Problem:

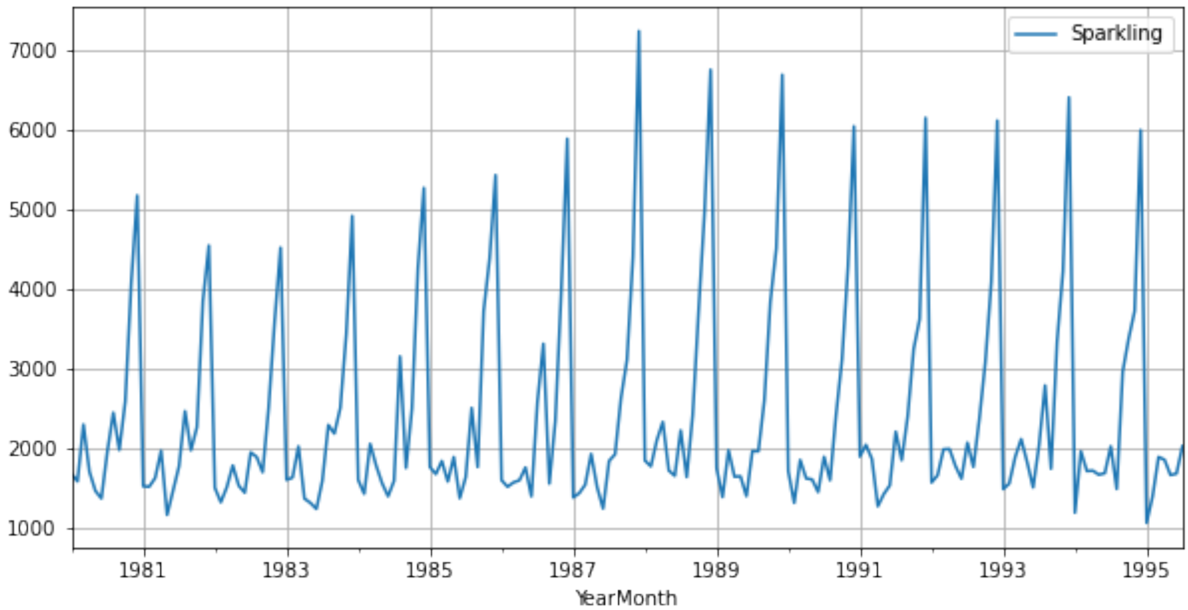
For this particular assignment, the data of different types of wine sales in the 20th century is to be analysed. Both of these data are from the same company but of different wines. As an analyst in the ABC Estate Wines, you are tasked to analyse and forecast Wine Sales in the 20th century.

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

Sparkling 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)
```

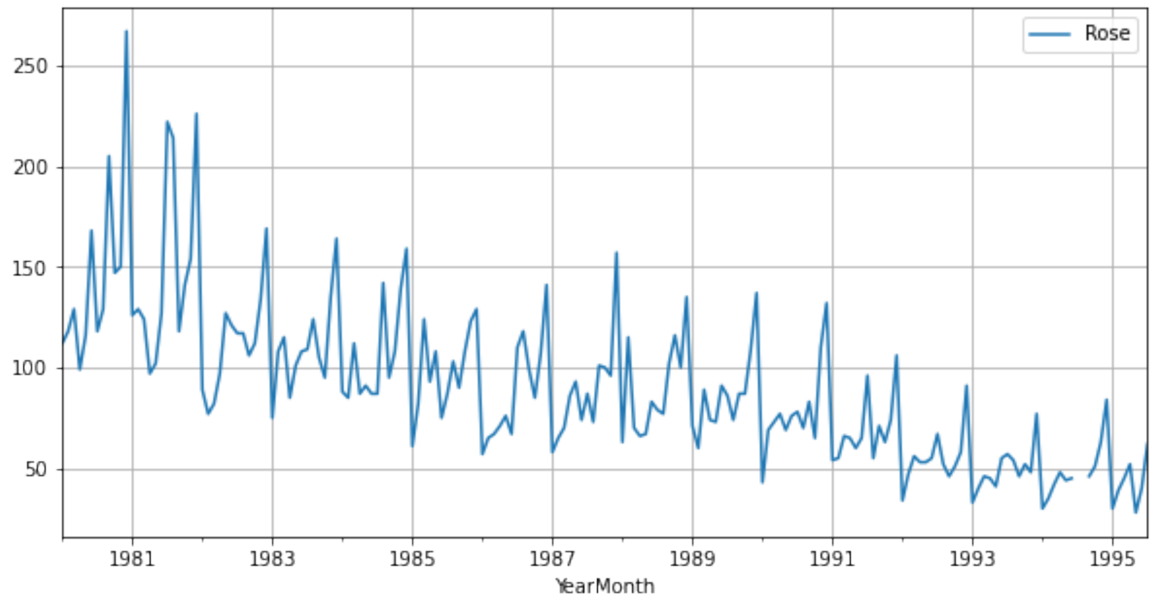
YearMonth	Sparkling
1980-01-01	1686
1980-02-01	1591
1980-03-01	2304
1980-04-01	1712
1980-05-01	1471



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



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

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

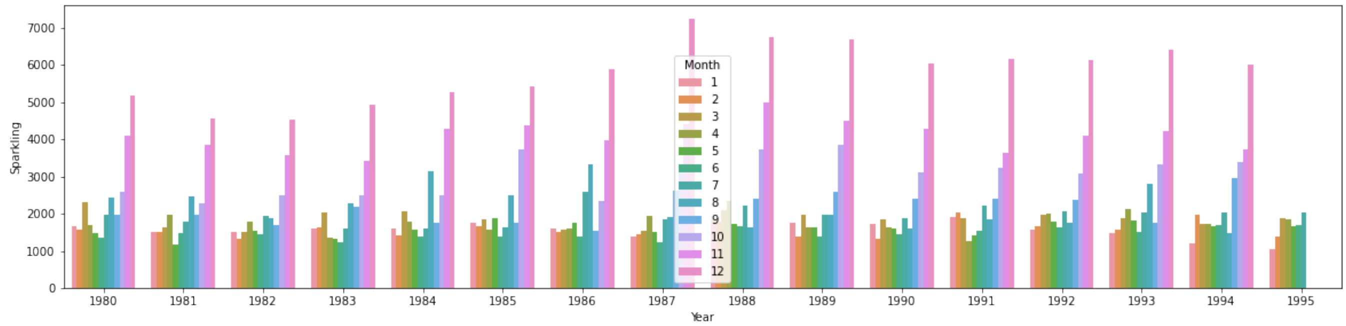
Missing values from the data were:

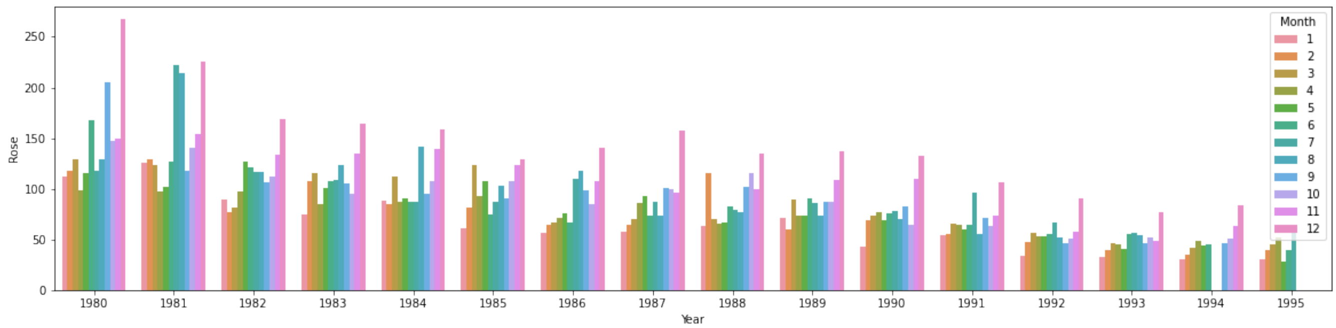
```
Sparkling    0
dtype: int64

