BUSINESS REPORT
Time Series Forecasting
Rose Wine Dataset

Submitted By:
Ayush
Agarwal

Index

Problem: - For this particular assignment, the data of different types of wine sales in the 20th century is to be analyzed. 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 analyze and forecast Wine Sales in the 20th century.

Dataset 2: -

(Rose Wines)

2.1 : - Read the data as an appropriate Time Series data and plot the data.

Converting the data into appropriate time series data our dataset will look like this:

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

Table No. 2.1.1

```
['1980-01-01', '1980-02-01', '1980-03-01', '1980-04-01',
 '1980-05-01', '1980-06-01', '1980-07-01', '1980-08-01',
 '1980-09-01', '1980-10-01',
 ...
 '1994-10-01', '1994-11-01', '1994-12-01', '1995-01-01',
 '1995-02-01', '1995-03-01', '1995-04-01', '1995-05-01',
 '1995-06-01', '1995-07-01'],
```

Fig No. 2.1.1

To understand this time series properly we will plot the data,

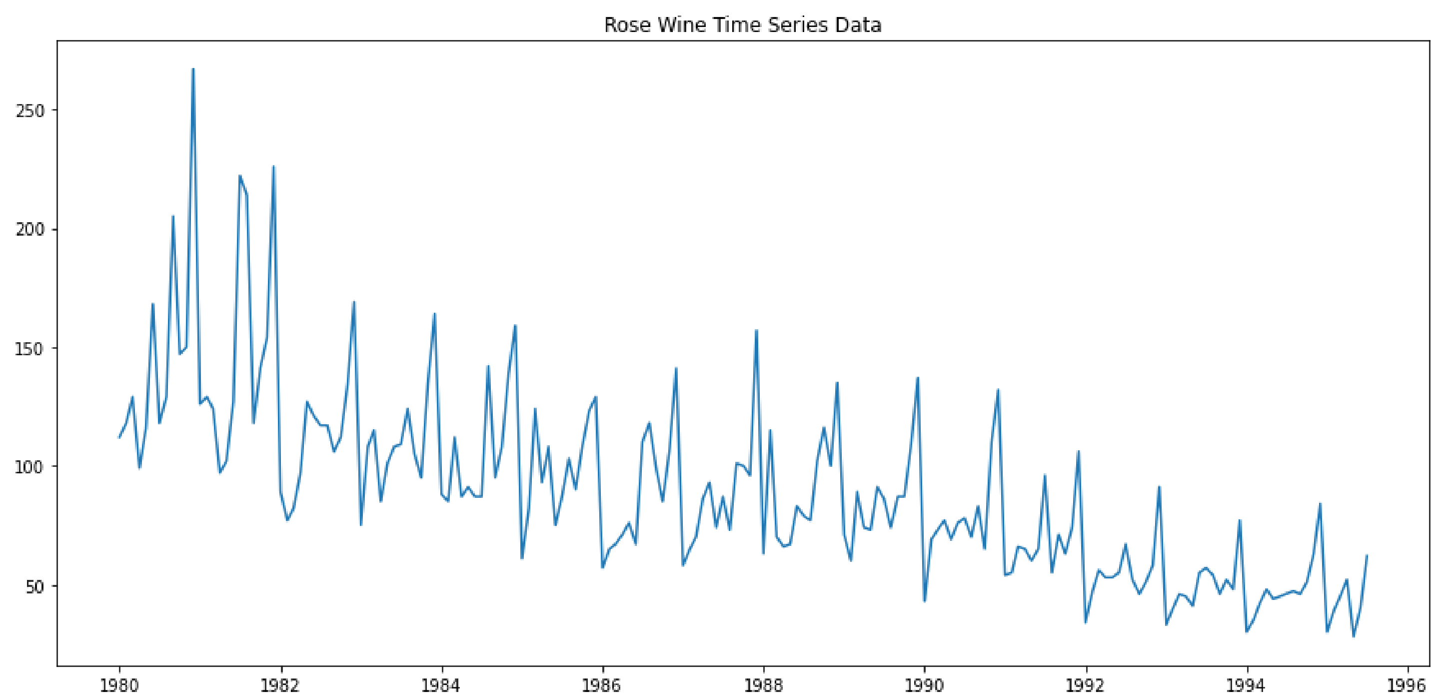


Fig. No. 2.1.2

2.2: - Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

Exploratory Data Analysis: -

a.) Shape of the dataset: -

Rose wine data attribute have 187 records from **01-01-1980** to **01-07-1995**.

b.) Checking for null values: -

2 null values are present.

```
1994-07-01       NaN
1994-08-01       NaN
```

For imputing null values, we will use interpolation with spline method:

```
1994-07-01   46.153199
1994-08-01   47.211982
```

c.) Checking Descriptive Statistics of the Dataset: -

	count	mean	std	min	25%	50%	75%	max
Rose	187.0	89.927087	39.224153	28.0	62.5	85.0	111.0	267.0

