# Exploration of Red Wine Data by Mohamed Hassan

## Introduction

In this project we will explores the univariate, bivariate, & multivariate relationships between variables using Exploratory Data Analysis (EDA) techniques in R. To do so we are going to use a tidy data (http://vita.had.co.nz/papers/tidy-data.pdf) that is created - using red wine samples - on 2009 by P.cortez and al, the dataset is related to variants of the Portuguese "Vinho Verde" (https://en.wikipedia.org/wiki/Vinho_Verde) wine.

## About the Dataset

This dataset is public available for research. The details are described in [Cortez et al., 2009].

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. ISSN: 0167-9236.

Available at:

- Elsevier (http://dx.doi.org/10.1016/j.dss.2009.05.016)
- Pre-press (pdf) (http://www3.dsi.uminho.pt/pcortez/winequality09.pdf)
- bib (http://www3.dsi.uminho.pt/pcortez/dss09.bib)

## **Number of Instances**: Red Wine - 1599 obersvations

## **Number of Attributes**: 11 + output attribute

## Variable Description

- **Fixed acidity**: most acids involved with wine or fixed or nonvolatile (do not evaporate readily).
- **Volatile acidity**: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste.
- **Citric acid**: found in small quantities, citric acid can add 'freshness' and flavor to wines.
- **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.

- **Chlorides**: the amount of salt in the wine.
- **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.

- **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.

- **Density**: the density of water is close to that of water depending on the percent alcohol and sugar content.
- **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.
- **Sulphates**: a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant.
- **Alcohol**: the percent alcohol content of the wine.
- **Quality**: output variable (based on sensory data, score between 0 and 10).

## Variable Information

– **Input variables (based on physicochemical tests):**

- fixed acidity (tartaric acid - g / dm^3)
- volatile acidity (acetic acid - g / dm^3)
- citric acid (g / dm^3)
- residual sugar (g / dm^3)
- chlorides (sodium chloride - g / dm^3
- free sulfur dioxide (mg / dm^3)
- total sulfur dioxide (mg / dm^3)
- density (g / cm^3)
- pH
- sulphates (potassium sulphate - g / dm3)
- alcohol (% by volume)

– **Output variable (based on sensory data):**

- quality (score between 0 and 10)

## What property makes good red wine?

- In this project we try to answer this question by exploring the red wine dataset.

```
##     X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1   1           7.4             0.70        0.00            1.9     0.076
## 2   2           7.8             0.88        0.00            2.6     0.098
## 3   3           7.8             0.76        0.04            2.3     0.092
## 4   4          11.2             0.28        0.56            1.9     0.075
## 5   5           7.4             0.70        0.00            1.9     0.076
## 6   6           7.4             0.66        0.00            1.8     0.075
## 7   7           7.9             0.60        0.06            1.6     0.069
## 8   8           7.3             0.65        0.00            1.2     0.065
## 9   9           7.8             0.58        0.02            2.0     0.073
## 10 10           7.5             0.50        0.36            6.1     0.071
##    free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                   11                   34  0.9978 3.51      0.56     9.4
## 2                   25                   67  0.9968 3.20      0.68     9.8
## 3                   15                   54  0.9970 3.26      0.65     9.8
## 4                   17                   60  0.9980 3.16      0.58     9.8
## 5                   11                   34  0.9978 3.51      0.56     9.4
## 6                   13                   40  0.9978 3.51      0.56     9.4
## 7                   15                   59  0.9964 3.30      0.46     9.4
## 8                   15                   21  0.9946 3.39      0.47    10.0
## 9                    9                   18  0.9968 3.36      0.57     9.5
## 10                  17                  102  0.9978 3.35      0.80    10.5
##    quality
## 1        5
## 2        5
## 3        5
## 4        6
## 5        5
## 6        5
## 7        5
## 8        7
## 9        7
## 10       5
```

