TIME SERIES FORECASTING PROJECT

Data Analysis Report

Prepared By

JAI GOUTHAM

Submitted on
28-03-2021

PROJECT OBJECTIVE

Problem Statement:

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 training and test. The test data should start in 1991.
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.
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.

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

Data Insights and plotting:

	YearMonth	Rose
0	1980-01	112.000000
1	1980-02	118.000000
2	1980-03	129.000000
3	1980-04	99.000000
4	1980-05	116.000000

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

Description of dataset:

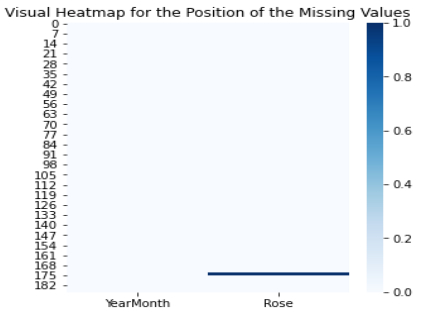
	count	mean	std	min	25%	50%	75%	max
Rose	187.0	89.914439	39.238325	28.0	62.5	85.0	111.0	267.0

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

Checking for Missing Values:

The Sparkling wine dataset does contain '0' missing/Null values and Rose wine dataset contains 2 Nan values.

```
The number of null values present in the Sparking dataset is
 YearMonth    0
Sparkling     0
dtype: int64


The number of null values present in the Sparking dataset is
 YearMonth    0
Rose          2
dtype: int64
```



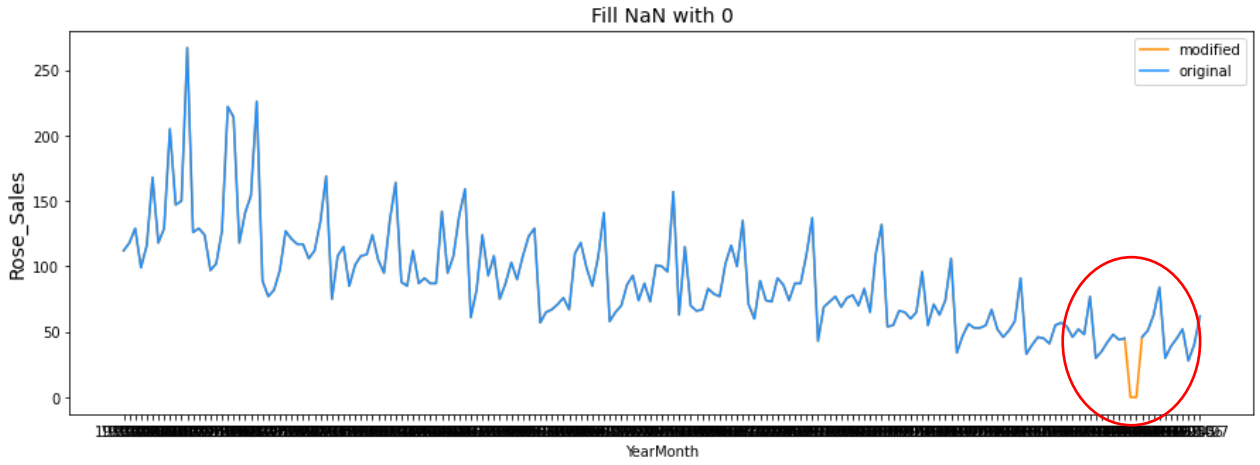
We can see that Rose variable has missing values.

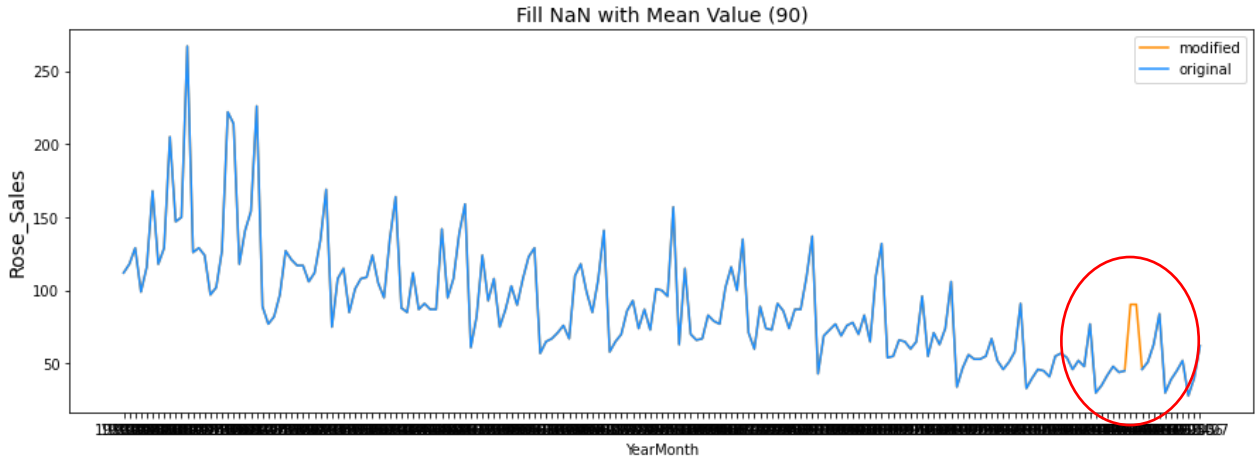
- **Option 1: Fill NaN with Mean Value**

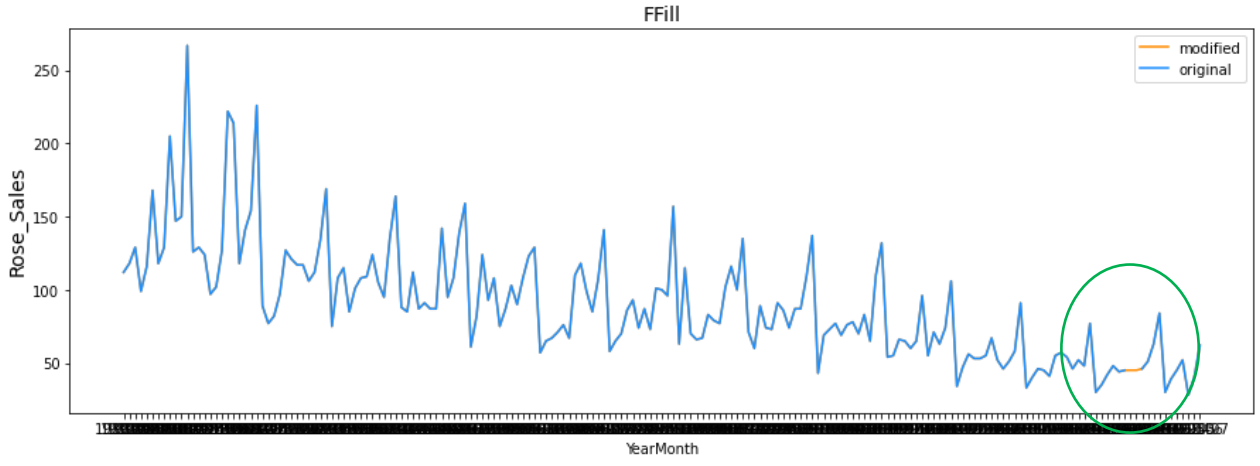
  Also in this example, we can see that filling NaNs with the mean value is also not sufficient.
- **Option 3: Fill NaN with Last Value with .ffill()**

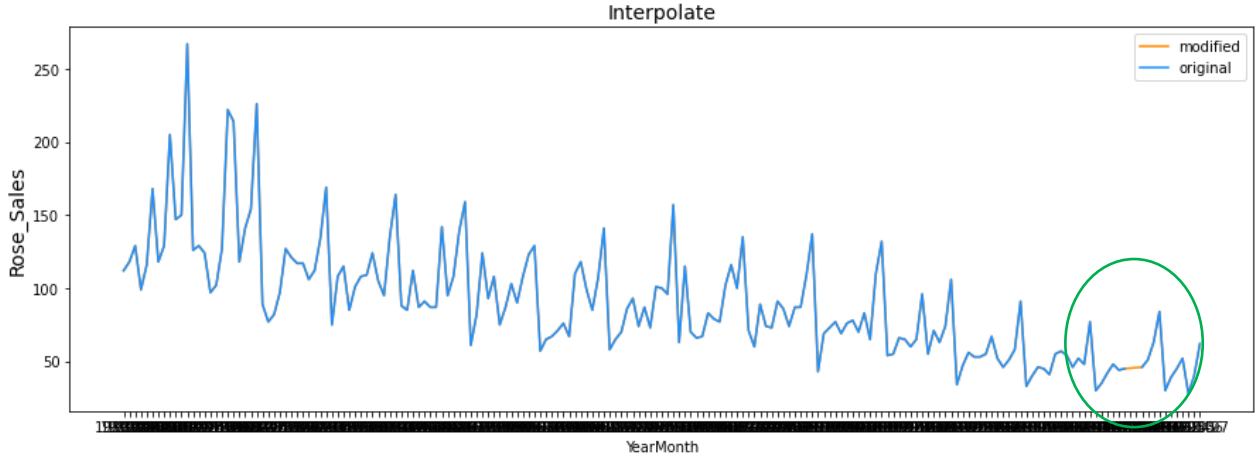
  Filling NaNs with the last value is already a little bit better in this case.
- **Option 4: Fill NaN with Linearly Interpolated Value with .interpolate()**

  Filling NaNs with the interpolated values is the best option in this small examlple but it requires knowledge of the neighouring values.


Fill NaN with 0
modified
original
Rose_Sales
YearMonth


Fill NaN with Mean Value (90)
modified
original
Rose_Sales
YearMonth


FFill
modified
original
Rose_Sales
YearMonth


Interpolate
modified
original
Rose_Sales
YearMonth

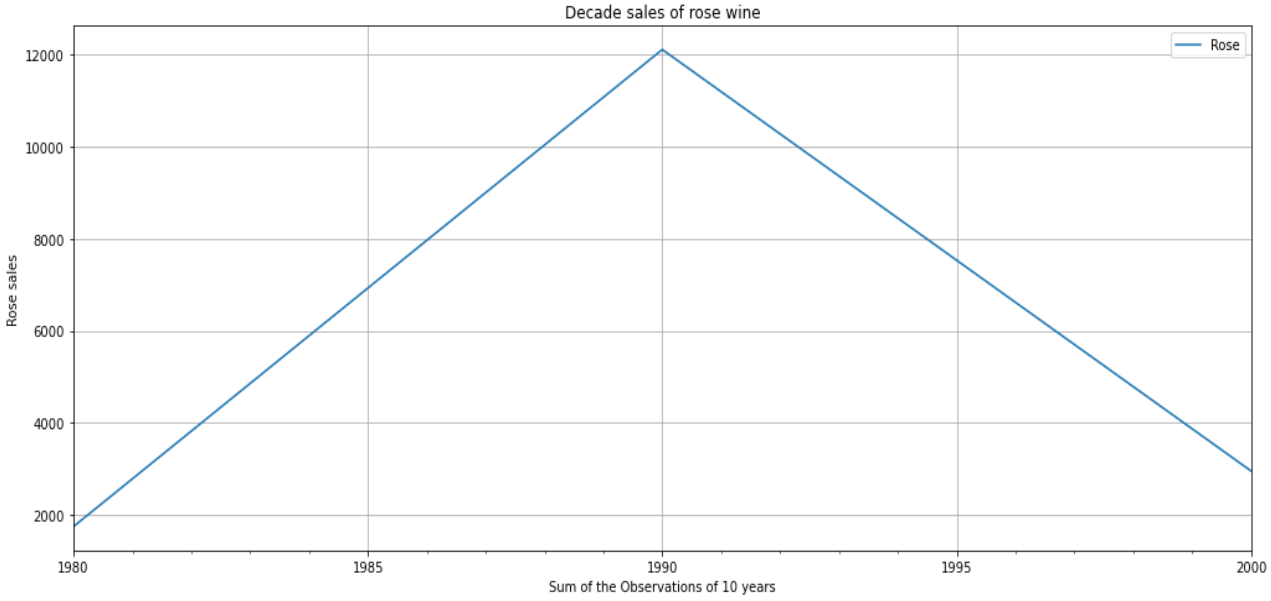
After Imputation (Interpolate method):

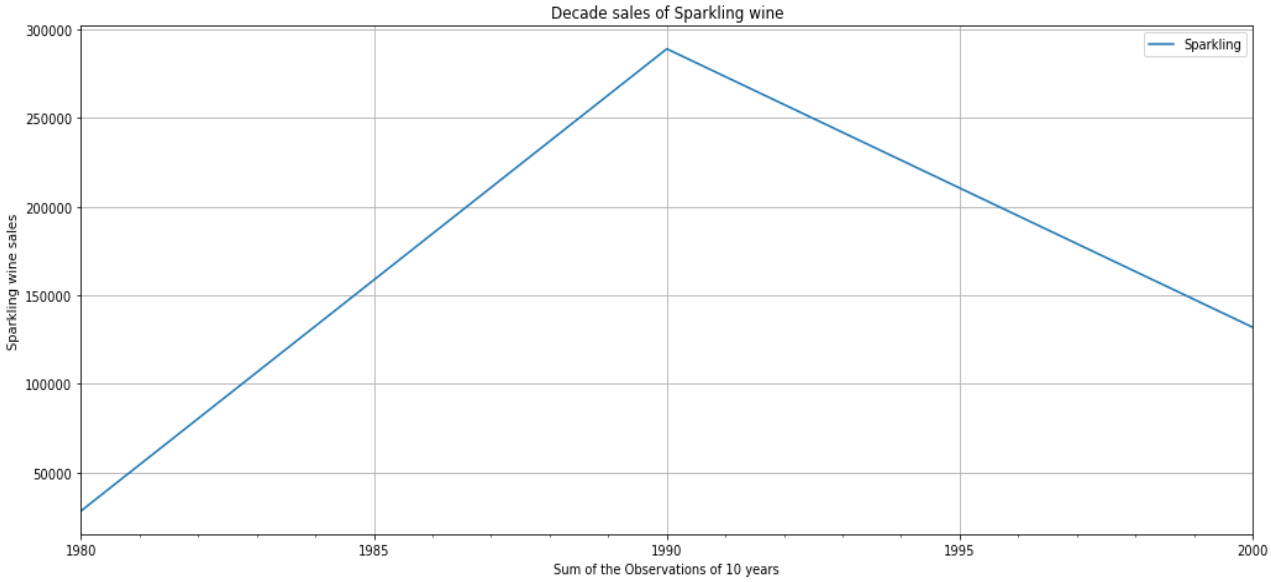
```
The number of null values present in the Sparking dataset is
 YearMonth    0
Sparkling    0
dtype: int64
```

```
The number of null values present in the Sparking dataset is
 YearMonth    0
Rose         0
dtype: int64
```

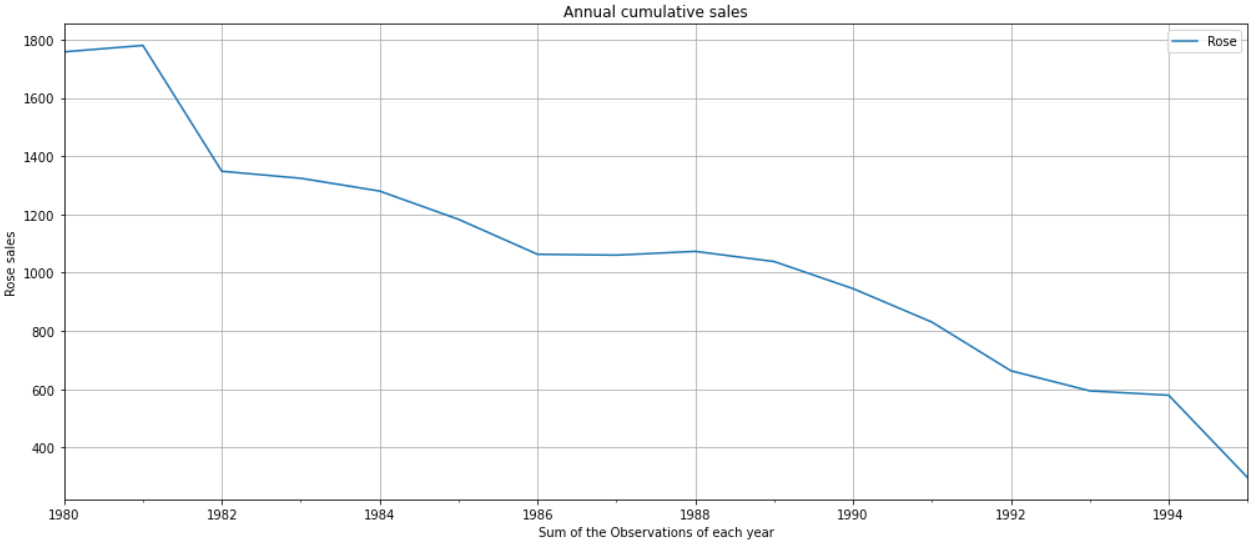
Plotting the data:

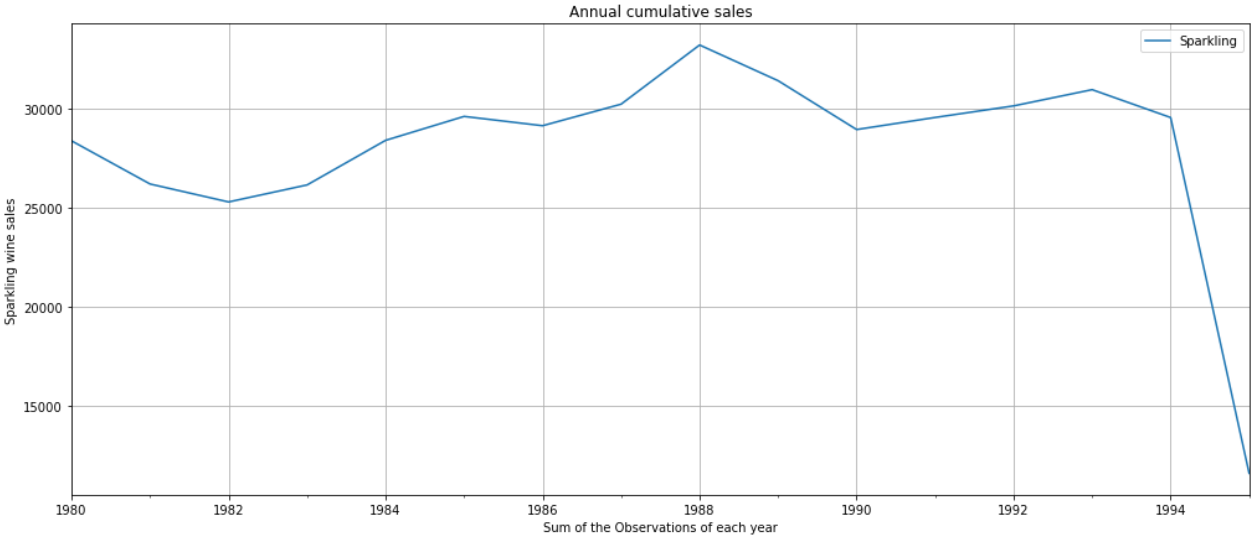
Decade plot for Rose wine and Sparkling wine dataset:



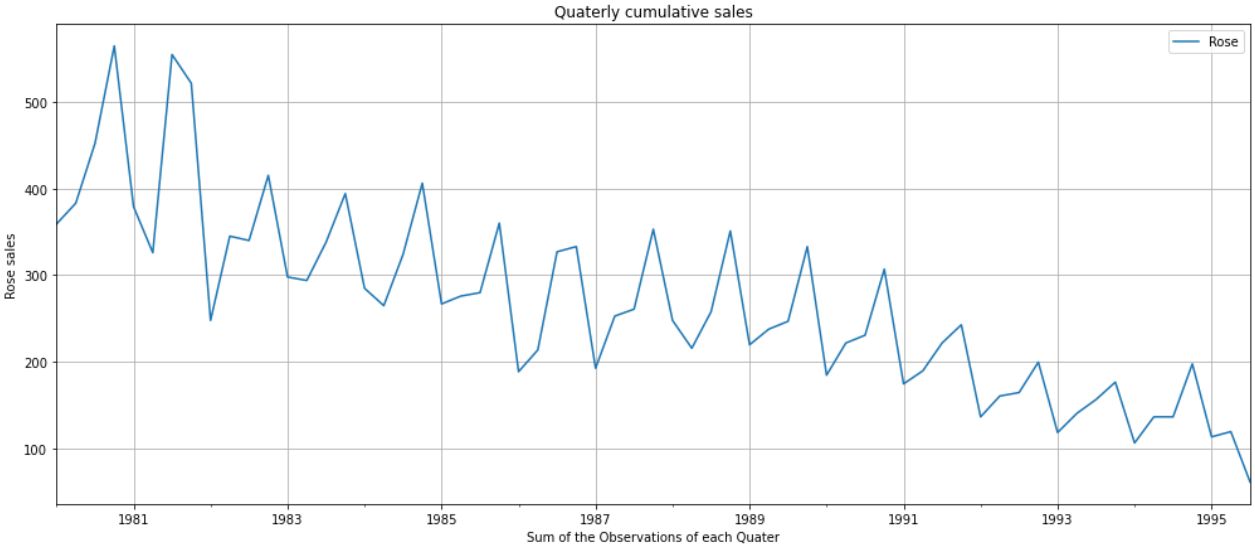


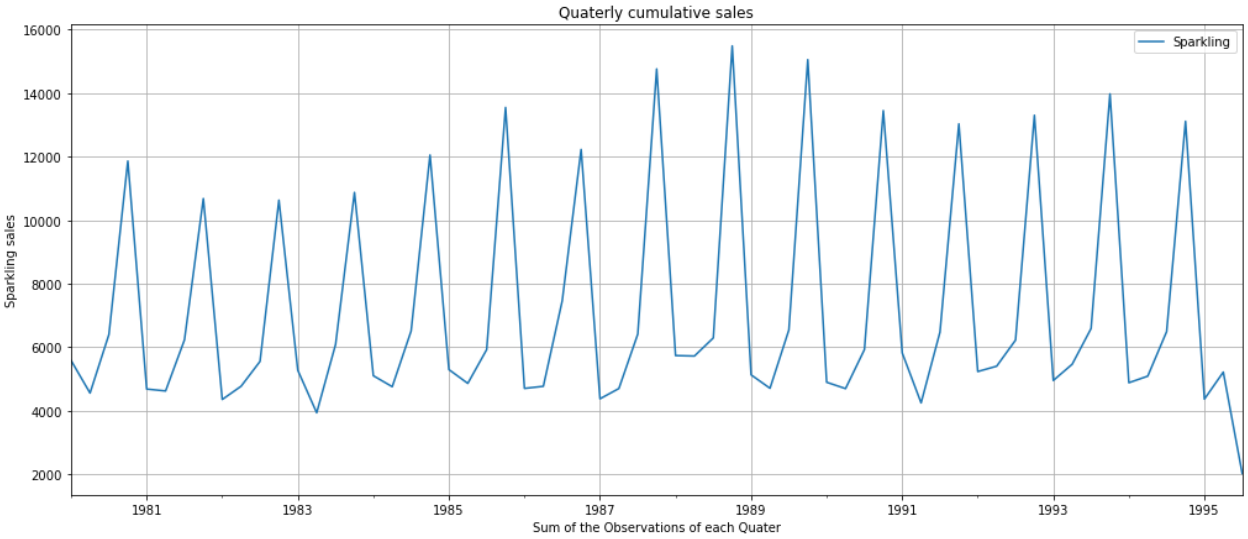
Annual cumulative sales plot for Rose wine and Sparkling wine dataset:



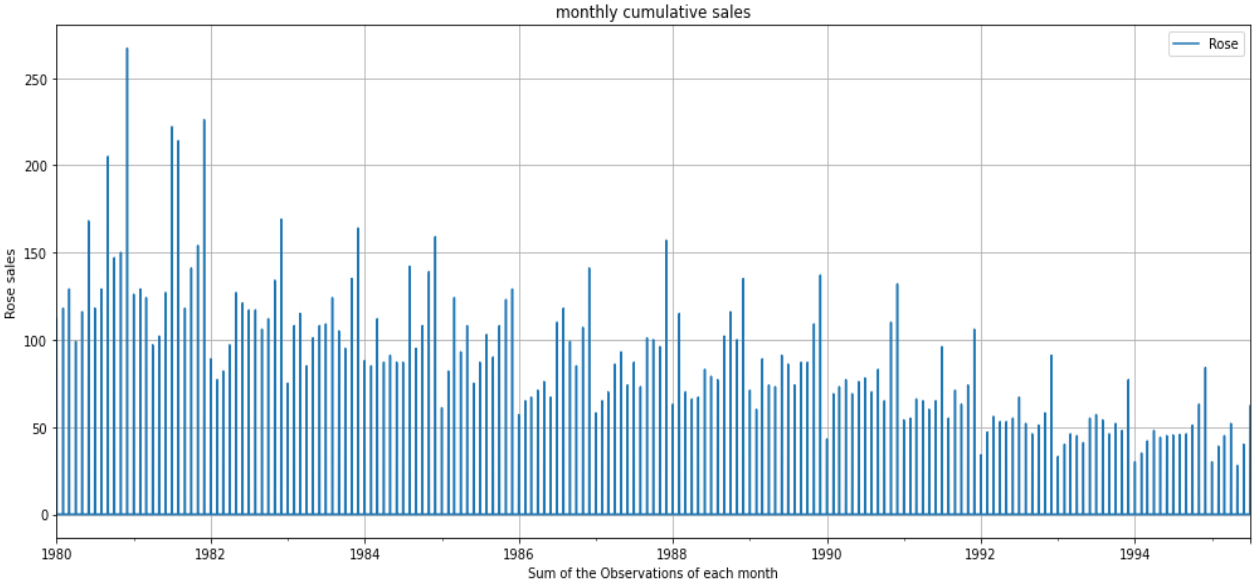


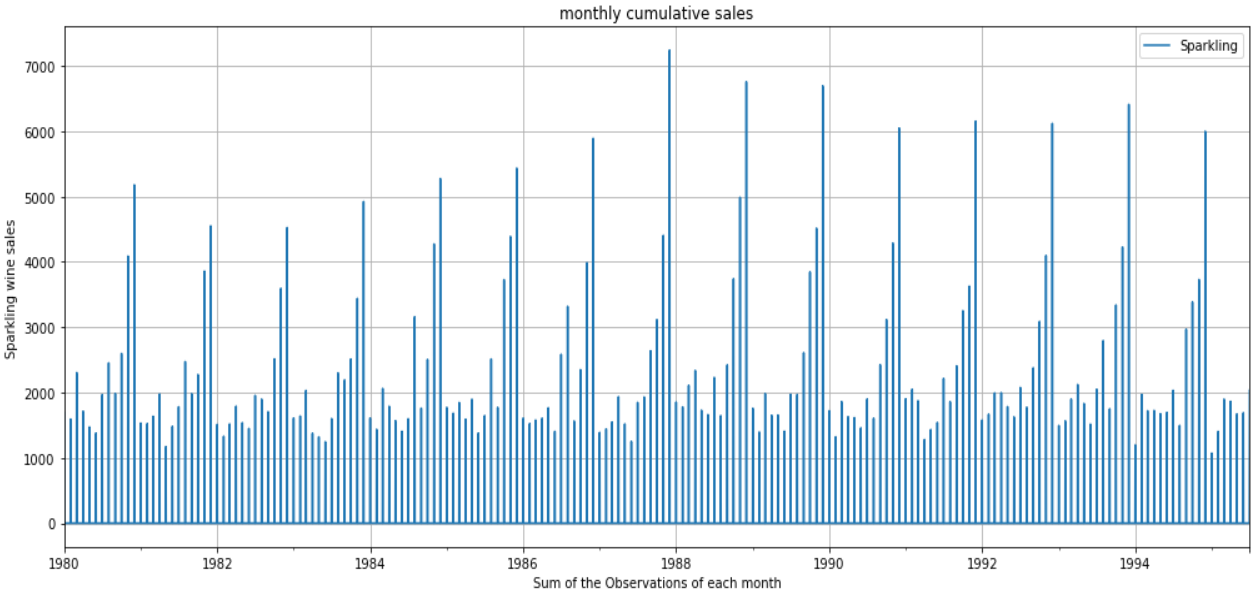
Quarterly sales plot for Rose wine and Sparkling wine dataset:





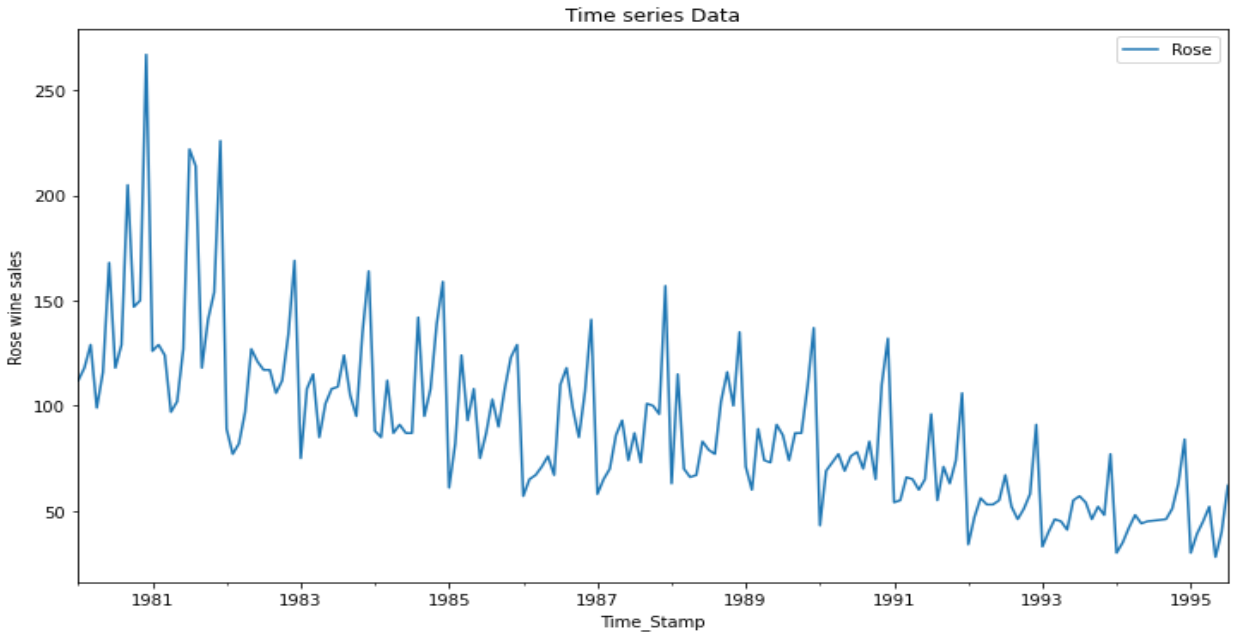
Monthly sales plot for Rose wine and Sparkling wine dataset:

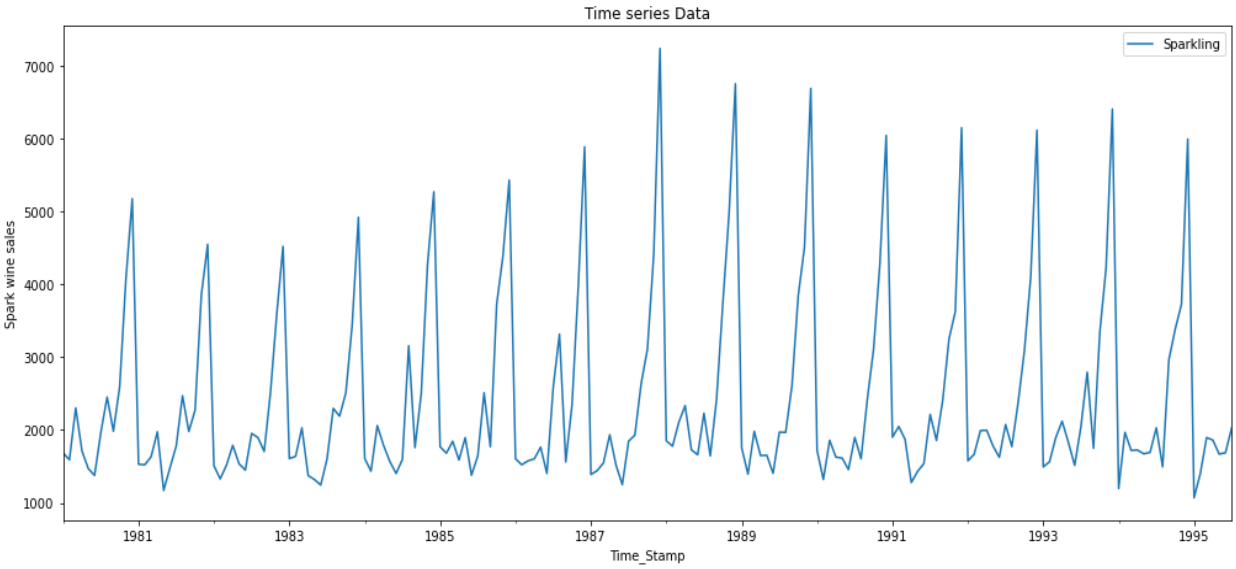




1.2 Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

Time Series plot for Rose wine and Sparkling wine dataset:

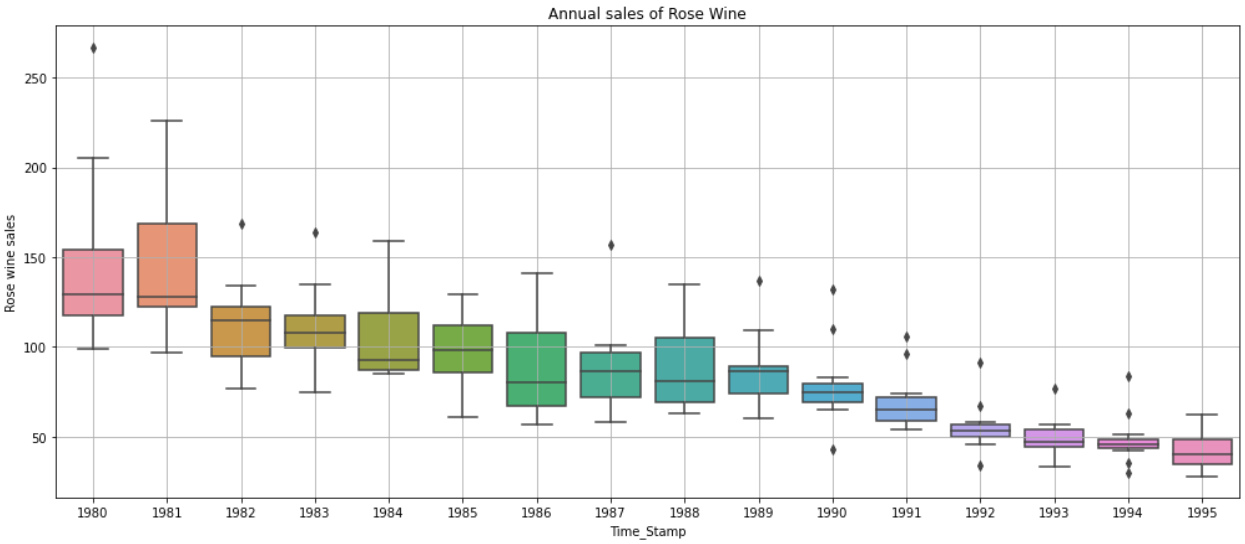


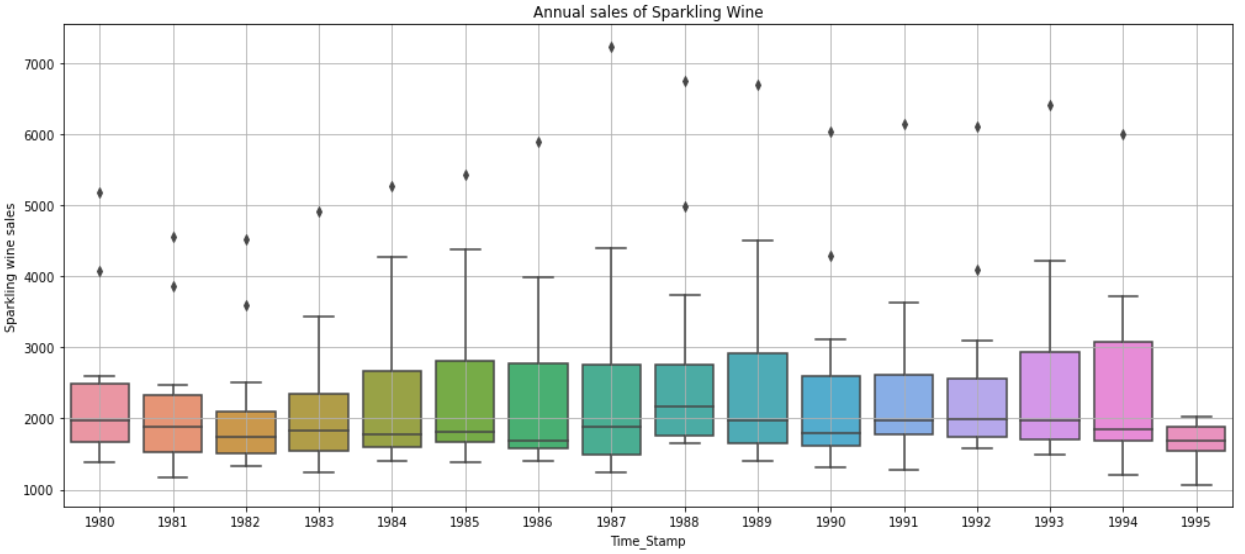


Inference from the Time series plot:

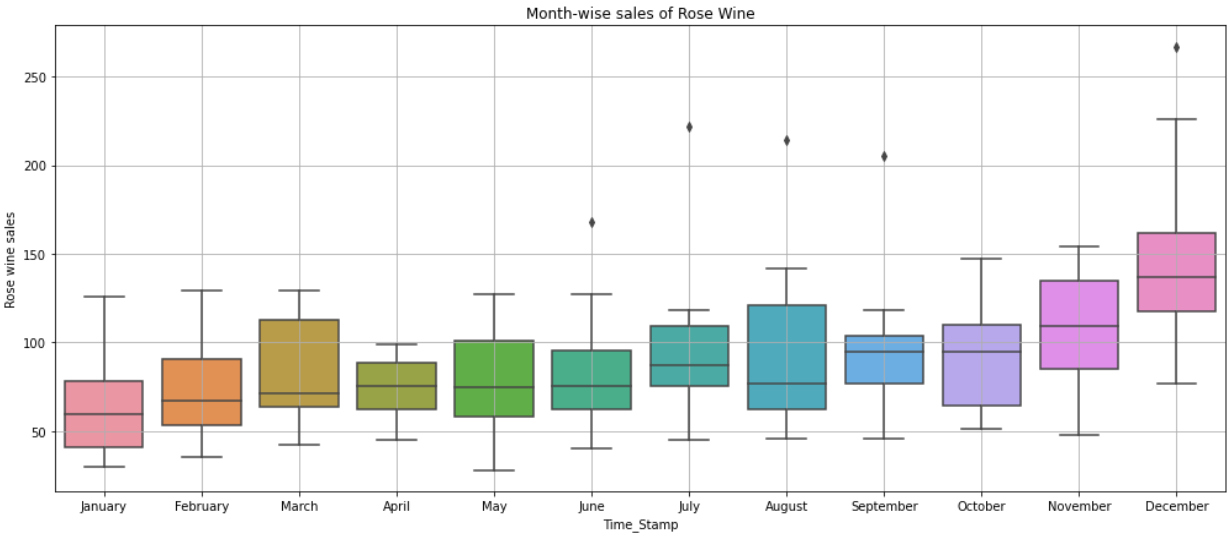
- We notice that there is a decreasing trend in the initial years which stabilizes after few years and again shows a decreasing trend in the rose wine time series plot.
- We also observe seasonality in the data trend and pattern seem to repeat on yearly basis on the rose wine time series plot.
- We notice that there is not much trend in the Sparkling wine time series plot.
- In sparkling wine time series, the seasonality seems to have a pattern on yearly basis.

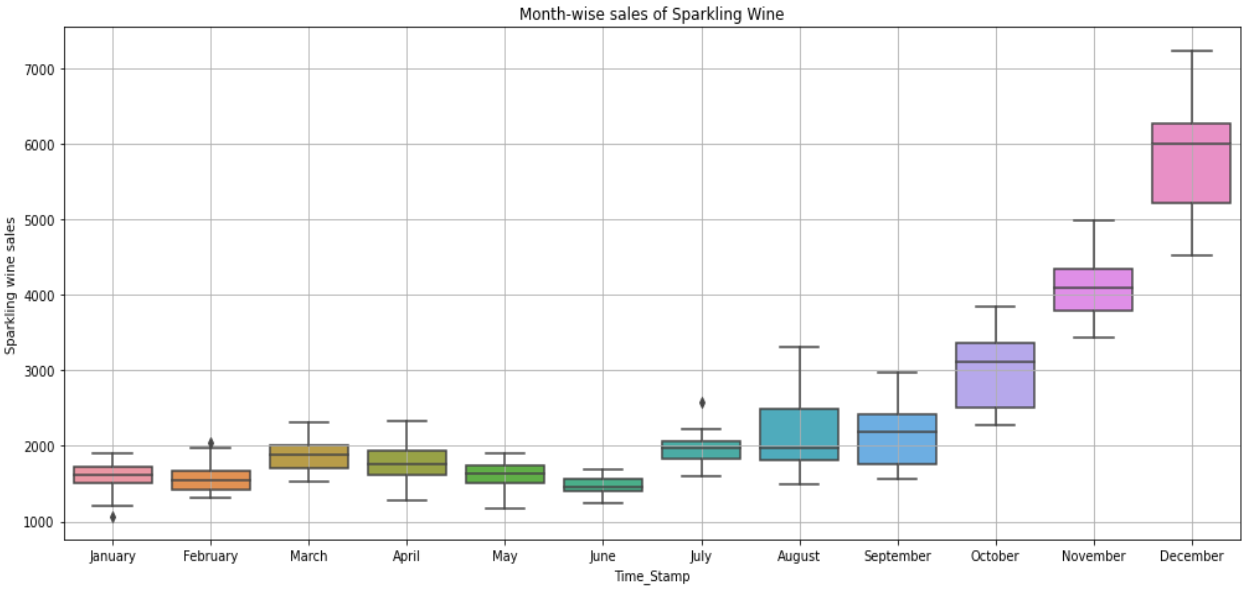
Annual Sales wise box plot for Rose wine and Sparkling wine dataset:





Monthly Sales wise box plot for Rose wine and Sparkling wine dataset:

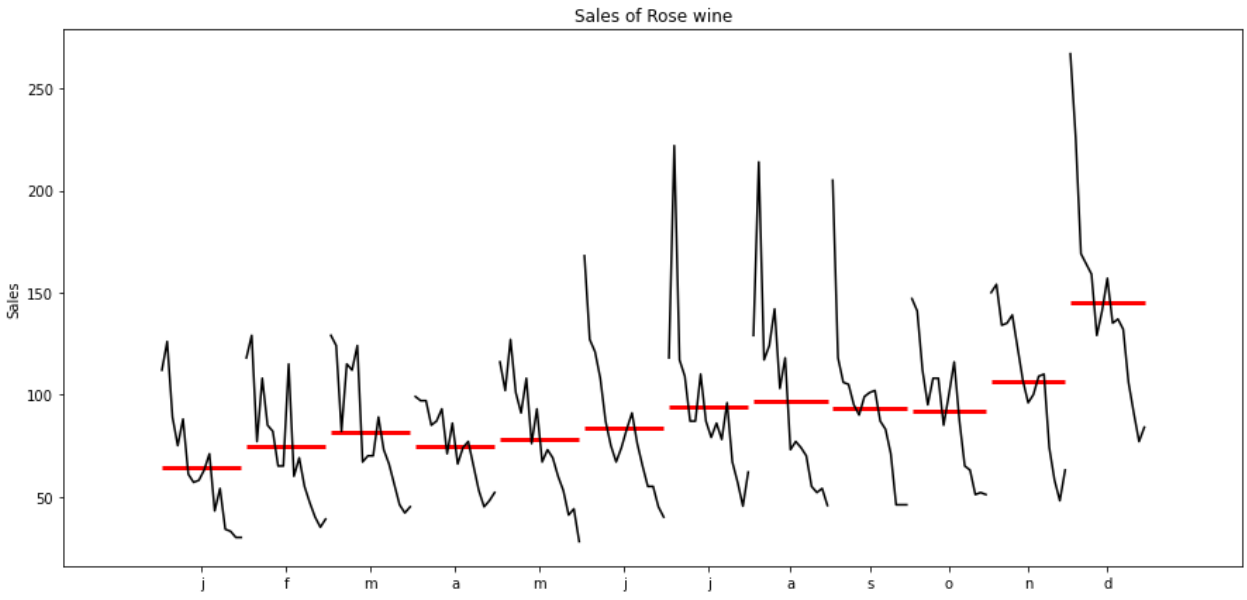


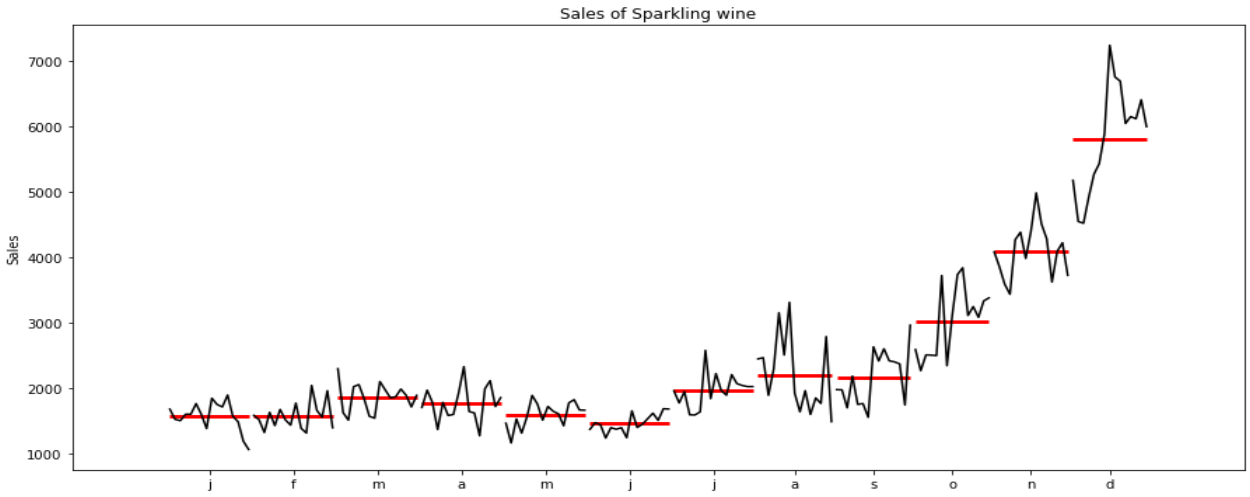


Inference from the Time series box plot:

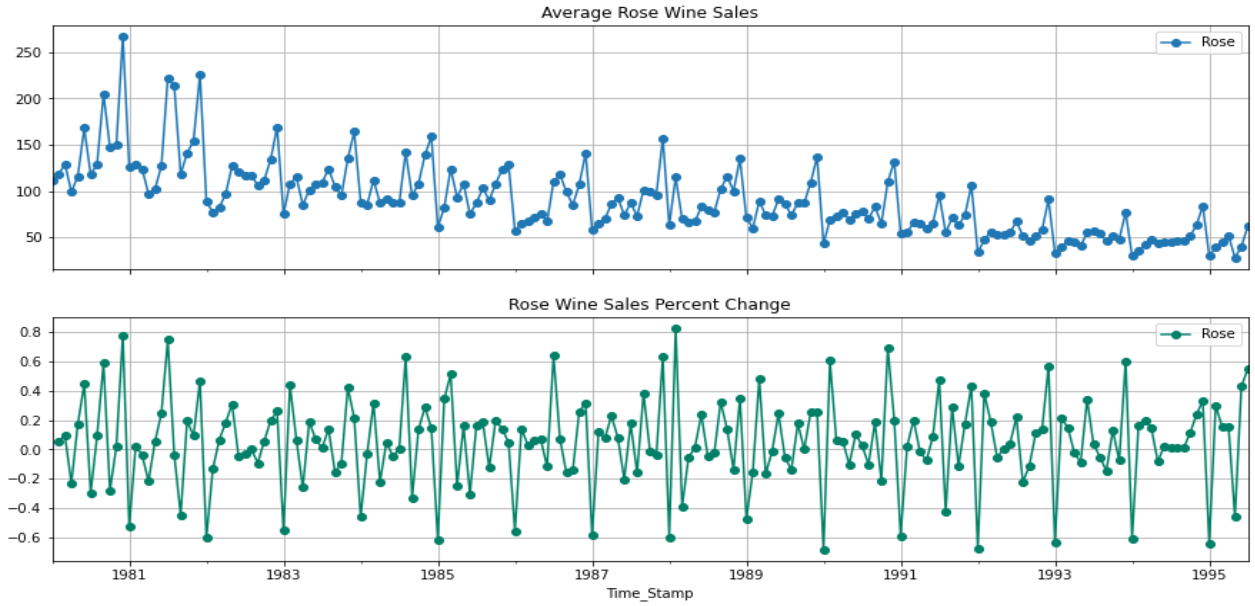
- We observed in the Time Series plot, the year wise boxplots over here also indicate a measure of downward trend for rose wine dataset.
- In the rose wine dataset, the highest sales is recorded in the year '1981'.
- We see that the sales of Rose wine have some outliers for certain years.
- December seems to have the highest sales of Rose wine and there is also outlier in June, July, August and September months
- In sparkling wine Time Series plot, the boxplots over here also do not indicate trend.
- we see that the sales of Sparkling wine have some outliers for almost all years except 1995.
- We also observe December month has the highest sales value for Sparkling wine

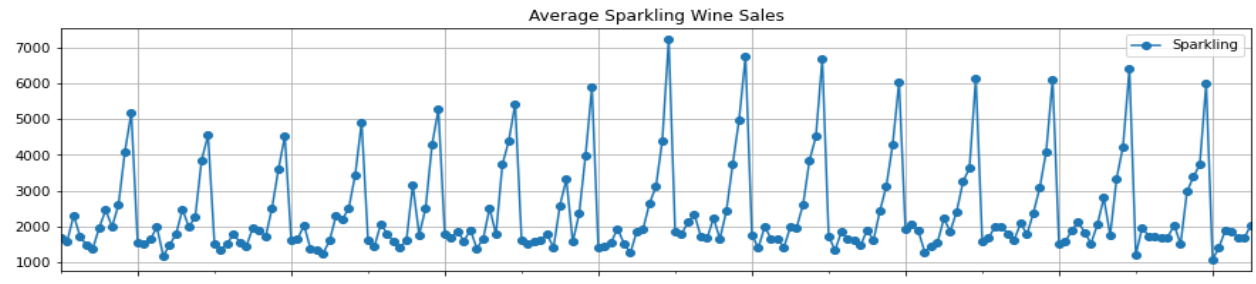
Time series Month plot for sales across years and month:

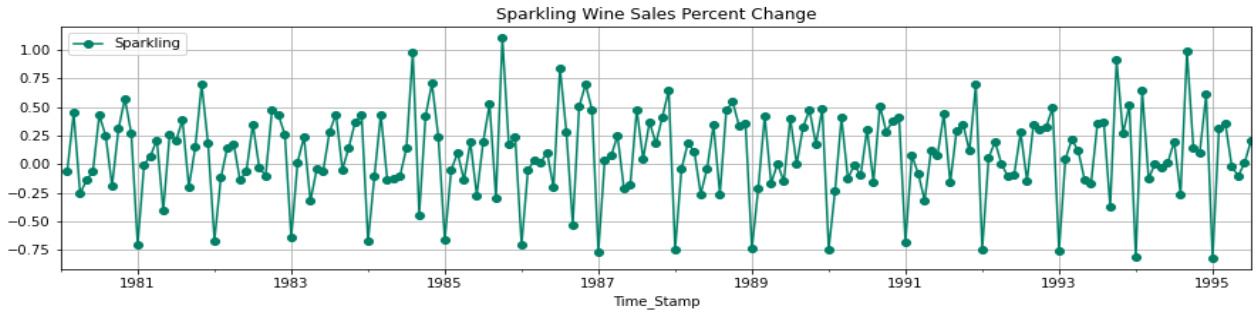




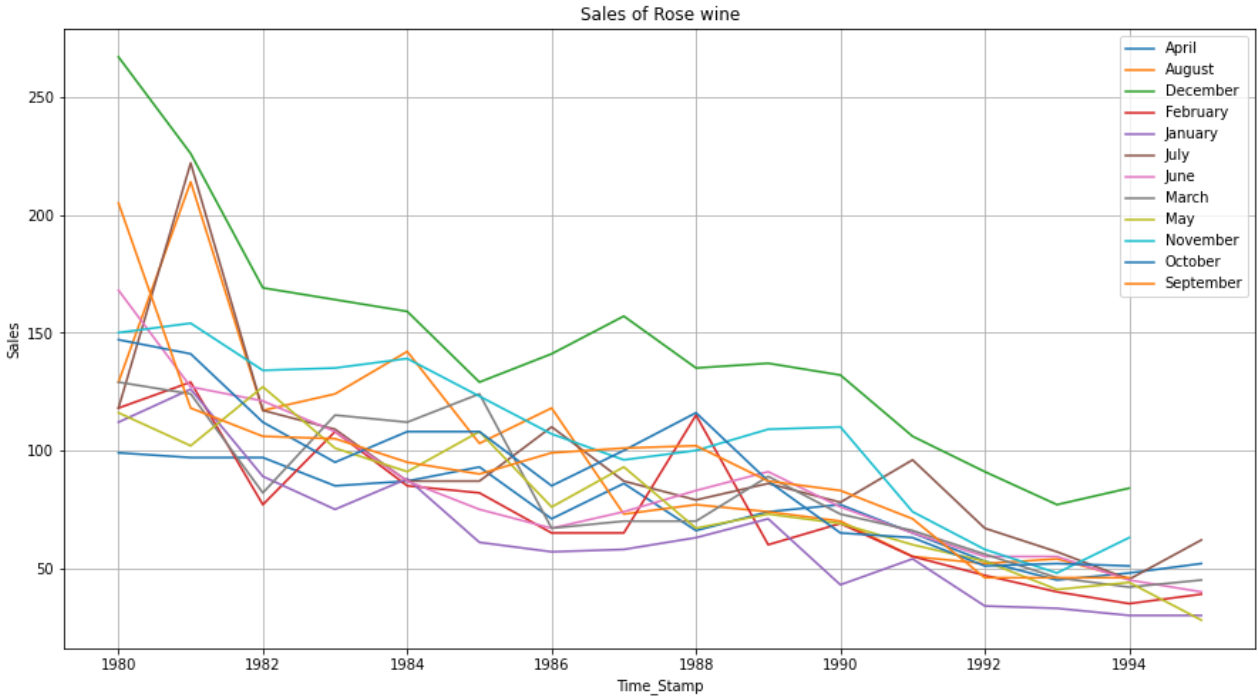
Time series Month plot for sales across years and month:

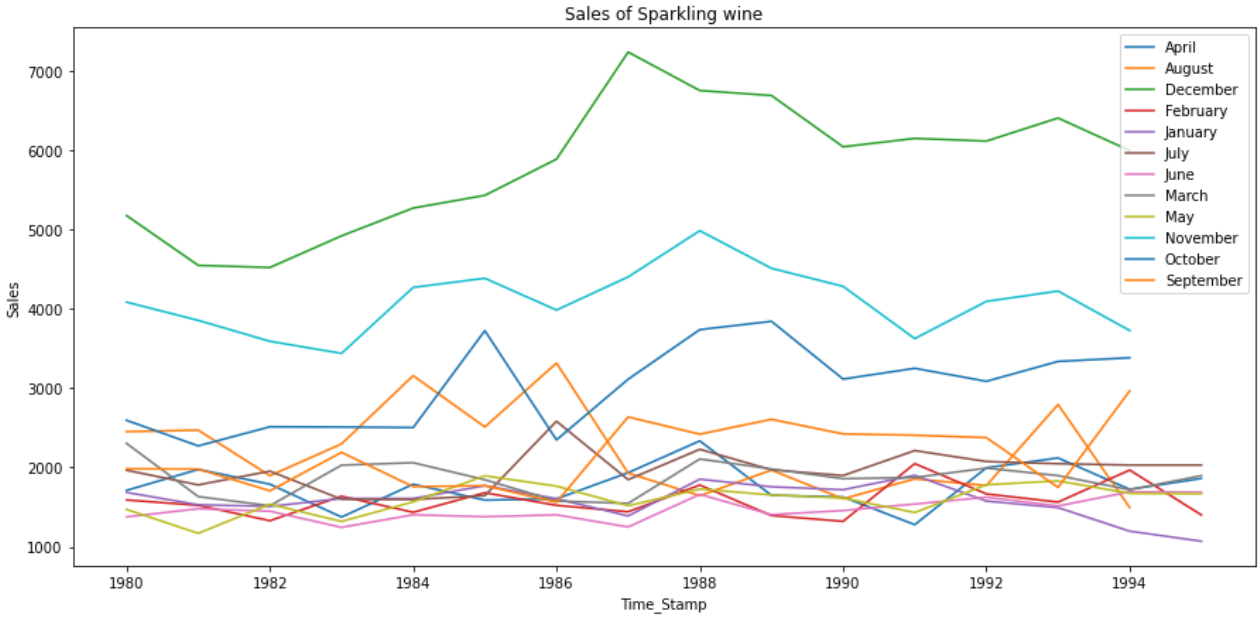






Cummulative-Month plot for sales across years and month:

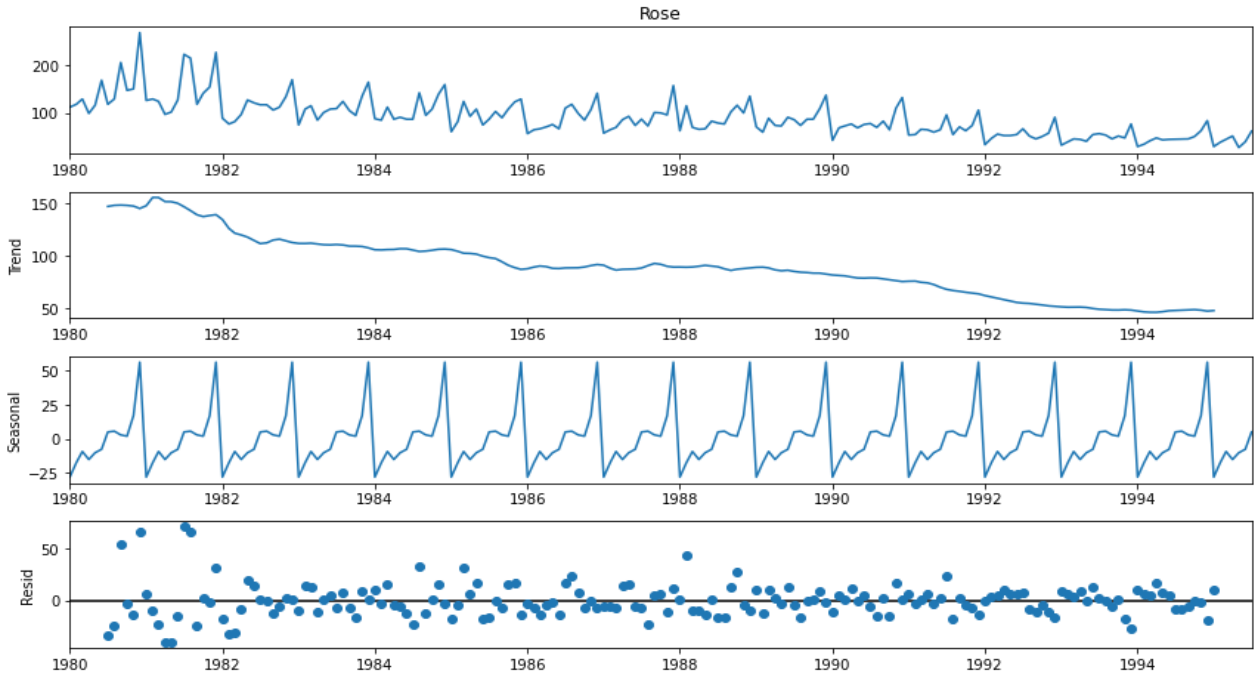




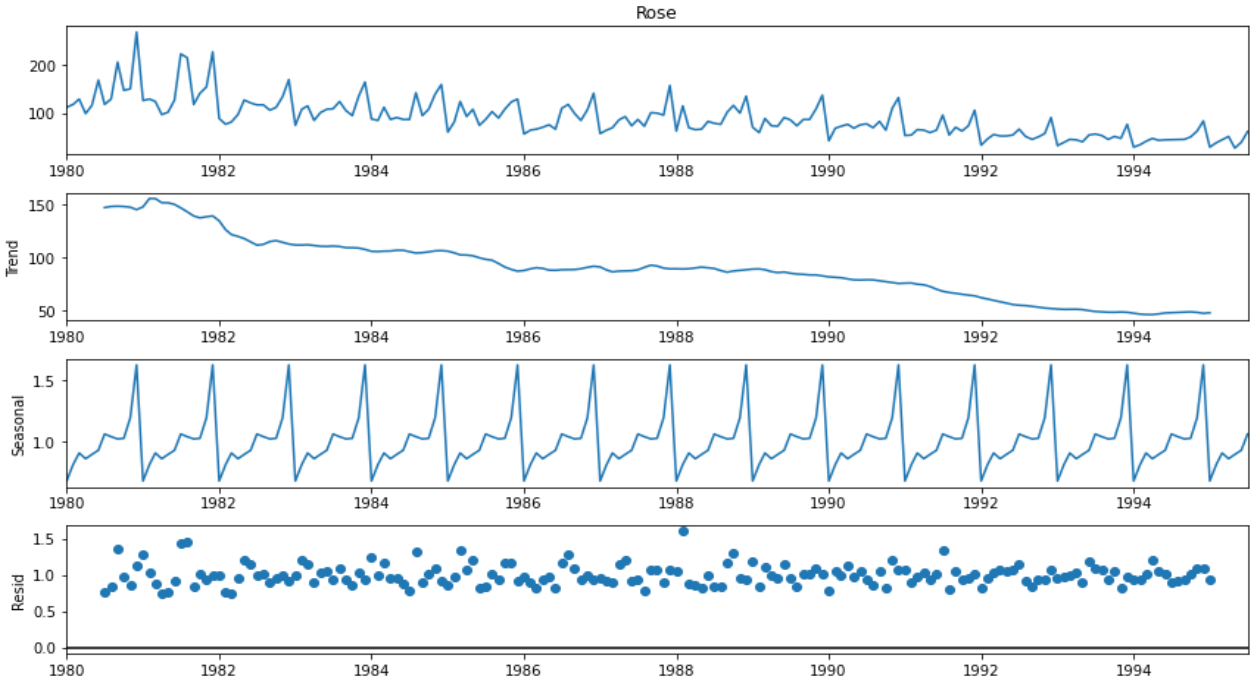
Inference from the Years and Month Time series plot:

- The median values keep increasing from January to December months in rose wine sales dataset.
- The Average Sales value also shows a decreasing trend in rose wine dataset.
- The median values are stable from January to June and has an increasing trend from July to December months in sparkling wine sales dataset.
- The Average Sales value does not show a trend in sparkling wine dataset.

