**Name: Vyshnavi Gajjela**

**Student Id:11552425**

**Assignment-2**

**Logistic Regression Report**

Logistic regression is a supervised learning. It is mainly used for classification, used when the dependent variable is categorical. The dependent variable is binary either 0 or 1.

**Types of Logistic Regression:**

**Binary Logistic Regression:** Here the categorial response has only two possible outcomes.

**Multinomial Logistic Regression:** It has three or more classes for choosing.

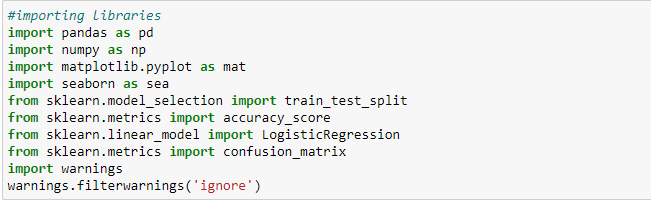
**Code:**

**Algorithm used:** Logistic Regression

The dataset winequality-red is given to us.It contains all the features of red wine.

**Libraries:** Importing all the required libraries

1. Pandas is used for data analysis.
2. Numpy is used for supporting array objects
3. Matplotlib is used for visualizing the plots for viewing
4. Sklearn is used to implement machine learning functions
5. Seaborn is used for data visualization
6. Logistic Regression is used for developing the model
7. train\_test\_split is used to split the dataset for testing the data and train data.
8. Confusion matrix is used as performance measuring.

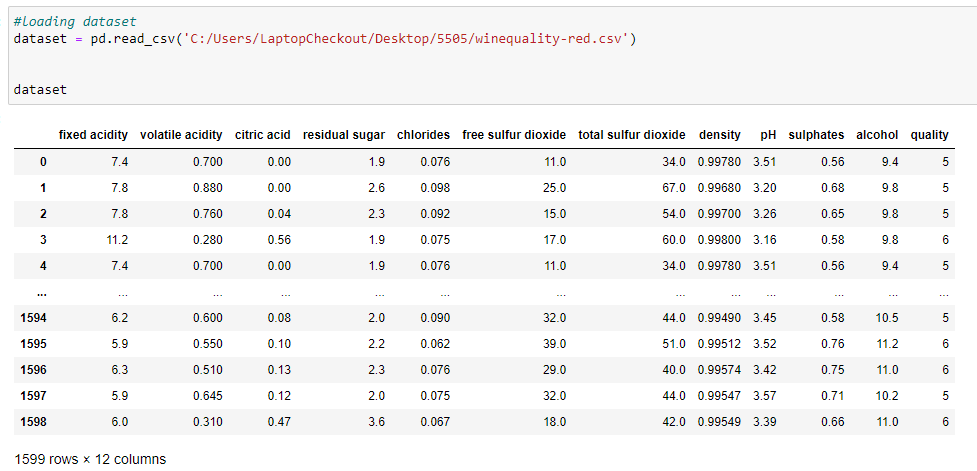


**Data Description:**

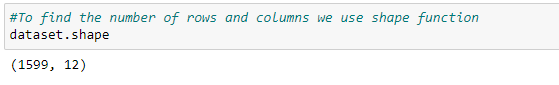
* For this assignment we used winequality-red dataset. It has the below columns:

1. fixed acidity
2. volatile acidity
3. citric acid
4. residual sugar
5. chlorides
6. free sulfur dioxide
7. total sulfur dioxide
8. density
9. pH
10. sulphates
11. alcohol

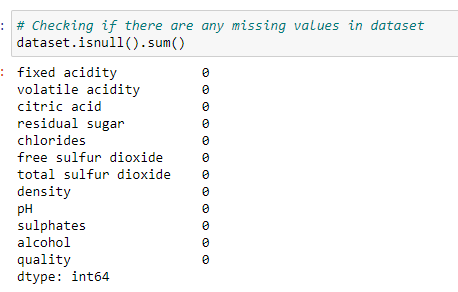
* The below is for loading dataset to pandas data frames.



