Time Series Forecasting-Sparkling

Pradeep Kumar Mishra

PGP-DSBA Online

Jun_B_21

Date: 16:feb:2022

Content View

List of Figures :

List of Tables :

Problem Statement - 1

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 and Rose.csv

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

Sample of the dataset :

Head datasets :

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

Table-01

Tail Datasets :

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

Table-02

Types of variables and missing values in the dataset :

```
<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
```

- From the above results we can see that there is no missing value present in the dataset.
- There are a total of 187 rows .

Note: We can see in the datasets. YearMonth variable does not format properly so first we use manual add date column.

```
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')
```

Final datasets :

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

Table-03

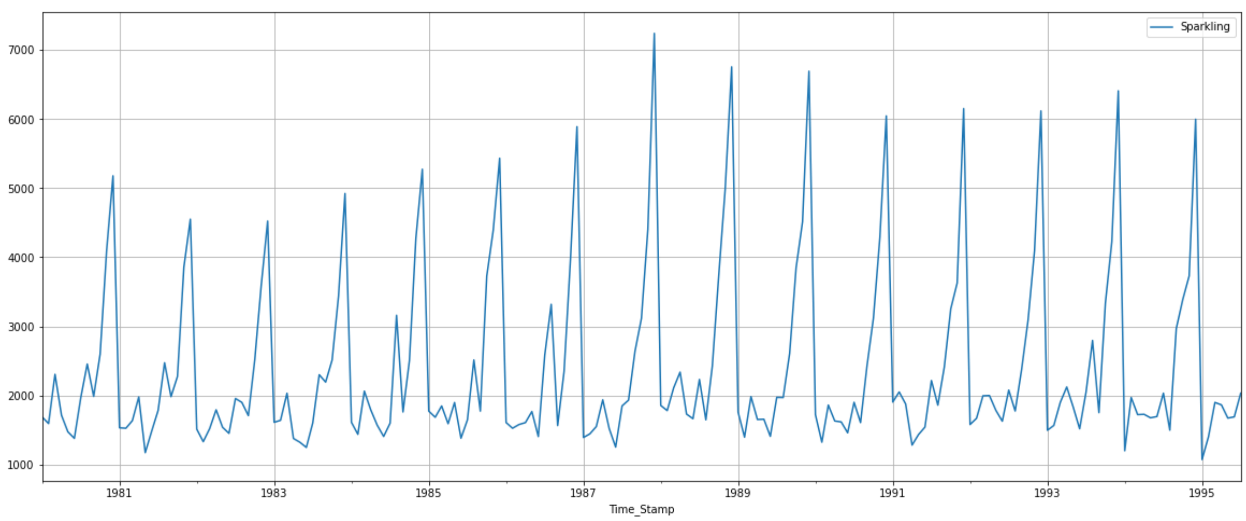

Figure - 01

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

Decompose the datasets :

Model = additive :

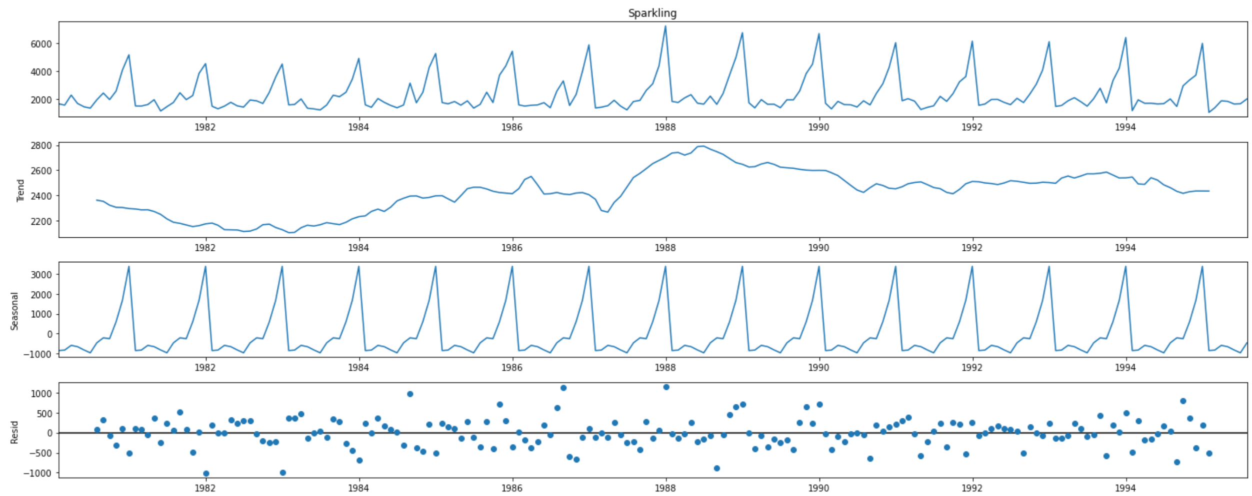

Figure-02

Model= multiplicative :

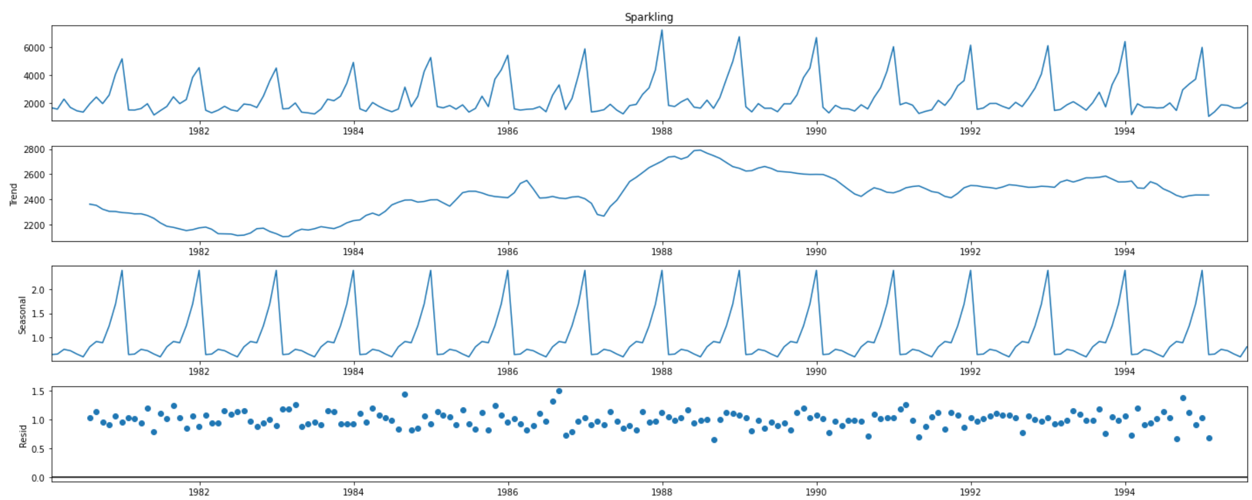

Figure-03

Remarks :

- We can see that the trend is upward.
- For the seasonality, not sure if there is multiplicative or additive seasonality we will see in other graphs.

Pivot Table :