Rose     2
dtype: int64

Rose     0
dtype: int64
```

Year-month wise bar chart:





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

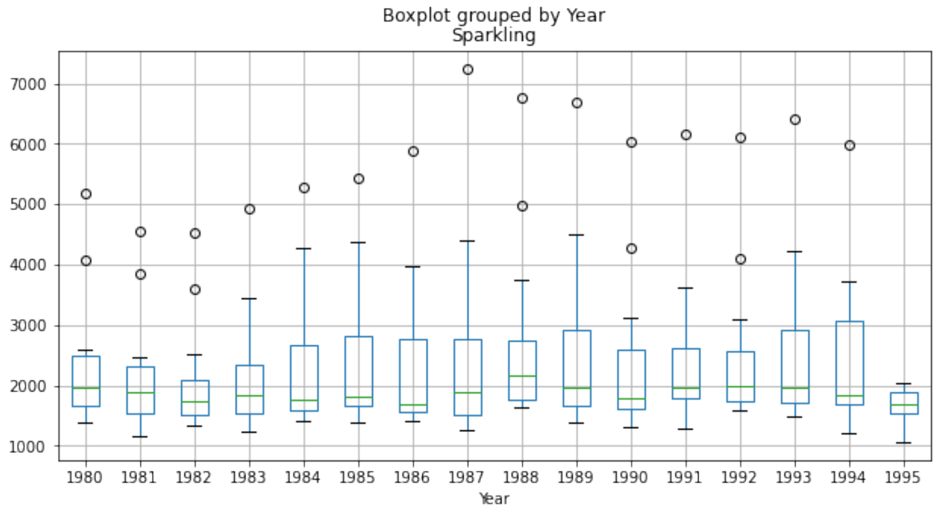
'Sparkling:'

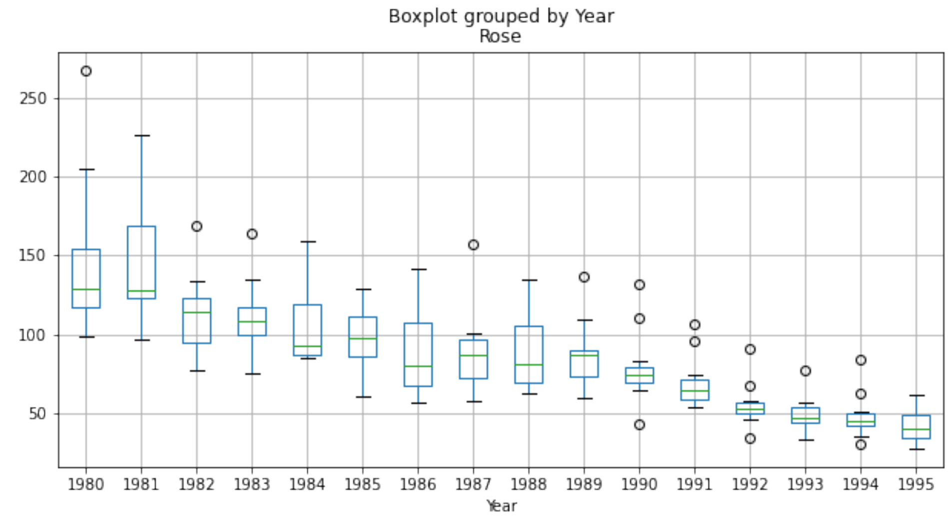
Month Year	1	2	3	4	5	6	7	8	9	10	11	12
1980	1686.0	1591.0	2304.0	1712.0	1471.0	1377.0	1966.0	2453.0	1984.0	2596.0	4087.0	5179.0
1981	1530.0	1523.0	1633.0	1976.0	1170.0	1480.0	1781.0	2472.0	1981.0	2273.0	3857.0	4551.0
1982	1510.0	1329.0	1518.0	1790.0	1537.0	1449.0	1954.0	1897.0	1706.0	2514.0	3593.0	4524.0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

'Rose:'

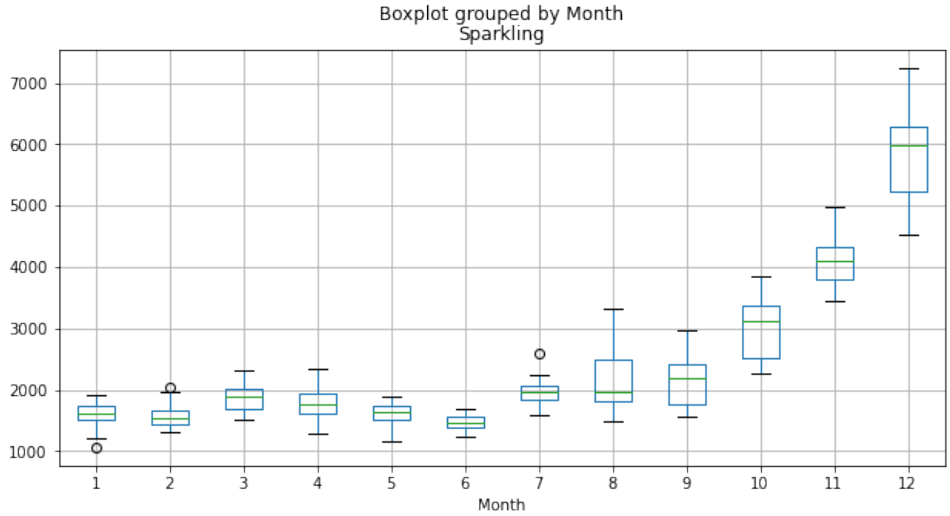
Month Year	1	2	3	4	5	6	7	8	9	10	11	12
1980	112.0	118.0	129.0	99.0	116.0	168.0	118.0	129.0	205.0	147.0	150.0	267.0
1981	126.0	129.0	124.0	97.0	102.0	127.0	222.0	214.0	118.0	141.0	154.0	226.0
1982	89.0	77.0	82.0	97.0	127.0	121.0	117.0	117.0	106.0	112.0	134.0	169.0
1983	75.0	108.0	115.0	85.0	101.0	108.0	109.0	124.0	105.0	95.0	135.0	164.0
1984	88.0	85.0	112.0	87.0	91.0	87.0	87.0	142.0	95.0	108.0	139.0	159.0
1985	61.0	82.0	124.0	93.0	108.0	75.0	87.0	103.0	90.0	108.0	123.0	129.0
1986	57.0	65.0	67.0	71.0	76.0	67.0	110.0	118.0	99.0	85.0	107.0	141.0
1987	58.0	65.0	70.0	86.0	93.0	74.0	87.0	73.0	101.0	100.0	96.0	157.0
1988	63.0	115.0	70.0	66.0	67.0	83.0	79.0	77.0	102.0	116.0	100.0	135.0
1989	71.0	60.0	89.0	74.0	73.0	91.0	86.0	74.0	87.0	87.0	109.0	137.0
1990	43.0	69.0	73.0	77.0	69.0	76.0	78.0	70.0	83.0	65.0	110.0	132.0
1991	54.0	55.0	66.0	65.0	60.0	65.0	96.0	55.0	71.0	63.0	74.0	106.0
1992	34.0	47.0	56.0	53.0	53.0	55.0	67.0	52.0	46.0	51.0	58.0	91.0
