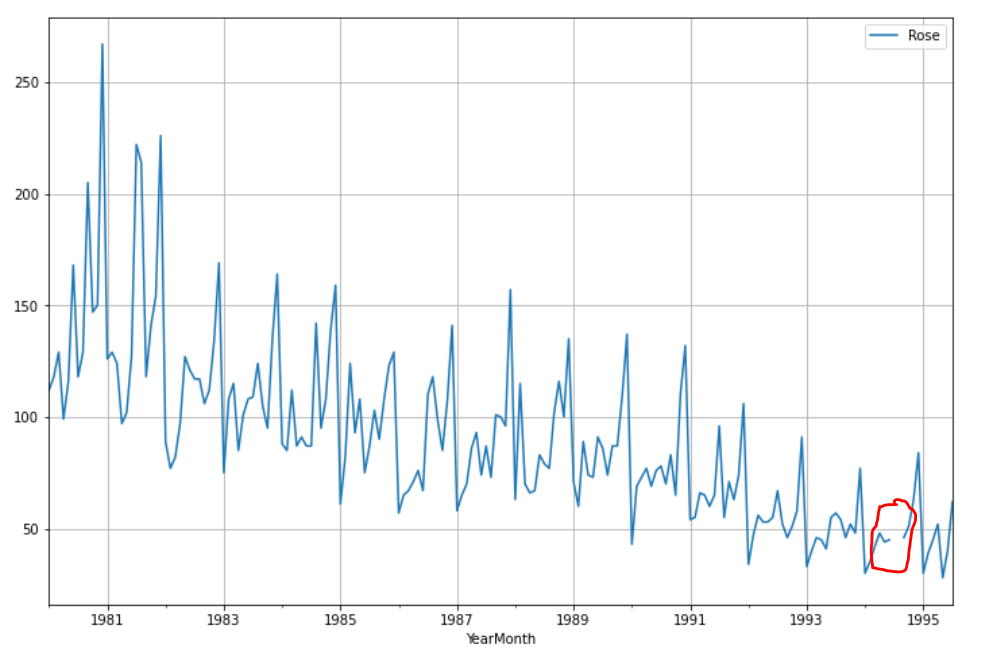
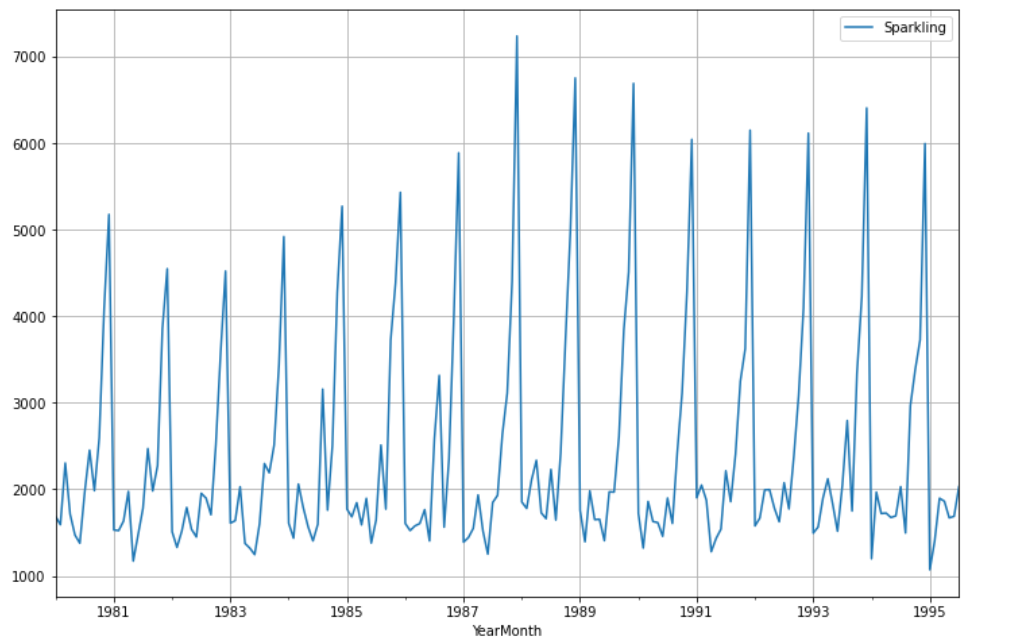
**For this particular assignment, the data of different types of wine sales in the 20th century is to be analysed. Both of these data are from the same company but of different wines. As an analyst in the ABC Estate Wines, you are tasked to analyse and forecast Wine Sales in the 20th century.**

**Q1: Read the data as an appropriate Time Series data and plot the data.**

Ans**:** After reading the data set both for Sparkling and Rose Wine:

1. When we check for the data type of the Year Month column which is considered as object in python so to change that while we are reading the file, we have to use parse\_date parameter present inside the read csv command.
2. Once that is achieved and once again if the data type for the Year Month column is checked it will be data time data type.
3. Once the data is read the plot is built for Sparkling and Rose Wine which looks something like this:

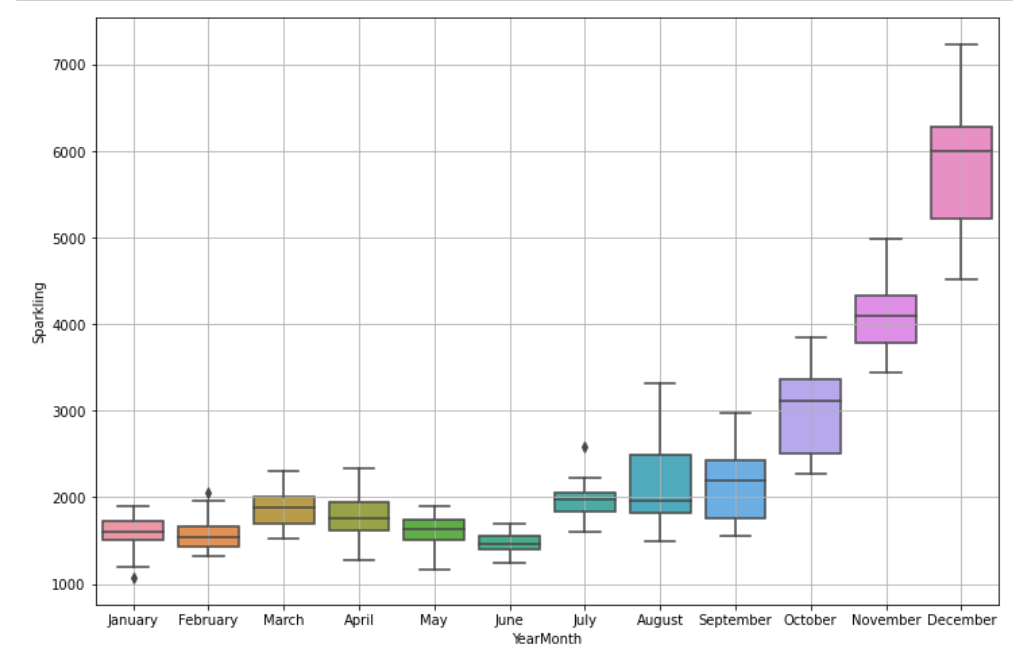


1. The red circle drawn for the Rose Wine graph shows the missing values for the year 1994.

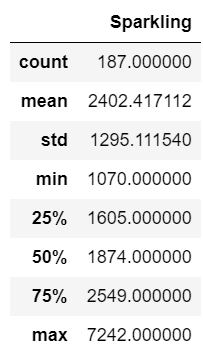
**Q2: Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.**

Ans: **Sparkling Wine EDA**:

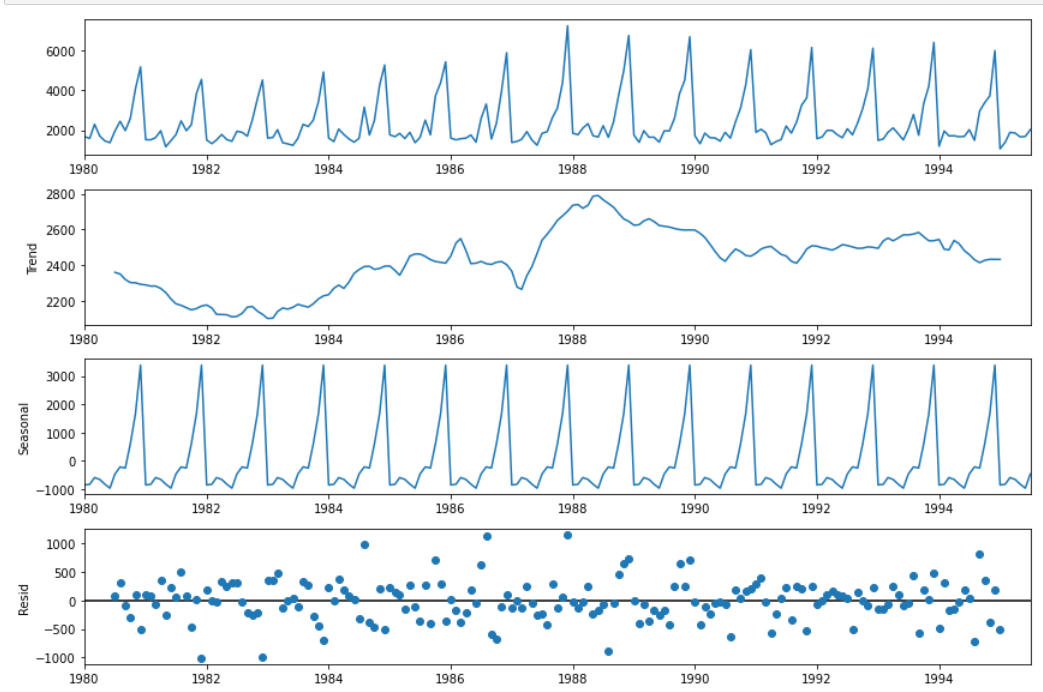
1. There were no missing or null values for the Sparkling Wine data set.
2. The outliers plot for the Sparkling Wine looks something like this



