Red Wine EDA with R

Tania Lincoln

November 12, 2017

Introduction

This project is an exploratory investigation of the quality of red wines. I will explore variables, structures, and patterns in the data. The project is open ended, like a stream of consciousness analysis where questions are posed and answers explored. At this time, machine learning is not a component of the project. In this analysis, I will add a summary and plots to provide a final reflection.

The data set contains ~16,000 red wines with 12 variables of the properties of the wine. Wine experts rated each of the wines from 0 (very bad) to 10 (very excellent).

Import Libraries

Import core libraries used in the analysis.

```r
library("ggplot2")
library("knitr")
library("dplyr")
library("PerformanceAnalytics")
library("corrplot")
library("Hmisc")
library("dplyr")
library("GGally")
```

Load Data

Read the cvs file, assign it to the wine dataset.

```r
wine_org <- read.csv('../../../downloads/wineQualityReds.csv', sep = ',')
```

Examine column names, basics statistics

To gain familiarity with the data and to understand what questions to ask, we'll look at some basic statistics.

We take a look at the column names in the dataset to get an understanding of what it contains.

```
##  [1] "X"                    "fixed.acidity"        "volatile.acidity"    
##  [4] "citric.acid"          "residual.sugar"       "chlorides"           
##  [7] "free.sulfur.dioxide"  "total.sulfur.dioxide" "density"             
## [10] "pH"                   "sulphates"            "alcohol"
## [13] "quality"
```

There appears to be a lot of names I don't understand. We will do some web research to understand these more.

Next we examine the datatypes and values.

```
## '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 ...
```

All of these are numbers. X seems like a row number.

We continue to look at basic statisitcs, examine the ranges and quantiles. We'll look at this more closely later. Note that the top quality rating assigned to a wine is 8, lowest is 3.

```
##        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  
```

Let's look at a few rows to tie all of basic exploration together.

```
##   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
##   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
##   quality
## 1       5
## 2       5
## 3       5
## 4       6
## 5       5
## 6       5
```

There are 12 variables and 1599 observations to examine. 11 of these variable are possible drivers for determining wine quality.

Based on the below definitions, it seem that there are several variables that could be related.

- pH and acid columns
- residual sugar, alcohol, and density
- sulfur and sulphates

Data Definition

- X - row number for each observation
- Fixed Acidity - Measurement of titrable acids and free hydrogen ions.

- Volatile Acidity - Acids produced by microbial action like fermentation. High levels can cause off-flavors or aromas.
- Citric Acid - A type of acid that produces bright citrus flavors.
- Residual Sugar - The amount of remaining sugar (leftover from the fermentation process) in the wine.
- Chlorides - The amount of salt in the wine
- Free Sulfur Dioxide - This acts like an antimicrobial and antioxidant
- Total Sulfur Dioxide - This acts like an antimicrobial and antioxidant
- Density - the meausurement of alcohol and sugar content in relation to water.
- pH - The measurement of the strength of the acid: 0 (acid) - 13 (base). This causes a brightness vs softer taste.
- Sulphates - A wine additive that acts like an antimicrobrial and antioxidant.
- Alcohol - The percent of alcohol in the wine
- Quality - Blind taste testing value assigned to the wine

Basic data cleansing

I've chosen to remove the row number column, x, from the dataset. We take a look at the dataset again to make sure it's removed.

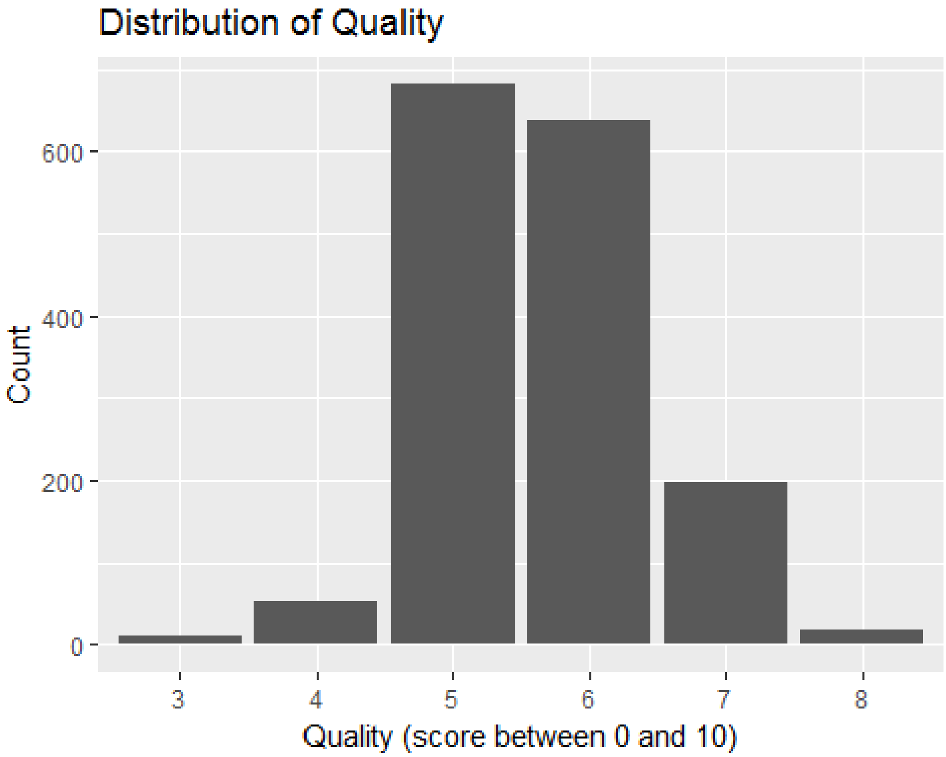
```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.4             0.70        0.00            1.9     0.076
## 2           7.8             0.88        0.00            2.6     0.098
## 3           7.8             0.76        0.04            2.3     0.092
## 4          11.2             0.28        0.56            1.9     0.075
## 5           7.4             0.70        0.00            1.9     0.076
## 6           7.4             0.66        0.00            1.8     0.075
##   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
##   quality
## 1       5
## 2       5
## 3       5
## 4       6
## 5       5
## 6       5
```

Basic Visualizations of single data elements (univariate anlaysis)

Based on the descriptions above, let's look at the variables that seem to contribute to taste: Volatile acidity, Citric Acid, Residual Sugar, Chlorides, and Quality. I think these would determine wine taste quality. Let's look at quality first to get understanding of how many wines and the ratings that were assigned.

Quality

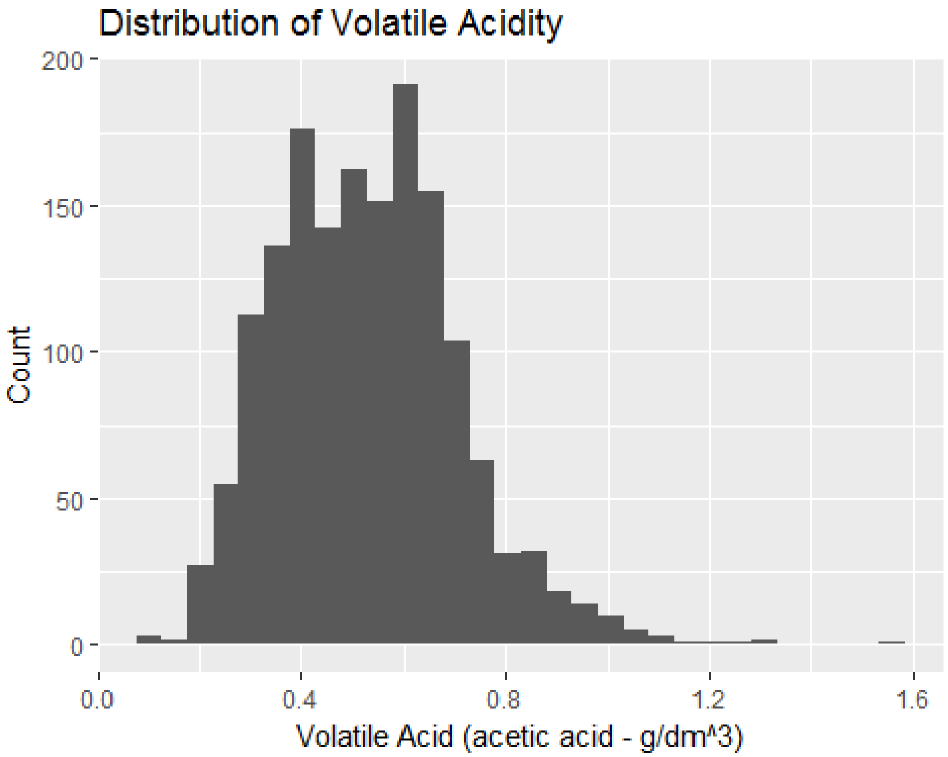
Here we will look at how the wines are distributed across Look at wine quality.



There are not very many bad or good wines. Nothing exceeds 8. There seem to be a lot of middle player wines and very few bad wines or good wines. I think fewer numbers of the extremes in our dataset might make it harder to find a definitive answer.