1993	33.0	40.0	46.0	45.0	41.0	55.0	57.0	54.0	46.0	52.0	48.0	77.0
1994	30.0	35.0	42.0	48.0	44.0	45.0	NaN	NaN	46.0	51.0	63.0	84.0
1995	30.0	39.0	45.0	52.0	28.0	40.0	62.0	NaN	NaN	NaN	NaN	NaN

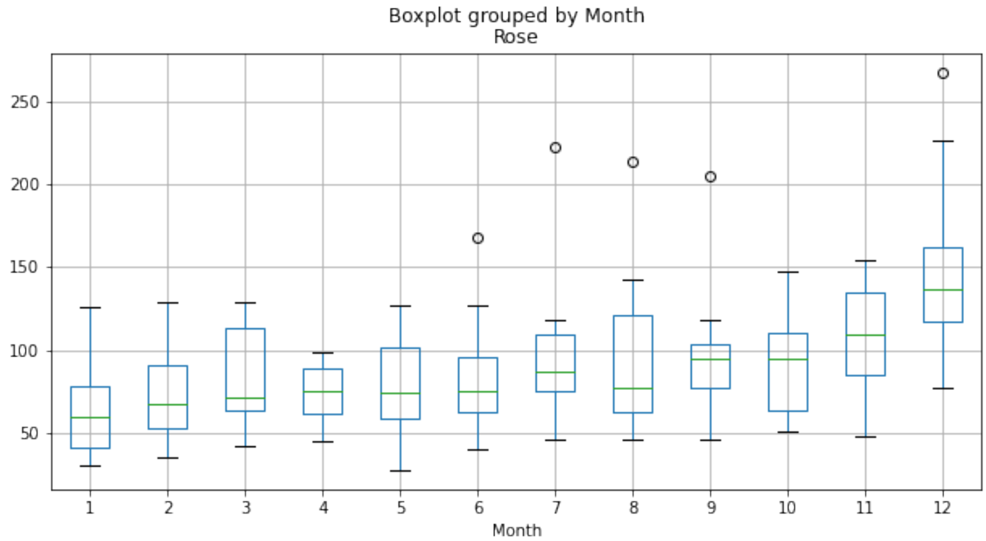
Yearly Boxplots





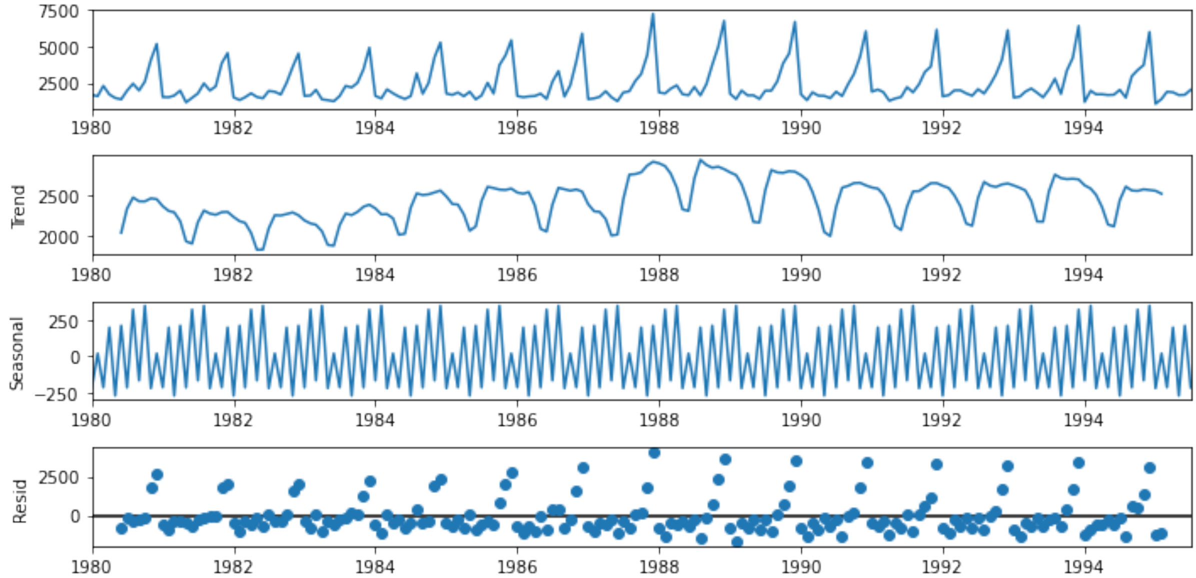
Monthly Boxplots:



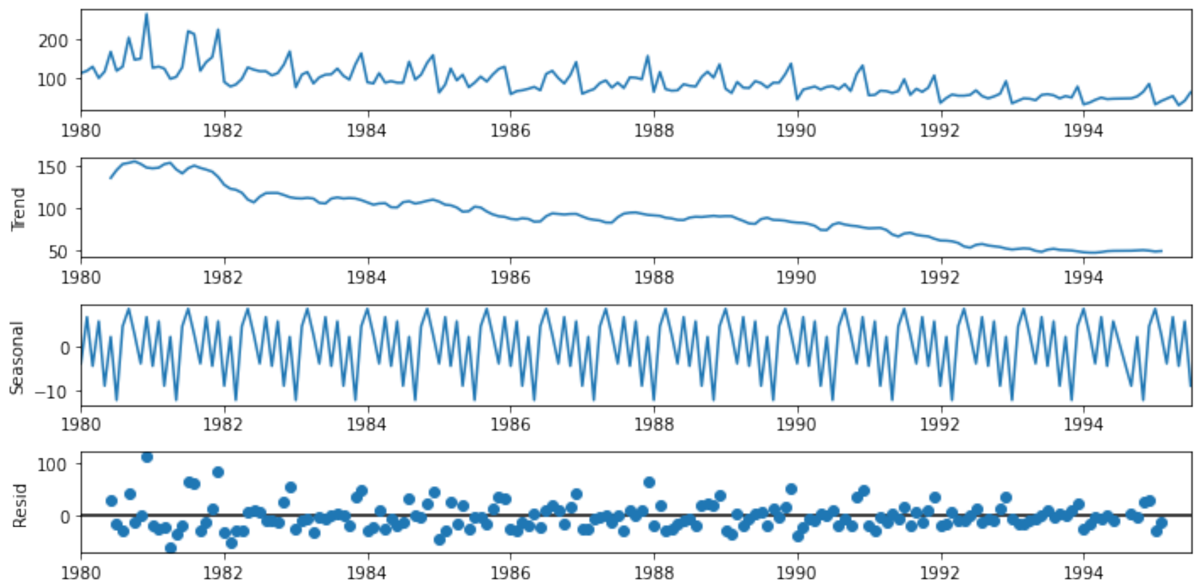


Additive Decomposition:

Sparkling:

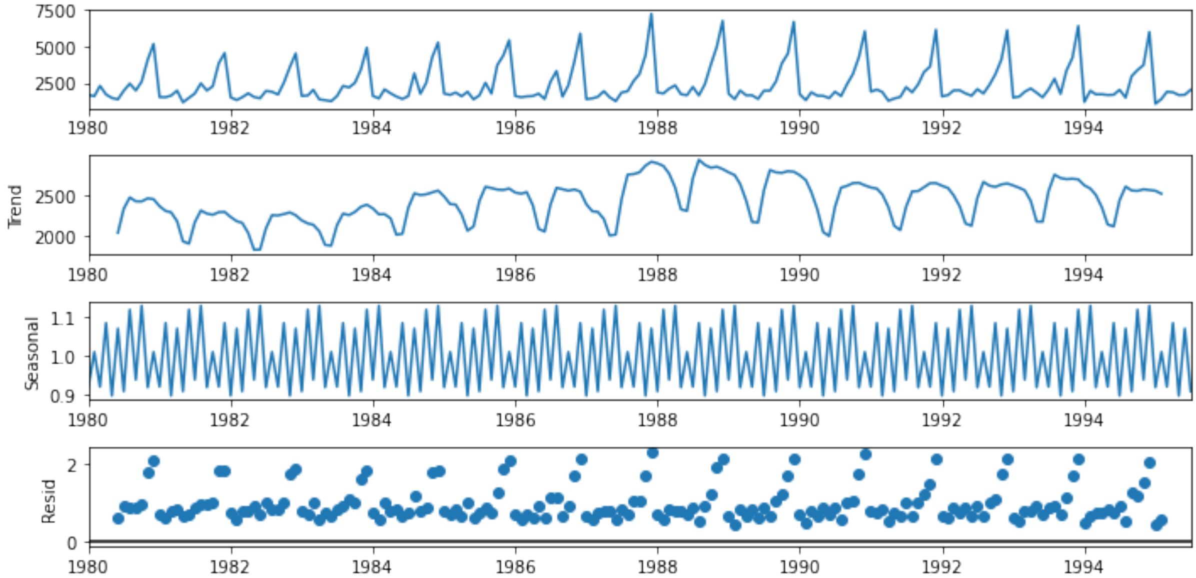


Rose:

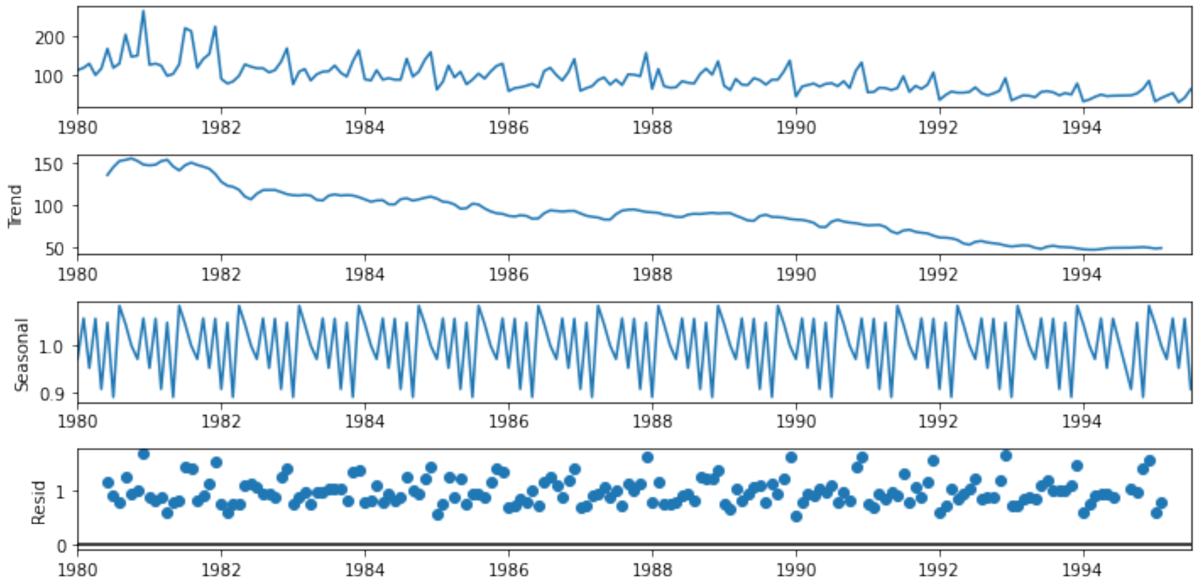


Multiplicative:

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 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

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

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

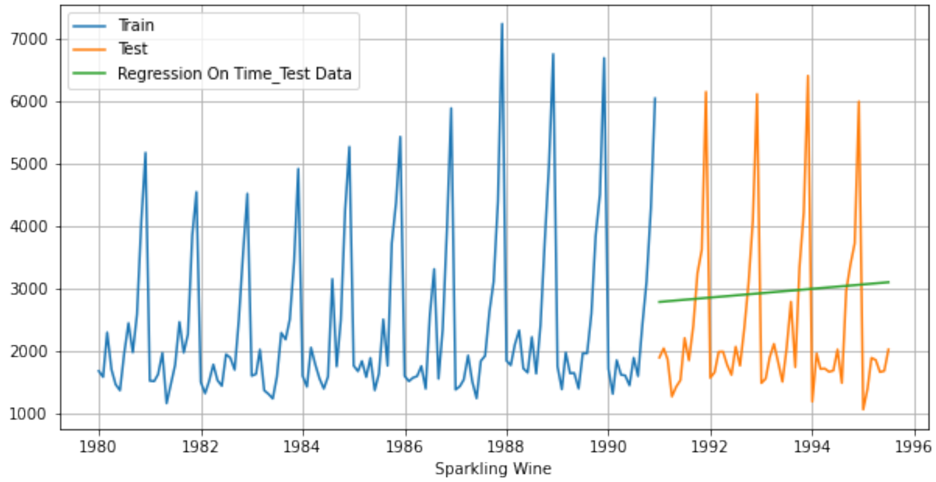
```
Test Data: (55, 3)
```

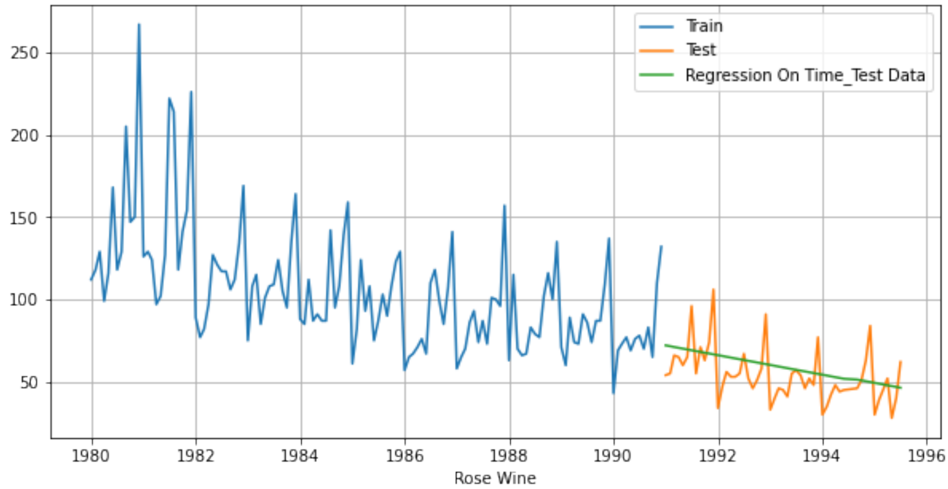
YearMonth	Rose	Year	Month
1991-01-01	54.0	1991	1
1991-02-01	55.0	1991	2
1991-03-01	66.0	1991	3
1991-04-01	65.0	1991	4
1991-05-01	60.0	1991	5

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

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. Please do try to build as many models as possible and as many iterations of models as possible with different parameters.

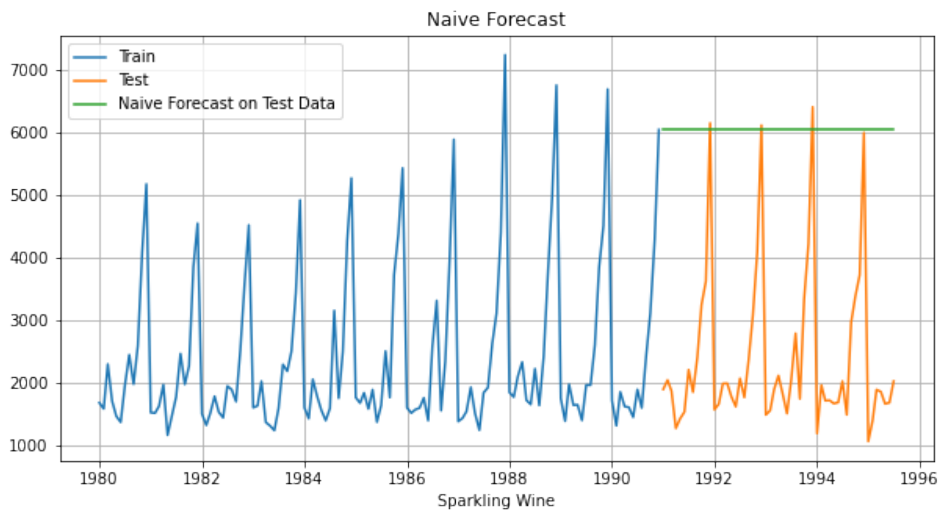
Model 1: Linear Regression: $\hat{y}$ t+1 = β y + c

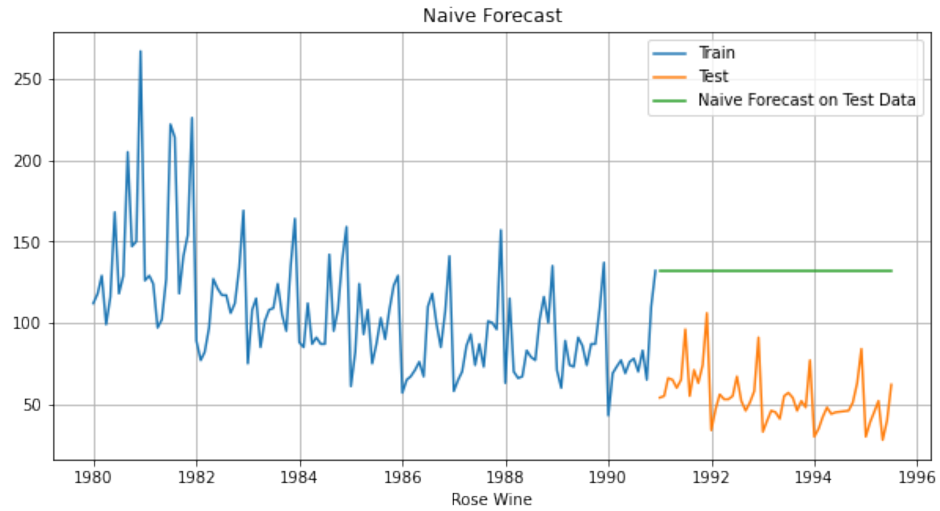




Model 2: Naive Approach: ŷ t+1=yt

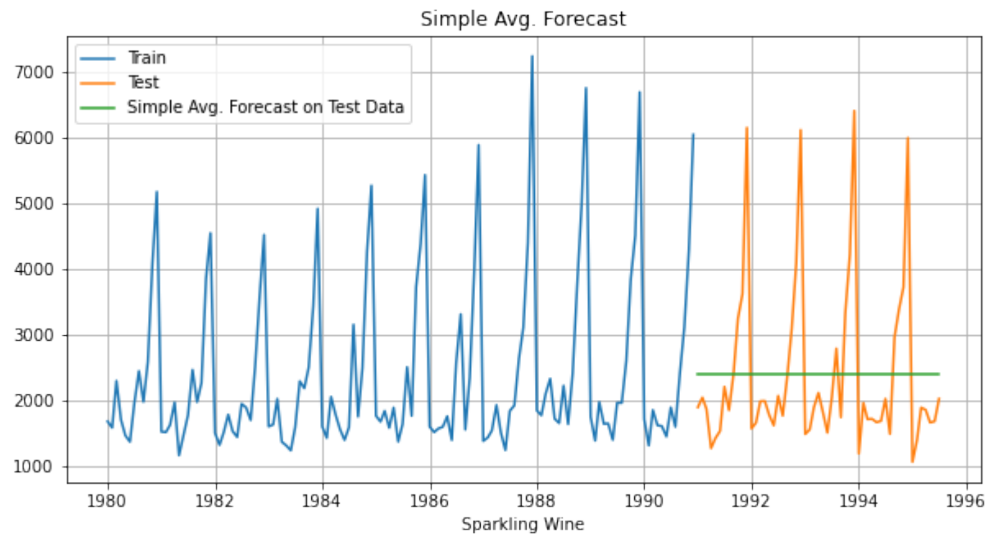
For this particular naive model, we say that the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow and since the prediction of tomorrow is same as today, therefore the prediction for day after tomorrow is also today.