1. There are no such major outliers observed in the data set.
2. The description for the Wine looks something like this:

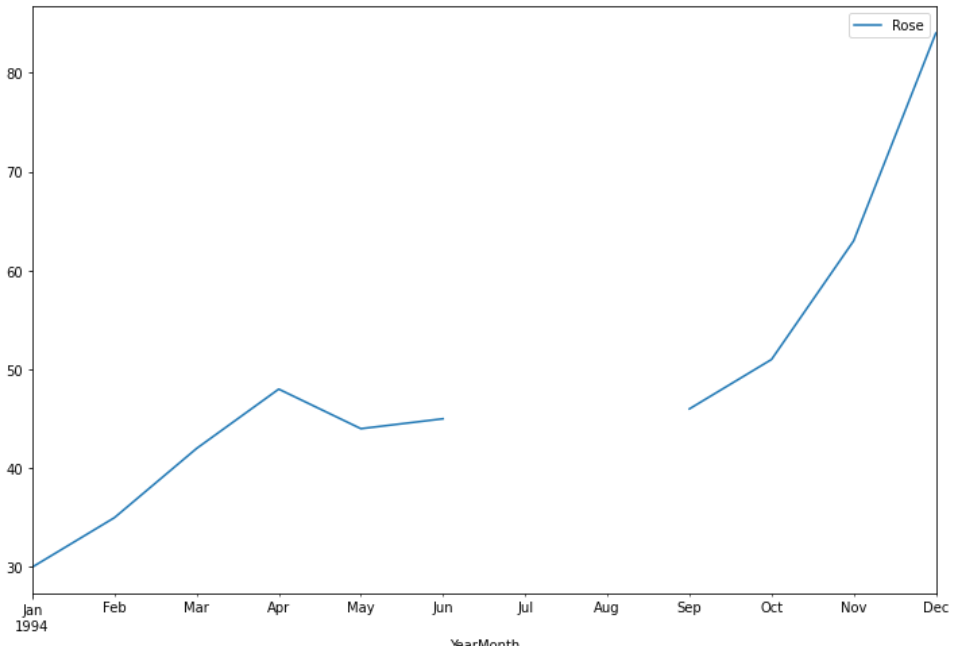


1. From the decomposition plot we can observe that the trend is not that much constant.
2. The Seasonality is constant and the most residual points are near the centre axis

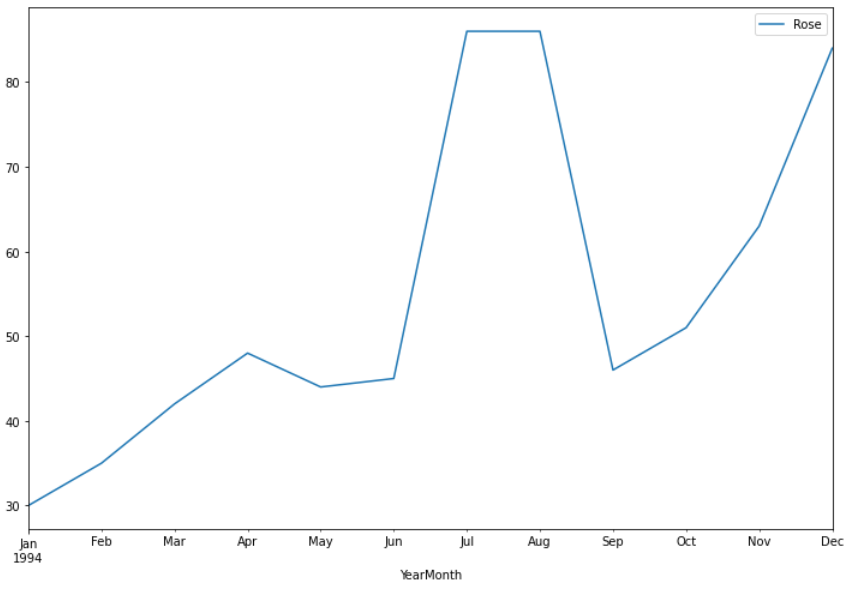


**Rose Wine EDA:**

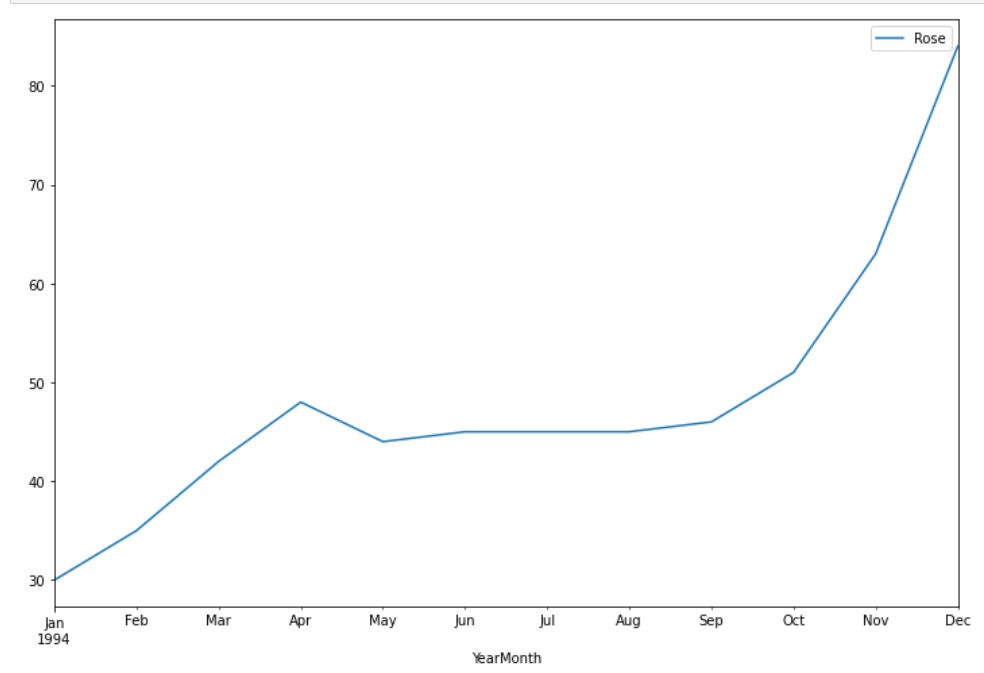
1. In the data set there were missing values in the year 1994 and as we know that if there are missing values in the data set then we cannot plot the decomposition chart.
2. So, first approach which comes in our mind is to use the median or mode and fill the missing data.
3. But when we only check the data for 1994, we observe that



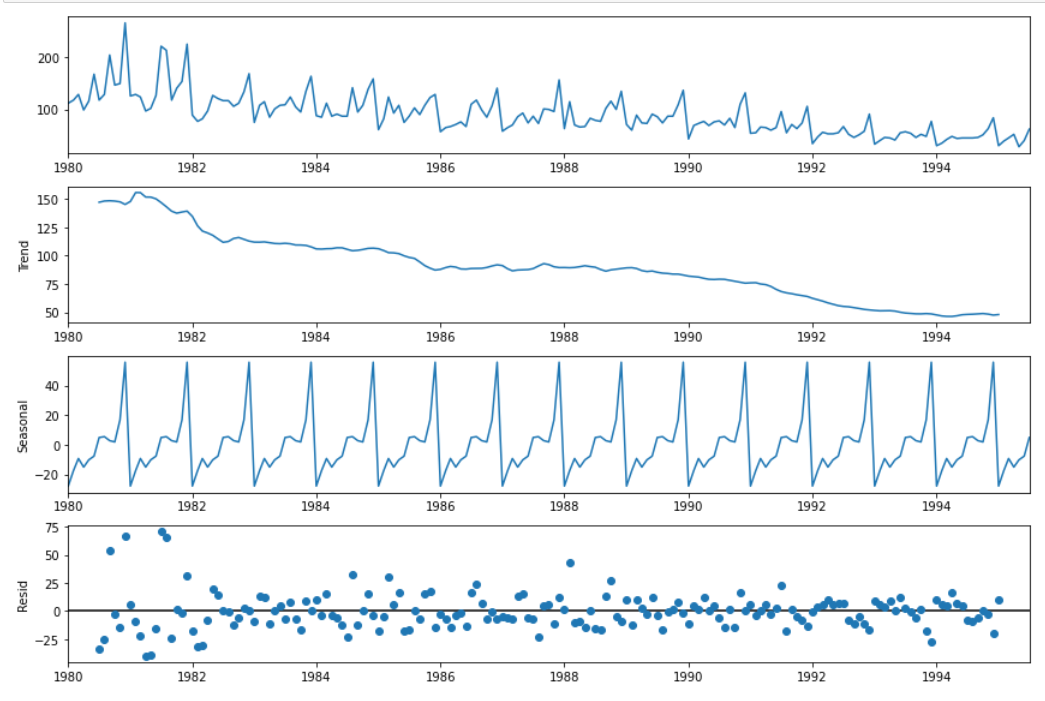
1. The median and mode which we will get from this data set will be 86 and 87 respectively.
2. Now, if we replace missing values with the median or mode value then the graph will look for year 1994 like this



1. Now this sudden spike is not at all explainable and can give us wrong observation so to avoid this when we check the data, we observe that for the year 1994 the graph is steadily going up so for that if we check the previous months data for this year, we can use the same values which in this case we will be taking into consideration the value 45.
2. So, when we use the value 45 the graph looks something like this.



1. As we can observe the above graph this can be explained quite easily as there is no sudden spike.
2. The decomposition plot for Rose Wine looks something like this

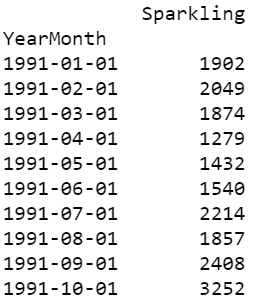
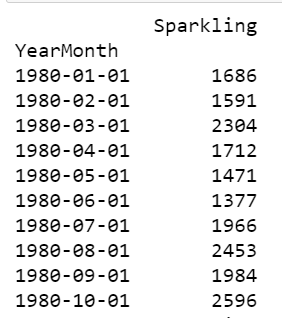


1. The trend for Rose wine is kind of on the declining side

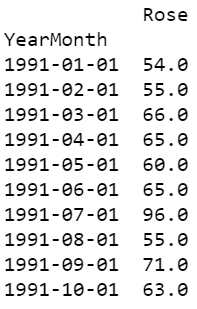
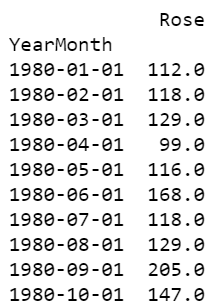
**Q3: Split the data into training and test. The test data should start in 1991.**

Ans: The data when split into training and testing data we have the data something like this.

