TIME SERIES FORECASTING

SOUNDARYA KSC

22-05-2022 PGP-DSBA

TABLE OF CONTENTS

LIST OF TABLES

Problem:

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

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

The data consist of different types of wine sales for ABC Estate Wines from the year 1980 to 1995. Here we have two datasets namely sparkle and rose. Reading the CSV file and checking the head() of the dataset.

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

Figure 1: Sparkle

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

Figure 2: Rose

Plotting the time series to understand the behaviour of the sparkling data

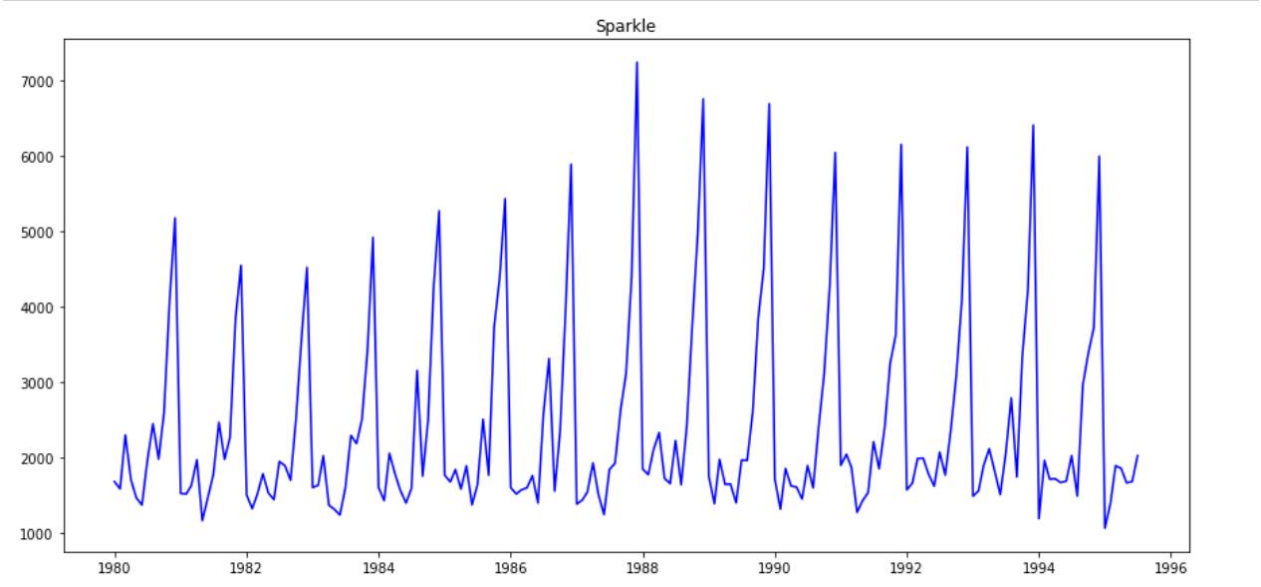


Figure 3: Sparkle time slot graph

Observation:

1. We notice that there is no much trend in the plot.

2. The seasonality seems to have a pattern on the yearly basis.
3. There are no missing values in the dataset.

Plotting the time series to understand the behaviour of the Rose data:

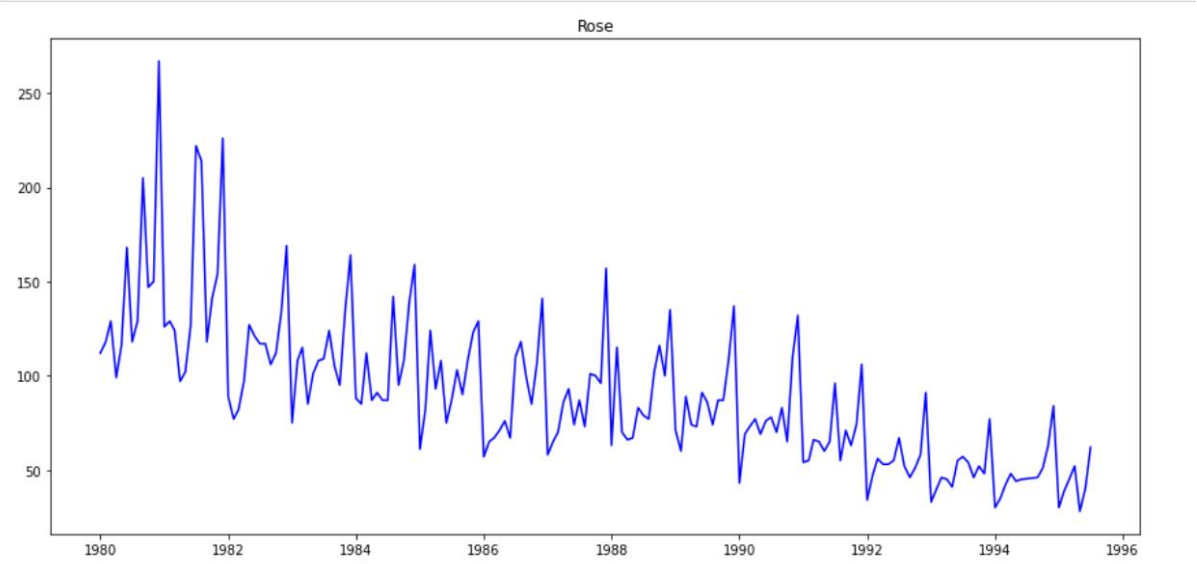


Figure 4: Rose time slot graph

Observation:

1. The sales of rose wines from 1990 to 1995. We can see the downward trend in sales.
2. We can see there is decreasing trend in the beginning which stabilizes after years and again shows the decreasing trend.
3. We can observe seasonality in the data trend and pattern repeat on yearly basis.
4. From the above graph, we could notice some missing values in the year 1994 and it must be imputed.

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

The descriptive statistics for Rose

	count	mean	std	min	25%	50%	75%	max
Rose	185.0	90.394595	39.175344	28.0	63.0	86.0	112.0	267.0

Figure 5: Rose descriptive statistics

The descriptive statistics for Sparkle

	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

Figure 6: Sparkle descriptive statistics

Checking data type of data features for rose

```
<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   Rose    185 non-null    float64
dtypes: float64(1)
memory usage: 2.9 KB
```

Figure 7: DT Rose

The dataset consists of 185 non null count and it contain missing values,it must be imputed

Checking data type of data features for sparkling

```
<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
```

Figure 8: DT Sparkling

The dataset consists of 187 observations and there are no missing values.

Imputing the missing value of Rose dataset

Checking for the null values are present in the data, from the above graph reference we could find the null data present in the year 1994.

YearMonth	Rose
1994-01-01	30.0
1994-02-01	35.0
1994-03-01	42.0
1994-04-01	48.0
1994-05-01	44.0
1994-06-01	45.0
1994-07-01	NaN
1994-08-01	NaN
1994-09-01	46.0
1994-10-01	51.0
1994-11-01	63.0
1994-12-01	84.0

Figure 9 rose missing value

1. From the observation, there are missing values in the month of July and august in 1994.
2. The missing values can be imputed using pandas function interpolate() method.
3. interpolate() function is basically used to fill NA values in the data frame or series. But, this is a very powerful function to fill the missing values. It uses various interpolation technique to fill the missing values rather than hard-coding the value.

Yearly sale boxplot for sparkling wine

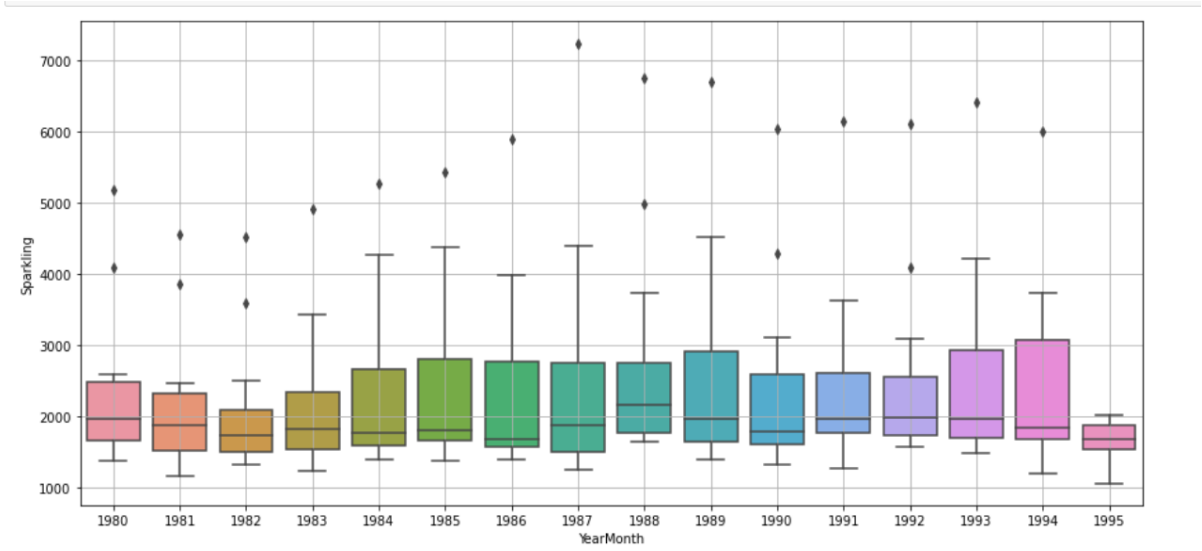


Figure 10 Yearly sale boxplot for sparkling

Observations:

1. The sale of sparkling wine has outliers for almost all the years except 1995
2. The highest mean sale for the sparkling is show in year 1988 and the lowest sales are in the year of 1995
3. The sale of sparkling, doesn't show any trend so they must work on it.
4. There is no increasing or decreasing trend in sale throughout the given time series.

Yearly sale boxplot for Rose wine

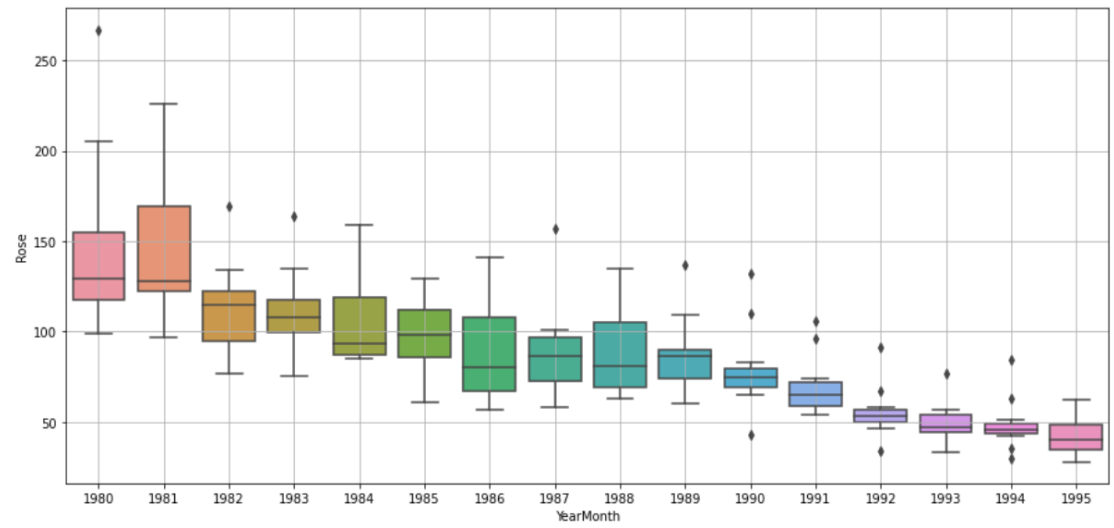


Figure 11: yearly boxplot for rose wine

Observation:

1. We see few outliers in the above graph.
2. From the yearly plot, we see that the box plot indicate downward trend.

Monthly sale boxplot for Rose wine

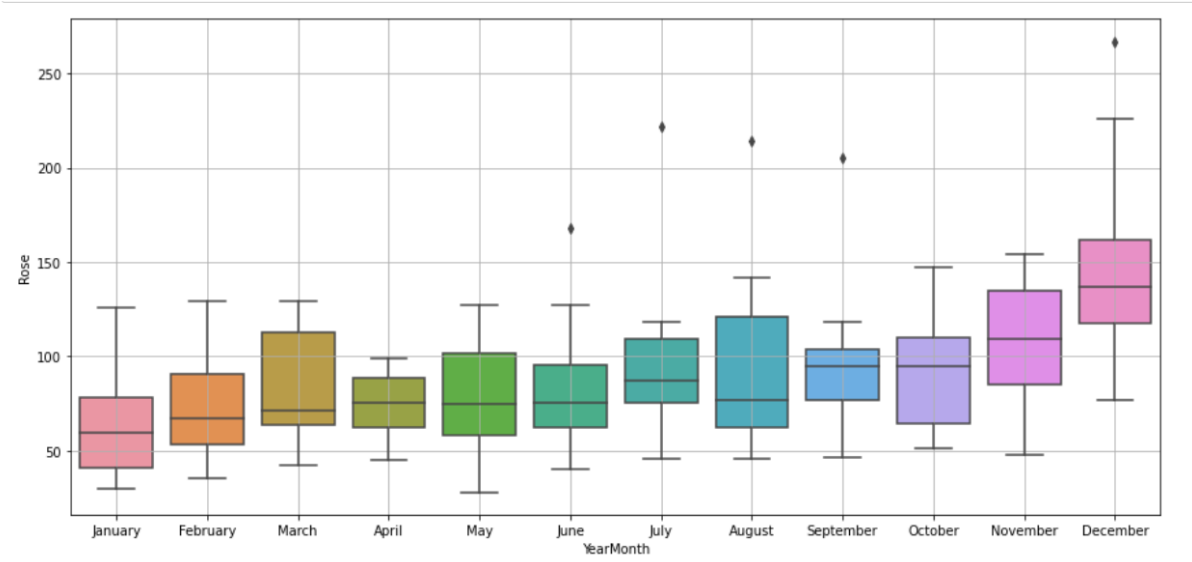


Figure 12 monthly sale boxplot rose

Observation:

1. We could see few outliers in the above graph. The outliers are present in June, July, august, September, December.
2. The highest monthly sale of wine is in the month of November - December.
3. December has the highest sales of the wine

Monthly sale boxplot for Sparkling wine

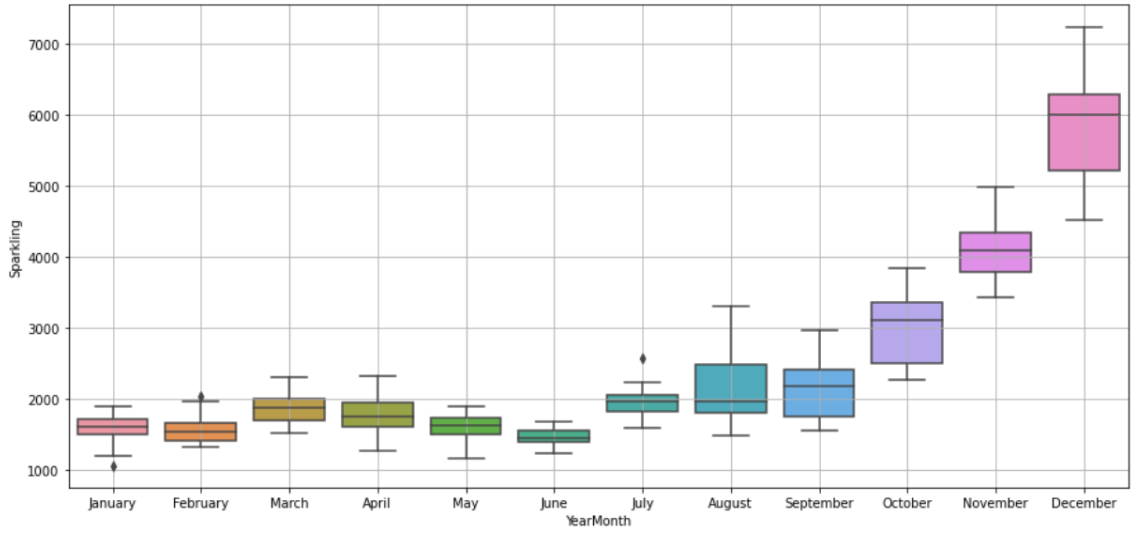


Figure 13 monthly boxplot sparkling

Observation:

1. We could see few outliers in the month of January, February, July.
2. From the above plot/graph, we could see the highest sale in the month of December.

Plotting the Empirical cumulative distribution for sparkling and rose wine:

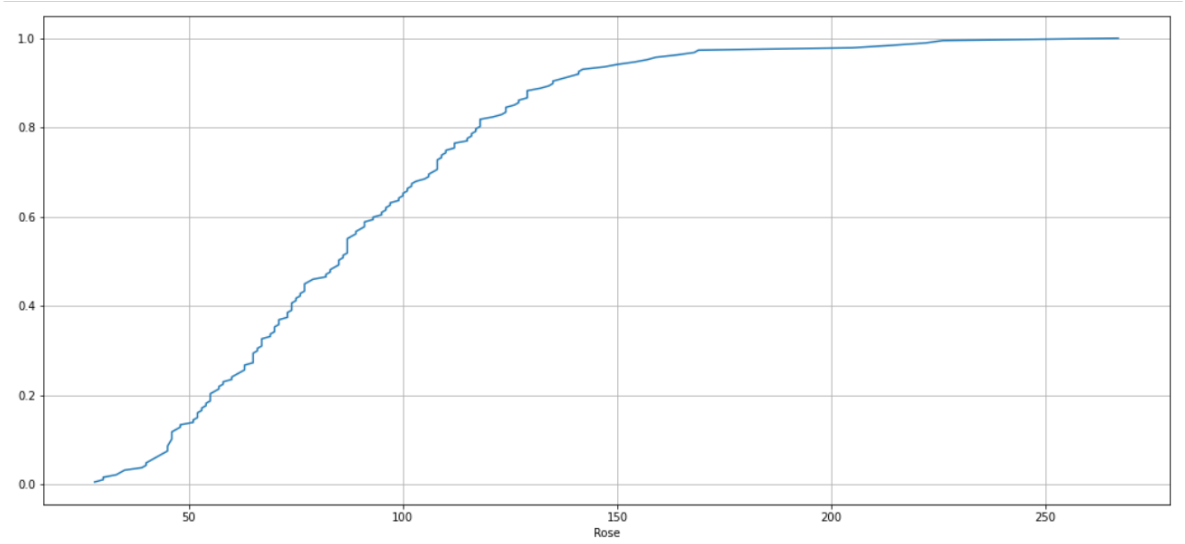


Figure 14 ECD Rose

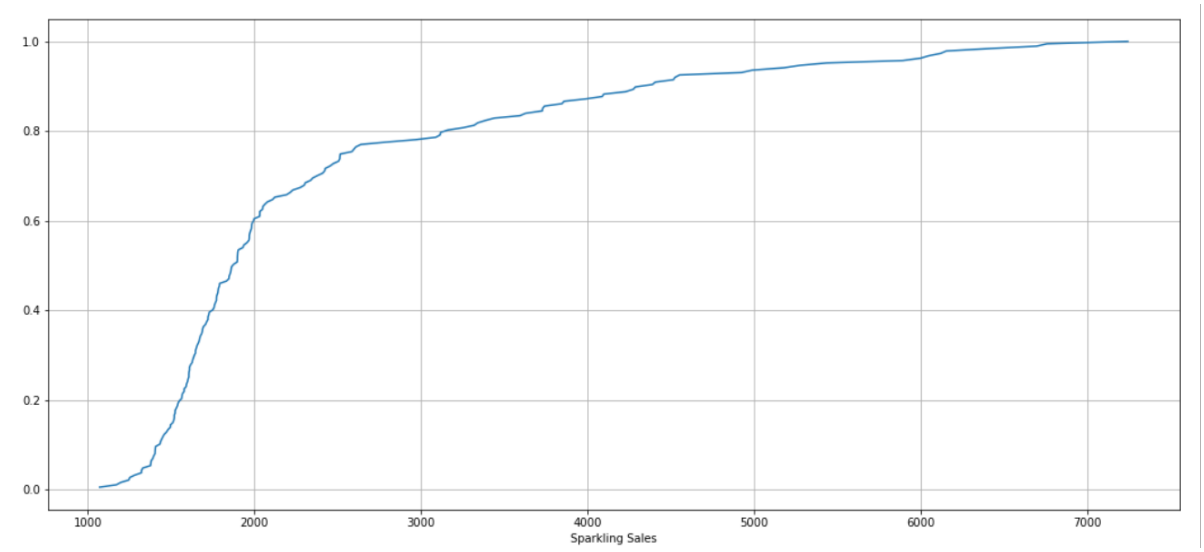


Figure 15 ECD Sparkling

- The percentage of data points refers to the what number of sales.

