TIME SERIES FORECASTING PROJECT

For this particular assignment, the data of different types of wine sales in the 20th century is to be analyzed. 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.

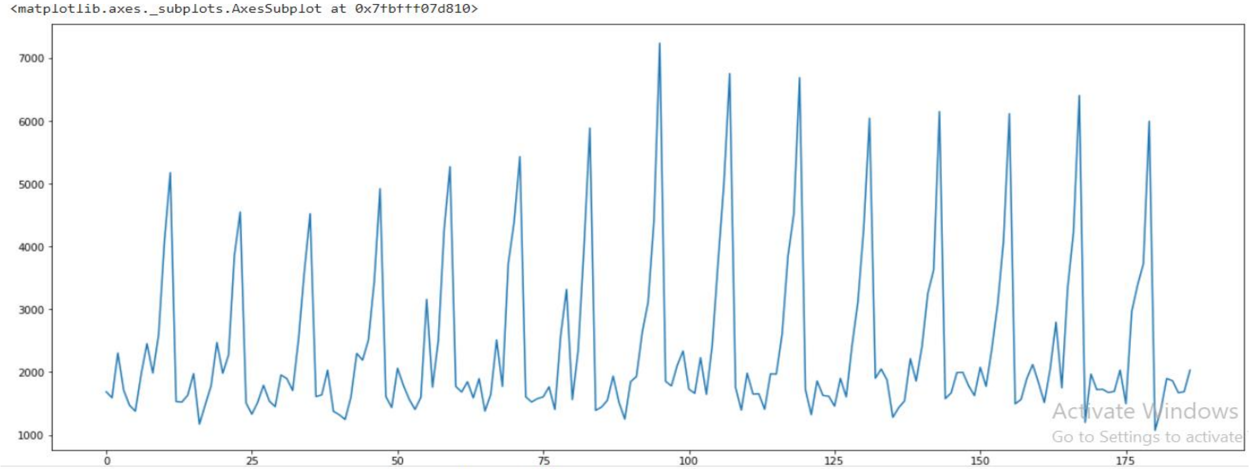
SPARKLING.CSV

1. Read the data as an appropriate Time Series data and plot the data.
   Time Series is a sequence of observations recorded at regular time intervals.

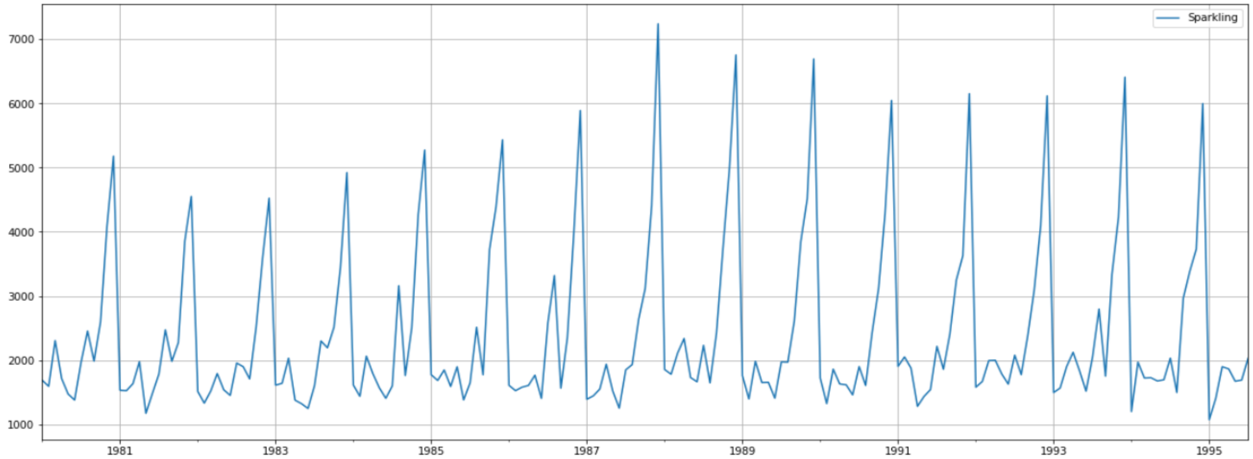
Time_Stamp	YearMonth	Sparkling
1980-01-31	1980-01	1686
1980-02-29	1980-02	1591
1980-03-31	1980-03	2304
1980-04-30	1980-04	1712
1980-05-31	1980-05	1471

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

<matplotlib.axes._subplots.AxesSubplot at 0x7fbfff07d810>



Given data is not time. So we parse the date range and create a timestamp. We also notice the increasing trend in the initial years.

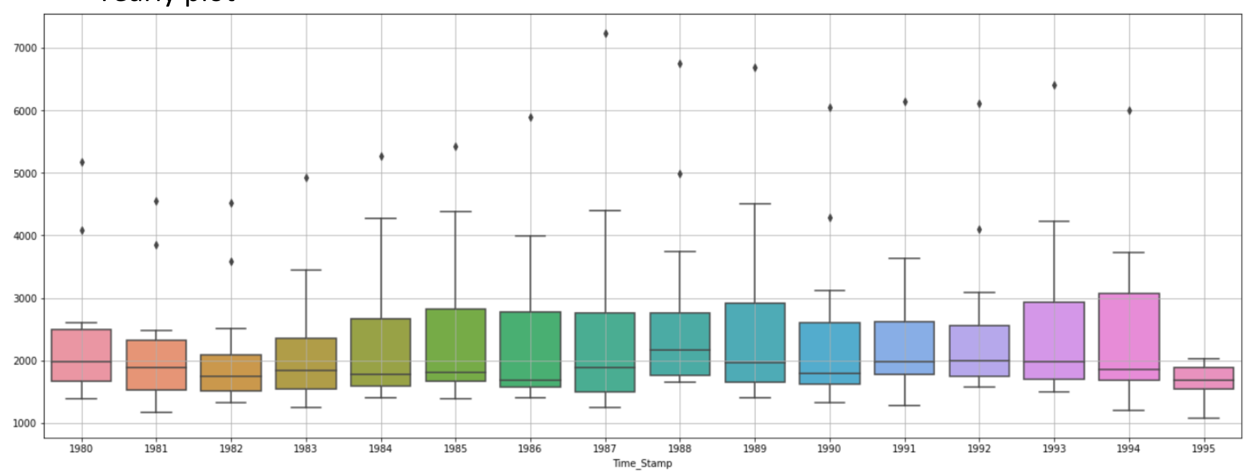


- Data consist of 187 data points
- It seems to be contain seasonality

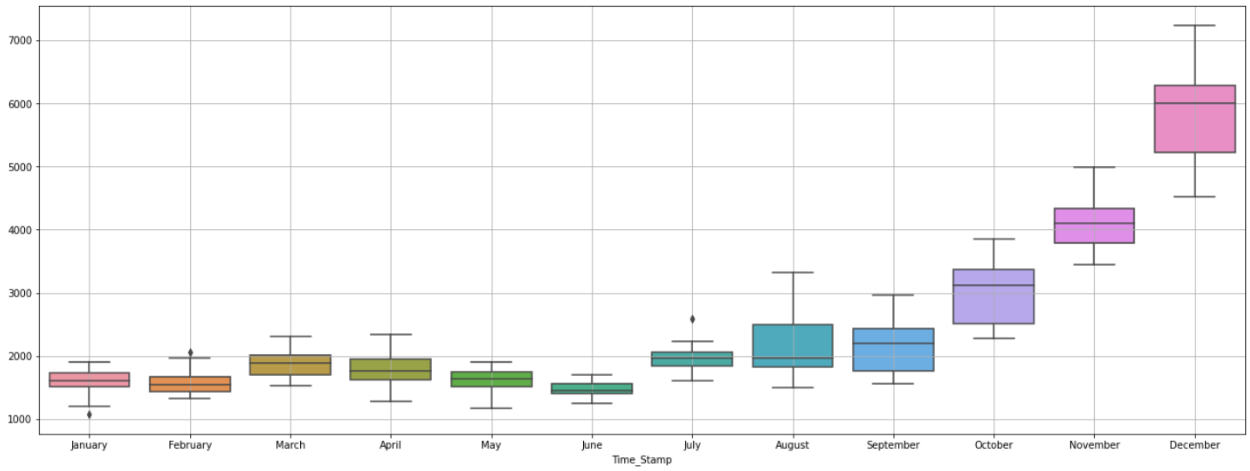
2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

From 1981 to 1988 there is an increase in the sparkling data. After that, there is a decrease or fall. Seasonality is seen from the stable fluctuations repeating over the data.