## Feature Names and Summary Statistics

Let’s run some basic functions to examine the structure and schema of the dataset.

```
## 'data.frame':    1599 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.
5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.06
5 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.3
5 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57
0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

This red wine dataset contains 1,599 obersvations with 13 variables on the chemical properties of the wine.

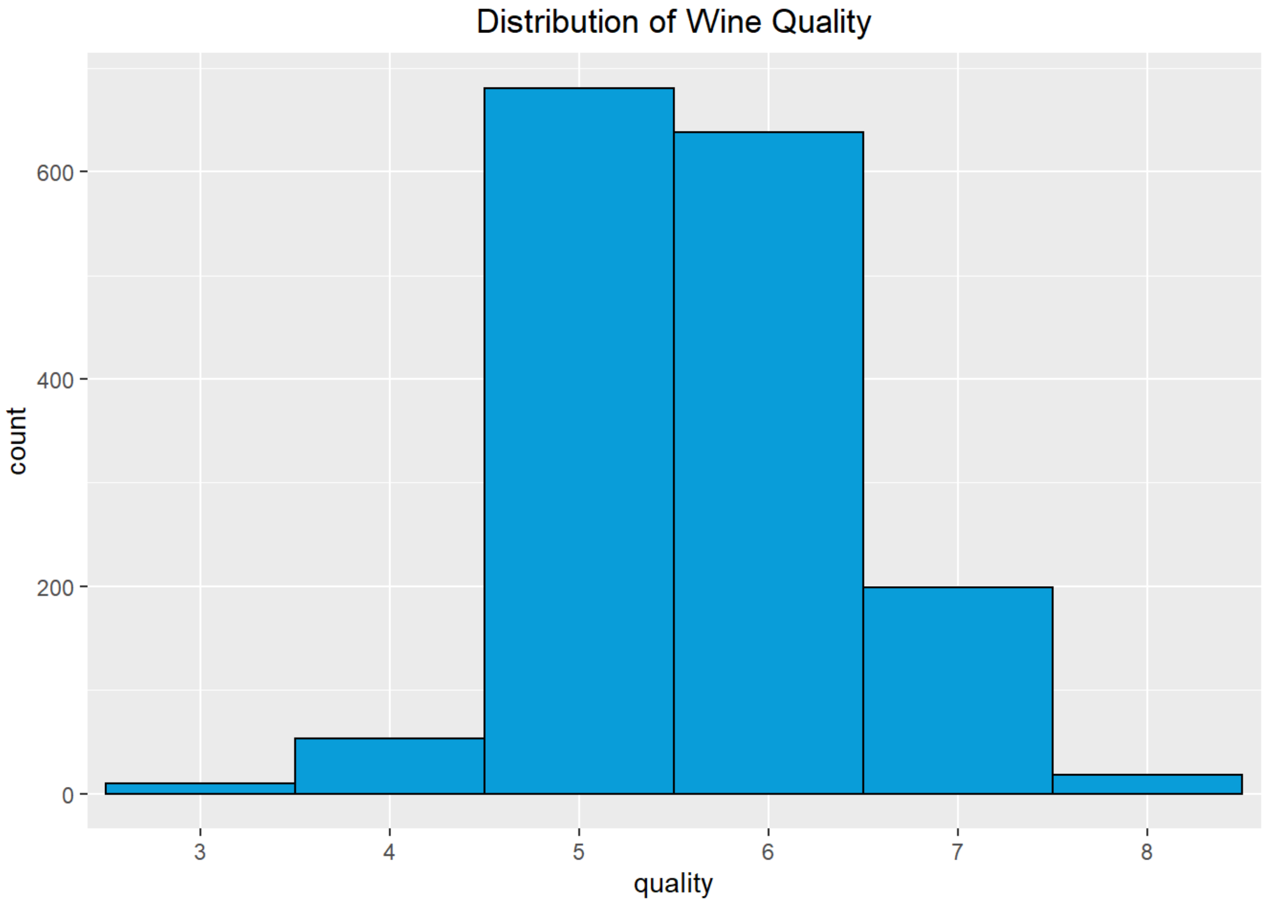
```
##        X          fixed.acidity   volatile.acidity  citric.acid
##  Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000
##  1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090
##  Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260
##  Mean   : 800.0   Mean   : 8.32   Mean   :0.5278   Mean   :0.271
##  3rd Qu.:1199.5   3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420
##  Max.   :1599.0   Max.   :15.90   Max.   :1.5800   Max.   :1.000
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.900   Min.   :0.01200   Min.   : 1.00
##  1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00
##  Median : 2.200   Median :0.07900   Median :14.00
##  Mean   : 2.539   Mean   :0.08747   Mean   :15.87
##  3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00
##  Max.   :15.500   Max.   :0.61100   Max.   :72.00
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  6.00       Min.   :0.9901   Min.   :2.740   Min.   :0.3300
##  1st Qu.: 22.00       1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500
##  Median : 38.00       Median :0.9968   Median :3.310   Median :0.6200
##  Mean   : 46.47       Mean   :0.9967   Mean   :3.311   Mean   :0.6581
##  3rd Qu.: 62.00       3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300
##  Max.   :289.00       Max.   :1.0037   Max.   :4.010   Max.   :2.0000
##     alcohol         quality
##  Min.   : 8.40   Min.   :3.000
##  1st Qu.: 9.50   1st Qu.:5.000
##  Median :10.20   Median :6.000
##  Mean   :10.42   Mean   :5.636
##  3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :14.90   Max.   :8.000
```

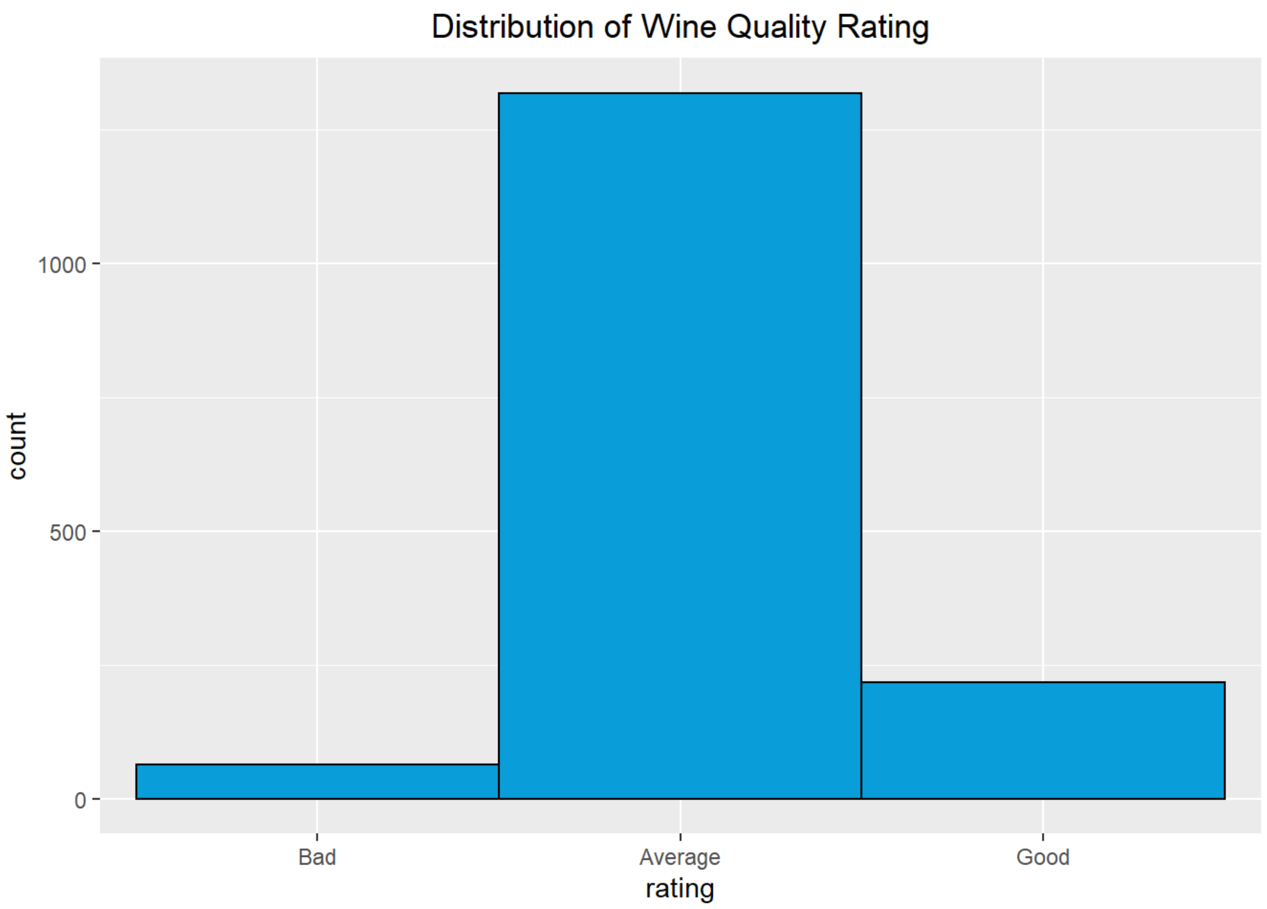
## Univariate Plots Section

In order to catch the meaning of each variable in the dataset, in this section we will seek to explore each attribute individually in what is known as univariate analysis. The most relevant attributes of this dataset are consolidated at the end of this section.

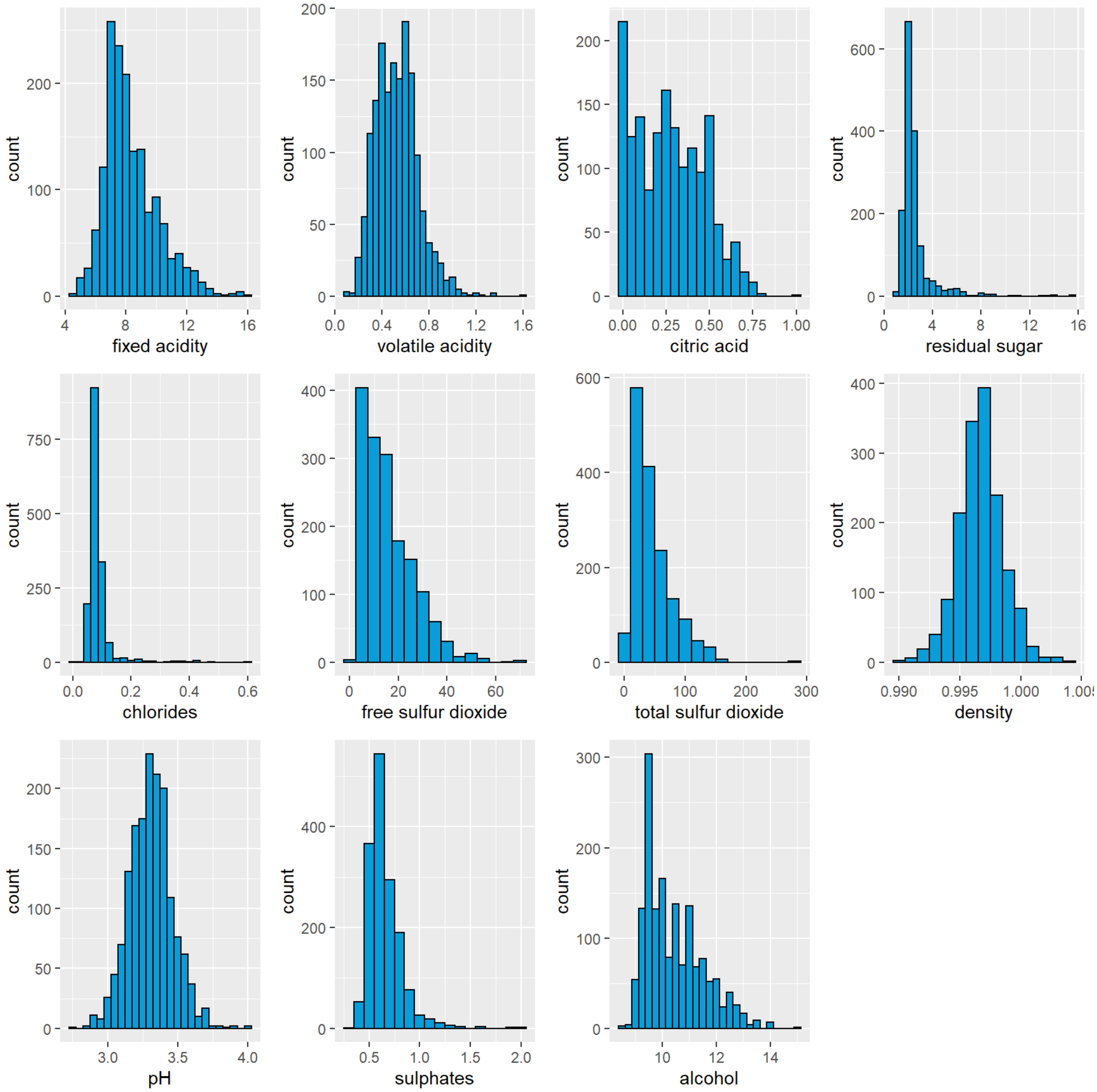
## Quality Distribution

The wine quality grade is a discrete number. from summary statistics and plot the distribution of red wine qualityis ranged from 3 to 8. The median value is at 6.In this sample, there were no evaluations lower than 3 or greater than 8. Where, 3.93% were evaluated with low scores (below 5), 82.50% were evaluated with average scores (5 and 6) and 13.57% were evaluated with high scores (above 6).


Distribution of Wine Quality
count
0
200
400
600
quality
3
4
5
6
7
8



## Distribution of Other Chemical Properties

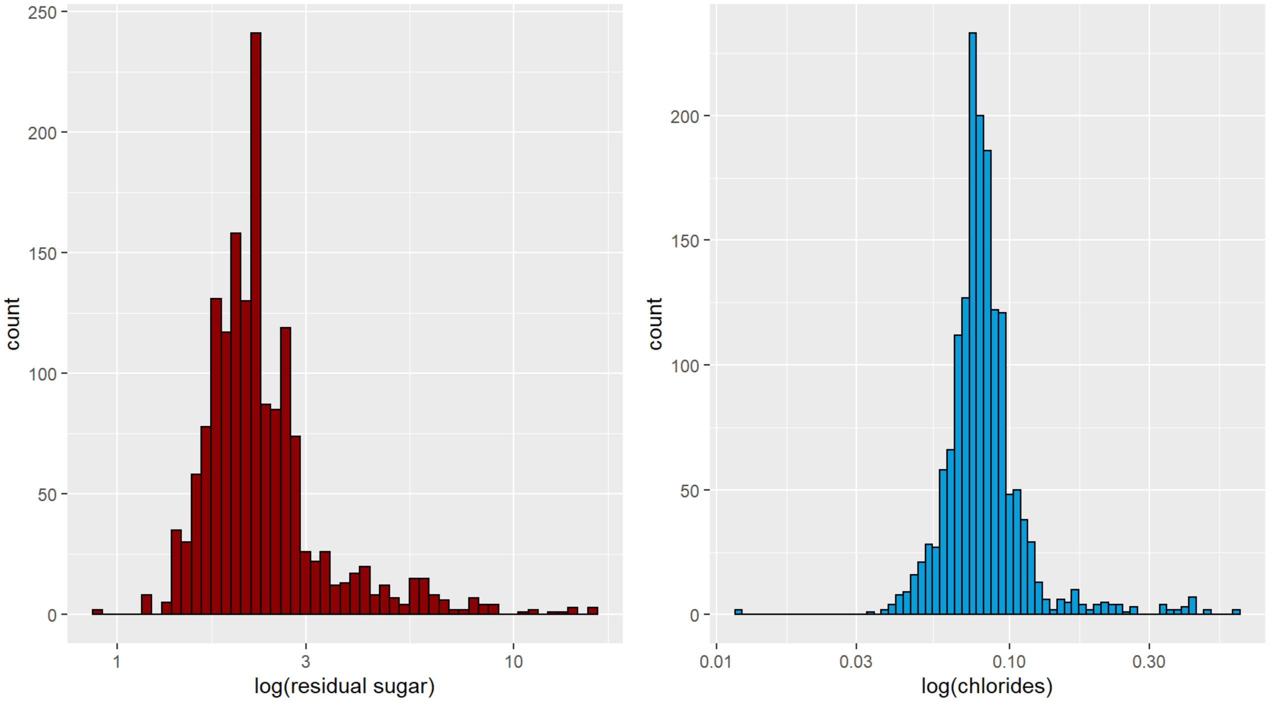


- The distribution of **Fixed Acidity** is positively skewed. The median is 7.9 g/dm³ with high concentration of wines with Fixed Acidity but due to some outliers, the mean has been dragged to 8.32 g/dm³.
- The distribution of **Volatile Acidity** looks like Bimodal with two peaks around 0.4 and 0.6 g/dm³.
- **Citric acid** has different characteristics which have a multimodal distribution, ranging from 0 g/dm³ to 1 g/dm³, with a peak at 0 g/dm³, another at 0.24 g/dm³ and one at 0.48 g/dm³.
- The distribution of **Residual Sugar** is again positively skewed with high peaks at around 2.2 g/dm³ with many outliers present at the higher ranges.
- For **Chlorides** also, we see a similar distribution like Residual Sugar.
- For **Free Sulphur Dioxide**, there is a high peak at 7 mg/dm³ but then it again follows the same

positively skewed long tailed patterns with some outliers in the high range.

- **Total Sulphur Dioxide** also follows a similar pattern of Free Sulphur Dioxide.
- For the **Density** variable, we see something new for the first time. This Variable has almost a perfect Normal Distribution.
- **pH** also has a very Normally distributed shape.
- **Sulphates** also exhibit a similar long tailed distribution like Chlorides or Free/Total Sulphur Dioxide. It has relatively less outliers.
- **Alcohol** also follows a skewed distribution but here the skewness is less than that of Chlorides or Residual Sugars.

**I'm going to try a transform residual sugar and chlorides to have better visualizations of the data**



## Univariate Analysis

### – What is the structure of your dataset?

In this dataset, there are 1599 rows wines with 11 chemical properties (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides,free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol) and quality which varying from 0 (very bad) to 10 (very excellent)as variables. I added a new column called 'rating'. Quality is categorical variable and others are numerical variables which reflect the physical and chemical properties of the wine. Most of the wines in this dataset belong to the 'average' quality with very few 'low' and 'high' quality wines. It may be challenging to build a predictive model as I don't have enough data for the high and low quality wines.

**Other observations:**

- The majority of wines were classified and ‘medium quality’ (between 4 to 7 out of ten);
- Wines are all very similar in terms of density;
- There’s a wide range of alcohol percentage, from 8% to 14%.

– **What is/are the main feature(s) of interest in your dataset?**

Quality is the main feature of interest. The objective of the analysis is to determine the features that influence wine quality the most, and then building a predictive model of quality using these variables.

– **What other features in the dataset do you think will help support your investigation into your feature(s) of interest?**

