TIME SERIES FORECASTING

BUSINESS REPORT

THAKUR ARUN SINGH

SEPTEMBER 2021

This Business Report shall provide detailed explanation of how we approached each problem given in the assignment. It shall also provide relative resolution and explanation with regards to the problems

CONTENTS

Problem 1:

For this particular assignment, the data of different types of wine sales in the 20th century is to be analyzed. 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 analyze and forecast Wine Sales in the 20th century..

PROBLEM 1.1

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

Resolution:

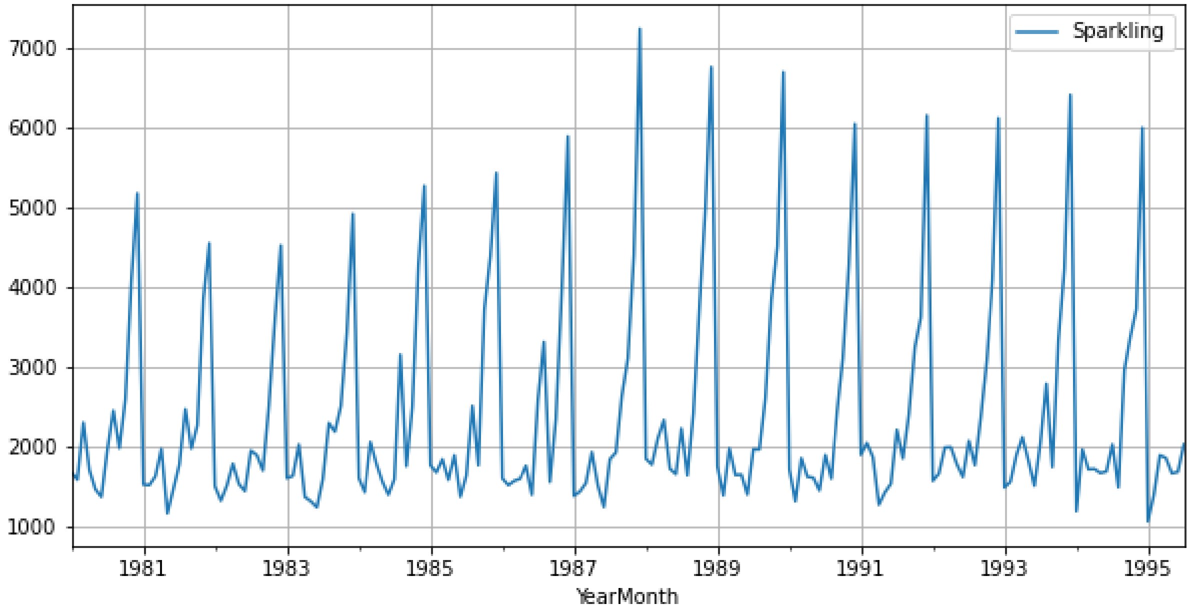
First, we import all the necessary libraries seaborn, numpy, pandas, sklearn etc to perform our analysis

Next, we import the data set "Sparkling" and "Rose"

Sparkling Dataset

```
DatetimeIndex(['1980-01-01', '1980-02-01', '1980-03-01', '1980-04-01',
               '1980-05-01', '1980-06-01', '1980-07-01', '1980-08-01',
               '1980-09-01', '1980-10-01',
               ...
               '1994-10-01', '1994-11-01', '1994-12-01', '1995-01-01',
               '1995-02-01', '1995-03-01', '1995-04-01', '1995-05-01',
               '1995-06-01', '1995-07-01'],
              dtype='datetime64[ns]', name='YearMonth', length=187, freq=None)
```

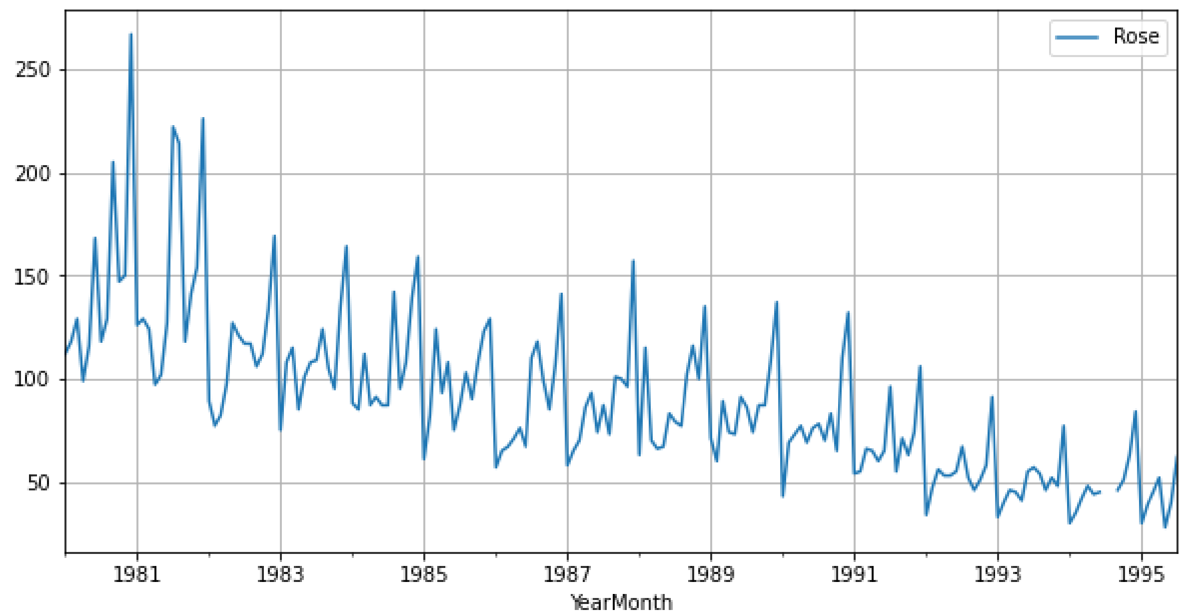
	Sparkling
YearMonth	
1980-01-01	1686
1980-02-01	1591
1980-03-01	2304
1980-04-01	1712
1980-05-01	1471



Rose Dataset

```
DatetimeIndex(['1980-01-01', '1980-02-01', '1980-03-01', '1980-04-01',
               '1980-05-01', '1980-06-01', '1980-07-01', '1980-08-01',
               '1980-09-01', '1980-10-01',
               ...
               '1994-10-01', '1994-11-01', '1994-12-01', '1995-01-01',
               '1995-02-01', '1995-03-01', '1995-04-01', '1995-05-01',
               '1995-06-01', '1995-07-01'],
              dtype='datetime64[ns]', name='YearMonth', length=187, freq=None)
```

	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



PROBLEM 1.2

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

Resolution:

Sparkling:

	Sparkling
count	187.000
mean	2402.417
std	1295.112
min	1070.000
25%	1605.000
50%	1874.000
75%	2549.000
max	7242.000

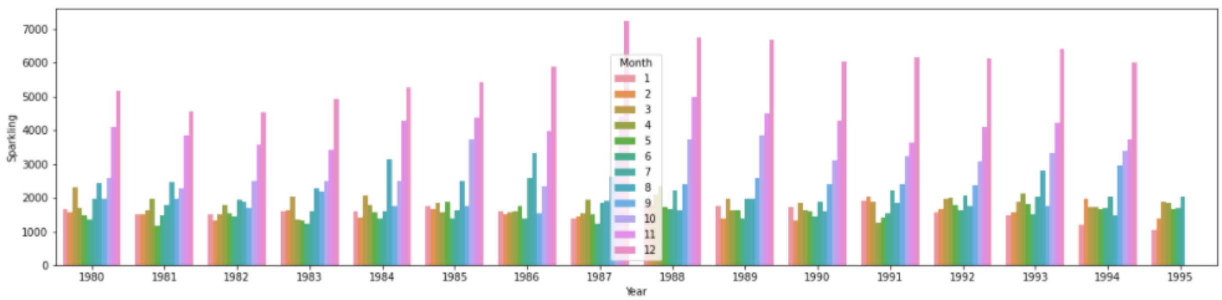
Rose:

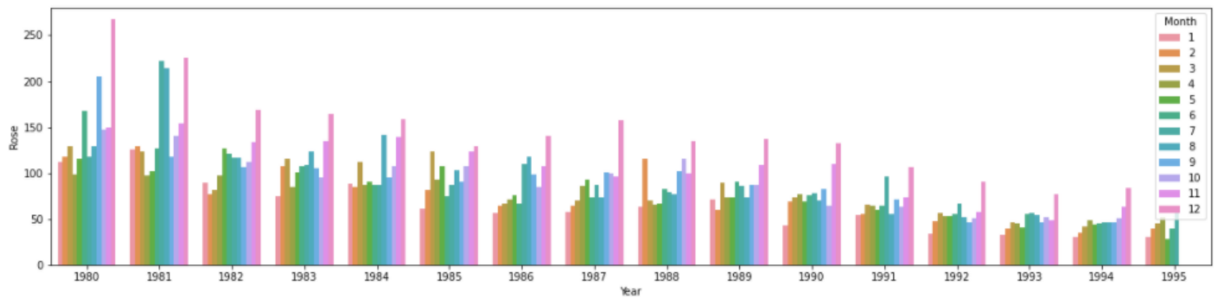
	Rose
count	185.000
mean	90.395
std	39.175
min	28.000
25%	63.000
50%	86.000
75%	112.000
max	267.000

```
Sparkling    0
dtype: int64

Rose    2
dtype: int64

Rose    0
dtype: int64

(187, 1)
```

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

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



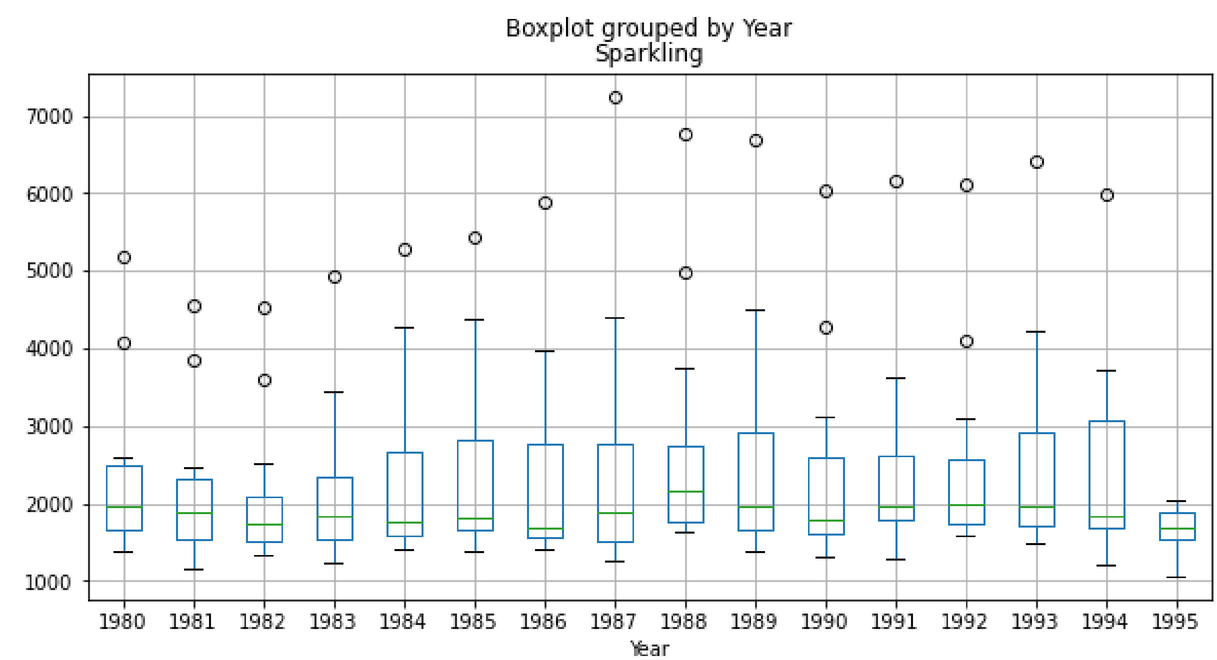


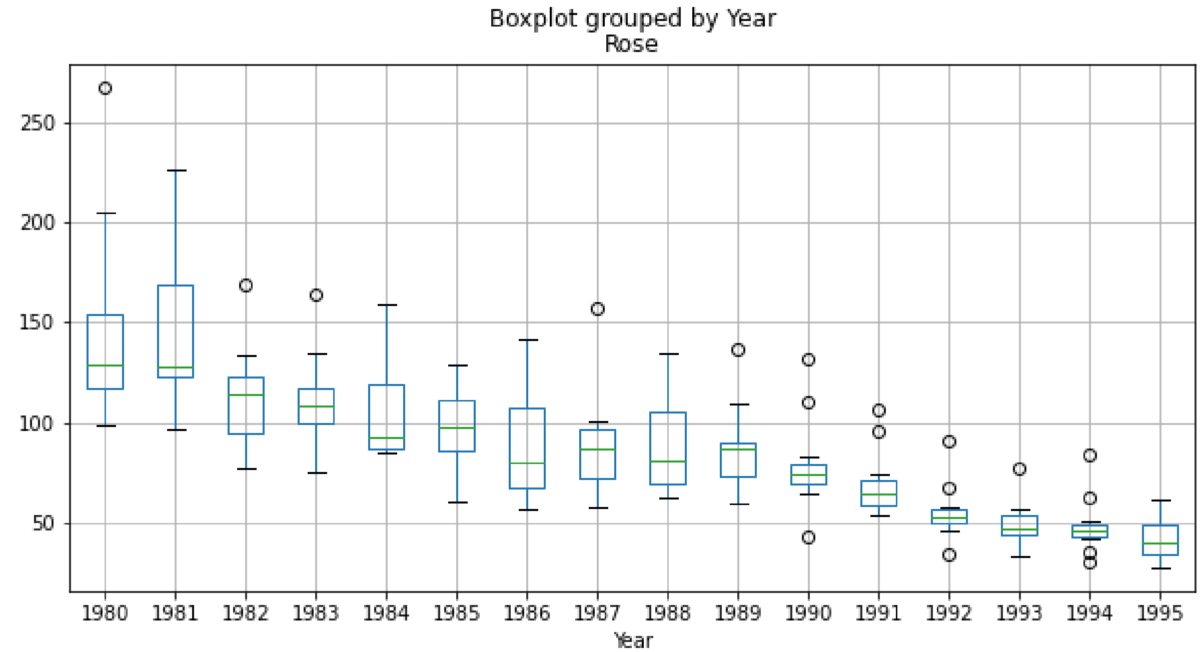
Below Pivot shows the sales made for a month in particular year:

Sparkling												
Month	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Year												
1980	1686	1591	2304	1712	1471	1377	1966	2453	1984	2596	4087	5179
1981	1530	1523	1633	1976	1170	1480	1781	2472	1981	2273	3857	4551
1982	1510	1329	1518	1790	1537	1449	1954	1897	1706	2514	3593	4524
1983	1609	1638	2030	1375	1320	1245	1600	2298	2191	2511	3440	4923
1984	1609	1435	2061	1789	1567	1404	1597	3159	1759	2504	4273	5274
1985	1771	1682	1846	1589	1896	1379	1645	2512	1771	3727	4388	5434
1986	1606	1523	1577	1605	1765	1403	2584	3318	1562	2349	3987	5891
1987	1389	1442	1548	1935	1518	1250	1847	1930	2638	3114	4405	7242
1988	1853	1779	2108	2336	1728	1661	2230	1645	2421	3740	4988	6757
1989	1757	1394	1982	1650	1654	1406	1971	1968	2608	3845	4514	6694
1990	1720	1321	1859	1628	1615	1457	1899	1605	2424	3116	4286	6047
1991	1902	2049	1874	1279	1432	1540	2214	1857	2408	3252	3627	6153
1992	1577	1667	1993	1997	1783	1625	2076	1773	2377	3088	4096	6119
1993	1494	1564	1898	2121	1831	1515	2048	2795	1749	3339	4227	6410
1994	1197	1968	1720	1725	1674	1693	2031	1495	2968	3385	3729	5999
1995	1070	1402	1897	1862	1670	1688	2031	NaN	NaN	NaN	NaN	NaN

