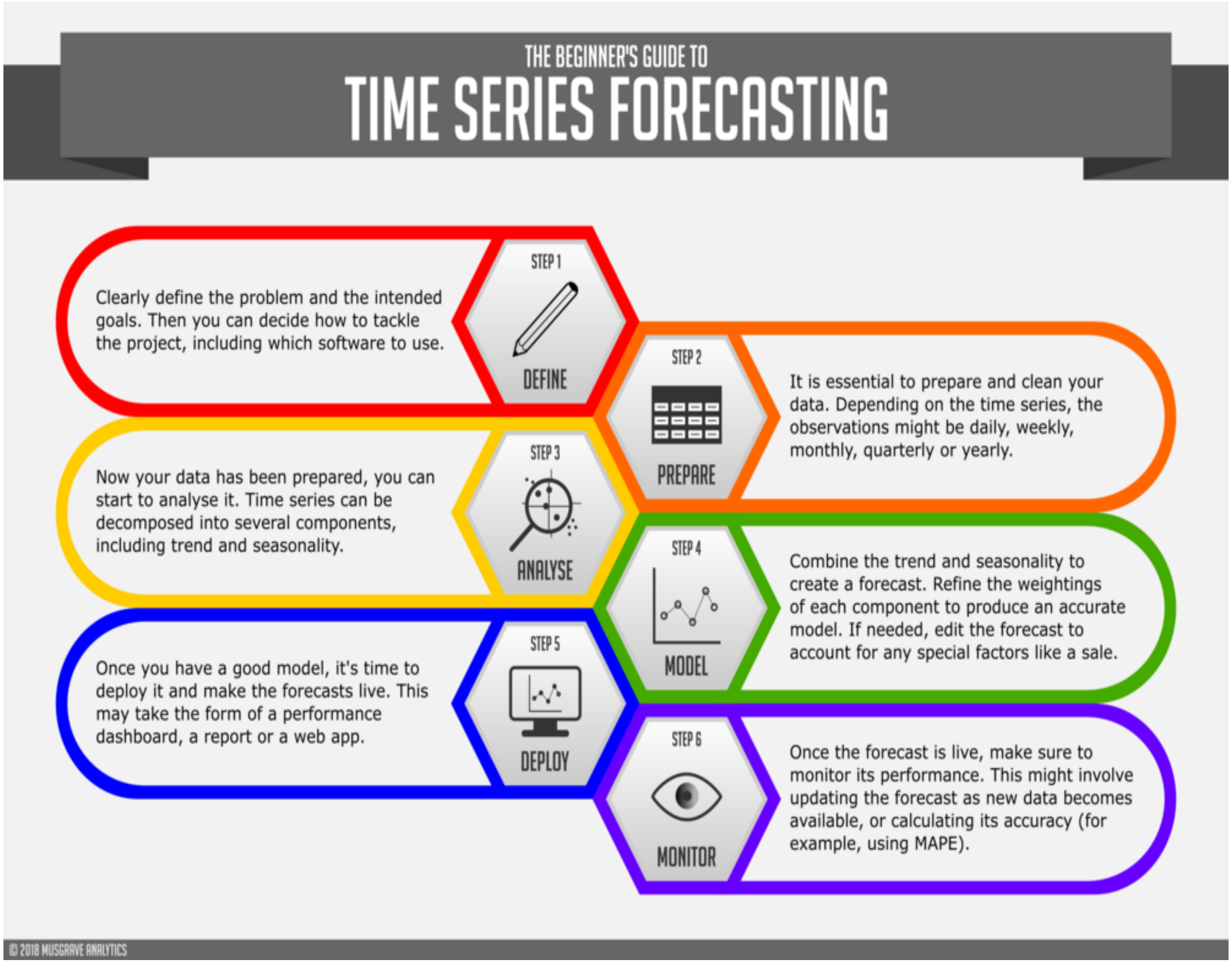
BARNALI PAUL

PROJECT: TIME SERIES FORECASTING

MARCH 20, 2022



CONTENT

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.

Data set for the Problem: Sparkling.csv and Rose.csv

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

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

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

4. **Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other additional models such as regression, naïve forecast models, simple average models, moving average models should also be built on the training data and check the performance on the test data using RMSE.**

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.

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.

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.

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

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.

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

OBJECTIVE:

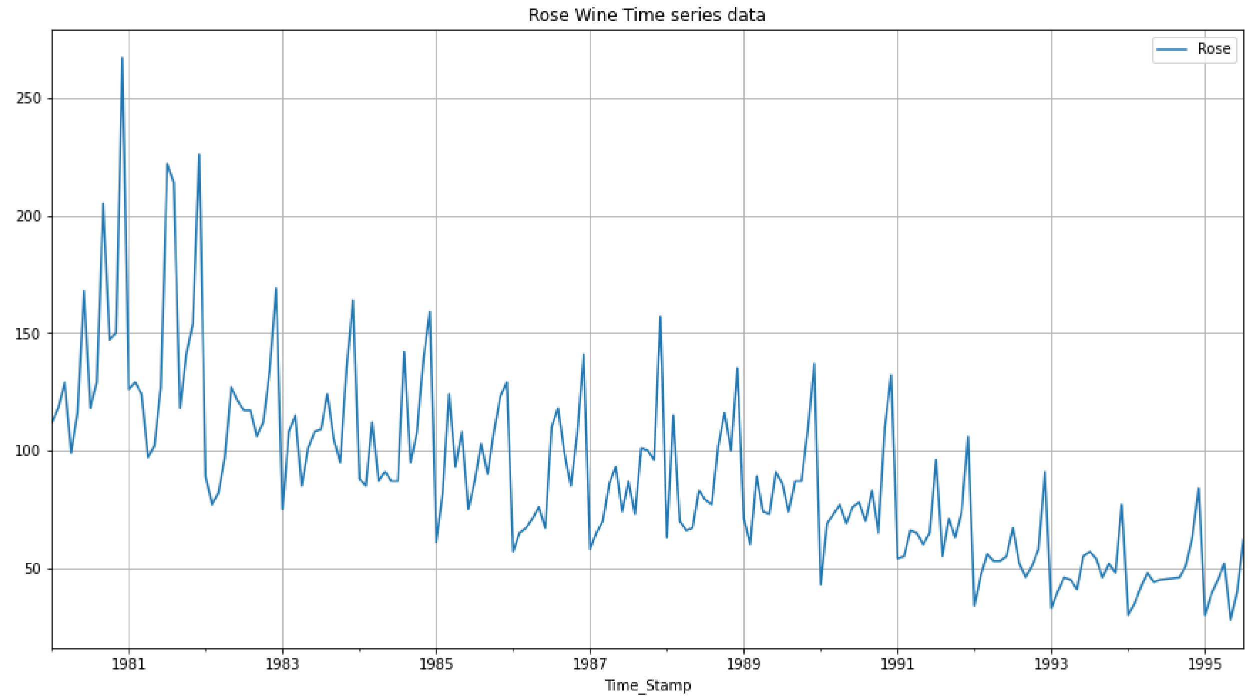
- We have the month wise sales data of two types of wines (Rose and Sparkling) for 15 years.
- We are going to read the time series data and apply different exponential smoothing models, ARIMA and SARIMA models (auto/manual) to forecast the future wine sales.
- We are also going to analyze results from the different models and build a final model to predict the future sales using the model that provides the best results

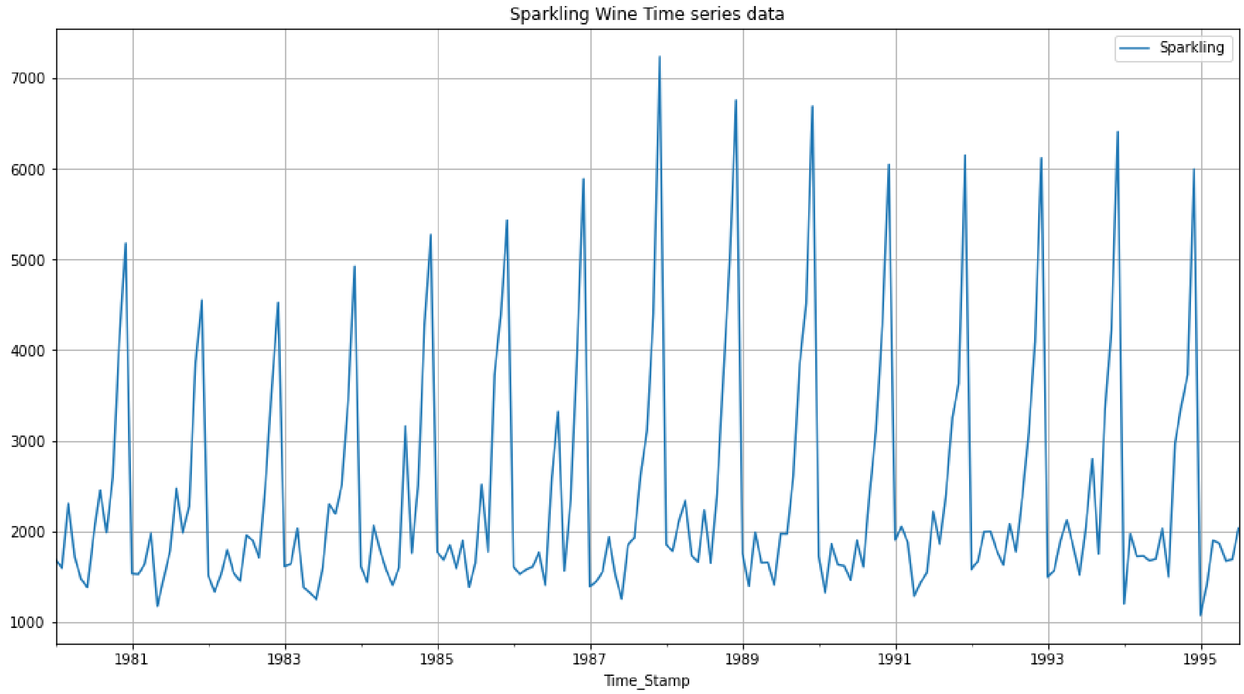
Problem 1. Read the data as an appropriate Time Series data and plot the data.

Solution:

The two datasets with Sales data of two types of wine “Rose” and “Sparkling” are imported into Jupyter notebook using pd.read_csv command.

The sales data of the two data sets are plotted as follows:





<u>Rose Wine</u>

- There is a decreasing trend in the initial years which stabilizes after a few years and again shows a decreasing trend.
- Also Observed seasonality in the data trend and pattern seem to repeat on yearly basis.

<u>Sparkling Wine:</u>

- There is not much trend in the plot.
- The seasonality seems to have a pattern on a yearly basis.

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

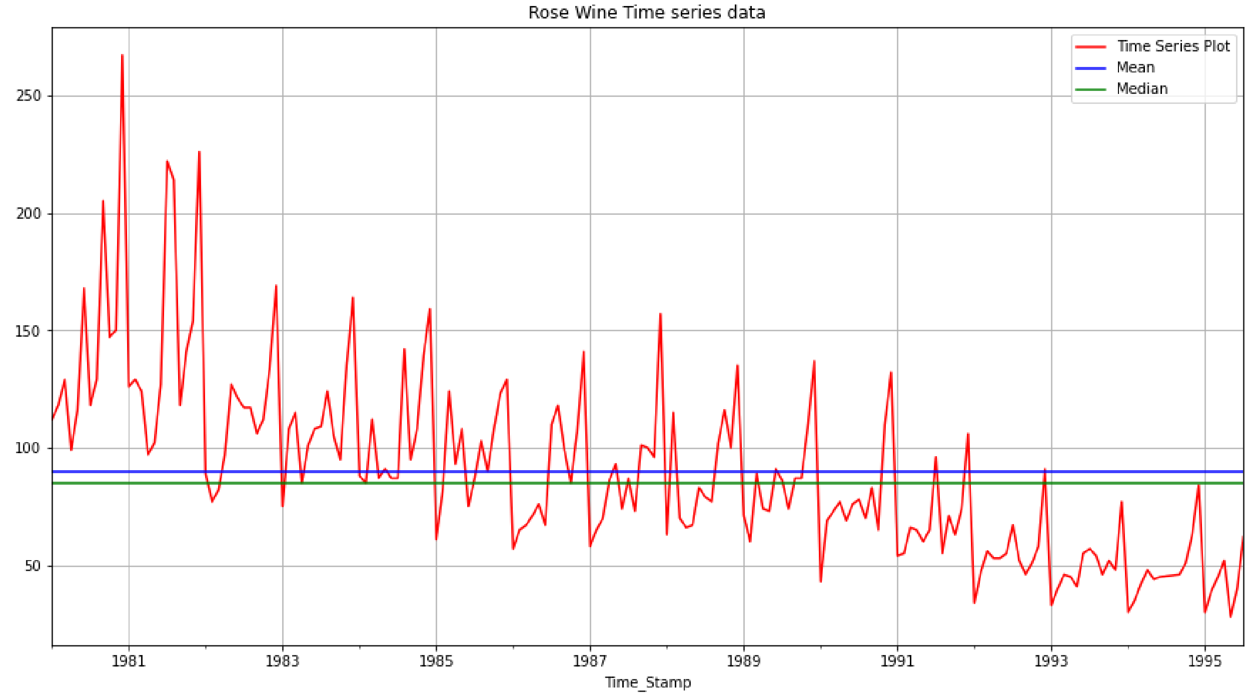
Solution:

We have 187 data points in Rose wine data with two missing values and 187 data points in Sparkling wine data. The basic measures of descriptive statistics tell us how the Sales have varied across years.

Missing Values in **Rose Wine** time series data is interpolated using Linear Interpolation function. Linear Interpolation simply means to estimate a missing value by connecting dots in a straight line in increasing order. In short, It estimates the unknown value in the same increasing order from previous values.

Descriptive Statistics of Rose wine data (without taking time component into account):

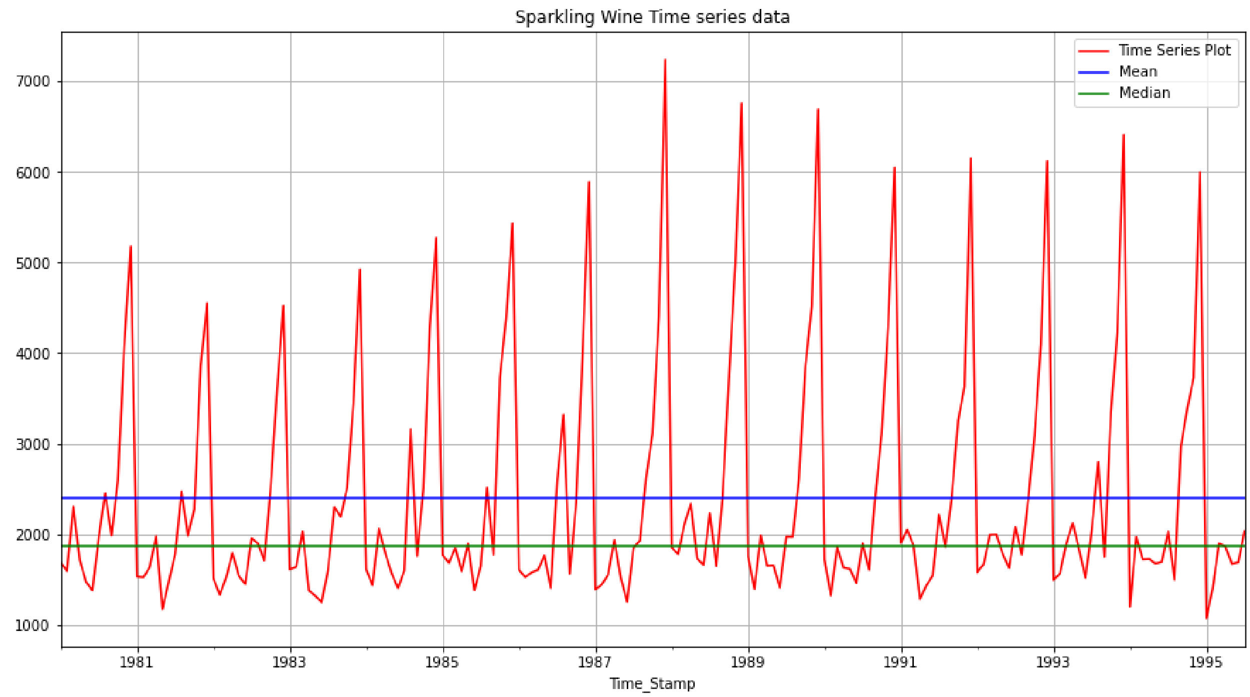
	Rose
count	187.000
mean	89.914
std	39.238
min	28.000
25%	62.500
50%	85.000
75%	111.000
max	267.000



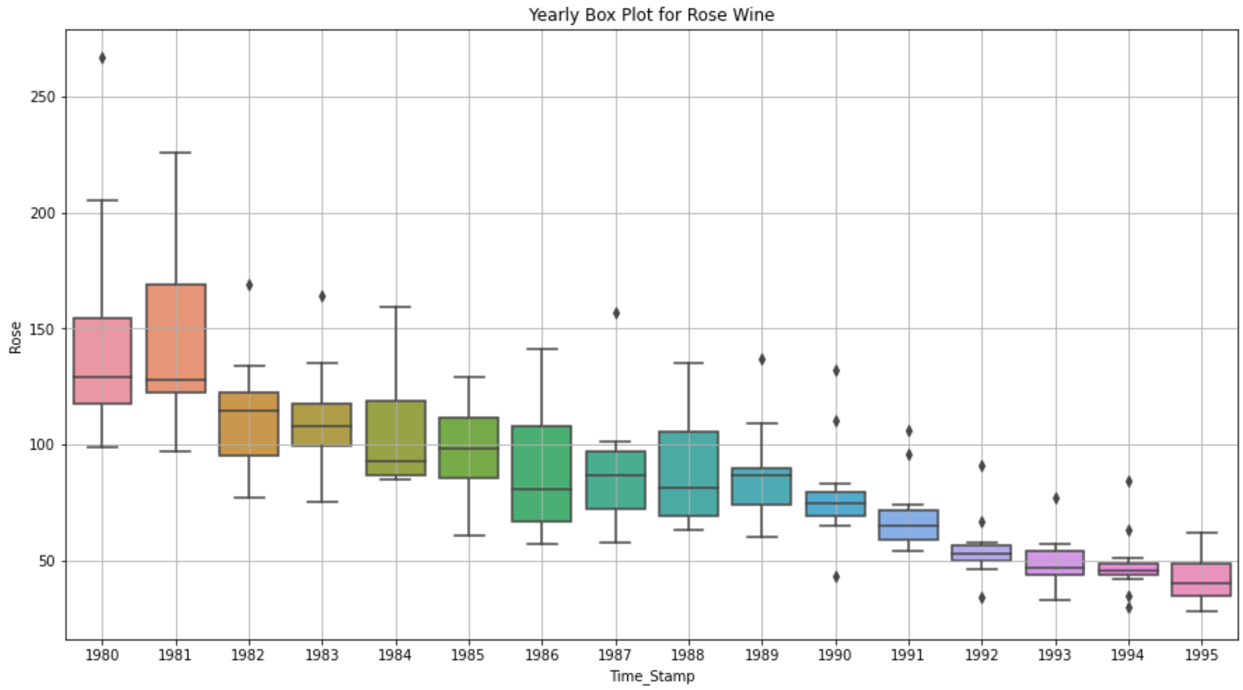
Descriptive Statistics of Sparkling wine data (without taking time component into account):

