TIME SERIES FORECASTING PROJECT REPORT

KIRAN.N

GREAT LEARNING

Table of Contents

List Of Tables

List Of Figures

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.

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

Time Series data present in csv file is read into a pandas Data Frame using read_csv() function. This normally loads data into a dataframe. To inform pandas that current data is a time series data we pass a parameter 'parse_dates' with the time series column YearMonth as a value. Also, we make our time series reference as the index.

The current Time series data, Sparkling.csv has the sales information of Sparkling wines from January 1980 to July 1995 total 187 rows.

The current Time series data, Rose.csv has the sales information of Sparkling wines from January 1980 to July 1995 total 187 rows.

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

Table 1: Sparking Data Set Sample

	Rose
YearMonth	
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

Table 2: Rose Data Set Sample

Following figures show the Time series plot of Sparkling and Rose wine sales information.

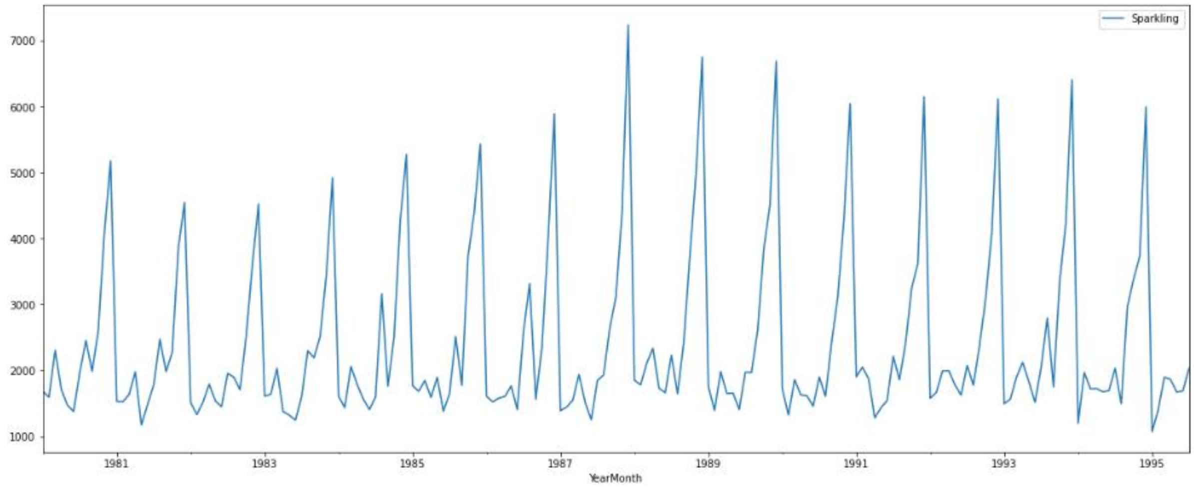

Figure 1: Time Series Data Plot of Sparkling Data

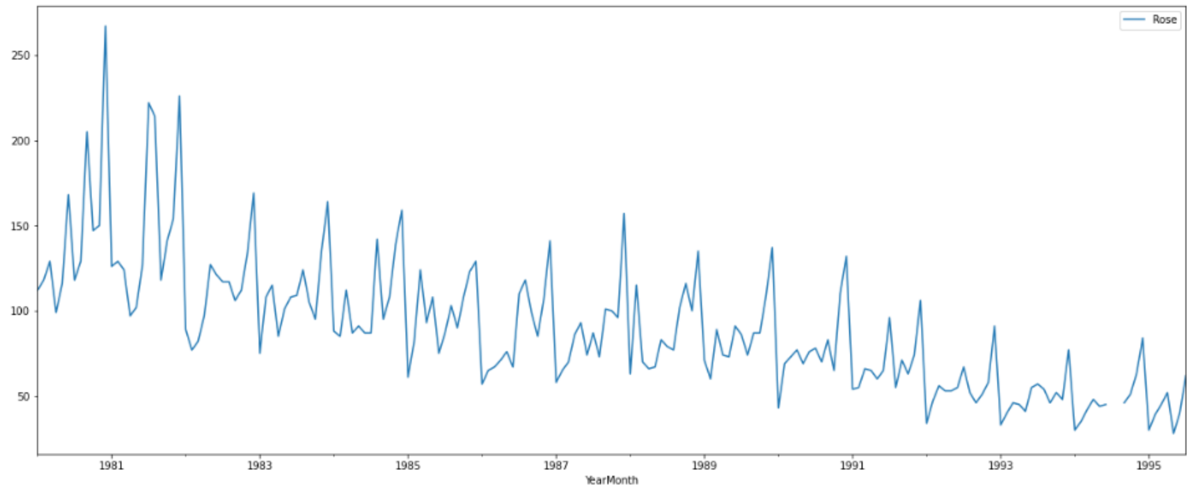

Figure 2: Time Series Data Plot of Rose Data

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

Size Of Dataset

```
In [21]:  df1.shape
Out[21]: (187, 1)

In [22]:  df2.shape
Out[22]: (187, 1)
```

From the above output we observe there are total 187 rows of data in each dataset.

Data Type & Null Check

- Sparkling Dataset

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
Data columns (total 1 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Sparkling  187 non-null    int64
dtypes: int64(1)
```

The Sparkling column present in data set is of integer type and there are no null values present in the dataset.

- Rose dataset

