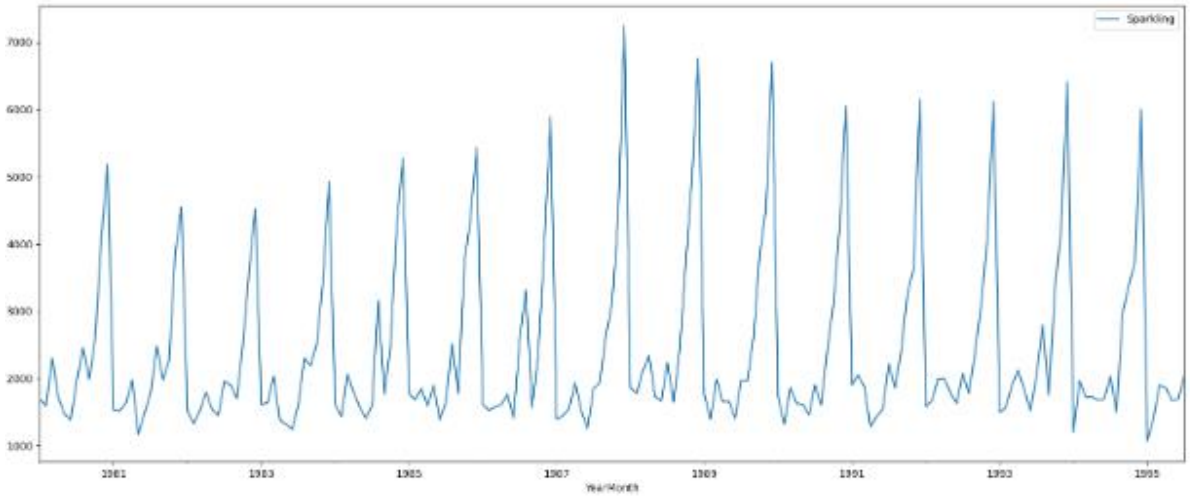
TIME SERIES FORECASTING PROJECT

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.

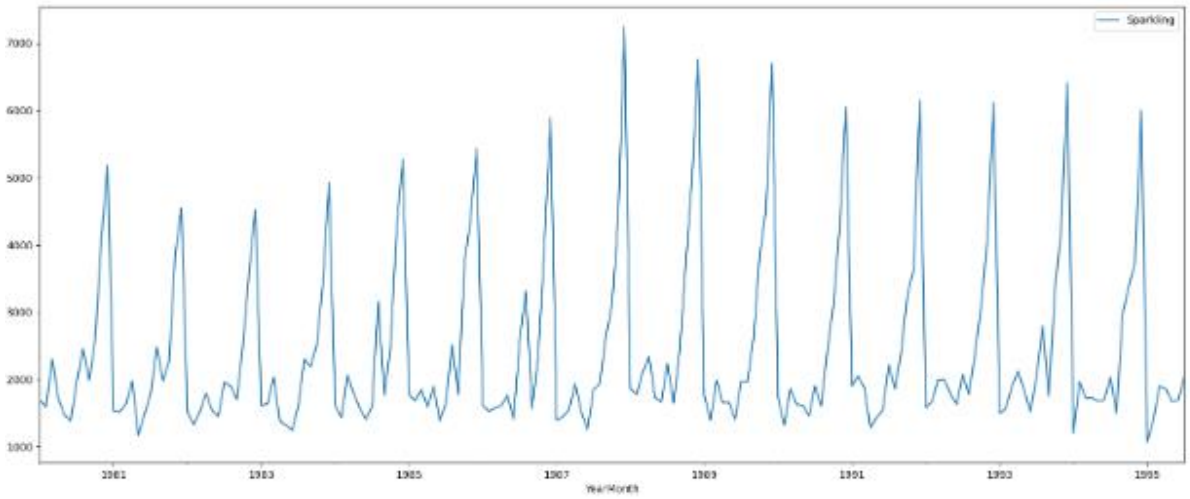
Sparkling.csv

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.

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



Given data is not time. So we pass the parse_date parameter to convert as timestamp. We also notice the increasing trend in the initial years.