Decomposition of Rose wine dataset time series plot: (Additive)



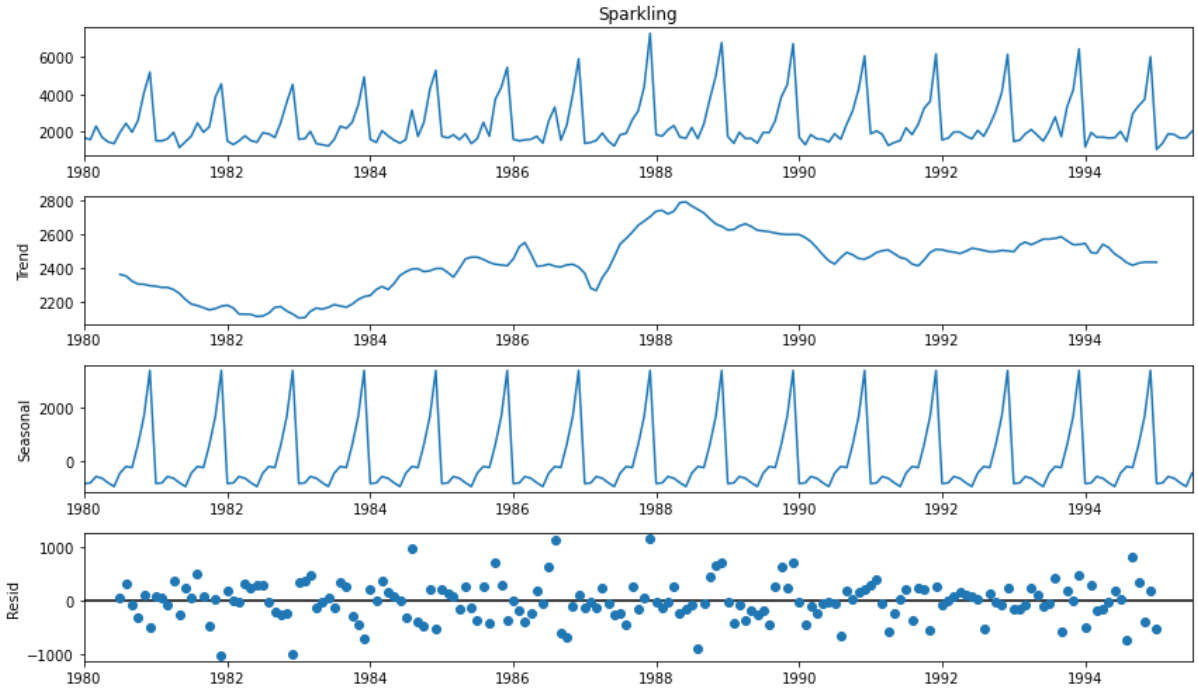
Decomposition of Rose wine dataset time series plot: (Multiplicative)



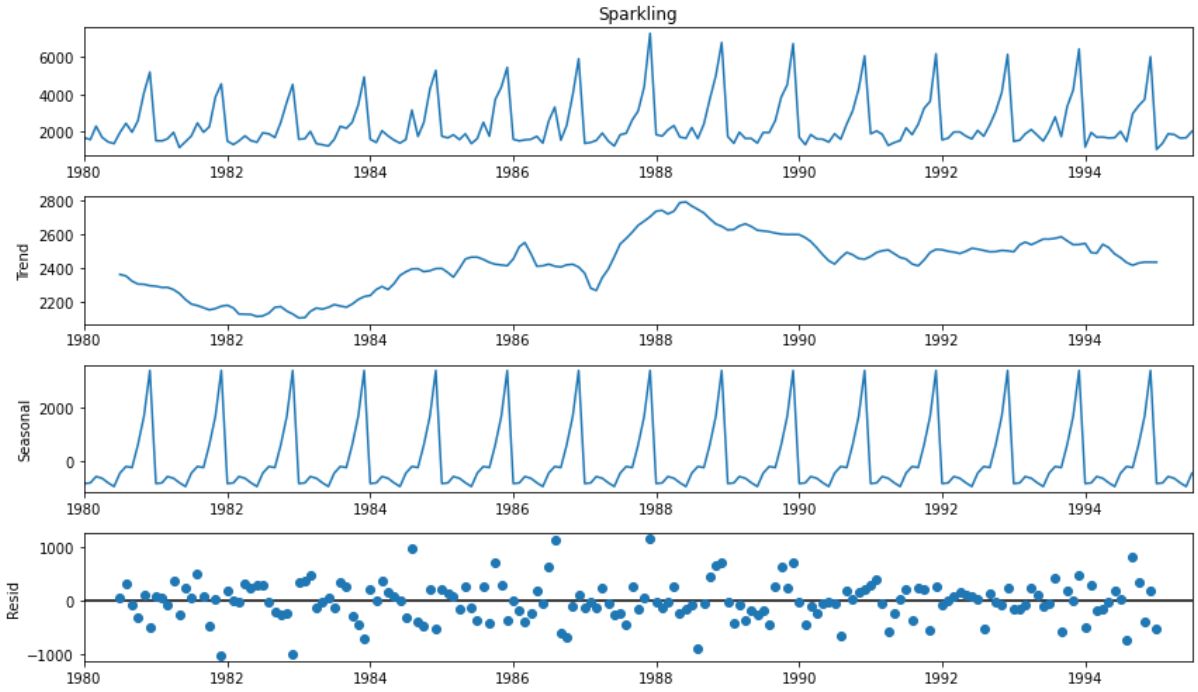
Inference from rose wine series decomposition plot:

- For additive we see the residual values are around 0 and for Multiplicative model we see the residual are around 1
- We cannot see any definite patterns in the residuals.
- The Rose wine time series is a multiplicative series.

Decomposition of Sparkling wine dataset time series plot: (Additive)



Decomposition of Rose wine dataset time series plot: (Multiplicative)



Inference from rose wine series decomposition plot:

- For additive we see the residual values are around 0 and for Multiplicative model we see the residual are around 1
- We cannot see any definite patterns in the residuals.
- The Sparkling wine time series is an Additive series.

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

Training Data is till the end of 1990. Test Data is from the beginning of 1991 to the last time stamp provided.

```
train_rose=df_rose[df_rose.index.year < 1991]
test_rose=df_rose[df_rose.index.year >= 1991]
train_spark=df_spark[df_spark.index.year < 1991]
test_spark=df_spark[df_spark.index.year >= 1991]
```

First few rows of Training Data

Time_Stamp	Rose
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0

Last few rows of Training Data

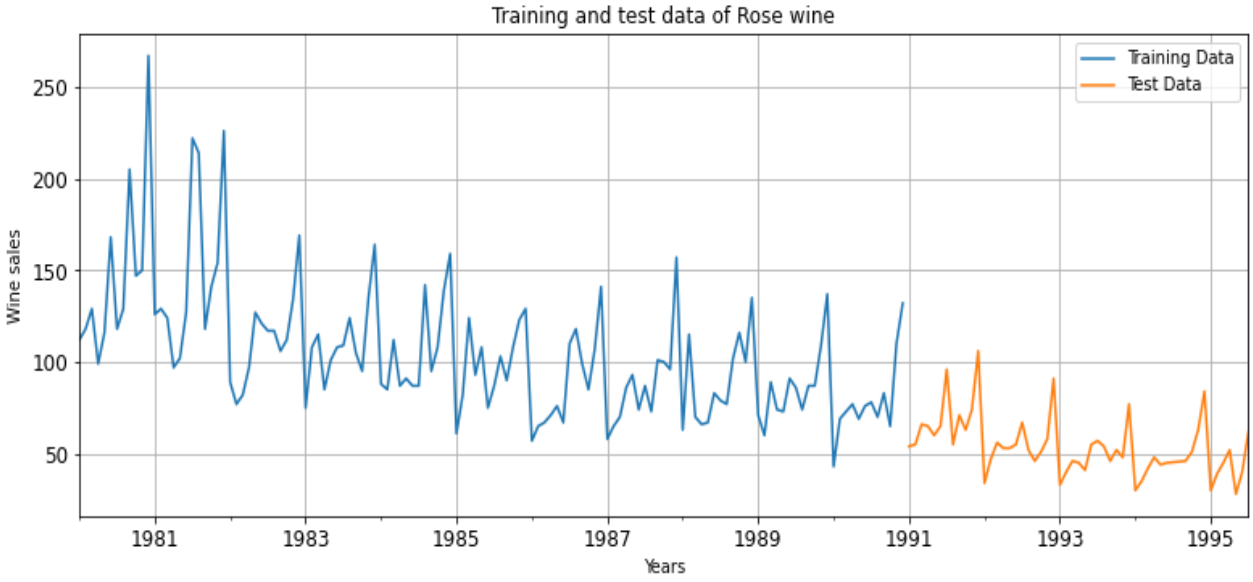
Time_Stamp	Rose
1990-08-01	70.0
1990-09-01	83.0
1990-10-01	65.0
1990-11-01	110.0
1990-12-01	132.0

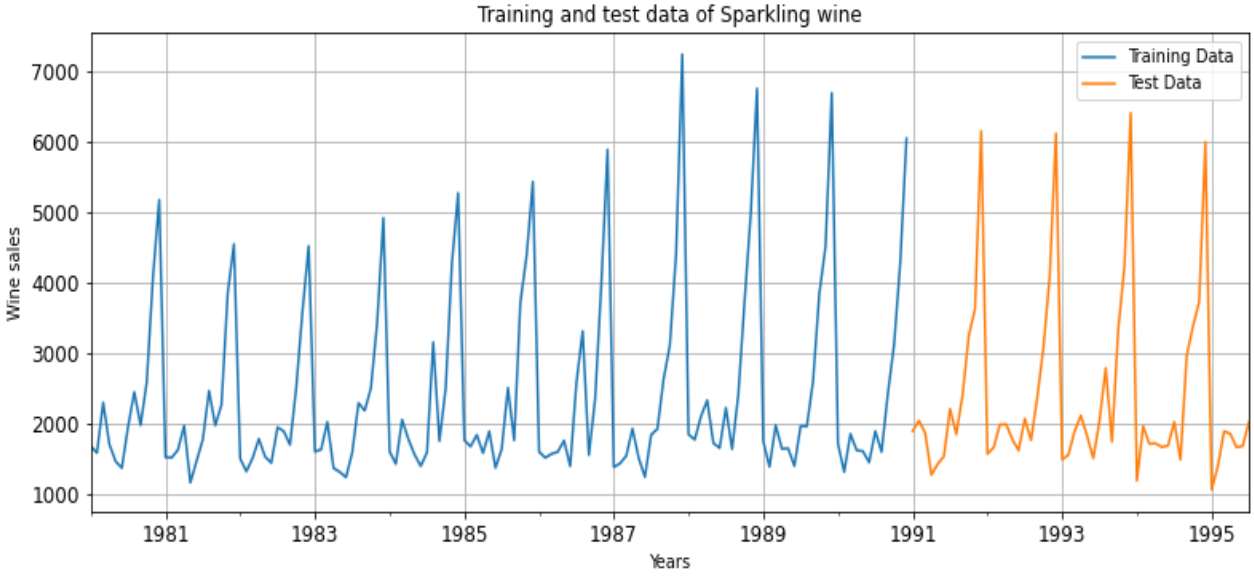
First few rows of Test Data

Time_Stamp	Rose
1991-01-01	54.0
1991-02-01	55.0
1991-03-01	66.0
1991-04-01	65.0
1991-05-01	60.0

Last few rows of Test Data

Time_Stamp	Rose
1995-03-01	45.0
1995-04-01	52.0
1995-05-01	28.0
1995-06-01	40.0
1995-07-01	62.0





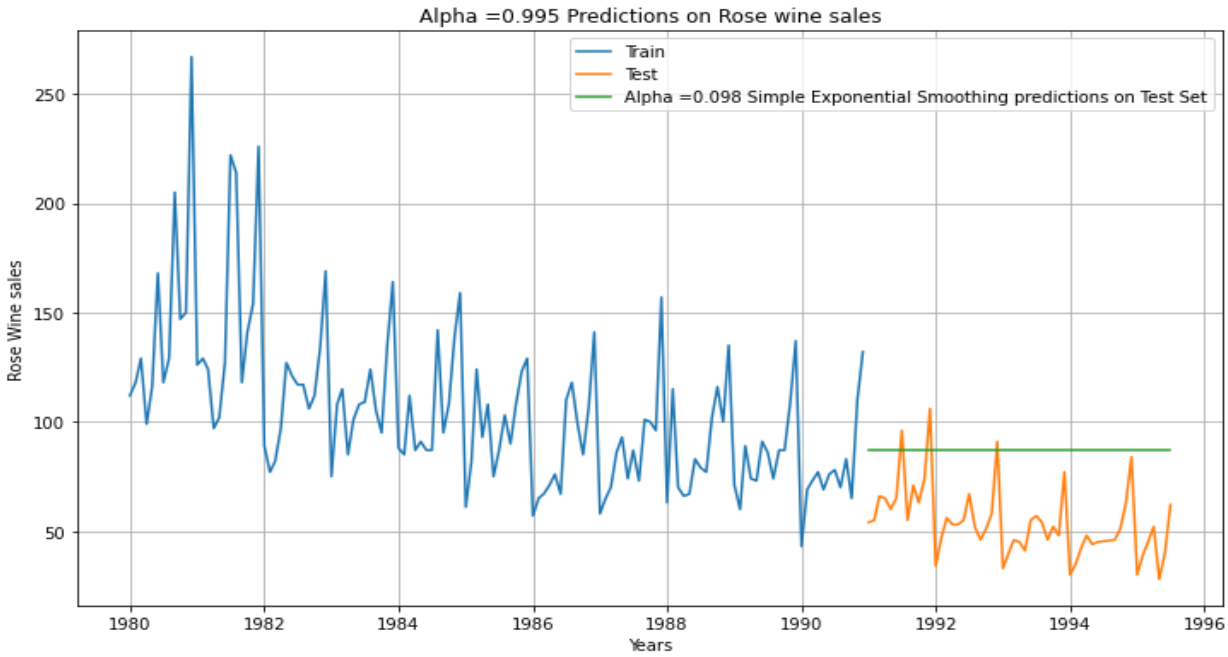
4. Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression, naïve forecast models, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE.

Method 1.1: Simple Exponential Smoothing for Rose wine Data

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. Parameter 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.

Note: Here, there is both trend and seasonality in the data. So, we should have directly gone for the Triple Exponential Smoothing but Simple Exponential Smoothing and the Double Exponential Smoothing models are built over here to get an idea of how the three types of models compare in this case.

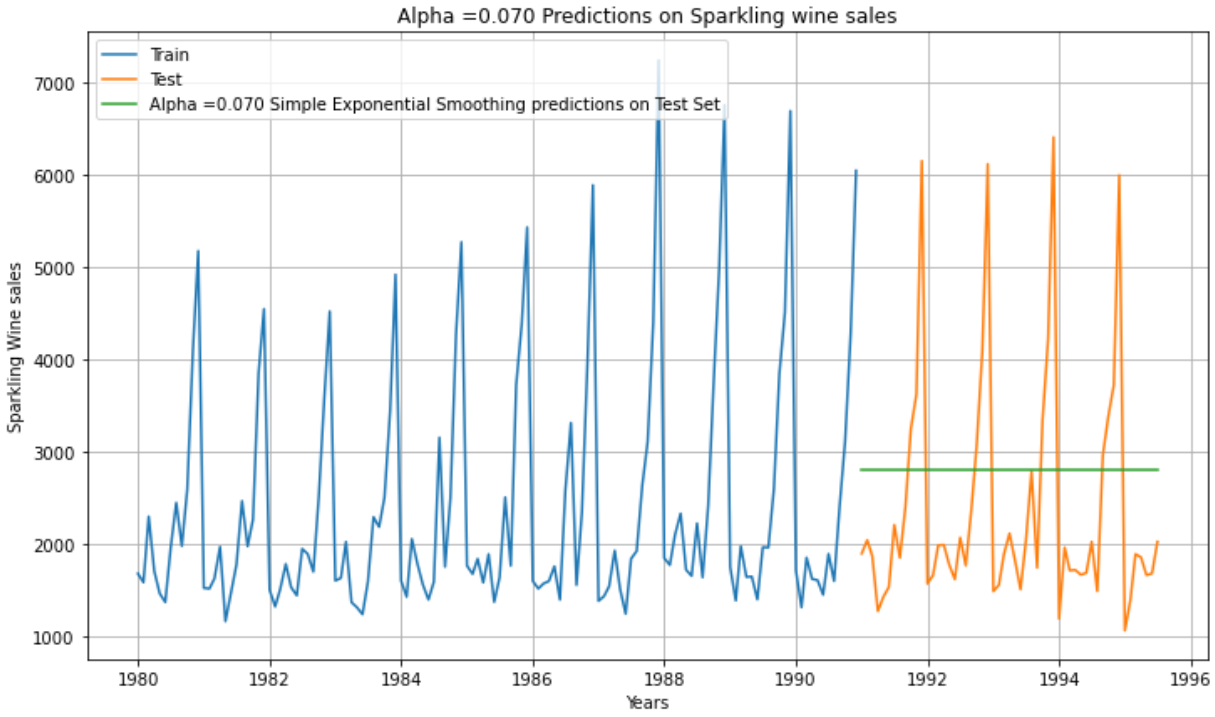
```
{'smoothing_level': 0.09874983698117956,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 134.38702481818487,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```



	Test RMSE	Test MAPE
Alpha=0.098,SimpleExponentialSmoothing	36.796241	63.88

Method 1.2: Simple Exponential Smoothing for Sparkling wine Data

```
{'smoothing_level': 0.07029120765764557,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 1764.0137060346985,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```



	Test RMSE	Test MAPE
Alpha=0.070,SimpleExponentialSmoothing	1338.008384	47.11

DOUBLE EXPONENTIAL SMOOTHING:

- One of the drawbacks of the simple exponential smoothing is that the model does not do perform 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, L_tLt is given by: $L_t=\alpha Y_t+(1-\alpha)\ F_t$Lt=αY t+(1−α) Ft
- Trend equation is given by $T_t=\beta(L_t-L_{t-1})+(1-\beta)T_{t-1}$Tt=β(Lt−Lt−1)+(1−β)Tt−1

Here, αα and ββ are the smoothing constants for level and trend, respectively,

- 0 <αα < 1 and 0 < ββ < 1.

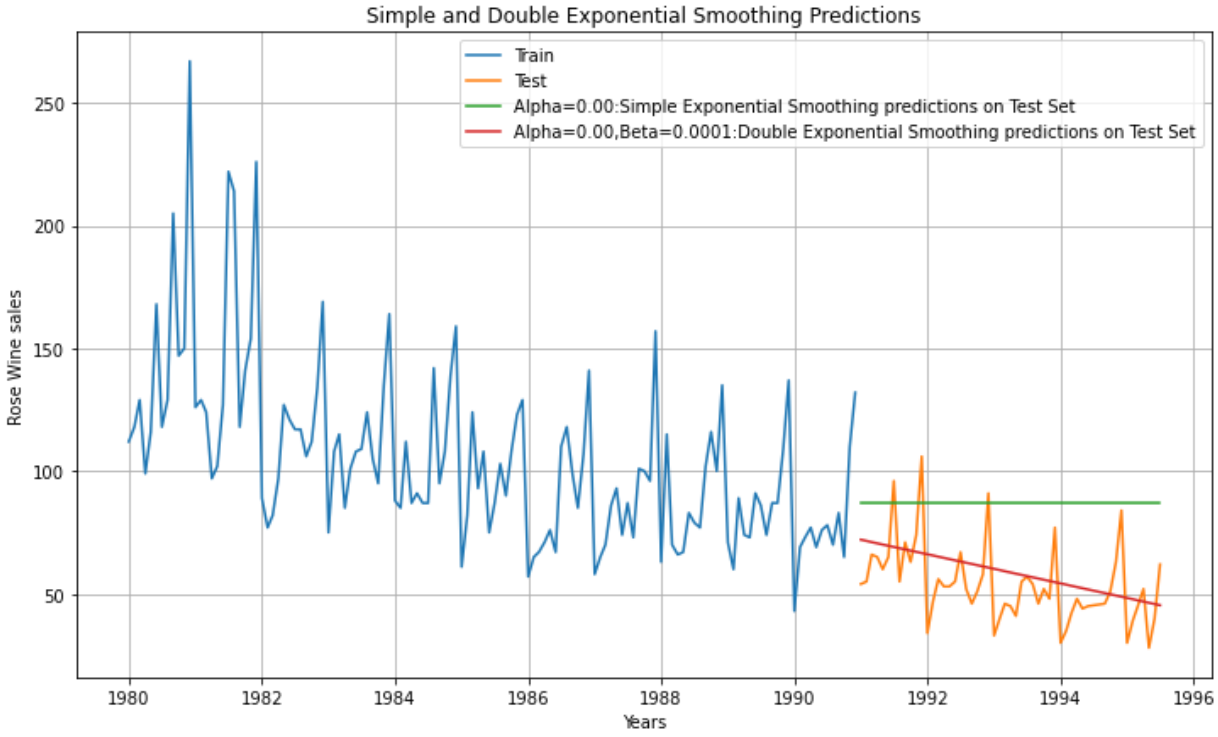
The forecast at time t + 1 is given by

- $F_{t+1}=L_t+T_t$Ft+1=Lt + Tt
- $F_{t+n}=L_t+nT_t$

Method 2.1: Double Exponential Smoothing for Rose wine Data

```
{'smoothing_level': 1.4901161193847656e-08,
 'smoothing_trend': 1.6610391146660035e-10,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 137.81553690867275,
 'initial_trend': -0.4943781897068274,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Time_Stamp	Rose	predict
1991-01-01	54.0	72.063238
1991-02-01	55.0	71.568859
1991-03-01	66.0	71.074481
1991-04-01	65.0	70.580103
1991-05-01	60.0	70.085725

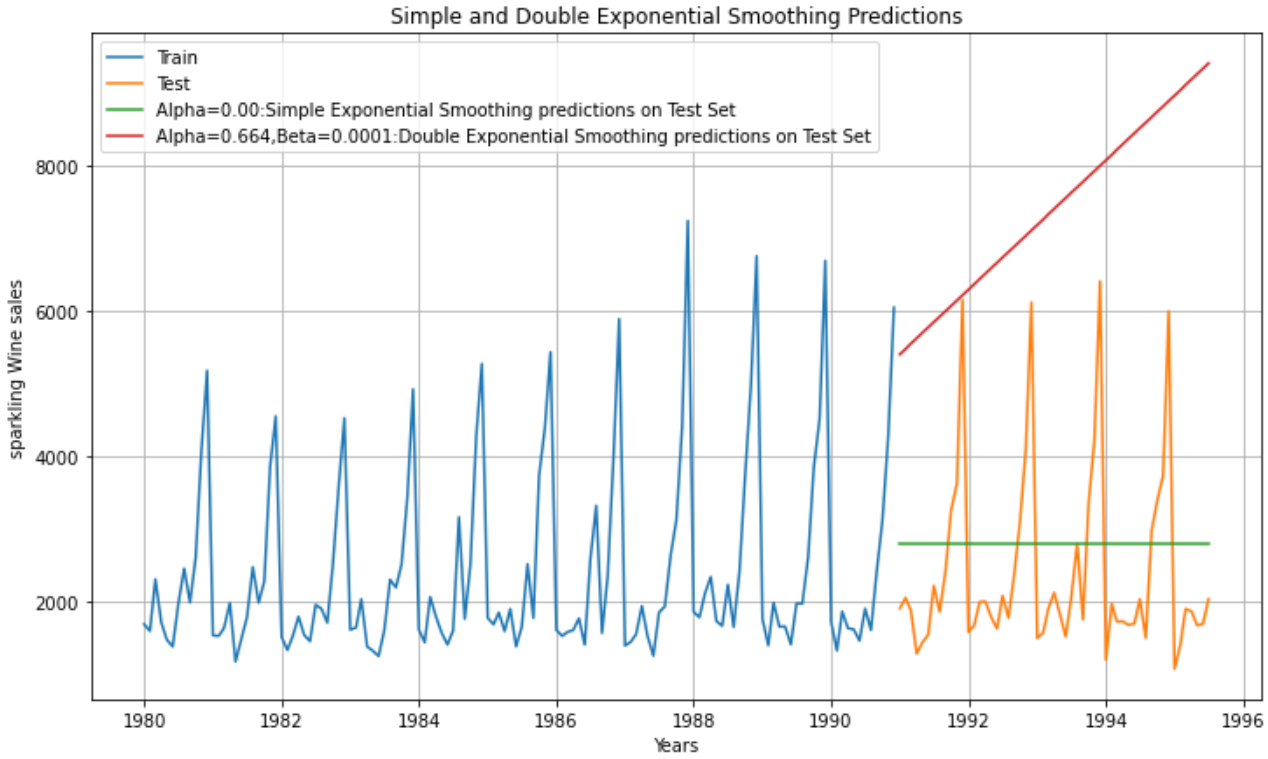


	Test RMSE	Test MAPE
Alpha=0.098,SimpleExponentialSmoothing	36.796241	63.88
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	36.879551	63.70

Method 2.2: Double Exponential Smoothing for Sparkling wine Data

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

Time_Stamp	Sparkling	predict
1991-01-01	1902	5401.733026
1991-02-01	2049	5476.005230
1991-03-01	1874	5550.277433
1991-04-01	1279	5624.549637
1991-05-01	1432	5698.821840



	Test RMSE	Test MAPE
Alpha=0.070,SimpleExponentialSmoothing	1338.008384	47.11
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	1777.734773	67.16

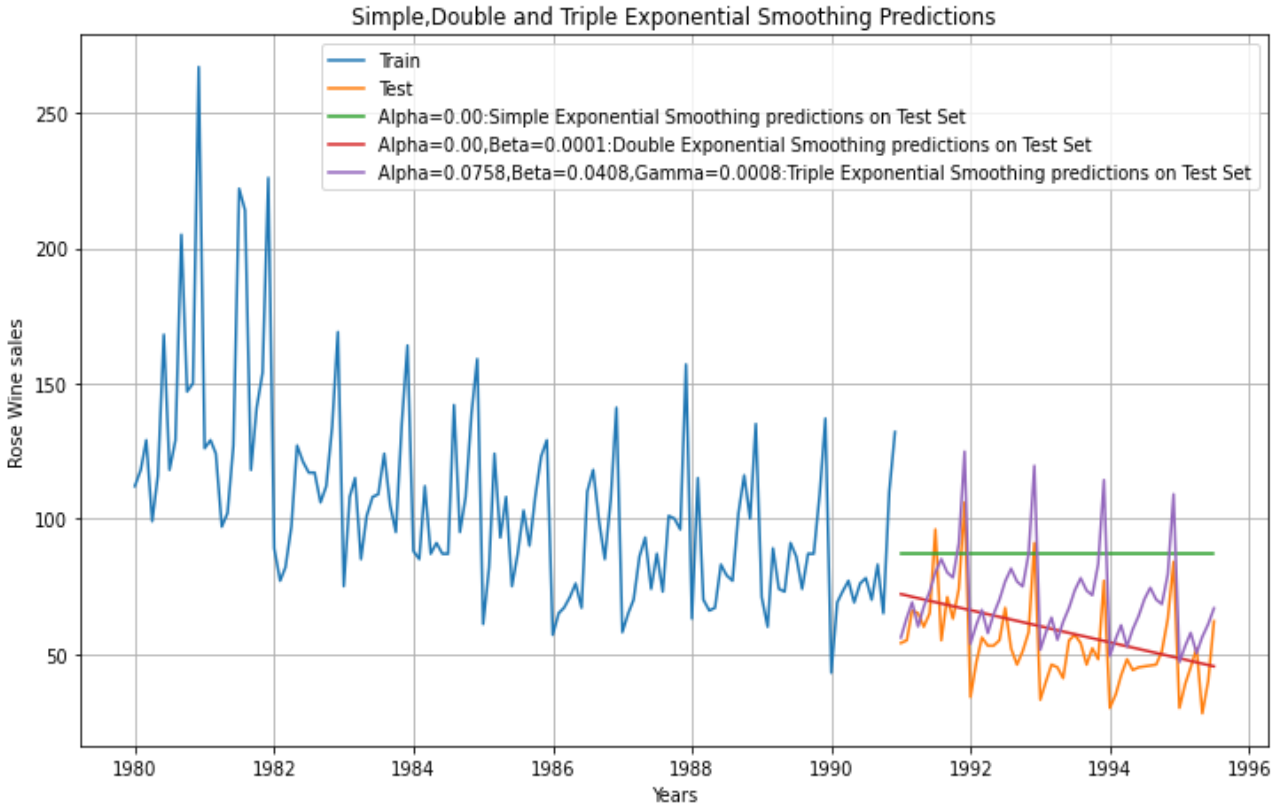
Holt-Winters - ETS(A, A, M) - Holt Winter's linear method

The three aspects of the time series behavior—value, trend, and seasonality—are expressed as three types of exponential smoothing, so Holt-Winters is called triple exponential smoothing.

The model predicts a current or future value by computing the combined effects of these three influences. The model requires several parameters: one for each smoothing (α, β, γ), the length of a season, and the number of periods in a season.

Method 3.1: Triple Exponential Smoothing for Rose wine Dataset:

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

{'smoothing_level': 0.07580378115501289, 'smoothing_trend': 0.04082731831671567, 'smoothing_seasonal': 0.0008792861232047841,
'damping_trend': nan, 'initial_level': 163.87796236599962, 'initial_trend': -0.9559811417358383, 'initial_seasons': array([0.68
432572, 0.77587329, 0.84828062, 0.74119702, 0.83386517,
       0.90761668, 0.99838676, 1.06374484, 1.00486364, 0.9847888 ,
       1.14803087, 1.58276201]), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}
```



Time_Stamp	Rose	predict
1991-01-01	54.0	56.036899
1991-02-01	55.0	63.349325
1991-03-01	66.0	69.021015
1991-04-01	65.0	60.120355
1991-05-01	60.0	67.380193

	Alpha Values	Beta Values	Gamma Values	Test RMSE	Test MAPE
233	0.3	0.4	0.4	1.037178e+01	13.38
151	0.2	0.6	0.2	9.565988e+00	13.41
10	0.1	0.2	0.1	9.223504e+00	13.42
223	0.3	0.3	0.4	1.016402e+01	13.58
11	0.1	0.2	0.2	9.496152e+00	13.70