```
<class 'pandas.core.frame.DataFrame'>
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
```

The Rose column present in data set is of integer type and there are 2 null values present in the dataset. Using bfill() we are replacing null values present in the dataset.

Descriptive Statistics

	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

Table 3: Descriptive Statistics of Sparkling Dataset

	Rose
count	187.000000
mean	89.919786
std	39.232269
min	28.000000
25%	62.500000
50%	85.000000
75%	111.000000
max	267.000000

Table 4: Descriptive Statistics of Rose Dataset

Univariate Analysis

Box Plots by Year

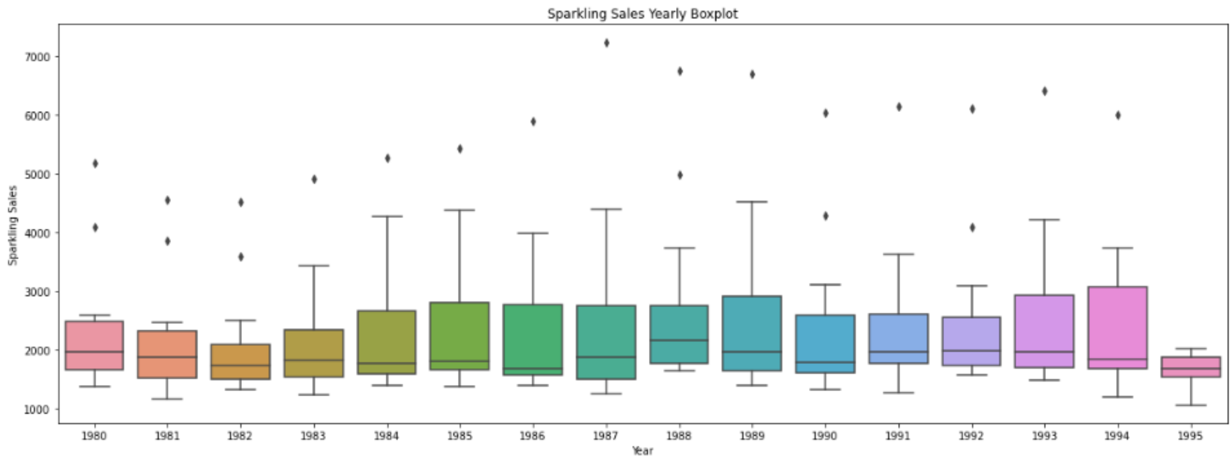

Figure 3: Sparkling Sales Yearly Boxplot

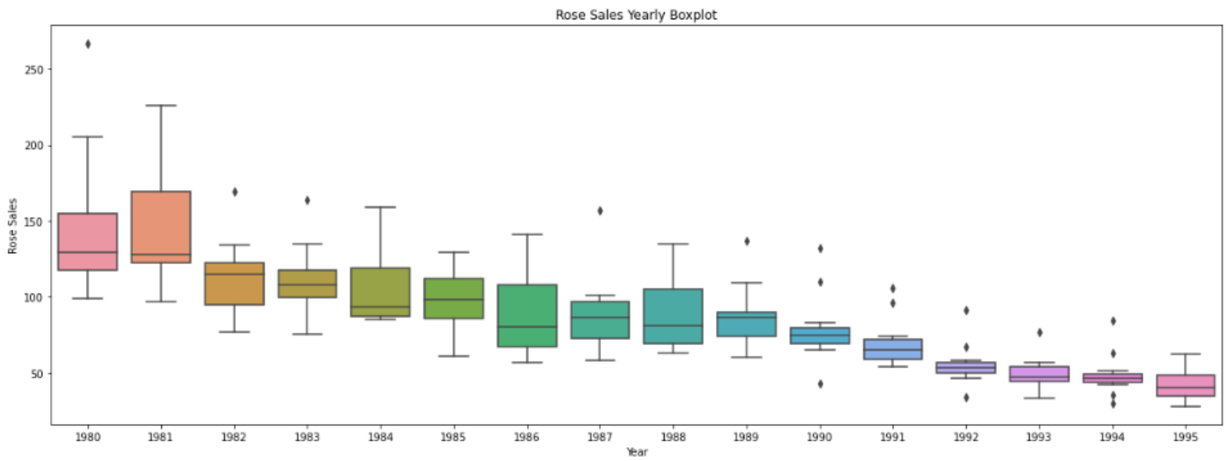

Figure 4: Rose Sales Yearly Boxplot

Box Plots by Month

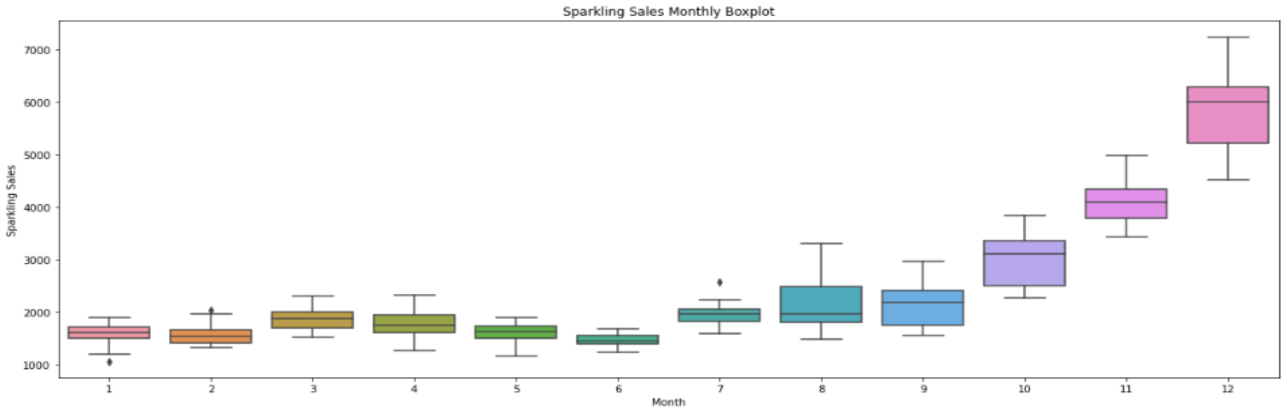

Figure 5: Sparkling Sales Monthly Boxplot

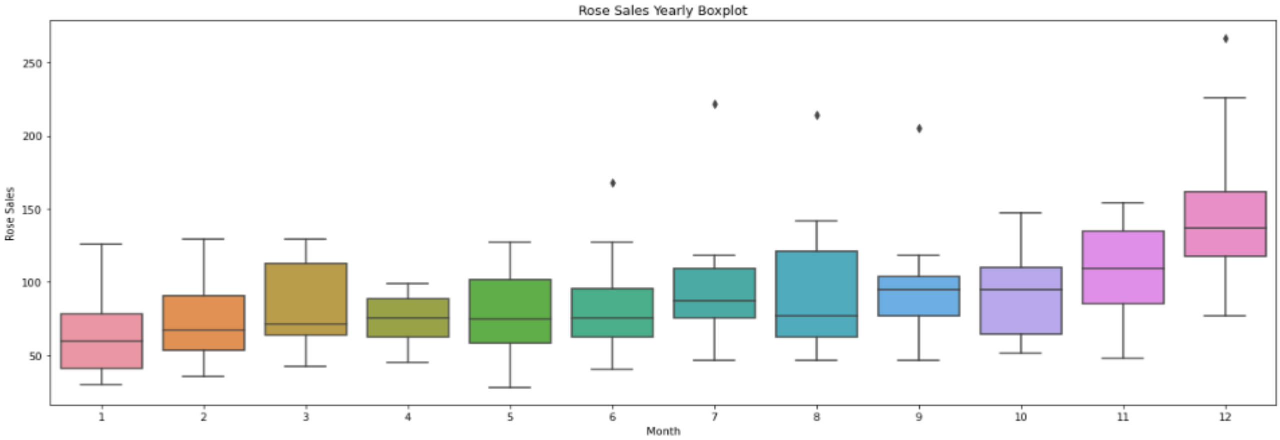

Figure 6: Rose Sales Monthly Boxplot

From the above monthly plots, we observe sales during December month are high compared to other months.

Also, sale of Rose wine is decreasing on year-on-year basis.

Time Series Decomposition

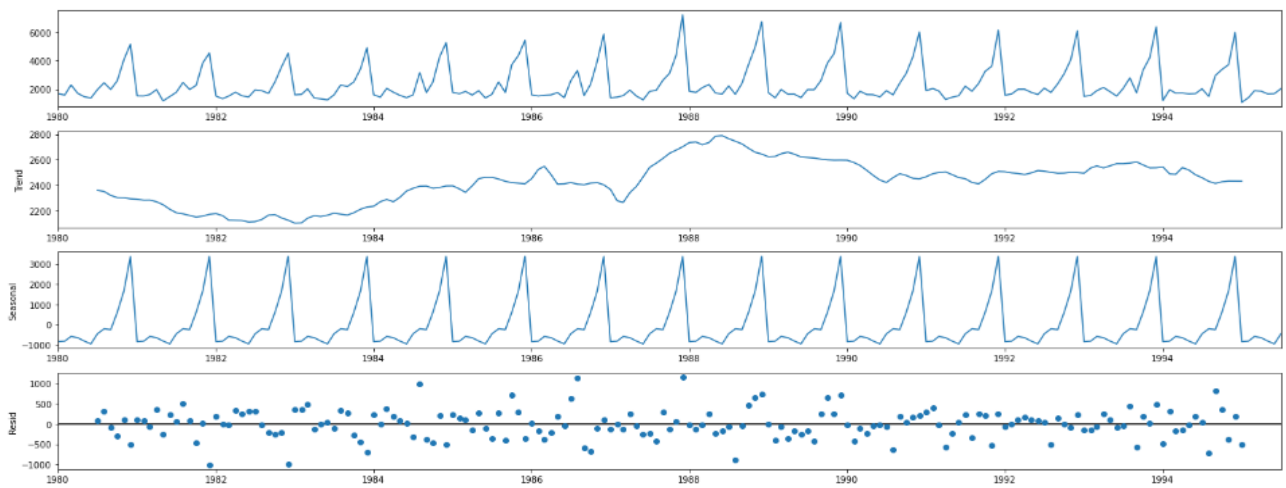

Figure 7: Additive Decomposition of Sparkling Data

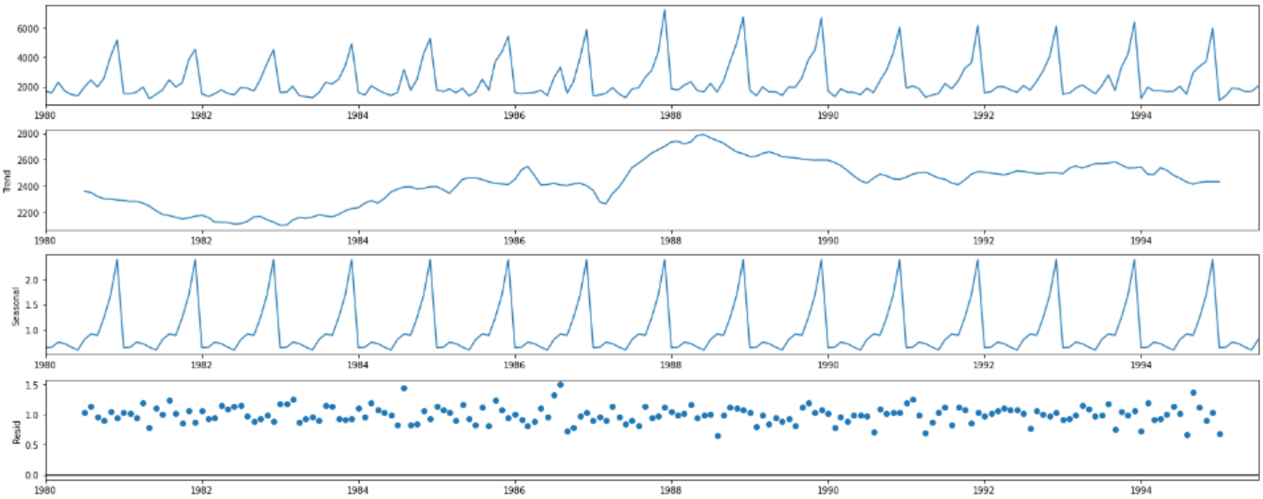

Figure 8: Multiplicative Decomposition of Sparkling Data

We have decomposed the Time series data in Additive and Multiplicative decomposition in Fig 7 and Fig 8 respectively. Observing both the decomposition patterns, Residual component in Additive decomposition still shows some kind of pattern and data points are spread across while Residual component in Multiplicative decomposition does not show any pattern and data points are spread evenly.