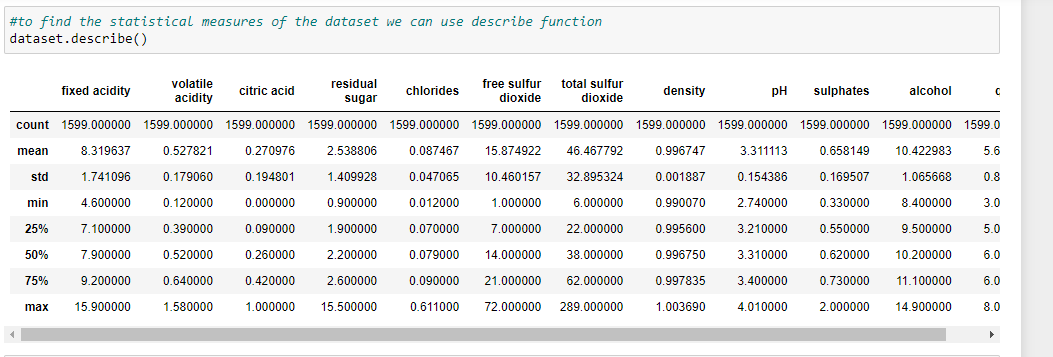
* **Shape** : This shape function is used to find the number of rows and columns of the dataset



* If there are any missing values in the dataset, can be checked using the below code.

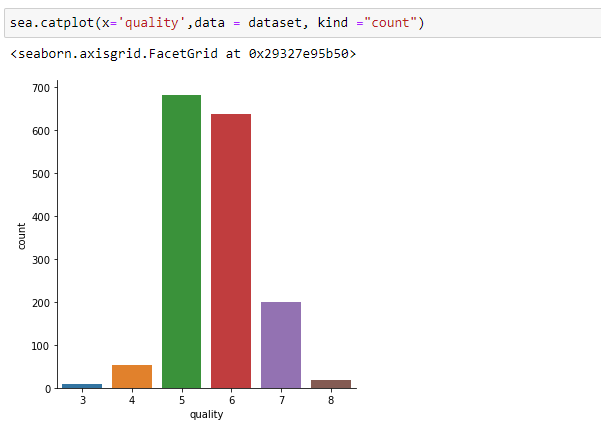


* describe : The describe function is used find the statistical measures.



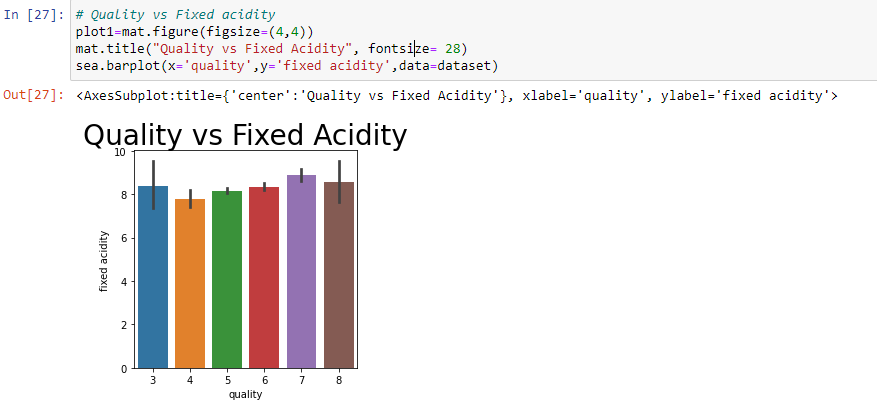
**Data Visualization:**

* We classify the wine based on the quality, if the quality of wine is above 6 then it is treated as good wine and if it is below it is not a good quality wine.

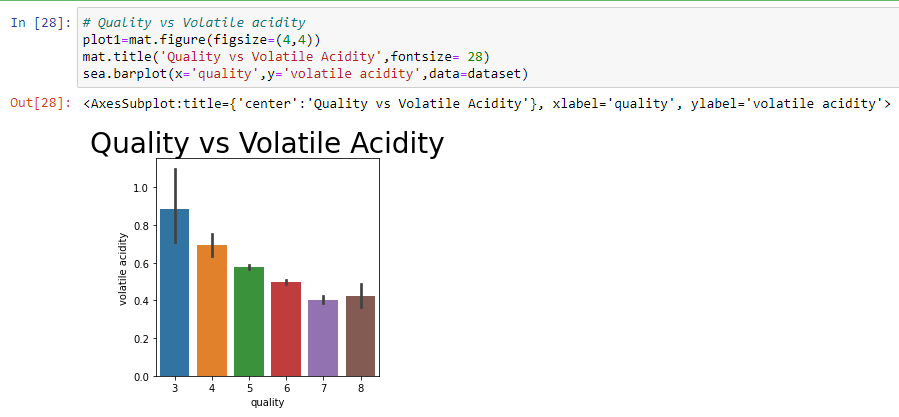


* The below are the visualizations using seaborn, matplotlib showing quality vs 11 predicator variables