	Test RMSE	Test MAPE
Alpha=0.098,SimpleExponentialSmoothing	36.796241	63.88
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	36.879551	63.70
Alpha=0.3,Beta=0.4,Gamma=0.4:TripleExponentialSmoothing	11.827223	18.66

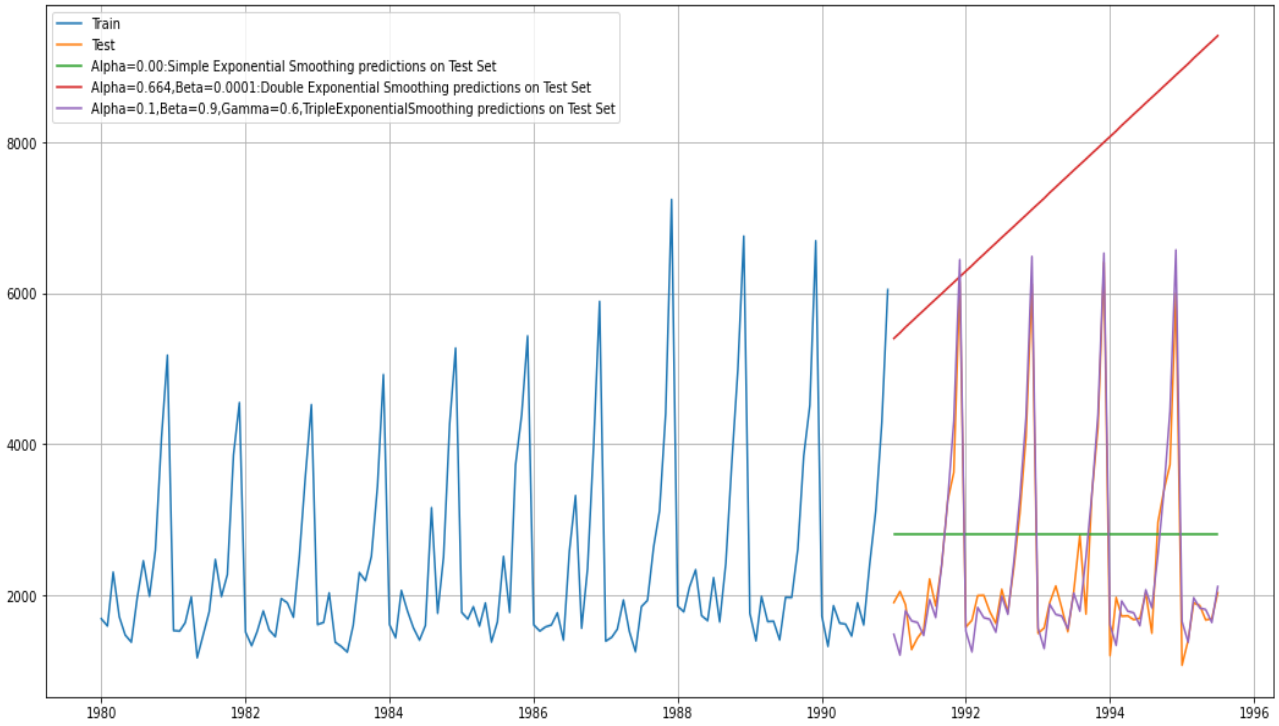
Method 3.2: Triple Exponential Smoothing for Sparkling wine Dataset:

==Holt Winters model Exponential Smoothing Estimated Parameters ==

{'smoothing_level': 0.11235974440805609, 'smoothing_trend': 0.03742154913668688, 'smoothing_seasonal': 0.4932616459048464, 'damping_trend': nan, 'initial_level': 1640.2806120050896, 'initial_trend': -3.261533670070838, 'initial_seasons': array([45.86595538, -48.96808341, 662.32406973, 73.10075169,
-168.81341007, -262.13208801, 326.10174942, 813.36401315,
344.51476989, 956.12012048, 2446.68553948, 3538.12189099]), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}

Time_Stamp	Sparkling	predict
1991-01-01	1902	1474.966680
1991-02-01	2049	1169.991432
1991-03-01	1874	1658.920133
1991-04-01	1279	1504.953983
1991-05-01	1432	1417.648032

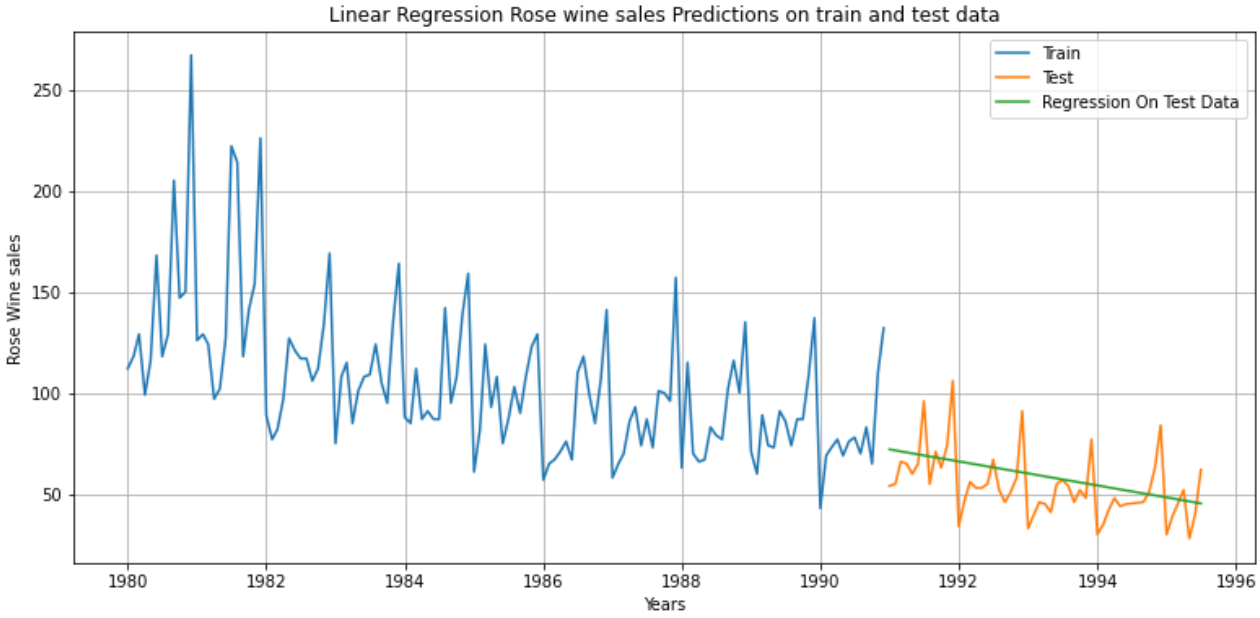
	Alpha Values	Beta Values	Gamma Values	Test RMSE	Test MAPE
85	0.1	0.9	0.6	3.384584e+02	9.93
139	0.2	0.4	1.0	3.388406e+02	10.59
110	0.2	0.2	0.1	3.437154e+02	11.06
30	0.1	0.4	0.1	3.463027e+02	11.50
200	0.3	0.1	0.1	3.723909e+02	11.73



	Test RMSE	Test MAPE
Alpha=0.070,SimpleExponentialSmoothing	1338.008384	47.11
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	1777.734773	67.16
Alpha=0.1,Beta=0.9,Gamma=0.6:TripleExponentialSmoothing	520.011735	18.27

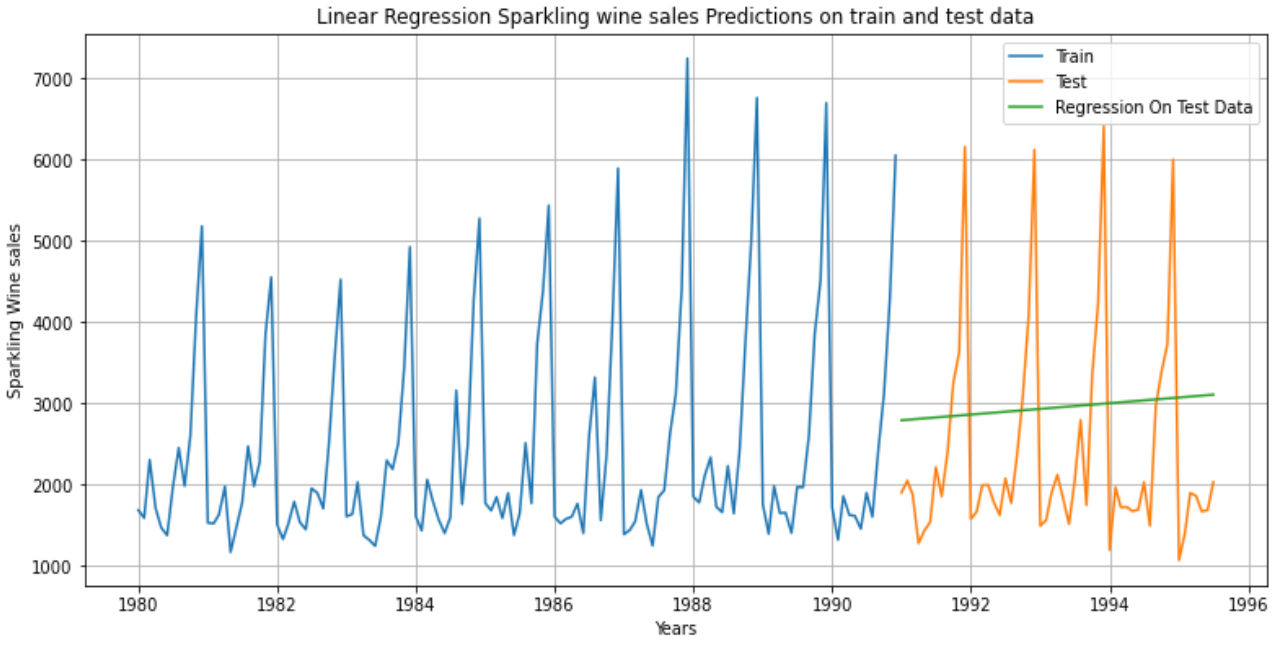
Method 4.1: Linear Regression for Rose wine Dataset:

Now that our training and test data has been modified, let us go ahead use $LinearRegression$ to build the model on the training data and test the model on the test data.



	Test RMSE	Test MAPE
Alpha=0.098,SimpleExponentialSmoothing	36.796241	63.88
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	36.879551	63.70
Alpha=0.3,Beta=0.4,Gamma=0.4:TripleExponentialSmoothing	11.827223	18.66
Linear Regression	15.268955	22.82

Method 4.2: Linear Regression for Sparkling wine Dataset:

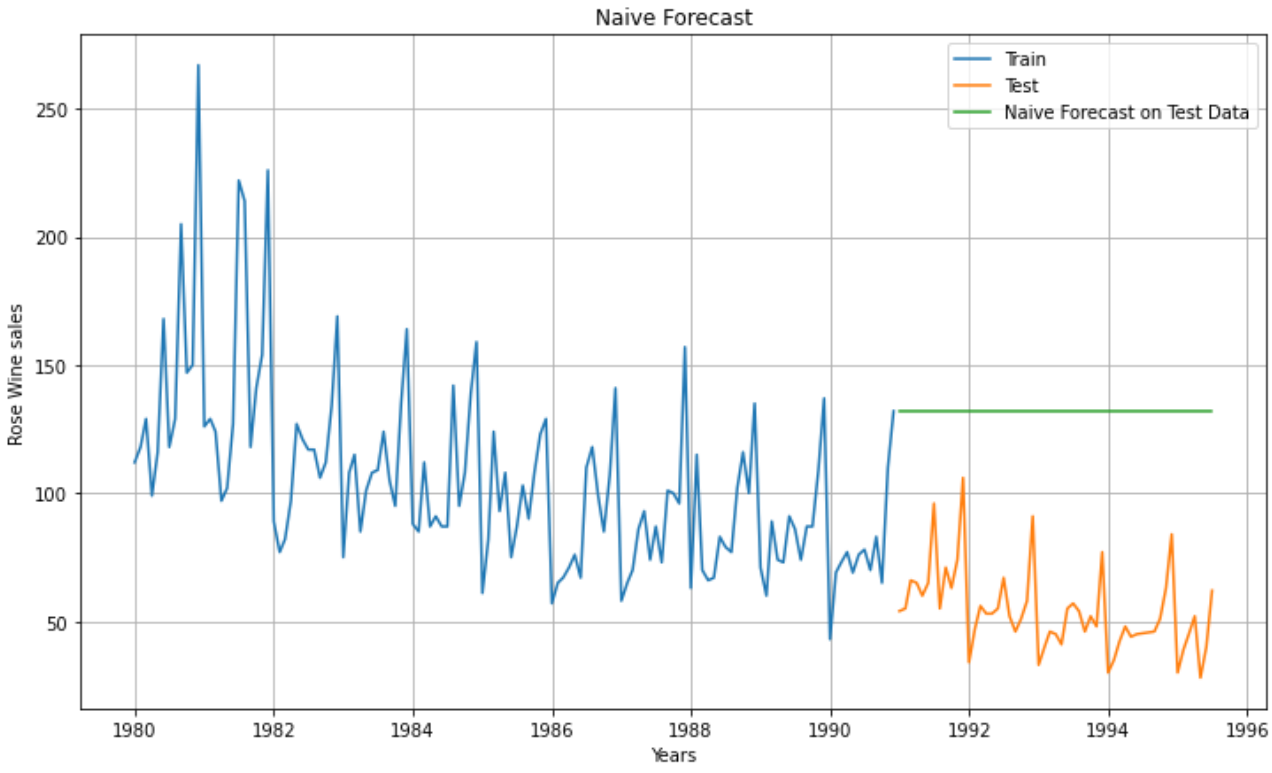


	Test RMSE	Test MAPE
Alpha=0.070,SimpleExponentialSmoothing	1338.008384	47.11
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	1777.734773	67.16
Alpha=0.1,Beta=0.9,Gamma=0.6:TripleExponentialSmoothing	520.011735	18.27
Linear Regression	1389.135175	50.15

Model : Naive Approach: $\hat{y}_{t+1}=y_t$ y^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.

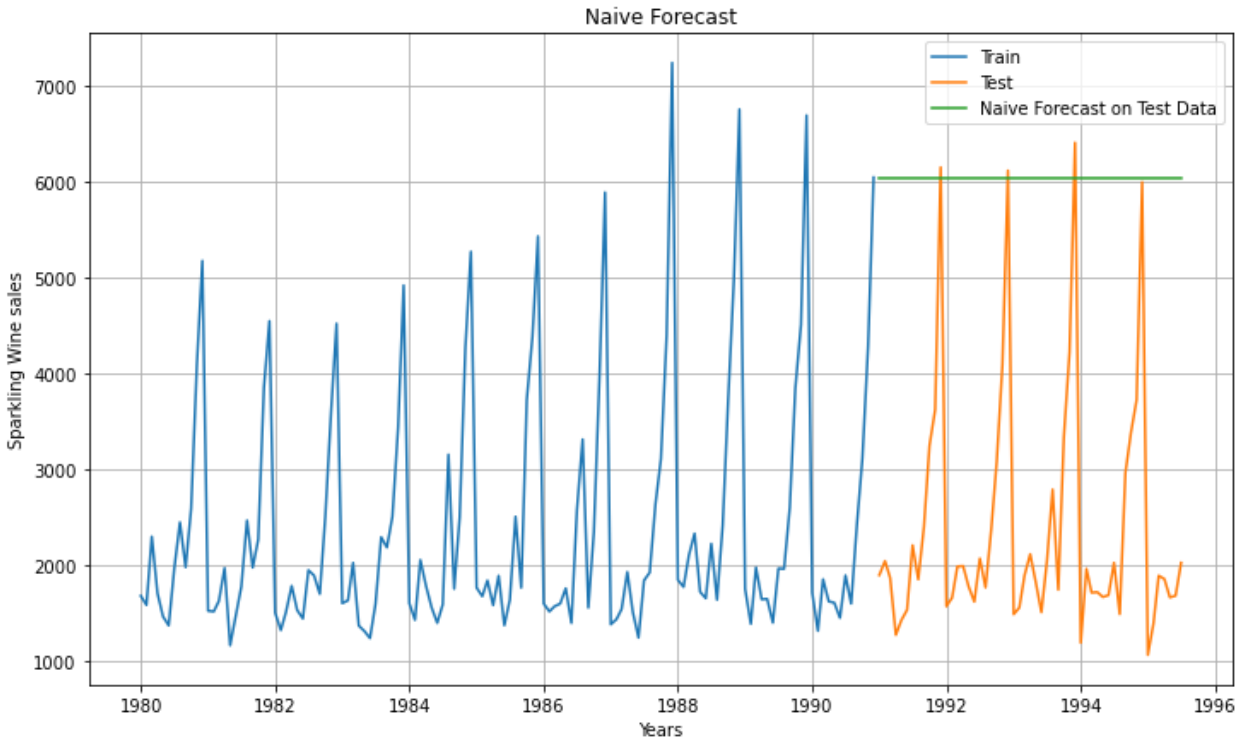
Method 5.1: Naïve model for Rose wine Dataset:



Naive method forecast on the Test Data, RMSE is 79.719

	Test RMSE	Test MAPE
Alpha=0.098,SimpleExponentialSmoothing	36.796241	63.88
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	36.879551	63.70
Alpha=0.3,Beta=0.4,Gamma=0.4:TripleExponentialSmoothing	11.827223	18.66
Linear Regression	15.268955	22.82
NaiveModel	79.718773	145.10

Method 5.2: Naïve model for Sparkling wine Dataset:



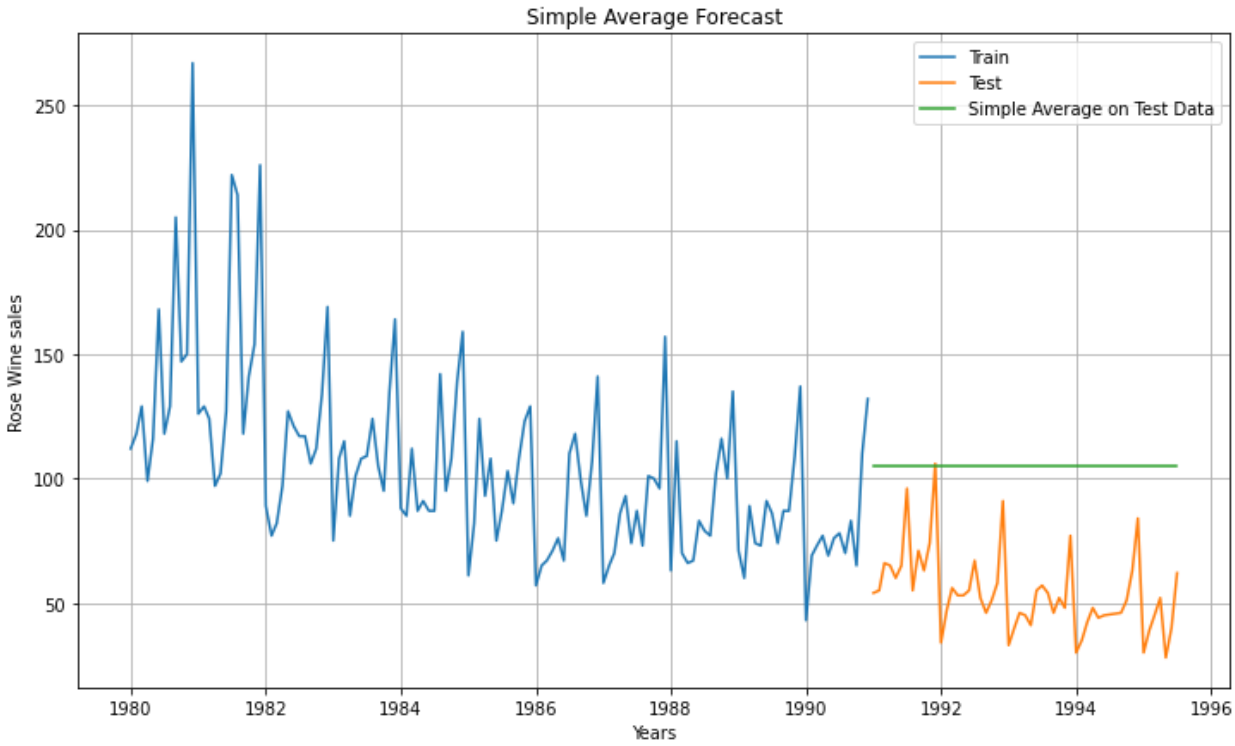

```
For Naive forecast on the Test Data,  RMSE is 3864.279
```

	Test RMSE	Test MAPE
Alpha=0.070,SimpleExponentialSmoothing	1338.008384	47.11
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	1777.734773	67.16
Alpha=0.1,Beta=0.9,Gamma=0.6:TripleExponentialSmoothing	520.011735	18.27
Linear Regression	1389.135175	50.15
NaiveModel	3864.279352	152.87

Method 6.1: Simple Average model for Rose wine Dataset:

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

	Rose	mean_forecast
Time_Stamp		
1991-01-01	54.0	104.939394
1991-02-01	55.0	104.939394
1991-03-01	66.0	104.939394
1991-04-01	65.0	104.939394
1991-05-01	60.0	104.939394



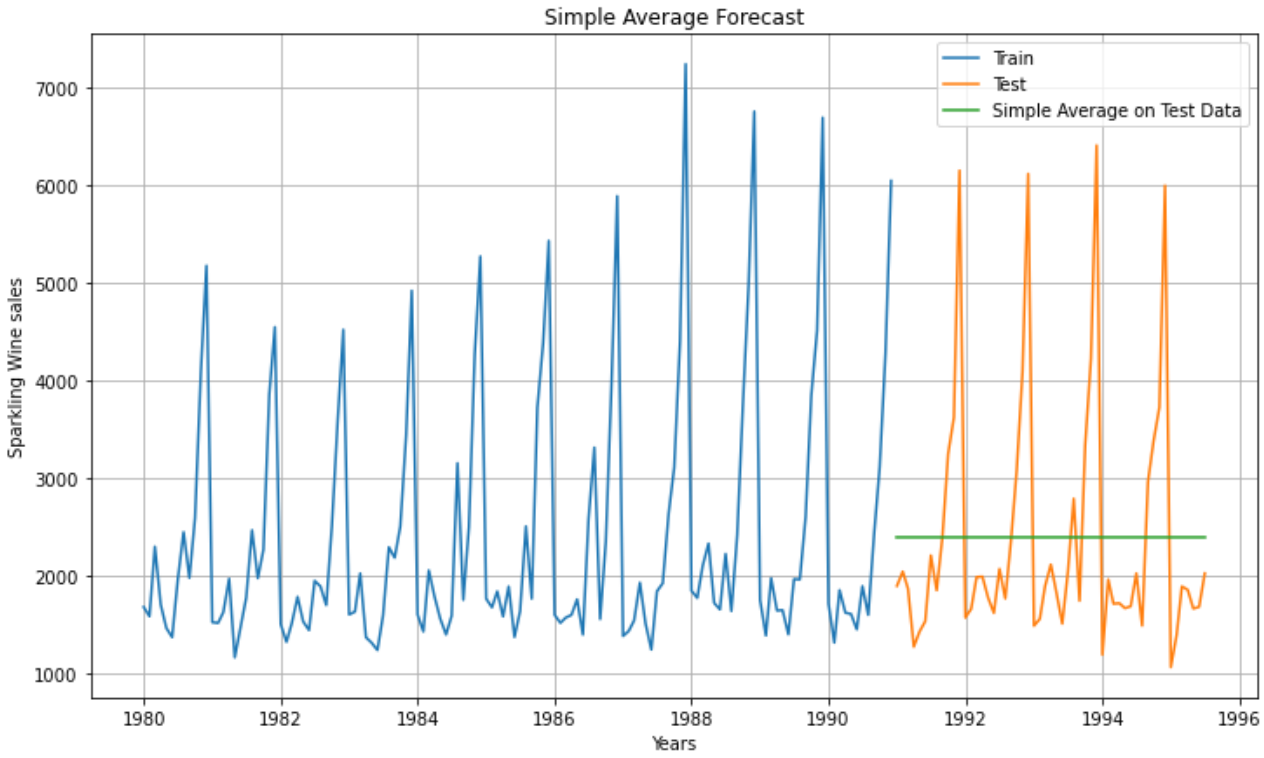

```
For Simple Average forecast on the Test Data,  RMSE is 53.461
For Simple Average forecast on the Test Data,  MAPE is 94.930
```

	Test RMSE	Test MAPE
Alpha=0.098,SimpleExponentialSmoothing	36.796241	63.88
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	36.879551	63.70
Alpha=0.3,Beta=0.4,Gamma=0.4:TripleExponentialSmoothing	11.827223	18.66
Linear Regression	15.268955	22.82
NaiveModel	79.718773	145.10
SimpleAverageModel	53.460570	94.93

Method 6.2: Simple Average model for Sparkling wine Dataset:

```
1 sa_test_spark['mean_forecast'] = train_spark['Sparkling'].mean()
2 sa_test_spark.head()
```

Time_Stamp	Sparkling	mean_forecast
1991-01-01	1902	2403.780303
1991-02-01	2049	2403.780303
1991-03-01	1874	2403.780303
1991-04-01	1279	2403.780303
1991-05-01	1432	2403.780303




```
For Simple Average forecast on the Test Data,  RMSE is 1275.082
For Simple Average forecast on the Test Data,  MAPE is 38.900
```

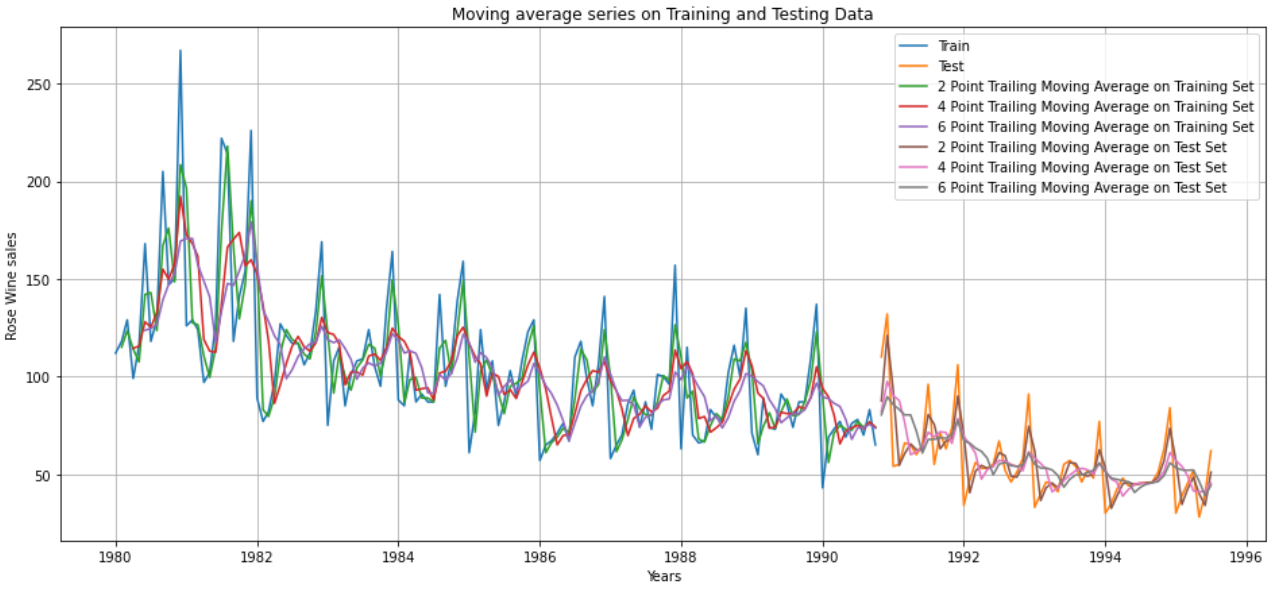
	Test RMSE	Test MAPE
Alpha=0.070,SimpleExponentialSmoothing	1338.008384	47.11
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	1777.734773	67.16
Alpha=0.1,Beta=0.9,Gamma=0.6:TripleExponentialSmoothing	520.011735	18.27
Linear Regression	1389.135175	50.15
NaiveModel	3864.279352	152.87
SimpleAverageModel	1275.081804	38.90

Method 7.1: Moving Average model for Rose wine Dataset:

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 maximum accuracy (or the minimum error) over here, we are going to average over the entire data.

Time_Stamp	Rose	Trailing_2	Trailing_4	Trailing_6
1980-01-01	112.0	NaN	NaN	NaN
1980-02-01	118.0	115.0	NaN	NaN
1980-03-01	129.0	123.5	NaN	NaN
1980-04-01	99.0	114.0	114.5	NaN
1980-05-01	116.0	107.5	115.5	NaN

Time_Stamp	Rose	Trailing_2	Trailing_4	Trailing_6
1990-11-01	110.0	87.5	82.00	80.333333
1990-12-01	132.0	121.0	97.50	89.666667
1991-01-01	54.0	93.0	90.25	85.666667
1991-02-01	55.0	54.5	87.75	83.166667
1991-03-01	66.0	60.5	76.75	80.333333



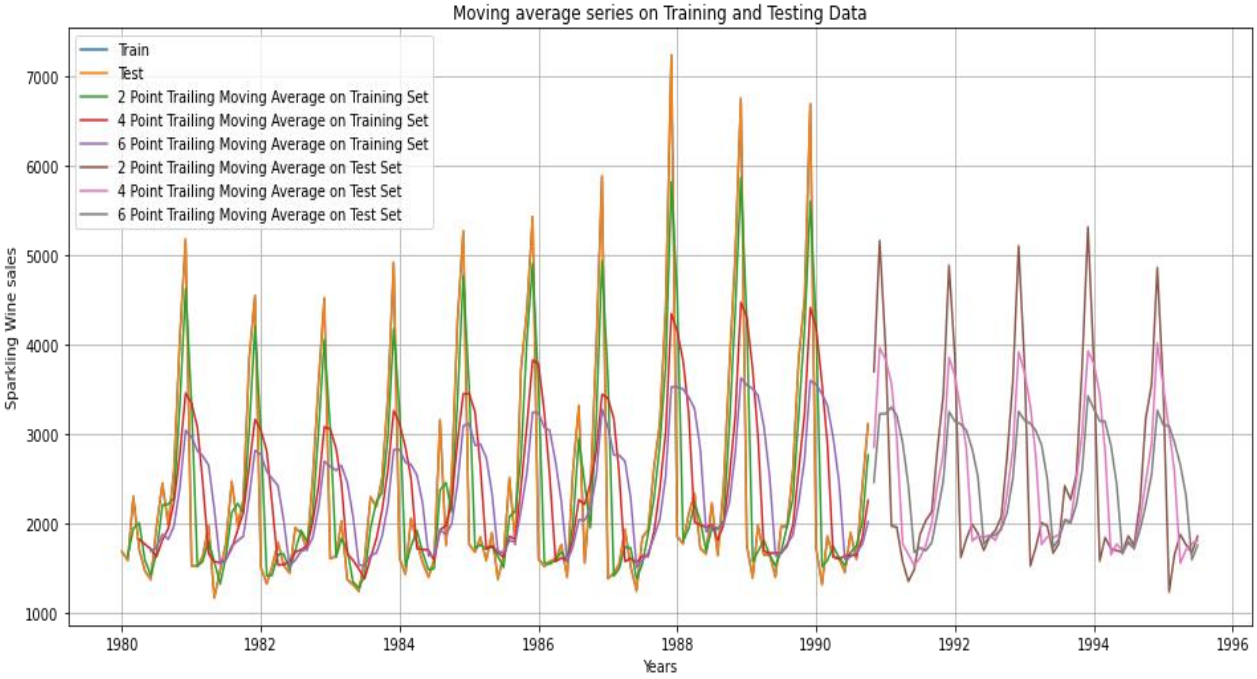

```
For 2 point Moving Average Model forecast on the Training Data,  RMSE is 11.801
For 4 point Moving Average Model forecast on the Training Data,  RMSE is 15.367
For 6 point Moving Average Model forecast on the Training Data,  RMSE is 15.862
```

	Test RMSE	Test MAPE
Alpha=0.098,SimpleExponentialSmoothing	36.796241	63.88
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	36.879551	63.70
Alpha=0.3,Beta=0.4,Gamma=0.4:TripleExponentialSmoothing	11.827223	18.66
Linear Regression	15.268955	22.82
NaiveModel	79.718773	145.10
SimpleAverageModel	53.460570	94.93
2pointTrailingMovingAverage	11.801043	13.56
4pointTrailingMovingAverage	15.367212	19.97
6pointTrailingMovingAverage	15.862350	21.49

Method 7.2: Moving Average model for Sparkling wine Dataset:

Time_Stamp	Sparkling	Trailing_2	Trailing_4	Trailing_6
1980-01-01	1686	NaN	NaN	NaN
1980-02-01	1591	1638.5	NaN	NaN
1980-03-01	2304	1947.5	NaN	NaN
1980-04-01	1712	2008.0	1823.25	NaN
1980-05-01	1471	1591.5	1769.50	NaN

Time_Stamp	Sparkling	Trailing_2	Trailing_4	Trailing_6
1990-11-01	4286	3701.0	2857.75	2464.500000
1990-12-01	6047	5166.5	3968.25	3229.500000
1991-01-01	1902	3974.5	3837.75	3230.000000
1991-02-01	2049	1975.5	3571.00	3304.000000
1991-03-01	1874	1961.5	2968.00	3212.333333




```
For 2 point Moving Average Model forecast on the Training Data,  RMSE is 811.179
For 4 point Moving Average Model forecast on the Training Data,  RMSE is 1184.213
For 6 point Moving Average Model forecast on the Training Data,  RMSE is 1337.201
```

	Test RMSE	Test MAPE
Alpha=0.070,SimpleExponentialSmoothing	1338.008384	47.11
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	1777.734773	67.16
Alpha=0.1,Beta=0.9,Gamma=0.6:TripleExponentialSmoothing	520.011735	18.27
Linear Regression	1389.135175	50.15
NaiveModel	3864.279352	152.87
SimpleAverageModel	1275.081804	38.90
2pointTrailingMovingAverage	811.178937	19.30
4pointTrailingMovingAverage	1184.213295	35.81
6pointTrailingMovingAverage	1184.213295	43.94

For this data, we had seasonality in the data, so by definition Triple Exponential Smoothing is supposed to work better than the Simple Exponential Smoothing as well as the Double Exponential Smoothing.

