Problem 2 Time Series Forecasting - Rose

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

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

Data set :

YearMonth	Rose
1980-01	112
1980-02	118
1980-03	129
1980-04	99
1980-05	116
1980-06	168
1980-07	118
1980-08	129

We are provided with the above data set of 187 rows and 02 columns. Of the above columns, one column is object data type and one is integer data type.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 187 entries, 0 to 186
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   YearMonth  187 non-null    object 
 1   Rose       185 non-null    float64
dtypes: float64(1), object(1)
memory usage: 3.0+ KB
```

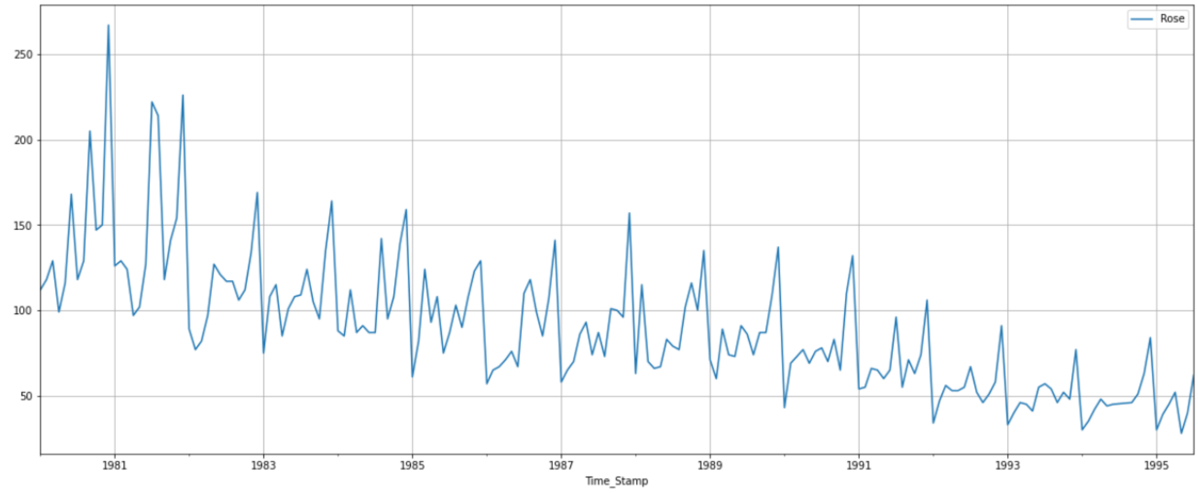
There are **two** Null values in the given dataset.

```
YearMonth    0
Rose         2
dtype: int64
```

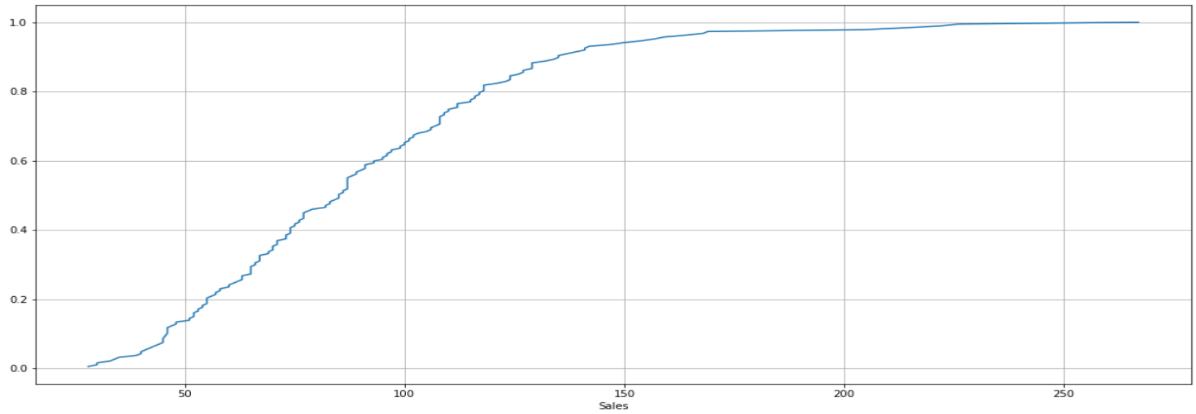
We have read the **YearMonth** column as date type and assign it as index.

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-31 to 1995-07-31
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   Rose    187 non-null    float64
dtypes: float64(1)
memory usage: 2.9 KB
```

By plotting the Time Series to understand the behaviour of the data. We have the following curve



The given data has downward trend and it has seasonality associated with it.

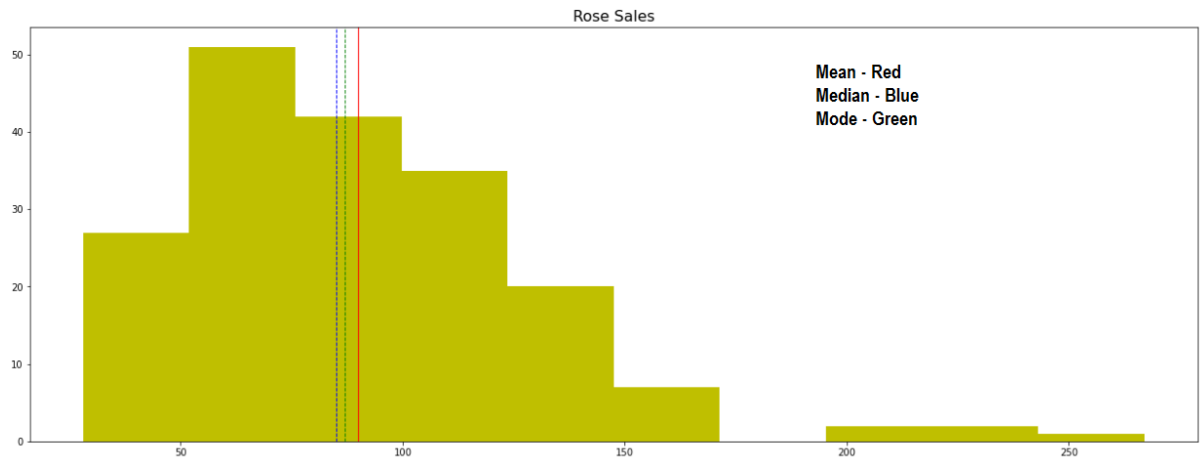


From the above plot , we can see that 60% of the values lie below value 900 and 80% of values lie below 120 respectively.

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

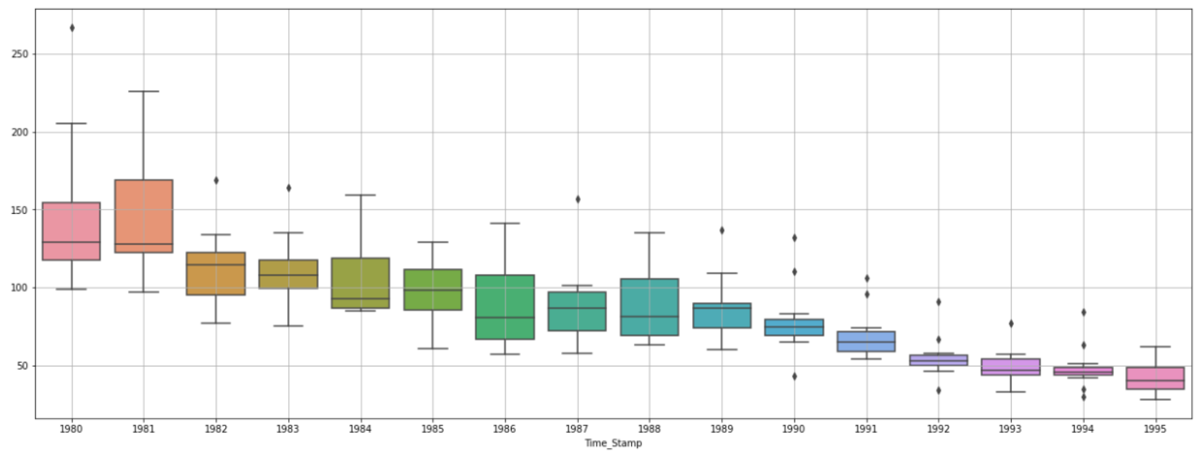
Descriptive statistics of the given time series:

