Jigsaw-MISB Bocconi

Graded Assignment - 03

Executive Summary

Red Wine Quality

(Linear Regression Model)

Author: Rohan Shetty (EPBA4)

Problem Statement

- In this project, I will analyze the ***Red Wine Data*** and try to understand which variables are responsible to ***predict the quality of wine on a scale of 0 to 10***.
- First I will try to get a feel of the variables on their own and then I will try to find out the ***correlation*** between them and the Wine Quality.
- Finally I will create a ***linear regression model*** to predict the outcome of a test set data.

Database Description

Citation Request:
Please include this citation if you plan to use this database:

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.
Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Source:
Paulo Cortez, University of Minho, Guimarães,
Portugal, http://www3.dsi.uminho.pt/pcortez
A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal
@2009

Kaggle URL: https://www.kaggle.com/piyushgoyal443/red-wine-dataset/data

Database Description

Data Set Information:

The datasets is related to red variants of the Portuguese "Vinho Verde" wine. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

- Number of Instances: red wine - **1599**
- Number of Attributes: **11**
- Missing Attribute Values: **None**

Description of attributes:

Input Variables:

fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol

Output variable:

quality (score between 0 and 10)

Preliminary Analysis

Dataset Structure:
The Red Wine Dataset had **1599 rows** and **13 columns** originally. After I added a new column called ‘rating’, the number of columns became 14. Here our categorical variable is ‘quality’, and the rest of the variables are numerical variables which reflect the physical and chemical properties of the wine.
I also see that in this dataset, most of the wines belong to the ‘average’ quality with very few ‘bad’ and ‘good’ ones. Now this raises my doubt if this dataset is a complete one or not.
For the lack of these data, it might be challenging to build a predictive model as I don’t have enough data for the Good Quality and the Bad Quality wines.

Point of Interest:
My main point of interest in this dataset is the ‘quality’. I would like to determine which factors determine the quality of a wine.

Preliminary Analysis

My initial thoughts and Hypothesis:

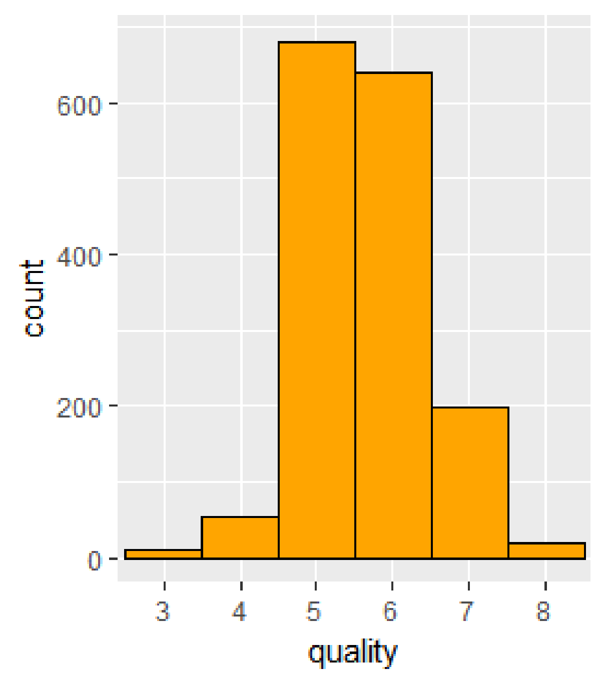
- Without analyzing the data, I think maybe the acidity (fixed, volatile or citric) will change the quality of wine based on their values.
- I also think the residual sugar will have an effect on the wine quality as sugar determines how sweet the wine will be and may adversely affect the taste of the wine.
- Lastly, I think alcohol content will also determine how strong the wine is.

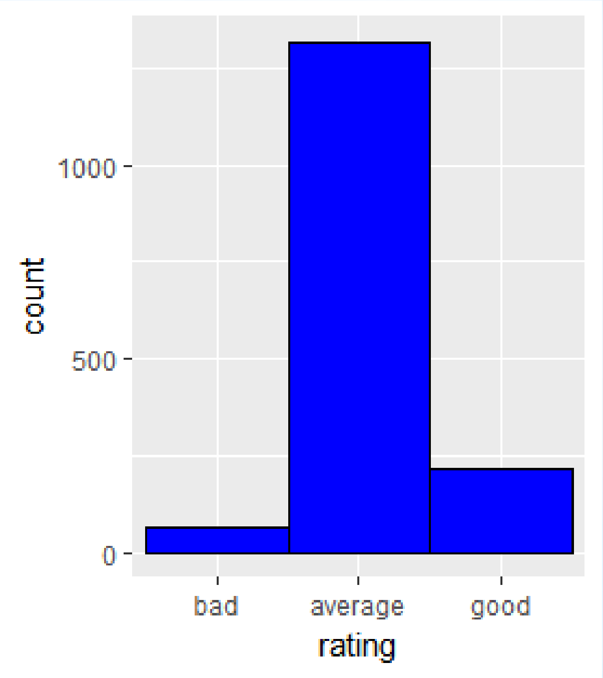
Analysis

- Firstly, I am going to plot the distribution of each of the variable.

- Based on the distribution shape, i.e. Normal, Positive Skew or Negative Skew, I can get some sense what to expect when I plot different variables against each other.

