Red wine quality exploratory data analysis

This analysis explores the relationship between concentration of chemicals and the quality of red wine. The dataset is related to red variants of the Portuguese "Vinho Verde" wine. For more details, consult: http://www.vinhoverde.pt/en/

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):

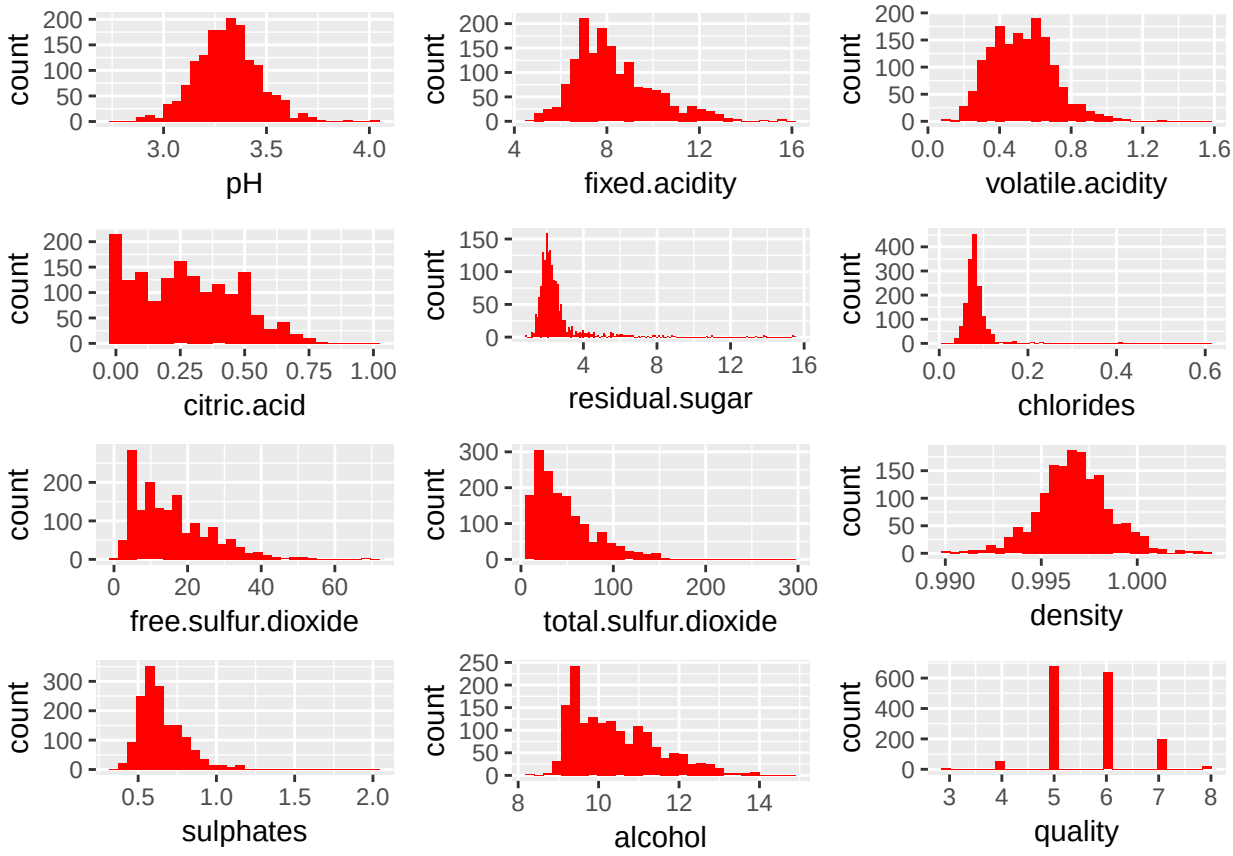
- quality (score between 0 and 10)

My target variable is "quality"" and my goal is to explore correlations between physiochemical variables and quality with the hope of predicting quality from physiochemical properties of wine.

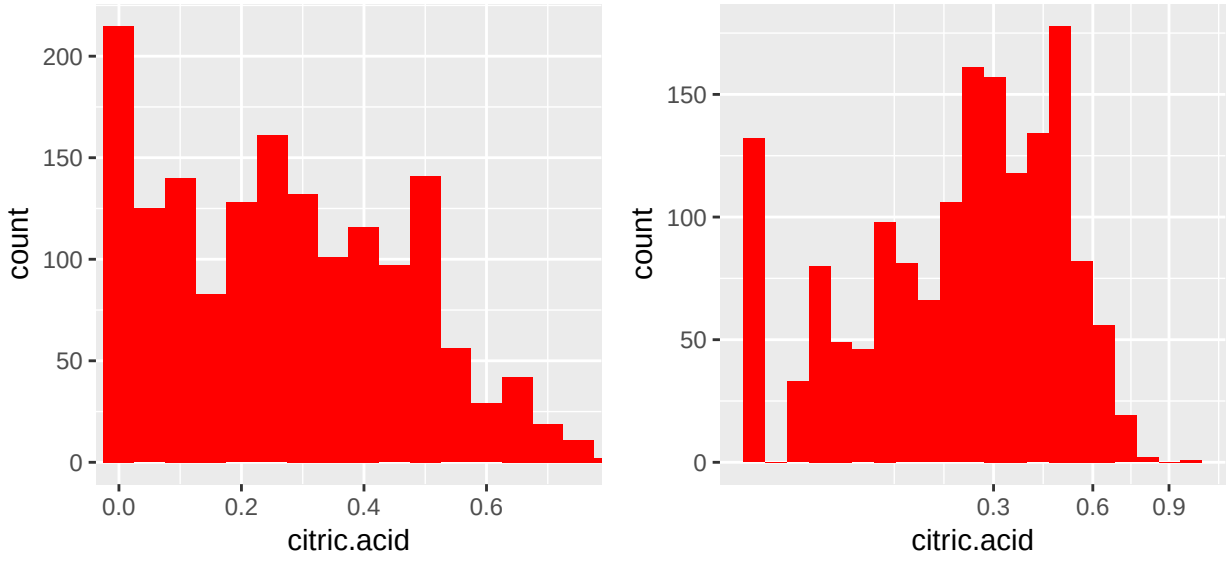
Univariate Plots Section

```
## '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 0.065 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 3.35 ...
##  $ 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 ...
```

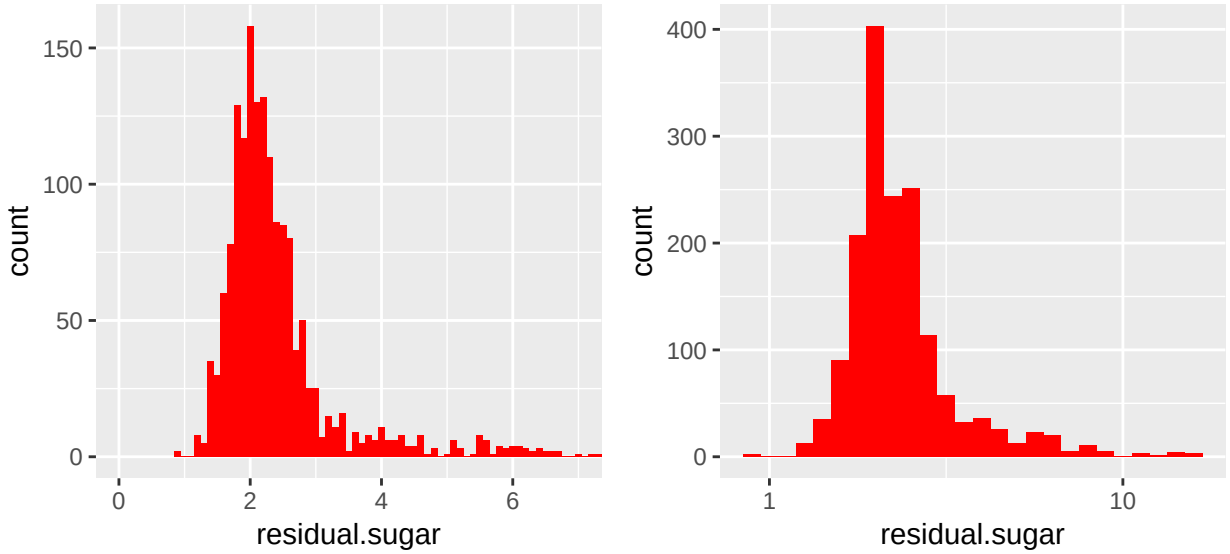
```
##        X          fixed.acidity   volatile.acidity  citric.acid
##  Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000
##  1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090
##  Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260
##  Mean   : 800.0   Mean   : 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
```