- To understand the spread of the data, we use plotting.
- Boxplot helps to check the outliers in each year and month.
- Yearly plot

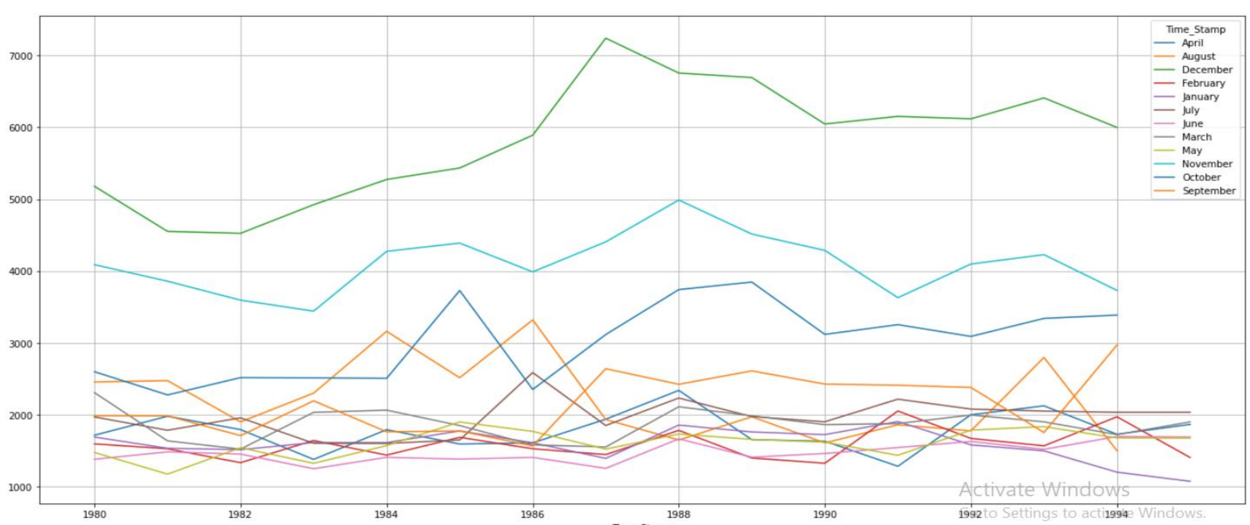


- Boxplot indicates the trend being present in the data.
- We can clearly see some of the outliers in the plot.
- Monthly plot
- The box plot for various months is plotted
- Monthly plot contains outliers in the month of January, February and July.

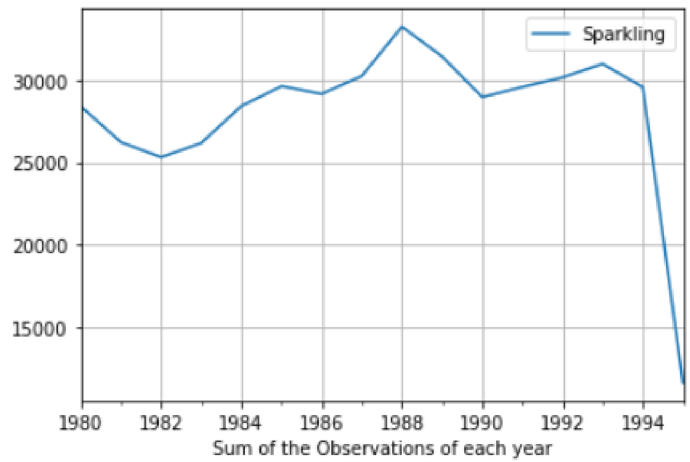


- Plot for different months and different years

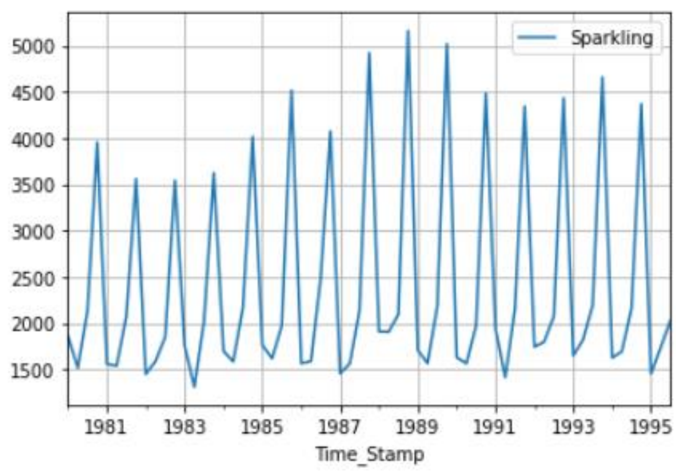
Time_Stamp	April	August	December	February	January	July	June	March	May	November	October	September
Time_Stamp												
1980	1712.0	2453.0	5179.0	1591.0	1686.0	1966.0	1377.0	2304.0	1471.0	4087.0	2596.0	1984.0
1981	1976.0	2472.0	4551.0	1523.0	1530.0	1781.0	1480.0	1633.0	1170.0	3857.0	2273.0	1981.0
1982	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



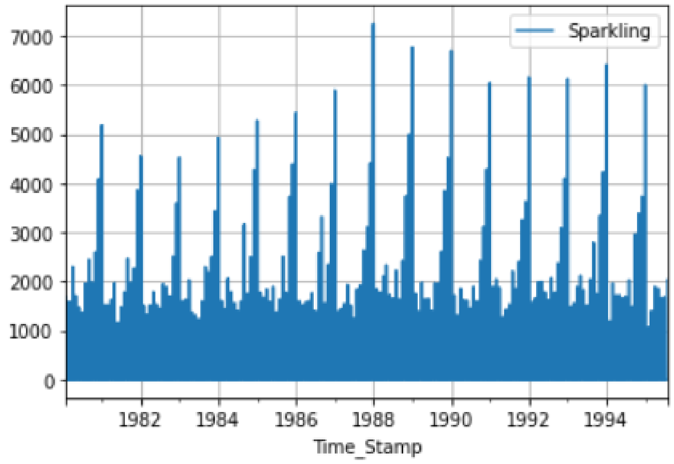
- December records have the high number of wine sales followed by November and October.
- May, June and July have low number of wine sales.
- Yearly Plot – aggregate the time series from an annual perspective and summing up the observations



- The plot shows that in 1982 there is a fall in the wine sales and a rise in 1984 and fall in 1986 followed by a maximum rise in 1988. In 1993 there is an increase and in 1994 there is a steep downfall is observed.
- Quarterly plot – aggregate the time series from a quarterly perspective and sum the observations of each quarter.

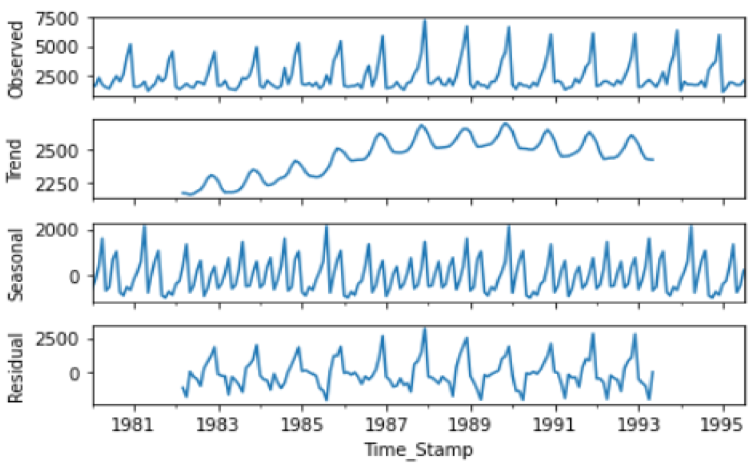


- High rise is found in 1988
- Daily plot –aggregate the data from a daily perspective
- Resampling can also be used for better overview



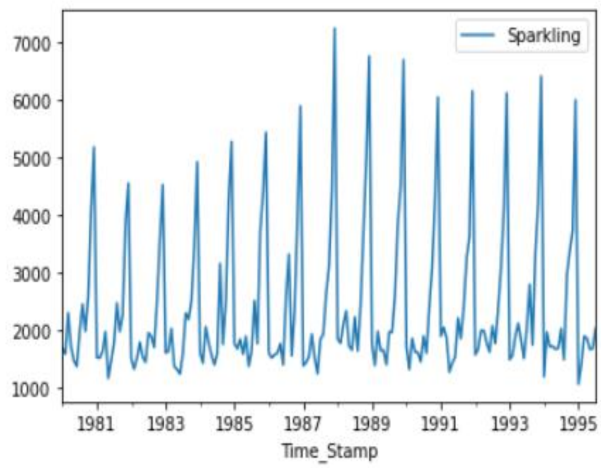

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 187 entries, 0 to 186
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   YearMonth   187 non-null    object
 1   Sparkling   187 non-null    int64
 2   Time_Stamp  187 non-null    datetime64[ns]
dtypes: datetime64[ns](1), int64(1), object(1)
memory usage: 4.5+ KB
```

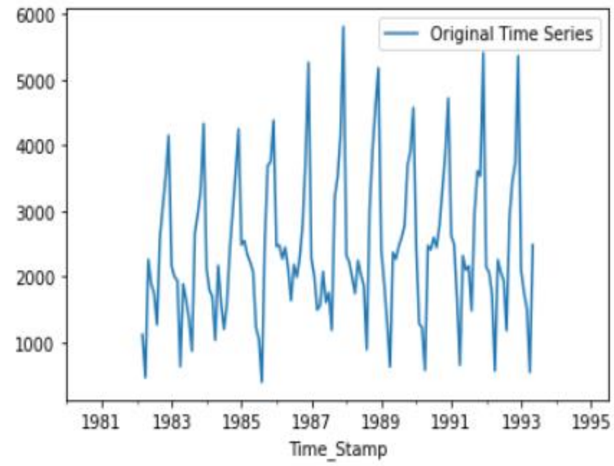
	Sparkling
count	187.000000
mean	2402.417112
std	1295.111540
min	1070.000000
25%	1605.000000
50%	1874.000000
75%	2549.000000
max	7242.000000



- From the decomposition, there is seasonality in the data.

