Sparkling Wine

ANALYSIS AND PREDICTION

SUHAS PAWAR | TIME SERIES FORECASTING

Problem Statement:

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

Data set for the Problem: Sparkling.csv

Please do perform the following questions on each of these two data sets separately.

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

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

First of all we will import all the necessary libraries like Pandas, Numpy, seaborn, os to set the path and Statsmodels as base library for all the models.

After setting the path we will import Sparkling dataset

We will read the dataset First by using pandas library.

Then We will Apply head and tail to see how data looks first and last 5 rows.

First 5 Rows:

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

Last 5 Rows:

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

Null Value check:

We will check for null condition, to see whether Dataset has null value or not.

```
YearMonth    0
Sparkling    0
dtype: int64
```

No null values are present in the dataset.

Shape of the Data:

```
(187, 2)
```

So Dataset has 187 rows and 2 Columns.

Data Info:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 187 entries, 0 to 186
Data columns (total 2 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   YearMonth   187 non-null    object
 1   Sparkling   187 non-null    int64 
dtypes: int64(1), object(1)
memory usage: 3.0+ KB
```

We can see that YearMonth column is of Object Data type while Sparkling is having int data type.

Here we make an assumption that the date starts and ends as mentioned below

We are assuming the Date Starts from 1980-01-31 to 1995-07-31.

