

A dark blue vertical bar is positioned on the left side of the slide. A blue arrow-shaped box points to the right from this bar, containing the text 'Problem -2'. In the bottom-left corner, there are several thin, curved, light blue lines that sweep upwards and to the right.

Problem -2

Rose wine Sales Analysis using Time Series

Kanhaiya Awasthi

PGP-DSBA SEPT GREAT LEARNING

Problem 2:-

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.

Data set for the Problem: [Rose.csv](#)

Please do perform the following questions on each of these two data sets separately.

1. Read the data as an appropriate Time Series data and plot the data.
2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.
3. Split the data into training and test. The test data should start in 1991.
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.
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 $\alpha = 0.05$.
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.
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.
8. Build a table with all the models built along with their corresponding parameters and the respective RMSE values 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.
10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

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

We have used Pandas Read CSV Function to read the Data Given in CSV format by passing following arguments.

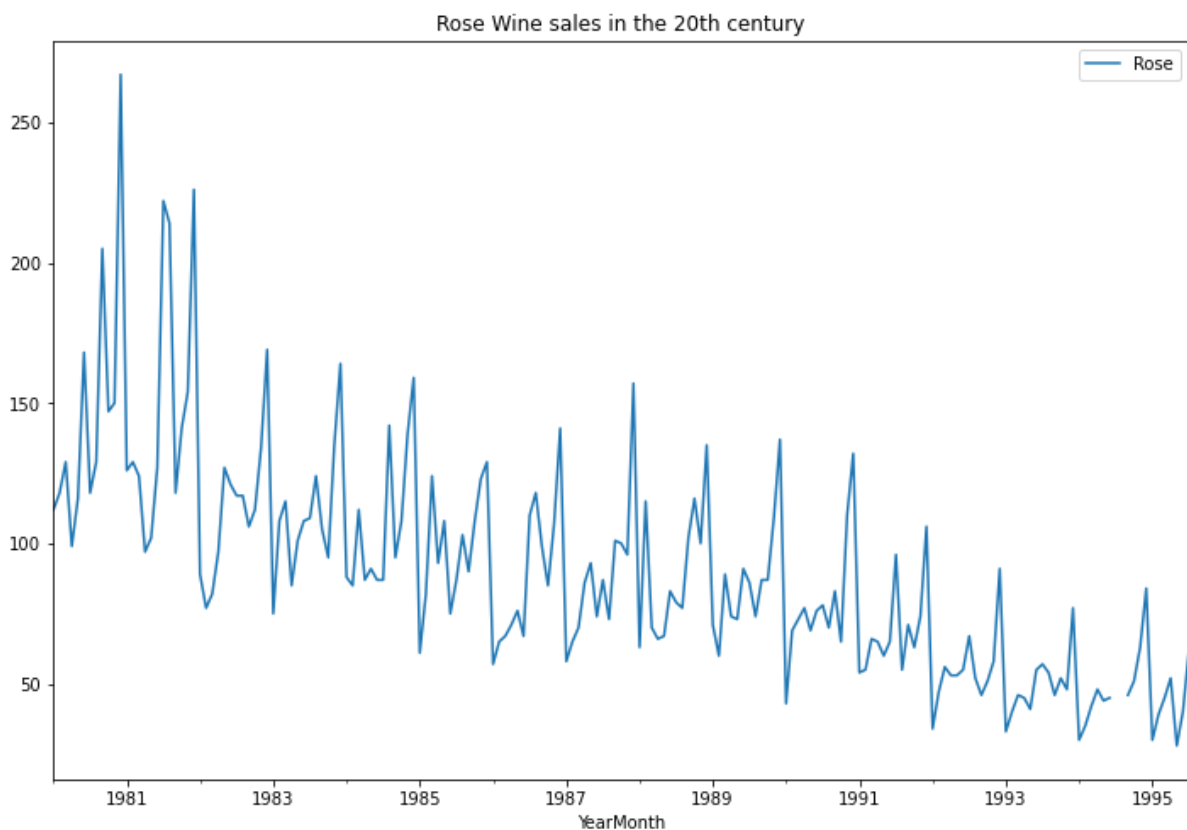
```
df = pd.read_csv('Rose.csv', parse_dates=['YearMonth'],  
                index_col='YearMonth')
```

Since the data is time series data so here Continuation of date matters a lot so we have made YearMonth Column as our index and by passing parse_date = YearMonth , we are telling python that YearMonth column is our Time series column, if we will not pass this argument then python will automatically select the data and find the column which consist of Datetime values.

- Let's have a look on first five rows Time series 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

- Time Series Data Plot



In initial View, the above time series plot has decreasing trend, we can also see seasonality in a plot too.

We can see that between 1994 & 1995 the series looks broken so we will check for the missing values using is null values check function.

```
] : df.isnull().sum()
```

```
] : Rose      2  
    dtype: int64
```

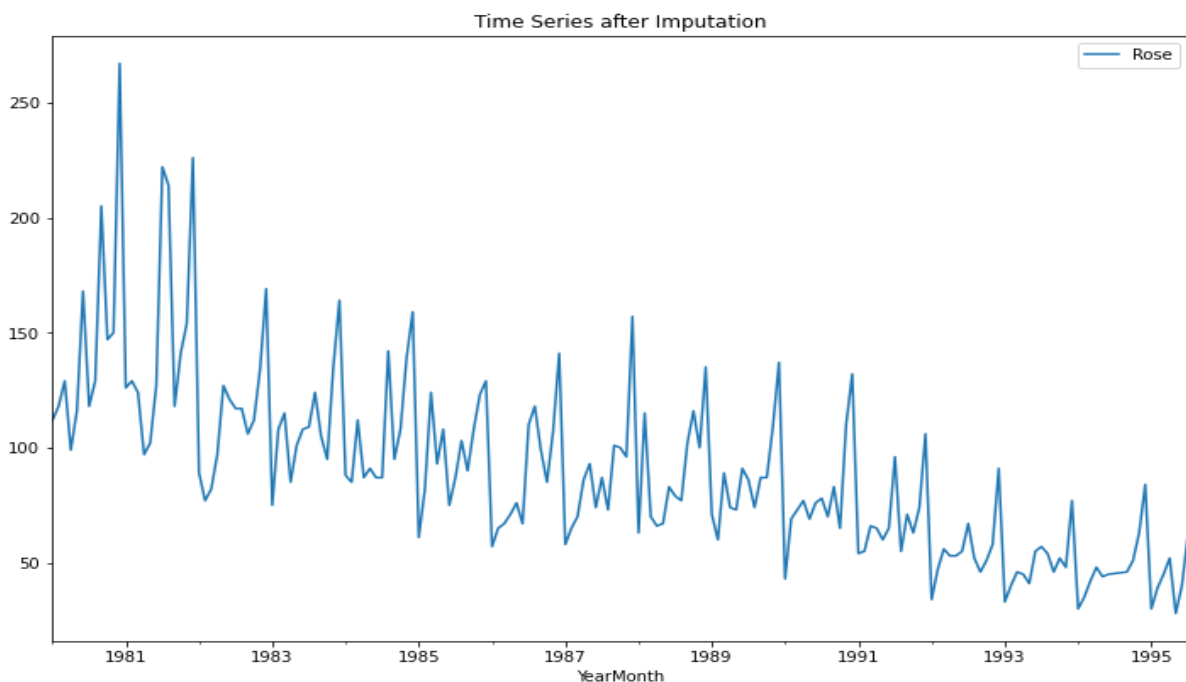
```
] : row_null = df.isnull()  
    row_has_null = row_null.any(axis=1)  
    rows_having_null = df[row_has_null]  
    rows_having_null
```

```
] :  
  
              Rose  
YearMonth  
1994-07-01  NaN  
1994-08-01  NaN
```

Using the above code, we found that in July 1994 & August 1994 the values are missing, so we have to find out best values for imputation.

We have imputed missing values by Interpolation method, Interpolation method draws a line between points and adjust them accordingly.

After Imputation Time series Gap is filled and Our New values are

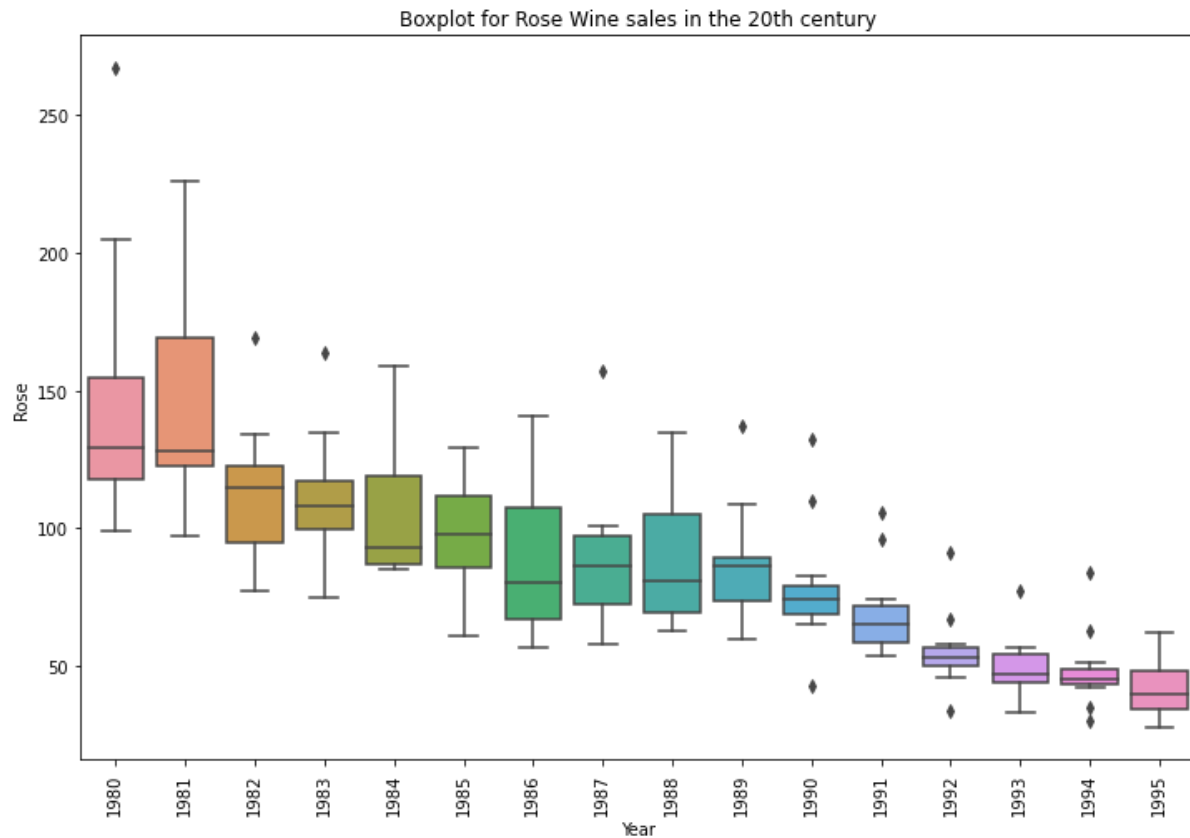


After Imputation Time series Gap is filled and time series is now continuous time series .

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

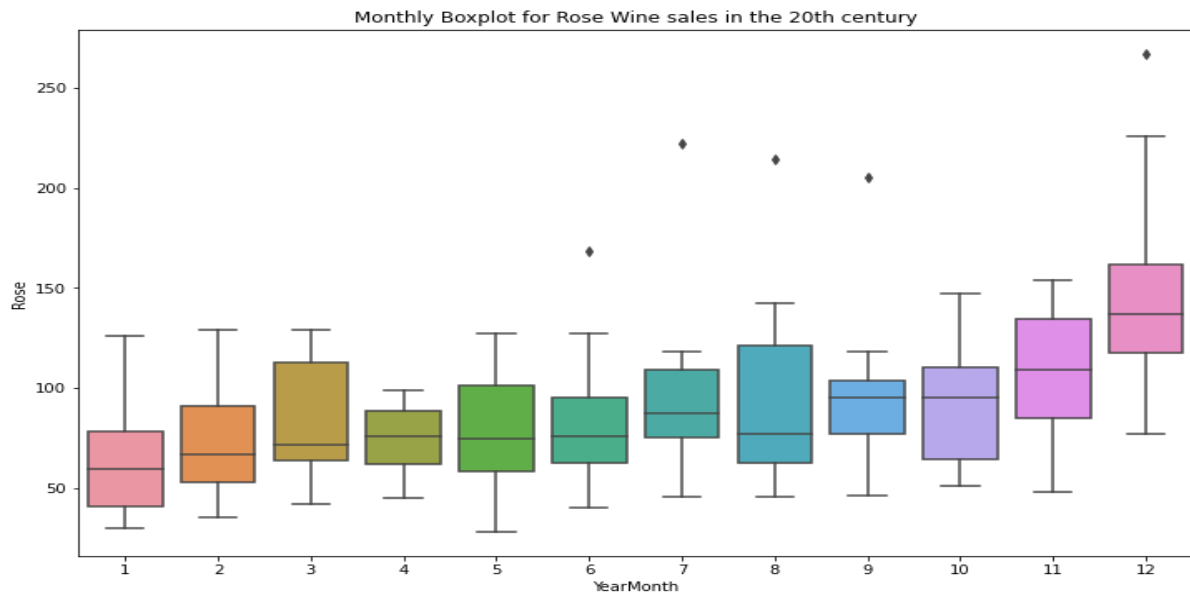
The Given data set tells us about monthly sales of Rose wine of Company ABC Estate Wines, so we will plot this Data across Various years & time frames.

- **Boxplot of yearly Rose Wine Sales: -**



As we saw in the previous plot, there is decreasing trend, Also there are some years where the hike or drop in Sales is more as seen in the outliers.

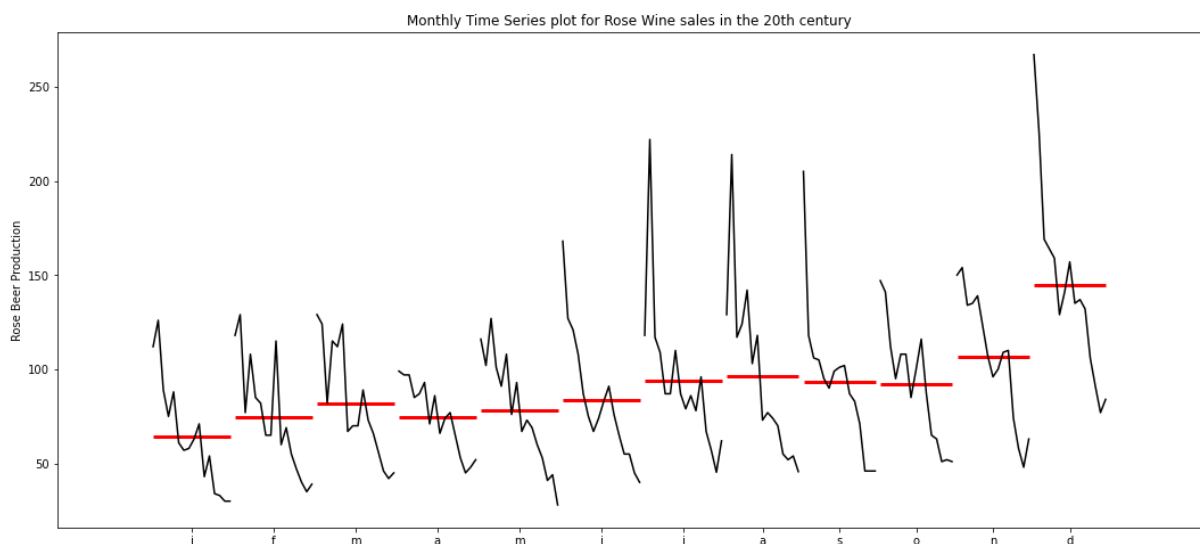
- **Boxplots for Each Calendar Month Across all Years: -**



The boxplot for monthly Sales for each month across years show very few outliers only for June to Sept, & December Months. this shows that in few years, there is a higher sale in particular Month.

Also we can Notice that There are higher median sales is recorded in the Month of December & Lowest median sales is recorded in the Month of January across all years.

