2021

Time Series Forecasting

Prajoth

Great Learning

Table of Contents

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.

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

Checking the head and tail of Dataset:

Sparkling:

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

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

Observation:

- Dataset has 187 rows and 2 columns.
- Dataset has both numerical and categorical values.
- Dataset is no null Values.
- Dataset is no duplicate values

Rose:

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

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

Inferences:

- Dataset has 187 rows and 2 columns.
- Dataset has both numerical and categorical values.
- Dataset has 2 null Values which is to be treated using Interpolate(Interpolation is mostly used to impute missing values in the dataframe or series while preprocessing data).
- There is no duplicate values.

Check the basic measures of descriptive statistics of the Time Series:

	Sparkling
count	187.000
mean	2402.417
std	1295.112
min	1070.000
25%	1605.000
50%	1874.000
75%	2549.000
max	7242.000

	Rose
count	185.000
mean	90.395
std	39.175
min	28.000
25%	63.000
50%	86.000
75%	112.000
max	267.000

Observation:

- Sparkling has min sales 1070 and max sales 7242 per month.
- Rose has min sales 28 and max sales 267 per month.

- Mean and Median is almost same for rose dataset which is to be normally distributed

Creating the Time Stamps and adding to the data frame to make it a Time Series Data:

Timestamp is created from years(1980-1995) with the frequency of month.Timestamp column loaded in dataset.

	YearMonth	Sparkling	Time_Stamp
0	1980-01	1686	1980-01-31
1	1980-02	1591	1980-02-29
2	1980-03	2304	1980-03-31
3	1980-04	1712	1980-04-30
4	1980-05	1471	1980-05-31

	YearMonth	Rose	Time_Stamp
0	1980-01	112.0	1980-01-31
1	1980-02	118.0	1980-02-29
2	1980-03	129.0	1980-03-31
3	1980-04	99.0	1980-04-30
4	1980-05	116.0	1980-05-31

Dropping Yearmonth Column:

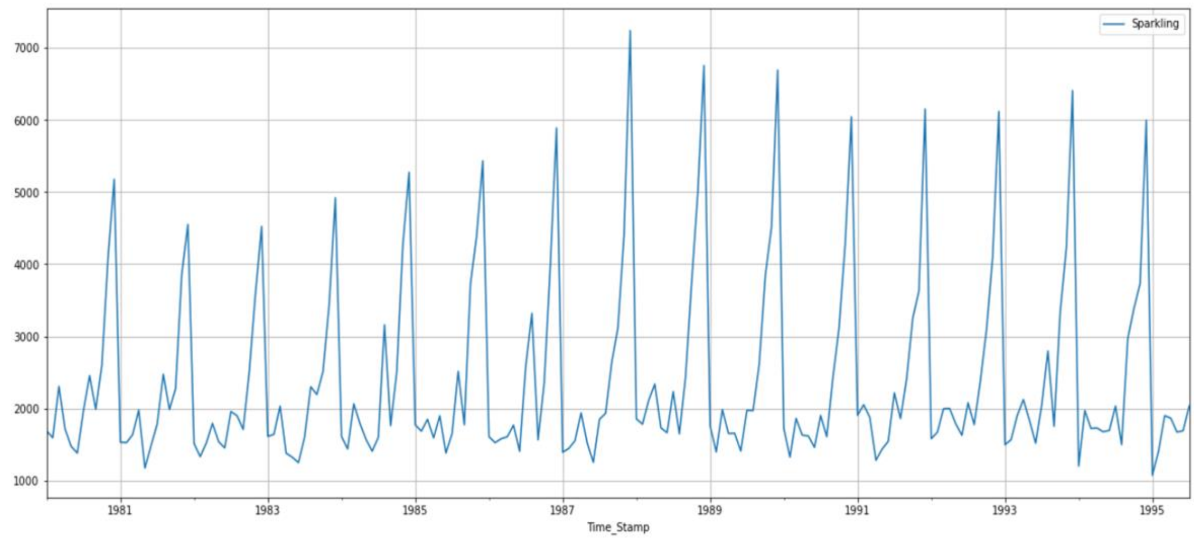
- We removed year-month column for further forecasting data.

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

	Rose
Time_Stamp	
1980-01-31	112.0
1980-02-29	118.0
1980-03-31	129.0
1980-04-30	99.0
1980-05-31	116.0

<u>Plot the Time Series to understand the behaviour of the data:</u>

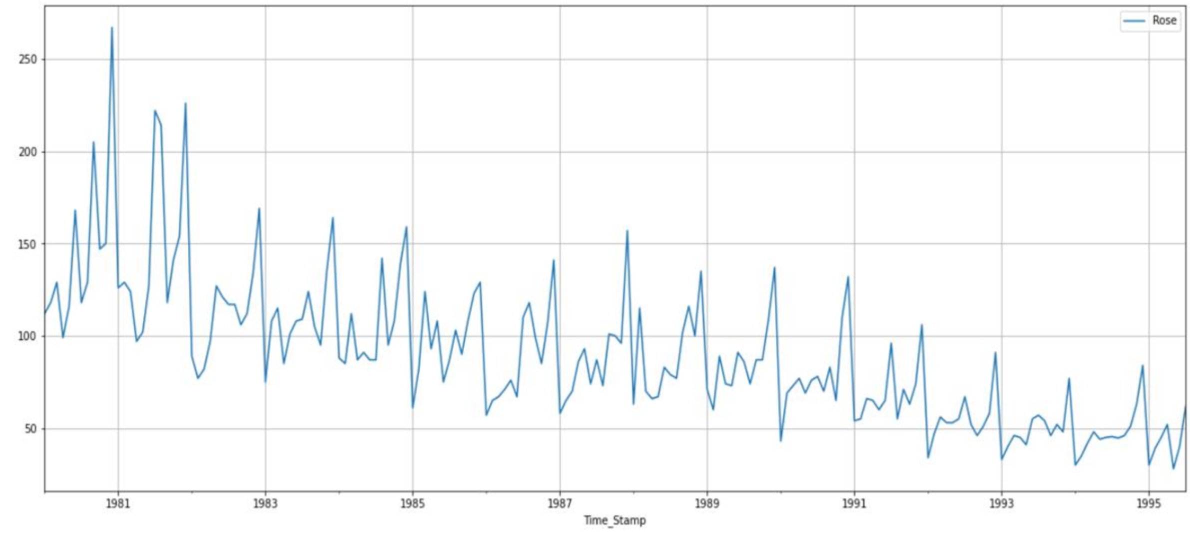
Sparkling:



Inferences:

- As per plot data has both trend and seasonality but there is no gradual drop in sales.It has moderate sales from start to end.

Rose:



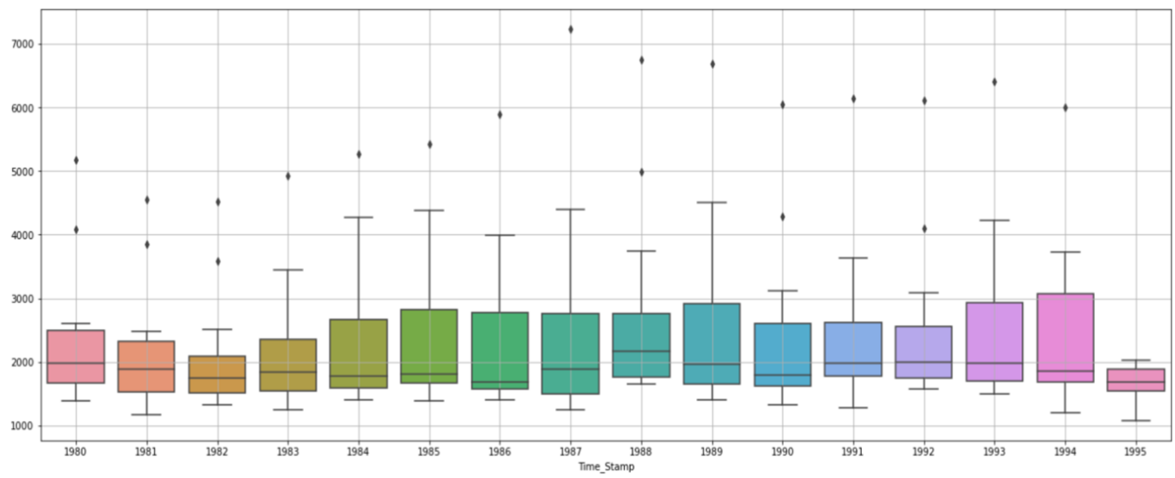
Inferences:

- As per plot data has both trend and seasonality but there is gradual drop in sales.It has decreasing trend from beginning.

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

Yearly Boxplot:

Sparkling:

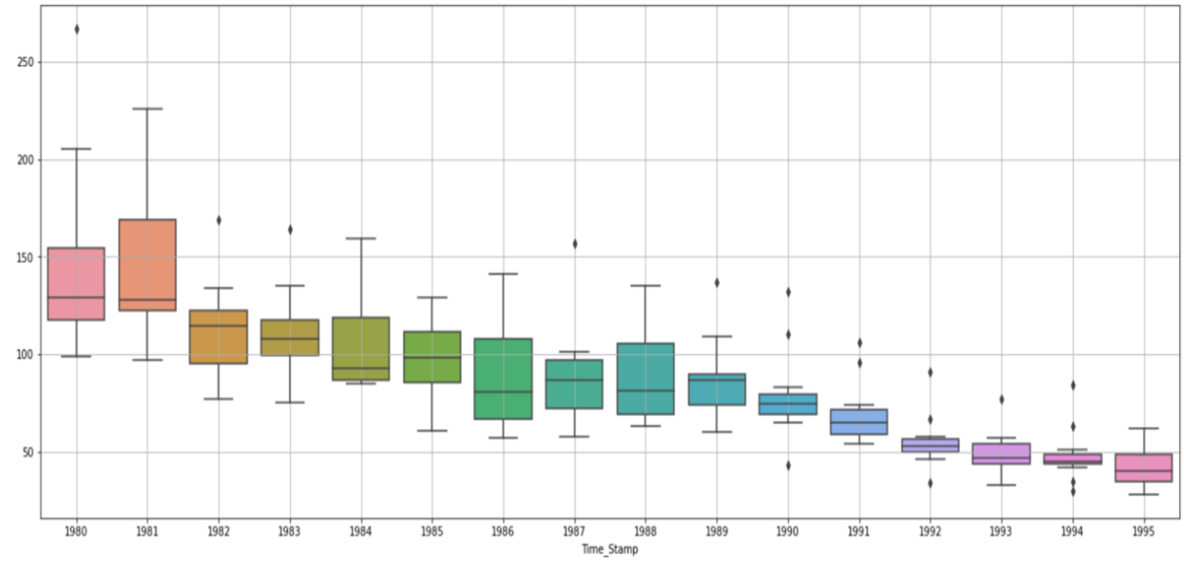


Inferences:

- There are outliers in data.
- Looking at median value sales are good at year 1988.
- Sales are low at 1995 which has 7 month of data apart.

- In year 1994 sales is good compared to previous years but has low median.

Rose:

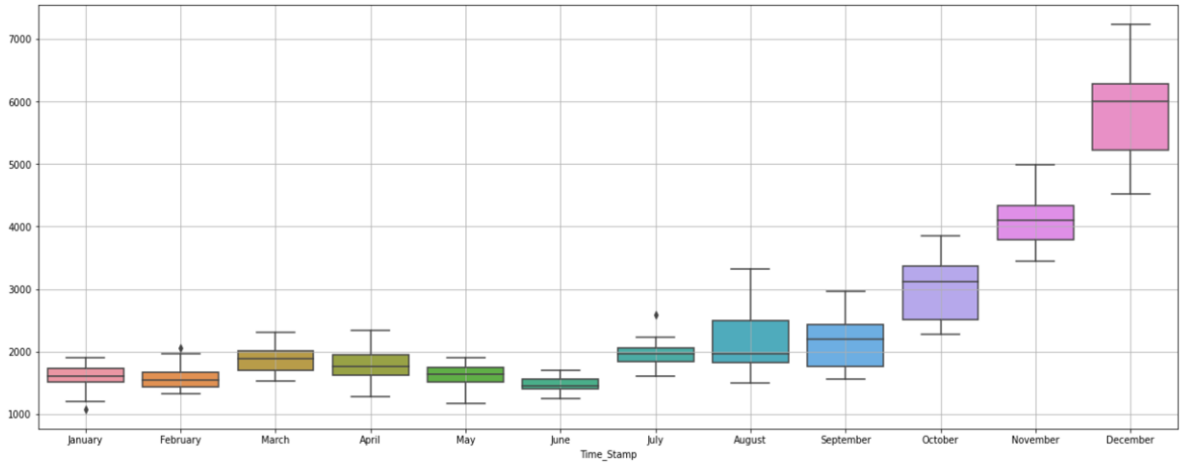


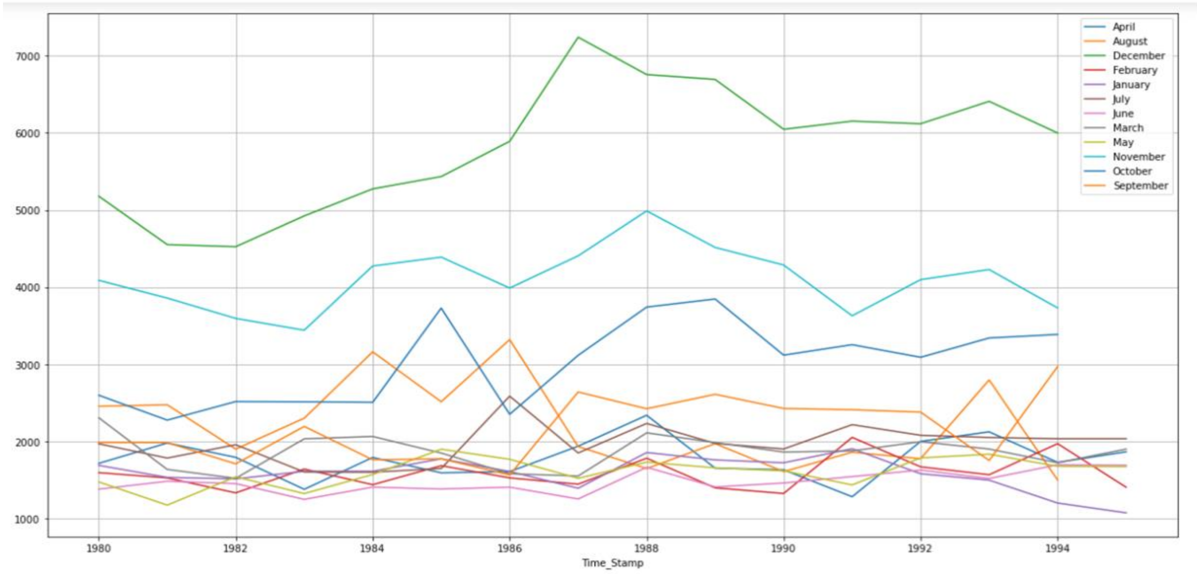
Observation:

- There are outliers in data.
- Sales has decreasing trend from beginning.
- Looking at median 1980 has good sales compared to previous years.
- Sales very low in 1995
- The yearly boxplots also shows that the Sales have decreased towards the last few years.