Plotting a graph of monthly sales of rose wine across the years

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

Table 1 Rose monthly sales trend

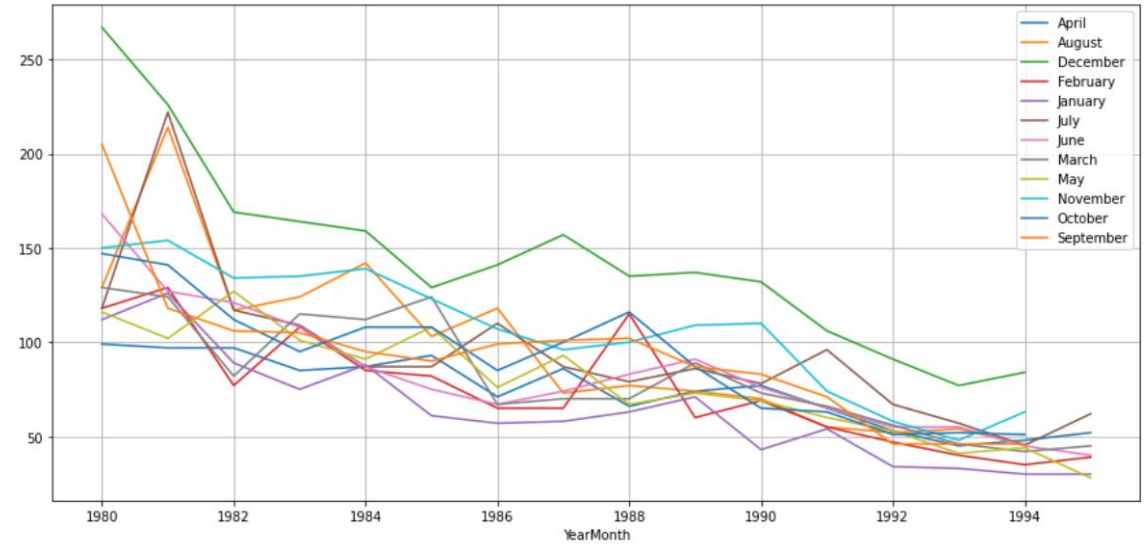


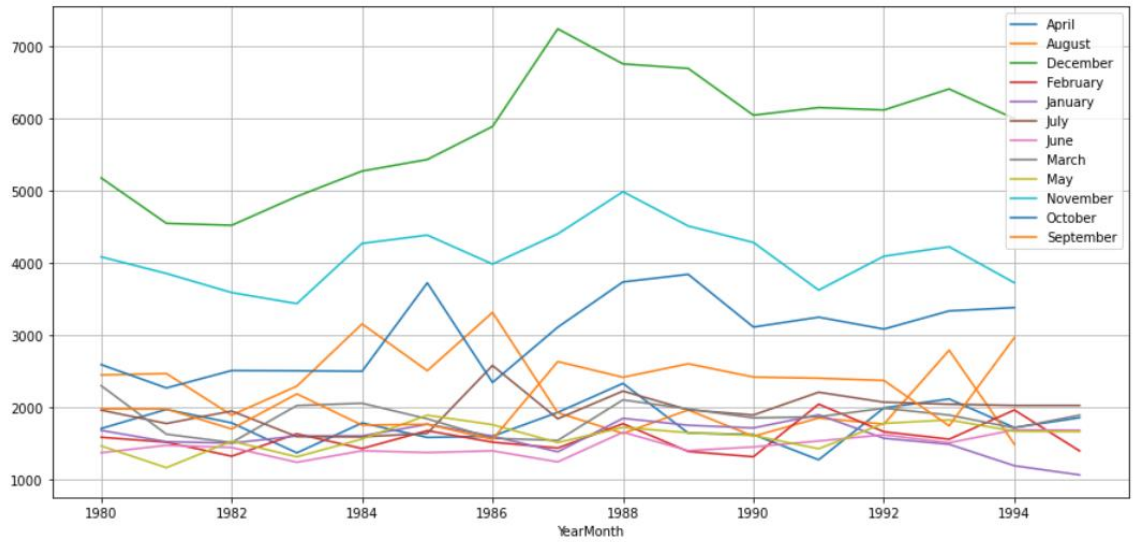
Figure 16 rose year month

Observation:

The colours indicates the month of years from January to December. The highest sales is in the month of December, January, February and may relatively lower sales.

Plotting a graph of monthly sales of Sparkling wine across the years

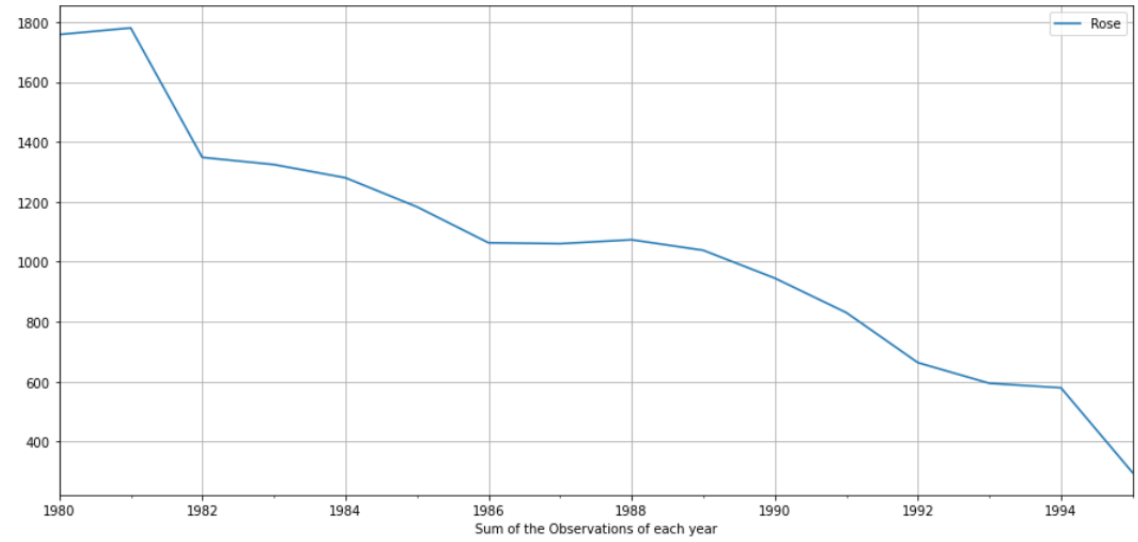
YearMonth	April	August	December	February	January	July	June	March	May	November	October	September
YearMonth												
1980	1712.0	2453.0	5179.0	1591.0	1686.0	1966.0	1377.0	2304.0	1471.0	4087.0	2596.0	1984.0
1981	1976.0	2472.0	4551.0	1523.0	1530.0	1781.0	1480.0	1633.0	1170.0	3857.0	2273.0	1981.0
1982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



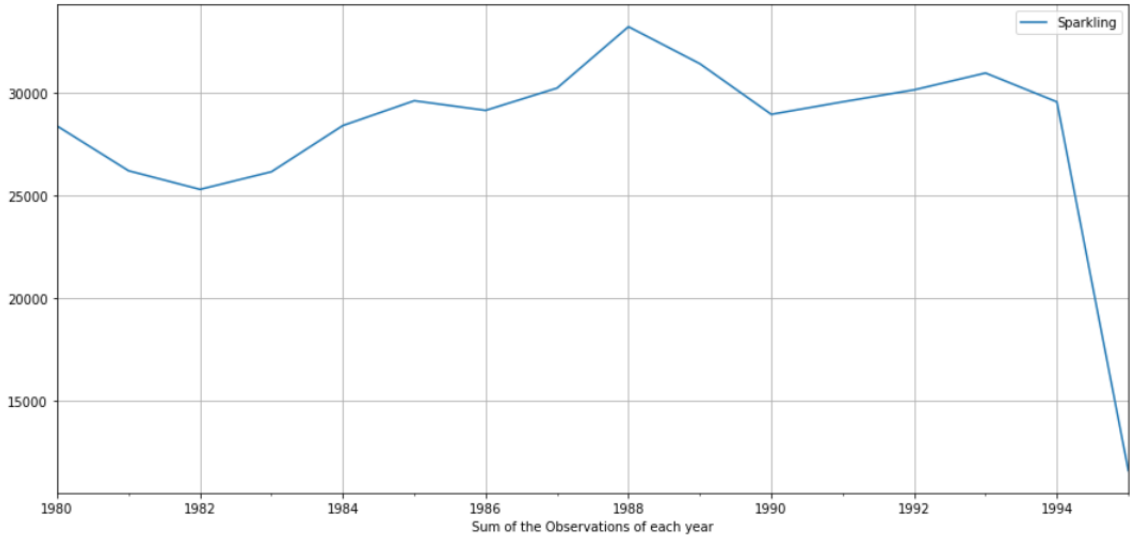
Observation:

- The highest sales of sparkling in the month of December and the lowest in the month of June for all the years.

Sum of sales of sparkling across different year:

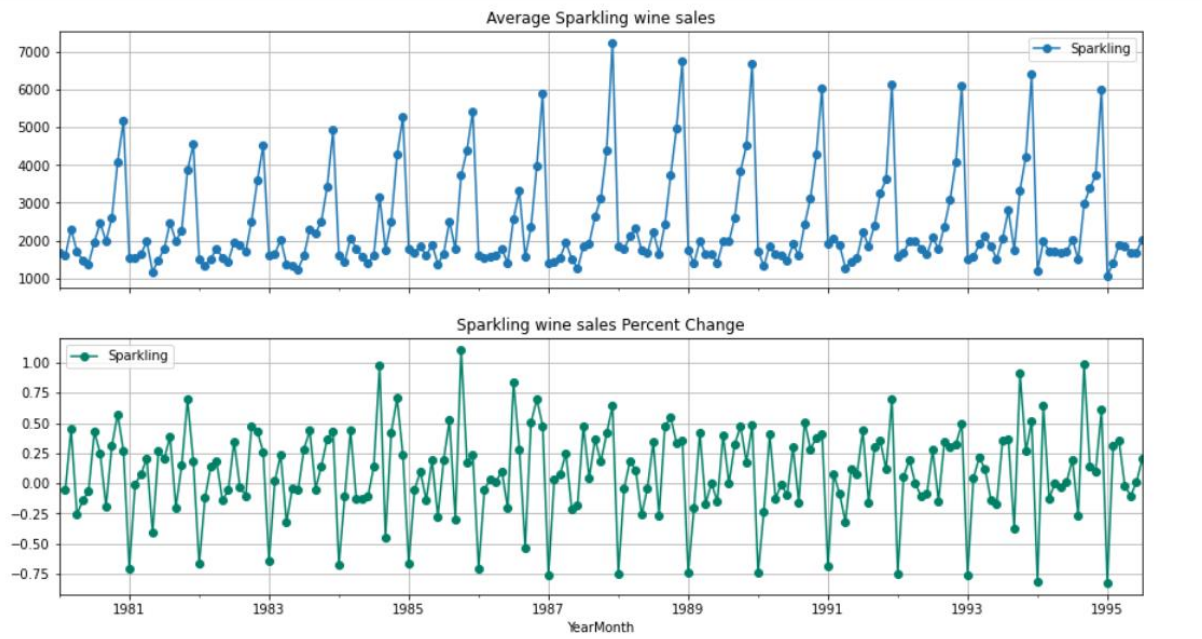


We can see that the sales of rose wine decreases from 1980-1982 and there is gradual decreasing trend.



We can see that sale of sparkling wine decreases from1980 to 1982 and steep decrease in sale from 1994.

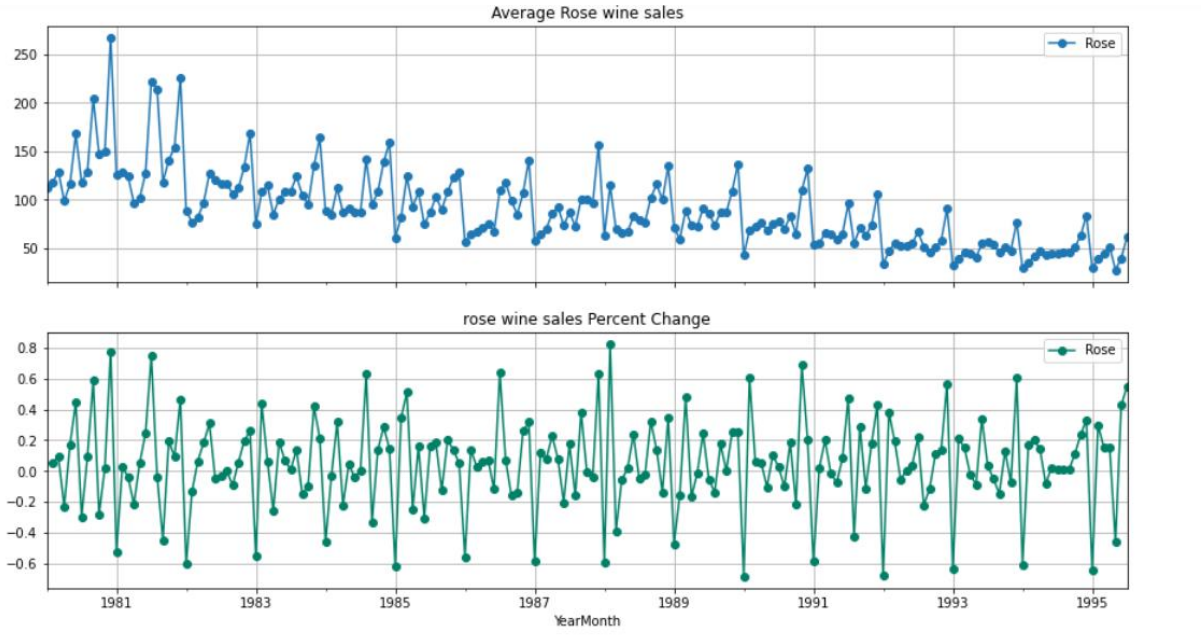
Plot the average sales per month and percentage change of sales of sparkling wine:



Observation:

1. The above two graph tell us the average sales and the percentage changes of sales with respect to the time for sparkling.
2. The average sales doesn't show any trend.

Plot the average sales per month and percentage change of sales of rose wine:

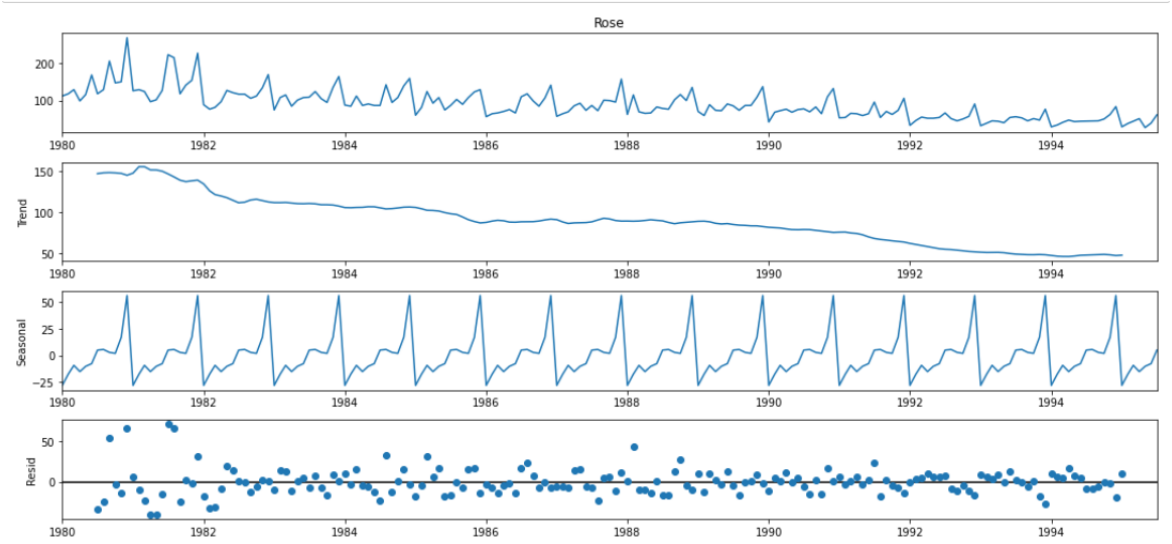


Observation:

1. The above two graph tell us the average sales and the percentage change of sales with respect to the time for rose wine.
2. The Average sales values shows decreasing trend.

Decompose the time series and plot the different components for rose

Additive Decomposition for rose



Observation:

1. There is seasonality and we observe decreasing trend.
2. Residual are located around 0
3. The residual ranging from -25b to +75 we can see some pattern in the residual so further decomposing to multiplicative model to minimize the residuals.

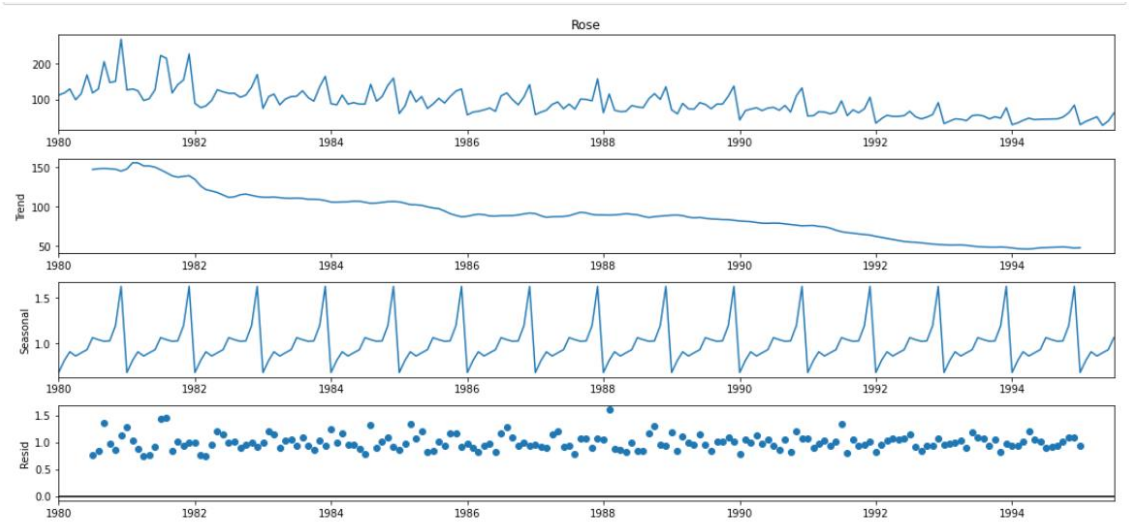
The first 12 month Additive trend, seasonality, residual value of rose wine:

```
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    147.083333
1980-08-01    148.125000
1980-09-01    148.375000
1980-10-01    148.083333
1980-11-01    147.416667
1980-12-01    145.125000
Name: trend, dtype: float64
```

```
Seasonality
 YearMonth
1980-01-01   -27.908647
1980-02-01   -17.435632
1980-03-01    -9.285830
1980-04-01   -15.098330
1980-05-01   -10.196544
1980-06-01    -7.678687
1980-07-01     4.896908
1980-08-01     5.499686
1980-09-01     2.774686
1980-10-01     1.871908
1980-11-01    16.846908
1980-12-01    55.713575
Name: seasonal, dtype: float64
```

```
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   -33.980241
1980-08-01   -24.624686
1980-09-01    53.850314
1980-10-01    -2.955241
1980-11-01   -14.263575
1980-12-01    66.161425
Name: resid, dtype: float64
```

Multiplicative Decomposition for rose



Observation:

1. For the multiplicative series, we residual around 1. The residual are minimized and multiplicative model is best for fit for decomposition.
2. There is seasonality and we observe decreasing trend.

