BUSINESS REPORT

ON

TIME SERIES FORECASTING

By Kshitij Nishant

Table of Contents

Contents

List of Figures

Figures

This particular report is for Sparkling.csv

Problem:

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

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

[So to start with I have made report one dataset at a time, starting with Sparkling.csv]

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

Answer:

Head:

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

Tail:

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

I converted them into date format:

	YearMonth	Sparkling	Date
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 have converted the data into date format and given the column name as Date.
I also dropped the column YearMonth as we got the month year and date format in one column named Time_Stamp:

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

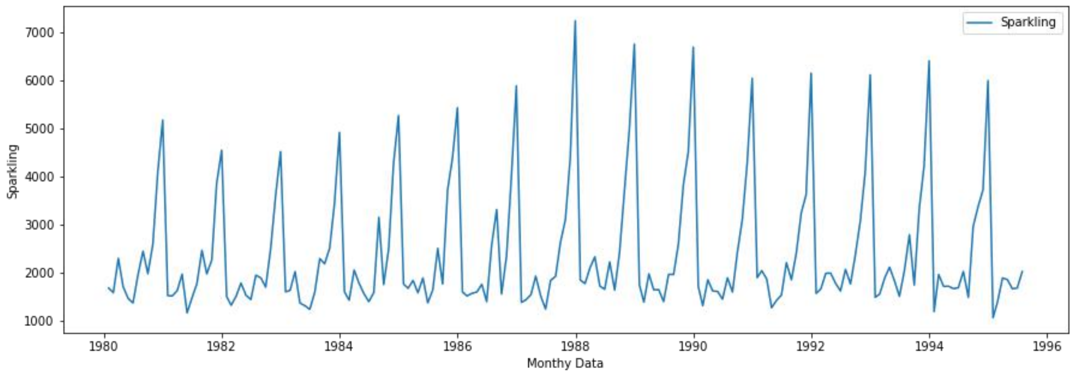

Fig1

1. Sparkling wine sales show no much trend in the yearly sale.
2. Sparkling wine sales shows seasonality which has yearly pattern.

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

Answer:

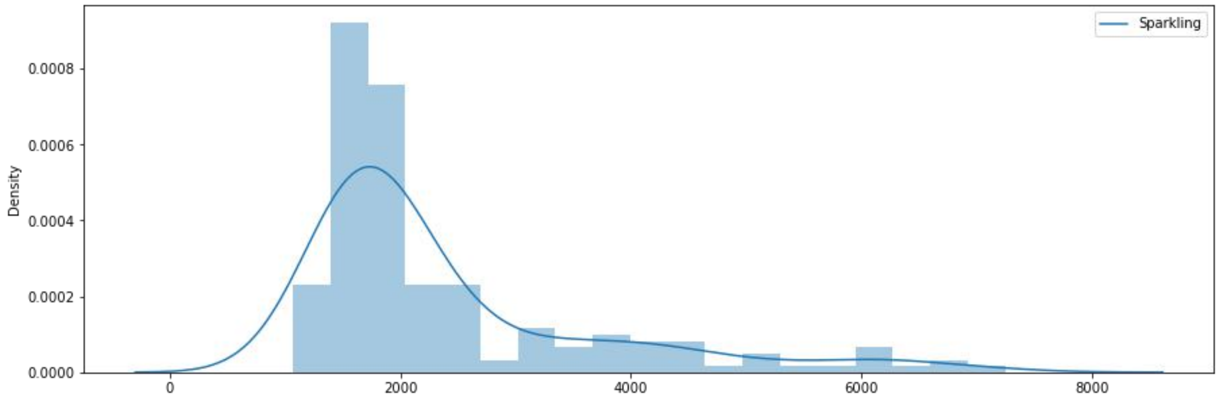

Fig 2

Data is skewed towards left.

Description of data:

	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

• There are 187 observations which represent the monthly sales of respective wines form the year 1980 to July 1995.

• The data has two variables the year/month of sales and the sales for the respective month of the year.

• Mean, min, max values for sparkling wine sales are greater than rose wine sales.

<u>Shape of data: </u> (187, 2)

<u>Null value check:</u> There are no null values in data set Sparkling.

Distribution of sale of wine-Sparkling in each year via BoxPlot:

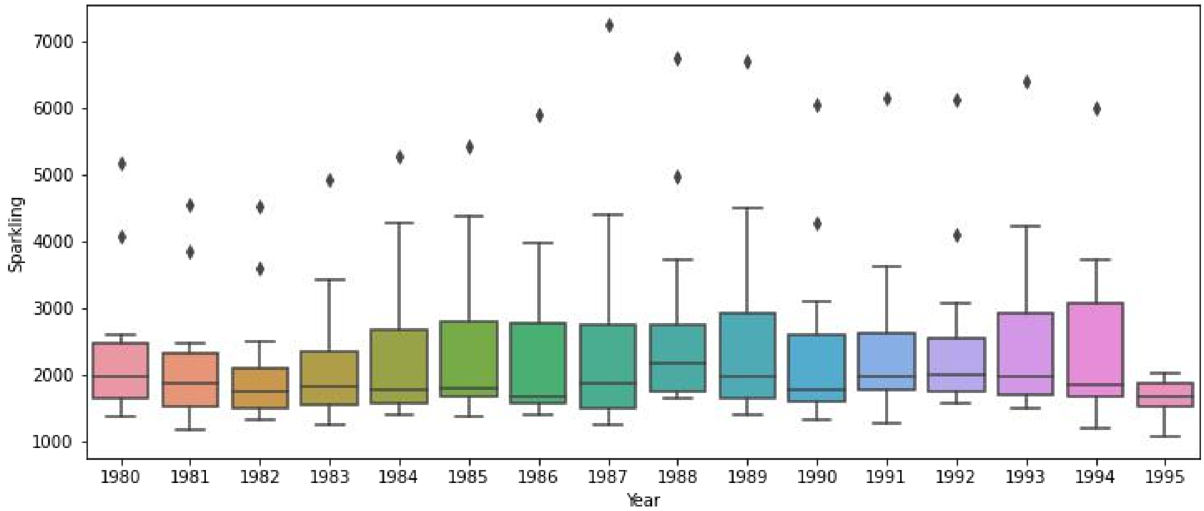

Fig3

- Outliers are present when looking at corresponding year wise data.

Distribution of sale of wine-Sparkling in each year via BarPlot:

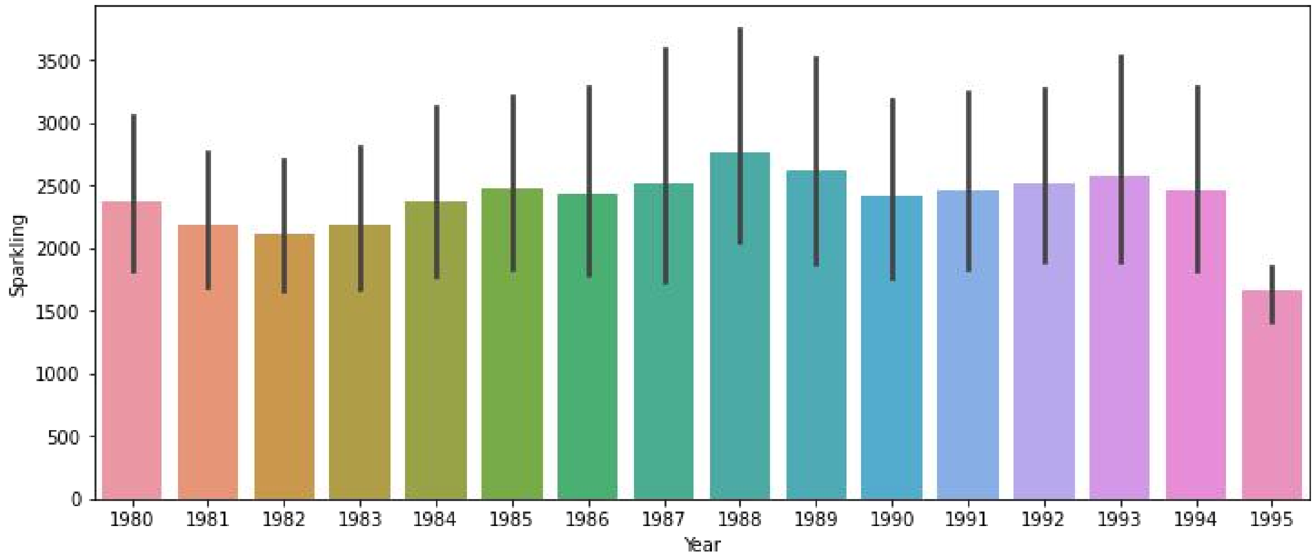

Fig4

- Data seems to have more or less same sales across the year. 1988 has recorded maximum sales.

Distribution of sale of wine-Sparkling in each Month via BarPlot:

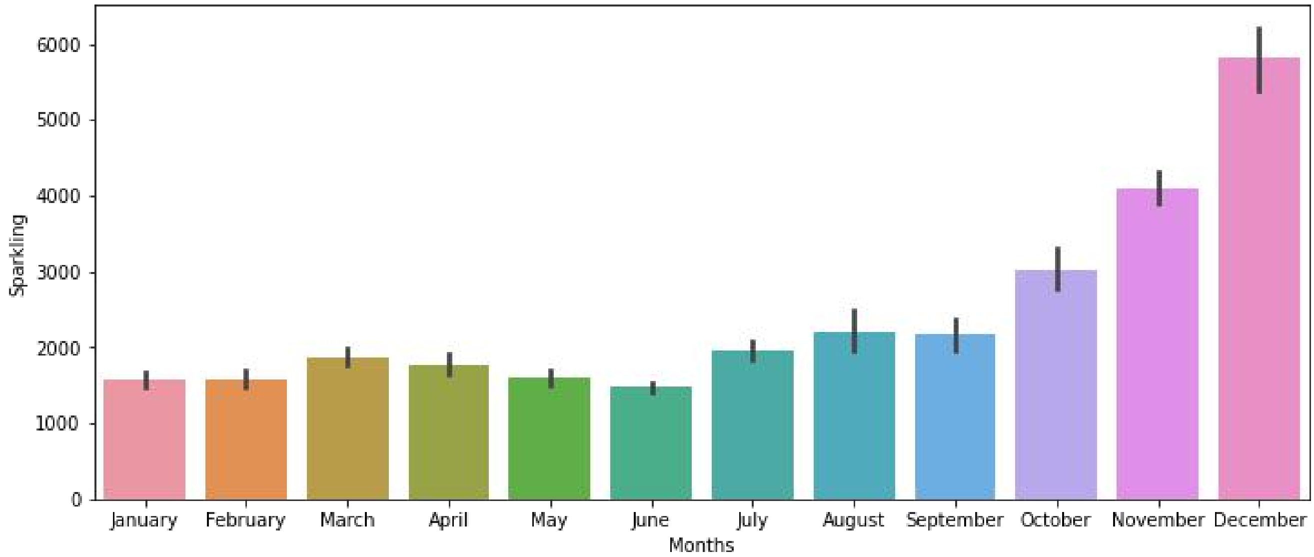

Fig5

- December have greatest amount of sales across all the months followed be November and Oct.

- Greater in sales may be due to the celebration in year end.

Distribution of sale of wine-Sparkling in each Month via BoxPlot:

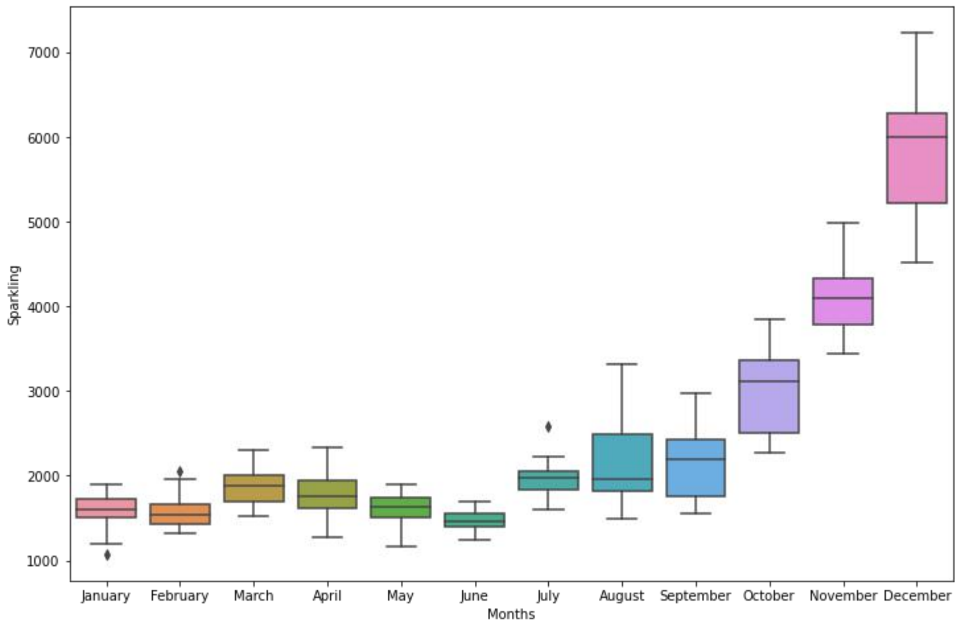

Fig6

- Box plot is also shows us that December has recorded most number of sales.

Distribution of average sale of wine-Sparkling of each Day via BarPlot:

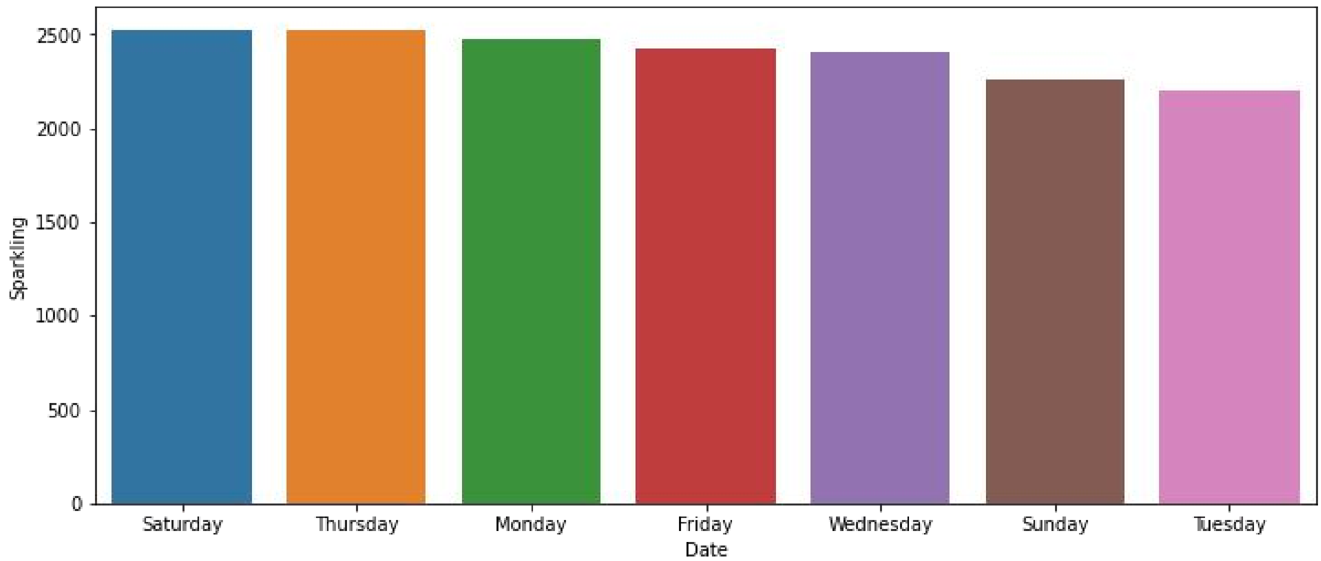

Fig7

- Saturday registers highest average sales of beer throughout the whole week.