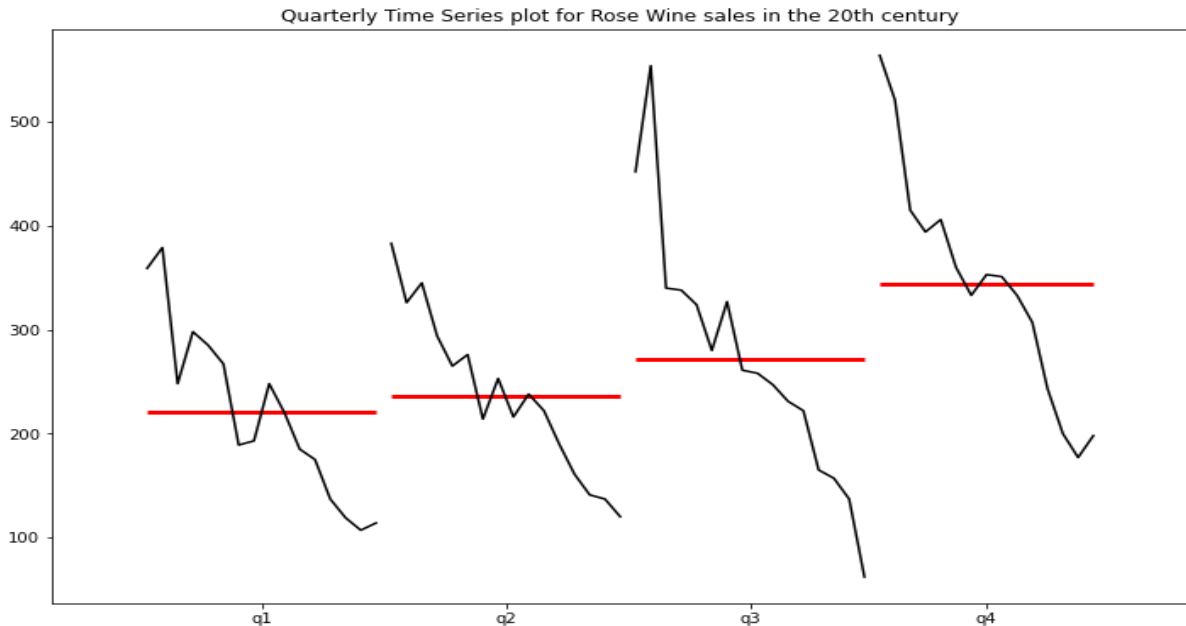
- **Month plot of the Time Series Data of Rose Wine**



The above Month Plot shows that after April Average wine sales are increasing, Peak is noticed in the month of December every year. So there is an yearly seasonality can be noticed.

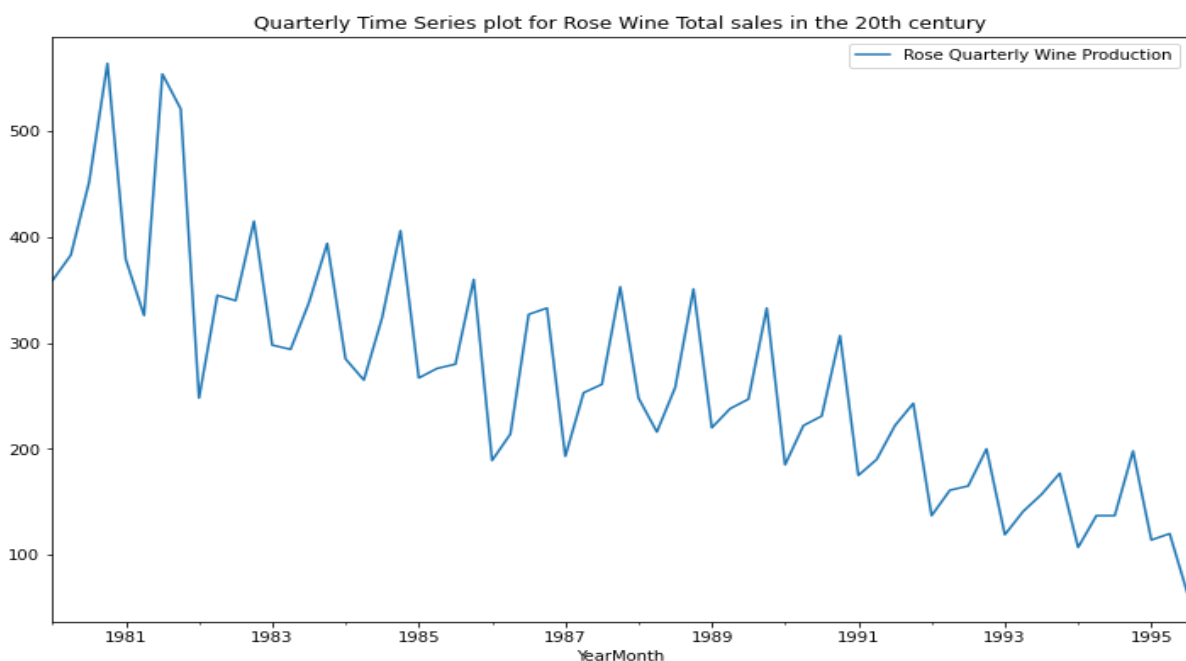
- **Conversion of Monthly time series Data into other periods:** - Since Our data is given Month-wise to convert this Data in Daily frequency, Quarterly frequency , We will use resample method of pandas data frame.

- **Quarter wise Time Series Plot:** - We have resampled Our Data to every Quarters & plot the QTR Plot as below.



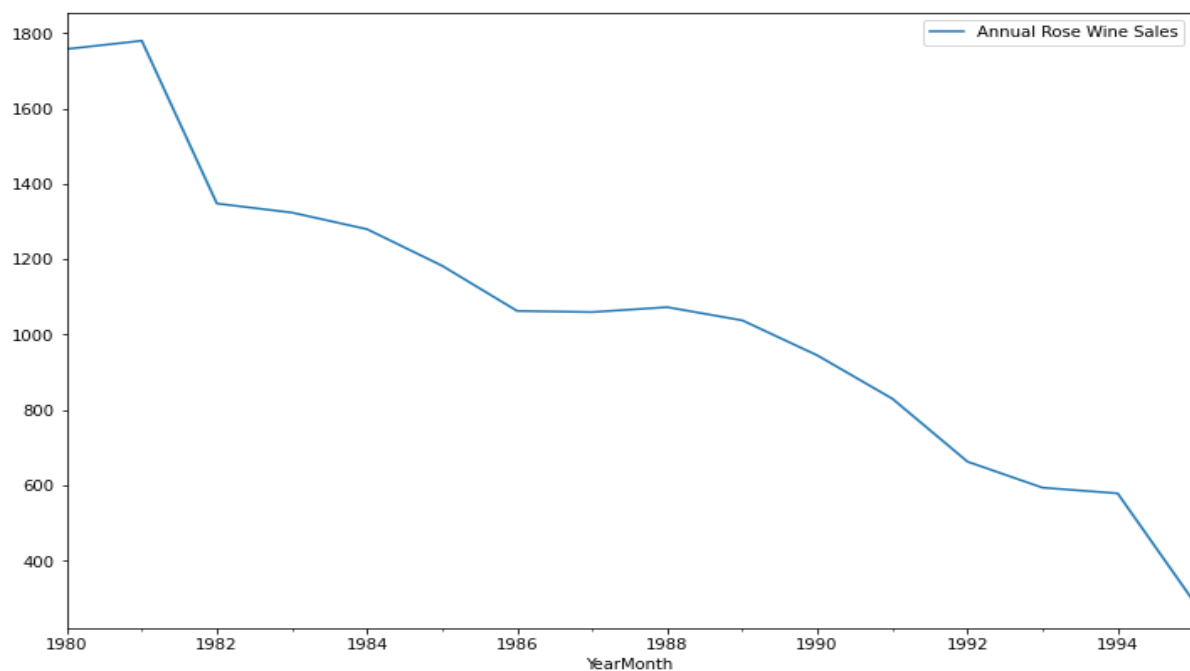
It can be observed from QTR plot Average Sales are highest in 4th QTR every year & average sale is lowest in every 1st QTR across all the years, this highest peak in 4th QTR Might be due to Christmas and other festive seasons.

- **Quarter wise Sales Time series Plot :-**



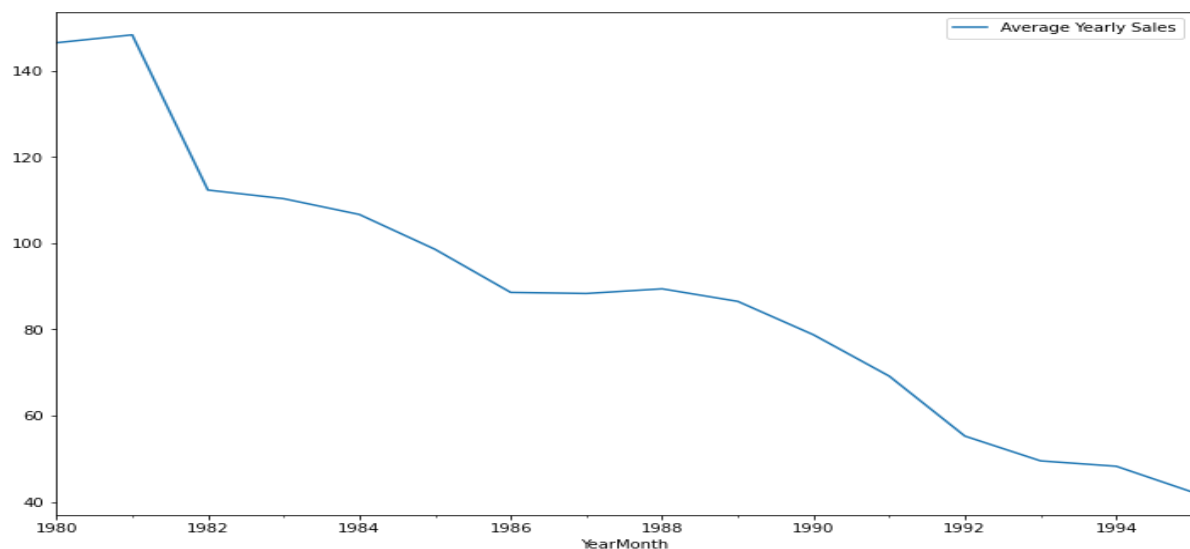
We can see from above Quarter wise time series plot that Trend in Every QTR is decreasing & yearly seasonality can be observed. As we resample data in higher frequencies seasonality will be smoothened & a clear trend can be observed.

- **Yearly Sum Time series Plot: -**



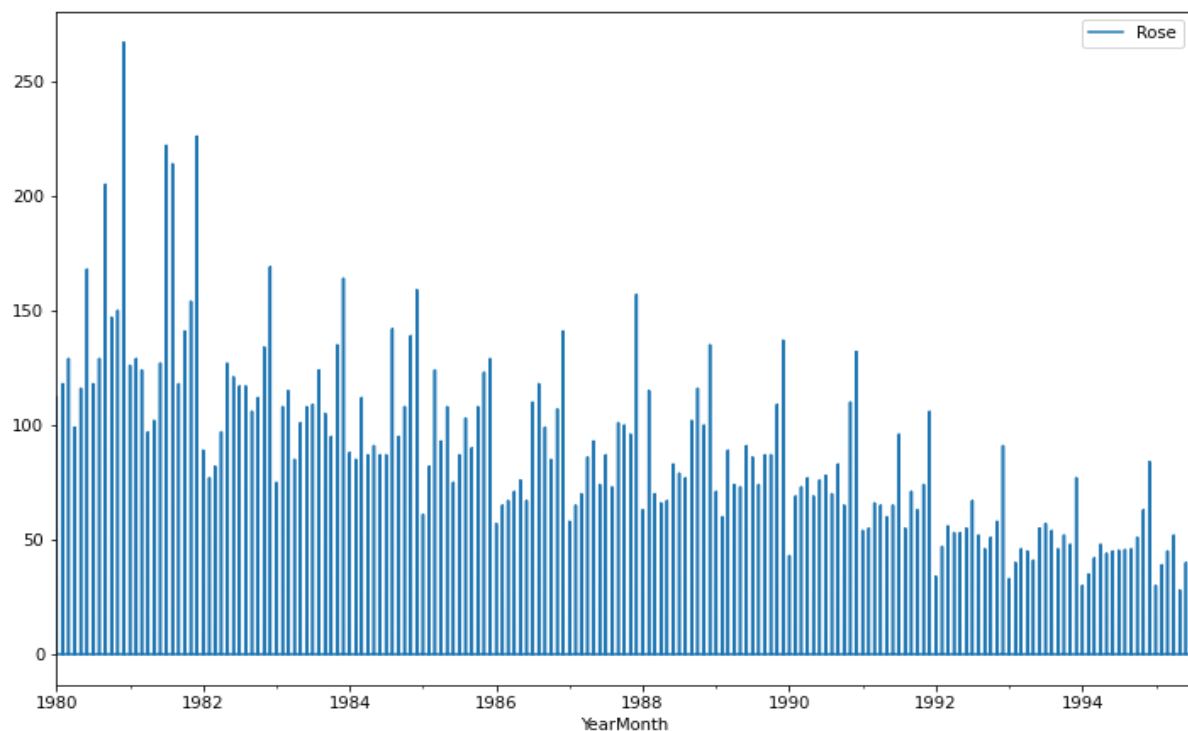
After resampling of our data as yearly we can see a downward pattern in wine sales, as noticed wine sales have dropped as the years passed.

- **Average Yearly Time series Plot: -**



The resampled annual figures have smoothened out the seasonality variations and we are able to see only the year on year trends in sales (both annual totals as well as monthly average for each year). The downward trend can be observed .

- **Daily Time Series Plot:** - After resampling our data to daily frequencies we have plotted time series plot as below. In below plot higher spikes can be seen for the Higher Daily sales for that particular bin.

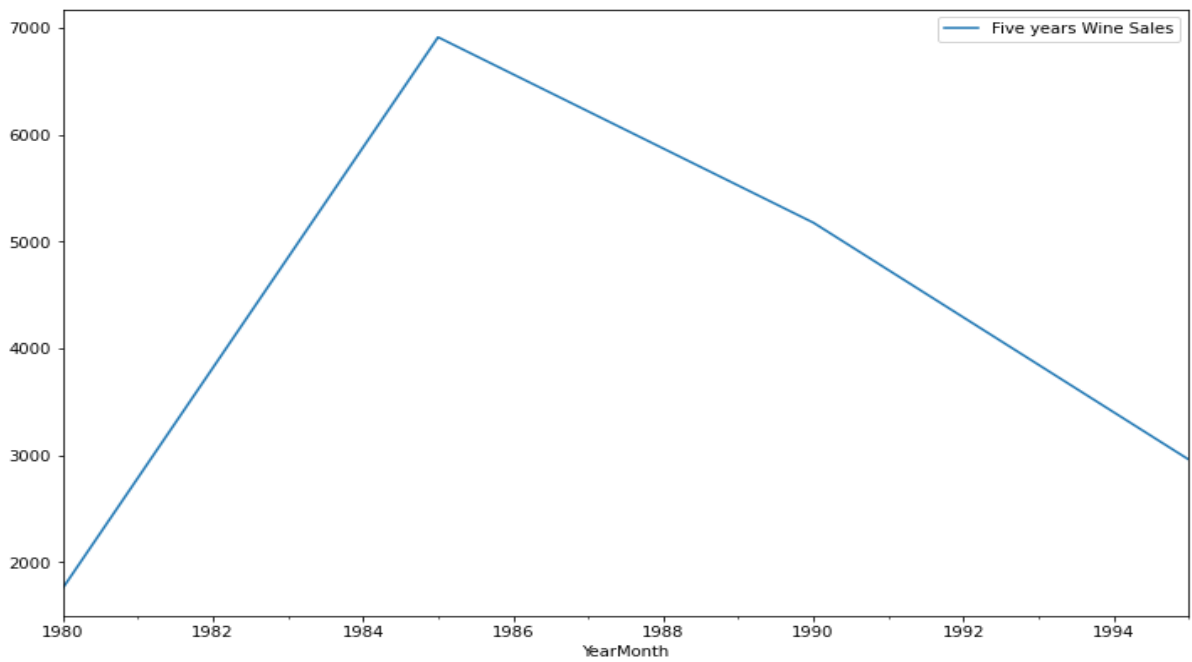


This daily graph doesn't contains much information.