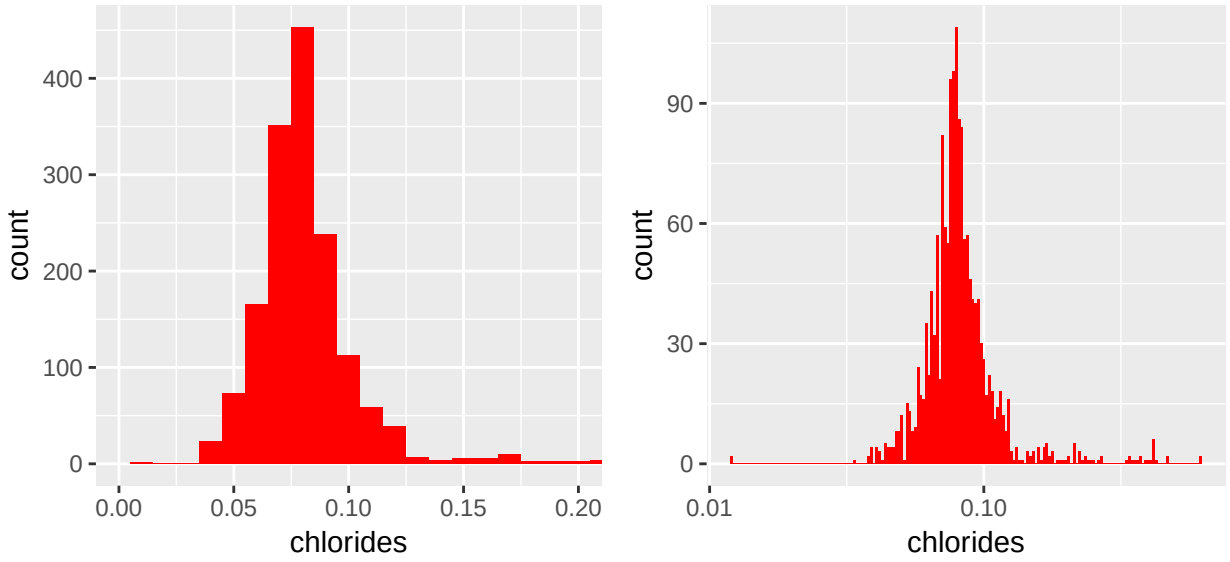
After an initial analysis of the distributions of the variable, we should focus on those with long tails, outliers, or non-normal distributions.



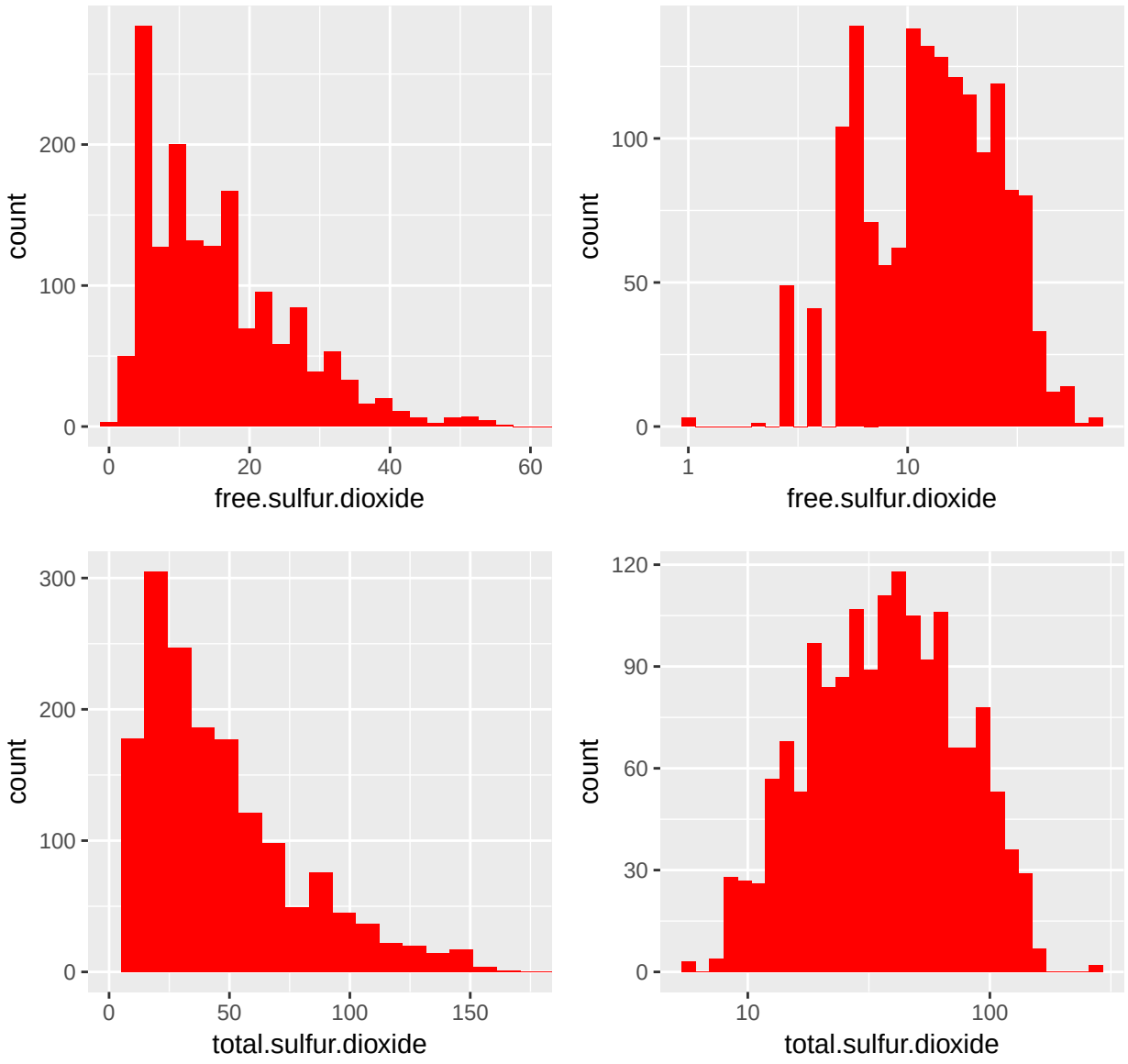
Distribution of citric acid is positively skewed. Mean citric acid content is 0.2709756 and its standard deviation is 0.1948011. However, both square root and logarithmic transformations failed to transfrom the data correctly. A simple cut off of values more than 0.75 was performed.