In this first moment of univariate analysis, it is difficult to identify an independent variable that relates well to quality. Perhaps the alcohol concentration along with blends of other variables can be used to determine quality of wine. Acidity (fixed, volatile or citric) values may change the quality of wine. pH may also have some effect on the quality. I like to see if pH is affected by the different acids present in the wine. Residual Sugar may have an effect on the wine quality as more or less sugar can change the taste of the wine.

– **Did you create any new variables from existing variables in the dataset?**

I have created rating as my new variables from wine quality.A rating was also created for the quality, low (below 5), average (5 and 6) and high (above 6).

– **Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?**

1. The following variables can be better interpreted after a logarithmic transformation because they are right skewed and the transformations allowed better visualizations of the data :

- Residual sugar;
- Chlorides.

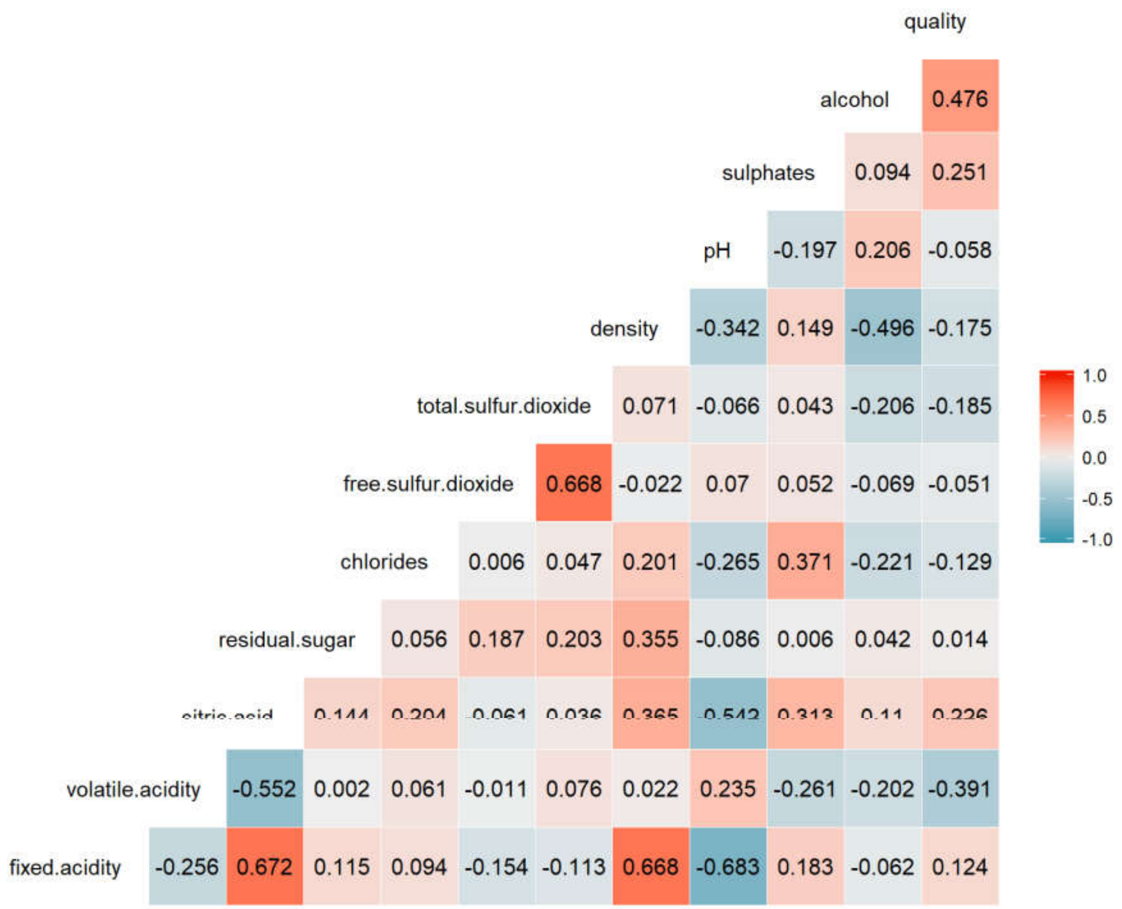
   They all look closer to a normal distribution after a logarithmic transformation.

2. The distribution of citric acid presented two unusual peaks which standed out of an otherwise normal distribution.

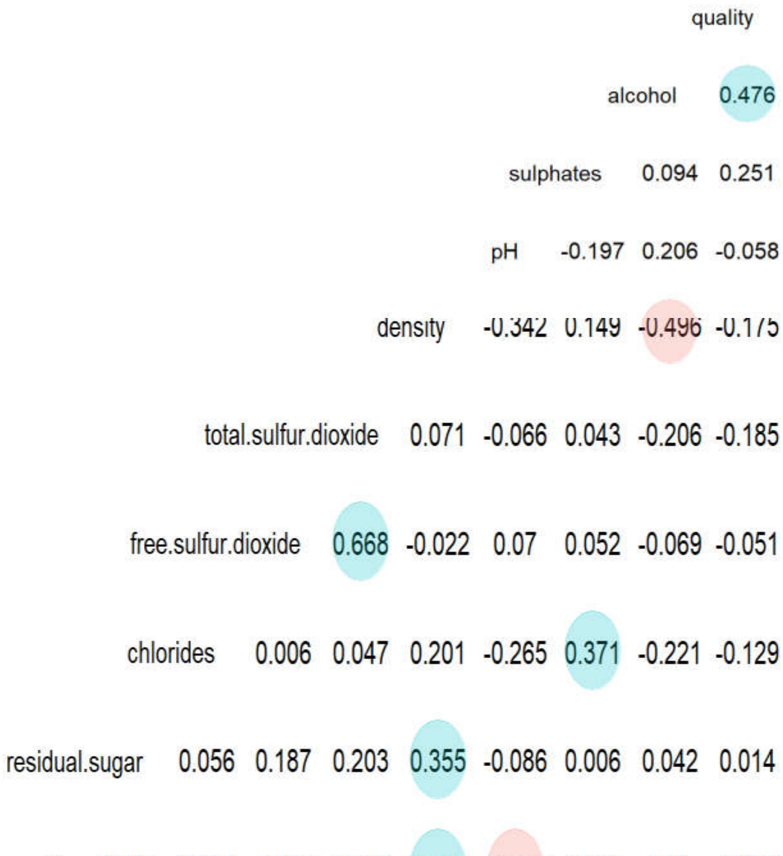
## Bivariate Analysis

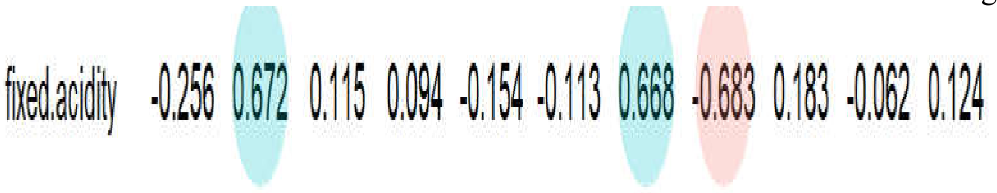
In this section, I will start exploring the relationships among all variables in our dataset, directing the analysis to find out which attributes are mainly related to our target variable: the quality score.

Correlation Graph Between Variables



Strong Correlation between Between Variables





**This shows the variables that correlate most highly with quality. These are:**

**1- Positive Correlation**

- alcohol
- sulphates
- citric.acid

**2- Negative Correlation**

- volatile.acidity
- total.sulphur.dioxide (weak negative)
- density(weak negative)
- pH (weak negative)

I also see that alcohol which is highly correlated with quality, has negative correlation with volatile.acidity

**Besides these, other properties that have strong correlation between them are:**

**1- Strong Positive Correlation**

- Density x Fixed Acidity (r = 0.67)
- Fixed Acidity x Citric Acid (r = 0.67)
- Free Sulfur Dioxide x Total Sulfur Dioxide (r = 0.67)
- pH x Volatile Acidity (r = 0.23) (We know that when decreasing pH the acidity increases)

**2- Strong Negative Correlation**

- pH x Fixed Acidity (r = -0.68)
- Citric Acid x Volatile Acidity (r = -0.55)
- pH x Citric acid (r = -0.54)
- Alcohol x Density (r = -0.50)

**From the correlation matrix we notice the only relevant attributes to quality are:**

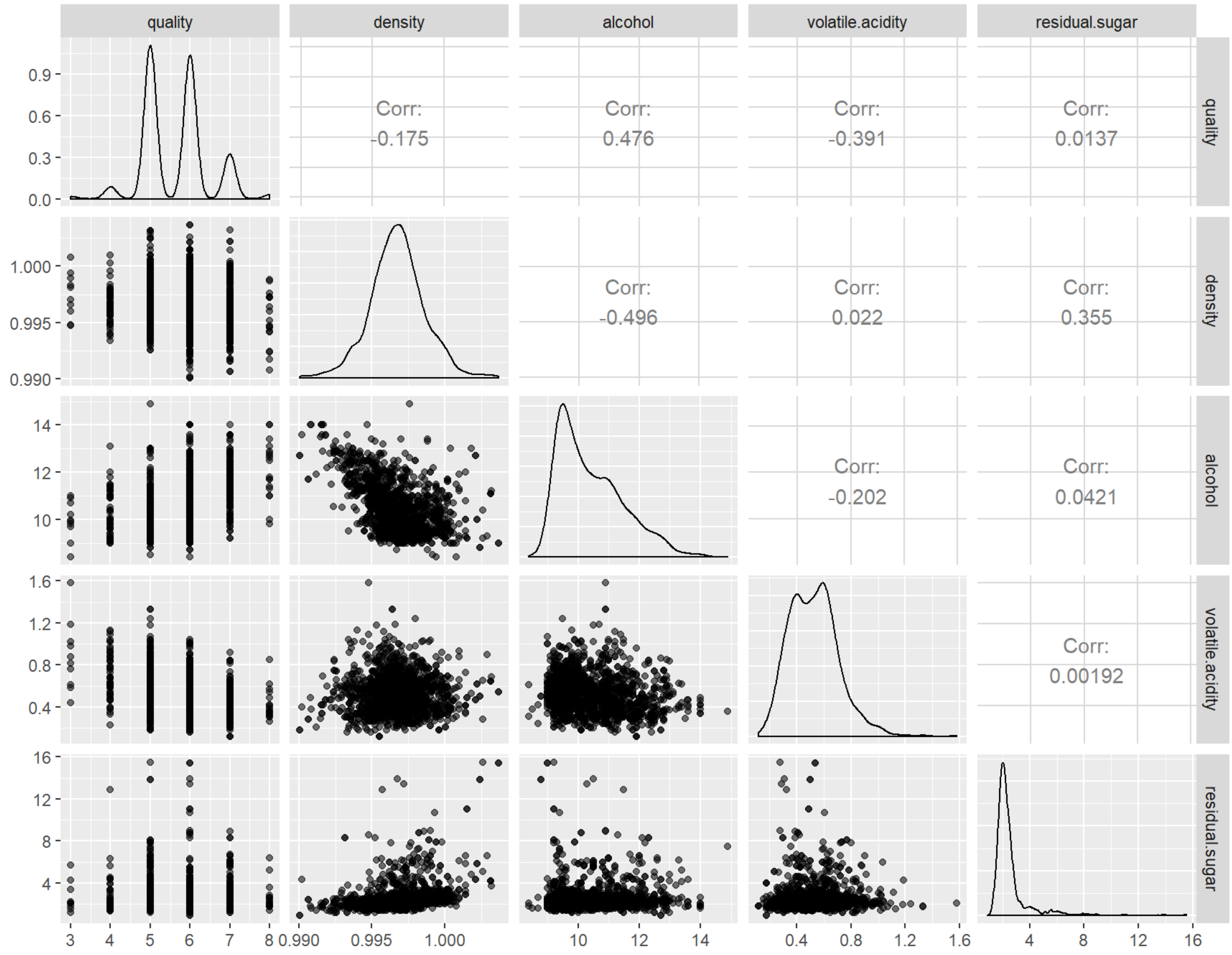
i. The percentage of alcohol and
ii. The volatile acidity.

There are some relevant correlation among other variables as well, and even though they don’t contribute explicitly to the target variable, it will be interesting to take a look on them trying to find out some new information or insight.

In order to select which of them we will work with, we should analyze the matrix above considering the previous univariate analysis for each attribute. In this case, we can draw the following assumptions:

- The wine quality is positively correlated to alcohol and negatively correlated to the volatile acidity. It does make sence with our previous findings, since the volatile acidity is responsible for an unpleasent taste in the wine.
- It is quite obvious that the pH attribute is negatively correlated to acidity, since the lower values in the pH scale means a higher acidity. Since it is a well-known fact, there is no sense to explore these relationships here.
- There seems to be some correlation between sulphates and chlorides. However, none of them presents a relevant correlation with quality. Furthermore, from the foregoing description it is possible to infer that both attributes are derived from the same mineral, so that we will not consider these attributes in the next analysis.
- It is interesting to observe there are some relationship between density and other attributes such as alcohol, residual sugar, citric acid, and fixed acidity.
- It also seems to be evident that the free and total forms of sulfur dioxide will be related to each other. The same stands for the different acidity types. Regarding this last, in order to simplify this analysis, we will now consider only the volatile acidity due to its impact on quality levels.

**Based on this, our next task is to create some pairplots relating all these chosen attributes:**



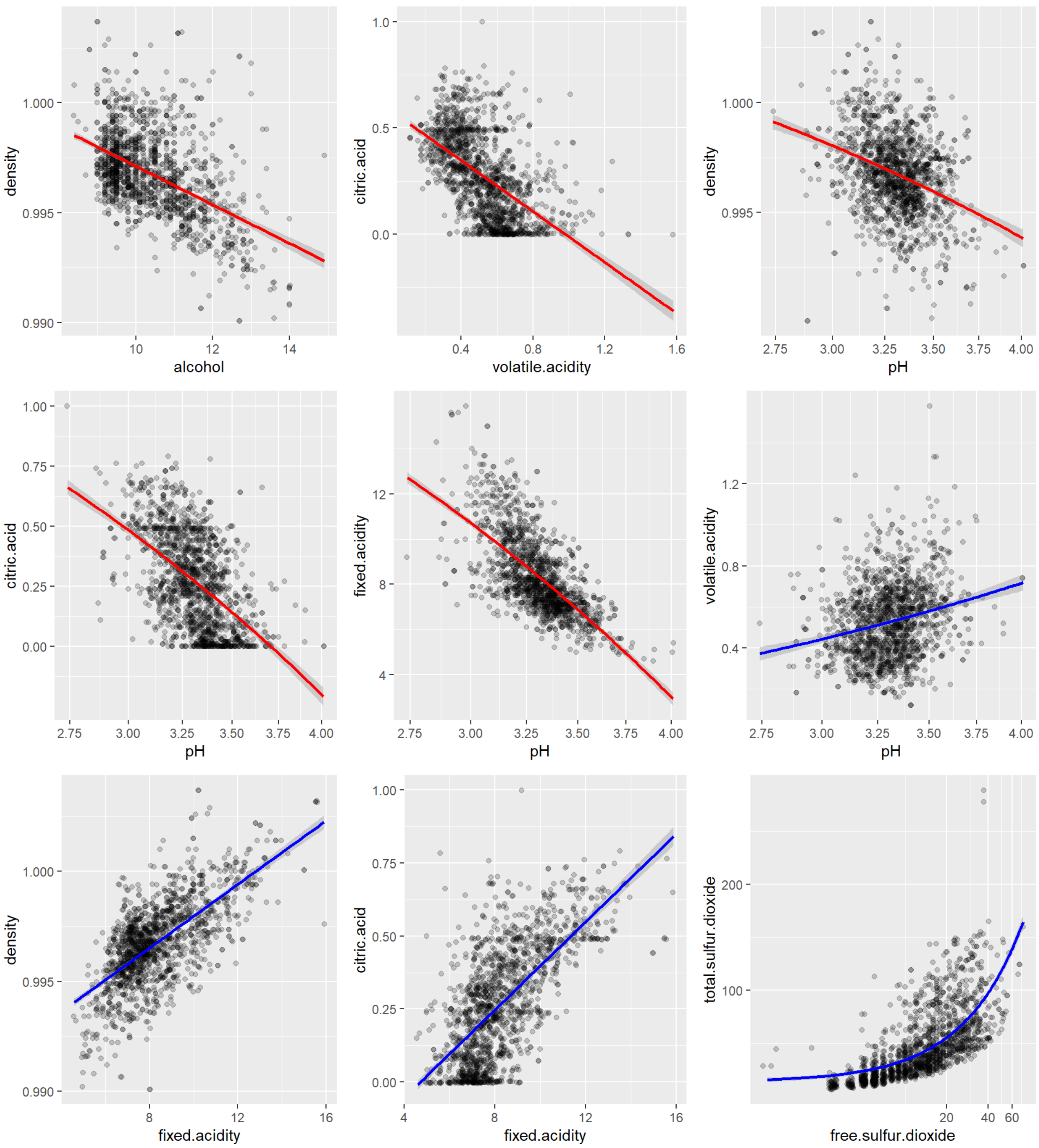
The pairplots above summarize most of the information we have gotten until now, as for example the unbalanced distribution of quality – ie, there are more wines scored as 5 or 6 than those below and above these values.

## Bivariate Analysis

– **Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?**

- Correlation plot helped to understand the correlation among different features.
- Quality is strongly correlated positively with alcohol ,sulfates and citric acid , and negatively with volatile acidity. Good wines have lower pH values, which also goes with having more fixed and citric acid and density.

– **Did you observe any interesting relationships between the other features (not the main feature (s) of interest)?**



- Citric acid and fixed acidity have a strong positive correlation of 0.67, while citrict acid and volatile acidity have a moderate negative correlation of -0.55
- Density and fixed acidity are two features with strong positive correlation of 0.67
- Negative correlation between alcohol and density
- An expected strong negative correlation between pH and fixed and citric acid
- A surprising positive correlation between pH and volatile acidity, since a higher pH value means less acidity, but a higher volatile acidity means more acidity.

**– What was the strongest relationship you found?**