- **5 year Time series Data & Plot :** -

Five years Wine Sales

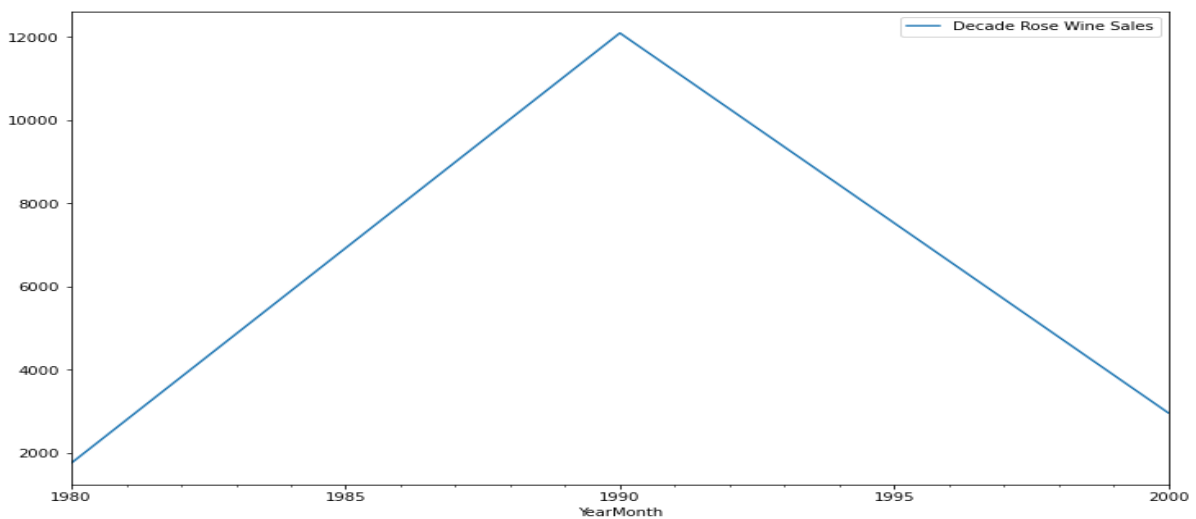
YearMonth	
1980-12-31	1758.0
1985-12-31	6915.0
1990-12-31	5179.0
1995-12-31	2962.0



From above plot it can be observed that for each 5 years' time frame up to 1985 the Rose Wine sales are having increasing trend but after 1985 there is a downward Trend.

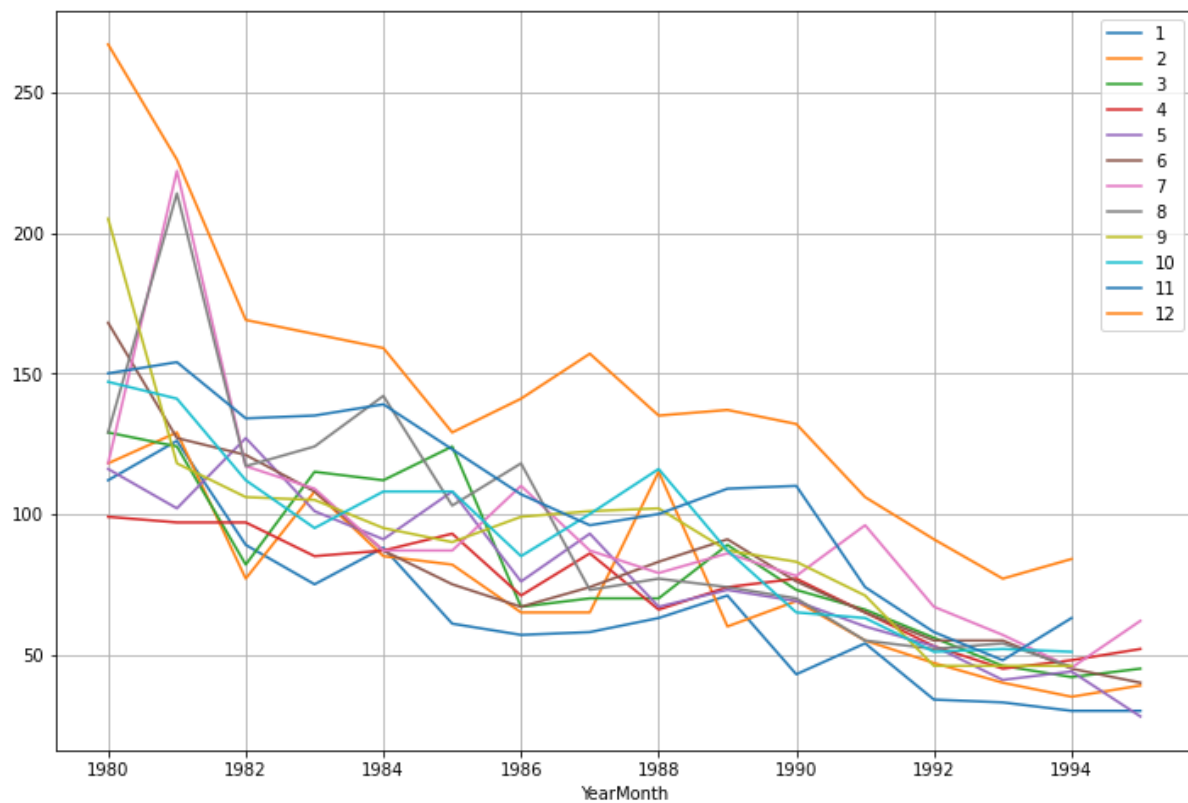
- **Wine sales time series plot across Decades: -**

Decade Rose Wine Sales	
YearMonth	
1980-12-31	1758.0
1990-12-31	12094.0
2000-12-31	2962.0



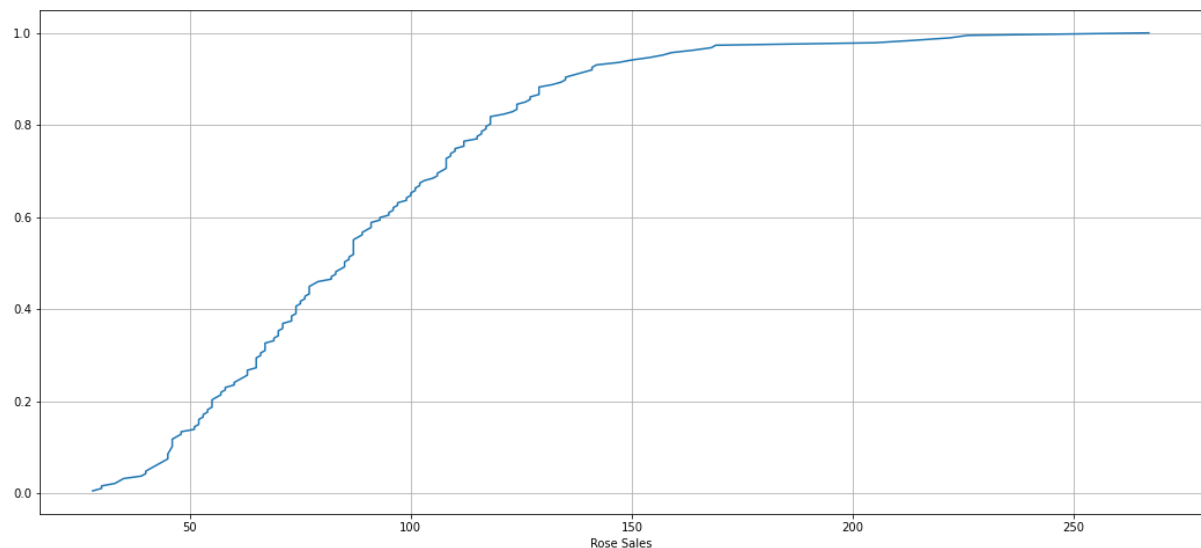
From above time series plot it can be observed that seasonal fluctuation has smoothened down & in first decade Rose Wine sales have seen a drastic Increasing trend & in second decade the sales are decreasing.

- **Monthly Trend Plot across Years: -**



As we can see from above plots that highest Monthly wine sales is recorded in the Month of December this might be due to the Christmas and New Year.

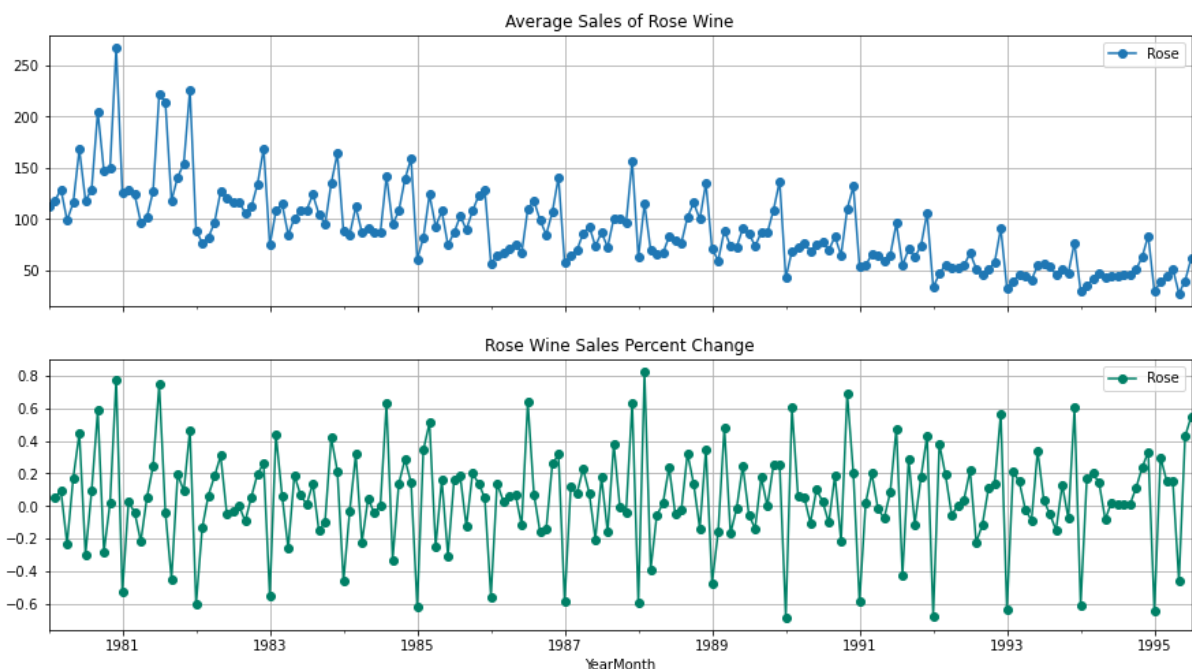
- **Empirical Cumulative Distribution plot :-**



Above particular graph tells us what percentage of data points refer to what number of Sales. In the above graph 90 % of data points just a little less than 150 units' sales & next 20 % of data points consist of next approx 100 units of rose Wine sales.

- **Plot For the average Wine Sales per month and the month on month percentage change of Rose Wine Sales: -**

The below two graphs tell us the Average 'Wine Sales' and the Percentage change of 'Wine Sales' with respect to the time.

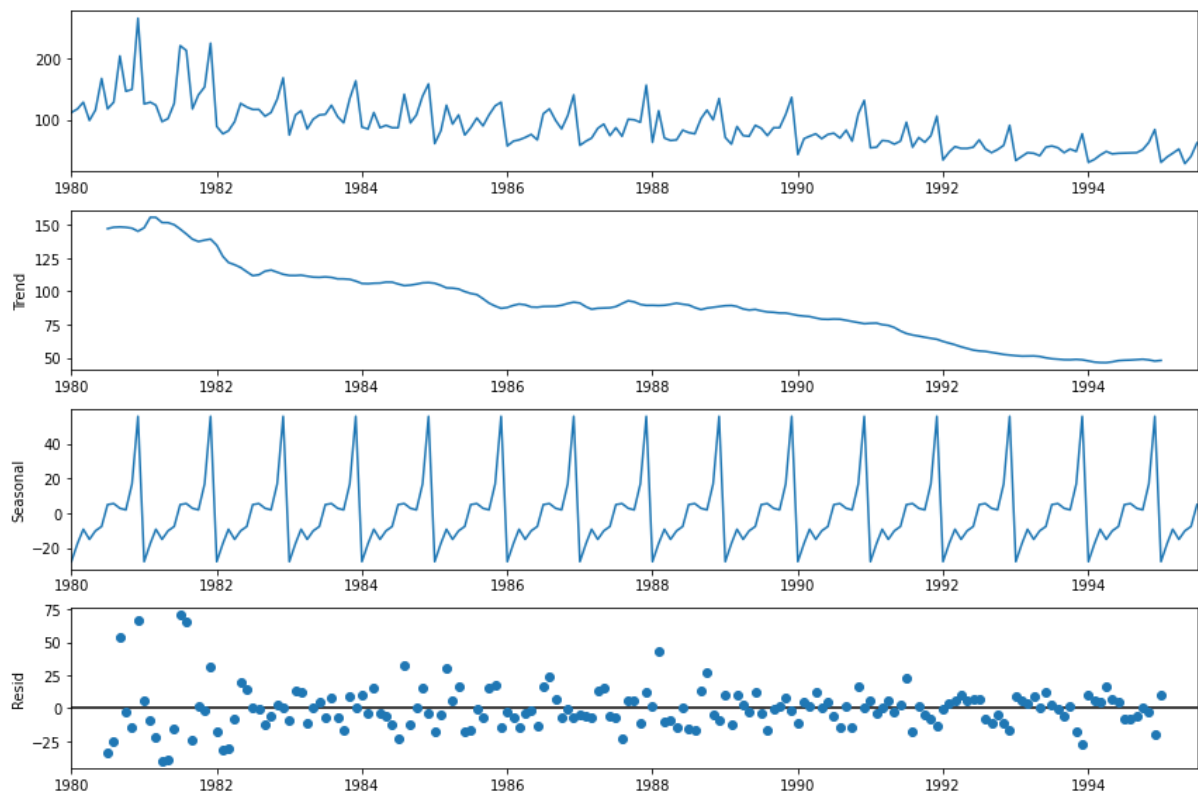


As we can see from Graph that in December 1984 the maximum sales is around 160 units & there was approx. 20 % change in sales recorded compared to November 1984. In January 1995 the change in Sales from previous month is more than -60%.

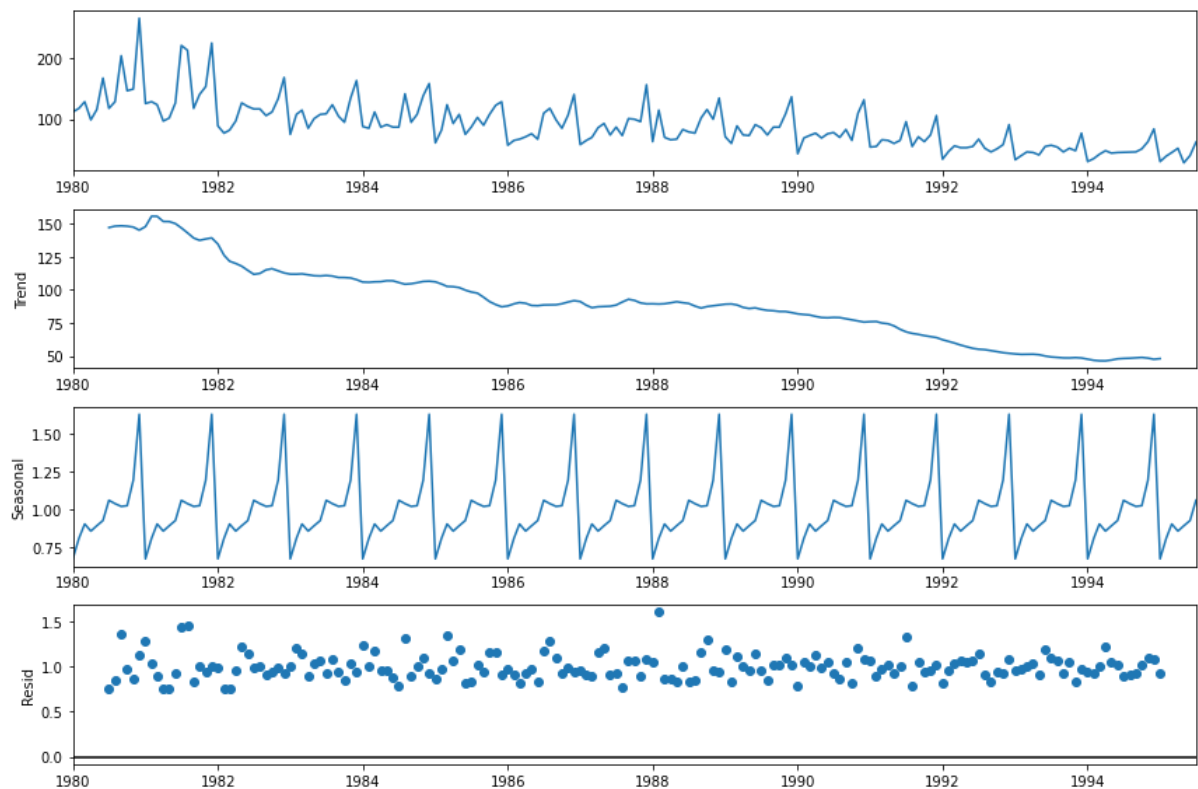
- **Time Series Decomposition:** - For Checking the Trend, Seasonality & Residuals Time we have to make Series Decomposition, this can be done in two ways Additive Decomposition & Multiplicative Decomposition.
- **Additive Decomposition:** - As per the 'additive' decomposition, we see that there is a decreasing trend. Yearly seasonality can also be seen. but if we see at residuals, a pattern can be seen in place of random distribution. So we might have to try for Multiplicative Decomposition.