Time_Stamp	1	2	3	4	5	6	7	8	9	10	11	12
Time_Stamp												
1980	1686.0	1591.0	2304.0	1712.0	1471.0	1377.0	1966.0	2453.0	1984.0	2596.0	4087.0	5179.0
1981	1530.0	1523.0	1633.0	1976.0	1170.0	1480.0	1781.0	2472.0	1981.0	2273.0	3857.0	4551.0
1982	1510.0	1329.0	1518.0	1790.0	1537.0	1449.0	1954.0	1897.0	1706.0	2514.0	3593.0	4524.0
1983	1609.0	1638.0	2030.0	1375.0	1320.0	1245.0	1600.0	2298.0	2191.0	2511.0	3440.0	4923.0
1984	1609.0	1435.0	2061.0	1789.0	1567.0	1404.0	1597.0	3159.0	1759.0	2504.0	4273.0	5274.0
1985	1771.0	1682.0	1846.0	1589.0	1896.0	1379.0	1645.0	2512.0	1771.0	3727.0	4388.0	5434.0
1986	1606.0	1523.0	1577.0	1605.0	1765.0	1403.0	2584.0	3318.0	1562.0	2349.0	3987.0	5891.0
1987	1389.0	1442.0	1548.0	1935.0	1518.0	1250.0	1847.0	1930.0	2638.0	3114.0	4405.0	7242.0
1988	1853.0	1779.0	2108.0	2336.0	1728.0	1661.0	2230.0	1645.0	2421.0	3740.0	4988.0	6757.0
1989	1757.0	1394.0	1982.0	1650.0	1654.0	1406.0	1971.0	1968.0	2608.0	3845.0	4514.0	6694.0
1990	1720.0	1321.0	1859.0	1628.0	1615.0	1457.0	1899.0	1605.0	2424.0	3116.0	4286.0	6047.0
1991	1902.0	2049.0	1874.0	1279.0	1432.0	1540.0	2214.0	1857.0	2408.0	3252.0	3627.0	6153.0
1992	1577.0	1667.0	1993.0	1997.0	1783.0	1625.0	2076.0	1773.0	2377.0	3088.0	4096.0	6119.0
1993	1494.0	1564.0	1898.0	2121.0	1831.0	1515.0	2048.0	2795.0	1749.0	3339.0	4227.0	6410.0
1994	1197.0	1968.0	1720.0	1725.0	1674.0	1693.0	2031.0	1495.0	2968.0	3385.0	3729.0	5999.0
1995	1070.0	1402.0	1897.0	1862.0	1670.0	1688.0	2031.0	NaN	NaN	NaN	NaN	NaN

Table- 4

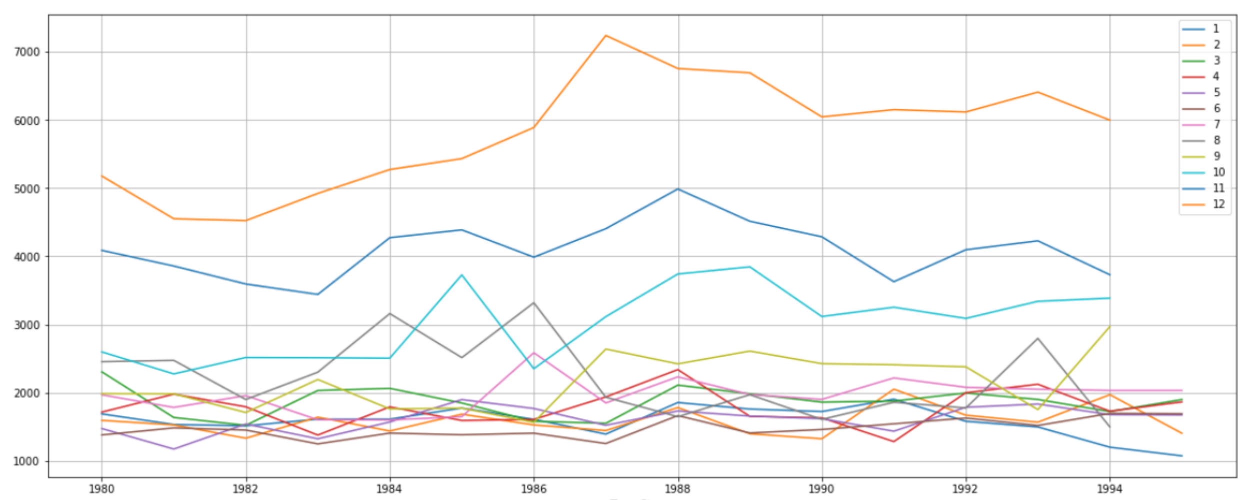

Figure-04

Note : By seeing the above graph we can see that some lines are crossing each other so we can say there is no additive seasonality.

Check the residual and normality:

For the multiplicative :

Residual = 0.9997456359115033

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

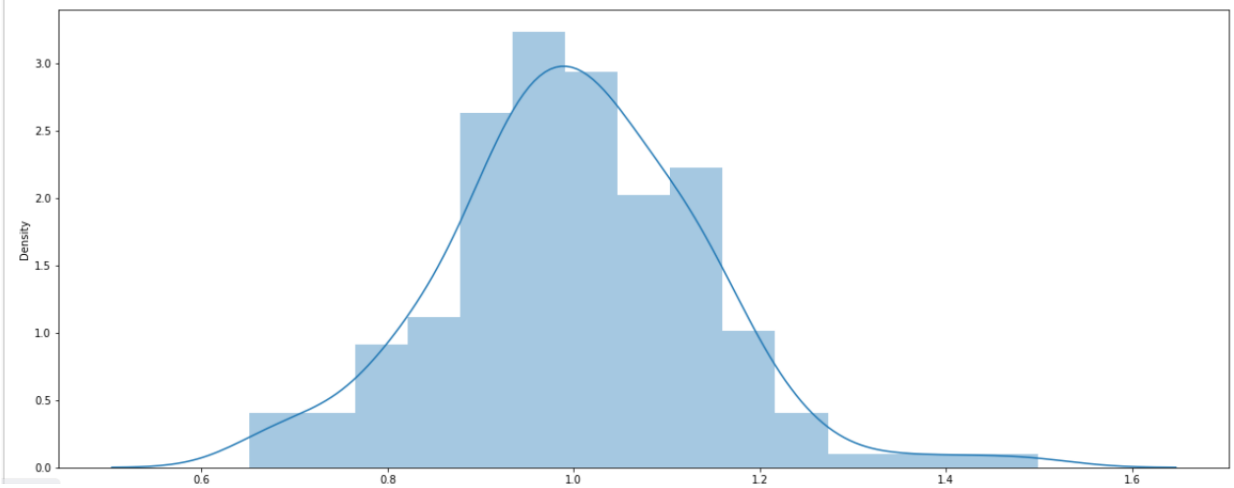

Figure-05

Remarks : for the multiplicative seasonality error mean = 1 and data should be normally distributed.

Note :

- P value is greater than 0.05 so null hypothesis is not rejected. Residual normally distributed.
- Residual mean =1 both conditions are valid so we can say that seasonality is multiplicative.

Boxplot for yearly : To check trends :

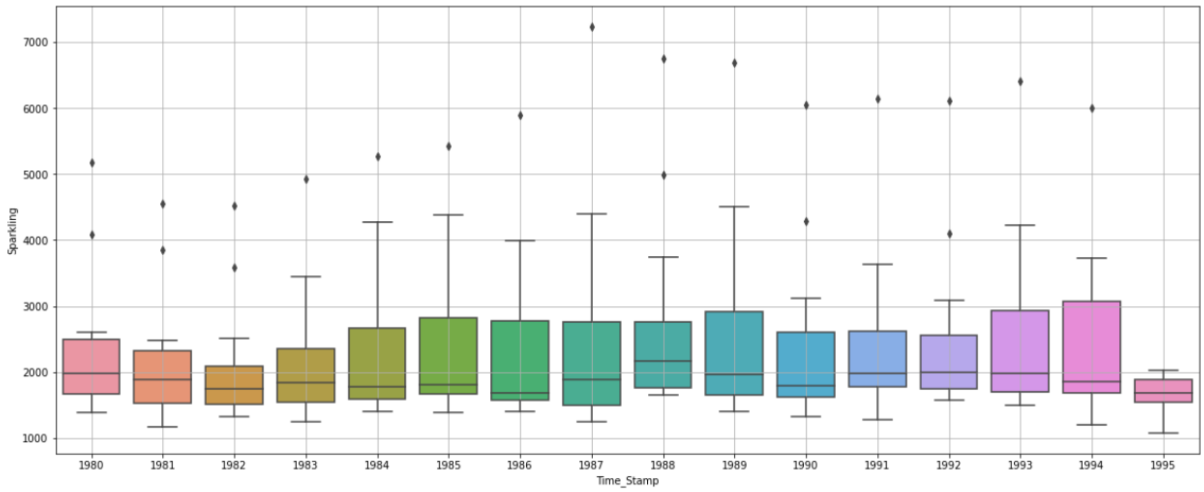

Figure-06

Note :

looks upwards trends.

