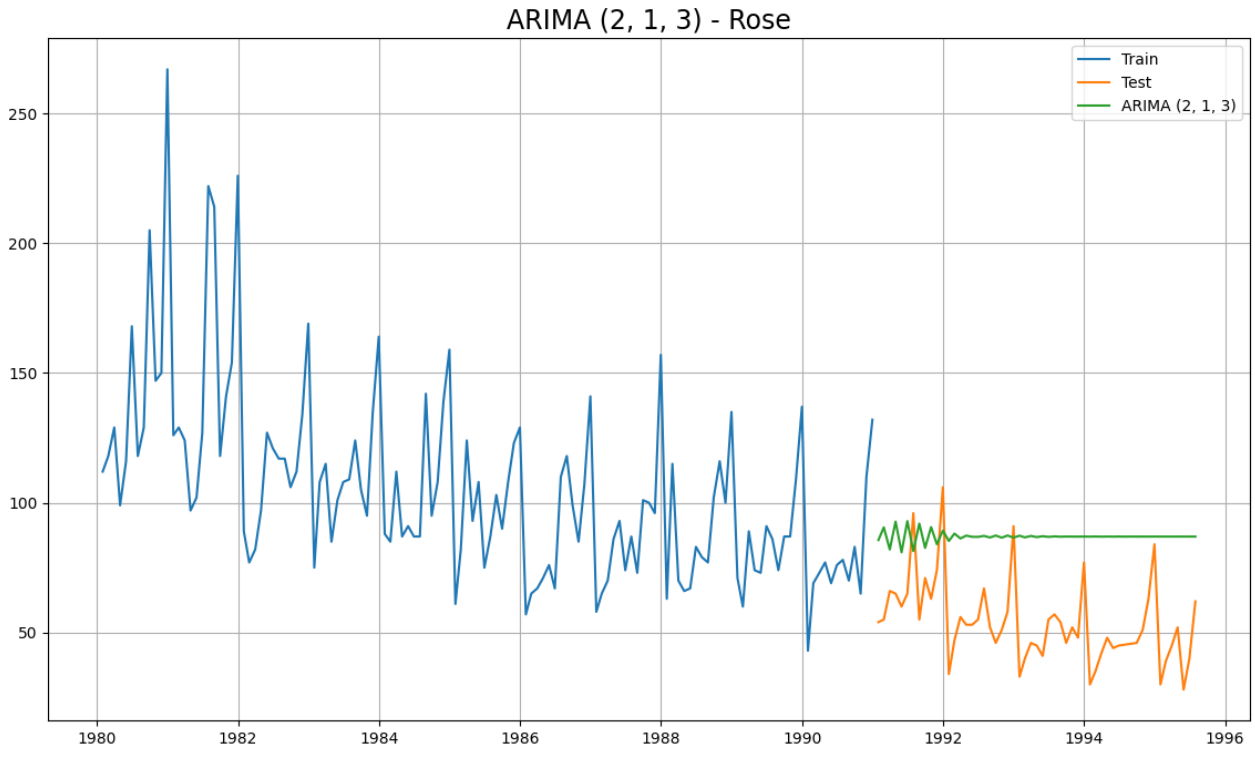
Build an Automated version of an ARIMA model for which the best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC).

```
                               SARIMAX Results
==========================================================================================
Dep. Variable:                      Rose   No. Observations:                  132
Model:                 ARIMA(2, 1, 3)   Log Likelihood                -631.348
Date:                Sat, 13 Apr 2024   AIC                           1274.695
Time:                        07:22:50   BIC                           1291.946
Sample:                    01-31-1980   HQIC                          1281.705
                         - 12-31-1990
Covariance Type:                  opg
==========================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.6779      0.084    -20.034      0.000      -1.842      -1.514
ar.L2         -0.7288      0.084     -8.702      0.000      -0.893      -0.565
ma.L1          1.0447      0.644      1.622      0.105      -0.217       2.307
ma.L2         -0.7718      0.134     -5.775      0.000      -1.034      -0.510
ma.L3         -0.9046      0.584     -1.549      0.121      -2.049       0.240
sigma2       858.8436    541.924      1.585      0.113    -203.308    1920.995
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                24.45
Prob(Q):                              0.88   Prob(JB):                         0.00
Heteroskedasticity (H):               0.40   Skew:                             0.71
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.57
===================================================================================
```

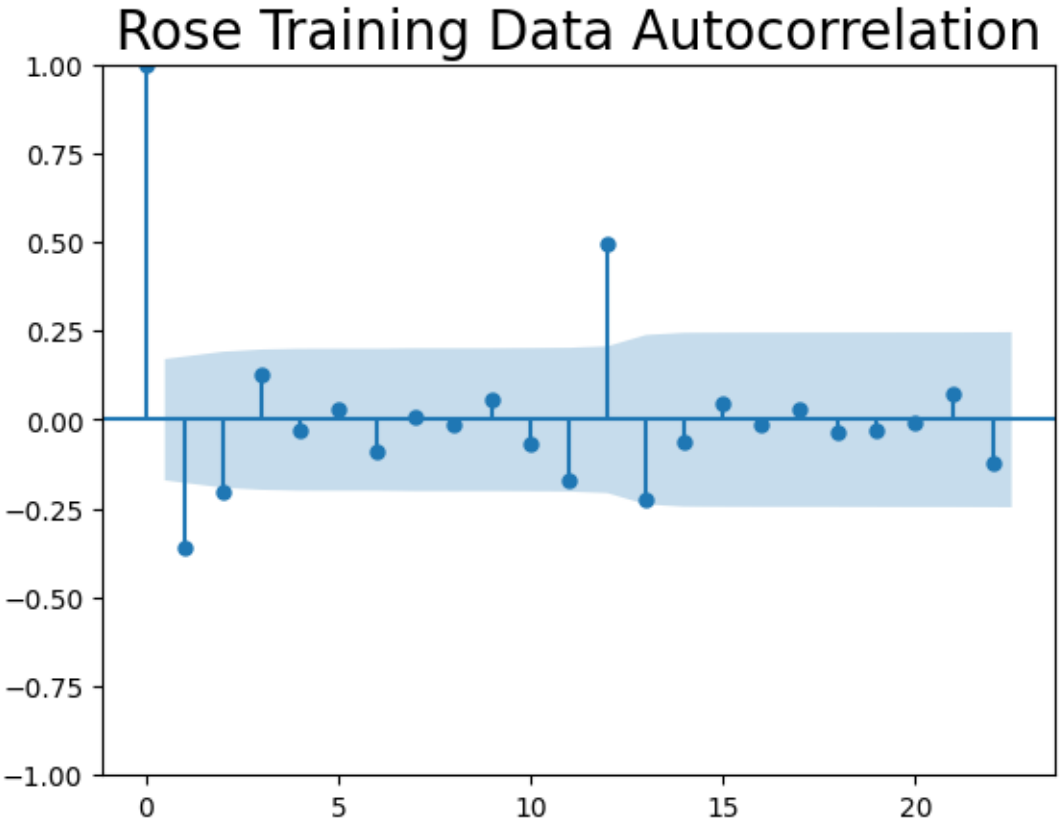
Rose Train Diagnostics plot.

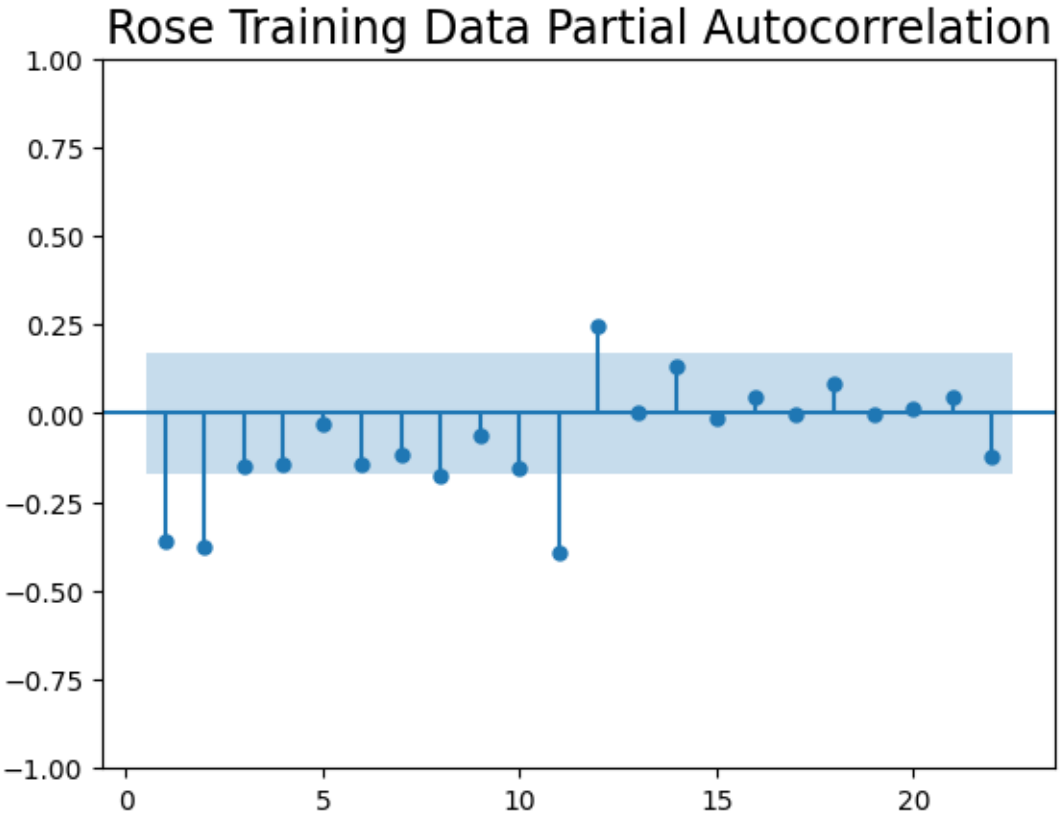


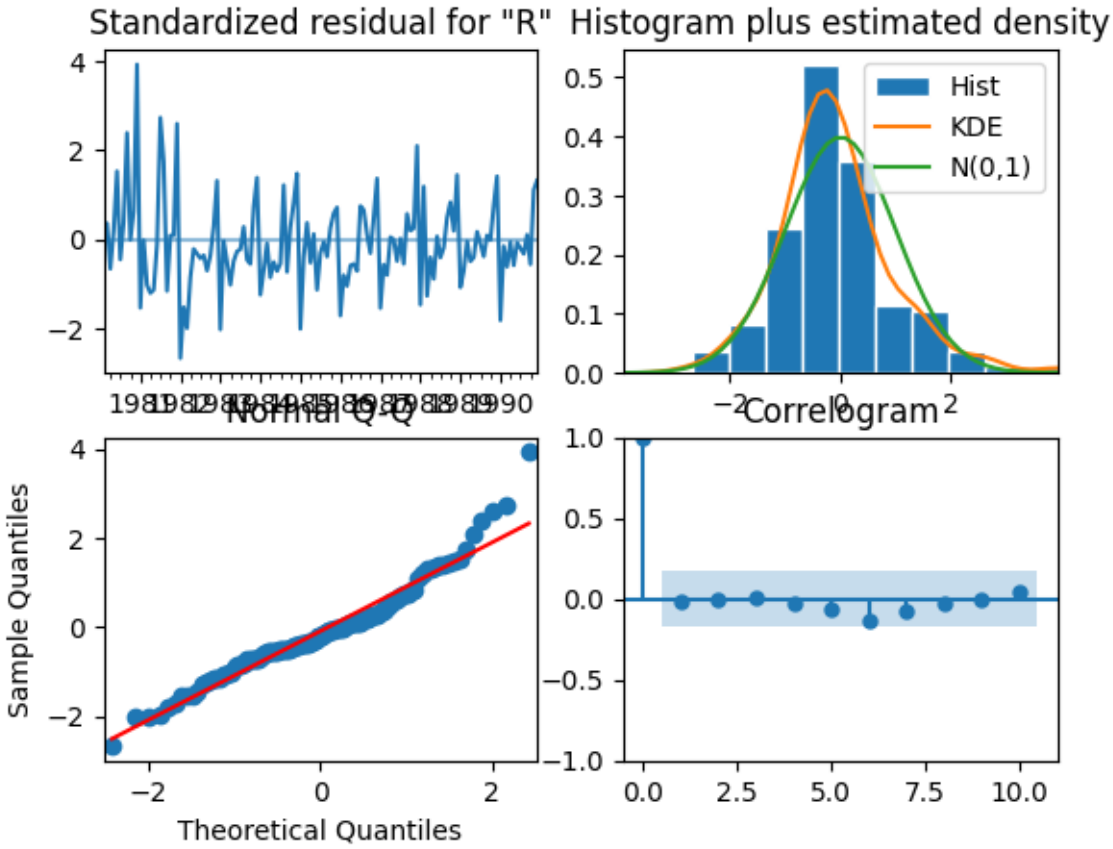
Predict on the Rose Test Set using this model and evaluate the model.



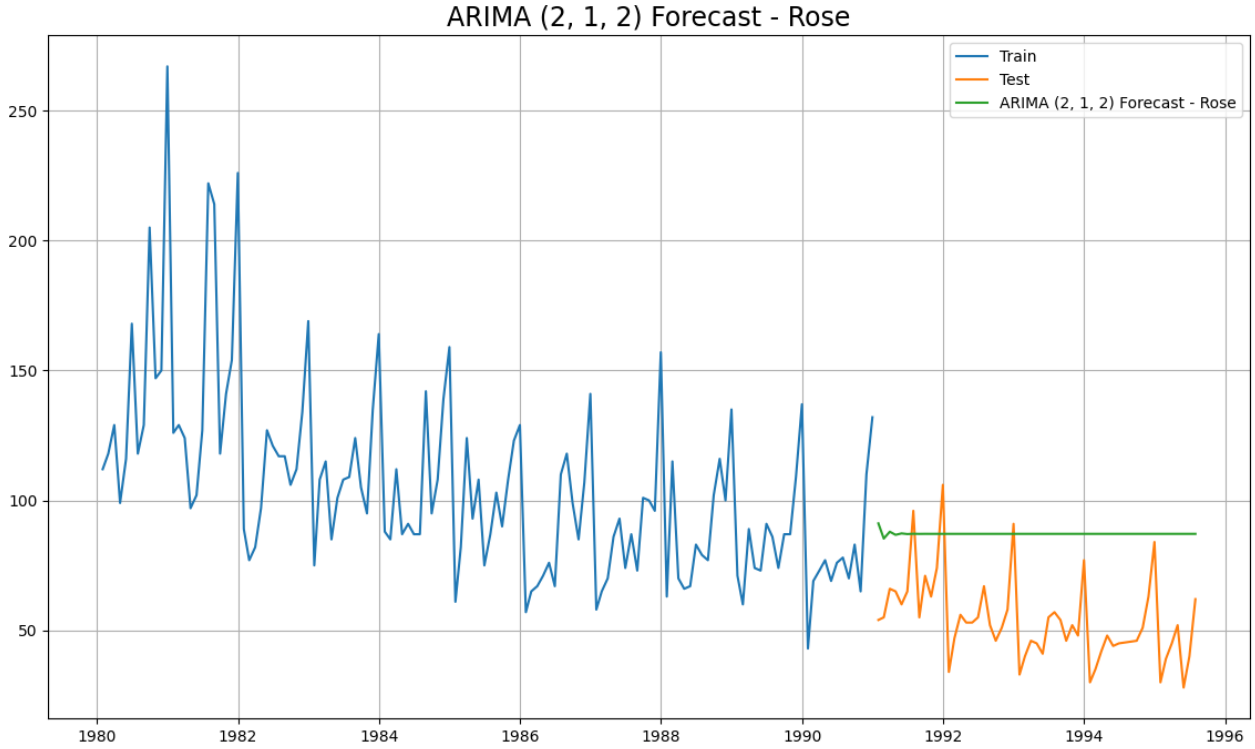
Build a version of the ARIMA model for which the best parameters are selected by looking at the ACF and the PACF plots on ROSE dataset



