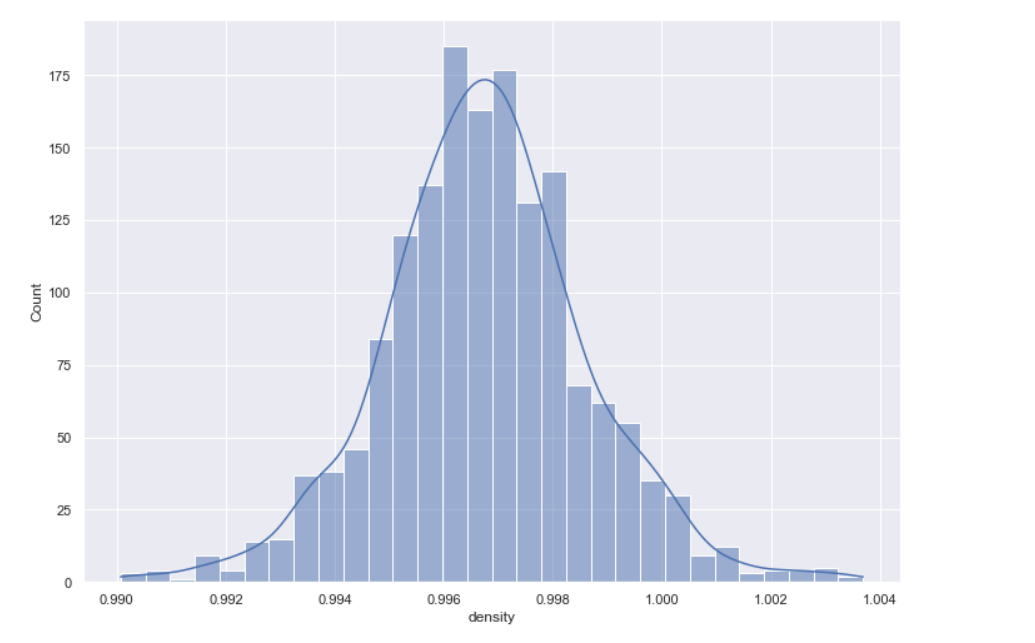
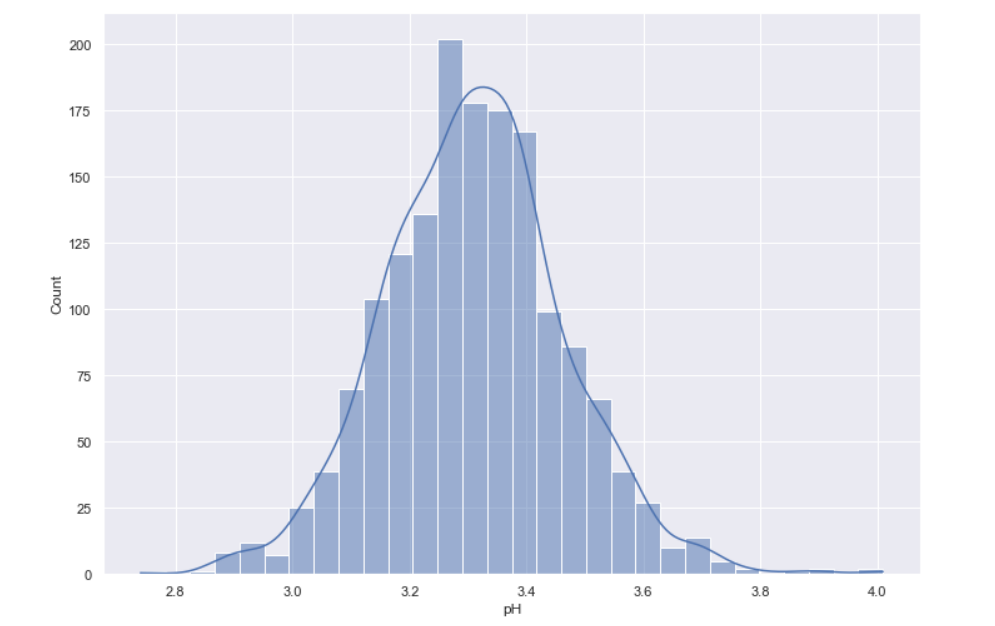
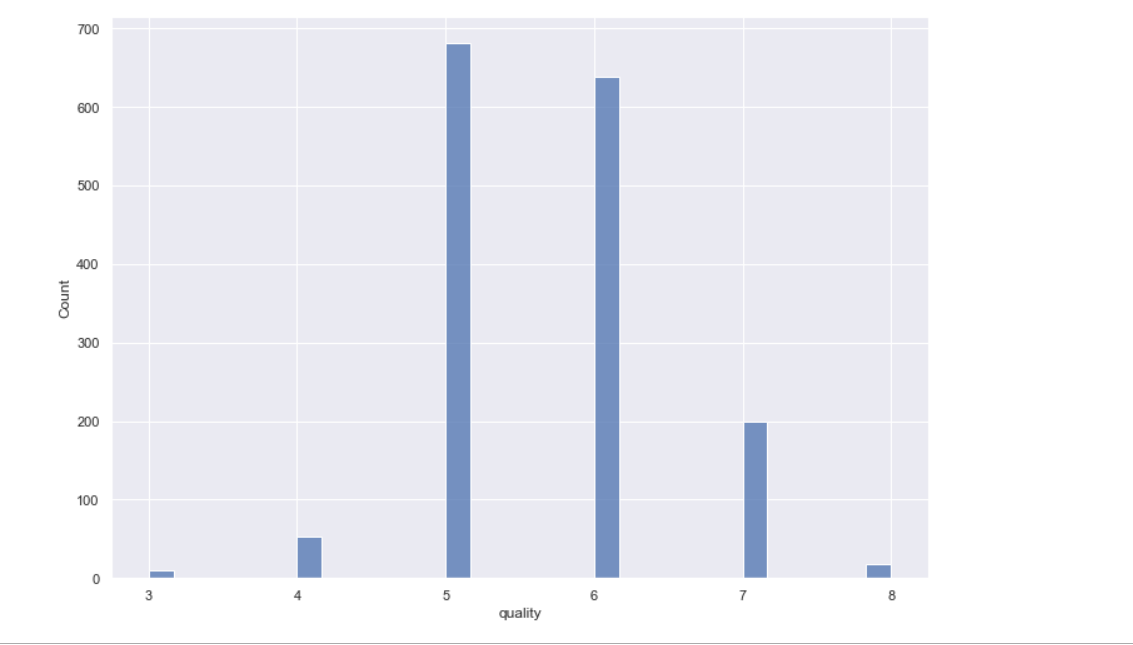
**PART B**