Boxplot for Month : To check seasonality

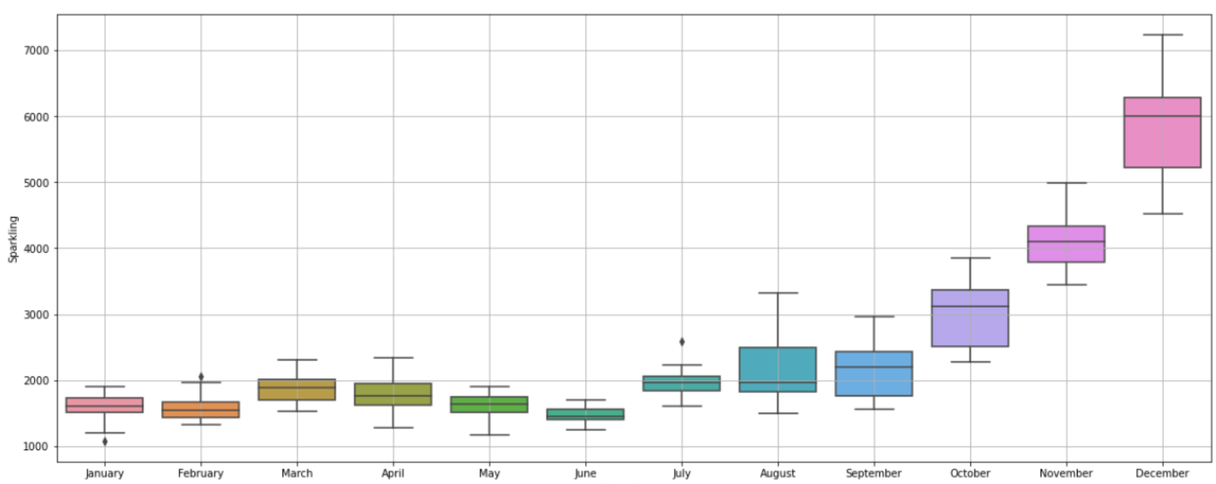

Figure- 07

Note: jan to jun looks constant sales, from july to december sales increasing and in the december sales is highest. So we can say that datasets have seasonality.

Month plot :

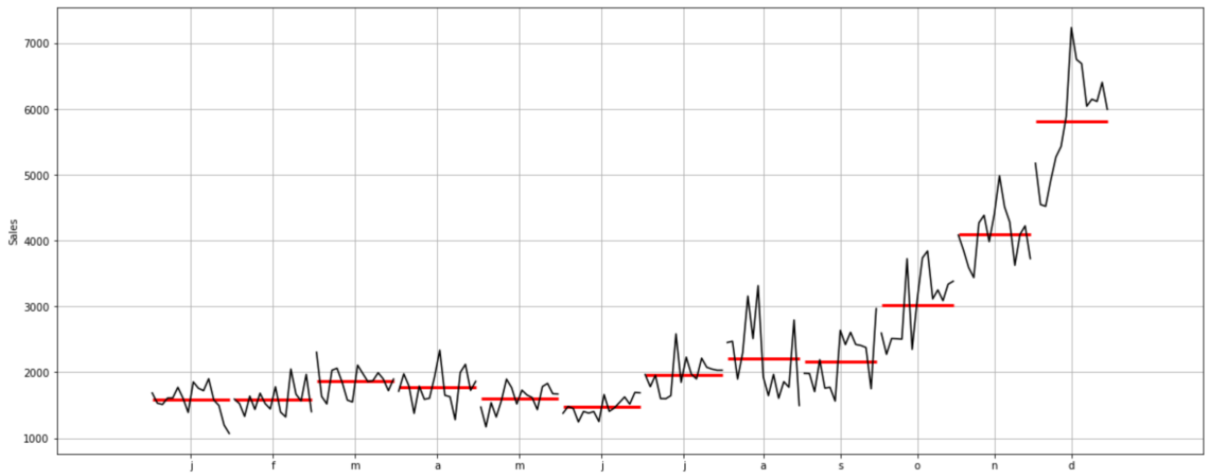

Figure-08

Note : Some month patterns look similar and some of the patterns look different.

So by this pattern we can not justify that data have seasonality.

1.3 Split the data into training and test. The test data should start in 1999.

```
First few rows of Training Data
```

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

```
Last few rows of Training Data
```

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

Table - 05

First few rows of Test Data

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

Last few rows of Test Data

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

Table-06

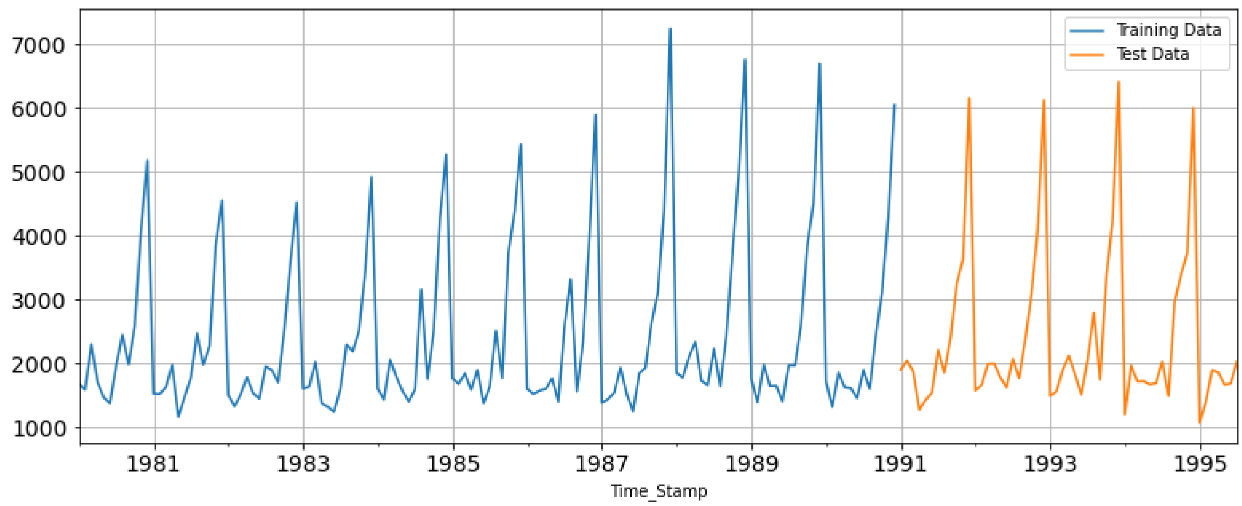

Figure-09

1.4 Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other additional models such as regression, naïve forecast models, simple average models, moving average models should also be built on the training data and check the performance on the test data using RMSE.

Linear Regression

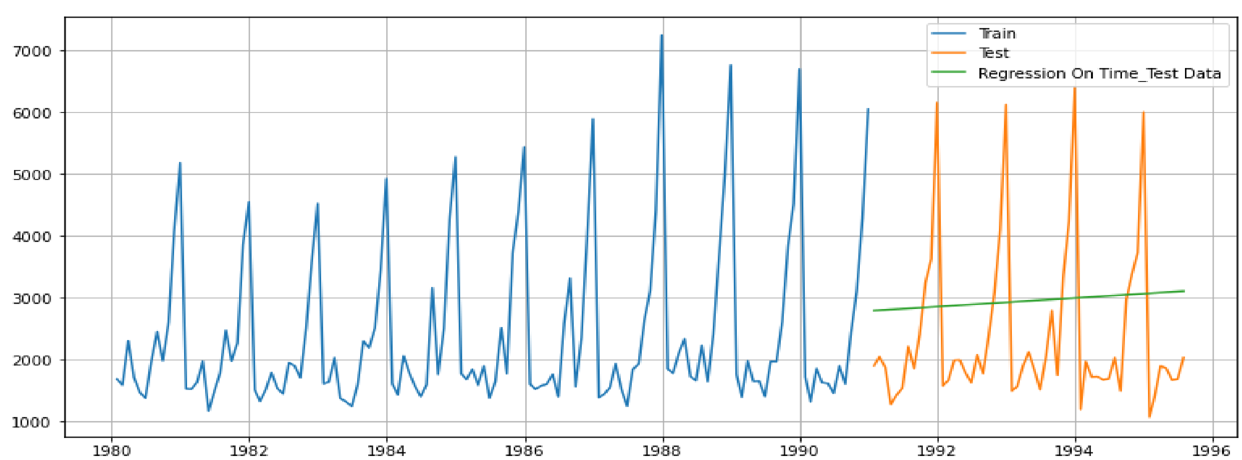

Figure-10

	Test RMSE
RegressionOnTime	1389.135175

naïve forecast models :