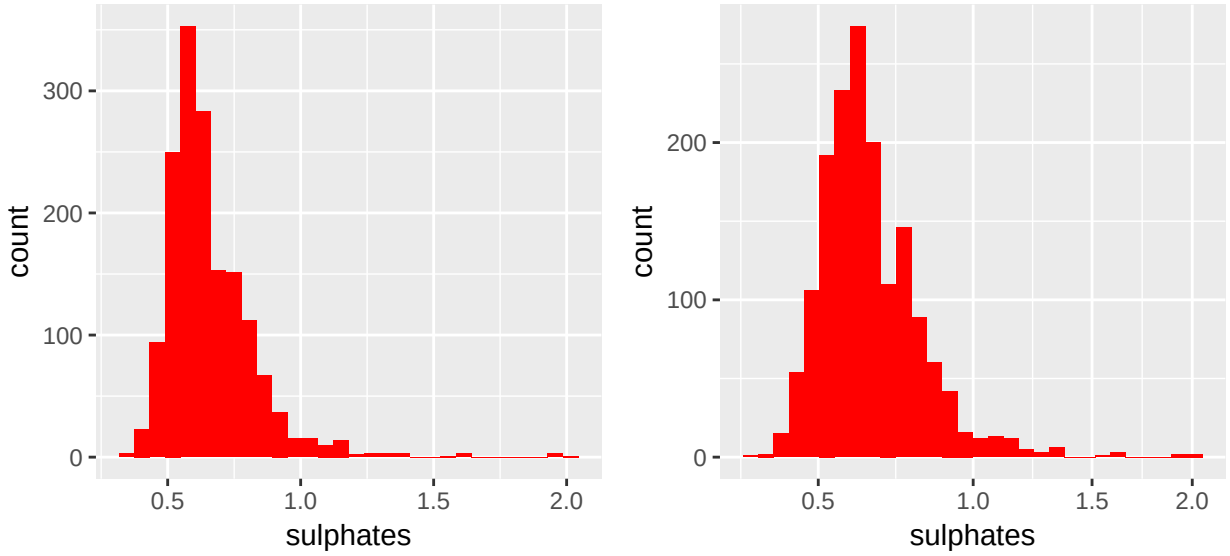
Distribution of residual sugar demonstrates a long tail to the right side of the distribution. Mean residual sugar content is 2.5388055 and its standard deviation is 1.4099281. log 10 transformation was performed (right hand side figure) to make the data a bit more manageable.



Chloride distribution also shows a tail to the right. However, no transformation significantly changed the data. Mean chloride content is 0.0874665 and its standard deviation is 0.0470653. A simple cut off of values more than 0.17 was performed.



Distribution of free and total sulfur dioxide was positively skewed. Mean free sulfur dioxide content is 15.8749218 and its standard deviation is 10.460157. Mean total sulfur dioxide content is 46.4677924 and its standard deviation is 32.8953245.log 10 transformation was performed (right hand side figures) on both.



Mean total sulphates content is 0.6581488 and its standard deviation is 0.169507. Sulphates distribution had a long tail, which was cut off at values larger than 1.2.

There are 1599 samples in the dataset. This dataset is very neat and tidy without NA values. Most parameters are normally distributed. Citric acid, free sulfuric acid and total sulfur dioxide are positively skweved. log10 transformation was performed on residual sugar, total, and free sulfur dioxide. It is worth mentioning that the quality(0-10) variable is discrete with 3 as the minimum and 8 as the maximum. However, more than 80% of the samples are graded 5 or 6 giving the quality variable a bell-shaped distribution.

Univariate Analysis

What is the structure of your dataset?

There are 1599 wines listed in the dataset with 12 varaibles (fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates, alcohol, and quality). All variables are numerical except for the quality variable, which is integer. There are no categorical variable. Observetions:

- Most variables are normally distibuted.
- The quality variable can take any value between 0 and 10. However, the minimum in the dataset is 3 and the maximum is 8.
- More than 80% of wines have a quality of 5 or 6.

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

The most interesting feature is the pH since theoretically, it lumps many other physiochemical quantities in it. Alcohol content might also have some impact on quality. In addition, the quality variable is also interesting to explore and is the main dependent variable in this analysis.

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

All other variables (fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, and sulphates) can help us analyze the data.

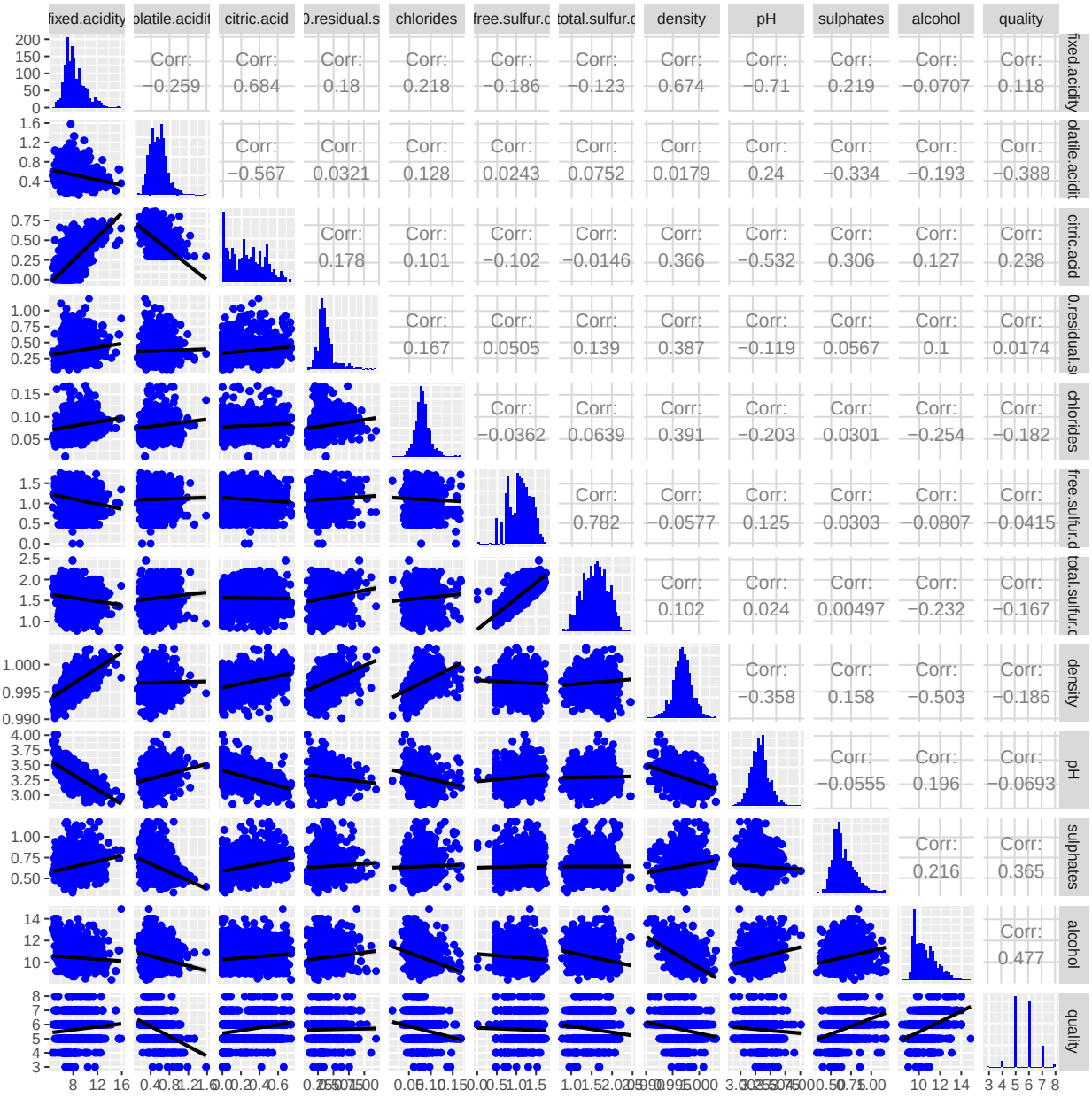
Did you create any new variables from existing variables in the dataset?

