WINE QUALITY EDA

- MEGHA SINGHAL

Problem Statement

- The dataset is related to red and white variants of the Portuguese "Vinho Verde" wine. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available .

- Purpose:

To analyze the dataset to summarize the main characteristics in order to determine the quality of wine.

Approach:

- Performed basic EDA.

Data Description

1 - fixed acidity: most acids involved with wine or fixed or nonvolatile (do not evaporate readily)

2 - volatile acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste

3 - citric acid: found in small quantities, citric acid can add 'freshness' and flavor to wines

4 - residual sugar: the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet

5 - chlorides: the amount of salt in the wine

6 - free sulfur dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine

7 - total sulfur dioxide: amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine

8 - density: the density of water is close to that of water depending on the percent alcohol and sugar content

9 - pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale

10 - sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant

11 - alcohol: the percent alcohol content of the wine

Output variable (based on sensory data): 12 - quality (score between 0 and 10)

Exploratory data analysis.

-There are 6497 records and 12 columns in WINE QUALITY DATASET

-Stepwise Univariate , Bivariate and Multivariate Analysis is performed to know how each variable affects outcome variable.

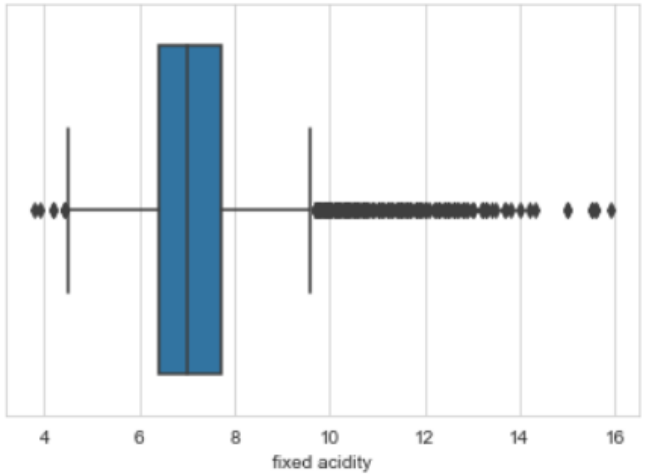
Boxlots are great for summarizing the shape of many datasets. For Univariate analysis Boxplots are used to determine outliers and skewness in dataset.

Bar plots are constructed for such cases, and the inferences noted down.

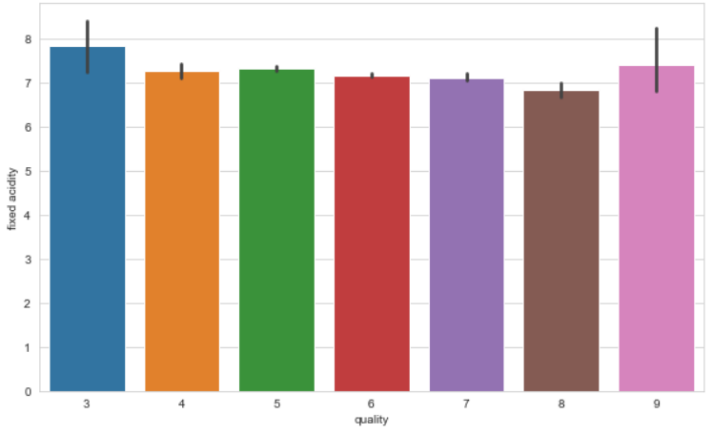
Scatter plots are constructed for numerical variables along with the respective conclusions from them.

This procedure forms bulk of EDA and gives business insights necessary to drive into problem statement, being summarized henceforth.

Fixed acidity

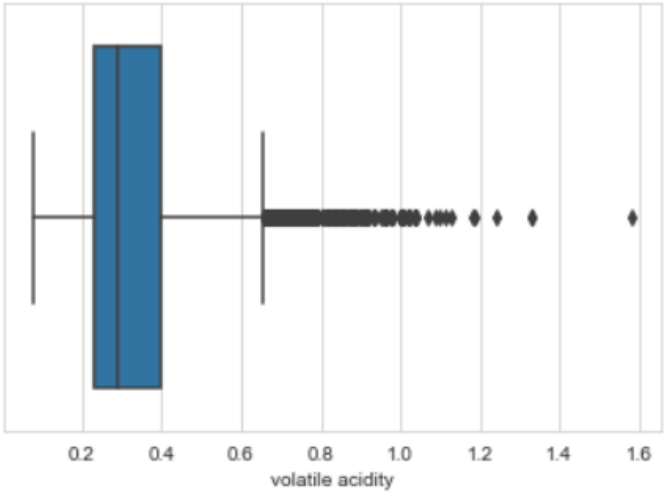


- The distribution of fixed.acidity data is lightly right skewed with minimum value of 3.80, maximum of 15 and median of 7 and mean of 7.215.

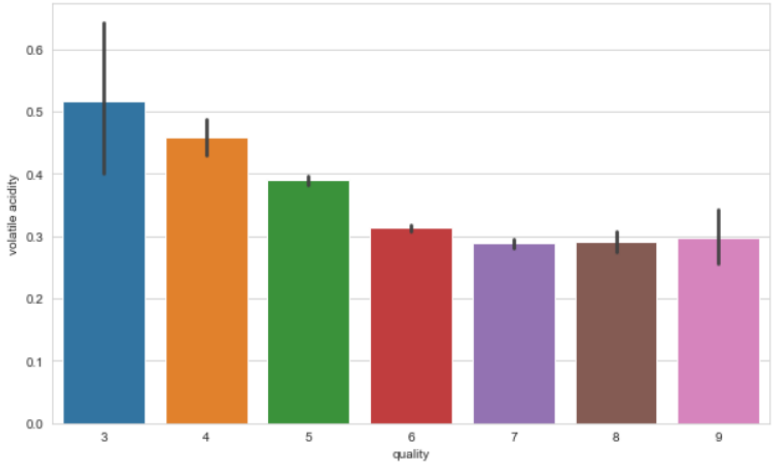


- Here we see that fixed acidity does not give any specification to classify the quality.