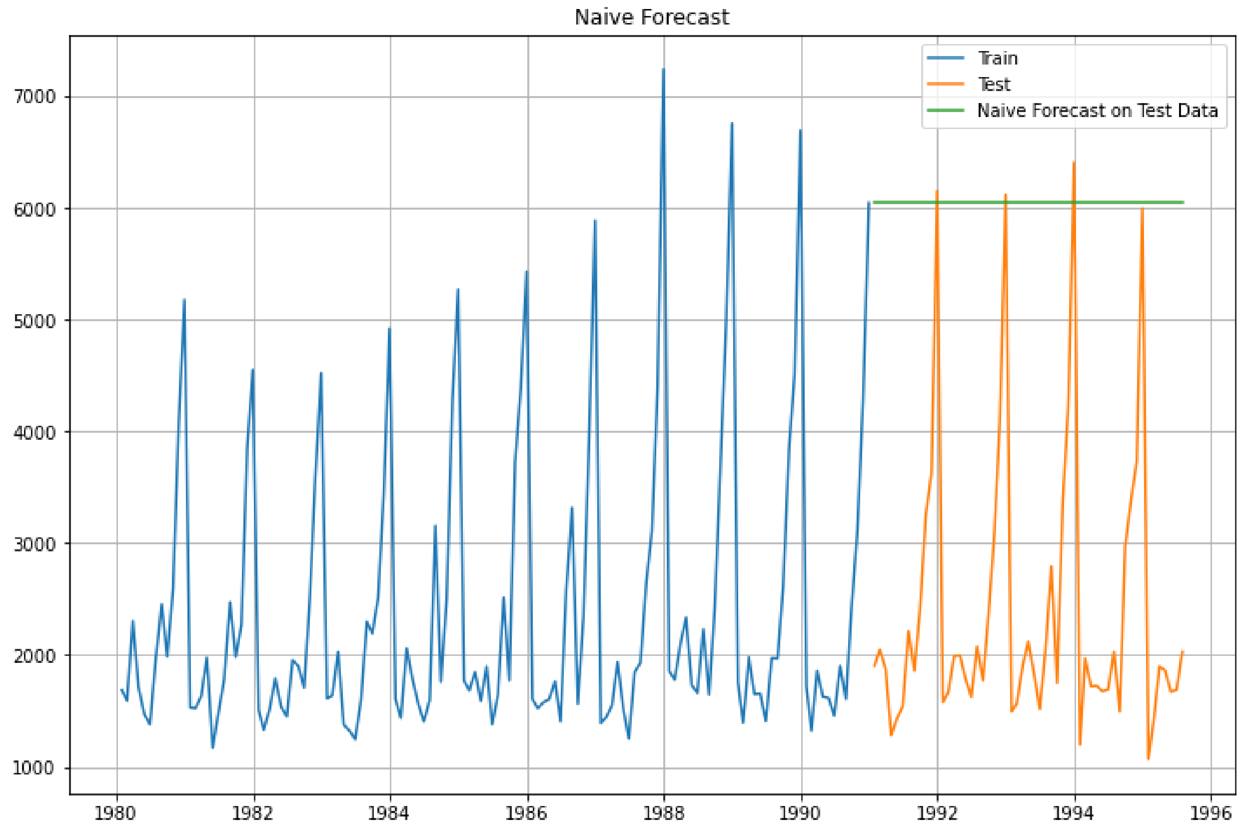

Figure-11

For Naive forecast on the Test Data, RMSE is 3864.279.

simple average models :

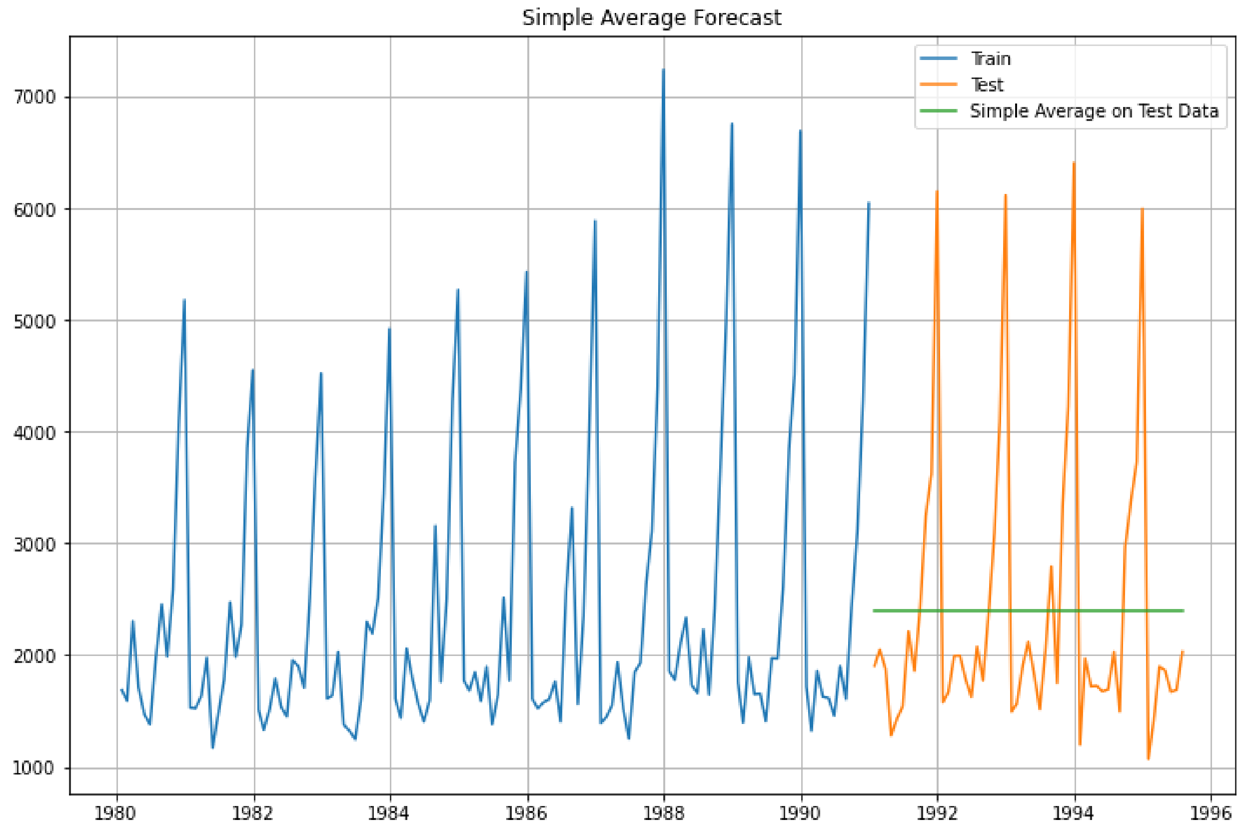

Figure-12

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

Moving Average :

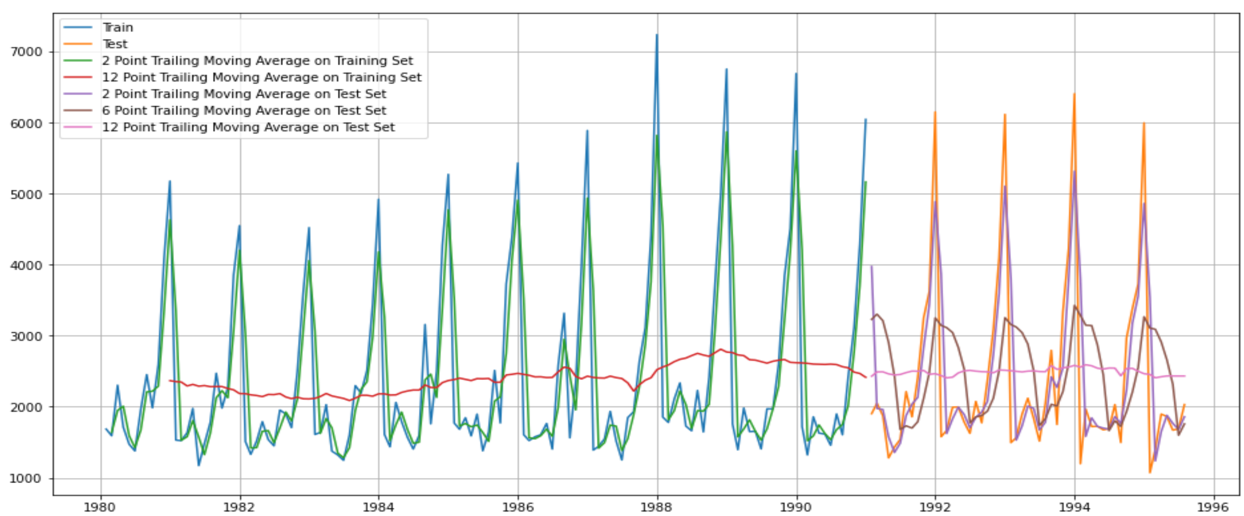

Figure-13

```
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
For 12 point Moving Average Model forecast on the Training Data,  RMSE is 1267.925
```