Table No. 2.2.1

d.) Checking mean and median value comparison along with dataset plot: -

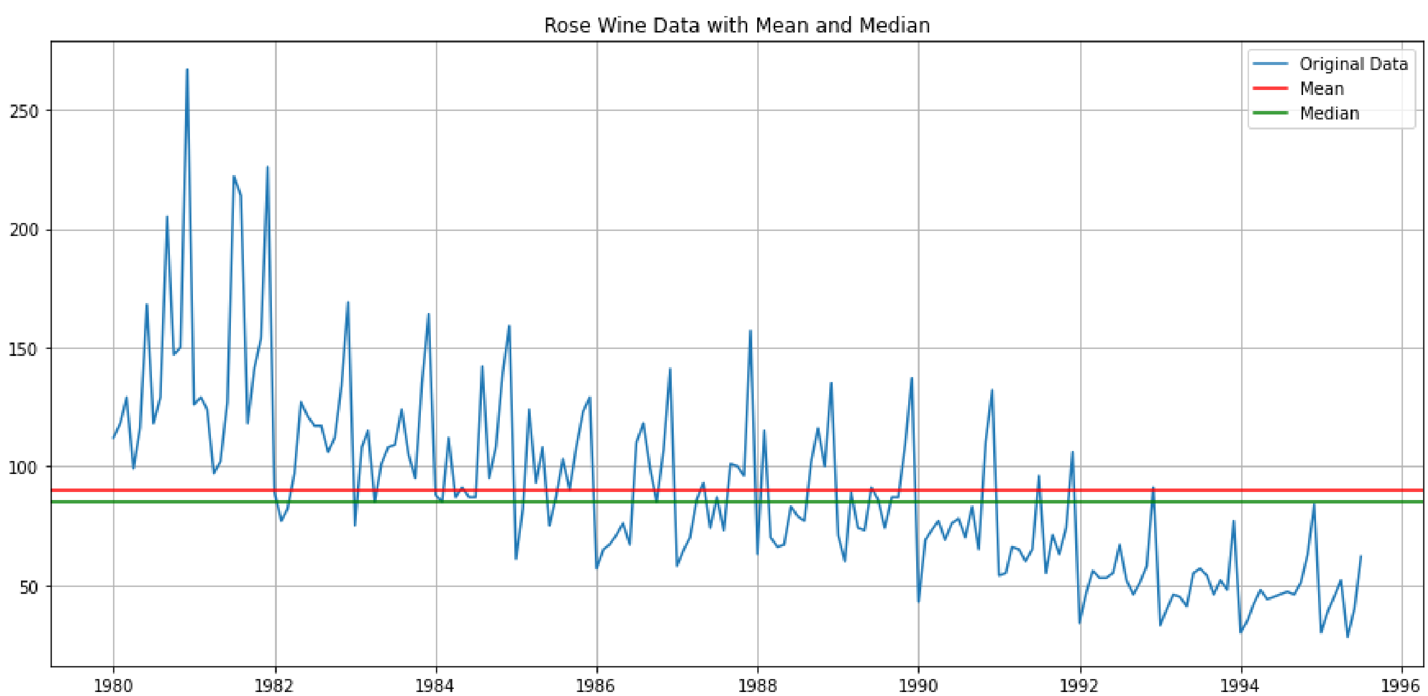


Fig. No. 2.2.1

e.) Checking yearly boxplot of the dataset: -

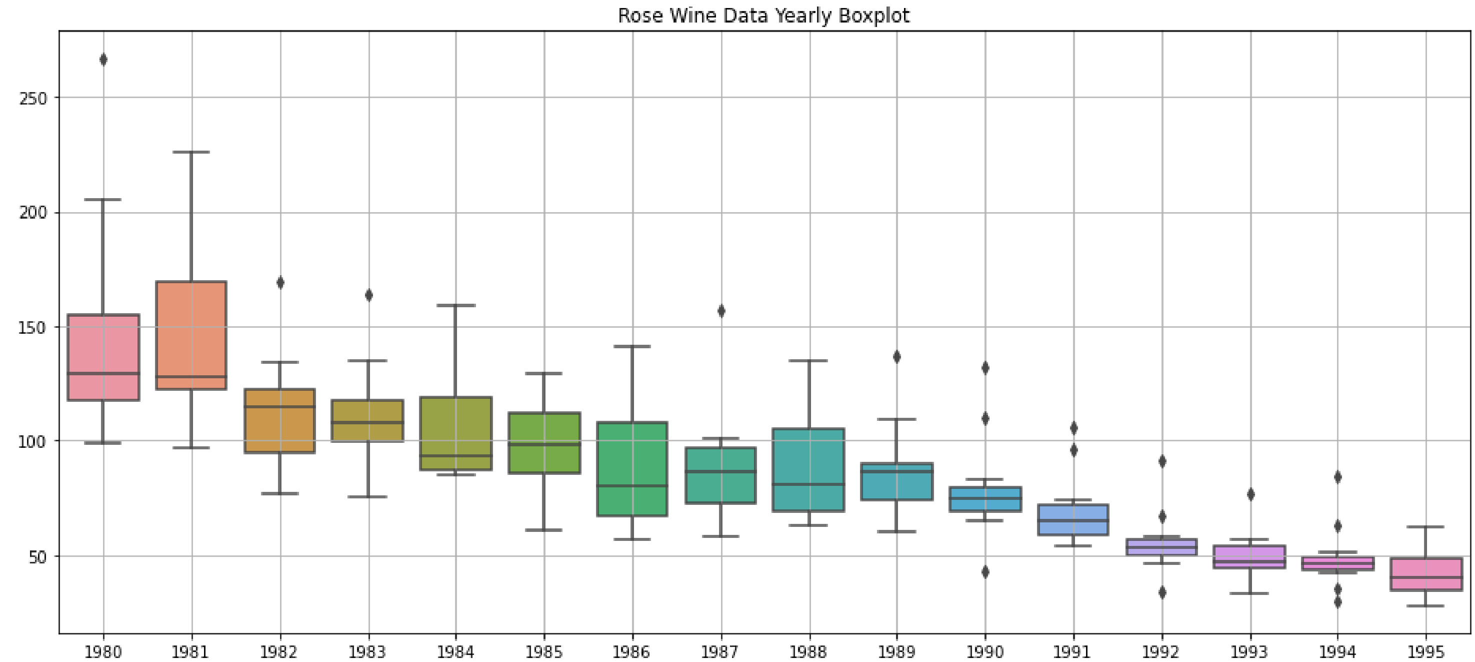


Fig. No. 2.2.2

f.) Checking monthly boxplot of the dataset: -

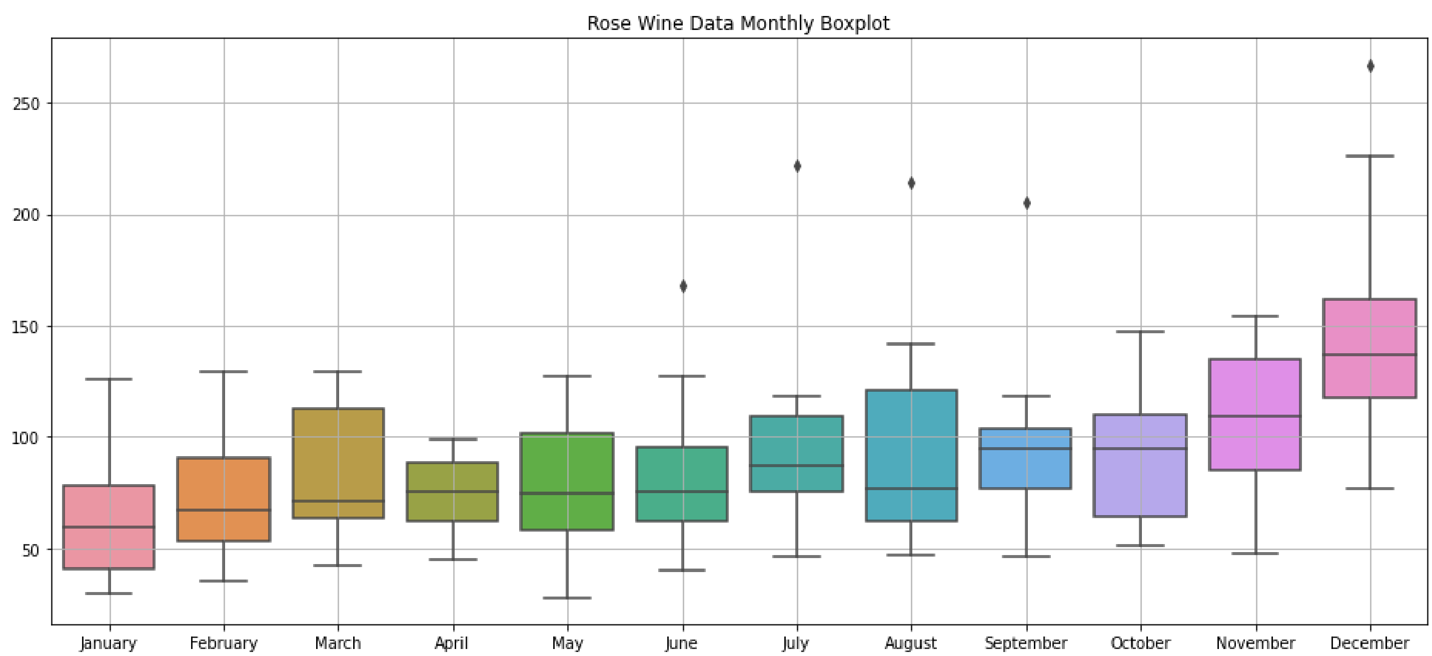


Fig. No. 2.2.3

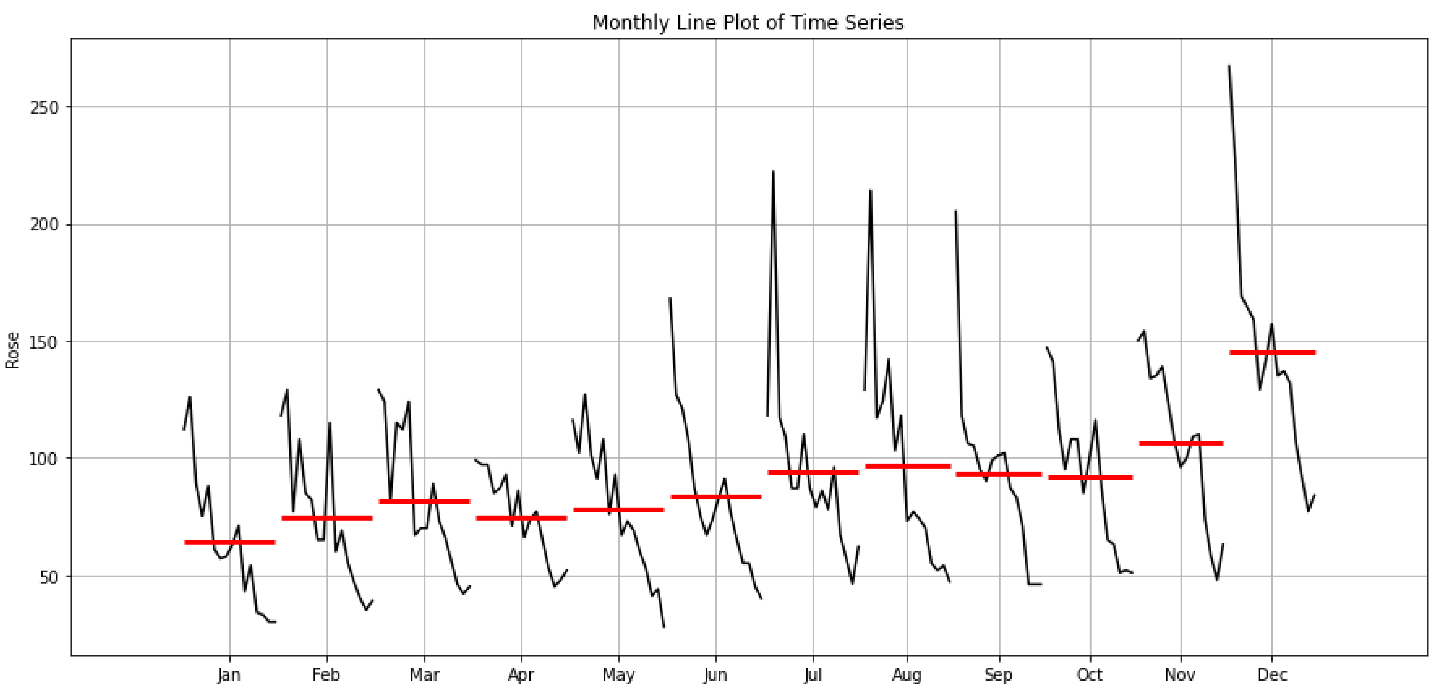
g.) Comparison between monthly and yearly data using line plot: -

I.) Monthly and Yearly Table: -

YearMonth	April	August	December	February	January	July	June	March	May	November	October	September
YearMonth												
1980	99.0	129.000000	267.0	118.0	112.0	118.000000	168.0	129.0	116.0	150.0	147.0	205.0
1981	97.0	214.000000	226.0	129.0	126.0	222.000000	127.0	124.0	102.0	154.0	141.0	118.0
1982	97.0	117.000000	169.0	77.0	89.0	117.000000	121.0	82.0	127.0	134.0	112.0	106.0
1983	85.0	124.000000	164.0	108.0	75.0	109.000000	108.0	115.0	101.0	135.0	95.0	105.0
1984	87.0	142.000000	159.0	85.0	88.0	87.000000	87.0	112.0	91.0	139.0	108.0	95.0
1985	93.0	103.000000	129.0	82.0	61.0	87.000000	75.0	124.0	108.0	123.0	108.0	90.0
1986	71.0	118.000000	141.0	65.0	57.0	110.000000	67.0	67.0	76.0	107.0	85.0	99.0
1987	86.0	73.000000	157.0	65.0	58.0	87.000000	74.0	70.0	93.0	96.0	100.0	101.0
1988	66.0	77.000000	135.0	115.0	63.0	79.000000	83.0	70.0	67.0	100.0	116.0	102.0
1989	74.0	74.000000	137.0	60.0	71.0	86.000000	91.0	89.0	73.0	109.0	87.0	87.0
1990	77.0	70.000000	132.0	69.0	43.0	78.000000	76.0	73.0	69.0	110.0	65.0	83.0
1991	65.0	55.000000	106.0	55.0	54.0	96.000000	65.0	66.0	60.0	74.0	63.0	71.0
1992	53.0	52.000000	91.0	47.0	34.0	67.000000	55.0	56.0	53.0	58.0	51.0	46.0
1993	45.0	54.000000	77.0	40.0	33.0	57.000000	55.0	46.0	41.0	48.0	52.0	46.0
1994	48.0	47.211982	84.0	35.0	30.0	46.153199	45.0	42.0	44.0	63.0	51.0	46.0
1995	52.0	NaN	NaN	39.0	30.0	62.000000	40.0	45.0	28.0	NaN	NaN	NaN

Table No. 2.2.2

II.) Monthly Line Plot with Respect to Every Year: -



III.) Yearly Line Plot with Respect to Every Month: -

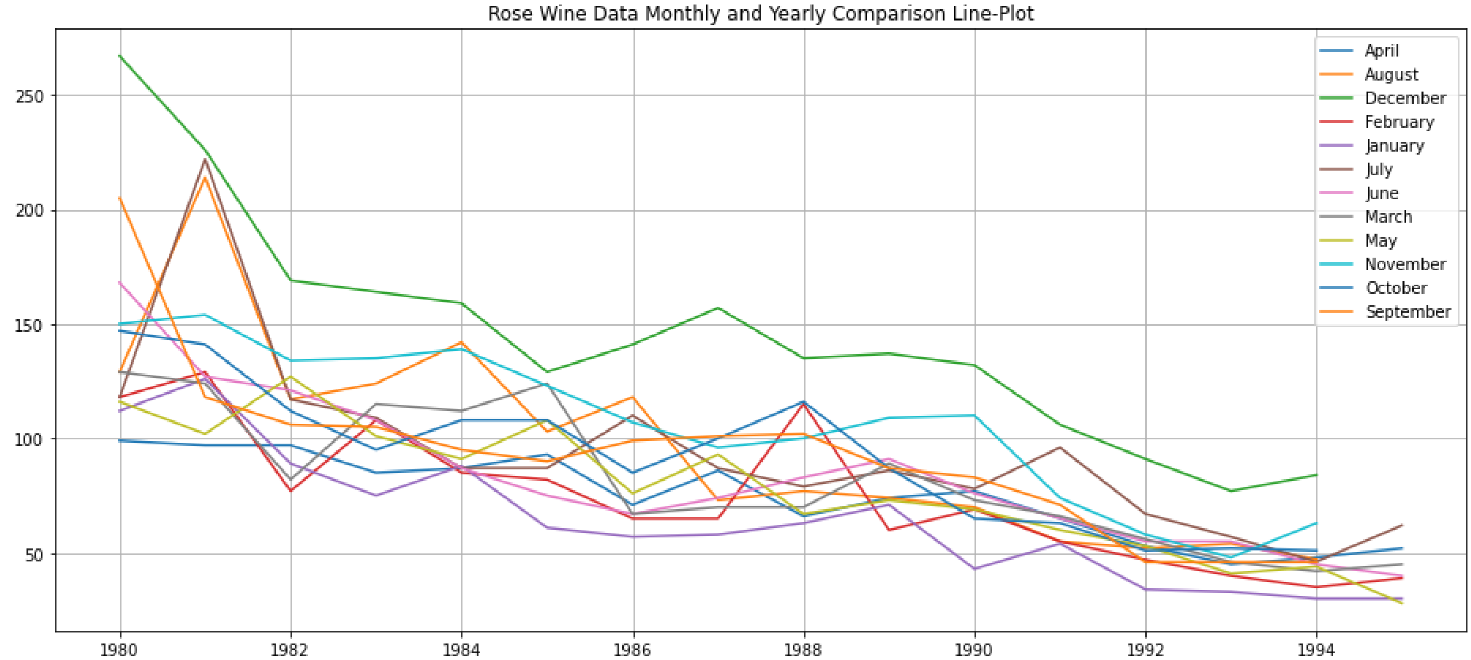


Fig. No. 2.2.4

Insights from the Exploratory Data Analysis: -

- Rose Wine production data is present from Jan-1980 to July-1995.
- The data has 2 null values and it is imputed using interpolation.
- In descriptive statistics it is observed that data has outliers; outliers represent the variation in wine production within the months or years.
- The variation within the month is very less but we can observe that there is production downfall yearby year.
- Only December month produce high volume as compare to othermonths.