The quality variable was changes into a factor variable to provide the necessary type for certain types of figures.

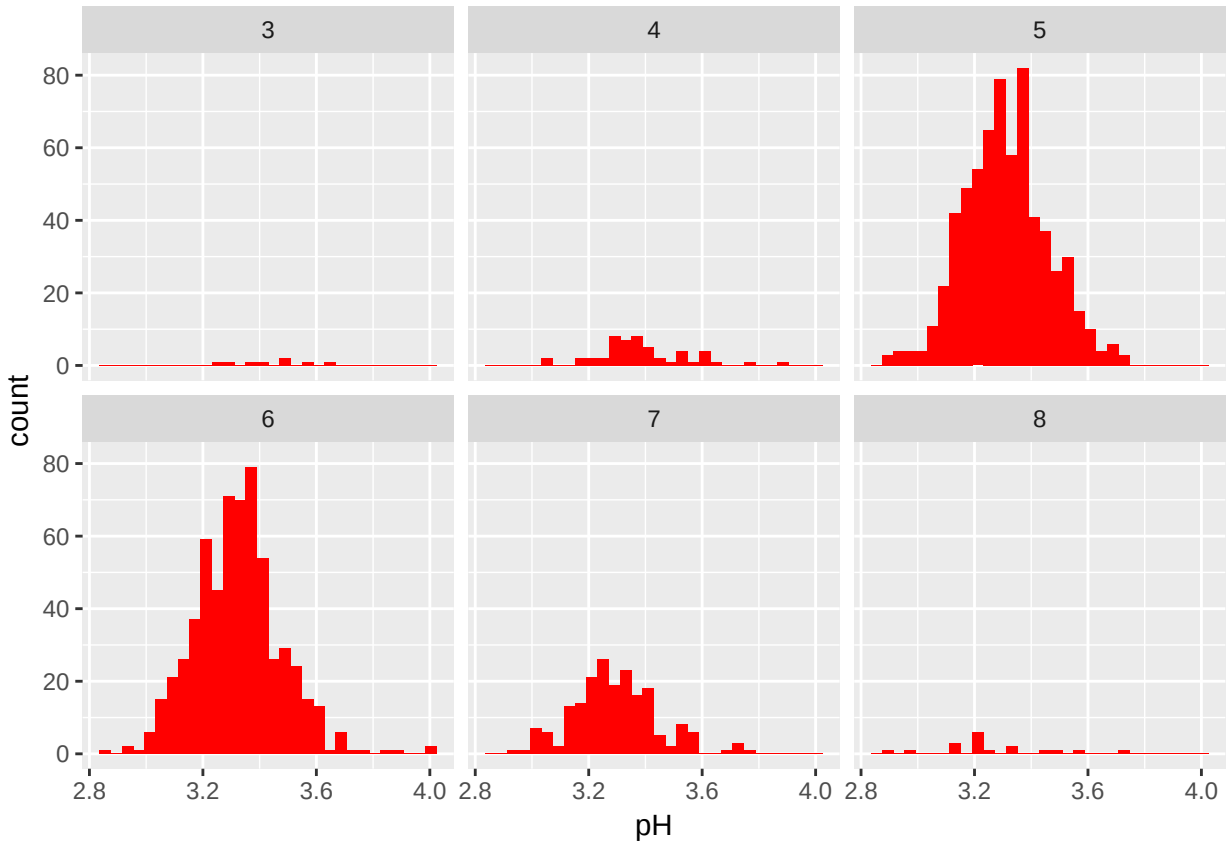
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

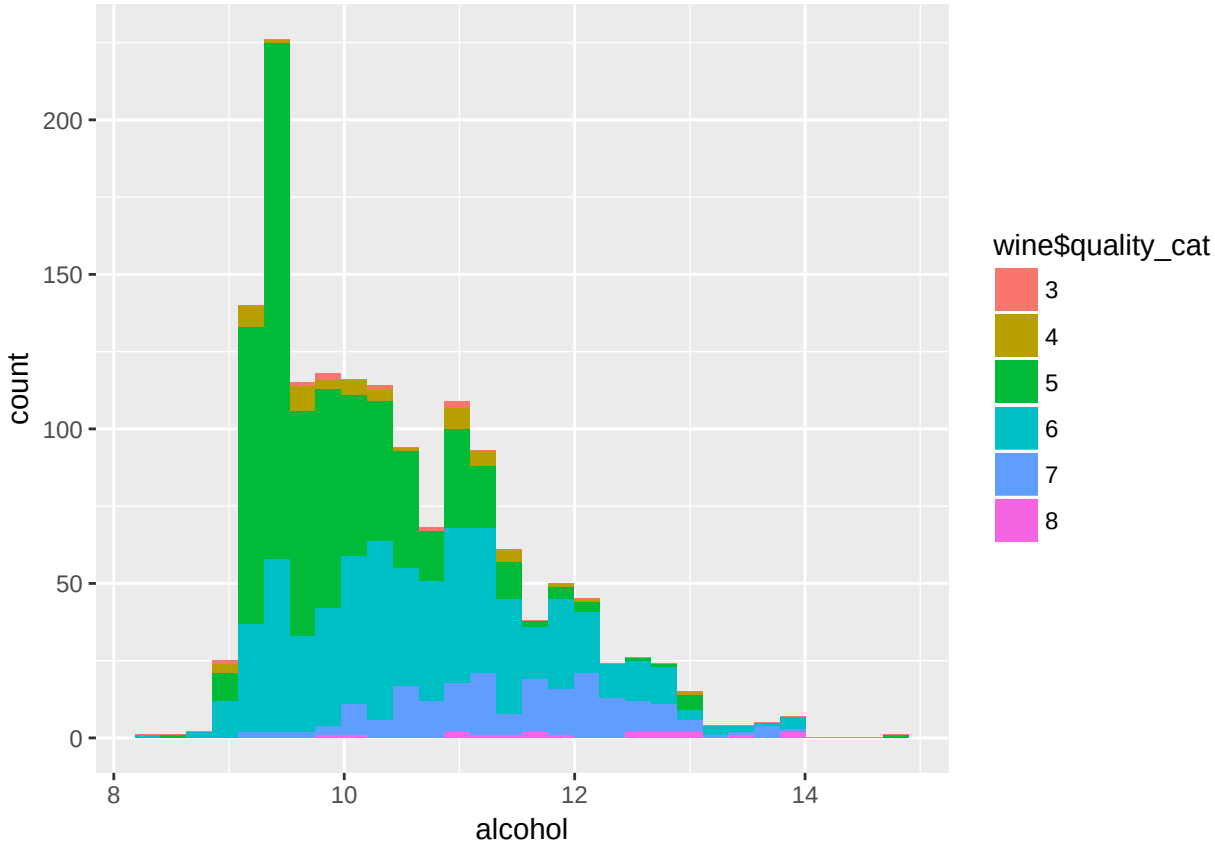
Later in the analysis some outliers were detected that were removed to simplify the analysis. In addition, log 10 transformation was performed on several varaibles.