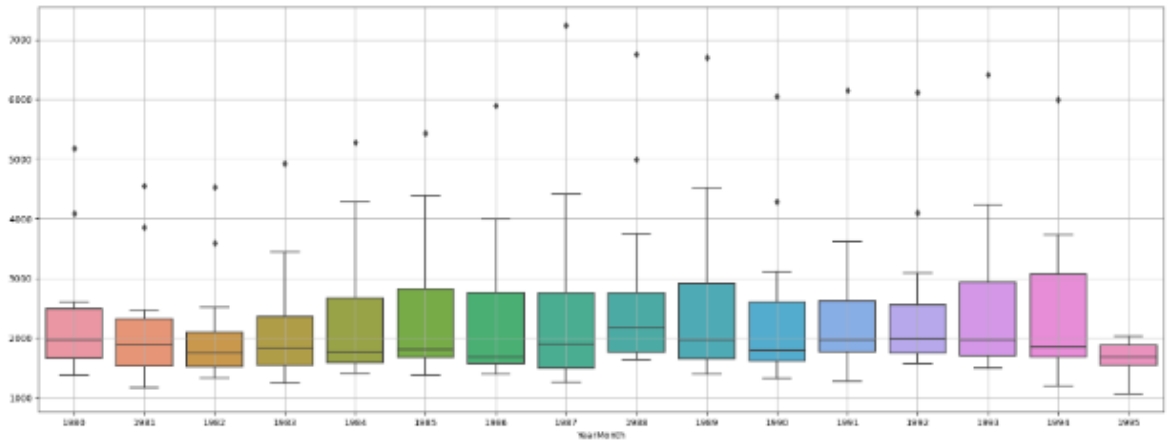

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
Data columns (total 1 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Sparkling  187 non-null    int64
dtypes: int64(1)
memory usage: 2.9 KB
```

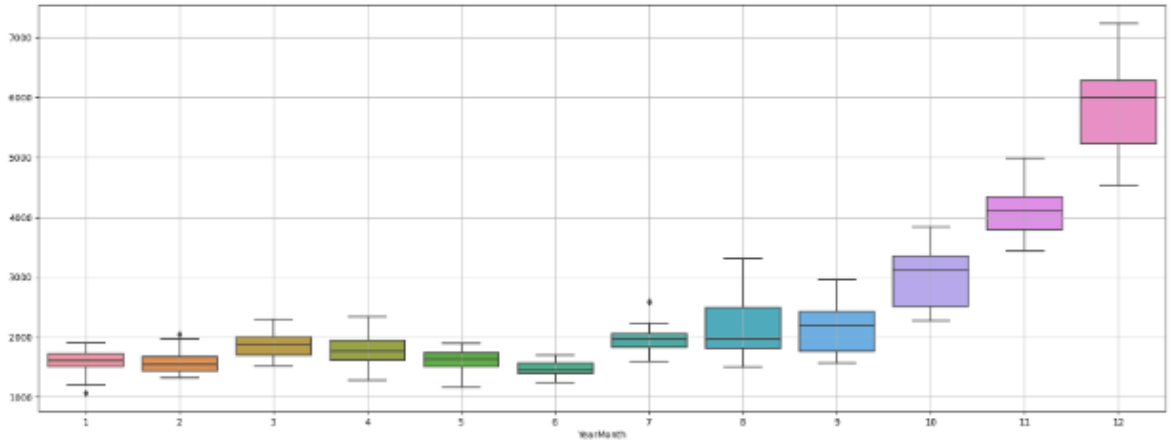
- Data consist of 187 data points
- It seems to be contain seasonality

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 and Monthly plot

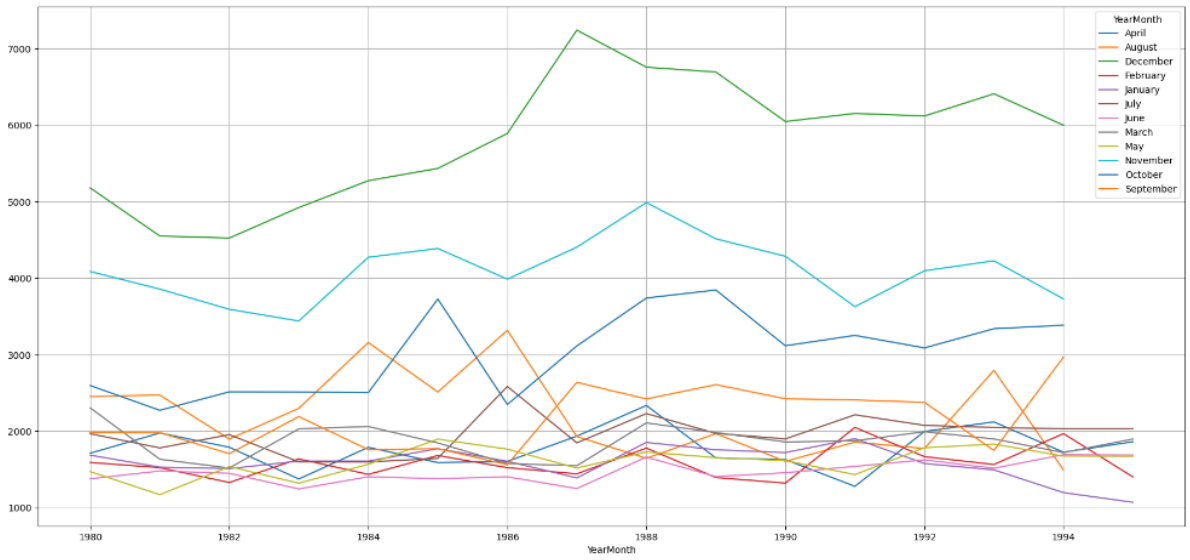


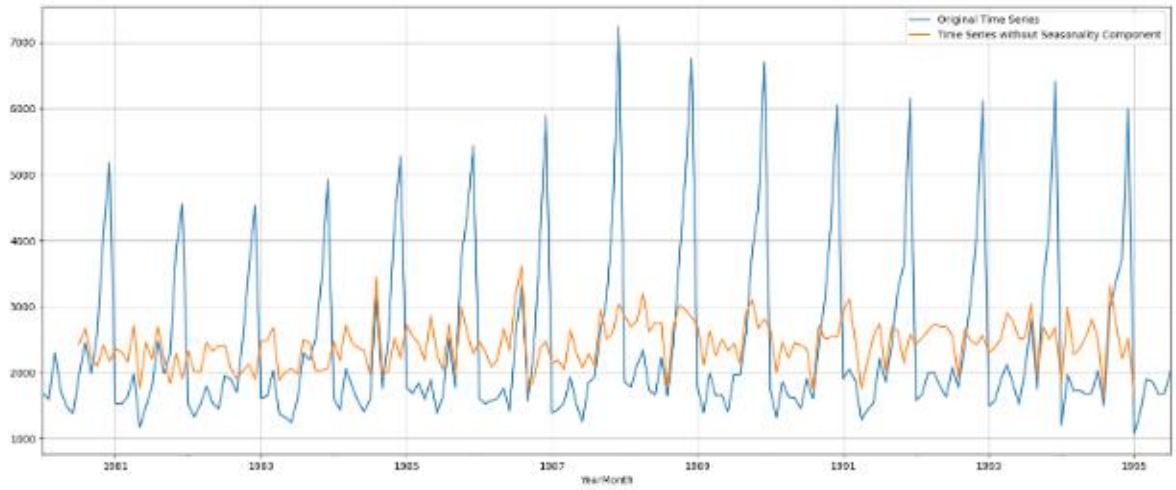


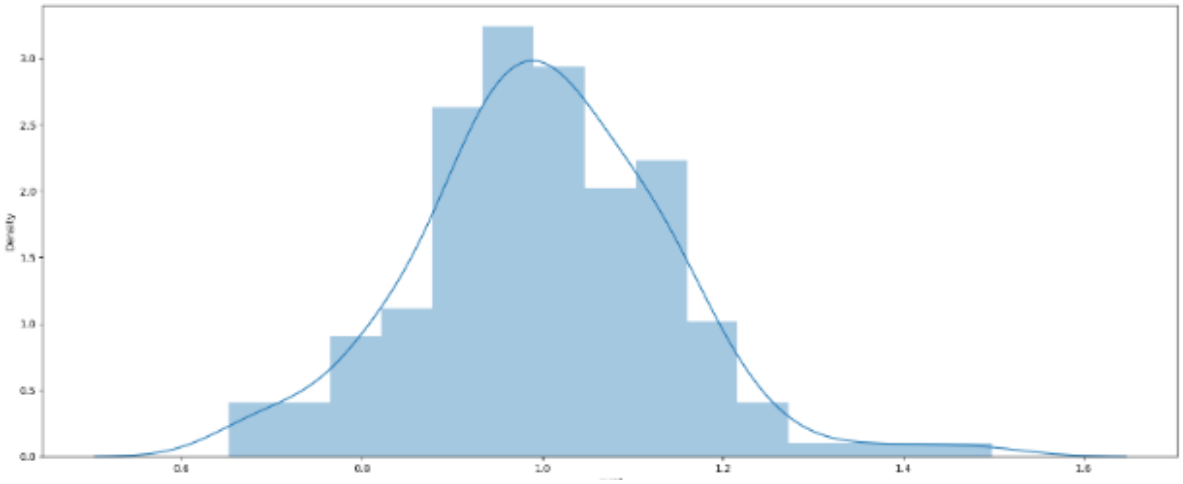
- Boxplot indicates the trend being present in the data.
- We can clearly see some of the outliers in the plot
- The box plot for various months is plotted
- Monthly plot contains outliers in the month of January, February and July.

- Plot for different month and different years

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

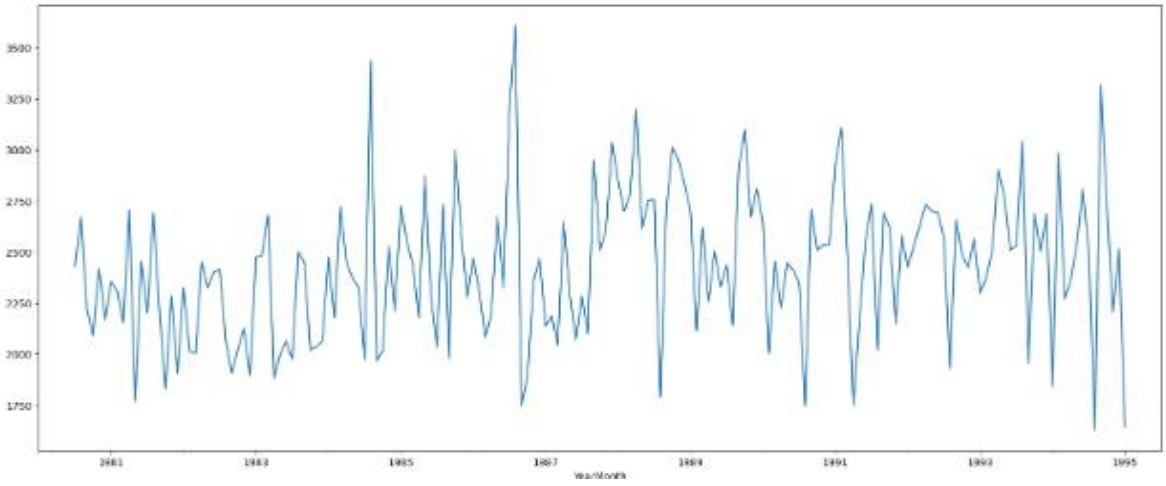