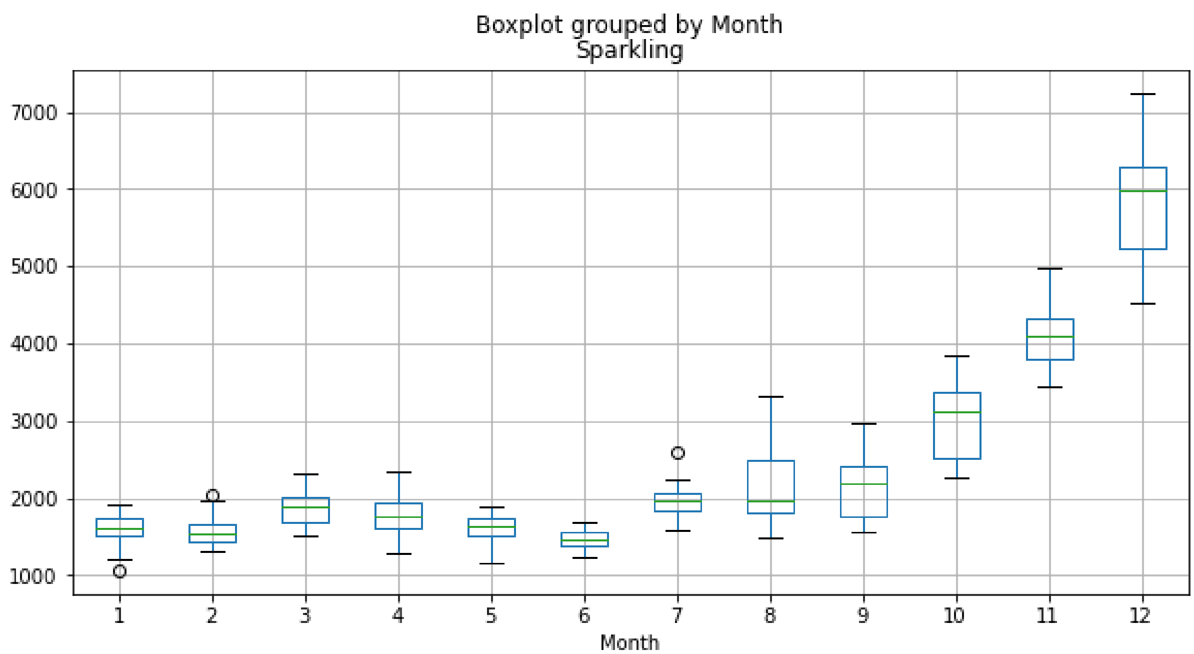
Rose												
Month	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Year												
1980	112	118	129	99	116	168	118	129	205	147	150	267
1981	126	129	124	97	102	127	222	214	118	141	154	226
1982	89	77	82	97	127	121	117	117	106	112	134	169
1983	75	108	115	85	101	108	109	124	105	95	135	164
1984	88	85	112	87	91	87	87	142	95	108	139	159
1985	61	82	124	93	108	75	87	103	90	108	123	129
1986	57	65	67	71	76	67	110	118	99	85	107	141
1987	58	65	70	86	93	74	87	73	101	100	96	157
1988	63	115	70	66	67	83	79	77	102	116	100	135
1989	71	60	89	74	73	91	86	74	87	87	109	137
1990	43	69	73	77	69	76	78	70	83	65	110	132
1991	54	55	66	65	60	65	96	55	71	63	74	106
1992	34	47	56	53	53	55	67	52	46	51	58	91
1993	33	40	46	45	41	55	57	54	46	52	48	77
1994	30	35	42	48	44	45	46	46	46	51	63	84
1995	30	39	45	52	28	40	62	NaN	NaN	NaN	NaN	NaN

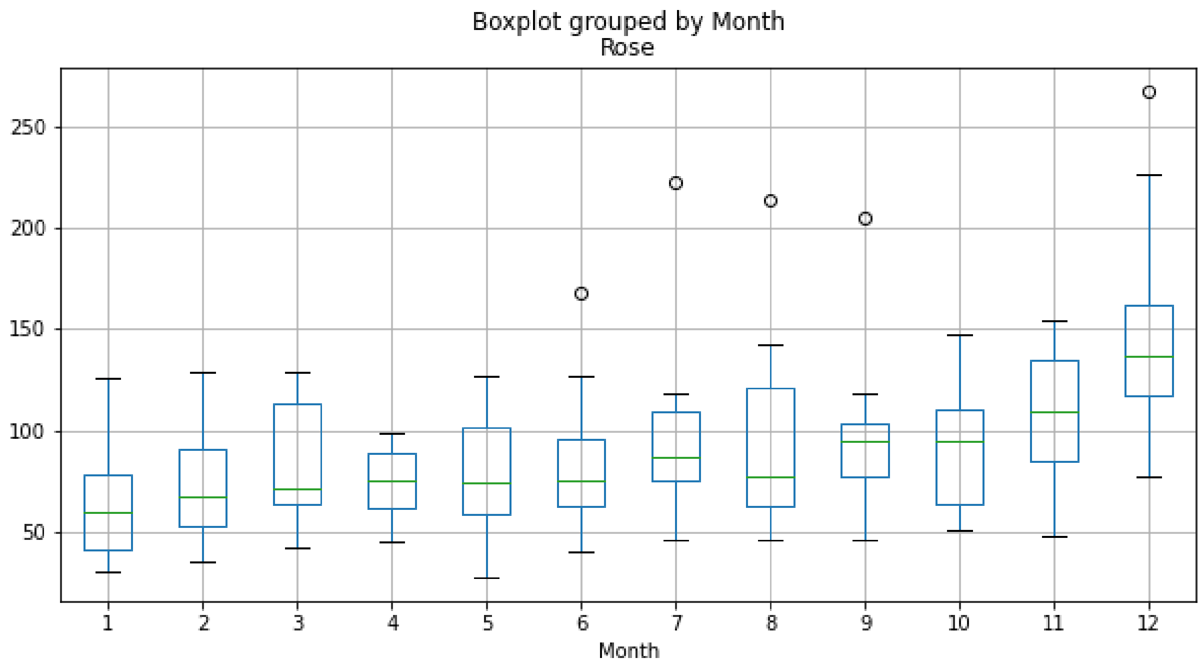
Yearly Boxplots





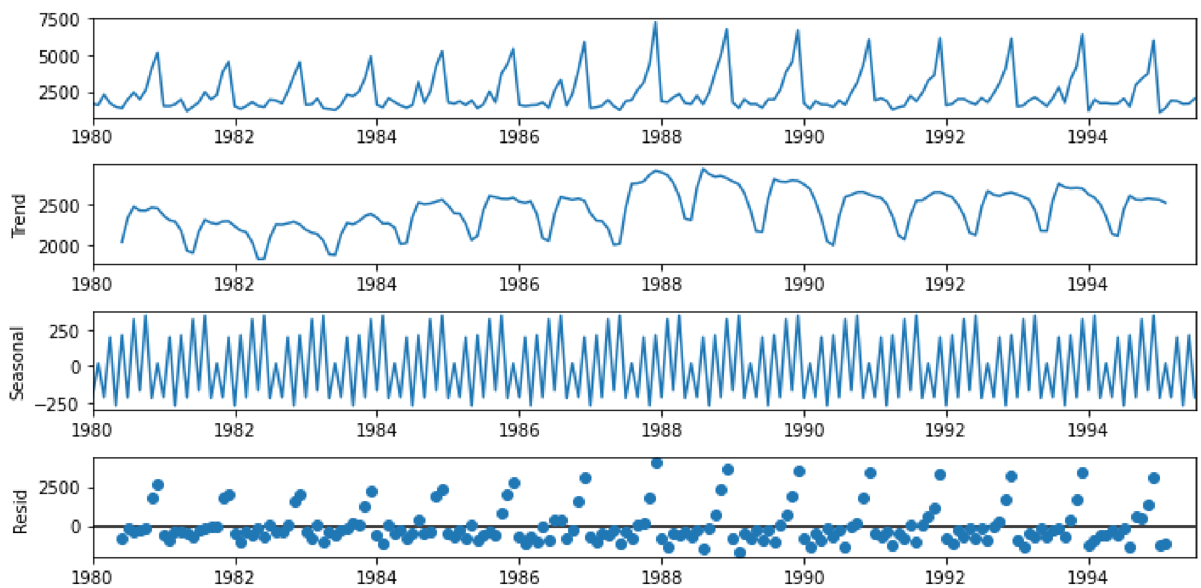
Monthly Boxplots



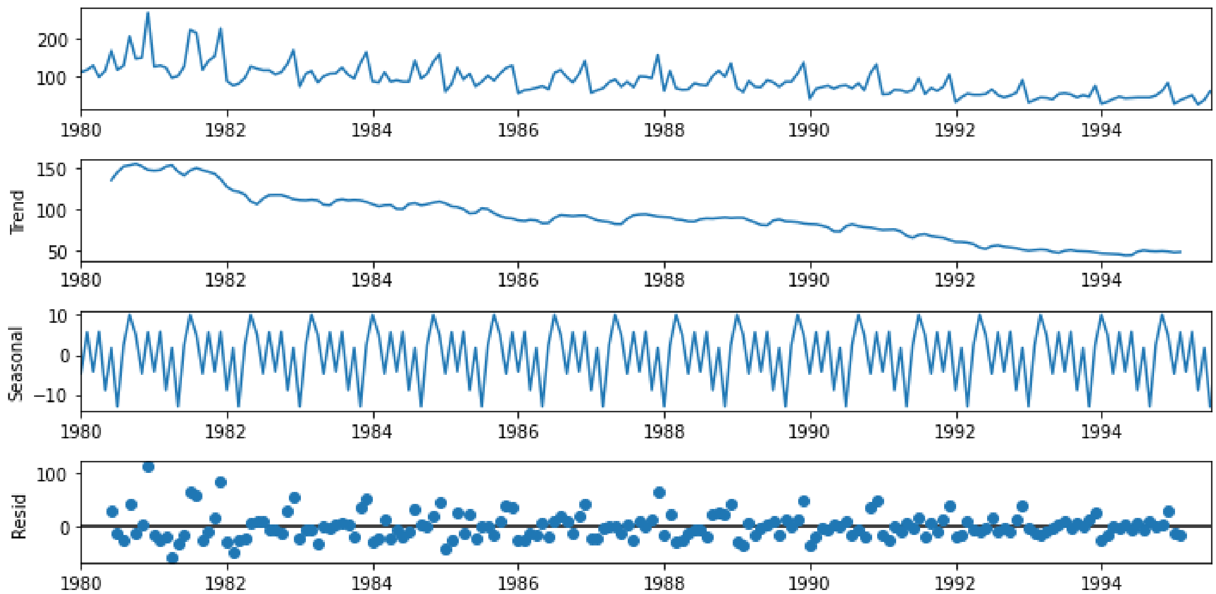


Additive Decomposition:

Sparkling:

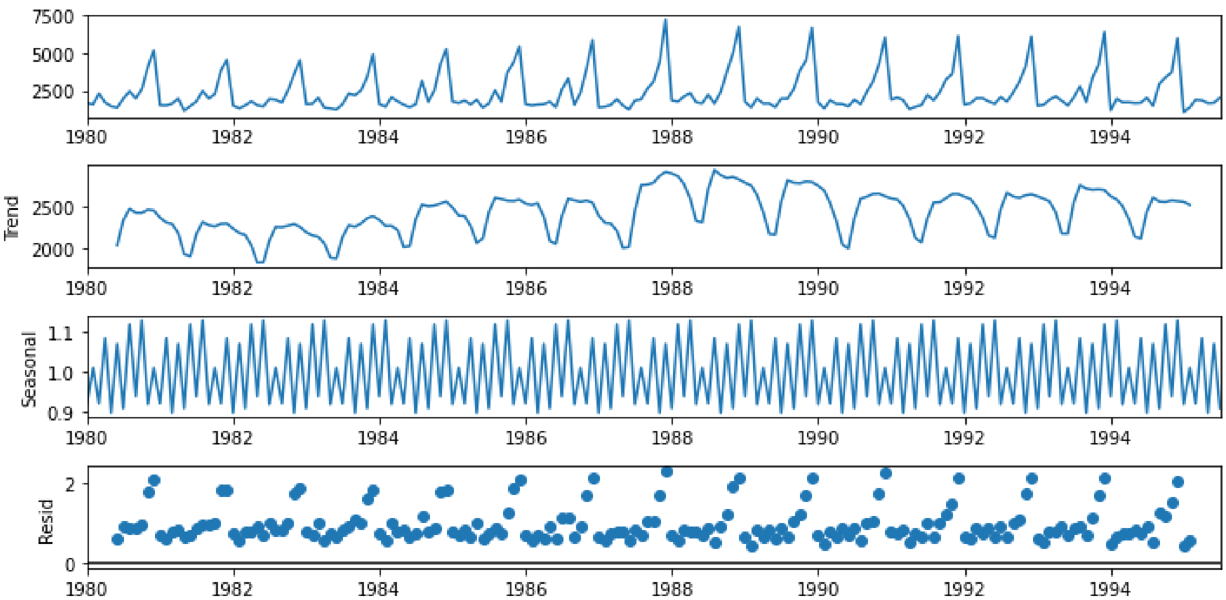


Rose:

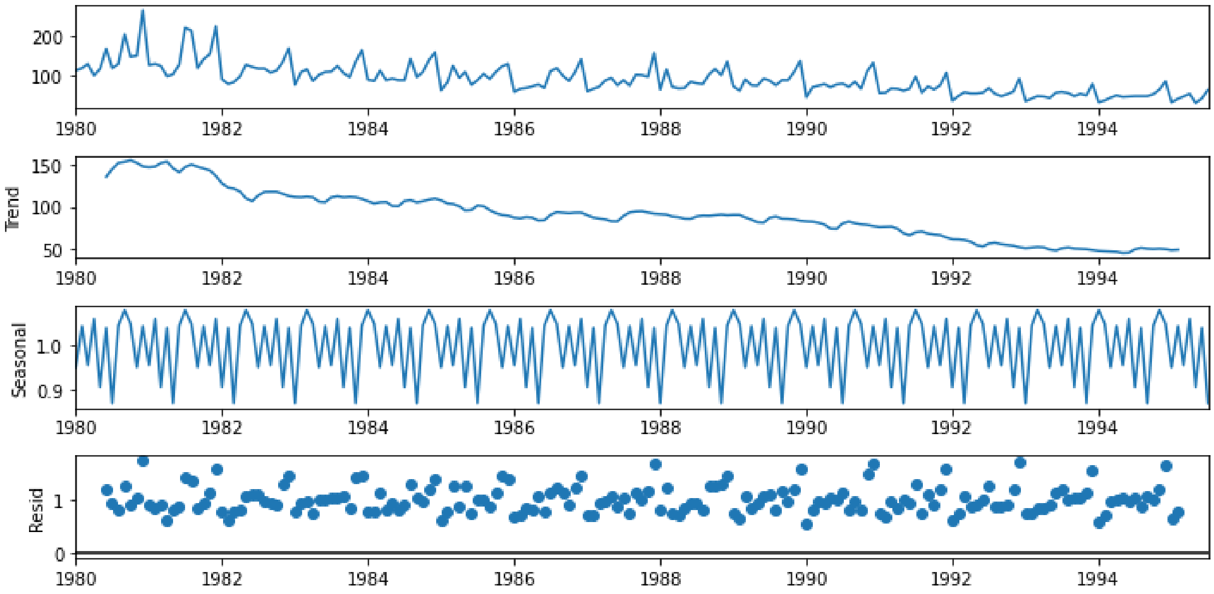


Multiplicative:

Sparkling:



Rose:



Summary Sparkling Dataset:

- Sparkling dataset doesn't show a visible trend however it shows seasonality, also if observed from additive decomposition the residual is catching some pattern.
- Multiplicative decomposition on the other hand seems to dictate on the series as the scale of the residual plot had decreased considerably
- Monthly bar plots showed that the sales are higher towards the last months than the earlier.

Summary Rose Dataset:

- Rose dataset show a clear decreasing trend as well as seasonality, multiplicative decomposition
- dictates the series the noise is reduced considerably in it also the seasonal patterns increase and decrease in the size across difference years
- The sales tend to go up during the July-August and also during end of the year.