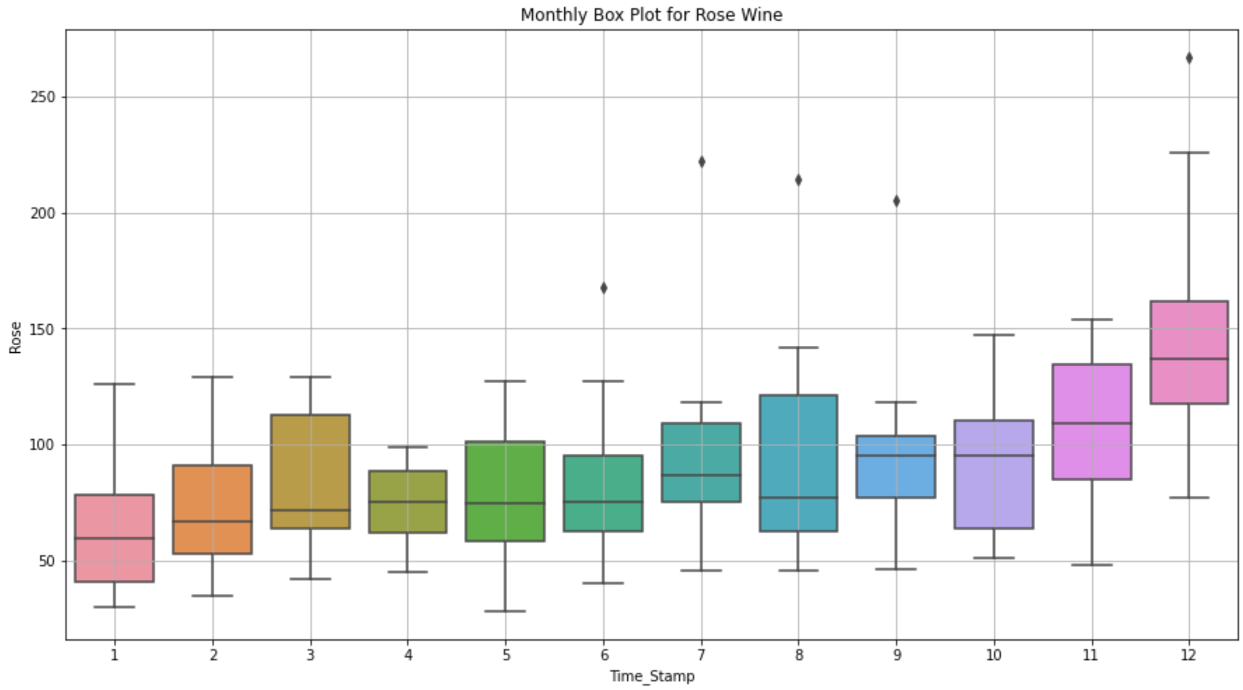
	Sparkling
count	187.000
mean	2402.417
std	1295.112
min	1070.000
25%	1605.000

50%	1874.000
75%	2549.000
max	7242.000



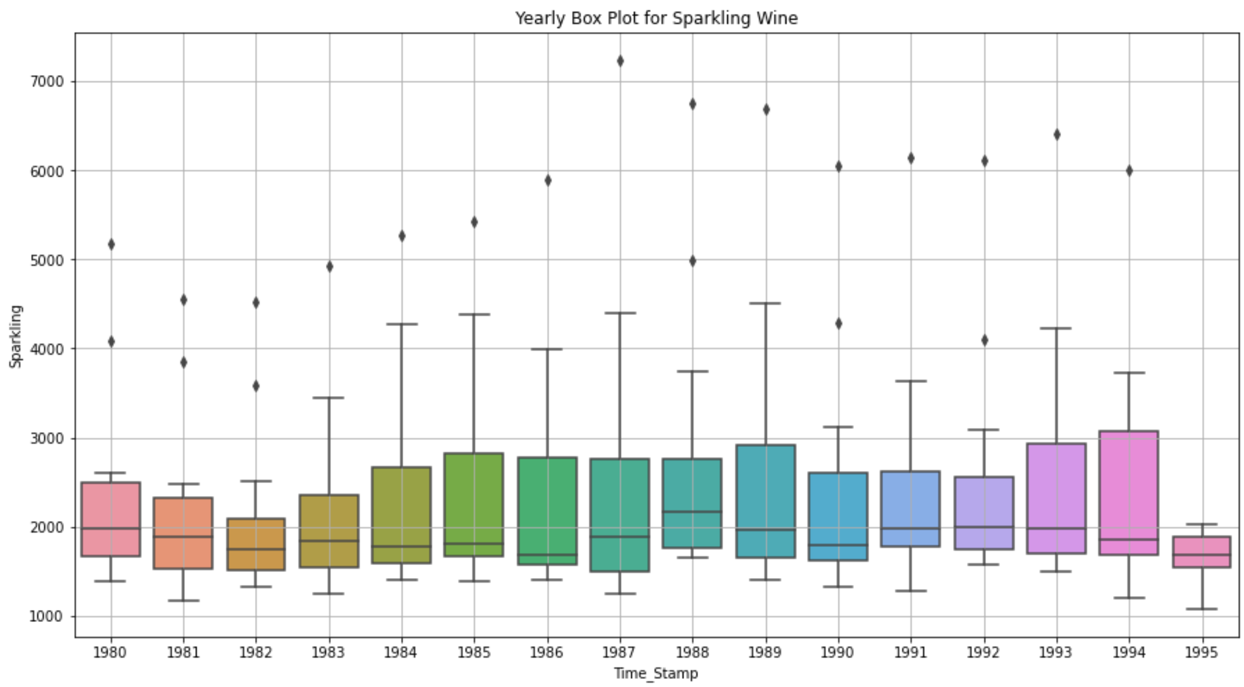
We should look at the box plots year wise and month wise of Rose wine timeseries data:

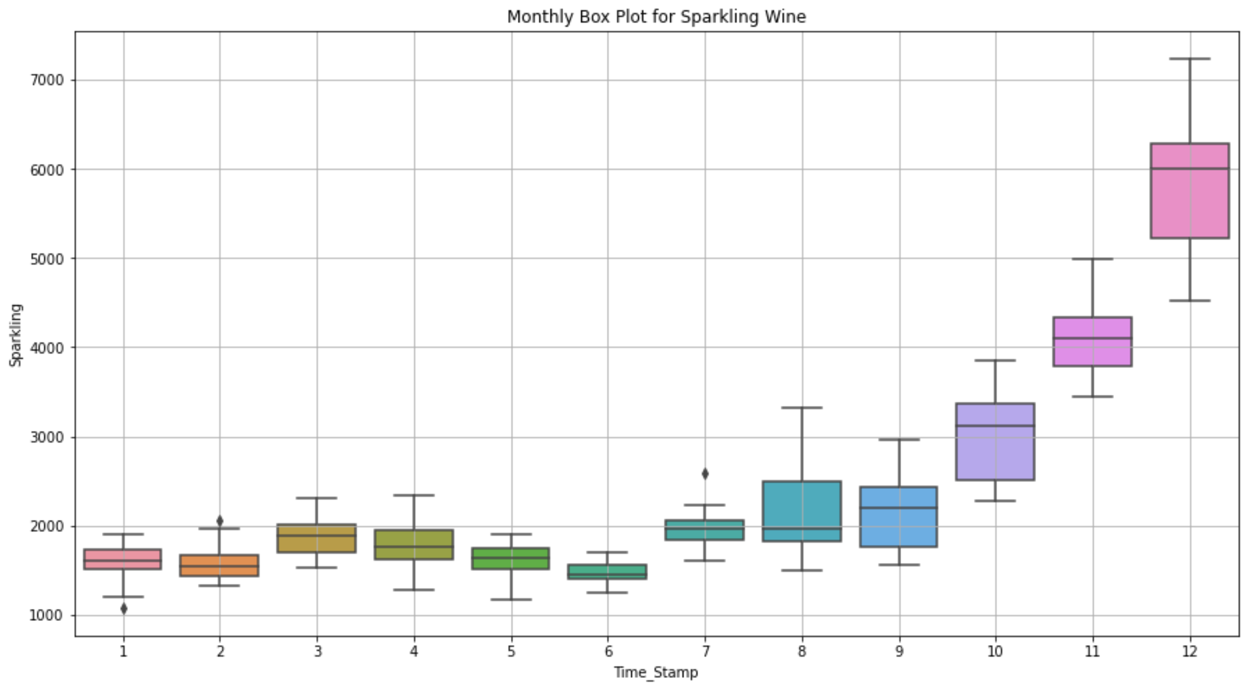




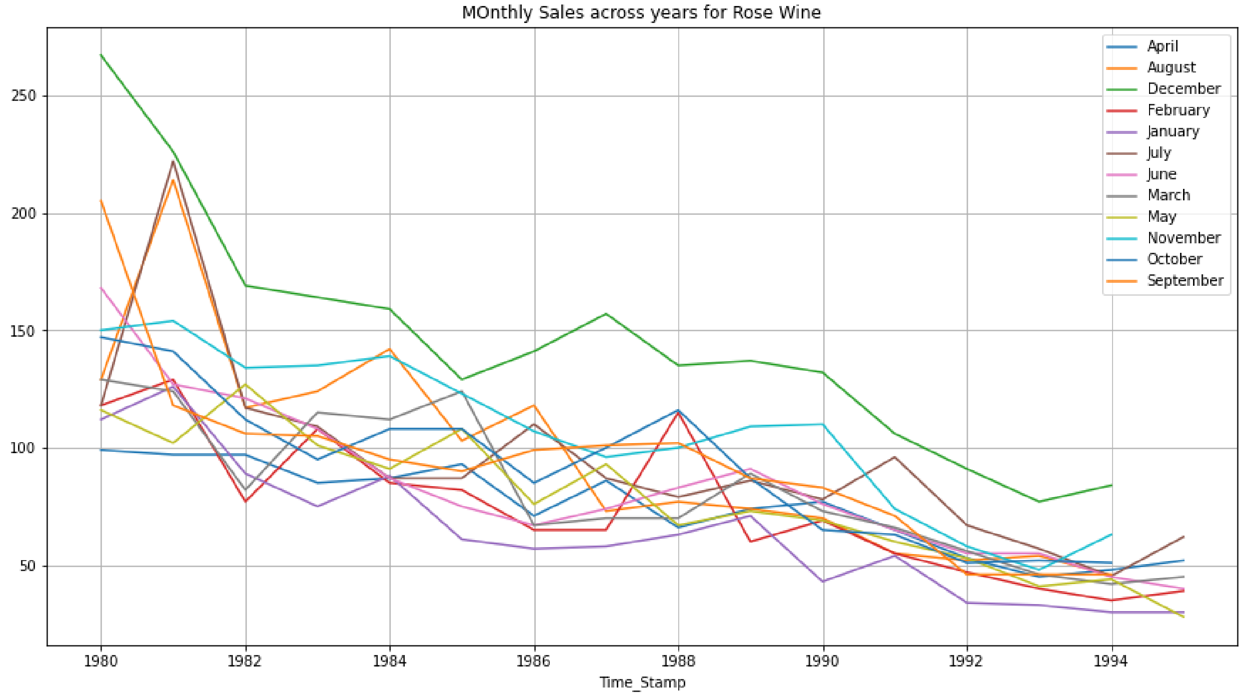
- As observed in the Time Series plot, the year wise boxplots over here also indicate a measure of downward trend.
- Also, we see that the sales of Rose wine have some outliers for certain years.
- December seems to have the highest sales of Rose wine and there are also outlier in June, July, August and September months.

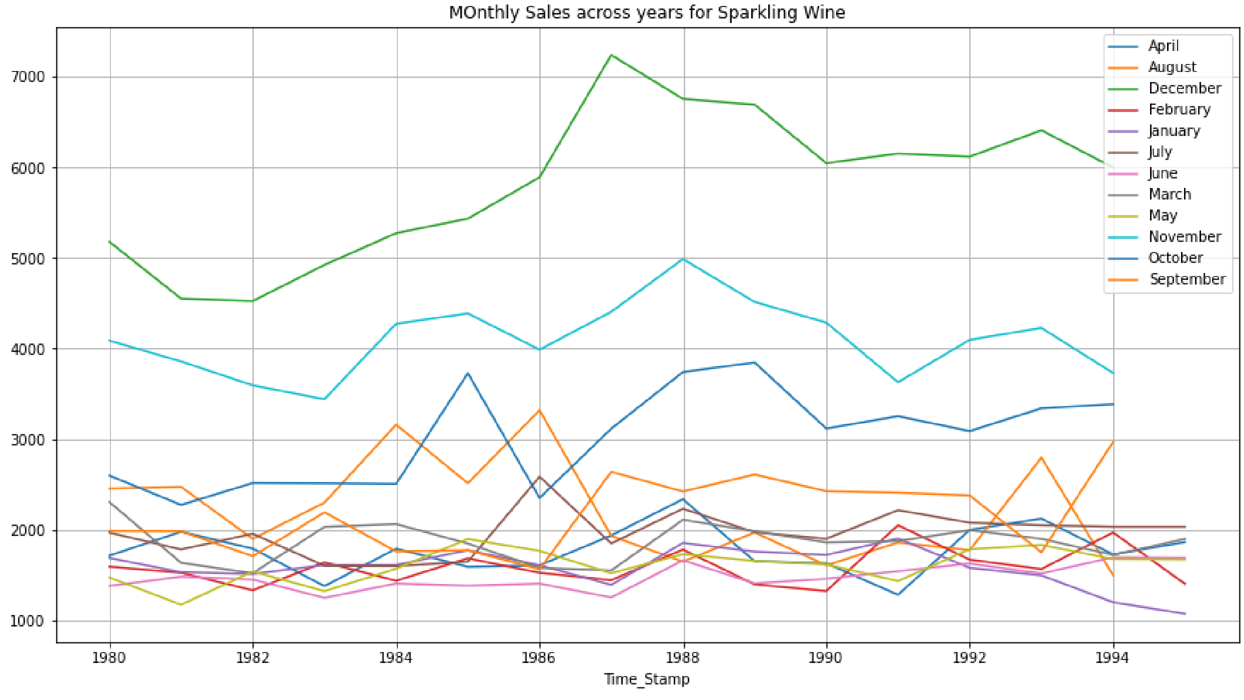
We should look at the box plots year wise and month wise of Sparkling wine timeseries data:



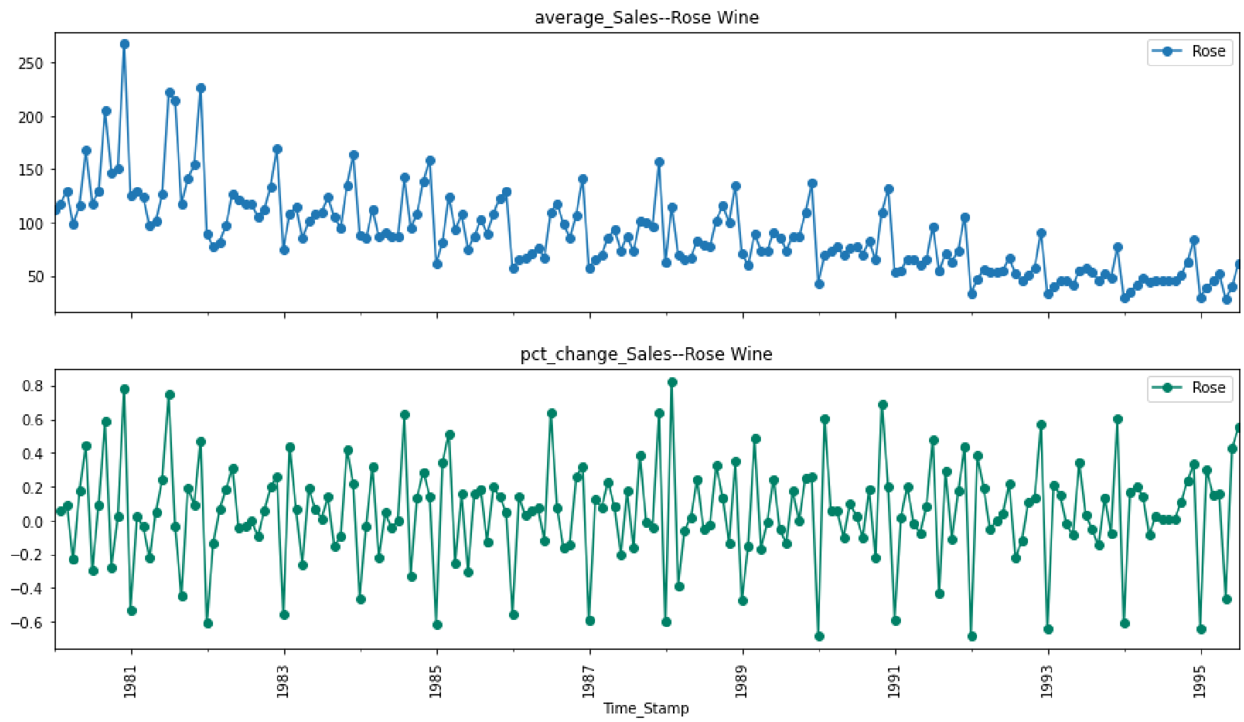


- As observed in the Time Series plot, the boxplots over here also do not indicate trend.
- Also, we see that the sales of Sparkling wine has some outliers for almost all years except 1995.
- We also observe December has the highest sales value for Sparkling wine.