Simple exponential smoothing :

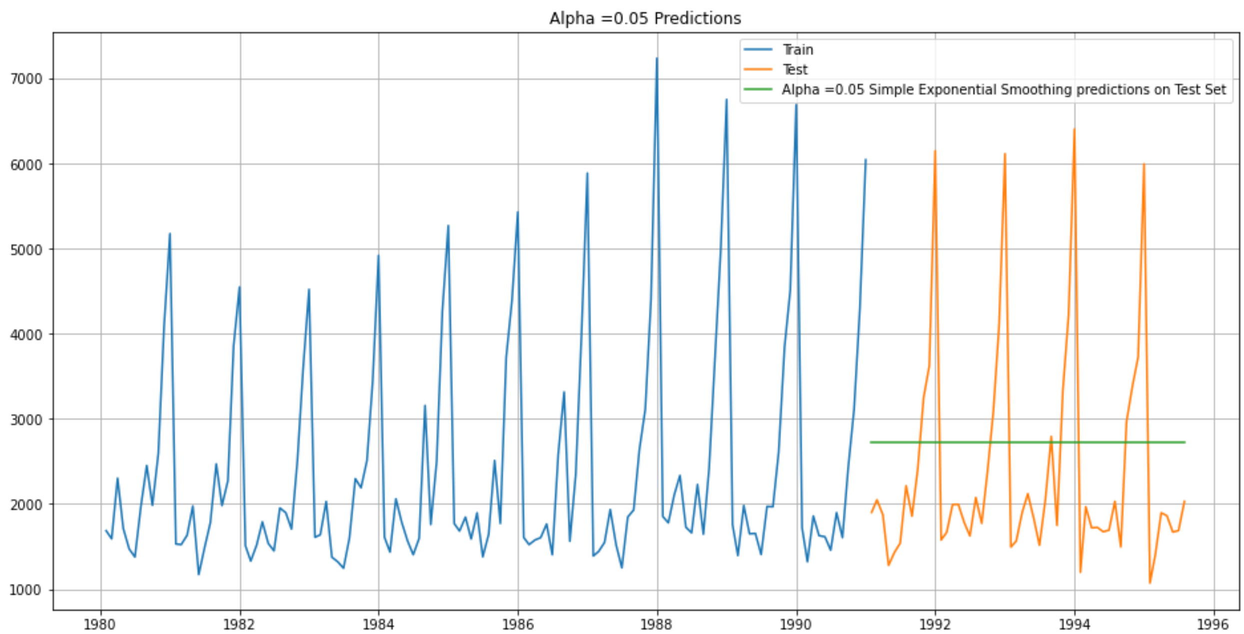

Figure-14

For Alpha =0.05 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 1316.035

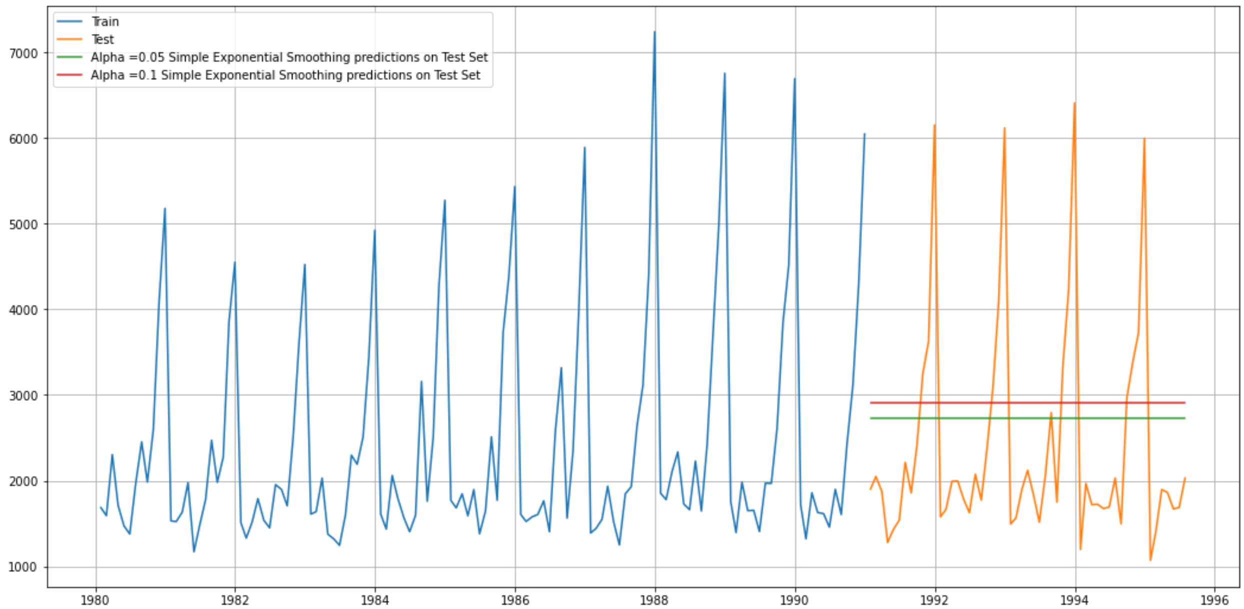

Figure-15

Alpha=0.1,SimpleExponentialSmoothing 1375.393398

Double exponential smoothing :

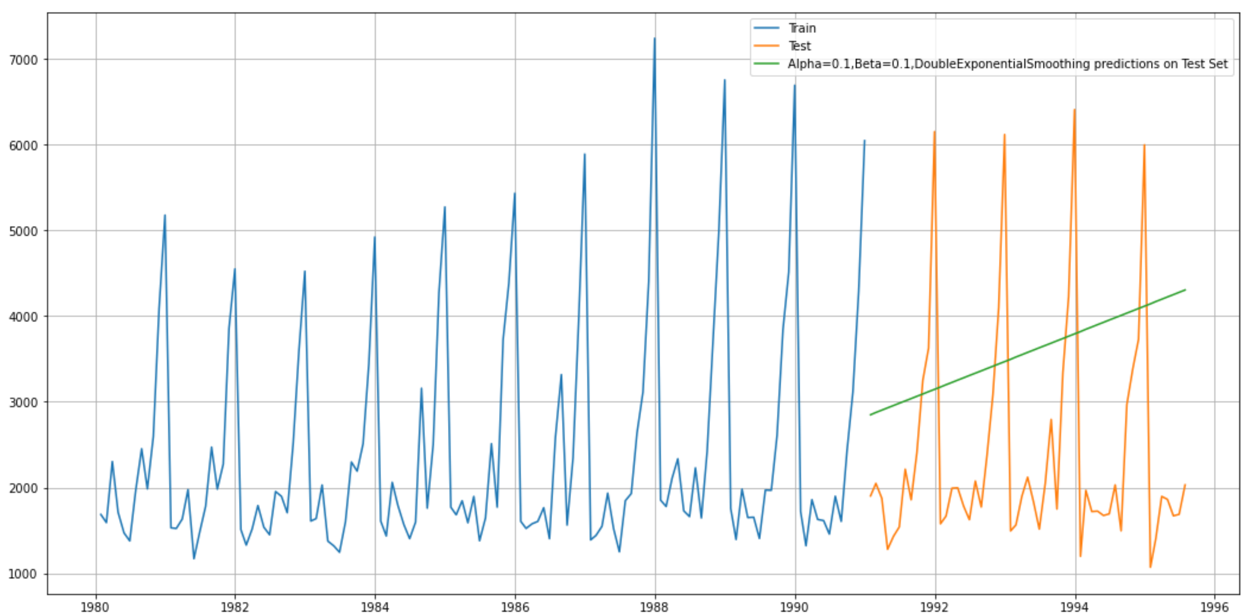

Figure - 16

Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing 1778.564670

Triple exponential smoothing :