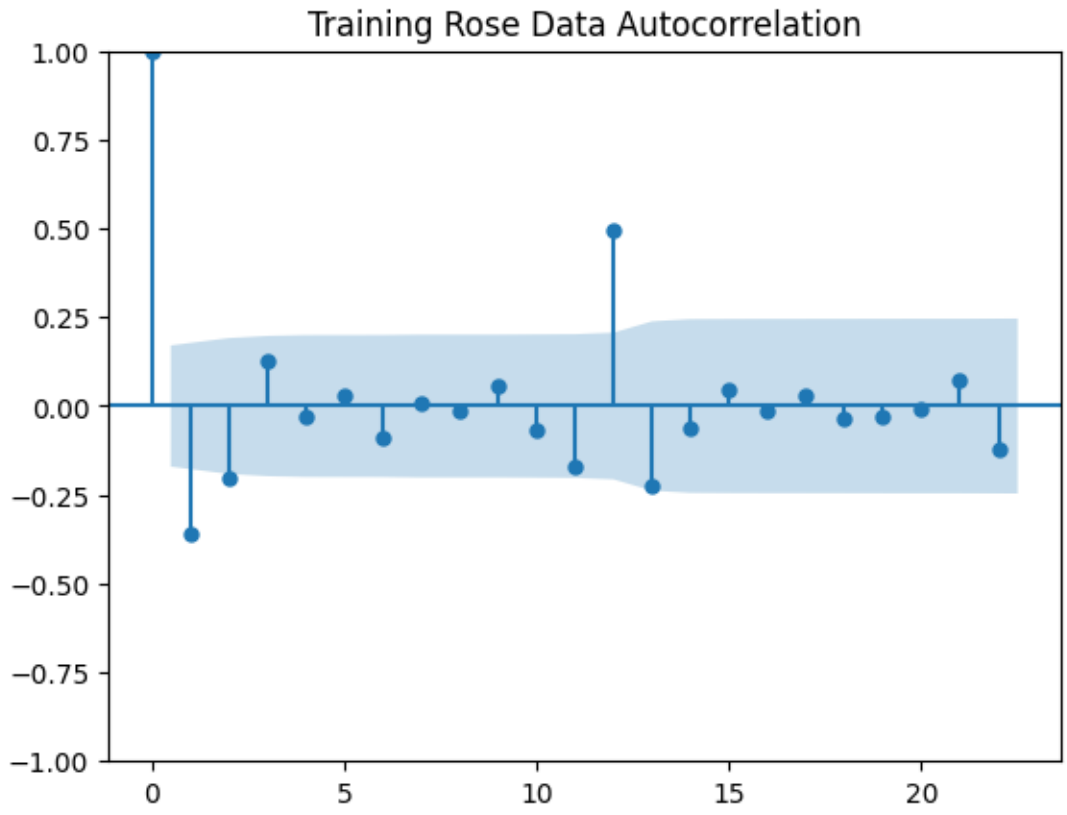


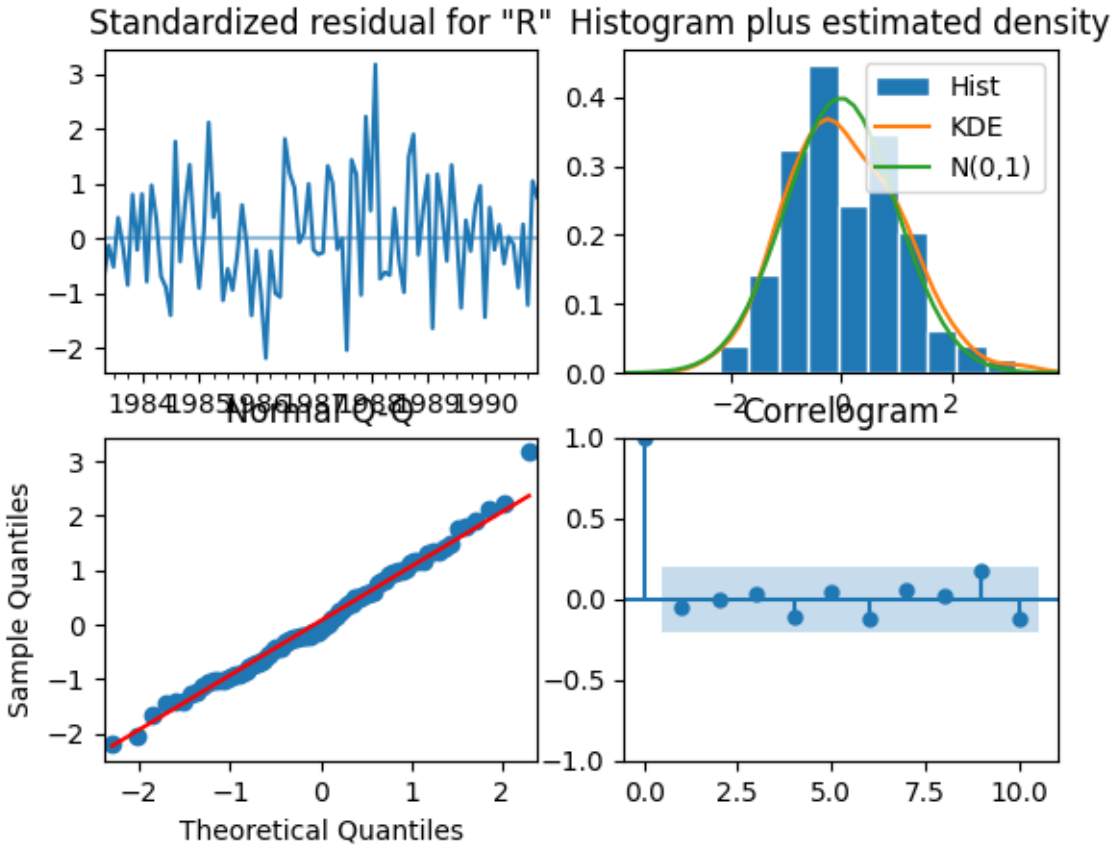


Predict on the Test Set using this model and evaluate the model.

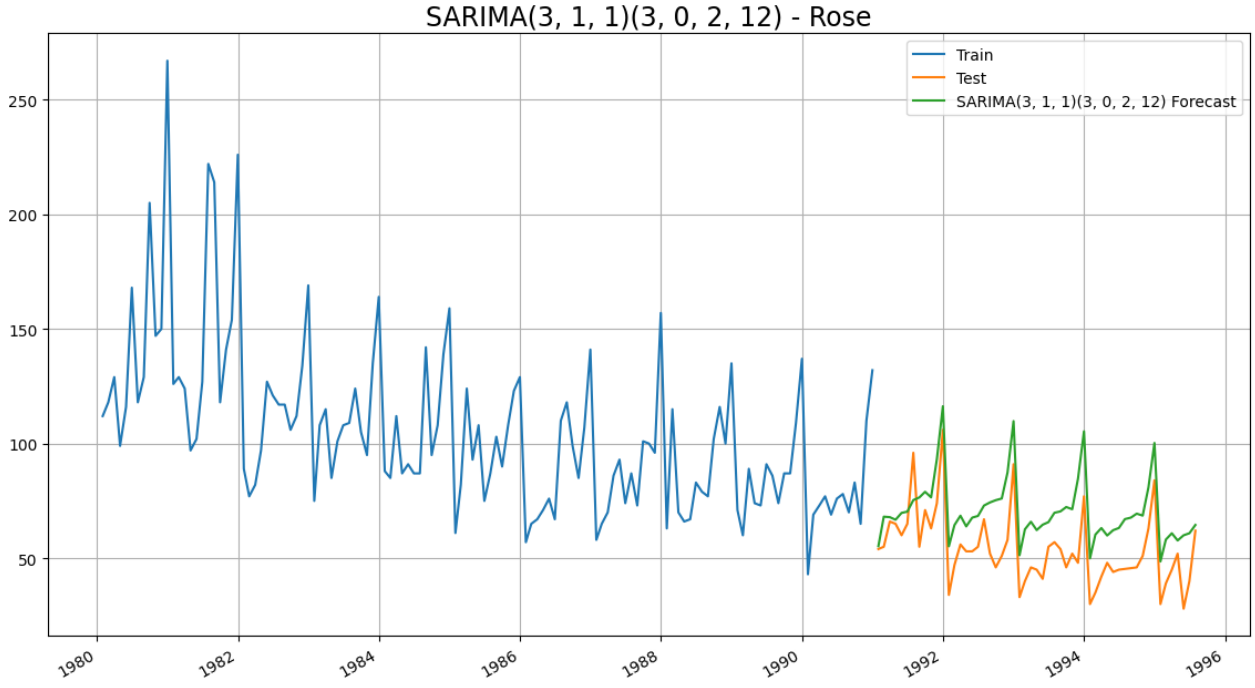


Build an Automated version of a SARIMA model for which the best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC) - ROSE DATA

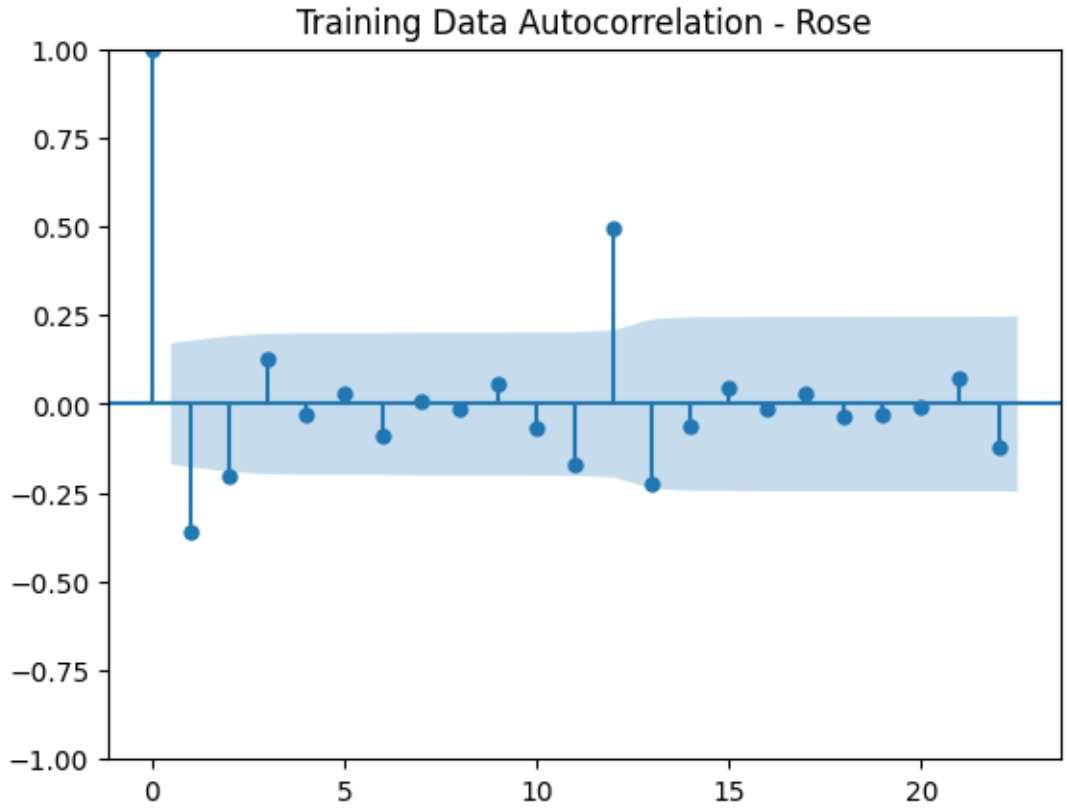


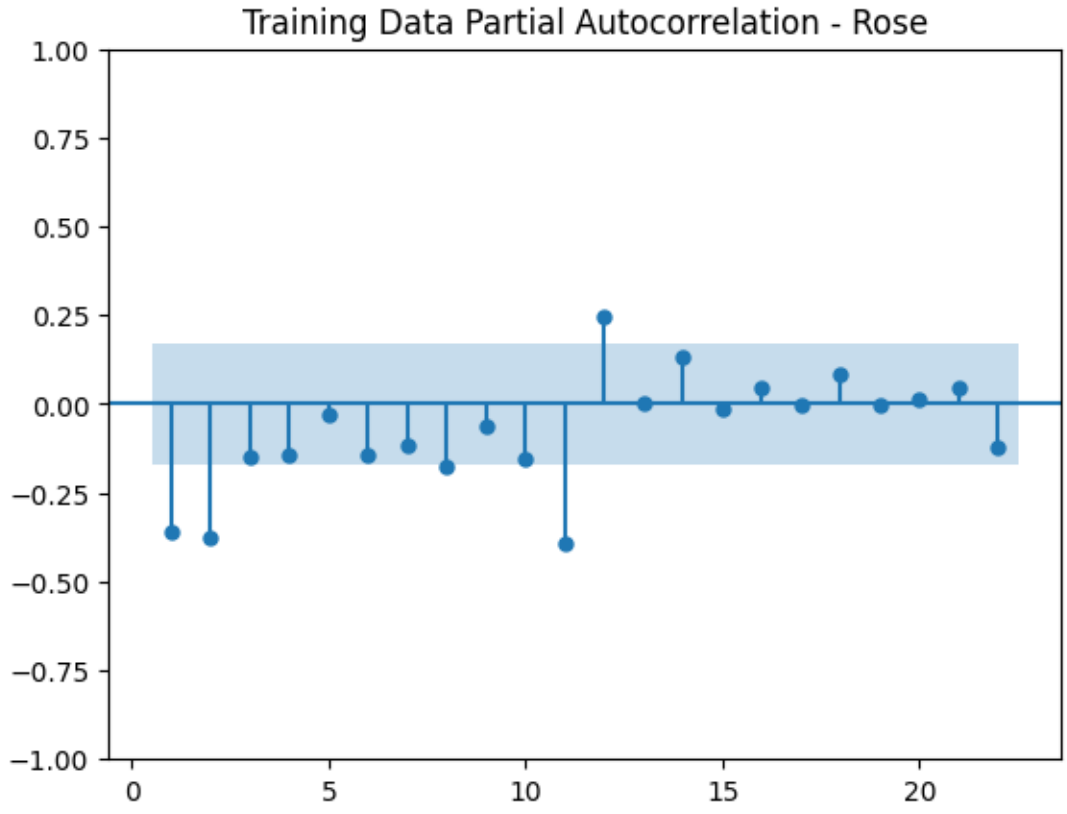


Predict on the Test Set using this model and evaluate the model.



Build a version of the SARIMA model for which the best parameters are selected by looking at the ACF and the PACF plots. - Seasonality at 12





Here, we have taken alpha=0.05.

We are going to take the seasonal period as 12 We are taking the p value to be 2 and the q value also to be 2 as the parameters same as the ARIMA model.

1.The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 0.