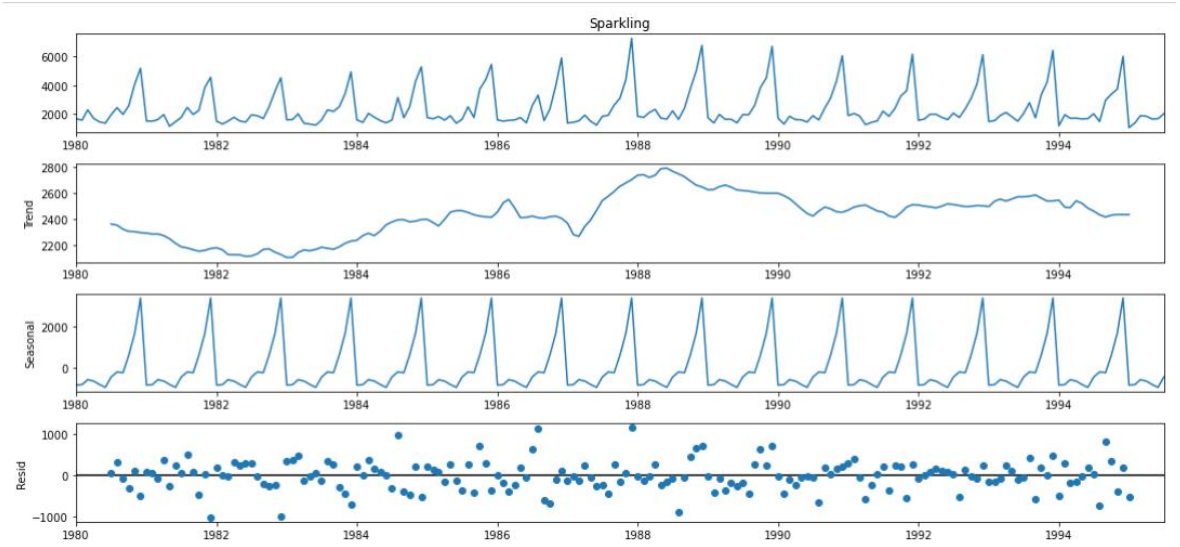
The first 12 month multiplicative trend, seasonality, residual value of rose wine:

```
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    147.083333
1980-08-01    148.125000
1980-09-01    148.375000
1980-10-01    148.083333
1980-11-01    147.416667
1980-12-01    145.125000
Name: trend, dtype: float64
```

```
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     70.835599
1980-08-01    315.999487
1980-09-01    -81.864401
1980-10-01   -307.353290
1980-11-01    109.891154
1980-12-01   -501.775513
Name: resid, dtype: float64
```

```
Seasonality
 YearMonth
1980-01-01    0.670111
1980-02-01    0.806163
1980-03-01    0.901164
1980-04-01    0.854024
1980-05-01    0.889415
1980-06-01    0.923985
1980-07-01    1.058038
1980-08-01    1.035881
1980-09-01    1.017648
1980-10-01    1.022573
1980-11-01    1.192349
1980-12-01    1.628646
Name: seasonal, dtype: float64
```

Additive Decomposition for Sparkling:



Observation:

The residual are around 0 from the plot of the residual in the decomposition.

The above graph shows seasonality and there is no proper trend.

The first 12 month multiplicative trend, seasonality, residual value of sparkling wine:

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	2360.666667
1980-08-01	2351.333333
1980-09-01	2320.541667
1980-10-01	2303.583333
1980-11-01	2302.041667
1980-12-01	2293.791667

Name: trend, dtype: float64

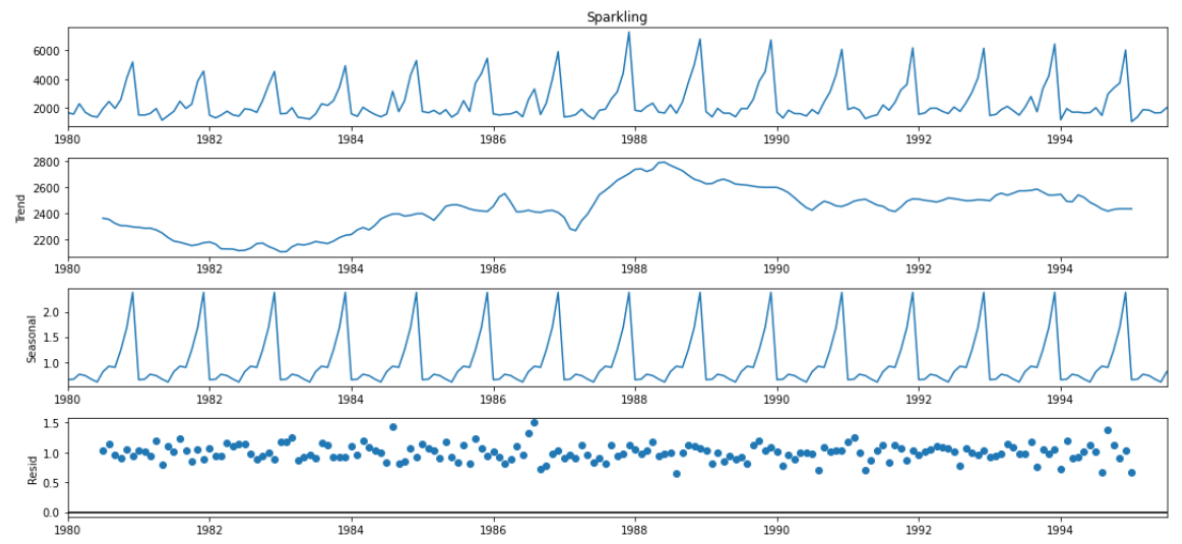
YearMonth	Seasonality
1980-01-01	-854.260599
1980-02-01	-830.350678
1980-03-01	-592.356630
1980-04-01	-658.490559
1980-05-01	-824.416154
1980-06-01	-967.434011
1980-07-01	-465.502265
1980-08-01	-214.332821
1980-09-01	-254.677265
1980-10-01	599.769957
1980-11-01	1675.067179
1980-12-01	3386.983846

Name: seasonal, dtype: float

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	70.835599
1980-08-01	315.999487
1980-09-01	-81.864401
1980-10-01	-307.353290
1980-11-01	109.891154
1980-12-01	-501.775513

Name: resid, dtype: float64

Multiplicative decomposition for Sparkling



Observation:

The residual are mostly around 1 to 1.5. The residuals are minimized and multiplicative model is best fit for decomposition.

There is seasonality and we don't observe proper trend.

The first 12 month multiplicative trend, seasonality, residual value of sparkling wine:

```
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    2360.666667
1980-08-01    2351.333333
1980-09-01    2320.541667
1980-10-01    2303.583333
1980-11-01    2302.041667
1980-12-01    2293.791667
Name: trend, dtype: float64
```

```
Seasonality
 YearMonth
1980-01-01    0.649843
1980-02-01    0.659214
1980-03-01    0.757440
1980-04-01    0.730351
1980-05-01    0.660609
1980-06-01    0.603468
1980-07-01    0.809164
1980-08-01    0.918822
1980-09-01    0.894367
1980-10-01    1.241789
1980-11-01    1.690158
1980-12-01    2.384776
Name: seasonal, dtype: float64
```

```
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    1.029230
1980-08-01    1.135407
1980-09-01    0.955954
1980-10-01    0.907513
1980-11-01    1.050423
1980-12-01    0.946770
Name: resid, dtype: float64
```

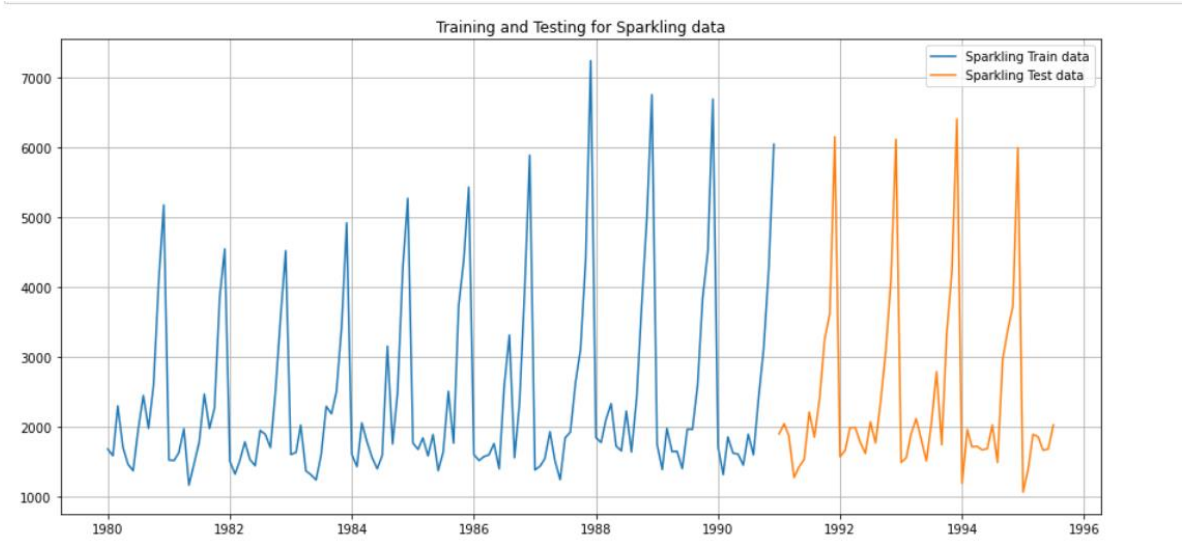
3. Split the data into training and test. The test data should start in 1991.

Splitting your dataset is essential for an unbiased evaluation of prediction performance. In most cases, it's enough to split your dataset randomly into three subsets: The training set is applied to train, or fit, your model.

Train and test dataset shape details

```
(132, 1)
(55, 1)
```

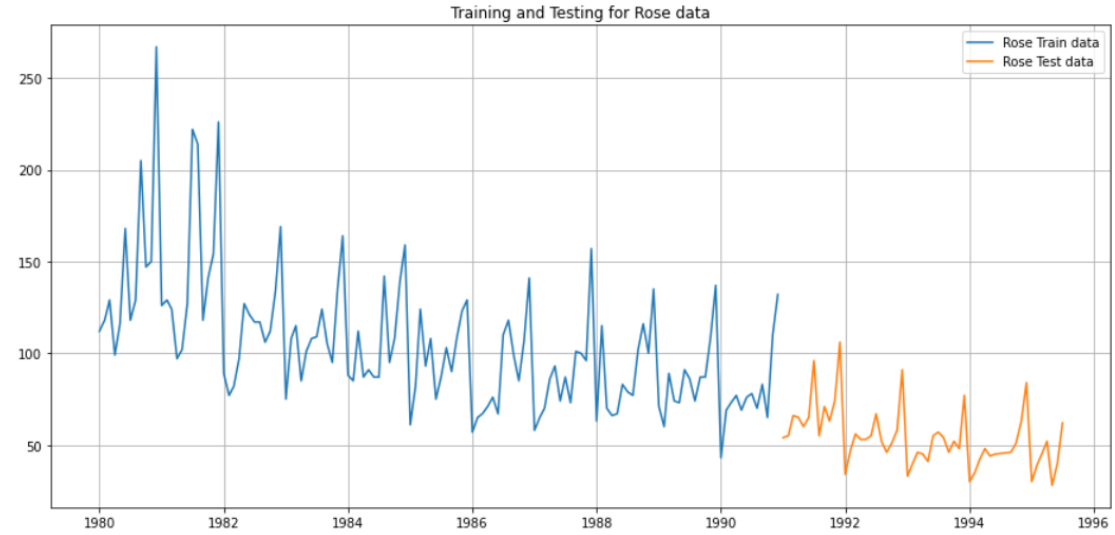
Plotting the training and testing for sparkling data



Splitting the rose data

```
(132, 1)
(55, 1)
```

Plotting training and testing for rose data



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

Building different models on the training data and forecast the values on test data and check performance of each model using root mean squared error(RMSE) values. The model with the minimum RMSE value would be our best fit model.

The different models that we are going to build as follows:

1. Linear regression model
2. Naïve approach
3. Simple average
4. Moving average
5. Simple exponential smoothing
6. Double exponential smoothing
7. Triple exponential smoothing

We could check out the performance for both rose and sparkling

Model-1: Linear regression model

Linear regression is perhaps one of the most well known and well understood algorithms in statistics and machine learning.

y=mx+c, where y is the dependent variable, m is slope, x is the independent variable and c is the intercept for a given line.

Linear regression for sparkling wine

Creation of train and test instances

```
Training Time instance for Sparkling
 [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33,
34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64,
65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95,
96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 12
1, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132]
Test Time instance for Sparkling
 [133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 15
7, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 1
82, 183, 184, 185, 186, 187]
```

The first and last few data for training and testing for sparkling

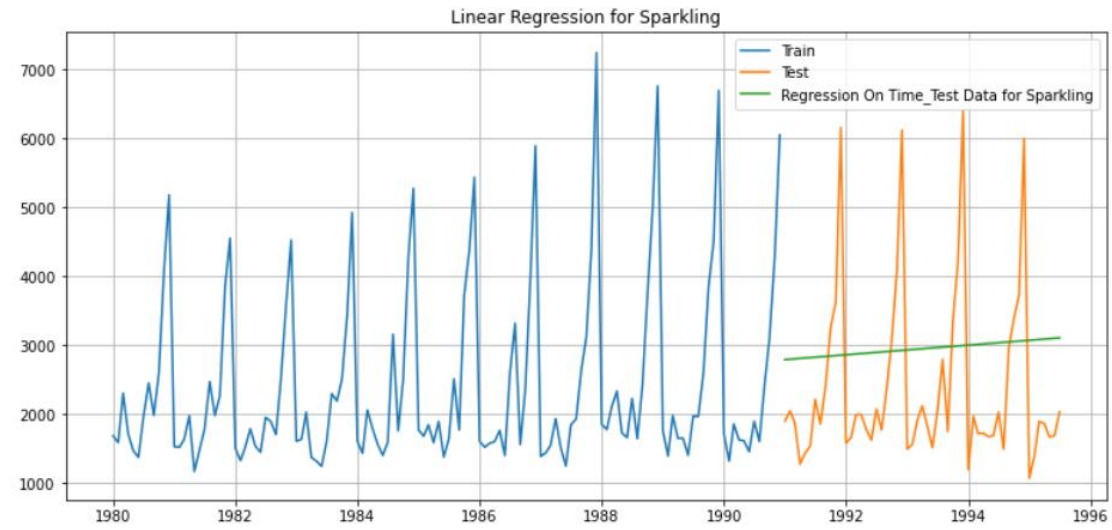
```
First few rows of Training Data for Sparkling
            Sparkling  time
YearMonth
1980-01-01       1686     1
1980-02-01       1591     2
1980-03-01       2304     3
1980-04-01       1712     4
1980-05-01       1471     5

Last few rows of Training Data for Sparkling
            Sparkling  time
YearMonth
1990-08-01       1605   128
1990-09-01       2424   129
1990-10-01       3116   130
1990-11-01       4286   131
1990-12-01       6047   132

First few rows of Test Data for Sparkling
            Sparkling  time
YearMonth
1991-01-01       1902   133
1991-02-01       2049   134
1991-03-01       1874   135
1991-04-01       1279   136
1991-05-01       1432   137


 Last few rows of Test Data for Sparkling
             Sparkling  time
 YearMonth
 1995-03-01       1897   183
 1995-04-01       1862   184
 1995-05-01       1670   185
 1995-06-01       1688   186
 1995-07-01       2031   187
```

Plotting linear regression for sparkling data



Checking the performance of the model by calculating the RMSE value:

	Test RMSE-Sparkling
Linear Regression	1389.135175

Linear regression for rose wine

Creation of training and test instances

```
Training Time instance for Rose
 [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33,
34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64,
65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95,
96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 12
1, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132]
Test Time instance for Rose
 [133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 15
7, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 1
82, 183, 184, 185, 186, 187]
```

The first and last few data for training and testing

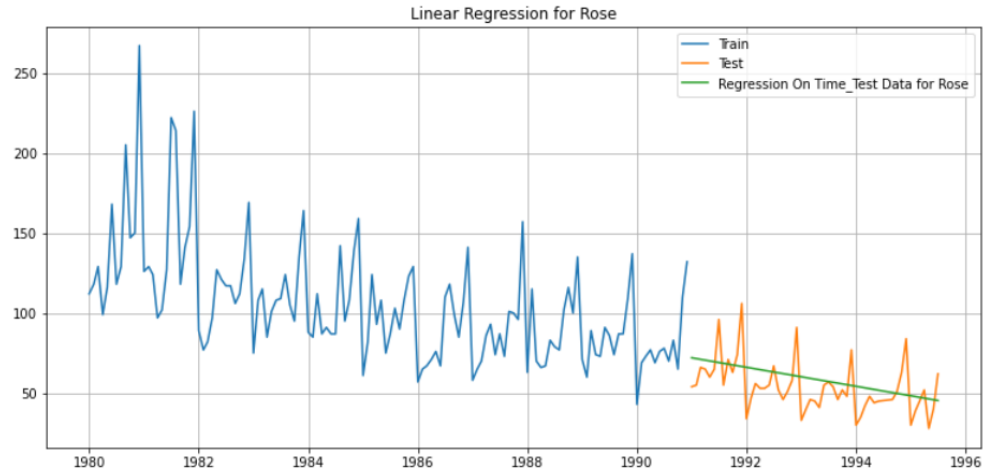
```
First few rows of Training Data for Rose
              Rose  time
YearMonth
1980-01-01  112.0     1
1980-02-01  118.0     2
1980-03-01  129.0     3
1980-04-01   99.0     4
1980-05-01  116.0     5

Last few rows of Training Data for Rose
              Rose  time
YearMonth
1990-08-01   70.0   128
1990-09-01   83.0   129
1990-10-01   65.0   130
1990-11-01  110.0   131
1990-12-01  132.0   132
```

```
First few rows of Test Data for Rose
             Rose  time
YearMonth
1991-01-01  54.0   133
1991-02-01  55.0   134
1991-03-01  66.0   135
1991-04-01  65.0   136
1991-05-01  60.0   137

Last few rows of Test Data for Rose
             Rose  time
YearMonth
1995-03-01  45.0   183
1995-04-01  52.0   184
1995-05-01  28.0   185
1995-06-01  40.0   186
1995-07-01  62.0   187
```

Plotting Linear regression for rose data



Checking the performance of the model by calculating the RMSE value.

```
For RegressionOnTime forecast on the Test Data for Rose,  RMSE is 15.269
```

	Test RMSE-Rose
Linear Regression	15.268955

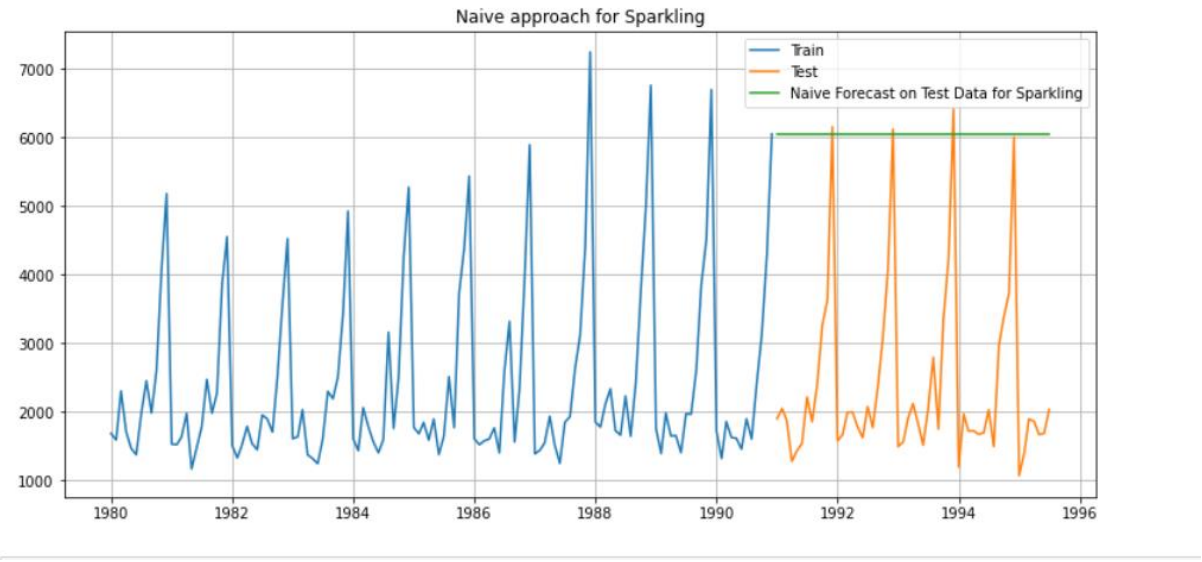
Model 2 : Naïve Approach

A naive approach consists of calculating a histogram of angles, assuming the accumulation of points corresponding to the directions of interest will result in visible peaks. In sparkling wine data, the value of train dataset was 6047.

```
YearMonth
1991-01-01    6047
1991-02-01    6047
1991-03-01    6047
1991-04-01    6047
1991-05-01    6047
Name: naive, dtype: int64
```