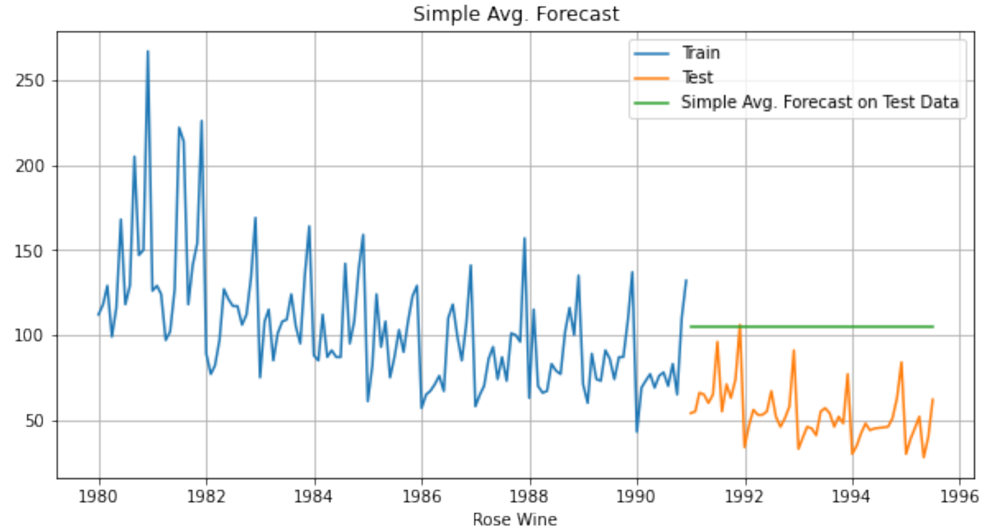




Method 3: Simple Average:

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

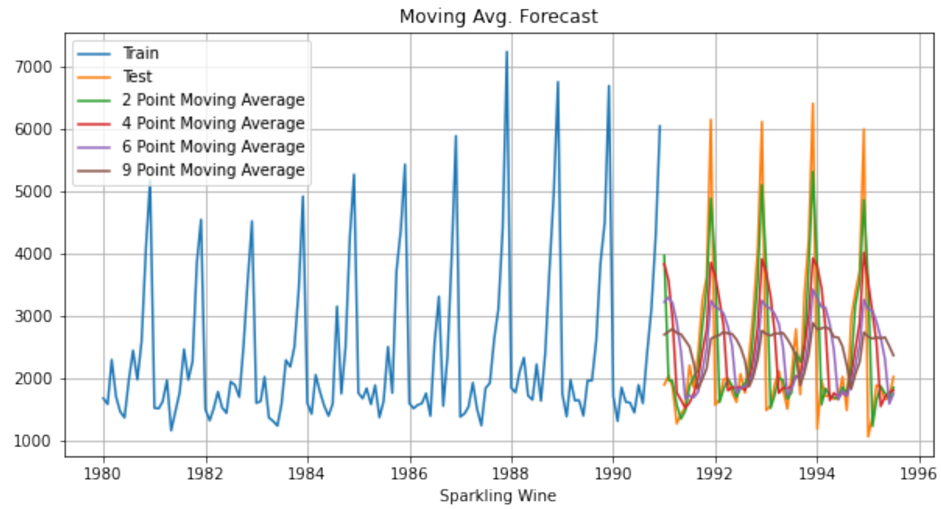


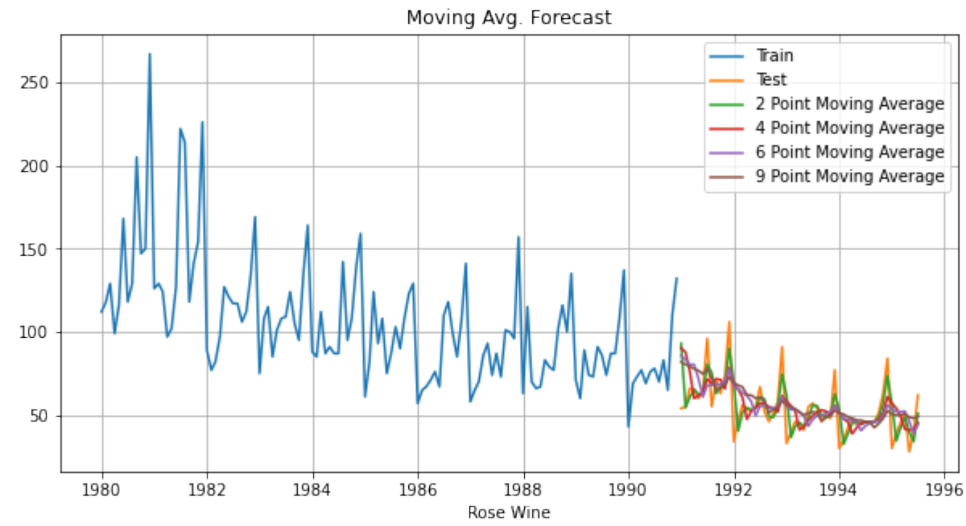


Method 4: Moving Average(MA)

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

$\hat{y}_{t+1}=\alpha Y_t+(1-\alpha)\hat{y}_t$ 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 forecating the short term avarage value or level and the other for capturing the trend.

Intercept or Level equation, $\hat{y}_t$ is given by: $\hat{y}_t=\alpha y_t+(1-\alpha)\hat{y}_t$ Trend equation is given by $T_t=\beta(\hat{y}_t-\hat{y}_{t-1})+(1-\beta)T_{t-1}$ 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

$F_{t+1}=\hat{y}_t+T_t$ $F_{t+n}=\hat{y}_t+nT_t$

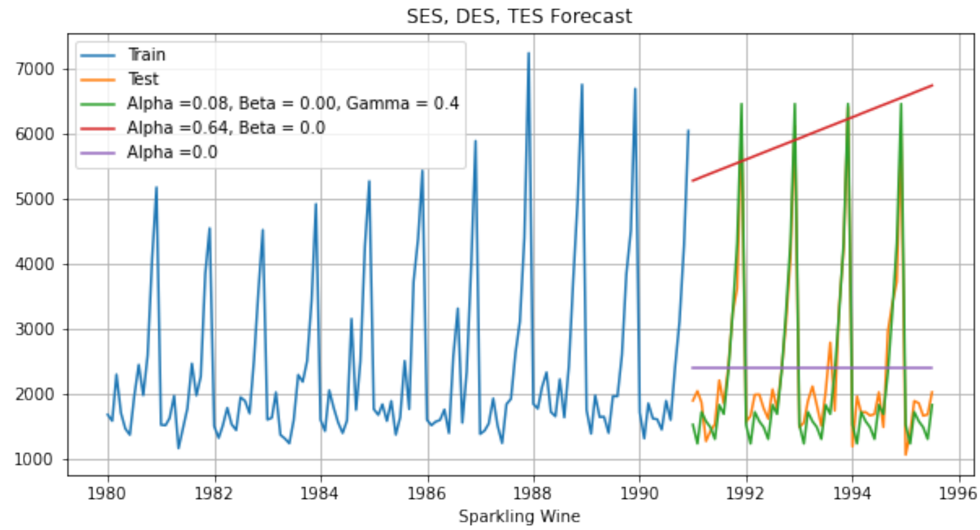
Though our Sparkling data doesn't seem to have a visible trend we are still going to build this model for the project. Rose data has a clear trend from the plot above

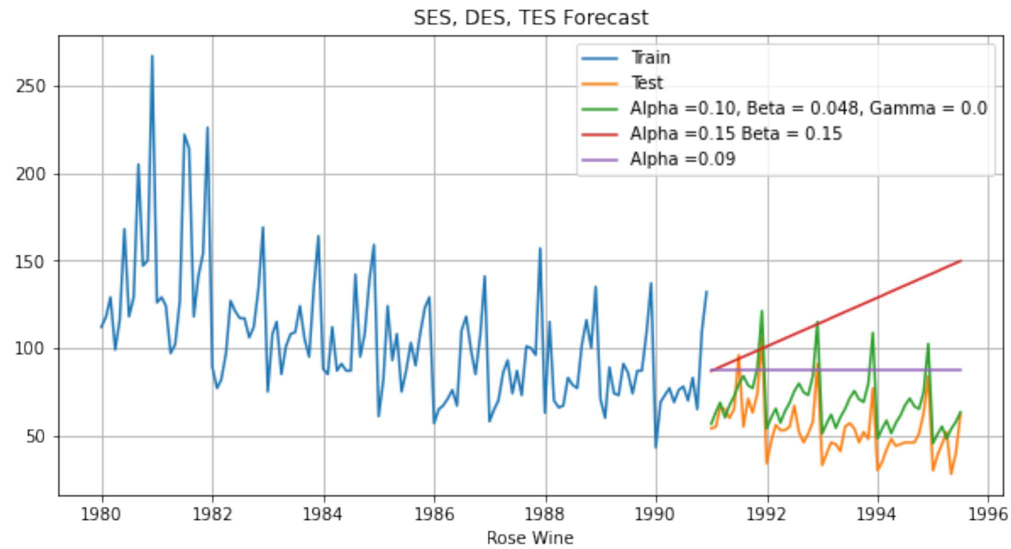
Inference

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.

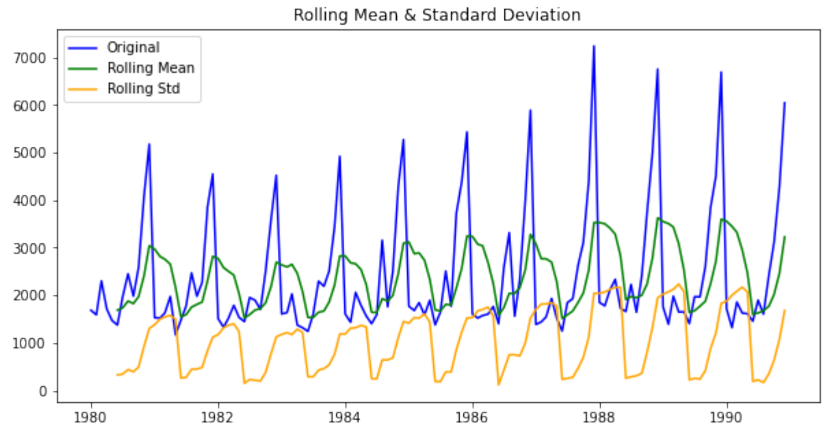
We will use ETS(A, A, A) Holt Winter's linear method with additive trend and seasonality for Sparkling data and ETS(A, A, M) Holt Winter's linear method with additive trend and multiplicative seasonality for Rose wine data. We will call it Triple Exponential Smoothing(TES)





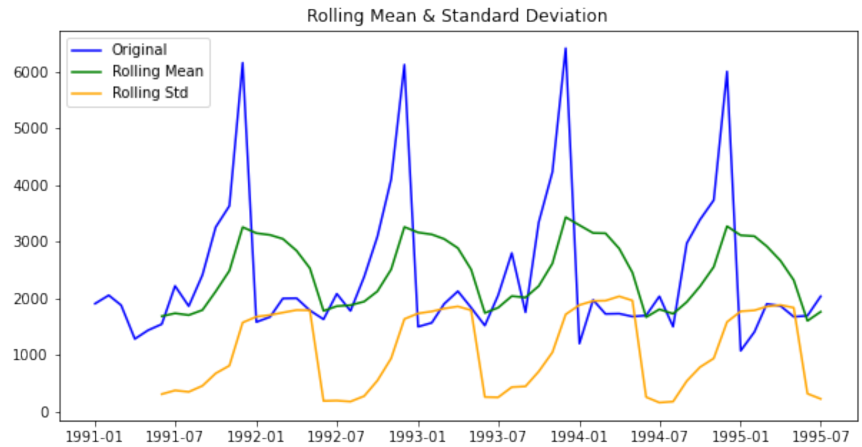
5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05.

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:

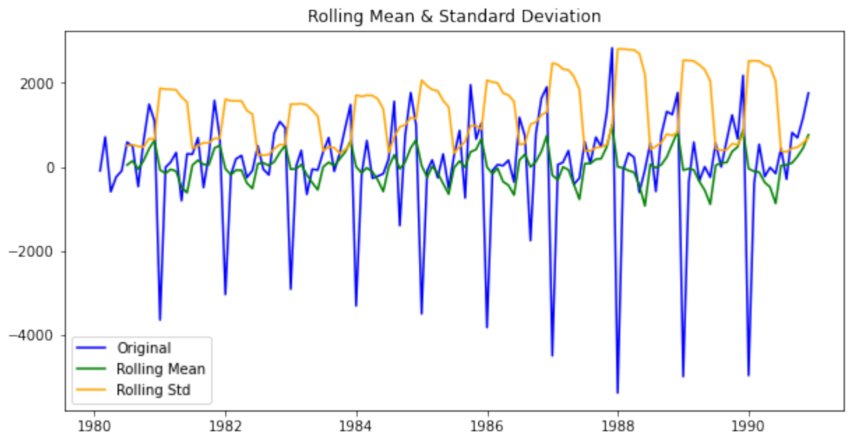