```
##
##  Pearson's product-moment correlation
##
## data:  redwine$volatile.acidity and redwine$citric.acid
## t = -26.489, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5856550 -0.5174902
## sample estimates:
##        cor
## -0.5524957
```

```
##
##  Pearson's product-moment correlation
##
## data:  redwine$fixed.acidity and redwine$pH
## t = -37.366, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7082857 -0.6559174
## sample estimates:
##        cor
## -0.6829782
```

```
##
##  Pearson's product-moment correlation
##
## data:  redwine$free.sulfur.dioxide and redwine$total.sulfur.dioxide
## t = 35.84, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6395786 0.6939740
## sample estimates:
##       cor
## 0.6676665
```

We can say that the following variables have relatively higher correlations to wine quality:

- **Positively**

1. alcohol
2. sulphates
3. citric acid

- **Negatively**

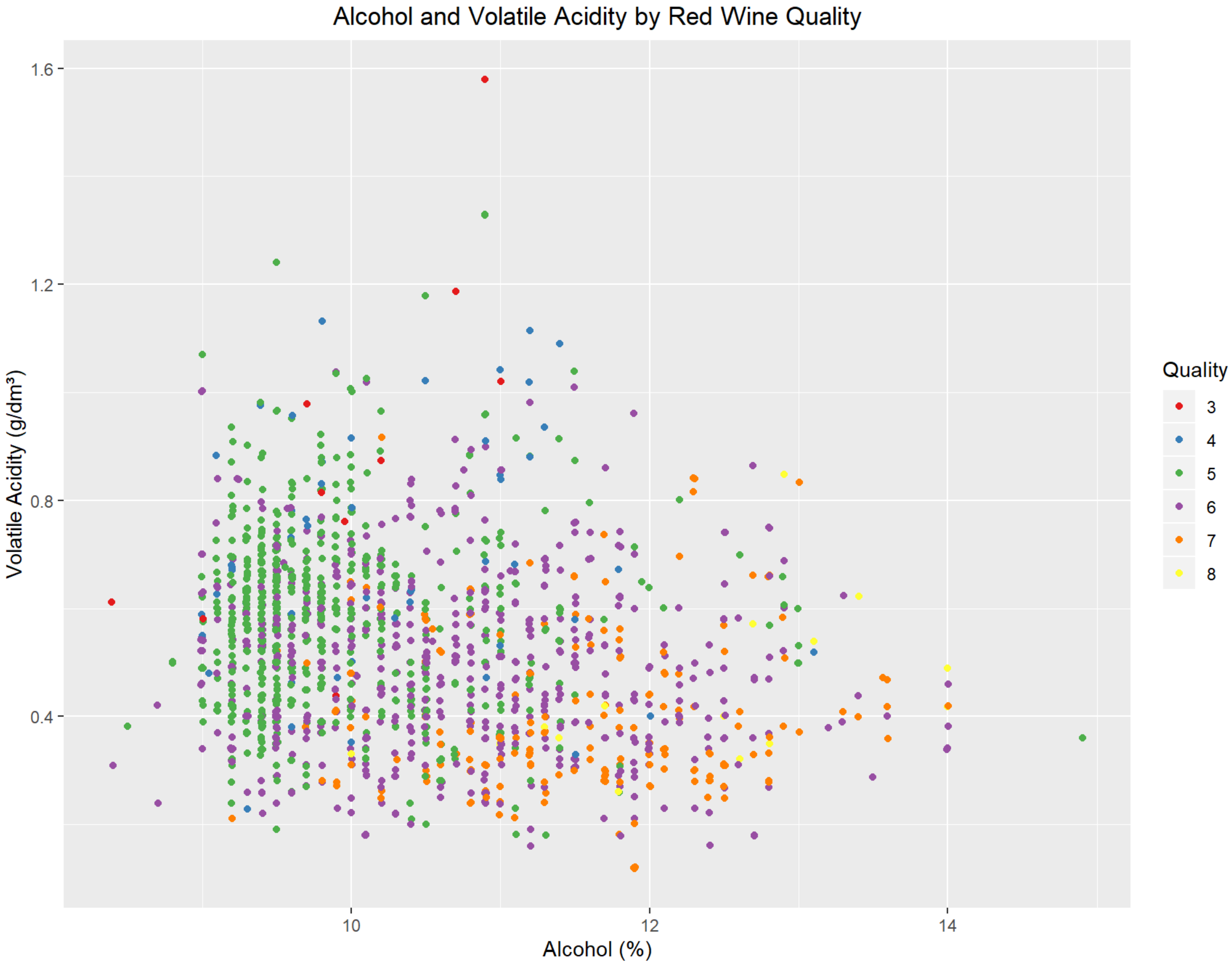
1. volatile acidity

Between other features, the strongest relationship appeared to be pH and fixed acidity, which had a negative correlation of -0.683 also between volatile acidity and citric acid(-0.55) and for positive correlation between free sulfur dioxide and total sulfur dioxide (0.667).

## Multivariate Plots Section

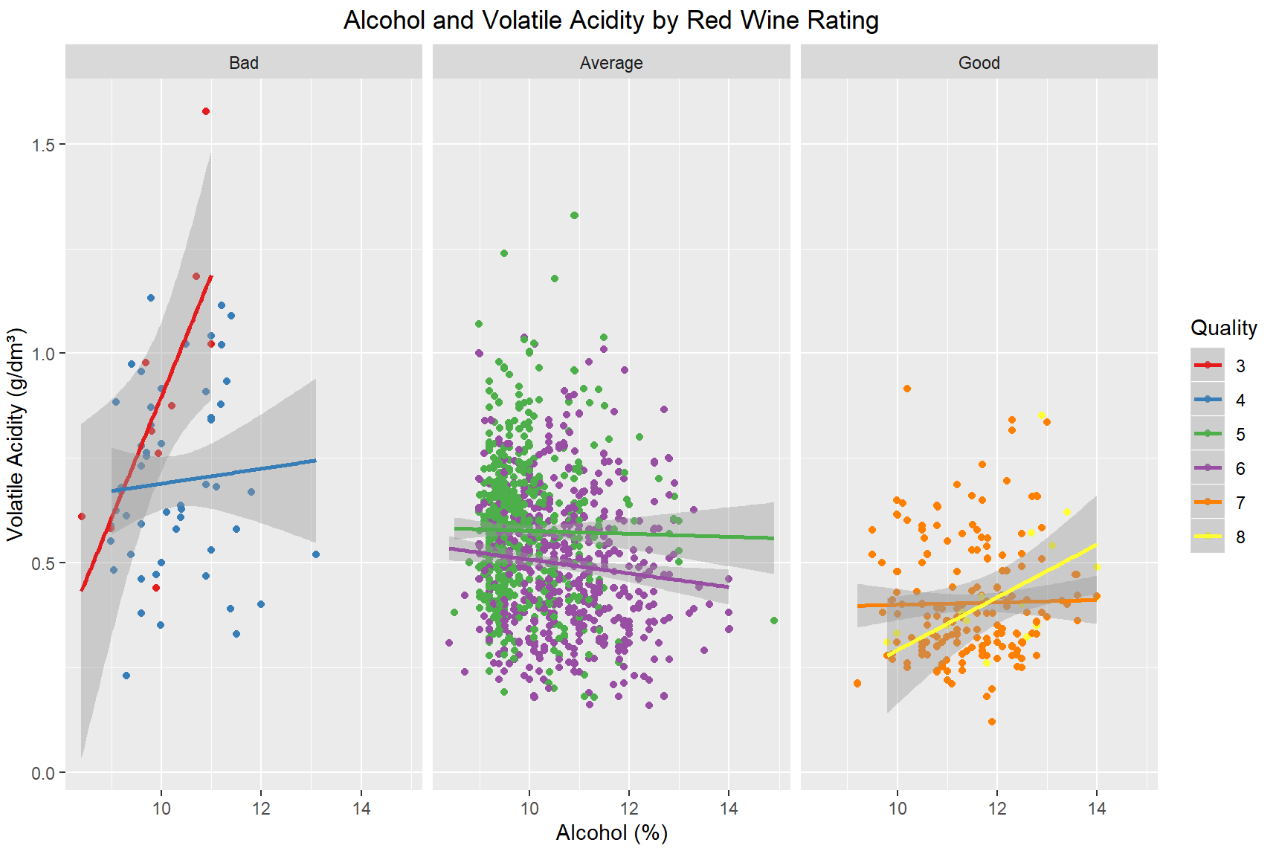
### Main Chemical Property vs Wine Quality

With different colors, we can add another dimension into the plot. There are 4 main features.Alcohol, volatile acidity are the top two factor that affect wine quality.

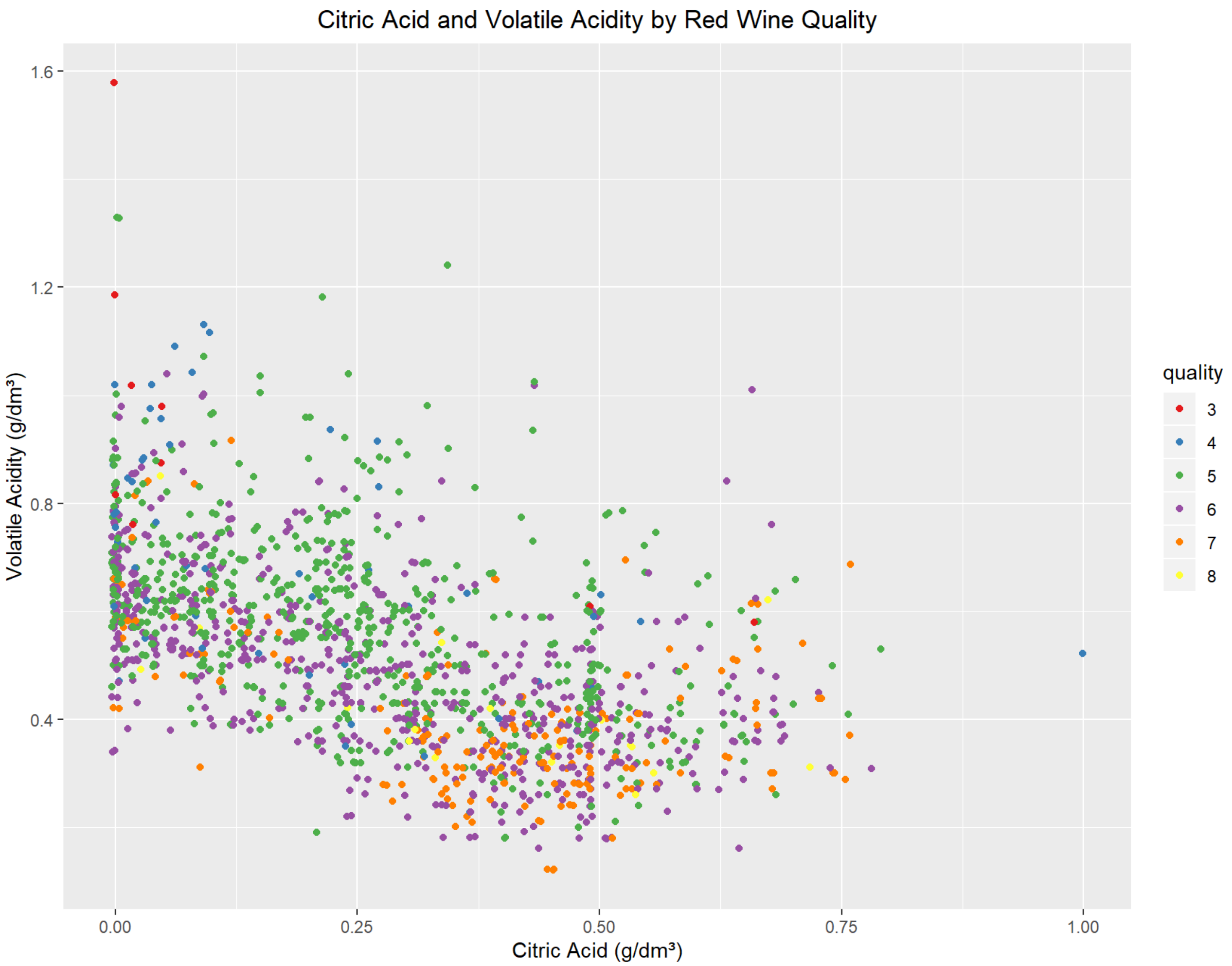


The figure looks over ploted, since the wine quality are discrete numbers. We can use jitter plot to alleviate this problem

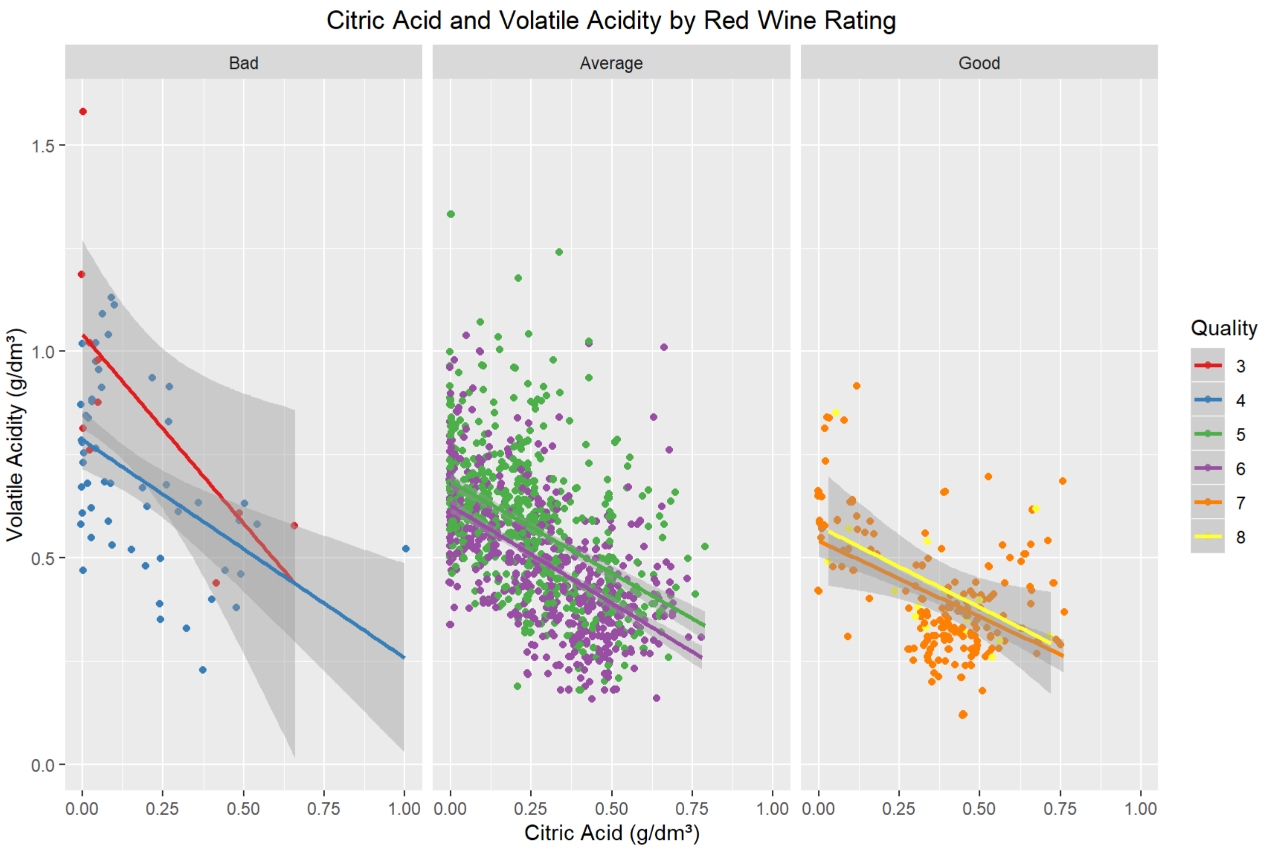
Now let’s look at the two variables with the strongest correlations with quality plotted against each other and colored by quality. I used a color scheme which accentuates the lowest and highest rated wines to help clarify the relationships present. While there are some exceptions, it is easy to see two main regions: the lowest quality wines tended to have lower alcohol percentages and higher volatile acidity concentrations, while the higher quality wines had higher alcohol percentages and lower volatile acidity concentrations, in general.



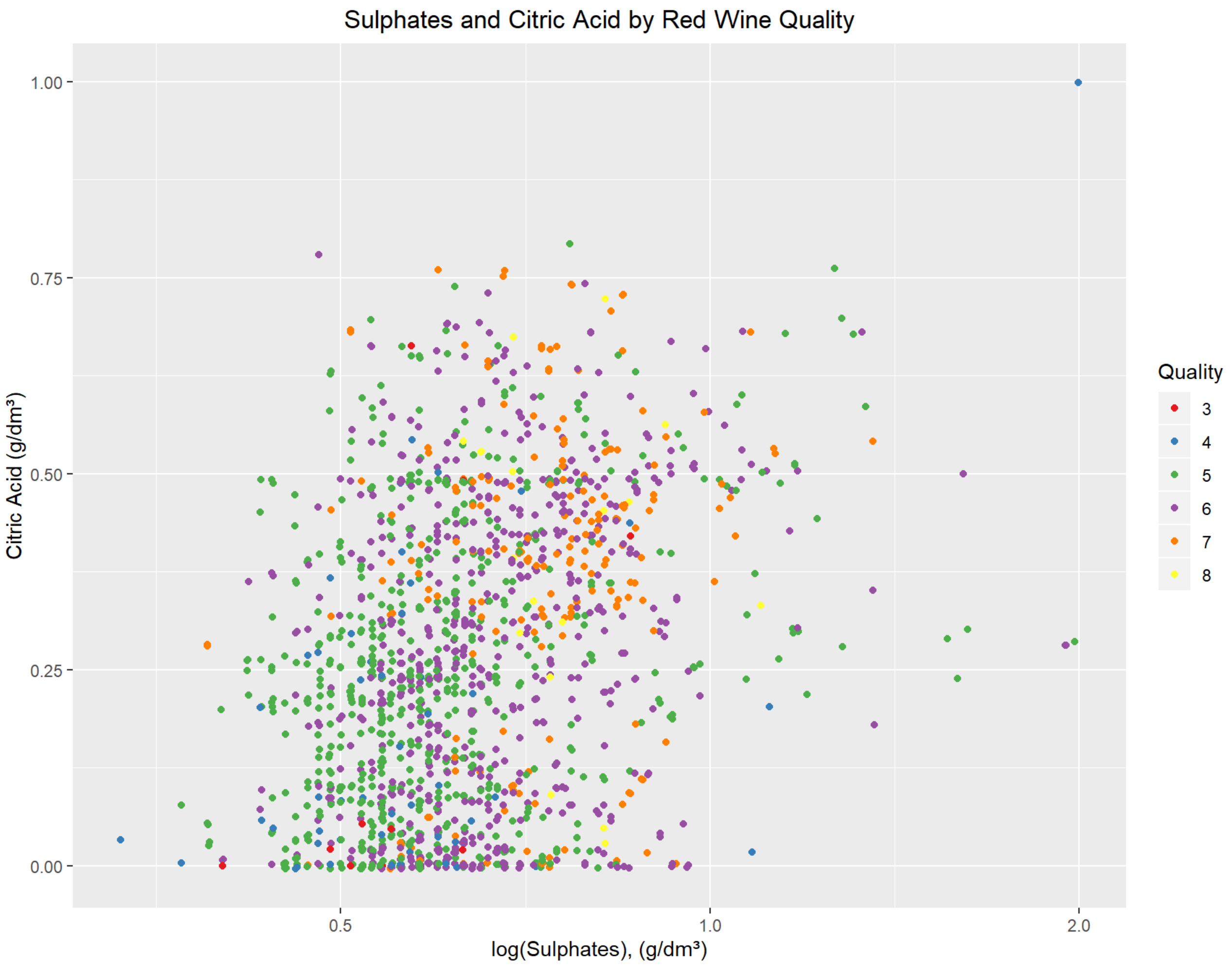
We can see higher quality wine have higher alcohol and lower volatile acidity.so, alchol and volatile acidity’s effect on wine qaulity can be observed.



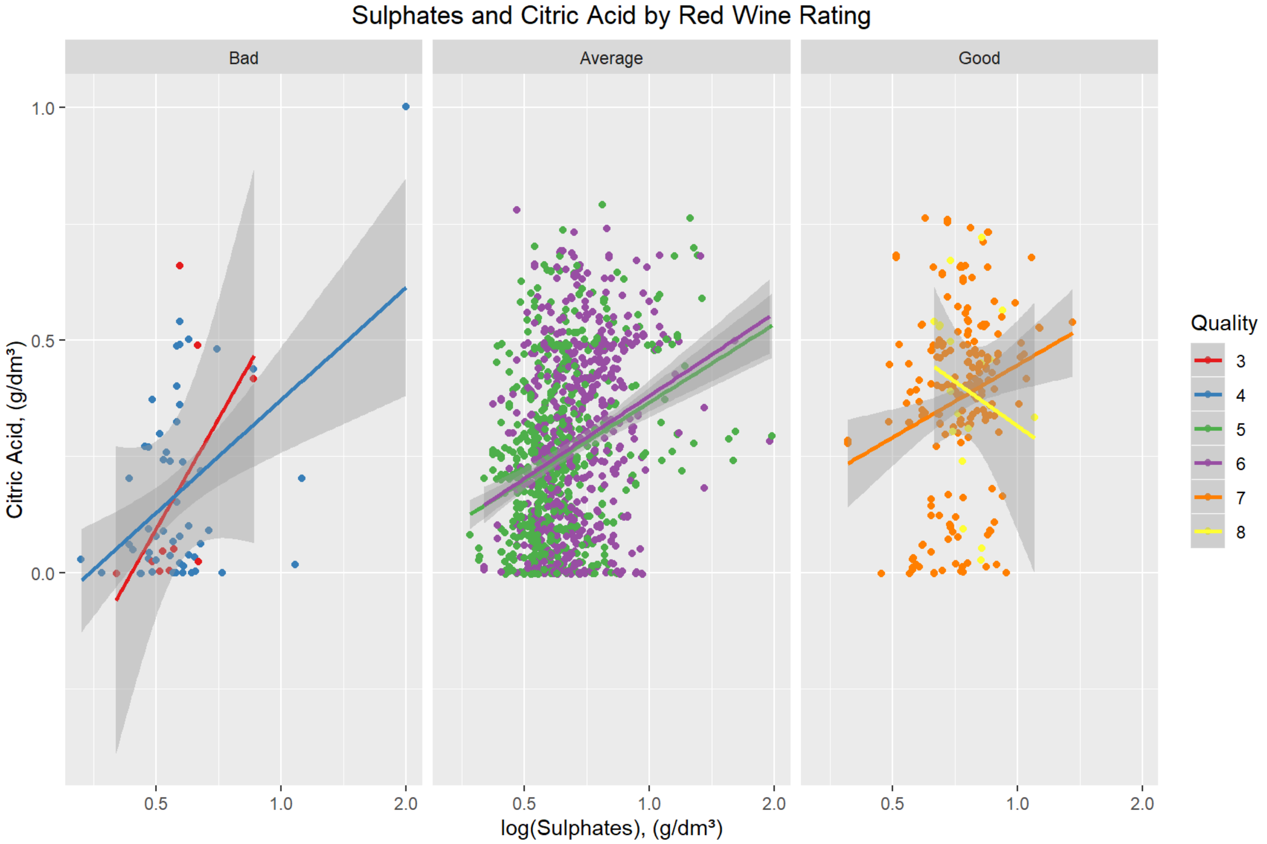
Finally, we can create a similar plot to examine volatile acidity and citric acid colored by quality. Here there isn't as quite as clear of a delineation between the low and high rated wines.



The highest rated wines tended to have higher citric acid concentrations and low volatile acidity concentrations, and the lower rated wines tended to have lower citric acid concentrations and higher volatile acidity concentrations.