Volatile Acidity

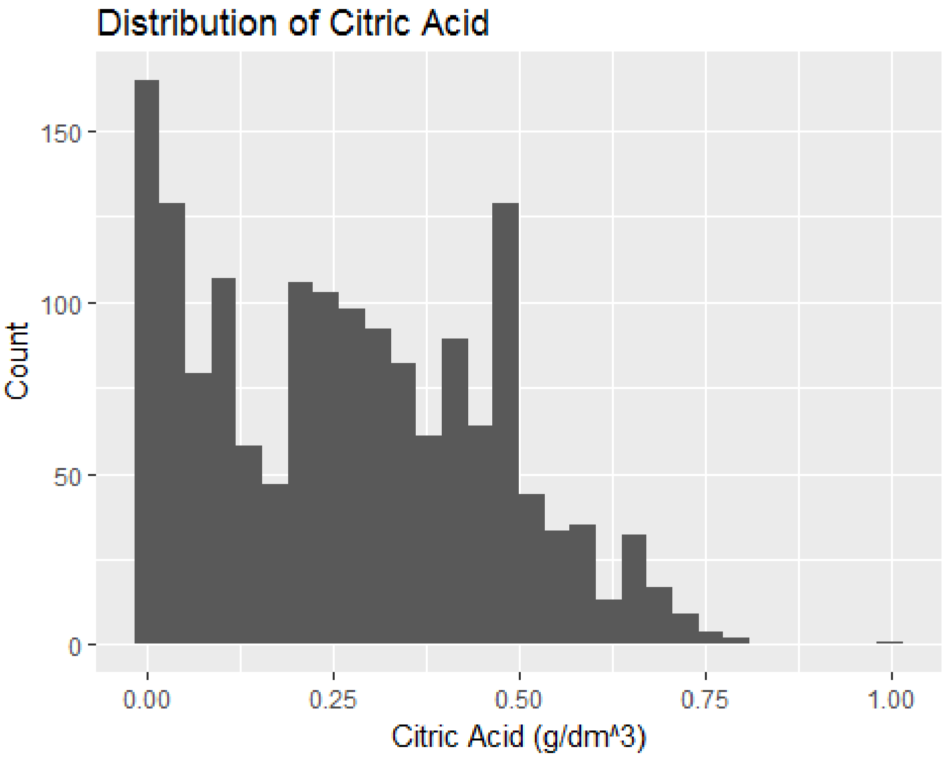
Next we take a look at volatile acid to see if it follows a similar pattern. It's the first of the factors that impact flavor.



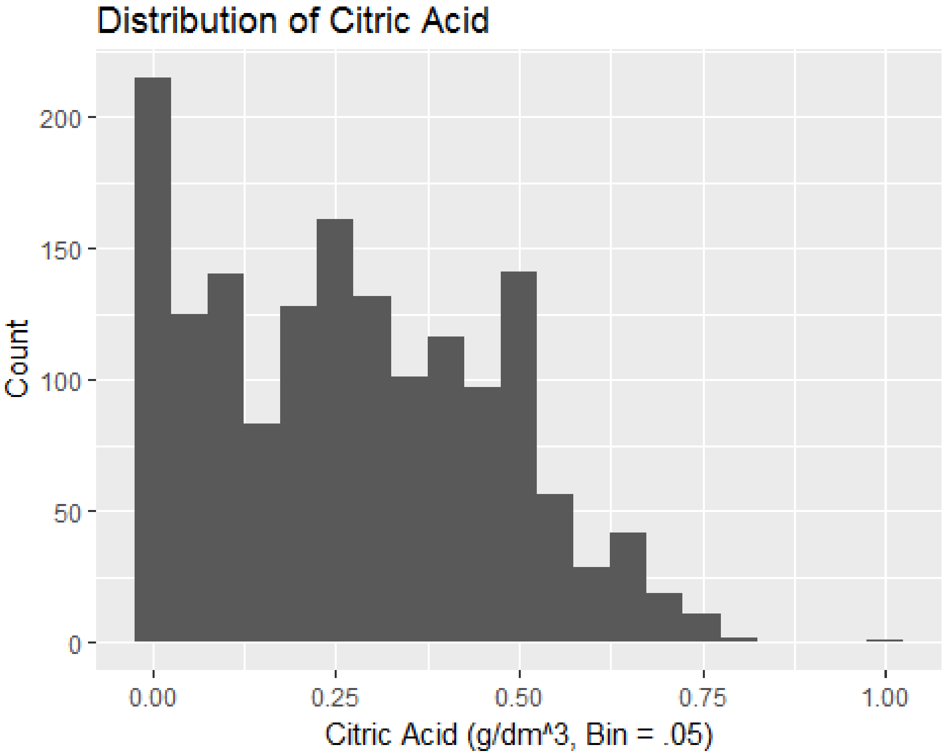
There is a drop a pretty sharp drop off around .7. I think low count representation at these numbers align with the definition that volatile acidity adds off flavors to the wine. A lot of wines seem to fall in the .6 bin.

Citric Acid

Here is the distribution for citric acid. I'm curious to see if it follows a similar pattern.



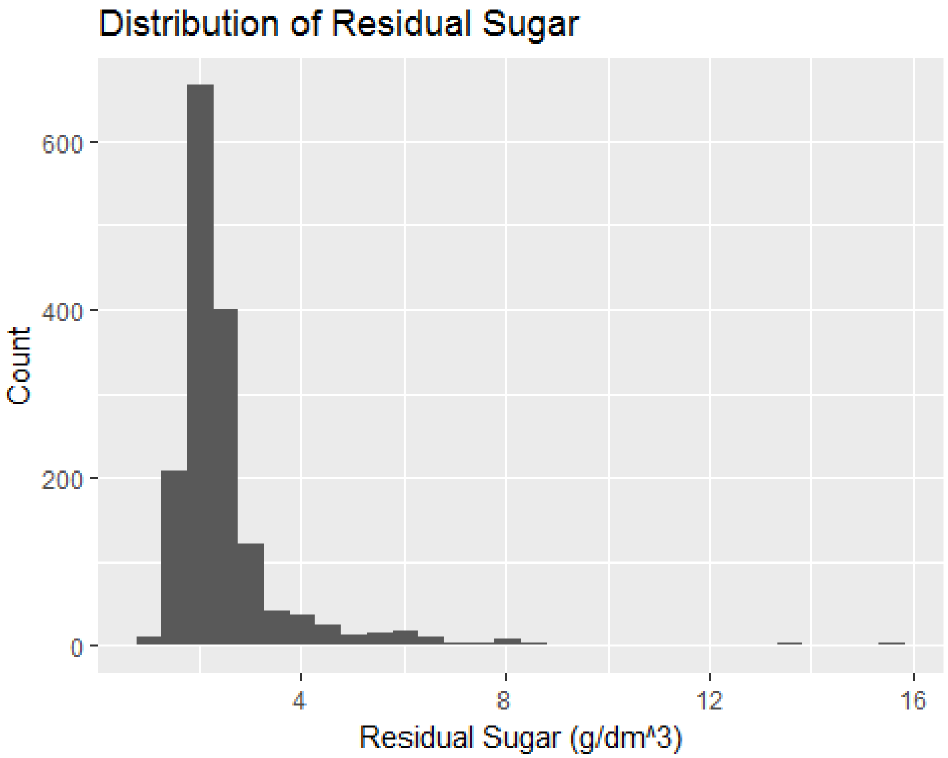
I'm not sure what this means, the distribution seems really varied. We'll try playing with the bin size to .05.



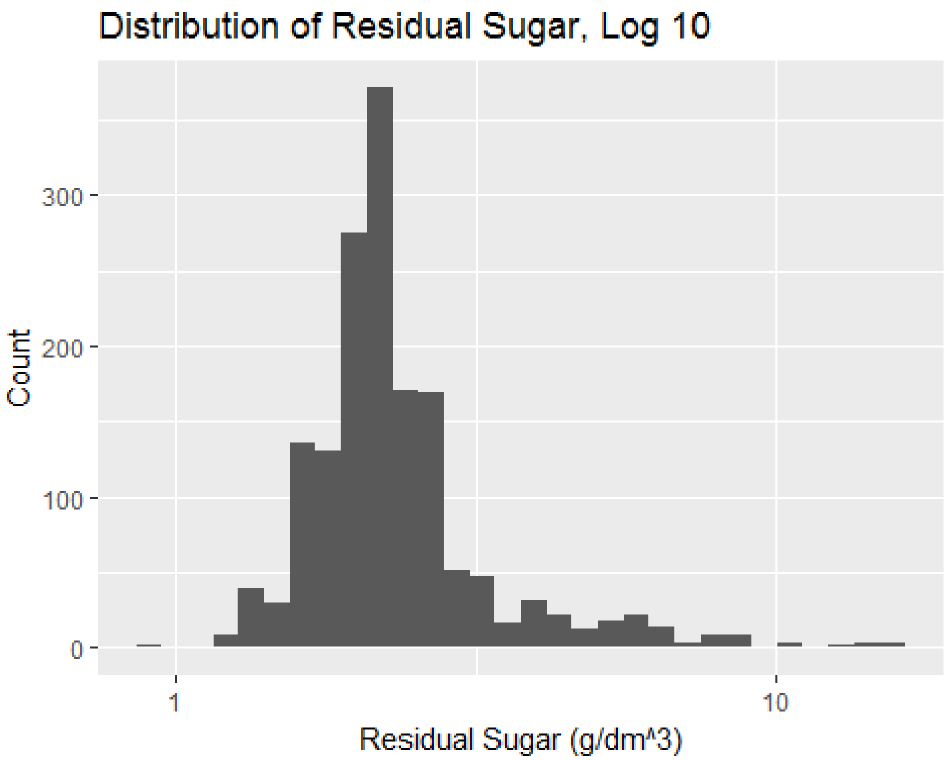
The bin size change doesn't seem significant.

