**LOGISTIC REGRESSION**

**About the model**:

**Logistic regression**: logistic regression is a statistical analytic approach for predicting a binary outcome, such as yes or no. A logistic regression model analyzes the relationship between one or more existing independent variables to predict a dependent data variable.

**Dataset:** red wine quality data set

I have taken this data set from Kaggle this data set has observations 1599 and 12 columns  
<https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>

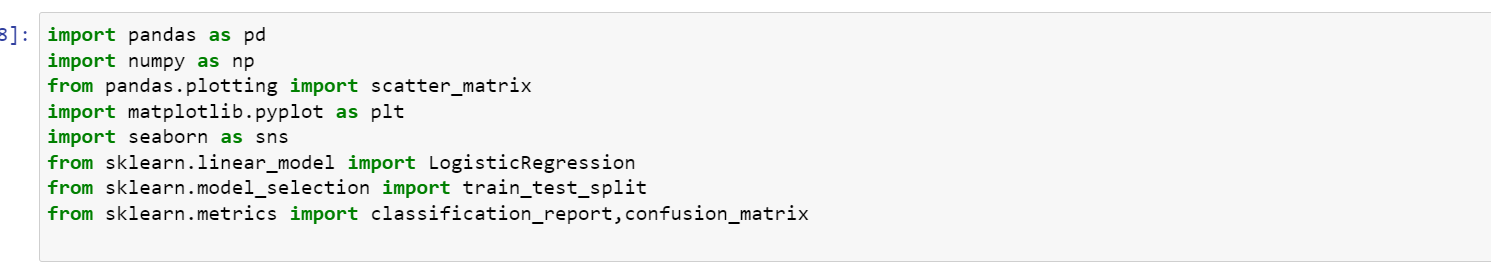
**Independent variable:-** : 'fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar', 'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density', 'pH', 'sulphates', 'alcohol' .

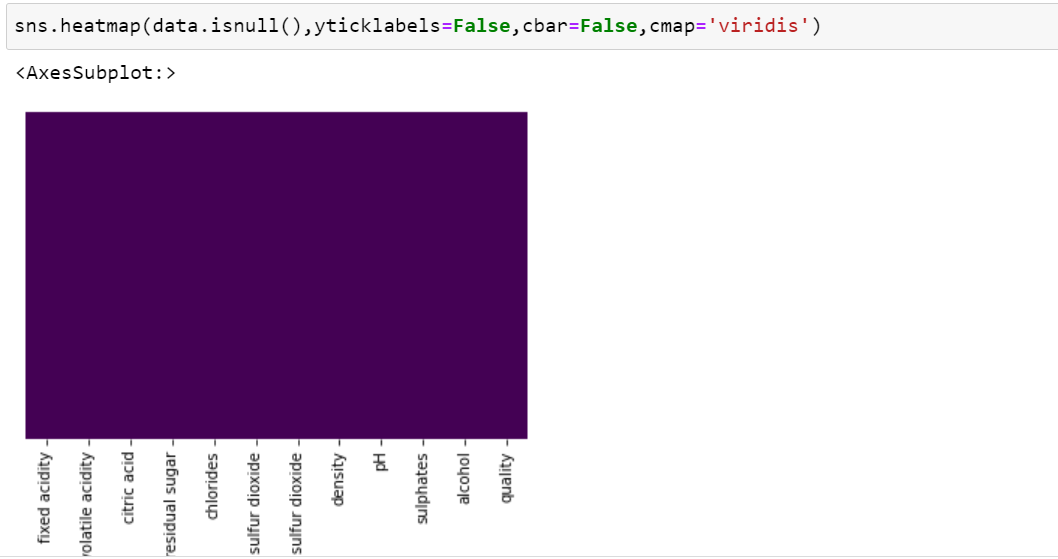
**Dependent variable:-** Quality.