However, since this was a model building exercise, we had gone on to build different models on the data and have compared these models with the best RMSE value on the test data.

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.

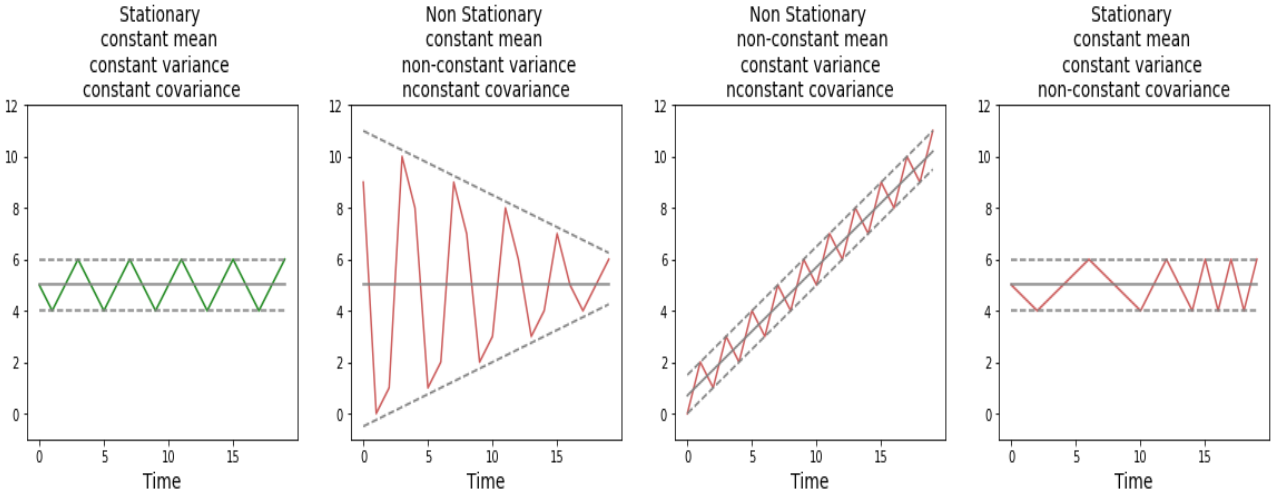
Stationarity

Some Time-series models, such as such as ARIMA, assume that the underlying data is stationary. Stationarity describes that the time-series has

- constant mean and mean are not time-dependent
- constant variance and variance are not time-dependent
- constant covariance and covariance are not time-dependent

If a time series has a specific (stationary) behaviour over a given time interval, then it can be assumed that the time series will behave the same at a later time.

Time series **with trend and/or seasonality are NON stationary**. Trend indicates that the mean is not constant over time and seasonality indicates that the variance is not constant over time.



The check for stationarity can be done via three different approaches:

1. **visually**: plot time series and check for trends or seasonality
2. **basic statistics**: split time series and compare the mean and variance of each partition
3. **statistical test**: Augmented Dickey Fuller test

Let's do the **visual check** first. We can see the features of variable Rose have non-constant mean and non-constant variance. Therefore, **none of these seem to be stationary**.

Statistical test to find the Stationarity

Augmented Dickey-Fuller (ADF) test is a type of statistical test called a unit root test. Unit roots are a cause for non-stationarity.

- **Null Hypothesis (H0)**: Time series has a unit root. (Time series is **not stationary**).
- **Alternate Hypothesis (H1)**: Time series has no unit root (Time series is **stationary**).

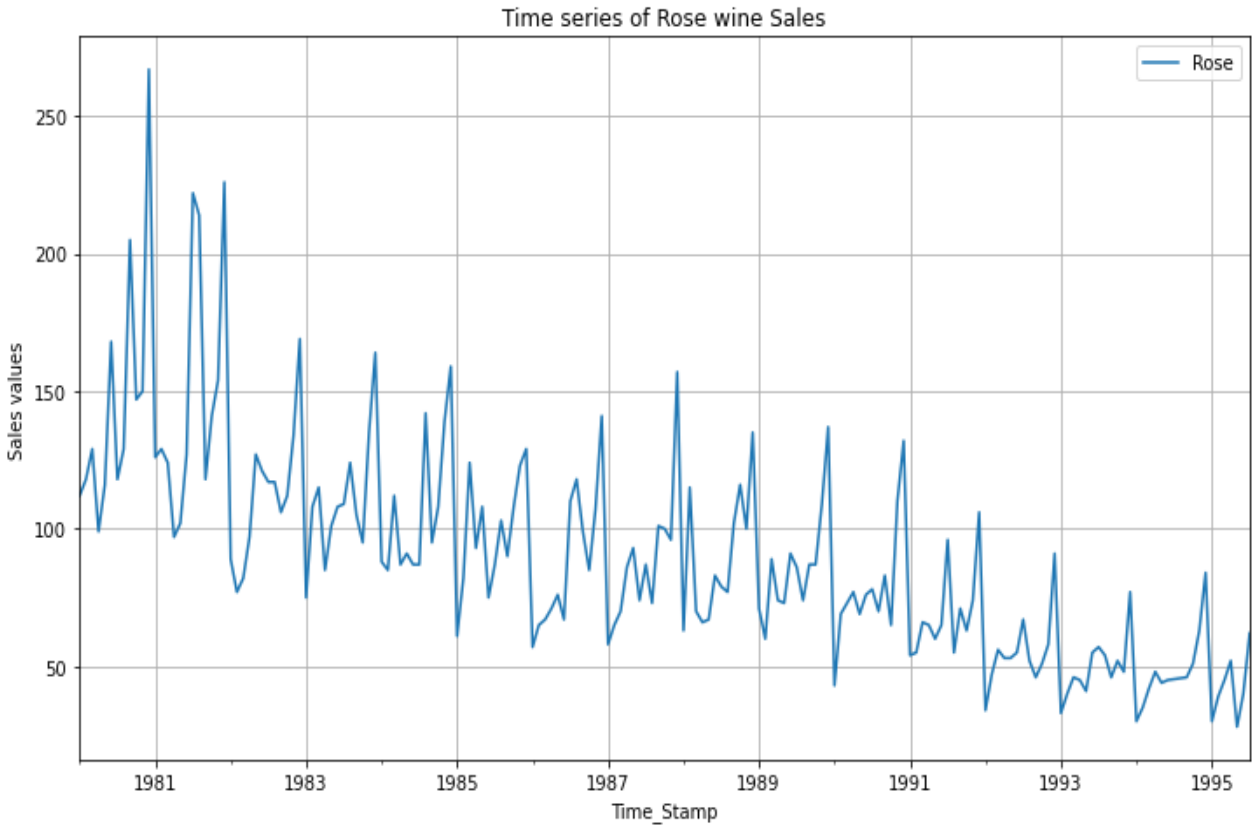
If the **null hypothesis can be rejected**, we can conclude that the **time series is stationary**.

There are two ways to rejects the null hypothesis:

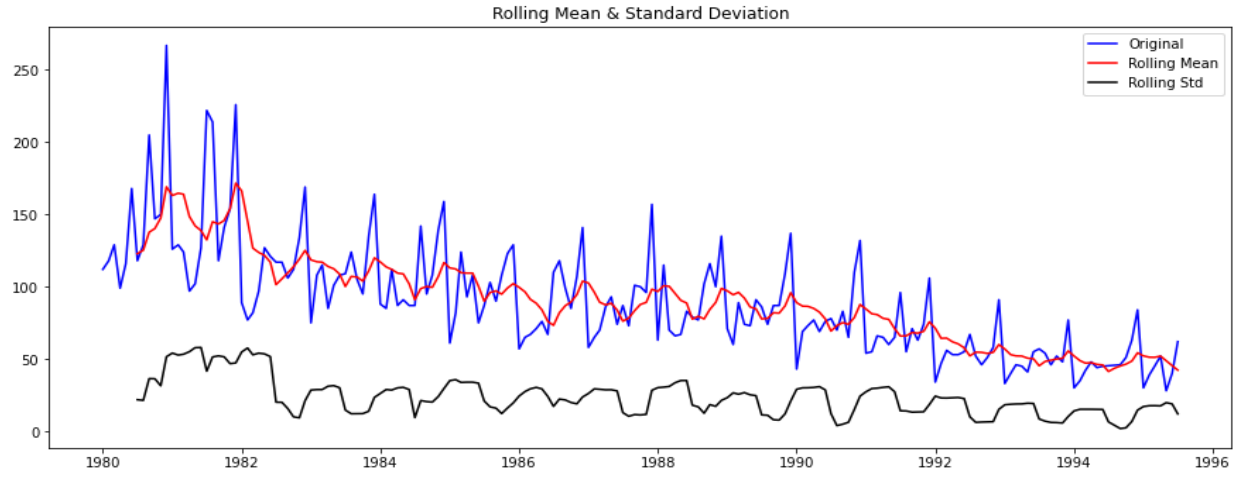
On the one hand, the null hypothesis can be rejected if the p-value is below a set significance level. The defaults significance level is 5%

- **p-value > significance level (default: 0.05)**: Fail to reject the null hypothesis (H0), the data has a unit root and is non-stationary.
- **p-value <= significance level (default: 0.05)**: Reject the null hypothesis (H0), the data does not have a unit root and is stationary.

Visual and Statistical Check to find Stationarity in Rose wine Dataset:

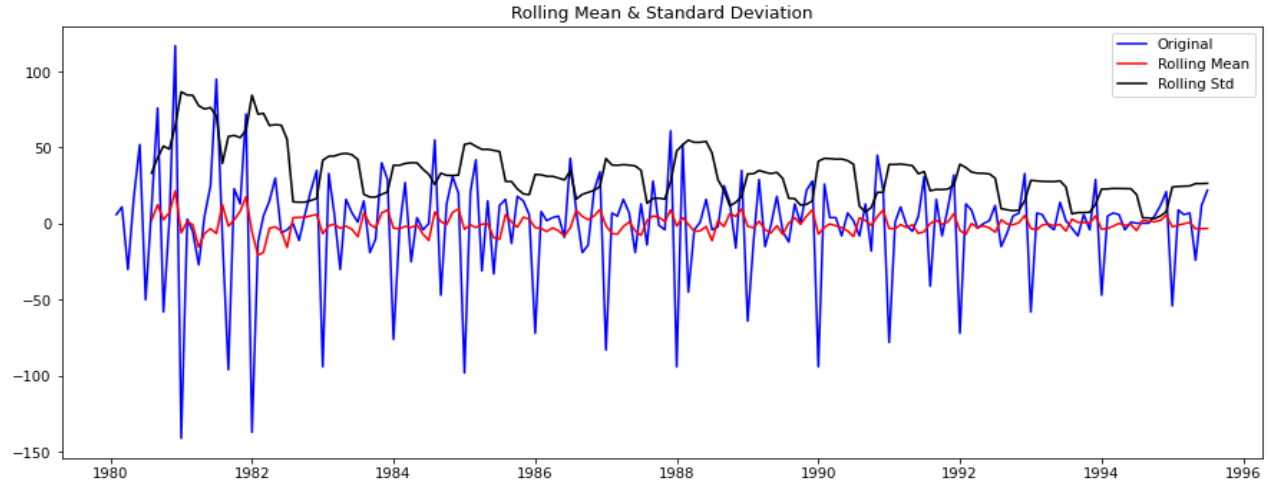


Visually we can able to find the Trend and Seasonality in the Rose wine time series plot. So it may be a Non-Stationary Time series plot.




```
Results of Dickey-Fuller Test:
Test Statistic                  -1.876699
p-value                          0.343101
#Lags Used                      13.000000
Number of Observations Used    173.000000
Critical Value (1%)             -3.468726
Critical Value (5%)             -2.878396
Critical Value (10%)            -2.575756
dtype: float64
```

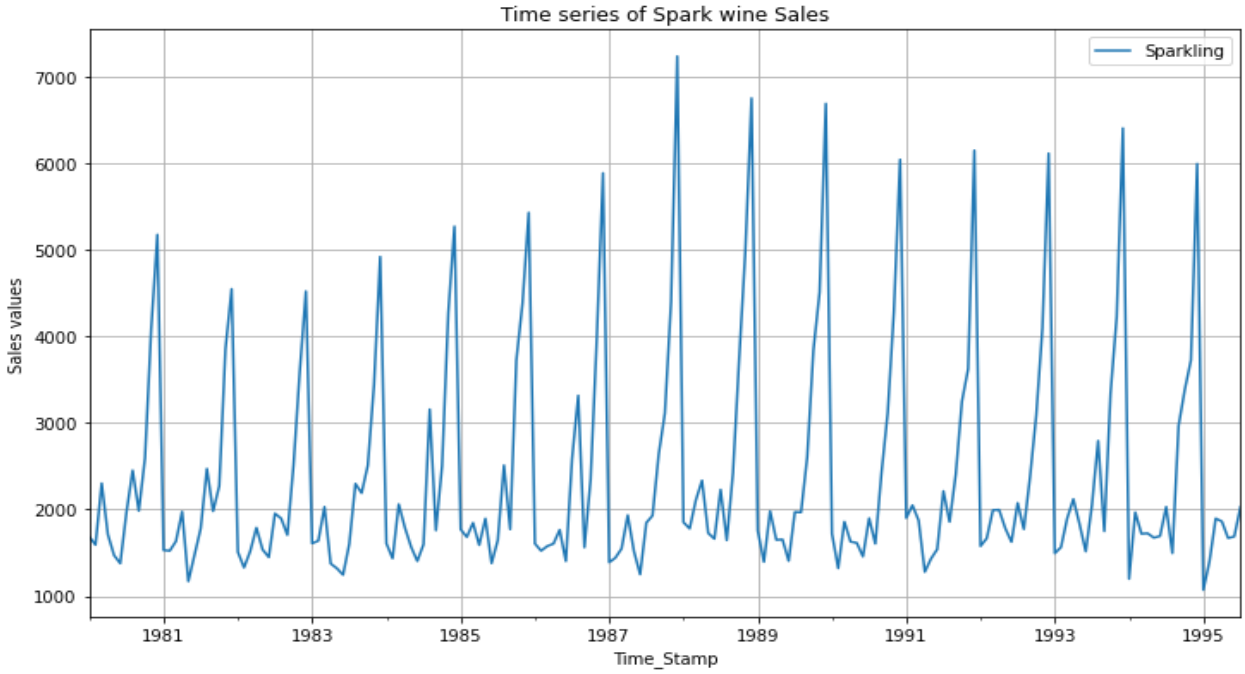
We see that the 'P-Value' is greater than (0.05) at 5% significant level the Time Series is non-stationary. Let us take a difference of order 1 and check whether the Time Series is stationary or not.

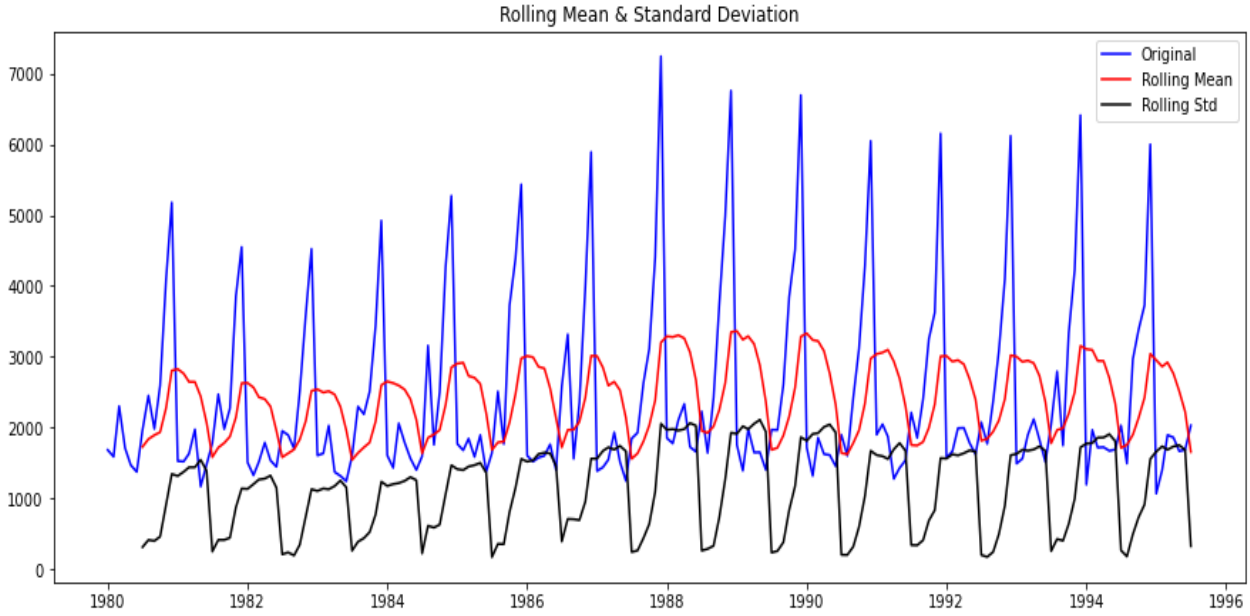



```
Results of Dickey-Fuller Test:
Test Statistic                -8.044392e+00
p-value                        1.810895e-12
#Lags Used                     1.200000e+01
Number of Observations Used    1.730000e+02
Critical Value (1%)           -3.468726e+00
Critical Value (5%)           -2.878396e+00
Critical Value (10%)          -2.575756e+00
dtype: float64
```

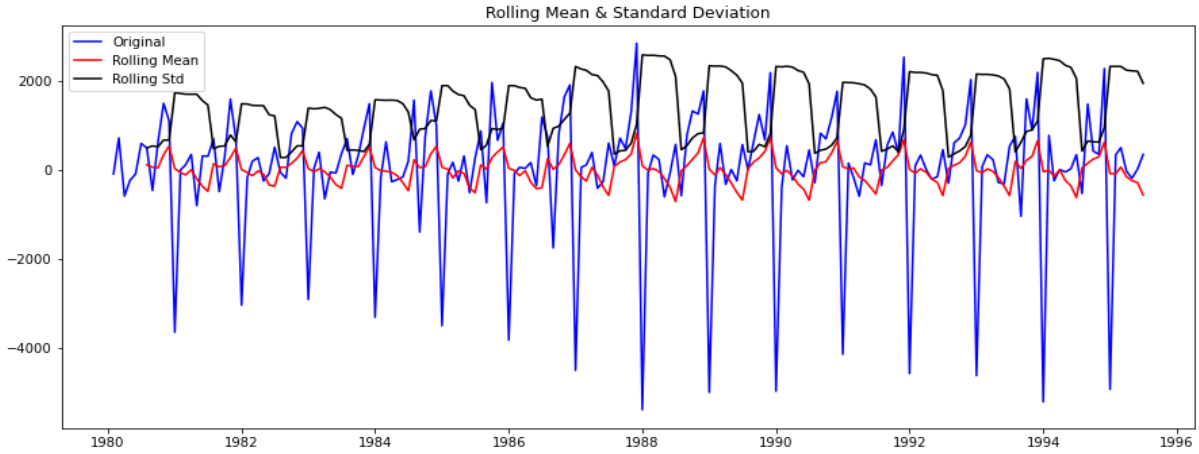
Now, the null hypothesis can be rejected because the p-value is below a set significance level. The defaults significance level is 5% **p-value <= significance level (default: 0.05)**: So, Reject the null hypothesis (H0), the data does not have a unit root and is stationary.

Visual and Statistical Check to find Stationarity in Sparkling wine Dataset:






```
Results of Dickey-Fuller Test:
Test Statistic                  -1.360497
p-value                          0.601061
#Lags Used                      11.000000
Number of Observations Used    175.000000
Critical Value (1%)             -3.468280
Critical Value (5%)             -2.878202
Critical Value (10%)            -2.575653
dtype: float64
```




```
Results of Dickey-Fuller Test:
Test Statistic                  -45.050301
p-value                           0.000000
#Lags Used                       10.000000
Number of Observations Used     175.000000
Critical Value (1%)              -3.468280
Critical Value (5%)              -2.878202
Critical Value (10%)             -2.575653
dtype: float64
```

Now, the null hypothesis can be rejected because the p-value is below a set significance level. The defaults significance level is 5% **p-value <= significance level (default: 0.05)**: So, Reject the null hypothesis (H0), the data does not have a unit root and is stationary.

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.

Method 8.1: Automated ARIMA model for Rose wine Dataset:

The data has some seasonality so ideally, we should build a SARIMA model. But for demonstration purposes we are building an ARIMA model both by looking at the minimum AIC criterion and by looking at the ACF and the PACF plots.

```
Some parameter combinations for the Model...
Model: (0, 1, 1)
Model: (0, 1, 2)
Model: (0, 1, 3)
Model: (1, 1, 0)
Model: (1, 1, 1)
Model: (1, 1, 2)
Model: (1, 1, 3)
Model: (2, 1, 0)
Model: (2, 1, 1)
Model: (2, 1, 2)
Model: (2, 1, 3)
Model: (3, 1, 0)
Model: (3, 1, 1)
Model: (3, 1, 2)
Model: (3, 1, 3)
```

```
ARIMA(0, 1, 0) - AIC:1335.1526583086775
ARIMA(0, 1, 1) - AIC:1280.726183046448
ARIMA(0, 1, 2) - AIC:1276.8353734911866
ARIMA(0, 1, 3) - AIC:1278.0742599150858
ARIMA(1, 1, 0) - AIC:1319.3483105802602
ARIMA(1, 1, 1) - AIC:1277.775753553521
ARIMA(1, 1, 2) - AIC:1277.3592281129256
ARIMA(1, 1, 3) - AIC:1279.312639992571
ARIMA(2, 1, 0) - AIC:1300.6092611744193
ARIMA(2, 1, 1) - AIC:1279.0456894093354
ARIMA(2, 1, 2) - AIC:1279.2986939365205
ARIMA(2, 1, 3) - AIC:1281.1962260431185
ARIMA(3, 1, 0) - AIC:1299.4787391543089
ARIMA(3, 1, 1) - AIC:1279.6059618578536
ARIMA(3, 1, 2) - AIC:1280.96924676004
ARIMA(3, 1, 3) - AIC:1273.1940974617016
```

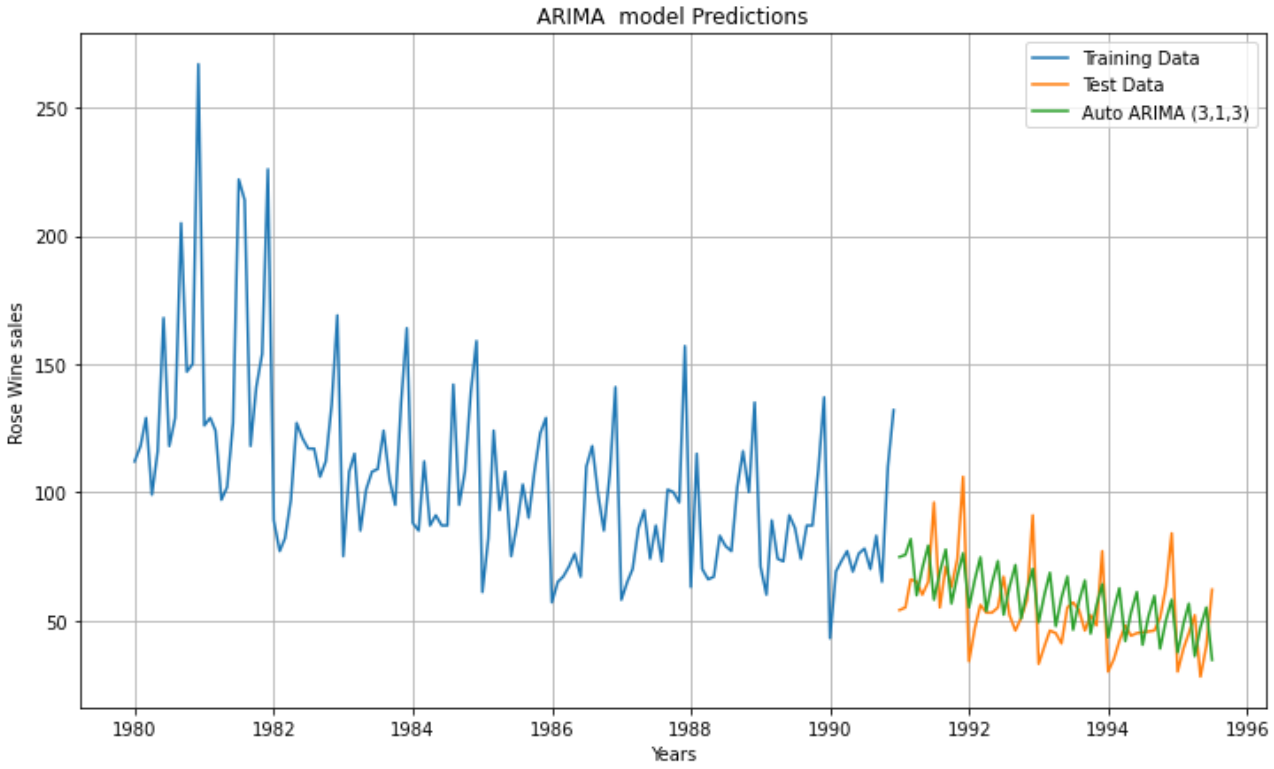
ARIMA Model Results

Dep. Variable:	D.Rose	No. Observations:	131
Model:	ARIMA(3, 1, 3)	Log Likelihood	-628.597
Method:	css-mle	S.D. of innovations	28.356
Date:	Sat, 27 Mar 2021	AIC	1273.194
Time:	22:11:41	BIC	1296.196
Sample:	02-01-1980 - 12-01-1990	HQIC	1282.541

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.4906	0.088	-5.548	0.000	-0.664	-0.317
ar.L1.D.Rose	-0.7243	0.086	-8.411	0.000	-0.893	-0.556
ar.L2.D.Rose	-0.7218	0.087	-8.342	0.000	-0.891	-0.552
ar.L3.D.Rose	0.2763	0.085	3.234	0.001	0.109	0.444
ma.L1.D.Rose	-0.0151	0.045	-0.339	0.735	-0.102	0.072
ma.L2.D.Rose	0.0151	0.044	0.340	0.734	-0.072	0.102
ma.L3.D.Rose	-1.0000	0.046	-21.901	0.000	-1.089	-0.911

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-0.5011	-0.8661j	1.0006	-0.3335
AR.2	-0.5011	+0.8661j	1.0006	0.3335
AR.3	3.6142	-0.0000j	3.6142	-0.0000
MA.1	1.0000	-0.0000j	1.0000	-0.0000
MA.2	-0.4925	-0.8703j	1.0000	-0.3320
MA.3	-0.4925	+0.8703j	1.0000	0.3320



For Auto ARIMA (3,1,3) forecast on the Test Data, RMSE is 15.986
For Auto ARIMA (3,1,3) forecast on the Test Data, MAPE is 26.080

Method 8.2: Automated ARIMA model for Sparkling wine Dataset:

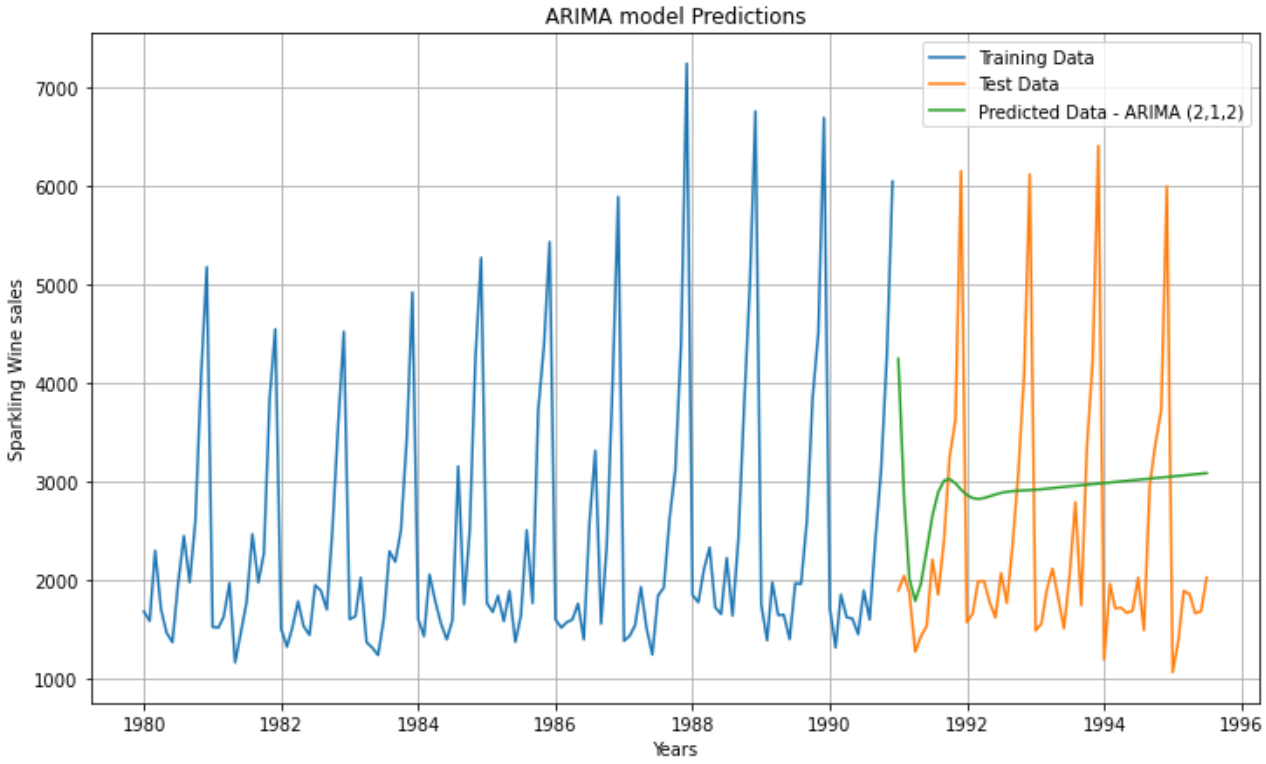
ARIMA Model Results

Dep. Variable:	D.Sparkling	No. Observations:	131
Model:	ARIMA(2, 1, 2)	Log Likelihood	-1099.313
Method:	css-mle	S.D. of innovations	1013.755
Date:	Sat, 27 Mar 2021	AIC	2210.626
Time:	22:11:47	BIC	2227.877
Sample:	02-01-1980 - 12-01-1990	HQIC	2217.636

	coef	std err	z	P>|z|	[0.025	0.975]
const	5.5845	0.519	10.753	0.000	4.567	6.602
ar.L1.D.Sparkling	1.2698	0.075	17.040	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	-1.9957	0.043	-46.821	0.000	-2.079	-1.912
ma.L2.D.Sparkling	0.9957	0.043	23.291	0.000	0.912	1.079

Roots

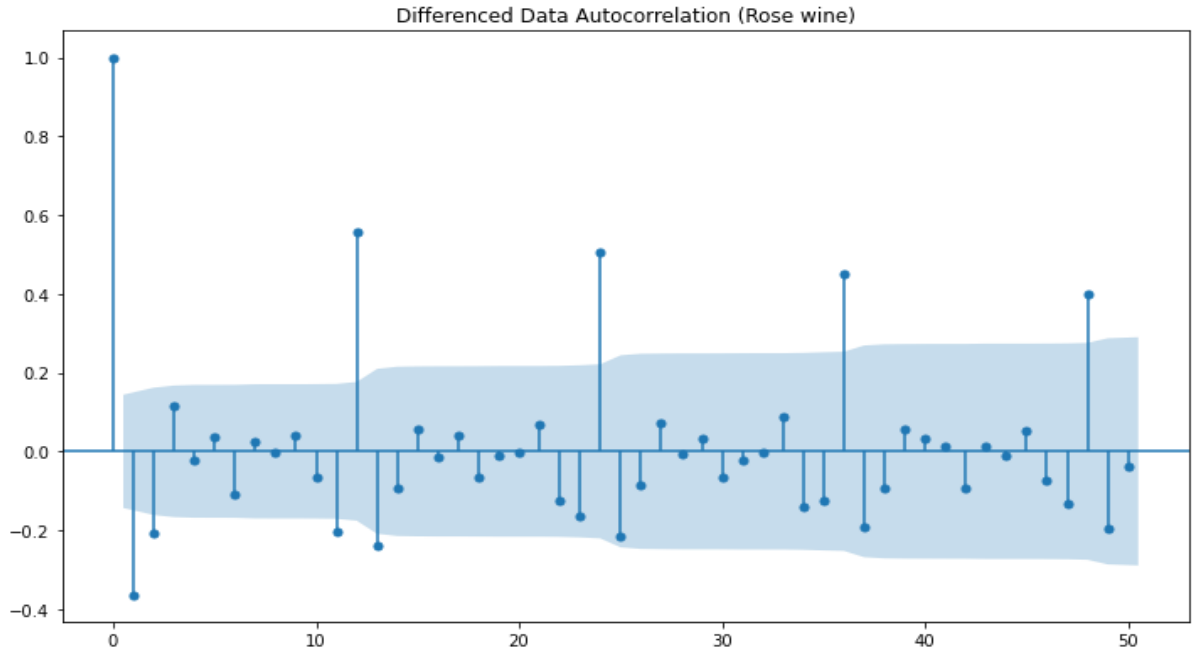
	Real	Imaginary	Modulus	Frequency
AR.1	1.1335	-0.7074j	1.3361	-0.0888
AR.2	1.1335	+0.7074j	1.3361	0.0888
MA.1	1.0000	+0.0000j	1.0000	0.0000
MA.2	1.0043	+0.0000j	1.0043	0.0000

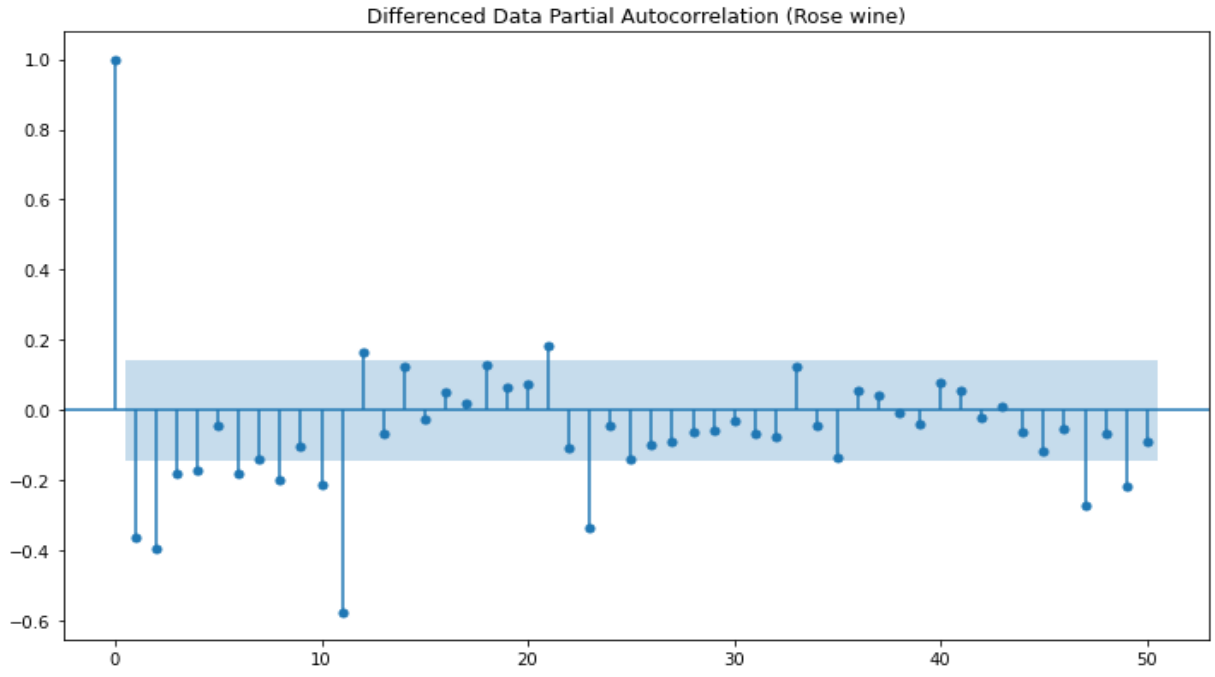


For Auto ARIMA (2,1,2) forecast on the Test Data, RMSE is 1374.037
For Auto ARIMA (2,1,2) forecast on the Test Data, MAPE is 48.330

Plotting ACF and PACF for rose wine dataset.

Let us look at the ACF and the PACF plots once more.





Here, we have taken alpha=0.05.

- The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 2.
- The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off to 4.

By looking at the above plots, we can say that both the PACF and ACF plot cuts-off at lag 4.

Model: Creating Auto SARIMA model

Let us look at the ACF plot once more to understand the seasonal parameter for the SARIMA model. We see that there can be a seasonality of 12. Anyway we will run our auto SARIMA models by setting seasonality both as 6 and 12. Setting the seasonality as 6 for the first iteration of the auto SARIMA model.

Method 9.1: Automated SARIMA model for Rose wine Dataset :Seasonality 6

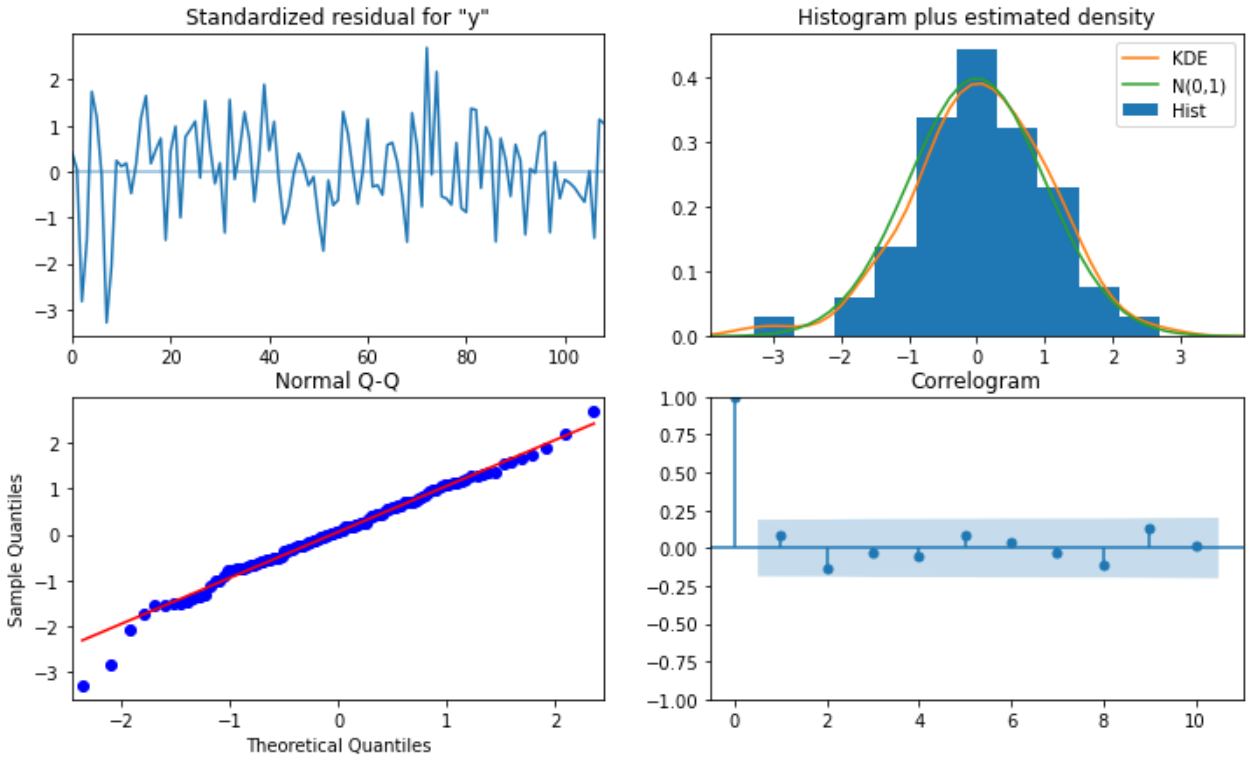
```
Examples of some parameter combinations for Model...
Model: (0, 1, 1)(0, 0, 1, 6)
Model: (0, 1, 2)(0, 0, 2, 6)
Model: (1, 1, 0)(1, 0, 0, 6)
Model: (1, 1, 1)(1, 0, 1, 6)
Model: (1, 1, 2)(1, 0, 2, 6)
Model: (2, 1, 0)(2, 0, 0, 6)
Model: (2, 1, 1)(2, 0, 1, 6)
Model: (2, 1, 2)(2, 0, 2, 6)
```

	param	seasonal	AIC
53	(1, 1, 2)	(2, 0, 2, 6)	1041.655818
26	(0, 1, 2)	(2, 0, 2, 6)	1043.600261
80	(2, 1, 2)	(2, 0, 2, 6)	1045.220359
71	(2, 1, 1)	(2, 0, 2, 6)	1051.673461
44	(1, 1, 1)	(2, 0, 2, 6)	1052.778470

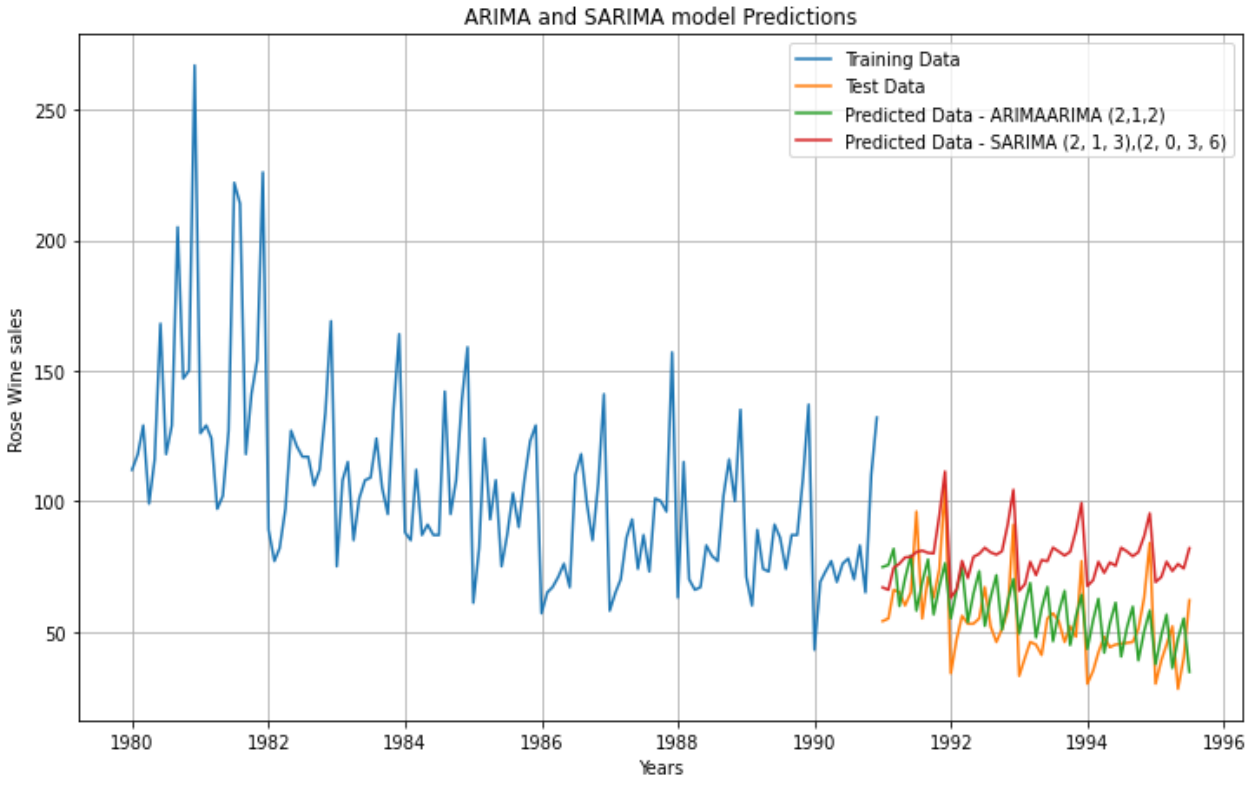
```
                                 SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(2, 1, 3)x(2, 0, 3, 6)   Log Likelihood                -464.872
Date:                           Sat, 27 Mar 2021   AIC                            951.744
Time:                                   22:12:16   BIC                            981.349
Sample:                                        0   HQIC                           963.750
                                             - 132
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5028      0.083     -6.084      0.000      -0.665      -0.341
ar.L2         -0.6628      0.084     -7.920      0.000      -0.827      -0.499
ma.L1         -0.3713    232.018     -0.002      0.999    -455.119     454.377
ma.L2          0.2033    145.848      0.001      0.999    -285.654     286.061
ma.L3         -0.8320    192.997     -0.004      0.997    -379.099     377.435
ar.S.L6       -0.0838      0.049     -1.720      0.085      -0.179       0.012
ar.S.L12       0.8099      0.052     15.466      0.000       0.707       0.913
ma.S.L6        0.1702      0.248      0.687      0.492      -0.315       0.656
ma.S.L12      -0.5646      0.199     -2.836      0.005      -0.955      -0.174
ma.S.L18       0.1709      0.143      1.198      0.231      -0.109       0.451
sigma2       260.7821   6.05e+04      0.004      0.997   -1.18e+05    1.19e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.72   Jarque-Bera (JB):                 4.77
Prob(Q):                              0.40   Prob(JB):                         0.09
Heteroskedasticity (H):               0.54   Skew:                            -0.36
Prob(H) (two-sided):                  0.06   Kurtosis:                         3.73
===================================================================================
```

```
array([ 66.90008129,  65.98947584,  74.43749292,  76.04017001,
        78.41455163,  78.63347519,  80.4808883 ,  81.01381416,
        80.15515623,  79.95248074,  94.75074252, 111.3218431 ,
        62.96142188,  66.24329563,  77.07357513,  70.52267262,
        78.62593752,  79.77698954,  82.11262477,  80.35035508,
        79.48085385,  80.76881938,  91.13485368, 104.33310351,
        65.65746831,  68.20139388,  76.72200786,  71.64472789,
        77.38464251,  76.96569147,  82.19027262,  80.66581723,
        79.12774449,  80.5797525 ,  88.58218464,  99.27306879,
        67.47096787,  69.70191547,  76.73741038,  72.47230712,
        76.46258537,  75.2426412 ,  82.12246236,  80.69882108,
        78.87508737,  80.40406679,  86.54558728,  95.31169383,
        68.98001241,  70.90205605,  76.75430308,  73.17387677,
        75.88916302,  74.1665893 ,  81.94554831])
```

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	66.900081	16.350100	34.854473	98.945689
1	65.989476	16.481270	33.686780	98.292172
2	74.437493	16.587192	41.927193	106.947793
3	76.040170	16.709772	43.289619	108.790721
4	78.414552	16.710386	45.662797	111.166306



From the model diagnostics plot, we can see that all the individual diagnostics plots almost follow the theoretical numbers and thus we cannot develop any pattern from these plots.




```
For Auto SARIMA (2,1,3),(2,0,3,6) forecast on the Test Data,  RMSE is 27.125
For Auto SARIMA (2,1,3),(2,0,3,6) forecast on the Test Data,  MAPE is 47.310
```

Method 9.2: Automated SARIMA model for Rose wine Dataset :Seasonality 12

```
Examples of some parameter combinations for Model
Model: (0, 1, 1)(0, 0, 1, 12)
Model: (0, 1, 2)(0, 0, 2, 12)
Model: (1, 1, 0)(1, 0, 0, 12)
Model: (1, 1, 1)(1, 0, 1, 12)
Model: (1, 1, 2)(1, 0, 2, 12)
Model: (2, 1, 0)(2, 0, 0, 12)
Model: (2, 1, 1)(2, 0, 1, 12)
Model: (2, 1, 2)(2, 0, 2, 12)
```

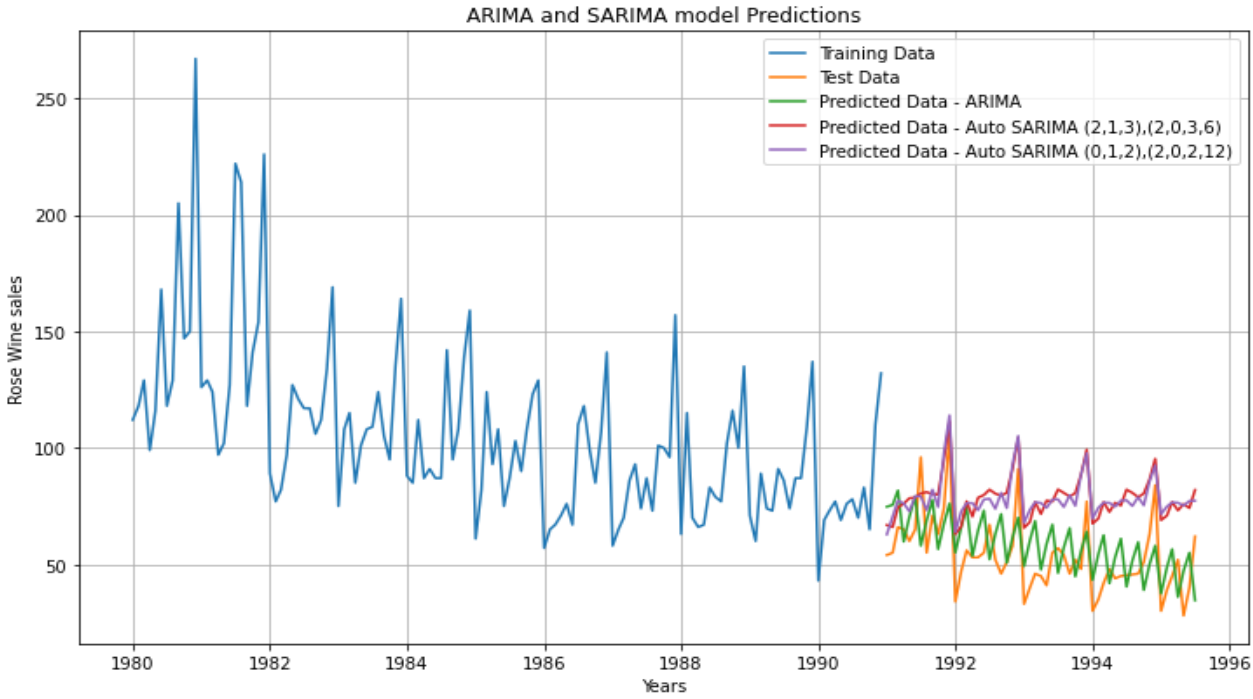
	param	seasonal	AIC
26	(0, 1, 2)	(2, 0, 2, 12)	887.937509
53	(1, 1, 2)	(2, 0, 2, 12)	889.902650
80	(2, 1, 2)	(2, 0, 2, 12)	890.668798
69	(2, 1, 1)	(2, 0, 0, 12)	896.518161
78	(2, 1, 2)	(2, 0, 0, 12)	897.346444

```
                                 SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 12)   Log Likelihood                -436.969
Date:                            Sun, 28 Mar 2021   AIC                            887.938
Time:                                    12:39:11   BIC                            906.448
Sample:                                         0   HQIC                           895.437
                                              - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.8427    189.705     -0.004      0.996    -372.658     370.972
ma.L2         -0.1573     29.804     -0.005      0.996     -58.571      58.257
ar.S.L12       0.3467      0.079      4.375      0.000       0.191       0.502
ar.S.L24       0.3023      0.076      3.996      0.000       0.154       0.451
ma.S.L12       0.0767      0.133      0.577      0.564      -0.184       0.337
ma.S.L24      -0.0726      0.146     -0.498      0.618      -0.358       0.213
sigma2       251.3136   4.77e+04      0.005      0.996   -9.32e+04    9.37e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.10   Jarque-Bera (JB):                 2.33
Prob(Q):                              0.75   Prob(JB):                         0.31
Heteroskedasticity (H):               0.88   Skew:                             0.37
Prob(H) (two-sided):                  0.70   Kurtosis:                         3.03
===================================================================================

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

For Auto SARIMA (0,1,2),(2,0,2,12) forecast on the Test Data, RMSE is 26.928
For Auto SARIMA (0,1,2),(2,0,2,12) forecast on the Test Data, MAPE is 46.600

```
1 predicted_auto_SARIMA_12_rose.predicted_mean
```

```
array([ 62.86726279,  70.54118967,  77.35641038,  76.2088136 ,
        72.74739753,  79.23086511,  79.21765403,  73.26209495,
        82.0739462 ,  74.67271597,  97.77998502, 113.97566465,
        63.85264079,  72.74761998,  76.28626616,  76.30908219,
        73.26798954,  77.8756539 ,  78.17510638,  73.87057954,
        80.6803955 ,  74.16637739,  92.98112907, 105.1152729 ,
        67.85648924,  73.41236844,  76.69969505,  76.36065012,
        74.25979069,  77.81745168,  77.91727923,  74.62432125,
        79.64942448,  75.1533519 ,  88.66259091,  97.76603397,
        69.69442556,  74.30991484,  76.51949397,  76.4088434 ,
        74.76104484,  77.3875488 ,  77.51269362,  75.06961174,
        78.87066649,  75.34245853,  85.71448363,  92.53923472,
        71.54213935,  74.82207351,  76.58201018,  76.44114289,
        75.23468553,  77.2209029 ,  77.29447231])
```

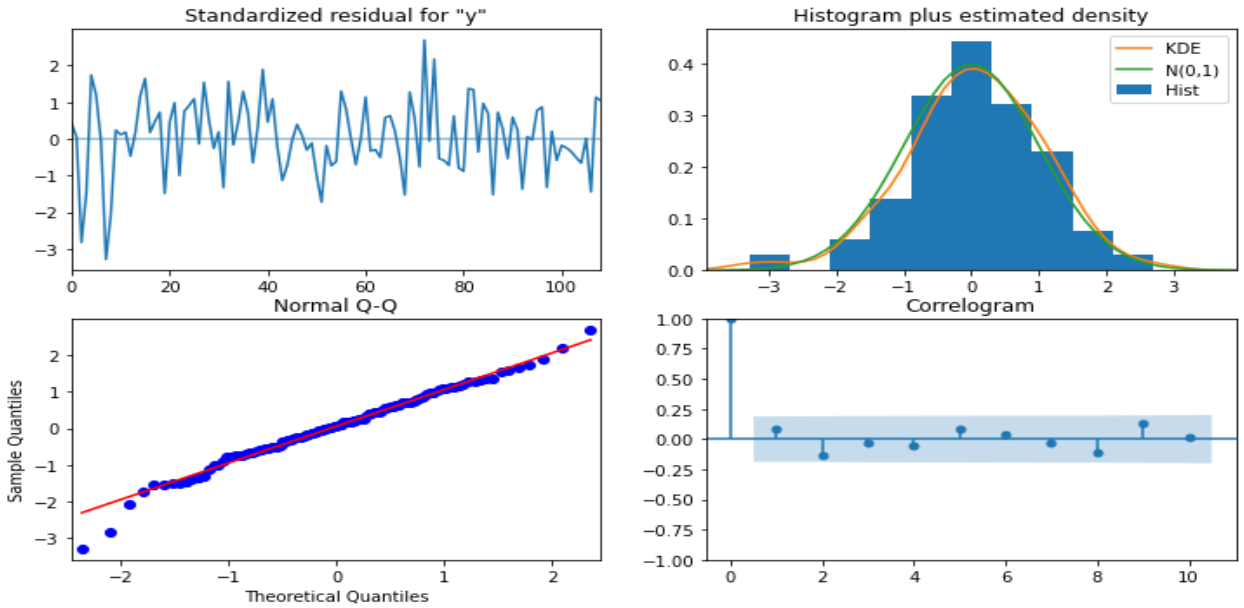



```
For Auto SARIMA (0,1,2),(2,0,2,12) forecast on the Test Data,  RMSE is 26.928
For Auto SARIMA (0,1,2),(2,0,2,12) forecast on the Test Data,  MAPE is 46.600
```

Method 9.3: Automated SARIMA model - Sparkling wine Dataset :Seasonality 6

```
Examples of some parameter combinations for Model...
Model: (0, 1, 1)(0, 0, 1, 6)
Model: (0, 1, 2)(0, 0, 2, 6)
Model: (1, 1, 0)(1, 0, 0, 6)
Model: (1, 1, 1)(1, 0, 1, 6)
Model: (1, 1, 2)(1, 0, 2, 6)
Model: (2, 1, 0)(2, 0, 0, 6)
Model: (2, 1, 1)(2, 0, 1, 6)
Model: (2, 1, 2)(2, 0, 2, 6)
```

	param	seasonal	AIC
53	(1, 1, 2)	(2, 0, 2, 6)	1727.678710
26	(0, 1, 2)	(2, 0, 2, 6)	1727.888819
80	(2, 1, 2)	(2, 0, 2, 6)	1729.192582
17	(0, 1, 1)	(2, 0, 2, 6)	1741.641477
44	(1, 1, 1)	(2, 0, 2, 6)	1743.374728

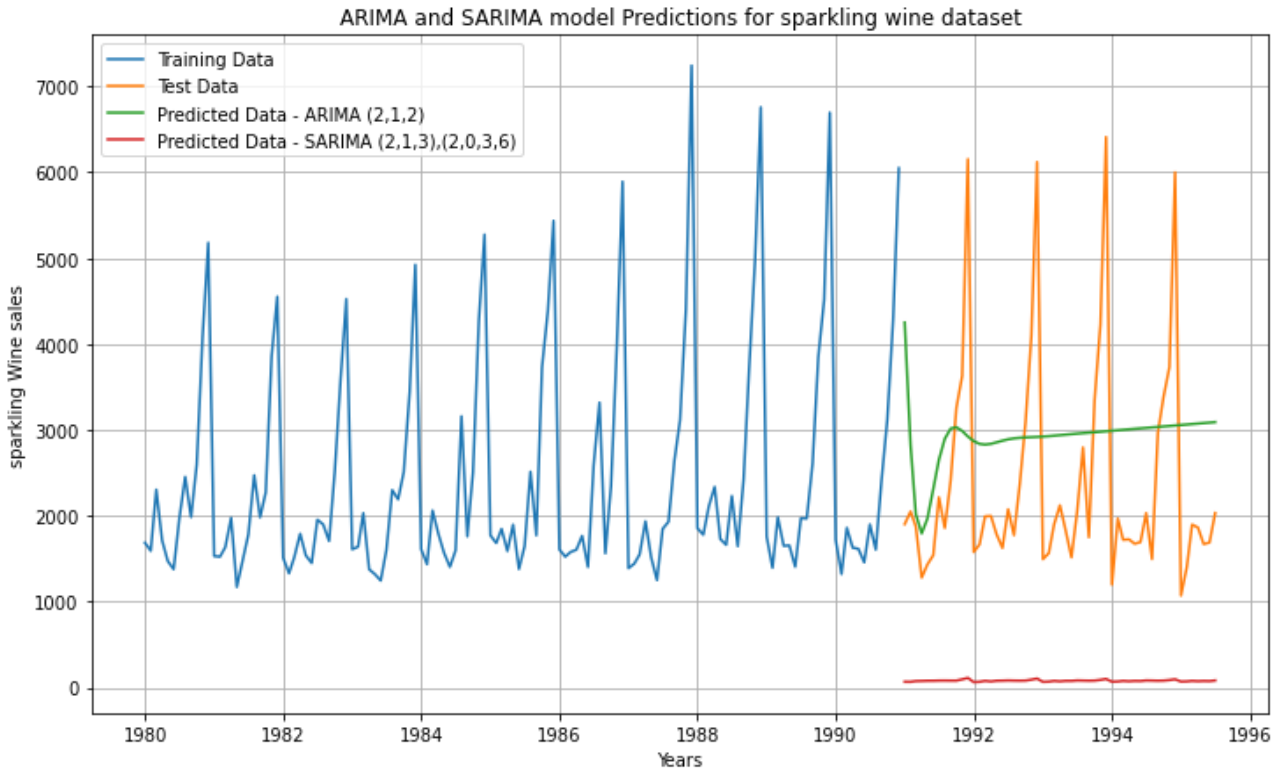


SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(2, 1, 3)x(2, 0, 3, 6)	Log Likelihood	-464.872
Date:	Sun, 28 Mar 2021	AIC	951.744
Time:	12:40:04	BIC	981.349
Sample:	0	HQIC	963.750
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5028	0.083	-6.084	0.000	-0.665	-0.341
ar.L2	-0.6628	0.084	-7.920	0.000	-0.827	-0.499
ma.L1	-0.3713	232.018	-0.002	0.999	-455.119	454.377
ma.L2	0.2033	145.848	0.001	0.999	-285.654	286.061
ma.L3	-0.8320	192.997	-0.004	0.997	-379.099	377.435
ar.S.L6	-0.0838	0.049	-1.720	0.085	-0.179	0.012
ar.S.L12	0.8099	0.052	15.466	0.000	0.707	0.913
ma.S.L6	0.1702	0.248	0.687	0.492	-0.315	0.656
ma.S.L12	-0.5646	0.199	-2.836	0.005	-0.955	-0.174
ma.S.L18	0.1709	0.143	1.198	0.231	-0.109	0.451
sigma2	260.7821	6.05e+04	0.004	0.997	-1.18e+05	1.19e+05

Ljung-Box (L1) (Q):	0.72	Jarque-Bera (JB):	4.77
Prob(Q):	0.40	Prob(JB):	0.09
Heteroskedasticity (H):	0.54	Skew:	-0.36
Prob(H) (two-sided):	0.06	Kurtosis:	3.73



For Auto SARIMA (2,1,3),(2,0,3,6) forecast on the Test Data, RMSE is 2643.865
For Auto SARIMA (2,1,3),(2,0,3,6) forecast on the Test Data, MAPE is 96.720

Method 9.4: Automated SARIMA model - Sparkling wine Dataset Seasonality 12