Volatile acidity

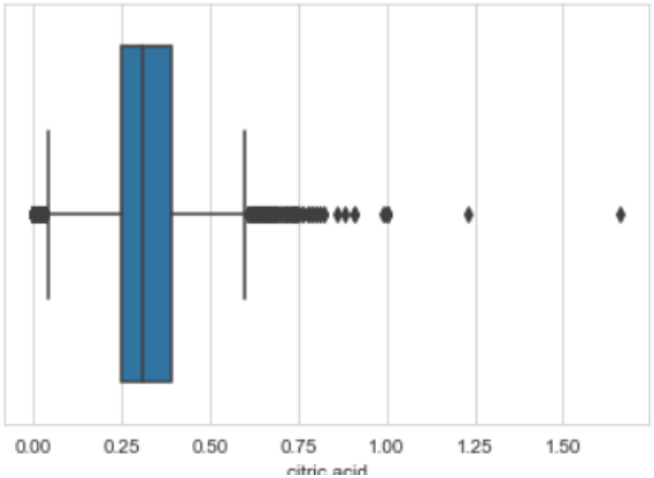


- The distribution of volatile.acidity data is right skewed ,with minimum value of 0.08, maximum of 1.58 and median of 0.29 and mean of 0.33.

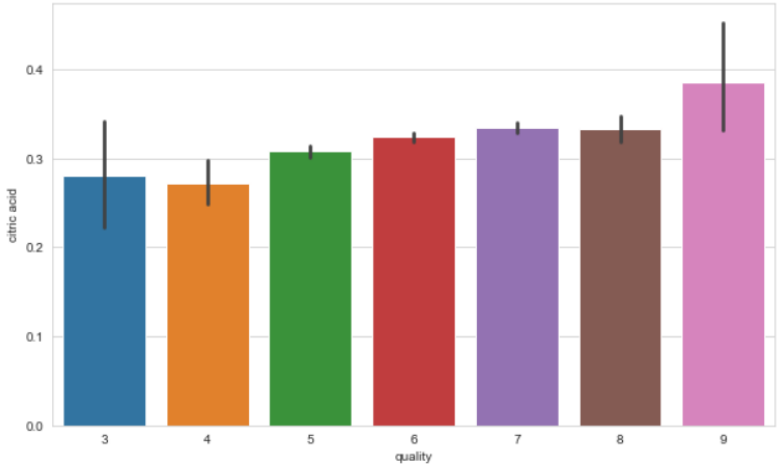


- Here we see that its quite a downing trend in the volatile acidity as we go higher the quality.

Citric Acid

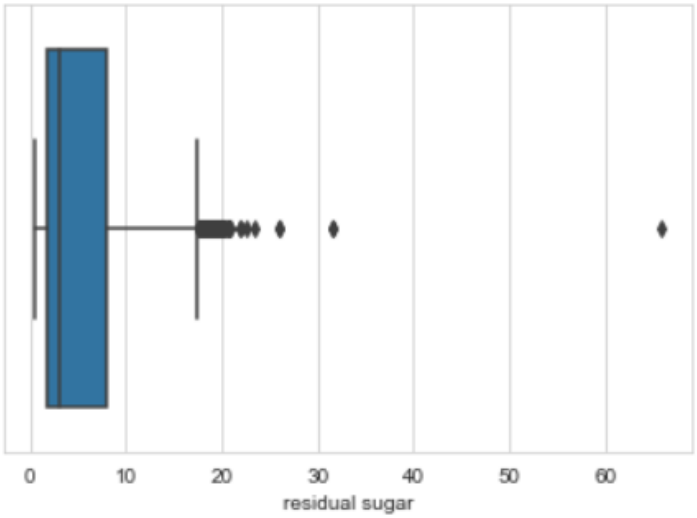


- The distribution of citric.acid data is right skewed.

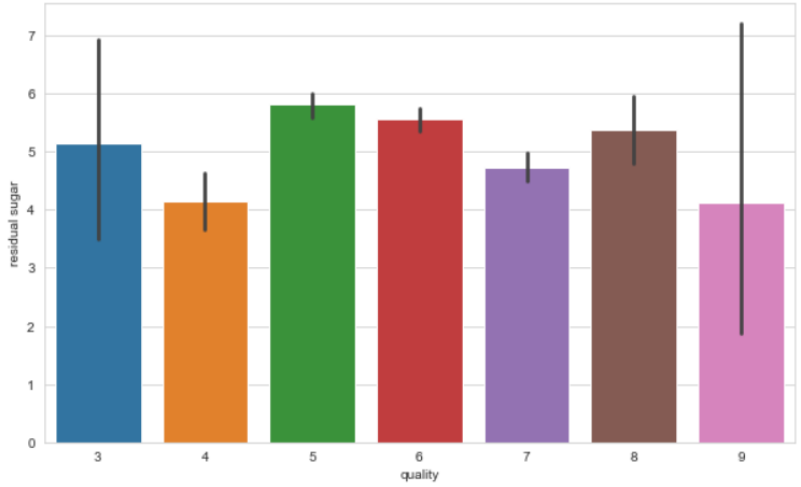


- Here we see that its quite a upward trend in the citric acid as we go higher the quality.

