Data Analyst Nanodegree Project 4: Red Wine Quality Evaluation

Submitted by S. Teodorovich

Introduction

I enjoy red wine but have no idea how to properly determine what makes a wine good or not other than the purely subjective application of "I don't know if this is good wine, but I like it." With that in mind, I took this dataset in the hope that it might give me a bit more insight into the actual physical properties that help determine wine quality.

The dataset was downloaded from https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityReds.csv.

We begin by loading necessary packages & libraries, followed by reading the Data:

Section 1: Univariate Plots

A good first step is always to do get a brief summary of the data set. I'll start with the variable names, followed by a summary of the variable data types, and last with an example of the actual data in tabular form:

```
##  [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"
```

We see that the first variable is called *X*, which is just an identifier and not actually relevant data. With that in mind, we can eliminate it from the dataset:

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

X Has been removed. Now to go ahead and run the rest of the dataset summaries:

```
## 'data.frame':    1599 obs. of  12 variables:
##  $ 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 ...
```

```
##   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
```

The first thing I see is that both Free and Total Sulphur Dioxide are listed as "num" yet in the sample it seems they are "int". It might be a good idea to check if this is correct or not:

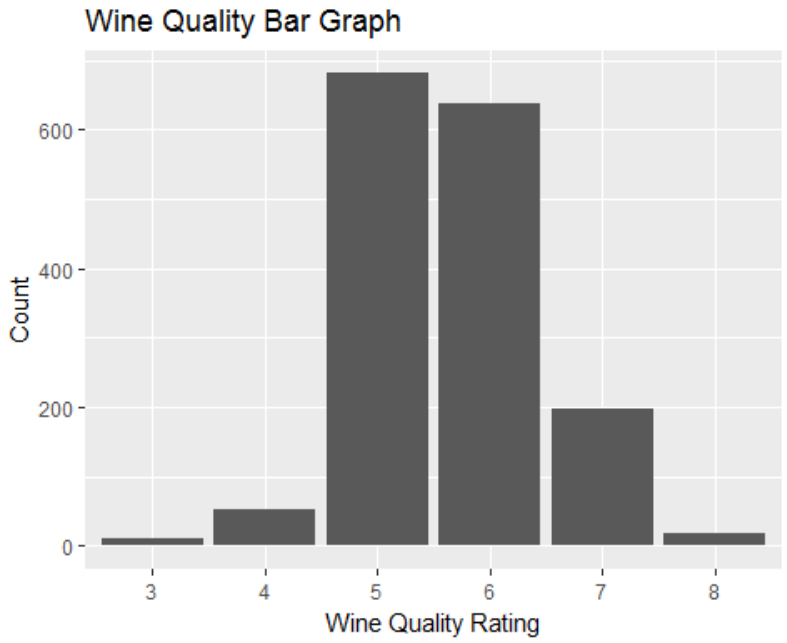
```
## [1] 40.5  5.5 37.5 37.5
```

```
## [1] 77.5 77.5
```

It seems that these are indeed numbers.

At this point we should begin to look at the variables in a bit more detail by creating histograms to check their distribution and shape.

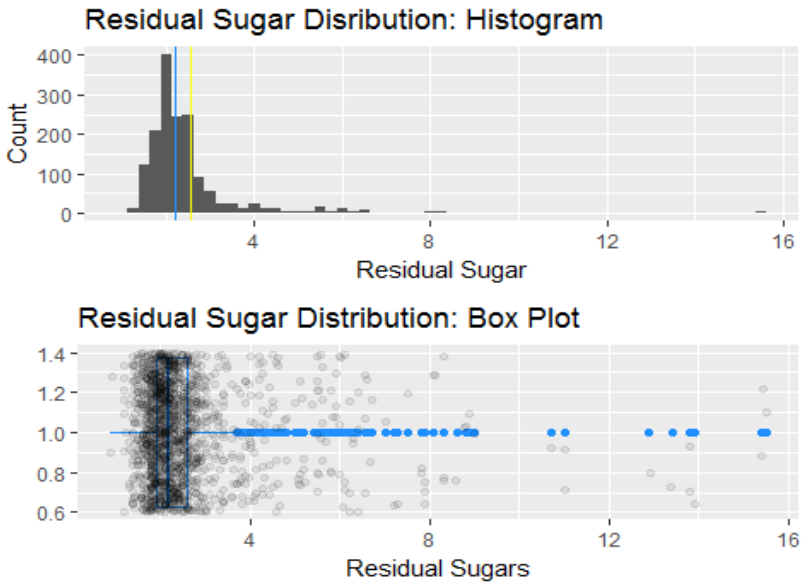
First, we'll give quality a lookover to see just how these wines stack up:




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.636   6.000   8.000
```

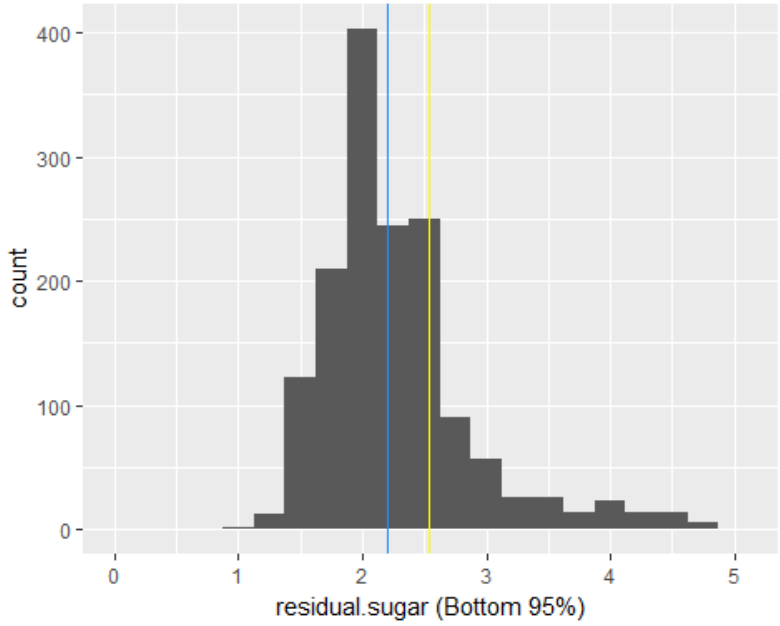
Wine quality is scored on a scale of 1 - 10, but all of our wines fall in the range of 3 - 8, suggesting that none of these wines were either exceptionally good or bad. Further, the distribution does somewhat approximate a normal curve, which is expected as one would think that there would be relatively few good or bad wines compared to medium quality.

Now let's start looking at the variables that describe the physio-chemical components of the wine, and which will serve as the focus of our efforts to determine what influences quality. We'll begin with Residual Sugar:



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```

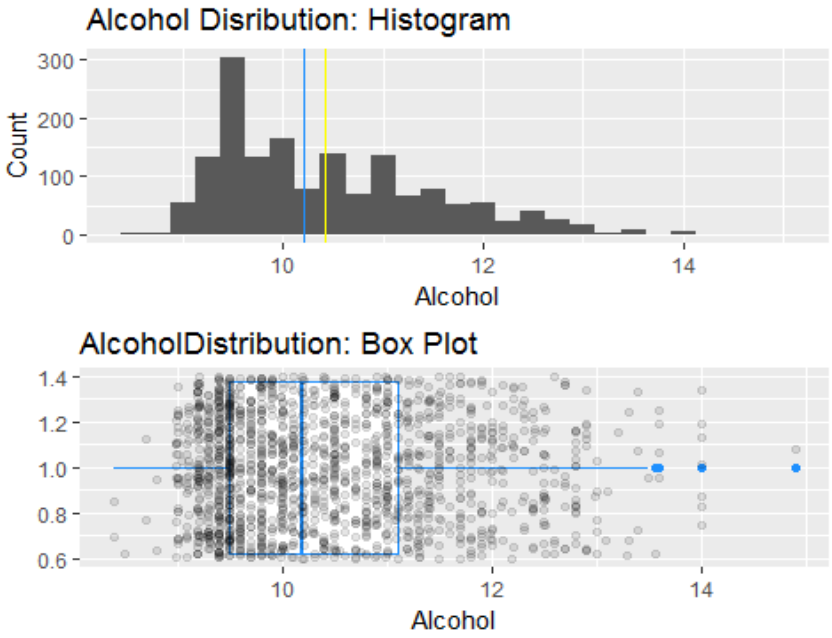
Now that is a positively skewed distribution with a very long tail. It may be more helpful to chop this a bit to remove some of the extreme outliers and take a peek at only the bottom 95%




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.274   2.500   5.000
```

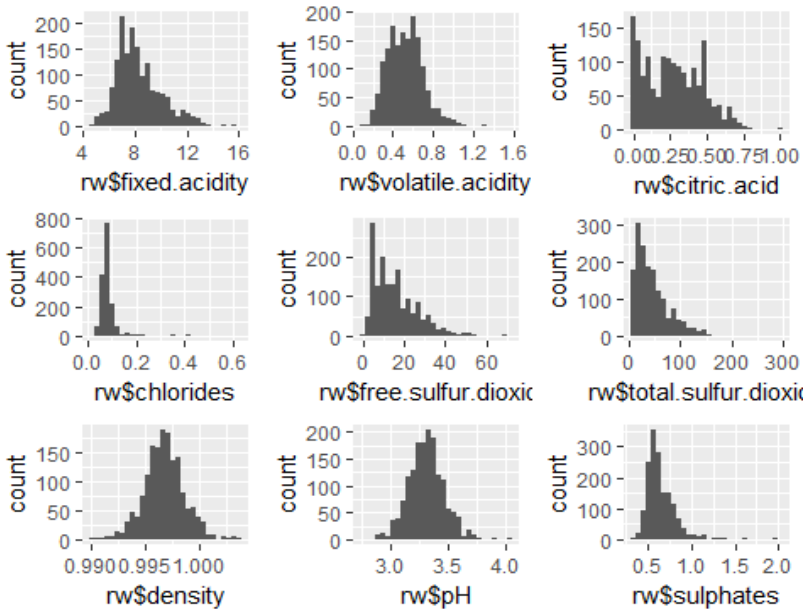
This is a much nicer distribution.