```
Results of Dickey-Fuller Test:
Test Statistic                  -1.790189
p-value                          0.385343
#Lags Used                      11.000000
Number of Observations Used     43.000000
Critical Value (1%)             -3.592504
Critical Value (5%)             -2.931550
Critical Value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

since the,
Null Hypothesis H0: The series is non stationary
Alternate Hypothesis H1: The series is stationary

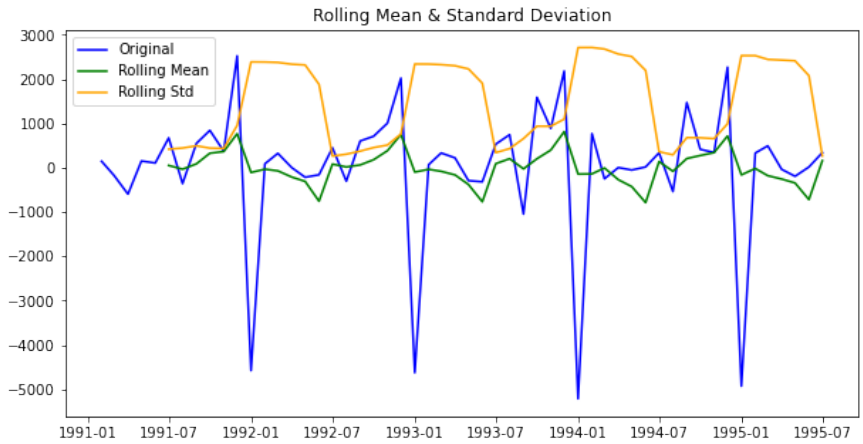
we cannot reject the null as the p values is greater than 0.05 (significance level) from the Augmented Dickey Fuller test above for both Train and Test of Sparkling Wine dataset
We can correct the non-stationarity by using multiple methods like taking differences at various level, using logged transformed series etc.
Here we will take difference of level 1 of the original series.

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
```

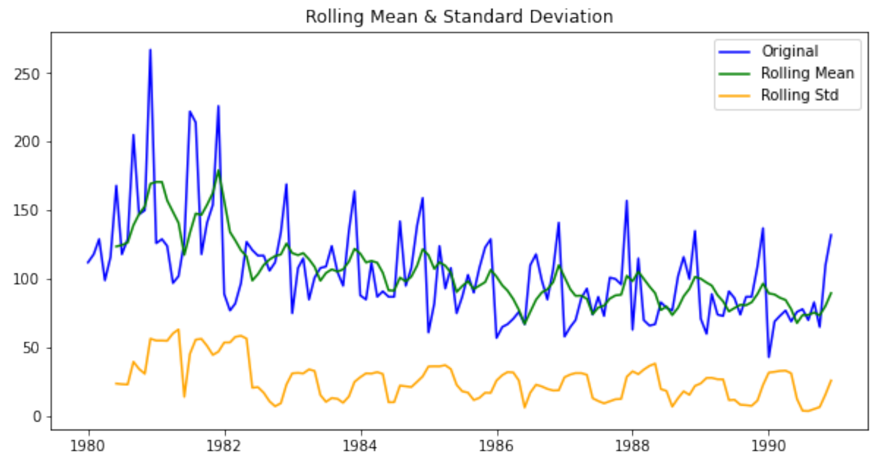
Differenced Sparkling Test set:




```
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
```

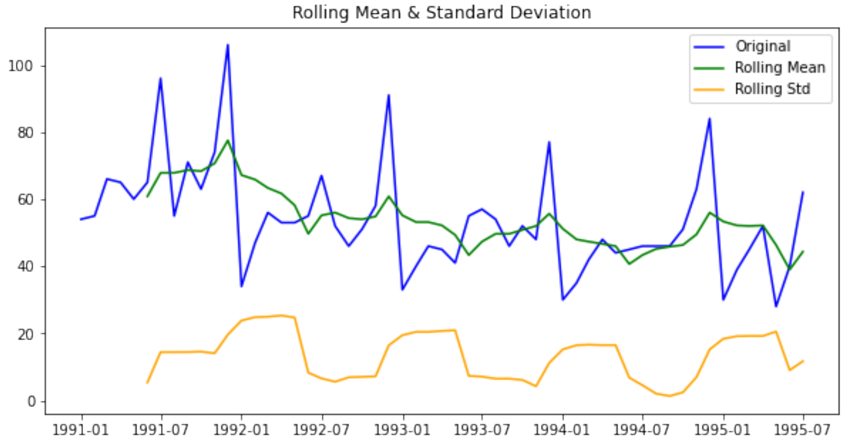
We can now see that the p –value < than 0.05 so we can reject the null-hypothesis and accept the alternate. So we say the series is stationary

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:

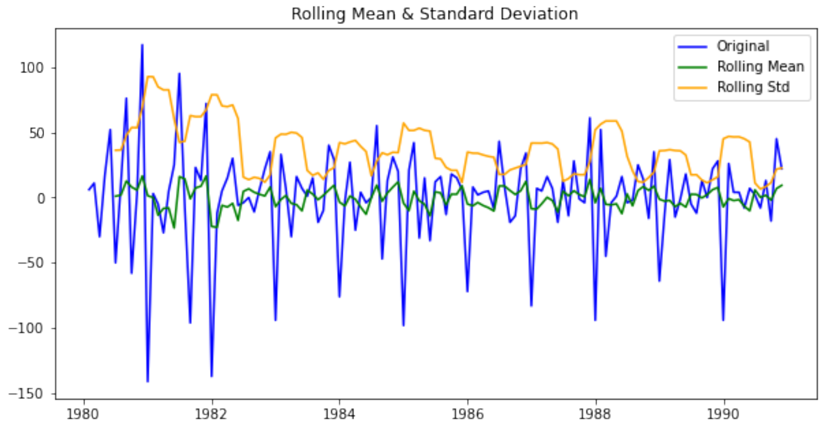



```
Results of Dickey-Fuller Test:
Test Statistic                 -4.464772
p-value                         0.000228
#Lags Used                     11.000000
Number of Observations Used    43.000000
Critical Value (1%)            -3.592504
Critical Value (5%)            -2.931550
Critical Value (10%)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

6. Bu 6. Build an automated version of the ARIMA 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.

ARIMA

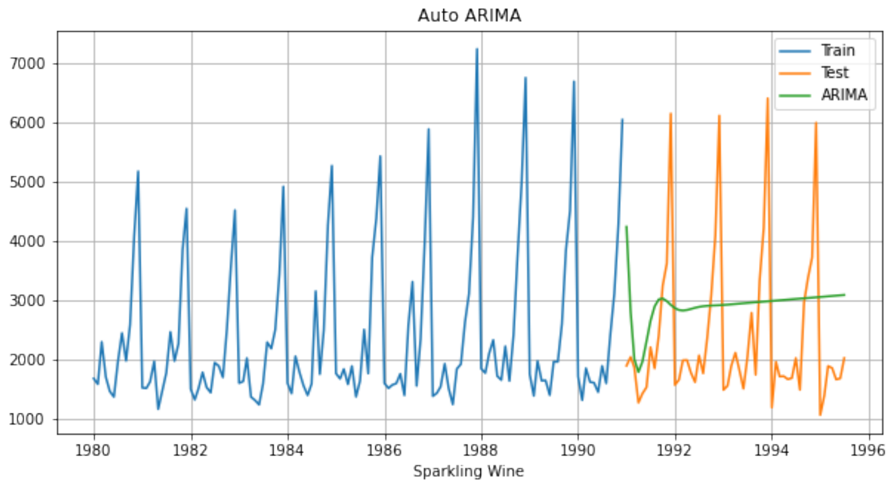
AIC score for both Sparkling and Rose wine dataset for different models is below:

	param	AIC_Sparkling
8	(2, 1, 2)	2210.616439
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)	2264.906437
3	(1, 1, 0)	2268.528061
0	(0, 1, 0)	2269.582796

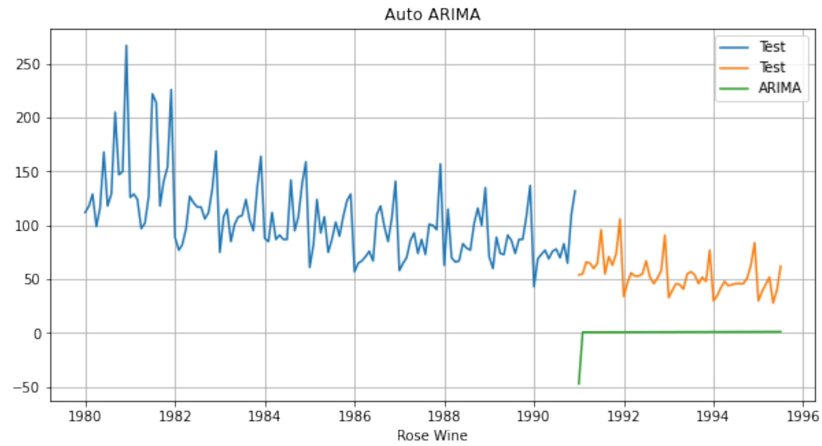
	param	AIC_Rose
2	(0, 1, 2)	1276.835372
5	(1, 1, 2)	1277.359225
4	(1, 1, 1)	1277.775748
7	(2, 1, 1)	1279.045689
8	(2, 1, 2)	1279.298694
1	(0, 1, 1)	1280.726183
6	(2, 1, 0)	1300.609261
3	(1, 1, 0)	1319.348311
0	(0, 1, 0)	1335.152658

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           1011.656
Date:                Thu, 29 Oct 2020   AIC                           2210.616
Time:                        21:11:20   BIC                           2227.868
Sample:                    02-01-1980   HQIC                          2217.626
                         - 12-01-1990
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 5.5857      0.516     10.820      0.000       4.574       6.598
ar.L1.D.Sparkling     1.2699      0.074     17.046      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    -1.9999      0.042    -47.100      0.000      -2.083      -1.917
ma.L2.D.Sparkling     0.9999      0.042     23.583      0.000       0.917       1.083
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.1336           -0.7074j            1.3362           -0.0888
AR.2            1.1336           +0.7074j            1.3362            0.0888
MA.1            1.0000           -0.0006j            1.0000           -0.0001
MA.2            1.0000           +0.0006j            1.0000            0.0001
-----------------------------------------------------------------------------
```



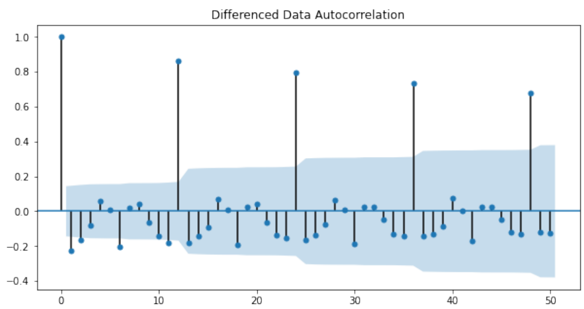

