Project Time series forecasting

Parnoshree Chatterjee

A. Sparkling.csv Dataset:

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

Answer:

Table 1. Head 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 2. Tail of the dataset

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

- In the first question we are reading the date set and checking the head and tail.
- It is to understand what is the start and end date of the time series.
- The above Sparkling data set starts from 1980-1991.
- We have two variables with name YearMonth and Sparkling
- Sparkling is the target variable.
- As we are trying to find the sales for Sparkling wine it is important to check and understand what are the variables, we will be working with

Table 3. Creating the Time Stamps and adding to the data frame to make it a Time Series Data

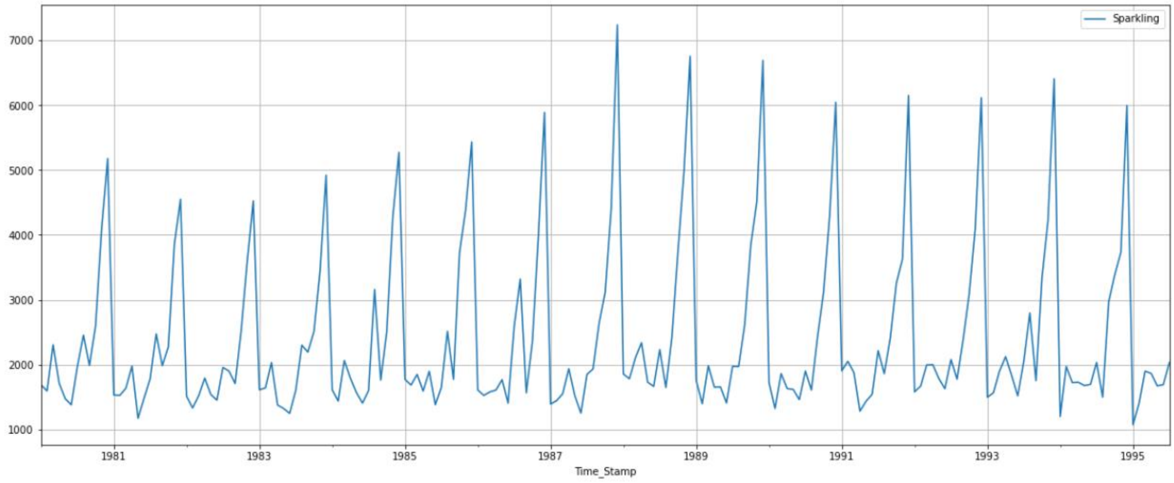
```
DatetimeIndex(['1980-01-31', '1980-02-29', '1980-03-31', '1980-04-30',
               '1980-05-31', '1980-06-30', '1980-07-31', '1980-08-31',
               '1980-09-30', '1980-10-31',
               ...
               '1994-10-31', '1994-11-30', '1994-12-31', '1995-01-31',
               '1995-02-28', '1995-03-31', '1995-04-30', '1995-05-31',
               '1995-06-30', '1995-07-31'],
              dtype='datetime64[ns]', length=187, freq='M')
```

Table 4. Adding Time stamp

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

Fig 1. Plotting the same graph from the second data frame with the date-time modifications.



- We are creating the time stamps and adding to the data frame to make it a time series data.
- As we know in order to find the seasonality, trend and Residual we have to create the data with time stamps.
- We have added time stamp as one of the variables.
- So we have two variables now Time stamp and Sparkling.
- We have plotted a graph above which talks about the second data frame with date and time modifications.

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

Answer:

- First we check the table for monthly sparks throughout the year using monthly data.
- There are no duplicates in the data set.
- We don't have null values either.
- We describe the function to check the description of the data set.
- We have standard deviation of 1295.11, mean of 2402.02

Table 5. For monthly Sparks across years

Time_Stamp	April	August	December	February	January	July	June	March	May	November	October	September
Time_Stamp												
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

There are no duplicates or null values in the data set.

The shape of the data set is 187,1

Table 6 . Describe function

	Sparkling
count	187.000
mean	2402.417
std	1295.112
min	1070.000
25%	1605.000
50%	1874.000
75%	2549.000
max	7242.000

Fig 2. Plot for Monthly Sparks throughout year

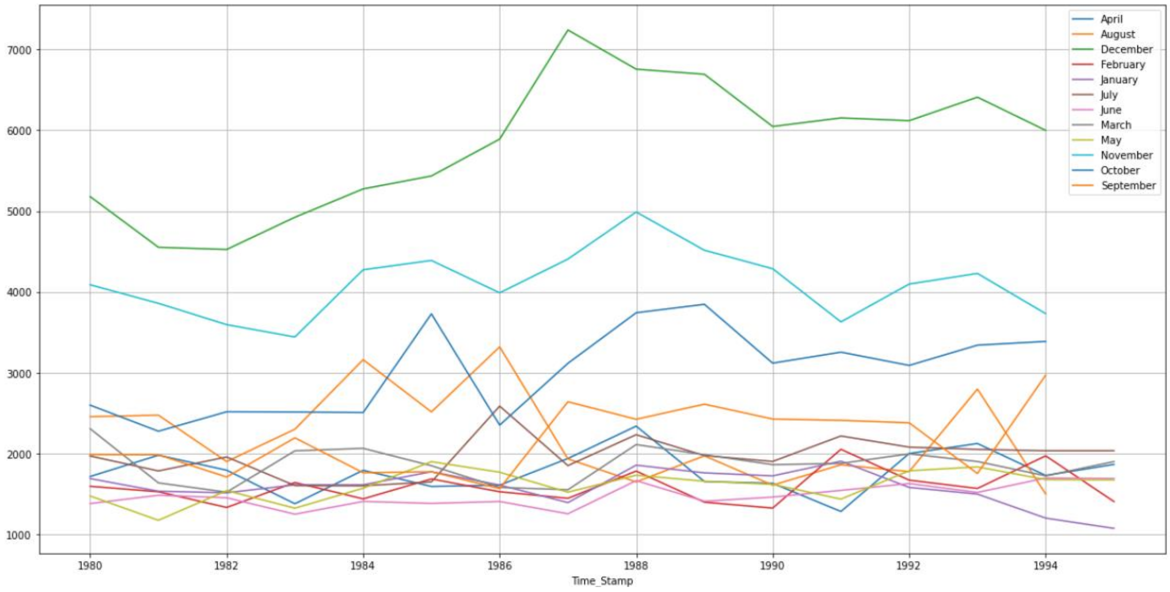


Fig 3. Plot for Yearly box plot

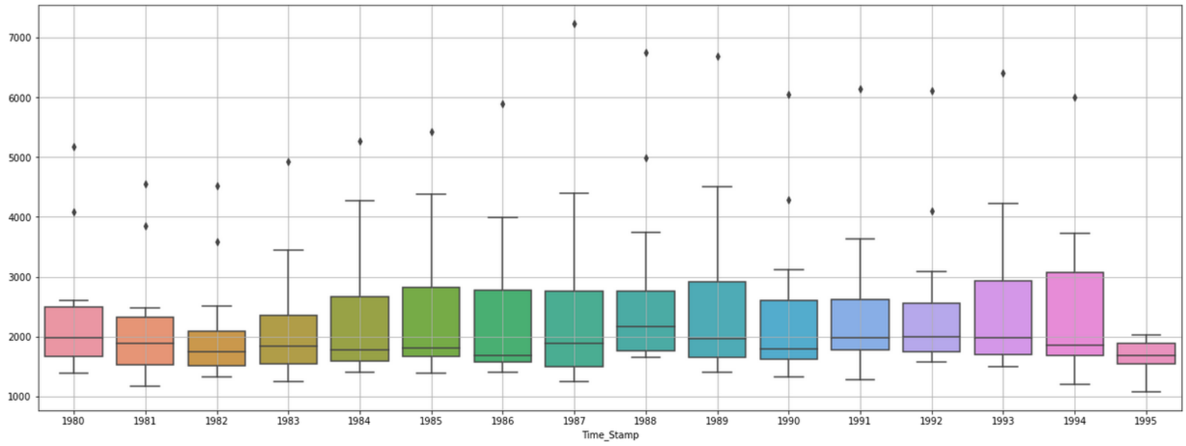


Fig 4. Plot for Monthly Boxplot

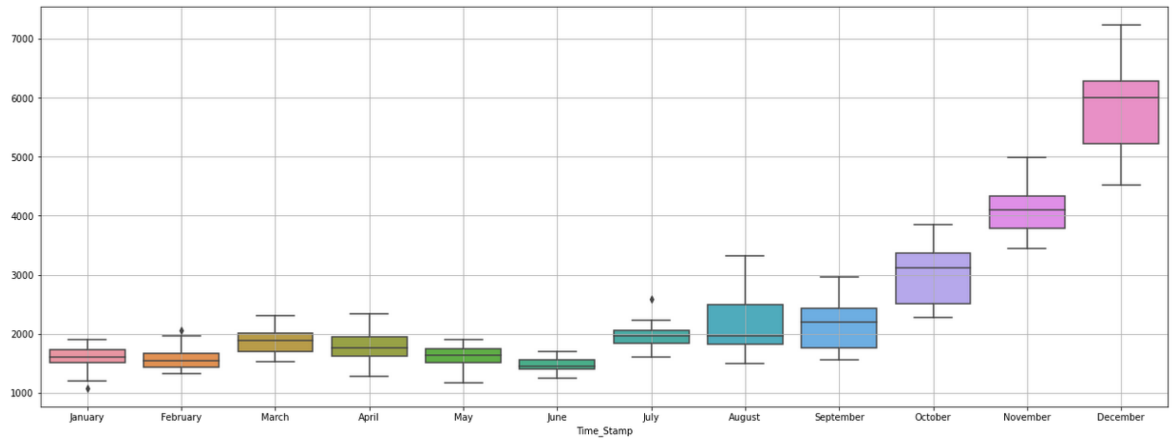
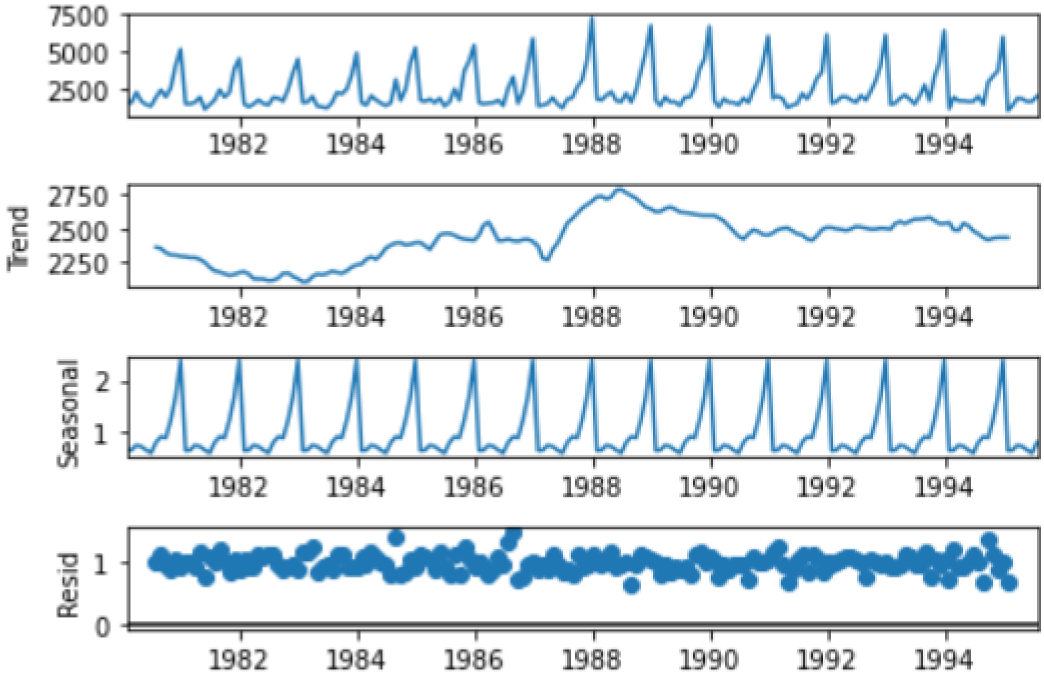


Fig 5. Plot for Decomposition



Decomposition Trend, Seasonality and Residual:

- Next we plot a graph to check monthly and yearly sparks.
- We have also used box plot to show the monthly and yearly sparks.
- As we see there are no outliers in the data set.
- We have plotted a decomposition graph.
- A **decomposition** of a **graph** is a collection of edge-disjoint subgraphs of such that every edge of belongs to exactly one . If each is a path or a cycle in , then is called a path **decomposition** of . If each is a path in , then is called an acyclic path **decomposition.**
- We can see the trend, seasonality and Residual in the above graph of Fig 5.
- We can also see the values for trend, seasonality and Residual in Table 7.

Table 7. Table showing Trend, Seasonality and Residual

```
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
1980-11-30    2302.041667
1980-12-31    2293.791667
Name: trend, dtype: float64

Seasonality
 Time_Stamp
1980-01-31    0.649843
1980-02-29    0.659214
1980-03-31    0.757440
1980-04-30    0.730351
1980-05-31    0.660609
1980-06-30    0.603468
1980-07-31    0.809164
1980-08-31    0.918822
1980-09-30    0.894367
1980-10-31    1.241789
1980-11-30    1.690158
1980-12-31    2.384776
Name: seasonal, dtype: float

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    1.029230
1980-08-31    1.135407
1980-09-30    0.955954
1980-10-31    0.907513
1980-11-30    1.050423
1980-12-31    0.946770
Name: resid, dtype: float64
```

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

Answer:

Shape of train and test data set is

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

Table 8. First few and Last few rows of training and testing dataset:

```
First few rows of Training Data
            Sparkling
Time_Stamp
1980-01-31       1686
1980-02-29       1591
1980-03-31       2304
1980-04-30       1712
1980-05-31       1471

Last few rows of Training Data
            Sparkling
Time_Stamp
1990-08-31       1605
1990-09-30       2424
1990-10-31       3116
1990-11-30       4286
1990-12-31       6047

First few rows of Test Data
            Sparkling
Time_Stamp
1991-01-31       1902
1991-02-28       2049
1991-03-31       1874
1991-04-30       1279
1991-05-31       1432

Last few rows of Test Data
            Sparkling
Time_Stamp
1995-03-31       1897
1995-04-30       1862
1995-05-31       1670
1995-06-30       1688
1995-07-31       2031
```

- We have split the data into Train and Test the above table 8 show the first and last few rows of train and test data.
- The shape of the dataset has changed to 132,1 and 55,1
- We have made sure the split for test data starts from 1991 as per the requirements.

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

Model 1: Linear Regression

Table 9. For Training Time Instance and Test Time Instance

```
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, 3
3, 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, 12
0, 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, 1
57, 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]
```

- Here our first model is Linear regression I will be briefly explaining what are the steps used during the modelling process.
- In statistics, linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression
- The above table 9 shows us the training and testing time instance.
- The below table 10 talks about the first and last few rows of training and testing data set. It is important to check the values to understand the difference made to the dataset.

Table 10. First few and Last few rows of training and testing dataset:

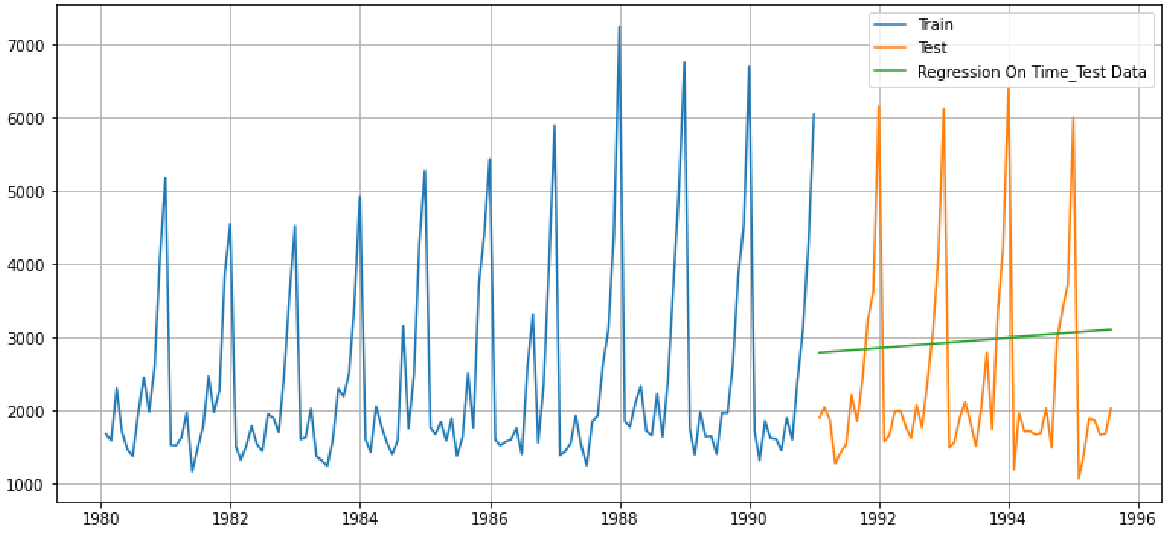
```
First few rows of Training Data
             Sparkling  time
Time_Stamp
1980-01-31       1686     1
1980-02-29       1591     2
1980-03-31       2304     3
1980-04-30       1712     4
1980-05-31       1471     5

Last few rows of Training Data
             Sparkling  time
Time_Stamp
1990-08-31       1605   128
1990-09-30       2424   129
1990-10-31       3116   130
1990-11-30       4286   131
1990-12-31       6047   132

First few rows of Test Data
             Sparkling  time
Time_Stamp
1991-01-31       1902   133
1991-02-28       2049   134
1991-03-31       1874   135
1991-04-30       1279   136
1991-05-31       1432   137

Last few rows of Test Data
             Sparkling  time
Time_Stamp
1995-03-31       1897   183
1995-04-30       1862   184
1995-05-31       1670   185
1995-06-30       1688   186
1995-07-31       2031   187
```

Fig 6. Test_Predictions_Model1



Test RMSE for Linear Regression

	Test RMSE
RegressionOnTime	1389.135175

Model evaluation using RMSE

```
For RegressionOnTime forecast on the Test Data, RMSE is 1389.135
```

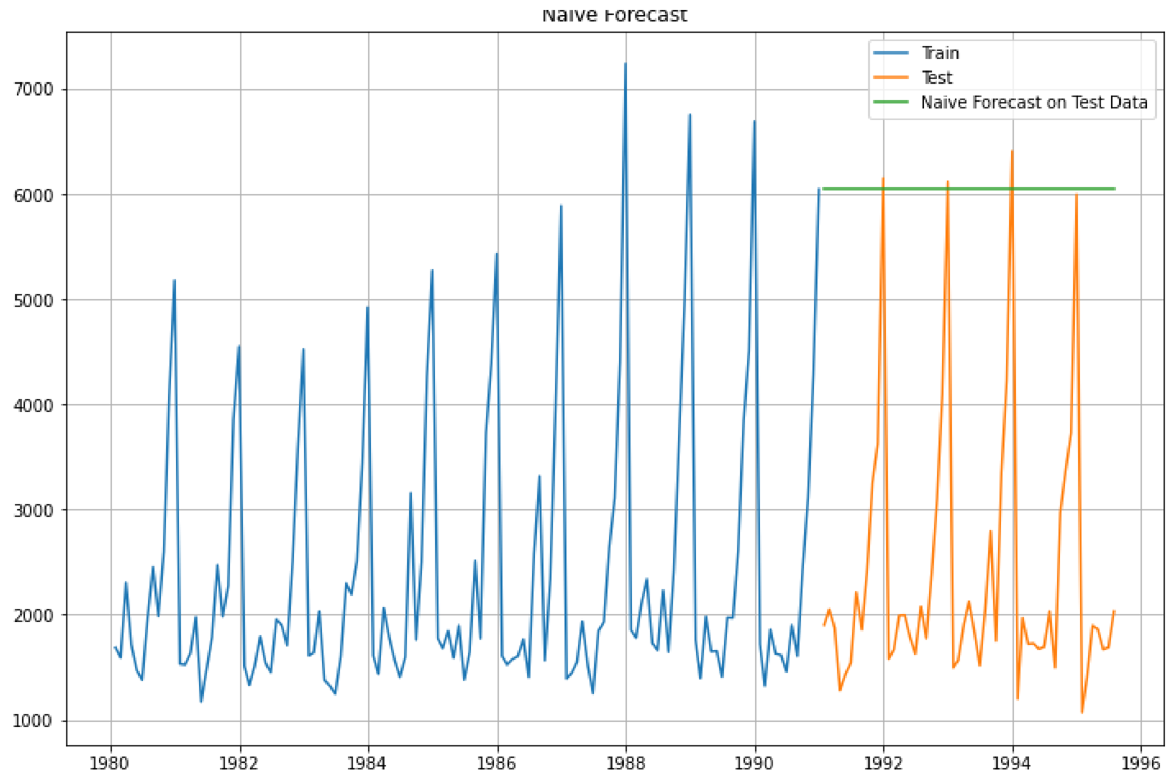
- The above figure 6 show the test prediction mode1.
- We can see the data set has seasonality and trend.
- The root mean square error for the LR model is 1389.13

Model 2: Naive Approach

Head of the data set

```
Time_Stamp
1991-01-31    6047
1991-02-28    6047
1991-03-31    6047
1991-04-30    6047
1991-05-31    6047
Name: naive, dtype: int64
```

***Fig 7.* Plot for Naive Approach**



RMSE score for the first 2 models

	Test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352

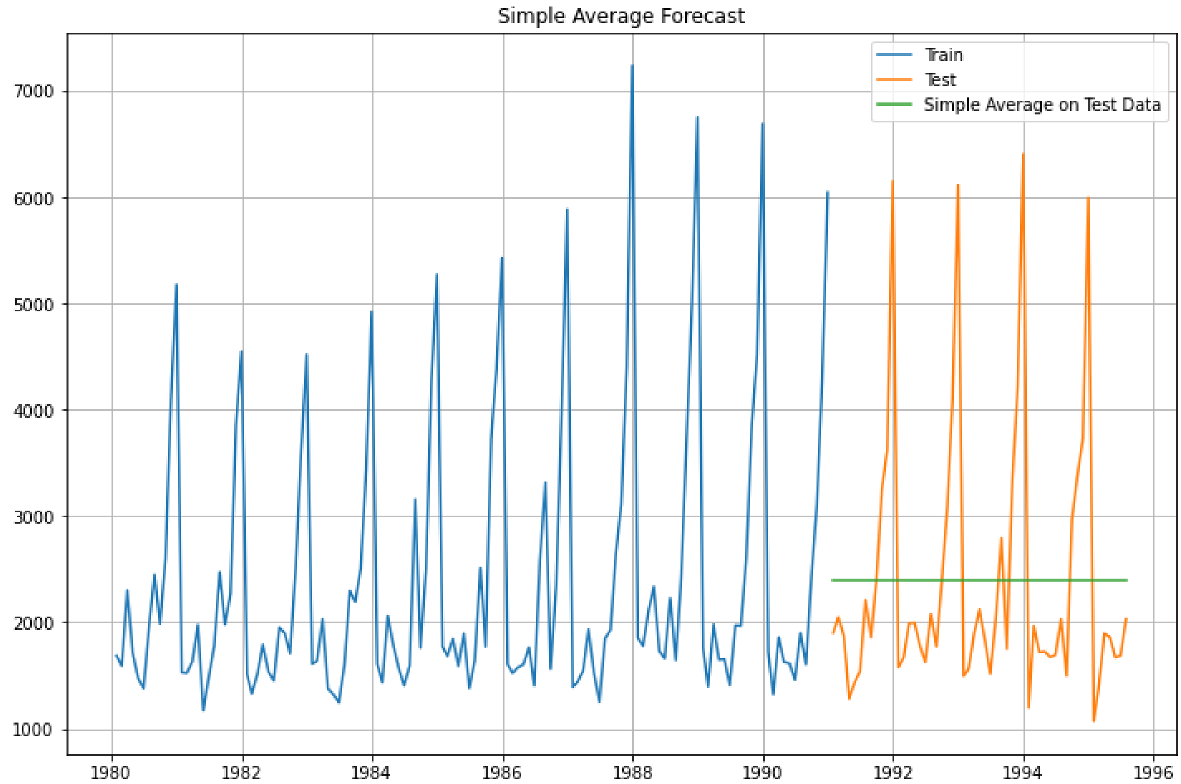
- A **model** in which minimum amounts of effort and manipulation of data are used to prepare a forecast. Most often naïve models used are random walk (current value as a forecast of the next period) and seasonal random walk (value from the same period of prior year as a forecast for the same period of forecasted year.)
- We have read the data head to check the values.
- Fig 7 shows the trend on the test and train data.
- The RMSE for Naïve Model is 3864 which is higher than the LR model.
- We have noted the value for both the models above for a better comparision.

Method 3: Simple Average

Head of the data set

	Sparkling	mean_forecast
Time_Stamp		
1991-01-31	1902	2403.780303
1991-02-28	2049	2403.780303
1991-03-31	1874	2403.780303
1991-04-30	1279	2403.780303
1991-05-31	1432	2403.780303

Fig 8. Plot for Simple Average



Model evaluation using RMSE

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

RMSE for Simple Average:

	Test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804

- The **simple average** of a set of observations is computed as the sum of the individual observations divided by the number of observations in the set.
- We have read the data head to check the values.
- Fig 8. shows the trend on the test and train data.
- The RMSE for Simple average is 1275 which is lower than the LR and Naïve model.
- We have noted the value for all three models above for a better comparison.

Method 4: Moving Average (MA)

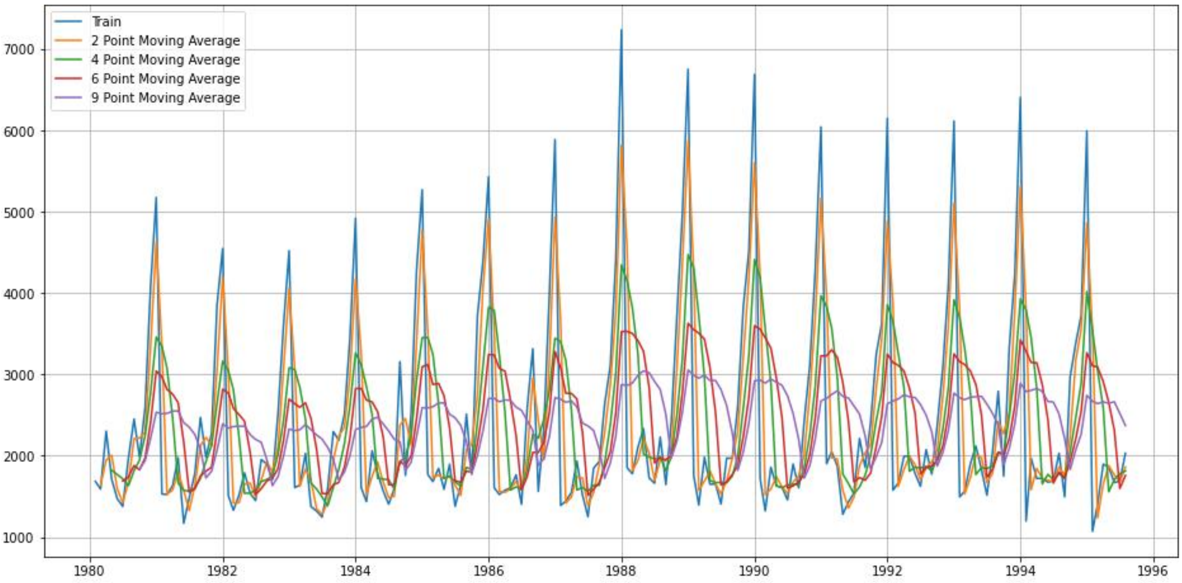
Head of the data set

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

Head after adding Trailing_2 to Trailing_9

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

Fig 9. Plot for Moving Average

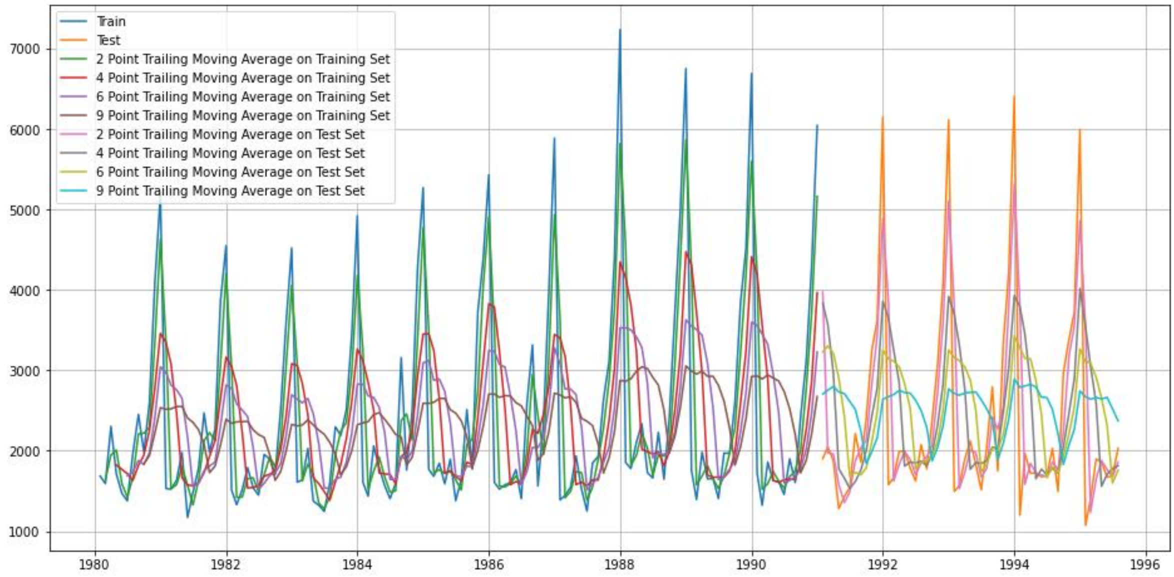


Creating train and test set

- In statistics, a moving average is a calculation to analyze data points by creating a series of averages of different subsets of the full data set. It is also called a moving mean or rolling mean and is a type of finite impulse response filter. Variations include: simple, and cumulative, or weighted forms.

- We have read the data head to check the values.
- Fig 9. shows the trend on the test and train data.

Fig 10. Plotting on both Train and Test



Model Evaluation

```
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
```

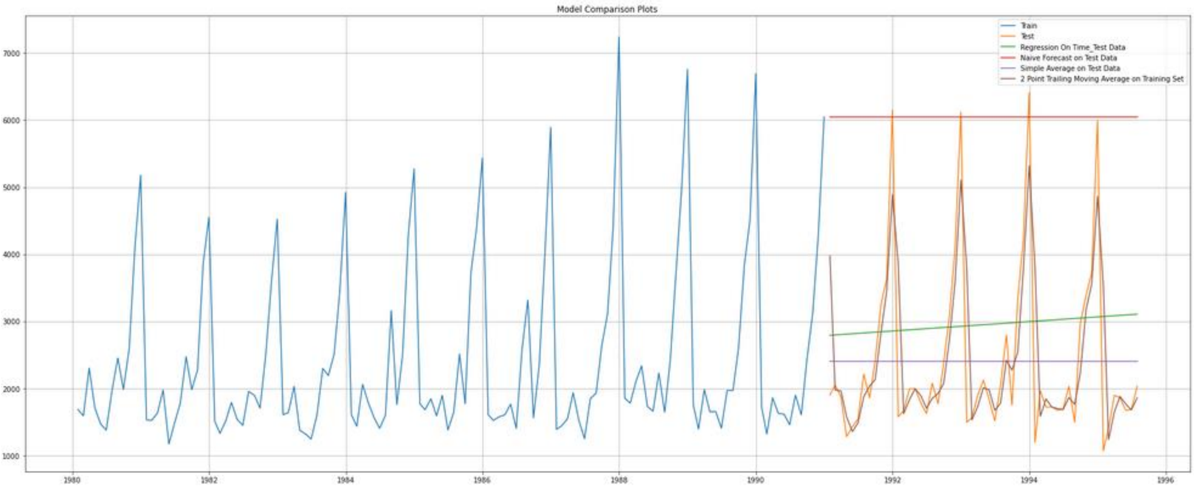
Table 11. Test RMSE table for all the models above.

	Test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315

- 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 1128.927
- For 9 point moving average model forecast on the training data RMSE is 1346.278
- We have noted the value for all the models above in the Table 11 for a better comparison.
- We have also plotted all the models so far to show the comparison as of now the best model performance is for 2 point moving average model forecast on the training data RMSE is 813.401

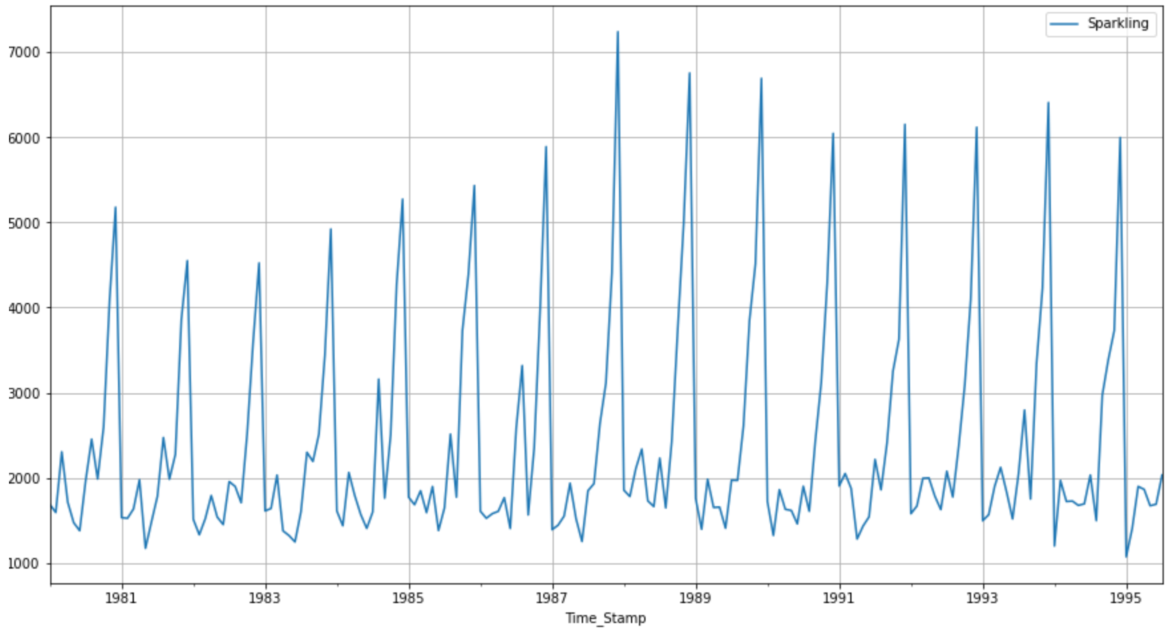
Before we go on to build the various Exponential Smoothing models, let us plot all the models and compare the Time Series plots.

<u>Fig 11. All the above models</u>



Method 5: Simple Exponential Smoothing

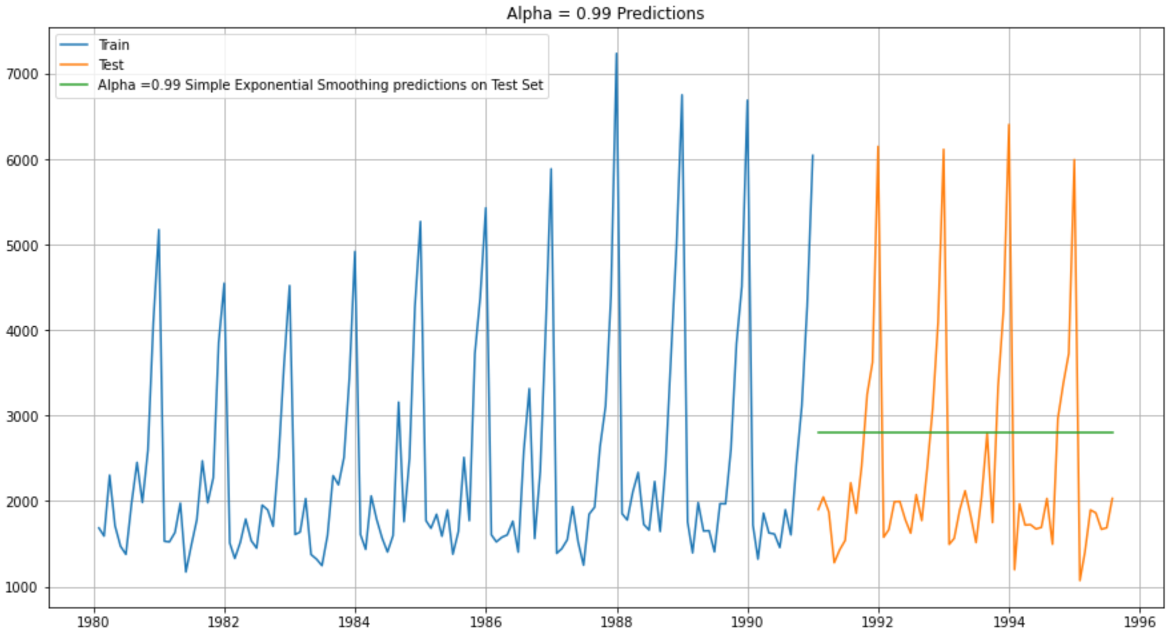
<u>Fig 12. Plot for SES</u>



<u>Table for Smoothing level and Trend.</u>

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

<u>Fig 13. SES - ETS(A, N, N) - Simple Exponential Smoothing with additive error</u>



<u>RMSE SES With Additive error</u>

```
SES RMSE: 1338.0083844916467
SES RMSE (calculated using statsmodels): 1338.0083844916464
```

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

{'smoothing_level': 0.1112722693612978 2, 'smoothing_trend': 0.01236080366981065, 'smoothing_seasonal': 0.4607176701154181 4,
'damping_trend': nan, 'initial_level': 2356.5780356927266, 'initial_trend': -0.1026206744222035, 'initial_seasons': array
([-636.2332476 , -722.98331145, -398.64399888, -473.43056513,
       -808.42484318, -815.34991251, -384.23076216,   72.99484233,
       -237.44231456,  272.32602717, 1541.3773669 , 2590.07686414]), 'use_boxcox': False, 'lamda': None, 'remove_bias': Fal
se}
```

- Exponential smoothing is a rule of thumb technique for smoothing time series data using the exponential window function. Whereas in the simple moving average the past observations are weighted equally, exponential functions are used to assign exponentially decreasing weights over time. It is an easily learned and easily applied

procedure for making some determination based on prior assumptions by the user, such as seasonality. Exponential smoothing is often used for analysis of time-series data.
We have read the data head to check the values.

- Fig 13. shows the trend on the test and train data for SES.
- The RMSE for SES is 1338.00 which is the average of other models too.
- We have noted the value for all three models above for a better comparison.

Holt-Winters - ETS(A, A, A) - Holt Winter's linear method with additive errors

Table for smoothing level:


```
{'smoothing_level': 0.11127226936129782, 'smoothing_trend': 0.01236080366981065, 'smoothing_seasonal': 0.46071767011541814,
'damping_trend': nan, 'initial_level': 2356.5780356927266, 'initial_trend': -0.1026206744222035, 'initial_seasons': array
([-636.2332476 , -722.98331145, -398.64399888, -473.43056513,
       -808.42484318, -815.34991251, -384.23076216,   72.99484233,
       -237.44231456,  272.32602717, 1541.3773669 , 2590.07686414]), 'use_boxcox': False, 'lamda': None, 'remove_bias': Fal
se}
```

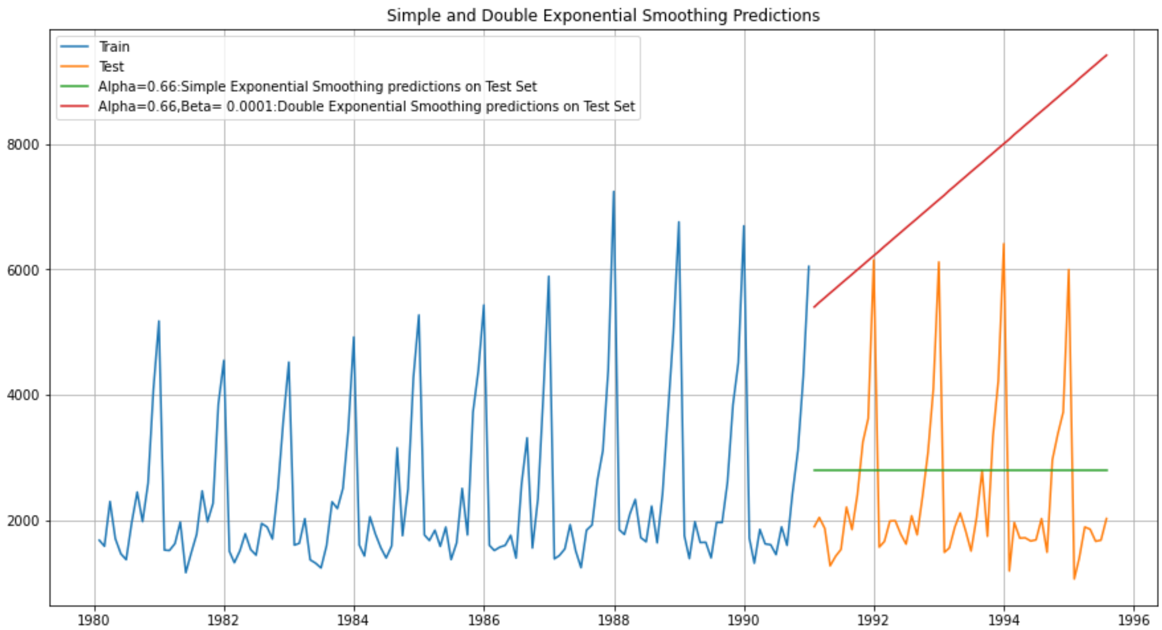

- *Holt - ETS(A, A, N) - Holt's linear method with additive errors Double Exponential Smoothing*
- *One of the drawbacks of the simple exponential smoothing is that the model does not do well in the presence of the trend.*
- *This model is an extension of SES known as Double Exponential model which estimates two smoothing parameters.*
- *Applicable when data has Trend but no seasonality.*
- *Two separate components are considered: Level and Trend.*
- *Level is the local mean.*
- *One smoothing parameter α corresponds to the level series*
- *A second smoothing parameter β corresponds to the trend series.*
- *Double Exponential Smoothing uses two equations to forecast future values of the time series, one for forecasting the short term average value or level and the other for capturing the trend.*

- As an extension of **Holt's** exponential smoothing that captures seasonality. This method produces exponentially smoothed values for the level of the forecast, the trend of the forecast, and the seasonal adjustment to the forecast.

Holt model Exponential Smoothing Estimated Parameters


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

Fig 13. Plot for Simple and Double Exponential Smoothing Predictions



DES RMSE:

```
DES RMSE: 5291.8798332269125
```

Holt-Winters - ETS(A, A, A) - Holt Winter's linear method with additive errors

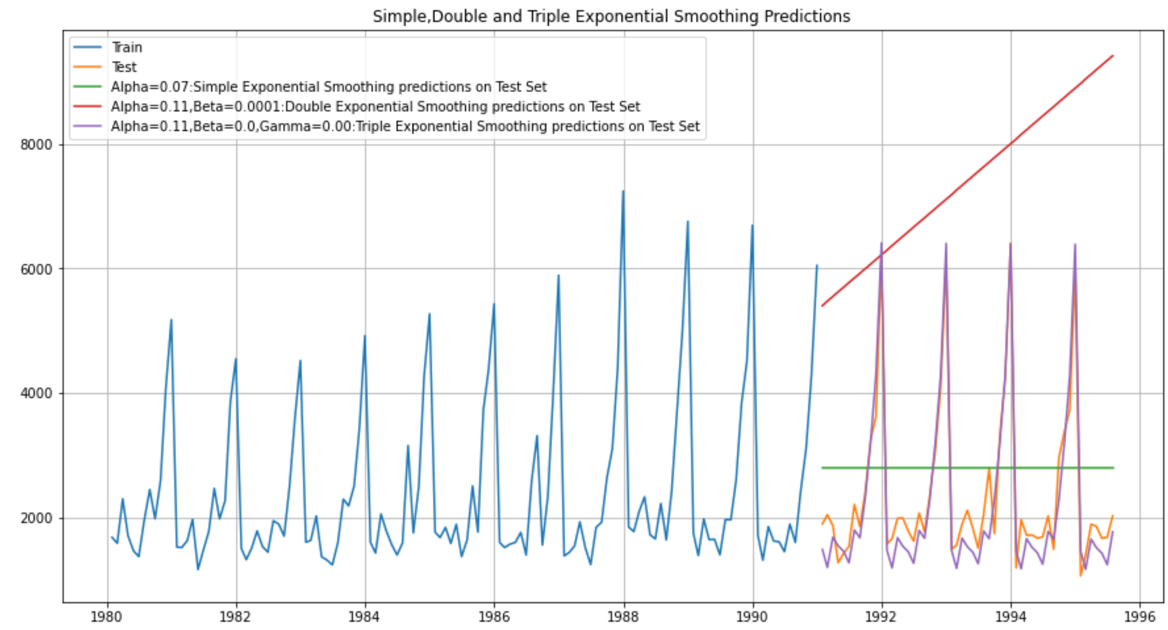
Table for smoothing level:

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

{'smoothing_level': 0.11127226936129782, 'smoothing_trend': 0.01236080366981065, 'smoothing_seasonal': 0.46071767011541814,
'damping_trend': nan, 'initial_level': 2356.5780356927266, 'initial_trend': -0.1026206744222035, 'initial_seasons': array
([-636.2332476 , -722.98331145, -398.64399888, -473.43056513,
       -808.42484318, -815.34991251, -384.23076216,   72.99484233,
       -237.44231456,  272.32602717, 1541.3773669 , 2590.07686414]), 'use_boxcox': False, 'lamda': None, 'remove_bias': Fal
se}
```

- Fig 13. shows the trend on the test and train data for SES.
- The RMSE for DES is 5291.00 which is higher than all the models.
- We have noted the value for all models above for a better comparison.

Fig 14. Plot for Simple, Double and Triple Exponential Smoothing Predictions




```
TES RMSE: 378.95173454983535
```

Hence Triple exponential smoothing is the best performing model.

	Test RMSE
Alpha=0.99,SES	1338.008384
Alpha=0.66,Beta=0.0001:DES	5291.879833
Alpha=0.11,Beta=0.01,Gamma=0.46:TES	378.951735

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

- The RMSE for TES is 378.95 which is lowest RMSE when compared to all the models.
- We have noted the value for all models above for a better comparison.
- Fig 14 shows the graph for SES, DES and TES.

Table for smoothing level:

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

{'smoothing_level': 0.11107159018710593, 'smoothing_trend': 0.04936515256239578, 'smoothing_seasonal': 0.36215619523685727,
'damping_trend': nan, 'initial_level': 2356.541733581815, 'initial_trend': -9.17995980894553, 'initial_seasons': array([0.7
1328861, 0.68355166, 0.90390669, 0.80572594, 0.65638171,
       0.65461713, 0.8865469 , 1.13392754, 0.91872967, 1.21236472,
       1.86777061, 2.37039694]), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}
```

Fig 15. Plot for Simple, Double and Triple Exponential Smoothing Predictions

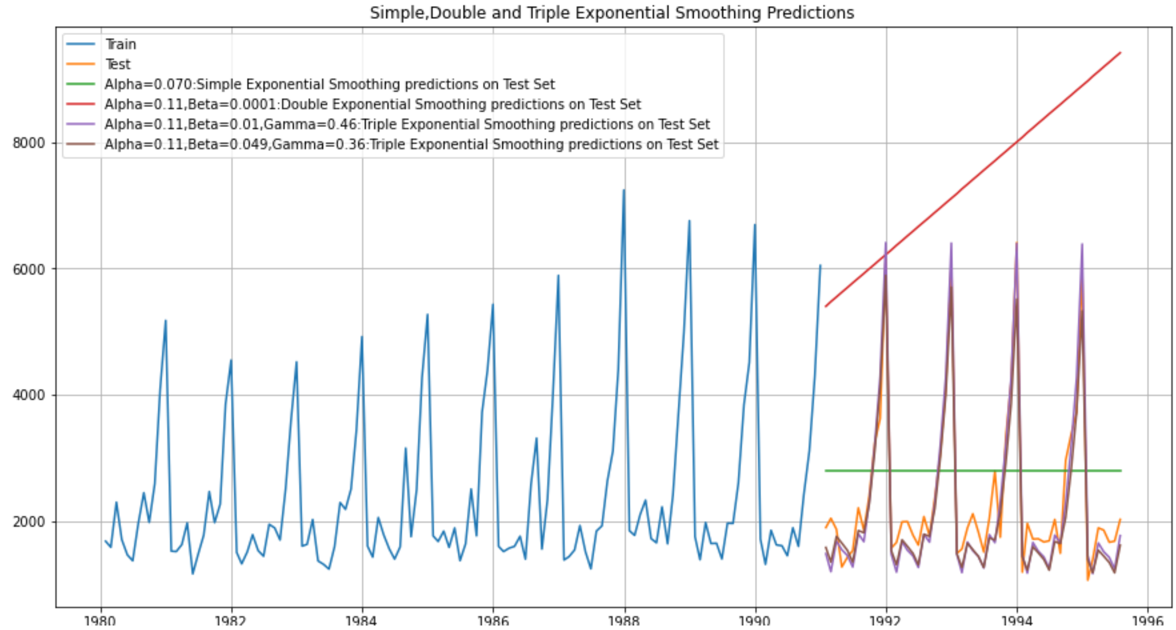


Table 14. Report model accuracy

```
TES_am RMSE: 403.21056676066024
```

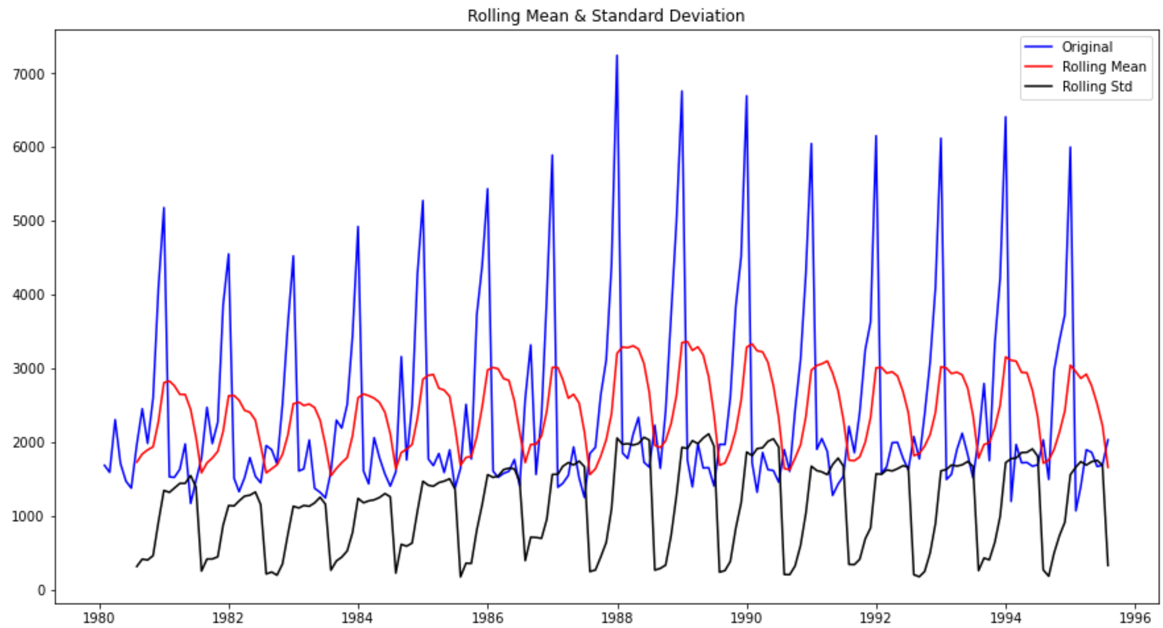
	Test RMSE
Alpha=0.99,SES	1338.008384
Alpha=0.66,Beta=0.0001:DES	5291.879833
Alpha=0.11,Beta=0.01,Gamma=0.46:TES	378.951735
Alpha=0.74,Beta=2.73e-06,Gamma=5.2e-07,Gamma=0:TES	403.210567

- The model accuracy for TES_am RMSE is 403.21 which is higher than the TES.
- Fig 14 shows the graph for SES, DES and TES predictions.
-

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.

Answer:

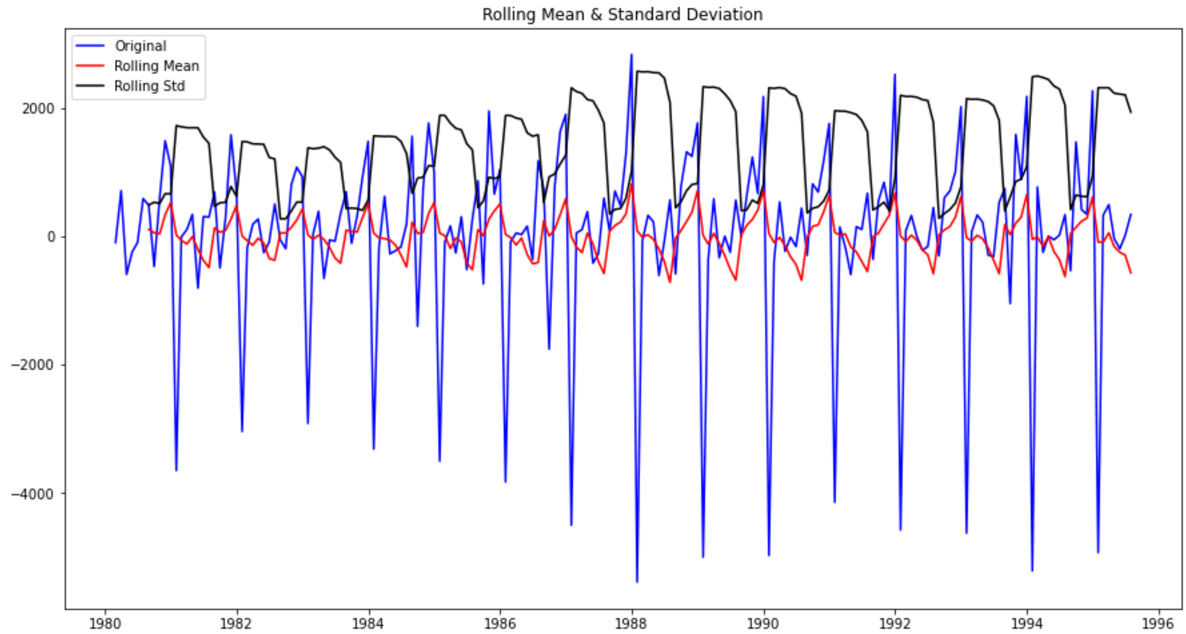
Fig 15. Plot for Rolling mean and standard mean




```
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% significant level the Time Series is non-stationary.
- Let us take a difference of order 1 and check whether the Time Series is stationary or not.

Fig 16. Plot for difference of Rolling mean and standard mean




```
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 at α= 0.05 the Time Series is indeed stationary.
- Plot the Autocorrelation and the Partial Autocorrelation function plots on the whole data.

Fig 17. Auto correlation:

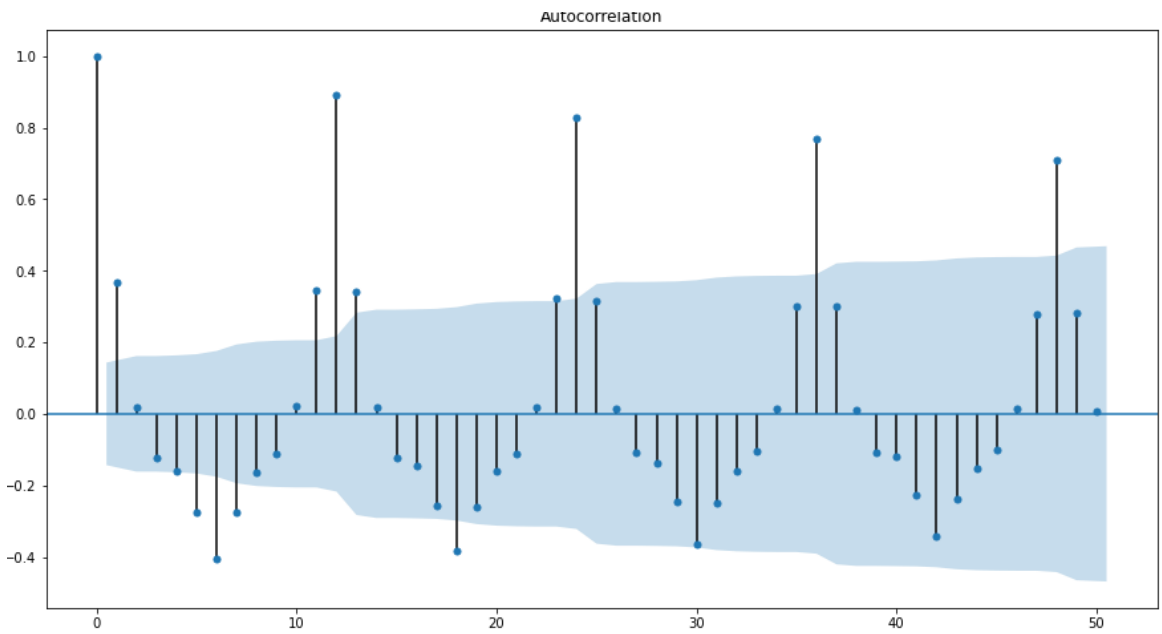


Fig 18. Differenced data Autocorrelation:

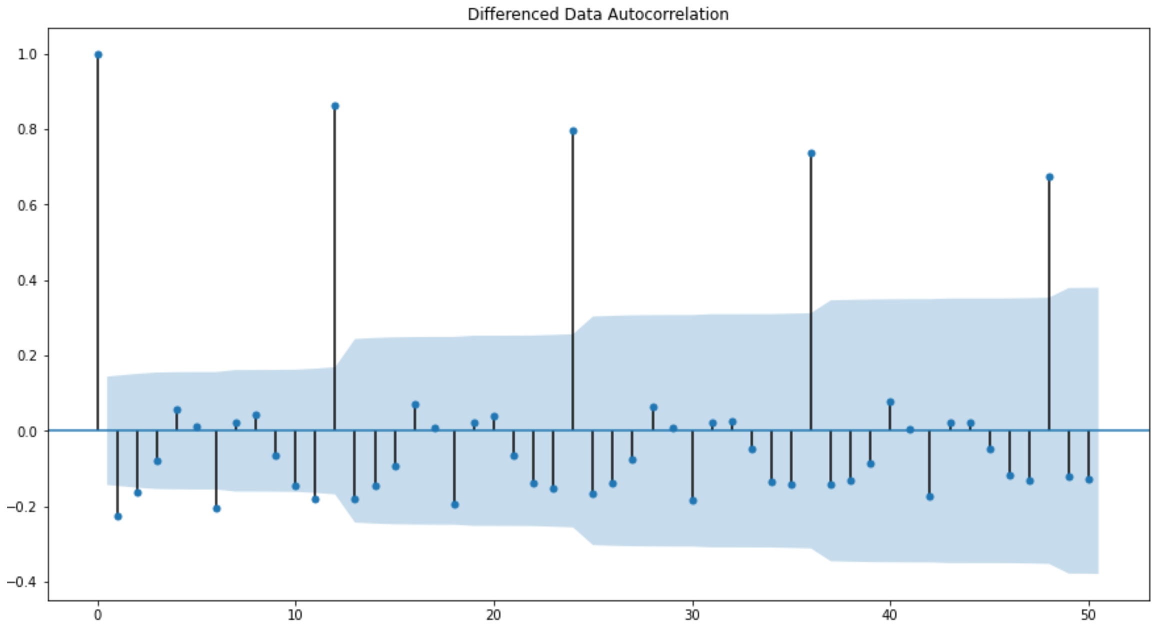


Fig 19. Plot for Partial Autocorrelation

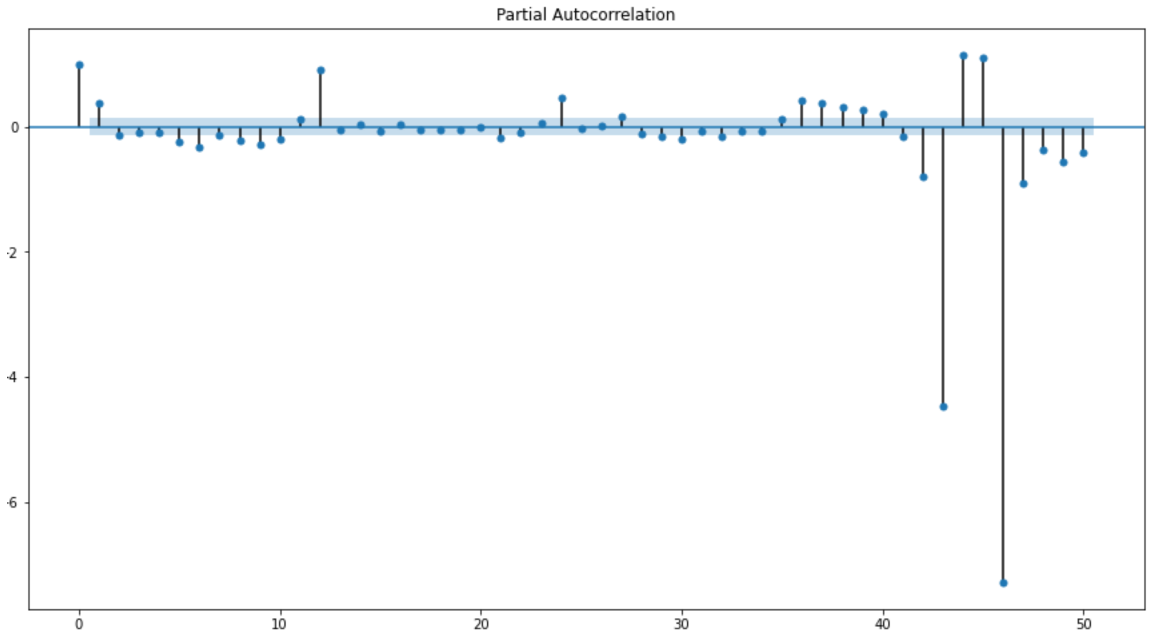
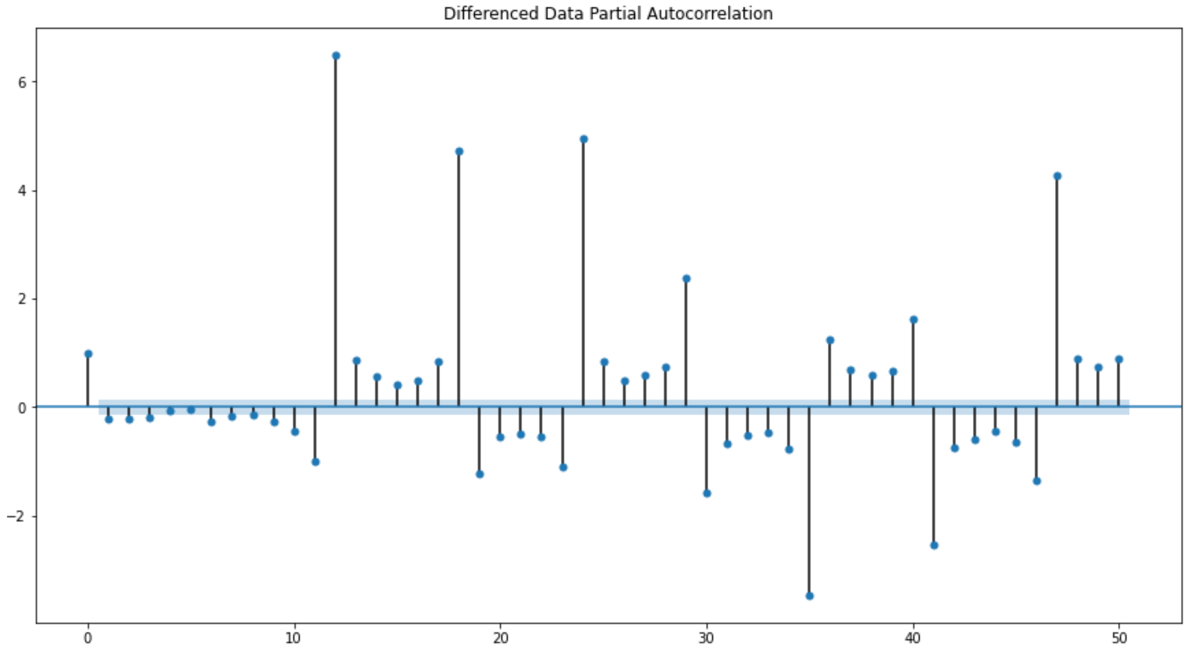


Fig 20. Differenced Data Partial Autocorrelation:



From the above plots, we can say that there seems to be a seasonality in the data.

- We have plotted different graphs for Rolling and difference mean, Autocorrelation and partial auto correlation.
- The graphs show trend and seasonality in the data set.
- However the correlation graphs show seasonality but no trend.
- We don't see any trend in the data set.

Table 14. First few and Last few rows of training and testing dataset:

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

Last few rows of Training Da

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

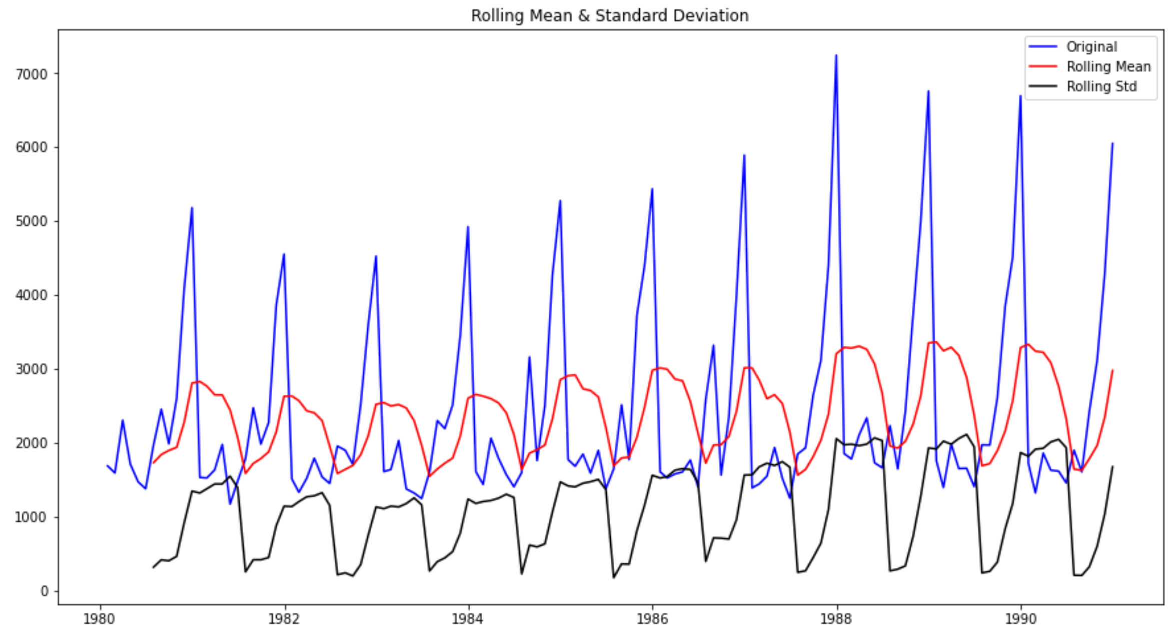
First few rows of Test Data

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

Last few rows of Test Data

	Sparkling
Time_Stamp	
1995-03-31	1897
1995-04-30	1862
1995-05-31	1670

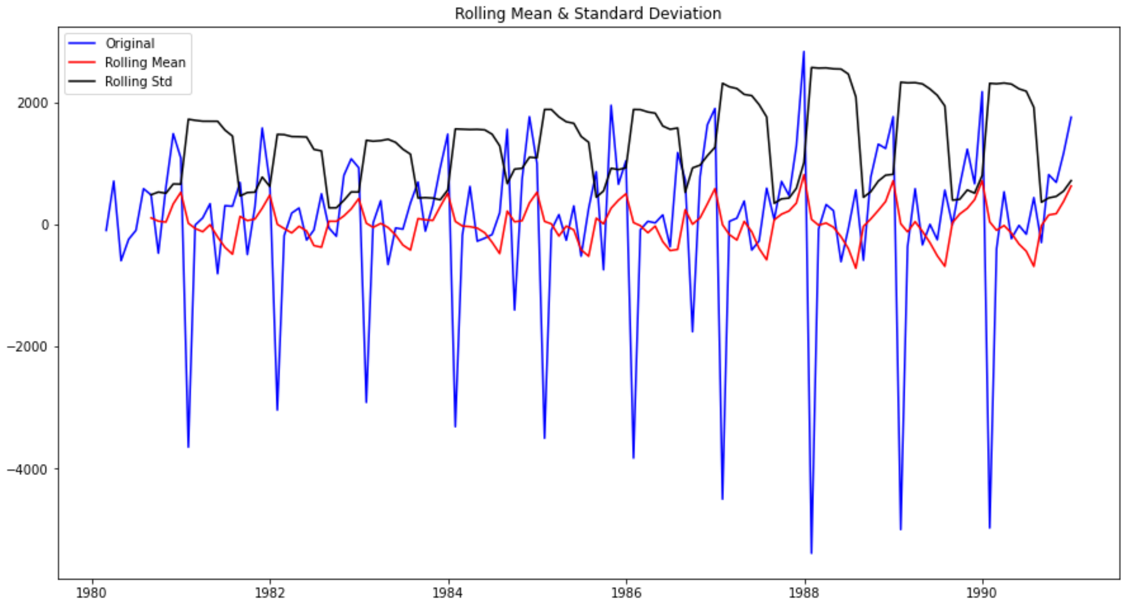
Fig 20. Plot for of Rolling mean and standard deviation




```
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
```

- We can see the rolling mean and standard deviation show trend in the data but no seasonality.

<u>Fig 21. Plot for difference of Rolling mean and standard deviation</u>




```
Results of Dickey-Fuller Test:
Test Statistic                 -8.005007e+00
p-value                         2.280104e-12
#Lags Used                      1.100000e+01
Number of Observations Used     1.190000e+02
Critical Value (1%)            -3.486535e+00
Critical Value (5%)            -2.886151e+00
Critical Value (10%)           -2.579896e+00
dtype: float64
```

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.

Answer:

Considering d=(0,1)

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

AIC of ARIMA Model:

```
ARIMA(0, 0, 2) - AIC:2245.343217655809
ARIMA(1, 0, 0) - AIC:2247.3482759501685
ARIMA(1, 0, 1) - AIC:2245.9490908876764
ARIMA(1, 0, 2) - AIC:2246.0121932465745
ARIMA(2, 0, 0) - AIC:2244.7999145664703
ARIMA(2, 0, 1) - AIC:2236.590818513248
ARIMA(2, 0, 2) - AIC:2200.904428580393
```

PARAM AIC

	param	AIC
8	(2, 0, 2)	2200.904429
7	(2, 0, 1)	2236.590819
6	(2, 0, 0)	2244.799915
1	(0, 0, 1)	2245.268851
2	(0, 0, 2)	2245.343218
4	(1, 0, 1)	2245.949091
5	(1, 0, 2)	2246.012193
3	(1, 0, 0)	2247.348276
0	(0, 0, 0)	2271.203212

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:            D.Sparkling   No. Observations:                  131
Model:                 ARIMA(2, 1, 1)   Log Likelihood               -1111.180
Method:                       css-mle   S.D. of innovations           1148.859
Date:                Tue, 06 Apr 2021   AIC                           2232.360
Time:                        12:39:07   BIC                           2246.736
Sample:                    02-29-1980   HQIC                          2238.202
                         - 12-31-1990
=================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------
const                 6.2031      3.807      1.629      0.103      -1.259      13.66
ar.L1.D.Sparkling     0.5026      0.087      5.753      0.000       0.331       0.67
ar.L2.D.Sparkling    -0.1910      0.088     -2.182      0.029      -0.363      -0.01
ma.L1.D.Sparkling    -1.0000      0.019    -51.615      0.000      -1.038      -0.96
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.3156           -1.8721j            2.2881           -0.1525
AR.2            1.3156           +1.8721j            2.2881            0.1525
MA.1            1.0000           +0.0000j            1.0000            0.0000
-----------------------------------------------------------------------------
```

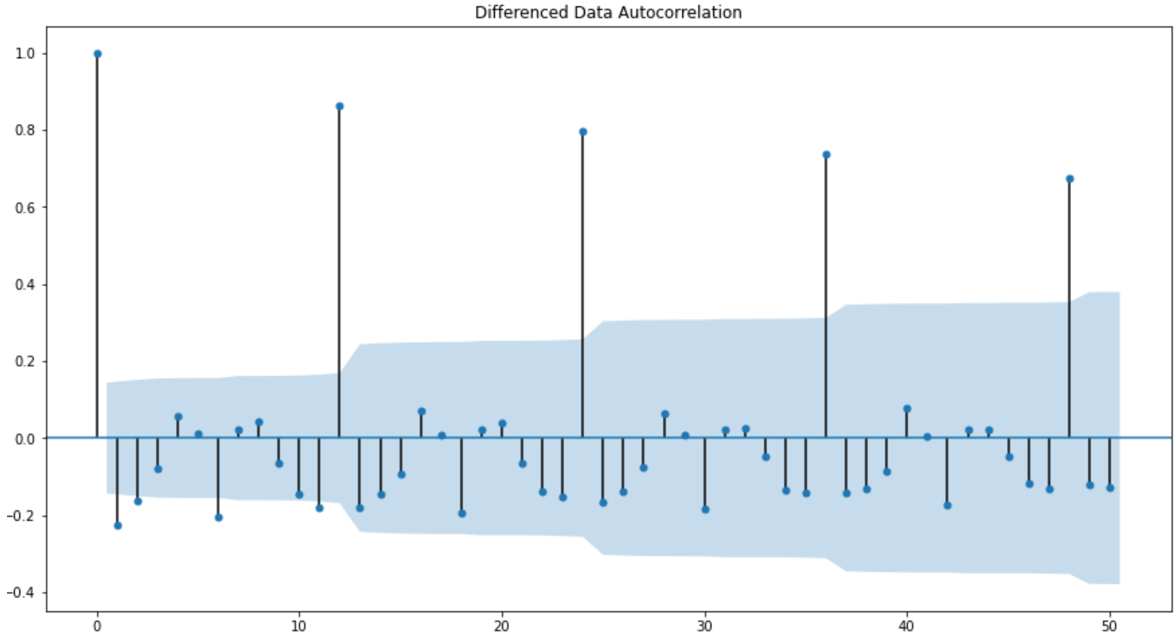
Predict on the Test Set using this model and evaluate the model.

```
1418.2020848039745
```

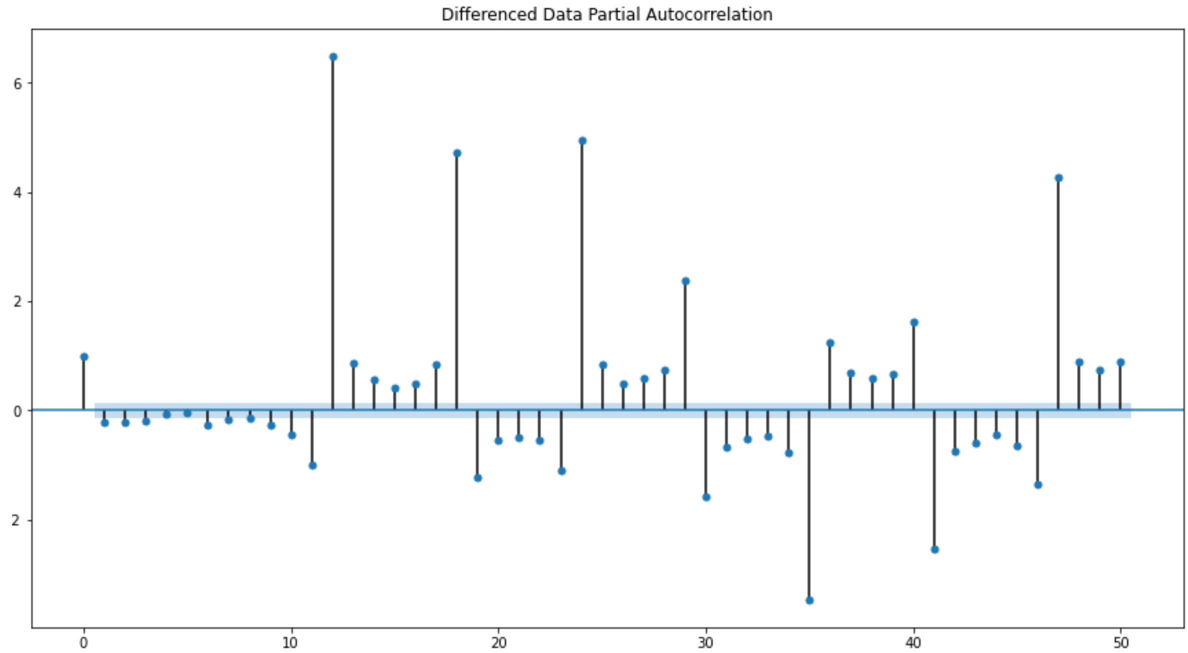
	Test RMSE
ARIMA(2,1,1)	1418.202085

Build a version of the ARIMA model for which the best parameters are selected by looking at the ACF and the PACF plots.

Fig 21. Differenced Data Autocorrelation ARIMA



<u>Fig 22. Differenced Data Partial Autocorrelation ARIMA</u>



- 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 the ACF plot cuts-off to 0.
- By looking at the above plots, we can say that both the PACF and ACF plot cuts-off at lag 0.

- We can see the AIC of ARIMA models are in thousands which Is good however we need to make sure the values are consistent.4

- Hence we will be considering Manual ARIMA where the RMSE value is 4779 which is very high which shows the model is not stable.

Manual Arima model:

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:            D.Sparkling   No. Observations:                  131
Model:                 ARIMA(0, 1, 0)   Log Likelihood               -1132.791
Method:                           css   S.D. of innovations           1377.911
Date:                Tue, 06 Apr 2021   AIC                           2269.583
Time:                        12:39:08   BIC                           2275.333
Sample:                    02-29-1980   HQIC                          2271.919
                         - 12-31-1990
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         33.2901    120.389      0.277      0.782    -202.667     269.248
==============================================================================
```

Predict on the Test Set using this model and evaluate the model.

```
4779.15429919654
```

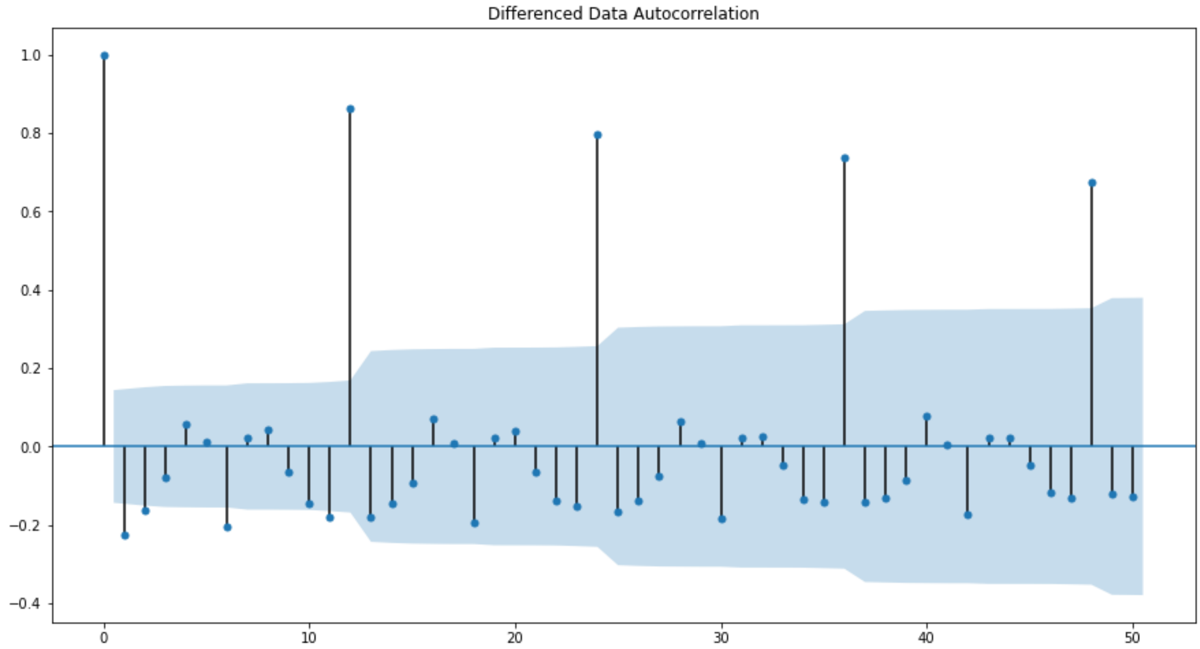
	Test RMSE
ARIMA(2,1,1)	1418.202085
ARIMA(0,1,0)	4779.154299

- We see that there is difference in the RMSE values for both the models, but remember that the second model is a much simpler model.

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

- Autoregressive integrated moving average In statistics and econometrics, and in particular in time series analysis, an autoregressive integrated moving average model is a generalization of an autoregressive moving average model. Both of these models are fitted to time series data either to better understand the data or to predict future points in the series.

Fig 23. Differenced Data Partial Autocorrelation SARIMA



- Setting the seasonality as 5 for the first iteration of the auto SARIMA model.

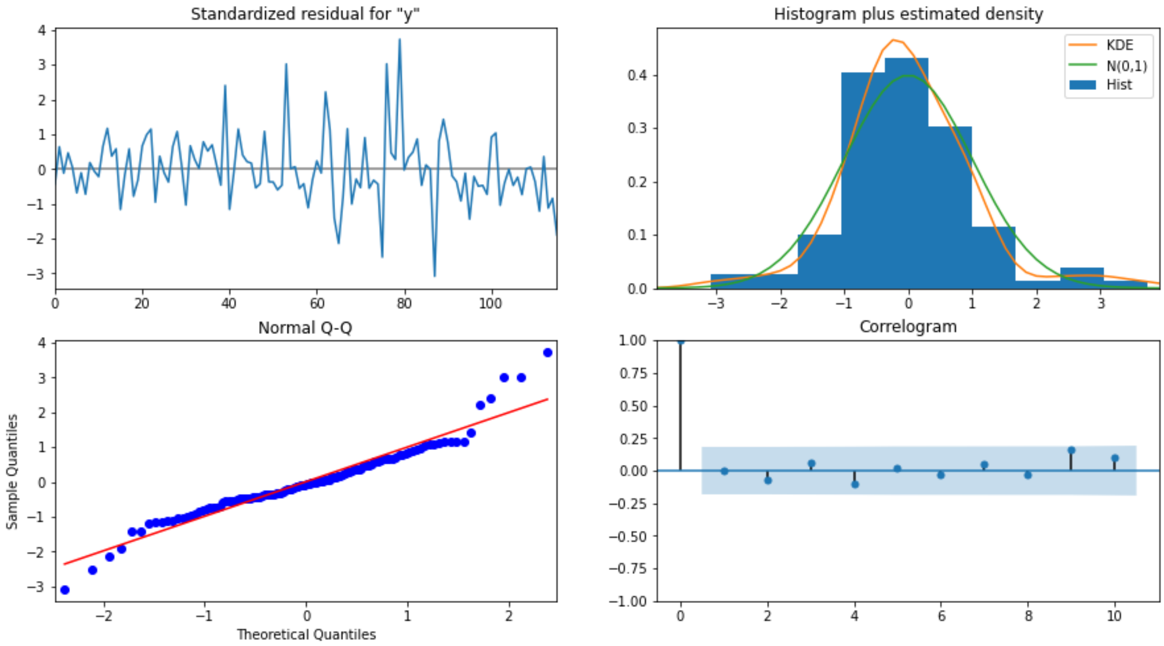
```
Examples of some parameter combinations for Model...
Model: (0, 0, 1)(0, 1, 1, 5)
Model: (0, 0, 2)(0, 1, 2, 5)
Model: (1, 0, 0)(1, 1, 0, 5)
Model: (1, 0, 1)(1, 1, 1, 5)
Model: (1, 0, 2)(1, 1, 2, 5)
Model: (2, 0, 0)(2, 1, 0, 5)
Model: (2, 0, 1)(2, 1, 1, 5)
Model: (2, 0, 2)(2, 1, 2, 5)
```

	param	seasonal	AIC
74	(2, 0, 2)	(0, 1, 2, 5)	1942.994484
50	(1, 0, 2)	(1, 1, 2, 5)	1959.901587
47	(1, 0, 2)	(0, 1, 2, 5)	1961.106356
53	(1, 0, 2)	(2, 1, 2, 5)	1961.539073
23	(0, 0, 2)	(1, 1, 2, 5)	1961.620828

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 6)   Log Likelihood                -856.944
Date:                           Tue, 06 Apr 2021   AIC                           1727.888
Time:                                   12:39:33   BIC                           1747.163
Sample:                                        0   HQIC                          1735.713
                                           - 132
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -1.0093      0.175     -5.753      0.000      -1.353      -0.665
ma.L2         -0.1219      0.131     -0.930      0.353      -0.379       0.135
ar.S.L6        0.0022      0.026      0.084      0.933      -0.049       0.053
ar.S.L12       1.0396      0.018     58.246      0.000       1.005       1.075
ma.S.L6        0.0428      0.144      0.298      0.765      -0.238       0.324
ma.S.L12      -0.6202      0.090     -6.877      0.000      -0.797      -0.443
sigma2       1.18e+05   1.84e+04      6.409      0.000    8.19e+04    1.54e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                38.96
Prob(Q):                              0.97   Prob(JB):                         0.00
Heteroskedasticity (H):               2.85   Skew:                             0.58
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.59
===================================================================================

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

Fig 24. Different graphs for standard residual for Y



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

Predict on the Test Set using this model and evaluate the model.

Summary frame:

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1375.673492	384.090298	622.870340	2128.476643
1	1116.710838	392.854482	346.730201	1886.691475
2	1667.608326	395.428052	892.583585	2442.633066
3	1528.351433	397.988074	748.309143	2308.393724
4	1372.244861	400.531787	587.216984	2157.272739

RMSE:

```
601.2724145005159
```

	Test RMSE
ARIMA(2,1,1)	1418.202085
ARIMA(0,1,0)	4779.154299
SARIMA(0,1,2)(2,0,2,6)	601.272415

- Here we can see the SARIMA model is performing much better than the ARIMA model at 601.2

Setting the seasonality as 10 for the second iteration of the auto SARIMA model.

```
Examples of some parameter combinations for Model...
Model: (0, 0, 1)(0, 1, 1, 10)
Model: (0, 0, 2)(0, 1, 2, 10)
Model: (1, 0, 0)(1, 1, 0, 10)
Model: (1, 0, 1)(1, 1, 1, 10)
Model: (1, 0, 2)(1, 1, 2, 10)
Model: (2, 0, 0)(2, 1, 0, 10)
Model: (2, 0, 1)(2, 1, 1, 10)
Model: (2, 0, 2)(2, 1, 2, 10)
```

AIC score

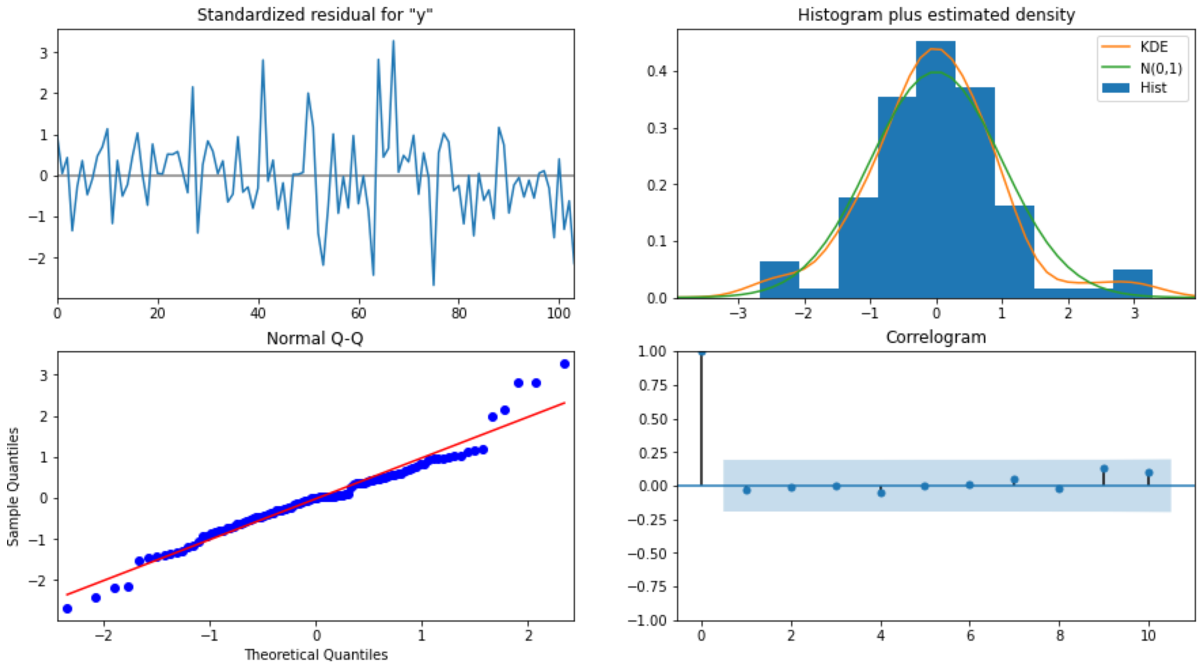
	param	seasonal	AIC
77	(2, 0, 2)	(1, 1, 2, 10)	1656.765529
53	(1, 0, 2)	(2, 1, 2, 10)	1697.395957
80	(2, 0, 2)	(2, 1, 2, 10)	1698.932153
47	(1, 0, 2)	(0, 1, 2, 10)	1700.098448
50	(1, 0, 2)	(1, 1, 2, 10)	1702.062080

- We can see the AIC for SARIMA model is also performing in 1656 and increasing with the Param and seasonal.
- If the predictors consist only of lagged values of Y, it is a pure autoregressive ("self-regressed") model, which is just a special case of a regression model and which could be fitted with standard regression software.
- For example, a first-order autoregressive ("AR(1)") model for Y is a simple regression model in which the independent variable is just Y lagged by one period (LAG(Y,1) in Stat graphics or Y_LAG1 in Regress
- If some of the predictors are lags of the errors, an ARIMA model it is NOT a linear regression model, because there is no way to specify "last period's error" as an independent variable: the errors must be computed on a period-to-period basis when the model is fitted to the data.

Model details

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(1, 1, 2)x(2, 0, 2, 12)   Log Likelihood                -770.040
Date:                            Tue, 06 Apr 2021   AIC                           1556.080
Time:                                    12:40:24   BIC                           1577.235
Sample:                                         0   HQIC                          1564.651
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.6386      0.291     -2.194      0.028      -1.209      -0.068
ma.L1         -0.0981      0.257     -0.382      0.702      -0.601       0.405
ma.L2         -0.7246      0.174     -4.153      0.000      -1.067      -0.383
ar.S.L12       0.7625      0.595      1.282      0.200      -0.403       1.928
ar.S.L24       0.2940      0.618      0.476      0.634      -0.918       1.506
ma.S.L12      -0.3121      0.589     -0.530      0.596      -1.466       0.842
ma.S.L24      -0.2691      0.360     -0.748      0.455      -0.975       0.436
sigma2      1.521e+05   2.06e+04      7.403      0.000    1.12e+05    1.92e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.08   Jarque-Bera (JB):                12.42
Prob(Q):                              0.78   Prob(JB):                         0.00
Heteroskedasticity (H):               1.54   Skew:                             0.35
Prob(H) (two-sided):                  0.21   Kurtosis:                         4.54
===================================================================================
```

Fig 25. Different graphs for standard residual for Y



Predicted auto SARIMA

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1317.893111	390.156571	553.200283	2082.585938
1	1308.037906	403.451661	517.287180	2098.788631
2	1607.837807	403.465693	817.059581	2398.616033
3	1598.855363	408.960375	797.307757	2400.402969
4	1377.308640	409.817024	574.082032	2180.535248

RMSE:

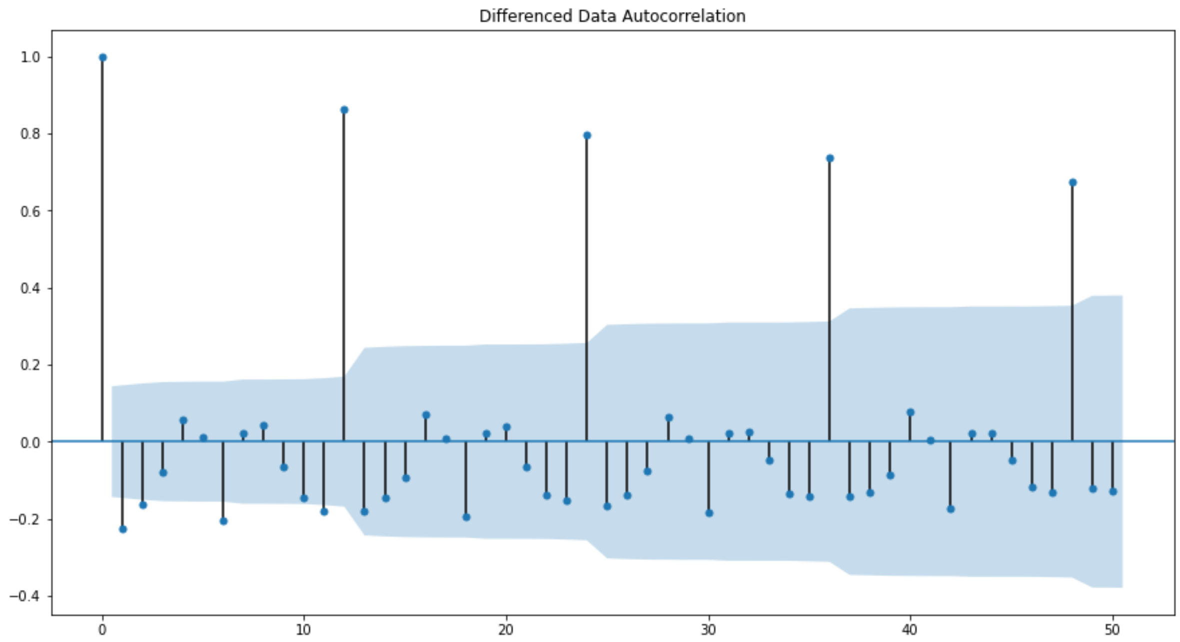
```
548.0176751478789
```

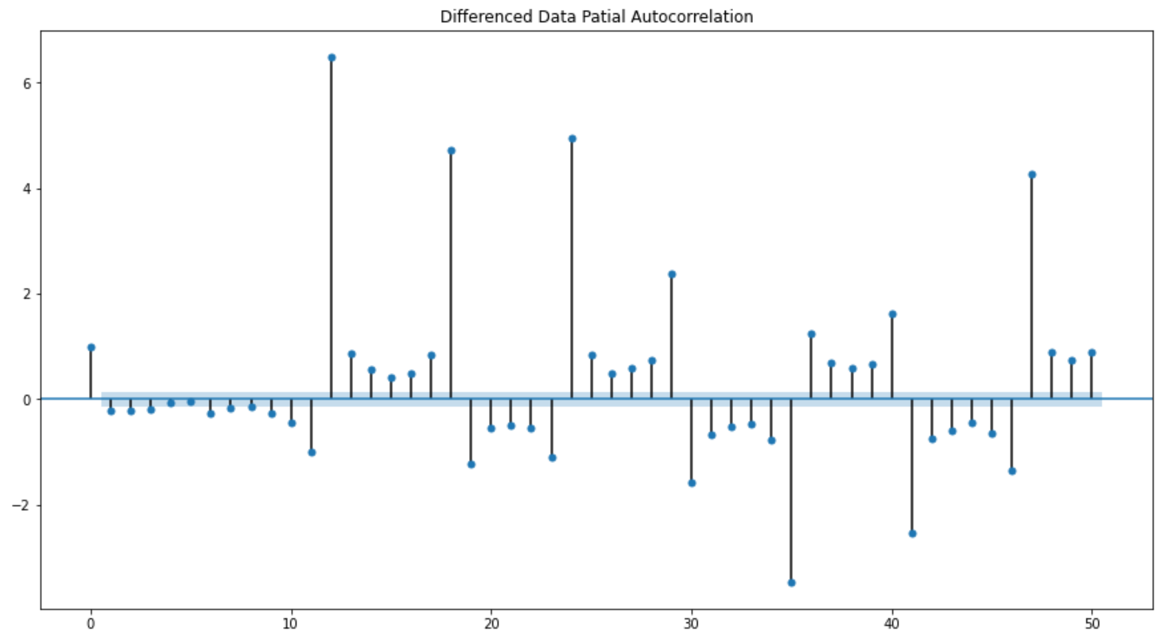
	Test RMSE
ARIMA(2,1,1)	1418.202085
ARIMA(0,1,0)	4779.154299
SARIMA(0,1,2)(2,0,2,6)	601.272415
SARIMA(1,1,2)(2,0,2,12)	548.017675

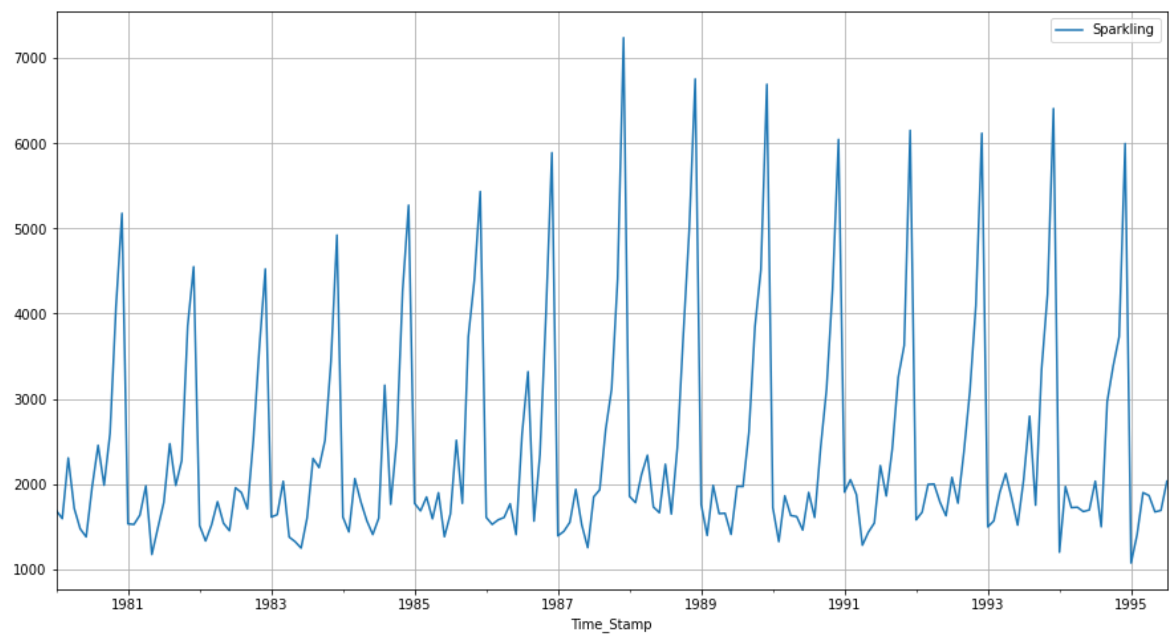
- Here we cam see the predicted auto SARIMA RMSE values have decreased to 548.

- We can see out of all the models SARIMA is the best performing model.
- We are building a model of SARIMA with seasonality 6 to check the best paremeters.

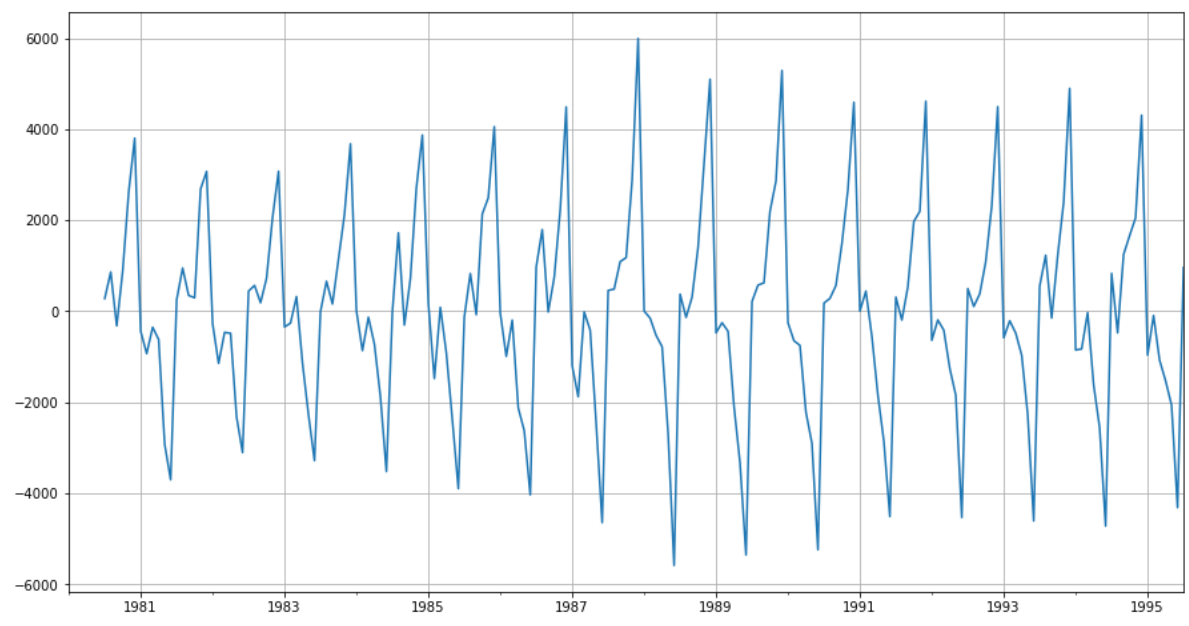
Build a version of the SARIMA model for which the best parameters are selected by looking at the ACF and the PACF plots. - Seasonality at 6.

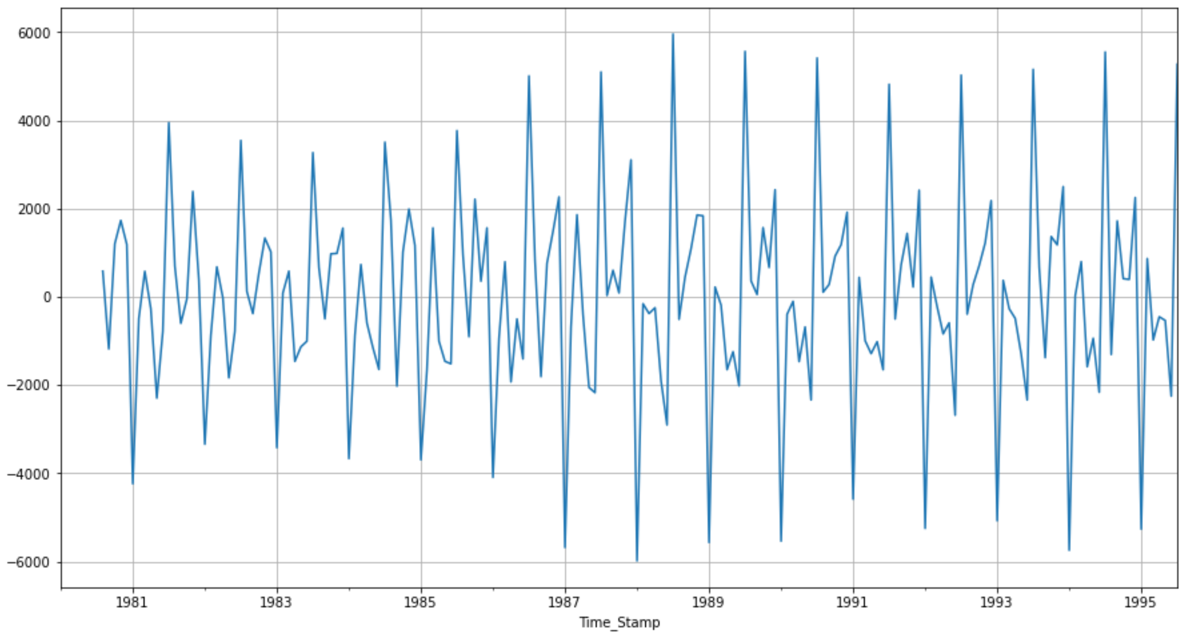


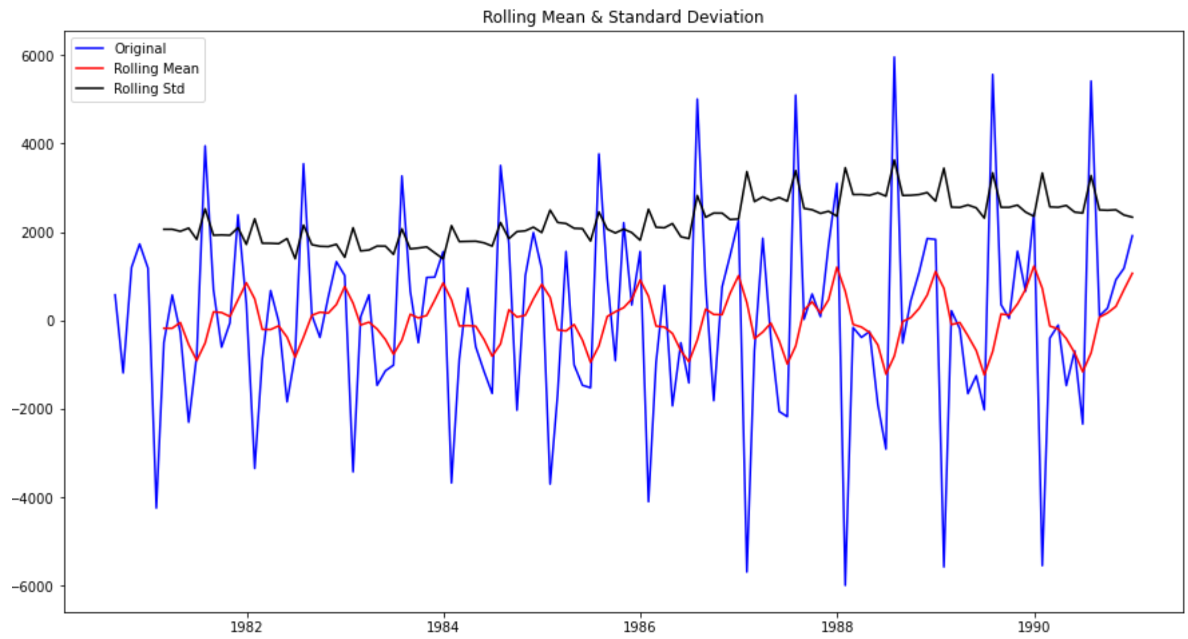




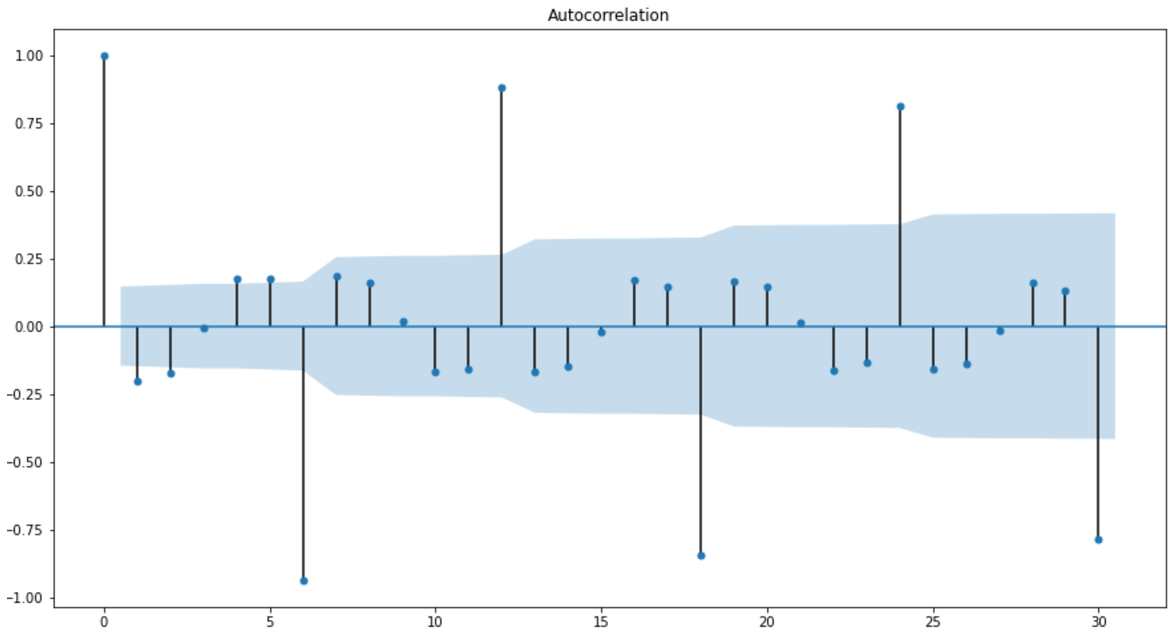
We see that there is a trend and a seasonality. So, now we take a seasonal differencing and check the series.