Residual Sugar

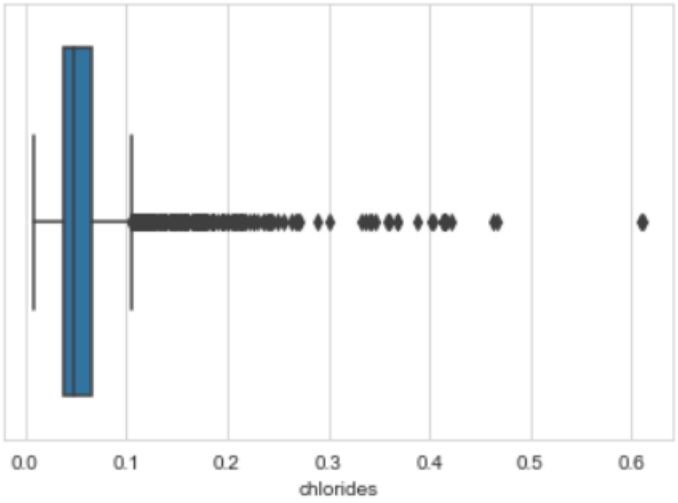


- The distribution of residual sugar data is right skewed (long-tailed) with minimum value of 0.6, maximum of 65 (many outlayers) and median of 3 and mean of 5.44.

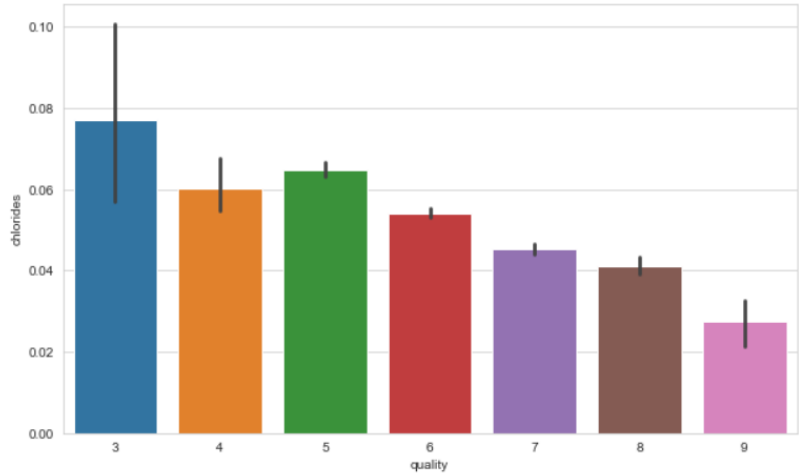


- Here we see that residual sugar does not give any specification to classify the quality.

Chlorides

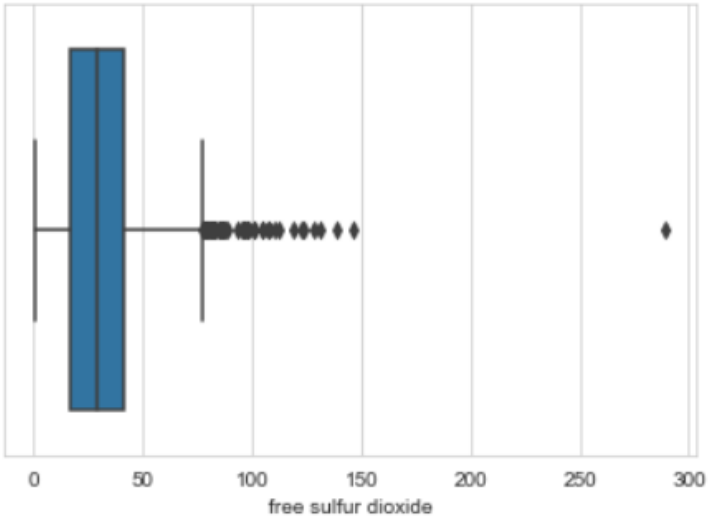


- The distribution of chlorides data is also right skewed (long-tailed) and very similar to distribution of residual.sugar with minimum value of 0.009, maximum of 0.611 (many outlayers) and median of 0.047 and mean of 0.005.

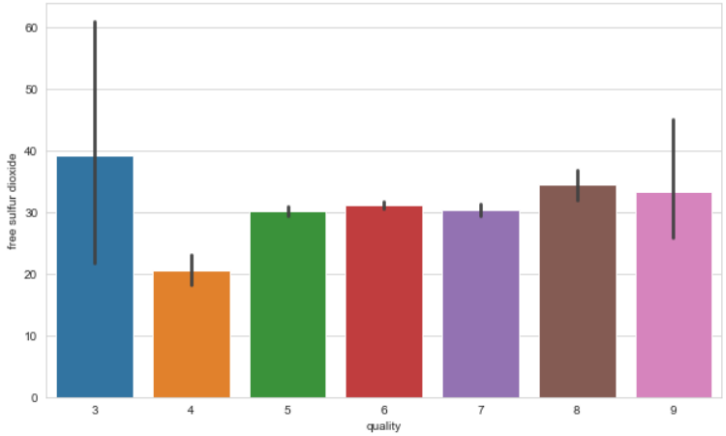


- Composition of chloride also go down as we go higher in the quality of the wine