Hence Multiplicative decomposition is the right way of decomposition for Sparkling dataset.

Individual Components output is present in IPYNB file.

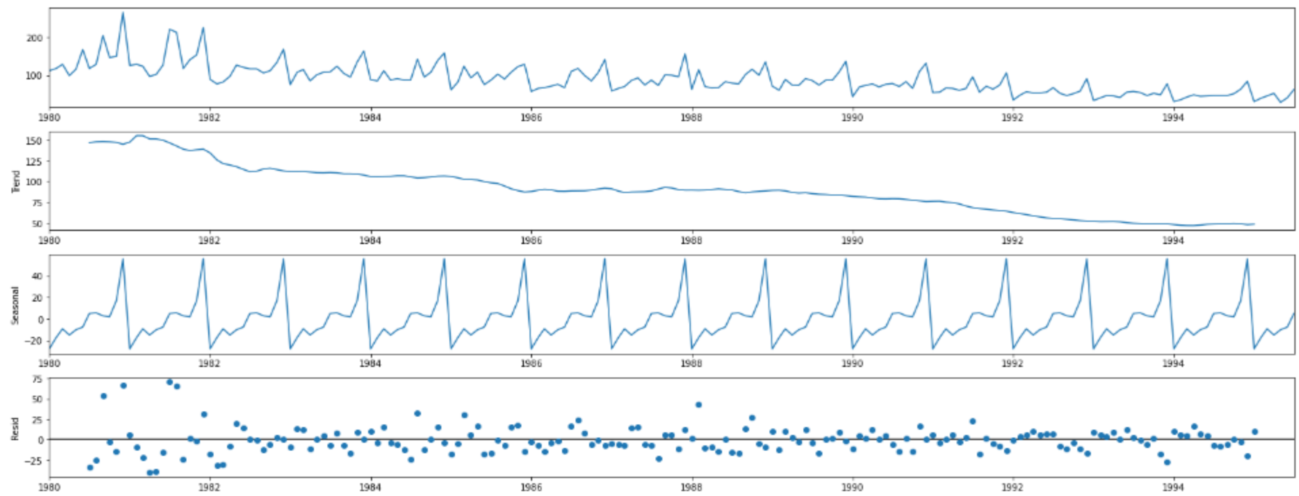

Figure 9: Additive Decomposition of Rose Data

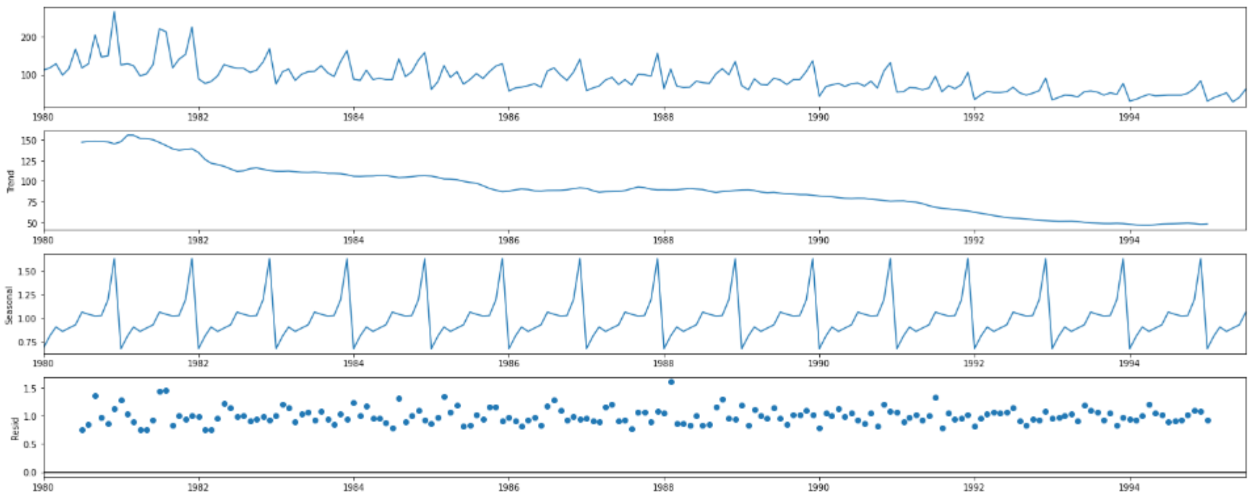

Figure 10: Multiplicative Decomposition of Rose Data

We have decomposed the Time series data in Additive and Multiplicative decomposition in Fig 9 and Fig 10 respectively. Observing both the decomposition patterns, Residual component in Additive decomposition still shows some kind of pattern and data points are spread across while Residual component in Multiplicative decomposition does not show any pattern and data points are spread evenly.

Hence Multiplicative decomposition is the right way of decomposition for Sparkling dataset.

Individual Components output is present in IPYNB file.

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

The regular approach to split the data into Train and Test dataset was to use TrainTestSplit which randomly splits the data train and test dataset. Currently we are dealing with Timeseries data which cannot be split randomly, here we split the data into train and test dataset based on a date. In the current problem all timeseries data before 1991 is taken as train data and test data starts from 1991.

After splitting the data into train and test data in both Sparkling and Rose dataset, Train dataset has 132 rows and test data set has 55 rows.