**Data pre-processing**:- the data set had no missing values and they were normally distributed I have checked this by using isnull() method, and I have explored the data set by using .describe()method.

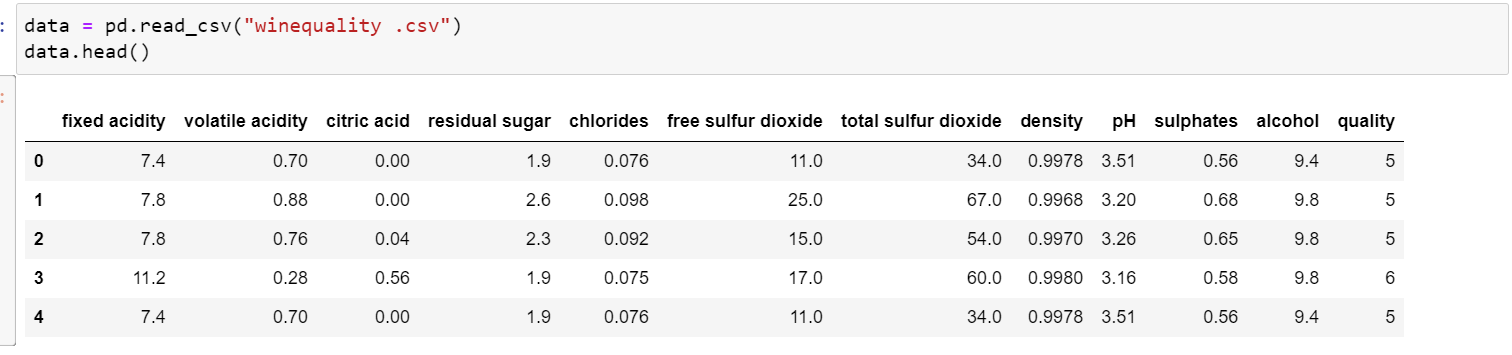
I have imported the file winequality into my jupyter notebook by using pandas to start my analysis and named the data frame as data.

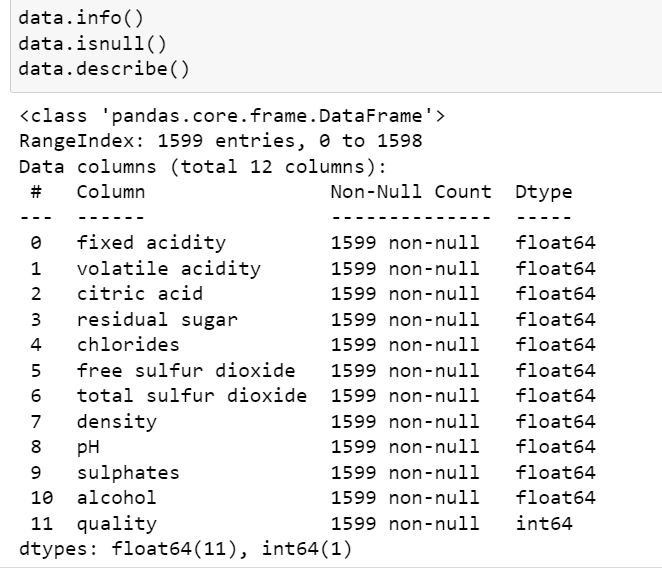
Also I have imported other libraries which Im going to use to create a logistic regression model.



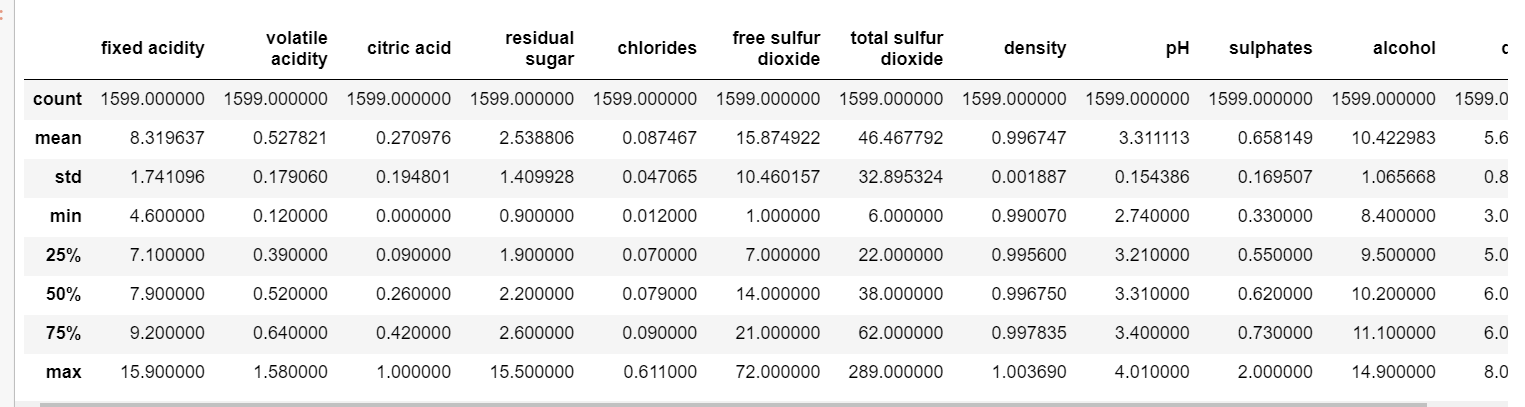


The above code shows there if there are any missing values in any column it will get marked on the above heat map, as we can see there are no spots in the heatmap hence there are no missing values.





We can observe that there are no missing values in all columns.



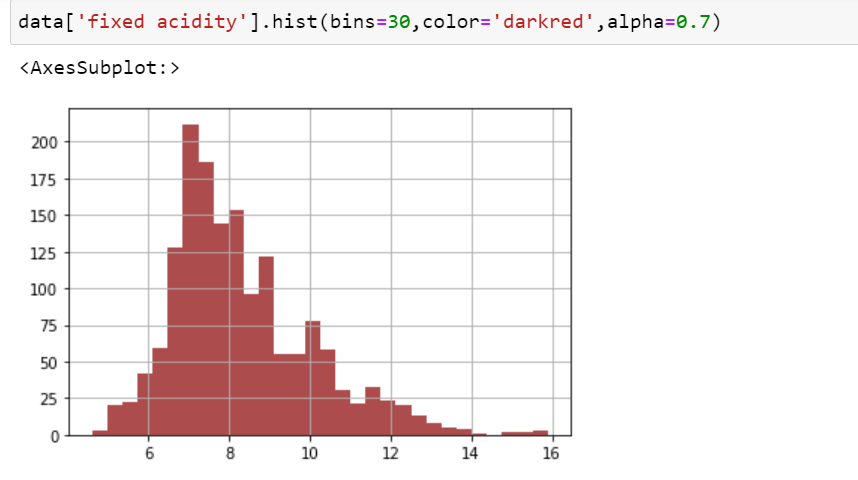
The above screenshot shows the result of .describe method which shows the mean std and other descriptive statistics related to the data set.

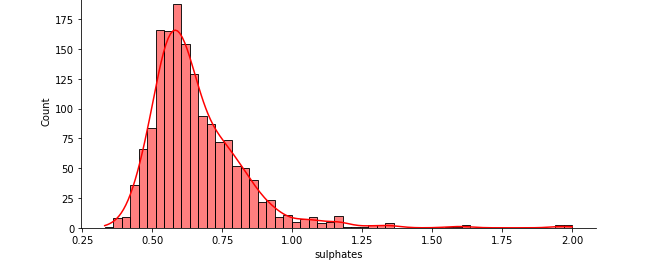
**Data visualizations:-**

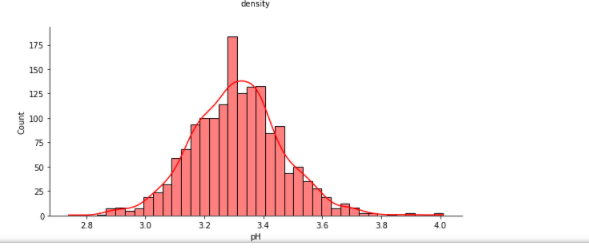
Before going to finally create the model I have used visualization to understand the data set more effectively.