INSIGHTS FROM EDA

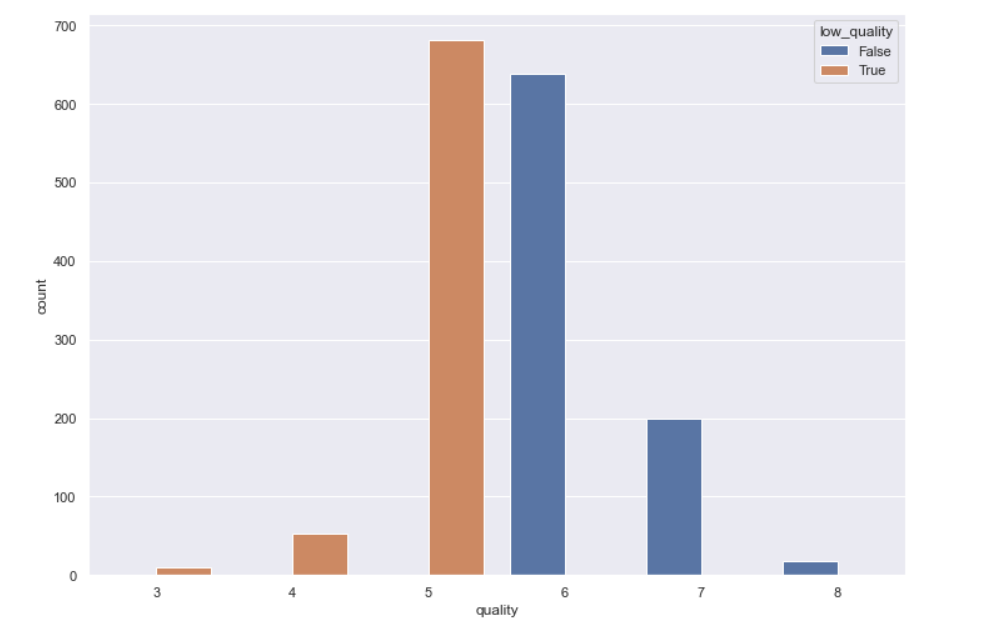
1.The histogram reveals that density and pH are normally distributed but the rest of the variables are more or less skewed.The quality dependent variable has a semi-normal discrete distribution.