```
In [39]:  train_spark.shape
Out[39]: (132, 1)

In [40]:  test_spark.shape
Out[40]: (55, 1)
```

```
In [41]:  train_rose.shape
Out[41]: (132, 1)

In [42]:  test_rose.shape
Out[42]: (55, 1)
```

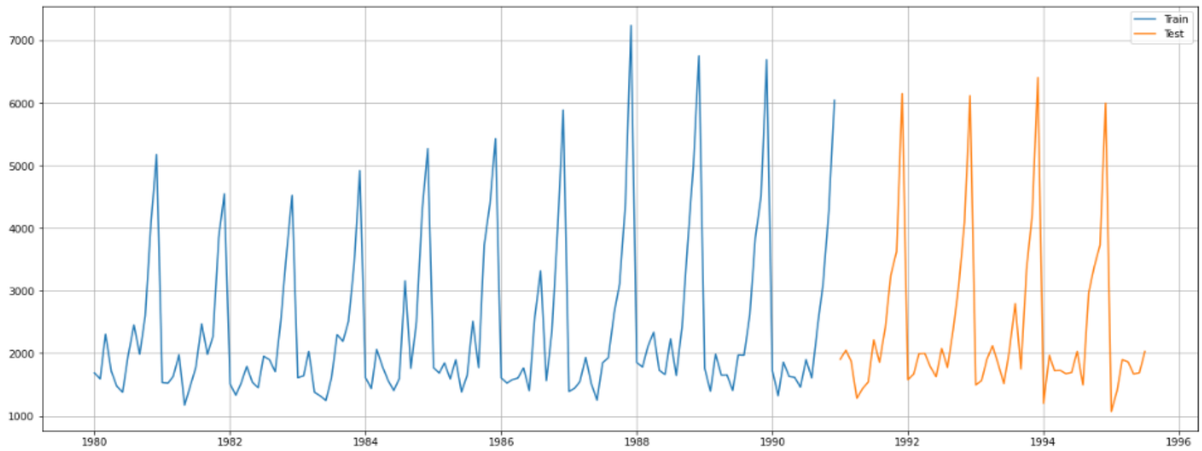

Figure 11: Sparkling Dataset After Train and Test Split

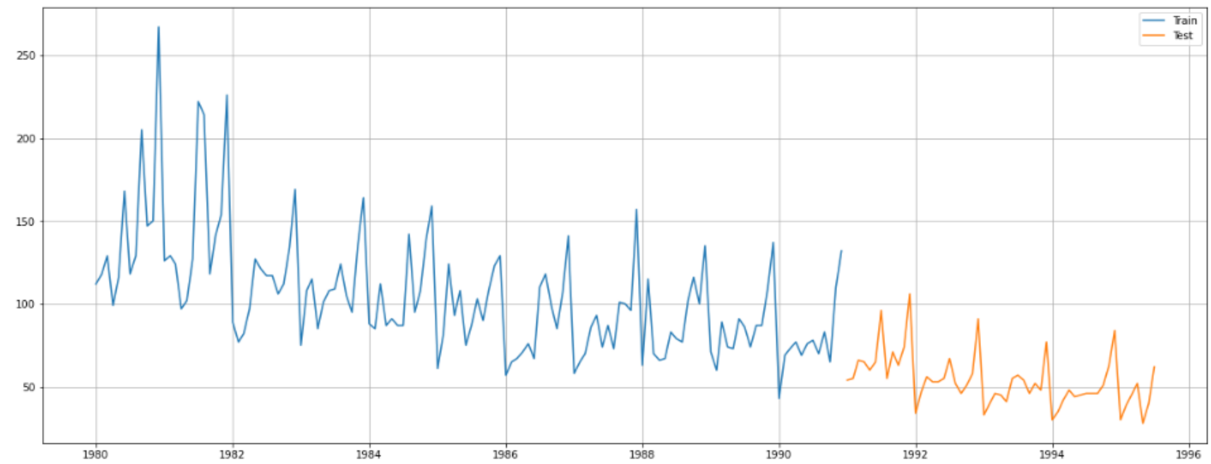

Figure 12: Rose Dataset After Train and Test Split

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.

After splitting the given dataset into test and train dataset, we have built Linear Regression Model, Naïve Forecast Model and Simple average model for Forecasting purpose.

Basic Forecast

	Test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804

Table 5: RMSE Values of Sparkling Data

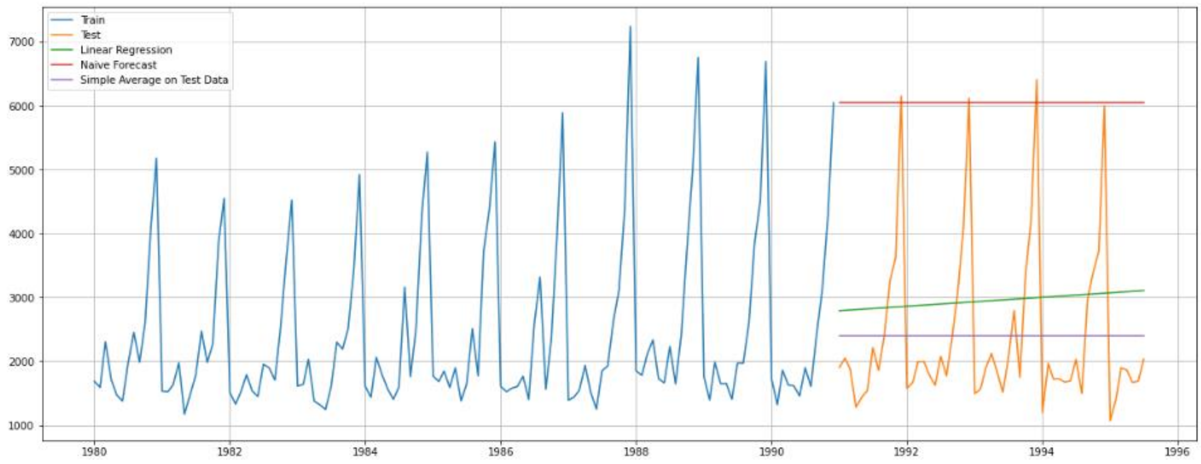

Figure 13: Sparkling Data Forecast Plot

	Test RMSE
RegressionOnTime	15.262509
NaiveModel	79.699093
SimpleAverageModel	53.440426

Table 6: RMSE Values of Rose Data

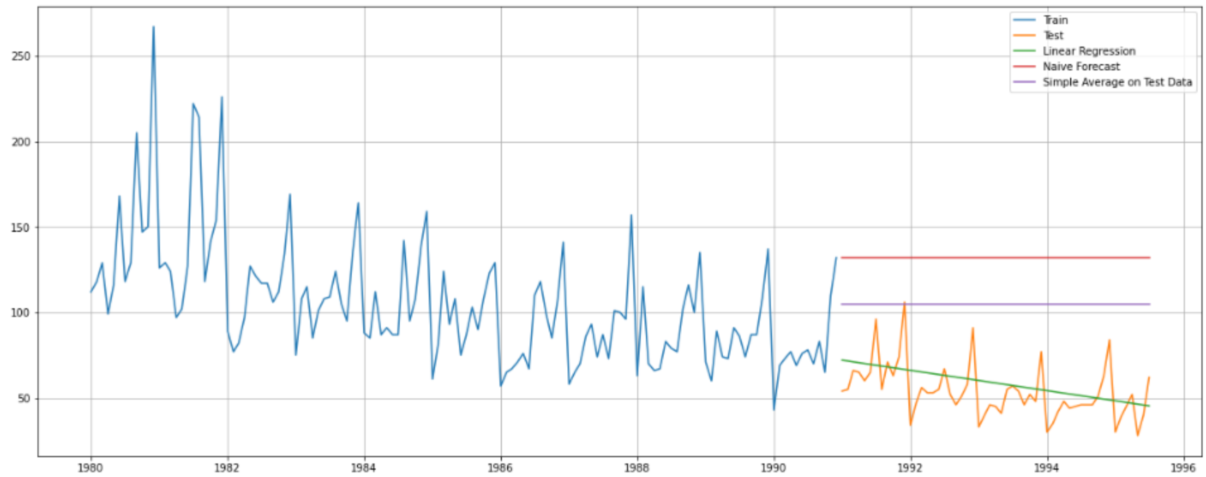

Figure 14: Rose Data Forecast Plot

Moving Average Forecast

	Test RMSE
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315

Table 7: Moving Average RMSE Values of Sparkling Data

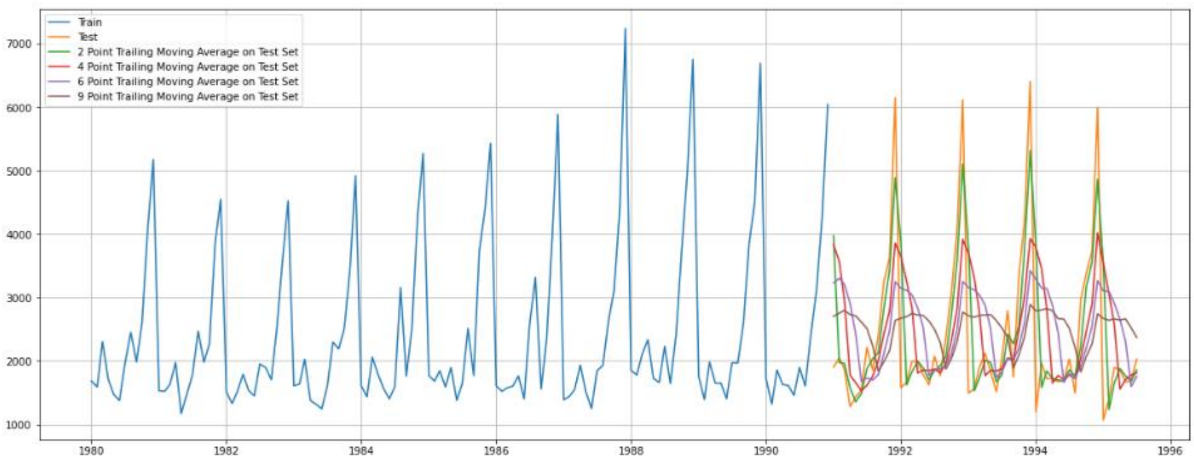

Figure 15: Moving Average Sparkling Data Forecast Plot

	Test RMSE
2pointTrailingMovingAverage	11.529409
4pointTrailingMovingAverage	14.448930
6pointTrailingMovingAverage	14.560046
9pointTrailingMovingAverage	14.724503

Table 8: Moving Average RMSE Values of Rose Data

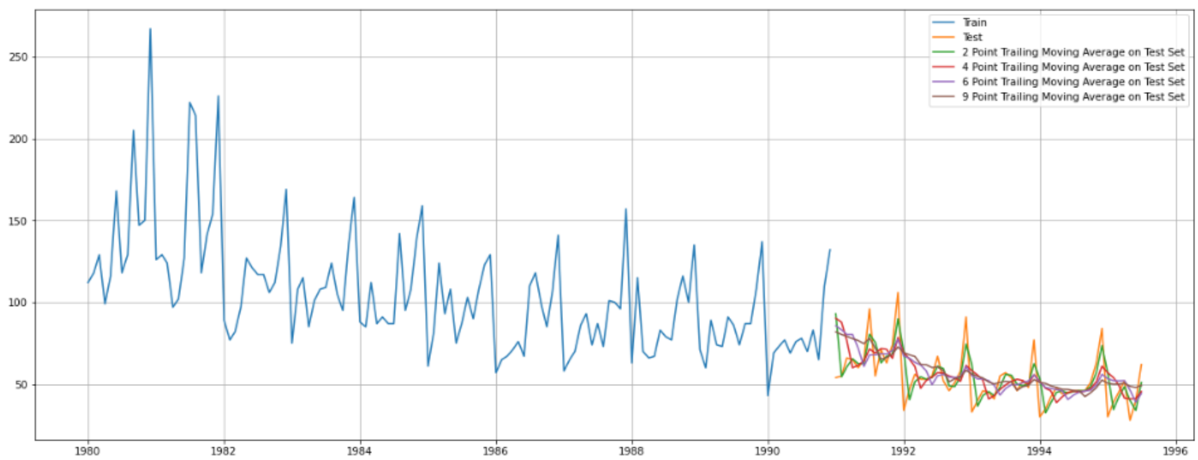

Figure 16: Moving Average Rose Data Forecast Plot

Exponential Smoothening Forecast

	Test RMSE
Simple Exponential Smoothing	1338.008384
Double Exponential Smoothing	5291.879833
TES With Additive Seasonality	378.951023
TES With Multiplicative Seasonality	404.286809

Table 9: Exponential Smoothening RMSE values of Sparkling Data

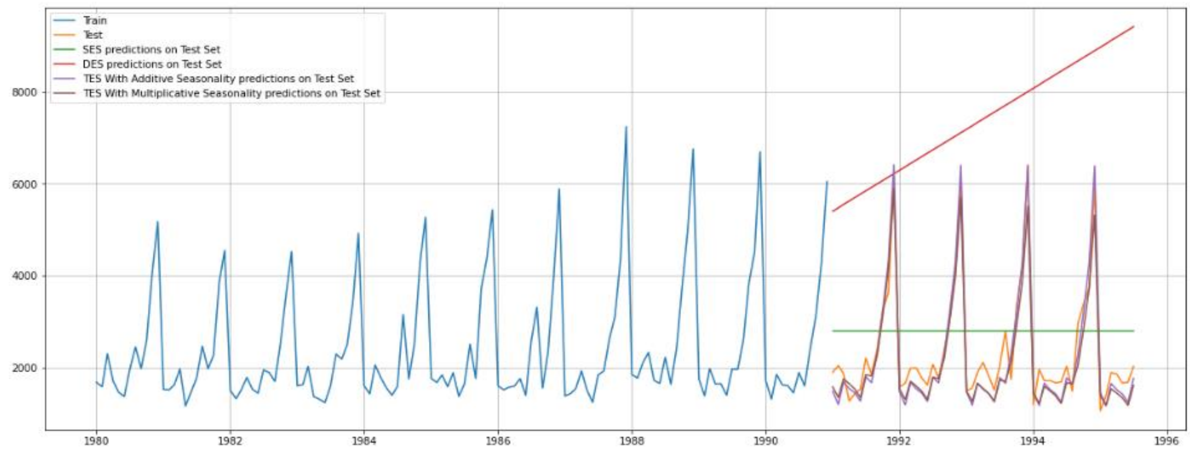

Figure 17: Exponential Smoothening Sparkling Data Forecast Plot

	Test RMSE
Simple Exponential Smoothing	36.775787
Double Exponential Smoothing	15.262498
TES With Additive Seasonality	14.237386
TES With Multiplicative Seasonality	20.132468

Table 10: Exponential Smoothening RMSE values of Rose Data

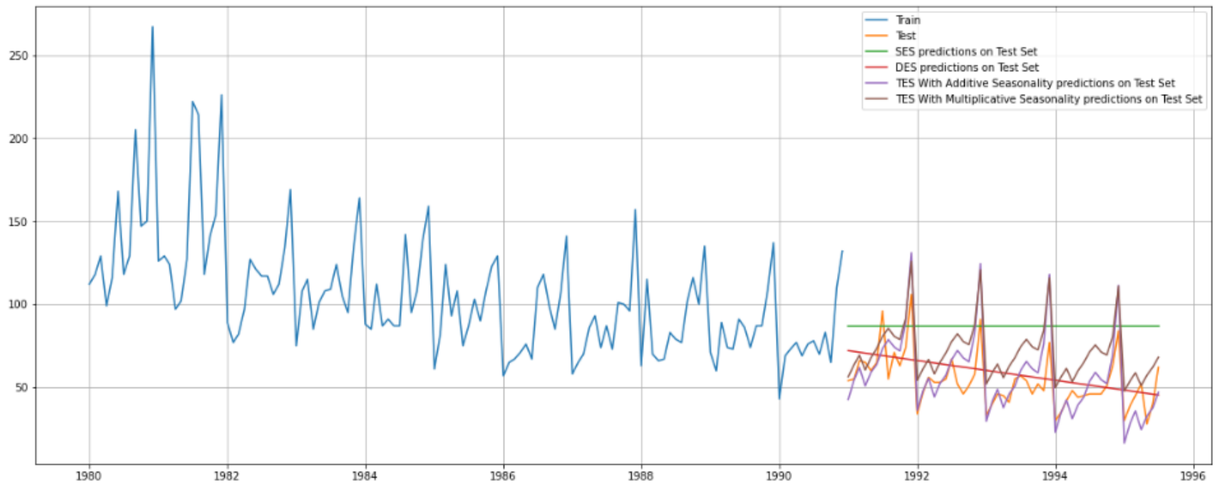

Figure 18: Exponential Smoothening Rose Data Forecast Plot

From the above forecasts,

- Sparkling data has highest RMSE for Double Exponential Smoothening Model and lowest RMSE for Triple Exponential Smoothening Model with Additive Seasonality. So Triple Exponential Smoothening Model with Additive Seasonality is better for the given Sparkling data.
- Rose data has highest RMSE for Naïve Forecast Model and lowest RMSE for 2-point Trailing Moving Average Model. So, 2-point Trailing Moving Average Model is better for the given Rose data.

Following table gives the Exponential smoothening parameters for each of the models.

	Alpha	Beta	Gama
Single Exponential Smoothening	0.07	-	-
Double Exponential Smoothening	0.67	0.0001	-
Triple Exponential Smoothening (Add)	0.11	0.01	0.46
Triple Exponential Smoothening (Mul)	0.11	0.04	0.36

Table 11: Exponential Smoothening Parameters for Sparkling Data

	Alpha	Beta	Gama
Single Exponential Smoothening	0.098	-	-
Double Exponential Smoothening	0	0.16	-
Triple Exponential Smoothening (Add)	0.08	0.0002	0.003
Triple Exponential Smoothening (Mul)	0.07	0.04	0.00007

Table 12: Exponential Smoothening Parameters for Rose Data

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.

The Augmented Dickey-Fuller test is a 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 alpha value where alpha = 0.05.

- We see that at 5% significant level the Sparkling Time Series data is non-stationary. (p-value = 0.567)