Formula for Additive Time series :-

$$Y = \text{Trend} + \text{Seasonality} + \text{residuals}$$



- **Multiplicative Time Series Decomposition:** - As per the 'Multiplicative' decomposition, we see that there is decreasing trend in the Rose wine sales . There is a seasonality as well. In this residuals are randomly distributed so Wine sales data is multiplicative in nature.



Formula For Multiplicative Time Series:-

$$Y = \text{Trend} * \text{Seasonality} * \text{Residuals}$$

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

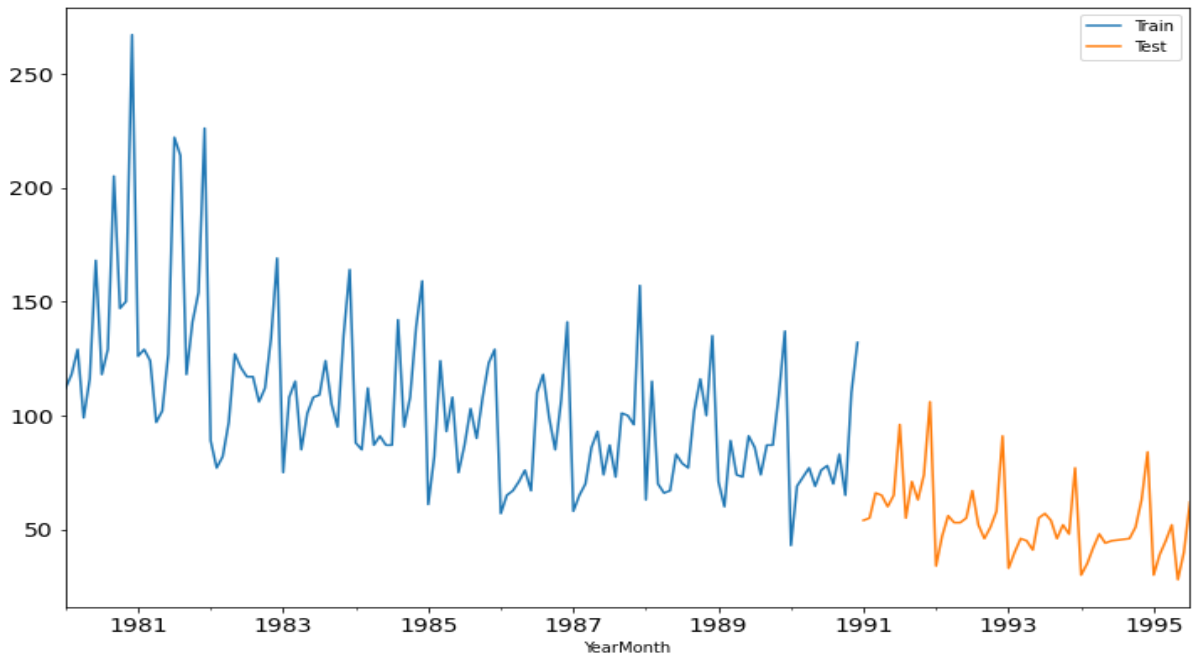
We have used following code for making train test split where Test data is started from 1991. As per shape 132 Data points are kept for the training of our model & 55 data points are kept for testing of above model.

```
: train = df[df.index.year < 1991]
: test = df[df.index.year >= 1991]

: train.shape, test.shape

: ((132, 1), (55, 1))
```

- **Train Test Split Graph:-**



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.

- **Building different models and comparing the accuracy metrics.**
- **Model 1: Linear Regression:-** In linear regression model, we are going to regress the 'Sales' variable against the order of the occurrence. For this we need to modify our training data before fitting it into a linear regression. We have to generate time instances for training and test data.

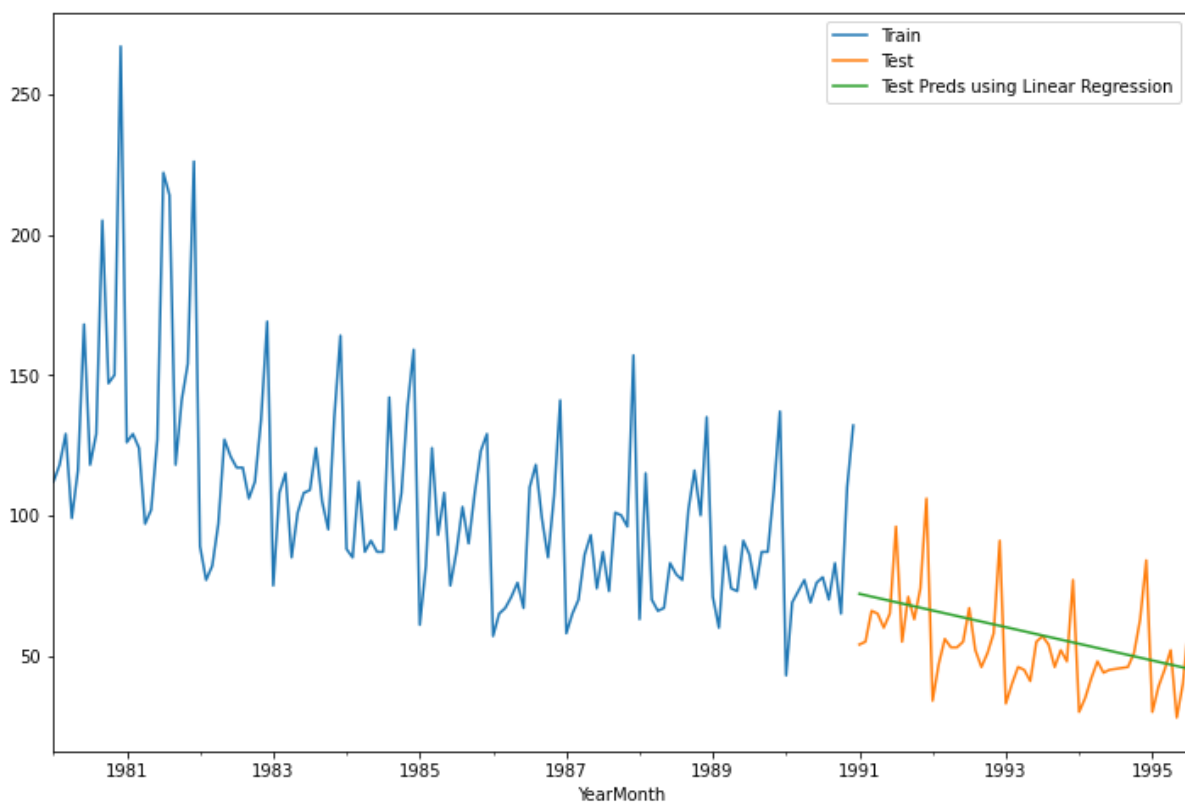
```
[61]: train_time = [i+1 for i in range(len(train))] # 1 to 132
      test_time = [i+133 for i in range(len(test))] # 133 to 187
      print('Training Time instance', '\n', train_time)
      print('Test Time instance', '\n', test_time)

Training Time instance
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132]
Test Time instance
[133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187]
```

- After that we have fitted our linear regression model against time instances variables & found the below predicted results against test data. (first few rows of prediction are shown .)

Rose time Reg On Time Instances of Rose wine data			
YearMonth			
1991-01-01	54.0	133	72.06
1991-02-01	55.0	134	71.57
1991-03-01	66.0	135	71.07
1991-04-01	65.0	136	70.58
1991-05-01	60.0	137	70.09

- Let's See Predictions Graphically.



As we know linear regression finds a best fit line against all data points, green line shows the regression line which is predicting the Rose wine sales . As from above regression line we can see that this is not a better predictor of test data, There are various errors, so we measure RMSE (Root Mean Square Errors) on test data.

- RMSE of regression

Model	Test RMSE
Reg On Time Instances of Rose wine data	15.27

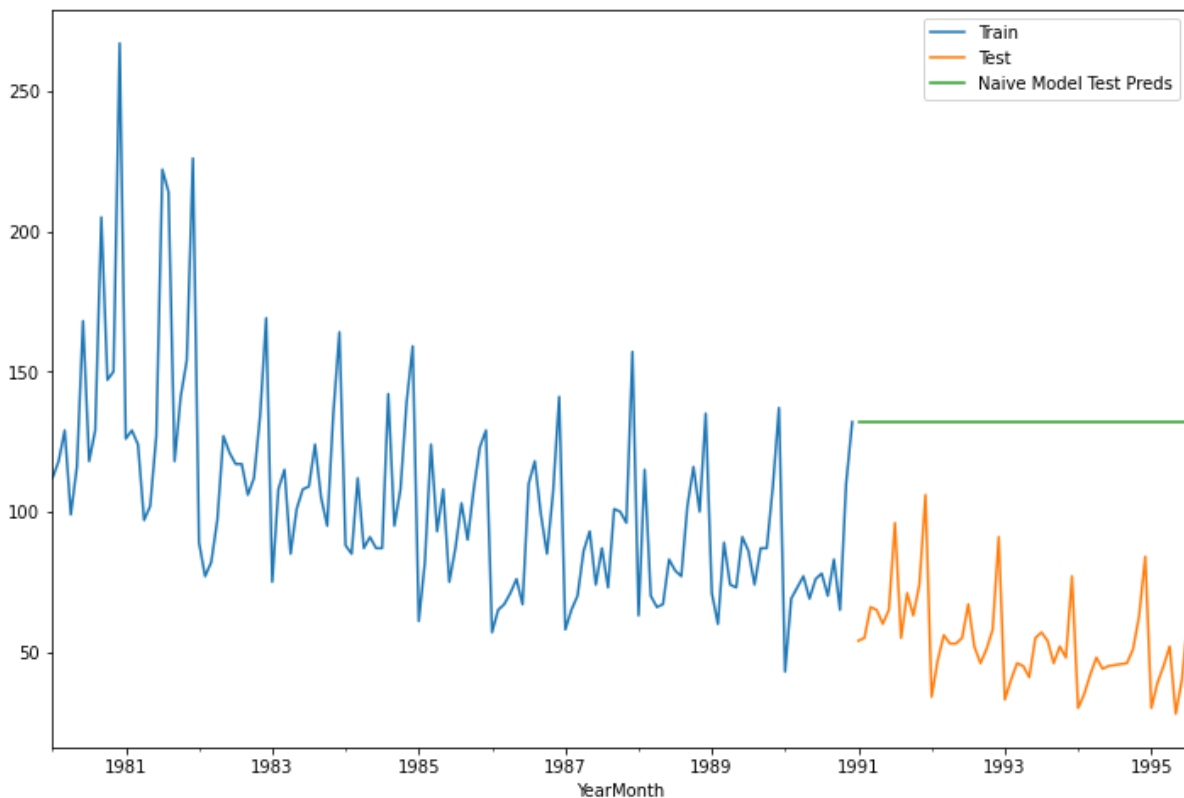
We will compare this RMSE with further Models.

- Model 2: Naive Approach:** - As Naïve approach is an Estimating technique in which the last period's actuals are used as this period's forecast, 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.

YearMonth	Prediction
01-01-1991	132.00
01-02-1991	132.00
01-03-1991	132.00
01-04-1991	132.00
01-05-1991	132.00

As we have seen that on December 1990 , there are 132 units of Rose wine sales so the sales as on Jan 1991 will also be 132 units & feb 1991 is also the same units .

This is a very simple model so we will see this graphically.



As we can see the green line is naïve model prediction and it's a flat line , It is clear from the Visuals that there is a lot of error in the prediction so will have calculated RMSE against test Data.

Model	Test RMSE
Reg On Time Instances of Rose wine data	15.27
NaiveModel	79.72

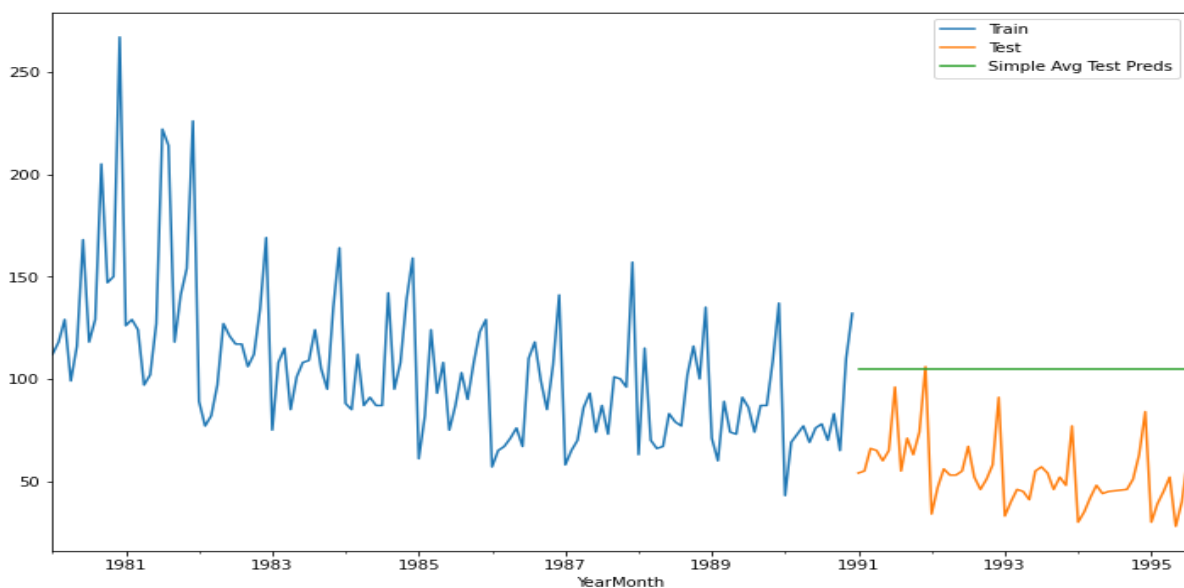
As we can see that the RMSE of Naïve model is around 3 times of linear regression so, the Naïve model is not a suitable model for this analysis.

- **Model 3: Simple Average:** - In this model we will predict the future as the average value of the training data.

YearMonth	Rose	mean_for recast
01-01-1991	54.00	104.94
01-02-1991	55.00	104.94
01-03-1991	66.00	104.94
01-04-1991	65.00	104.94
01-05-1991	60.00	104.94

As we can see that train data average value is 104.94 units, so we are forecasting the same for our next test data observations.

- Let's see Our forecast visually.



As we can see that the green line shows the average value as prediction for the next periods. This is also not best fitting with our test data, so let's see the RMSE Value.