Monthly Boxplot:

Sparkling:

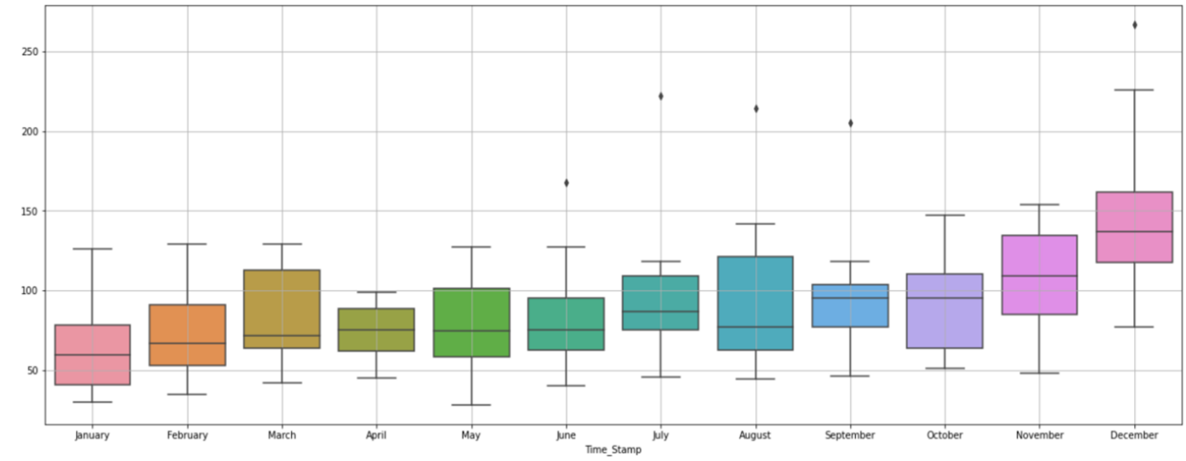


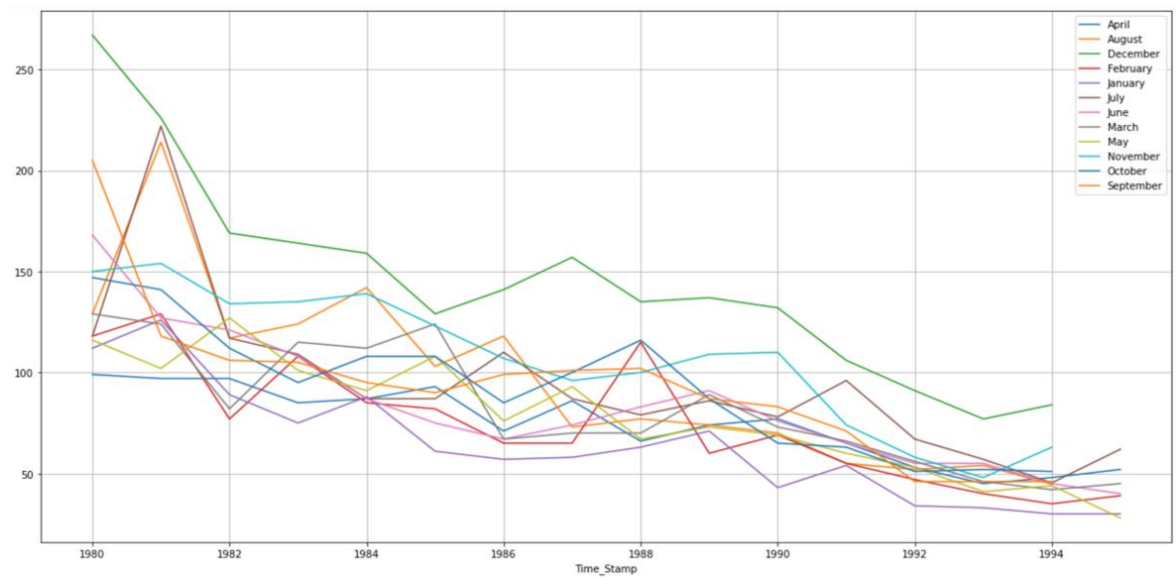


Inference:

- Sales are gradually increasing from month July to December which is second half of every year. Sales are very low in first half of every year.
- December has recorded higher number of sales and January has lower sales.

Rose:





:

Inference:

- Customer preferred to buy rose wines in month of December which is recorded higher number of sales.
- January has recorded lower number of wine sales.

Decompose the Time Series:

Decomposition is a forecasting technique that separates or decomposes historical data into different components and uses them to create a forecast that is more accurate than a simple trend line

Additive model is useful when the seasonal variation is relatively constant over time.

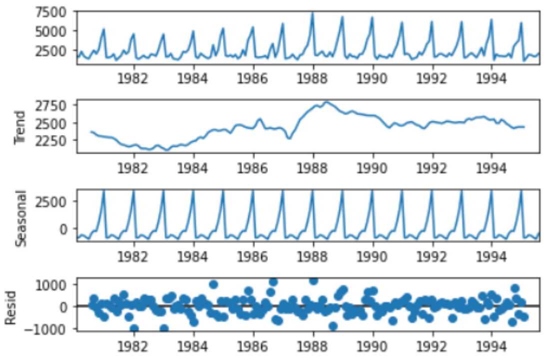
Additive= Trend + Seasonal + Random

Multiplicative model is useful when the seasonal variation increases over time

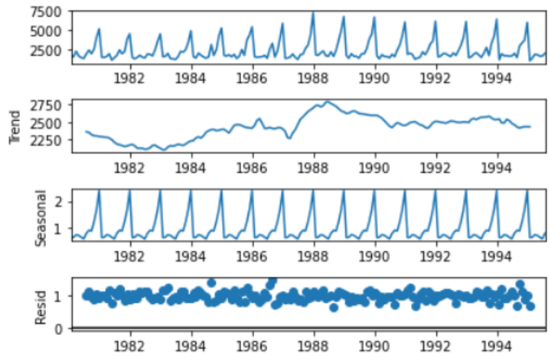
Multiplicative = Trend * Seasonal * Random

Sparkling:

Additive



Multiplicative

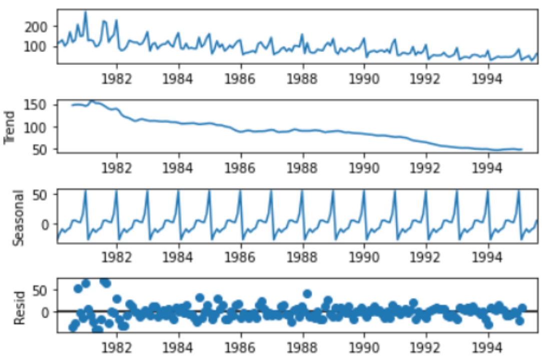


Inferences:

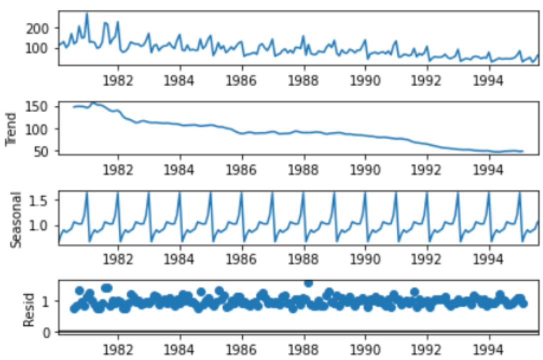
- There is trend and seasonality in wine sales.
- Residuals is around 0 for additive and 1 for multiplicative.

Rose:

Additive



Multiplicative



Inferences:

- There is decreasing trend and seasonality has well.
- Residuals is around 0 for additive and 1 for multiplicative.

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

- Data is splitted into train and test data for building models. Train data has data from beginning to 1990(132 rows and 1 column) and test data has data from 1991 to 1995(55 rows and 1 column)
- In time series machine learning analysis, **our observations are not independent**, and thus we cannot split the data randomly as we do in non-time-series analysis. The characteristics of time series data, such as autoregressive nature, trend, seasonality, or cyclicality, would not allow a random split to be valid

Sparkling:

Head and tail of Train data:

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

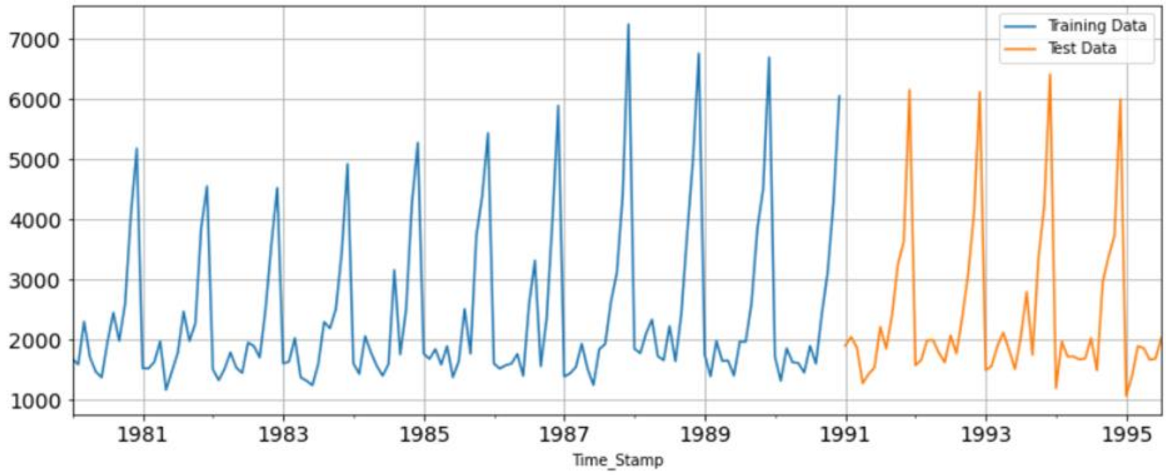
Time_Stamp	Sparkling
1990-08-31	1605
1990-09-30	2424
1990-10-31	3116
1990-11-30	4286
1990-12-31	6047

Head and tail of Test data:

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

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

Plotting graph for Train and Test Data for Sparkling:



Rose:

Head and tail of Train data:

Time_Stamp	Rose
1980-01-31	112.0
1980-02-29	118.0
1980-03-31	129.0
1980-04-30	99.0
1980-05-31	116.0

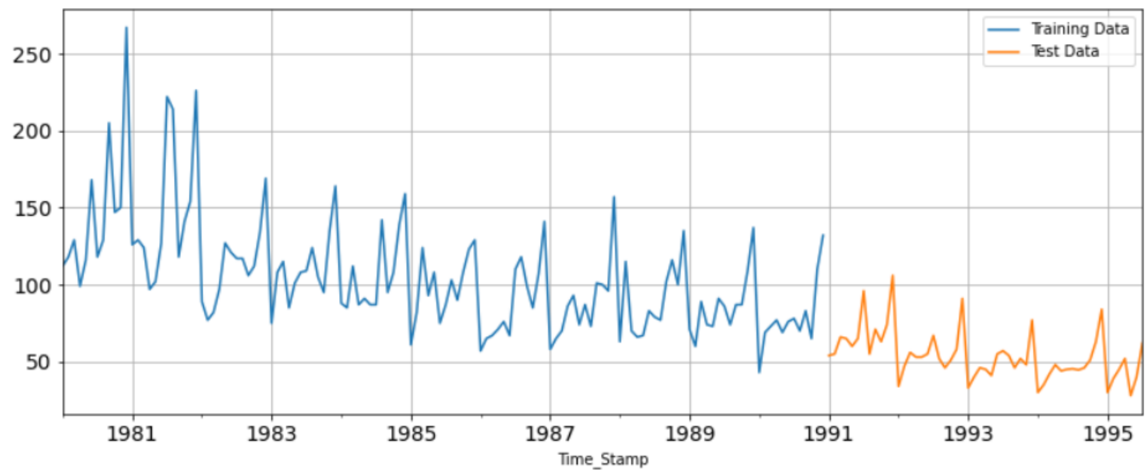
Time_Stamp	Rose
1990-08-31	70.0
1990-09-30	83.0
1990-10-31	65.0
1990-11-30	110.0
1990-12-31	132.0

Head and tail of Test data:

Time_Stamp	Rose
1991-01-31	54.0
1991-02-28	55.0
1991-03-31	66.0
1991-04-30	65.0
1991-05-31	60.0

Time_Stamp	Rose
1995-03-31	45.0
1995-04-30	52.0
1995-05-31	28.0
1995-06-30	40.0
1995-07-31	62.0

Plotting graph for Train and Test Data for Rose:



4. Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data

- Other models such as regression, naive forecast models, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE.

Building different models and comparing the accuracy metrics

Model 1: Linear Regression:

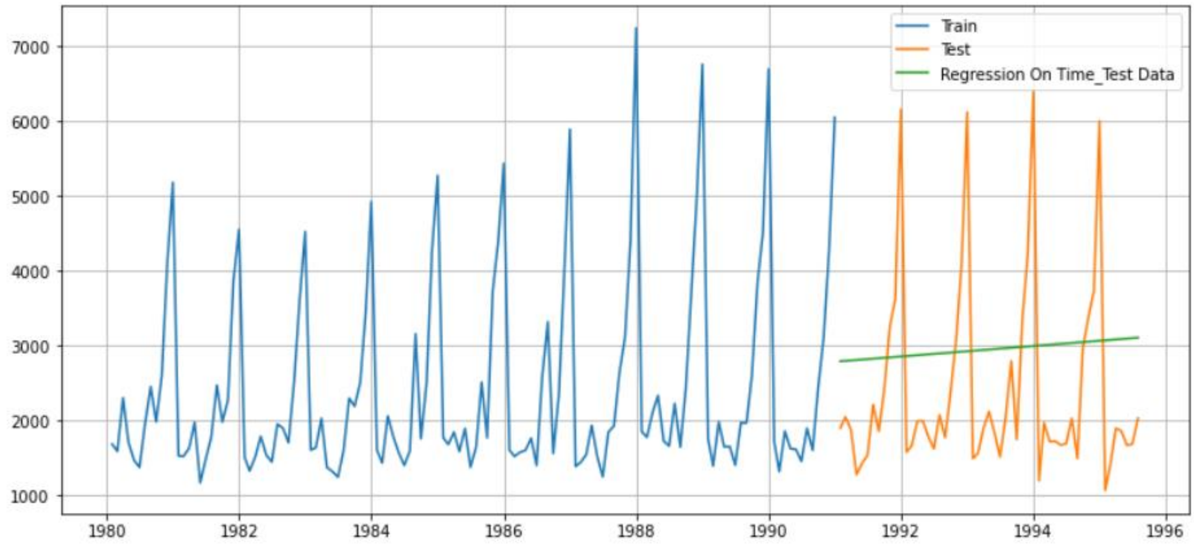
The linear regression algorithm learns how to make a weighted sum from its input features. For two features, we would have:

target = weight_1 * feature_1 + weight_2 * feature_2 + bias

During training, the regression algorithm learns values for the parameters weight_1, weight_2, and bias that best fit the target. (This algorithm is often called ordinary least squares since it chooses values that minimize the squared error between the target and the predictions.) The weights are also called regression coefficients and the bias is also called the intercept because it tells you where the graph of this function crosses the y-axis.