```
Seasonality
                Sparkling
Time_Stamp
1980-01-31   -599.686853
1980-02-29   -192.465699
1980-03-31    398.543916
1980-04-30   1582.018275
1980-05-31   -648.994545
1980-06-30   -496.103519
1980-07-31    686.684942
1980-08-31   1027.184942
1980-09-30   -711.002558
1980-10-31   -860.050635
1980-11-30   -486.569866
1980-12-31   -628.170827
```

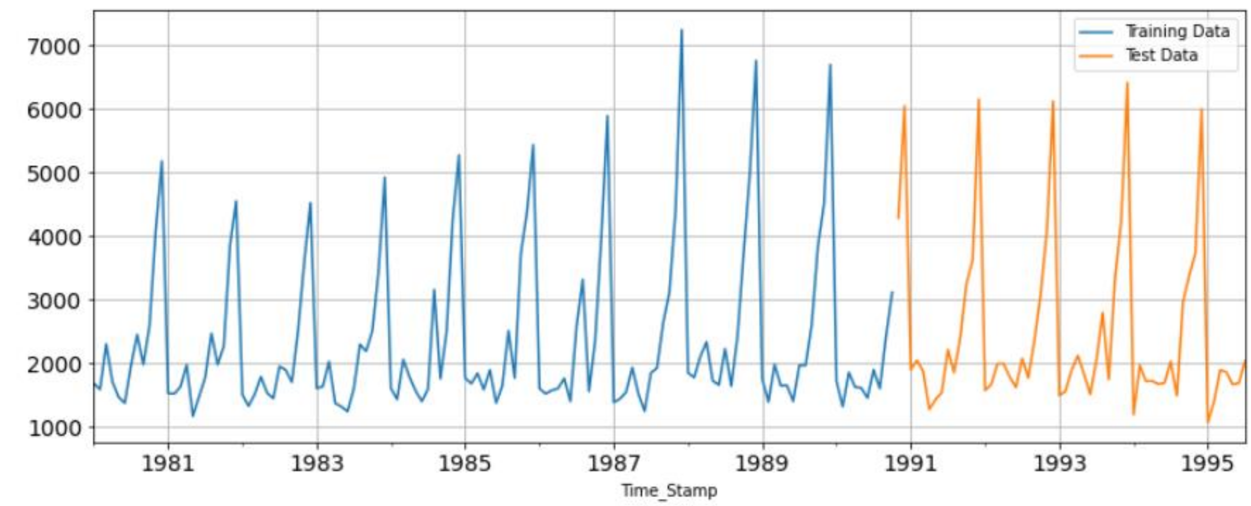




3. Split the data into training and test. The test data should start in 1991.

Train and test shapes

```
(130, 1)
(57, 1)
```



- The test data starts from 1991
- It is difficult to predict the future if the past is not happened. From the above split, we are predicting similar to the past data.

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.

Model1: Linear Regression

- Regress the “Sparkling” variable against the order of occurrence.

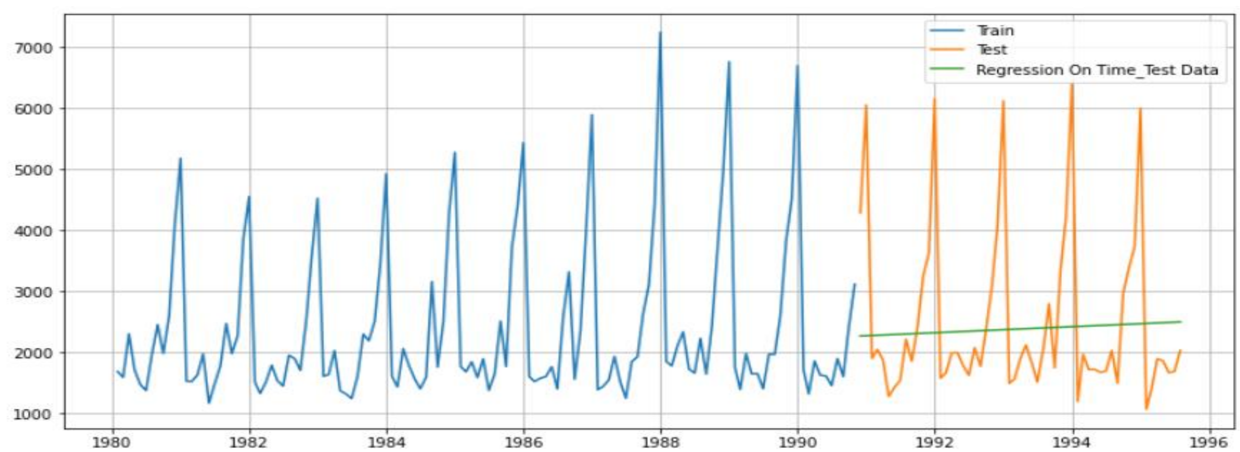
- Modifying the training set
- Generate the numerical instance order for both training and test set
- Printing the head and tail of test and train data

```
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
1990-06-30       1457   126
1990-07-31       1899   127
1990-08-31       1605   128
1990-09-30       2424   129
1990-10-31       3116   130
```

```
First few rows of Test Data
            Sparkling  time
Time_Stamp
1990-11-30       4286    43
1990-12-31       6047    44
1991-01-31       1902    45
1991-02-28       2049    46
1991-03-31       1874    47

Last few rows of Test Data
            Sparkling  time
Time_Stamp
1995-03-31       1897    95
1995-04-30       1862    96
1995-05-31       1670    97
1995-06-30       1688    98
1995-07-31       2031    99
```

- Linear Regression is built on the training and test dataset

1995-01-31	1070	93	2474.280747
1995-02-28	1402	94	2478.389977
1995-03-31	1897	95	2482.499207
1995-04-30	1862	96	2486.608437
1995-05-31	1670	97	2490.717666
1995-06-30	1688	98	2494.826896
1995-07-31	2031	99	2498.936126



- Defining the accuracy metrics
- Evaluating the model

