**TIME SERIES FORECASTING**

**Gaurav Sood**

**PGP-DSBA Online July 2021**

**Date: 20th February, 2022**

**Table of Contents**

**Problem 1: Time Series Forecasting (Sparkling Dataset) ……………………………………….7**

* 1. Read the data as an appropriate Time Series data and plot the data..................................……7
  2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition…………………………………………………………………………....…….………10
  3. Split the data into training and test. The test data should start in 1991….………………...….…15
  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…………………….…………………..……17
  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……………………………...………………..23
  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…………………………………………………………………..……….24
  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………………………………………………27
  8. Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data…………………………………………………...……...…33
  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…………...…34
  10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales…………………………...……………………………….36

**Problem 2: Time Series Forecasting (Rose Dataset) …………………………………………….**

* 1. Read the data as an appropriate Time Series data and plot the data..................................….38
  2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition…………………………………………………………………………....…….………41
  3. Split the data into training and test. The test data should start in 1991….………………...…….47
  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…………………….…………………..……48
  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……………………………...………………..55
  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…………………………………………………………………..……….57
  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………………………………………………60
  8. Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data…………………………………………………...……...…67
  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…………...…68
  10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales…………………………...……………………………….69

**List of Tables**

Table no.1: Dataset Sample………………………………………………………………………………………………………………

Table no.2: Dataset Sample…………………..…………………………………………………………………………………………..

Table no.3: EDA…………………………….…………………………………………………………………………………………………..

Table no.4: Null Value Check…………………………………………………………………………………………………………….

Table no.5: Data Description………..………………………………………………………………………………………………...

Table no.6:

Table no.7:

Table no.8:

Table no.9:

Table no.10:

Table no. 11

Table no. 12

Table no. 13

Table no. 14

Table no. 15

Table no. 16

**List of Figures**

Figure no. 1: ……………………………………………………………………………………………………………………………………

Figure no. 2: ……………………………………………………………………………………………………………………………………

Figure no. 3: ……………………………………………………………………………………………………………………………………

Figure no. 4: ……………………………………………………………………………………………………………………………………

Figure no. 5: ……………………………………………………………………………………………………………………………………

Figure no. 6: ……………………………………………………………………………………………………………………………………

Figure no. 7: ……………………………………………………………………………………………………………………………………

Figure no. 8: ……………………………………………………………………………………………………………………………………

Figure no. 9: ……………………………………………………………………………………………………………………………………

Figure no. 10: ………………………………………………………………………………………………………………………………….

Figure no. 11: ………………………………………………………………………………………………………………………………….

Figure no. 12: ………………………………………………………………………………………………………………………………….

Figure no. 13: ………………………………………………………………………………………………………………………………….

Figure no. 14: ………………………………………………………………………………………………………………………………….

Figure no. 15: ………………………………………………………………………………………………………………………………….

Figure no. 16: ………………………………………………………………………………………………………………………………….

Figure no. 17: ………………………………………………………………………………………………………………………………….

Figure no. 18: ………………………………………………………………………………………………………………………………….

Figure no. 19: ………………………………………………………………………………………………………………………………….

Figure no. 20: ………………………………………………………………………………………………………………………………….

Figure no. 21: ………………………………………………………………………………………………………………………………….

Figure no. 22: ………………………………………………………………………………………………………………………………….

Figure no. 23: ………………………………………………………………………………………………………………………………….

Figure no. 24: ………………………………………………………………………………………………………………………………….

Figure no. 25: ………………………………………………………………………………………………………………………………….

Figure no. 26: ………………………………………………………………………………………………………………………………….

**Problem 1: Time Series Forecasting (Sparkling Dataset)**

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.**

First, we load all the necessary libraries for model building.

Then, we read the head and tail of the dataset to check whether the data has been properly fed.

**Sample of the Dataset**

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

**Table no. 1: Dataset Sample**

| **YearMonth** | **Sparkling** |
| --- | --- |
| **182** | 1995-03 | 1897 |
| **183** | 1995-04 | 1862 |
| **184** | 1995-05 | 1670 |
| **185** | 1995-06 | 1688 |
| **186** | 1995-07 | 2031 |

**Table no. 2: Dataset Sample**

**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 Sparkling 187 non-null int64

dtypes: int64(1), object(1)

memory usage: 3.0+ KB

**Table no. 3: Data Info**

**Checking for Missing Values**

False

**Table no. 4: Missing Value Check**

There are no missing values in the given dataset.

**Creating Time Stamp**

DatetimeIndex(['1980-01-31', '1980-02-29', '1980-03-31', '1980-04-30',

'1980-05-31', '1980-06-30', '1980-07-31', '1980-08-31',

'1980-09-30', '1980-10-31',

...

'1994-10-31', '1994-11-30', '1994-12-31', '1995-01-31',

'1995-02-28', '1995-03-31', '1995-04-30', '1995-05-31',

'1995-06-30', '1995-07-31'],

dtype='datetime64[ns]', length=187, freq='M')

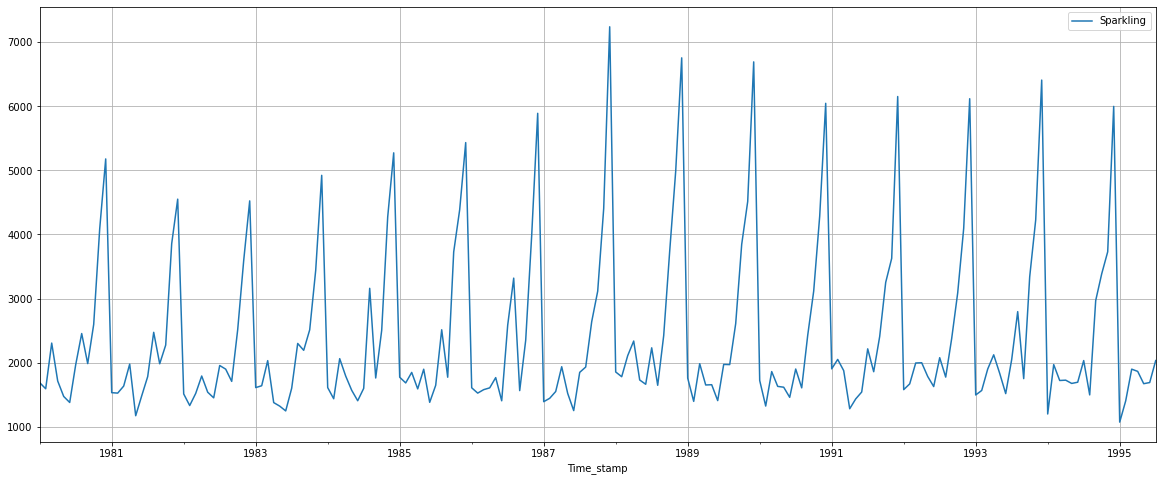
**Adding Time Stamp to Dataframe**

| **YearMonth** | **Sparkling** | **Time\_stamp** |
| --- | --- | --- |
| **0** | 1980-01 | 1686 | 1980-01-31 |
| **1** | 1980-02 | 1591 | 1980-02-29 |
| **2** | 1980-03 | 2304 | 1980-03-31 |
| **3** | 1980-04 | 1712 | 1980-04-30 |
| **4** | 1980-05 | 1471 | 1980-05-31 |

**Time Stamp as Index**

|  | **Sparkling** |
| --- | --- |
| **Time\_stamp** |  |
| **1980-01-31** | 1686 |
| **1980-02-29** | 1591 |
| **1980-03-31** | 2304 |
| **1980-04-30** | 1712 |
| **1980-05-31** | 1471 |

**Plotting the Dataset**

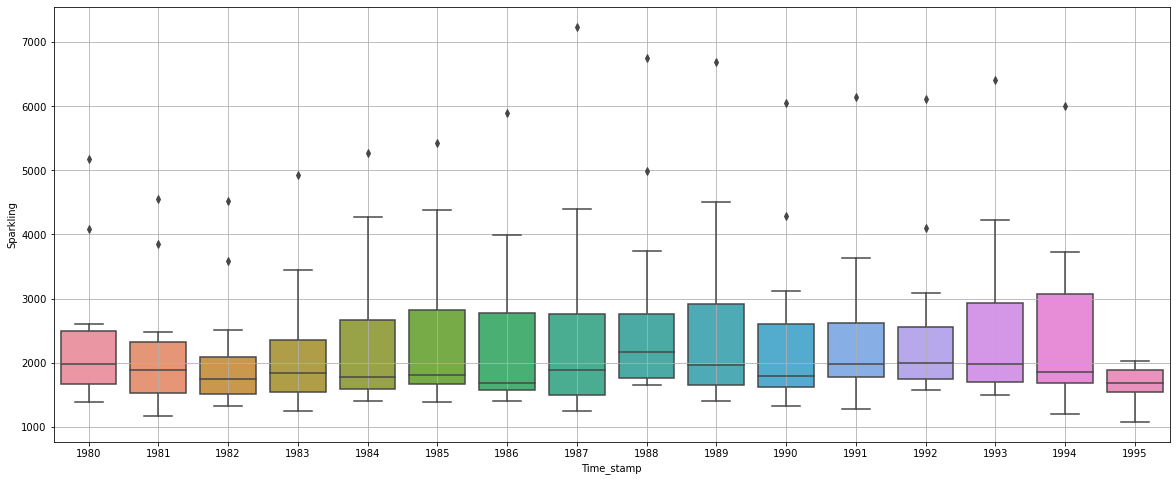
****

**Dataset Summary**

|  | **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 |

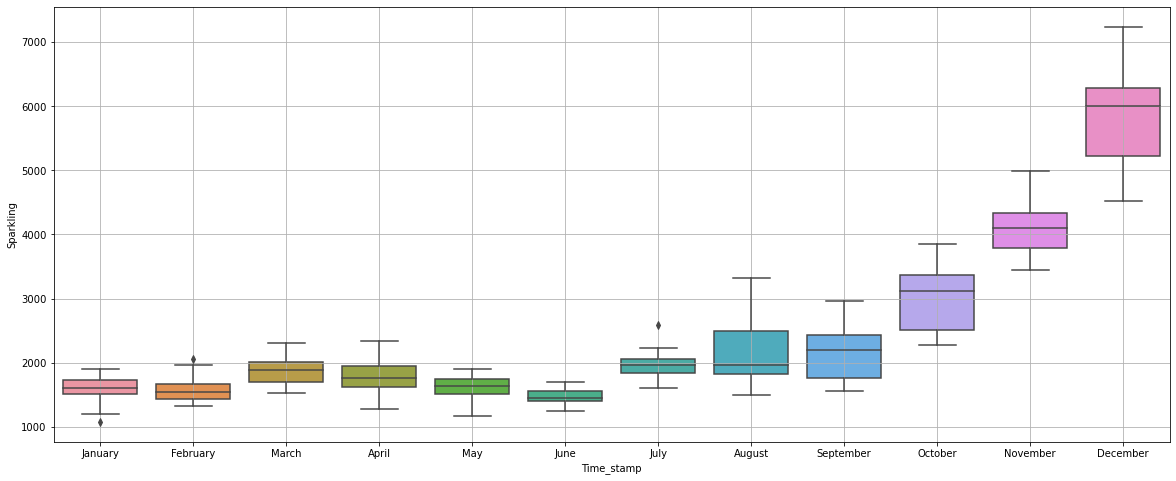
* 1. **Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.**

Sales over the years:



Median value across the years are the same.

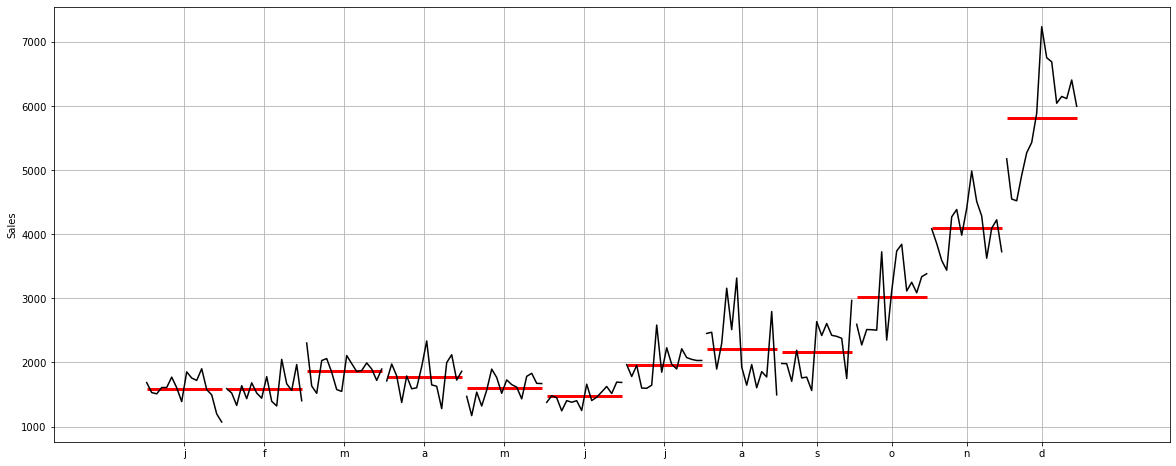
**Monthly Boxplot**



Sales increase in the second half of the year from July till December, then drop off in January.

The month of June records the lowest sales of the year.

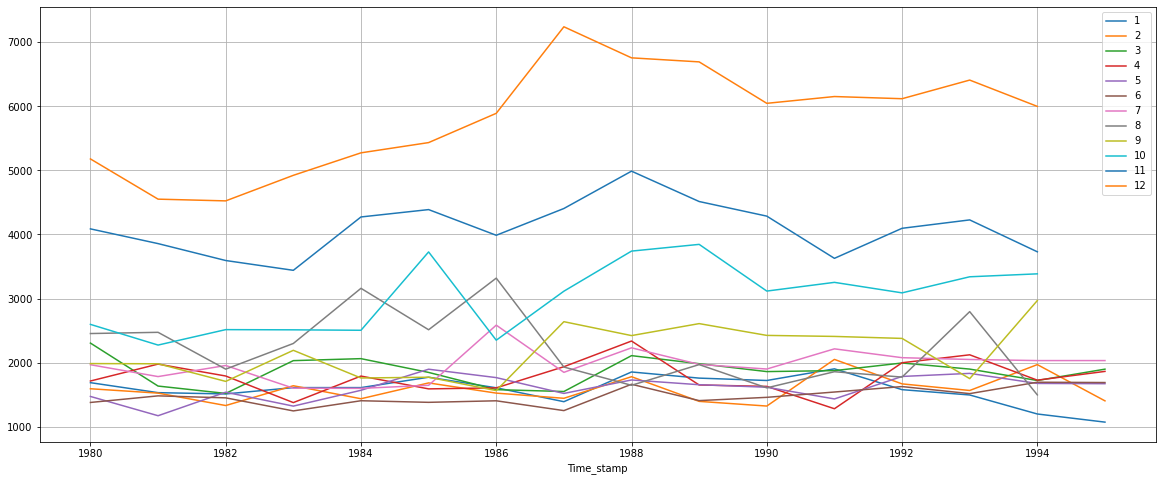
Sales across different years and within different months over the years:



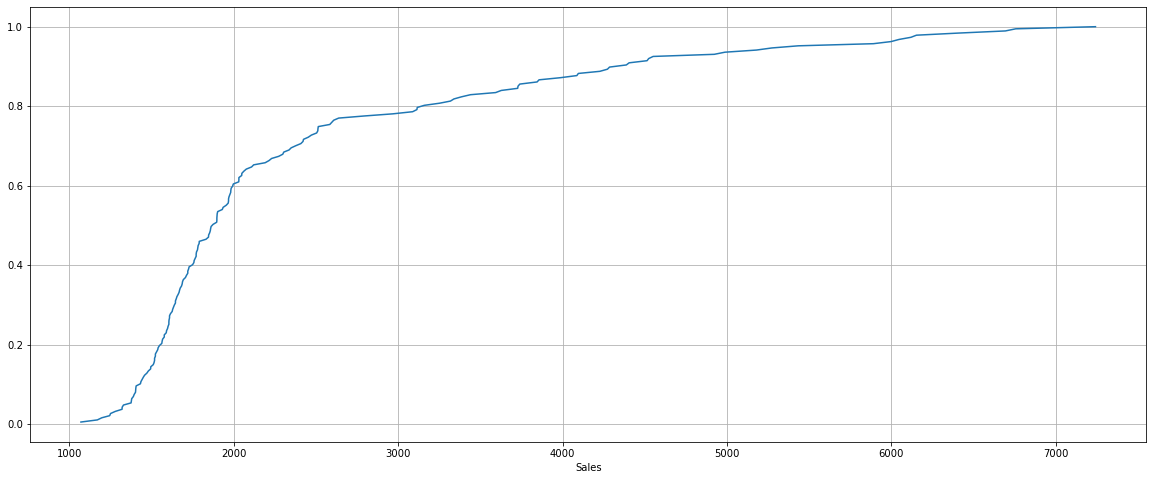
This plot shows us the behavior of time series across months.

**Plotting Graph of Monthly Sales:**

| **Time\_stamp** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Time\_stamp** |  |  |  |  |  |  |  |  |  |  |  |  |
| **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 |

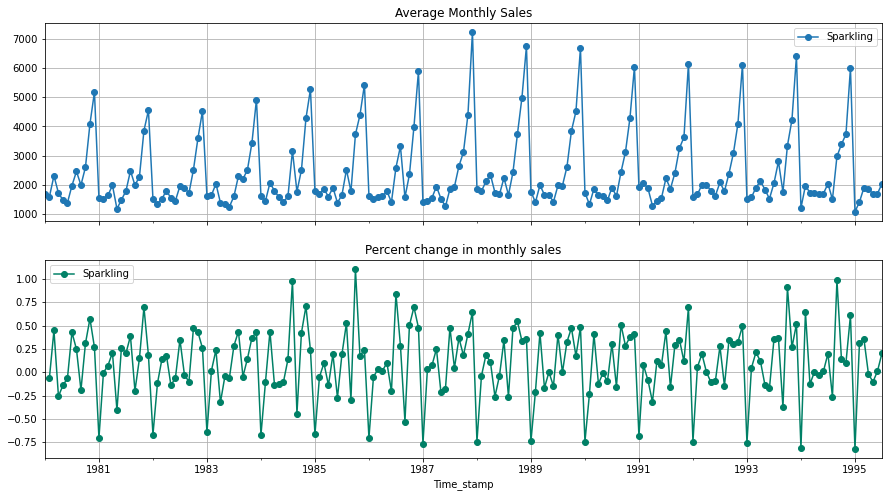


**Empirical Cumulative Distribution:**

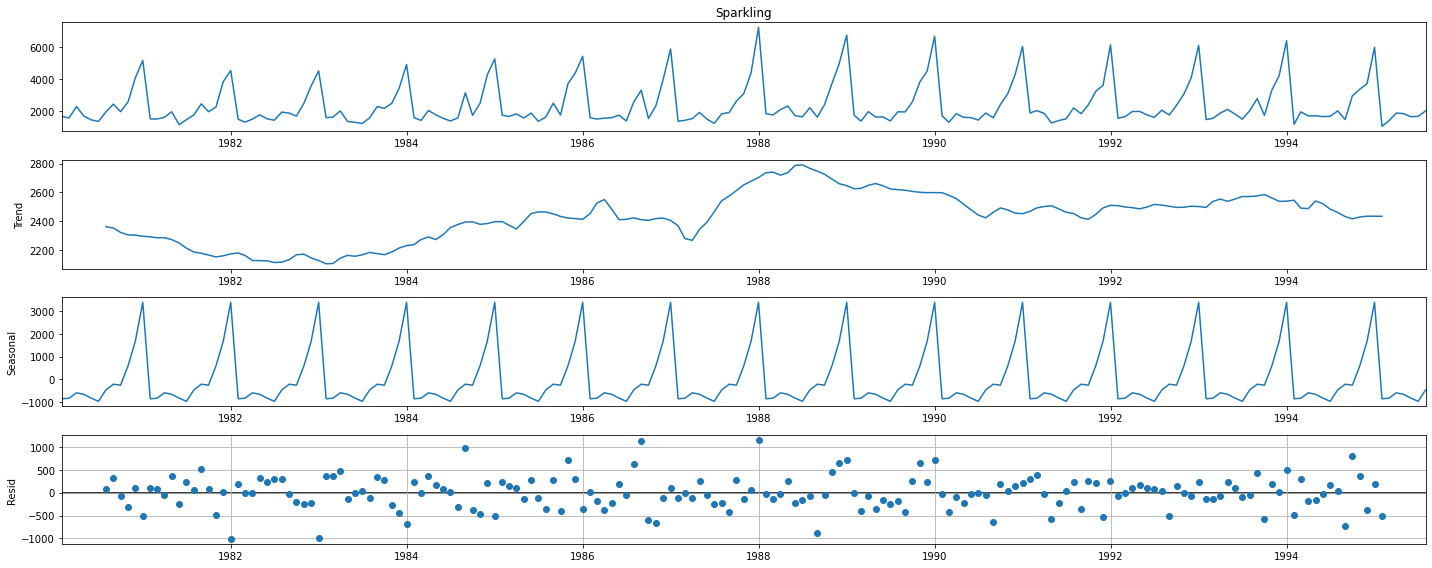


This graph tells us what percentage datapoint refers to what number of sales.

Only 20% of sales are between 3000 to 6000.