MOnthly Sales across years for Rose Wine
April
August
December
February
January
July
June
March
May
November
October
September
250
200
150
100
50
1980
1982
1984
1986
1988
1990
1992
1994
Time_Stamp


MOnthly Sales across years for Sparkling Wine
April
August
December
February
January
July
June
March
May
November
October
September
7000
6000
5000
4000
3000
2000
1000
1980
1982
1984
1986
1988
1990
1992
1994
Time_Stamp

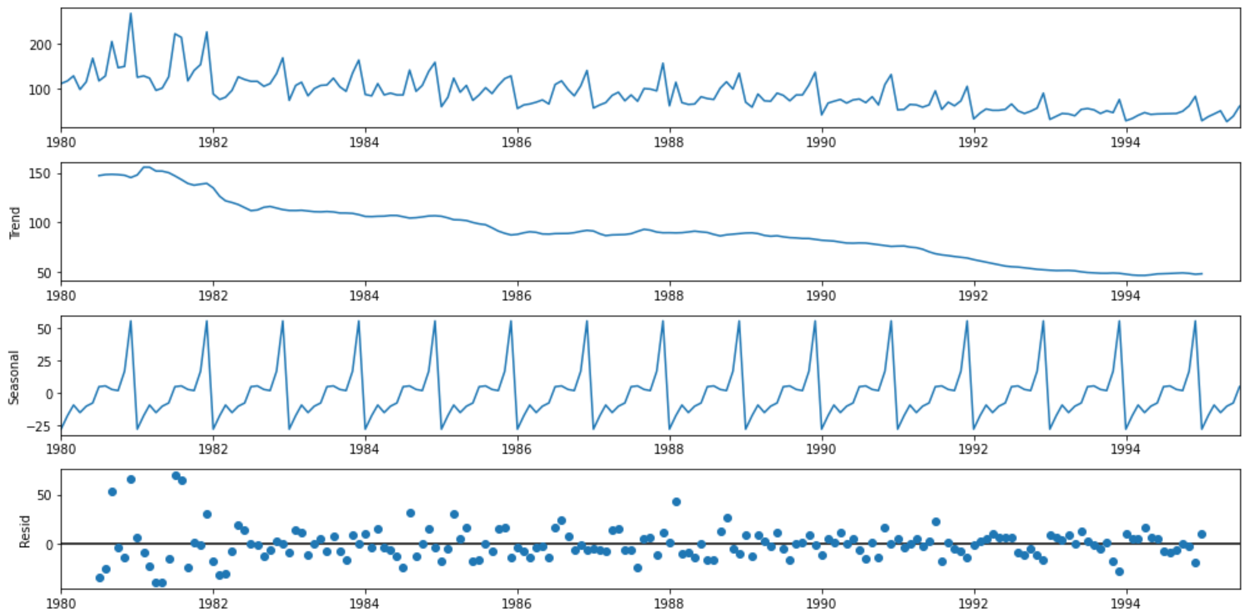
- The above Line plots of Year/month wise sales data of Rose and Sparkling wine that December month has the highest Sales and January, February and March months show lower sales values.



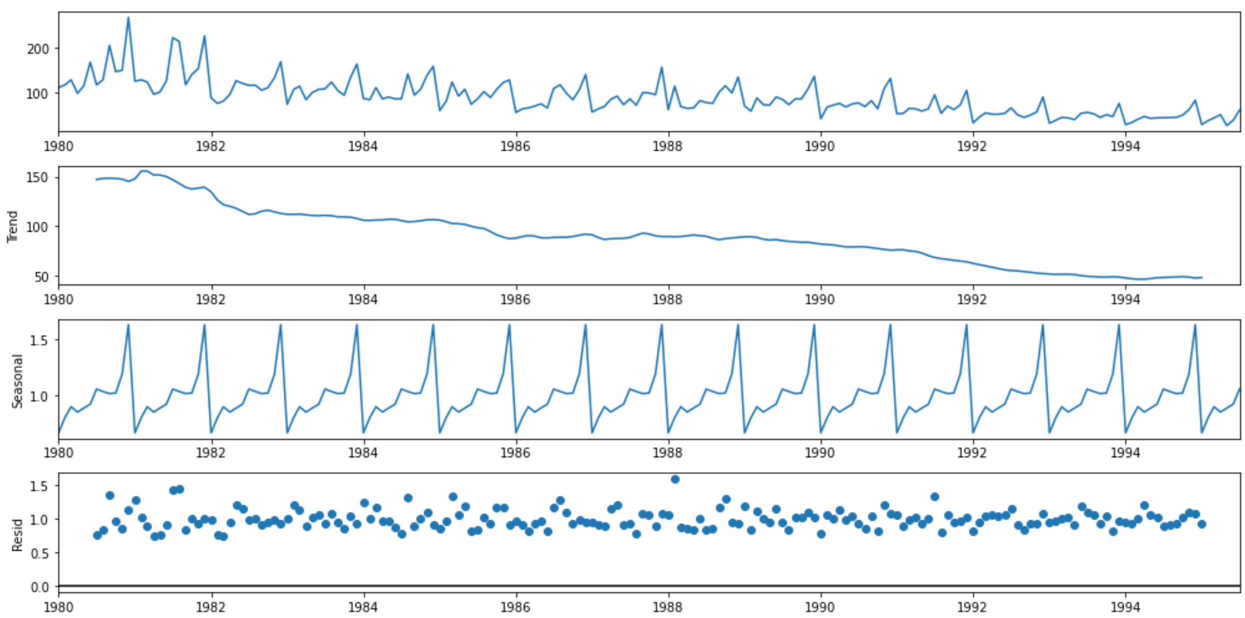
- The Average Sales value in Rose wine time series also shows a decreasing trend whereas Percentage change in Sales does not show any trend.

- The Average Sales value in Sparkling wine time series does not show any trend whereas Percentage change in Sales also does not show any trend.

Rose Wine time series addition decomposition:

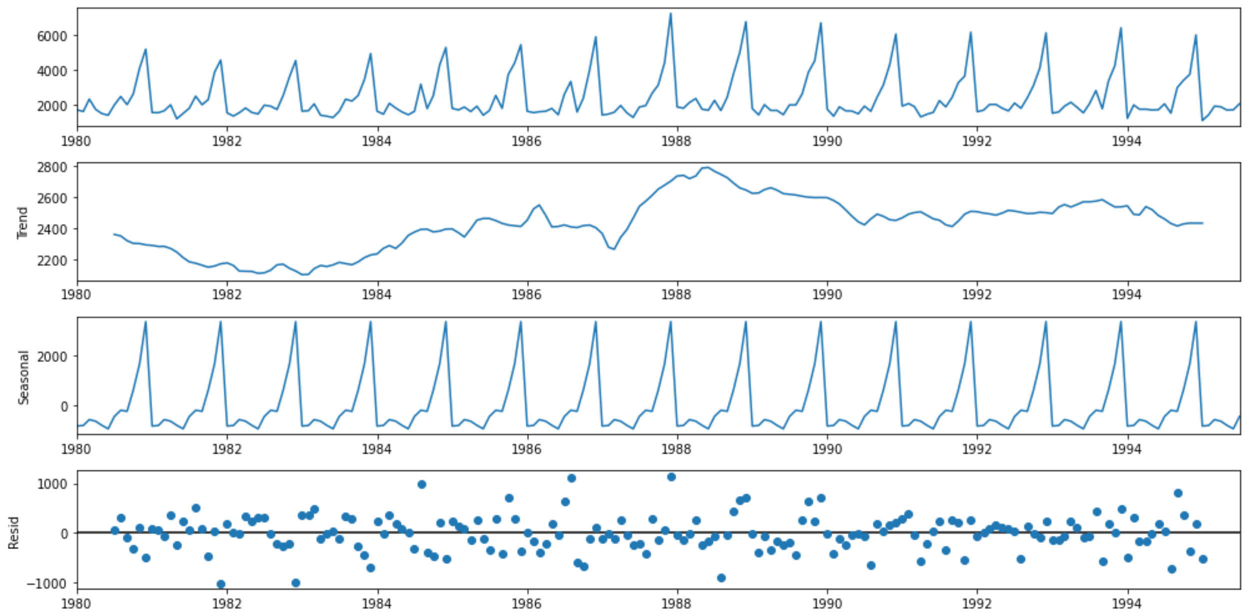


Rose Wine time series multiplicative decomposition:

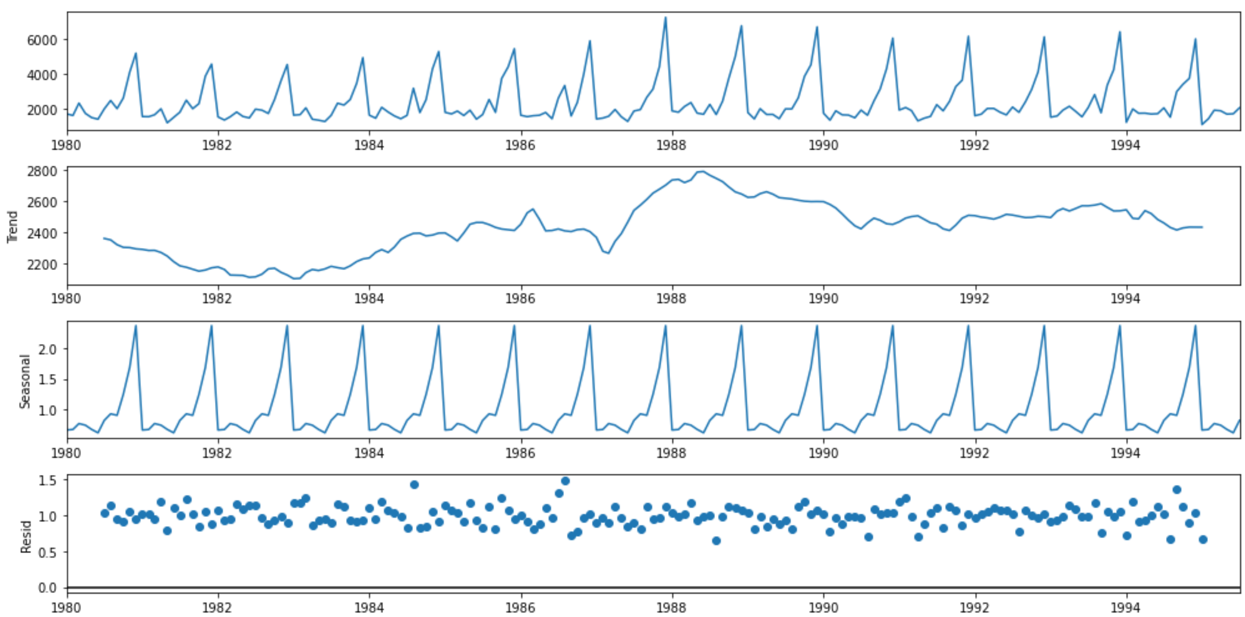


- For additive we see the residual values are around 0 and for Multiplicative model we see the residuals are around 1.

Sparkling Wine time series addition decomposition:



Sparkling Wine time series multiplicative decomposition:

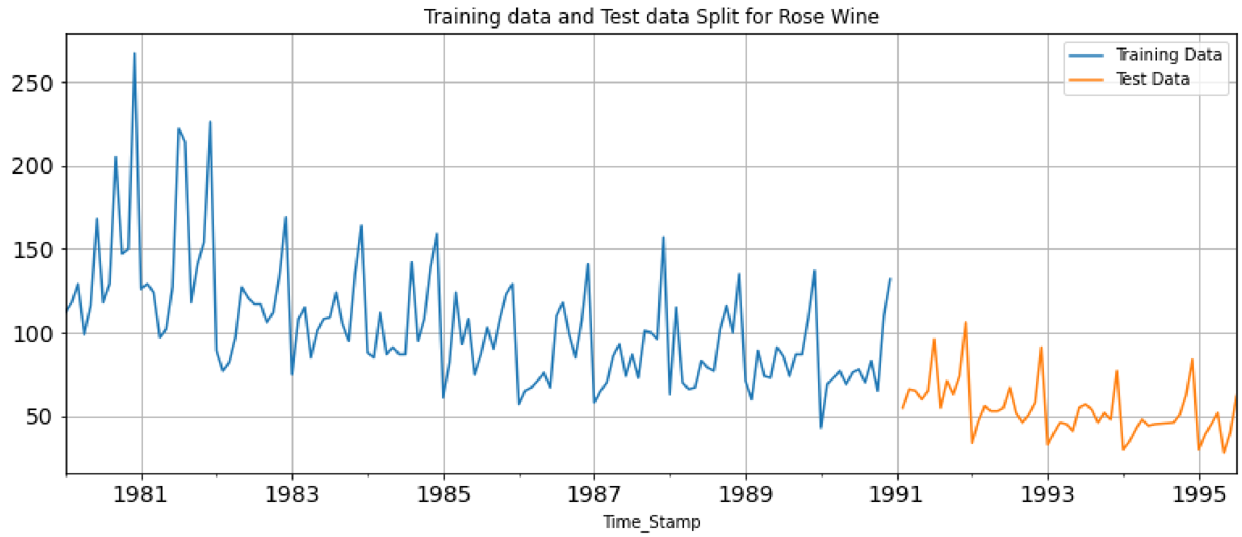


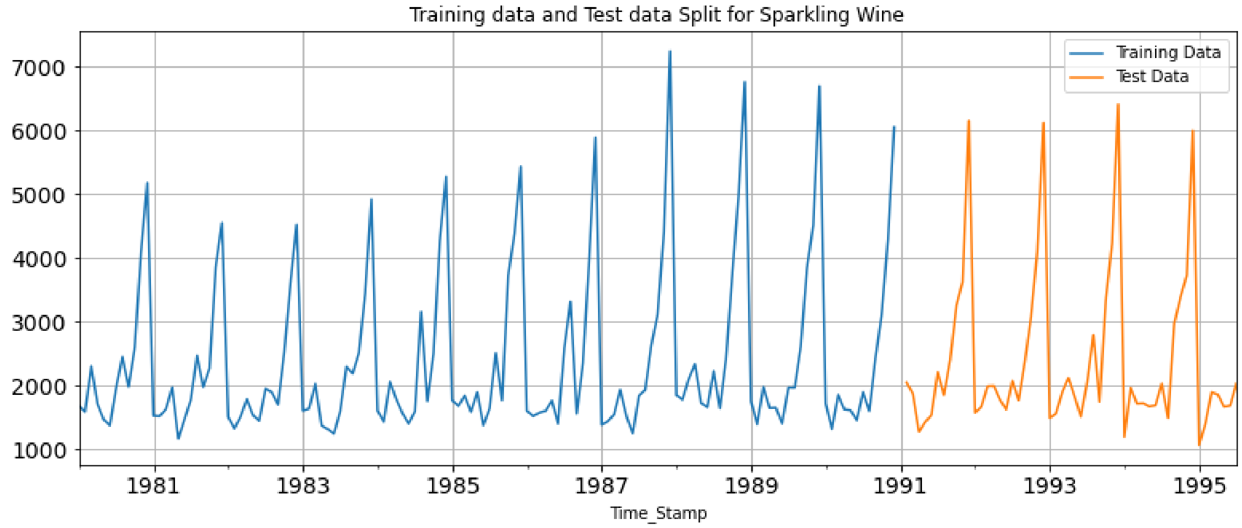
- For additive we see the residual values are around 0 and for Multiplicative model we see the residuals are around 1.

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

Solution:

- The **Train data of Rose wine** sales has been split for data up to 1990 and has 132 data points
- The **Train data of Sparkling wine** sales has been split for data up to 1990 and has 132 data points.
- The **Test data of Rose wine** sales has been split for data from 1991 and has 55 data points.
- The **Test data of Sparkling wine** sales has been split for data from 1991 and has 55 data points.