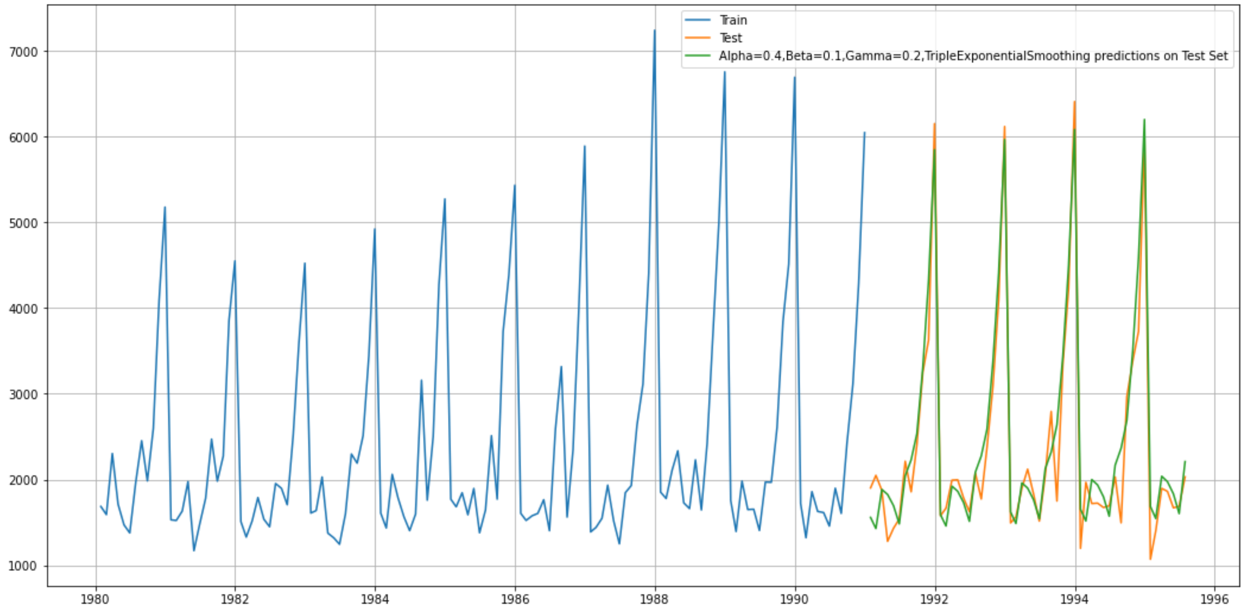

Figure - 17

Alpha=0.4,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing 336.715250

1.5 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.

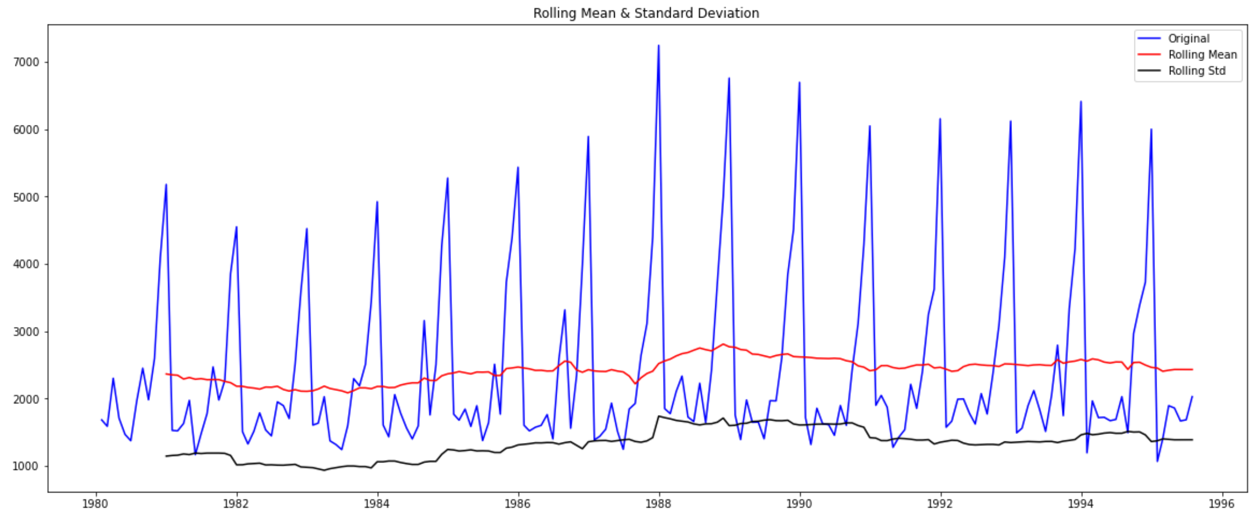

Figure - 18

```
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
```

Note :

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

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

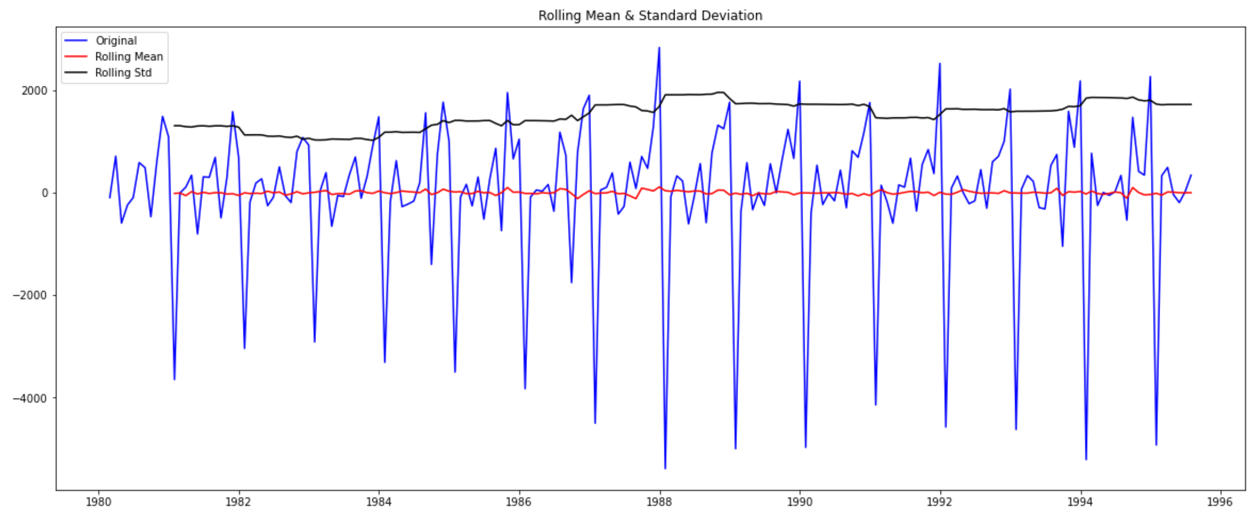

Figure - 19

```
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
```

Note :

After differencing We see that at α = 0.05 the Time Series is indeed stationary.

1.6 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.

