

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 -1'. In the bottom-left corner, there are several thin, curved, light blue lines that sweep upwards and to the right.

Problem -1

Sparkling wine Sales Analysis & Prediction using Time Series Forecasting

Kanhaiya Awasthi

PGP- DSBA SEPT GREAT LEARNING

Problem -1: -

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: [Sparkling.csv](#) and [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.

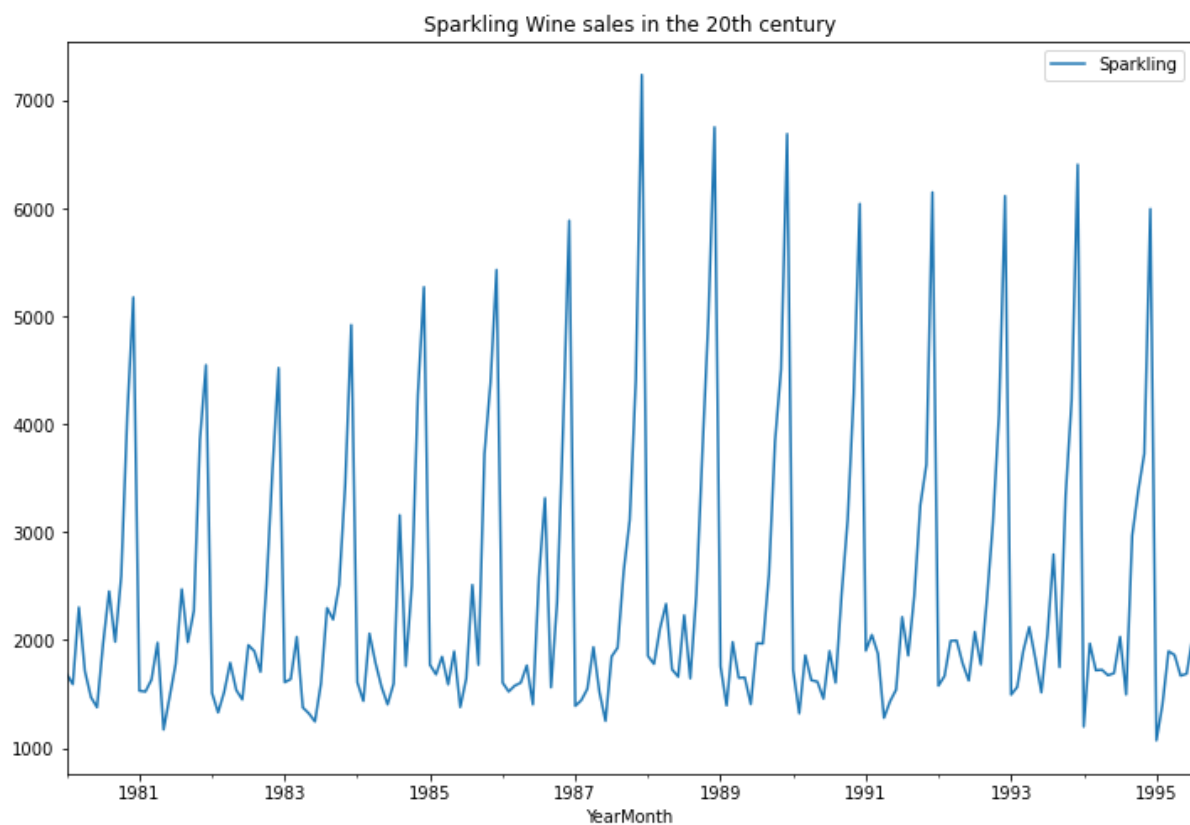
```
df = pd.read_csv('Sparkling.csv', parse_dates=['YearMonth'],  
                index_col='YearMonth')
```

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

Sparkling	
YearMonth	
1980-01-01	1686
1980-02-01	1591
1980-03-01	2304
1980-04-01	1712
1980-05-01	1471

- Time Series Data Plot**

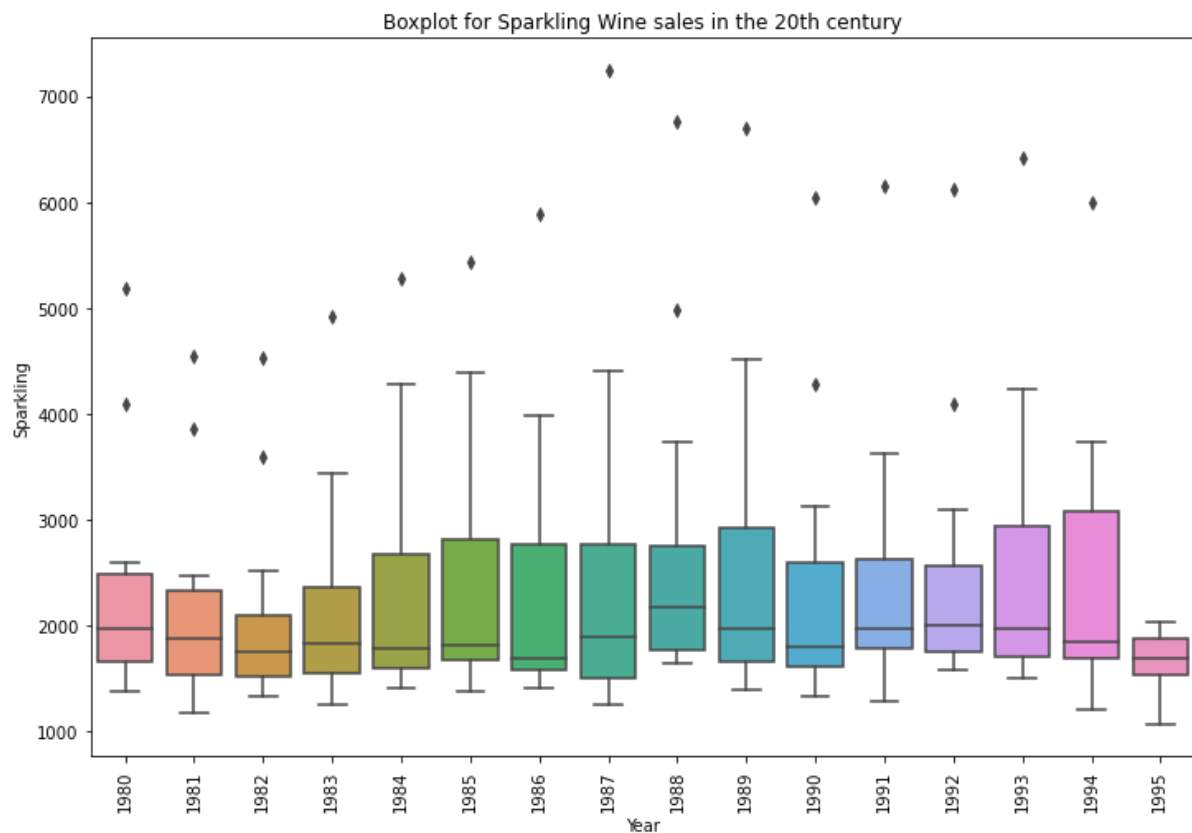


In initial View, the above time series plot has increasing trend up to 1988 & then there is decreasing trend is observed, we can also see seasonality in a plot too.

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

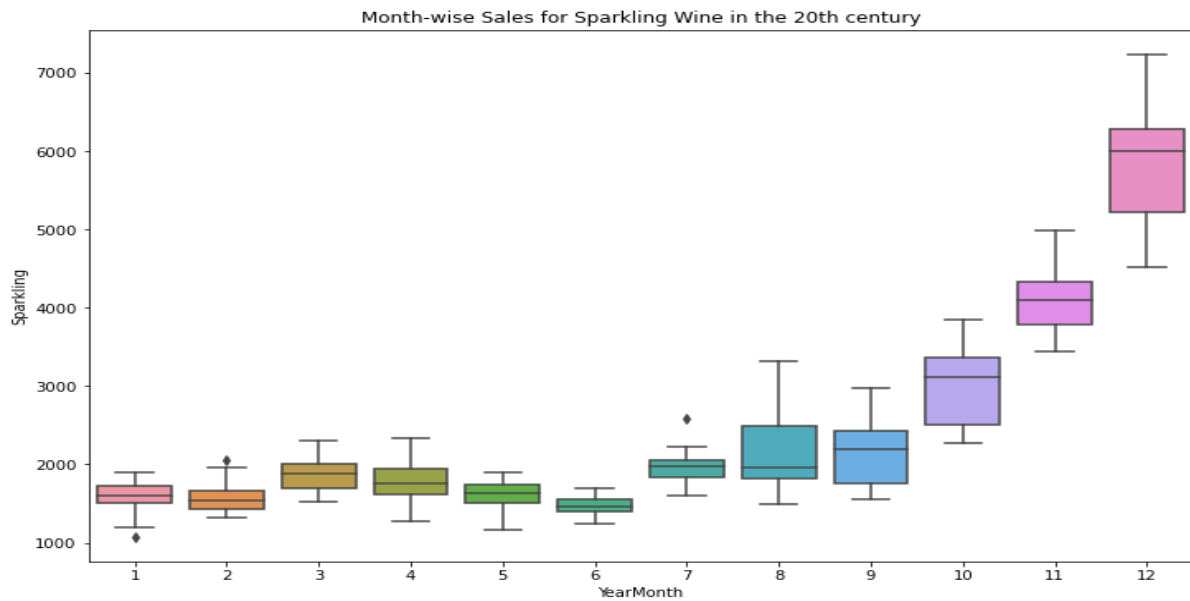
The Given data set tells us about monthly sales of Sparkling wine of Company ABC Estate Wines , so we will plot this Data across Various years & time frames .

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



As we saw in the previous plot, there is an increasing trend in the initial years upto 1988 after which the sales seems to decrease till 1990 and then is on a Slight upward & Downward trend. Also there are some years where the hike or drop in Sales is more as seen in the outliers.

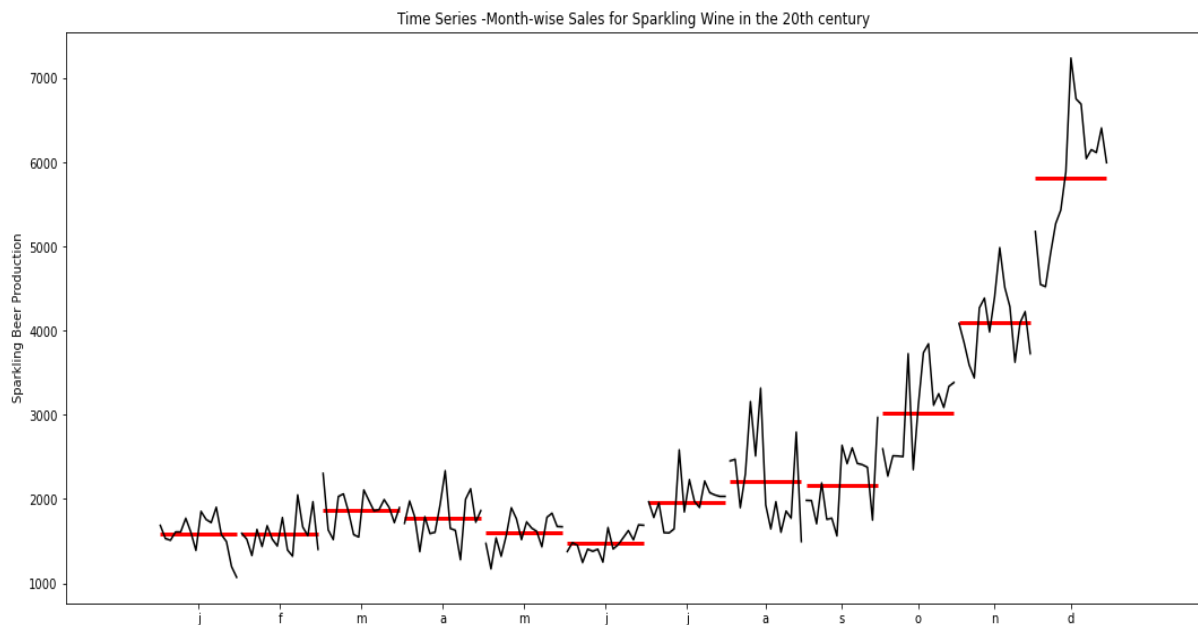
- **Boxplots for Each Calendar Month Across all Years: -**



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

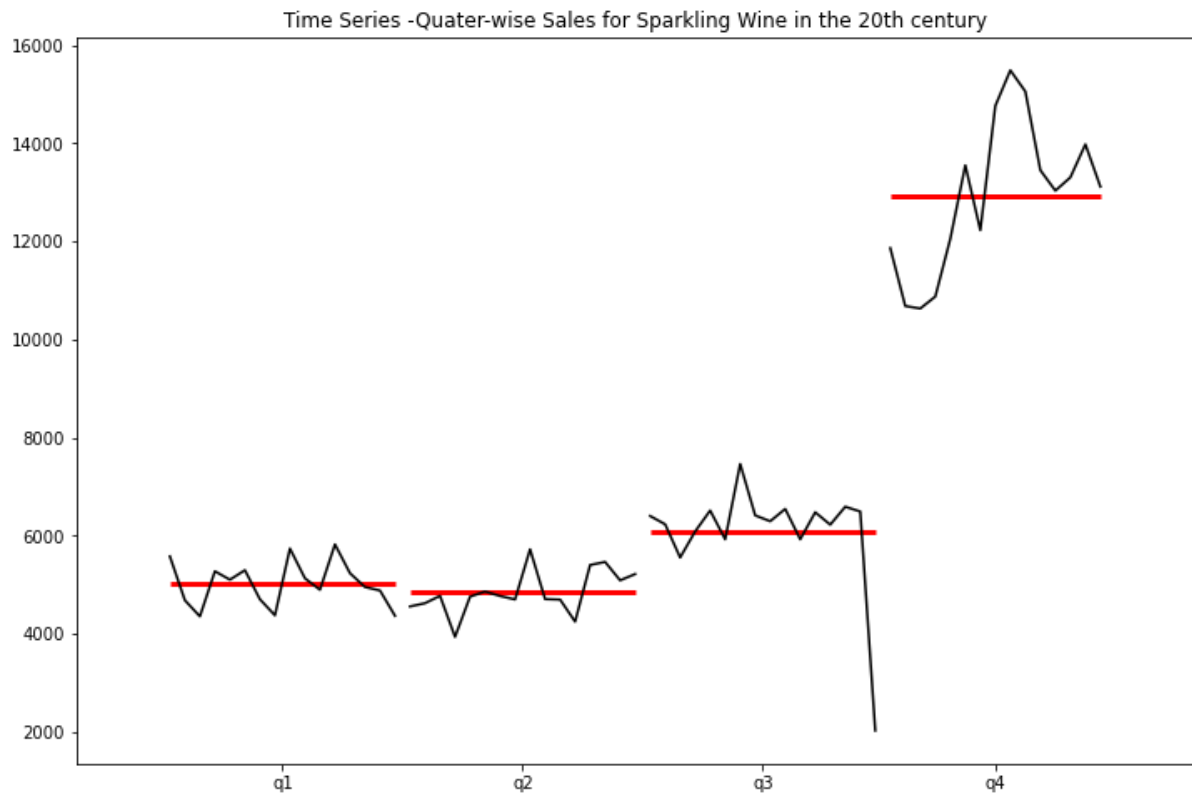
Also we can Notice that There are higher median sales is recorded in the Month of December & Lowest median sales is recorded in the Month of June across all years.