Bivariate Plots Section



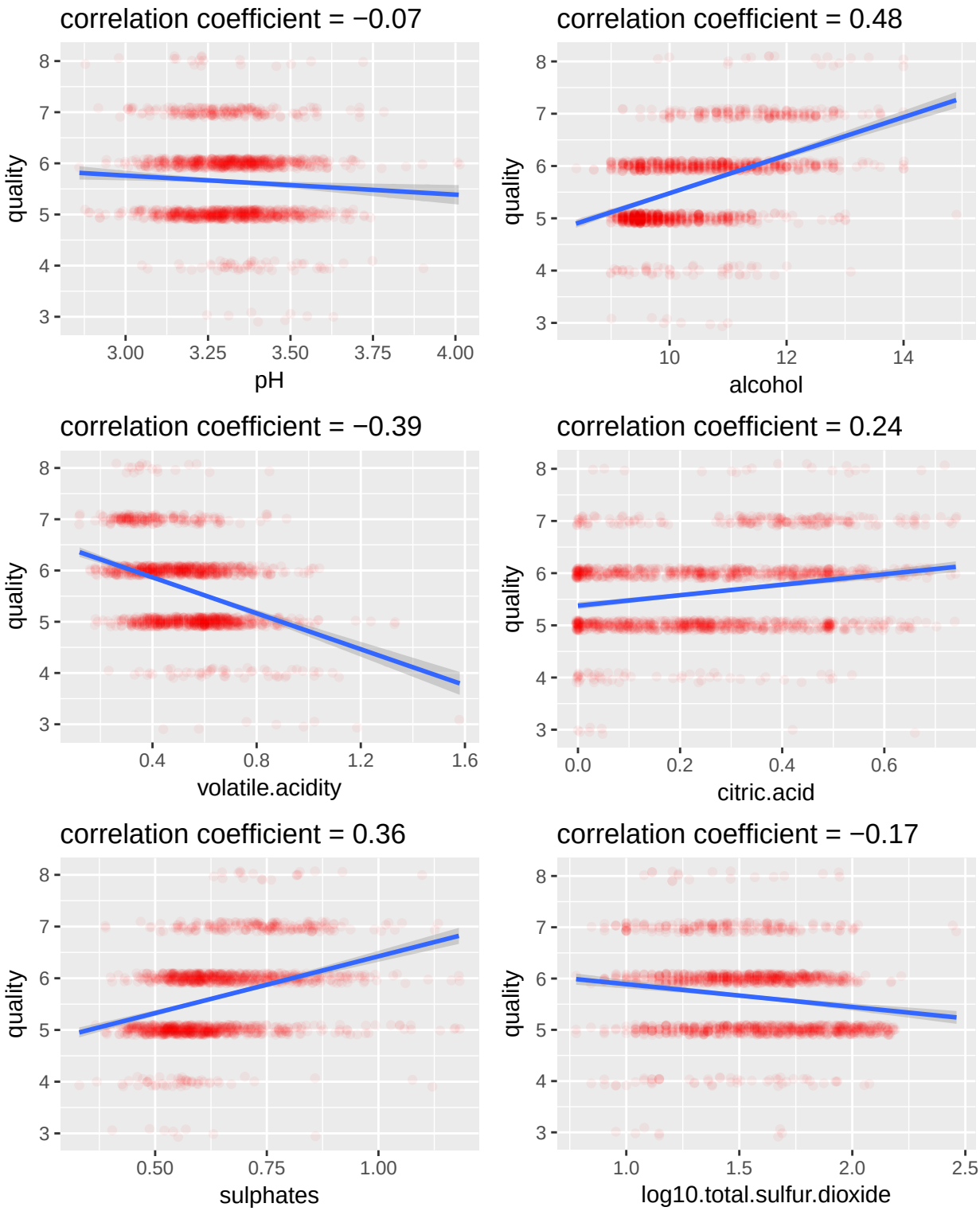
Based on correlation coefficients calculated in this step and my initial idea, the relationship between quality with ph,alcohol, volatile acidity, citric acid, sulphates, and log10 of total sulfur dioxide was chosen to be examined more in depth.


3
4
5
6
7
8
count
pH
0
20
40
60
80
2.8
3.2
3.6
4.0

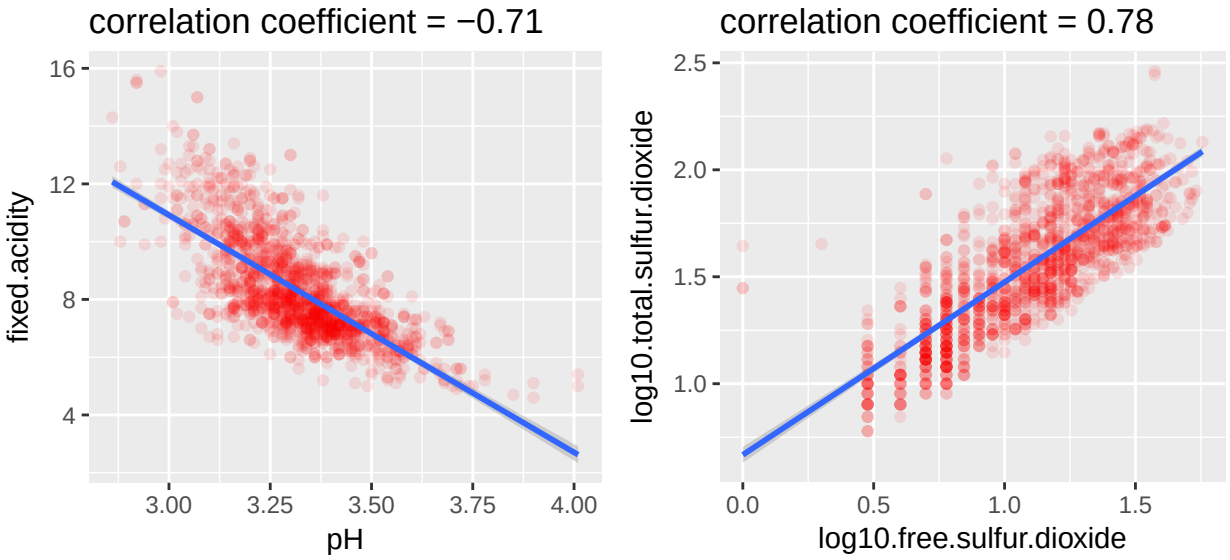


pH shows no understandable relationship with quality.

Center of distribution of alcohol content moves toward higher values as quality increases.

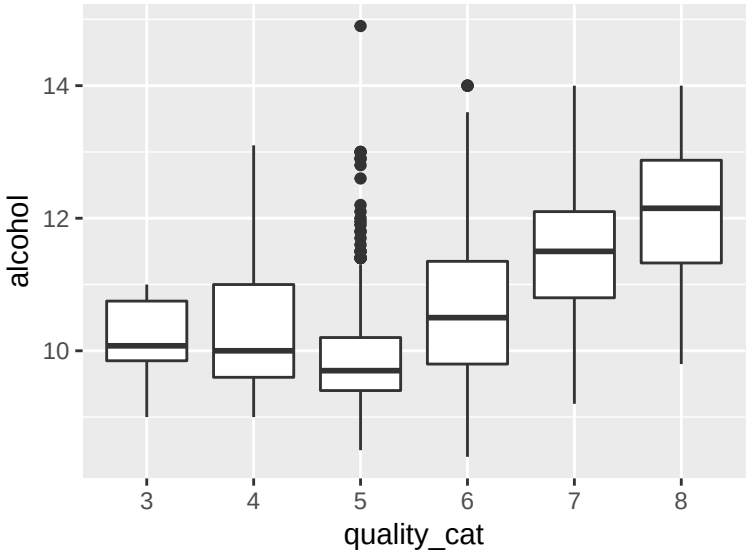


Based on this analysis, alcohol, volatile acidity, and sulphates are promising variables that might explain variability of the quality variable.