**Sparkling Wine training and testing data**



Rose Wine Training and Testing Data

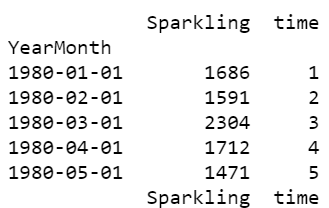


**Q4**: **Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data.  
Other models such as regression, naïve forecast models, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE.**

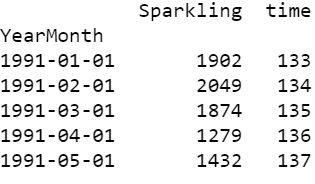
Ans: When the training and testing data is achieved by splitting, we start building different models to check which model provides us with better RMSE [Root Mean Squared Error]. As the name suggests that RMSE means error so as much as error is possible should be less. The purpose of these models is that which model gives us less RMSE value.

**Linear Regression:**

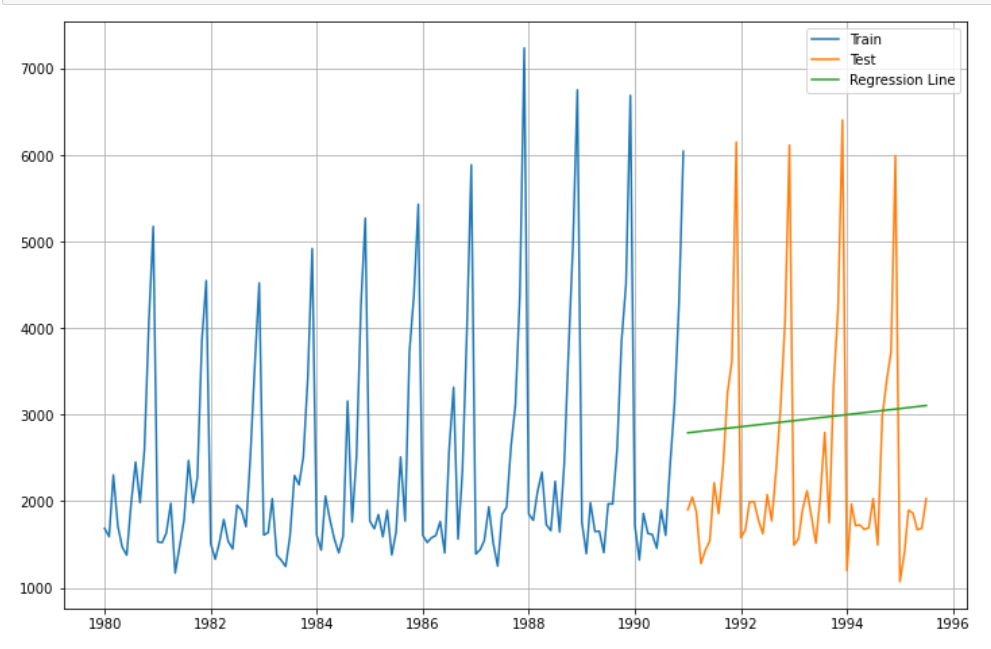
1. In Linear Regression first step is to assign the index starting from 1 to the training data which will look something like this and store it in the new feature named time.



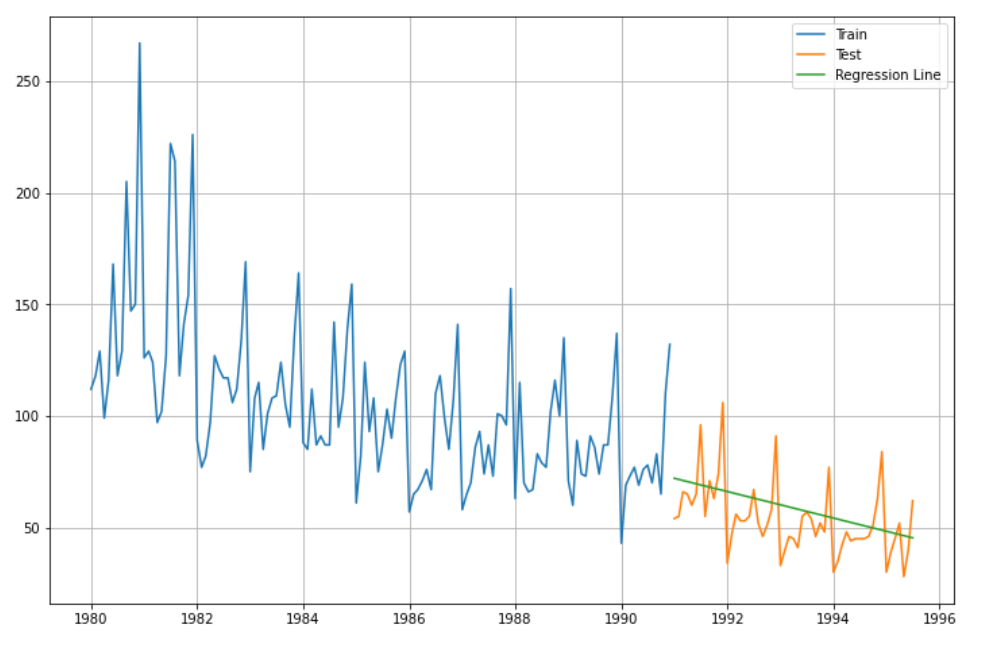
1. From where the time ends in training data which in the above case is 132 then from 133 to the last row of the main data set in this the last row number is 187 these index numbers are assigned to testing data.



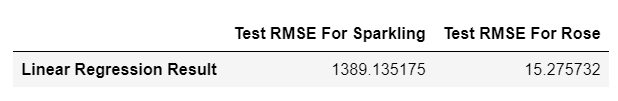
1. When this is done the data is fitted for the training data.
2. Then the predicted values are calculated for the testing data.
3. When this is achieved, we plot the graph for the training data, testing data and the predicted values of the testing data.
4. Sparkling Wine Plot:



1. Rose Wine Plot:

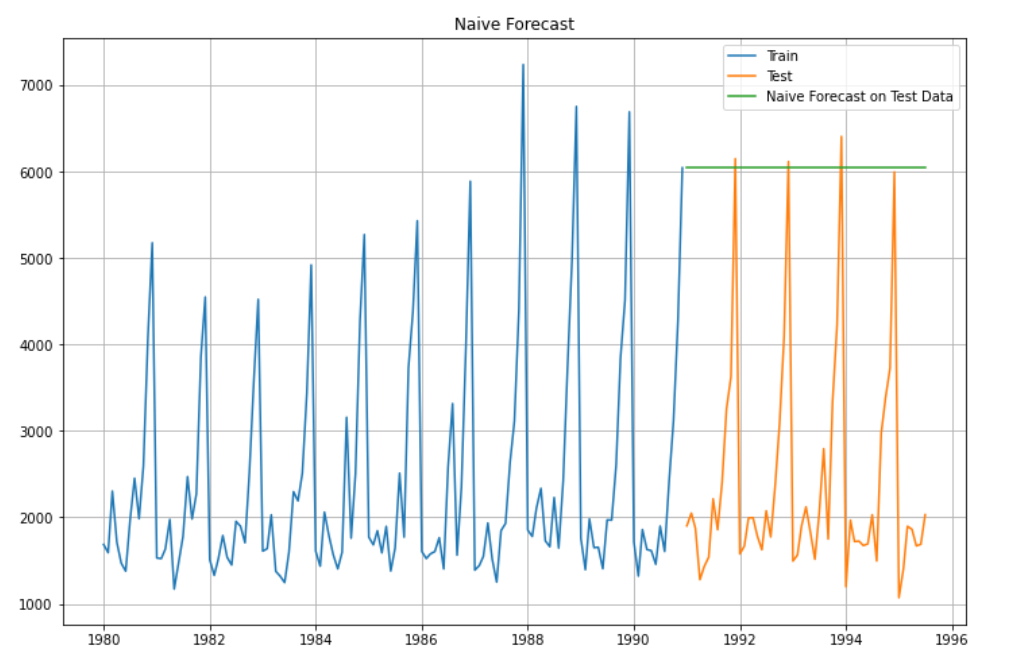


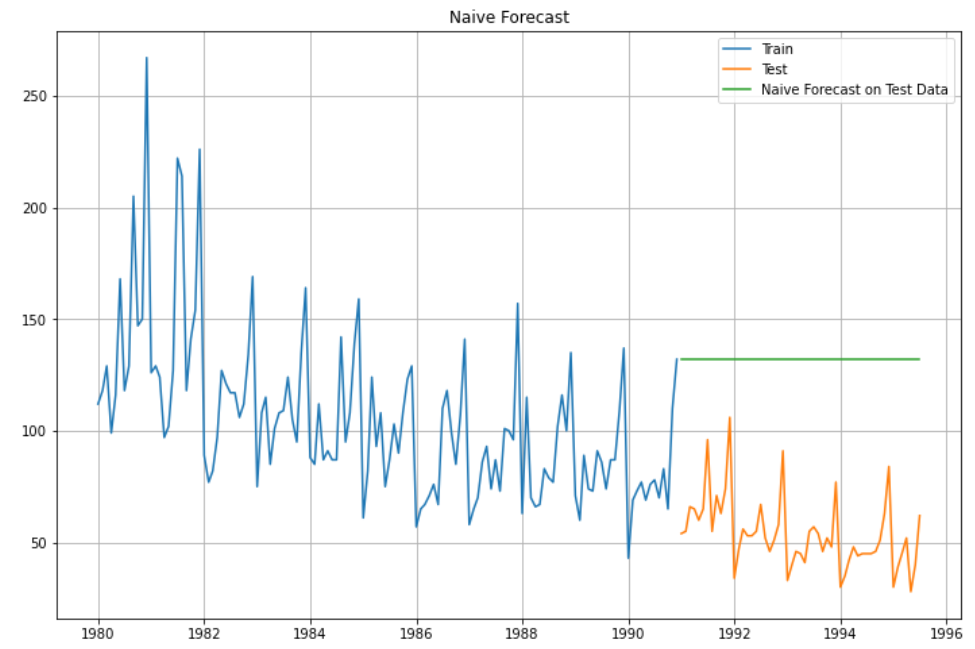
1. As we can observe the plots for both Sparkling and Rose Wine is that this model cannot be taken into consideration as it is kind not able to predict the data properly for the data sets.
2. Checking for the RMSE values we observe the values.