ARIMA :

AIC Values : In Ascending order

	param	AIC
10	(2, 1, 2)	2210.618562
15	(3, 1, 3)	2225.661559
14	(3, 1, 2)	2228.928204
11	(2, 1, 3)	2229.358094
9	(2, 1, 1)	2232.360490
2	(0, 1, 2)	2232.783098
3	(0, 1, 3)	2233.016605
6	(1, 1, 2)	2233.597647

Table-07

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:            D.Sparkling   No. Observations:                  131
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1099.309
Method:                       css-mle   S.D. of innovations           1012.730
Date:                Sat, 12 Feb 2022   AIC                           2210.619
Time:                        23:16:11   BIC                           2227.870
Sample:                    02-29-1980   HQIC                          2217.628
                         - 12-31-1990
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 5.5843      0.518     10.790      0.000       4.570       6.599
ar.L1.D.Sparkling     1.2700      0.074     17.048      0.000       1.124       1.416
ar.L2.D.Sparkling    -0.5604      0.074     -7.620      0.000      -0.704      -0.416
ma.L1.D.Sparkling    -1.9978      0.042    -47.093      0.000      -2.081      -1.915
ma.L2.D.Sparkling     0.9978      0.042     23.501      0.000       0.915       1.081
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.1333           -0.7073j            1.3359           -0.0888
AR.2            1.1333           +0.7073j            1.3359            0.0888
MA.1            1.0004           +0.0000j            1.0004            0.0000
MA.2            1.0019           +0.0000j            1.0019            0.0000
-----------------------------------------------------------------------------
```

Note : All lags are significant. We can expect the result to be good.

RMSE of Test data : ARIMA(2,1,2)   1374.546024

SARIMA :

AIC Values : In Ascending order

	param	seasonal	AIC
287	(2, 1, 1)	(2, 0, 2, 6)	2004.405208
62	(0, 1, 2)	(2, 0, 2, 6)	2004.527202
187	(1, 1, 2)	(2, 0, 2, 6)	2006.914723
37	(0, 1, 1)	(2, 0, 2, 6)	2007.195159
87	(0, 1, 3)	(2, 0, 2, 6)	2007.742155

Table-08

```
                                      SARIMAX Results
===============================================================================================
Dep. Variable:                                    y   No. Observations:                  132
Model:             SARIMAX(2, 1, 1)x(2, 0, [1, 2], 6)   Log Likelihood                -864.020
Date:                              Sat, 12 Feb 2022   AIC                           1744.041
Time:                                      20:40:29   BIC                           1766.138
Sample:                                           0   HQIC                          1753.012
                                              - 132
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.0502      0.122      0.411      0.681      -0.189       0.290
ar.L2         -0.1086      0.121     -0.897      0.370      -0.346       0.129
ma.L1         -0.8593      0.079    -10.936      0.000      -1.013      -0.705
ar.S.L6        0.0019      0.025      0.079      0.937      -0.046       0.050
ar.S.L12       1.0421      0.016     66.124      0.000       1.011       1.073
ma.S.L6        0.0130      0.137      0.095      0.924      -0.255       0.281
ma.S.L12      -0.6389      0.089     -7.187      0.000      -0.813      -0.465
sigma2      1.468e+05   1.44e+04     10.184      0.000    1.19e+05    1.75e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                38.11
Prob(Q):                              0.90   Prob(JB):                         0.00
Heteroskedasticity (H):               2.79   Skew:                             0.51
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.61
===================================================================================
```

RMSE of test data :

SARIMA(2,1,1)(2,0,2,6) 636.214759642355

	param	seasonal	AIC
50	(1, 1, 2)	(1, 0, 2, 12)	1555.584248
53	(1, 1, 2)	(2, 0, 2, 12)	1556.076790
26	(0, 1, 2)	(2, 0, 2, 12)	1557.121579
23	(0, 1, 2)	(1, 0, 2, 12)	1557.160507
77	(2, 1, 2)	(1, 0, 2, 12)	1557.340402

Table-09

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(1, 1, 2)x(1, 0, 2, 12)   Log Likelihood                -770.792
Date:                            Tue, 15 Feb 2022   AIC                           1555.584
Time:                                    13:53:23   BIC                           1574.095
Sample:                                         0   HQIC                          1563.083
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.6283      0.255     -2.464      0.014      -1.128      -0.128
ma.L1         -0.1040      0.225     -0.463      0.644      -0.545       0.337
ma.L2         -0.7277      0.154     -4.736      0.000      -1.029      -0.427
ar.S.L12       1.0439      0.014     72.834      0.000       1.016       1.072
ma.S.L12      -0.5550      0.098     -5.663      0.000      -0.747      -0.363
ma.S.L24      -0.1354      0.120     -1.133      0.257      -0.370       0.099
sigma2      1.506e+05   2.03e+04      7.401      0.000    1.11e+05     1.9e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.04   Jarque-Bera (JB):                11.72
Prob(Q):                              0.84   Prob(JB):                         0.00
Heteroskedasticity (H):               1.47   Skew:                             0.36
Prob(H) (two-sided):                  0.26   Kurtosis:                         4.48
===================================================================================
```

RMSE of test data :

SARIMA(1,1,2)(1,0,2,12) 528.611364

1.7 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.

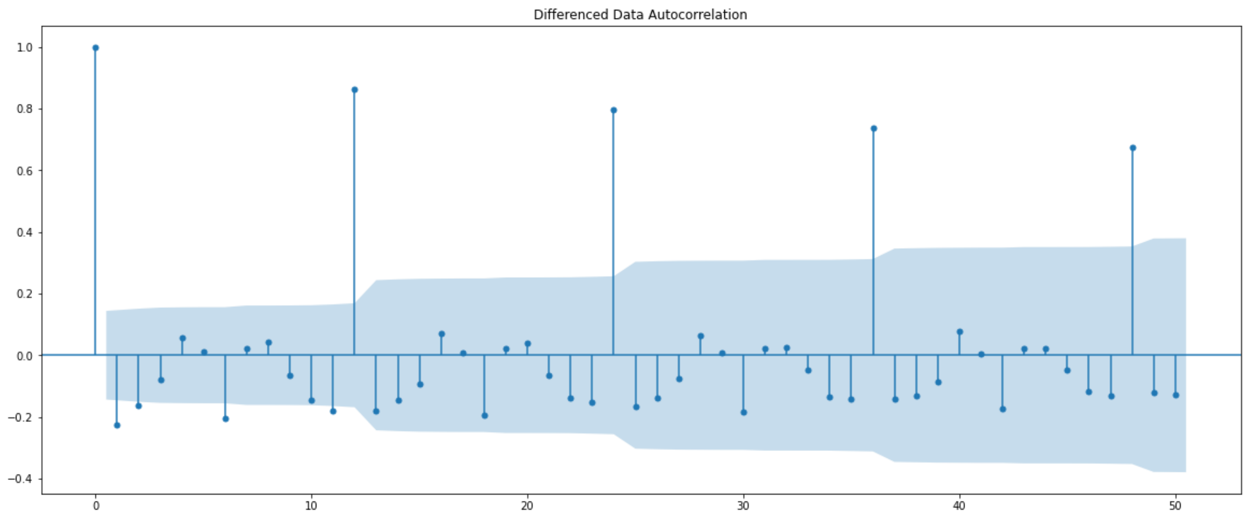

Figure - 20

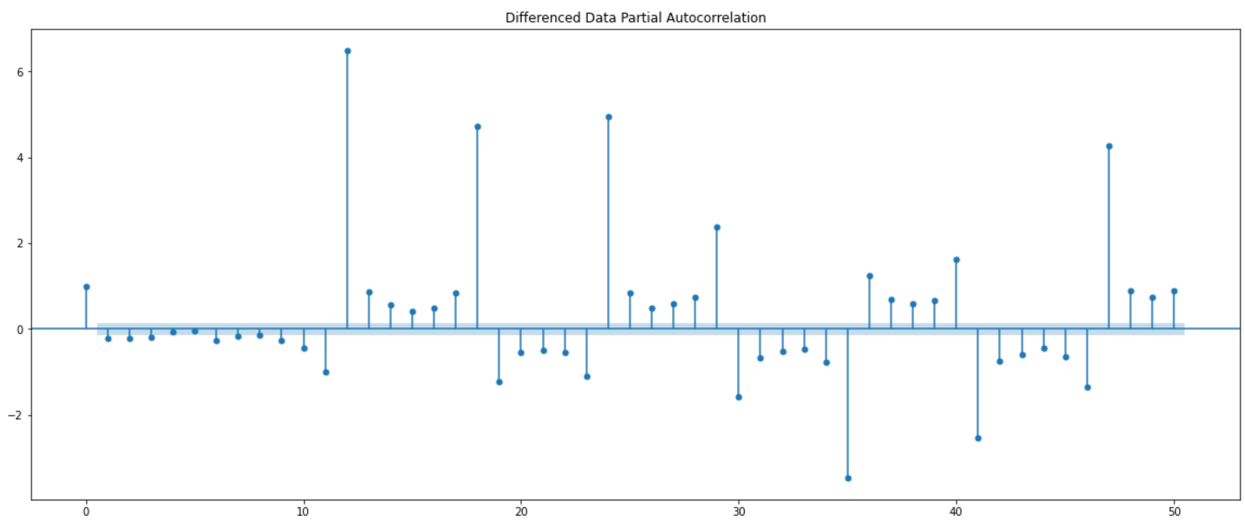

Figure-21

```
                              ARIMA Model Results
==============================================================================
Dep. Variable:            D.Sparkling   No. Observations:                  131
Model:                 ARIMA(0, 1, 0)   Log Likelihood               -1132.791
Method:                           css   S.D. of innovations           1377.911
Date:                Tue, 15 Feb 2022   AIC                           2269.583
Time:                        17:45:36   BIC                           2275.333
Sample:                    02-29-1980   HQIC                          2271.919
                         - 12-31-1990
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         33.2901    120.389      0.277      0.782    -202.667     269.248
==============================================================================
```

