1.Sparkling wine

Problem-

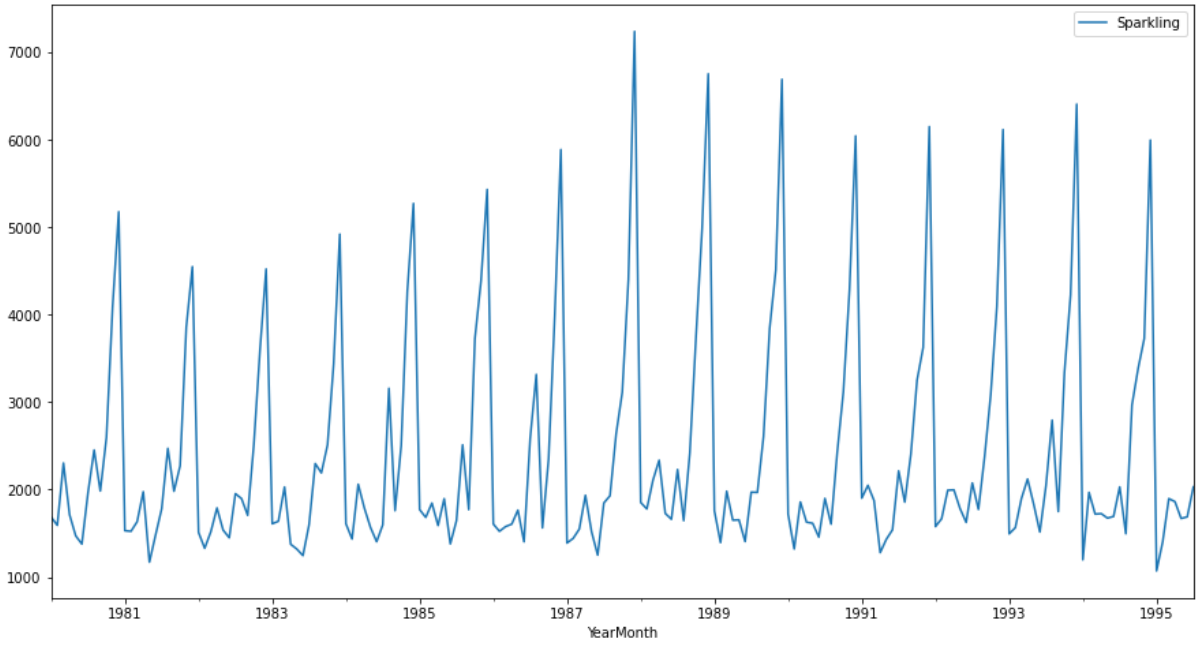
As an analyst in the ABC Estate Wines, we have to analyse and forecast Sparkling Wine Sales in the 20th century.

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

Solution- Data has been read as a time series data after converting the 'yearmonth' column from object to datetime64 by using parse_dates argument.

- The 'yearmonth' has been set as an index.

Plot-



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

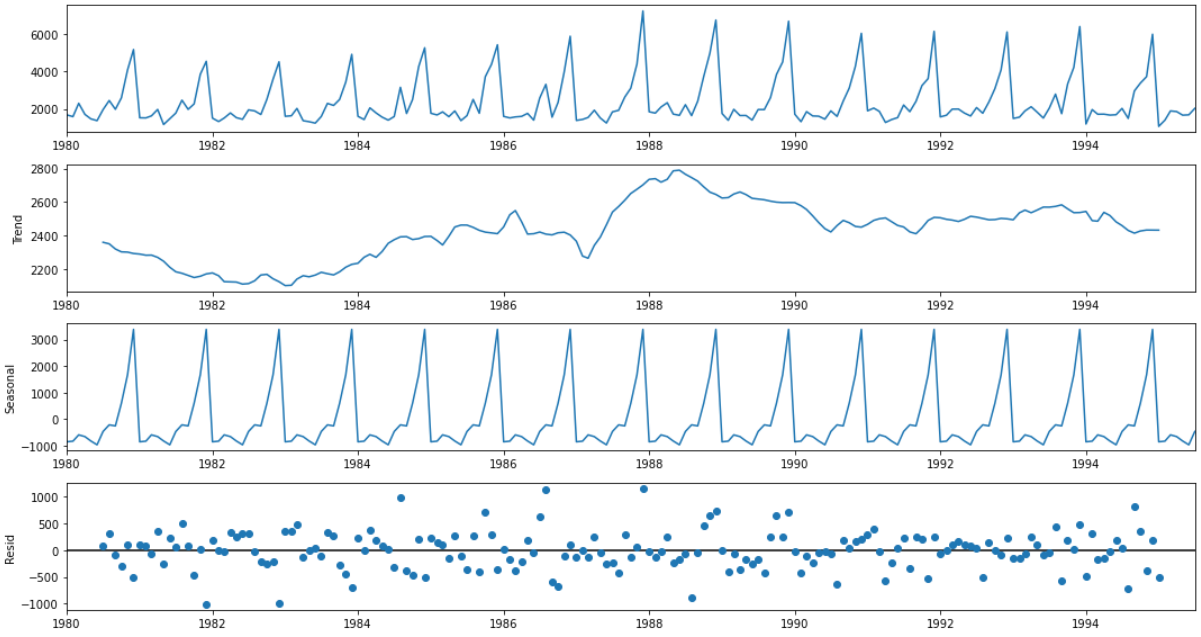
Solution - On looking at the plot we can make out that the data is having a seasonality but no trend. For better visualization and look more in detail we are going to decompose data into 3 components-
i) Trend
ii) Seasonal
iii) Error or Irregular component or random component

Trend and Seasonal are systemic components and take part in the forecasting while irregular components do not take part.

There are 2 methods of decomposition:-

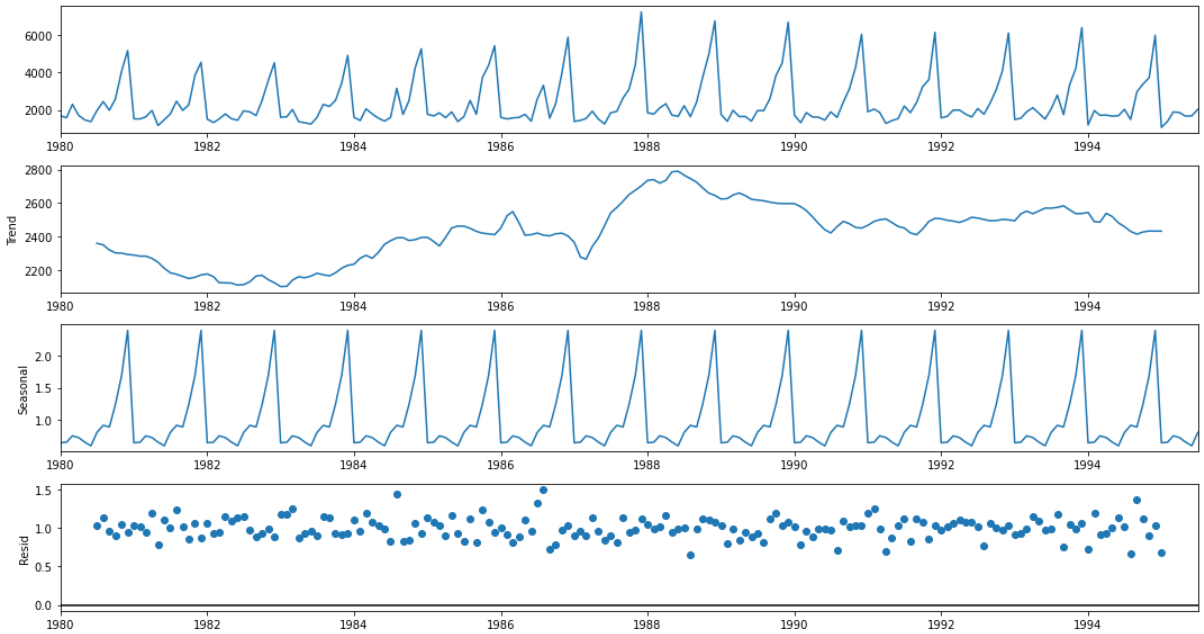
1) Additive method
2) Multiplicative method

1) Additive method-
   Observation = Trend + Seasonality + Error



Here we can see that there is no trend but there is seasonality and error is showing some kind of pattern i.e this decomposition has not taken out all the information from data.

2) Multiplicative method-
   Observation = Trend * Seasonality * Error



Here there is seasonality and no trend and error is showing no pattern i.e all the information are extracted from the data.

- As error in the Additive method is showing some kind of pattern and error in the Multiplicative method is showing no pattern, decomposition done by the multiplicative method is better than additive method.

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

Solution- IN time series, most recent datas are kept as test set and older datas are kept as training set.Here data collected after 1991 are kept as test data.

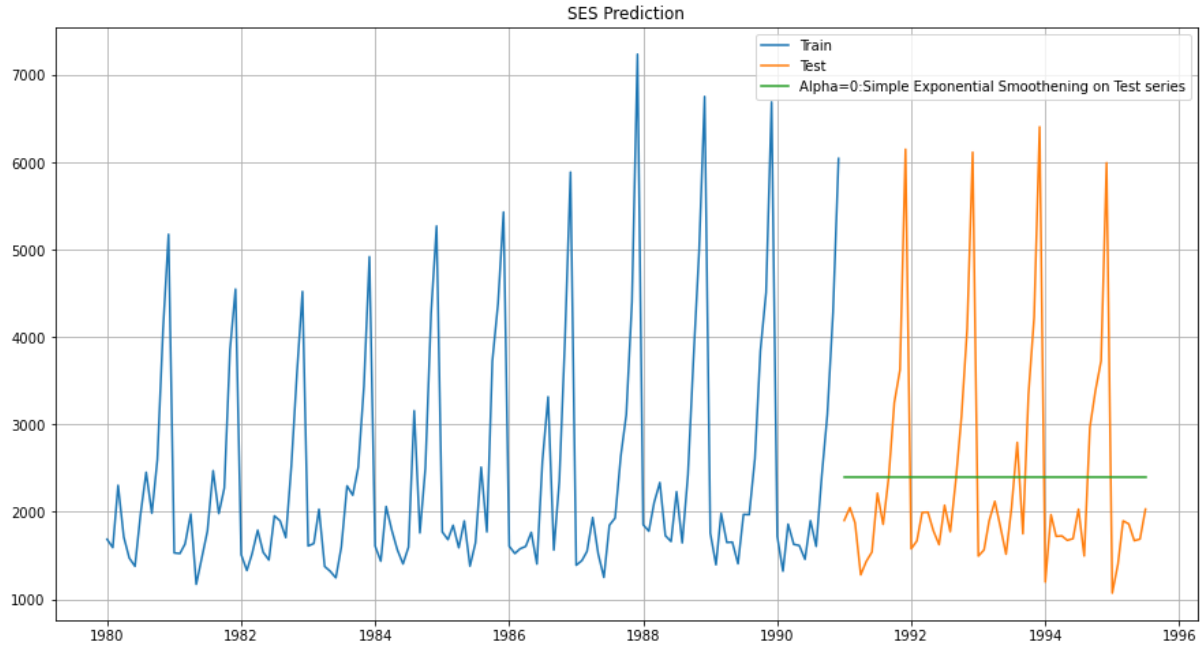
Question 4 - Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression,naïve forecast models and simple average models. should also be built on the training data and check the performance on the test data using RMSE.