```
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')
```

Total 187 entries needs to be added to Original Dataframe while we need to remove YearMonth Column.

And need to make This new Column as Index.

	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

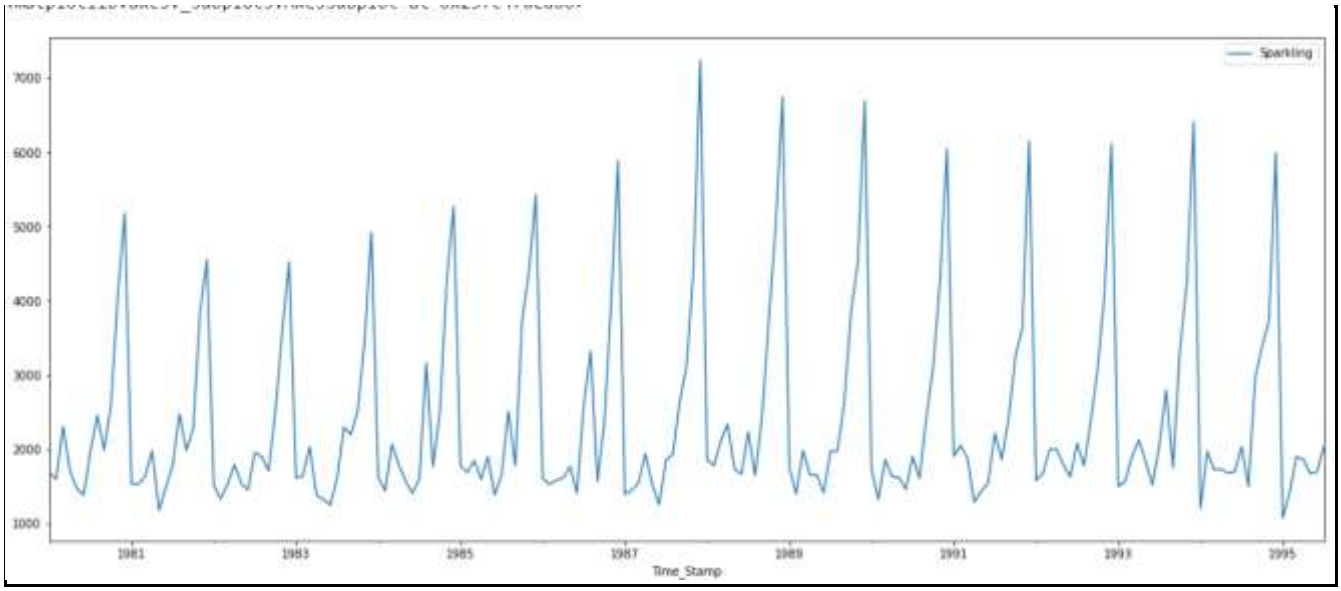
We can see the timestamp is incorporated in dataset, we will make this as index. We only need the column Sparkling Sales. Other Time_Stamp will be index column.

Time_Stamp	Sparkling
1980-01-31	1686
1980-02-29	1591
1980-03-31	2304
1980-04-30	1712
1980-05-31	1471
...	...
1995-03-31	1897
1995-04-30	1862
1995-05-31	1670
1995-06-30	1688
1995-07-31	2031

187 rows × 1 columns

We will plot the dataset of Sparkling wine sales.

Lets plot the time series



By looking and Above plot we must say there is seasonality present.

In above plot we can see some spikes at the end of every year, but we are little bit Confused about trend, like from 1982 to 1987 there is Some positive trend, but from there onwards downward trend.Lets do further EDA to get more insights.

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

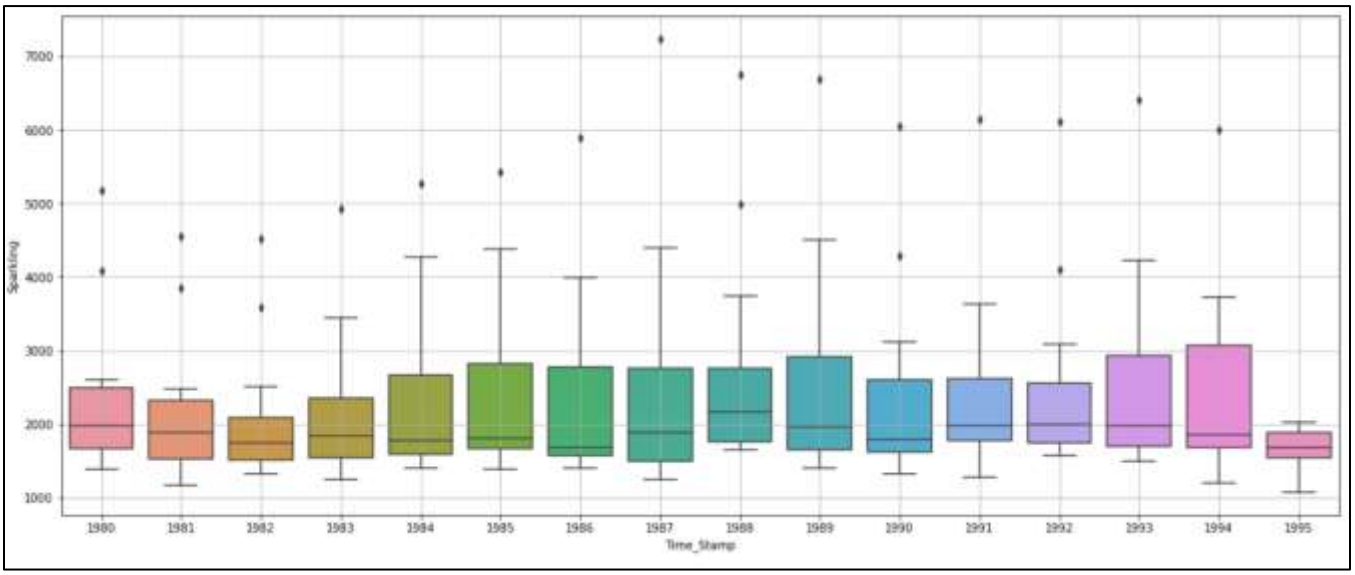
Description:

	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 above describe Function we can see that Highest Sales of Sparkling wine is 7242.Mean and Median are having much difference.

Total 187 rows we have enties from Jan 1980 to July 1995.

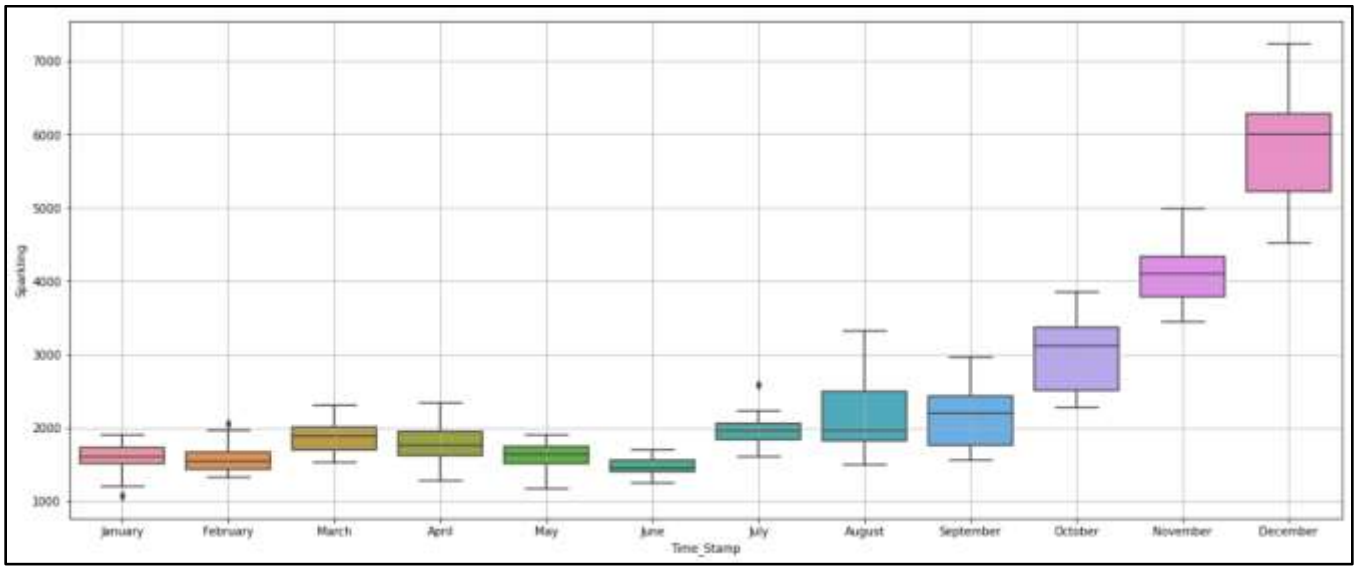
Yearly plot:



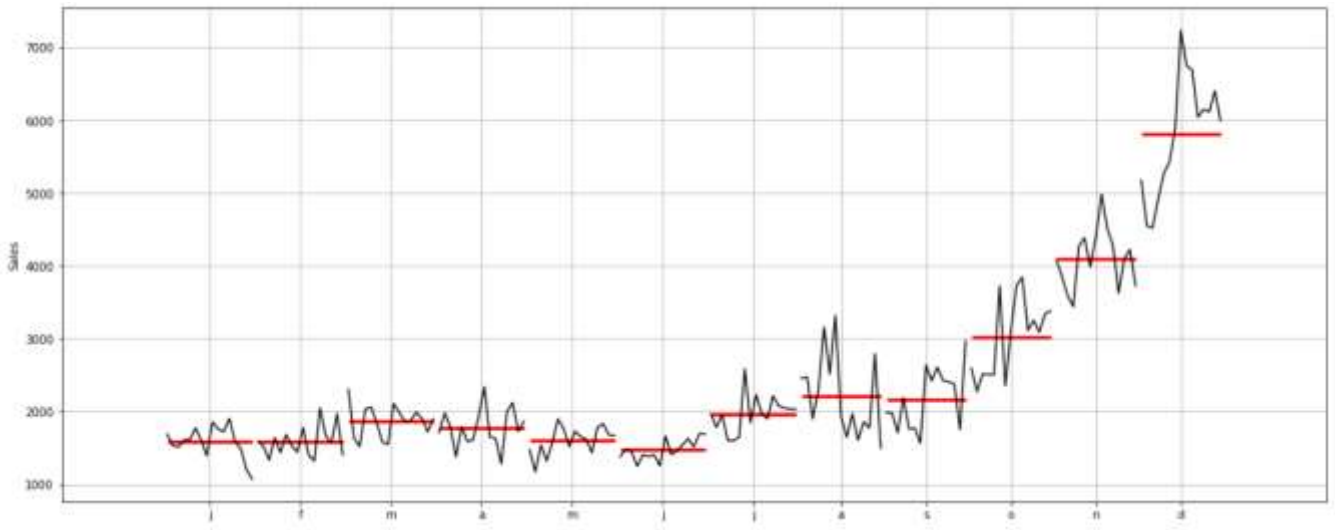
We can see from above graph the median value for all years is remains around 2000. It means the sales of Sparkling Wine is Almost same over the years, that is not Positive or Negative trend.
Highest Sales of Sparkling wine is occurred in year 1987, and in 1995 one of the month has brought Lowest sale of wines.

Lets check for Monthly plot

Monthly plot:



Yes finally we get Something notable, In above plot Sales has been Increasd from August to December. Till this point Sales are around 2000. so We need Higher stock or Production from August to December.October, November , December has sales more than 2000 and it is almost 3000,4000,6000.

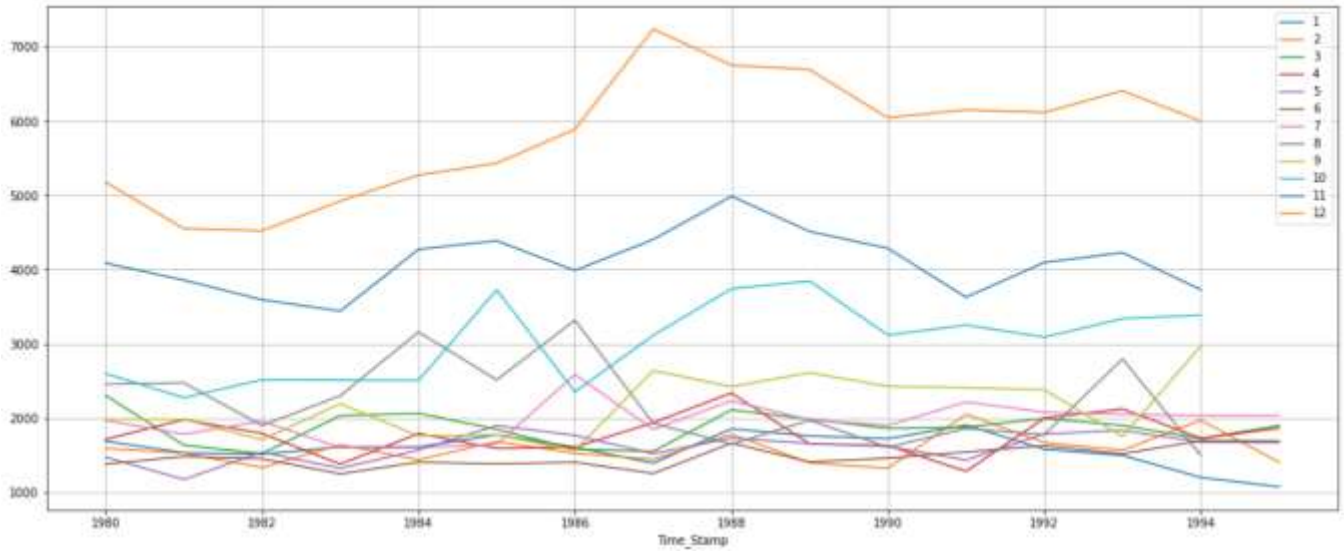


Highlight The maximum sales:

We can see above, The maximum Sales of Month in a All years.

itamp	1	2	3	4	5	6	7	8	9	10	11	
itamp												
1980	1686.000000	1591.000000	2304.000000	1712.000000	1471.000000	1377.000000	1966.000000	2453.000000	1984.000000	2596.000000	4087.000000	5179.00
1981	1530.000000	1523.000000	1633.000000	1976.000000	1170.000000	1480.000000	1781.000000	2472.000000	1981.000000	2273.000000	3857.000000	4551.00
1982	1510.000000	1329.000000	1518.000000	1790.000000	1537.000000	1449.000000	1954.000000	1897.000000	1706.000000	2514.000000	3593.000000	4524.00
1983	1609.000000	1638.000000	2030.000000	1375.000000	1320.000000	1245.000000	1600.000000	2298.000000	2191.000000	2511.000000	3440.000000	4923.00
1984	1609.000000	1435.000000	2061.000000	1789.000000	1567.000000	1404.000000	1597.000000	3159.000000	1759.000000	2504.000000	4273.000000	5274.00
1985	1771.000000	1682.000000	1846.000000	1589.000000	1896.000000	1379.000000	1645.000000	2512.000000	1771.000000	3727.000000	4368.000000	5434.00
1986	1606.000000	1523.000000	1577.000000	1605.000000	1765.000000	1403.000000	2584.000000	3318.000000	1562.000000	2349.000000	3967.000000	5891.00
1987	1389.000000	1442.000000	1548.000000	1935.000000	1518.000000	1250.000000	1847.000000	1930.000000	2638.000000	3114.000000	4405.000000	7242.00
1988	1853.000000	1779.000000	2108.000000	2336.000000	1728.000000	1661.000000	2230.000000	1645.000000	2421.000000	3740.000000	4988.000000	6757.00
1989	1757.000000	1394.000000	1962.000000	1650.000000	1654.000000	1406.000000	1971.000000	1968.000000	2608.000000	3845.000000	4514.000000	6694.00
1990	1720.000000	1321.000000	1859.000000	1628.000000	1615.000000	1457.000000	1899.000000	1605.000000	2424.000000	3116.000000	4286.000000	6047.00
1991	1902.000000	2049.000000	1874.000000	1279.000000	1432.000000	1540.000000	2214.000000	1657.000000	2408.000000	3252.000000	3627.000000	6153.00
1992	1577.000000	1667.000000	1993.000000	1997.000000	1783.000000	1625.000000	2076.000000	1773.000000	2377.000000	3088.000000	4096.000000	6119.00
1993	1494.000000	1564.000000	1898.000000	2121.000000	1831.000000	1515.000000	2048.000000	2795.000000	1749.000000	3339.000000	4227.000000	6410.00
1994	1197.000000	1968.000000	1720.000000	1725.000000	1674.000000	1693.000000	2031.000000	1495.000000	2968.000000	3385.000000	3729.000000	5999.00
1995	1070.000000	1402.000000	1897.000000	1862.000000	1670.000000	1688.000000	2031.000000	nan	nan	nan	nan	

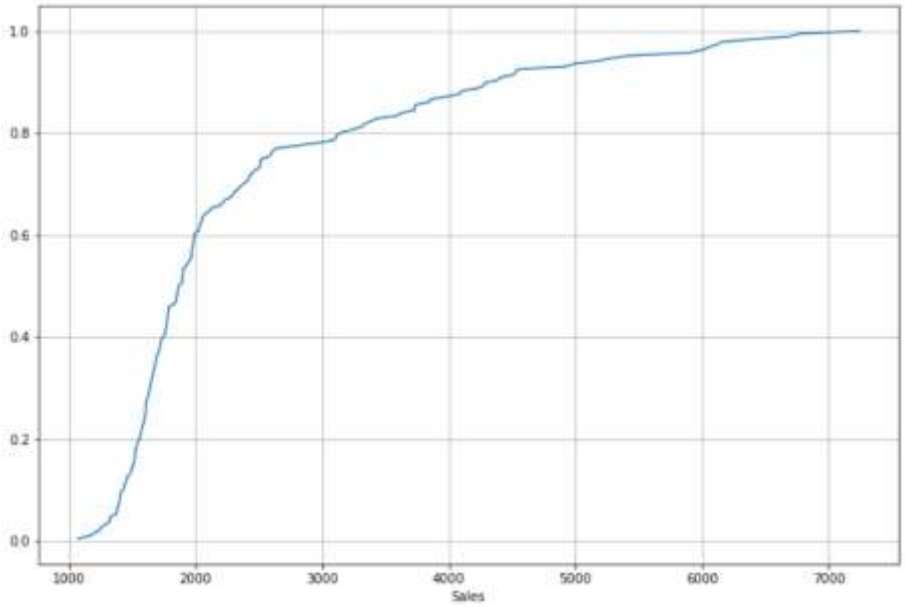
We can see above, The maximum Sales of Month in a all years.



From 9th month sales are going High thats what we can see in above plot, This is true for all the mentioned time series. In The Year the 1987 December the Sales are Highest it can be seen.

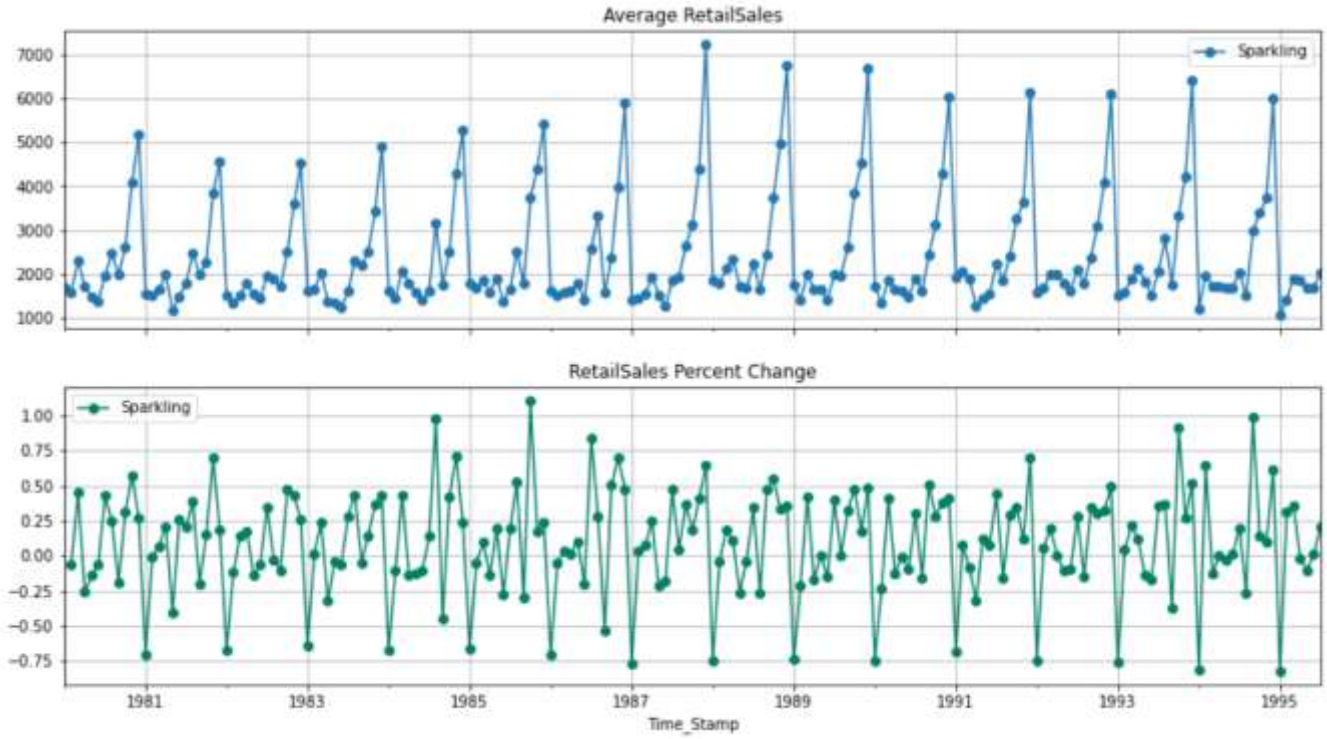
The orange line is for December month, It is Alienated from Other lines.

Empirical Cumulative Distribution.



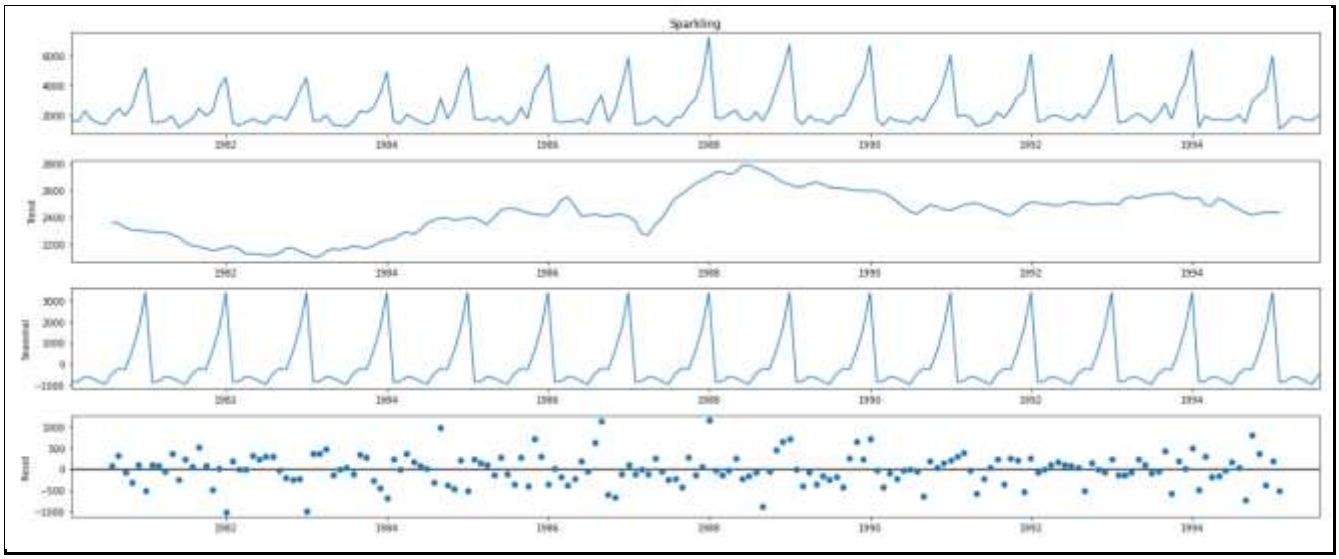
This particular graph tells us what percentage of data points refer to what number of Sales. 60% of the sales are below 2000. Maximum sales is greater than 7000. Only 20% of whole data tells us that sales are more than 3000-6000..

Plot of average Sales per month and the month on month percentage change of Sales.



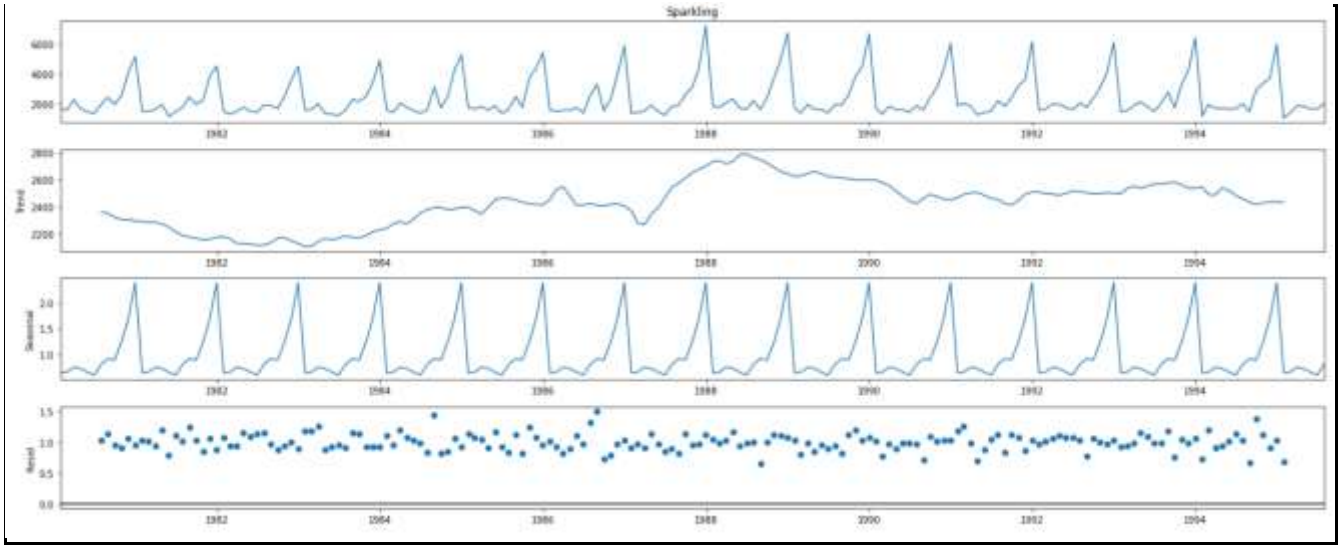
Decompose the Time Series and plot the different components.

Additive Decomposition:



We were not Clear about the trend and Seasonality, but here one thing we can notice There is strong Seasonality, Positive Negative trends. One more thing to Notice in Additive Decomposition the Residuals are showing patterns, it is not permissible actually not a good choice, but before saying anything lets move to Multiplicative Decomposition.

Multiplicative Decomposition:



For the multiplicative decomposition series, we see that a lot of residuals are located around 1.

Almost Same Trend And Seasonality, Also residual is showing patterns so we will choose Additive decomposition as its alwyas easy to go with Additive Model, why to choose complicate models? when both are giving almost same results.

```
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: float64
```

```
Residual
 Time_Stamp
1980-01-31         NaN
1980-02-29         NaN
1980-03-31         NaN
1980-04-30         NaN
1980-05-31         NaN
1980-06-30         NaN
1980-07-31    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.

Training Data is till the end of 1990. Test Data is from the beginning of 1991 to the last time stamp provided.

