TIME SERIES FORECASTING PROJECT BUSINESS REPORT

Date – 25/12/2022

Table of contents

Contents

List of Figures:

Contents

Problem1-Sparkling:

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 Description:

1. YearMonth : Date and year
2. Sparkling : Sale of Sparkling wine

Sample of the dataset:

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

We have Converted column YearMonth into index.

Exploratory Data Analysis:

Let us check the types of variables in the data frame.

As we have changed YearMonth column into index so only one column present in the data which is the 'Sparkling' and data type of this column is integer.

Checking for missing values in the dataset:

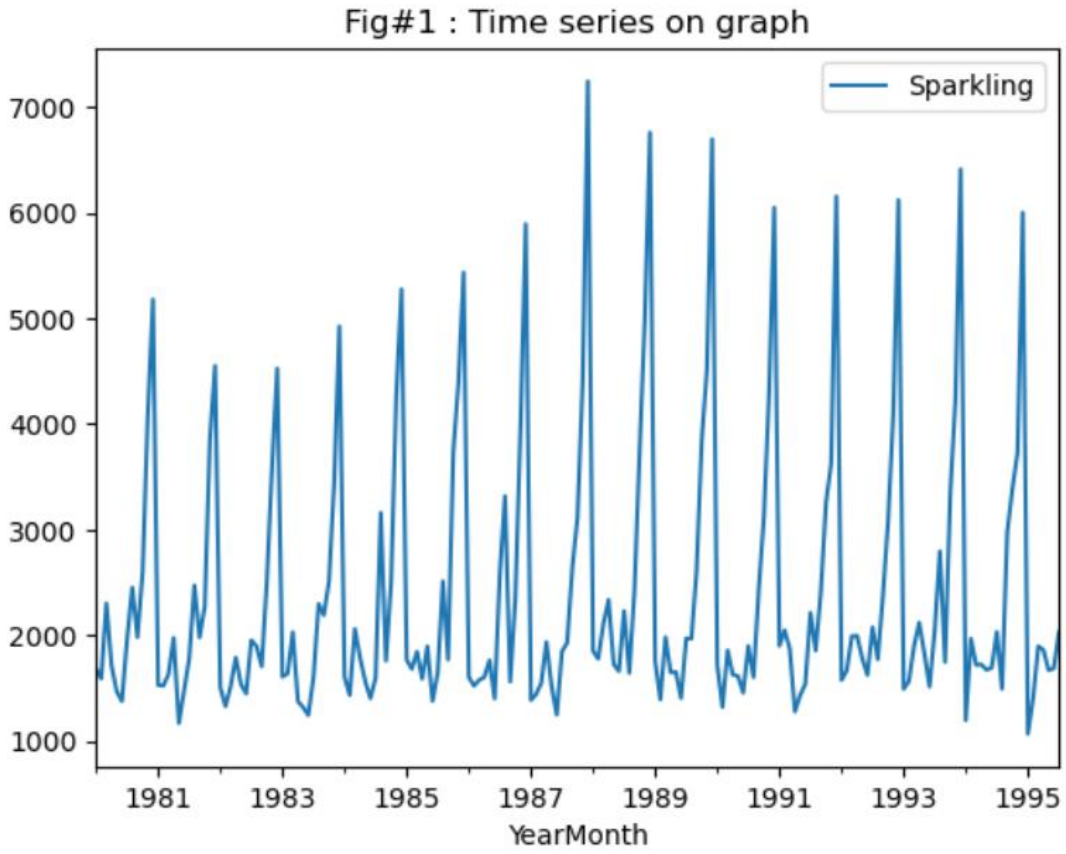
```
Data columns (total 1 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Sparkling  187 non-null    int64
dtypes: int64(1)
memory usage: 2.9 KB
```

From the above results we can see that there is no missing value present in the dataset.

Q1 Read the data as an appropriate Time Series data and plot the data.

Sample of the time series data.

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

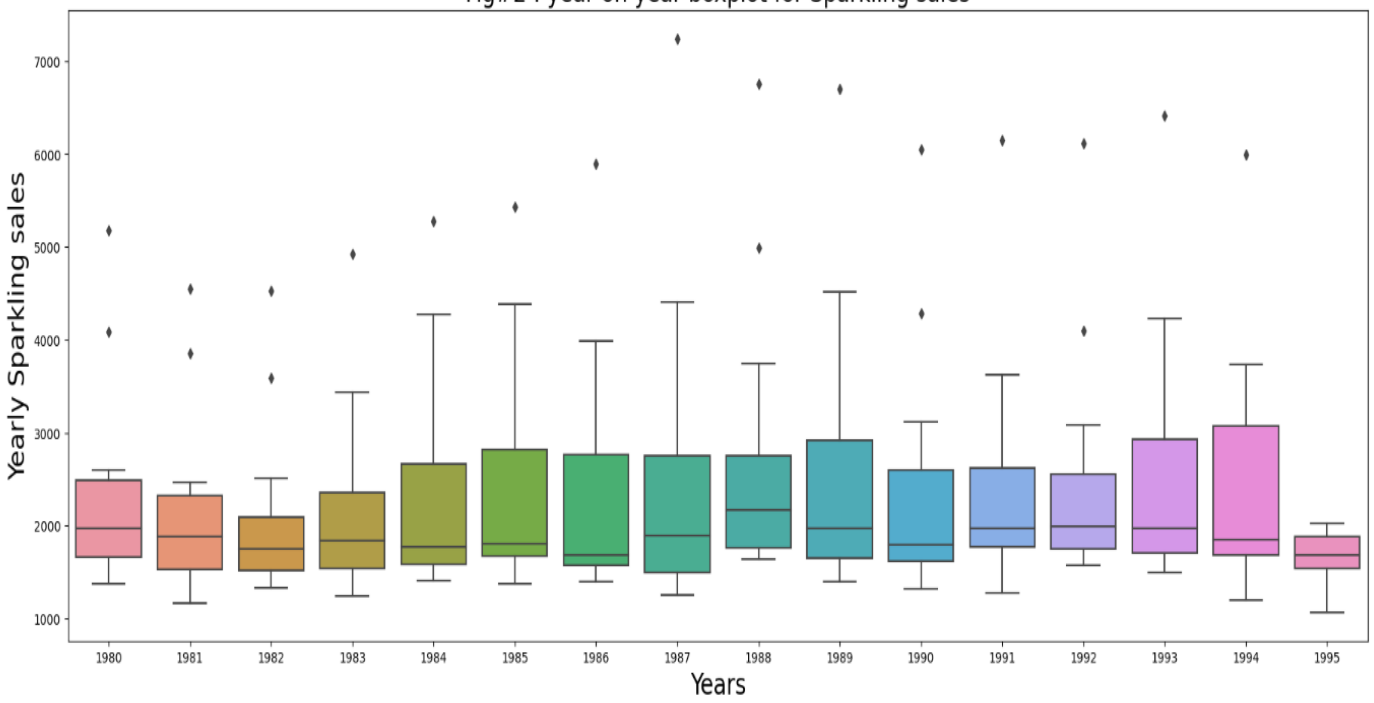


As per the graph Sparkling wine sales hit maximum numbers in year 1988.

Q2 Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

We have built different plot with different time combination.

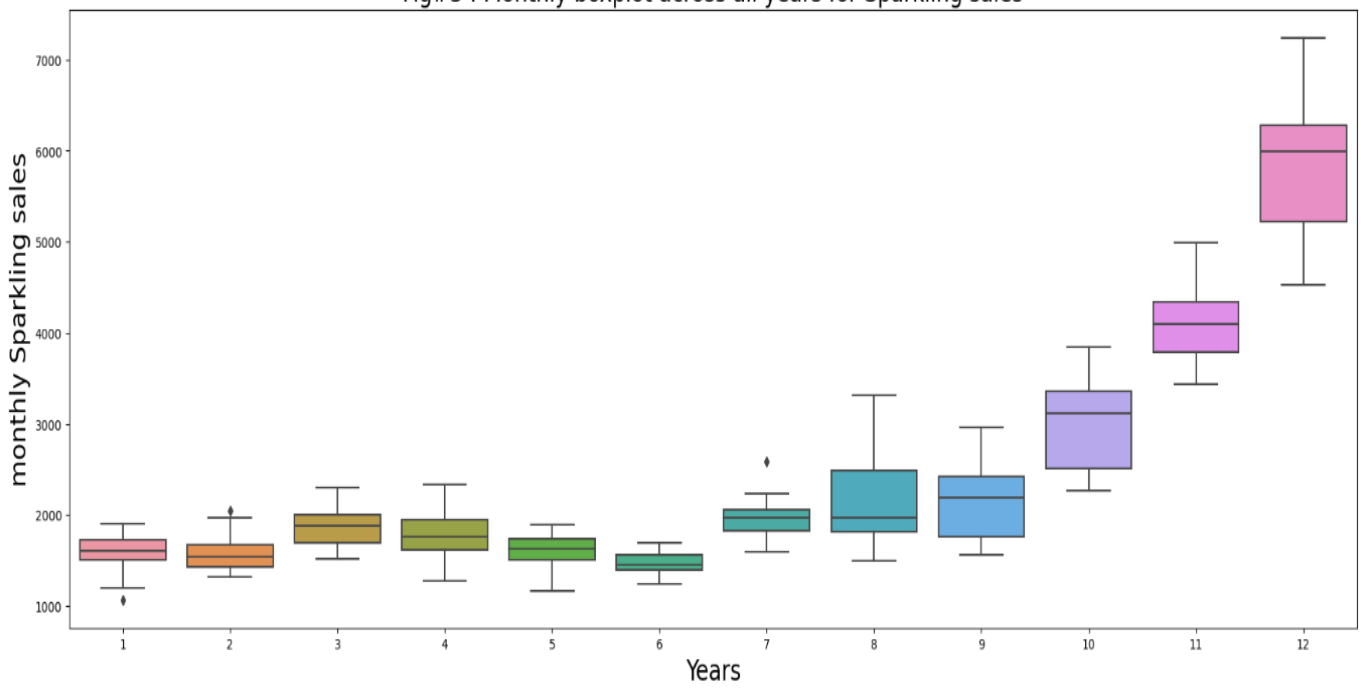
Fig#2 : year on year boxplot for Sparkling sales



~ We can see that we have data from past 16 years.

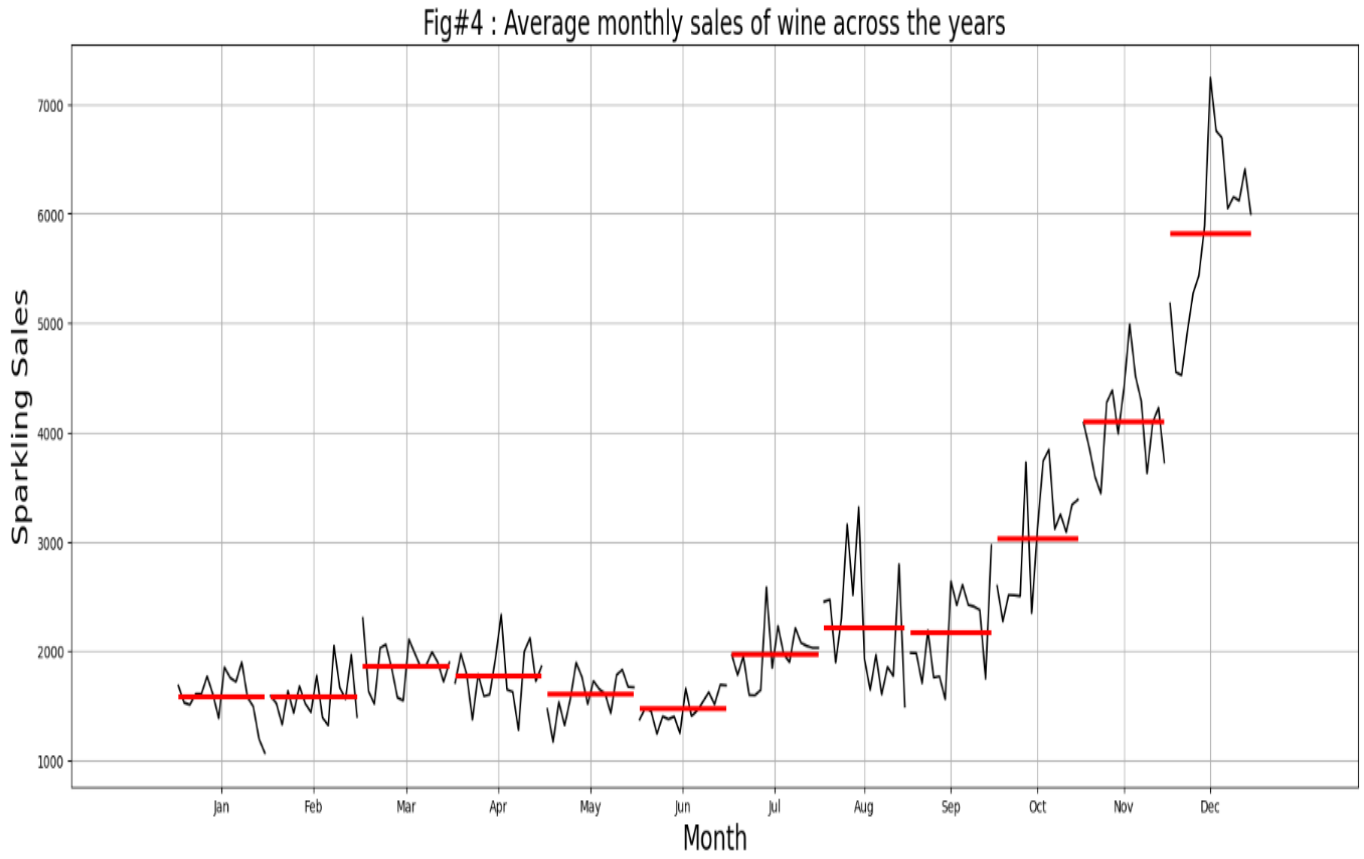
~ As we got to know from the Time Series plot, the boxplots over here indicate no such trend present. Also, we see that the sales of wine have some outliers for most of years.

Fig#3 : Monthly boxplot across all years for Sparkling sales

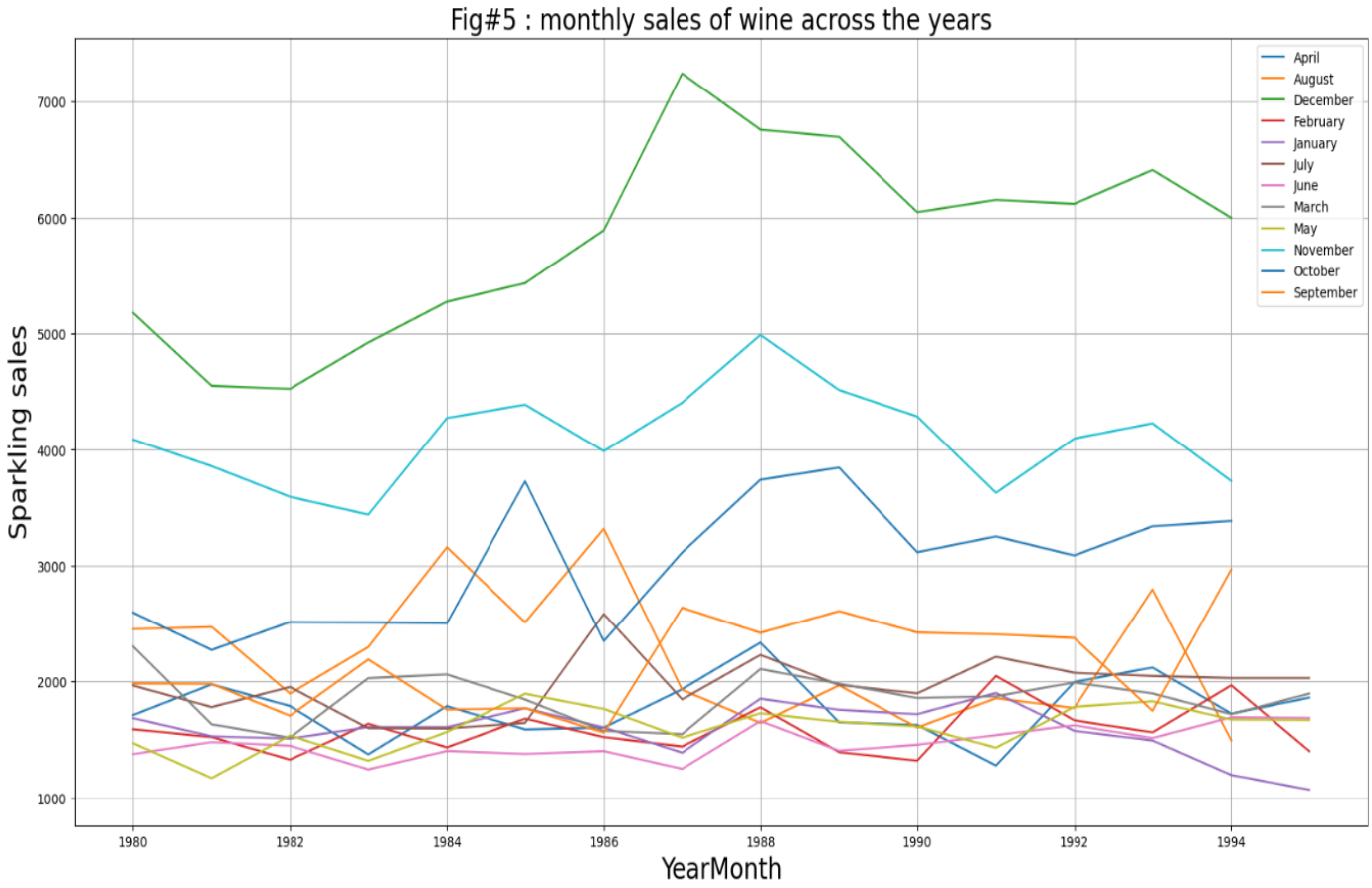


~ The boxplots for the monthly production for different years have outliers.

~ We can see that sales of Sparkling wine increases at the end of the year.

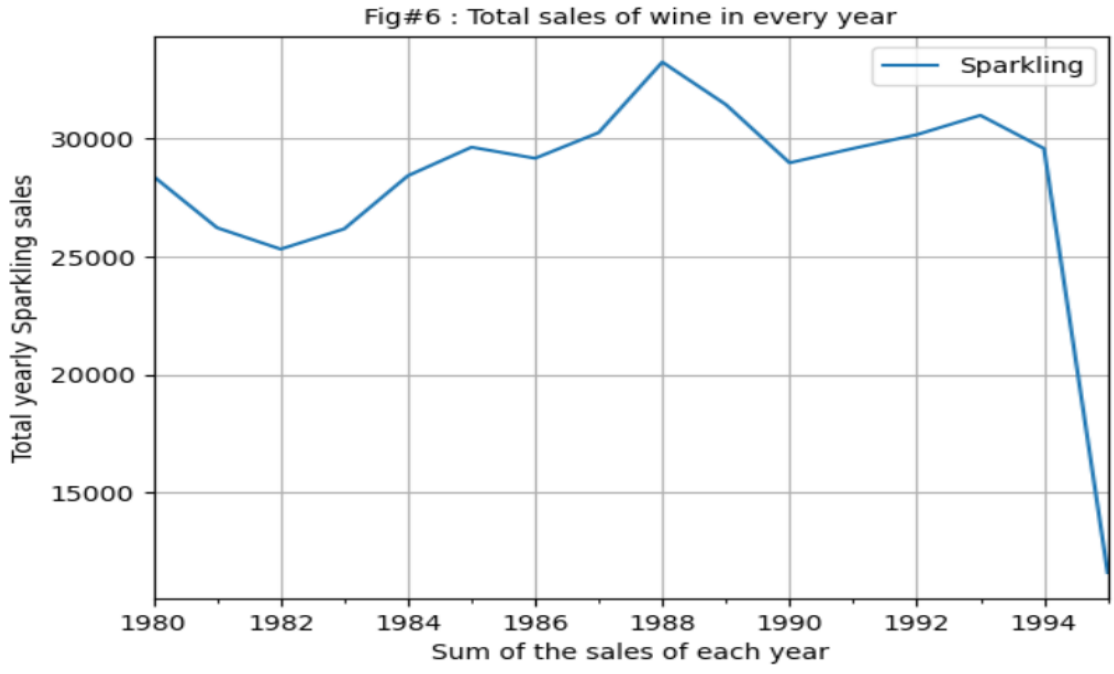


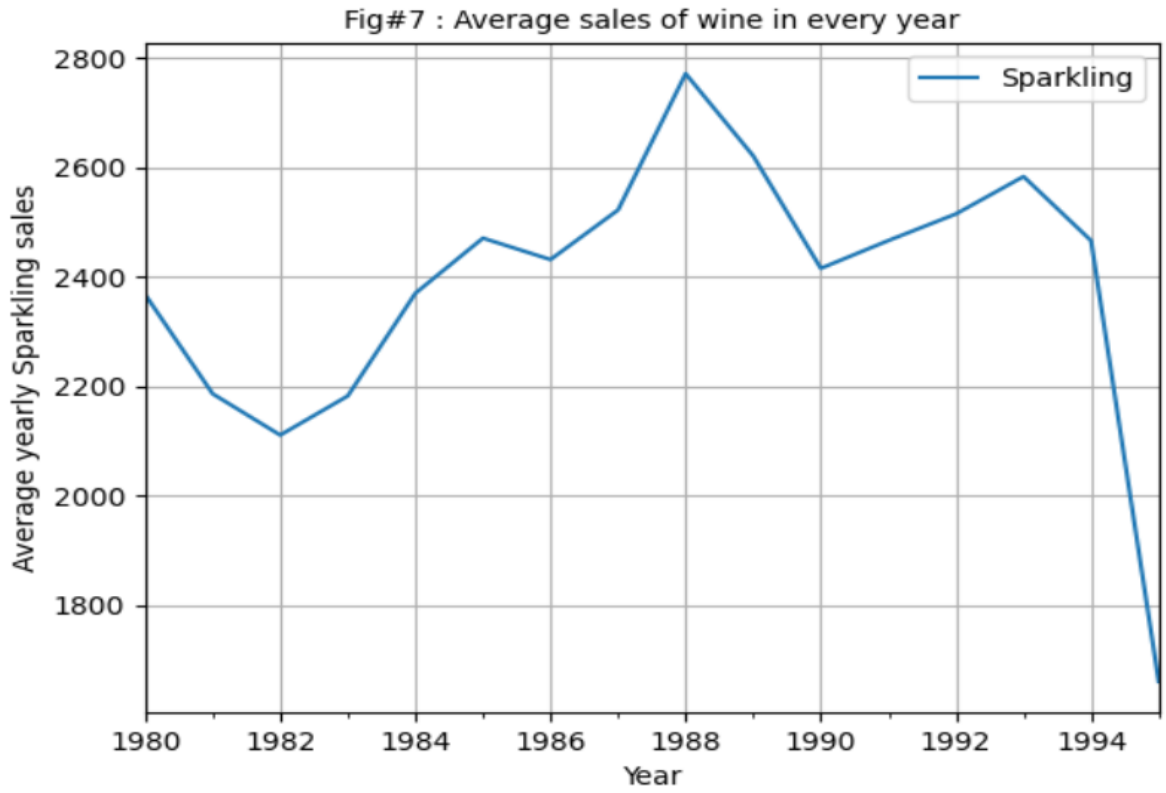
~ Here the red line shows the average wine sales for all the years in a particular month.



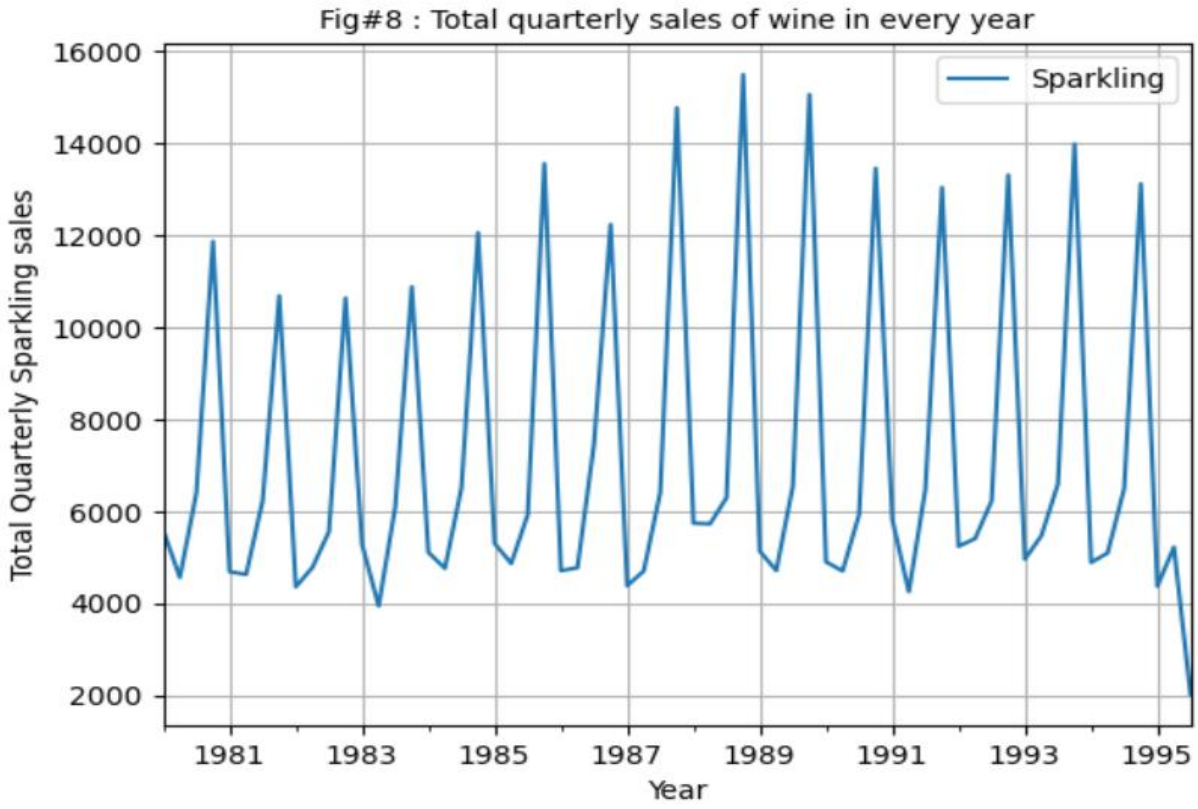
We can see that sales are maximum for December month.

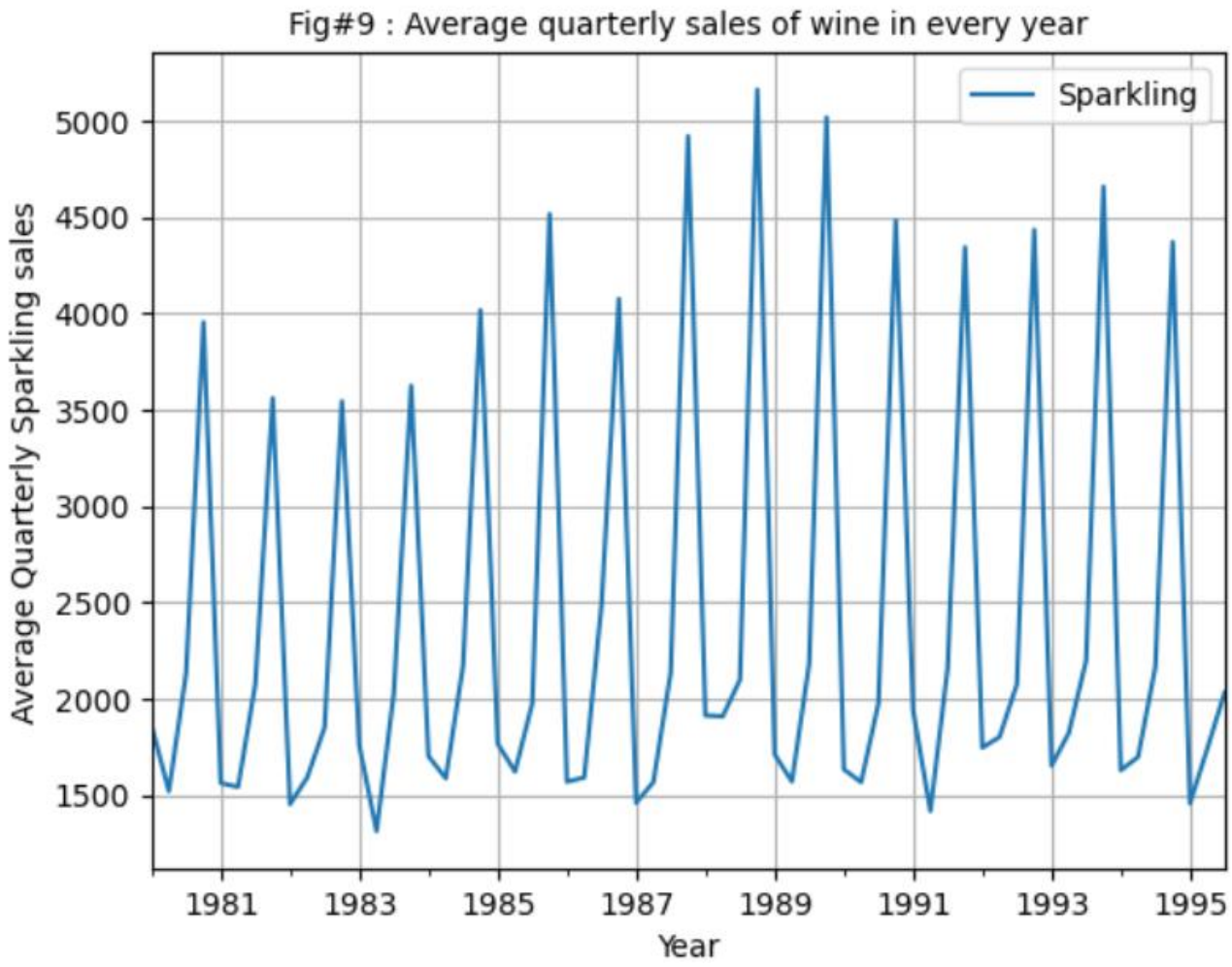
Yearly plot





Quarterly plot

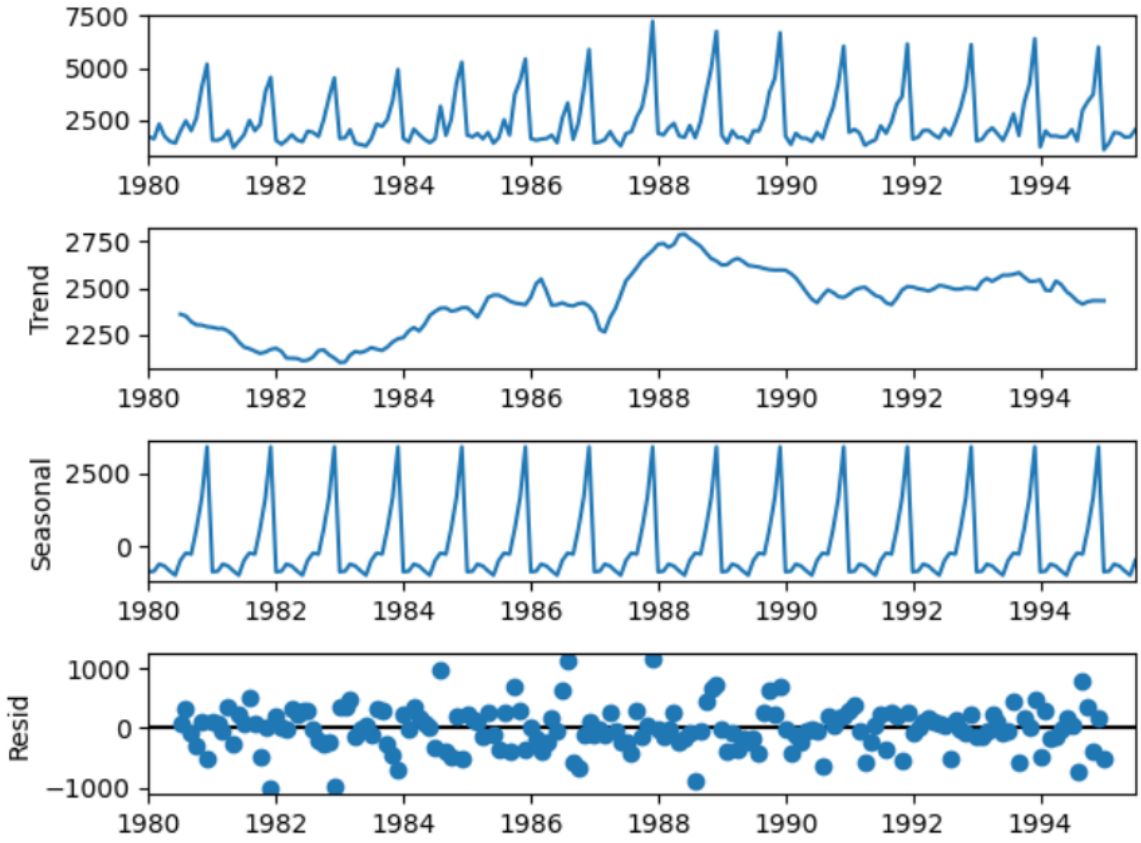




Decompose the Time Series

We will decompose the time series with both model additive and multiplicative.

