Time Series Forecasting Project

by Anisha Sharma

PGPDSBA.O.Mar23.A

Great Learning

Table of Contents

1.Read the data as an appropriate Time Series data and plot the data.	2
2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.	9
3. Split the data into training and test. The test data should start in 1991.	2
4. Build all the 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.	16
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.	4

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.	11
7. Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.	2
8. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.	3
9. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales. Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.	5

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

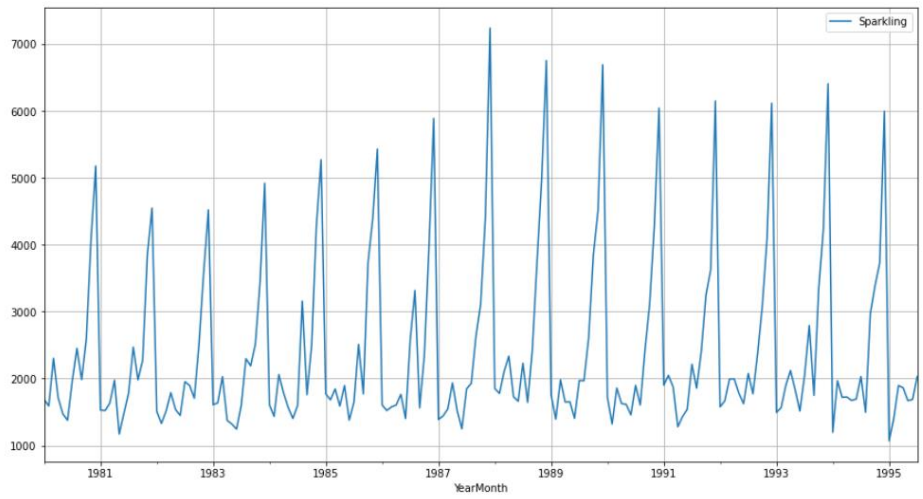
1. Reading the Data have imported both the data series in two different ways. As we can observe, each entry has an YearMonth value with it, which is not really a datapoint, but an index for the sales entry. So in reality the datasets have a single column that contains the quantity of wines sold in that particular month. Here, while reading the datasets I have given the argument in a way so that it parses the first column which is date column, and indicates to the system that this is a one column series through squeeze.

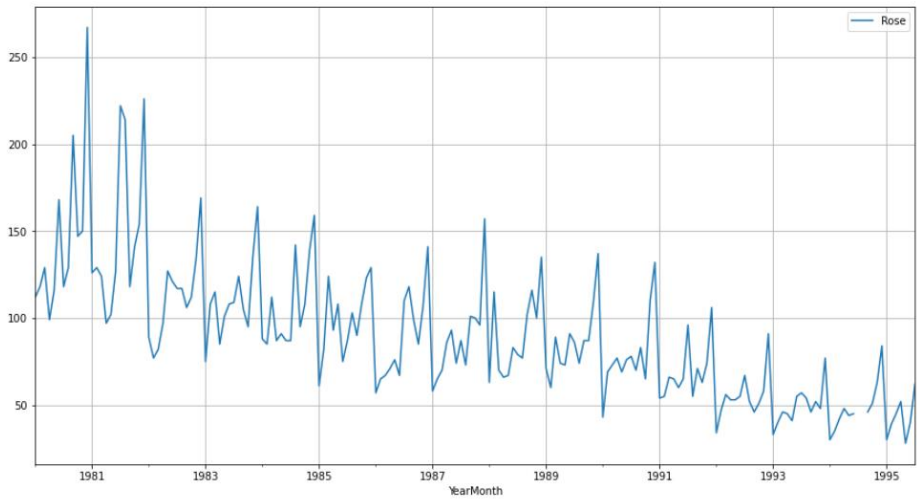
	YearMonth	Sparkling
0	1980-01-01	1686
1	1980-02-01	1591
2	1980-03-01	2304
3	1980-04-01	1712
4	1980-05-01	1471

	Rose	YearMonth
0	112.0	1980-01-31
1	118.0	1980-02-29
2	129.0	1980-03-31
3	99.0	1980-04-30
4	116.0	1980-05-31

Reading Wine Datasets It can be observed that both the datasets have data starting from January 1980 going till July 1995, so there are 187 entries in totality in each dataset.

1.Plotting the Data Now that I have uploaded the datasets with no arguments (and hence uploaded the datasets without parsing the dates here), I will need to provide a timestamp value by ourselves. In addition to that I have removed theYearMonthvariable and added a time stamp to the dataset myself. I have plotted both the time series below.





As we can observe from the above plots, the sales for Rose Wine are showing a declining trend and the sales for Sparkling wines are showing an upward trend. There is a certain seasonality element that is visible in the graphs. We will explore the trend and seasonality further during decomposition, where we will be able to view a much detailed report on these two factors.

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

EDA analysis:

1.Null value check, I got this output. We have 0 null values in Sparkling and 2 in Rose.

```
Sparkling    0    Rose    2
dtype: int64      dtype: int64
```

As we can see, the Rose dataset contains 2 Null values and there are no Null values for Sparkling dataset; I addressed the Rose dataset Null values using linear interpolation so as to obtain the imputed values in place of the values that are missing. Post the imputation, I confirmed that there are no more Null values in the Rose dataset.

2. Duplicate value check: There are no duplicate entries in the datasets as each value correspond to a different time index, so basically these are all sales figures for different months.

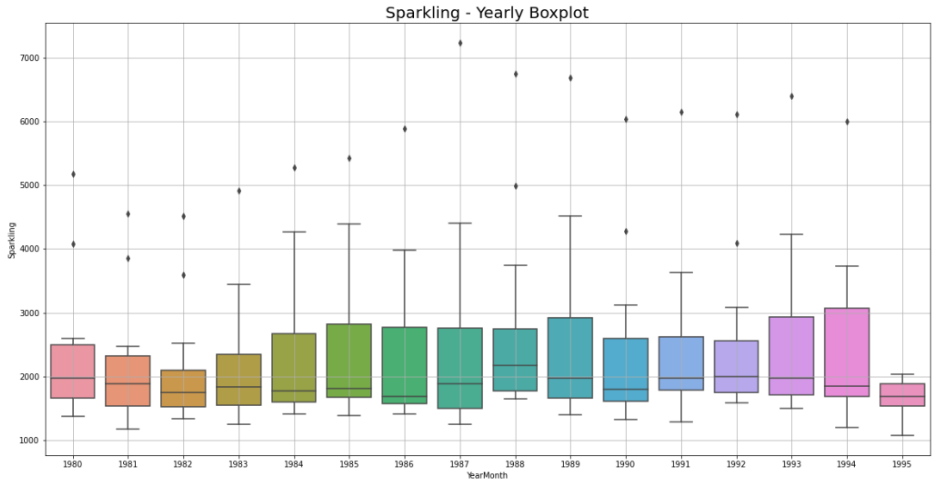
	count	mean	std	min	25%	50%	75%	max
Sparkling	187.0	2402.417112	1295.11154	1070.0	1605.0	1874.0	2549.0	7242.0

