Time Series Forecasting-Rose

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	Rose
0	1980-01	112.0
1	1980-02	118.0
2	1980-03	129.0
3	1980-04	99.0
4	1980-05	116.0

Table-01

Tail Datasets :

	YearMonth	Rose
182	1995-03	45.0
183	1995-04	52.0
184	1995-05	28.0
185	1995-06	40.0
186	1995-07	62.0

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   Rose       185 non-null    float64
dtypes: float64(1), object(1)
memory usage: 3.0+ KB
```

- From the above results we can see that there is some 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	Rose
1980-01-31	112
1980-02-29	118
1980-03-31	129
1980-04-30	99
1980-05-31	116

Table-03

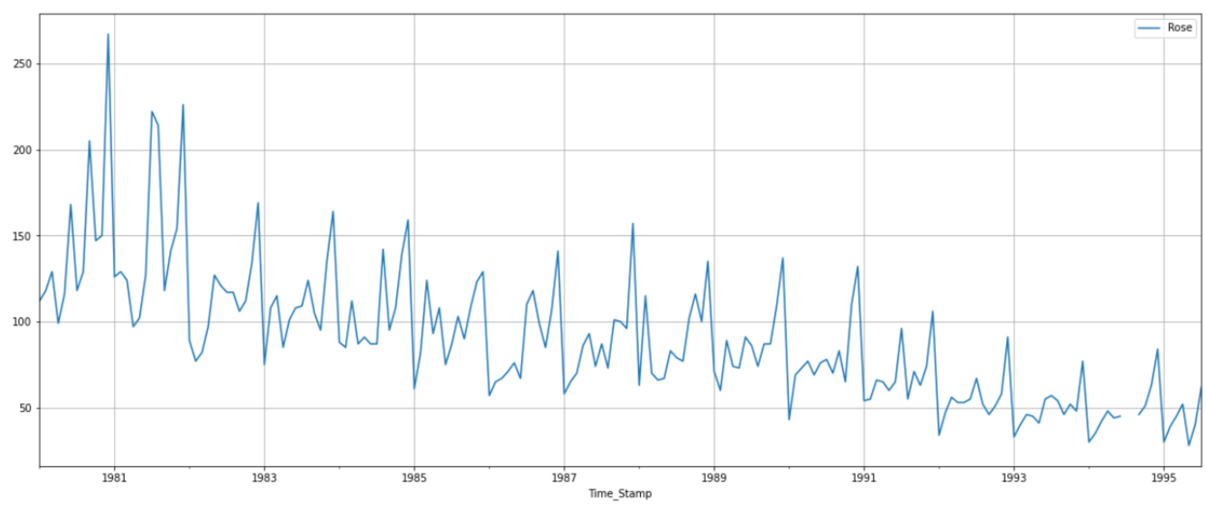

Figure - 01

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

Decompose the datasets :

Model = additive :

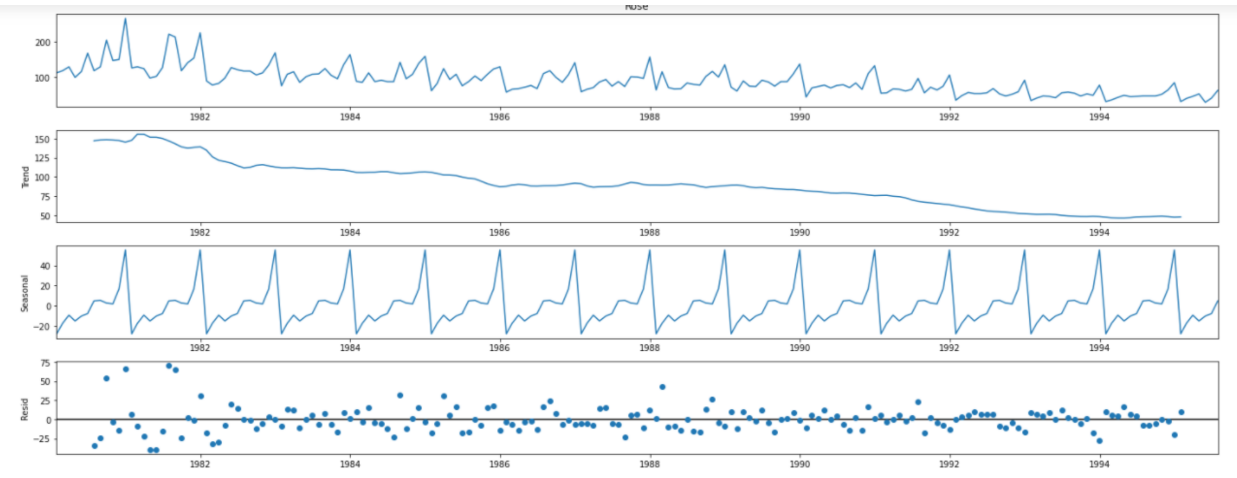

Figure-02

Model= multiplicative :

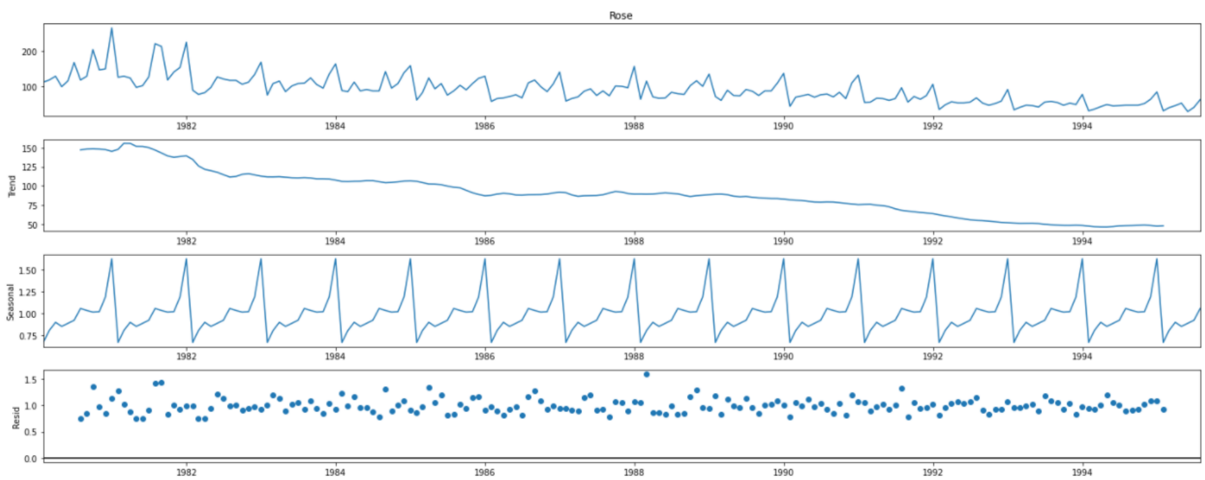

Figure-03

Remarks :

- We can see that the trend is downward.
- 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	112.0	118.0	129.0	99.0	116.0	168.0	118.0	129.0	205.0	147.0	150.0	267.0
1981	126.0	129.0	124.0	97.0	102.0	127.0	222.0	214.0	118.0	141.0	154.0	226.0
1982	89.0	77.0	82.0	97.0	127.0	121.0	117.0	117.0	106.0	112.0	134.0	169.0
1983	75.0	108.0	115.0	85.0	101.0	108.0	109.0	124.0	105.0	95.0	135.0	164.0
1984	88.0	85.0	112.0	87.0	91.0	87.0	87.0	142.0	95.0	108.0	139.0	159.0
1985	61.0	82.0	124.0	93.0	108.0	75.0	87.0	103.0	90.0	108.0	123.0	129.0
1986	57.0	65.0	67.0	71.0	76.0	67.0	110.0	118.0	99.0	85.0	107.0	141.0
1987	58.0	65.0	70.0	86.0	93.0	74.0	87.0	73.0	101.0	100.0	96.0	157.0
1988	63.0	115.0	70.0	66.0	67.0	83.0	79.0	77.0	102.0	116.0	100.0	135.0
1989	71.0	60.0	89.0	74.0	73.0	91.0	86.0	74.0	87.0	87.0	109.0	137.0
1990	43.0	69.0	73.0	77.0	69.0	76.0	78.0	70.0	83.0	65.0	110.0	132.0
1991	54.0	55.0	66.0	65.0	60.0	65.0	96.0	55.0	71.0	63.0	74.0	106.0
1992	34.0	47.0	56.0	53.0	53.0	55.0	67.0	52.0	46.0	51.0	58.0	91.0
1993	33.0	40.0	46.0	45.0	41.0	55.0	57.0	54.0	46.0	52.0	48.0	77.0
1994	30.0	35.0	42.0	48.0	44.0	45.0	46.0	46.0	46.0	51.0	63.0	84.0
1995	30.0	39.0	45.0	52.0	28.0	40.0	62.0	NaN	NaN	NaN	NaN	NaN

Table- 4

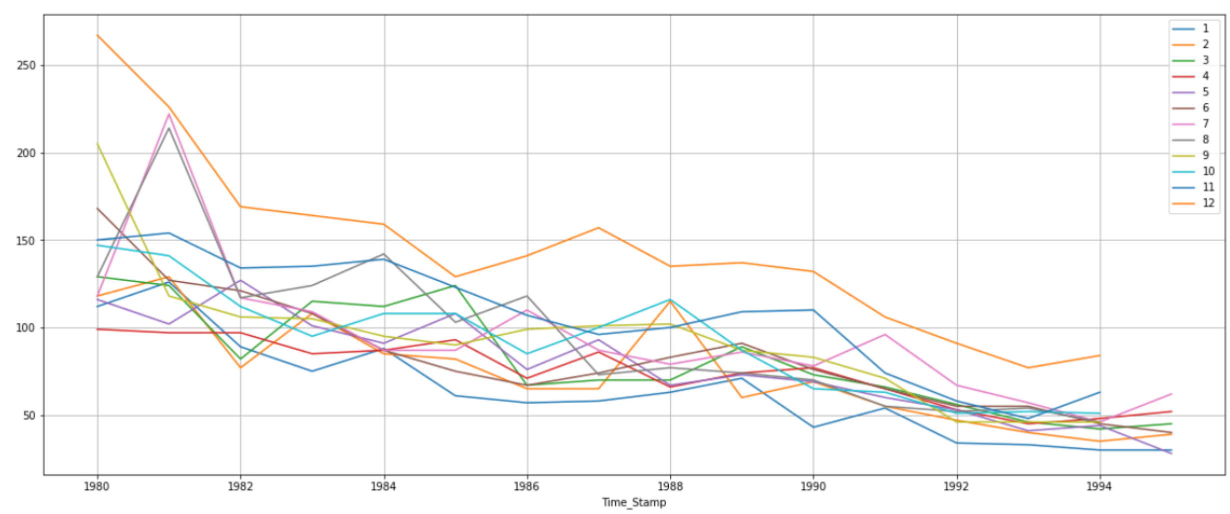

Figure-04

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.

Note : There are some missing values so we imputed using backfill.

Check the residual and normality:

For the multiplicative :

Residual = 0.9994553586957378