Analysis





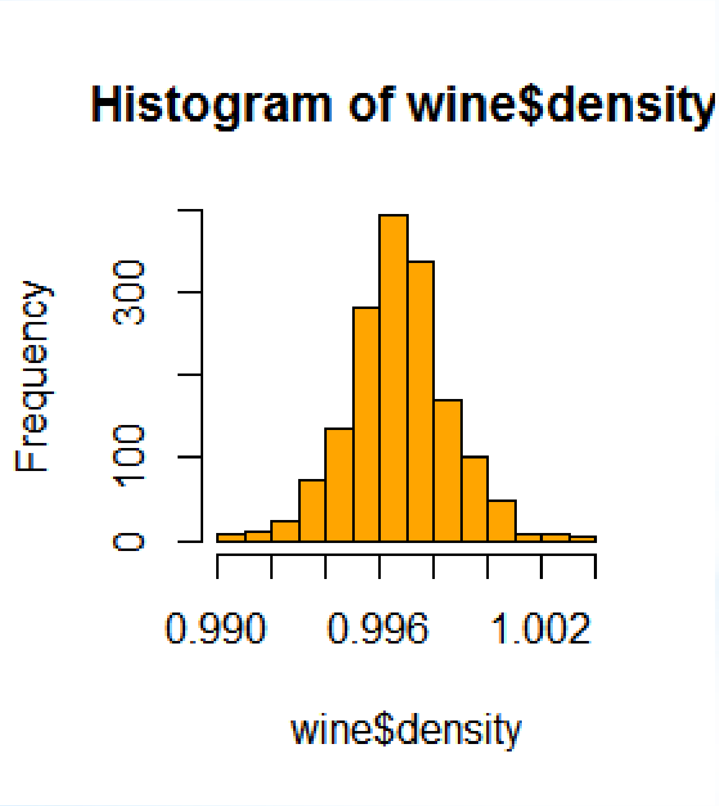
- One thing I can see from the above two plots is most of the wines in the dataset are average quality wines.
- As the good quality and the poor quality wines are almost like outliers here, it might be difficult to get an accurate model of the Wine Quality.

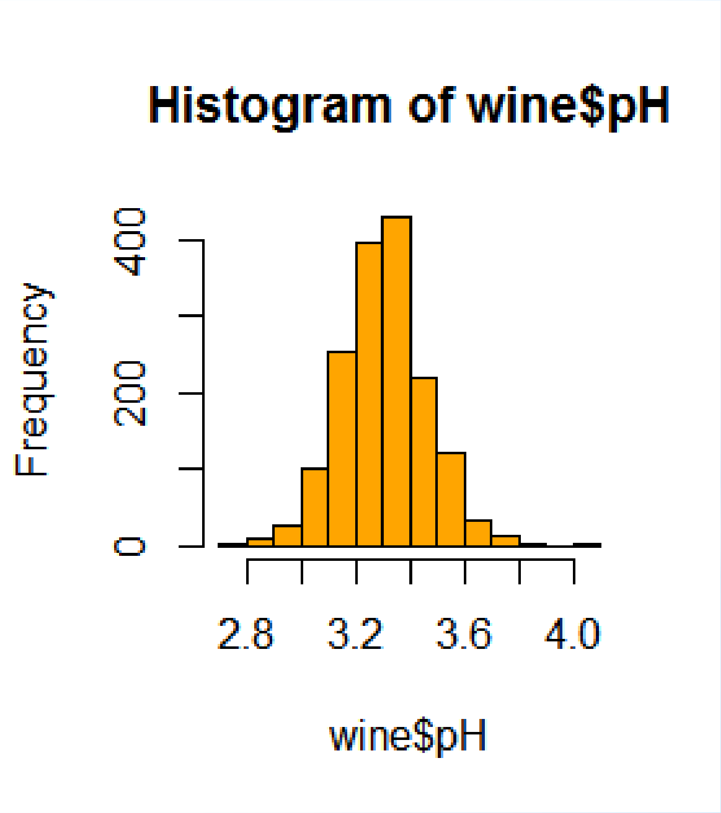
Let's look at the other plots.

Analysis

Distribution and Outliers

- Density and pH seems normally distributed with few outliers.

