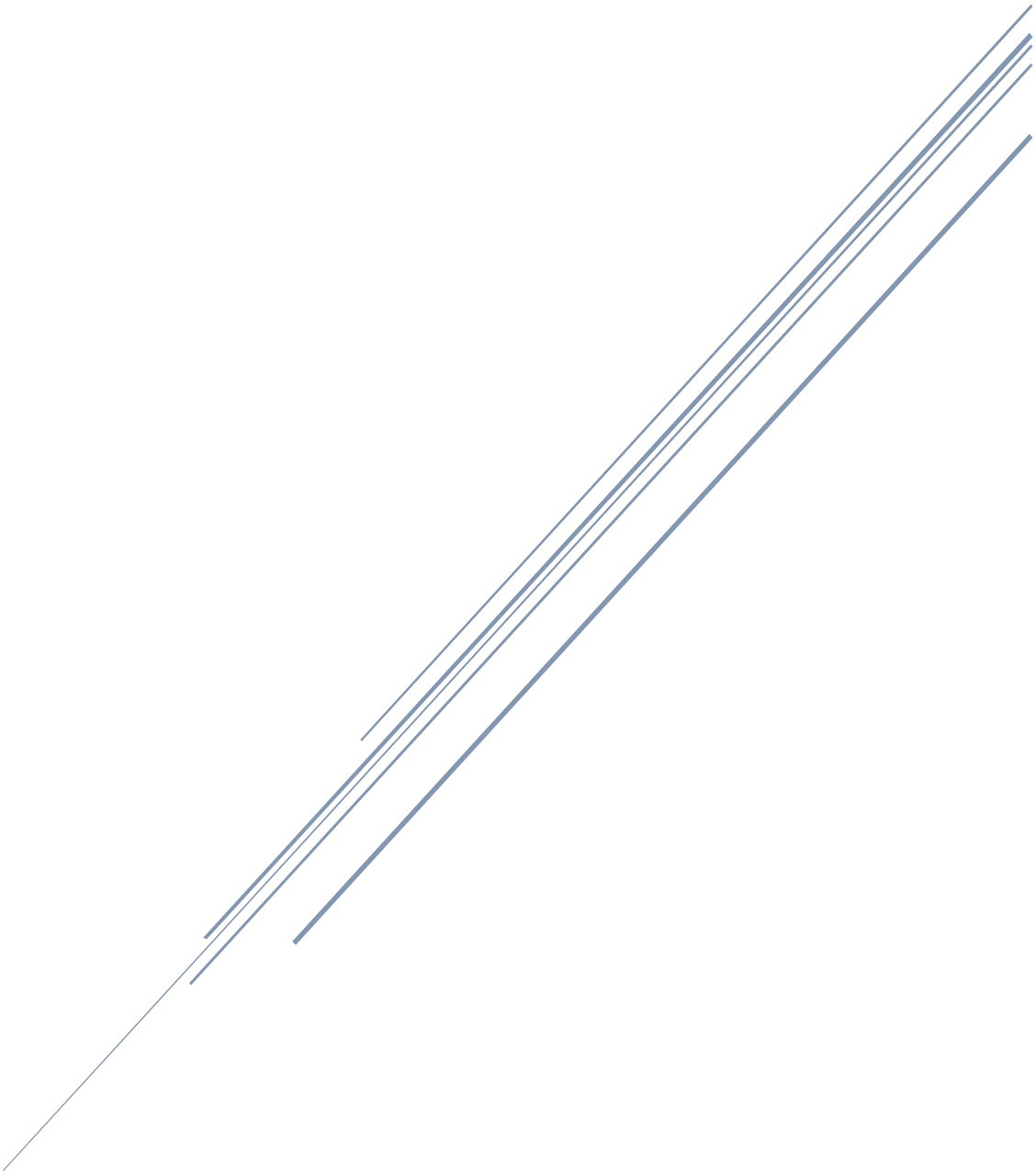
TIME SERIES FORCASTING - BUSINESS PROJECT

By Vinish Vincent



Great Learnings
PGP-DSBA - 23

Table of content

Table of Content

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**

Please do perform the following questions on each of these two data sets separately.

1.1 Read the data as an appropriate Time Series data and plot the data.
1.2 Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.
1.3 Split the data into training and test. The test data should start in 1991.
1.4 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.
1.5 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.
1.6 Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.
1.7 Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.
1.8 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.
1.9 Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

Problem Statement Part 1 – Sparkling Wine:

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

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

Solution:

Table 1: **Data top 5 records**

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

Table 2: **Data last 5 records**

YearMonth	Sparkling
1995-03-01	1897
1995-04-01	1862
1995-05-01	1670
1995-06-01	1688
1995-07-01	2031

Table 3: **Data types and info**

Data Type

```
Sparkling    int64
dtype: object
```

Data Info

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
Data columns (total 1 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Sparkling  187 non-null    int64
dtypes: int64(1)
memory usage: 2.9 KB
```

Table 4: **Data description**

	Sparkling
count	187.000000
mean	2402.417112
std	1295.111540
min	1070.000000
25%	1605.000000
50%	1874.000000
75%	2549.000000
max	7242.000000

The dataset consists of a single column named **"Sparkling,"** which contains **integer** values. This data represents a **time series of sparkling wine sales**, with the values recorded as counts or quantities

The data covers a **time period from January 1980 to July 1995, spanning 187 months**. This dataset represents a historical time series, allowing for the analysis of trends and patterns in sparkling wine data over this time frame

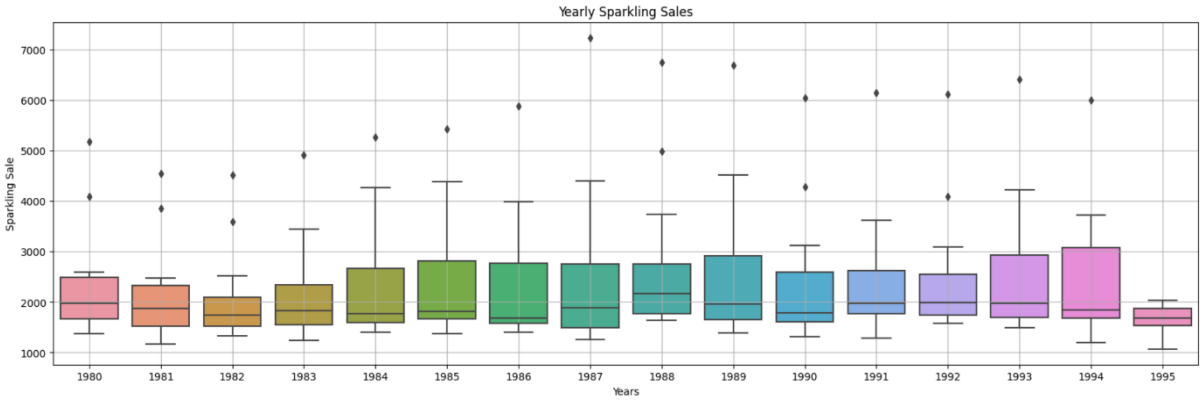
The summary statistics reveal a wide range of sparkling wine counts, with a **minimum of 1070** and a **maximum of 7242**

The dataset have **no null values** and we can proceed with EDA

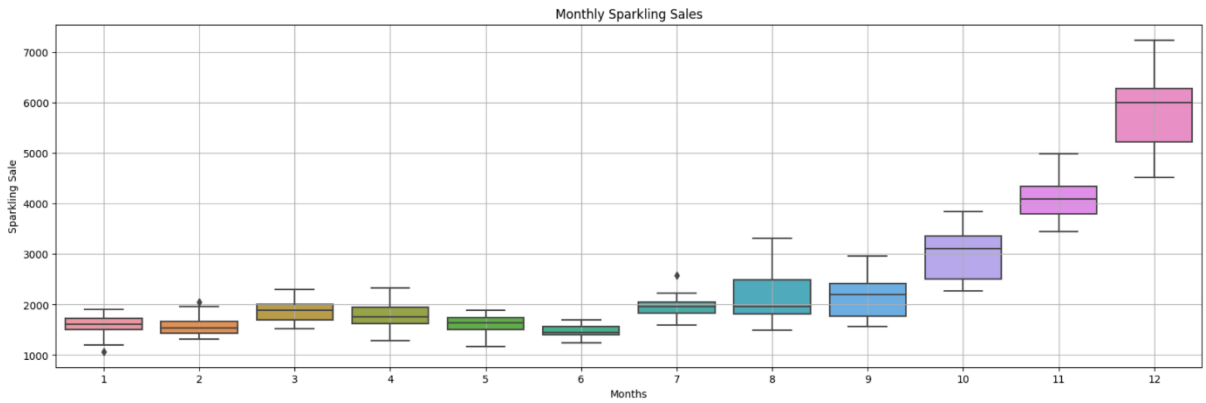
1.2 Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

Solution:

Graph 1: **Boxplot of Sales across Years**



Graph 2: **Boxplot of Sales across months**



Monthly sales of sparkling wine shows seasonal patterns, with **higher sales** typically occurring in the **later months of the year** (e.g., November and December), possibly due to holiday celebrations. Conversely, **sales** are **lower** during the **early months of the year** (e.g., January and February)

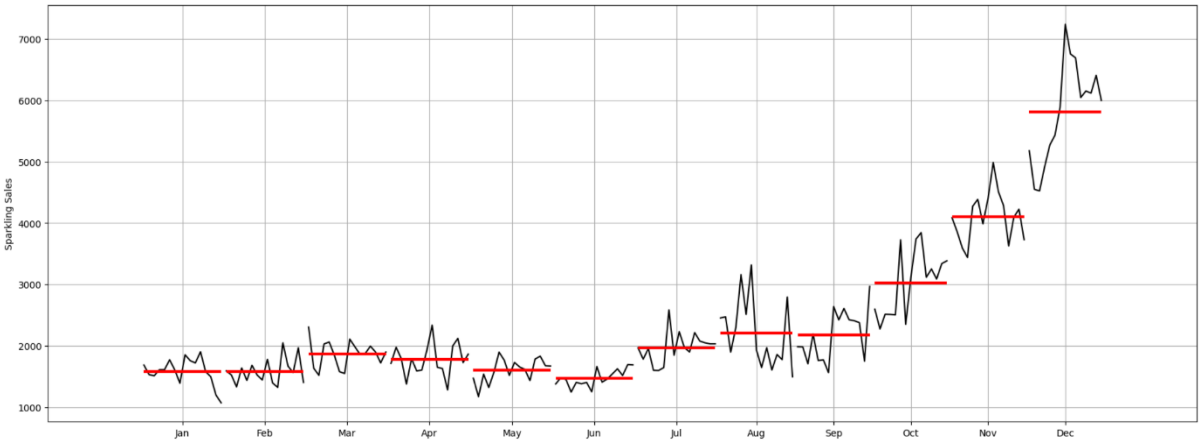
Over the years, there appears to be a **general increase in sales**, with **some fluctuations**. This suggests that there may be a long-term growth trend in sparkling wine consumption, possibly driven by changing consumer preferences or market dynamics

Table 5: **Pivot Table Months vs Year**

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

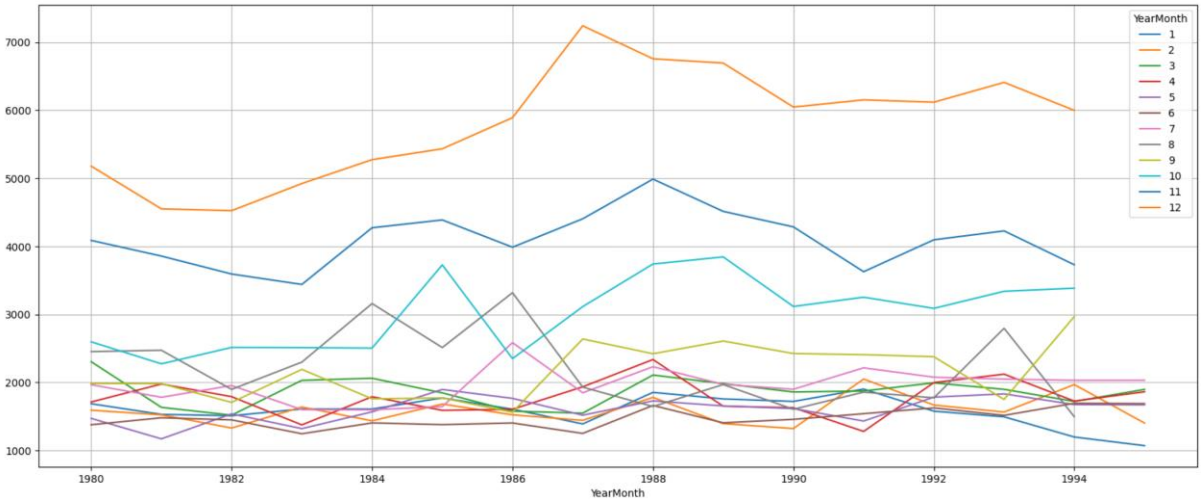
Across the years, the **months of July and October** consistently show **higher sales** figures compared to other months, indicating a potential recurring pattern or specific factors driving sales during these months

Graph 3: **Monthly Plot for Time Series**



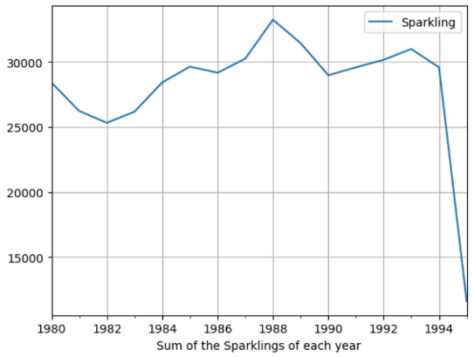
- The red lines indicate the average sales for the month

Graph 4: **Monthly Sales across Years**

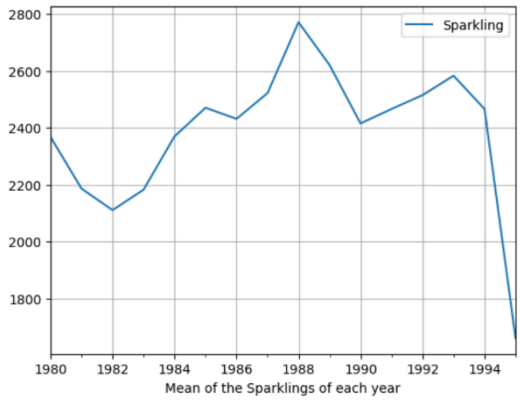


The first three months with **highest sales of Sparkling wine** is in month of **December, November, and October**

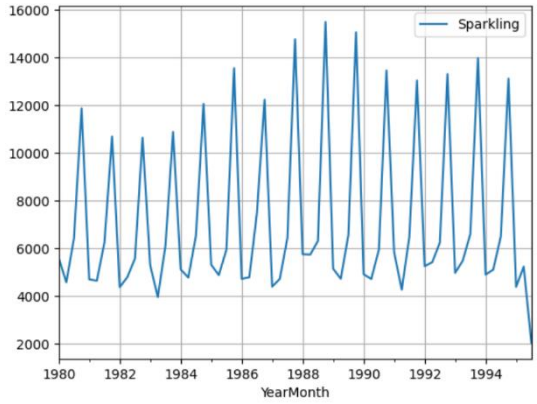
Graph 5: **Sum of the Sparkling wine of each year**



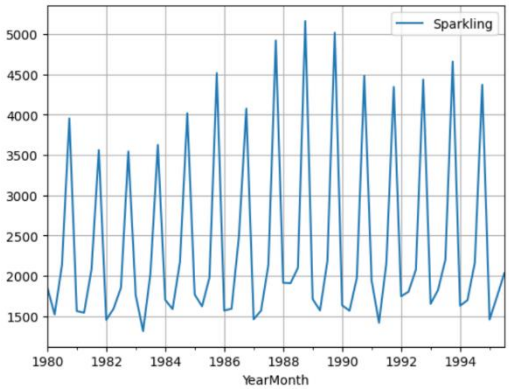
Graph 6: **Mean of the Sparkling wine of each year**



Graph 7: **Quarterly sum of the Sparkling wine of each year**



Graph 8: **Quarterly mean of the Sparkling wine of each year**



The **yearly sum of sparkling wine sales** shows **fluctuations** over the years, with **1984** having the **highest total sales at 28,431** and **1982** having the **lowest at 25,321**. This insight highlights the variations in annual sales

Examining quarterly sales reveals more granular trends. For example, the **first quarter of 1980** had **total sales** of **5,581**, indicating a strong start to the year. This insight allows for a closer look at seasonality within each year

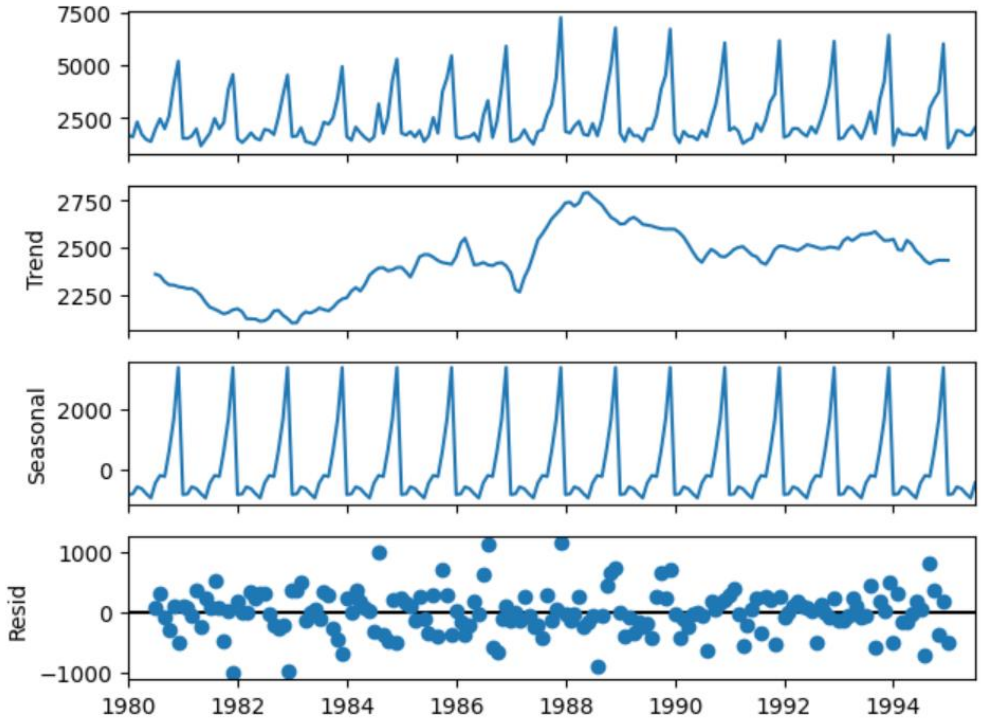
The mean quarterly sales data gives an average sales figure for each quarter. For instance, the **average sales in the first quarter of 1980 were 1,860.33**, providing insights into sales patterns within quarters across the years

Decomposition of Time Series

Here we use additive model and multiplicative model

Additive Model

Graph 9: **Decomposition of Time series using Additive Model**



It seen that Time series is divided into 3 parts **Trend, Seasonal and Residual [Noise]**. In seasonal it is seen that a **constant increase above 2500 points**. This shows there is constant factor which is getting added in the series.

Residual is between 1000 to -1000 points, but more concentrated over 0 points. There are some outliers seen in the residual part.

Table 6: **Trend Aspect of Additive Model**

```
Trend
 YearMonth
1980-01-01        NaN
1980-02-01        NaN
1980-03-01        NaN
1980-04-01        NaN
1980-05-01        NaN
1980-06-01        NaN
1980-07-01    2360.67
1980-08-01    2351.33
1980-09-01    2320.54
1980-10-01    2303.58
1980-11-01    2302.04
1980-12-01    2293.79
Name: trend, dtype: float64
```

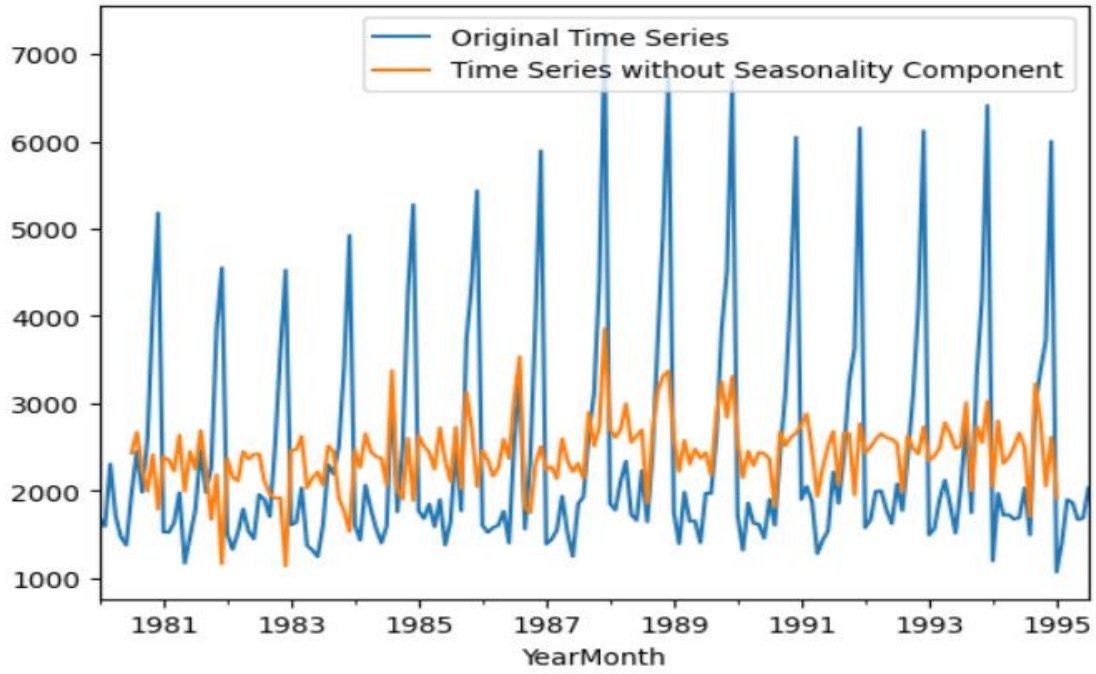
Table 7: **Seasonality Aspect of Additive Model**

```
Seasonality
 YearMonth
1980-01-01    -854.26
1980-02-01    -830.35
1980-03-01    -592.36
1980-04-01    -658.49
1980-05-01    -824.42
1980-06-01    -967.43
1980-07-01    -465.50
1980-08-01    -214.33
1980-09-01    -254.68
1980-10-01     599.77
1980-11-01    1675.07
1980-12-01    3386.98
Name: seasonal, dtype: float64
```

Table 8: **Residual Aspect of Additive Model**

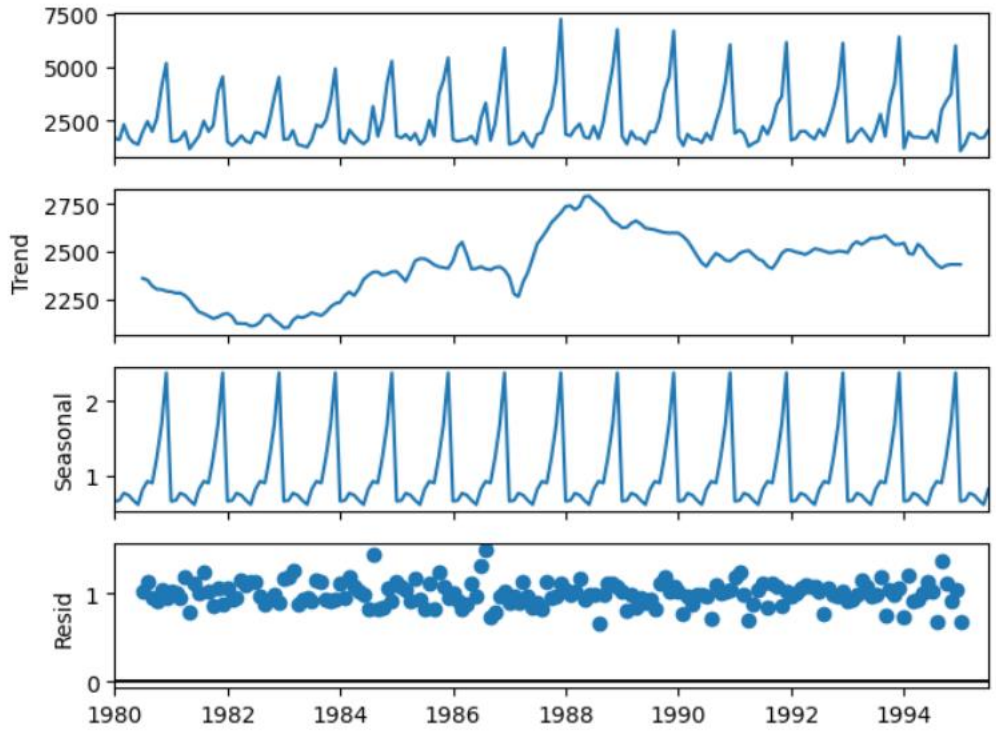
```
Residual
 YearMonth
1980-01-01       NaN
1980-02-01       NaN
1980-03-01       NaN
1980-04-01       NaN
1980-05-01       NaN
1980-06-01       NaN
1980-07-01     70.84
1980-08-01    316.00
1980-09-01    -81.86
1980-10-01   -307.35
1980-11-01    109.89
1980-12-01   -501.78
Name: resid, dtype: float64
```

Graph 10: **Time series without the seasonality component**



Multiplicative model

Graph 11: **Decomposition of Time series using Multiplicative Model**



It seen that Time series is divided into 3 parts Trend Seasonal and Residual [Noise]. In **seasonal** it is seen that a **constant increase above 2 points**. This shows there is constant percentage factor which is getting multiplied in the series.

Residual is more concentrated at 1 points which minimum. There are some outliers seen in the residual part

Table 9: **Trend Aspect of Multiplicative Model**

YearMonth	Trend
1980-01-01	NaN
1980-02-01	NaN
1980-03-01	NaN
1980-04-01	NaN
1980-05-01	NaN
1980-06-01	NaN
1980-07-01	2360.67
1980-08-01	2351.33
1980-09-01	2320.54
1980-10-01	2303.58
1980-11-01	2302.04
1980-12-01	2293.79

Name: trend, dtype: float64

Table 10: **Seasonality Aspect of Multiplicative Model**

YearMonth	Seasonality
1980-01-01	0.65
1980-02-01	0.66
1980-03-01	0.76
1980-04-01	0.73
1980-05-01	0.66
1980-06-01	0.60
1980-07-01	0.81
1980-08-01	0.92
1980-09-01	0.89
1980-10-01	1.24
1980-11-01	1.69
1980-12-01	2.38

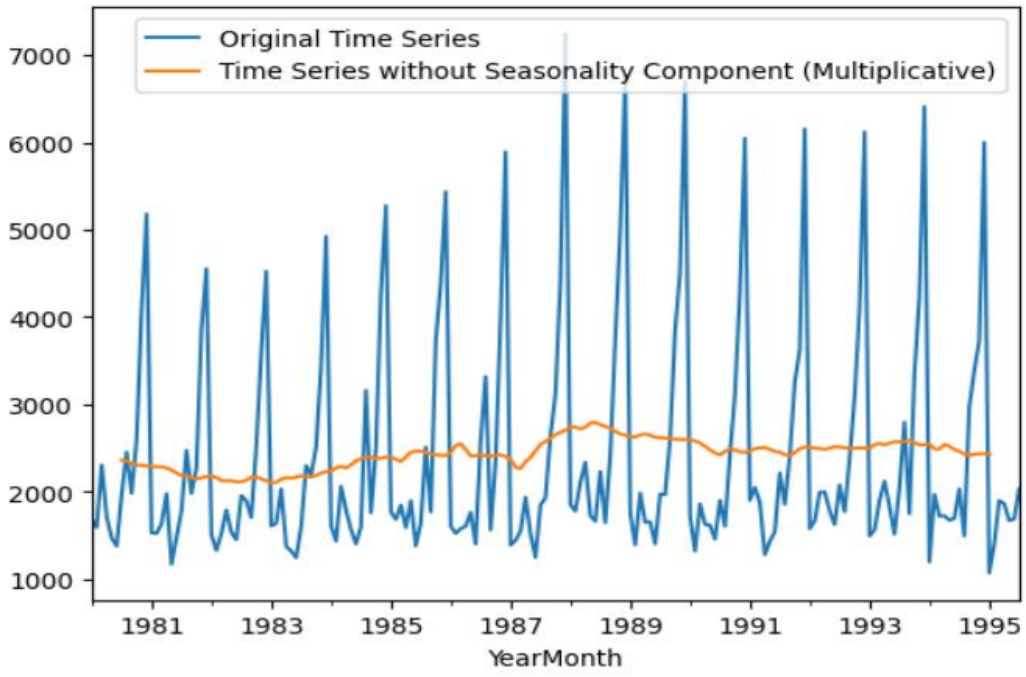
Name: seasonal, dtype: float64

Table 11: **Residual Aspect of Multiplicative Model**

YearMonth	Residual
1980-01-01	NaN
1980-02-01	NaN
1980-03-01	NaN
1980-04-01	NaN
1980-05-01	NaN
1980-06-01	NaN
1980-07-01	1.03
1980-08-01	1.14
1980-09-01	0.96
1980-10-01	0.91
1980-11-01	1.05
1980-12-01	0.95

Name: resid, dtype: float64

Graph 12: **Time series without the seasonality component**



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

Solution:

We take the whole data and divided it into train and test. By indexing the year less 1991 for train dataset, while more than or equal to 1991 as test dataset.

There were 187 data points in the whole dataset, while splitting the training dataset has 132 data points and 55 data points as test data set.

Table 12: **First five rows of Training Data**

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

Table 13: **Last five rows of Training Data**

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

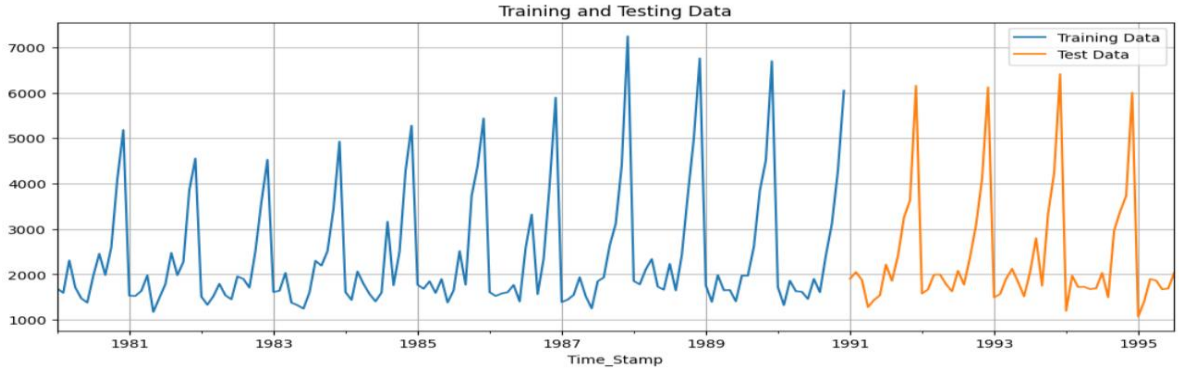
Table 14: **First five rows of Test Data**

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

Table 15: **Last five rows of Test Data**

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

Graph 13: **Graphical representation of the Training and Test data set**



1.4 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

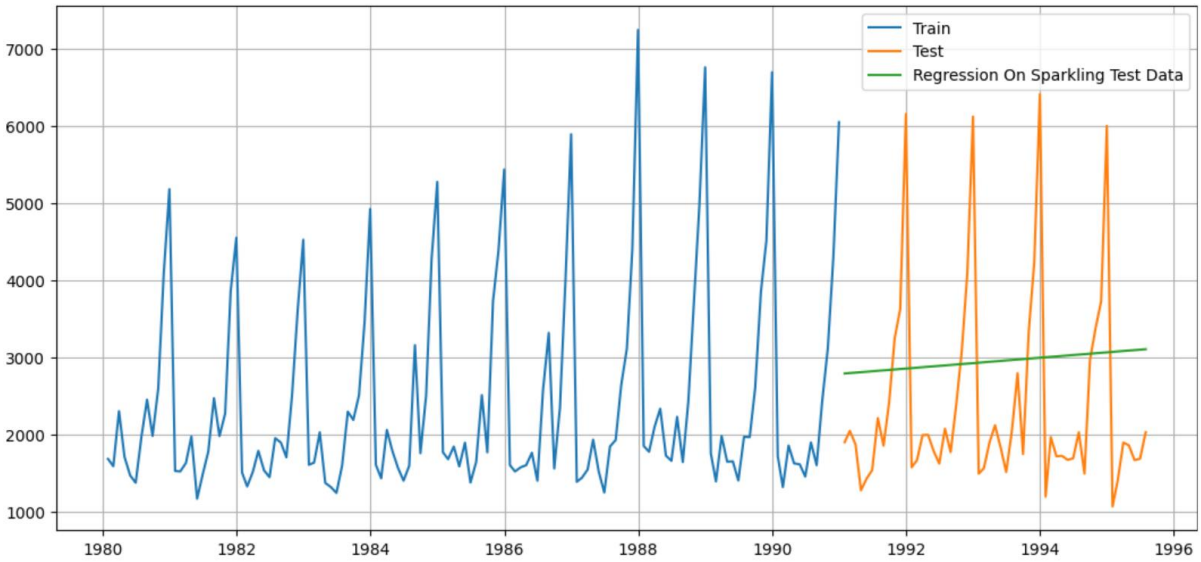
Solution:

Model 1: Linear Regression

We imported Linear Regression from sklearn. This model is based on Linear Regression method to forecast the data

```
Training Time instance
 [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, ]  .....  [128, 129, 130, 131, 132]
Test Time instance
 [133, 134, 135, 136, 137, 138, 139, 140,  .....  [183, 184, 185, 186, 187]
```

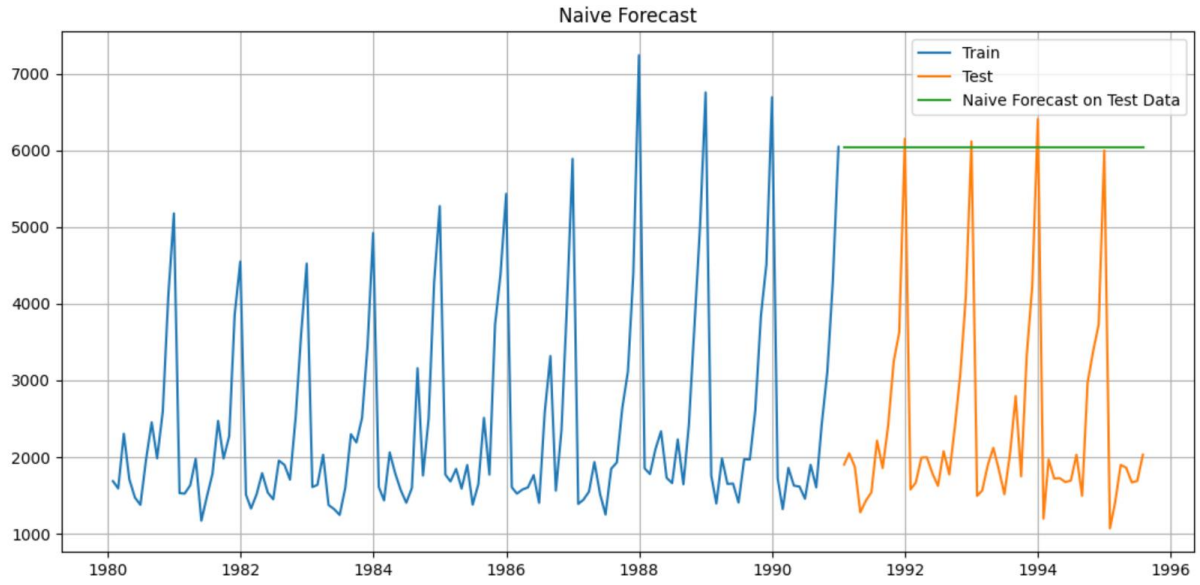
Graph 14: **Linear Regression model on Time Series**



RSME for lr_model1 is: 1389.135174897992

Model 2: Naive Approach

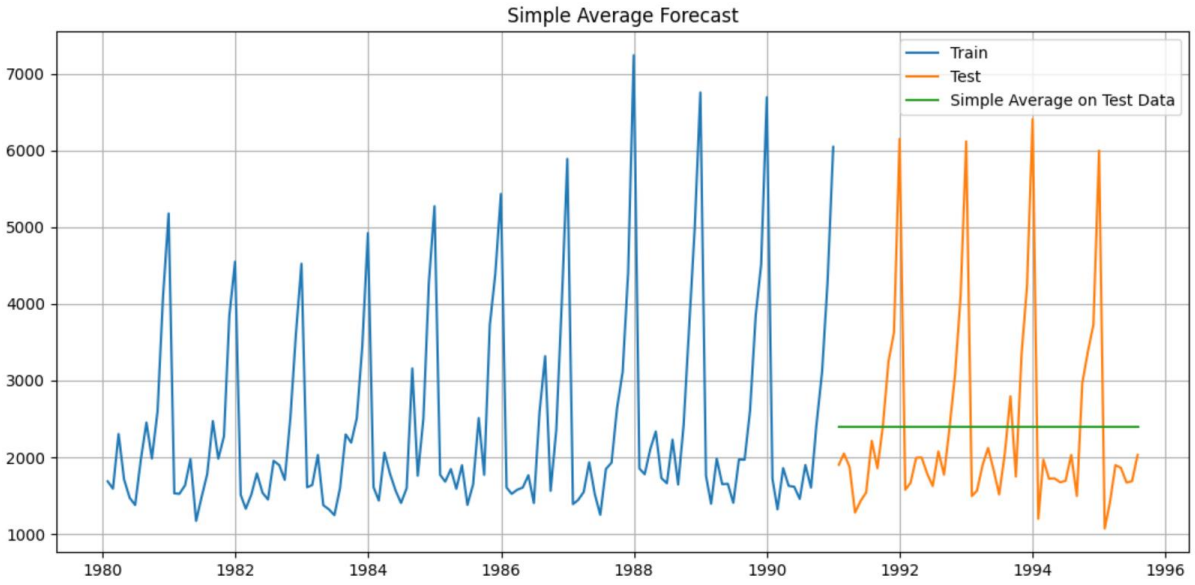
Graph 15: **Naïve Forecast model on Time Series**



RMSE for Naïve model is: 3864.2793518443914

Model 3: Simple Average

Graph 16: **Simple Average model on Time Series**



RMSE for Simple Average model is: 1275.0818036965309

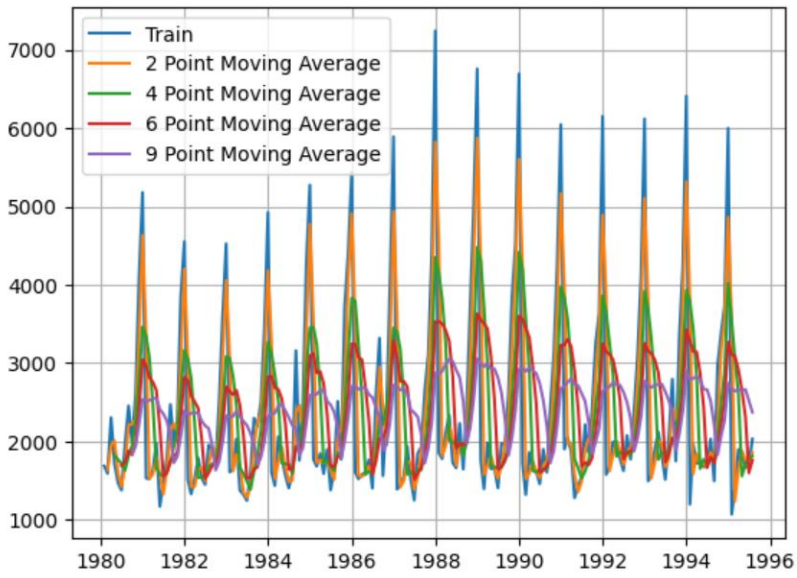
Model 4: Moving Average

This method uses averaging to forecast the values based on window sizes. The window keeps on moving with the size constant for newer points to be forecasted

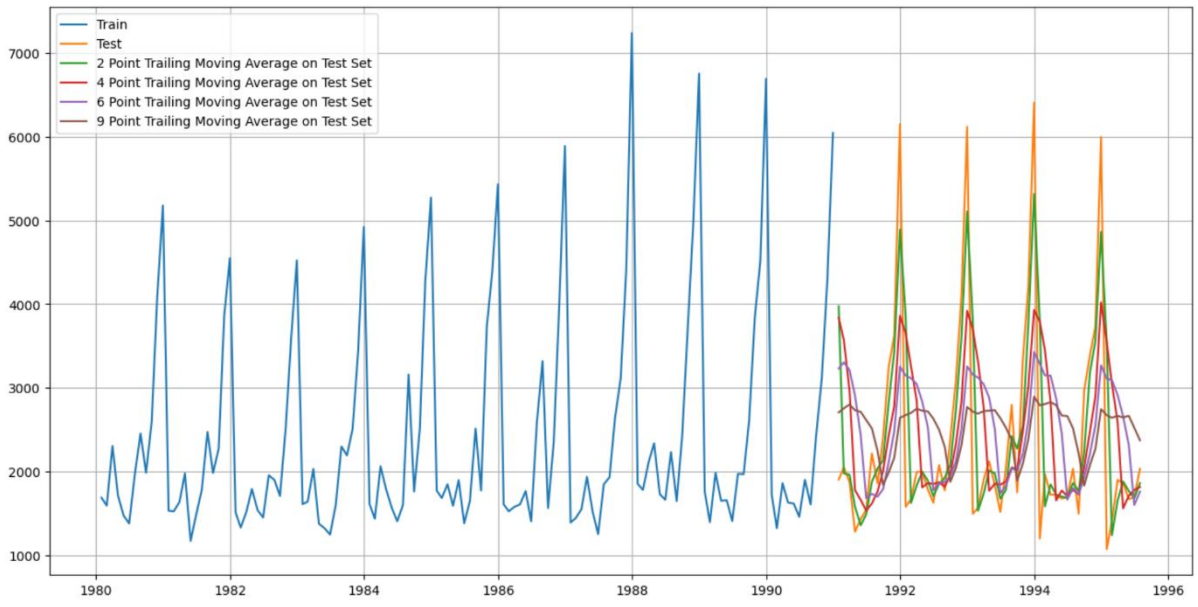
Table 16: **Considering window size as 2, 4, 6, and 9**

Time_Stamp	Sparkling	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1980-01-31	1686	NaN	NaN	NaN	NaN
1980-02-29	1591	1638.5	NaN	NaN	NaN
1980-03-31	2304	1947.5	NaN	NaN	NaN
1980-04-30	1712	2008.0	1823.25	NaN	NaN
1980-05-31	1471	1591.5	1769.50	NaN	NaN

Graph 17: **Moving Average model on Time Series**



Graph 18: **Moving Average model on Time Series (Test Data)**



For 2 point Moving Average Model forecast on the Data, RMSE is 813.401

For 4 point Moving Average Model forecast on the Data, RMSE is 1156.590

For 6 point Moving Average Model forecast on the Data, RMSE is 1283.927

For 9 point Moving Average Model forecast on the Data, RMSE is 1346.278

Method 5: Simple Exponential Smoothing [SES]

Exponential smoothing averages or exponentially weighted moving averages consist of forecast based on previous period's data with exponentially declining influence on the older observations.

Table 17: **Best possible values of parameters in SES**

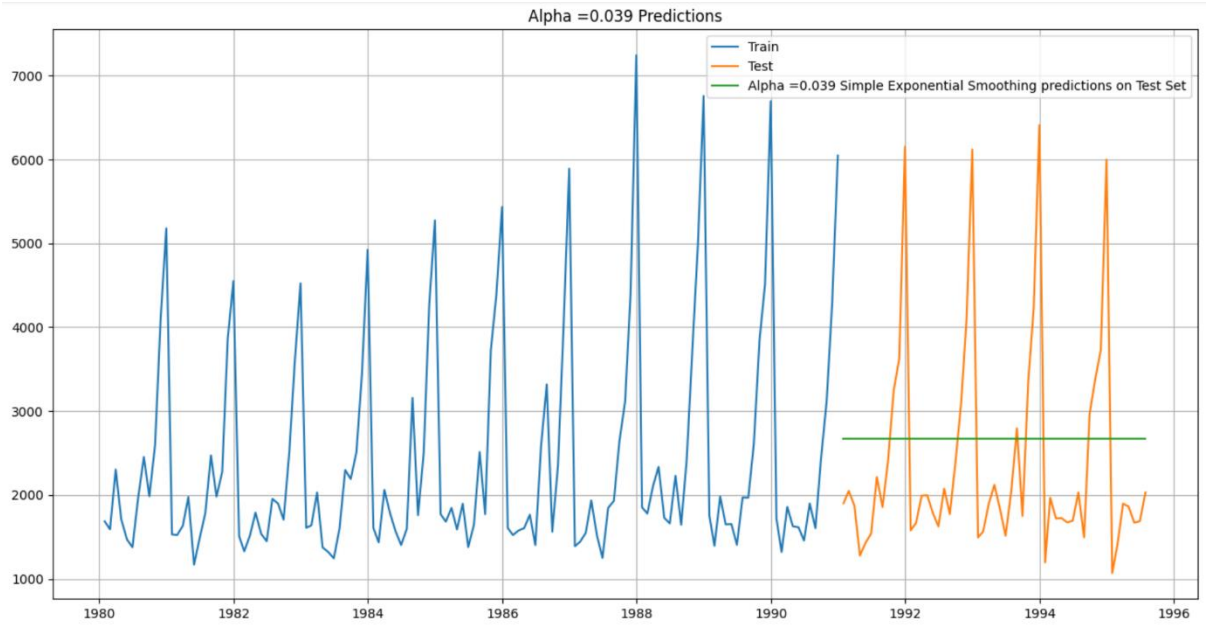
```
{'smoothing_level': 0.03953488372093023,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 1686.0,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

And then using these parameters we predict the values of test dataset

Table 18: **Predicted test values using best parameters in SES**

Time_Stamp	Sparkling	predict
1991-01-31	1902	2676.676366
1991-02-28	2049	2676.676366
1991-03-31	1874	2676.676366
1991-04-30	1279	2676.676366
1991-05-31	1432	2676.676366

Graph 19: **Simple Exponential Smoothing model on Time Series**



For Alpha =0.039 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 1304.927

Method 6: Double Exponential Smoothing (Holt's Model)

This model considers level and trend while forecasting values

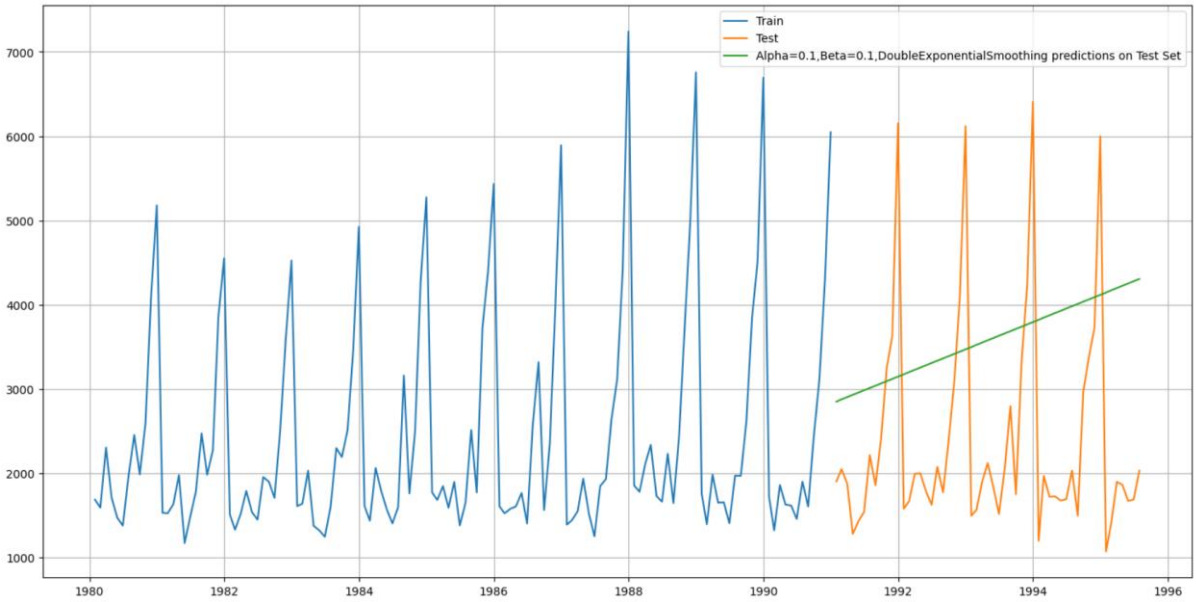
We use a combination of different alpha and beta values and consider the best value and build a model on it

Table 19: **Series of different Alpha and Beta value in DES model with Train and Test RMSE score (sorted)**

	Alpha Values	Beta Values	Train RMSE	Test RMSE
0	0.1	0.1	1382.520870	1778.564670
1	0.1	0.2	1413.598835	2599.439986
10	0.2	0.1	1418.041591	3611.763322
2	0.1	0.3	1445.762015	4293.084674
20	0.3	0.1	1431.169601	5908.185554
...	...	...	...	...
98	1.0	0.9	1985.368445	57823.177011
79	0.8	1.0	1872.711054	57990.117908
89	0.9	1.0	1948.020916	59008.254331
99	1.0	1.0	2077.672157	59877.076519
19	0.2	1.0	2325.013004	60749.773505

100 rows × 4 columns

Graph 20: **Double Exponential Smoothing model with Alpha =0.1, Beta=0.1 on Time Series**



For Alpha =0.1, Beta=0.1, Double Exponential Smoothing Model forecast on the Test Data, RMSE is 1778.564670

Method 7: Triple Exponential Smoothing (Holt - Winter's Model)[TES]

Holt-Winters - ETS(A, A, A) - Holt Winter's linear method with additive errors. Three parameters, Level, Trend and Seasonality are accounted for in this model.

Here it use multiplicative [seasonal], Additive [trend] and residual also.

Here we apply SES on the train and test data. And automatic fit with best possible value of smoothing level (alpha), and no value for trend and seasonality.

Table 20: Best possible values of parameters in TES

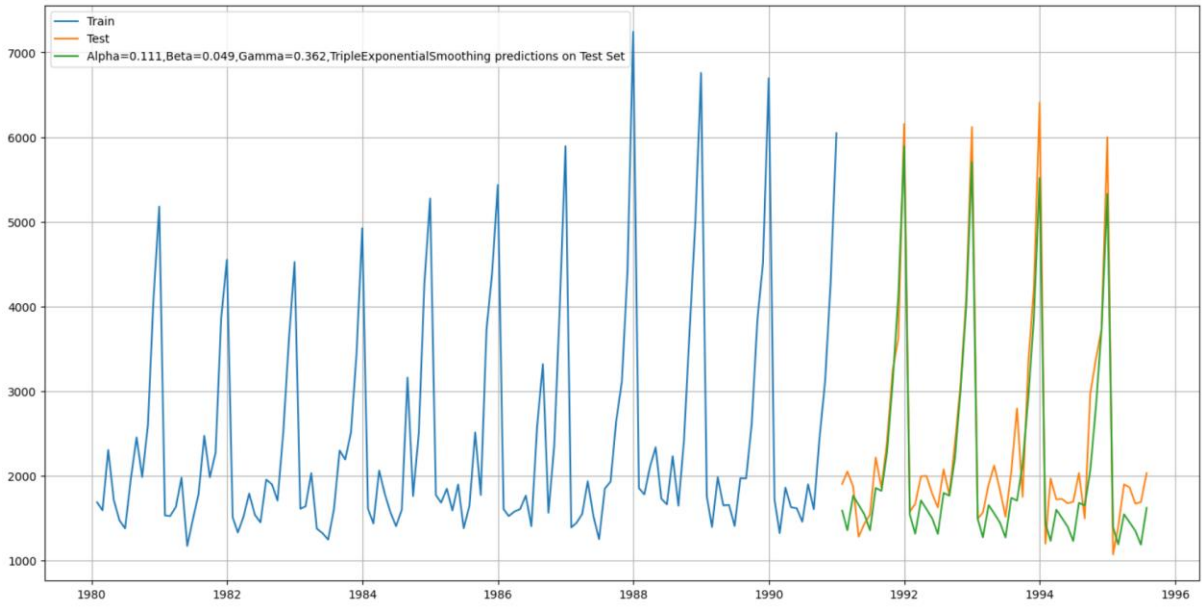
```
{'smoothing_level': 0.11119949831569428,
 'smoothing_trend': 0.049430920023313805,
 'smoothing_seasonal': 0.3620525701498937,
 'damping_trend': nan,
 'initial_level': 2356.5264391986907,
 'initial_trend': -9.443690175376352,
 'initial_seasons': array([0.71325627, 0.68332509, 0.90537798, 0.80561841, 0.65639659,
       0.65451508, 0.88690241, 1.13423953, 0.91927727, 1.21396745,
       1.86941738, 2.3734461 ]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

And then using these parameters we predict the values of test dataset

Table 21: Predicted test values using best parameters in TES

Time_Stamp	Sparkling	auto_predict
1991-01-31	1902	1587.685845
1991-02-28	2049	1356.590237
1991-03-31	1874	1763.121866
1991-04-30	1279	1656.379813
1991-05-31	1432	1542.186697

Graph 21: **Triple Exponential Smoothing model on Time Series**



Alpha=0.111, Beta=0.049, Gamma=0.362, Triple Exponential Smoothing Model forecast on the Test Data, RMSE is 403.706

Table 22: RMSE score with Alpha=0.111, Beta=0.049, Gamma=0.362 in TES model

	Test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315
Alpha=0.039,SimpleExponentialSmoothing	1304.927405
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	1778.564670
Alpha=0.111,Beta=0.049,Gamma=0.362,TripleExponentialSmoothing	403.706228

The higher the alpha , beta and gamma value more weightage is given to the more recent observation. That means, what happened recently will happen again.

Now we try using different Alpha, Beta and Gamma values and try fit the model and find the best Lowest RMSE value

Table 23: **Series of different Alpha, Beta, and Gamma value in SES model with Train and Test RMSE score**

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE
301	0.4	0.1	0.2	384.467709	317.434302
211	0.3	0.2	0.2	388.544148	329.037543
200	0.3	0.1	0.1	388.220071	337.080969
110	0.2	0.2	0.1	398.482510	340.186457
402	0.5	0.1	0.3	396.598057	345.913415

And we got Alpha =0.4, Beta=0.1 and Gamma =0.2 has lesser RMSE score than others

Graph 22: **Triple Exponential Smoothing model with Alpha =0.4, Beta=0.1, Gamma=0.2 on= Time Series**

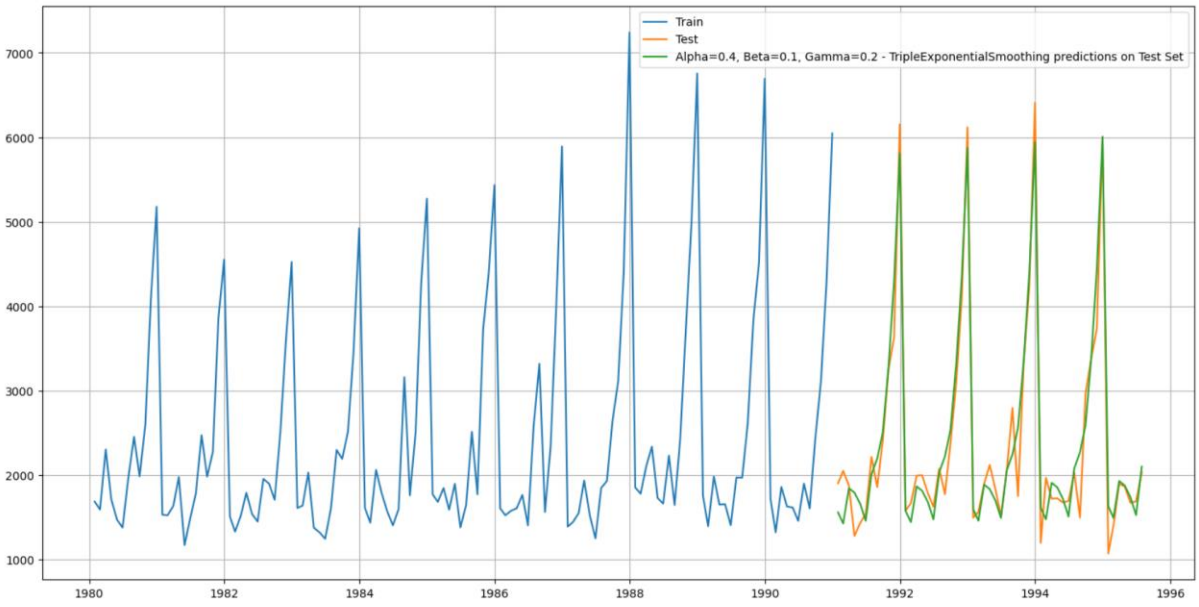


Table 24: RMSE score with Alpha =0.4, Beta=0.1, Gamma=0.2 in TES model

	Test RMSE
Alpha=0.4,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing	317.434302
Alpha=0.111,Beta=0.049,Gamma=0.362,TripleExponentialSmoothing	403.706228
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
SimpleAverageModel	1275.081804
6pointTrailingMovingAverage	1283.927428
Alpha=0.039,SimpleExponentialSmoothing	1304.927405
9pointTrailingMovingAverage	1346.278315
RegressionOnTime	1389.135175
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	1778.564670
NaiveModel	3864.279352

1.5 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:

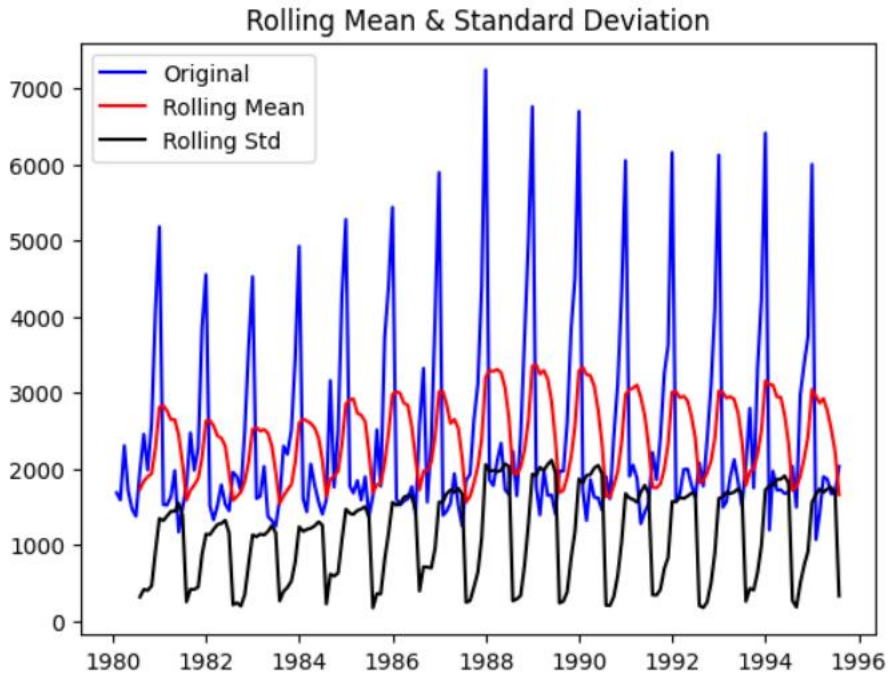
The Augmented Dickey-Fuller test is an unit root test which determines whether there is a unit root and subsequently whether the series is non-stationary.

The hypothesis in a simple form for the ADF test is:

- H0 : The Time Series has a unit root and is thus non-stationary.
- H1 : The Time Series does not have a unit root and is thus stationary.

We would want the series to be stationary for building ARIMA models and thus we would want the p-value of this test to be less than the α value.

Graph 23: **Graphically representation of whether the Time series is stationary or not along with Augmented Dickey-Fuller test results**




```
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
```

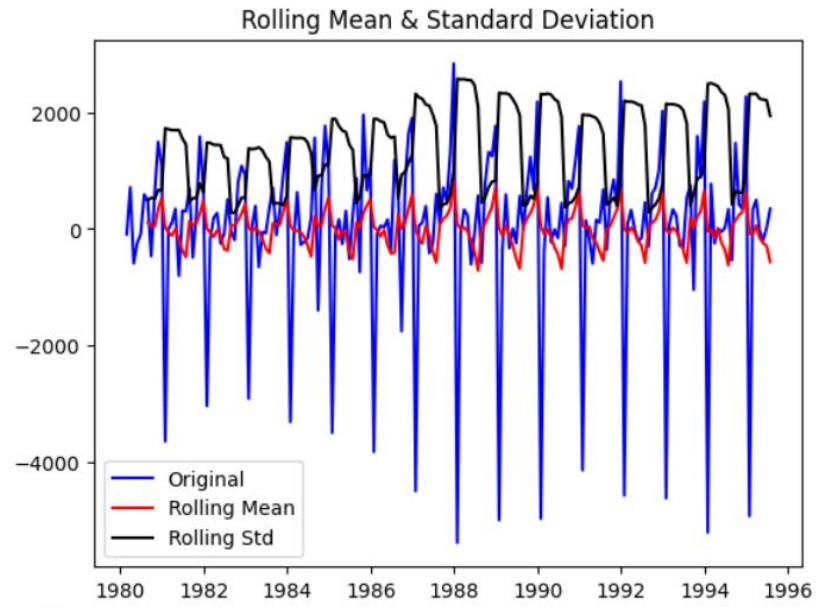
Insights:

Here the p value greater than alpha value 0.05, so we fail to reject null hypothesis.

We see that at 5% significant level the Time Series is non-stationary.

So to make time series stationary we differentiation on the Sparkling variable value once

Graph 24: Time series Differentiation of Stationary




```
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
```

Insights:

Here the p value less than alpha value 0.05, so we reject null hypothesis.

We see that at 5% significant level the Time Series is stationary.

Hence the time series is now stationary after taking 1 order derivative.

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

Solution:

ARIMA Model

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

Table 25: **Parameter combinations for ARIMA Model**

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

We create an empty data frame with parameters and AIC [Akaike Information Criteria] and then fit the ARIMA Model and sort it lowest possible AIC values

Table 26: **Lowest AIC value with parameters**

	param	AIC
7	(2, 0, 1)	2236.590860
6	(2, 0, 0)	2244.811782
1	(0, 0, 1)	2245.312136
2	(0, 0, 2)	2245.347184
4	(1, 0, 1)	2246.005400
5	(1, 0, 2)	2246.935700
3	(1, 0, 0)	2247.358829
8	(2, 0, 2)	2248.277281
0	(0, 0, 0)	2271.205819

Now we fit this parameter [p=2, d=0, q=1] in train dataset and later find prediction on test data

Table 27: Summary Report on ARIMA model

```
                               SARIMAX Results
==========================================================================================
Dep. Variable:                  Sparkling   No. Observations:                  132
Model:                 ARIMA(2, 0, 1)   Log Likelihood               -1113.295
Date:                Sun, 10 Sep 2023   AIC                           2236.591
Time:                        14:00:19   BIC                           2251.005
Sample:                    01-31-1980   HQIC                          2242.448
                         - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const       2399.4586    118.215     20.297      0.000    2167.762    2631.155
ar.L1          1.2375      0.138      8.938      0.000       0.966       1.509
ar.L2         -0.5293      0.124     -4.266      0.000      -0.772      -0.286
ma.L1         -0.8080      0.156     -5.174      0.000      -1.114      -0.502
sigma2      1.233e+06   1.37e+05      9.016      0.000    9.65e+05     1.5e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.03   Jarque-Bera (JB):                26.42
Prob(Q):                              0.86   Prob(JB):                         0.00
Heteroskedasticity (H):               2.40   Skew:                             0.80
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.49
===================================================================================

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

Here it is seen p value is less than 0.05. So it parameter can be used for further predictions

The RMSE value on test data using parameters p=2, d=0, q=1 is 1269.345

SARIMA Model

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

Table 28: Parameter combinations for SARIMA Model

```
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)
```

We create an empty data frame with parameters and AIC [Akaike Information Criteria] and then fit the SARIMA Model and sort it lowest possible AIC values

Table 29: Lowest AIC value with parameters

	param	seasonal	AIC
53	(1, 1, 2)	(2, 0, 2, 6)	1727.678698
26	(0, 1, 2)	(2, 0, 2, 6)	1727.888809
17	(0, 1, 1)	(2, 0, 2, 6)	1741.696452
44	(1, 1, 1)	(2, 0, 2, 6)	1743.374728
71	(2, 1, 1)	(2, 0, 2, 6)	1744.040769

Now we fit this parameter [p=1, d=1, q=2] and seasonal [P=2, D=0, Q=2] with 6 cycle in train dataset and later find prediction on test data

Table 30: **Summary Report on SARIMA model**

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1, 1, 2)x(2, 0, 2, 6)	Log Likelihood	-855.839
Date:	Sun, 10 Sep 2023	AIC	1727.679
Time:	14:01:08	BIC	1749.707
Sample:	0	HQIC	1736.621
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.6449	0.286	-2.256	0.024	-1.205	-0.085
ma.L1	-0.1069	0.250	-0.428	0.669	-0.596	0.383
ma.L2	-0.7005	0.202	-3.470	0.001	-1.096	-0.305
ar.S.L6	-0.0045	0.027	-0.165	0.869	-0.057	0.049
ar.S.L12	1.0361	0.018	56.076	0.000	1.000	1.072
ma.S.L6	0.0676	0.152	0.444	0.657	-0.231	0.366
ma.S.L12	-0.6122	0.093	-6.589	0.000	-0.794	-0.430
sigma2	1.448e+05	1.71e+04	8.465	0.000	1.11e+05	1.78e+05

Ljung-Box (L1) (Q):	0.09	Jarque-Bera (JB):	25.24
Prob(Q):	0.77	Prob(JB):	0.00
Heteroskedasticity (H):	2.63	Skew:	0.47
Prob(H) (two-sided):	0.00	Kurtosis:	5.09

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

The RMSE value on test data using parameters [p=1, d=1, q=2] and seasonal [P=2, D=0, Q=2] with 6 cycle is 626.9452

We predict using these parameters and Seasonal for test data

Table 31: Automatic Predicted SARIMA model with its limits

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1330.393189	380.564773	584.499940	2076.286437
1	1177.260495	392.117175	408.724954	1945.796037
2	1625.929590	392.311510	857.013160	2394.846019
3	1546.281100	397.711790	766.780315	2325.781885
4	1308.737968	398.931580	526.846440	2090.629497

Insights:

From automatic ARIMA and SARIMA models, SARIMA model is best and from that where the p=1, d=1, q=2 while for seasonality P=2, D=0, Q=2 with 6 seasonality cycle is best with low RMSE value.

Table 32: RMSE Score for all the models (sorted)

	Test RMSE
Alpha=0.4,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing	317.434302
Alpha=0.111,Beta=0.049,Gamma=0.362,TripleExponentialSmoothing	403.706228
SARIMA(1, 1, 2)(2, 0, 2, 6)	626.945220
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
ARIMA(2,0,1)	1269.345658
SimpleAverageModel	1275.081804
6pointTrailingMovingAverage	1283.927428
ARIMA(2,1,2)	1299.979749
Alpha=0.039,SimpleExponentialSmoothing	1304.927405
9pointTrailingMovingAverage	1346.278315
RegressionOnTime	1389.135175
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	1778.564670
NaiveModel	3864.279352

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

Table 33: **All the models with test RMSE**

	Test RMSE
Alpha=0.4,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing	317.434302
Alpha=0.111,Beta=0.049,Gamma=0.362,TripleExponentialSmoothing	403.706228
SARIMA(1, 1, 2)(2, 0, 2, 6)	626.945220
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
ARIMA(2,0,1)	1269.345658
SimpleAverageModel	1275.081804
6pointTrailingMovingAverage	1283.927428
ARIMA(2,1,2)	1299.979749
Alpha=0.039,SimpleExponentialSmoothing	1304.927405
9pointTrailingMovingAverage	1346.278315
RegressionOnTime	1389.135175
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	1778.564670
NaiveModel	3864.279352

Insights:

- Triple Exponential Smoothing with Alpha=0.4, Beta=0.1, Gamma=0.2 and Alpha=0.111, Beta=0.049, Gamma=0.362 have lowest RMSE score on test data
- Model SARIMA (1,1,2)(2,0,2,6) has third lowest test RMSE score among all the models

1.8 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:

The optimum model is Triple Exponential Smoothing with Alpha=0.4, Beta=0.1, Gamma=0.2. So we use these parameters and see it works on whole data

Model Evaluation and Predictions

The RMSE score for whole data is 317.434 and it predict for 12 months

Table 34: Exponential Smoothing Model Summary

ExponentialSmoothing Model Results

Dep. Variable:	Sparkling	No. Observations:	187
Model:	ExponentialSmoothing	SSE	32101160.793
Optimized:	True	AIC	2285.966
Trend:	Additive	BIC	2337.664
Seasonal:	Additive	AICC	2290.037
Seasonal Periods:	12	Date:	Sun, 10 Sep 2023
Box-Cox:	False	Time:	14:01:09
Box-Cox Coeff.:	None		

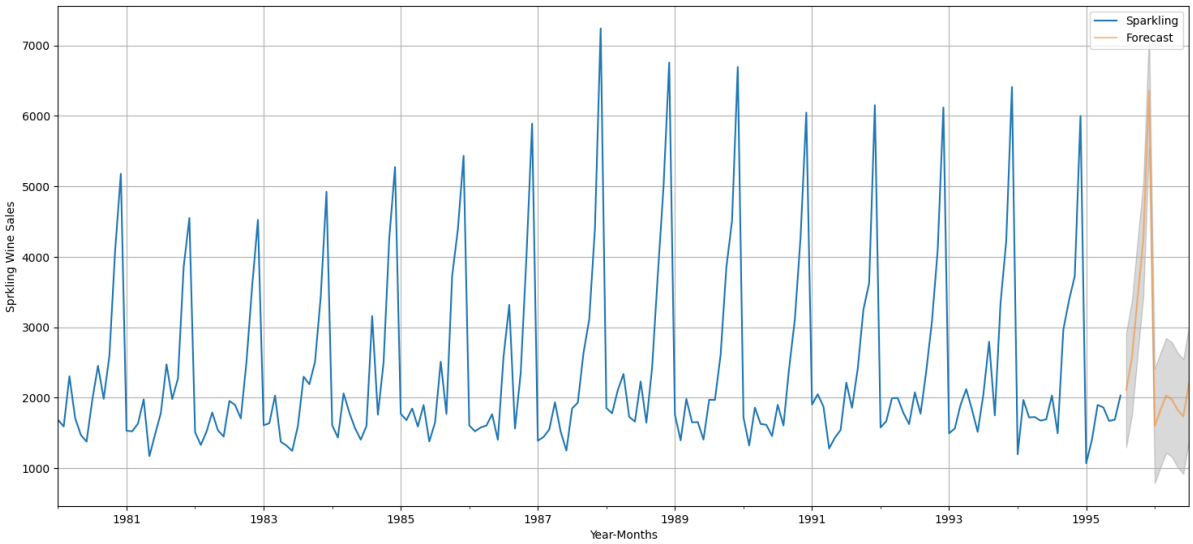
	coeff	code	optimized
smoothing_level	0.4000000	alpha	False
smoothing_trend	0.1000000	beta	False
smoothing_seasonal	0.2000000	gamma	False
initial_level	2581.5049	l.0	True
initial_trend	-24.748267	b.0	True
initial_seasons.0	-847.59268	s.0	True
initial_seasons.1	-912.02233	s.1	True
initial_seasons.2	-513.84796	s.2	True
initial_seasons.3	-604.20054	s.3	True
initial_seasons.4	-862.41028	s.4	True
initial_seasons.5	-905.96348	s.5	True
initial_seasons.6	-386.06083	s.6	True
initial_seasons.7	210.67460	s.7	True
initial_seasons.8	-122.09207	s.8	True
initial_seasons.9	520.76398	s.9	True
initial_seasons.10	1779.9805	s.10	True
initial_seasons.11	2868.3402	s.11	True

RSME for full model is: 414.32349855829955

Table 35: Forecasted values for the next 12 months with confidence interval

	lower_CI	prediction	upper_ci
1995-08-31	1294.848664	2109.045832	2923.243000
1995-09-30	1745.600182	2559.797350	3373.994518
1995-10-31	2608.090578	3422.287746	4236.484914
1995-11-30	3436.709104	4250.906272	5065.103440
1995-12-31	5542.431556	6356.628723	7170.825891
1996-01-31	785.673491	1599.870659	2414.067827
1996-02-29	1009.662511	1823.859678	2638.056846
1996-03-31	1217.325295	2031.522463	2845.719630
1996-04-30	1160.571989	1974.769157	2788.966325
1996-05-31	1012.987405	1827.184573	2641.381741
1996-06-30	919.808859	1734.006027	2548.203195
1996-07-31	1392.557109	2206.754277	3020.951445

Graph 25: Predictions for 12 months in Future using Triple Exponential Smoothing (with confidence interval, alph=0.05)



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

Recommendations:

- Promote summer and other seasonal sales by offering enticing discounts and incentives to attract more customers, especially during the first half of the year, from May to August
- Recognize the seasonality of sparkling wine sales and leverage this by providing special offers during the months of May, June, and July. These offers could include discounts, freebies, or promotions like free movie tickets or club memberships with the purchase of specific wine quantities to stimulate sales
- Plan inventory management effectively from September to December to meet the expected high demand during this period
- Boost year-round sales by implementing promotional campaigns from January to June, ensuring a consistent flow of customers throughout the year

Problem Statement Part 2 – Rose Wine:

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.

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

Solution:

Table 36: **Data top 5 records**

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

Table 37: **Data last 5 records**

	YearMonth	Rose
182	1995-03	45.0
183	1995-04	52.0
184	1995-05	28.0
185	1995-06	40.0
186	1995-07	62.0

Table 38: **Data info**

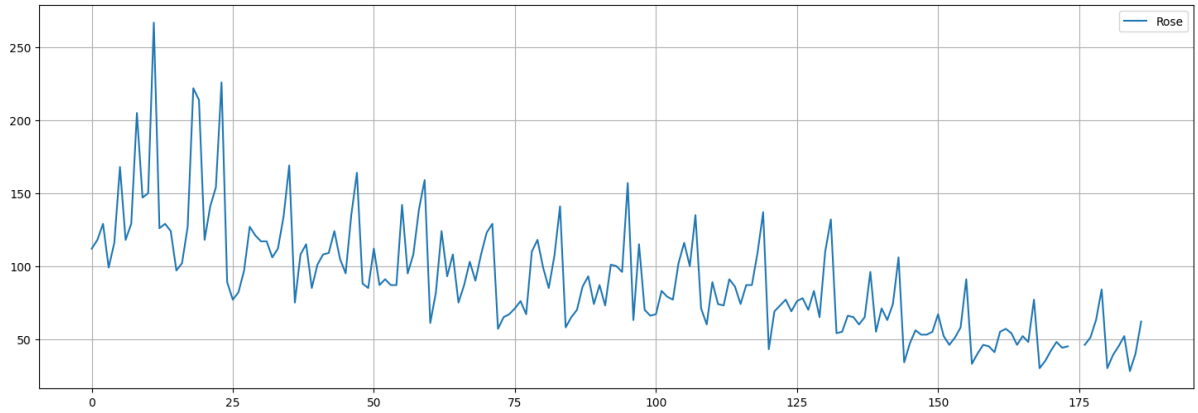
Data Info

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

Table 39: **Data description**

	Rose
count	185.000000
mean	90.394595
std	39.175344
min	28.000000
25%	63.000000
50%	86.000000
75%	112.000000
max	267.000000

Graph 26: Rose wine dataset



Insights:

- The dataset covers a period from 1980 to 1995, comprising a total of 187 entries. However, there are two missing values in the 'Rose' column
- There is significant variability in monthly rose wine sales, with the mean sales volume around 90 units. Sales range from a minimum of 28 units to a maximum of 267 units, indicating seasonal fluctuations
- The median (50th percentile) sales volume is 86 units, which suggests that sales tend to be around this level on an average month
- The data highlights that there are periods of peak sales, as indicated by the 75th percentile value of 112 units. These peaks could be related to specific seasons or events that drive higher demand for rose wine

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

Solution:

Looking for Missing values

```
Missing Values in the Dataset
_____________________________
Rose has 2 missing values, which is 1.07% of total data
```

Table 40: **Index of missing values**

	Rose
Time_Stamp	
1994-07-31	NaN
1994-08-31	NaN

Graph 27: Interpolating the missing value using linear method

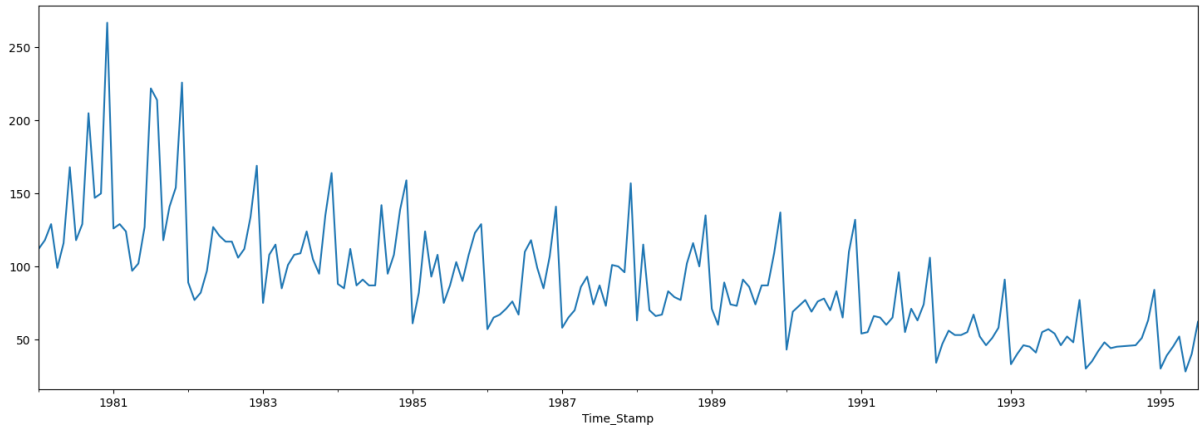
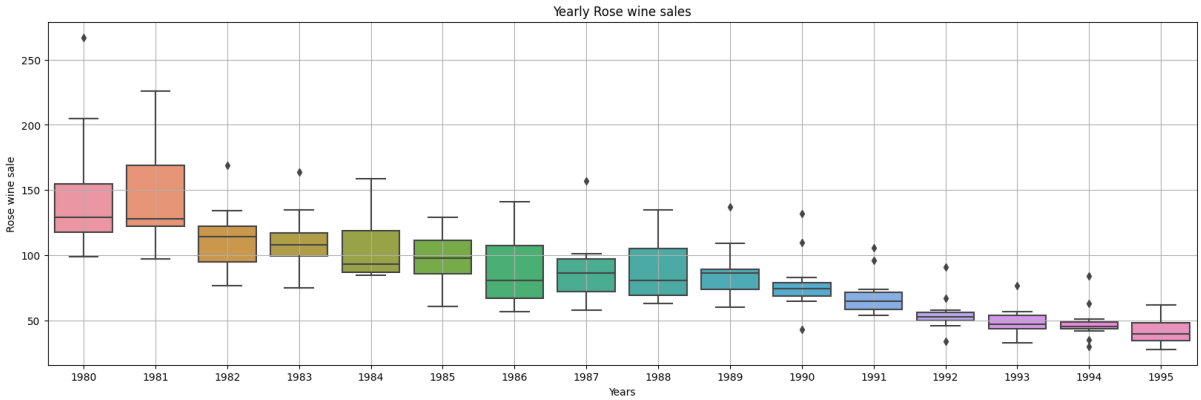


Table 41: **Interpolated missing values**

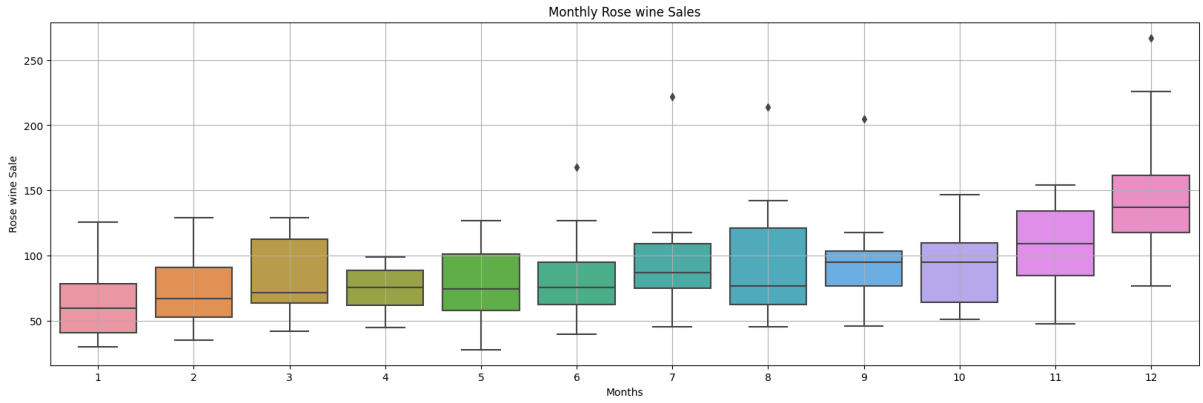
Time_Stamp	Rose
1994-07-31	45.333333
1994-08-31	45.666667

Graph 28: **Boxplot of Sales across Years**



In the provided data, it's evident that sales initially exhibited a strong performance, followed by a gradual decline over time

Graph 29: **Boxplot of Sales across months**



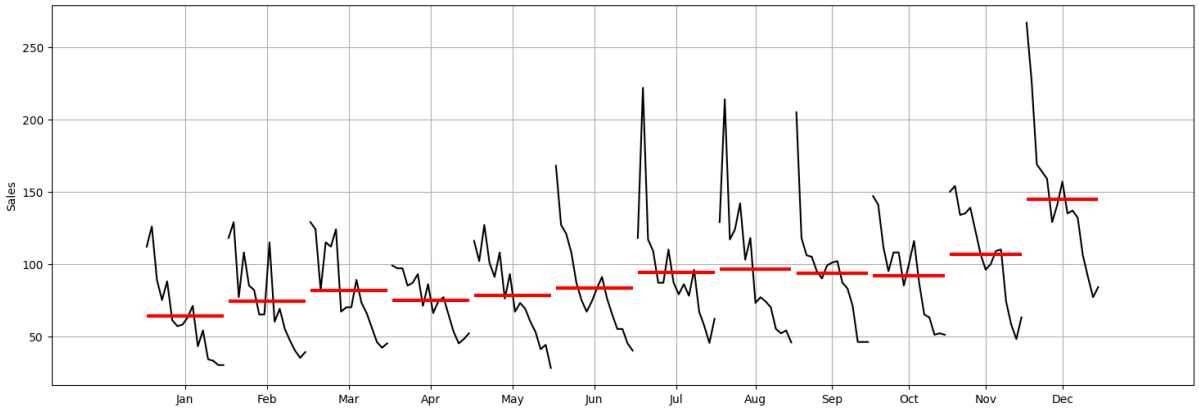
In this data, it's evident that sales tend to surge as the year comes to a close

Table 42: **Pivot Table Months vs Year**

Time_Stamp	1	2	3	4	5	6	7	8	9	10	11	12
Time_Stamp												
1980	112.0	118.0	129.0	99.0	116.0	168.0	118.000000	129.000000	205.0	147.0	150.0	267.0
1981	126.0	129.0	124.0	97.0	102.0	127.0	222.000000	214.000000	118.0	141.0	154.0	226.0
1982	89.0	77.0	82.0	97.0	127.0	121.0	117.000000	117.000000	106.0	112.0	134.0	169.0
1983	75.0	108.0	115.0	85.0	101.0	108.0	109.000000	124.000000	105.0	95.0	135.0	164.0
1984	88.0	85.0	112.0	87.0	91.0	87.0	87.000000	142.000000	95.0	108.0	139.0	159.0
1985	61.0	82.0	124.0	93.0	108.0	75.0	87.000000	103.000000	90.0	108.0	123.0	129.0
1986	57.0	65.0	67.0	71.0	76.0	67.0	110.000000	118.000000	99.0	85.0	107.0	141.0
1987	58.0	65.0	70.0	86.0	93.0	74.0	87.000000	73.000000	101.0	100.0	96.0	157.0
1988	63.0	115.0	70.0	66.0	67.0	83.0	79.000000	77.000000	102.0	116.0	100.0	135.0
1989	71.0	60.0	89.0	74.0	73.0	91.0	86.000000	74.000000	87.0	87.0	109.0	137.0
1990	43.0	69.0	73.0	77.0	69.0	76.0	78.000000	70.000000	83.0	65.0	110.0	132.0
1991	54.0	55.0	66.0	65.0	60.0	65.0	96.000000	55.000000	71.0	63.0	74.0	106.0
1992	34.0	47.0	56.0	53.0	53.0	55.0	67.000000	52.000000	46.0	51.0	58.0	91.0
1993	33.0	40.0	46.0	45.0	41.0	55.0	57.000000	54.000000	46.0	52.0	48.0	77.0
1994	30.0	35.0	42.0	48.0	44.0	45.0	45.333333	45.666667	46.0	51.0	63.0	84.0
1995	30.0	39.0	45.0	52.0	28.0	40.0	62.000000	NaN	NaN	NaN	NaN	NaN

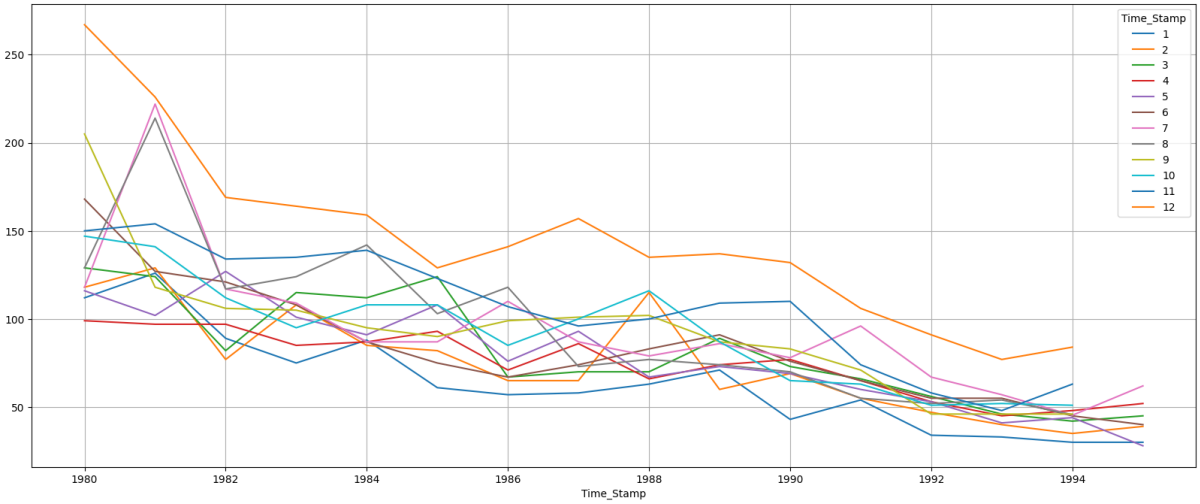
The data exhibits clear seasonal patterns, with higher values observed during specific months. For example, there are consistent peaks in various months across the years, suggesting seasonality in the underlying phenomenon

Graph 30: **Monthly Plot for Time Series**



- The red lines indicate the average sales for the month

Graph 31: **Monthly Sales across Years**



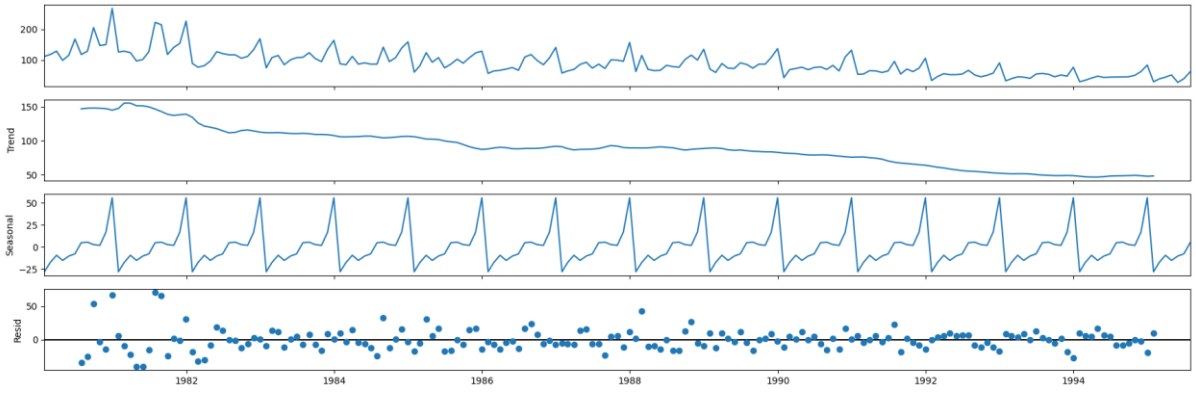
February month has the decent sales among all of the months

Decomposition of Time Series

Here we use additive model and multiplicative model

Additive Model

Graph 32: **Decomposition of Time series using Additive Model**



Insights:

- As per the 'additive' decomposition, we see that there is a decreasing trend in sales of Rose wine starting from 1980 to 1994 sales has fallen only.
- Seasonality is also clearly visible from the seasonal graph where trend lines are forming the peaks with different height every year.
- Residuals seems to be scattered from the 0 level. Indicating that the series is not additive

Table 43: **Trend Aspect of Additive Model**

```
Trend
 Time_Stamp
1980-01-31       NaN
1980-02-29       NaN
1980-03-31       NaN
1980-04-30       NaN
1980-05-31       NaN
1980-06-30       NaN
1980-07-31    147.08
1980-08-31    148.12
1980-09-30    148.37
1980-10-31    148.08
1980-11-30    147.42
1980-12-31    145.12
Freq: M, Name: trend, dtype: float64
```

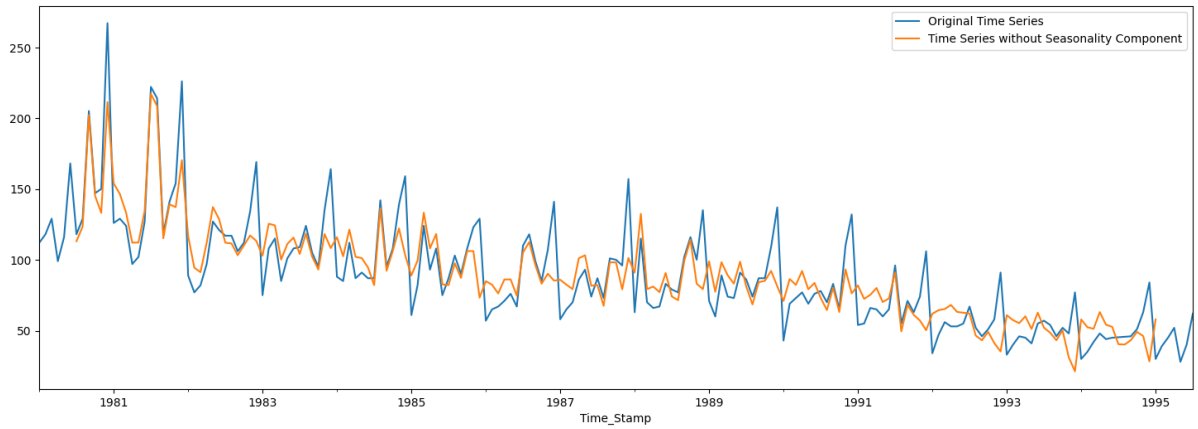
Table 44: **Seasonality Aspect of Additive Model**

```
Seasonality
 Time_Stamp
1980-01-31   -27.91
1980-02-29   -17.44
1980-03-31    -9.29
1980-04-30   -15.10
1980-05-31   -10.20
1980-06-30    -7.68
1980-07-31     4.90
1980-08-31     5.50
1980-09-30     2.77
1980-10-31     1.87
1980-11-30    16.85
1980-12-31    55.71
```

Table 45: **Residual Aspect of Additive Model**

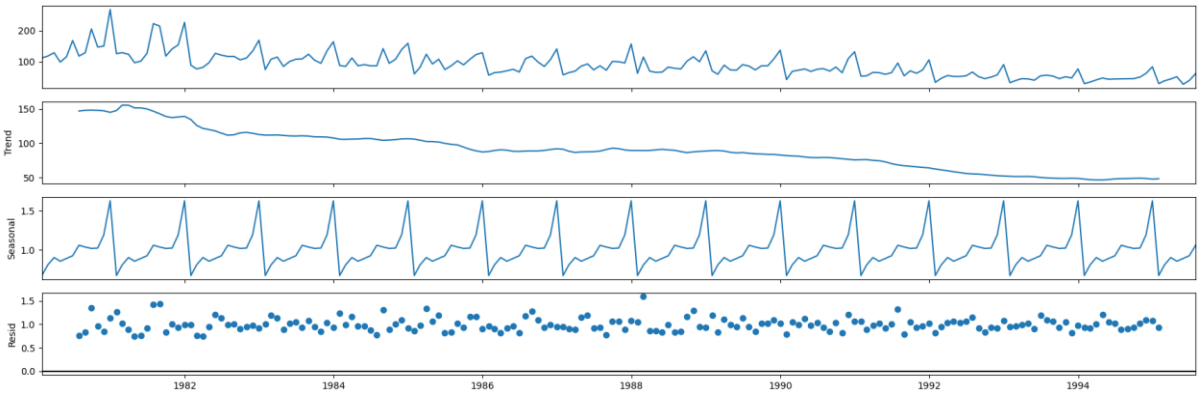
```
Residual
 Time_Stamp
1980-01-31      NaN
1980-02-29      NaN
1980-03-31      NaN
1980-04-30      NaN
1980-05-31      NaN
1980-06-30      NaN
1980-07-31   -33.98
1980-08-31   -24.62
1980-09-30    53.85
1980-10-31    -2.96
1980-11-30   -14.26
1980-12-31    66.16
```

Graph 33: **Time series without the seasonality component**



Multiplicative model

Graph 34: **Decomposition of Time series using Multiplicative Model**



The trend and seasonality are present same as in case of additive model. But residuals plot is clearly showing the concentration of data towards 1 point. Hence it can be concluded that series is multiplicative

Table 46: **Trend Aspect of Multiplicative Model**

Trend

Time_Stamp	
1980-01-31	NaN
1980-02-29	NaN
1980-03-31	NaN
1980-04-30	NaN
1980-05-31	NaN
1980-06-30	NaN
1980-07-31	147.08
1980-08-31	148.12
1980-09-30	148.37
1980-10-31	148.08
1980-11-30	147.42
1980-12-31	145.12

Table 47: **Seasonality Aspect of Multiplicative Model**

Seasonality

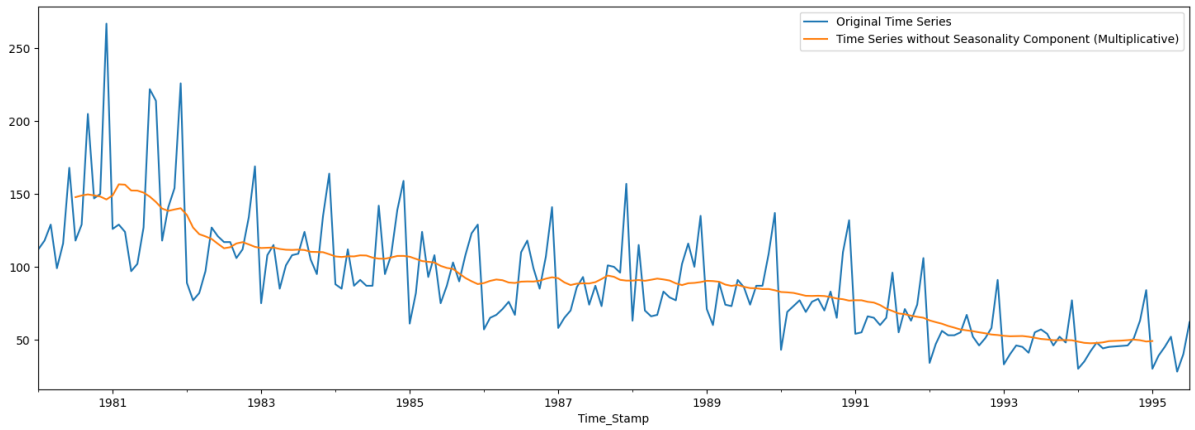
Time_Stamp	
1980-01-31	0.67
1980-02-29	0.81
1980-03-31	0.90
1980-04-30	0.85
1980-05-31	0.89
1980-06-30	0.92
1980-07-31	1.06
1980-08-31	1.04
1980-09-30	1.02
1980-10-31	1.02
1980-11-30	1.19
1980-12-31	1.63

Table 48: **Residual Aspect of Multiplicative Model**

Residual

Time_Stamp	
1980-01-31	NaN
1980-02-29	NaN
1980-03-31	NaN
1980-04-30	NaN
1980-05-31	NaN
1980-06-30	NaN
1980-07-31	0.76
1980-08-31	0.84
1980-09-30	1.36
1980-10-31	0.97
1980-11-30	0.85
1980-12-31	1.13

Graph 35: **Time series without the seasonality component**



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

Solution:

We take the whole data and divided it into train and test. By indexing the year less 1991 for train dataset, while more than or equal to 1991 as test dataset.

There were 187 data points in the whole dataset, while splitting the training dataset has 132 data points and 55 data points as test data set.

Table 49: **First five rows of Training Data**

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

Table 50: **Last five rows of Training Data**

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

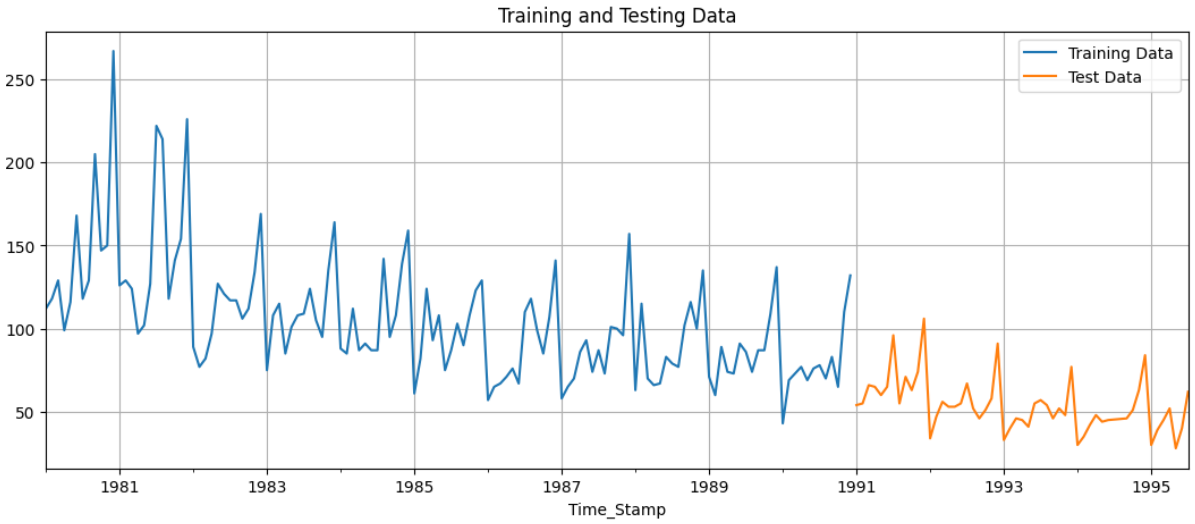
Table 51: **First five rows of Test Data**

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

Table 52: **Last five rows of Test Data**

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

Graph 36: **Graphical representation of the Training and Test data set**



2.4 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

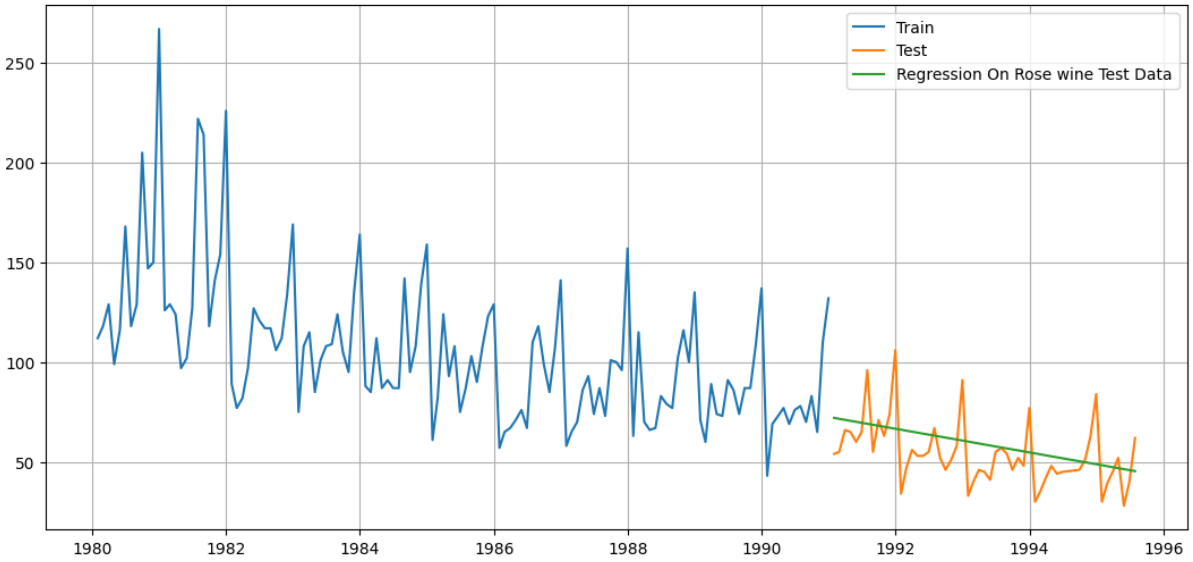
Solution:

Model 1: Linear Regression

We imported Linear Regression from sklearn. This model is based on Linear Regression method to forecast the data

```
Training Time instance
 [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, ]  .....  [128, 129, 130, 131, 132]
Test Time instance
 [133, 134, 135, 136, 137, 138, 139, 140,  .....  [183, 184, 185, 186, 187]
```

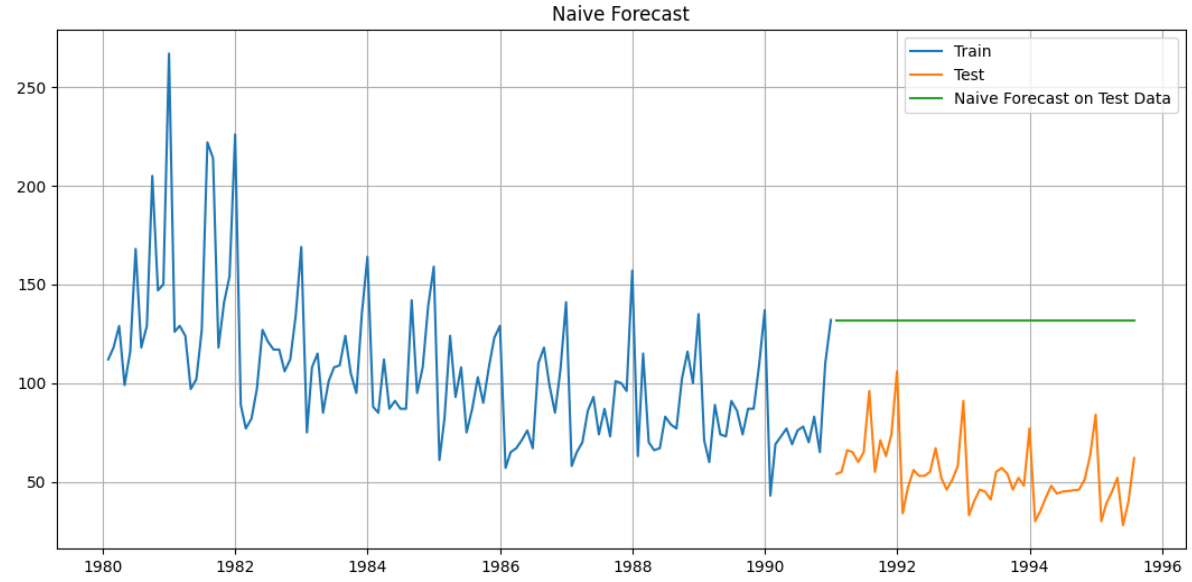
Graph 37: **Linear Regression model on Time Series**



RSME for lr_model is: 15.268

Model 2: Naive Approach

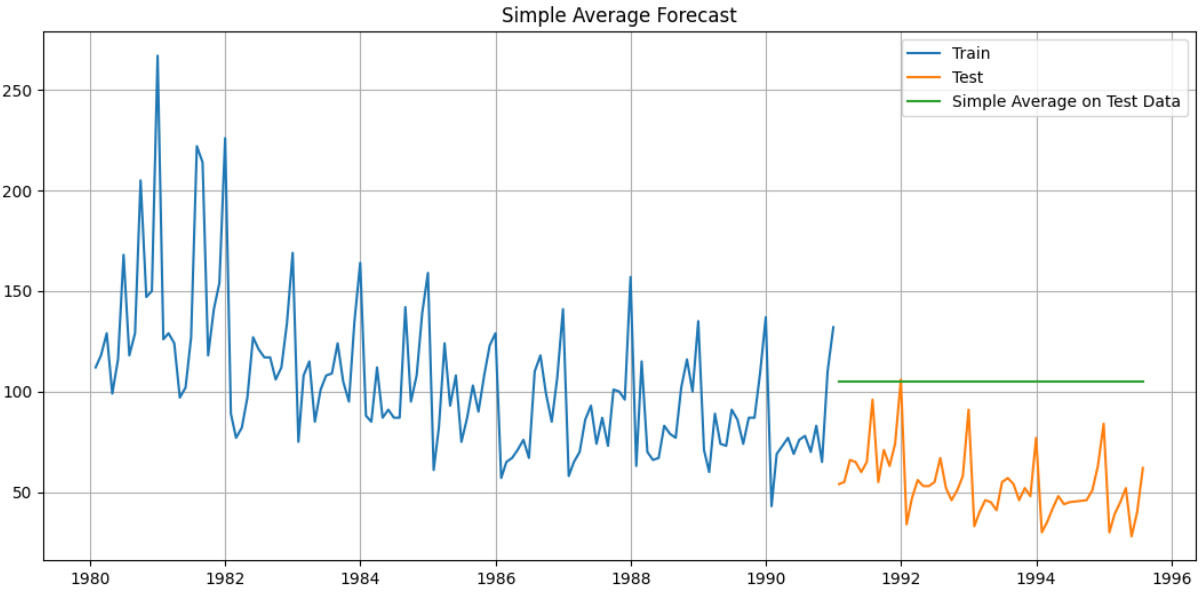
Graph 38: **Naïve Forecast model on Time Series**



RMSE for Naïve model is: 79.718

Model 3: Simple Average

Graph 39: **Simple Average model on Time Series**



RMSE for Simple Average model is: 53.460

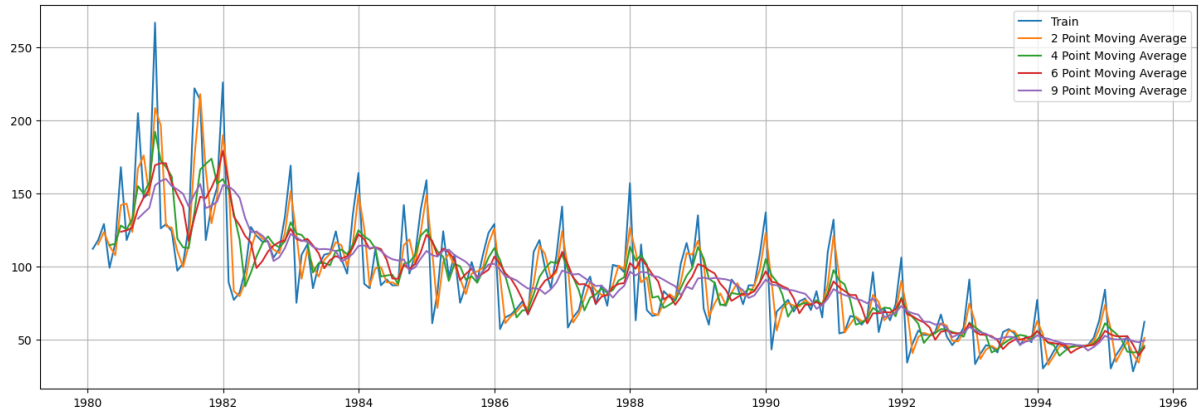
Model 4: Moving Average

This method uses averaging to forecast the values based on window sizes. The window keeps on moving with the size constant for newer points to be forecasted

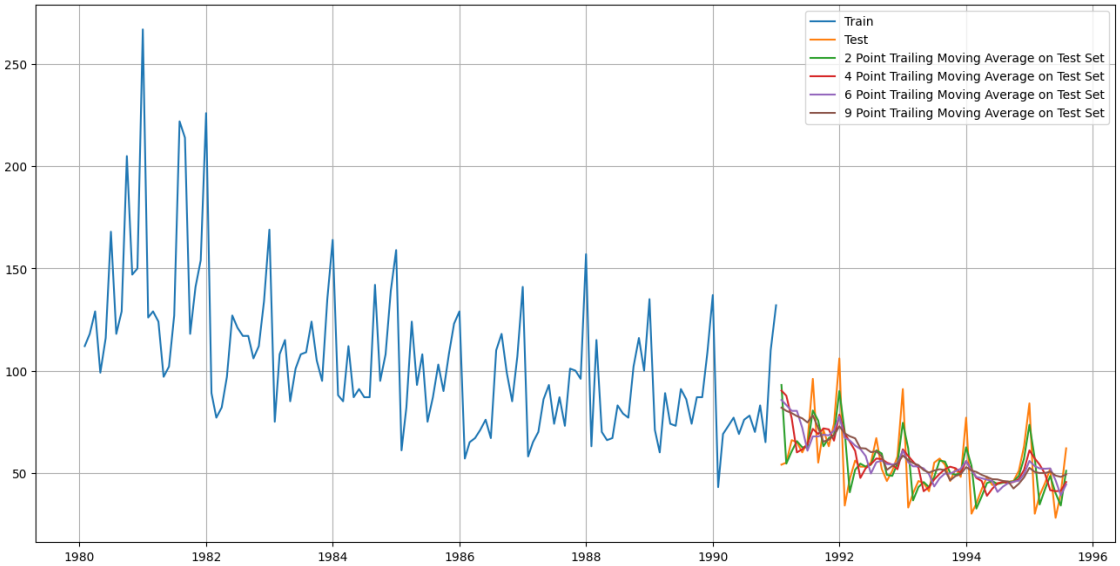
Table 53: **Considering window size as 2, 4, 6, and 9**

Time_Stamp	Rose	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1980-01-31	112.0	NaN	NaN	NaN	NaN
1980-02-29	118.0	115.0	NaN	NaN	NaN
1980-03-31	129.0	123.5	NaN	NaN	NaN
1980-04-30	99.0	114.0	114.5	NaN	NaN
1980-05-31	116.0	107.5	115.5	NaN	NaN

Graph 40: **Moving Average model on Time Series**



Graph 41: **Moving Average model on Time Series (Test Data)**



For 2 point Moving Average Model forecast on the Training Data, RMSE is **11.529**

For 4 point Moving Average Model forecast on the Training Data, RMSE is **14.451**

For 6 point Moving Average Model forecast on the Training Data, RMSE is **14.566**

For 9 point Moving Average Model forecast on the Training Data, RMSE is **14.728**

Method 5: Simple Exponential Smoothing [SES]

Exponential smoothing averages or exponentially weighted moving averages consist of forecast based on previous period's data with exponentially declining influence on the older observations.

Table 54: **Best possible values of parameters in SES**

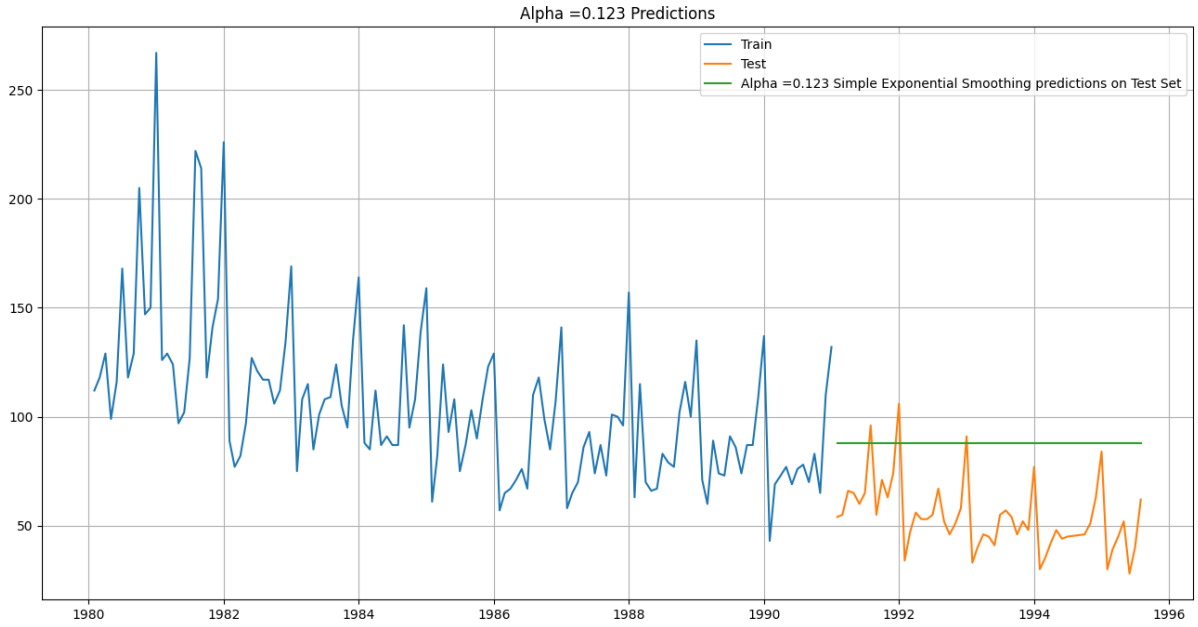
```
{'smoothing_level': 0.12362013466760018,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 112.0,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

And then using these parameters we predict the values of test dataset

Table 55: **Predicted test values using best parameters in SES**

	Rose	predict
Time_Stamp		
1991-01-31	54.0	87.983765
1991-02-28	55.0	87.983765
1991-03-31	66.0	87.983765
1991-04-30	65.0	87.983765
1991-05-31	60.0	87.983765

Graph 42: **Simple Exponential Smoothing model on Time Series**



For Alpha =0.123 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 37.592

Method 6: Double Exponential Smoothing (Holt's Model)

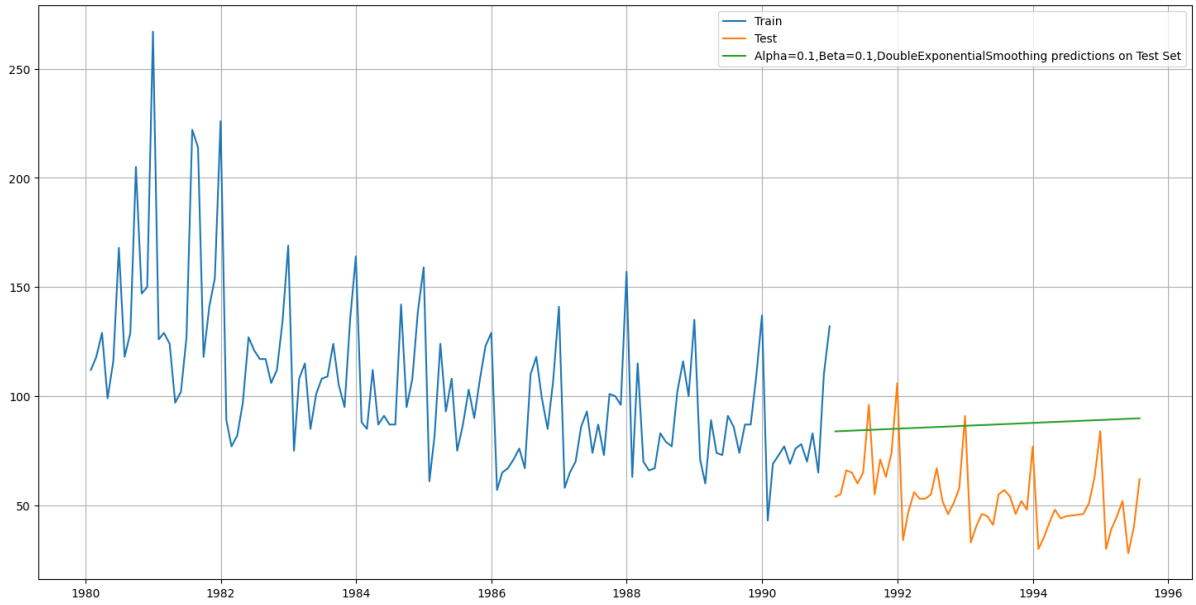
This model considers level and trend while forecasting values

We use a combination of different alpha and beta values and consider the best value and build a model on it

Table 56: **Series of different Alpha and Beta value in DES model with Train and Test RMSE score (sorted)**

	Alpha Values	Beta Values	Train RMSE	Test RMSE
0	0.1	0.1	34.439111	36.923416
1	0.1	0.2	33.450729	48.688648
10	0.2	0.1	33.097427	65.731702
2	0.1	0.3	33.145789	78.156641
20	0.3	0.1	33.611269	98.653317

Graph 43: **Double Exponential Smoothing model with Alpha =0.1, Beta=0.1 on Time Series**



For Alpha =0.1, Beta=0.1, Double Exponential Smoothing Model forecast on the Test Data, **RMSE is 36.923**

Method 7: Triple Exponential Smoothing (Holt - Winter's Model)[TES]

Holt-Winters - ETS(A, A, A) - Holt Winter's linear method with additive errors. Three parameters, Level, Trend and Seasonality are accounted for in this model.

Here it use multiplicative [seasonal], Additive [trend] and residual also.

Here we apply SES on the train and test data. And automatic fit with best possible value of smoothing level (alpha), and no value for trend and seasonality.

Table 57: Best possible values of parameters in TES

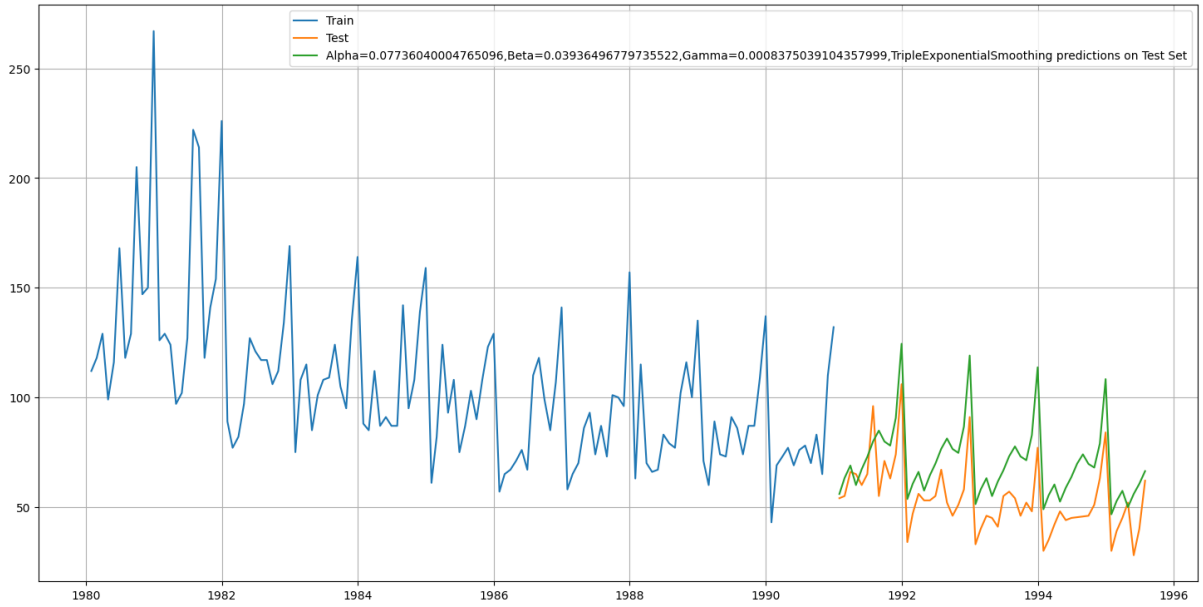
```
{'smoothing_level': 0.07736040004765096,
 'smoothing_trend': 0.03936496779735522,
 'smoothing_seasonal': 0.0008375039104357999,
 'damping_trend': nan,
 'initial_level': 156.90674503596637,
 'initial_trend': -0.9061396720042346,
 'initial_seasons': array([0.7142168 , 0.80982439, 0.88543128, 0.77363782, 0.87046319,
       0.94699283, 1.04196135, 1.11012703, 1.04835489, 1.0276963 ,
       1.19783562, 1.6514144 ]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

And then using these parameters we predict the values of test dataset

Table 58: **Predicted test values using best parameters in TES**

Time_Stamp	Rose	auto_predict
1991-01-31	54.0	55.942246
1991-02-28	55.0	63.240624
1991-03-31	66.0	68.899674
1991-04-30	65.0	60.007486
1991-05-31	60.0	67.257150

Graph 44: **Triple Exponential Smoothing model on Time Series**



Alpha=0.773, Beta=0.039, Gamma=0.000, Triple Exponential Smoothing Model forecast on the Test Data, **RMSE is 19.113**

Table 59: **RMSE score with Alpha=0.0773, Beta=0.0393, Gamma=0.0008 in TES model**

	Test RMSE
RegressionOnTime	15.268955
NaiveModel	79.718773
SimpleAverageModel	53.460570
2pointTrailingMovingAverage	11.529278
4pointTrailingMovingAverage	14.451403
6pointTrailingMovingAverage	14.566327
9pointTrailingMovingAverage	14.727630
Alpha=0.123,SimpleExponentialSmoothing	37.592212
Alpha=0.1,SimpleExponentialSmoothing	36.828033
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	36.923416
Alpha=0.0773,Beta=0.0393,Gamma=0.0008,TripleExponentialSmoothing	19.113110

The higher the alpha , beta and gamma value more weightage is given to the more recent observation. That means, what happened recently will happen again.

Now we try using different Alpha, Beta and Gamma values and try fit the model and find the best Lowest RMSE value

Table 60: **Series of different Alpha, Beta, and Gamma value in SES model with Train and Test RMSE score**

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE
10	0.1	0.2	0.1	19.770392	9.223504
11	0.1	0.2	0.2	20.253487	9.496152
151	0.2	0.6	0.2	23.129850	9.565988
12	0.1	0.2	0.3	20.871304	9.888106
142	0.2	0.5	0.3	23.656276	9.891550

And we got Alpha =0.1, Beta=0.2 and Gamma =0.1 has lesser RMSE score than others

Graph 45: **Triple Exponential Smoothing model with Alpha =0.1, Beta=0.2, Gamma=0.1 on Time Series**

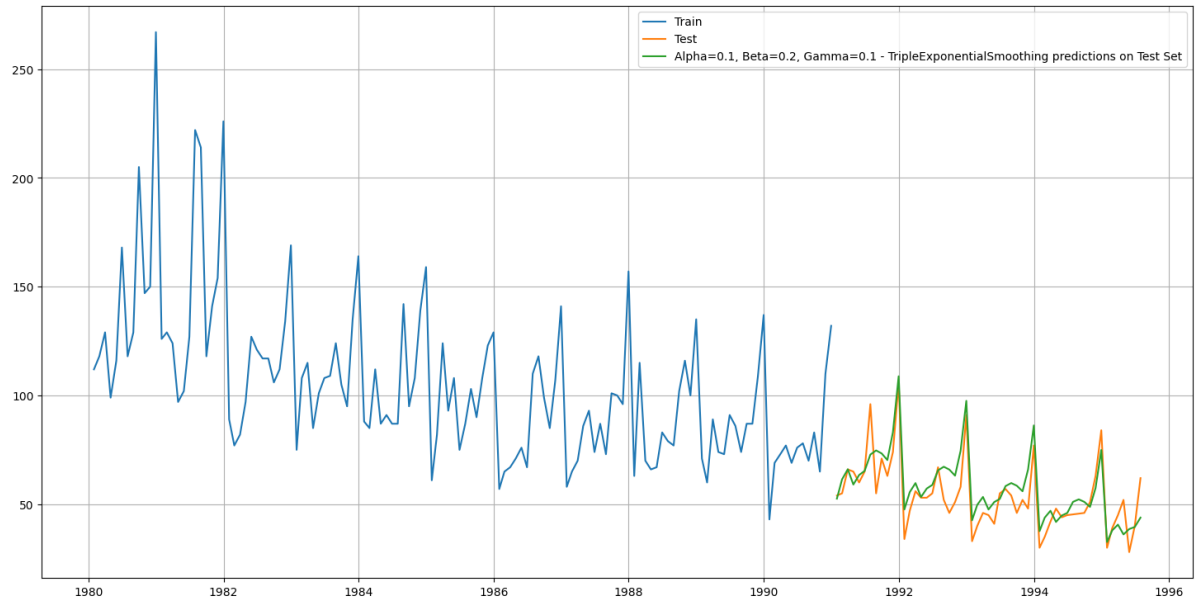


Table 61: **RMSE score with Alpha =0.1, Beta=0.2, Gamma=0.1 in TES model**

	Test RMSE
Alpha=0.1,Beta=0.2,Gamma=0.1,TripleExponentialSmoothing	9.223504
2pointTrailingMovingAverage	11.529278
4pointTrailingMovingAverage	14.451403
6pointTrailingMovingAverage	14.566327
9pointTrailingMovingAverage	14.727630
RegressionOnTime	15.268955
Alpha=0.0773,Beta=0.0393,Gamma=0.0008,TripleExponentialSmoothing	19.113110
Alpha=0.1,SimpleExponentialSmoothing	36.828033
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	36.923416
Alpha=0.123,SimpleExponentialSmoothing	37.592212
SimpleAverageModel	53.460570
NaiveModel	79.718773

2.5 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:

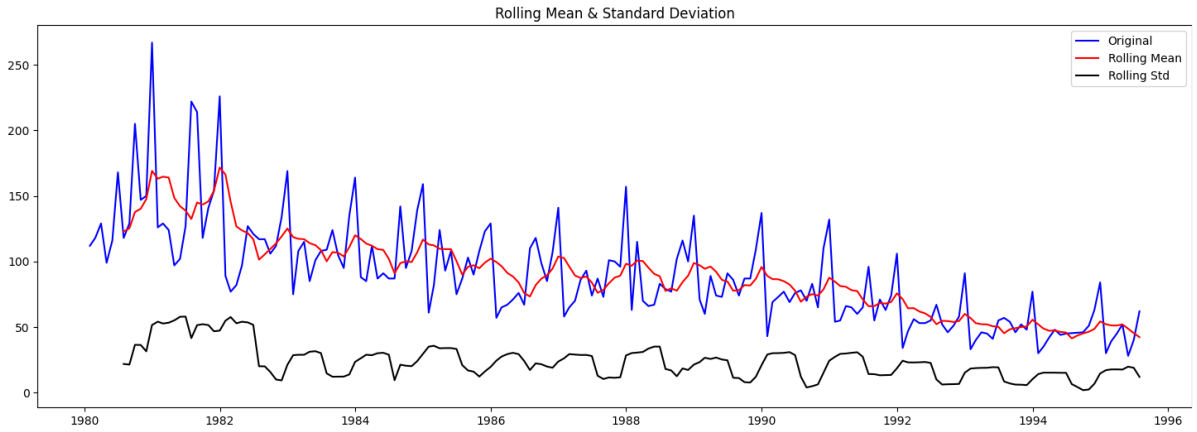
The Augmented Dickey-Fuller test is an unit root test which determines whether there is a unit root and subsequently whether the series is non-stationary.

The hypothesis in a simple form for the ADF test is:

- H0 : The Time Series has a unit root and is thus non-stationary.
- H1 : The Time Series does not have a unit root and is thus stationary.

We would want the series to be stationary for building ARIMA models and thus we would want the p-value of this test to be less than the α value.

Graph 46: **Graphically representation of whether the Time series is stationary or not along with Augmented Dickey-Fuller test results**

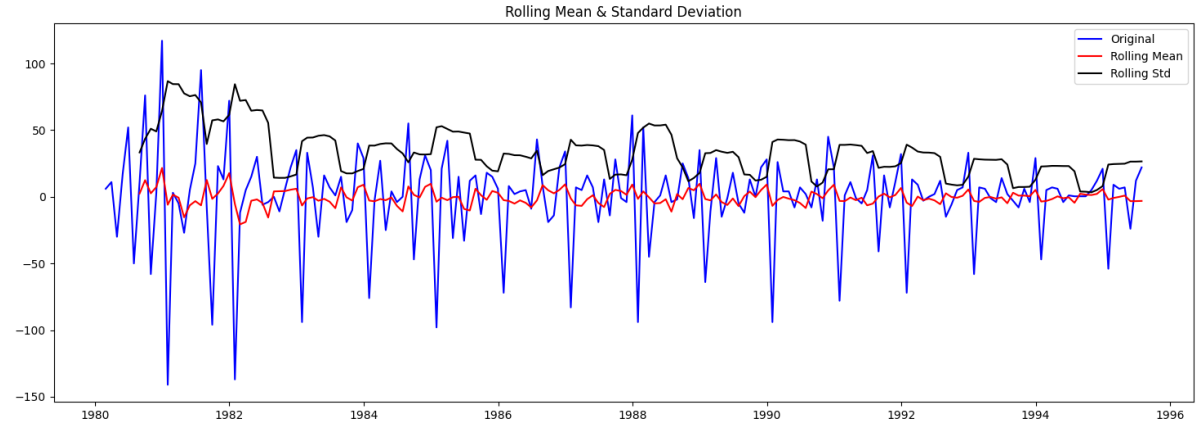



```
Results of Dickey-Fuller Test:
Test Statistic                  -1.876699
p-value                          0.343101
#Lags Used                      13.000000
Number of Observations Used    173.000000
Critical Value (1%)             -3.468726
Critical Value (5%)             -2.878396
Critical Value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

Insights:

The p-value is more than 0.05, hence we cannot reject the null. The data is not stationary, which means have to differentiate the time series

Graph 47: Time series Differentiation of Stationary




```
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
```

Insights:

Here the p value less than alpha value 0.05, so we reject null hypothesis.

We see that at 5% significant level the Time Series is stationary.

Hence the time series is now stationary after taking 1 order derivative.

2.6 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:

ARIMA Model

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

Table 62: **Parameter combinations for ARIMA Model**

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

We create an empty data frame with parameters and AIC [Akaike Information Criteria] and then fit the ARIMA Model and sort it lowest possible AIC values

Table 63: **Lowest AIC value with parameters**

	param	AIC
5	(1, 0, 2)	1292.053213
8	(2, 0, 2)	1292.248056
7	(2, 0, 1)	1292.937195
4	(1, 0, 1)	1294.510585
3	(1, 0, 0)	1301.546304
6	(2, 0, 0)	1302.347685
1	(0, 0, 1)	1305.468406
2	(0, 0, 2)	1306.587015
0	(0, 0, 0)	1324.899703

Now we fit this parameter [p=1, d=0, q=2] in train dataset and later find prediction on test data

Table 64: Summary Report on ARIMA model

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	ARIMA(1, 0, 2)	Log Likelihood	-641.027
Date:	Sun, 10 Sep 2023	AIC	1292.053
Time:	15:45:23	BIC	1306.467
Sample:	01-31-1980	HQIC	1297.910
	- 12-31-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
const	107.8405	24.779	4.352	0.000	59.275	156.406
ar.L1	0.9861	0.027	36.818	0.000	0.934	1.039
ma.L1	-0.6874	0.090	-7.622	0.000	-0.864	-0.511
ma.L2	-0.2007	0.093	-2.148	0.032	-0.384	-0.018
sigma2	960.8593	100.353	9.575	0.000	764.172	1157.547

Ljung-Box (L1) (Q):	0.05	Jarque-Bera (JB):	58.48
Prob(Q):	0.82	Prob(JB):	0.00
Heteroskedasticity (H):	0.36	Skew:	0.98
Prob(H) (two-sided):	0.00	Kurtosis:	5.61

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

Here it is seen p value is less than 0.05. So it parameter can be used for further predictions

The RMSE value on test data using parameters p=1, d=0, q=2 is 45.438

SARIMA Model

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

Table 65: Parameter combinations for SARIMA Model

```
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)
```

We create an empty data frame with parameters and AIC [Akaike Information Criteria] and then fit the SARIMA Model and sort it lowest possible AIC values

Table 66: **Lowest AIC value with parameters**

	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)	1045.230049
71	(2, 1, 1)	(2, 0, 2, 6)	1051.673461
44	(1, 1, 1)	(2, 0, 2, 6)	1052.778470

Now we fit this parameter [p=1, d=1, q=2] and seasonal [P=2, D=0, Q=2] with 6 cycle in train dataset and later find prediction on test data

Table 67: **Summary Report on SARIMA model**

```
                                 SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(1, 1, 2)x(2, 0, 2, 6)   Log Likelihood                -512.828
Date:                           Sun, 10 Sep 2023   AIC                           1041.656
Time:                                   15:45:38   BIC                           1063.685
Sample:                                        0   HQIC                          1050.598
                                             - 132
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5939      0.152     -3.899      0.000      -0.893      -0.295
ma.L1         -0.1954    122.598     -0.002      0.999    -240.482     240.091
ma.L2         -0.8046     98.686     -0.008      0.993    -194.226     192.617
ar.S.L6       -0.0625      0.035     -1.764      0.078      -0.132       0.007
ar.S.L12       0.8451      0.039     21.883      0.000       0.769       0.921
ma.S.L6        0.2227    162.166      0.001      0.999    -317.617     318.062
ma.S.L12      -0.7775    126.029     -0.006      0.995    -247.790     246.235
sigma2       335.1822   6.52e+04      0.005      0.996   -1.27e+05    1.28e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.07   Jarque-Bera (JB):                56.68
Prob(Q):                              0.78   Prob(JB):                         0.00
Heteroskedasticity (H):               0.47   Skew:                             0.52
Prob(H) (two-sided):                  0.02   Kurtosis:                         6.26
===================================================================================

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

The RMSE value on test data using parameters [p=1, d=1, q=2] and seasonal [P=2, D=0, Q=2] with 6 cycle is 26.133

We predict using these parameters and Seasonal for test data

Table 68: Automatic Predicted SARIMA model with its limits

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	62.838344	18.848519	25.895925	99.780763
1	67.628830	19.300294	29.800948	105.456711
2	74.745289	19.412843	36.696815	112.793763
3	71.323939	19.475788	33.152096	109.495782
4	76.016004	19.484068	37.827933	114.204075

Insights:

From automatic ARIMA and SARIMA models, SARIMA model is best and from that where the p=1, d=1, q=2 while for seasonality P=2, D=0, Q=2 with 6 seasonality cycle is best with low RMSE value.

Table 69: RMSE Score for all the models (sorted)

	Test RMSE
Alpha=0.1,Beta=0.2,Gamma=0.1,TripleExponentialSmoothing	9.223504
2pointTrailingMovingAverage	11.529278
4pointTrailingMovingAverage	14.451403
6pointTrailingMovingAverage	14.566327
9pointTrailingMovingAverage	14.727630
RegressionOnTime	15.268955
Alpha=0.0773,Beta=0.0393,Gamma=0.0008,TripleExponentialSmoothing	19.113110
SARIMA(1, 1, 2)(2, 0, 2, 6)	26.133357
Alpha=0.1,SimpleExponentialSmoothing	36.828033
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	36.923416
ARIMA(0,1,2)	37.306480
Alpha=0.123,SimpleExponentialSmoothing	37.592212
ARIMA(1,0,2)	45.438589
SimpleAverageModel	53.460570
NaiveModel	79.718773

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

Table 70: **All the models with test RMSE**

	Test RMSE
Alpha=0.1,Beta=0.2,Gamma=0.1,TripleExponentialSmoothing	9.223504
2pointTrailingMovingAverage	11.529278
4pointTrailingMovingAverage	14.451403
6pointTrailingMovingAverage	14.566327
9pointTrailingMovingAverage	14.727630
RegressionOnTime	15.268955
Alpha=0.0773,Beta=0.0393,Gamma=0.0008,TripleExponentialSmoothing	19.113110
SARIMA(1, 1, 2)(2, 0, 2, 6)	26.133357
Alpha=0.1,SimpleExponentialSmoothing	36.828033
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	36.923416
ARIMA(0,1,2)	37.306480
Alpha=0.123,SimpleExponentialSmoothing	37.592212
ARIMA(1,0,2)	45.438589
SimpleAverageModel	53.460570
NaiveModel	79.718773

Insights:

- Triple Exponential Smoothing with Alpha=0.1, Beta=0.2, Gamma=0.1 and 2 point Trailing Moving Average have lowest RMSE score on test data
- 4 point Trailing Moving Average has third lowest test RMSE score among all the models

2.8 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:

The optimum model is Triple Exponential Smoothing with Alpha=0.1, Beta=0.2, Gamma=0.1. So we use these parameters and see it works on whole data

Model Evaluation and Predictions

The RMSE score for whole data is 17.023 and it predict for 12 months

Table 71: **Exponential Smoothing Model Summary**

ExponentialSmoothing Model Results

Dep. Variable:	Rose	No. Observations:	187
Model:	ExponentialSmoothing	SSE	54193.825
Optimized:	True	AIC	1092.143
Trend:	Additive	BIC	1143.841
Seasonal:	Multiplicative	AICC	1096.214
Seasonal Periods:	12	Date:	Sun, 10 Sep 2023
Box-Cox:	False	Time:	15:45:39
Box-Cox Coeff.:	None		

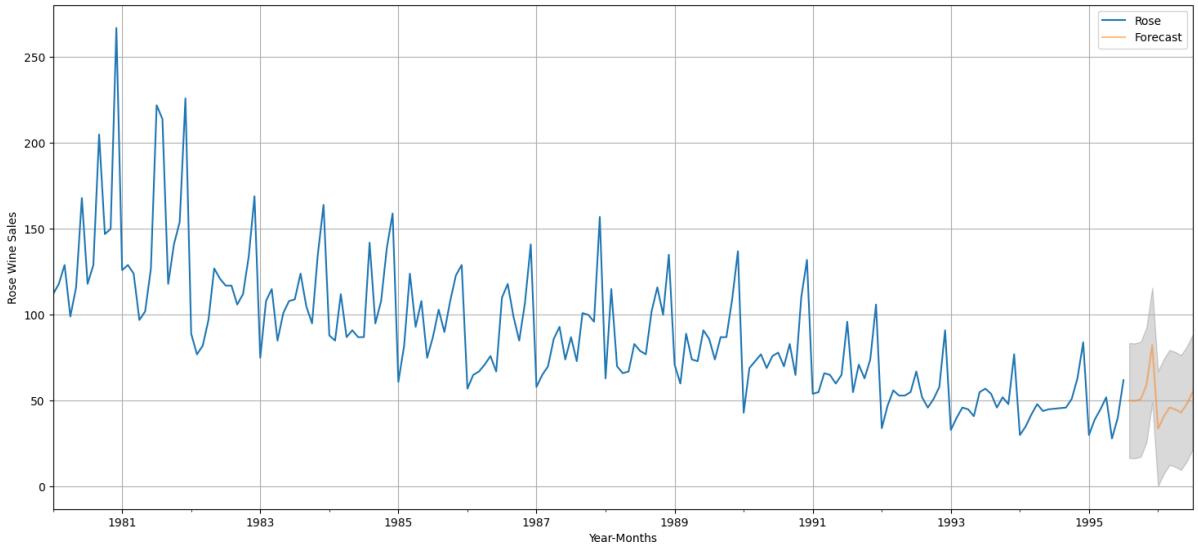
	coeff	code	optimized
smoothing_level	0.1000000	alpha	False
smoothing_trend	0.2000000	beta	False
smoothing_seasonal	0.1000000	gamma	False
initial_level	137.09623	l.0	True
initial_trend	1.1820048	b.0	True
initial_seasons.0	0.8083344	s.0	True
initial_seasons.1	0.8790882	s.1	True
initial_seasons.2	0.9589981	s.2	True
initial_seasons.3	0.8334805	s.3	True
initial_seasons.4	0.9277028	s.4	True
initial_seasons.5	1.0396304	s.5	True
initial_seasons.6	1.1671270	s.6	True
initial_seasons.7	1.2314734	s.7	True
initial_seasons.8	1.1271652	s.8	True
initial_seasons.9	1.1112289	s.9	True
initial_seasons.10	1.2543064	s.10	True
initial_seasons.11	1.7770756	s.11	True

RSME for full model is: **17.023**

Table 72: **Forecasted values for the next 12 months with confidence interval**

	lower_CI	prediction	upper_ci
1995-08-31	16.636544	50.084259	83.531973
1995-09-30	16.427462	49.875177	83.322892
1995-10-31	17.384897	50.832612	84.280326
1995-11-30	25.743386	59.191101	92.638816
1995-12-31	48.902700	82.350415	115.798129
1996-01-31	0.266862	33.714576	67.162291
1996-02-29	7.341521	40.789236	74.236951
1996-03-31	12.639114	46.086829	79.534543
1996-04-30	11.477366	44.925081	78.372795
1996-05-31	9.635348	43.083063	76.530778
1996-06-30	14.550330	47.998045	81.445760
1996-07-31	21.447403	54.895117	88.342832

Graph 48: **Predictions for 12 months in Future using Triple Exponential Smoothing (with confidence interval, alpha=0.05)**



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

Recommendations:

- Rose wine sales shown a decrease in trend on year-on-year basis
- December month has the highest sales in a year for Rose wine as well
- The company should stimulate sales during the summer and other seasons by introducing discounts and enticing offers, with a focus on the first three months and July, August, and September
- Sales of rose wine exhibit distinct seasonality
- Notably, high sales for rose wine are observed during the months of March, August, October, November, and December
- To meet increased demand during these peak months, the company should proactively plan and maintain sufficient stock levels
- To boost sales during periods of lower demand, the company should strategize and implement promotional campaigns

Both Wine Comparison

- The sales figures for Rose wine are notably lower when compared to Sparkling wine.
- Considering the recent trends, the company may find it beneficial to prioritize and invest more in its Sparkling wine business, as it appears to have a stronger market presence compared to Rose wine

Thank you