**Additive Decomposition:**



From the above plot, the trend is quite unclear as it is fluctuating (increasing, decreasing, then again increasing). Meanwhile, a strong seasonality is observed. Residual has some patterns. Multiplicative model needs to be analysed.

**Multiplicative Decomposition:**



For this model, the trend and seasonality are the same. Residual has some patterns.

So, Additive Model is considered for further analysis.

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 2360.666667

1980-08-31 2351.333333

1980-09-30 2320.541667

1980-10-31 2303.583333

Name: trend, dtype: float64

seasonality

Time\_stamp

1980-01-31 -854.260599

1980-02-29 -830.350678

1980-03-31 -592.356630

1980-04-30 -658.490559

1980-05-31 -824.416154

1980-06-30 -967.434011

1980-07-31 -465.502265

1980-08-31 -214.332821

1980-09-30 -254.677265

1980-10-31 599.769957

Name: seasonal, dtype: float64

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 70.835599

1980-08-31 315.999487

1980-09-30 -81.864401

1980-10-31 -307.353290

Name: resid, dtype: float64

. . .

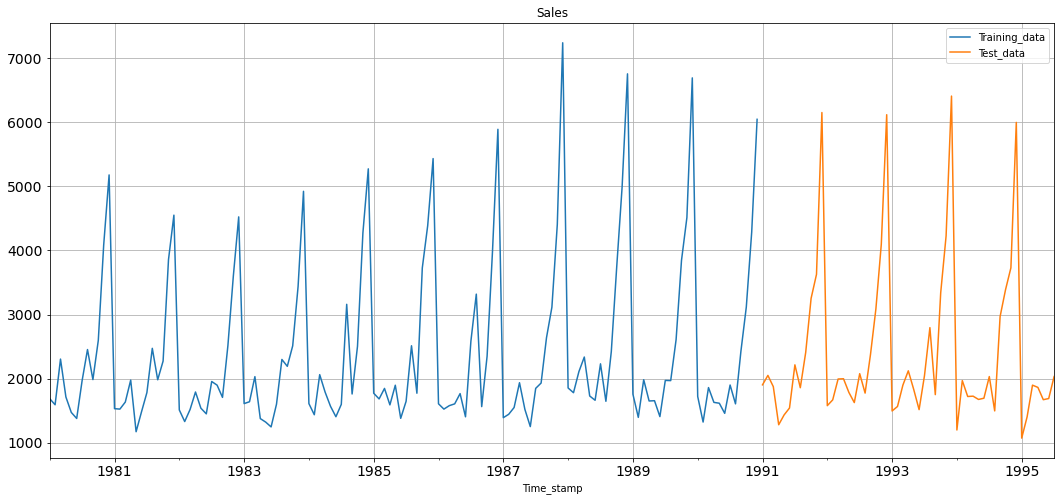
* 1. **Split the data into training and test. The test data should start in 1991.**

Dataset was split between Training and Test dataset. The test data starts from 1991.

**Shape of Training and Test Dataset**

(132, 1)

(55, 1)



* 1. **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.**

**Model 1: Linear Regression**

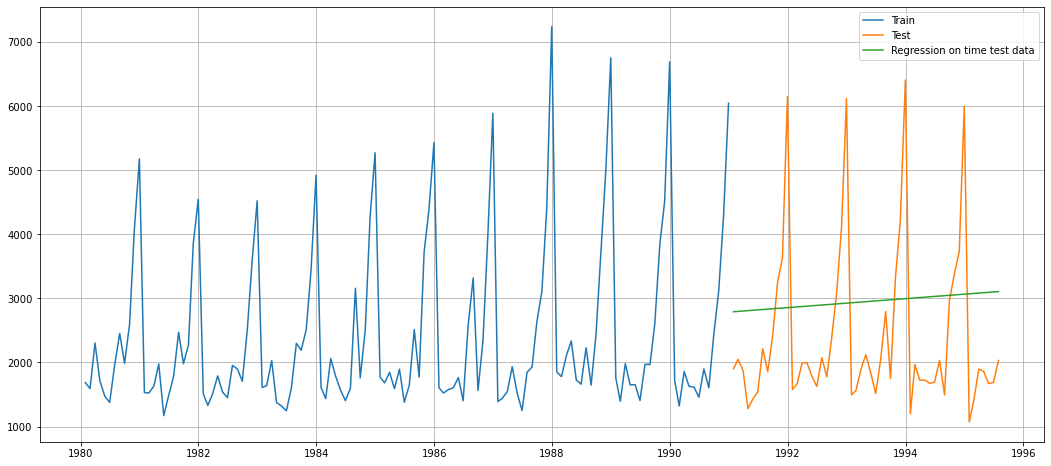
For this, we are going to regress the ‘Sales’ variable against the order of occurrence. For this we need to modify our training data before fitting it into a Linear Regression.

Training Time instance

[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, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132]

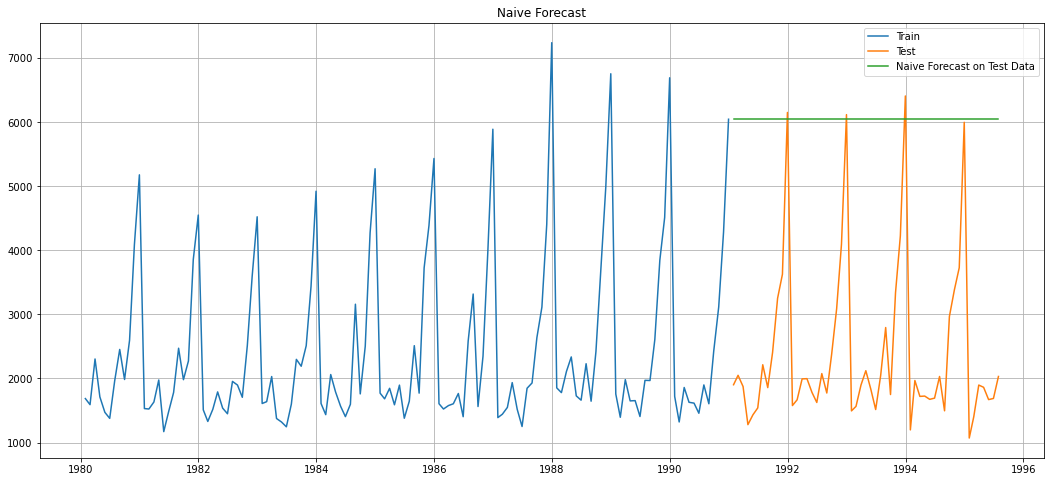
Test Time instance

[133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187]



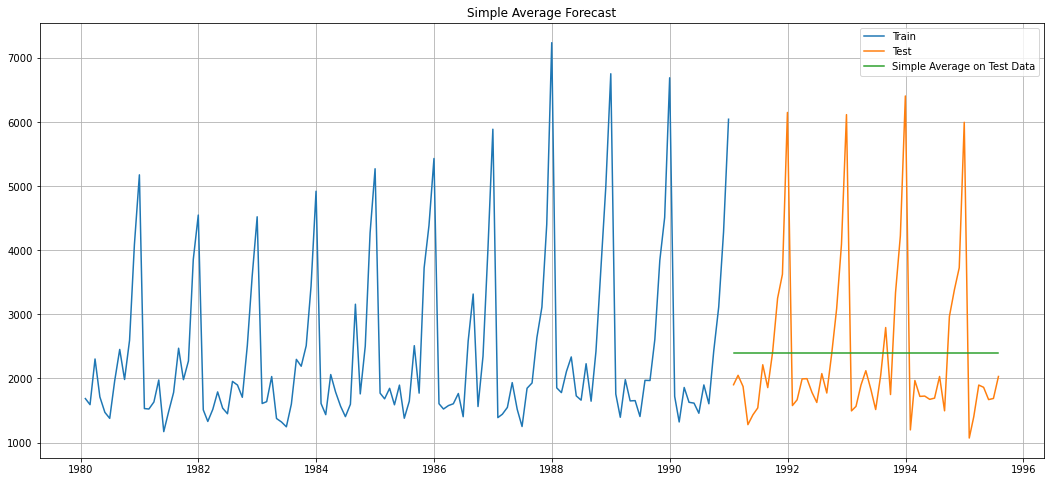
For RegressionOnTime forecast on the Test Data, RMSE is 1389.135

**Model 2: Naïve Approach**



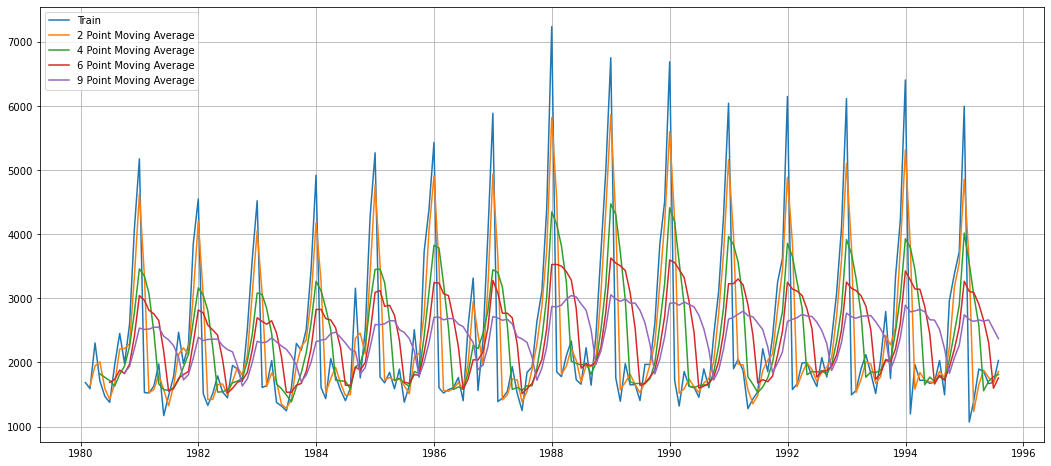
For Naive forecast on the Test Data, RMSE is 3864.279

**Model 3: Simple Average**

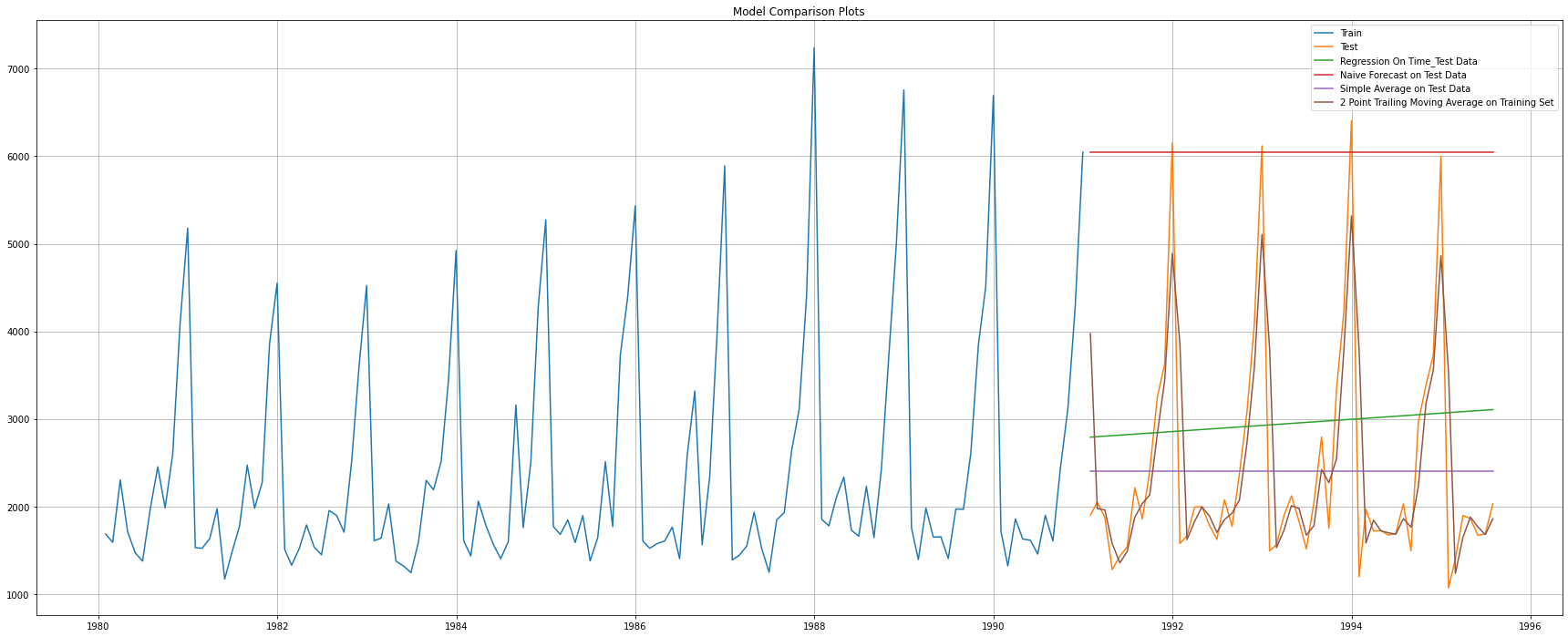


For Simple Average forecast on the Test Data, RMSE is 1275.082

**Model 4: Moving Average**







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

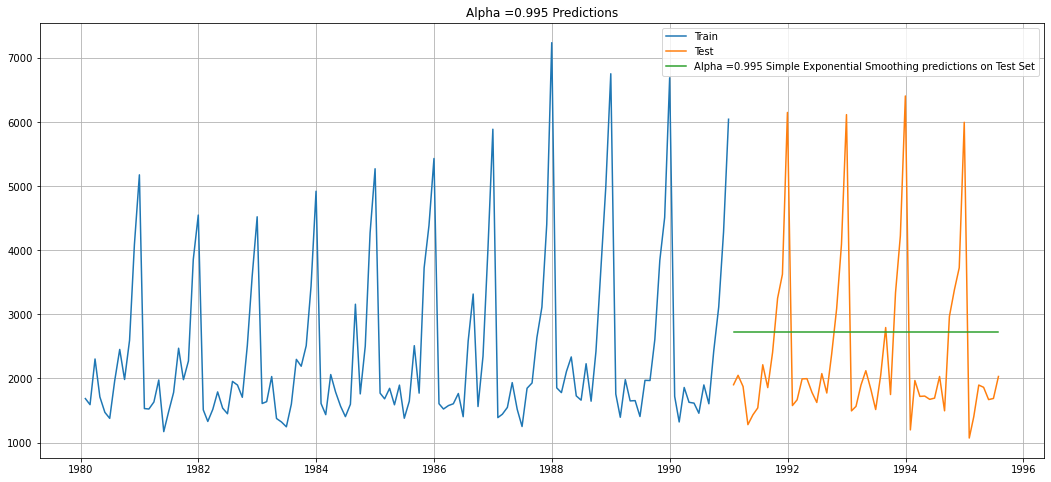
For 4 point Moving Average Model forecast on the Training Data, RMSE is 1156.590

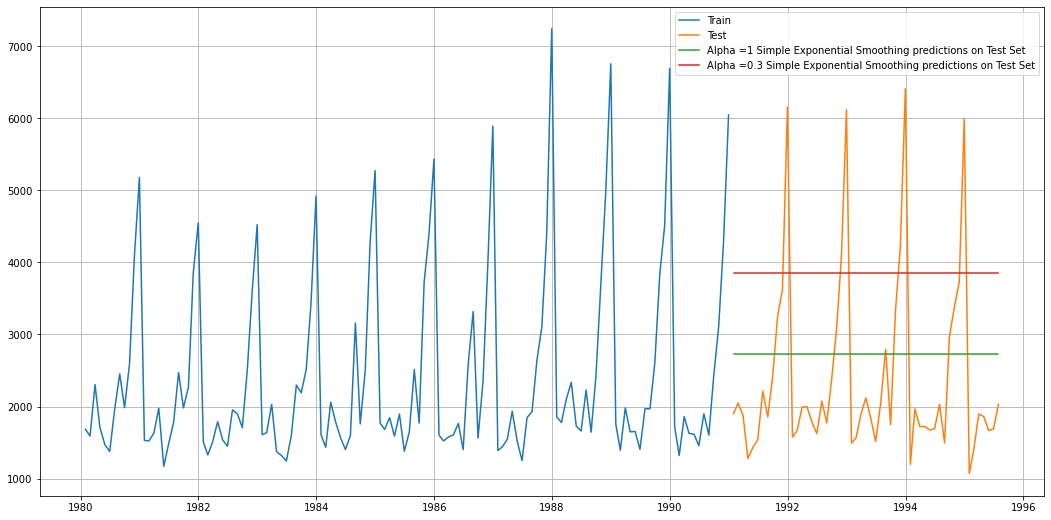
For 6 point Moving Average Model forecast on the Training Data, RMSE is 1283.927

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

. . .

**Model 5: Simple Exponential Smoothing**

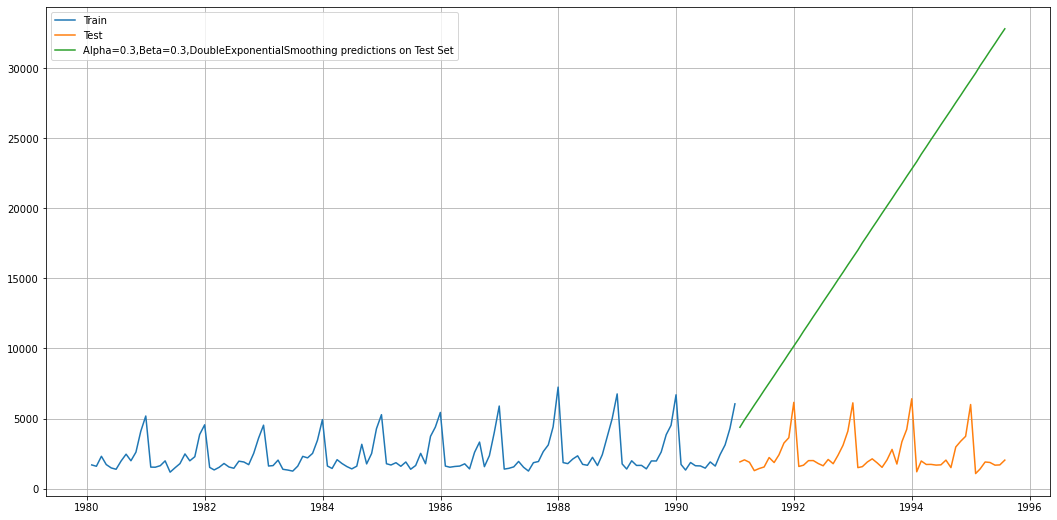




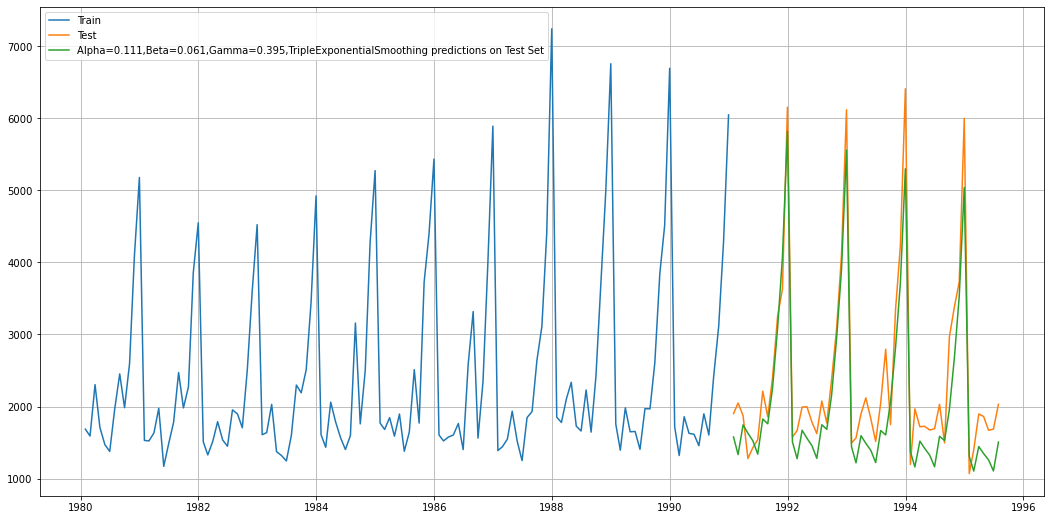
For Alpha =0.995 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 1316.035

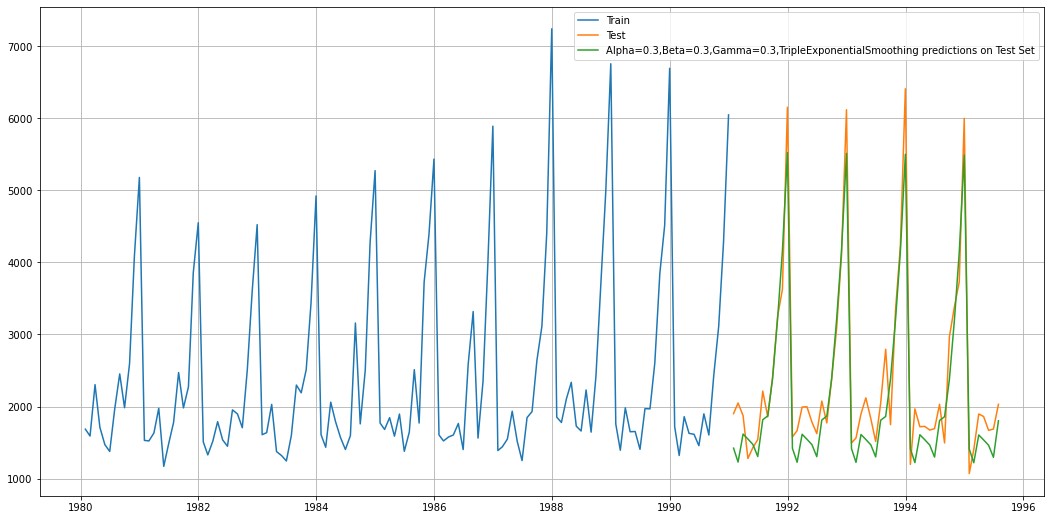
. . .