Free Sulpher Dioxide

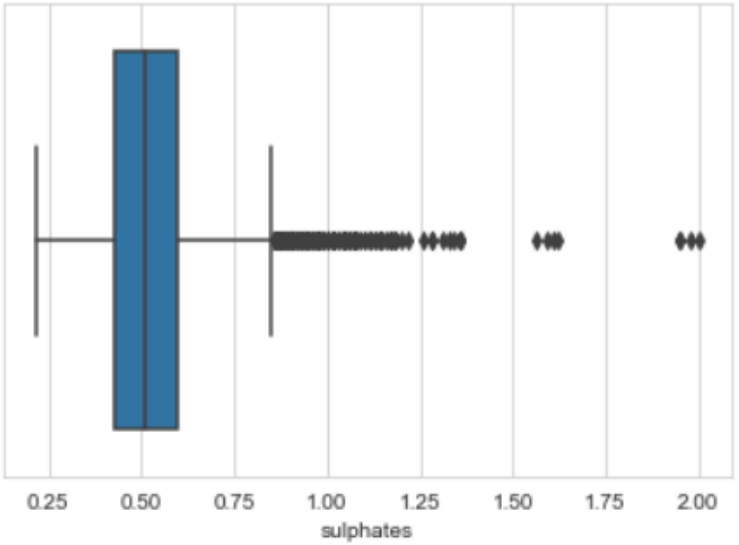


- The distribution of sulpher dioxide data is also right skewed (long-tailed) with minimum value of 1, maximum of 289 (many outlayers) and median of 29 and mean of 30.5.

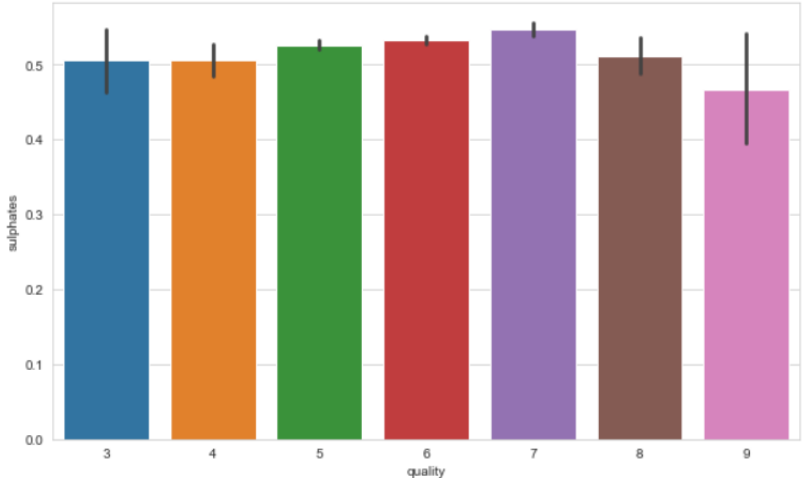


- Here we see that residual sugar does not give any specification to classify the quality.

Sulphate

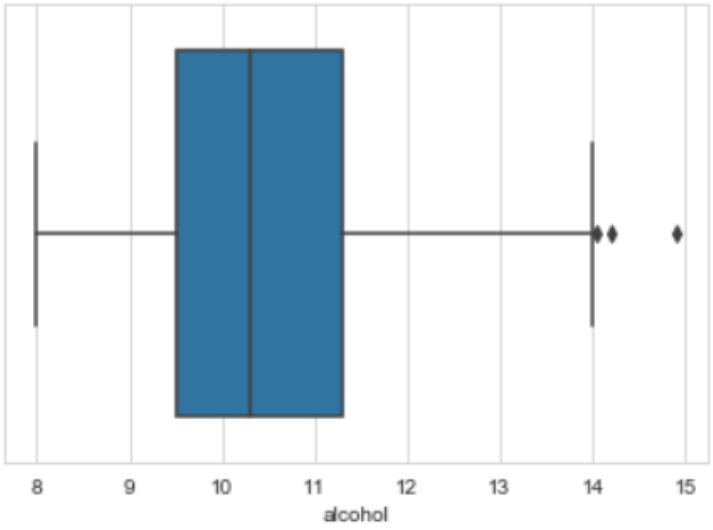


- The distribution of sulphates data is right skewed (long-tailed) with minimum value of 0.22, maximum of 2 and median of 0.51 and mean of 0.53.

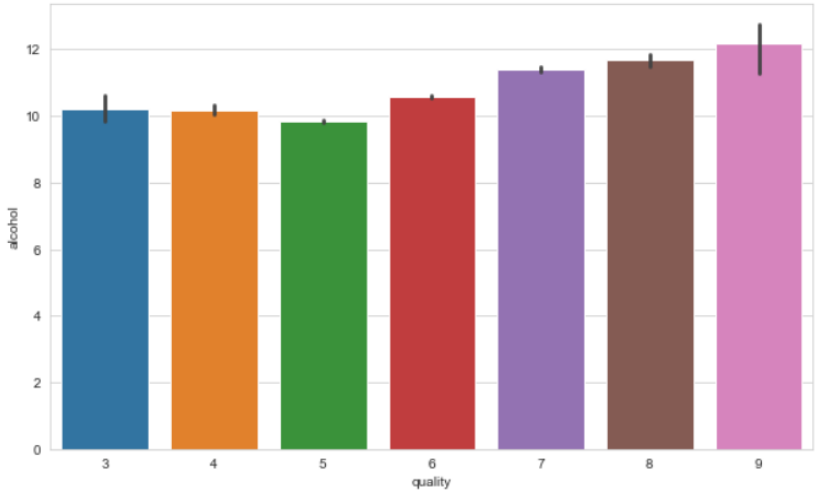


- -The average sulphates value decreases as the quality level increases.

Alcohol



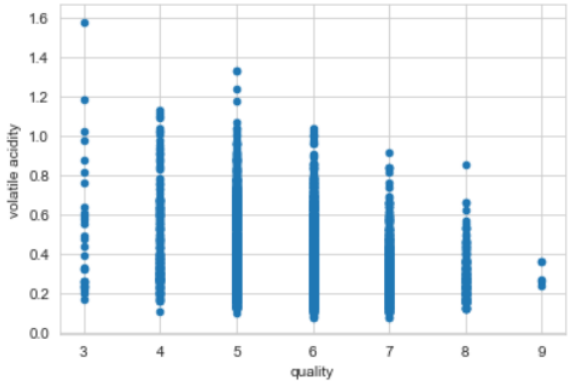
- The distribution of alcohol data is right skewed but does not have many outliers with minimum value of 8.0, maximum of 14.9 and median of 10.3 and mean of 10.49.



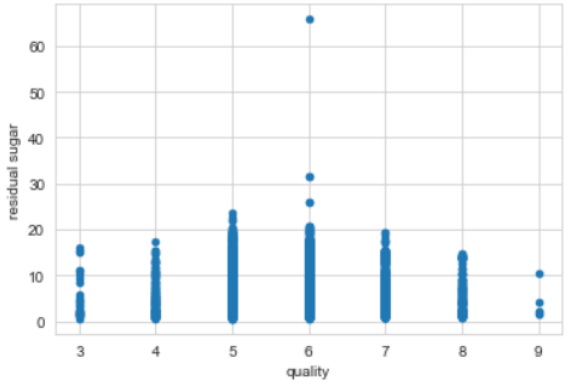
- Alcohol level also goes higher as the quality of wine increases

Based on scatterplots of quality against different feature variables, which of the following is most likely to have a positive impact on quality?

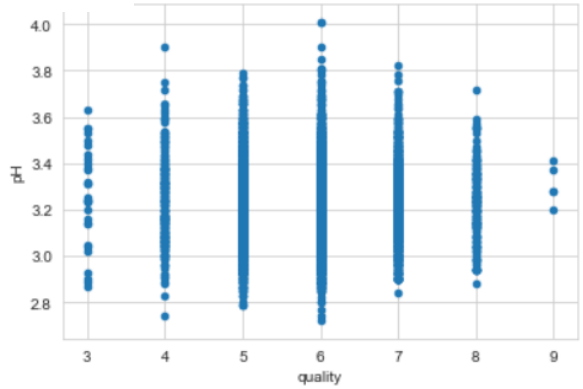
Volatile acidity



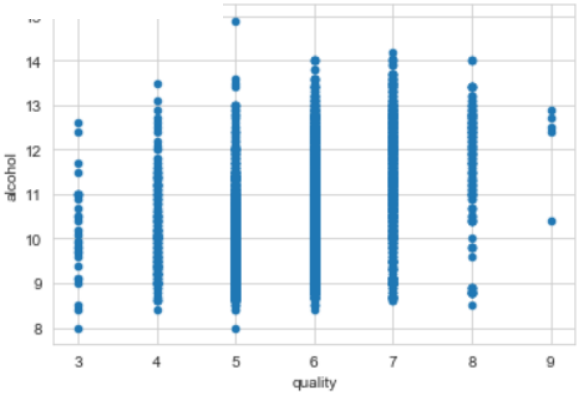
Residual Sugar



PH

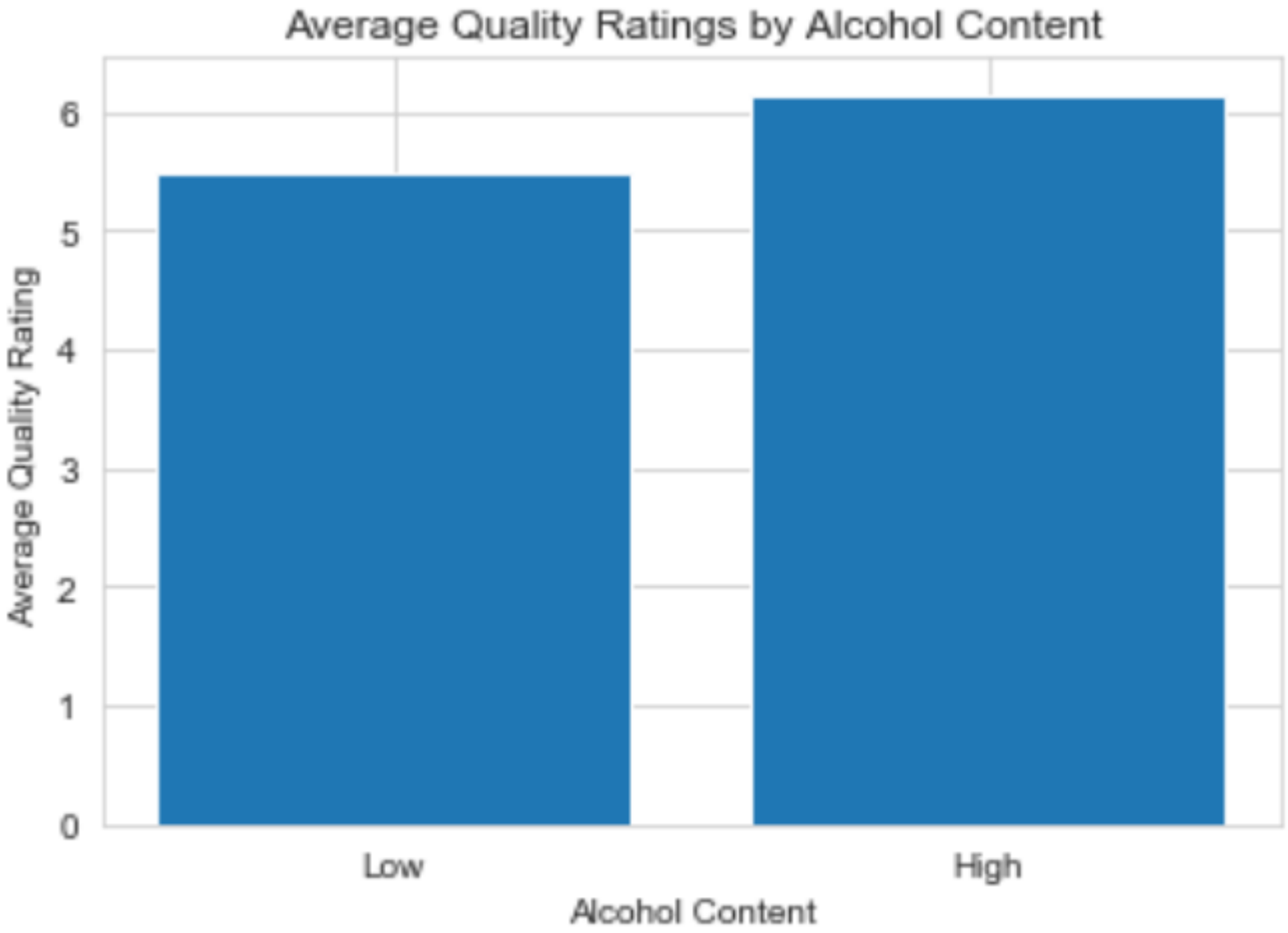


Alcohol



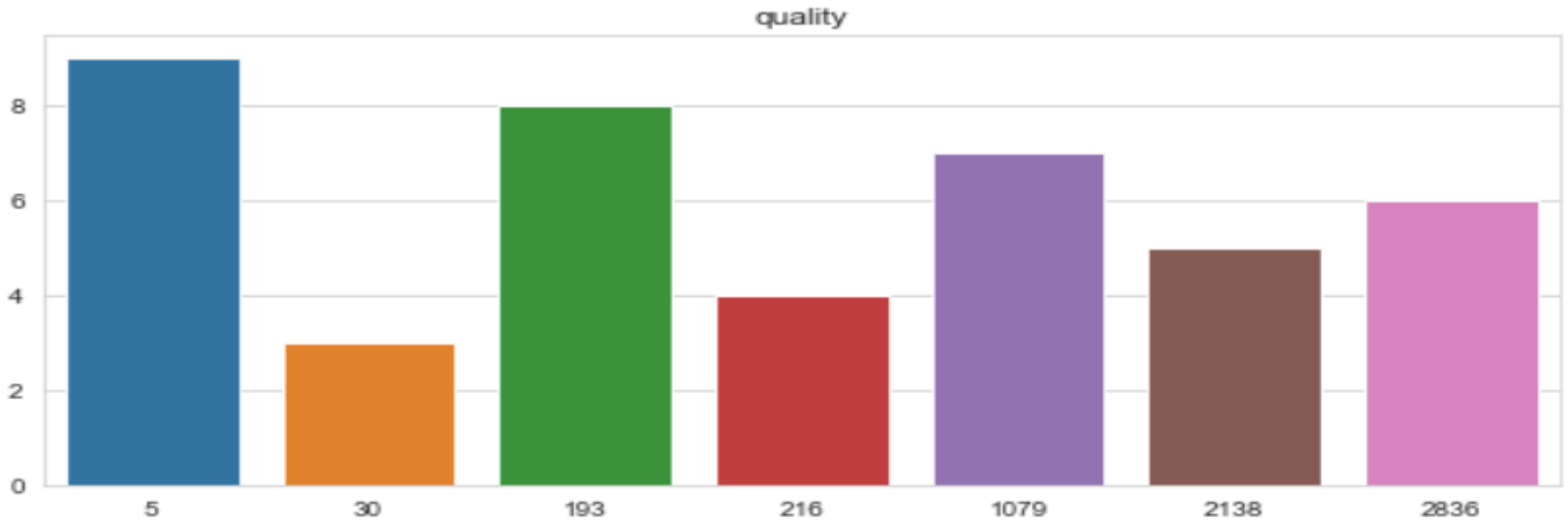
- ALCOHOL is most likely to have a positive impact on quality.

Do wines with higher alcoholic content receive better ratings?



- Yes, Wines with higher alcoholic content generally receive better ratings

How many wines with particular quality ?

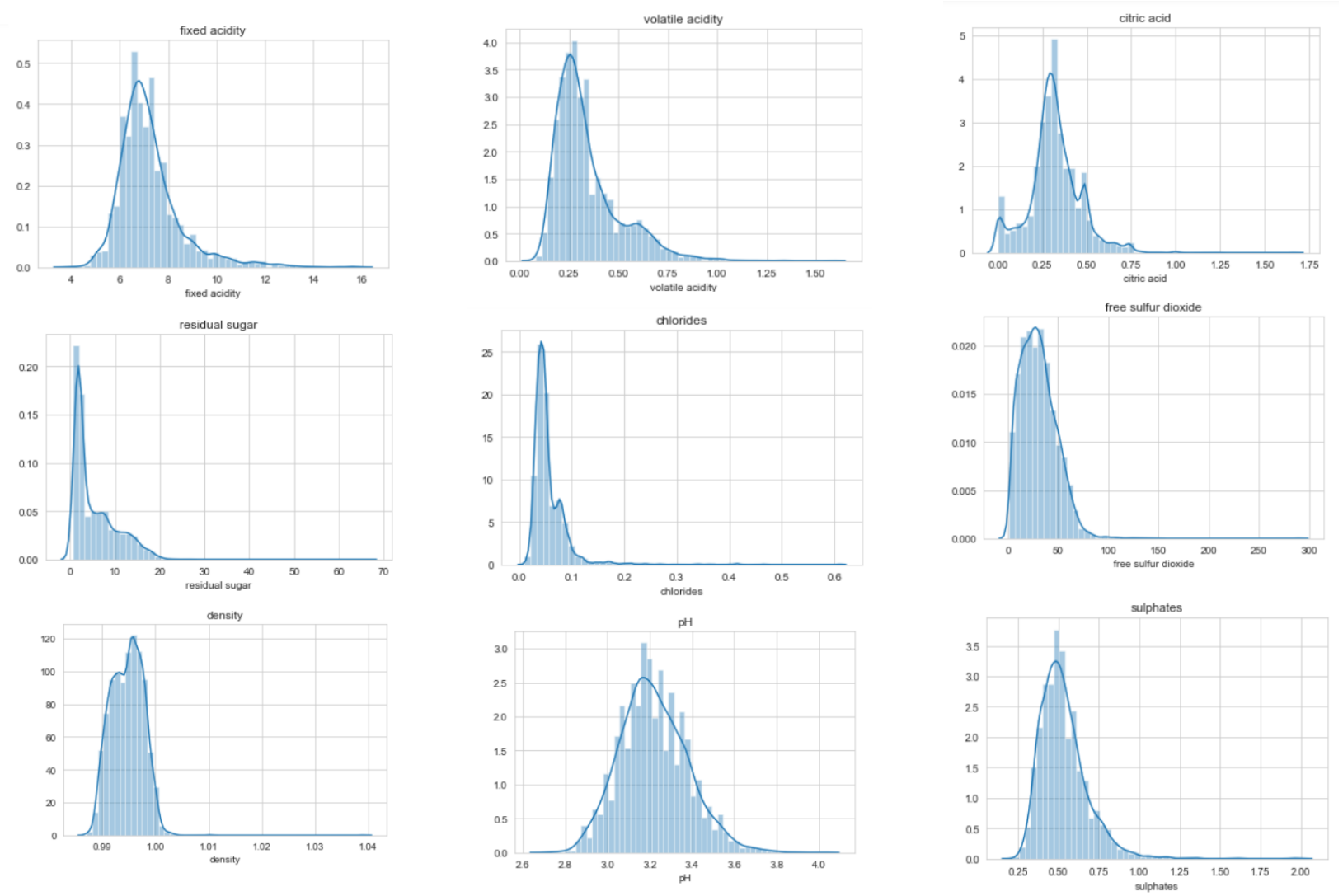


All features in the data set contain numerical values 8 different quality levels have been determined for wines; 3, 4, 5, 6, 7, 8,9 The lowest value is 3 and the highest value is 9.

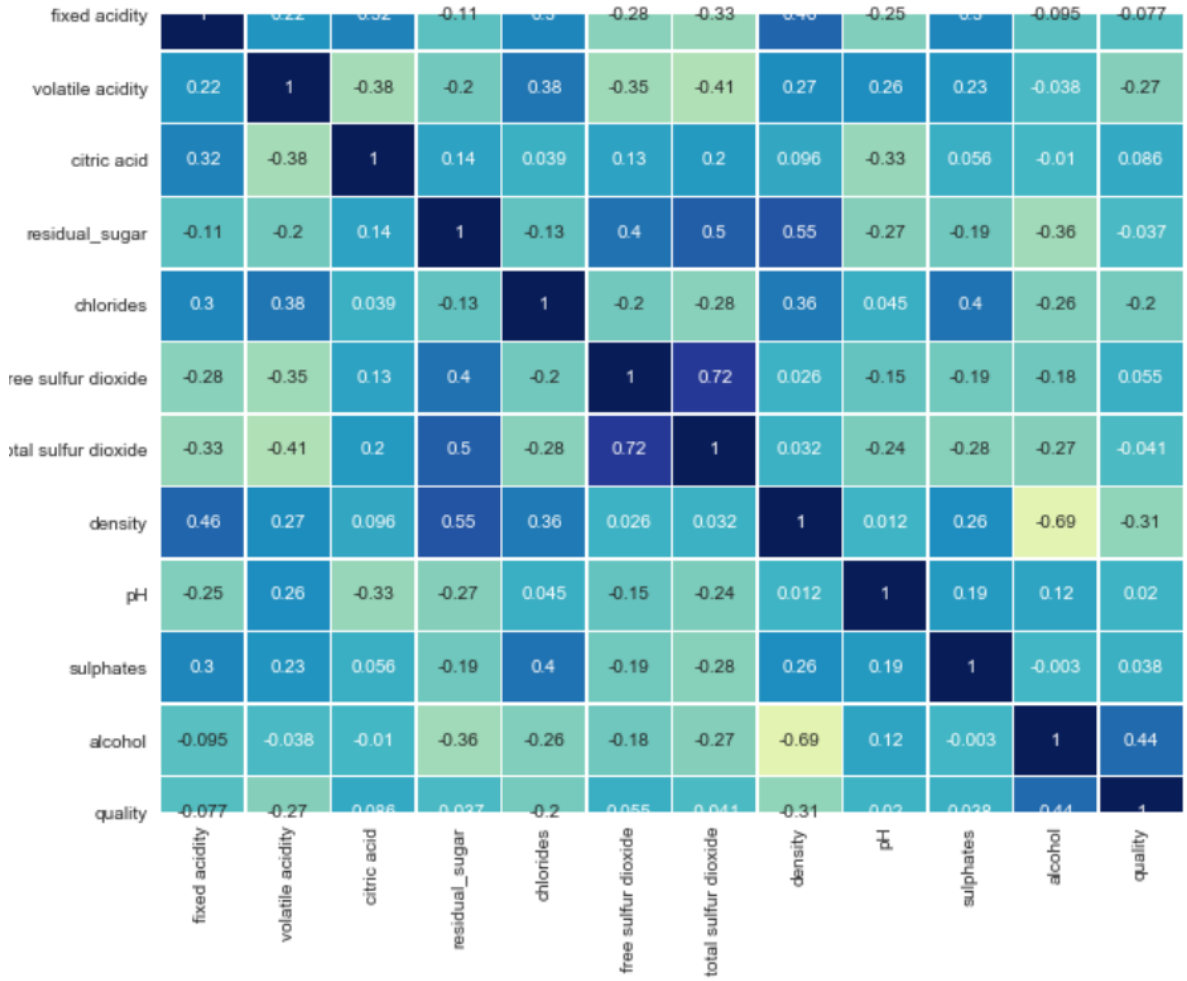
Class distribution of the samples in the data set is unstable. Unbalanced class distribution is undesirable for machine learning models. Considering this imbalance, the 'stratify' feature will be used when leaving the data set for training and testing.