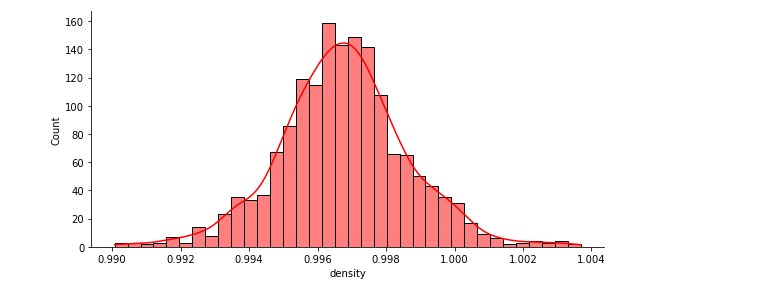


By looking at the above screenshot we can observe that there are more values of wine quality which is in between 5 and 6.

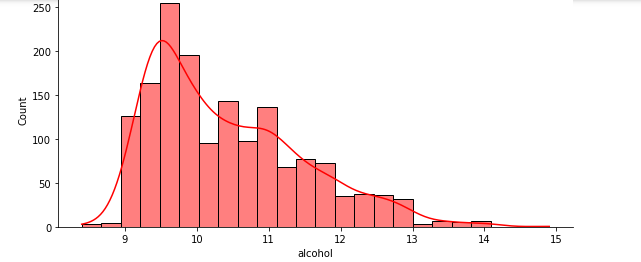


The above visualization is about fixed acidity. We can see that there are more than 225 observation which has value in between 7 and 9. It is even normally distributed. 



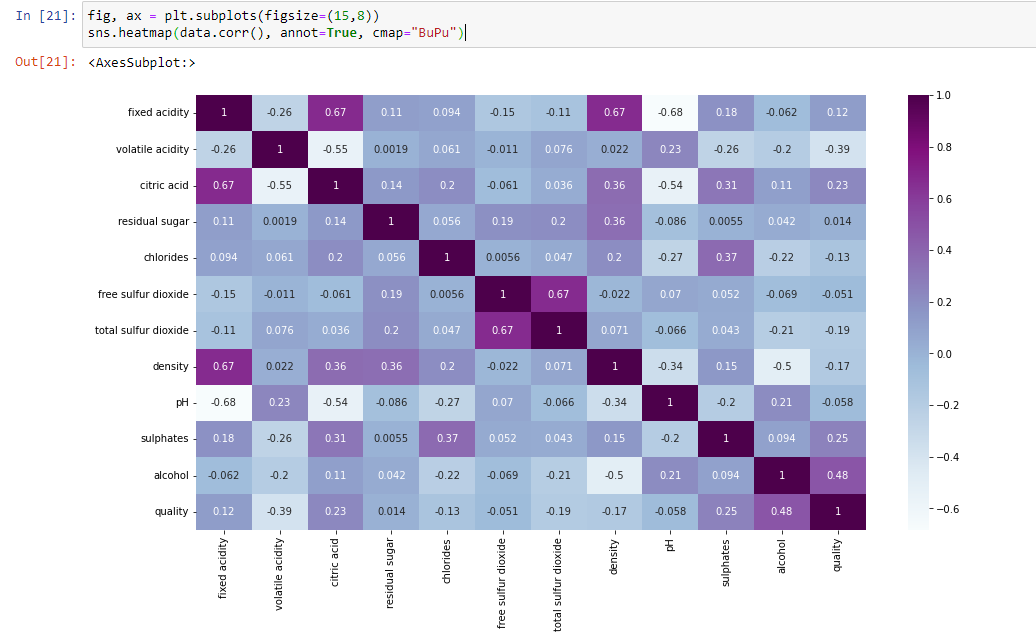


The above visualization shows the histogram of density independent variable. Its evident that this variable is normally distributed which is good for analysis.