```
For RegressionOnTime forecast on the Test Data,  RMSE is 1374.550
```

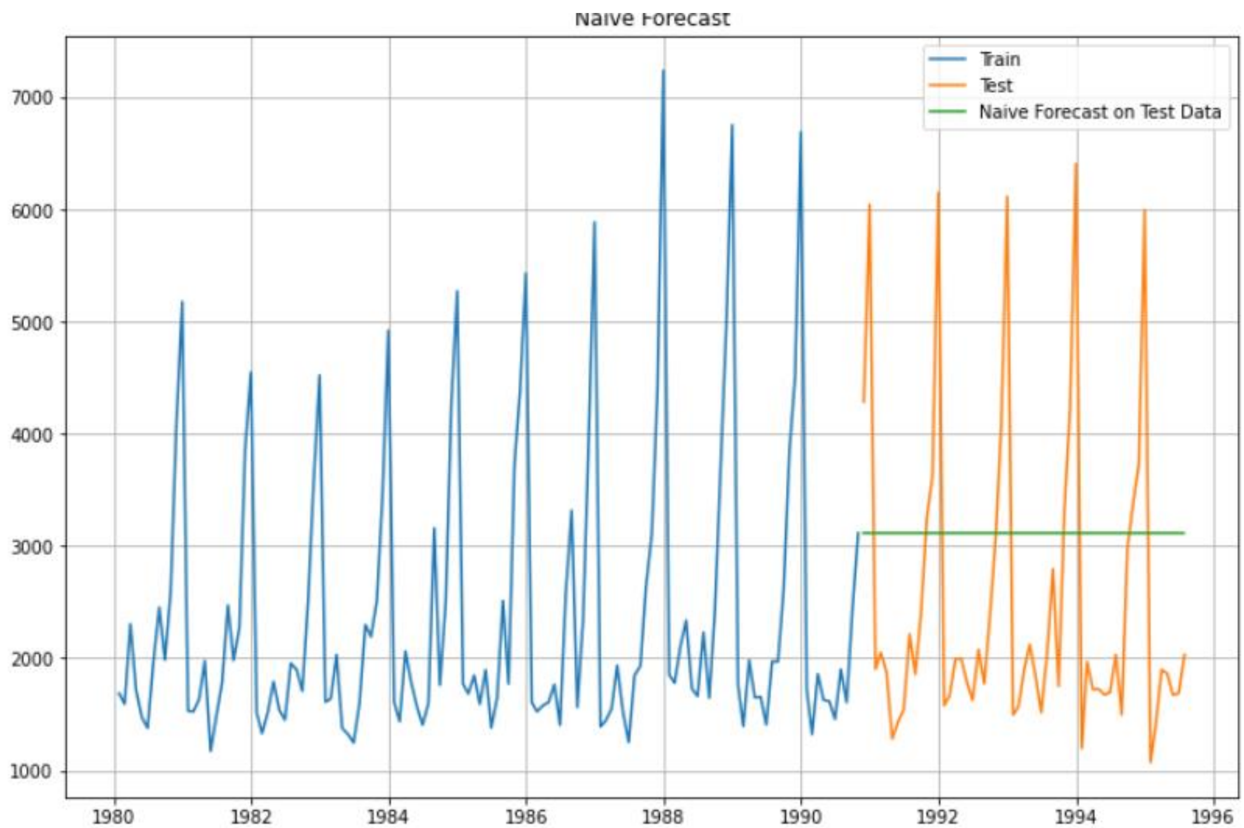
	Test RMSE
RegressionOnTime	1374.550202

Model2 – Naïve model

We say that the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow

```
129
```

```
Time_Stamp
1990-11-30    3116
1990-12-31    3116
1991-01-31    3116
1991-02-28    3116
1991-03-31    3116
Name: naive, dtype: int64
```

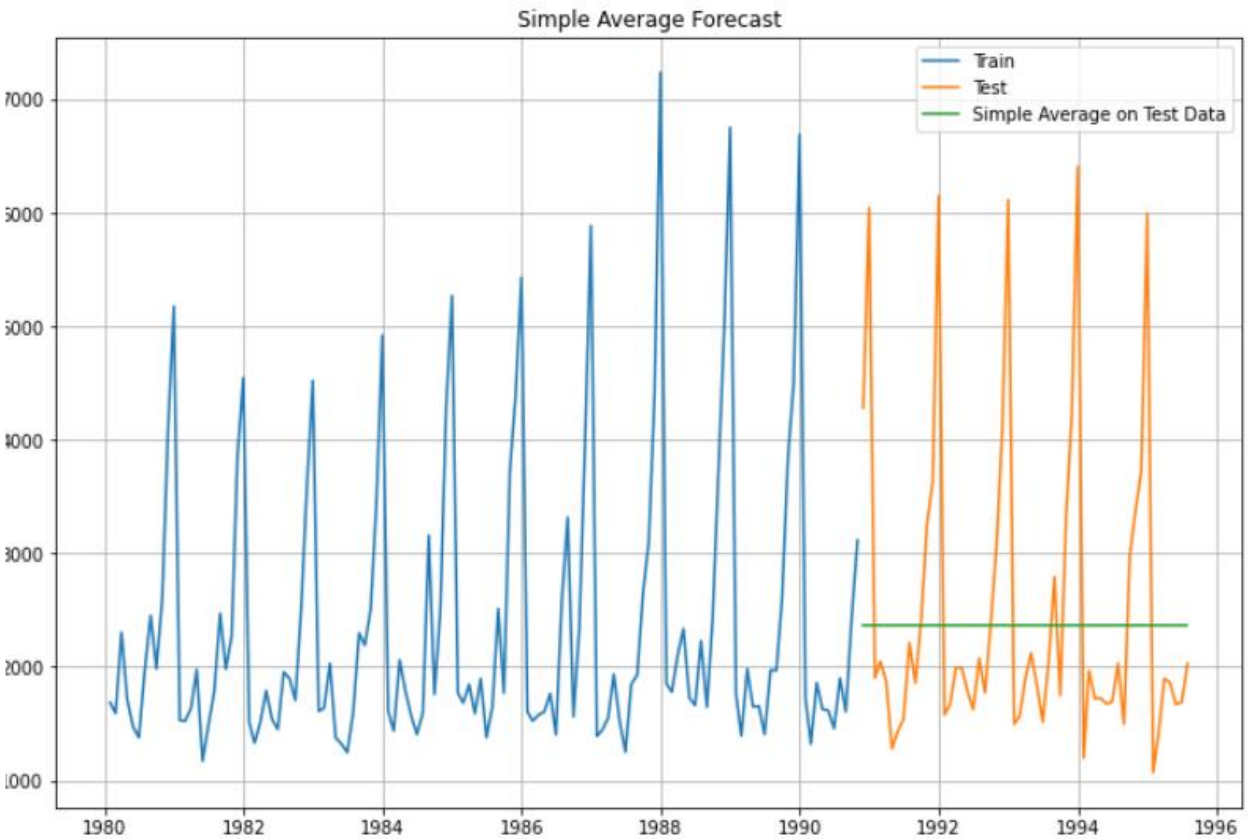


```
For RegressionOnTime forecast on the Test Data,  RMSE is 1496.445
```

	Test RMSE
RegressionOnTime	1374.550202
NaiveModel	1496.444629

Model3 – Simple Average – Forecast using the average of training values

Time_Stamp	Sparkling	mean_forecast
1990-11-30	4286	2361.276923
1990-12-31	6047	2361.276923
1991-01-31	1902	2361.276923
1991-02-28	2049	2361.276923
1991-03-31	1874	2361.276923

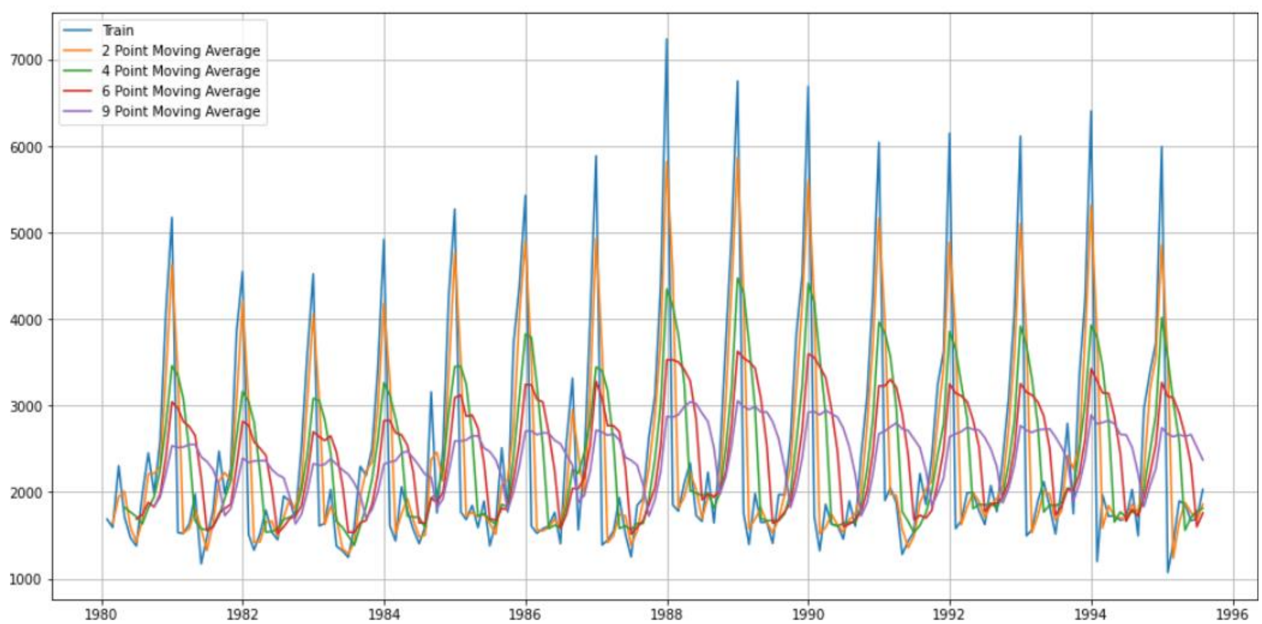