To cope with the unbalanced distribution of classes, classes with few samples are replicated to have the same number of samples as classes with multiple samples.

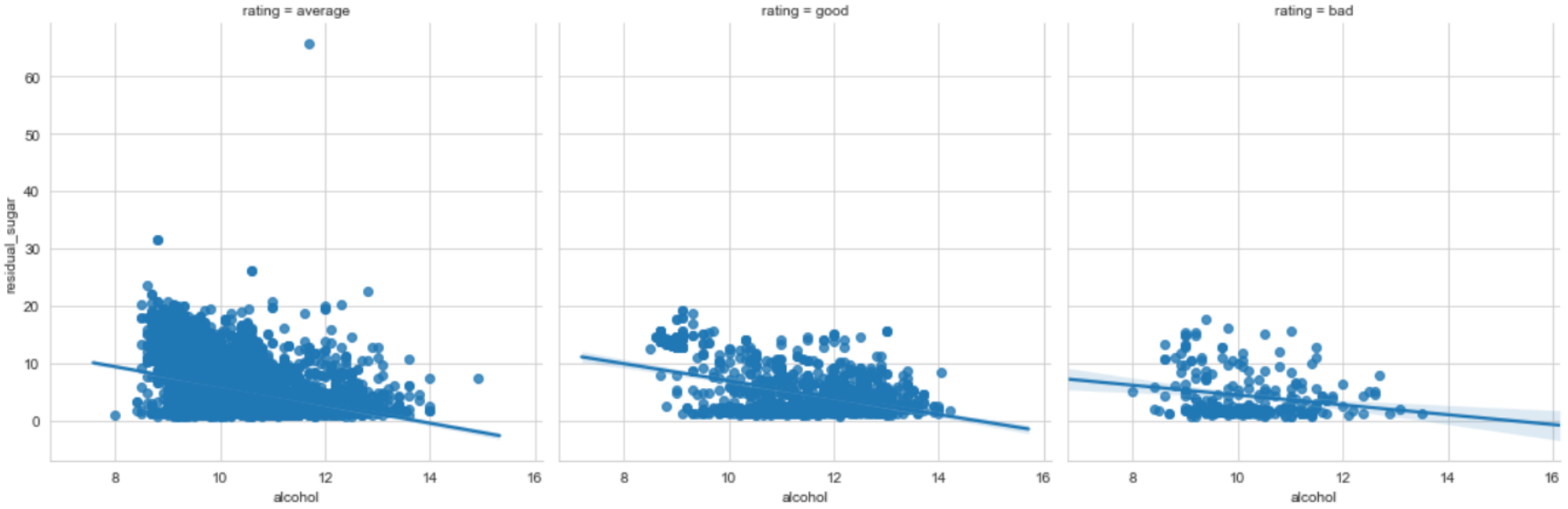
Density plot to check linearity



“pH” column appears to be normally distributed remaining all independent variables are right skewed/positively skewed.



As can be seen from the above heat-map, there is a positive correlation between residual sugar and density, also there is a negative correlation between alcohol and density.



-The linear regression plots above for different wine quality ratings (bad, average & good) shows the regression between alcohol and residual sugar content of the red wine.

-We can observe from the trendline that, for good and average wine types the residual sugar content remains almost constant irrespective of alcohol content value. Whereas for bad quality wine, the residual sugar content increases gradually with the increase in alcohol content.

-This analysis can help in manufacturing the good quality wine with continuous monitoring and controlling the alcohol and residual sugar content.

Conclusion

-We observed the key factors that determine and affects the quality of the red wine. Wine quality is ultimately a subjective measure. The ordered factor 'quality' was not very helpful and to overcome this, so we created another variable called 'rating'.

-Our analysis showed that high quality wines typically have high alcohol content (b/w 10 to 14%) and lower residual sugar (0-20 units).

-This new information suggests that higher quality wines uses moderately sweet grapes and longer fermentation process.

-Too much S02 or volatile acidity means lower quality of wine.

-A right amount of citric acid is a correct balance of a quality wine.

-The usage of this analysis will help to understand whether by modifying the variables, it is possible to increase the quality of the wine on the market. If you can control your variables, then you can predict the quality of your wine and obtain more profits.

Thank You!