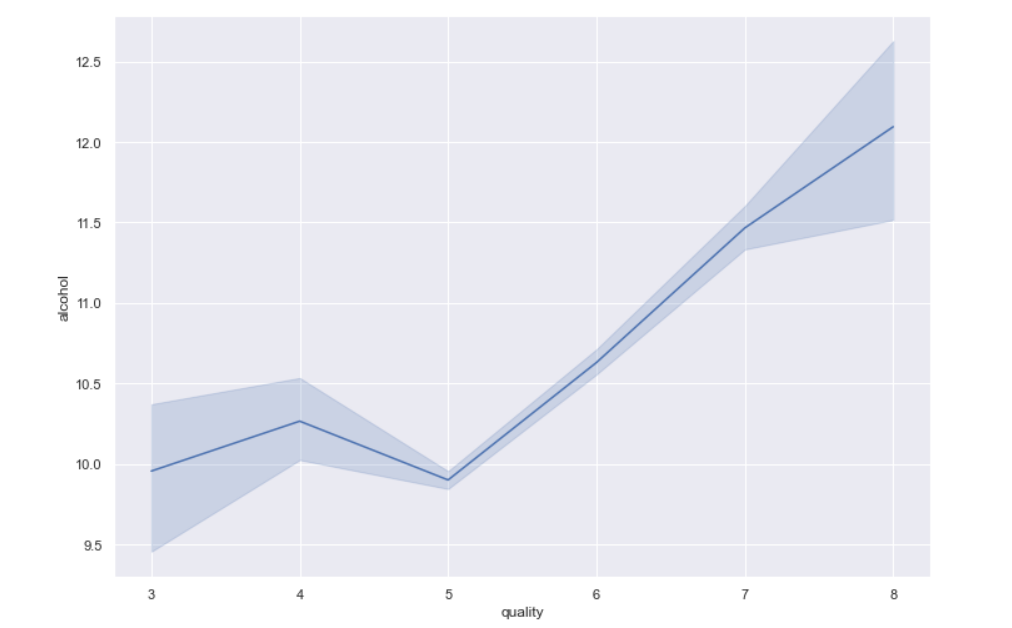




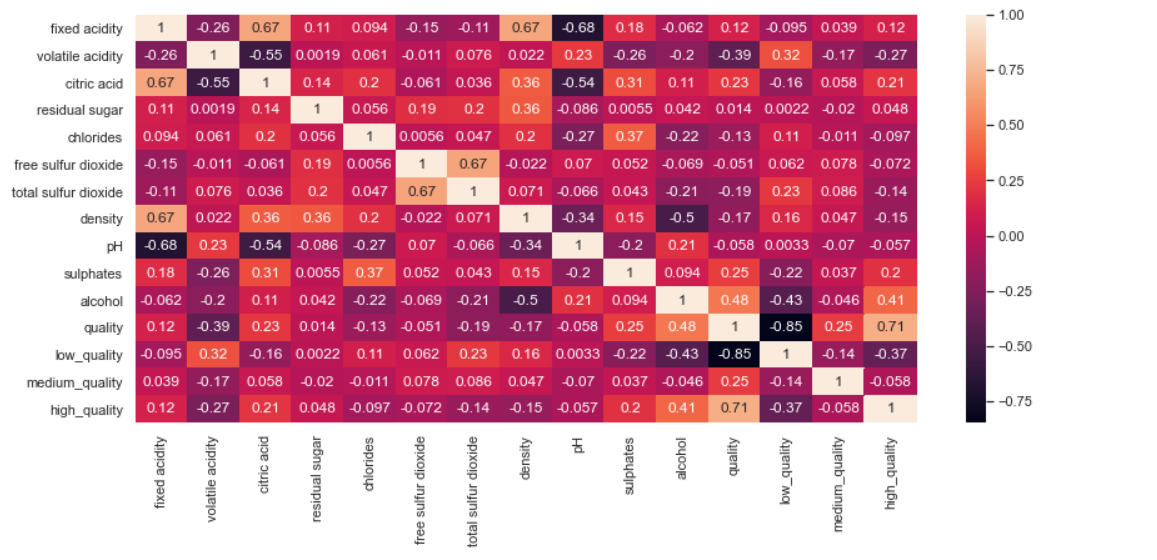
2. In the dataset, the minimum quality value is 3 and maximum is 8.The majority of red wines have either 5 or 6 ranking for their quality. About 200 red wines have quality ranking as 7. The rest are either 3,4, or 8.The average quality of the red wines is 5.63. The quality variable can be split as low quality if [0-5], medium [5-7],high[7-10].