**Model 6: Double Exponential Smoothing (Holt’s Model)**



**Model 7: Triple Exponential Smoothing (Holt-Winter’s Model)**





For Alpha=0.111,Beta=0.061,Gamma=0.395, Triple Exponential Smoothing Model forecast on the Test Data, RMSE is 469.768

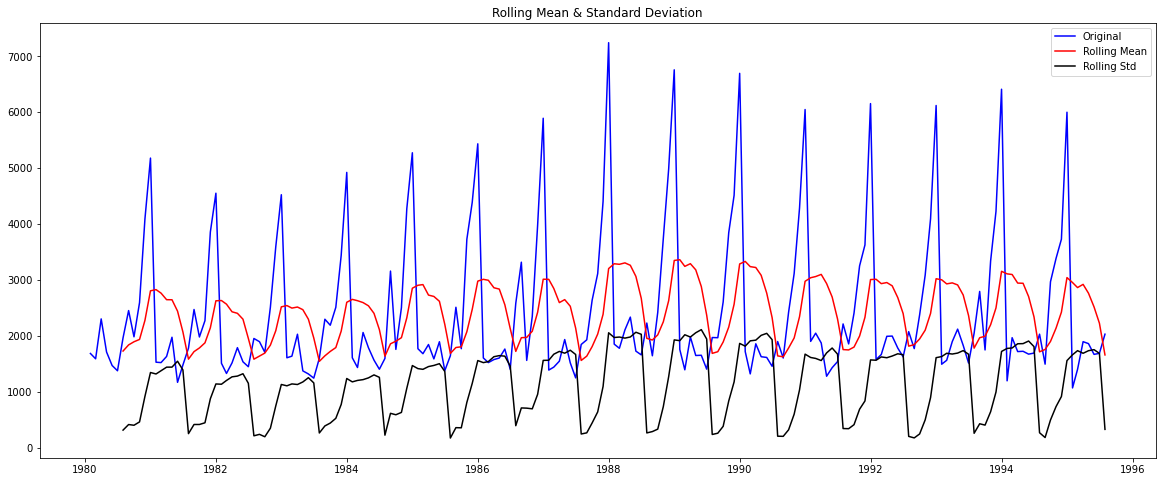
. . .

* 1. **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.**

**Dickey-Fuller Test**

Null Hypothesis H0 – Series is not Stationary.

Alternative Hypothesis H1 – Series is Stationary.



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

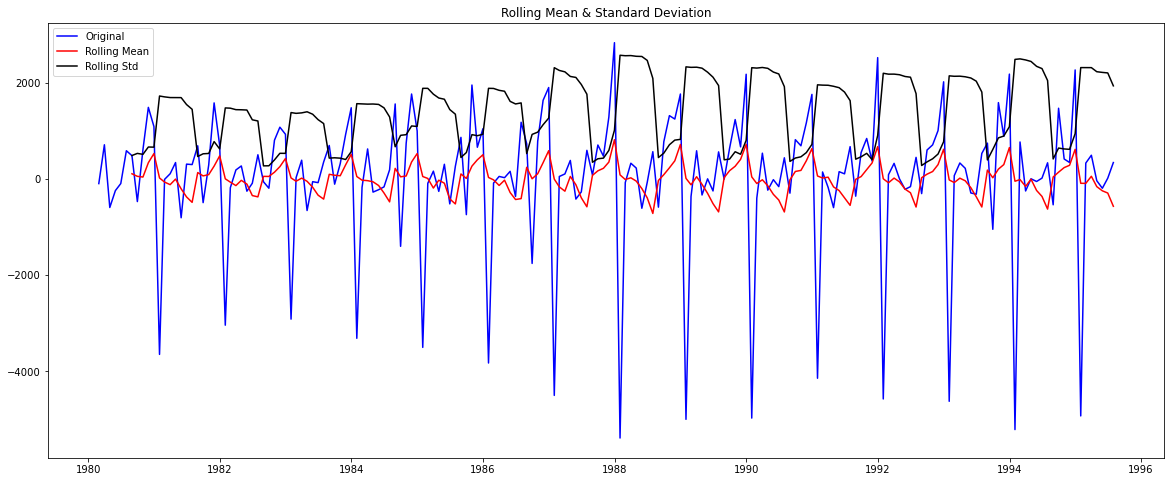
Critical Value (5%) -2.878202

Critical Value (10%) -2.575653

dtype: float64

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

Let us take a difference of order 1 and check if the Time Series is stationary or not.



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

We see that after taking a difference of order 1 the series has become stationary at alpha = 0.05.

* 1. **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.**

**Automated ARIMA Model**

The following loop helps us in getting a combination of different parameters of p and q in the range of 0 and 2.

We have kept the value of d as 1 as we need to take a difference of the series to make it stationary.

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)

Model calculated for different p and q values and sorted with lowest AIC values:

| **param** | **AIC** |
| --- | --- |
| **8** | (2, 1, 2) | 2210.626049 |
| **7** | (2, 1, 1) | 2232.360490 |
| **2** | (0, 1, 2) | 2232.783098 |
| **5** | (1, 1, 2) | 2233.597647 |
| **4** | (1, 1, 1) | 2235.013945 |
| **6** | (2, 1, 0) | 2262.035600 |
| **1** | (0, 1, 1) | 2264.906439 |
| **3** | (1, 1, 0) | 2268.528061 |
| **0** | (0, 1, 0) | 2269.582796 |

ARIMA Model Results

==============================================================================

Dep. Variable: D.Sparkling No. Observations: 131

Model: ARIMA(2, 1, 2) Log Likelihood -1099.313

Method: css-mle S.D. of innovations 1013.755

Date: Sun, 20 Feb 2022 AIC 2210.626

Time: 10:08:47 BIC 2227.877

Sample: 02-29-1980 HQIC 2217.636

- 12-31-1990

=====================================================================================

coef std err z P>|z| [0.025 0.975]

-------------------------------------------------------------------------------------

const 5.5845 0.519 10.753 0.000 4.567 6.602

ar.L1.D.Sparkling 1.2698 0.075 17.040 0.000 1.124 1.416

ar.L2.D.Sparkling -0.5601 0.074 -7.617 0.000 -0.704 -0.416

ma.L1.D.Sparkling -1.9957 0.043 -46.821 0.000 -2.079 -1.912

ma.L2.D.Sparkling 0.9957 0.043 23.291 0.000 0.912 1.079

Roots

=============================================================================

Real Imaginary Modulus Frequency

-----------------------------------------------------------------------------

AR.1 1.1335 -0.7074j 1.3361 -0.0888

AR.2 1.1335 +0.7074j 1.3361 0.0888

MA.1 1.0000 +0.0000j 1.0000 0.0000

MA.2 1.0043 +0.0000j 1.0043 0.0000

-----------------------------------------------------------------------------

Predicting on Test set using this model and evaluating the model:

RMSE: 1374.0370092580251

**Automated SARIMA Model**

The following loop helps us in getting a combination of different parameters of p, q, P, Q in the range of 0 and 3.

We have kept the value of d in the range (1,2) and D in the range (0,1).

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)

Model calculated for different p, q, d, P, Q, D values and sorted by least AIC values:

|  | **param** | **seasonal** | **AIC** |
| --- | --- | --- | --- |
| **50** | (1, 1, 2) | (1, 0, 2, 12) | 1555.584247 |
| **53** | (1, 1, 2) | (2, 0, 2, 12) | 1555.934563 |
| **26** | (0, 1, 2) | (2, 0, 2, 12) | 1557.121564 |
| **23** | (0, 1, 2) | (1, 0, 2, 12) | 1557.160507 |
| **77** | (2, 1, 2) | (1, 0, 2, 12) | 1557.340403 |

SARIMAX Results

==========================================================================================

Dep. Variable: y No. Observations: 132

Model: SARIMAX(1, 1, 2)x(1, 0, 2, 12) Log Likelihood -770.792

Date: Sun, 20 Feb 2022 AIC 1555.584

Time: 10:11:05 BIC 1574.095

Sample: 0 HQIC 1563.083

- 132

Covariance Type: opg

==============================================================================

coef std err z P>|z| [0.025 0.975]

------------------------------------------------------------------------------

ar.L1 -0.6282 0.255 -2.463 0.014 -1.128 -0.128

ma.L1 -0.1041 0.225 -0.463 0.643 -0.545 0.337

ma.L2 -0.7276 0.154 -4.734 0.000 -1.029 -0.426

ar.S.L12 1.0439 0.014 72.840 0.000 1.016 1.072

ma.S.L12 -0.5550 0.098 -5.663 0.000 -0.747 -0.363

ma.S.L24 -0.1354 0.120 -1.133 0.257 -0.370 0.099

sigma2 1.506e+05 2.03e+04 7.401 0.000 1.11e+05 1.9e+05

===================================================================================

Ljung-Box (L1) (Q): 0.04 Jarque-Bera (JB): 11.72

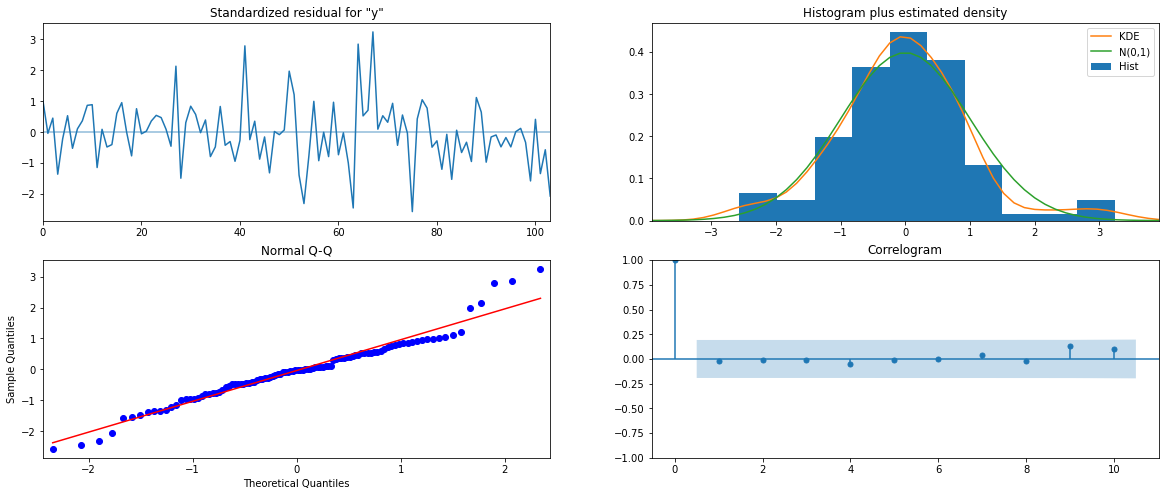
Prob(Q): 0.84 Prob(JB): 0.00

Heteroskedasticity (H): 1.47 Skew: 0.36

Prob(H) (two-sided): 0.26 Kurtosis: 4.48

===================================================================================

**Diagnostic Plot:**



Predicting on Test set using this model and evaluating the model:

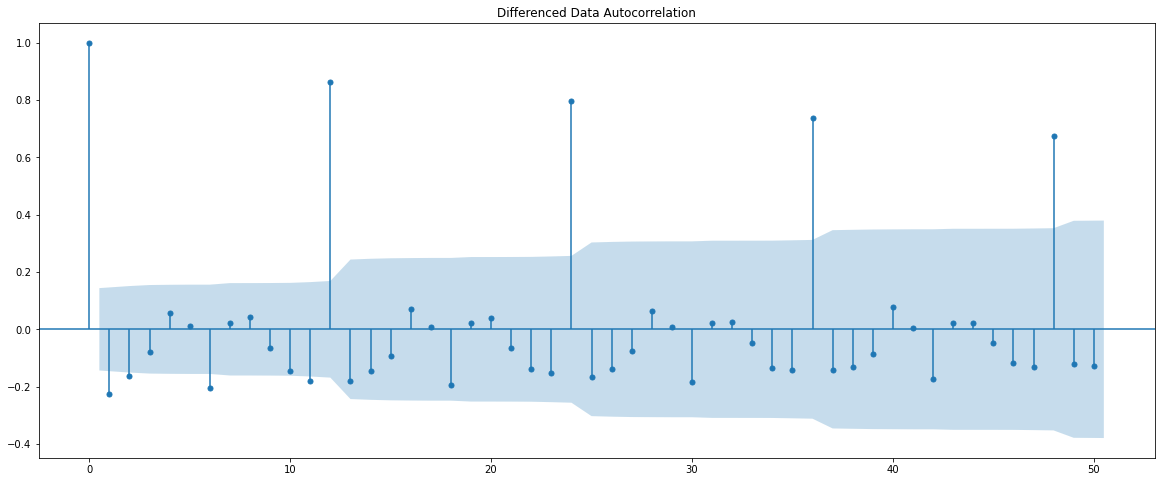
Model Forecast:

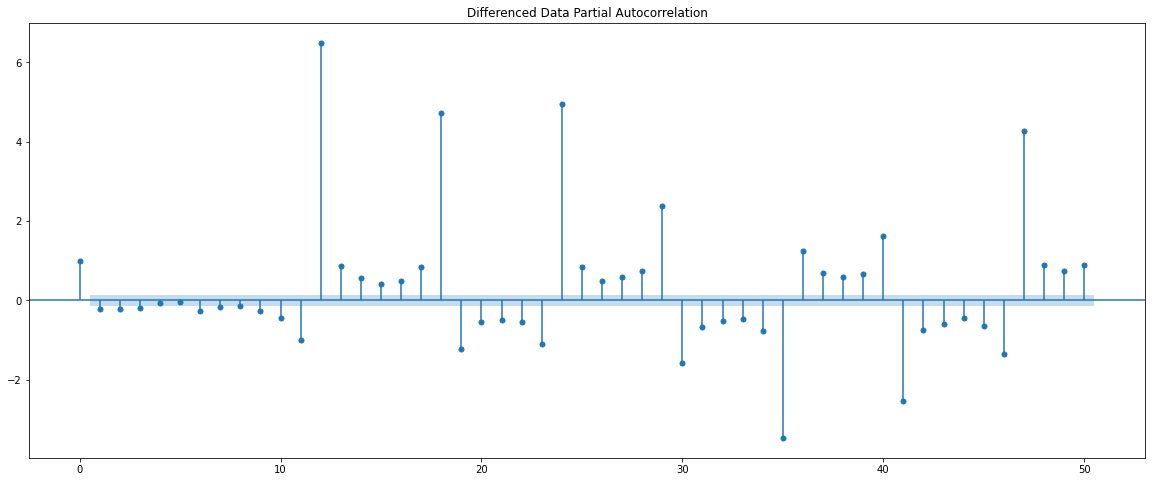
| **y** | **mean** | **mean\_se** | **mean\_ci\_lower** | **mean\_ci\_upper** |
| --- | --- | --- | --- | --- |
| **0** | 1327.386179 | 388.344789 | 566.244378 | 2088.527979 |
| **1** | 1315.110447 | 402.007725 | 527.189784 | 2103.031110 |
| **2** | 1621.588566 | 402.001331 | 833.680435 | 2409.496698 |
| **3** | 1598.867173 | 407.239040 | 800.693322 | 2397.041023 |
| **4** | 1392.687924 | 407.969111 | 593.083159 | 2192.292689 |

RMSE: 528.6216014879402

* 1. **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.**

**ARIMA Model**





Here, we have taken alpha = 0.05.

The Auto-Regressive parameter in an ARIMA model is ‘p’ which comes from the significant lag before which the PACF plot cuts off to 0.

The Moving Average parameter in an ARIMA model is ‘q’ which comes from the significant lag before which the ACF plot cuts off to 0.

By observing the plots we can see that the PACF and the ACF plot cuts off at lag 3 and 2.

ARIMA Model Results

==============================================================================

Dep. Variable: D.Sparkling No. Observations: 131

Model: ARIMA(3, 1, 2) Log Likelihood -1107.464

Method: css-mle S.D. of innovations 1106.238

Date: Sun, 20 Feb 2022 AIC 2228.928

Time: 10:12:17 BIC 2249.054

Sample: 02-29-1980 HQIC 2237.106

- 12-31-1990

=====================================================================================

coef std err z P>|z| [0.025 0.975]

-------------------------------------------------------------------------------------

const 5.9844 3.643 1.643 0.100 -1.156 13.125

ar.L1.D.Sparkling -0.4420 1.28e-05 -3.46e+04 0.000 -0.442 -0.442

ar.L2.D.Sparkling 0.3079 4.63e-05 6645.786 0.000 0.308 0.308

ar.L3.D.Sparkling -0.2501 3.81e-05 -6560.861 0.000 -0.250 -0.250

ma.L1.D.Sparkling -0.0008 0.020 -0.040 0.968 -0.039 0.037

ma.L2.D.Sparkling -0.9992 0.020 -51.220 0.000 -1.037 -0.961

Roots

=============================================================================

Real Imaginary Modulus Frequency

-----------------------------------------------------------------------------

AR.1 -1.0000 -0.0000j 1.0000 -0.5000

AR.2 1.1156 -1.6594j 1.9996 -0.1558

AR.3 1.1156 +1.6594j 1.9996 0.1558

MA.1 1.0000 +0.0000j 1.0000 0.0000

MA.2 -1.0008 +0.0000j 1.0008 0.5000

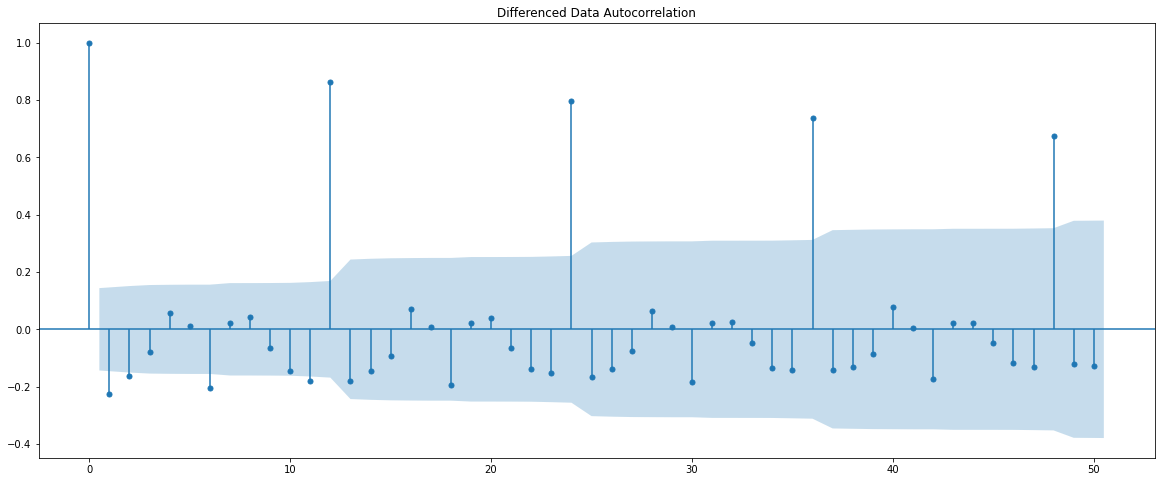
-----------------------------------------------------------------------------

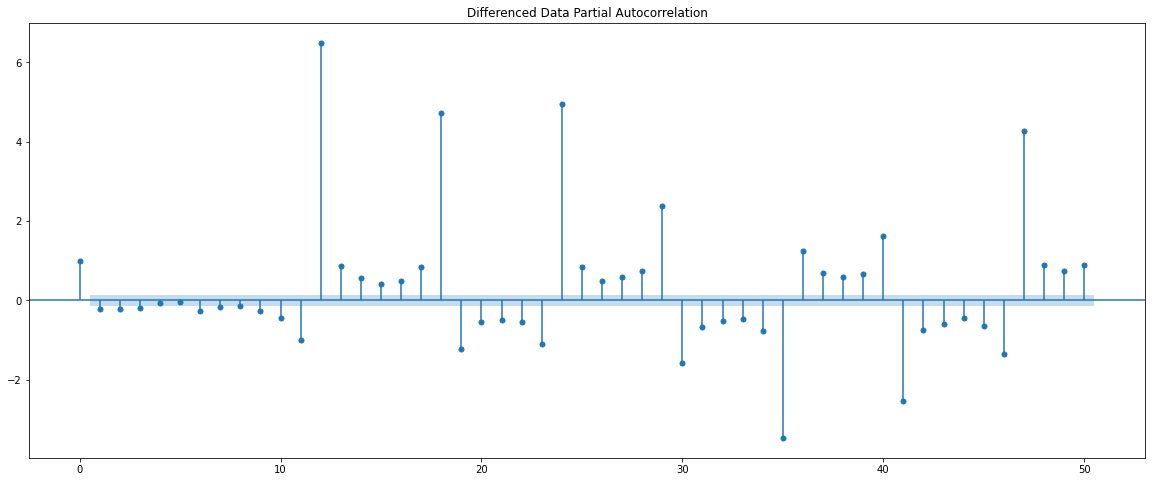
Predicting on Test set using this model and evaluating the model:

RMSE: 1379.0491231340152

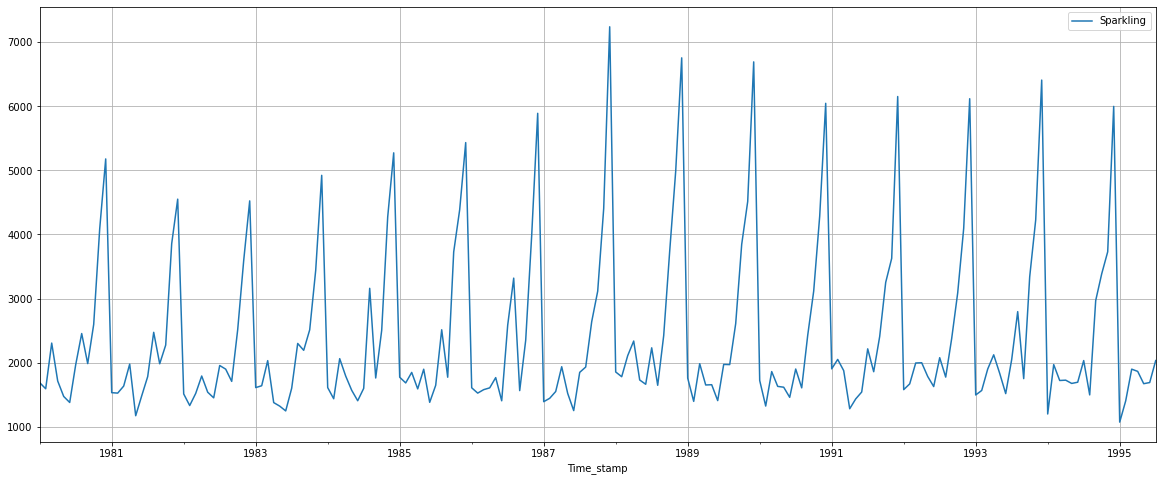
. . .