How about we take a look at Alcohol next:



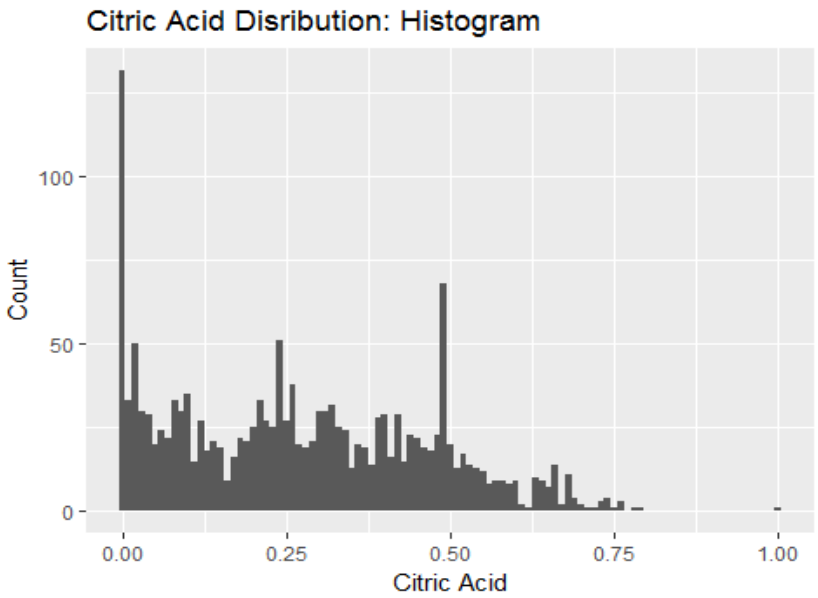
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```

Alcohol also shows a positive skew, but not nearly as pronounced as that for Residual Sugar. Since we have nine more elements to review, it might be easier to set them up in a 3x3 grid for convenience:

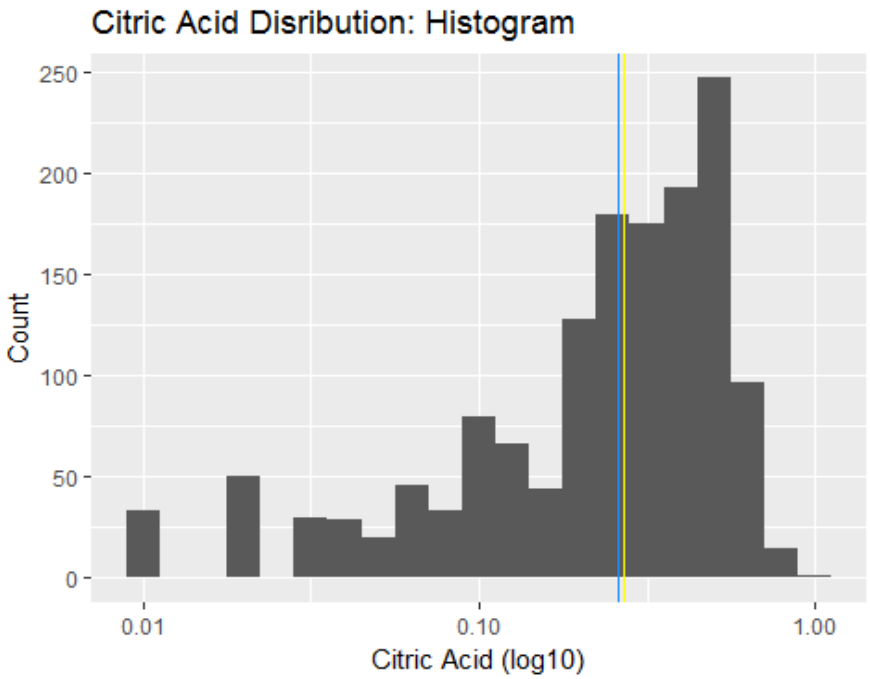


We see that Density and pH both have what appear to be normal distribution, while the rest show some degree of positive skew, particularly Chlorides, which is not only highly skewed, but also has a very narrow range other than the long tail. Citric Acid seems to be out of whack, too, with almost a bi-modal distribution with a positive skew thrown in for fun. We'll take a closer look at both of these. starting with Citric Acid

We'll begin by adjusting bin width to a smaller range:



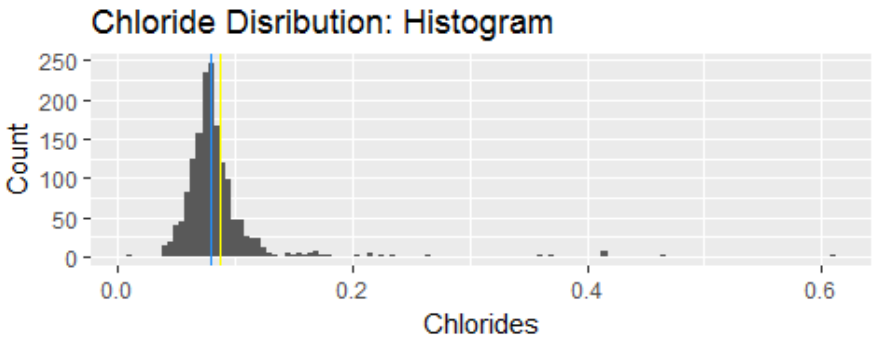
There are an awful lot of wines with 0 g/L of Citric Acid, with spikes at around 0.02, 0.24, and 0.49. Maybe adjusting this further by changing the concentration to a logarithmic scale and expanding the binwidth by a bit:

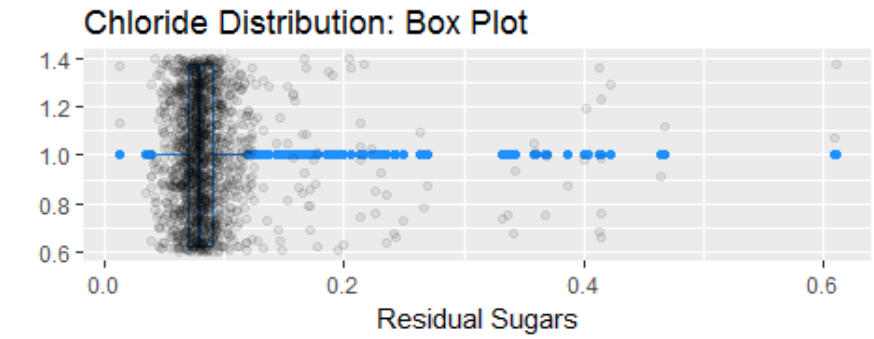



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```

with a log scale it seems like this is now a negative skew. There's definitely something odd going on here, but we'll move on for now.

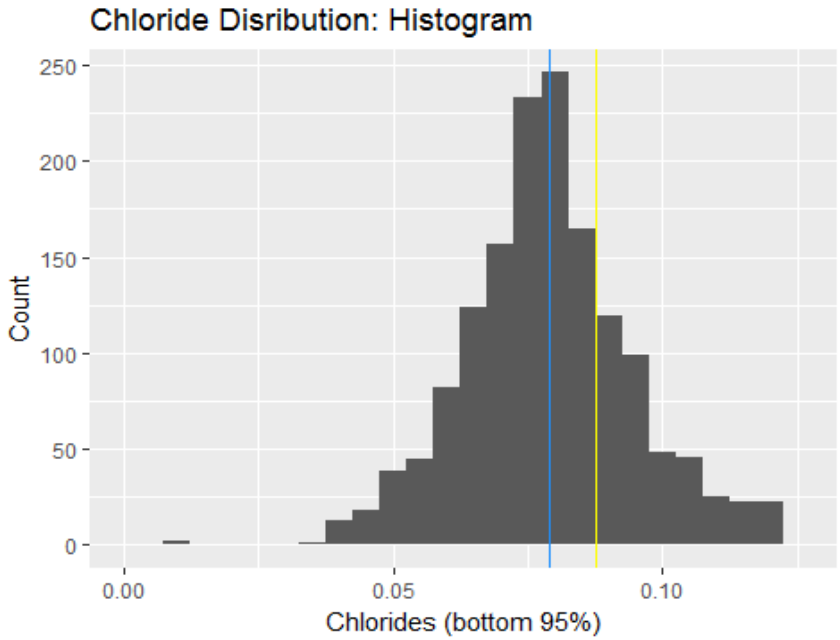
Next we'll start digging into Chlorides a bit more to see what may be happening, because it there are quite a few of these trailing way out there. Let's start by getting an idea of the extent of the outliers in the distribution:





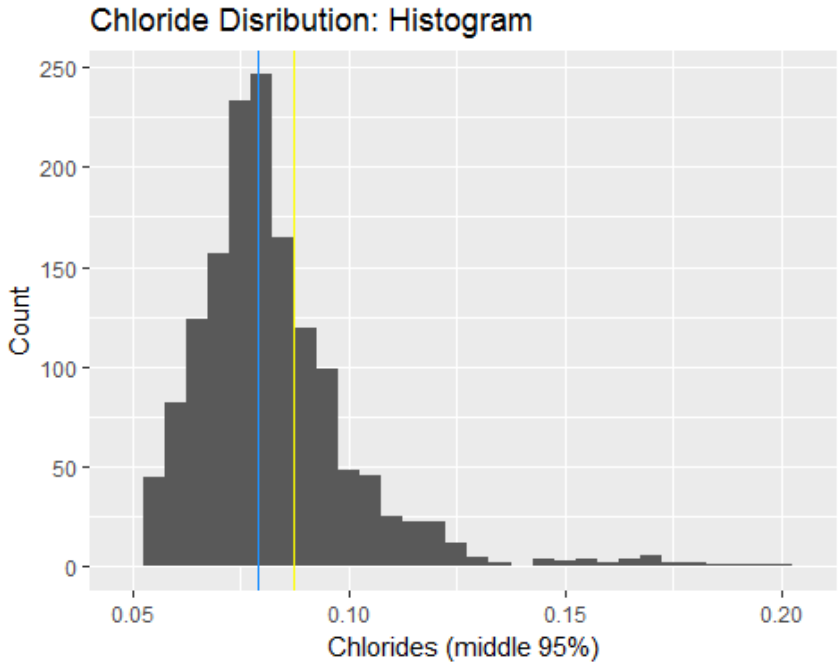

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```

Okay, that's a considerable amount of outliers for what otherwise does seem like a normal distribution. At this point it's probably ab good idea to take just the bottom 95%:



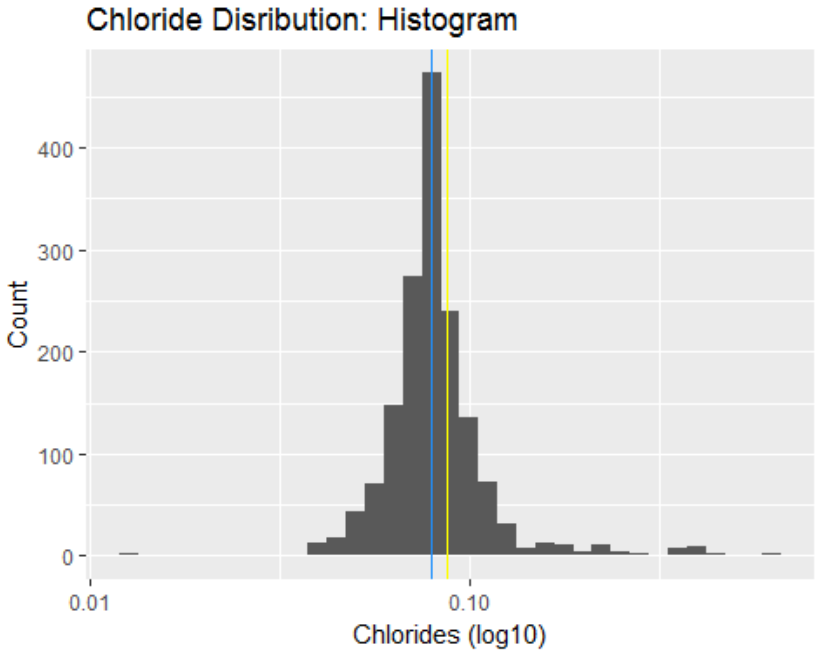

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07800 0.07914 0.08800 0.12600
```

Well, it seems there's a lone sample out on the far low end, so maybe changing this to the middle 95% is a better idea:




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07800 0.07914 0.08800 0.12600
```

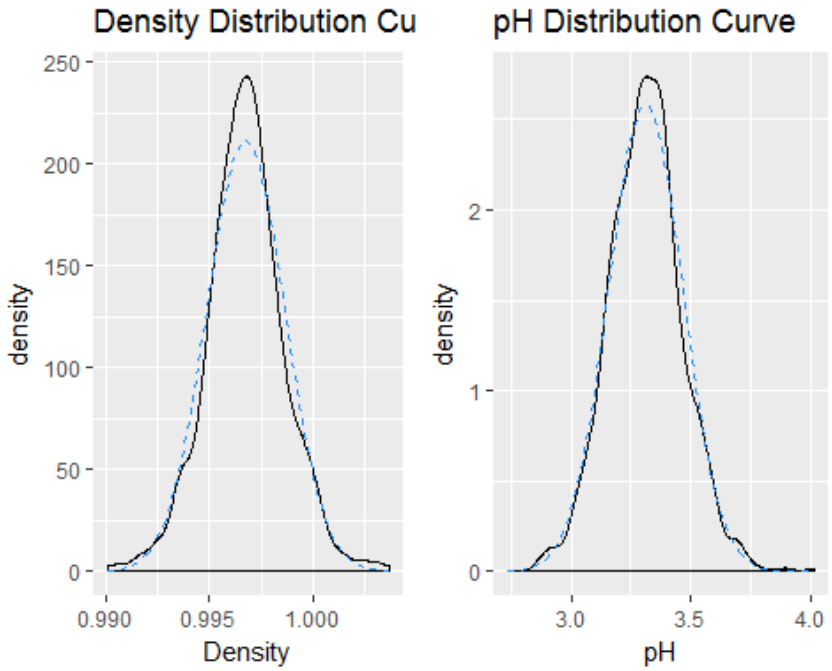
There's still a tail out there, so why not try adjusting the Chloride scale to a log:




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```

Now that the scale is logarithmic, it looks a lot better.

There are a couple of other variables that also show positive skew, such as Free and Total Sulfur Dioxide. But let's not look at those right now. Instead, I'm interested in taking a peek at Density and pH since both seem to have normal distribution. It would be nice to check it to confirm:



Those are both pretty close normal distribution For all intents and purposes we can be confident that Density and pH are normally distributed.

Univariate Analysis

What is the structure of your dataset?

The Red Wine Quality dataset contains 12 variables and 1599 observations. Of the variables, 11 are direct measurements of the physical and chemical properties of wine. One variable is Quality, which is a subjective score given to the wine as a whole.

Two variables, Density and pH show normal distributions, Eight, Residual Sugar, Alcohol, Fixed Acidity, Volatile Acidity, Chlorides, Free Sulphur Dioxide, Total Sulphur Dioxide, and Sulphates all show some degree of positive skew, while Citric Acid seems to have an almost bi-modal distribution.

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

My hypothesis is that Alcohol and Residual Sugar are the most important qualities in determining wine quality.

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

I'm also of the opinion that Acidity, in the forms of Volatile, Fixed, and Citric Acids as well as pH contribute to wine quality.

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?

Oh yeah. Chlorides and Residual Sugar both had very strong positive skew, and Citric Acid showed a very bizzare bi-modal distribution along with a possible positive skew. In all cases I adjusted the range (eliminating outliers) and scale (moving from normal to log10). This helped normalize Residual Sugar and Chlorides, though Citric Acid remains a head-scratcher.

Section 2: Bivariate Plots

We begin by comparing all of the variables against each other in a table:

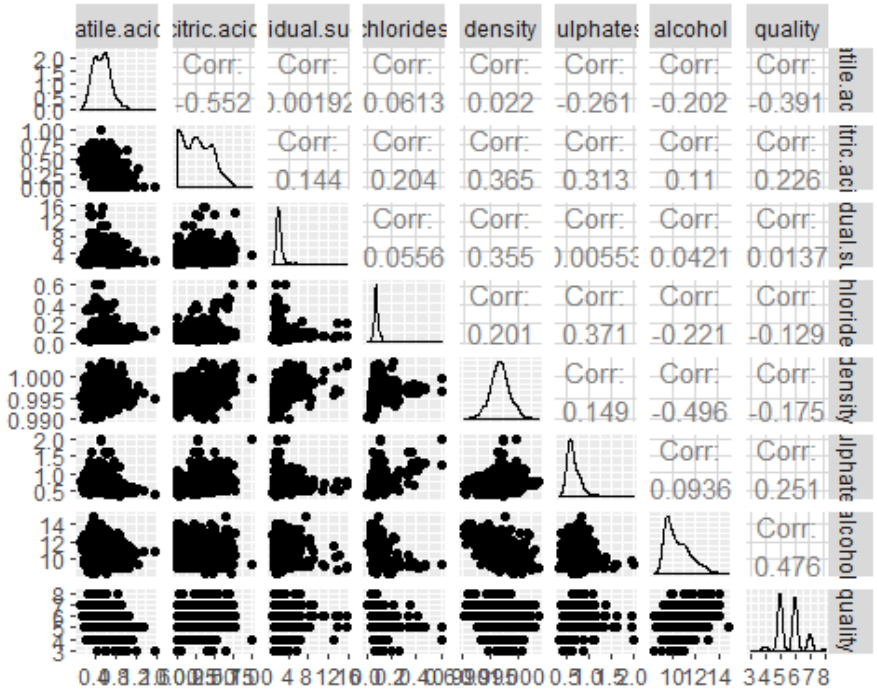
```
##                       fixed.acidity volatile.acidity citric.acid
## fixed.acidity                 1.000           -0.256       0.672
## volatile.acidity             -0.256            1.000      -0.552
## citric.acid                   0.672           -0.552       1.000
## residual.sugar                0.115            0.002       0.144
## chlorides                     0.094            0.061       0.204
## free.sulfur.dioxide          -0.154           -0.011      -0.061
## total.sulfur.dioxide         -0.113            0.076       0.036
## density                       0.668            0.022       0.365
## pH                           -0.683            0.235      -0.542
## sulphates                     0.183           -0.261       0.313
## alcohol                      -0.062           -0.202       0.110
## quality                       0.124           -0.391       0.226
##                       residual.sugar chlorides free.sulfur.dioxide
## fixed.acidity                  0.115     0.094               -0.154
## volatile.acidity               0.002     0.061               -0.011
## citric.acid                    0.144     0.204               -0.061
## residual.sugar                 1.000     0.056                0.187
## chlorides                      0.056     1.000                0.006
## free.sulfur.dioxide            0.187     0.006                1.000
## total.sulfur.dioxide           0.203     0.047                0.668
## density                        0.355     0.201               -0.022
## pH                            -0.086    -0.265                0.070
## sulphates                      0.006     0.371                0.052
## alcohol                        0.042    -0.221               -0.069
## quality                        0.014    -0.129               -0.051
##                       total.sulfur.dioxide density     pH sulphates alcohol
## fixed.acidity                       -0.113   0.668 -0.683     0.183  -0.062
## volatile.acidity                     0.076   0.022  0.235    -0.261  -0.202
```

```
## citric.acid                           0.036   0.365 -0.542     0.313   0.110
## residual.sugar                        0.203   0.355 -0.086     0.006   0.042
## chlorides                             0.047   0.201 -0.265     0.371  -0.221
## free.sulfur.dioxide                   0.668  -0.022  0.070     0.052  -0.069
## total.sulfur.dioxide                  1.000   0.071 -0.066     0.043  -0.206
## density                               0.071   1.000 -0.342     0.149  -0.496
## pH                                   -0.066  -0.342  1.000    -0.197   0.206
## sulphates                             0.043   0.149 -0.197     1.000   0.094
## alcohol                              -0.206  -0.496  0.206     0.094   1.000
## quality                              -0.185  -0.175 -0.058     0.251   0.476
##                       quality
## fixed.acidity           0.124
## volatile.acidity       -0.391
## citric.acid             0.226
## residual.sugar          0.014
## chlorides              -0.129
## free.sulfur.dioxide    -0.051
## total.sulfur.dioxide   -0.185
## density                -0.175
## pH                     -0.058
## sulphates               0.251
## alcohol                 0.476
## quality                 1.000
```

The number of variables makes this table a bit too much to take in at once, however, a review of the Quality column indicates that Alcohol (0.476) and Volatile Acidity (-0.391) have the closest correlation with wine Quality.

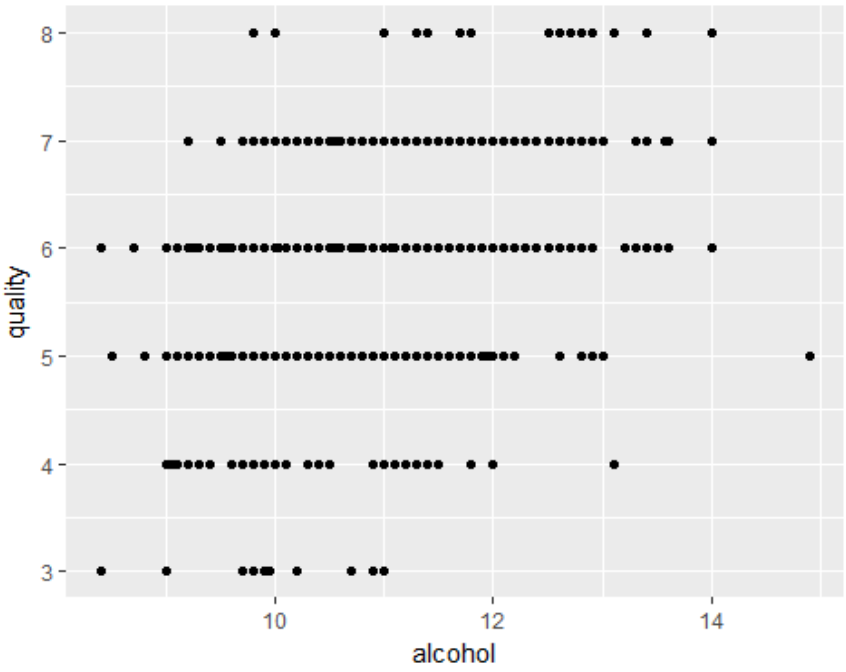
Let's give a look over at a series of scatterplots:

Okay, that's almost more intimidating than the table. It may be a good idea to trim the number of variables to make this a bit easier to digest. Given the correlations I think it might be good to go back to the table and consider focusing on those variables with the greatest correlation to Quality. With that in mind, I'll remove pH, Free Sulphur Dioxide, Fixed Acidity, and Total Sulphur Dioxide from the matrix. Even though Residual Sugar has the lowest correlation to quality in the table, since it was one of the factors I initially considered to be important to quality, I will keep it.

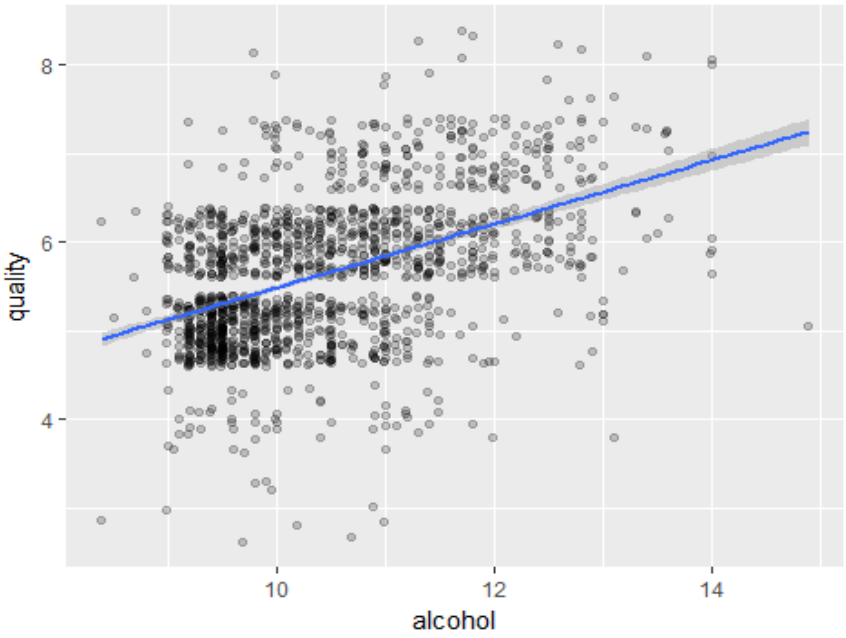


Now that's much easier to review. Density and Citric Acid both have some high correlative values with other variables, bu in terms of Quality it seems that Volatile Acid, Citric Acid, Sulphates and Alcohol are the most relevant, each with a correlation coefficient greater than +/- 0.225

Getting a closer look at these is a good idea. Let's start with Alcohol, which has the highest correlative value at 0.476

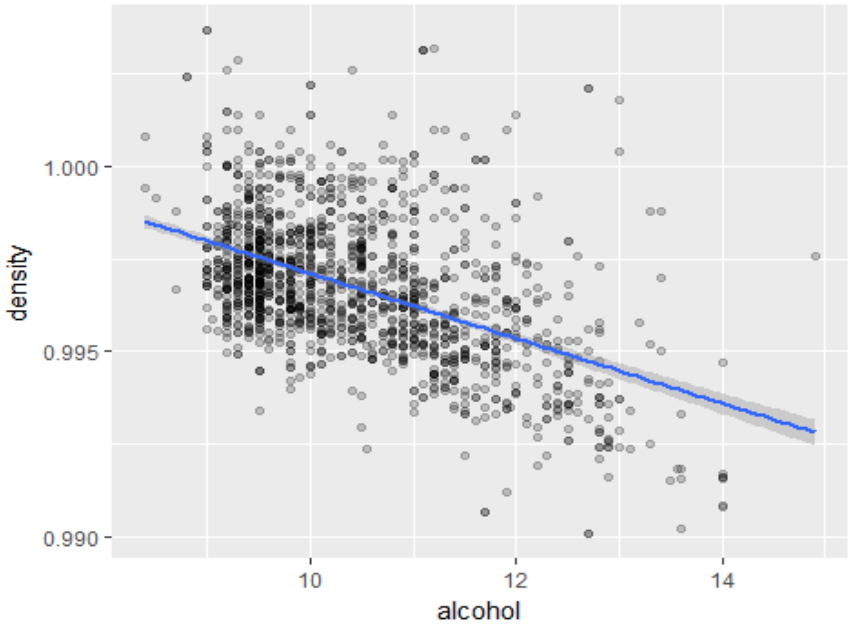


We'll clean this up a bit and add a trendline:

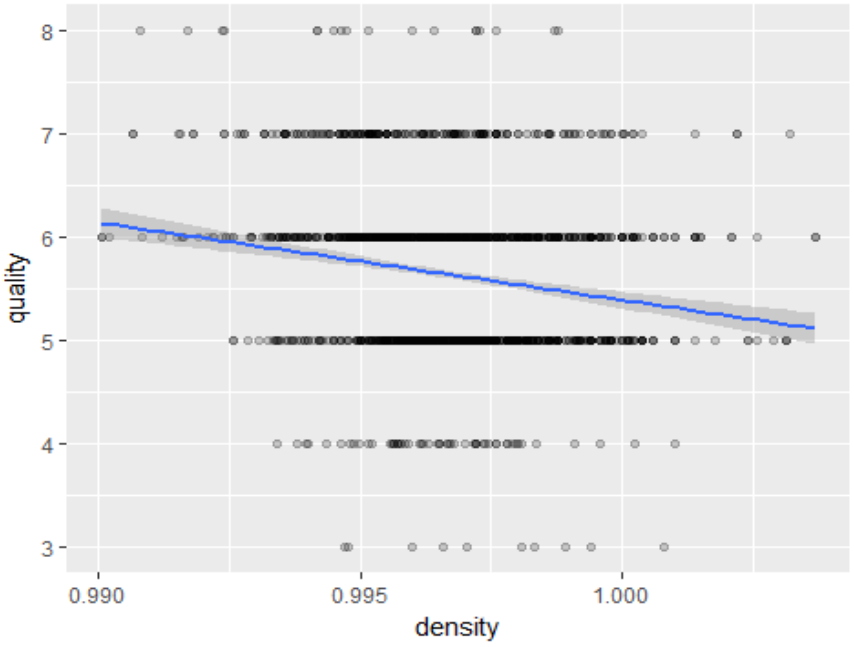


Two things jump out from this: The first is the unambiguous trendline showing an increase in quality with an increase in alcohol, and secondly that there is a very high concentration of wines in the 9-10% Alcohol concentration range.

One of the things that we all should remember from high school chemistry is that alcohol is less dense than water, therefore we would expect wine density to be inversely proportional to alcohol level. And, a look at the table confirms this as Density has a correlation of -0.496 with Alcohol. Thus, it may be a good idea to take a look at these two variables together:

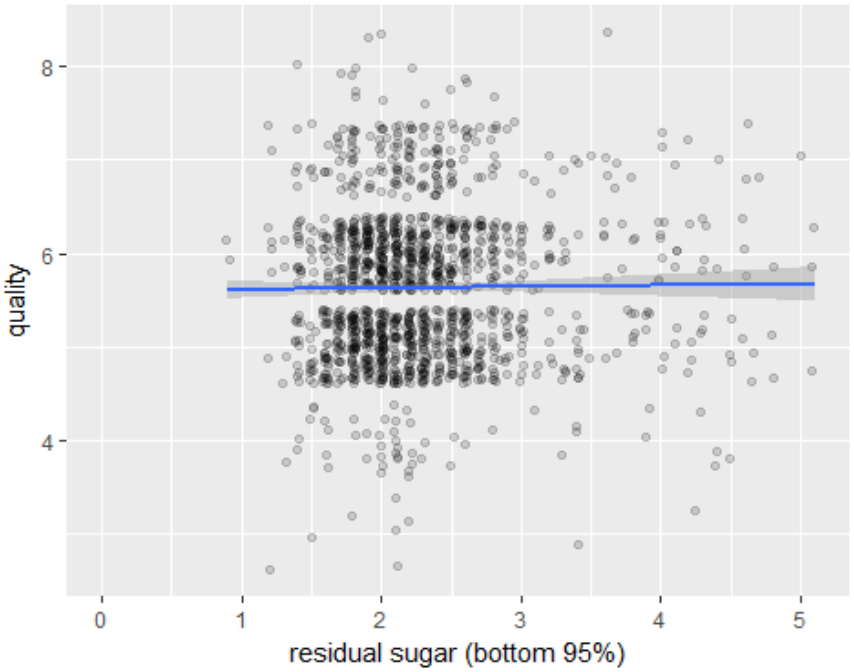


This graph is no surprise. And, given this correlation, let's look at Density and Qualityl:



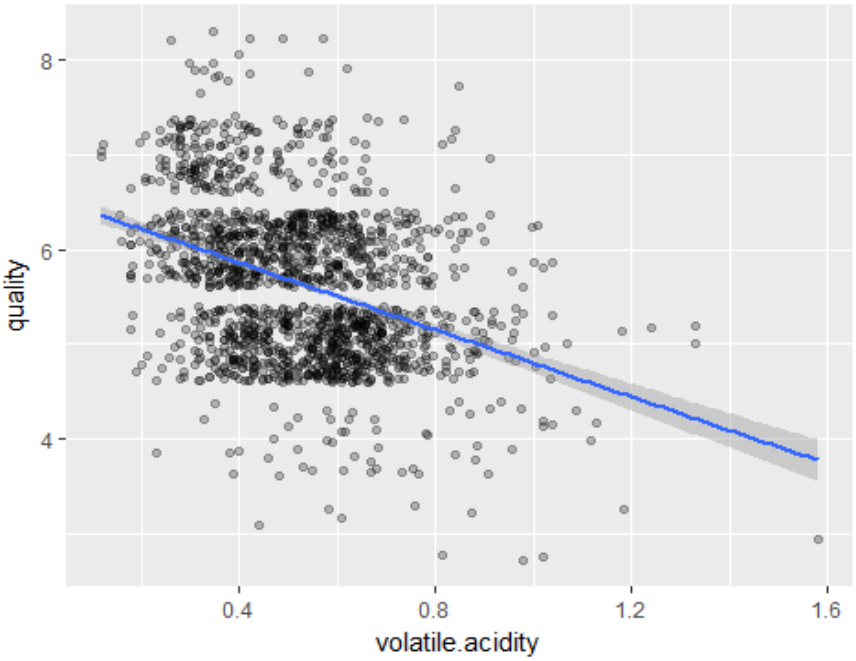
Again, the trendline does seem to be consistent given the Alcohol vs Quality graph and the relationship between Density and Alcohol, however, the data points seem to indicate that Density isn't a strong factor.

Aside from Alcohol, my initial suspicion was that Residual Sugar was an important factor in determining wine quality. Let's run this to see how it measures against Quality:



Clearly my suspicion was wrong, because that trendline is pretty flat. And, as it turns out, almost all wines have a concentration of between 1.5 and 2.5 g/L, while higher sugar levels seem to be about evenly distributed among the different quality levels.

My secondary suspicion was that acidity would have some impact on quality. As both the table and the graph matrix showed, Volatile Acidity had the second highest correlative score (-0.391) with Quality. Let's give that a look:



Well, at first look it seems that there is definitely something here, though it appears to be more a matter of high Volatile Acidity having a negative impact than anything else, as most of the wines have between 0.3 and 0.6, and seem to be distributed more or less evenly among wines of quality 5 & 6.

Now it's all fine and good to look at these graphs and see trendlines indicating correlation, but exactly how big an impact do Alcohol and Volatile Acidity have on the quality of wine? To check that we can do a quick test to find the R-squared value and measure impact. We'll start with Alcohol:

```
##
## Call:
## lm(formula = as.numeric(quality) ~ alcohol, data = rw)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8442 -0.4112 -0.1690  0.5166  2.5888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.87497    0.17471   10.73   <2e-16 ***
## alcohol      0.36084    0.01668   21.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7104 on 1597 degrees of freedom
## Multiple R-squared:  0.2267, Adjusted R-squared:  0.2263
## F-statistic: 468.3 on 1 and 1597 DF,  p-value: < 2.2e-16
```

We see R-squared indicates a 22.67% impact, which, given the presence of 11 distinct factors, seems to be quite a bit. My original hypothesis was that Residual Sugar also was important, though the correlation table/graph seems to deny that. Just for fun, let's check:

```
##
## Call:
## lm(formula = as.numeric(quality) ~ residual.sugar, data = rw)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6609 -0.6334  0.3580  0.3690  2.3729
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.616055   0.041616 134.950   <2e-16 ***
## residual.sugar 0.007865   0.014331   0.549    0.583
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8077 on 1597 degrees of freedom
## Multiple R-squared:  0.0001886,  Adjusted R-squared:  -0.0004375
## F-statistic: 0.3012 on 1 and 1597 DF,  p-value: 0.5832
```

Sure enough, The R squared value comes in at just under two tenths of a percent at 0.019% Clearly there is nothing going on here. So, how about acidity? Let's look at Volatile Acidity which had the second highest correlation to quality:

```
##
## Call:
## lm(formula = as.numeric(quality) ~ volatile.acidity, data = rw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.79071 -0.54411 -0.00687  0.47350  2.93148
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       6.56575    0.05791  113.39   <2e-16 ***
## volatile.acidity -1.76144    0.10389  -16.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7437 on 1597 degrees of freedom
## Multiple R-squared:  0.1525, Adjusted R-squared:  0.152
## F-statistic: 287.4 on 1 and 1597 DF,  p-value: < 2.2e-16
```

Here the R-squared value shows a 15.25% impact. That's not as big as I would have expected, but then again there are three additional measures of acidity: Fixed Acidity, Citric Acid, and pH. Let's give those a run:

Fixed Acidity:

```
##
## Call:
## lm(formula = as.numeric(quality) ~ fixed.acidity, data = rw)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8248 -0.6061  0.1925  0.4341  2.5550
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.15732    0.09789  52.684  < 2e-16 ***
## fixed.acidity  0.05754    0.01152   4.996  6.5e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8016 on 1597 degrees of freedom
## Multiple R-squared:  0.01539,    Adjusted R-squared:  0.01477
## F-statistic: 24.96 on 1 and 1597 DF,  p-value: 6.496e-07
```