Solution- Exponential smoothing methods- This method consists of flattening time series data. Exponential smoothing averages or exponentially weighted moving averages consist of forecasts based on previous periods data with exponentially declining influence on the older observations. It consists of parameters error, trend and seasonality. One or more parameters control how fast the weights decay. These parameters have values between 0 and 1.

1)Simple exponential smoothing- This method is suitable for forecasting data with no clear trend and seasonal pattern. The forecasting i single exponential smoothing is given by:

$$F_{t+1}=\alpha Y_t+(1-\alpha)F_t$$

Parameter α is the smoothing constant and its value lies between 0 and 1. Since the model uses only one constant, it is called single exponential smoothing.



Green line shows forecasting done using single exponential smoothing with α=0

RMSE- Root mean squared error

RMSE of Simple exponential smoothing- 1275.08

2) Double exponential smoothing- Also known as HOLT Exponential smoothing.This model estimates 2 smoothing parameters. This model is applicable when data has trends but no seasonality.

- In this 2 separate components are considered : level and trend. Level is the local mean.
- Double exponential smoothing uses 2 equations for forecasting, one for forecasting the level and other for forecasting the trend.
- Level equation-

  $$L_t=\alpha Y_t+(1-\alpha)F_t$$

  α= Smoothing constant for level, values lie between 0 and 1.

- Trend equation-

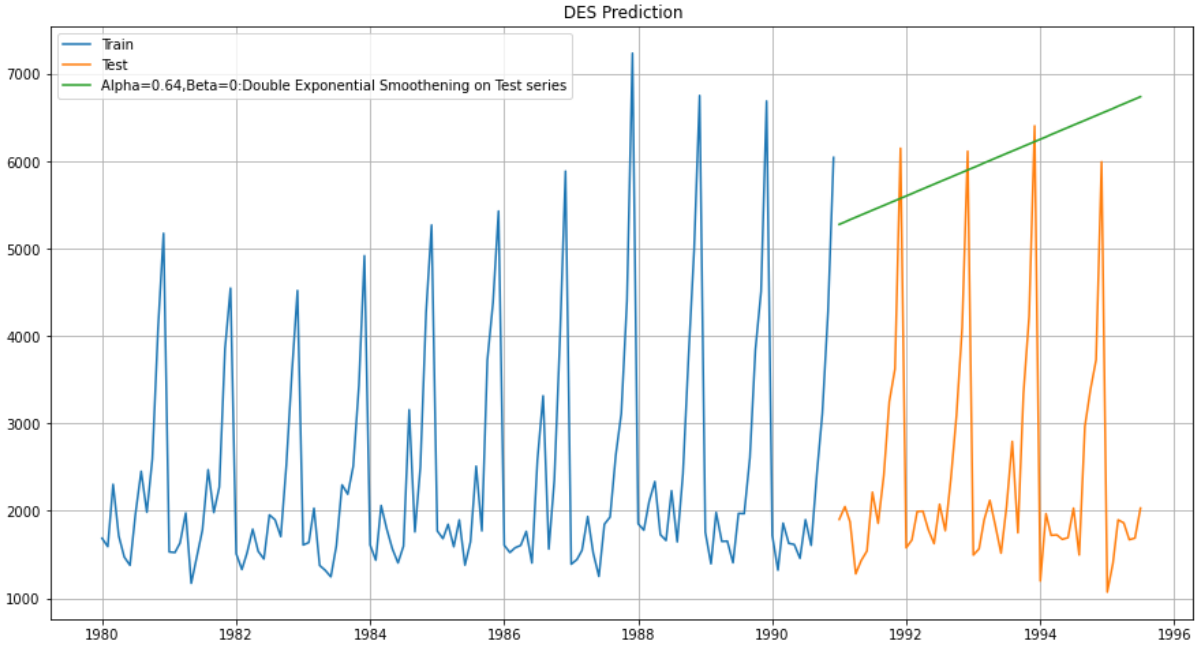
  $$T_t=\beta(L_t-L_{t-1})+(1-\beta)T_{t-1}$$

  β = Smoothing constant for trend and its value lies between 0 and 1.

- The forecast at time t+1 is given by

  $$F_{t+1}=L_t+T_t$$

$$F_{t+n}=L_t+nT_t$$



Green line shows forecasting done by double exponential smoothing with α= 0.64 and β=0.

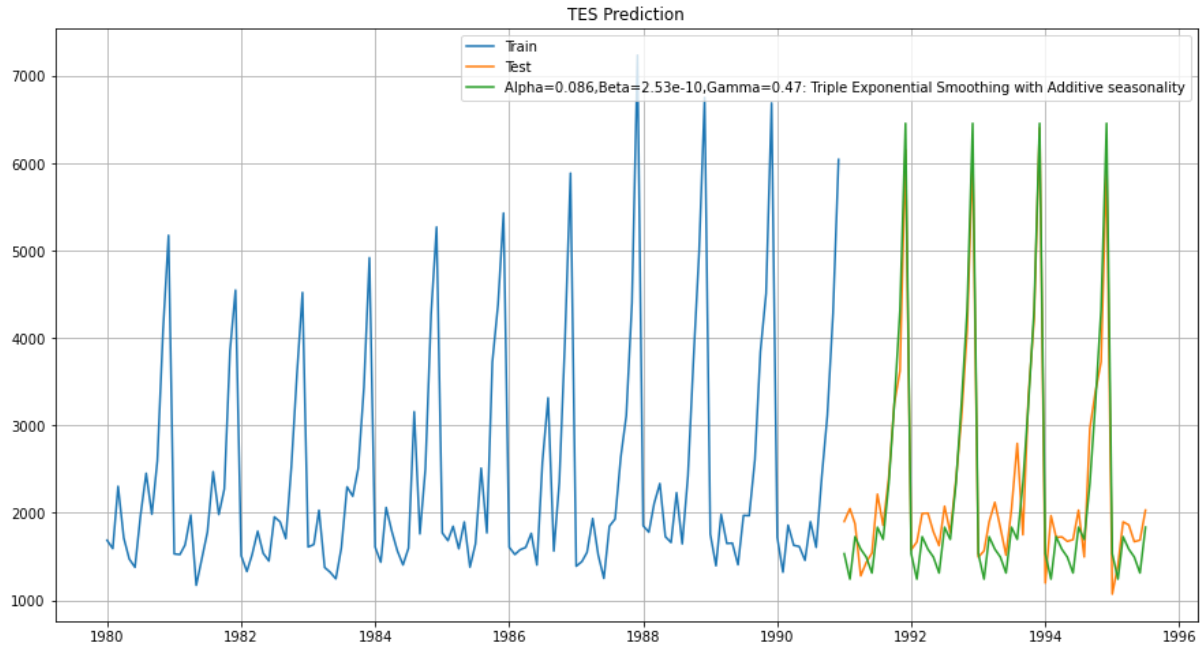
-RMSE of double exponential smoothing = 3851

3)Triple exponential smoothing/ HOLT-Winters exponential smoothing - This model is applicable when the data have trend and seasonality. This model has 3 smoothing parameters: level, trend and seasonality.

α = Smoothing constant for level
β = Smoothing constant for trend
γ = Smoothing constant for seasonality

- It is done by both additive seasonality and multiplicative seasonality

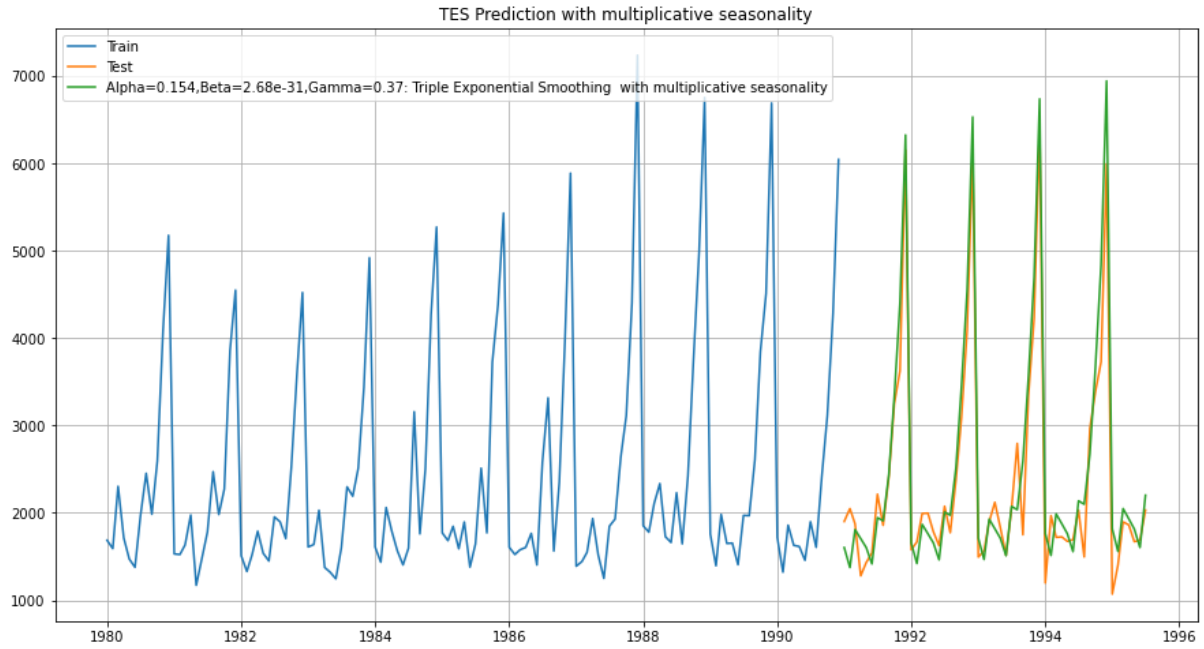
i) Triple exponential smoothing with additive seasonality



Green line shows a forecast done by Triple exponential smoothing with additive seasonality and α= 0.86 , β=2.5e-10, γ=0.47.