As shown here most of the data distributed coming from the quality 5 to 7 with positive trend as sulphates contents increase the citic acid increase and this is for all wine quality except the quality number 8 and this is due to outliers.



We can see higher quality wine have higher sulphates (x-axis), higher citric acidity (y-axis) ,speacially for quality Number 5 and 6.

**We can see higher quality wine have higher alcohol,lower volatile acidity and higher sulphates.**

## Multivariate Analysis

In this section, I investigated the alcohol vs volatile acidity, citric acid vs volatile acidity, and sulphates vs citric acid analyzed by quality.

My focus here was to understand how the quality scores change these correlations.

**– Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?**

- Looking at alcohol plotted against volatile acidity and colored by quality rating helped to visualize the strongest relationships involving quality, even though alcohol and volatile acidity were weakly correlated (-0.202) themselves. The highest quality wines tended to have high alcohol percentages and low volatile acidity concentrations.This was already expected from the bivariate plotting section though.

- High Alcohol and Sulaphate content seems to produce better wines. Citric Acid, even though weakly correlated plays a part in improving the wine quality and influence the quality of wine positively, while adding volatile acid influences the quality of wine negatively.

- From pairplot , There is a very interesting relation between density, alcohol, residual sugar and quality. In general, quality increases as alcohol increases, density decreases and residual sugar decreases. These variables were amongst the most important predictors in the linear model built.

– **Were there any interesting or surprising interactions between features?**

- Higher Citric Acid and low Volatile Acid seems to produce better Wines. Correlation plot was showing that citric acid influence the quality of wine but from the plots above, we can observe that citric acid alone doen’t influence the quality that much.

– **OPTIONAL: Did you create any models with your dataset?Discuss the strengths and limitations of your model.**

## Linear Multivariable Model

Linear multivariable model was created to predict the wine quality based on chemical properties.

The features are selected incrementally in order of how strong the correlation between this feature and wine quality.

I will use combinations of two and more variables for the multiple regression model predicting the quality of red wine.

- First combination consists of all the variables that increase the quality with their increasing levels (m1).
- Next combination is density and fixed.acidity as its visual representation implied its value for predicting the quality (m2).
- Next goes volatile.acidity, as this variable has the highest negative correlation coefficient with the quality variable (m3).
- And the last combination consists of pH, total.sulfur.dioxide and free.sulfur.dioxide, based on the last step of the previous EDA (m4).