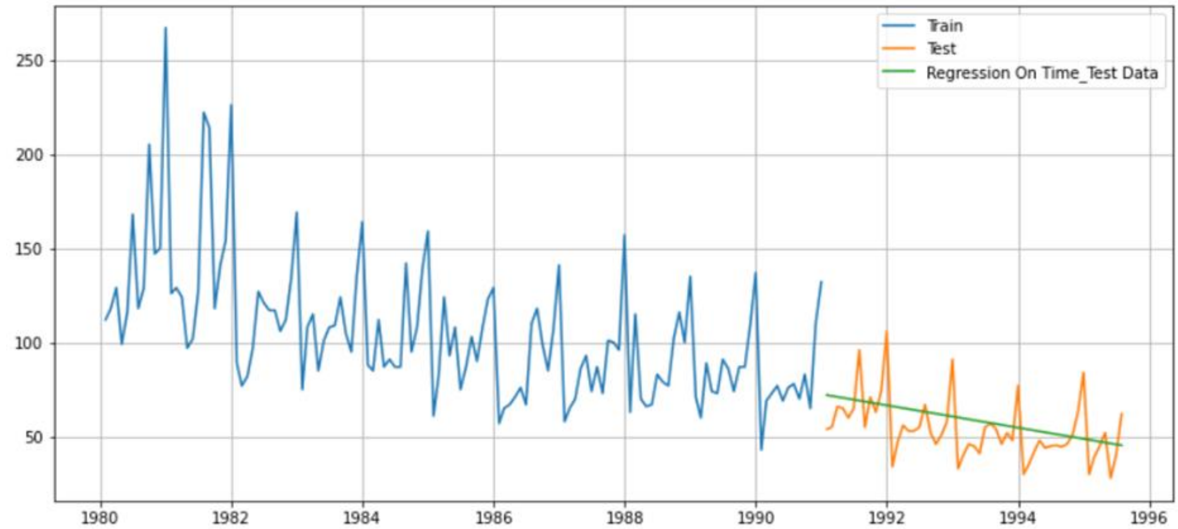
We see that we have successfully the generated the numerical time instance order for both the training and test set. Now we will add these values in the training and test set to build models.

Sparkling:



- Linear Regression model intersects the test data which is gradually increasing upwards.
- RMSE Value for linear regression model is 1389.135175

Rose:

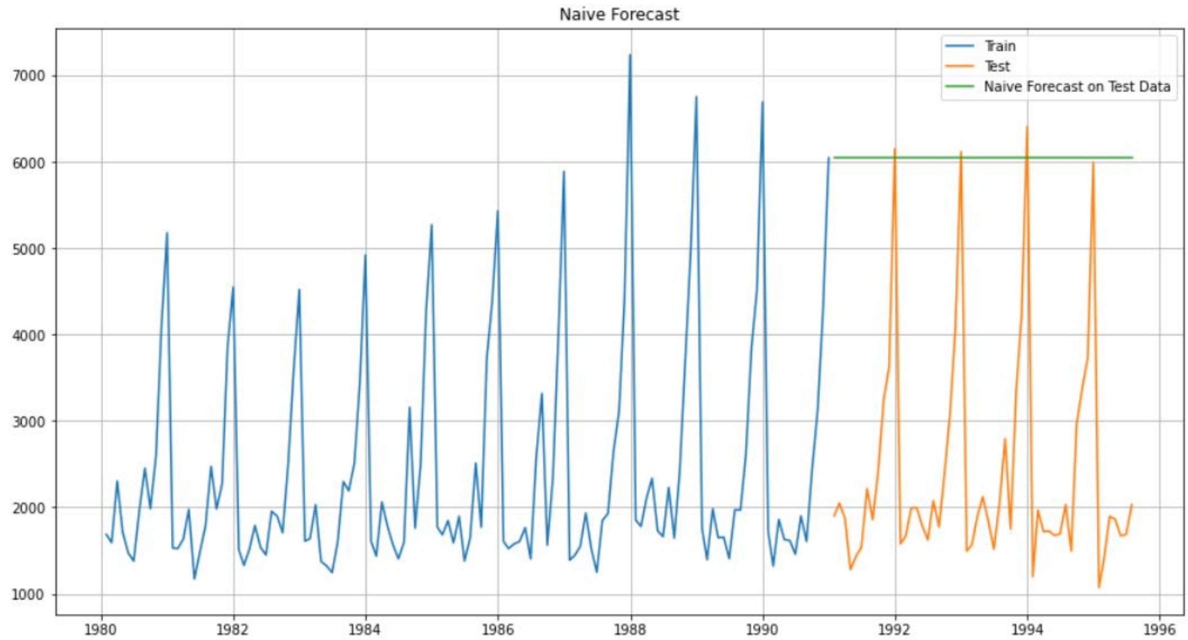


- Linear Regression model intersects the test data which is gradually decreasing downwards.
- RMSE Value for linear regression model is 15.27552

Model 2: Naive Approach

- For this particular naive model, we say that the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow and since the prediction of tomorrow is same as today,therefore the prediction for day after tomorrow is also today.

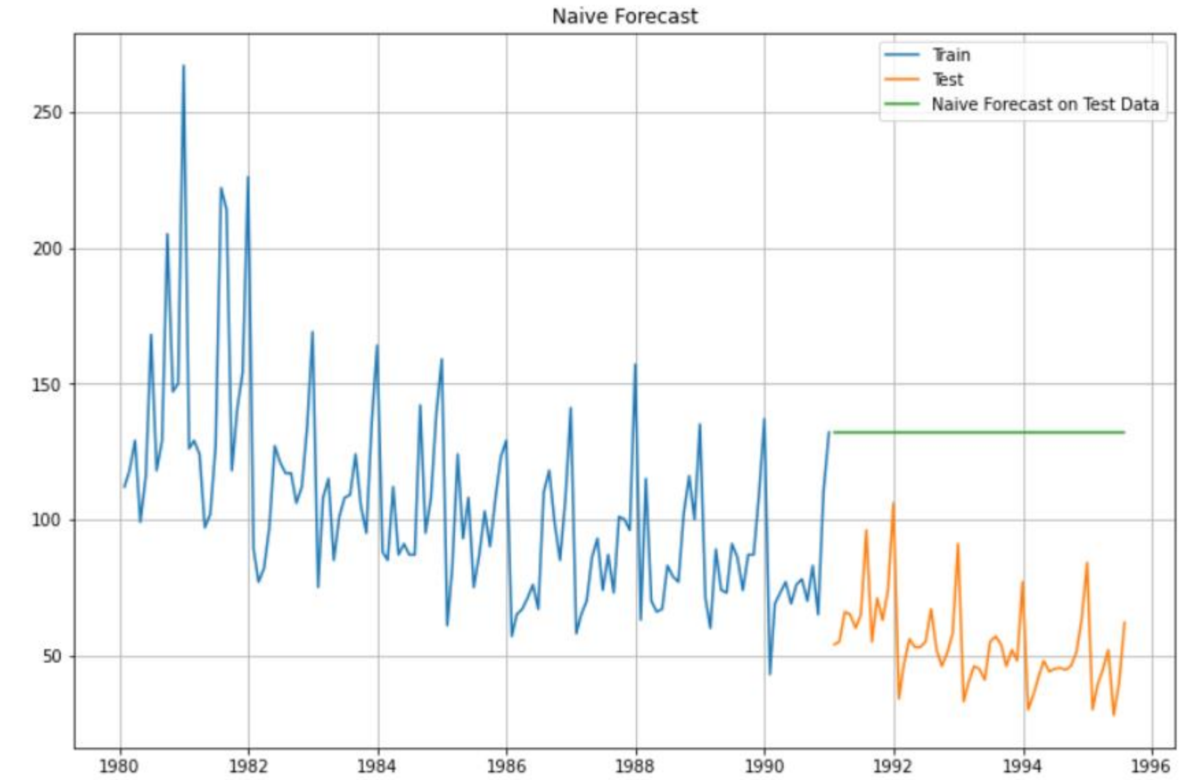
Sparking:



Inference:

- Naive Approach model intersects the test data which is flat line. This model has no impact in time series forecasting.
- RMSE Value for linear regression model is 3864.279352

Rose:



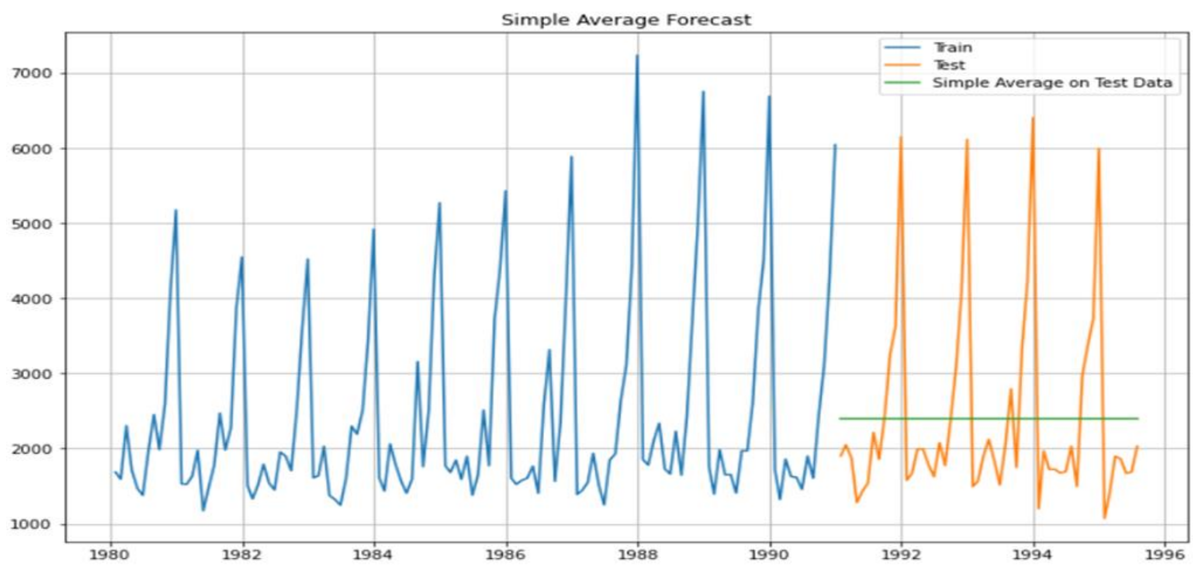
Inference:

- Naive Approach model not intersects the test data which is flat line. This model has no impact in time series forecasting.
- RMSE Value for linear regression model is 79.738146

Method 3: Simple Average

- Simple Average is the sum of the numbers divided by the total number of values in the set.

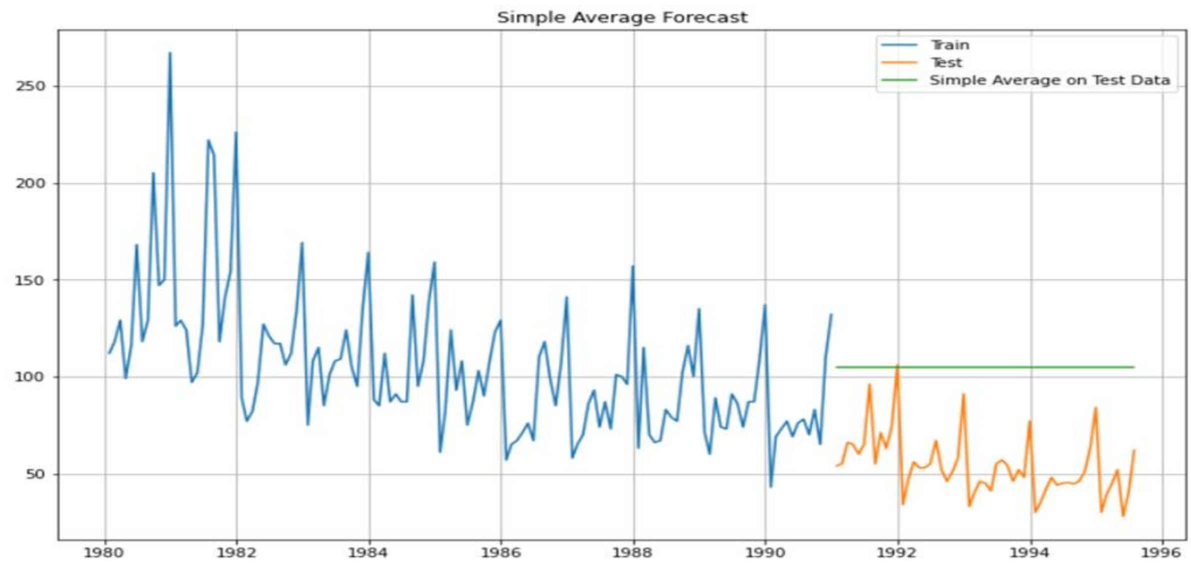
Sparkling:



Inference:

- Simple Average intersects the test data which is flat line. Which is better than linear tegression and Naive Model.
- RMSE Value for linear regression model is 1275.081804

Rose:



Inference:

- Simple Average not intersecting the test data which is flat line. This model will not impact for forecasting
- RMSE Value for linear regression model is 53.480456

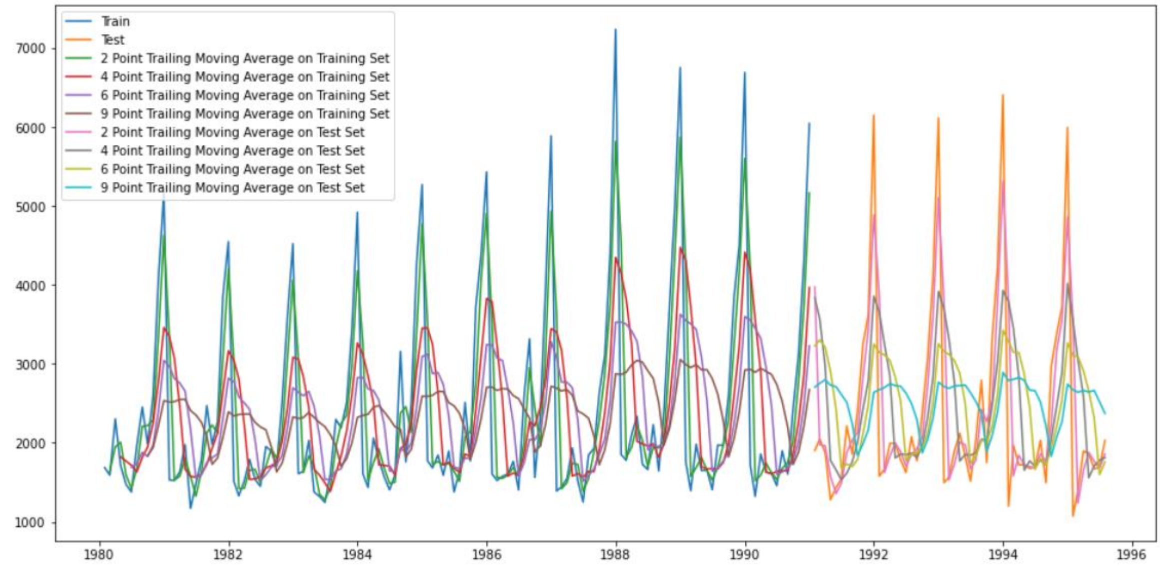
Method 4: Moving Average(MA):

- A simple moving average (SMA) is an arithmetic moving average calculated by adding recent prices and then dividing that figure by the number of time periods in the calculation average. ... Short-term averages respond quickly to changes in the price of the underlying security, while long-term averages are slower to react.