```
For Simple Average forecast on the Test Data,  RMSE is 1368.747
```

	Test RMSE
RegressionOnTime	1374.550202
NaiveModel	1496.444629
SimpleAverageModel	1368.746717

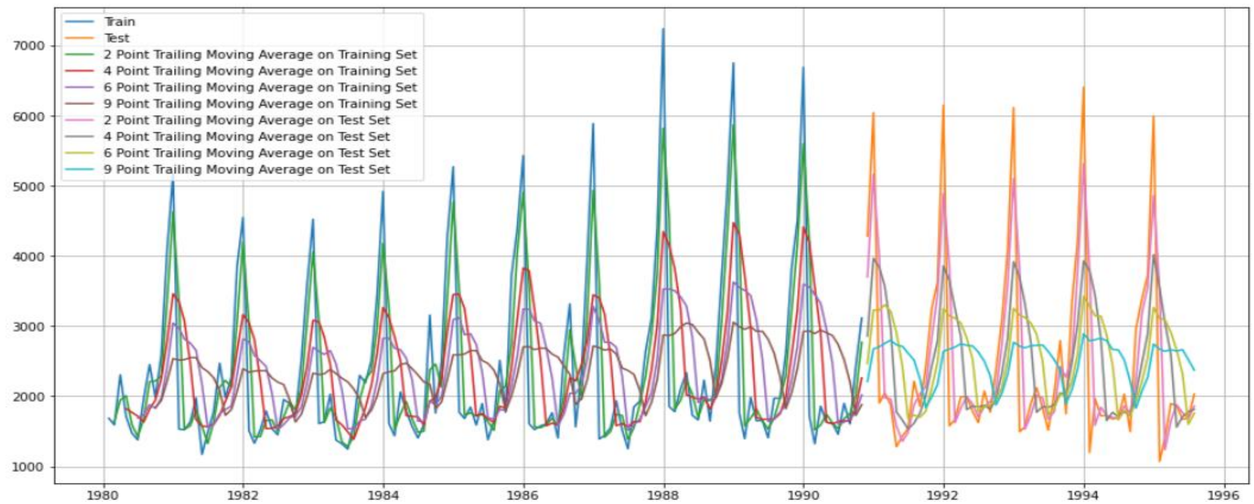
Model4- Moving Average – Calculating the rolling means (or moving averages) for different intervals. The best interval can be determined by the maximum accuracy (or the minimum error) over here.

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



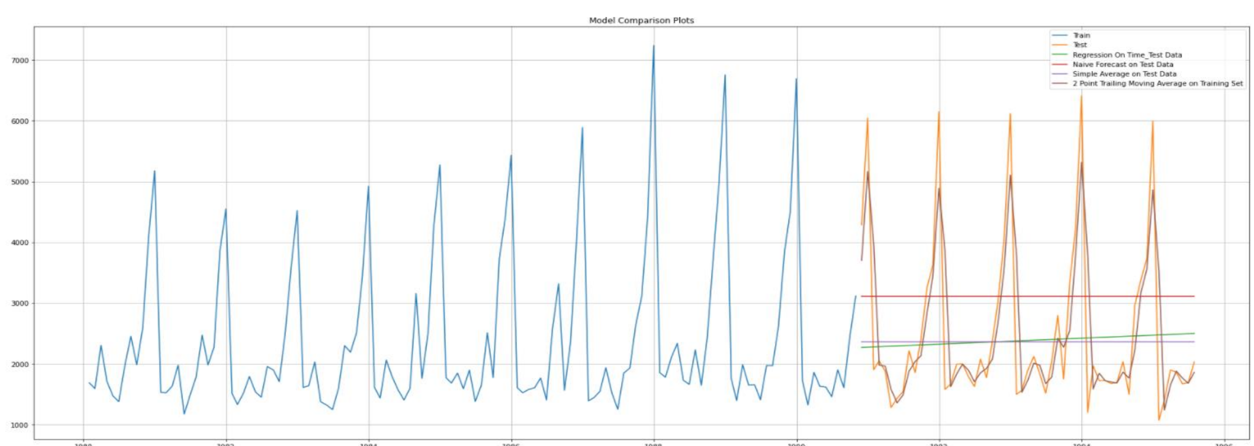