\

Plotting graph for Naïve approach for sparkling:



Checking the performance of the model by calculating the RMSE value

```
For RegressionOnTime forecast on the Test Data for Sparkling,  RMSE is 3864.279
```

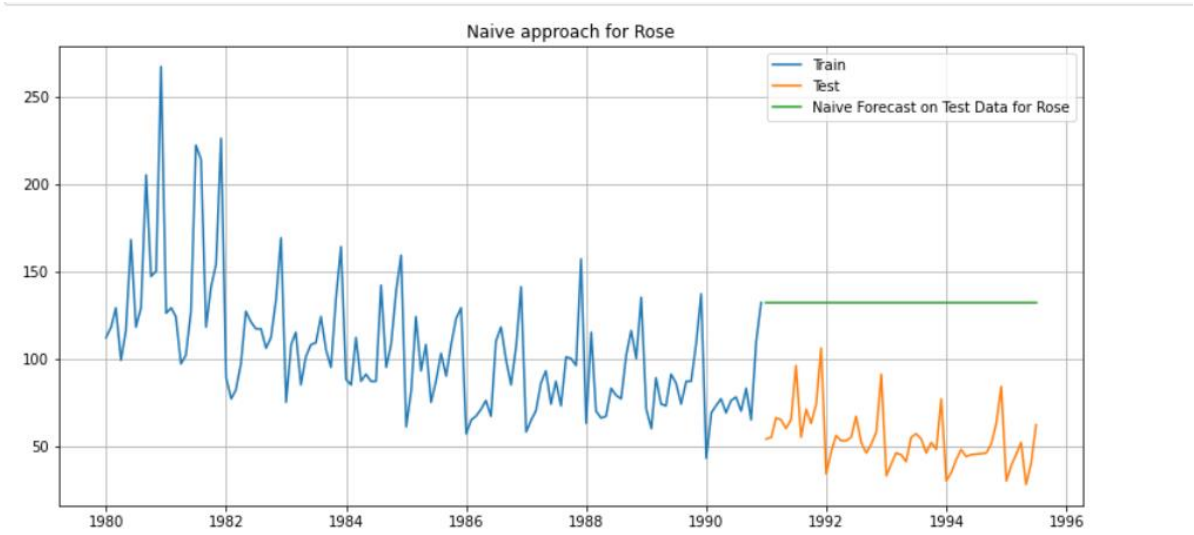
	Test RMSE-Sparkling
Linear Regression	1389.135175
Naive Approach	3864.279352

Naïve approach for Rose

The last value of train dataset was 132. The model will predict 132 for every instance of the test data.

```
YearMonth
1991-01-01    132.0
1991-02-01    132.0
1991-03-01    132.0
1991-04-01    132.0
1991-05-01    132.0
Name: naive, dtype: float64
```

Plotting graph for Naïve Approach for rose



Checking the performance of the model by calculating the RMSE value

```
For RegressionOnTime forecast on the Test Data for Rose,  RMSE is 79.719
```

	Test RMSE-Rose
Linear Regression	15.268955
Naive Approach	79.718773

Model 3 : Simple Average

The simple average approach of time series forecasting is a very simple method of forecasting the values. We average the data by months or quarters or years and then calculate the average for the period.

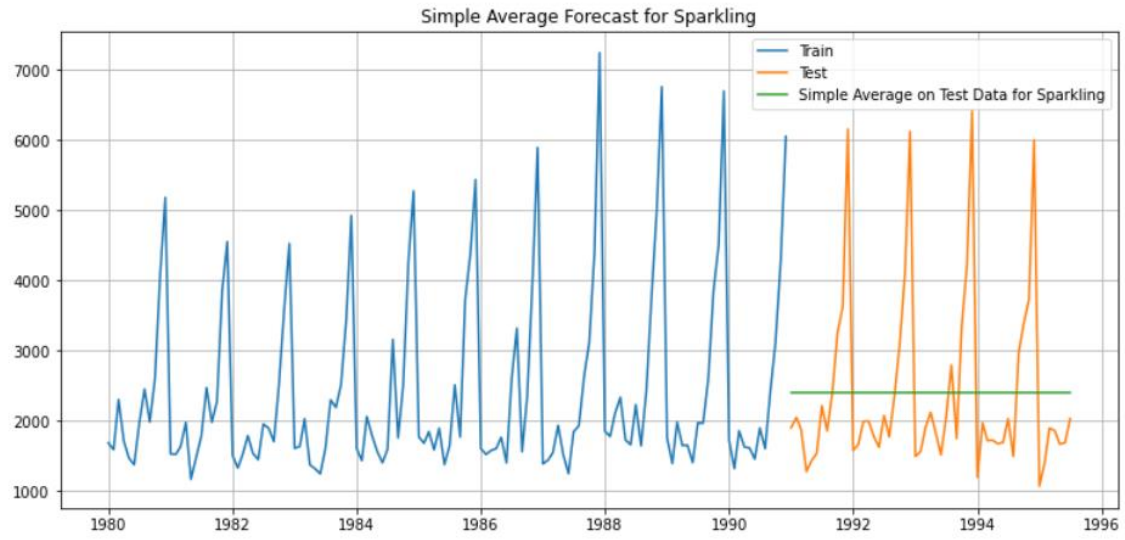
The average of training data by months is considered for forecasting for the testing data.

Simple average for sparkling

In the case of data, the average of training data is 2403.780.

YearMonth	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

Plotting graph for simple average for sparkling



From the plot above, we could see that the predictions on test were made of the same value as average of train data which is 2403.780. checking the performance by using RMSE value.

```
For Simple Average forecast on the Test Data for Sparkling,  RMSE is 1275.082
```

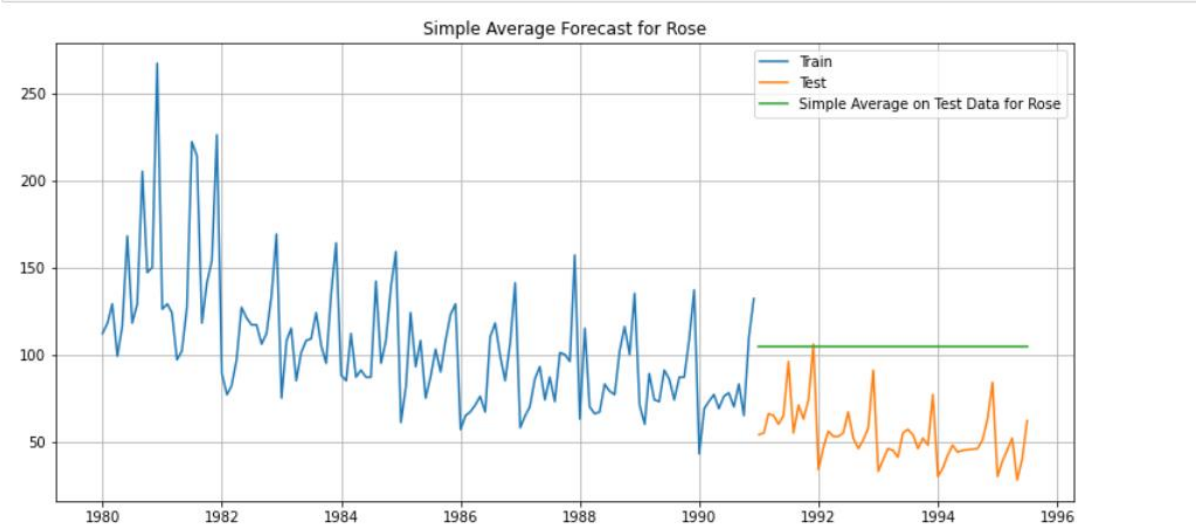
	Test RMSE-Sparkling
Linear Regression	1389.135175
Naive Approach	3864.279352
Simple Average	1275.081804

Simple average for Rose

The average of sales of training data for rose is calculated which is the predicted forecast value for the test data. Our predicted values for the entire data set wil be 104.93

YearMonth	Rose	mean_forecast
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

Plotting graph for simple average for rose



From the plot above we can see that the predictions on test were made of the same value as the average of train data which is 104.93

```
For Simple Average forecast on the Test Data for Rose,  RMSE is 53.461
```

	Test RMSE-Rose
Linear Regression	15.268955
Naive Approach	79.718773
Simple Average	53.460570

Model 4 : Moving Average method

A moving average is a technical indicator that investors and traders use to determine the trend direction of securities. It is calculated by adding up all the data points during a specific period and dividing the sum by the number of time periods. Moving averages help technical traders to generate trading signals.

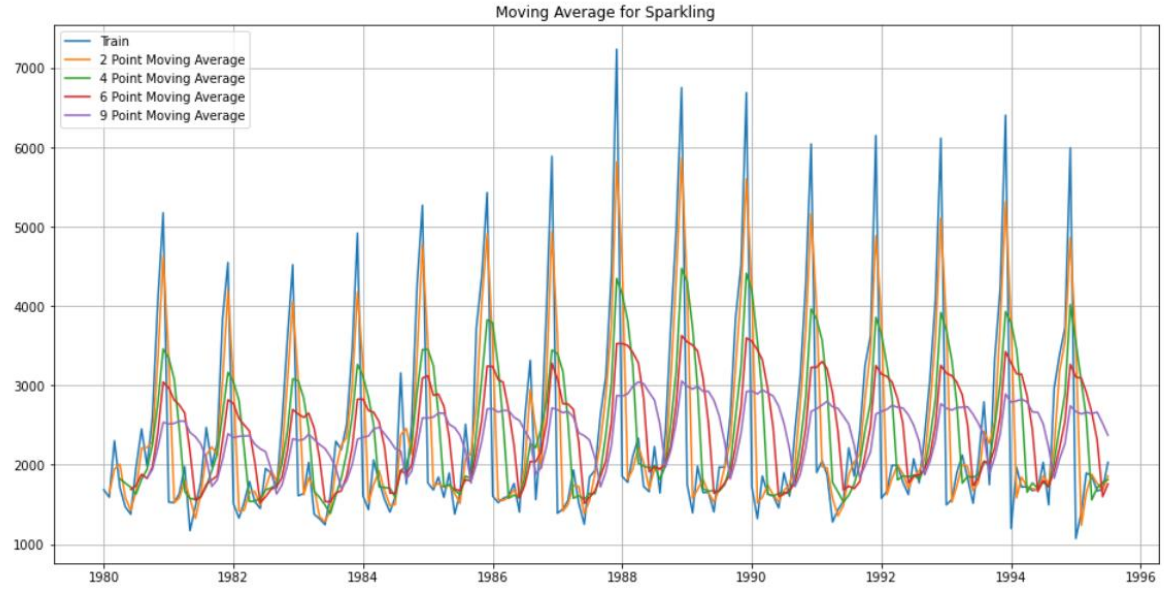
Moving average for sparkling

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

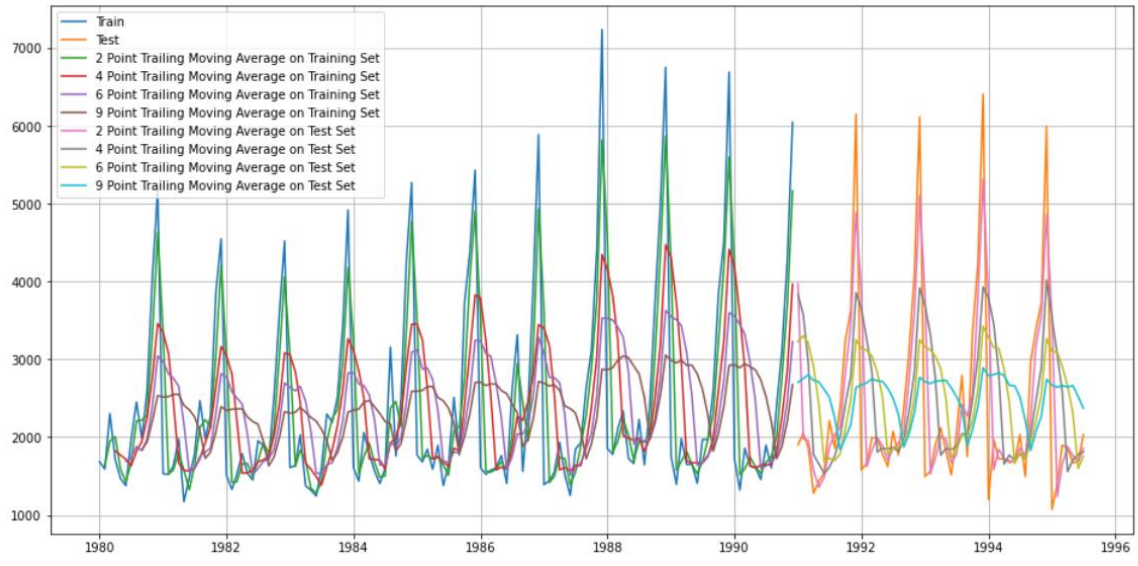
Checking the head of the moving averages for all four rolling intervals

Plotting the graph for moving average for entire sparkling data

```
Text(0.5, 1.0, 'Moving Average for Sparkling')
```



Plotting the graph for moving average for sparkling of both train and test data




```
For 2 point Moving Average Model forecast on the Training Data for Sparkling,  RMSE is 813.401
For 4 point Moving Average Model forecast on the Training Data for Sparkling,  RMSE is 1156.590
For 6 point Moving Average Model forecast on the Training Data for Sparkling,  RMSE is 1283.927
For 9 point Moving Average Model forecast on the Training Data for Sparkling,  RMSE is 1346.278
```

	Test RMSE-Sparkling
Linear Regression	1389.135175
Naive Approach	3864.279352
Simple Average	1275.081804
2point Trailing Moving Average	813.400684
4point Trailing Moving Average	1156.589694
6point Trailing Moving Average	1283.927428
9point Trailing Moving Average	1346.278315

Moving average for rose

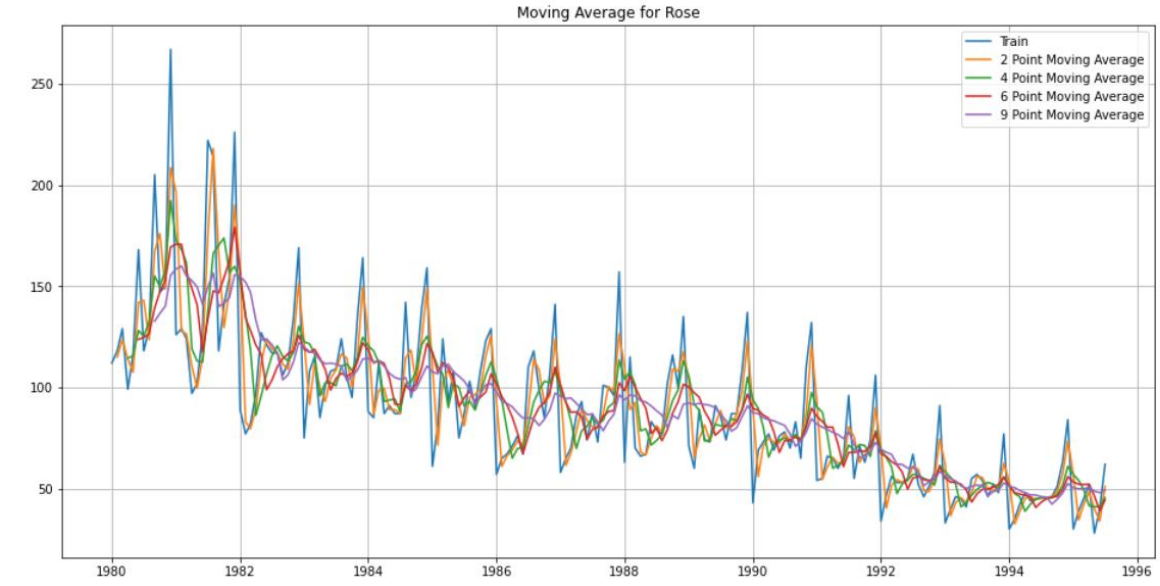
Checking the head of the moving averages for all the 4 rolling intervals

YearMonth	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

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

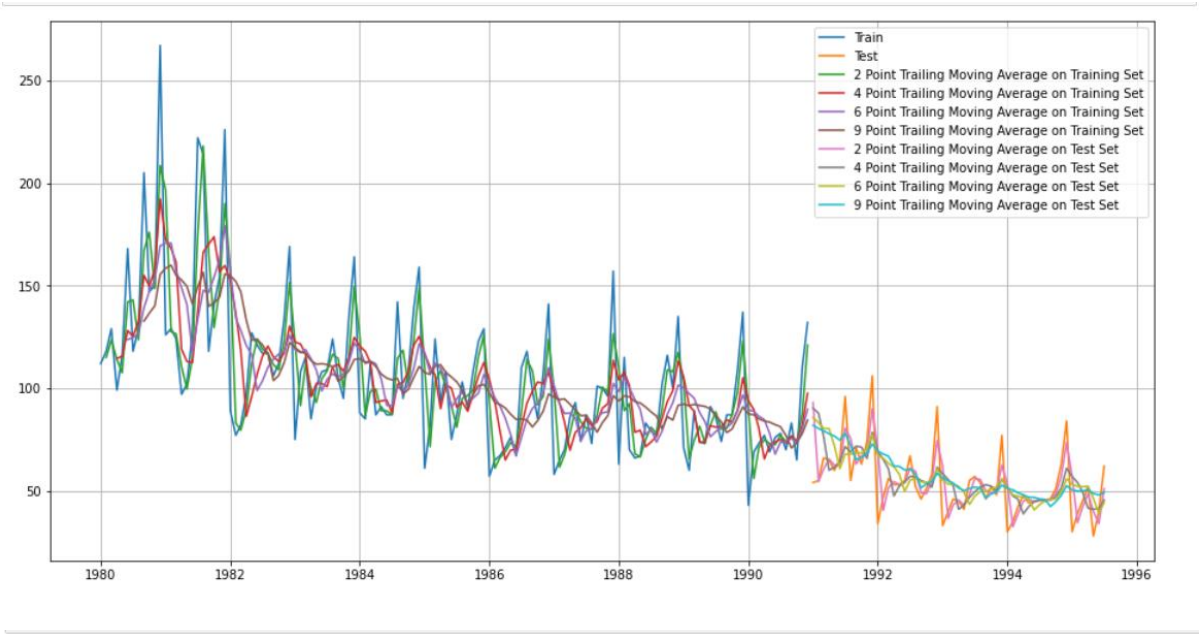
Plotting the graph for moving average for entire rose data

```
Text(0.5, 1.0, 'Moving Average for Rose')
```



Now creating a train and test data and calculating the moving average on train and forecasting on train for all 4 groups of rolling averages of 2,4,6, and 8 and calculating the RMSE on the test data for the same.

Plotting the graph for moving average for rose of both train and test data




```
For 2 point Moving Average Model forecast on the Training Data for Rose,  RMSE is 11.529
For 4 point Moving Average Model forecast on the Training Data for Rose,  RMSE is 14.451
For 6 point Moving Average Model forecast on the Training Data for Rose,  RMSE is 14.566
For 9 point Moving Average Model forecast on the Training Data for Rose,  RMSE is 14.728
```

	Test RMSE-Rose
Linear Regression	15.268955
Naive Approach	79.718773
Simple Average	53.460570
2point Trailing Moving Average	11.529278
4point Trailing Moving Average	14.451403
6point Trailing Moving Average	14.566327
9point Trailing Moving Average	14.727630

The RMSE for moving average with 2 point rolling is performing better.