Residual Sugar

Here we take a closer look at the residual sugar.



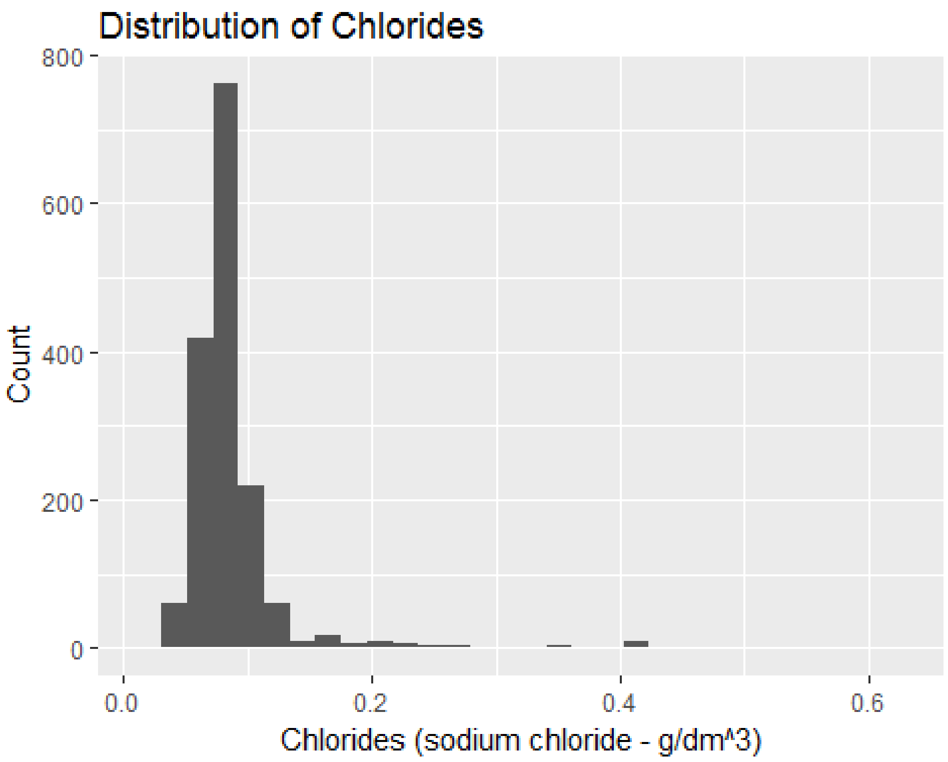
With there being such a positive skew, we will try plotting this again using a log 10 scale.



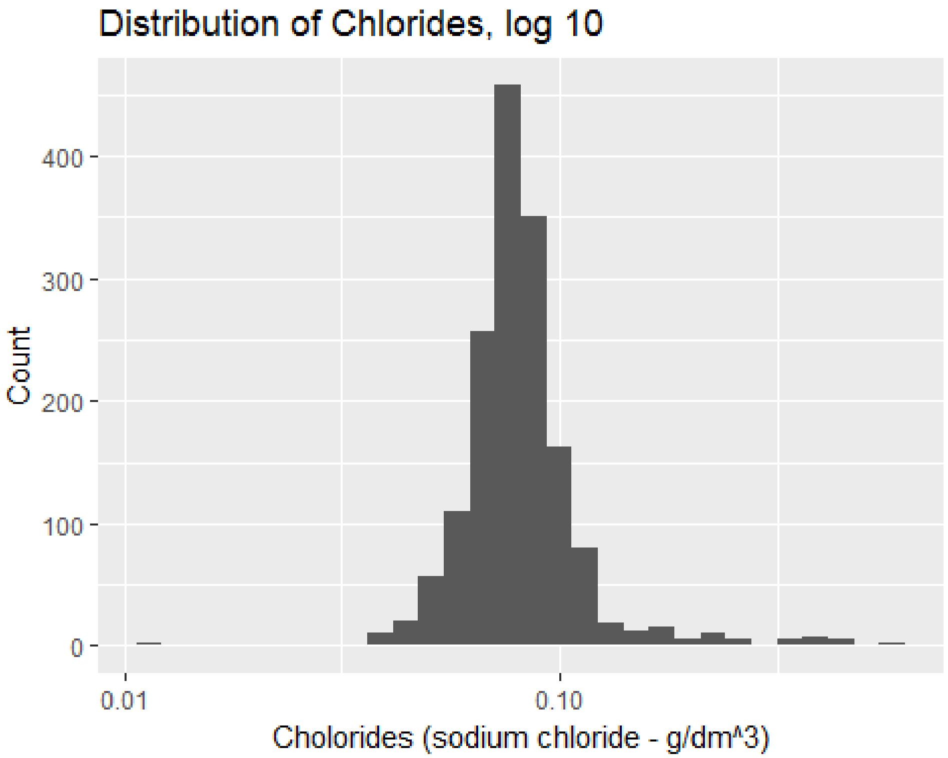
This distribution looks more normalized.

Chlorides

Here the distribution for chlorides seems similar to sugar, positively skewed and several outliers.



We take a look at the same data using the log_10 overlay.



This chart appears to be more normalized as well.

I think looking at these variables individually helps with understanding what's in the data, but it doesn't seem very useful in determining what makes a good wine. Next, we'll take a look at a couple of variables together.

Exploring Basic Questions (bivariate analysis)

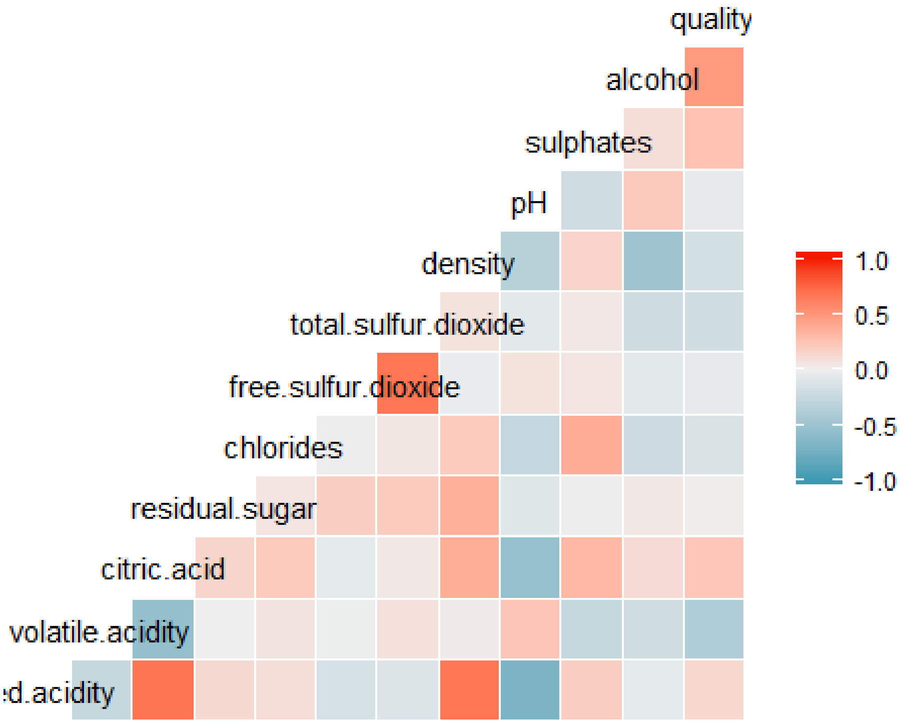
Based on some of the column descriptions, several variables could be related to eachother. We'll examine these more closely to see if that's true and if these combinations help determine wine quality.

- pH and acid columns
- residual sugar, alcohol, and density
- sulfur and sulphates variables