```
ShapiroResult(statistic=0.9488479495048523, pvalue=5.949785190750845e-06)
```

```
C:\Users\Pradeep Mishra\anaconda3\lib\site-packages\seaborn\distributions.py:2557: FutureWarning: `distplot` is a dep
recated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-le
vel function with similar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

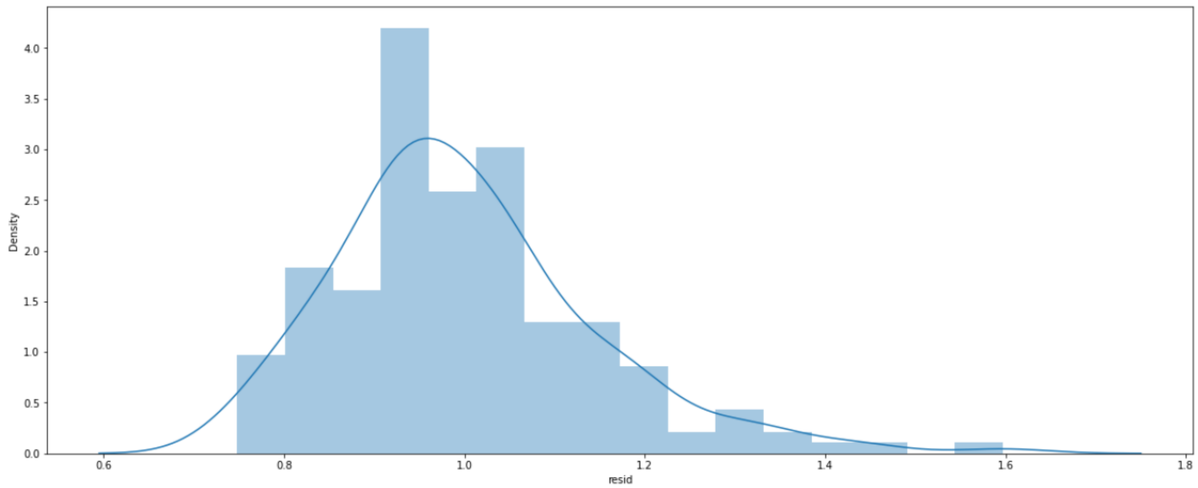

Figure-05

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

Note :

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

Bxplot for yearly : To check trends :

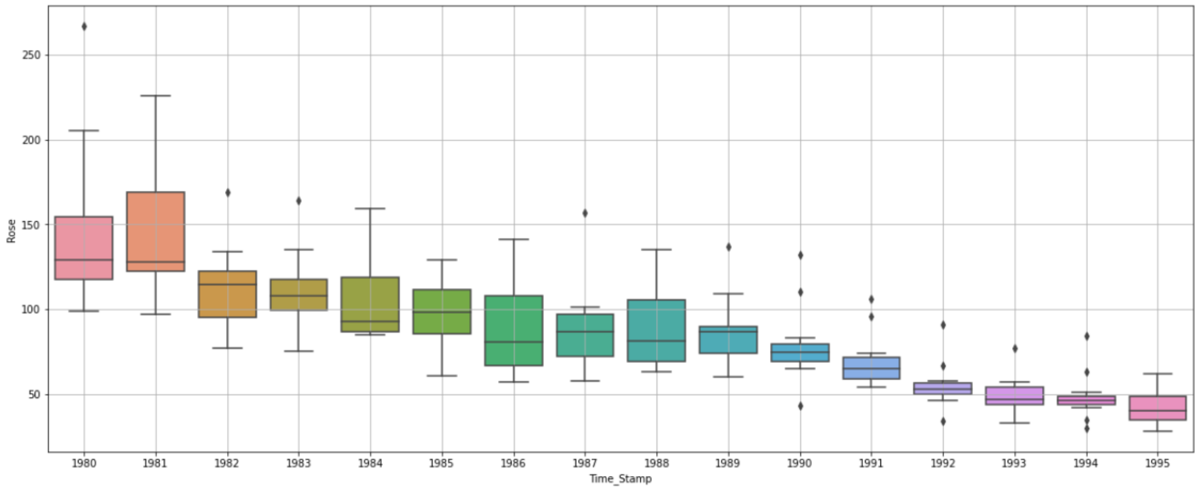

Figure-06

Note :

looks downward trends.

Boxplot for Month : To check seasonality

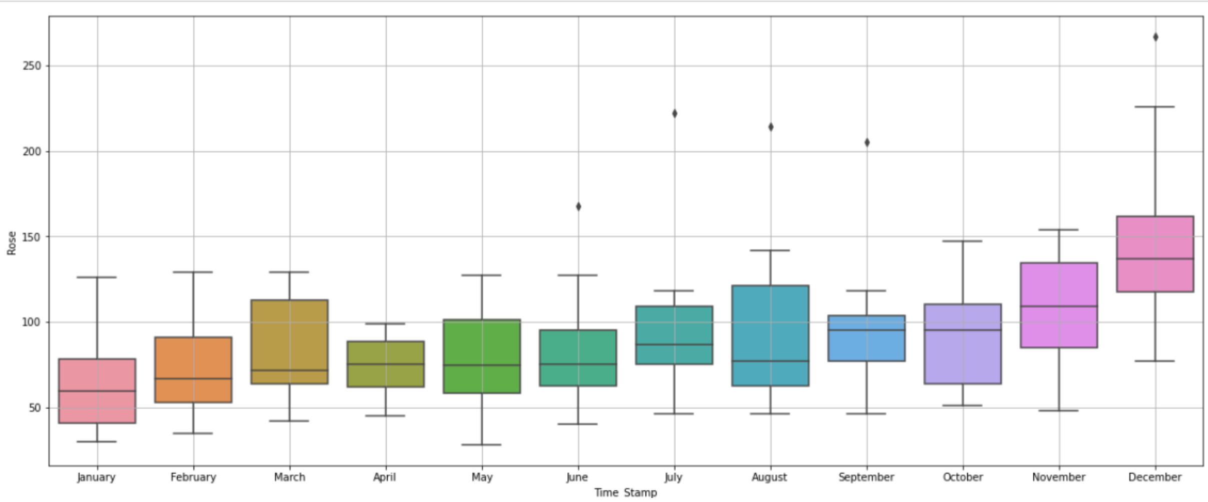

Figure- 07

Note: datasets have no seasonality.