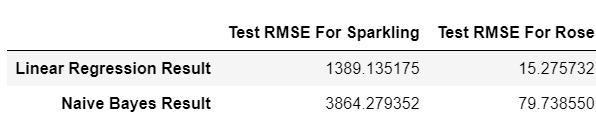
**Naïve Forecast:**

1. For the Naïve forecast the plots for the Sparkling and Rose Wine looked something like this





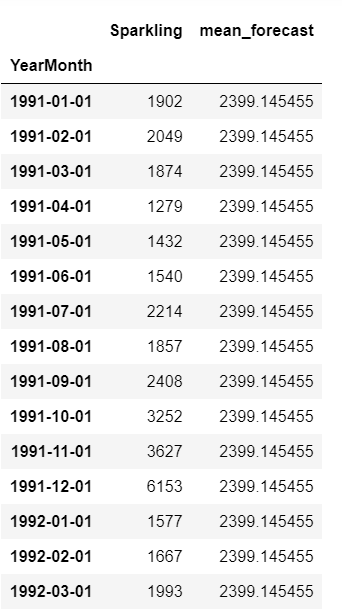
1. Now if we observe this plot the green line indicates the predicted values of the test data which does not look that much convincing and cannot be considered as a better modelling technique as the Forecast line is linear in nature which does not help us to predict anything.
2. The RMSE values for both wines were:



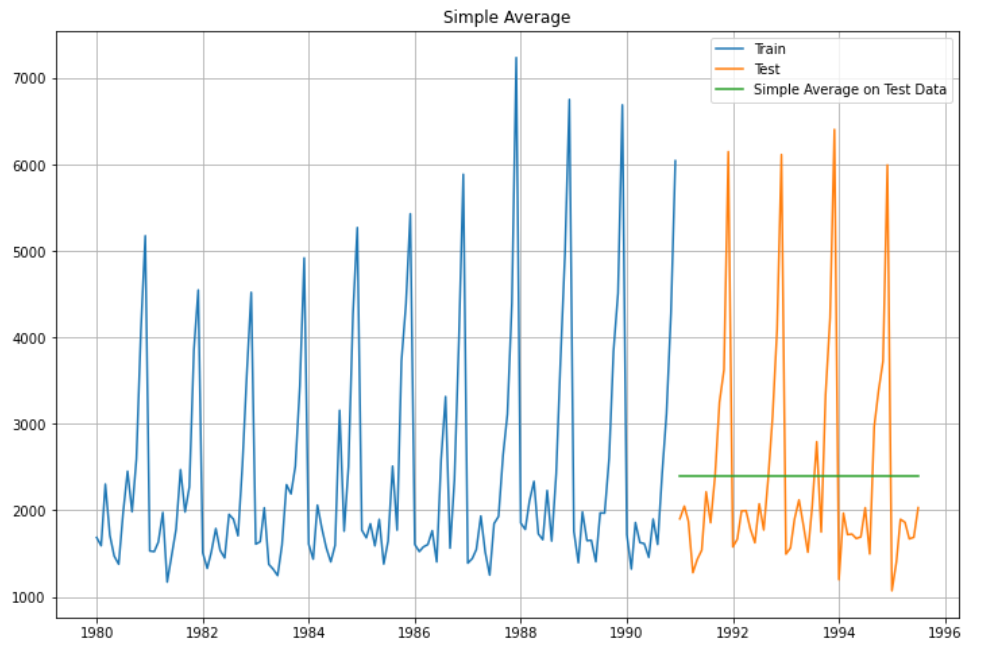
1. If we compare RMSE values of Naïve Forecast with the RMSE value of Linear Regression currently Linear Regression is performing better as compared to Naïve

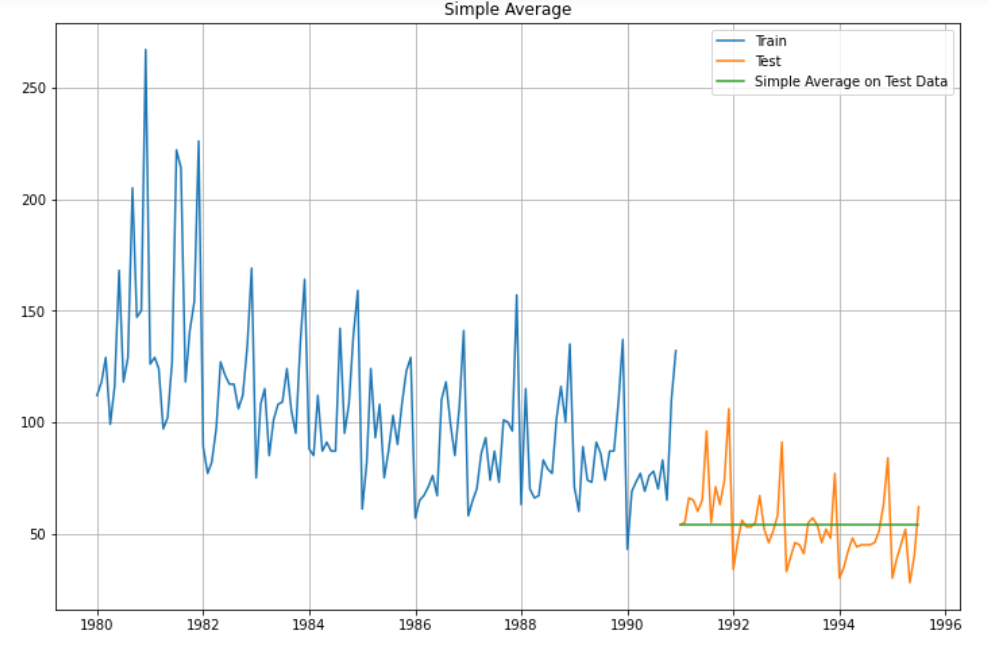
**Simple Average:**

1. As the name suggests in Simple Average, we simply take the average of the data set

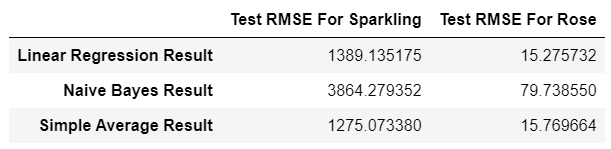


1. As we can see from the above data set, we take the average and create a new column named mean forecast and this mean forecast is for testing data set.
2. The same mean forecast is done for the Rose Wine as well and here also the mean forecast column is created.
3. Now, as we can incur from the mean forecast that as the values are same the line created will also be linear in nature.
4. The plot of Sparkling and Rose Wine would be able to say a better story.





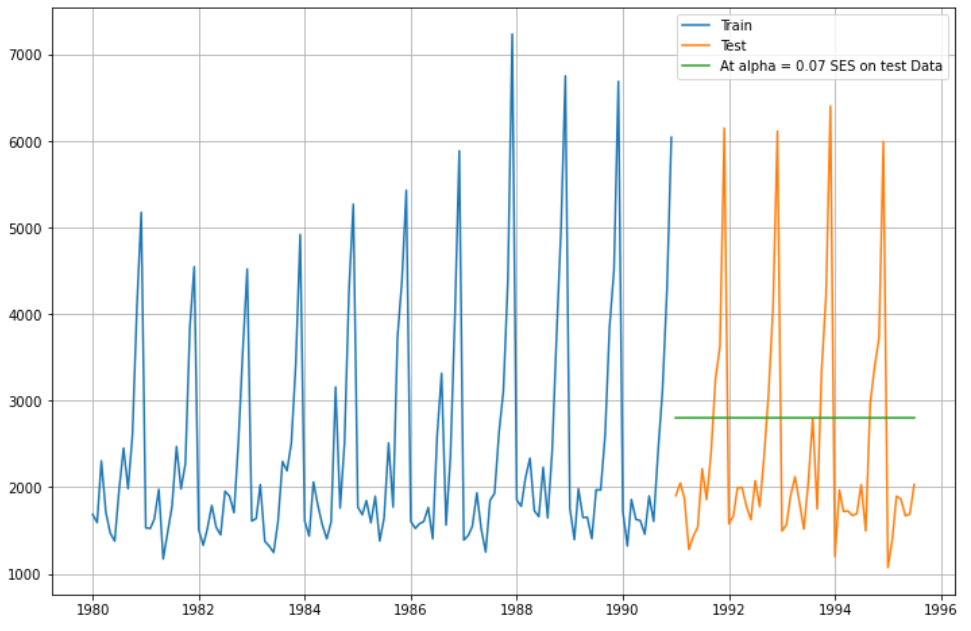
1. The RMSE value for the Simple Average is as follows:

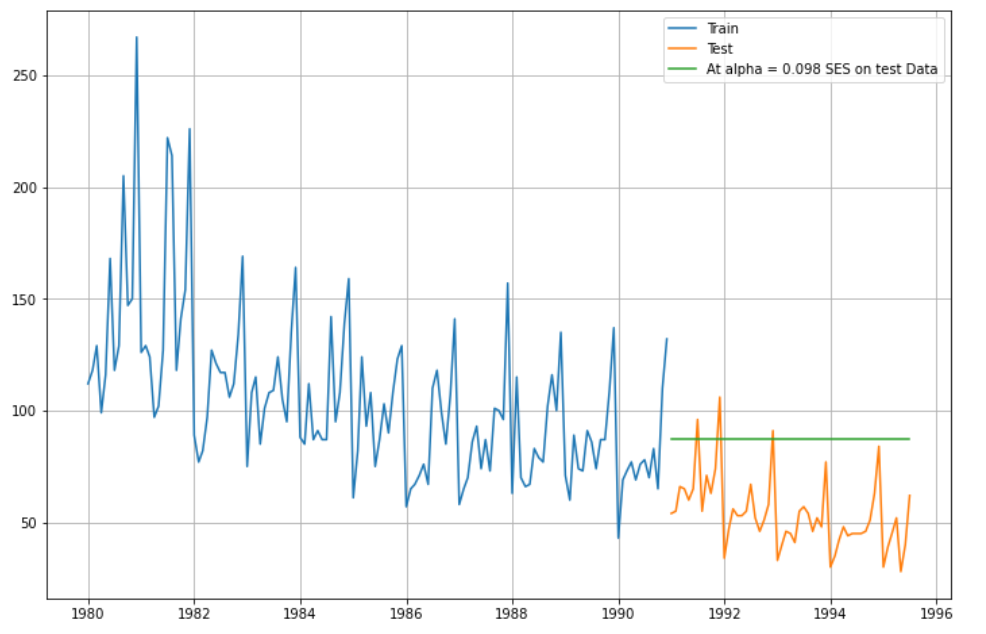


1. Now as we can incur from these values the Simple Average Model currently has the least RMSE value for the Sparkling wine and Linear Regression has the least RMSE value for the Rose Wine.