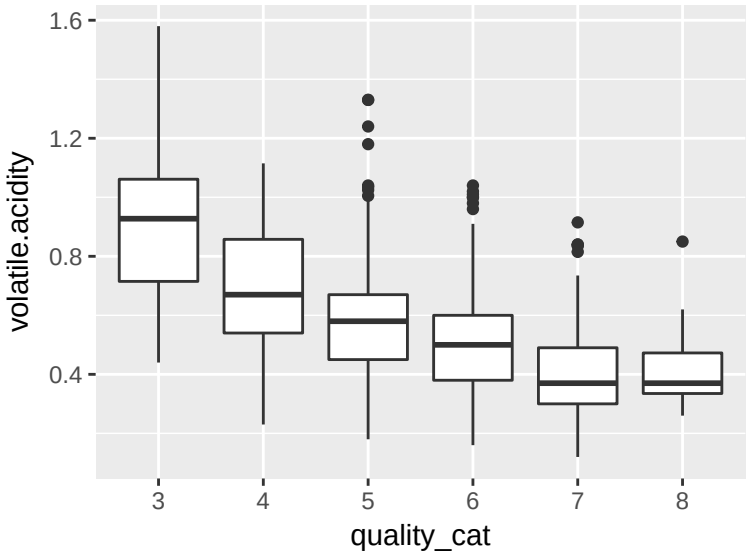


Among all other variables, pH and fixed acidity are strongly correlated, which is expected. pH is a measure of acidity and lower values of pH indicate acidic environment. Similarly, free and total sulfur dioxide are correlated. Free sulfur is a portion of total sulfur and the correlation is expected.

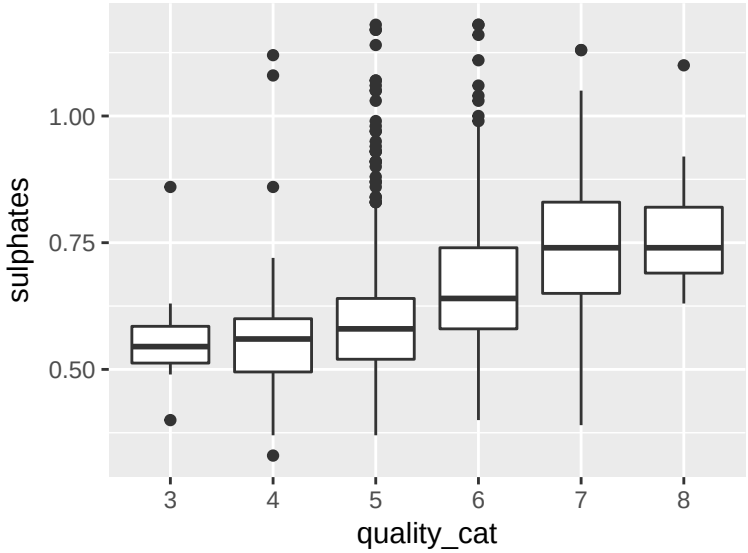
Since the varaibility of the data is high, box plots might be more helpful as they show the middle 50% of the data.



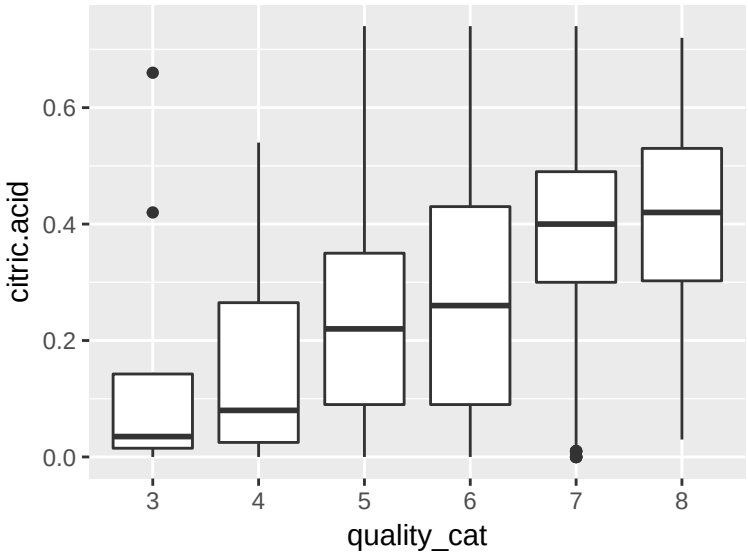
There is a clear relationship between alcohol and quality especially in the higher end of quality. Box plots are more understandable than scatter plots here.



There is also a clear negative relationship between volatile acidity and quality.



The relationship of sulphates and quality is more subtle compared to the previous two.



There is also a positive relationship between citric acid and quality. However, the variance of data is a bit higher than previous variables.

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?

The variable of interest in this study is quality. The most significant positivie correlation was between alcohol content and quality. The most significant negative correlation was between volatile acidity and quality.

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

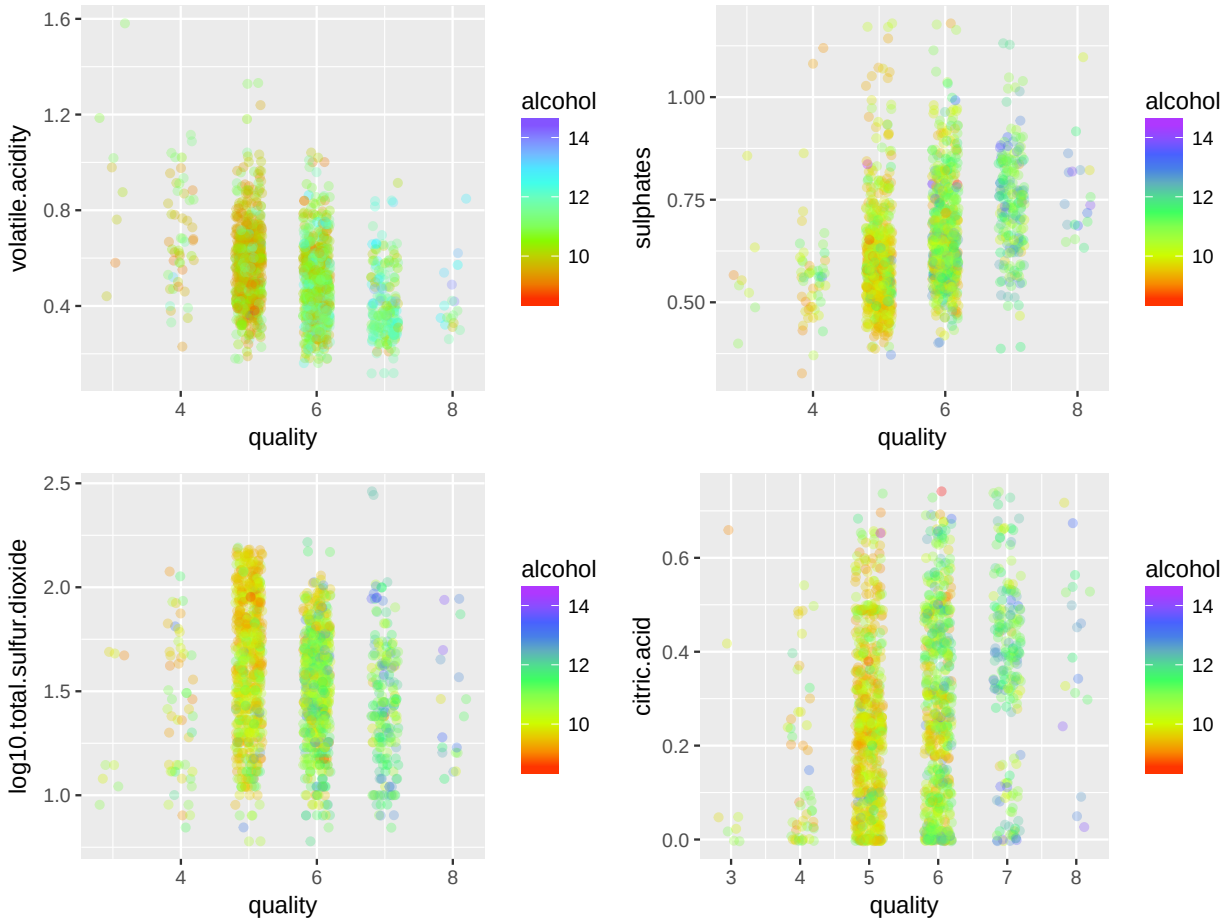
Overall the most significant correlation between all pairs of variables was between total sulfur dioxide and free sulfur dioxide followed by the correlation between pH and fixed acidity, which is logical since pH is a measure of acidity. More interestingly, density was highly correlated with fixed acidity, which I did not expect.