- We see that at 5% significant level the Rose Time Series data is non-stationary. (p-value = 0.756)

Let us take one level of differencing to see whether the series becomes stationary.

- We see that at α = 0.05 the Sparkling Time Series with one level of differencing is indeed stationary. (p-value = 8.47 e-11)
- We see that at α = 0.05 the Sparkling Time Series with one level of differencing is indeed stationary. (p-value = 3.89 e-08)

Above we have considered training data, also complete dataset is not stationary for both Sparkling and Rose but with one level of differencing they are stationary.

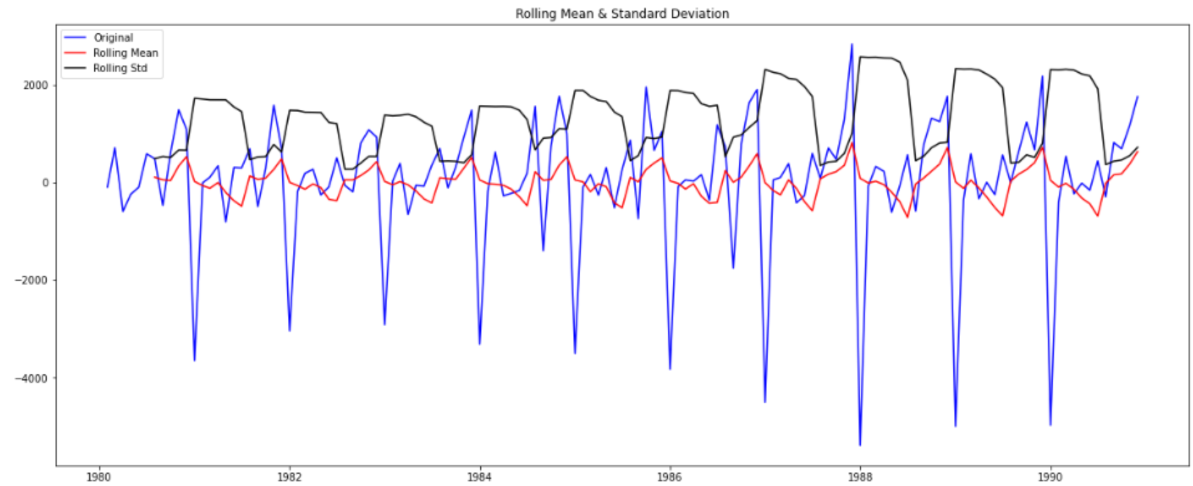

Figure 19: Rolling Statistics plot of Sparkling Data with one level of differencing

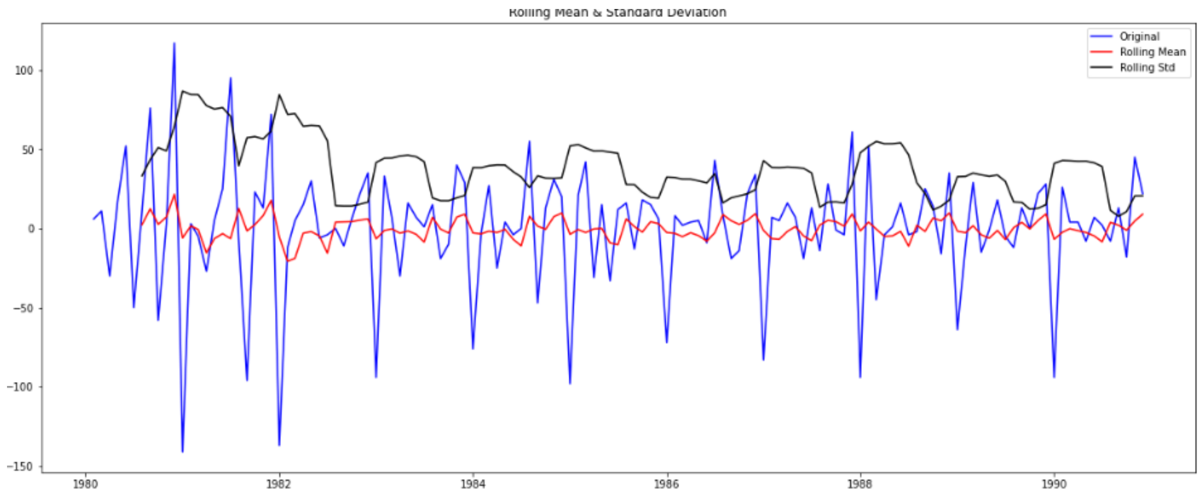

Figure 20: Rolling Statistics plot of Rose Data with one level of differencing

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.

ARIMA models can be built keeping the Akaike Information Criterion (AIC) in mind as well. In this case, we choose the 'p' and 'q' values to determine the AR and MA orders respectively which gives us the lowest AIC value. Lower the AIC better is the model.

After building the ARIMA mode, optimal values for p, d, q with lowest AIC is:

- Sparkling data -> (2,1,2) with AIC of 2210.61
- Rose data -> (0,1,2) with AIC of 1276.83

We will Plot PACF plot to find the seasonality factor before proceeding with SARIMA model.

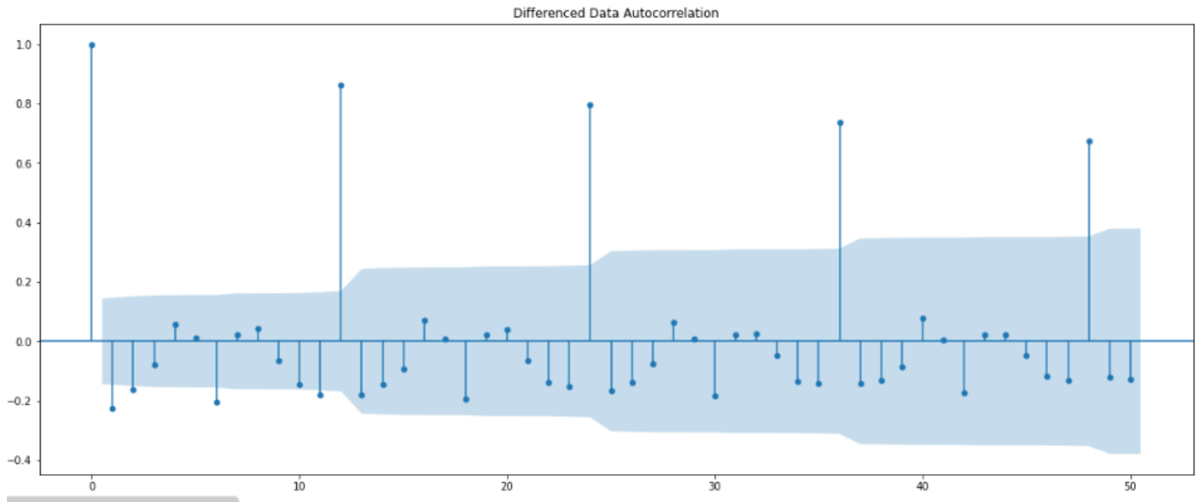

Figure 21: ACF Plot of Sparkling Data

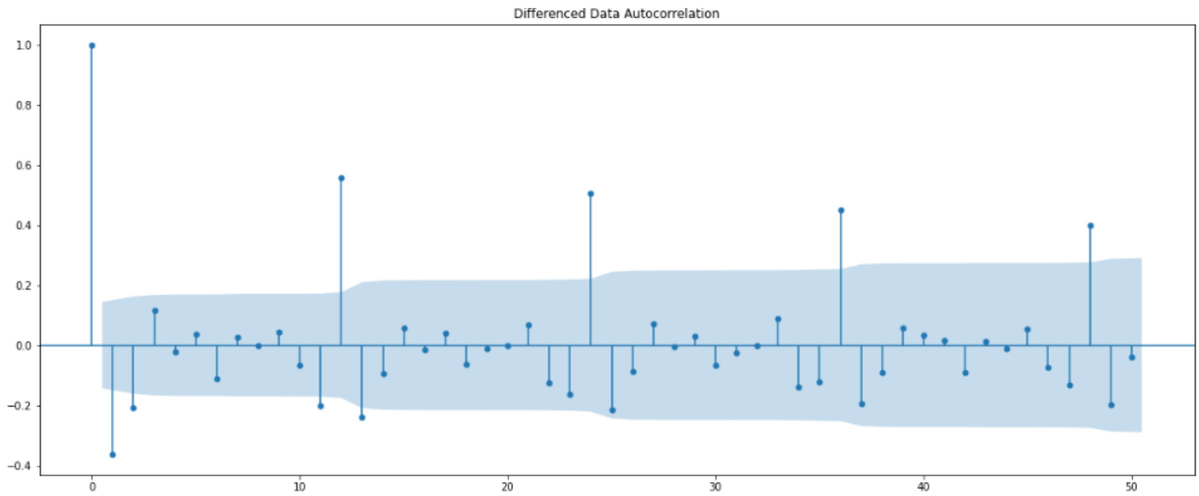

Figure 22: ACF Plot of Rose Data

From the above 2 plots we observe the seasonality of 6 as well as 12 for both Sparkling and Rose Data. We will proceed with 6.

Since the Sparkling data and Rose data with difference level equal to seasonal factor (6) is stationary we take D as 0.

After building the ARIMA mode, optimal values for (p, d, q) and (P, D, Q, seasonal factor) with lowest AIC is:

- Sparkling data -> (0,1,2) (2, 0, 2, 6) with AIC of 1727.88
- Rose data -> (1,1,2) (2, 0, 2, 6) with AIC of 1041.65

	ARIMA (p, d, q)	SARIMA (p, d, q) (P, D, Q, Seasonal Factor)
Sparkling Data	(2, 1, 2)	(0,1,2) (2, 0, 2, 6)
Rose Data	(0, 1, 2)	(1,1,2) (2, 0, 2, 6)

Table 13: Summarising ARIMA - SARIMA Optimal Values

Sparkling Data

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:            D.Sparkling   No. Observations:                  131
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1099.309
Method:                       css-mle   S.D. of innovations           1012.730
Date:                Wed, 06 Jul 2022   AIC                           2210.619
Time:                        13:12:43   BIC                           2227.870
Sample:                    02-01-1980   HQIC                          2217.628
                         - 12-01-1990
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 5.5843      0.518     10.790      0.000       4.570       6.599
ar.L1.D.Sparkling     1.2700      0.074     17.048      0.000       1.124       1.416
ar.L2.D.Sparkling    -0.5604      0.074     -7.620      0.000      -0.704      -0.416
ma.L1.D.Sparkling    -1.9978      0.042    -47.093      0.000      -2.081      -1.915
ma.L2.D.Sparkling     0.9978      0.042     23.501      0.000       0.915       1.081
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.1333           -0.7073j            1.3359           -0.0888
AR.2            1.1333           +0.7073j            1.3359            0.0888
MA.1            1.0004           +0.0000j            1.0004            0.0000
MA.2            1.0019           +0.0000j            1.0019            0.0000
-----------------------------------------------------------------------------
```