Additive Model



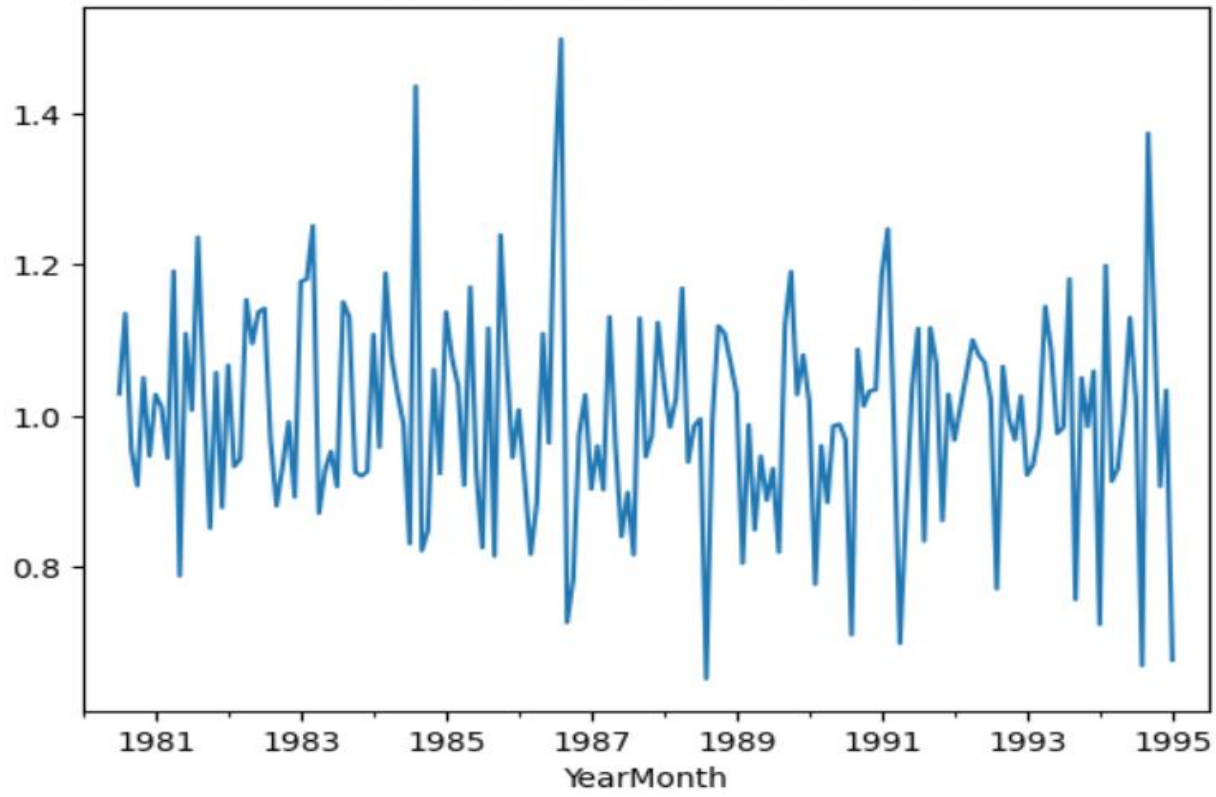
~ We can see that seasonality present in the time series data.

~ We can see that residuals are random as there is no trend present in the residuals.

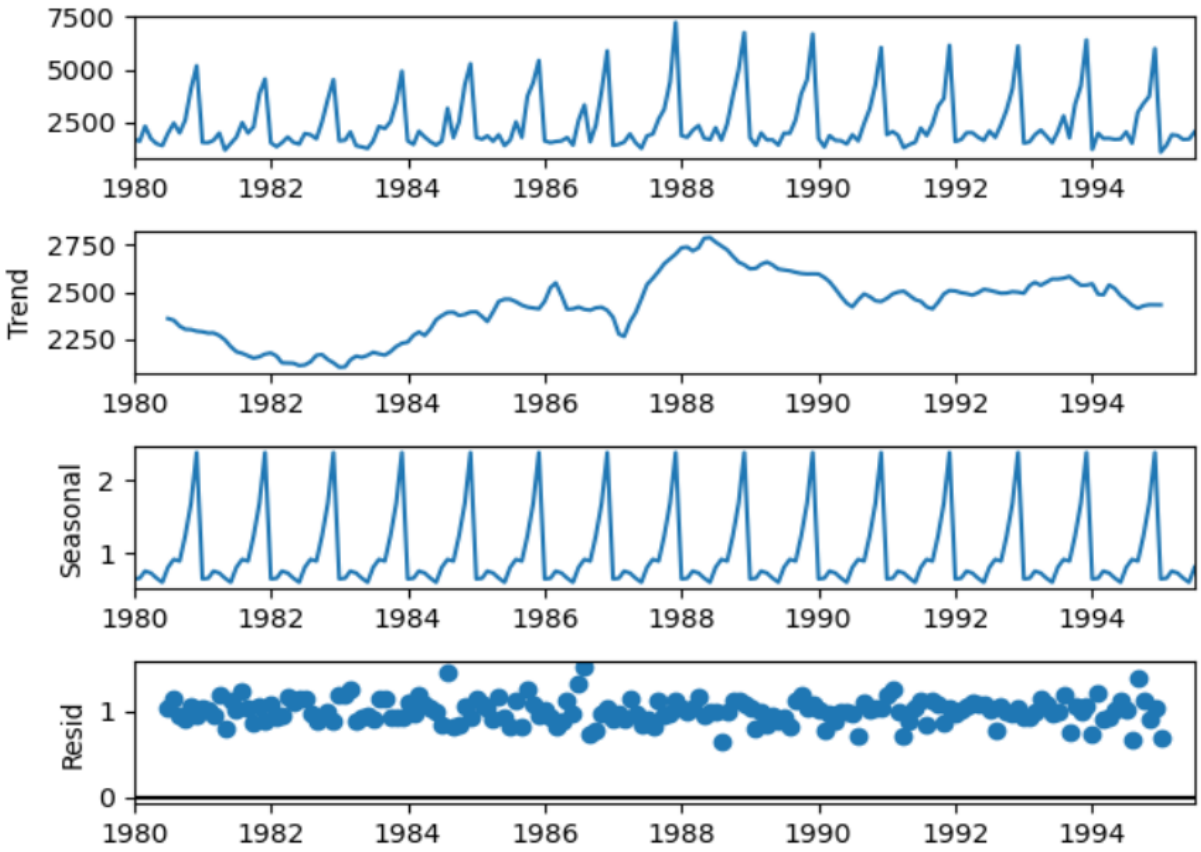
```
Trend
 YearMonth
1980-01-01             NaN
1980-02-01             NaN
1980-03-01             NaN
1980-04-01             NaN
1980-05-01             NaN
1980-06-01             NaN
1980-07-01     2360.666667
1980-08-01     2351.333333
1980-09-01     2320.541667
1980-10-01     2303.583333
1980-11-01     2302.041667
1980-12-01     2293.791667
Name: trend, dtype: float64

Seasonality
 YearMonth
1980-01-01     -854.260599
1980-02-01     -830.350678
1980-03-01     -592.356630
1980-04-01     -658.490559
1980-05-01     -824.416154
1980-06-01     -967.434011
1980-07-01     -465.502265
1980-08-01     -214.332821
1980-09-01     -254.677265
1980-10-01      599.769957
1980-11-01     1675.067179
1980-12-01     3386.983846
Name: seasonal, dtype: float64
```

```
Residual
 YearMonth
1980-01-01           NaN
1980-02-01           NaN
1980-03-01           NaN
1980-04-01           NaN
1980-05-01           NaN
1980-06-01           NaN
1980-07-01     70.835599
1980-08-01    315.999487
1980-09-01    -81.864401
1980-10-01   -307.353290
1980-11-01    109.891154
1980-12-01   -501.775513
Name: resid, dtype: float64
```

Fig#10 :Additive models Residual plot



Multiplicative Model

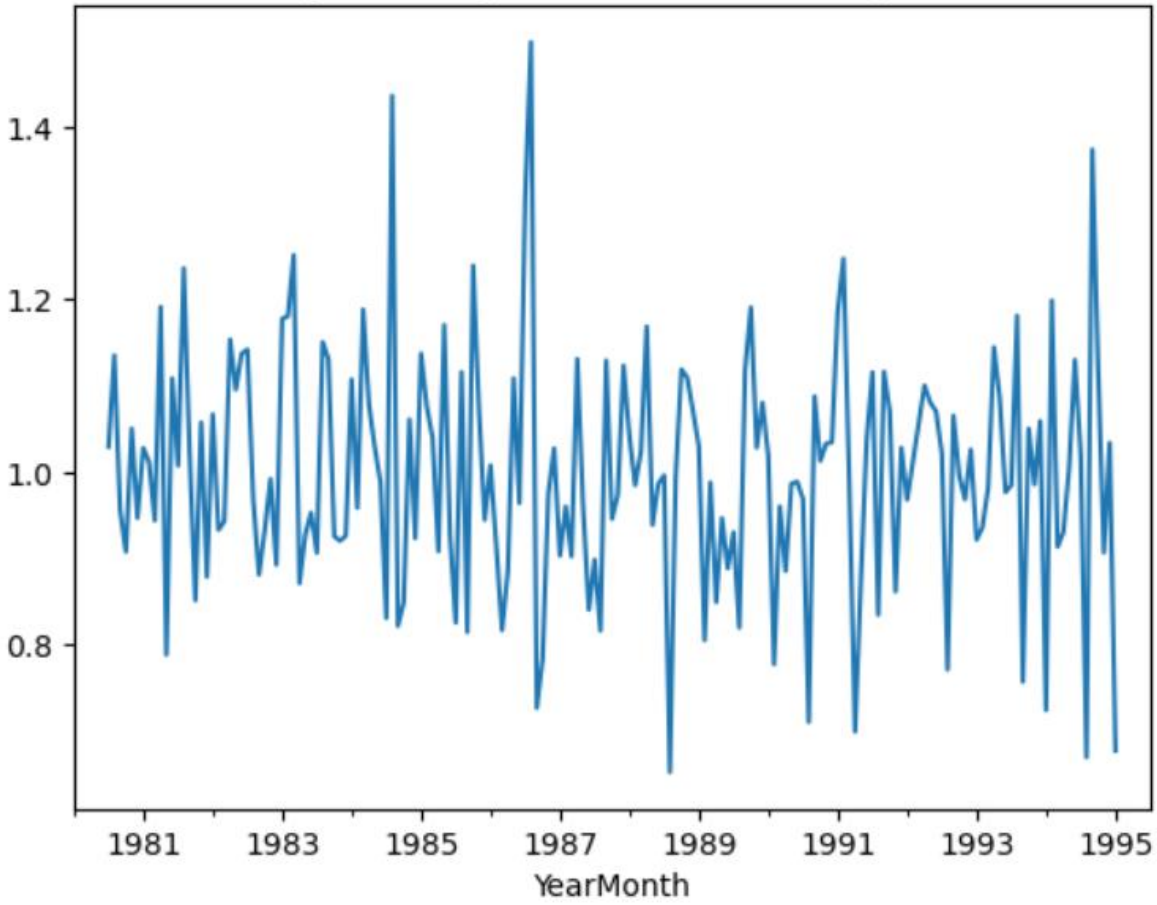


As per multiplicative model We can see that seasonality present in the time series data.

```
Trend
 YearMonth
1980-01-01              NaN
1980-02-01              NaN
1980-03-01              NaN
1980-04-01              NaN
1980-05-01              NaN
1980-06-01              NaN
1980-07-01     2360.666667
1980-08-01     2351.333333
1980-09-01     2320.541667
1980-10-01     2303.583333
1980-11-01     2302.041667
1980-12-01     2293.791667
Name: trend, dtype: float64

Seasonality
 YearMonth
1980-01-01     0.649843
1980-02-01     0.659214
1980-03-01     0.757440
1980-04-01     0.730351
1980-05-01     0.660609
1980-06-01     0.603468
1980-07-01     0.809164
1980-08-01     0.918822
1980-09-01     0.894367
1980-10-01     1.241789
1980-11-01     1.690158
1980-12-01     2.384776
Name: seasonal, dtype: float64
```

```
Residual
 YearMonth
1980-01-01          NaN
1980-02-01          NaN
1980-03-01          NaN
1980-04-01          NaN
1980-05-01          NaN
1980-06-01          NaN
1980-07-01     1.029230
1980-08-01     1.135407
1980-09-01     0.955954
1980-10-01     0.907513
1980-11-01     1.050423
1980-12-01     0.946770
Name: resid, dtype: float64
```

Fig#11 :Multiplicative models Residual plot



~ We can see that in both residual's plot residuals are random as it does not follow any pattern and so there is no clear merit to specifically choose multiplicative decomposition as additive is good enough.

~ We can say that there is no clear trend present.

Q3 Split the data into training and test. The test data should start in 1991.

We have split the data into train, test and below is the sample of that data.

First few rows of Training Data

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

Last few rows of Training Data

YearMonth	Sparkling
1990-08-01	1605
1990-09-01	2424
1990-10-01	3116
1990-11-01	4286
1990-12-01	6047

First few rows of Test Data

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

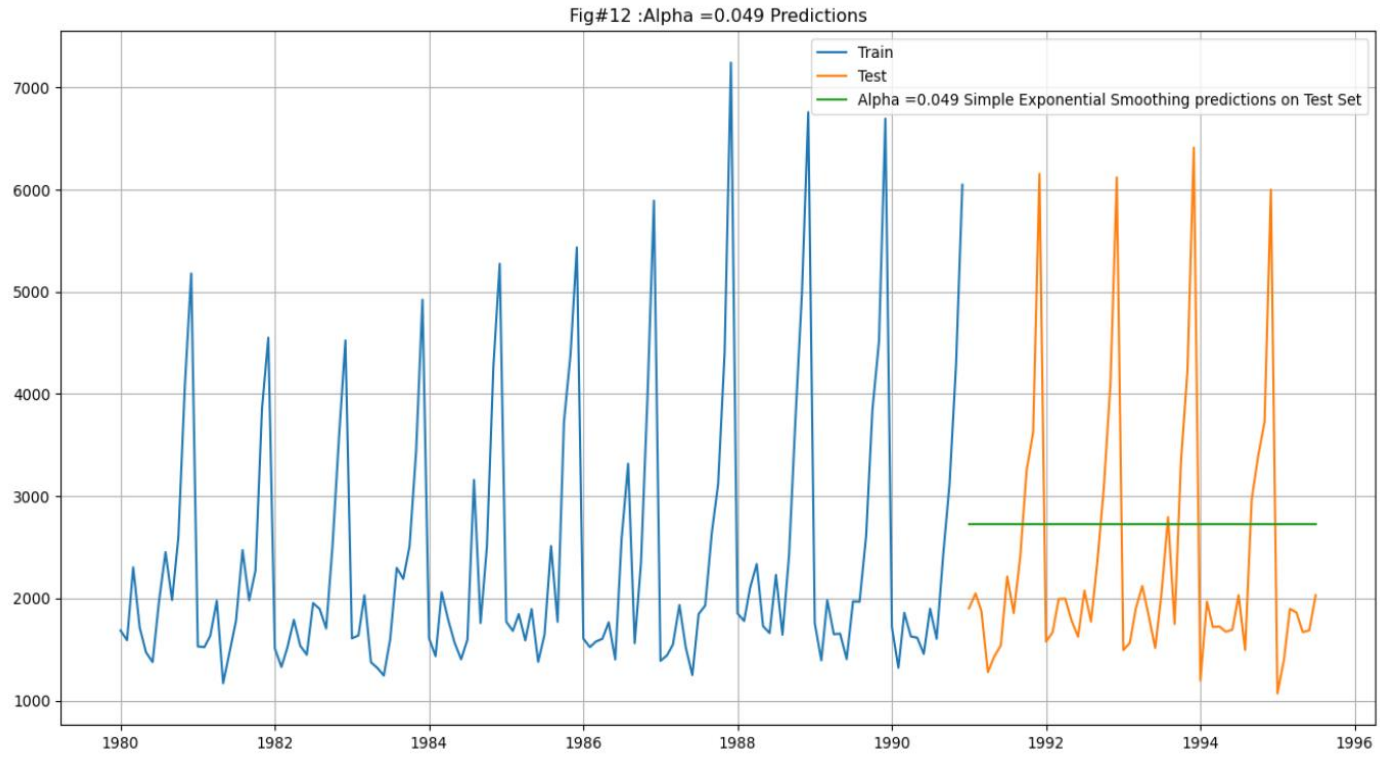
Last few rows of Test Data

YearMonth	Sparkling
1995-03-01	1897
1995-04-01	1862
1995-05-01	1670
1995-06-01	1688
1995-07-01	2031

Q4 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.

Model 1: Simple Exponential Smoothing

First, we are building SES model by using default Alpha value (smoothing level) which is – 0.049



For Alpha =0.049 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 1316.035

Checking different alpha values:

After trying with different alpha values, we got following RMSE.

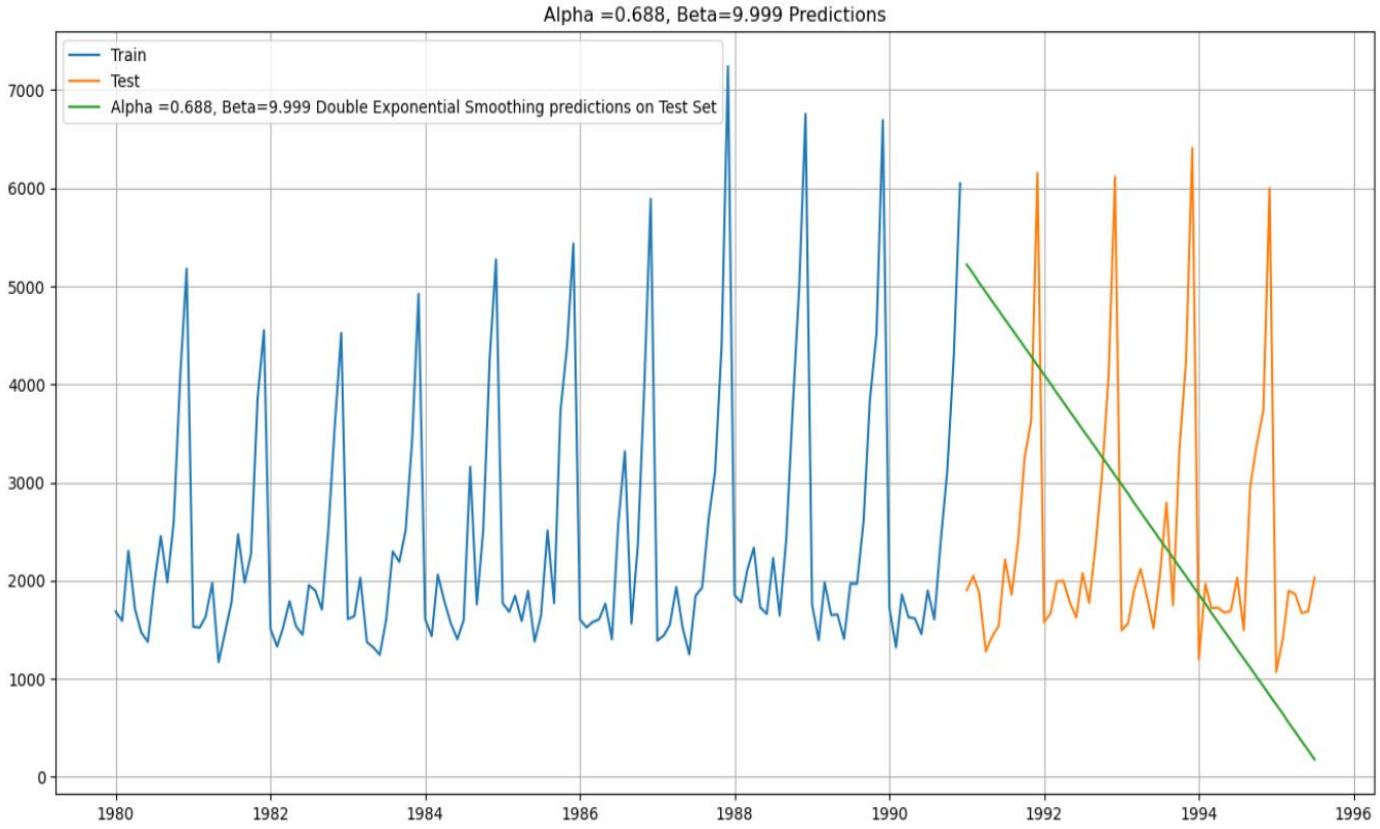
	Alpha Values	Train RMSE	Test RMSE
0	0.3	1359.511747	1935.507132
1	0.4	1352.588879	2311.919615
2	0.5	1344.004369	2666.351413
3	0.6	1338.805381	2979.204388
4	0.7	1338.844308	3249.944092
5	0.8	1344.462091	3483.801006
6	0.9	1355.723518	3686.794285

~ After checking different value of alpha, we are unable to get RMSE less than 1316.035 which we obtained when Alpha = 0.049

~ With Alpha =0.049 we can say that past observations have a large influence on forecasts.

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

First, we are building DES model by using default Alpha value (smoothing level) which is – 0.688 and Beta value (smoothing trend) is - 9.999



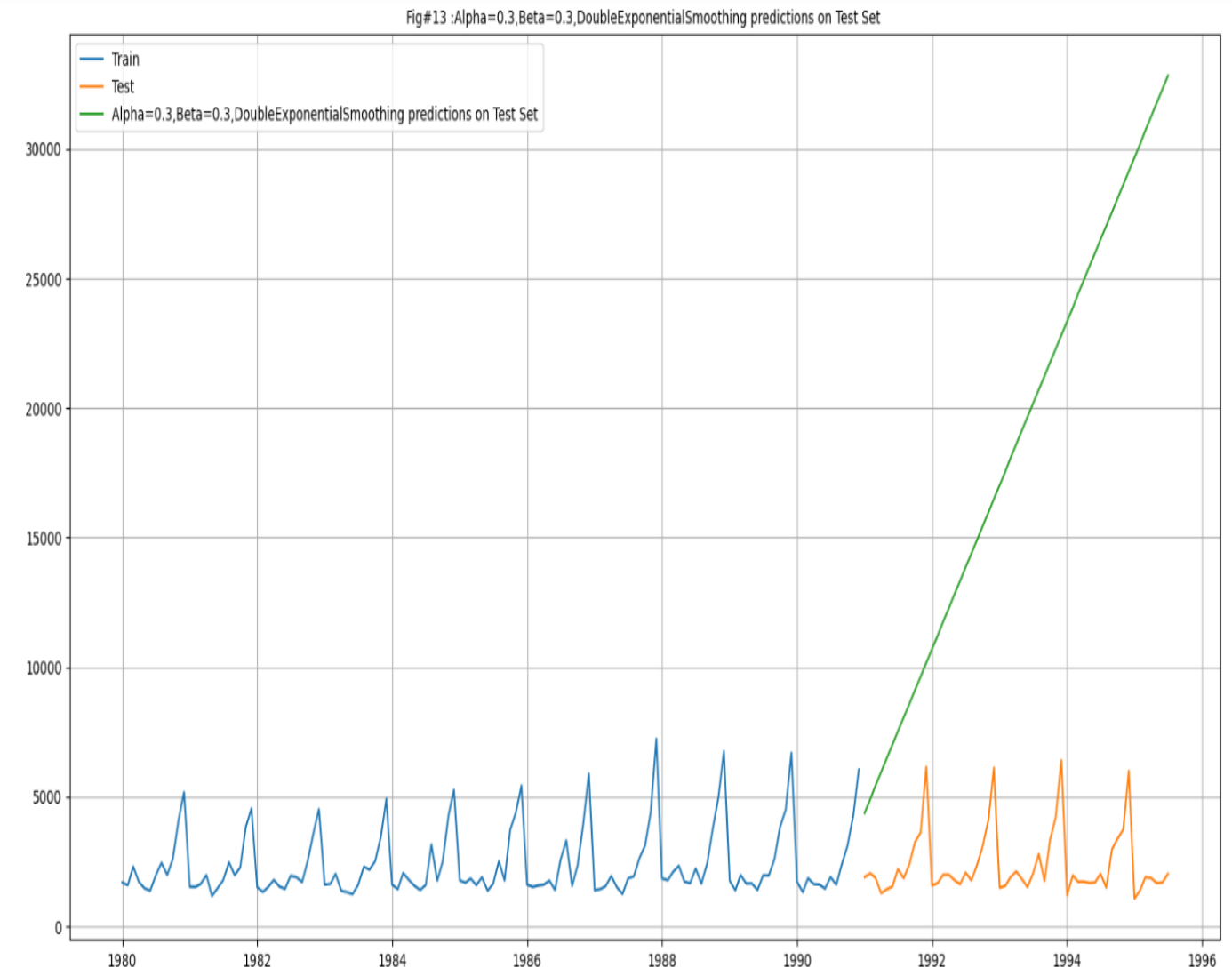
For Alpha =0.688, Beta=9.999 Double Exponential Smoothing Model forecast on the Test Data, RMSE is 2007.239

Checking different alpha & Beta values:

After trying with different alpha & Beta values, we got following RMSE.