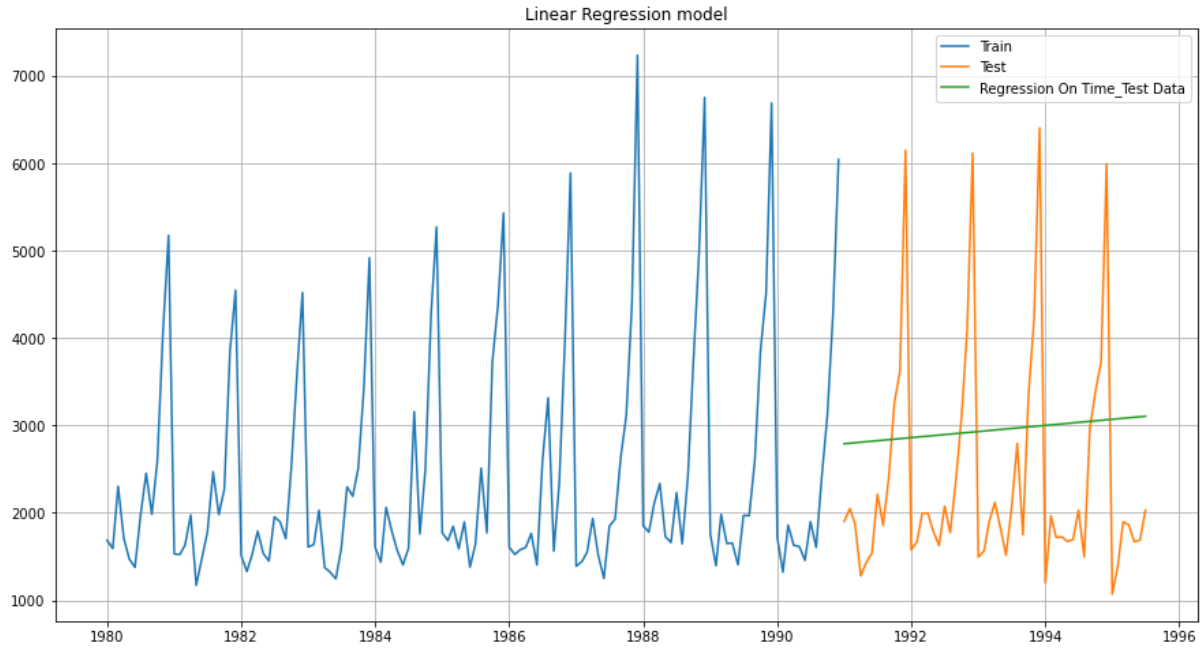
- RMSE of Triple exponential smoothing with additive seasonality = 362.72

ii)Triple exponential smoothing with multiplicative seasonality



Green line shows a forecast done by triple exponential smoothing with multiplicative seasonality and α= 0.154 , β=2.68e-31, γ=0.37.
-RMSE of triple exponential smoothing with multiplicative seasonality= 383

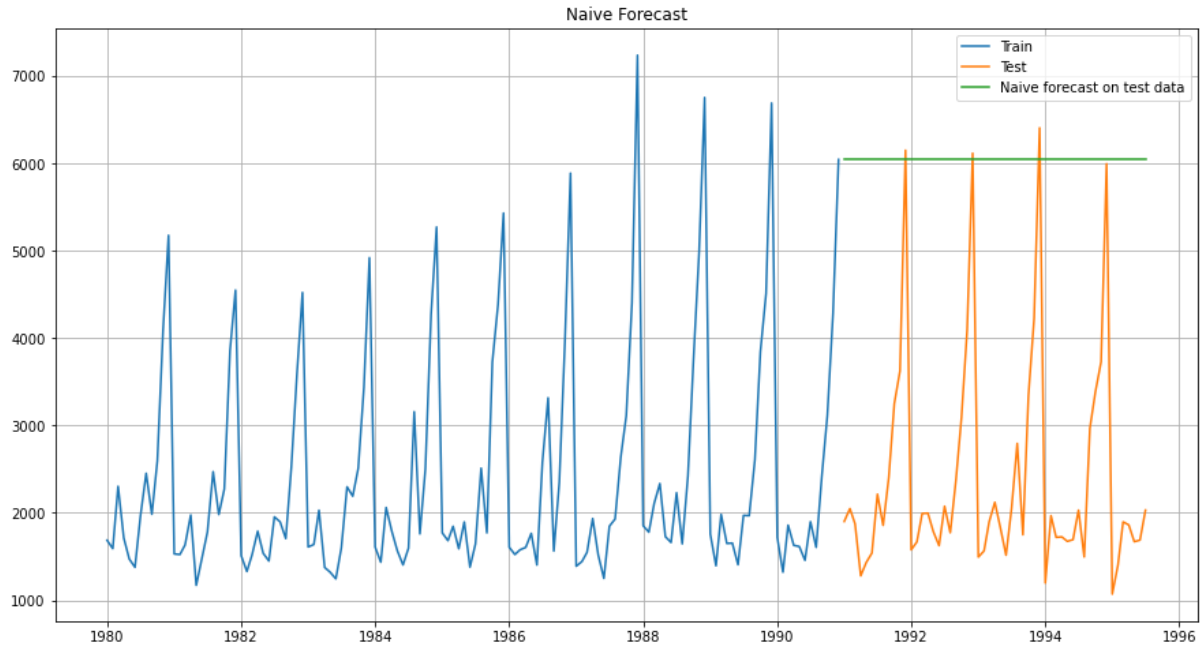
4) Linear Regression model- For this model we have regressed the 'Sparkling' variable against the order of occurrence. For this we have modified our training data before fitting it into a linear regression model.



Green line shows forecasting done by linear regression model on test data

- RMSE of Linear regression model = 1389

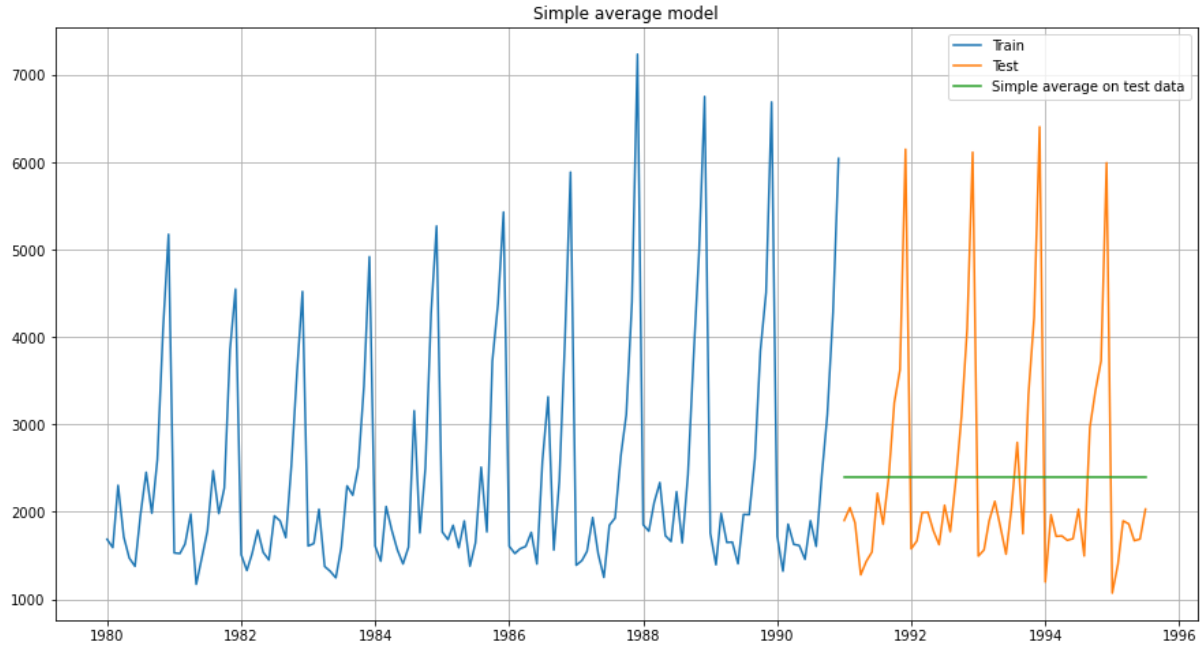
5) Naive model- This model works by forecasting the last value in the train set i.e it will forecast today's value to tomorrow and tomorrow's value to the day after tomorrow. So the forecasting of future days will be the same as today.



Green line shows Naive forecast done on test data by taking the last value of train set which is 6047.

- RMSE of Naive model = 3864

6) Simple average model- This model works by forecasting the mean average of observations in a train set.



Green line shows forecasting done by simple mean average model on train set by taking average mean of 2403.8.

- RMSE of simple average model = 1275.08

Question 5 -

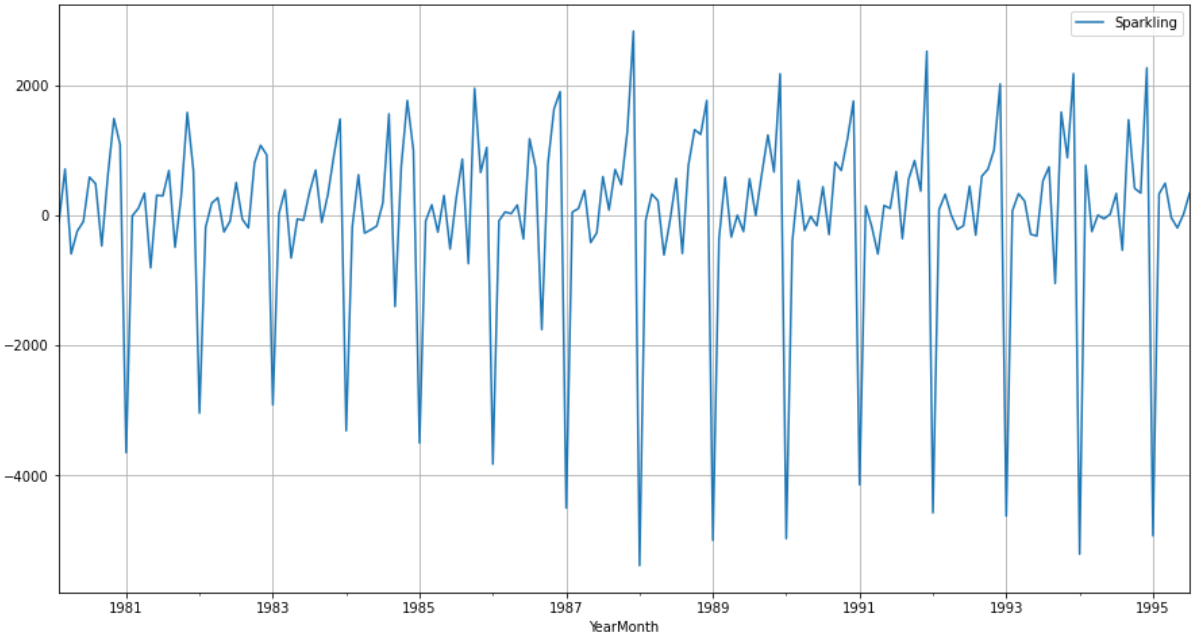
Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment.

Solution- The stationarity of a series is determined by the 'Augmented Dickey-Fuller test(ADF)' which is an unit root test. It determines whether there is a unit root and subsequently whether the series is non-stationary. The hypothesis for ADF test are:

Null hypothesis,H0 - the time series has unit root and is thus non- stationary.
Alternate hypothesis,H1- The time series doesn't have a unit root and thus is stationary.

Significance level,= 0.05
For a series to be stationary p value has to lower than significance level of 0.05.

- Here p value is 0.75, which is higher than significance level and thus time series is not stationary
- The series is made stationary by using one level of differencing and after application of one level of differencing p value is 0 and thus the series has become stationary.
- Plot after making series stationary-



-

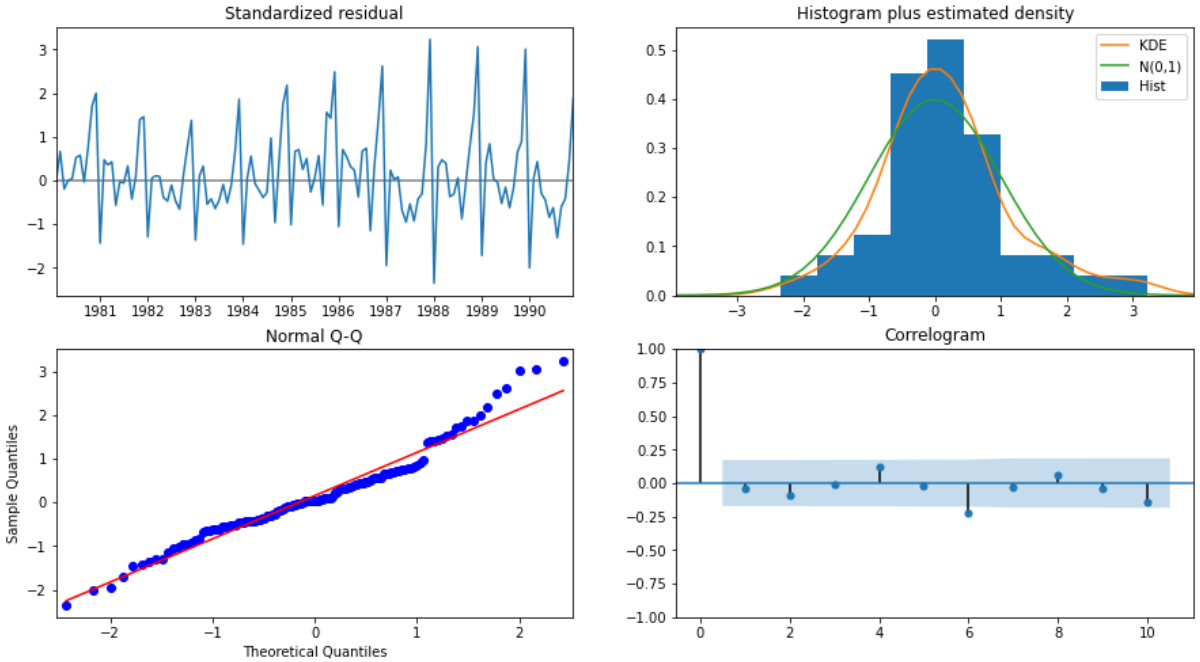
Question 6-

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.

Solution-

1) Automated ARIMA model-
-ARIMA stands for Autoregressive integrated moving average.
- An ARIMA model consists of Autoregressive(AR) part and Moving average(MA) part
-The parameters of ARIMA model are:
p- The number of lag observations included in the model, also called the lag order.
d- The number of times that the raw observations are different, also called the degree of difference.
q- The size of the moving average window,also called the order of moving average.

- This is known as Box-Jenkins methodology for building the ARIMA model
- Here the automated ARIMA model is formed by finding out the lowest Akaike Information Criterion(AIC). Lower the AIC better is the model.
- By using itertools ARIMA model with order(p=2,d=1,q=2) has the lowest AIC of 2213.50.
- Diagnostic plot-

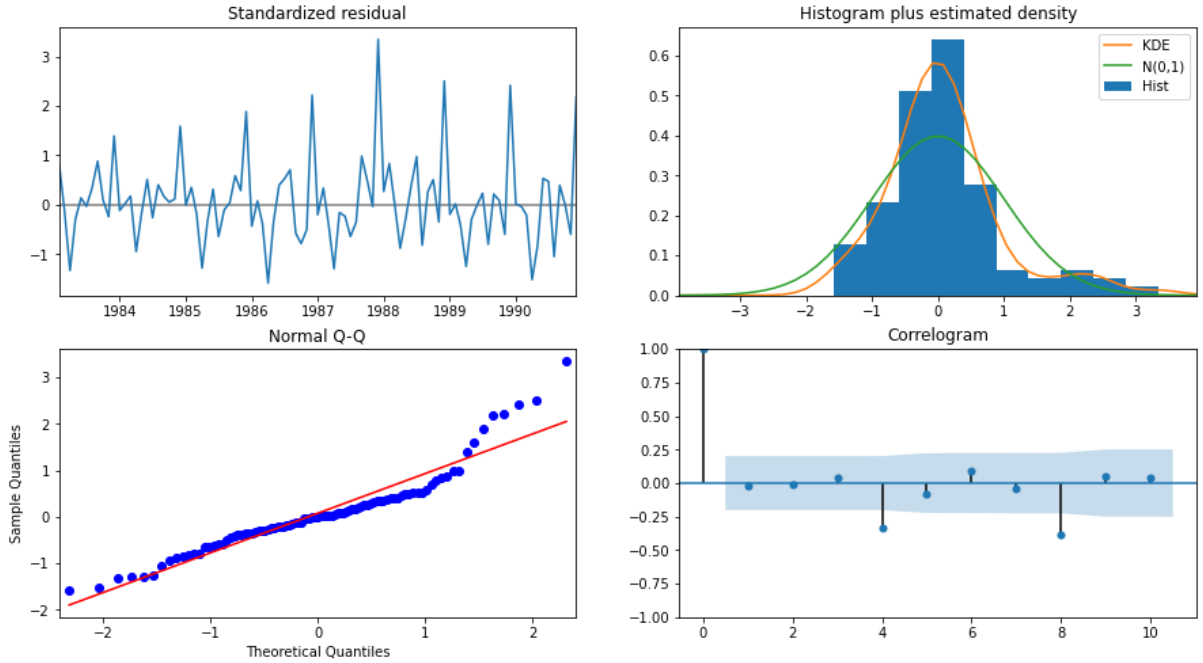


- RMSE of Automated ARIMA model- 1299.98

2) Automated Seasonal ARIMA model- For a Seasonal Auto-Regressive Integrated Moving Average we have to take care of four parameters such as AR (p), MA (q), Seasonal AR (P) and Seasonal MA (Q) with the correct of differencing (d)

and seasonal differencing (D). Here, the 'F' parameter indicates the seasonality/seasonal effects over a particular period.

- Here SARIMA model is formed by order(3,1,3) and seasonal order(3,0,0,11)
- Diagnostic plot-



- RMSE of automated SARIMA model= 1190.9.

Question 7-

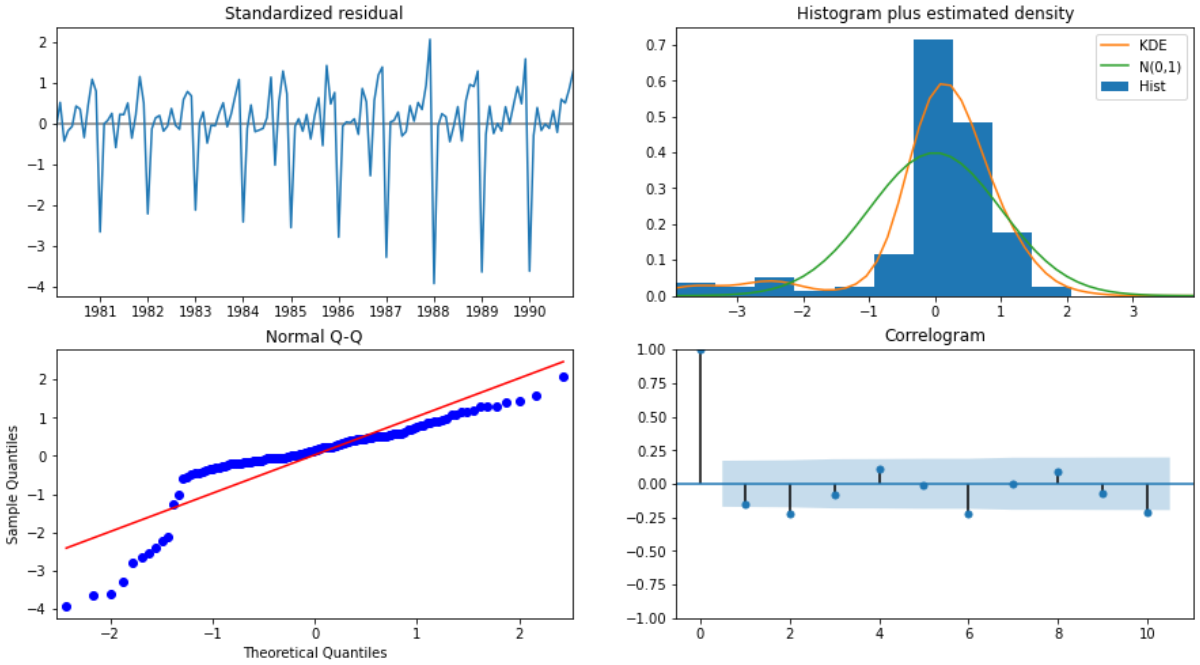
Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.

Solution -

i) Manual ARIMA model- In Manual ARIMA model the AR(p) order is selected by where the partial auto-correlation plot cuts off and Ma(q) order is selected by where Auto-correlation plot cuts off.

-Here by looking at ACF ,p=0 and by looking at PACF, q=0, . So the manual ARIMA model is formed by using order(0,1,0).

- Diagnostic plot-



- RMSE of manual ARIMA model = 3864.28

ii) Manual SARIMA model- Here Manual SARIMA model is formed by taking order(3,1,3) and seasonal order(0,0,11,22).

- RMSE of Manual SARIMA model = 3083.67.

Question 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.

Solution -

	Test RMSE
Alpha=0,SES	1275.080000
Alpha=0.64,Beta=0:DES	3851.000000
Alpha=0.086,Beta=2.53e-10,Gamma=0.47:TES	362.720000

Alpha=0.154,Beta=2.68e-31,Gamma=0.37:TES	383.000000
Linear Regression model	1389.000000
Naive model	3864.000000
Simple average model	1275.081804
AUTO ARIMA model(order=(2,1,2))	1299.980000
Auto SARIMA model(order=(3,1,3),seasonal_order=(3,0,0,11))	1190.090000
Manual ARIMA model(order=(0,1,0))	3864.280000
Manual sARIMA model(order=(3,1,3),seasonal_order=(0,0,11,22))	3083.670000

Inference- TES model with Alpha=0.154, Beta= 2.68e-31 and Gamma =0.37 have lowest RMSE and thus is the best model.

Question 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.

Solution- Auto SARIMA model is chosen here to make forecasts on full data because in comparison with other models it has lower RMSE.

- The Auto SARIMA model is applied and the forecast has been done for the next 12 months.

Sparkling	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-01	2480.209187	1284.002784	-36.390026	4996.808401
1995-09-01	2520.060771	1298.895698	-25.728017	5065.849559
1995-10-01	2030.446888	1317.188466	-551.195065	4612.088842
1995-11-01	3711.571893	1317.158844	1129.987997	6293.155790
1995-12-01	2568.991184	1320.028727	-18.217580	5156.199947
1996-01-01	1930.918664	1326.656176	-669.279660	4531.116988

1996-02-01	2412.938121	1326.613178	-187.175928	5013.052171
1996-03-01	2840.326128	1329.406922	234.736441	5445.915815
1996-04-01	2092.727728	1333.909762	-521.687364	4707.142821
1996-05-01	2169.646160	1333.959311	-444.866048	4784.158367
1996-06-01	2849.595267	1336.356710	230.384245	5468.806288
1996-07-01	2093.504192	1409.987466	-670.020460	4857.028844

- RMSE of Automated SARIMA model on full data = 1204.2

Question 10-

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

Solution-

Forecasting of model on full dataset shows-

- Maximum sale in the month of November.
- Medium sale in the months of August,September,December,February,March and June.
- Lower Sale in October,April,May and July.
- Lowest sale in January.

Suggestion- In the month of January wine companies can give maximum discounts to increase sale value.

-In the lower sales month they can give some discounts.

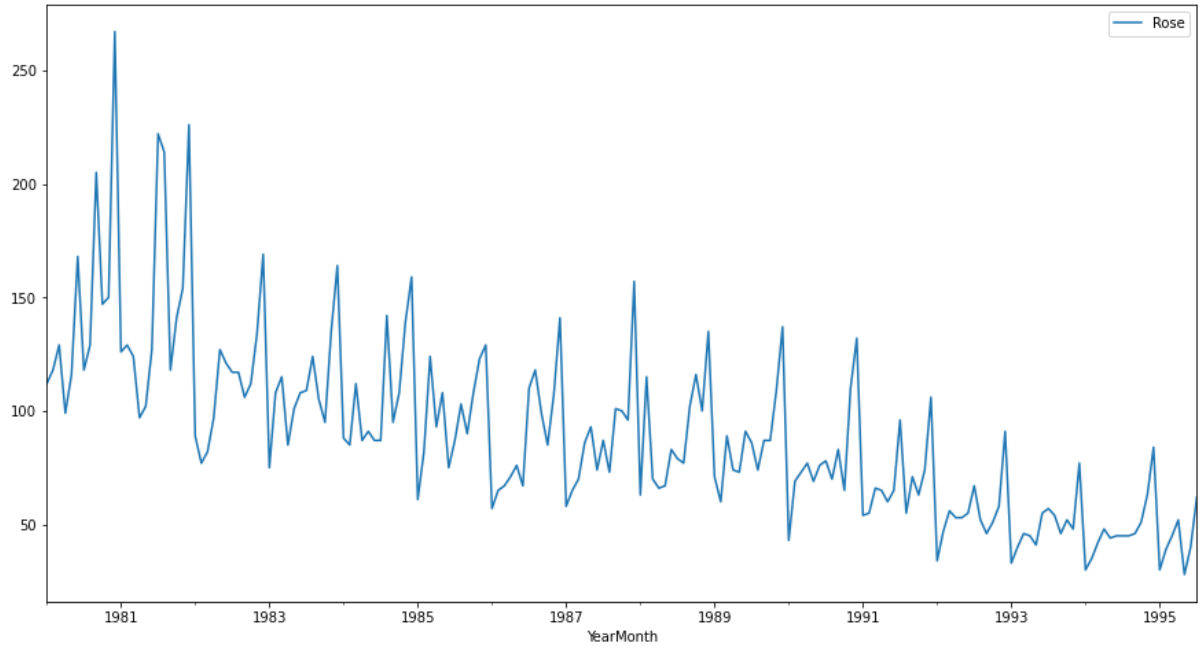
- In the maximum sales month there is no need to give any discounts.

2)Rose wine-

Question 1-

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

Solution-
-Data has been read as a time series.
-There are 2 missing values which have been replaced by using the forward fill method.
-Plot-

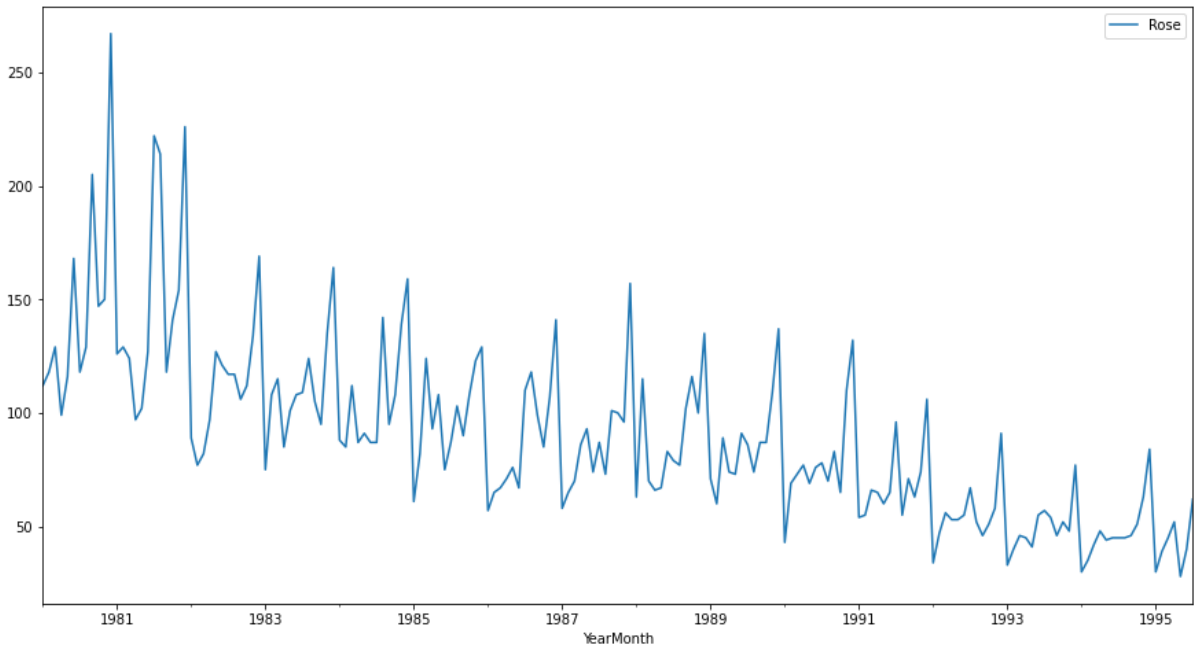


Question 2-

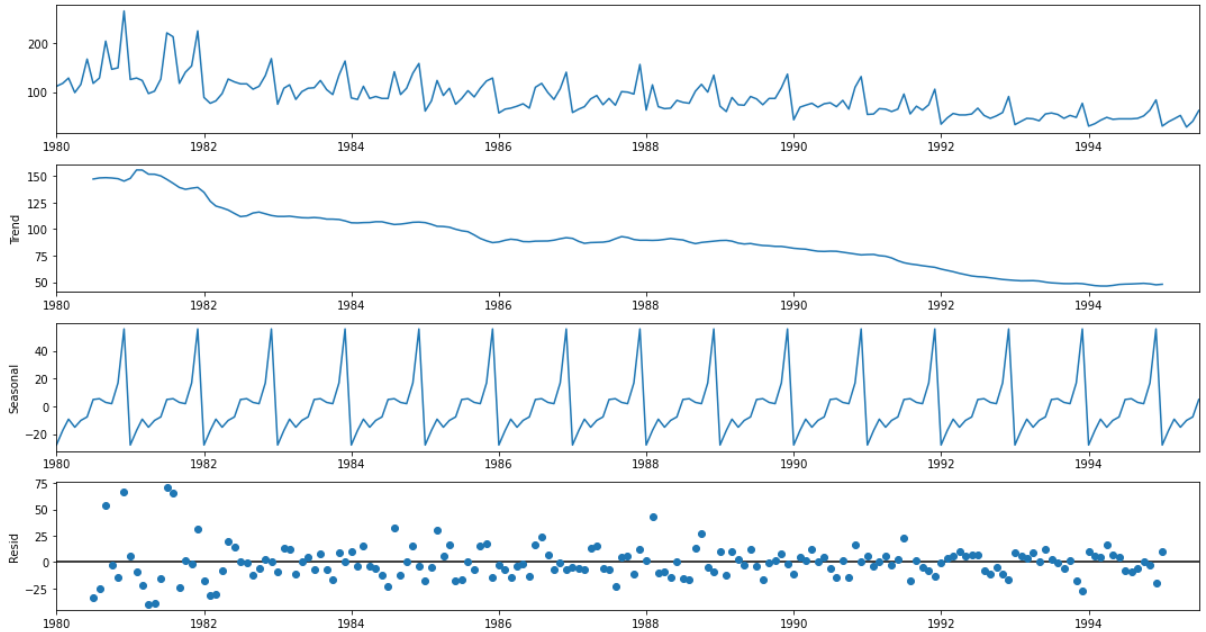
Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

Solution-

- First step of exploratory data analysis in a time series is plotting data.
- Plot-



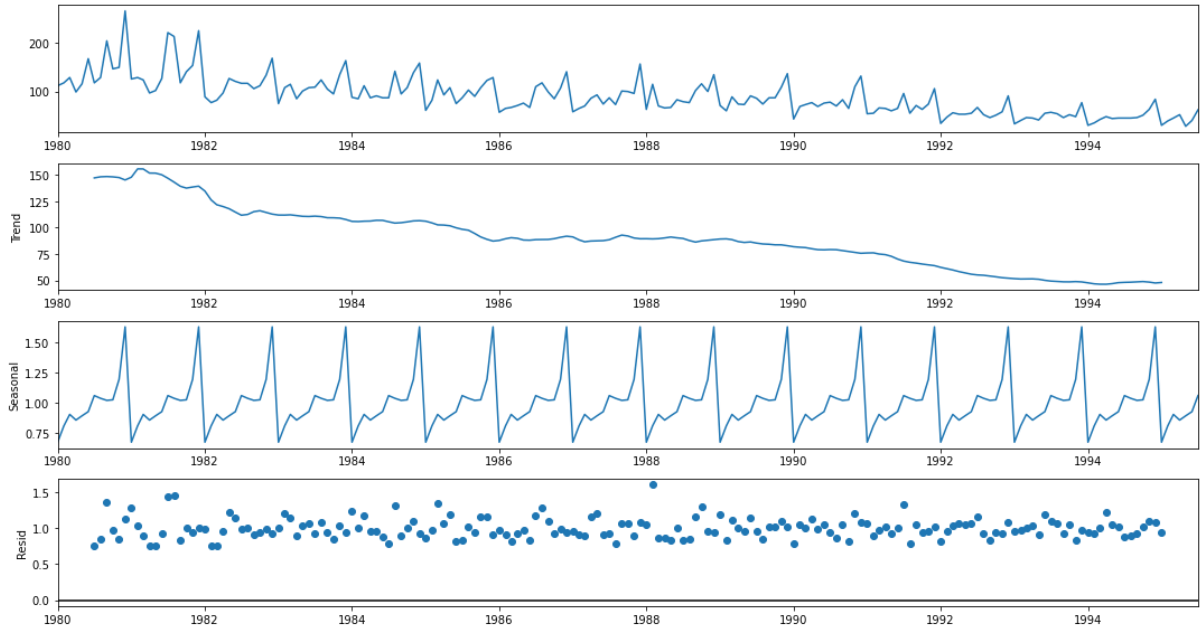
-By looking at the plot we can see that there is both trend and seasonality in the time series.
- We can look more in detail in the time series by performing decomposition
i) Decomposition with Additive method-



Inference- There are both trend and seasonality in this series

- Error is showing some pattern i.e all information is not completely extracted from the observations.

ii)Decomposition with multiplicative method-



Inference- Downward trend and seasonality are present

-Error is not showing any pattern i.e all information is extracted from the series.

-In comparison to the Additive method, the Multiplicative method is better in performing decomposition.

Question 3-

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

Solution-

Whole data has been split into training sets and test sets.

Training set- Data collected from 1980/01/01 till 1990/12/01

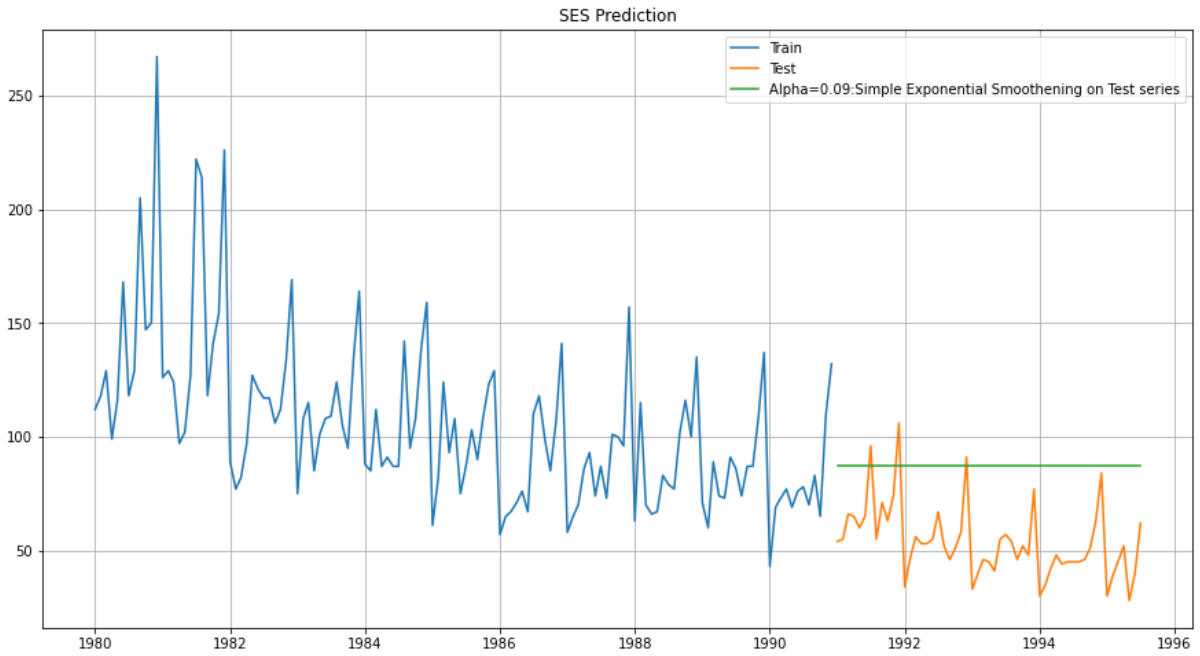
Testing set- Data collected from 1991/01/01 onwards.

Question 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 and simple average models. should also be built on the training data and check the performance on the test data using RMSE.

Solution-

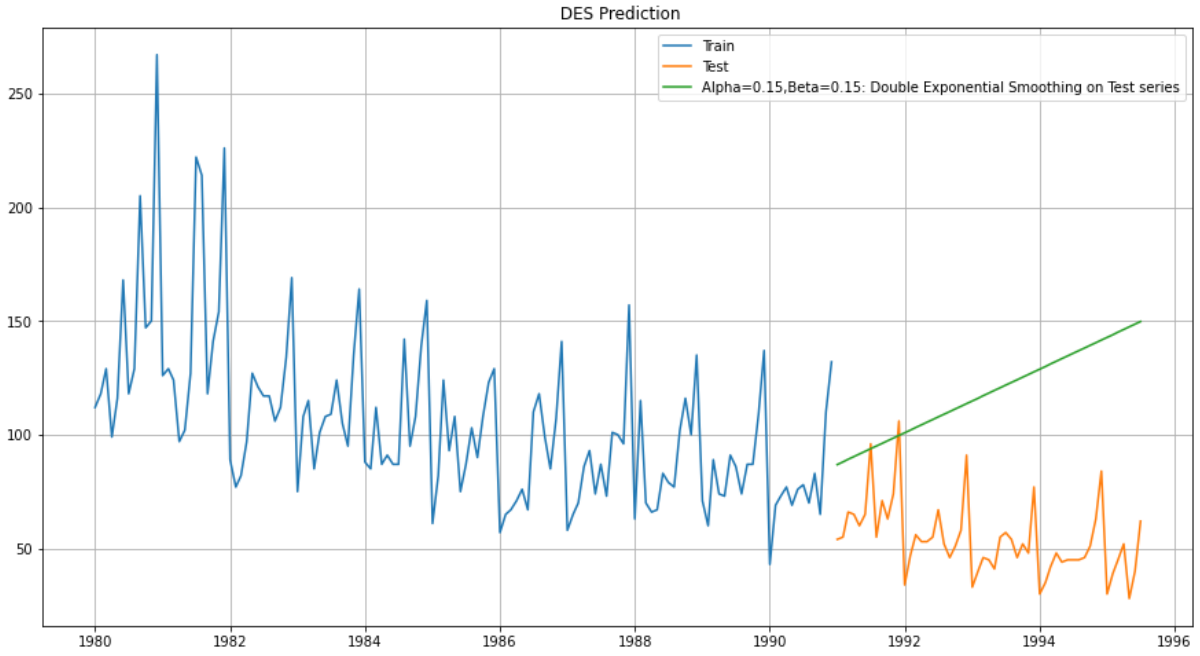
1) Simple exponential smoothing(SES)-



Inference- Green line shows forecasting of single exponential smoothing on test data with alpha=0.09.

- RMSE of SES = 37

2) Double exponential smoothing(DES)-

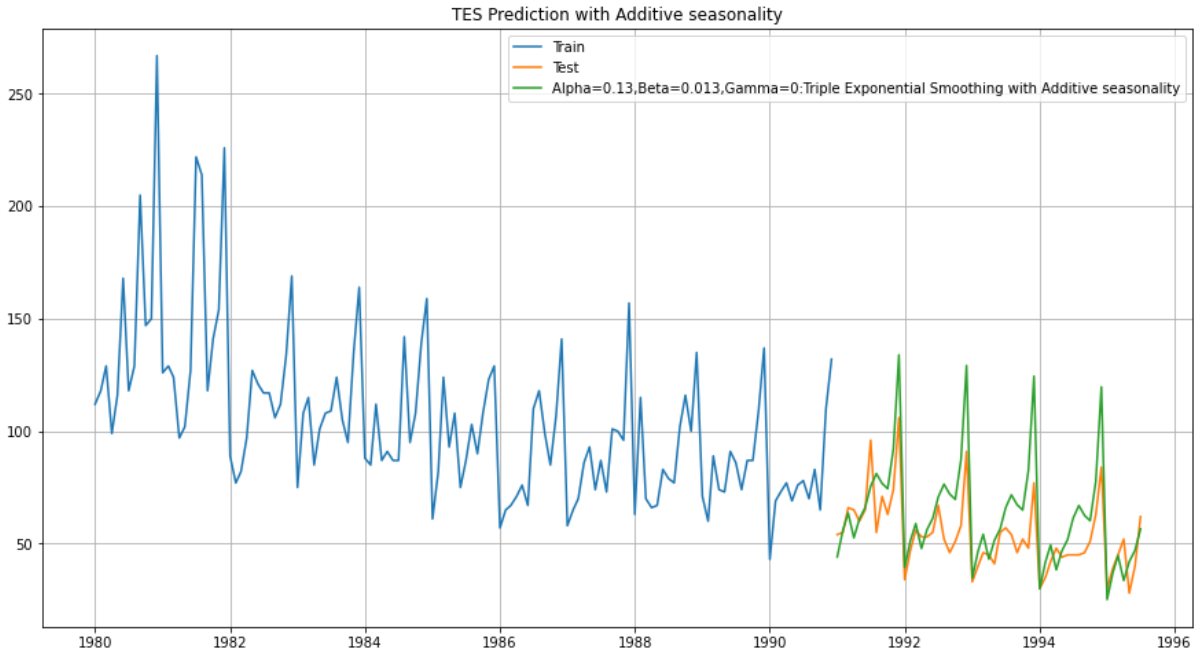


Inference- Green line shows forecasting done by double exponential smoothing on test set with Alpha=0.15 and Beta=0.15

- RMSE of DES = 70.6

3) Triple exponential smoothing(TES)-

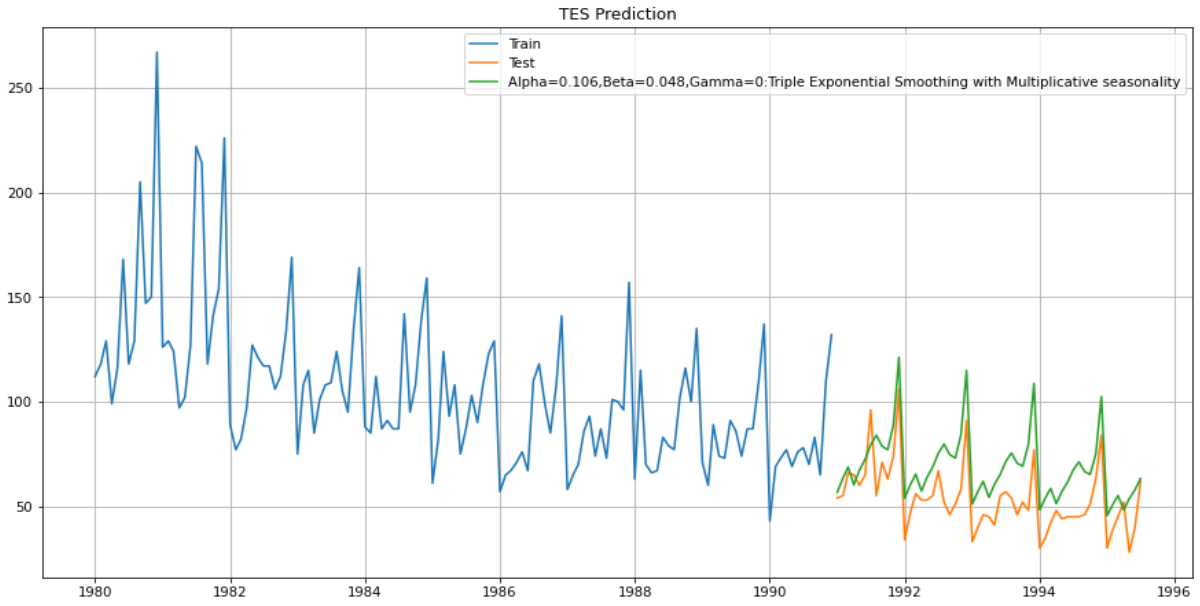
i) With additive seasonality-



Inference- Green line shows forecasting done by TES on test data with Alpha=0.13, Beta=0.013 and Gamma= 0.