```
ShapiroResult(statistic=0.9859988689422607, pvalue=0.07802142202854156)
```

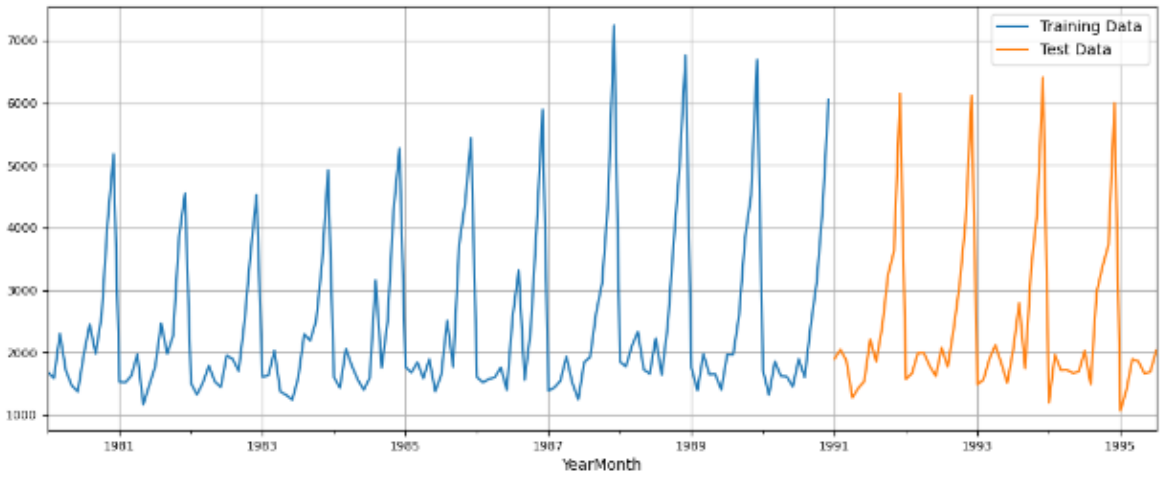
- Deseasonlized plot



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

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

Train and Test Shapes



- The test data starts from 1991
- 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 model, Moving Average model etc., should also be built on the training data and check the performance on the test data using RMSE.**

Model 1: Linear Regression:

- Regression the 'Sparkling' variable against the order of occurrence.
- Modifying the training set
- Generate the numerical instance order for both training the test set
- Printing the head and tail of test and train data

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

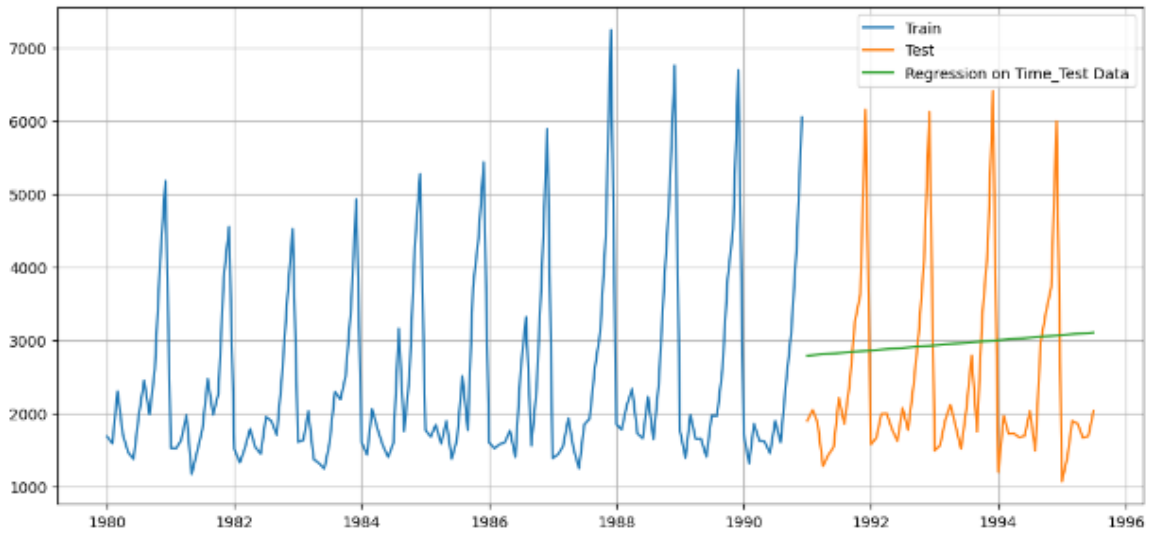
```
Last few rows of Training Data
            Sparkling  time
YearMonth
1990-08-01       1605   128
1990-09-01       2424   129
1990-10-01       3116   130
1990-11-01       4286   131
1990-12-01       6047   132


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

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

- Linear Regression is built on the training and test dataset

	Sparkling	time	RegOnTime
YearMonth			
1991-01-01	1902	133	2791.652093
1991-02-01	2049	134	2797.484752
1991-03-01	1874	135	2803.317410
1991-04-01	1279	136	2809.150069
1991-05-01	1432	137	2814.982727
1991-06-01	1540	138	2820.815386
1991-07-01	2214	139	2826.648044



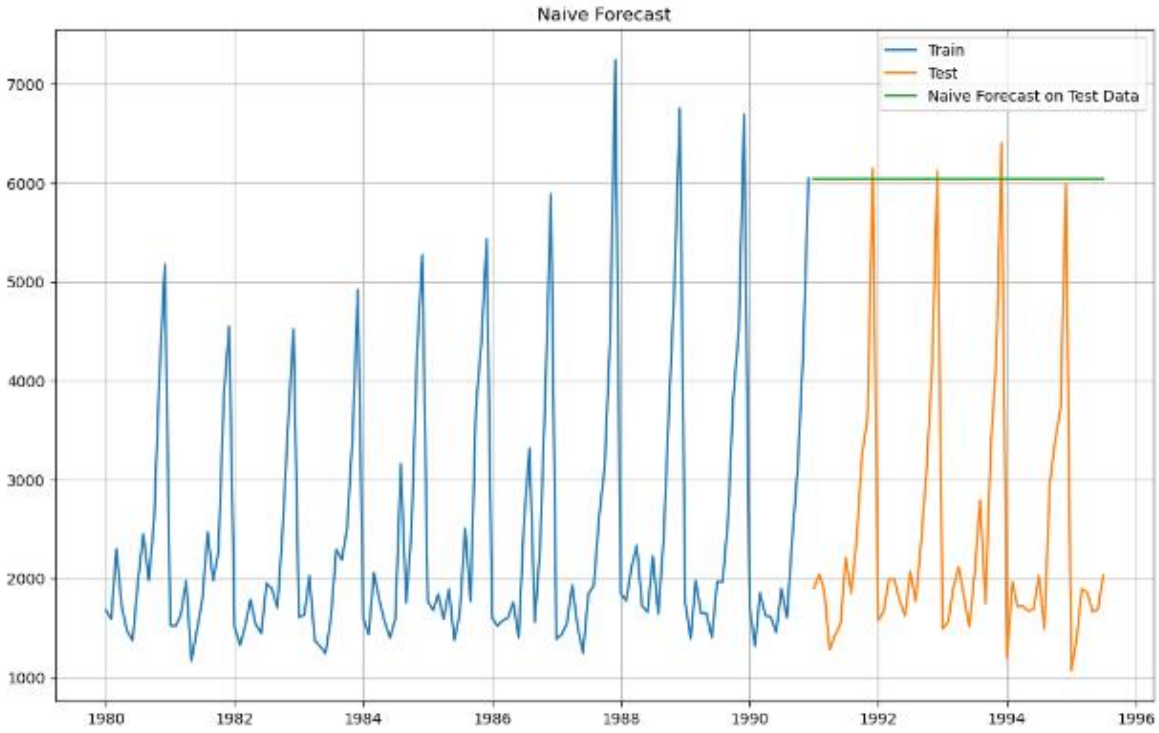
- Defining the accuracy metrics
- Evaluating the model

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

	Test RMSE
RegressionOnTime	1389.135175

Model 2: Naïve Approach:

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

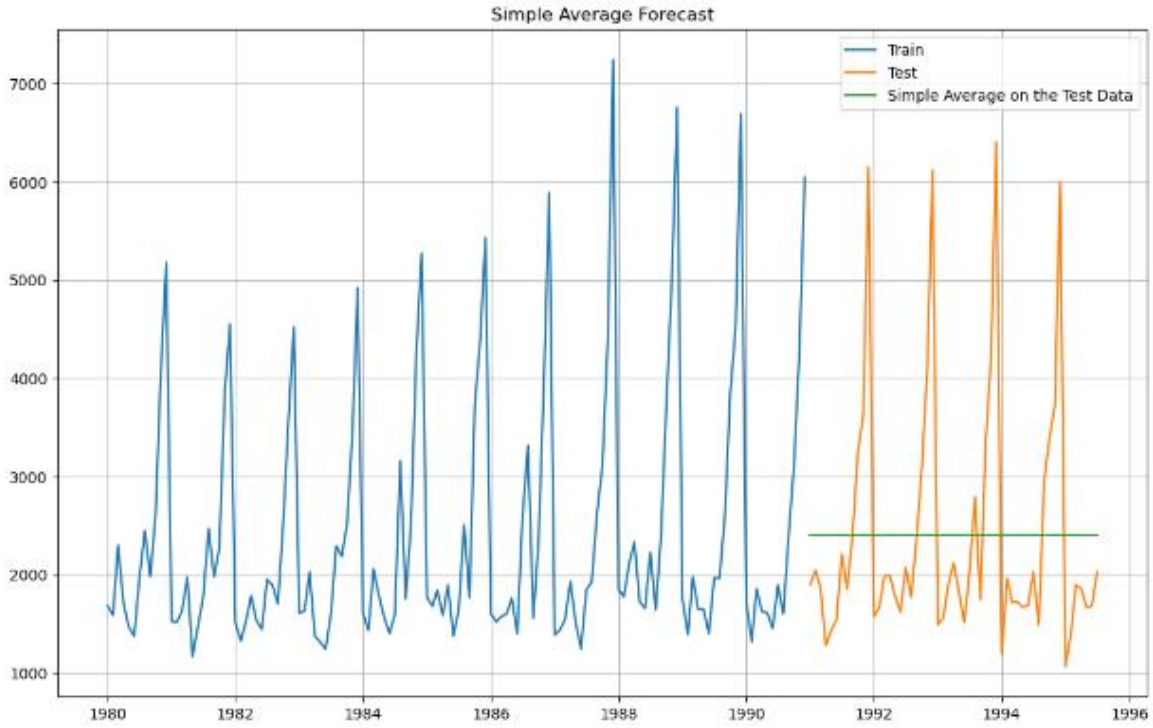


For Naive forecast on the Test Data, RMSE is 3864.279

	Test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352

Model 3: Simple Average Model:

YearMonth	Sparkling	mean_forecast
1991-01-01	1902	2403.780303
1991-02-01	2049	2403.780303
1991-03-01	1874	2403.780303
1991-04-01	1279	2403.780303
1991-05-01	1432	2403.780303

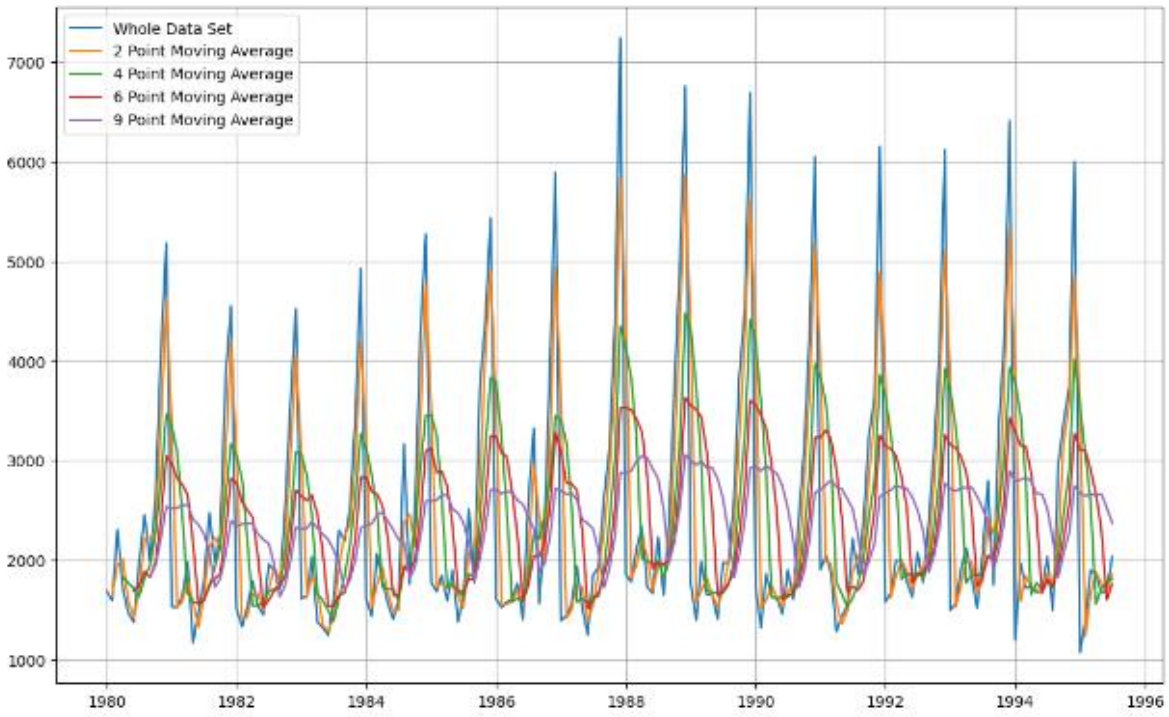



```
For Simple Average Forecast on the Test Data, RMSE is 1275.082
```

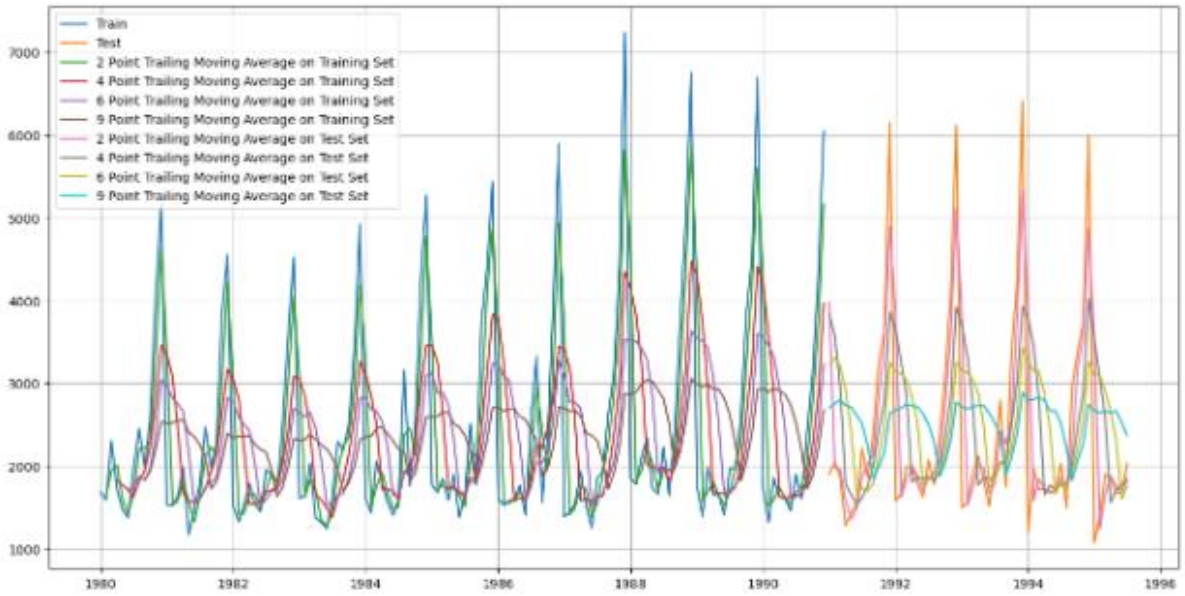
	Test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804

Model 4: Moving Average (MA) – 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.

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



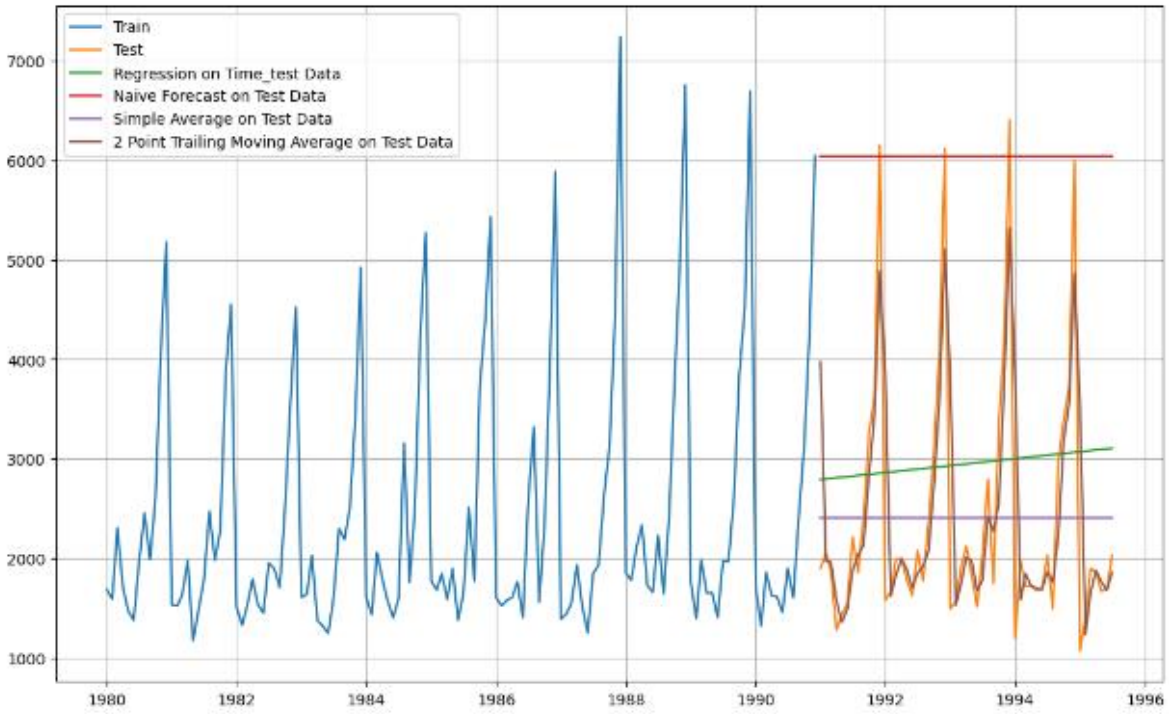
Let us split the data into train and test and plot this Time Series. The window of the moving average is need to be carefully selected as too big a window will result is not having any test set as the whole series might bet averaged over.




```
For 2 point Moving Average Model forecast on the Test data, RMSE is 813.401
For 4 point Moving Average Model forecast on the Test data, RMSE is 1156.590
For 6 point Moving Average Model forecast on the Test data, RMSE is 1283.927
For 9 point Moving Average Model forecast on the Test data, RMSE is 1346.278
```