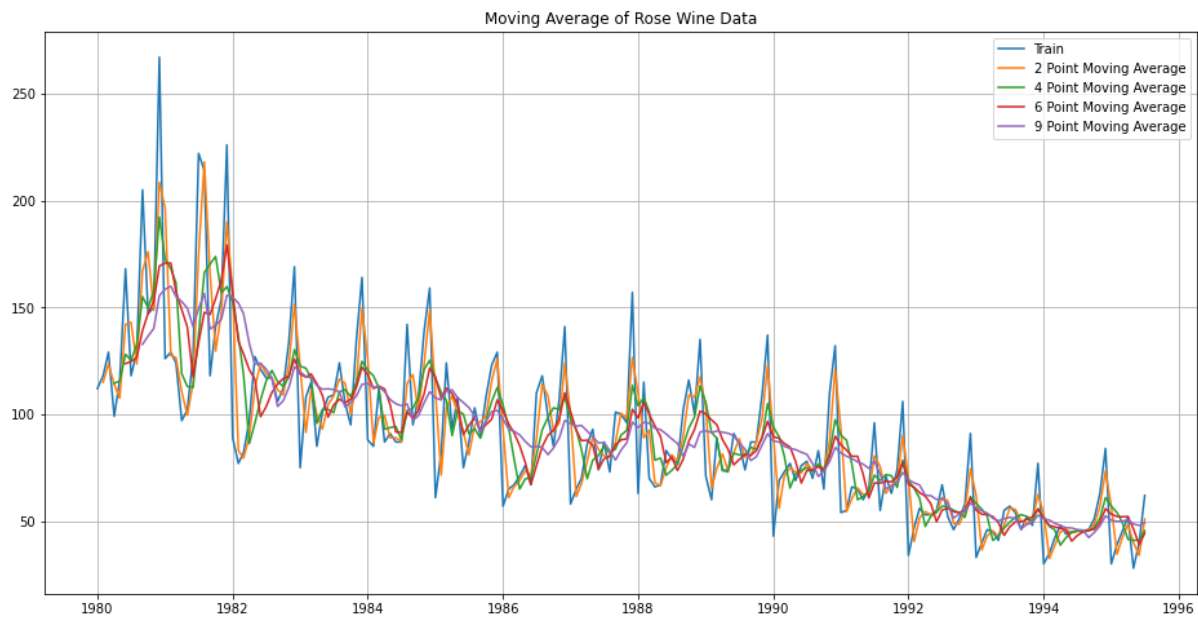
Model	Test RMSE
Reg On Time Instances of Rose wine data	15.27
NaiveModel	79.72
SimpleAverageModel	53.46

We can see from above table that RMSE is second lowest among all the three models, but it could also be better let's try other models too.

- **Model 4: Moving Average:** - In the moving average model, we will calculate rolling means at different intervals. The best interval can be determined by the minimum error. So we have calculated Moving average /Rolling means of 2 data points, 4 data points, 6 & 9 data points.

YearMonth	Sparkling	2 point Trailing Average	4 point Trailing Average	6 point Trailing Average	9 point Trailing Average
01-01-1980	112.00	NaN	NaN	NaN	NaN
01-02-1980	118.00	115.00	NaN	NaN	NaN
01-03-1980	129.00	123.50	NaN	NaN	NaN
01-04-1980	99.00	114.00	114.50	NaN	NaN
01-05-1980	116.00	107.50	115.50	NaN	NaN
01-06-1980	168.00	142.00	128.00	123.67	NaN
01-07-1980	118.00	143.00	125.25	124.67	NaN
01-08-1980	129.00	123.50	132.75	126.50	NaN
01-09-1980	205.00	167.00	155.00	139.17	132.67
01-10-1980	147.00	176.00	149.75	147.17	136.56
01-11-1980	150.00	148.50	157.75	152.83	140.11
01-12-1980	267.00	208.50	192.25	169.33	155.44
01-01-1981	126.00	196.50	172.50	170.67	158.44
01-02-1981	129.00	127.50	168.00	170.67	159.89
01-03-1981	124.00	126.50	161.50	157.17	155.00

We take rolling means and found above results, let's plot them all on training and test data as below

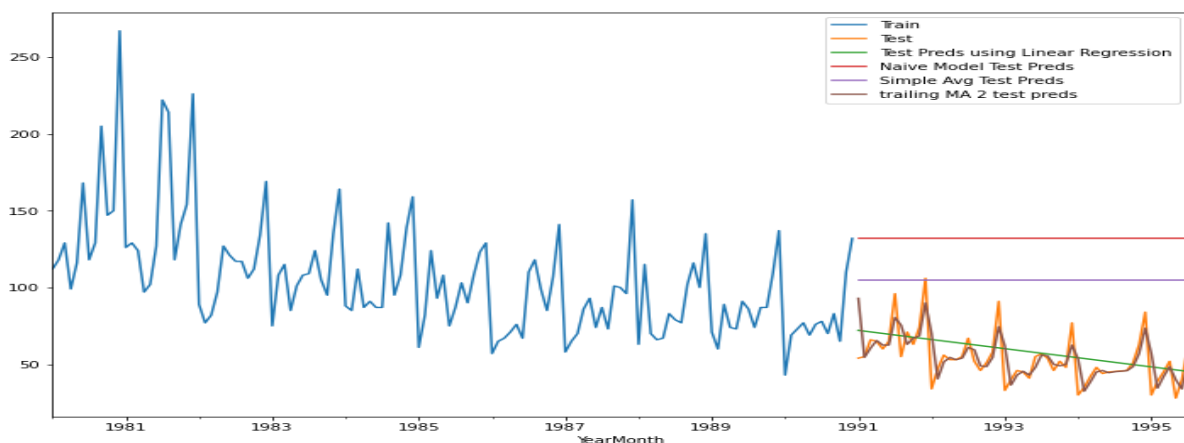


We can clearly see that 2 point moving average is best fitting with training and test data & then subsequent 4, 6 & 9 points trailings respectively. Let's Check RMSE for all the Moving averaged data.

Model	Test RMSE
Reg On Time Instances of Rose wine data	15.27
NaiveModel	79.72
SimpleAverageModel	53.46
2pointTrailingMovingAverage	11.53
4pointTrailingMovingAverage	14.45
6pointTrailingMovingAverage	14.57
9pointTrailingMovingAverage	14.73

As we can see from above table 2 point moving average is having least RMSE among all and it is also fitting best with the test data.

- Now before Moving Further Models let's plot all the above 4 Models in one plot and look at visually.



From above figure 2 point moving average model is fitting best.

- **Building of various exponential smoothing Models:** - as we have seen that from above analysis that our dataset is having all the three components level, Trend, Seasonality. So there might be Triple exponential smoothing will work out at best, for the sake of our analysis we are going to build all the Exponential smoothing models.
- **Model 5: Simple Exponential Smoothing:** - In this model we consider that our data is having only level component, there is no seasonality & trend, so we will calculate best alpha value using iterative Brute force method.

We have fitted simple exponential smoothing model to our training data, by default model we have found alpha value = 0.09 which is approximately 0.1 .

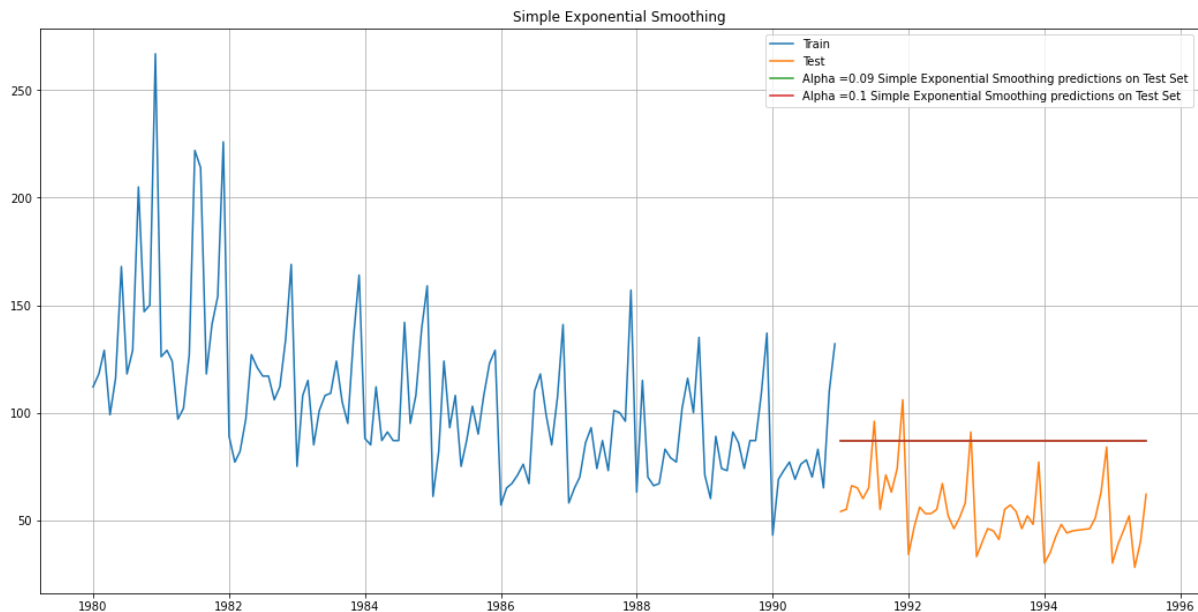
```
{'smoothing_level': 0.09874985002628338,  
'smoothing_slope': nan,  
'smoothing_seasonal': nan,  
'damping_slope': nan,  
'initial_level': 134.38726392126804,  
'initial_slope': nan,  
'initial_seasons': array([], dtype=float64),  
'use_boxcox': False,  
'lamda': None,  
'remove_bias': False}
```

We have found below prediction from base model. Here alpha value is close to zero (0.1).

	Rose	predict
YearMonth		
1991-01-01	54.0	87.104998
1991-02-01	55.0	87.104998
1991-03-01	66.0	87.104998
1991-04-01	65.0	87.104998
1991-05-01	60.0	87.104998

The above prediction gives a straight line, let's try brute force method for finding best Alpha value as per least RMSE value.

- By Brute force method SES & base SES model we have found following plot .



By brute force method we got alpha value as 0.1 as in base model it is 0.09 both values are equal , so lines are overlapping on each other , as per graph none of the Simple exponential smoothing model is best fitting with test data let's check their RMSE Values below .

Model	Test RMSE
Reg On Time Instances of Rose wine data	15.27
NaiveModel	79.72
SimpleAverageModel	53.46
2pointTrailingMovingAverage	11.53
4pointTrailingMovingAverage	14.45
6pointTrailingMovingAverage	14.57
9pointTrailingMovingAverage	14.73
Alpha=0.09,SimpleExponentialSmoothing	36.80
Alpha=0.1,SimpleExponentialSmoothing	36.83

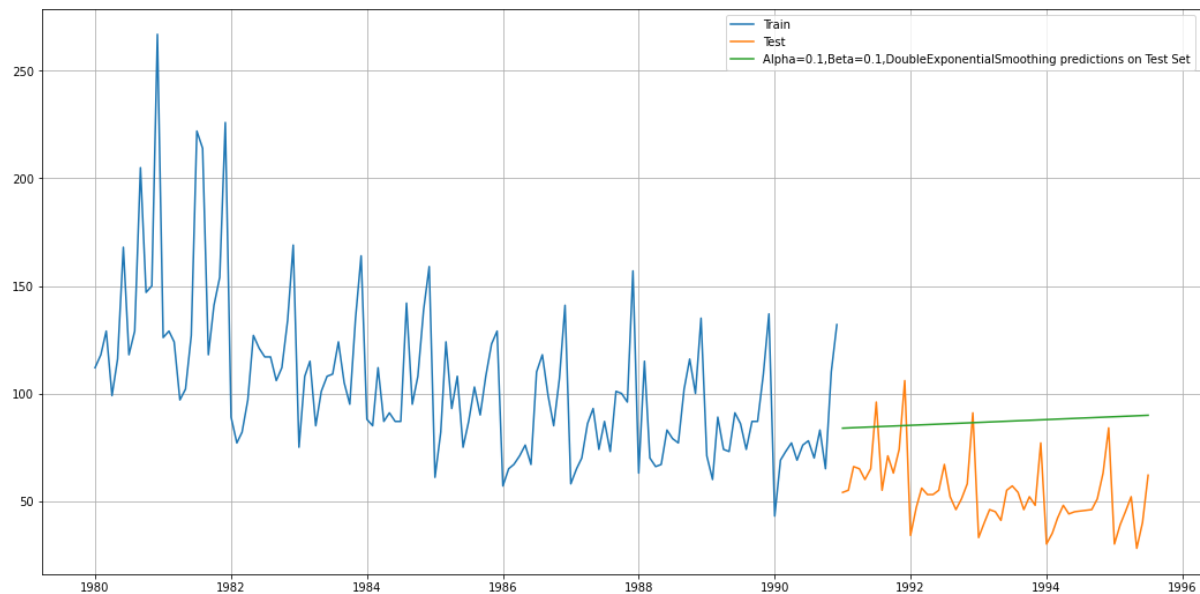
Amongst both simple exponential smoothening models alpha =0.1 value is giving comparatively less RMSE. Simple exponential smoothing is not looking good predictor of future Rose wine sales .

- **Model 6: Double Exponential Smoothing (Holt's Model):-** In Holt's model two parameters level α & trend β coefficients are estimated .

We have fitted Holt's Model & By using brute force method, we have run a loop and estimated best alpha & beta values as below.

$$\text{Level } \alpha = 0.1 \text{ \& trend } \beta = 0.1$$

Let's plot the Double exponential smoothing predictions.



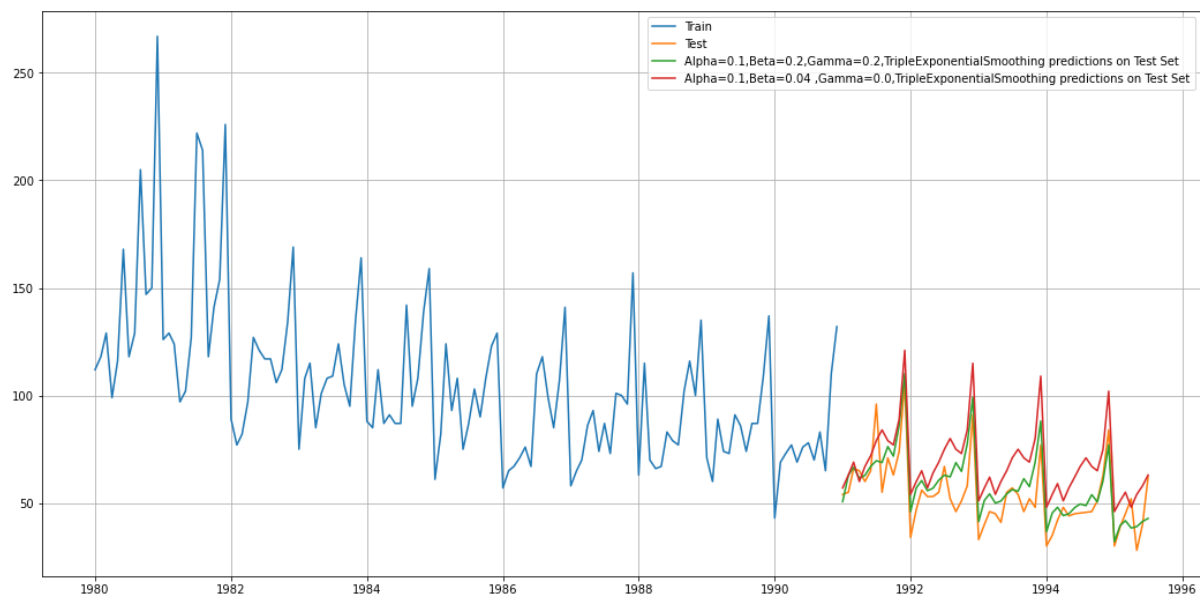
In above graph green line is showing Double exponential smoothing predictions on test sets , visually it is not best forecasting our test data of Rose Wine , let's Check corresponding RMSE.

Model	Test RMSE
Reg On Time Instances of Rose wine data	15.27
NaiveModel	79.72
SimpleAverageModel	53.46
2pointTrailingMovingAverage	11.53
4pointTrailingMovingAverage	14.45
6pointTrailingMovingAverage	14.57
9pointTrailingMovingAverage	14.73
Alpha=0.09,SimpleExponentialSmoothing	36.80
Alpha=0.1,SimpleExponentialSmoothing	36.83
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	36.92

As in last line we can see that Holt's model is giving us comparatively high value of RMSE i.e. 36.92 but it is comparable with simple exponential smoothing model & so we can say that it's not a good predictor/forecaster of our future Sales of Rose Wine.

- **Model 7: Triple Exponential Smoothing (Holt - Winter's Model) :-** This model account that the Rose wine dataset is having all the three components , Level , Trend, Seasonality too. This means Rose wine sales are dependent on parameters α , β and γ . So we will calculate all the three parameters α , β and γ here.

We have fitted Holt-Winter's Model to our training data and found below base parameters, Alpha=0.1, Beta=0.048, Gamma=0.0 & found that our RMSE is 17.92 , Now we have to try Brute force method by running various combination of values of all the three parameters . We found below values of Alpha=0.1, Beta=0.2, Gamma=0.2 .