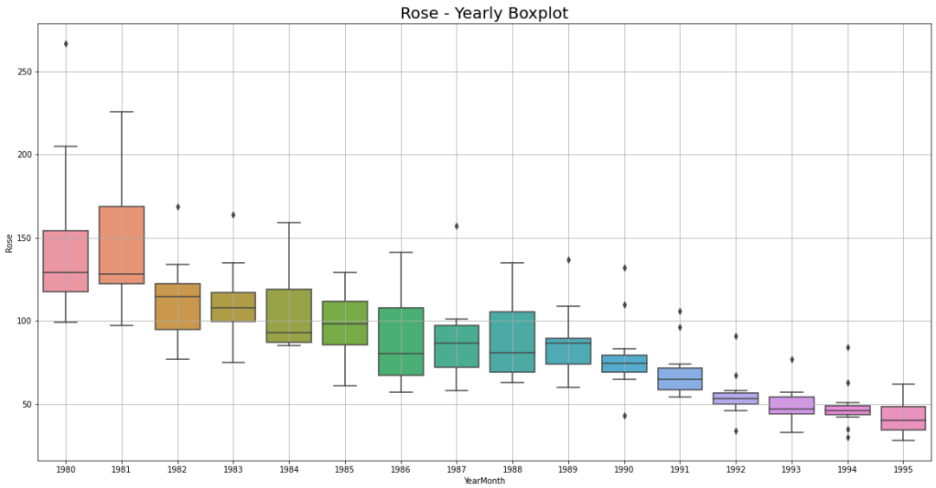
	count	mean	std	min	25%	50%	75%	max
Rose	185.0	90.394595	39.175344	28.0	63.0	86.0	112.0	267.0

As we can see from the above, both the wine sales time series data look like they are skewed. There is High Standard Deviation for both the time series since the Min and Max have significant differences

between them. Moreover, there is difference between the mean and the median for the same reason of skewness. As mentioned earlier, there are in total 187 records in both the datasets.

3.Outlier detection: Following are the yearly boxplots for the two wine sales time-series

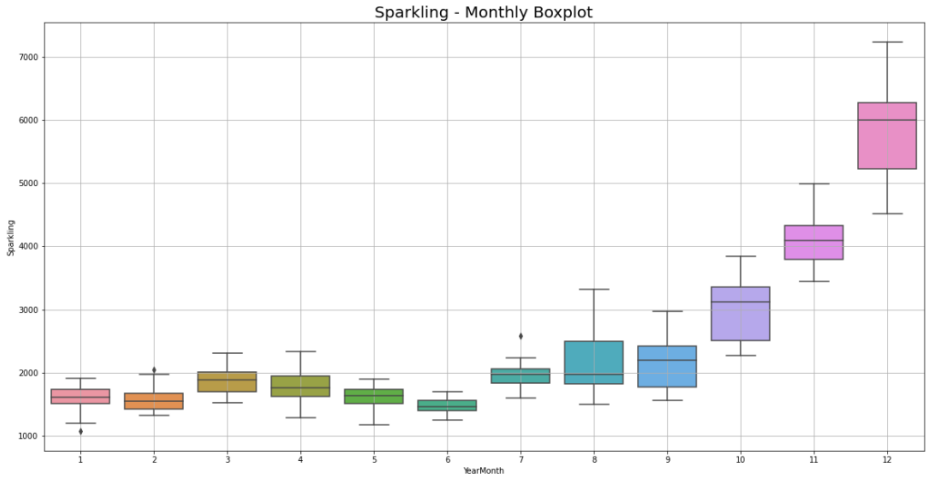


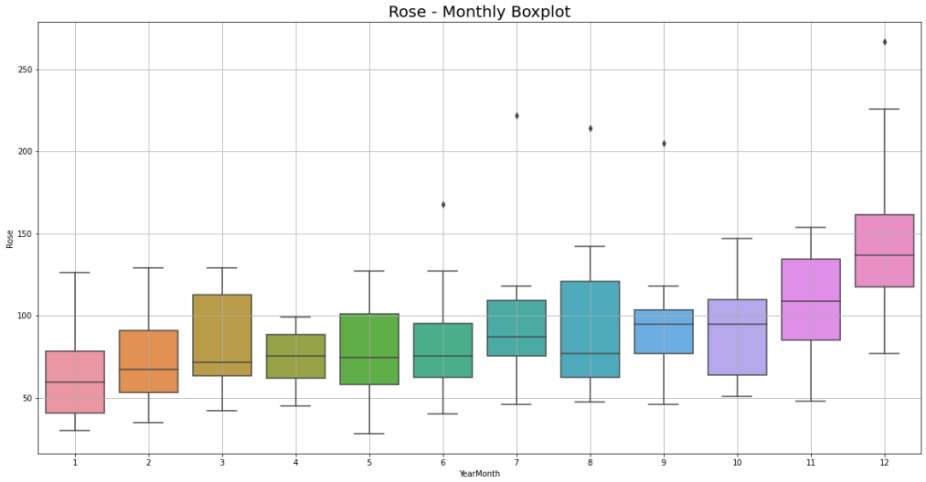


As we can observe from the above plots, Rose wine has mostly a downward sales trend. The highest sales for Rose wine can be observed in 1981 and the lowest sales in 1994(because the 1995 sales seem to be doing well, considering the data is only till July month and reaching to the 1994 level already in 7 months itself). The highest variation in monthly sales for Rose wine seems to be in the year 1981 and on the year 1994 there seems to be the lowest variation in monthly sales. As we can observe, the Sparkling Wine sales have a variation each year, the years 1985and 1986 seem to be the years with the least variation, so the 2 years show certain consistency in terms of sales. The highest sales for Sparkling Wine seems to happen in the year 1994 and the lowest in the year 1982. Based on the 1995 data of 7 months (till July), it is difficult to comment on the sales performance of that year. The Sparkling wine sales appear to be going down from the year 1980 and have started increasing from 10the year 1983. The variation in Sparkling Wine sales seem to be increasing for the period1983-1986, while the highest variation in Sparkling wine sales is in the year 1994. There is clear skewness that can be observed for

Sparkling wine sales for all the years, except maybe in 1981. There are outliers in the yearly sales data, however as it is a Time Series, we can ignore the outlier data.