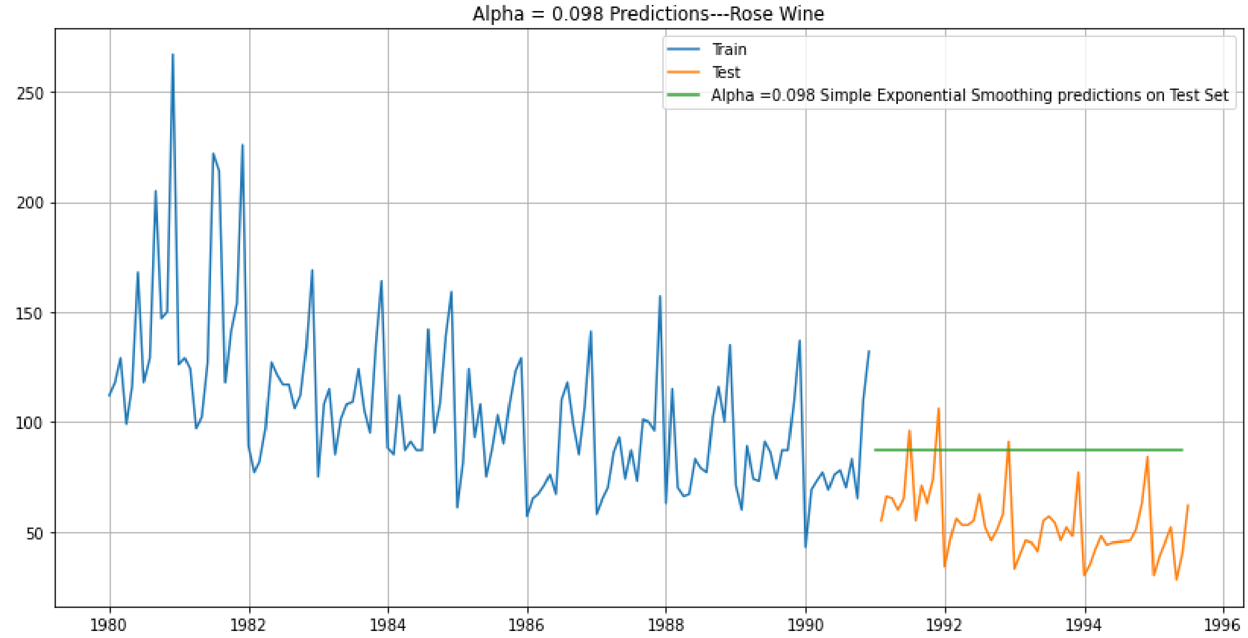




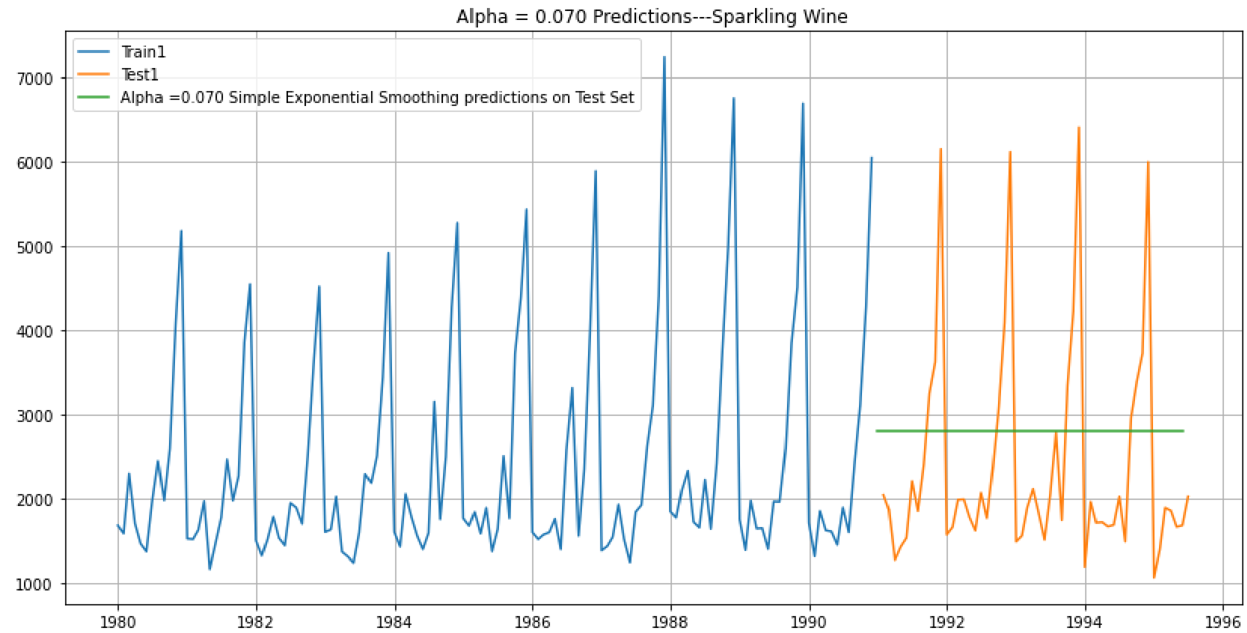
Problem 4. Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression, naïve forecast models, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE. Please do try to build as many models as possible and as many iterations of models as possible with different parameters.

Solution:

1) **Simple Exponential Smoothing :**

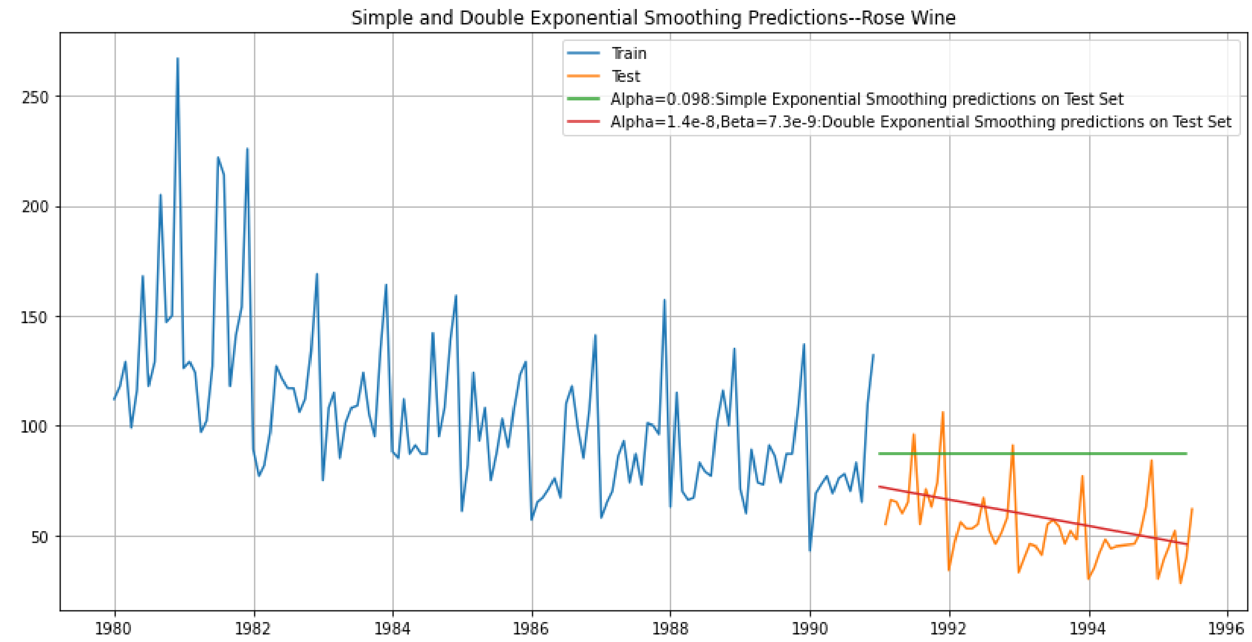


- While trying to forecast Model Evaluation for @ α = 0.098 : Simple Exponential Smoothing, We get an RMSE score of 36.80 for the Test data.

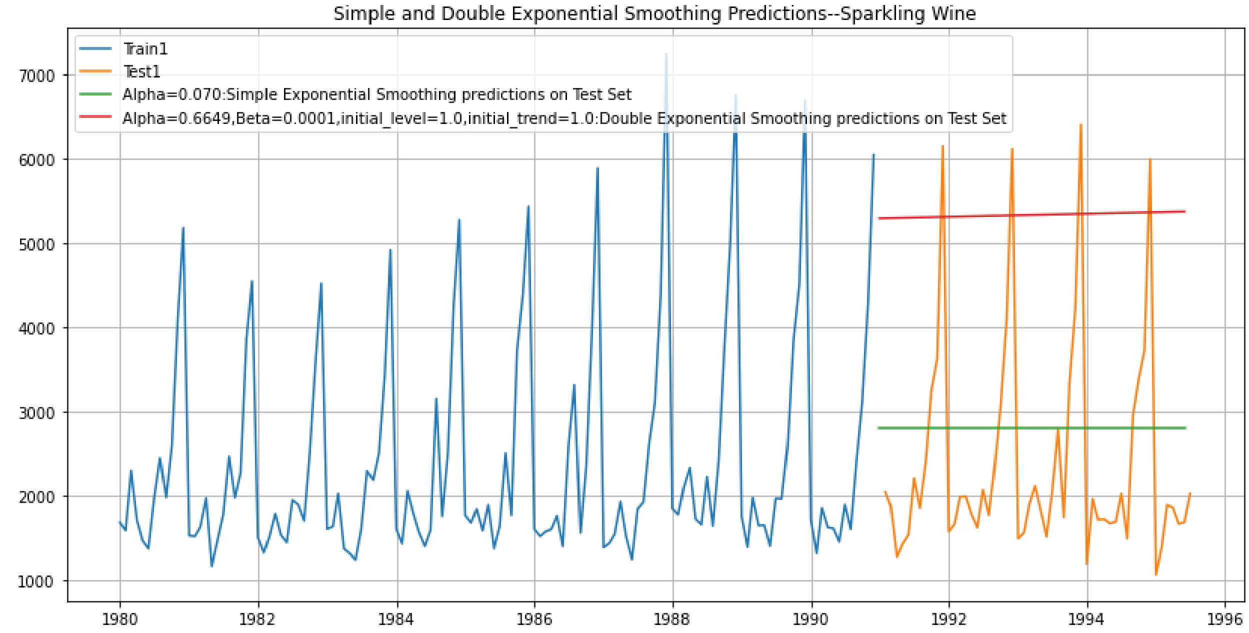


- While trying to forecast Model Evaluation for @ α = 0.070 : Simple Exponential Smoothing, We get an RMSE score of 1344 for the Test data.

2) Double Exponential Smoothing:



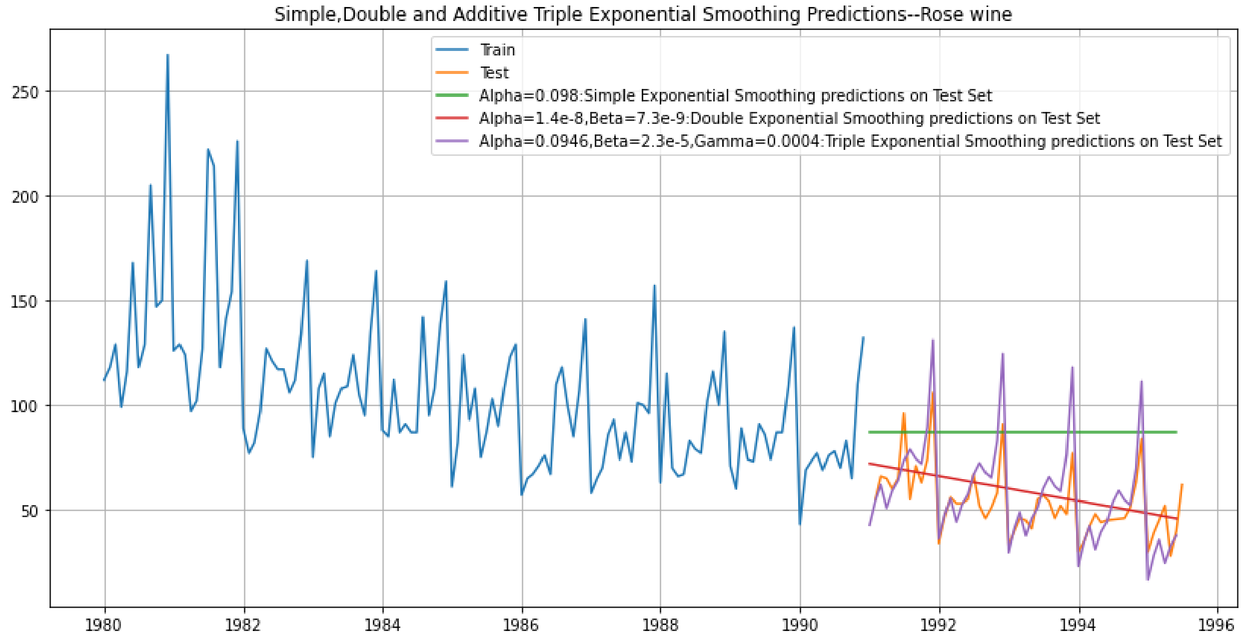
RMSE: 15.36



RMSE: 3193

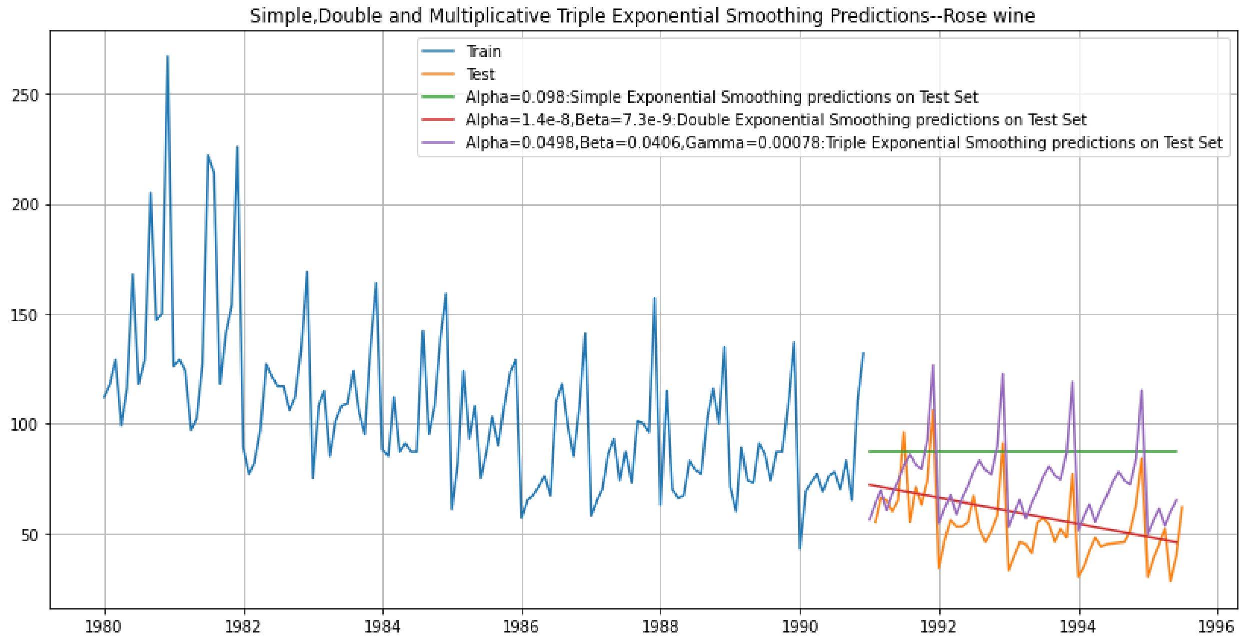
3) Triple Exponential Smoothing:

Additive trend and seasonality type: Rose wine



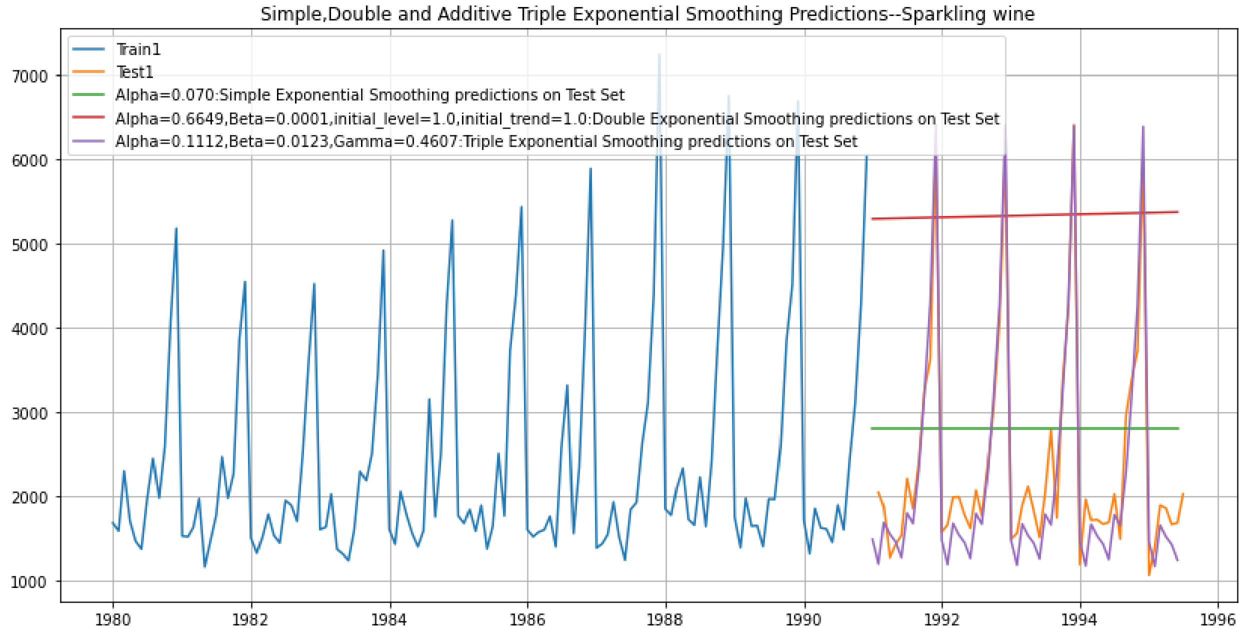
RMSE: 27.11

Multiplicative trend and seasonality type: Rose wine



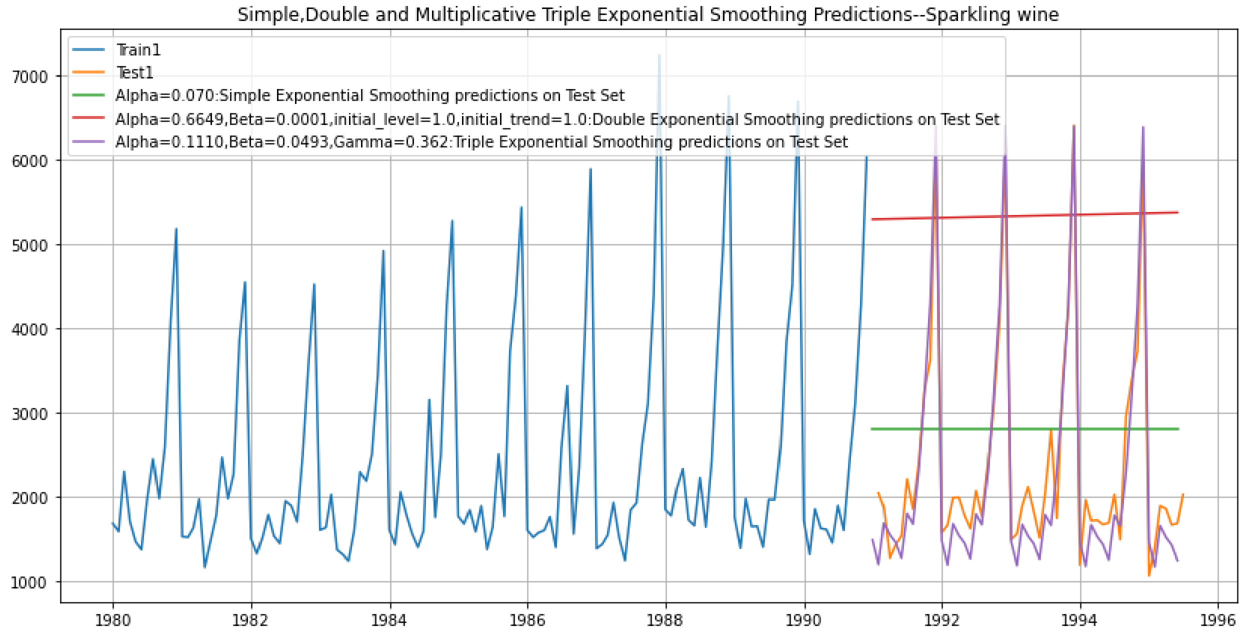
RMSE: 29.53

Additive trend and seasonality type: Sparkling wine



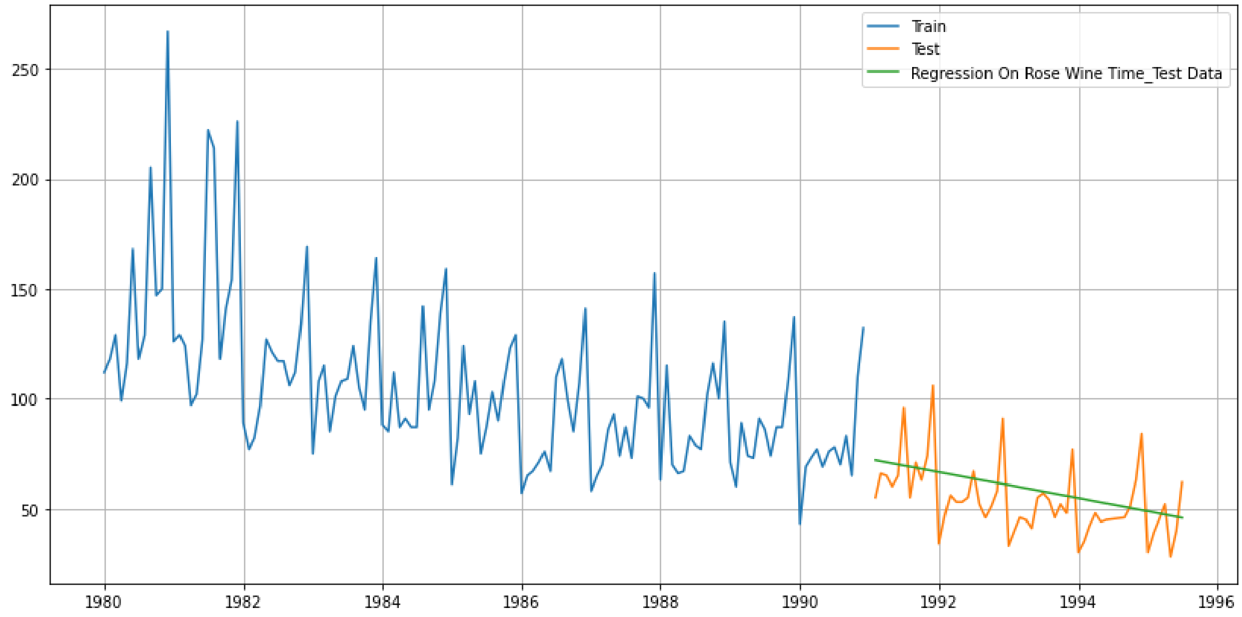
RMSE: 1575

Multiplicative trend and seasonality type: Sparkling wine

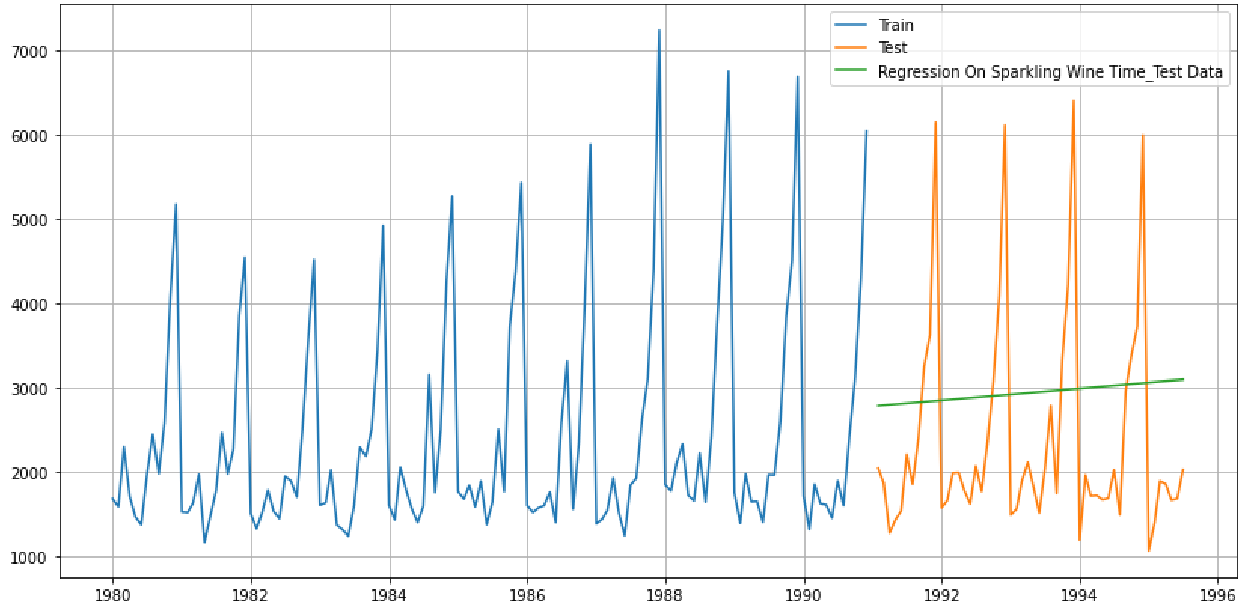


RMSE: 1448

4) Linear Regression Model

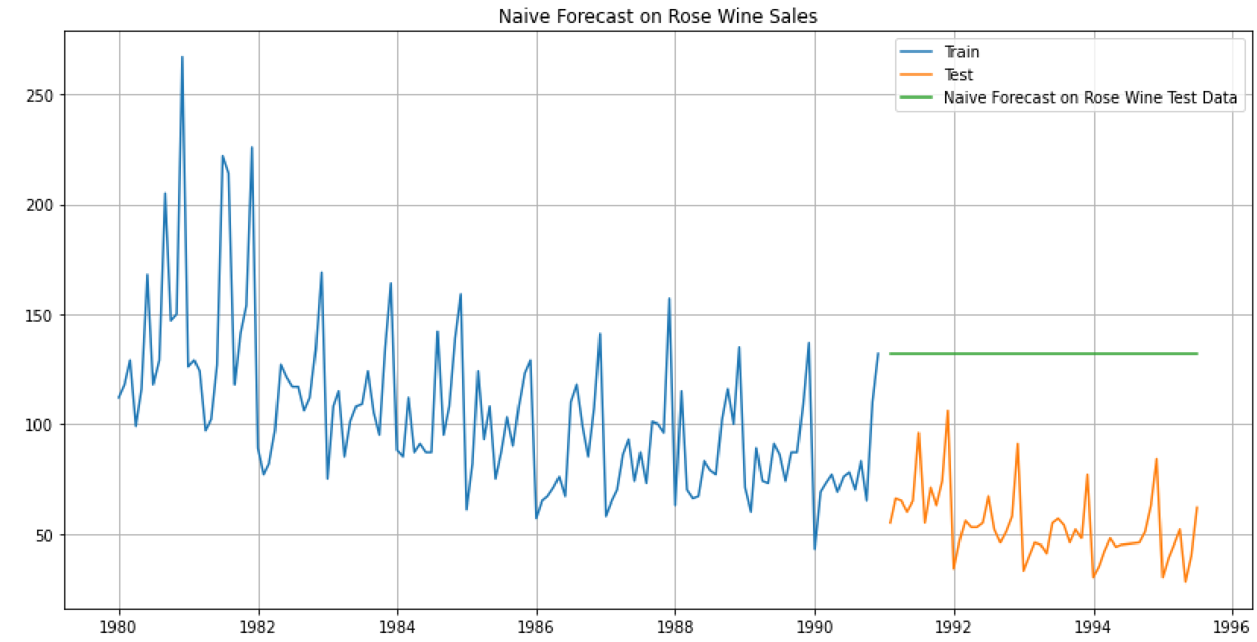


RMSE: 15.36

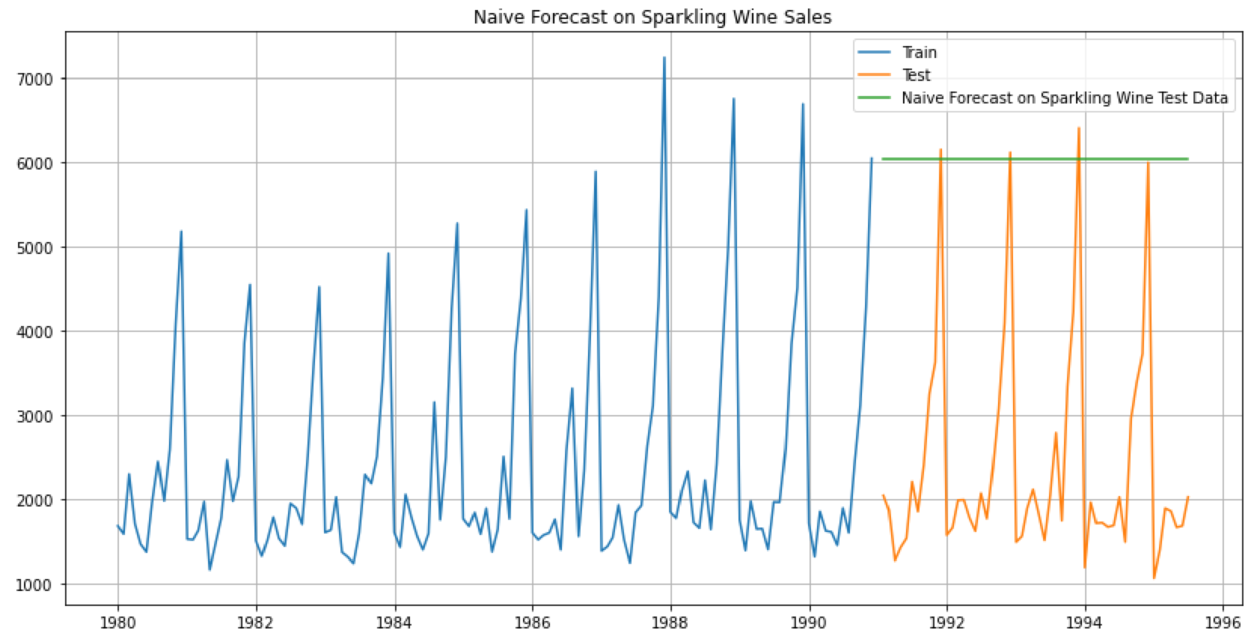


RMSE: 1394

5) Naive Approach



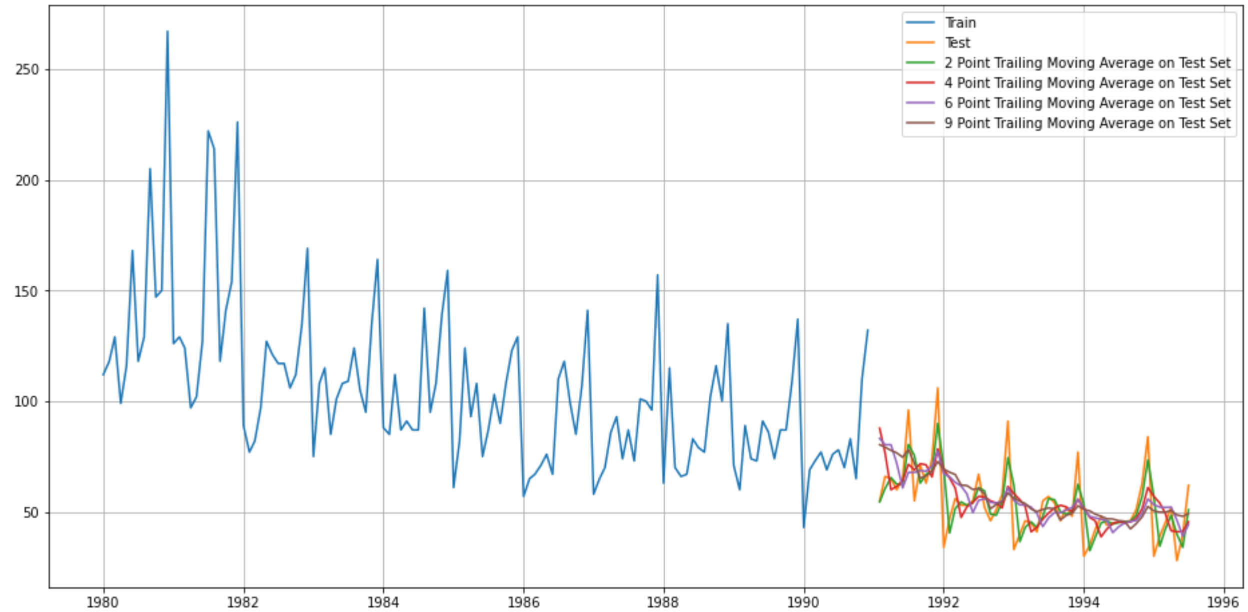
RMSE: 79.7



RMSE: 3858

6) Moving Average Model

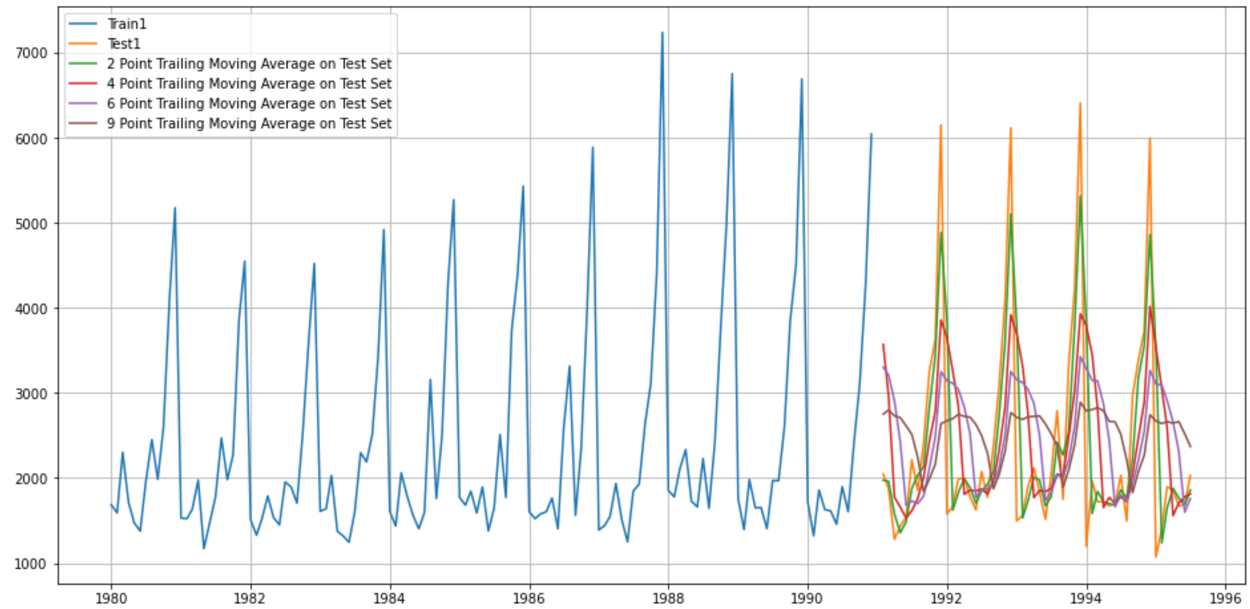
Rose Wine:



RMSE:

2pointTrailingMovingAverage	10.354667
4pointTrailingMovingAverage	13.725015
6pointTrailingMovingAverage	14.054791
9pointTrailingMovingAverage	14.370674

Sparkling Wine:



RMSE:

2pointTrailingMovingAverage	770.928742
4pointTrailingMovingAverage	1137.137053
6pointTrailingMovingAverage	1283.096993
9pointTrailingMovingAverage	1354.277938

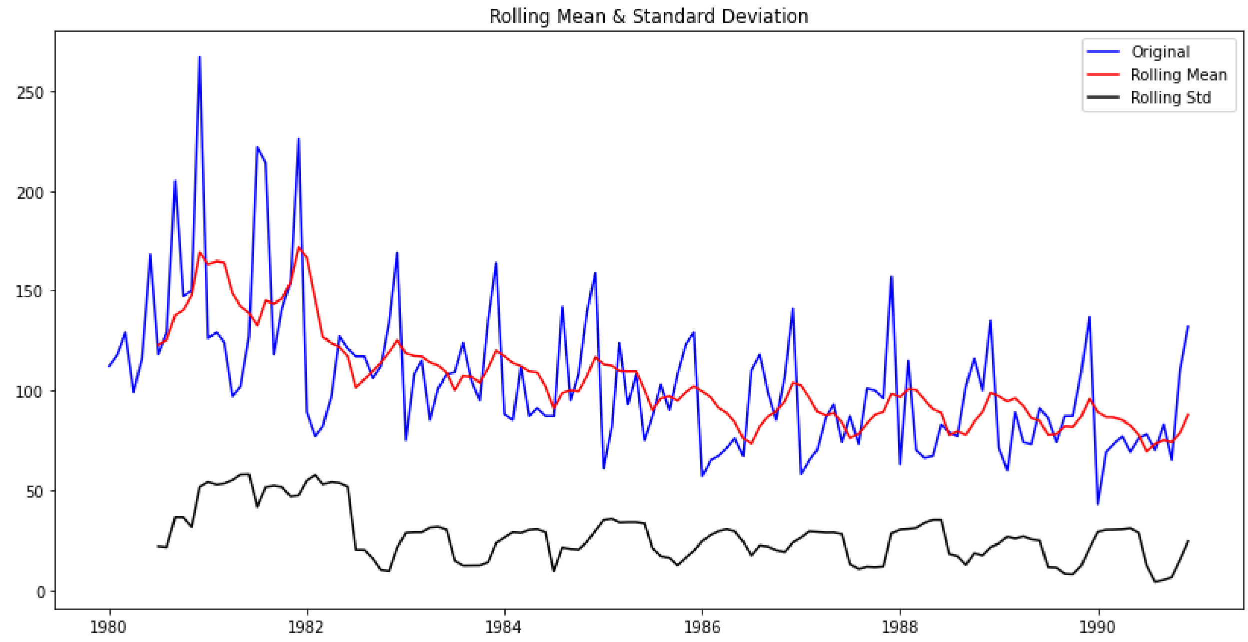
Problem 5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity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

Solution

We test the training data for stationarity using the Augmented Dickey-Fuller (ADF) test at α= 0.05.