Month plot :

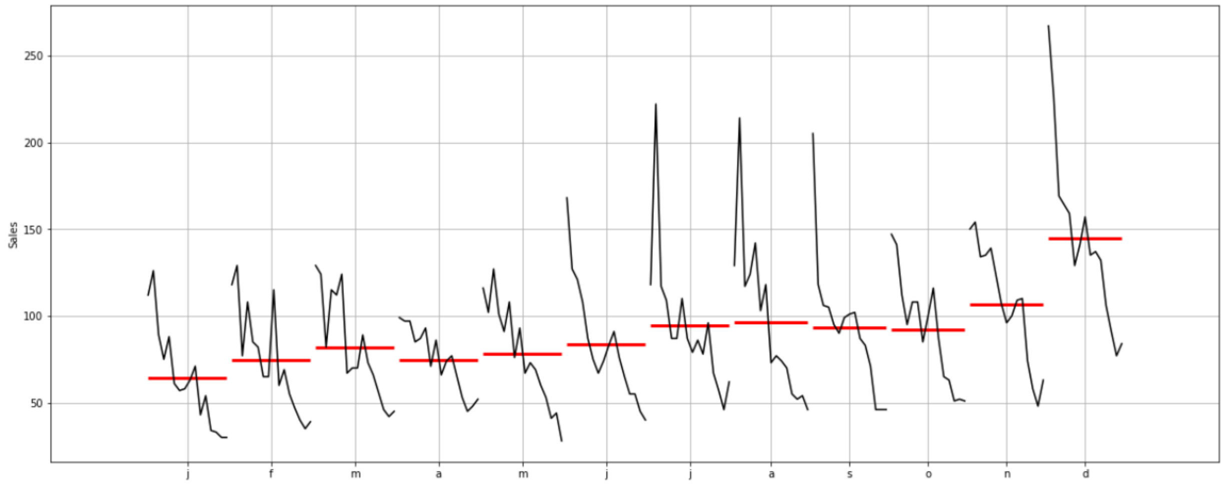

Figure-08

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	Rose
1980-01-31	112
1980-02-29	118
1980-03-31	129
1980-04-30	99
1980-05-31	116

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

Time_Stamp	Rose
1990-08-31	70
1990-09-30	83
1990-10-31	65
1990-11-30	110
1990-12-31	132

```
First few rows of Test Data
```

Table - 05

First few rows of Test Data

Time_Stamp	Rose
1991-01-31	54
1991-02-28	55
1991-03-31	66
1991-04-30	65
1991-05-31	60

Last few rows of Test Data

Time_Stamp	Rose
1995-03-31	45
1995-04-30	52
1995-05-31	28
1995-06-30	40
1995-07-31	62

Table-06

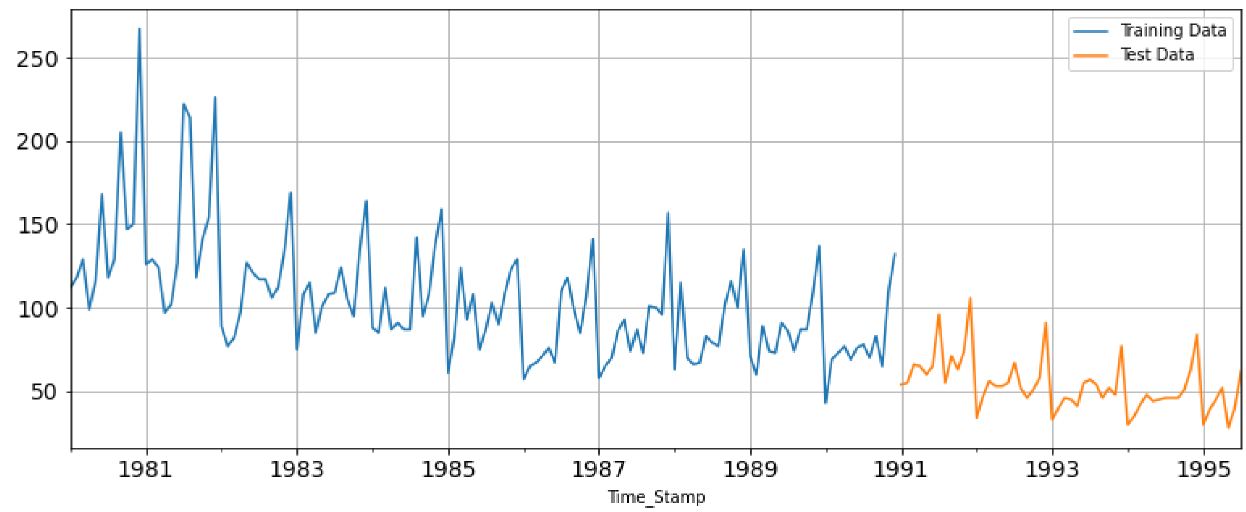

Figure-09

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

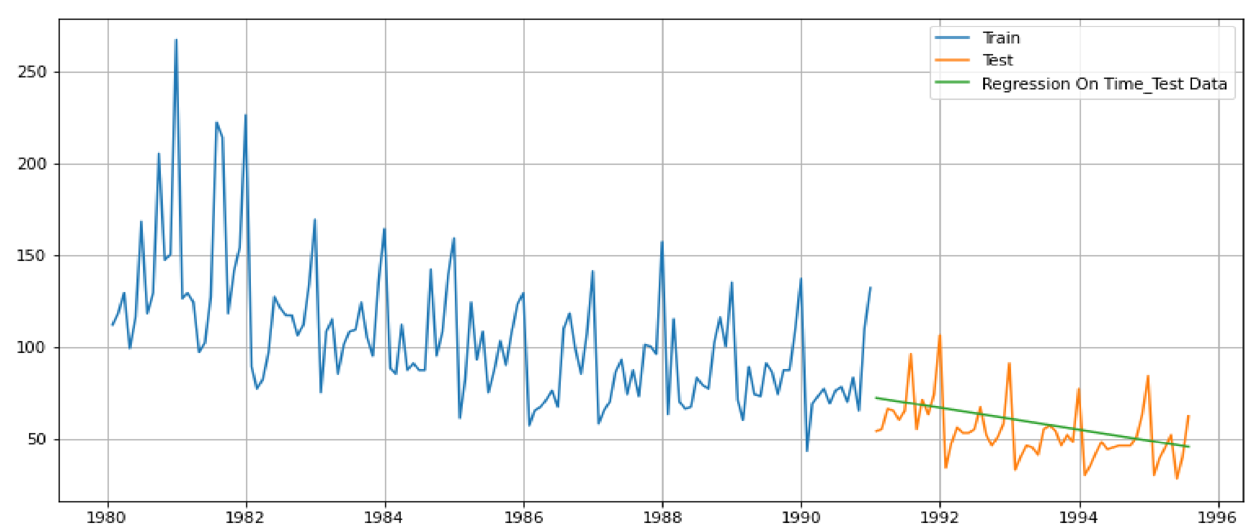

Figure-10

	Test RMSE
RegressionOnTime	15.262509

naïve forecast models :