We look at a correlation matrix to see if these columns have a relationship to each other that is statistically significant.

```
##                      fixed.acidity volatile.acidity citric.acid
## fixed.acidity             1.000000        -0.256131    0.671703
## volatile.acidity         -0.256131         1.000000   -0.552496
## citric.acid               0.671703        -0.552496    1.000000
## residual.sugar            0.114777         0.001918    0.143577
## chlorides                 0.093705         0.061298    0.203823
## free.sulfur.dioxide      -0.153794        -0.010504   -0.060978
## total.sulfur.dioxide     -0.113181         0.076470    0.035533
## density                   0.668047         0.022026    0.364947
## pH                       -0.682978         0.234937   -0.541904
## sulphates                 0.183006        -0.260987    0.312770
## alcohol                  -0.061668        -0.202288    0.109903
## quality                   0.124052        -0.390558    0.226373
##                      residual.sugar chlorides free.sulfur.dioxide
## fixed.acidity              0.114777  0.093705           -0.153794
## volatile.acidity           0.001918  0.061298           -0.010504
## citric.acid                0.143577  0.203823           -0.060978
## residual.sugar             1.000000  0.055610            0.187049
## chlorides                  0.055610  1.000000            0.005562
## free.sulfur.dioxide        0.187049  0.005562            1.000000
## total.sulfur.dioxide       0.203028  0.047400            0.667666
## density                    0.355283  0.200632           -0.021946
## pH                        -0.085652 -0.265026            0.070377
## sulphates                  0.005527  0.371260            0.051658
## alcohol                    0.042075 -0.221141           -0.069408
## quality                    0.013732 -0.128907           -0.050656
##                      total.sulfur.dioxide   density        pH sulphates
## fixed.acidity                   -0.113181  0.668047 -0.682978  0.183006
## volatile.acidity                 0.076470  0.022026  0.234937 -0.260987
## citric.acid                      0.035533  0.364947 -0.541904  0.312770
## residual.sugar                   0.203028  0.355283 -0.085652  0.005527
## chlorides                        0.047400  0.200632 -0.265026  0.371260
## free.sulfur.dioxide              0.667666 -0.021946  0.070377  0.051658
## total.sulfur.dioxide             1.000000  0.071269 -0.066495  0.042947
## density                          0.071269  1.000000 -0.341699  0.148506
## pH                              -0.066495 -0.341699  1.000000 -0.196648
## sulphates                        0.042947  0.148506 -0.196648  1.000000
## alcohol                         -0.205654 -0.496180  0.205633  0.093595
## quality                         -0.185100 -0.174919 -0.057731  0.251397
##                          alcohol   quality
## fixed.acidity        -0.061668  0.124052
## volatile.acidity     -0.202288 -0.390558
## citric.acid           0.109903  0.226373
## residual.sugar        0.042075  0.013732
## chlorides            -0.221141 -0.128907
## free.sulfur.dioxide  -0.069408 -0.050656
## total.sulfur.dioxide -0.205654 -0.185100
## density              -0.496180 -0.174919
## pH                    0.205633 -0.057731
## sulphates             0.093595  0.251397
## alcohol               1.000000  0.476166
## quality               0.476166  1.000000
```

This is hard to read, it's a lot of text. We'll use a visual variation of the correlation chart, we can also see where there is a statistical significance.



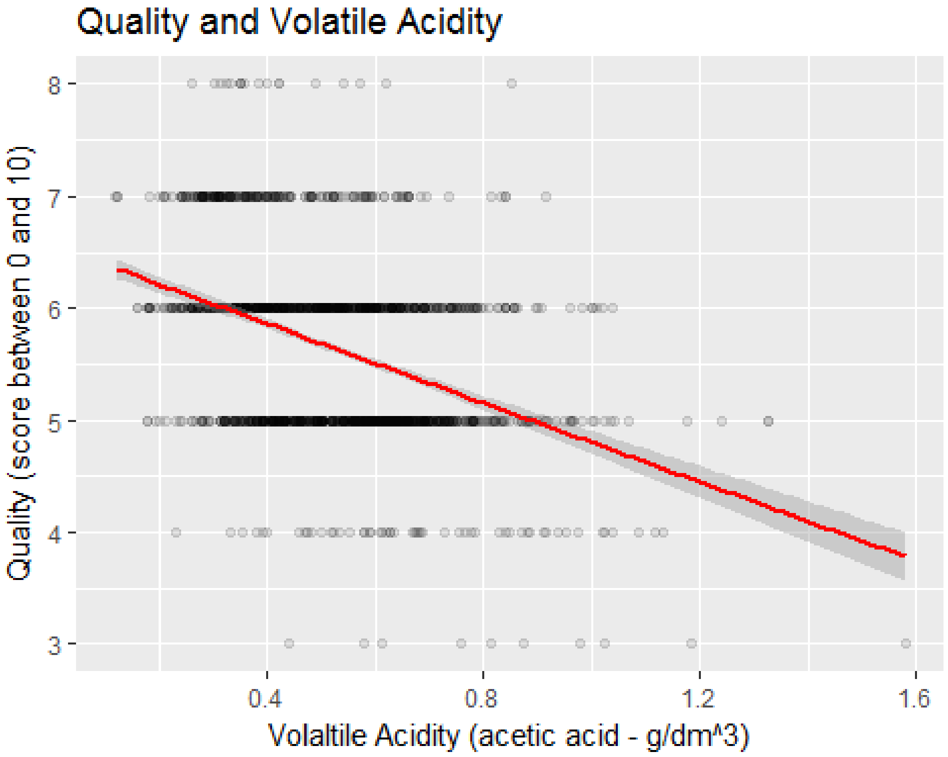
Strong statistical significance is seen in the following correlations:

- fixed acidity --> citric acid, density, pH.
- volitile acidity --> citric acid, quality

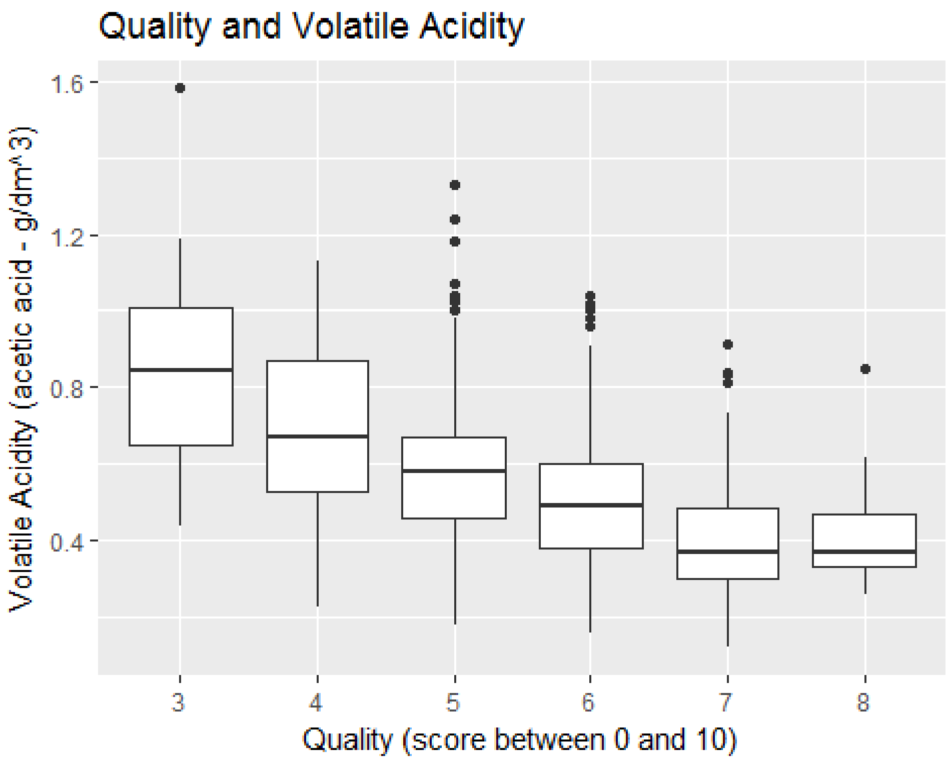
- citric acid --> volitile acid, fixed acid, pH
- free sulfur dioxide --> total sulfur dioxide
- density --> fixed acidity, pH, alcohol
- pH --> fixed acidity, citric acid, density
- alcohol --> density, quality
- quality --> alchohol, volitile acid

We were mostly correct about related variables except for *all* acids significantly contributing to pH; residual sugar impacting density; Sulfur and Sulphates are not related. I was not suprised to see that volatile acids was a high predictor since this can cause off-flavors. I was surprised to see alcohol as a high predictor for wine quality.

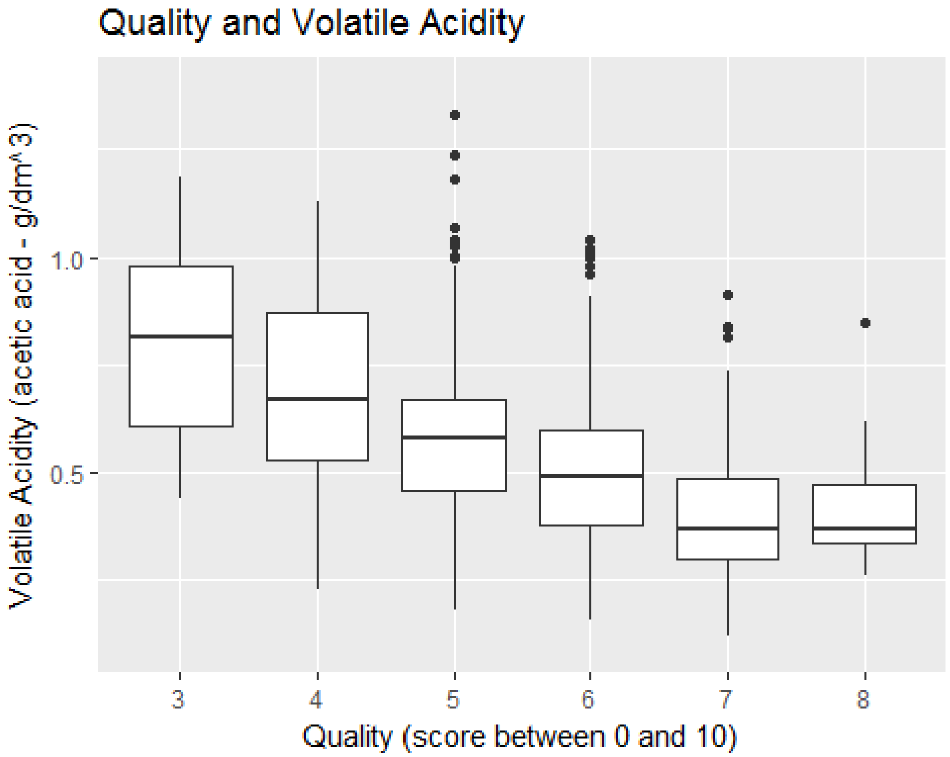
Let's try to determine if quality decreases and volatile acidity increase.