PROBLEM 1.3

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

Resolution:

YearMonth	Sparkling	Year	Month
1990-08-01	1605	1990	8
1990-09-01	2424	1990	9
1990-10-01	3116	1990	10
1990-11-01	4286	1990	11
1990-12-01	6047	1990	12

```
Train Data: (132, 3)
```

YearMonth	Rose	Year	Month
1990-08-01	70.0	1990	8
1990-09-01	83.0	1990	9
1990-10-01	65.0	1990	10
1990-11-01	110.0	1990	11
1990-12-01	132.0	1990	12

```
Train Data: (132, 3)
```

YearMonth	Sparkling	Year	Month
1991-01-01	1902	1991	1
1991-02-01	2049	1991	2
1991-03-01	1874	1991	3
1991-04-01	1279	1991	4
1991-05-01	1432	1991	5

```
Test Data: (55, 3)
```

YearMonth	Rose	Year	Month
1991-01-01	54.0	1991	1
1991-02-01	55.0	1991	2
1991-03-01	66.0	1991	3
1991-04-01	65.0	1991	4
1991-05-01	60.0	1991	5

```
Test Data: (55, 3)
```

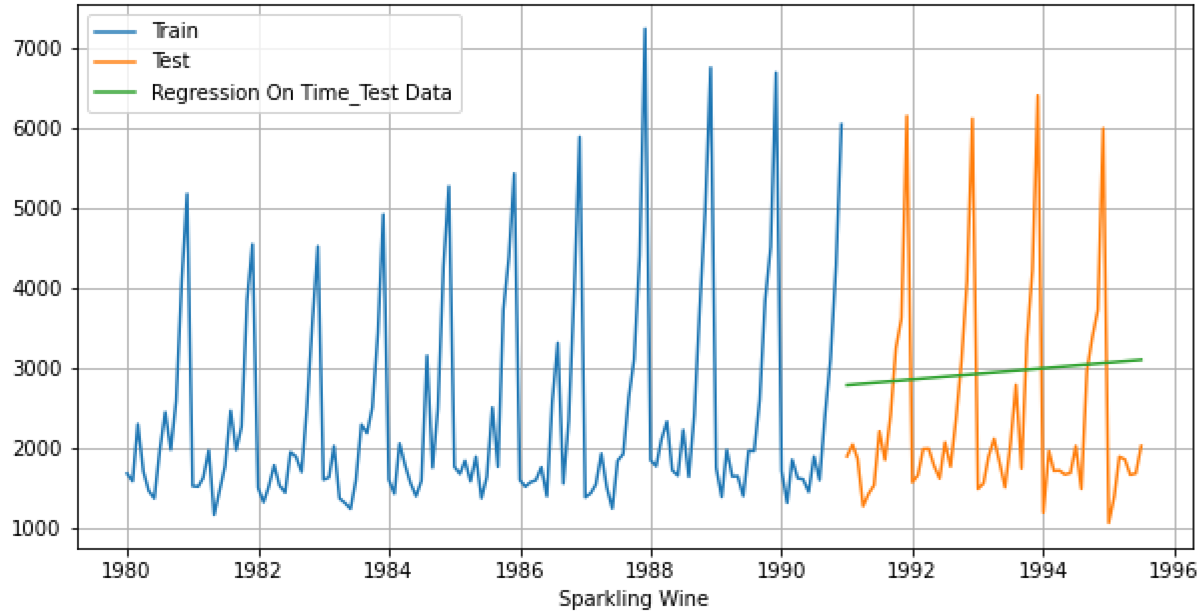
PROBLEM 1.4

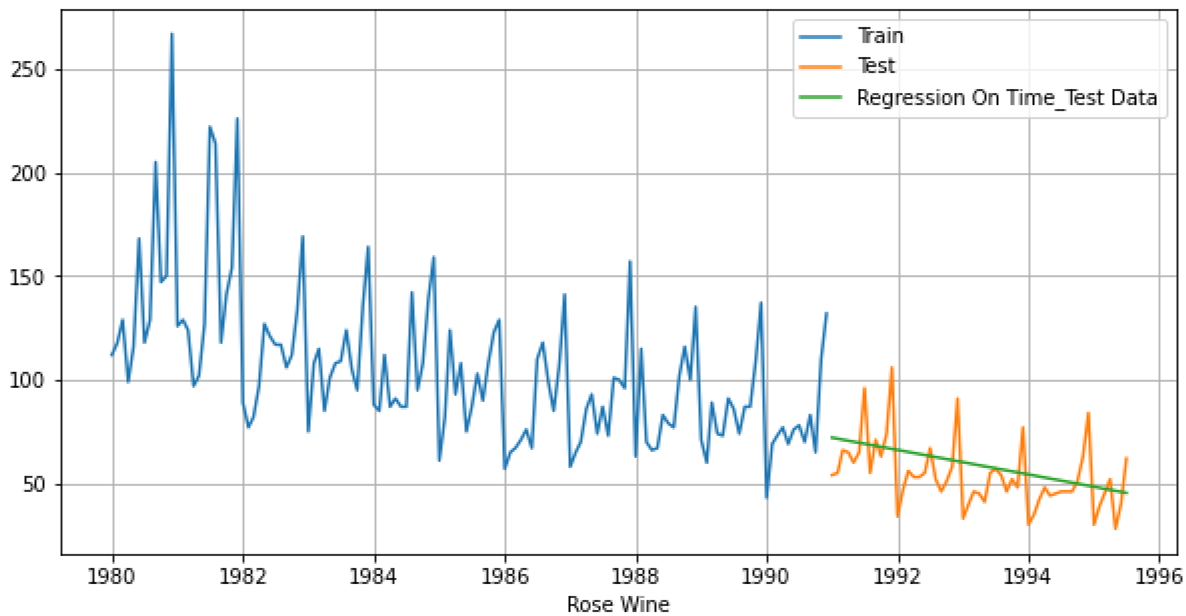
Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data.

Other models such as regression, naïve forecast models, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE

Resolution:

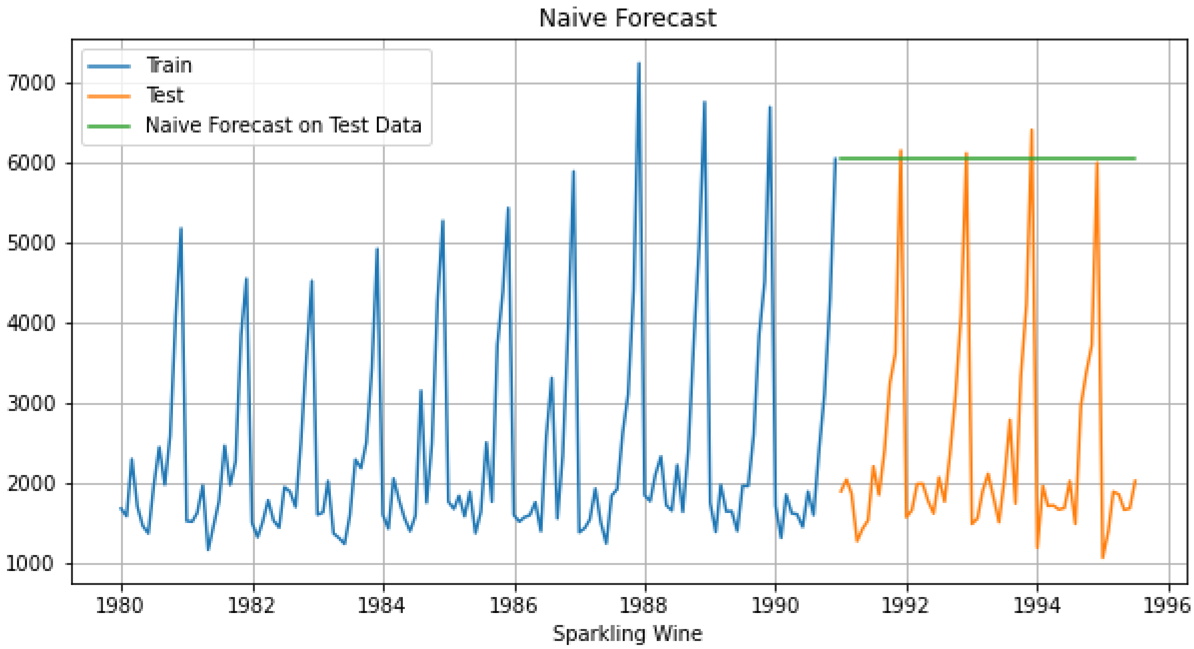
Model 1: Linear Regression: $\hat{y}$ t+1 = β y + c

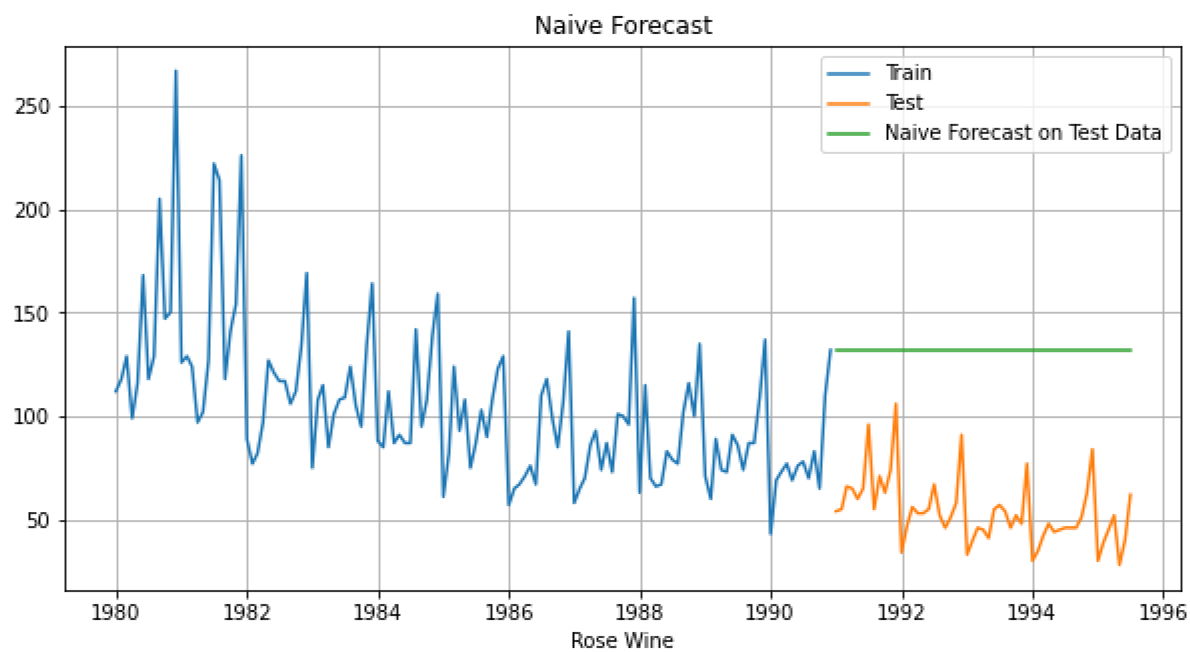

Train
Test
Regression On Time_Test Data
7000
6000
5000
4000
3000
2000
1000
1980
1982
1984
1986
1988
1990
1992
1994
1996
Sparkling Wine


Train
Test
Regression On Time_Test Data
250
200
150
100
50
1980
1982
1984
1986
1988
1990
1992
1994
1996
Rose Wine

Model 2: Naive Approach: ŷ t+1=yt

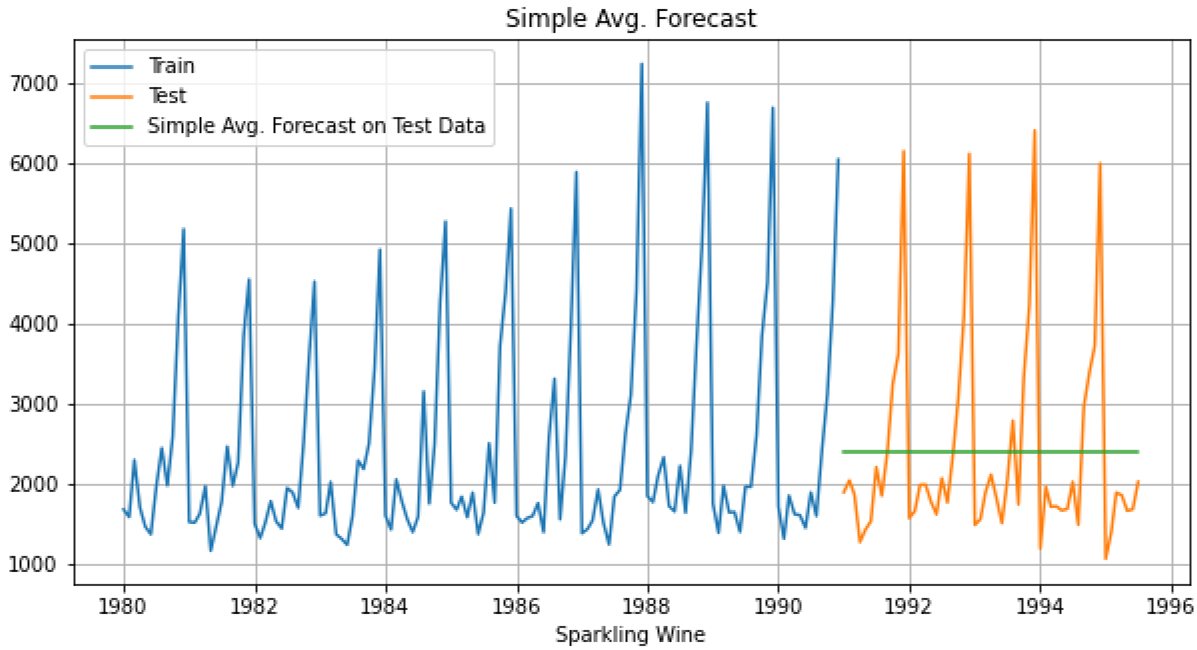
For this particular naive model, we say that the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow and since the prediction of tomorrow is same as today, therefore the prediction for day after tomorrow is also today.

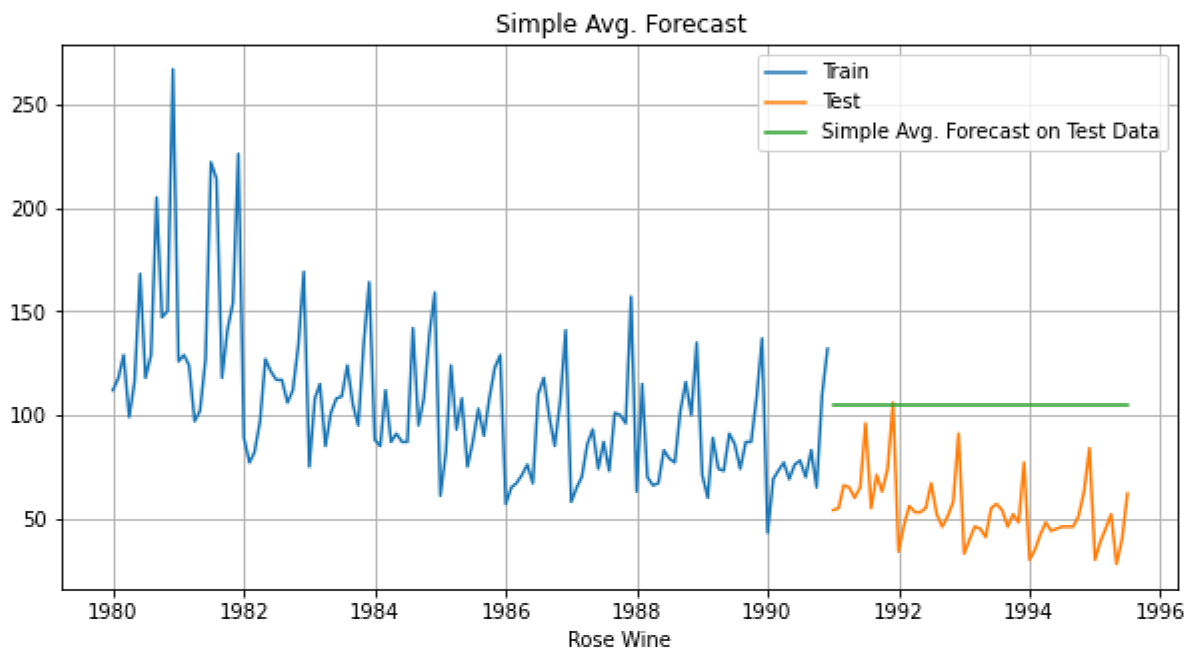




Method 3: Simple Average:

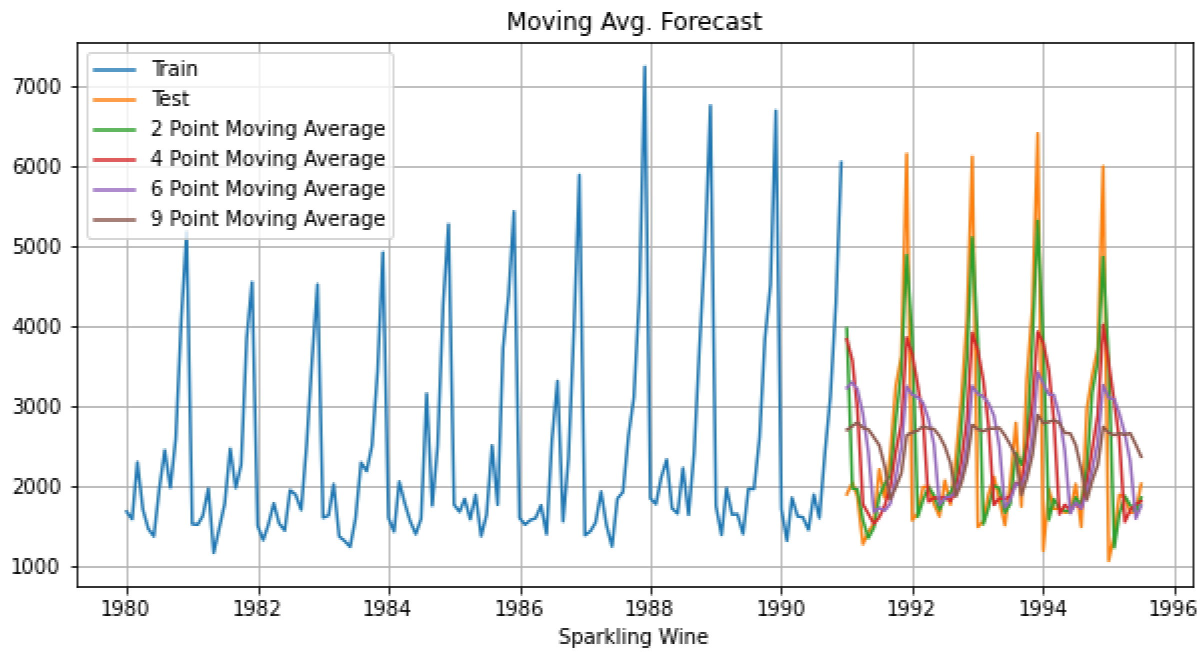
For this particular simple average method, we will forecast by using the average of the training values.