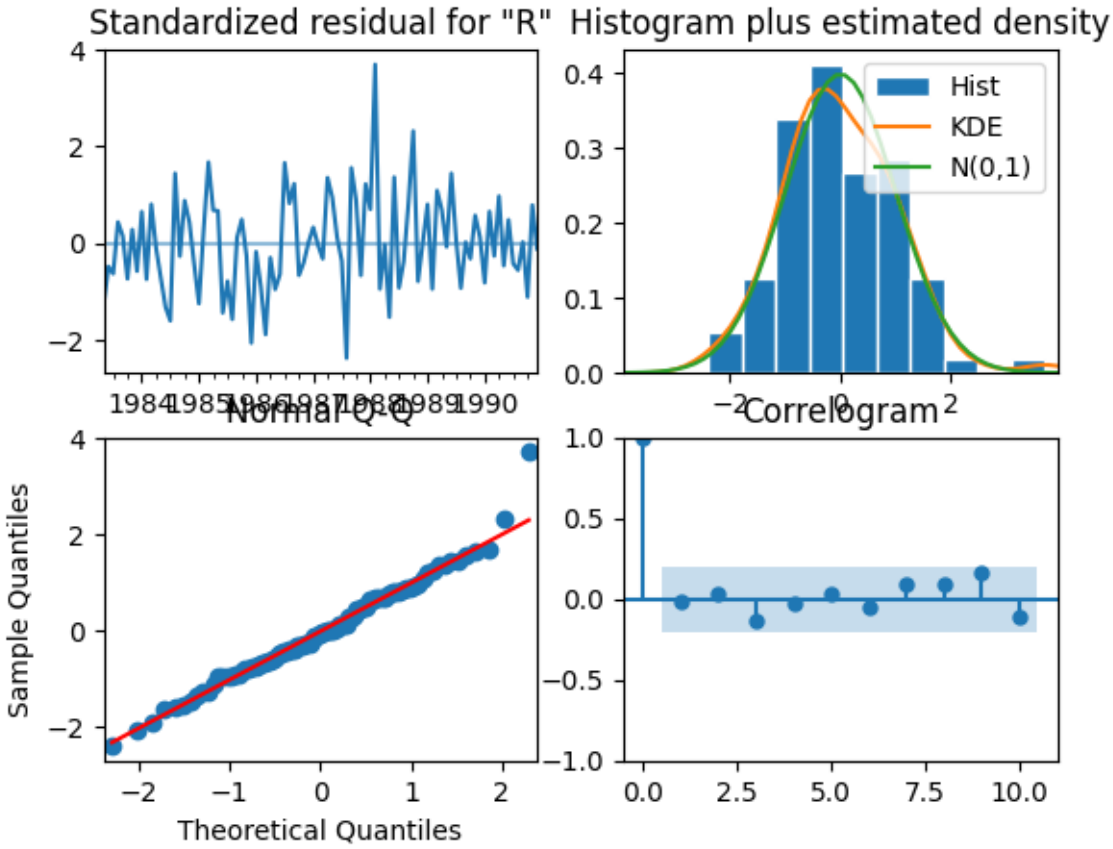
2.The Moving-Average parameter in an SARIMA model is 'Q' which comes from the significant lag after which the ACF plot cuts-off to 3.

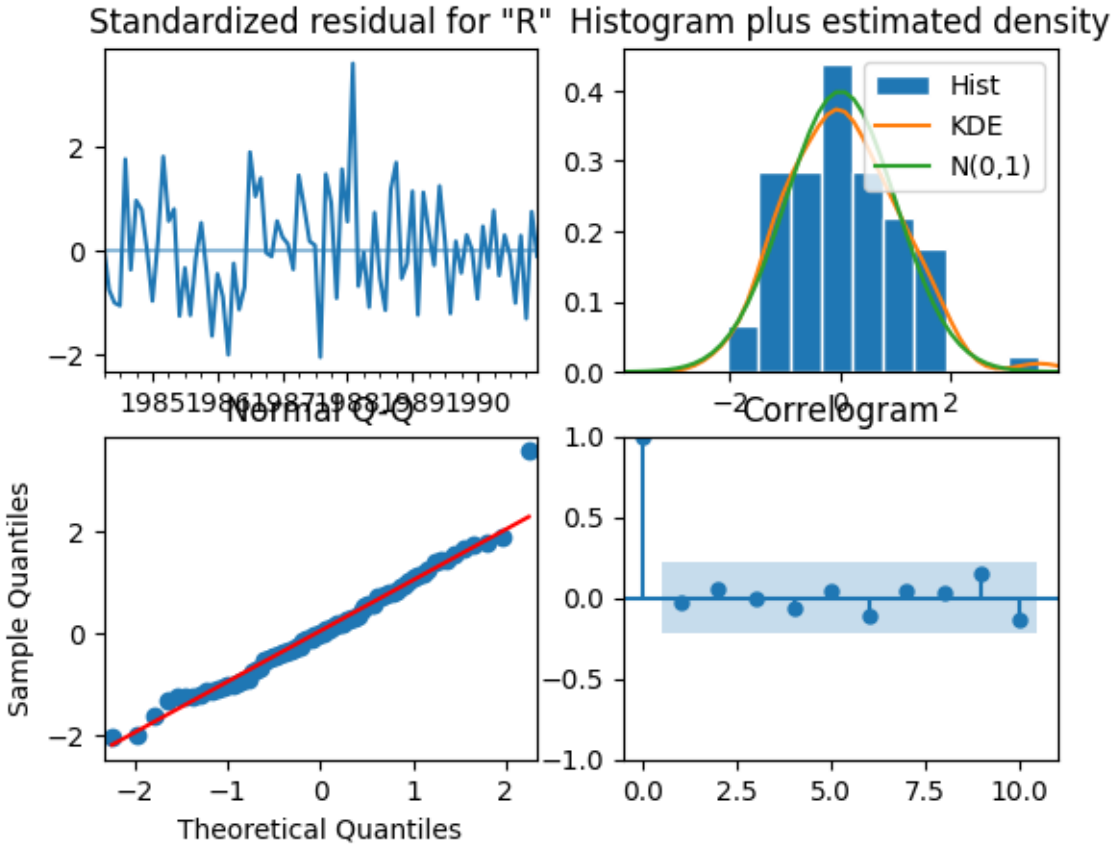
SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	SARIMAX(2, 1, 2)x(2, 1, 2, 12)	Log Likelihood	-379.498
Date:	Sat, 13 Apr 2024	AIC	776.996
Time:	07:47:03	BIC	799.692
Sample:	01-31-1980 - 12-31-1990	HQIC	786.156
Covariance Type:	opg		

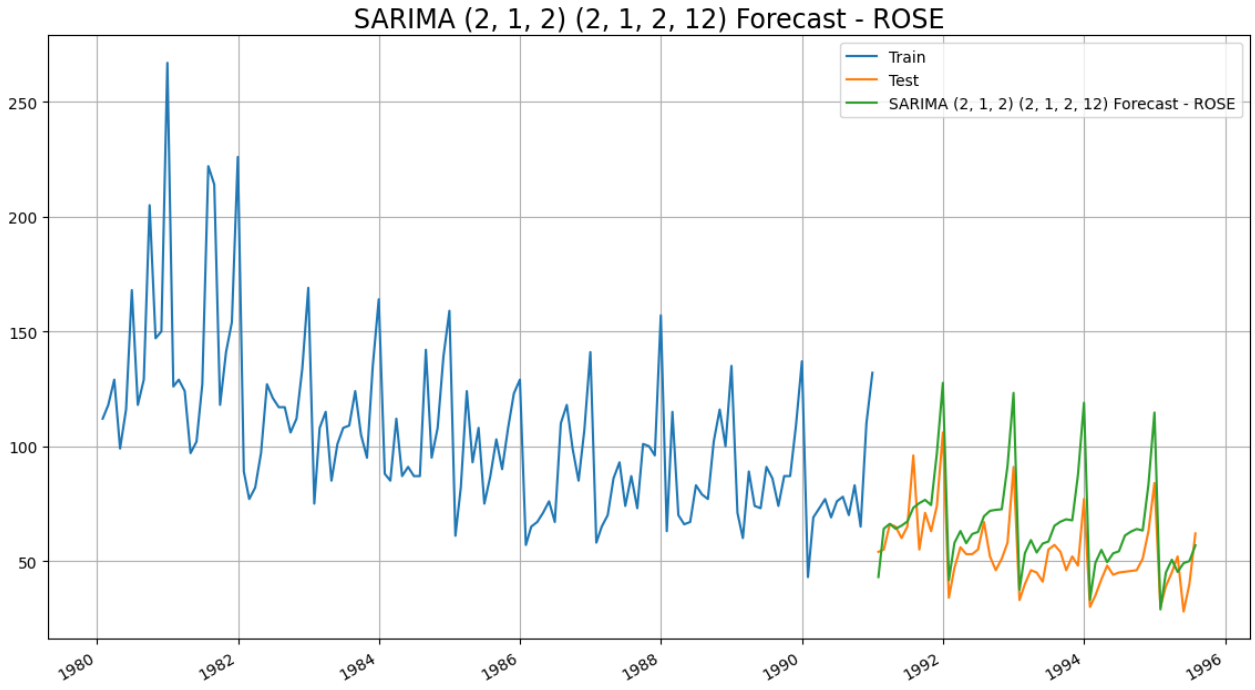
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.8551	0.146	-5.838	0.000	-1.142	-0.568
ar.L2	-0.0022	0.125	-0.017	0.986	-0.247	0.242
ma.L1	-0.0128	0.193	-0.066	0.947	-0.392	0.366
ma.L2	-1.0600	0.168	-6.294	0.000	-1.390	-0.730
ar.S.L12	0.0347	0.185	0.187	0.851	-0.328	0.397
ar.S.L24	-0.0459	0.029	-1.598	0.110	-0.102	0.010
ma.S.L12	-0.7223	0.333	-2.172	0.030	-1.374	-0.071
ma.S.L24	-0.0772	0.212	-0.364	0.716	-0.493	0.339
sigma2	171.0334	54.128	3.160	0.002	64.945	277.122

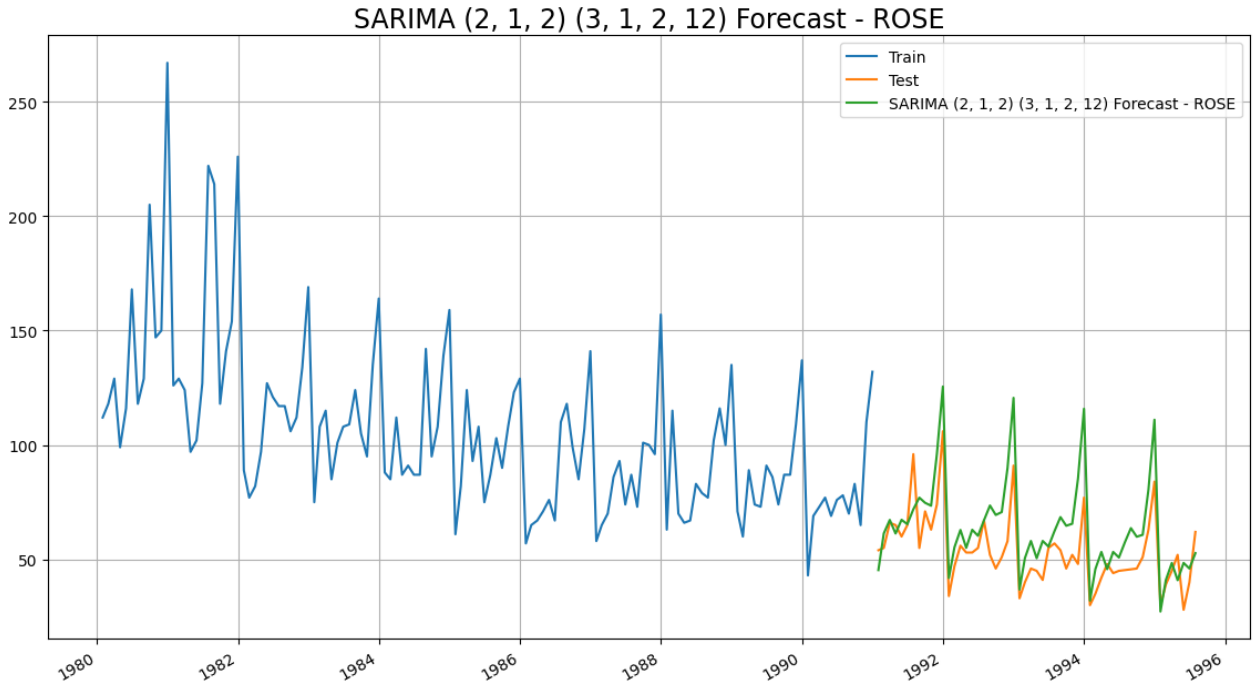
Ljung-Box (L1) (Q):	0.03	Jarque-Bera (JB):	7.06
Prob(Q):	0.86	Prob(JB):	0.03
Heteroskedasticity (H):	0.87	Skew:	0.45
Prob(H) (two-sided):	0.71	Kurtosis:	4.01


Standardized residual for "R"
Histogram plus estimated density
Hist
KDE
N(0,1)
Normal Q-Q
Correlogram
Sample Quantiles
Theoretical Quantiles


Standardized residual for "R"
Histogram plus estimated density
Hist
KDE
N(0,1)
Normal Q-Q
Correlogram
Sample Quantiles
Theoretical Quantiles

Predict on the Test Set using this model and evaluate the model.





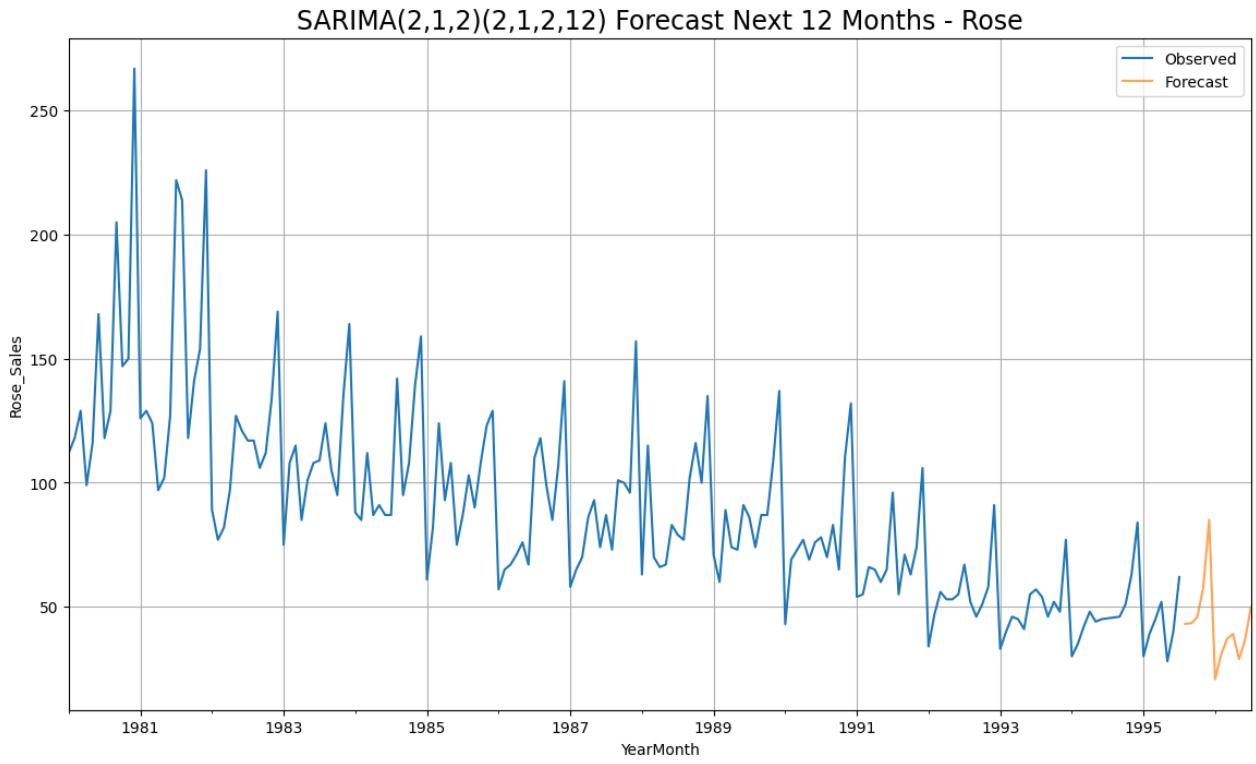
1.6 Compare the performance of the models

Building the most optimum model on the Full Data.

Here, we have a scenario where our training data was stationary but our full data was not stationary. So, we will use the same parameters as our training data but with adding a level of differencing which is needed for the data to be stationary.