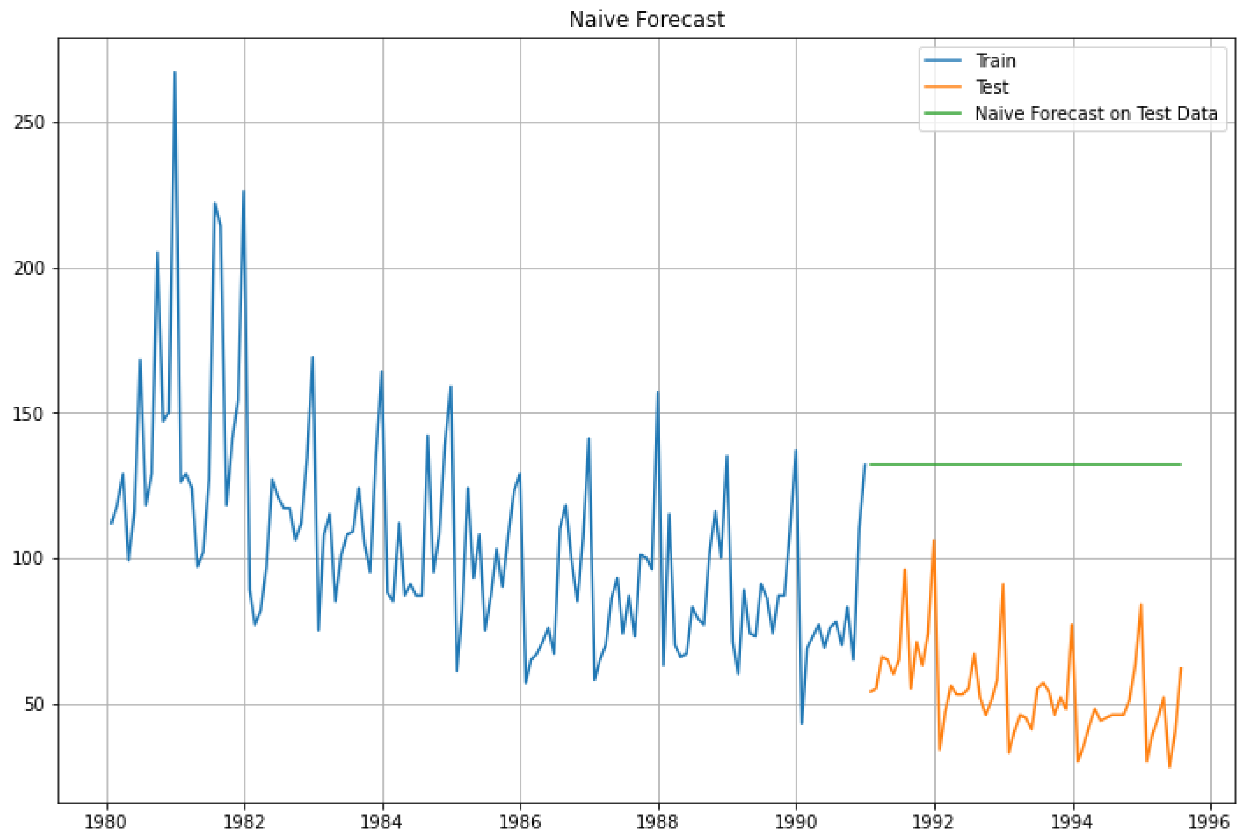

Figure-11

For Naive forecast on the Test Data, RMSE is 79.699

simple average models :

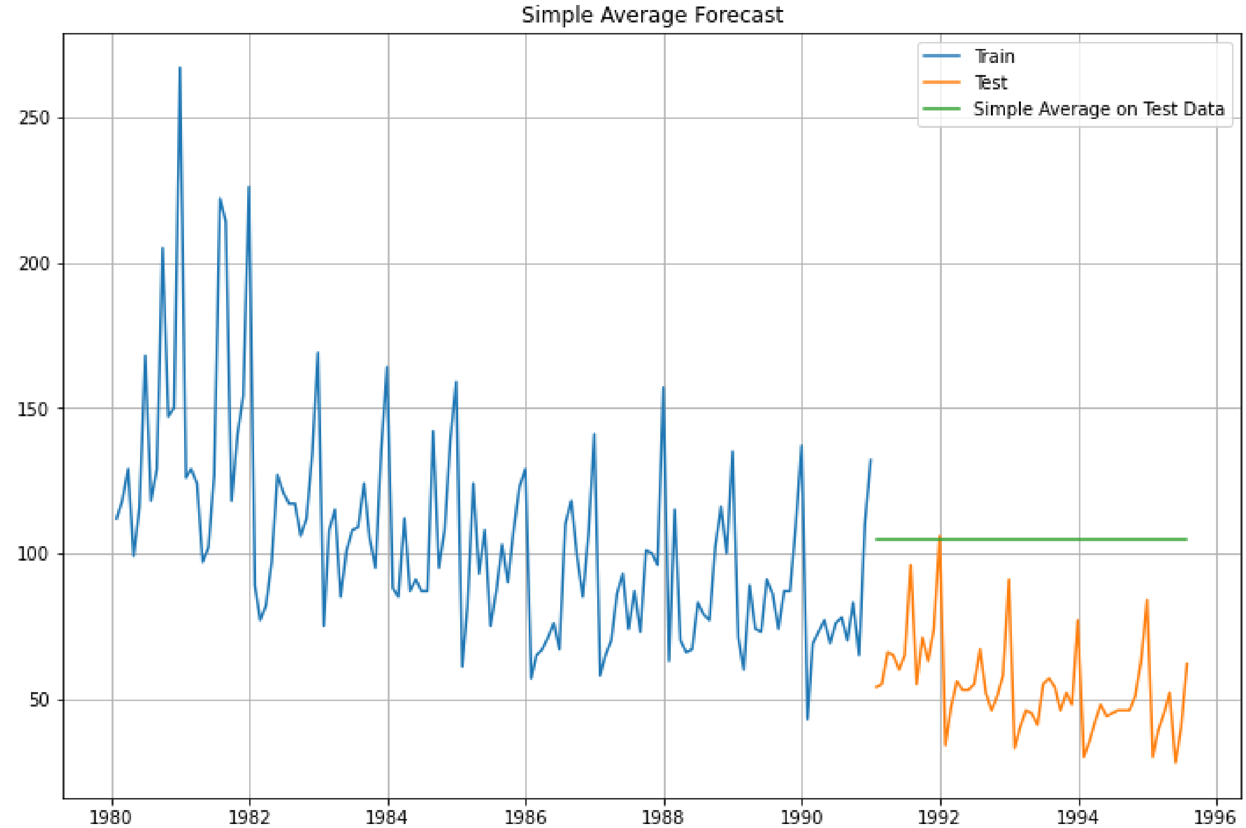

Figure-12

For Simple Average forecast on the Test Data, RMSE is 53.440

Moving Average :

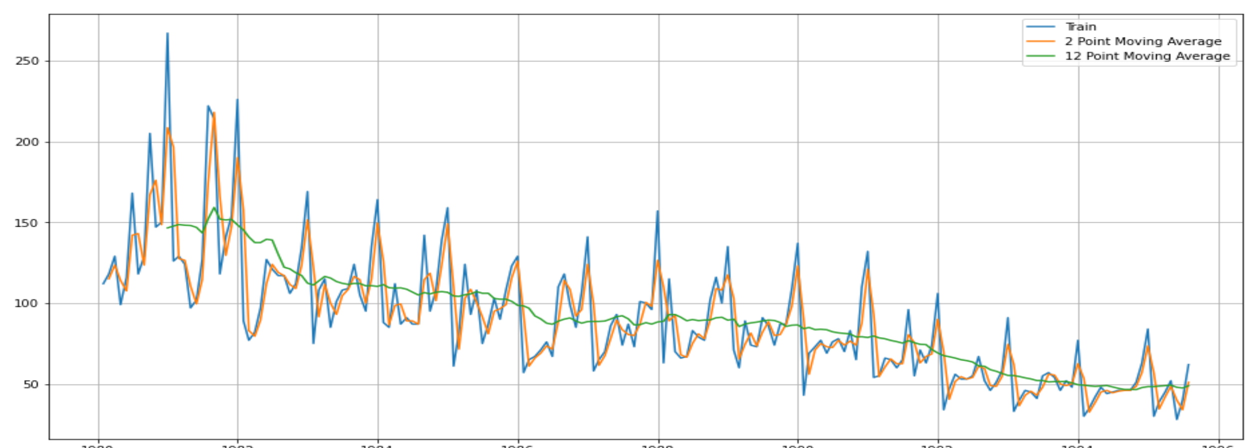

Figure-13

```
For 2 point Moving Average Model forecast on the Training Data,  RMSE is 11.529
For 4 point Moving Average Model forecast on the Training Data,  RMSE is 14.449
For 6 point Moving Average Model forecast on the Training Data,  RMSE is 14.560
For 9 point Moving Average Model forecast on the Training Data,  RMSE is 14.725
For 12 point Moving Average Model forecast on the Training Data,  RMSE is 15.234
```