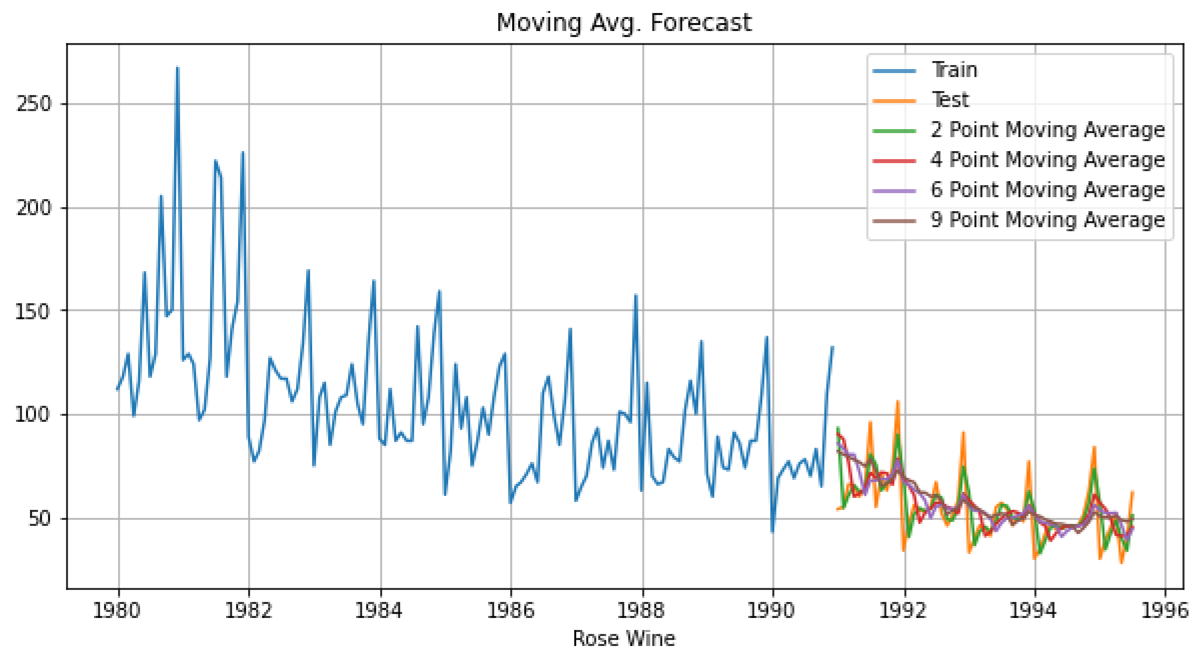




Method 4: Moving Average(MA)

For the moving average model, we are going to calculate rolling means (or moving averages) for different intervals. The best interval can be determined by the minimum error. The below plot shows the forecast for different rolling means:





Method 5: Exponential Smoothing methods

Exponential smoothing methods consist of flattening time series data. Exponential smoothing averages or exponentially weighted moving averages consist of forecast based on previous periods data with exponentially declining influence on the older observations.

Simple Exponential Smoothing (SES): The simplest of the exponentially smoothing methods is naturally called simple exponential smoothing (SES). This method is suitable for forecasting data with no clear trend or seasonal pattern. In Single ES, the forecast at time (t + 1) is given by Winters, 1960

$\hat{y}t+1=\alpha Yt+(1-\alpha)\hat{y}t$ Parameter α is called the smoothing constant and its value lies between 0 and 1. Since the model uses only one smoothing constant, it is called Single Exponential Smoothing.

Sparkling data doesn't show visible trend however it shows seasonality, Rose data on the other hand shows both trend and seasonality, all the Exponential models will still be built on both the datasets.

Double Exponential Smoothing(DES): One of the drawbacks of the simple exponential smoothing is that the model does not do well in the presence of the trend. This model is an extension of SES known as Double Exponential model which estimates two smoothing parameters. Applicable when data has Trend but no seasonality. Two separate components are considered: Level and Trend. Level is the local mean. One smoothing parameter α corresponds to the level series A second smoothing parameter β corresponds to the trend series. Double Exponential Smoothing uses two equations to forecast future values of the time series, one for forecasting the short term average value or level and the other for capturing the trend.

Intercept or Level equation, $\hat{y}t$ is given by: $\hat{y}t=\alpha yt+(1-\alpha)\hat{y}t$ Trend equation is given by

$Tt=\beta(\hat{y}t-\hat{y}t-1)+(1-\beta)Tt-1$ Here, α and β are the smoothing constants for level and trend, respectively,

$0 <\alpha < 1$ and $0 < \beta < 1$.

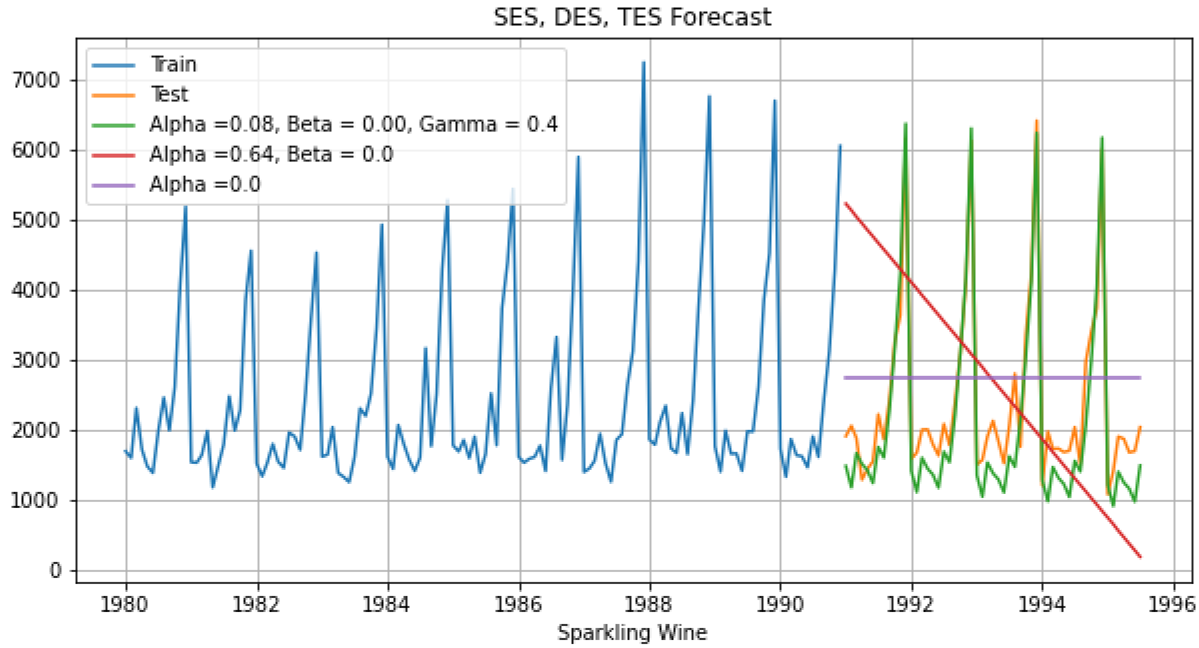
The forecast at time t + 1 is given by

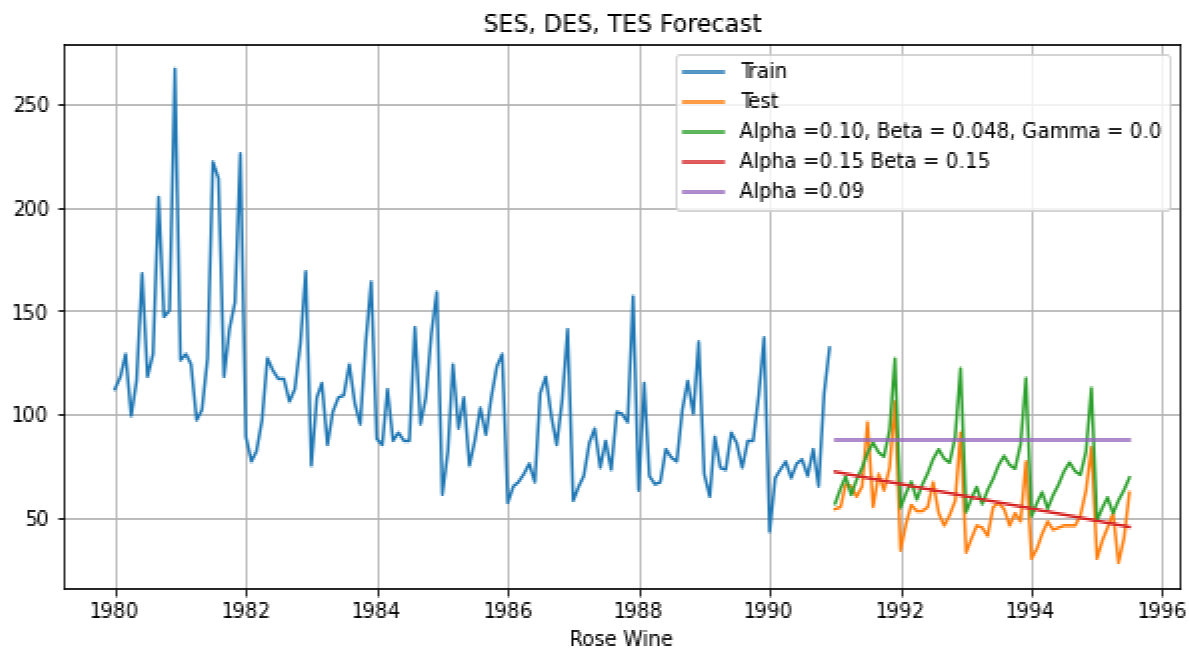
$Ft+1=\hat{y}t+Tt$ $Ft+n=\hat{y}t+nTt$

Though our Sparkling data doesn't seem to have a visible trend we are still going to build this model for the project. Rose data has a clear trend from the plot above

Inference

- Here, we see that the Double Exponential Smoothing model has picked up the trend component as well (see the below fig.)
- Our data has seasonality too so we will include one more smoothing parameter for seasonality which is gamma.
- We will use ETS (A, A, A) Holt Winter's linear method with additive trend and seasonality for Sparkling data and ETS (A, A, M) Holt Winter's linear method with additive trend and multiplicative seasonality for Rose wine data. We will call it Triple Exponential Smoothing (TES)





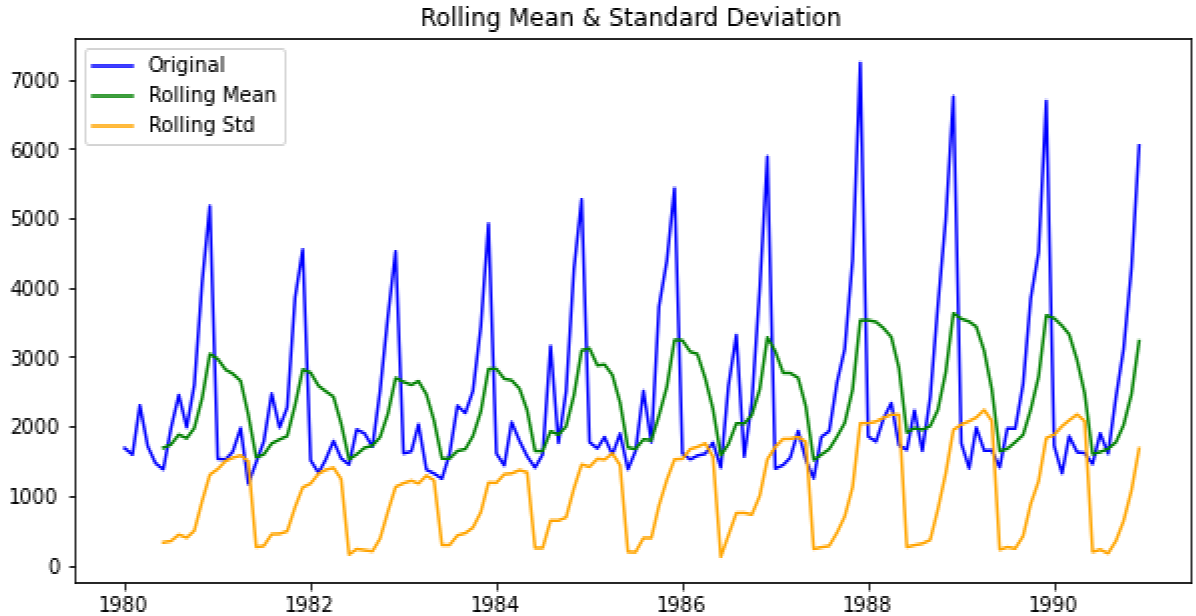
PROBLEM 1.5

Check for the stationary 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 stationary and comment.

Note: Stationary should be checked at alpha = 0.05.