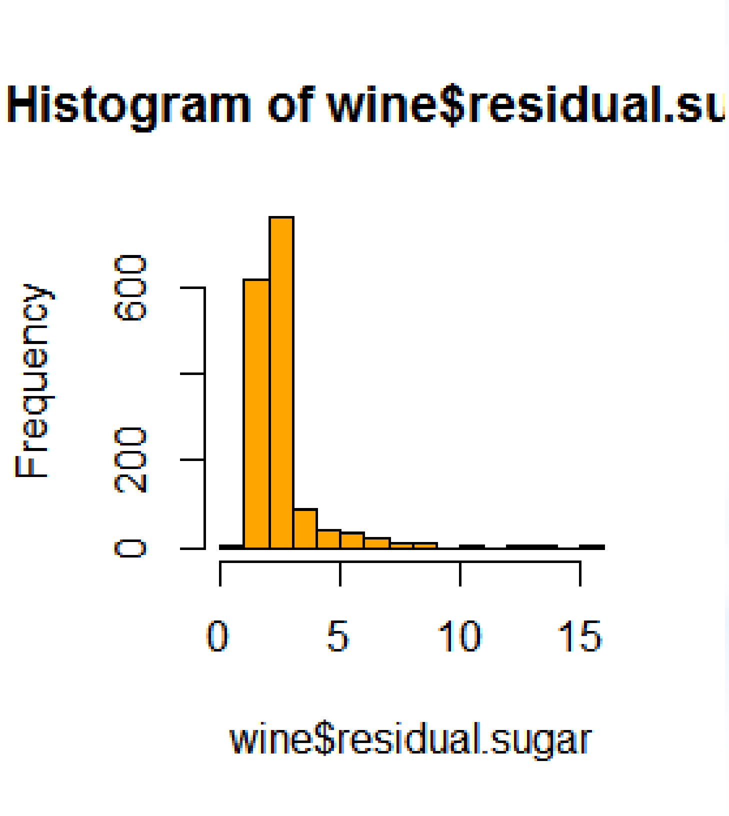


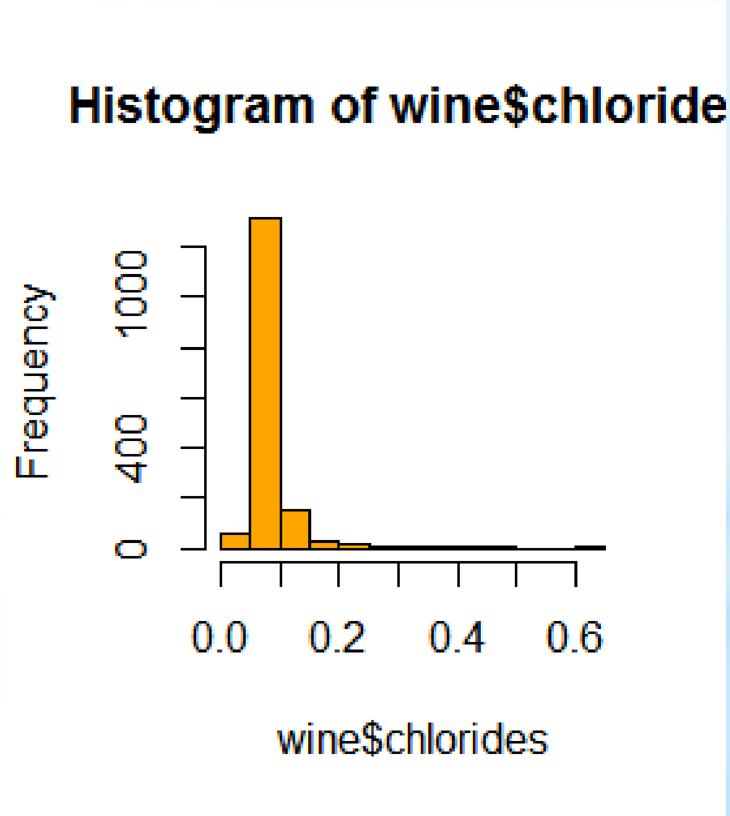


Analysis

Distribution and Outliers

- Residual sugar and Chloride seems to have extreme outliers.