```
Rose Data:
                             ARIMA Model Results
==============================================================================
Dep. Variable:                 D.Rose   No. Observations:                  130
Model:                 ARIMA(0, 1, 2)   Log Likelihood                -636.749
Method:                       css-mle   S.D. of innovations             30.402
Date:                Thu, 29 Oct 2020   AIC                           1281.498
Time:                        21:11:20   BIC                           1292.968
Sample:                    03-01-1980   HQIC                          1286.158
                         - 12-01-1990
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const            0.0091      0.005      2.009      0.044       0.000       0.018
ma.L1.D.Rose    -1.9981      0.038    -51.939      0.000      -2.074      -1.923
ma.L2.D.Rose     0.9981      0.039     25.915      0.000       0.923       1.074
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
MA.1            1.0000           +0.0000j            1.0000            0.0000
MA.2            1.0019           +0.0000j            1.0019            0.0000
-----------------------------------------------------------------------------
```

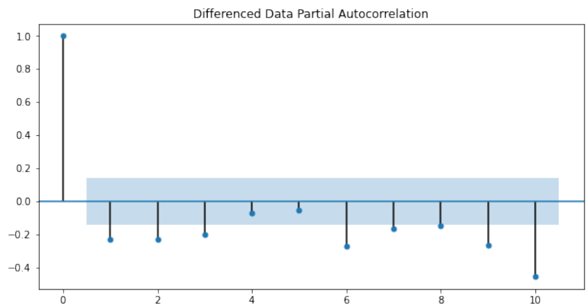


7. Build ARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.

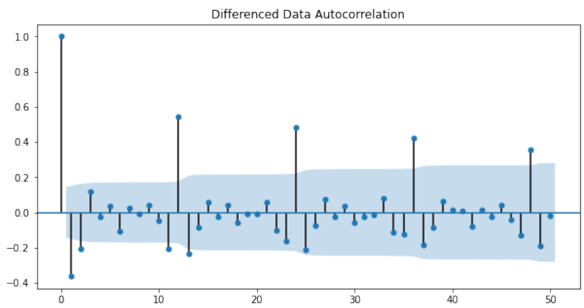
ARIMA

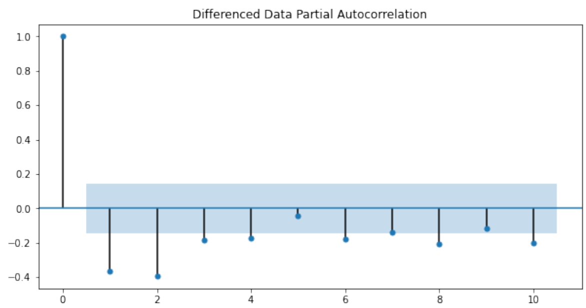
Sparkling Dataset:





Rose Dataset:





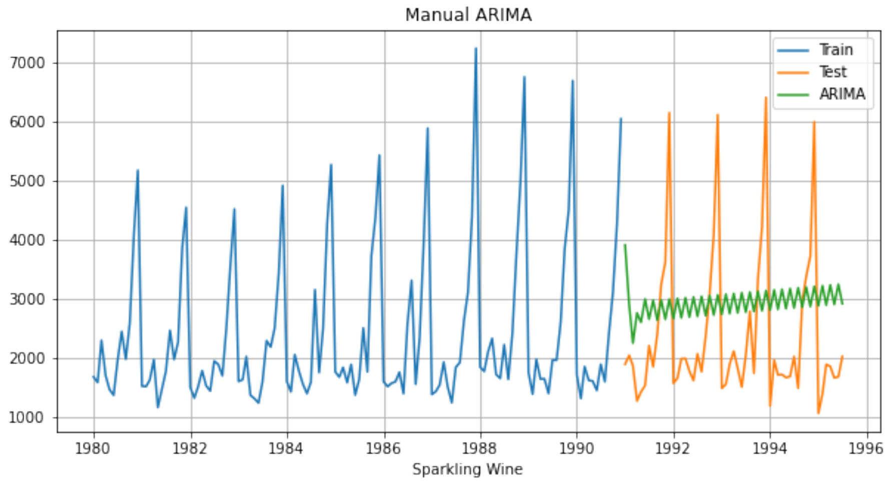
Here, we have taken alpha=0.05.

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.

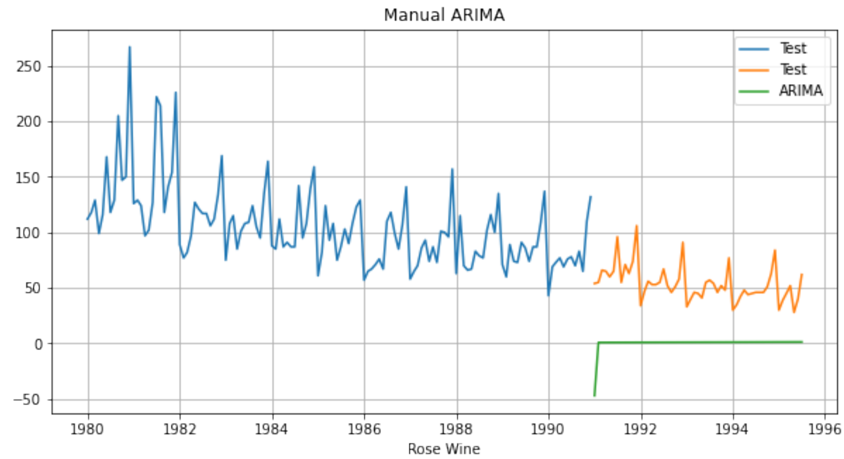
By looking at the above plots for Sparkling data, we can say that both the PACF cuts off at 3 and ACF plot cuts-off at lag 2.

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           1105.900
Date:                Thu, 29 Oct 2020   AIC                           2228.927
Time:                        21:12:00   BIC                           2249.053
Sample:                    02-01-1980   HQIC                          2237.105
                         - 12-01-1990
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 5.9754   8.93e-05   6.69e+04      0.000       5.975       5.976
ar.L1.D.Sparkling    -0.4419        nan        nan        nan         nan         nan
ar.L2.D.Sparkling     0.3079        nan        nan        nan         nan         nan
ar.L3.D.Sparkling    -0.2502        nan        nan        nan         nan         nan
ma.L1.D.Sparkling    -0.0002        nan        nan        nan         nan         nan
ma.L2.D.Sparkling    -0.9998        nan        nan        nan         nan         nan
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1           -1.0000           -0.0000j            1.0000           -0.5000
AR.2            1.1153           -1.6592j            1.9992           -0.1558
AR.3            1.1153           +1.6592j            1.9992            0.1558
MA.1            1.0000           +0.0000j            1.0000            0.0000
MA.2           -1.0002           +0.0000j            1.0002            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:                Fri, 30 Oct 2020   AIC                           1283.753
Time:                        18:50:08   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           -0.4115j            1.1770           -0.0569
AR.2            1.1027           +0.4115j            1.1770            0.0569
AR.3           -0.6862           -1.6643j            1.8003           -0.3122
AR.4           -0.6862           +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
-----------------------------------------------------------------------------
```



6. Build an automated version of the 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.

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
236	(3, 1, 2)	(3, 0, 0, 12)	1387.234717
253	(3, 1, 3)	(3, 0, 1, 12)	1387.322106
220	(3, 1, 1)	(3, 0, 0, 12)	1387.788332
237	(3, 1, 2)	(3, 0, 1, 12)	1388.602607
221	(3, 1, 1)	(3, 0, 1, 12)	1388.681480
...	...	...	...
35	(0, 1, 2)	(0, 0, 3, 12)	7611.935696
227	(3, 1, 2)	(0, 0, 3, 12)	7691.792919
119	(1, 1, 3)	(1, 0, 3, 12)	8630.041823
247	(3, 1, 3)	(1, 0, 3, 12)	8767.539933
23	(0, 1, 1)	(1, 0, 3, 12)	NaN

	param	seasonal	AIC_Rose
222	(3, 1, 1)	(3, 0, 2, 12)	774.400286
238	(3, 1, 2)	(3, 0, 2, 12)	774.880945
220	(3, 1, 1)	(3, 0, 0, 12)	775.426699
221	(3, 1, 1)	(3, 0, 1, 12)	775.495331
252	(3, 1, 3)	(3, 0, 0, 12)	775.561019
...	...	...	...
215	(3, 1, 1)	(1, 0, 3, 12)	NaN
231	(3, 1, 2)	(1, 0, 3, 12)	NaN
235	(3, 1, 2)	(2, 0, 3, 12)	NaN
239	(3, 1, 2)	(3, 0, 3, 12)	NaN
247	(3, 1, 3)	(1, 0, 3, 12)	NaN

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:

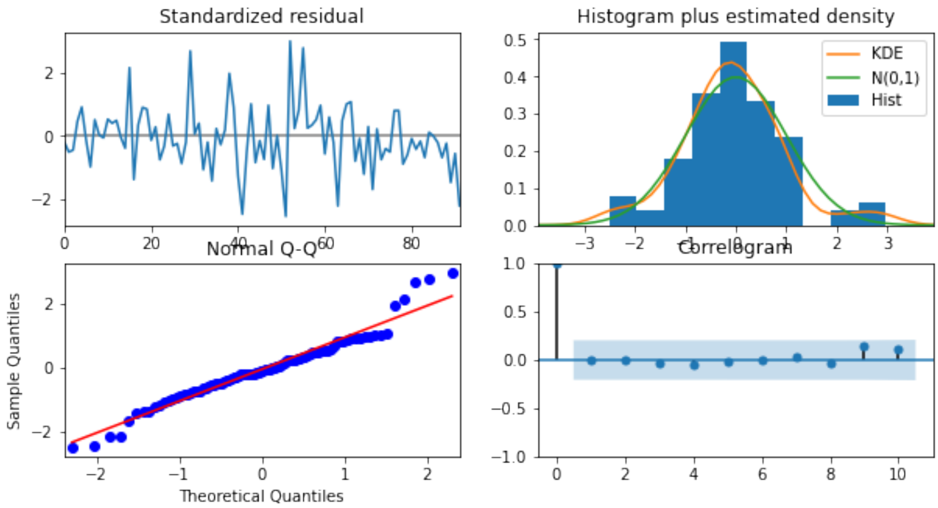
```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(3, 1, 2)x(3, 0, [], 12)   Log Likelihood                -684.617
Date:                           Tue, 03 Nov 2020   AIC                           1387.235
Time:                                   12:24:27   BIC                           1409.931
Sample:                                        0   HQIC                          1396.395
                                           - 132
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5373      0.338     -1.588      0.112      -1.201       0.126
ar.L2          0.0257      0.187      0.137      0.891      -0.340       0.392
ar.L3          0.0785      0.130      0.605      0.545      -0.176       0.333
ma.L1         -0.3365      0.294     -1.143      0.253      -0.914       0.241
ma.L2         -0.7978      0.344     -2.321      0.020      -1.472      -0.124
ar.S.L12       0.5712      0.103      5.541      0.000       0.369       0.773
ar.S.L24       0.2606      0.117      2.223      0.026       0.031       0.490
ar.S.L36       0.2126      0.111      1.915      0.055      -0.005       0.430
sigma2      1.449e+05   2.95e+04      4.912      0.000    8.71e+04    2.03e+05
===================================================================================
Ljung-Box (Q):                       27.31   Jarque-Bera (JB):                 8.81
Prob(Q):                              0.94   Prob(JB):                         0.01
Heteroskedasticity (H):               1.17   Skew:                             0.36
Prob(H) (two-sided):                  0.67   Kurtosis:                         4.33
===================================================================================
```

Rose Data:

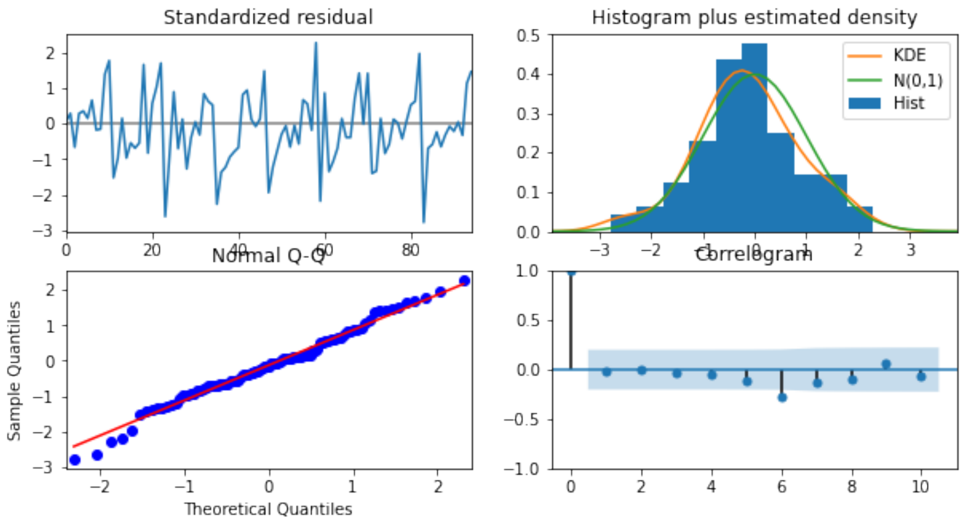
```
                                     SARIMAX Results
===============================================================================================
Dep. Variable:                                     y   No. Observations:                  132
Model:             SARIMAX(3, 1, 1)x(3, 0, [1, 2], 11)   Log Likelihood                -437.103
Date:                               Tue, 03 Nov 2020   AIC                            894.205
Time:                                       12:24:30   BIC                            919.744
Sample:                                            0   HQIC                           904.525
                                               - 132
Covariance Type:                                 opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.1883      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    -12.307      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      -0.0041      0.115     -0.036      0.971      -0.229       0.221
ma.S.L11       0.1809      0.448      0.404      0.686      -0.697       1.059
ma.S.L22      -0.1735      0.234     -0.742      0.458      -0.632       0.285
sigma2       565.7150     98.705      5.731      0.000     372.258     759.172
===================================================================================
Ljung-Box (Q):                      131.48   Jarque-Bera (JB):                 0.17
Prob(Q):                              0.00   Prob(JB):                         0.92
Heteroskedasticity (H):               0.91   Skew:                            -0.06
Prob(H) (two-sided):                  0.80   Kurtosis:                         3.17
===================================================================================
```

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.