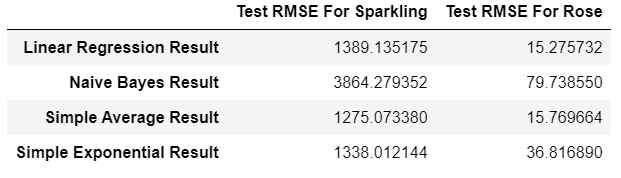
**Simple Exponential:**

1. Now, we have the simple exponential model which has the values for alpha, beta and gamma.
2. In this case we have only alpha value or smoothening level value.
3. The values forecasted for the Simple Exponential will be same as the alpha value is only for smoothening the plot.
4. Once we calculate the forecasted values the plot is built for the Sparkling and Rose Wine.
5. The alpha value for Sparkling is 0.07 and for Rose is 0.098





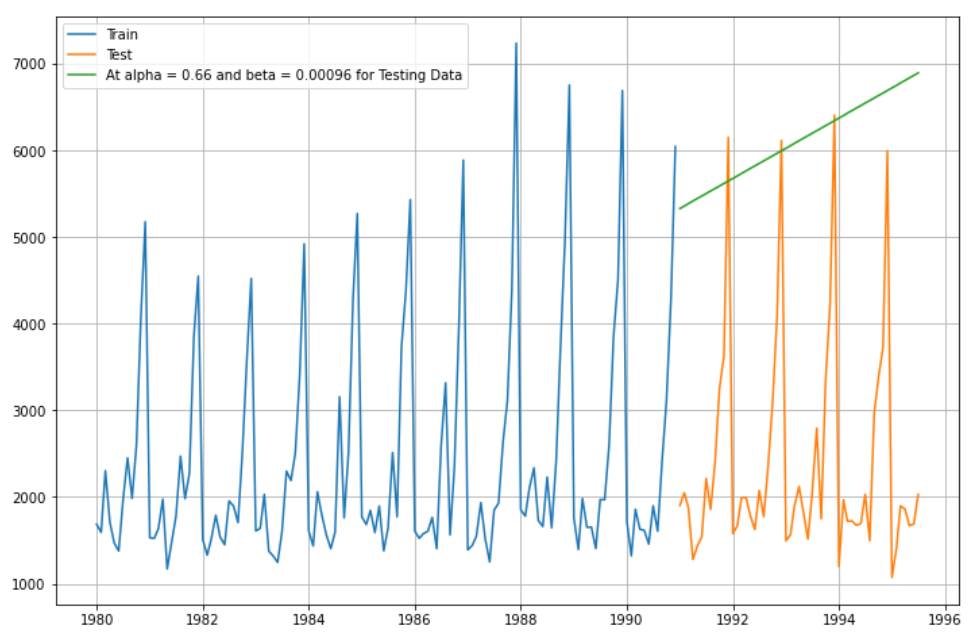
1. The plots also forecast a linear line which does not explain anything about the data predicted.
2. The RMSE values for these are:

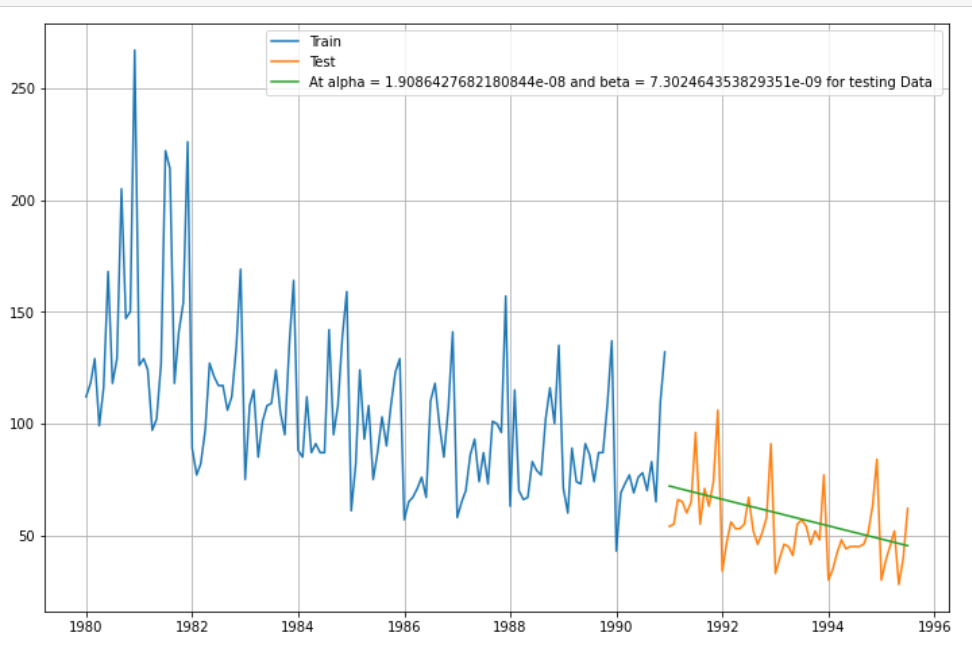


1. Simple Exponential RMSE values are better than Naïve but not that great when we compare it with other RMSE values.

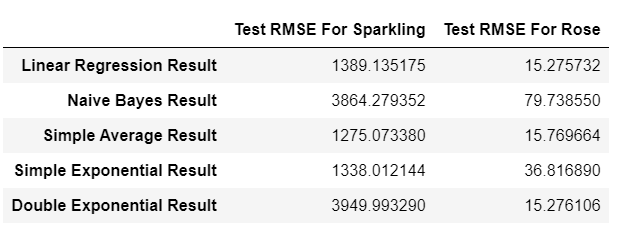
**Double Exponential Smoothening:**

1. Now, we have the double exponential model which has the values for alpha, beta and gamma.
2. In this case we have both alpha and beta value or smoothening level value and smoothening trend.
3. For Sparkling the alpha and beta values are 0.66 and 0.00096 and for the Rose the alpha and beta values are 1.9086427682180844e-08 and 7.302464353829351e-09.
4. The plots for Sparkling and Rose Wines are as follows:





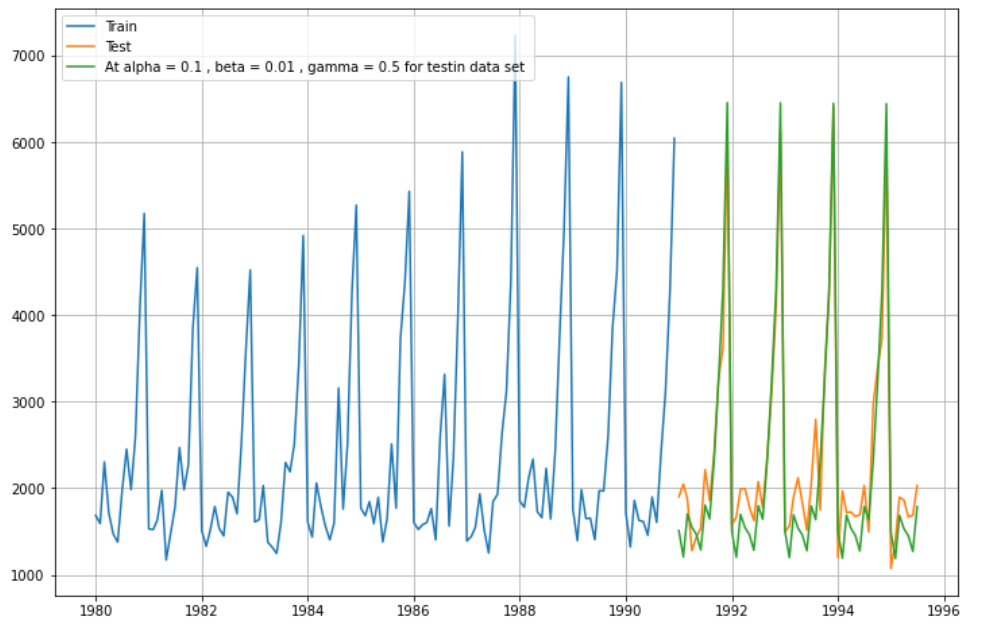
1. Now as we can observe from the plots due to the addition of the beta values the graph is trying to explain the trend of the predicted data.
2. The RMSE values for this are:

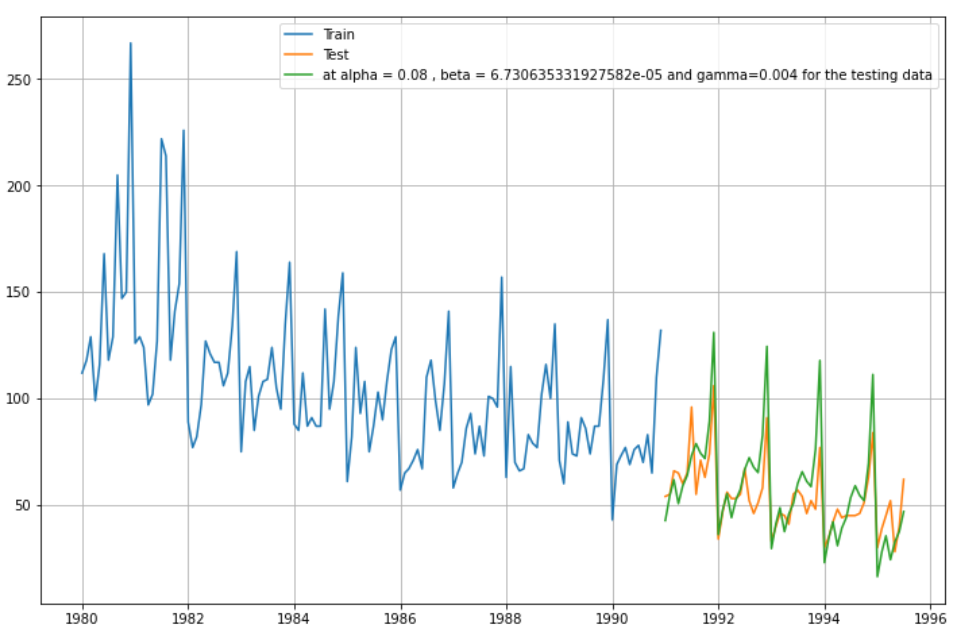


1. Double Exponential shows the trend of the data but the RMSE value for Sparkling lies on the higher side while for Rose the value is on the lower side.