Citric Acid:

```
##
## Call:
## lm(formula = as.numeric(quality) ~ citric.acid, data = rw)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0011 -0.5976  0.1021  0.5057  2.5901
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.38172    0.03372 159.610   <2e-16 ***
## citric.acid  0.93845    0.10104   9.288   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7869 on 1597 degrees of freedom
## Multiple R-squared:  0.05124,    Adjusted R-squared:  0.05065
## F-statistic: 86.26 on 1 and 1597 DF,  p-value: < 2.2e-16
```

pH:

```
##
## Call:
## lm(formula = as.numeric(quality) ~ pH, data = rw)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6817 -0.6394  0.3032  0.3878  2.4874
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.6359     0.4332  15.320   <2e-16 ***
## pH           -0.3020     0.1307  -2.311    0.021 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8065 on 1597 degrees of freedom
## Multiple R-squared:  0.003333,   Adjusted R-squared:  0.002709
## F-statistic:  5.34 on 1 and 1597 DF,  p-value: 0.02096
```

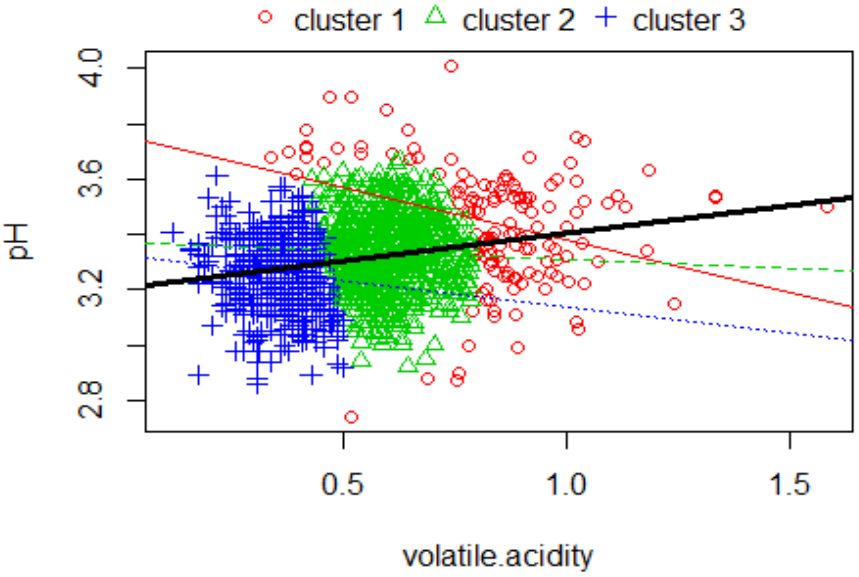
Clearly, neither pH nor Citric Acid have much impact, at 0.33% and 5%, respectively. But Fixed Acidity also contributes 15.39% to the quality, so this along with Volatile Acidity and Alcohol make up over 50% of the impact on overall wine quality.

Something in the Acid?

One of the things I noticed when reviewing both the graph matrix and the correlation table was that pH had inconsistent correlation with the other three acid components. Since pH is a general measure of acidity with an inverse relation between pH measurement and degree of acidity, one would expect a

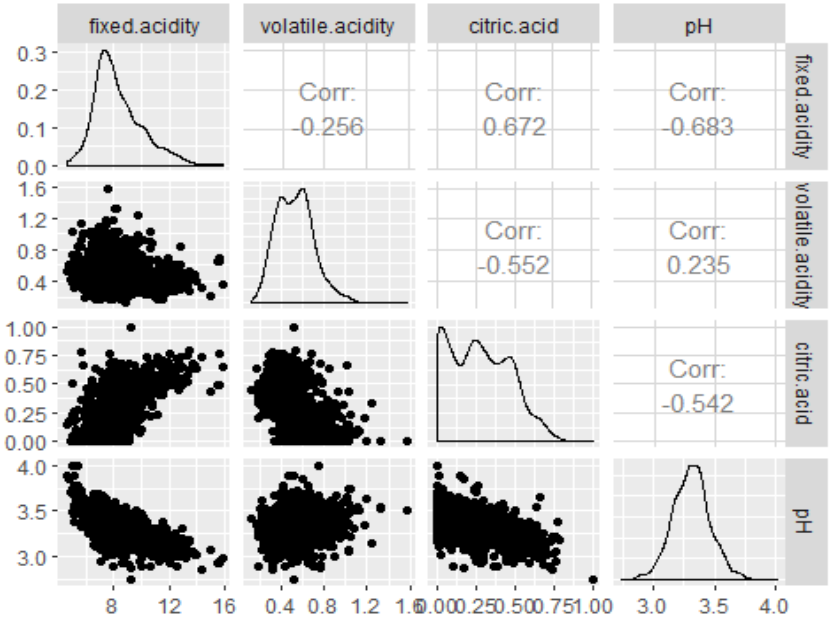
negative correlation between pH and Volatile, Fixed, and Citric Acid. However, pH is actually somewhat positively correlated with Volatile Acidity, with a correlation factor of 0.235. How can this be?

My suspicion is the presence of Simpson's Paradox, which I will check now:



Well, look at that! Volatile Acidity is comprised of three separate clusters, each of which shows a negative correlation with pH, but when aggregated has a positive correlation. Simpson's confirmed!

And, just for grins and giggles, let's run a graph matrix of just the acids to see how they all relate to one another:



Well, unsurprisingly there seems to be fairly strong correlations across the board.

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?

My initial hypothesis was that Alcohol and Sugar were the two most important factors in determining wine quality. It turns out I was half-right. Alcohol appears to be the single biggest factor, contributing about 22.5% to overall quality, while sugar is irrelevant, at only 0.019%.

Beyond that, my next suspicion was that acidity plays a part, and sure enough both Fixed and Volatile Acidity contribute just over 15% each. Combined with alcohol, these three factors have over 50% impact on wine quality.

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

Looking at the relationships between the factors I found that there was a Simpson's Paradox hiding in the acids. As we all know from high school chemistry, pH is a direct measurement of acidity, with the lower the pH the more acidic. Yet the relationship between pH and Volatile Acidity showed a positive correlation - which would mean either something odd was happening in Volatile Acidity, or there is something about wine that messes up the pH-Acid relationship. A look into it showed that Volatile Acidity was actually an aggregation of three separate subsets of acids, each of which was negatively correlated with pH, but which combined gave a false positive correlation.

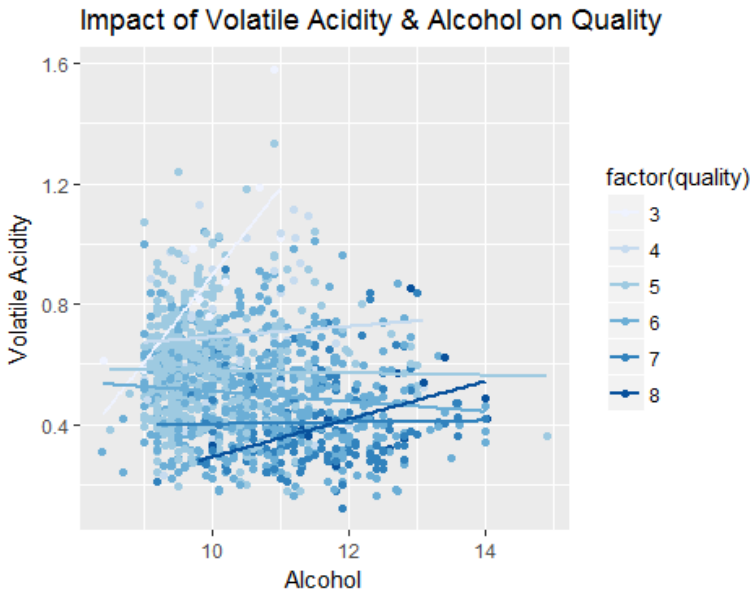
How cool is it to find something like that hiding in your wine?

What was the strongest relationship you found?

There were several strongly correlated factors. First, the acids all shared strong correlations to one another, which is expected. As did Alcohol and Density (again, as expected). However, in terms of qine quality, it was Alcohol and Volatile Acidity which had the strongest relationship.