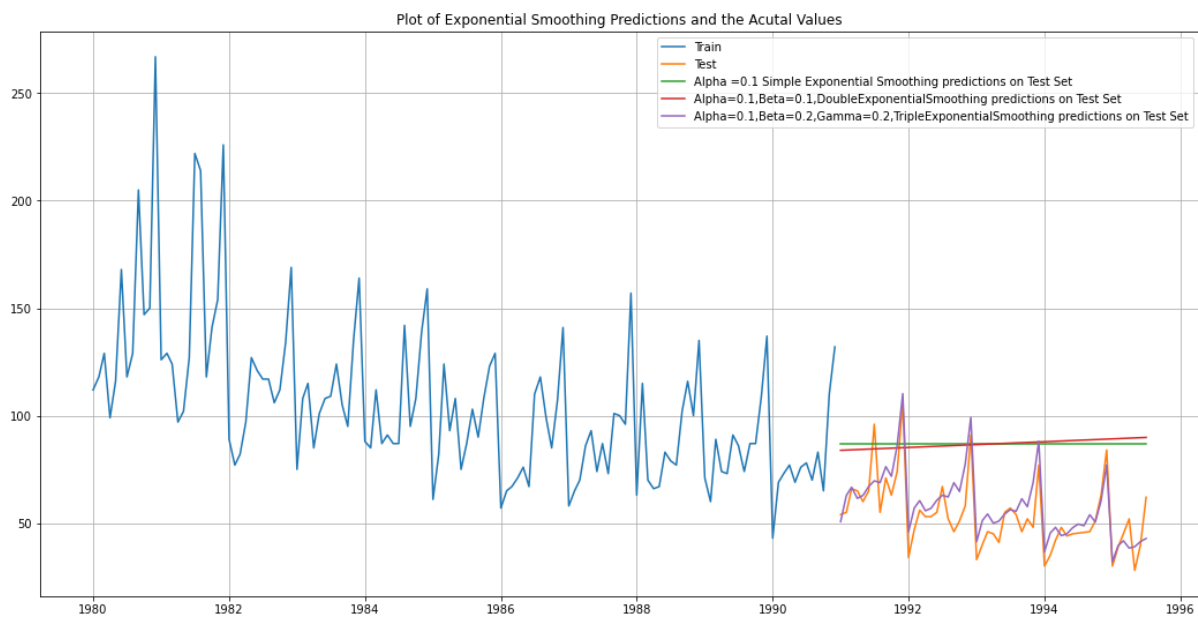


We have seen that model found by brute force method is giving lowest values of RMSE.

Model	Test RMSE
Reg On Time Instances of Rose wine data	15.27
NaiveModel	79.72
SimpleAverageModel	53.46
2pointTrailingMovingAverage	11.53
4pointTrailingMovingAverage	14.45
6pointTrailingMovingAverage	14.57
9pointTrailingMovingAverage	14.73
Alpha=0.09, Simple Exponential Smoothing	36.80
Alpha=0.1, Simple Exponential Smoothing	36.83
Alpha=0.1, Beta=0.1, Double Exponential Smoothing	36.92
Alpha=0.1, Beta=0.04, Gamma=0.0, Triple Exponential Smoothing	↑ 17.39
Brute force - Alpha=0.1, Beta=0.2, Gamma=0.2, Triple Exponential Smoothing	↓ 9.64

This means Rose wine sales depends on below best forecast parameters Alpha=0.1, Beta=0.2, Gamma=0.2 .

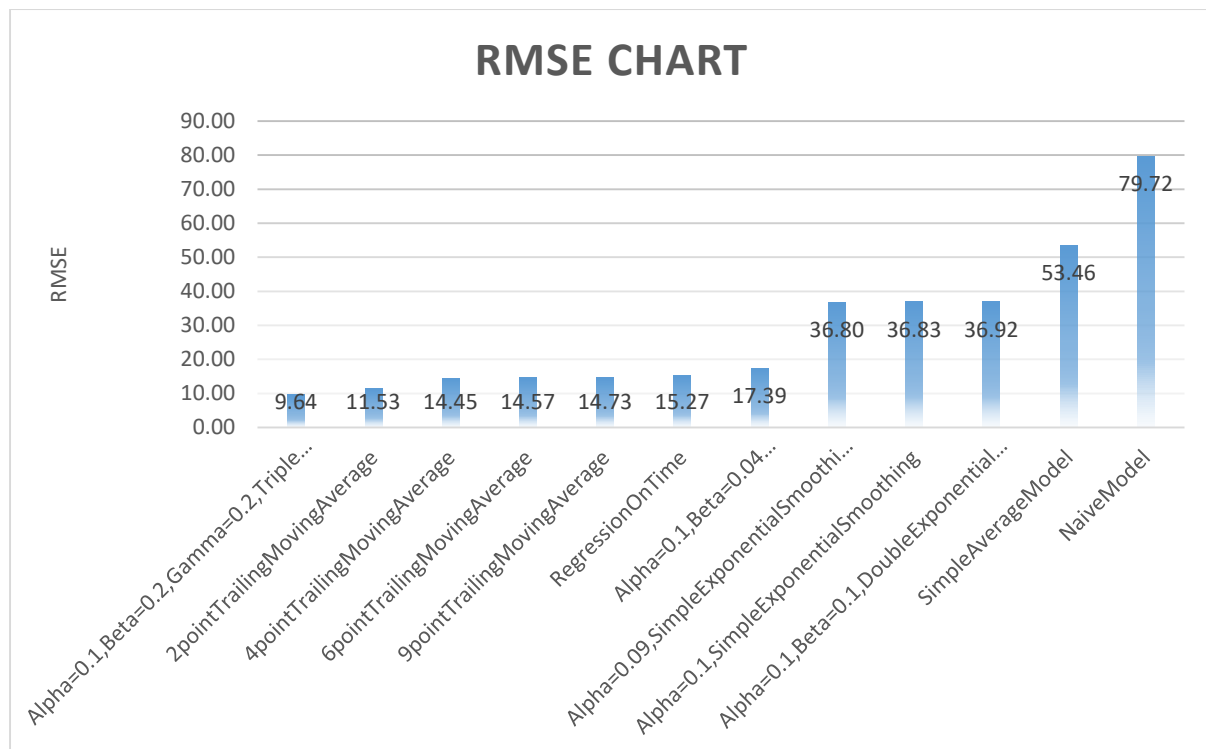
- Plotting combined all the three Smoothing Models.



We can see that amongst all smoothing model triple exponential smoothing is performing better .

- We have sorted all the Models in ascending order of RMSE .

Model	Test RMSE
Alpha=0.1,Beta=0.2,Gamma=0.2,TripleExponentialSmoothing	9.64
2pointTrailingMovingAverage	11.53
4pointTrailingMovingAverage	14.45
6pointTrailingMovingAverage	14.57
9pointTrailingMovingAverage	14.73
RegressionOnTime	15.27
Alpha=0.1,Beta=0.04 ,Gamma=0.00001,TripleExponentialSmoothing	17.39
Alpha=0.09,SimpleExponentialSmoothing	36.80
Alpha=0.1,SimpleExponentialSmoothing	36.83
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	36.92
SimpleAverageModel	53.46
NaiveModel	79.72



By considering RMSE till now our best performing model is Alpha=0.1, Beta=0.2, Gamma=0.2 Tripple exponential smoothing.

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 $\alpha = 0.05$.

We will check stationarity of the Rose wine sales data by using Augmented Dicky Fuller test on test data, below are the Null and alternate hypothesis of the test .

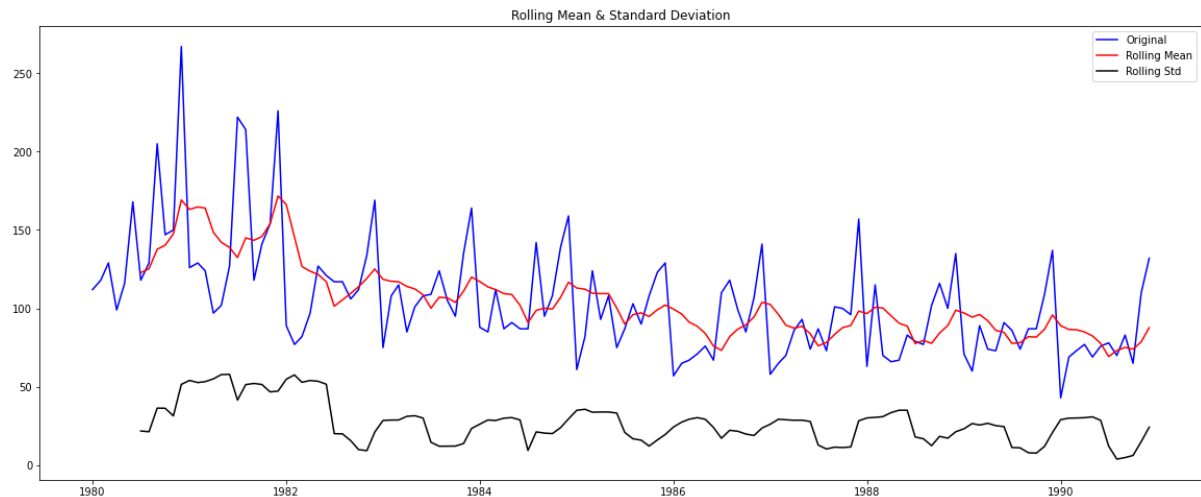
Null hypothesis H_0

: unit root is present in Rose wine sales data, i. e. time series is Not stationary .

Alternate hypothesis H_a

: unit root is not present in Rose wine sales data, i. e. time series is stationary.

- After Performing ADF Test we found below results

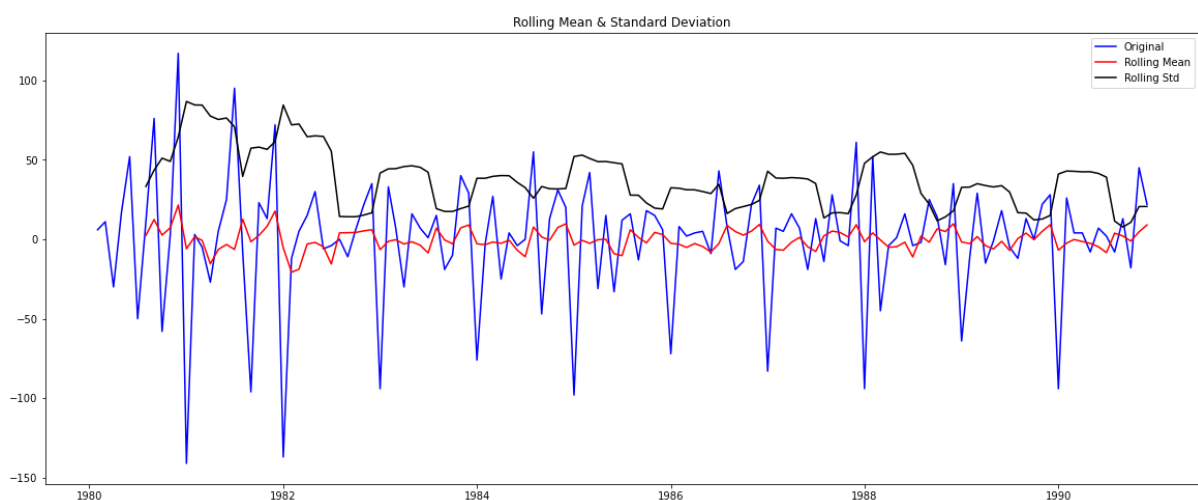


```
Results of Dickey-Fuller Test:
Test Statistic      -2.164250
p-value             0.219476
#Lags Used          13.000000
Number of Observations Used 118.000000
Critical Value (1%) -3.487022
Critical Value (5%) -2.886363
Critical Value (10%) -2.580009
dtype: float64
```

as we can see from above ADF output that p-value is 0.219 which is greater than level of significance $\alpha = 0.05$, so we failed to reject the null hypothesis thus **time series is not stationary**.

Stationarity means there should be constant mean & constant variance (No Trend & Constant Seasonality) to make the time series stationary we will take a difference of order 1 and check whether the Time Series is stationary or not.

- After taking a differencing of Order 1 and performing ADF Test again we found below results & Graph.



```
Results of Dickey-Fuller Test:
Test Statistic      -6.592372e+00
p-value             7.061944e-09
#Lags Used          1.200000e+01
Number of Observations Used  1.180000e+02
Critical Value (1%)   -3.487022e+00
Critical Value (5%)   -2.886363e+00
Critical Value (10%)  -2.580009e+00
dtype: float64
```

from above graph we can notice that in the Rose wine sales across various months & years trend pattern has gone away & a clear seasonality can be seen. Also here the P-value is less than level of significance $\alpha = 0.05$, so we reject the null hypothesis and conclude that **time series is 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.

- **Build an automated version of the ARIMA :-** We will calculate the (p,d,q) values for building an ARIMA Model ,where: p is the number of autoregressive terms, d is the number of non-seasonal differences needed for stationarity and q is order of Moving average .

We had run a loop with various ranges of values of p, d,q and according to Lowest AIC Value we have selected the best values & fit it in to the ARIMA Model .

▪ Automated ARIMA AIC Values

param	AIC
(0, 1, 2)	↓ 1276.84
(1, 1, 2)	↓ 1277.36
(1, 1, 1)	↓ 1277.78
(2, 1, 1)	↓ 1279.05
(2, 1, 2)	↓ 1279.30
(0, 1, 1)	↓ 1280.73
(2, 1, 0)	→ 1300.61
(1, 1, 0)	↑ 1319.35
(0, 1, 0)	↑ 1335.15

As we can see from the above table that lowest AIC Values is found on parameters (0,1,2) , so we will take 0 no of autoregressive terms & 1st order of non seasonal differencing and 2nd order of Moving Average.

Below are the results of fitted ARIMA model .

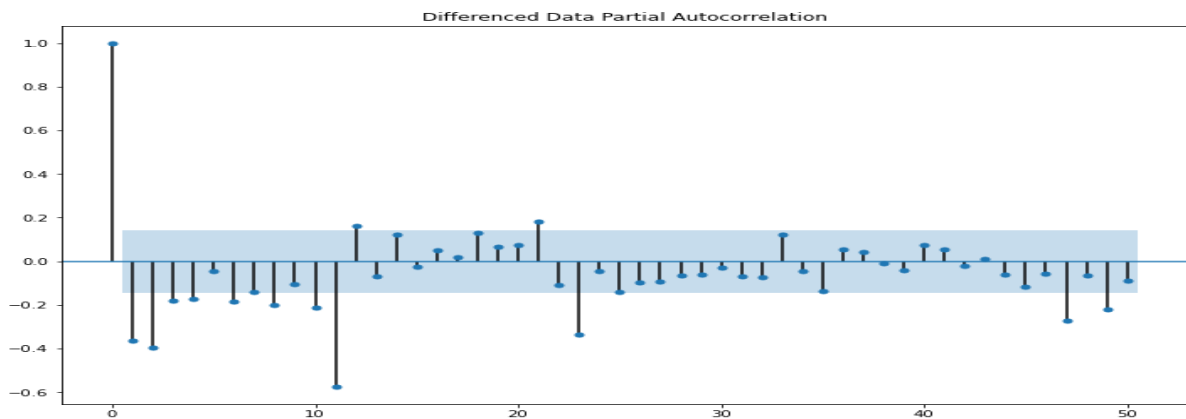
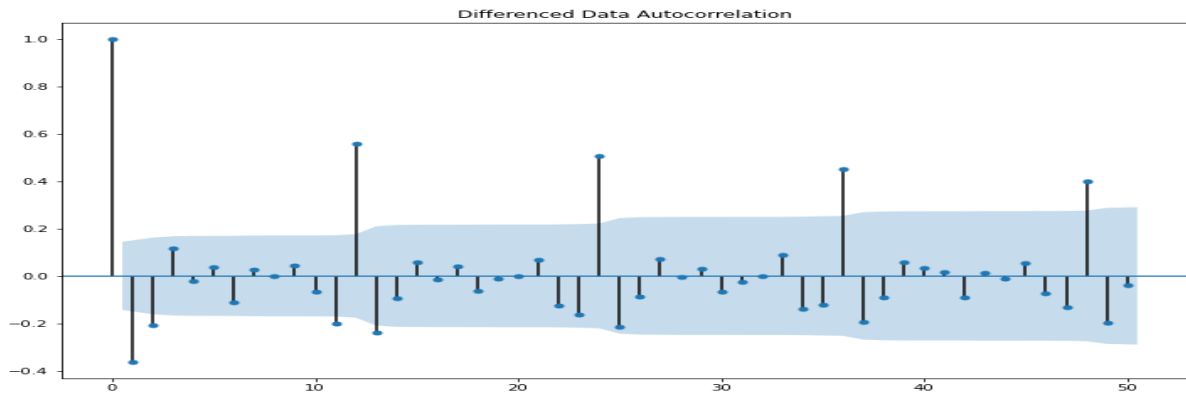
ARIMA Model Results						
Dep. Variable:	D.Rose	No. Observations:	131			
Model:	ARIMA(0, 1, 2)	Log Likelihood	-634.418			
Method:	csm-mle	S.D. of innovations	30.167			
Date:	Sun, 23 May 2021	AIC	1276.835			
Time:	00-01-1900	BIC	1288.336			
Sample:	29222	HQIC	1281.509			
	-2003					
	coef	std err	z	P> z 	[0.025	0.975]
const	-0.4885	0.085	-5.742	0	-0.655	-0.322
ma.L1.D.Rose	-0.7601	0.101	-7.499	0	-0.959	-0.561
ma.L2.D.Rose	-0.2398	0.095	-2.518	0.012	-0.427	-0.053

The light green marked are the coefficient of 2 Moving average for the Rose wine data & there is no auto regressive coefficient is present because of unavailability of AR Component , The dark green marked values are p values 2 of them are less than level of significance =0.05, So we can say that all the AR & MA components are significant.