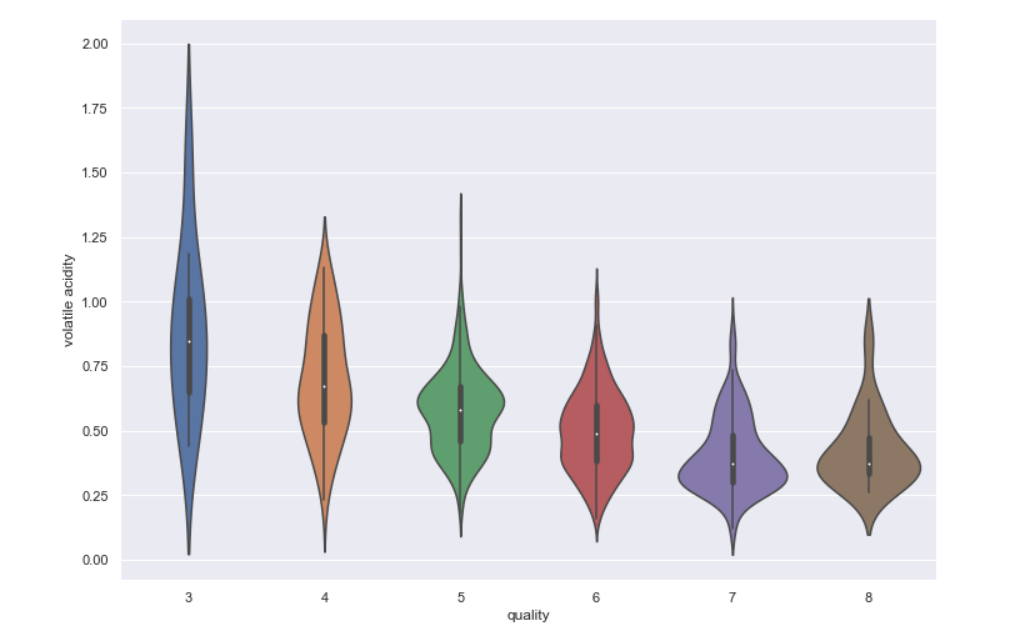


3. The alcohol content of the red wine increases with respect to the quality of the wine

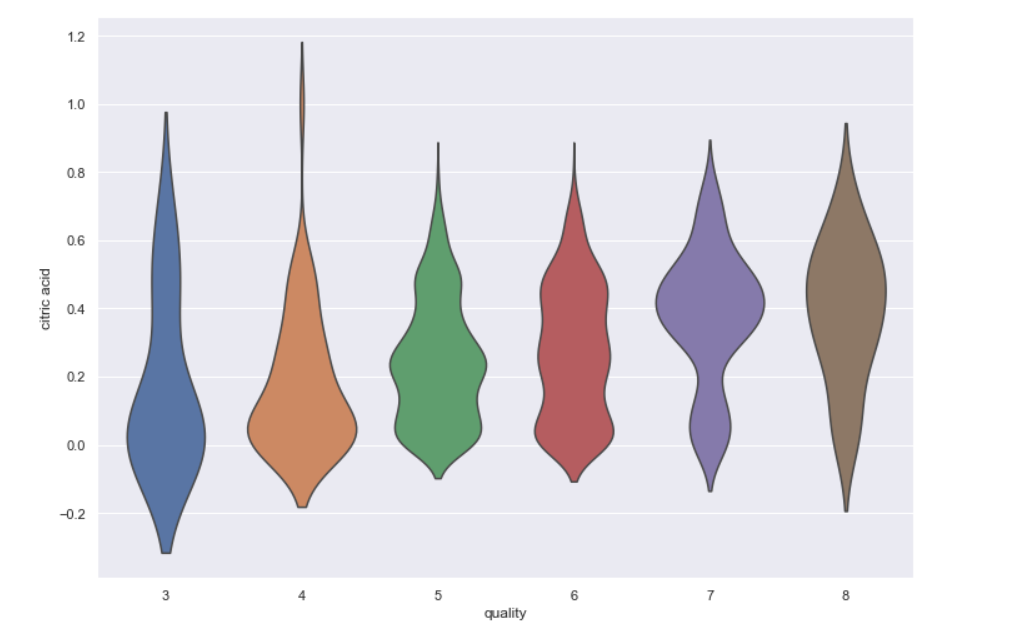


4. The correlation matrix shows that fixed acidity is highly positively correlated with citric acid and density.total\_sulfur dioxide is highly positively correlated with free\_sulfur\_dioxide.pH is highly negatively correlated with fixed acidity.citric acid is negatively correlated with volatile acidity and pH.



5. Less volatile acidity in a sample results in higher wine quality.

6.The bigger citric acid level in a sample on average the better quality of the sample is.



1.What are the assumptions of Linear Regression?

* The relationship between dependent variable y and independent variable x exists is linear.
* For every value of the independent variable x, there is an expected value of the dependent variable y and these values are normally distributed.
* y is a continuous random variable and x values are fixed and not random.
* The sampling errors associated with the expected value of y are associated to be an independent random variable distributed normally with mean 0 and constant standard deviation.

2.How can we evaluate a Regression model?Define each metric and its interpretation.

There are 3 mian metrics for model evaluation in Regression:

* R square/adjusted R square

R square measures how much variability in dependent variable can be explianed by the model.It is the square of the Correlation Coefficient (R) and that is why it is called R square.

R square is calculated by the sum of squared of prediction error divided by the total sum of the square which replaces the calculated prediction with mean.R square value is between 0 to 1 and a bigger value indicates a better fit between prediction and actual value.

* Mean Square Error(MSE)/Root Mean Square Error(RMSE)