If the data is non-stationary, use appropriate measures to stationarize the data and then check for stationarity using the Augmented Dickey-Fuller (ADF) Test at α = 0.05.

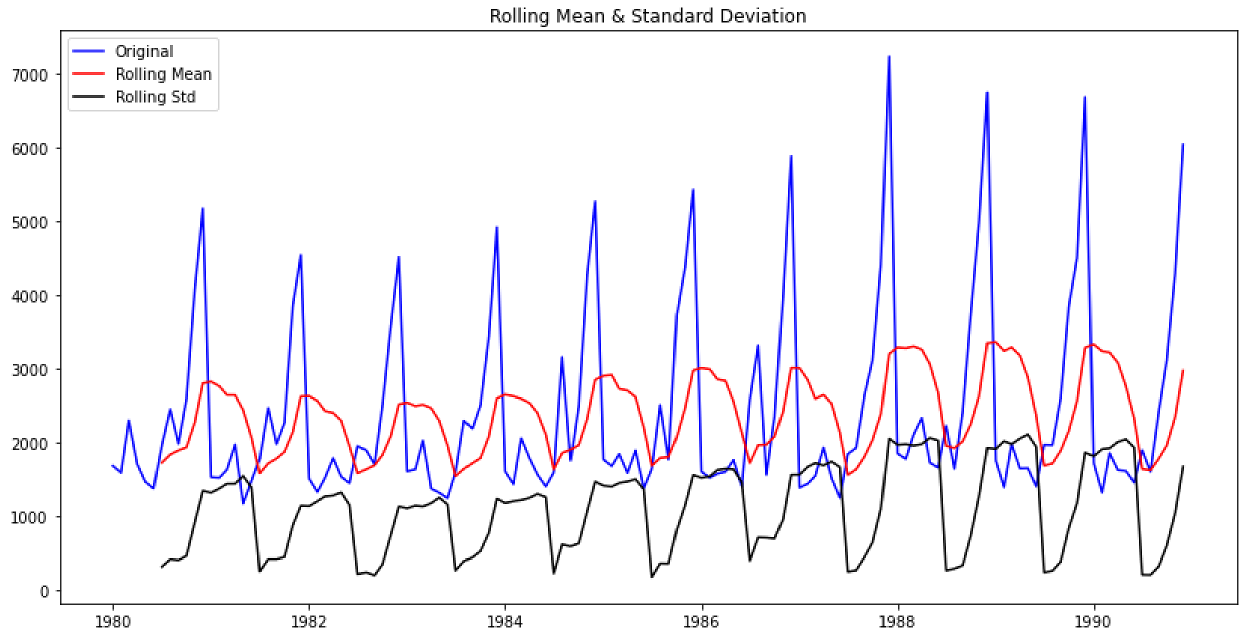
Rose Wine time series:



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

Sparkling Wine Time series:

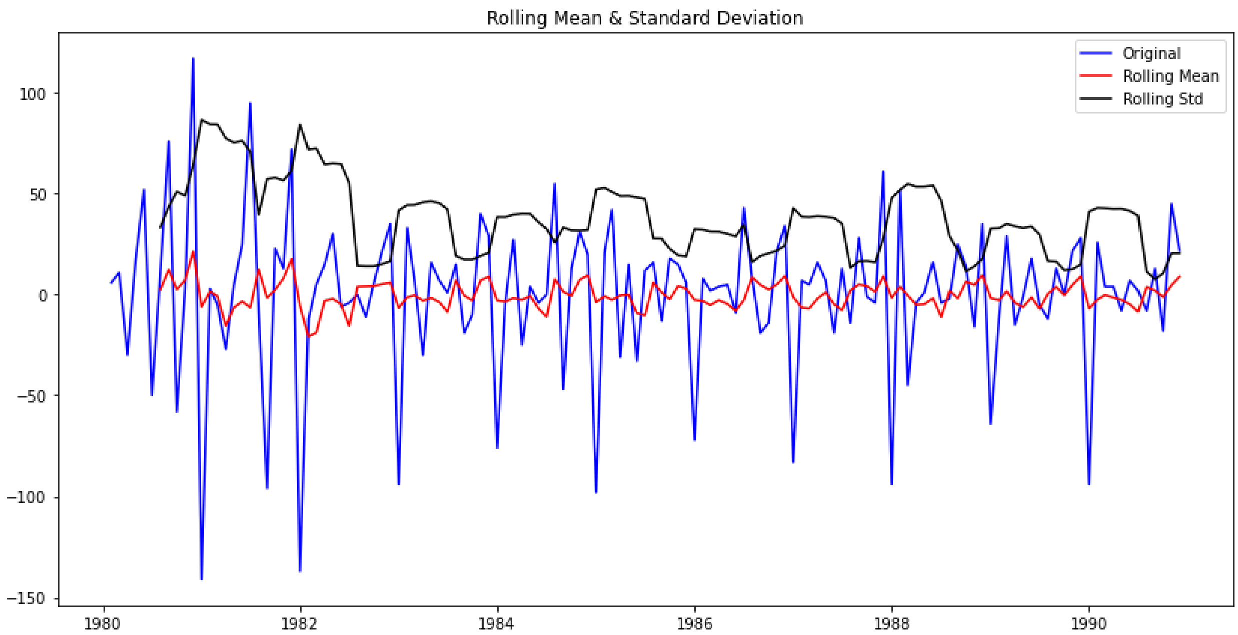


Results of Dickey-Fuller Test:

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

- P-value for both Rose and Sparkling wine time series data is high and more than 0.05. Thus, we fail to reject the null hypothesis. That means both time series data are non-stationary.
- Differencing by level (1) to stationarize both time series data Rose and Sparkling wine and again checking for stationarity using Dickey-Fuller Test:

Rose Wine Time Series:



Results of Dickey-Fuller Test:

Test Statistic -6.592372e+00

p-value 7.061944e-09

#Lags Used 1.200000e+01

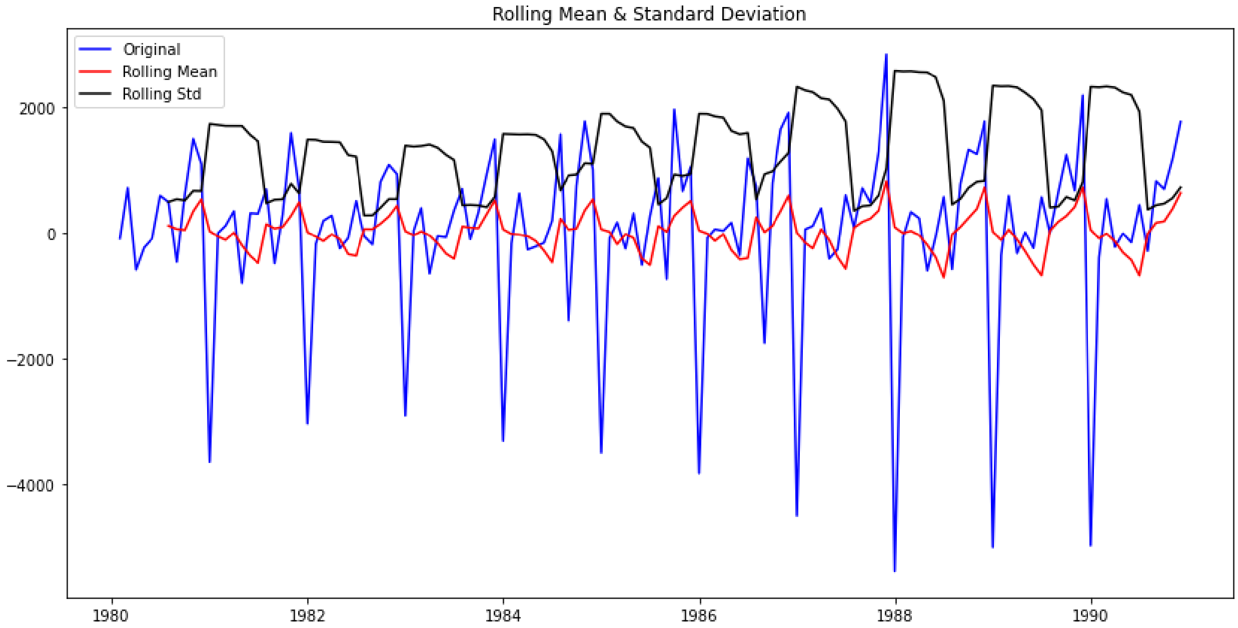
Number of Observations Used 1.180000e+02

Critical Value (1%) -3.487022e+00

Critical Value (5%) -2.886363e+00

Critical Value (10%) -2.580009e+00

Sparkling Wine Time Series:



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

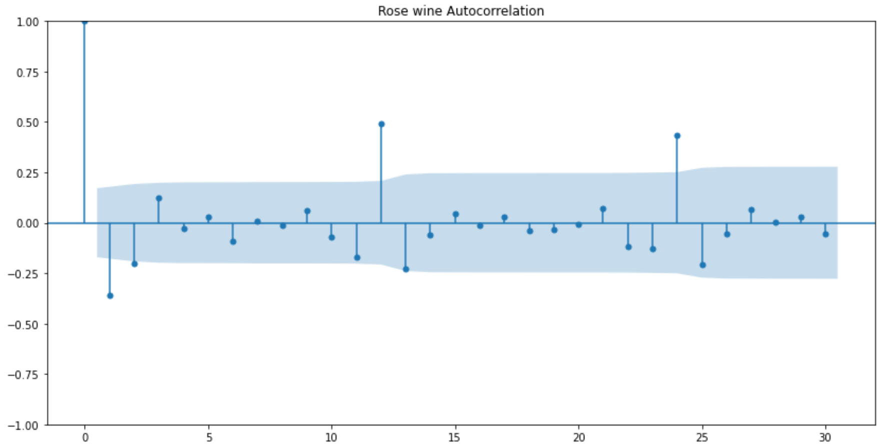
- **After differencing both time series data by level (1), p-value for stationarity hypothesis test is less than 0.05. Thus both time series data are now stationary data.**

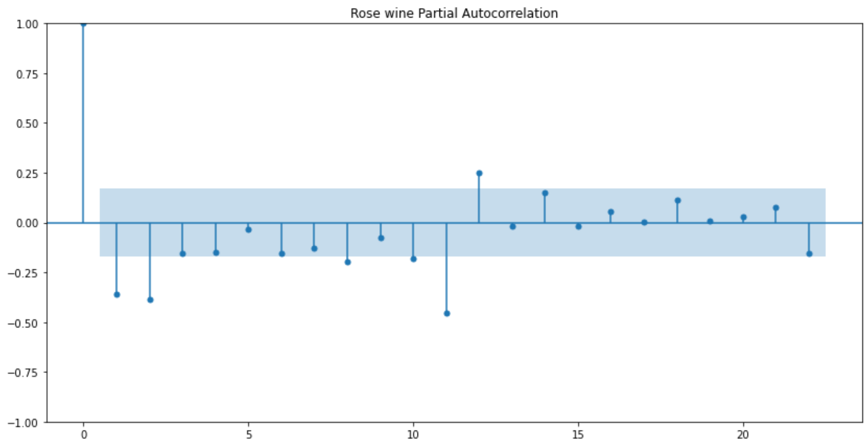
Problem 6. Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.

Solution

Rose Wine Time series:

We checked the Stationarity of the train data of Rose sales data at alpha 0.05 and observe from the result table that data was not stationary hence we took a difference of order 1 and found the train data was stationary then.





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

	param	AIC
11	(2, 1, 3)	1274.695121
15	(3, 1, 3)	1278.659536
2	(0, 1, 2)	1279.671529
6	(1, 1, 2)	1279.870723
3	(0, 1, 3)	1280.545376

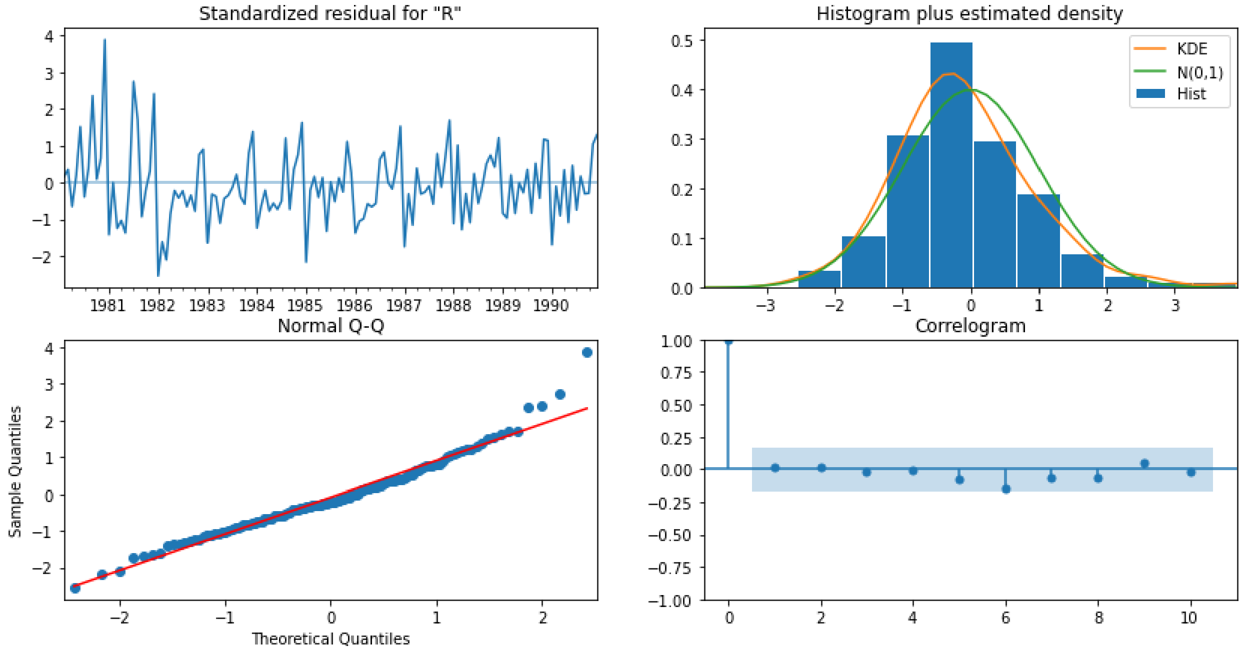
Lowest AIC: 1274.695 for (2,1,3)

```
auto_ARIMA = ARIMA(train, order=(2,1,3))

results_auto_ARIMA = auto_ARIMA.fit()

print(results_auto_ARIMA.summary())
```

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(2, 1, 3)   Log Likelihood                -631.348
Date:                Sun, 20 Mar 2022   AIC                           1274.695
Time:                        15:03:38   BIC                           1291.946
Sample:                    01-01-1980   HQIC                          1281.705
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.6777      0.084    -20.037      0.000      -1.842      -1.514
ar.L2         -0.7285      0.084     -8.701      0.000      -0.893      -0.564
ma.L1          1.0445      0.650      1.606      0.108      -0.230       2.319
ma.L2         -0.7724      0.134     -5.772      0.000      -1.035      -0.510
ma.L3         -0.9049      0.590     -1.533      0.125      -2.062       0.252
sigma2       859.0300    547.623      1.569      0.117    -214.292    1932.351
==============================================================================
```

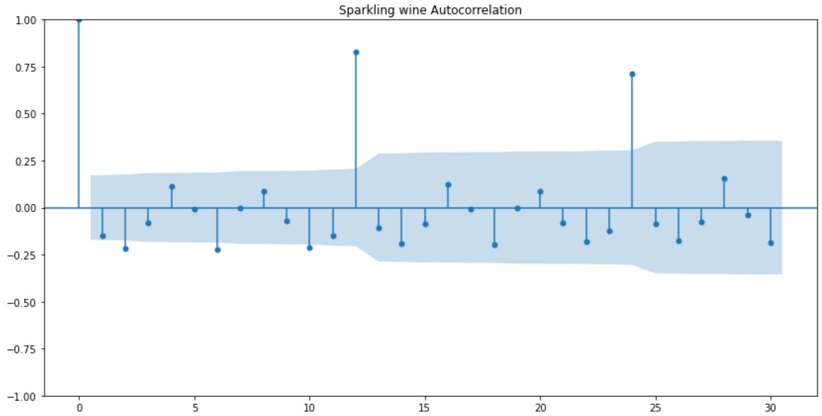


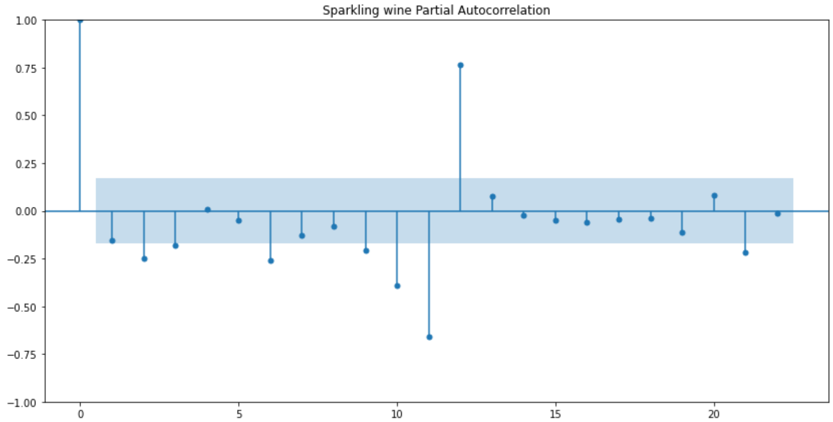
RMSE:

ARIMA(2,1,3)	36.716972	76.827795

Sparkling Wine Time series:

We checked the Stationarity of the train data of Sparkling sales data at alpha 0.05 and observe from the result table that data was not stationary hence we took a difference of order 1 and found the train data was stationary then.





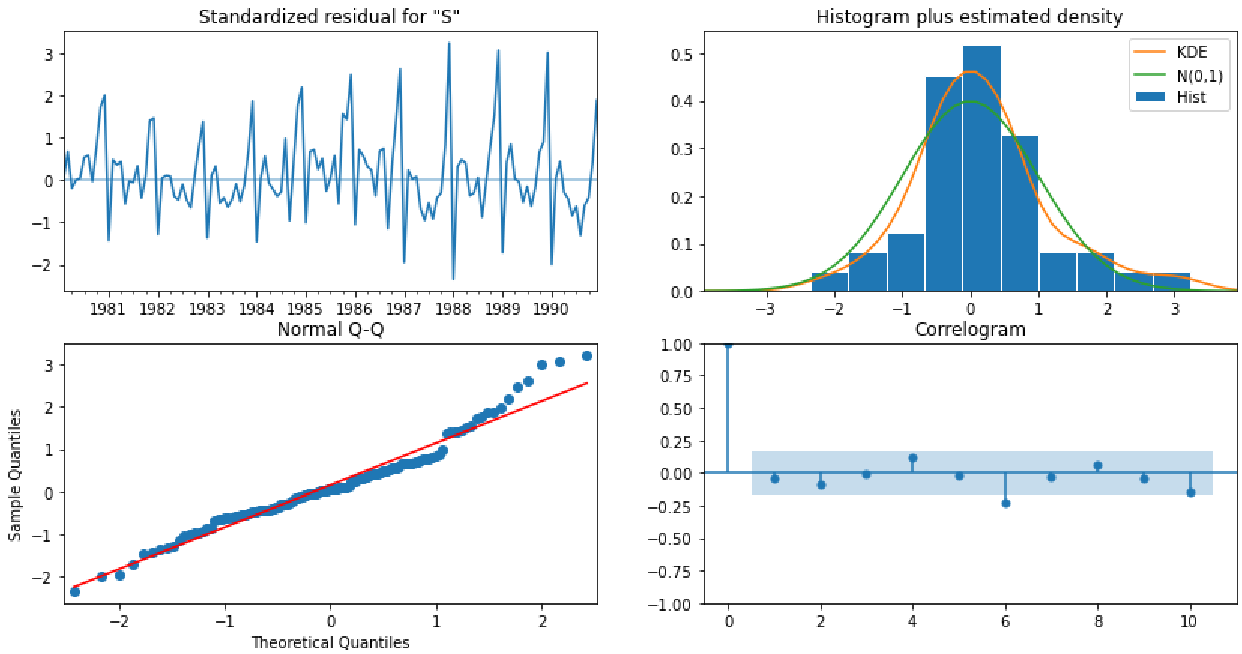
Build an Automated version of an ARIMA model for which the best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC).

AIC	Param
2213.509213	(2, 1, 2)
2221.452079	(3, 1, 3)
2230.769089	(3, 1, 2)

2232.901099	(2, 1, 3)
2233.777626	(2, 1, 1)

Lowest AIC: 2213 for (2,1,2)

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1101.755
Date:                Sun, 20 Mar 2022   AIC                           2213.509
Time:                        15:03:45   BIC                           2227.885
Sample:                    01-01-1980   HQIC                          2219.351
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.3121      0.046     28.782      0.000       1.223       1.401
ar.L2         -0.5593      0.072     -7.740      0.000      -0.701      -0.418
ma.L1         -1.9917      0.109    -18.214      0.000      -2.206      -1.777
ma.L2          0.9999      0.110      9.108      0.000       0.785       1.215
sigma2      1.099e+06      2e-07   5.51e+12      0.000     1.1e+06     1.1e+06
===================================================================================
```