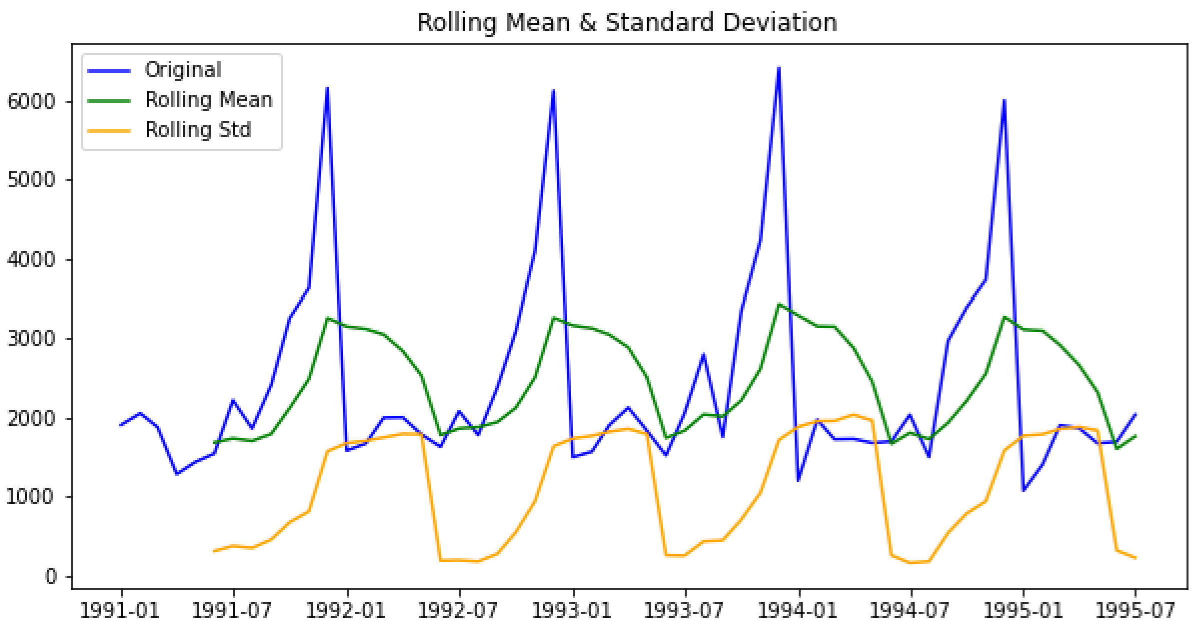
Resolution:

Sparkling Train set:




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

Sparkling Test set:

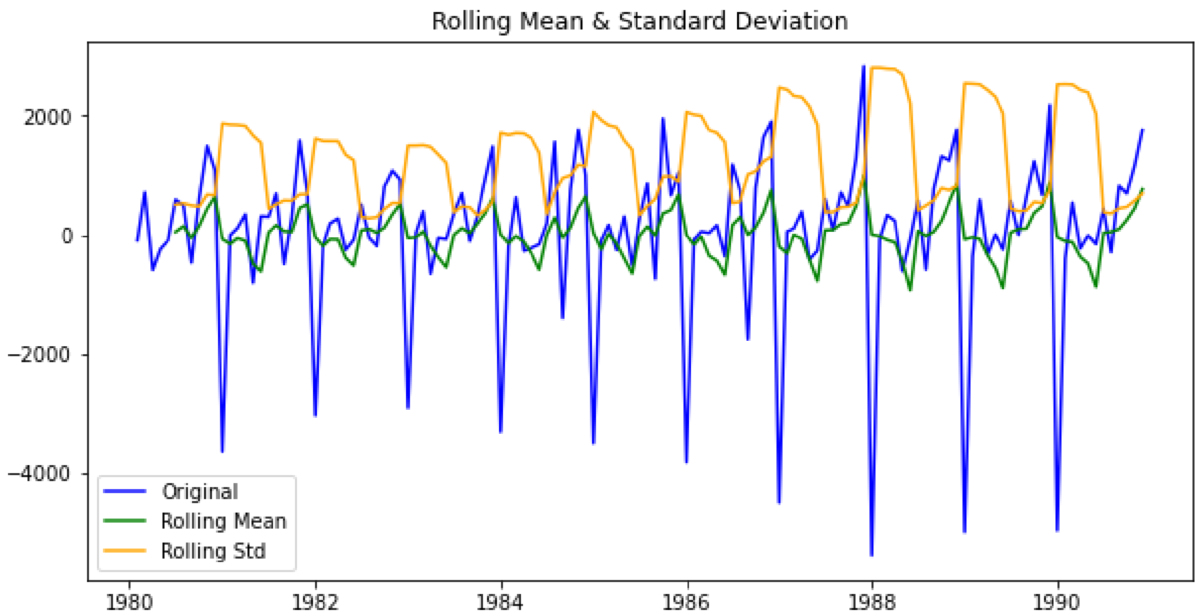



```
Results of Dickey-Fuller Test:
Test Statistic                  -1.790189
p-value                          0.385343
#Lags Used                      11.000000
Number of Observations Used     43.000000
Critical Value (1%)             -3.592504
Critical Value (5%)             -2.931550
Critical Value (10%)            -2.604066
dtype: float64
```

Since the Null Hypothesis H0 : The series is non-stationary Alternate Hypothesis H1: The series is stationary

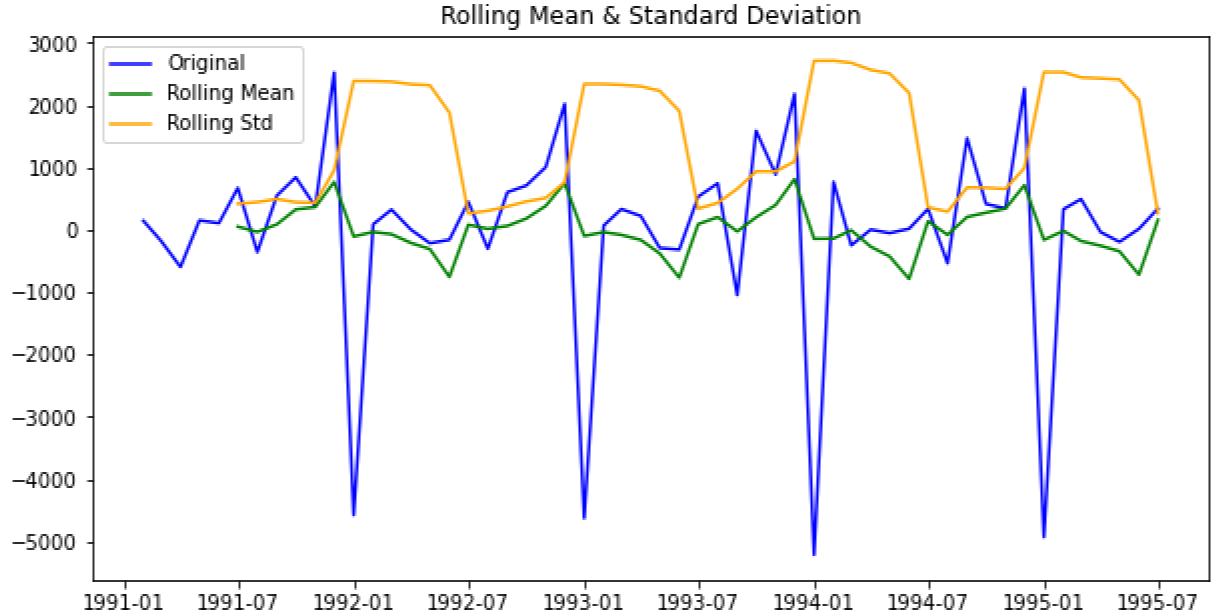
We cannot reject the null as the p values for both of series is greater than 0.05 (significance level) from the Augmented Dickey Fuller test above

Differenced Sparkling Train set:




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

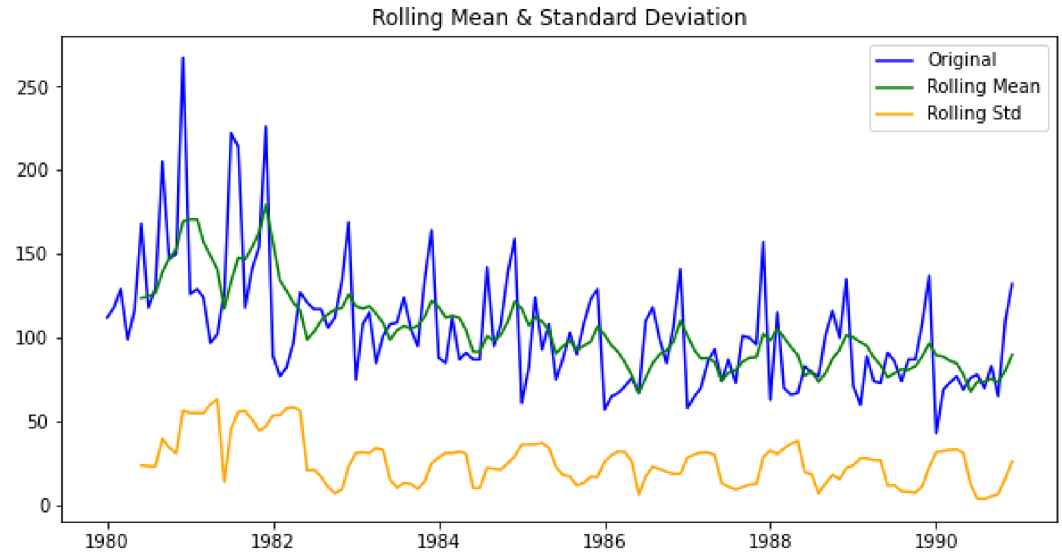
Differenced Sparkling Test set:




```
Results of Dickey-Fuller Test:
Test Statistic                 -7.050414e+00
p-value                         5.545252e-10
#Lags Used                      1.100000e+01
Number of Observations Used     4.200000e+01
Critical Value (1%)            -3.596636e+00
Critical Value (5%)            -2.933297e+00
Critical Value (10%)           -2.604991e+00
dtype: float64
```

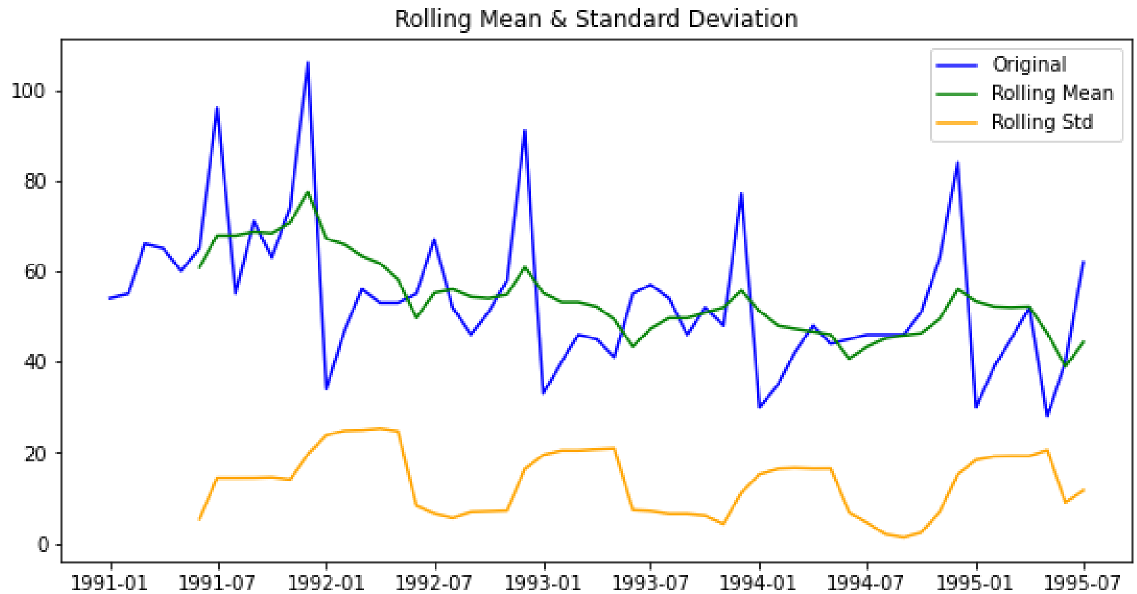
We can now see that the p –value < than 0.05 so we can reject the null-hypothesis and accept the alternate. So we say the series is stationary.

Rose Train Set:




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

Rose Test Set:



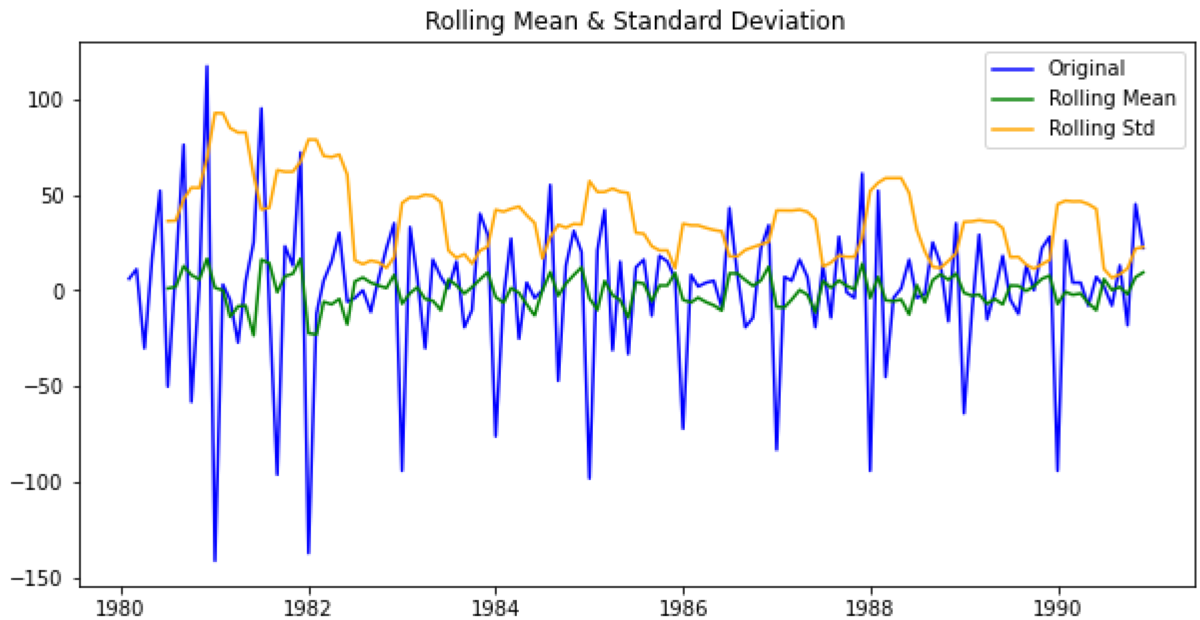

```
Results of Dickey-Fuller Test:
Test Statistic                 -4.464772
p-value                         0.000228
#Lags Used                     11.000000
Number of Observations Used    43.000000
Critical Value (1%)            -3.592504
Critical Value (5%)            -2.931550
Critical Value (10%)           -2.604066
dtype: float64
```

Since the Null Hypothesis H0: The series is non-stationary Alternate Hypothesis H1: The series is stationary we cannot reject the null as the p values is greater than 0.05 (significance level) from the Augmented Dickey Fuller test above Train set of Rose Wine dataset, on the contrary we can reject the null as the p values is less than 0.05 (significance level) from the Augmented Dickey Fuller test above Test set of Rose Wine dataset

We can correct the non-stationary by using multiple methods like taking differences at various level, using logged transformed series etc.

Here we will take difference of level 1 of the original train series and we will use the train dataset as is.

Differenced Rose Train set:




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

PROBLEM 1.6

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.

Resolution:

ARIMA

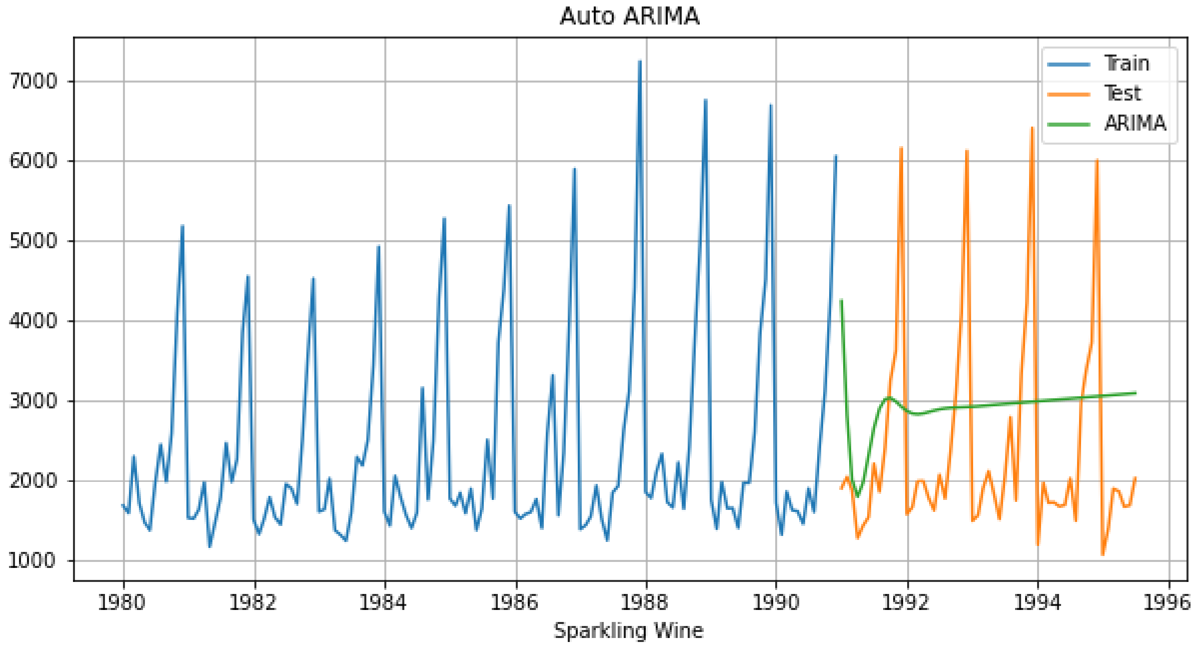
AIC score for both Sparkling and Rose wine dataset for different models is below:

	param	AIC_Sparkling
8	(2, 1, 2)	2210.616954
7	(2, 1, 1)	2232.360490
2	(0, 1, 2)	2232.783098
5	(1, 1, 2)	2233.597647
4	(1, 1, 1)	2235.013945
6	(2, 1, 0)	2262.035601
1	(0, 1, 1)	2264.906439
3	(1, 1, 0)	2268.528061
0	(0, 1, 0)	2269.582796

	param	AIC_Rose
2	(0, 1, 2)	1276.835372
5	(1, 1, 2)	1277.359223
4	(1, 1, 1)	1277.775747
7	(2, 1, 1)	1279.045689
8	(2, 1, 2)	1279.298694
1	(0, 1, 1)	1280.726183
6	(2, 1, 0)	1300.609261
3	(1, 1, 0)	1319.348311
0	(0, 1, 0)	1335.152658