What was the strongest relationship you found?

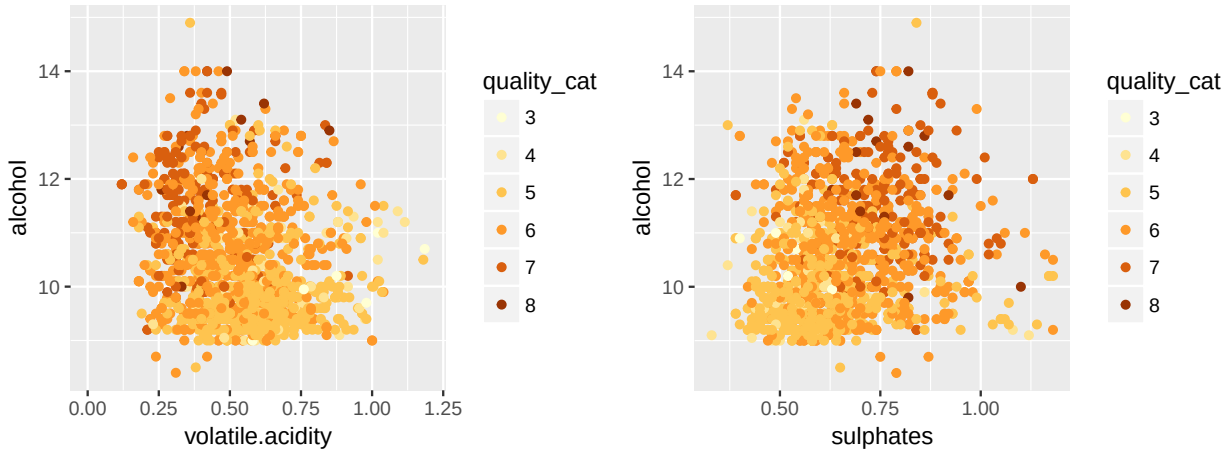
The strongest correlation was between total sulfur dioxide and free sulfur dioxide with the coefficient of 0.78.

Multivariate Plots Section

Alcohol content proved to be the most significant parameter to explain the variance of the quality variable. Therefore, other variables were plotted in conjuction with alcohol to examine if they are capable of describing the remaining varability in the data. In the following plots, if the data point in each column show a trend with respect to the variable plotted on the y axis, that parameter can help us explain some variance in the data.



Based on the previous plots, alcohol content is the best variable to describe the variance in the data. in addition to alcohol, volatile acidity and sulphates were the next important variables that describe some of the variance. In contrast, citiric acid and total sulfur dioxide are not helpful variables to determine wine quality.



Finally, we can observe that higher alcohol content, lower volatile acidity, and higher sulphate content correlate with higher quality of wine.

Multivariate Analysis

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?

As expected based on the correlation analysis, alcohol content is the most helpful parameter to understand quality of wine based in this dataset. After that, volatile acidity and sulphates are the next important parameters.

Were there any interesting or surprising interactions between features?

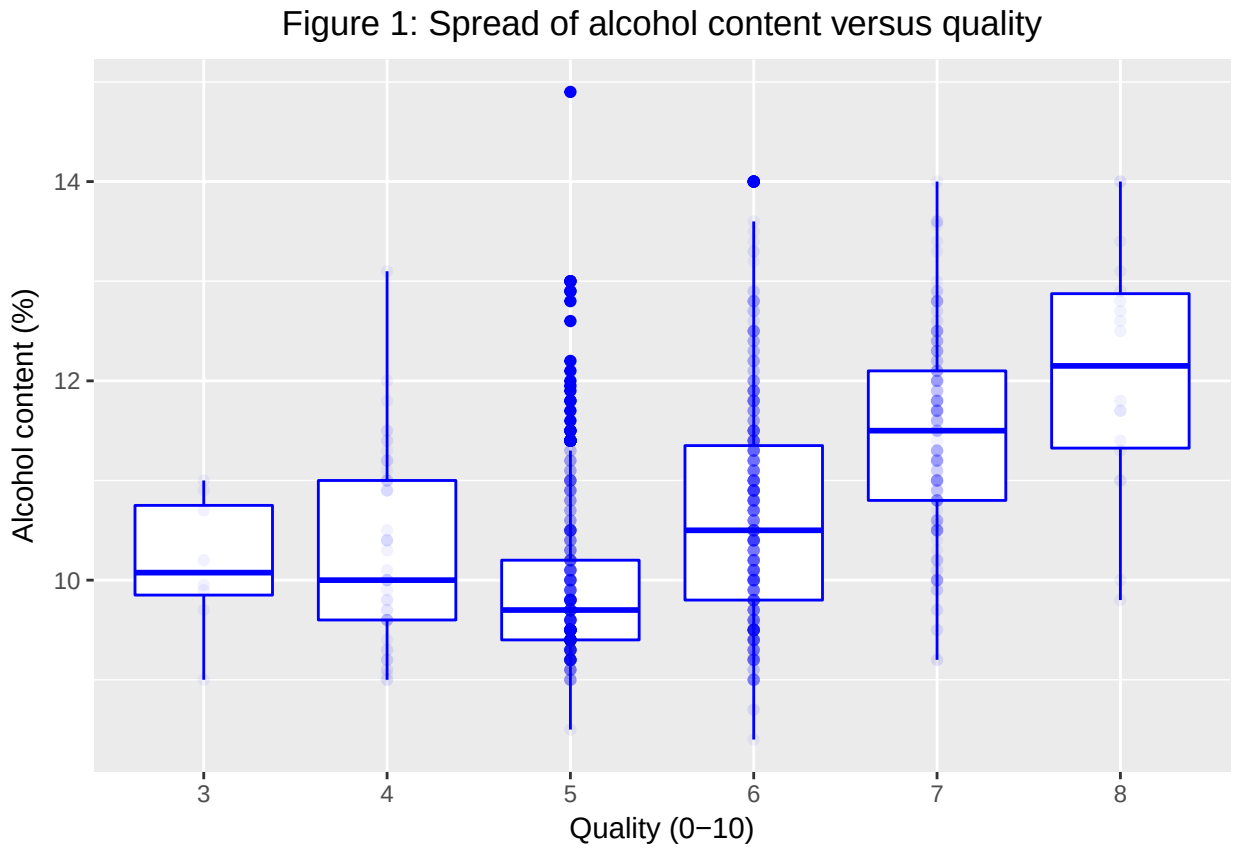
Nothing surprising.

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

I tried a neural network model to classify the wines. However, the accuracy was just above 62%, which is not acceptable. The features of the data appear inadequate for modeling based on this short, preliminary attempt. I should mention that I did not spend a lot of time on the modeling part and I might be mising a more suitable model.