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

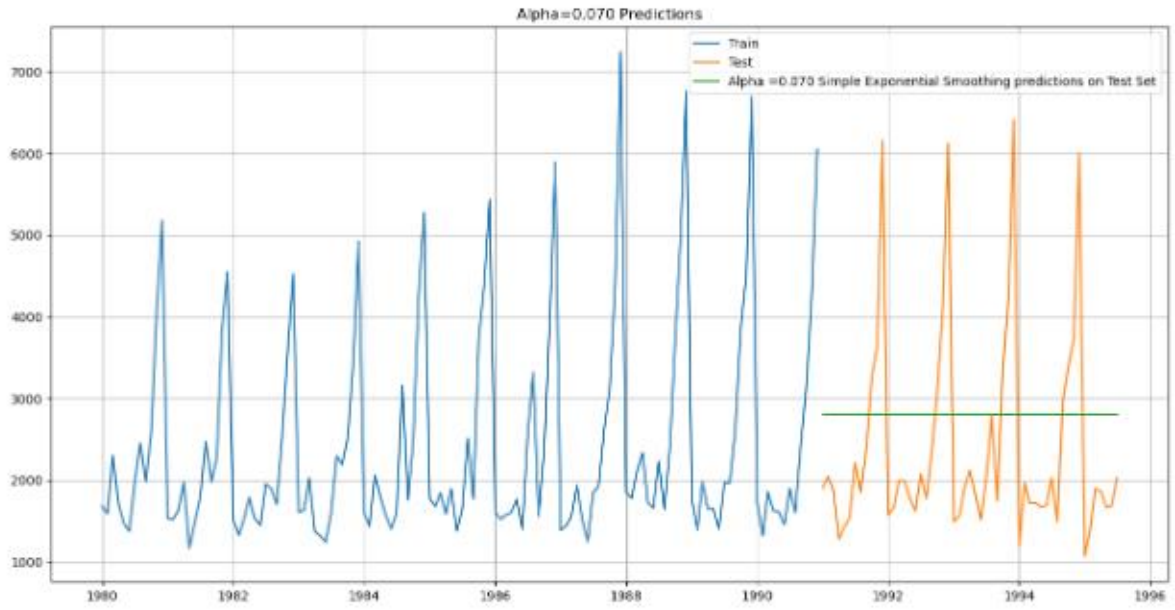
Before we go on to build the various Exponential smoothing models, let us plot all the models and compare the Time Series plots



Model – 5 – Exponential Smoothing

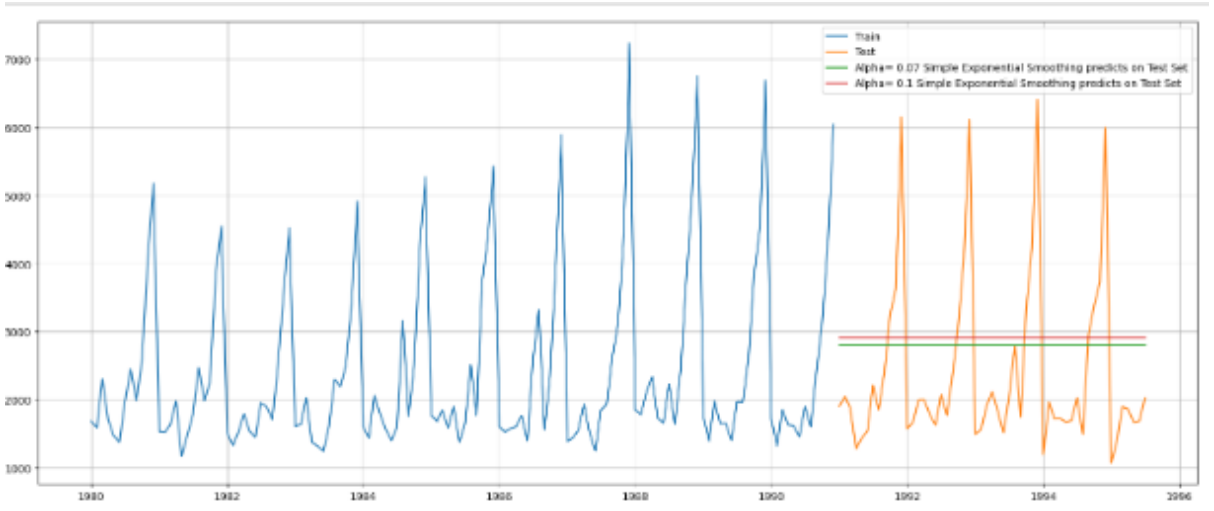
YearMonth	Sparkling	predict
1991-01-01	1902	2804.650301
1991-02-01	2049	2804.650301
1991-03-01	1874	2804.650301
1991-04-01	1279	2804.650301
1991-05-01	1432	2804.650301

```
{'smoothing_level': 0.07028442075641193,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 1763.8402828521703,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

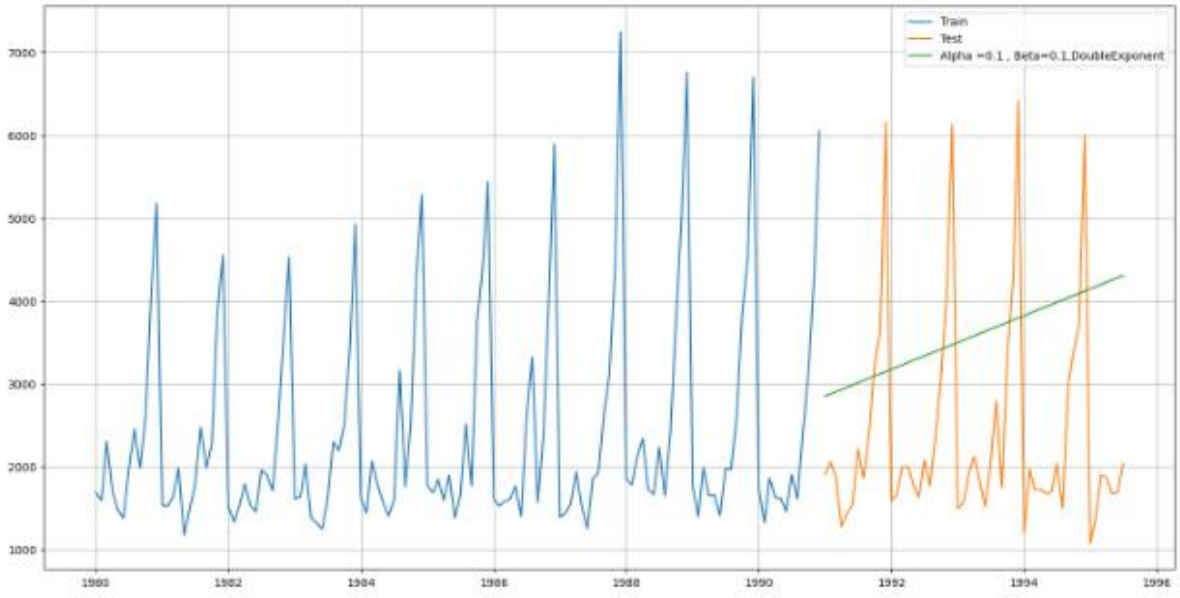


	Test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2PointTrailingMovingAverage	813.400684
4PointTrailingMovingAverage	1156.589694
6PointTrailingMovingAverage	1283.927428
9PointTrailingMovingAverage	1346.278315
Alpha=0.070,SimpleExponentialsSmoothing	1338.000861

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



- Double Exponential Model

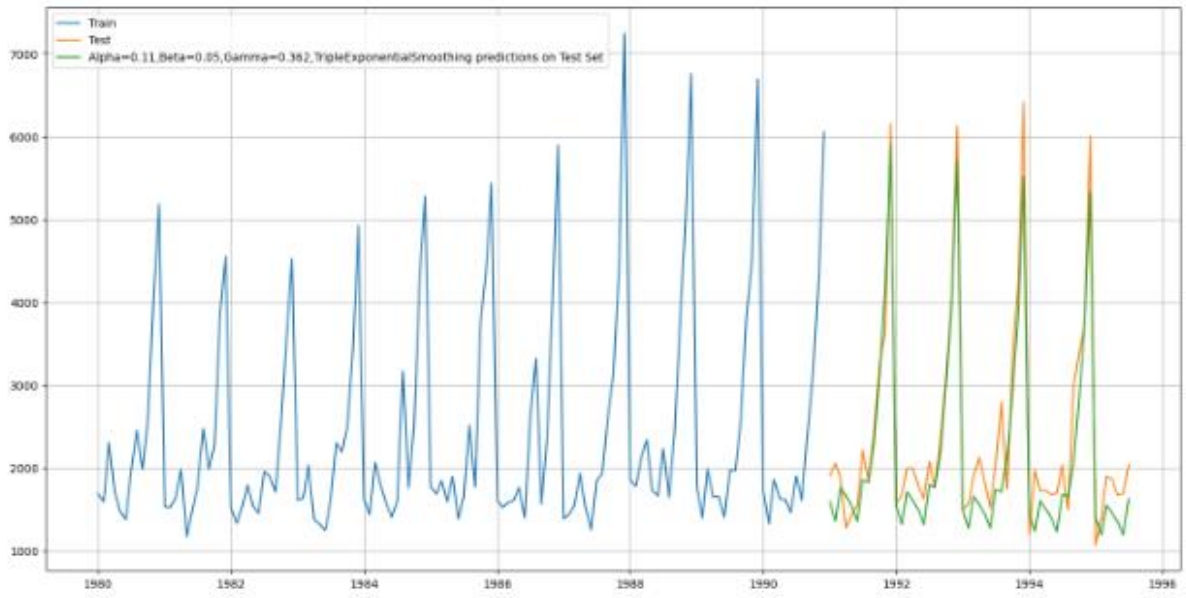


	Test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2PointTrailingMovingAverage	813.400684
4PointTrailingMovingAverage	1156.589694
6PointTrailingMovingAverage	1283.927428
9PointTrailingMovingAverage	1346.278315
Alpha=0.070,SimpleExponentialSmoothing	1338.000861
Alpha=0.3,SimpleExponentialSmoothing	1375.393335
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	1777.734773

- Triple Exponential Smoothing (Holt – Winter Model)

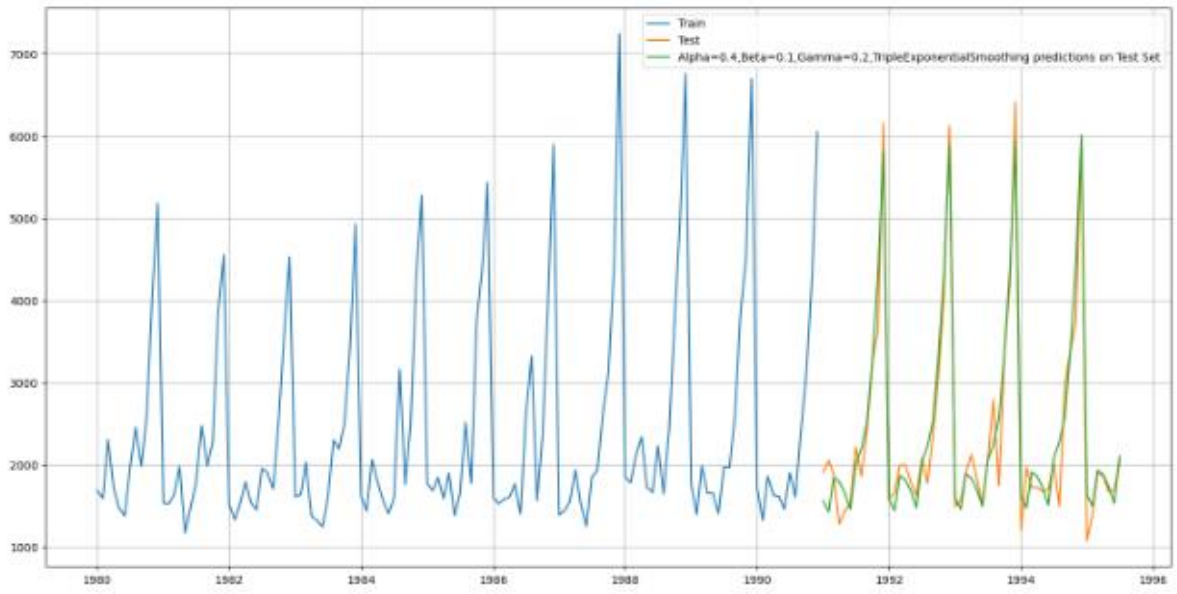
```
{'smoothing_level': 0.11101471561088701,
 'smoothing_trend': 0.0493145907614654,
 'smoothing_seasonal': 0.36244934537370843,
 'damping_trend': nan,
 'initial_level': 2356.496908624238,
 'initial_trend': -9.809526161838415,
 'initial_seasons': array([0.713711  , 0.68278724, 0.90458411, 0.8053878 , 0.65
571739,
        0.65388935, 0.88616088, 1.13350811, 0.91894498, 1.21186447,
        1.87099202, 2.37505867]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