	Rose
count	187.000
mean	89.914
std	39.238
min	28.000
25%	62.500
50%	85.000
75%	111.000
max	267.000



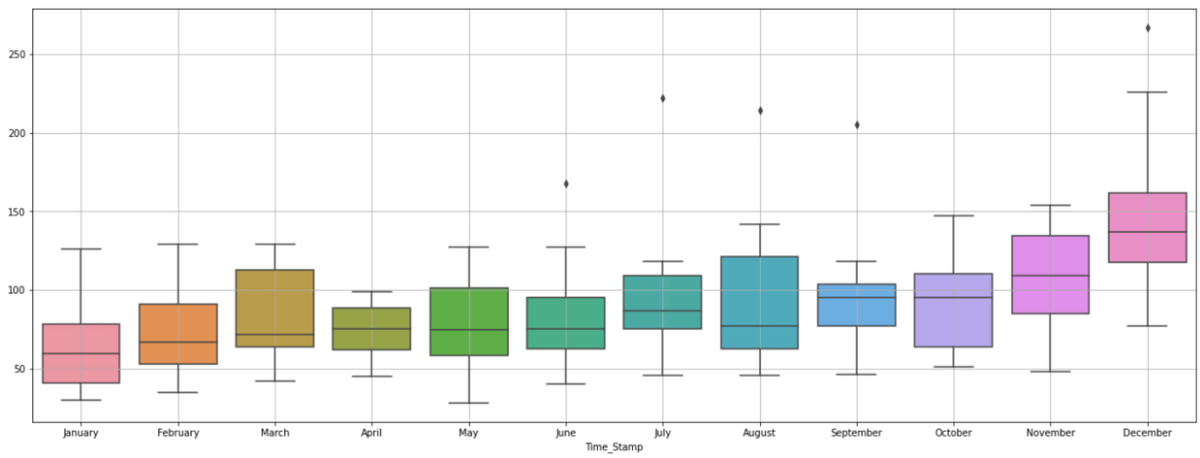
The given data set has mean of value – '89.914 and median value –'85'

Spread of sales across different years:

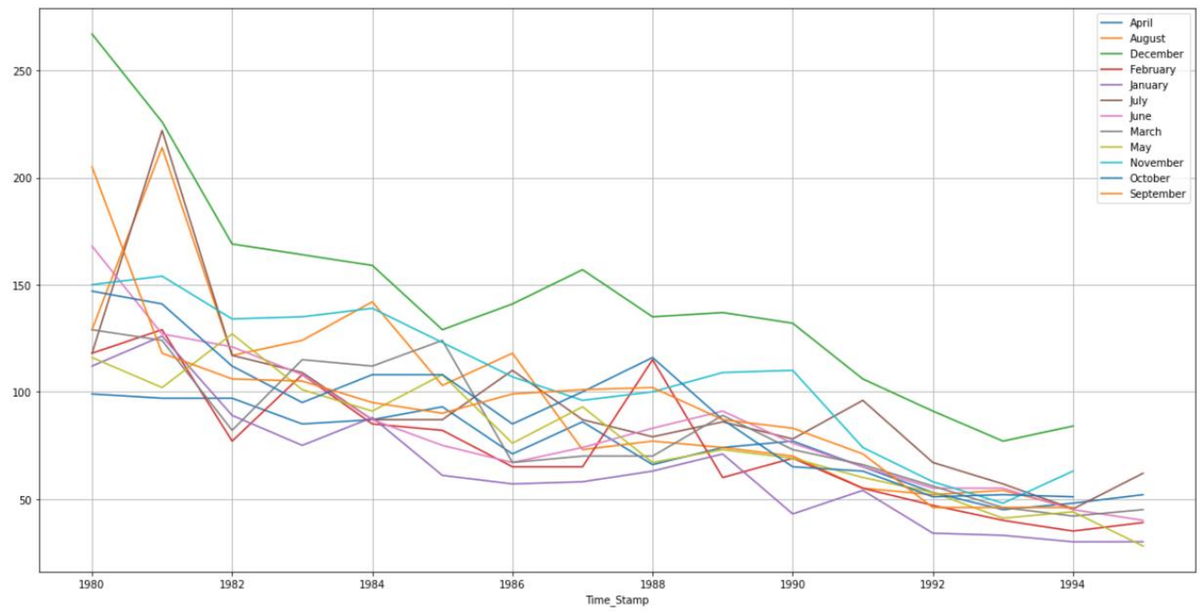


We can see that sales have are decreased from start to last. All most all years are showing outlier values of the data set.

Spread of sales across different months:



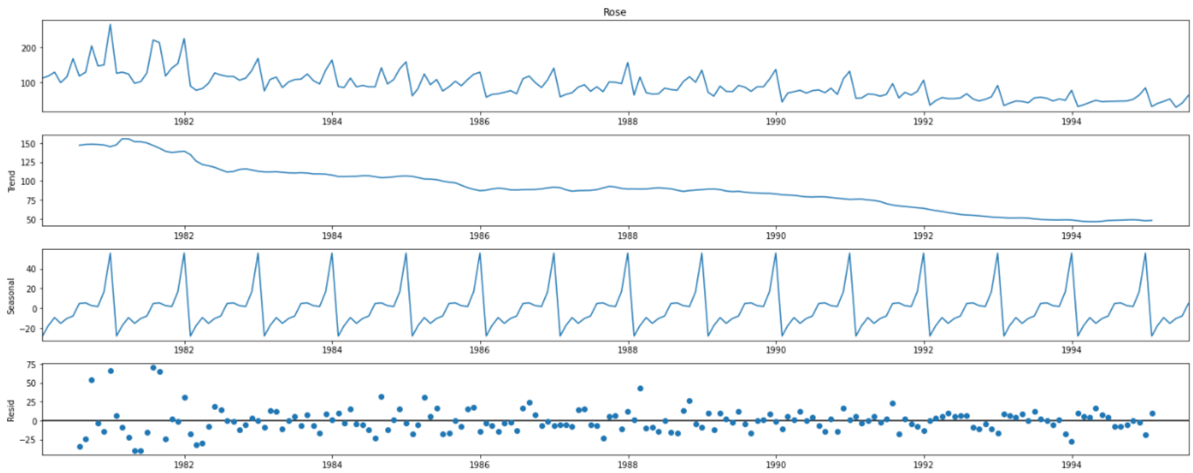
We can understand that **December month** is having the highest sales among all the months.



From above plot also, we can see that December has the highest sales across years.