Wines are densely populated between .2 - .8 and show a decrease in quality I'll try using a box plot next to see if it's easier to understand.

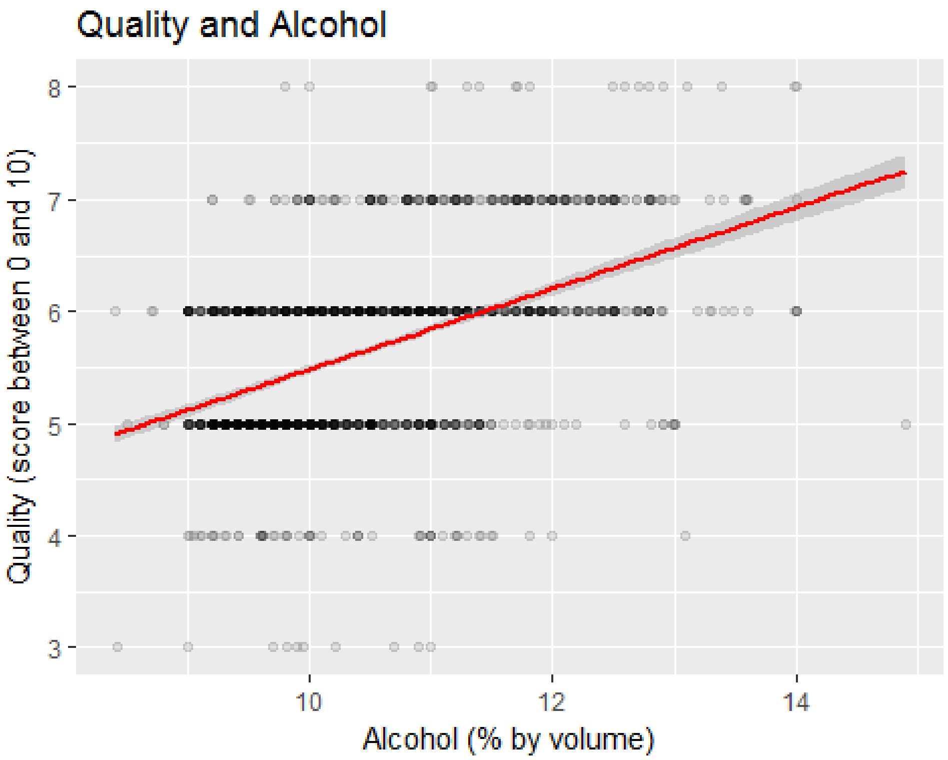


This ia a lot cleaner. You can see that as volatile acidity declines qualitiy goes up. There are several outliers. Let's plot it again without the point at 1.6



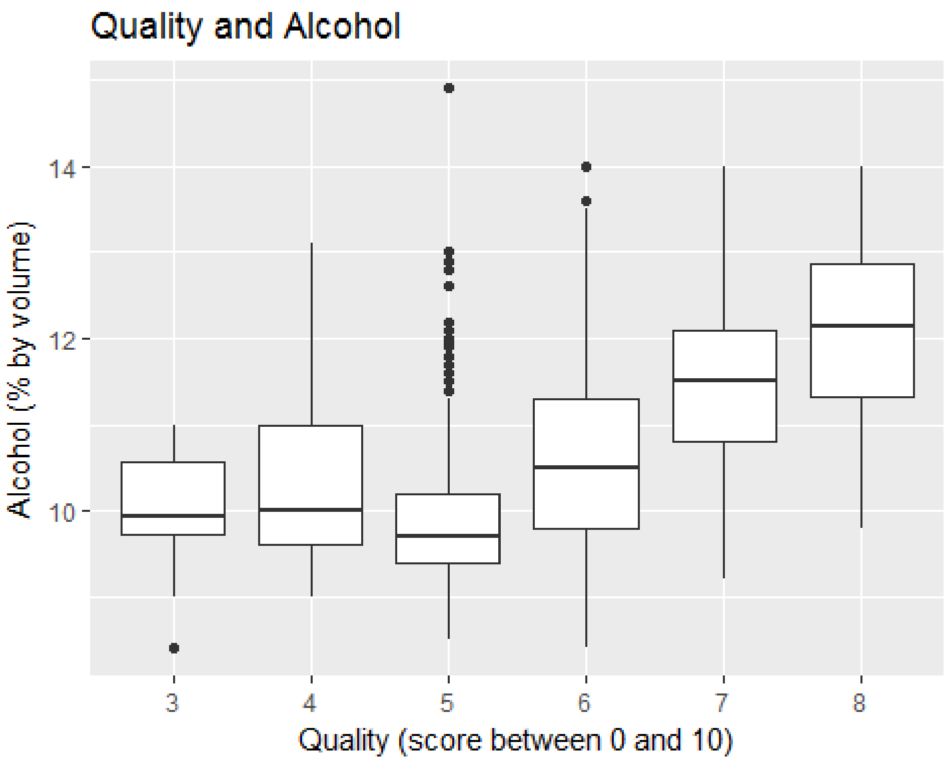
This plot shows that the range of volatile acid is wider in the lower quality wines. However, there are several outliers in the better quality wines. This must mean that there are other factors involved in wine quality.

Next, let's look at Wine quality and whether it increases as alcohol content increases. I think this one is interesting and funny!

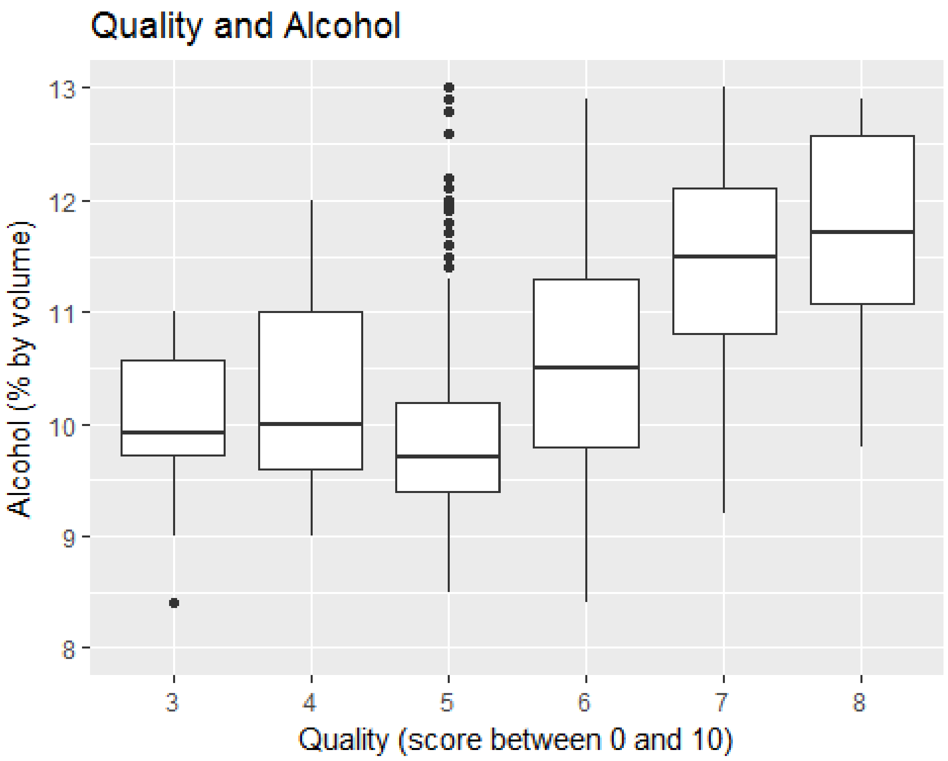


It does seem to increase in quality. The plot is very dense at 5 and 6, but that's because we have so many wines in these ratings.

I think a boxplot would help us understand how the data is being distributed with alcohol and quality.