4. Monthly Sales Across Years - Rose ,Wine Monthly Sales Across Years - Sparkling Wine





As can be observed from the above two sets of tables and graphs, the months of December seems to be the month that drives the highest sales figures for both Rose and Sparkling Wines. The second highest sales for Sparkling being in November while Rose wine shows a mixed trend, with highest sales in August or July for certain years. We can observe a seasonality element in the graphs for both Rose and Sparkling wines.

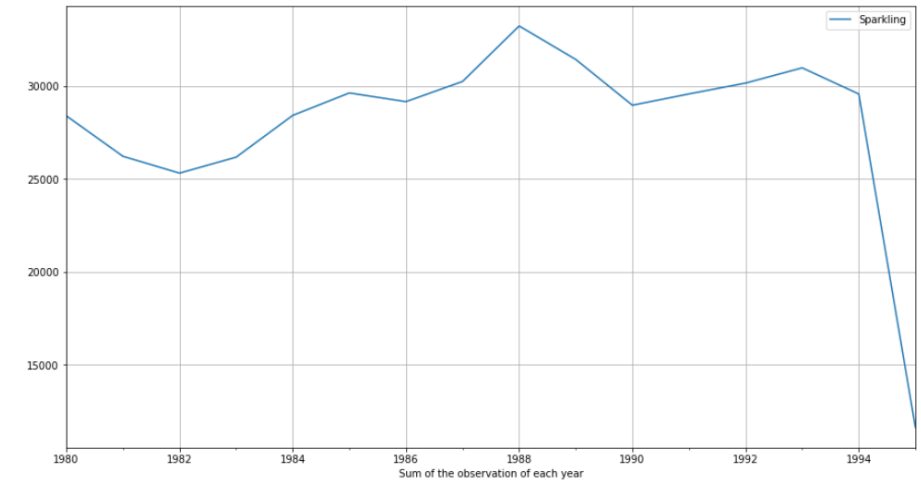
5.Yearly Sum of Observations The yearly sum of sales numbers can be observed in the following tables and graphs:

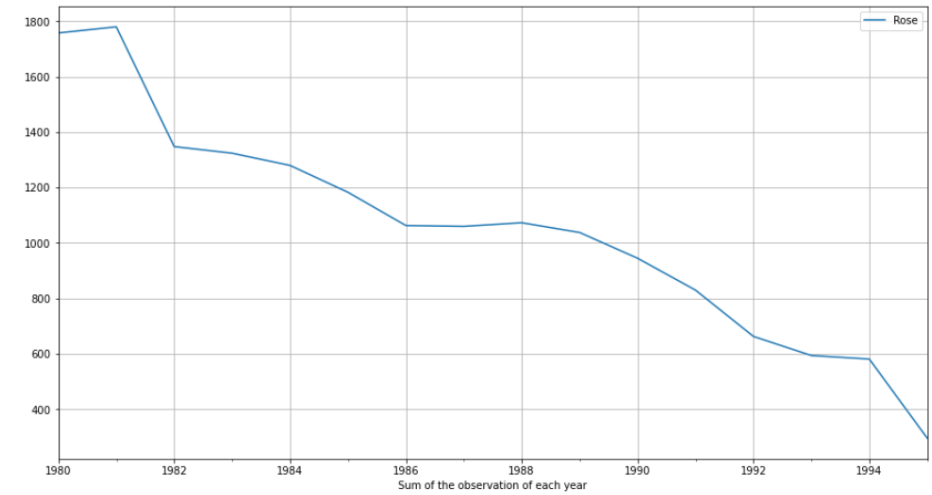
Sparkling

YearMonth	
1980-12-31	28406
1981-12-31	26227
1982-12-31	25321
1983-12-31	26180
1984-12-31	28431

Rose

YearMonth	
1980-12-31	1758.0
1981-12-31	1780.0
1982-12-31	1348.0
1983-12-31	1324.0
1984-12-31	1280.0

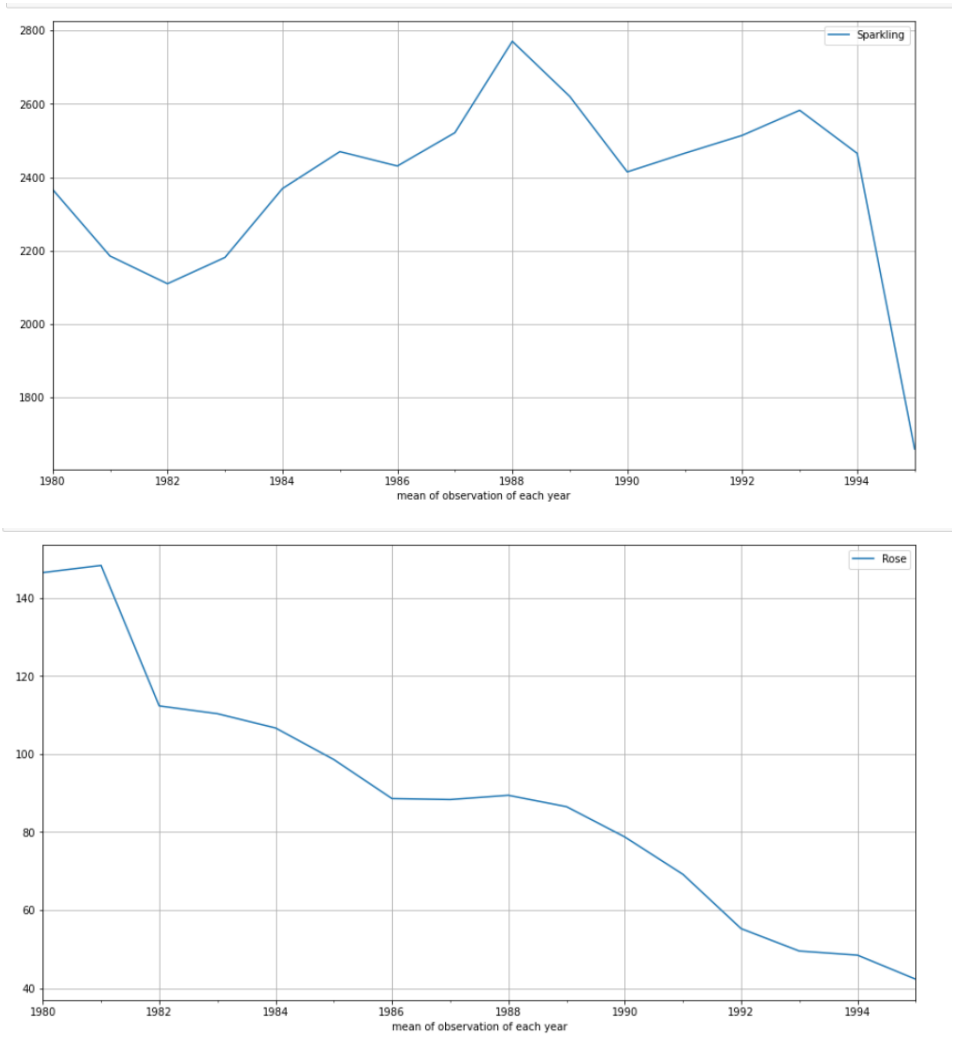




As can be observed from the above summation tables and the plotted graphs, Rose wine annual sales year on year observe a downward sales trend. While the sales figures for