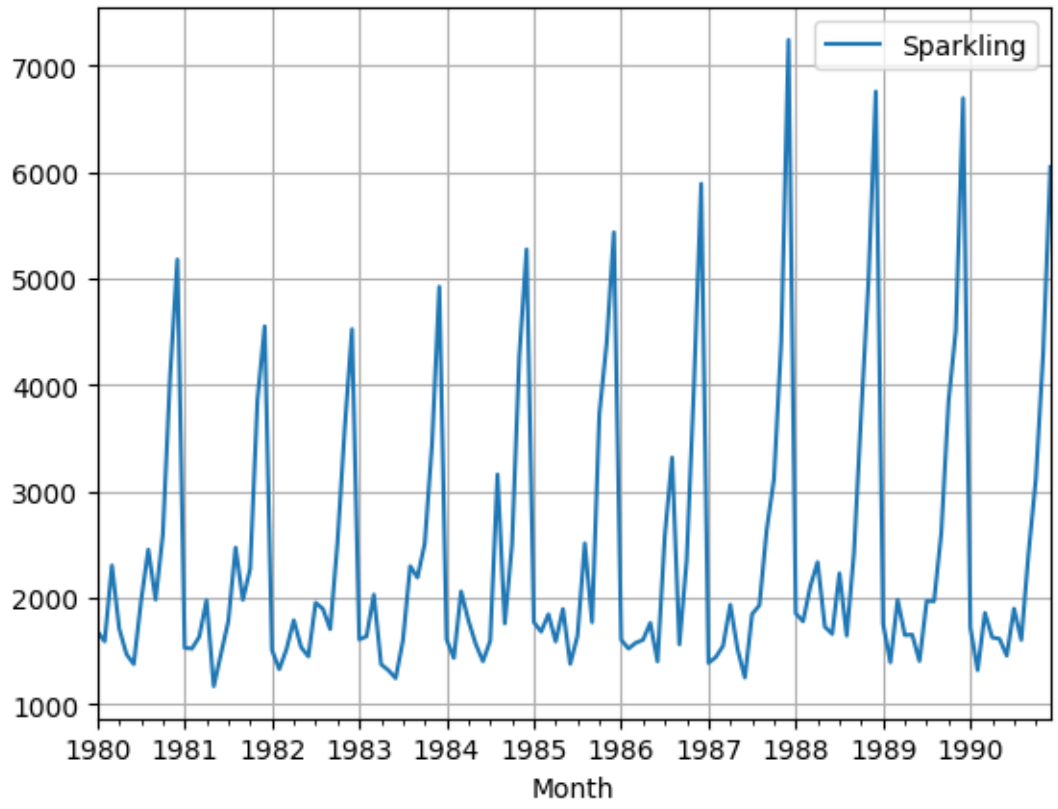
```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                                Rose   No. Observations:                  187
Model:             SARIMAX(2, 1, 2)x(2, 1, 2, 12)   Log Likelihood                -587.531
Date:                            Sat, 13 Apr 2024   AIC                           1193.062
Time:                                    09:24:04   BIC                           1219.976
Sample:                                01-31-1980   HQIC                          1203.997
                                     - 07-31-1995
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.8649      0.101     -8.561      0.000      -1.063      -0.667
ar.L2          0.0340      0.091      0.376      0.707      -0.143       0.211
ma.L1          0.0892    435.502      0.000      1.000    -853.480     853.658
ma.L2         -0.9108    396.671     -0.002      0.998    -778.372     776.550
ar.S.L12       0.0719      0.166      0.434      0.664      -0.252       0.396
ar.S.L24      -0.0357      0.017     -2.046      0.041      -0.070      -0.001
ma.S.L12      -0.6869      0.222     -3.093      0.002      -1.122      -0.252
ma.S.L24      -0.0549      0.151     -0.365      0.715      -0.350       0.240
sigma2       158.9059   6.92e+04      0.002      0.998   -1.35e+05    1.36e+05
==========================================================================================
Ljung-Box (L1) (Q):                   0.05   Jarque-Bera (JB):                10.12
Prob(Q):                              0.83   Prob(JB):                         0.01
Heteroskedasticity (H):               0.53   Skew:                             0.35
Prob(H) (two-sided):                  0.03   Kurtosis:                         4.08
==========================================================================================
```

Evaluate the model on the whole data and predict 12 months into the future



	Test RMSE Rose	Test RMSE Sparkling	Test MAPE Rose
RegressionOnTime	15.268955	1389.135175	NaN
NaiveModel	79.718773	3864.279352	NaN
SimpleAverageModel	53.460570	1275.081804	NaN
2pointTrailingMovingAverage	11.529278	813.400684	NaN
4pointTrailingMovingAverage	14.451403	1156.589694	NaN
6pointTrailingMovingAverage	14.566327	1283.927428	NaN
9pointTrailingMovingAverage	14.727630	1346.278315	NaN
Simple Exponential Smoothing	36.796225	1338.004623	NaN
Double Exponential Smoothing	15.270968	5291.879833	NaN
Triple Exponential Smoothing (Additive Season)	14.243240	378.626241	NaN
Triple Exponential Smoothing (Multiplicative Season)	19.113110	403.706228	NaN
ARIMA(2,1,3)	36.815186	NaN	75.843732
ARIMA(2,1,2)	36.871197	NaN	76.056213
SARIMA(3, 1, 1)(3, 0, 2, 12)	18.882146	NaN	36.376501
SARIMA(2,1,2)(3,1,2,12)	15.360839	NaN	22.964890

ARIMA / SARIMA Modelling on SPARKLING dataset

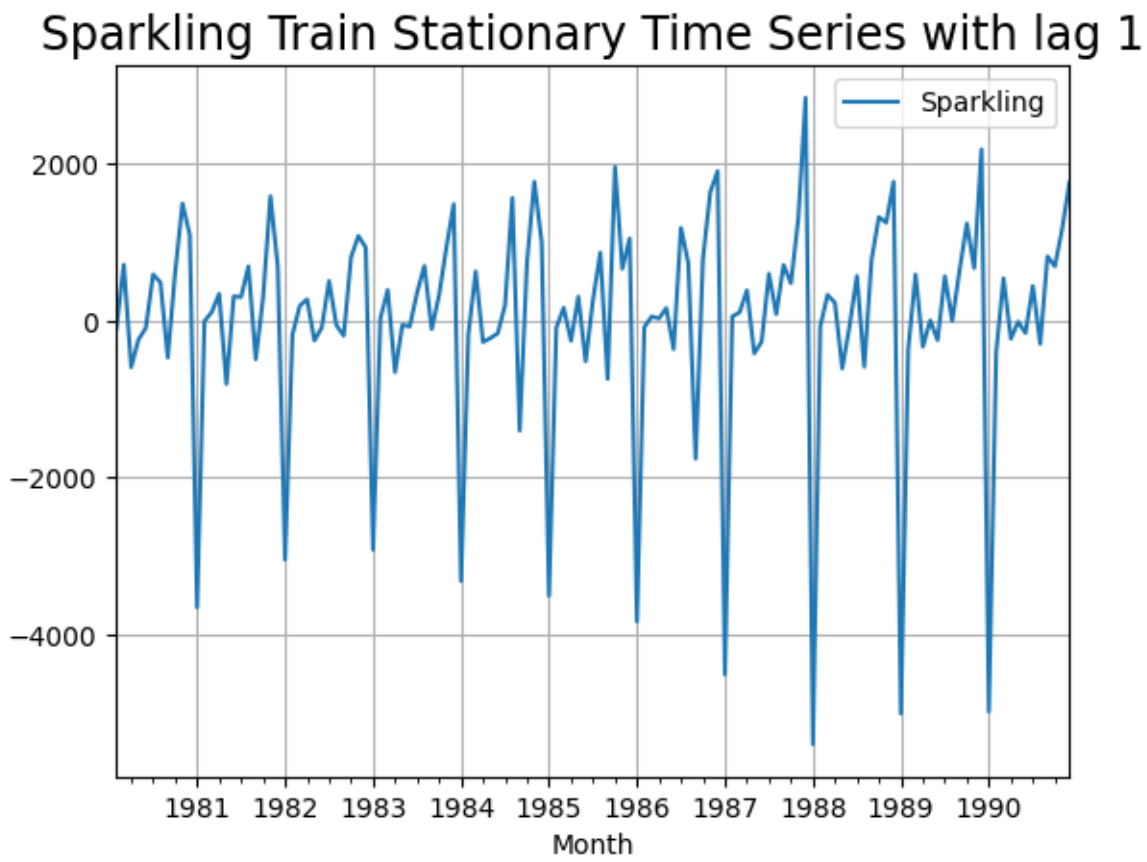


The training data is non-stationary at 95% confidence level. Let us take a first level of differencing to stationarize the Time Series.

```
[ ]  dftest = adfuller(strain.diff().dropna(),regression='ct')
     print('DF test statistic is %3.3f' %dftest[0])
     print('DF test p-value is' ,dftest[1])
     print('Number of lags used' ,dftest[2])
```

```
DF test statistic is -7.968
DF test p-value is 8.4792106555143e-11
Number of lags used 11
```

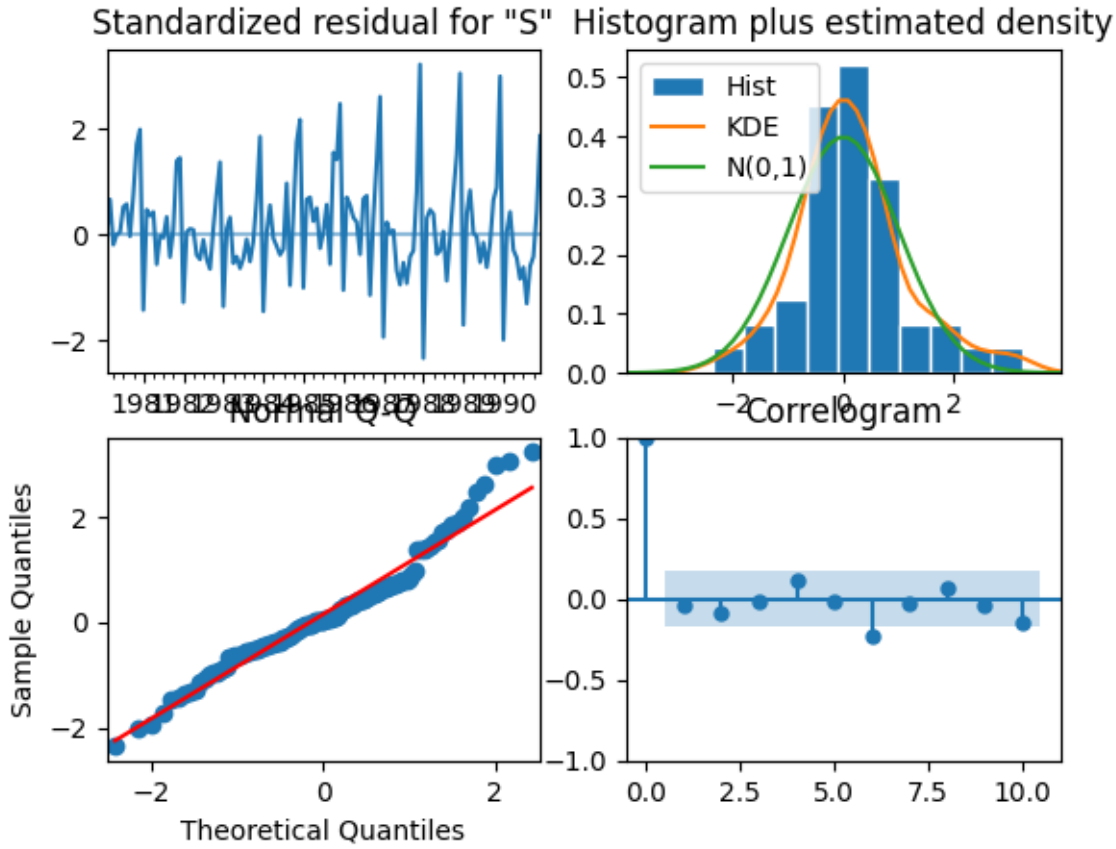
Training data is now Stationary Now, let us go ahead and plot the differenced training data.



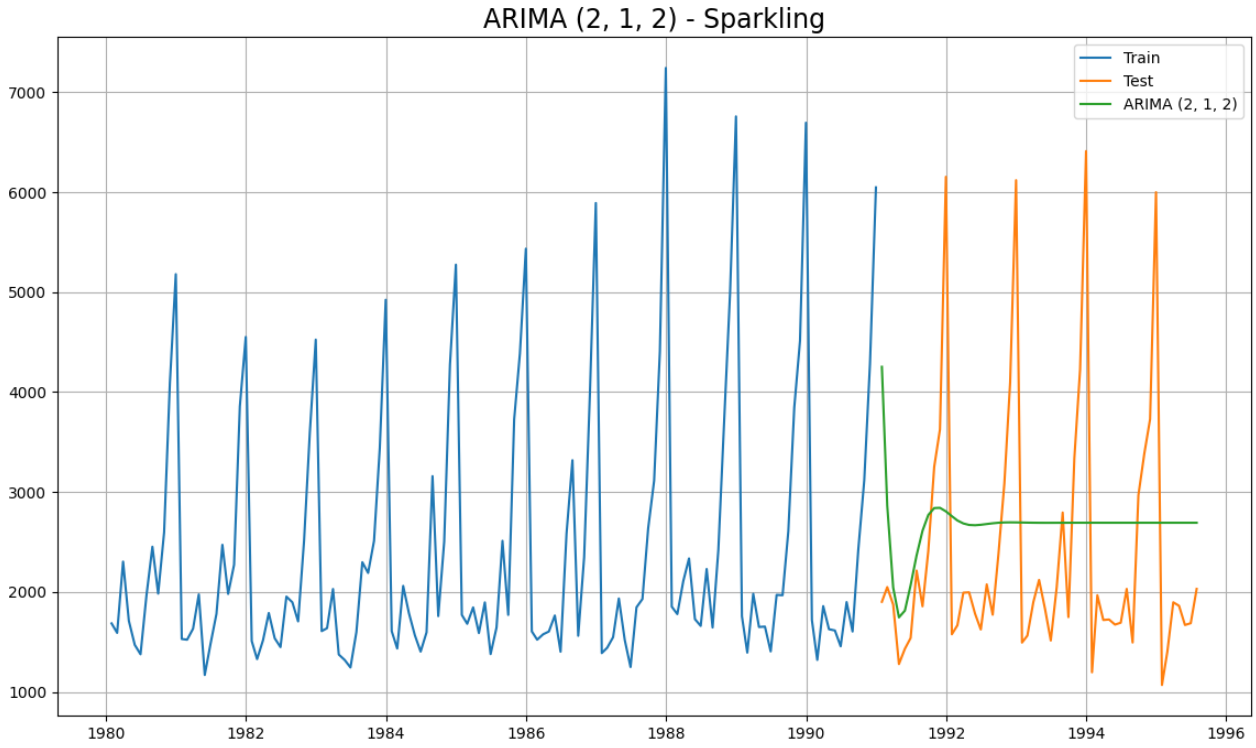
Build an Automated version of an ARIMA model for which the best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC) – SPARKLING

```
                               SARIMAX Results
==========================================================================================
Dep. Variable:                  Sparkling   No. Observations:                  132
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1101.755
Date:                Sat, 13 Apr 2024   AIC                           2213.509
Time:                        09:38:42   BIC                           2227.885
Sample:                  01-31-1980   HQIC                          2219.351
                       - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.3121      0.046     28.782      0.000       1.223       1.401
ar.L2         -0.5593      0.072     -7.740      0.000      -0.701      -0.418
ma.L1         -1.9917      0.109    -18.215      0.000      -2.206      -1.777
ma.L2          0.9999      0.110      9.108      0.000       0.785       1.215
sigma2      1.099e+06      2e-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
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 3.27e+27. Standard errors may be unstable.
```

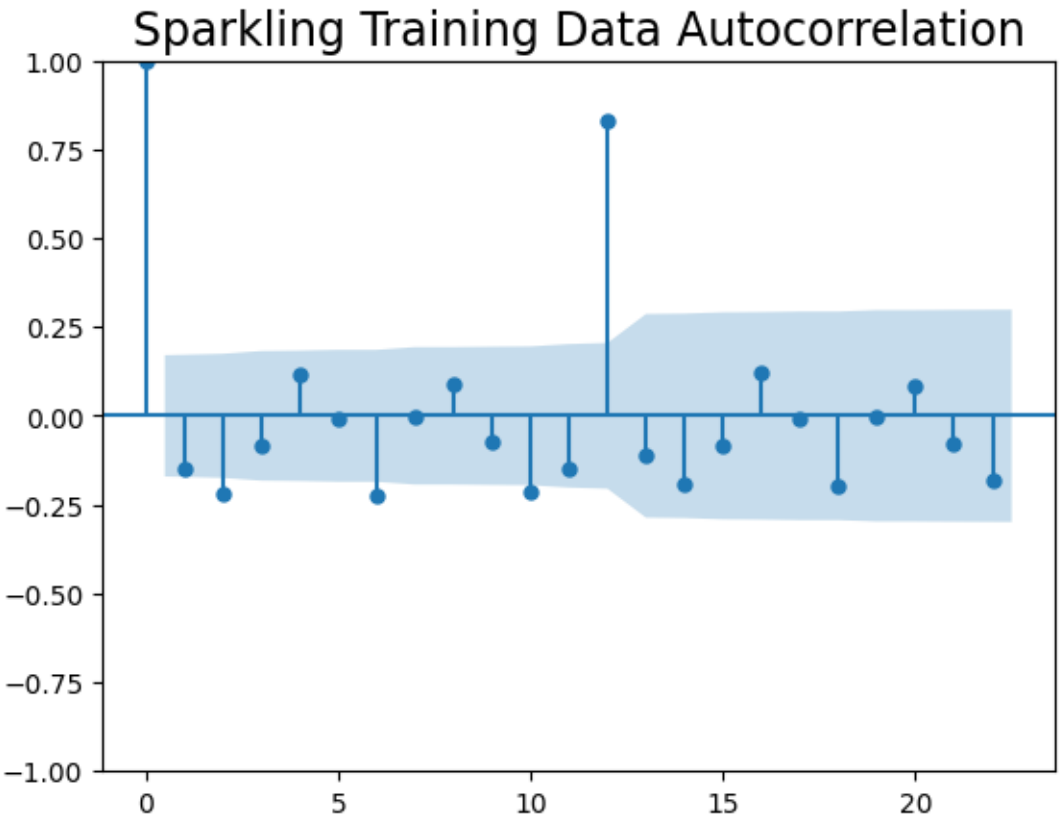
Diagnostics plot - Sparkling

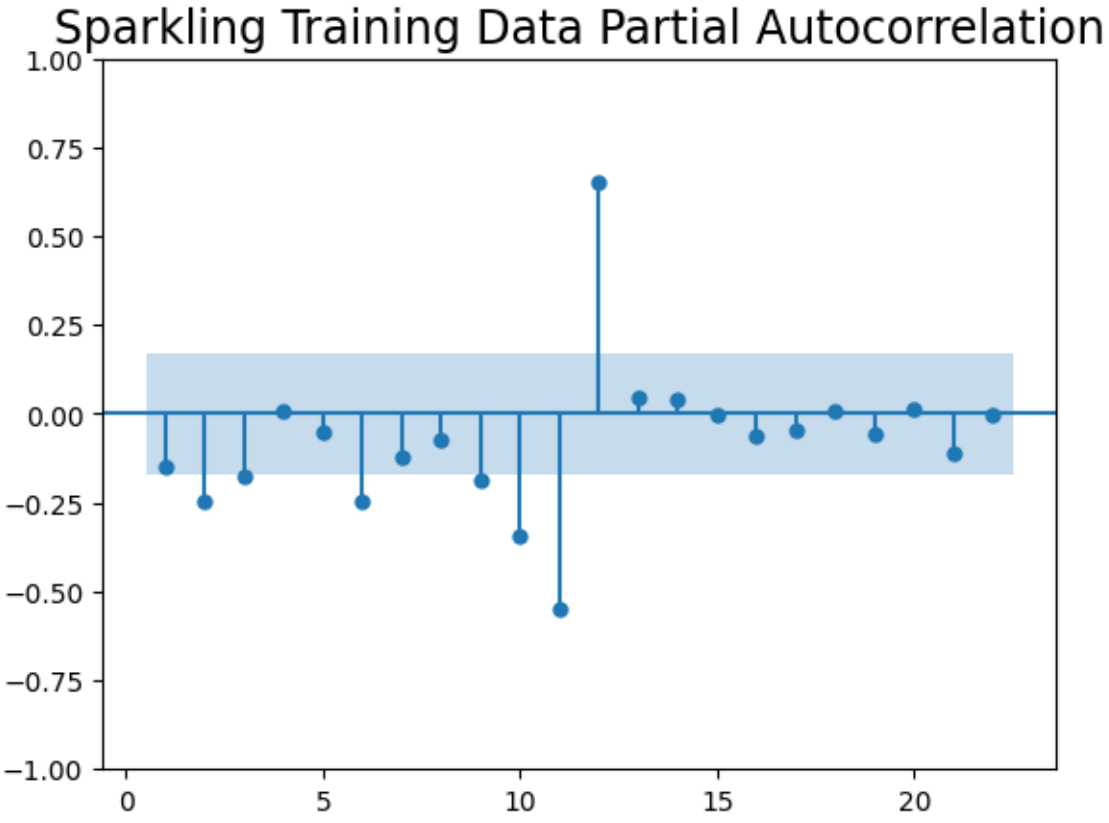


Predict on the Test Set using this model and evaluate the model.



Build a version of the ARIMA model for which the best parameters are selected by looking at the ACF and the PACF plots - SPARKLING

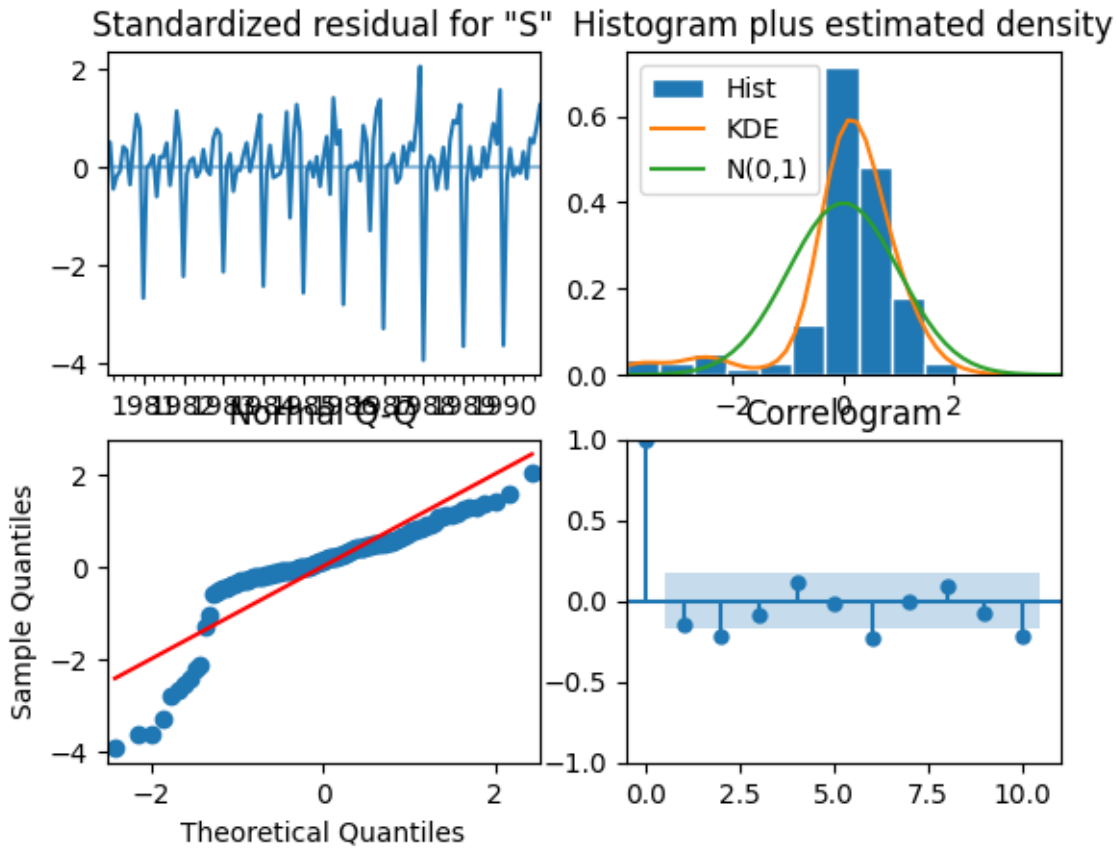




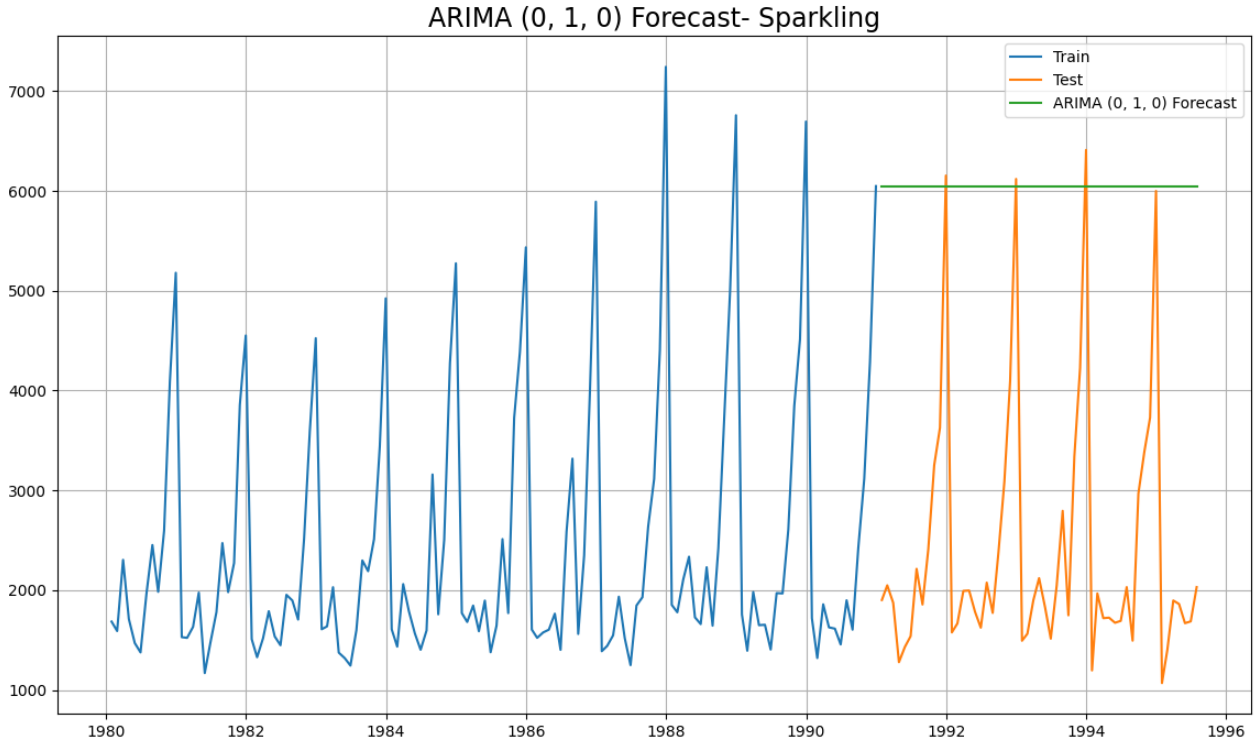
By looking at the above plots, we will take the value of p and q to be 0 and 0 respectively.

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(0, 1, 0)   Log Likelihood               -1132.832
Date:                Sat, 13 Apr 2024   AIC                           2267.663
Time:                        09:41:52   BIC                           2270.538
Sample:                    01-31-1980   HQIC                          2268.831
                         - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
sigma2      1.885e+06   1.29e+05     14.658      0.000    1.63e+06    2.14e+06
===================================================================================
Ljung-Box (L1) (Q):                   3.07   Jarque-Bera (JB):               198.83
Prob(Q):                              0.08   Prob(JB):                         0.00
Heteroskedasticity (H):               2.46   Skew:                            -1.92
Prob(H) (two-sided):                  0.00   Kurtosis:                         7.65
===================================================================================

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

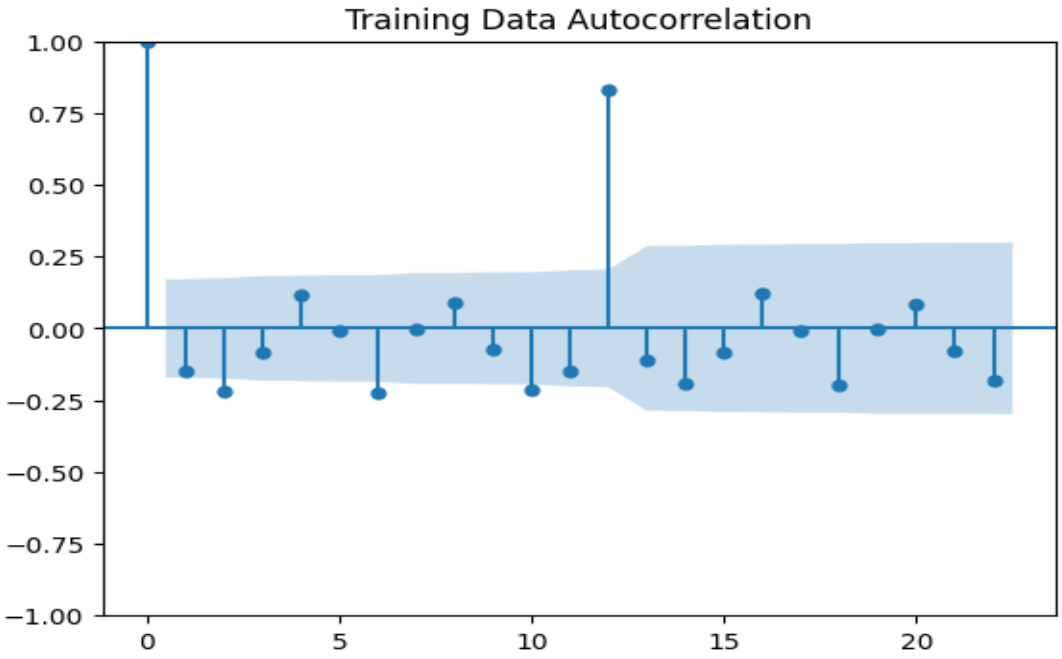
Let us analyse the residuals from the various diagnostics plot.

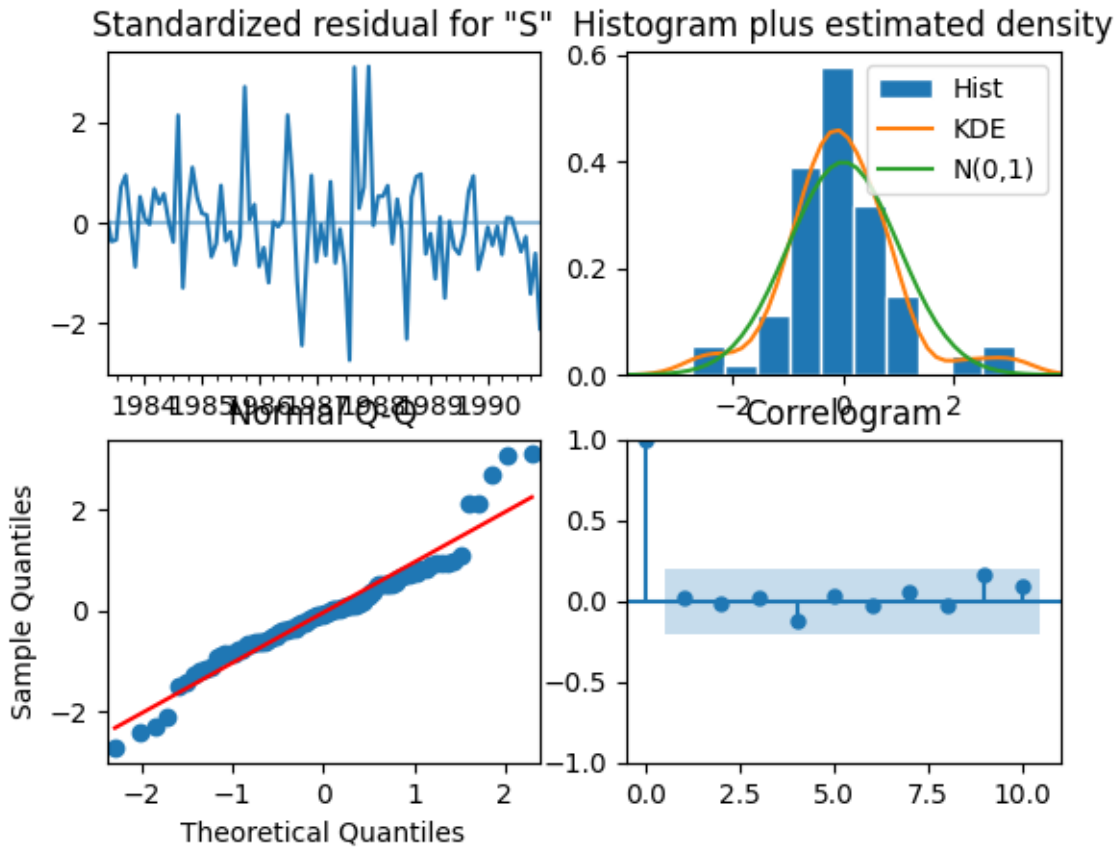


Predict on the Test Set using this model and evaluate the model.