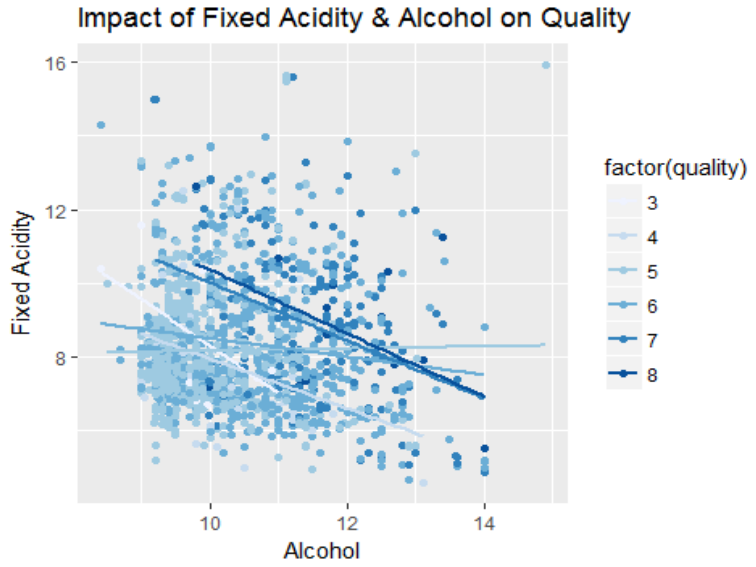
Section 3: Multivariate Plots

As mentioned several times eariler, the two factors most strongly correlated with Quality were Alcohol and Volatile Acidity. Let's go ahead and plot this:



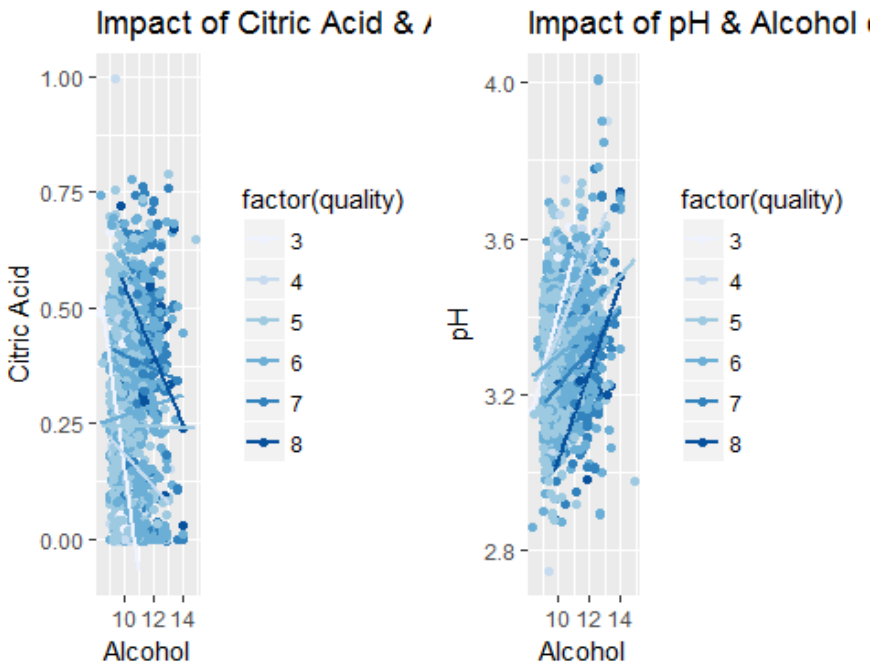
From this it appears that there is some sort of relationship between Alcohol and Volatile Acidity present in high and low quality wines (scores of 8 and 3) in which the increase in both improves wine, but which is neglibile in medium quality wines (scores of 4-6).

The R-squared analysis indicated that that Fixed Acidity contributed nearly as much to quality as Volatile Acidity, so let's run that alongside Alcohol:



This seems to show a different relationship, where decreased Fixed Acidity in the presence of higher Alcohol had a positive impact on wine, and again seems restriced to both high (score 7 & 8) and low (score 3) wines.

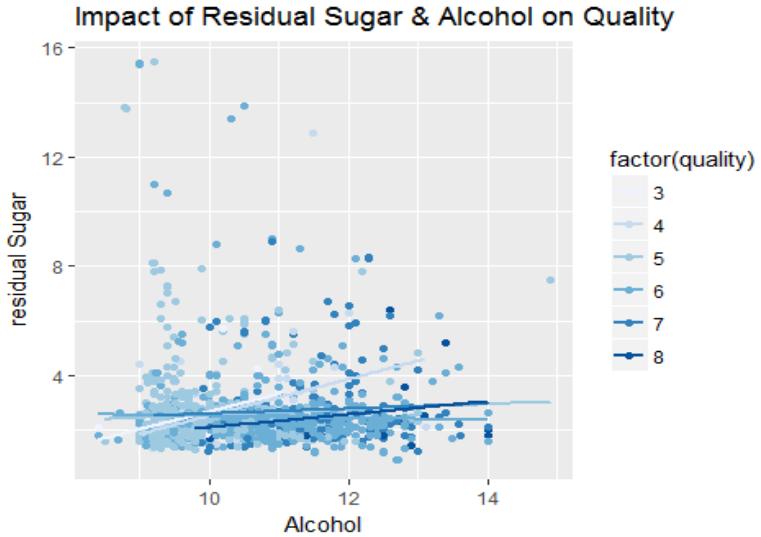
We may as well look at the remaining acidity measurements just to be thorough.



Okay, these are a bit squeezed, but there are patterns visible in both. It appears that Citric Acid is similar to Fixed Acidity in that its impact on quality is inverseley proportional to the presence of Alcohol, particluarly at high (score 7 & 8) and low (score 3 & 4) wines. It also seems that it's impact may be greater than Fixed Acidity.

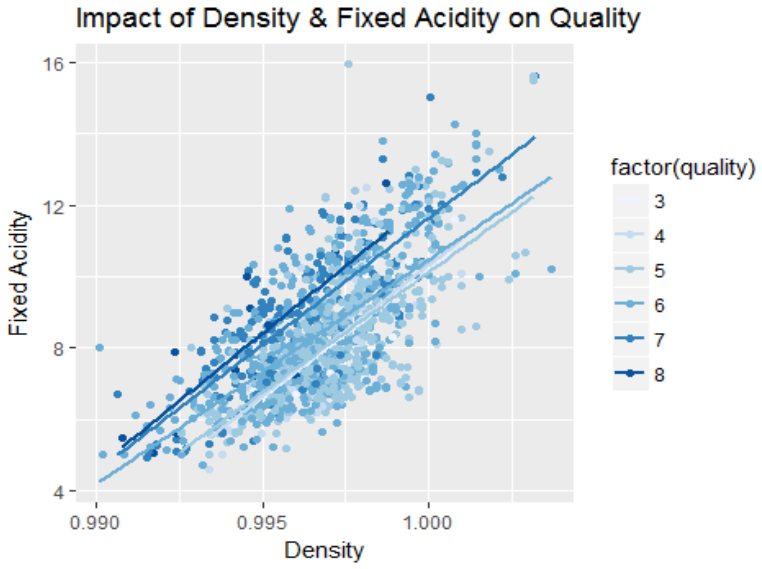
pH, on the other hand, has a similar directly proportional impact, much like Volatile Acidity, although it appears to be present in all levels of wine quality. This suggests that, in general, acidity is a significant factor, though it is difficult to tell which of the acidic properties are at play.

Now, even though we've already beaten the idea that Residual Sugar has any impact, since it was part of my initial hypothesis, I feel duty-bound to include an analysis using it and Alcohol:



Well, this seems to be nothing more than a stack of flat lines, other than at the low end of wine quality. This isn't much of a surprise.

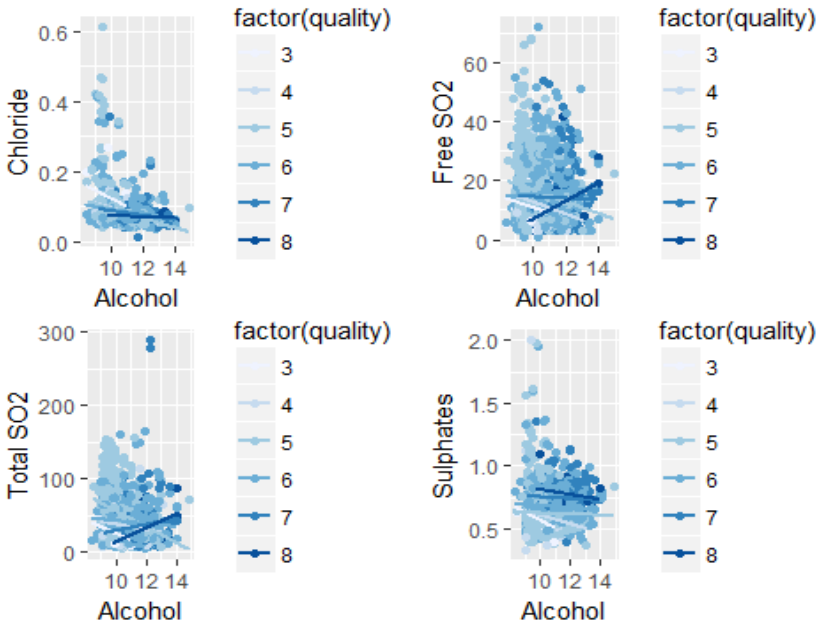
Curiosity has me reviewing the correlation table, where I notice that the highest correlative factor between un-related elements (i.e. not acids to acids, or free and total sulfur dioxide) is between Density and Fixed Acidity with a value of 0.668. Naturally, I have to check those out and see whether their interplay has anything to do with quality at all:



Well, there certainly isn't much ambituity here, is there? Across all wine quality scores there is a very clear and very strong positive relationship between Fixed Acidity and Density. Given that Density is greatly impaced by the level of Alcohol, this certainly not surprising, as we saw earlier how Fixed Acidity had an inverse relationship to Alcohol.