Sparkling wine show a dip initially with sales picking up from the year 1982 right up to the year 1988 and then observing another dip in the sales. The steep drop post 1994 for both Rose and Sparkling wine is because of the relatively less (half year data - till July)data available for the year 1995.

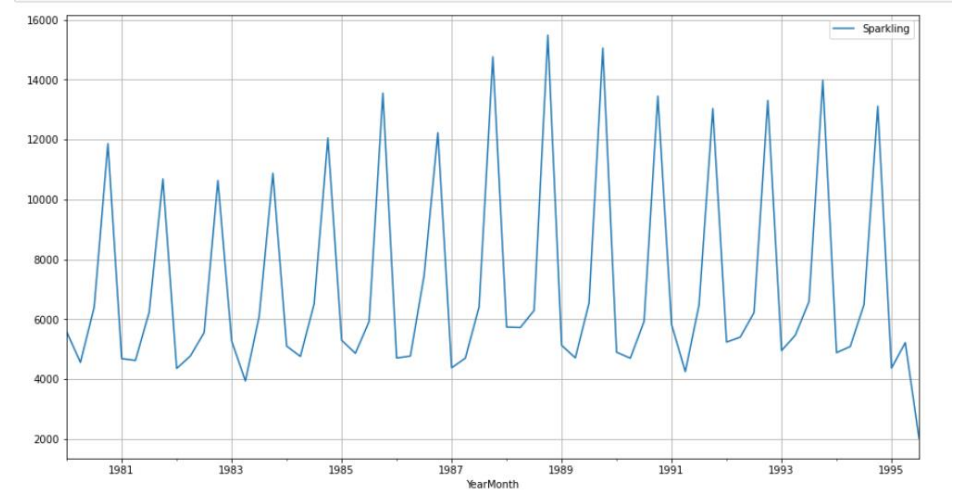
6.Mean of Observations of each Year:

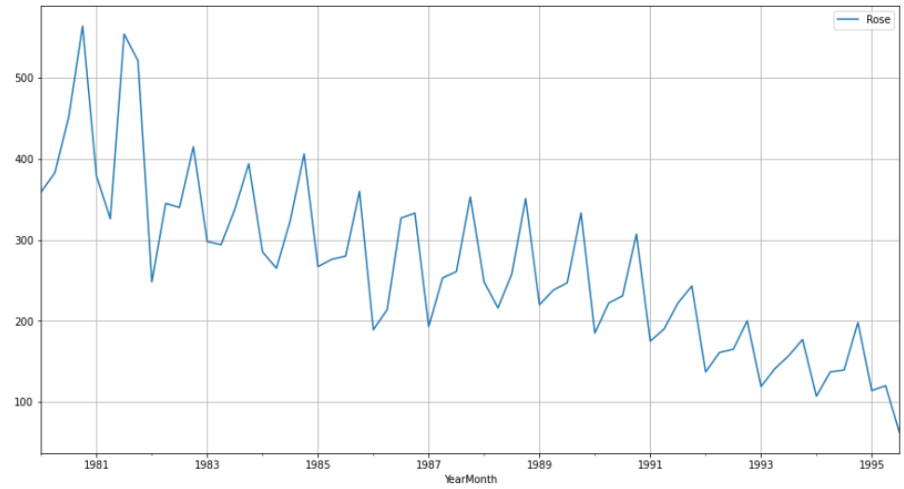


Mean of Observations of each Year From the above tables and graphs, we can confirm the observations from the previous section. An added observation would be that the mean sales for Rose wine are much lesser than that of Sparkling wines.

7.Sum of Observations of each Quarter The quarterly sum of sales numbers can be observed in the following tables and graphs:

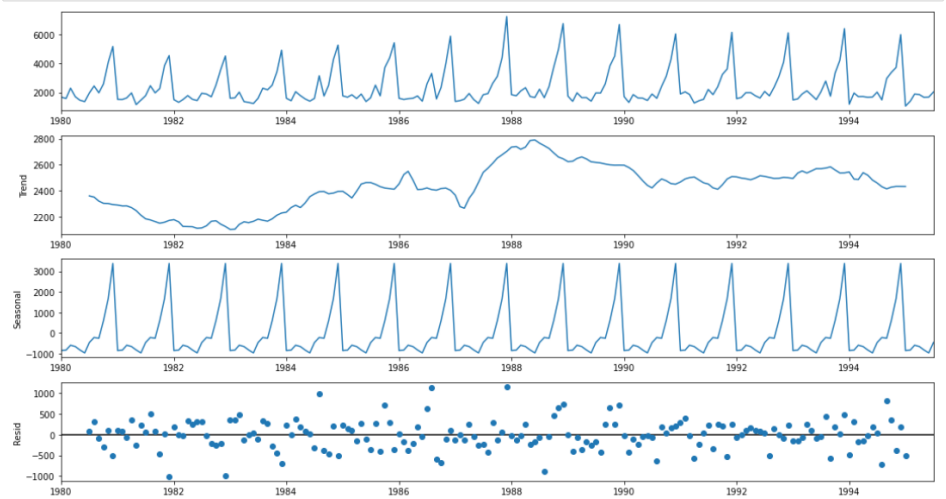
Sum of Observations of each Quarter From the below tables and graphs showing the quarterly sum of sales figures for both the datasets, we can observe that the Quarterly sales show a downward trend for Rose wineand an upward trend for Sparkling wine. Also there is a slight element of seasonality in both the time series datasets.