- RMSE of TES with additive seasonality = 362.72.

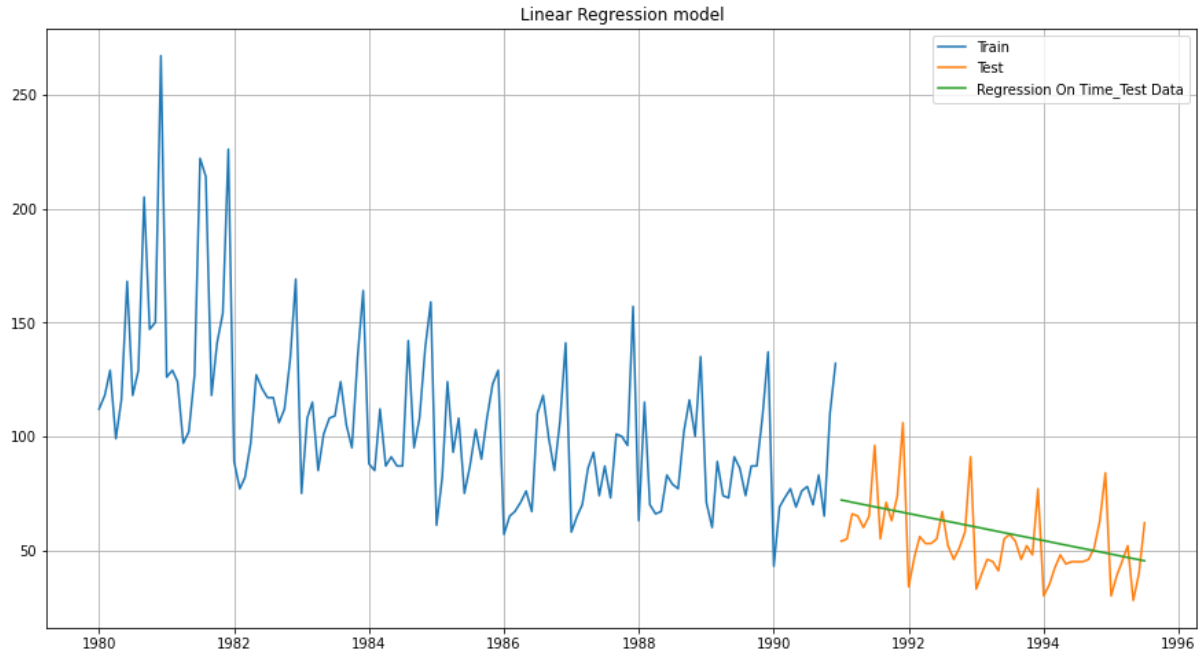
ii) With multiplicative seasonality-



Inference - Green colour line shows forecasting done by TES on test data with Alpha=0.106, Beta=0.048, Gamma=0.

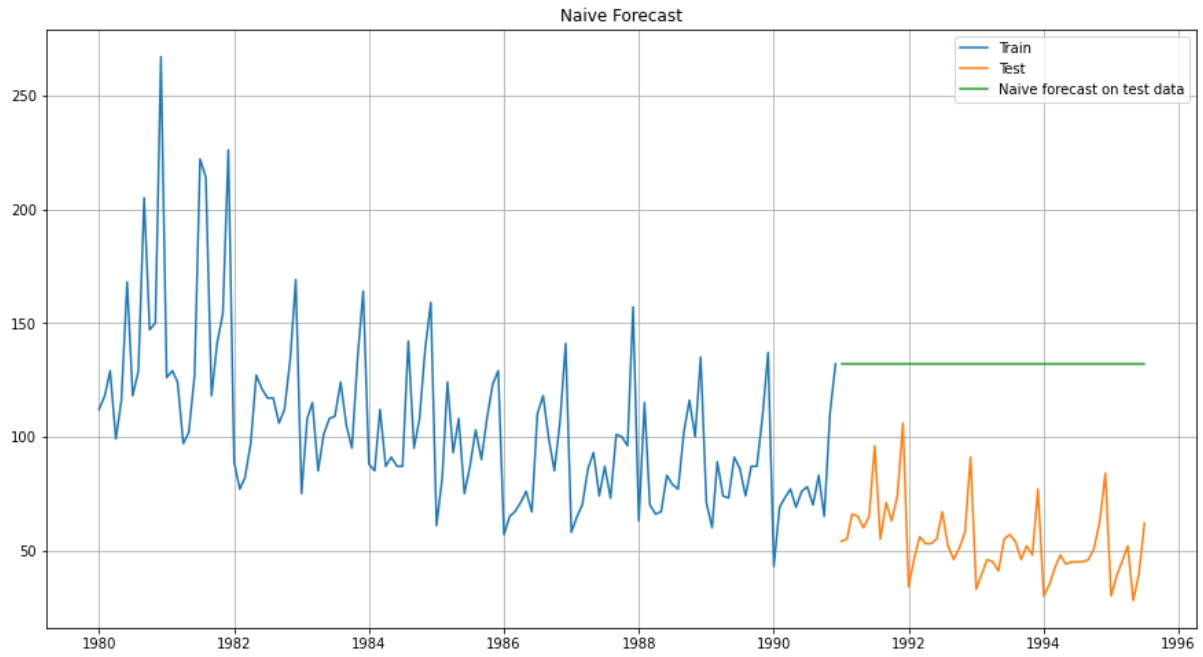
- RMSE of TES model with multiplicative seasonality = 17

4)Linear regression model-



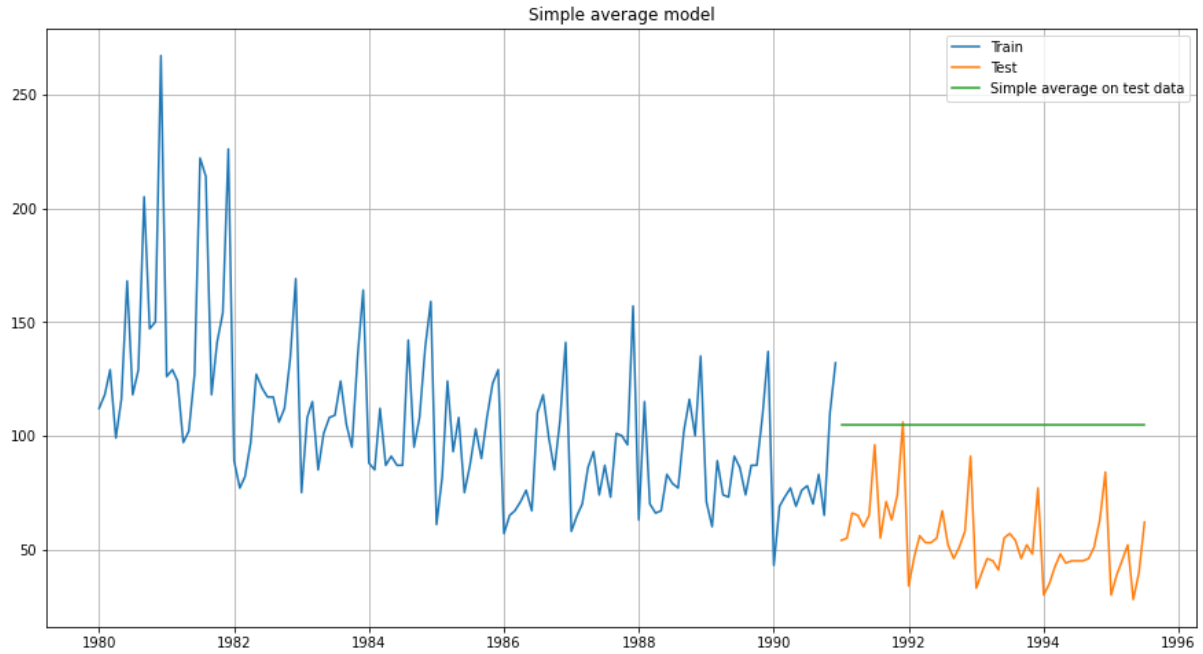
Green line shows forecasting done by linear regression model on test data
-RMSE of Linear regression model = 15

5) Naive forecast model-



Green line shows forecast done by Naive model on test data
-RMSE of Naive forecast model = 80

6) Simple mean average model-



Green colour line shows forecast done by simple mean average model on test data

-RMSE of Simple mean average = 53.48

Question 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.

Solution-

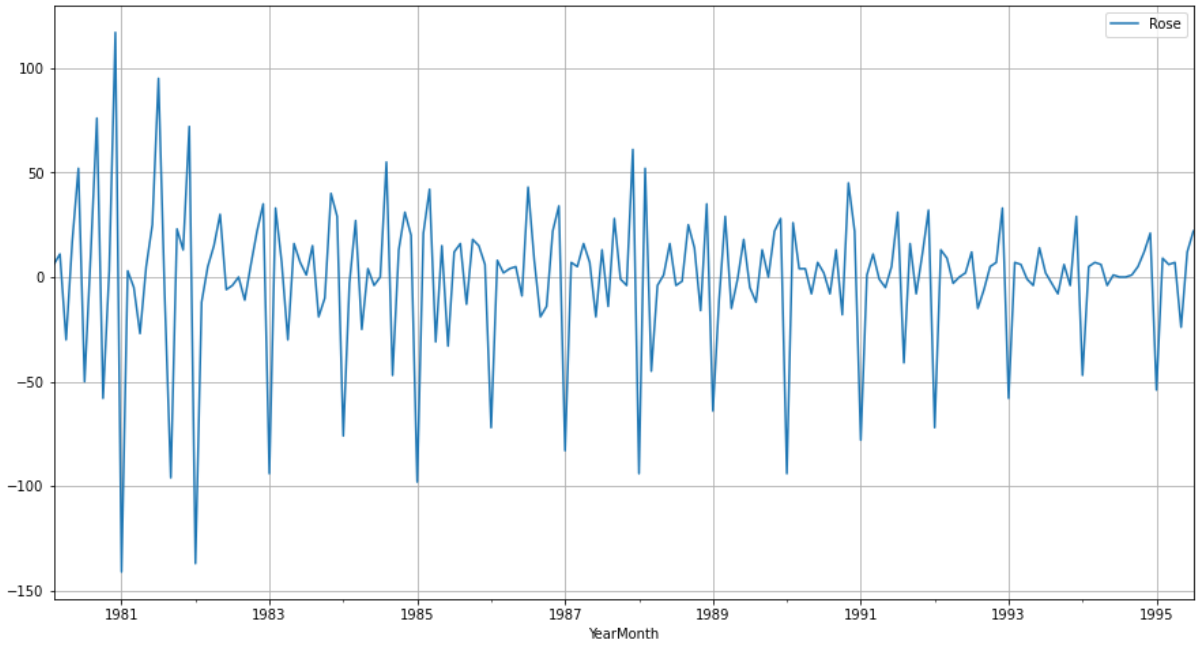
The stationarity of a series is determined by the 'Augmented Dickey-Fuller test(ADF)' which is an unit root test. It determines whether there is a unit root and subsequently whether the series is non-stationary. The hypothesis for ADF test are:

Null hypothesis,H0 - the time series has unit root and is thus non- stationary.
Alternate hypothesis,H1- The time series doesn't have a unit root and thus is stationary.