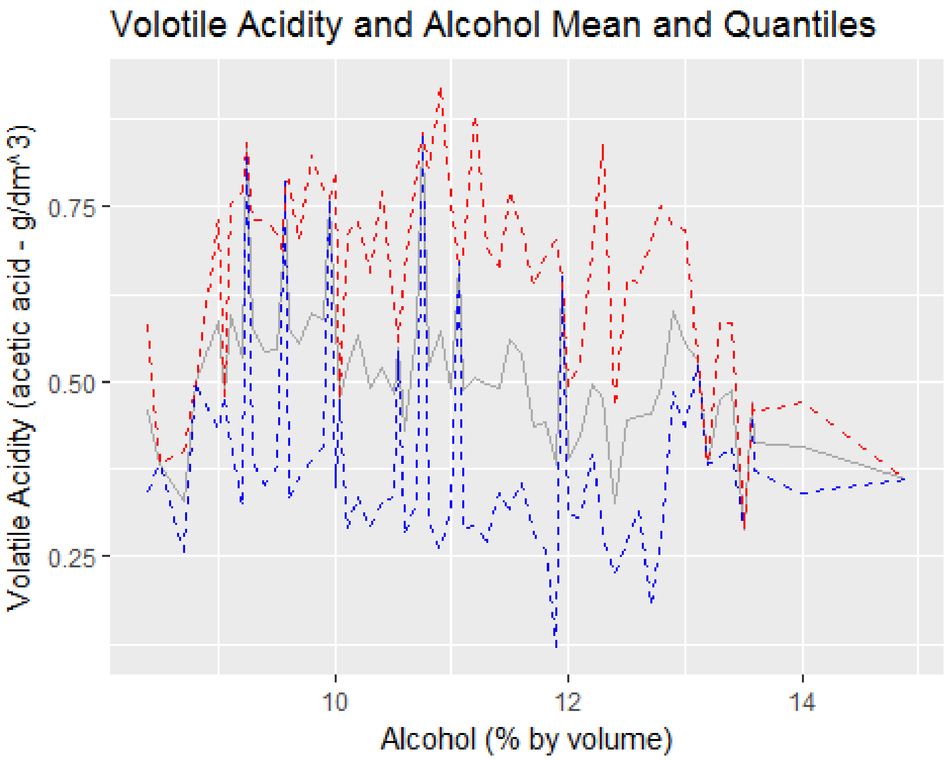


It seems that quality increases when alcohol increases around 6 and up. This could be hard to see with the outliers, so we will plot again with some of those removed.



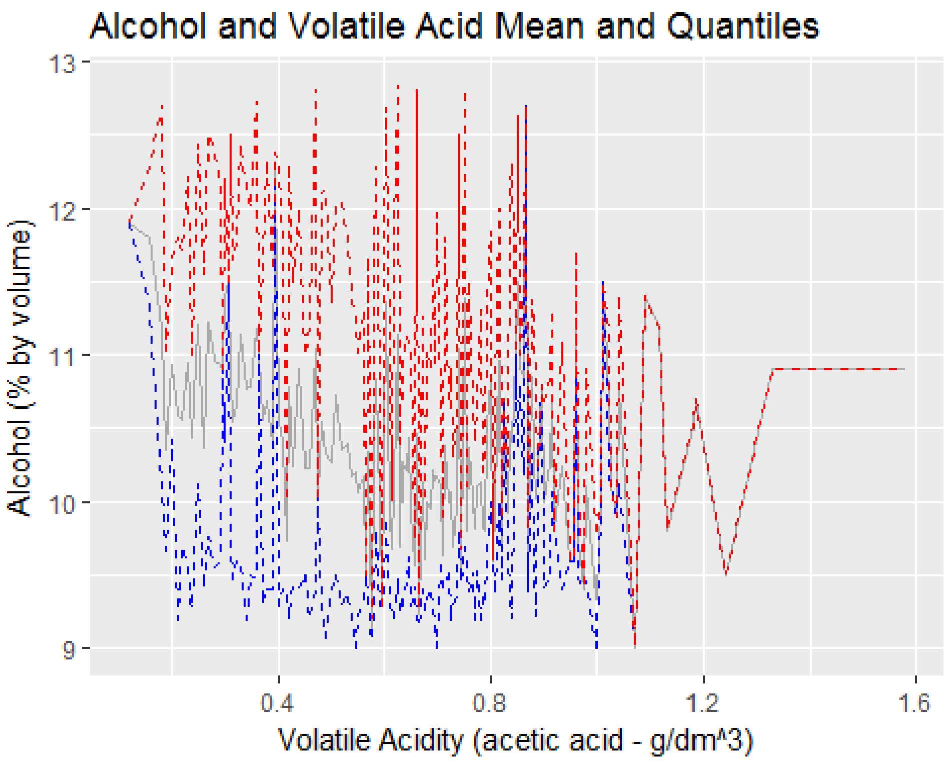
The ranges seem more varied than I expected. 5s have a lot of outliers and there is a dip in the mean. Maybe this means that volatile acid and alcohol work togther in wine quality or there are other contributing factors.

We could plot the mean, lower and uppper quantiles of alcohol and volatile acidity in a line graph to see if this could show us the a relatioinship between these two variables.



It's interesting to look at, but pretty difficult for me to undestand.

In the next plot we switch axis to see what that looks any better.



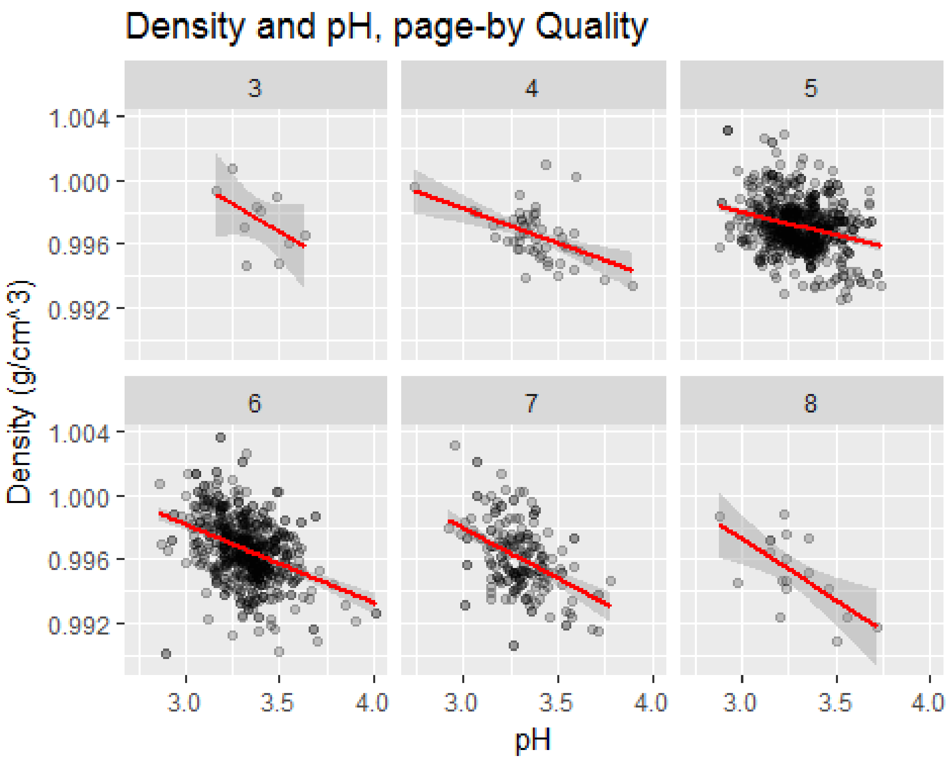
This is even harder to understand. I don't think either of these plots were helpful.

Exploring more advanced questions (multivariate analysis)

It's clear that wine quality declines with the presence of more volatile acid. With alchohol being more varied, let's see if other factors that influence alcohol or vice versa contribute to wine quality. Of these, I think volitile acide, density, and pH are insteresting since they show the highest statistical significance to alcohol or quality.

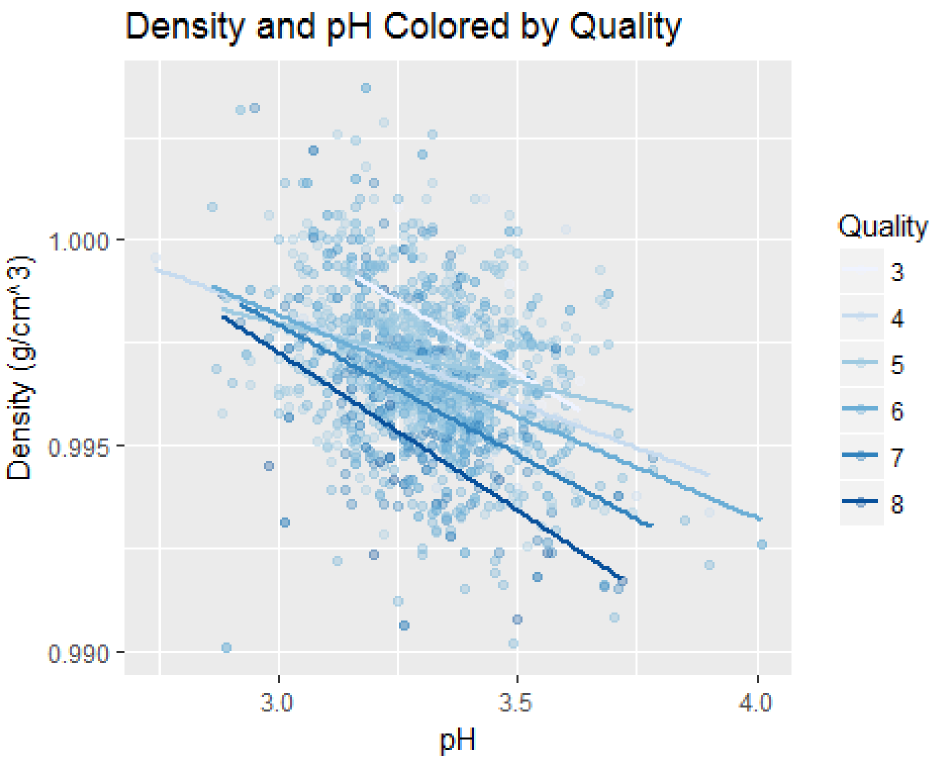
pH and Density

Let's look at pH and density together. Does it differ with quality?



It's interesting to see how things are clustered, but this seems hard to understand. I think we learn that there are a lot of wines that are 5s and 6s, which isn't new.