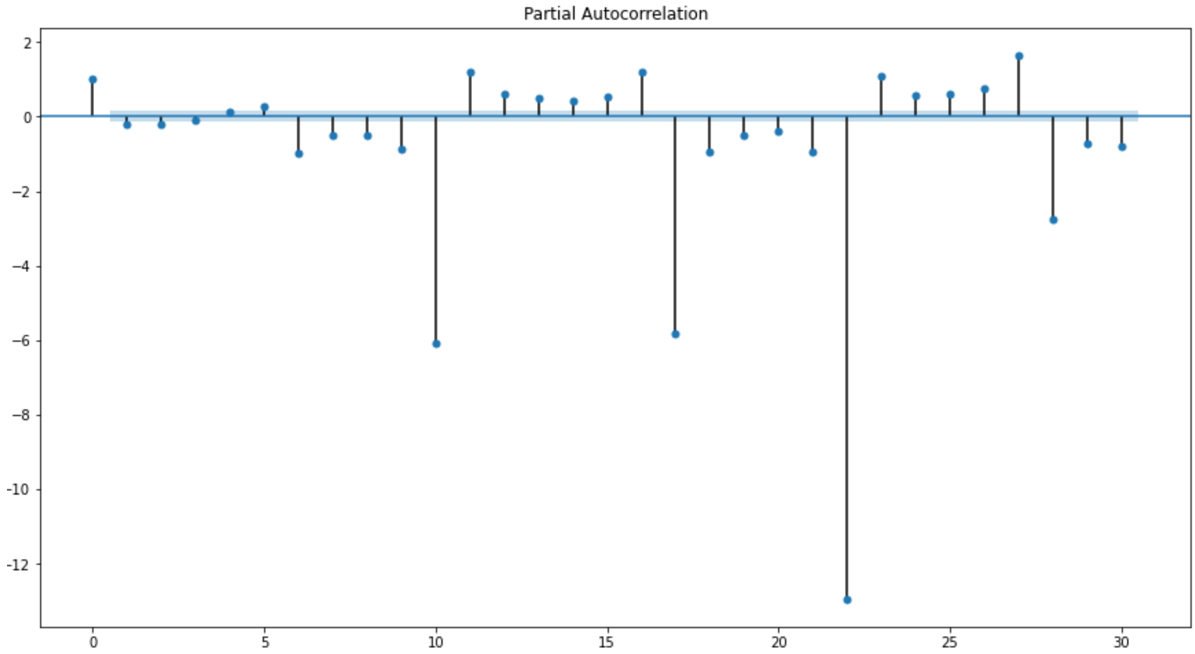






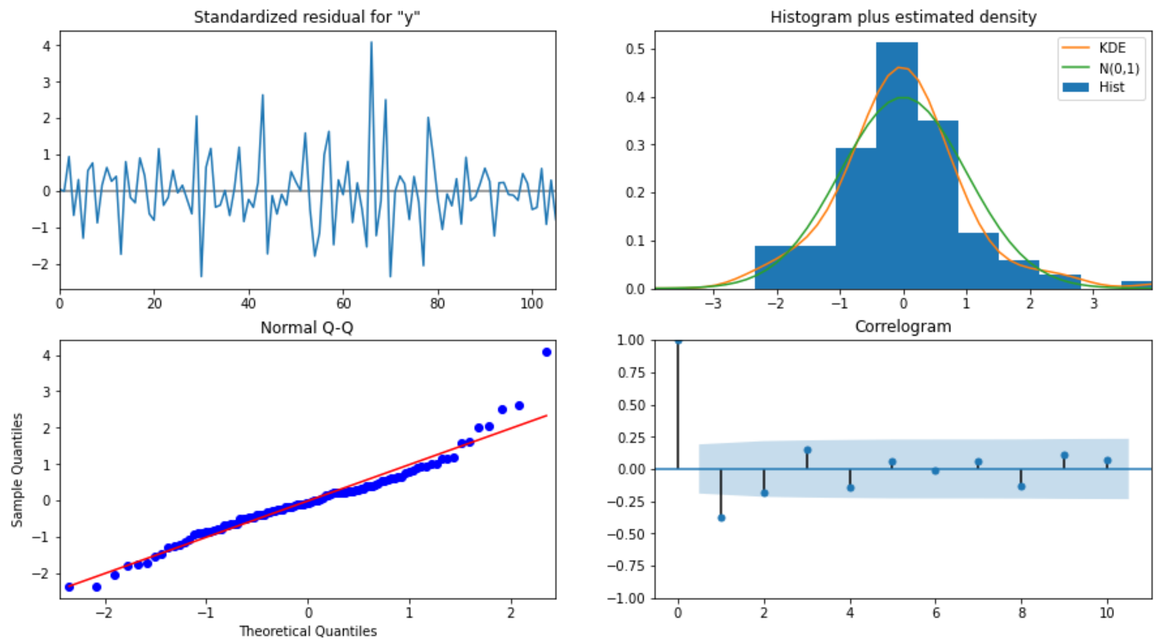

```
Results of Dickey-Fuller Test:
Test Statistic                -7.017242e+00
p-value                        6.683657e-10
#Lags Used                     1.300000e+01
Number of Observations Used    1.110000e+02
Critical Value (1%)           -3.490683e+00
Critical Value (5%)           -2.887952e+00
Critical Value (10%)          -2.580857e+00
dtype: float64
```


Autocorrelation
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
-1.00
0
5
10
15
20
25
30


Partial Autocorrelation
2
0
-2
-4
-6
-8
-10
-12
0
5
10
15
20
25
30

Summary:

```
                                      SARIMAX Results
============================================================================================
Dep. Variable:                                   y   No. Observations:                   132
Model:             SARIMAX(0, 1, 0)x(1, 1, [1, 2, 3], 6)   Log Likelihood               -811.726
Date:                             Tue, 06 Apr 2021   AIC                            1633.452
Time:                                     12:40:26   BIC                            1646.770
Sample:                                          0   HQIC                           1638.850
                                             - 132
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.S.L6       -1.0176      0.015    -68.669      0.000      -1.047      -0.989
ma.S.L6        0.0335      0.176      0.190      0.850      -0.312       0.379
ma.S.L12      -0.4659      0.081     -5.771      0.000      -0.624      -0.308
ma.S.L18       0.0764      0.164      0.465      0.642      -0.246       0.399
sigma2      2.608e+05   2.85e+04      9.148      0.000    2.05e+05    3.17e+05
===================================================================================
Ljung-Box (L1) (Q):                  15.59   Jarque-Bera (JB):                33.69
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.72   Skew:                             0.68
Prob(H) (two-sided):                  0.34   Kurtosis:                         5.41
===================================================================================
```



Predict on the Test Set using this model and evaluate the model.

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	907.444637	510.716558	-93.541423	1908.430696
1	530.133636	722.261572	-885.473033	1945.740305
2	1125.542692	884.585866	-608.213748	2859.299131
3	933.852547	1021.431609	-1068.116619	2935.821712
4	743.729140	1141.995143	-1494.540211	2981.998492

	Test RMSE
ARIMA(2,1,1)	1418.202085
ARIMA(0,1,0)	4779.154299
SARIMA(0,1,2)(2,0,2,6)	601.272415
SARIMA(1,1,2)(2,0,2,12)	548.017675
SARIMA(0,1,0)(1,1,3,6)	1914.603838

- The values are listed in the above table here we have considered seasonality as 6 and we can see the best performance is noticed at seasonality 12
- We can see the model is best performing at seasonality 12 and 601 at seasonality 6 which also is better than the other models.

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.

Answer

	Test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315

SARIMA(1,1,2)(2,0,2,12)	548.017675
SARIMA(0,1,2)(2,0,2,6)	601.272415
ARIMA(2,1,1)	1418.202085
SARIMA(0,1,0)(1,1,3,6)	1914.603838
ARIMA(0,1,0)	4779.154299

- In the above table we have listed all the models to check which is the best performing model.
- The table shows SARIMA at seasonality 12 of 548.01 and 601.27 at seasonality 6. ‘
- The moving average model is comparatively performing better than the LR, Naïve model with 813.40
- Hence in order to build the best mode we will be considering the SARIMA model with seasonality 12.

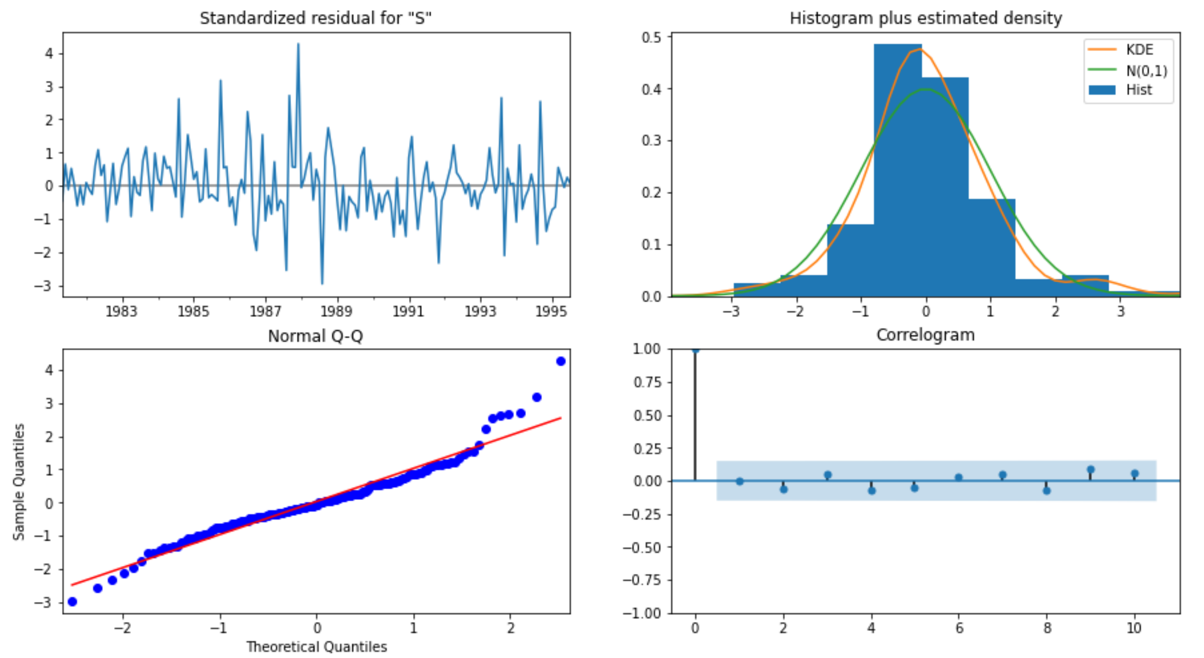
9) Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

Answer:

SARIMAX model:

```
                               SARIMAX Results
==========================================================================================
Dep. Variable:                           Sparkling   No. Observations:                 187
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 6)   Log Likelihood               -1258.196
Date:                             Tue, 06 Apr 2021   AIC                           2530.392
Time:                                     12:40:27   BIC                           2552.384
Sample:                                 01-31-1980   HQIC                          2539.315
                                      - 07-31-1995
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.8220      0.077    -10.657      0.000      -0.973      -0.671
ma.L2         -0.1099      0.079     -1.384      0.167      -0.266       0.046
ar.S.L6        0.0071      0.018      0.408      0.683      -0.027       0.041
ar.S.L12       1.0170      0.012     87.912      0.000       0.994       1.040
ma.S.L6       -0.0482      0.087     -0.556      0.578      -0.218       0.122
ma.S.L12      -0.6362      0.068     -9.325      0.000      -0.770      -0.502
sigma2      1.388e+05   1.09e+04     12.711      0.000    1.17e+05     1.6e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                56.45
Prob(Q):                              0.97   Prob(JB):                         0.00
Heteroskedasticity (H):               1.24   Skew:                             0.62
Prob(H) (two-sided):                  0.42   Kurtosis:                         5.52
===================================================================================
```

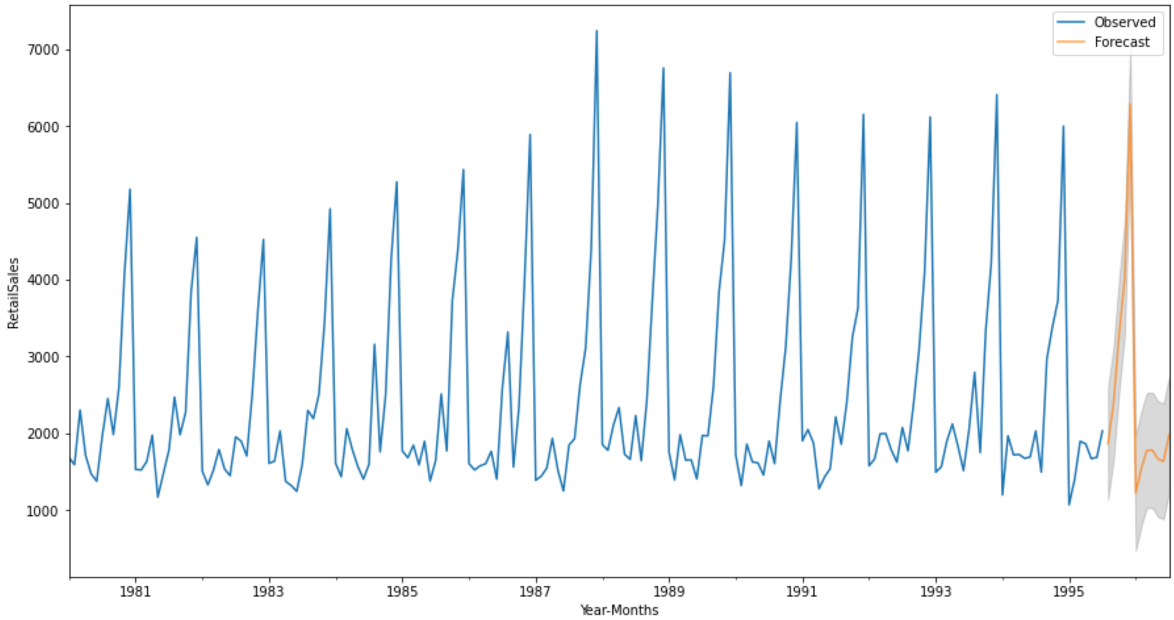
Fig 26. Different graphs for standard residual for Y



Sparkling	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-31	1864.155296	372.561946	1133.947301	2594.363292
1995-09-30	2393.415652	378.421515	1651.723111	3135.108192
1995-10-31	3285.359346	379.271120	2542.001609	4028.717083
1995-11-30	4017.456507	380.118832	3272.437285	4762.475728
1995-12-31	6286.072721	380.964664	5539.395700	7032.749742

Full model RMSE:

```
RMSE of the Full Model 531.9798801443804
```



	Test RMSE
ARIMA(2,1,1)	1418.202085
ARIMA(0,1,0)	4779.154299
SARIMA(0,1,2)(2,0,2,6)	601.272415
SARIMA(1,1,2)(2,0,2,12)	548.017675
SARIMA(0,1,0)(1,1,3,6)	1914.603838

- We can see the full model RMSE is 531.97 which is good enough we can understand the model is not performing at its best however we can consider it to check the sparkling wine sales.

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

Answer:

- The company should be using the SARIMA model with seasonality 12 to check the sales of the Sparkling wine.
- The best measures company should be considering is it to make sure the sales are published by different sources of media options.
- We know in order to get the customer base we need to be manufacturing the best quality wines with affordable prices.
- Having expensive wine reduces the sale as not everyone can afford such expensive drinks.
- The quality of the wines should be maintained.
- As we know the older the drink the expensive it is. We need to make sure there is enough stock which is saved in the backend to supply when necessary.
- Proper testing and tasting should happen to understand the likes an dislikes of the drink.
- We can use the above SARIMA model and focus on the customer base depending on the seasonal data.