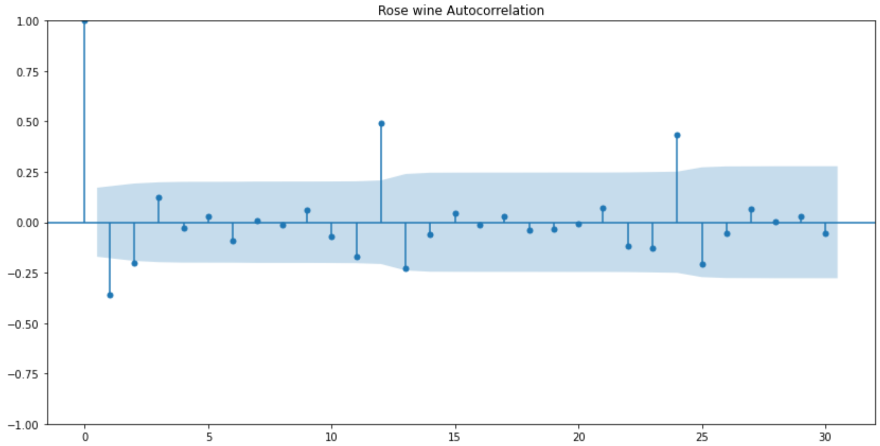
RMSE:

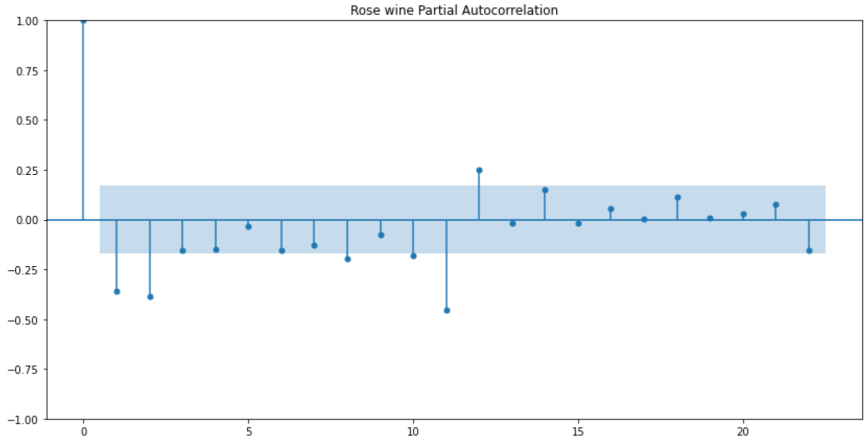
ARIMA(2,1,2)	1307.775680	45.931042

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

Solution: Manual ARIMA

Rose Wine Time Series:



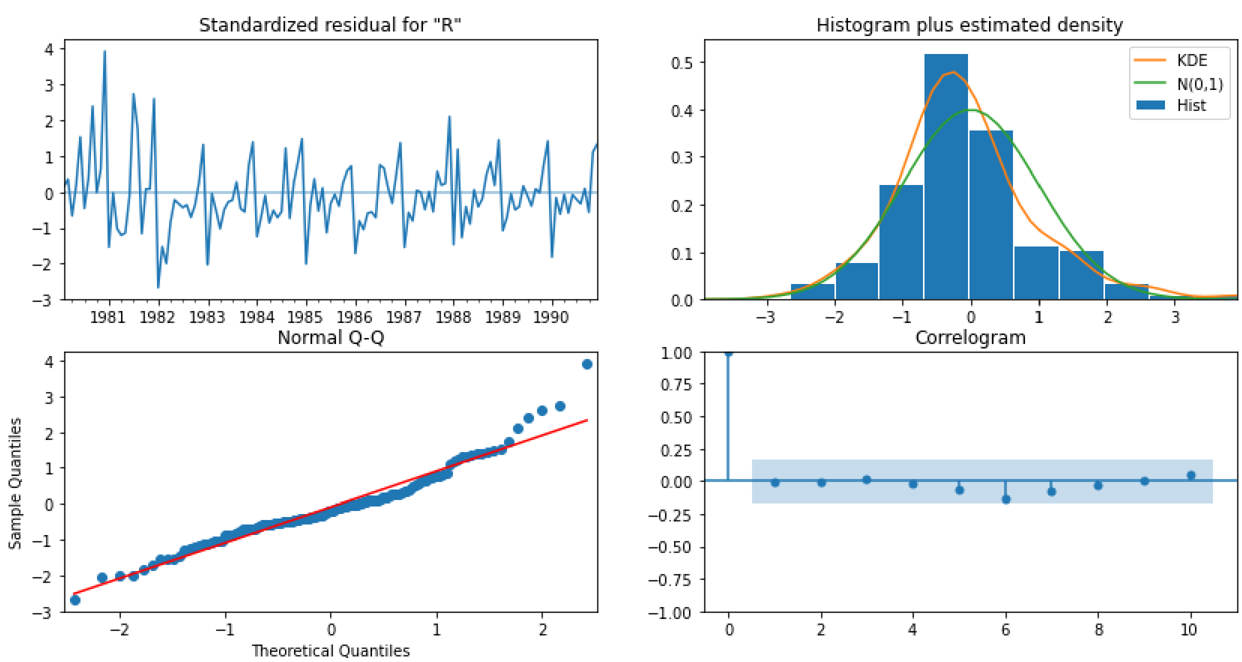


Referring to the ACF and PACF plots of Rose wine:

- The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off to 2.
- The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 2.

- By looking at the above plots, we will take the value of p and q to be 2 and 2 respectively.

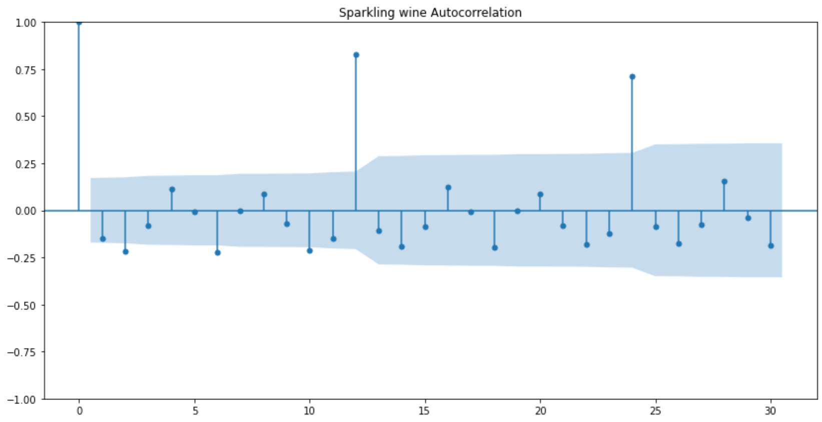
```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(2, 1, 2)   Log Likelihood                -635.935
Date:                Sun, 20 Mar 2022   AIC                           1281.871
Time:                        15:03:47   BIC                           1296.247
Sample:                    01-01-1980   HQIC                          1287.712
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.4540      0.469     -0.969      0.333      -1.372       0.464
ar.L2          0.0001      0.170      0.001      0.999      -0.334       0.334
ma.L1         -0.2541      0.459     -0.554      0.580      -1.154       0.646
ma.L2         -0.5984      0.430     -1.390      0.164      -1.442       0.245
sigma2       952.1601     91.424     10.415      0.000     772.973    1131.347
--------------------------------------------------------------------------------
```

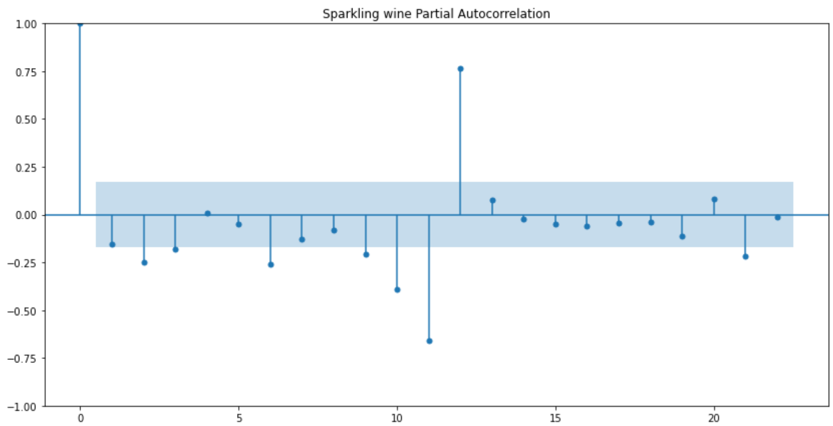


RMSE:

ARIMA(2,1,2)	36.945939	76.862961

Sparkling Wine Time series:





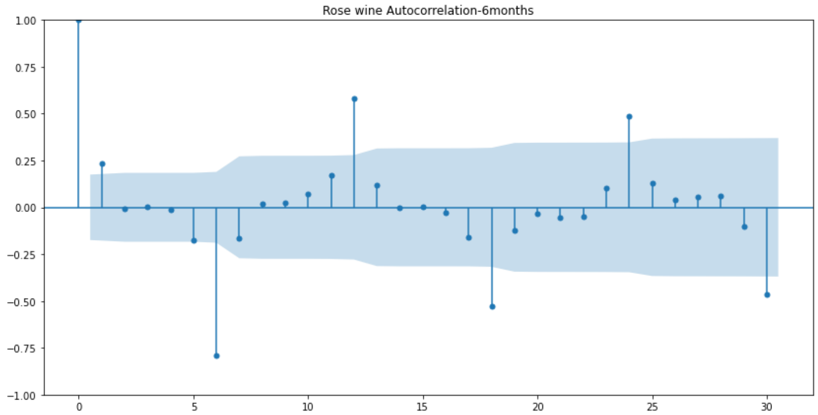
Referring to the ACF and PACF plots of Sparkling wine:

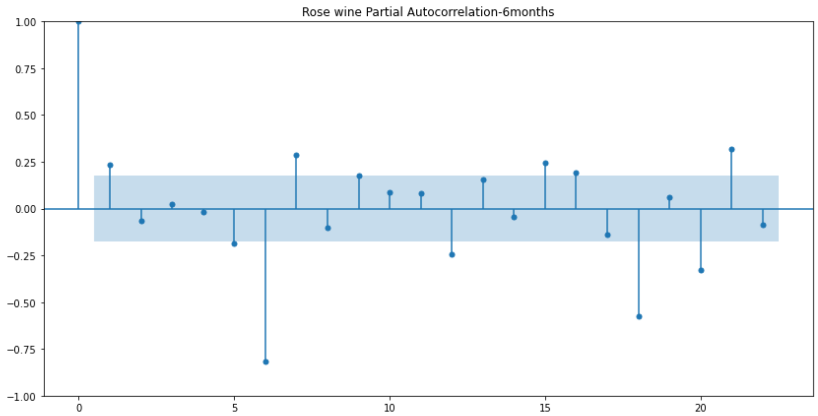
- The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off to 2.
- The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 2.
- By looking at the above plots, we will take the value of p and q to be 2 and 2 respectively.

Here, (p,d,q) is (2,1,2). Same as Automated ARIMA model. So the RMSE will also be Same.

Manual SARIMA Model for Rose Wine:

We derived the Seasonal parameters based on the seasonal cut-offs on the 6 months series plot and chose Seasonal AR parameter P value 2, Seasonal Moving average parameter Q value 9, differencing 1 and Seasonal period 6.



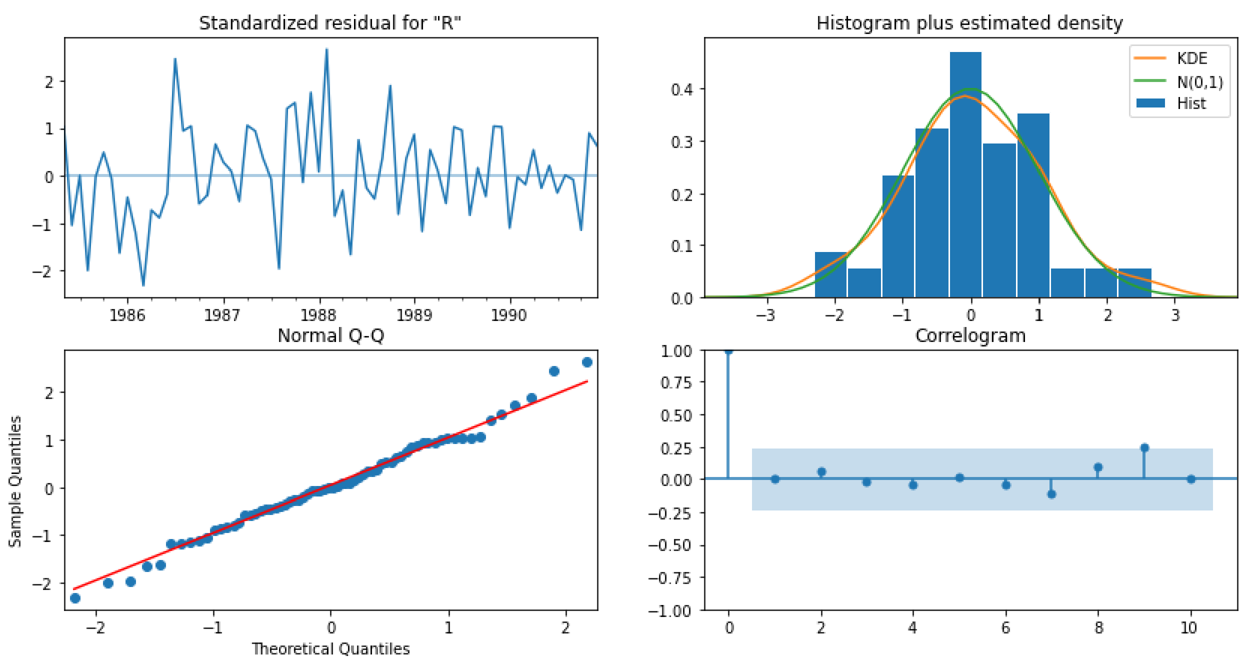


Building SARIMA model on order=(2,1,2), seasonal_order=(2, 1, 9, 6):

```
                                      SARIMAX Results
==========================================================================================================
Dep. Variable:                                         Rose   No. Observations:                    132
Model:            SARIMAX(2, 1, 2)x(2, 1, [1, 2, 3, 4, 5, 6, 7, 8, 9], 6)   Log Likelihood          -282.910
Date:                                      Sun, 20 Mar 2022   AIC                                597.821
Time:                                              15:05:11   BIC                                633.333
Sample:                                          01-01-1980   HQIC                               611.892
                                               - 12-01-1990
Covariance Type:                                        opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.9397      0.217     -4.322      0.000      -1.366      -0.514
ar.L2         -0.0719      0.172     -0.419      0.675      -0.408       0.265
ma.L1          0.0583     19.434      0.003      0.998     -38.032      38.149
ma.L2         -0.9427     18.298     -0.052      0.959     -36.805      34.920
ar.S.L6       -1.1278      0.778     -1.451      0.147      -2.652       0.396
ar.S.L12      -0.1736      0.763     -0.228      0.820      -1.668       1.321
ma.S.L6        0.5893    254.486      0.002      0.998    -498.194     499.373
ma.S.L12      -1.2519     93.189     -0.013      0.989    -183.898     181.395
ma.S.L18      -0.0541    102.461     -0.001      1.000    -200.874     200.765
ma.S.L24       0.0233     65.478      0.000      1.000    -128.312     128.359
ma.S.L30      -0.2304     80.433     -0.003      0.998    -157.876     157.415
ma.S.L36       0.0435     42.236      0.001      0.999     -82.737      82.824
ma.S.L42       0.2734     50.713      0.005      0.996     -99.122      99.669
ma.S.L48       0.4042     41.300      0.010      0.992     -80.542      81.350
ma.S.L54       0.1336     20.754      0.006      0.995     -40.543      40.811
sigma2        94.3162   1.63e+04      0.006      0.995   -3.18e+04     3.2e+04
===================================================================================
```