Distribution of daily sale of wine-Sparkling of each day via BarPlot:

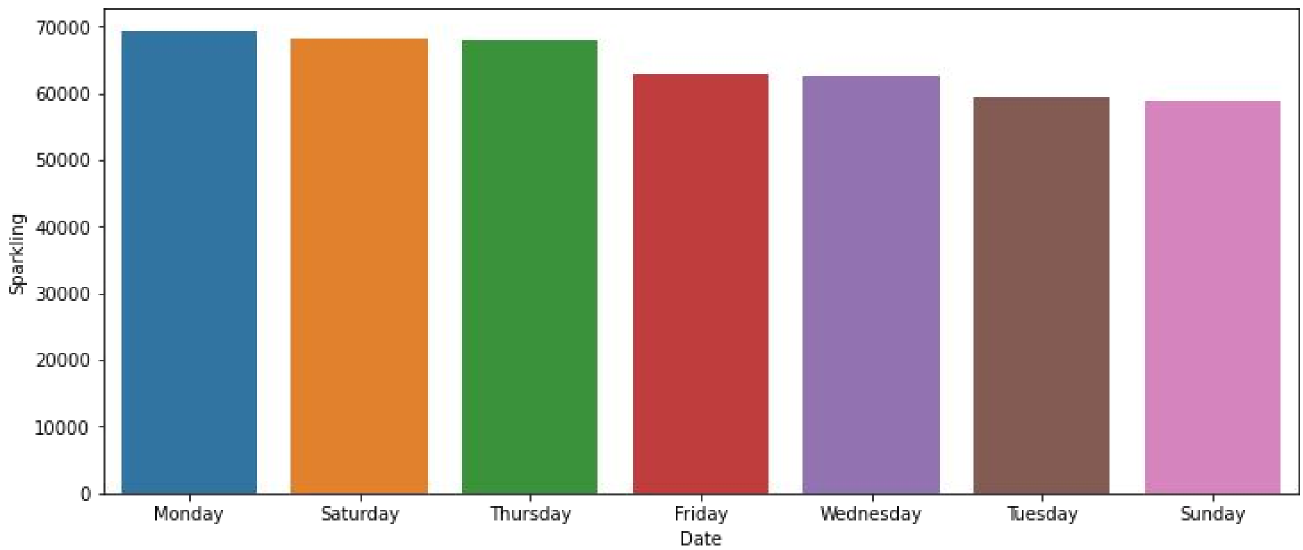

Fig8

- Monday has the highest sales overall.

Time series monthplot to understand the spread of Sparkling Sales across different years and within different months across years:

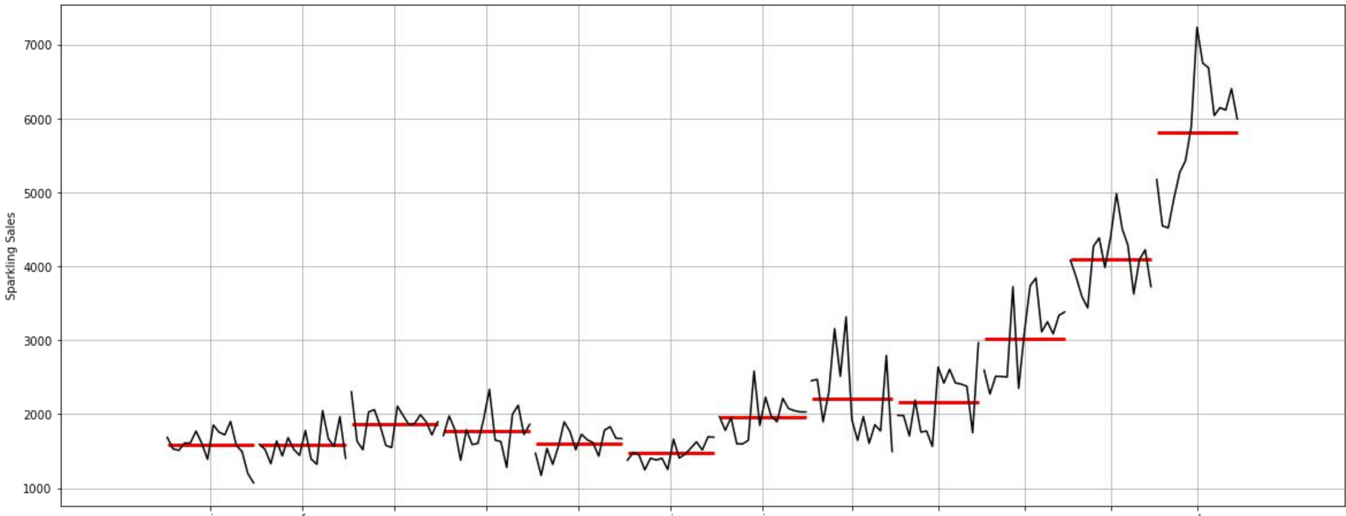

Fig9

Graph of monthly Sparkling's Sales across years:

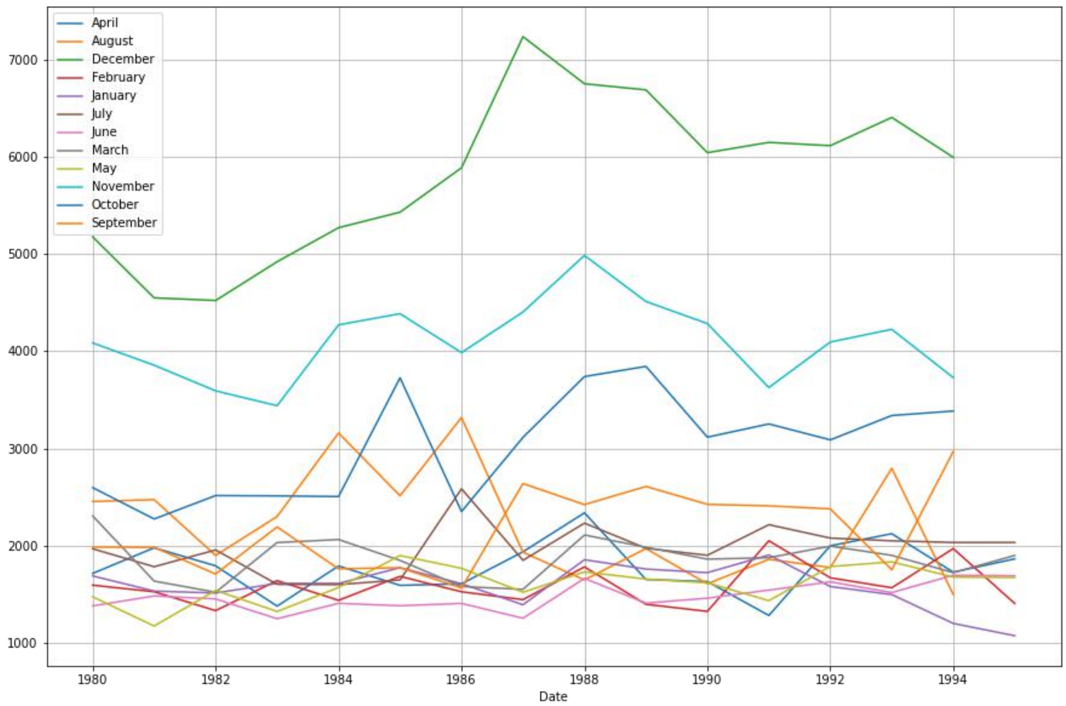

Fig 10

- Dec registers the highest amount of sales.

Empirical Cumulative Distribution:

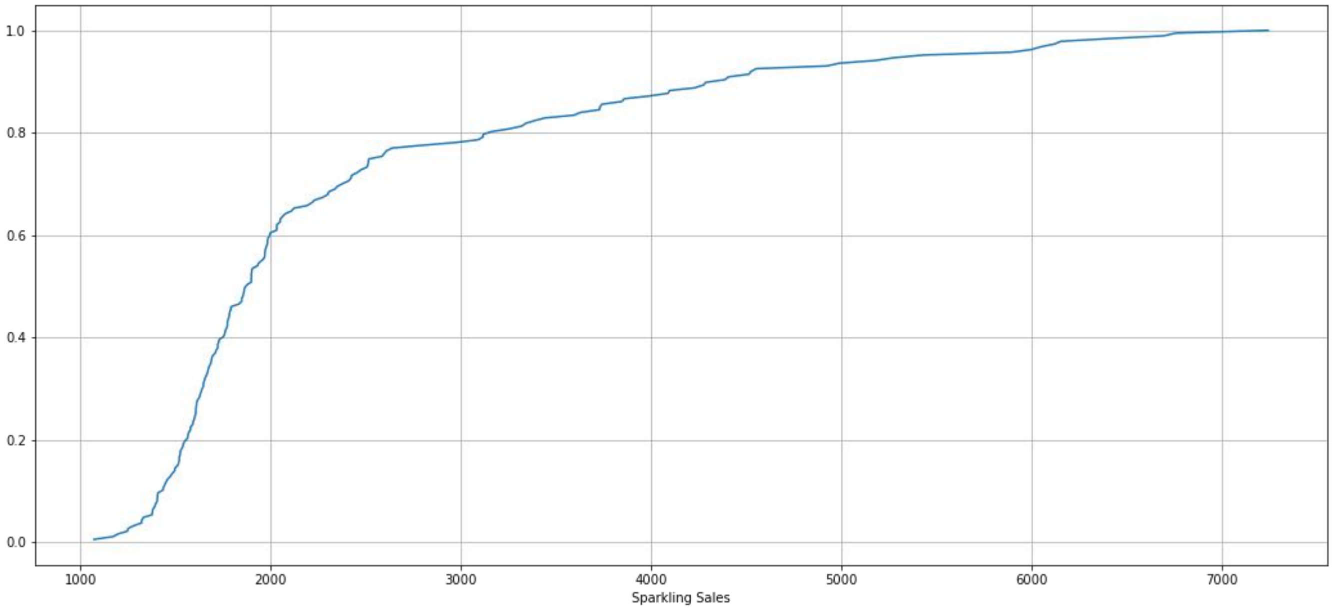

Fig 11

Average Sparkling Sales per month and the month on month percentage change of Sparkling Sales:

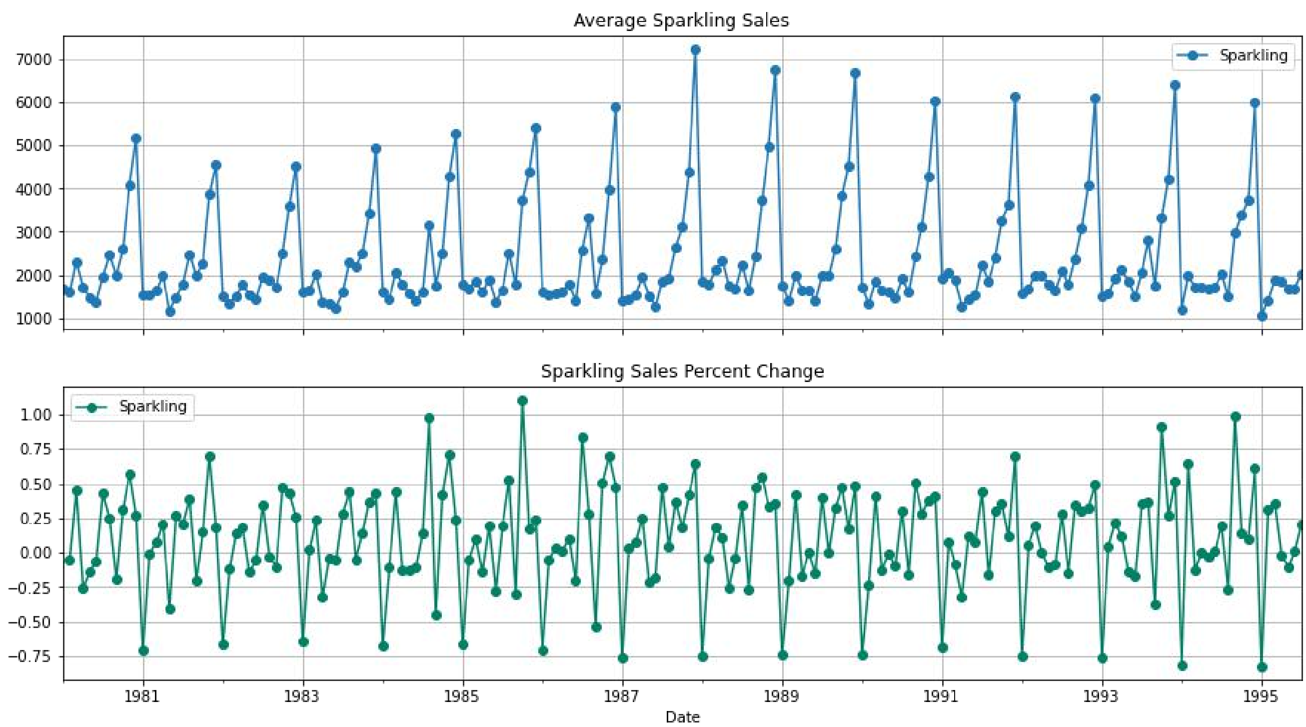

Fig 12

- The median values are stable from January to June and has an increasing trend from July to December.

- The Average Sales value does not show a trend.