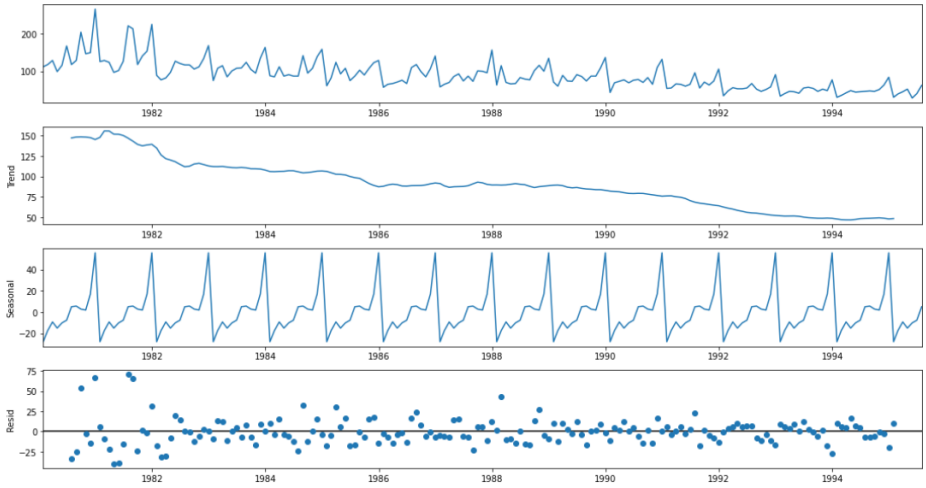




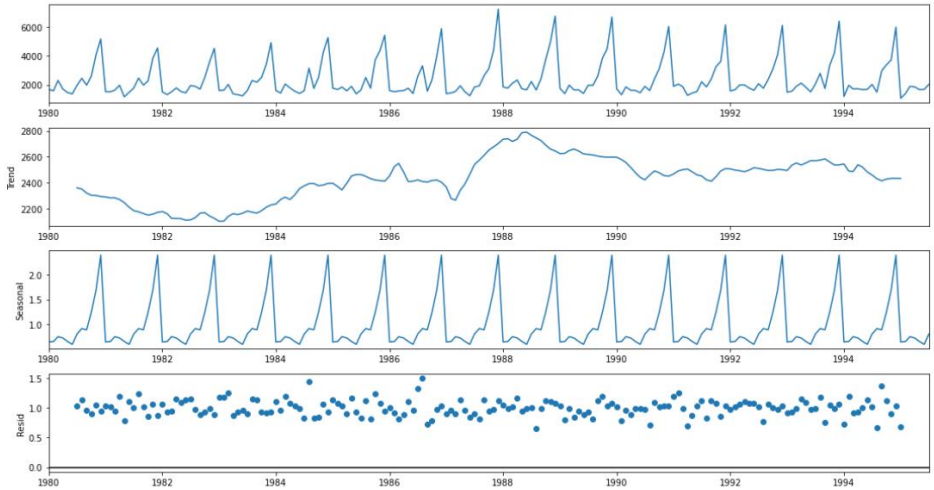
8.Decomposition I have provided the decomposed elements for both the Time Series below:

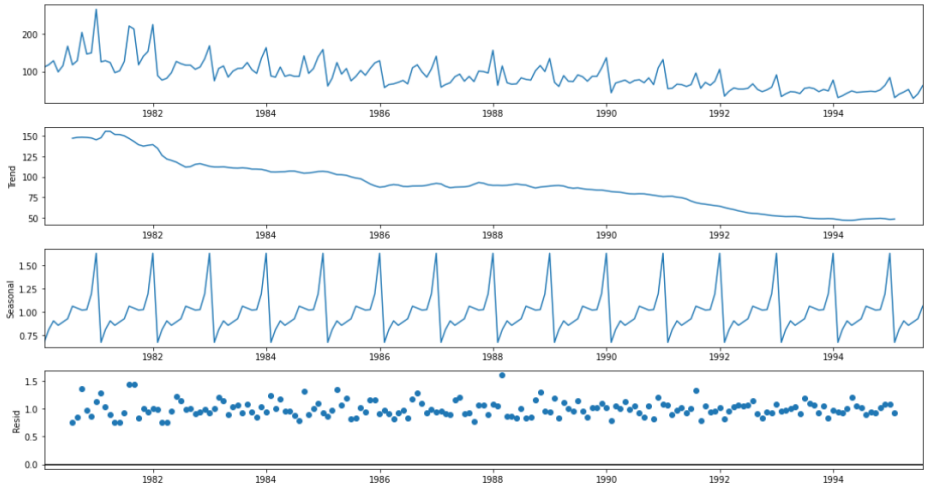
ADD:




Trend
Seasonal
Resid

MUL:


Trend
Seasonal
Resid


Trend
Seasonal
Resid

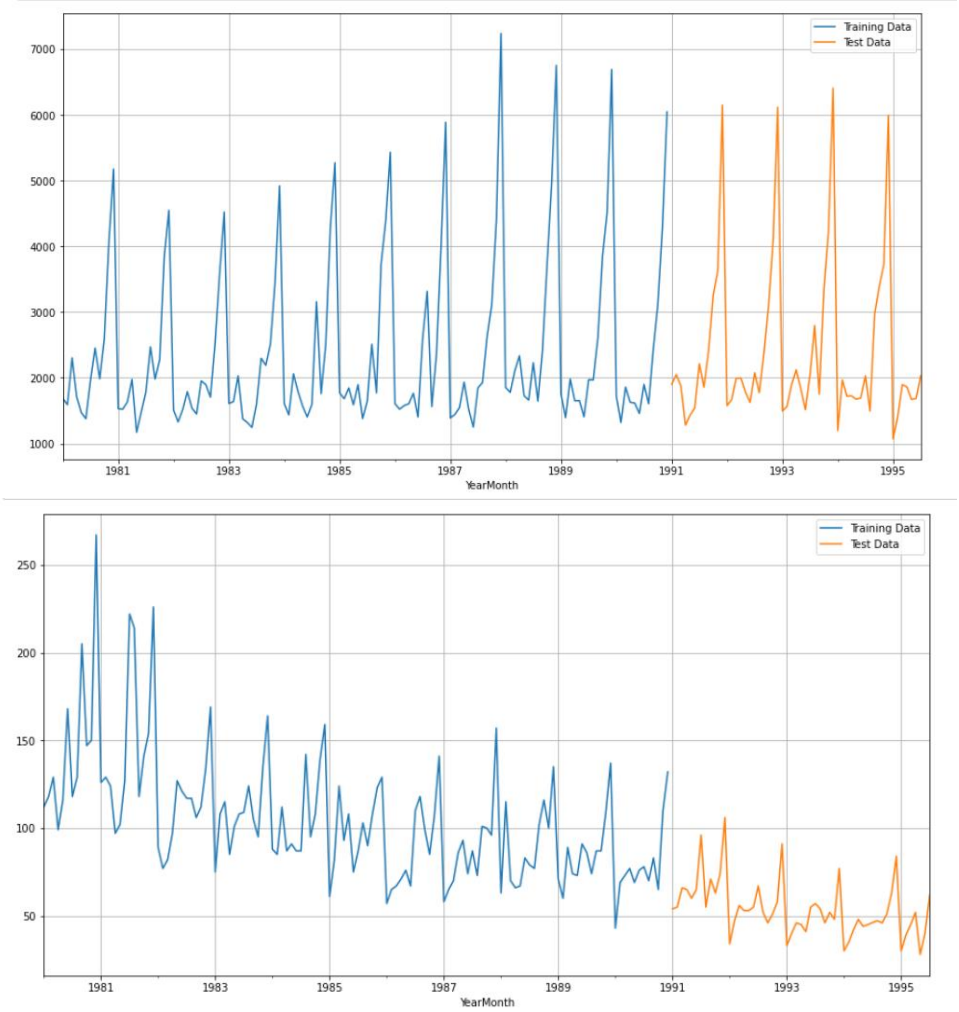
Multiplicative Decomposition - Rose and Sparkling Wine We can see the decomposition of the two time series above. I have tried with both additive and multiplicative decomposition for both time series so that I can determine if the wine datasets are a multiplicative or additive series.