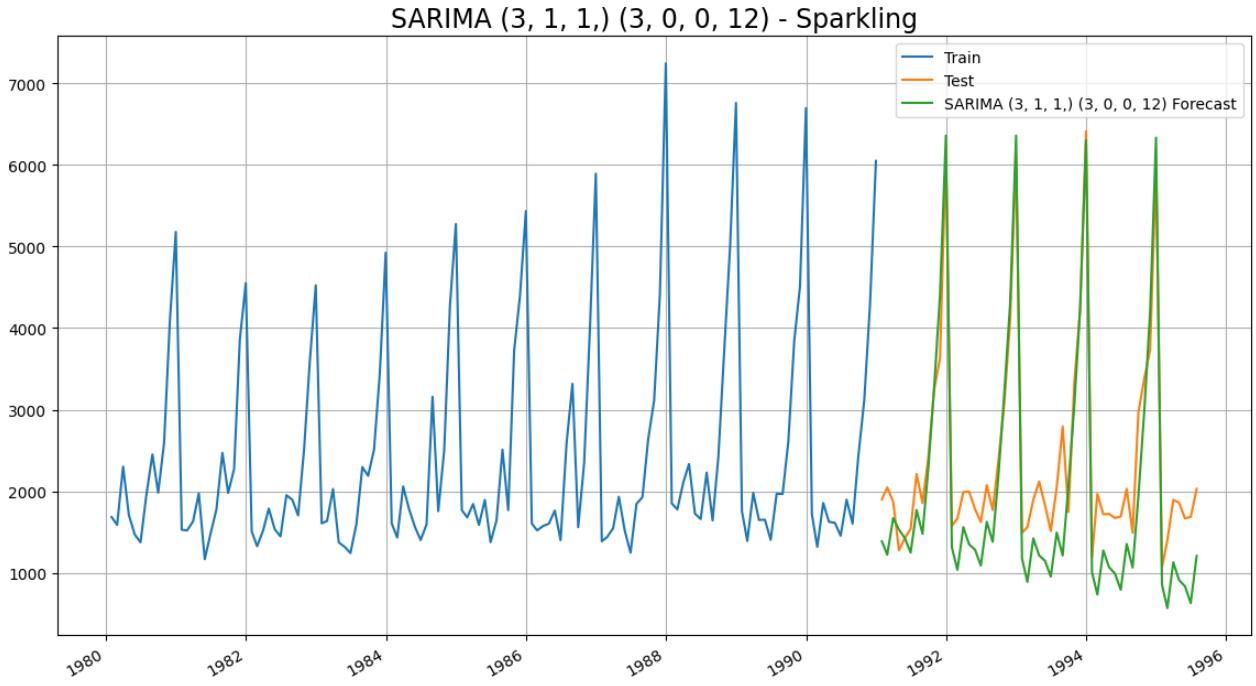


Build an Automated version of a SARIMA model for which the best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC) - SPARKLING

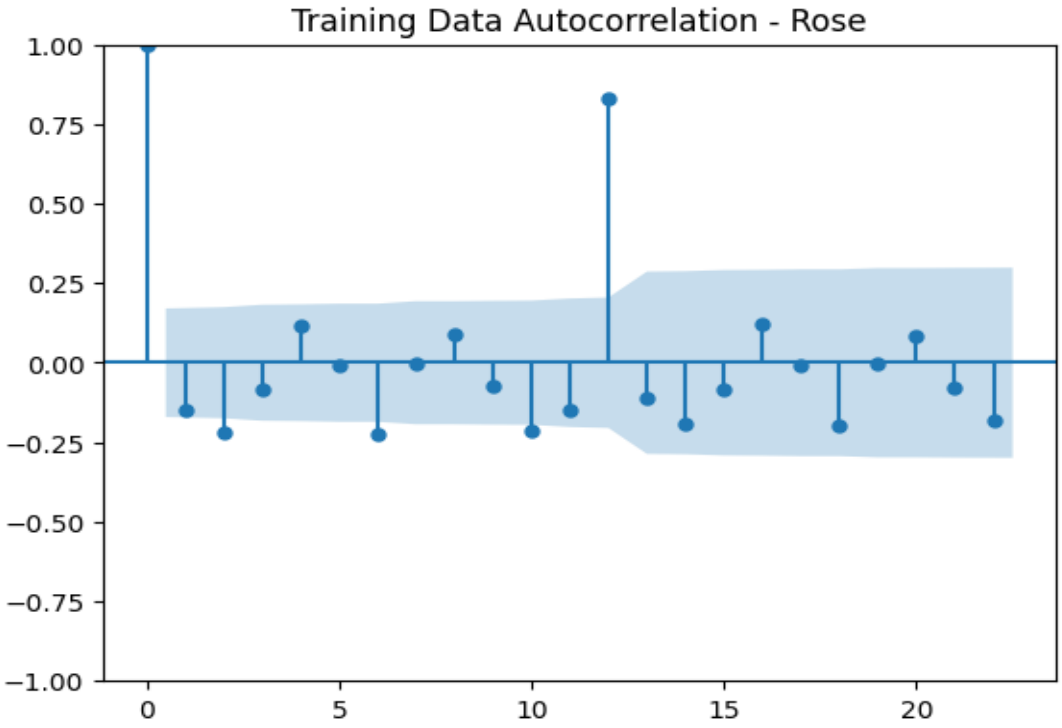


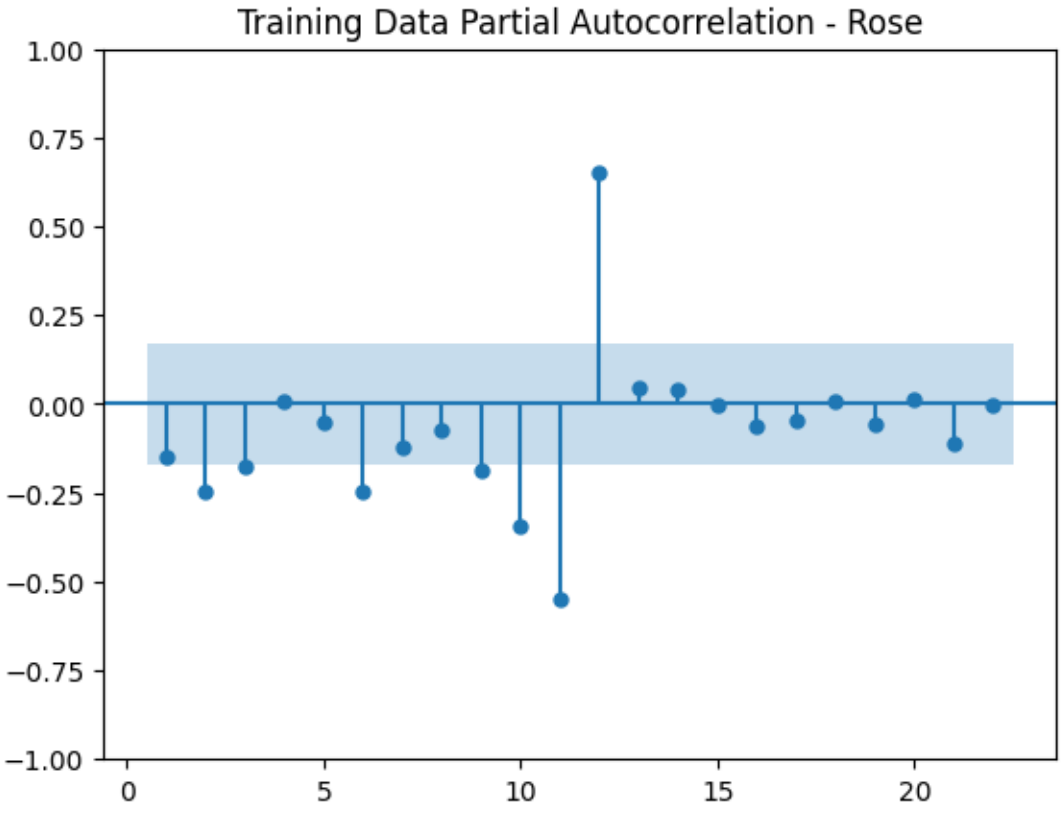


Predict on the Test Set using this model and evaluate the model.

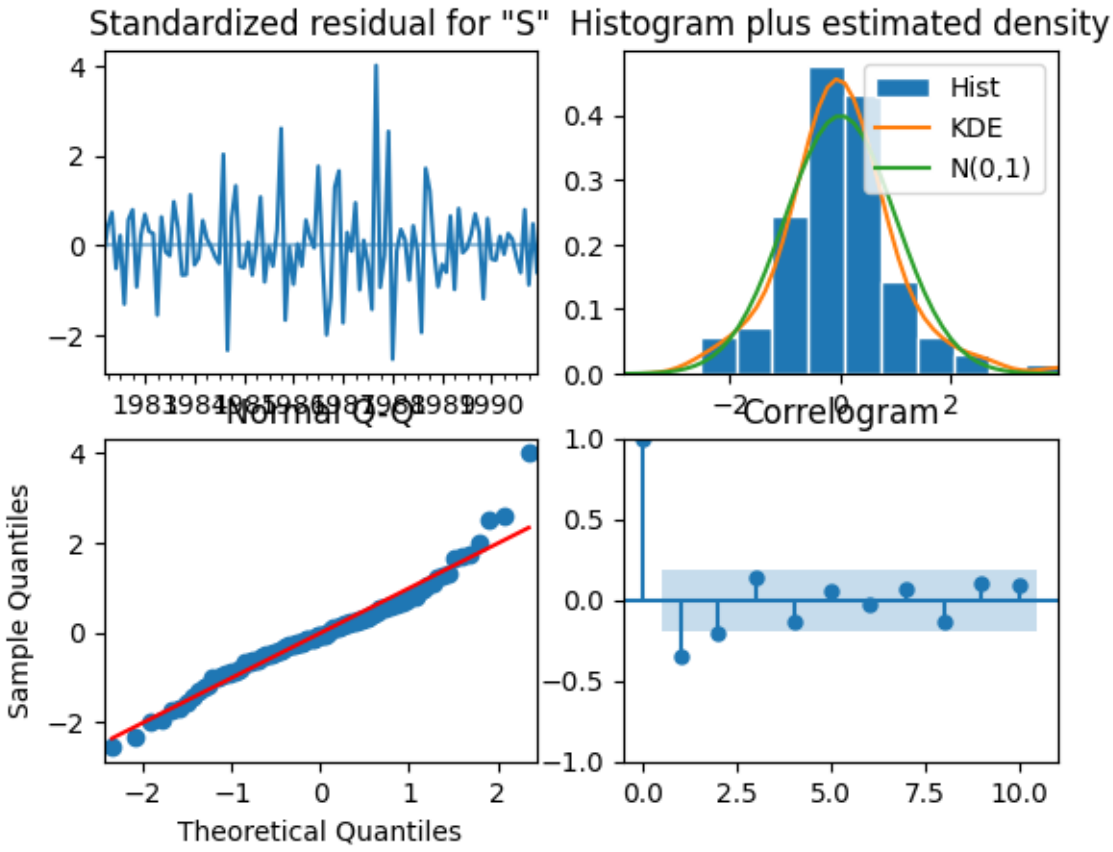


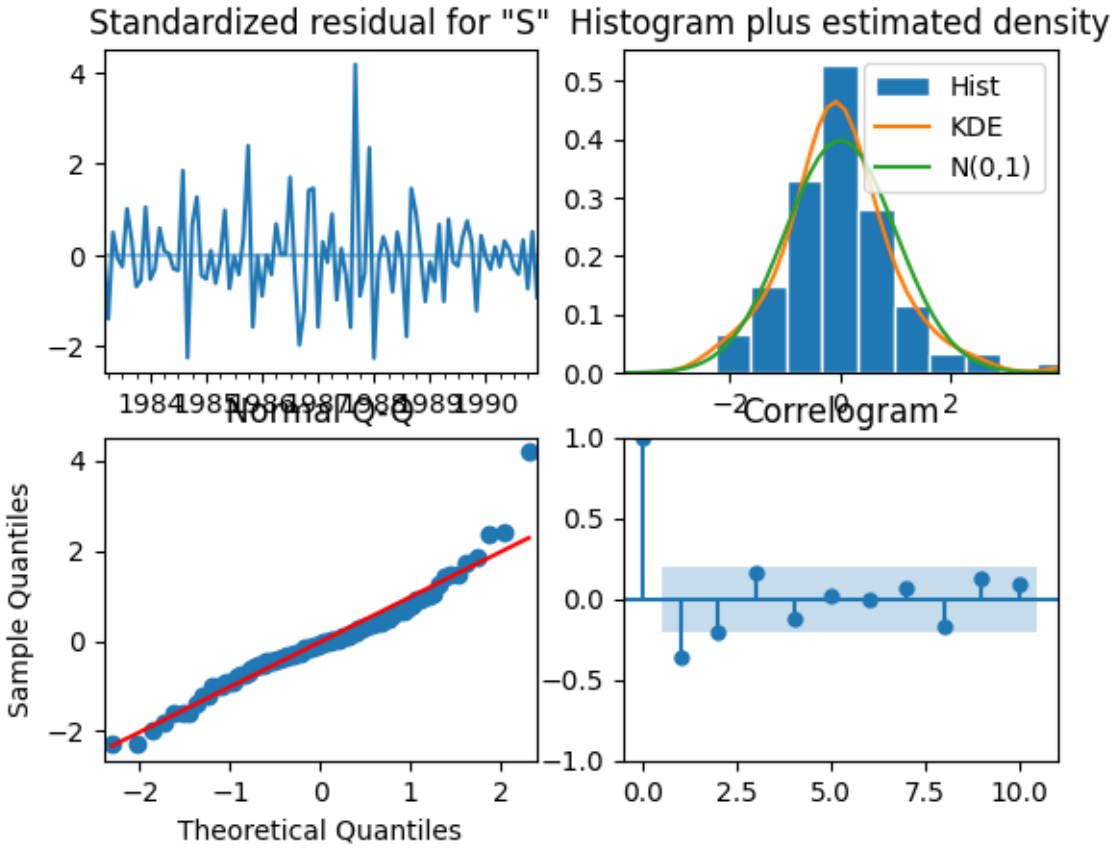
Build a version of the SARIMA model for which the best parameters are selected by looking at the ACF and the PACF plots. - Seasonality at 12 - SPARKLING