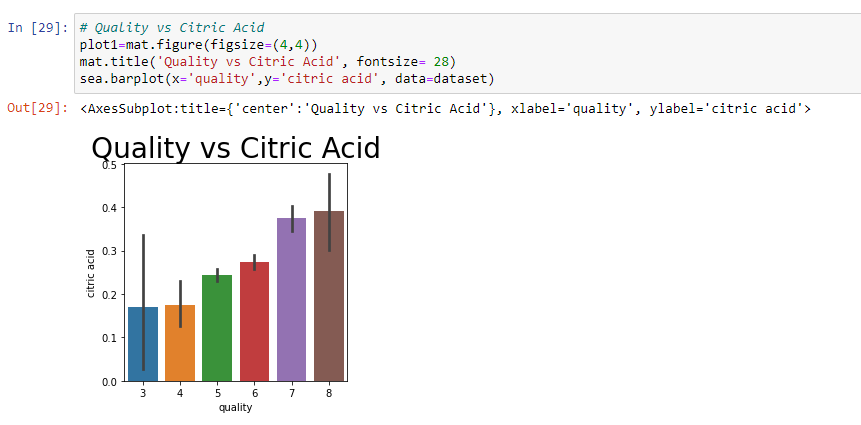
1. Quality vs Fixed Acidity



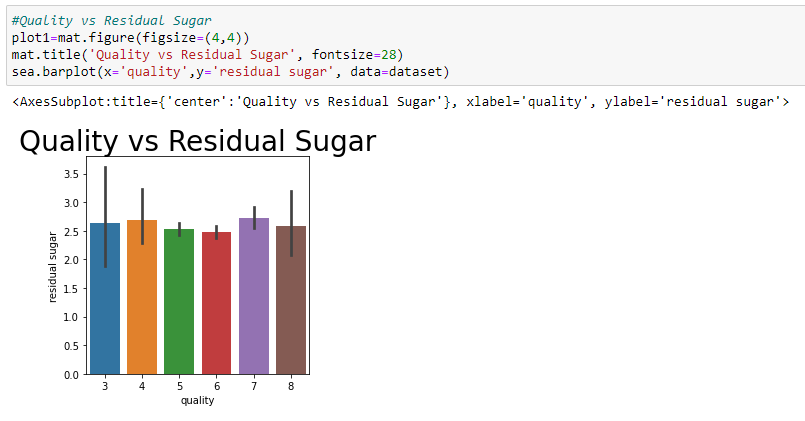
1. Quality vs Volatile Acidity



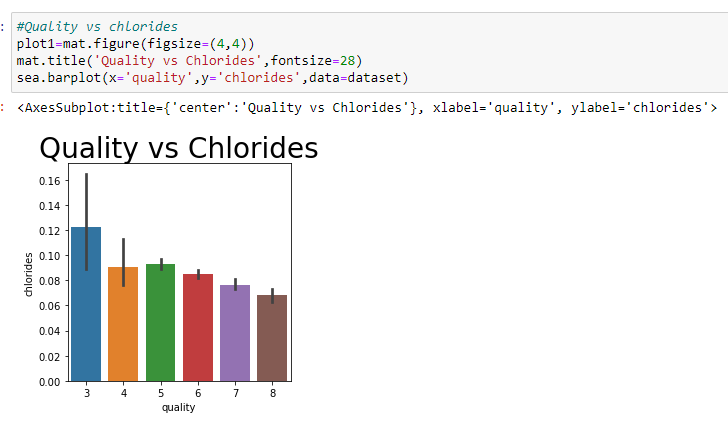
1. Quality vs Citric Acid



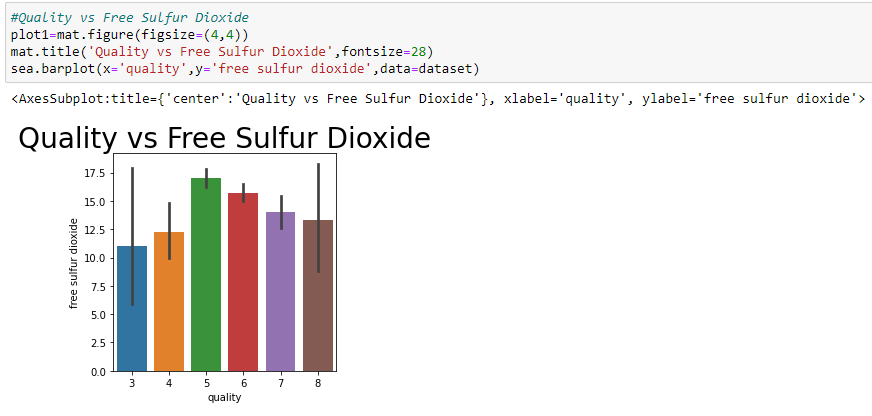
* Quality vs Residual sugar