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



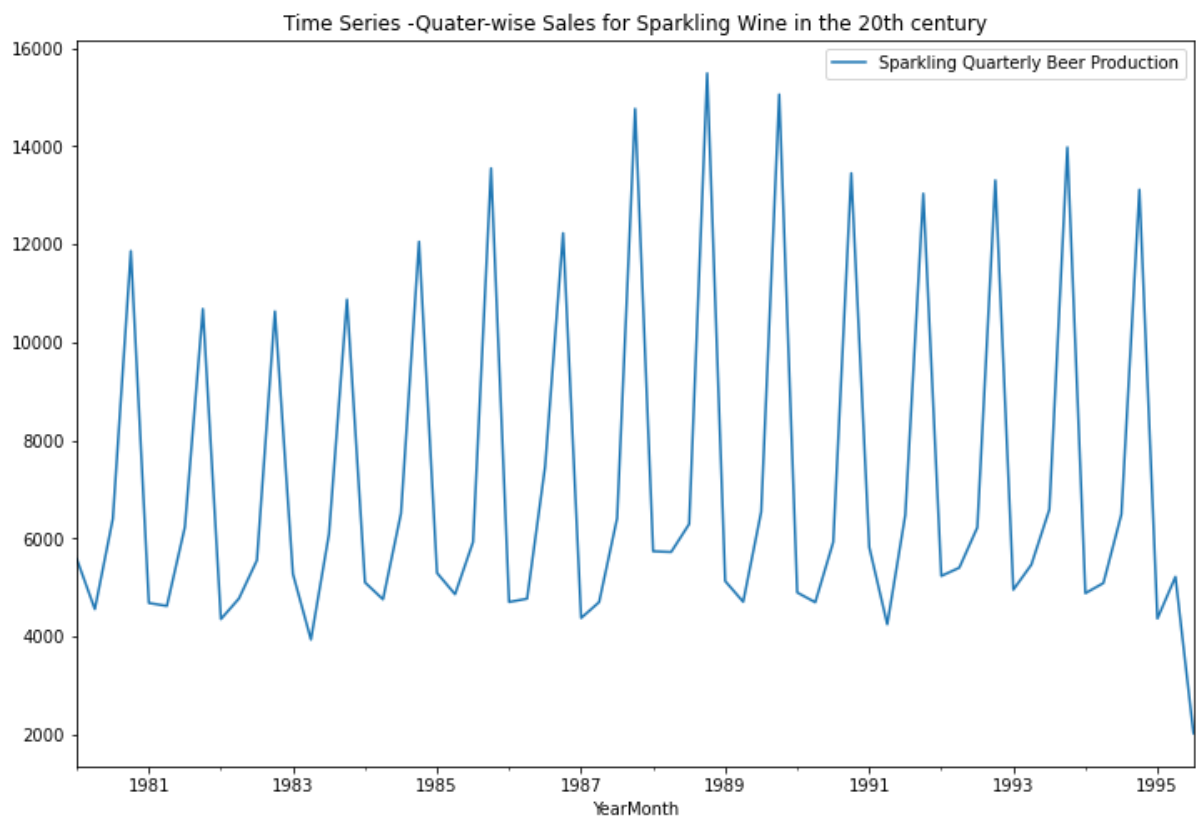
The above Month Plot shows that after June Average wine sales are increasing, Peak is noticed in the month of December every year. So there is slight yearly seasonality can be seen.

- **Conversion of Monthly time series Data into other periods:** - Since Our data is given Month-wise to convert this Data in Daily frequency, Quaterly frequency , We will use resample method of pandas dataframe.
- **Quarter wise Time Series Plot:** - We have resampled Our Data to every Quarters & plot the QTR Plot as below.



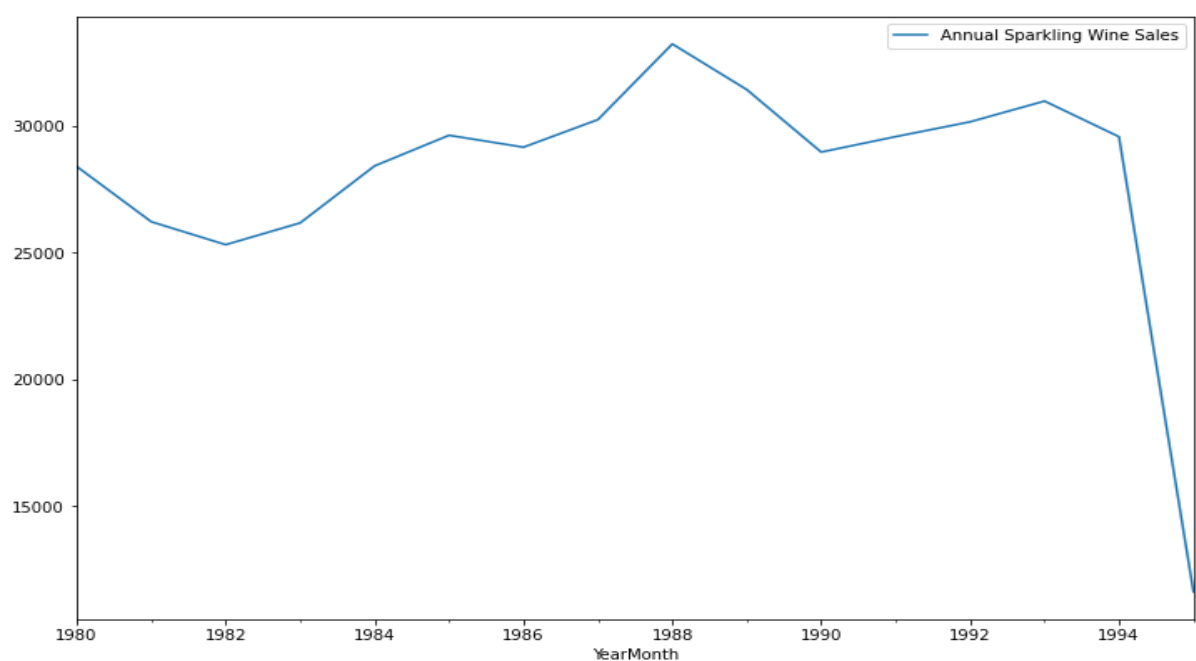
It can be observed from QTR plot Average Sales are highest in 4th QTR every year & average sale are lowest in every 2nd QTR across all the years, This highest peak in 4th QTR Might be due to Christmas and other festive seasons.

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



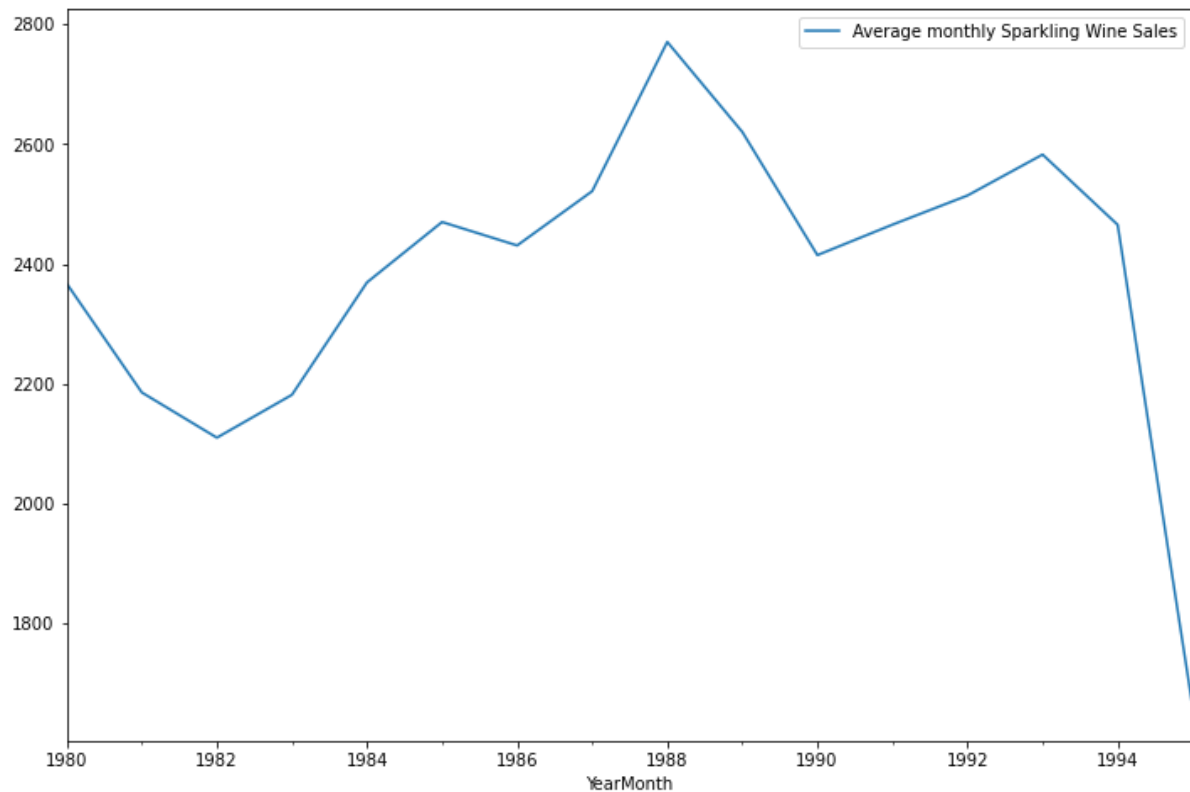
We can see from above Quarter wise time series plot that Trend in Every QTR is Increasing & yearly seasonality can be observed. As we resample data in higher frequencies seasonality will be smoothed & a clear trend can be observed.

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



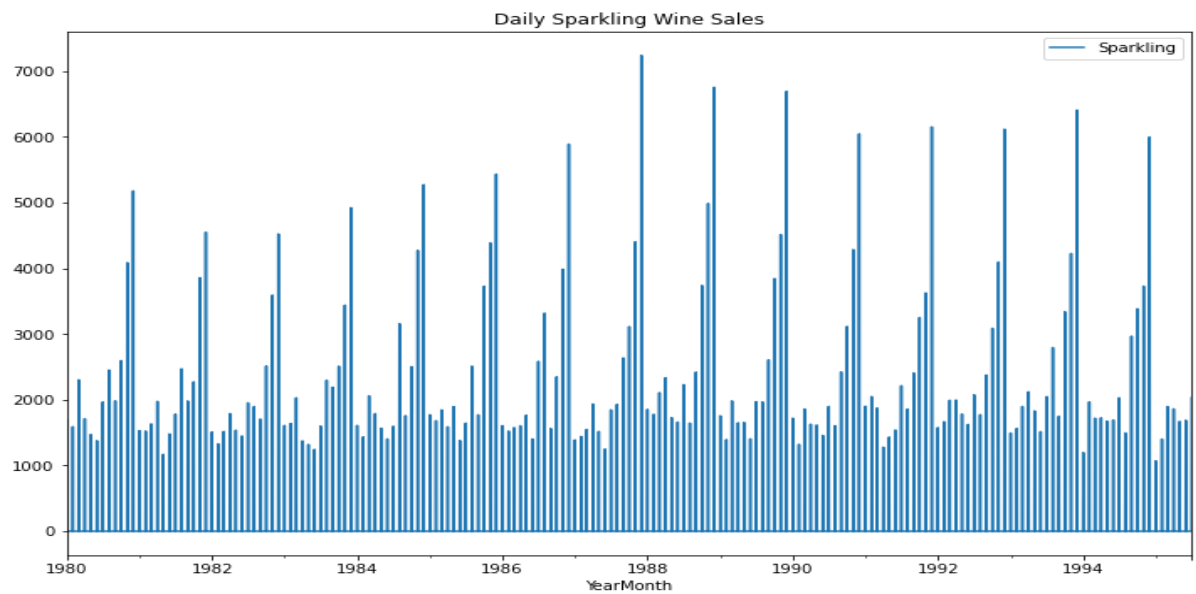
After resampling of our data as yearly we can see a Zig – Zag pattern in wine sales, As noticed wine sales have increased up to year 1988, after that sales have dropped.

- **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).

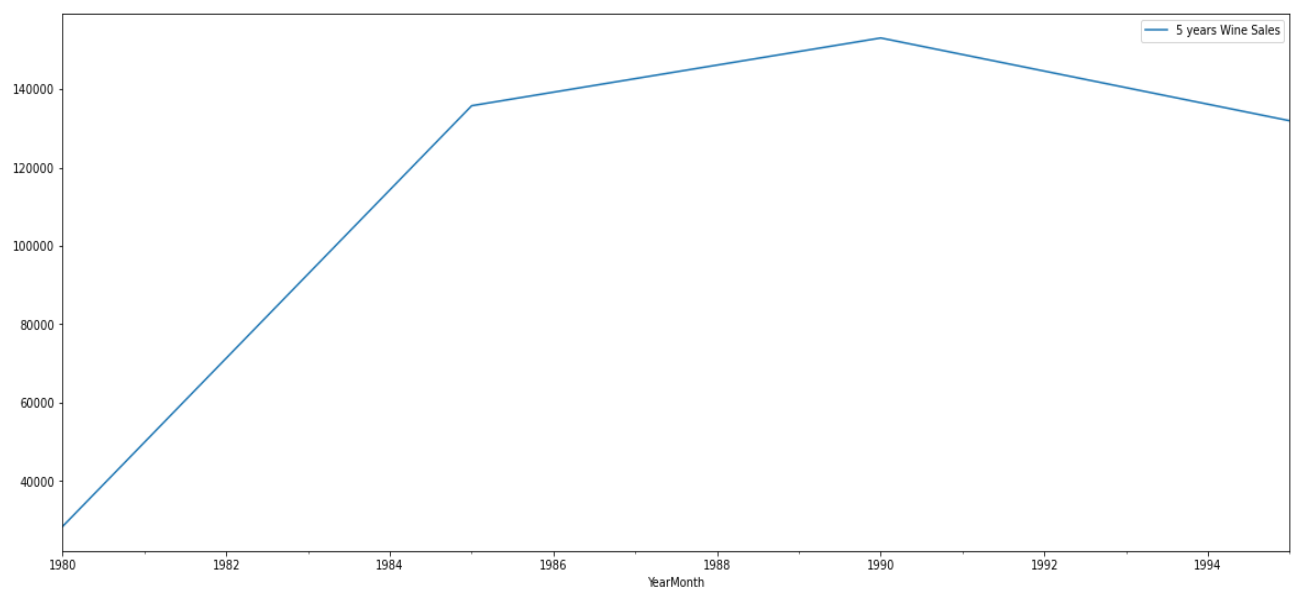
- **Daily Time Series Plot:-** After resampling our data to daily frequencies we have plotted time series plot as below. In below plot higher spikes can be seen for the Higer Daily sales for that particular bin.