- A trailing average may also be referred to as a **moving average**. ... Calculate the average of the first three months' data if you are using a three-month trailing period. If your data begins in January, calculate the average of January, February and March. This figure becomes the three-month trailing average for March.

- Let us split the data into train and test and plot this Time Series. The window of the moving average is need to be carefully selected as too big a window will result in not having any test set as the whole series might get averaged over .

Sparkling:

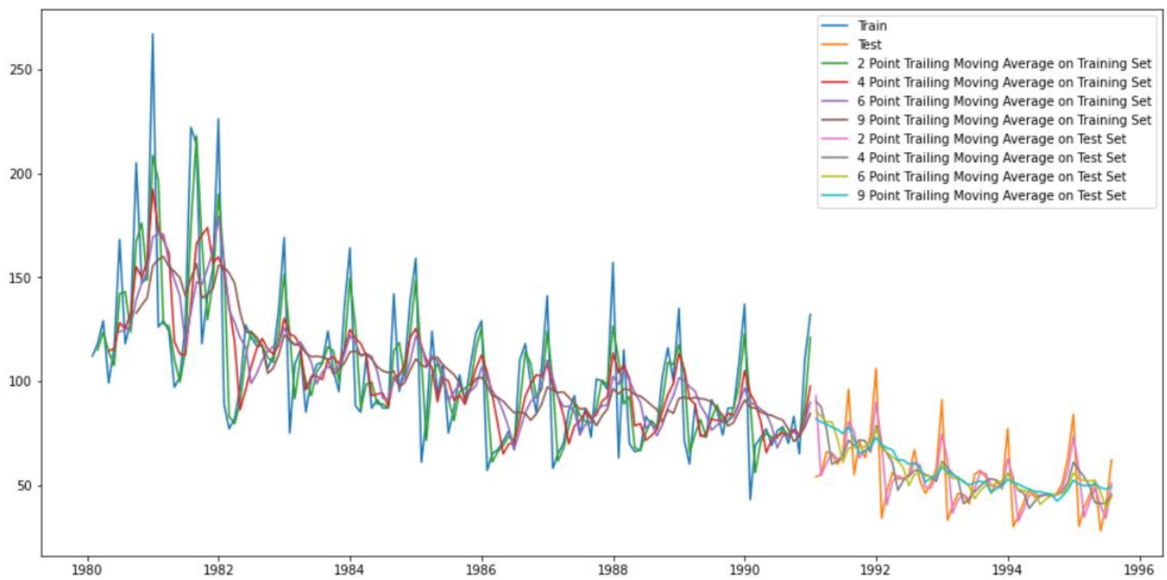



```
For 2 point Moving Average Model forecast on the Training Data,  RMSE is 813.401
For 4 point Moving Average Model forecast on the Training Data,  RMSE is 1156.590
For 6 point Moving Average Model forecast on the Training Data,  RMSE is 1283.927
For 9 point Moving Average Model forecast on the Training Data,  RMSE is 1346.278
```

Inference:

- Two trailing moving average is performing well with train data with low RMSE value compared to other trailing values.

Rose:



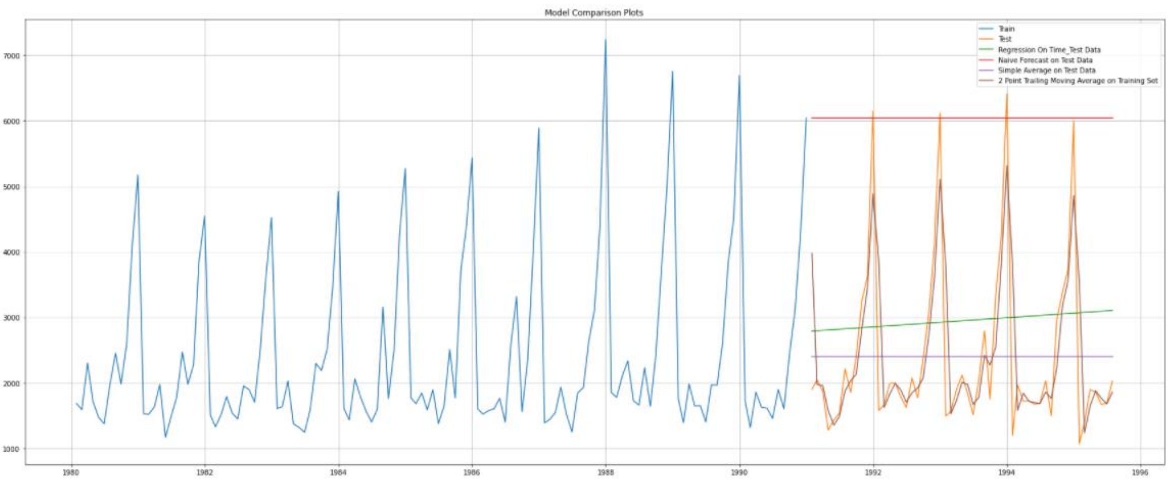
```
For 2 point Moving Average Model forecast on the Training Data,  RMSE is 11.530
For 4 point Moving Average Model forecast on the Training Data,  RMSE is 14.456
For 6 point Moving Average Model forecast on the Training Data,  RMSE is 14.571
For 9 point Moving Average Model forecast on the Training Data,  RMSE is 14.731
```

Inference:

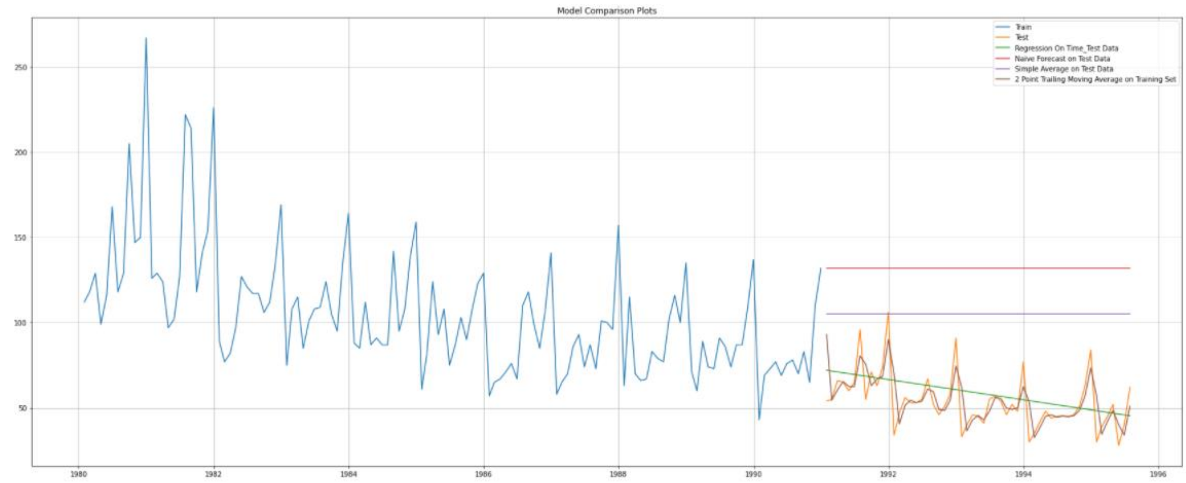
- Two trailing moving average is performing well with train data as well as test data with low RMSE value compared to other trailing values.

Before we go on to build the various Exponential Smoothing models, let us plot all the models and compare the Time Series plots

Sparkling:



Rose:



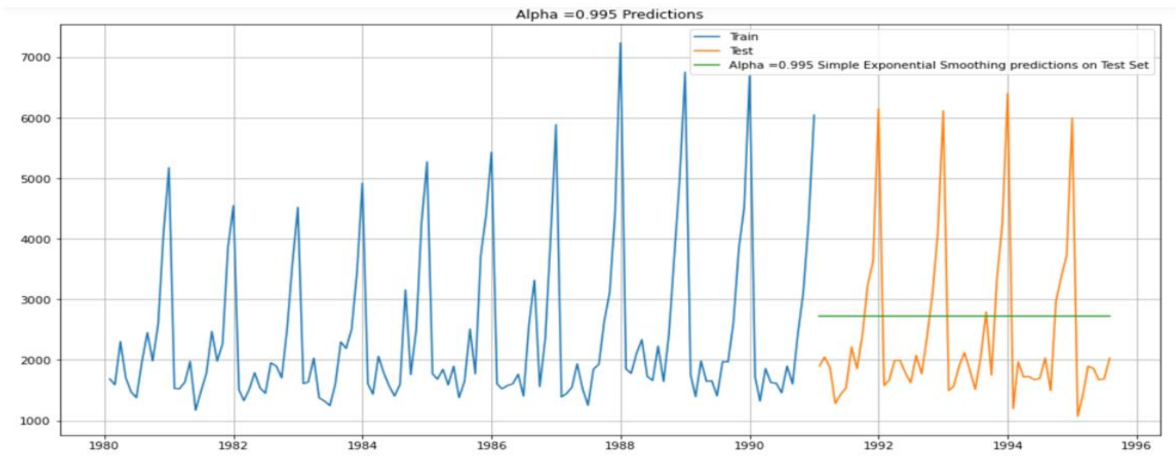
Inferences:

- Among 4 models two trailing moving average seems to be working better for both train and test data with low RMSE Value.

Method 5: Simple Exponential Smoothing

- The simplest of the exponentially smoothing methods is naturally called simple exponential smoothing (SES)[13]. This method is suitable for forecasting data with no clear trend or seasonal pattern.

Sparkling:

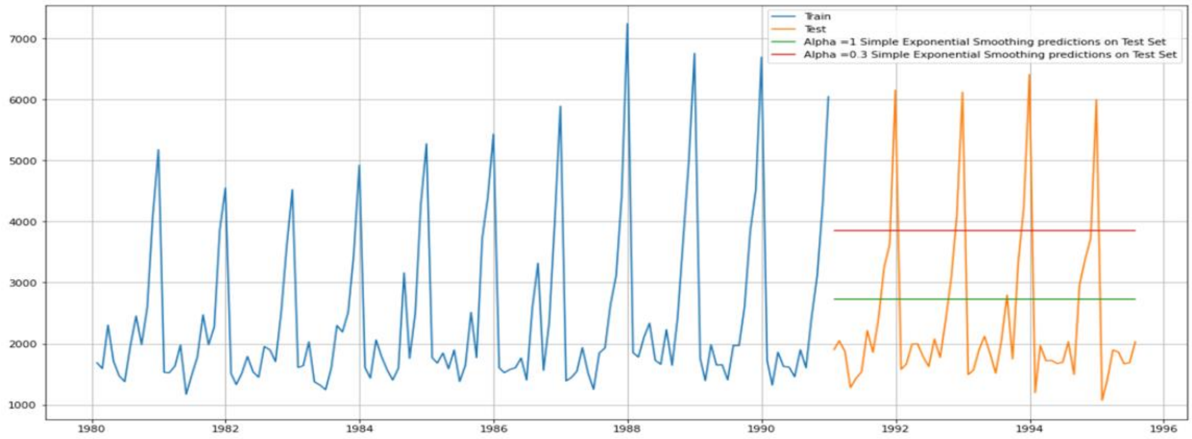


- This model is build for alpha=0.995 which has RSME value 1316.035487

Setting different alpha values:

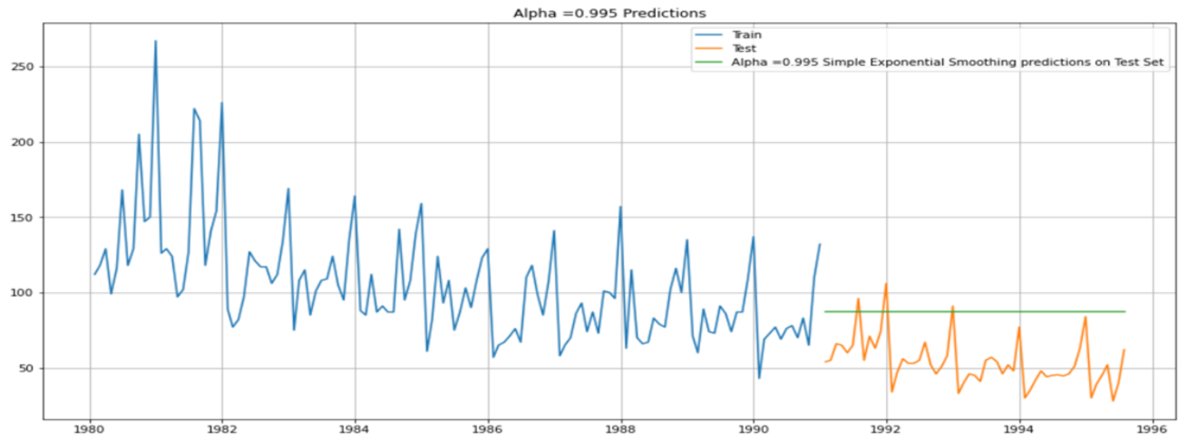
- Remember, the higher the alpha value more weightage is given to the more recent observation. That means, what happened recently will happen again.

- Alpha=0.3 which has low train and test value to build a model. RMSE value for Simple Exponential Smoothing is 1935.507132 which is not performing well compared to previous model.

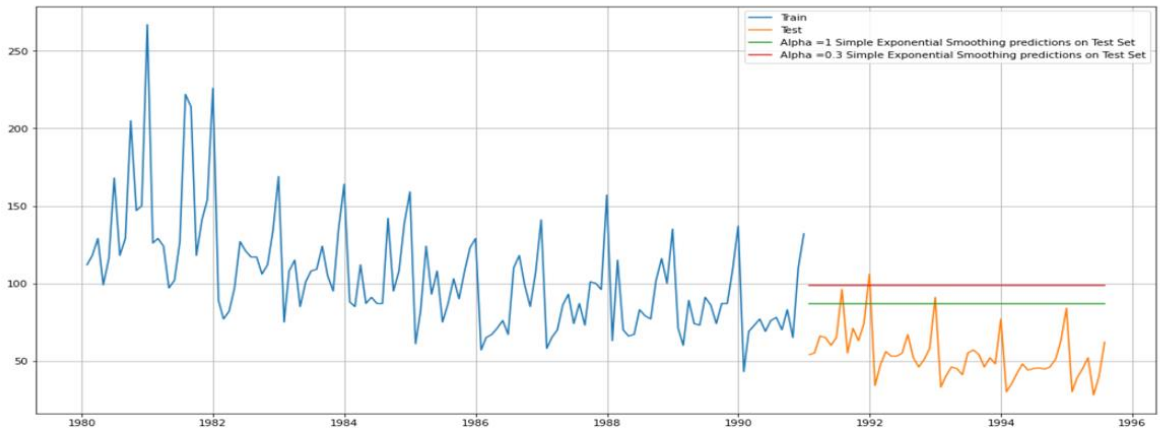


Rose:

- This model is build for alpha=0.995 which has RSME value 36.816502.