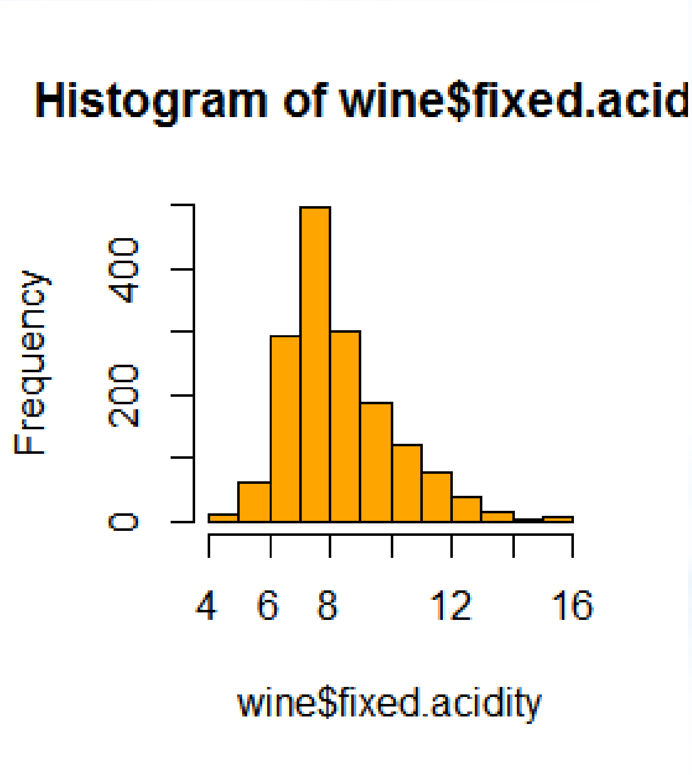


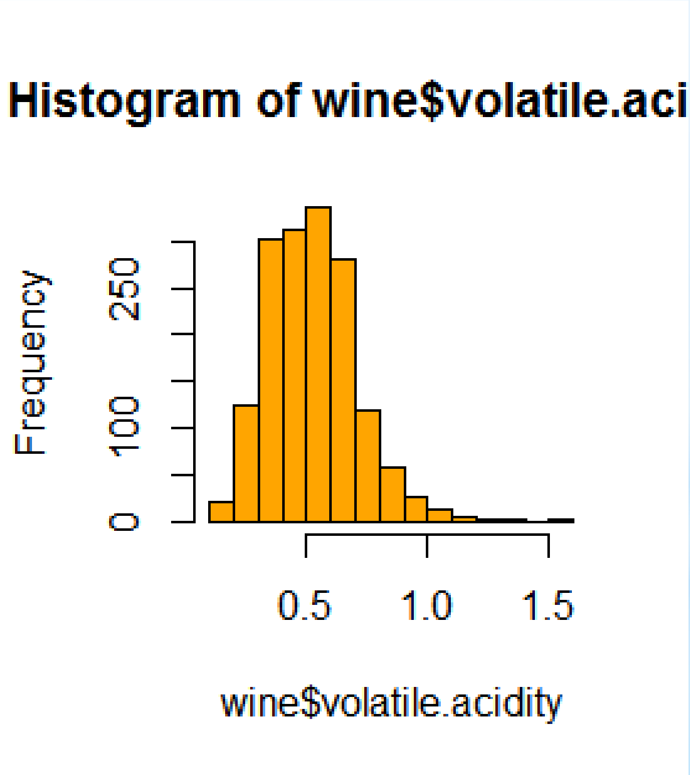


Analysis

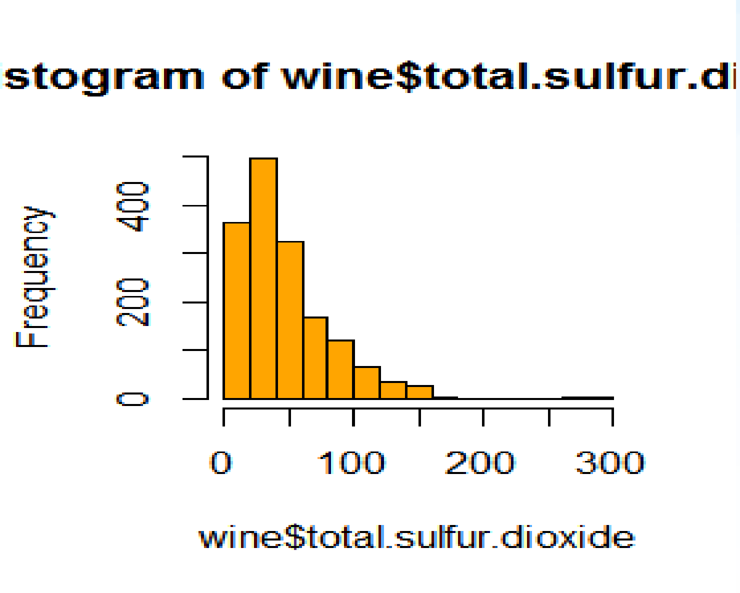
Distribution and Outliers

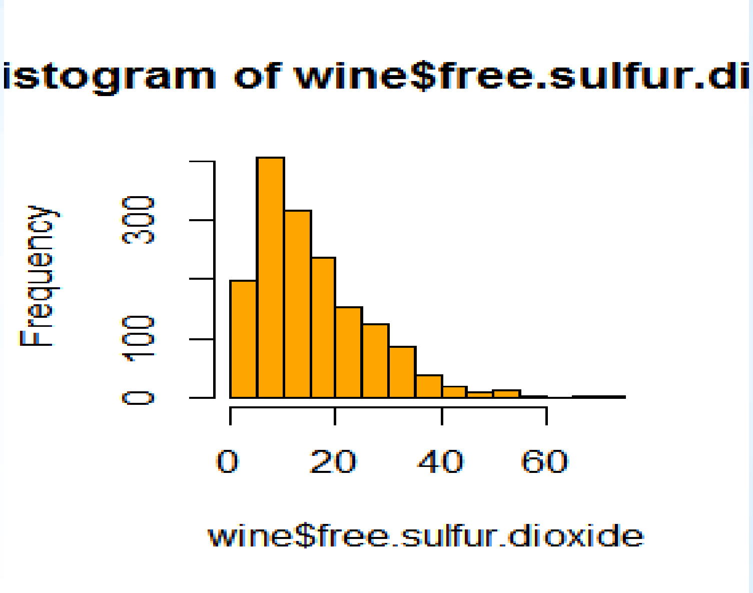
- Fixed and volatile acidity, total and free sulfur dioxides, alcohol and sulphate seem to be long-tailed for the outliers present.