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

5 years Wine Sales

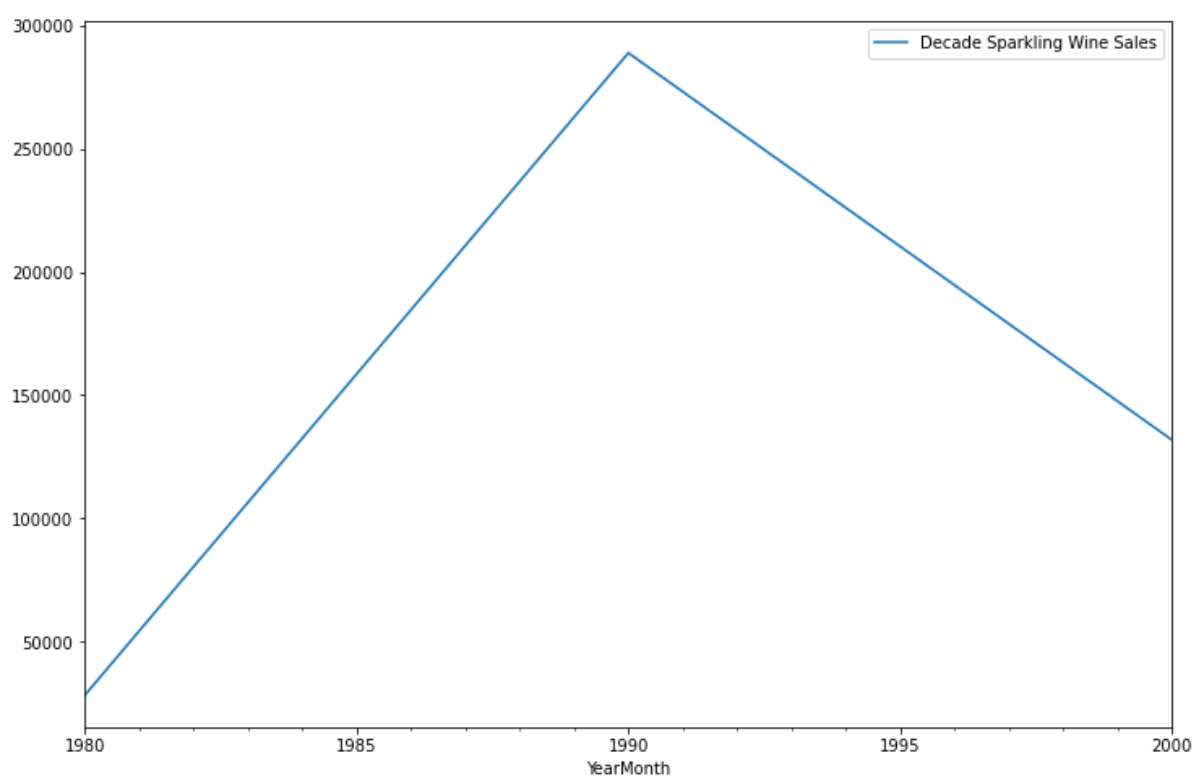
YearMonth	
1980-12-31	28406
1985-12-31	135799
1990-12-31	153094
1995-12-31	131953



From above plot it can be observed that for each 5 years' time frame up to 1990 the Wine sales are having increasing trend but after 1990 there is a slight drop.

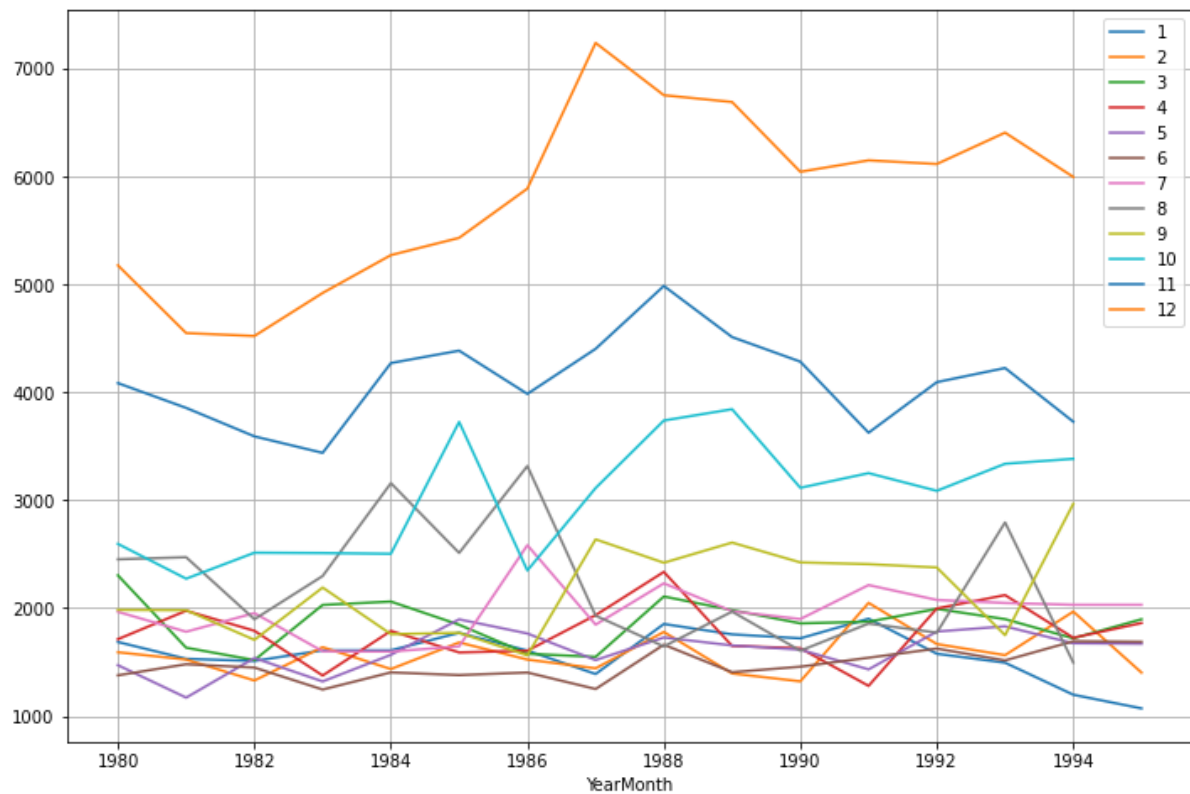
- **Wine sales time series plot across Decades: -**

Decade Sparkling Wine Sales	
YearMonth	
1980-12-31	28406
1990-12-31	288893
2000-12-31	131953



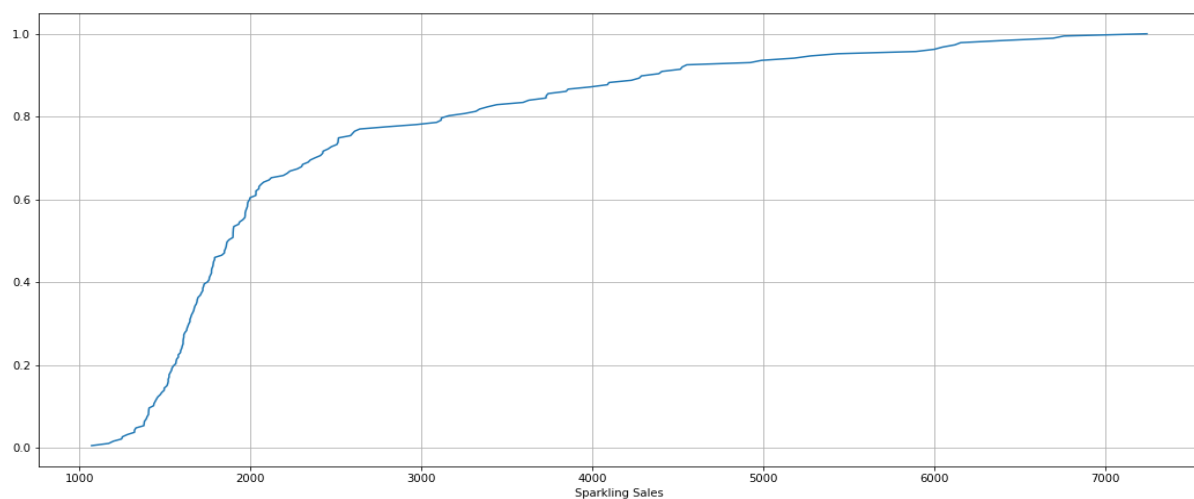
From above time series plot it can be observed that seasonal fluctuation has smoothened down & in first decade sales have seen a drastic Increasing trend & in second decade the sales trend is decreasing.

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



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

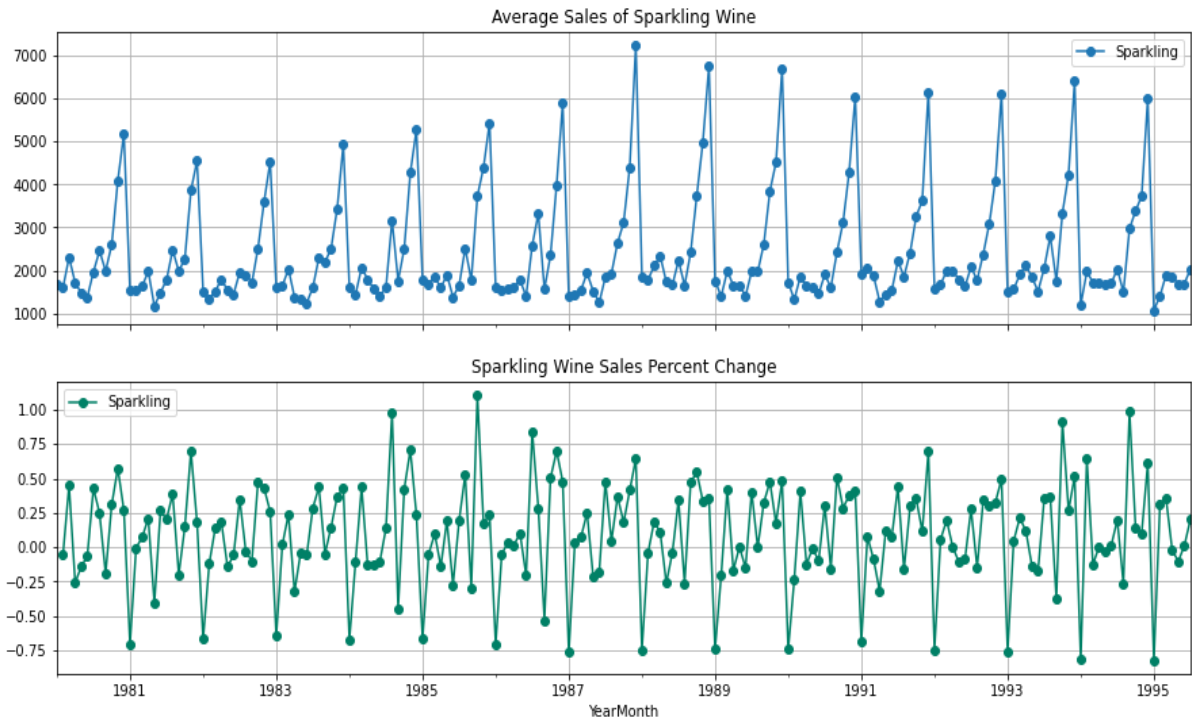
- **Empirical Cumulative Distribution plot :-**



Above particular graph tells us what percentage of data points refer to what number of Sales. in the above graph 80 % of data points just a little more than 3000 units' sales & next 20 % of data points consist of next approx 4000 units of Wine sales.

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

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

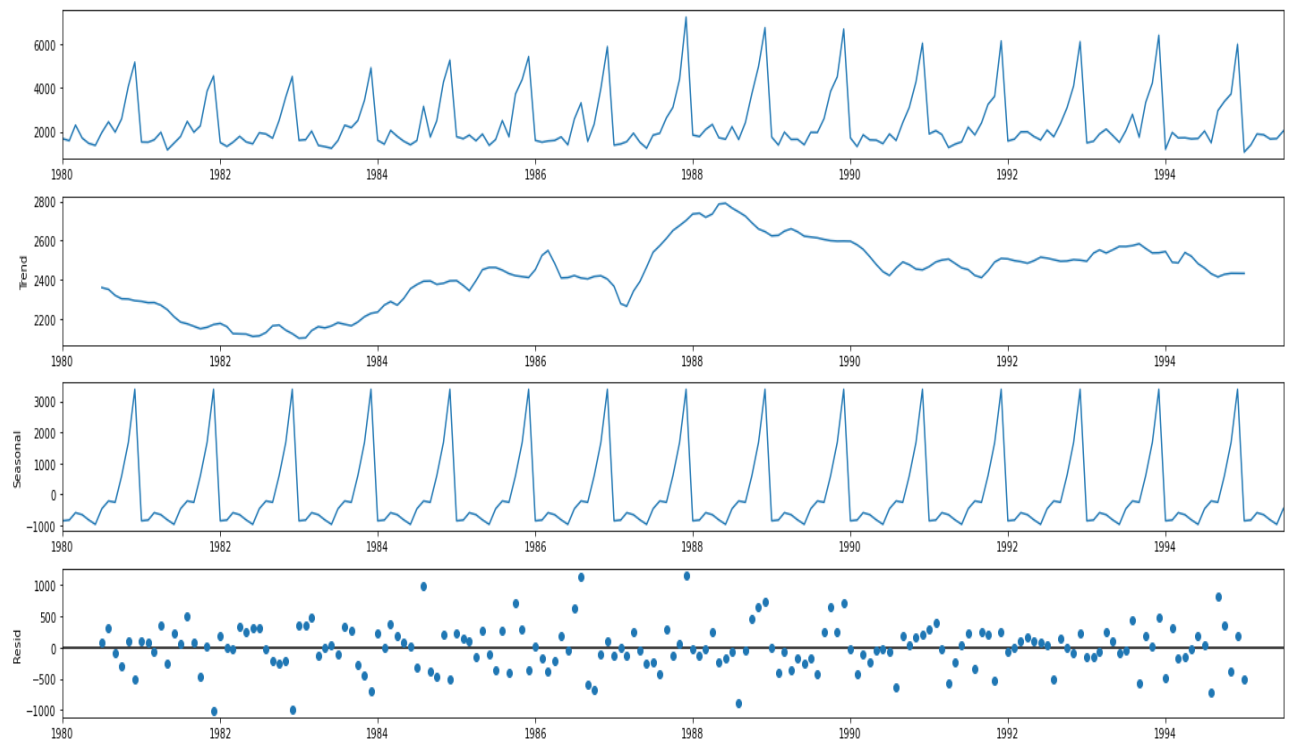


As we can see from Graph that in December 1980 the maximum sales is around 5100 & there was approx. 60 % change in sales recorded compared to November 1980. In December 1985 the change in Sales from previous month is more than 110%.

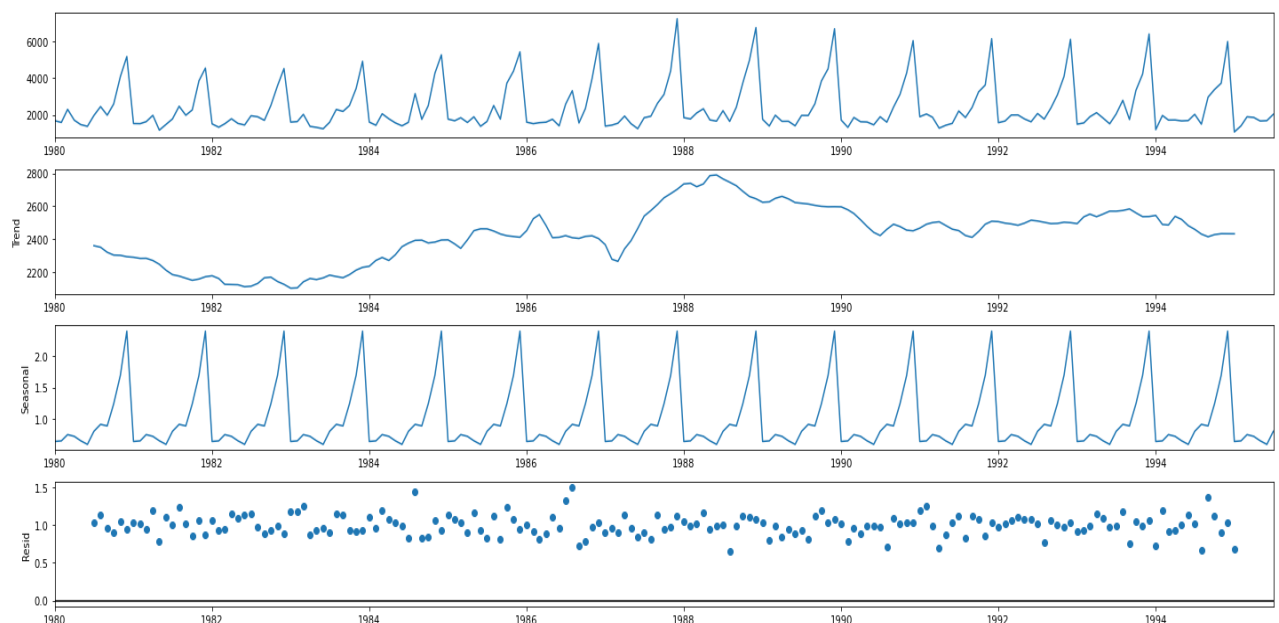
- **Time Series Decomposition:** - For Checking the Trend, Seasonality & Residuals Time we have to make Series Decomposition, this can be done in two ways Additive Decomposition & Multiplicative Decomposition.
- **Additive Decomposition:** - As per the 'additive' decomposition, we see that there is a pronounced trend in the earlier years of the data upto 1989. There is a seasonality as well. but if we see at residuals there 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 a pronounced trend in the earlier years of the data upto 1989. There is a seasonality as well. In this residuals are randomly distributed so Wine sales data is multiplicative in nature.

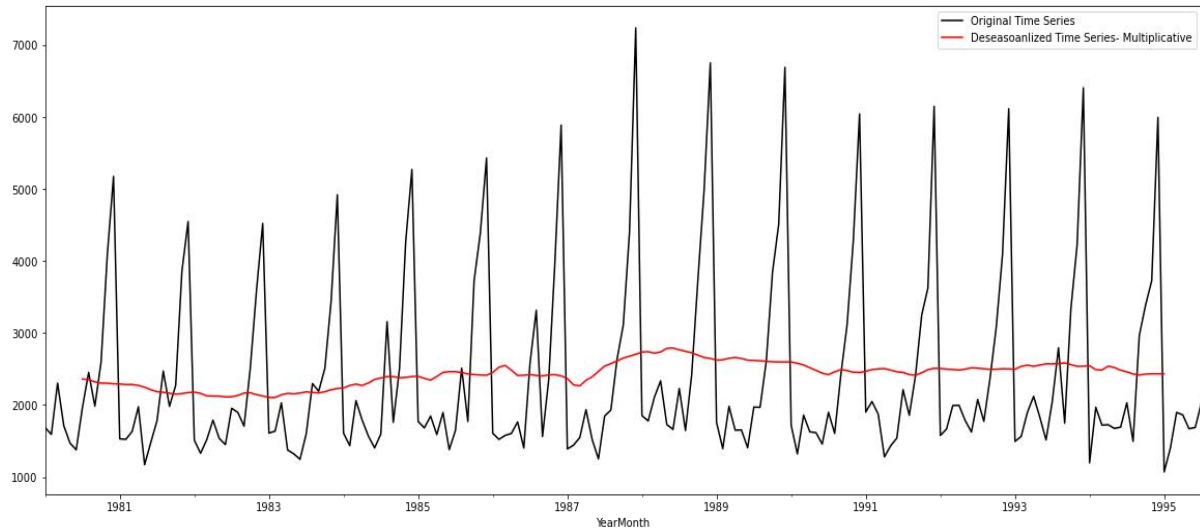


Formula For Multiplicative Time Series:-

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

We will De-seasonalize our time series by removing Seasonal Component from Time series .

- **De-Seasonalize Time Series: -**



In Above Multiplicative De-seasonalized Time series Trend & residual components can be clearly seen.

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

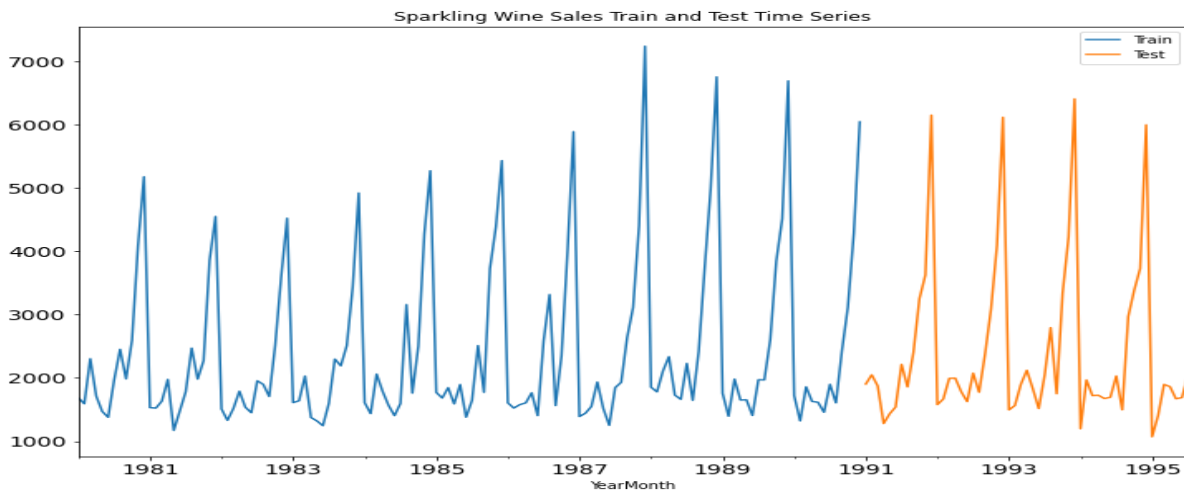
We have used following code for making train test split where Test data is started from 1991. As per shape 132 Data points are kept for the training of our model & 55 data points are kept for testing of above model.

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

```
: train.shape, test.shape
```

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

- **Train Test 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)
```

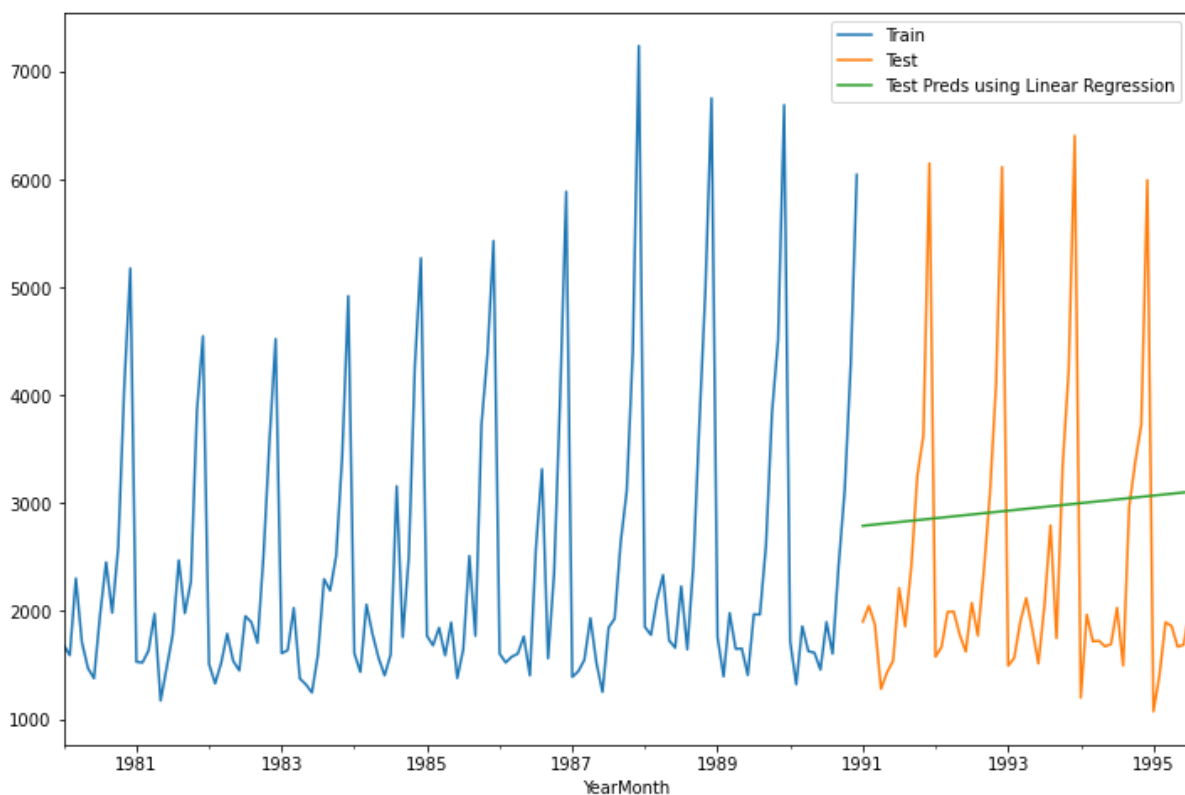
Training Time instance
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 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 .)

	Sparkling	time	RegOnTime
YearMonth			
1991-01-01	1902	133	2791.652093
1991-02-01	2049	134	2797.484752
1991-03-01	1874	135	2803.317410
1991-04-01	1279	136	2809.150069
1991-05-01	1432	137	2814.982727

- 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. 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

Test RMSE	
Linear Regression OnTime instance	1389.135175

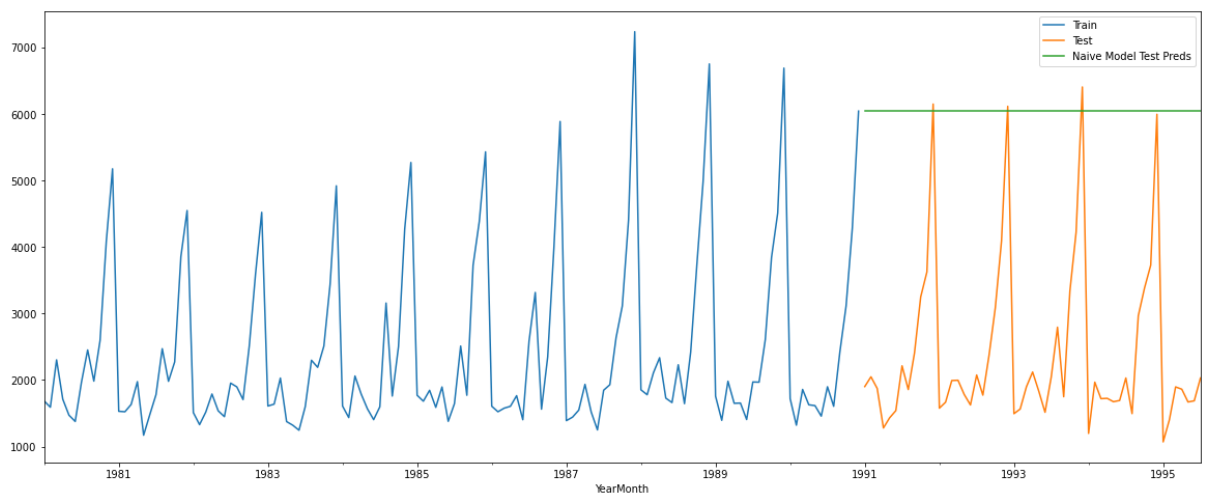
We will compare this RMSE With further Models.

- **Model 2: Naïve 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.

Year Month	Prediction
01-01-1991	6047
01-02-1991	6047
01-03-1991	6047
01-04-1991	6047
01-05-1991	6047

As we have seen that on December 1990 , there are 6047 units of Sparkling wine sales so the sales as on jan 1991 will also be 6047 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.

Test RMSE	
Linear Regression OnTime instance	1389.135175
NaiveModel	3864.279352

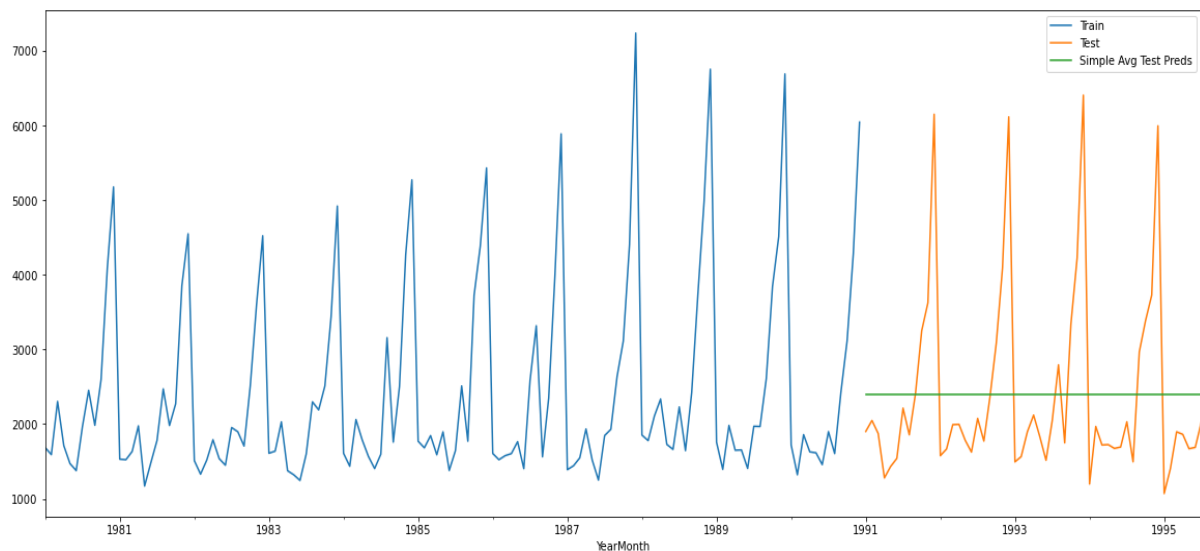
As we can see that the RMSE of Naïve model is around 3 times of linear regression so , the Naïve model is not a suitable model for this analysis.

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

YearMonth	Sparkling	Simple Average forecast
01-01-1991	1902	2403.78
01-02-1991	2049	2403.78
01-03-1991	1874	2403.78
01-04-1991	1279	2403.78
01-05-1991	1432	2403.78

As we can see that train data average value is 2403.78 units, so we are forecasting the same for our 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.

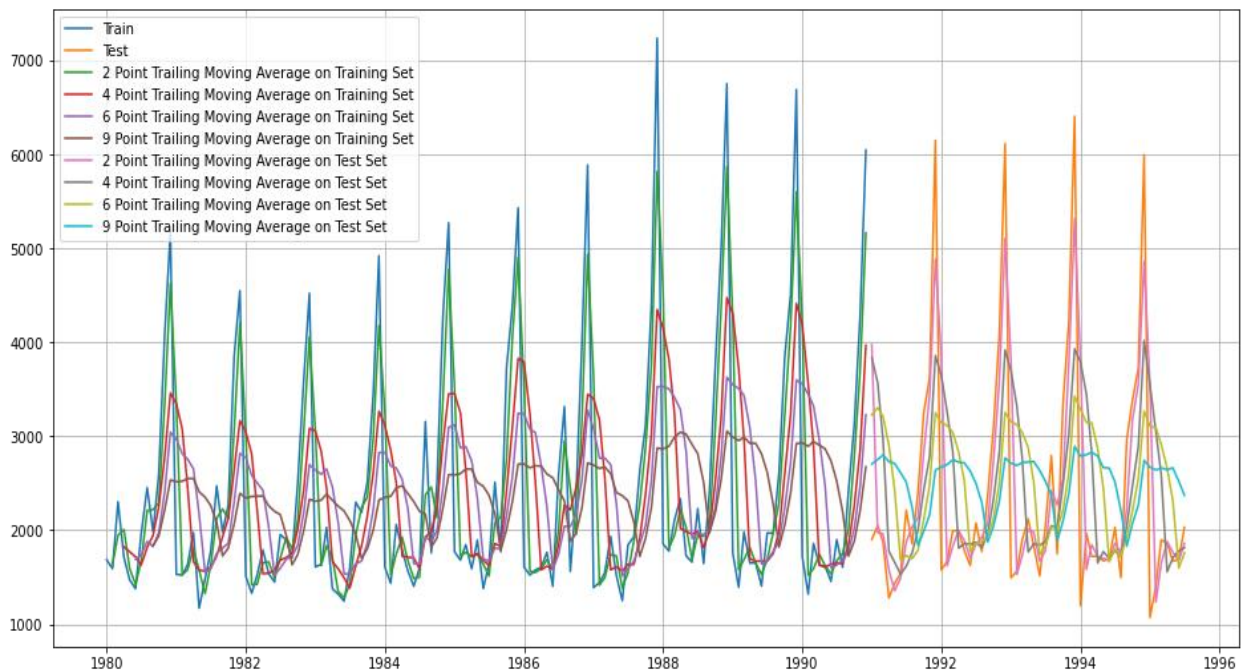
Test RMSE	
Linear Regression OnTime instance	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804

We can see from above table that RMSE is lowest among all the two 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	1686.00	NaN	NaN	NaN	NaN
01-02-1980	1591.00	1638.50	NaN	NaN	NaN
01-03-1980	2304.00	1947.50	NaN	NaN	NaN
01-04-1980	1712.00	2008.00	1823.25	NaN	NaN
01-05-1980	1471.00	1591.50	1769.50	NaN	NaN
01-06-1980	1377.00	1424.00	1716.00	1690.17	NaN
01-07-1980	1966.00	1671.50	1631.50	1736.83	NaN
01-08-1980	2453.00	2209.50	1816.75	1880.50	NaN
01-09-1980	1984.00	2218.50	1945.00	1827.17	1838.22
01-10-1980	2596.00	2290.00	2249.75	1974.50	1939.33
01-11-1980	4087.00	3341.50	2780.00	2410.50	2216.67
01-12-1980	5179.00	4633.00	3461.50	3044.17	2536.11
01-01-1981	1530.00	3354.50	3348.00	2971.50	2515.89
01-02-1981	1523.00	1526.50	3079.75	2816.50	2521.67
01-03-1981	1633.00	1578.00	2466.25	2758.00	2550.11

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

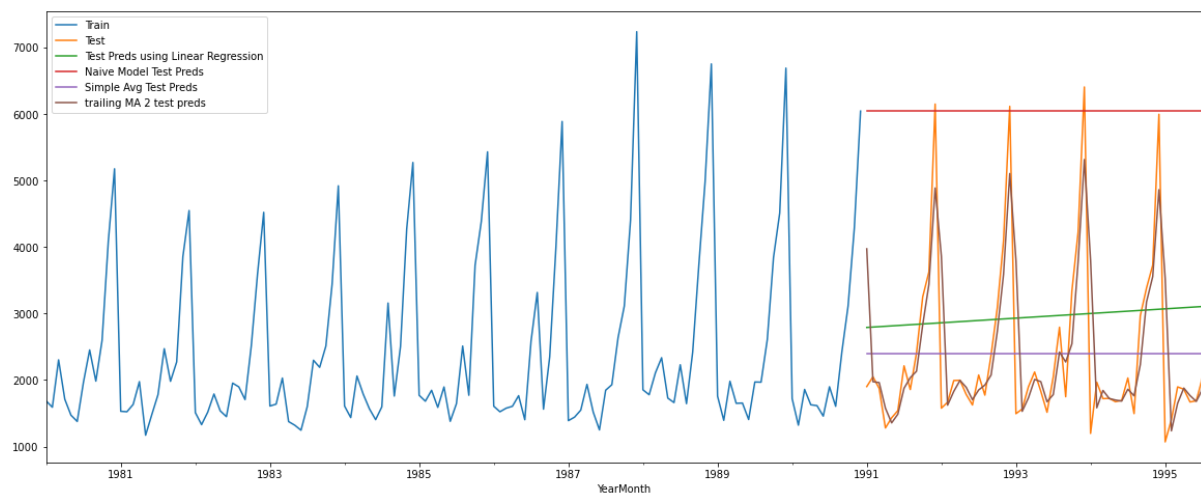


We can clearly see that 2 point moving average is best fitting with training and test data & then subsequent 4 , 6 & 9 points trailings respectively . Let's Check RMSE for all the Moving averaged data.

	Test RMSE
Linear Regression OnTime instance	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315

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.

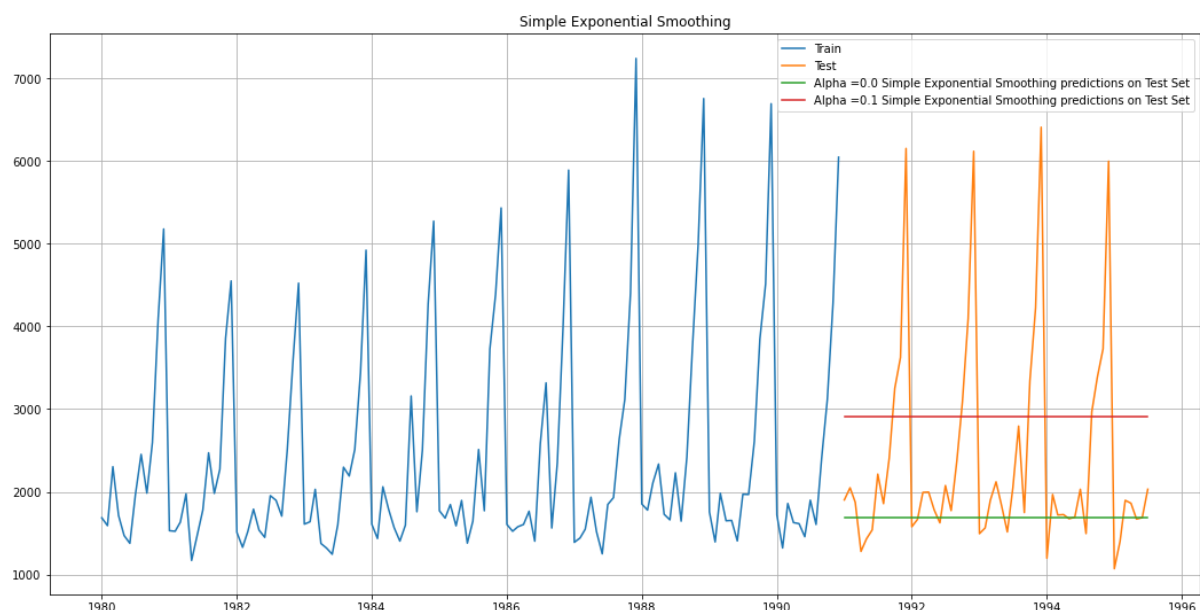
- **Model 5: Simple Exponential Smoothing:** - In this model we consider that our data is having only level component, there is no seasonality & trend, so we will calculate best alpha value using iterative Brute force method.

We have fitted simple exponential smoothing model to our training data and found below prediction from base model. Here alpha value is close to zero (0.0).

	Sparkling	predict
YearMonth		
1991-01-01	1902	2403.790103
1991-02-01	2049	2403.790103
1991-03-01	1874	2403.790103
1991-04-01	1279	2403.790103
1991-05-01	1432	2403.790103

The above prediction as similar to Simple average method because here no level is considered, 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 0.0 , as per graph none of the Simple exponential smoothing model is best fitting with test data let's check their RMSE Values below .

	Test RMSE
Linear Regression OnTime instance	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315
Alpha=0.00,SimpleExponentialSmoothing	1275.081839
Alpha=0.1,SimpleExponentialSmoothing	1375.393398

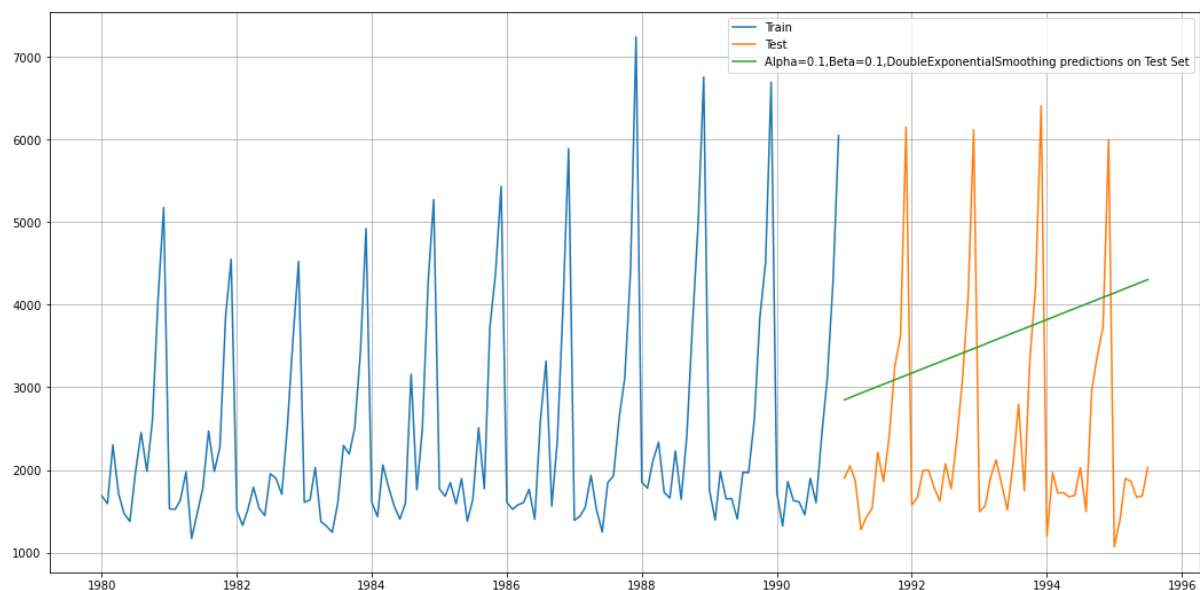
Amongst both simple exponential smoothening models alpha =0.0 value is giving comparatively less RMSE.

- **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.

Level $\alpha = 0.1$ & trend $\beta = 0.1$

Let's plot the Double exponential smoothing predictions.



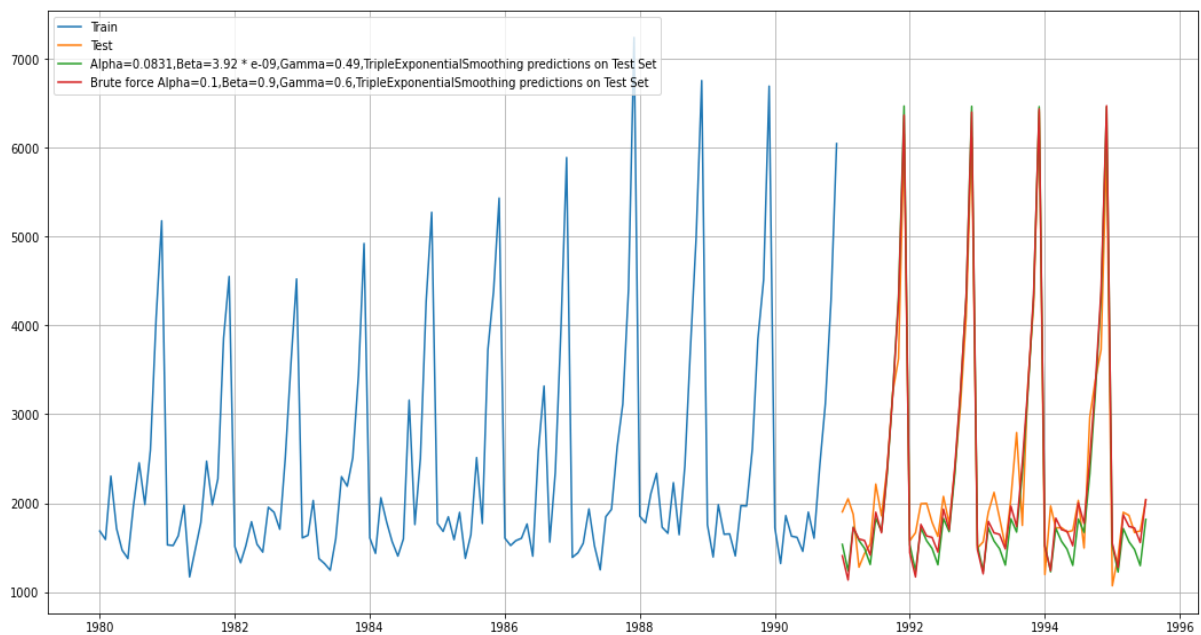
In above graph green line is showing Double exponential smoothing predictions on test sets , visually it is not best forecasting our test data, let's Check corresponding RMSE.

	Test RMSE
Linear Regression OnTime instance	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315
Alpha=0.00,SimpleExponentialSmoothing	1275.081839
Alpha=0.1,SimpleExponentialSmoothing	1375.393398
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	1778.564670

As in last line we can see that Holt's model is giving us comparatively high value of RMSE i.e. 1778.5 & so we can say that it's not a good predictor/forecaster of our future data.

- Model 7: Triple Exponential Smoothing (Holt - Winter's Model) :-** This model account that the Sparkling wine dataset is having all the three components , Level , Trend, Seasonality too. This means Sparkling 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=0.0831, Beta=3.92 * e-09, Gamma=0.49 & found that our RMSE is 362 , 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=0.9,Gamma=0.6 .

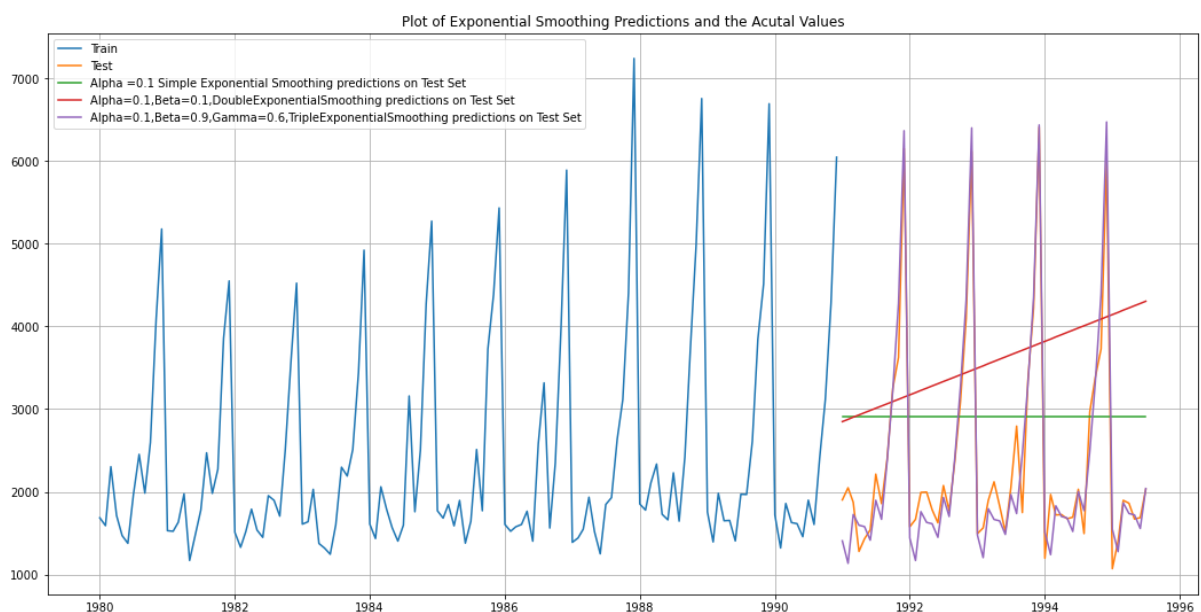


We have seen that model found by brute force method is giving lowest values of RMSE.

Model	Test RMSE
Alpha=0.0831,Beta=3.92 * e-09,Gamma=0.49,TripleExponentialSmoothing	362.74
Alpha=0.1,Beta=0.9,Gamma=0.6,TripleExponentialSmoothing	338.45

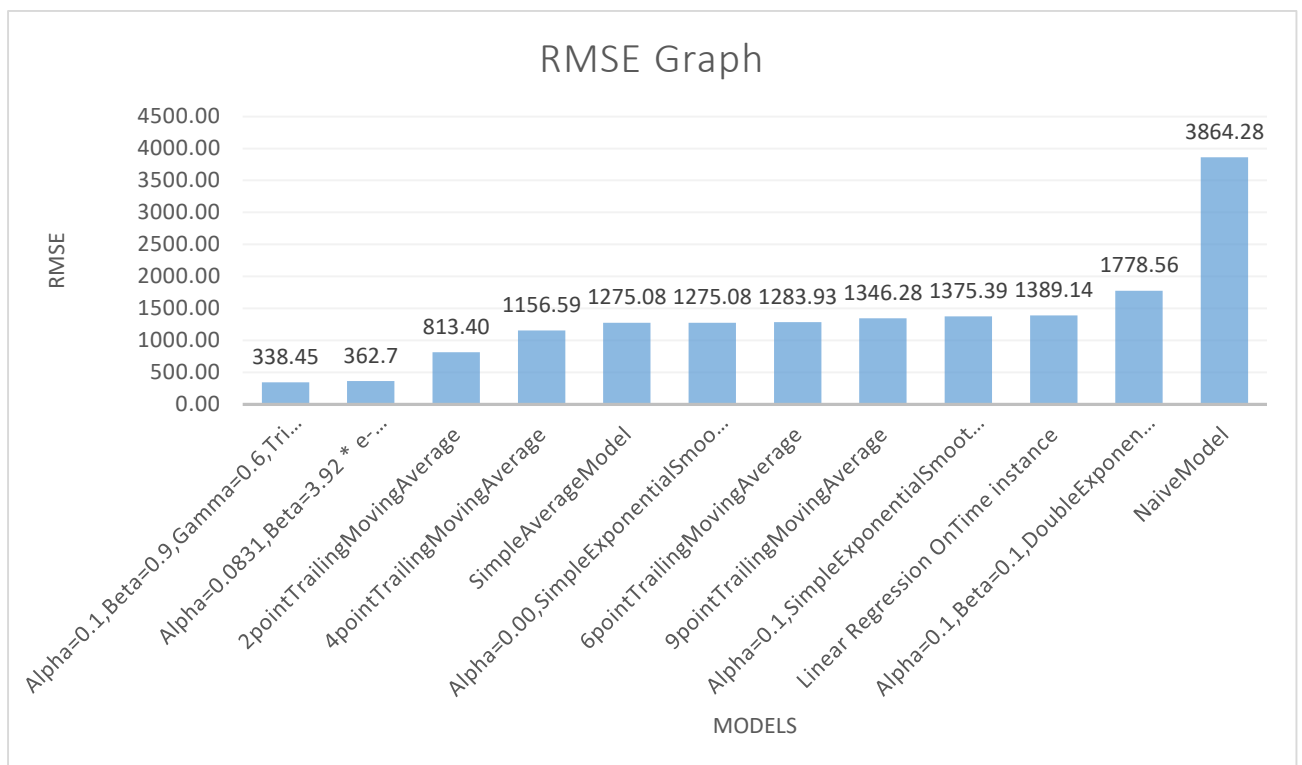
This means Sparkling wine sales depends on below best forecast parameters **Alpha=0.1, Beta=0.9, Gamma=0.6**.

- Plotting combined all the three Smoothing Models.



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

Model	Test RMSE
Alpha=0.1,Beta=0.9,Gamma=0.6,TripleExponentialSmoothing	↓ 338.45
Alpha=0.0831,Beta=3.92 * e-09,Gamma=0.49,TripleExponentialSmoothing	↓ 362.74
2pointTrailingMovingAverage	↓ 813.40
4pointTrailingMovingAverage	↓ 1156.59
SimpleAverageModel	↓ 1275.08
Alpha=0.00,SimpleExponentialSmoothing	↓ 1275.08
6pointTrailingMovingAverage	↓ 1283.93
9pointTrailingMovingAverage	↓ 1346.28
Alpha=0.1,SimpleExponentialSmoothing	↓ 1375.39
Linear Regression OnTime instance	↓ 1389.14
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	→ 1778.56
NaiveModel	↑ 3864.28



By considering RMSE till now our best performing model is Alpha=0.1, Beta=0.9, Gamma=0.6 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 Sparkling wine sales data by using Augmented Dicky Fuller test on test data, below are the Null and alternate hypothesis of the test .

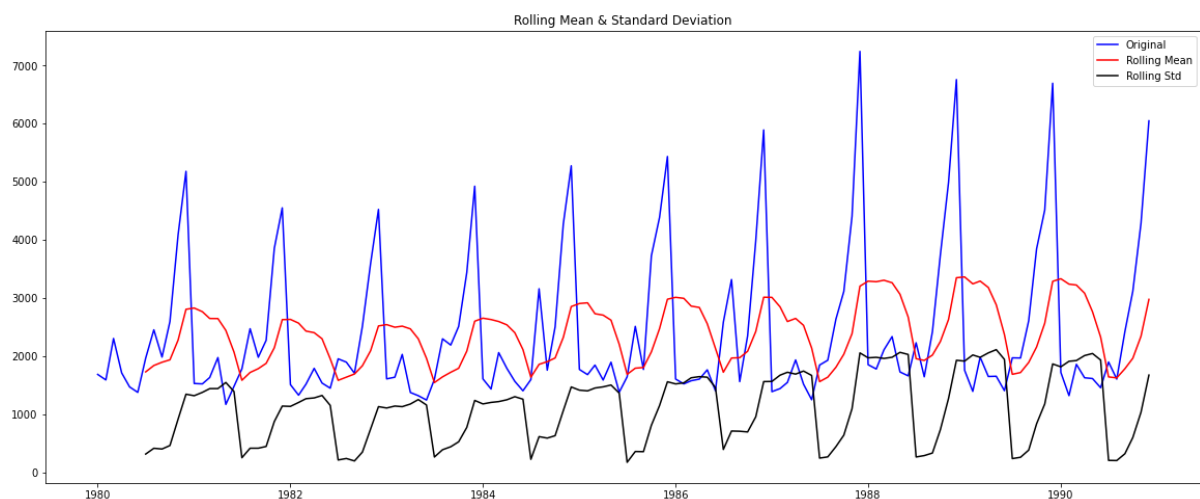
Null hypothesis H_0

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

Alternate hypothesis H_a

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

- After Performing ADF Test we found below results

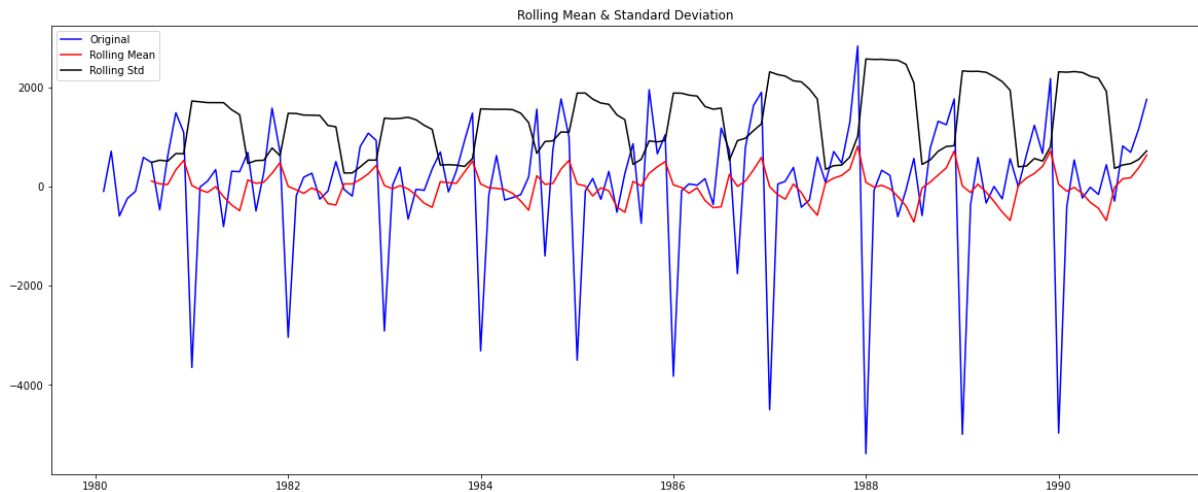


```
Results of Dickey-Fuller Test:
Test Statistic          -1.208926
p-value                  0.669744
#Lags Used              12.000000
Number of Observations Used 119.000000
Critical Value (1%)      -3.486535
Critical Value (5%)      -2.886151
Critical Value (10%)     -2.579896
dtype: float64
```

as we can see from above ADF output that p-value is 0.669 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.



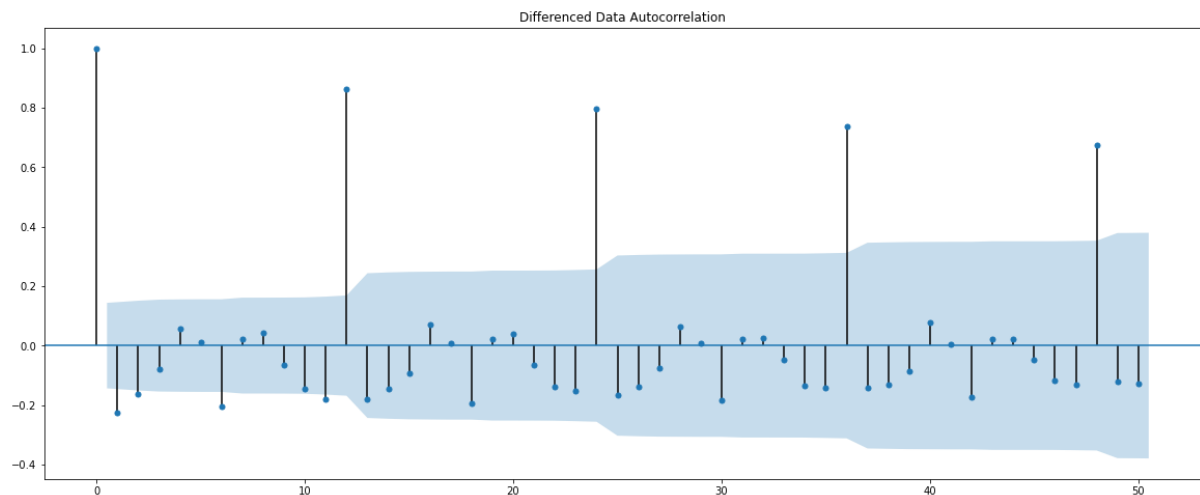
Results of Dickey-Fuller Test:

Test Statistic	-8.005007e+00
p-value	2.280104e-12
#Lags Used	1.100000e+01
Number of Observations Used	1.190000e+02
Critical Value (1%)	-3.486535e+00
Critical Value (5%)	-2.886151e+00
Critical Value (10%)	-2.579896e+00
dtype:	float64

from above graph we can notice that in the Sparkling 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
(2, 1, 2)	2210.62
(2, 1, 1)	2232.36
(0, 1, 2)	2232.78
(1, 1, 2)	2233.60
(1, 1, 1)	2235.01
(2, 1, 0)	2262.04
(0, 1, 1)	2264.91
(1, 1, 0)	2268.53
(0, 1, 0)	2269.58

As we can see from the above table that lowest AIC Values is found on parameters (2,1,2) , so we will take 2 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

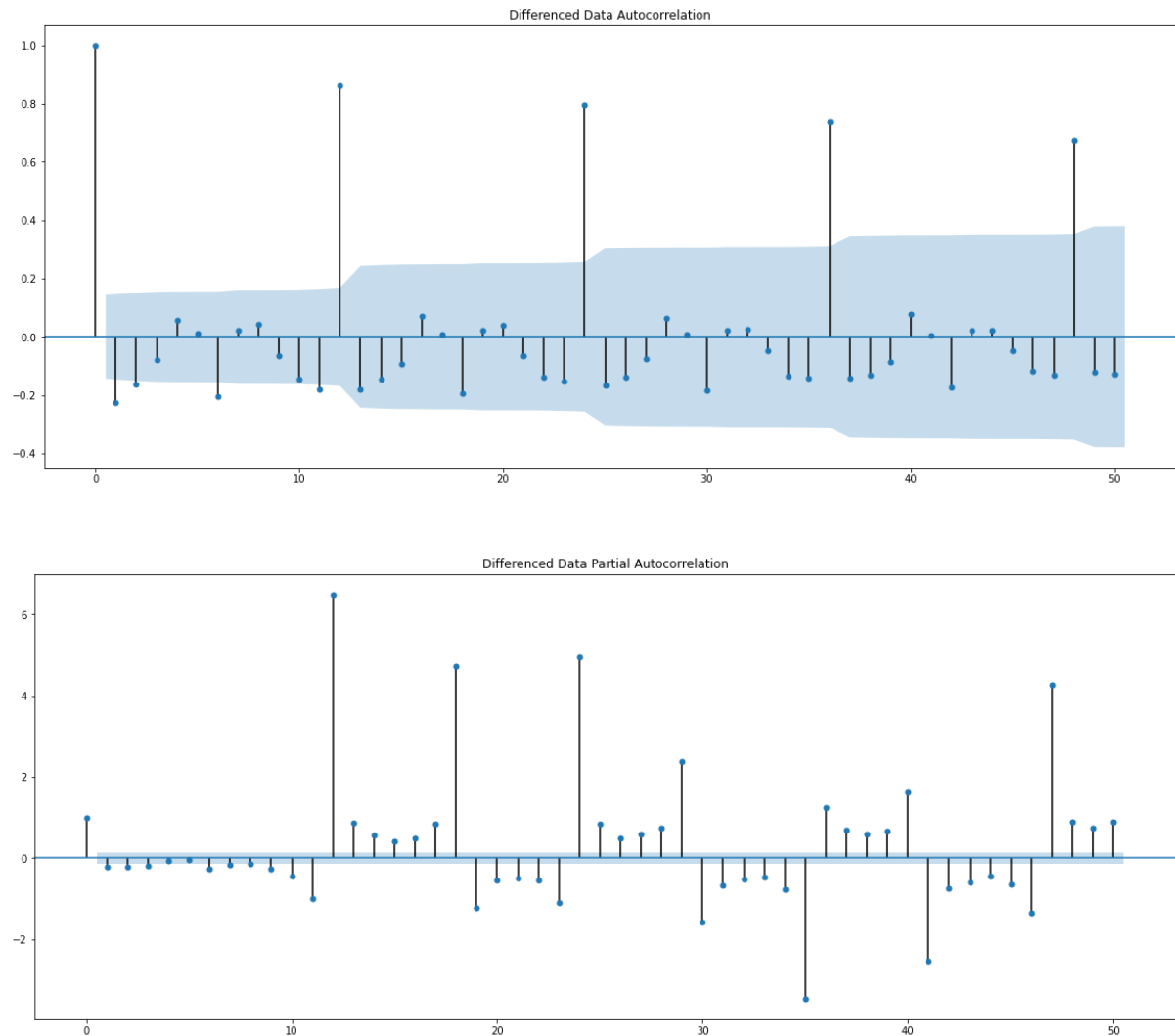
Dep. Variable:	D.Sparkling	No. Observations:	131			
Model:	ARIMA(2, 1, 2)	Log Likelihood	-1099.31			
Method:	css-mle	S.D. of innovations	1012.606			
Date:	Fri, 21 May 2021	AIC	2210.618			
Time:	00:48:04	BIC	2227.869			
Sample:	02-01-1980	HQIC	2217.628			
	-2003					
	coef	std err	z	P> z	[0.025	0.975]
const	5.5857	0.517	10.814	0	4.573	6.598
ar.L1.D.Sparkling	1.2699	0.074	17.046	0	1.124	1.416
ar.L2.D.Sparkling	-0.5601	0.074	-7.617	0	-0.704	-0.416
ma.L1.D.Sparkling	-1.998	0.042	-47.168	0	-2.081	-1.915
ma.L2.D.Sparkling	0.998	0.042	23.548	0	0.915	1.081
Roots						
	Real	Imaginary	Modulus	Frequency		
AR.1	1.1336	-0.7073j	1.3362	-0.0888		
AR.2	1.1336	+0.7073j	1.3362	0.0888		
MA.1	1.0006	+0.0000j	1.0006	0		
MA.2	1.0014	+0.0000j	1.0014	0		

The light green marked are the coefficient of 2 Auto regressive & 2 Moving average for the Sparkling wine data , The dark green marked values are p values which all are less than level of significance =0.05, So we can say that all the AR & MA components are significant.

Model	Test RMSE
ARIMA(2,1,2)	1374.678

The root mean square value is 1374.6 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 .

param	seasonal	AIC
(1, 1, 2)	(1, 0, 2, 12)	↓ 1555.58
(1, 1, 2)	(2, 0, 2, 12)	↓ 1556.08
(0, 1, 2)	(2, 0, 2, 12)	↑ 1557.12
(0, 1, 2)	(1, 0, 2, 12)	↑ 1557.16
(2, 1, 2)	(2, 0, 2, 12)	↑ 1557.84

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

SARIMAX Results						
Dep. Variable:	Sparkling	No. Observations:	132			
Model:	SARIMAX(1, 1, 2)x(1, 0, 2, 12)	Log Likelihood	-770.792			
Date:	Fri, 21 May 2021	AIC	1555.584			
Time:	00:50:44	BIC	1574.095			
Sample:	01-01-1980	HQIC	1563.083			
	-2003					
Covariance Type:	opg					
	coef	std err	z	P> z 	[0.025	0.975]
ar.L1	-0.6281	0.255	-2.463	0.014	-1.128	-0.128
ma.L1	-0.1041	0.225	-0.463	0.643	-0.545	0.337
ma.L2	-0.7276	0.154	-4.734	0	-1.029	-0.426
ar.S.L12	1.0439	0.014	72.838	0	1.016	1.072
ma.S.L12	-0.555	0.098	-5.663	0	-0.747	-0.363
ma.S.L24	-0.1354	0.12	-1.133	0.257	-0.37	0.099
sigma2	1.51E+05	2.03E+04	7.4	0	1.11E+05	1.90E+05

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.

Sparkling	mean	mean_se	mean_ci_lower	mean_ci_upper
1991-01-01	1327.375384	388.346117	566.230982	2088.519785
1991-02-01	1315.100332	402.011042	527.173168	2103.027496
1991-03-01	1621.588540	402.004648	833.673908	2409.503172
1991-04-01	1598.853714	407.242446	800.673187	2397.034242
1991-05-01	1392.689005	407.972726	593.077156	2192.300855

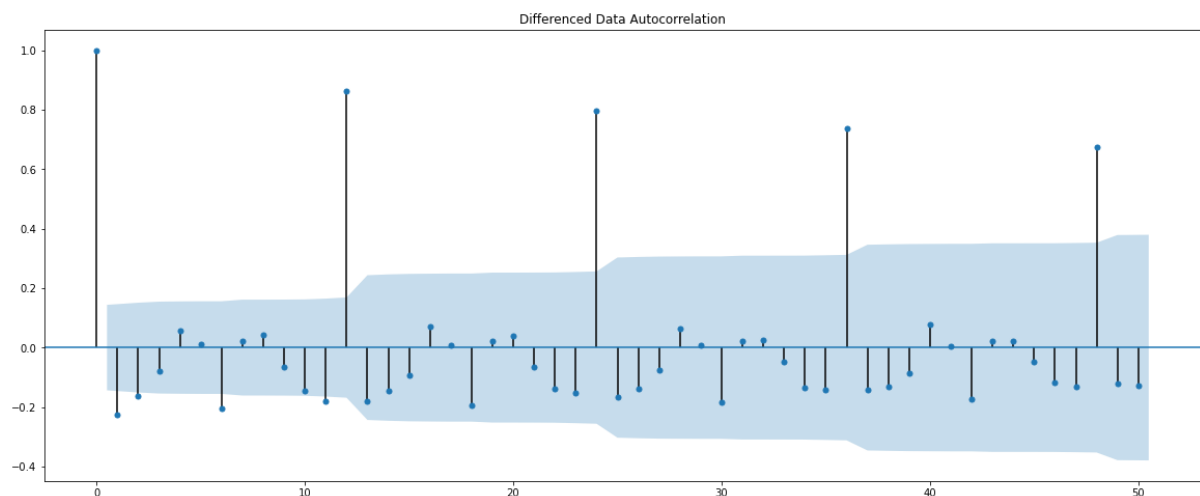
- SARIMA RMSE:-

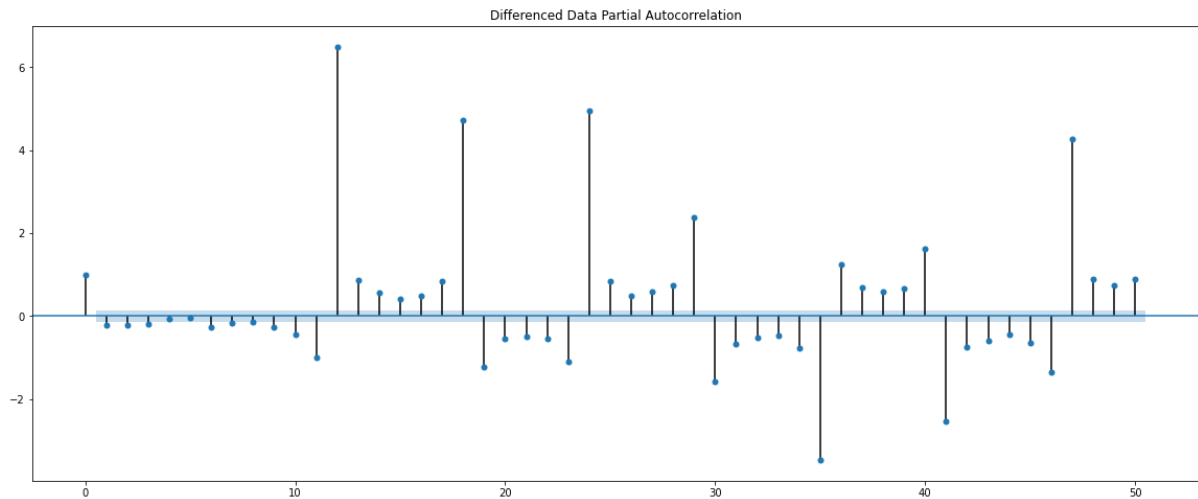
Model	Test RMSE
ARIMA(2,1,2)	1374.678
SARIMA(1,1,2)(2,0,2,12)	528.62352

The RMSE is 528.6, which is comparatively lower, so SARIMA model can be better forecaster for future Sparkling 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 Sparkling 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 3.
- 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) -> (3,1,2) values ,we have built our Arima Model and found below results.

Manual ARIMA Model Results						
Dep. Variable:	D.Sparkling	No. Observations:	131			
Model:	ARIMA(3, 1, 2)	Log Likelihood	-1107.46			
Method:	css-mle	S.D. of innovations	1106.121			
Date:	Fri, 21 May 2021	AIC	2228.927			
Time:	00-01-1900	BIC	2249.054			
Sample:	29222	HQIC	2237.106			
	-2003					
	coef	std err	z	P> z 	[0.025	0.975]
const	5.9846	3.643	1.643	0.1	-1.156	13.125
ar.L1.D.Sparkling	-0.442	5.85E-06	-75500	0	-0.442	-0.442
ar.L2.D.Sparkling	0.3079	1.53E-05	20200	0	0.308	0.308
ar.L3.D.Sparkling	-0.2501	1.32E-05	-18900	0	-0.25	-0.25
ma.L1.D.Sparkling	-0.0006	0.02	-0.028	0.978	-0.04	0.039
ma.L2.D.Sparkling	-0.9994	0.02	-49.289	0	-1.039	-0.96

From above output we have noticed that our AIC value is 2228.9, which is comparable with other models , also we can notice that except Moving average L1- lag 1 component all the p values are significant . for each Auto regressive and Moving average all the coefficient are given.

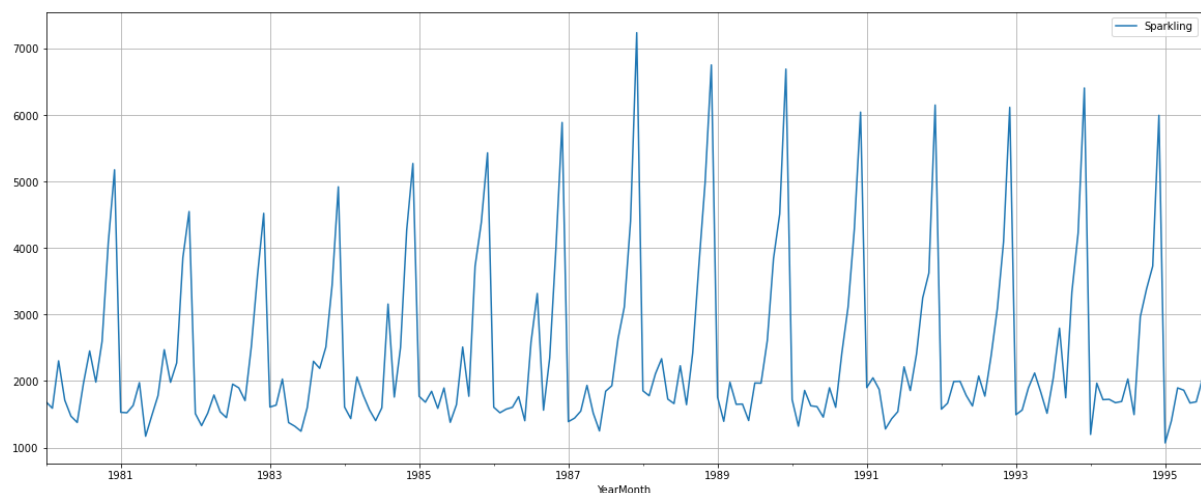
- Now we have Predicted on the Test Set using Manual Arima model and evaluated the model using RMSE , Below are the results (refer

Model	Test RMSE
Manual ARIMA(3,1,2)	1378.98

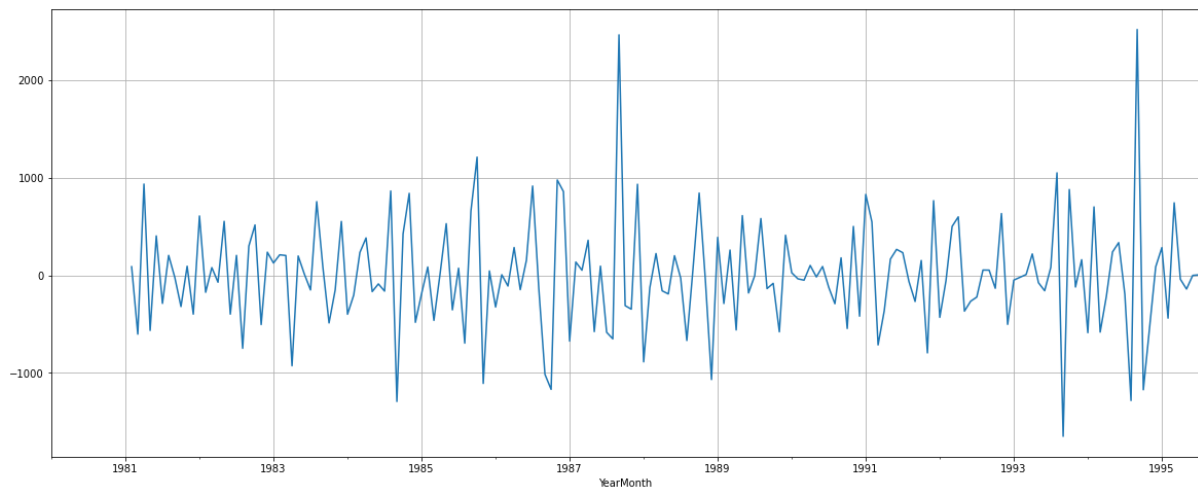
This RMSE Value is slightly higher than our automated ARIMA Model.

- **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.

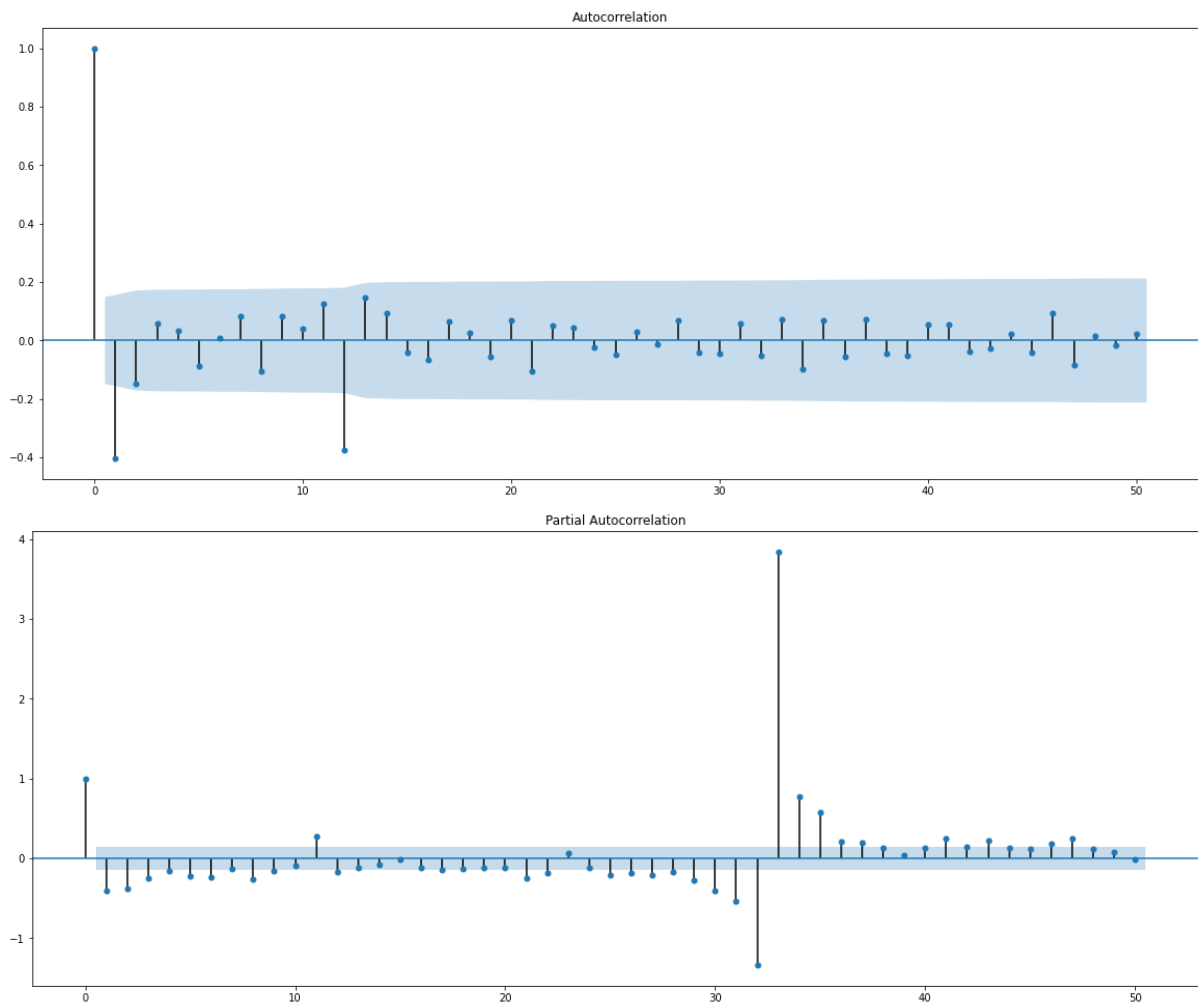


Now looking at differenced series at seasonal interval 12 below we can notice that there is almost no trend present only seasonality present in the data .



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 3, 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).

SARIMAX Results						
Dep. Variable:	Sparkling	No. Observations:	132			
Model:	SARIMAX(1, 1, 2)x(3, 1, [1], 12)	Log Likelihood	-613.167			
Date:	Sat, 22 May 2021	AIC	1242.334			
Time:	01:43:08	BIC	1261.588			
Sample:	01-01-1980	HQIC	1250.064			
	-2003					
Covariance Type:	opg					
	coef	std err	z	P> z 	[0.025	0.975]
ar.L1	-0.5743	0.32	-1.795	0.073	-1.201	0.053
ma.L1	-0.1641	0.274	-0.6	0.549	-0.7	0.372
ma.L2	-0.7412	0.207	-3.586	0	-1.146	-0.336
ar.S.L12	-0.5371	0.995	-0.54	0.59	-2.488	1.414
ar.S.L24	-0.2611	0.391	-0.667	0.505	-1.028	0.506
ar.S.L36	-0.1228	0.185	-0.663	0.507	-0.486	0.24
ma.S.L12	1.19E-01	1.00E+00	0.119	0.905	1.85E+00	2.09E+00
sigma2	1.83E+05	3.06E+04	5.963	0	1.23E+05	2.43E+05

From above output we have noticed that our AIC value is 1242, which is comparatively better with other models, also we can notice that except Moving average L1- lag 1, L2 –Lag 2 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
Manual SARIMA(1,1,2)(3,1,1,12)	350.83963

We have gained a very good RMSE value & 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 combined all the Models along with their respective results of the RMSE & created a data frame as below .

Model	Test RMSE
Alpha=0.1,Beta=0.9,Gamma=0.6,TripleExponentialSmoothing	↓ 338.45
Manual SARIMA(1,1,2)(3,1,1,12)	↓ 350.84
Alpha=0.0831,Beta=3.92 * e-09,Gamma=0.49,TripleExponentialSmoothing	↓ 362.74
SARIMA(1,1,2)(2,0,2,12)	↓ 528.62
2pointTrailingMovingAverage	↓ 813.40
4pointTrailingMovingAverage	↓ 1156.59
SimpleAverageModel	↓ 1275.08
Alpha=0.00,SimpleExponentialSmoothing	↓ 1275.08
6pointTrailingMovingAverage	↓ 1283.93
9pointTrailingMovingAverage	↓ 1346.28
ARIMA(2,1,2)	↓ 1374.68
Alpha=0.1,SimpleExponentialSmoothing	↓ 1375.39
Manual ARIMA(3,1,2)	↓ 1378.98
Linear Regression OnTime instance	↓ 1389.14
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	→ 1778.56
NaiveModel	↑ 3864.28

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.

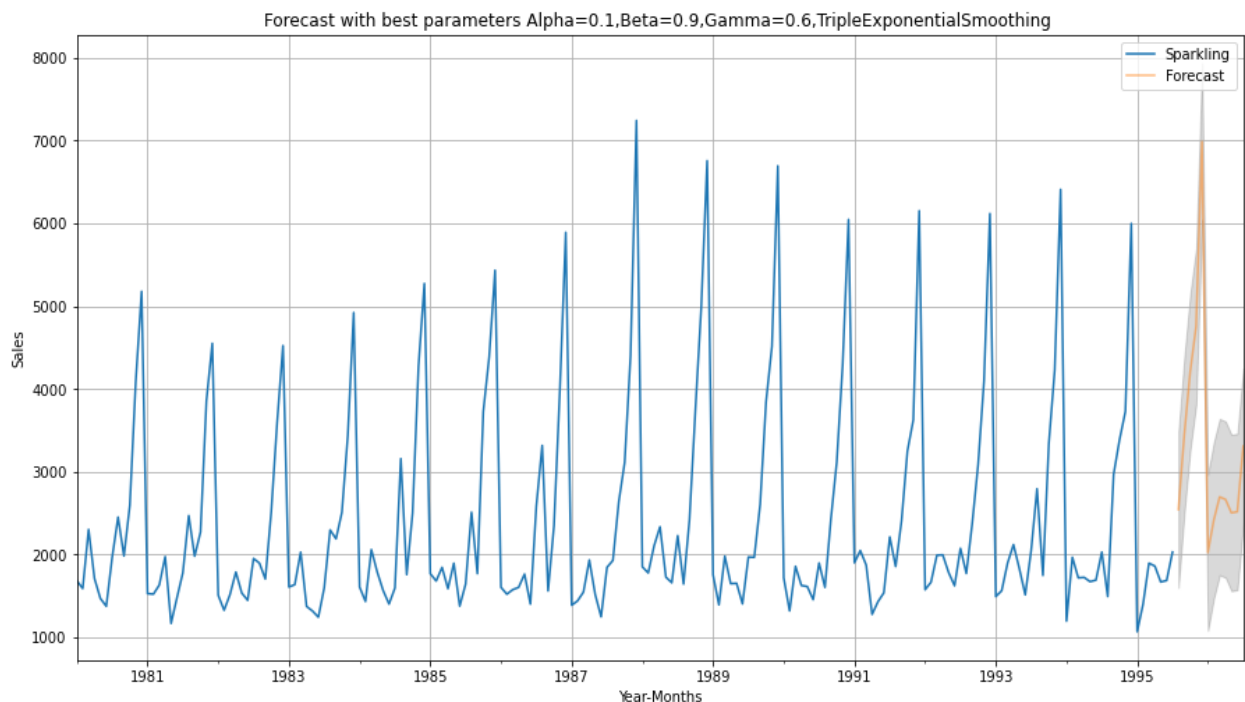
As we can see from above RMSE Table that Alpha=0.1, Beta=0.9, Gamma=0.6, TripleExponentialSmoothing having lowest RMSE So it is the best Model & 2nd model is SARIMA model with (1,1,2) (2,0,2,12) because of second lowest RMSE -350.84 . So we will Build top 2 Models for predicting the Sparkling wine sales for next 12 Months.

- **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 Sparkling wine sales by passing Trend as additive & seasonality as additive and found the Full Model RMSE -> 480.2
- 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	1600.89	2544.76	3488.63
01-09-1995	2506.60	3450.47	4394.34
01-10-1995	3263.10	4206.97	5150.84
01-11-1995	3811.83	4755.70	5699.57
01-12-1995	6044.32	6988.19	7932.06
01-01-1996	1081.40	2025.26	2969.13
01-02-1996	1471.02	2414.89	3358.76
01-03-1996	1753.53	2697.39	3641.26
01-04-1996	1722.49	2666.36	3610.23
01-05-1996	1560.45	2504.32	3448.19
01-06-1996	1574.31	2518.18	3462.05
01-07-1996	2364.64	3308.51	4252.38

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 around 7000 units band.



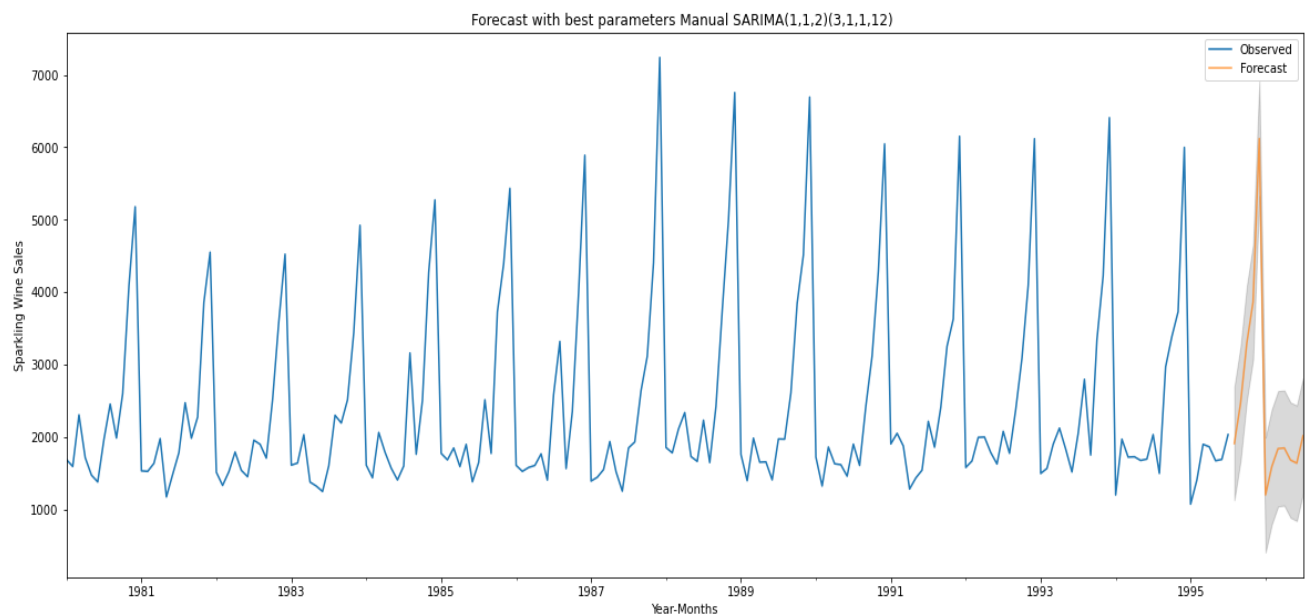
- Prediction for next 12 Month using Manual SARIMA (3,1,2) (3,1,1,12)
- We have fitted Our Model to Full dataset of Sparkling wine sales by passing below parameters and found Full Model RMSE -> 555.85

```
model = SARIMAX(df['Sparkling'], order=(1,1,2),
seasonal_order=(3,1,1,12), 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_low er	mean_ci_upper
01-08-1995	1908.37	398.80	1126.74	2690.00
01-09-1995	2463.01	403.29	1672.58	3253.44
01-10-1995	3295.40	403.46	2504.63	4086.17
01-11-1995	3862.59	404.76	3069.28	4655.89
01-12-1995	6121.56	404.79	5328.19	6914.93
01-01-1996	1197.77	405.38	403.24	1992.29
01-02-1996	1587.26	405.54	792.42	2382.11
01-03-1996	1836.93	405.93	1041.31	2632.54
01-04-1996	1846.61	406.17	1050.53	2642.70
01-05-1996	1679.37	406.50	882.64	2476.10
01-06-1996	1636.30	406.78	839.03	2433.56
01-07-1996	2013.26	407.08	1215.39	2811.13

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 forecasted Graph we can infer that average Sparkling sales are somewhat increasing from last year.

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 Sparkling wine of company ABC Estate Wines, so from analysis we have noticed that Sparkling wine sales are increasing up to year 1988 & after that it is having decreasing trend there is a clear seasonality can also be seen.

Since in previous questions we have built a table sorted in increasing RMSE and found that **Triple Exponential Smoothing Model Alpha=0.1, Beta=0.9, Gamma=0.6** is having lowest RMSE of 339 this means that for building the Predictive model for Sparkling wine sales you have to take smoothing factor for Level is 0.1 & Smoothing factor for Trend is 0.9. That infers Trend factor is playing biggest role in predicting next year Sparkling wine sales & since seasonality smoothing factor is 0.6 so seasonality is playing less important role compared to trend. RMSE is the comparable factor, Lower the RMSE means that best optimized Model.

We can also observe the following.

-  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 Sparkling 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 decade The trend is increasing & in second decade trend is decreasing.
-  In the month of June average wine sales are lowest

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

- ✓ As the ABC Estate Wines Company's Sparkling Wine highest sale is in 4th Quarter so company can increase the price of wine along with offering free packaged snacks with it.
- ✓ 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 December people will buy more wine may be due to Christmas & New year party, so company should maintain sufficient stock and distribution channels to sell more, also they can offer Combo packages with other brands of ABC Estate Wines.
- ✓ 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 and make good ad films, Post it on media platform.
- ✓ If we talk about Indian Prospective direct ad films & marketing is prohibited for wines so they can introduce one Cross product in the market to increase their sales (Like Tuborg have introduced their soda in the market).

- ✓ Most of the peoples are free on their weekends , so company can offer weekly discount and offers so that Sales can be increased.