```
##
## Calls:
## m1: lm(formula = as.numeric(quality) ~ alcohol * sulphates * citric.acid *
##     fixed.acidity, data = redwine)
## m2: lm(formula = as.numeric(quality) ~ alcohol + sulphates + citric.acid +
##     fixed.acidity + density + alcohol:sulphates + alcohol:citric.acid +
##     sulphates:citric.acid + alcohol:fixed.acidity + sulphates:fixed.acidity
+
##     citric.acid:fixed.acidity + fixed.acidity:density + alcohol:sulphates:ci
tric.acid +
##     alcohol:sulphates:fixed.acidity + alcohol:citric.acid:fixed.acidity +
##     sulphates:citric.acid:fixed.acidity + alcohol:sulphates:citric.acid:fixe
d.acidity,
##     data = redwine)
## m3: lm(formula = as.numeric(quality) ~ alcohol + sulphates + citric.acid +
##     fixed.acidity + density + volatile.acidity + alcohol:sulphates +
##     alcohol:citric.acid + sulphates:citric.acid + alcohol:fixed.acidity +
##     sulphates:fixed.acidity + citric.acid:fixed.acidity + fixed.acidity:dens
ity +
##     alcohol:sulphates:citric.acid + alcohol:sulphates:fixed.acidity +
##     alcohol:citric.acid:fixed.acidity + sulphates:citric.acid:fixed.acidity
+
##     alcohol:sulphates:citric.acid:fixed.acidity, data = redwine)
## m4: lm(formula = as.numeric(quality) ~ alcohol + sulphates + citric.acid +
##     fixed.acidity + density + volatile.acidity + pH + total.sulfur.dioxide
+
##     free.sulfur.dioxide + alcohol:sulphates + alcohol:citric.acid +
##     sulphates:citric.acid + alcohol:fixed.acidity + sulphates:fixed.acidity
+
##     citric.acid:fixed.acidity + fixed.acidity:density + pH:total.sulfur.diox
ide +
##     pH:free.sulfur.dioxide + total.sulfur.dioxide:free.sulfur.dioxide +
##     alcohol:sulphates:citric.acid + alcohol:sulphates:fixed.acidity +
##     alcohol:citric.acid:fixed.acidity + sulphates:citric.acid:fixed.acidity
+
##     pH:total.sulfur.dioxide:free.sulfur.dioxide + alcohol:sulphates:citric.a
cid:fixed.acidity,
##     data = redwine)
##
## ============================================================================
=========================
##                                                          m1         m
2          m3           m4
## ----------------------------------------------------------------------------
-------------------------
##   (Intercept)                                             4.004    -18.407
-30.327      -49.730
##                                                          (8.287)   (52.025)
```

```
(50.640)     (52.164)
##   alcohol                                              -0.248     -0.03
1     -0.006        0.166
##                                                        (0.791)    (0.78
8)    (0.767)      (0.772)
##   sulphates                                            -1.259      1.73
3      1.544        1.326
##                                                       (11.824)   (11.869)
(11.550)     (11.605)
##   citric.acid                                          24.816     31.95
4     33.083       36.242*
##                                                       (18.651)   (18.861)
(18.354)     (18.386)
##   fixed.acidity                                         0.078     10.12
0      8.308        7.842
##                                                        (1.148)    (5.64
6)    (5.497)      (5.526)
##   alcohol x sulphates                                   0.400      0.15
1      0.199        0.126
##                                                        (1.120)    (1.12
1)    (1.091)      (1.098)
##   alcohol x citric.acid                                -1.790     -2.48
3     -2.711       -3.038
##                                                        (1.769)    (1.78
4)    (1.736)      (1.742)
##   sulphates x citric.acid                             -55.541*   -64.136*
-62.580*     -61.781*
##                                                       (25.491)   (25.795)
(25.101)     (25.161)
##   alcohol x fixed.acidity                               0.004     -0.02
9     -0.014       -0.039
##                                                        (0.111)    (0.11
1)    (0.108)      (0.108)
##   sulphates x fixed.acidity                            -0.679     -1.02
5     -0.781       -0.833
##                                                        (1.639)    (1.64
0)    (1.596)      (1.602)
##   citric.acid x fixed.acidity                          -3.359     -4.13
0     -4.076       -4.447*
##                                                        (2.308)    (2.32
5)    (2.262)      (2.263)
##   alcohol x sulphates x citric.acid                     4.484      5.231
*      5.177*       5.190*
##                                                        (2.400)    (2.42
7)    (2.362)      (2.369)
##   alcohol x sulphates x fixed.acidity                   0.052      0.08
2      0.048        0.067
##                                                        (0.158)    (0.15
8)    (0.154)      (0.155)
```

```
##   alcohol x citric.acid x fixed.acidity                 0.273      0.34
9      0.342        0.385
##                                                       (0.221)    (0.22
2)    (0.216)      (0.216)
##   sulphates x citric.acid x fixed.acidity               6.979*     7.874
*     7.330*       7.300*
##                                                       (3.168)    (3.19
1)    (3.105)      (3.109)
##   alcohol x sulphates x citric.acid x fixed.acidity    -0.597*    -0.675
*    -0.621*      -0.634*
##                                                       (0.302)    (0.30
4)    (0.295)      (0.296)
##   density                                                         19.74
5     31.965       55.838
##                                                                  (52.053)
(50.668)     (52.270)
##   fixed.acidity x density                                         -9.68
5     -7.953       -7.355
##                                                                   (5.58
4)    (5.436)      (5.465)
##   volatile.acidit
y                                                         -1.106***    -0.985
***
#
#
(0.117)      (0.119)
##   p
H
        -1.566***
#
#
             (0.375)
##   total.sulfur.dioxid
e                                                                    -0.072*
*
#
#
             (0.024)
##   free.sulfur.dioxid
e                                                                     -0.207*
*
#
#
             (0.066)
##   pH x total.sulfur.dioxid
e                                                                0.021**
#
#
             (0.007)
```