Simple exponential smoothing :

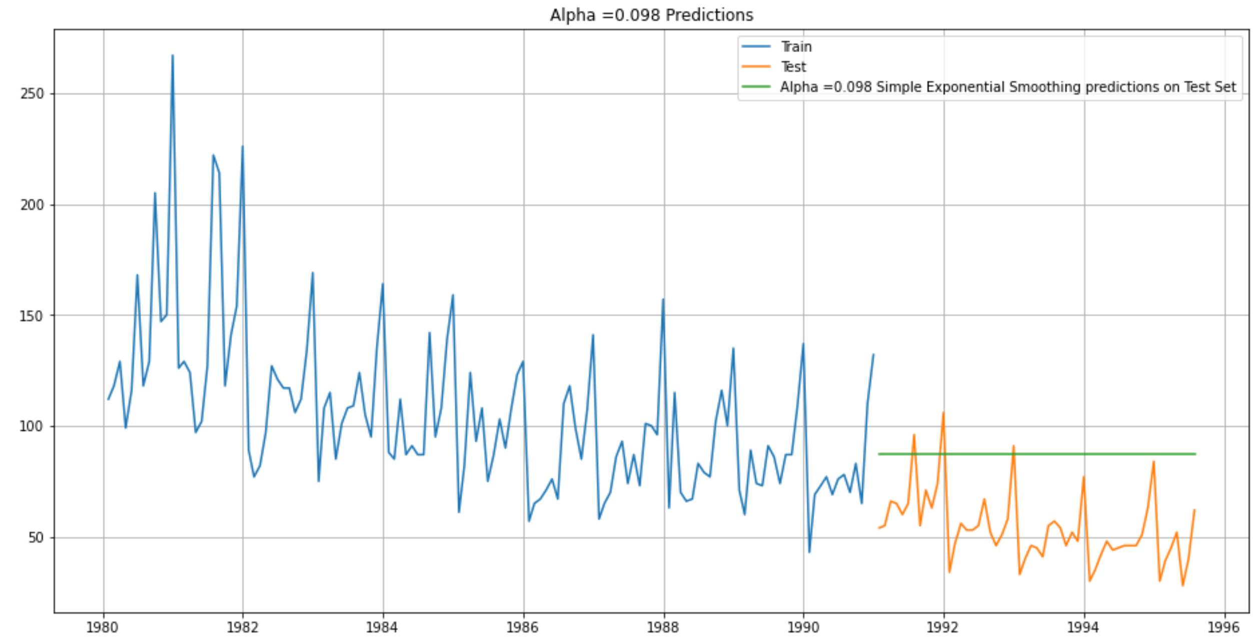

Figure-14

For Alpha =0.098 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 36.776

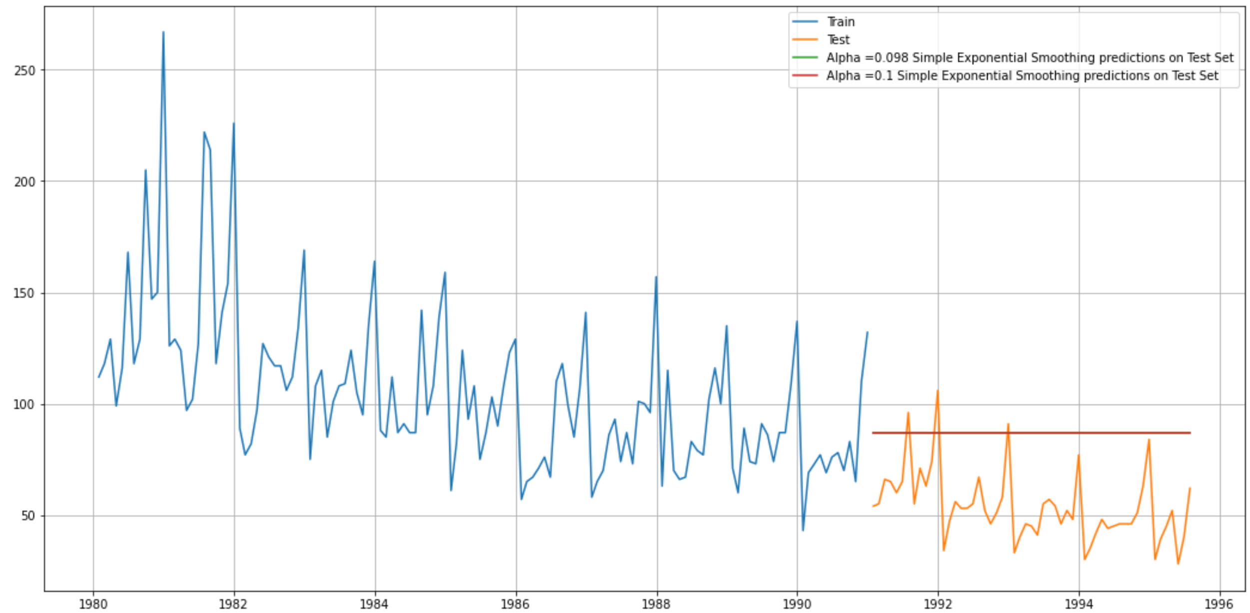

Figure-15

Alpha=0.1,SimpleExponentialSmoothing RMSE = 36.807579

Double exponential smoothing :

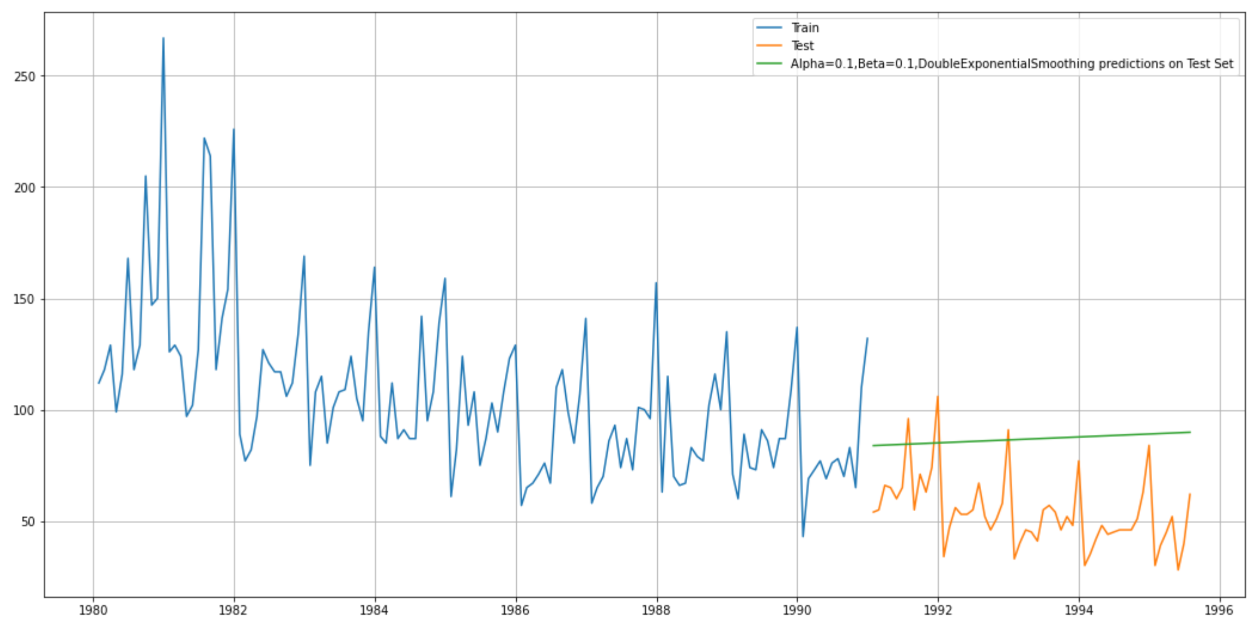

Figure - 16

Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing 36.902316

Triple exponential smoothing :