RMSE of test data :

ARIMA(0,1,0) 4779.154299

```
                                     SARIMAX Results
=========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(0, 1, 0)x(0, 1, 0, 6)   Log Likelihood               -1130.492
Date:                           Tue, 15 Feb 2022   AIC                           2262.984
Time:                                   17:45:36   BIC                           2265.804
Sample:                                        0   HQIC                          2264.129
                                           - 132
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
sigma2      4.842e+06    5.1e+05      9.495      0.000    3.84e+06    5.84e+06
===================================================================================
Ljung-Box (L1) (Q):                   1.89   Jarque-Bera (JB):                 4.17
Prob(Q):                              0.17   Prob(JB):                         0.12
Heteroskedasticity (H):               1.96   Skew:                            -0.05
Prob(H) (two-sided):                  0.03   Kurtosis:                         3.89
===================================================================================
```

RMSE of test data :

SARIMA(0,1,0)(0,1,0,6) 27078.593877

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

	Test RMSE
Alpha=0.4,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing	336.715250
Alpha=0.0.111,Beta=0.0617,Gamma=0.395,TripleExponentialSmoothing	469.767970
SARIMA(1,1,2)(1,0,2,12)	528.611364
SARIMA(2,1,1)(2,0,2,6)	636.214760
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
12pointTrailingMovingAverage	1267.925330
SimpleAverageModel	1275.081804
6pointTrailingMovingAverage	1283.927428
Alpha=0.05,SimpleExponentialSmoothing	1316.035487
9pointTrailingMovingAverage	1346.278315
ARIMA(2,1,2)	1374.546024
Alpha=0.1,SimpleExponentialSmoothing	1375.393398
RegressionOnTime	1389.135175
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	1778.564670
SARIMA(0,1,1)(3,0,1,6)	1999.383783
SARIMA(0,1,1)(3,0,1,6)	1999.383783
ARIMA(7,1,0)	2308.994154
NaiveModel	3864.279352

Table - 10

Note : We can say that the triple exponential(alpha=0.4, Beta=0.1, gamma=0.395) gives least RMSE.

1.9 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.

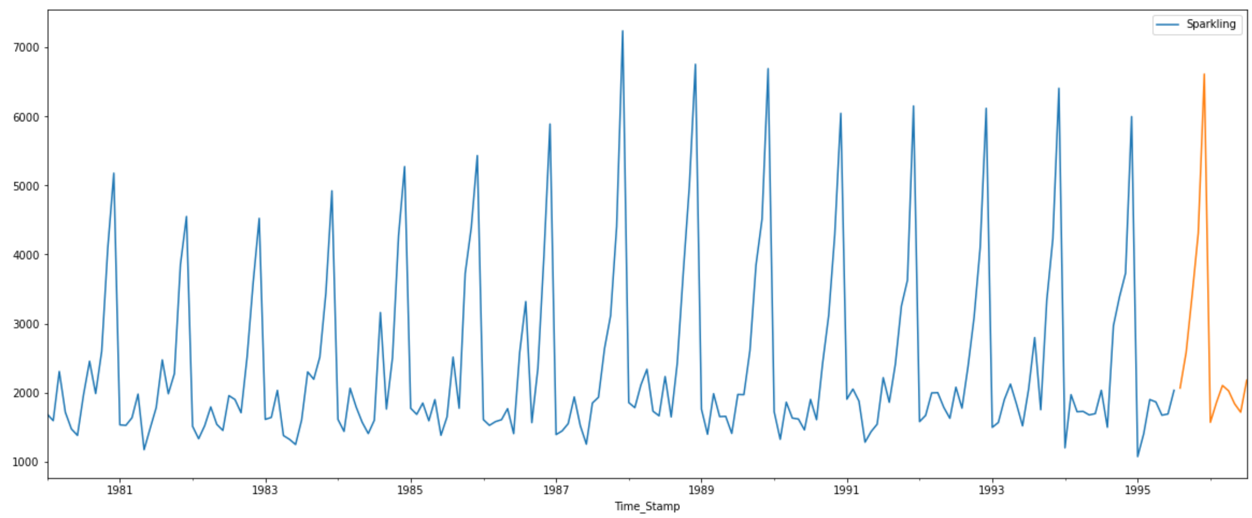

Figure- 22

RMSE of the Full Model 377.29032542281715

	lower_CI	prediction	upper_ci
1995-08-31	1321.896024	2063.370030	2804.844037
1995-09-30	1838.303763	2579.777769	3321.251776
1995-10-31	2676.612337	3418.086343	4159.560350
1995-11-30	3567.115379	4308.589385	5050.063392
1995-12-31	5874.310141	6615.784148	7357.258154

Table-10

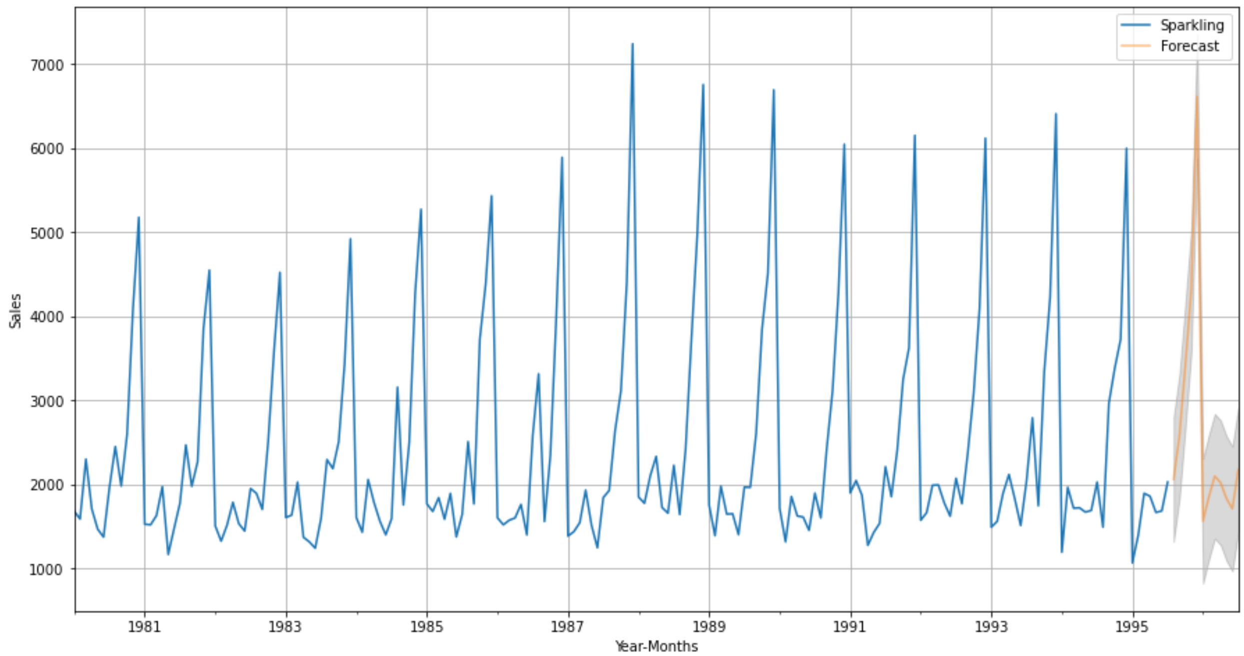

Figure - 23

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

- As Seen in the yearly plot, the upward trend is very slight. I suggest that we try to increase the sales.
- need to run promotional marketing campaigns or evaluate if we need to tie up with an alternate agency. It will increase sales.
- From Jan to June sales are low so we should try to increase the sales.
- From July to dec sales are increasing so we can try to increase the sales more.