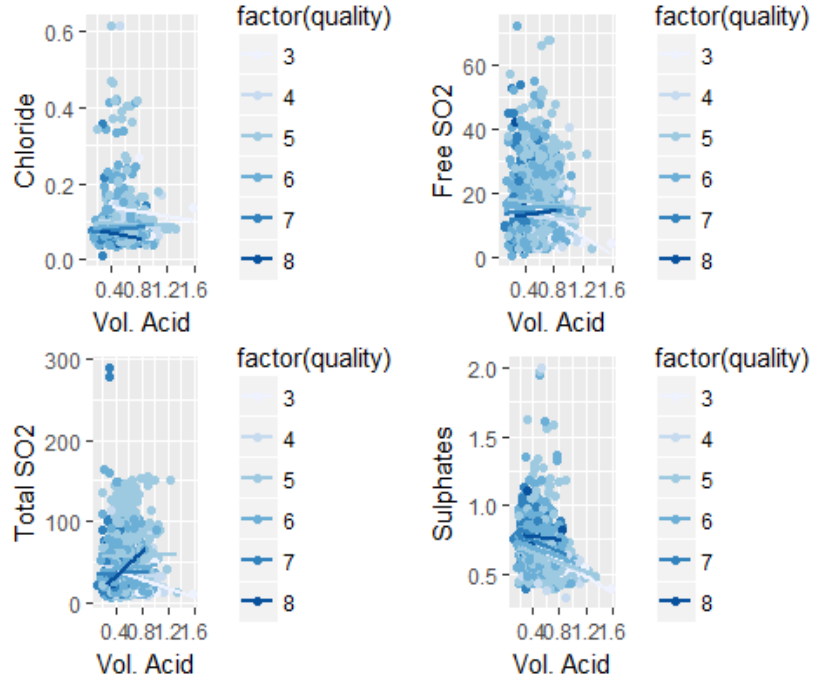
Interesting.

Lastly, it may be valuable to consider how some of the thus far neglected factors, such as sulphates, chlorides, and free & total sulfur dioxide may be involved. I'll begin by running each of these along with alcohol:



It's hard to see for sure, but it appears all of these variable have some impact, though the extent is difficult to easily see..

And last, just out of sheer curiosity I'll run these with Volatile Acidity:



Once again, it appears that there may be some impact, but it is minimal and difficult to identify.

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?

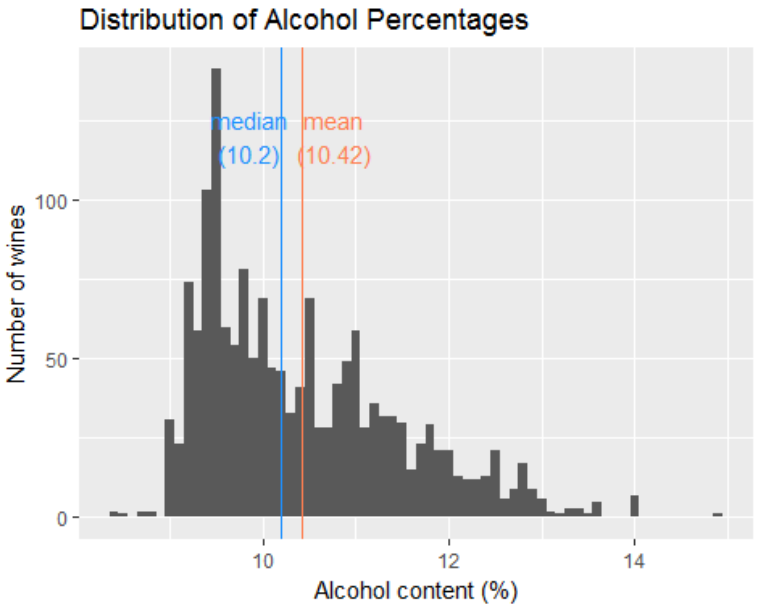
These plots helped underscore the importance of acidity in the quality of wine.

Were there any interesting or surprising interactions between features?

The strong relationship that Density has with Fixed Acidity on quality is very interesting. Further, the direct effect that alcohol concentration has on density in general helps support my initial hypothesis regarding Alcohol's importance, but it also does raise the question of what else may impact wine's density.

Section 4: Final Plots and Summary

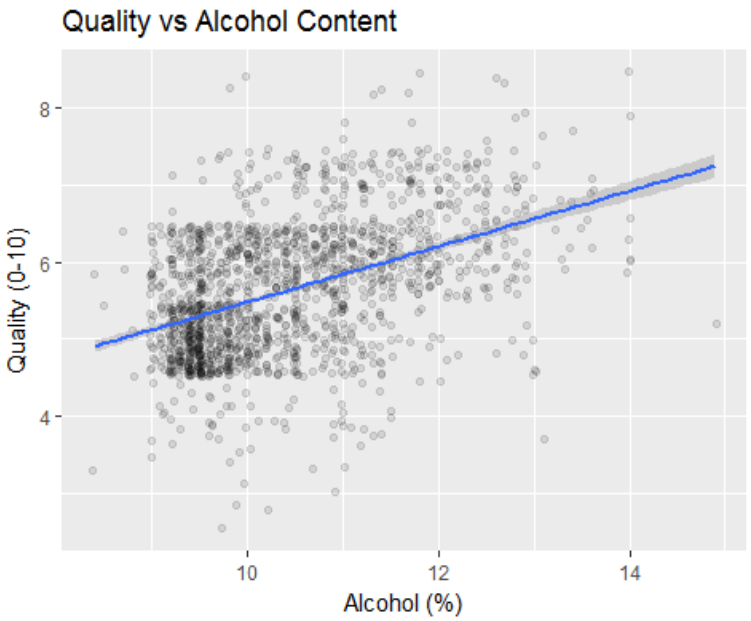
Plot One



Description One

Alcohol content in wine shows a positive skew (i.e. the Mean lies at a higher value than the Median, and there is a long tail trailing toward higher values). The majority of wines have alcohol content between 9 and 11%, with very few greater than 12%

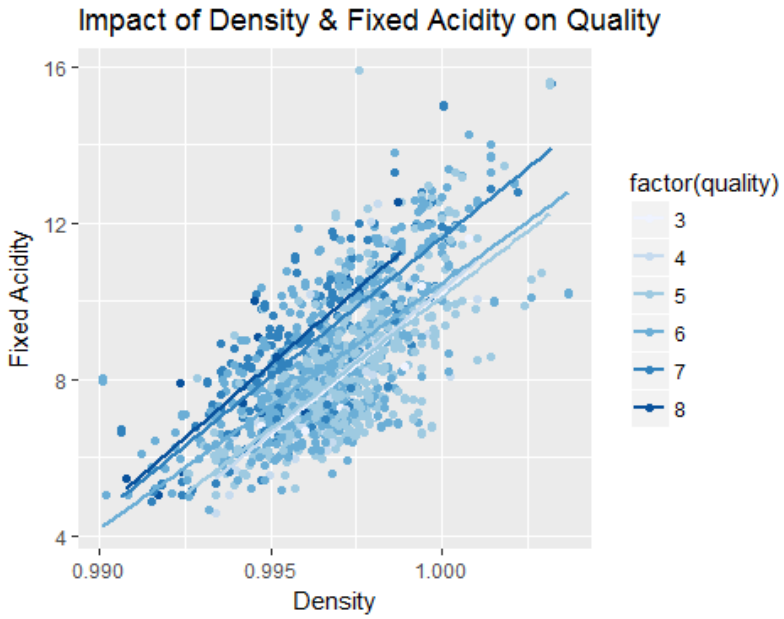
Plot Two



Description Two

This plot shows a clear positive correlation between alcohol content and wine quality, particularly as alcohol rises to between 10 and 12%. Wines below 10% alcohol tend to range in the medium to low quality range, though this may be more due to the fact that there are much fewer wines with greater than 10% alcohol than there are below 10%.

Plot Three



Description Three

This provides a clear view of the close, directly proportional relationship between Density and Fixed Acidity on quality. However, since both of these variables are influenced by several factors, it serves to emphasise the deeply complex chemical nature of wine, and the delicate relationship between all of the basic chemical components present.

Section 5: Reflection

I approached this analysis as someone who enjoys red wine on occasion, but who has no knowledge or experience in determining quality beyond the most basic subjective test of whether I liked it or not. However, I also knew that there existed properties within the wine that affected the overall quality, and that how those worked to make a wine good or bad was objective.

With that in mind, I first took note of the eleven properties measured: four measurements of differng acidic qualities, two measurements of sulfur dioxide and related sulphates, density and alcohol, sugars, and chlorides. Based on my limited knowledge of wines, my initial suspicion was that alcohol and sugars were the most important factors in quality, with acids playing a secondary role.

The first step in checking my hypothesis was a review of the descriptive statistics and distribution of each element. This showed that, for the most part, all had some degree of positive skew, other than Density and pH, which had normal distribution, and Citric Acid, which ad a very strange bi-modal distribution with a positive skew.

The next step was to do bi-variate analysis to look for initial correlation between the factors, which showed that while alcohol did have a correlation to quality, sugar did not, and that acids may play a bigger role than intended. However, it also exposed a puzzling result where pH seemed to have a positive correlation to Volatile Acidity, rather than the negative correlation expected. A bit of digging confirmed the presence of Simpson's Paradox to explain this strange relationship.

The multivariate analysis confirmed that alcohol and acidity are important factors in determining quality, though the interplay between them is complicated. Further, while the other elements (such as sulphates or chlorides) do not appear to have an obvious role in quality, there may be relationships and interactions which are present - such as how these may influence the density of the wine, but which I could not determine.

If I were to continue in this analysis, I think I would attempt to focus more on how the variables influence the density of the wine, as well as teasing out how variables like the sulfur dioxides or chloride impact overall acidity. These certainly play a role in wine quality, and I would be curious to understand how. Further, this dataset covered generic "Red" wine. And as we know, Red wine contains many different types, such as Cabernet, Pinot Noir, Sangiovese, Merlot, Malbec, etc. These different wines have different characteristics, and the components of what makes a good Pinot Noir may be very different than what makes a good Sangiovese or a good Malbec. Getting a more precise dataset, in which wine varietals are identified, would be quite interesting.