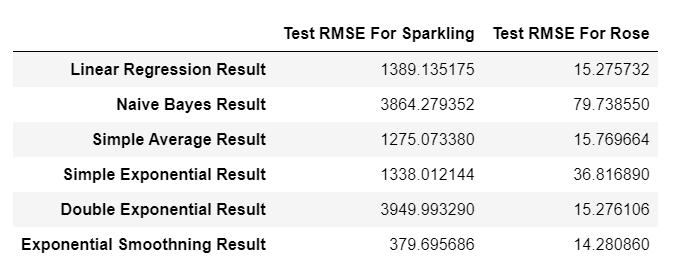
**Exponential Smoothening:**

1. Now, we have the exponential smoothening model which has the values for alpha, beta and gamma.
2. In this case we have all the parameters alpha, beta and gamma value or smoothening level value and smoothening trend and smoothening seasonality.
3. The values for alpha, beta and gamma calculated for Sparkling is 0.1, 0.01 and 0.5 and for the Rose is 0.08, 6.730635331927582e-05 and 0.04
4. Now when we calculate the forecasting data for the testing data set the plots for Sparkling and Rose looks something like this:





1. Now, as we can observe with the addition of new parameter the seasonality the prediction of the testing data is looking kind of on track as it is kind of able to replicate the testing data plot which is better as compared to the previous models we have checked until now.
2. The RMSE values for these are as follows:



1. The Exponential Smoothening shows the least RMSE values for both Sparkling and Rose Wine indicating that currently the exponential smoothening model is performing better predictions

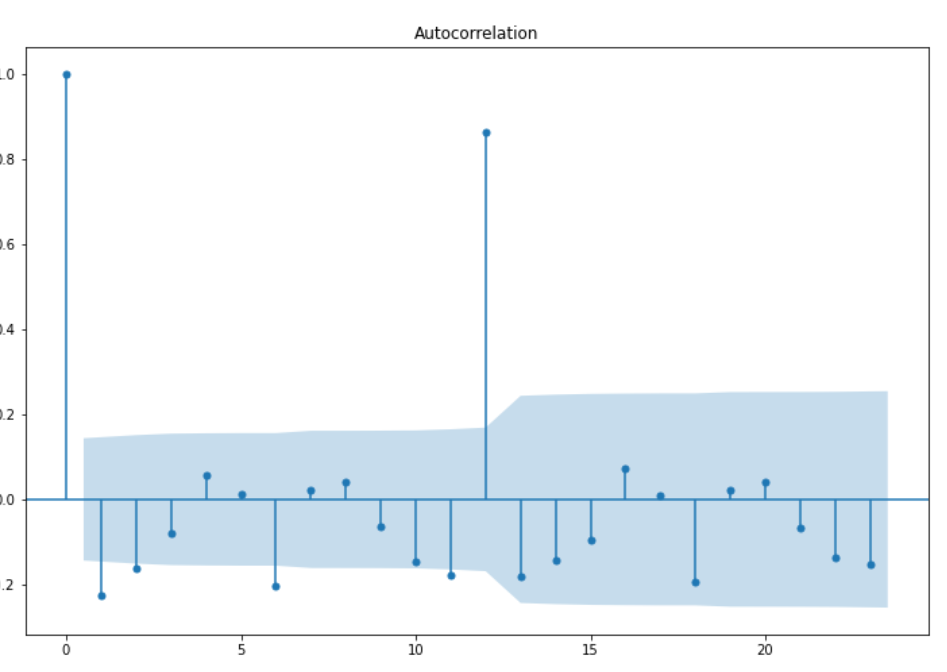
**Q5: Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment.  
Note: Stationarity should be checked at alpha = 0.05.**

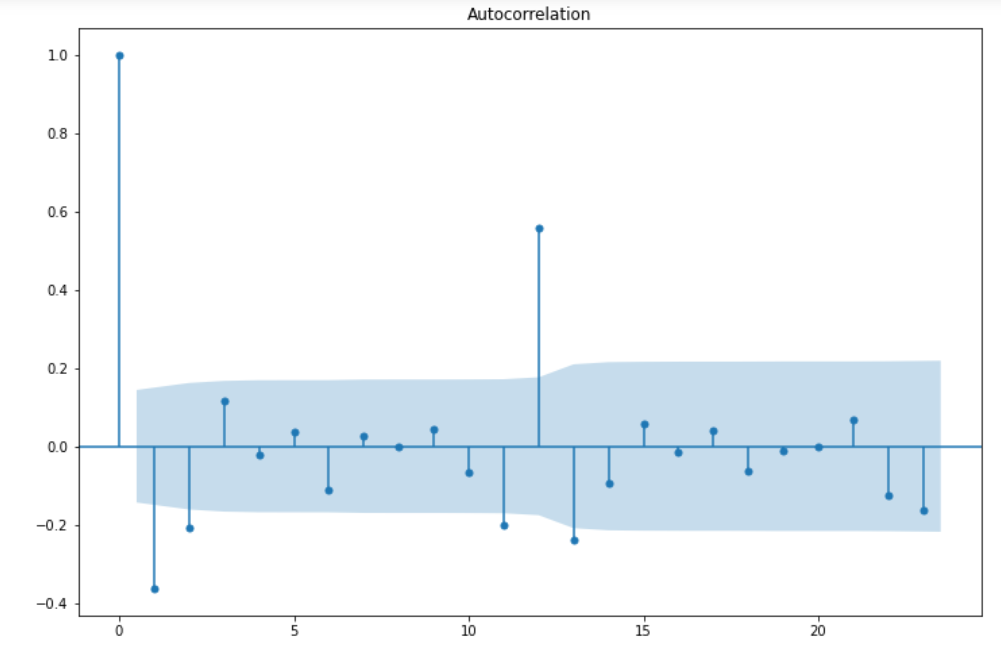
Ans: When we talk about the hypothesis testing, we have our Null Hypothesis and Alternative Hypothesis as:

H0 [Null Hypothesis]: Time Series data is not stationary

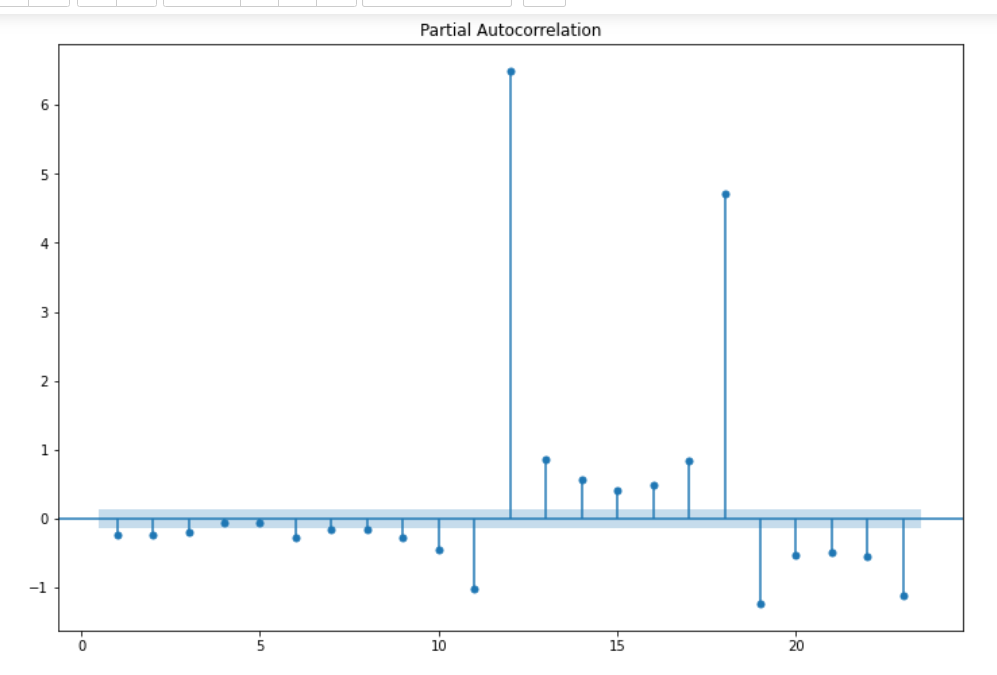
H1 [Alternate Hypothesis]: Time Series data is stationary

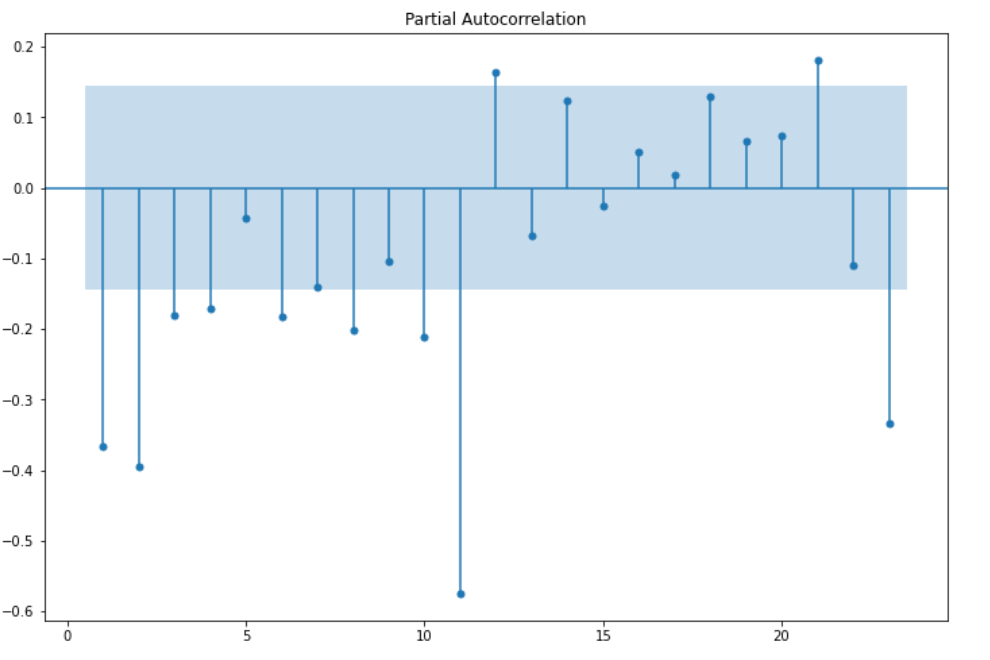
1. As we can see from the plots that the data is not stationary and when we check for the p- values for both Sparkling and Rose wine data set we get the values of 0.70 and 0.46 respectively.
2. Now, if the p-value > 0.05 then we fail to reject the null hypothesis and if the p value < 0.05 we reject the null hypothesis
3. As we can incur from the p values above that both the values for Sparkling and Rose is greater than 0.05 which means that we fail to reject null hypothesis indicating that the data is not stationary.
4. To make the data stationary we have to do the difference between the values like second data point minus first data point and third data point minus the first data point like this.
5. Once we do that for both our data set, we get the updated p values for sparkling and rose wine as 0 and 3.028272263688236e-11 respectively.
6. Now if we observe these values, they are < 0.05 which means that we reject the Null hypothesis which means that the data is stationary now.
7. If we have to check for the stationarity of data set at alpha= 0.05 then we can take the help of auto correlation function plots and partial auto correlation function plots.
8. The acf plot for Sparkling and Rose Wine are:





1. The pacf plots for the Sparkling and Rose Wine are:

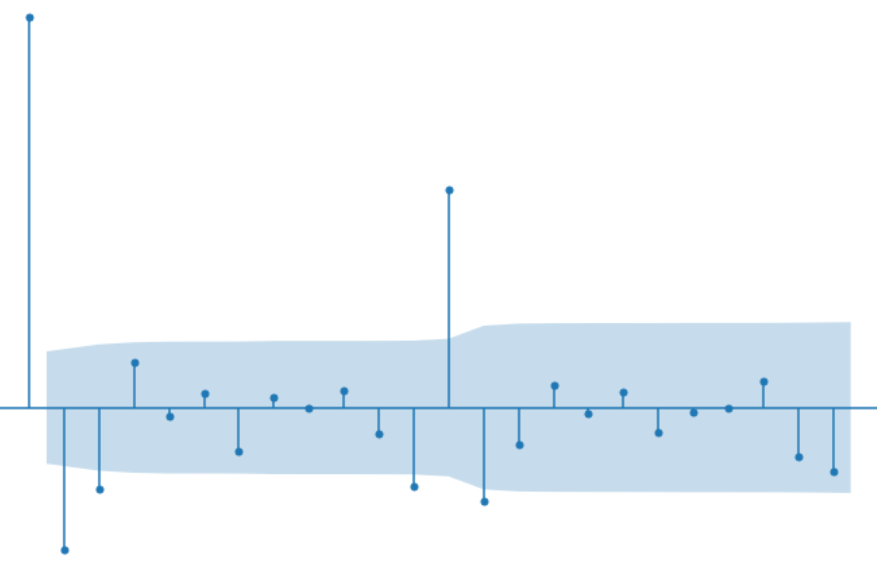




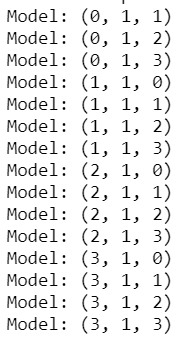
**Q6: Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.**

Ans: **ARIMA Model:**

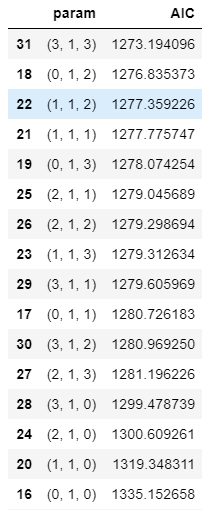
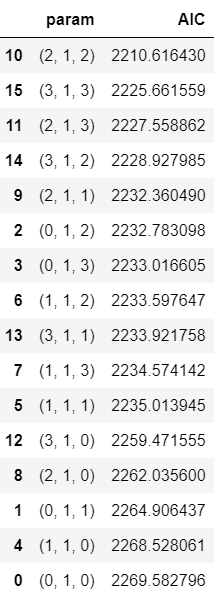
1. ARIMA or Auto Regressive Moving Average is used with the multiple parameters so that we can achieve the lowest AIC value.
2. Lowest AIC means that score is considered to be the best combinations for the variables p, q and d.
3. The variable p is for the Auto regressive function, q is for the moving average and d is the difference.
4. If the data is not stationary difference is used for the making the data stationary.
5. The values for p and q are taken by observing the PACF plot and ACF plot respectively.
6. The cut off values are taken from the pacf and acf plot.



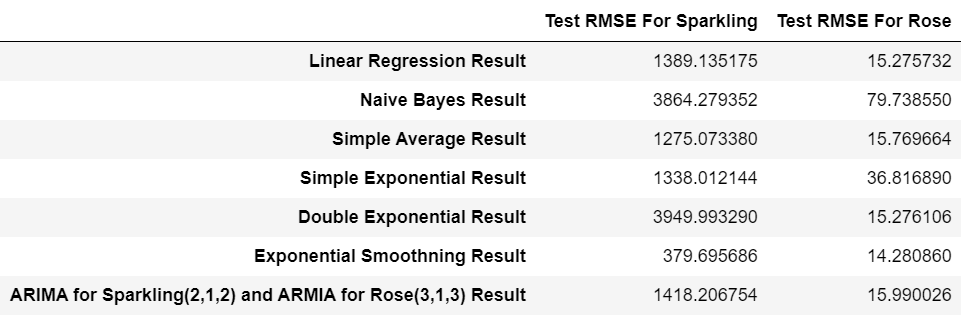
1. As we can check the above graph is an acf plot the first line is outside the blue region is not considered for the cut off so the next line which starts from the blue region from there we check how many lines are inside the blue region. As soon as the any line comes out of the blue region the number of lines before that will be considered as cut off number and that will be given to p and q.
2. After getting the values for p, d and q we get the combinations which are something like this:



1. After calculating the AIC values for all these combinations, we get a data table now as we can observe from below table that the AIC value is least for the combination (2,1,2). So, we will be going ahead with this combination for the Sparkling Wine and for Rose Wien we will be taking the combination of (3,1,3)



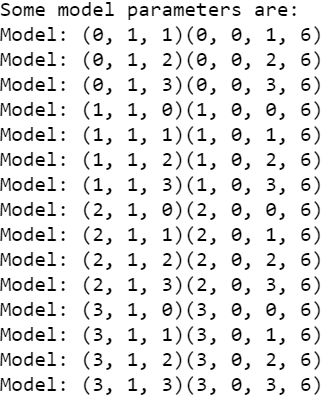
1. Now, we will be taking up these combinations for Sparkling and Rose wine and forecast the data and check for the RMSE values for it.
2. The RMSE values for these are as follows:



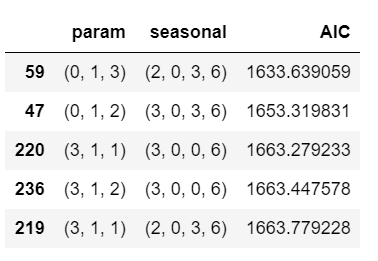
1. For a better understanding we have taken all the previous model results to compare if ARIMA model has performed better as compared to other models or not.
2. As we can observe that it has not performed that great and cannot be considered a good model to predict the testing data.

**SARIMA Model:**

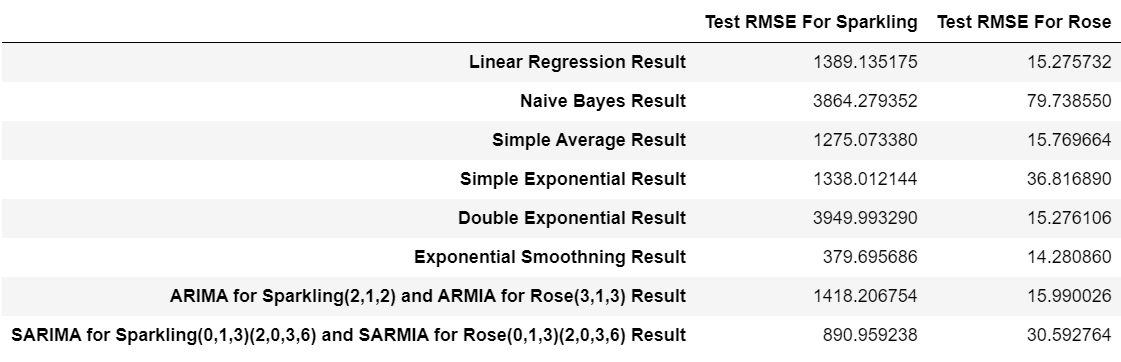
1. Now, we check the model performance with the SARIMA parameters combinations.
2. The SARIMA is same as ARIMA which means that we have the parameters p and q but, in this case, we have D and not d.
3. This D is also for the difference but in SARIMA. In this the p, D and q are combined with the combination of (p, D, q) and (p, D, q, S).
4. Where S stands for seasonality. This number S can be checked from the pacf and acf plots.
5. The pattern has to be observed in these plots. Meaning, that we have check from what point does the graph is showing kind of similar pattern.
6. Once that is identified which in our case is 6.
7. The combinations of the parameters look something like this:



1. Now, we have to check the same thing the lowest AIC value and the AIC values looks something like this for Sparkling and Rose Wine is:



1. The lowest AIC value in which we get for the combination of (0,1,3) (2,0,3,6) for Sparkling and (0,1,3) (2,0,3,6) for Rose.
2. Now let’s check for its RMSE value and compare it with RMSE values of previously built model:

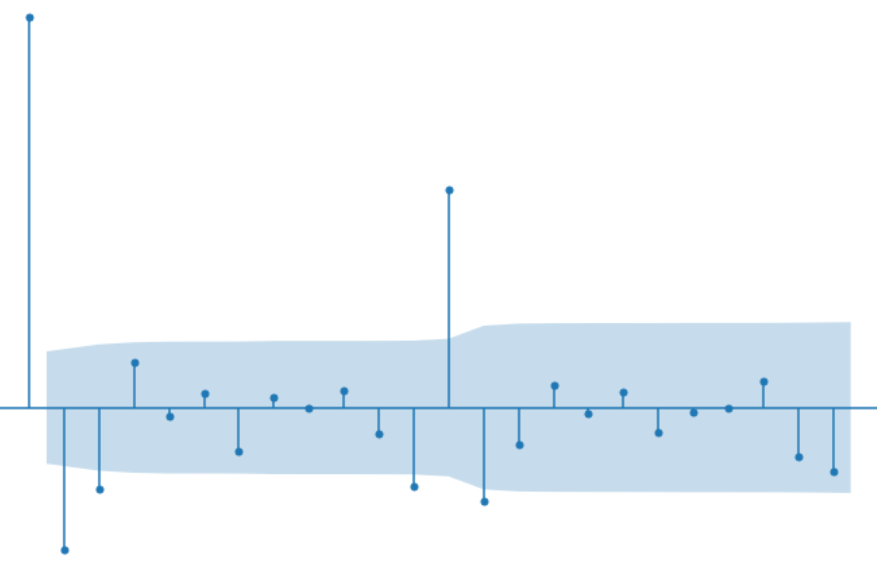


1. These values for SARIMA are not that much great as compared to other models.

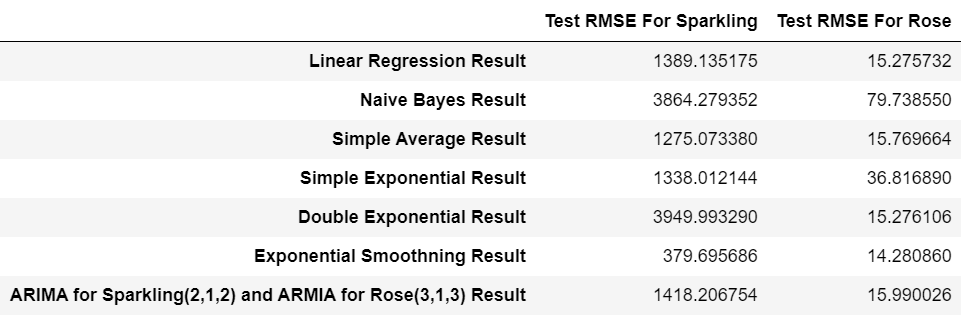
**Q7: Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.**

Ans: **ARIMA Model:**

1. ARIMA or Auto Regressive Moving Average is used with the multiple parameters so that we can achieve the lowest AIC value.
2. Lowest AIC means that score is considered to be the best combinations for the variables p, q and d.
3. The variable p is for the Auto regressive function, q is for the moving average and d is the difference.
4. If the data is not stationary difference is used for the making the data stationary.
5. The values for p and q are taken by observing the PACF plot and ACF plot respectively.
6. The cut off values are taken from the pacf and acf plot.