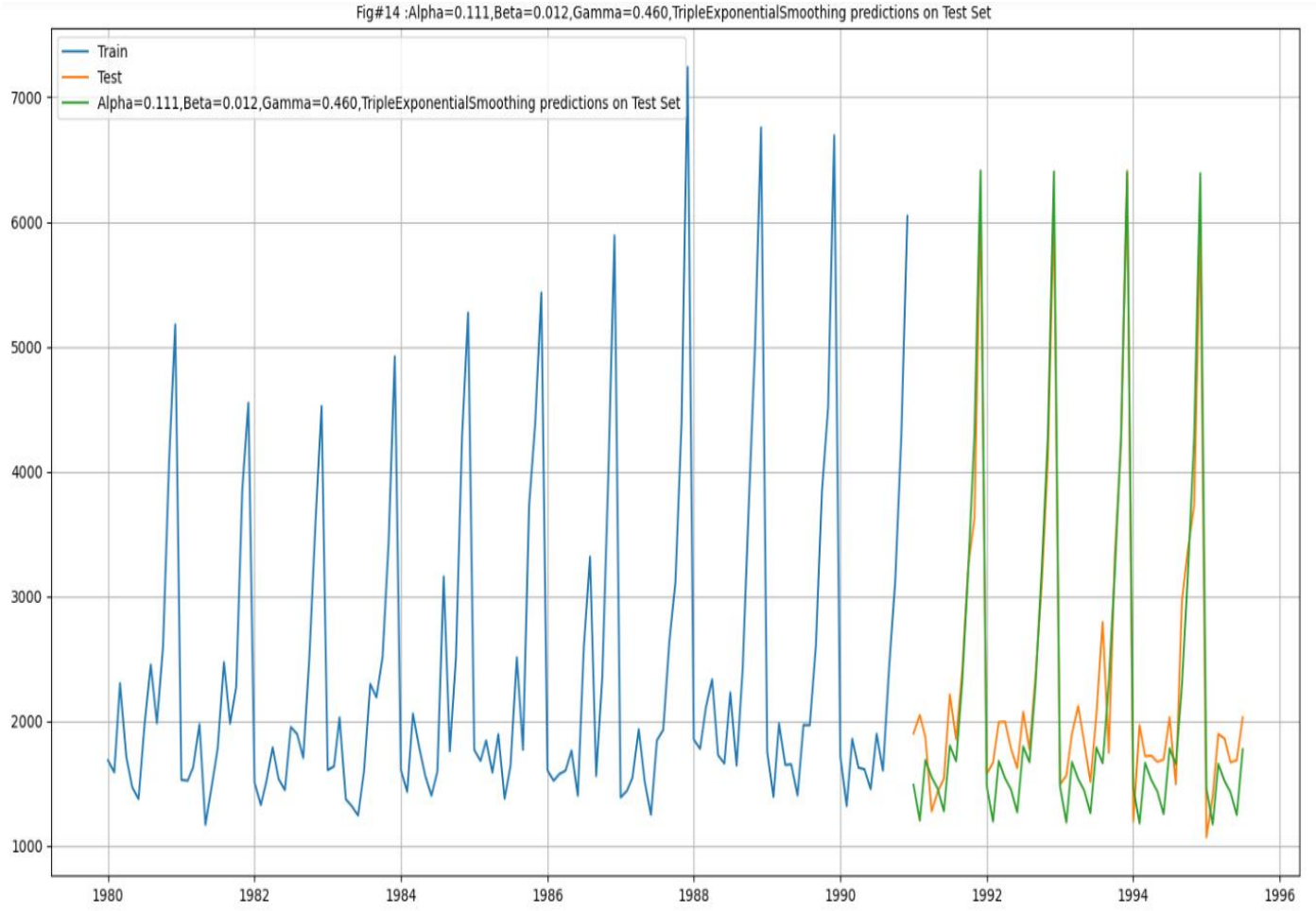
	Alpha Values	Beta Values	Train RMSE	Test RMSE
0	0.3	0.3	1592.292788	18259.110704
8	0.4	0.3	1569.338606	23878.496940
1	0.3	0.4	1682.573828	26069.841401
16	0.5	0.3	1530.575845	27095.532414
24	0.6	0.3	1506.449870	29070.722592

We can see that RMSE value is not less than 2007.239 which we obtained when Alpha =0.688, Beta=9.999



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

First, we are building TES model by using default Alpha value (smoothing level) which is – 0.111, Beta value (smoothing trend) is - 0.012 & Gamma value (smoothing seasonal) is- 0.460



For Alpha=0.111, Beta=0.012, Gamma=0.460, Triple Exponential Smoothing Model forecast on the Test Data, RMSE is 378.951

Checking different alpha, Beta & Gamma values:

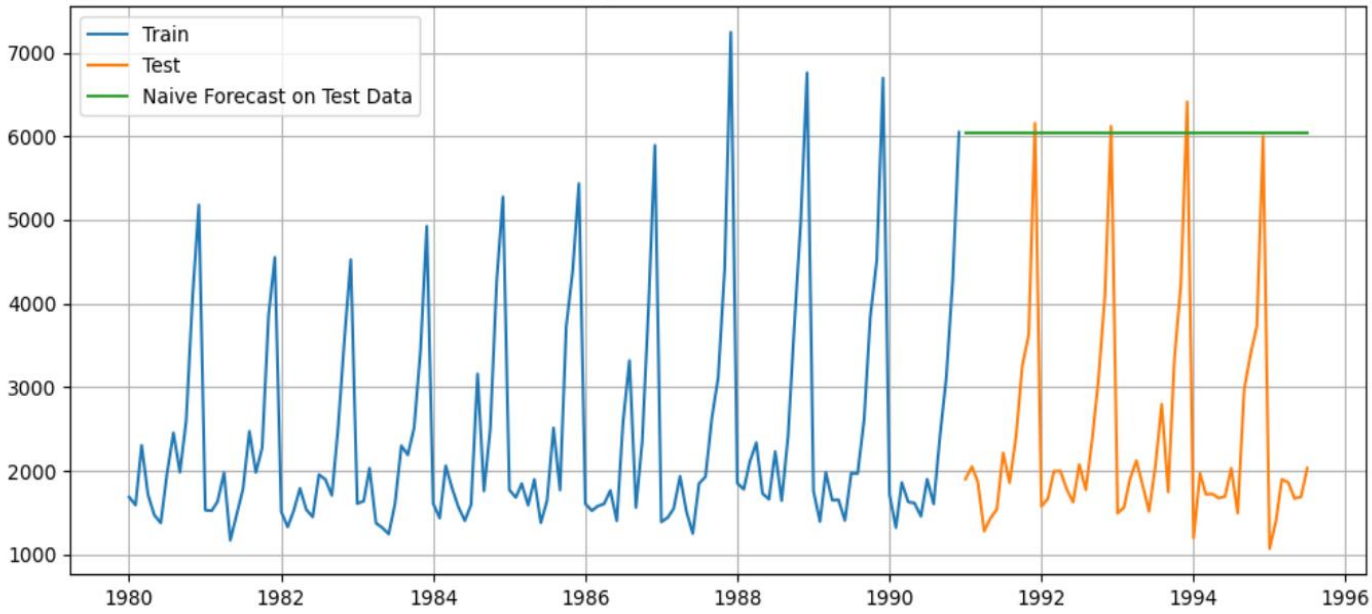
After trying with different alpha, Beta & gamma values, we got following RMSE.

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE
129	0.5	0.3	0.4	477.403185	649.197748
7	0.3	0.3	1.0	538.858393	687.206158
64	0.4	0.3	0.3	464.061379	734.718132
464	1.0	0.5	0.3	660.104855	769.360784
196	0.6	0.3	0.7	569.835997	926.412440

After checking different values of alpha, Beta & Gamma we are unable to get RMSE less than 378.95 which we obtained when Alpha=0.111, Beta=0.012, Gamma=0.460

Model 4: Naive forecast

Fig#15: Naive Forecast



For naive forecast on the Test Data, RMSE is 3864.279

Model 5: Linear Regression

For regression first we need to add numerical time instance order in the data and here is the sample of that data.

First few rows of Training Data

YearMonth	Sparkling	time
1980-01-01	1686	1
1980-02-01	1591	2
1980-03-01	2304	3
1980-04-01	1712	4
1980-05-01	1471	5

Last few rows of Training Data

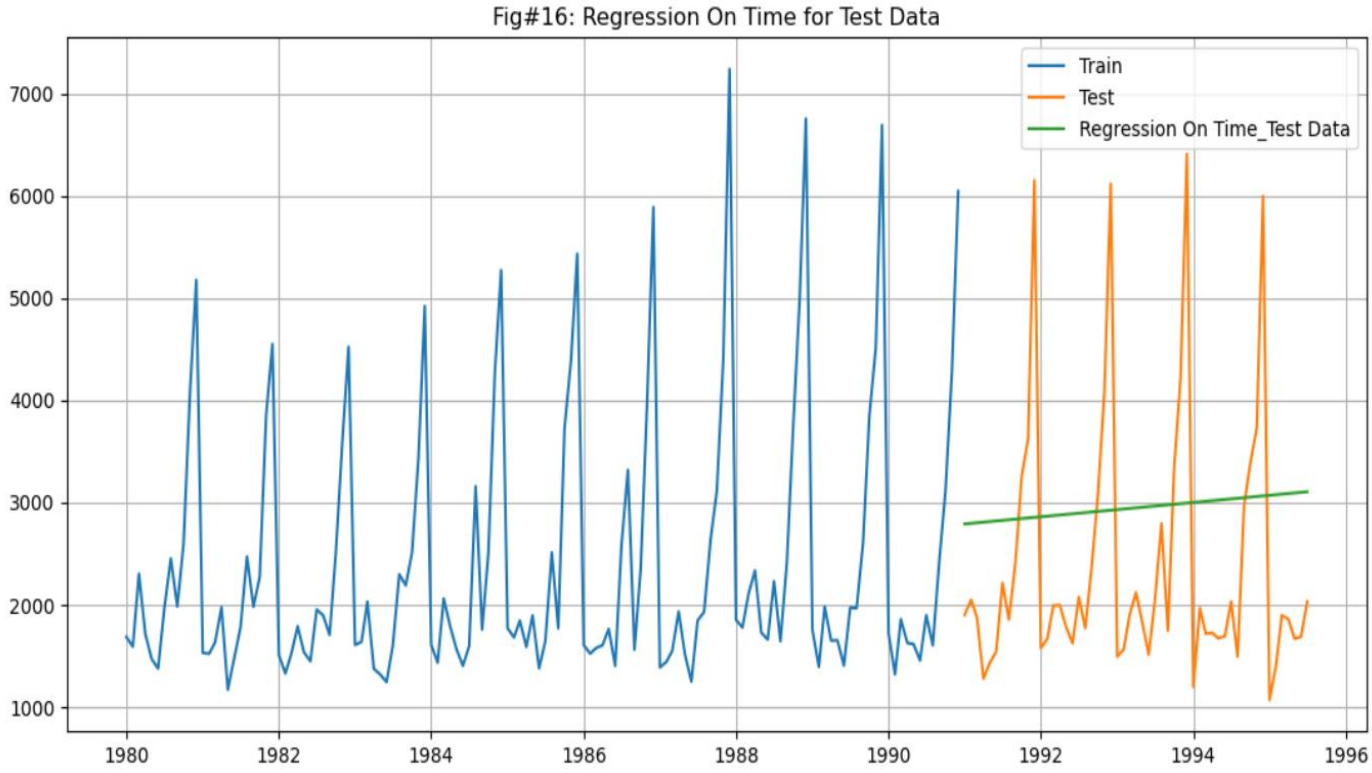
YearMonth	Sparkling	time
1990-08-01	1605	128
1990-09-01	2424	129
1990-10-01	3116	130
1990-11-01	4286	131
1990-12-01	6047	132

First few rows of Test Data

YearMonth	Sparkling	time
1991-01-01	1902	133
1991-02-01	2049	134
1991-03-01	1874	135
1991-04-01	1279	136
1991-05-01	1432	137

Last few rows of Test Data

YearMonth	Sparkling	time
1995-03-01	1897	183
1995-04-01	1862	184
1995-05-01	1670	185
1995-06-01	1688	186
1995-07-01	2031	187



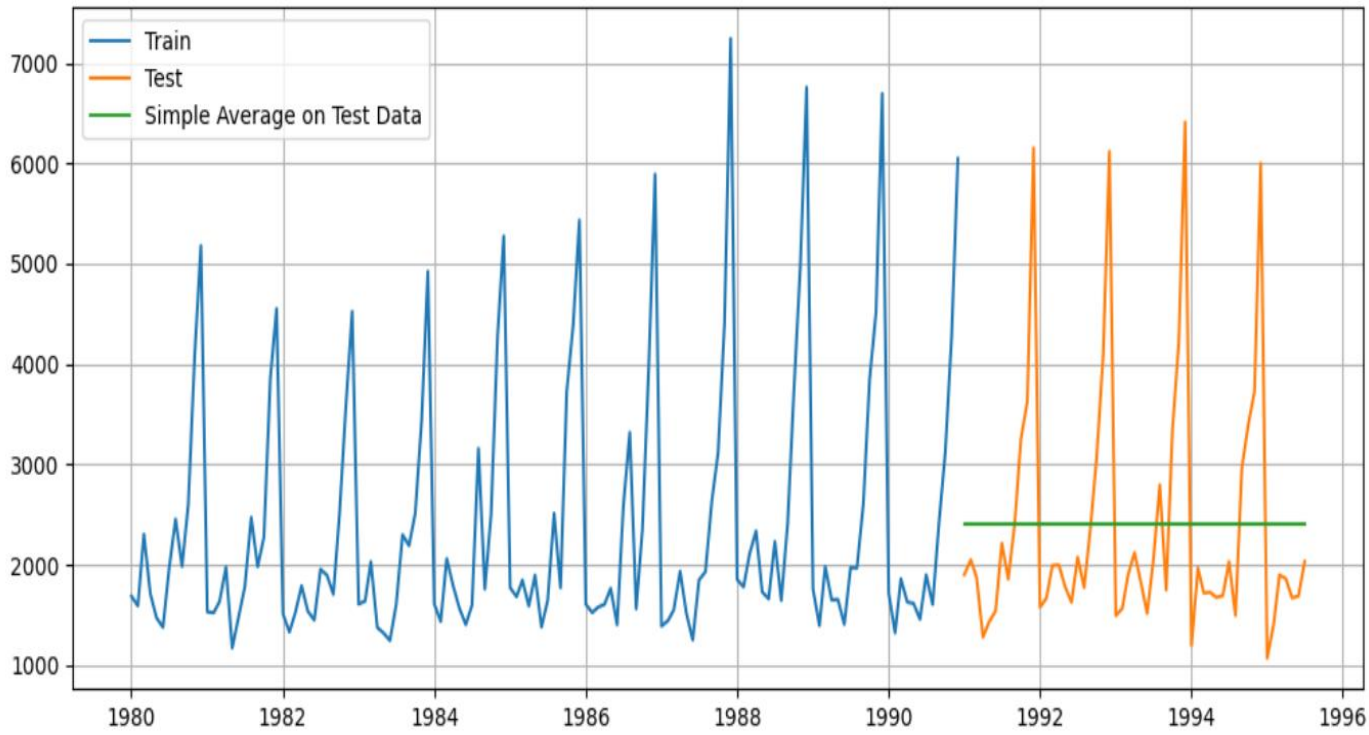
For model Regression on Time forecast on the Test Data, RMSE is 1389.135

Model 6 : Simple Average

Below is the mean forecast as per Simple Average

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

Fig:17 Simple Average Forecast

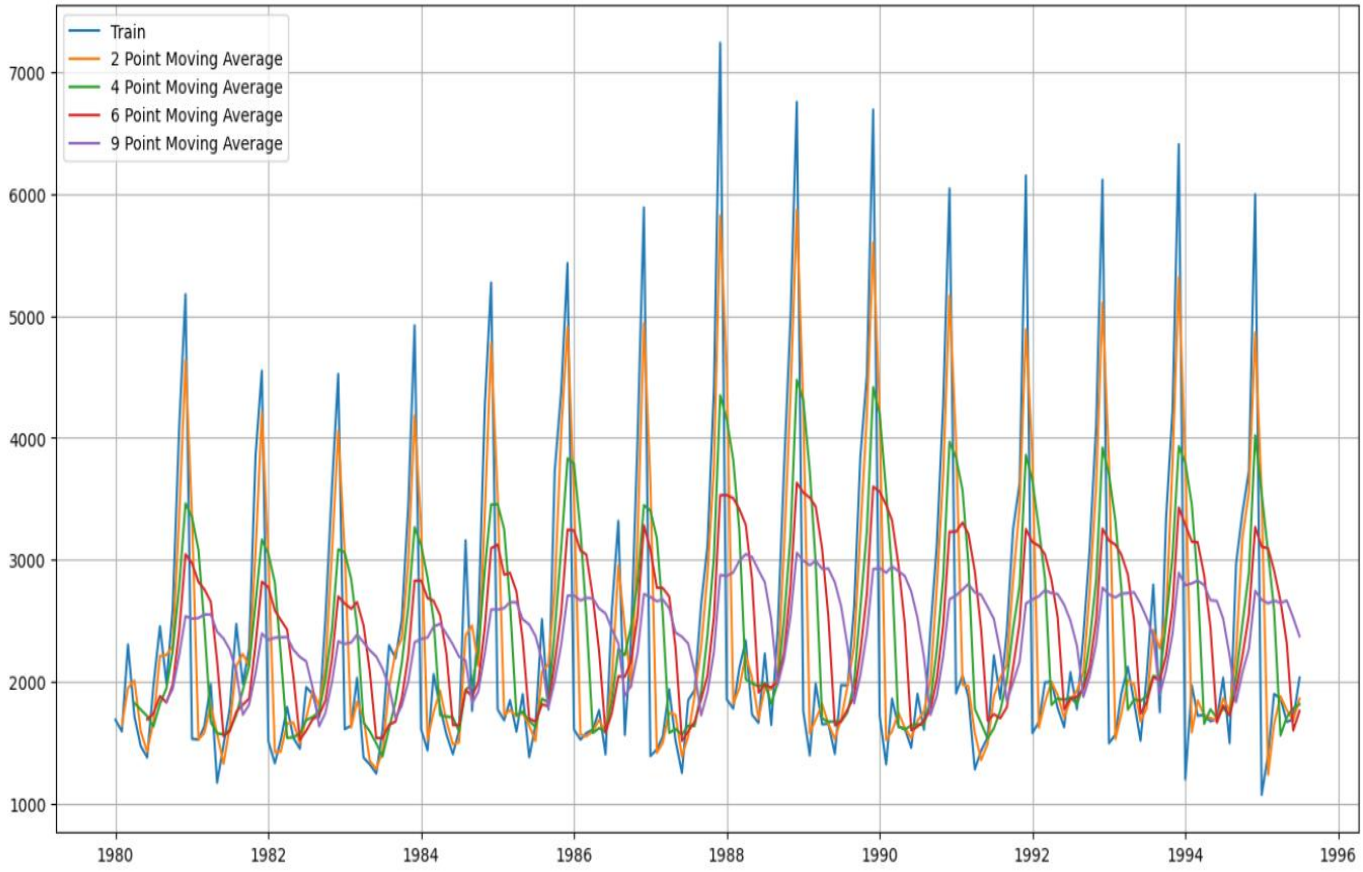


For Simple Average forecast on the Test Data, RMSE is 1275.082

Model 11 : Moving Average(MA)

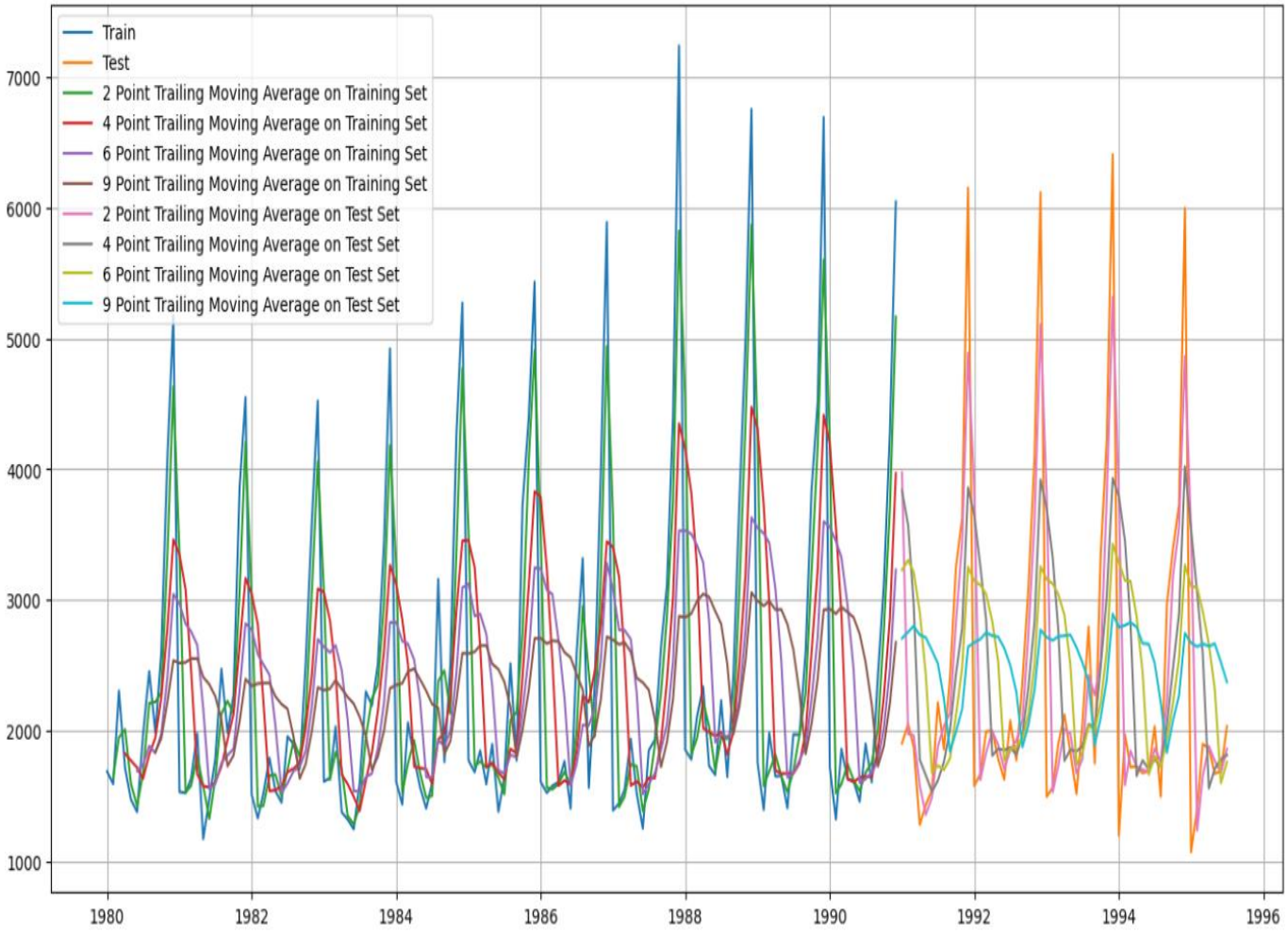
Trailing moving averages

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



Sample of trailing Moving Average train data

YearMonth	Sparkling	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1990-08-01	1605	1752.0	1644.00	1677.166667	2199.777778
1990-09-01	2424	2014.5	1846.25	1771.333333	1725.333333
1990-10-01	3116	2770.0	2261.00	2019.333333	1880.444444
1990-11-01	4286	3701.0	2857.75	2464.500000	2209.888889
1990-12-01	6047	5166.5	3968.25	3229.500000	2675.222222



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

So far, we have tried 7 different models and Triple Exponential Smoothing suits very well on our data because seasonality present in the data which considered in the Triple Exponential Smoothing.

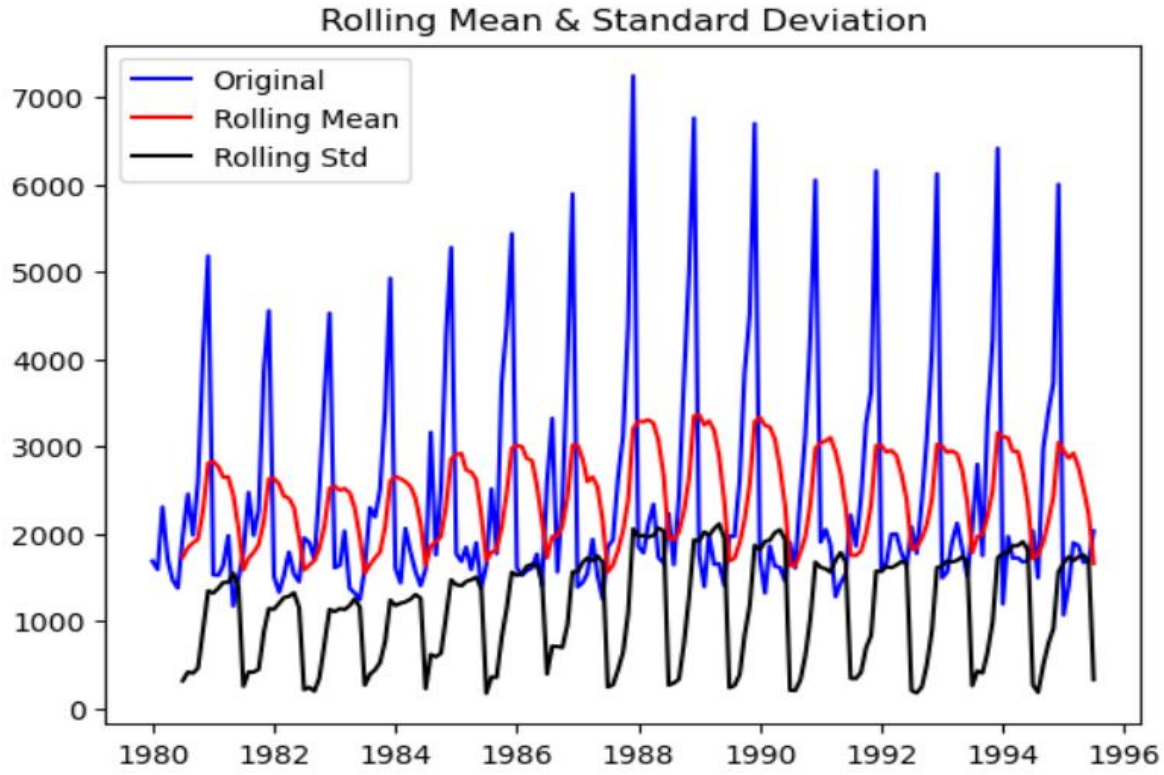
Q5 Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05.

To check the stationarity of the data ADF test can be used which is a hypothesis test and following is the null & alternate hypothesis for this test.

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.

Check for stationarity of the whole Time Series data

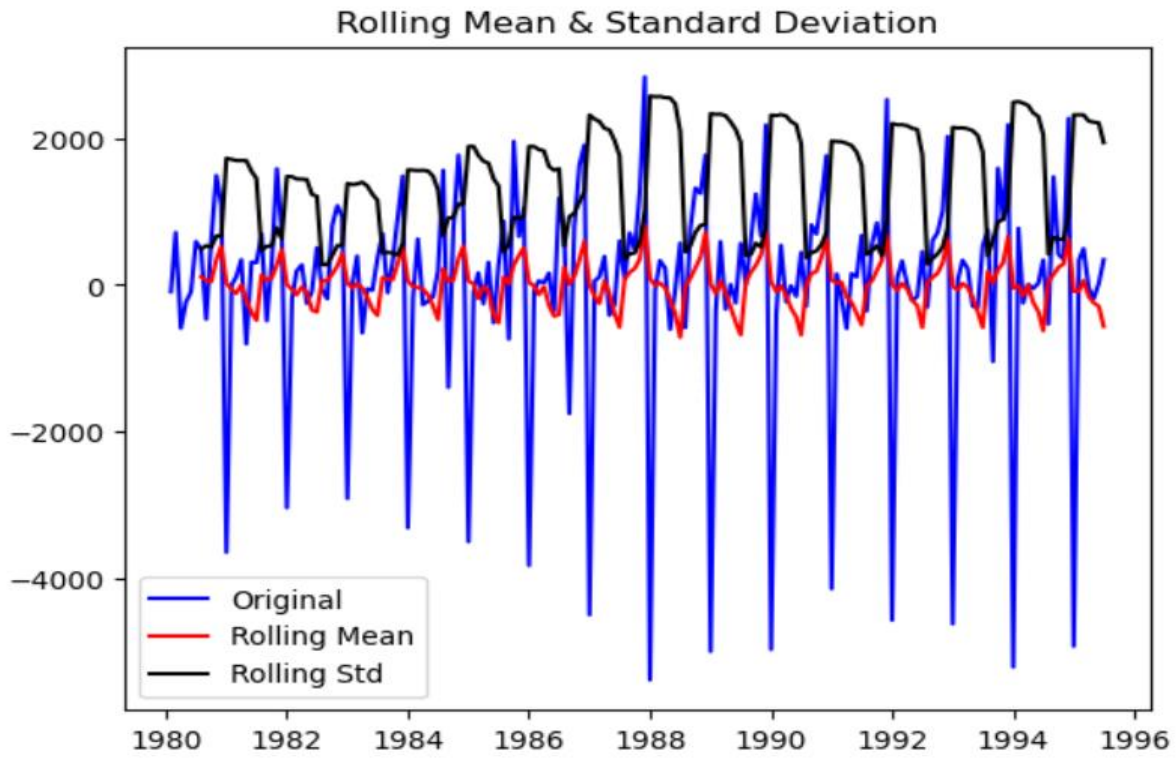


Results of Dickey-Fuller Test:

```
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
```

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.

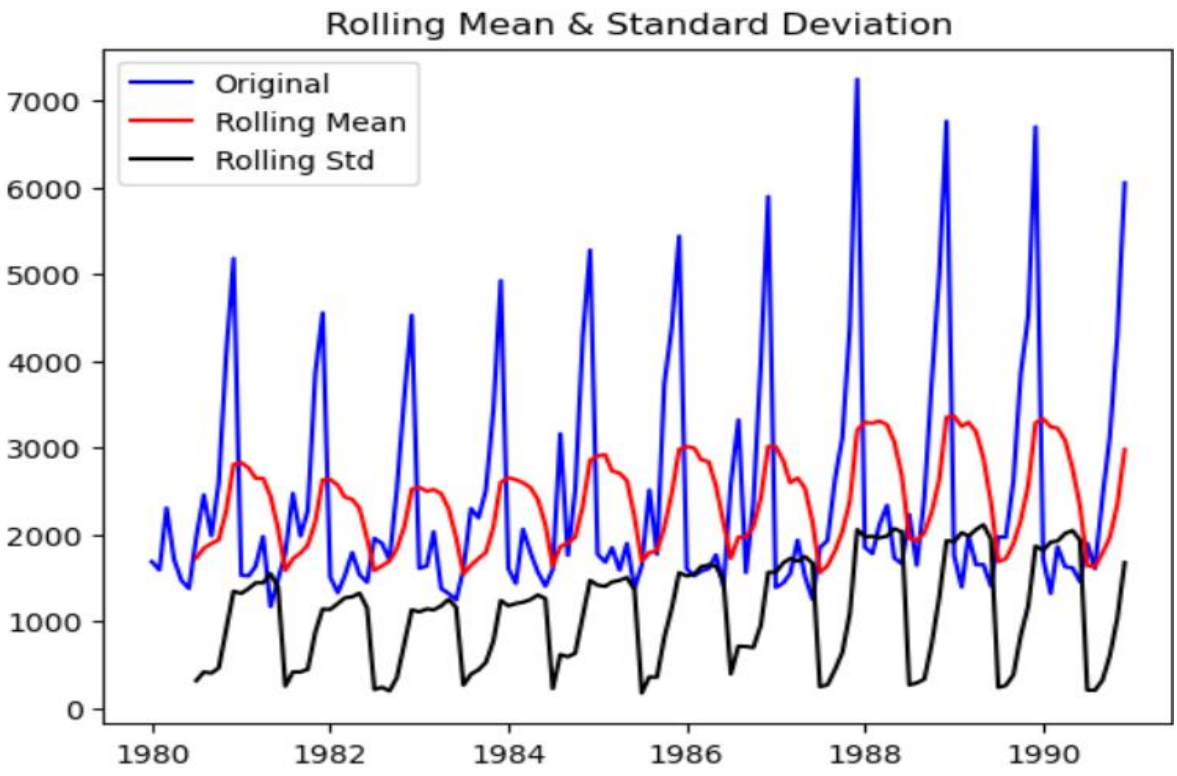


Results of Dickey-Fuller Test:

```
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 $\alpha = 0.05$ the Time Series become stationary.

Check for stationarity of the Training Data Time Series:

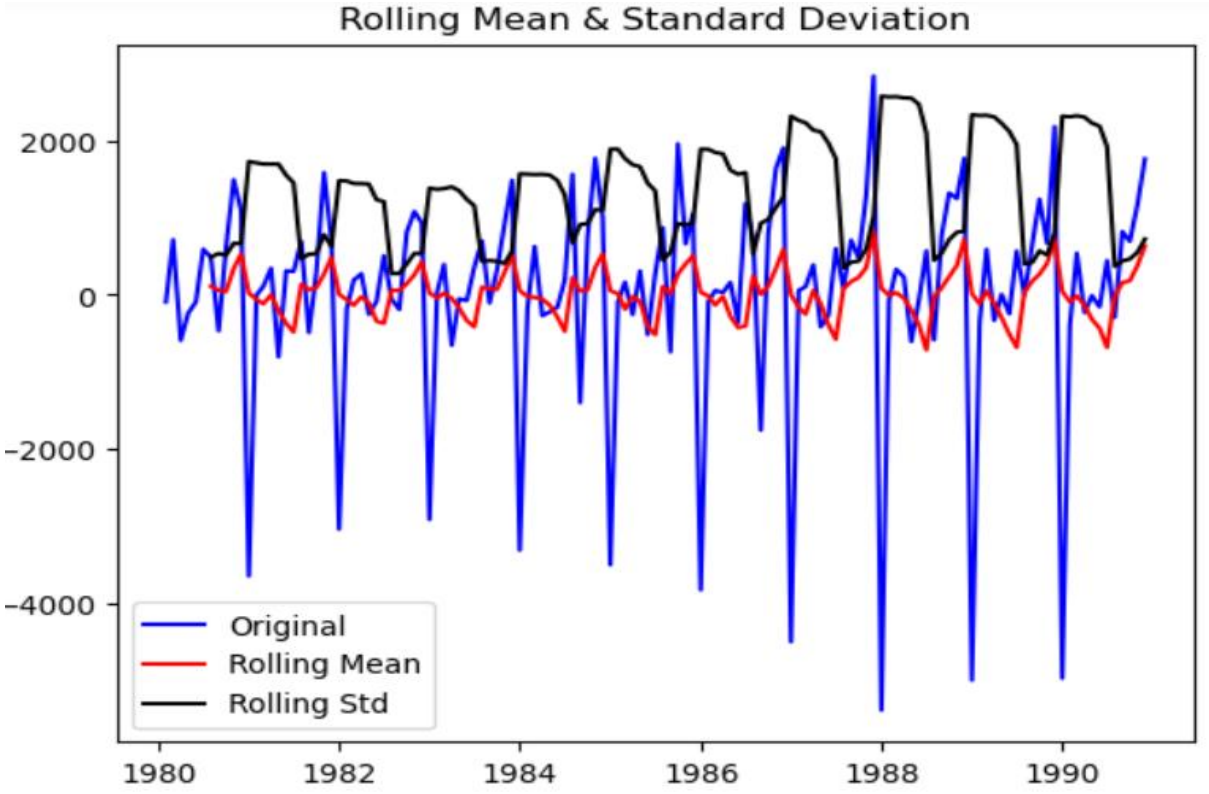


Results of Dickey-Fuller Test:

```
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
dtype: float64
```

We see that the train data is not stationary at $\alpha = 0.05$

Let us take a difference of order 1 and check whether train data is stationary or not



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 train data have become stationary at $\alpha = 0.05$

Q6 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.

Model 7 : Automated ARIMA Model

We are taking all combination of p, d, q values from 0 to 2 and below is the AIC of these (p, d, q) values.

	param	AIC
8	(2, 1, 2)	2213.509212
7	(2, 1, 1)	2233.777626
2	(0, 1, 2)	2234.408323
5	(1, 1, 2)	2234.527200
4	(1, 1, 1)	2235.755095
6	(2, 1, 0)	2260.365744
1	(0, 1, 1)	2263.060016
3	(1, 1, 0)	2266.608539
0	(0, 1, 0)	2267.663036

We can see that lowest AIC value generated when p=2, d=1, q=2 so we are building ARIMA model with these p, d, q values.

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	132
Model:	ARIMA(2, 1, 2)	Log Likelihood	-1101.755
Date:	Mon, 26 Dec 2022	AIC	2213.509
Time:	12:46:45	BIC	2227.885
Sample:	01-01-1980	HQIC	2219.351
	- 12-01-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.3121	0.046	28.782	0.000	1.223	1.401
ar.L2	-0.5593	0.072	-7.741	0.000	-0.701	-0.418
ma.L1	-1.9917	0.109	-18.217	0.000	-2.206	-1.777
ma.L2	0.9999	0.110	9.109	0.000	0.785	1.215
sigma2	1.099e+06	1.99e-07	5.51e+12	0.000	1.1e+06	1.1e+06

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

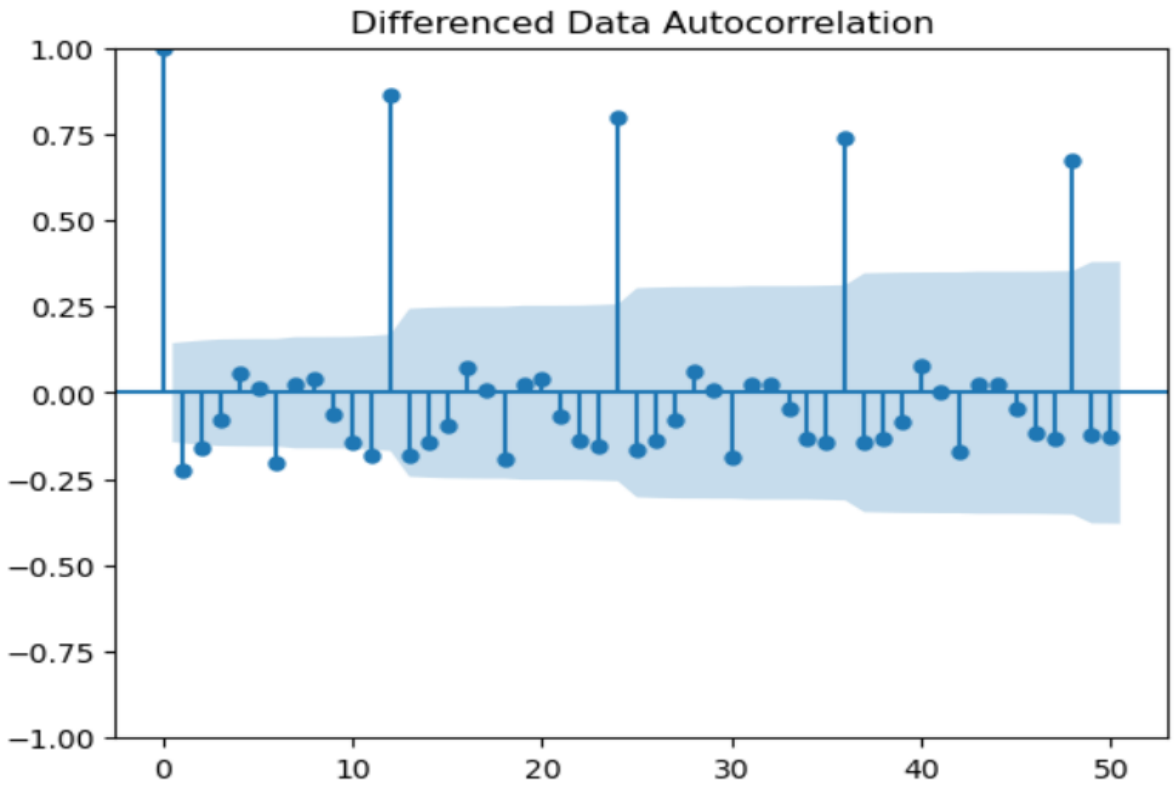
Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 8.24e+27. Standard errors may be unstable.

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

Automated ARIMA model (2, 1, 2) RMSE score = 1299.9800412702375

Model 8 : Automated SARIMA Model

Let us look at the ACF plot to understand the seasonal parameter for the SARIMA model.



~ We can see that there is a seasonality.

~ We see that there can be a seasonality of 6 as well as 12. We will run our auto SARIMA models by setting seasonality both as 6 and 12.

Setting the seasonality as 6 for the first iteration of the auto SARIMA model

Examples of some parameter combinations for Model...

```
Model: (0, 1, 1)(0, 0, 1, 6)
Model: (0, 1, 2)(0, 0, 2, 6)
Model: (1, 1, 0)(1, 0, 0, 6)
Model: (1, 1, 1)(1, 0, 1, 6)
Model: (1, 1, 2)(1, 0, 2, 6)
Model: (2, 1, 0)(2, 0, 0, 6)
```

```
Model: (2, 1, 1)(2, 0, 1, 6)
Model: (2, 1, 2)(2, 0, 2, 6)
```

Top 5 lowest AIC values:

	param	seasonal	AIC
53	(1, 1, 2)	(2, 0, 2, 6)	1727.666982
26	(0, 1, 2)	(2, 0, 2, 6)	1727.888803
17	(0, 1, 1)	(2, 0, 2, 6)	1741.703672
44	(1, 1, 1)	(2, 0, 2, 6)	1743.379778
71	(2, 1, 1)	(2, 0, 2, 6)	1744.040750

We can see that lowest AIC value generated when p=1, d=1, q=2 & P=2, D=0, Q=2, S=6 so we are building SARIMA model with these p, d, q & P, D, Q, S values.

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                  132
Model:             SARIMAX(1, 1, 2)x(2, 0, 2, 6)   Log Likelihood                -855.833
Date:                           Mon, 26 Dec 2022   AIC                            1727.667
Time:                                   12:47:49   BIC                            1749.696
Sample:                                        0   HQIC                           1736.609
                                             - 132
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.6444      0.286     -2.256      0.024      -1.204      -0.085
ma.L1          0.3820      0.368      1.038      0.299      -0.339       1.103
ma.L2         -1.1367      0.357     -3.180      0.001      -1.837      -0.436
ar.S.L6       -0.0044      0.027     -0.164      0.870      -0.057       0.049
ar.S.L12       1.0361      0.018     56.101      0.000       1.000       1.072
ma.S.L6        0.0673      0.152      0.443      0.658      -0.231       0.365
ma.S.L12      -0.6126      0.093     -6.595      0.000      -0.795      -0.431
sigma2      8.922e+04   5.05e+04      1.768      0.077   -9703.822    1.88e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.09   Jarque-Bera (JB):                 25.26
Prob(Q):                              0.77   Prob(JB):                          0.00
Heteroskedasticity (H):               2.63   Skew:                              0.47
Prob(H) (two-sided):                  0.00   Kurtosis:                          5.09
===================================================================================

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

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

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1330.505332	380.584338	584.573736	2076.436928
1	1177.544582	392.138322	408.967595	1946.121570
2	1626.050444	392.332657	857.092567	2395.008321
3	1546.469801	397.727590	766.938049	2326.001553
4	1309.113097	398.948350	527.188699	2091.037495

Automated SARIMA model (1,1,2) (2,0,2,6) RMSE score = 626.47

Setting the seasonality as 12 for the second iteration of the auto SARIMA model

```
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 lowest AIC values:

	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)	1557.121563
23	(0, 1, 2)	(1, 0, 2, 12)	1557.160507
77	(2, 1, 2)	(1, 0, 2, 12)	1557.340403

We can see that lowest AIC value generated when p=1, d=1, q=2 & P=1, D=0, Q=2, S=12 so we are building SARIMA model with these p, d, q & P, D, Q, S values.

```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(1, 1, 2)x(1, 0, 2, 12)   Log Likelihood                -770.792
Date:                            Mon, 26 Dec 2022   AIC                           1555.584
Time:                                    12:49:19   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     -2.464      0.014      -1.128      -0.128
ma.L1         -0.1040      0.225     -0.463      0.644      -0.545       0.337
ma.L2         -0.7276      0.154     -4.736      0.000      -1.029      -0.427
ar.S.L12       1.0439      0.014     72.838      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
===================================================================================

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

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

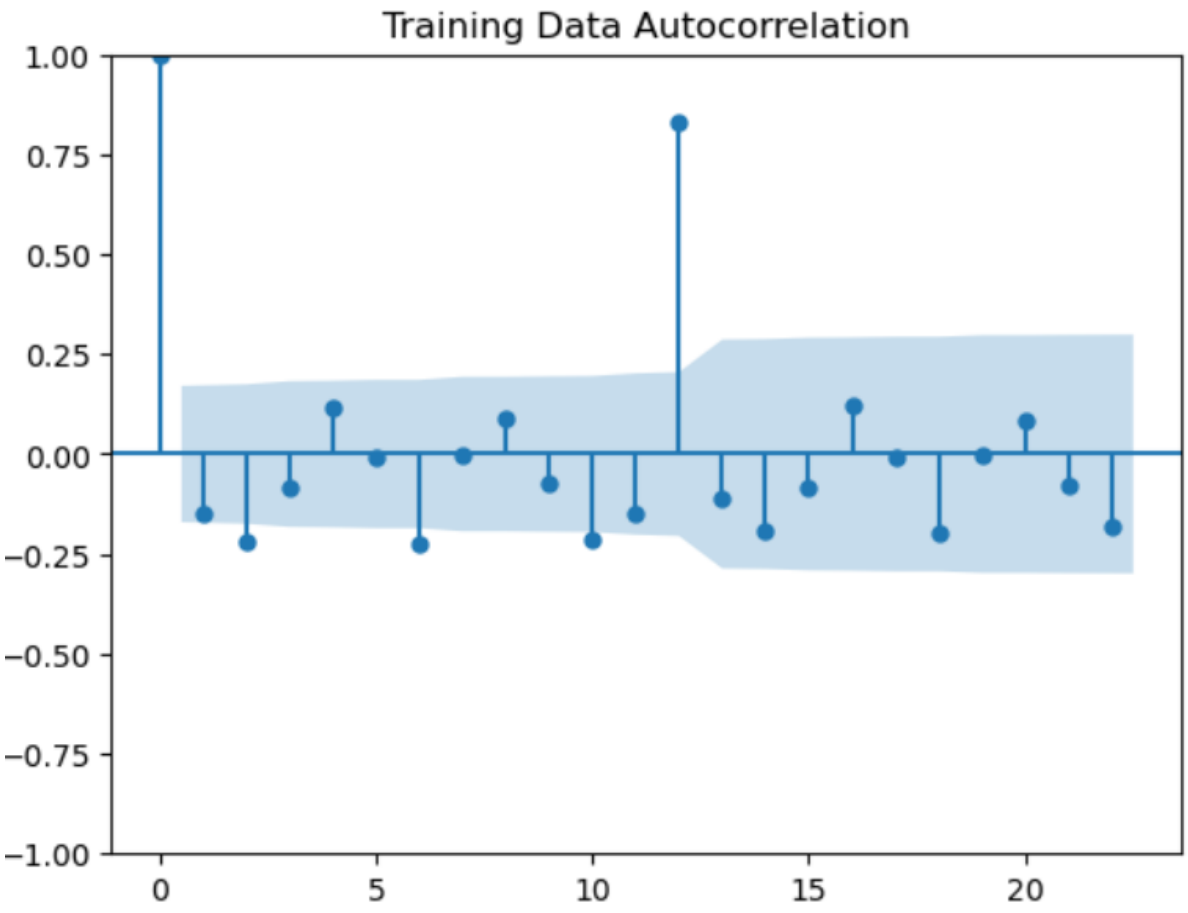
y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1327.322207	388.342361	566.185166	2088.459247
1	1315.065275	402.008370	527.143348	2102.987202
2	1621.521423	402.001984	833.612014	2409.430833
3	1598.817127	407.241740	800.637983	2396.996271
4	1392.629222	407.972228	593.018348	2192.240095

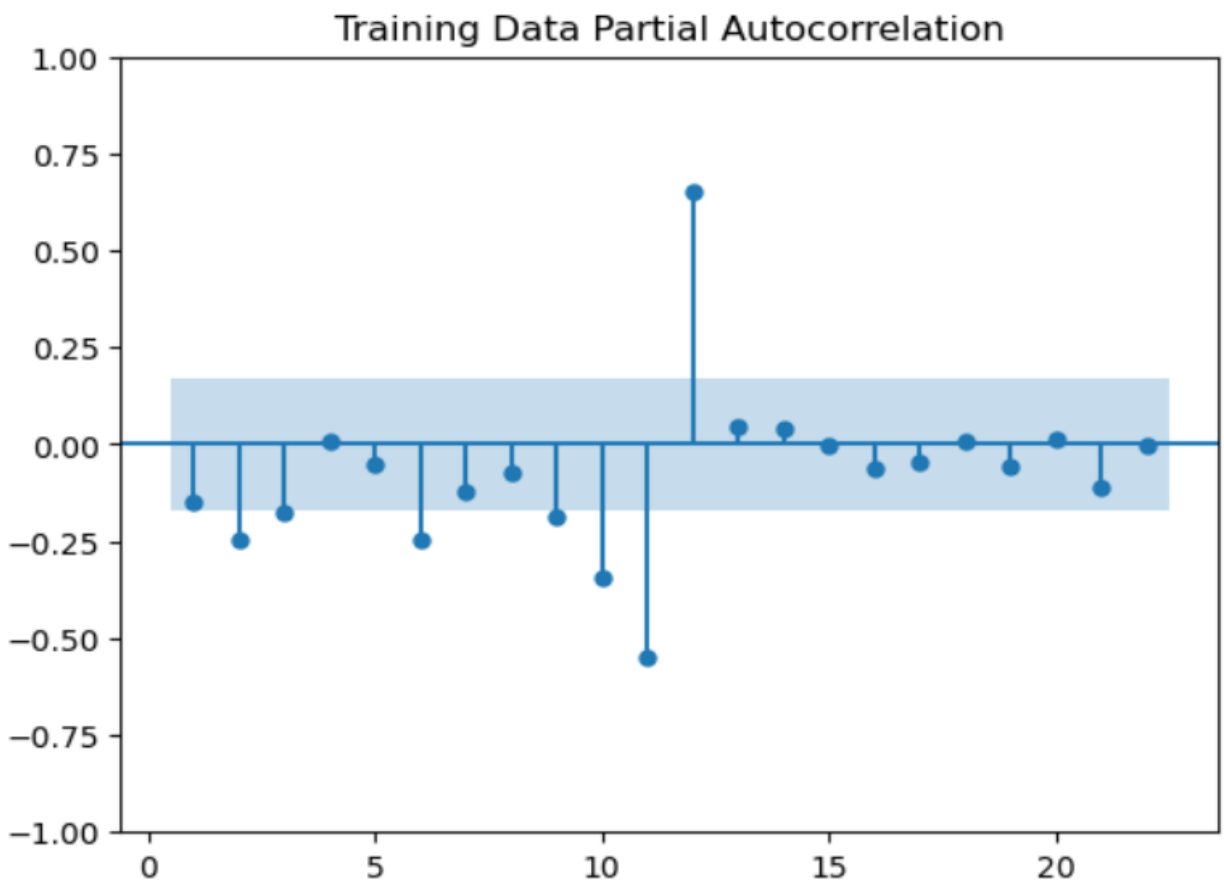
Automated SARIMA model (1,1,2) (1,0,2,12) RMSE score = 528.66

We see that the RMSE value have reduced further when the seasonality parameter was changed to 12.

Q7 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.

Model 9: ARIMA model as per ACF and the PACF cut-off.





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 0.

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	132
Model:	ARIMA(0, 1, 0)	Log Likelihood	-1132.832
Date:	Mon, 26 Dec 2022	AIC	2267.663
Time:	12:49:36	BIC	2270.538
Sample:	01-01-1980 - 12-01-1990	HQIC	2268.831
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
sigma2	1.885e+06	1.29e+05	14.658	0.000	1.63e+06	2.14e+06

Ljung-Box (L1) (Q):	3.07	Jarque-Bera (JB):	198.83
Prob(Q):	0.08	Prob(JB):	0.00
Heteroskedasticity (H):	2.46	Skew:	-1.92
Prob(H) (two-sided):	0.00	Kurtosis:	7.65

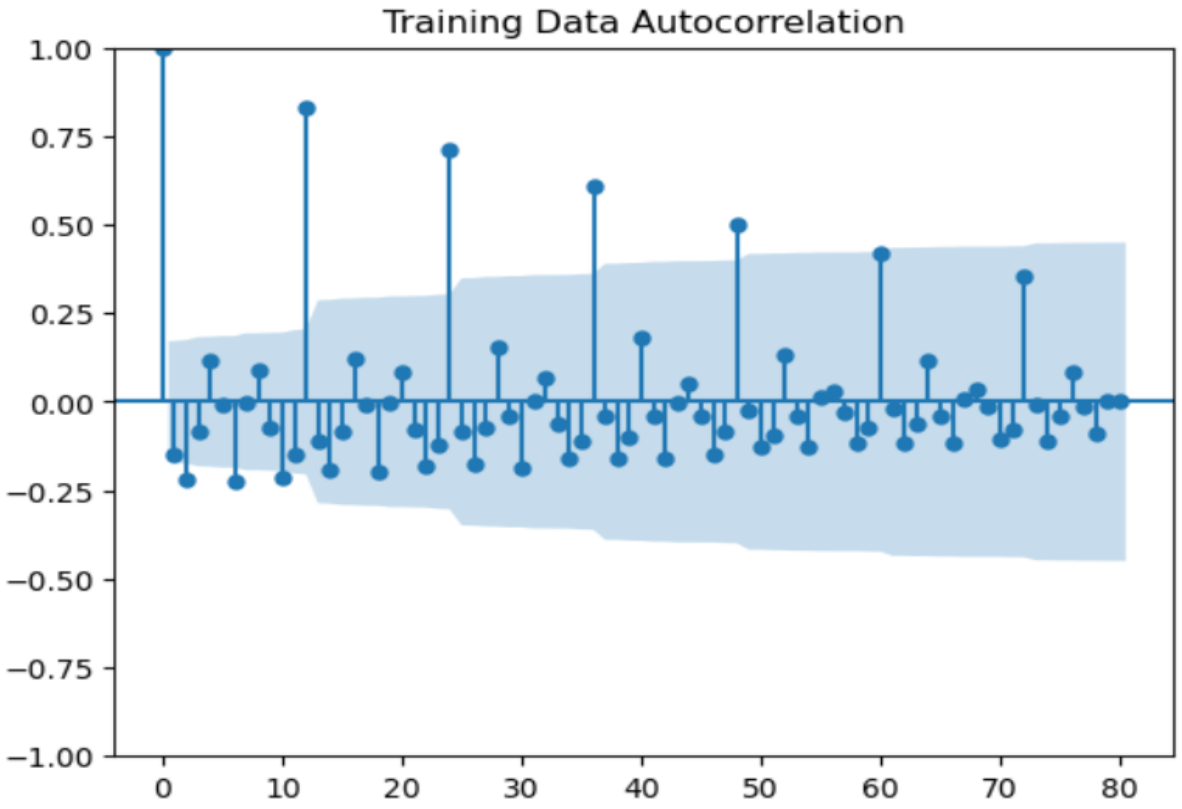
Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

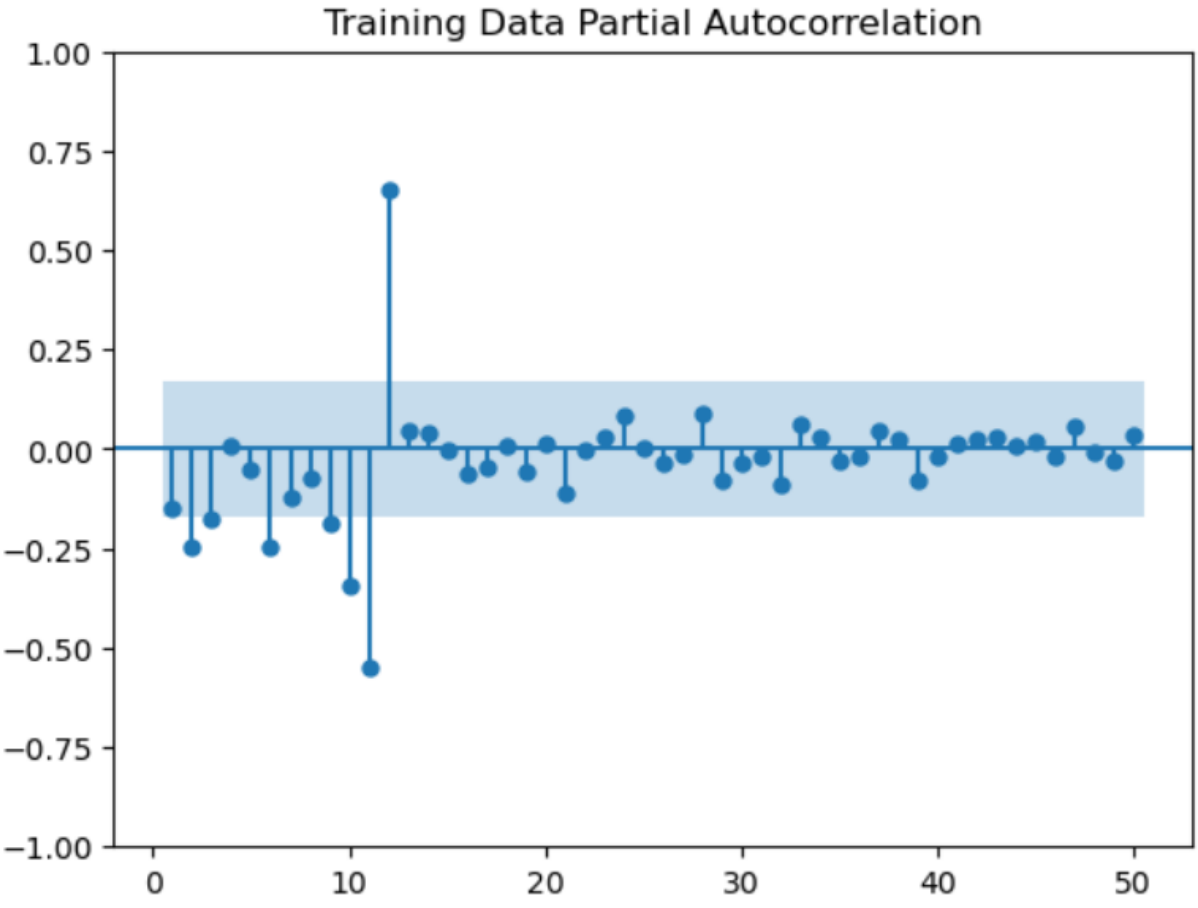
Predict on the Test Set using this model and evaluate the model

Automated SARIMA model (0, 1, 0) RMSE score = 3864.27

Model 10: SARIMA model as per ACF and the PACF cut-off

Let us look at the ACF and the PACF plots once more.


Training Data Autocorrelation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
10
20
30
40
50
60
70
80


Training Data Partial Autocorrelation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
10
20
30
40
50

Here, we have taken alpha=0.05.

We are going to take the seasonal period as 12. We are taking the p value to be 0 and the q value also to be 0 as the parameters same as the ARIMA model.

The Auto-Regressive parameter in a SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 0.

The Moving-Average parameter in a SARIMA model is 'Q' which comes from the significant lag after which the ACF plot cuts-off to 4.

```
                                     SARIMAX Results
==========================================================================================================
Dep. Variable:                                      y   No. Observations:                  132
Model:             SARIMAX(0, 1, 0)x(0, 1, [1, 2, 3, 4], 12)   Log Likelihood                -540.038
Date:                                Mon, 26 Dec 2022   AIC                             1090.076
Time:                                        12:50:03   BIC                             1101.319
Sample:                                             0   HQIC                            1094.542
                                                  - 132
Covariance Type:                                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.S.L12      -0.4508      0.420     -1.074      0.283      -1.274       0.372
ma.S.L24      -0.0624      0.262     -0.239      0.811      -0.575       0.450
ma.S.L36      -0.1686      0.243     -0.694      0.488      -0.645       0.308
ma.S.L48      -0.1186      0.297     -0.400      0.689      -0.700       0.463
sigma2      2.829e+05   1.21e+05      2.342      0.019    4.62e+04     5.2e+05
==============================================================================
Ljung-Box (L1) (Q):                   6.78   Jarque-Bera (JB):                17.59
Prob(Q):                              0.01   Prob(JB):                         0.00
Heteroskedasticity (H):               0.29   Skew:                             0.82
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.83
==============================================================================

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

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

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1457.309895	540.491101	397.966802	2516.652987
1	1137.104669	763.100057	-358.543959	2632.753297
2	1621.987570	934.083916	-208.783265	3452.758404
3	1337.133393	1078.287452	-776.271177	3450.537964
4	1375.211644	1205.360898	-987.252305	3737.675593

Manual SARIMA model (0,1,0) (0,1,4,12) RMSE score = 624.51

Q8 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.

We have created a Data Frame named as - resultsDf and below is the output of that data.

	Test RMSE
Alpha=0.049,SimpleExponentialSmoothing	1316.035487
Alpha =0.688, Beta=9.999, DoubleExponentialSmoothing	2007.238526
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	18259.110704
Alpha=0.111,Beta=0.012,Gamma=0.460,TripleExponentialSmoothing	378.951023
NaiveModel	3864.279352
RegressionOnTime	1389.135175
SimpleAverageModel	1275.081804
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315
ARIMA(2,1,2)	1299.980041
SARIMA(1,1,2)(2,0,2,6)	626.473600
SARIMA(1,1,2)(1,0,2,12)	528.664707
ARIMA(0,1,0)	3864.279352
SARIMA(0,1,0)(0,1,4,12)	624.511337

Q9 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.

We have built multiple models and based on the minimum RMSE values these are the best models.

First: Alpha=0.111, Beta=0.012, Gamma=0.460, Triple Exponential Smoothing

Second: SARIMA (1,1,2) (1,0,2,12)

So will build these 2 models on the Full data.

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

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	187
Model:	SARIMAX(1, 1, 2)x(1, 0, 2, 12)	Log Likelihood	-1173.413
Date:	Sun, 25 Dec 2022	AIC	2360.827
Time:	13:11:43	BIC	2382.309
Sample:	01-01-1980	HQIC	2369.551
	- 07-01-1995		
Covariance Type:	opg		

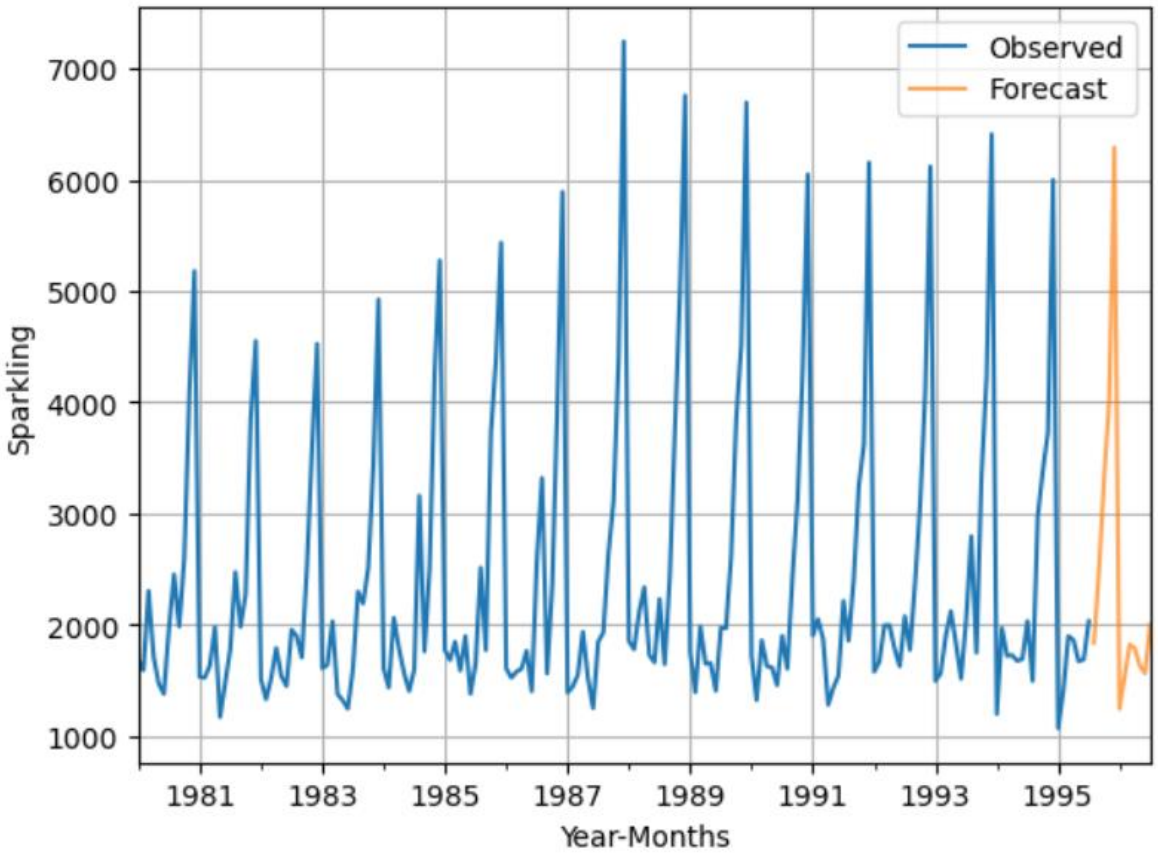
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.6609	0.242	-2.733	0.006	-1.135	-0.187
ma.L1	-0.2739	0.200	-1.368	0.171	-0.666	0.118
ma.L2	-0.8112	0.227	-3.576	0.000	-1.256	-0.367
ar.S.L12	1.0157	0.012	84.457	0.000	0.992	1.039
ma.S.L12	-1.3873	0.338	-4.102	0.000	-2.050	-0.724
ma.S.L24	-0.1461	0.146	-1.001	0.317	-0.432	0.140
sigma2	5.948e+04	1.84e+04	3.232	0.001	2.34e+04	9.56e+04

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	27.47
Prob(Q):	0.96	Prob(JB):	0.00
Heteroskedasticity (H):	1.03	Skew:	0.52
Prob(H) (two-sided):	0.93	Kurtosis:	4.76

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

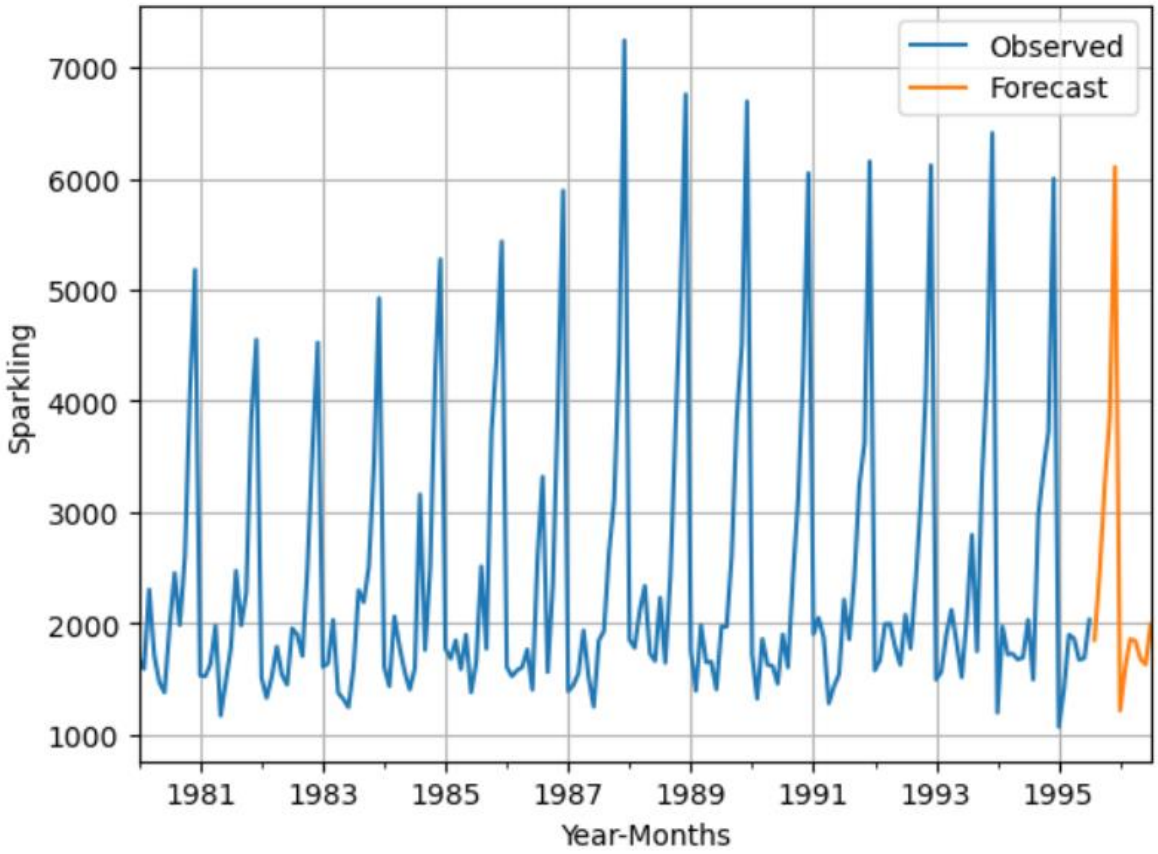
Sparkling	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-01	1836.358925	379.711435	1092.138187	2580.579663
1995-09-01	2489.607781	384.474678	1736.051258	3243.164303
1995-10-01	3324.589777	384.580437	2570.825972	4078.353583
1995-11-01	4020.231720	386.338203	3263.022756	4777.440683
1995-12-01	6290.002543	386.392767	5532.686636	7047.318449
1996-01-01	1244.698734	387.302892	485.599014	2003.798455
1996-02-01	1533.143141	387.531253	773.595842	2292.690440
1996-03-01	1821.710238	388.158130	1060.934283	2582.486192
1996-04-01	1788.499051	388.498154	1027.056661	2549.941441
1996-05-01	1627.572403	389.017140	865.112819	2390.031986
1996-06-01	1563.325860	389.413000	800.090406	2326.561315
1996-07-01	2000.709048	389.887486	1236.543618	2764.874478

RMSE of SARIMA model built on full data is- 539.9824921681942



Alpha=0.111, Beta=0.012, Gamma=0.460, Triple Exponential Smoothing Model

	lower_CI	prediction	upper_ci
1995-08-01	1127.564038	1851.014405	2574.464772
1995-09-01	1731.765860	2455.216227	3178.666594
1995-10-01	2522.446436	3245.896803	3969.347170
1995-11-01	3150.038433	3873.488800	4596.939167
1995-12-01	5379.092987	6102.543354	6825.993721



RMSE of Triple Exponential Smoothing Model built on full data is- 368.11

So, we can see that Triple Exponential Smoothing Model perform very well on the data as it has given low RMSE.

Q10 Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

1. We have built total 11 models and stored Test RMSE values in the data frame - resultsDf.

2. After checking All RMSE values we can say that Triple Exponential Smoothing gives the lowest RMSE Value.

3. As per the final plot on complete data we can say that our model is able to forecast very well.

4. We have seen that past data has more influence in the forecast so more the past data better the forecast.

5. we have seen that sales of wine goes up at the end of the year possibly due to festival seasons so for the other months where sales are low company can provide different offers.

6. From our Full data Model we can see that sales for the end of year 1995 is higher compare to the other months and same we obtained in our original data.

7. Other model such as Automated SARIMA model also doing good because it crosschecked different combination of parameters and with lowest AIC value we build the model, as AIC take care of complexity of model and RMSE take care of Accuracy.

8. Based on the SARIMA model which built on Full data we got Jarque-Bera test's P-value which is 0 so we reject the H0 & we found that Data is not normal.

9. Based on the Ljung-Box test's P-value we failed to reject the H0 so we can say that residuals are independent.

10. With the help of next 12 months forecast sales company should make some plan to do more sales in those months where sales are low compare to end of the year they should focus more on the months before September.

Problem2-Rose

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 Description:

1. YearMonth : Date and year
2. Rose : Sales of Rose wine

Sample of the dataset:

YearMonth	Rose
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0

We have Converted column YearMonth into index.

Exploratory Data Analysis:

Let us check the types of variables in the data frame.

As we have changed YearMonth column into index so only one column present in the data which is the 'Rose' and data type of this column is integer.

Checking for missing values in the dataset:

```
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   Rose    185 non-null    float64
dtypes: float64(1)
memory usage: 2.9 KB
```

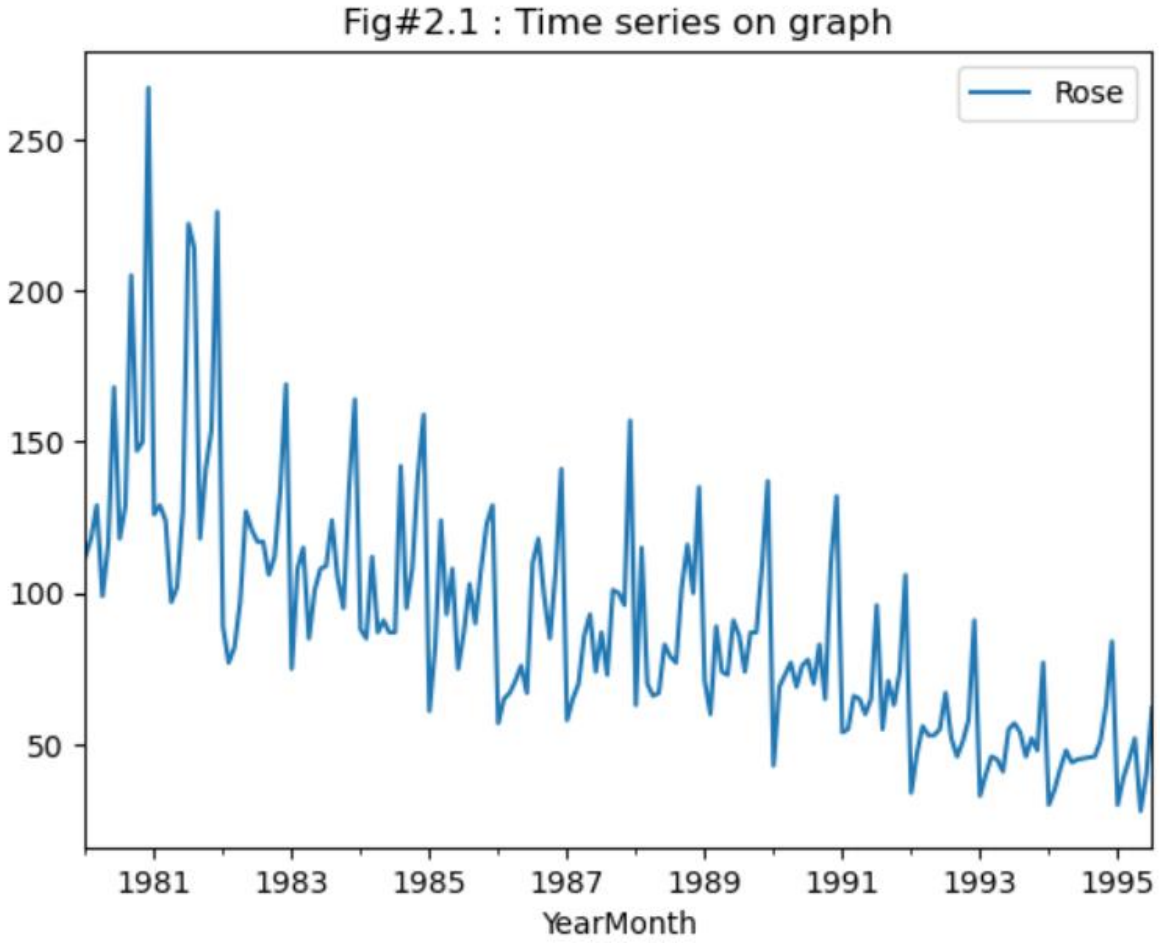
We can see that there are null values present in the data so will impute them.

Q2.1 Read the data as an appropriate Time Series data and plot the data.

Sample of the time series data.

YearMonth	Rose
1995-03-01	45.0
1995-04-01	52.0
1995-05-01	28.0
1995-06-01	40.0
1995-07-01	62.0

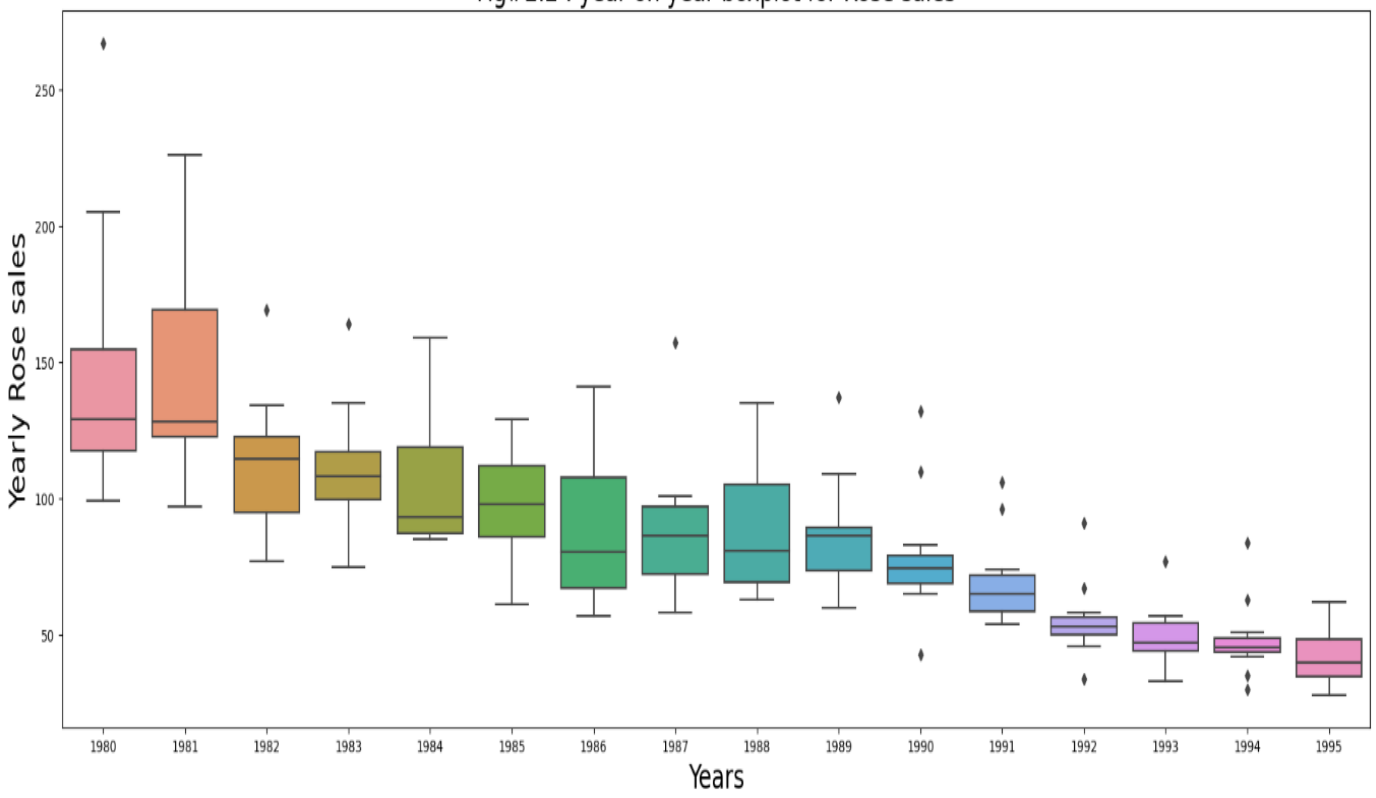
We are imputing the null values by using interpolate python function.



As per the graph Rose wine sales hit maximum numbers in year 1981.

Q2.2 Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

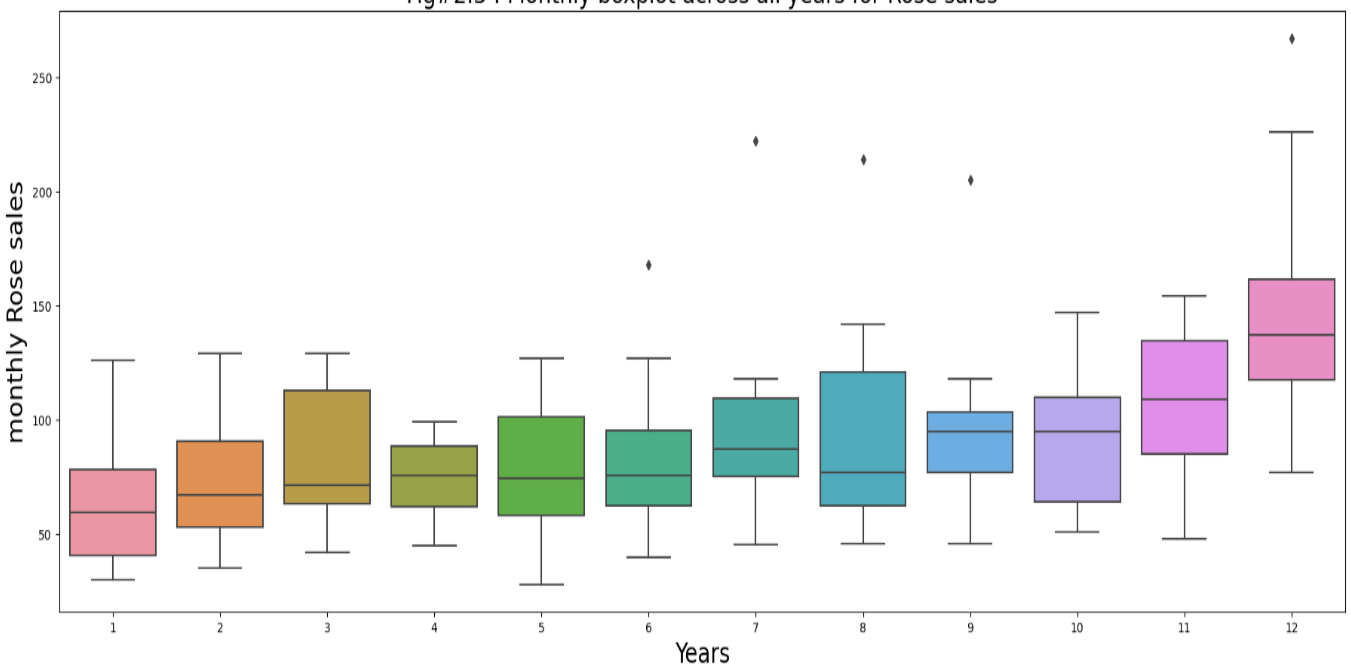
Fig#2.2 : year on year boxplot for Rose sales



We can see that we have data from past 16 years.

As we got to know from the Time Series plot, the boxplots over here indicate trend present. Also, we see that the sales of wine have some outliers for most of years.

Fig#2.3 : Monthly boxplot across all years for Rose sales

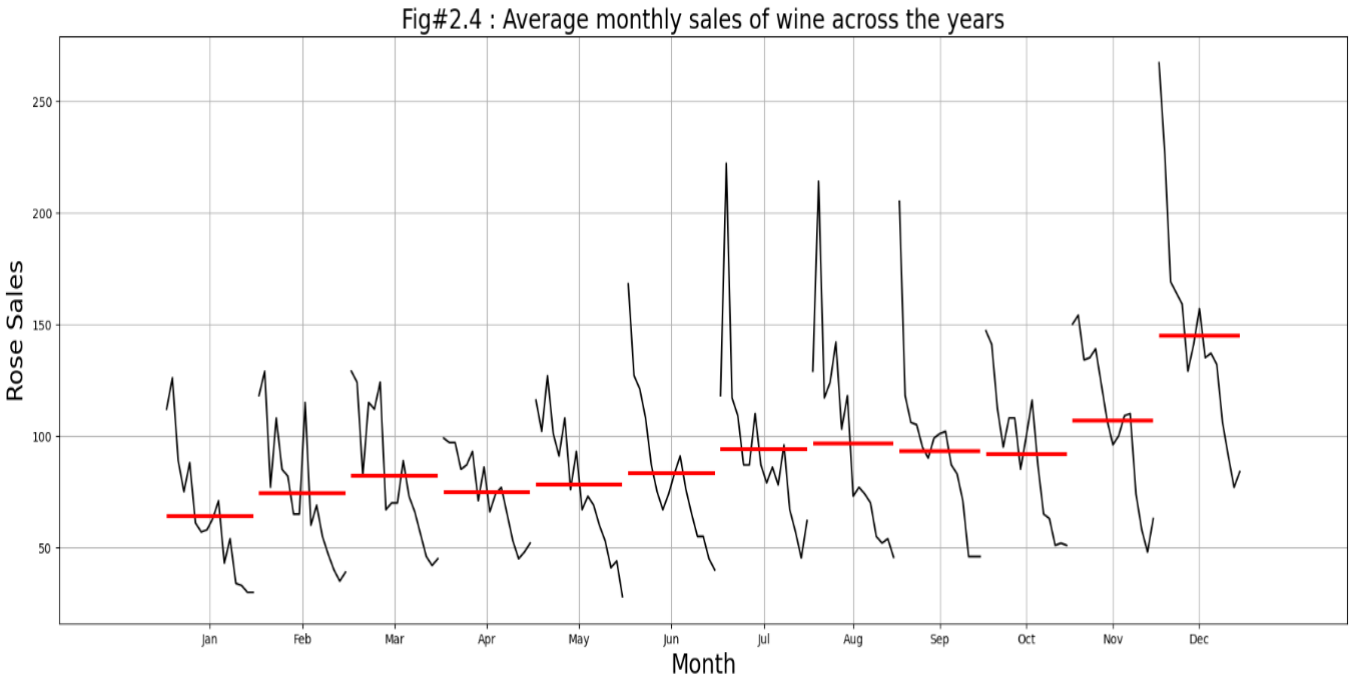


The boxplots for the monthly production for different years have outliers.

We can see that sales of Rose wine have some increase at the end of the year.

Plot a month plot of the given Time Series

Fig#2.4 : Average monthly sales of wine across the years



Here the red line shows the average wine sales for all the years in a particular month.

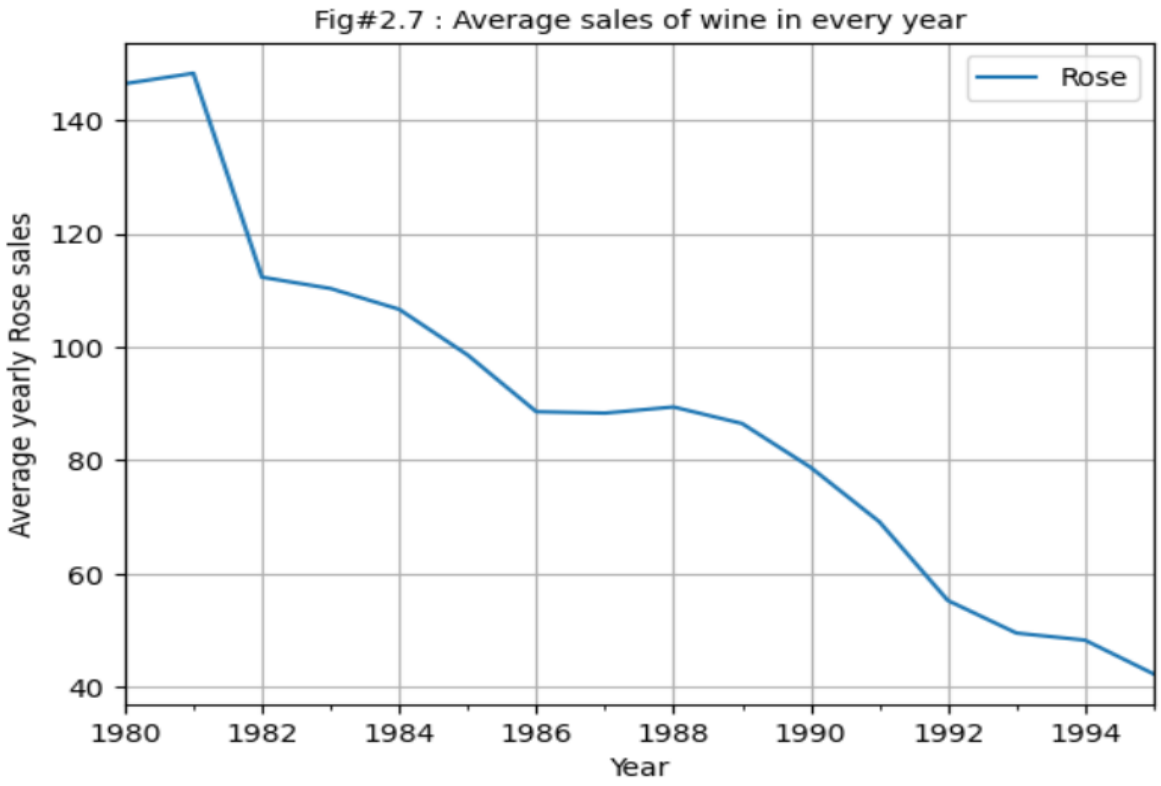
Fig#2.5 : monthly sales of wine across the years

We can see that sales are maximum for December month.

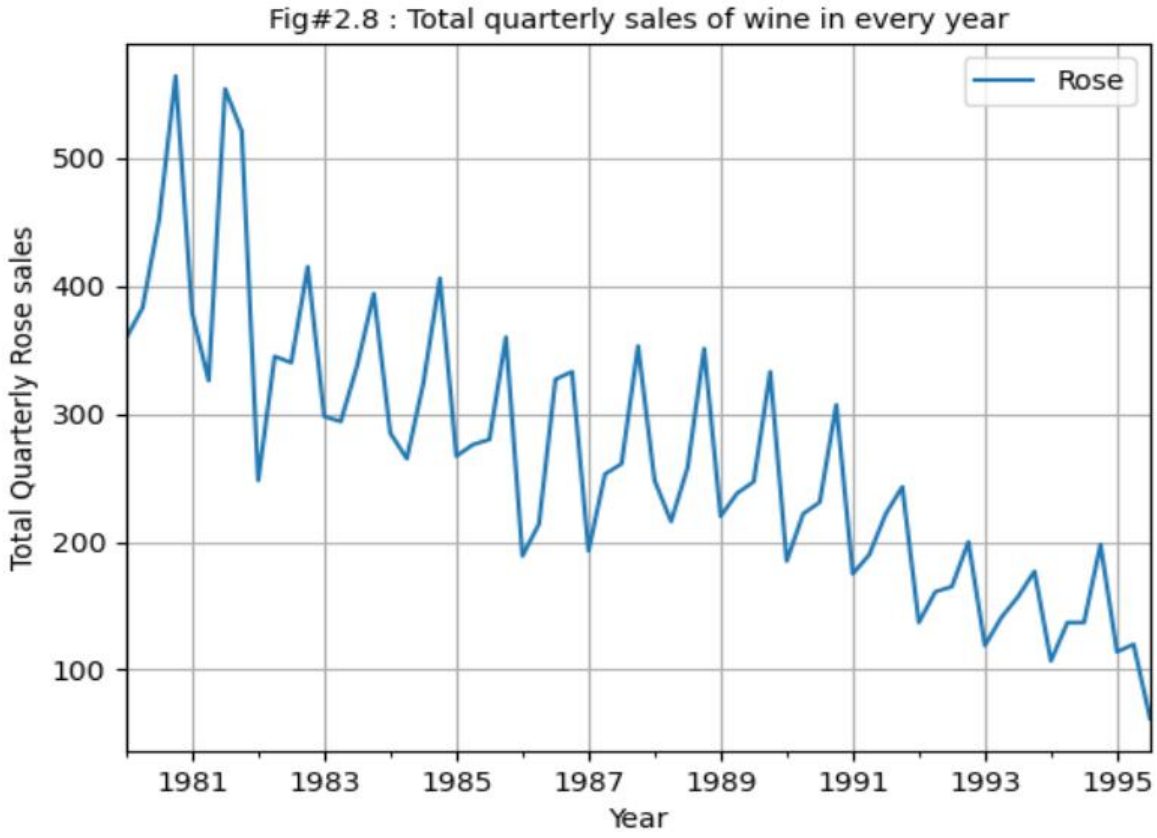
Read this monthly data into a quarterly and yearly format

Yearly plot

Fig#2.6 : Total sales of wine in every year

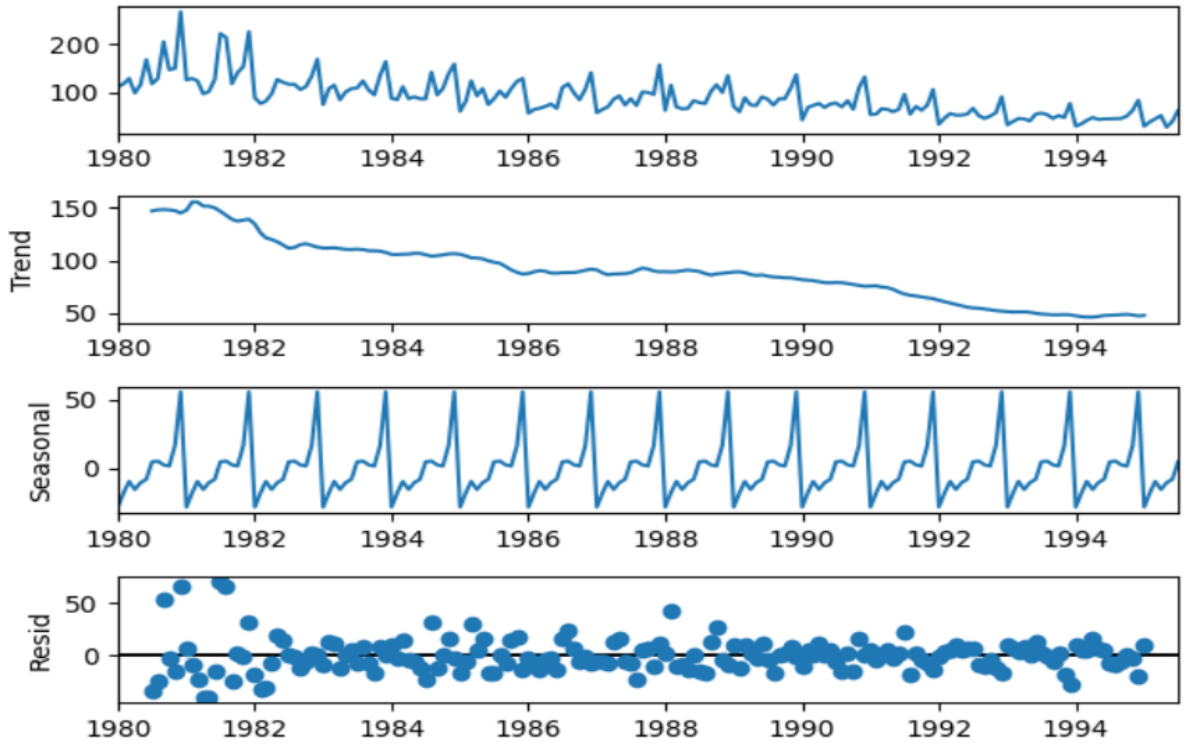


Quarterly plot



Decompose the Time Series

Additive Model



We can see that trend & seasonality present in the time series data.

We can see that residuals are random

Trend

```
 YearMonth
1980-01-01           NaN
1980-02-01           NaN
1980-03-01           NaN
1980-04-01           NaN
1980-05-01           NaN
1980-06-01           NaN
1980-07-01    147.083333
1980-08-01    148.125000
1980-09-01    148.375000
1980-10-01    148.083333
1980-11-01    147.416667
1980-12-01    145.125000
Name: trend, dtype: float64
```

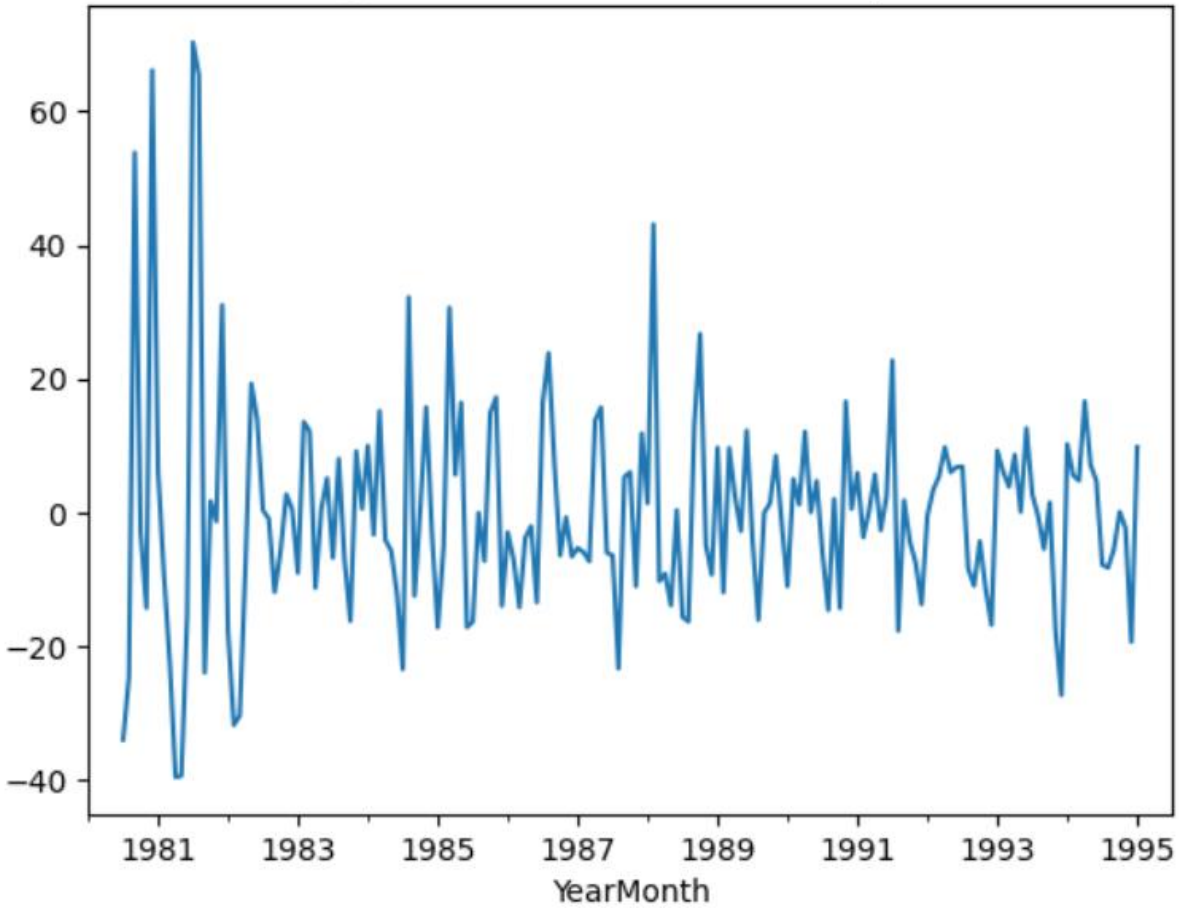
Seasonality

```
 YearMonth
1980-01-01   -27.908647
1980-02-01   -17.435632
1980-03-01    -9.285830
1980-04-01   -15.098330
1980-05-01   -10.196544
1980-06-01    -7.678687
1980-07-01     4.896908
1980-08-01     5.499686
1980-09-01     2.774686
1980-10-01     1.871908
1980-11-01    16.846908
1980-12-01    55.713575
Name: seasonal, dtype: float64
```

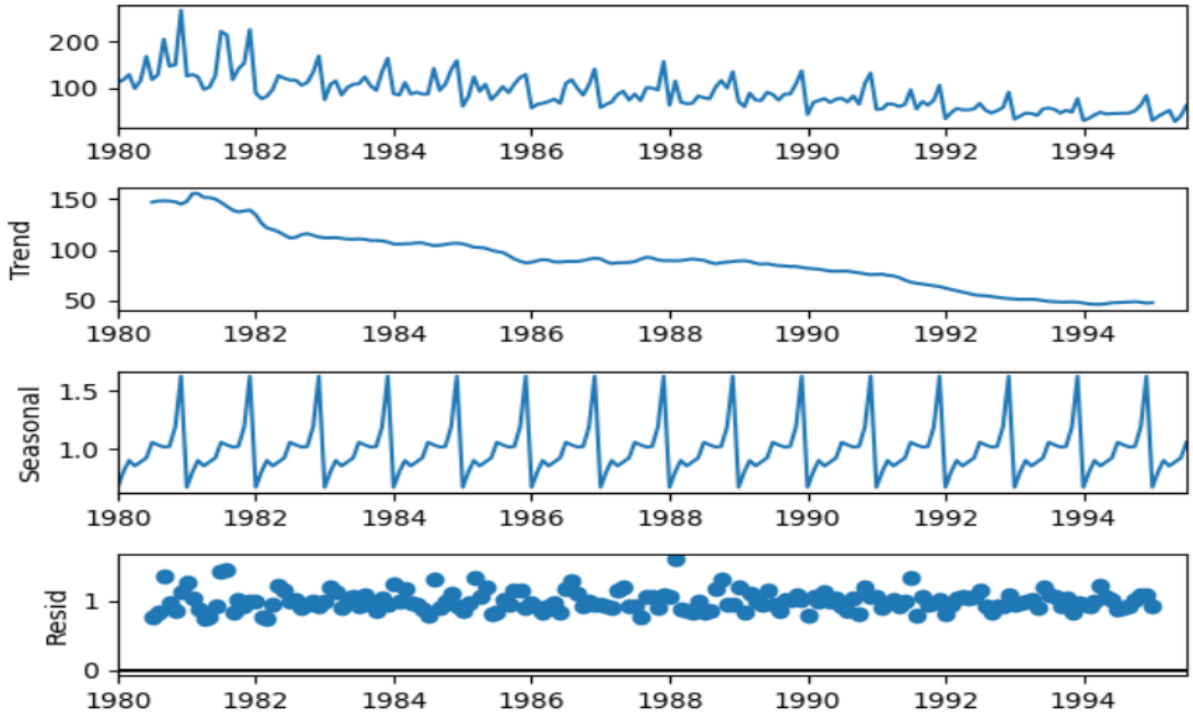
Residual

```
 YearMonth
1980-01-01          NaN
1980-02-01          NaN
1980-03-01          NaN
1980-04-01          NaN
1980-05-01          NaN
1980-06-01          NaN
1980-07-01   -33.980241
1980-08-01   -24.624686
1980-09-01    53.850314
1980-10-01    -2.955241
1980-11-01   -14.263575
1980-12-01    66.161425
Name: resid, dtype: float64
```

Fig#2.10 :Additive models Residual plot



Multiplicative Model



as per multiplicative model We can see that Trend & seasonality present in the time series data

Trend

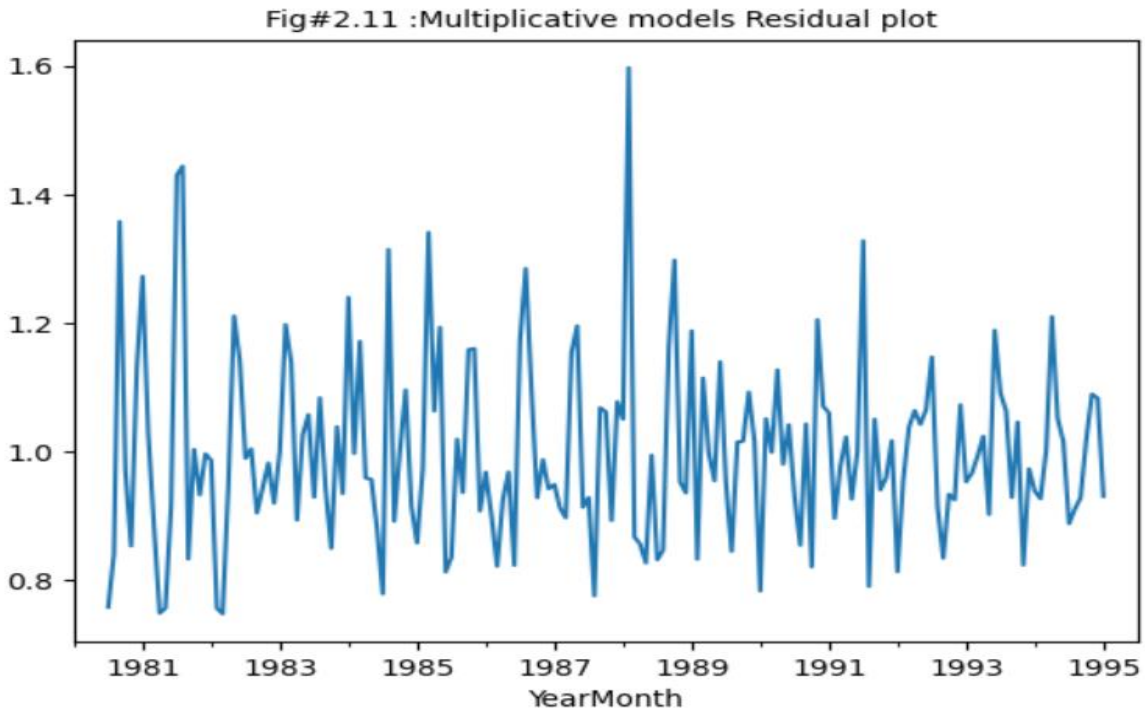
```
 YearMonth
1980-01-01           NaN
1980-02-01           NaN
1980-03-01           NaN
1980-04-01           NaN
1980-05-01           NaN
1980-06-01           NaN
1980-07-01    147.083333
1980-08-01    148.125000
1980-09-01    148.375000
1980-10-01    148.083333
1980-11-01    147.416667
1980-12-01    145.125000
Name: trend, dtype: float64
```

Seasonality

```
 YearMonth
1980-01-01    0.670111
1980-02-01    0.806163
1980-03-01    0.901164
1980-04-01    0.854024
1980-05-01    0.889415
1980-06-01    0.923985
1980-07-01    1.058038
1980-08-01    1.035881
1980-09-01    1.017648
1980-10-01    1.022573
1980-11-01    1.192349
1980-12-01    1.628646
Name: seasonal, dtype: float64
```

Residual

```
 YearMonth
1980-01-01         NaN
1980-02-01         NaN
1980-03-01         NaN
1980-04-01         NaN
1980-05-01         NaN
1980-06-01         NaN
1980-07-01    0.758258
1980-08-01    0.840720
1980-09-01    1.357674
1980-10-01    0.970771
1980-11-01    0.853378
1980-12-01    1.129646
Name: resid, dtype: float64
```



We can see that in both residual's plot residuals are random as it does not follow any pattern and so there is no clear merit to specifically choose multiplicative decomposition as additive is good enough.

Q2.3 Split the data into training and test. The test data should start in 1991.

First few rows of Training Data

YearMonth	Rose
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0

Last few rows of Training Data

YearMonth	Rose
1990-08-01	70.0
1990-09-01	83.0
1990-10-01	65.0
1990-11-01	110.0
1990-12-01	132.0

First few rows of Test Data

YearMonth	Rose
1991-01-01	54.0
1991-02-01	55.0
1991-03-01	66.0
1991-04-01	65.0
1991-05-01	60.0

Last few rows of Test Data

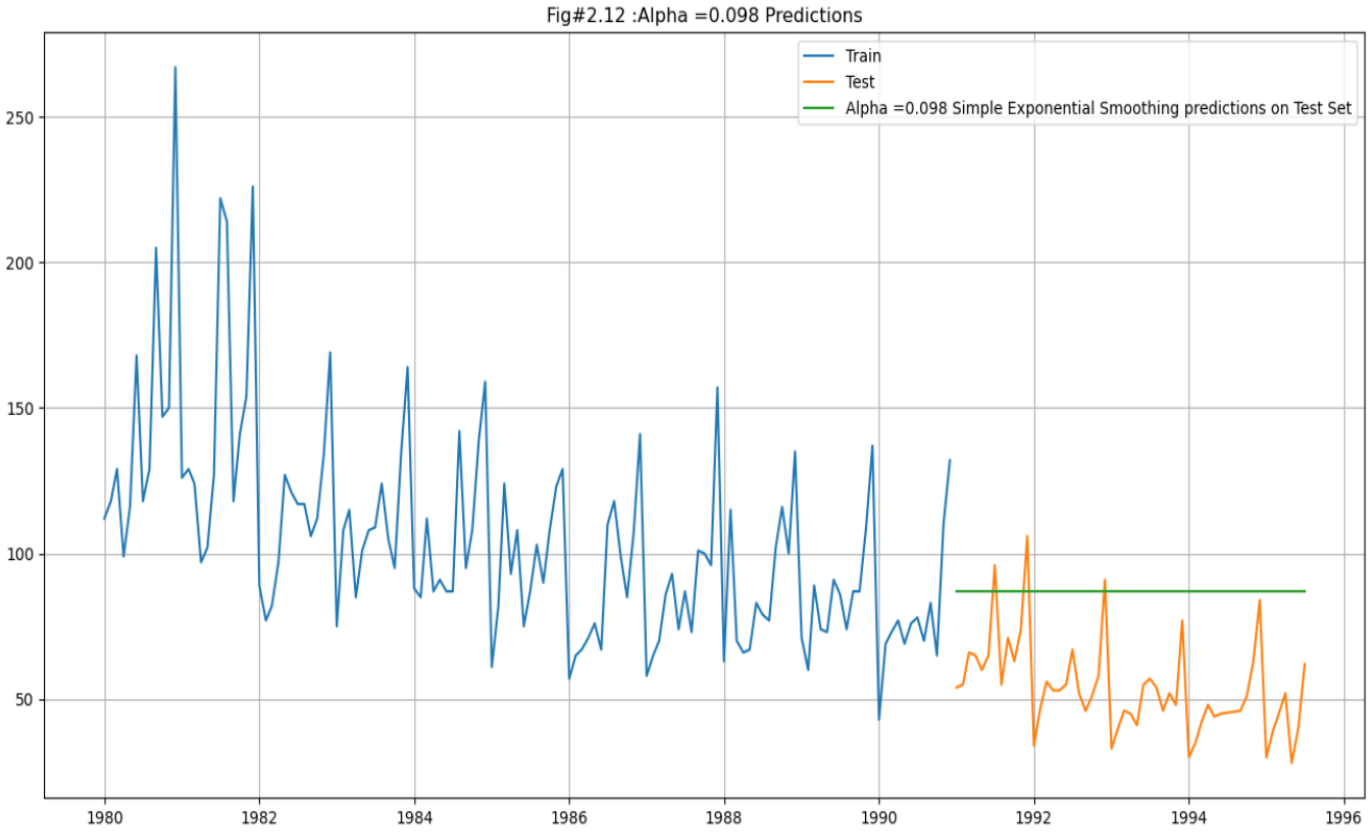
YearMonth	Rose
1995-03-01	45.0
1995-04-01	52.0
1995-05-01	28.0
1995-06-01	40.0
1995-07-01	62.0

Q2.4 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.

Model 1: Simple Exponential Smoothing

First, we are building SES model by using default Alpha value (smoothing level) which is - 0.098

YearMonth	Rose	predict
1991-01-01	54.0	87.104983
1991-02-01	55.0	87.104983
1991-03-01	66.0	87.104983
1991-04-01	65.0	87.104983
1991-05-01	60.0	87.104983



For Alpha =0.098 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 36.796

checking different alpha values

After trying with different alpha values, we got following RMSE

	Alpha Values	Train RMSE	Test RMSE
0	0.3	32.470164	47.504821
1	0.4	33.035130	53.767406
2	0.5	33.682839	59.641786
3	0.6	34.441171	64.971288
4	0.7	35.323261	69.698162
5	0.8	36.334596	73.773992
6	0.9	37.482782	77.139276

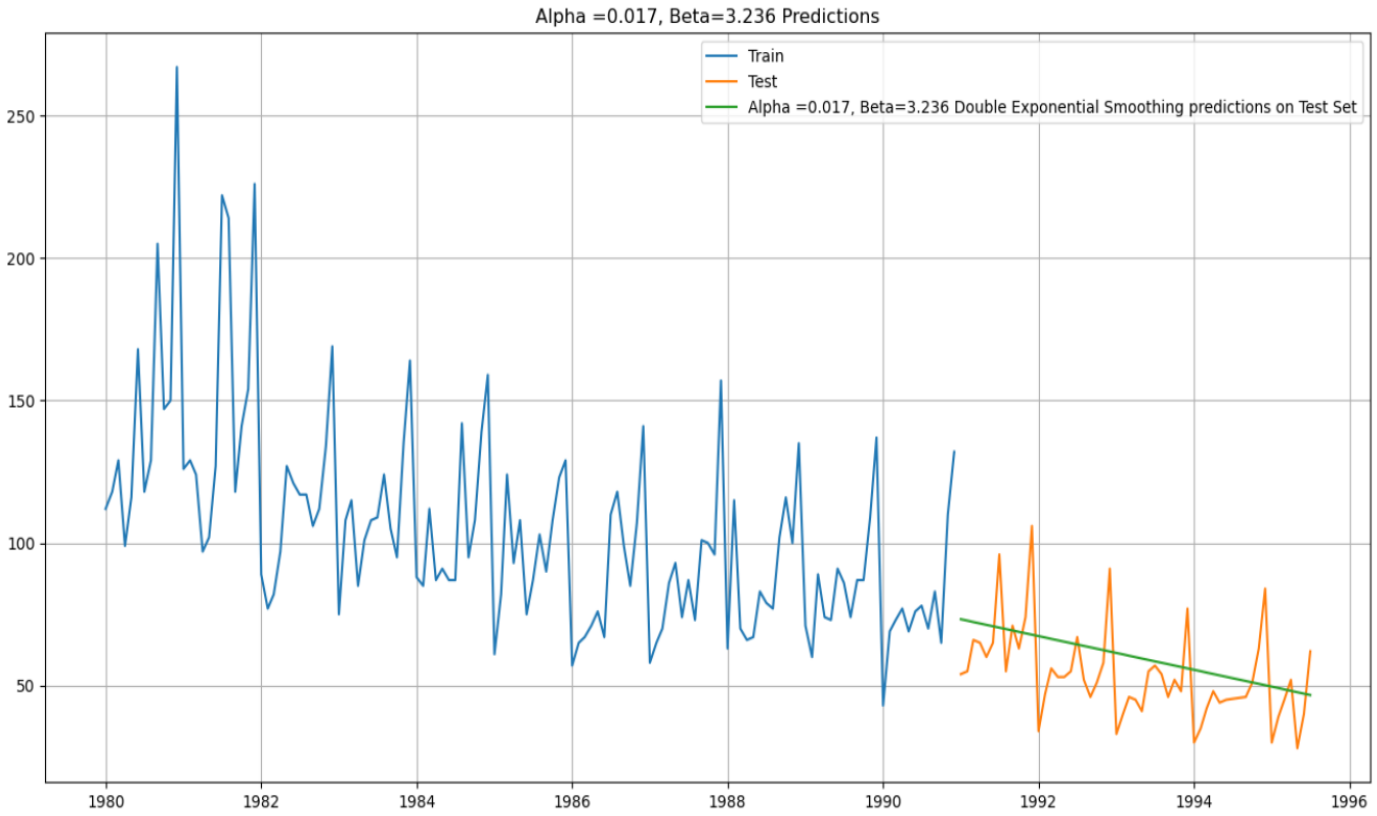
After checking different value of alpha, we are unable to get RMSE less than 36.79 which we obtained when Alpha = 0.098

With Alpha =0.098 we can say that past observations have a large influence on forecasts.

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

First, we are building DES model by using default Alpha value (smoothing level) which is – 0.017 and Beta value (smoothing trend) is - 3.236

	Rose	predict
YearMonth		
1991-01-01	54.0	73.259732
1991-02-01	55.0	72.767150
1991-03-01	66.0	72.274569
1991-04-01	65.0	71.781987
1991-05-01	60.0	71.289405

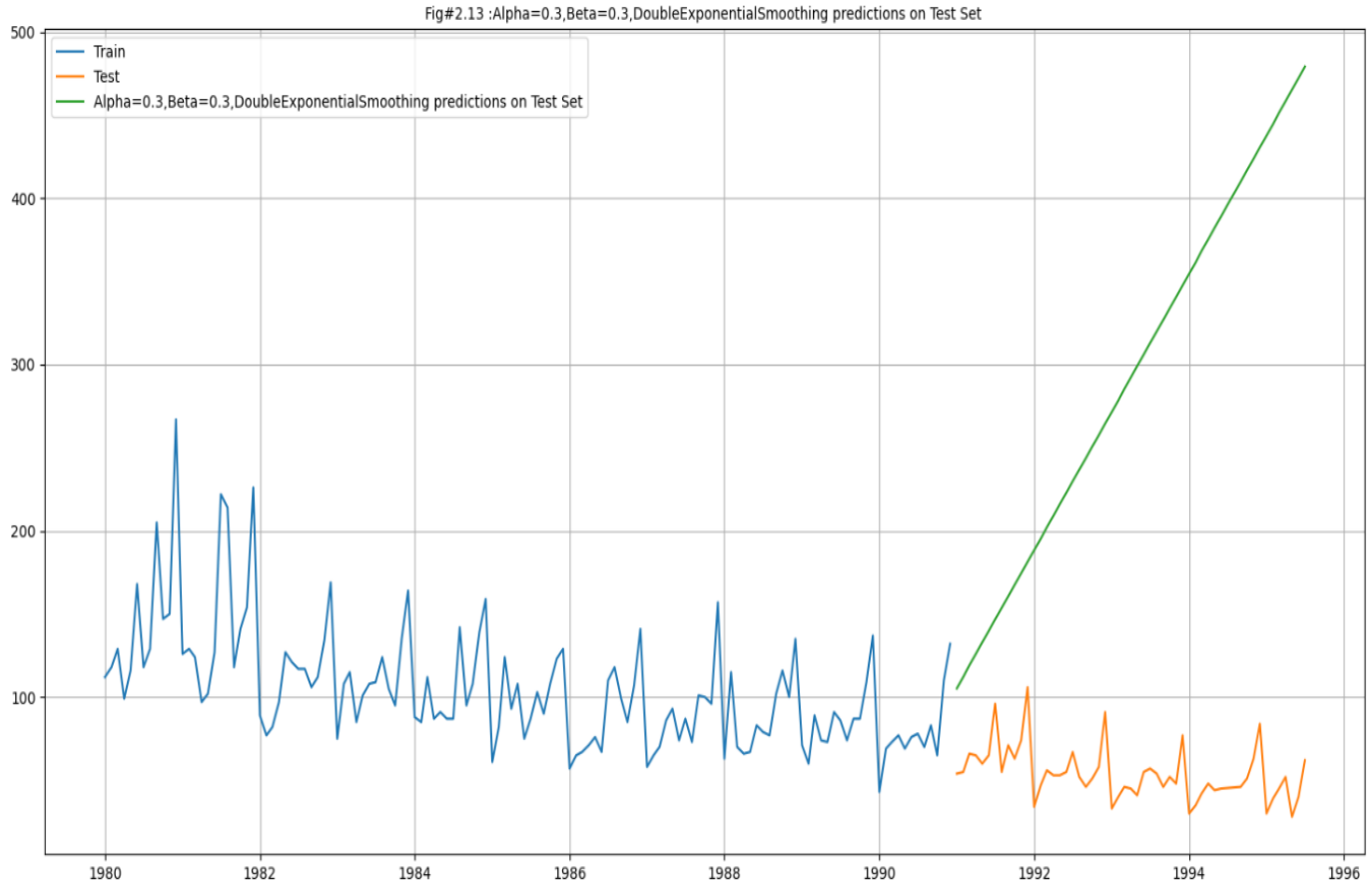


For Alpha =0.017, Beta=3.236 Double Exponential Smoothing Model forecast on the Test Data, RMSE is 15.707

Checking different alpha & Beta values:

After trying with different alpha & Beta values, we got following RMSE.

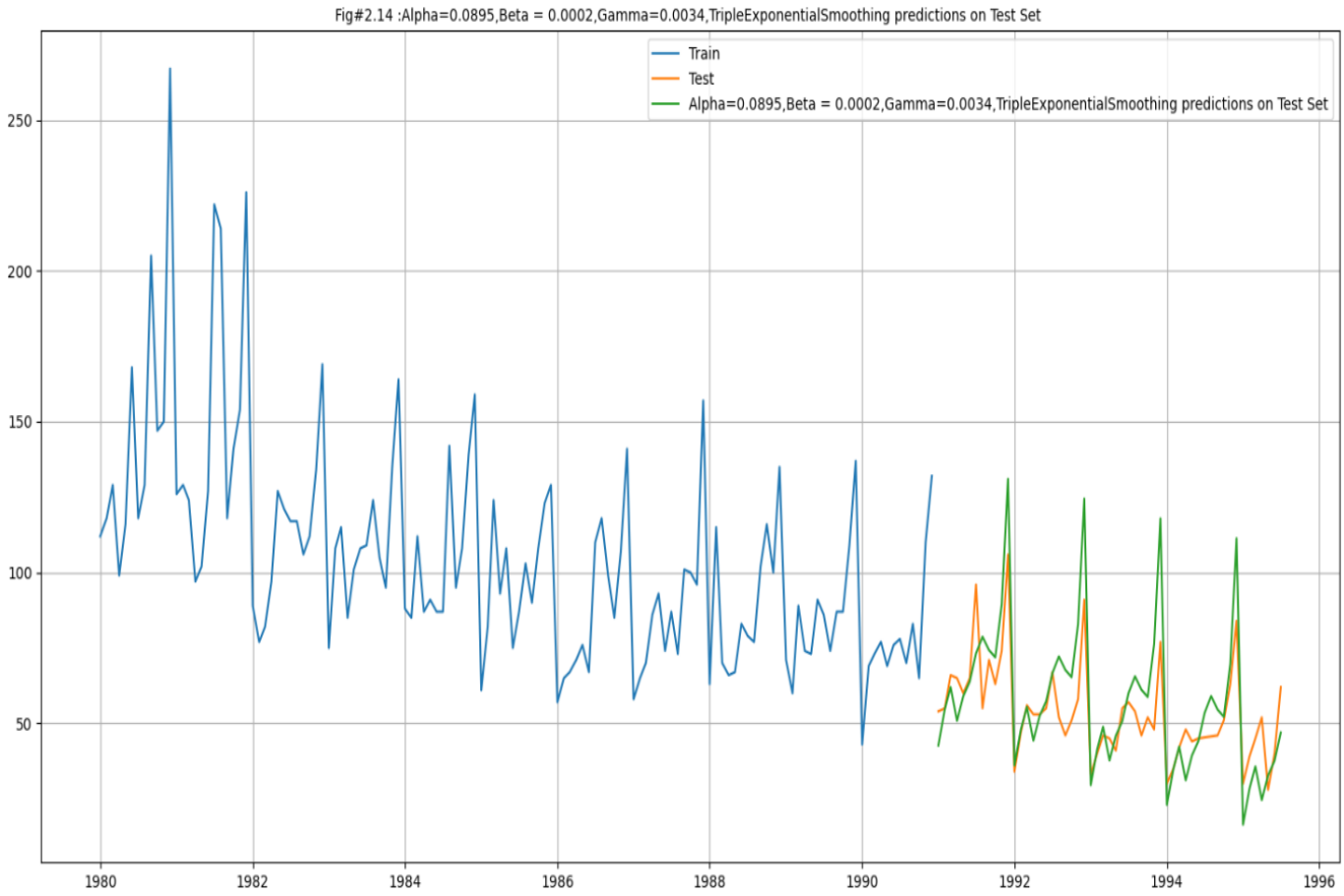
	Alpha Values	Beta Values	Train RMSE	Test RMSE
0	0.3	0.3	35.944983	265.567594
8	0.4	0.3	36.749123	339.306534
1	0.3	0.4	37.393239	358.750942
16	0.5	0.3	37.433314	394.272629
24	0.6	0.3	38.348984	439.296033



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

First, we are building TES model by using default Alpha value (smoothing level) which is – 0.0895, Beta value (smoothing trend) is - 0.0002 & Gamma value (smoothing seasonal) is- 0.0034

YearMonth	Rose	auto_predict
1991-01-01	54.0	42.684928
1991-02-01	55.0	54.564005
1991-03-01	66.0	61.995209
1991-04-01	65.0	50.852018
1991-05-01	60.0	59.034271



For Alpha=0.0895, Beta = 0.0002, Gamma=0.0034, Triple Exponential Smoothing Model fore cast on the Test Data, RMSE is 14.250

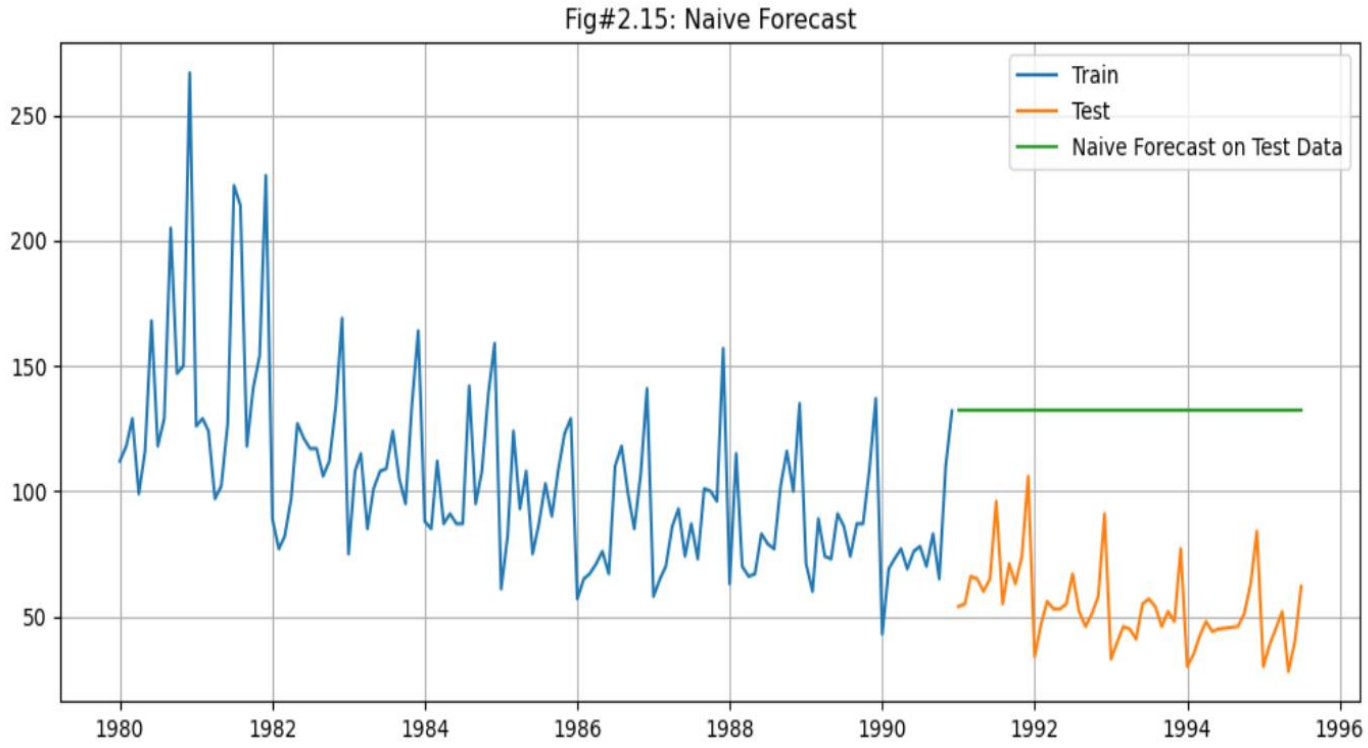
Checking different alpha, Beta & Gamma values:

After trying with different alpha, Beta & gamma values, we got following RMSE.

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE
1	0.3	0.3	0.4	25.308660	12.723156
0	0.3	0.3	0.3	24.279693	12.950648
64	0.4	0.3	0.3	25.447354	13.293449
129	0.5	0.3	0.4	27.934042	17.883742
200	0.6	0.4	0.3	29.455453	23.408349

After checking different values of alpha, Beta & Gamma we are able to get RMSE less than 14.24

Model 4: Naive forecast



For Naive forecast on the Test Data, RMSE is 79.719

Model 5: Linear Regression

For regression first we need to add numerical time instance order in the data and here is the sample of that data.

First few rows of Training Data

YearMonth	Rose	time
1980-01-01	112.0	1
1980-02-01	118.0	2
1980-03-01	129.0	3
1980-04-01	99.0	4
1980-05-01	116.0	5

Last few rows of Training Data

YearMonth	Rose	time
1990-08-01	70.0	128

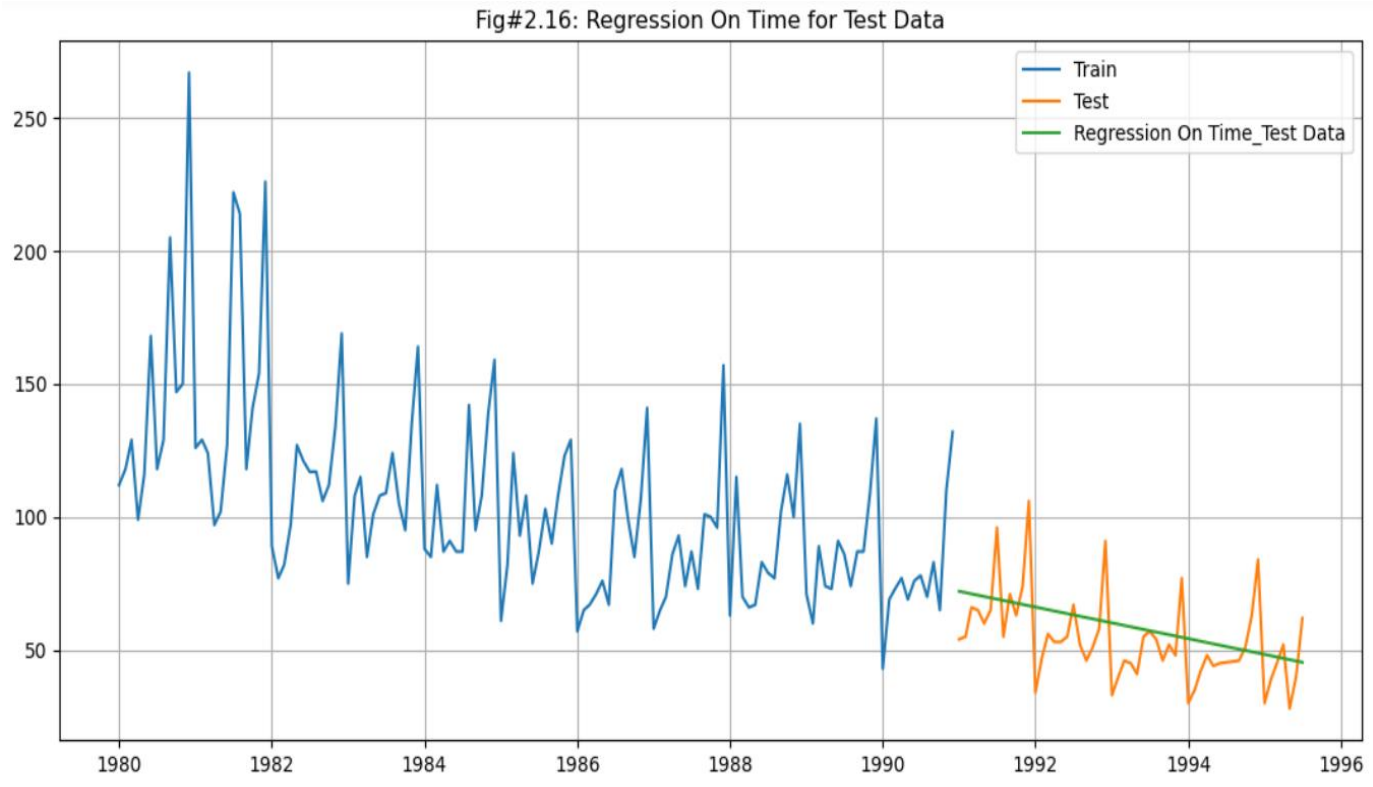
```
1990-09-01   83.0   129
1990-10-01   65.0   130
1990-11-01  110.0   131
1990-12-01  132.0   132
```

First few rows of Test Data

```
             Rose  time
YearMonth
1991-01-01  54.0   133
1991-02-01  55.0   134
1991-03-01  66.0   135
1991-04-01  65.0   136
1991-05-01  60.0   137
```

Last few rows of Test Data

```
             Rose  time
YearMonth
1995-03-01  45.0   183
1995-04-01  52.0   184
1995-05-01  28.0   185
1995-06-01  40.0   186
1995-07-01  62.0   187
```

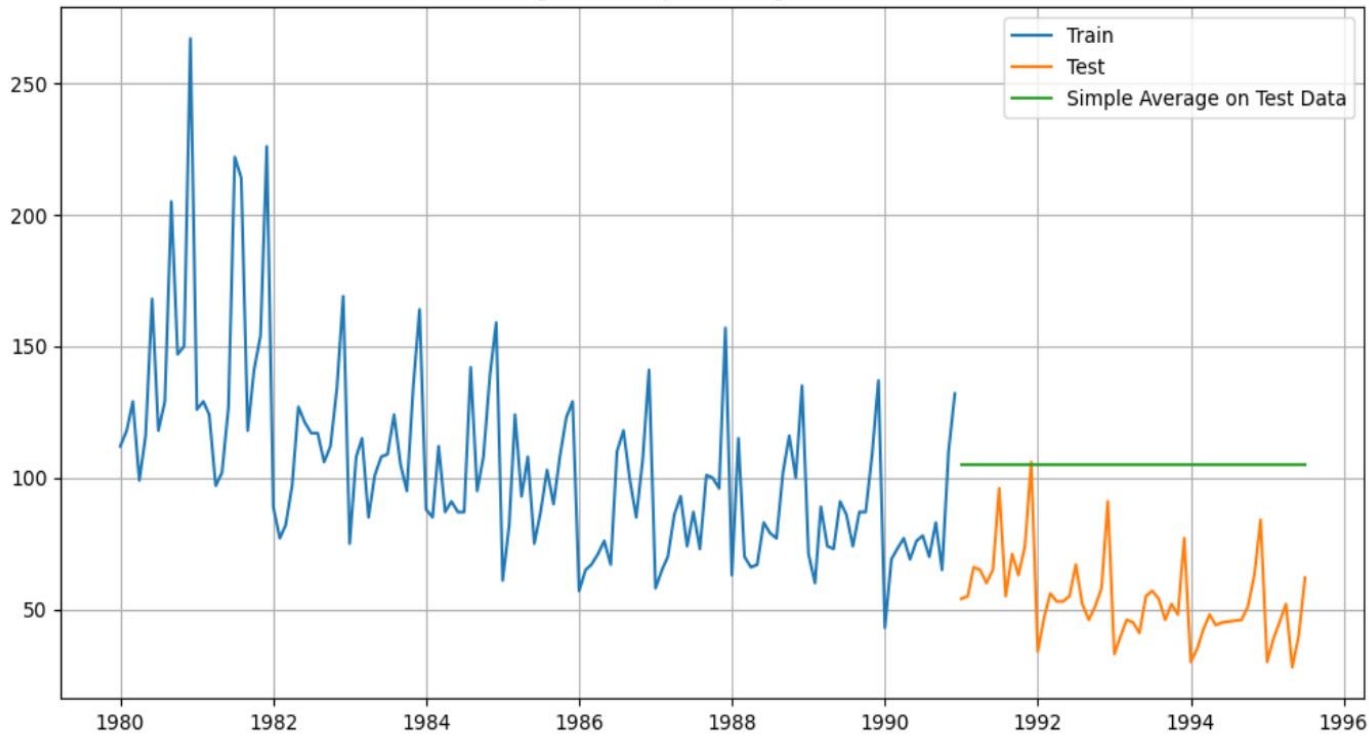


For Regression on Time forecast on the Test Data, RMSE is 15.269

Model 6 : Simple Average

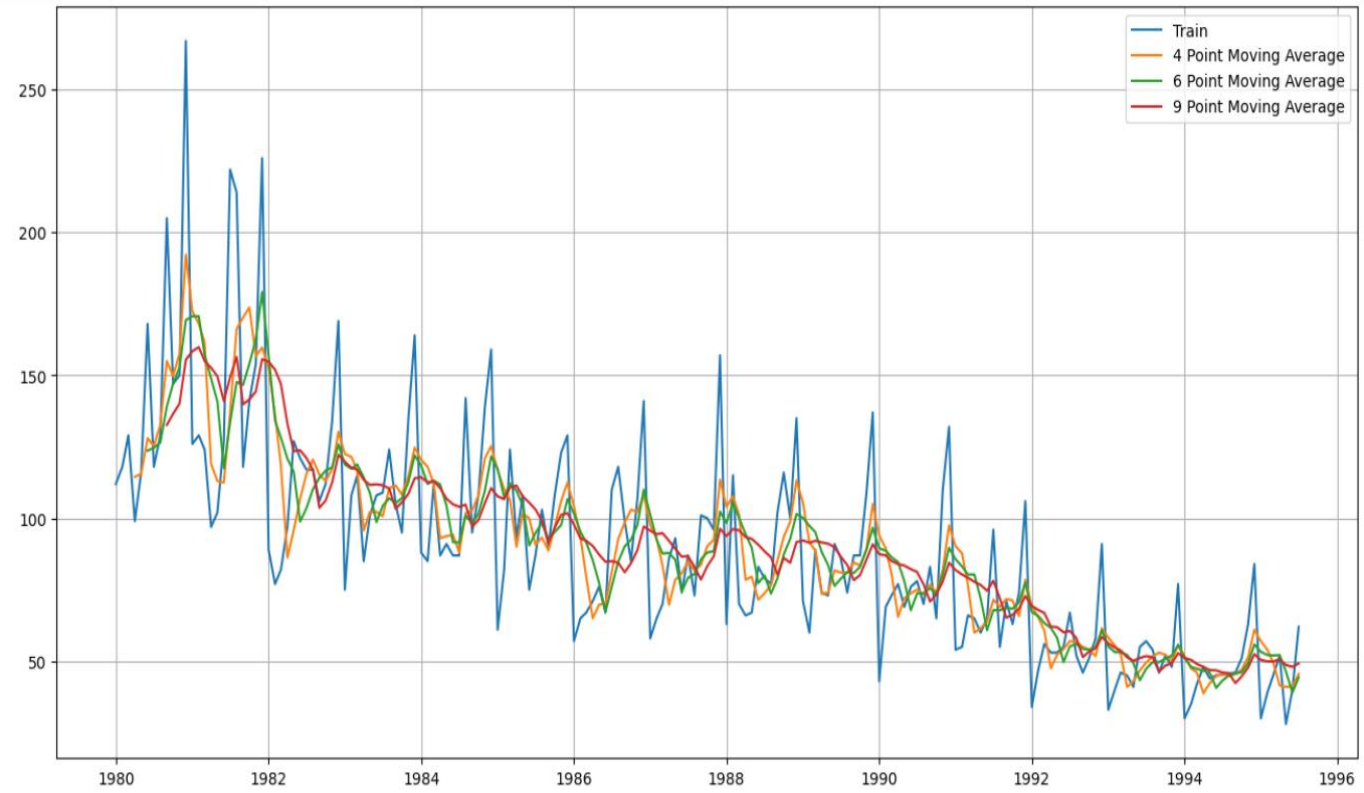
YearMonth	Rose	mean_forecast
1991-01-01	54.0	104.939394
1991-02-01	55.0	104.939394
1991-03-01	66.0	104.939394
1991-04-01	65.0	104.939394
1991-05-01	60.0	104.939394

Fig:2.17 Simple Average Forecast

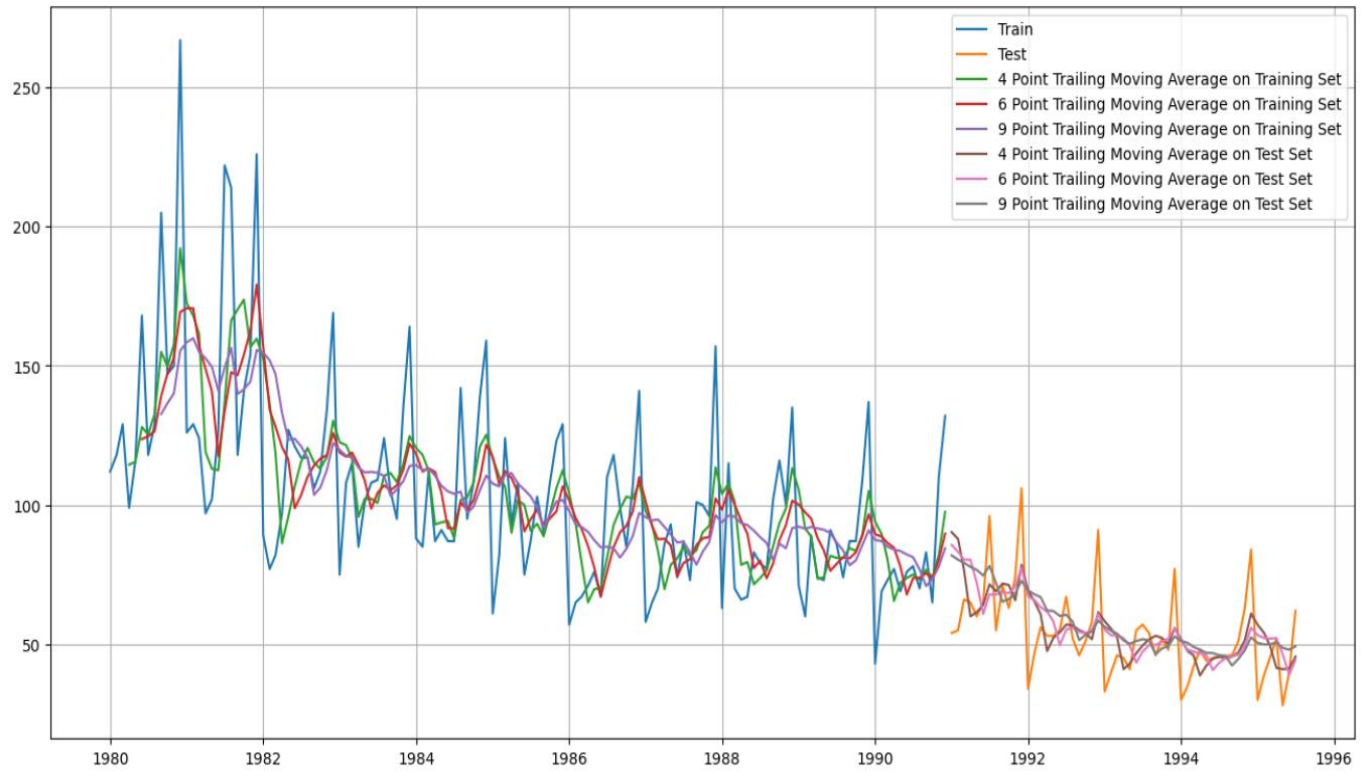


For Simple Average forecast on the Test Data, RMSE is 53.461

Model 11 : Moving Average(MA)



YearMonth	Rose	Trailing_4	Trailing_6	Trailing_9
1990-08-01	70.0	73.25	73.833333	76.888889
1990-09-01	83.0	76.75	75.500000	70.888889
1990-10-01	65.0	74.00	73.500000	73.333333
1990-11-01	110.0	82.00	80.333333	77.888889
1990-12-01	132.0	97.50	89.666667	84.444444



For 4 point Moving Average Model forecast on the Training Data, RMSE is 14.451
For 6 point Moving Average Model forecast on the Training Data, RMSE is 14.566
For 9 point Moving Average Model forecast on the Training Data, RMSE is 14.728

So far, we have tried 7 different models and Triple Exponential Smoothing suits very well on our data with lowest RMSE value

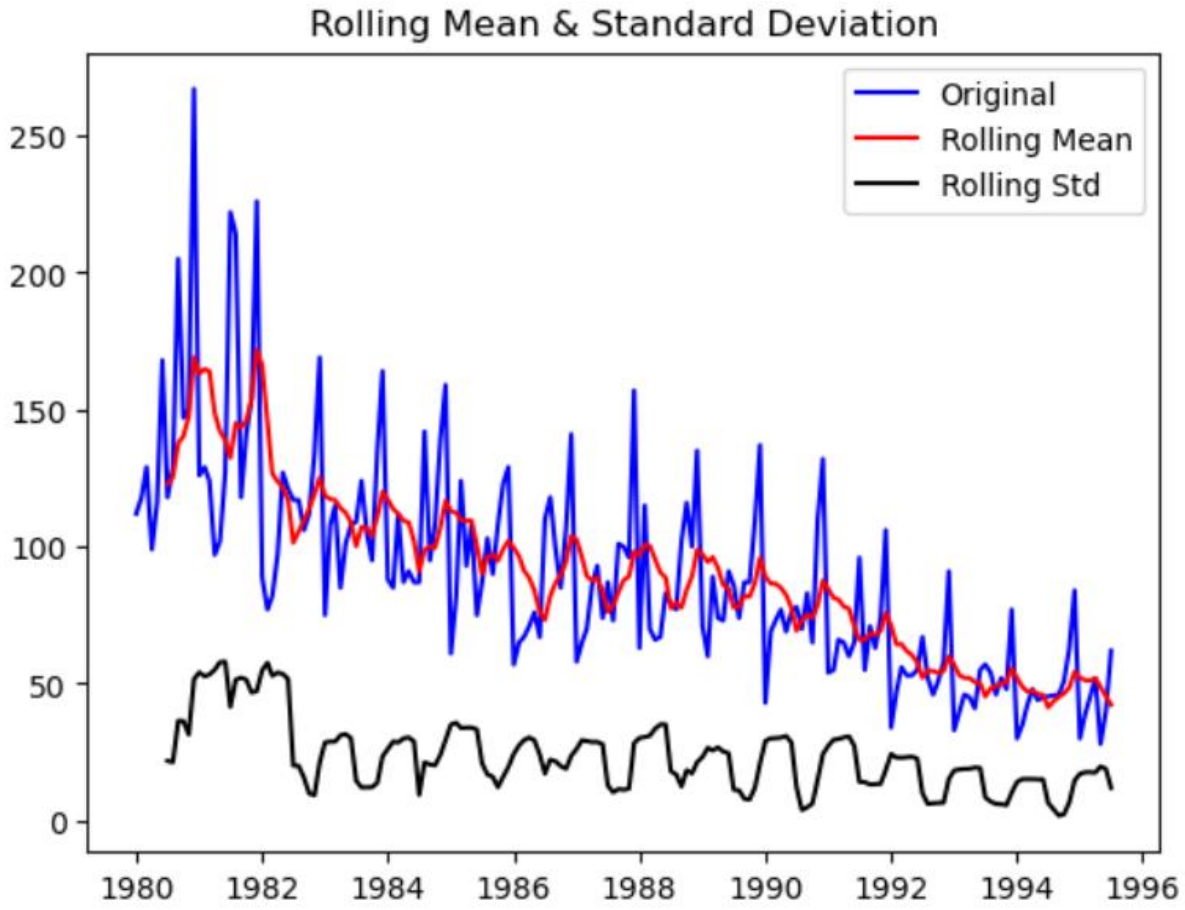
Q2.5 Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05.

To check the stationarity of the data ADF test can be used which is a hypothesis test and following is the null & alternate hypothesis for this test.

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.

Check for stationarity of the whole Time Series data.

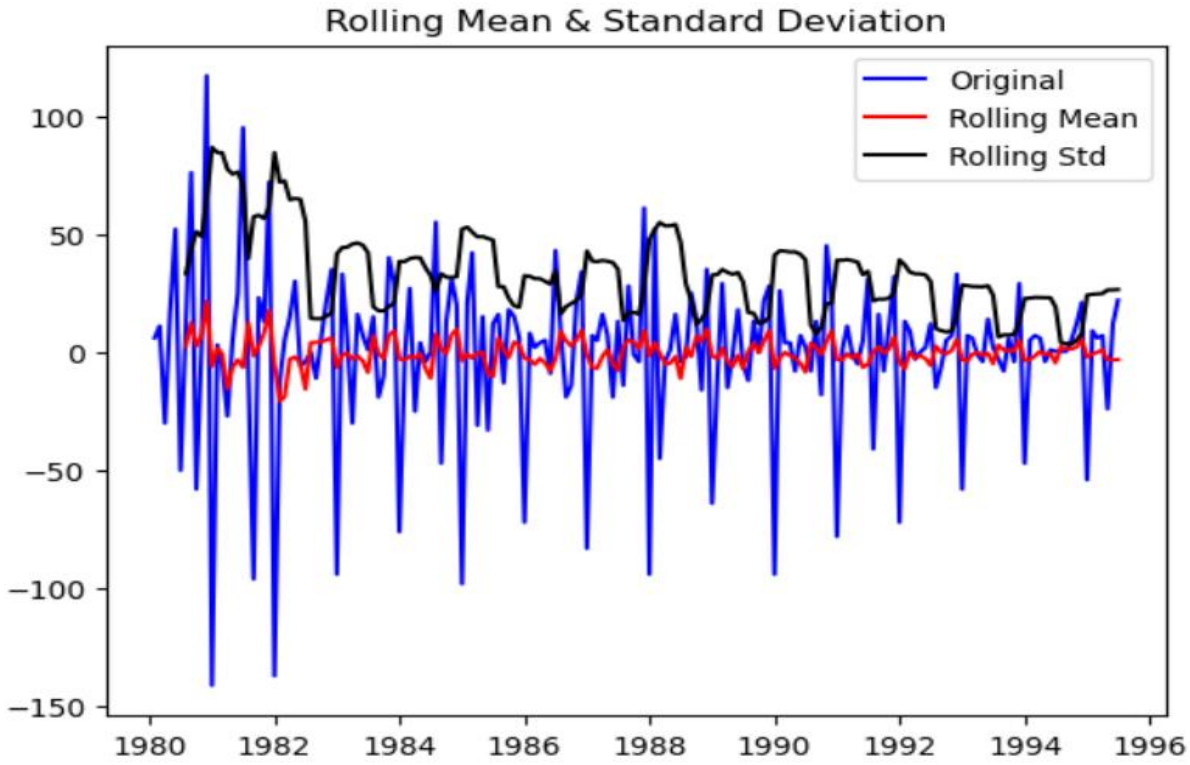


Results of Dickey-Fuller Test:

Test Statistic	-1.876699
p-value	0.343101
#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

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.

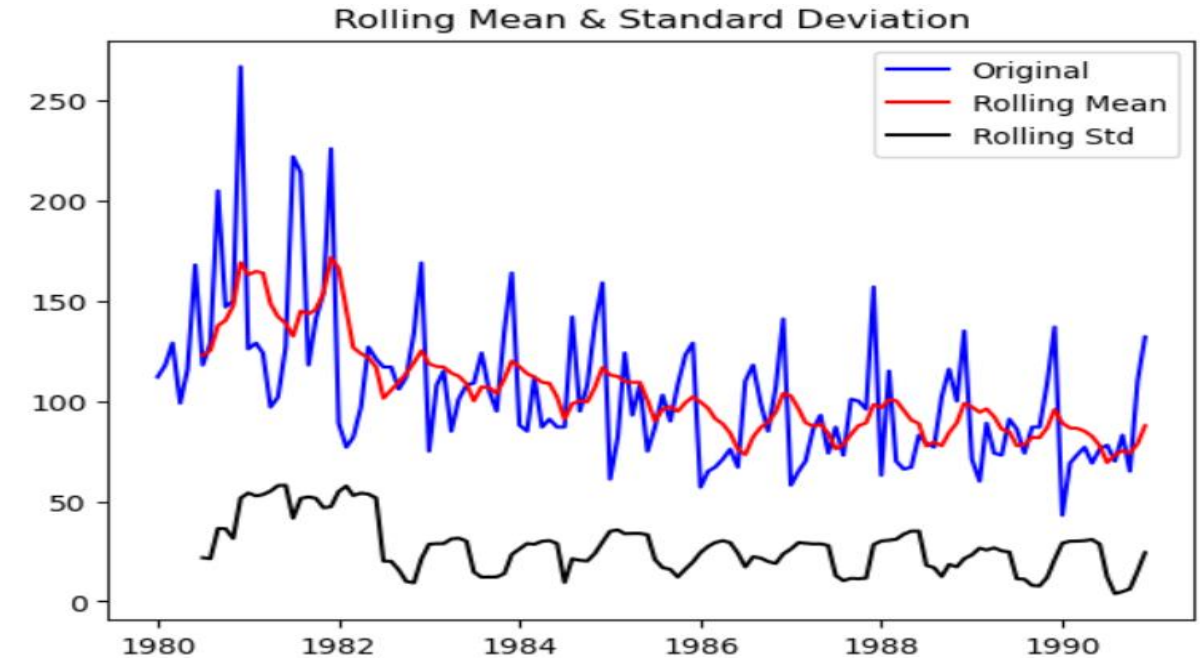


Results of Dickey-Fuller Test:

Test Statistic	-8.044392e+00
p-value	1.810895e-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

We see that at $\alpha = 0.05$ the Time Series become stationary.

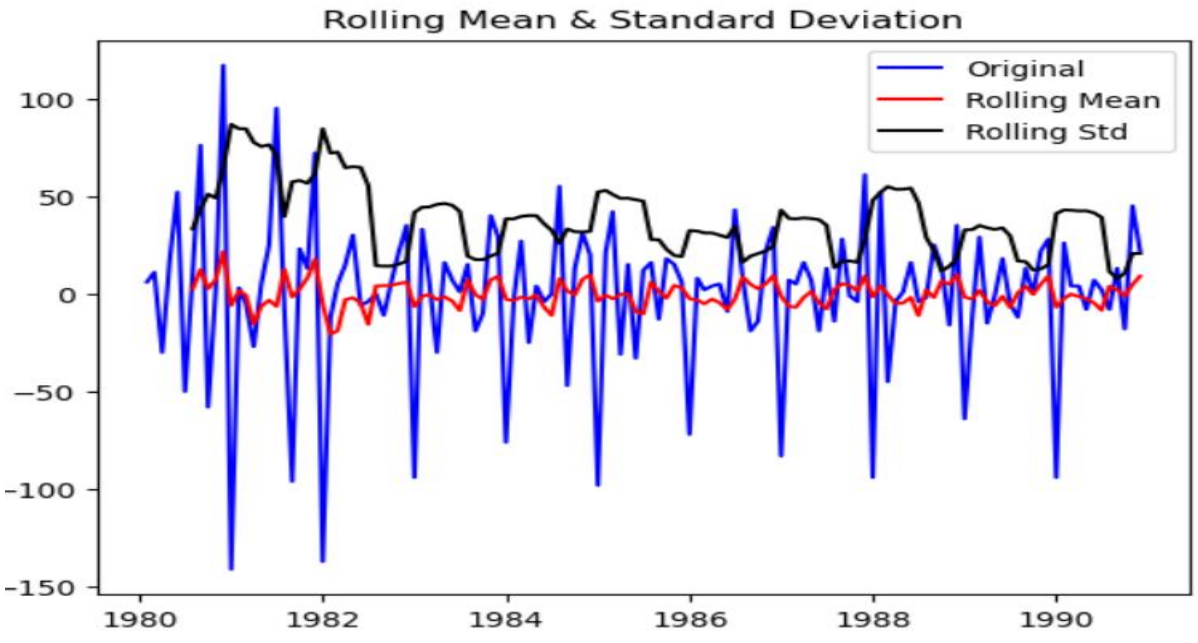
Check for stationarity of the Training Data Time Series



Results of Dickey-Fuller Test:

Test Statistic	-2.164250
p-value	0.219476
#Lags Used	13.000000
Number of Observations Used	118.000000
Critical Value (1%)	-3.487022
Critical Value (5%)	-2.886363
Critical Value (10%)	-2.580009

We see that the train data is not stationary at $\alpha = 0.05$.



Results of Dickey-Fuller Test:

```
Test Statistic                 -6.592372e+00
p-value                         7.061944e-09
#Lags Used                      1.200000e+01
Number of Observations Used     1.180000e+02
Critical Value (1%)            -3.487022e+00
Critical Value (5%)            -2.886363e+00
Critical Value (10%)           -2.580009e+00
```

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

Q2.6 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.

Model 7 : Automated ARIMA Model

We are taking all combination of p, d, q values from 0 to 2 and below is Some parameter combinations of these (p, d, q) values

```
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)
```

Below is the AIC of these (p, d, q) values

	param	AIC
2	(0, 1, 2)	1279.671529
5	(1, 1, 2)	1279.870723
4	(1, 1, 1)	1280.574230
7	(2, 1, 1)	1281.507862
8	(2, 1, 2)	1281.870722
1	(0, 1, 1)	1282.309832
6	(2, 1, 0)	1298.611034
3	(1, 1, 0)	1317.350311
0	(0, 1, 0)	1333.154673

We can see that lowest AIC value generated when p=0, d=1, q=2 so we are building ARIMA model with these p, d, q values.

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	ARIMA(0, 1, 2)	Log Likelihood	-636.836
Date:	Mon, 26 Dec 2022	AIC	1279.672
Time:	22:09:00	BIC	1288.297
Sample:	01-01-1980 - 12-01-1990	HQIC	1283.176
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.6970	0.072	-9.689	0.000	-0.838	-0.556
ma.L2	-0.2042	0.073	-2.794	0.005	-0.347	-0.061
sigma2	965.8407	88.305	10.938	0.000	792.766	1138.915

Ljung-Box (L1) (Q):	0.14	Jarque-Bera (JB):	39.24
Prob(Q):	0.71	Prob(JB):	0.00
Heteroskedasticity (H):	0.36	Skew:	0.82
Prob(H) (two-sided):	0.00	Kurtosis:	5.13

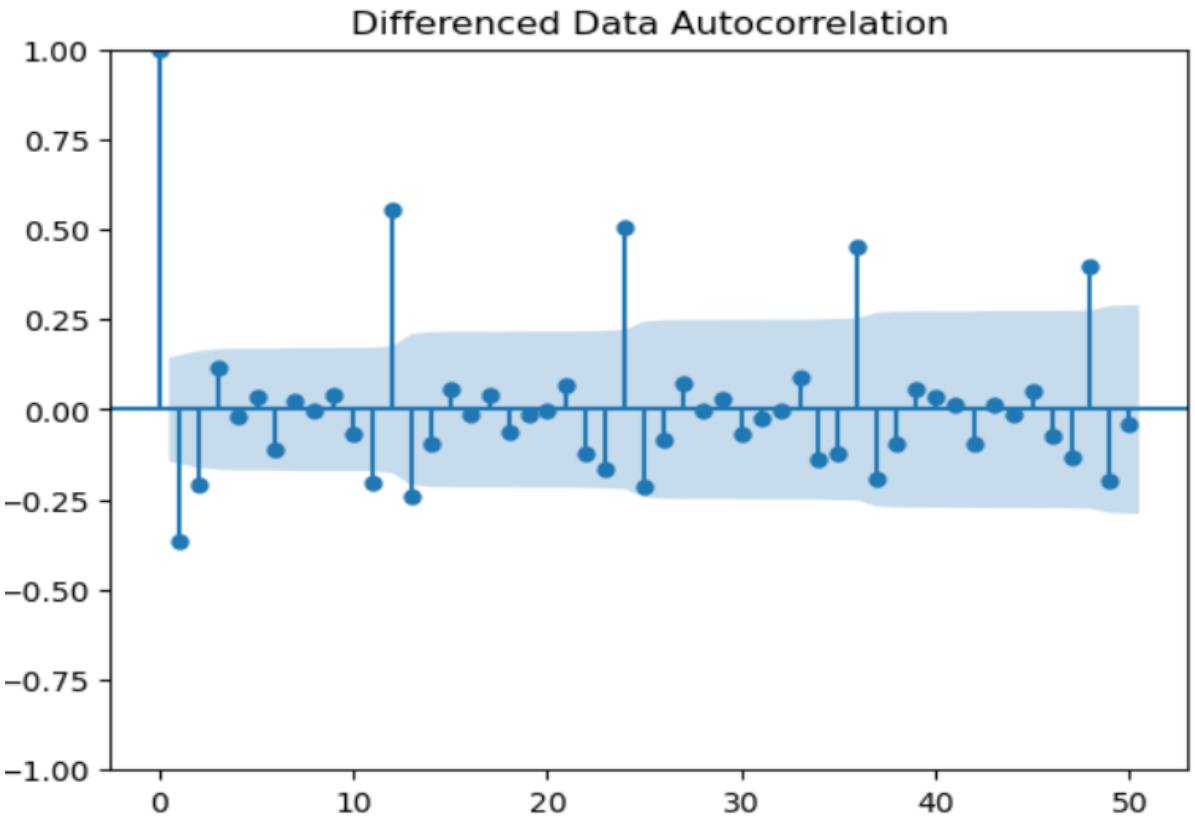
Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

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

Automated ARIMA model (0, 1, 2) RMSE score = 37.30

Model 8 : Automated SARIMA Model

Let us look at the ACF plot to understand the seasonal parameter for the SARIMA model



We can see that there is a seasonality at 12.

Setting the seasonality as 12 for the auto SARIMA model

```
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 lowest AIC values:

	param	seasonal	AIC
26	(0, 1, 2)	(2, 0, 2, 12)	887.937509
53	(1, 1, 2)	(2, 0, 2, 12)	889.899655
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

We can see that lowest AIC value generated when p=0, d=1, q=2 & P=2, D=0, Q=2, S=12 so we are building SARIMA model with these p, d, q & P, D, Q, S values.

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                  1
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 12)   Log Likelihood                -436.9
Date:                            Mon, 26 Dec 2022   AIC                             887.9
Time:                                    22:09:58   BIC                             906.4
Sample:                                         0   HQIC                            895.4
                                              - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.8427    189.814     -0.004      0.996    -372.871     371.185
ma.L2         -0.1573     29.821     -0.005      0.996     -58.605      58.290
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
===================================================================================

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

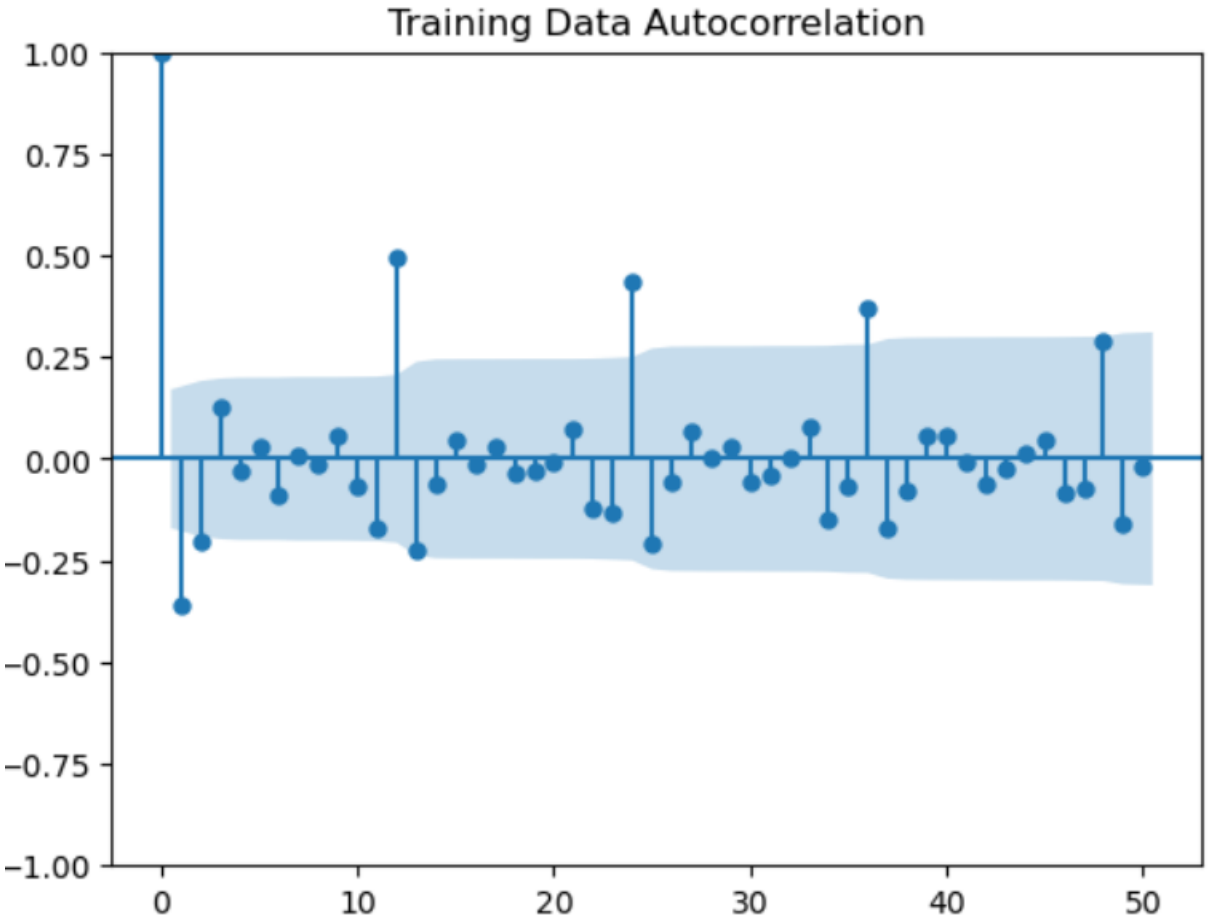
Predict on the Test Set using this model and evaluate the model

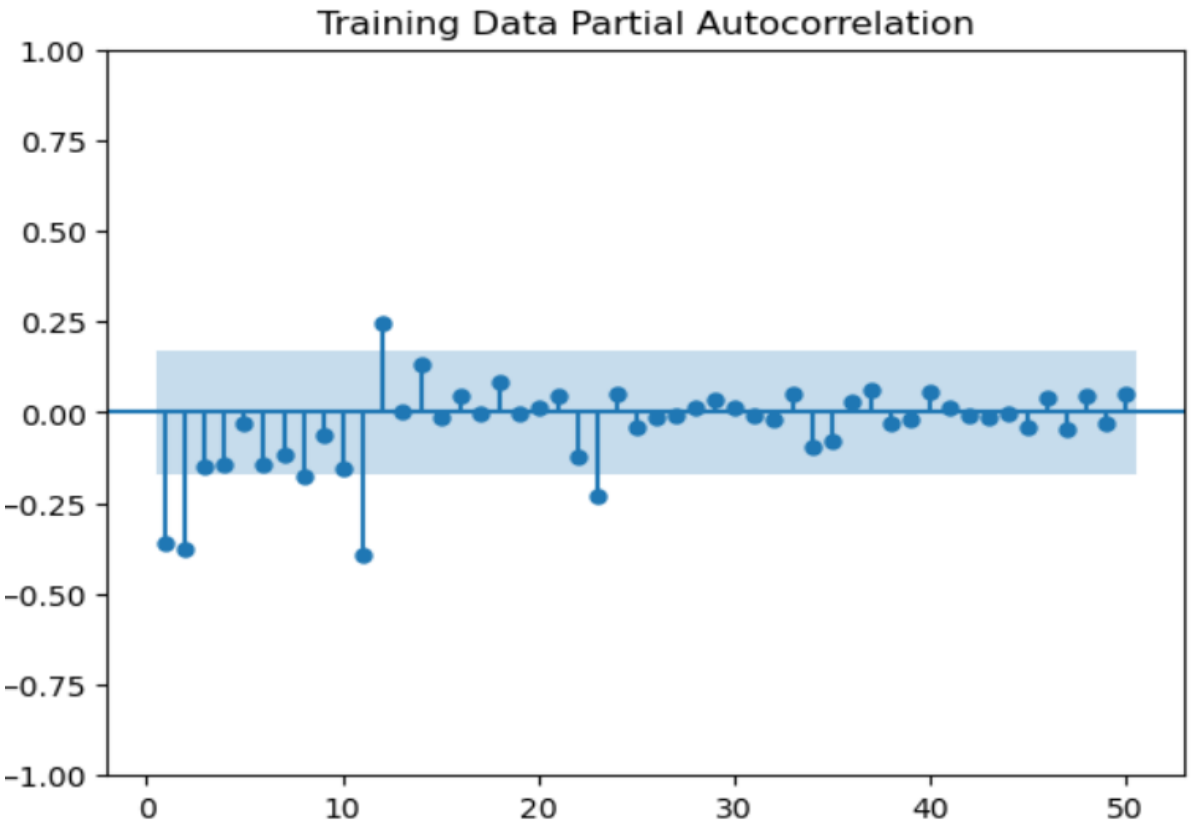
y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	62.867264	15.928501	31.647976	94.086552
1	70.541190	16.147659	38.892360	102.190019
2	77.356411	16.147656	45.707586	109.005236
3	76.208814	16.147656	44.559989	107.857638
4	72.747398	16.147656	41.098573	104.396223

Automated SARIMA model (**0,1,2) (1,0,2,12**) RMSE score = 26.92

Q2.7 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

Model 9: ARIMA model as per ACF and the PACF cut-off.





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 2.

The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 1

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	ARIMA(2, 1, 1)	Log Likelihood	-636.754
Date:	Mon, 26 Dec 2022	AIC	1281.508
Time:	22:10:06	BIC	1293.009
Sample:	01-01-1980	HQIC	1286.181
	- 12-01-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1834	0.078	2.353	0.019	0.031	0.336
ar.L2	-0.0959	0.106	-0.907	0.364	-0.303	0.111
ma.L1	-0.9075	0.058	-15.663	0.000	-1.021	-0.794
sigma2	964.2240	88.516	10.893	0.000	790.735	1137.713

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	40.53
Prob(Q):	0.96	Prob(JB):	0.00
Heteroskedasticity (H):	0.36	Skew:	0.87
Prob(H) (two-sided):	0.00	Kurtosis:	5.10

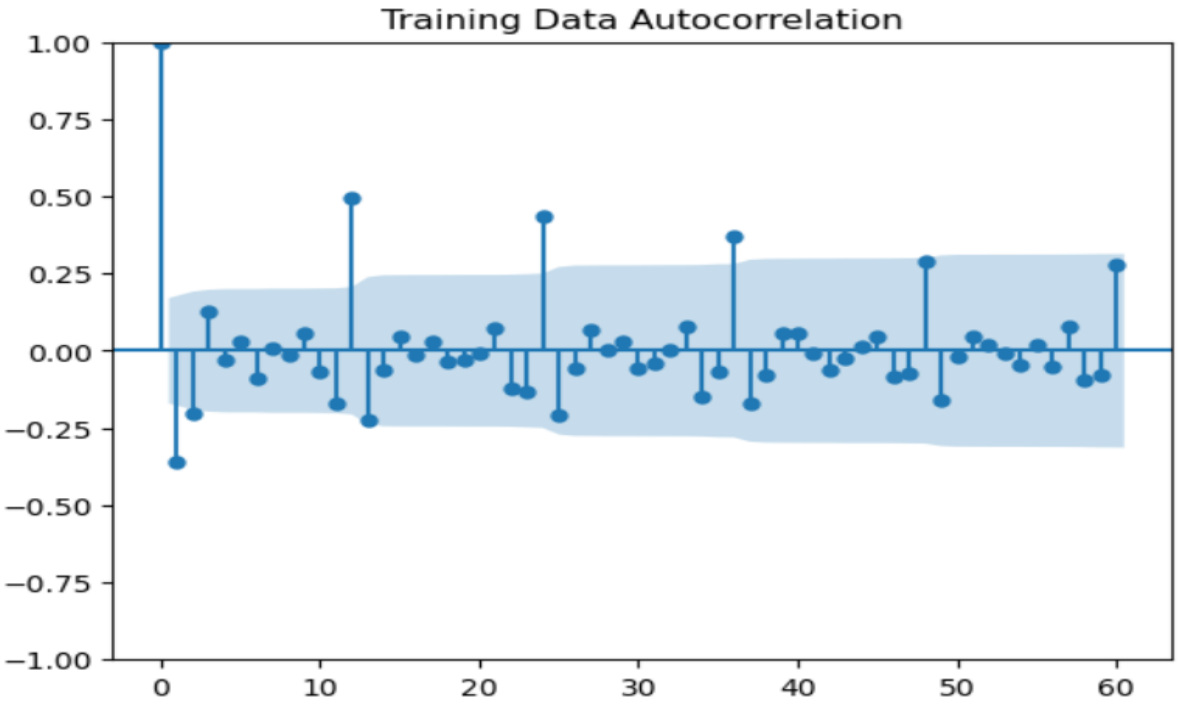
Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

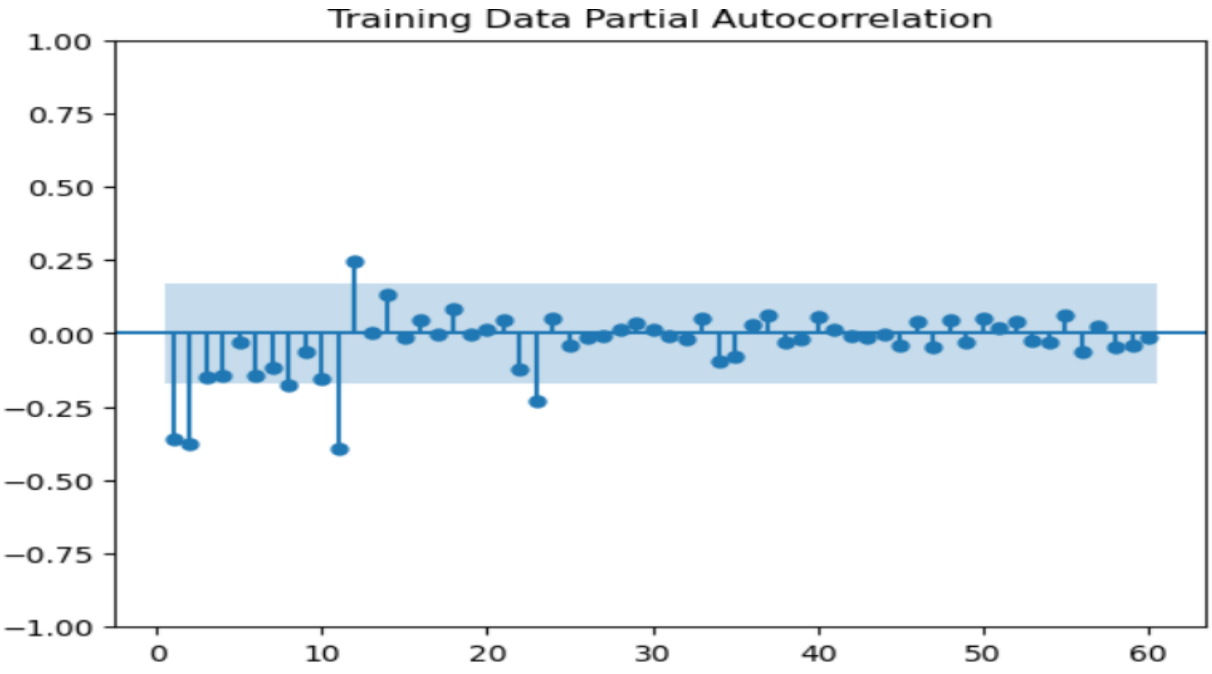
Predict on the Test Set using this model and evaluate the model

Automated ARIMA model (**2,1,1**) RMSE score = 36.83

Model 10: SARIMA model as per ACF and the PACF cut-off

Let us look at the ACF and the PACF plots once more.





Here, we have taken alpha=0.05.

We are going to take the seasonal period as 12. We are taking the p value to be 2 and the q value also to be 1 as the parameters same as the ARIMA model.

The Auto-Regressive parameter in a SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 1.

The Moving-Average parameter in a SARIMA model is 'Q' which comes from the significant lag after which the ACF plot cuts-off to 3.

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                  132
Model:             SARIMAX(2, 1, 1)x(1, 1, [1, 2, 3], 12)   Log Likelihood               -1795.937
Date:                                Mon, 26 Dec 2022   AIC                             3607.875
Time:                                        22:10:14   BIC                             3627.030
Sample:                                             0   HQIC                            3615.560
                                                  - 132
Covariance Type:                                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.0863         -0       -inf      0.000       0.086       0.086
ar.L2         -0.3071         -0        inf      0.000      -0.307      -0.307
ma.L1         -0.5760   1.73e-32  -3.33e+31      0.000      -0.576      -0.576
ar.S.L12      -0.2848   1.09e-35  -2.62e+34      0.000      -0.285      -0.285
ma.S.L12   -6.084e+13   1.37e-32  -4.45e+45      0.000   -6.08e+13   -6.08e+13
ma.S.L24   -5.378e+13   1.33e-45  -4.04e+58      0.000   -5.38e+13   -5.38e+13
ma.S.L36   -1.974e+13   5.06e-46   -3.9e+58      0.000   -1.97e+13   -1.97e+13
sigma2      7.796e-10   3.04e-10      2.563      0.010    1.83e-10    1.38e-09
===================================================================================
Ljung-Box (L1) (Q):                   2.92   Jarque-Bera (JB):                 3.56
Prob(Q):                              0.09   Prob(JB):                         0.17
Heteroskedasticity (H):               1.85   Skew:                            -0.30
Prob(H) (two-sided):                  0.12   Kurtosis:                         3.83
===================================================================================

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

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

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	15.901788	1.698699e+09	-3.329388e+09	3.329388e+09
1	81.208099	1.907076e+09	-3.737800e+09	3.737800e+09
2	57.752404	1.926575e+09	-3.776018e+09	3.776018e+09
3	73.426243	1.984908e+09	-3.890349e+09	3.890349e+09
4	63.471574	2.097352e+09	-4.110734e+09	4.110734e+09

Manual SARIMA model (2,1,1), (1,1,3,12) RMSE score = 23.46

Q2.8 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.

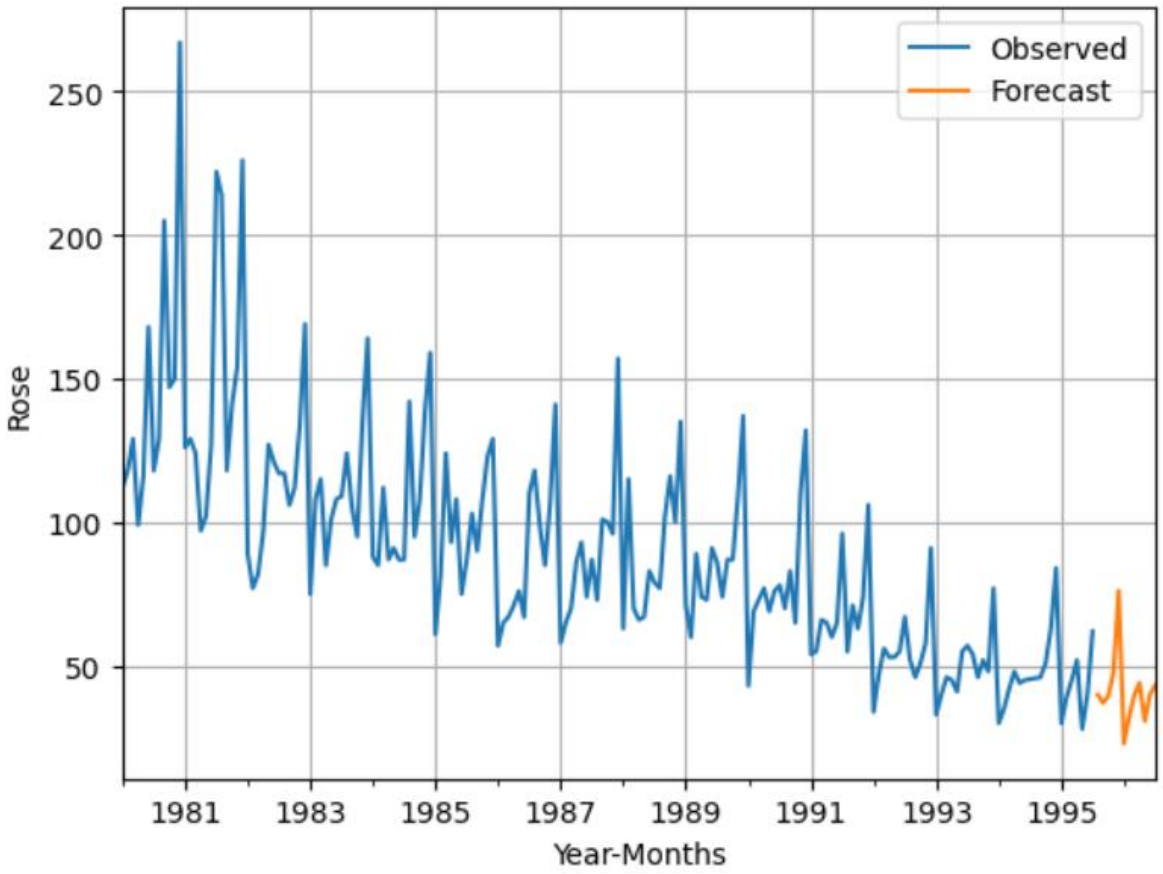
We have created a Data Frame named as - resultsDf and below is the output of that data.

	Test RMSE
Alpha=0.098,SimpleExponentialSmoothing	36.796227
Alpha =0.017, Beta=3.236, DoubleExponentialSmoothing	15.707052
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	265.567594
Alpha=0.0895,Beta = 0.0002,Gamma=0.0034,TripleExponentialSmoothing	14.249661
Alpha=0.3,Beta = 0.3,Gamma=0.4,TripleExponentialSmoothing	12.723156
NaiveModel	79.718773
RegressionOnTime	15.268955
SimpleAverageModel	53.460570
4pointTrailingMovingAverage	14.451403
6pointTrailingMovingAverage	14.566327
9pointTrailingMovingAverage	14.727630
ARIMA(0,1,2)	37.306480
SARIMA(1,1,2)(1,0,2,12)	26.928362
ARIMA(2,1,1)	36.833530
SARIMA(2,1,1)(1,1,3,12)	23.462849

Q2.9 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.

We have built multiple models and based on the minimum RMSE values
Triple Exponential Smoothing is the best models.

Building the Triple Exponential Smoothing model on the Full Data



Next 12 months forecasts:

	lower_CI	prediction	upper_ci
1995-08-01	-2.693344	39.913880	82.521105
1995-09-01	-5.546417	37.060808	79.668033
1995-10-01	-3.168022	39.439203	82.046428
1995-11-01	5.034250	47.641475	90.248700
1995-12-01	33.512938	76.120163	118.727388
1996-01-01	-19.667110	22.940115	65.547340
1996-02-01	-10.458781	32.148444	74.755669
1996-03-01	-2.811685	39.795540	82.402765
1996-04-01	1.441612	44.048836	86.656061
1996-05-01	-11.819076	30.788149	73.395374
1996-06-01	-2.635458	39.971766	82.578991
1996-07-01	0.594201	43.201426	85.808651

RMSE of Triple Exponential Smoothing Model built on full data is- 21.68

So, we can see that Triple Exponential Smoothing Model perform very well on the data as it has low RMSE.

Q2.10 Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales

1. After checking All RMSE values we can say that Triple Exponential Smoothing gives the lowest RMSE Value.

2. As per the final plot on complete data we can say that our model is able to forecast very well.

3. We have seen that past data has more influence in the forecast so more the past data better the forecast.

4. we have seen that sales of wine went up at the end of the year possibly due to festival seasons so for the other months where sales are low company can provide different offers.

5. Based on the SARIMA model which built on Full data we got Jarque-Bera test's P-value is not less than 0.05 so we failed to reject the H0 & we found that Data is normal.

6. Based on the Ljung-Box test's P-value we failed to reject the H0 so we can say that residuals are independent.

7. With the help of next 12 months forecast sales company can make plan to increase the sales for those months where forecast shows less sales.

8. Sales of Rose wine decreasing continuously so need to check what went wrong in wine sales.

9. Across the year Month 'April' sales are very low so need to check what is the impact of April month sales on overall sales.

10. Maximum drop in Average wine sales occur after year 1981 and since then it kept decreasing.

THE END