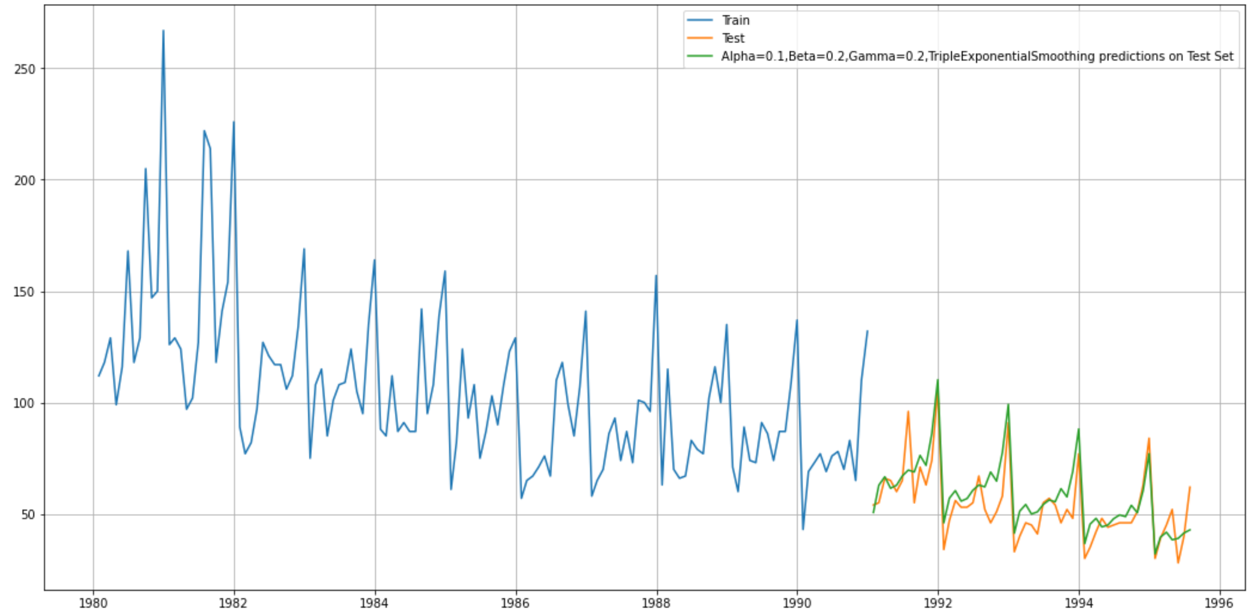

Figure - 17

Alpha=0.1,Beta=0.2,Gamma=0.2,TripleExponentialSmoothing 9.633969

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.

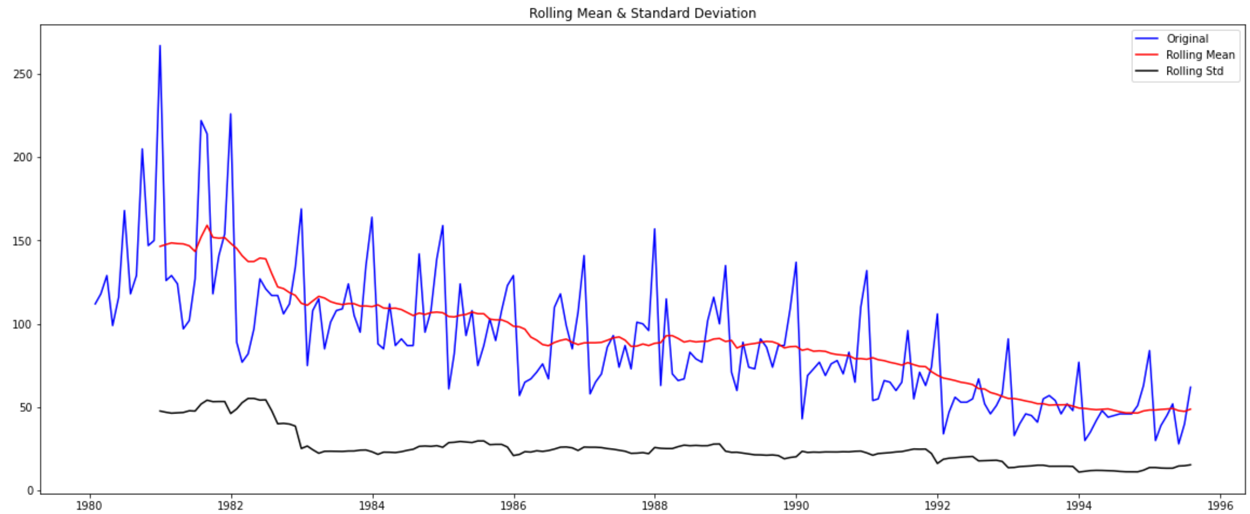

Figure - 18

```
Results of Dickey-Fuller Test:
Test Statistic                  -1.877440
p-value                          0.342747
#Lags Used                      13.000000
Number of Observations Used    173.000000
Critical Value (1%)             -3.468726
Critical Value (5%)             -2.878396
Critical Value (10%)            -2.575756
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.

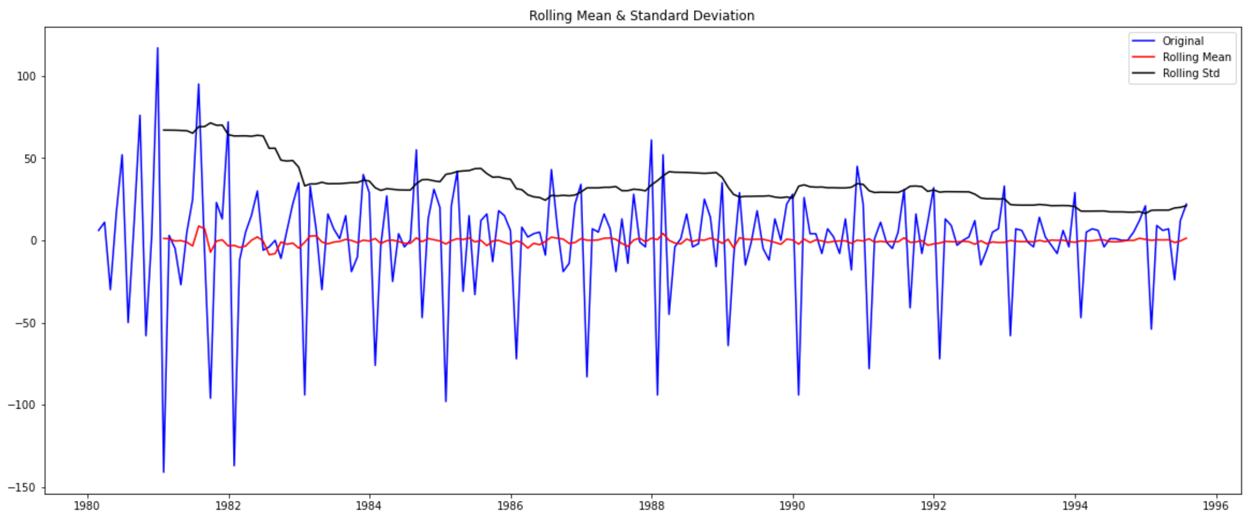

Figure - 19

```
Results of Dickey-Fuller Test:
Test Statistic                  -8.044614e+00
p-value                          1.808550e-12
#Lags Used                       1.200000e+01
Number of Observations Used      1.730000e+02
Critical Value (1%)             -3.468726e+00
Critical Value (5%)             -2.878396e+00
Critical Value (10%)            -2.575756e+00
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
15	(3, 0, 3)	1290.913289
6	(1, 0, 2)	1292.053213
10	(2, 0, 2)	1292.248056
7	(1, 0, 3)	1292.929011
9	(2, 0, 1)	1292.937195
13	(3, 0, 1)	1293.883000
14	(3, 0, 2)	1294.247915
11	(2, 0, 3)	1294.248232
5	(1, 0, 1)	1294.510585
12	(3, 0, 0)	1296.776953

Table-07

```
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(2, 1, 3)   Log Likelihood                -631.347
Date:                Wed, 16 Feb 2022   AIC                           1274.695
Time:                        11:35:56   BIC                           1291.946
Sample:                    01-31-1980   HQIC                          1281.705
                         - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.6781      0.084    -20.035      0.000      -1.842      -1.514
ar.L2         -0.7289      0.084     -8.703      0.000      -0.893      -0.565
ma.L1          1.0450      0.685      1.527      0.127      -0.297       2.387
ma.L2         -0.7716      0.137     -5.636      0.000      -1.040      -0.503
ma.L3         -0.9046      0.622     -1.455      0.146      -2.123       0.314
sigma2       858.3595    576.845      1.488      0.137    -272.237    1988.956
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                24.45
Prob(Q):                              0.88   Prob(JB):                         0.00
Heteroskedasticity (H):               0.40   Skew:                             0.71
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.57
===================================================================================
```

Note : All lags are not significant. We can expect the result not to be good.