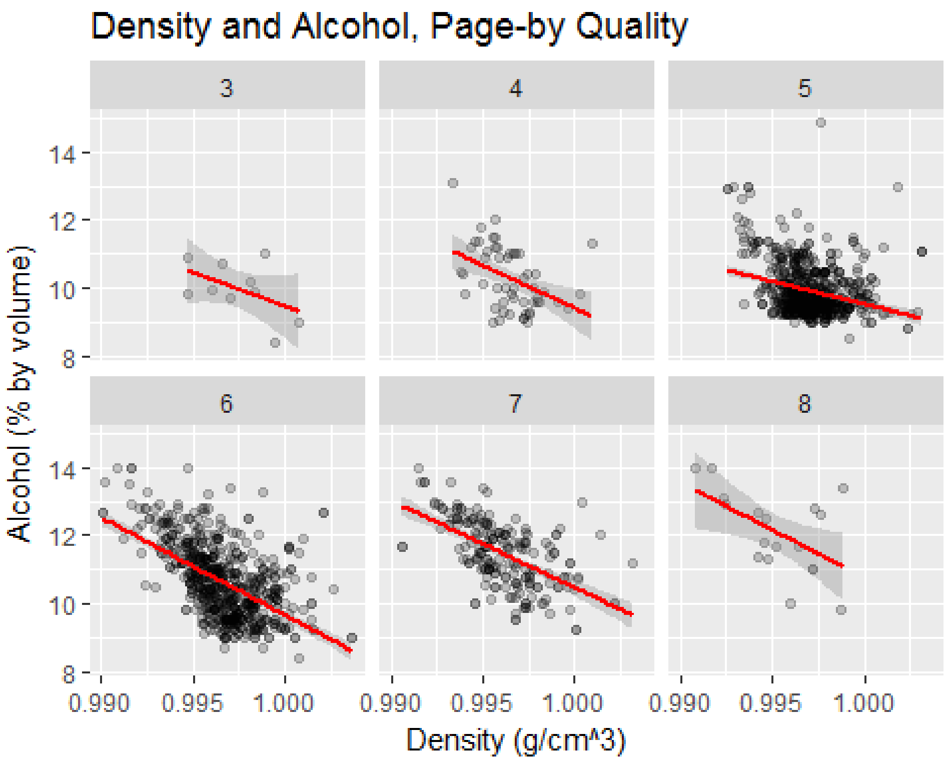
In the next plot, we try adding all quality factors to the grid to better compare these together.



In this graph, it's easier to see that as density decreases, pH increases and that best quality wines have lower density values.

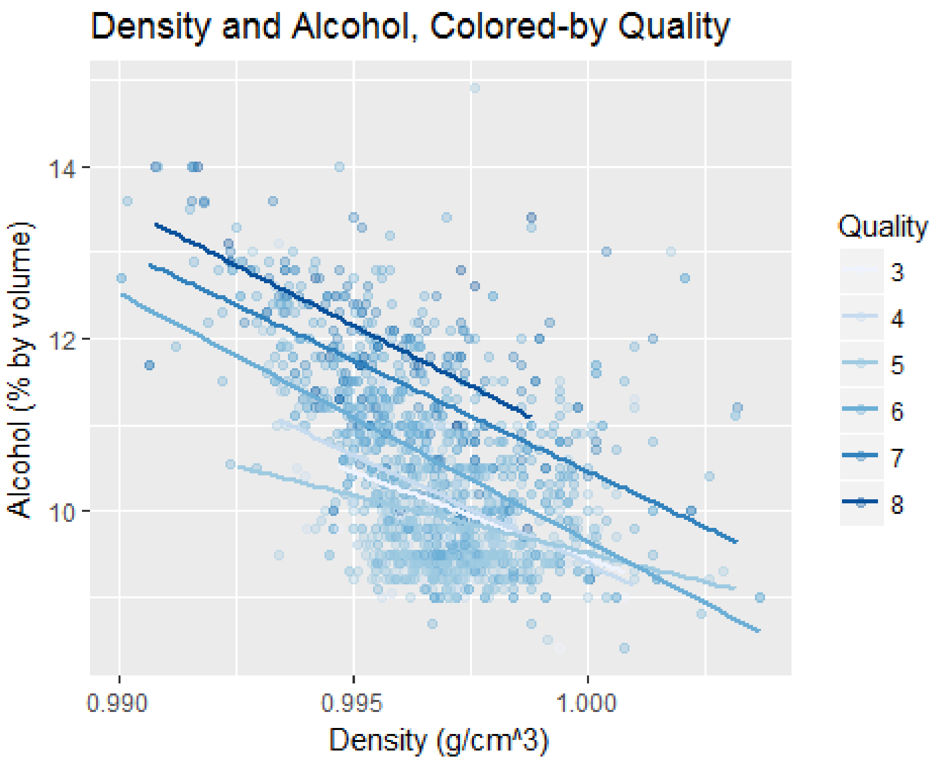
Alcohol and Density

Next, let's look at alcohol and density by quality.



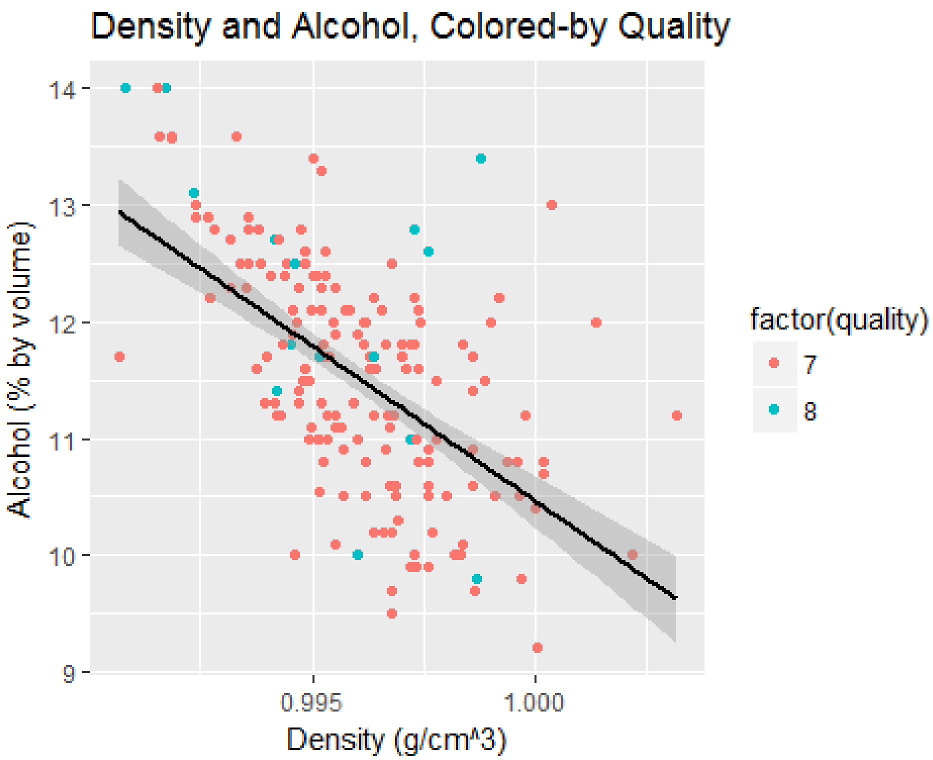
As alcohol increases, density increases 5s have clusterd strongly below the regression line of the graph for alcohol content.

Next, we look at all wines in one graph with a quality color.



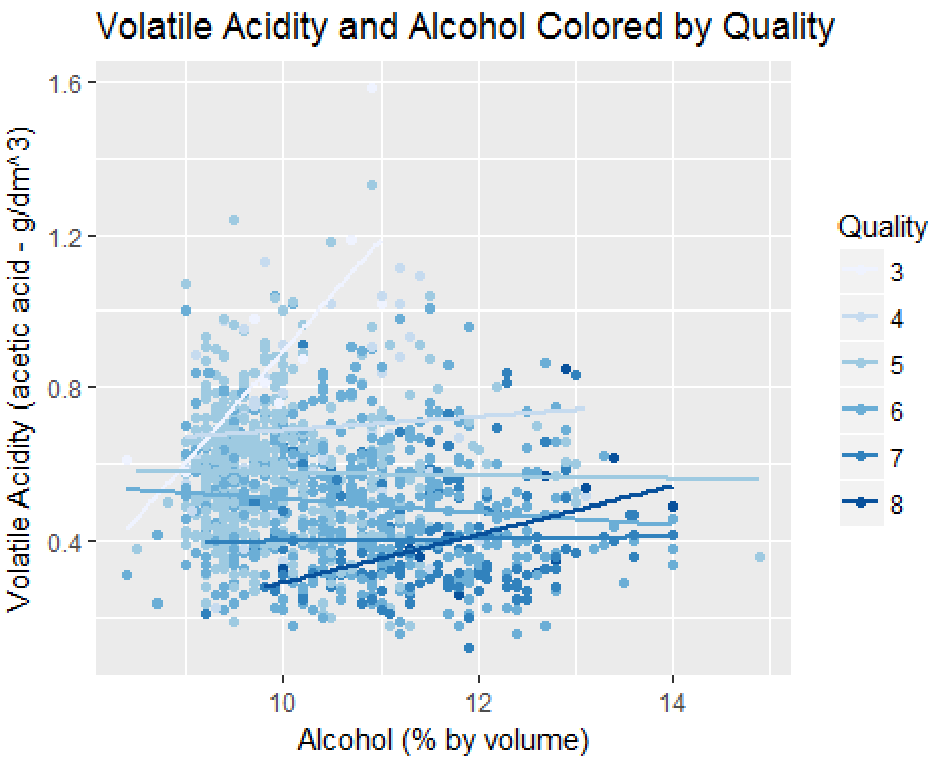
Wines of higher quality have more alcohol content which we learned earlier in the correlation graph and the bi-variate analysis. We can also see that density for the higher quality wines max outs below 1.