Using Different Alpha values:



- Remember, the higher the alpha value more weightage is given to the more recent observation. That means, what happened recently will happen again.
- Alpha=0.3 which has low train and test value to build a model. RMSE value for Simple Exponential Smoothing is 47.524854 which is not performing well compared to previous model.

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

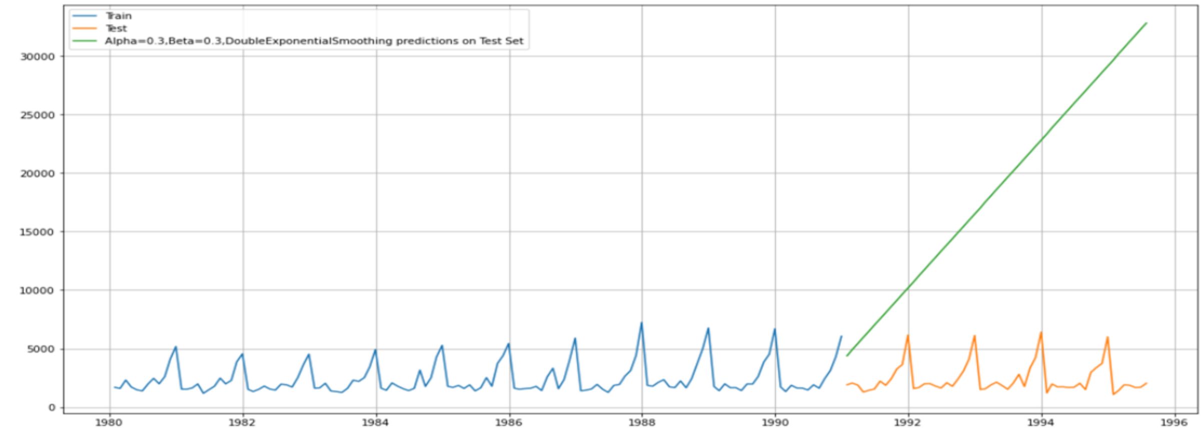
- Double Exponential Smoothing is an extension to Exponential Smoothing that explicitly adds support for trends in the univariate time series.

- In addition to the alpha parameter for controlling smoothing factor for the level, an additional smoothing factor is added to control the decay of the influence of the change in trend called beta (b).

- The method supports trends that change in different ways: an additive and a multiplicative, depending on whether the trend is linear or exponential respectively.

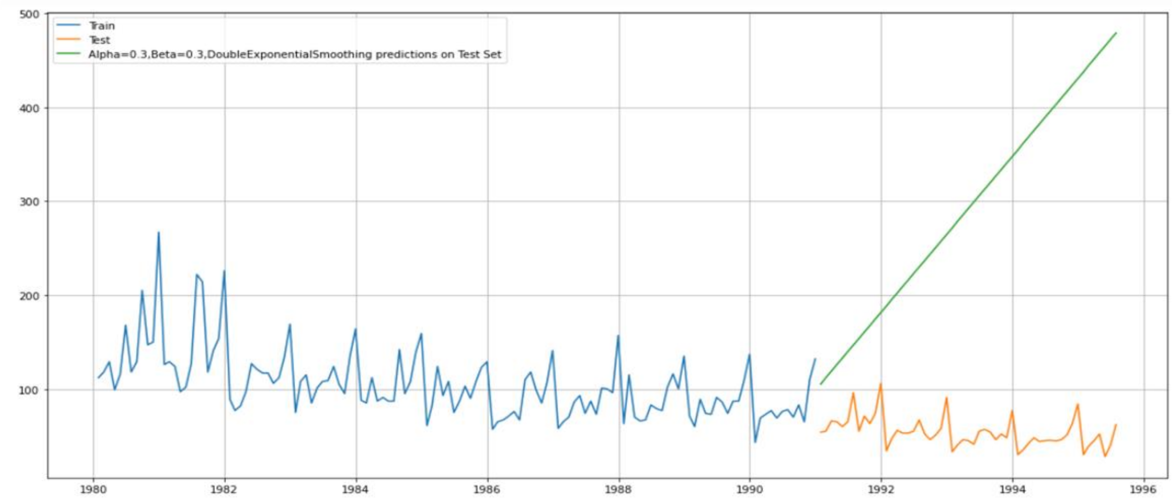
- Double Exponential Smoothing with an additive trend is classically referred to as Holt's linear trend model.

Sparkling:



- This model is build for both alpha=0.3 and Beta= 0.3 which has RMSE Value 18259.110704 which has high compared to all models.

Rose:



- This model is build for both alpha=0.3 and Beta= 0.3 which has RMSE Value 47.524854 which has high compared to all models.

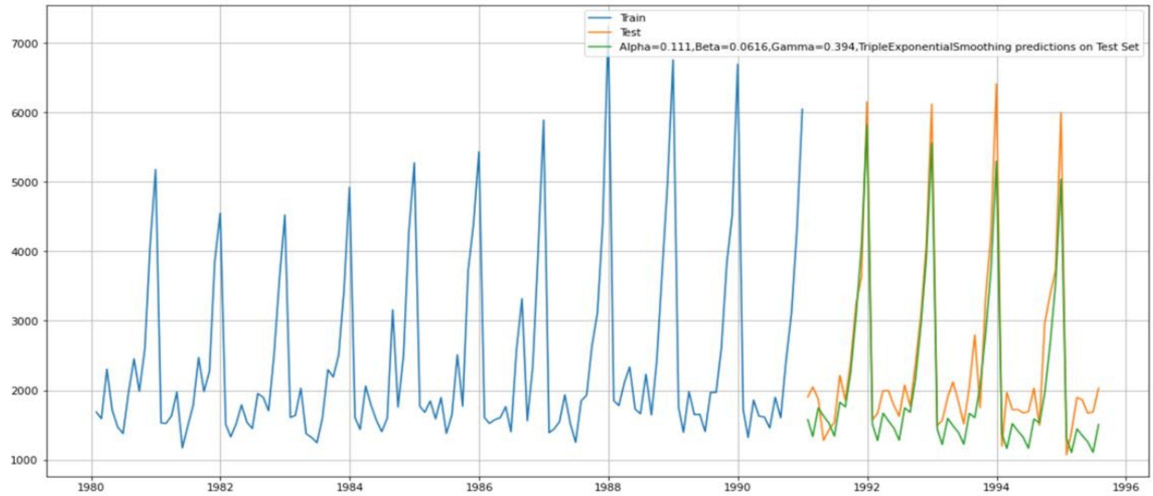
Method 7: Triple Exponential Smoothing (Holt - Winter's Model)

- Triple exponential smoothing is **used to handle the time series data containing a seasonal component**. This method is based on three smoothing equations: stationary component, trend, and seasonal. Both seasonal and trend can be additive or multiplicative. ... Seasonal change smoothing factor.
- TES explicitly adds support to the univariate time series for seasonality; it is also referred to as Holt-Winters Exponential Smoothing on the name of two contributors Charles Holt and Peter Winters.
- The Holt-Winters exponential smoothing model permits the level, trend and seasonality patterns to change over time as it is an adaptive method.

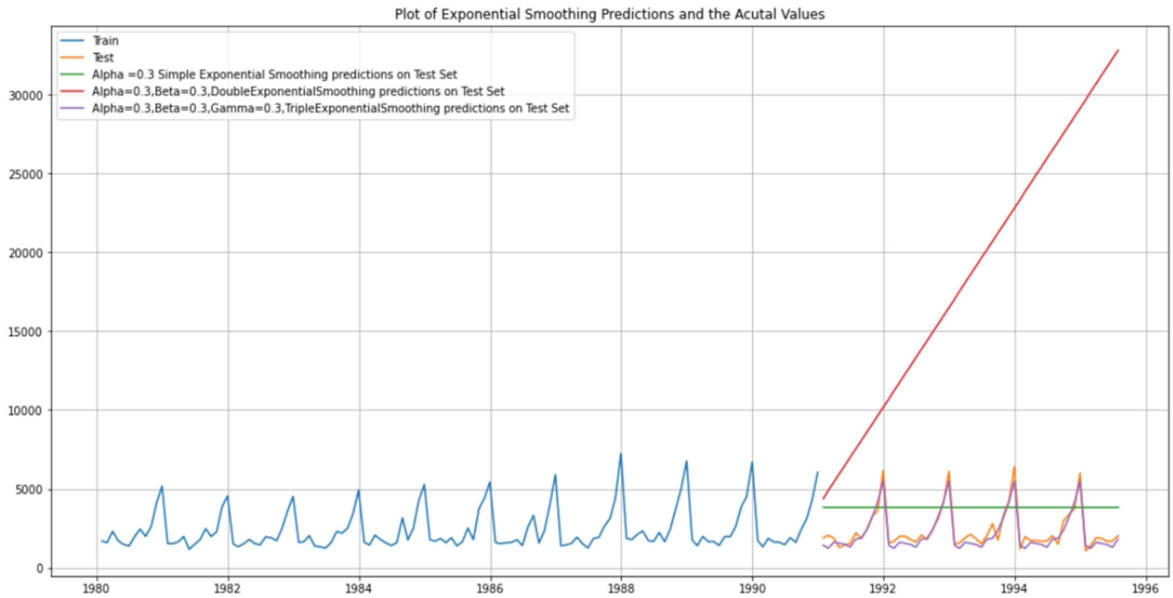
- Beside the two smoothing factors, alpha and beta, an additional new factor is introduced, called gamma in order to control/determine the impact on the seasonal element.

- In correspondence with the trend, seasonality can be modeled in the particular of additive or multiplicative process for the linear and exponential variation in the seasonality.

Sparkling:

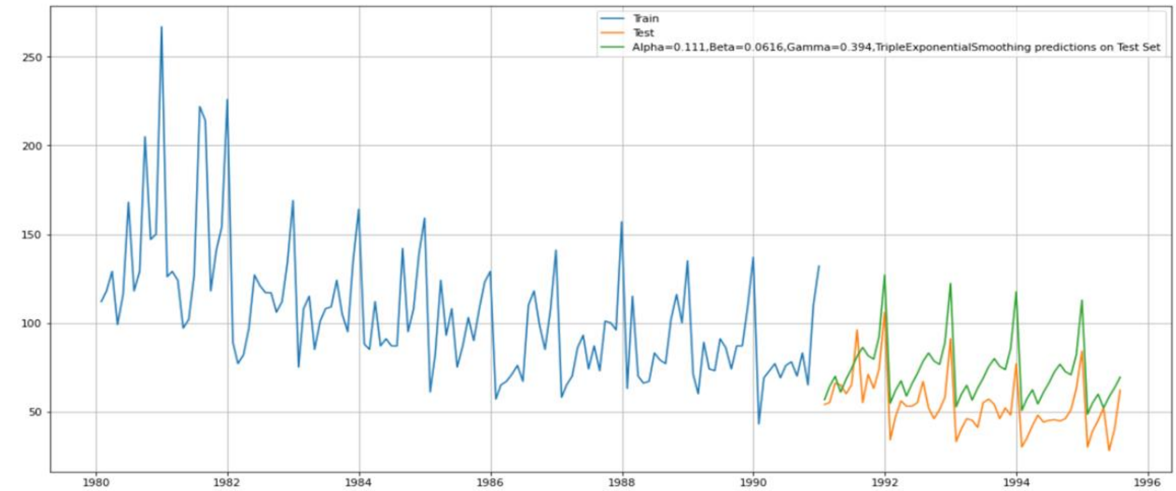


- This model is build for both alpha=0.111, Beta= 0.0615 and Gamma=0.394 which has RMSE Value 469.767970 which has low compared to all models.
- For this data, we had both trend and seasonality so by definition Triple Exponential Smoothing is supposed to work better than the Simple Exponential Smoothing as well as the Double Exponential Smoothing. However, since this was a model building exercise we had gone on to build different models on the data and have compared these model with the best RMSE value on the test data.

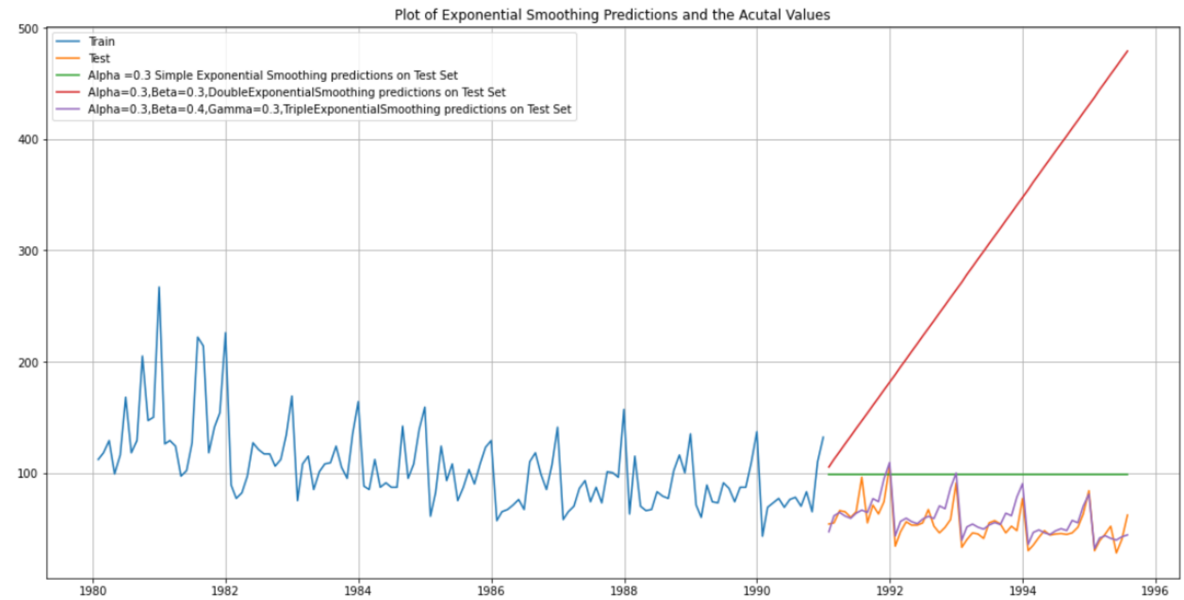


- This model is build for both alpha=0.3, Beta= 0.3 and Gamma=0.3 which has RMSE Value 392.786198 which has low compared to all models

Rose:



- This model is build for both alpha=0.111, Beta= 0.0615 and Gamma=0.394 which has RMSE Value 21.046473 which has low compared to all models
- For this data, we had both trend and seasonality so by definition Triple Exponential Smoothing is supposed to work better than the Simple Exponential Smoothing as well as the Double Exponential Smoothing. However, since this was a model building exercise we had gone on to build different models on the data and have compared these model with the best RMSE value on the test data.