Significance level,= 0.05
For a series to be stationary p value has to lower than significance level of 0.05.

- Here p value is 0.46, which is higher than significance level and thus null hypothesis is not rejected and the series is not stationary.

- The series is made stationary by using one level of differencing. After using one level of differencing, p value is 3.08e-11 and thud the series has become stationary
- Plot after making series stationary-



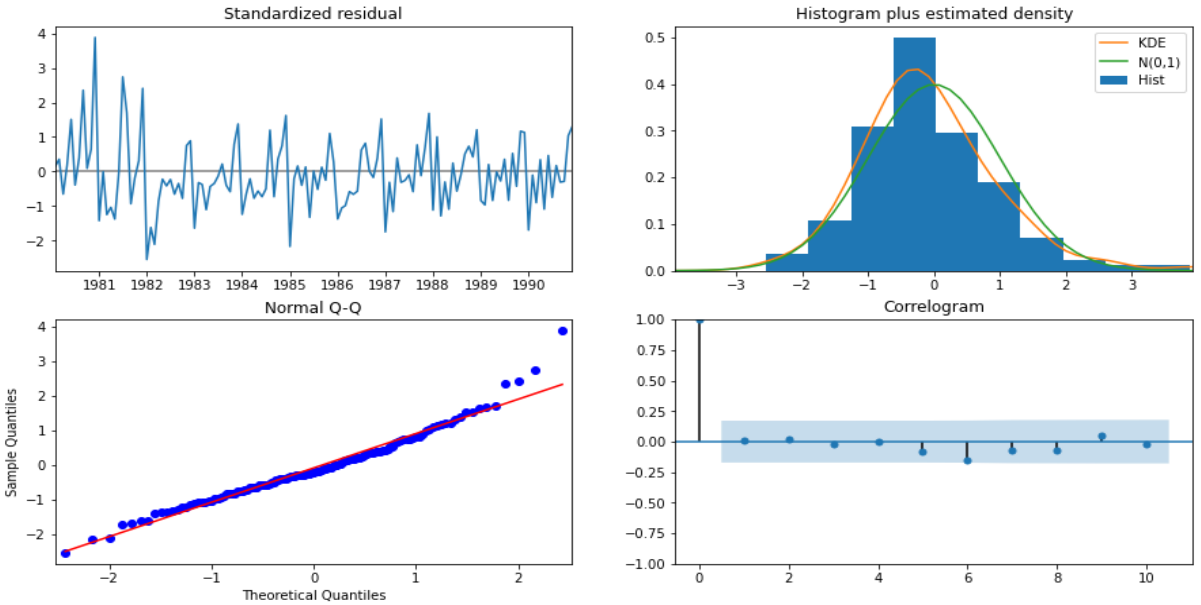
Question 6-

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.

Solution -

1) Automated ARIMA model- It is formed by using order(2,1,3)

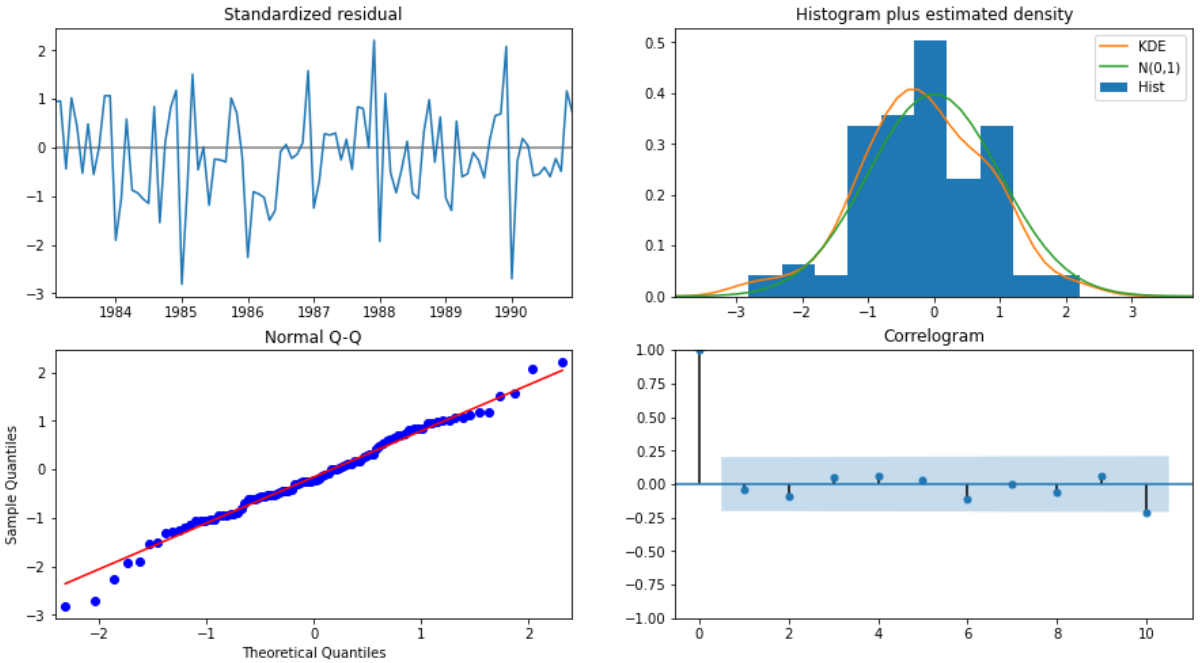
- Diagnostic plot-



-RMSE of automated ARIMA plot = 36.84

2)Automated SARIMA plot- It is formed by using order(3,1,3) and seasonal order(3,0,1,11).

-Diagnostic plot-



-RMSE of Automated SARIMA model = 34.36

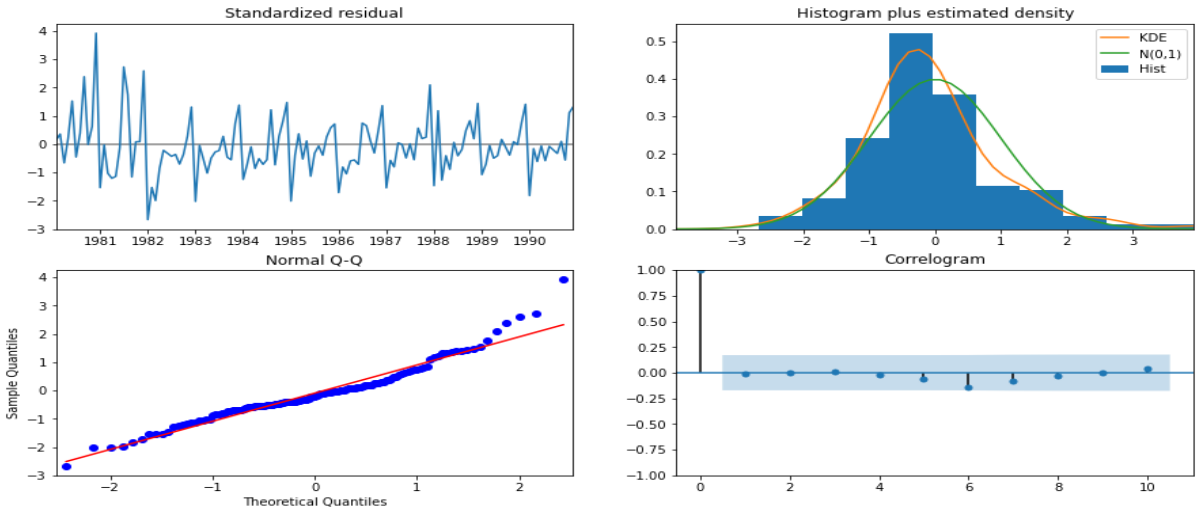
Question 7-

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.

Solution-

1) Manual ARIMA model- Formed by using order(2,1,2)

- Diagnostic plot-



-RMSE of manual ARIMA model = 36.89

2) Manual SARIMA model- Formed by using order(2,1,2) and seasonal order(0,0,11,22).
    - RMSE of manual SARIMA model = 66.61

Question 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.

Solution-

Dataframe-

	Test RMSE
SES: Alpha=0.09	37.000000
DES: Alpha=0.15, Beta=0,15	70.600000
TES: Alpha=0.13, Beta=0.013, Gamma=0	362.720000
TES: Alpha=0.106, Beta=0.048, Gamma=0	17.000000
Linear Regression	15.000000
Naive model	80.000000
Simple average model	53.480857
Auto ARIMA model(order(2,1,3))	36.840000
Auto SARIMA model(order(3,1,3),seasonal order(3,0,1,11))	34.360000
manual ARIMA model(order(2,1,2)	36.890000
manual SARIMA model(order(2,1,2),seasonal order(0,0,11,22))	66.610000

Inference- Linear Regression model is the model with lowest RMSE and is thus better than other models for forecasting.

Question 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.

Solution-

-An Automated SARIMA model is chosen here because of its lower RMSE.

- Model is formed and forecasting is done for the next 12 months with confidence intervals.

Rose	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-01	51.709634	21.089768	10.374448	93.0448
1995-09-01	51.049711	21.363923	9.177192	92.922230
1995-10-01	38.184378	21.404736	-3.768134	80.136889
1995-11-01	44.181390	21.627924	1.791437	86.571342
1995-12-01	52.799143	21.835805	10.001752	95.596534
1996-01-01	48.955481	21.945114	5.943848	91.967113
1996-02-01	46.406292	21.946433	3.392074	89.420510
1996-03-01	46.341953	22.005910	3.211161	89.472744
1996-04-01	51.629679	22.237744	8.044501	95.214858
1996-05-01	49.696122	22.634348	5.333615	94.058628
1996-06-01	45.560215	23.115619	0.254434	90.865997
1996-07-01	47.486408	23.118318	2.175337	92.797478

-RMSE of full data on this model is 45.35

QUESTION 10-

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

Solution- An automated SARIMA model is formed on full data.

- Forecasting of this model in the next 12 months shows that there will be more wine in August ,September,December and April.
- There will be less wine in October.
- Suggestions- To increase the sale of wine, companies can give some discounts on wine in October. There is no need for any discounts in August,September,December and April.