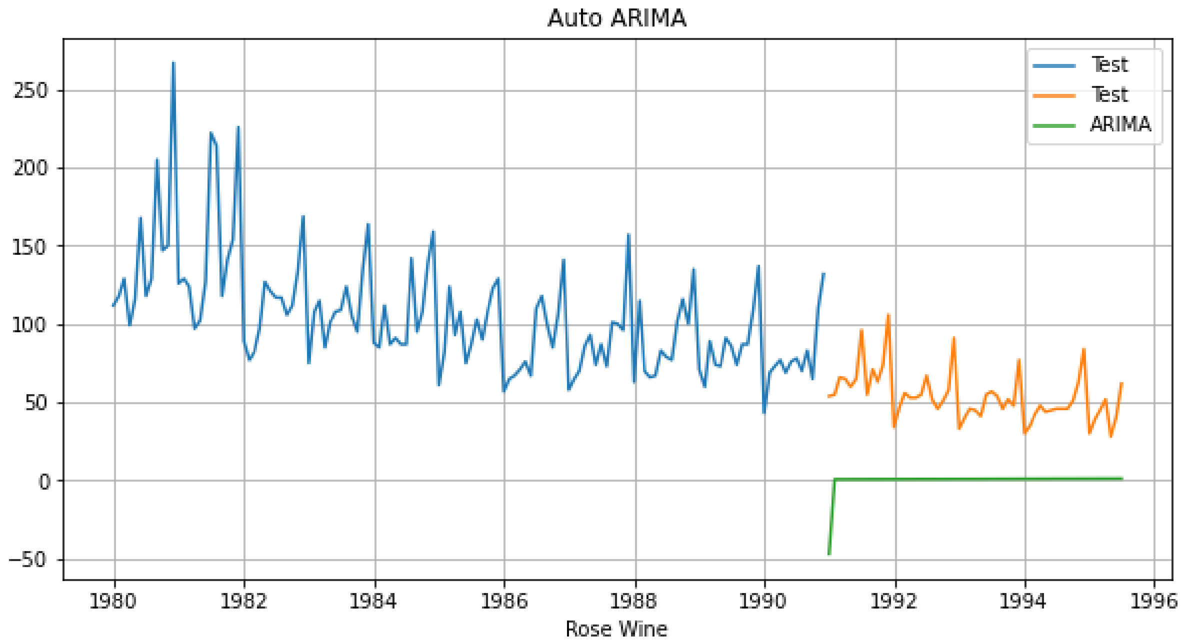
An automated model of (2,1,2) will be built on sparkling wine data and (0,1,2) on rose wine data. Both are of difference order 1.

```
Sparkling Data:
                              ARIMA Model Results
==============================================================================
Dep. Variable:            D.Sparkling   No. Observations:                  131
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1099.308
Method:                       css-mle   S.D. of innovations           1011.985
Date:                Sun, 05 Sep 2021   AIC                           2210.617
Time:                        14:09:34   BIC                           2227.868
Sample:                    02-01-1980   HQIC                          2217.627
                         - 12-01-1990
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 5.5860      0.516     10.825      0.000       4.575       6.597
ar.L1.D.Sparkling     1.2698      0.074     17.045      0.000       1.124       1.416
ar.L2.D.Sparkling    -0.5601      0.074     -7.617      0.000      -0.704      -0.416
ma.L1.D.Sparkling    -1.9993      0.042    -47.149      0.000      -2.082      -1.916
ma.L2.D.Sparkling     0.9993      0.042     23.584      0.000       0.916       1.082
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.1335           -0.7075j            1.3361           -0.0888
AR.2            1.1335           +0.7075j            1.3361            0.0888
MA.1            1.0002           +0.0000j            1.0002            0.0000
MA.2            1.0006           +0.0000j            1.0006            0.0000
-----------------------------------------------------------------------------
```



Rose Data:

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:                 D.Rose   No. Observations:                  130
Model:                 ARIMA(0, 1, 2)   Log Likelihood                -636.754
Method:                       css-mle   S.D. of innovations             30.440
Date:                Sun, 05 Sep 2021   AIC                           1281.508
Time:                        14:09:51   BIC                           1292.978
Sample:                    03-01-1980   HQIC                          1286.169
                         - 12-01-1990
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const            0.0091      0.005      1.996      0.046       0.000       0.018
ma.L1.D.Rose    -1.9955      0.039    -51.366      0.000      -2.072      -1.919
ma.L2.D.Rose     0.9955      0.039     25.457      0.000       0.919       1.072
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
MA.1            1.0000           +0.0000j            1.0000            0.0000
MA.2            1.0045           +0.0000j            1.0045            0.0000
-----------------------------------------------------------------------------
```



From the ACF plot we see a significant seasonal correlation after every 11th interval Setting the seasonality as 12 for the first iteration of the auto SARIMA model.

AIC scores for SARIMAX model

	param	seasonal	AIC_Sparkling
3	(0, 1, 3)	(3, 0, 3, 12)	2182.456353
14	(3, 1, 2)	(3, 0, 3, 12)	2677.681332
10	(2, 1, 2)	(3, 0, 3, 12)	2813.241309
7	(1, 1, 3)	(3, 0, 3, 12)	2953.601648
1	(0, 1, 1)	(3, 0, 3, 12)	3036.436346
8	(2, 1, 0)	(3, 0, 3, 12)	3043.170885
5	(1, 1, 1)	(3, 0, 3, 12)	3091.984506
0	(0, 1, 0)	(3, 0, 3, 12)	3172.429930
4	(1, 1, 0)	(3, 0, 3, 12)	3264.185626
12	(3, 1, 0)	(3, 0, 3, 12)	3303.282950
11	(2, 1, 3)	(3, 0, 3, 12)	3347.568890
9	(2, 1, 1)	(3, 0, 3, 12)	3363.739589
13	(3, 1, 1)	(3, 0, 3, 12)	3382.091626
2	(0, 1, 2)	(3, 0, 3, 12)	3665.477652
6	(1, 1, 2)	(3, 0, 3, 12)	3909.418096
15	(3, 1, 3)	(3, 0, 3, 12)	6460.315416

	param	seasonal	AIC_Rose
2	(0, 1, 2)	(3, 0, 3, 12)	2145.523921
10	(2, 1, 2)	(3, 0, 3, 12)	2945.075027
4	(1, 1, 0)	(3, 0, 3, 12)	3321.965396
3	(0, 1, 3)	(3, 0, 3, 12)	3345.176296
7	(1, 1, 3)	(3, 0, 3, 12)	3396.093427
15	(3, 1, 3)	(3, 0, 3, 12)	3482.006817
6	(1, 1, 2)	(3, 0, 3, 12)	3484.017259
1	(0, 1, 1)	(3, 0, 3, 12)	3490.096143
12	(3, 1, 0)	(3, 0, 3, 12)	3492.860663
11	(2, 1, 3)	(3, 0, 3, 12)	3498.600279
9	(2, 1, 1)	(3, 0, 3, 12)	3547.731254
0	(0, 1, 0)	(3, 0, 3, 12)	3549.313989
5	(1, 1, 1)	(3, 0, 3, 12)	3595.673963
13	(3, 1, 1)	(3, 0, 3, 12)	3712.496109
14	(3, 1, 2)	(3, 0, 3, 12)	3723.046367
8	(2, 1, 0)	(3, 0, 3, 12)	3857.771586

An automated SARIMA model of (3,1,2) will be built on sparkling wine data and (3,1,1) on rose wine data. both are of difference order 1 and seasonality 12.

Sparkling Data:

```
                                     SARIMAX Results                                     
=========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(3, 1, 2)x(3, 0, [], 12)   Log Likelihood                -696.287
Date:                           Sun, 05 Sep 2021   AIC                           1410.574
Time:                                   15:02:33   BIC                           1433.270
Sample:                                        0   HQIC                          1419.734
                                           - 132                                         
Covariance Type:                             opg                                         
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.5592      0.126    -12.411      0.000      -1.805      -1.313
ar.L2         -1.4484      0.118    -12.273      0.000      -1.680      -1.217
ar.L3         -0.4105      0.102     -4.008      0.000      -0.611      -0.210
ma.L1          1.1987      0.111     10.770      0.000       0.981       1.417
ma.L2          0.9999      0.123      8.109      0.000       0.758       1.242
ar.S.L12       0.4951      0.089      5.549      0.000       0.320       0.670
ar.S.L24       0.2824      0.096      2.944      0.003       0.094       0.470
ar.S.L36       0.2827      0.103      2.744      0.006       0.081       0.485
sigma2      2.046e+05    8.2e-07   2.49e+11      0.000    2.05e+05    2.05e+05
===================================================================================
Ljung-Box (L1) (Q):                   1.67   Jarque-Bera (JB):                23.20
Prob(Q):                              0.20   Prob(JB):                         0.00
Heteroskedasticity (H):               0.96   Skew:                             0.71
Prob(H) (two-sided):                  0.91   Kurtosis:                         5.01
===================================================================================
```

Note:

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

[2] Covariance matrix is singular or near-singular, with condition number 2.3e+26. Standard errors may be unstable.

Rose Data:

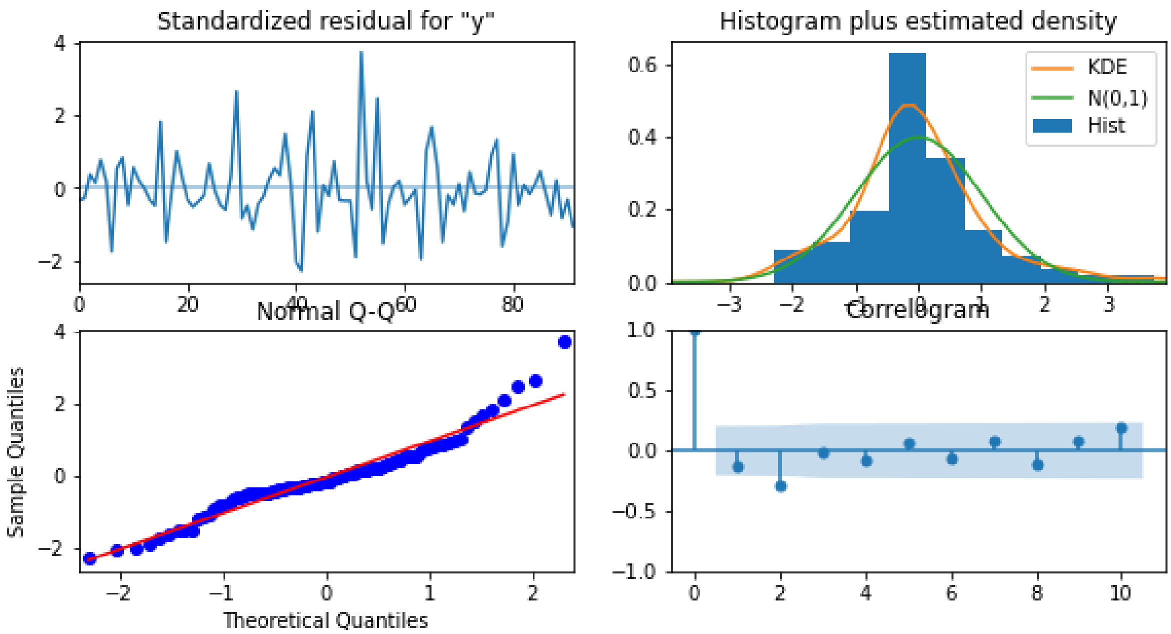
```
                                      SARIMAX Results                                      
===========================================================================================
Dep. Variable:                                   y   No. Observations:                  132
Model:             SARIMAX(3, 1, 1)x(3, 0, [1, 2], 11)   Log Likelihood                -437.103
Date:                             Sun, 05 Sep 2021   AIC                            894.205
Time:                                     15:02:44   BIC                            919.744
Sample:                                          0   HQIC                           904.525
                                             - 132                                         
Covariance Type:                               opg                                         
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.1884      0.121      1.551      0.121      -0.050       0.426
ar.L2          0.0208      0.123      0.170      0.865      -0.219       0.261
ar.L3          0.0146      0.140      0.104      0.917      -0.259       0.288
ma.L1         -0.9339      0.076    -12.309      0.000      -1.083      -0.785
ar.S.L11      -0.2380      0.421     -0.565      0.572      -1.063       0.587
ar.S.L22      -0.0357      0.170     -0.210      0.834      -0.369       0.298
ar.S.L33      -0.0042      0.115     -0.036      0.971      -0.229       0.221
ma.S.L11       0.1810      0.448      0.404      0.686      -0.697       1.059
ma.S.L22      -0.1736      0.234     -0.742      0.458      -0.632       0.285
sigma2       565.7116     98.703      5.731      0.000     372.258     759.165
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                 0.17
Prob(Q):                              0.90   Prob(JB):                         0.92
Heteroskedasticity (H):               0.91   Skew:                            -0.06
Prob(H) (two-sided):                  0.80   Kurtosis:                         3.17
===================================================================================
```

Note:

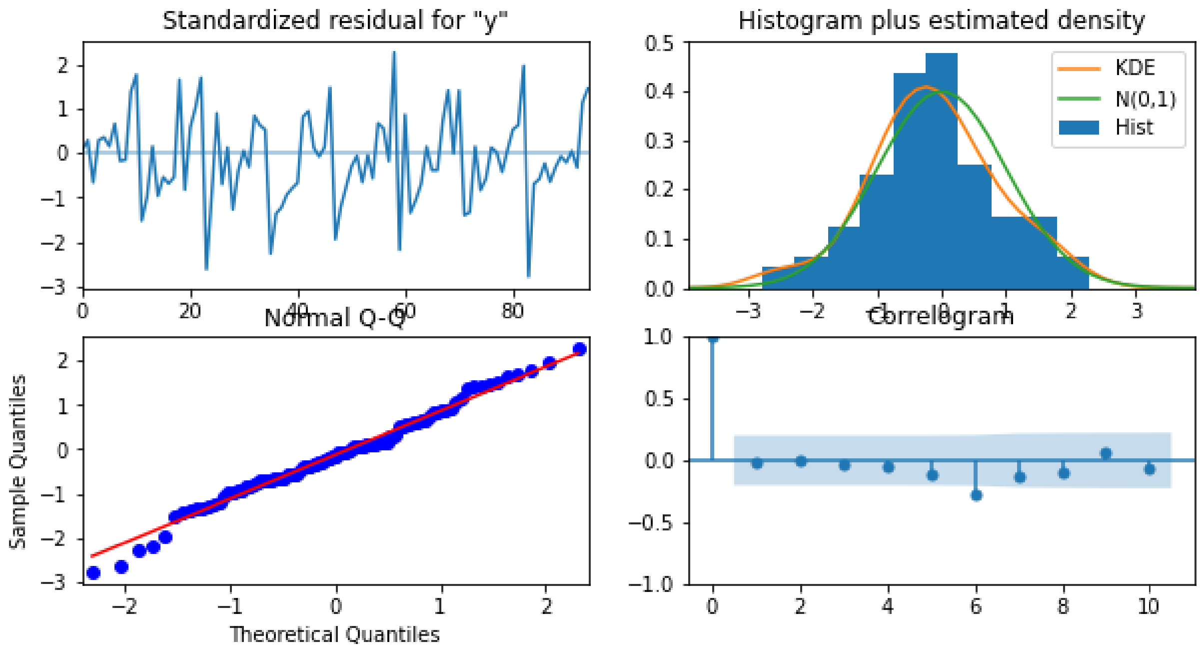
[1] Covariance matrix calculated using the outer product of gradients (complex-step)

Diagnostic plots for Auto SARIMA model are as below:

Sparkling Data:



Rose Data:



Sparkling Dataset Diagnostic:

From the diagnostic plots we see that the assumptions of Normality, heteroscedasticity as seems to be getting satisfied as well the series show randomness and no auto correlation between the residuals

Rose Dataset Diagnostic:

The plot shows randomness of the residual also the assumption of normality and heteroscedasticity is satisfied, it shows no auto correlation until lag 5, then shows a rise in significance at 6.

Though visual plots satisfy most assumptions the test proves it wrong seen from the summary of SARIMAX model for both the dataset.

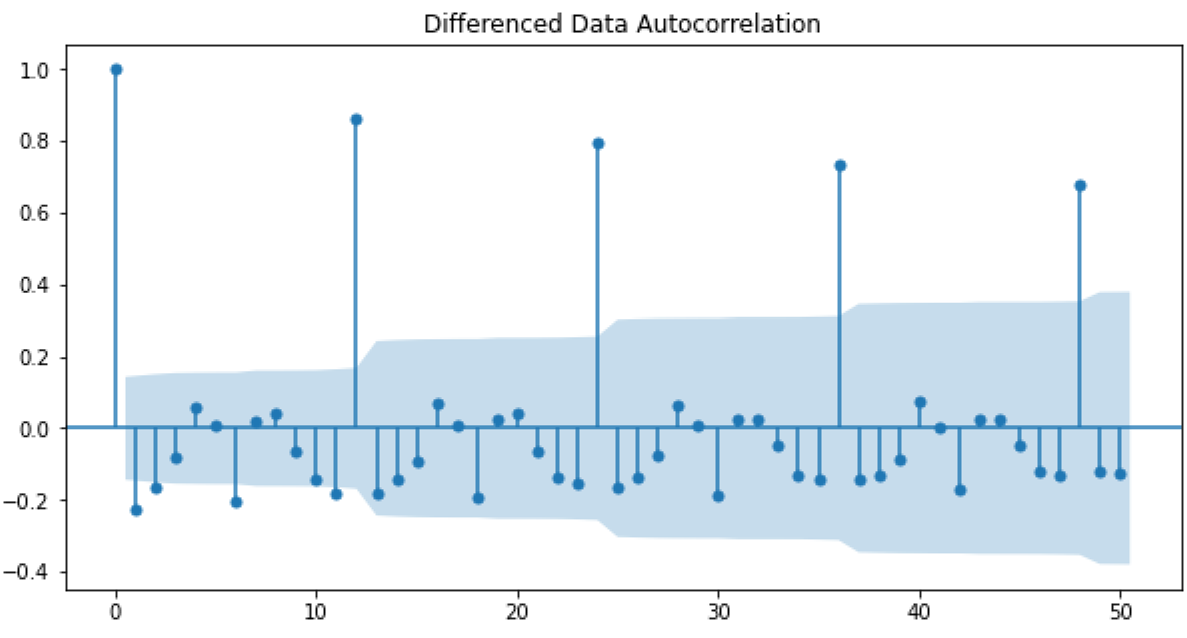
PROBLEM 1.7

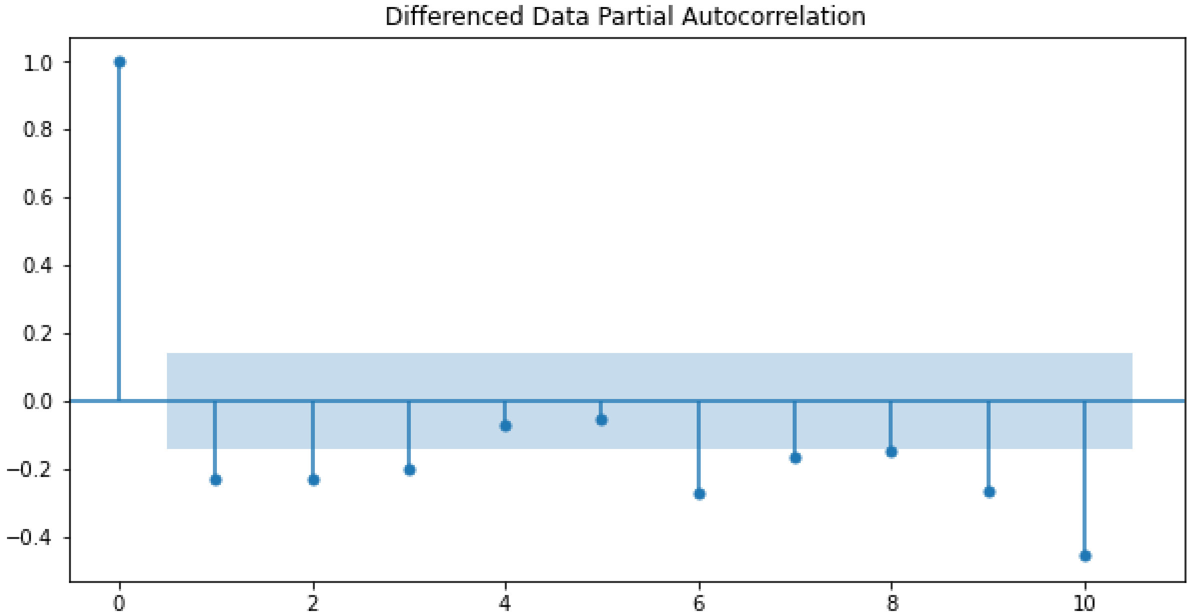
Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.

Resolution:

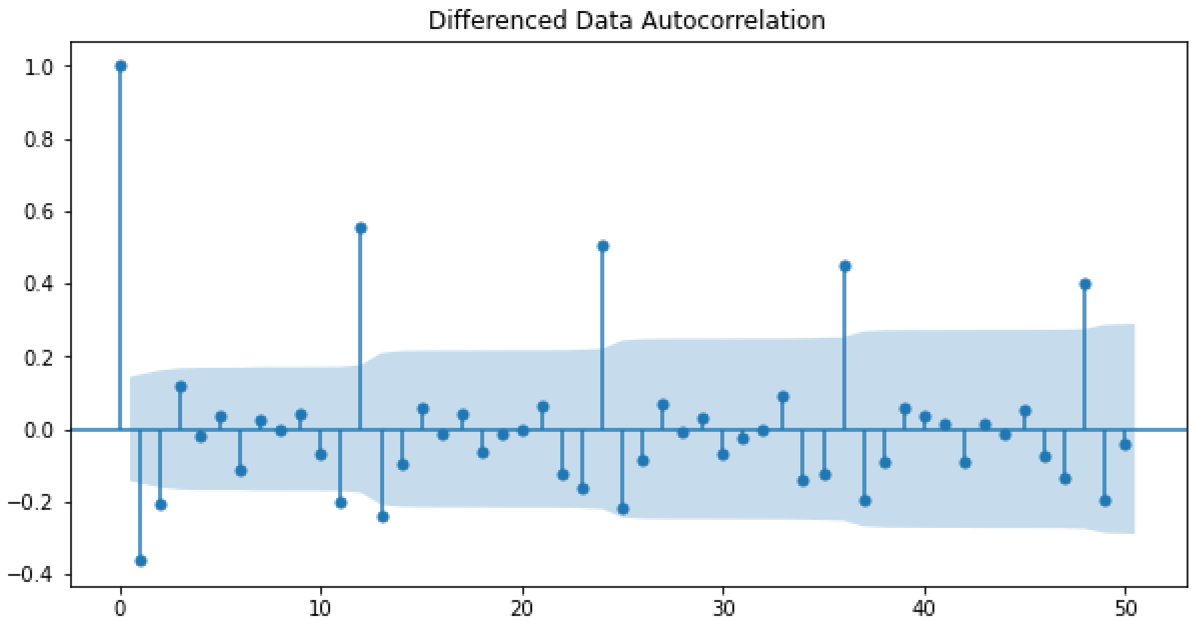
ARIMA

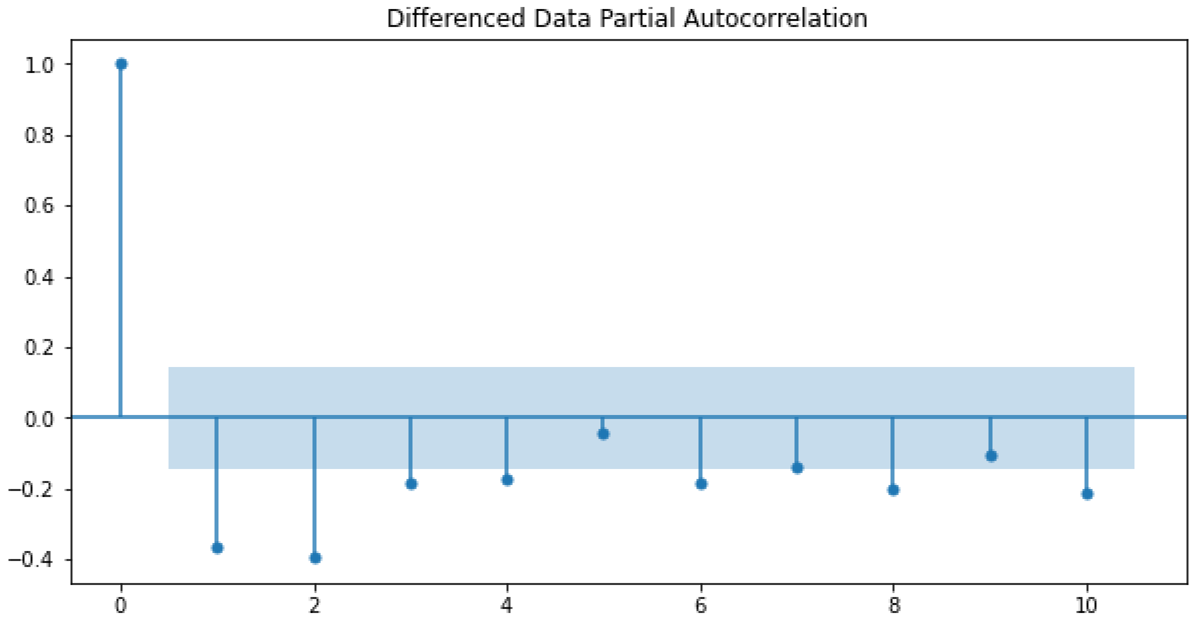
Sparkling Dataset:





Rose Dataset:

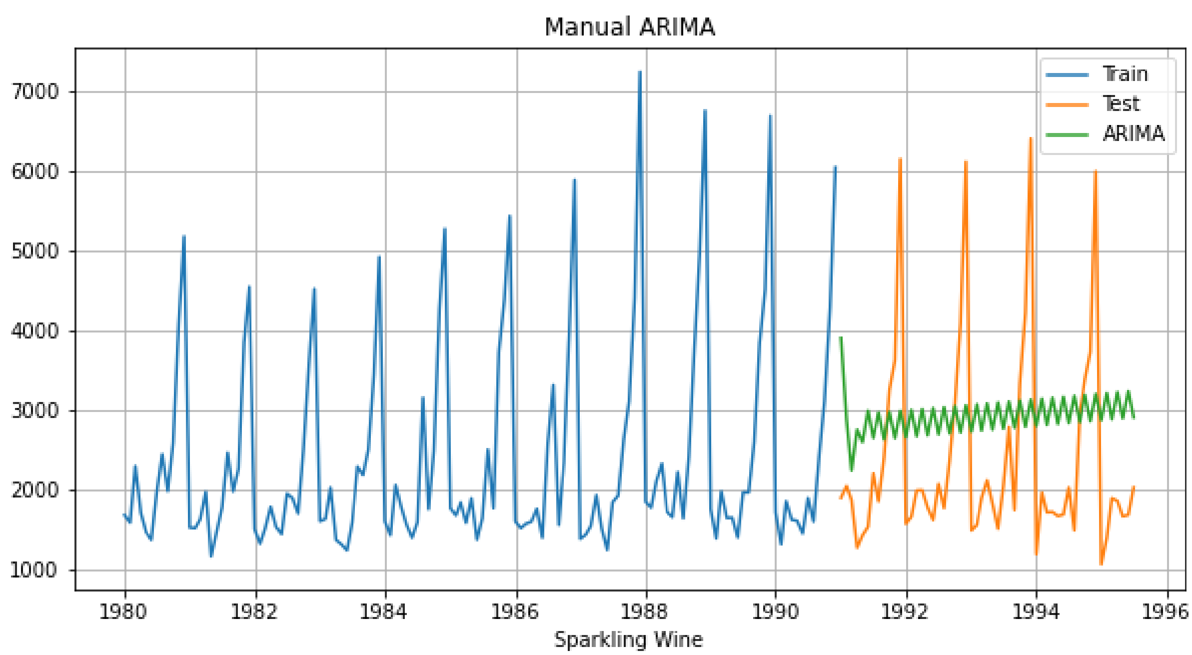




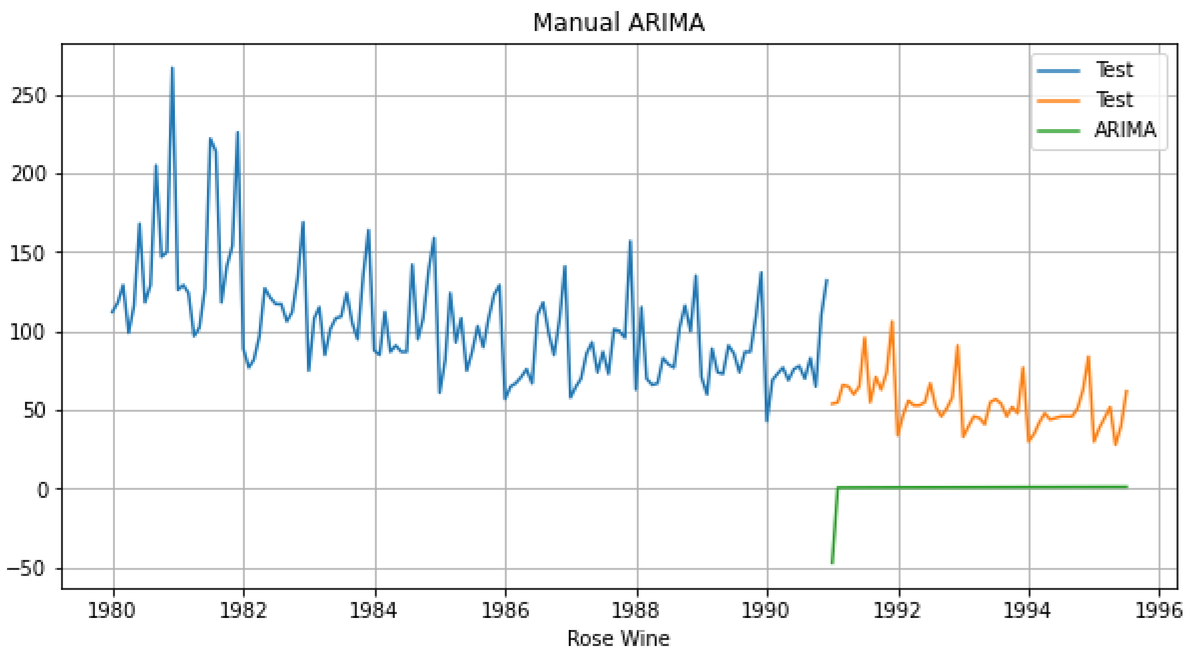
- Here, we have taken alpha=0.05.
- The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off to 0. The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 0.
- By looking at the above plots for Sparkling data, we can say that both the PACF cuts off at 3 and ACF plot cuts-off at lag 2.
- By looking at the above plots for Rose data, we can say that PACF cuts off at 4 and ACF plot cuts-off at lag 2.

Sparkling Data:

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:             D.Sparkling   No. Observations:                  131
Model:                  ARIMA(3, 1, 2)   Log Likelihood               -1107.464
Method:                        css-mle   S.D. of innovations           1106.033
Date:                 Sun, 05 Sep 2021   AIC                           2228.928
Time:                         14:12:29   BIC                           2249.054
Sample:                     02-01-1980   HQIC                          2237.106
                          - 12-01-1990
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 5.8816        nan        nan        nan         nan         nan
ar.L1.D.Sparkling    -0.4422        nan        nan        nan         nan         nan
ar.L2.D.Sparkling     0.3075   7.77e-06   3.96e+04      0.000       0.308       0.308
ar.L3.D.Sparkling    -0.2503        nan        nan        nan         nan         nan
ma.L1.D.Sparkling    -0.0004      0.028     -0.013      0.990      -0.055       0.054
ma.L2.D.Sparkling    -0.9996      0.028    -36.010      0.000      -1.054      -0.945
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1           -1.0000           -0.0000j            1.0000           -0.5000
AR.2            1.1145           -1.6595j            1.9990           -0.1559
AR.3            1.1145           +1.6595j            1.9990            0.1559
MA.1            1.0000           +0.0000j            1.0000            0.0000
MA.2           -1.0004           +0.0000j            1.0004            0.5000
-----------------------------------------------------------------------------
```



```
Rose:
                             ARIMA Model Results
==============================================================================
Dep. Variable:                 D.Rose   No. Observations:                  131
Model:                 ARIMA(4, 1, 2)   Log Likelihood                -633.876
Method:                       css-mle   S.D. of innovations             29.793
Date:                Sun, 05 Sep 2021   AIC                           1283.753
Time:                        14:12:39   BIC                           1306.754
Sample:                    02-01-1980   HQIC                          1293.099
                         - 12-01-1990
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const           -0.1905      0.576     -0.331      0.741      -1.319       0.938
ar.L1.D.Rose     1.1685      0.087     13.391      0.000       0.997       1.340
ar.L2.D.Rose    -0.3562      0.132     -2.693      0.007      -0.616      -0.097
ar.L3.D.Rose     0.1855      0.132      1.402      0.161      -0.074       0.445
ar.L4.D.Rose    -0.2227      0.091     -2.443      0.015      -0.401      -0.044
ma.L1.D.Rose    -1.9506        nan        nan        nan         nan         nan
ma.L2.D.Rose     1.0000        nan        nan        nan         nan         nan
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.1027           -0.4116j            1.1770           -0.0569
AR.2            1.1027           +0.4116j            1.1770            0.0569
AR.3           -0.6863           -1.6643j            1.8003           -0.3122
AR.4           -0.6863           +1.6643j            1.8003            0.3122
MA.1            0.9753           -0.2209j            1.0000           -0.0355
MA.2            0.9753           +0.2209j            1.0000            0.0355
-----------------------------------------------------------------------------
```



AIC for sparkling data is the lowest for the model (3,1,2), also we saw the from ACF and PACG plots that the cut off of p and q are at 3 and 2 resp. so we conclude that the auto SARIMAX and the manual SARIMAX models are the same.

SARIMA

For Rose data let's build a model at the p and q cut off at 4, 2 respectively.

Manual SARIMAX Summary on Rose data:

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(4, 1, 2)x(3, 0, 2, 12)   Log Likelihood                -371.081
Date:                            Sun, 05 Sep 2021   AIC                            766.161
Time:                                    15:04:24   BIC                            796.292
Sample:                                         0   HQIC                           778.317
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.7986      0.188     -4.250      0.000      -1.167      -0.430
ar.L2         -0.0110      0.159     -0.069      0.945      -0.322       0.300
ar.L3         -0.1475      0.153     -0.963      0.336      -0.448       0.153
ar.L4         -0.2441      0.108     -2.269      0.023      -0.455      -0.033
ma.L1         -0.0887      0.186     -0.476      0.634      -0.454       0.276
ma.L2         -0.7649      0.183     -4.186      0.000      -1.123      -0.407
ar.S.L12       0.7670      0.165      4.638      0.000       0.443       1.091
ar.S.L24       0.0839      0.149      0.565      0.572      -0.207       0.375
ar.S.L36       0.0765      0.093      0.823      0.410      -0.106       0.259
ma.S.L12      -0.5259      0.288     -1.824      0.068      -1.091       0.039
ma.S.L24      -0.2331      0.230     -1.014      0.311      -0.684       0.218
sigma2       181.2903     39.761      4.559      0.000     103.359     259.221
===================================================================================
Ljung-Box (L1) (Q):                   0.04   Jarque-Bera (JB):                 0.93
Prob(Q):                              0.85   Prob(JB):                         0.63
Heteroskedasticity (H):               1.24   Skew:                             0.25
Prob(H) (two-sided):                  0.56   Kurtosis:                         2.99
===================================================================================
```

PROBLEM 1.8

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

Resolution:

	Test_Spark RMSE
Regression	1389.135175
NaiveModel	3864.279352
SimpleAvg	1275.081804
MovingAvg2	813.400684
MovingAvg4	1156.589694
MovingAvg6	1283.927428
MovingAvg9	1346.278315
SES	1316.034674
DES	2007.238526
TES	473.954404
Auto ARIMA (2,1,2)	1375.028279
Manual ARIMA (3,1,2)	1375.097144
Auto SARIMA (3,1,2)(3,0,0,12)	1918.728383

	Test_Rose RMSE
Regression	15.262509
NaiveModel	79.699093
SimpleAvg	53.440426
MovingAvg2	11.529409
MovingAvg4	14.448930
MovingAvg6	14.560046
MovingAvg9	14.724503
SES	36.775774
DES	15.262495
TES	20.906585
Auto ARIMA (0,1,2)	56.351188
Manual ARIMA (4,1,2)	33.930217
Auto SARIMA (3,1,1)(3,0,2,12)	38.034261

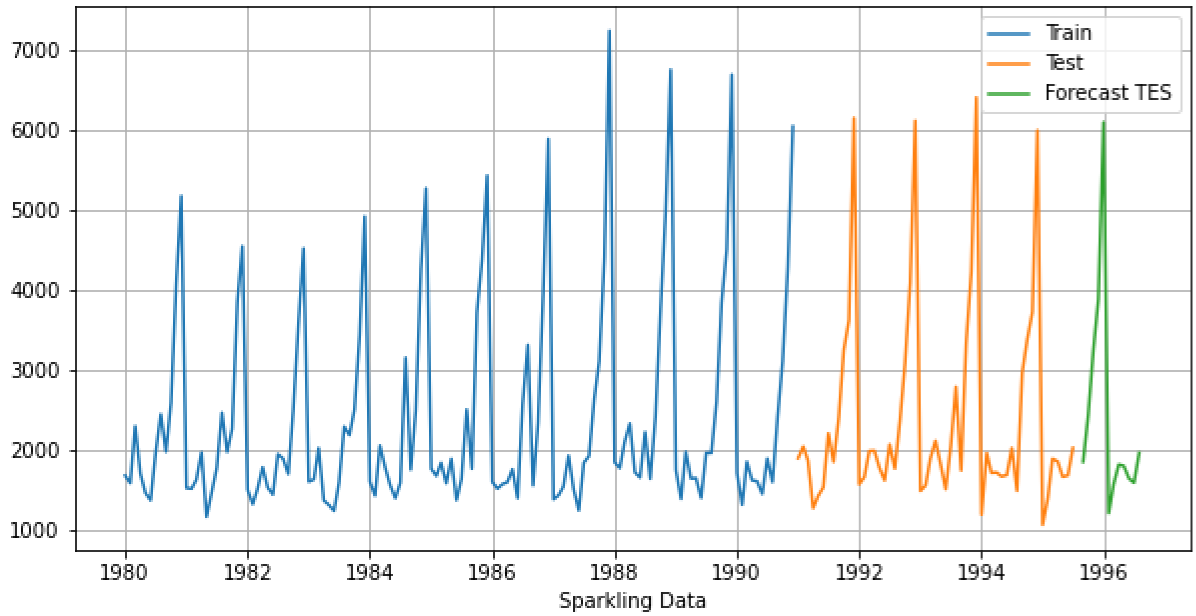
PROBLEM 1.9

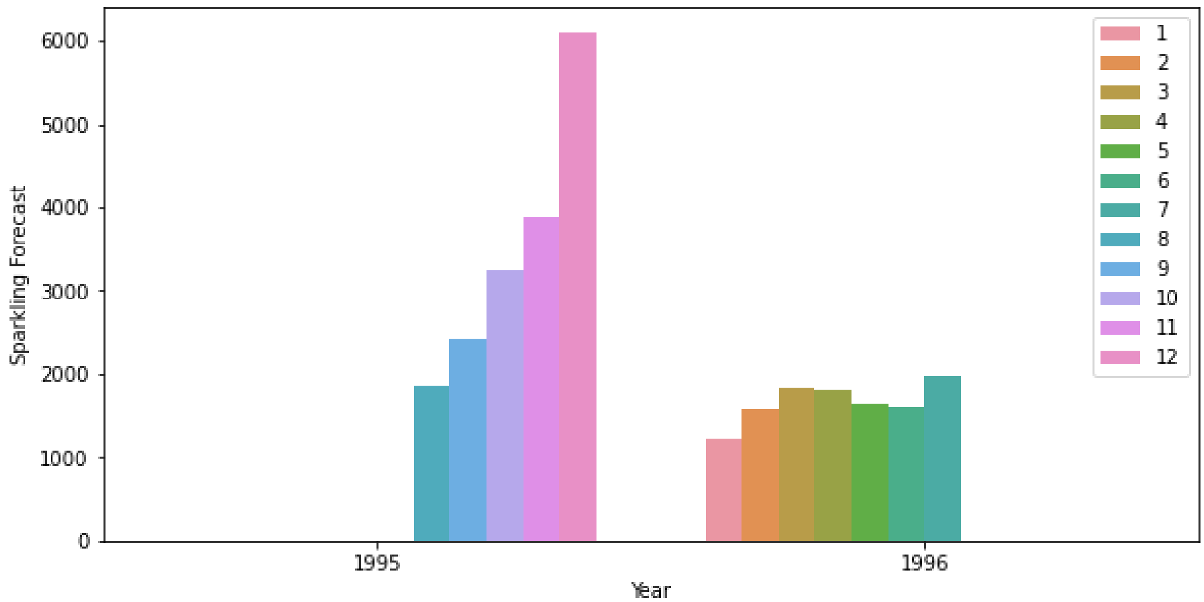
Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

Resolution:

For Sparkling dataset, we see that Triple Exponential smoothing gives the best forecast, so we will move forward with that for forecasting

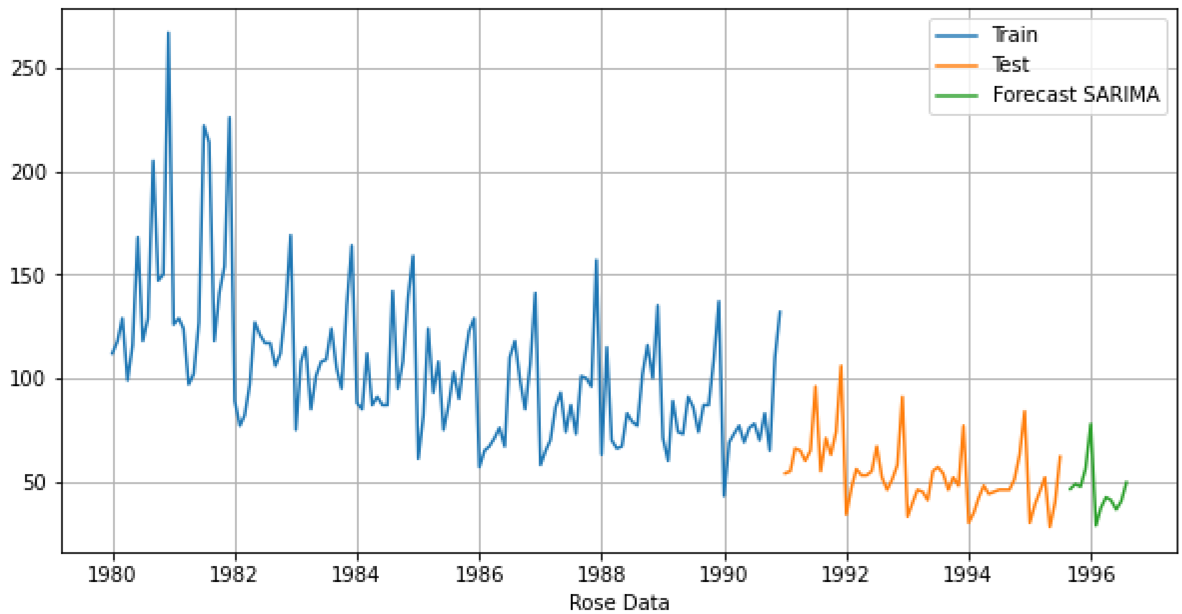
Time	Sparkling Forecast	lower CI	upper CI
1995-08-31	1858.037942	1075.060737	2641.015147
1995-09-30	2432.697346	1649.720141	3215.674551
1995-10-31	3246.134445	2463.157240	4029.111649
1995-11-30	3888.467711	3105.490506	4671.444916
1995-12-31	6099.865815	5316.888610	6882.843019
1996-01-31	1216.015599	433.038394	1998.992804
1996-02-29	1576.041044	793.063839	2359.018249
1996-03-31	1824.637312	1041.660107	2607.614517
1996-04-30	1806.325944	1023.348739	2589.303148
1996-05-31	1648.757666	865.780461	2431.734871
1996-06-30	1595.635315	812.658110	2378.612520
1996-07-31	1969.866380	1186.889175	2752.843585

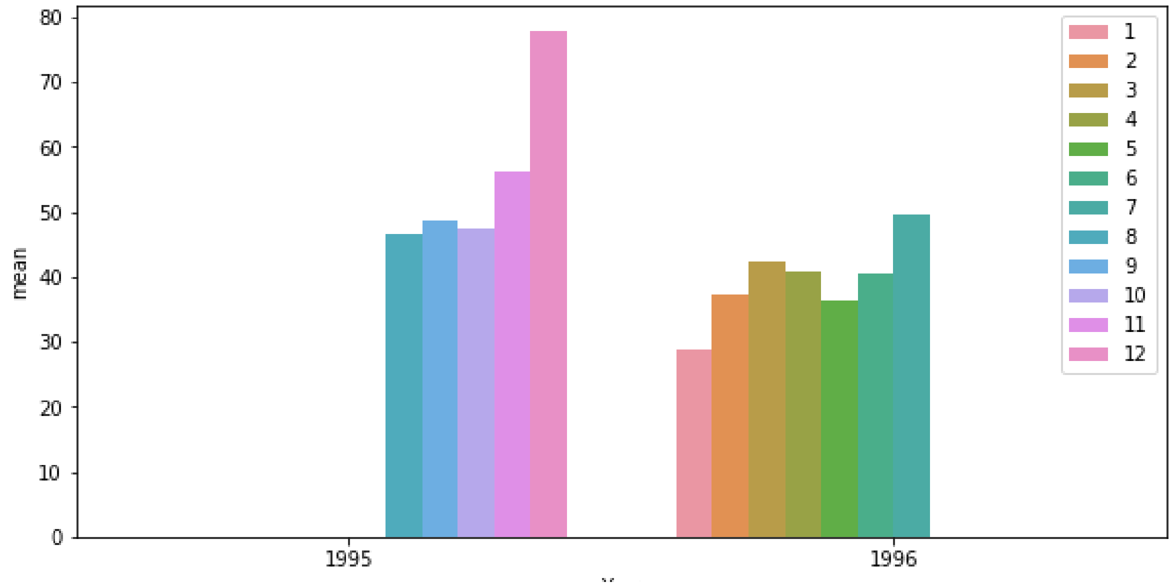




For Rose dataset rolling avg shows the best RMSE, however since the window chosen was very small(2,4,6,9) it was natural it was going to work well on Test set. The other model which gave the best RMSE was TES and Manual SARIMAX (4,1,2)(3,0,2,12). We will built a final model on the entire Rose dataset using SARIMAX.

y	mean	mean_se	mean_ci_lower	mean_ci_upper
Time				
1995-08-31	46.413681	11.969449	22.953992	69.873369
1995-09-30	48.793911	12.039961	25.196021	72.391800
1995-10-31	47.508266	12.108541	23.775961	71.240571
1995-11-30	56.269275	12.121077	32.512402	80.026149
1995-12-31	77.863686	12.121551	54.105882	101.621490
1996-01-31	28.708561	12.214416	4.768746	52.648376
1996-02-29	37.191299	12.374841	12.937057	61.445541
1996-03-31	42.402183	12.562169	17.780784	67.023582
1996-04-30	40.944350	12.728683	15.996590	65.892110
1996-05-31	36.441184	12.842221	11.270894	61.611474
1996-06-30	40.438159	12.933916	15.088148	65.788169
1996-07-31	49.552354	13.022241	24.029231	75.075477





PROBLEM 1.10

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

Resolution:

Sparkling Wine data:

- TES (Triple Exponential Smoothing) has worked the best for the forecast with lowest RMSE on test data
- You can see from the above chart that the forecast for next 12 months is slightly over the sales of the previous 12 months however, there isn't a considerable increase.
- Observed from the month wise bar plots previously, we can say that the sales of Sparkling wine tend to go up in last two months probably because it's a holiday season than the rest and its lowest around Jun and July
- ABC can take various measures to increase the sales towards the beginning and mid of the year, it can introduce promotional activities or discounts during the low sales period.
- ABC can tie up with events like concerts, weddings etc. and do some sponsorships to boost sales during the slack

Rose Wine data:

- We chose manual SARIMAX model to predict for the Rose wine data. The model was passed the cut offs found through ACF and PACF plots of q and p respectively and seasonality of 12 as the plots showed a patterned significance after 11 lags.
- You can see from the above plot for Rose wine data the forecast for 1996 is more or less same as of for 1995.
- Observed from the monthly bar plot sales shows an increasing trend from August towards December, it’s on the lower side beginning of the year
- ABC can take sought promotional activities and implement some discounts during the first half of the year

The End

Thakur Arun Singh

***************************^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^*************************