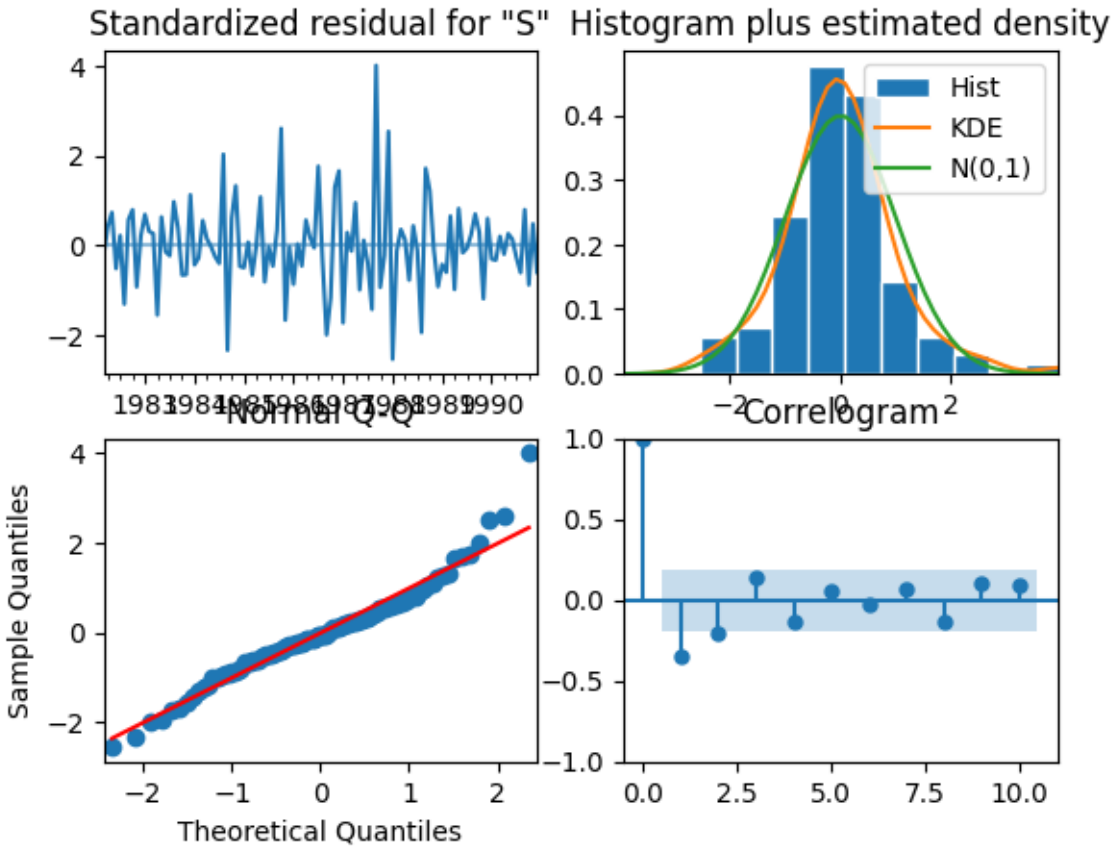


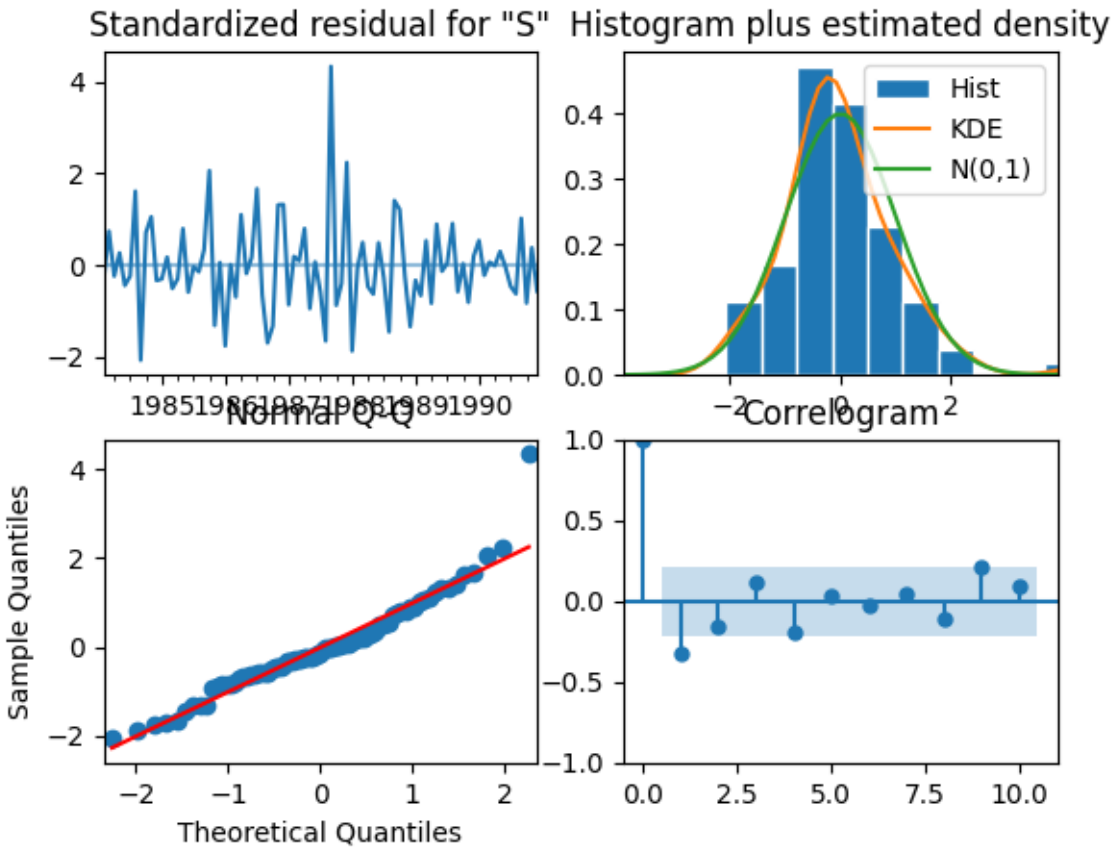


```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                           Sparkling   No. Observations:                  132
Model:             SARIMAX(0, 1, 0)x(1, 1, [1], 12)   Log Likelihood                -811.162
Date:                              Sat, 13 Apr 2024   AIC                           1628.324
Time:                                      10:05:37   BIC                           1636.315
Sample:                                  01-31-1980   HQIC                          1631.563
                                       - 12-31-1990
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.S.L12       0.1482      0.223      0.664      0.507      -0.289       0.586
ma.S.L12      -0.5732      0.217     -2.640      0.008      -0.999      -0.148
sigma2      2.577e+05   2.63e+04      9.806      0.000    2.06e+05    3.09e+05
===================================================================================
Ljung-Box (L1) (Q):                  13.54   Jarque-Bera (JB):                 27.17
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.73   Skew:                             0.59
Prob(H) (two-sided):                  0.36   Kurtosis:                         5.19
===================================================================================

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

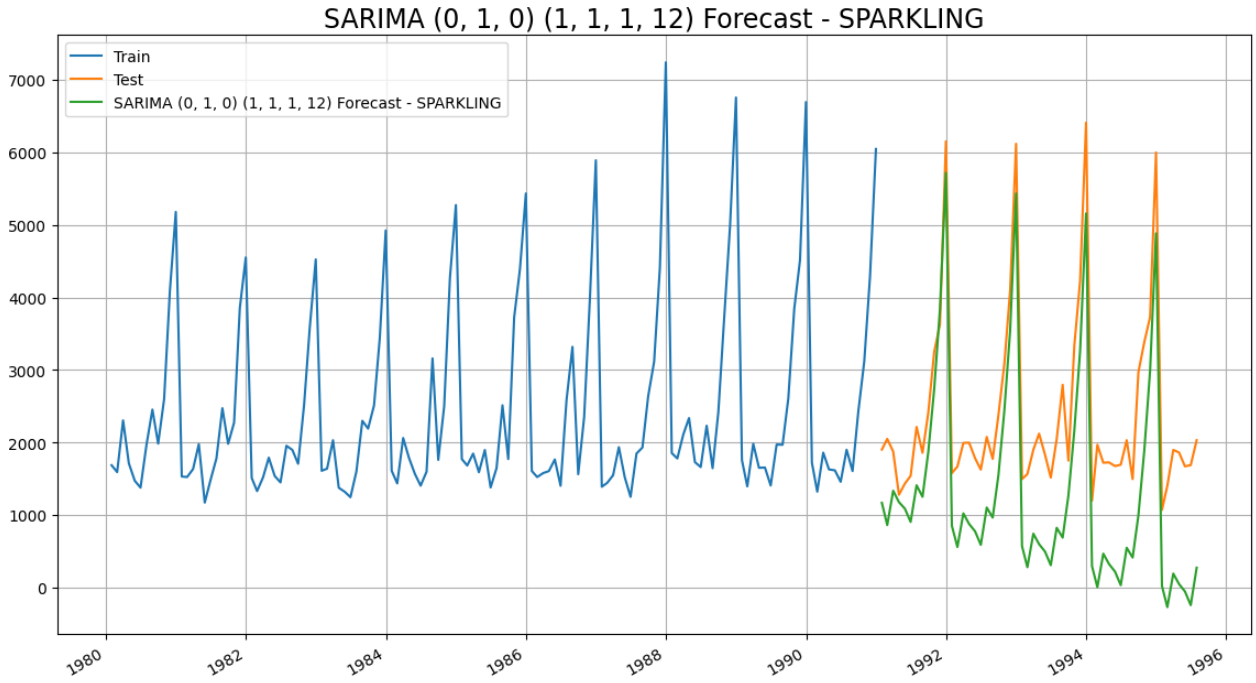


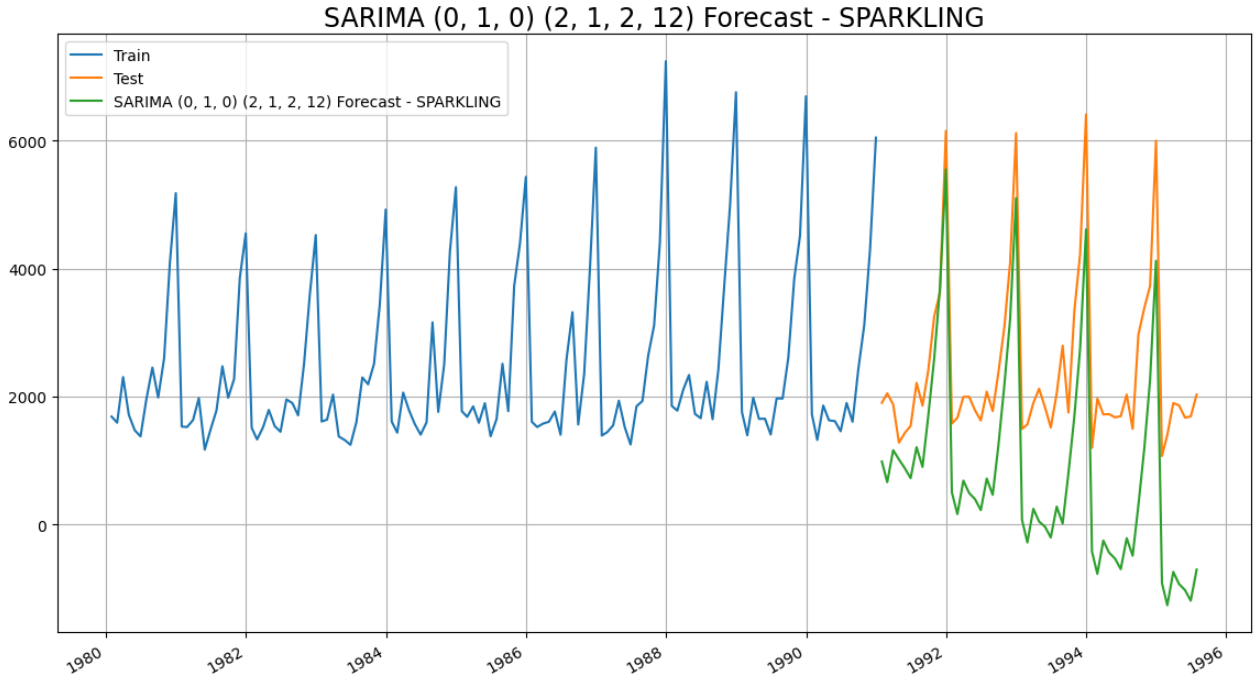

Standardized residual for "S"
Histogram plus estimated density
Hist
KDE
N(0,1)
Normal Q-Q
Correlogram
Sample Quantiles
Theoretical Quantiles

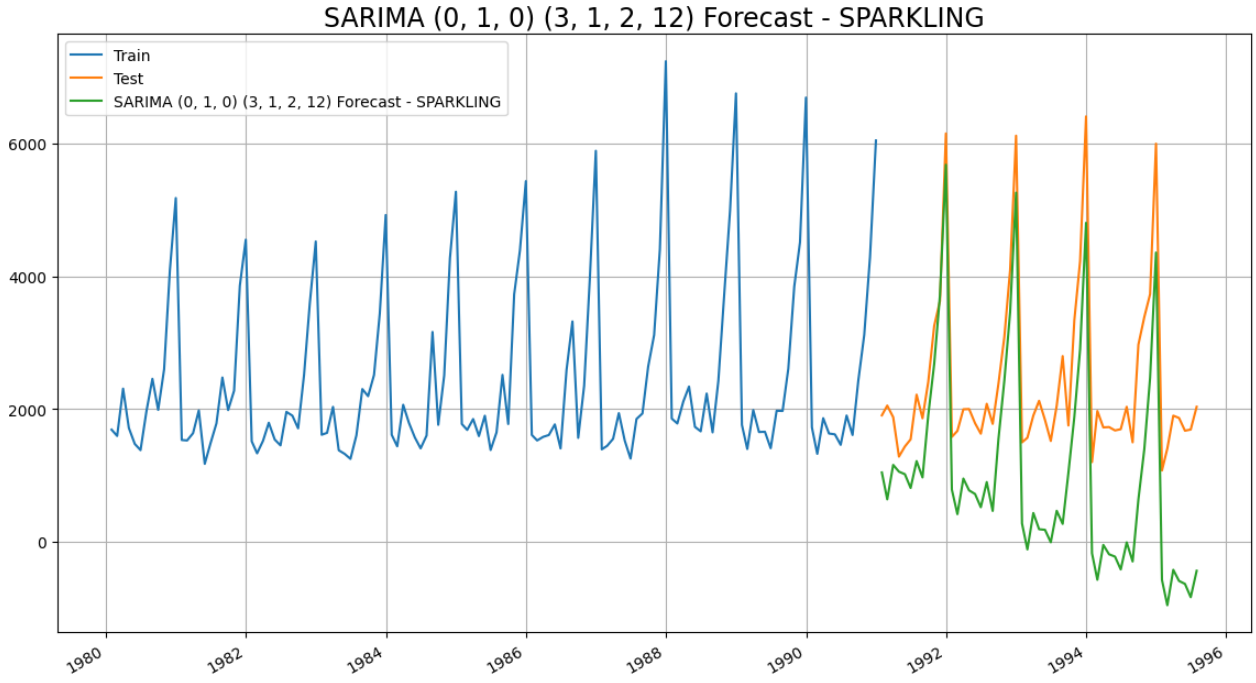

Standardized residual for "S"
Histogram plus estimated density
Hist
KDE
N(0,1)
Normal Q-Q
Correlogram
Sample Quantiles
Theoretical Quantiles



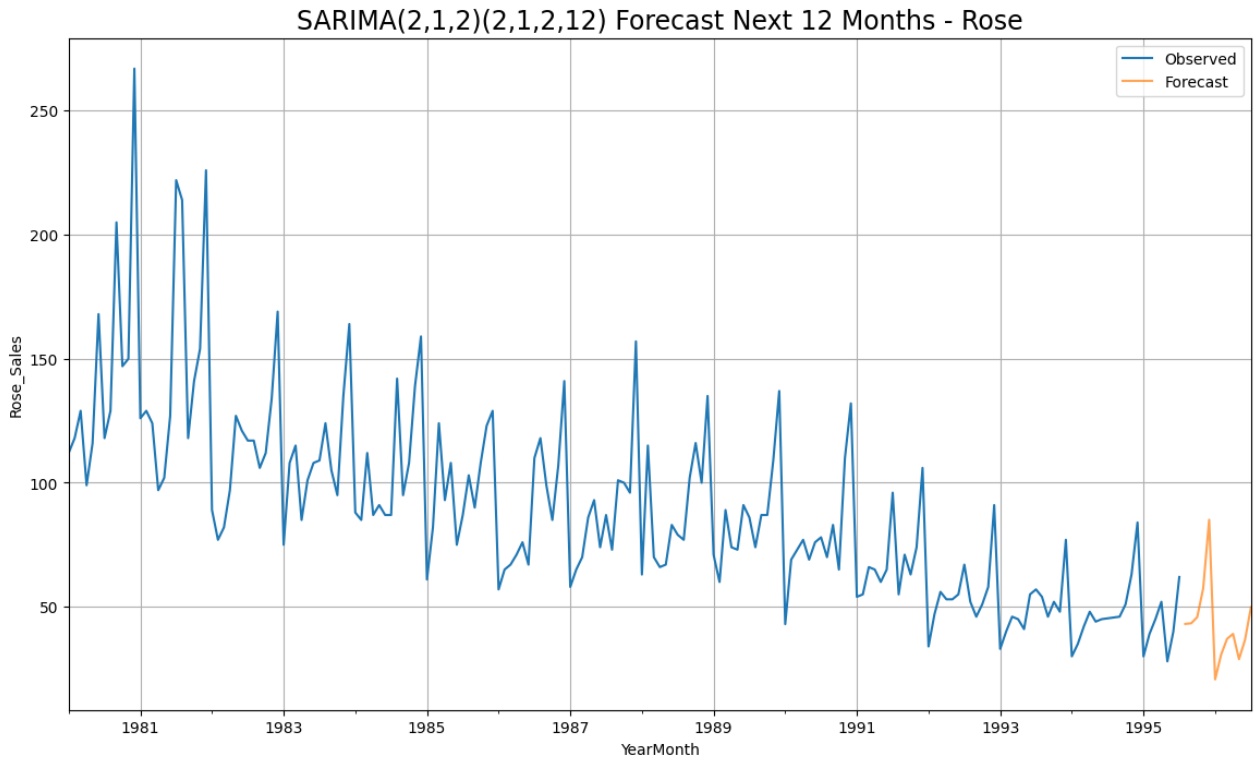
Predict on the Test Set using this model and evaluate the model.

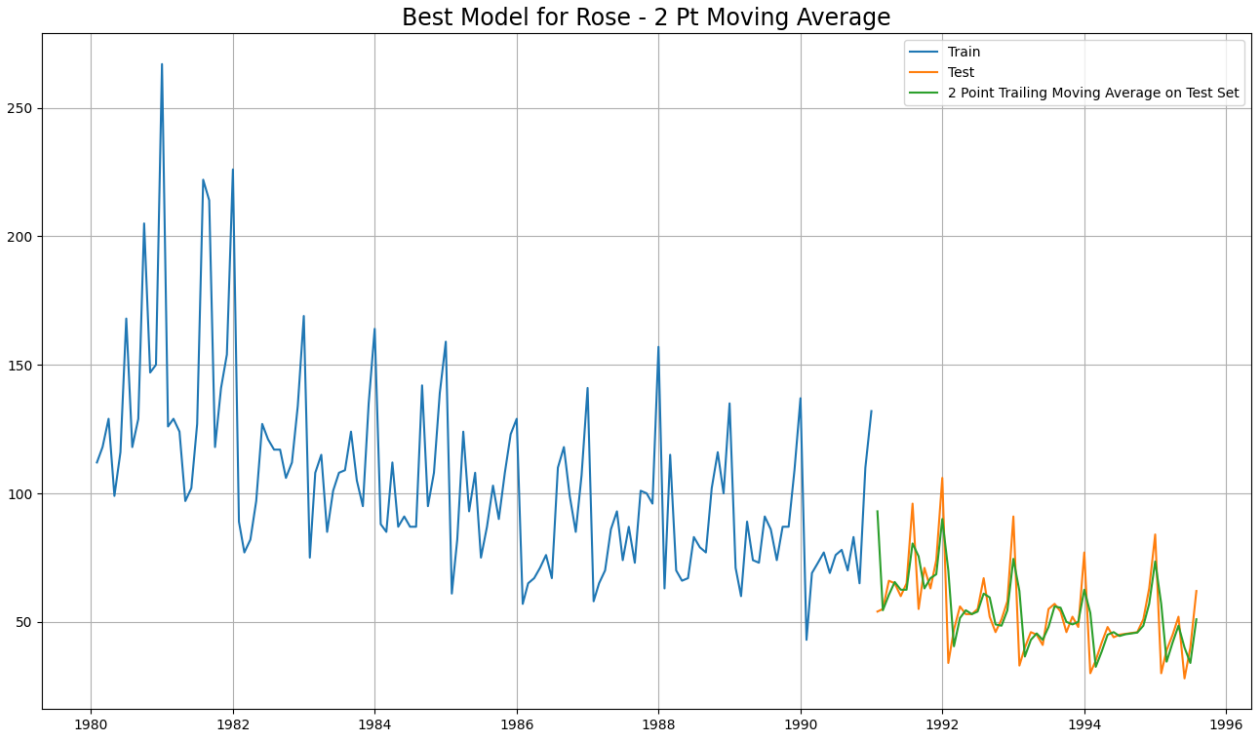



SARIMA (0, 1, 0) (2, 1, 2, 12) Forecast - SPARKLING
Train
Test
SARIMA (0, 1, 0) (2, 1, 2, 12) Forecast - SPARKLING
6000
4000
2000
0
1980
1982
1984
1986
1988
1990
1992
1994
1996


SARIMA (0, 1, 0) (3, 1, 2, 12) Forecast - SPARKLING
Train
Test
SARIMA (0, 1, 0) (3, 1, 2, 12) Forecast - SPARKLING