Table 14: Auto ARIMA Model Result Summary of Sparkling Data

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 6)   Log Likelihood                -856.944
Date:                            Wed, 06 Jul 2022   AIC                           1727.889
Time:                                    15:06:45   BIC                           1747.164
Sample:                                         0   HQIC                          1735.713
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.7851      0.103     -7.655      0.000      -0.986      -0.584
ma.L2         -0.0976      0.112     -0.871      0.384      -0.317       0.122
ar.S.L6        0.0022      0.026      0.084      0.933      -0.049       0.053
ar.S.L12       1.0396      0.018     58.254      0.000       1.005       1.075
ma.S.L6        0.0427      0.143      0.298      0.766      -0.238       0.324
ma.S.L12      -0.6202      0.090     -6.878      0.000      -0.797      -0.443
sigma2      1.475e+05   1.42e+04     10.372      0.000     1.2e+05    1.75e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                38.96
Prob(Q):                              0.97   Prob(JB):                         0.00
Heteroskedasticity (H):               2.85   Skew:                             0.58
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.59
===================================================================================

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

Table 15: Auto SARIMA Model Result Summary of Sparkling Data

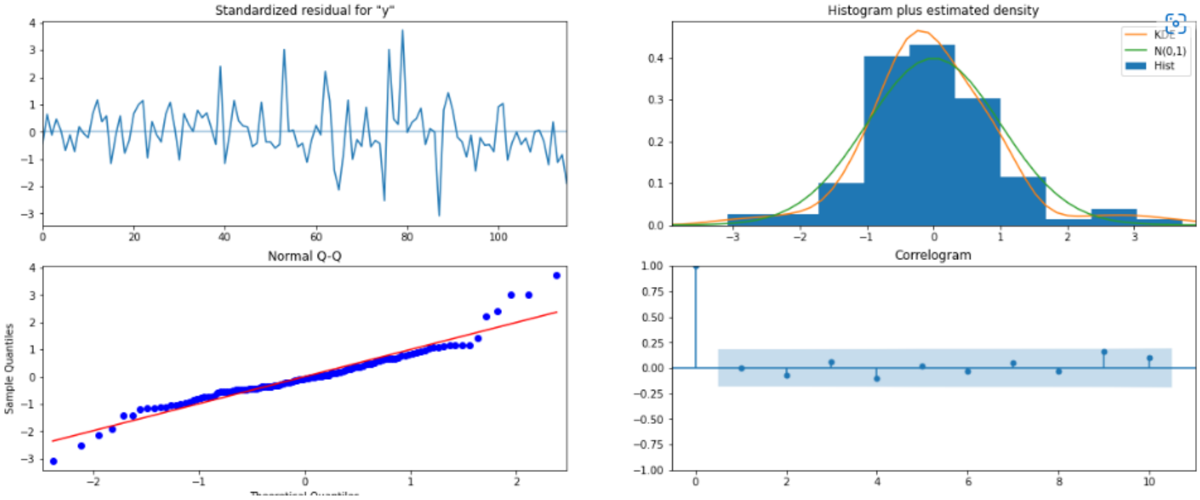

Figure 23: Auto SARIMA Model Diagnostic Plot of Sparkling Data

	RMSE
SARIMA(0, 1, 2)(2, 0, 2, 6)-AIC	601.122857
ARIMA(2, 1, 2)-AIC	1374.546024

Table 16: Auto ARIMA - SARIMA RMSE values of Sparkling Data

Rose Data

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:                 D.Rose   No. Observations:                  131
Model:                 ARIMA(0, 1, 2)   Log Likelihood                -634.418
Method:                       css-mle   S.D. of innovations             30.167
Date:                Wed, 06 Jul 2022   AIC                           1276.835
Time:                        13:12:44   BIC                           1288.336
Sample:                    02-01-1980   HQIC                          1281.509
                         - 12-01-1990
================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const         -0.4886      0.085     -5.742      0.000      -0.655      -0.322
ma.L1.D.Rose  -0.7601      0.101     -7.499      0.000      -0.959      -0.561
ma.L2.D.Rose  -0.2398      0.095     -2.518      0.012      -0.427      -0.053
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
MA.1            1.0001           +0.0000j            1.0001            0.0000
MA.2           -4.1695           +0.0000j            4.1695            0.5000
-----------------------------------------------------------------------------
```

Table 17: Auto ARIMA Model Result Summary of Rose Data

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(1, 1, 2)x(2, 0, 2, 6)   Log Likelihood                -512.828
Date:                            Wed, 06 Jul 2022   AIC                           1041.656
Time:                                    15:08:58   BIC                           1063.685
Sample:                                         0   HQIC                          1050.598
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5940      0.152     -3.900      0.000      -0.892      -0.295
ma.L1         -0.1954    939.337     -0.000      1.000   -1841.262    1840.872
ma.L2         -0.8047    755.878     -0.001      0.999   -1482.298    1480.689
ar.S.L6       -0.0626      0.035     -1.764      0.078      -0.132       0.007
ar.S.L12       0.8451      0.039     21.884      0.000       0.769       0.921
ma.S.L6        0.2226    775.183      0.000      1.000   -1519.108    1519.554
ma.S.L12      -0.7774    602.586     -0.001      0.999   -1181.824    1180.269
sigma2       335.2013    3.9e+05      0.001      0.999   -7.64e+05    7.65e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.07   Jarque-Bera (JB):                56.68
Prob(Q):                              0.78   Prob(JB):                         0.00
Heteroskedasticity (H):               0.47   Skew:                             0.52
Prob(H) (two-sided):                  0.02   Kurtosis:                         6.26
===================================================================================

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

Table 18: Auto SARIMA Model Result Summary of Rose Data

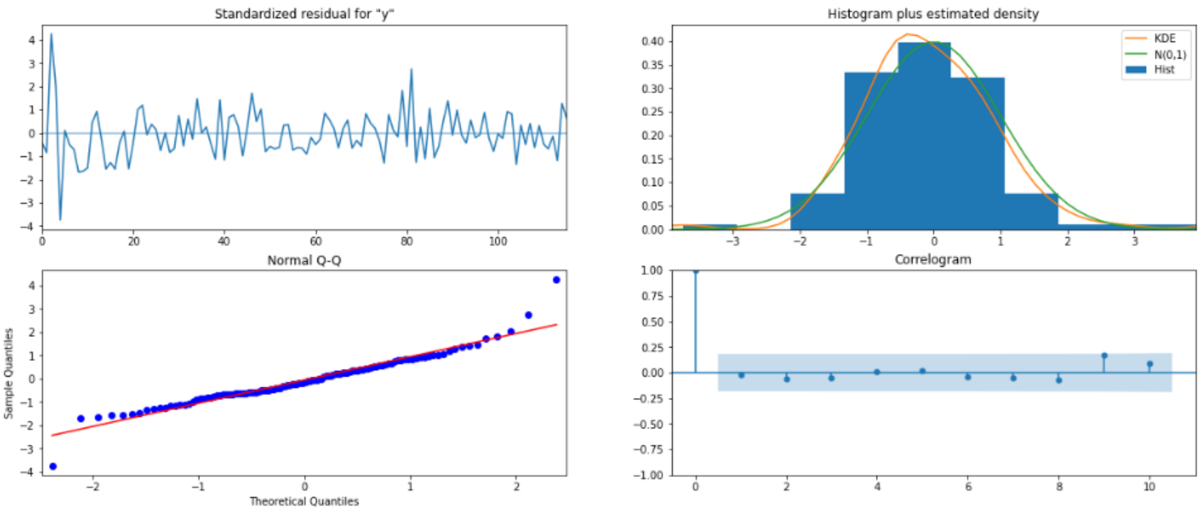

Figure 24: SARIMA Model Diagnostic Plot of Rose Data

	RMSE
SARIMA(1, 1, 2)(2, 0, 2, 6)-AIC	26.111408
ARIMA(0, 1, 2)-AIC	15.611357

Table 19: ARIMA - SARIMA RMSE values of Rose Data

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.

For both Sparkling and Rose data, data as it is was not stationary but data with 1 level of differencing is stationary so d = 1.

Sparkling data

Let us plot ACF and PACF plot and find the values for p and q based on the cut off

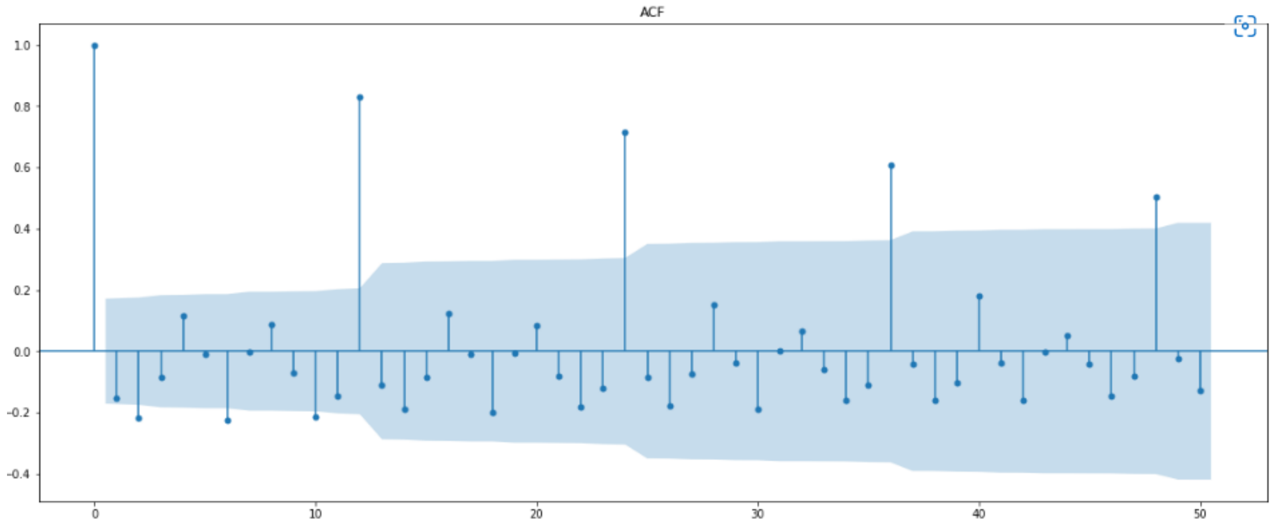

Figure 25: ACF Plot of Sparkling Training Data with 1 Level Differencing

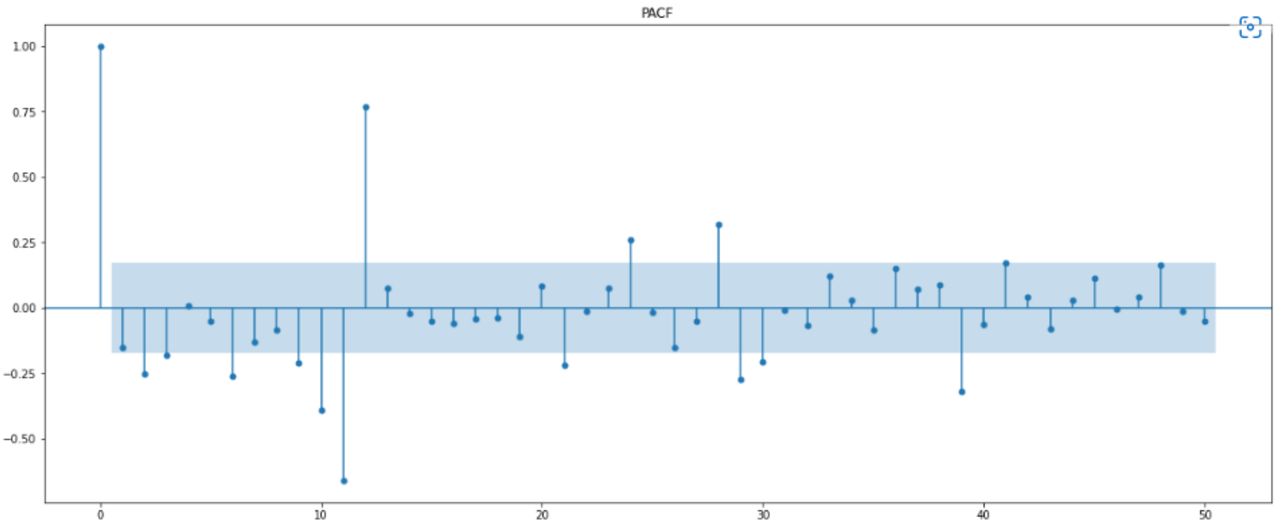

Figure 26: PACF Plot of Sparkling Training Data with 1 Level Differencing

- 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.

So, for ARIMA model (p, d, q) is (0, 1, 0)

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:            D.Sparkling   No. Observations:                  131
Model:                 ARIMA(0, 1, 0)   Log Likelihood               -1132.791
Method:                           css   S.D. of innovations           1377.911
Date:                Wed, 06 Jul 2022   AIC                           2269.583
Time:                        16:35:23   BIC                           2275.333
Sample:                    02-01-1980   HQIC                          2271.919
                         - 12-01-1990
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         33.2901    120.389      0.277      0.782    -202.667     269.248
==============================================================================
```

Table 20: Manual ARIMA Model Result Summary of Sparkling Data

Since we observe a seasonality of 12 we plot a ACF and PACF plot for Data with level of difference equal to 12 to find P and Q based on the cut off.