RMSE of Test data : ARIMA(2,1,3) 36.79705233275287

SARIMA :

AIC Values : In Ascending order

	param	seasonal	AIC
606	(4, 1, 4)	(1, 1, 1, 6)	1300.083427
32	(0, 1, 1)	(1, 1, 2, 6)	1303.285643
157	(1, 1, 1)	(1, 1, 2, 6)	1303.900592
115	(0, 1, 4)	(3, 1, 0, 6)	1307.399805
457	(3, 1, 3)	(1, 1, 2, 6)	1307.807674

Table-08

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                y   No. Observations:                   132
Model:             SARIMAX(4, 1, 4)x(1, 1, [1], 6)   Log Likelihood                 -520.116
Date:                          Tue, 15 Feb 2022   AIC                            1062.232
Time:                                  23:35:00   BIC                            1092.330
Sample:                                       0   HQIC                           1074.447
                                          - 132
Covariance Type:                            opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.7041      0.121     -5.821      0.000      -0.941      -0.467
ar.L2         -0.8000      0.102     -7.849      0.000      -1.000      -0.600
ar.L3         -0.6397      0.100     -6.421      0.000      -0.835      -0.444
ar.L4         -0.0555      0.087     -0.639      0.523      -0.226       0.115
ma.L1          0.0217    199.186      0.000      1.000    -390.377     390.420
ma.L2       8.261e-08    997.038   8.29e-11      1.000   -1954.159    1954.159
ma.L3         -0.0217    190.887     -0.000      1.000    -374.153     374.110
ma.L4         -1.0000   1027.440     -0.001      0.999   -2014.744    2012.744
ar.S.L6       -0.9062      0.025    -36.972      0.000      -0.954      -0.858
ma.S.L6        0.3741      0.099      3.766      0.000       0.179       0.569
sigma2       477.7764   4.91e+05      0.001      0.999   -9.62e+05    9.63e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.30   Jarque-Bera (JB):                 8.48
Prob(Q):                              0.58   Prob(JB):                         0.01
Heteroskedasticity (H):               0.39   Skew:                             0.23
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.26
===================================================================================
```

RMSE of test data :

SARIMA(4,1,4)(1,1,1,6) 17.145498

	param	seasonal	AIC
26	(0, 1, 2)	(2, 0, 2, 12)	887.937509
53	(1, 1, 2)	(2, 0, 2, 12)	889.903048
80	(2, 1, 2)	(2, 0, 2, 12)	890.668798
69	(2, 1, 1)	(2, 0, 0, 12)	896.518161
78	(2, 1, 2)	(2, 0, 0, 12)	897.346444

Table-09

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                    y   No. Observations:                  132
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 12)   Log Likelihood                -436.969
Date:                              Tue, 15 Feb 2022   AIC                            887.938
Time:                                      23:40:22   BIC                            906.448
Sample:                                           0   HQIC                           895.437
                                              - 132
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.8427    189.943     -0.004      0.996    -373.124     371.439
ma.L2         -0.1573     29.841     -0.005      0.996     -58.645      58.330
ar.S.L12       0.3467      0.079      4.375      0.000       0.191       0.502
ar.S.L24       0.3023      0.076      3.996      0.000       0.154       0.451
ma.S.L12       0.0767      0.133      0.577      0.564      -0.184       0.337
ma.S.L24      -0.0726      0.146     -0.498      0.618      -0.358       0.213
sigma2       251.3137   4.77e+04      0.005      0.996   -9.33e+04    9.38e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.10   Jarque-Bera (JB):                 2.33
Prob(Q):                              0.75   Prob(JB):                         0.31
Heteroskedasticity (H):               0.88   Skew:                             0.37
Prob(H) (two-sided):                  0.70   Kurtosis:                         3.03
===================================================================================
```

RMSE of test data :

SARIMA(0,1,2)(2,0,2,12) 26.907439

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.

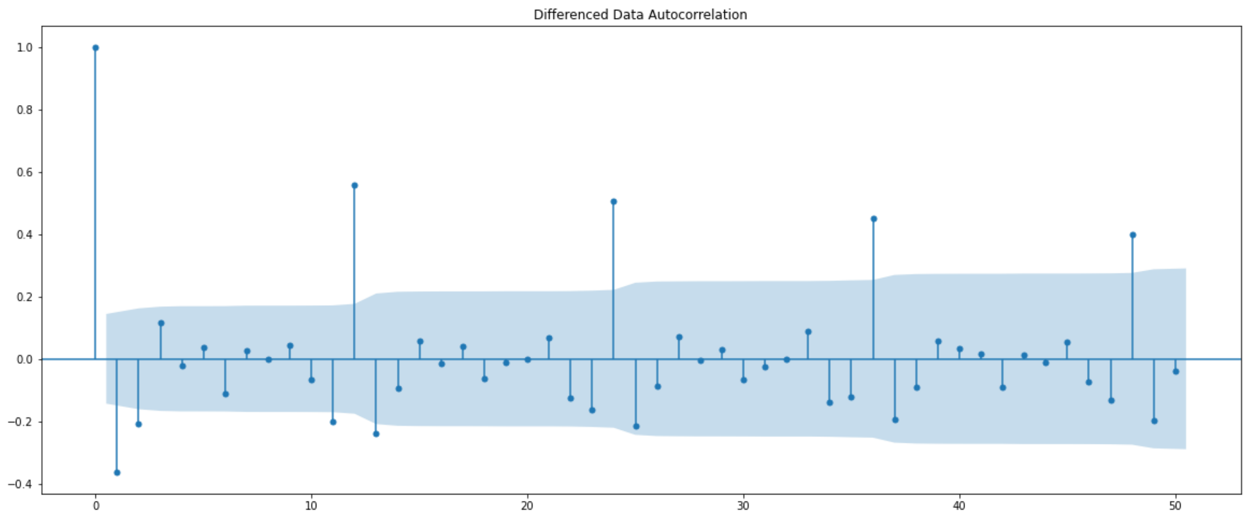

Figure - 20

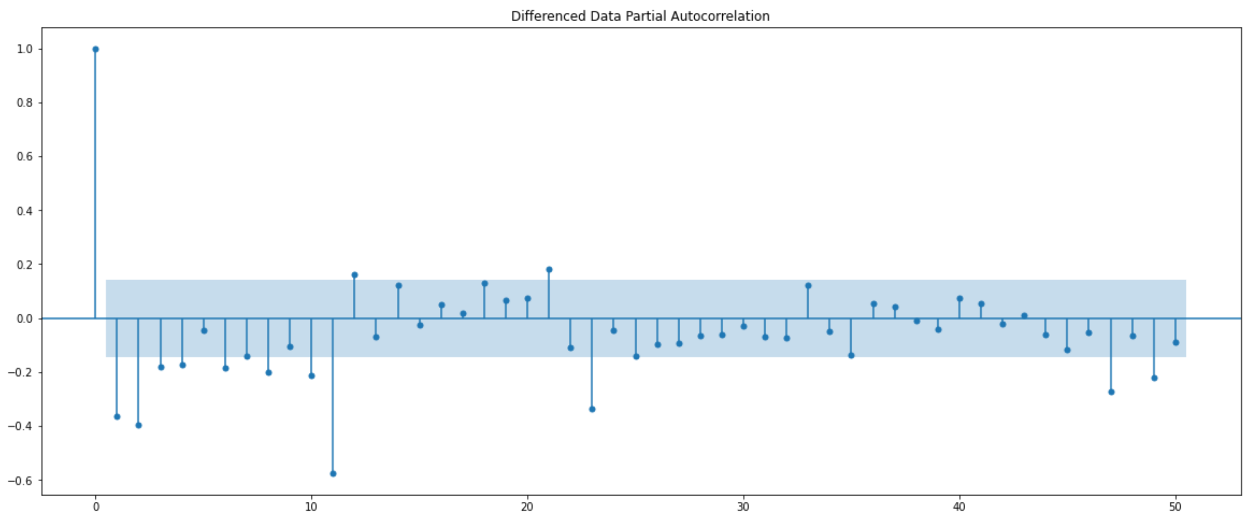

Figure-21

```
                                    SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(1, 1, 0)   Log Likelihood                -656.675
Date:                Tue, 15 Feb 2022   AIC                           1317.350
Time:                        23:48:33   BIC                           1323.101
Sample:                    01-31-1980   HQIC                          1319.687
                         - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.3555      0.067     -5.274      0.000      -0.488      -0.223
sigma2      1321.6677    125.729     10.512      0.000    1075.243    1568.092
===================================================================================
Ljung-Box (L1) (Q):                   2.56   Jarque-Bera (JB):                19.96
Prob(Q):                              0.11   Prob(JB):                         0.00
Heteroskedasticity (H):               0.32   Skew:                            -0.34
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.79
===================================================================================
```