6000
4000
2000
0
1980
1982
1984
1986
1988
1990
1992
1994
1996

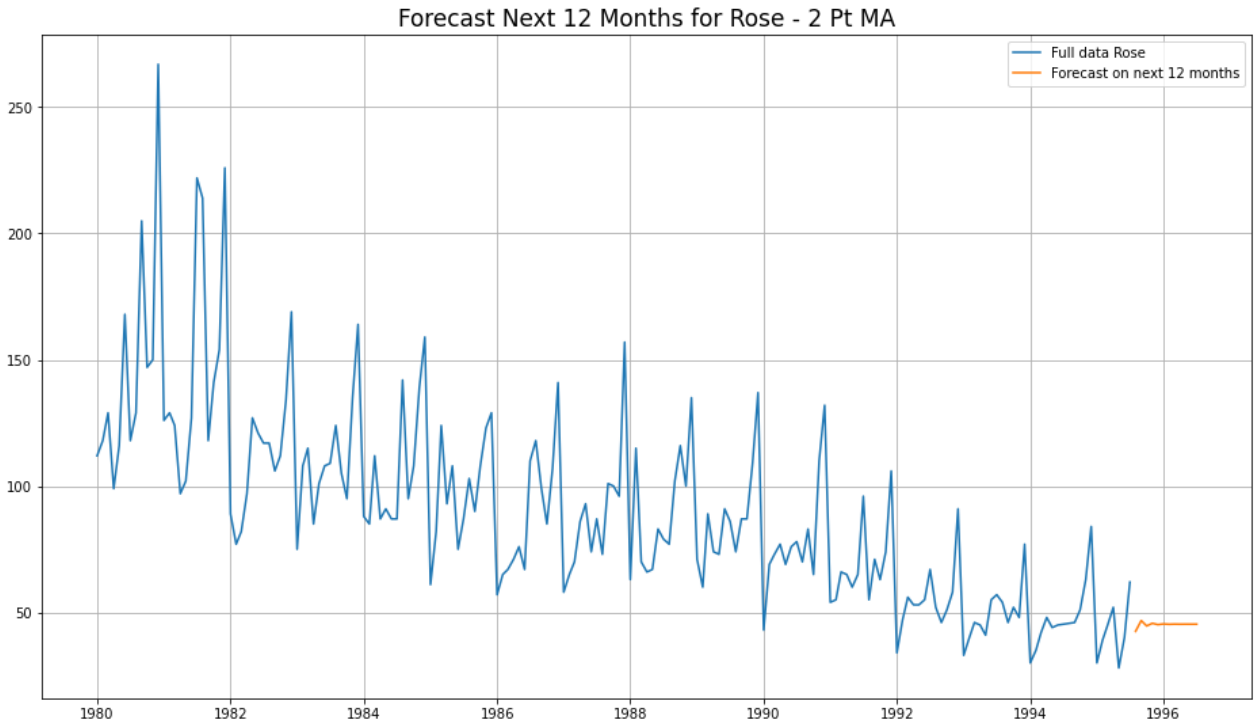
Build a version of the SARIMA model for which the best parameters are selected by looking at the ACF and the PACF plots. - Seasonality at 12 – ROSE



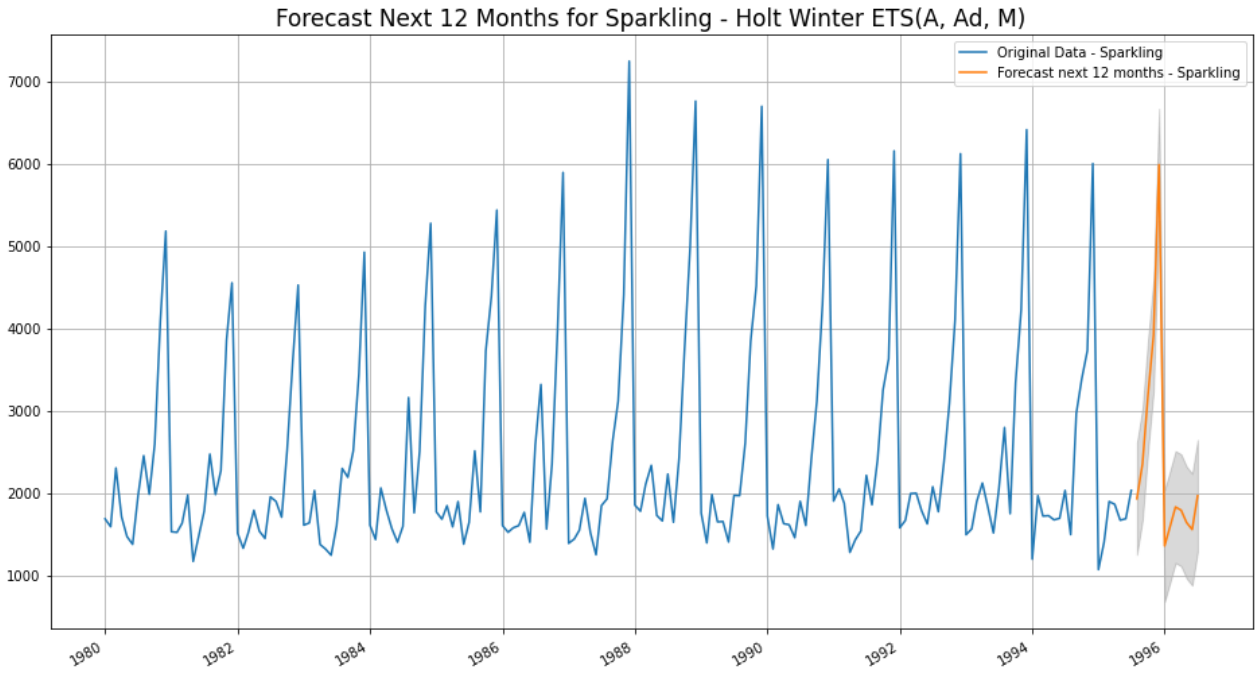


FORECAST ON NEXT 12 MONTHS – ROSE

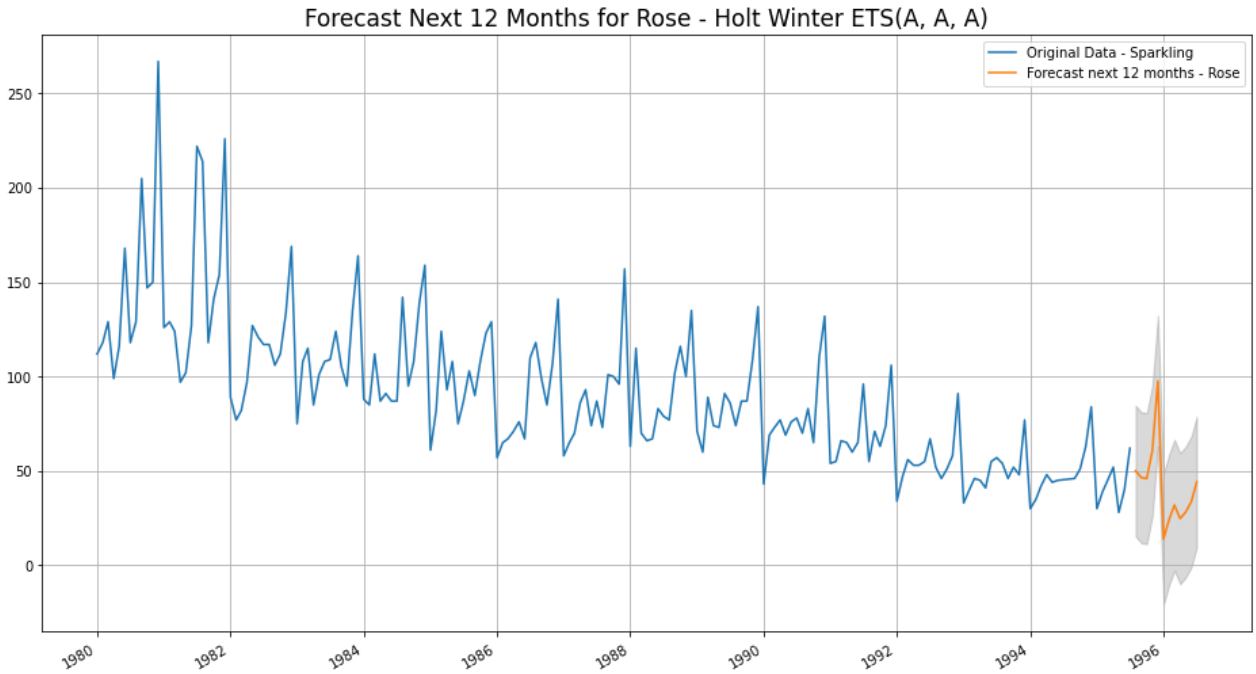
(Using 2 Pt Moving Average Model)



Holt-Winters - ETS(A, A, M) - Best Model for Sparkling Till Now



Building the second most optimum model on ROSE - TES ETS(A, A, A)



1.7 Insights and recommendations-

1. We have loaded the Sparkling dataset.csv to dfs dataset and rose.csv to dfr.

2. Performed EDA to check whether there are any missing values and outliers present in the dataset

3. After applying EDA we have split the data into train and test. Train is our sample data and Test are predicted data (actual data).Training Data is till the end of 1990. Test Data is from the beginning of 1991 to the last time stamp provided

4. Building different models and prediction the accuracy or doing model evaluation using RMSE-

- Linear Regression model
- Naïve Model
- Simple average model
- Simple exponential model
- Trailing moving average model
- Double exponential model
- Triple exponential model

5. Built the automated ARIMA/SARIMA MODEL

6, many different forecasting algorithms and analysis methods can be applied to extract the relevant information that is required. Regardless of using Autoregressive algorithms to determine the trend patterns for forecasting or the ARIMA model to deduce the correlation pattern of the data, it all depends on the application use cases and the complexity. Since most time series forecasting analyses are trivial, choosing the easiest and simplest model is the best way to look at it.

7. So we have read the problem described it.performed the various models and evaluated the models

8. In the month of December the sales for Sparkling Wine increases have more demand those other months.

9 .Matching season and customers demand and trend.

10. In-store Tastings and Events can attract the customers to your store and can increase sales.

11. We can also see in the year 1981, 1983 and 1994 the Wine sales in the month of October, November remained constant.