```
Examples of some parameter combinations for Model
Model: (0, 1, 1)(0, 0, 1, 12)
Model: (0, 1, 2)(0, 0, 2, 12)
Model: (1, 1, 0)(1, 0, 0, 12)
Model: (1, 1, 1)(1, 0, 1, 12)
Model: (1, 1, 2)(1, 0, 2, 12)
Model: (2, 1, 0)(2, 0, 0, 12)
Model: (2, 1, 1)(2, 0, 1, 12)
Model: (2, 1, 2)(2, 0, 2, 12)
```

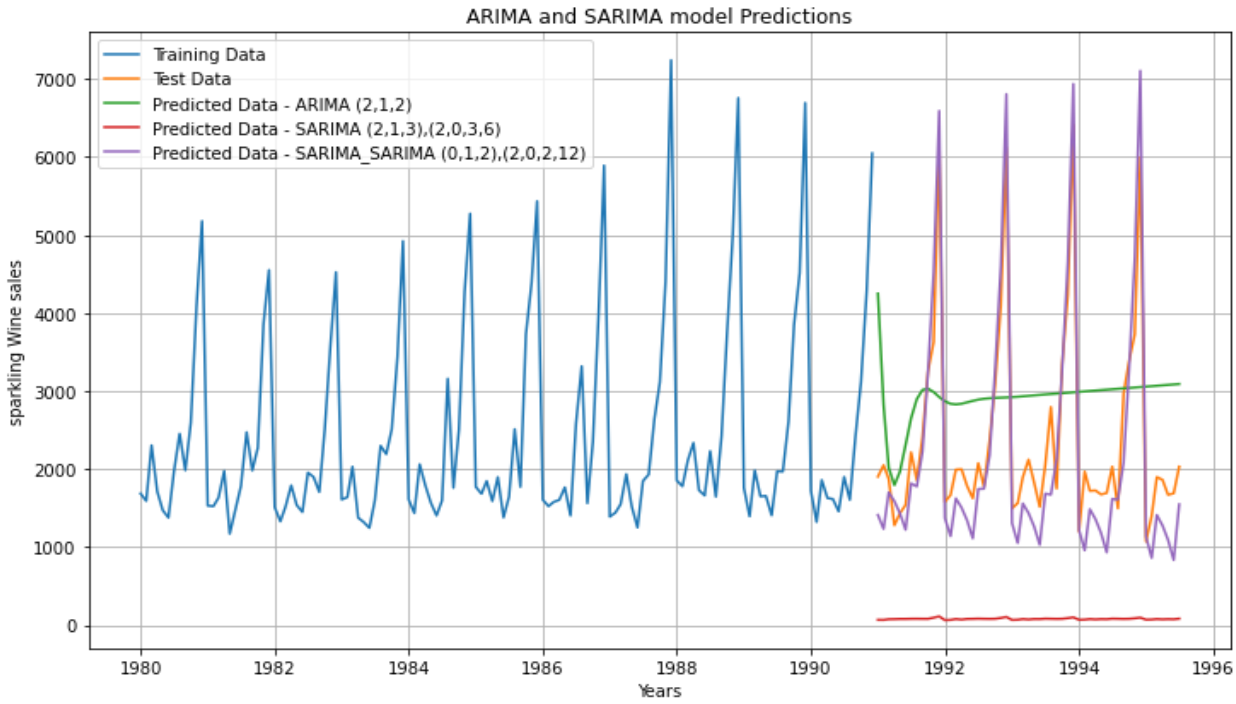
	param	seasonal	AIC
26	(0, 1, 2)	(2, 0, 2, 12)	887.937509
53	(1, 1, 2)	(2, 0, 2, 12)	889.902650
80	(2, 1, 2)	(2, 0, 2, 12)	890.668798
69	(2, 1, 1)	(2, 0, 0, 12)	896.518161
78	(2, 1, 2)	(2, 0, 0, 12)	897.346444

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(0, 1, 2)x(2, 0, 2, 12)	Log Likelihood	-771.561
Date:	Sun, 28 Mar 2021	AIC	1557.122
Time:	12:40:54	BIC	1575.632
Sample:	0	HQIC	1564.621
	- 132		
Covariance Type:	opg		

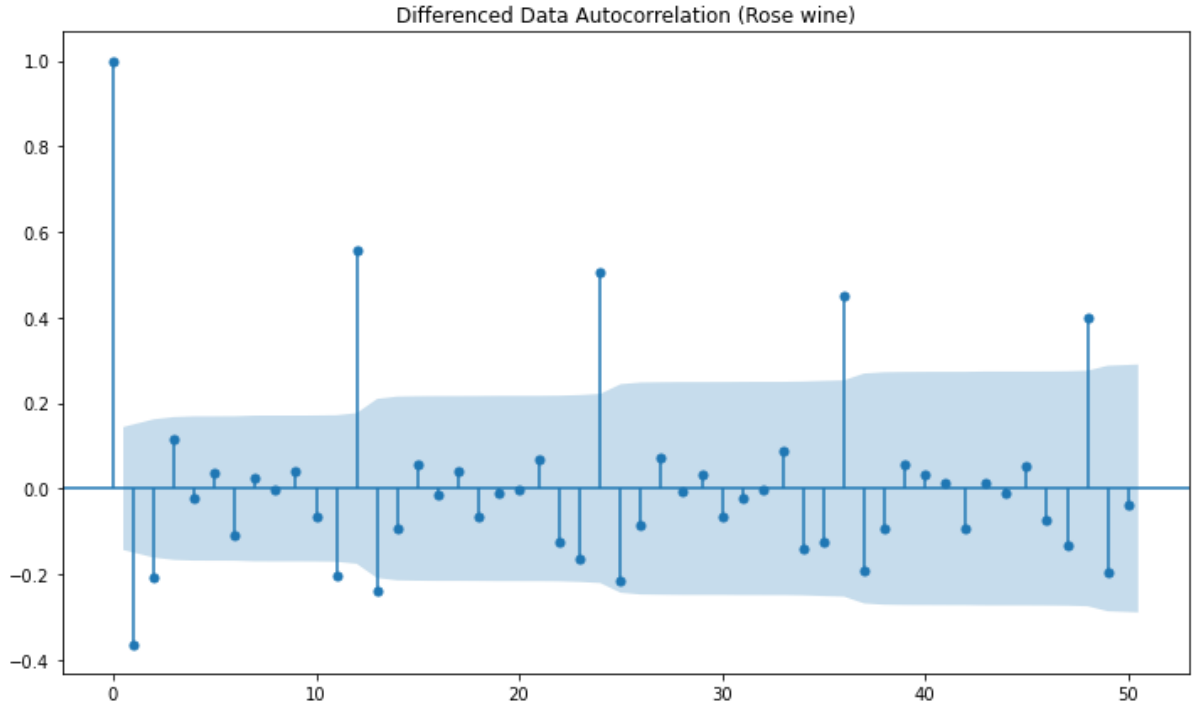
	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.7771	0.101	-7.680	0.000	-0.975	-0.579
ma.L2	-0.1239	0.121	-1.022	0.307	-0.361	0.114
ar.S.L12	0.6981	0.751	0.929	0.353	-0.774	2.171
ar.S.L24	0.3599	0.783	0.460	0.646	-1.174	1.894
ma.S.L12	1.6415	3.717	0.442	0.659	-5.643	8.926
ma.S.L24	-1.5933	2.664	-0.598	0.550	-6.815	3.628
sigma2	2.915e+04	9.51e+04	0.306	0.759	-1.57e+05	2.16e+05

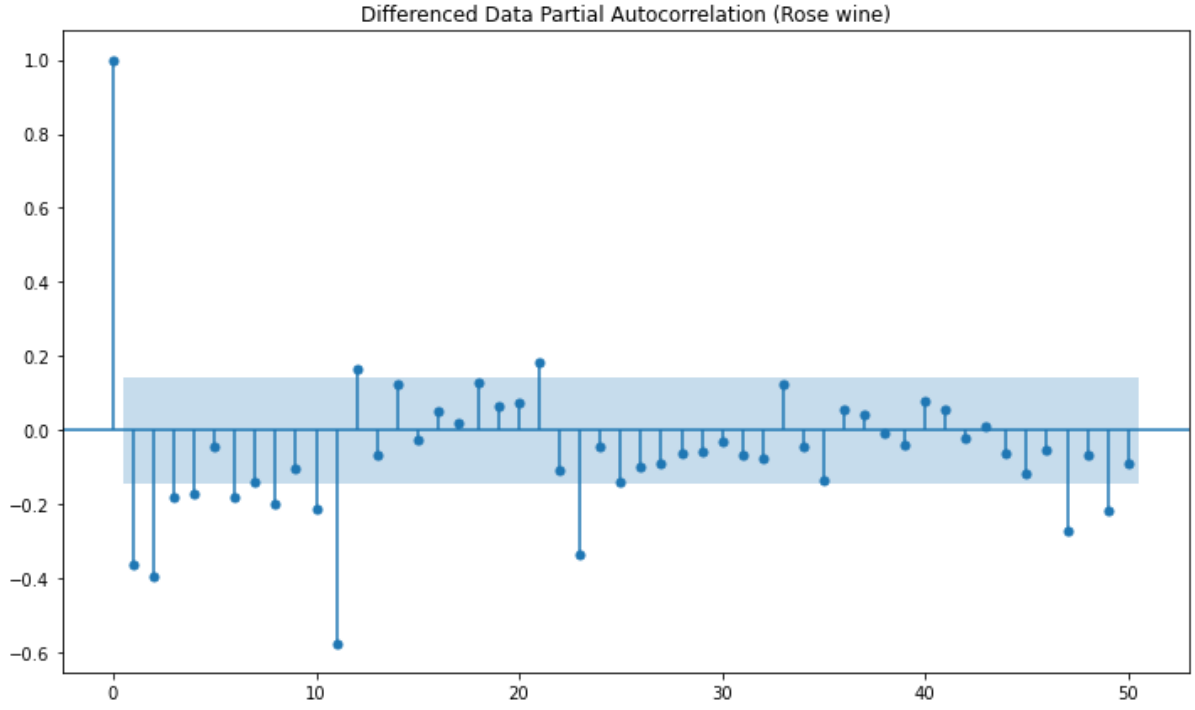
Ljung-Box (L1) (Q):	0.03	Jarque-Bera (JB):	22.07
Prob(Q):	0.85	Prob(JB):	0.00
Heteroskedasticity (H):	1.45	Skew:	0.49
Prob(H) (two-sided):	0.28	Kurtosis:	5.04




```
For Auto SARIMA (0,1,2),(2,0,2,12) forecast on the Test Data,  RMSE is 526.477
For Auto SARIMA (0,1,2),(2,0,2,12) forecast on the Test Data,  MAPE is 18.480
```

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.





Here, we have taken alpha=0.05.

We are going to take the seasonal period as 12. We will keep the p(4) and q(2) parameters same as the ARIMA model.

Method 10.1: Manual ARIMA model for Rose wine Dataset:

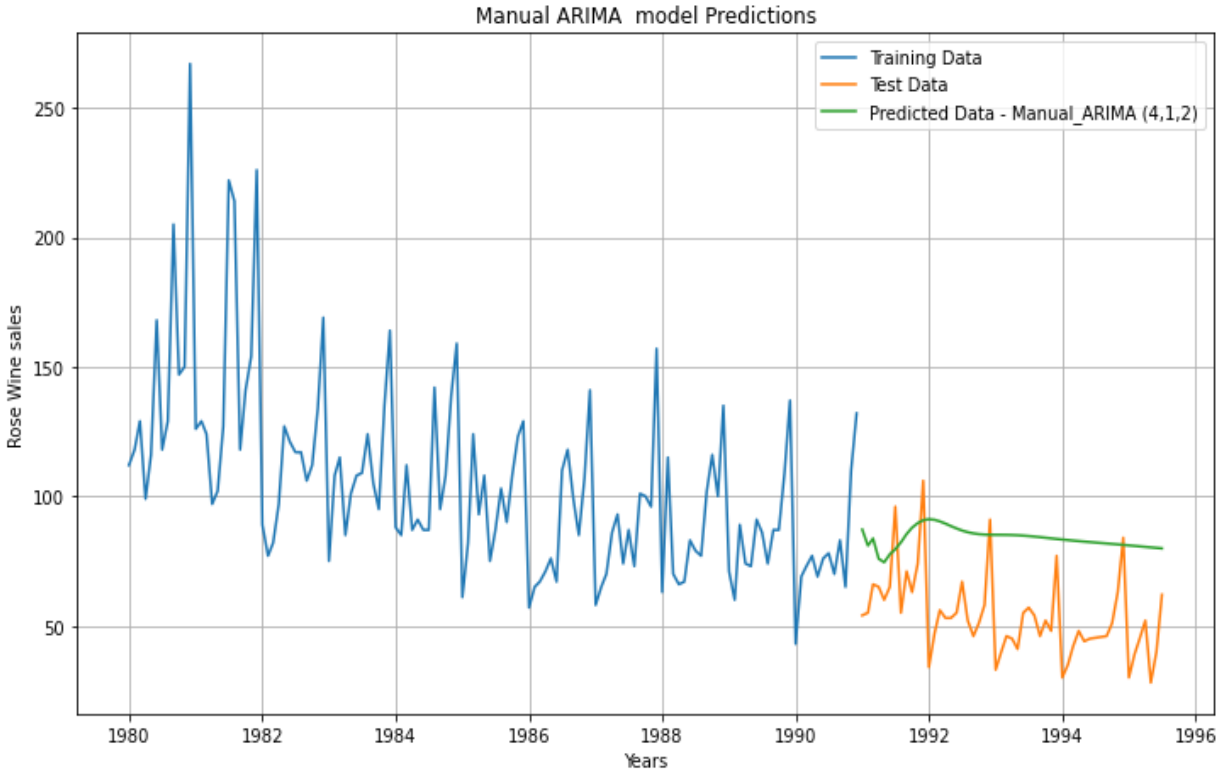
ARIMA Model Results

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, 28 Mar 2021	AIC	1283.753
Time:	12:40:56	BIC	1306.754
Sample:	02-01-1980 - 12-01-1990	HQIC	1293.099

	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	-2.692	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	-0.6863	-1.6644j	1.8003	-0.3122
AR.4	-0.6863	+1.6644j	1.8003	0.3122
MA.1	0.9753	-0.2209j	1.0000	-0.0355
MA.2	0.9753	+0.2209j	1.0000	0.0355



For Manual_ARIMA (4,1,2) forecast on the Test Data, RMSE is 33.950
For Manual_ARIMA (4,1,2) forecast on the Test Data, MAPE is 58.420

Method 10.2: Manual ARIMA model for Sparkling wine Dataset:

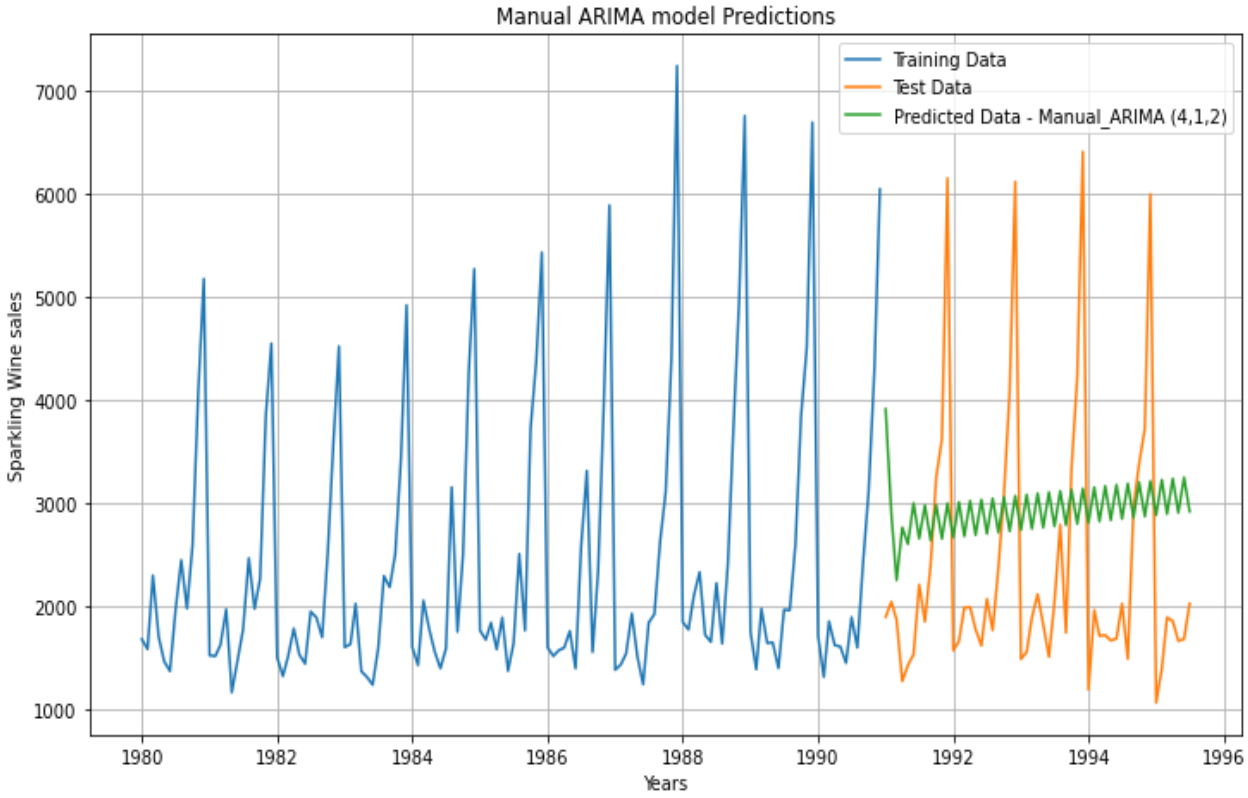
ARIMA Model Results

Dep. Variable:	D.Sparkling	No. Observations:	131
Model:	ARIMA(3, 1, 2)	Log Likelihood	-1107.464
Method:	css-mle	S.D. of innovations	1106.238
Date:	Sun, 28 Mar 2021	AIC	2228.928
Time:	12:41:00	BIC	2249.054
Sample:	02-01-1980 - 12-01-1990	HQIC	2237.106

	coef	std err	z	P>|z|	[0.025	0.975]
const	5.9844	3.643	1.643	0.100	-1.156	13.125
ar.L1.D.Sparkling	-0.4420	1.28e-05	-3.46e+04	0.000	-0.442	-0.442
ar.L2.D.Sparkling	0.3079	4.63e-05	6645.786	0.000	0.308	0.308
ar.L3.D.Sparkling	-0.2501	3.81e-05	-6560.861	0.000	-0.250	-0.250
ma.L1.D.Sparkling	-0.0008	0.020	-0.040	0.968	-0.039	0.037
ma.L2.D.Sparkling	-0.9992	0.020	-51.220	0.000	-1.037	-0.961

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-1.0000	-0.0000j	1.0000	-0.5000
AR.2	1.1156	-1.6594j	1.9996	-0.1558
AR.3	1.1156	+1.6594j	1.9996	0.1558
MA.1	1.0000	+0.0000j	1.0000	0.0000
MA.2	-1.0008	+0.0000j	1.0008	0.5000



For Manual_ARIMA (4,1,2) forecast on the Test Data, RMSE is 1379.049
For Manual_ARIMA (4,1,2) forecast on the Test Data, MAPE is 49.320

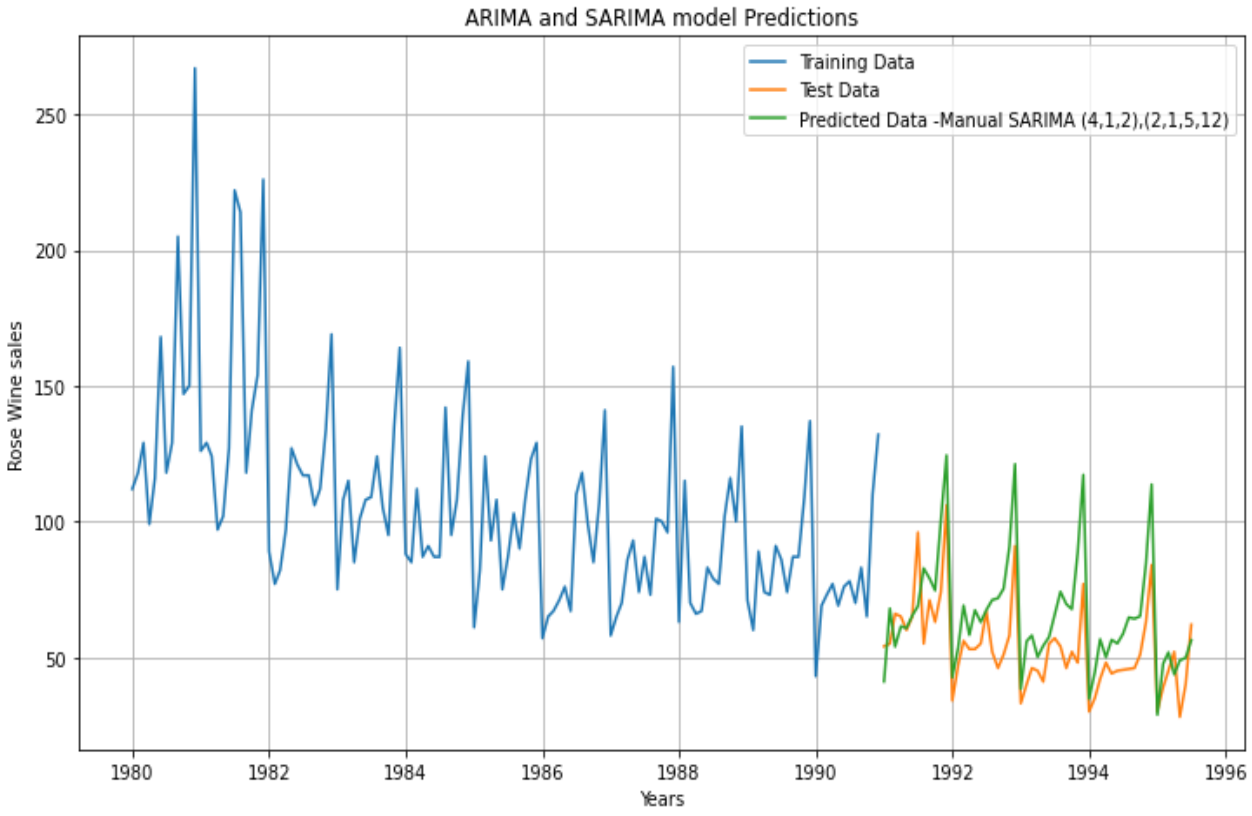
Method 10.3: Manual SARIMA model for Rose wine Dataset:

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(4, 1, 2)x(2, 1, [1, 2, 3, 4, 5, 6, 7, 8, 9], 12)	Log Likelihood	-13.779
Date:	Sun, 28 Mar 2021	AIC	63.558
Time:	12:44:04	BIC	64.988
Sample:	0	HQIC	53.914
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.3696	4.515	-0.303	0.762	-10.218	7.479
ar.L2	-1.3327	4.231	-0.315	0.753	-9.625	6.959
ar.L3	-1.0268	6.093	-0.169	0.866	-12.970	10.916
ar.L4	-0.2775	1.680	-0.165	0.869	-3.570	3.015
ma.L1	-1.3103	2.157	-0.608	0.543	-5.537	2.917
ma.L2	0.3103	6.889	0.045	0.964	-13.192	13.812
ar.S.L12	0.7600	25.851	0.029	0.977	-49.908	51.427
ar.S.L24	0.3366	34.616	0.010	0.992	-67.510	68.183
ma.S.L12	-0.2526	27.869	-0.009	0.993	-54.875	54.370
ma.S.L24	0.2583	5.948	0.043	0.965	-11.399	11.916
ma.S.L36	-0.1349	4.764	-0.028	0.977	-9.471	9.201
ma.S.L48	0.0038	3.377	0.001	0.999	-6.614	6.622
ma.S.L60	0.0020	0.290	0.007	0.994	-0.566	0.570
ma.S.L72	-0.0001	0.070	-0.002	0.998	-0.137	0.137
ma.S.L84	-0.0002	0.165	-0.001	0.999	-0.324	0.324
ma.S.L96	0.0011	0.314	0.004	0.997	-0.614	0.616
ma.S.L108	-0.0019	0.994	-0.002	0.998	-1.950	1.947
sigma2	1.6104	0.172	9.349	0.000	1.273	1.948

Ljung-Box (L1) (Q):	2.08	Jarque-Bera (JB):	0.32
Prob(Q):	0.15	Prob(JB):	0.85
Heteroskedasticity (H):	5.73	Skew:	0.43
Prob(H) (two-sided):	0.19	Kurtosis:	2.51



For Manual_SARIMA (4,1,2),(2, 1, 5, 12) forecast on the Test Data, RMSE is 17.239
For Manual_SARIMA (4,1,2),(2, 1, 5, 12) forecast on the Test Data, MAPE is 26.450

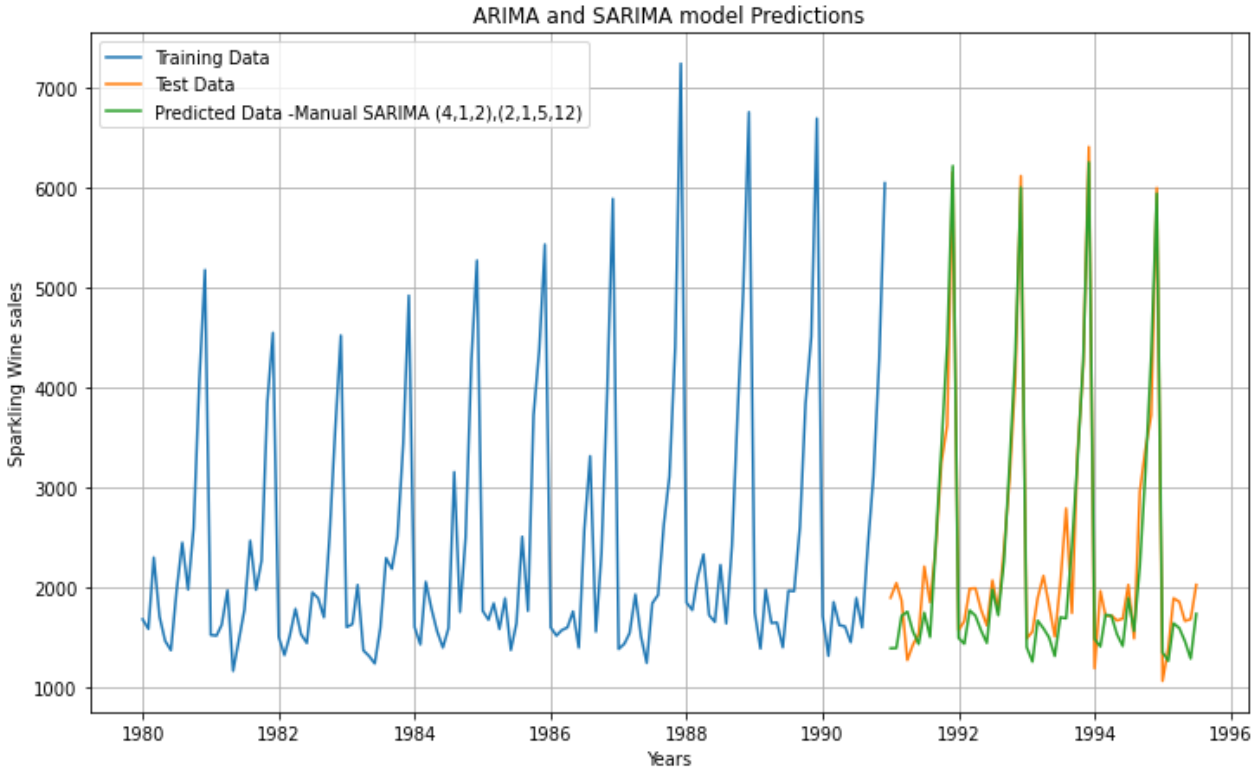
Method 10.4: Manual SARIMA model for Sparkling wine Dataset:

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(4, 1, 2)x(2, 1, [1, 2, 3, 4, 5], 12)	Log Likelihood	-417.930
Date:	Sun, 28 Mar 2021	AIC	863.860
Time:	15:05:05	BIC	892.215
Sample:	0	HQIC	874.853
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5047	0.332	-1.521	0.128	-1.155	0.146
ar.L2	0.1800	0.206	0.875	0.381	-0.223	0.583
ar.L3	0.1264	0.214	0.590	0.555	-0.293	0.546
ar.L4	-0.0404	0.266	-0.152	0.879	-0.561	0.480
ma.L1	-0.1231	0.266	-0.462	0.644	-0.645	0.399
ma.L2	-0.7976	0.274	-2.912	0.004	-1.334	-0.261
ar.S.L12	-0.9811	0.406	-2.417	0.016	-1.777	-0.186
ar.S.L24	-0.2323	0.423	-0.549	0.583	-1.061	0.596
ma.S.L12	0.9127	0.602	1.517	0.129	-0.267	2.092
ma.S.L24	-0.1126	0.547	-0.206	0.837	-1.185	0.960
ma.S.L36	-0.3066	0.437	-0.701	0.483	-1.164	0.550
ma.S.L48	0.1311	0.386	0.339	0.734	-0.626	0.889
ma.S.L60	-0.1692	0.550	-0.308	0.758	-1.247	0.909
sigma2	1.295e+05	5.47e-06	2.37e+10	0.000	1.3e+05	1.3e+05

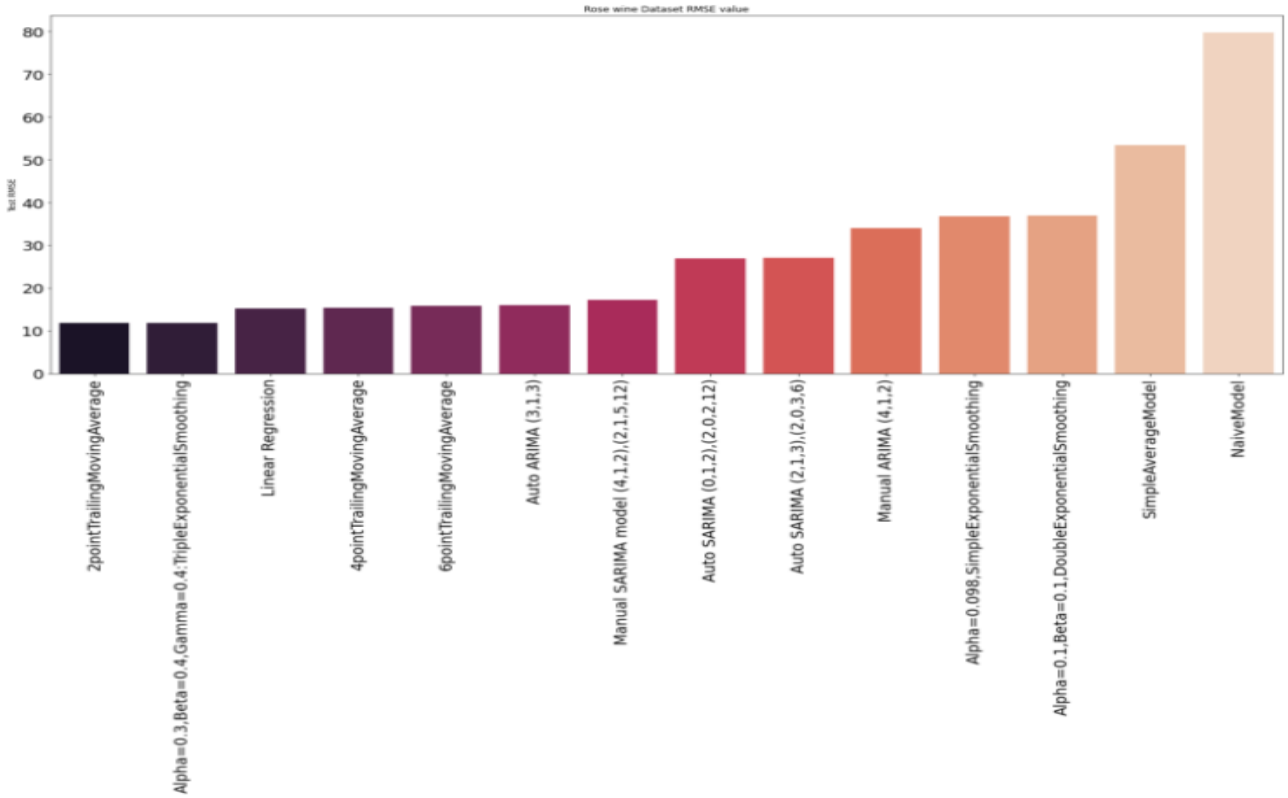
Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	4.18
Prob(Q):	0.97	Prob(JB):	0.12
Heteroskedasticity (H):	0.31	Skew:	0.54
Prob(H) (two-sided):	0.02	Kurtosis:	3.79



For Manual_SARIMA (4,1,2),(2, 1, 5, 12) forecast on the Test Data, RMSE is 348.604
For Manual_SARIMA (4,1,2),(2, 1, 5, 12) forecast on the Test Data, MAPE is 11.210

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.

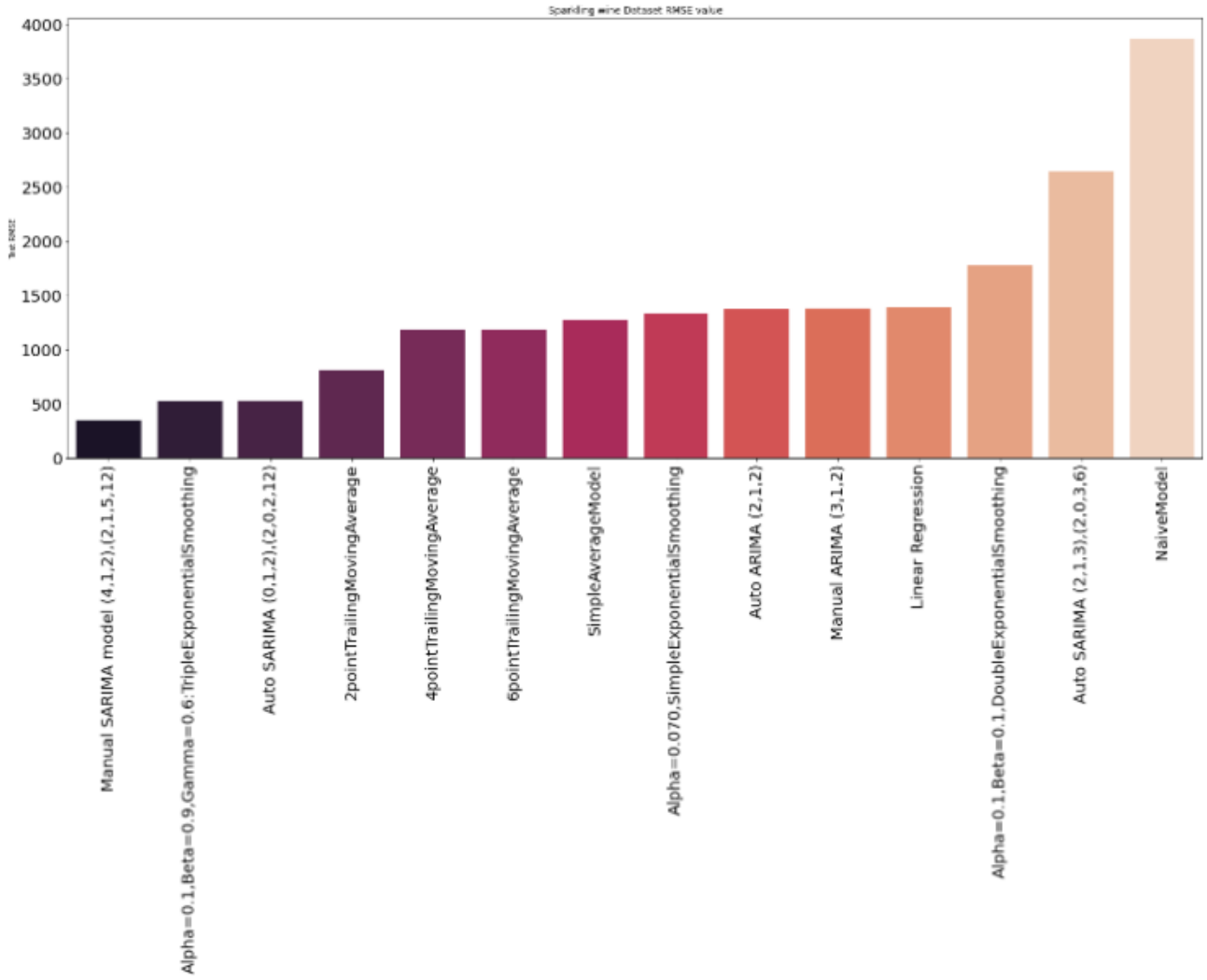
RMSE plot for the Rose wine Dataset:



RMSE table for the Rose wine Dataset:

	Test RMSE	Test MAPE
2pointTrailingMovingAverage	11.801043	13.56
Alpha=0.3,Beta=0.4,Gamma=0.4:TripleExponentialSmoothing	11.827223	18.66
Linear Regression	15.268955	22.82
4pointTrailingMovingAverage	15.367212	19.97
6pointTrailingMovingAverage	15.862350	21.49
Auto ARIMA (3,1,3)	15.986441	26.08
Manual SARIMA model (4,1,2),(2,1,5,12)	17.238792	26.45
Auto SARIMA (0,1,2),(2,0,2,12)	26.928361	46.60
Auto SARIMA (2,1,3),(2,0,3,6)	27.124836	47.31
Manual ARIMA (4,1,2)	33.950457	58.42
Alpha=0.098,SimpleExponentialSmoothing	36.796241	63.88
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	36.879551	63.70
SimpleAverageModel	53.460570	94.93
NaiveModel	79.718773	145.10

RMSE plot for the Sparkling wine Dataset:



	Test RMSE	Test MAPE
Manual SARIMA model (4,1,2),(2,1,5,12)	348.604318	11.21
Alpha=0.1,Beta=0.9,Gamma=0.6:TripleExponentialSmoothing	520.011735	18.27
Auto SARIMA (0,1,2),(2,0,2,12)	526.476536	18.48
2pointTrailingMovingAverage	811.178937	19.30
4pointTrailingMovingAverage	1184.213295	35.81
6pointTrailingMovingAverage	1184.213295	43.94
SimpleAverageModel	1275.081804	38.90
Alpha=0.070,SimpleExponentialSmoothing	1338.008384	47.11
Auto ARIMA (2,1,2)	1374.037009	48.33
Manual ARIMA (3,1,2)	1379.049123	49.32
Linear Regression	1389.135175	50.15
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	1777.734773	67.16
Auto SARIMA (2,1,3),(2,0,3,6)	2643.864992	96.72
NaiveModel	3864.279352	152.87

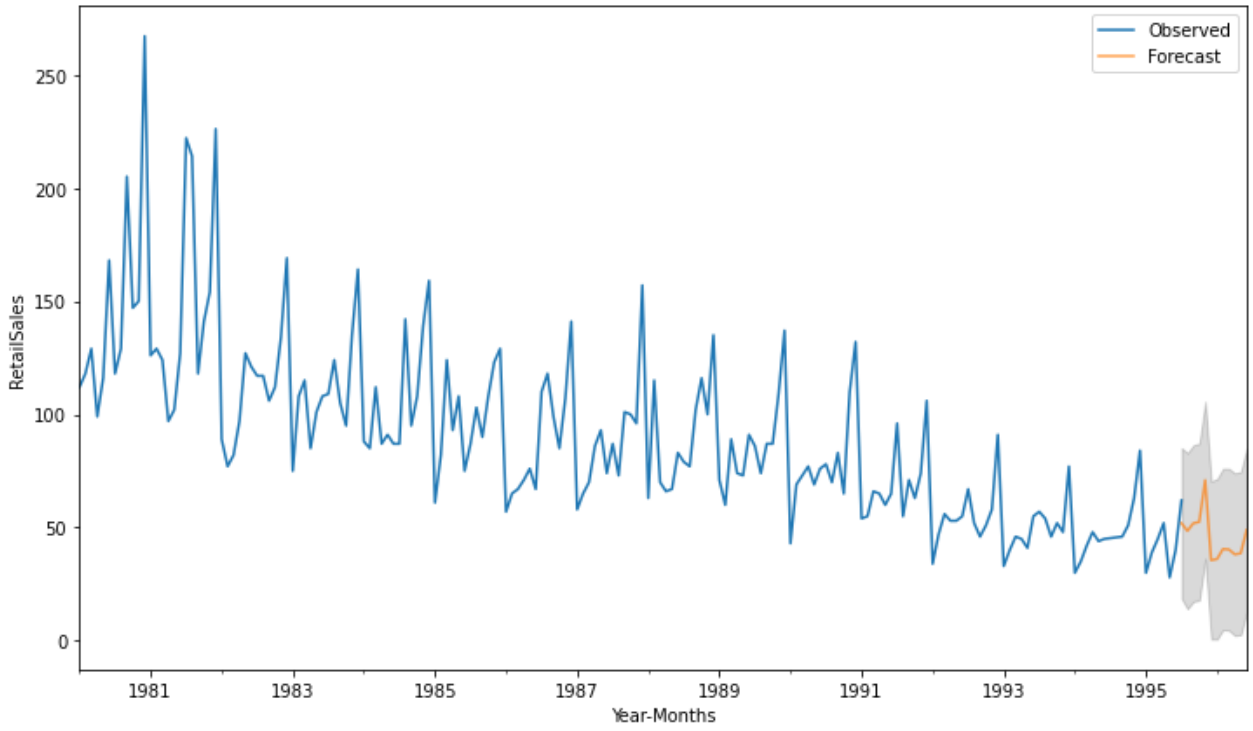
9. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

- In this particular we have built several models and went through a model building exercise. This particular exercise has given us an idea as to which particular model gives us the least error on our test set for this data.
- But in Time Series Forecasting, we need to be very vigil about the fact that after we have done this exercise, we need to build the model on the whole data. Remember, the training data that we have used to build the model stops much before the data ends.
- In order to forecast using any of the models built, we need to build the models again (this time on the complete data) with the same parameters. For this we will go ahead and build only the top 2 models which gave us the best accuracy (least RMSE).
- For rose wine dataset the Triple Exponential Smoothing (or) Holts winters method and SARIMA model is having low RMSE value. Here we will go for both method and we will check the future 12 months prediction.

Model 1: SARIMA model for Rose wine Dataset:

```
full_data_model = sm.tsa.statespace.SARIMAX(df_rose_model2['Rose'],
                                order=(0,1,2),
                                seasonal_order=(2, 0, 2, 6),
                                enforce_stationarity=False,
                                enforce_invertibility=False)
results_full_data_model = full_data_model.fit(maxiter=1000)
print(results_full_data_model.summary())

prediction1 = results_full_data_model.get_forecast(steps=12)
```

```
                                 SARIMAX Results
==========================================================================================
Dep. Variable:                               Rose   No. Observations:                  187
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 6)   Log Likelihood                -734.147
Date:                           Sun, 28 Mar 2021   AIC                           1482.294
Time:                                   16:12:53   BIC                           1504.286
Sample:                               01-01-1980   HQIC                          1491.218
                                    - 07-01-1995
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.7297      0.070    -10.350      0.000      -0.868      -0.592
ma.L2         -0.1899      0.066     -2.883      0.004      -0.319      -0.061
ar.S.L6       -0.0496      0.029     -1.687      0.092      -0.107       0.008
ar.S.L12       0.8766      0.030     29.411      0.000       0.818       0.935
ma.S.L6        0.1746      0.235      0.744      0.457      -0.286       0.635
ma.S.L12      -0.7825      0.194     -4.033      0.000      -1.163      -0.402
sigma2       283.5320     58.864      4.817      0.000     168.161     398.903
==========================================================================================
Ljung-Box (L1) (Q):                   0.19   Jarque-Bera (JB):                297.19
Prob(Q):                              0.66   Prob(JB):                         0.00
Heteroskedasticity (H):               0.17   Skew:                             0.45
Prob(H) (two-sided):                  0.00   Kurtosis:                         9.40
==========================================================================================
```




```
rmse = mean_squared_error(df_rose['Rose'],results_full_data_model.fittedvalues,squared=False)
print('The RMSE of the Full Rose wine Model in SARIMA',rmse)
```

```
The RMSE of the Full Rose wine Model in SARIMA 28.050813052318645
```

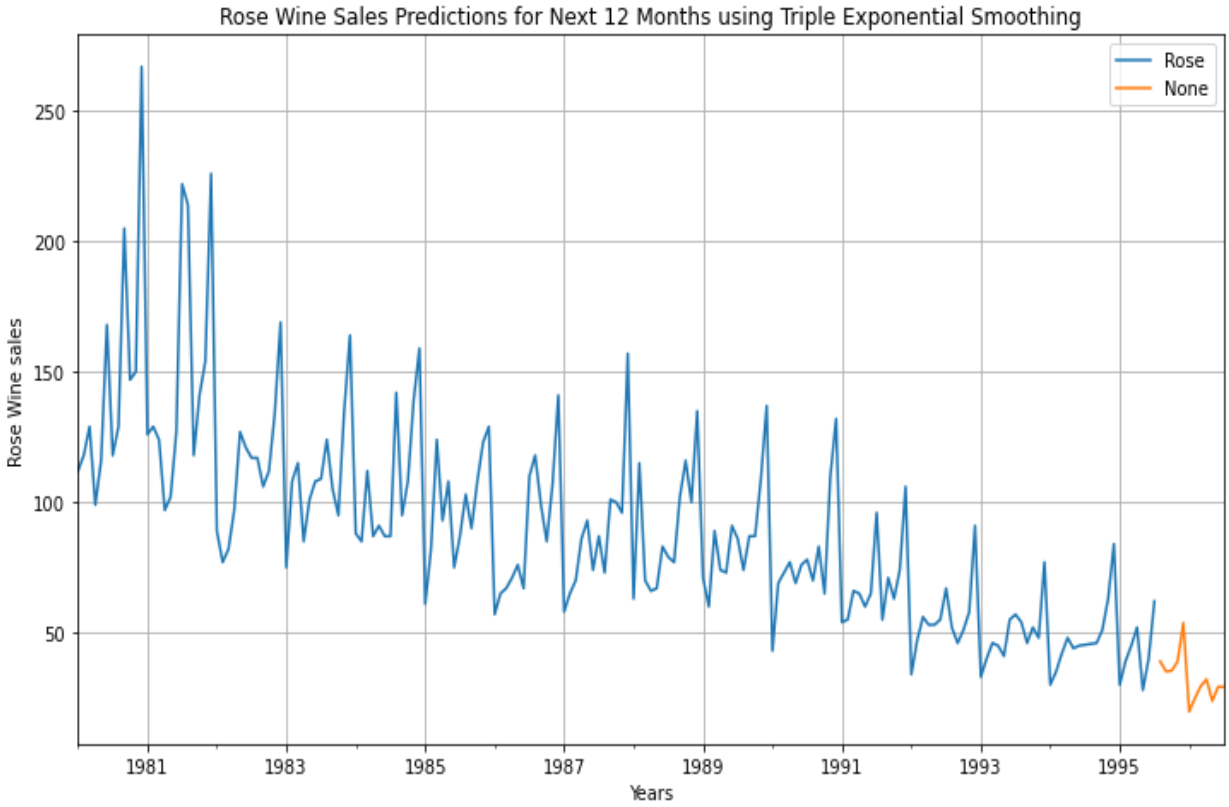
Model 2: Triple Exponential Smoothing model for Rose wine Dataset:

Fitting the model and predicting the forecast for 12 months

```
df_rose_model = ExponentialSmoothing(df_rose_model,trend='additive',seasonal='multiplicative').fit(smoothing_level=0.3,
                                                                                                   smoothing_trend=0.4,
                                                                                                   smoothing_seasonal=0.4)
prediction = df_rose_model.forecast(steps=12)

RMSE_rose_model = metrics.mean_squared_error(df_rose['Rose'],df_rose_model.fittedvalues,squared=False)
```

The Smoothing parameters are

```
{'smoothing_level': 0.3,
 'smoothing_trend': 0.4,
 'smoothing_seasonal': 0.4,
 'damping_trend': nan,
 'initial_level': 47.91744149392783,
 'initial_trend': 0.3135190639361388,
 'initial_seasons': array([2.32638821, 2.43334074, 2.54550317, 1.94840495, 2.27475456,
        3.03668439, 3.08392563, 3.3179567 , 2.81243263, 2.88303416,
        3.05978364, 4.77440092]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```




```
The RMSE score for Triple Exponential Smoothing is: 21.97
```

Inference from the Final Models of Rose wine Data:

- Both the model which has performed well in test sets with minimum RMSE values, but in full 12 months predications the Triple Exponential Smoothing model performs well in Prediction.
- The RMSE value for Triple Exponential smoothing is 21.97, but for SARIMA model the RMSE is 28.05.
- On comparing both the models both the Triple Exponential Smoothing and SARIMA model are considered in predicting the future for Rose wine dataset.

Model 1.1: SARIMA model for Sparkling wine Dataset:

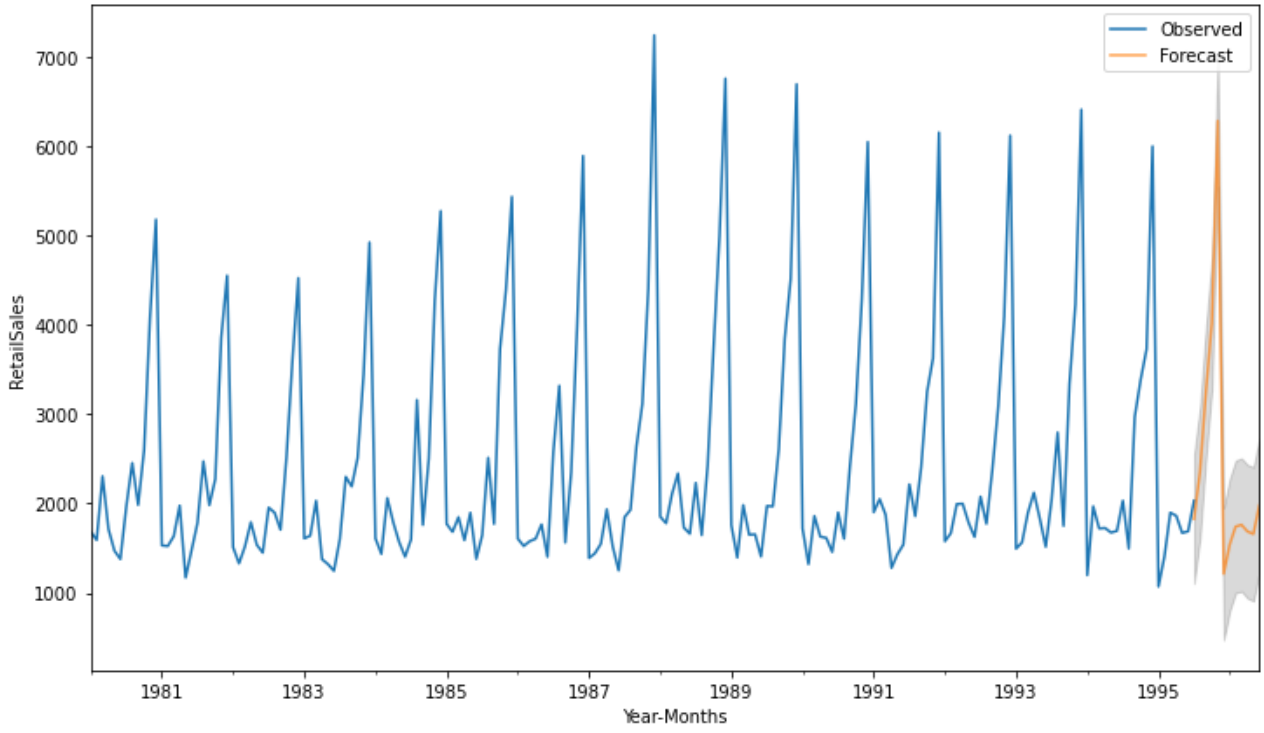
```
full_data_model = sm.tsa.statespace.SARIMAX(df_rose_model2['Rose'],
                                order=(0,1,2),
                                seasonal_order=(2, 0, 2, 6),
                                enforce_stationarity=False,
                                enforce_invertibility=False)
results_full_data_model = full_data_model.fit(maxiter=1000)
print(results_full_data_model.summary())

prediction1 = results_full_data_model.get_forecast(steps=12)
```

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	187
Model:	SARIMAX(4, 1, 2)x(2, 0, 2, 6)	Log Likelihood	-1249.372
Date:	Sun, 28 Mar 2021	AIC	2520.744
Time:	16:13:04	BIC	2555.238
Sample:	01-01-1980 - 07-01-1995	HQIC	2534.741
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.0386	0.741	0.052	0.958	-1.414	1.491
ar.L2	-0.1134	0.134	-0.846	0.397	-0.376	0.149
ar.L3	0.0220	0.109	0.201	0.840	-0.192	0.236
ar.L4	-0.1074	0.093	-1.151	0.250	-0.290	0.075
ma.L1	21.8537	404.479	0.054	0.957	-770.911	814.619
ma.L2	-25.3912	448.133	-0.057	0.955	-903.716	852.934
ar.S.L6	0.0092	0.017	0.546	0.585	-0.024	0.042
ar.S.L12	1.0177	0.011	96.183	0.000	0.997	1.038
ma.S.L6	-0.0976	0.097	-1.008	0.313	-0.287	0.092
ma.S.L12	-0.6562	0.070	-9.381	0.000	-0.793	-0.519
sigma2	211.2209	7454.158	0.028	0.977	-1.44e+04	1.48e+04

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	38.76
Prob(Q):	0.97	Prob(JB):	0.00
Heteroskedasticity (H):	1.32	Skew:	0.56
Prob(H) (two-sided):	0.30	Kurtosis:	5.06




```
rmse = mean_squared_error(df_spark['Sparkling'],results_full_data_model.fittedvalues,squared=False)
print('The RMSE of the Full Sparkling wine Model in SARIMA',rmse)
```

```
The RMSE of the Full Sparkling wine Model in SARIMA 533.246413330493
```

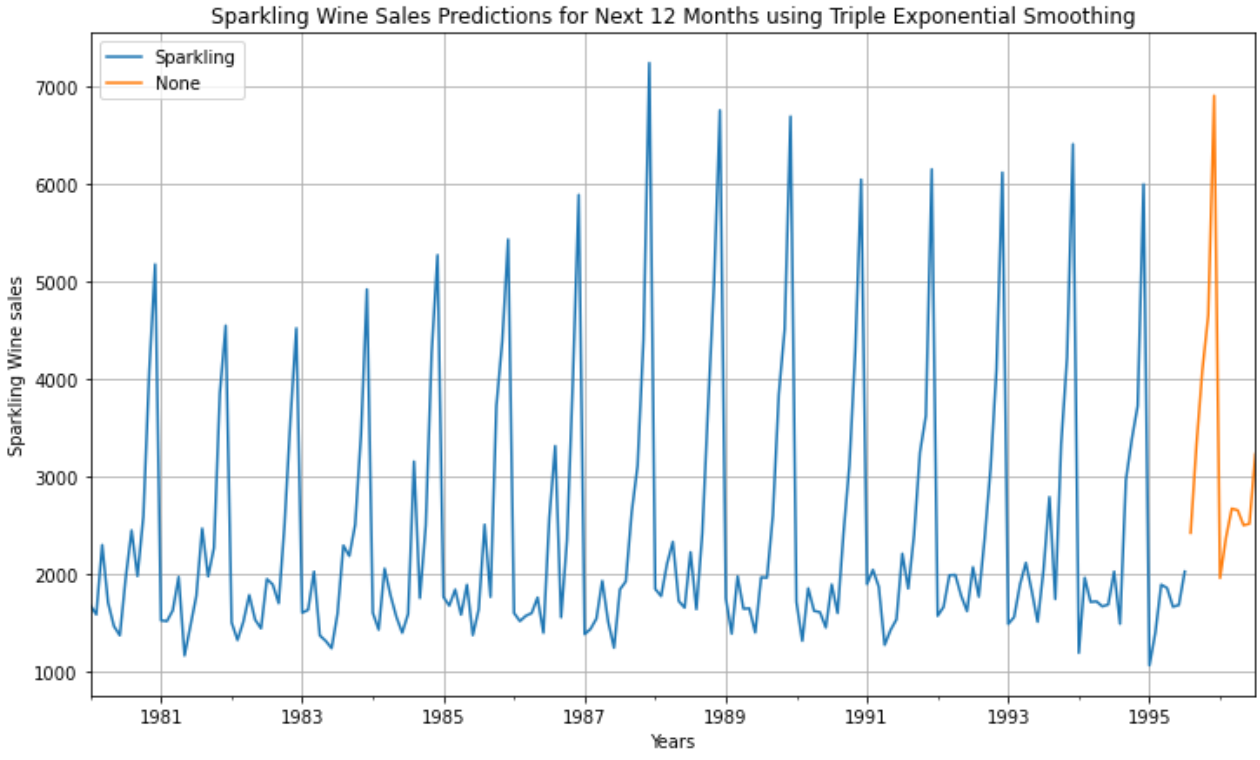
Model 2.1: Triple Exponential Smoothing model for Sparkling wine Dataset:

Fitting the model and predicting the forecast for 12 months

```
df_spark_model = ExponentialSmoothing(df_spark_model,trend='additive',seasonal='additive').fit(smoothing_level=0.1,
                                                                                               smoothing_trend=0.9,
                                                                                               smoothing_seasonal=0.6)
prediction = df_spark_model.forecast(steps=12)

RMSE_spark_model = metrics.mean_squared_error(df_spark['Sparkling'],df_spark_model.fittedvalues,squared=False)
```

The Smoothing parameters are

```
{'smoothing_level': 0.1,
 'smoothing_trend': 0.9,
 'smoothing_seasonal': 0.6,
 'damping_trend': nan,
 'initial_level': 1417.45818473436,
 'initial_trend': 13.9090842925432,
 'initial_seasons': array([ 198.01624311,  119.09780667,  637.27575775,  306.06125755,
        -109.80767012, -160.84986074,  311.04764462,  804.34293346,
         352.12500494,  923.34537014, 2426.88747331, 3485.27105127]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```




```
The RMSE score for Full Sparkling wine model Triple Exponential Smoothing is: 465.5687
```

Inference from the Final Models of Sparkling wine Data:

- Both the model which has performed well in test sets with minimum RMSE values, but in full 12 months predications the SARIMA model performs well in Prediction.
- The RMSE value for Triple Exponential smoothing is 465.56, but for SARIMA model the RMSE is 533.24.
- On comparing both the models performs well we will consider both the Triple Exponential Smoothing and SARIMA model for Sparkling wine dataset in predicting the future.

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

Business Insights and Suggestions for Rose wine Dataset:

- The graph showing the predictions for 12 months period from start='1995-07- 01', end='1996-07-01' shows that predicted numbers will follow the past trend years.
- **BEST CASE SENARIO**: From the SARIMA model highest sales will touch 75 units and lowest sales can be 40 units over a period of 12 months into the future.
- **WORST CASE SENARIO**: From the Triple Exponential Smoothing model highest sales will touch 55 units and lowest sales can be 20 units over a period of 12 months into the future.
- 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. December month shows the highest Sales across the year while the value has come down through the years from 1980-1994.
- The Sales of Rose wine is seasonal and also has trend, **hence the company cannot have the same stock through the year.**
- The company should use the prediction results and capitalize on the **high demand and Low demand seasons to ensure the source and supply at demands situation.**
- In off seasons like months of Feb, March, April, May, company can give good **discounts or some promotional offers** like free movies, club memberships etc on purchase of certain units of wine to boost the sales.
- Company may also **introduce more variants of Rose wine** as per customers feedback on taste of this particular wine.

Business Insights and Suggestions for Sparkling wine Dataset:

- The graph showing the predictions for 12 months period from start='1995-07- 01', end='1996-07-01' shows that predicted numbers will follow the past trend years.
- **BEST CASE SENARIO**: From the Triple Exponential Smoothing model highest sales will touch 7000 units and lowest sales can be 2000 units over a period of 12 months into the future.
- **WORST CASE SENARIO**: From the SARIMA model highest sales will touch 6400 units and lowest sales can be 1200 units over a period of 12 months into the future.
- 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 Sparling 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 company should use the prediction results and **capitalize on the high demand seasons and ensure to source and supply the high demand.**
- In off seasons months of May, June, July, company can give good discounts, but from the Prediction results, **the company no need to spend too much to increase the sales of the Sparkling wine.**