Decompose the Time Series:

Additive Decomposition –



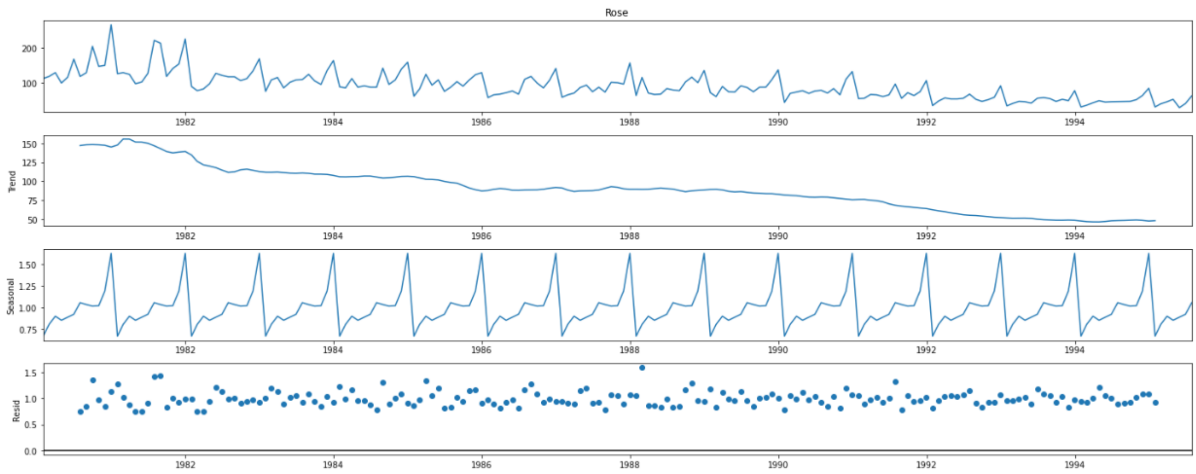

```
Trend
 Time_Stamp
1980-01-31    NaN
1980-02-29    NaN
1980-03-31    NaN
1980-04-30    NaN
1980-05-31    NaN
Name: trend, dtype: float64
```

```
Seasonality
 Time_Stamp
1980-01-31    -27.908647
1980-02-29    -17.435632
1980-03-31     -9.285830
1980-04-30    -15.098330
1980-05-31    -10.196544
Name: seasonal, dtype: float64
```

```
Residual
 Time_Stamp
1980-01-31    NaN
1980-02-29    NaN
1980-03-31    NaN
1980-04-30    NaN
1980-05-31    NaN
Name: resid, dtype: float64
```

As per the 'additive' decomposition, we see that there is a decreased trend from starting to the last. There is a seasonality as well. We see that the residuals are located around 0 from the plot of the residuals in the decomposition.

Multiplicative Decomposition:




```
Trend
 Time_Stamp
1980-01-31    NaN
1980-02-29    NaN
1980-03-31    NaN
1980-04-30    NaN
1980-05-31    NaN
Name: trend, dtype: float64
```

```
Seasonality
 Time_Stamp
1980-01-31    0.670111
1980-02-29    0.806163
1980-03-31    0.901164
1980-04-30    0.854024
1980-05-31    0.889415
Name: seasonal, dtype: float64
```

```
Residual
 Time_Stamp
1980-01-31    NaN
1980-02-29    NaN
1980-03-31    NaN
1980-04-30    NaN
1980-05-31    NaN
Name: resid, dtype: float64
```

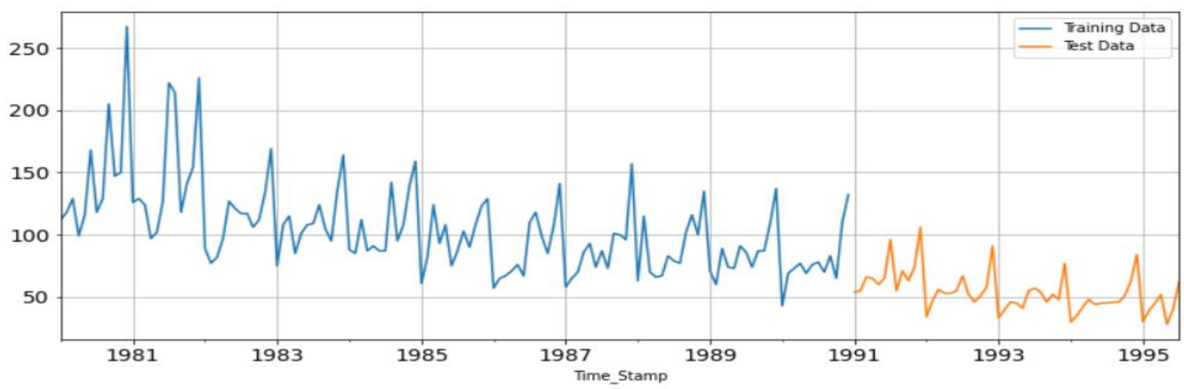
As per the 'Multiplicative' decomposition, we see that there is a decreased trend from starting to the last. There is a seasonality as well. We see that the residuals are located around 1 from the plot of the residuals in the decomposition.

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

```
Shape of Training Data
 (132, 1)

Shape of Testing Data
 (55, 1)
```

```
First few rows of Training Data
                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

Last few rows of Training Data
                Rose
Time_Stamp
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
```

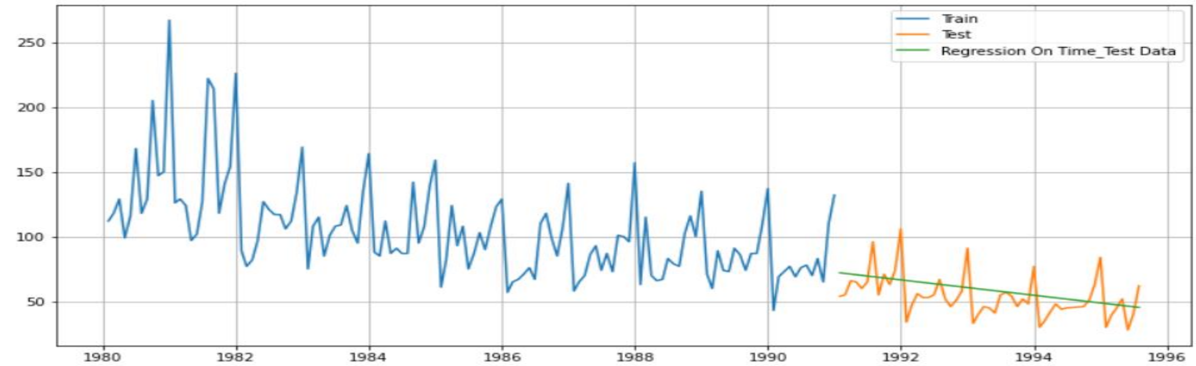
```
First few rows of Test Data
                Rose
Time_Stamp
1991-01-31   54.0
1991-02-28   55.0
1991-03-31   66.0
1991-04-30   65.0
1991-05-31   60.0

Last few rows of Test Data
                Rose
Time_Stamp
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
```



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

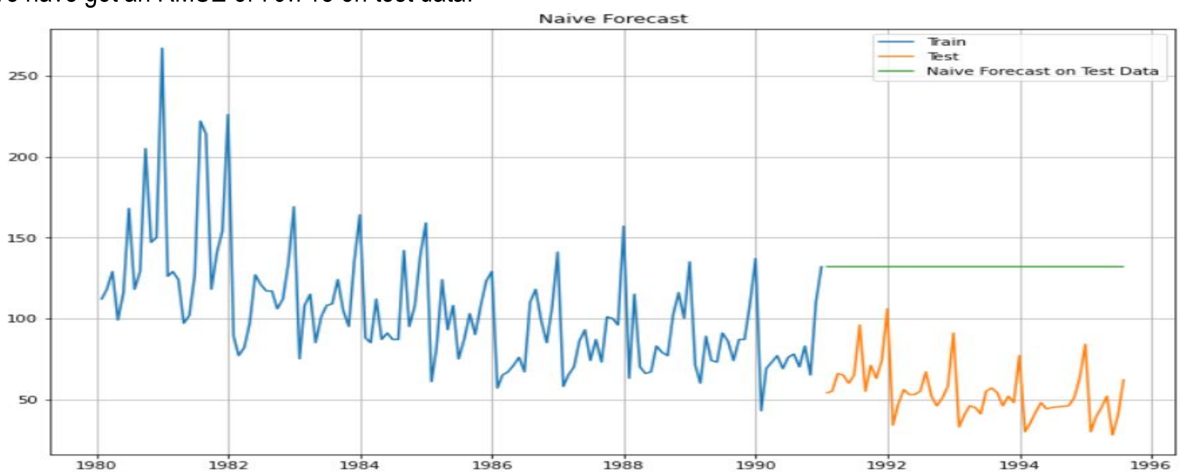
<u>Linear Regression Model :</u>

We have got an RMSE of 15.268 on test data.