Model	Test RMSE
ARIMA(0,1,2)	15.62

The root mean square value is 15.62 which is comparable with other models, we can try out other models too.

- **Build an automated version of the SARIMA :-** In SARIMA model we have to calculate the values of (p,d,q) (P,D,Q,s) where ,p and seasonal P: indicate number of autoregressive terms (lags of the stationarized series), d and seasonal D: indicate differencing that must be done to stationarize series, q and seasonal Q: indicate number of moving average terms (lags of the forecast errors),s: indicates seasonal length in the data.



We have noticed from the previous analysis that we have taken first order differencing for making our company data stationary & we are using an iterative brute force approach for selecting our best (p,d,q) (P,D,Q,s) parameters. Also we can notice from above plots that there will be seasonality of 12 months because significance pattern is repeating after every 12 months.

After running the loop for various combination of parameters we found below AIC Criterion sorted in ascending order .

seasonal	AIC	
(3, 1, 4)	(2, 0, 2, 12)	↓ 872.61
(2, 1, 4)	(2, 0, 2, 12)	↑ 878.33
(4, 1, 4)	(2, 0, 0, 12)	↑ 878.60
(2, 1, 3)	(2, 0, 2, 12)	↑ 879.22
(4, 1, 3)	(2, 0, 0, 12)	↑ 879.95

By choosing the best parameters i.e. (3,1,4) (2,0,2,12) , we have found below SARIMA Results.

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132			
Model:	SARIMAX(3, 1, 4)x(2, 0, [1, 2], 12)	Log Likelihood	-424.304			
Date:	Sun, 23 May 2021	AIC	872.609			
Time:	03:26:53	BIC	904.109			
Sample:	01-01-1980	HQIC	885.364			
	-2003					
Covariance Type:	opg					
	coef	std err	z	P> z	[0.025	0.975]
ar.L1	0.0214	0.266	0.08	0.936	-0.501	0.543
ar.L2	-0.2026	0.219	-0.926	0.354	-0.631	0.226
ar.L3	-0.5875	0.239	-2.457	0.014	-1.056	-0.119
ma.L1	-0.9328	163.688	-0.006	0.995	-321.755	319.889
ma.L2	0.1572	11.062	0.014	0.989	-21.524	21.839
ma.L3	0.5524	36.702	0.015	0.988	-71.382	72.487
ma.L4	-7.77E-01	1.27E+02	-0.006	0.995	2.50E+02	2.49E+02
ar.S.L12	3.79E-01	1.13E-01	3.36	0.001	1.58E-01	5.99E-01
ar.S.L24	0.3025	0.092	3.279	0.001	0.122	0.483
ma.S.L12	0.0429	0.167	0.257	0.798	-0.285	0.371
ma.S.L24	-0.1001	0.144	-0.696	0.486	-0.382	0.182
sigma2	227.3685	37200	0.006	0.995	-72700	73200

We haven't used any exogenous variables here , Yellow marked values are coefficients of Auto Regression & Moving Average terms . Green marked values shows that these terms are significant.

We have predicted the following on test data after that we check this on test data & calculate the RMSE value.

Rose	mean	mean_se	mean_ci_lower	mean_ci_upper
1991-01-01	61.545619	15.152003	31.848239	91.242999
1991-02-01	73.808659	15.224329	43.969523	103.647795
1991-03-01	75.096241	15.232328	45.241426	104.951055
1991-04-01	73.484951	15.480564	43.143603	103.826300
1991-05-01	70.188128	15.491703	39.824949	100.551307

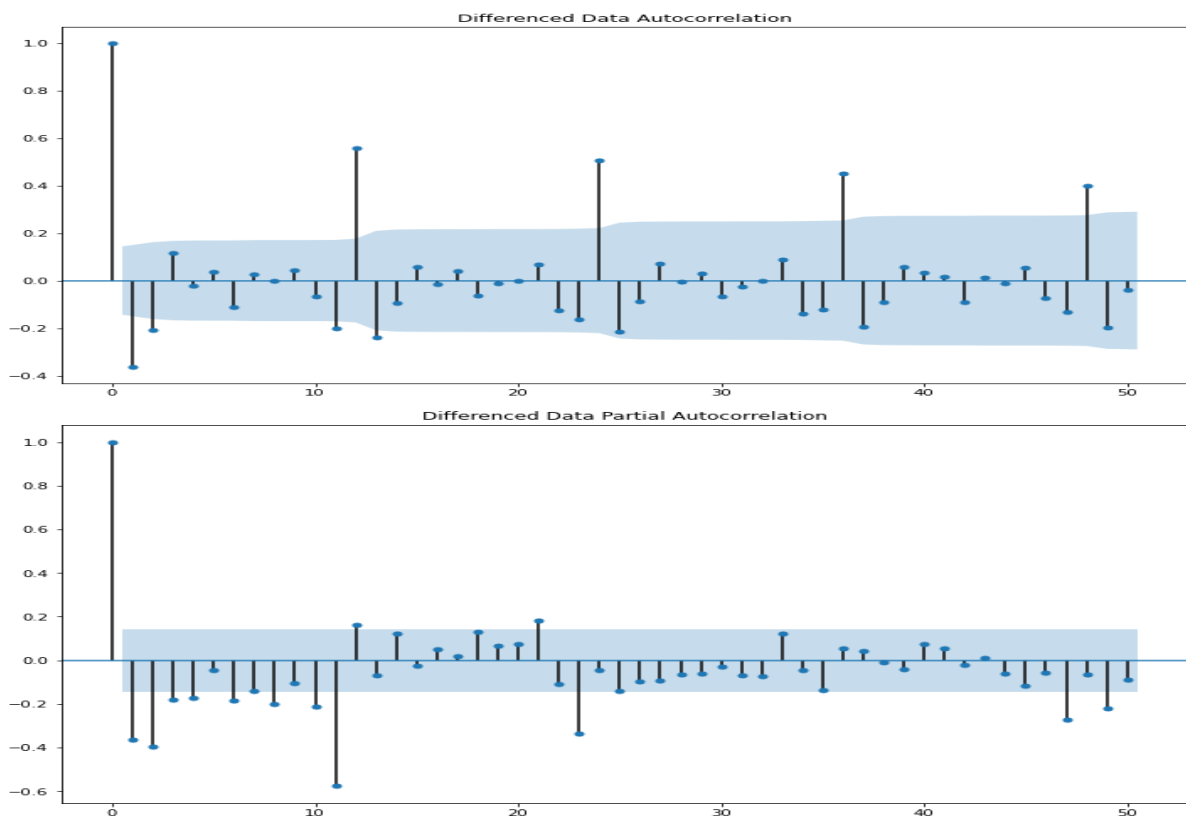
- SARIMA RMSE:-

Model	Test RMSE
AUTO ARIMA(0,1,2)	↓ 15.62
AUTO SARIMA(3,1,4)(2,0,2,12)	↑ 25.74

We haven't gain much on RMSE , so SARIMA model may be better forecaster for future Rose wine sales.

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.

- Build a version of the ARIMA model for which the best parameters are selected by looking at the ACF and the PACF plots.
- As we have noticed that Rose wine time series data is not stationary, in previous steps we have taken 1 st order differencing to make our data stationary.
- Now let's have a closer look on ACF & PACF Plots.



- The Auto-Regressive parameter is 'p' which comes from the significant lag before which the PACF plot cuts-off is 4.
- The Moving-Average parameter is 'q' which comes from the significant lag before the ACF plot cuts-off is 2.
- Order of differencing is 1.

- Now from the (p,d,q) -> (4,1,2) values ,we have built our Arima Model and found below summary results.

Dep. Variable:	D.Rose	No. Observations:	131
Model:	ARIMA(4, 1, 2)	Log Likelihood	-633.876
Method:	css-mle	S.D. of innovations	29.793
Date:	Sat, 22 May 2021	AIC	1283.753
Time:	00-01-1900	BIC	1306.754
Sample:	29222	HQIC	1293.099
	-2003		

	coef	std err	z	P> z	[0.025	0.975]
const	-0.1905	0.576	-0.331	0.741	-1.319	0.938
ar.L1.D.Rose	1.1685	0.087	13.391	0	0.997	1.34
ar.L2.D.Rose	-0.3562	0.132	-2.693	0.007	-0.616	-0.097
ar.L3.D.Rose	0.1855	0.132	1.402	0.161	-0.074	0.445
ar.L4.D.Rose	-0.2227	0.091	-2.443	0.015	-0.401	-0.044
ma.L1.D.Rose	-1.9506	nan	nan	nan	nan	nan
ma.L2.D.Rose	1.00E+00	nan	nan	nan	nan	nan
Roots						
	Real	Imaginary	Modulus	Frequency		
AR.1	1.1027	-0.4116j	1.177	-0.0569		
AR.2	1.1027	+0.4116j	1.177	0.0569		
AR.3	-0.6862	-1.6643j	1.8003	-0.3122		
AR.4	-0.6862	+1.6643j	1.8003	0.3122		
MA.1	0.9753	-0.2209j	1	-0.0355		
MA.2	0.9753	+0.2209j	1	0.0355		

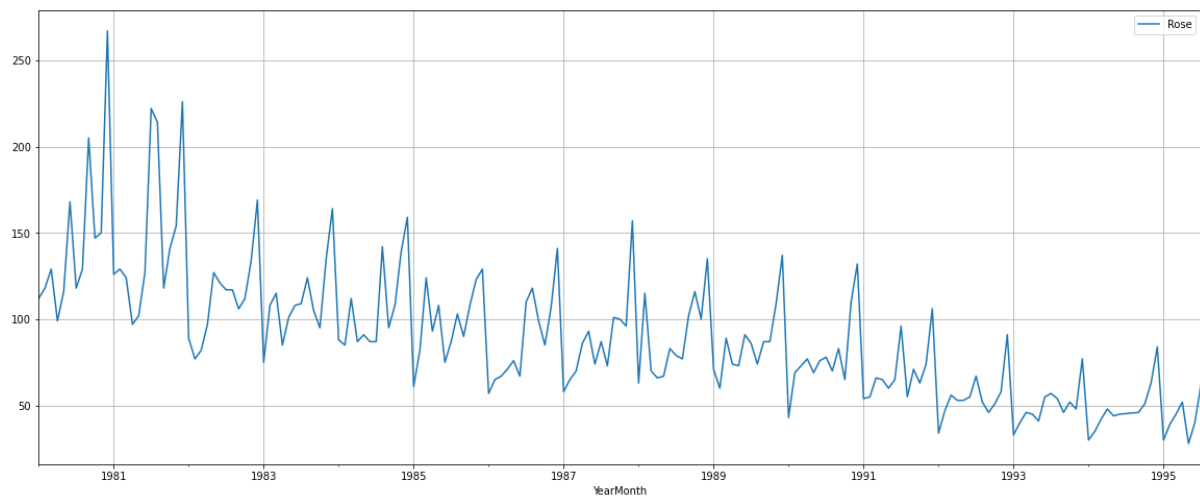
- From above output we have noticed that our AIC value is 1283, which is comparable with other models , also we can notice that except Moving average L2- lag 2 component all the p values are not significant . for each Auto regressive and Moving average all the coefficient are given.
- Now we have Predicted on the Test Set using Manual Arima model and evaluated the model using RMSE , Below are the results (refer jupyter Notebook for Coding)

Model	Test RMSE
AUTO ARIMA(0,1,2)	↓ 15.62
AUTO SARIMA(3,1,4)(2,0,2,12)	→ 25.74
Manual ARIMA(4,1,2)	↑ 33.95

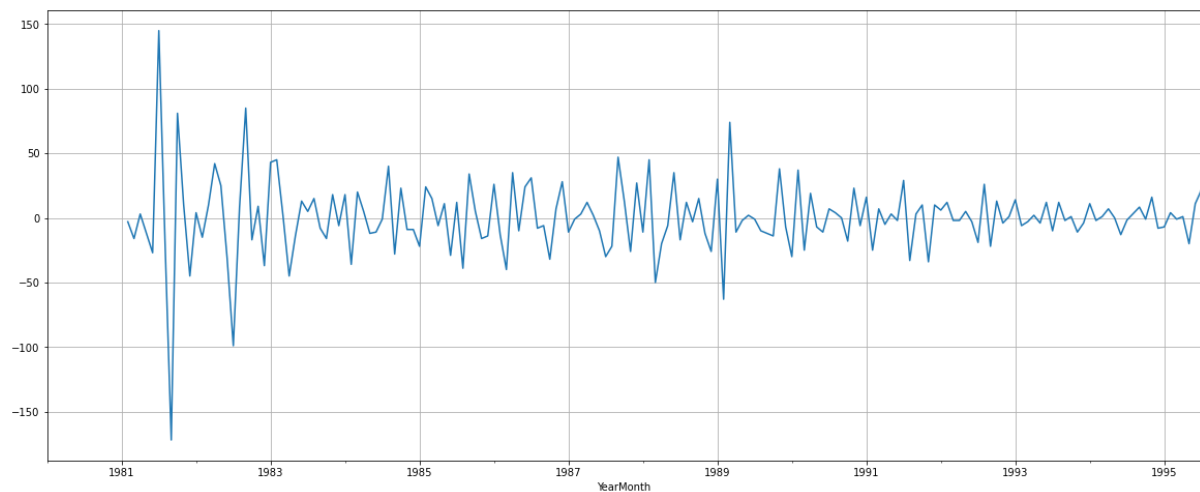
This RMSE Value marked in green is higher than our automated ARIMA & SARIMA Models.

- **Building a SARIMA model for which the best parameters are selected by looking at the ACF and the PACF plots.**

We see from our previous ACF plot at the seasonal interval (12) significance is repeating. So, we take a seasonal differencing of the original series. Before that let us look at the original series.

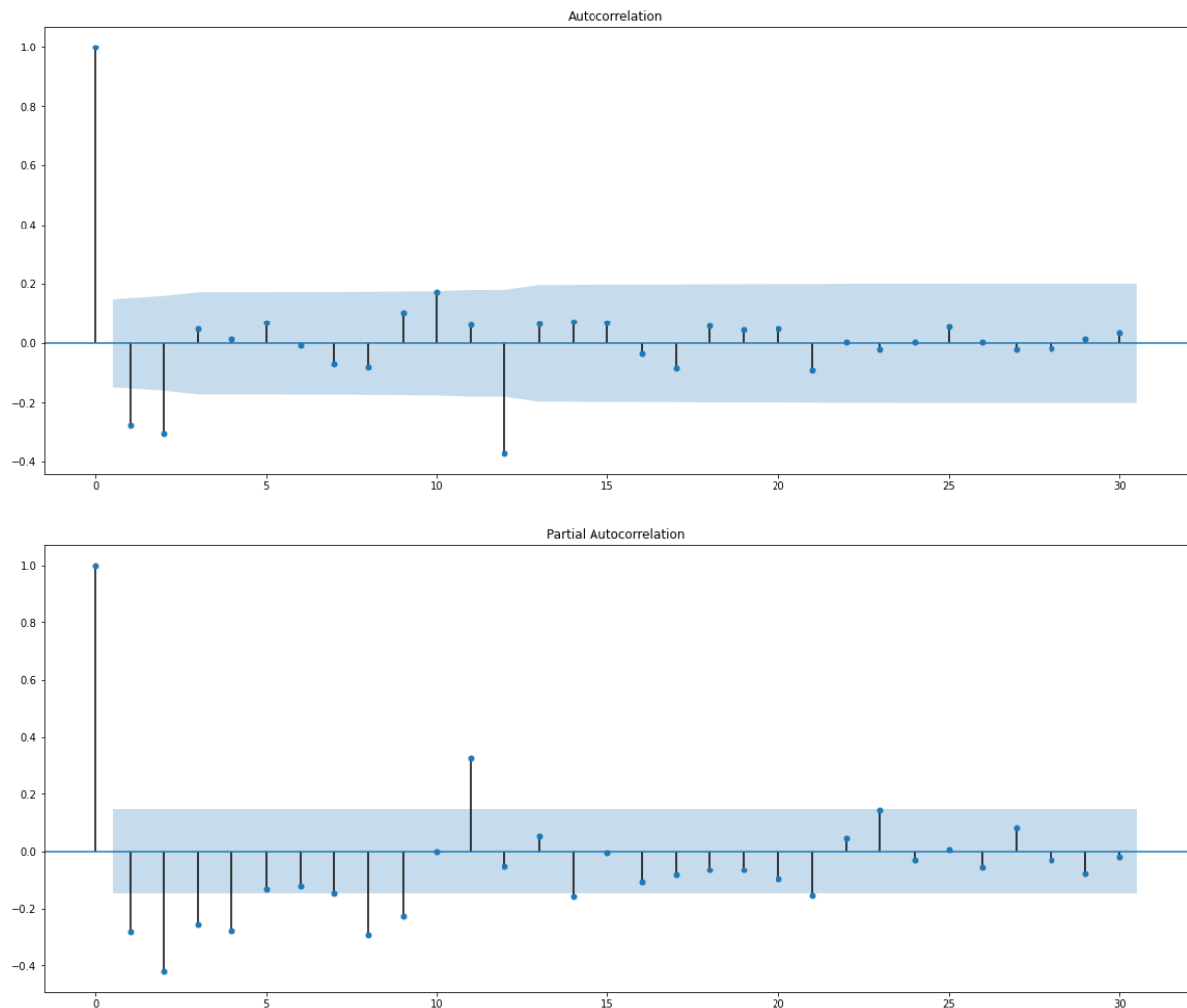


Now looking at differenced series at seasonal interval 12 below we can notice that there is almost no trend present only seasonality present in the data.



After above steps we have checked the test stationarity after differencing order 1, we found that our time series is stationary.

- Now Let's check the PACF & ACF Plots of new modified time series



- The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off is 0, The Moving-Average parameter in an SARIMA model is 'Q' which comes from the significant lag after which the ACF plot cuts-off is 1 we have to check the ACF and the PACF plots only at multiples of 12 (since 12 is the seasonal period).





Manual SARIMA Results			
Dep. Variable:	Rose	No. Observations:	132
Model:	SARIMAX(4, 1, 2)x(0, 1, [1], 12)	Log Likelihood	-446.102
Date:	Sun, 23 May 2021	AIC	908.203
Time:	03:57:27	BIC	929.358
Sample:	01-01-1980	HQIC	916.774
	-2003		

Covariance Type:	opg					
	coef	std err	z	P> z 	[0.025	0.975]
ar.L1	-0.8045	0.119	-6.775	0	-1.037	-0.572
ar.L2	0.039	0.14	0.277	0.781	-0.236	0.314
ar.L3	-0.2307	0.147	-1.566	0.117	-0.519	0.058
ar.L4	-0.1873	0.108	-1.74	0.082	-0.398	0.024
ma.L1	0.1433	138.57	0.001	0.999	-271.45	271.736
ma.L2	-0.8567	118.727	-0.007	0.994	-233.558	231.844
ma.S.L12	-5.41E-01	8.50E-02	-6.385	0	-7.07E-01	-3.75E-01
sigma2	296.7842	4.11E+04	0.007	0.994	8.03E+04	8.09E+04

Ljung-Box (Q):	30.27	Jarque-Bera (JB):	0.03
Prob(Q):	0.87	Prob(JB):	0.98
Heteroskedasticity (H):	0.55	Skew:	-0.02
Prob(H) (two-sided):	0.08	Kurtosis:	3.07

From above output we have noticed that our AIC value is 908, which is comparatively better with other models, also we can notice that except Moving average L1- lag 1 component all the p values are not significant. For each Auto regressive and Moving average all the coefficients are given marked in yellow.

Now we have predicated wine sales against the test data and found that the below RMSE.

Model	Test RMSE
AUTO ARIMA(0,1,2)	 15.62
AUTO SARIMA(3,1,4)(2,0,2,12)	 25.74
Manual ARIMA(4,1,2)	 33.95
Manual SARIMA(4,1,2)(0,1,1,12)	 15.91

We have gained a very good RMSE value which is comparable with Auto ARIMA Model & this Model might be one of the best model to predict the future wine sales.

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 have built a data frame of all model built with their descending order of RMSE as below.

Model	Test RMSE
Alpha=0.1,Beta=0.2,Gamma=0.2,TripleExponential Smoothing	↓ 9.64
2pointTrailingMovingAverage	↓ 11.53
4pointTrailingMovingAverage	↓ 14.45
6pointTrailingMovingAverage	↓ 14.57
9pointTrailingMovingAverage	↓ 14.73
Reg On Time Instances of Rose wine data	↓ 15.27
ARIMA(0,1,2)	↓ 15.62
Manual SARIMA(4,1,2)(0,1,1,12)	↓ 15.91
Alpha=0.1,Beta=0.04, Gamma=0.00001, TripleExponentialSmoothing	↓ 17.39
SARIMA(3,1,4)(2,0,2,12)	↓ 25.74
Manual ARIMA(4,1,2)	→ 33.95
Alpha=0.09,SimpleExponentialSmoothing	→ 36.80
Alpha=0.1,SimpleExponentialSmoothing	→ 36.83
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	→ 36.92
SimpleAverageModel	→ 53.46
NaiveModel	↑ 79.72

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.

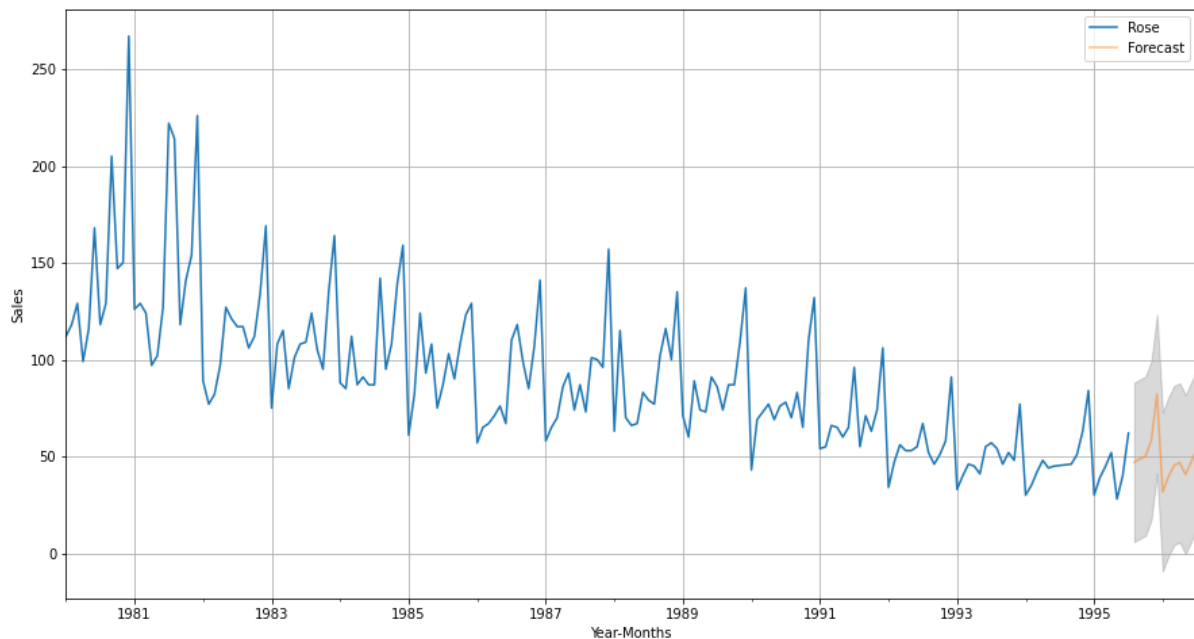
As we can see from above RMSE Table that Alpha=0.1, Beta=0.2, Gamma=0.2, TripleExponentialSmoothing having lowest RMSE of 9.64 So it is the best Model

- **By using Alpha=0.1, Beta=0.9, Gamma=0.6, Triple Exponential Smoothing: - Prediction for Next 12 Months.**
 - We have fitted Our Model to Full dataset of Rose wine sales by passing Trend as multiplicative & seasonality as multiplicative and found the Full Model RMSE -> 20.98
 - Next We have predicted for the Next 12 Months along with 95 %Confidence Intervals as below

YearMonth	lower_CI	prediction	upper_ci
01-08-1995	5.87	46.98	88.10
01-09-1995	7.53	48.65	89.76
01-10-1995	9.06	50.18	91.29
01-11-1995	17.11	58.22	99.34
01-12-1995	41.10	82.21	123.33
01-01-1996	-9.43	31.69	72.80
01-02-1996	-1.68	39.44	80.56
01-03-1996	4.27	45.39	86.50
01-04-1996	5.60	46.72	87.83
01-05-1996	-0.47	40.65	81.76
01-06-1996	6.10	47.22	88.33
01-07-1996	12.54	53.66	94.77

From above table we can see that predictions are marked in Green & Lower Confidence Interval Marked in red, Upper Confidence Intervals are Marked in Blue, That means our model is 95% confident that the prediction will lies in this range. let's see Our predictions graphically.

We can see that in December 1995 Our predicted sales are touching above 100 units band.



- **Prediction for next 12 Month using Manual SARIMA (4,1,2) (0,1,1,12)**

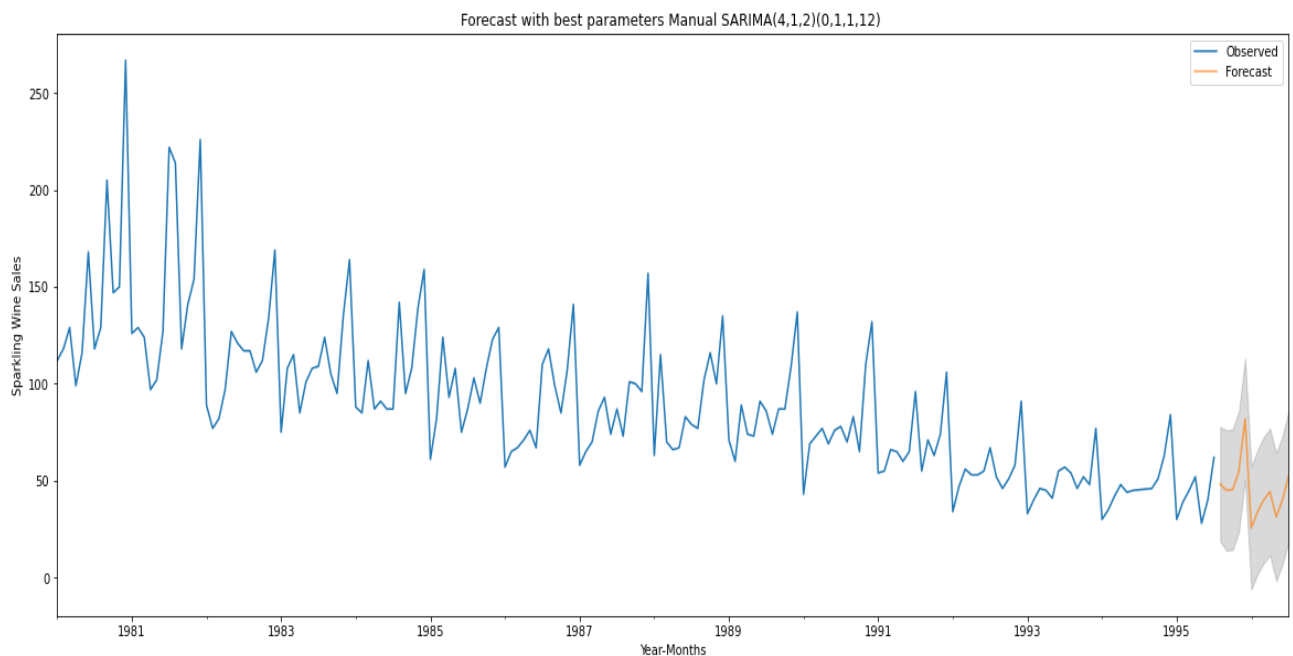
- We have fitted Our Model to Full dataset of Rose wine sales by passing below parameters and found Full Model RMSE -> 39.12

```
model = SARIMAX(df['Rose'], order=(4,1,2), seasonal_order=(0,1,1,12),
enforce_stationarity=False, enforce_invertibility=False)
```

- Next We have predicted for the Next 12 Months along with 95 %Confidence Intervals as below.

Year Month	Prediction	mean_se	mean_ci_lower	mean_ci_upper
01-08-1995	44.01	12.78	18.96	69.06
01-09-1995	45.34	13.02	19.82	70.87
01-10-1995	45.07	13.06	19.48	70.67
01-11-1995	57.67	13.11	31.97	83.37
01-12-1995	85.52	13.11	59.82	111.22
01-01-1996	21.31	13.32	-4.81	47.42
01-02-1996	31.64	13.34	5.50	57.78
01-03-1996	37.63	13.53	11.11	64.16
01-04-1996	39.66	13.55	13.11	66.22
01-05-1996	29.92	13.70	3.06	56.79
01-06-1996	37.50	13.73	10.59	64.41
01-07-1996	49.80	13.86	22.63	76.97

From above table we can see that predictions are marked in Green & Lower Confidence Interval Marked in red, Upper Confidence Intervals Are Marked in dark Blue & mean standard errors are marked in sky blue colour, The Confidence Interval means our model is 95% confident that the prediction will lies in this range. let's see Our predictions graphically.



From above Graph we can see that there is decreasing sales trend along with yearly seasonality, we can have a clear picture about how Our Rose brand Wine sales will be in Next 12 Months assuming that there will be no physical or natural obstacles, like lockdown, Curfew, Natural disaster etc.

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

As in the problem statement it is asked that you have to analyse the sales of Rose wine of company ABC Estate Wines, so from analysis we have noticed that Rose wine sales are having decreasing trend & clear seasonality.

Since in previous questions we have built a table sorted in increasing RMSE and found that **Triple Exponential Smoothing Model Alpha=0.1, Beta=0.2, Gamma=0.2 is having lowest RMSE of 9.64** this means that for building the Predictive model for Rose wine sales you have to take smoothing factor for Level is 0.1 & Smoothing factor for Trend is 0.2, seasonality smoothing factor is also 0.2. That infers Trend & Seasonality are playing equal role in predicting next year Rose wine sales. RMSE is the comparable factor, Lower the RMSE (Root Mean Square Error) means that best optimized Model.

We can also observe the following.

- ✚ As we have seen that sales trend is decreasing over the years so there might be an issue with the wine taste, Quality or Competitor Wine , So Company should focus on their quality , taste & smell etc.
- ✚ The 4th Quarter is having the highest sales across all the year, 3rd Quarter is having second highest sales this may be due to winter season, thus we can infer that in winter rose wine is more purchased by peoples
- ✚ Also from year- Month Boxplot we can notice that in the Month of December every year Wine sales are increasing, this might be due to Christmas, New Year and other festivals.
- ✚ In first five years, the trend is increasing & in next intervals of 5 years' trend is continually decreasing.
- ✚ In the month of January average wine sales are lowest across all years.
- ✚ In first Quarter Average wine sales are lowest across all years

Based on above findings we following are the measures that company should take.

- ✓ Company can make promotional activity for Rose wine on festive seasons and weekends.
- ✓ In the month of December wine demand is very high may be due to Christmas & New year party, so company should sell the wines on their MRP without discount.
- ✓ Company should maintain sufficient stock and distribution channels in the Month of December & for 1st Quarter they will have to reduce the stock to earn More profit.
- ✓ Company can also shake hand with leading restaurants and bars and give them the best discounts, in turn it will directly increase their sales
- ✓ In the month of April to June most people will be on summer vacation so company can target the tourist points and open a distribution centre there.
- ✓ Company can take various promotional activities, like lucky draw, win a chance to meet the celebrity etc.
- ✓ Company should also increase their presence on social media platforms.

- ✓ Company can also introduce on cross selling product with this Rose wine Brand to increase their sales.
- ✓ Company can also make home delivery of Rose Wine Brand to increase the reach to the Customers.
- ✓ Company can also make online surveys and do the sentiment analysis on their feedback to know where is the problem.
- ✓ Company can sponsor Cricket leagues, social events with Rose wine brand, so that people will make a good sentiment about rose wine.