```
##   pH x free.sulfur.dioxid
e                                                                 0.065**
#
#
             (0.020)
##   total.sulfur.dioxide x free.sulfur.dioxid
e                                                0.003***
#
#
             (0.001)
##   pH x total.sulfur.dioxide x free.sulfur.dioxid
e                                          -0.001***
#
#
             (0.000)
## -----------------------------------------------------------------------------
-------------------------
##   R-squared                                                0.333      0.34
2      0.377        0.391
##   N                                                     1599       1599       1
599         1599
## =============================================================================
=========================
##   Significance: *** = p < 0.001; ** = p < 0.01; * = p < 0.05
```

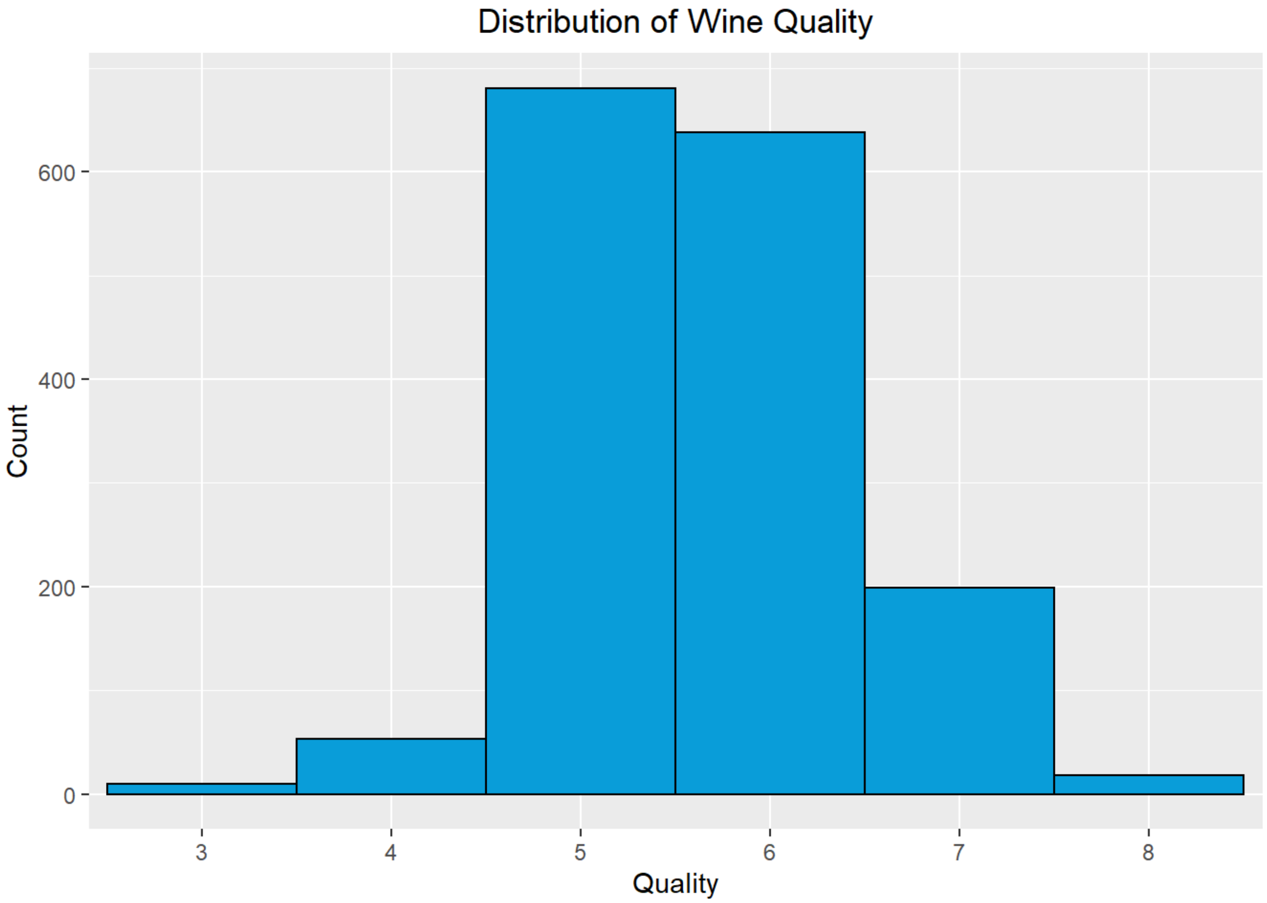
- The given model explains 39% of cases in the given dataset which is not a great linear model we obtained. The highest R-squared = 0.333(tells us this model is not good for explaining wine quality) is provided by the first combination of parameters (alcohol, sulphates, citric.acid, fixed.acidity). Next three sets of features add 0.009-0.058 to the previous R-squared value, which was a very low one. It indicates that a linear model probably is not the best fit for this dataset.

- This model has limitations. It is based on the limited data that does not provide very high (more than 8) and very low (less than 3) quality scores. Collecting the data with more cases of extreme scores, as well as additional data with existing low-quality scores (3 and 4), could significantly improve the model’s predictive power.

- Also,We are not entirely surprised by this limitations given that the correlation between any variable and quality was not very high, and 82.5% of the dataset was in average.

## Final Plots and Summary

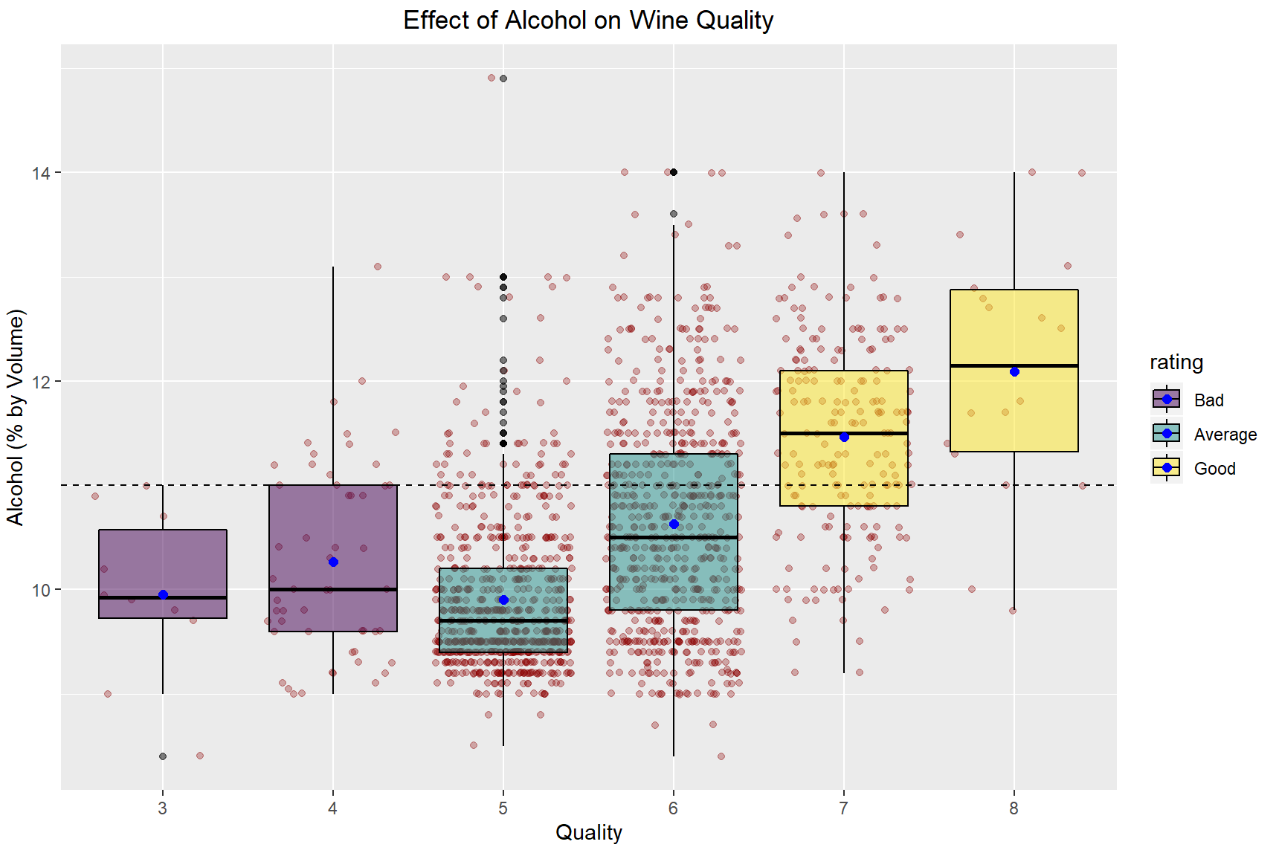
### Plot One

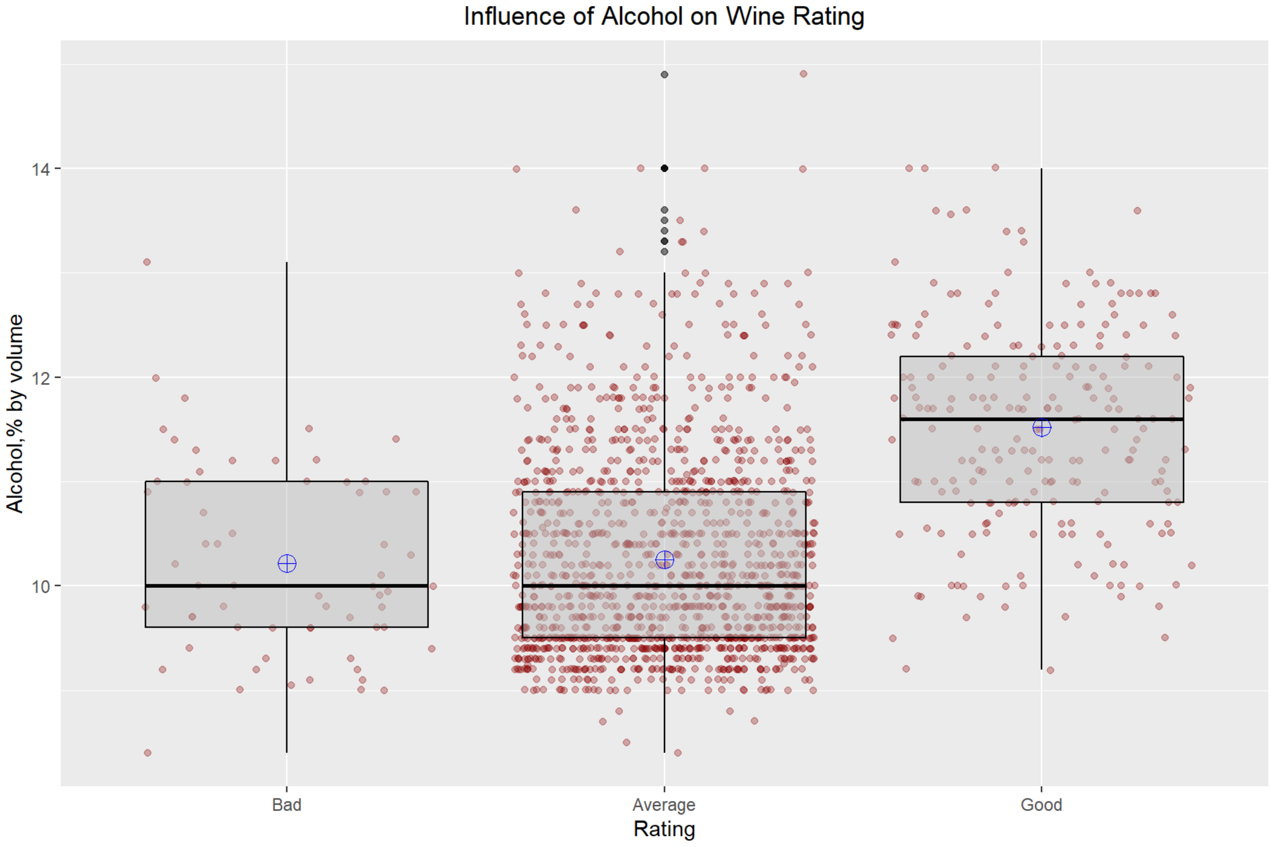


**Description**

Since quality is our main feature of interest, it’s important to pay special attention to it. We can see that the data is not balanced, there are much more normal wines than poor or excellent ones.Among 1599 observations of wines, 82.5% of wines received average quality score (5 or 6), 3.93% of wines received low quality score (3 or 4), and 13.57% of wines received high quality score (7 or 8). This can make it somewhat difficult to understand what makes a good wine, as we have such a limited selection, relative to average wines. Having more lesser qualities wines would also be helpful as it would provide a stronger comparison between what makes a bad wine versus a good wine.

### Plot Two


Effect of Alcohol on Wine Quality
Alcohol (% by Volume)
14
12
10
3
4
5
6
7
8
Quality
rating
Bad
Average
Good


Influence of Alcohol on Wine Rating
Alcohol,% by volume
14
12
10
Bad
Average
Good
Rating

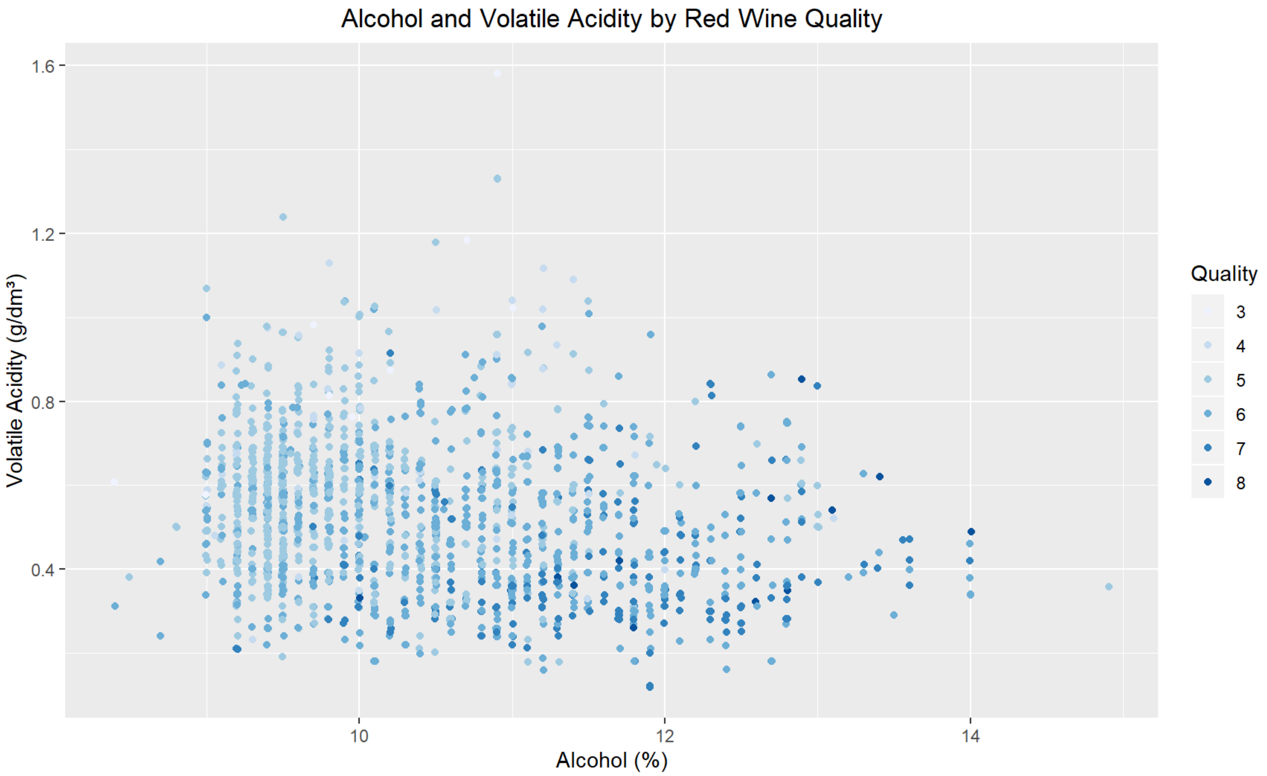
```
## redwine$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   8.400   9.725   9.925   9.955  10.575  11.000 
## -------------------------------------------------------- 
## redwine$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    9.00    9.60   10.00   10.27   11.00   13.10 
## -------------------------------------------------------- 
## redwine$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     8.5     9.4     9.7     9.9    10.2    14.9 
## -------------------------------------------------------- 
## redwine$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    8.40    9.80   10.50   10.63   11.30   14.00 
## -------------------------------------------------------- 
## redwine$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    9.20   10.80   11.50   11.47   12.10   14.00 
## -------------------------------------------------------- 
## redwine$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    9.80   11.32   12.15   12.09   12.88   14.00
```

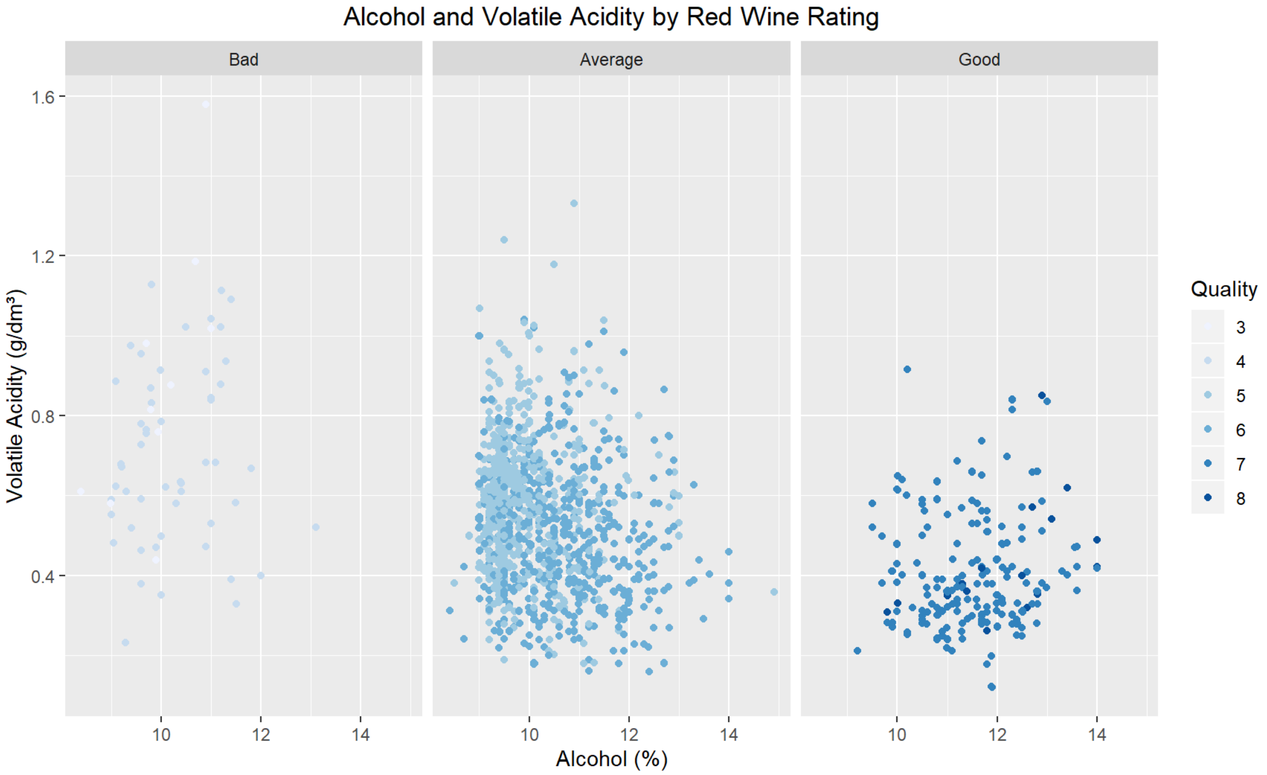
### Description

This plot tells us that Alcohol percentage has played a big role in determining the quality of Wines. Alcohol has the strongest correlation with quality, the higher the alcohol percentage, the better the wine quality.A 75% of good wine contain above 11% of alcohol, while 75% of average and bad quality ones have a alcohol percent bellow 11%.

In this dataset, even though most of the data pertains to average quality wine, we can see from the above plot that the mean and median coincides for all the boxes implying that for a particular Quality it is very normally distributed. So a very high value of the median in the best quality wines imply that almost all points have a high percentage of alcohol. For scores from 5 to 8, quality increases as alcohol increases, and for scores 3 and 4 the relation is the inverse. Alcohol has the largest correlation with quality among all the variables in the dataset, with a Pearson's correlation coefficient of 0.476 .

## Plot Three


Alcohol and Volatile Acidity by Red Wine Quality


Alcohol and Volatile Acidity by Red Wine Rating

This chart shows how quality improves as the alcohol content increases and the volitile acidity decreases. There is an overall trend of the colors getting darker as they go to the bottom right.

Alcohol by volume and volatile acidity were the two chemical properties most closely related to quality in red wine. Alcohol had a positive relationship with quality, perhaps due to a higher concentration of flavor in wines with higher alcohol percentages. Volatile acidity had a negative relationship with quality rating, due to the fact that higher concentrations can lead to undesirable vinegar-like flavors. As evidenced by the three distinct regions in the plot, the lowest quality wines tended to have lower alcohol percentages and higher volatile acidity concentrations, while the higher quality wines had higher alcohol percentages and lower volatile acidity concentrations, in general.

## Reflection

- The red wine dataset contains information on 1599 observations Portuguese Red Wines of 11 chemical properties plus a score of wine quality. Clearly, this dataset was designed to understand the quality of the wine by its chemical properties. The biggest difficulty was handling the complexity of the dataset. It was hard to keep track of all of the different relationships at play and to determine where to focus next. Because there were so many different potential directions to go with the analysis, it was hard for me to balance my desire to be thorough and consistent with trying to focus only on the important information. It was a good taste of the difficulties of dealing with complex datasets.