- This model is build for both alpha=0.3, Beta= 0.3 and Gamma=0.3 which has RMSE Value 10.950048 which has low compared to all models. Which seems working better for Time series forescasting

Inferences:

- In this particular we have built several models and went through a model building exercise. This particular exercise has given us an idea as to which particular model gives us the least error on our test set for this data.
- But in Time Series Forecasting, we need to be very vigil about the fact that after we have done this exercise we need to build the model on the whole data.
- Remember, the training data that we have used to build the model stops much before the data ends. In order to forecast using any of the models built, we need to build the models again (this time on the complete data) with the same parameters.

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

Check for stationarity of the whole Time Series data.

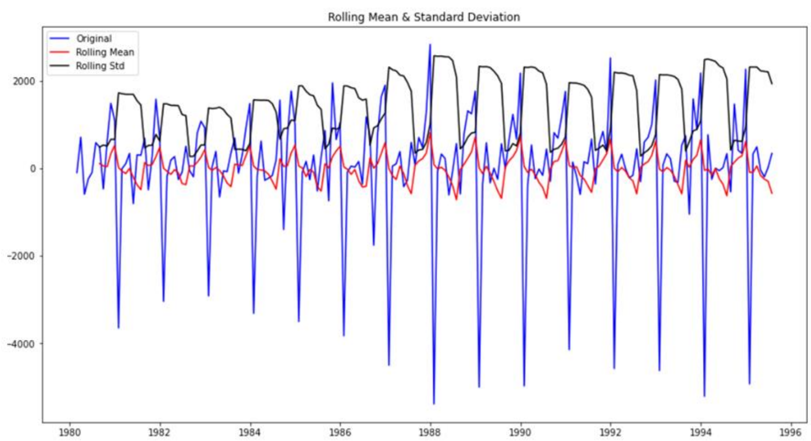
- A **stationary** time series is one whose properties do not depend on the time at which the series is observed. Thus, time series with trends, or with seasonality, are not stationary.The trend and seasonality will affect the value of the time series at different times.
- **Non-stationary** or have means, variances, and covariances that change over time.

The hypothesis in a simple form for the ADF test is:

Null Hypothesis : The Time Series has a unit root and is thus non-stationary.

Alternative Hypothesis : The Time Series does not have a unit root and is thus stationary.

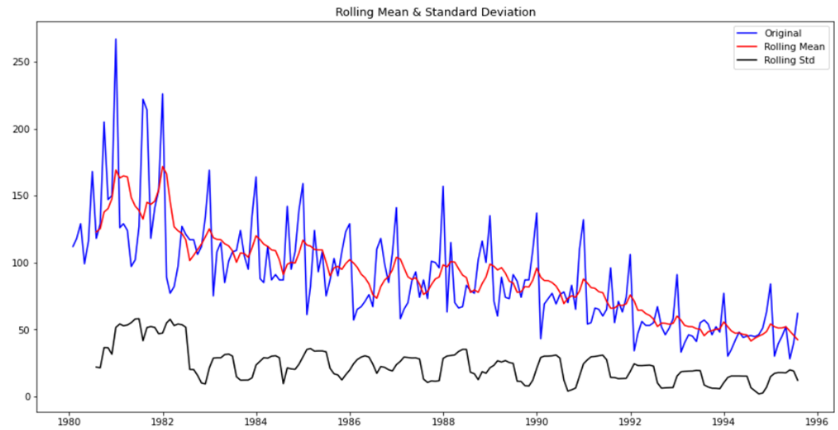
Sparkling:




```
Results of Dickey-Fuller Test:
Test Statistic                 -45.050301
p-value                          0.000000
#Lags Used                      10.000000
Number of Observations Used    175.000000
Critical Value (1%)             -3.468280
Critical Value (5%)             -2.878202
Critical Value (10%)            -2.575653
dtype: float64
```

➢ From the Plot we observed Alpha value is less than 0.05 so we reject the null hypothesis . So the time series is indeed stationary.

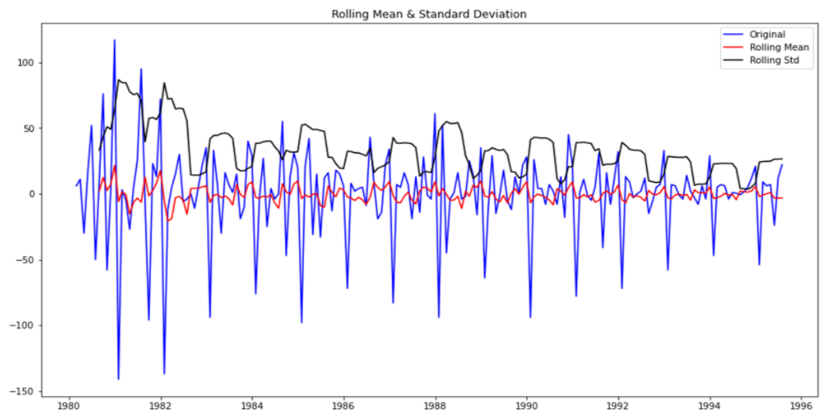
Rose:




```
Results of Dickey-Fuller Test:
Test Statistic                 -1.873656
p-value                         0.344554
#Lags Used                     13.000000
Number of Observations Used   173.000000
Critical Value (1%)            -3.468726
Critical Value (5%)            -2.878396
Critical Value (10%)           -2.575756
dtype: float64
```

➢ From the Plot we observed Alpha value is greater than 0.05 so we reject failed to reject null hypothesis. So time series is indeed stationary.

➢ If the series is non-stationary, stationary the Time Series by taking a difference of the Time Series. Then we can use this particular differenced series to train the ARIMA models. We do not need to worry about stationary for the Test Data because we are not building any models on the Test Data, we are evaluating our models over there. You can look at other kinds of transformations as part of making the time series stationary like taking logarithms.




```
Results of Dickey-Fuller Test:
Test Statistic                -8.044177e+00
p-value                        1.813173e-12
#Lags Used                     1.200000e+01
Number of Observations Used    1.730000e+02
Critical Value (1%)           -3.468726e+00
Critical Value (5%)           -2.878396e+00
Critical Value (10%)          -2.575756e+00
dtype: float64
```

➢ From the Plot we observed Alpha value is less than 0.05 so we reject the null hypothesis . So the time series is indeed stationary.

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.

We need build both automated ARIMA/SARIMA Model

<u>Arima Model:</u>

- An autoregressive integrated moving average, or ARIMA, is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends.
- A statistical model is autoregressive if it predicts future values based on past values.
- ARIMA makes use of lagged moving averages to smooth time series data
- Each component in ARIMA functions as a parameter with a standard notation.

For ARIMA models, a standard notation would be ARIMA with p, d, and q, where integer values substitute for the parameters to indicate the type of ARIMA model used. The parameters can be defined as:

p: the number of lag observations in the model; also known as the lag order.

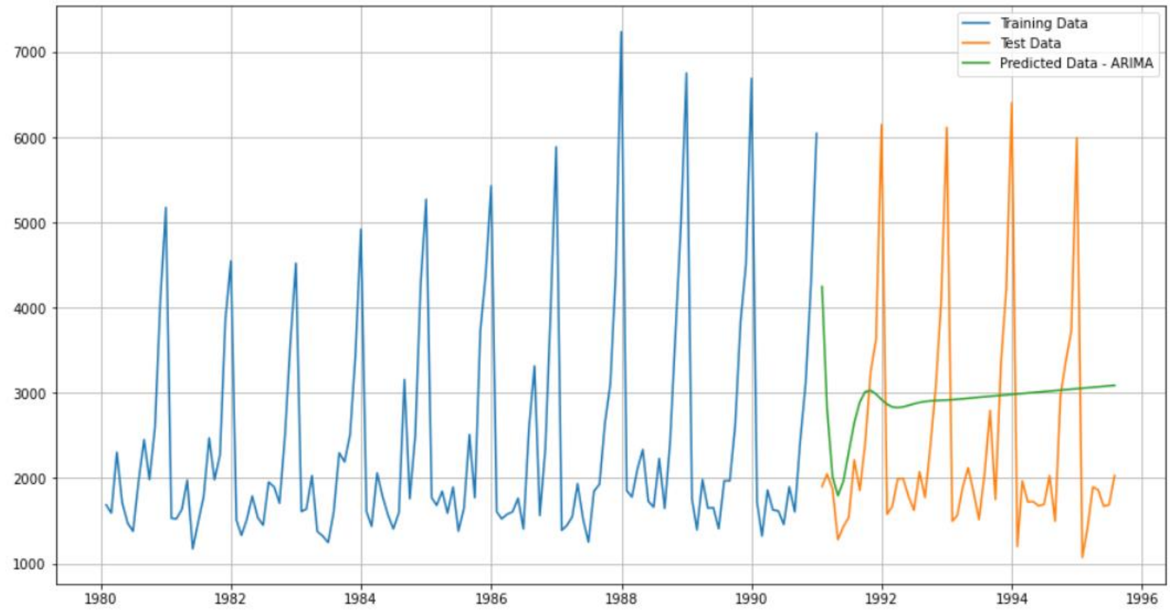
d: the number of times that the raw observations are differenced; also known as the degree of differencing.

q: the size of the moving average window; also known as the order of the moving average

Sparkling:

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:            D.Sparkling   No. Observations:                  131
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1099.313
Method:                       css-mle   S.D. of innovations           1013.755
Date:                Sat, 25 Dec 2021   AIC                           2210.626
Time:                        22:19:17   BIC                           2227.877
Sample:                    02-29-1980   HQIC                          2217.636
                         - 12-31-1990
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 5.5845      0.519     10.753      0.000       4.567       6.602
ar.L1.D.Sparkling     1.2698      0.075     17.040      0.000       1.124       1.416
ar.L2.D.Sparkling    -0.5601      0.074     -7.617      0.000      -0.704      -0.416
ma.L1.D.Sparkling    -1.9957      0.043    -46.821      0.000      -2.079      -1.912
ma.L2.D.Sparkling     0.9957      0.043     23.291      0.000       0.912       1.079
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.1335           -0.7074j            1.3361           -0.0888
AR.2            1.1335           +0.7074j            1.3361            0.0888
MA.1            1.0000           +0.0000j            1.0000            0.0000
MA.2            1.0043           +0.0000j            1.0043            0.0000
```

- From the automated Arima model we observed (2,1,2) has low AIC value 2210.

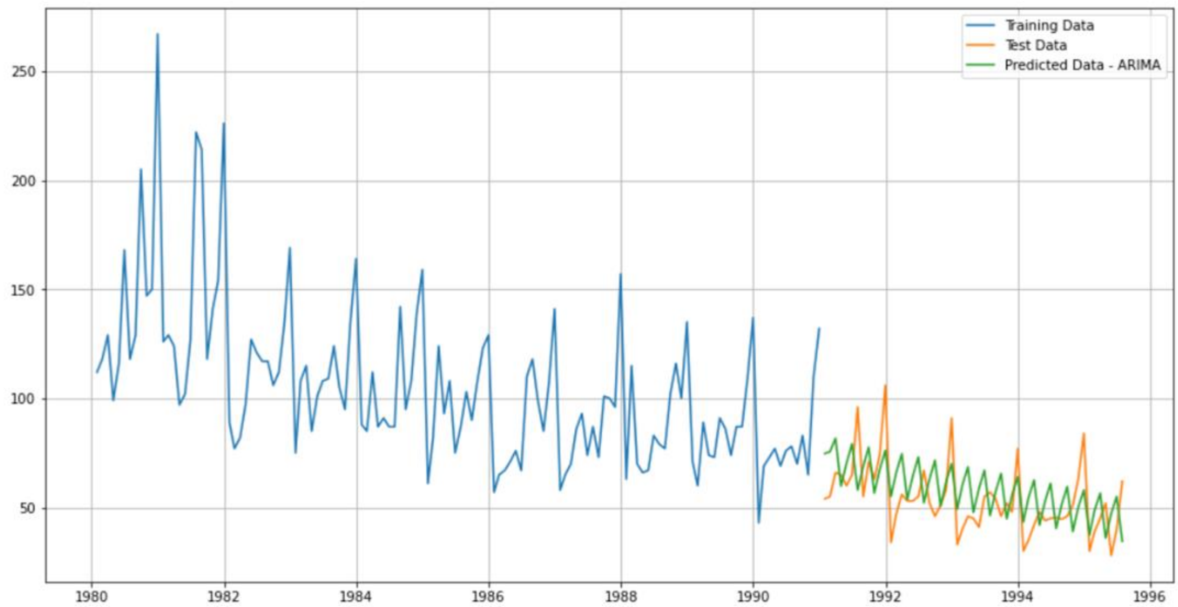


- So we calculated Arima model with (2,1,2) which has RMSE Value 1374.037

Rose:

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:                 D.Rose   No. Observations:                  131
Model:                 ARIMA(3, 1, 3)   Log Likelihood                -628.597
Method:                       css-mle   S.D. of innovations             28.356
Date:                Sat, 25 Dec 2021   AIC                           1273.194
Time:                        22:19:03   BIC                           1296.196
Sample:                    02-29-1980   HQIC                          1282.541
                         - 12-31-1990
===============================================================================
                   coef    std err          z      P>|z|      [0.025      0.97
-------------------------------------------------------------------------------
const           -0.4906      0.088     -5.548      0.000      -0.664      -0.3
ar.L1.D.Rose    -0.7243      0.086     -8.411      0.000      -0.893      -0.5
ar.L2.D.Rose    -0.7218      0.087     -8.342      0.000      -0.891      -0.5
ar.L3.D.Rose     0.2763      0.085      3.234      0.001       0.109       0.4
ma.L1.D.Rose    -0.0151      0.045     -0.339      0.735      -0.102       0.0
ma.L2.D.Rose     0.0151      0.044      0.340      0.734      -0.072       0.1
ma.L3.D.Rose    -1.0000      0.046    -21.901      0.000      -1.089      -0.9
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1           -0.5011           -0.8661j            1.0006           -0.3335
AR.2           -0.5011           +0.8661j            1.0006            0.3335
AR.3            3.6142           -0.0000j            3.6142           -0.0000
MA.1            1.0000           -0.0000j            1.0000           -0.0000
MA.2           -0.4925           -0.8703j            1.0000           -0.3320
MA.3           -0.4925           +0.8703j            1.0000            0.3320
-----------------------------------------------------------------------------
```

- From the automated sarima model we observed (2,1,2) has low AIC value 1273.



- So we calculated Arima model with (2,1,2) which has RMSE Value 15.994252. Model performs good but lower than triple exponential smoothing.

Sarima Model:

- Seasonal Autoregressive Integrated Moving Average, SARIMA or Seasonal ARIMA, is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component.
- It adds three new hyperparameters to specify the autoregression (AR), differencing (I) and moving average (MA) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality.
- Configuring a SARIMA requires selecting hyperparameters for both the trend and seasonal elements of the series.

Trend Elements

There are three trend elements that require configuration.

They are the same as the ARIMA model; specifically:

p: Trend autoregression order.

d: Trend difference order.

q: Trend moving average order.