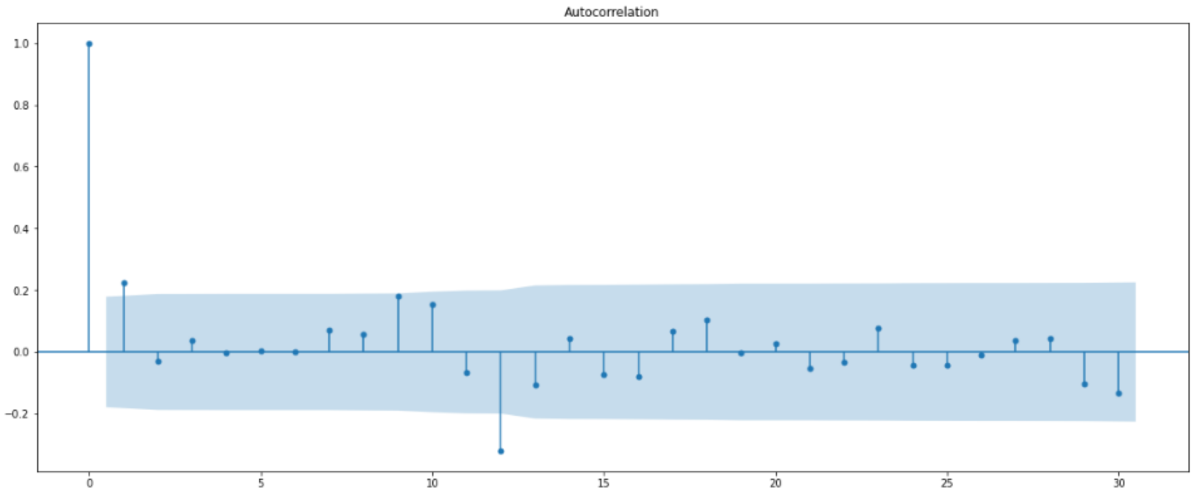

Figure 27: ACF Plot of Sparkling Training Data with 12 Level Differencing

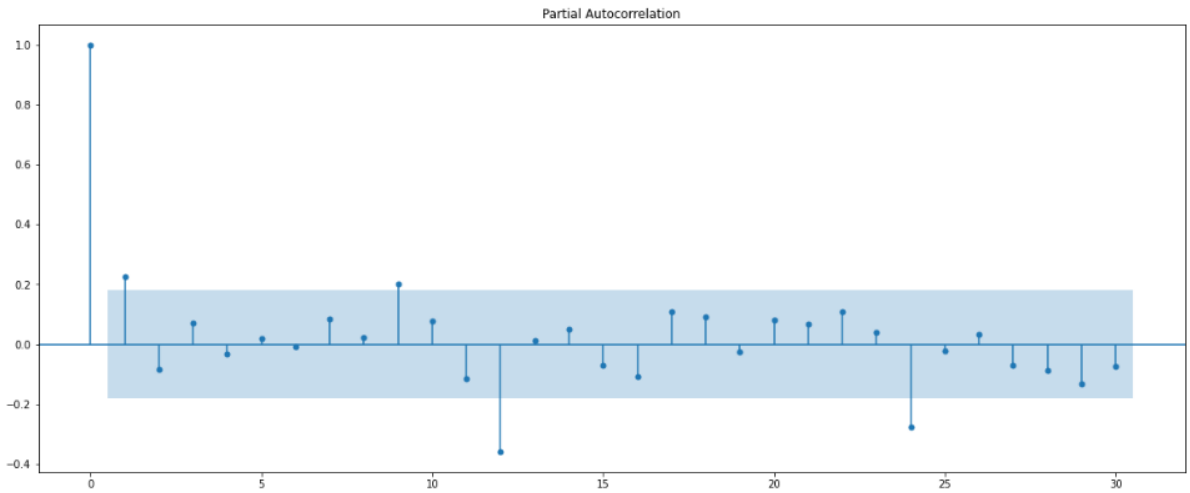

Figure 28: PACF Plot of Sparkling Training Data with 12 Level Differencing

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

So, for SARIMA model (p, d, q) (P, D, Q, seasonal Factor) is (0, 1, 0) (1, 0, 1, 12)

```
                                     SARIMAX Results                                      
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(0, 1, 0)x(1, 0, [1], 12)   Log Likelihood                -900.495
Date:                            Wed, 06 Jul 2022   AIC                           1806.991
Time:                                    17:16:17   BIC                           1815.303
Sample:                                         0   HQIC                          1810.365
                                            - 132                                         
Covariance Type:                              opg                                         
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.S.L12       1.0325      0.019     52.957      0.000       0.994       1.071
ma.S.L12      -0.5384      0.078     -6.896      0.000      -0.691      -0.385
sigma2      2.463e+05   2.34e+04     10.520      0.000       2e+05    2.92e+05
===================================================================================
Ljung-Box (L1) (Q):                  19.69   Jarque-Bera (JB):                31.97
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               1.88   Skew:                             0.66
Prob(H) (two-sided):                  0.05   Kurtosis:                         5.18
===================================================================================

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

Table 21: Manual SARIMA Model Result Summary of Sparkling Data

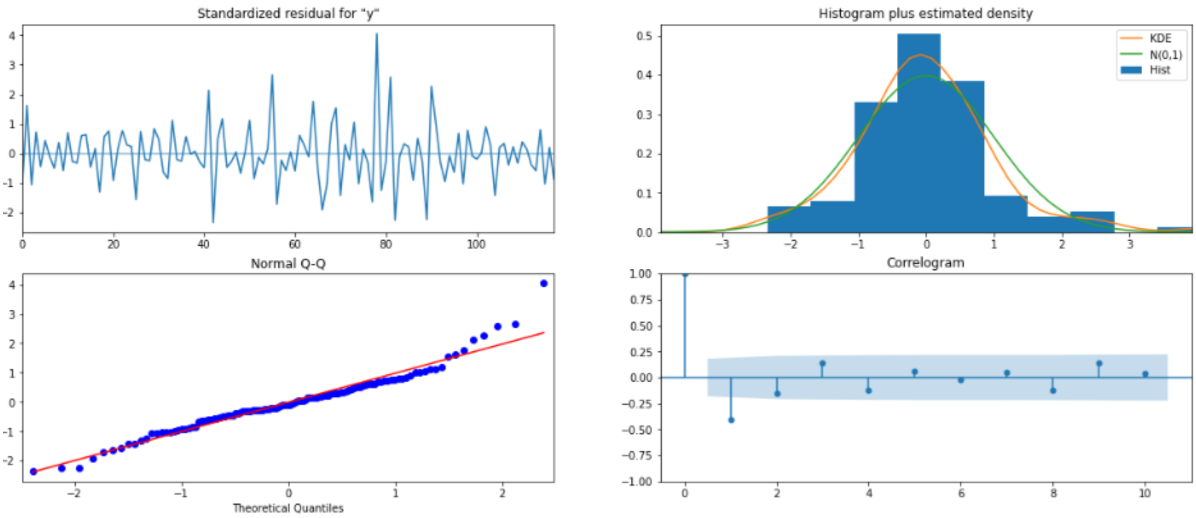

Figure 29: Manual SARIMA Model Diagnostic Plot of Sparkling Data

	RMSE
ARIMA(0, 1, 0)-Manual	4779.154299
SARIMA(0, 1, 0)(1, 0, 1, 12)-Manual	1787.706713

Table 22: Manual ARIMA - SARIMA RMSE values of Sparkling Data

Rose Data

Let us plot ACF and PACF plot and find the values for p and q based on the cut off

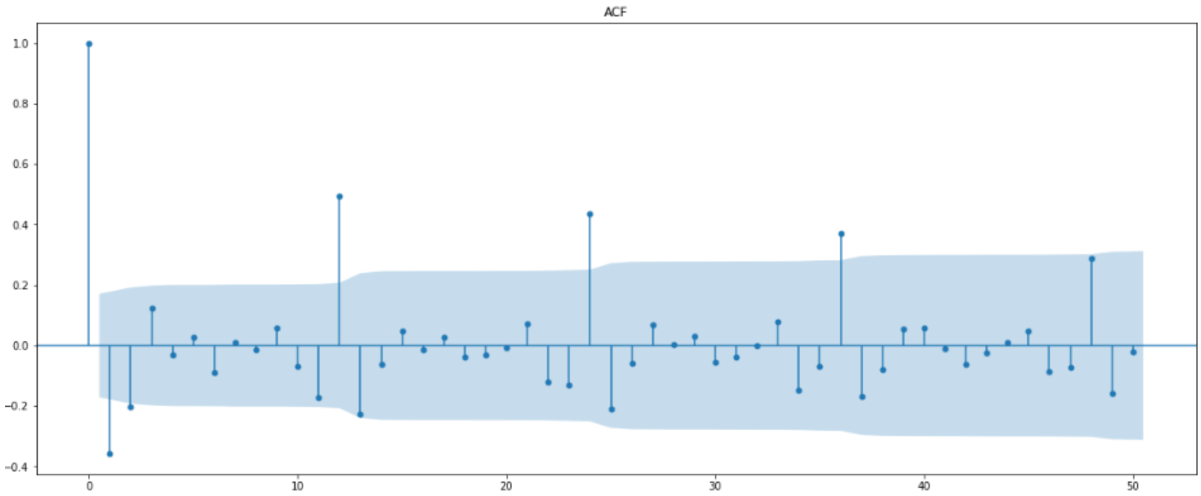

Figure 30: ACF Plot of Rose Training Data with 1 Level Differencing

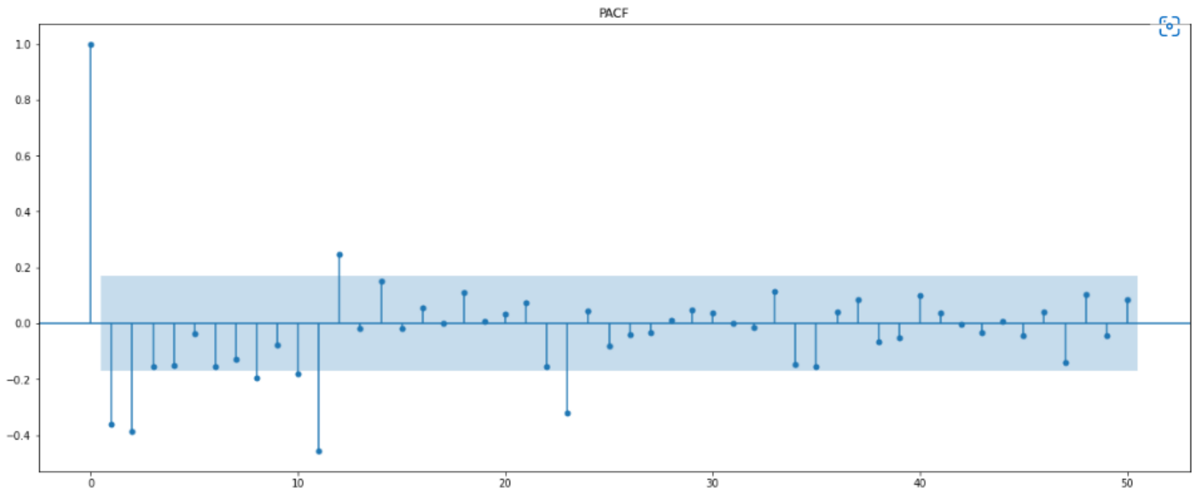

Figure 31: PACF Plot of Rose Training Data with 1 Level Differencing

- 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 2.

So, for ARIMA model (p, d, q) is (2, 1, 2)