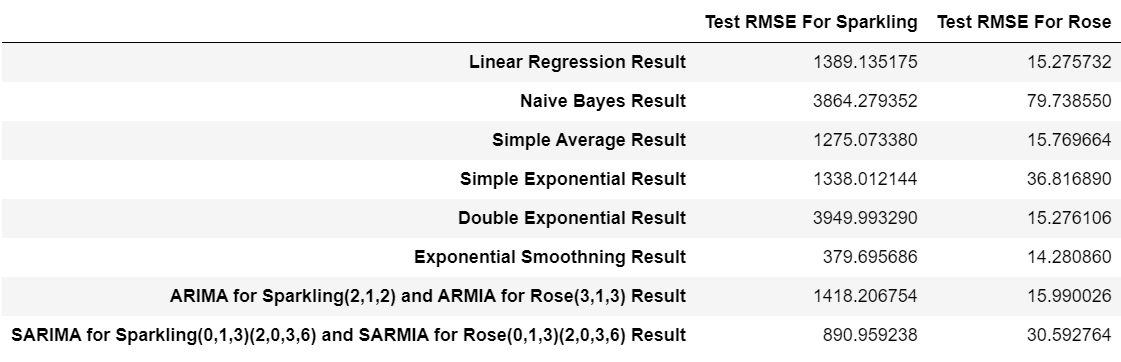


1. As we can check the above graph is an acf plot the first line is outside the blue region is not considered for the cut off so the next line which starts from the blue region from there we check how many lines are inside the blue region. As soon as the any line comes out of the blue region the number of lines before that will be considered as cut off number and that will be given to p and q.
2. The RMSE values for these are as follows:



**SARIMA Model:**

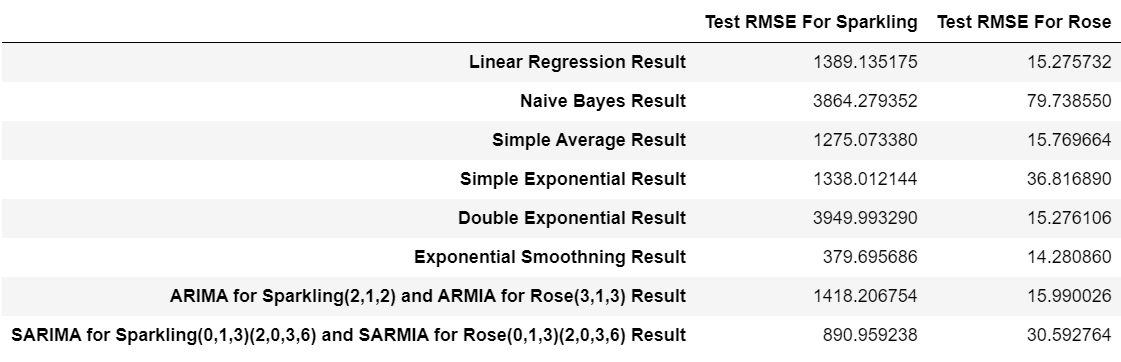
1. Now, we check the model performance with the SARIMA parameters combinations.
2. The SARIMA is same as ARIMA which means that we have the parameters p and q but, in this case, we have D and not d.
3. This D is also for the difference but in SARIMA. In this the p, D and q are combined with the combination of (p, D, q) and (p, D, q, S).
4. Where S stands for seasonality. This number S can be checked from the pacf and acf plots.
5. The pattern has to be observed in these plots. Meaning, that we have check from what point does the graph is showing kind of similar pattern.
6. Once that is identified which in our case is 6.
7. Now let’s check for its RMSE value and compare it with RMSE values of previously built model:



1. These values for SARIMA are not that much great as compared to other models.

**Q8: 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.**

Ans: After building all the models and calculating the RMSE values for Sparkling and Rose wine we get the table for the RMSE values which looks something like this:

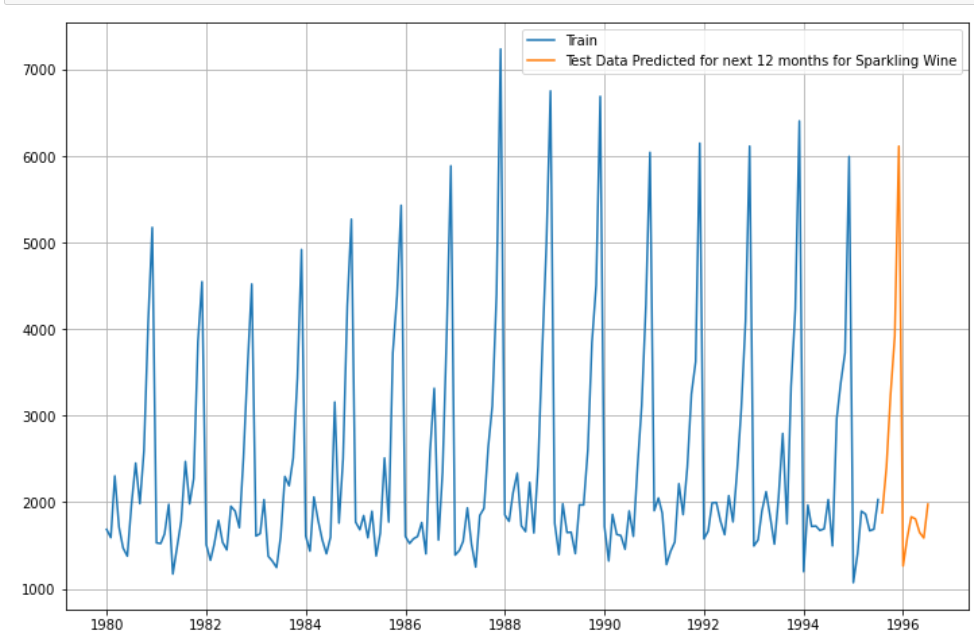


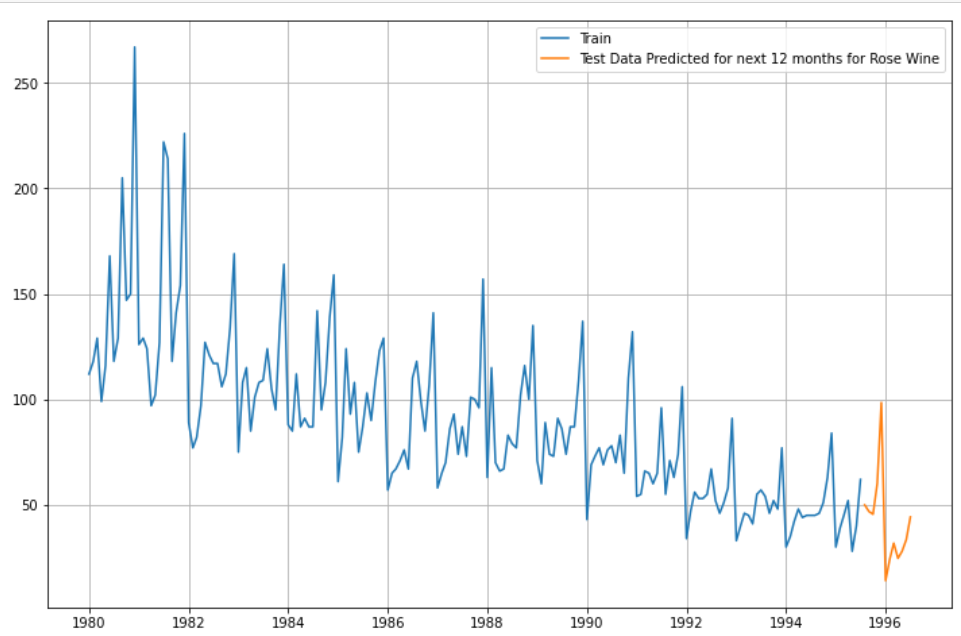
As we can observe the least RMSE value which we get is from the Exponential Smoothening Result that is for the rmse value of sparkling wine is 379.69 and for rose wine 14.28.

**Q9: 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.**

Ans:

1. Now, to predict the next 12 months of the data the models which we have built before we can conclude that the model which has lowest RMSE value should be considered which in our case is the Exponential Smoothening model.
2. We will be taking the next 12 months data into testing data so that the predictions can be checked.
3. The graphs for the sparkling and rose wine are respectively.





1. As we can see the predictions have come out quite well.

Q10: **Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.**

Ans: Comments from the model building:

1. After building different models we can conclude that the best model which can be used for forecasting is exponential smoothening.
2. There are two wines which are sparkling and rose respectively.
3. Now if we talk about the sparkling wine the sales of the wine show a kind of similar pattern which indicates that the sparkling wine is most sold out during the times of October and December.
4. As these months are considered to be the kind of new year months.
5. So, the sparkling wine sales shows an increase in these months.
6. During the other months the company can launch some promotional offers such as a party pack offer in which the customer can get two wine bottles and with that some cash back as well.
7. This offer might sound outdated but the drinks like wine are mainly sold during the new years or any occasion etc.
8. So, people might not like to buy wine except if there is any occasion.
9. As the company is kind of popular, they can invest their amount in beer or other drinks which will help them decrease there loses during the times when there are less sales of wine.
10. As some drinks like beer is kind of everyone’s go to drink and people like it weather or not there is a occasion or just normal casual day.
11. So, the loss which we incur from wine business can be recovered from our beer sales.