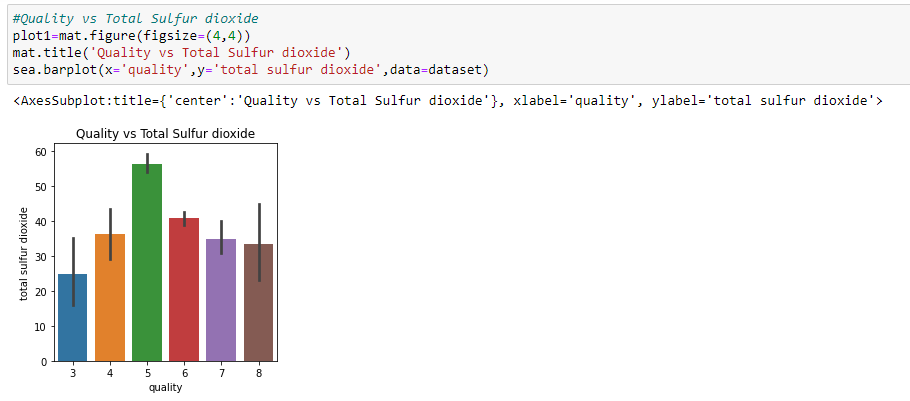
* Quality vs Chlorides



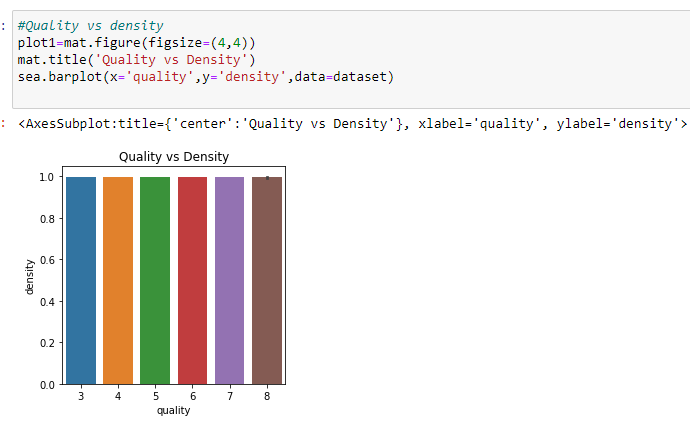
* Quality vs Free sulfur dioxide



* Quality vs Total sulfur dioxide



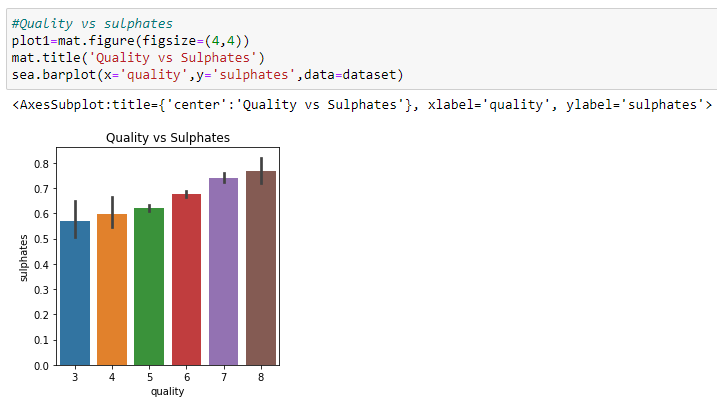
* Quality vs density



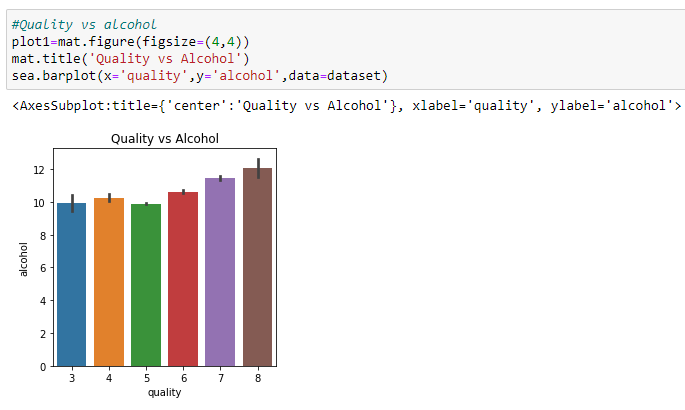