Maybe with focus on the good wines in the next chart we can see significant details.



Nothing jumps out to me in this chart that wasn't in the chart above.

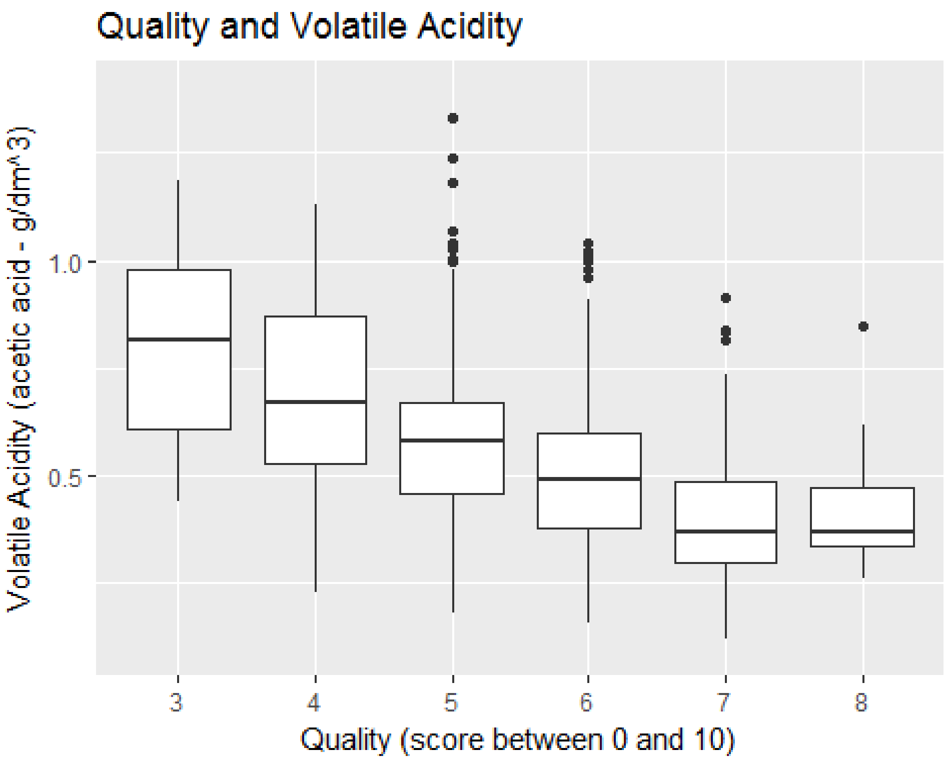
Let's take a look at volaltile acid and alchohol together in a different way than we did before.



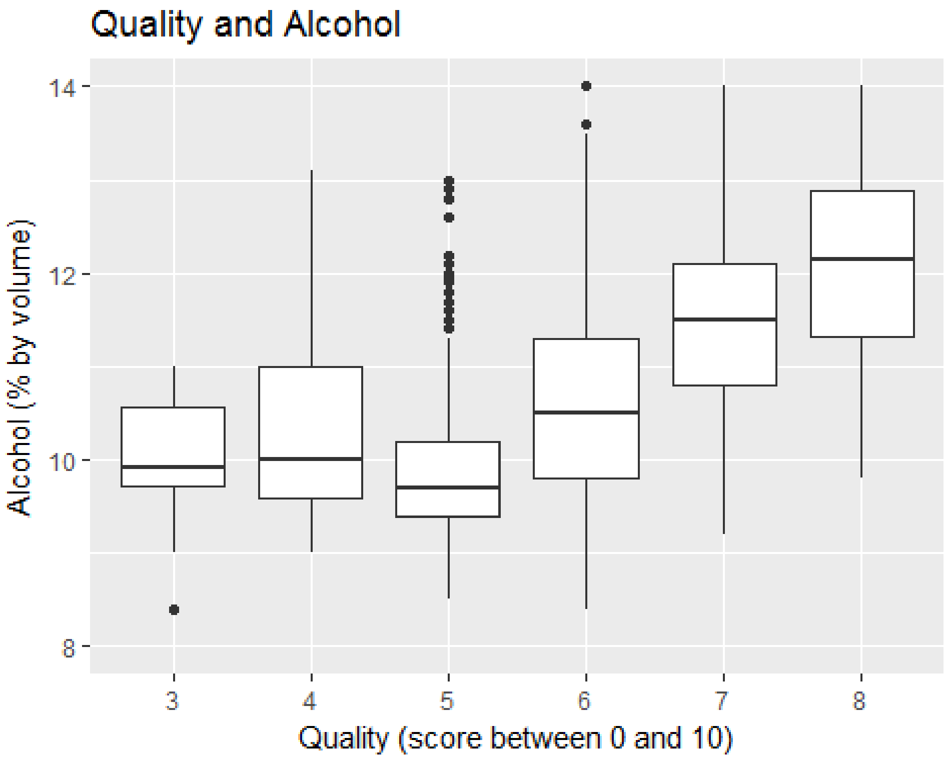
Volatile acidity and alcohol quantity are strong contributors to wine quality. Looking at them plotted together, it seems that lower quality wines are plotted on the left hand side and more quality wines are plotted to the lower right. There are some outliers on both axis.

Final Plots and Summary

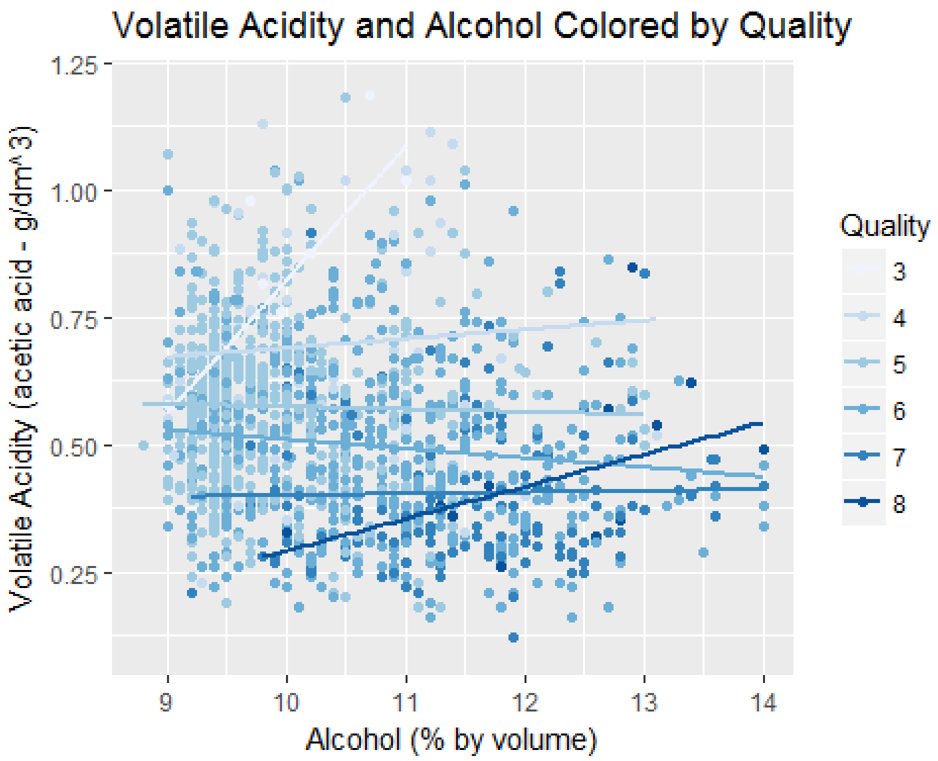
It is clear that wine quality increases when there is a decrease in volatile acidity. From the wine makers manual, we learned that volatile acidity can produce off flavors like vinegar and humans are sensitive to this taste.



This plot shows how wine quality increases when there is more alcohol present in wines of quality 6 and above. I don't know if alcohol imparts a different taste or if this is just expectation of alcoholic drinks.



Volatile acidity and alcohol quantity are strong contributors to wine quality when we look at them plotted together. We have strong representation of 5s and 6s which are also more left and middle, vertically. Most of the regression lines are fairly flat, except for 3's and 8's, which I thought was interesting.



Reflection

Even though volatile acidity and alcohol are statistically significant to wine quality, I wasn't able to see this easily with the data. I would have liked more data points of higher quality wines to see if a better pattern emerged. The data analysis could have used variable reduction, there appeared to be a lot of variables that were correlated to eachother. I think this made the analysis inaccurate and caused some factors to have a stronger sway in the data.

Acknowlegments

For data definitions:

- Practical Winery
- Waterhouse UC Davis
- Wine Maker Academy

For R tutorials:

- R Tutorials
- R Markdown

- Filtering dataframe
- Coloring Plot Points