**SARIMA Model**

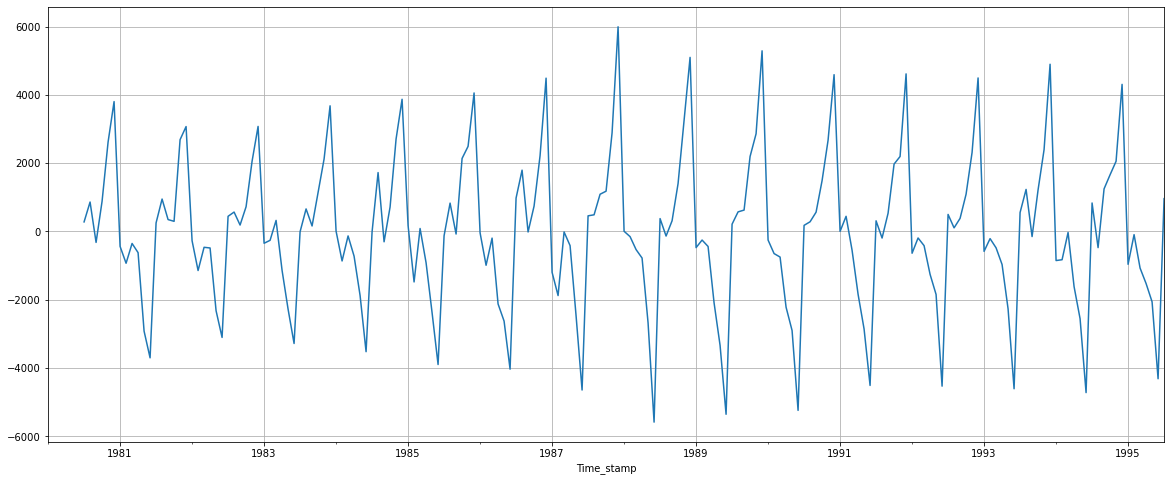


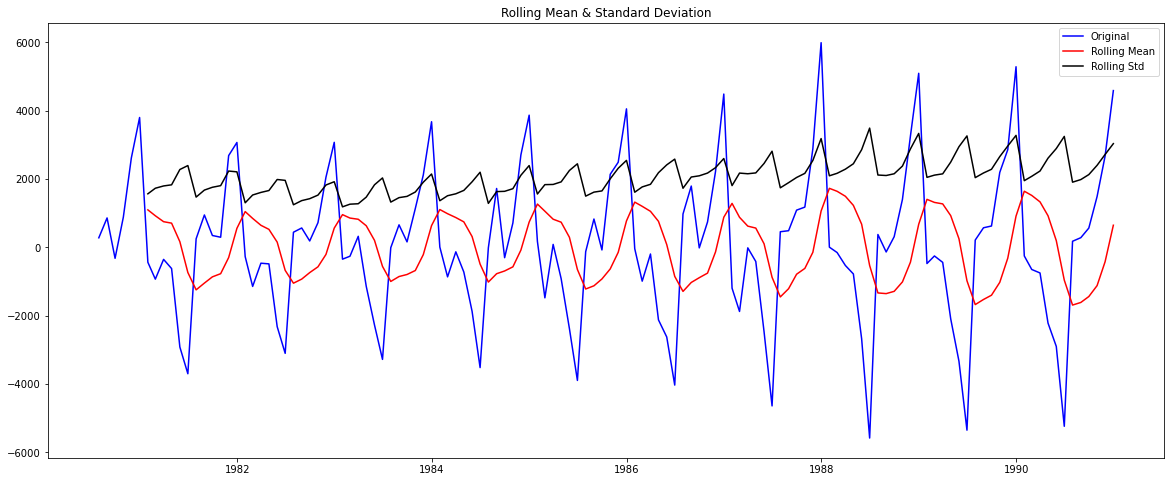


**Plotting the Dataset:**



We can see that there is seasonality present. So, we take a seasonal differencing and then check the series.





Results of Dickey-Fuller Test:

Test Statistic -8.181919e+00

p-value 8.088278e-13

#Lags Used 6.000000e+00

Number of Observations Used 1.190000e+02

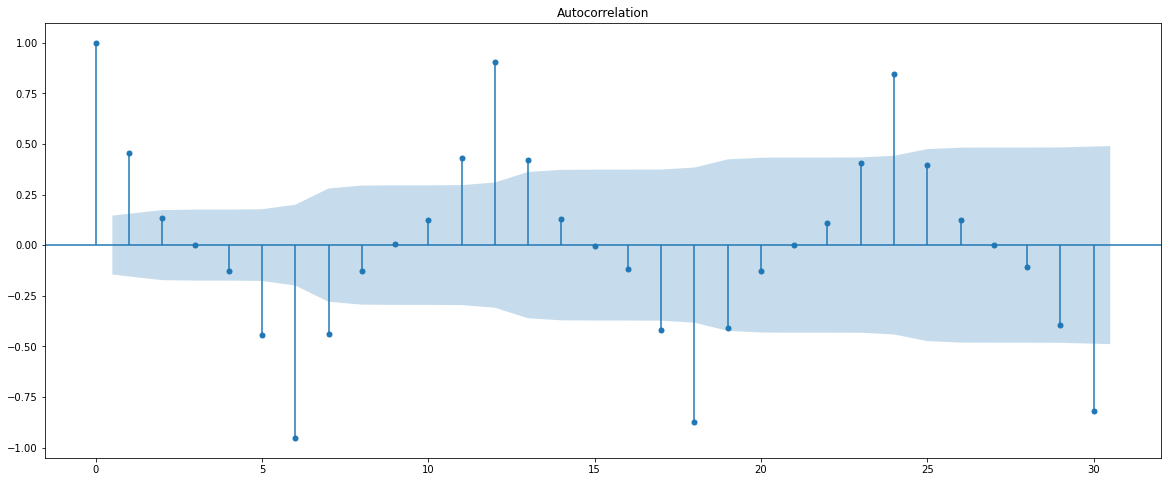
Critical Value (1%) -3.486535e+00

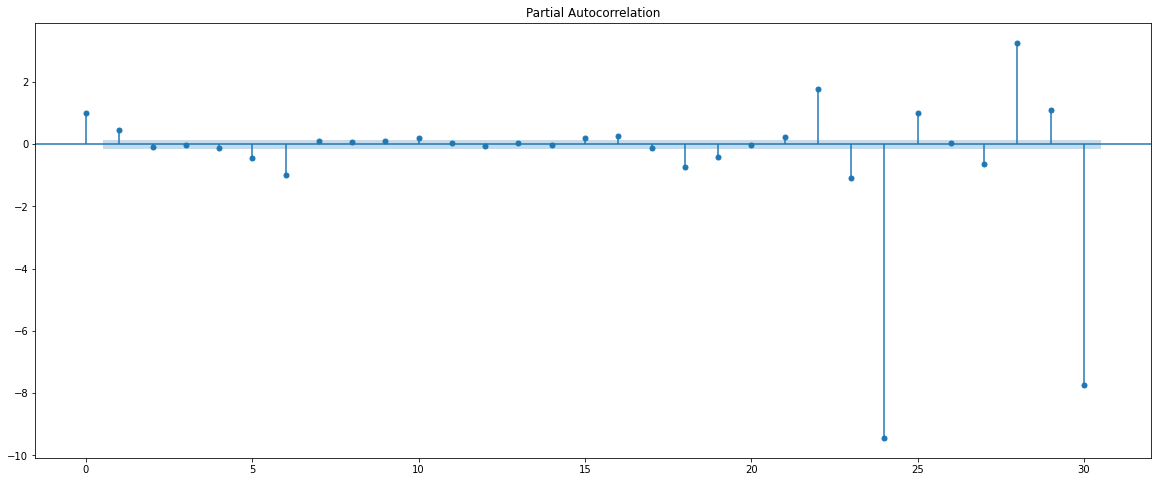
Critical Value (5%) -2.886151e+00

Critical Value (10%) -2.579896e+00

dtype: float64

**ACF and PACF plot after differencing:**





Here, we have taken alpha = 0.05.

We are going to take the seasonal period as 6. We will keep the p(1) and q(1) parameters the same as the ARIMA model.

The Auto-Regressive parameter in an ARIMA model is ‘P’ which comes from the significant lag after which the PACF plot cuts off to 0.

The Moving Average parameter in a SARIMA model is ‘Q’ which comes from the significant lag after which the ACF plot cuts off to 0.

Since 6 is the seasonal period, we must remember to check the ACF and PACF plots only at multiples of 6.

By observing the plots we can see that the ACF and the PACF cut off at 1 and 1.

SARIMAX Results

===========================================================================================

Dep. Variable: y No. Observations: 132

Model: SARIMAX(3, 1, 2)x(1, 1, [1], 6) Log Likelihood -866.759

Date: Sun, 20 Feb 2022 AIC 1749.519

Time: 10:14:19 BIC 1771.547

Sample: 0 HQIC 1758.461

- 132

Covariance Type: opg

==============================================================================

coef std err z P>|z| [0.025 0.975]

------------------------------------------------------------------------------

ar.L1 -0.0416 0.896 -0.046 0.963 -1.798 1.715

ar.L2 -0.0162 0.272 -0.060 0.953 -0.548 0.516

ar.L3 0.0884 0.140 0.630 0.529 -0.187 0.363

ma.L1 -0.6756 0.956 -0.706 0.480 -2.550 1.199

ma.L2 -0.3244 0.884 -0.367 0.714 -2.057 1.408

ar.S.L6 -0.9784 0.025 -39.614 0.000 -1.027 -0.930

ma.S.L6 -0.1195 0.132 -0.907 0.365 -0.378 0.139

sigma2 1.746e+05 3.86e-06 4.53e+10 0.000 1.75e+05 1.75e+05

===================================================================================

Ljung-Box (L1) (Q): 0.01 Jarque-Bera (JB): 10.80

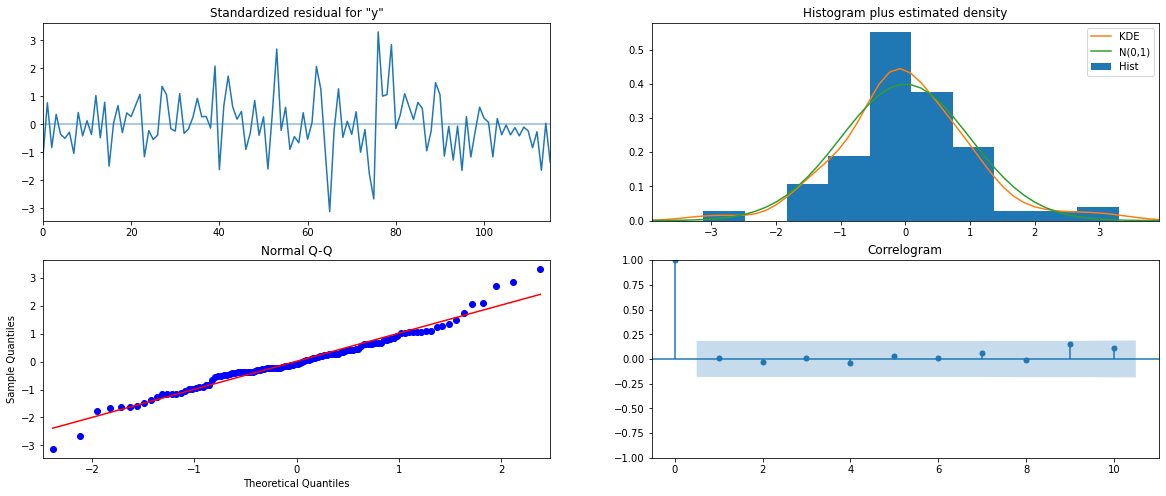
Prob(Q): 0.93 Prob(JB): 0.00

Heteroskedasticity (H): 1.72 Skew: 0.23

Prob(H) (two-sided): 0.09 Kurtosis: 4.42

===================================================================================

**Diagnostic Plot:**



Predicting on the Test Set using this model and evaluating the model:

RMSE: 329.88649987548496

Forecasting test set with confidence interval:

| **y** | **mean** | **mean\_se** | **mean\_ci\_lower** | **mean\_ci\_upper** |
| --- | --- | --- | --- | --- |
| **0** | 1529.776845 | 419.586764 | 707.401900 | 2352.151791 |
| **1** | 1386.732195 | 436.989129 | 530.249240 | 2243.215150 |
| **2** | 1863.791987 | 437.026609 | 1007.235573 | 2720.348400 |
| **3** | 1741.536788 | 438.833155 | 881.439608 | 2601.633968 |
| **4** | 1712.828898 | 439.027078 | 852.351637 | 2573.306158 |

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

| **Test RMSE** |
| --- |
| **SARIMA(3,1,2)(1,1,1,6) based on ACF & PACF** | 329.886500 |
| **Alpha=0.3,Beta=0.3,Gamma=0.3,TripleExponentialSmoothing** | 392.786198 |
| **Alpha=0.111,Beta=0.061,Gamma=0.395,TripleExponentialSmoothing** | 469.767970 |
| **SARIMA (1,1,2)(1,0,2,12)** | 528.621601 |
| **2pointTrailingMovingAverage** | 813.400684 |
| **4pointTrailingMovingAverage** | 1156.589694 |
| **SimpleAverageModel** | 1275.081804 |
| **6pointTrailingMovingAverage** | 1283.927428 |
| **Alpha=0.995,SimpleExponentialSmoothing** | 1316.035487 |
| **9pointTrailingMovingAverage** | 1346.278315 |
| **ARIMA (2,1,2)** | 1374.037009 |
| **ARIMA (3,1,2) based on ACF & PACF** | 1379.049123 |
| **RegressionOnTime** | 1389.135175 |
| **Alpha=0.3,SimpleExponentialSmoothing** | 1935.507132 |
| **NaiveModel** | 3864.279352 |
| **Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing** | 18259.110704 |

* 1. **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.**

From the above analysis, the SARIMA model based on the ACF, PACF plot would be the optimum model for this dataset because it has a low RMSE value.

SARIMAX Results

===========================================================================================

Dep. Variable: y No. Observations: 132

Model: SARIMAX(3, 1, 2)x(1, 1, [1], 6) Log Likelihood -866.759

Date: Sun, 20 Feb 2022 AIC 1749.519

Time: 10:14:19 BIC 1771.547

Sample: 0 HQIC 1758.461

- 132

Covariance Type: opg

==============================================================================

coef std err z P>|z| [0.025 0.975]

------------------------------------------------------------------------------

ar.L1 -0.0416 0.896 -0.046 0.963 -1.798 1.715

ar.L2 -0.0162 0.272 -0.060 0.953 -0.548 0.516

ar.L3 0.0884 0.140 0.630 0.529 -0.187 0.363

ma.L1 -0.6756 0.956 -0.706 0.480 -2.550 1.199

ma.L2 -0.3244 0.884 -0.367 0.714 -2.057 1.408

ar.S.L6 -0.9784 0.025 -39.614 0.000 -1.027 -0.930

ma.S.L6 -0.1195 0.132 -0.907 0.365 -0.378 0.139

sigma2 1.746e+05 3.86e-06 4.53e+10 0.000 1.75e+05 1.75e+05

===================================================================================

Ljung-Box (L1) (Q): 0.01 Jarque-Bera (JB): 10.80

Prob(Q): 0.93 Prob(JB): 0.00

Heteroskedasticity (H): 1.72 Skew: 0.23

Prob(H) (two-sided): 0.09 Kurtosis: 4.42

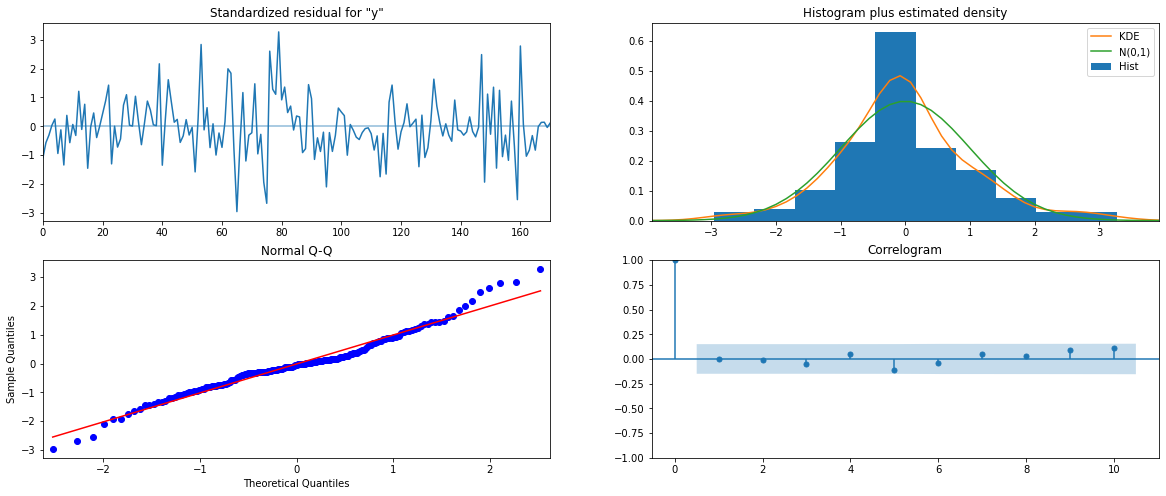
===================================================================================

Warnings:

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

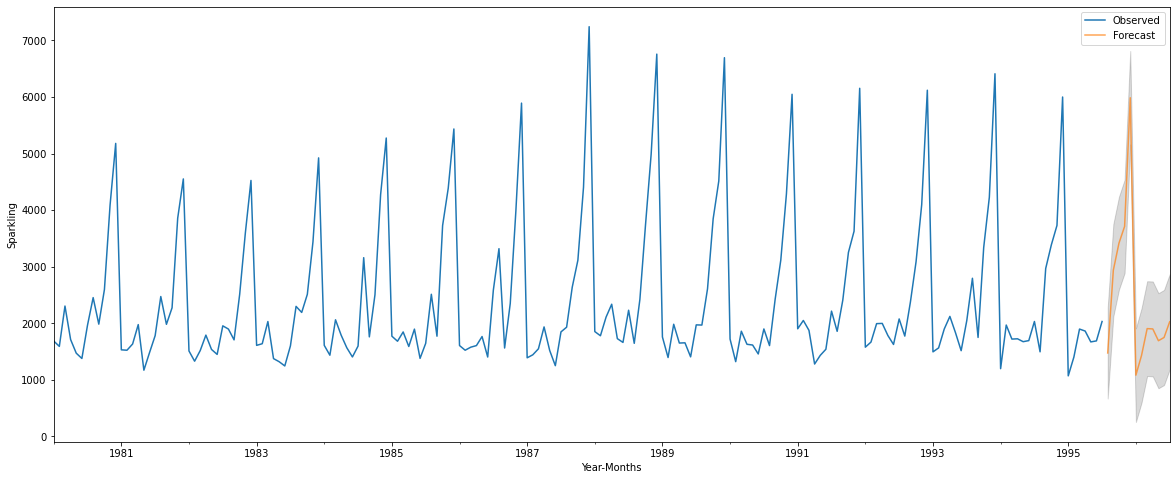
[2] Covariance matrix is singular or near-singular, with condition number 7.73e+26. Standard errors may be unstable.