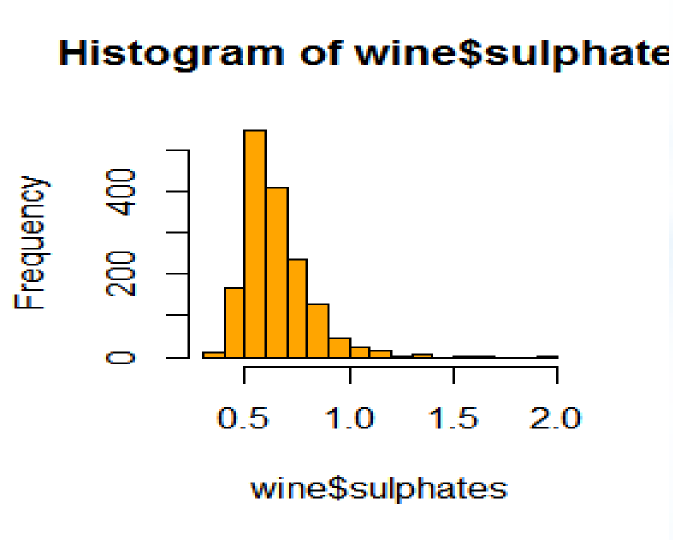


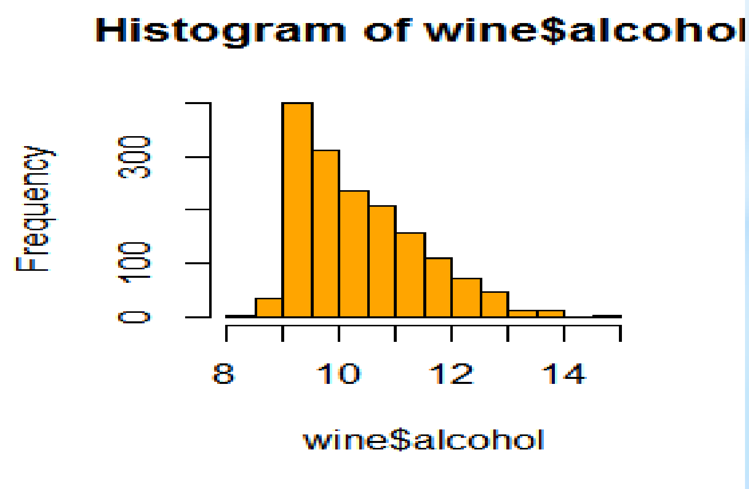


Analysis


stogram of wine$total.sulfur.di
Frequency
0
200
400
0
100
200
300
wine$total.sulfur.dioxide


istogram of wine$free.sulfur.di
Frequency
0
100
300
0
20
40
60
wine$free.sulfur.dioxide


Histogram of wine$sulphate
Frequency
0
200
400
0.5
1.0
1.5
2.0
wine$sulphates


Histogram of wine$alcohol
Frequency
0
100
300
8
10
12
14
wine$alcohol

Analysis

Checking co-relation of each variable against red wine quality

fixed.acidity	0.12405165
volatile.acidity	-0.39055778
citric.acid	0.22637251
residual.sugar	0.01373164
chlorides	-0.12890656
free.sulfur.dioxide	-0.05065606
total.sulfur.dioxide	-0.18510029
density	-0.17491923
pH	-0.05773139
sulphates	0.25139708
alcohol	0.47616632

- **Volatile acidity,** and **alcohol** have strong influence on quality
- Even **Sulphate** and **Citric acid** also has some good amount of co-relation with quality of red wine.

Analysis

Created Box plots between these variables to see if anything is missed from the correlation table.

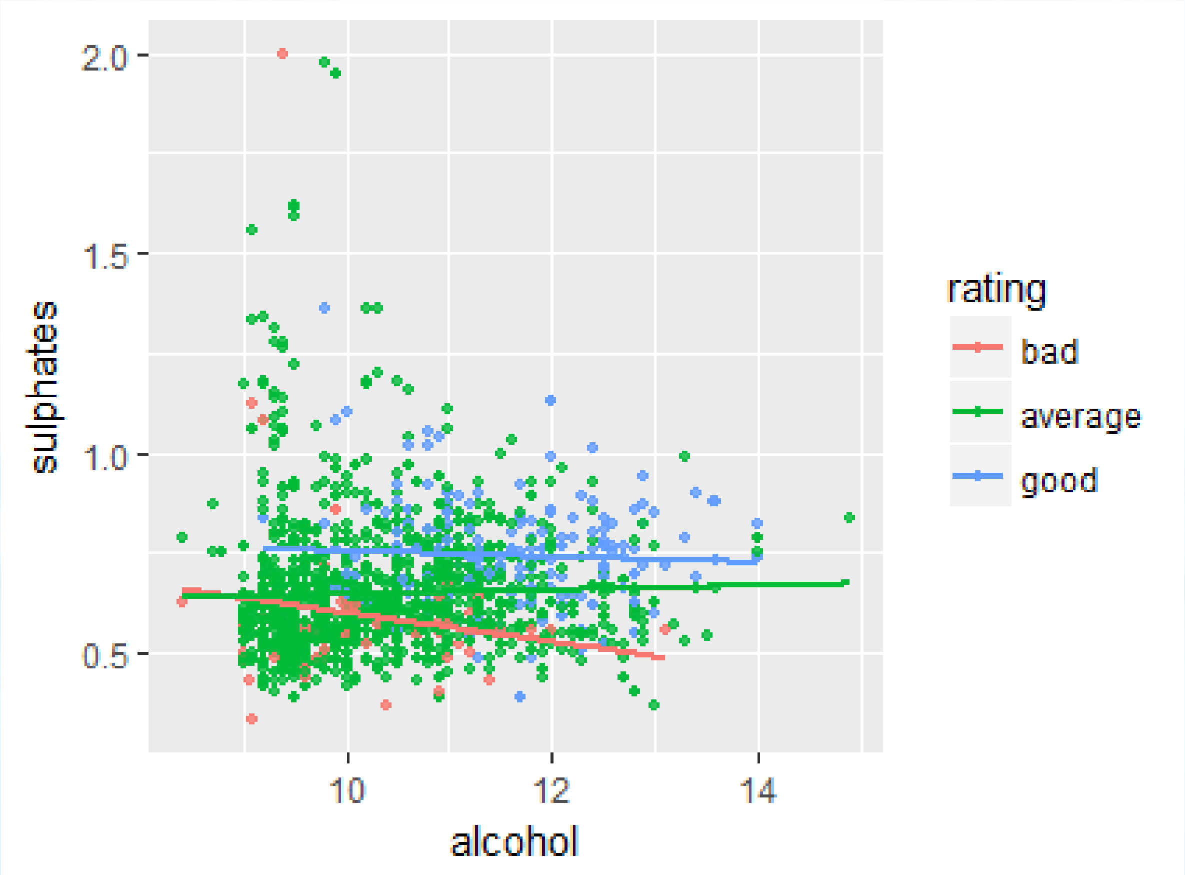
Observations:

- Fixed Acidity seems to have almost no effect on quality.
- Volatile Acidity seems to have a negative correlation with the quality.
- Citric acid seems to have a positive correlation with Wine Quality. Better wines have higher Citric Acid.
- Better wines seem to have higher alcohol percentages. On applying simple linear regression it seemed alcohol contributes about 22% change in quality of wine.
- Even though weakly correlated, it seems that lower percent of Chloride seems to produce better wines.
- Better wines seem to have lower densities. But then again, this may be due to the higher alcohol content in them.
- Residual sugar almost has no effect on the wine quality.
- Better wines seems to have lower pH level i.e. more acidic.
- It seems that better red wines have a stronger concentration of Sulphate.

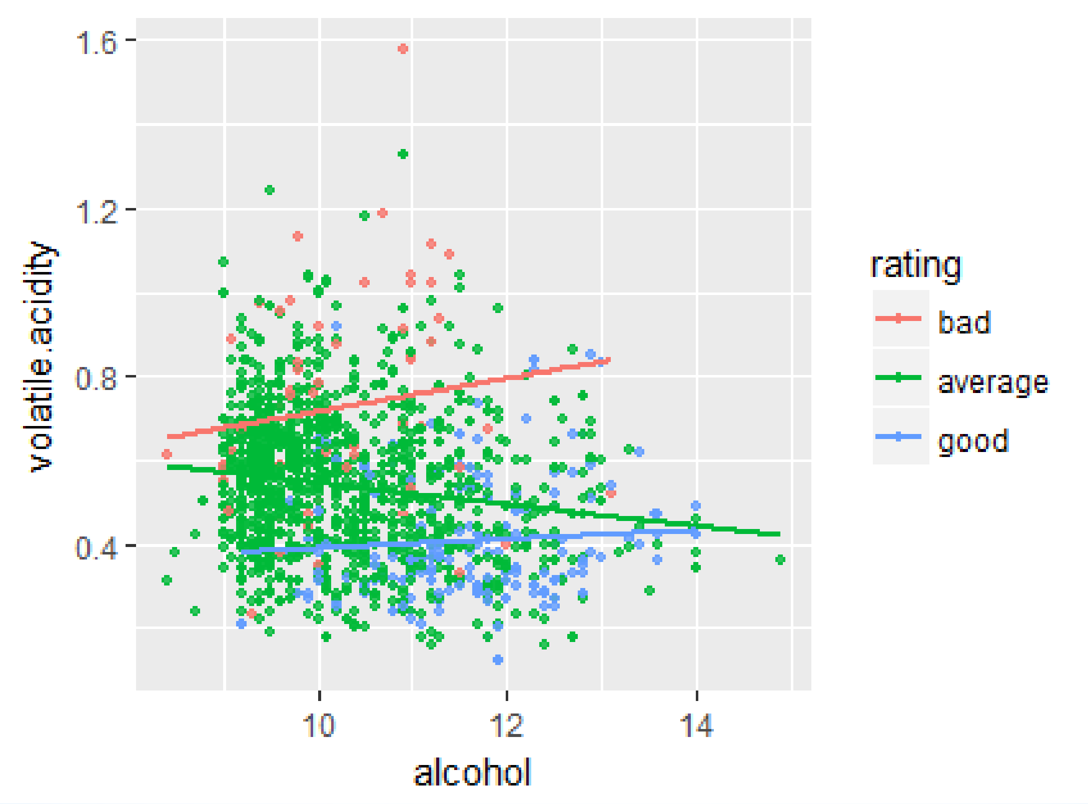
Analysis

We know alcohol is a major contributor but contributes to only 22% change in quality of wine.

I tried to plot couple of multivariate graphs to see if other strongly correlated variables influence change in quality of wine



Analysis



Observation:

- High alcohol with considerable high concentration of sulphates seem to produce better wine.
- High alcohol with considerable low concentration of volatile acidity seem to produce better wine.

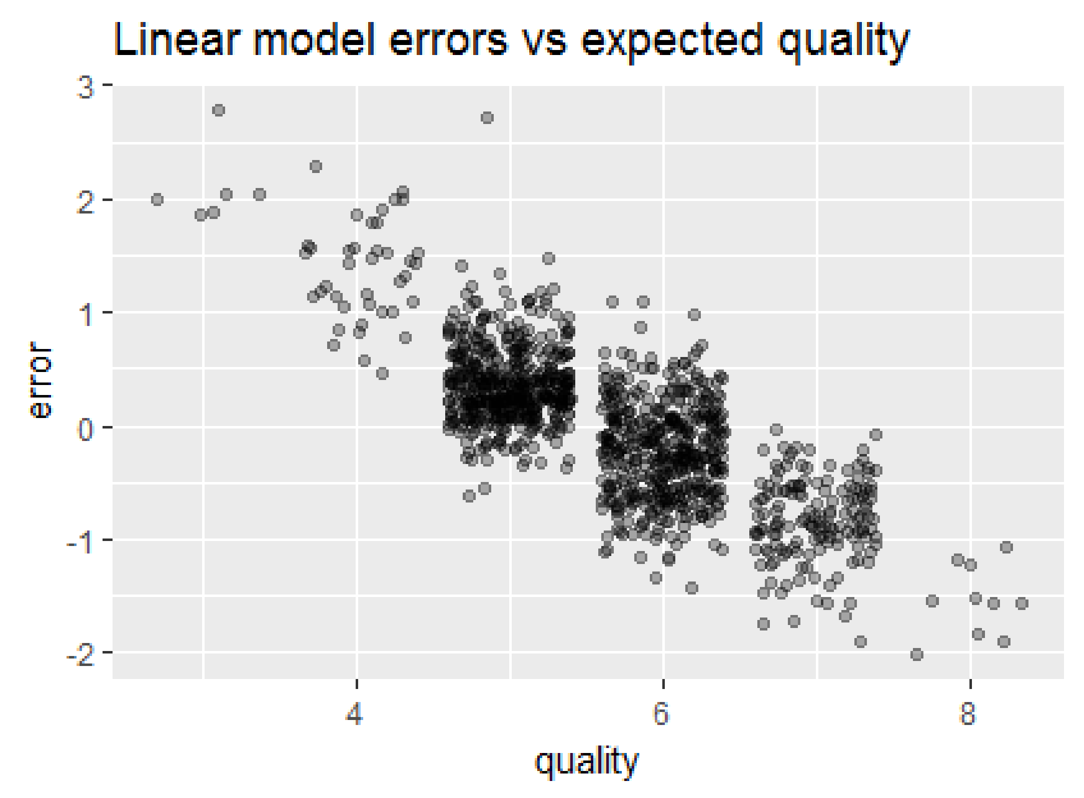
Linear Regression Model

Model Description:

- I took all the variables which are most strongly correlated with the quality of the wine and generated a linear model with them.
- I tried multiple models but observed a low confidence level and this can be due to the fact that our dataset comprised mainly of ‘Average’ quality wines and there were very few data about the ‘Good’ and the ‘Bad’ quality wines.
- The final linear regression model to predict target variable “quality” comprised of variables such as alcohol, citric acid, sulphates and acidity (fixed and volatile).
- However, due to low confidence level, the model could predict only 35% of change in quality of wine.

Linear Regression Model

- Here, we see that the error is much more dense in the 'Average' quality section than the 'Good' and the 'Bad' quality wines.
- This is evident from the fact that most of our dataset contains 'Average' quality wines and there is not too many data in the extreme ranges.



Linear Regression Model

Predicted VS Actual:

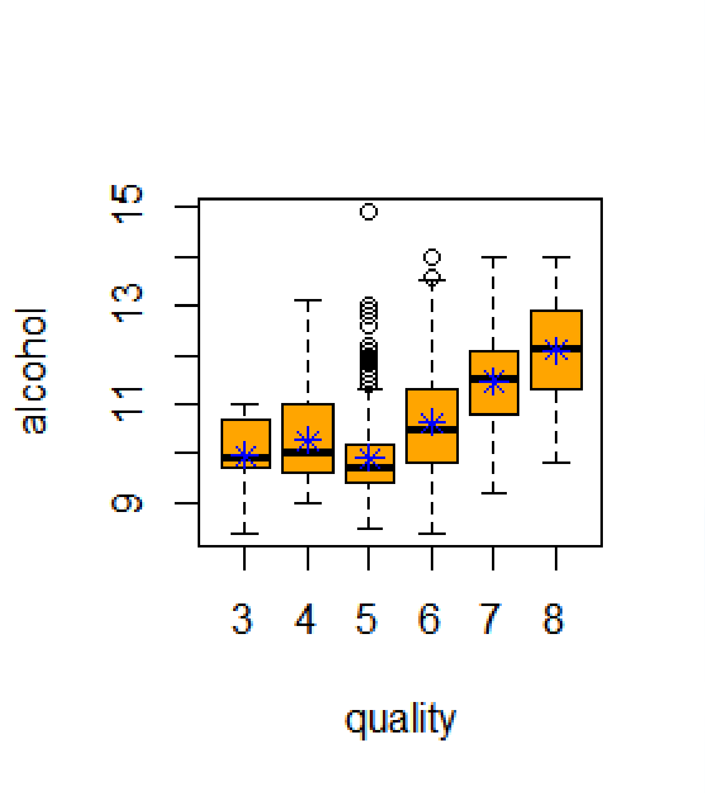
Based on my final model, I predicted wine quality of test data.

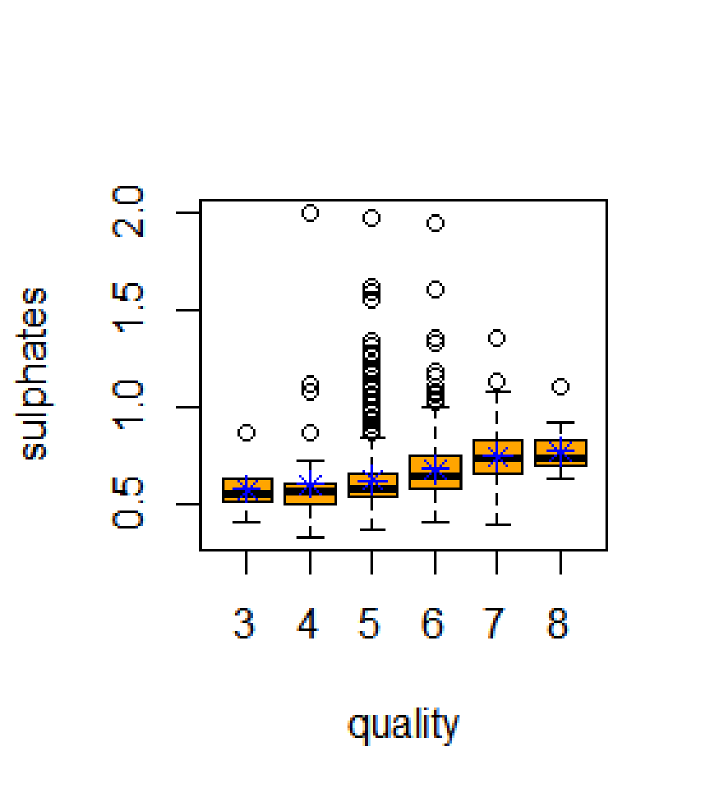
Below are the observations for the first 6 records of the test data:

#	***predicted***	***actual***	***error***
#6	**5.155359**	**5**	**0.1553588**
#7	**5.188282**	**5**	**0.1882817**
#11	**5.123794**	**5**	**0.1237940**
#14	**5.582424**	**5**	**0.5824239**
#15	**5.331526**	**5**	**0.3315263**
#20	**5.676508**	**6**	**-0.3234922**

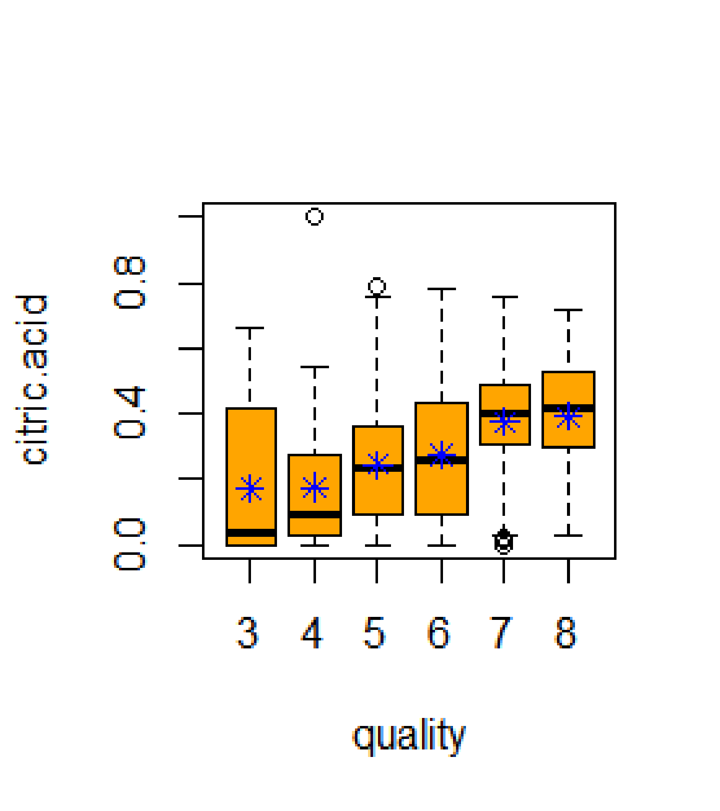
Summary / Implications

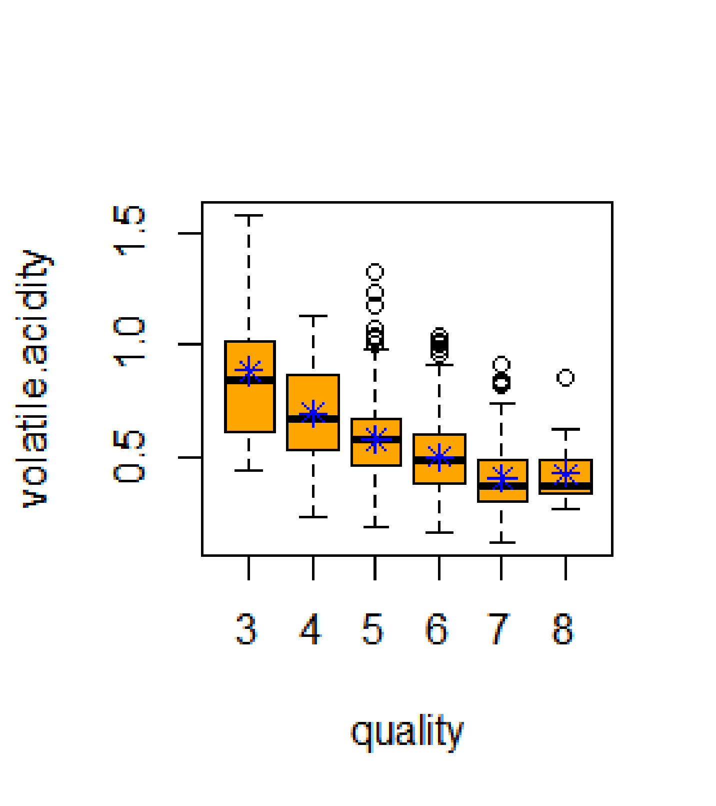
- High Alcohol, Acidity and Sulphate content played an important role in determining quality of wine.
- Citric Acid, even though weakly correlated plays a part in improving the wine quality.





Summary / Implications





Summary / Implications

- However, on this data set, the main struggle was to get a higher confidence level when predicting factors that are responsible for the production of different quality of wines especially the ‘Good’ and the ‘Bad’ ones.
- As the data was very centralized towards the ‘Average’ quality, my training set did not have enough data to accurately build a model which can predict the quality of a wine.
- So in future if I can get a dataset about Red Wines with more complete information then I can build my models more effectively.

Thank You