RMSE of test data :

ARIMA(1,1,0) 74.025930

```
                                     SARIMAX Results
=========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(1, 1, 0)x(0, 1, 0, 6)   Log Likelihood                -672.458
Date:                           Tue, 15 Feb 2022   AIC                           1348.917
Time:                                   23:54:52   BIC                           1354.557
Sample:                                        0   HQIC                          1351.208
                                           - 132
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.3453      0.083     -4.175      0.000      -0.507      -0.183
sigma2      3005.5811    308.165      9.753      0.000    2401.590    3609.573
===================================================================================
Ljung-Box (L1) (Q):                   1.53   Jarque-Bera (JB):                20.71
Prob(Q):                              0.22   Prob(JB):                         0.00
Heteroskedasticity (H):               0.30   Skew:                            -0.07
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.00
===================================================================================
```

RMSE of test data :

SARIMA(1,1,0)(0,1,0,6) 334.497445

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=0.1,Beta=0.2,Gamma=0.2,TripleExponentialSmoothing	9.633969
2pointTrailingMovingAverage	11.529409
4pointTrailingMovingAverage	14.448930
6pointTrailingMovingAverage	14.560046
9pointTrailingMovingAverage	14.724503
12pointTrailingMovingAverage	15.234402
RegressionOnTime	15.262509
SARIMA(4,1,4)(1,1,1,6)	17.145498
Alpha=0.065,Beta=0.0519,Gamma=3.879136202038614e-06,TripleExponentialSmoothing	20.995338
SARIMA(0,1,2)(2,0,2,12)	26.907439
Alpha=0.098,SimpleExponentialSmoothing	36.775774
ARIMA(2,1,3)	36.797052
Alpha=0.1,SimpleExponentialSmoothing	36.807579
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	36.902316
SimpleAverageModel	53.440426
ARIMA(1,1,0)	74.025930
NaiveModel	79.699093
SARIMA(1,1,0)(0,1,0,6)	334.497445

Table - 10

Note : We can say that the triple exponential(alpha=0.1, Beta=0.2, gamma=0.2) 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.

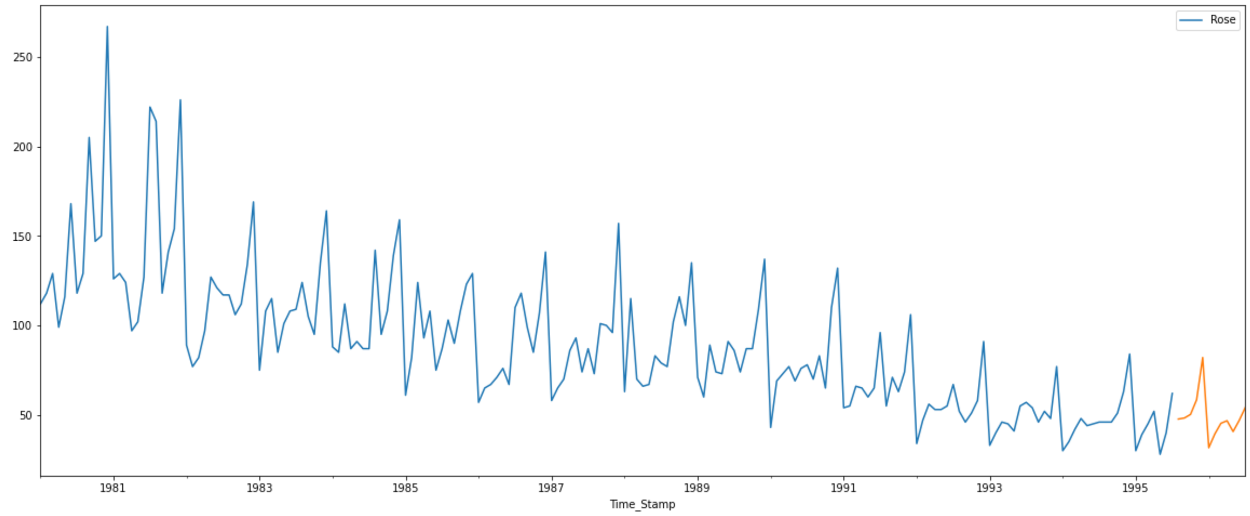

Figure - 22

RMSE of the Full Model 17.40194318433039

	lower_CI	prediction	upper_ci
1995-08-31	13.514598	47.695360	81.876121
1995-09-30	14.103116	48.283878	82.464639
1995-10-31	16.087505	50.268267	84.449028
1995-11-30	24.256156	58.436918	92.617679
1995-12-31	47.893523	82.074285	116.255046

Table-10

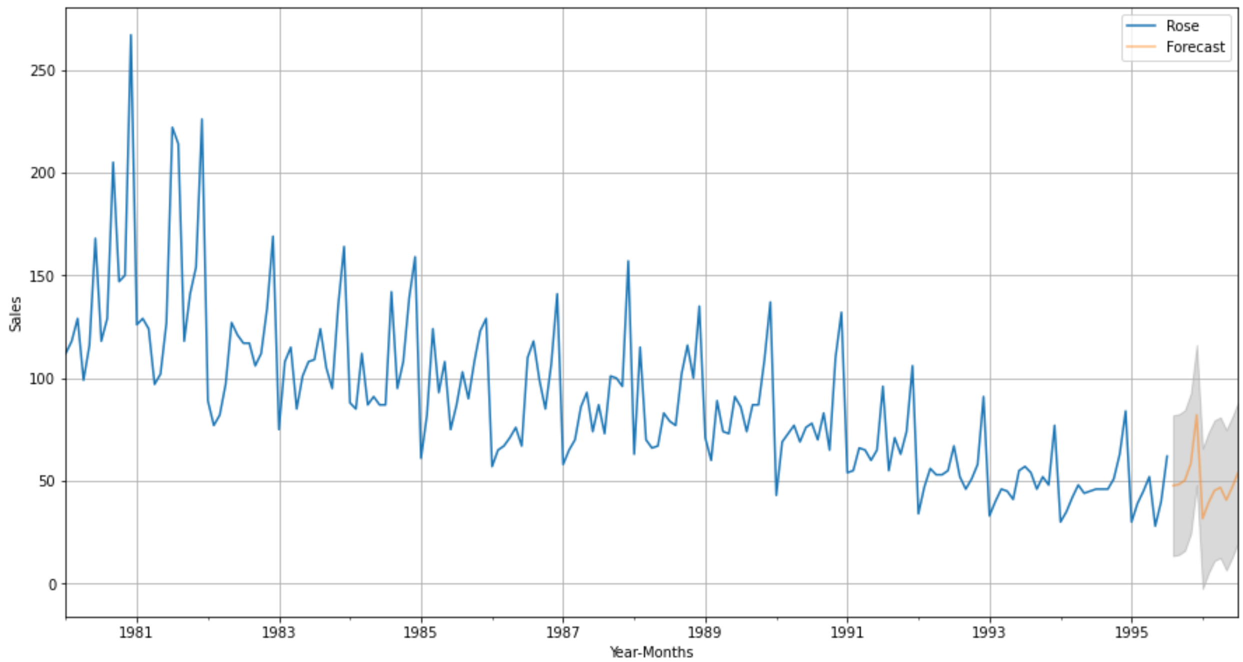

Figure - 23

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 we see in the decomposing the datasets trend is going downward so we need to look up seriously why this is happening and try to make sales in an upward trend.
- In December, sales are high.
- need to run promotional marketing campaigns or evaluate if we need to tie up with an alternate agency. It will increase sales.