Additive Decomposition of dataset:

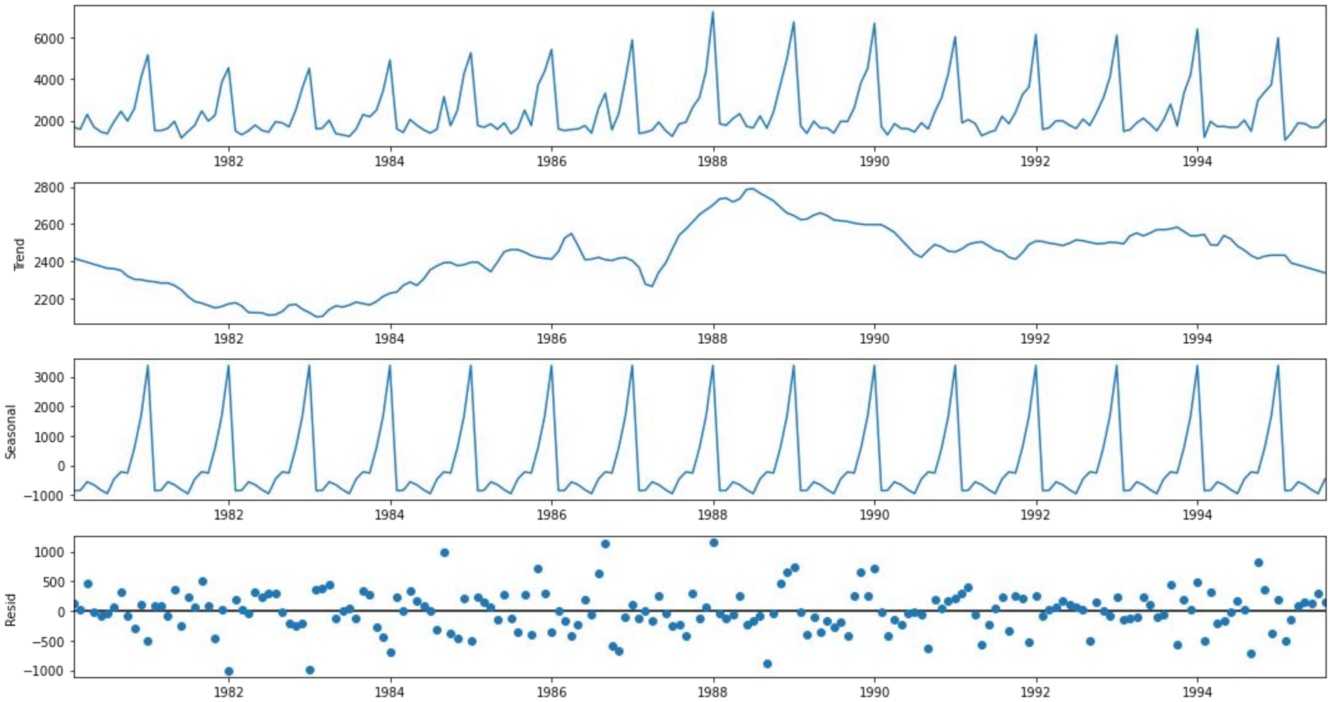

Fig 13

Multiplicative Decomposition of dataset:

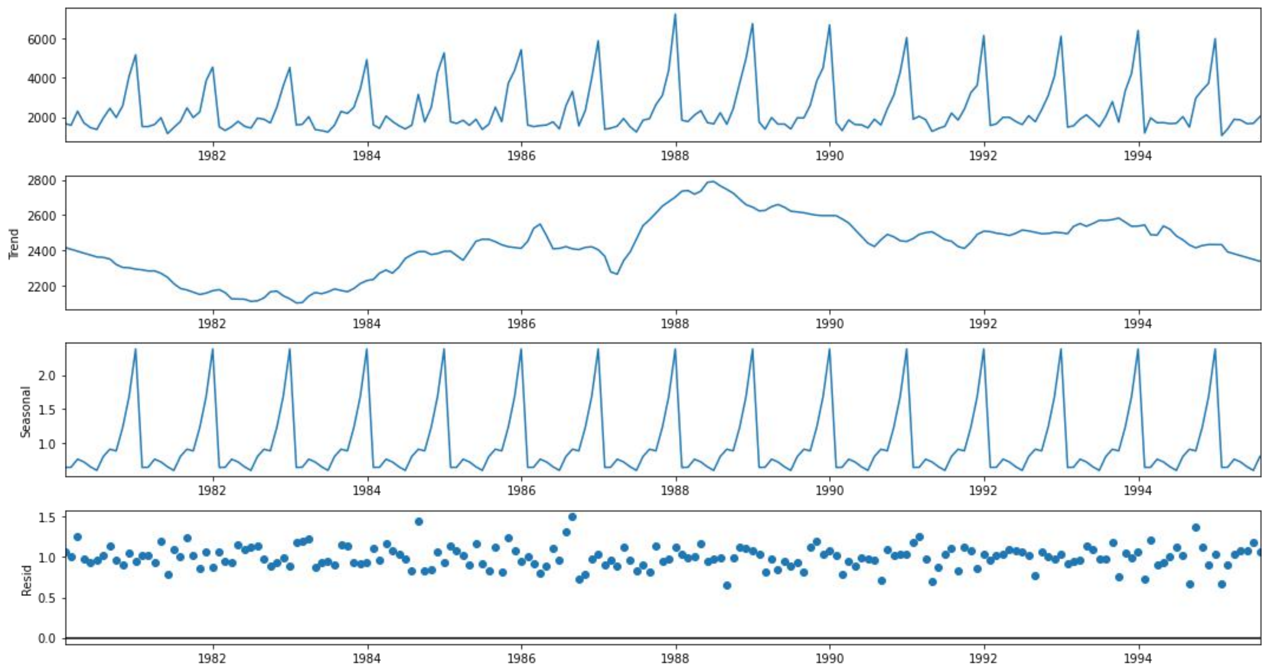

Fig 14

- For additive we see the residual values don’t make any pattern and there is no increasing treand or seasonality but for Multiplicative model we see the residual make some form of pattern.
- So I decided to choose additive model is better for forecasting Sparkling.csv.

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

<u>Answer:</u>

Tail of Train Set:

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

Head of Test Set:

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

Train-Test Plot

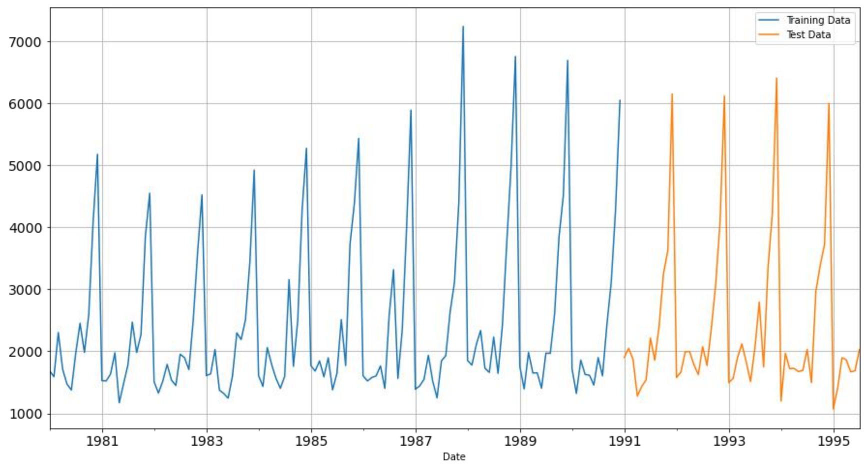

Fig15

4.Build all the 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.

<u>Answer:</u>

<u>*Model 1: Linear Regression*</u>

Train set after Predictions:

	Sparkling	train_time	RegOnTime
Date			
1980-01-31	1686	1	2021.741171
1980-02-29	1591	2	2027.573830
1980-03-31	2304	3	2033.406488
1980-04-30	1712	4	2039.239147
1980-05-31	1471	5	2045.071805

Test Set after Predictions:

	Sparkling	test_time	RegOnTime
Date			
1991-01-31	1902	133	2791.652093
1991-02-28	2049	134	2797.484752
1991-03-31	1874	135	2803.317410
1991-04-30	1279	136	2809.150069
1991-05-31	1432	137	2814.982727

Plotting Training,Testing and values obtained from regression model:

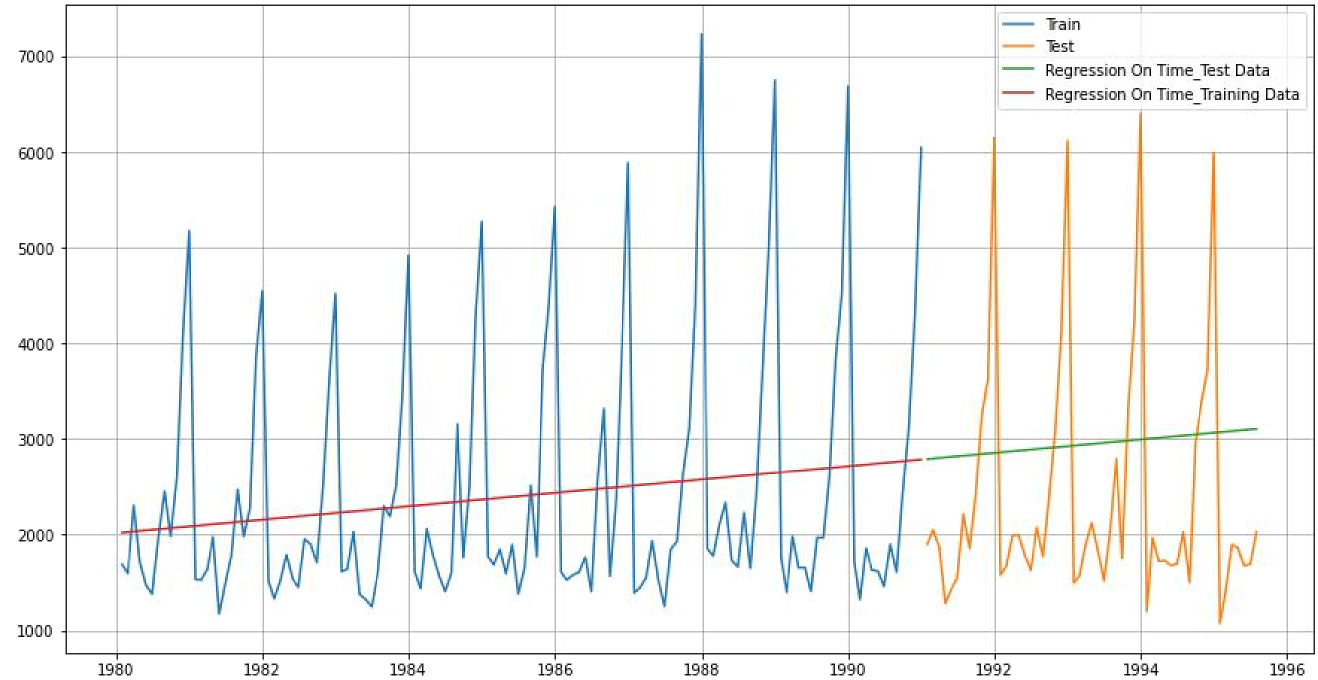

Fig16

Model Evaluation:

For RegressionOnTime forecast on the Training Data, RMSE is 1279.322 and MAPE is 40.05

For RegressionOnTime forecast on the Test Data, RMSE is 1389.135 and MAPE is 50.15

Model 2: Naive Approach

Train set Head after Naive Predictions:

Date
1980-01-31 6047
1980-02-29 6047
1980-03-31 6047
1980-04-30 6047
1980-05-31 6047

Test Set Head after Naive Predictions:

Date
1991-01-31 6047
1991-02-28 6047
1991-03-31 6047
1991-04-30 6047
1991-05-31 6047

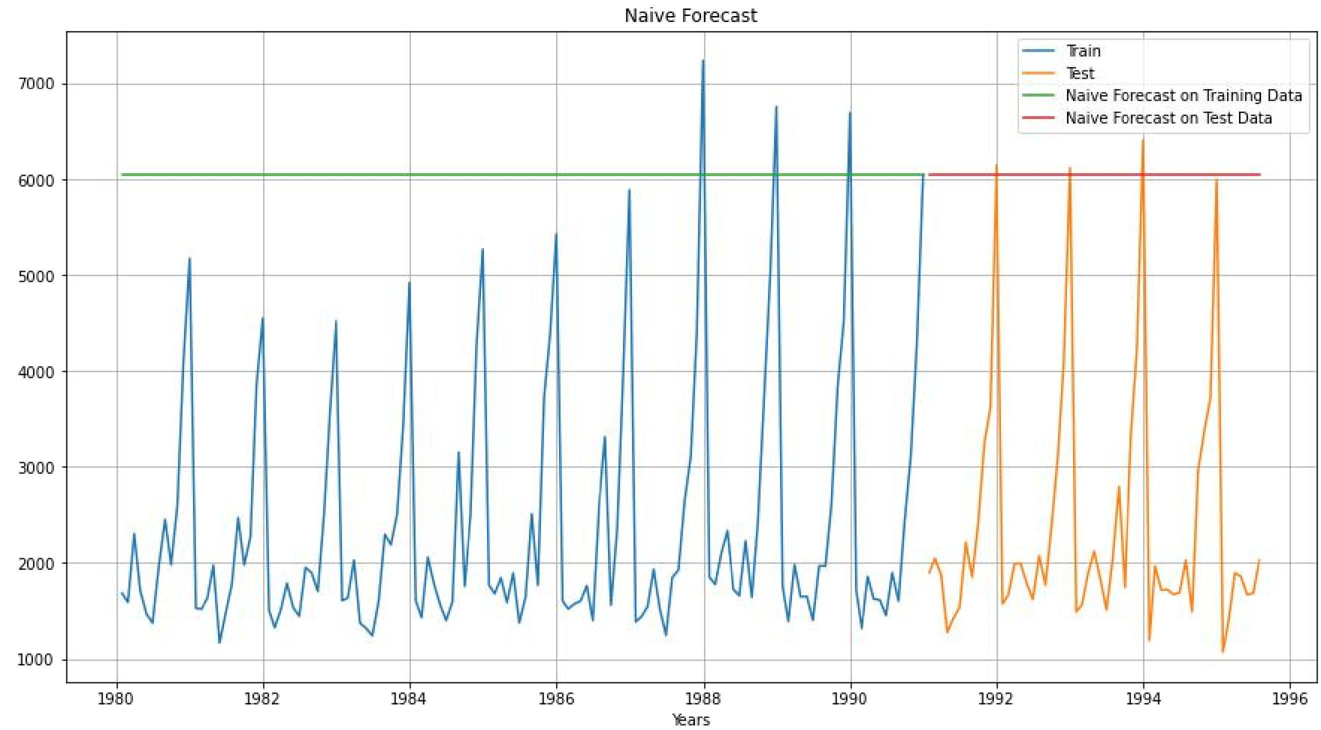

Fig 17

Model Evaluation:

For Naive Model forecast on the Training Data, RMSE is 3867.701 and MAPE is 153.17

For RegressionOnTime forecast on the Test Data, RMSE is 3864.279 and MAPE is 152.87

Method 3: Simple Average

Simple Avg Train set:

Date	Sparkling	mean_forecast
1980-01-31	1686	2403.780303
1980-02-29	1591	2403.780303
1980-03-31	2304	2403.780303
1980-04-30	1712	2403.780303
1980-05-31	1471	2403.780303

Simple Avg Test Set:

Date	Sparkling	mean_forecast
1991-01-31	1902	2403.780303
1991-02-28	2049	2403.780303
1991-03-31	1874	2403.780303
1991-04-30	1279	2403.780303
1991-05-31	1432	2403.780303