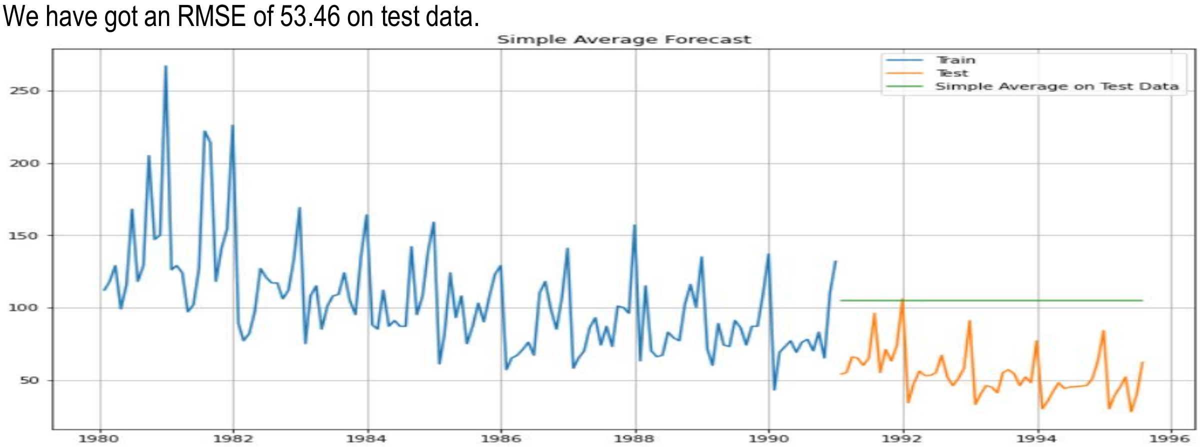
<u>Naive Model:</u>

We have got an RMSE of 79.718 on test data.



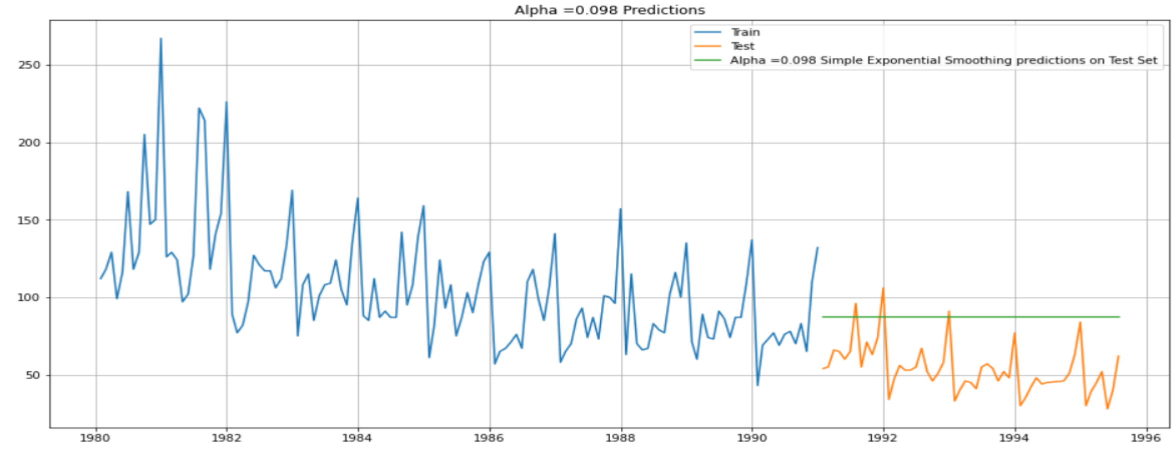
<u>Simple Average Method :</u>

We have got an RMSE of 53.46 on test data.



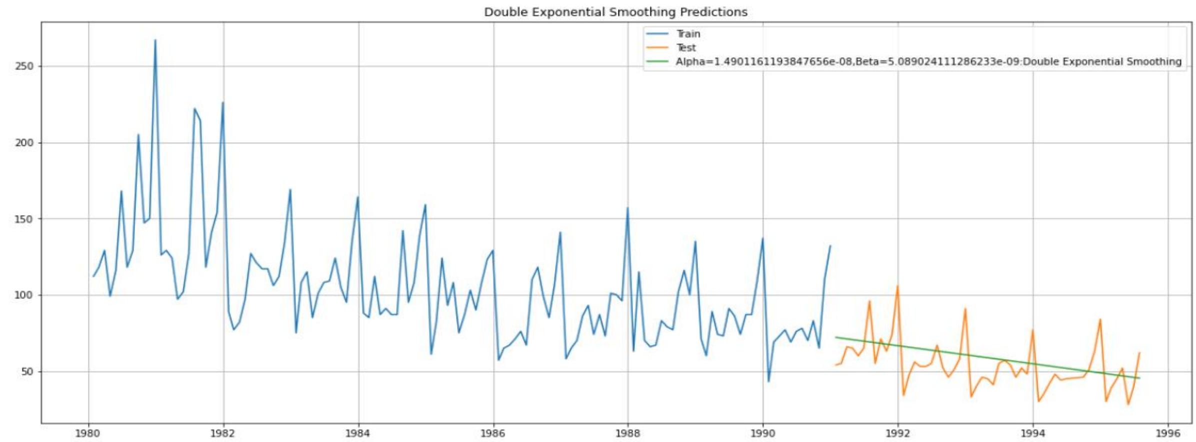
Simple Exponential Smoothing:

We have got an RMSE of 36.796 on test data.



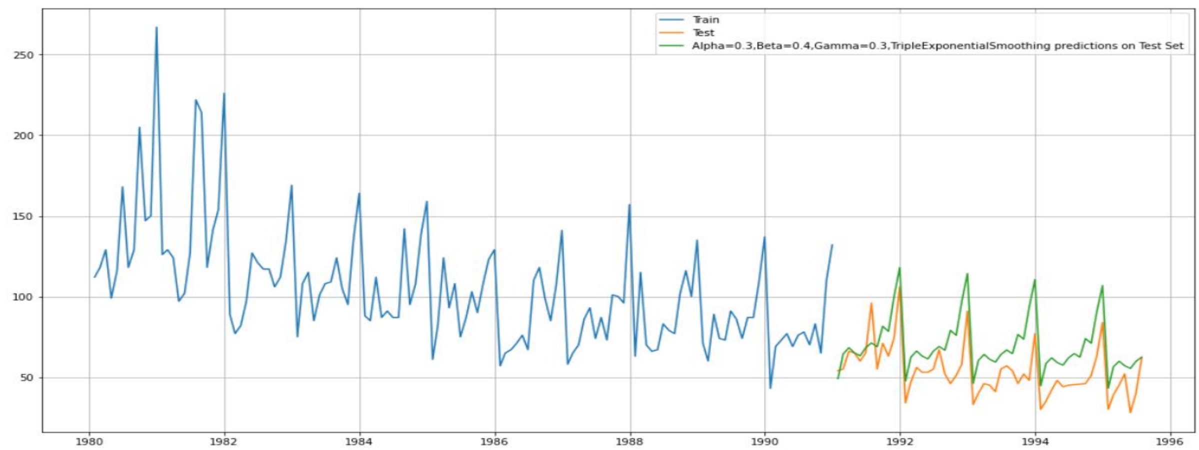
Double Exponential Smoothing :

We have got an RMSE of 15.268 on test data.