7. Build 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.

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:                             Tue, 03 Nov 2020   AIC                            766.161
Time:                                     12:26:52   BIC                            796.292
Sample:                                          0   HQIC                           778.317
                                             - 132
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.7987      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         -0.7650      0.183     -4.186      0.000      -1.123      -0.407
ar.S.L12       0.7670      0.165      4.637      0.000       0.443       1.091
ar.S.L24       0.0838      0.149      0.565      0.572      -0.207       0.375
ar.S.L36       0.0764      0.093      0.823      0.411      -0.106       0.259
ma.S.L12      -0.5258      0.288     -1.824      0.068      -1.091       0.039
ma.S.L24      -0.2330      0.230     -1.013      0.311      -0.684       0.218
sigma2       181.3252     39.762      4.560      0.000     103.392     259.258
===================================================================================
Ljung-Box (Q):                       32.58   Jarque-Bera (JB):                 0.93
Prob(Q):                              0.79   Prob(JB):                         0.63
Heteroskedasticity (H):               1.24   Skew:                             0.25
Prob(H) (two-sided):                  0.56   Kurtosis:                         2.99
===================================================================================
```

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

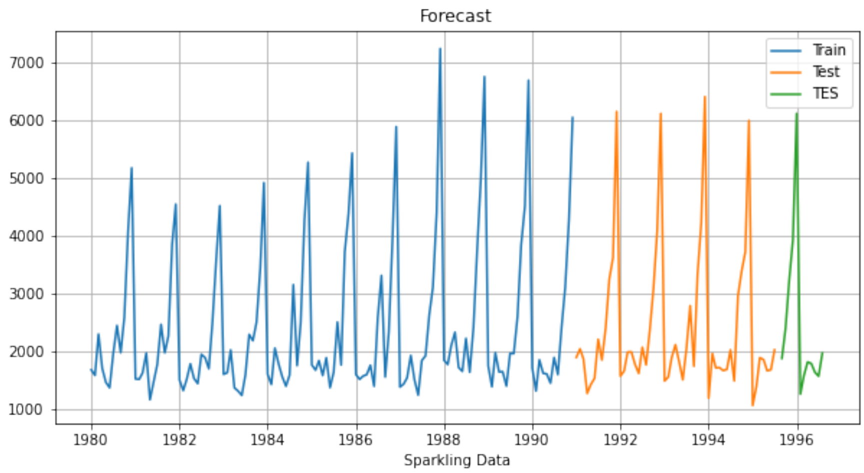
	Test_Spark RMSE
Regression	1389.135175
NaiveModel	3864.279352
SimpleAvg	1275.081804
MovingAvg2	813.400684
MovingAvg4	1156.589694
MovingAvg6	1283.927428
MovingAvg9	1346.278315
SES	1275.081808
DES	3851.279016
TES	362.722421
Auto ARIMA (2,1,2)	1375.217459
Manual ARIMA (3,1,2)	1378.503207
Auto SARIMA (3,1,2)(3,0,0,12)	542.992268

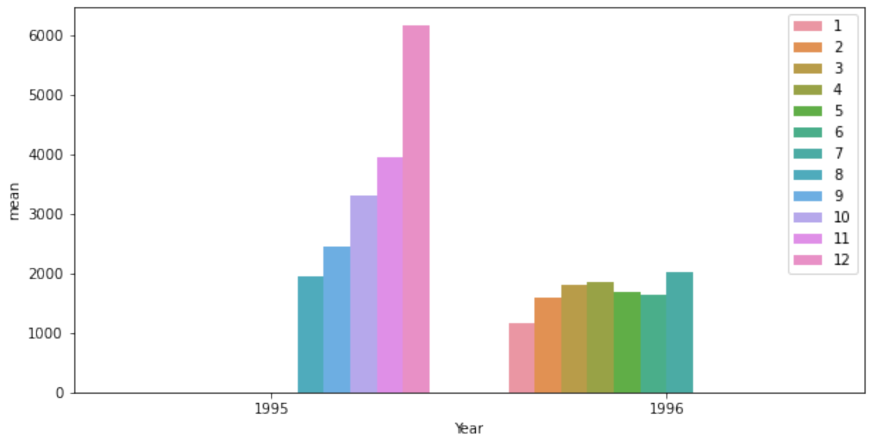
	Test_Rose RMSE
Regression	15.262509
NaiveModel	79.699093
SimpleAvg	53.440426
MovingAvg2	11.529409
MovingAvg4	14.448930
MovingAvg6	14.560046
MovingAvg9	14.724503
SES	36.775789
DES	70.549148
TES	17.345537
Auto ARIMA (0,1,2)	56.295815
Manual ARIMA (4,1,2)	33.930714
Auto SARIMA (3,1,1)(3,0,2,12)	38.033934
Manual SARIMA (4,1,2)(3,0,2,12)	18.292515

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.

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

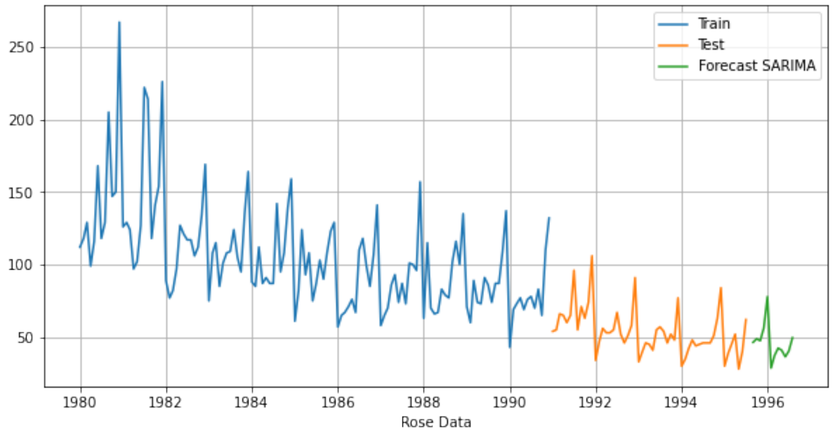
Time	Sparkling Forecast	lower CI	upper CI
1995-08-31	1884.976769	1098.923918	2671.029620
1995-09-30	2402.258496	1616.205645	3188.311348
1995-10-31	3245.977232	2459.924381	4032.030084
1995-11-30	3932.213204	3146.160352	4718.266055
1995-12-31	6119.724082	5333.671230	6905.776933
1996-01-31	1266.116913	480.064062	2052.169764
1996-02-29	1583.646638	797.593787	2369.699490
1996-03-31	1821.829048	1035.776197	2607.881900
1996-04-30	1795.729426	1009.676575	2581.782277
1996-05-31	1643.054809	857.001958	2429.107661
1996-06-30	1576.941975	790.889124	2362.994826
1996-07-31	1975.093831	1189.040980	2761.146683

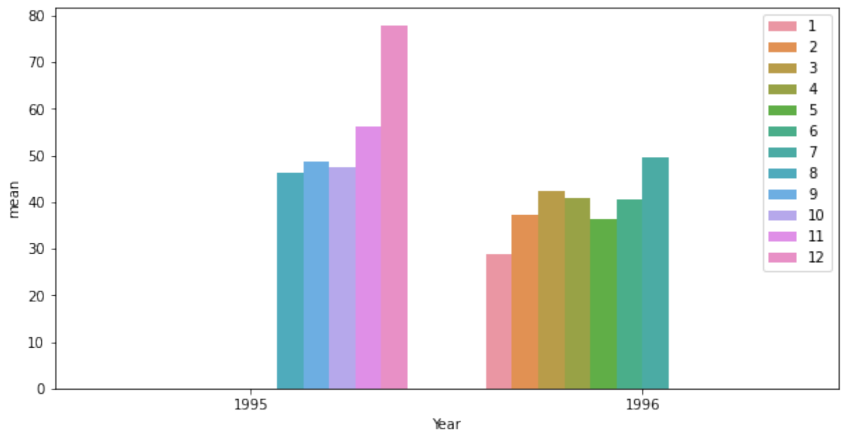




For Rose dataset rolling avg. shows the best RMSE, however since the window chosen was very small (2,4,6,9) it was natural it was going to work well on Test set. The other model which gave the best RMSE was TES and Manual SARIMAX (4,1,2)(3,0,2,12). We will built a final model on the entire Rose dataset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

y	mean	mean_se	mean_ci_lower	mean_ci_upper
Time				
1995-08-31	46.413218	11.968861	22.954681	69.871754
1995-09-30	48.794426	12.039313	25.197806	72.391047
1995-10-31	47.508517	12.107956	23.777360	71.239674
1995-11-30	56.269481	12.120556	32.513628	80.025334
1995-12-31	77.865551	12.121030	54.108768	101.622334
1996-01-31	28.706389	12.213962	4.767464	52.645315
1996-02-29	37.190388	12.374242	12.937319	61.443457
1996-03-31	42.401945	12.561326	17.782199	67.021690
1996-04-30	40.940838	12.727661	15.995080	65.886595
1996-05-31	36.442687	12.841096	11.274601	61.610773
1996-06-30	40.435392	12.932768	15.087632	65.783152
1996-07-31	49.548096	13.021093	24.027223	75.068969





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

Sparkling Wine data:

1. TES (Triple Exponential Smoothing) has worked the best for the forecast with lowest RMSE on test data

2. 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.
3. 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
4. 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.
5. ABC can tie up with events like concerts, weddings etc. and do some sponsorships to boost sales during the slack

Rose Wine data:

6. 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.
7. You can see from the above plot for Rose wine data the forecast for 1996 is more or less same as of for 1995.
8. 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
9. ABC can take sought promotional activities and implement some discounts during the first half of the year