* Quality vs pH



* Quality vs Sulphates

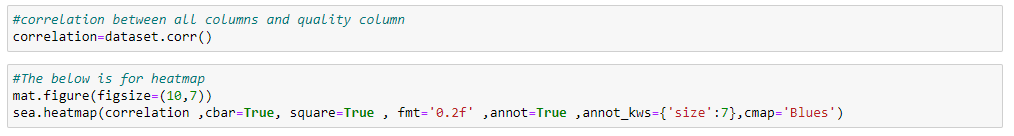


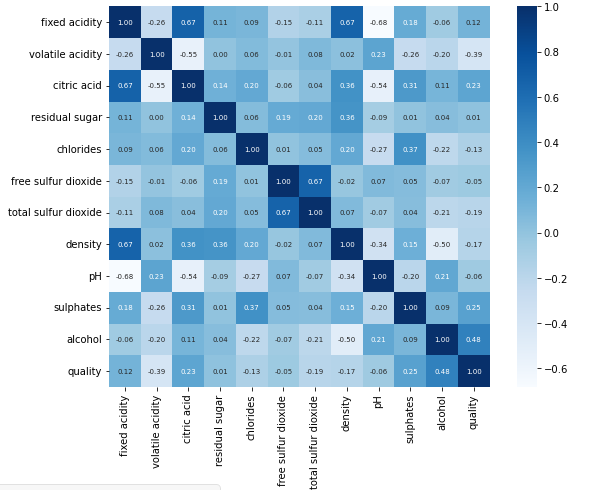
* Quality vs Alcohol



**Correlation matrix:**

The correlation matrix depicts how the dataset variables are related to one another.