```
For 2 point Moving Average Model forecast on the Training Data,  RMSE is 811.179
For 4 point Moving Average Model forecast on the Training Data,  RMSE is 1184.213
For 6 point Moving Average Model forecast on the Training Data,  RMSE is 1337.201
For 9 point Moving Average Model forecast on the Training Data,  RMSE is 1422.653
```

Let us split the data into train and test and plot this Time Series. The window of the moving average is need to be carefully selected as too big a window will result in not having any test set as the whole series might get averaged over.



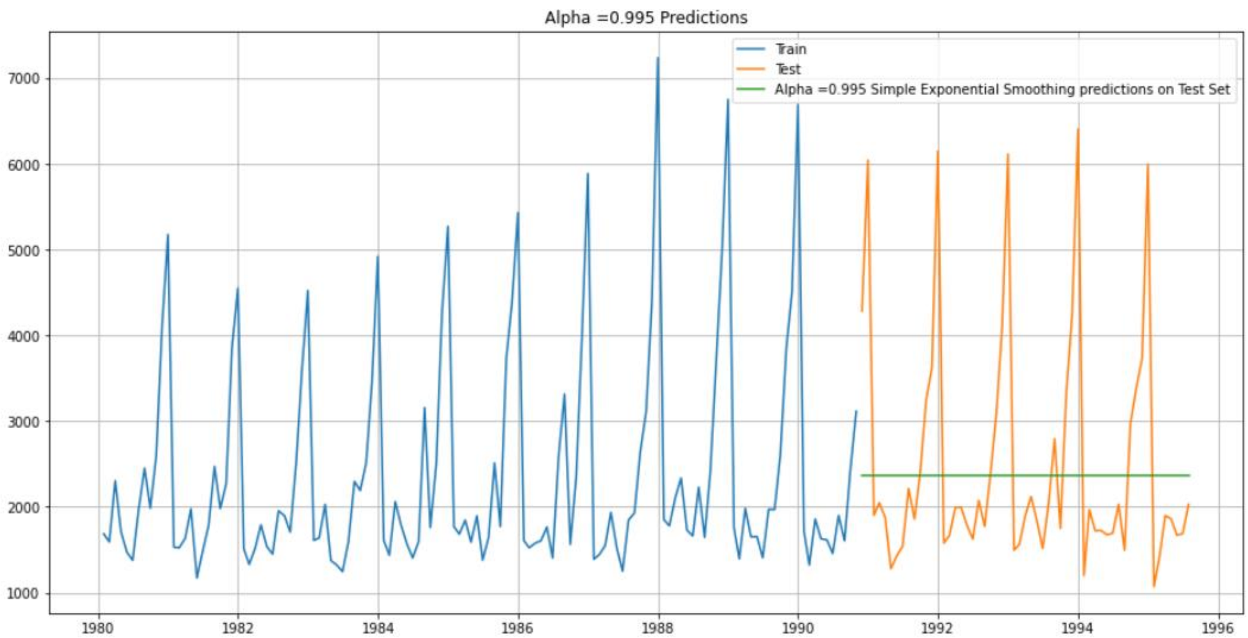
	Test RMSE
RegressionOnTime	1374.550202
NaiveModel	1496.444629
SimpleAverageModel	1368.746717
2pointTrailingMovingAverage	811.178937
4pointTrailingMovingAverage	1184.213295
6pointTrailingMovingAverage	1337.200524
9pointTrailingMovingAverage	1422.653281

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



Model -5- Exponential Smoothing

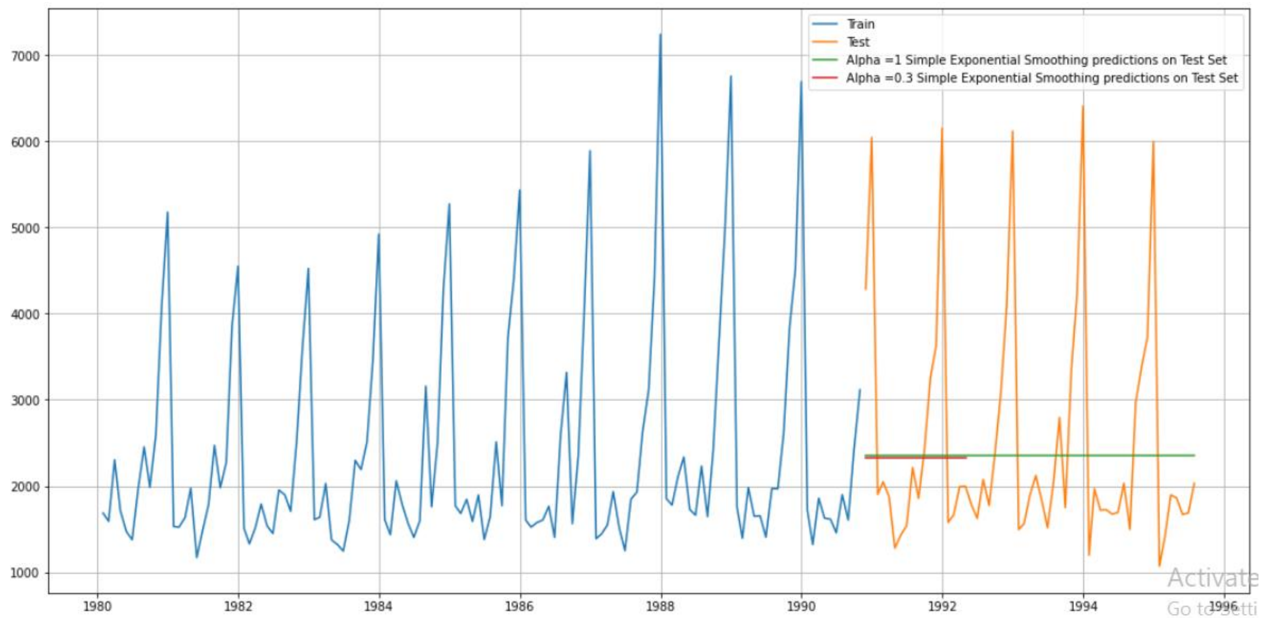
Time_Stamp	Sparkling	predict
1990-11-30	4286	2361.278901
1990-12-31	6047	2361.278901
1991-01-31	1902	2361.278901
1991-02-28	2049	2361.278901
1991-03-31	1874	2361.278901

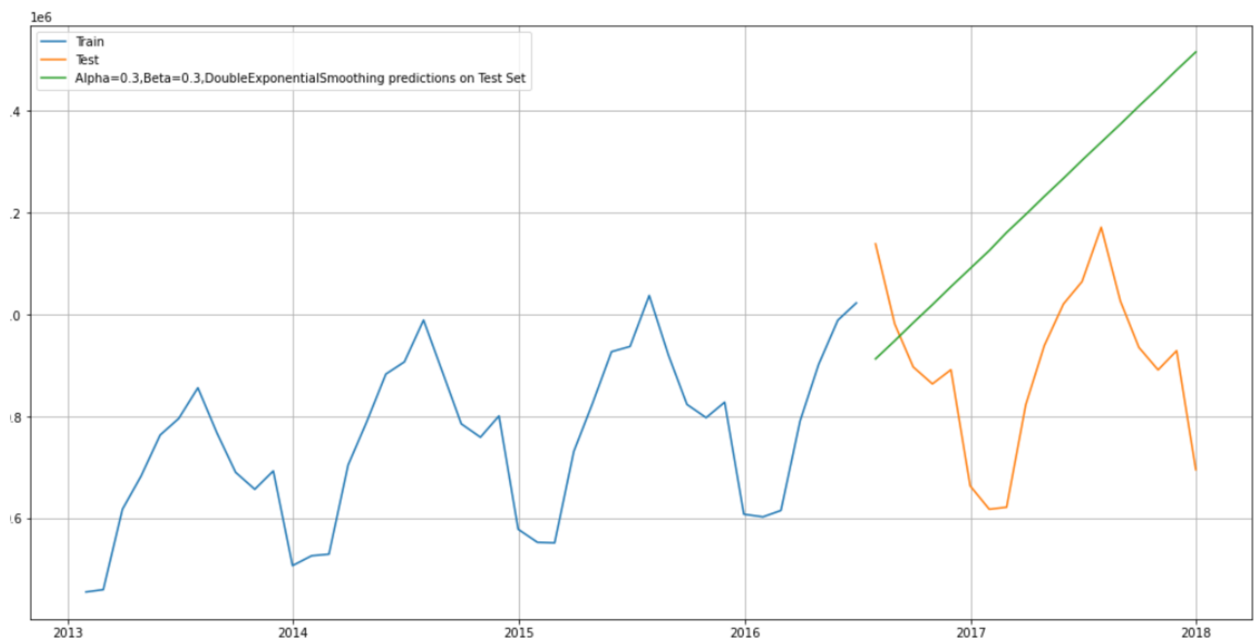


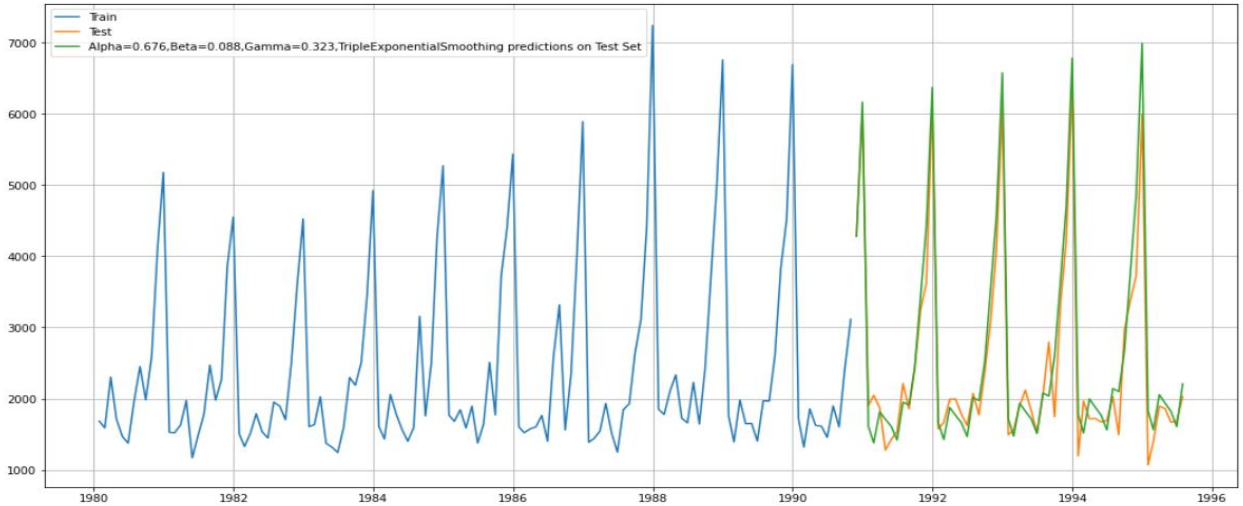

```
For Alpha =0.995 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 1368.747
```

	Test RMSE
RegressionOnTime	1374.550202
NaiveModel	1496.444629
SimpleAverageModel	1368.746717
2pointTrailingMovingAverage	811.178937
4pointTrailingMovingAverage	1184.213295
6pointTrailingMovingAverage	1337.200524
9pointTrailingMovingAverage	1422.653281
Alpha=0.995,SimpleExponentialSmoothing	1368.746522

Setting different alpha values. Higher the alpha, the more weightage is given to more recent observation.



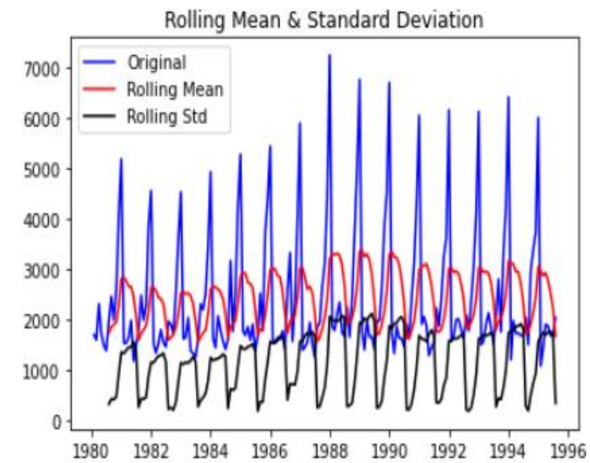




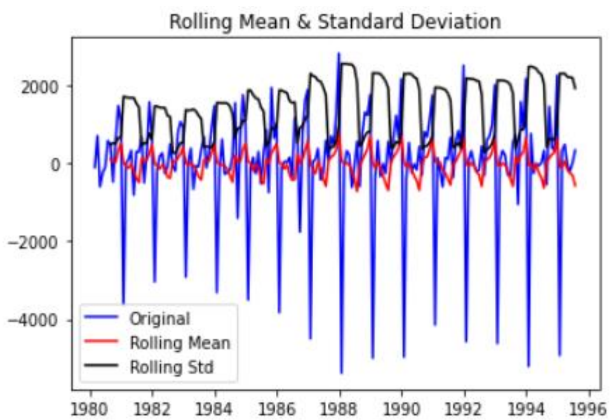
	Test RMSE
RegressionOnTime	1374.550202
NaiveModel	1496.444629
SimpleAverageModel	1368.746717
2pointTrailingMovingAverage	811.178937
4pointTrailingMovingAverage	1184.213295
6pointTrailingMovingAverage	1337.200524
9pointTrailingMovingAverage	1422.653281
Alpha=0.995,SimpleExponentialSmoothing	1368.746522
Alpha=0.676,Beta=0.088,Gamma=0.323,TripleExponentialSmoothing	388.974278

	Test RMSE
Alpha=0.676,Beta=0.088,Gamma=0.323,TripleExponentialSmoothing	388.974278
2pointTrailingMovingAverage	811.178937
4pointTrailingMovingAverage	1184.213295
6pointTrailingMovingAverage	1337.200524
Alpha=0.995,SimpleExponentialSmoothing	1368.746522
Alpha=0.995,SimpleExponentialSmoothing	1368.746522
SimpleAverageModel	1368.746717
RegressionOnTime	1374.550202
9pointTrailingMovingAverage	1422.653281
NaiveModel	1496.444629

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.




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



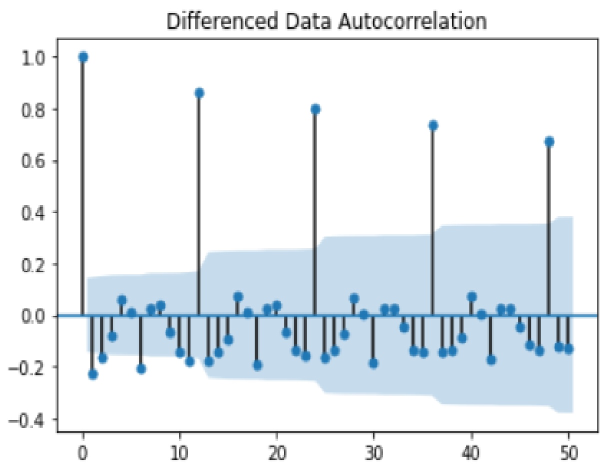