Final Plots and Summary

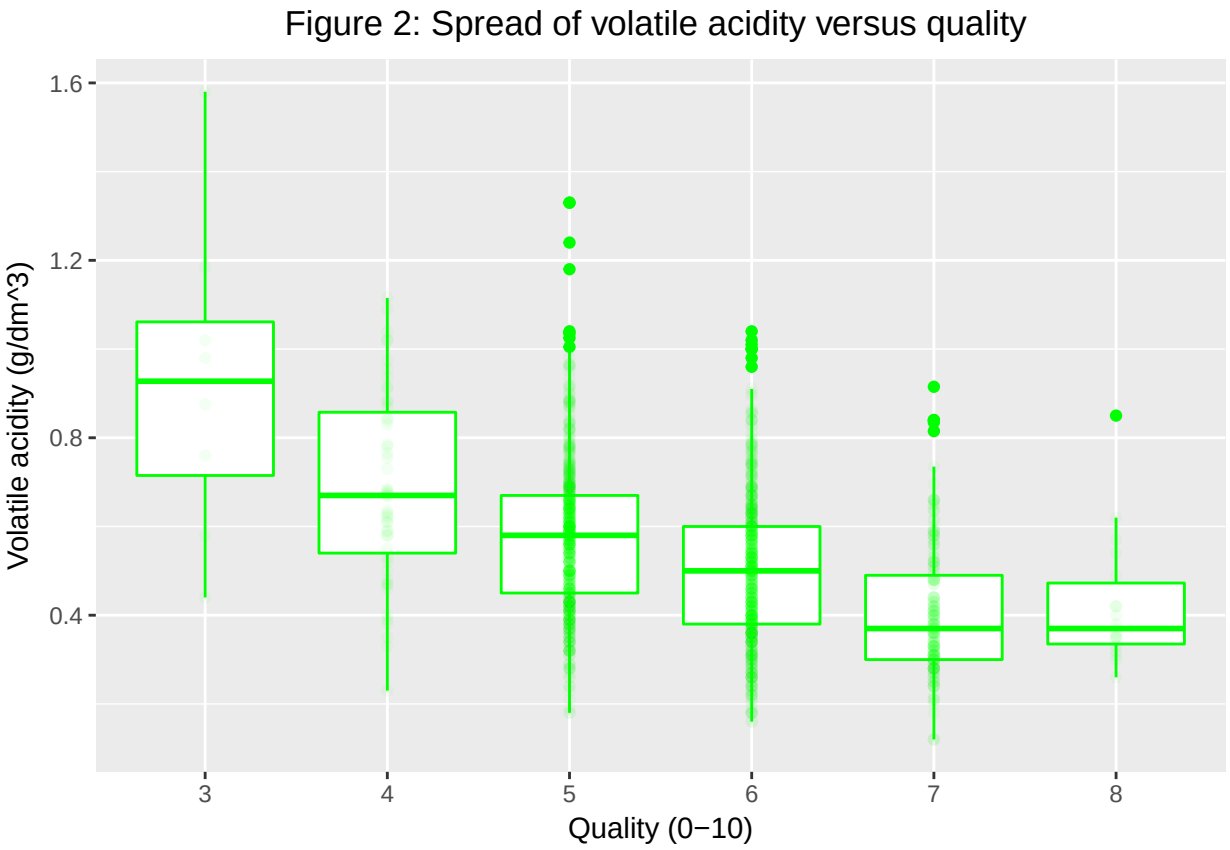
Plot One



Description One

Figure 1 shows the boxplot of alcohol content in each group. The median alcohol content of the low quality (3-4) wines in the database is around 10%. However, the median alcohol content increases linearly for the higher quality (5-8) wines. We can conclude that there is a relationship between quality and alcohol content especially at the high quality end.

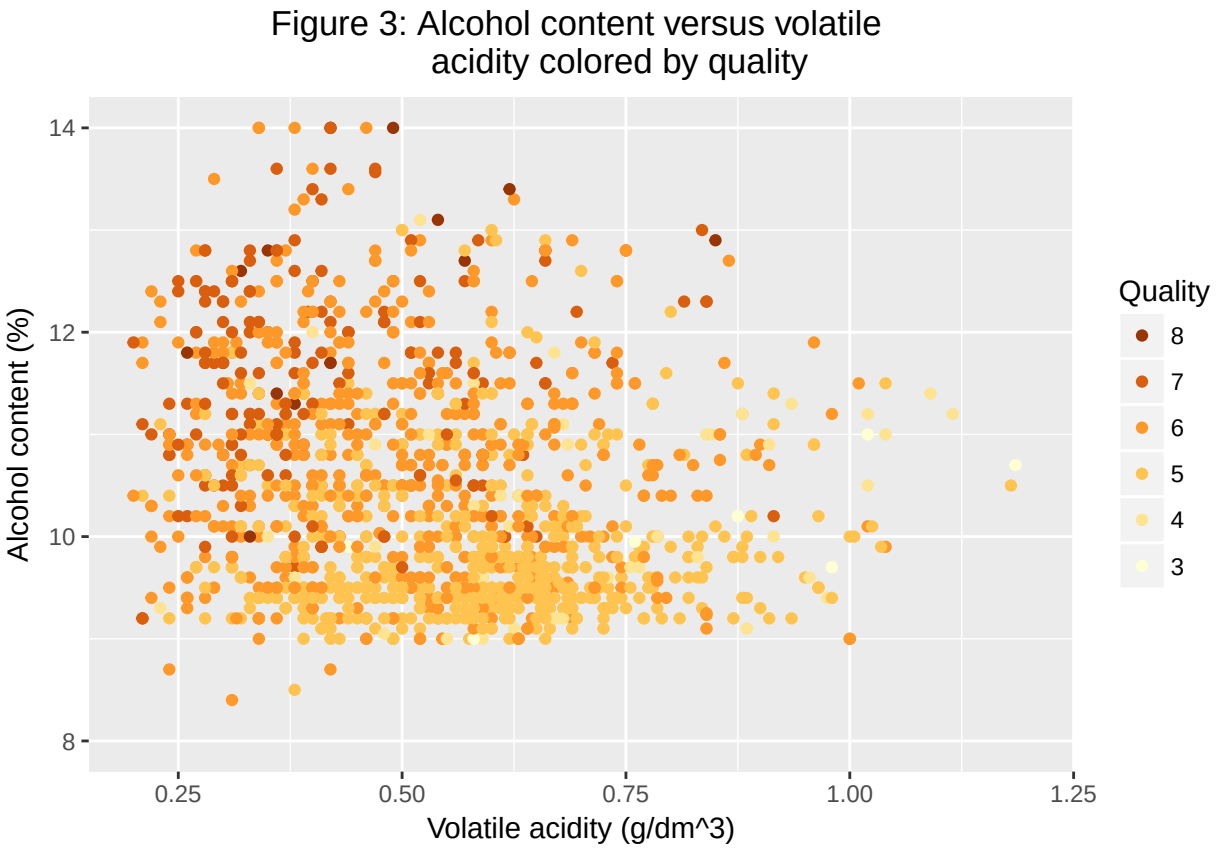
Plot Two



Description Two

There is a negative correlation between volatile acidity and quality of a wine as shown in Figure 2. The median volatile acidity decreases almost linearly with quality.

Plot Three



Description Three

Figure 3 demonstrate the combined effect of volatile acidity and alcohol content on quality. It is visible that most wines with quality of 7 and 8 have more than 10% alcohol content and less than 0.6 g/dmˆ3 volatile acidity. In addition, most wines with quality less than 6 have less than 10% alcohol content.

Reflection

There are many factors that contribute to quality of a wine. In addition, quality can vary significantly due to personal preferences. Although many of the chemicals showed negligible correlation with the wine qulaity, alcohol content, sulphates, and volatile acidity concentrations showed some correlation with wine quality. However, my initial assesment is that many other features contribute to wine quality that are not represented in the dataset. This project was very interesting to me and I understood the gist of an statement I heard sometime, which goes like “A lot of data and a desire to find an answer does not guarantee finding one”. I struggled with finding a good combination of variables to explain the variability of the quality since these were not easily visible in the dataset. I was able to use box plots to reduce variance and see some trends in the data. However, I believe there are missing features that can explain the variability of the quality variable better. My idea is to add some other variables like dryness and sweetness to the dataset. The wine industry already has a method to measure these types of variables and these variables with enrich the dataset.