- My goal was to determine which chemical properties had the strongest effect on perceived red wine quality. I read up on information about each property so I understood overall implications as I looked at the dataset further. I started by exploring the relationship individual variables with quality look for any interesting distributions and to get a feel for the ranges of values and seeing which ones correlated most highly with the quality rating. After looking at the distributions of some variables, I looked at the relationship between two- and, eventually, three-variable combinations by calculated the correlation coefficients for each combination of variables in order to determine the strengths of the relationships between the variables using correlation matrix and a scatterplot matrix to decide what relationships I wanted to investigate between variables, particularly those involving quality. We are interested in the correlation between the features and wine quality.

- The wine quality is more complex. It does not have an obvious driver. Most of the data visualization in this project was done on the 4 features that have the highest correlation coefficient: alcohol(0.476), volatile acidity(-0.391), sulphates(0.251), citric acid(0.226). Based on these findings, I explored the data further, concentrating on the effect of alcohol, volatile acidity, sulphates content and citric acid.

  In one hand it was possible to see the correlation between some chemical properties (primarily Alcohol percentage and Volatile Acidity) and wine quality. This was expected, common sense tells us that people usually like “strong” wines (in terms of Alcohol) and dislike high levels of acidity, which can lead to an unpleasant, vinegar taste.

  One interesting observation was that some wines didn’t have citric acid at all. I initially thought it was maybe due to incomplete data but then I researched further about wines. I saw that citric acid actually is added to some wines to increase the acidity. So it’s obvious that some wines would not have Citric Acid at all. I tried to figure out the effect of each individual acid on the overall pH of the wine. Here I found out a very strange phenomenon where I saw that for volatile acids, the pH was increasing with acidity which was against everything I learned in my Science classes.

I was surprised by the dramatic effect on the correlation that removing sulphate outliers had on correlation. A graphic lesson to learn about outliers. The log10 histograms were interesting in the way they created normal distributions for several factors. The use of the "normal" and the log10 histograms were helpful in terms of identifying anomalies in the distributions (e.g. bimodal peaks).

- My findings showed that the key factor affecting the quality of wine to have most good red wines have mainly high alcohol, sulphate and citric acid levels and low volatile acidity even though we had a limited data of 1599 observations. In that dataset, I noticed that the dataset is highly unbalanced 82.5% of the wines are of average quality between (grade 5 & 6). If we could have a dataset of more observations in the future for improvement can be made if more data can be collected on both low-quality (grade 3,4) and high-quality (grade 7,8) wine and uniform quality of wines, it has fewer data points. Then we will be able to perform a better analysis. We can be more certain about whether there is a significant correlation between a chemical component and the wine quality.
- A linear model for predicting quality was built for the combination of variables would make a good predictive model, I spent more time playing with the data, making more models and seeing how the attributes work together to make good quality wine, but it performed poorly as its coefficient of determination was very low, indicating that the dataset did not behave very much linearly. The process of evaluating wines is very subjective, and experts can be biased by their histories and preferences, making the relation between quality and the other variables too complex to be explained by a linear model. The best model was able to account only 39% of the variance. It is possible that the failure of this analysis in creating a linear model can be explained by the dataset consisting on much more 'Average' wines (1319) than Bad (63) or Good (217) ones. In the future, a different set of quality prediction models could be applied, and an evaluation of the best fit could be performed perhaps conduct a nonlinear regression modelling seeking for better results.

## References

http://vita.had.co.nz/papers/tidy-data.pdf (http://vita.had.co.nz/papers/tidy-data.pdf)

https://en.wikipedia.org/wiki/Vinho_Verde (https://en.wikipedia.org/wiki/Vinho_Verde)

http://dx.doi.org/10.1016/j.dss.2009.05.016 (http://dx.doi.org/10.1016/j.dss.2009.05.016)

http://www3.dsi.uminho.pt/pcortez/winequality09.pdf (http://www3.dsi.uminho.pt/pcortez/winequality09.pdf)

http://www3.dsi.uminho.pt/pcortez/dss09.bib (http://www3.dsi.uminho.pt/pcortez/dss09.bib)

https://www.datamentor.io/r- (https://www.datamentor.io/r-)

https://stackoverflow.com/questions/40675778/center-plot-title-in-ggplot2 (https://stackoverflow.com/questions/40675778/center-plot-title-in-ggplot2) http://blog.revolutionanalytics.com/2015/09/resizing-plots-in-the-r-kernel-for-jupyter-notebooks.html (http://blog.revolutionanalytics.com/2015/09/resizing-plots-in-the-r-kernel-for-jupyter-notebooks.html)

https://briatte.github.io/ggcorr/#controlling-the-main-geometry (https://briatte.github.io/ggcorr/#controlling-the-main-geometry)

https://stackoverflow.com/questions/6328771/changing-values-when-converting-column-type-to-numeric (https://stackoverflow.com/questions/6328771/changing-values-when-converting-column-type-to-numeric)

https://ggplot2.tidyverse.org/reference/scale_brewer.html (https://ggplot2.tidyverse.org/reference/scale_brewer.html)

https://www.dummies.com/programming/r/how-to-convert-a-factor-in-r/ (https://www.dummies.com/programming/r/how-to-convert-a-factor-in-r/)

https://github.com/ (https://github.com/)