YearMonth	Sparkling	auto_predict
1991-01-01	1902	1587.923122
1991-02-01	2049	1356.650595
1991-03-01	1874	1763.350752
1991-04-01	1279	1656.524633
1991-05-01	1432	1542.386930



	Test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2PointTrailingMovingAverage	813.400684
4PointTrailingMovingAverage	1156.589694
6PointTrailingMovingAverage	1283.927428
9PointTrailingMovingAverage	1346.278315
Alpha=0.070,SimpleExponentialSmoothing	1338.000861
Alpha=0.3,SimpleExponentialSmoothing	1375.393335
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	1777.734773
Alpha=0.11,Beta=0.05,Gamma=0.362,TripleExponentialSmoothing	402.936179

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

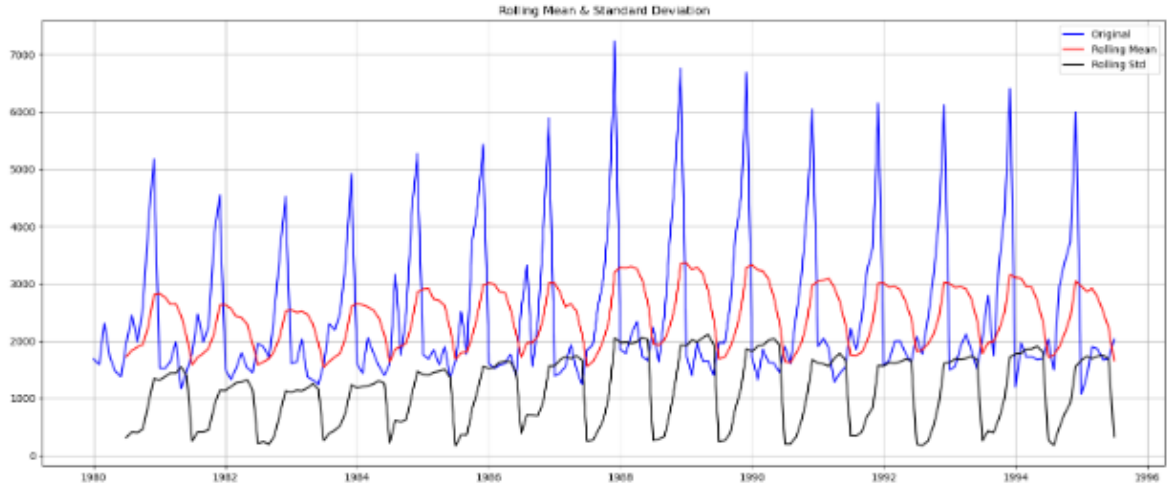


	Test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2PointTrailingMovingAverage	813.400684
4PointTrailingMovingAverage	1156.589694
6PointTrailingMovingAverage	1283.927428
9PointTrailingMovingAverage	1346.278315
Alpha=0.070,SimpleExponentialSmoothing	1338.000861
Alpha=0.3,SimpleExponentialSmoothing	1375.393335
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	1777.734773
Alpha=0.11,Beta=0.05,Gamma=0.362,TripleExponentialSmoothing	402.936179
Alpha=0.4,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing	317.434302

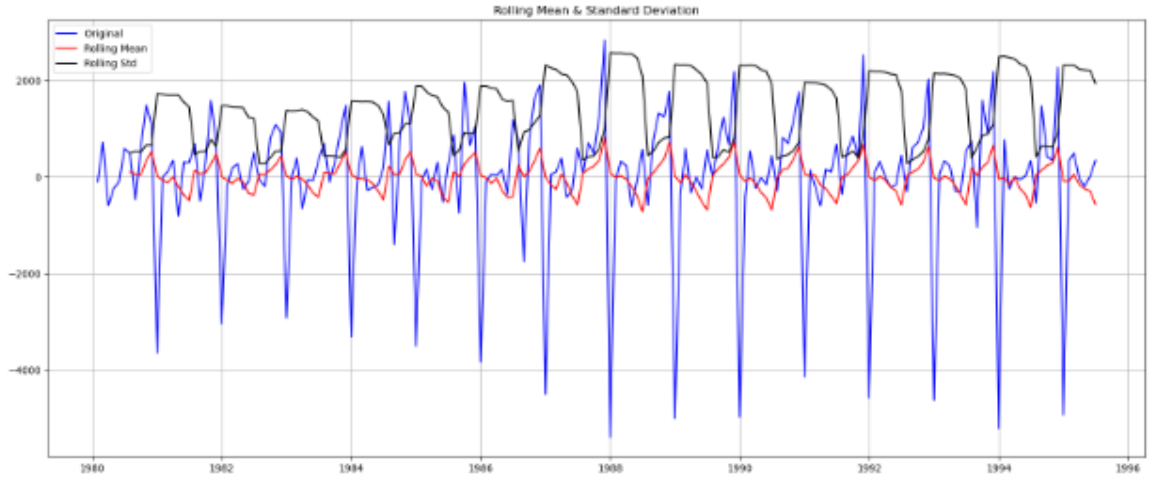
	Test RMSE
Alpha=0.4,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing	317.434302
Alpha=0.11,Beta=0.05,Gamma=0.362,TripleExponentialSmoothing	402.936179
2PointTrailingMovingAverage	813.400684
4PointTrailingMovingAverage	1156.589694
SimpleAverageModel	1275.081804
6PointTrailingMovingAverage	1283.927428
Alpha=0.070,SimpleExponentialsSmoothing	1338.000861
9PointTrailingMovingAverage	1346.278315
Alpha=0.3,SimpleExponentialSmoothing	1375.393335
RegressionOnTime	1389.135175
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	1777.734773
NaiveModel	3864.279352

- After sorting the values lowest RMSE in ascending order, Triple Exponential Smoothing is the best model with Alpha = 0.4, Beta =0.1 and Gamma=0.2 having RMSE as 317.434.