Decomposition: -

I.) **Additive: -**

After decomposing dataset using Additive model we get trend, seasonal and residual(error) plot,

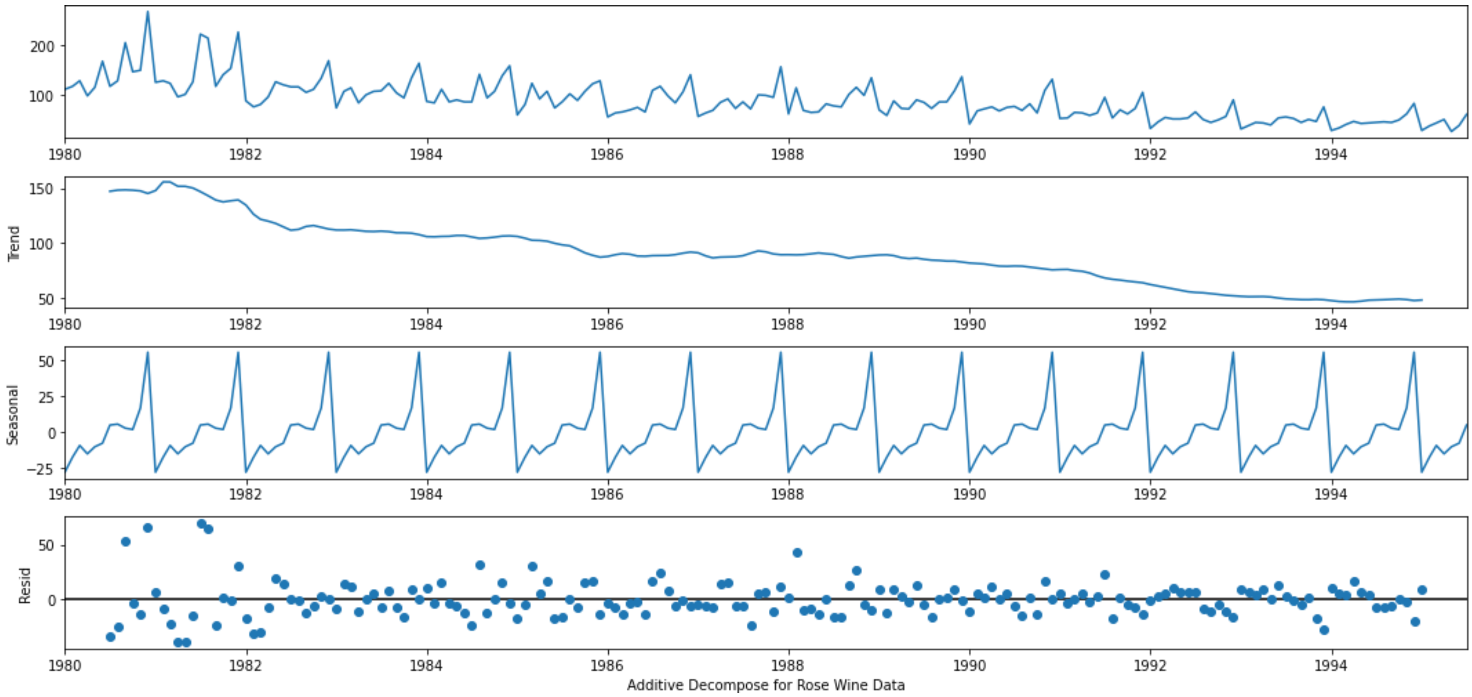


Fig No. 2.2.5

Insights from Additive Decomposition: -

- Strong trend is observed.
- Seasonality is observed.
- Residual (Error) lying within the range of -50 and 50.

II.) **Multiplicative: -**

After decomposing dataset using Multiplicative model we get trend, seasonal and residual(error) plot,

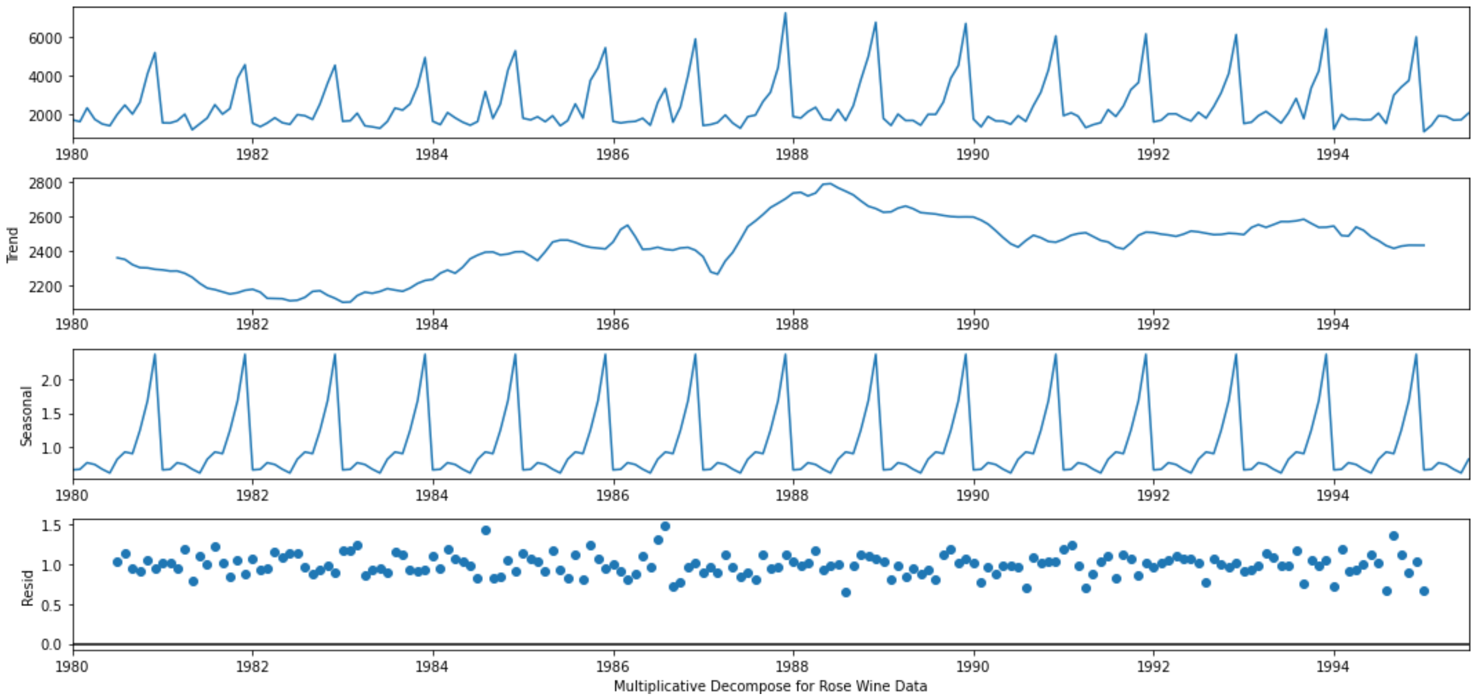


Fig No. 2.2.6

<u>Insights from Multiplicative Decomposition: -</u>

- Trend is not observed.
- Seasonality is observed.
- Residual (Error) lying within the range of 0.5 and 1.5, here this is percentage error.

2.3 : - Split the data into training and test. The test data should start in 1991.

Splitting the dataset into train and test :

```
Top 5 Rows of Train Data
              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
```

```
Top 5 Rows of Test Data
             Rose
YearMonth
1991-01-01   54.0
1991-02-01   55.0
1991-03-01   66.0
1991-04-01   65.0
1991-05-01   60.0
```

- Train Dataset having range from **<u>Jan-1980 to Dec-1990</u>** i.e., 132 records.
- Test Dataset having range from **<u>Jan-1991 to Jul-1995</u>** i.e., 55 records.

Plot of train and test dataset,

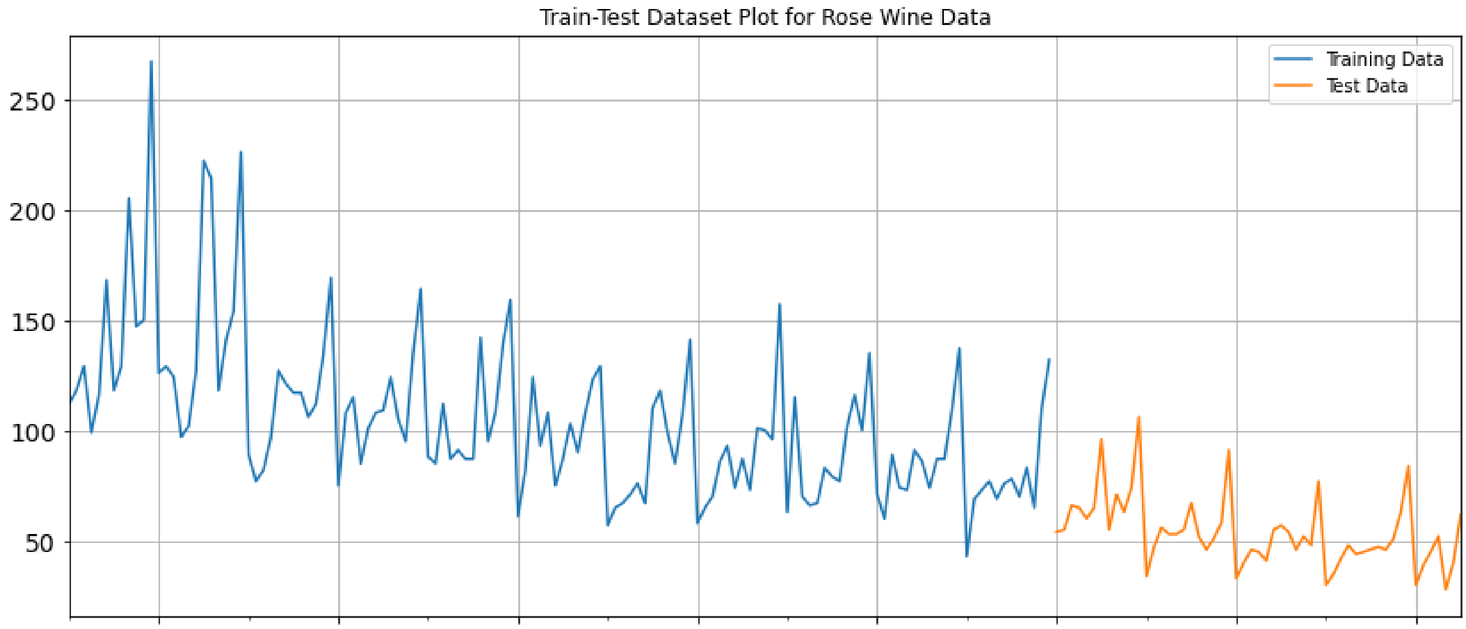


Fig. No. 2.3.1

2.4 : - Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other additional models such as regression, naïve forecast models, simple average models, moving average models should also be built on the training data and check the performance on the test data using RMSE.

First, we will evaluate on **Linear Regression, Naïve Model, Simple Average, Moving Average** and then on **exponential smoothing.**

Linear Regression

After adding date range column in an ordinal format to the regression model as independent variable we can forecast accordingly, dataset after adding date range it will look like,

Top 5 Rows for Linear Regression Train

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

Top 5 Rows for Linear Regression Test

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

After training the dataset on train dataset and predicting it on test dataset we got our predicted values which can be visualize via this plot,

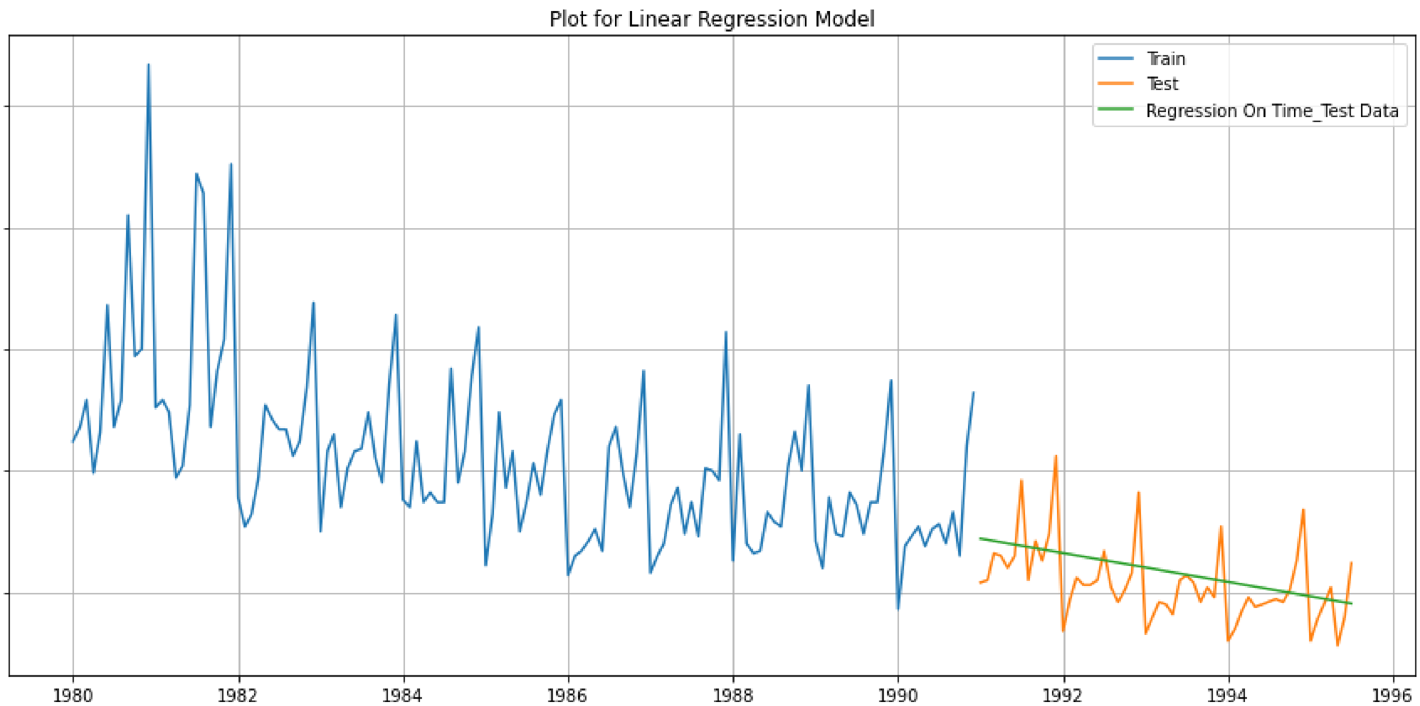


Fig No. 2.4.1

Model Evaluation: -

We can evaluate the model by calculating the RSME (Root Mean Square Error) on Test Data, minimum the RSME better the model and for this model RSME would be,

	Test RMSE
RegressionOnTime	15.255492

Root Mean Square Error of Linear Regression Model for Test Data is 15.2554

Naïve Model

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

In naïve model, predicted values are of the train dataset last value. for this model forecast values would be 132.

After getting the predicted values we can visualize our data,

Plot for Naive Model

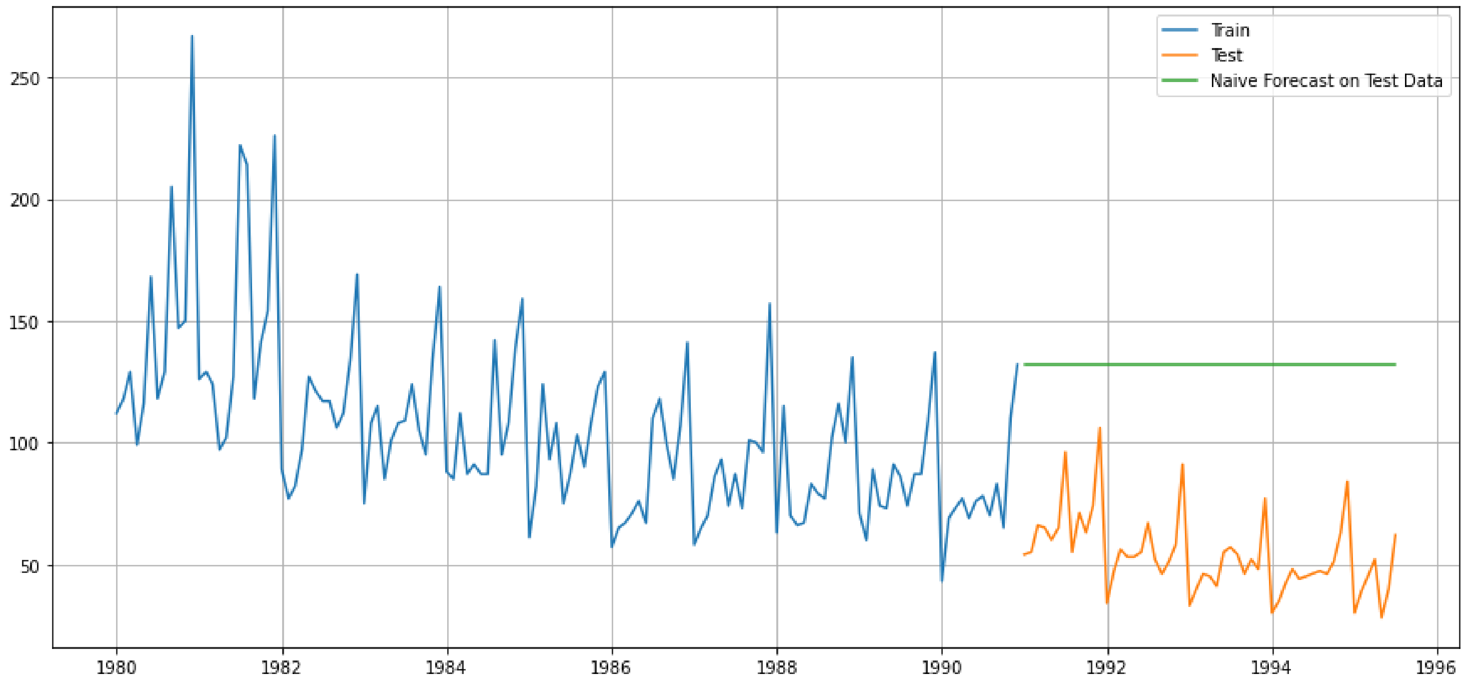


Fig No. 2.4.2

Model Evaluation: -

For this model RSME would be,

	Test RMSE
RegressionOnTime	15.255492
NaiveOnTime	79.672475

Root Mean Square Error of Naive Model for Test Data is 79.6724.

Simple Average Model

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

For simple average model, the mean of train dataset is 104.9393.

After getting the predicted values we can visualize our data,

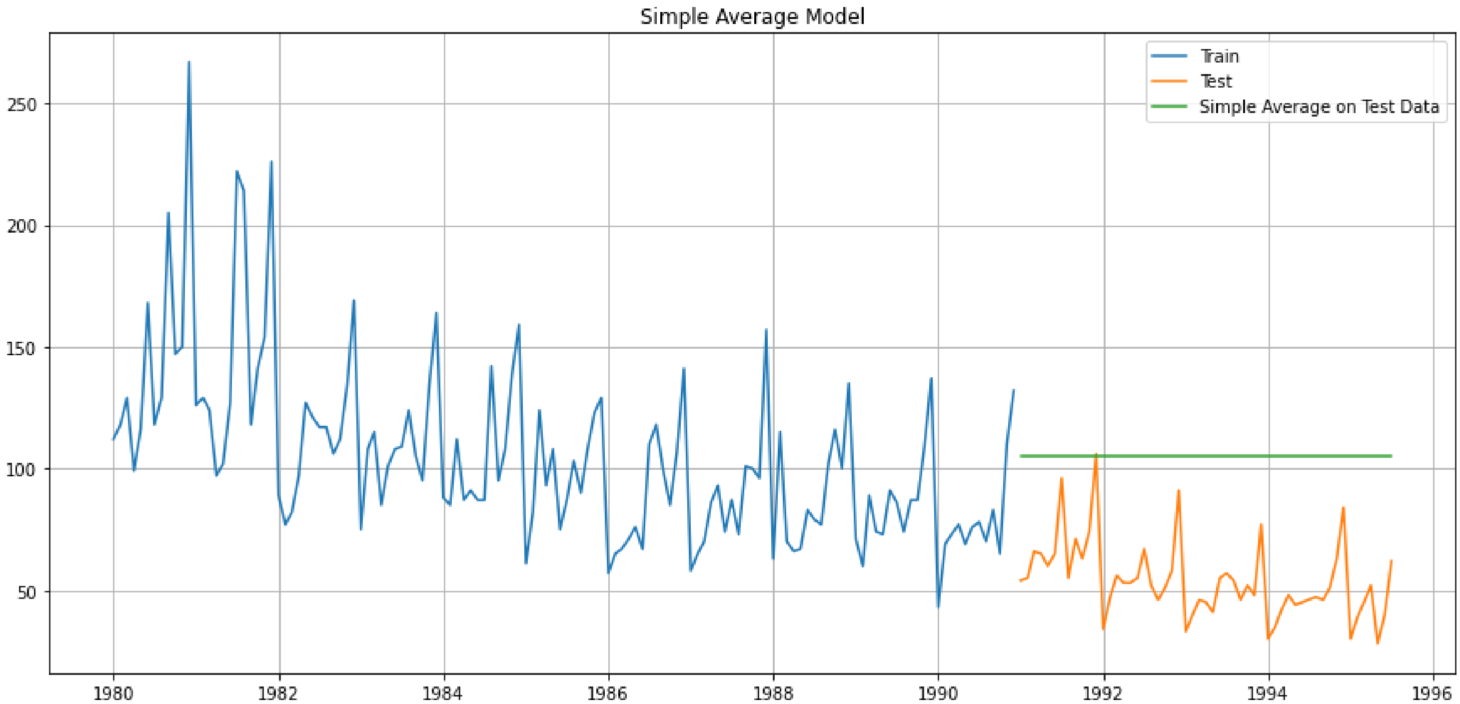


Fig No. 2.4.3

Model Evaluation: -

For this model RSME would be,

	Test RMSE
RegressionOnTime	15.255492
NaiveOnTime	79.672475
SimpleAverage	53.413298

Root Mean Square Error of Simple Average Model for Test Data is 53.4132.

Moving Average Model

YearMonth	Rose	Trailing2	Trailing4	Trailing6	Trailing9
1980-01-01	112.0	NaN	NaN	NaN	NaN
1980-02-01	118.0	115.0	NaN	NaN	NaN
1980-03-01	129.0	123.5	NaN	NaN	NaN
1980-04-01	99.0	114.0	114.5	NaN	NaN
1980-05-01	116.0	107.5	115.5	NaN	NaN

A moving average is defined as an average of fixed number of items in the time series which move through the series by dropping the top items of the previous averaged group and adding the next in each successive average.
Here we took averages of 2,4,6 and 9 successive records.

After getting the predicted values on each moving average rolling parameter we can visualize our data,

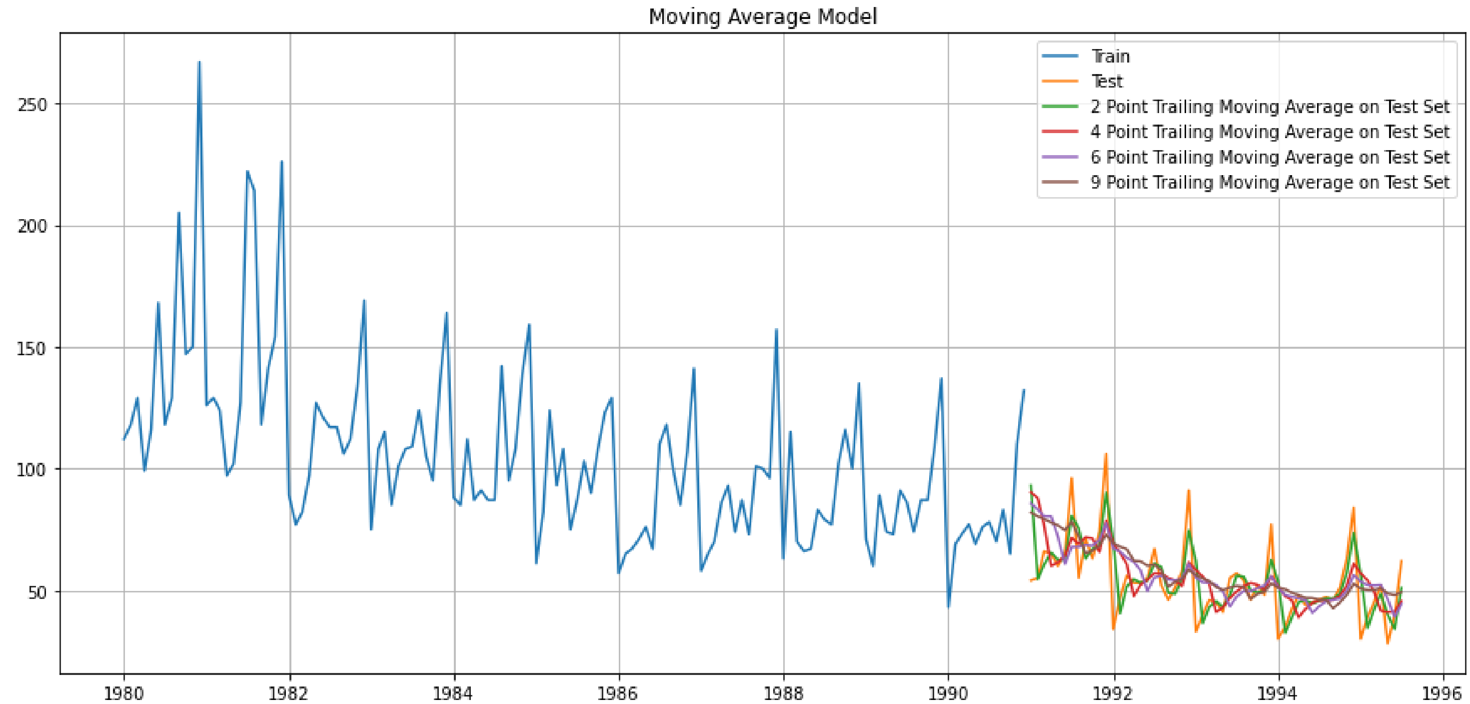


Fig No. 2.4.4

Here it is difficult to say that which rolling parameter is best to analyze, we will check RSME values for each parameter.

Model Evaluation: -

For this model RSME would be,

	Test RMSE
RegressionOnTime	15.255492
NaiveOnTime	79.672475
SimpleAverage	53.413298
2 Point Trailing on Test Data	11.529985
4 Point Trailing on Test Data	14.444375
6 Point Trailing on Test Data	14.554986
9 Point Trailing on Test Data	14.721520

Root Mean Square Error of Moving Average Model for Test Data for 2 Point Trailing Average is 11.5299.

We can also compare all forecasting parameter using plot,

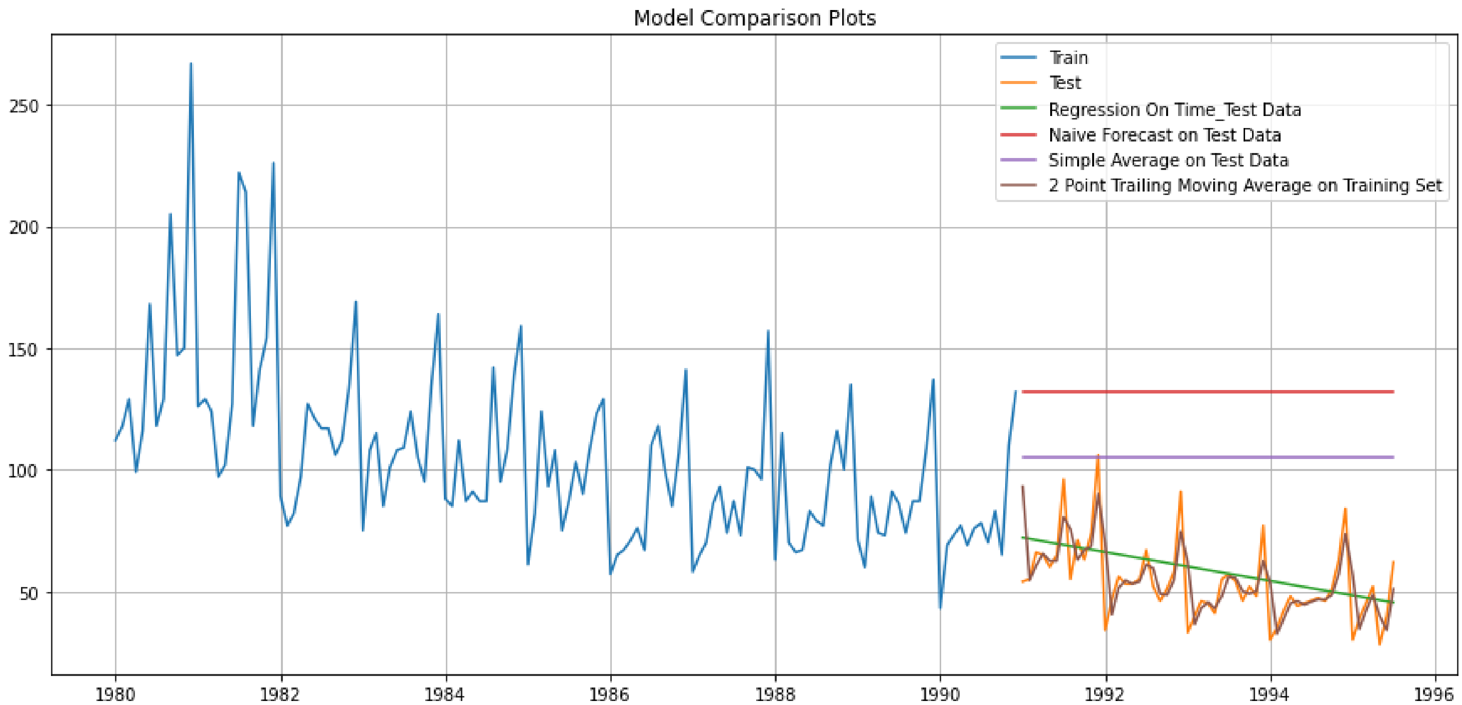


Fig No. 2.4.5

From the above plot it is clearly seen that **2-point trailing moving average** have best fitted line to test dataset, and for more better forecast result and RSME value we will go for Exponential Smoothing Method.

Simple Exponential Smoothing

For more better model we will evaluate our model using exponential smoothing, and simple exponential smoothing is one of them, here we consider smoothing level only and after fitting our model we will get best parameter to analyze further,

```
{'smoothing_level': 0.09874983698117956,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 134.38702481818487,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Best Parameters for Simple Exponential Smoothing

```
1991-01-01    87.104997
1991-02-01    87.104997
1991-03-01    87.104997
1991-04-01    87.104997
1991-05-01    87.104997
Freq: MS, dtype: float64
```

Top 5 Rows of Predicted Values on Test Data for Simple Exponential Smoothing

After prediction on best parameters, we will plot the data for better understanding,

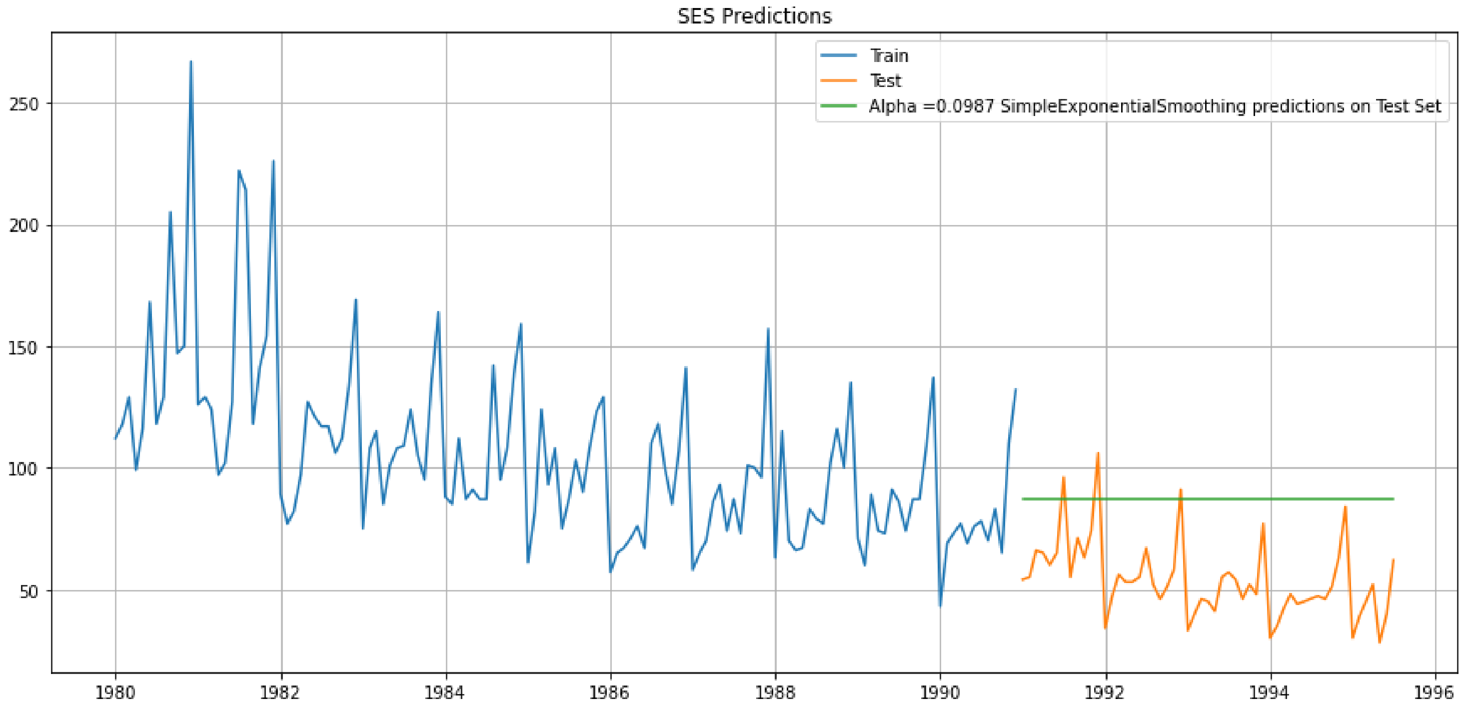


Fig. No. 2.4.6

Model Evaluation: -

For this model RSME would be,

	Test RMSE
Alpha=0.0987 SimpleExponentialSmoothing	36.748402

Root Mean Square Error of Simple Exponential Smoothing Model for Test Data is 36.748

Double Exponential Smoothing

Here we consider smoothing level and smoothing trend, after fitting our model we will get best parameter to analyze further,

```
{'smoothing_level': 1.4901161193847656e-08,
 'smoothing_trend': 1.6610391146660035e-10,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 137.81553690867275,
 'initial_trend': -0.4943781897068274,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Best Parameters for Double Exponential Smoothing

```
1991-01-01    72.063238
1991-02-01    71.568859
1991-03-01    71.074481
1991-04-01    70.580103
1991-05-01    70.085725
Freq: MS, dtype: float64
```

Top 5 Rows of Predicted Values on Test Data for Double Exponential Smoothing

After prediction on best parameters, we will plot the data for better understanding,

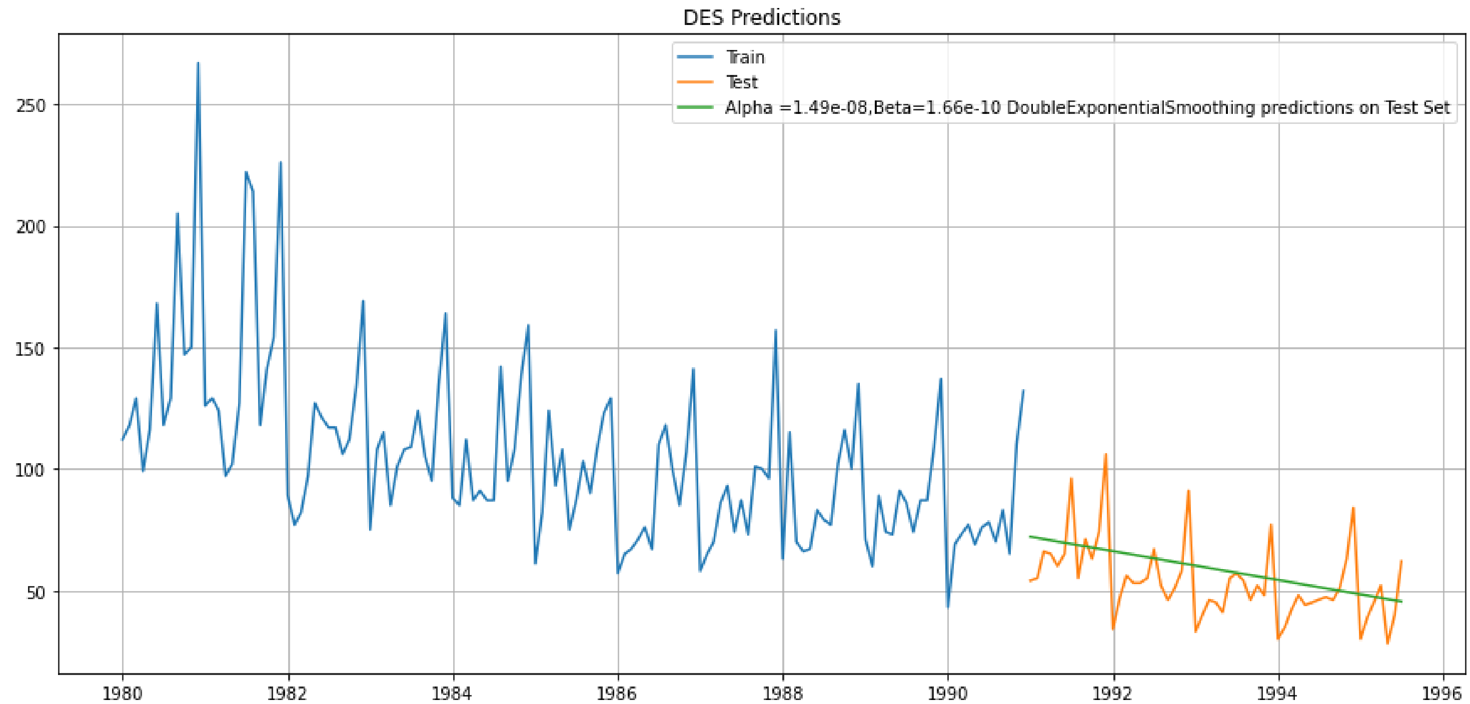


Fig. No. 2.4.7

Model Evaluation: -
For this model RSME would be,

	Test RMSE
Alpha=0.0987 SimpleExponentialSmoothing	36.748402
Alpha =1.49e-08,Beta=1.66e-10 DoubleExponentialSmoothing	15.255480

Root Mean Square Error of Double Exponential Smoothing for Test Data is 15.2554, it is clearly seen that RSME is improved for this model.

Triple Exponential Smoothing

Here we consider smoothing level, smoothing trend and as well as smoothing seasonality, after fitting our model we will get best parameter to analyze further,

```
{'smoothing_level': 0.0715106306609405,
 'smoothing_trend': 0.04529179757535142,
 'smoothing_seasonal': 7.244325029450242e-05,
 'damping_trend': nan,
 'initial_level': 130.40839142502193,
 'initial_trend': -0.77985743179386,
 'initial_seasons': array([0.86218996, 0.977675  , 1.0687727 , 0.93403881, 1.050625
        1.14410977, 1.25836944, 1.33937772, 1.26778766, 1.24131254,
        1.44724625, 1.99553681]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Best Parameters for Triple Exponential Smoothing

After prediction on best parameters, we will plot the data for better understanding,

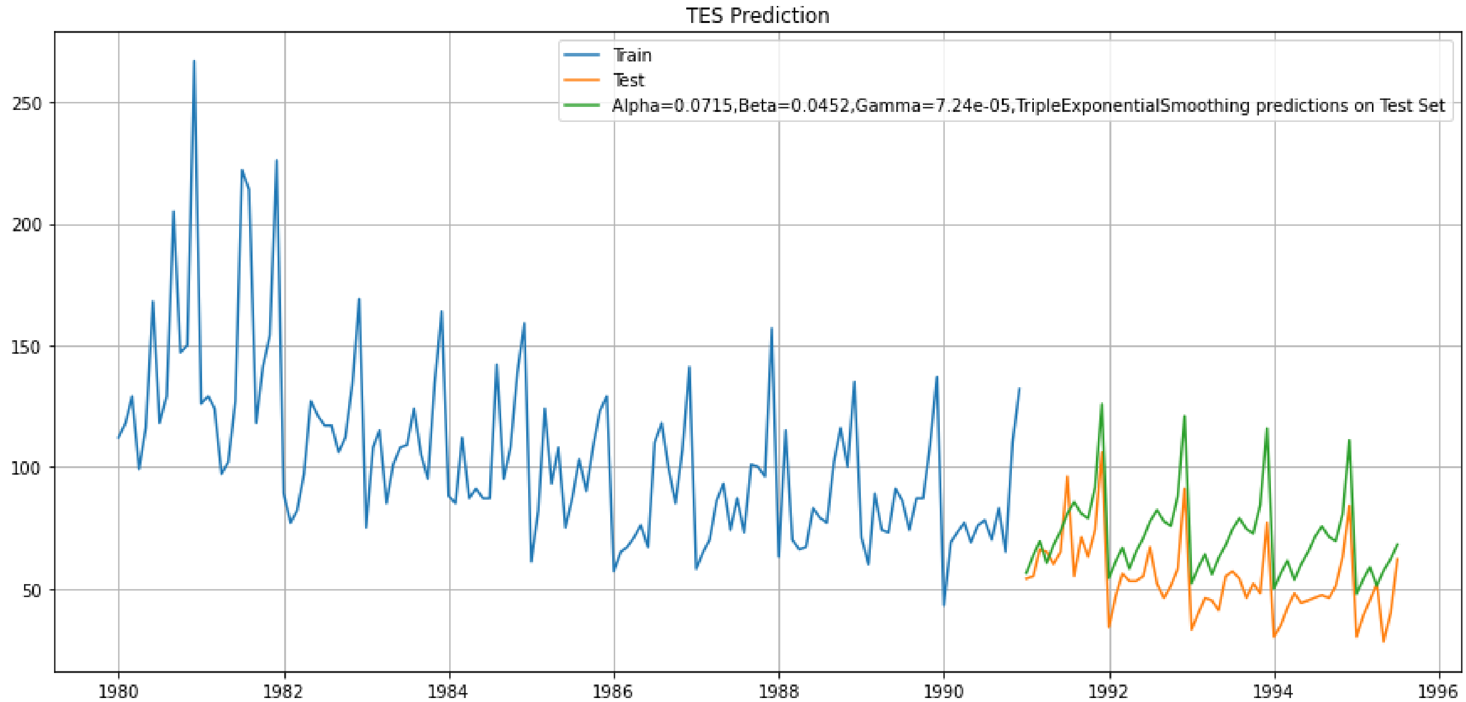


Fig. No. 2.4.7

Model Evaluation: -
For this model RSME would be,

	Test RMSE
Alpha=0.0987 SimpleExponentialSmoothing	36.748402
Alpha =1.49e-08,Beta=1.66e-10 DoubleExponentialSmoothing	15.255480
Alpha=0.0715,Beta=0.0452,Gamma=7.24e-05 TripleExponentialSmoothing	20.097325

Root Mean Square Error of Triple Exponential Smoothing for Test Data is 404.2868, and it is clearly seen that this smoothing expresses the best forecast model as it is showing minimum RSME value.

For better understanding we will visualize all three smoothing model along with train and test dataset,

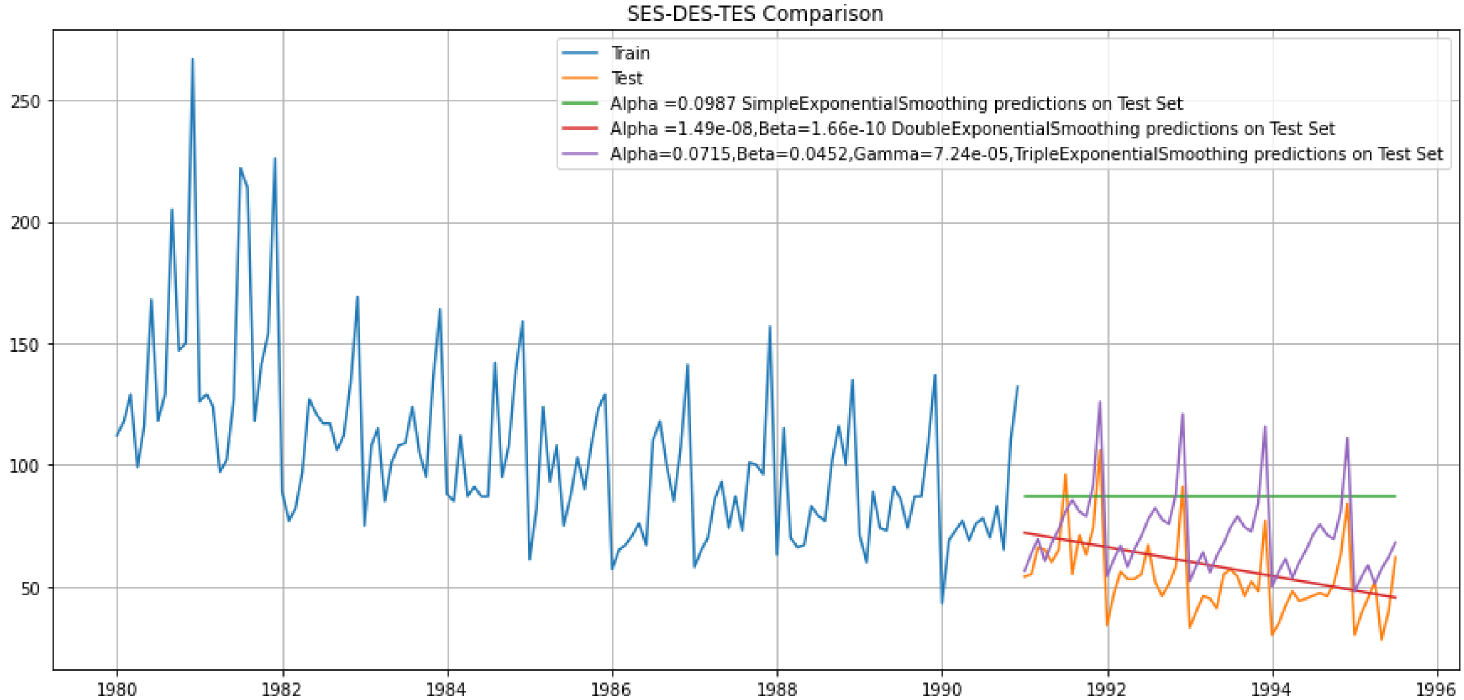


Fig. No. 2.4.8

We are end up with all model like Linear Regression, Naïve Model, Simple Average, Moving Average and all Exponential smoothing models now compare all RSME together and stat which would be the best model for the prediction,

	Test RMSE
RegressionOnTime	15.255492
NaiveOnTime	79.672475
SimpleAverage	53.413298
2 Point Trailing on Test Data	11.529985
4 Point Trailing on Test Data	14.444375
6 Point Trailing on Test Data	14.554986
9 Point Trailing on Test Data	14.721520
Alpha=0.0987 SimpleExponentialSmoothing	36.748402
Alpha =1.49e-08,Beta=1.66e-10 DoubleExponentialSmoothing	15.255480
Alpha=0.0715,Beta=0.0452,Gamma=7.24e-05 TripleExponentialSmoothing	20.097325

Table No. 2.4.1

Conclusion: -

It can be concluded that 2-point Trailing using Moving Average is the best model for forecasting.

2.5 : - 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.

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 α value i.e., 0.05.

If we found p-value greater than α value than we don't have enough evidence to reject the null hypothesis and its stats that data is non stationary.
After applying Augmented Dickey-Fuller we found certain results i.e.,

```
Rose Data test statistic is -2.394
Rose Data test p-value is 0.3830431487073731
Number of lags used 12
Number of Observation Used 174
Critical Values {'1%': -4.011763737803776, '5%': -3.4360292512258863, '10%': -3.1420436590266103}
```

We observed p-value for the dataset is **0.3830** which is greater than significance value (0.05) so in that case we can say that our dataset is non stationary.
To make dataset stationary we will take first order differencing and apply Dickey-Fuller test again,
After applying Dickey-Fuller again result we got,

```
Rose Data test statistic is -8.402
Rose Data test p-value is 8.415846763882622e-12
Number of lags used 11
Number of Observation Used 174
Critical Values {'1%': -4.011763737803776, '5%': -3.4360292512258863, '10%': -3.1420436590266103}
```

Here p-value is less than the level of significance hence we can say that our dataset become stationary after first order differencing.

YearMonth	Rose	Differencing
1980-01-01	112.0	NaN
1980-02-01	118.0	6.0
1980-03-01	129.0	11.0
1980-04-01	99.0	-30.0
1980-05-01	116.0	17.0

We can see new dataset here after first order differencing,
Here is a plot of stationary dataset,

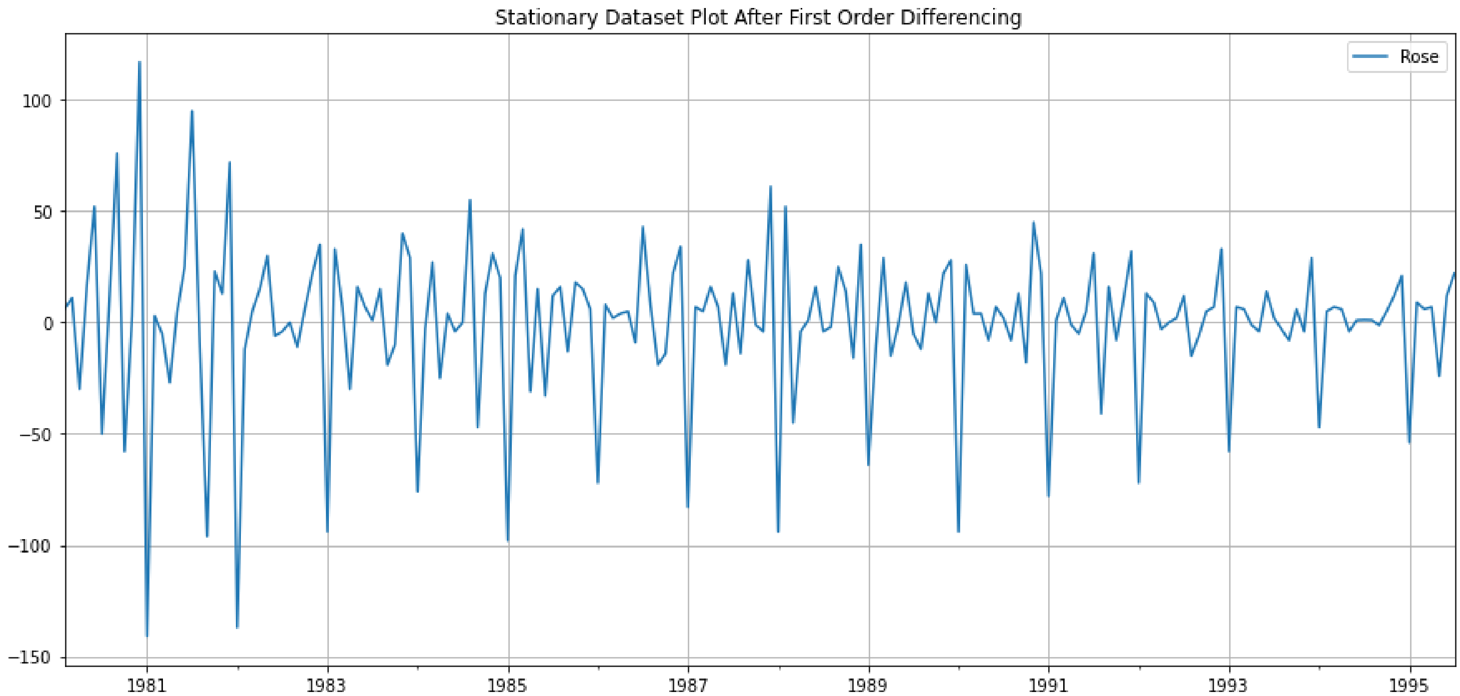


Fig. No. 2.5.1

2.6 : - 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.

For building ARIMA/SARIMA model on train dataset first we have to check stationarity, so after applying dickey-fuller on train dataset parameter we got,

```
Train Dataset test statistic is -1.686
Train Dataset test p-value is 0.7569093051047064
Number of lags used 13
```

The training data is non-stationary at 95% confidence level. Let us take a first level of differencing to stationaries the Time Series.

Apply dickey-fuller after first level differencing on test data,

```
Train Dataset test statistic is -6.804
Train Dataset test p-value is 3.894831356782412e-08
Number of lags used 12
```

Now, let us go ahead and plot the differenced training data.

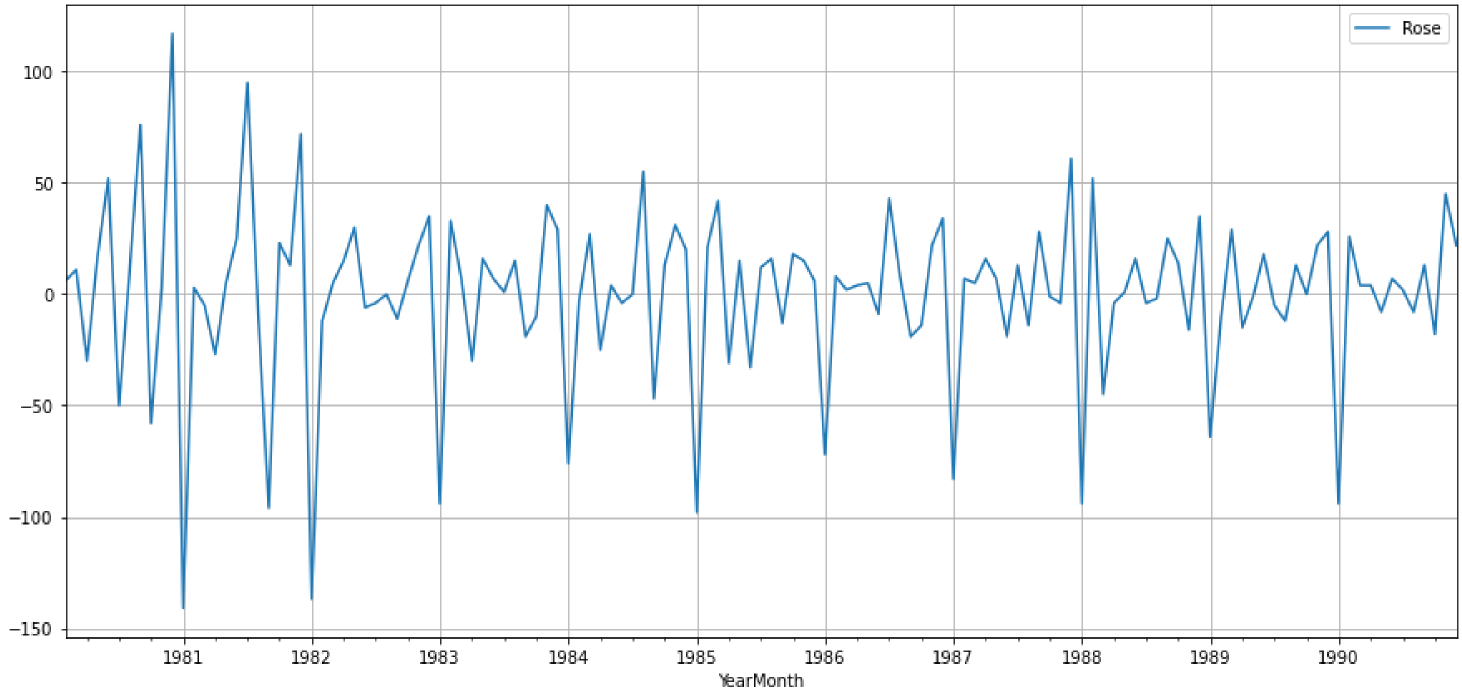


Fig No. 2.6.1

Automated Version of the ARIMA Model

For automated version of ARIMA model we fixed ranges for defined parameter that is p, d, q.

- The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot = range (0,4), i.e., (0,1,2,3)
- The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot = range (0,4), i.e., (0,1,2,3)
- The differencing parameter in an ARIMA model is 'd' which comes from making dataset stationary=range (1,2), i.e., 1.

After applying itertools to make different combination and fetch Akaike Information Criteria (AIC) for train data,

	param	AIC
11	(2, 1, 3)	1274.695412
15	(3, 1, 3)	1278.667917
2	(0, 1, 2)	1279.671529
6	(1, 1, 2)	1279.870723
3	(0, 1, 3)	1280.545376

ARIMA Model with p=2, d=1, q=3 having the lowest AIC value i.e., **1274.6954**
Let's check the summary of the model with these parameters.

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(2, 1, 3)   Log Likelihood                -631.348
Date:                Sun, 13 Nov 2022   AIC                           1274.695
Time:                        18:28:50   BIC                           1291.947
Sample:                    01-01-1980   HQIC                          1281.705
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.6783      0.084    -19.999      0.000      -1.843      -1.514
ar.L2         -0.7291      0.084     -8.687      0.000      -0.894      -0.565
ma.L1          1.0446      0.618      1.691      0.091      -0.166       2.255
ma.L2         -0.7720      0.132     -5.858      0.000      -1.030      -0.514
ma.L3         -0.9045      0.560     -1.616      0.106      -2.002       0.192
sigma2       860.3101    519.823      1.655      0.098    -158.525    1879.145
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                24.51
Prob(Q):                              0.87   Prob(JB):                         0.00
Heteroskedasticity (H):               0.40   Skew:                             0.71
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.57
===================================================================================
```

Fig No. 2.6.2

Diagnostics Plot: -

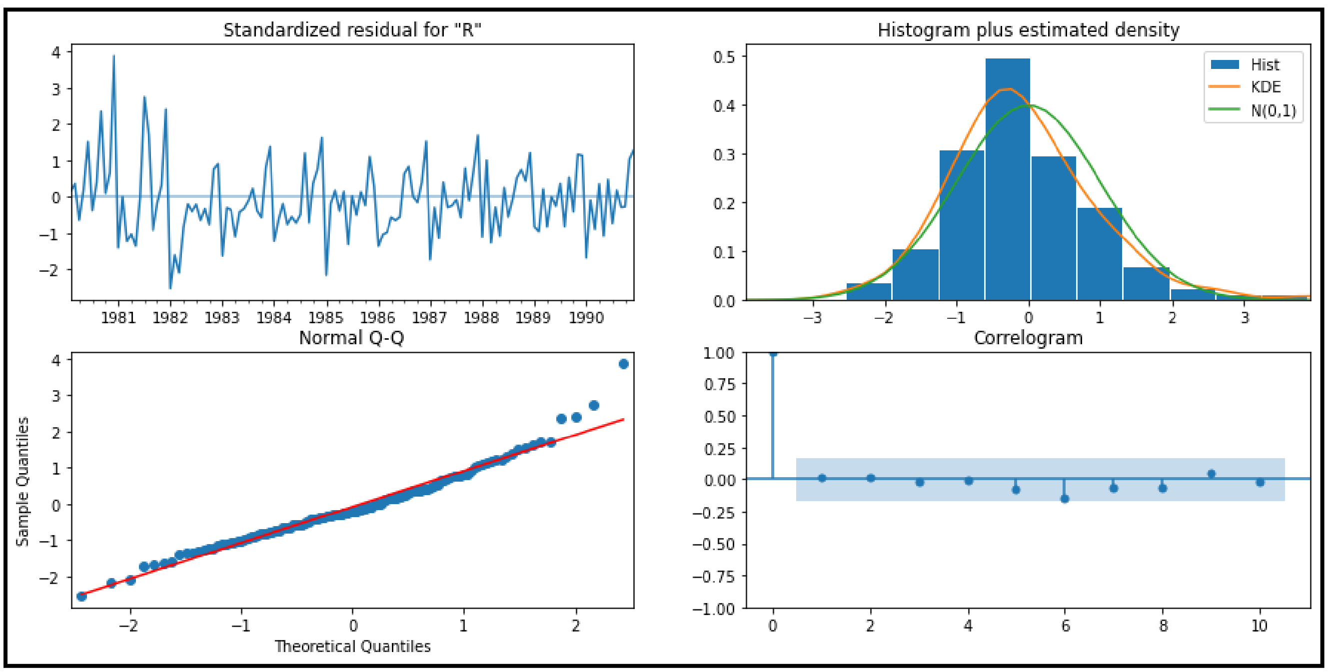


Fig No. 2.6.3

Model Evaluation: -

After getting lowest AIC value of train dataset, we will evaluate our model on test dataset,

	RMSE	MAPE
ARIMA(2,1,3)	36.765327	75.663695

RMSE for automated ARIMA model is 36.765 and MAPE is 75.663

Automated Version of the SARIMA Model

For automated version of SARIMA model, we fixed ranges for defined parameter that is p, d, q and P, D, Q.

- The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot = range (0,3), i.e., (0,1,2)
- The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot = range (0,3), i.e., (0,1,2)
- The differencing parameter in an ARIMA model is ‘d’ which comes from making dataset stationary=range (1,2), i.e., 1.
- The Auto-Regressive parameter in a SARIMA model is 'P' which comes from the significant lag before which the PACF plot = range (0,3), i.e., (0,1,2)
- The Moving-Average parameter in a SARIMA model is 'Q' which comes from the significant lag before the ACF plot = range (0,3), i.e., (0,1,2)
- The differencing parameter in a SARIMA model is ‘D’ which comes from making dataset stationary=range (0,1), i.e., 0.
- With seasonal factor=12

After applying itertools to make different combination and fetch Akaike Information Criteria (AIC) for train data,

	param	seasonal	AIC
26	(0, 1, 2)	(2, 0, 2, 12)	887.937509
53	(1, 1, 2)	(2, 0, 2, 12)	889.902849
80	(2, 1, 2)	(2, 0, 2, 12)	890.668798
69	(2, 1, 1)	(2, 0, 0, 12)	896.518161
78	(2, 1, 2)	(2, 0, 0, 12)	897.346444

SARIMA Model with p=0, d=1, q=2 and P=2, D=0, Q=2 and seasonality of 12 having the lowest AIC value i.e., **887.9375.**

Diagnostics Plot: -

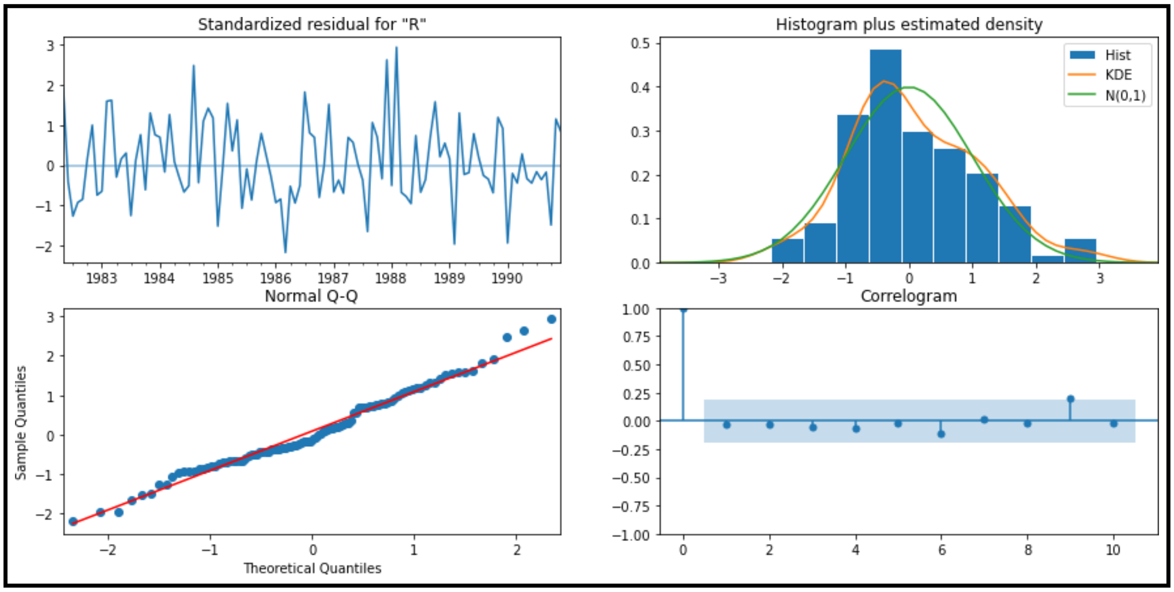


Fig No. 2.6.4

Fig No. 2.6.5

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                  Rose   No. Observations:               132
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 12)   Log Likelihood                -436.969
Date:                            Sun, 13 Nov 2022   AIC                            887.938
Time:                                    18:37:56   BIC                            906.448
Sample:                                01-01-1980   HQIC                           895.437
                                     - 12-01-1990
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.8427    189.892     -0.004      0.996    -373.024     371.339
ma.L2         -0.1573     29.833     -0.005      0.996     -58.629      58.314
ar.S.L12       0.3467      0.079      4.375      0.000       0.191       0.502
ar.S.L24       0.3023      0.076      3.996      0.000       0.154       0.451
ma.S.L12       0.0767      0.133      0.577      0.564      -0.184       0.337
ma.S.L24      -0.0726      0.146     -0.498      0.618      -0.358       0.213
sigma2       251.3137   4.77e+04      0.005      0.996   -9.33e+04    9.38e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.10   Jarque-Bera (JB):                 2.33
Prob(Q):                              0.75   Prob(JB):                         0.31
Heteroskedasticity (H):               0.88   Skew:                             0.37
Prob(H) (two-sided):                  0.70   Kurtosis:                         3.03
===================================================================================
```

Model Evaluation: -

After getting lowest AIC value of train dataset, we will evaluate our model on test dataset,

	RMSE	MAPE
SARIMA(0,1,2)(2,0,2,12)	26.880861	54.7519

RMSE for automated SARIMA model is 26.8808 and MAPE is 54.7519

2.7 : - Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

Fig No. 2.7.4

Fig No. 2.7.6

	Test RMSE
RegressionOnTime	15.255492
NaiveOnTime	79.672475
SimpleAverage	53.413298
2 Point Trailing on Test Data	11.529985
4 Point Trailing on Test Data	14.444375
6 Point Trailing on Test Data	14.554986
9 Point Trailing on Test Data	14.721520
Alpha=0.0987 SimpleExponentialSmoothing	36.748402
Alpha =1.49e-08,Beta=1.66e-10 DoubleExponentialSmoothing	15.255480
Alpha=0.0715,Beta=0.0452,Gamma=7.24e-05 TripleExponentialSmoothing	20.097325

	RMSE	MAPE
ARIMA Auto(2,1,3)	36.765327	75.663695
SARIMA Auto(0,1,2)(2,0,2,12)	26.880861	54.751900
ARIMA Manual(2,1,2)	36.823420	75.880580
SARIMA Manual(2,1,2)(1,0,1,12)	21.493603	43.555189

Table No. 2.8.1

2.8 : - 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.

Based on model building exercise the best model was SARIMA Automated model, now imposing the same parameters on complete dataset.

After imposing the most optimum model on the complete dataset, summary would be look like,

Fig No. 2.9.1

```
                                      SARIMAX Results
============================================================================================
Dep. Variable:                                 Rose   No. Observations:                  187
Model:             SARIMAX(2, 1, 2)x(1, 0, [1], 12)   Log Likelihood                -733.287
Date:                              Sat, 12 Nov 2022   AIC                           1480.573
Time:                                      21:23:49   BIC                           1502.565
Sample:                                  01-01-1980   HQIC                          1489.496
                                       - 07-01-1995
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.1284      0.073     15.435      0.000       0.985       1.272
ar.L2         -0.2660      0.078     -3.418      0.001      -0.418      -0.113
ma.L1         -1.9568   1711.289     -0.001      0.999   -3356.021    3352.108
ma.L2          1.0000   1749.032      0.001      1.000   -3427.040    3429.040
ar.S.L12       0.9192      0.021     44.622      0.000       0.879       0.960
ma.S.L12      -1.0000   1748.978     -0.001      1.000   -3428.934    3426.934
sigma2       245.8682      7.481     32.868      0.000     231.207     260.530
===================================================================================
Ljung-Box (L1) (Q):                   0.11   Jarque-Bera (JB):               278.35
Prob(Q):                              0.74   Prob(JB):                         0.00
Heteroskedasticity (H):               0.14   Skew:                            -0.43
Prob(H) (two-sided):                  0.00   Kurtosis:                         9.19
===================================================================================
```

Diagnostics Plot: -

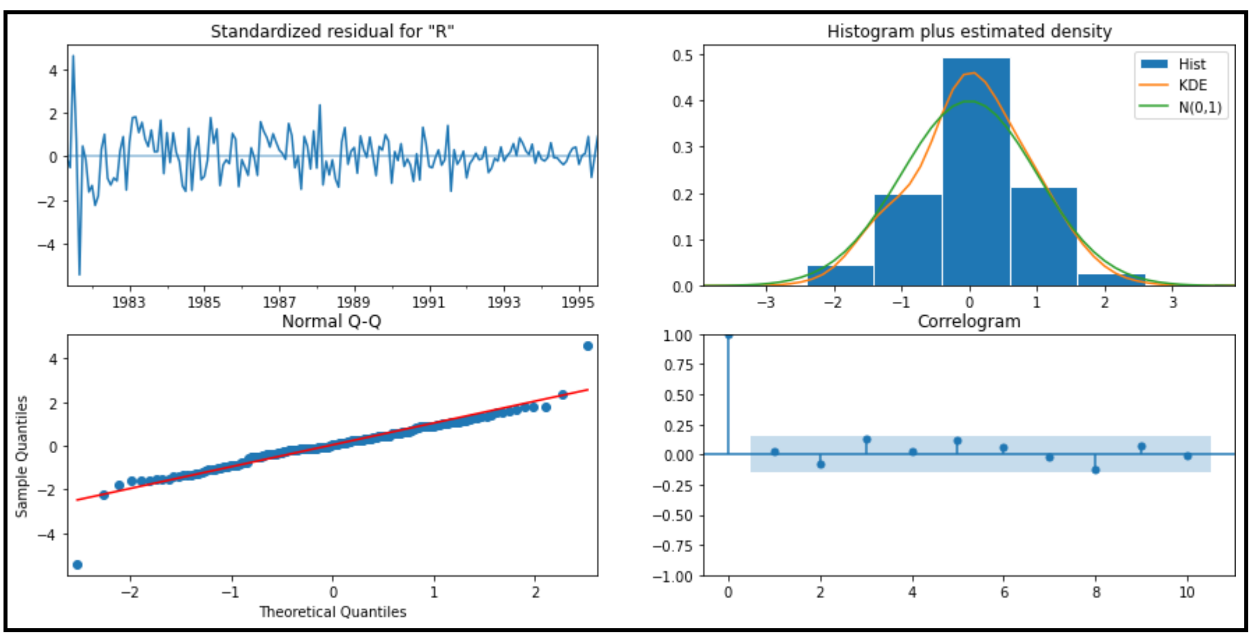


Fig No. 2.9.2

Predicting 12 months into the future with appropriate confidence intervals/bands, and dataset for next forecasted 12 months would be,

Rose	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-01	53.771428	16.275959	21.871135	85.671722
1995-09-01	47.310910	16.541422	14.890317	79.731502
1995-10-01	45.246755	16.544079	12.820957	77.672554
1995-11-01	51.992864	16.549458	19.556522	84.429206
1995-12-01	70.819703	16.552164	38.378058	103.261347
1996-01-01	34.174184	16.592095	1.654276	66.694092
1996-02-01	39.486188	16.678239	6.797440	72.174937
1996-03-01	42.717077	16.872337	9.647904	75.786250
1996-04-01	37.008950	17.147931	3.399623	70.618277
1996-05-01	41.458367	17.496379	7.166094	75.750641
1996-06-01	43.815019	17.905449	8.720984	78.909055
1996-07-01	46.854087	18.361877	10.865470	82.842704

Table No. 2.9.1

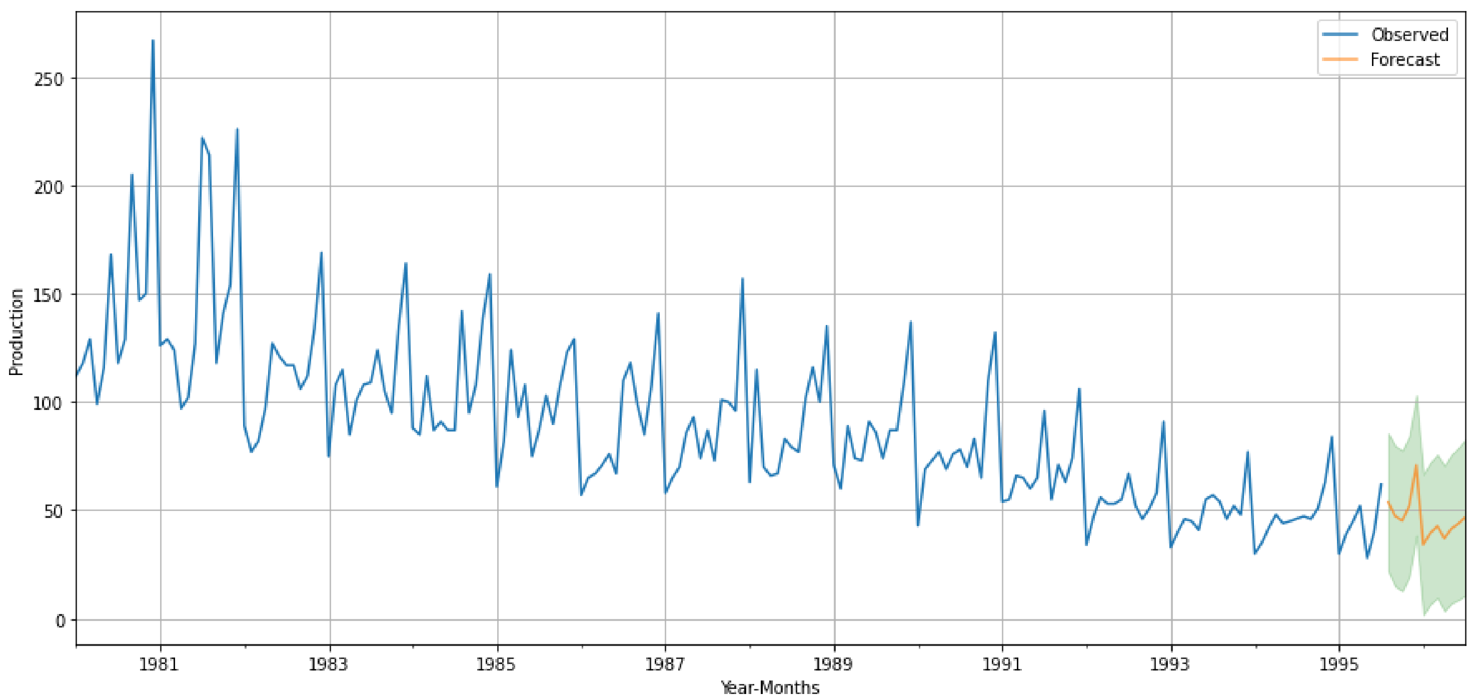


Fig No. 2.9.3

Here is the dataset and plot for next 12 month forecasted values with appropriate interval band.

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

- Year on year rose production is decreasing.
- 2-point trailing is the best forecasting model or this dataset.
- 2^{nd} best forecasting model is 4-point trailing.
- After decomposition we observed there is both seasonality and trend in the dataset.
- KDE plot is almost same for all model.