Model 5 : Simple Exponential Smoothing (SES) – ETS(A,N,N)

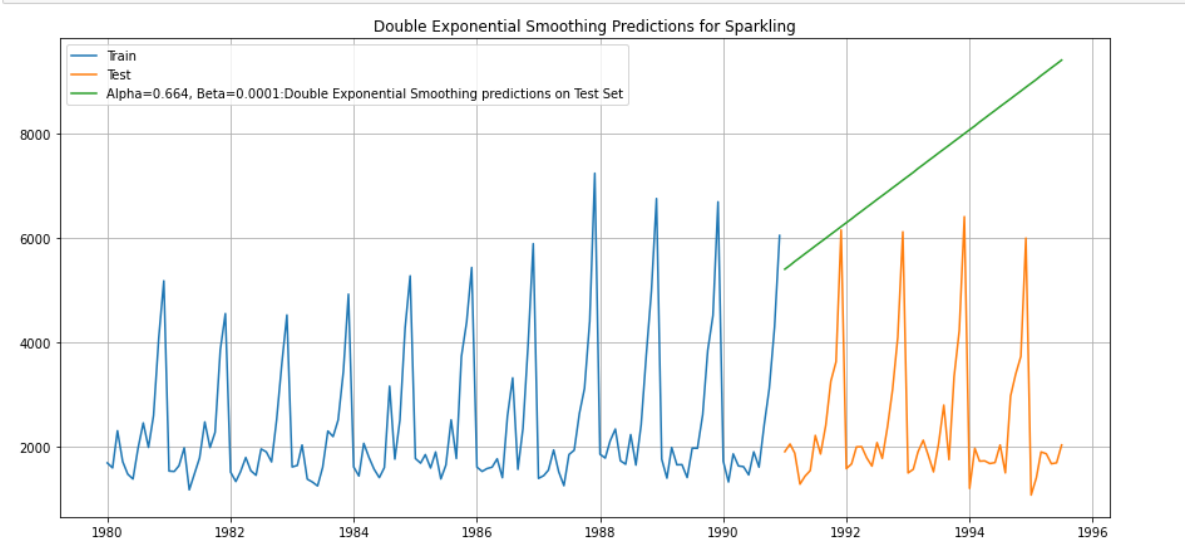
Single Exponential Smoothing, SES for short, also called Simple Exponential Smoothing, is a time series forecasting method for univariate data without a trend or seasonality. It requires a single parameter, called alpha (a), also called the smoothing factor or smoothing coefficient.

```
{'smoothing_level': 0.049607360581862936,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 1818.535750008871,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

As simple exponential smoothing model only works with the level, trend and seasonality parameters are blank. Smoothing level parameters (Alpha) = 0.0496

YearMonth	Sparkling	predict
1991-01-01	1902	2724.932624
1991-02-01	2049	2724.932624
1991-03-01	1874	2724.932624
1991-04-01	1279	2724.932624
1991-05-01	1432	2724.932624

Plotting the graph of forecast on the test data and look at the predictions:



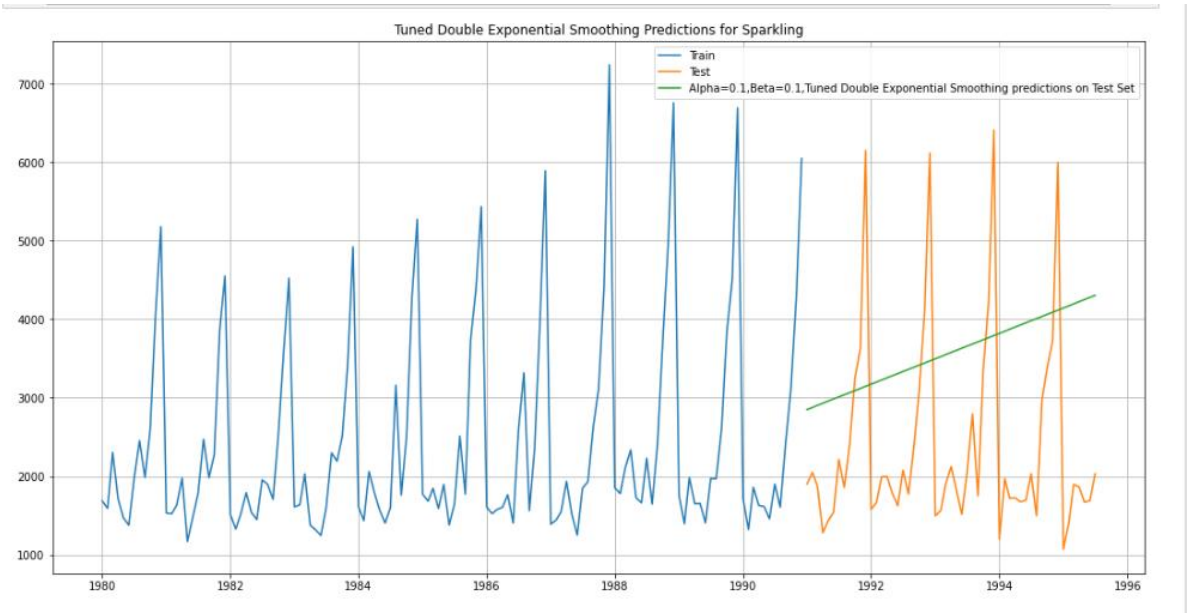

```
DES RMSE: 5291.8798332269125
```

	Test RMSE-Sparkling
Linear Regression	1389.135175
Naive Approach	3864.279352
Simple Average	1275.081804
2point Trailing Moving Average	813.400684
4point Trailing Moving Average	1156.589694
6point Trailing Moving Average	1283.927428
9point Trailing Moving Average	1346.278315
Alpha=0.049:Simple Exponential Smoothing	1316.035487
Alpha=0.02:Tuned Simple Exponential Smoothing	1279.495201
Alpha=0.66,Beta=0.0001:DoubleExponentialSmoothing	5291.879833

Tuned Simple Exponential smoothing for sparkling

	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

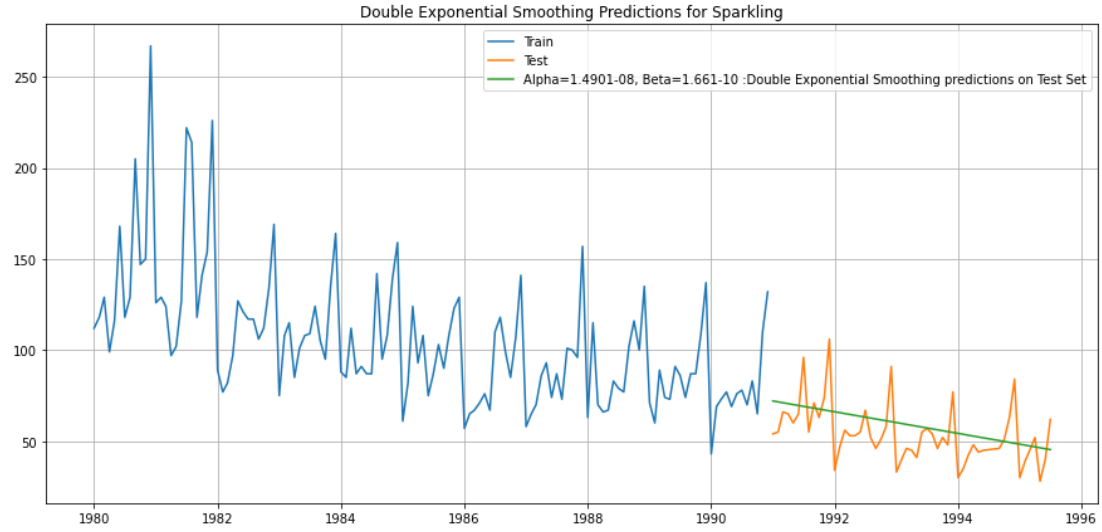
Plotting the graph for tuned simple exponential smoothing predictions for sparkling:



A prediction of approx. 2500 is made for the test data using the tuned simple exponential smoothing model. The RMSE for test is reduced from 1316.035 to 1279.495

	Test RMSE-Sparkling
Linear Regression	1389.135175
Naive Approach	3864.279352
Simple Average	1275.081804
2point Trailing Moving Average	813.400684
4point Trailing Moving Average	1156.589694
6point Trailing Moving Average	1283.927428
9point Trailing Moving Average	1346.278315
Alpha=0.049:Simple Exponential Smoothing	1316.035487
Alpha=0.02:Tuned Simple Exponential Smoothing	1279.495201
Alpha=0.66,Beta=0.0001:DoubleExponentialSmoothing	5291.879833
Alpha=0.1,Beta=0.1:Tuned Double Exponential Smoothing	1778.564670

Double exponential smoothing predications for sparkling



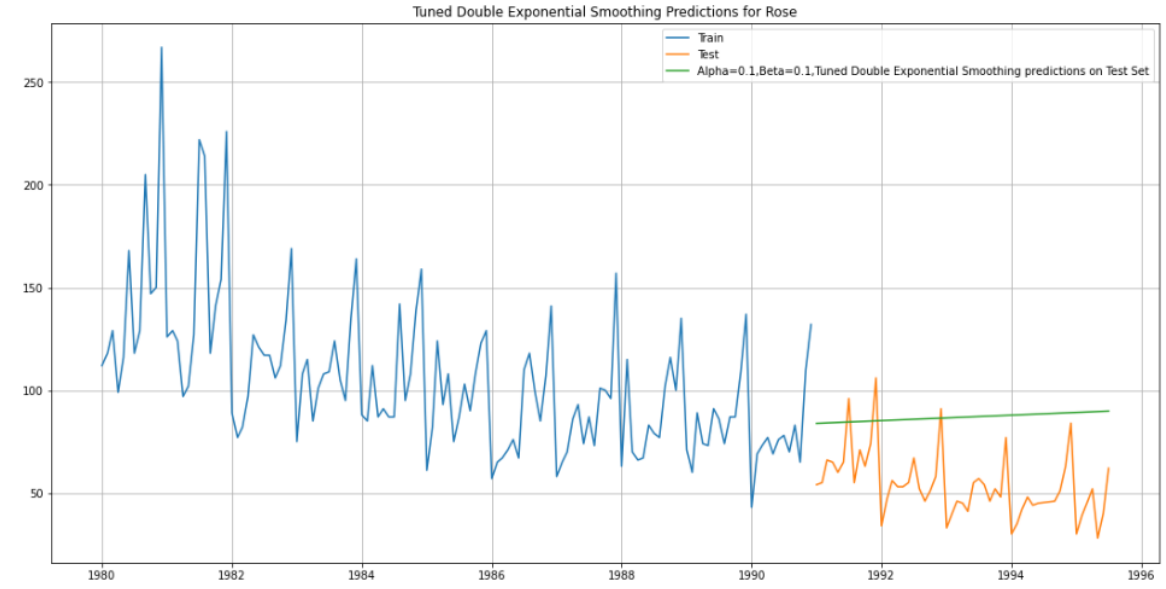
```
DES RMSE: 15.268943764436564
```

	Test RMSE-Rose
Linear Regression	15.268955
Naive Approach	79.718773
Simple Average	53.460570
2point Trailing Moving Average	11.529278
4point Trailing Moving Average	14.451403
6point Trailing Moving Average	14.566327
9point Trailing Moving Average	14.727630
Alpha= 0.098:Simple Exponential Smoothing	36.796227
Alpha=0.07:Tuned Simple Exponential Smoothing	36.435772
Alpha=1.4901-08,,Beta=1.661-10:Double Exponential Smoothing	15.268944

Tuned double exponential smoothing for rose

	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

Plotting the graph for tuned double exponential smoothing predictions for sparkling

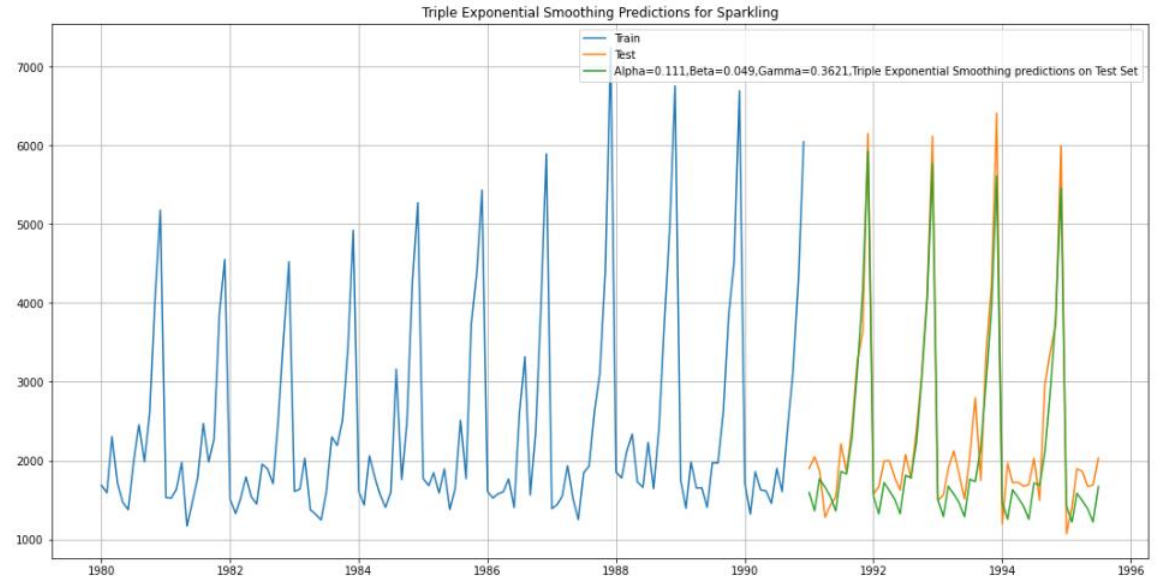


	Test RMSE-Rose
Linear Regression	15.268955
Naive Approach	79.718773
Simple Average	53.460570
2point Trailing Moving Average	11.529278
4point Trailing Moving Average	14.451403
6point Trailing Moving Average	14.566327
9point Trailing Moving Average	14.727630
Alpha= 0.098:Simple Exponential Smoothing	36.796227
Alpha=0.07:Tuned Simple Exponential Smoothing	36.435772
Alpha=1.4901-08,,Beta=1.661-10:Double Exponential Smoothing	15.268944
Alpha=0.1,,Beta=0.1:Tuned Double Exponential Smoothing	36.923416

Triple exponential smoothing predictions

```
{'smoothing_level': 0.11106668752955826,
 'smoothing_trend': 0.04936072355729082,
 'smoothing_seasonal': 0.3621821387810734,
 'damping_trend': nan,
 'initial_level': 2360.4089797373545,
 'initial_trend': 0.9992288111047797,
 'initial_seasons': array([0.71936124, 0.6984697 , 0.90024844, 0.80991063, 0.66820986,
        0.66898271, 0.87875613, 1.11648842, 0.90067181, 1.17297733,
        1.82687893, 2.27815792]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

YearMonth	Sparkling	auto_predict
1991-01-01	1902	1591.299973
1991-02-01	2049	1360.408886
1991-03-01	1874	1767.949510
1991-04-01	1279	1661.619432
1991-05-01	1432	1547.414170



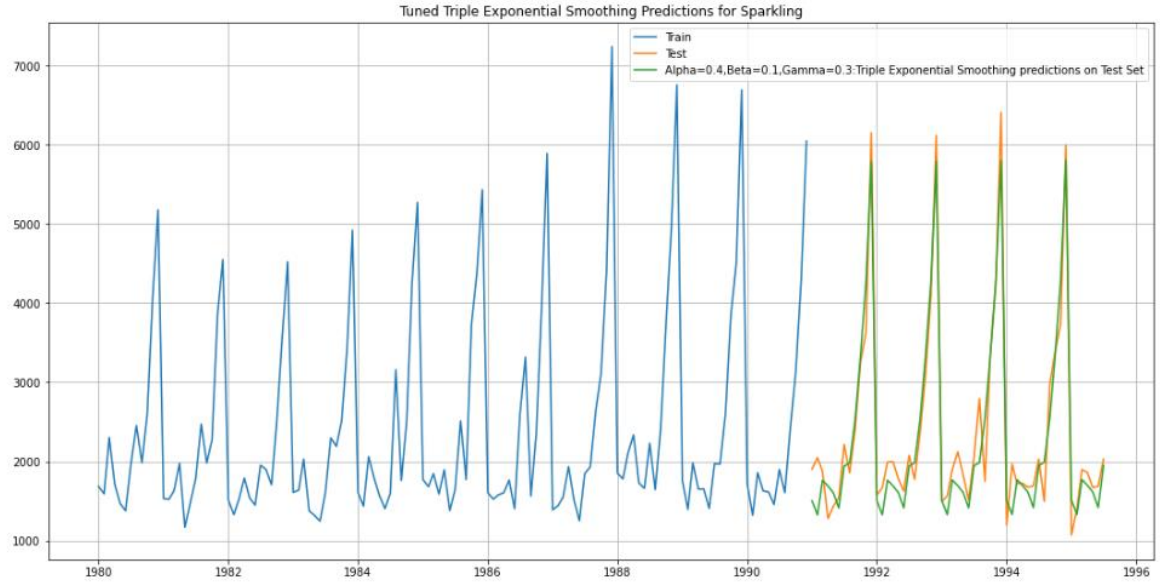

```
TES RMSE: 380.3984781419219
```

	Test RMSE-Sparkling
Linear Regression	1389.135175
Naive Approach	3864.279352
Simple Average	1275.081804
2point Trailing Moving Average	813.400684
4point Trailing Moving Average	1156.589694
6point Trailing Moving Average	1283.927428
9point Trailing Moving Average	1346.278315
Alpha=0.049:Simple Exponential Smoothing	1316.035487
Alpha=0.02:Tuned Simple Exponential Smoothing	1279.495201
Alpha=0.66,Beta=0.0001:DoubleExponentialSmoothing	5291.879833
Alpha=0.1,Beta=0.1:Tuned Double Exponential Smoothing	1778.564670
Alpha=0.111,Beta=0.049,Gamma=0.362:Triple Exponential Smoothing	380.398478

Tuned triple exponential smoothing predictions for sparkling

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE
302	0.4	0.1	0.3	381.106645	326.579641
201	0.3	0.1	0.2	375.956510	342.464413
110	0.2	0.2	0.1	395.987244	345.931571
131	0.2	0.4	0.2	401.704682	349.425739
222	0.3	0.3	0.3	396.692796	353.602587

Plotting the graph of forecast on the test data and look at the predictions for rose

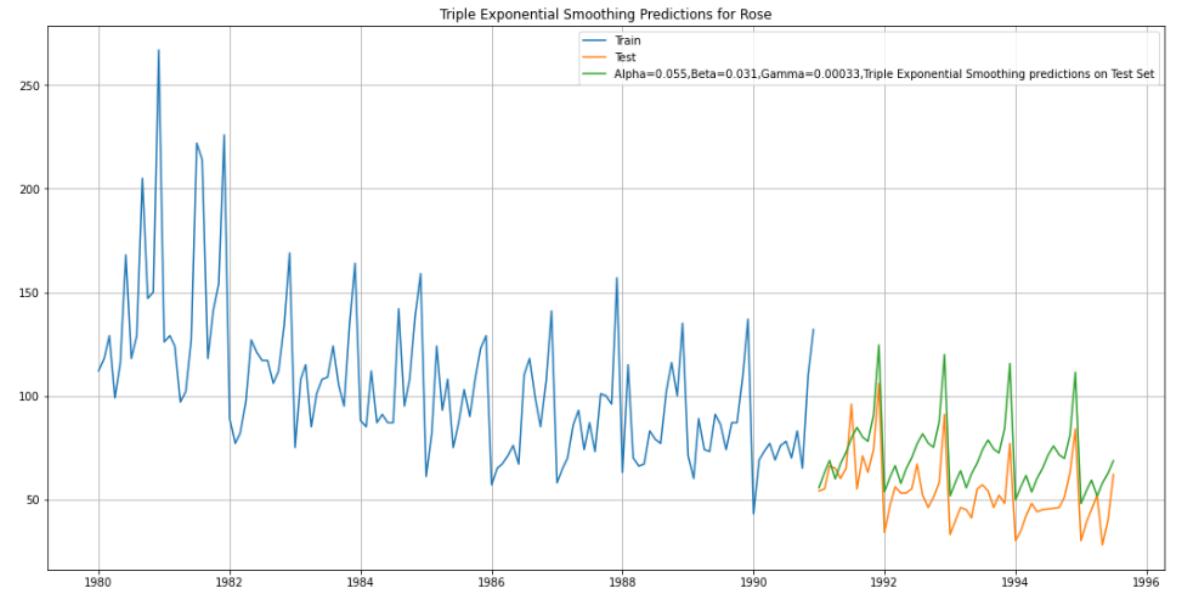


	Test RMSE-Sparkling
Linear Regression	1389.135175
Naive Approach	3864.279352
Simple Average	1275.081804
2point Trailing Moving Average	813.400684
4point Trailing Moving Average	1156.589694
6point Trailing Moving Average	1283.927428
9point Trailing Moving Average	1346.278315
Alpha=0.049:Simple Exponential Smoothing	1316.035487
Alpha=0.02:Tuned Simple Exponential Smoothing	1279.495201
Alpha=0.66,Beta=0.0001:DoubleExponentialSmoothing	5291.879833
Alpha=0.1,Beta=0.1:Tuned Double Exponential Smoothing	1778.564670
Alpha=0.111,Beta=0.049,Gamma=0.362:Triple Exponential Smoothing	380.398478
Alpha=0.4,Beta=0.01,Gamma=0.3:Tuned Triple Exponential Smoothing	326.579641

Triple Exponential smoothing for rose

```
{'smoothing_level': 0.05509258651447915,
 'smoothing_trend': 0.03163443011388579,
 'smoothing_seasonal': 0.00033441920536960617,
 'damping_trend': nan,
 'initial_level': 162.24448448772696,
 'initial_trend': 0.9924159109944972,
 'initial_seasons': array([0.69939026, 0.79380649, 0.86893412, 0.75865299, 0.85377453,
        0.9282575 , 1.02003364, 1.08767274, 1.03068915, 1.00761385,
        1.17626069, 1.61916255]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

YearMonth	Rose	auto_predict
1991-01-01	54.0	55.663816
1991-02-01	55.0	62.993228
1991-03-01	66.0	68.738503
1991-04-01	65.0	59.835212
1991-05-01	60.0	67.118704

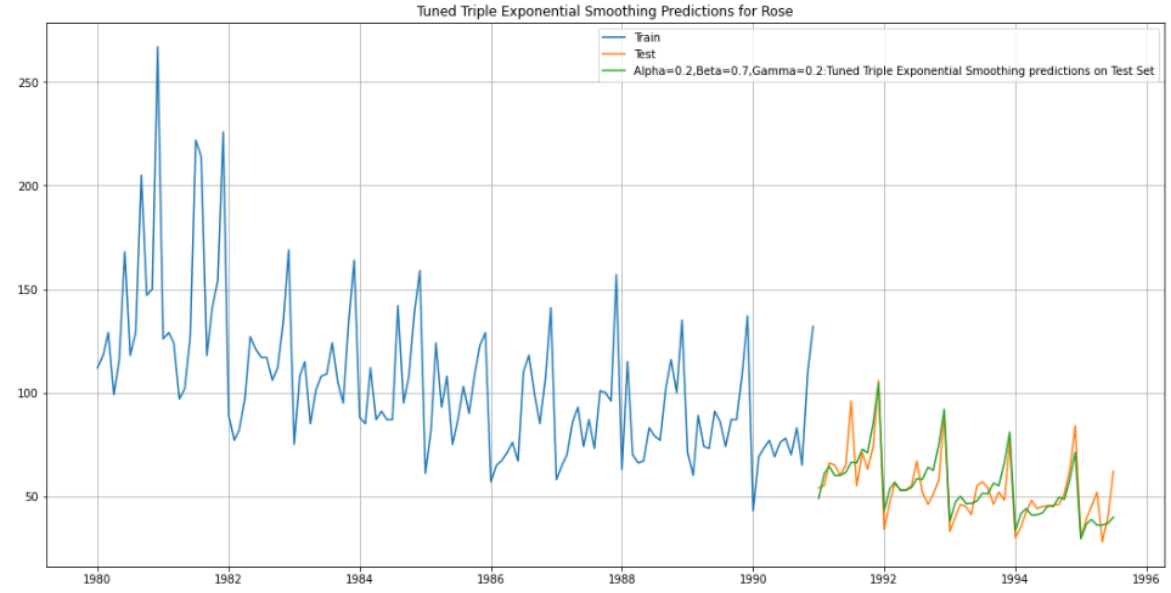


```
TES RMSE: 19.98744868511025
```

	Test RMSE-Rose
Linear Regression	15.268955
Naive Approach	79.718773
Simple Average	53.460570
2point Trailing Moving Average	11.529278
4point Trailing Moving Average	14.451403
6point Trailing Moving Average	14.566327
9point Trailing Moving Average	14.727630
Alpha= 0.098:Simple Exponential Smoothing	36.796227
Alpha=0.07:Tuned Simple Exponential Smoothing	36.435772
Alpha=1.4901-08,,Beta=1.661-10:Double Exponential Smoothing	15.268944
Alpha=0.1,,Beta=0.1:Tuned Double Exponential Smoothing	36.923416
Alpha=0.055,Beta=0.031,Gamma=0.00033:Triple Exponential Smoothing	19.987449

Tuned triple exponential predictions for rose

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE
161	0.2	0.7	0.2	24.042290	8.702460
215	0.3	0.2	0.6	26.940472	10.206702
10	0.1	0.2	0.1	19.647823	11.133402
11	0.1	0.2	0.2	20.172839	11.403697
12	0.1	0.2	0.3	20.828952	11.826158



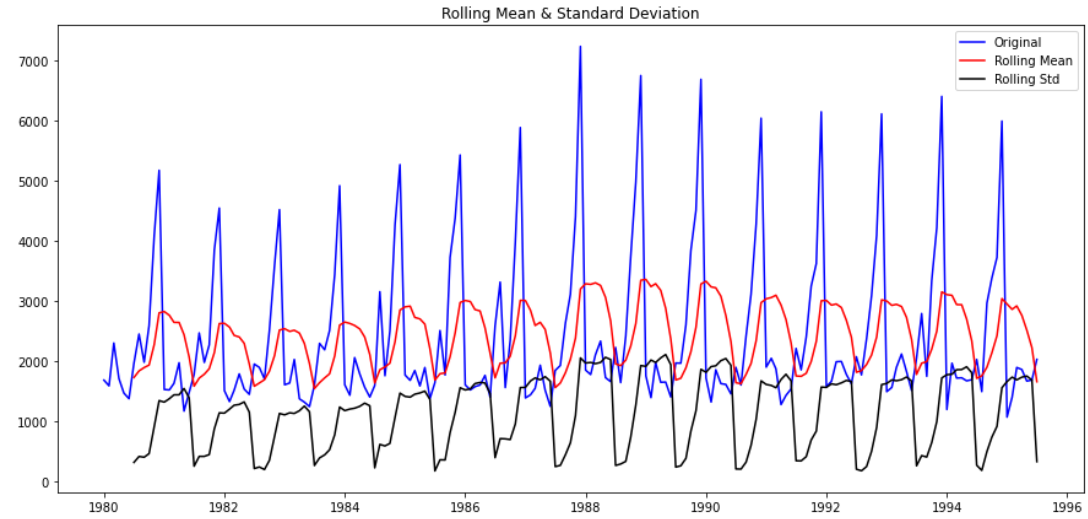
	Test RMSE-Rose
Linear Regression	15.268955
Naive Approach	79.718773
Simple Average	53.460570
2point Trailing Moving Average	11.529278
4point Trailing Moving Average	14.451403
6point Trailing Moving Average	14.566327
9point Trailing Moving Average	14.727630
Alpha= 0.098:Simple Exponential Smoothing	36.796227
Alpha=0.07:Tuned Simple Exponential Smoothing	36.435772
Alpha=1.4901-08,,Beta=1.661-10:Double Exponential Smoothing	15.268944
Alpha=0.1,,Beta=0.1:Tuned Double Exponential Smoothing	36.923416
Alpha=0.055,Beta=0.031,Gamma=0.00033:Triple Exponential Smoothing	19.987449
Alpha=0.2,Beta=0.7,Gamma=0.2:Tuned Triple Exponential Smoothing	8.702460

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.

Augmented Dickey Fuller test (ADF Test) is a common statistical test used to test whether a given Time series is stationary or not. It is one of the most commonly used statistical test when it comes to analyzing the stationary of a series. In ARIMA time series forecasting, the first step is to determine the number of differencing required to make the series stationary.

Since testing the stationarity of a time series is a frequently performed activity in autoregressive models, the ADF test along with KPSS test is something that you need to be fluent in when performing time series analysis.Another point to remember is the ADF test is fundamentally a statistical significance test. That means, there is a hypothesis testing involved with a null and alternate hypothesis and as a result a test statistic is computed and p-values get reported. It is from the test statistic and the p-value, you can make an inference as to whether a given series is stationary or not.

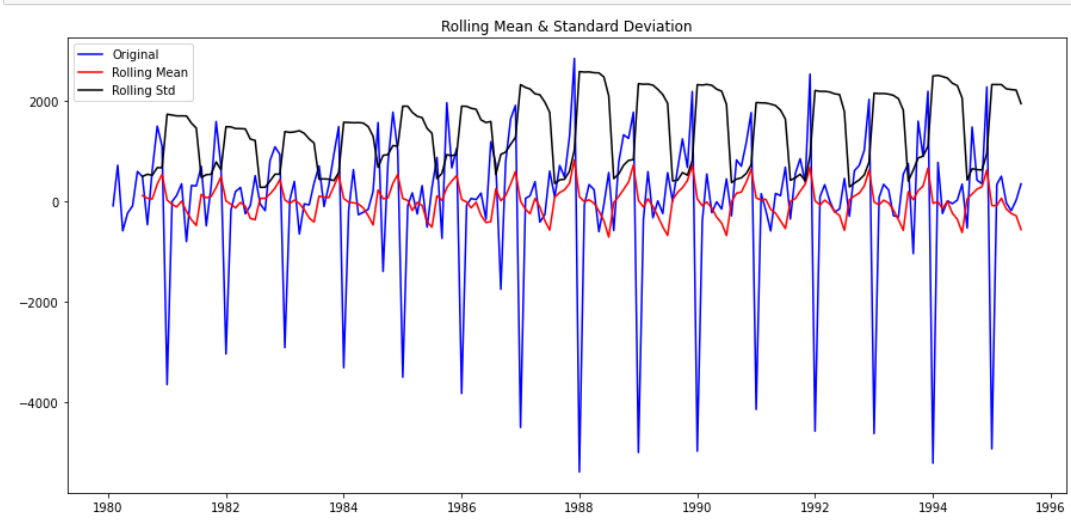
Test for stationarity of the series for sparkling data:




```
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
```

The P-value is 0.601061 is greater than the alpha. Therefore, we fail to reject null hypothesis. The series is non stationary. Let us take difference of order 1 and check whether the time series is stationary or not.

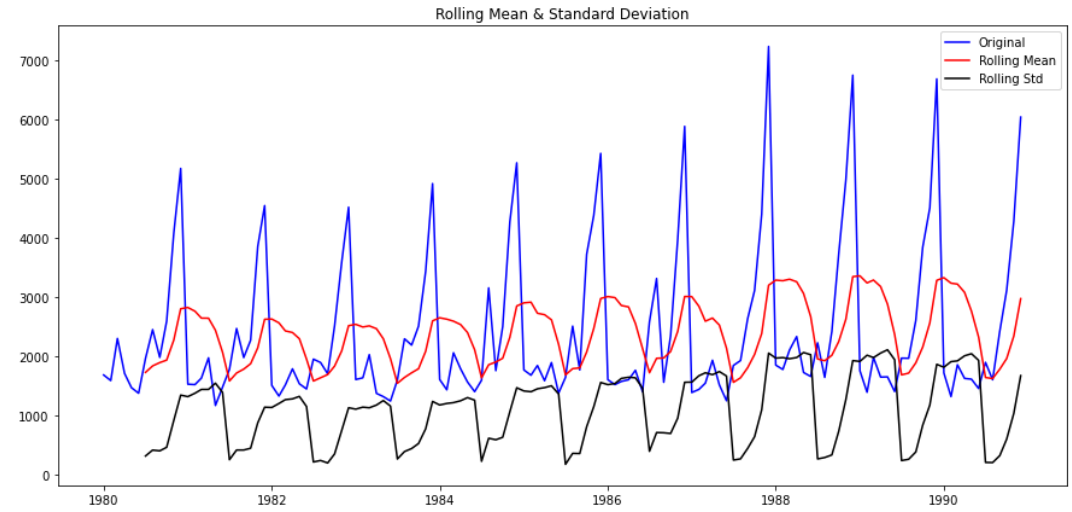
Checking the plot and stationary of time series with difference of order of one and checking with augumented dickey fuller test.




```
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
```

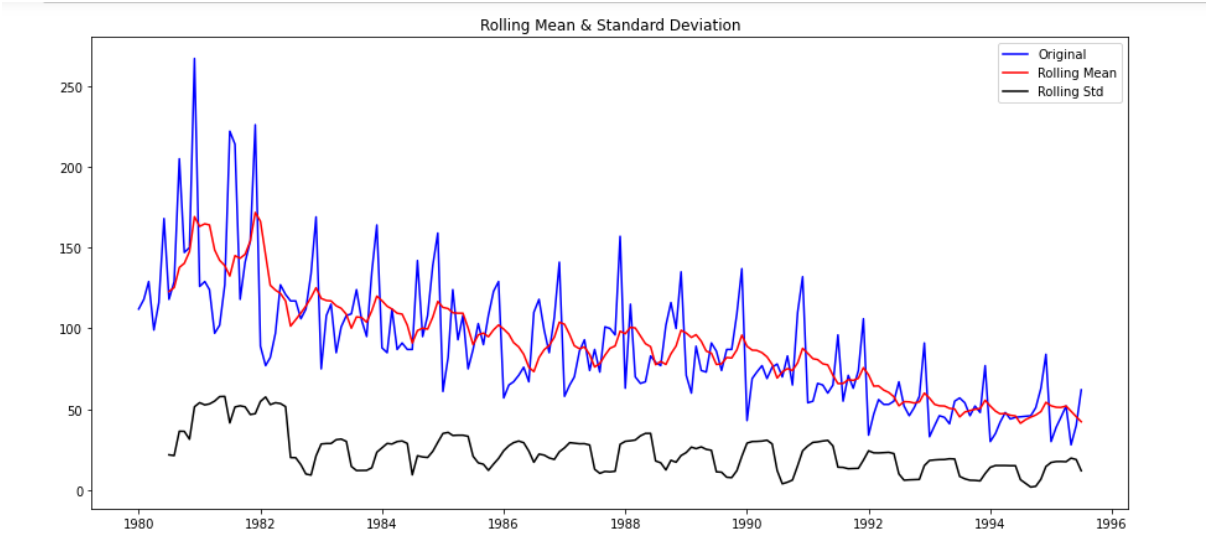
From the above graph series, the p-values is 0.000 which is less than the alpha (0.05). therefore, we can reject the null hypothesis. Therefore, the series at difference of order 1 is stationary for sparkling dataset.

Test for stationarity of the series for Rose data:




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

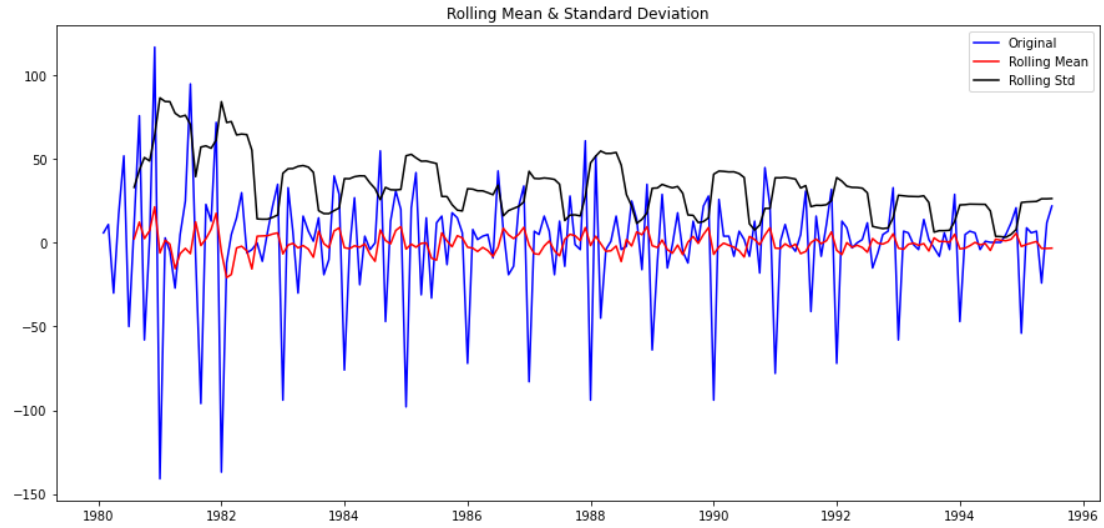
Here the p-values is 0.669744 which is greater than alpha, Therefore, we fail to reject null hypothesis. The series is non stationary. Let us take difference of order 1 and check whether the time series is stationary or not.




```
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
```

Since p – value is 0.343101 which is greater than alpha, we fail to reject the null hypothesis. Therefore, it is non stationary.

Checking the plot and stationary of time series with difference of order of one and checking with augumented dickey fuller test.



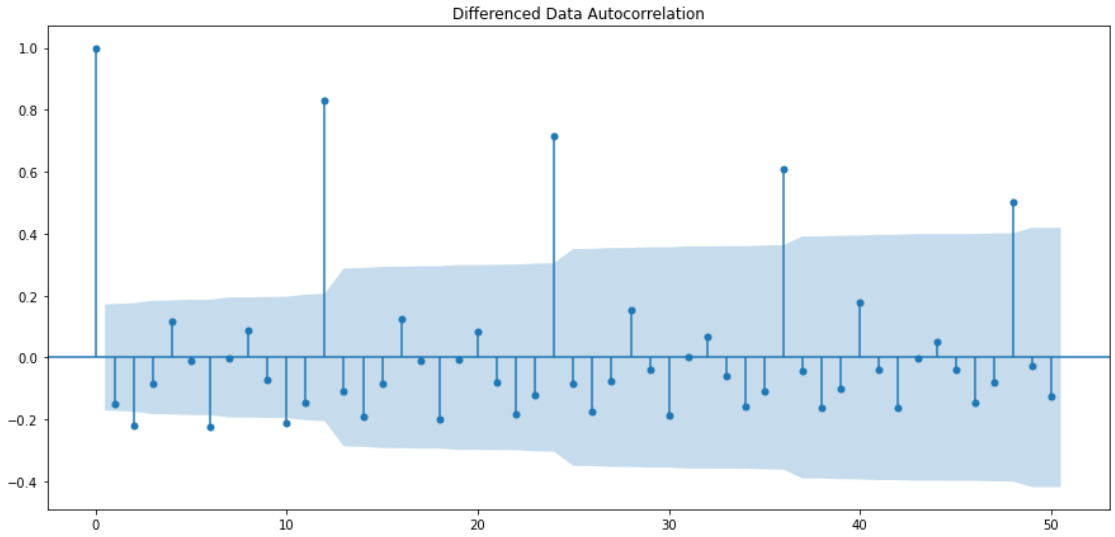

```
Results of Dickey-Fuller Test:
Test Statistic                -8.044392e+00
p-value                        1.810895e-12
#Lags Used                     1.200000e+01
Number of Observations Used    1.730000e+02
Critical Value (1%)           -3.468726e+00
Critical Value (5%)           -2.878396e+00
Critical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

For the above test, p-values is lesser than the alpha (0.05), therefore we can reject the null hypothesis. The series at difference of order 1 is stationary for rose dataset.

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.

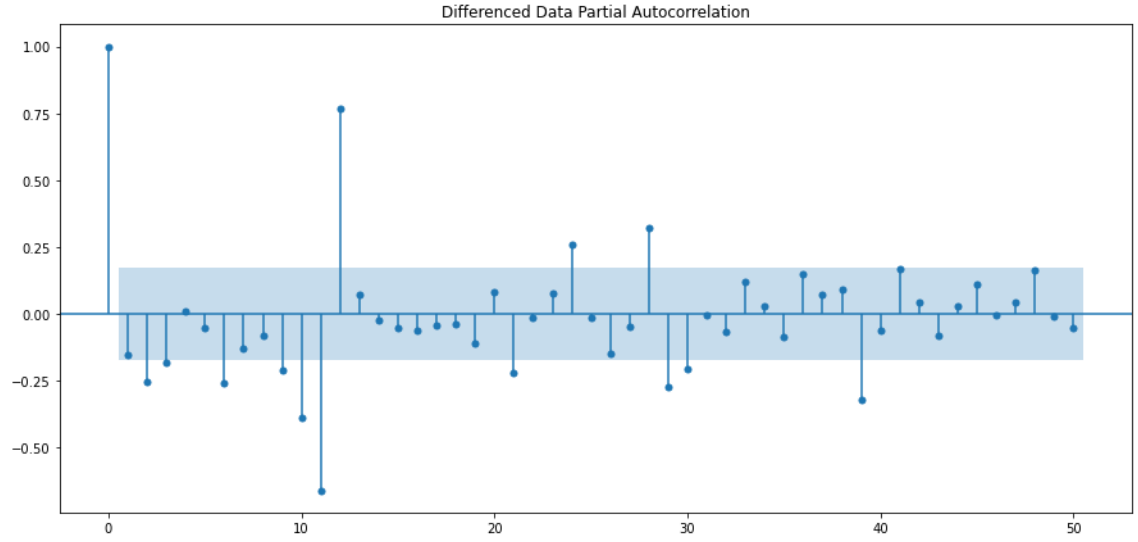
An autoregressive integrated moving average model is a form of regression analysis that gauges the strength of one dependent variable relative to other changing variables. The models goal is to predict future securities or financial market moves by examining the differences between values in the series instead of through actual values.

Plotting the autocorrelation and the partial autocorrelation function plot for sparkling



By looking at the above plots, we can say that of ACF, we can say that the moving average parameter in an ARIMA model is q which comes from the significant lag before the ACF plot cut off to 0

Next plotting PACF on the difference train stationary dataset for sparkling




```
ARIMA(0, 1, 0) - AIC:1333.1546729124348
ARIMA(0, 1, 1) - AIC:1282.3098319748312
ARIMA(0, 1, 2) - AIC:1279.6715288535818
ARIMA(1, 1, 0) - AIC:1317.3503105381492
ARIMA(1, 1, 1) - AIC:1280.5742295380032
ARIMA(1, 1, 2) - AIC:1279.870723423191
ARIMA(2, 1, 0) - AIC:1298.6110341605004
ARIMA(2, 1, 1) - AIC:1281.5078621868474
ARIMA(2, 1, 2) - AIC:1281.8707222264284
```

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

	param	AIC
2	(0, 1, 2)	1279.671529
5	(1, 1, 2)	1279.870723
4	(1, 1, 1)	1280.574230
7	(2, 1, 1)	1281.507862
8	(2, 1, 2)	1281.870722
1	(0, 1, 1)	1282.309832
6	(2, 1, 0)	1298.611034
3	(1, 1, 0)	1317.350311
0	(0, 1, 0)	1333.154673

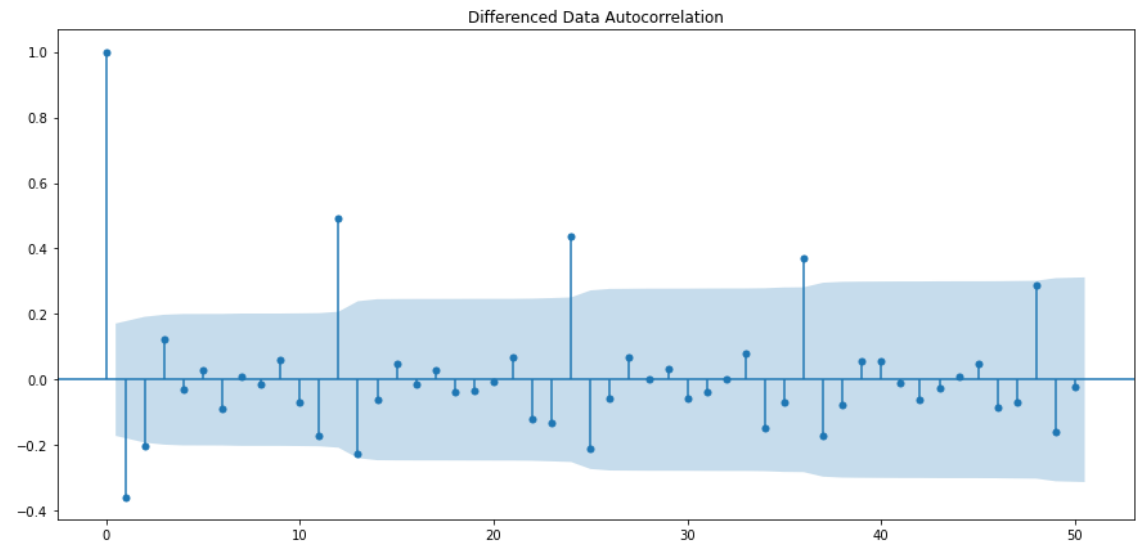
From the above table we can see that the lowest AIC value is obtained from the combination where p=2, d=1, q=1

```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                      Rose   No. Observations:                  132
Model:                 ARIMA(0, 1, 2)   Log Likelihood                -636.836
Date:                Fri, 20 May 2022   AIC                           1279.672
Time:                        11:34:54   BIC                           1288.297
Sample:                    01-01-1980   HQIC                          1283.176
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.6970      0.072     -9.689      0.000      -0.838      -0.556
ma.L2         -0.2042      0.073     -2.794      0.005      -0.347      -0.061
sigma2       965.8407     88.305     10.938      0.000     792.766    1138.915
===================================================================================
Ljung-Box (L1) (Q):                   0.14   Jarque-Bera (JB):                39.24
Prob(Q):                              0.71   Prob(JB):                         0.00
Heteroskedasticity (H):               0.36   Skew:                             0.82
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.13
===================================================================================

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

```
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)
```

Plotting the autocorrelation and the partial autocorrelation function plot for Rose



	Test RMSE-Rose
Linear Regression	15.268955
Naive Approach	79.718773
Simple Average	53.460570
2point Trailing Moving Average	11.529278
4point Trailing Moving Average	14.451403
6point Trailing Moving Average	14.566327
9point Trailing Moving Average	14.727630
Alpha= 0.098:Simple Exponential Smoothing	36.796227
Alpha=0.07:Tuned Simple Exponential Smoothing	36.435772
Alpha=1.4901-08,,Beta=1.661-10:Double Exponential Smoothing	15.268944
Alpha=0.1,,Beta=0.1:Tuned Double Exponential Smoothing	36.923416
Alpha=0.055,Beta=0.031,Gamma=0.00033:Triple Exponential Smoothing	19.987449
Alpha=0.2,Beta=0.7,Gamma=0.2:Tuned Triple Exponential Smoothing	8.702460
ARIMA(0,1,2) AIC criteria	15.619203

```
37.30647971665308
```

From the above table we can see that the lowest AIC value is obtained from the combination where p=0, d=1, q=2

Considering this order, the ARIMA model is built on the train data and the summary is checked.

SARIMAX model

SARIMAX(Seasonal Auto-Regressive Integrated Moving Average with exogenous factors) is an updated version of the ARIMA model. we can say SARIMAX is a seasonal equivalent model like SARIMA and Auto ARIMA. it can also deal with external effects. This feature of the model differs from other models

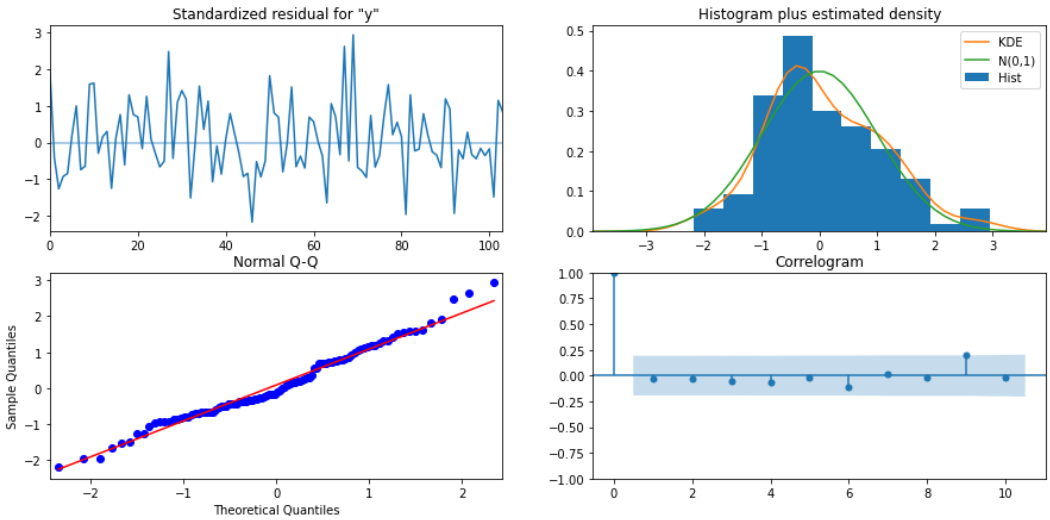
```
                                 SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 12)   Log Likelihood                -436.969
Date:                            Fri, 20 May 2022   AIC                            887.938
Time:                                    11:37:31   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    189.943     -0.004      0.996    -373.124     371.439
ma.L2         -0.1573     29.841     -0.005      0.996     -58.645      58.330
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       251.3137   4.77e+04      0.005      0.996   -9.33e+04    9.38e+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).
```

AIC is 887.938

	param	seasonal	AIC
26	(0, 1, 2)	(2, 0, 2, 12)	887.937509
53	(1, 1, 2)	(2, 0, 2, 12)	889.903048
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

Checking a diagnostic on the residual



y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	62.867265	15.928501	31.647977	94.086553
1	70.541190	16.147659	38.892360	102.190020
2	77.356411	16.147657	45.707586	109.005236
3	76.208814	16.147657	44.559989	107.857639
4	72.747398	16.147657	41.098573	104.396223

```
26.928362030237995
```

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.

An autoregressive integrated moving average, or ARIMA, is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends. A statistical model is autoregressive if it predicts future values based on past values. For example, an ARIMA model might seek to predict a stock's future prices based on its past performance or forecast a company's earnings based on past periods.

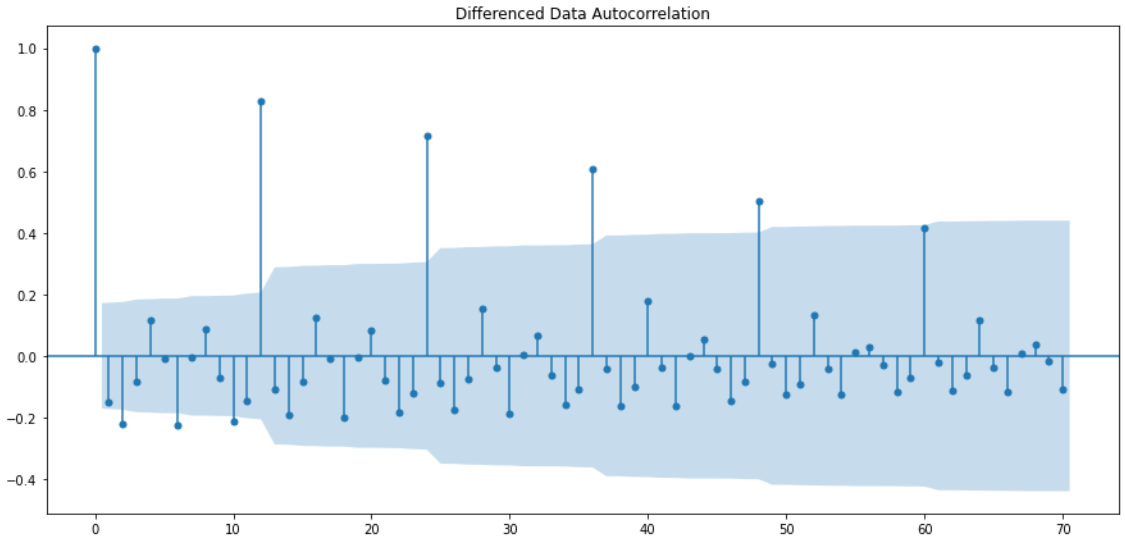
ARIMA Model for Sparkling

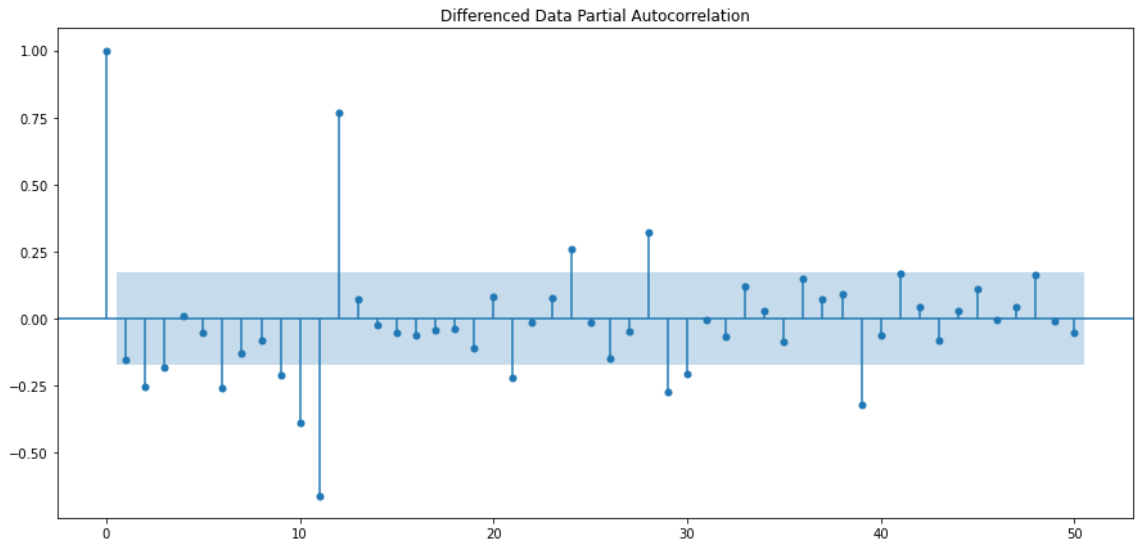
Start with building the ACF and PACF plots for the train data with difference 1. We will check for the values of p and q by looking at the plot and where the cut off point are located in the plot.

We can select the order p for AR model based on significant spikes from the PACF plot.one more indication of the AR process is that the ACF plot decays more slowly. In contrast to the AR model, we can select the order q for model from ACF if this plot has a sharp cut off after lag.

To find the value of p, we look into PACF plot. We can see that the cut off point is the first point. We will take p where the last point was not yet cut by confidence interval. Hence, p=0.

To find the value of q, we look into the autocorrelation plot, we see that first point cuts the confidence interval, therefore q=0.m so our order would be p=0, d=1 and q=0. We take this order and build an ARIMA model on the training data.





Summary of manual ARIMA model

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                Sparkling   No. Observations:                  132
Model:                 ARIMA(0, 1, 0)   Log Likelihood               -1132.832
Date:                Fri, 20 May 2022   AIC                           2267.663
Time:                        14:30:04   BIC                           2270.538
Sample:                    01-01-1980   HQIC                          2268.831
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
sigma2      1.885e+06   1.29e+05     14.658      0.000    1.63e+06    2.14e+06
===================================================================================
Ljung-Box (L1) (Q):                   3.07   Jarque-Bera (JB):               198.83
Prob(Q):                              0.08   Prob(JB):                         0.00
Heteroskedasticity (H):               2.46   Skew:                            -1.92
Prob(H) (two-sided):                  0.00   Kurtosis:                         7.65
===================================================================================

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

AIC for this model is 2267.663, the model looks good. We can check the performance with the help of RMSE score. With this ARIMA model.

The RMSE score on test data is 3864.279

	Test RMSE-Sparkling
Linear Regression	1389.135175
Naive Approach	3864.279352
Simple Average	1275.081804
2point Trailing Moving Average	813.400684
4point Trailing Moving Average	1156.589694
6point Trailing Moving Average	1283.927428
9point Trailing Moving Average	1346.278315
Alpha=0.049:Simple Exponential Smoothing	1316.035487
Alpha=0.02:Tuned Simple Exponential Smoothing	1279.495201
Alpha=0.66,Beta=0.0001:DoubleExponentialSmoothing	5291.879833
Alpha=0.1,Beta=0.1:Tuned Double Exponential Smoothing	1778.564670
Alpha=0.111,Beta=0.049,Gamma=0.362:Triple Exponential Smoothing	380.398478
Alpha=0.4,Beta=0.01,Gamma=0.3:Tuned Triple Exponential Smoothing	326.579641
ARIMA(0,1,0) Manual plot	3864.279352
SARIMA(0,1,0)(2,1,4,12) Manual plot	937.540131

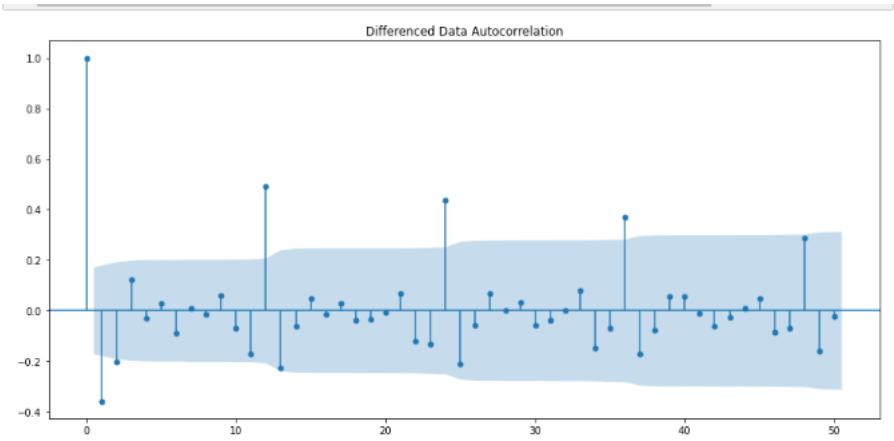
ARIMA model for rose:

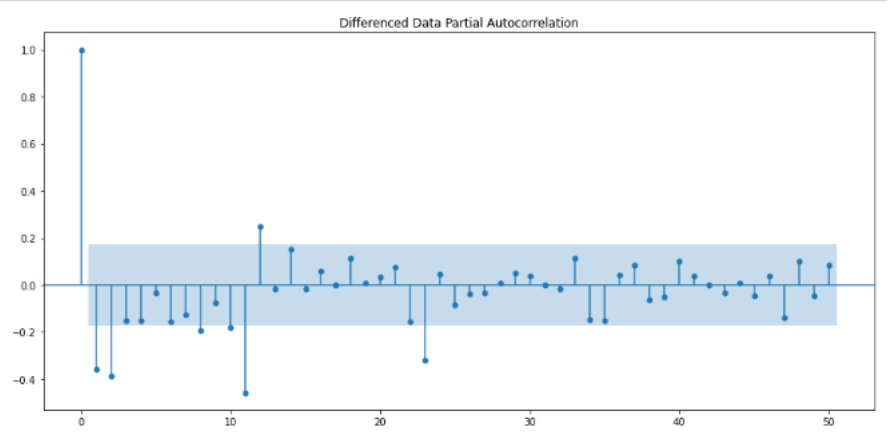
Start with building the ACF and PACF plots for the train data with difference 1. We will check for the values of p and q by looking at the plot and where the cut off point are located in the plot.

We can select the order p for AR model based on significant spikes from the PACF plot.one more indication of the AR process is that the ACF plot decays more slowly. In contrast to the AR model, we can select the order q for model from ACF if this plot has a sharp cut off after lag.

To find the value of p, we look into PACF plot. We can see that the cut off point is the first point. We will take p where the last point was not yet cut by confidence interval. Hence, p=0.

To find the value of q, we look into the autocorrelation plot, we see that first point cuts the confidence interval, therefore q=0.m so our order would be p=0, d=1 and q=0. We take this order and build an ARIMA model on the training data.





SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	ARIMA(2, 1, 2)	Log Likelihood	-635.935
Date:	Fri, 20 May 2022	AIC	1281.871
Time:	14:34:56	BIC	1296.247
Sample:	01-01-1980	HQIC	1287.712
	- 12-01-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.4540	0.469	-0.969	0.333	-1.372	0.464
ar.L2	0.0001	0.170	0.001	0.999	-0.334	0.334
ma.L1	-0.2541	0.459	-0.554	0.580	-1.154	0.646
ma.L2	-0.5984	0.430	-1.390	0.164	-1.442	0.245
sigma2	952.1601	91.424	10.415	0.000	772.973	1131.347

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	34.16
Prob(Q):	0.88	Prob(JB):	0.00
Heteroskedasticity (H):	0.37	Skew:	0.79
Prob(H) (two-sided):	0.00	Kurtosis:	4.94

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

AIC is 1281.871. the model looks good with all the coefficients and p values. We can check the performance with the help of RMSE score. With this ARIMA model.

The RMSE score on test data is 36.871

SARIMA Models:

We will now build SARIMA models based on the ACF and PACF plot of training data cut off points and then forecast on test data. Finally, we will use RMSE values to assess the model's performance.

SARIMA Models:

We have already seen from the ACF plot that the seasonality for sparkling wine train dataset after one difference is 12. So, we will take seasonality as 12.

Checking the summary for sparkling for manual SARIMA model – Rose

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(0, 1, 2)x(2, 0, 2, 12)	Log Likelihood	-436.969
Date:	Fri, 20 May 2022	AIC	887.938
Time:	11:37:31	BIC	906.448
Sample:	0	HQIC	895.437
	- 132		
Covariance Type:	opg		

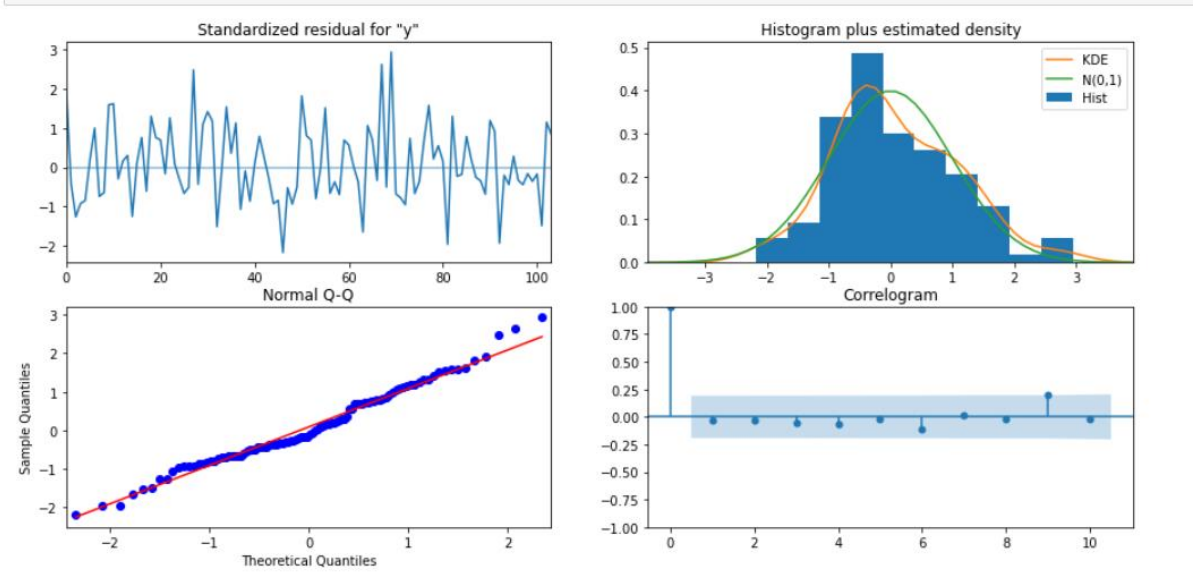
	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.8427	189.943	-0.004	0.996	-373.124	371.439
ma.L2	-0.1573	29.841	-0.005	0.996	-58.645	58.330
ar.S.L12	0.3467	0.079	4.375	0.000	0.191	0.502
ar.S.L24	0.3023	0.076	3.996	0.000	0.154	0.451
ma.S.L12	0.0767	0.133	0.577	0.564	-0.184	0.337
ma.S.L24	-0.0726	0.146	-0.498	0.618	-0.358	0.213
sigma2	251.3137	4.77e+04	0.005	0.996	-9.33e+04	9.38e+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).

AIC – 887.938.

Checking a diagnostic on the residual:

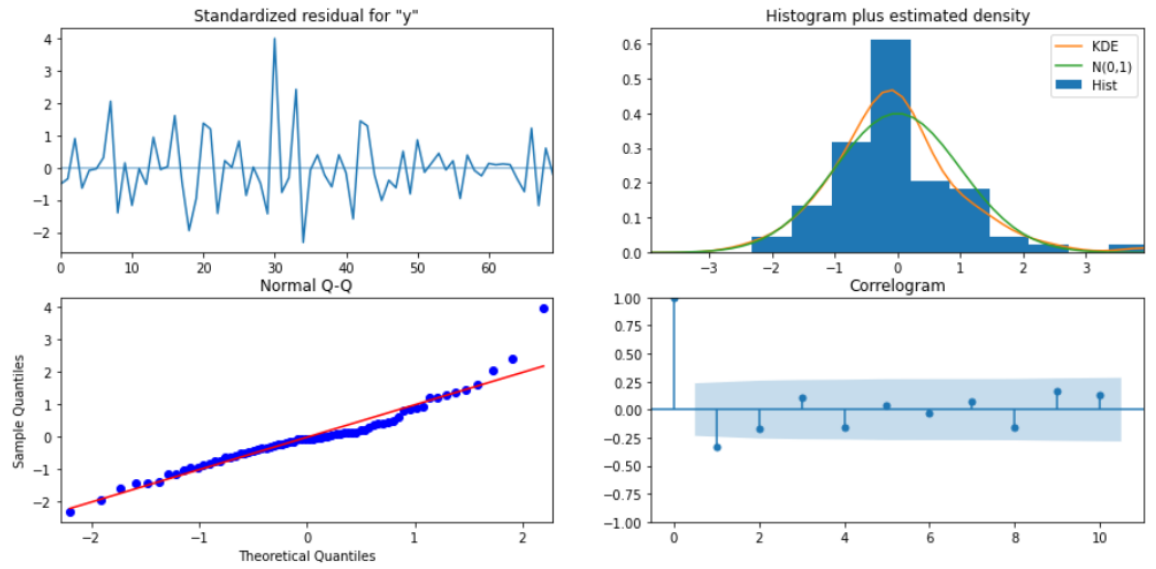


Checking the summary for sparkling for manual SARIMA model – sparkling

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                  132
Model:             SARIMAX(0, 1, 0)x(2, 1, [1, 2, 3, 4], 12)   Log Likelihood        -538.663
Date:                                Fri, 20 May 2022   AIC                        1091.326
Time:                                        14:32:27   BIC                        1107.066
Sample:                                             0   HQIC                       1097.578
                                                  - 132
Covariance Type:                                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.S.L12      -0.5734      0.253     -2.266      0.023      -1.070      -0.077
ar.S.L24      -0.5548      0.108     -5.147      0.000      -0.766      -0.344
ma.S.L12       0.3449      0.391      0.882      0.378      -0.422       1.111
ma.S.L24       0.5798      0.191      3.040      0.002       0.206       0.954
ma.S.L36      -0.5033      0.117     -4.306      0.000      -0.732      -0.274
ma.S.L48      -0.0809      0.349     -0.232      0.816      -0.764       0.602
sigma2      2.044e+05   1.02e-06      2e+11      0.000    2.04e+05    2.04e+05
===================================================================================
Ljung-Box (L1) (Q):                   7.81   Jarque-Bera (JB):                32.02
Prob(Q):                              0.01   Prob(JB):                         0.00
Heteroskedasticity (H):               0.32   Skew:                             0.95
Prob(H) (two-sided):                  0.01   Kurtosis:                         5.72
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 1.95e+27. Standard errors may be unstable.
```