Triple Exponential Smoothing:

We have got an RMSE of 10.945 on test data.



Comparing RMSE values for all the above three models , we have got the following table

	Test RMSE
RegressionOnTime	15.268955
NaiveModel	79.718773
SimpleAverageModel	53.460570
Alpha=0.098,SimpleExponentialSmoothing	36.796243
Alpha=1.4901161193847656e-08,Beta=5.089024111286233e-09:Double Exponential Smoothing	15.268954
Alpha=0.075,Beta=0.040,Gamma=0.0004, Triple Exponential Smoothing	19.381887
Alpha=0.3,Beta=0.4,Gamma=0.3,TripleExponentialSmoothing	10.945435

We have built several models got an idea as to which particular model gives us the least error on our test set for this data. As the dataset has both trend and seasonality , Triple Exponential Smoothing works best with this model among all the above models.

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.

Augmented Dickey –Fuller test is used to test whether a time is non-stationary.

Null hypothesis Ho : Time series is non stationary
Alternative hypothesis Ha : Time series is stationary.
Rejection of null hypothesis implies that the series is stationary.

For the dataset, we have following results :

```
Results of Dickey-Fuller Test:
Test Statistic                  -1.876699
p-value                          0.343101
#Lags Used                      13.000000
Number of Observations Used    173.000000
Critical Value (1%)             -3.468726
Critical Value (5%)             -2.878396
Critical Value (10%)            -2.575756
dtype: float64
```

As the p-value is greater than 0.05 , we fail to reject the null hypothesis. So the time series is non stationary. Let us take a difference of order 1 and check whether the Time Series is stationary or not.

```
Results of Dickey-Fuller Test:
Test Statistic                -8.044392e+00
p-value                        1.810895e-12
#Lags Used                     1.200000e+01
Number of Observations Used    1.730000e+02
Critical Value (1%)           -3.468726e+00
Critical Value (5%)           -2.878396e+00
Critical Value (10%)          -2.575756e+00
dtype: float64
```

We see that after the difference of order 1 , the time series is stationary.

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.

As the data shows seasonality , we use SARIMA model on the training data.

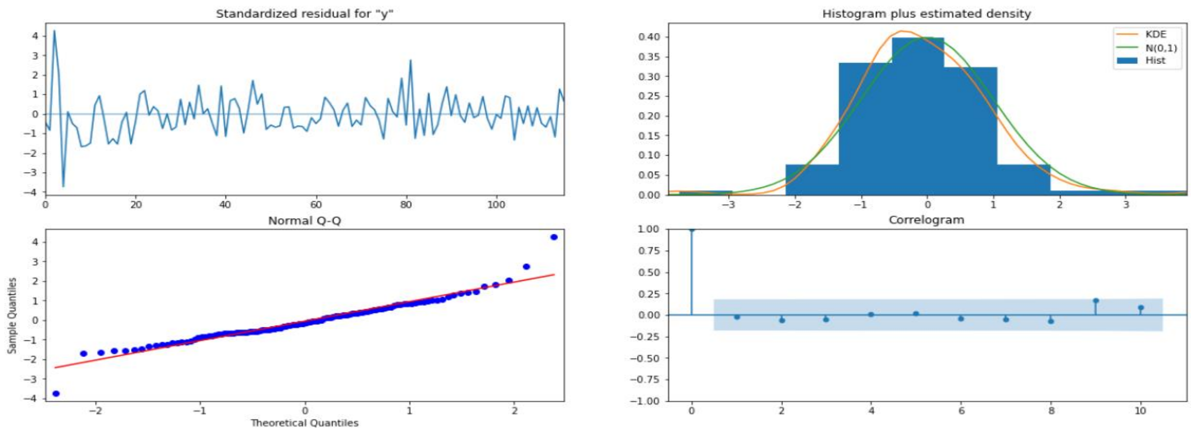
Seasonality as 6 for the model , we have got lowest AIC for on the training data for the model with paramaters

param	seasonal	AIC
(1, 1, 2)	(2, 0, 2, 6)	1041.655818
(0, 1, 2)	(2, 0, 2, 6)	1043.600261
(2, 1, 2)	(2, 0, 2, 6)	1045.220389
(2, 1, 1)	(2, 0, 2, 6)	1051.673461
(1, 1, 1)	(2, 0, 2, 6)	1052.778470

SARIMAX Results:

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                               y   No. Observations:                    132
Model:             SARIMAX(1, 1, 2)x(2, 0, 2, 6)   Log Likelihood                 -512.828
Date:                         Sun, 20 Dec 2020   AIC                            1041.656
Time:                                 20:41:36   BIC                            1063.685
Sample:                                      0   HQIC                           1050.598
                                         - 132
Covariance Type:                           opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5939      0.152     -3.914      0.000      -0.891      -0.296
ma.L1         -0.1954    188.566     -0.001      0.999    -369.777     369.387
ma.L2         -0.8046    151.765     -0.005      0.996    -298.258     296.649
ar.S.L6       -0.0625      0.035     -1.794      0.073      -0.131       0.006
ar.S.L12       0.8451      0.039     21.889      0.000       0.769       0.921
ma.S.L6        0.2226    188.635      0.001      0.999    -369.495     369.940
ma.S.L12      -0.7774    146.598     -0.005      0.996    -288.104     286.549
sigma2       335.1965      0.906    369.902      0.000     333.420     336.973
===================================================================================
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
===================================================================================
```

Plot Diagnostics:



We have an RMSE of value **26.134** on test data

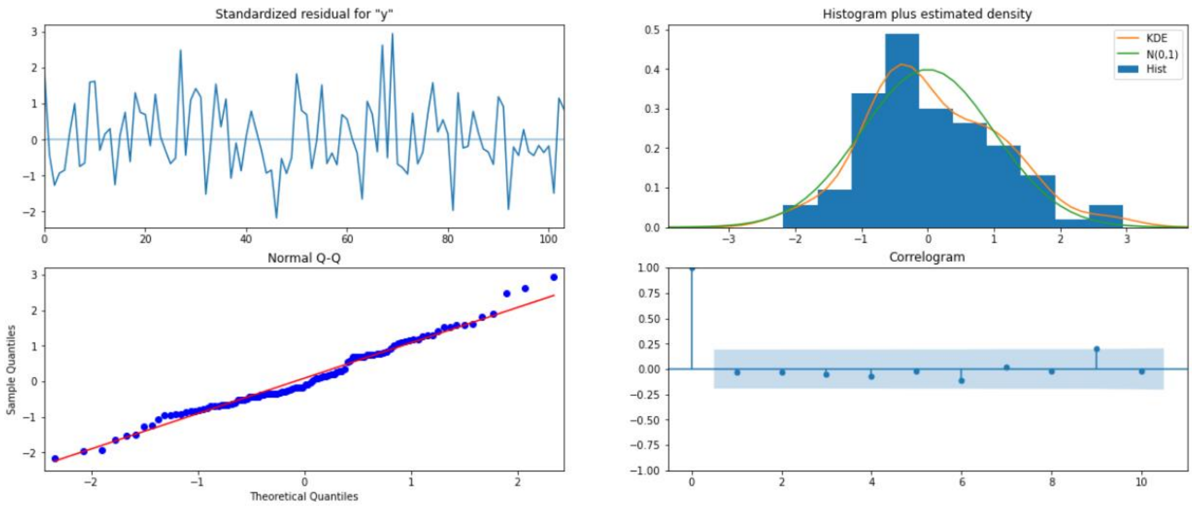
Seasonality as 12 for the model , we have got lowest AIC for on the training data for the model with paramaters

param	seasonal	AIC
(0, 1, 2)	(2, 0, 2, 12)	887.937509
(0, 1, 2)	(2, 0, 2, 12)	887.937509
(2, 1, 2)	(2, 0, 2, 12)	890.668798
(2, 1, 2)	(2, 0, 2, 12)	890.668798
(2, 1, 1)	(2, 0, 0, 12)	896.518161

SARIMAX Results:

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                 132
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 12)   Log Likelihood                -436.969
Date:                            Sun, 20 Dec 2020   AIC                            887.938
Time:                                    20:51:04   BIC                            906.448
Sample:                                         0   HQIC                           895.437
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.8427    189.512     -0.004      0.996    -372.279     370.593
ma.L2         -0.1573     29.773     -0.005      0.996     -58.512      58.197
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       251.3136   4.76e+04      0.005      0.996   -9.31e+04    9.36e+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
===================================================================================
```