**Diagnostic Plot:**



RMSE of the Full Model 571.8691917375559

Predicting 12 months into the future and plotting it with confidence interval:



* 1. **Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.**
* Sales of Sparkling Wine from 1980 to 1995 were analysed. Performing EDA gave us further insights.
* From 1981 to 1983, there is a downward trend, followed by an upward trend from 1983 to 1988, after which it again goes down till 1995.
* Median: 1900.
* The highest sale occurred in 1987, whereas the lowest occurred in 1995.
* The Monthly plot indicates that sales jump up heavily from October to December. One speculated reason for this rise in sales could be customers purchasing wine for the occasion of Christmas and New Year’s. Stocks have to be increased during this holiday season.
* Subsequently, January has recorded the lowest sales which is right after the December month of previous years.
* Trend and Seasonality are present in the dataset.
* Dataset was split for training and test set. Various models such as Linear Regression, Naïve Bayes, Simple Exponential Smoothing, Double Exponential Smoothing, Triple Exponential Smoothing, ARIMA and SARIMA built on the training data and tested on the test data.
* SARIMA model would be the best suited model for the data, due to the dataset containing seasonality. This was backed up by the RMSE value, as it had the lowest RMSE value of all the models. Hence, the SARIMA model was applied on full data.
* Sales for the next 12 months are predicted with confidence interval. Sales are still varying heavily across months, and there is a downward trend.
* Since shelf life of the wine is good, the company could just produce more wine which would meet the peak demand without additional resources during the months of highest sales.

**Problem 2: Time Series Forecasting (Rose Dataset)**

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.**

**Sample of the Dataset**

| **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 |

| **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 |

**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

**Checking for Missing Values**

YearMonth 0

Rose 2

dtype: int64

The dataset has 2 missing values. These values have to be imputed.

**Creating Time Stamp**

DatetimeIndex(['1980-01-31', '1980-02-29', '1980-03-31', '1980-04-30',

'1980-05-31', '1980-06-30', '1980-07-31', '1980-08-31',

'1980-09-30', '1980-10-31',

...

'1994-10-31', '1994-11-30', '1994-12-31', '1995-01-31',

'1995-02-28', '1995-03-31', '1995-04-30', '1995-05-31',

'1995-06-30', '1995-07-31'],

dtype='datetime64[ns]', length=187, freq='M')

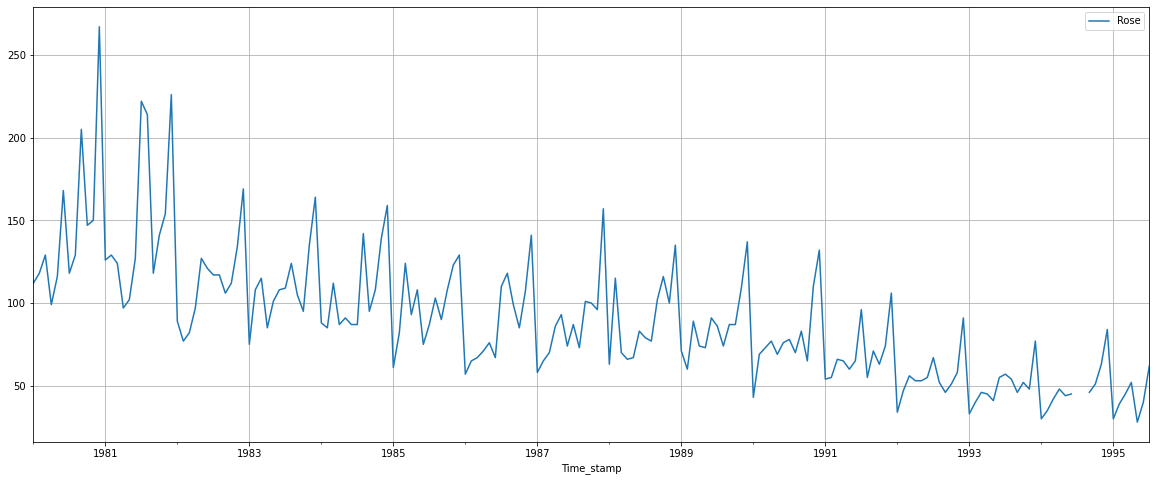
**Adding Time Stamp to Dataframe**

| **YearMonth** | **Rose** | **Time\_stamp** |
| --- | --- | --- |
| **0** | 1980-01 | 112.0 | 1980-01-31 |
| **1** | 1980-02 | 118.0 | 1980-02-29 |
| **2** | 1980-03 | 129.0 | 1980-03-31 |
| **3** | 1980-04 | 99.0 | 1980-04-30 |
| **4** | 1980-05 | 116.0 | 1980-05-31 |

**Time Stamp as Index**

|  | **Rose** |
| --- | --- |
| **Time\_stamp** |  |
| **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 |

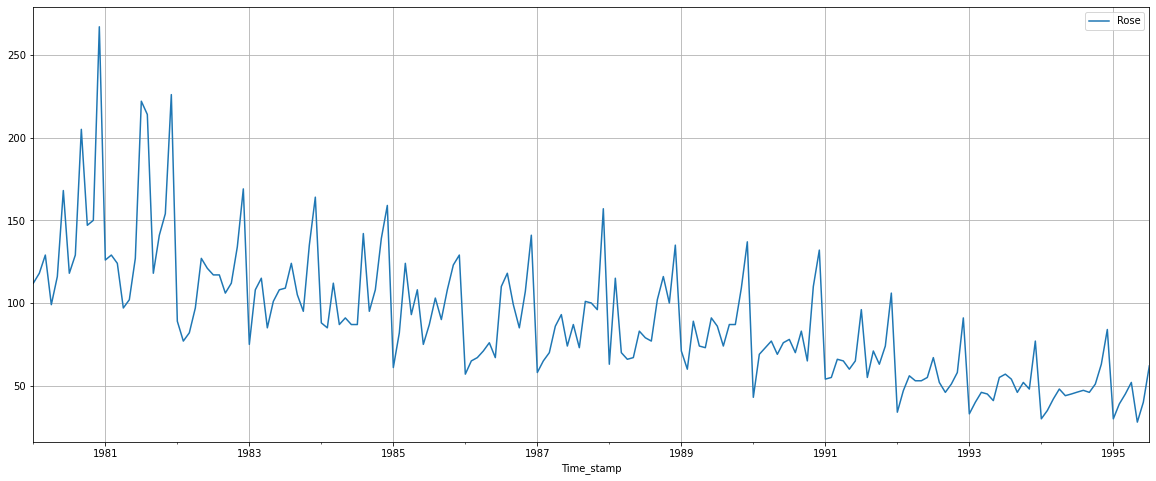
**Plotting the Dataset**

****

The data has a decreasing trend and seasonality.

Data has two missing values in the months of July and August in the year 1994. This is visible in the plot as well.

Interpolating the data with spline method, we plot it again to visualize the missing values.



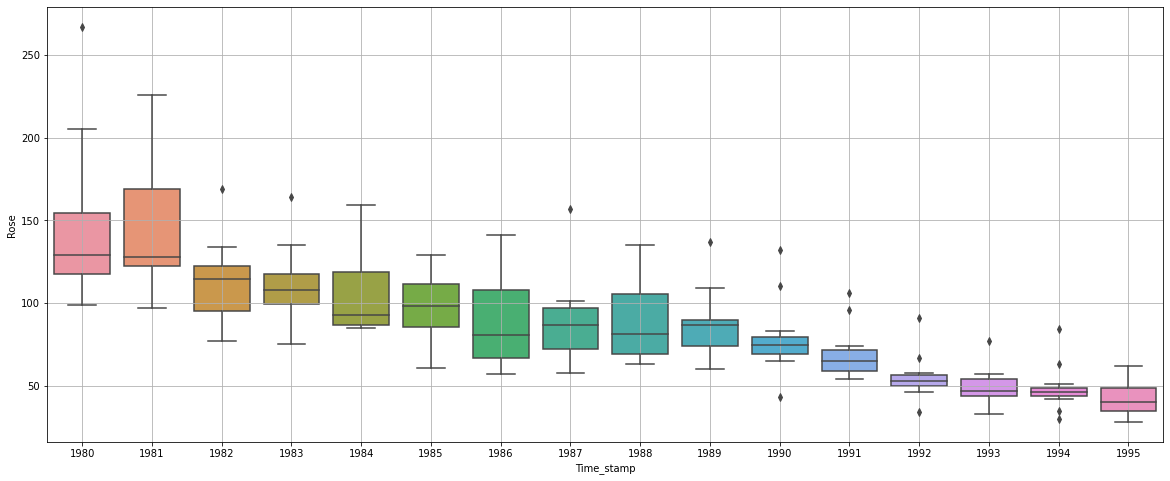
A continuous plot is observed.

**Data Description**

|  | **Rose** |
| --- | --- |
| **count** | 187.000000 |
| **mean** | 89.927152 |
| **std** | 39.224081 |
| **min** | 28.000000 |
| **25%** | 62.500000 |
| **50%** | 85.000000 |
| **75%** | 111.000000 |
| **max** | 267.000000 |

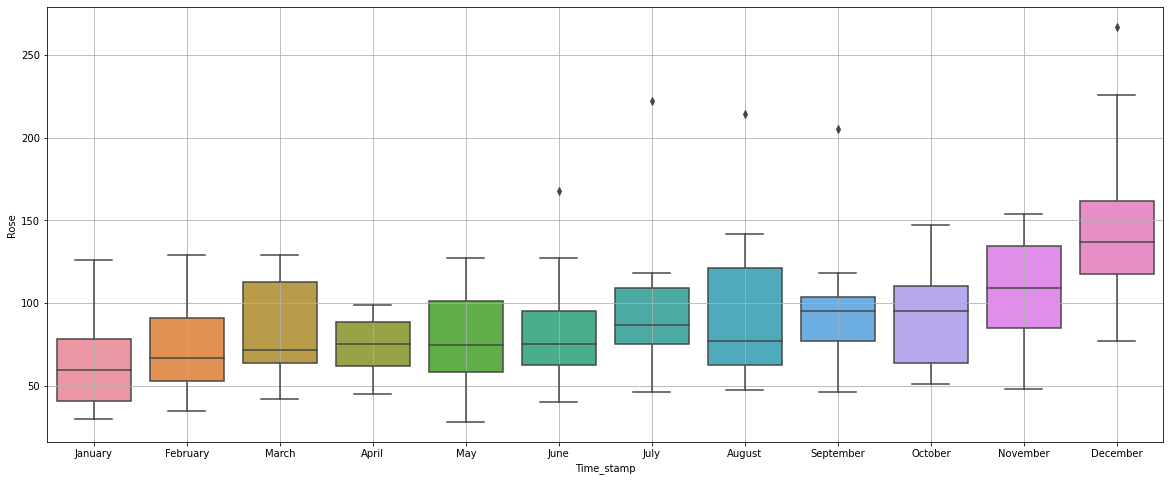
* 1. **Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.**

Sales over the years:

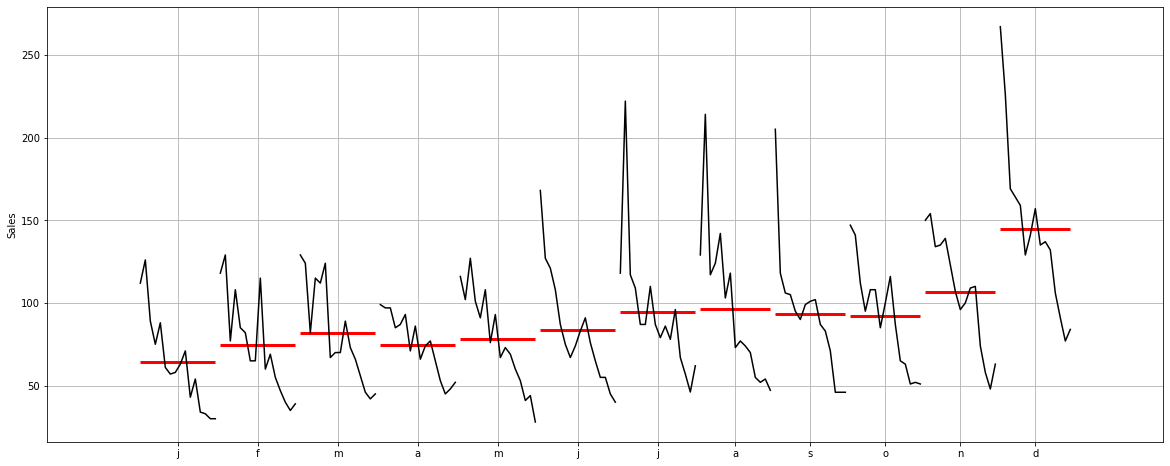
****

There is a noticeable decreasing trend.

**Monthly Boxplot**

****

Sales across different years and within different months over the years:

****

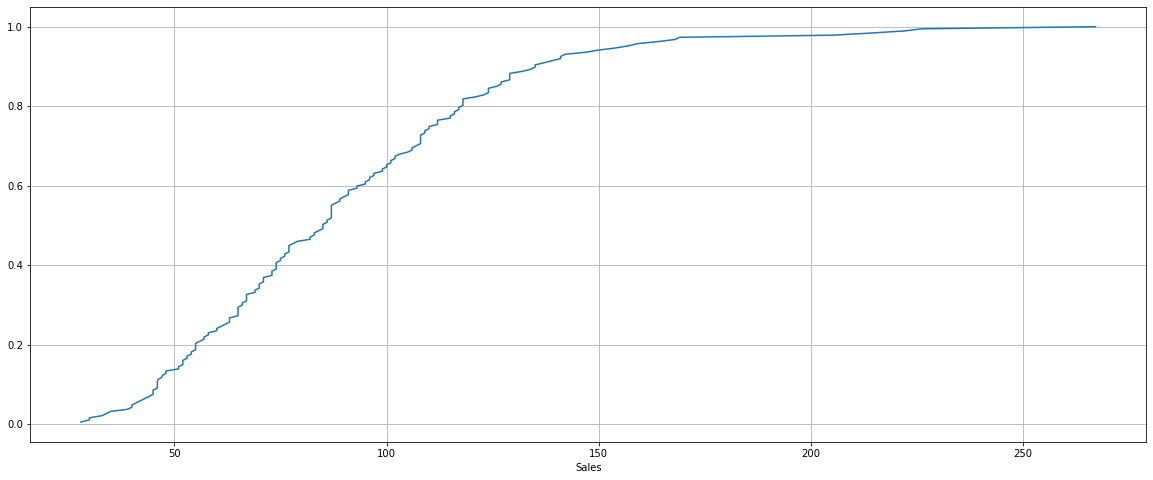
This plot shows us the behavior of time series across months.

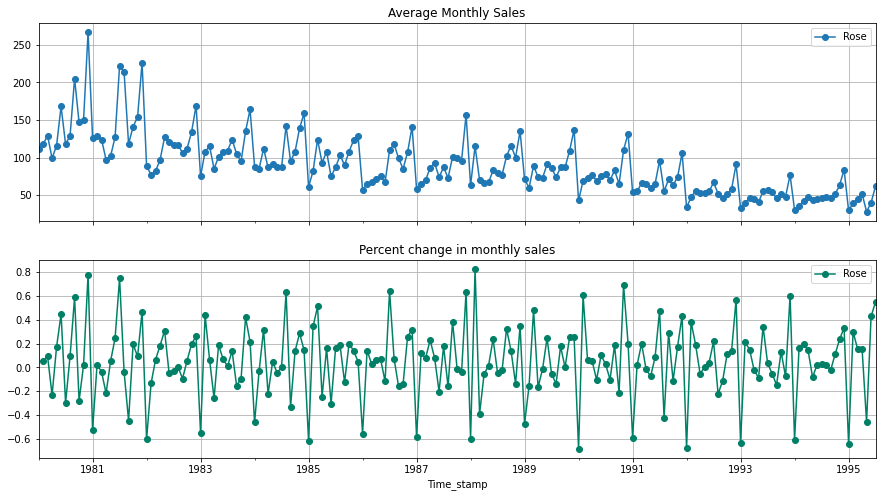
**Plotting Graph of Monthly Sales:**

| **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 | 46.155493 | 47.221907 | 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 |

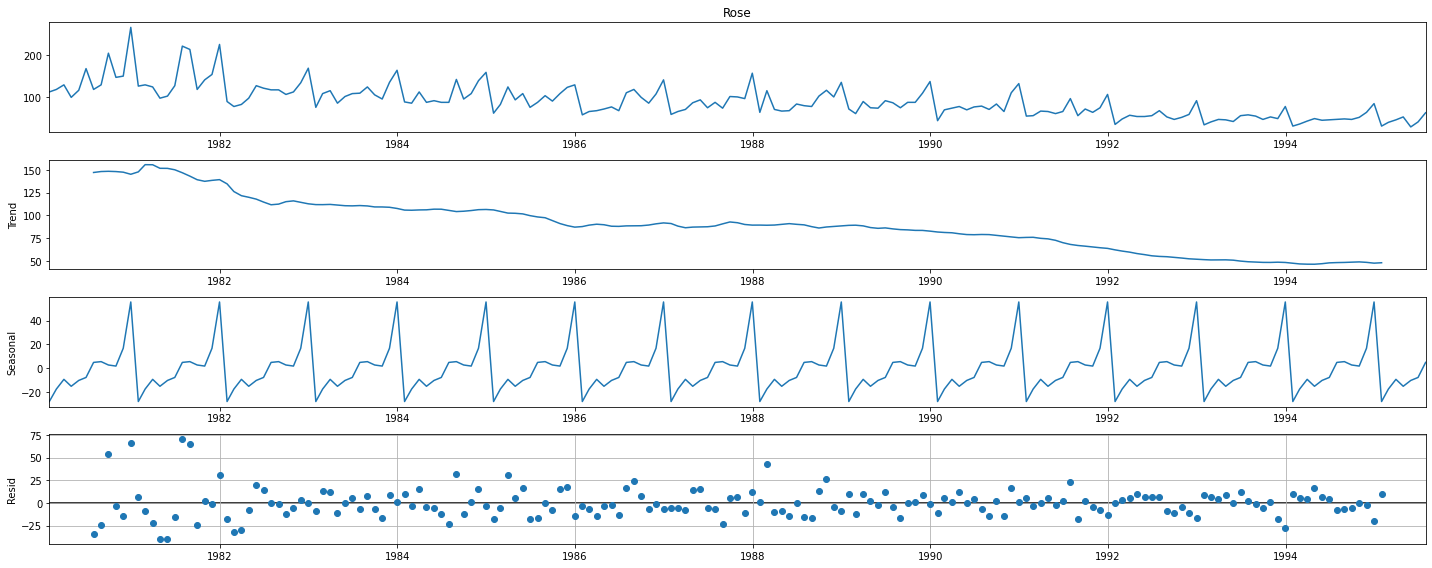
****

**Empirical Cumulative Distribution:**

****

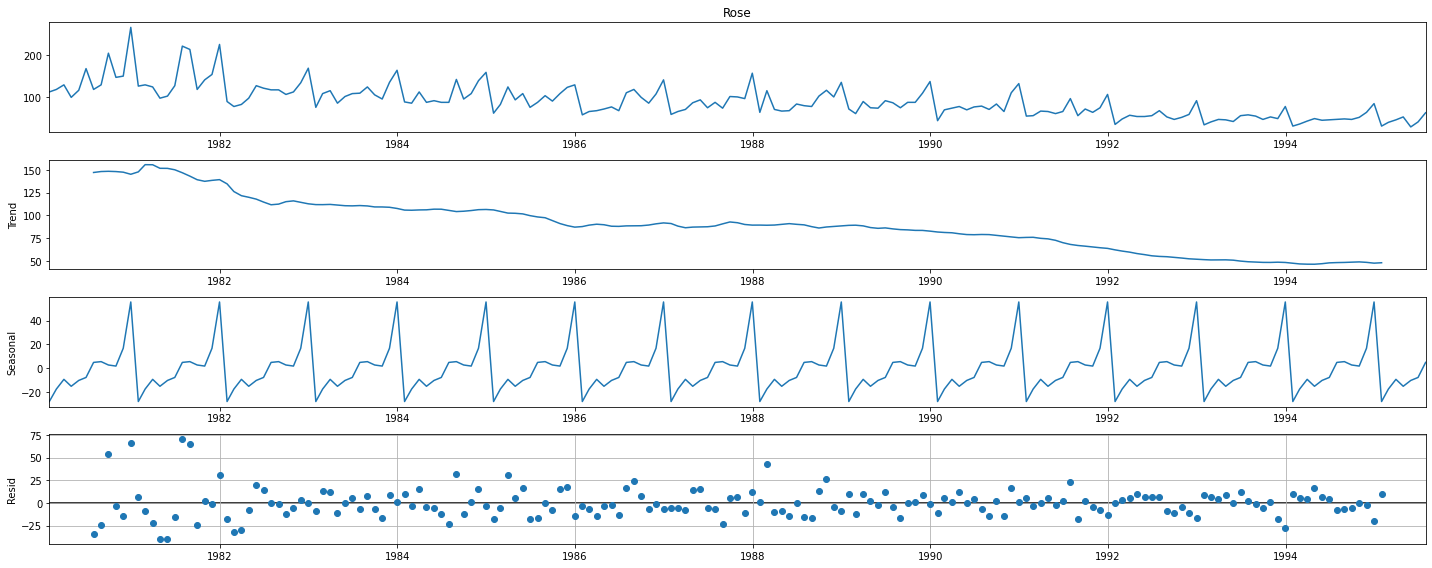
****

**Additive Decomposition:**

****

The above plot shows us that the data has decreasing trend and seasonality. Residual has some patterns. Multiplicative model can give us further insights.

**Multiplicative Decomposition:**

****

For this model, the trend and seasonality are the same. Residual has some patterns.

Additive Model is considered for further analysis.

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

1980-07-31 147.083333

1980-08-31 148.125000

1980-09-30 148.375000

1980-10-31 148.083333

Name: trend, dtype: float64

seasonality

Time\_stamp

1980-01-31 -27.921848

1980-02-29 -17.445147

1980-03-31 -9.299974

1980-04-30 -15.112474

1980-05-31 -10.210688

1980-06-30 -7.692831

1980-07-31 4.938518

1980-08-31 5.590168

1980-09-30 2.761485

1980-10-31 1.858708

Name: seasonal, dtype: float64

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 -34.021852

1980-08-31 -24.715168

1980-09-30 53.863515

1980-10-31 -2.942041

Name: resid, dtype: float64

. . .

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

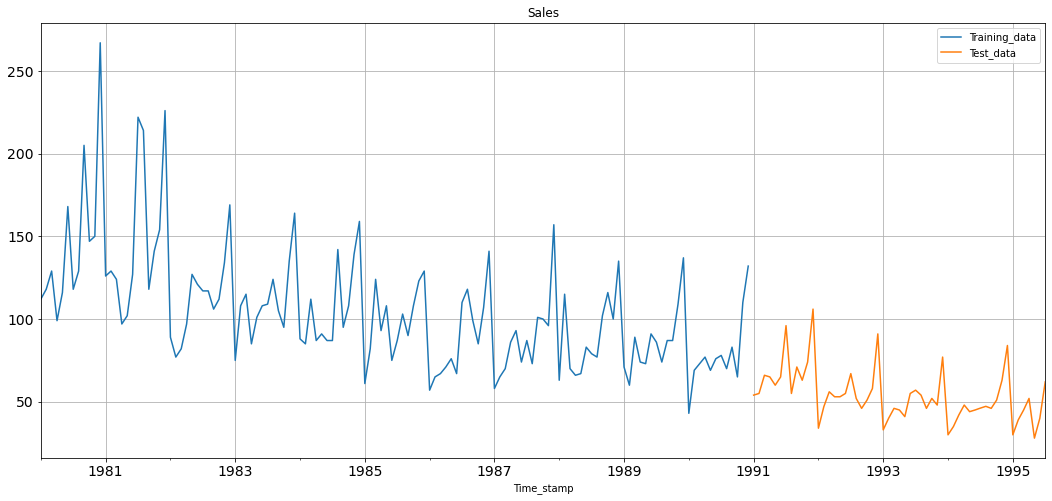
Dataset was split between Training and Test dataset. The test data starts from 1991.

**Shape of Training and Test Dataset**

(132, 1)

(55, 1)

**Plotting Training and Test Dataset:**

****

* 1. **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.**

**Model 1: Linear Regression**

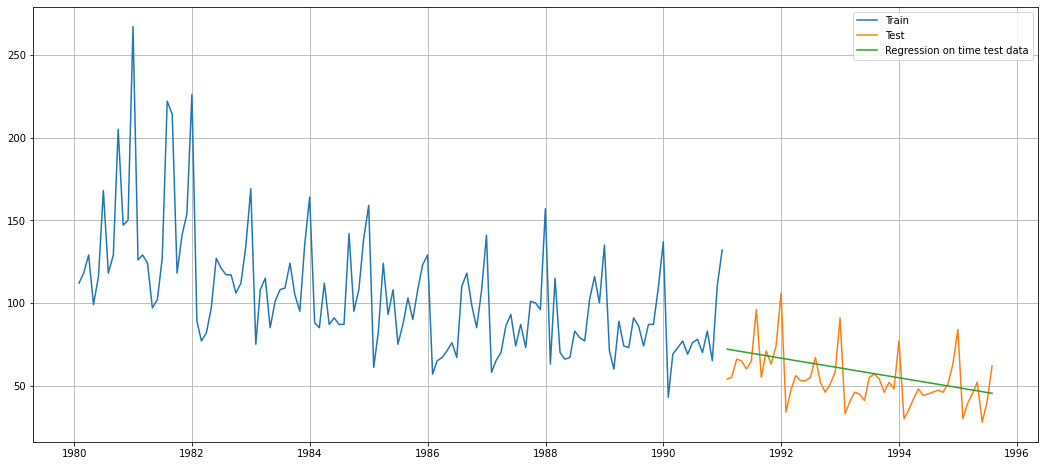
For this, we are going to regress the ‘Sales’ variable against the order of occurrence. For this we need to modify our training data before fitting it into a Linear Regression.

Training Time instance

[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, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132]

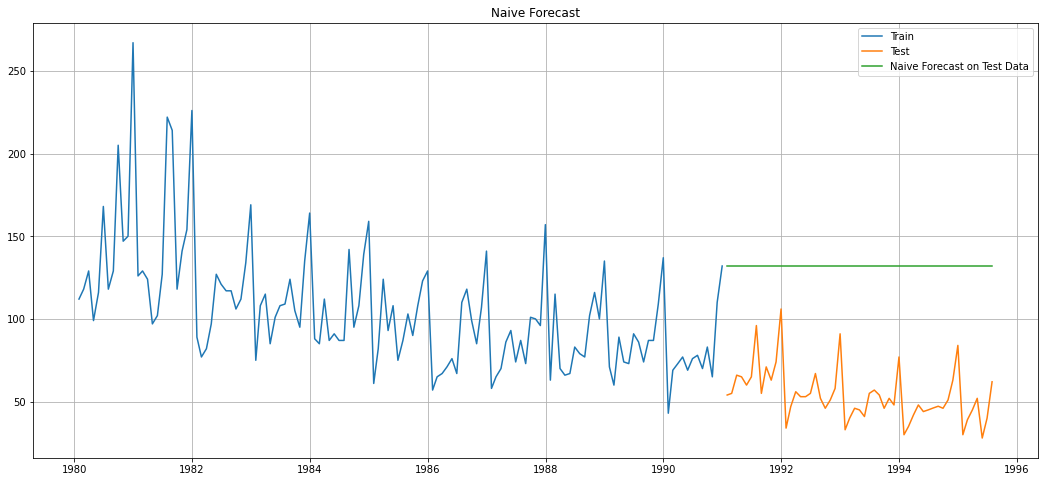
Test Time instance

[133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187]

****

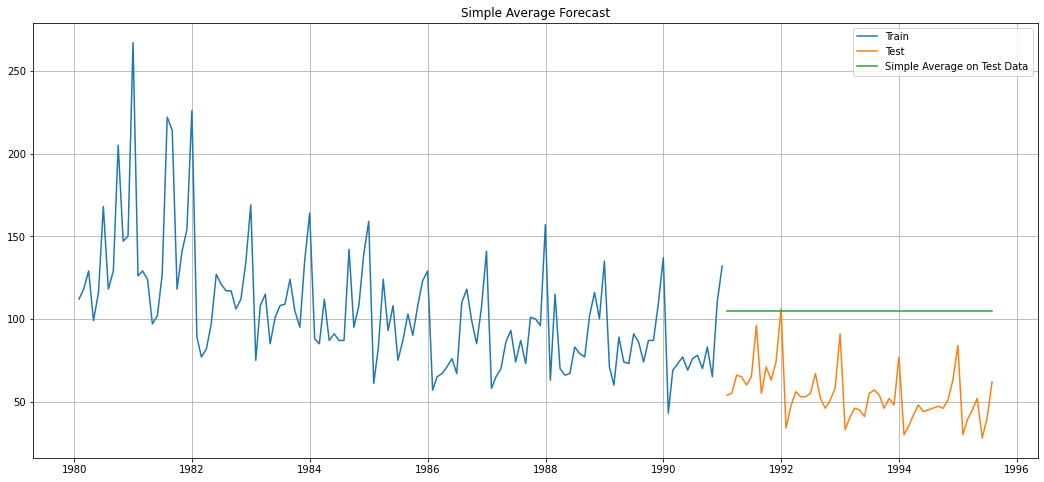
For RegressionOnTime forecast on the Test Data, RMSE is 15.255

**Model 2: Naïve Approach**

****

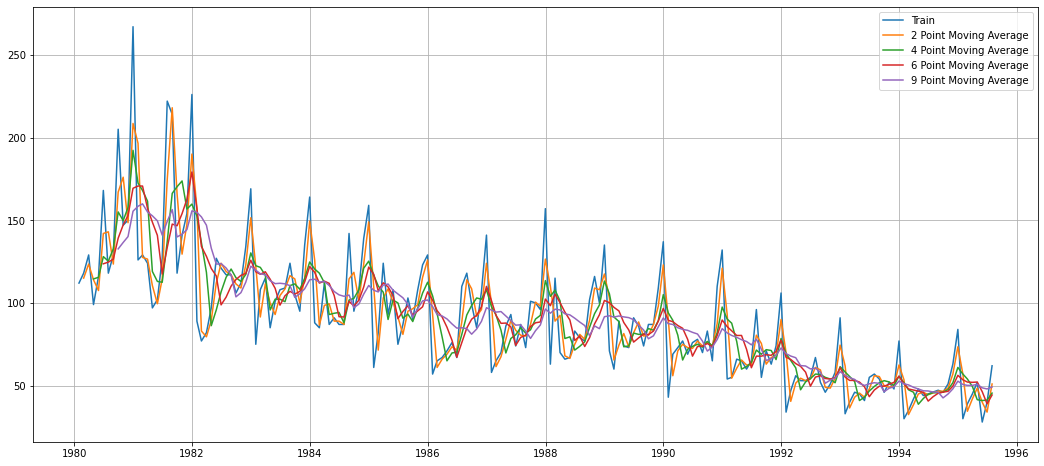
For Naive forecast on the Test Data, RMSE is 79.672

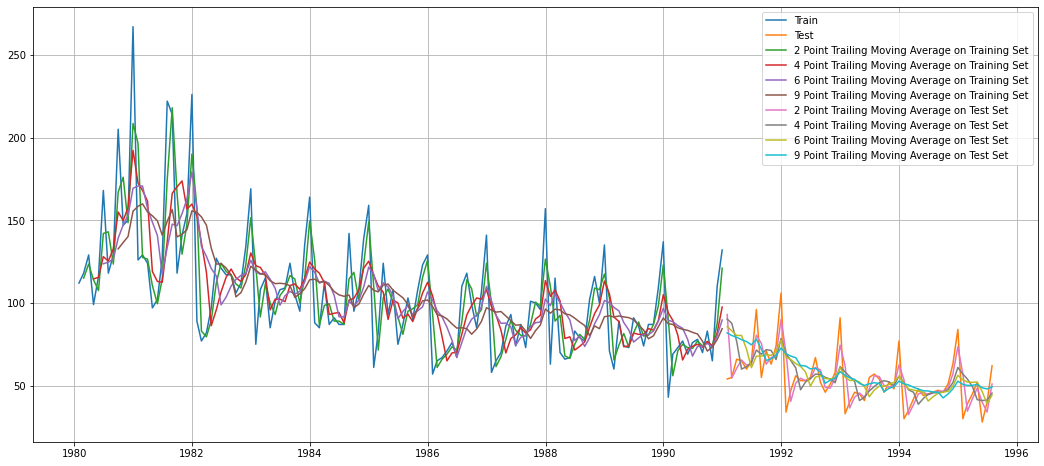
**Model 3: Simple Average**

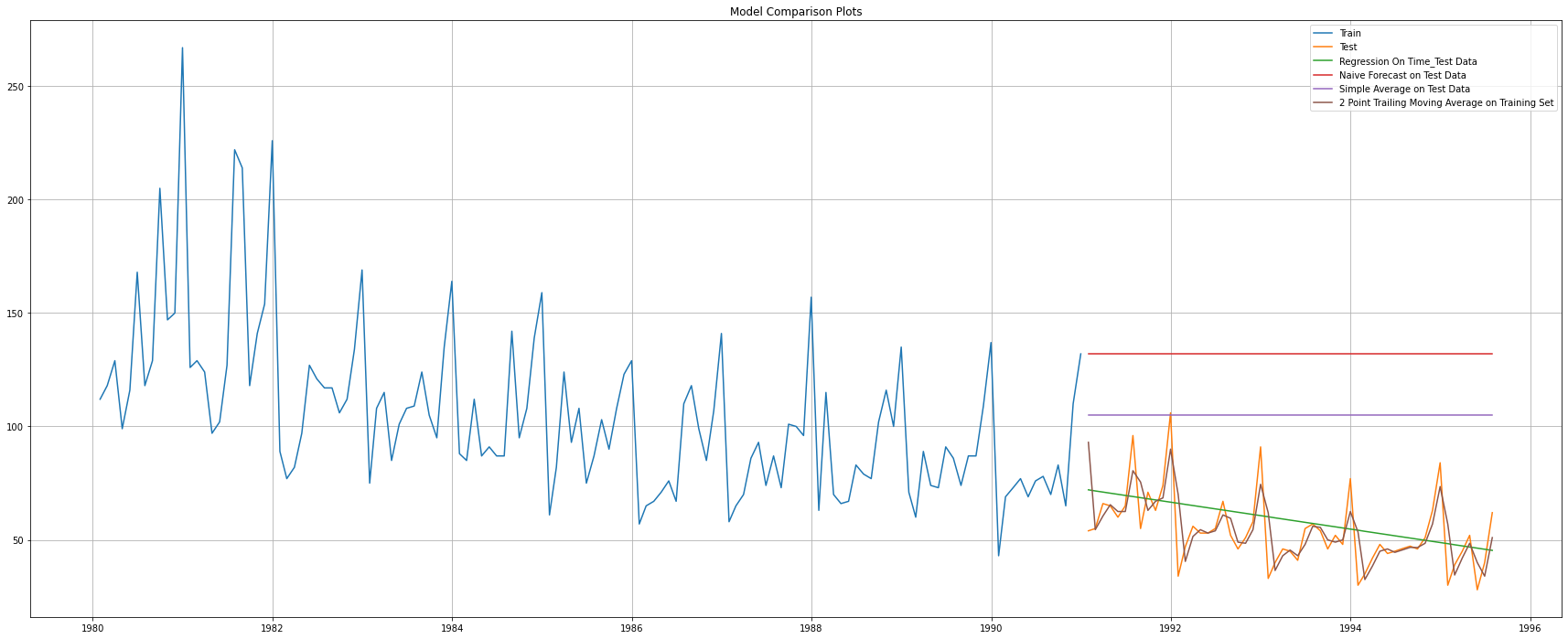
****

For Simple Average forecast on the Test Data, RMSE is 53.413

**Model 4: Moving Average**

****

****

****

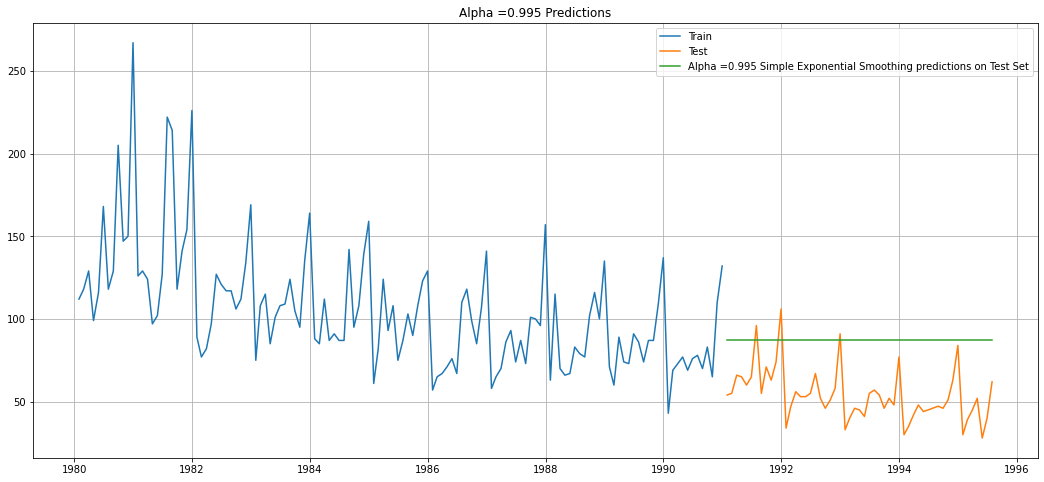
For 2 point Moving Average Model forecast on the Training Data, RMSE is 11.530

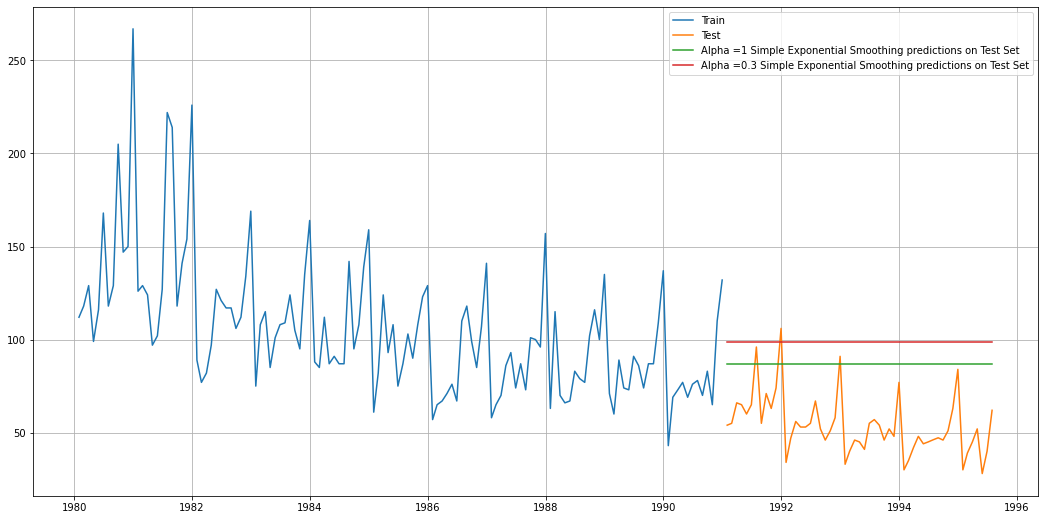
For 4 point Moving Average Model forecast on the Training Data, RMSE is 14.444

For 6 point Moving Average Model forecast on the Training Data, RMSE is 14.555

For 9 point Moving Average Model forecast on the Training Data, RMSE is 14.721

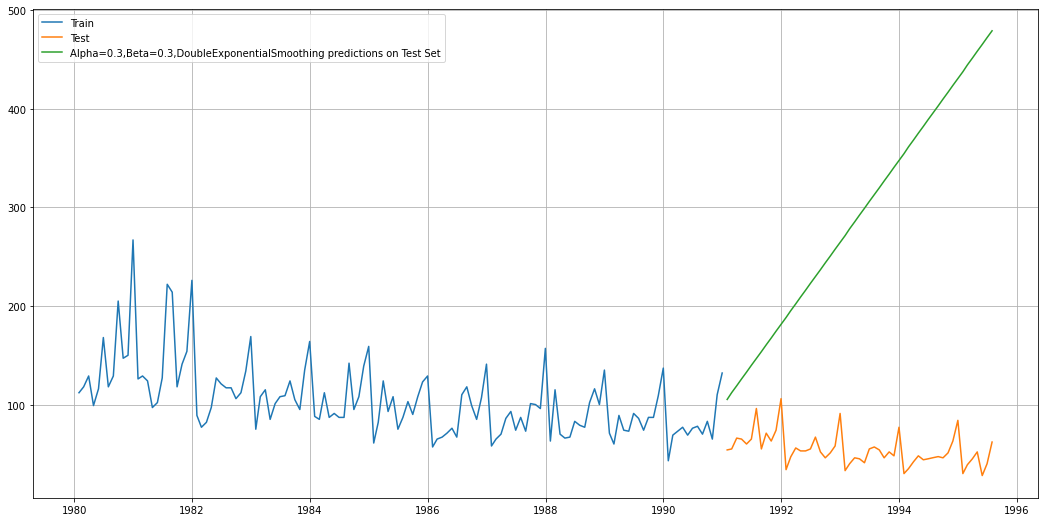
**Model 5: Simple Exponential Smoothing**

****

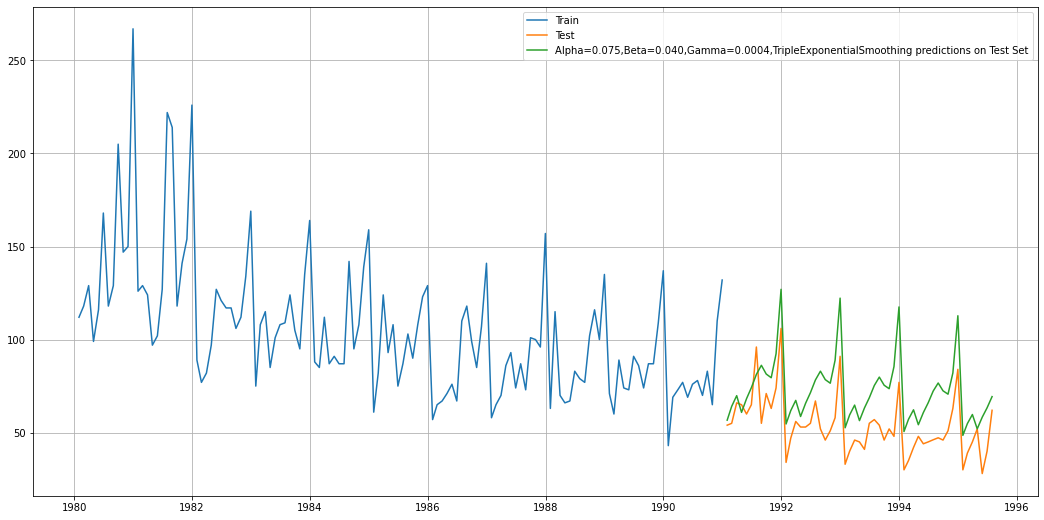
****

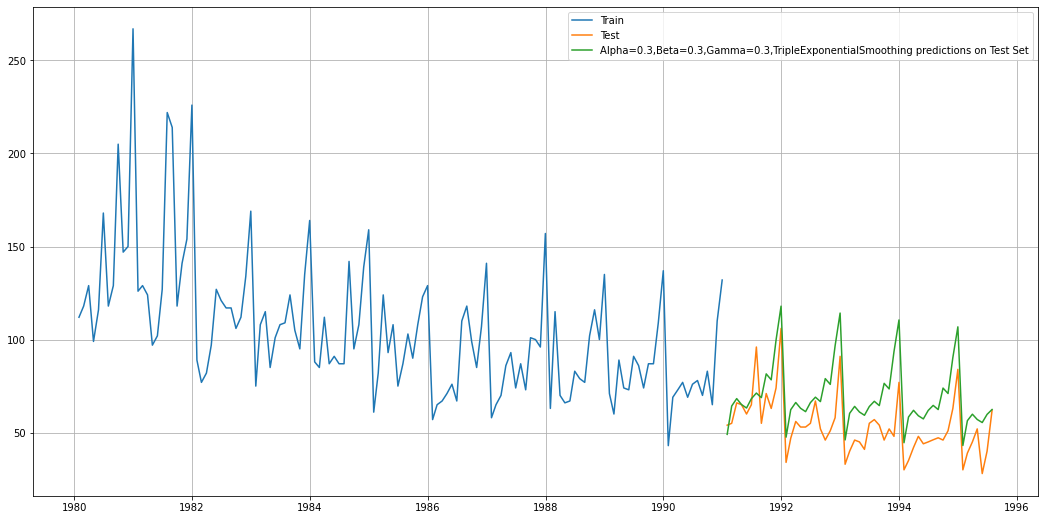
For Alpha =0.995 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 36.748

**Model 6: Double Exponential Smoothing (Holt’s Model)**

****

**Model 7: Triple Exponential Smoothing (Holt-Winter’s Model)**

****

****

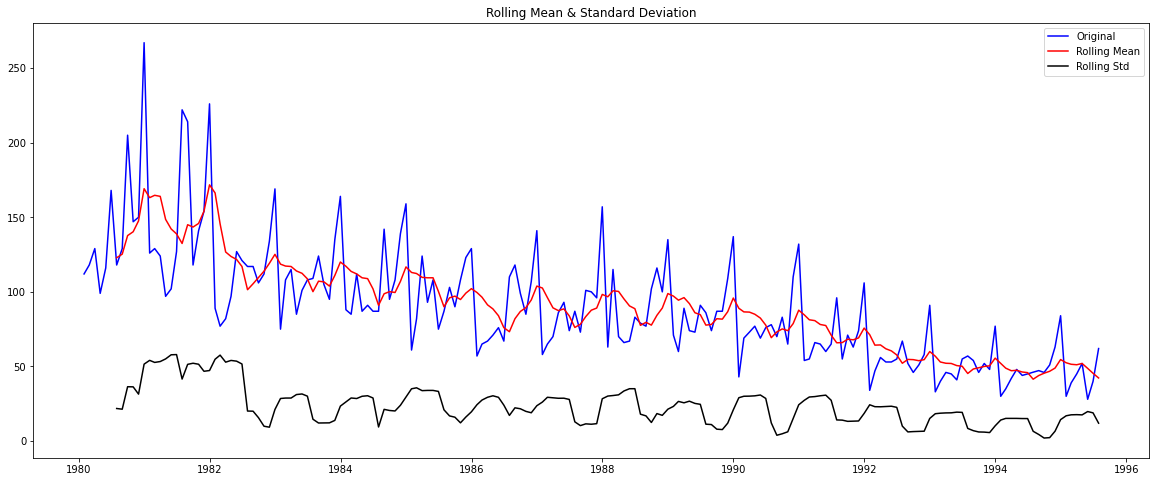
For Alpha=0.075,Beta=0.040,Gamma=0.0004, Triple Exponential Smoothing Model forecast on the Test Data, RMSE is 20.960

* 1. **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.**

**Dickey-Fuller Test**

Null Hypothesis H0 – Series is not Stationary.

Alternative Hypothesis H1 – Series is Stationary.

****

Results of Dickey-Fuller Test:

Test Statistic -1.880931

p-value 0.341084

#Lags Used 13.000000

Number of Observations Used 173.000000

Critical Value (1%) -3.468726

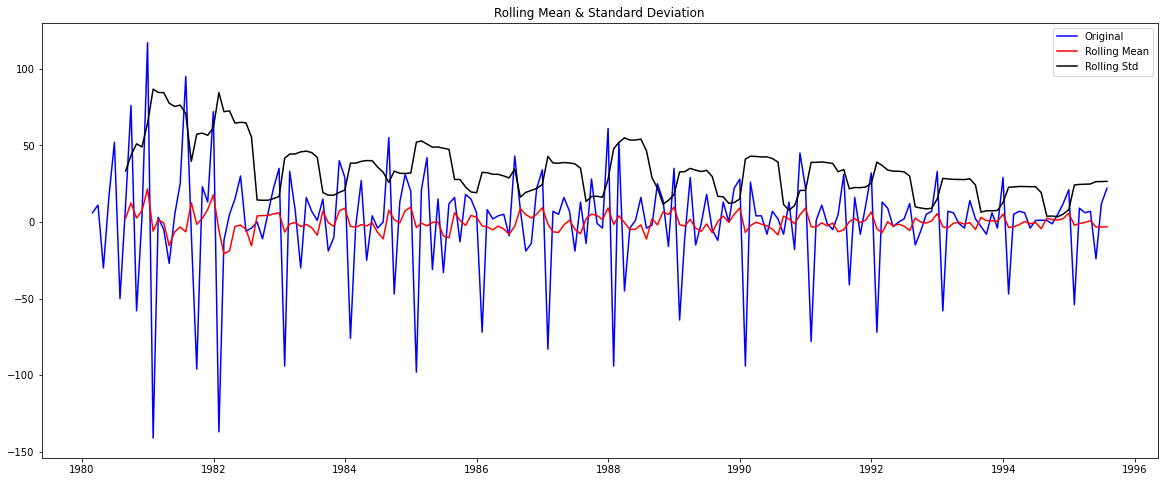
Critical Value (5%) -2.878396

Critical Value (10%) -2.575756

dtype: float64

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

Let us take a difference of order 1 and check if the Time Series is stationary or not.

****

Results of Dickey-Fuller Test:

Test Statistic -8.044820e+00

p-value 1.806363e-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

We see that after taking a difference of order 1 the series has become stationary at alpha = 0.05.

* 1. **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.**

**Automated ARIMA Model**

The following loop helps us in getting a combination of different parameters of p and q in the range of 0 and 2.

We have kept the value of d as 1 as we need to take a difference of the series to make it stationary.

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)

Model calculated for different p and q values and sorted with lowest AIC values:

| **param** | **AIC** |
| --- | --- |
| **2** | (0, 1, 2) | 1276.835373 |
| **5** | (1, 1, 2) | 1277.359228 |
| **4** | (1, 1, 1) | 1277.775754 |
| **7** | (2, 1, 1) | 1279.045689 |
| **8** | (2, 1, 2) | 1279.298694 |
| **1** | (0, 1, 1) | 1280.726183 |
| **6** | (2, 1, 0) | 1300.609261 |
| **3** | (1, 1, 0) | 1319.348311 |
| **0** | (0, 1, 0) | 1335.152658 |

ARIMA Model Results

==============================================================================

Dep. Variable: D.Rose No. Observations: 131

Model: ARIMA(0, 1, 2) Log Likelihood -634.418

Method: css-mle S.D. of innovations 30.167

Date: Sun, 20 Feb 2022 AIC 1276.835

Time: 05:41:49 BIC 1288.336

Sample: 02-29-1980 HQIC 1281.509

- 12-31-1990

================================================================================

coef std err z P>|z| [0.025 0.975]

--------------------------------------------------------------------------------

const -0.4886 0.085 -5.742 0.000 -0.655 -0.322

ma.L1.D.Rose -0.7601 0.101 -7.499 0.000 -0.959 -0.561

ma.L2.D.Rose -0.2398 0.095 -2.518 0.012 -0.427 -0.053

Roots

=============================================================================

Real Imaginary Modulus Frequency

-----------------------------------------------------------------------------

MA.1 1.0001 +0.0000j 1.0001 0.0000

MA.2 -4.1695 +0.0000j 4.1695 0.5000

-----------------------------------------------------------------------------

Predicting on Test set using this model and evaluating the model:

RMSE: 15.60419552386519

**Automated SARIMA Model**

The following loop helps us in getting a combination of different parameters of p, q, P, Q in the range of 0 and 3.

We have kept the value of d in the range (1,2) and D in the range (0,1).

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)

Model calculated for different p, q, d, P, Q, D values and sorted by least AIC values:

| **param** | **seasonal** | **AIC** |
| --- | --- | --- |
| **26** | (0, 1, 2) | (2, 0, 2, 12) | 887.937509 |
| **53** | (1, 1, 2) | (2, 0, 2, 12) | 889.902849 |
| **80** | (2, 1, 2) | (2, 0, 2, 12) | 890.668798 |
| **69** | (2, 1, 1) | (2, 0, 0, 12) | 896.518161 |
| **78** | (2, 1, 2) | (2, 0, 0, 12) | 897.346444 |

SARIMAX Results

==========================================================================================

Dep. Variable: y No. Observations: 132

Model: SARIMAX(0, 1, 2)x(2, 0, 2, 12) Log Likelihood -436.969

Date: Sun, 20 Feb 2022 AIC 887.938

Time: 05:43:54 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 189.892 -0.004 0.996 -373.024 371.339

ma.L2 -0.1573 29.833 -0.005 0.996 -58.629 58.314

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

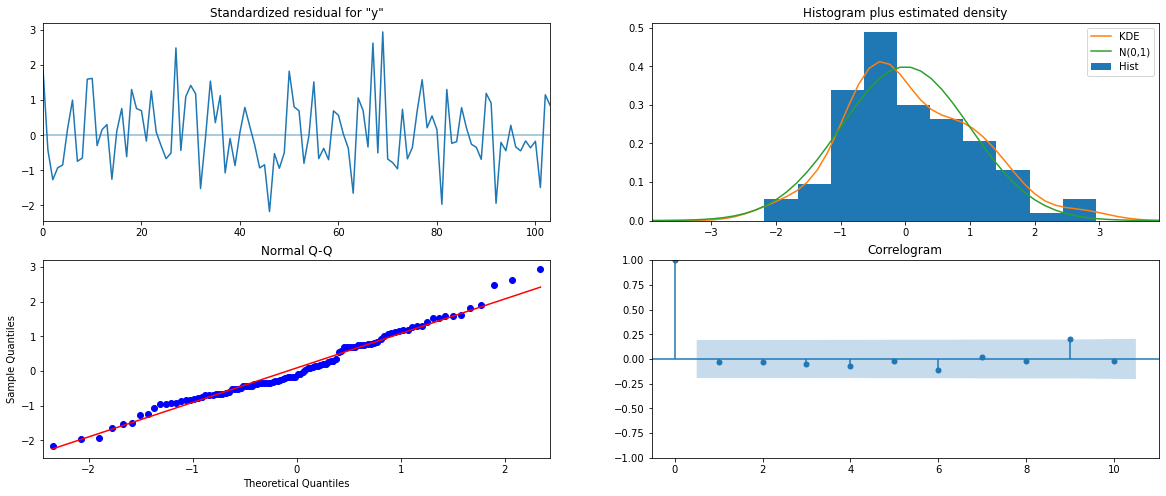
Prob(Q): 0.75 Prob(JB): 0.31

Heteroskedasticity (H): 0.88 Skew: 0.37

Prob(H) (two-sided): 0.70 Kurtosis: 3.03

===================================================================================

**Diagnostic Plot:**

****

Predicting on Test set using this model and evaluating the model:

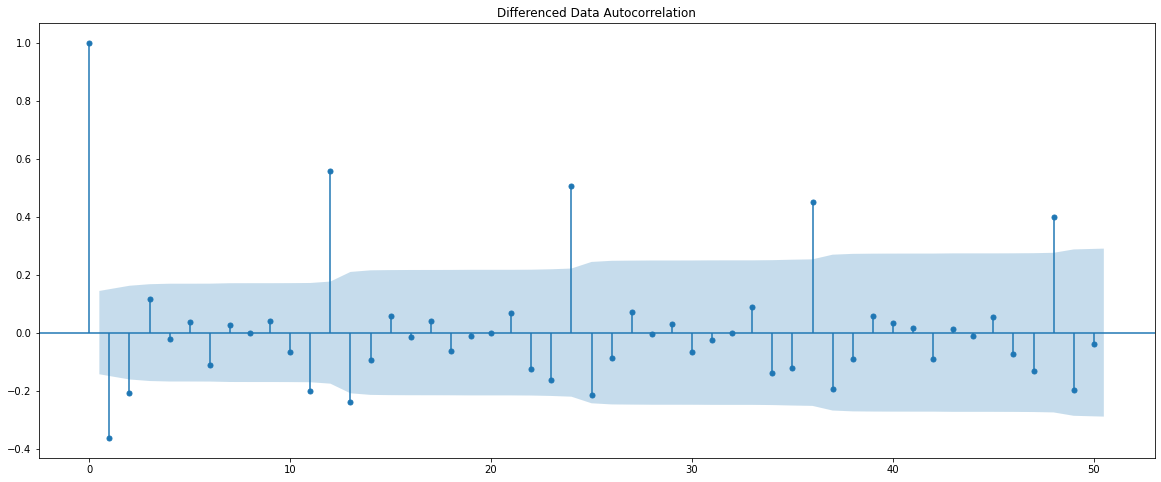
Forecasting the Model:

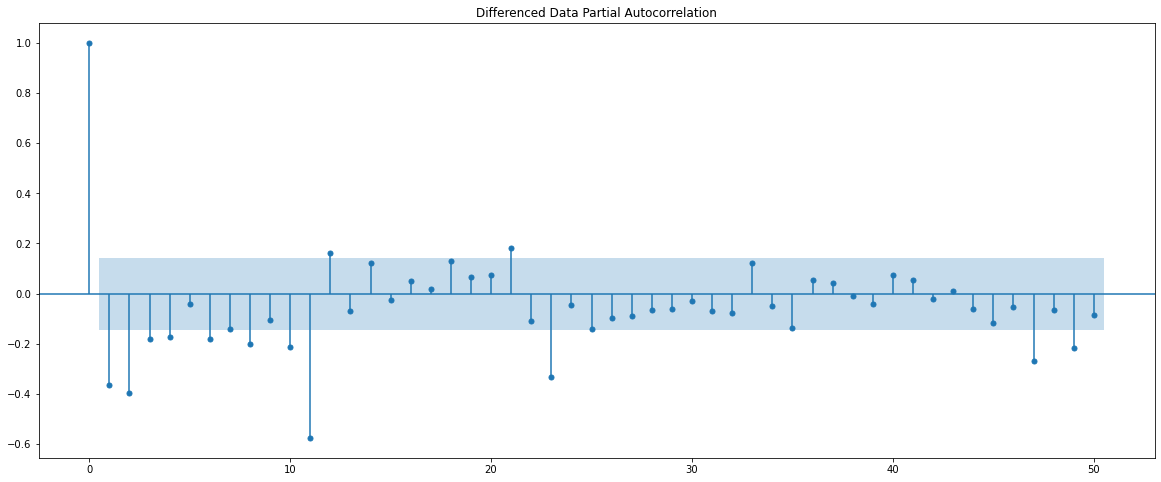
| **y** | **mean** | **mean\_se** | **mean\_ci\_lower** | **mean\_ci\_upper** |
| --- | --- | --- | --- | --- |
| **0** | 62.867264 | 15.928501 | 31.647976 | 94.086552 |
| **1** | 70.541190 | 16.147659 | 38.892361 | 102.190020 |
| **2** | 77.356411 | 16.147656 | 45.707586 | 109.005236 |
| **3** | 76.208814 | 16.147656 | 44.559989 | 107.857639 |
| **4** | 72.747398 | 16.147656 | 41.098573 | 104.396223 |

RMSE: 26.8806254304391

* 1. **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.**

**ARIMA Model**

****

****

Here, we have taken alpha = 0.05.

The Auto-Regressive parameter in an ARIMA model is ‘p’ which comes from the significant lag before which the PACF plot cuts off to 0.

The Moving Average parameter in an ARIMA model is ‘q’ which comes from the significant lag before which the ACF plot cuts off to 0.

By observing the plots we can see that the PACF and the ACF plot cuts off at lag 4 and 2.

ARIMA Model Results

==============================================================================

Dep. Variable: D.Rose No. Observations: 131

Model: ARIMA(4, 1, 2) Log Likelihood -633.876

Method: css-mle S.D. of innovations 29.793

Date: Sun, 20 Feb 2022 AIC 1283.753

Time: 05:44:50 BIC 1306.754

Sample: 02-29-1980 HQIC 1293.099

- 12-31-1990

================================================================================

coef std err z P>|z| [0.025 0.975]

--------------------------------------------------------------------------------

const -0.1905 0.576 -0.331 0.741 -1.319 0.938

ar.L1.D.Rose 1.1685 0.087 13.391 0.000 0.997 1.340

ar.L2.D.Rose -0.3562 0.132 -2.692 0.007 -0.616 -0.097

ar.L3.D.Rose 0.1855 0.132 1.402 0.161 -0.074 0.445

ar.L4.D.Rose -0.2227 0.091 -2.443 0.015 -0.401 -0.044

ma.L1.D.Rose -1.9506 nan nan nan nan nan

ma.L2.D.Rose 1.0000 nan nan nan nan nan

Roots

=============================================================================

Real Imaginary Modulus Frequency

-----------------------------------------------------------------------------

AR.1 1.1027 -0.4115j 1.1770 -0.0569

AR.2 1.1027 +0.4115j 1.1770 0.0569

AR.3 -0.6863 -1.6644j 1.8003 -0.3122

AR.4 -0.6863 +1.6644j 1.8003 0.3122

MA.1 0.9753 -0.2209j 1.0000 -0.0355

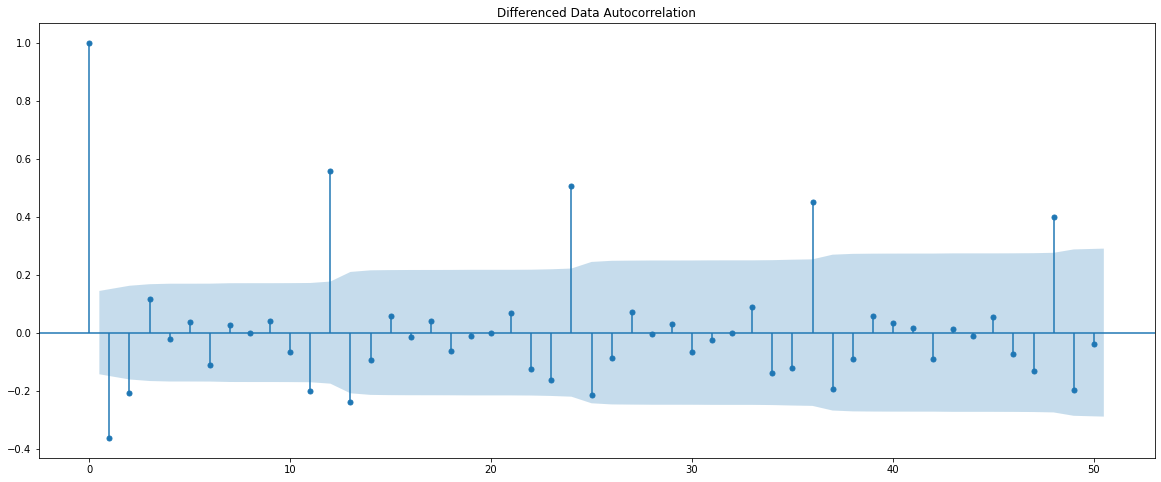
MA.2 0.9753 +0.2209j 1.0000 0.0355

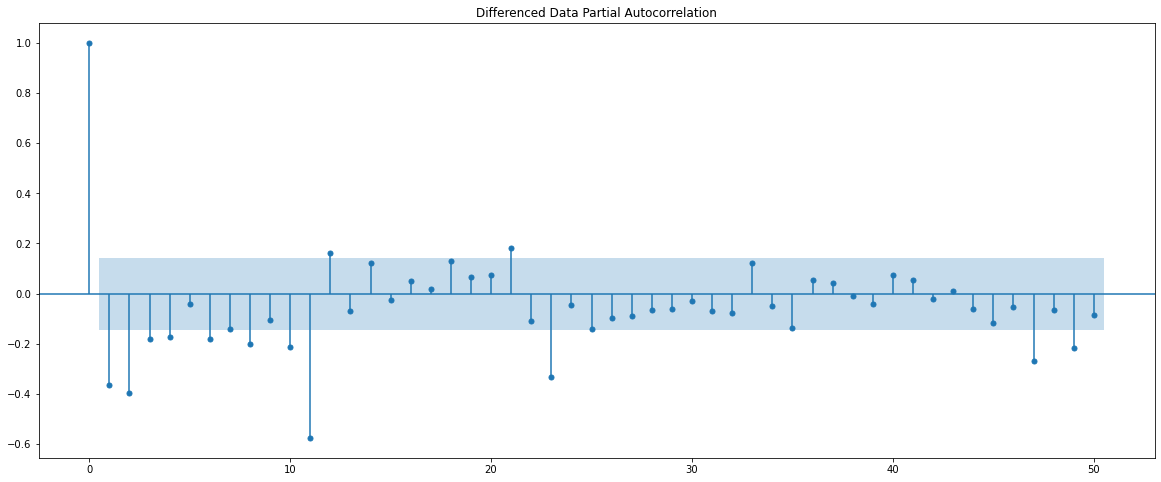
-----------------------------------------------------------------------------