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 Observation 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 Observation Used     175.000000
Critical value (1%)             -3.468280
Critical value (5%)             -2.878202
Critical value (10%)            -2.575653
```

- A Time Series is considered to be stationary when statistical properties such as the variance and (auto) correlation are constant over time.
- Stationarity means that the autocorrelation of lag ‘k’ depends on k, but not on time t.

- When the time series data is not stationary we need to convert it into stationary before applying model.
- Apply Augmented Dickey fuller test to check the validity of stationarity, p-value < 0.05 to confirm the stationarity.
  - Null Hypothesis : Time Series is Non-Stationary.
  - Alternate Hypothesis: Time Series is stationary.
- From the null and alternative 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 the null hypothesis – Stationary
- Let us take a difference of order 1 and check whether the Time Series is stationary or not
- At Alpha = 0.05 the Time Series is indeed stationary
- (d=1) 1st order differencing is done where the difference between the current and previous(1 lag before) series is taken and then checked for stationarity using the ADF(Augument Dicky Fueller) test. If difference time series is stationary, we preceed with AR modelling. Else we do (d=2) 2nd order difference, and this process repeats till we get a stationary time series
- The variance of a time series may also not be the same over time. To remove this kind of non-stationarity, we can transform the data. If the variance is increasing over time, then a log transformation can stabilize the variance.

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.**

- Autoregression means regression of a variable on itself which means Autoregressive model use previous time period values to predict the current time period values.
- One of the fundamental assumptions of an AR model is that the time series is assumed to be a stationary process.
- We look at the Partial Autocorrelations of a stationary Time Series to understand the order of Auto-Regressive model.

7. **Build ARIMA/SARIMA model based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.**

- For an AR model, 2 ways to identify order of 'p'.
  1) PACF Approach: The PACF method where the (partial Auto Correlation Function) Values cut off and become zero after a certain lag. PACF vanishes if there is no regression coefficient that fat back. The cut-off value where this happens can be taken as the order of AR as 'p'. This can be seen from a PACF plot.
  2) Lowest AIC Approach: The lowest Akaike Information Criteria(AIC) value compared among different orders of 'p' is considered.

- 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.

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	132
Model:	ARIMA(2, 1, 2)	Log Likelihood	-1101.755
Date:	Sat, 21 Oct 2023	AIC	2213.509
Time:	08:58:45	BIC	2227.885
Sample:	01-01-1980 - 12-01-1990	HQIC	2219.351
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.3120	0.046	28.780	0.000	1.223	1.401
ar.L2	-0.5593	0.072	-7.742	0.000	-0.701	-0.418
ma.L1	-1.9917	0.109	-18.217	0.000	-2.206	-1.777
ma.L2	0.9999	0.110	9.109	0.000	0.785	1.215
sigma2	1.099e+06	2e-07	5.51e+12	0.000	1.1e+06	1.1e+06

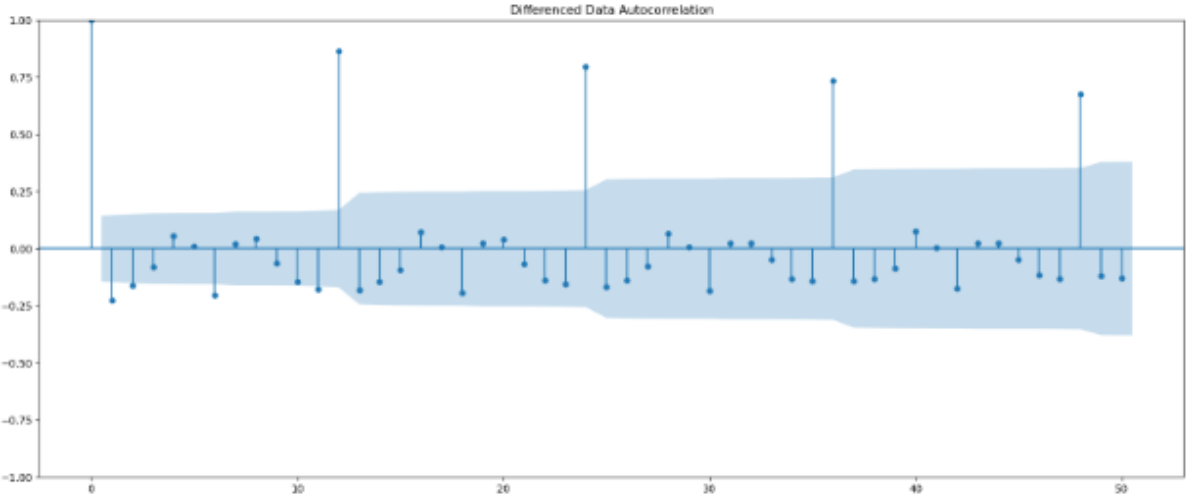
Ljung-Box (L1) (Q):	0.19	Jarque-Bera (JB):	14.46
Prob(Q):	0.67	Prob(JB):	0.00
Heteroskedasticity (H):	2.43	Skew:	0.61
Prob(H) (two-sided):	0.00	Kurtosis:	4.08

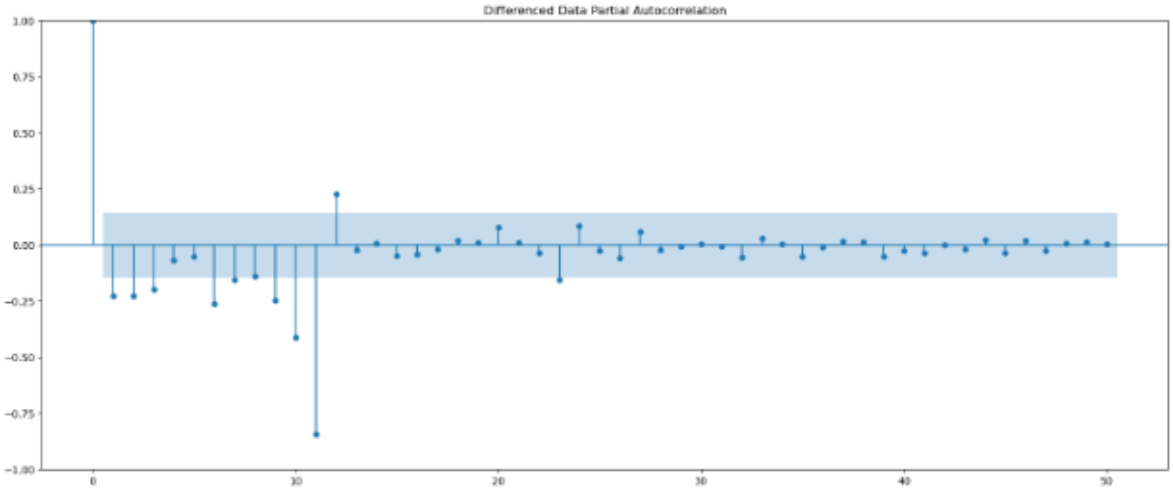
Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 5.46e+27. Standard errors may be unstable.

	RMSE
ARIMA(2,1,2)	1299.968134

```
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)
```

```
ARIMA(0, 1, 0) - AIC:2267.6630357855465
ARIMA(0, 1, 1) - AIC:2263.0600155918555
ARIMA(0, 1, 2) - AIC:2234.4083231303816
ARIMA(1, 1, 0) - AIC:2266.6085393190097
ARIMA(1, 1, 1) - AIC:2235.7550946734245
ARIMA(1, 1, 2) - AIC:2234.527200450945
ARIMA(2, 1, 0) - AIC:2260.365743968097
ARIMA(2, 1, 1) - AIC:2233.777626299803
ARIMA(2, 1, 2) - AIC:2213.509218065467
```





- Again we plot ACF to see and understand the sasonal parameter of SARIMA model.
- We see seasonality is 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 values.
- 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)
```

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 6)   Log Likelihood                -856.944
Date:                           Sat, 21 Oct 2023   AIC                           1727.889
Time:                                   08:59:11   BIC                           1747.164
Sample:                                        0   HQIC                          1735.713
                                             - 132
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.7851      0.103     -7.654      0.000      -0.986      -0.584
ma.L2         -0.0975      0.112     -0.870      0.384      -0.317       0.122
ar.S.L6        0.0022      0.026      0.084      0.933      -0.048       0.053
ar.S.L12       1.0396      0.018     58.251      0.000       1.005       1.075
ma.S.L6        0.0428      0.143      0.298      0.766      -0.238       0.324
ma.S.L12      -0.6203      0.090     -6.877      0.000      -0.797      -0.443
sigma2      1.475e+05   1.42e+04     10.371      0.000     1.2e+05    1.75e+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).
```

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1375.656633	384.084166	622.865501	2128.447765
1	1116.751594	392.851369	346.777059	1886.726128
2	1667.624322	395.424399	892.606741	2442.641903
3	1528.381331	397.983889	748.347242	2308.415421
4	1372.290762	400.527078	587.272113	2157.309410

	RMSE
ARIMA(2,1,2)	1299.968134
SARIMA(0,1,2)(2,0,2,6)	601.227607

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

```
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)
```

```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(0, 1, 2)x(1, 0, 2, 12)   Log Likelihood                -772.580
Date:                            Sat, 21 Oct 2023   AIC                           1557.161
Time:                                    08:59:53   BIC                           1573.027
Sample:                                         0   HQIC                          1563.588
                                              - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.7749      0.101     -7.707      0.000      -0.972      -0.578
ma.L2         -0.1321      0.123     -1.075      0.282      -0.373       0.109
ar.S.L12       1.0431      0.015     70.608      0.000       1.014       1.072
ma.S.L12      -0.5577      0.094     -5.932      0.000      -0.742      -0.373
ma.S.L24      -0.1183      0.120     -0.988      0.323      -0.353       0.116
sigma2      1.567e+05   1.89e+04      8.306      0.000     1.2e+05    1.94e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                21.23
Prob(Q):                              0.88   Prob(JB):                         0.00
Heteroskedasticity (H):               1.36   Skew:                             0.50
Prob(H) (two-sided):                  0.37   Kurtosis:                         4.97
===================================================================================

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

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(2,1,2)	1299.968134
SARIMA(0,1,2)(2,0,2,6)	601.227607
SARIMA(0,1,2)(1,0,2,12)	507.867500

- It is clear that SARIMA(0,1,2)(1,0,2,12) has the lower RMSE and ARIMA(2,1,2) has the higher value.

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

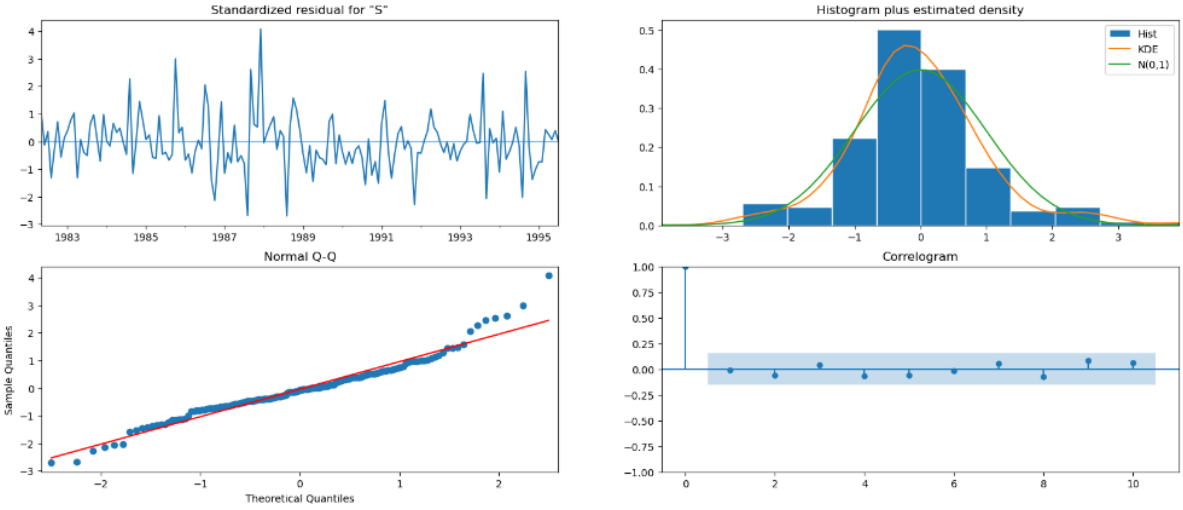
SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	187
Model:	SARIMAX(0, 1, 2)x(1, 0, 2, 12)	Log Likelihood	-1174.374
Date:	Sat, 21 Oct 2023	AIC	2360.749
Time:	08:59:54	BIC	2379.162
Sample:	01-01-1980 - 07-01-1995	HQIC	2368.226
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.9257	0.105	-8.778	0.000	-1.132	-0.719
ma.L2	-0.1223	0.086	-1.421	0.155	-0.291	0.046
ar.S.L12	1.0145	0.013	81.071	0.000	0.990	1.039
ma.S.L12	-0.5728	0.073	-7.899	0.000	-0.715	-0.431
ma.S.L24	-0.0592	0.081	-0.732	0.464	-0.218	0.099
sigma2	1.345e+05	1.43e+04	9.376	0.000	1.06e+05	1.63e+05

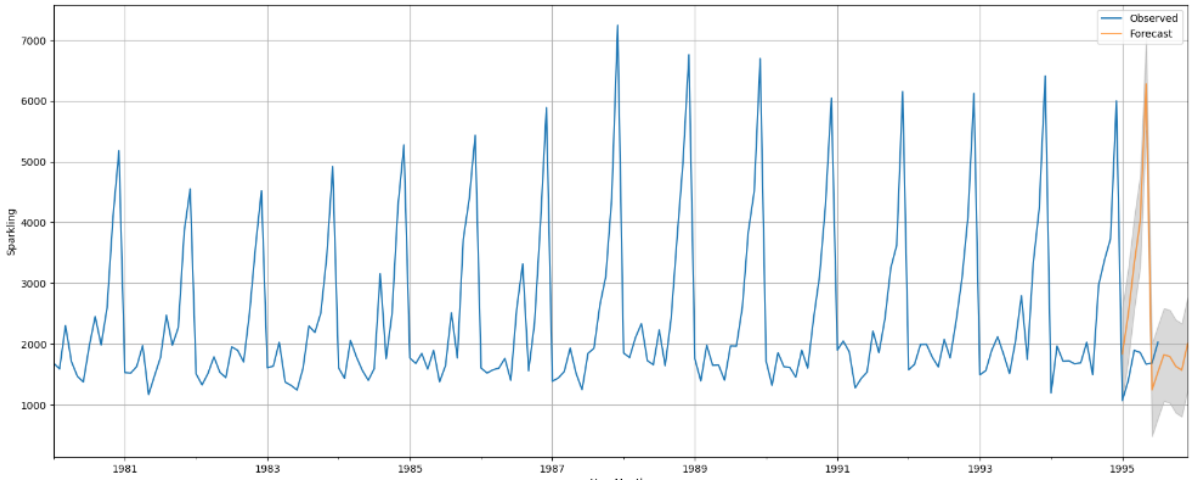
Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	41.23
Prob(Q):	0.91	Prob(JB):	0.00
Heteroskedasticity (H):	0.96	Skew:	0.60
Prob(H) (two-sided):	0.89	Kurtosis:	5.19

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



Sparkling	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-01	1841.603719	382.492352	1091.932485	2591.274952
1995-09-01	2492.873270	387.265254	1733.847320	3251.899220
1995-10-01	3324.301581	387.665509	2564.491145	4084.112016
1995-11-01	4015.785130	388.065351	3255.191018	4776.379242
1995-12-01	6278.887251	388.464786	5517.510262	7040.264240
1996-01-01	1247.430575	388.863827	485.271478	2009.589671
1996-02-01	1541.867501	389.262583	778.926859	2304.808144
1996-03-01	1826.029723	389.661209	1062.307787	2589.751660
1996-04-01	1794.780678	390.058977	1030.279131	2559.282226
1996-05-01	1634.583265	390.456361	869.302860	2399.863670
1996-06-01	1572.222224	390.853346	806.163743	2338.280705
1996-07-01	2005.679015	391.249930	1238.843243	2772.514787

```
RMSE of Full Model  534.763907243463
```



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 534.7639
- Plotting the forecast with the confidence band

- It is clear that SARIMA(0,1,2)(1,0,2,12) has the lowest RMSE and ARIMA(2,1,2) has the higher value.