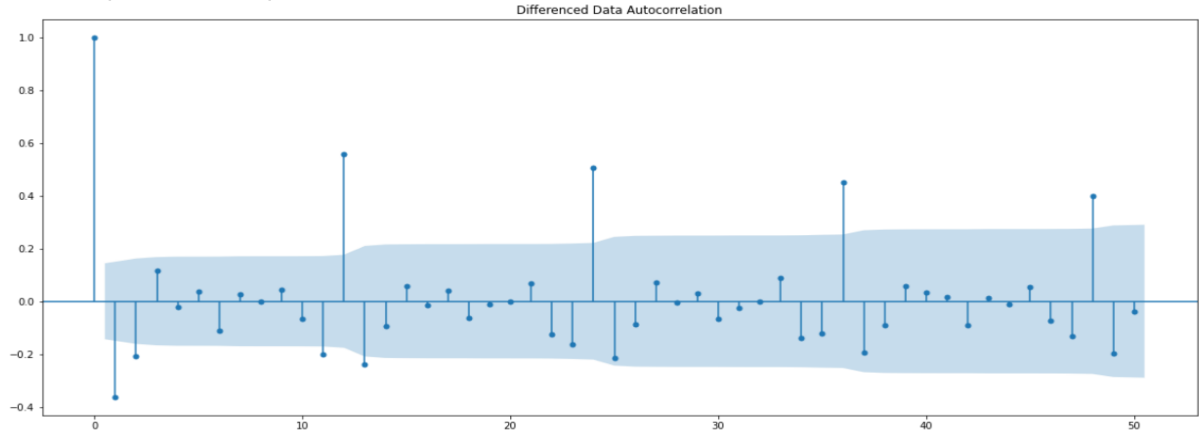
Plot Diagnostics:

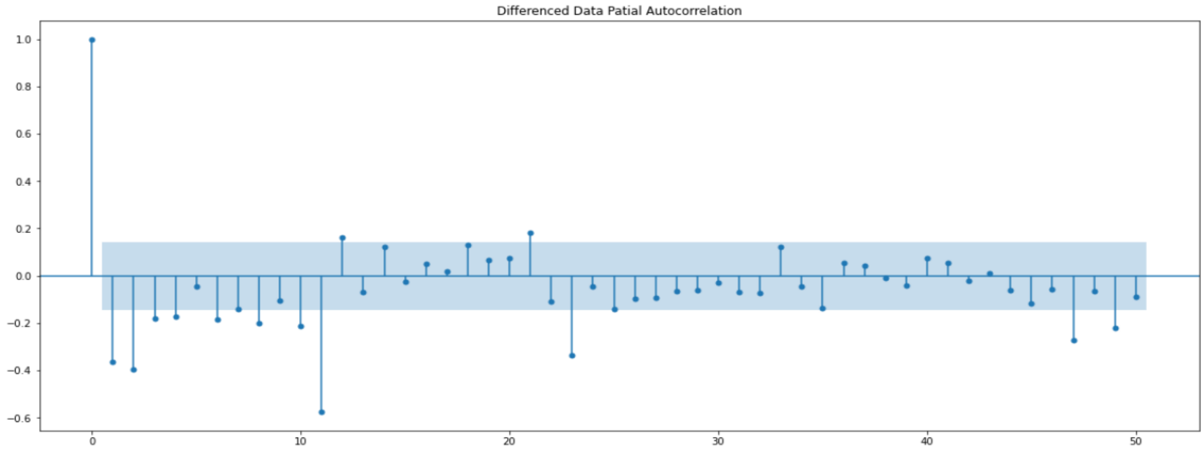


We have an RMSE of value **26.928** on the test data

2.7. 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.

We have got the following ACF and PACF plots

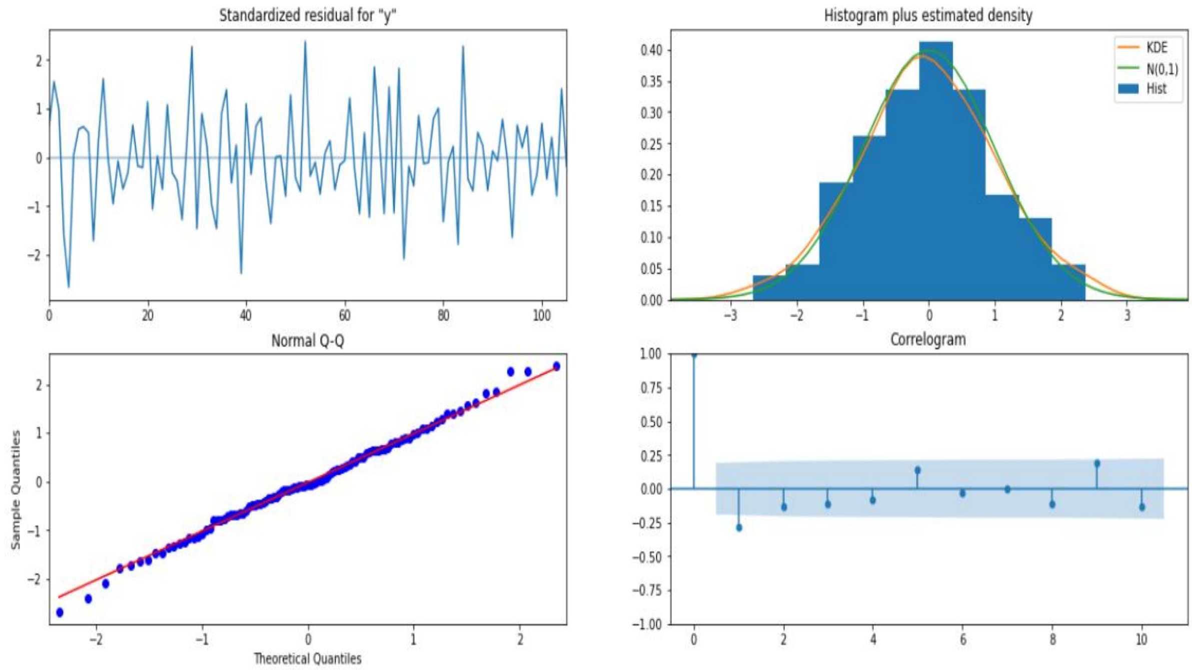




SARIMAX results:

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                 132
Model:             SARIMAX(0, 1, 0)x(1, 1, [1, 2, 3], 6)   Log Likelihood             -478.459
Date:                             Sun, 20 Dec 2020   AIC                            966.918
Time:                                     20:59:42   BIC                            980.235
Sample:                                          0   HQIC                           972.315
                                             - 132
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.S.L6       -0.8507      0.039    -22.083      0.000      -0.926      -0.775
ma.S.L6       -0.2404      0.119     -2.024      0.043      -0.473      -0.008
ma.S.L12      -0.5019      0.127     -3.961      0.000      -0.750      -0.254
ma.S.L18      -0.1041      0.104     -0.998      0.318      -0.308       0.100
sigma2       464.4001     69.420      6.690      0.000     328.339     600.461
===================================================================================
Ljung-Box (L1) (Q):                   8.65   Jarque-Bera (JB):                 0.00
Prob(Q):                              0.00   Prob(JB):                         1.00
Heteroskedasticity (H):               0.77   Skew:                            -0.01
Prob(H) (two-sided):                  0.45   Kurtosis:                         2.97
===================================================================================
```

Plot Diagnostics:



We have got an RMSE value 37.874 on the test data

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

We can summarize the results of all the different models through the following table:

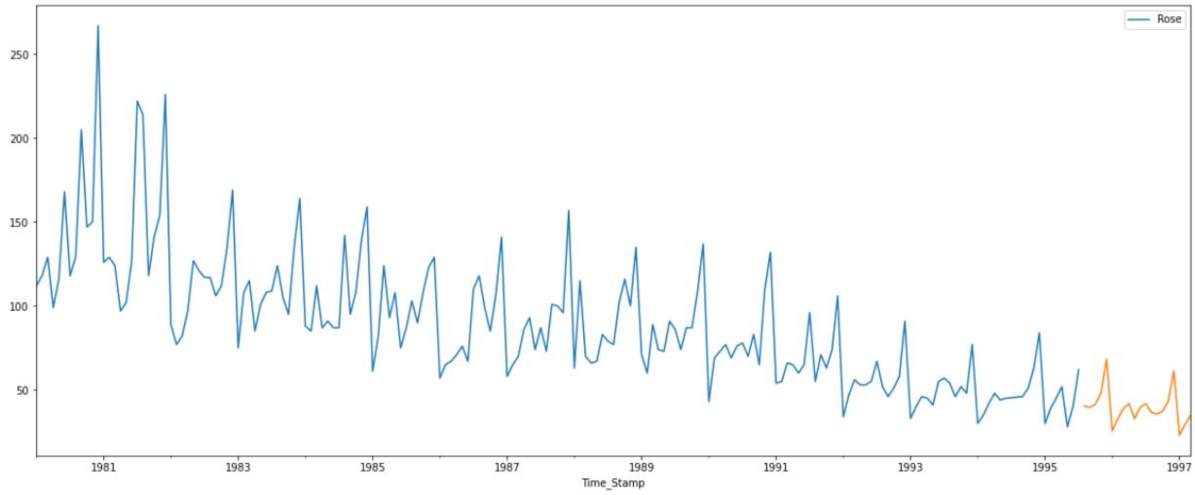
	Test RMSE
RegressionOnTime	15.268955
NaiveModel	79.718773
SimpleAverageModel	53.460570
Alpha=0.098,SimpleExponentialSmoothing	36.796243
Alpha=1.4901161193847656e-08,Beta=5.089024111286233e-09:Double Exponential Smoothing	15.268954
Alpha=0.075,Beta=0.040,Gamma=0.0004, Triple Exponential Smoothing	19.381887
Alpha=0.3,Beta=0.4,Gamma=0.3,TripleExponentialSmoothing	10.945435