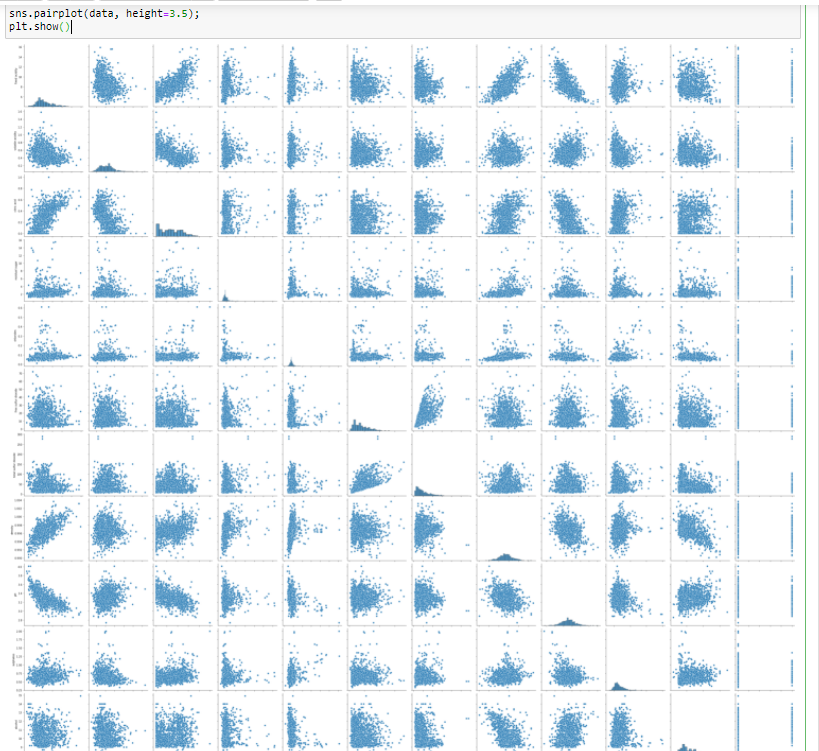
The above histogram shows the visualization of the alcohol variable we can observe that most of it values lies in between9 and 10.

I have drawn visualization for every independent variable which can be seen in ipynb file which I have submitted .



The above heat map is the correlation matrix for all the variables present in the data set.

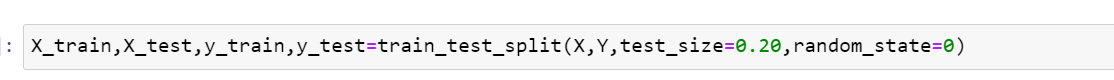
The below image shows the pair plot which I have drawn between every variable to check their relationship with each other and how they are related to each other, we can see that most of them are closely spread which means there are very less outliers. Hence I can proceed with next steps.



Splitting the data:-

I have split the model into 80% for training and 20% for testing as instructed in the assignment description

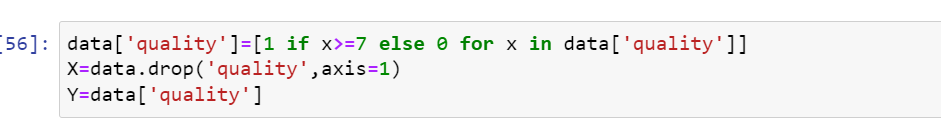
Here is the code where I have done that



I have set test\_size to 0.20. and random\_state to 0.