```
                              ARIMA Model Results
==============================================================================
Dep. Variable:                 D.Rose   No. Observations:                  131
Model:                 ARIMA(2, 1, 2)   Log Likelihood                -633.649
Method:                       css-mle   S.D. of innovations             29.975
Date:                Wed, 06 Jul 2022   AIC                           1279.299
Time:                        17:35:21   BIC                           1296.550
Sample:                    02-01-1980   HQIC                          1286.309
                         - 12-01-1990
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const           -0.4911      0.081     -6.076      0.000      -0.649      -0.333
ar.L1.D.Rose    -0.4383      0.218     -2.015      0.044      -0.865      -0.012
ar.L2.D.Rose     0.0269      0.109      0.246      0.806      -0.188       0.241
ma.L1.D.Rose    -0.3316      0.203     -1.633      0.102      -0.729       0.066
ma.L2.D.Rose    -0.6684      0.201     -3.332      0.001      -1.062      -0.275
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1           -2.0290           +0.0000j            2.0290            0.5000
AR.2           18.3389           +0.0000j           18.3389            0.0000
MA.1            1.0000           +0.0000j            1.0000            0.0000
MA.2           -1.4961           +0.0000j            1.4961            0.5000
-----------------------------------------------------------------------------
```

Table 23: Manual ARIMA Model Result Summary of Rose Data

Since we observe a seasonality of 12 we plot a ACF and PACF plot for Data with level of difference equal to 12 to find P and Q based on the cut off.

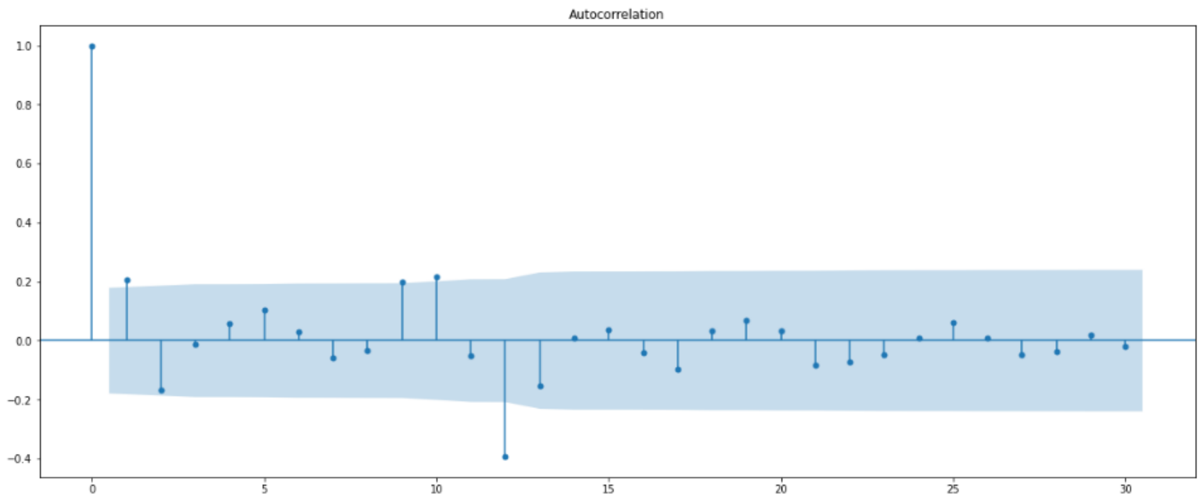

Figure 32: ACF Plot of Rose Training Data with 12 Level Differencing

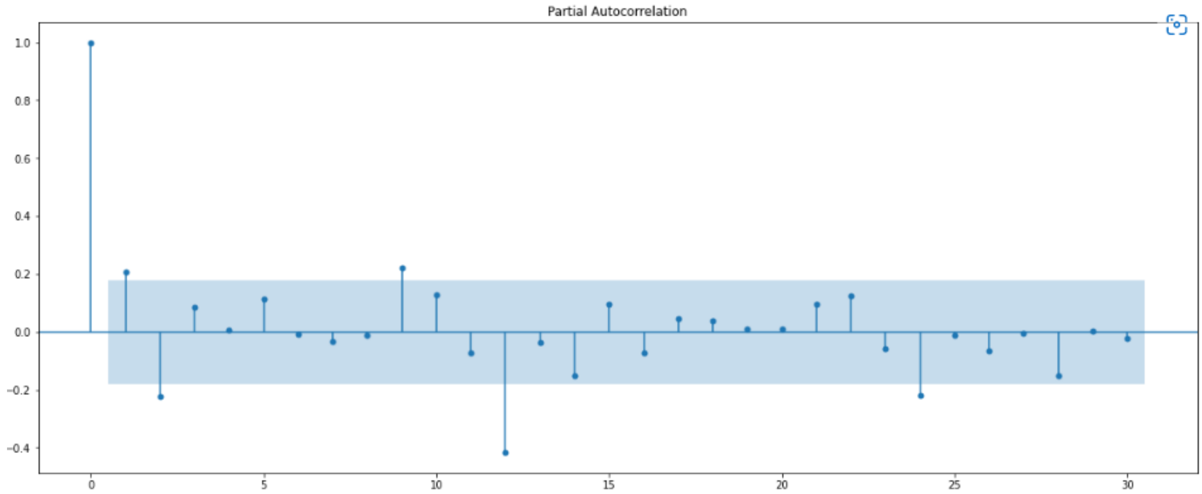

Figure 33: PACF Plot of Rose Training Data with 12 Level Differencing

- The Auto-Regressive parameter in a SARIMA model is 'P' which comes from the significant lag before which the PACF plot cuts-off to 2.
- The Moving-Average parameter in a SARIMA model is 'Q' which comes from the significant lag before the ACF plot cuts-off to 1.

So, for SARIMA model (p, d, q) (P, D, Q, seasonal Factor) is (2, 1, 2) (2, 0, 1, 12)

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(2, 1, 2)x(2, 0, [1], 12)   Log Likelihood                -441.189
Date:                            Wed, 06 Jul 2022   AIC                            898.378
Time:                                    19:03:27   BIC                            919.610
Sample:                                         0   HQIC                           906.982
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.4772      0.305      1.564      0.118      -0.121       1.075
ar.L2         -0.1667      0.104     -1.608      0.108      -0.370       0.037
ma.L1         -1.3270    391.036     -0.003      0.997    -767.744     765.090
ma.L2          0.3270    127.912      0.003      0.998    -250.377     251.031
ar.S.L12       0.3280      0.082      3.983      0.000       0.167       0.489
ar.S.L24       0.2831      0.070      4.04[illegible]00       0.146       0.420
ma.S.L12       0.1309      0.131      0.998      0.318      -0.126       0.388
sigma2       248.8255   9.73e+04      0.003      0.998    -1.9e+05    1.91e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                 2.96
Prob(Q):                              0.90   Prob(JB):                         0.23
Heteroskedasticity (H):               1.01   Skew:                             0.37
Prob(H) (two-sided):                  0.99   Kurtosis:                         3.34
===================================================================================

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

Table 24: Manual SARIMA Model Result Summary of Rose Data

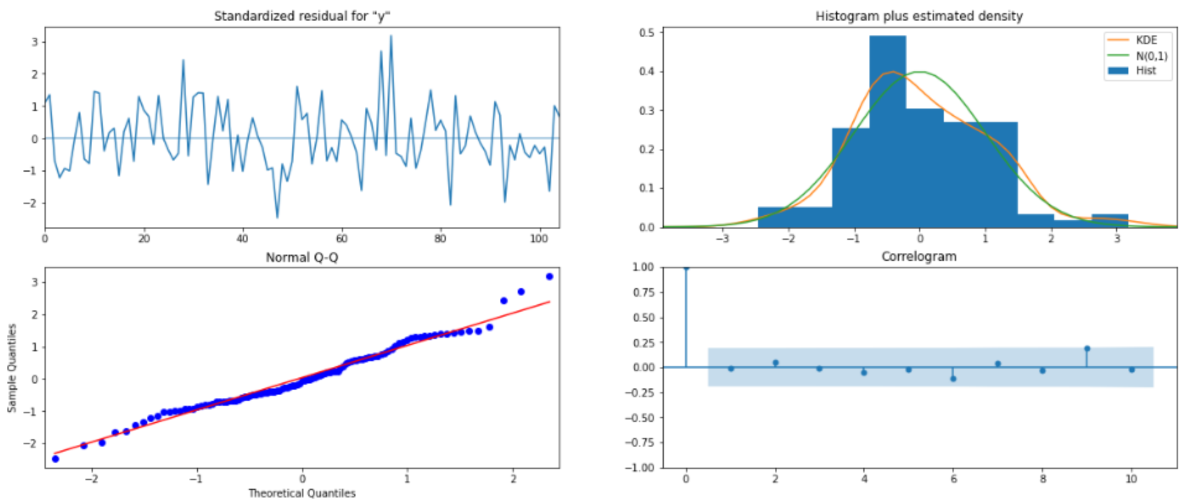

Figure 34: Manual SARIMA Model Diagnostic Plot of Rose Data

	RMSE
ARIMA(2, 1, 2)-Manual	15.348707
SARIMA(2, 1, 2)(2, 0, 1, 12)-Manual	28.199343

Table 25: Manual ARIMA - SARIMA RMSE values of Rose Data

Q8 Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

Sparkling Data

	RMSE
TES With Additive Seasonality	378.951023
TES With Multiplicative Seasonality	404.286809
SARIMA(0, 1, 2)(2, 0, 2, 6)-AIC	601.122857
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
SimpleAverageModel	1275.081804
6pointTrailingMovingAverage	1283.927428
Simple Exponential Smoothing	1338.008384
9pointTrailingMovingAverage	1346.278315
ARIMA(2, 1, 2)-AIC	1374.546024
RegressionOnTime	1389.135175
SARIMA(0, 1, 0)(1, 0, 1, 12)-Manual	1787.706713
NaiveModel	3864.279352
ARIMA(0, 1, 0)-Manual	4779.154299
Double Exponential Smoothing	5291.879833

Table 26: Sparkling Data RMSE Values on the Test Data

Rose Data

	RMSE
2pointTrailingMovingAverage	11.529409
TES With Additive Seasonality	14.237386
4pointTrailingMovingAverage	14.448930
6pointTrailingMovingAverage	14.560046
9pointTrailingMovingAverage	14.724503
Double Exponential Smoothing	15.262498
RegressionOnTime	15.262509
ARIMA(2, 1, 2)-Manual	15.348707
ARIMA(0, 1, 2)-AIC	15.611357
TES With Multiplicative Seasonality	20.132468
SARIMA(1, 1, 2)(2, 0, 2, 6)-AIC	26.111408
SARIMA(2, 1, 2)(2, 0, 1, 12)-Manual	28.199343
Simple Exponential Smoothing	36.775787
SimpleAverageModel	53.440426
NaiveModel	79.699093

Table 27: Rose Data RMSE Values on the Test Data

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

Sparkling Data

From the Table 27 we observe Triple Exponential Smoothing with additive seasonality is the optimal model for given Sparkling dataset which has least RMSE value compared to other models built.

So, using this model we will forecast the data for next 12 months.

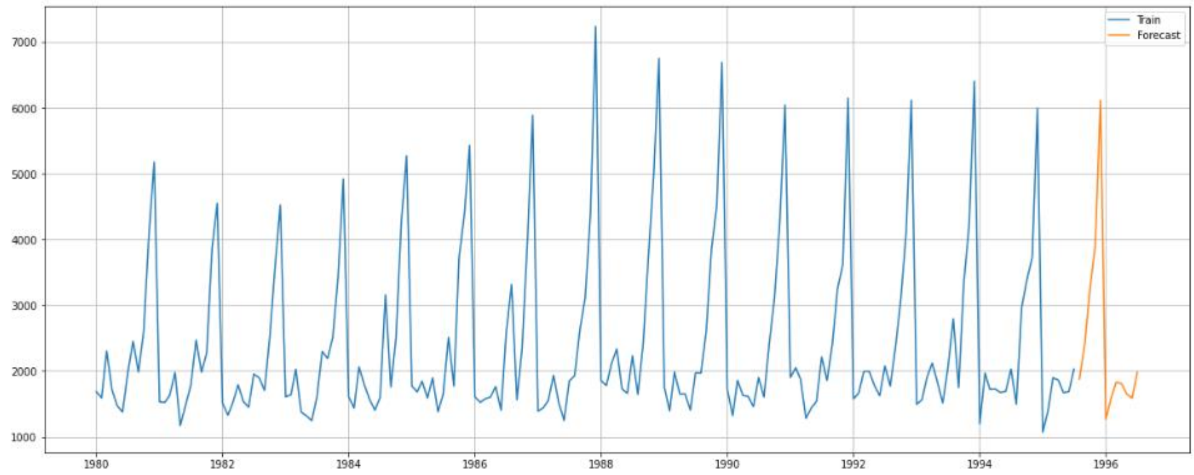

Figure 35: Sparkling Data Forecast Using Optimal Model

Rose Data

From the Table 28 we observe Triple Exponential Smoothing with additive seasonality is the 2nd optimal model for given Rose dataset which has least RMSE value compared to other models built.

2 Point Trailing Moving average was one with least RMSE.

Here using Triple Exponential Smoothing with additive seasonality model we will forecast the data for next 12 months.

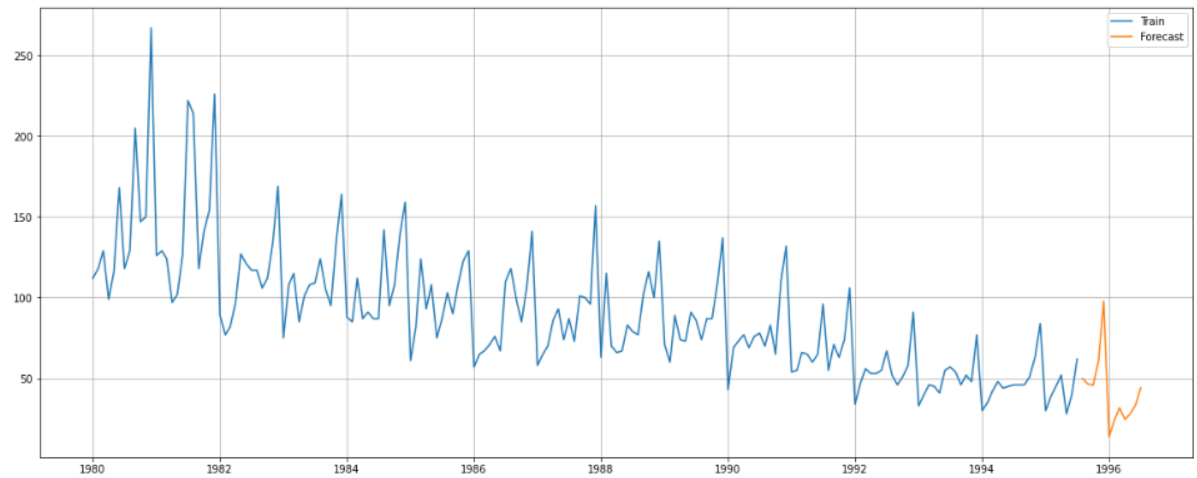

Figure 36: Rose Data Forecast Using Optimal Model

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

Here Triple Exponential Smoothing with additive seasonality model has been selected as optimal Model.

	Alpha	Beta	Gama
Triple Exponential Smoothening (Add) – Sparkling Data	0.11	0.01	0.46
Triple Exponential Smoothening (Mul) – Rose Data	0.07	0.04	0.00007

Figure 37: Optimal Model with Optimal Values

Triple Exponential Smoothing with additive seasonality had RMSE of 404.28 and 14.23 for Sparkling and Rose data respectively.

Suggestions:

- From Figure 4 we observe Sales of Rose is decreasing year by year, so company can give more offers on Rose wines and market about the product so that sales increase.
- From Figure 5 and Figure 6 we observe sales of Sparkling and Rose are more in December month Compared to other moths. So, they can increase their production in December month and give more offers and attract customers in other months.
- Market about the products to increase visibility.
- Promote the products by promoting health benefits of wine

Thank You