Train Data:

First Five rows:

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

Last Five rows:

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

Test Data:

First 5 rows:

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

Test Data Last 5 rows:

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

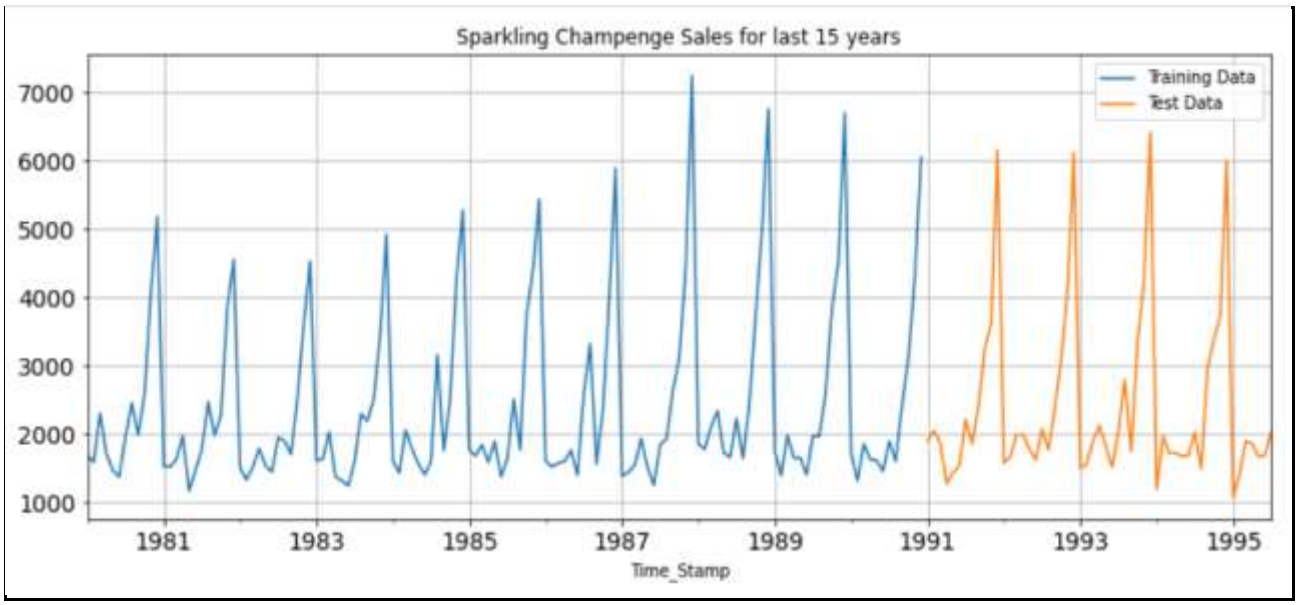
Shape of Train Data: 132 Rows and 1 Column

Shape of Test Data : 55 Rows and 1 Column

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

test data starts at Jan-1991 and It is till July-1995

Train and Test data plot:



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 and simple average models. should also be built on the training data and check the performance on the test data using RMSE.

1.Simple Exponential Smoothing

We have created the test and train data for the upcoming Models.

First we will create model on train Data, will test model performance on Test Data.

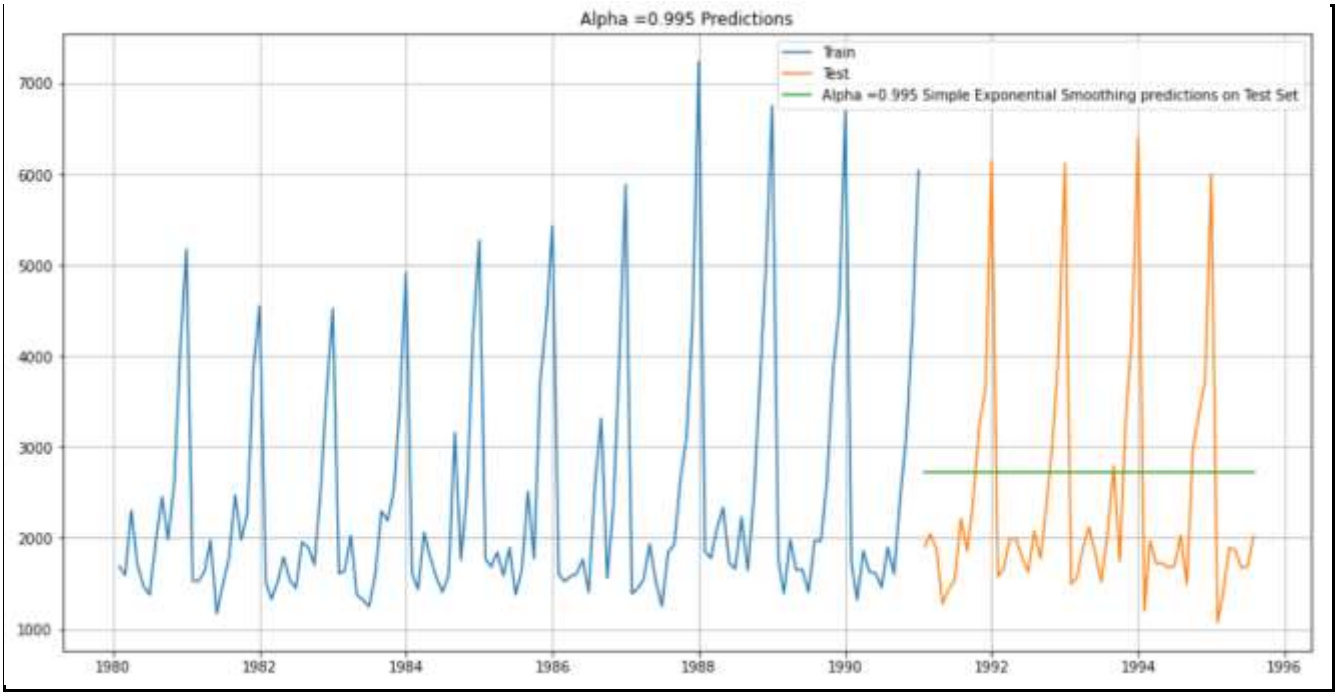
The parameters of Model Simple Exponential Smoothing

```
{'smoothing_level': 0.04960659880745982,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 1818.5047538435374,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Now lets predict on the test data

Time_Stamp	Sparkling	predict
1991-01-31	1902	2724.93
1991-02-28	2049	2724.93
1991-03-31	1874	2724.93
1991-04-30	1279	2724.93
1991-05-31	1432	2724.93

Plotting on both the Training and Test data:



Model Evaluation for α = 0.995 : Simple Exponential Smoothing

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

	Test RMSE
Alpha=0.995,Simple Exponential Model	1316.03

Setting different alpha values. the higher the alpha value more weightage is given to the more recent observation. That means, what happened recently will happen again. We will run a loop with different alpha values to understand which particular value works best for alpha on the test set.

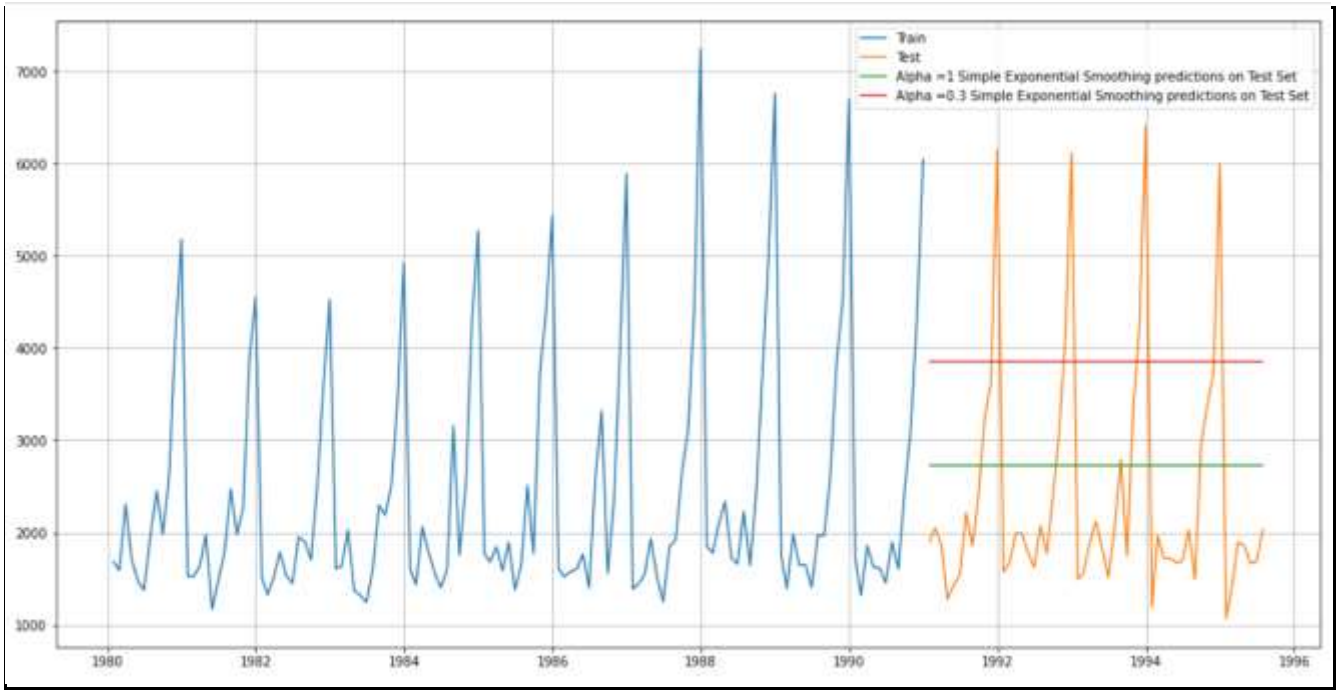
First we will define an empty dataframe to store our values from the loop

Alpha Values	Train RMSE	Test RMSE

Model Evaluation:

	Alpha Values	Train RMSE	Test RMSE
0	0.30	1359.51	1935.51
1	0.40	1352.59	2311.92
2	0.50	1344.00	2666.35
3	0.60	1338.81	2979.20
4	0.70	1338.84	3249.94
5	0.80	1344.46	3483.80
6	0.90	1355.72	3686.79
7	1.00	1373.08	3864.28

Plotting on both the Training and Test data



The Final Result of Simple Exponential Smoothing in form of RMSE:

	Test RMSE
Alpha=0.995,Simple Exponential Model	1316.03
Alpha=0.3,SimpleExponentialSmoothing	1935.51

Model 2 : Double Exponential Smoothing (Holt's Model)

Two parameters α and β are estimated in this model. Level and Trend are accounted for in this model.

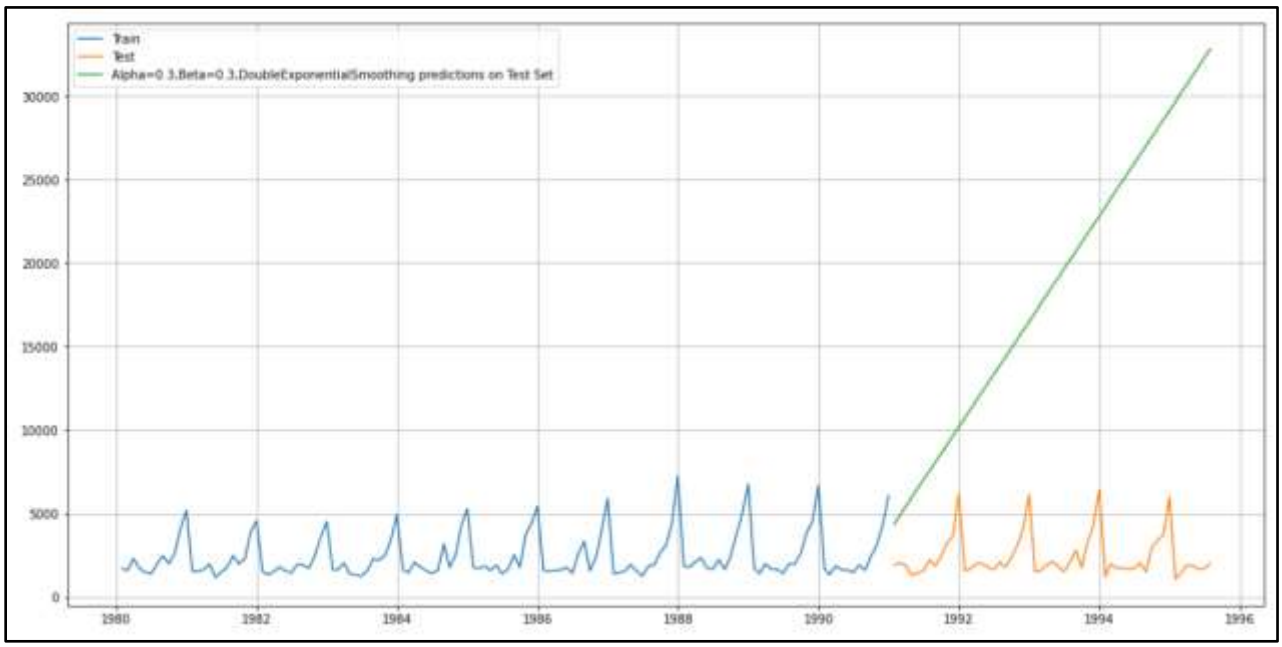
First We will create a model and then we will predict on test data.

We will check the results for Different values of Alpha and Beta . Let us sort the data frame in the ascending ordering of the 'Test RMSE' and the 'Test MAPE' values.

	Alpha Values	Beta Values	Train RMSE	Test RMSE
0	0.30	0.30	1592.29	18259.11
8	0.40	0.30	1569.34	23878.50
1	0.30	0.40	1682.57	26069.84
16	0.50	0.30	1530.58	27095.53
24	0.60	0.30	1506.45	29070.72

Test RMSE is lowest for Alpha=0.3 and Beta =0.3. hence we will select these values for better performance.

Plotting on both the Training and Test data



Test RMSE :

	Test RMSE
Alpha=0.995,Simple Exponential Model	1316.03
Alpha=0.3,SimpleExponentialSmoothing	1935.51
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	18259.11

For this the Test RMSE is too high. So we cannot go further with this model. Lets deep dive more, to check about other models.

Method 3: Triple Exponential Smoothing (Holt - Winter's Model)

Three parameters α , β and γ are estimated in this model. Level, Trend and Seasonality are accounted for in this model.

Autofit parameters for the Triple Exponential smoothing:

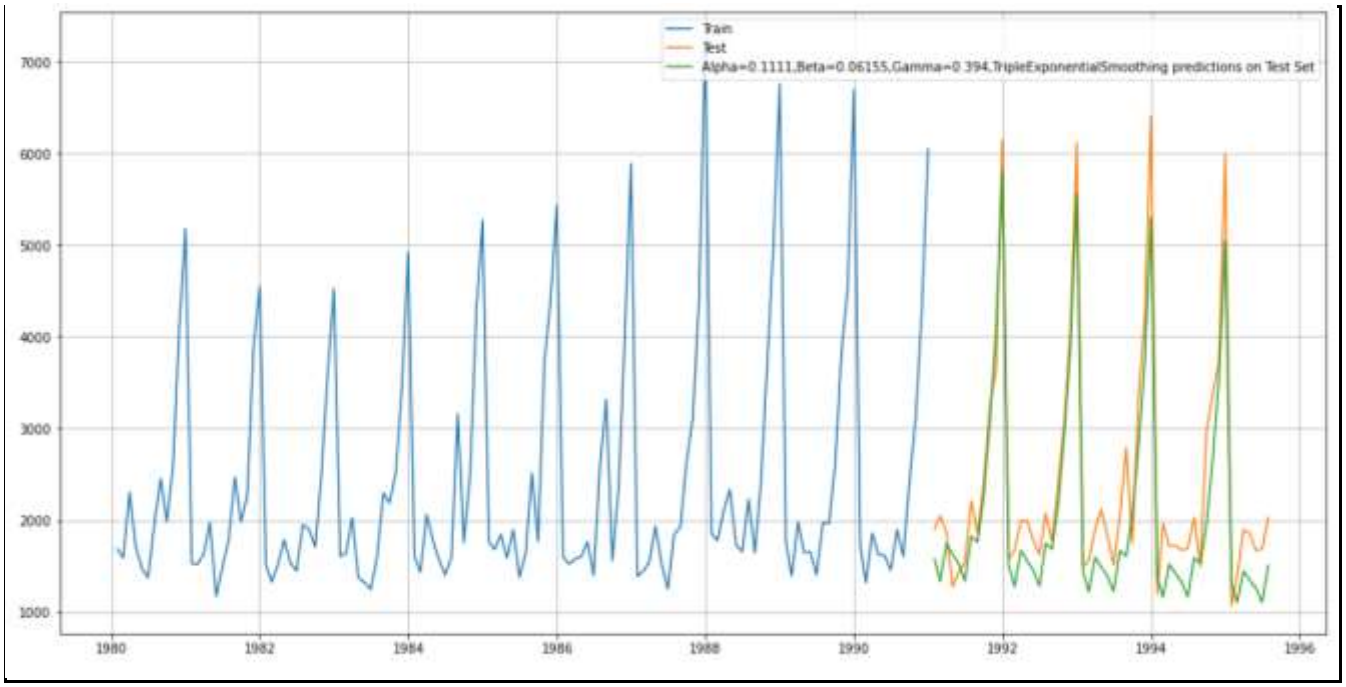
```
{'smoothing_level': 0.11110109539432127,
 'smoothing_trend': 0.06155956038741422,
 'smoothing_seasonal': 0.39402402538387826,
 'damping_trend': nan,
 'initial_level': 1637.1063954979618,
 'initial_trend': -9.145552698252029,
 'initial_seasons': array([1.05835222, 1.01510513, 1.40303179, 1.19830843, 0.96806777,
       0.97034883, 1.32237078, 1.70460903, 1.37084089, 1.80737708,
       2.83425686, 3.61646361]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

The above fit of the model is by the best parameters that Python thinks for the model. It uses a brute force method to choose the parameters.

Lets predict on the test data:

Time_Stamp	Sparkling	auto_predict
1991-01-31	1902	1577.39
1991-02-28	2049	1334.20
1991-03-31	1874	1746.25
1991-04-30	1279	1630.97
1991-05-31	1432	1523.58

Plotting on both the Training and Test using autofit



For Alpha=0.111,Beta=0.06155,Gamma=0.394, Triple Exponential Smoothing Model forecast on the Test Data, RMSE is 468.758

	Test RMSE
Alpha=0.995,Simple Exponential Model	1316.03
Alpha=0.3,SimpleExponentialSmoothing	1935.51
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	18259.11
Alpha=0.1111,Beta=0.06155,Gamma=0.394,TripleExponentialSmoothing	468.76

Lets define an empty dataframe to store our values from the loop

Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE

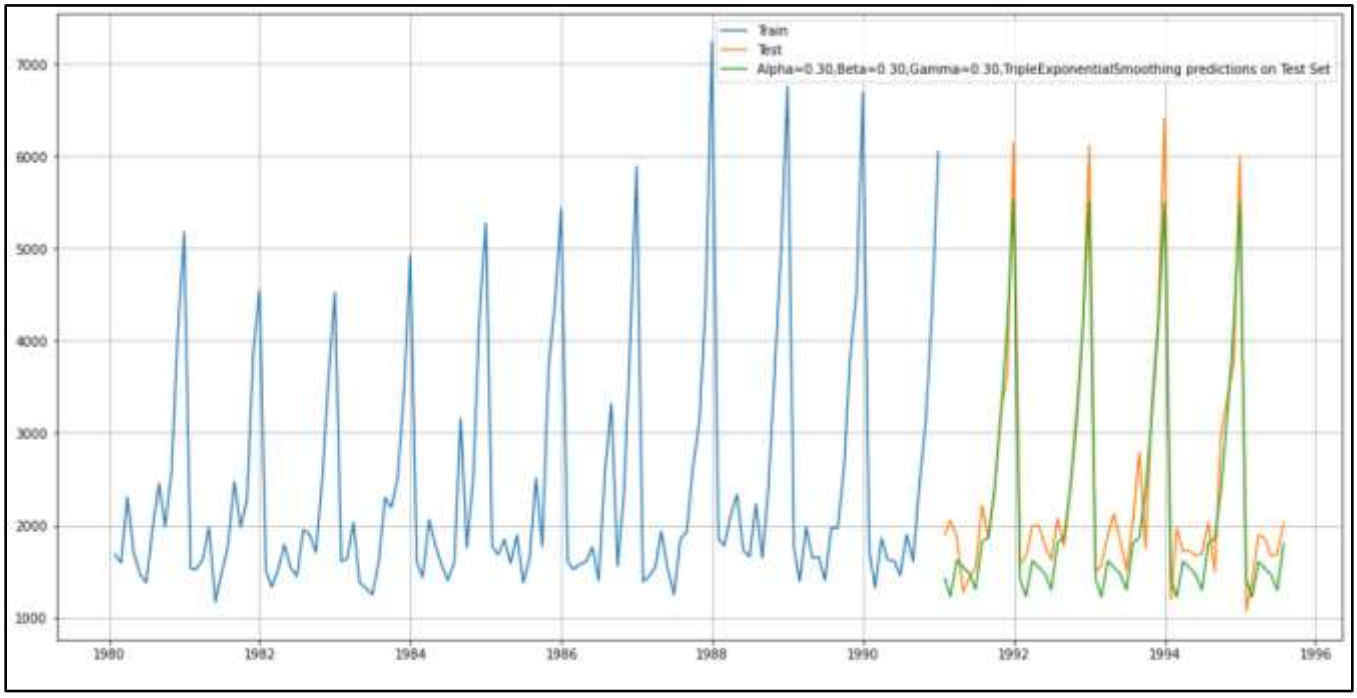
We will append all the results to it, by changing values of alpha, beta , gamma.

After Sorting the first five rows are

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE
0	0.30	0.30	0.30	404.51	392.79
8	0.30	0.40	0.30	424.83	410.85
65	0.40	0.30	0.40	435.55	421.41
296	0.70	0.80	0.30	700.32	518.19
130	0.50	0.30	0.50	498.24	542.18

So this model is giving RMSE 392.79 lowest so far, and Ideal values are 0.3,0.3,0.3 for Alpha, Beta , Gamma.

Plotting on both the Training and Test data using brute force alpha, beta and gamma determination



The sorted results of all the models so far:

	Test RMSE
Alpha=0.3,Beta=0.3,Gamma=0.3,TripleExponentialSmoothing	392.79
Alpha=0.1111,Beta=0.06155,Gamma=0.394,TripleExponentialSmoothing	468.76
Alpha=0.995,Simple Exponential Model	1316.03
Alpha=0.3,SimpleExponentialSmoothing	1935.51
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	18259.11

For this data, we had both trend and seasonality so by definition Triple Exponential Smoothing is supposed to work better than the Simple Exponential Smoothing as well as the Double Exponential Smoothing
Lets Build Other Models Also.

Model 4: Linear Regression

For this particular linear regression, we are going to regress the 'Sparkling-Sales' variable against the order of the occurrence. For this we need to modify our training data before fitting it into a linear regression.

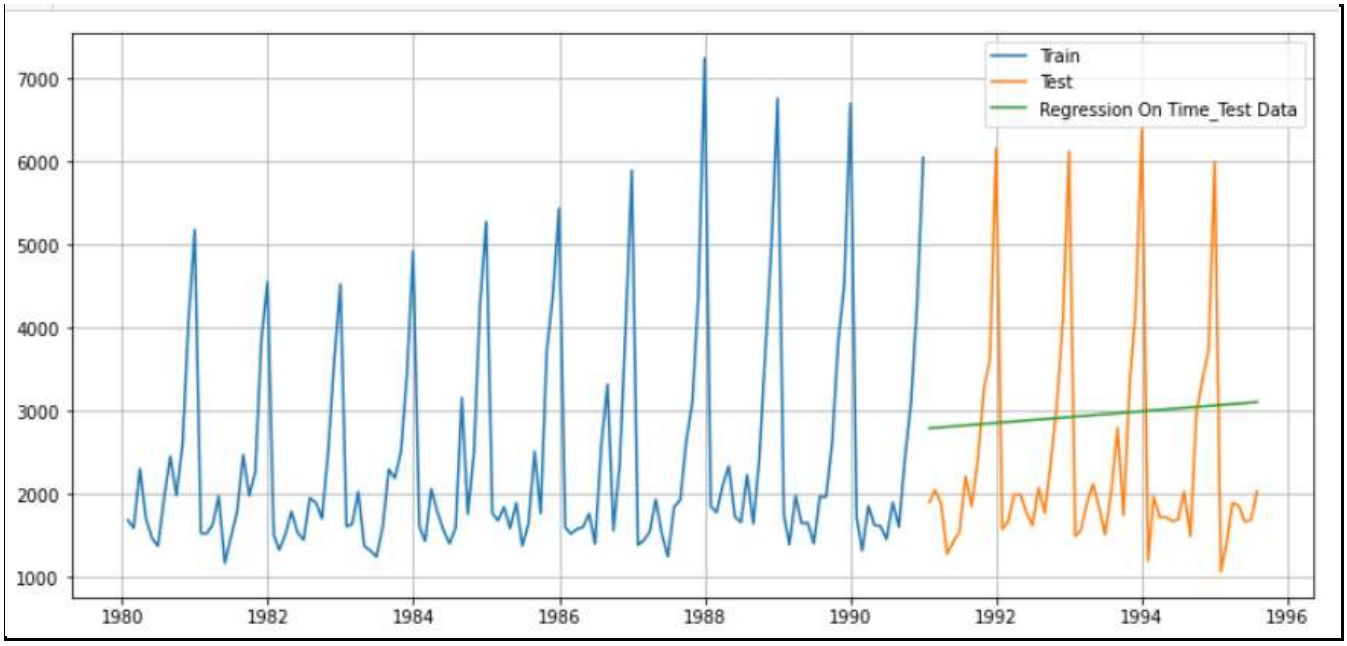
```
Training Time instance
 [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 3
4, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65,
66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97,
98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123,
124, 125, 126, 127, 128, 129, 130, 131, 132]
Test Time instance
 [133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157,
158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 18
3, 184, 185, 186, 187]
```

We see that we have successfully the generated the numerical time instance order for both the training and test set. Now we will add these values in the training and test set.

```
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
```

Now our training and test data has been modified, let us go ahead use Linear regression to build the model on the training data and test the model on the test data.

Prediction plot:



Model Evaluation

For RegressionOnTime forecast on the Test Data, RMSE is 1389.35

	Test RMSE
Alpha=0.995,Simple Exponential Model	1316.03
Alpha=0.3,SimpleExponentialSmoothing	1935.51
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	18259.11
Alpha=0.1111,Beta=0.06155,Gamma=0.394,TripleExponentialSmoothing	468.76
Alpha=0.3,Beta=0.3,Gamma=0.3,TripleExponentialSmoothing	392.79
RegressionOnTime	1389.14

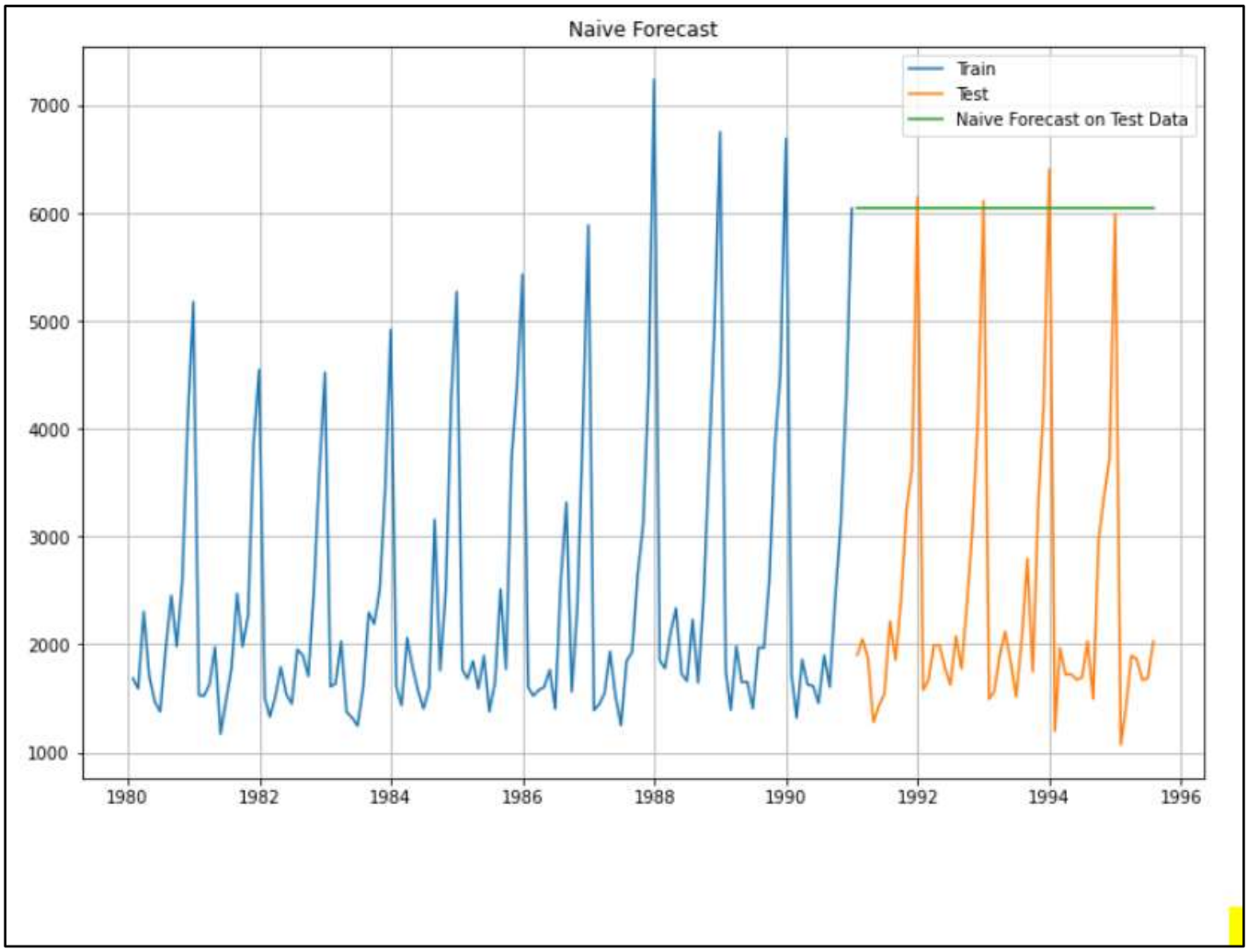
Model 5: Naive Approach:

For this particular naive model, we say that the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow and since the prediction of tomorrow is same as today,therefore the prediction for day after tomorrow is also today.

Prediction : First 5 rows:

```
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
```

Prediction Plot:



Model Evaluation

For RegressionOnTime forecast on the Test Data, RMSE is 3864.279

	Test RMSE
Alpha=0.995,Simple Exponential Model	1316.03
Alpha=0.3,SimpleExponentialSmoothing	1935.51
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	18259.11
Alpha=0.1111,Beta=0.06155,Gamma=0.394,TripleExponentialSmoothing	468.76
Alpha=0.3,Beta=0.3,Gamma=0.3,TripleExponentialSmoothing	392.79
RegressionOnTime	1389.14
NaiveModel	3864.28

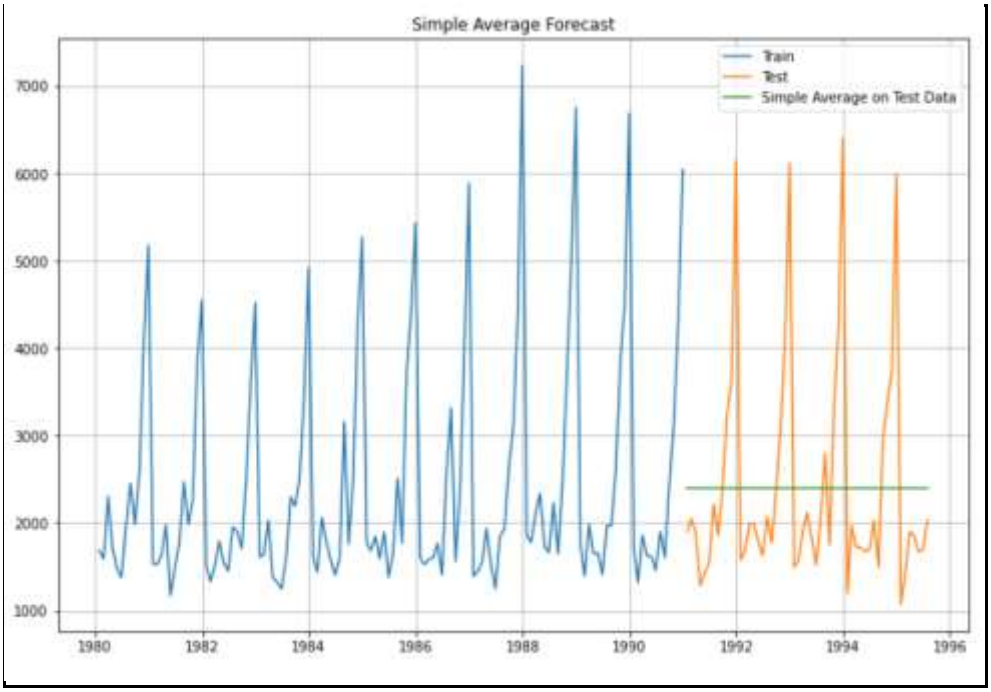
Model 6: Simple Average

For this particular simple average method, we will forecast by using the average of the training values.

Forecast:

Time_Stamp	Sparkling	mean_forecast
1991-01-31	1902	2403.78
1991-02-28	2049	2403.78
1991-03-31	1874	2403.78
1991-04-30	1279	2403.78
1991-05-31	1432	2403.78

Prediction Plot vs Test Data:



Model Evaluation

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

	Test RMSE
Alpha=0.995,Simple Exponential Model	1316.03
Alpha=0.3,SimpleExponentialSmoothing	1935.51
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	18259.11
Alpha=0.1111,Beta=0.06155,Gamma=0.394,TripleExponentialSmoothing	468.76
Alpha=0.3,Beta=0.3,Gamma=0.3,TripleExponentialSmoothing	392.79
RegressionOnTime	1389.14
NaiveModel	3864.28
SimpleAverageModel	1275.08

Model 7: Moving Average(MA)

For the moving average model, we are going to calculate rolling means (or moving averages) for different intervals. The best interval can be determined by the maximum accuracy (or the minimum error) over here.

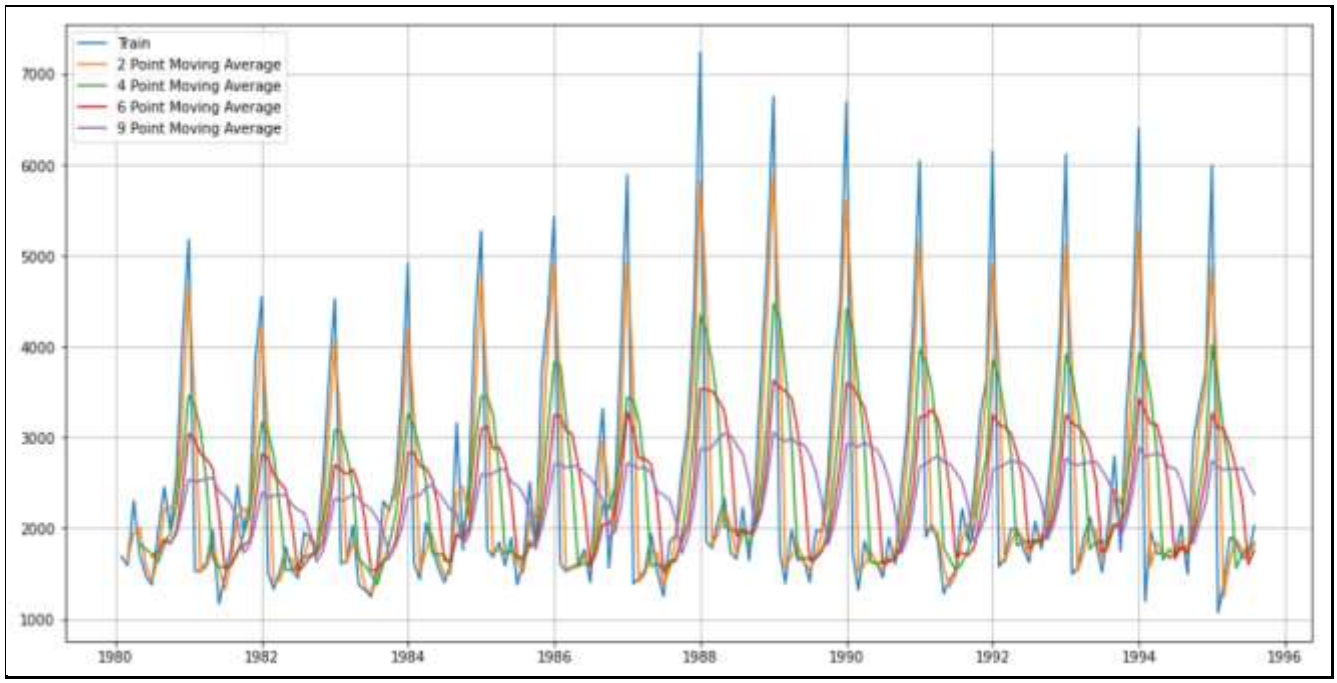
For Moving Average, we are going to average over the entire data.

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

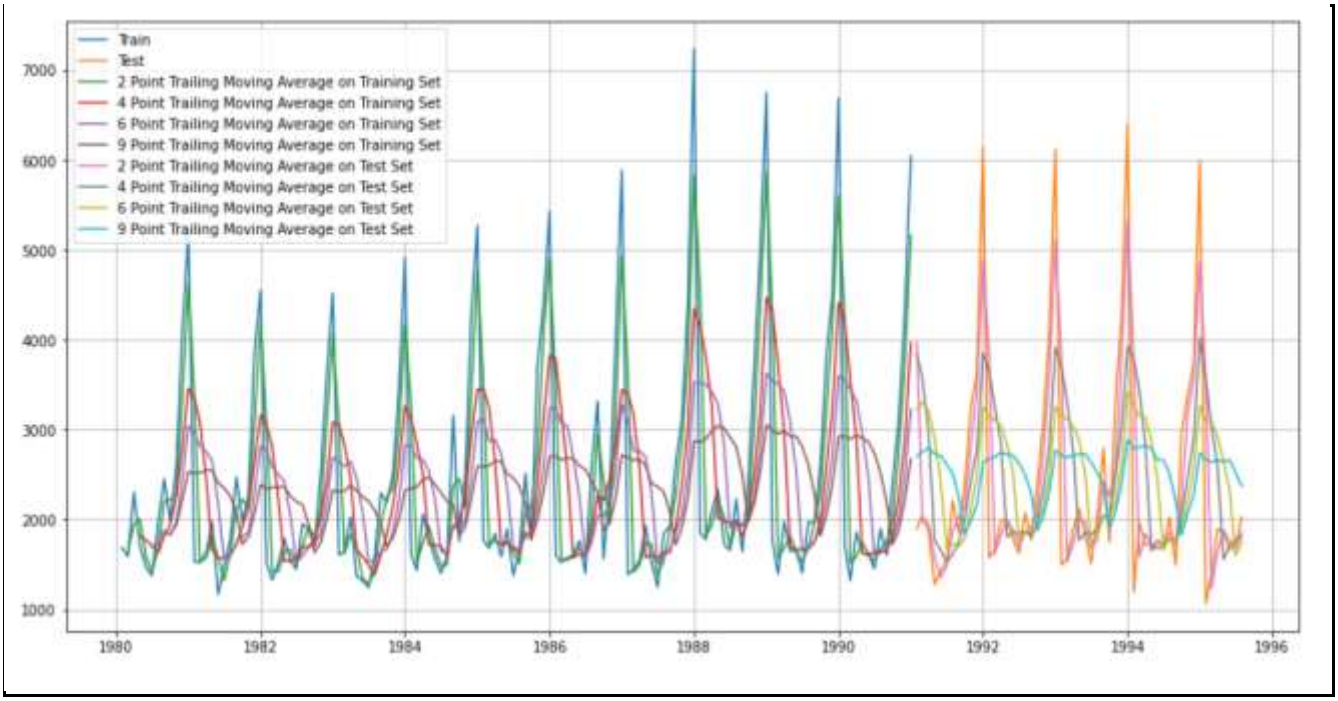
For Trailing Average,

Time_Stamp	Sparkling	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1980-01-31	1686	nan	nan	nan	nan
1980-02-29	1591	1638.50	nan	nan	nan
1980-03-31	2304	1947.50	nan	nan	nan
1980-04-30	1712	2008.00	1823.25	nan	nan
1980-05-31	1471	1591.50	1769.50	nan	nan

Plot for 2,4,6,9 Point Moving Average



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.



Model Evaluation

Done only on Test data

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

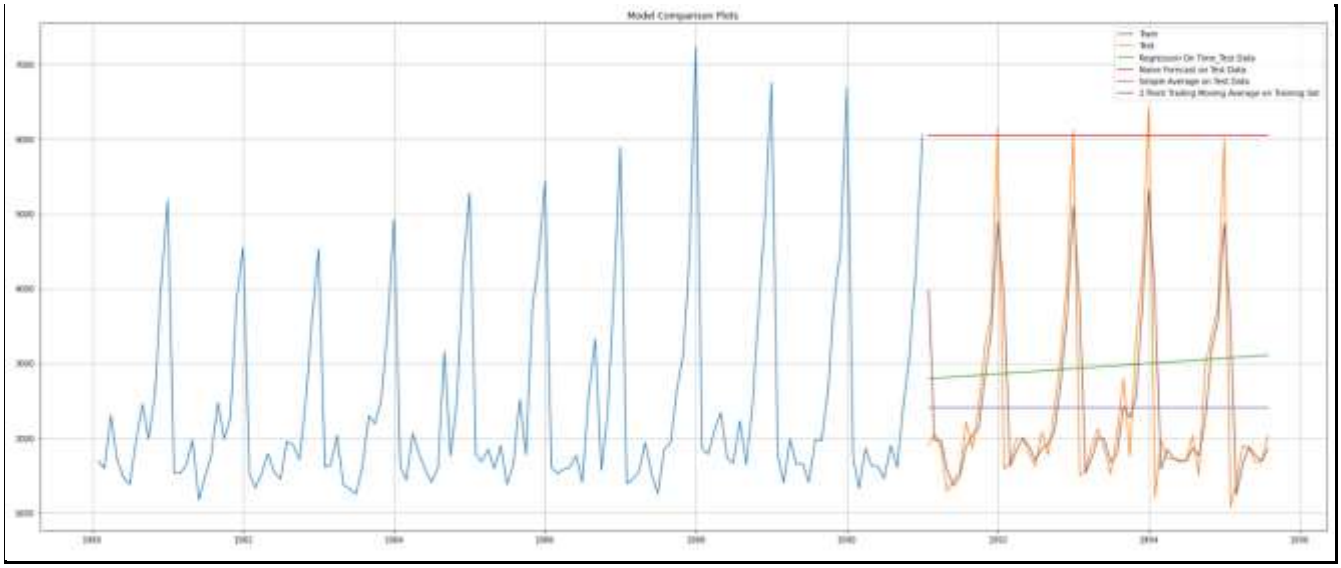
For 6 point Moving Average Model forecast on the Training Data, RMSE is 1283.927

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

Test RMSE results for All models:

	Test RMSE
Alpha=0.995,Simple Exponential Model	1316.03
Alpha=0.3,SimpleExponentialSmoothing	1935.51
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	18259.11
Alpha=0.1111,Beta=0.06155,Gamma=0.394,TripleExponentialSmoothing	468.76
Alpha=0.3,Beta=0.3,Gamma=0.3,TripleExponentialSmoothing	392.79
RegressionOnTime	1389.14
NaiveModel	3864.28
SimpleAverageModel	1275.08
2pointTrailingMovingAverage	813.40
4pointTrailingMovingAverage	1156.59
6pointTrailingMovingAverage	1283.93
9pointTrailingMovingAverage	1346.28

Plotting on both Training and Test data



Sorted by RMSE values on the Test Data:

	Test RMSE
Alpha=0.3,Beta=0.3,Gamma=0.3,TripleExponentialSmoothing	392.79
Alpha=0.1111,Beta=0.06155,Gamma=0.394,TripleExponentialSmoothing	468.76
2pointTrailingMovingAverage	813.40
4pointTrailingMovingAverage	1156.59
SimpleAverageModel	1275.08
6pointTrailingMovingAverage	1283.93
Alpha=0.995,Simple Exponential Model	1316.03
9pointTrailingMovingAverage	1346.28
RegressionOnTime	1389.14
Alpha=0.3,SimpleExponentialSmoothing	1935.51
NaiveModel	3864.28
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	18259.11

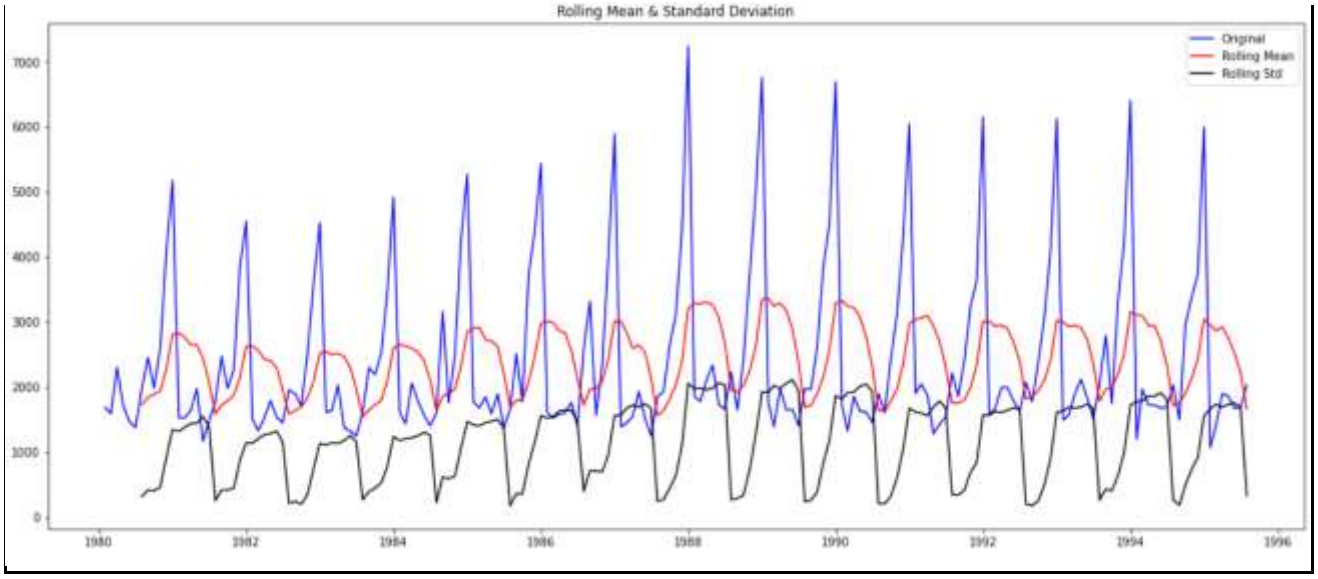
5.Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05.

Check for Stationarity:

Dicky Fuller Test

Null Hypothesis Ho- Series is not Stationary

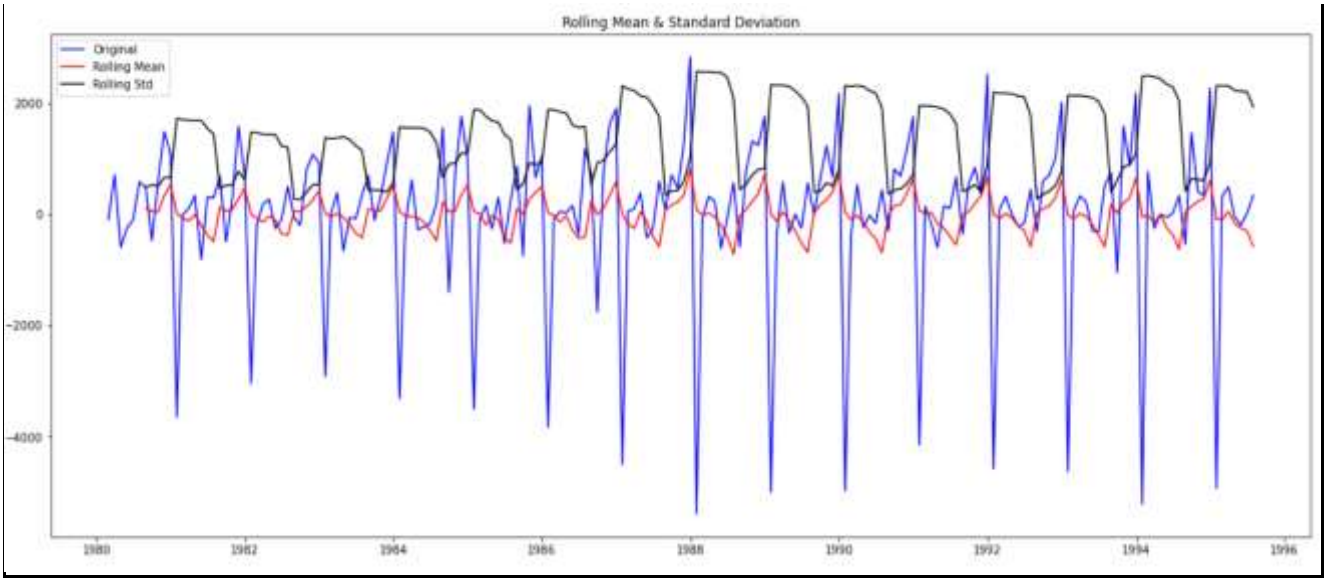
Alternative Hypothesis H1- Series is Stationary



P value is higher that 0.05 Hence Null Hypothesis is True, we will take Order 1 differencing to make series stationary.

```
Results of Dickey-Fuller Test:
Test Statistic                  -1.36
p-value                          0.60
#Lags Used                      11.00
Number of Observations Used    175.00
Critical Value (1%)             -3.47
Critical Value (5%)             -2.88
Critical Value (10%)            -2.58
dtype: float64
```

We check stationarity at initial level, but series is not stationary as P value is higher than 0.05 difference of order 1.




```
Results of Dickey-Fuller Test:
Test Statistic                 -45.05
p-value                          0.00
#Lags Used                      10.00
Number of Observations Used    175.00
Critical Value (1%)             -3.47
Critical Value (5%)             -2.88
Critical Value (10%)            -2.58
dtype: float64
```

We see that at P value is less than 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. 8

Automated Verion of ARIMA

The following loop helps us in getting a combination of different parameters of p and q in the range of 0 and 2. We have kept the value of d as 1 as we need to take a difference of the series to make it stationary.

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

We will Apply all values of p and q. check which combination is giving us the low AIC score.

```
ARIMA(0, 1, 0) - AIC:2269.582796371201
ARIMA(0, 1, 1) - AIC:2264.906436993431
ARIMA(0, 1, 2) - AIC:2232.7830976843675
ARIMA(1, 1, 0) - AIC:2268.5280605652692
ARIMA(1, 1, 1) - AIC:2235.0139453492384
ARIMA(1, 1, 2) - AIC:2233.5976471203803
ARIMA(2, 1, 0) - AIC:2262.035600101831
ARIMA(2, 1, 1) - AIC:2232.3604898820927
ARIMA(2, 1, 2) - AIC:2210.6165947934273
```

Sort the above AIC values in the ascending order to get the parameters for the minimum AIC value

	param	AIC
8	(2, 1, 2)	2210.62
7	(2, 1, 1)	2232.36
2	(0, 1, 2)	2232.78
5	(1, 1, 2)	2233.60
4	(1, 1, 1)	2235.01
6	(2, 1, 0)	2262.04
1	(0, 1, 1)	2264.91
3	(1, 1, 0)	2268.53
0	(0, 1, 0)	2269.58

ARIMA Model Results

Dep. Variable:	D.Sparkling	No. Observations:	131
Model:	ARIMA(2, 1, 2)	Log Likelihood	-1099.308
Method:	css-mle	S.D. of innovations	1011.613
Date:	Tue, 22 Jun 2021	AIC	2210.617
Time:	12:08:24	BIC	2227.868
Sample:	02-29-1980 - 12-31-1990	HQIC	2217.627

	coef	std err	z	P>|z|	[0.025	0.975]
const	5.5861	0.516	10.825	0.000	4.575	6.597
ar.L1.D.Sparkling	1.2699	0.074	17.047	0.000	1.124	1.416
ar.L2.D.Sparkling	-0.5602	0.074	-7.618	0.000	-0.704	-0.416
ma.L1.D.Sparkling	-2.0000	0.042	-47.103	0.000	-2.083	-1.917
ma.L2.D.Sparkling	1.0000	0.042	23.588	0.000	0.917	1.083

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.1335	-0.7073j	1.3361	-0.0888
AR.2	1.1335	+0.7073j	1.3361	0.0888
MA.1	1.0000	-0.0004j	1.0000	-0.0001
MA.2	1.0000	+0.0004j	1.0000	0.0001

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

	Test RMSE
ARIMA(2,1,2)	1375.26

Performance of Models so far:

	Test RMSE
Alpha=0.995,Simple Exponential Model	1316.03
Alpha=0.3,SimpleExponentialSmoothing	1935.51
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	18259.11
Alpha=0.1111,Beta=0.06155,Gamma=0.394,TripleExponentialSmoothing	468.76
Alpha=0.3,Beta=0.3,Gamma=0.3,TripleExponentialSmoothing	392.79
RegressionOnTime	1389.14
NaiveModel	3864.28
SimpleAverageModel	1275.08
2pointTrailingMovingAverage	813.40
4pointTrailingMovingAverage	1156.59
6pointTrailingMovingAverage	1283.93
9pointTrailingMovingAverage	1346.28
ARIMA(2,1,2)	1375.26

Still this point Triple exponential has performed the best.

Automated Version of SARIMA

In Model SARIMA, we are considering Seasonal P,D,Q,S.

It is extension of ARIMA model by considering Seasonality factor.

```
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)
```

Top 5 Parameters AIC score:

	param	seasonal	AIC
50	(1, 1, 2)	(1, 0, 2, 12)	1555.58
53	(1, 1, 2)	(2, 0, 2, 12)	1556.08
26	(0, 1, 2)	(2, 0, 2, 12)	1557.12
23	(0, 1, 2)	(1, 0, 2, 12)	1557.16
77	(2, 1, 2)	(1, 0, 2, 12)	1557.34

From the above result we will choose (1,1,2)(1,0,2,12).

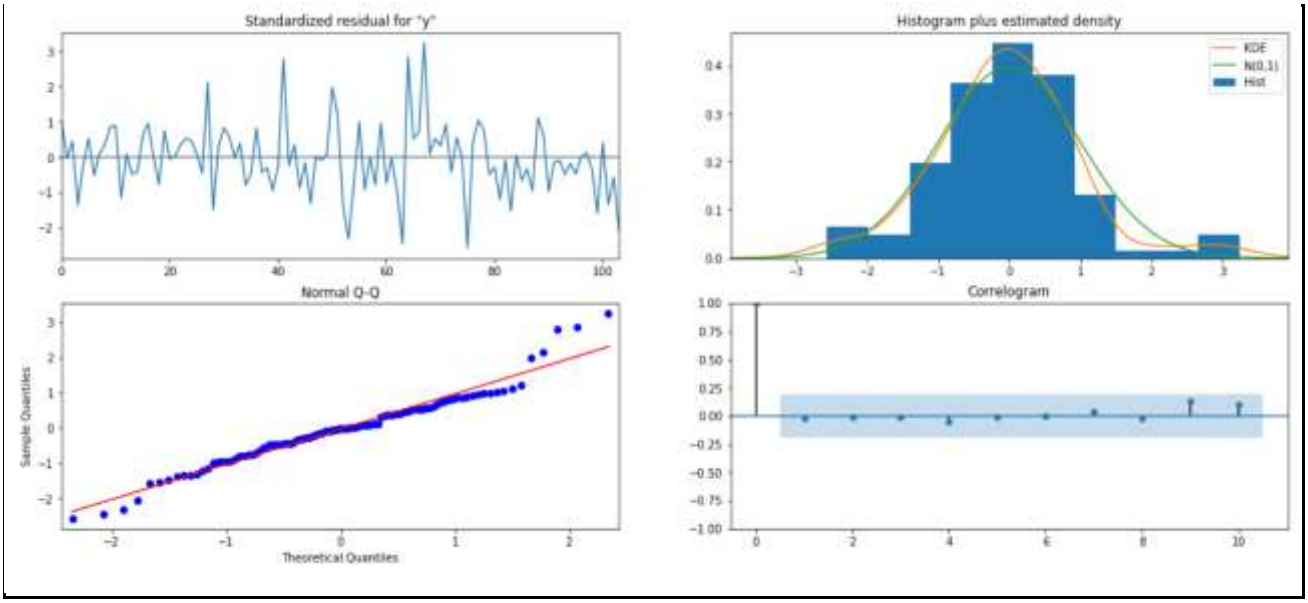
SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1, 1, 2)x(1, 0, 2, 12)	Log Likelihood	-770.792
Date:	Tue, 22 Jun 2021	AIC	1555.584
Time:	13:40:38	BIC	1574.095
Sample:	0	HQIC	1563.083
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.6279	0.255	-2.461	0.014	-1.128	-0.128
ma.L1	-0.1042	0.225	-0.463	0.643	-0.545	0.337
ma.L2	-0.7275	0.154	-4.731	0.000	-1.029	-0.426
ar.S.L12	1.0439	0.014	72.820	0.000	1.016	1.072
ma.S.L12	-0.5549	0.098	-5.662	0.000	-0.747	-0.363
ma.S.L24	-0.1354	0.120	-1.133	0.257	-0.370	0.099
sigma2	1.506e+05	2.04e+04	7.400	0.000	1.11e+05	1.9e+05

Ljung-Box (L1) (Q):	0.04	Jarque-Bera (JB):	11.72
Prob(Q):	0.84	Prob(JB):	0.00
Heteroskedasticity (H):	1.47	Skew:	0.36
Prob(H) (two-sided):	0.26	Kurtosis:	4.48

Results:



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

mean	mean_se	mean_ci_lower	mean_ci_upper
1327.34	388.36	566.17	2088.51
1315.07	402.03	527.10	2103.03
1621.60	402.02	833.65	2409.55
1598.81	407.26	800.59	2397.02
1392.71	407.99	593.06	2192.36

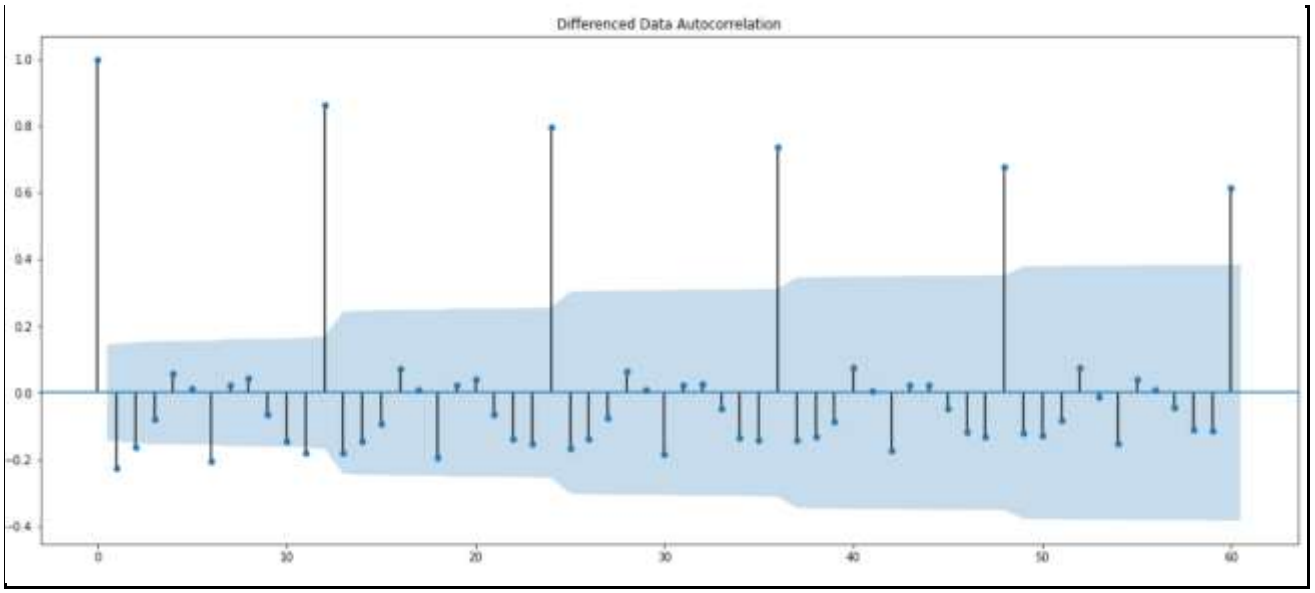
RMSE for SARIMA model is 528.64.

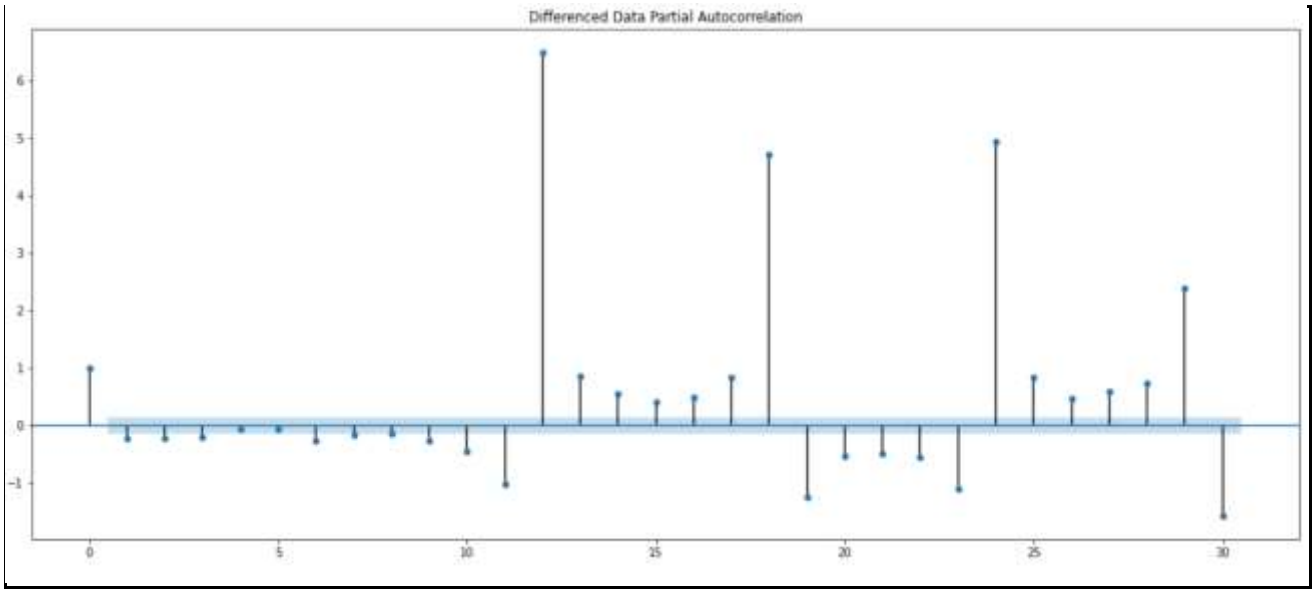
	Test RMSE
Alpha=0.995,Simple Exponential Model	1316.03
Alpha=0.3,SimpleExponentialSmoothing	1935.51
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	18259.11
Alpha=0.1111,Beta=0.06155,Gamma=0.394,TripleExponentialSmoothing	468.76
Alpha=0.3,Beta=0.3,Gamma=0.3,TripleExponentialSmoothing	392.79
RegressionOnTime	1389.14
NaiveModel	3864.28
SimpleAverageModel	1275.08
2pointTrailingMovingAverage	813.40
4pointTrailingMovingAverage	1156.59
6pointTrailingMovingAverage	1283.93
9pointTrailingMovingAverage	1346.28
ARIMA(2,1,2)	1375.26
SARIMA(1,1,2)(1,0,2,12)	528.64

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.

Manual ARIMA

Let us look at the ACF and the PACF plots





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 3 and 2.

ARIMA Model Results

Dep. Variable:	D.Sparkling	No. Observations:	131
Model:	ARIMA(3, 1, 2)	Log Likelihood	-1107.465
Method:	css-mle	S.D. of innovations	1106.521
Date:	Tue, 22 Jun 2021	AIC	2228.930
Time:	13:42:16	BIC	2249.057
Sample:	02-29-1980 - 12-31-1990	HQIC	2237.108

	coef	std err	z	P>|z|	[0.025	0.975]
const	5.8937	3.643	1.618	0.106	-1.246	13.033
ar.L1.D.Sparkling	-0.4420	8.84e-05	-5002.292	0.000	-0.442	-0.442
ar.L2.D.Sparkling	0.3078	0.000	932.303	0.000	0.307	0.308
ar.L3.D.Sparkling	-0.2502	0.000	-986.101	0.000	-0.251	-0.250
ma.L1.D.Sparkling	-0.0013	0.017	-0.076	0.939	-0.034	0.031
ma.L2.D.Sparkling	-0.9987	0.017	-59.919	0.000	-1.031	-0.966

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-1.0000	-0.0000j	1.0000	-0.5000
AR.2	1.1150	-1.6593j	1.9991	-0.1558
AR.3	1.1150	+1.6593j	1.9991	0.1558
MA.1	1.0000	+0.0000j	1.0000	0.0000
MA.2	-1.0013	+0.0000j	1.0013	0.5000

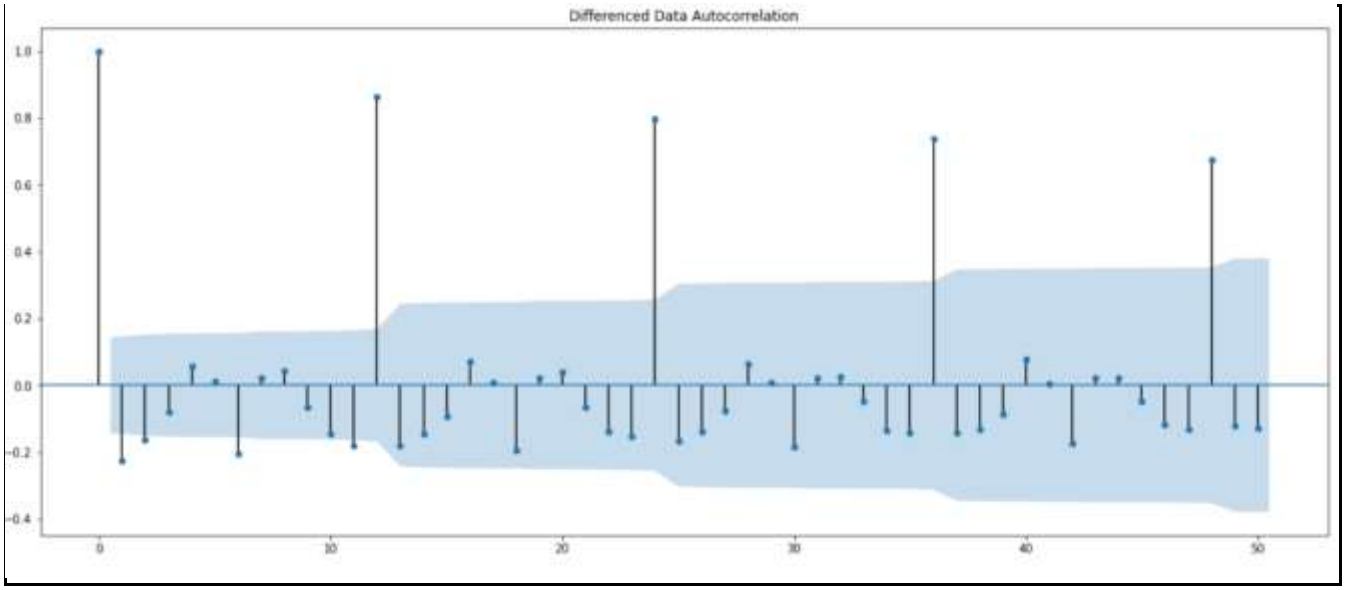
Predict on the Test Set using this model and evaluate the model.

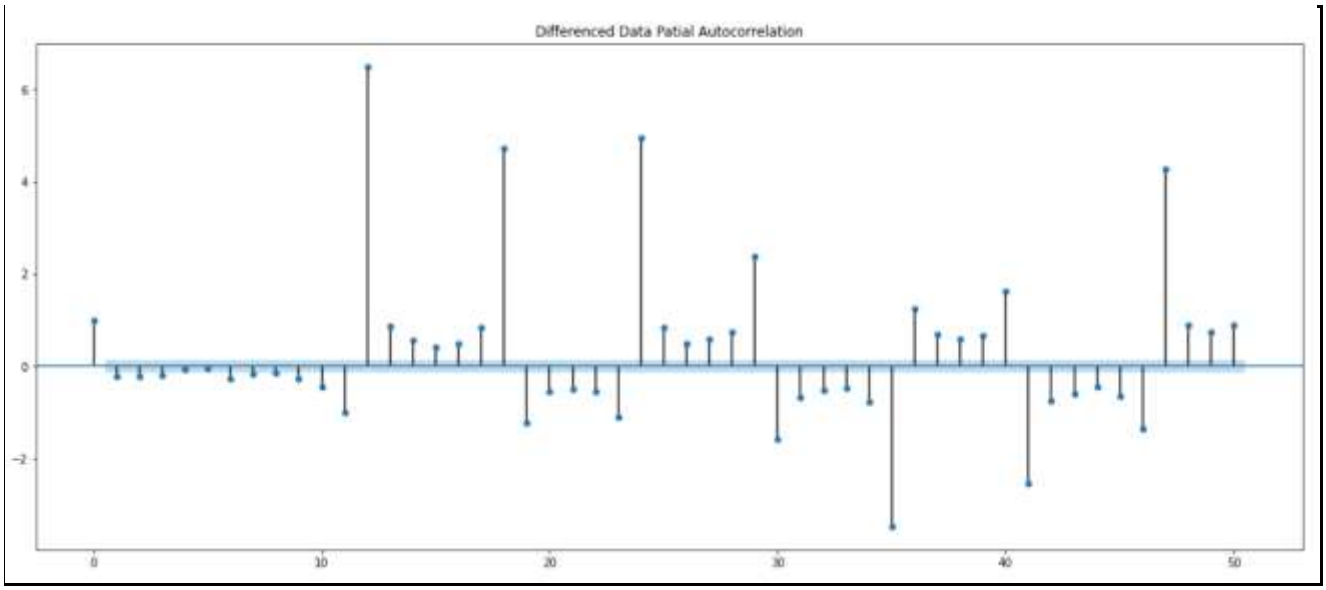
RMSE for Manual ARIMA model is 1375.80.

The Overall Results till now,

	Test RMSE
Alpha=0.995,Simple Exponential Model	1316.03
Alpha=0.3,SimpleExponentialSmoothing	1935.51
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	18259.11
Alpha=0.1111,Beta=0.06155,Gamma=0.394,TripleExponentialSmoothing	468.76
Alpha=0.3,Beta=0.3,Gamma=0.3,TripleExponentialSmoothing	392.79
RegressionOnTime	1389.14
NaiveModel	3864.28
SimpleAverageModel	1275.08
2pointTrailingMovingAverage	813.40
4pointTrailingMovingAverage	1156.59
6pointTrailingMovingAverage	1283.93
9pointTrailingMovingAverage	1346.28
ARIMA(2,1,2)	1375.26
SARIMA(1,1,2)(1,0,2,12)	528.64
ARIMA(3,1,2)	1375.86

SARIMA Model : Manually looking at ACF and PACF





The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 0. The Moving-Average parameter in an SARIMA model is 'q' which comes from the significant lag after which the ACF plot cuts-off to 0. Remember to check the ACF and the PACF plots only at multiples of 12 (since 12 is the seasonal period). By looking at the plots we see that the ACF and the PACF do not directly cut-off to 0.

This is a common problem while building models by looking at the ACF and the PACF plots. But we are able to explain the model.

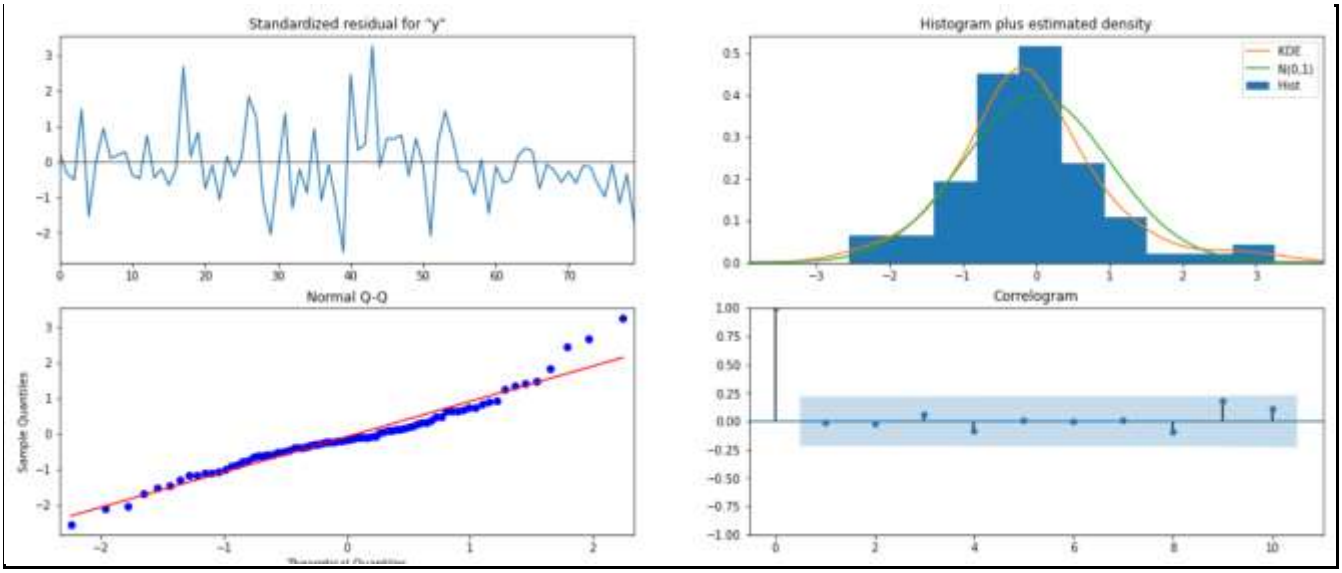
SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(3, 1, 2)x(3, 1, 2, 12)	Log Likelihood	-598.630
Date:	Tue, 22 Jun 2021	AIC	1219.260
Time:	13:42:41	BIC	1245.462
Sample:	0	HQIC	1229.765
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.7555	0.151	-5.008	0.000	-1.051	-0.460
ar.L2	0.1169	0.185	0.633	0.527	-0.245	0.479
ar.L3	-0.0520	0.143	-0.365	0.715	-0.332	0.228
ma.L1	0.0331	0.191	0.173	0.863	-0.341	0.407
ma.L2	-0.9669	0.156	-6.194	0.000	-1.273	-0.661
ar.S.L12	-0.7531	0.496	-1.518	0.129	-1.725	0.219
ar.S.L24	-0.6368	0.350	-1.818	0.069	-1.324	0.050
ar.S.L36	-0.2469	0.151	-1.640	0.101	-0.542	0.048
ma.S.L12	0.3712	0.491	0.756	0.450	-0.591	1.333
ma.S.L24	0.3466	0.365	0.949	0.343	-0.369	1.063
sigma2	1.791e+05	1.67e-06	1.07e+11	0.000	1.79e+05	1.79e+05

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	13.17
Prob(Q):	0.93	Prob(JB):	0.00
Heteroskedasticity (H):	0.66	Skew:	0.62
Prob(H) (two-sided):	0.29	Kurtosis:	4.55

Results Dignostics for Manual SARIMA:



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

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1510.10	425.19	676.73	2343.47
1	1431.60	440.19	568.84	2294.35
2	1850.32	440.26	987.42	2713.21
3	1781.85	441.01	917.49	2646.21
4	1550.30	441.00	685.96	2414.64

RMSE for Manual SARIMA is 329.55

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.

	Test RMSE
Alpha=0.995,Simple Exponential Model	1316.03
Alpha=0.3,SimpleExponentialSmoothing	1935.51
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	18259.11
Alpha=0.1111,Beta=0.06155,Gamma=0.394,TripleExponentialSmoothing	468.76
Alpha=0.3,Beta=0.3,Gamma=0.3,TripleExponentialSmoothing	392.79
RegressionOnTime	1389.14
NaiveModel	3864.28
SimpleAverageModel	1275.08
2pointTrailingMovingAverage	813.40
4pointTrailingMovingAverage	1156.59
6pointTrailingMovingAverage	1283.93
9pointTrailingMovingAverage	1346.28
ARIMA(2,1,2)	1375.26
SARIMA(1,1,2)(1,0,2,12)	528.64
ARIMA(3,1,2)	1375.86
SARIMA(3,1,2)(3,1,2,12)	329.56

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.

If we see all the RMSE, We can Say that SARIMA Model can perform Well for this series, This series has level, Seasonality and Trend also.

So we will Select SARIMA model, we will give train data as Whole Data. That is, Test and Train data earlier, Now all data will be used for Training for Purpose.

We are predicting here for Next 12 Months.

Building the most optimum model on the Full Data.

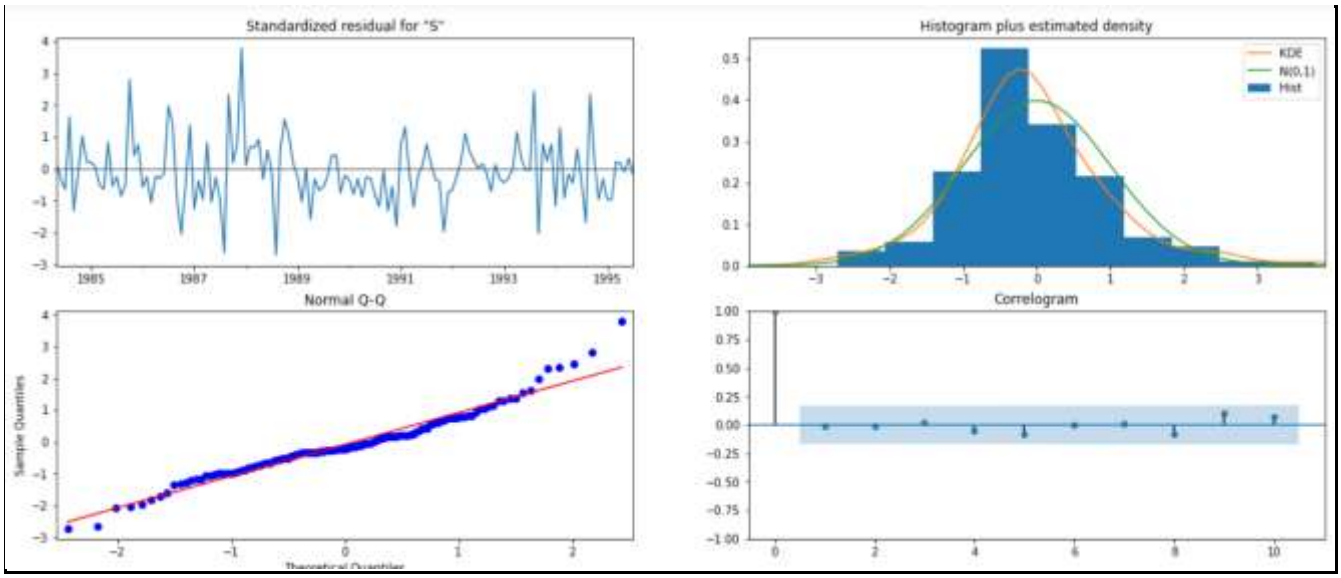
SARIMAX Results

Dep. Variable:		Sparkling	No. Observations:			187
Model:	SARIMAX(3, 1, 2)x(3, 1, 2, 12)		Log Likelihood			-1000.243
Date:	Tue, 22 Jun 2021		AIC			2022.487
Time:	13:47:56		BIC			2054.445
Sample:	01-31-1980		HQIC			2035.473
	- 07-31-1995					
Covariance Type:	opg					

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.8609	0.090	-9.545	0.000	-1.038	-0.684
ar.L2	0.0119	0.129	0.092	0.926	-0.241	0.265
ar.L3	-0.0766	0.102	-0.753	0.451	-0.276	0.123
ma.L1	0.0322	0.120	0.269	0.788	-0.203	0.267
ma.L2	-0.9677	0.098	-9.839	0.000	-1.161	-0.775
ar.S.L12	-0.6101	0.392	-1.555	0.120	-1.379	0.159
ar.S.L24	-0.4981	0.231	-2.160	0.031	-0.950	-0.046
ar.S.L36	-0.2472	0.109	-2.262	0.024	-0.461	-0.033
ma.S.L12	0.1229	0.396	0.311	0.756	-0.652	0.898
ma.S.L24	0.2488	0.266	0.936	0.349	-0.272	0.770
sigma2	1.562e+05	1.31e-06	1.19e+11	0.000	1.56e+05	1.56e+05

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	26.97
Prob(Q):	0.92	Prob(JB):	0.00
Heteroskedasticity (H):	0.56	Skew:	0.59
Prob(H) (two-sided):	0.05	Kurtosis:	4.84

Results Dignostics:

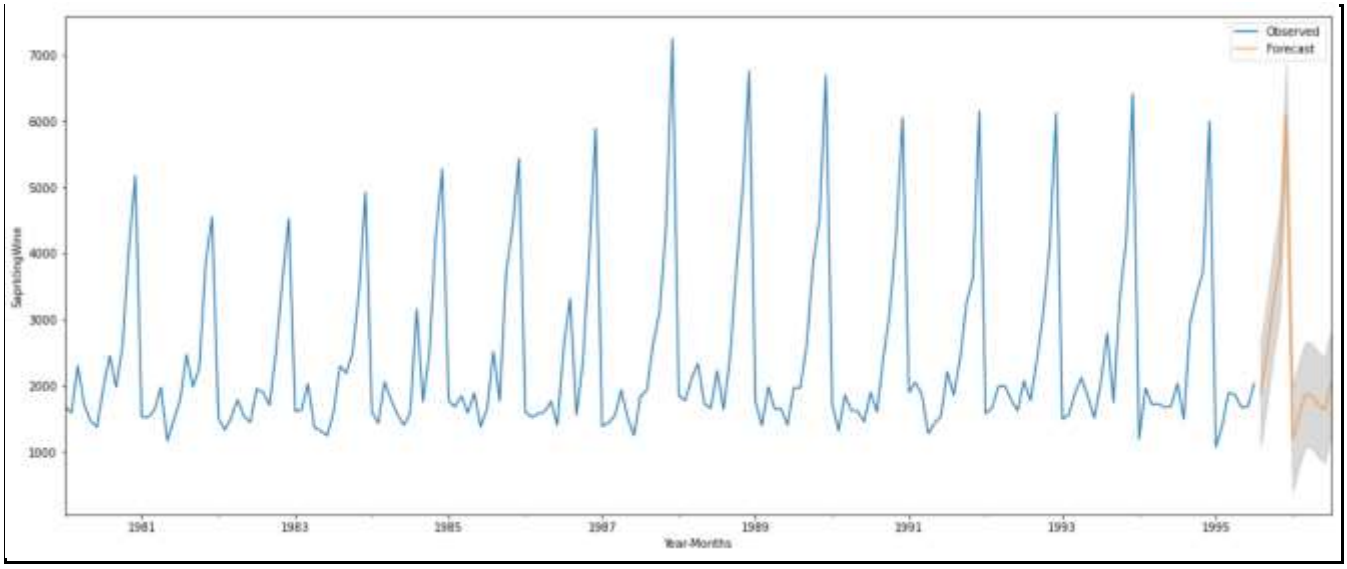


predict 12 months into the future.

Sparkling	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-31	1868.69	396.48	1091.60	2645.78
1995-09-30	2511.36	401.84	1723.76	3298.95
1995-10-31	3272.65	402.69	2483.40	4061.90
1995-11-30	3874.44	403.10	3084.38	4664.50
1995-12-31	6098.96	403.12	5308.85	6889.06
1996-01-31	1191.82	403.83	400.33	1983.31
1996-02-29	1557.11	403.84	765.59	2348.62
1996-03-31	1872.46	404.46	1079.73	2665.19
1996-04-30	1851.38	404.48	1058.61	2644.15
1996-05-31	1719.85	405.06	925.94	2513.76
1996-06-30	1631.71	405.09	837.74	2425.68
1996-07-31	2038.46	405.64	1243.43	2833.49

RMSE of the Full Model 578.9516037053722

plot the forecast along with the confidence band



In The above plot Orange line shows Sales of Sparkling wines in 12 Months.

10.Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales. Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present. 5

To Analyse the sales of Sparkling Wine we got hidden insights by performing Exploaratory Data Analysis. We performed EDA then we get to know

1.From 1981 to 1987 trend is upward, but from 1987 is both upward and downward.

2.Median Value of Sales are remains around 2000.

3.Highest Sales of Sparkling wine is occurred in year 1987, and in 1995 one of the month has brought Lowest sale of wines.

4.From Monthly plot, Sales has been Increasd from August to December. Till this point Sales are around 2000. so We need Higher stock or Production from August to December. October,November,December has sales more than 2000 and it is almost equal to 3000,4000,6000.

5.In month of december atleast we need 7242 Sparkling wines in stock as this is the highest number reached.

After Decomposition it is clear that Series has Trend and Seasonality both.

So we Applied the models like Simple, Double, Triple Exponential Smoothing to the Train data till 1991. and then tested it on Test data. Out of these Triple exponential smoothing seems to be quite effective as the series has all three factors Alpha, Beta, gamma. Season, trend and level. The RMSE is low for Triple exponential smoothing.

Further we have applied models like Linear Regression and Naive Bayes, per as we can see in the Dignostics it failed to predict sales, and RMSE is also high for these models.

After We moved to the models like Simple and Trailing Moving Average.This time it was little better than Linear Regression and Naive Bayes.

After That Finally we Applied ARIMA and SARIMA Automated by AIC scores and Manual ARIMA SARIMA by observing ACF and PACF plots. PACF has given p value while ACF has given as q value.

The series has Seasonality hence The best suited Model was SARIMA, Only 329 RMSE on the Test Data, and When Applied on Full data it has given RMSE 578, Which is still good as compare to Others.

So we Finalize This Model. Here we have taken P as 3 and Q as 2, D=1 as we got stationary series after one order differencing.

Finally, We Predict Sales for Next 12 Months, and in December we will need 6099 Sparkling wines atleast to fulfill the requirement.

In Future Company Should be ready with the Stock of the Year,Company should hire staff as per the Requirement. If Company is producing wines for December then they must be think of manpower required. Excess Production might hamper the Business, so as predicted Company should Target.

If Company want to Increase Revenue, then they have to Think for all year instead of focussing only on December.

THE END