SARIMA(1,1,2)(2,0,2,6)	26.134254
SARIMA(0,1,2)(2,0,2,12)	26.928361
SARIMA(0,1,0)(1,1,3,6)	37.874033

From above table, we can see that Triple Exponential Smoothing with Alpha = 0.3,Beta = 0.4 and Gamma = 0.3 has the lowest Test RMSE of value 10.945

2.9. 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.

As the Triple Exponential Smoothing with Alpha = 0.3,Beta = 0.4 and Gamma = 0.3 has the lowest Test RMSE of value 10.945, we use this model to prediction.

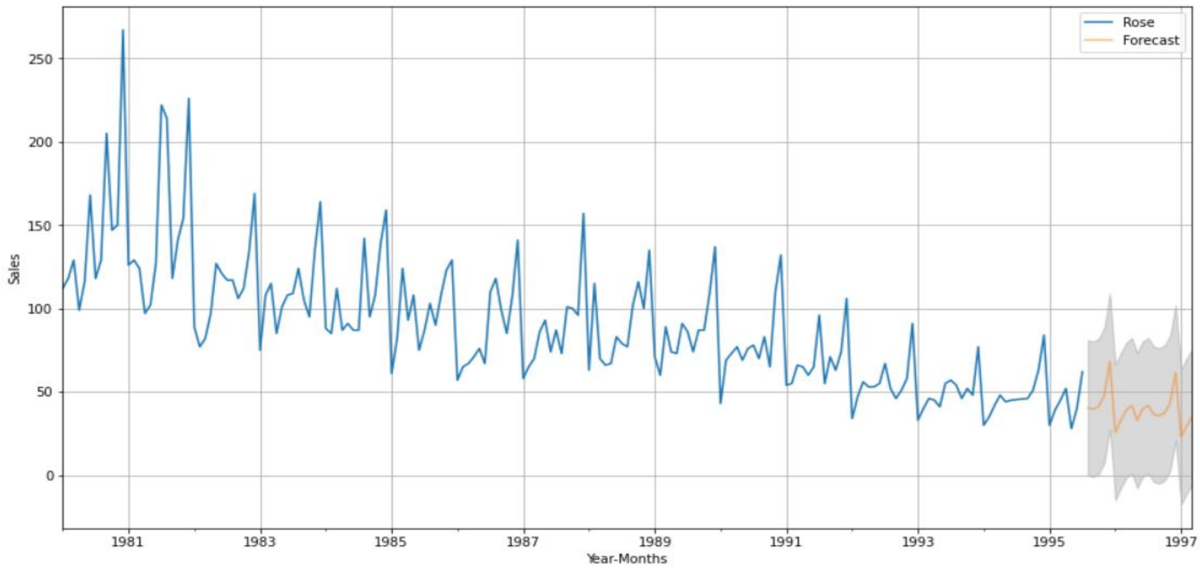
This model gives RMSE of **20.672** on the full data.



Confidence bands for prediction:

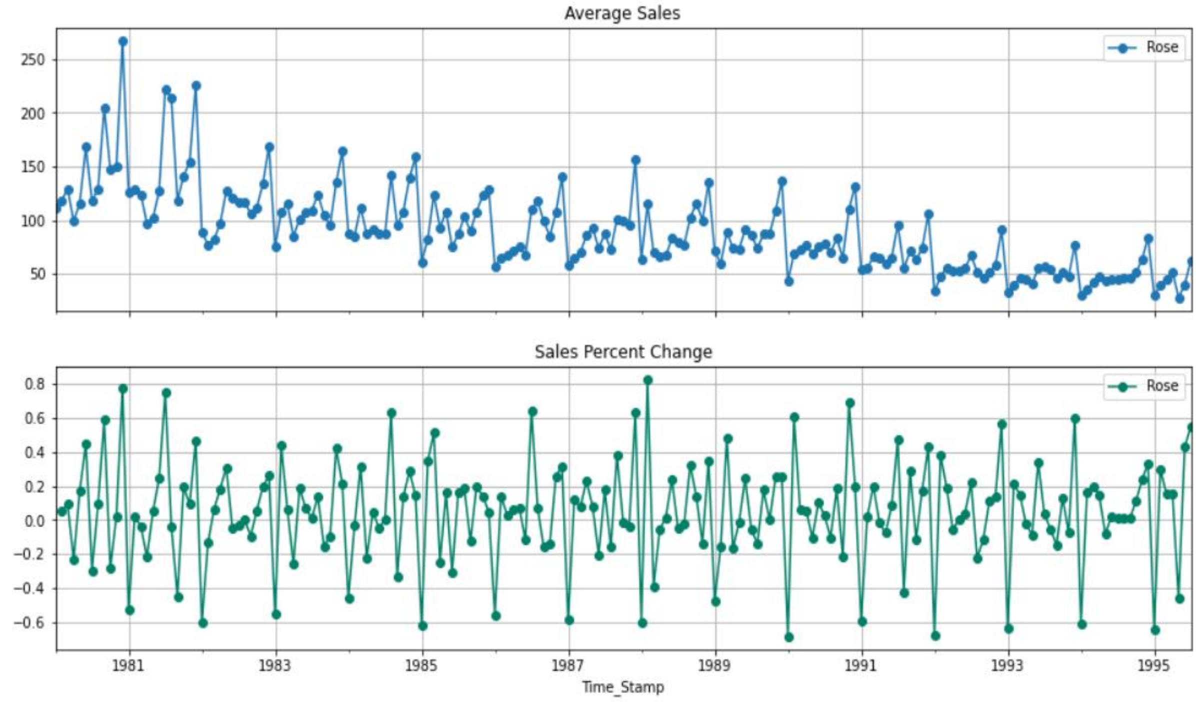
lower_CI	prediction	upper_ci
-0.145493	40.466297	81.078087
-1.088642	39.523148	80.134938
0.860742	41.472532	82.084323
7.399766	48.011557	88.623347
27.672910	68.284701	108.896491

Forecast along with the confidence band :

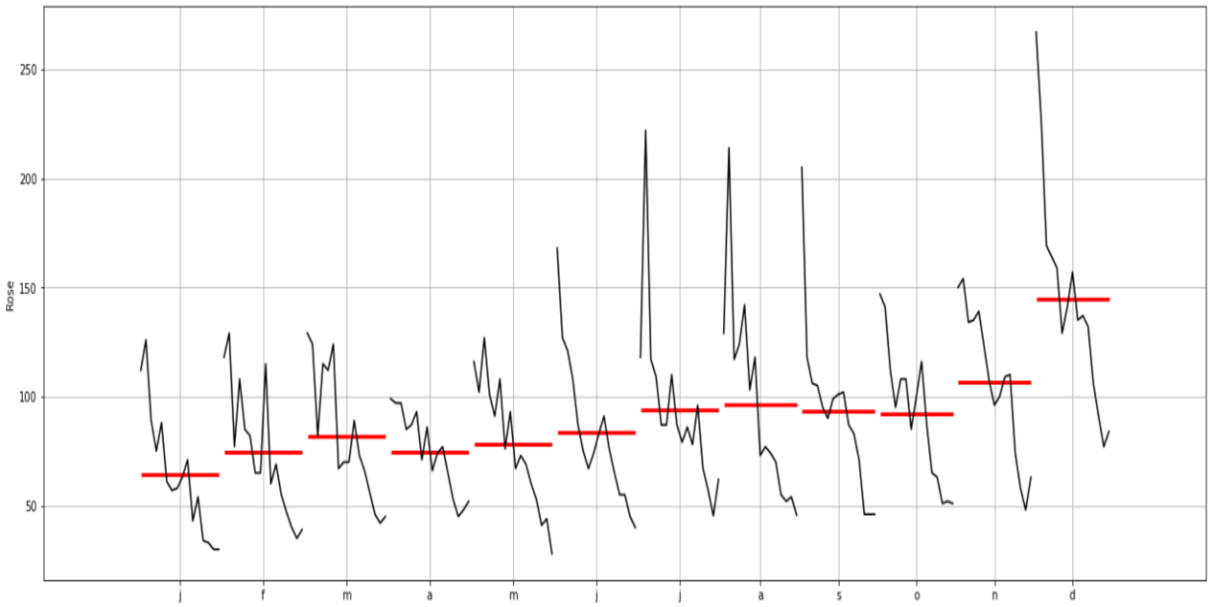


2.10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

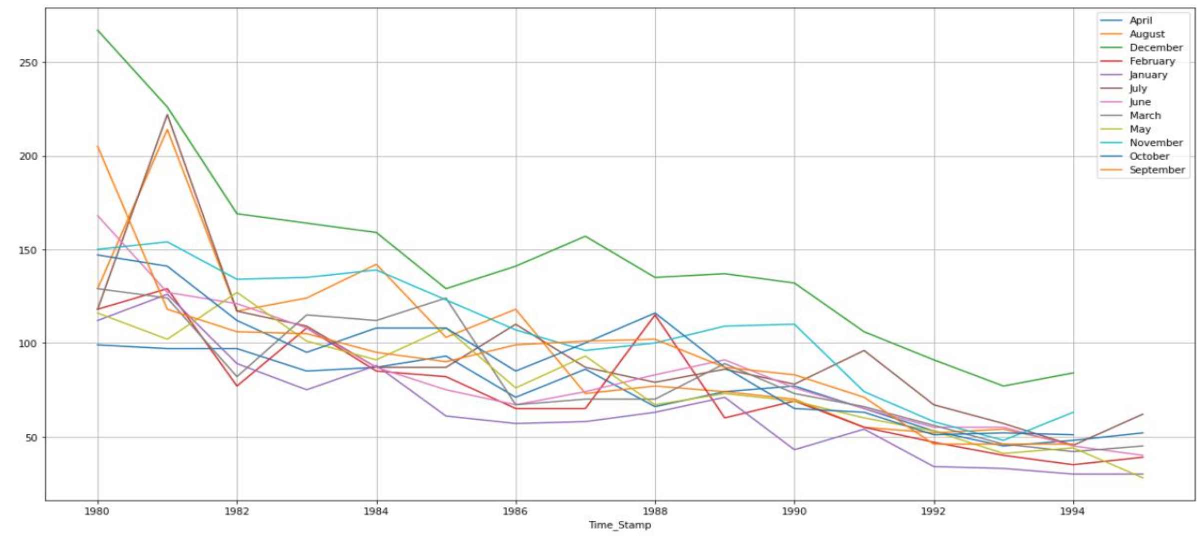
The Triple Exponential Smoothing model with parameters Alpha = 0.3,Beta = 0.4 and Gamma = 0.3 will be helpful in making best forecasts for the given time series data.



The average sales are decreasing year by year from 1981 to 1995 and also sales percent change is higher at the start and end of year.



Looking at sales for different months across all the years, there is minimal change in sales in april month and maximum change in july, august and December months. The company can use this data in maintain stocks for the respective periods based on the demand.



From the plot we can see that sales are higher in December month for all the years.

This could be because off festival events such as Christmas & New year. So the company can increase sales in this month by increasing sales qtys and providing any offers to attract more wine consumers.

Therefore from the above forecasting values based on the trend and seasonality of the time series data, the company can make best decisions in increasing it sales.