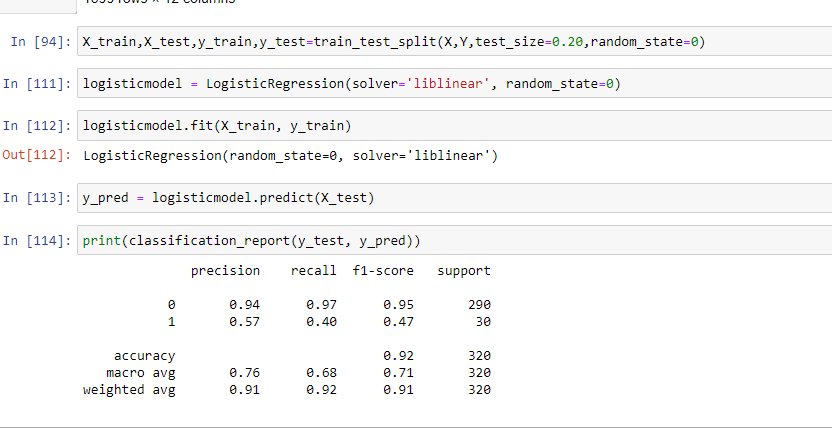
Wine quality is graded on a scale of 2 to 8. However, because we're going to create a model that predicts wine quality in two ranges (high quality, low quality). The binary values of these categories are transformed (0 and 1s).

Values greater than or equal to seven are converted to one, while all other values are turned to zero.

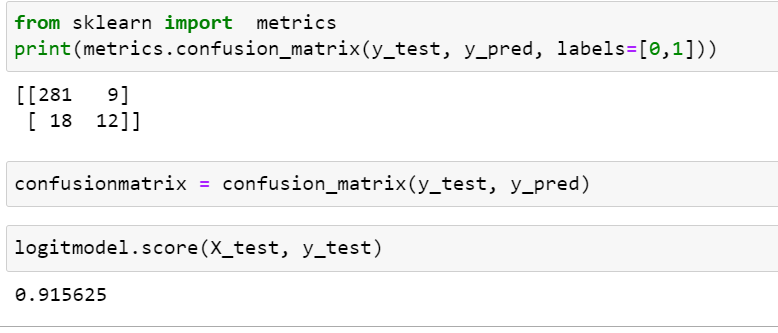


**Applying the model:-**

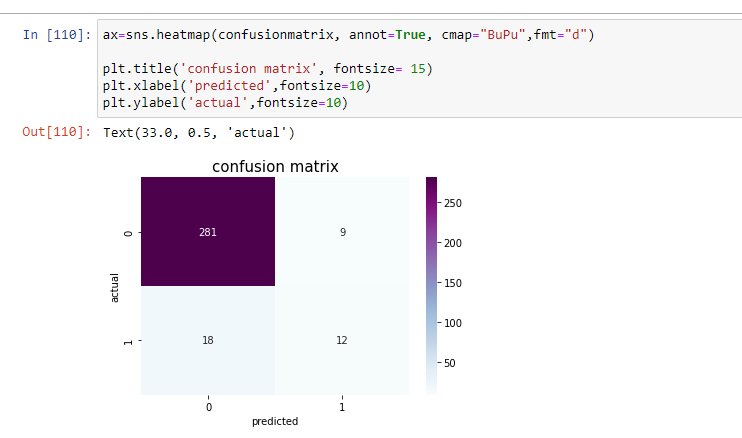
I have used the below model to train the data set .



The below image shows the confusion matrix value of the model predicted by using logistic regression.



The above image shows the accuracy of the model which is 91%



The image shows the confusion matrix of the model that is predicted.

**Conclusion**:- by the above accuracy we can see that the model has a accuracy pf **91**% which is very good model and we can assume that the model predicts accurate results with accuracy of **91**% of quality of wine.