Checking a diagnostic on the residual:



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.

All of the models have now been created. All their performance evaluated using RMSE values. Compare the different models based on their RMSE values to see which one is the best fit.

Sparkling Wine

	Test RMSE-Sparkling
Alpha=0.4,Beta=0.01,Gamma=0.3:Tuned Triple Exponential Smoothing	326.579641
Alpha=0.111,Beta=0.049,Gamma=0.362:Triple Exponential Smoothing	380.398478
2point Trailing Moving Average	813.400684
SARIMA(0,1,0)(2,1,4,12) Manual plot	937.540131
4point Trailing Moving Average	1156.589694
Simple Average	1275.081804
Alpha=0.02:Tuned Simple Exponential Smoothing	1279.495201
6point Trailing Moving Average	1283.927428
Alpha=0.049:Simple Exponential Smoothing	1316.035487
9point Trailing Moving Average	1346.278315
Linear Regression	1389.135175
Alpha=0.1,Beta=0.1:Tuned Double Exponential Smoothing	1778.564670
Naive Approach	3864.279352
ARIMA(0,1,0) Manual plot	3864.279352
Alpha=0.66,Beta=0.0001:DoubleExponentialSmoothing	5291.879833

The RMSE score of all the models performed above are list in ascending order, with least RMSE aa the best performing model. Tuned Triple Exponential smoothing with alpha =0.4, beta=0.1, gamma=0.3 is the best fit model which can be used to forecast the sales with least RMSE score of 326.579641

Rose wine

	Test RMSE-Rose
Alpha=0.2,Beta=0.7,Gamma=0.2:Tuned Triple Exponential Smoothing	8.702460
2point Trailing Moving Average	11.529278
4point Trailing Moving Average	14.451403
6point Trailing Moving Average	14.566327
9point Trailing Moving Average	14.727630
Alpha=1.4901-08,,Beta=1.661-10:Double Exponential Smoothing	15.268944
Linear Regression	15.268955
ARIMA(0,1,2) AIC criteria	15.619203
SARIMA(2,1,2)(2,1,4,12) Manual plot	16.931818
Alpha=0.055,Beta=0.031,Gamma=0.00033:Triple Exponential Smoothing	19.987449
SARIMA(0,1,2)(2,0,2,12) AIC criteria	26.928362
Alpha=0.07:Tuned Simple Exponential Smoothing	36.435772
Alpha= 0.098:Simple Exponential Smoothing	36.796227
ARIMA(2,1,2) Manual plot	36.871197
Alpha=0.1,,Beta=0.1:Tuned Double Exponential Smoothing	36.923416
Simple Average	53.460570
Naive Approach	79.718773

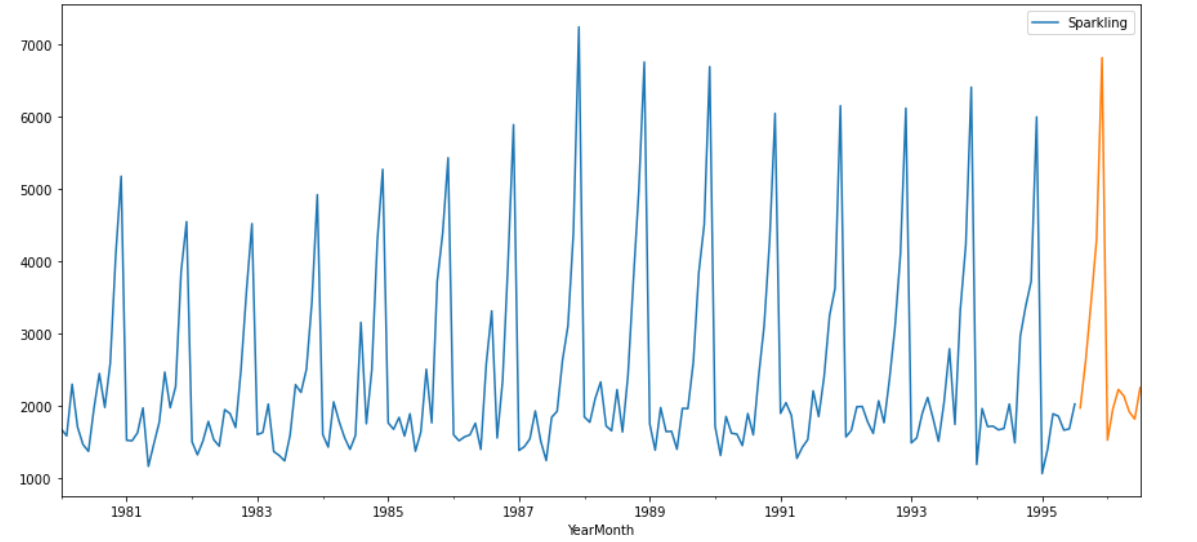
The RMSE score of all the models performed above are list in ascending order, with least RMSE aa the best performing model. Tuned Triple Exponential smoothing with alpha =0.2, beta=0.7, gamma=0.2 is the best fit model which can be used to forecast the sales with least RMSE score of 8.702

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.

We have built all the models and found out the best fit from all the models. We can use the best fit models to forecast values for the following 12 months with appropriate confidence levels.

Sparkling wine

```
RMSE: 391.21270100618614
```

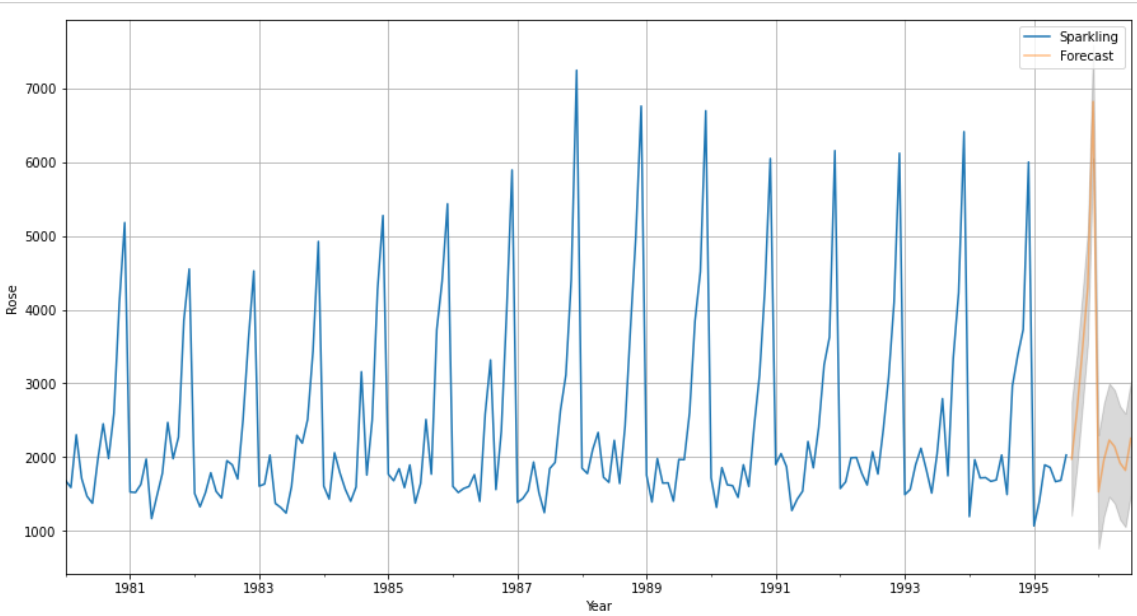


We will build with parameters Alpha = 0.4, Beta=0.1, and Gamma=0.3 on the sparkling dataset. The RMSE value is 391.212 and getting the predictions for 12 months into the future.

The predictions and lower & upper confidence intervals for 12 months in future is as below:

	lower_CI	prediction	upper_ci
1995-08-01	1210.026565	1977.230483	2744.434402
1995-09-01	1863.226093	2630.430011	3397.633930
1995-10-01	2676.175225	3443.379143	4210.583062
1995-11-01	3533.489521	4300.693440	5067.897358
1995-12-01	6051.103478	6818.307397	7585.511315

Plotting the graph for sparkling data and forecast along with the confidence band

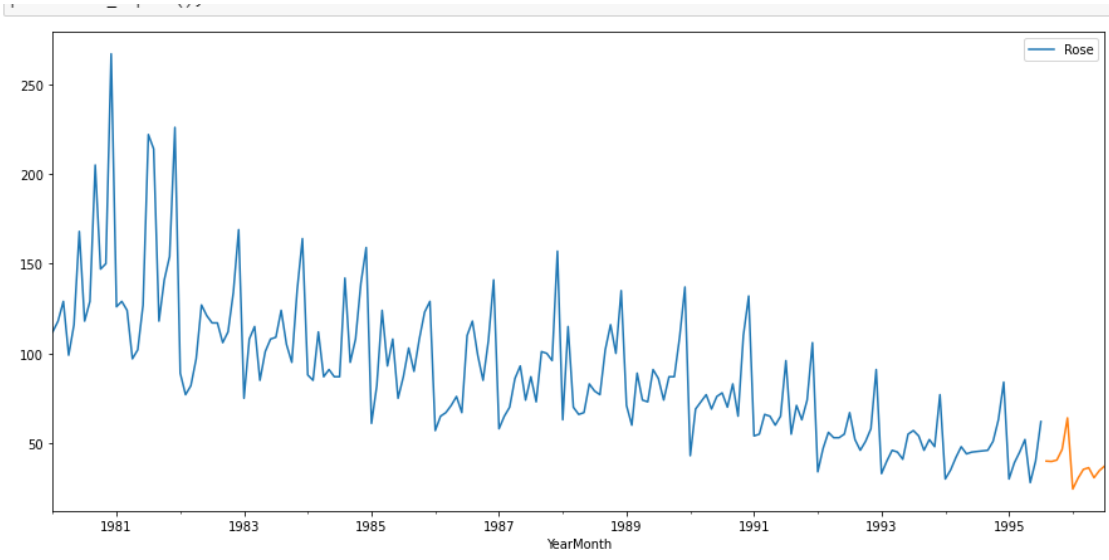


The predictions are consistent with the original dataset, as seen in the graph above. The seasonality and trend are maintained

```
RMSE: 20.681380709733475
```

Rose wine:

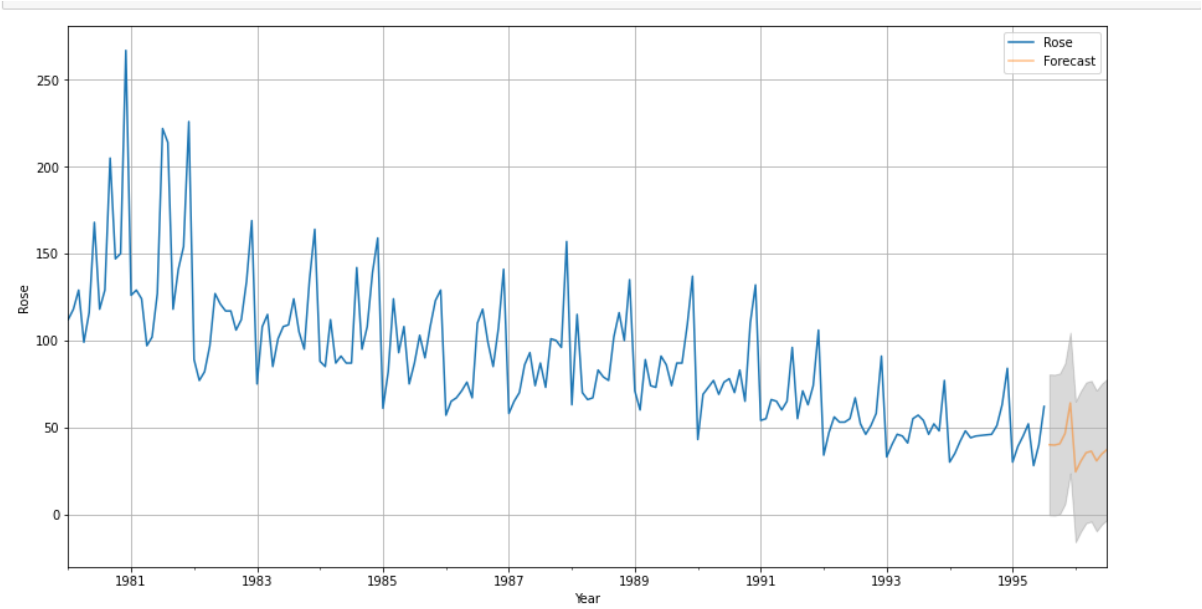
We will build with parameters Alpha = 0.2, Beta=0.7, and Gamma=0.2 on the rose dataset. The RMSE value is 20.6813 and getting the predictions for 12 months into the future.



The predictions for 12 months in future is as below:

	lower_CI	prediction	upper_ci
1995-08-01	-0.538512	39.978825	80.496163
1995-09-01	-0.744895	39.772443	80.289780
1995-10-01	0.034644	40.551981	81.069318
1995-11-01	5.969813	46.487150	87.004487
1995-12-01	23.535288	64.052625	104.569962

Plotting the graph for rose data and forecast along with the confidence band



The predictions are consistent with the original dataset, as seen in the graph above. The seasonality and trend are maintained

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

Inference and recommendations for Rose model

- Rose sales shows a decrease in trend compared to the previous years
- December month shows the highest sale across the year while the value has come down through the years from 1980-1994
- The models are built considering the trends and seasonality in to account and we see the future prediction is in line with the trend and seasonality in the previous years
- The company should use the prediction results and capitalize on high demand seasons and ensure to source and supply the high demand and also plan the low demand seasons to stock as per the demand.
- In the series of promotions, we could plan for promotion around a big event or like sponsorship. This helps us to drive business and generate the awareness of the brand to the audience at a bigger level.
- They show the increased sales, so they must plan according to the growth of sales. Pre-planned with their plans.
- The PR packages to influencers will also help to promote to the audience and the company could get customer reviews for indirect marketing, their reviews also helps us to improve the standard of the brand.
- As far as marketing, deals and discounts plays major role.

Inference and recommendations for Sparkling model

Sparkling sales shows the stabilized values and not much trend compared to previous years

December month shows the highest sales across the years

The models are built considering the trend and seasonality in to account and we see from the output that future prediction is in line with the trend and seasonality in the previous years

The sales of sparkling wine is seasonal, hence the company cannot have the same stock through the year. The predictions would help here to plan the stock need basis the forecasted sales.

The company should use the prediction results and capitalize on the high demand seasons and ensure to source, supply the high demand

The company should use the prediction result to plan accordingly to improve their business strategy.