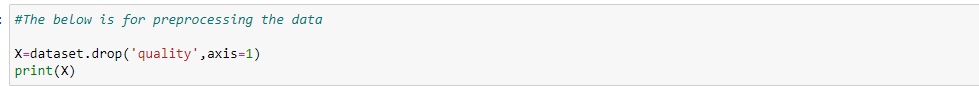


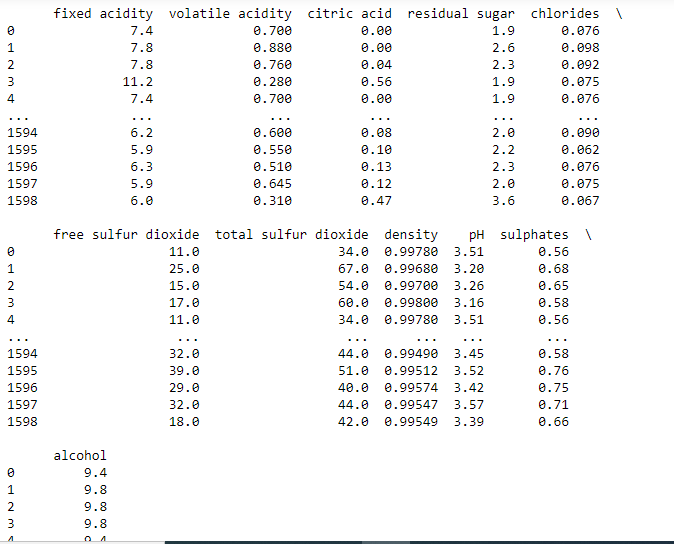


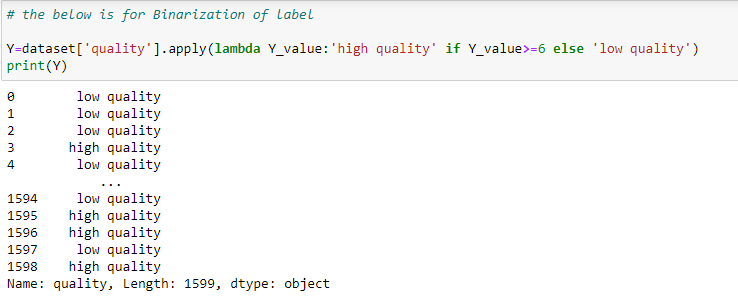
**Separating data and Label:**

All the data except data is stored in X and the quality data is stored in Y

Label binarization is performed to know the quality if it is high or low.



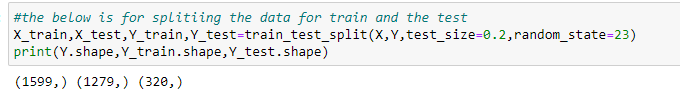




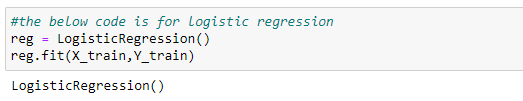
* **Train and Split:**

1. **test\_size**= 0.2: we split our dataset mainly into 2 parts i.e., train set and test set. The ratio of test set compare to dataset is 0.2 observations will be kept into test set. The test set should not be too large, if it is too large we will lack of data to train.
2. **train\_size:** if we have already used test\_size then the rest of data will automatically assigned to train\_size
3. **random\_state:** this is mainly used for random number generator. We can also place the instance of RandomState class. If we give it as 0 or blank, the RandomState instance used by np.random will be used.

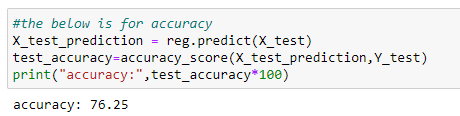
**We used 80% of data as training and 20% as testing**



* Model training with logistic regression



* Calculating the Accuracy score on testing data



**Result:**

Accuracy is : 76.25

The accuracy for this model is around 76.25%

* **Confusion matrix:**

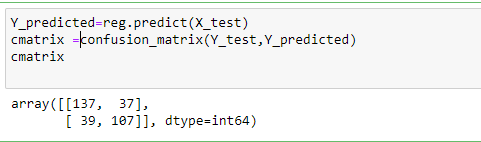
Confusion matrix is used as performance measurement for classification problem where output can be two or more classes. It has 4 different combinations of predicted and actual values.

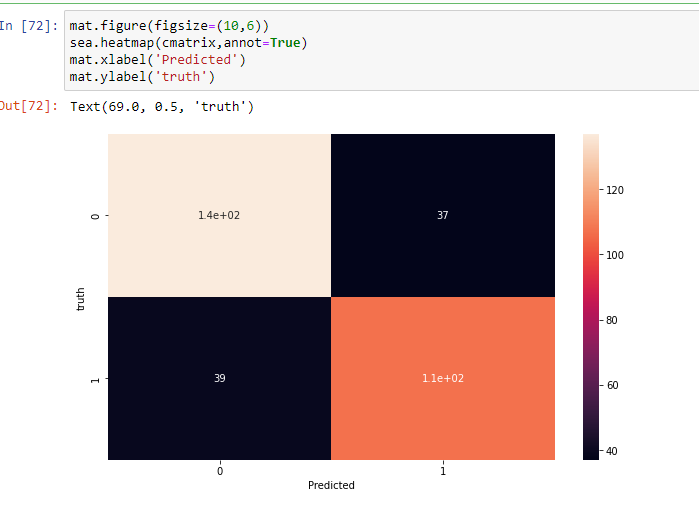
TP: It means True Positive, here the Predicted values are correctly predicted as actual positive

FP: Here the Predicted values incorrectly predicted an actual positive. i.e., Negative values predicted as positive

FN: It means False Negative, here the Positive values are predicted as negative

TN: It means True Negative, here the Predicted values correctly predicted as an actual negative





**Conclusion:**

I have 11 predictor variables , they are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol that predict one dependent variable quantity in this logistic regression model. Quality is inversely related to volatile, fixed acidity, chlorides, free and total sulfur dioxide, while the other variables are directly proportional.

This model also has accuracy which is 76.25%