Predicting on Test set using this model and evaluating the model:

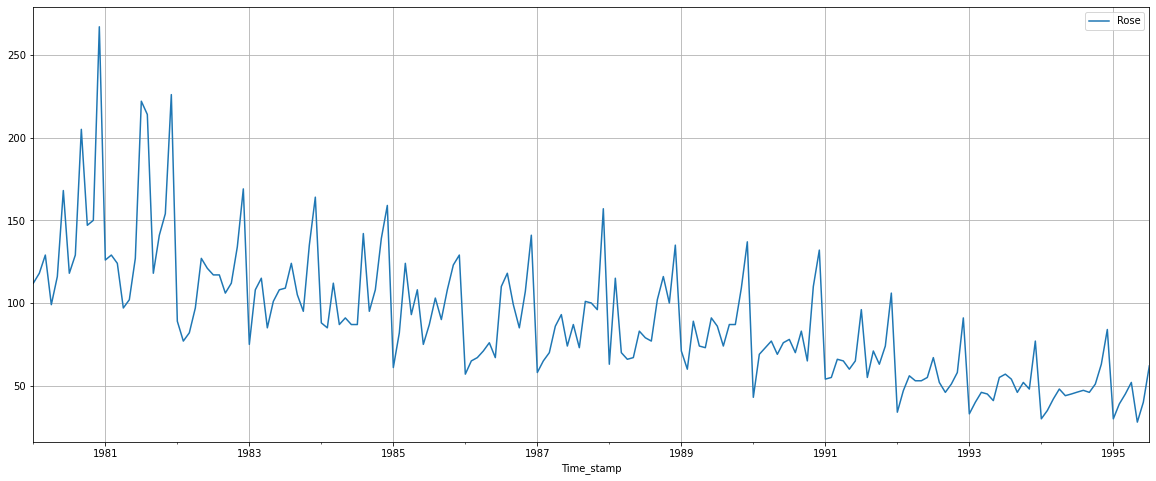
RMSE: 33.90482729155276

**SARIMA Model**

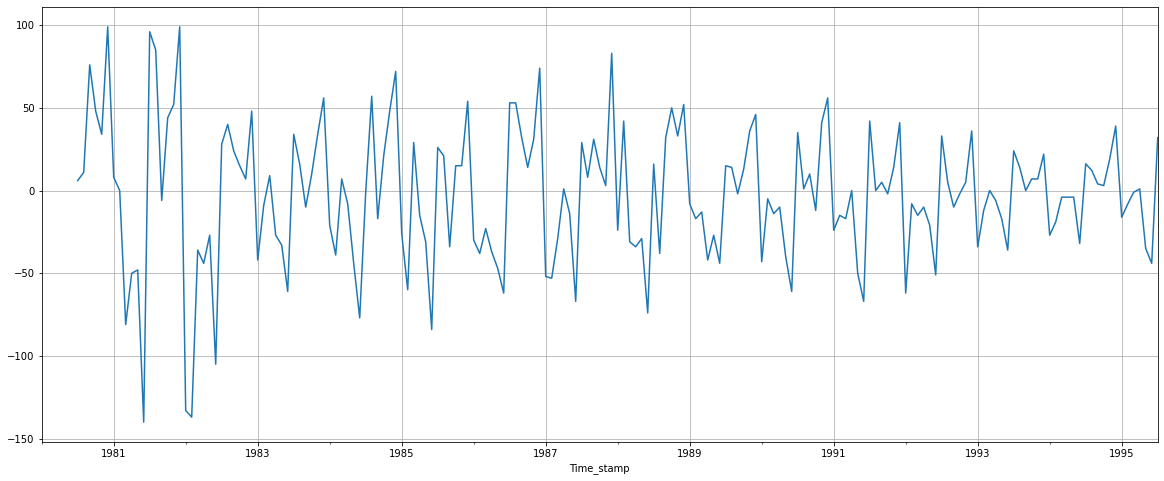
****

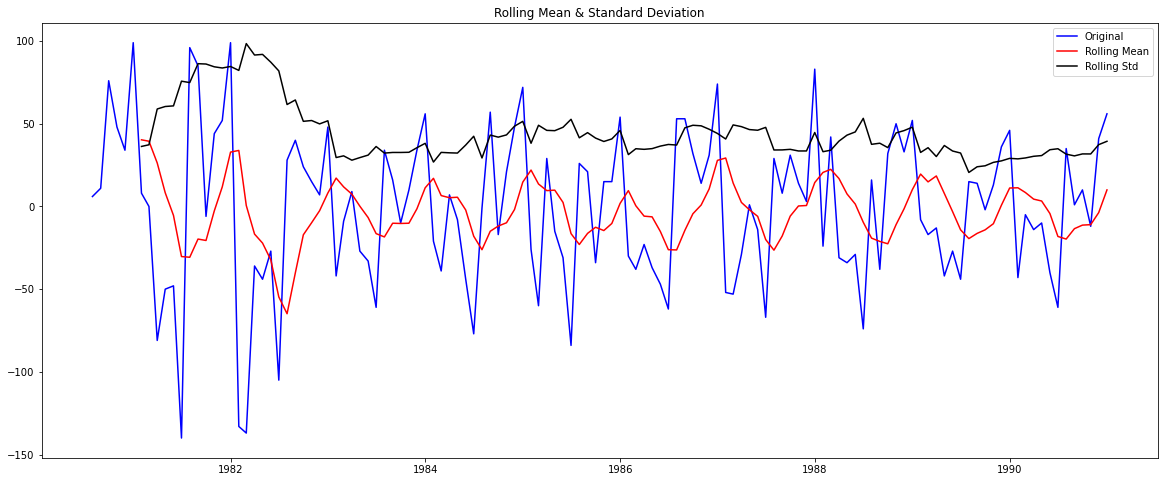
****

**Plotting the Dataset:**

****

We can see that there is seasonality. So, we take a seasonal differencing and then check the series.

****

****

Results of Dickey-Fuller Test:

Test Statistic -7.442449e+00

p-value 5.956534e-11

#Lags Used 7.000000e+00

Number of Observations Used 1.180000e+02

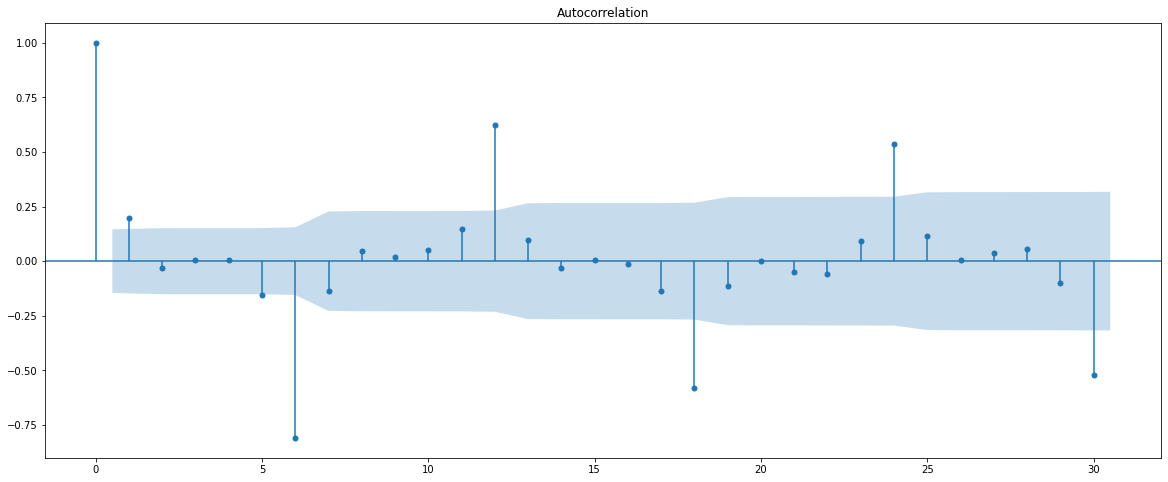
Critical Value (1%) -3.487022e+00

Critical Value (5%) -2.886363e+00

Critical Value (10%) -2.580009e+00

dtype: float64

**ACF and PACF plot after differencing:**

****

****

Here, we have taken alpha = 0.05.

We are going to take the seasonal period as 6. We will keep the p(1) and q(1) parameters the same as the ARIMA model.

The Auto-Regressive parameter in an ARIMA model is ‘P’ which comes from the significant lag after which the PACF plot cuts off to 0.

The Moving Average parameter in a SARIMA model is ‘Q’ which comes from the significant lag after which the ACF plot cuts off to 0.

Since 6 is the seasonal period, we must remember to check the ACF and PACF plots only at multiples of 6.

By observing the plots we can see that the ACF and the PACF cut off at 1 and 1.

SARIMAX Results

===========================================================================================

Dep. Variable: y No. Observations: 132

Model: SARIMAX(4, 1, 2)x(1, 1, [1], 6) Log Likelihood -526.344

Date: Sun, 20 Feb 2022 AIC 1070.688

Time: 05:46:22 BIC 1095.393

Sample: 0 HQIC 1080.716

- 132

Covariance Type: opg

==============================================================================

coef std err z P>|z| [0.025 0.975]

------------------------------------------------------------------------------

ar.L1 -0.0722 0.270 -0.267 0.789 -0.601 0.457

ar.L2 -0.4955 0.141 -3.515 0.000 -0.772 -0.219

ar.L3 -0.0486 0.121 -0.401 0.688 -0.286 0.189

ar.L4 -0.3135 0.107 -2.926 0.003 -0.524 -0.104

ma.L1 87.5064 3459.696 0.025 0.980 -6693.373 6868.386

ma.L2 -56.4405 2197.671 -0.026 0.980 -4363.797 4250.916

ar.S.L6 -0.7330 0.066 -11.163 0.000 -0.862 -0.604

ma.S.L6 -1.4121 0.208 -6.779 0.000 -1.820 -1.004

sigma2 0.0341 2.673 0.013 0.990 -5.206 5.274

===================================================================================

Ljung-Box (L1) (Q): 0.07 Jarque-Bera (JB): 23.99

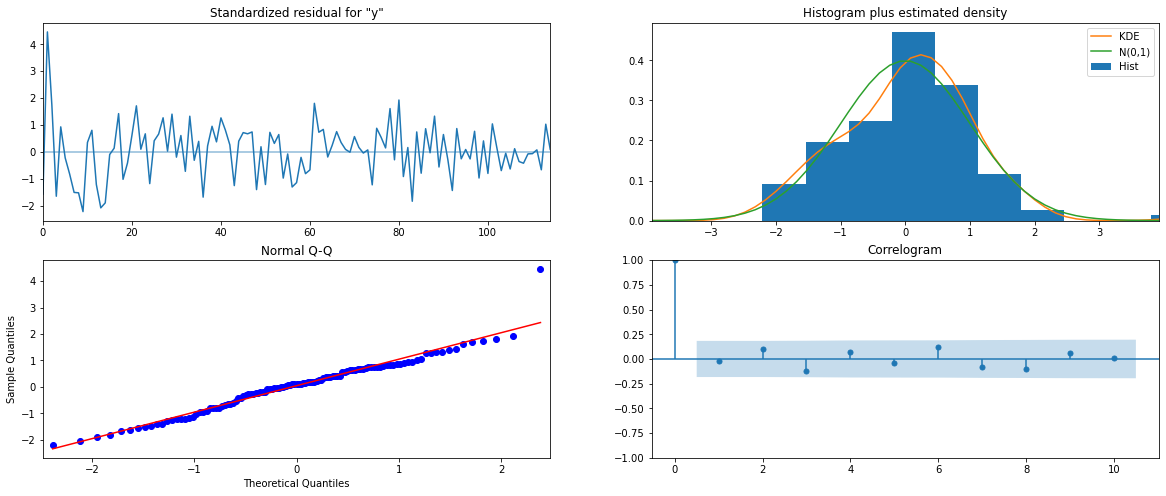
Prob(Q): 0.79 Prob(JB): 0.00

Heteroskedasticity (H): 0.36 Skew: 0.44

Prob(H) (two-sided): 0.00 Kurtosis: 5.06

===================================================================================

**Diagnostic Plot:**

****

Predicting on Test set using this model and evaluating the model:

Forecasting Test Set with confidence interval:

| **y** | **mean** | **mean\_se** | **mean\_ci\_lower** | **mean\_ci\_upper** |
| --- | --- | --- | --- | --- |
| **0** | 43.681048 | 22.984050 | -1.366863 | 88.728958 |
| **1** | 66.918041 | 23.988326 | 19.901786 | 113.934296 |
| **2** | 71.109578 | 24.245548 | 23.589177 | 118.629979 |
| **3** | 78.880894 | 24.588962 | 30.687415 | 127.074373 |
| **4** | 67.838140 | 24.693580 | 19.439612 | 116.236668 |

RMSE: 19.004805582081588

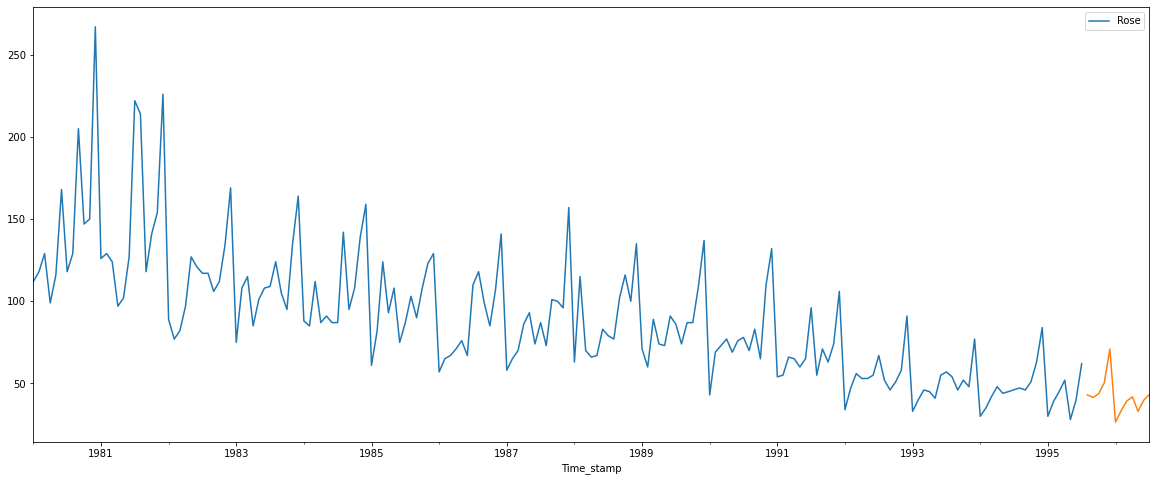
* 1. **Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.**

| **Test RMSE** |
| --- |
| **Alpha=0.3,Beta=0.3,Gamma=0.3,TripleExponentialSmoothing** | 10.935749 |
| **2pointTrailingMovingAverage** | 11.529994 |
| **4pointTrailingMovingAverage** | 14.444342 |
| **6pointTrailingMovingAverage** | 14.554944 |
| **9pointTrailingMovingAverage** | 14.721499 |
| **RegressionOnTime** | 15.255435 |
| **ARIMA(0,1,2)** | 15.604196 |
| **SARIMA(4,1,2)(1,1,1,6) based on ACF & PACF** | 19.004806 |
| **Alpha=0.075,Beta=0.040,Gamma=0.0004,TripleExponentialSmoothing** | 20.959970 |
| **SARIMA(0,1,2)(2,0,2,12)** | 26.880625 |
| **ARIMA (4,1,2) based on ACF & PACF** | 33.904827 |
| **Alpha=0.995,SimpleExponentialSmoothing** | 36.748147 |
| **Alpha=0.3,SimpleExponentialSmoothing** | 47.457057 |
| **SimpleAverageModel** | 53.413057 |
| **NaiveModel** | 79.672238 |
| **Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing** | 265.509912 |

* 1. **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.**

The best model would be the Triple Exponential Smoothing Model with multiplicative seasonality, with the parameters alpha = 0.3, beta = 0.3 and gamma = 0.3.

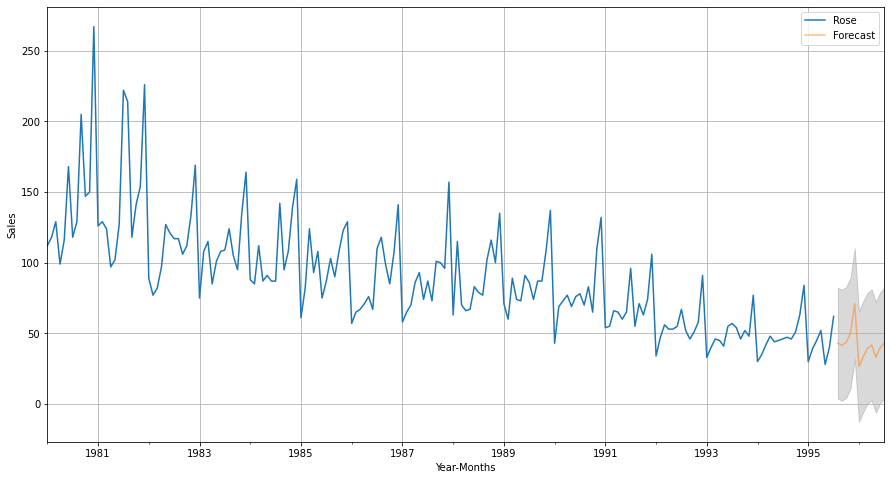
RMSE: 19.95.



We calculated the upper and lower confidence bands at 95% confidence level.

| **lower\_CI** | **prediction** | **upper\_ci** |
| --- | --- | --- |
| **1995-08-31** | 3.801628 | 43.008486 | 82.215344 |
| **1995-09-30** | 2.303920 | 41.510777 | 80.717635 |
| **1995-10-31** | 4.472139 | 43.678997 | 82.885855 |
| **1995-11-30** | 11.061360 | 50.268218 | 89.475076 |
| **1995-12-31** | 31.744941 | 70.951799 | 110.158657 |

Plotting the forecast along with the confidence band:



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

**Prediction:**

1995-08-31 43.008486

1995-09-30 41.510777

1995-10-31 43.678997

1995-11-30 50.268218

1995-12-31 70.951799

1996-01-31 26.448111

1996-02-29 33.419382

1996-03-31 39.277687

1996-04-30 41.829249

1996-05-31 33.017517

1996-06-30 39.751866

1996-07-31 43.310089

Freq: M, dtype: float64

**Describe Prediction:**

count 12.000000

mean 42.206015

std 10.956609

min 26.448111

25% 37.813111

50% 41.670013

75% 43.402316

max 70.951799

dtype: float64

. . .

* Sales of Rose Wine from 1980 to 1995 were analysed. Performing EDA gave us further insights.
* From 1981, Sales are found to be decreasing steadily till 1995.
* Median: 85.
* The highest sale occurred in 1980, whereas the lowest occurred in 1995.
* The Monthly plot indicates that sales jump up every November and December. One possible reason for this could be customers purchasing wine for the occasion of Christmas and New Year’s. Stocks have to be increased during this holiday season.
* Subsequently, January has recorded the lowest sales which is right after the December month of previous years.
* Trend and Seasonality are present in the dataset.
* Dataset was split for training and test set. Various models such as Linear Regression, Naïve Bayes, Simple Exponential Smoothing, Double Exponential Smoothing, Triple Exponential Smoothing, ARIMA and SARIMA built on the training data and tested on the test data.
* SARIMA model would be the best suited model for the data, due to the dataset containing seasonality. However, from RMSE values of various models were found and Triple Exponential Smoothing is determined to be the best for the dataset and so, it was applied on full data.
* Sales for the next 12 months are predicted with confidence interval. Sales are still varying heavily across months, and there is a downward trend.
* It is advised to conduct surveys about the wine and action to be taken accordingly.
* Quality of the wine needs improvement to bring sales back up.

**THE END**