As we can observe from the above, we can say that the wine time series are clearly multiplicative in nature and both have a seasonal component. We can also observe again that the Rose wine sales depict a downward sales trend and the Sparkling wine sales show an upward sales trend. The plots above clearly indicate that the Wine sales are unstable and not uniform, and they have an apparent seasonality trend. Moreover, the seasonal variation seems to be more in the case of Sparkling wine as compared to the Rose wine; while the sales variation seems to be more in case of Rose wine as compared to Sparkling wine.

3.Split the data into training and test.

The test data should start in 1991. I have split the time series datasets into Train and Test datasets below. It is given the question that the Test Data should start in 1991, so I have used 71% of the datasets for the Training dataset (instead of the usual 70%) and the rest of the dataset for Test datasets.

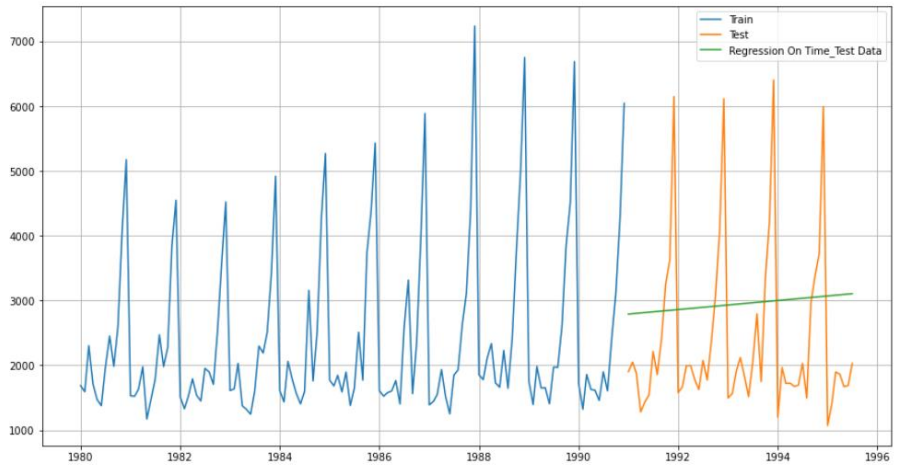
I have also confirmed that the Train dataset indeed ends in 1990, and the Test datasetindeed starts in 1991 by using the Head and Tail functions on the Training and Testdatasets. As we can observe, the size of the Train data frame is 132 observations and thatof the Test data frame is 55 observations.I have also plotted the Train and test data frames for both time series datasets below:



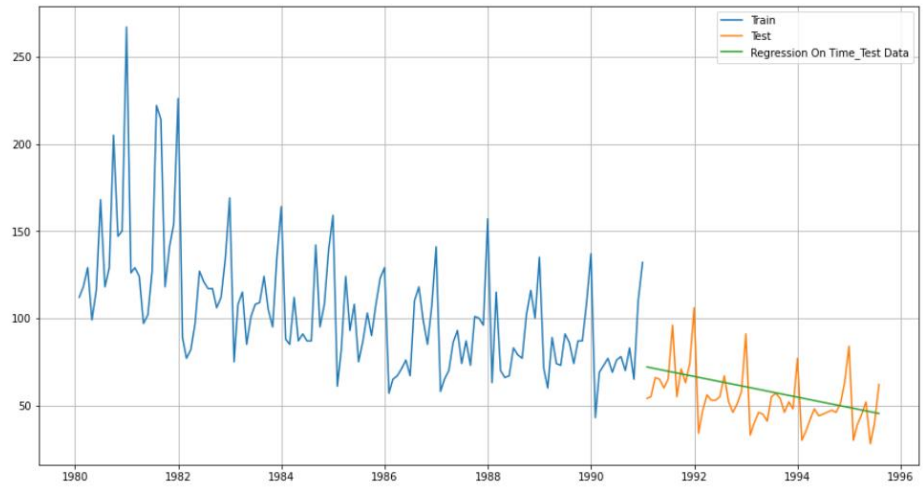
We can observe the training and test data in the above plots, the Orange part of the plotsdepicts the Train datasets (January '80 – December '90), and the Blue part of the plotsdepict the test datasets(January '91 – July '95).