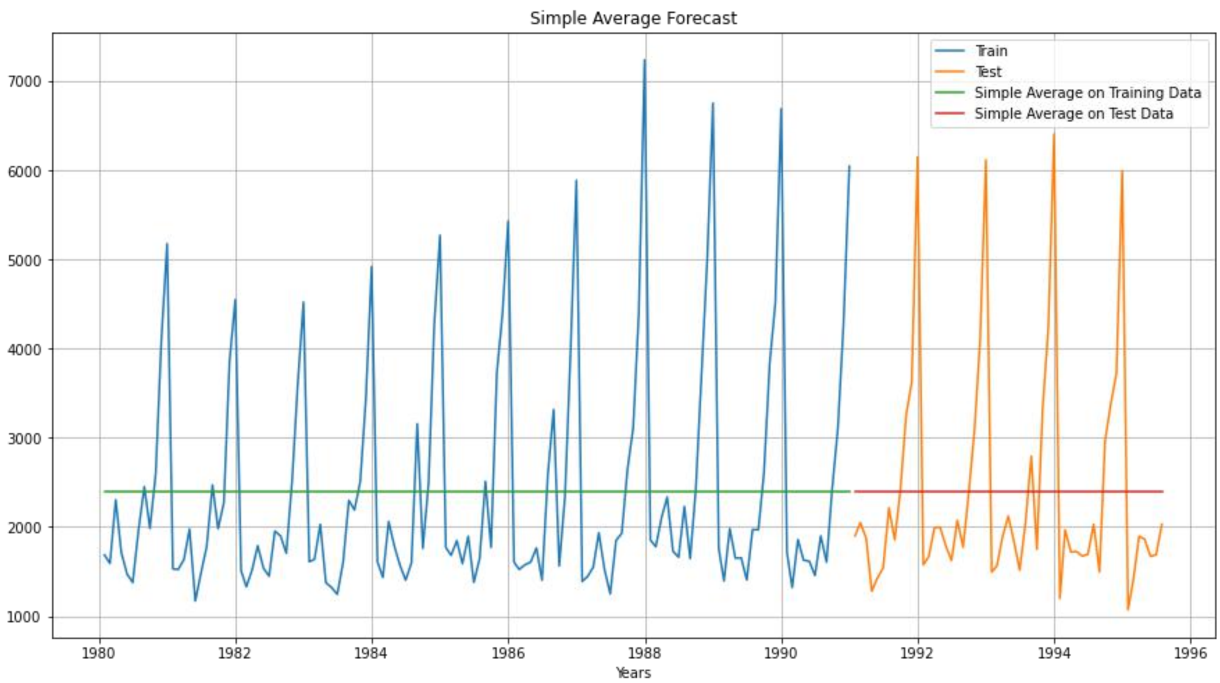

Fig 18

Model Evaluation of Simple Average:

For Simple Average Model forecast on the Training Data, RMSE is 1298.484 and MAPE is 40.36

For Simple Average forecast on the Test Data, RMSE is 1275.082 and MAPE is 38.90

Method 4: Moving Average(MA)

Moving Avg Train Set:

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

Trailing data set:

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

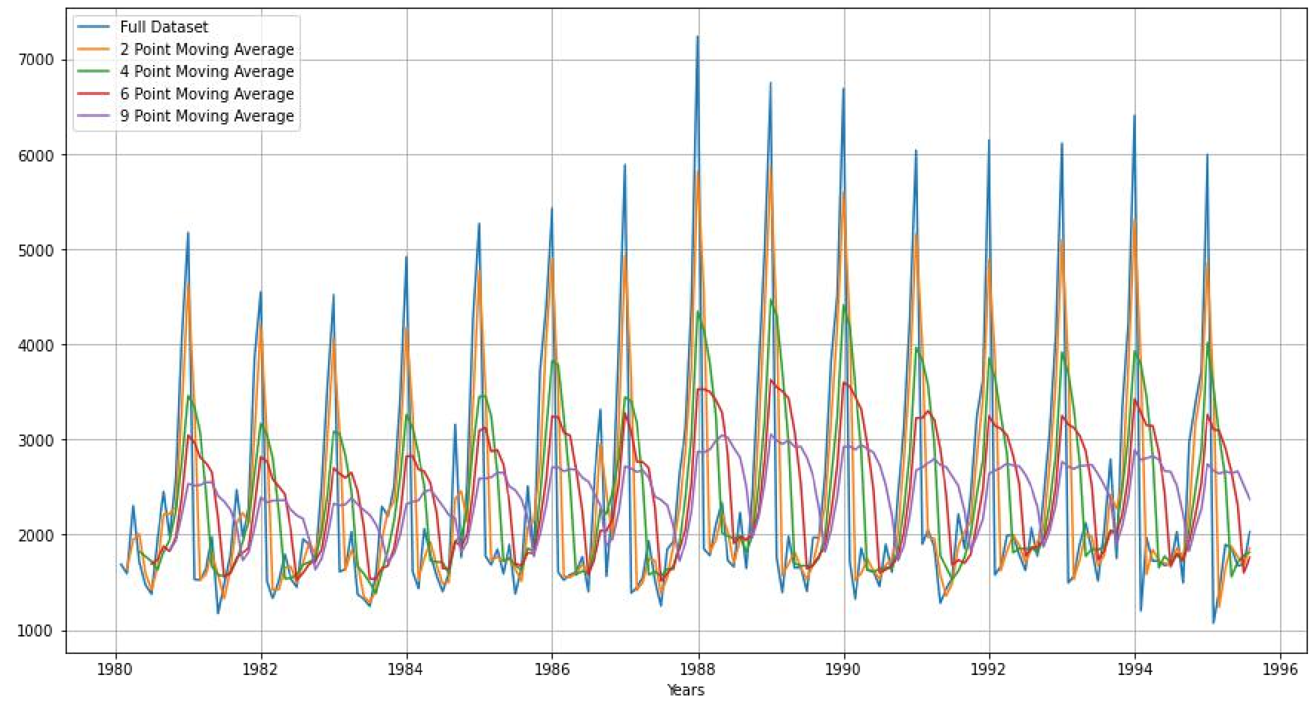

Fig 19

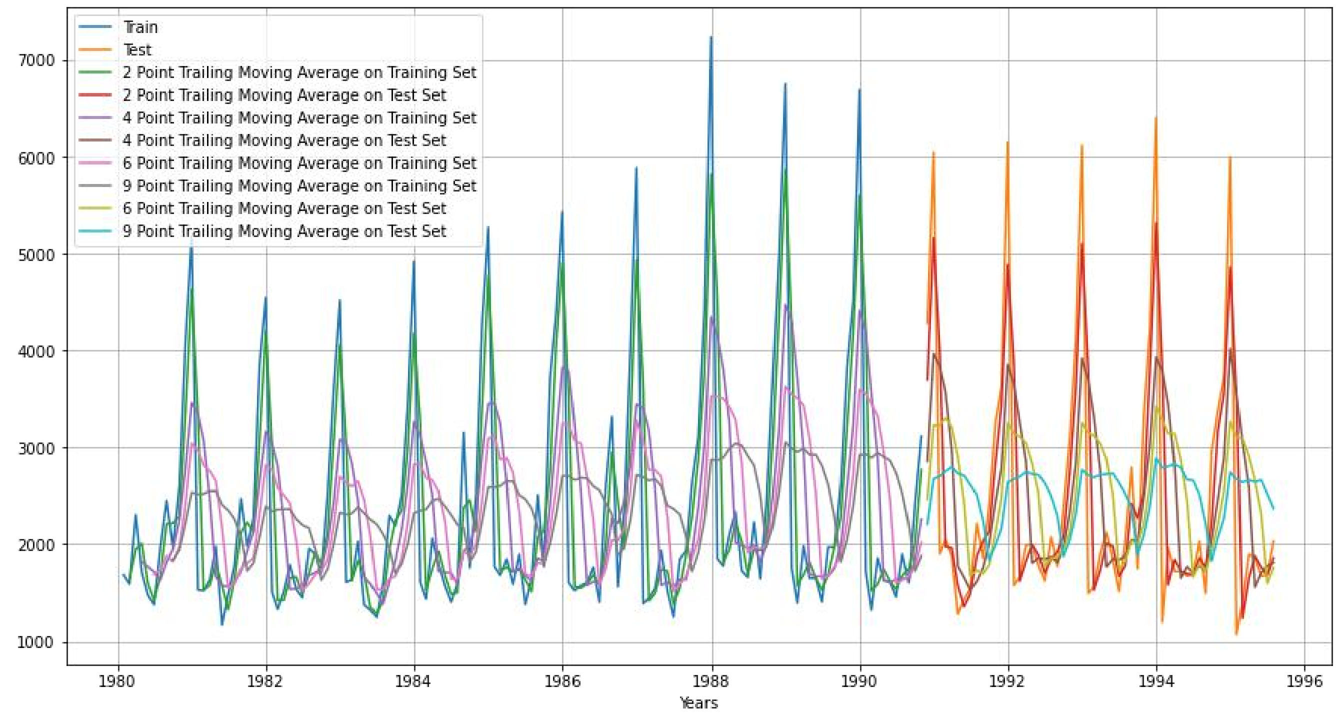

Fig 20

We see 2 point moving average gives best results here:

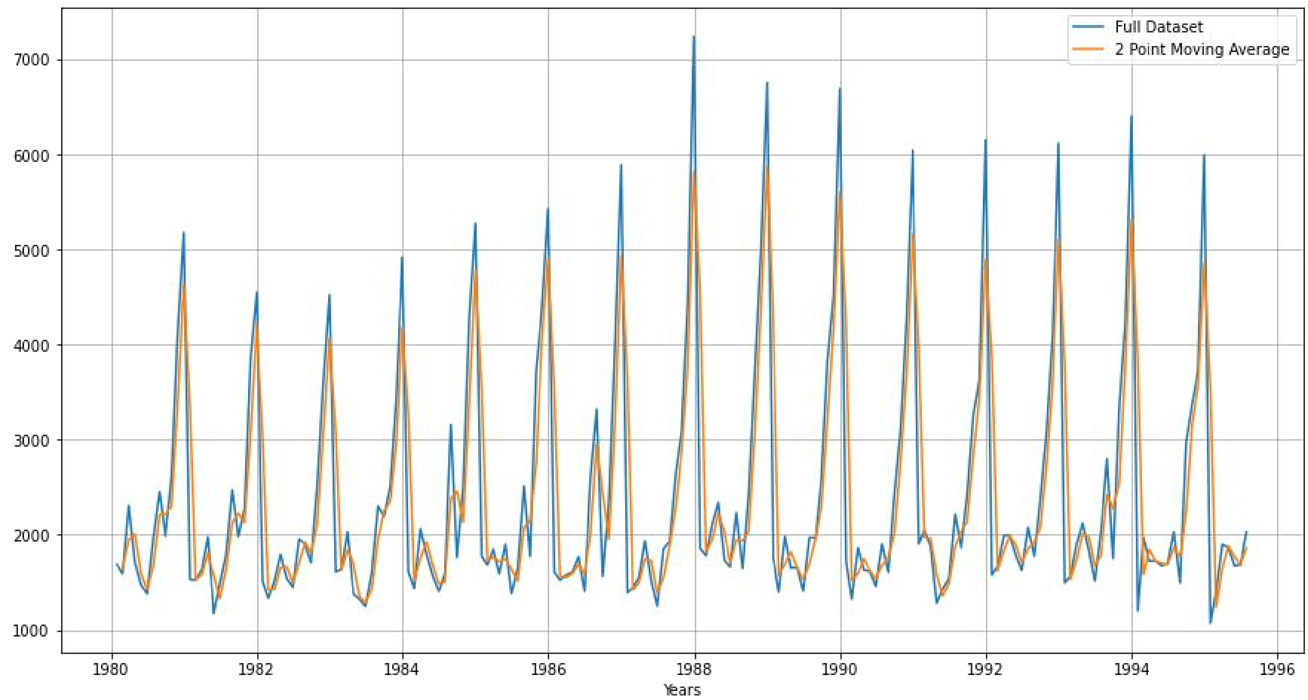

Fig 21

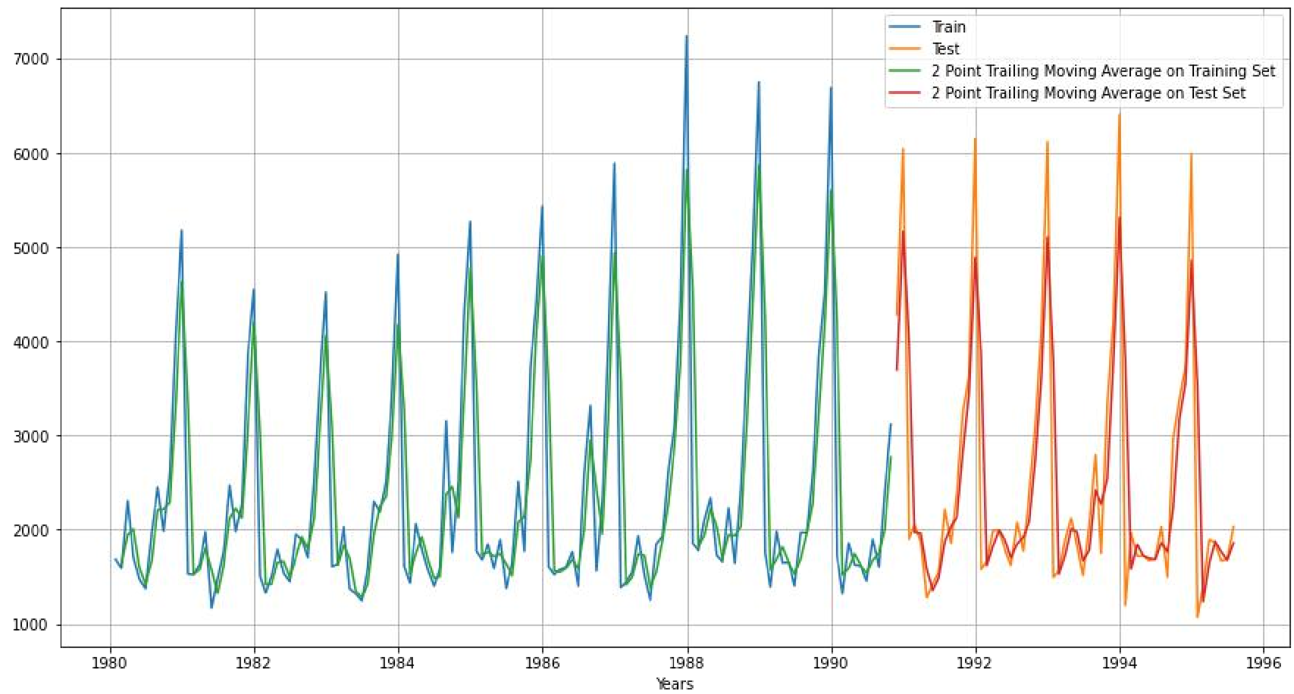

Fig 22

For 2 point Moving Average Model forecast on the Testing Data, RMSE is 813.401 and MAPE is 19.70

For 4 point Moving Average Model forecast on the Testing Data, RMSE is 1156.590 and MAPE is 35.96

For 6 point Moving Average Model forecast on the Testing Data, RMSE is 1283.927 and MAPE is 43.86

For 9 point Moving Average Model forecast on the Testing Data, RMSE is 1346.278 and MAPE is 46.86

Method 5: Simple Exponential Smoothing

SES Test Set after forecast:

	Sparkling	predict
Date		
1991-01-31	1902	2724.932624
1991-02-28	2049	2724.932624
1991-03-31	1874	2724.932624
1991-04-30	1279	2724.932624
1991-05-31	1432	2724.932624