```
Results of Dickey-Fuller Test:
Test Statistic                 -45.050301
p-value                          0.000000
#Lags Used                      10.000000
Number of Observations Used    175.000000
Critical Value (1%)             -3.468280
Critical Value (5%)             -2.878202
Critical Value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

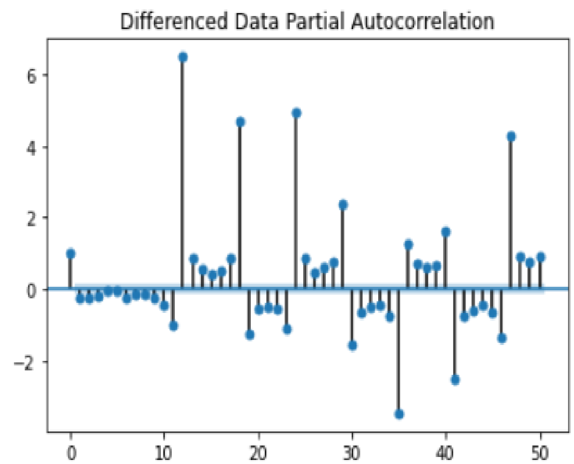
- When the time series data is not stationary we need to convert it into stationary before applying models.
- We use Augmented Dickey fuller test.
- It determines how strongly a time series is defined by the trend.
- From the null and alternate hypothesis, we define time series data is stationary or not.
- We see that 5% significant level the time series is non-stationarity
- P value >0.05 – Failed to reject null hypothesis - Stationary
- Let us take a difference of order 1 and check whether the Time Series is stationary or not
- At α = 0.05 the Time Series is indeed stationary.

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

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.

```
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)
```





- From the above plot, we see seasonality in the data.

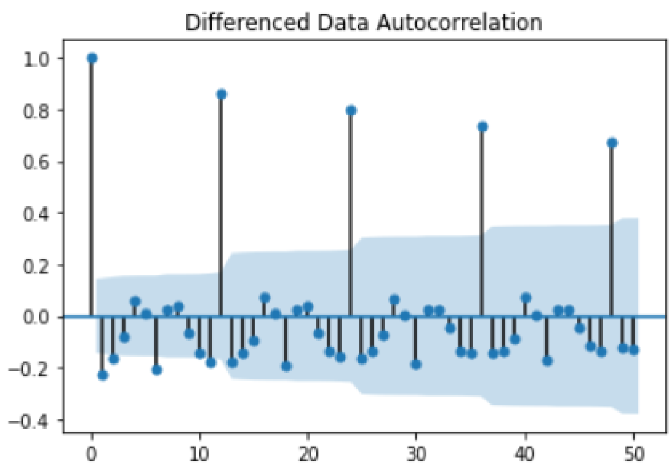
```
ARIMA(0, 1, 0) - AIC:2234.707214654729
ARIMA(0, 1, 1) - AIC:2228.15037491062
ARIMA(0, 1, 2) - AIC:2193.882351278575
ARIMA(1, 1, 0) - AIC:2233.142091199552
ARIMA(1, 1, 1) - AIC:2196.4628372747275
ARIMA(1, 1, 2) - AIC:2195.0953659379893
ARIMA(2, 1, 0) - AIC:2225.660614396246
ARIMA(2, 1, 1) - AIC:2193.86860259207
ARIMA(2, 1, 2) - AIC:2175.566357052884
```

	param	AIC
8	(2, 1, 2)	2175.566357
7	(2, 1, 1)	2193.868603
2	(0, 1, 2)	2193.882351
5	(1, 1, 2)	2195.095366
4	(1, 1, 1)	2196.462837
6	(2, 1, 0)	2225.660614
1	(0, 1, 1)	2228.150375
3	(1, 1, 0)	2233.142091
0	(0, 1, 0)	2234.707215

- If we have seasonality, then we should go for SARIMA model.
- We are building ARIMA model by looking at minimum AIC values and ACF and PACF plots.
- Sorting the AIC values to see the lower AIC value.

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:            D.Sparkling   No. Observations:                  129
Model:                 ARIMA(2, 1, 1)   Log Likelihood               -1091.934
Method:                       css-mle   S.D. of innovations           1128.325
Date:                Fri, 02 Jul 2021   AIC                           2193.869
Time:                        05:07:07   BIC                           2208.168
Sample:                    02-29-1980   HQIC                          2199.679
                         - 10-31-1990
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 4.2096      3.701      1.138      0.257      -3.044      11.463
ar.L1.D.Sparkling     0.4762      0.086      5.512      0.000       0.307       0.645
ar.L2.D.Sparkling    -0.1865      0.086     -2.168      0.032      -0.355      -0.018
ma.L1.D.Sparkling    -1.0000      0.020    -50.418      0.000      -1.039      -0.961
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.2766          -1.9319j            2.3156           -0.1571
AR.2            1.2766          +1.9319j            2.3156            0.1571
MA.1            1.0000          +0.0000j            1.0000            0.0000
-----------------------------------------------------------------------------
```

	RMSE
ARIMA(2,1,1)	1386.71183



- Again we plot ACF to see and understand the seasonal parameter of SARIMA model.
- We see seasonality in 6 as well as 12.
- We run SARIMA model by setting seasonality both as 6 and 12.
- First iteration by setting 6 as the seasonality
- We sort the AIC values to see the lowest of all vales.
- Next predicting the data using the SARIMA model and evaluating the model.
- We get the summary of the data

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

```
SARIMA(2, 1, 1)x(2, 0, 0, 6) - AIC:1731.8137132625177
SARIMA(2, 1, 1)x(2, 0, 1, 6) - AIC:1733.7098303257294
SARIMA(2, 1, 1)x(2, 0, 2, 6) - AIC:1710.4498904719412
SARIMA(2, 1, 2)x(0, 0, 0, 6) - AIC:2140.669395942271
SARIMA(2, 1, 2)x(0, 0, 1, 6) - AIC:2042.7000095525877
SARIMA(2, 1, 2)x(0, 0, 2, 6) - AIC:1850.8435518017636
SARIMA(2, 1, 2)x(1, 0, 0, 6) - AIC:2038.5071951549105
SARIMA(2, 1, 2)x(1, 0, 1, 6) - AIC:1927.69282782436
SARIMA(2, 1, 2)x(1, 0, 2, 6) - AIC:1796.3236930264027
```

	param	seasonal	AIC
53	(1, 1, 2)	(2, 0, 2, 6)	1694.325453
26	(0, 1, 2)	(2, 0, 2, 6)	1694.839212
80	(2, 1, 2)	(2, 0, 2, 6)	1695.565322
17	(0, 1, 1)	(2, 0, 2, 6)	1708.125767
44	(1, 1, 1)	(2, 0, 2, 6)	1710.045544

```
                                     Statespace Model Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  130
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 6)   Log Likelihood                -840.420
Date:                            Fri, 02 Jul 2021   AIC                           1694.839
Time:                                    05:11:33   BIC                           1713.993
Sample:                                         0   HQIC                          1702.612
                                            - 130
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -1.0171      0.153     -6.663      0.000      -1.316      -0.718
ma.L2         -0.0825      0.120     -0.685      0.493      -0.318       0.153
ar.S.L6        0.0073      0.022      0.326      0.744      -0.036       0.051
ar.S.L12       1.0571      0.017     62.698      0.000       1.024       1.090
ma.S.L6        0.0334      0.142      0.235      0.815      -0.245       0.312
ma.S.L12      -0.6723      0.086     -7.819      0.000      -0.841      -0.504
sigma2      1.187e+05    1.7e+04      6.990      0.000    8.55e+04    1.52e+05
===================================================================================
Ljung-Box (Q):                       25.24   Jarque-Bera (JB):                30.25
Prob(Q):                              0.97   Prob(JB):                         0.00
Heteroskedasticity (H):               2.99   Skew:                             0.44
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.37
===================================================================================
```

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	4756.012838	376.547622	4017.993060	5494.032616
1	7041.953214	381.352541	6294.515969	7789.390460
2	1569.005078	382.893759	818.547101	2319.463056
3	1246.112342	384.428713	492.645911	1999.578774
4	1805.475687	385.947855	1049.031792	2561.919583

	RMSE
ARIMA(2,1,1)	1386.711830
SARIMA(0,1,2)(2,0,2,6)	646.880691

- There is a huge gain in the RMSE value by including seasonal parameters
- Keeping 12 as seasonal parameter for second iteration

	param	seasonal	AIC
53	(1, 1, 2)	(2, 0, 2, 12)	1521.737955
50	(1, 1, 2)	(1, 0, 2, 12)	1521.949482
80	(2, 1, 2)	(2, 0, 2, 12)	1523.217832
77	(2, 1, 2)	(1, 0, 2, 12)	1523.524946
26	(0, 1, 2)	(2, 0, 2, 12)	1523.707298

```
                                     Statespace Model Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  130
Model:             SARIMAX(1, 1, 2)x(2, 0, 2, 12)   Log Likelihood                -752.869
Date:                            Fri, 02 Jul 2021   AIC                           1521.738
Time:                                    05:21:06   BIC                           1542.738
Sample:                                         0   HQIC                          1530.241
                                            - 130
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.6466      0.268     -2.415      0.016      -1.171      -0.122
ma.L1          0.2835      0.300      0.944      0.345      -0.305       0.872
ma.L2         -1.1683      0.331     -3.528      0.000      -1.817      -0.519
ar.S.L12       0.7532      0.508      1.482      0.138      -0.243       1.749
ar.S.L24       0.3249      0.541      0.601      0.548      -0.736       1.385
ma.S.L12      -0.9795      0.491     -1.997      0.046      -1.941      -0.018
ma.S.L24      -0.5626      0.670     -0.840      0.401      -1.876       0.750
sigma2      4.952e+04   2.51e+04      1.975      0.048     372.999    9.87e+04
===================================================================================
Ljung-Box (Q):                       20.68   Jarque-Bera (JB):                 8.05
Prob(Q):                              1.00   Prob(JB):                         0.02
Heteroskedasticity (H):               1.46   Skew:                             0.21
Prob(H) (two-sided):                  0.27   Kurtosis:                         4.31
===================================================================================
```

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	4695.855387	380.042427	3950.985918	5440.724857
1	7226.356271	388.895557	6464.134986	7988.577556
2	1584.427512	389.285007	821.442919	2347.412105
3	1417.348365	392.127133	648.793306	2185.903424
4	1828.714038	392.163488	1060.087725	2597.340351

8. Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

	RMSE
ARIMA(2,1,1)	1386.711830
SARIMA(0,1,2)(2,0,2,6)	646.880691
SARIMA(1,1,2)(2,0,2,12)	712.707894

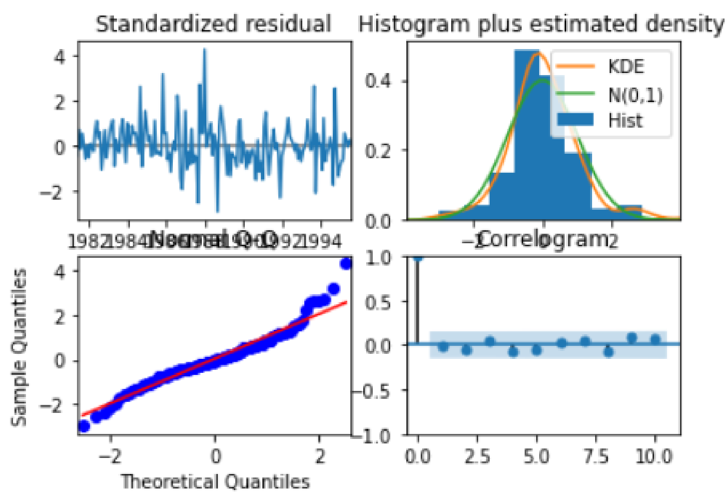
- It is clear that SARIMA(0,1,2)(2,0,2,6) has the lower RMSE and ARIMA(2,1,1) has the higher value.

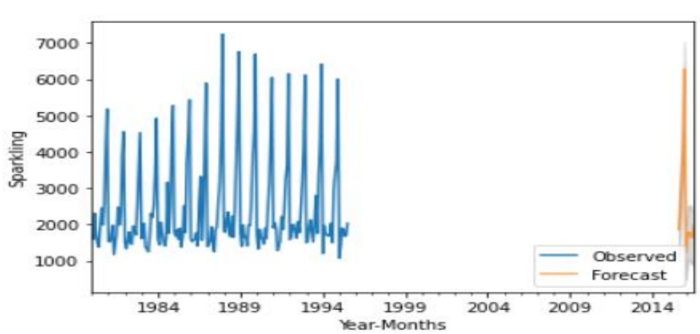
9. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

```
                                     Statespace Model Results
==========================================================================================
Dep. Variable:                          Sparkling   No. Observations:                  187
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 6)   Log Likelihood               -1258.200
Date:                            Fri, 02 Jul 2021   AIC                           2530.399
Time:                                    05:22:40   BIC                           2552.391
Sample:                                01-31-1980   HQIC                          2539.322
                                     - 07-31-1995
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.9475      0.106     -8.974      0.000      -1.154      -0.741
ma.L2         -0.1250      0.086     -1.445      0.148      -0.294       0.044
ar.S.L6        0.0073      0.018      0.414      0.679      -0.027       0.042
ar.S.L12       1.0171      0.012     87.817      0.000       0.994       1.040
ma.S.L6       -0.4433      0.105     -4.207      0.000      -0.650      -0.237
ma.S.L12      -0.9418      0.103     -9.141      0.000      -1.144      -0.740
sigma2      8.262e+04   1.24e+04      6.648      0.000    5.83e+04    1.07e+05
===================================================================================
Ljung-Box (Q):                       19.45   Jarque-Bera (JB):                56.39
Prob(Q):                              1.00   Prob(JB):                         0.00
Heteroskedasticity (H):               1.24   Skew:                             0.62
Prob(H) (two-sided):                  0.42   Kurtosis:                         5.52
===================================================================================
```

Sparkling	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-31	1863.656191	372.567001	1133.438287	2593.874095
1995-09-30	2393.459684	378.443212	1651.724617	3135.194750
1995-10-31	3285.461780	379.292653	2542.061841	4028.861719
1995-11-30	4017.185571	380.140202	3272.124465	4762.246676
1995-12-31	6286.473411	380.985873	5539.754821	7033.192001
1996-01-31	1220.495767	381.829677	472.123352	1968.868183
1996-02-29	1544.236886	381.967774	795.593805	2292.879967
1996-03-31	1777.915220	382.640494	1027.953633	2527.876807
1996-04-30	1781.404294	383.413983	1029.926696	2532.881892
1996-05-31	1665.476283	384.185923	912.485711	2418.466855
1996-06-30	1637.321580	384.956318	882.821062	2391.822098
1996-07-31	1980.909543	385.725179	1224.902085	2736.917002

```
RMSE of the Full Model 530.8502242211852
```





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

- To find the most optimum model, we run the model on the full data
- Correlogram, histogram, residual and quartiles are shown.
- We predict for the next 12 months for next years.
- We get forecast.
- RMSE of the full complete data is 530.8
- Plotting the forecast with the confidence band
- It is clear that SARIMA(0,1,2)(2,0,2,6) has the lower RMSE and ARIMA(2,1,1) has the higher value.