RMSE:

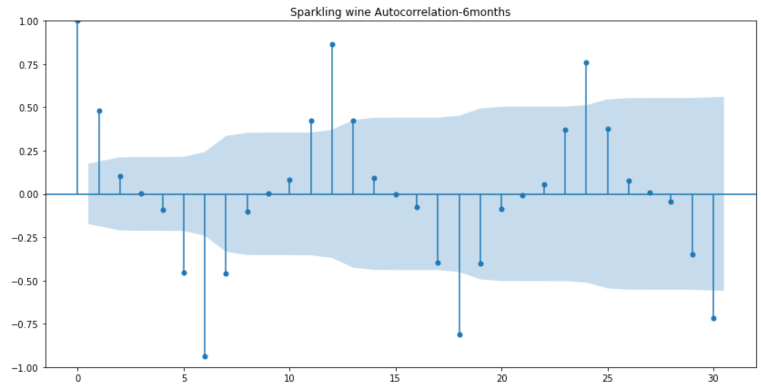
SARIMA(2,1,2)(2,1,9,6)	28.524310	39.621799

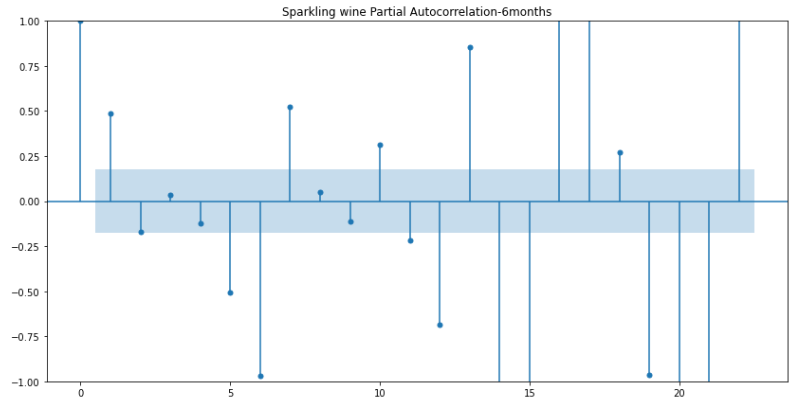


- Model diagnostics confirms that the model residuals are normally distributed. Standardized residual do not display any obvious seasonality, Histogram plus estimated density - The KDE plot has normal distribution , Normal Q-Q plot – There is an ordered distribution of residuals (blue dots) following the linear trend Correlogram – The time series residuals have low correlation with lagged versions of itself.

Manual SARIMA for Sparkling Wine:

We then derive the Seasonal parameters based on the seasonal cut-offs on the 6 months series plot and chose Seasonal AR parameter P value 1, Seasonal Moving average parameter Q value 2, differencing 1 and Seasonal period 6.



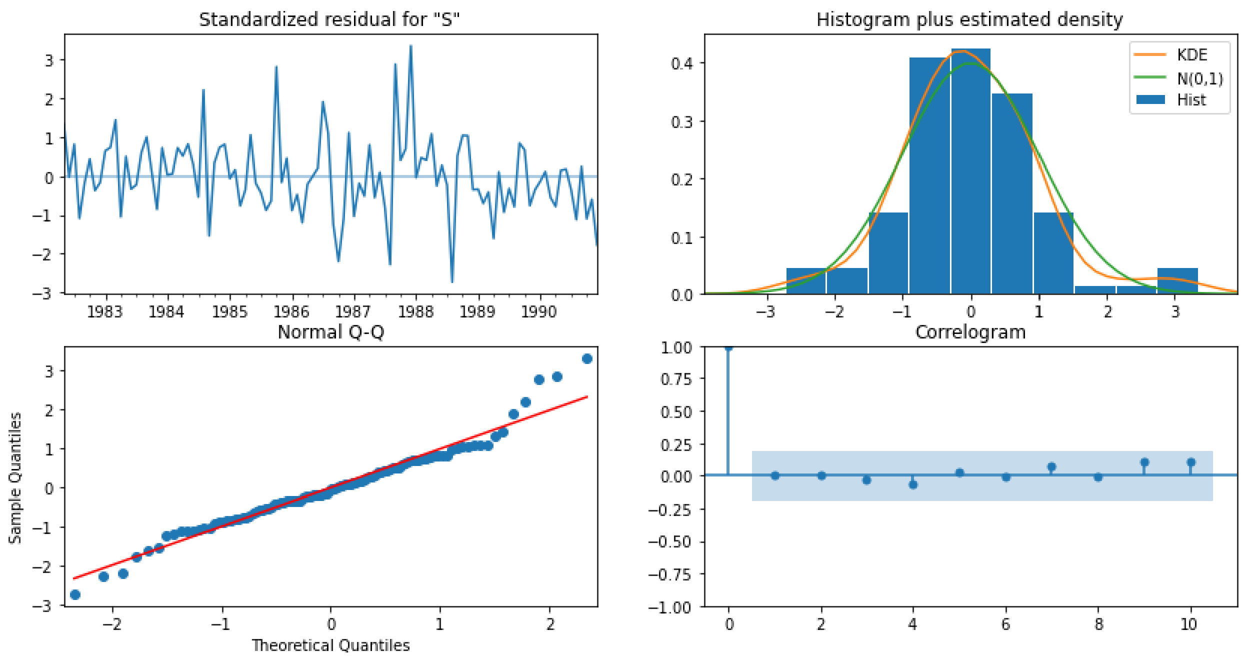


Building SARIMA model on order=(2,1,2), seasonal_order=(1, 1, 3, 6):

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                              Sparkling   No. Observations:              132
Model:             SARIMAX(2, 1, 2)x(1, 1, [1, 2, 3], 6)   Log Likelihood            -770.858
Date:                                Sun, 20 Mar 2022   AIC                       1559.716
Time:                                        15:05:22   BIC                       1583.515
Sample:                                    01-01-1980   HQIC                      1569.358
                                         - 12-01-1990
Covariance Type:                                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.7171      0.270     -2.659      0.008      -1.246      -0.188
ar.L2         -0.0388      0.143     -0.272      0.786      -0.319       0.241
ma.L1         -0.0533      0.241     -0.221      0.825      -0.526       0.420
ma.L2         -0.7395      0.205     -3.614      0.000      -1.141      -0.338
ar.S.L6       -1.0230      0.008   -124.082      0.000      -1.039      -1.007
ma.S.L6        0.3399      0.295      1.154      0.249      -0.237       0.917
ma.S.L12      -0.7353      0.191     -3.859      0.000      -1.109      -0.362
ma.S.L18       0.1576      0.172      0.916      0.360      -0.180       0.495
sigma2      1.242e+05   3.48e+04      3.565      0.000    5.59e+04    1.93e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                13.37
Prob(Q):                              0.99   Prob(JB):                         0.00
Heteroskedasticity (H):               1.17   Skew:                             0.45
Prob(H) (two-sided):                  0.65   Kurtosis:                         4.51
===================================================================================
```

RMSE:

SARIMA(2,1,2)(1,1,3,6)	1765.864840	34.292558



- Model diagnostics confirms that the model residuals are normally distributed. Standardized residual do not display any obvious seasonality, Histogram plus estimated density - The KDE plot has normal distribution , Normal Q-Q plot – There is an ordered distribution of residuals (blue dots) following the linear trend Correlogram – The time series residuals have low correlation with lagged versions of itself

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

Solution:

Rose Wine: RMSE values on different models

Model with parameters	RMSE	MAPE
Alpha=0.098,SES	36.861106	NaN
Alpha=1.4e-8,Beta=7.3e-9:DES	15.369605	NaN
Alpha=0.0946,Beta=2.3e-5,Gamma=0.0004:TES-Add	27.110218	NaN
Alpha=0.0498,Beta=0.0406,Gamma=0.00078:TES-Mul	29.530500	NaN
RegressionOnTime	15.369603	NaN
NaiveModel	79.750253	NaN
2pointTrailingMovingAverage	10.354667	NaN
4pointTrailingMovingAverage	13.725015	NaN
6pointTrailingMovingAverage	14.054791	NaN

9pointTrailingMovingAverage	14.370674	NaN
ARIMA(2,1,3)	36.716972	76.827795
ARIMA(2,1,2)	36.945939	76.862961
SARIMA(2,1,2)(2,1,9,6)	28.524310	39.621799

Sparkling Wine: RMSE values on different models

Models with parameters	**RMSE**	**MAPE**
Alpha=0.070,SES	1344.734375	NaN
Alpha=0.6649,Beta=0.0001,initial_level=1.0,initial_trend=1.0:DES	3193.863577	NaN
Alpha=0.1112,Beta=0.0123,Gamma=0.4607:TES-Add	1575.981106	NaN
Alpha=0.1110,Beta=0.0493,Gamma=0.362:TES-Mul	1448.376991	NaN
RegressionOnTime	1394.441182	NaN
NaiveModel	3858.888237	NaN
2pointTrailingMovingAverage	770.928742	NaN

4pointTrailingMovingAverage	1137.13705 3	NaN
6pointTrailingMovingAverage	1283.09699 3	NaN
9pointTrailingMovingAverage	1354.27793 8	NaN
ARIMA(2,1,2)	1307.77568 0	45.93104 2
ARIMA(2,1,2)	1307.77568 0	45.93104 2
SARIMA(2,1,2)(1,1,3,6)	1765.86484 0	34.29255 8

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

Solution:

Rose Wine Time series:

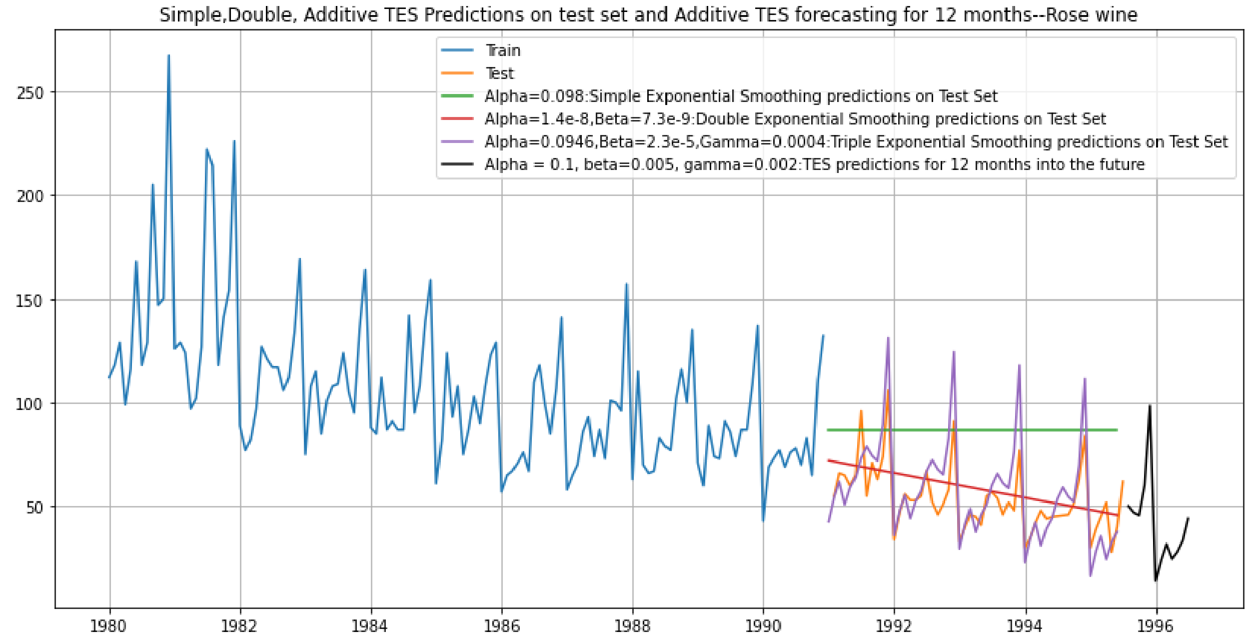
We observed from the RMSE scores that Triple Exponential with additive trend and seasonality would work better for the Rose Sales data where we had seasonality and Trend.

Building a model on the whole data using the parameters got from best fit:

We see that the best model is the Triple Exponential Smoothing (Holt-winter method) with additive trend and seasonality with parameters α = 0.1, β = 0.005 and γ = 0.002.

Forecasting for 12 months into the future using the above model:

1995-08-01	50.123621
1995-09-01	46.952806
1995-10-01	45.709683
1995-11-01	60.308535
1995-12-01	98.500361
1996-01-01	14.085685
1996-02-01	24.343326
1996-03-01	31.877575
1996-04-01	24.708757
1996-05-01	28.108126
1996-06-01	33.580485
1996-07-01	44.200468



Sparkling Wine Time series data:

We observed from the RMSE scores that Triple Exponential with multiplicative trend and seasonality would work better for the Sparkling Sales data where we had seasonality and Trend.

Building a model on the whole data using the parameters got from best fit:

We see that the best model is the Triple Exponential Smoothing (Holt-winter method) with multiplicative trend and seasonality with parameters $\alpha = 0.075$, $\beta = 0.065$ and $\gamma = 0.273$

Forecasting for 12 months into the future using the above model:

1995-08-01 1934.374456

1995-09-01 2354.471507

1995-10-01 3182.615708

1995-11-01 3922.288284

1995-12-01 5991.347796

1996-01-01 1358.902328

1996-02-01 1600.746579

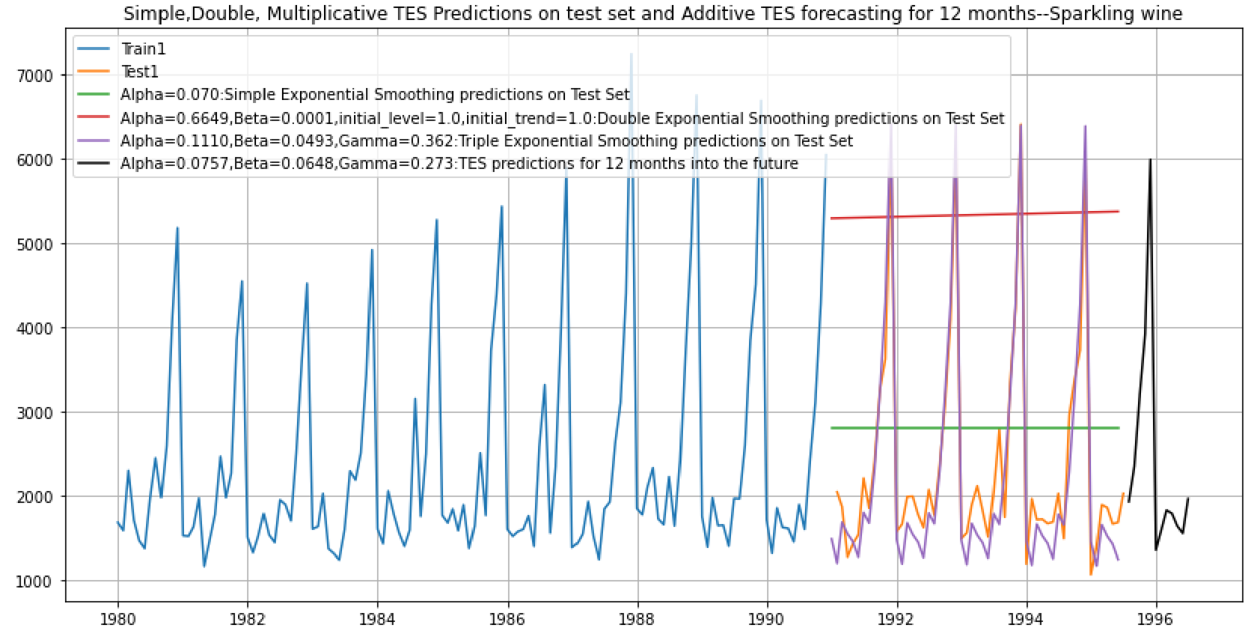
1996-03-01 1831.938577

1996-04-01 1792.725979

1996-05-01 1643.529843

1996-06-01 1557.747541

1996-07-01 1968.016093



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

<u>Inference and Recommendations – Rose Sales</u>

- Rose sales show a decrease in trend compared to the previous years.
- December month shows the highest Sales across the year while the value has come down through the years from 1980-1994.
- The models are built considering the Trend and Seasonality in to account and we see from the output plot that the future prediction is in line with the trend and seasonality in the previous years.
- The Sales of Rose wine is seasonal and also has a trend, hence the company cannot have the same stock through the year. The predictions would help here to plan the Stock need basis the forecasted sales.
- Apart from higher sale in November and December months, Rose sales will be above average in the summer months of July and August.
- The forecast also indicates that the year-on-year sale of wine is not showing an upward trend. The winery should investigate the low demand for Rose wine in market and make corrective actions in marketing and promotions.
- Identify retailers with low reorder or high refusal rates.
- Compare actual performance to goals and uncover sales opportunities.

Inference and Recommendations – Sparkling Sales

- Sparkling sales shows stabilized values and not much trend compared to previous years
- December month shows the highest Sales across the years from 1980-1994
- The models are built considering the Trend and Seasonality in to account and we see from the output plot that the future prediction is in line with the trend and seasonality in the previous years
- The Sales of Sparkling wine is seasonal, hence the company cannot have the same stock through the year. The predictions would help here to plan the Stock need basis the forecasted sales.
- The forecast also indicates that the year-on-year sale of wine is not showing an upward trend. The winery must adopt innovative marketing skills to improve the sale compared to previous years.
- The progress against sales targets can be tracked.
- Identify retailers with low reorder or high refusal rates.