Smoothing Level: 0.0496

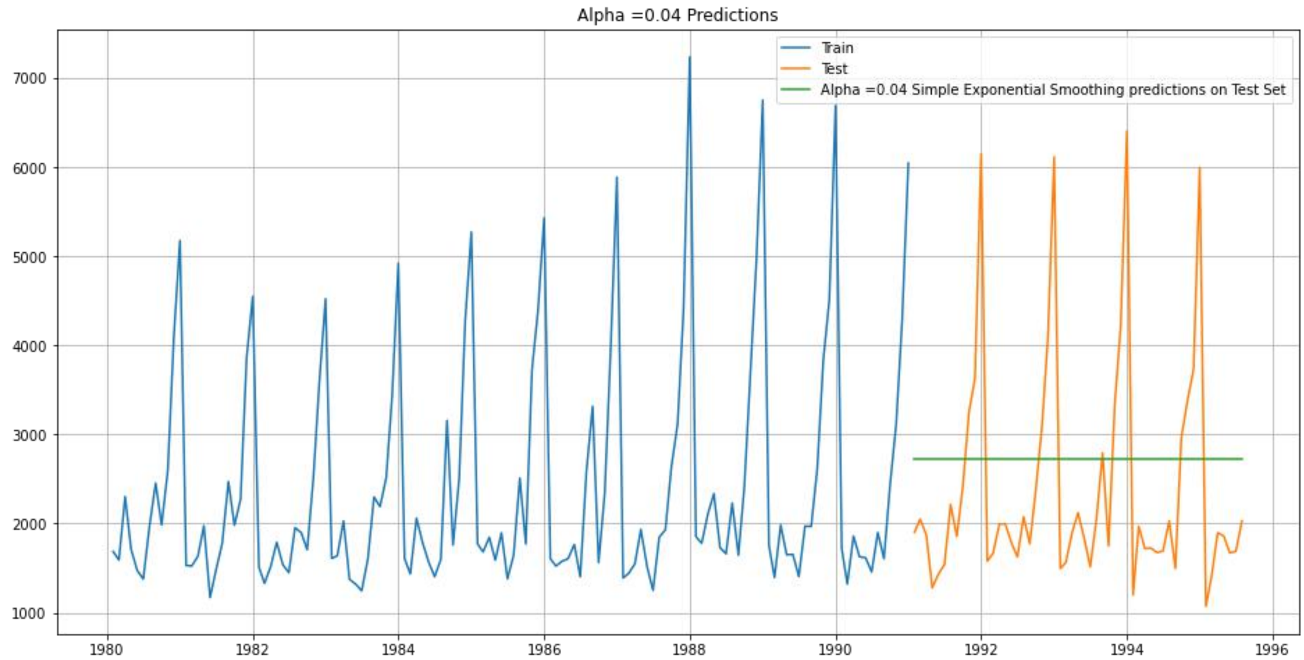

Fig 23

Model Evaluation for alpha = 0.04:

For Alpha =0.04 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 1316.035 and MAPE is 45.47

After SES Tuning:

	Alpha Values	Train RMSE	Test RMSE	Test MAPE
0	0.3	1359.511747	1935.507132	75.66
1	0.4	1352.588879	2311.919615	91.55
2	0.5	1344.004369	2666.351413	106.27
3	0.6	1338.805381	2979.204388	118.77
4	0.7	1338.844308	3249.944092	129.34
5	0.8	1344.462091	3483.801006	138.34
6	0.9	1355.723518	3686.794285	146.08

The RMSE for Alpha = 0.3 was not better than Alpha= 0.04

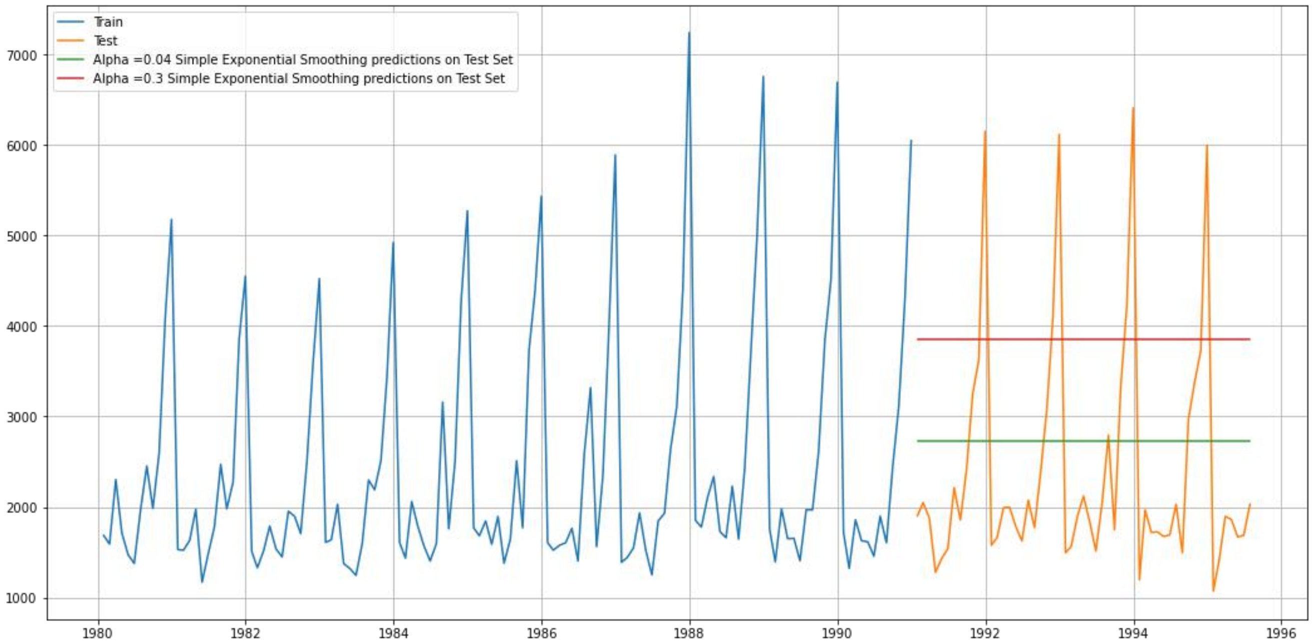

Fig 24

Method 6: Double Exponential Smoothing (Holt's Model)

Best Values after Predictions:

	Alpha Values	Beta Values	Train RMSE	Test RMSE	Test MAPE
0	0.3	0.3	1592.292788	18259.110704	675.28
8	0.4	0.3	1569.338606	23878.496940	886.00
1	0.3	0.4	1682.573828	26069.841401	960.18
16	0.5	0.3	1530.575845	27095.532414	1007.39
24	0.6	0.3	1506.449870	29070.722592	1082.18

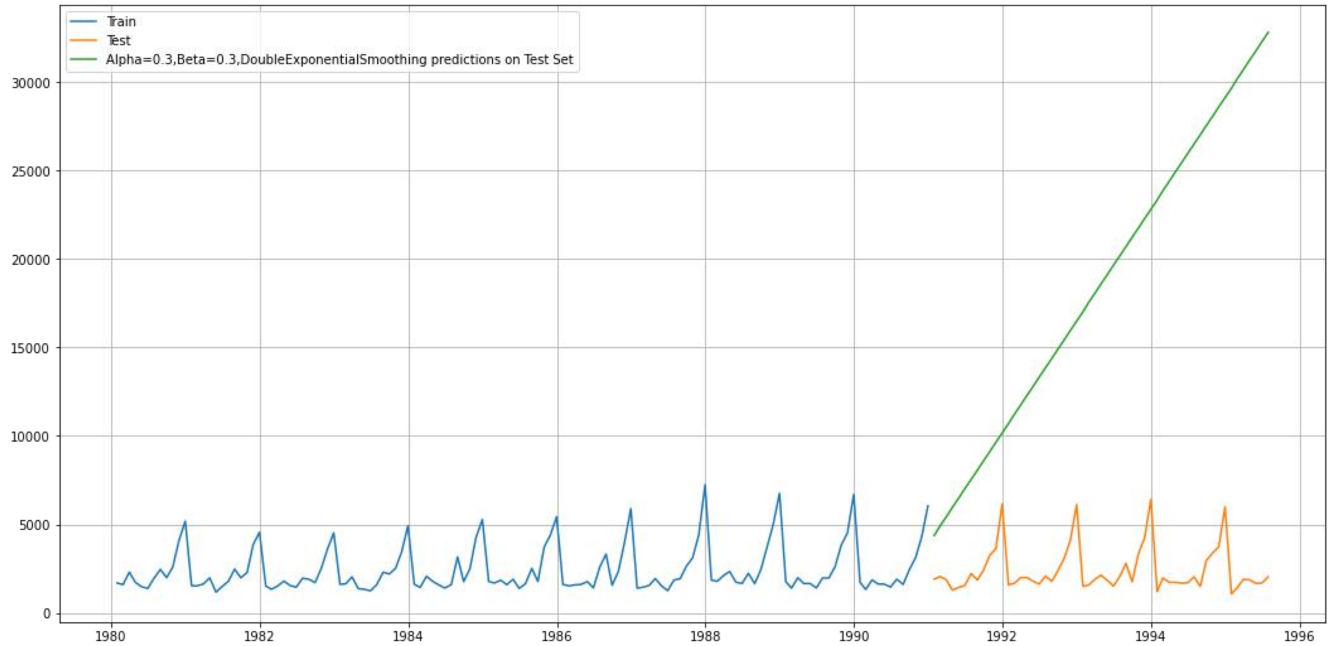

Fig 25

Method 7: Triple Exponential Smoothing (Holt - Winter's Model)

'smoothing_level': 0.11235974440805609,
'smoothing_trend': 0.03742154913668688,
'smoothing_seasonal': 0.4932616459048464

Train set after fitting values:

	Sparkling	auto_predict
Date		
1980-01-31	1686	1682.885034

1980-02-29	1591	1585.152555
1980-03-31	2304	2293.877876
1980-04-30	1712	1702.610588
1980-05-31	1471	1458.609608

Test Set after prediction:

	Sparkling	auto_predict
Date		
1991-01-31	1902	1474.966680
1991-02-28	2049	1169.991432
1991-03-31	1874	1658.920133
1991-04-30	1279	1504.953983
1991-05-31	1432	1417.648032

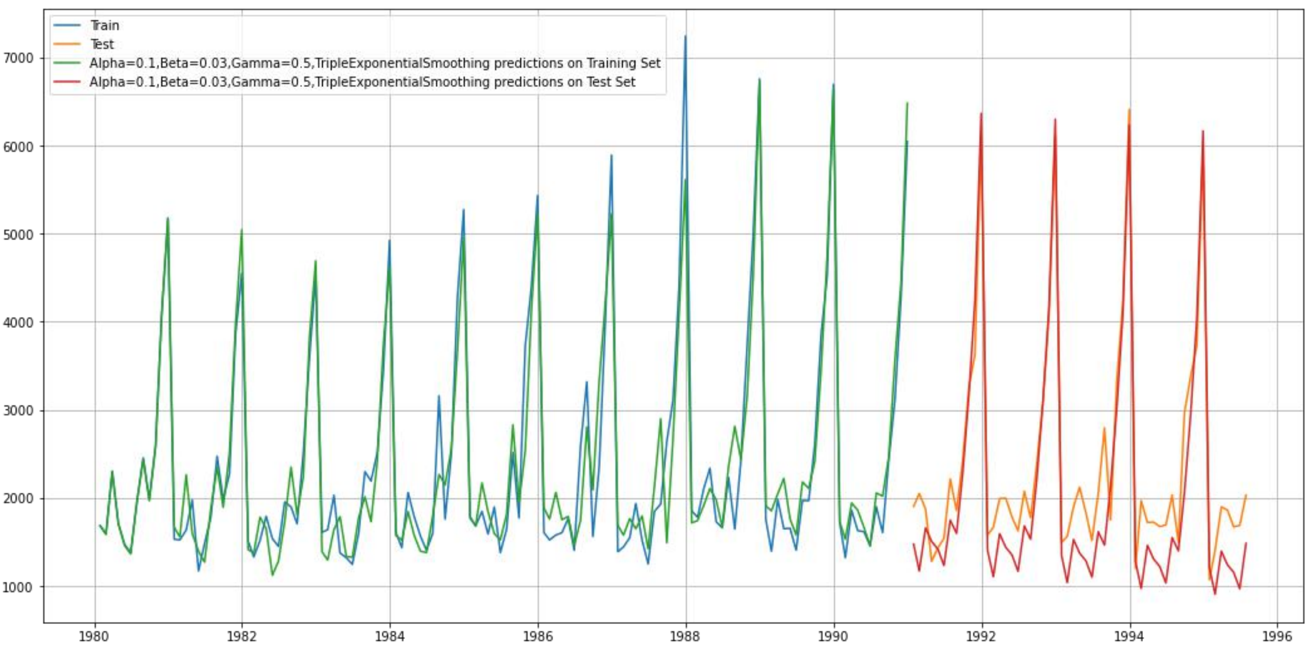

Fig 26

Model Evaluation:

For Alpha: 0.1,Beta: 0.03 and Gamma:0.5, Triple Exponential Smoothing Model forecast on the Training Data, RMSE is 376.279 MAPE is 10.85

For Alpha: 0.1,Beta: 0.03 and Gamma:0.5,Triple Exponential Smoothing Model forecast on the Test Data, RMSE is 473.152 MAPE is 16.53

Model Evaluation after tuning:

Best Params after Tuning:

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE	Test MAPE
3	0.03	0.03	0.33	410.406536	317.553880	9.59
2	0.03	0.03	0.23	426.249806	329.788722	9.98
364	0.33	0.03	0.13	435.742430	327.818278	10.00
607	0.53	0.03	0.23	435.932306	334.132762	10.10
195	0.13	0.63	0.83	448.545611	340.764988	10.34

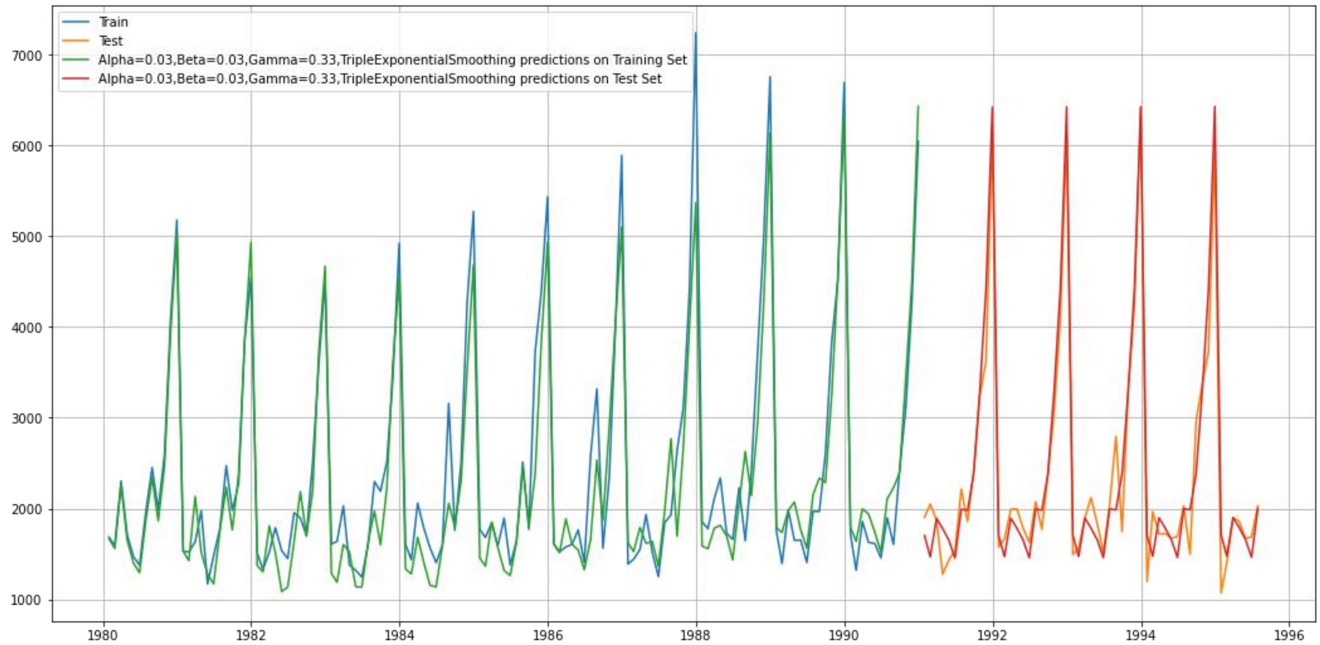

Fig 27

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.

Answer:

The Augmented Dickey-Fuller test is an unit root test which determines whether there is a unit root and subsequently whether the series is non-stationary.

The hypothesis in a simple form for the ADF test is:
• H0 : The Time Series has a unit root and is thus non-stationary.
• H1 : The Time Series does not have a unit root and is thus stationary.

We would want the series to be stationary for building ARIMA models and thus we would want the p-value of this test to be less than the α value.(0.05)

Check for stationarity of the Whole Data Time Series:

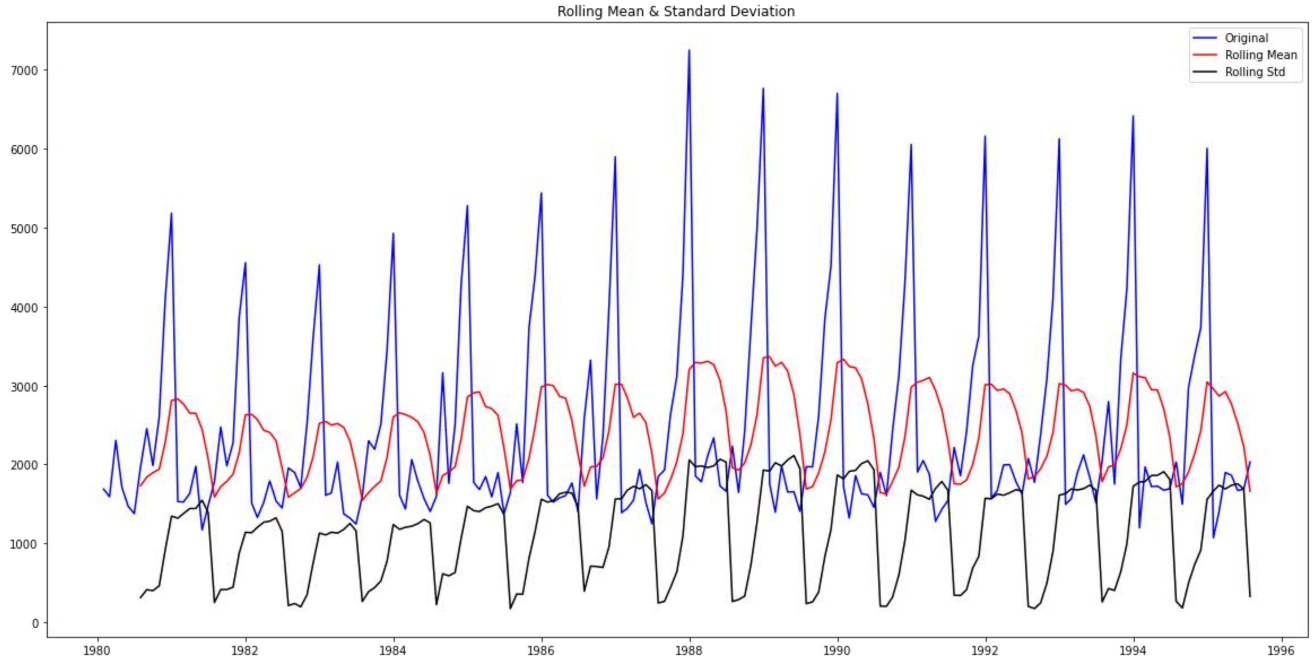

Fig 28

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

We see that at 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:

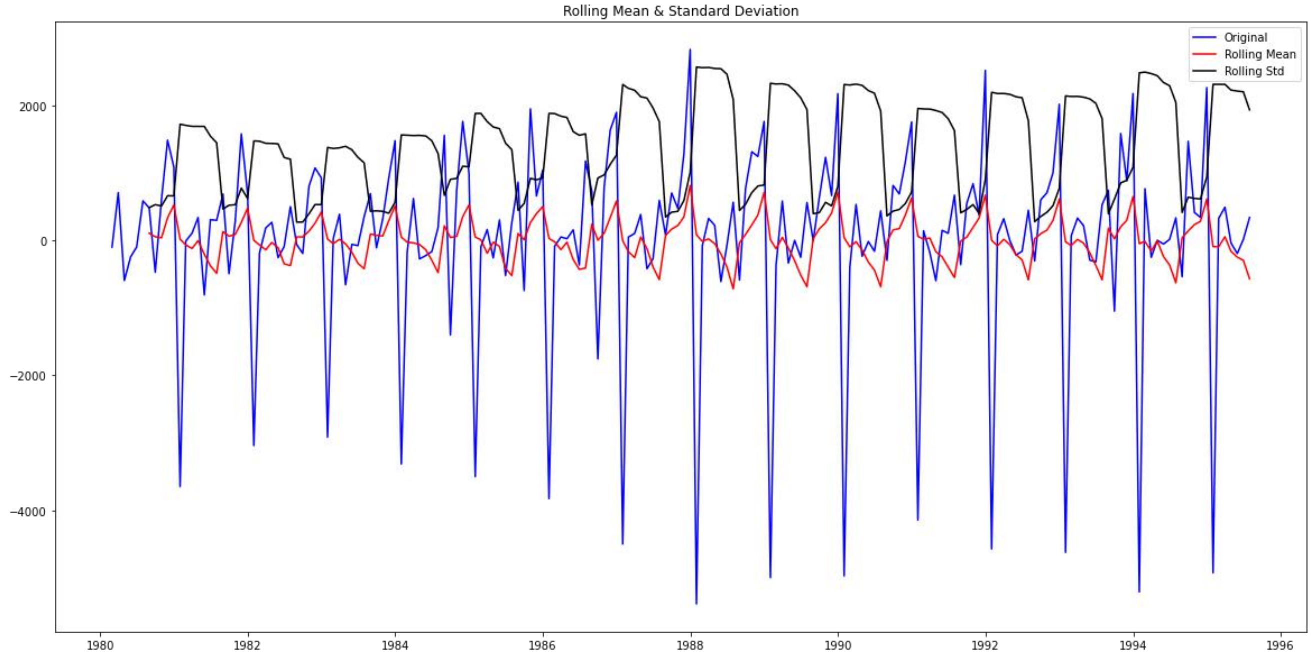

Fig 29

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

We see that at alpha = 0.05 when taking difference of order 1 the Time Series is indeed stationary.

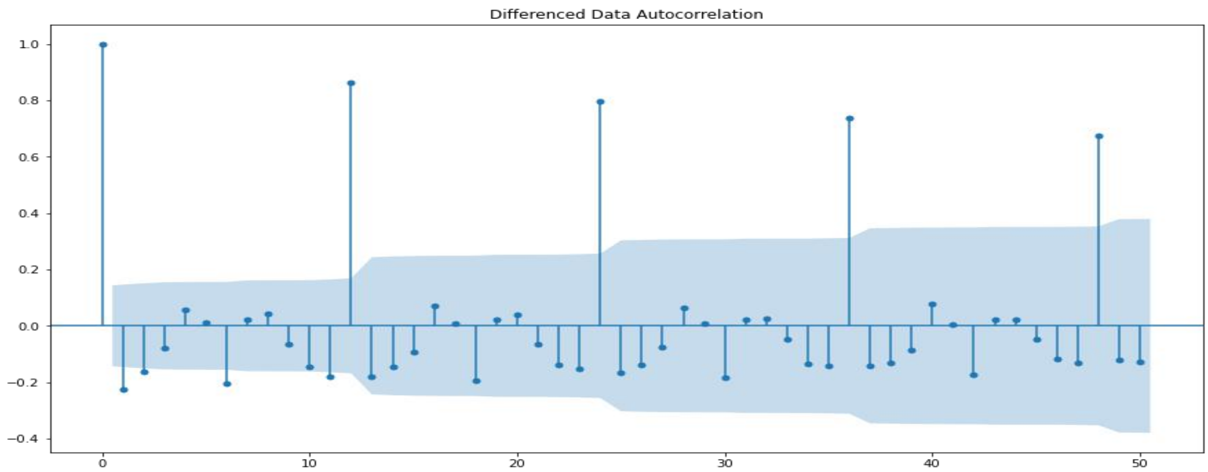

Fig 30

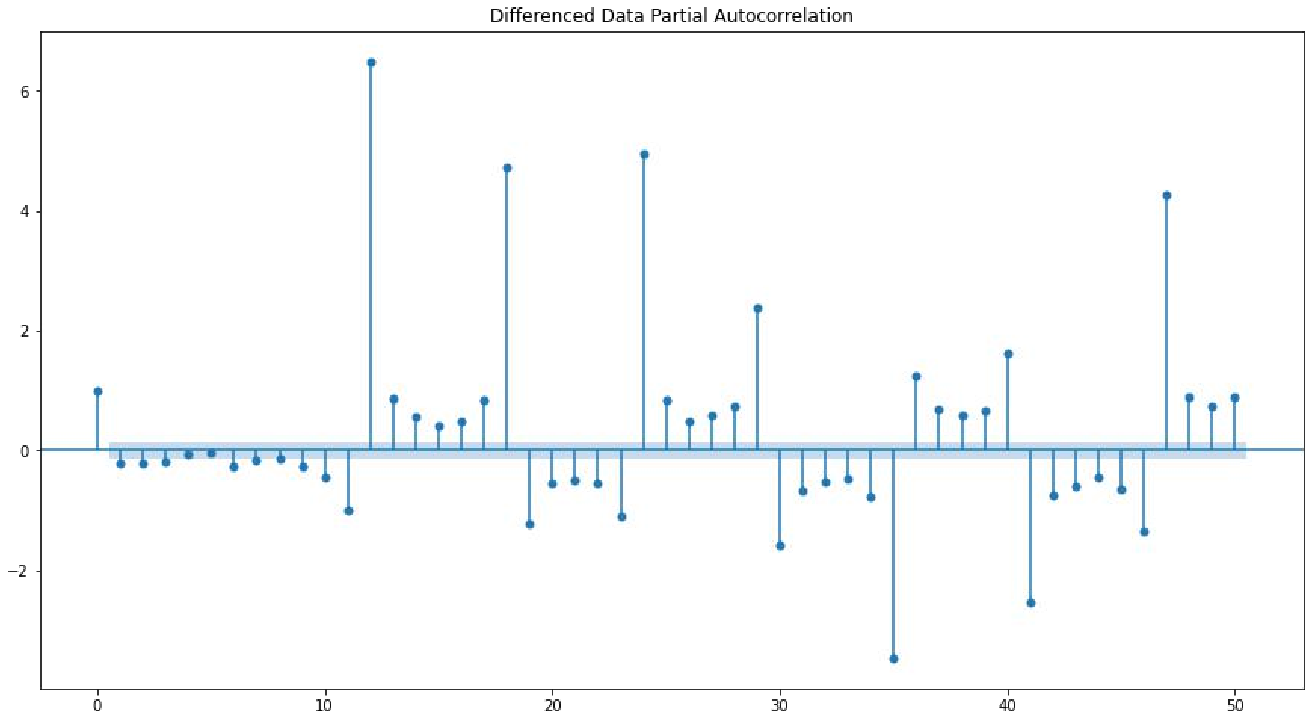
Differenced Data Partial Autocorrelation
6
4
2
0
−2
0
10
20
30
40
50

Fig 31

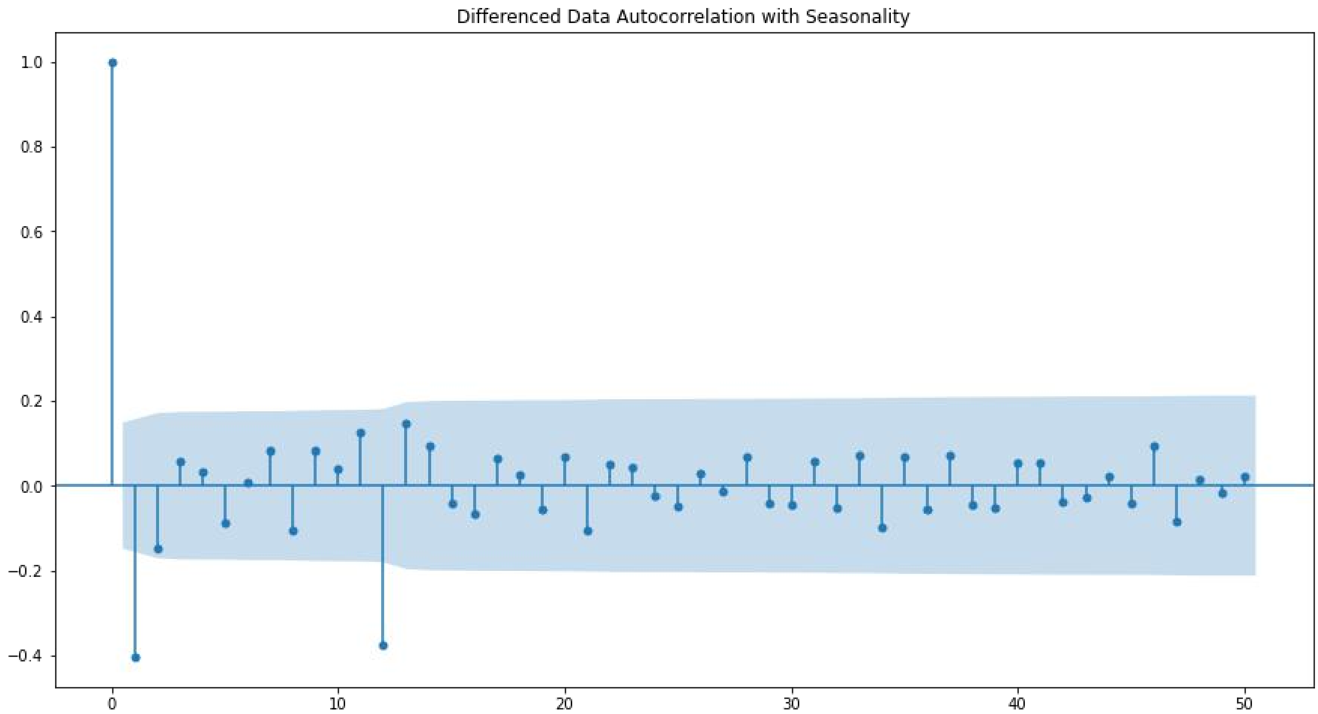
Differenced Data Autocorrelation with Seasonality
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
10
20
30
40
50

Fig 32

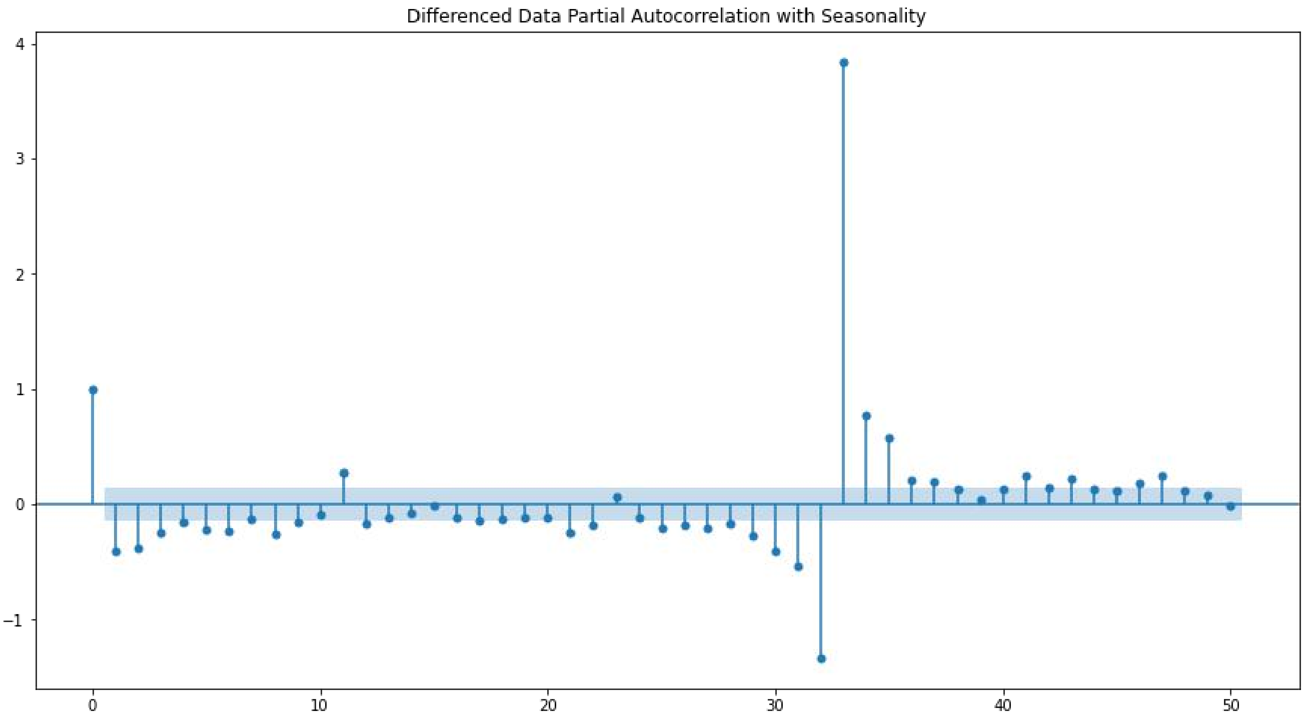

Fig 33

We observe the ACF plot for Sparkling Sales and observe seasonality at intervals 12, hence we run the Automated SARIMA models at seasonality 12.

When we build manual ARIMA model for Sparkling Sales based on the ACF and PACF plots. Hence we chose the AR parameter p = 3 and P = 1, Moving average parameter q = 2 and Q = 0 and d =1 and D= 1 based on the plots.

Check for stationarity of the Training Data Time Series:

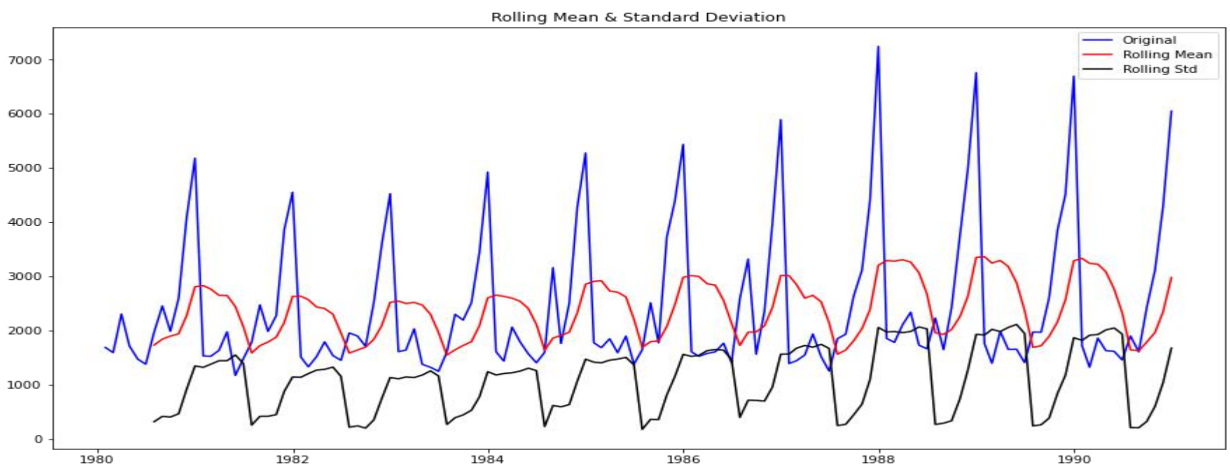

Fig 34

Results of Dickey-Fuller Test:

Test Statistic	-1.208926
p-value	0.669744
#Lags Used	12.000000
Number of Observations Used	119.000000
Critical Value (1%)	-3.486535
Critical Value (5%)	-2.886151
Critical Value (10%)	-2.579896

We see that the Train series is not stationary at alpha = 0.05.

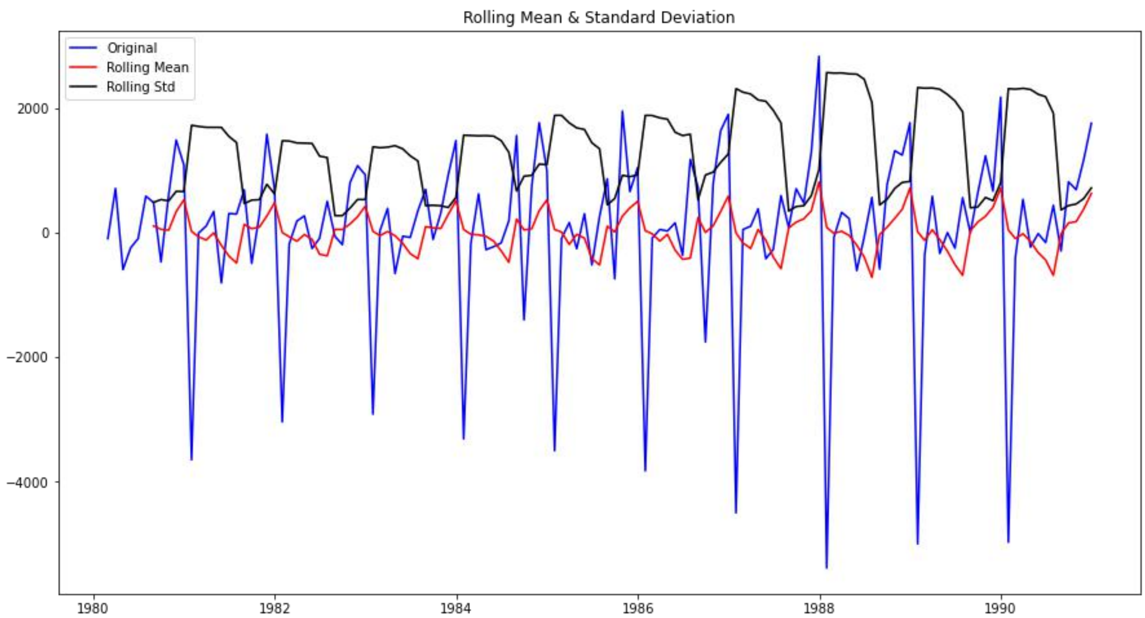

Fig 35

Results of Dickey-Fuller Test:

Test Statistic	-8.005007e+00
p-value	2.280104e-12
#Lags Used	1.100000e+01
Number of Observations Used	1.190000e+02
Critical Value (1%)	-3.486535e+00
Critical Value (5%)	-2.886151e+00
Critical Value (10%)	-2.579896e+00

We see that after taking a difference of order 1 the series have become stationary at alpha = 0.05

Note: If the series is non-stationary, stationarize the Time Series by taking a difference of the Time Series. Then we can use this particular differenced series to train the ARIMA models. We do not need to worry about stationarity for the Test Data because we are not building any models on the Test Data, we are evaluating our models over there

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.

Answer:

1. Automated version of ARIMA for which the best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC).

Some parameter combinations for the Model...
Model: (0, 1, 1)
Model: (0, 1, 2)
Model: (0, 1, 3)
Model: (0, 1, 4)
Model: (1, 1, 0)
Model: (1, 1, 1)
Model: (1, 1, 2)
Model: (1, 1, 3)
Model: (1, 1, 4)
Model: (2, 1, 0)
Model: (2, 1, 1)
Model: (2, 1, 2)
Model: (2, 1, 3)
Model: (2, 1, 4)
Model: (3, 1, 0)
Model: (3, 1, 1)
Model: (3, 1, 2)
Model: (3, 1, 3)
Model: (3, 1, 4)
Model: (4, 1, 0)
Model: (4, 1, 1)
Model: (4, 1, 2)
Model: (4, 1, 3)
Model: (4, 1, 4)

Best ARIMA Params by sorting lowest AIC to top:

	param	AIC
11	(2, 1, 2)	2210.617093
13	(2, 1, 4)	2220.220499
17	(3, 1, 3)	2225.661559
18	(3, 1, 4)	2226.054856
22	(4, 1, 3)	2226.954554

ARIMA SUMMARY:

```
                             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           1012.061
Date:                Sun, 20 Mar 2022   AIC                           2210.617
Time:                        22:02:22   BIC                           2227.868
Sample:                    02-29-1980   HQIC                          2217.627
                         - 12-31-1990
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 5.5859      0.516     10.818      0.000       4.574       6.598
ar.L1.D.Sparkling     1.2699      0.074     17.047      0.000       1.124       1.416
ar.L2.D.Sparkling    -0.5601      0.074     -7.617      0.000      -0.704      -0.416
ma.L1.D.Sparkling    -1.9991      0.042    -47.179      0.000      -2.082      -1.916
ma.L2.D.Sparkling     0.9991      0.042     23.594      0.000       0.916       1.082
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.1337           -0.7073j            1.3362           -0.0888
AR.2            1.1337           +0.7073j            1.3362            0.0888
MA.1            1.0001           +0.0000j            1.0001            0.0000
MA.2            1.0008           +0.0000j            1.0008            0.0000
-----------------------------------------------------------------------------
```

Test rmse for arima is 1375.0019866285338
Test mape for arima is 48.39

2. Automated version of a SARIMA model for which the best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC):

We observe the ACF plot for Sparkling Sales and observe seasonality at intervals 12, hence we run the Automated SARIMA models at seasonality 12.

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)

Best SARIMA Params by sorting lowest AIC to top:

	param	seasonal	AIC
50	(1, 1, 2)	(1, 0, 2, 12)	1555.584247
53	(1, 1, 2)	(2, 0, 2, 12)	1555.934563
26	(0, 1, 2)	(2, 0, 2, 12)	1557.121570
23	(0, 1, 2)	(1, 0, 2, 12)	1557.160507
77	(2, 1, 2)	(1, 0, 2, 12)	1557.340402

SARIMA SUMMARY:

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(1, 1, 2)x(1, 0, 2, 12)   Log Likelihood                -770.792
Date:                            Sun, 20 Mar 2022   AIC                           1555.584
Time:                                    22:31:43   BIC                           1574.095
Sample:                                         0   HQIC                          1563.083
                                            - 132
Covariance Type:                              opg
==================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
```

ar.L1	-0.6282	0.255	-2.463	0.014	-1.128	-0.128
ma.L1	-0.1041	0.225	-0.463	0.643	-0.545	0.337
ma.L2	-0.7276	0.154	-4.735	0.000	-1.029	-0.426
ar.S.L12	1.0439	0.014	72.836	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

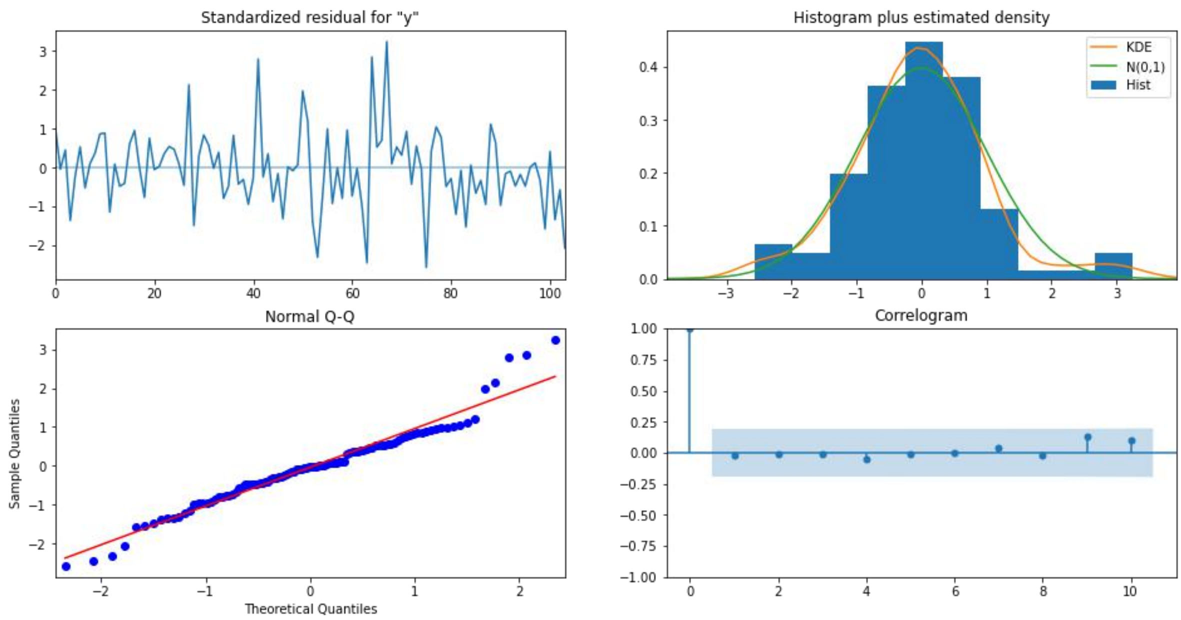

Fig 36

Inference from Model diagnostics confirms that:

- the model residuals are normally distributed Standardized residual – Do not display any obvious seasonality
- Histogram plus estimated density - The KDE plot of
- the residuals is similar with the normal distribution, hence the model residuals are normally distributed based
- Normal Q-Q plot – There is an ordered distribution of residuals (blue dots) following the linear trend of the samples taken from a standard normal distribution with N(0, 1)
- Correlogram – The time series residuals have low correlation with lagged versions of itself.

Test rmse for SARIMA is 528.6041848904955

Test mape for SARIMA is 18.89

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.

<u>Answer:</u>

1. Manual ARIMA Model

When we build manual ARIMA model for Sparkling Sales based on the ACF and PACF plots. Hence we chose the AR parameter p = 3, Moving average parameter q = 2 and d =1 based on the ACF/PACF plots.

```
ARIMA Model Results
==============================================================================
Dep. Variable:            D.Sparkling    No. Observations:                  131
Model:                 ARIMA(3, 1, 2)    Log Likelihood               -1107.464
Method:                       css-mle    S.D. of innovations           1106.107
Date:                Sun, 20 Mar 2022    AIC                           2228.927
Time:                        22:31:49    BIC                           2249.054
Sample:                    02-29-1980    HQIC                          2237.106
                         - 12-31-1990
=================================================================================
===
                        coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------
-----
const                 5.9849      3.643      1.643      0.100      -1.156      13.126
ar.L1.D.Sparkling    -0.4420   5.43e-06  -8.14e+04      0.000      -0.442      -0.442
ar.L2.D.Sparkling     0.3079   1.31e-05   2.34e+04      0.000       0.308       0.308
ar.L3.D.Sparkling    -0.2501   1.13e-05  -2.21e+04      0.000      -0.250      -0.250
ma.L1.D.Sparkling    -0.0005      0.021     -0.026      0.979      -0.041       0.040
ma.L2.D.Sparkling    -0.9995      0.021    -48.712      0.000      -1.040      -0.959
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1           -1.0000           -0.0000j            1.0000           -0.5000
AR.2            1.1156           -1.6594j            1.9996           -0.1558
AR.3            1.1156           +1.6594j            1.9996            0.1558
MA.1            1.0000           +0.0000j            1.0000            0.0000
MA.2           -1.0005           +0.0000j            1.0005            0.5000
-----------------------------------------------------------------------------
```

Test rmse for Manual arima is 1378.9863774088376
Test mape for Manual arima is 49.31

2. Manual SARIMA Model

We observe the ACF plot for Sparkling Sales and observe seasonality at intervals 12, hence we run the SARIMA models at seasonality 12.

When we build manual SARIMA model for Sparkling Sales based on the ACF and PACF plots. Hence we chose the AR parameter p = 3 and P = 1, Moving average parameter q = 2 and Q = 0 and d =1 and D= 1 based on the plots.

```
SARIMAX Results
==============================================================================
========
Dep. Variable:                          y   No. Observations:           132
Model:      SARIMAX(1, 1, 2)x(1, 0, 2, 12)   Log Likelihood          -770.792
Date:                     Sun, 20 Mar 2022   AIC                     1555.584
Time:                             22:31:53   BIC                     1574.095
Sample:                                  0   HQIC                    1563.083
                                     - 132
Covariance Type:                       opg
==============================================================================
              coef    std err        z      P>|z|     [0.025     0.975]
------------------------------------------------------------------------------
ar.L1      -0.6282     0.255    -2.463     0.014     -1.128     -0.128
ma.L1      -0.1041     0.225    -0.463     0.643     -0.545      0.337
ma.L2      -0.7276     0.154    -4.735     0.000     -1.029     -0.426
ar.S.L12    1.0439     0.014    72.836     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
```

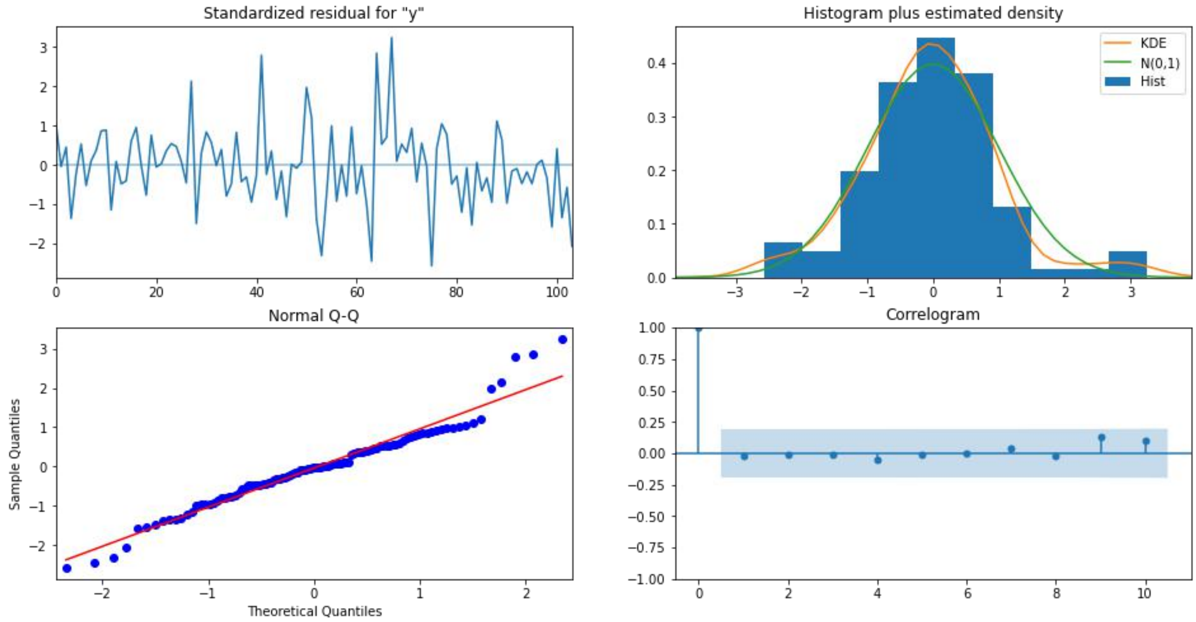

Fig 37

Model diagnostics confirms that the model residuals are normal y distributed. Standardized residual do not display any obviouss easonality,
Histogram plus estimated density - The KDE plot has normal distribution ,
Normal Q-Q plot – There is an ordered distribution of residual s (blue dots) following the linear trend ,
Correlogram – The time series residuals have low correlation with lagged versions of itself

Test rmse for Manual sarima is 528.6041848904955
Test mape for Manual sarima is 18.89

I built various models by tweaking the parameters by looking at the ACF and PACF plots of which I showed the best mode l in the report.

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.

Answer:

Sorted by RMSE values on the Test Data:

	Test RMSE	Test MAPE
Alpha=0.03,Beta=0.03,Gamma=0.33,TripleExponent	317.553880	9.59
Alpha=0.1,Beta=0.03,Gamma=0.5,TripleExponent	473.152417	16.53
SARIMA(1, 1, 2)(1, 0, 2, 12)	528.604185	18.89
SARIMA(1, 1, 2)(1, 0, 0, 12)	528.604185	18.89
2pointTrailingMovingAverage	813.400684	19.70
4pointTrailingMovingAverage	1156.589694	35.96
SimpleAverageModel	1275.081804	38.90
6pointTrailingMovingAverage	1283.927428	43.86
Alpha =0.04 Simple Exponential Smoothing Model	1316.035487	45.47
9pointTrailingMovingAverage	1346.278315	46.86
ARIMA (2,1,2)	1375.001987	48.39
ARIMA(3,1,2)	1378.986377	49.31
RegressionOnTime	1389.135175	50.15
NaiveModel	3864.279352	152.87
NaiveModel	3864.279352	152.87
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	18259.110704	675.28

Sorted by MAPE values on the Test Data:

	Test RMSE	Test MAPE
Alpha=0.03,Beta=0.03,Gamma=0.33,TripleExponenti...	317.553880	9.59
Alpha=0.1,Beta=0.03,Gamma=0.5,TripleExponential...	473.152417	16.53
SARIMA(1, 1, 2)(1, 0, 2, 12)	528.604185	18.89
SARIMA(1, 1, 2)(1, 0, 0, 12)	528.604185	18.89
2pointTrailingMovingAverage	813.400684	19.70
4pointTrailingMovingAverage	1156.589694	35.96
SimpleAverageModel	1275.081804	38.90
6pointTrailingMovingAverage	1283.927428	43.86
Alpha =0.04 Simple Exponential Smoothing Model	1316.035487	45.47
9pointTrailingMovingAverage	1346.278315	46.86
ARIMA (2,1,2)	1375.001987	48.39
ARIMA(3,1,2)	1378.986377	49.31
RegressionOnTime	1389.135175	50.15
NaiveModel	3864.279352	152.87
NaiveModel	3864.279352	152.87
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	18259.110704	675.28

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

Answer:

We see that the best model is the Triple Exponential Smoothin g with additive seasonality with the parameters alpha = 0.03, beta = 0.03 and gamma = 0.33.

RMSE when Model run on full dataset: 364.0751576684798
MAPE when Model run on full dataset : 10.93

Predictions on 12 months into future:

1995-08-31	1938.336822
1995-09-30	2419.476144
1995-10-31	3280.986121
1995-11-30	3998.450523
1995-12-31	6164.402641
1996-01-31	1336.296165
1996-02-29	1621.453894
1996-03-31	1856.928998
1996-04-30	1826.032100
1996-05-31	1677.731338
1996-06-30	1605.037257
1996-07-31	2031.547055

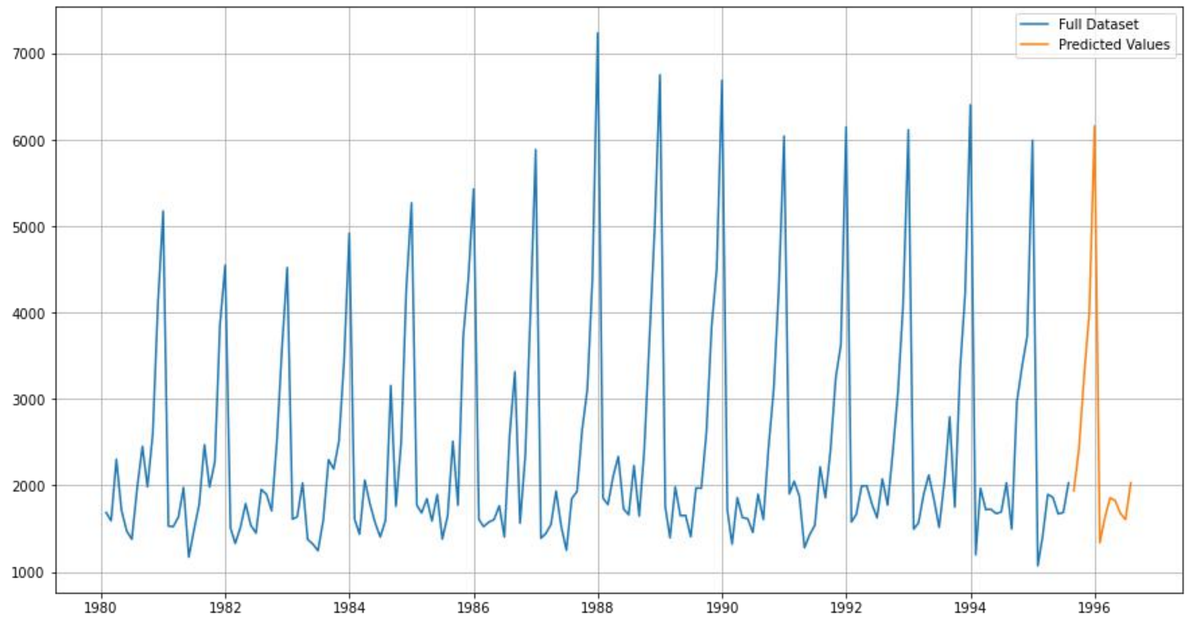

Fig 38

I have calculated the upper and lower confidence bands at 9
5% confidence level.
The percentile function under numpy lets us calculate these an
d adding and subtracting from the predictions gives us the nec
essary confidence bands for the predictions.

	lower_CI	prediction	upper_ci
1995-08-31	528.136857	1938.336822	7507.102631
1995-09-30	1009.276179	2419.476144	7988.241953
1995-10-31	1870.786156	3280.986121	8849.751930
1995-11-30	2588.250558	3998.450523	9567.216332
1995-12-31	4754.202676	6164.402641	11733.168450

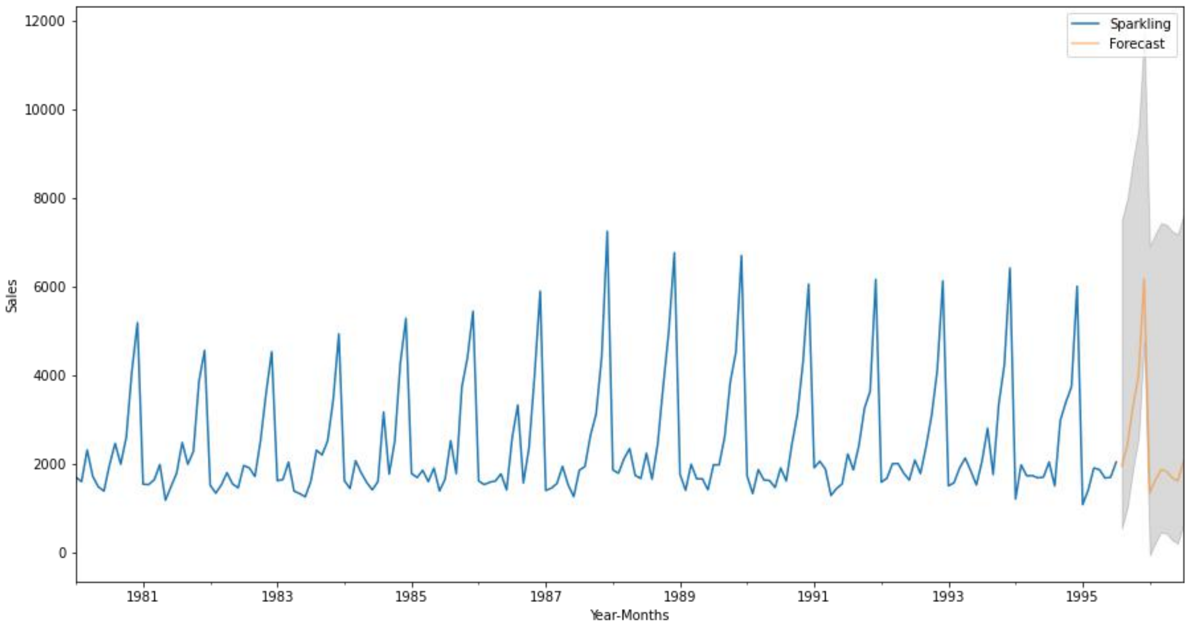
Sparkling
Forecast
12000
10000
8000
6000
4000
2000
0
Sales
1981
1983
1985
1987
1989
1991
1993
1995
Year-Months

Fig 39

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

Answer:

1. Triple Exponential Model is performing best in this case giving us the least error.

2. Looking at the bar plot, we can see that on December months the sales are highest. We can use this insights to increase our sales further.

3. We can introduce certain offers in November, December months to attaract more customers.

4. On Saturdays mean sales of the wine is highest. We can give certain offers to attract more customers.

5. Year 1988 has the highest sales recorded till data. We can go back to find out the reasons to which pushed the sales so much.

6. We can also see in the year 1981, 1983 and 1994 the Wine sales in the month of October, November remained constant after that it has starting fluctuating needs to pay attention after that.

7.Looking at the prediction, we can say that the sales figure will be more or less same as that of previous year. Hence some important measures have to be taken to increase the trend. As the trend has been more or less constant through out the years.

8. Both the models are built considering the Trend and Seasonality in to account and we see from the output plot that the future prediction is in line with the trend and seasonality in the previous years.

9.The company should use the prediction results and capitalize on the high demand seasons and ensure to
source and supply the high demand

10.The company should use the prediction results to plan the l ow demand seasons to stock as per the
demand.

11. The price of rose wine may be expensive than sparkling s o seasonal discounts can help improve the sales of rose wine. Products that are discounted should be highlighted so consumer s can see the savings prominently. Discounts can compel consu mers to buy.

12.As we know how the seasonality is in the prediction compa ny cannot have the same stock through the year. You should c reate a dynamic consumer experience with fresh point-of-sale materials and wellstocked displays. Displays need to look fresh and interesting and tell a compelling story about why the consumer should purchase the product.

12. Seasonal memberships and discounts can be introduced. Co nsumers get very excited about savings and appreciate discount s being passed on. Many prominent retailers also have loyalty programs or club member cards that create excitement. A club-member price brings consumers back and improve sales.

13. Events and tastings help draw consumers to your store and generate sales. Retailers with economies of scale successfully sample consumers on more profitable wines. Some even compa rison-taste customers on national brands that are more expensiv eto demonstrate they are offering a less expensive but superior product.

14.And bringing in celebrities, sommeliers or trade reps for tast ings can help create excitement and drive
traffic