While R square is a relative measure of how well the model fits dependent variable,Mean square error is an absolute measure of the goodness of the fit.MSE is calculated by the sum of square of prediction error which is real output minus predicted output and then divide by the number of data points.It gives us an absolute number on how much our predicted results deviate from the actual number.

* Mean Absolute Error(MAE)

MAE is similar to MSE.However, instead of the sum of square of error in MSE,MAE is taking the sum of the absolute value of error.Compare to MSE, MAE is more direct representation of sum of error terms.MSE gives larger penalization to big prediction error by square it while MAE treats all errors the same.

3.Can R squared be negative?

Yes

4.what is dummy variable trap?

The dummy variable trap is a scenario in which the independent variables are multicollinear - a scenario in which two or more variables are highly correlated;in simple terms one variable can be predicted from the others.To demonstrate the dummy variable trap,take the case of gender as an example.

5.Is One Hot Encoding different from Dummy Variables?

Dummy variable is a binary variable that indicates whether a separate categorical variabletakes on a specific value.One hot encoding technique is a function or code to create a dummy variable.

6.How is Polynomial Regression different from Linear Regression?

Polynomial Regression is a form of Linear Regression where only due to the Non-Linear relationship between dependent and independent variables we add some polynomial terms to linear regression to convert it into Polynomial Regression.

7.

Model.score() is used to get the accuracy of the model and model.predict is used predict the variable.

In (i) the variation of of x\_test is explained by the predicted value of x\_test.

In (ii) the variation of x\_train is explained by the predicted value of x\_train.

In (iii) and (iv) the accuracy of x\_test and y\_test is 90% and accuracy of x\_train and y\_train is 89%.