Seasonal Elements

There are four seasonal elements that are not part of ARIMA that must be configured; they are:

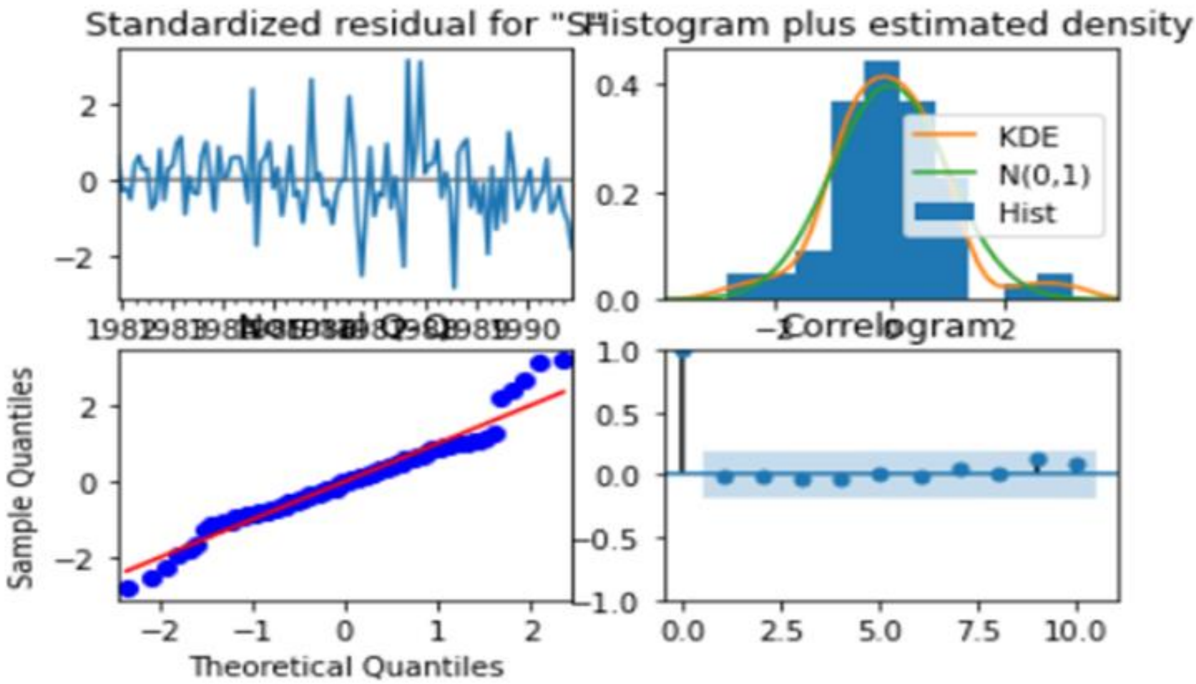
P: Seasonal autoregressive order.

D: Seasonal difference order.

Q: Seasonal moving average order.

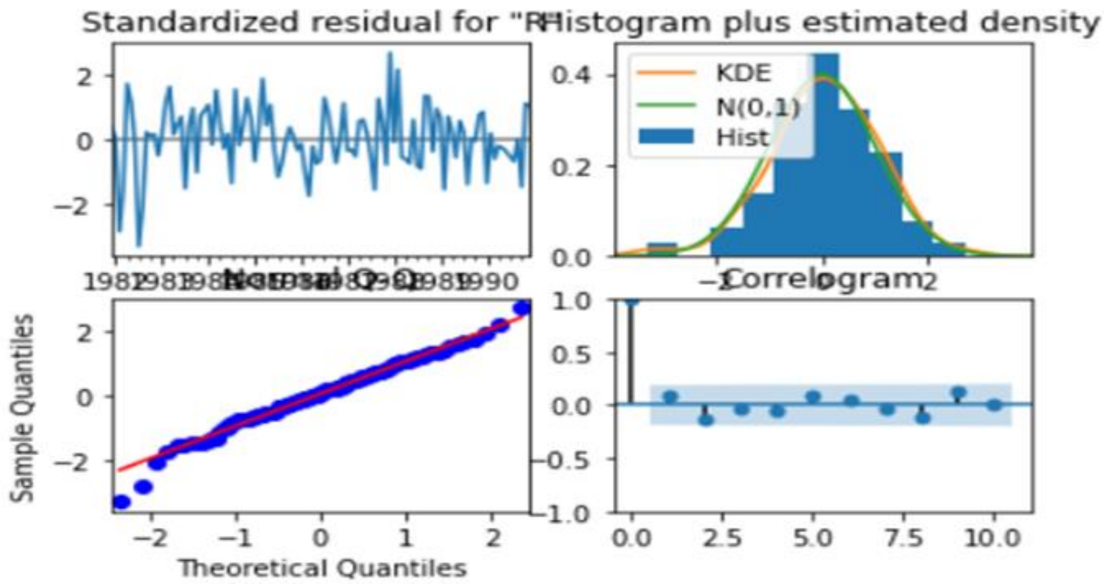
m: The number of time steps for a single seasonal period.

Sparkling:



- So we calculated Arima model with (2,1,3) (2,0,3,6) which has RMSE Value 825.67. Model performs good but lower than triple exponential smoothing.

Rose:

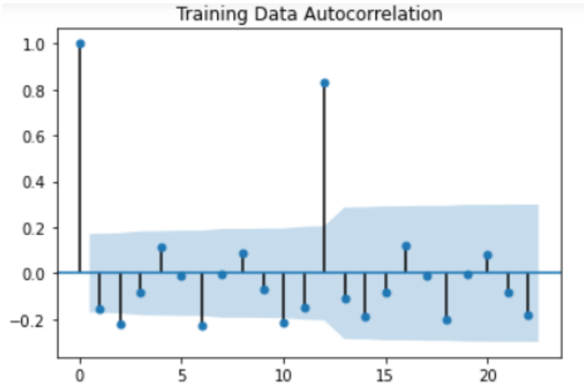


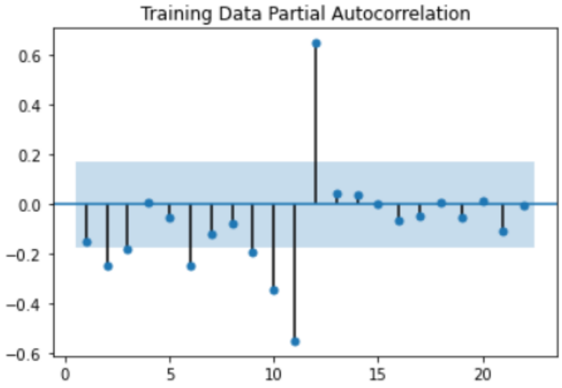
- So we calculated Arima model with (2,1,3) (2,0,3,6) which has RMSE Value 27.14. Model performs good but lower than triple exponential smoothing.

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

Sparkling:

Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data





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.

Manual Arima for Sparkling:

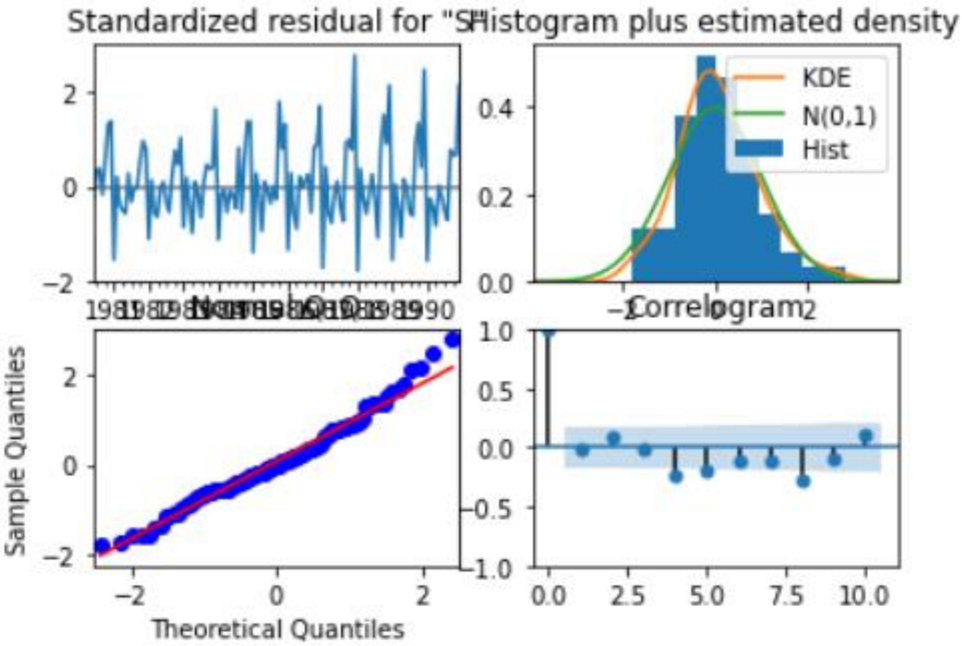
```
                             ARIMA Model Results
==============================================================================
Dep. Variable:            D.Sparkling   No. Observations:                  131
Model:                 ARIMA(3, 1, 3)   Log Likelihood               -1104.831
Method:                       css-mle   S.D. of innovations           1074.561
Date:                Sat, 25 Dec 2021   AIC                           2225.662
Time:                        23:34:47   BIC                           2248.663
Sample:                    02-29-1980   HQIC                          2235.008
                         - 12-31-1990
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 6.3212      4.207      1.503      0.133      -1.924      14.566
ar.L1.D.Sparkling    -0.7550      0.093     -8.099      0.000      -0.938      -0.572
ar.L2.D.Sparkling    -0.3887      0.110     -3.533      0.000      -0.604      -0.173
ar.L3.D.Sparkling     0.2143      0.089      2.395      0.017       0.039       0.390
ma.L1.D.Sparkling     0.3401      0.006     55.867      0.000       0.328       0.352
ma.L2.D.Sparkling    -0.3401      0.037     -9.308      0.000      -0.412      -0.268
ma.L3.D.Sparkling    -1.0000        nan        nan        nan         nan         nan
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1           -0.7463           -0.9244j            1.1880           -0.3581
AR.2           -0.7463           +0.9244j            1.1880            0.3581
AR.3            3.3064           -0.0000j            3.3064           -0.0000
MA.1            1.0000           -0.0000j            1.0000           -0.0000
MA.2           -0.6700           -0.7423j            1.0000           -0.3669
MA.3           -0.6700           +0.7423j            1.0000            0.3669
```

- From the Manual Arima model we observed (3,1,3) has low AIC value 2225.
- So we calculated Arima model with (3,1,3) which has RMSE Value 1425.09. Model performs good but lower than triple exponential smoothing.

Manual SARIMA for Sparkling:

```
                                SARIMAX Results
==============================================================================
Dep. Variable:                Sparkling   No. Observations:                  132
Model:                 SARIMAX(3, 1, 3)   Log Likelihood               -1070.887
Date:                  Sat, 25 Dec 2021   AIC                           2155.775
Time:                          23:35:08   BIC                           2175.684
Sample:                      01-31-1980   HQIC                          2163.864
                           - 12-31-1990
Covariance Type:                    opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.3554      0.140      2.537      0.011       0.081       0.630
ar.L2         -0.9352      0.062    -14.998      0.000      -1.057      -0.813
ar.L3          0.3664      0.123      2.972      0.003       0.125       0.608
ma.L1         -0.7979      0.100     -8.002      0.000      -0.993      -0.602
ma.L2          0.7965      0.146      5.442      0.000       0.510       1.083
ma.L3         -0.9642      0.144     -6.710      0.000      -1.246      -0.683
sigma2      1.519e+06   3.13e+05      4.860      0.000    9.07e+05    2.13e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.05   Jarque-Bera (JB):                 5.68
Prob(Q):                              0.82   Prob(JB):                         0.06
Heteroskedasticity (H):               2.87   Skew:                             0.47
Prob(H) (two-sided):                  0.00   Kurtosis:                         3.44
===================================================================================
```

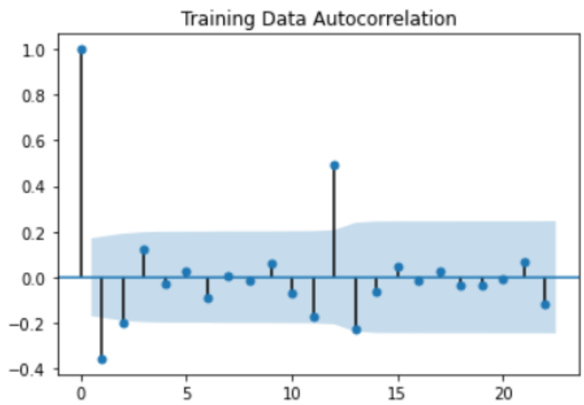
- From the Manual Sarima model we observed (3,1,3) has low AIC value 21555

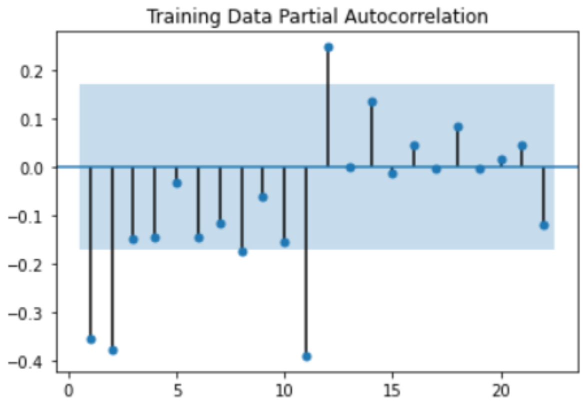


- So we calculated Arima model with (3,1,3) which has RMSE Value 1425.09. Model performs good but lower than triple exponential smoothing.

Rose:

Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data





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.

Manual ARIMA for Rose:

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:                 D.Rose   No. Observations:                  131
Model:                 ARIMA(3, 1, 3)   Log Likelihood                -628.597
Method:                       css-mle   S.D. of innovations             28.356
Date:                Sat, 25 Dec 2021   AIC                           1273.194
Time:                        22:25:24   BIC                           1296.196
Sample:                    02-29-1980   HQIC                          1282.541
                         - 12-31-1990
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const           -0.4906      0.088     -5.548      0.000      -0.664      -0.317
ar.L1.D.Rose    -0.7243      0.086     -8.411      0.000      -0.893      -0.556
ar.L2.D.Rose    -0.7218      0.087     -8.342      0.000      -0.891      -0.552
ar.L3.D.Rose     0.2763      0.085      3.234      0.001       0.109       0.444
ma.L1.D.Rose    -0.0151      0.045     -0.339      0.735      -0.102       0.072
ma.L2.D.Rose     0.0151      0.044      0.340      0.734      -0.072       0.102
ma.L3.D.Rose    -1.0000      0.046    -21.901      0.000      -1.089      -0.911
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1           -0.5011           -0.8661j            1.0006           -0.3335
AR.2           -0.5011           +0.8661j            1.0006            0.3335
AR.3            3.6142           -0.0000j            3.6142           -0.0000
MA.1            1.0000           -0.0000j            1.0000           -0.0000
MA.2           -0.4925           -0.8703j            1.0000           -0.3320
MA.3           -0.4925           +0.8703j            1.0000            0.3320
```

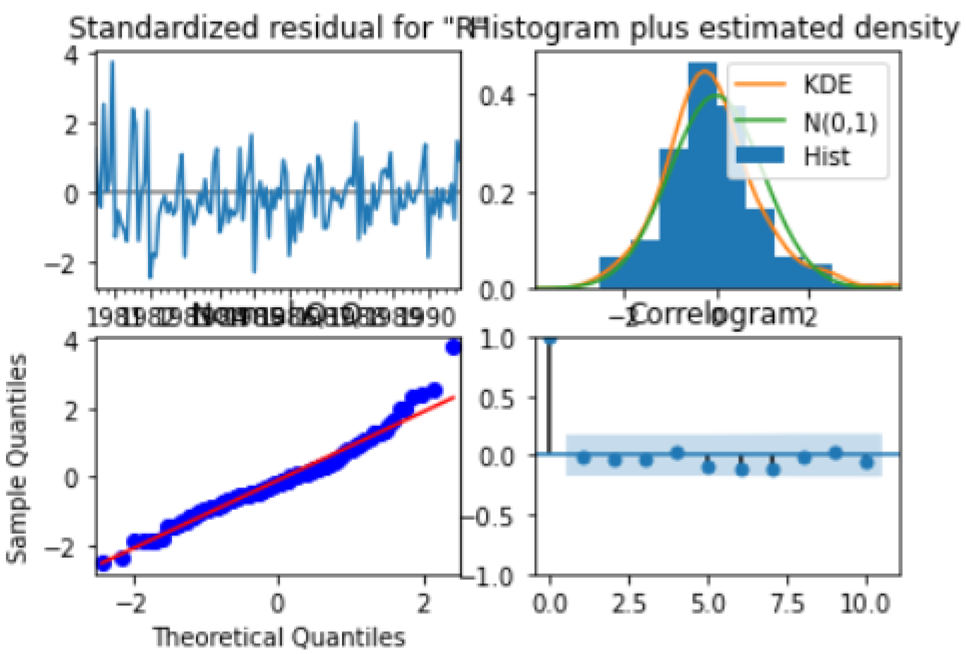
- From the Manual Arima model we observed (3,1,3) has low AIC value 12733.

- So we calculated Arima model with (3,1,3) which has RMSE Value 15.99425. Model performs good but lower than triple exponential smoothing.

Manual SARIMA for Rose:

```
                                SARIMAX Results                                
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:               SARIMAX(3, 1, 3)   Log Likelihood                -614.869
Date:                Sat, 25 Dec 2021   AIC                           1243.739
Time:                        22:25:26   BIC                           1263.648
Sample:                    01-31-1980   HQIC                          1251.827
                         - 12-31-1990                                         
Covariance Type:                  opg                                         
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.5771      0.087    -18.067      0.000      -1.748      -1.406
ar.L2         -0.6528      0.143     -4.574      0.000      -0.933      -0.373
ar.L3          0.1257      0.091      1.383      0.167      -0.053       0.304
ma.L1          0.9505     14.216      0.067      0.947     -26.912      28.812
ma.L2         -0.7057      2.241     -0.315      0.753      -5.099       3.687
ma.L3         -0.9147     13.953     -0.066      0.948     -28.262      26.433
sigma2       878.3546   1.34e+04      0.065      0.948   -2.54e+04    2.72e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                22.33
Prob(Q):                              0.91   Prob(JB):                         0.00
Heteroskedasticity (H):               0.37   Skew:                             0.68
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.55
===================================================================================
```

- From the Manual SARIMA model we observed (3,1,3) has low AIC value 1243.



- So we calculated SARIMA model with (3,1,3) (0,0,0,6)which has RMSE Value 15.994. Model performs good but lower than triple exponential smoothing.

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

Sparkling:

	Test RMSE	RMSE
RegressionOnTime	1389.135175	NaN
NaiveModel	3864.279352	NaN
SimpleAverageModel	1275.081804	NaN
2pointTrailingMovingAverage	813.400684	NaN
4pointTrailingMovingAverage	1156.589694	NaN
6pointTrailingMovingAverage	1283.927428	NaN
9pointTrailingMovingAverage	1346.278315	NaN
Alpha=0.995,SimpleExponentialSmoothing	1316.035487	NaN
Alpha=0.3,SimpleExponentialSmoothing	1935.507132	NaN
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	18259.110704	NaN
Alpha=0.111,Beta=0.0616,Gamma=0.394,TripleExponentialSmoothing	469.767970	NaN
Alpha=0.3,Beta=0.3,Gamma=0.3,TripleExponentialSmoothing	392.786198	NaN
ARIMA(2,1,2)	NaN	1374.037009
SARIMA(2,1,3)(2,0,3,6)	825.679766	NaN
ARIMA(3,1,3)	NaN	1425.093334
SARIMA(3,1,3)(0,0,0,6)	1286.652409	NaN

Rose:

	Test RMSE	RMSE
RegressionOnTime	15.275520	NaN
NaiveModel	79.738146	NaN
SimpleAverageModel	53.480456	NaN
2pointTrailingMovingAverage	11.529720	NaN
4pointTrailingMovingAverage	14.456379	NaN
6pointTrailingMovingAverage	14.570846	NaN
9pointTrailingMovingAverage	14.731252	NaN
Alpha=0.995,SimpleExponentialSmoothing	36.816502	NaN
Alpha=0.3,SimpleExponentialSmoothing	47.524854	NaN
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	265.591576	NaN
Alpha=0.111,Beta=0.0616,Gamma=0.394,TripleExponentialSmoothing	21.046473	NaN
Alpha=0.3,Beta=0.45,Gamma=0.3,TripleExponentialSmoothing	10.950048	NaN
ARIMA(3, 1, 3)	15.994252	NaN
SARIMA(2,1,3)(2,0,3,6)	27.148086	NaN
ARIMA(3,1,3)	NaN	15.994252
SARIMA(3,1,3)(0,0,0,6)	36.465056	NaN

Inference:

- For this data, we had both trend and seasonality so by definition Triple Exponential Smoothing is supposed to work better than the Simple Exponential Smoothing as well as the Double Exponential Smoothing.
- However, since this was a model building exercise we had gone on to build different models on the data and have compared these model with the best RMSE value on the test data.

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.

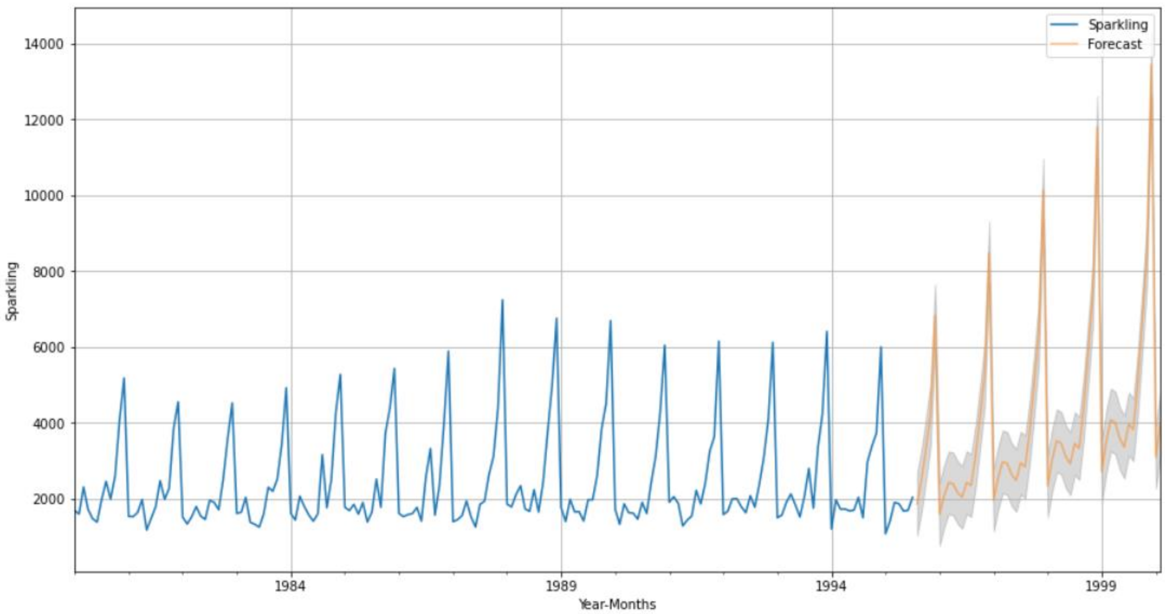
- In this particular we have built several models and went through a model building exercise. This particular exercise has given us an idea as to which particular model gives us the least error on our test set for this data. But in Time Series Forecasting, we need to be very vigil about the fact that after we have done this exercise we need to build the model on the whole data.
- Remember, the training data that we have used to build the model stops much before the data ends. In order to forecast using any of the models built, we need to build the models again (this time on the complete data) with the same parameters.

The two models to be built on the whole data are the following:

Alpha=0.111,Beta=0.06167,Gamma=0.394,TripleExponentialSmoothing

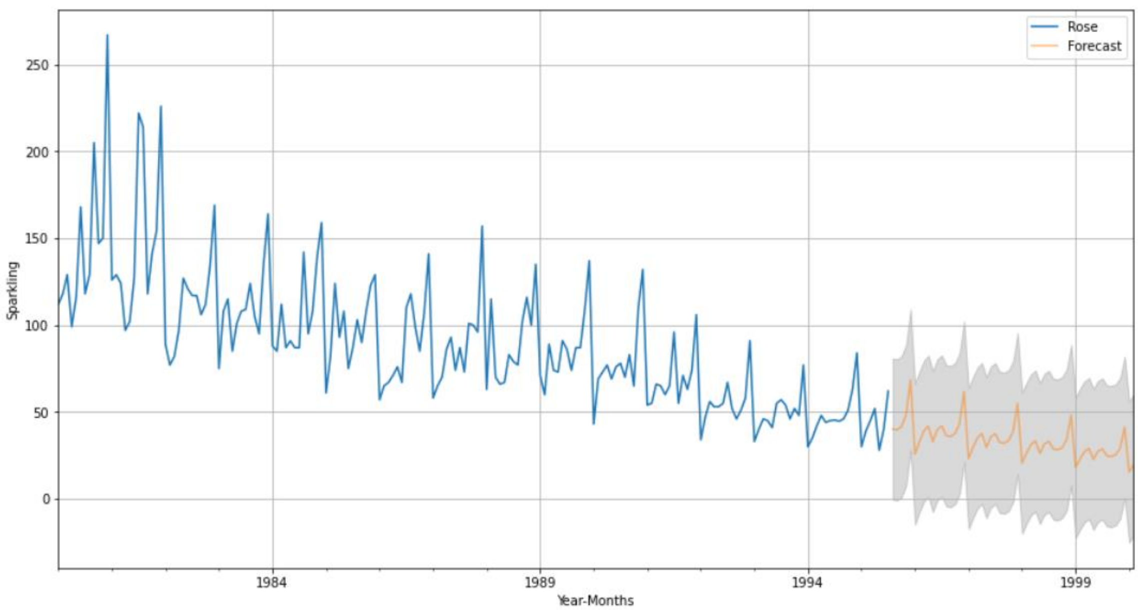
Alpha=0.3,Beta=0.3,Gamma=0.3,TripleExponentialSmoothing

Sparkling:



```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                            Sparkling   No. Observations:                187
Model:             SARIMAX(2, 1, 3)x(2, 0, 3, 6)   Log Likelihood               -1208.895
Date:                           Sun, 26 Dec 2021   AIC                           2439.790
Time:                                   00:08:22   BIC                           2473.889
Sample:                               01-31-1980   HQIC                          2453.633
                                    - 07-31-1995
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.4971      1.388     -0.358      0.720      -3.218       2.224
ar.L2          0.0956      0.938      0.102      0.919      -1.742       1.933
ma.L1         -0.4549      1.377     -0.330      0.741      -3.154       2.244
ma.L2         -0.7608      0.538     -1.414      0.157      -1.815       0.293
ma.L3          0.1194      1.129      0.106      0.916      -2.094       2.333
ar.S.L6        0.0096      0.018      0.521      0.602      -0.027       0.046
ar.S.L12       1.0172      0.011     93.113      0.000       0.996       1.039
ma.S.L6       -0.3161      0.197     -1.603      0.109      -0.703       0.070
ma.S.L12      -0.8705      0.113     -7.724      0.000      -1.091      -0.650
ma.S.L18      -0.0949      0.146     -0.650      0.515      -0.381       0.191
sigma2      9.507e+04   1.84e+04      5.159      0.000     5.9e+04    1.31e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                31.05
Prob(Q):                              0.96   Prob(JB):                         0.00
Heteroskedasticity (H):               1.14   Skew:                             0.50
Prob(H) (two-sided):                  0.64   Kurtosis:                         4.88
===================================================================================
```

Rose:



```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                Rose   No. Observations:                  187
Model:             SARIMAX(2, 1, 3)x(2, 0, 3, 6)   Log Likelihood                -675.256
Date:                             Sat, 25 Dec 2021   AIC                           1372.512
Time:                                     22:25:32   BIC                           1406.611
Sample:                                 01-31-1980   HQIC                          1386.355
                                      - 07-31-1995
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5257      0.061     -8.685      0.000      -0.644      -0.407
ar.L2         -0.6847      0.054    -12.742      0.000      -0.790      -0.579
ma.L1         -0.2432      0.072     -3.373      0.001      -0.385      -0.102
ma.L2          0.2349      0.074      3.188      0.001       0.090       0.379
ma.L3         -0.7576      0.075    -10.090      0.000      -0.905      -0.610
ar.S.L6       -0.0542      0.034     -1.619      0.105      -0.120       0.011
ar.S.L12       0.8638      0.034     25.307      0.000       0.797       0.931
ma.S.L6        0.1693      4.022      0.042      0.966      -7.713       8.052
ma.S.L12      -0.6534      3.337     -0.196      0.845      -7.193       5.886
ma.S.L18       0.1745      0.749      0.233      0.816      -1.293       1.642
sigma2       197.9963    793.627      0.249      0.803   -1357.483    1753.476
===================================================================================
Ljung-Box (L1) (Q):                   0.35   Jarque-Bera (JB):                18.17
Prob(Q):                              0.56   Prob(JB):                         0.00
Heteroskedasticity (H):               0.21   Skew:                            -0.30
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.51
```

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

- The recommendation to Business is to use “Triple Exponential model” to predict the wine sales for the future, as it errors minimum and predicts as good as the original sales.
- Consumers get very excited about savings and appreciate discounts being passed on. Many prominent retailers also have loyalty programs or club member cards that create excitement. A club-member price brings consumers back and gives you an edge on your competition.
- Need to arrange proper training to sales person, So they can convice the customer it seems to increase the sales.
- In month which has low sales of wines need to arrange business events, weddings party, and summer concert series, and these are all great ways to leverage a stunning location to increase revenue.
- The sales seem to be higher in the month of December and that obvious due to Christmas and new year celebrations. We can increase the production with proper advertisements will increase the sales. As per previous data sales will be decreased in beginning of the year so we need to deepdive and work on stratergies.
- Sales person should post product regurly in social media will definetly increase the sales.
- Wine sales is fluctuating over the month, So we need to get customer review based on that we make the decision.

- There seems to be sales not increasing it is sometimes increasing and sometimes decreasing, so business has to act on quickly to improve their process and strategy to increase sale of this wine.
- Wine outlet should adapt new trend to sale the wines. We need work deeply on cache advertisement,customer service and creating ambiance for customer