4. Build all the 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.

Linear regression: The extracts of Training and Test data for the Linear Regression for Sparkling dataset can be seen below:

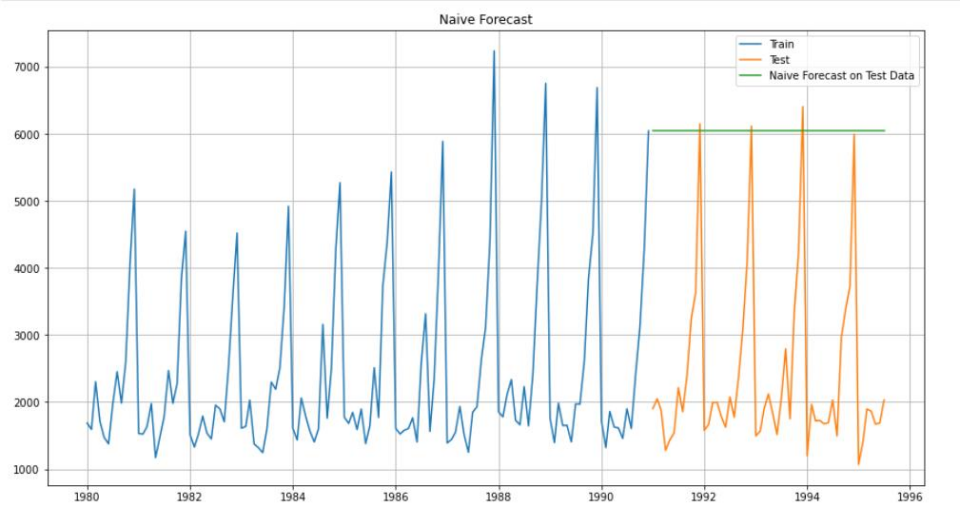


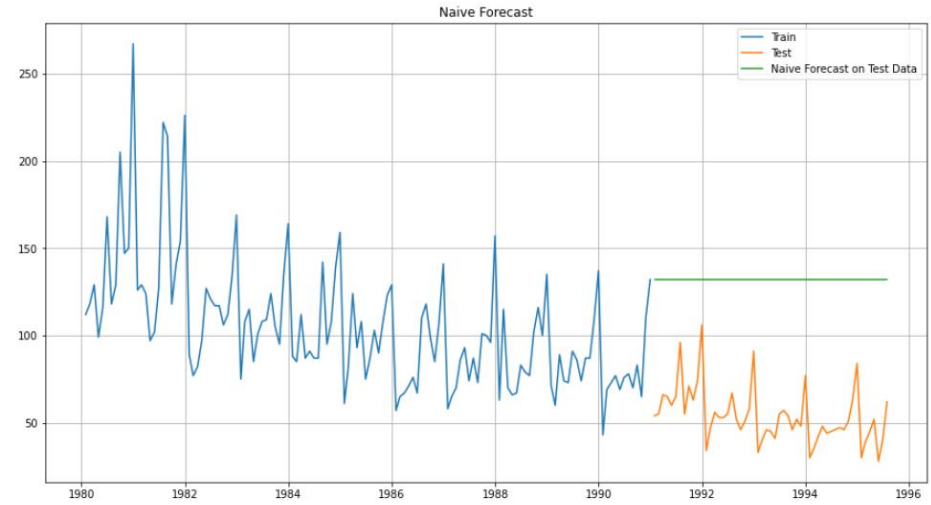
For Rose Dataset-



The Regression plots above depict the regression on test set as the blue line. As we can observe from the above plots and metrics, Rosewine sales show a downward trend, and the Sparkling wine sales show an upward trend.For RegressionOnTime forecast on the Test Data for Rose wine, RMSE = 15.25 and For RegressionOnTime forecast on the Test Data for Sparkling wine, RMSE = 1389.13

Naïve Model: for sparkling and rose respectively.



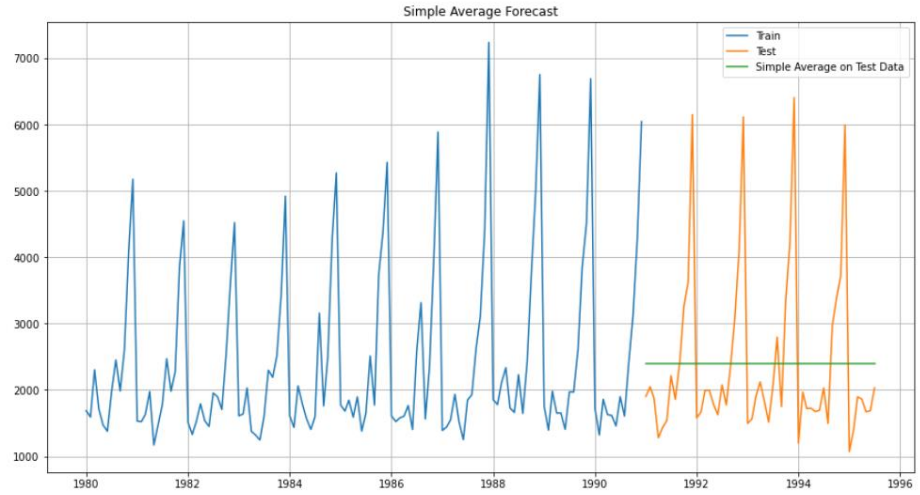


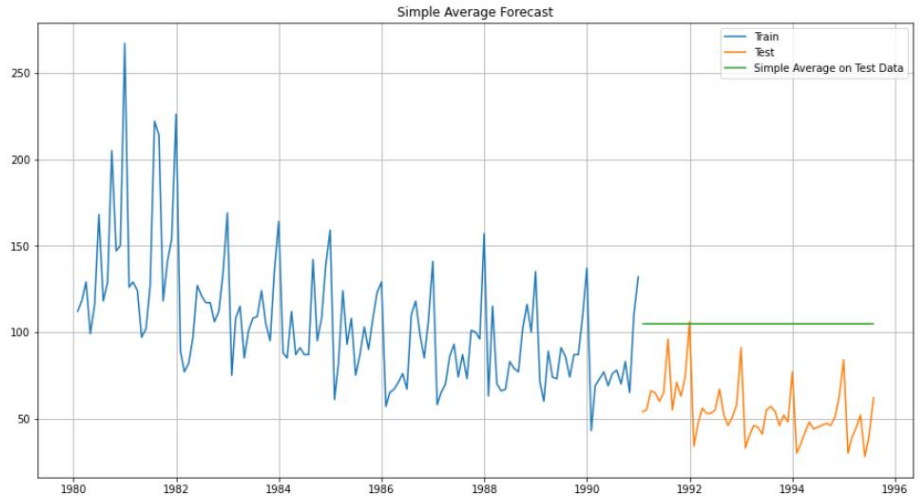
	Test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352

	Test RMSE
RegressionOnTime	15.255435
NaiveModel	79.672238

As can be seen from the Naïve model performance for Rose and Sparkling wine datasets above, the Naïve model is not suitable for any of the wine datasets since the forecasts depends on the previous last observation.

Simple Average model:





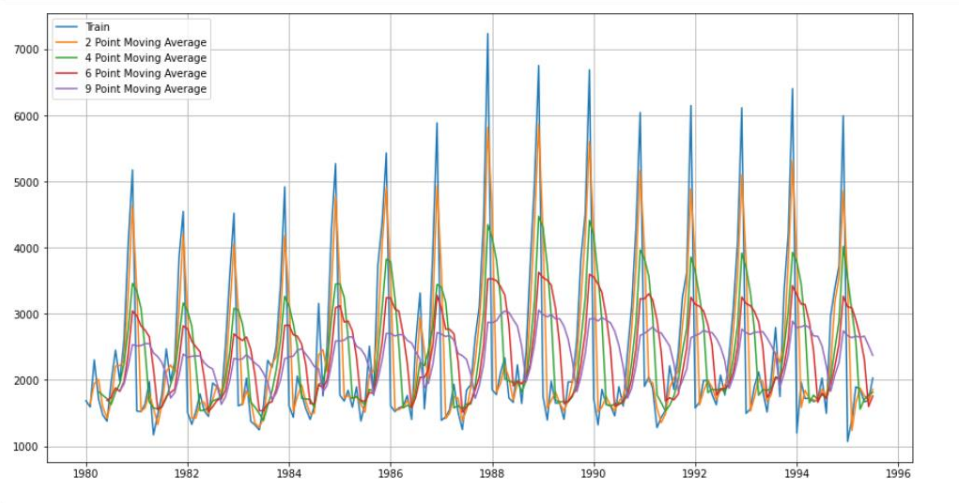
	Test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804

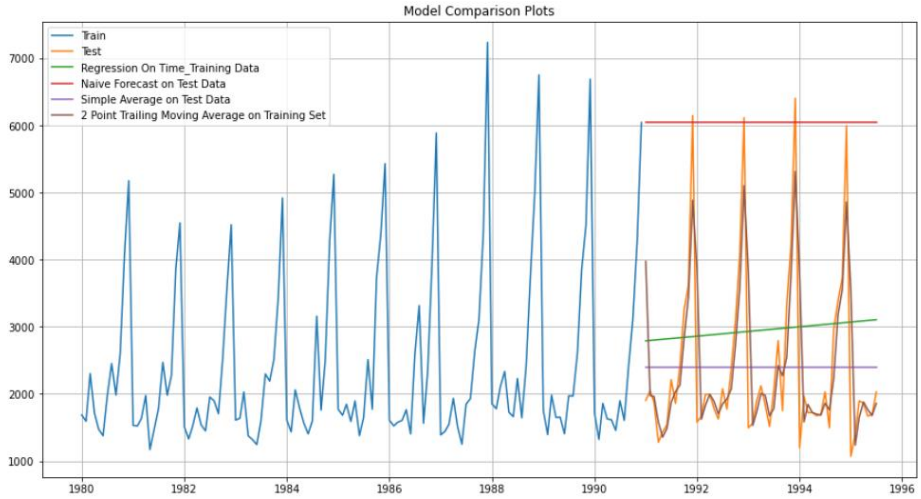
	Test RMSE
RegressionOnTime	15.255435
NaiveModel	79.672238
SimpleAverageModel	53.413057

As can be seen from the Simple Average model performance for Rose and Sparklingwine datasets above, the Linear Regression model has the best performance among allthe three models run till now for the Rose wine dataset; while the Simple Average modelshows the best performance among all the three models run till now for the Sparklingwine dataset.

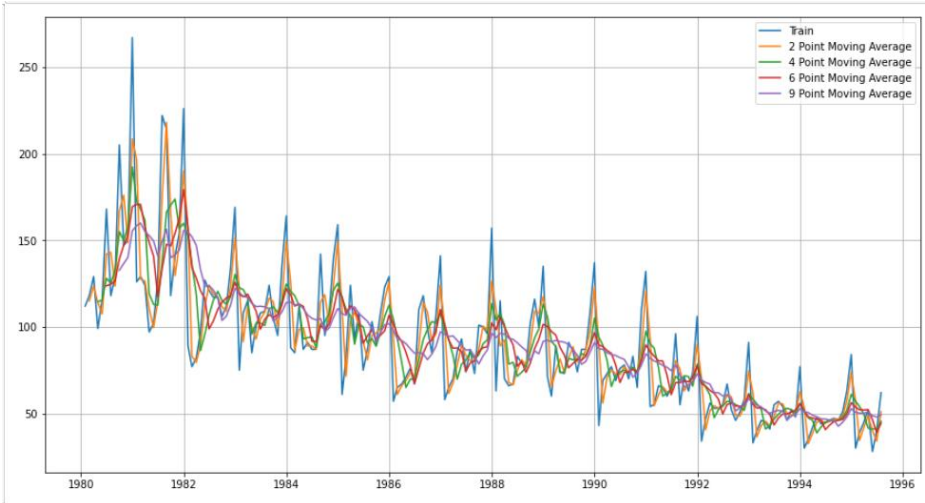
Moving average:

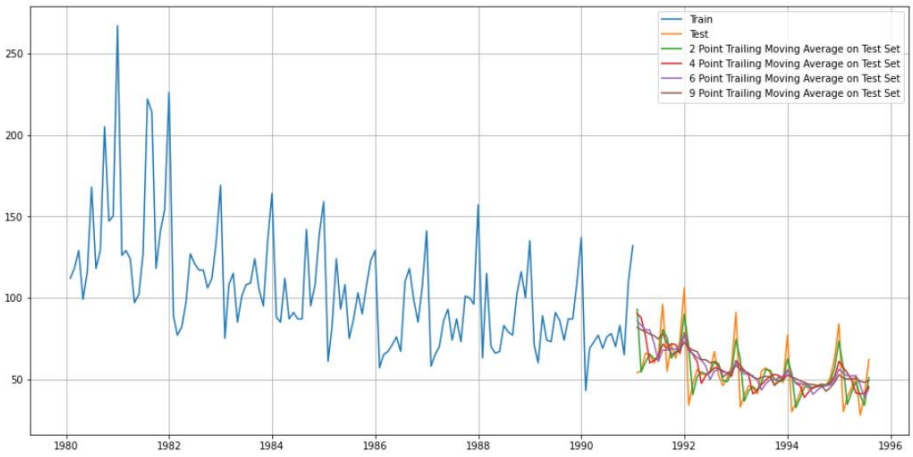
For sparkling





For Rose





As we can observe from the above plots, all of the trailing average plots show predictionvalues below the actual train and test data sets, and the 9 point trailing average plotshows the lowest prediction of all the plots. The closest prediction to actual data is shown by the 2 point trailing moving average model. This observation is corroborated by theRMSE scores for each of these moving average models.As can be seen from the summarized performance of all the models, the 2 point movingaverage has shown the best performance of all the models run on the Rose and Sparkling wine dataset.

Simple Exponential: