# Investigating White Wine Quality

Code

Justin Smith

05 May, 2020

# 1 Introduction

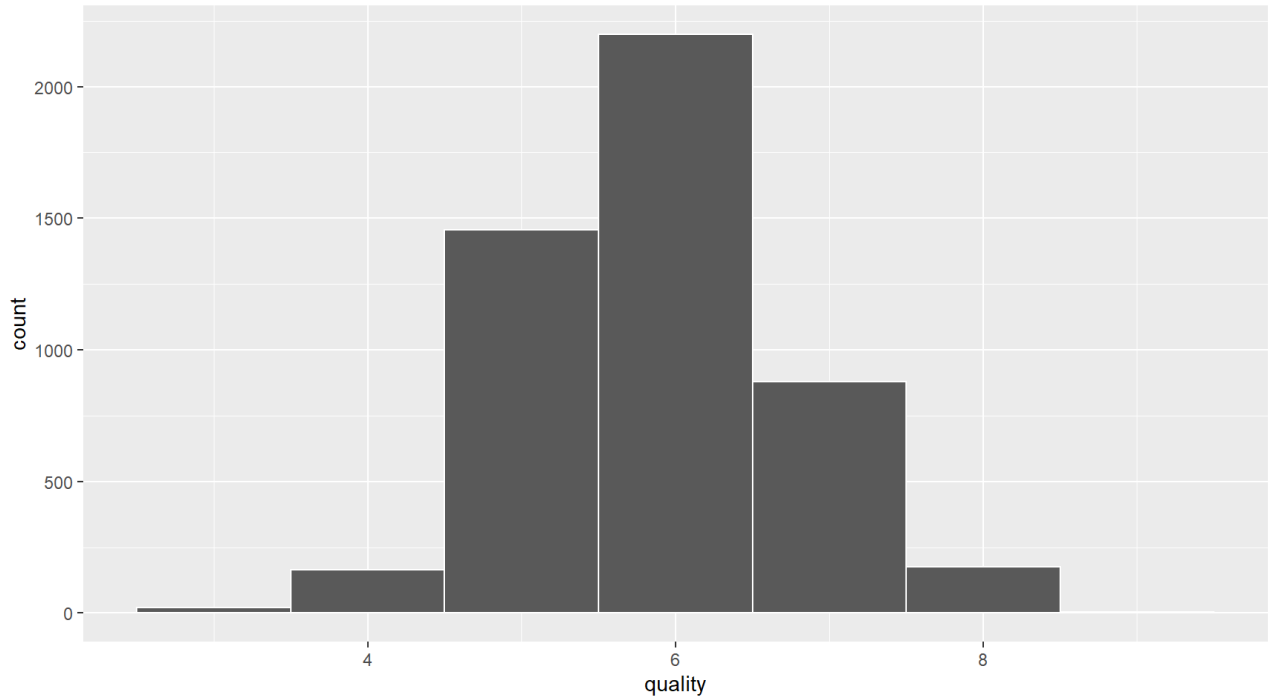
I will investigate how the chemical composition of white wine affects its quality rating. More information about the dataset can be found here (https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityInfo.txt)

# 2 Summary and Structure

```
## 'data.frame':    4898 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22
...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43
...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045
0.049 0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45
...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
```

```
##        X        fixed.acidity    volatile.acidity  citric.acid
##  Min.   :   1   Min.   : 3.800   Min.   :0.0800   Min.   :0.0000
##  1st Qu.:1225   1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700
##  Median :2450   Median : 6.800   Median :0.2600   Median :0.3200
##  Mean   :2450   Mean   : 6.855   Mean   :0.2782   Mean   :0.3342
##  3rd Qu.:3674   3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900
##  Max.   :4898   Max.   :14.200   Max.   :1.1000   Max.   :1.6600
##  residual.sugar     chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   : 0.600   Min.   :0.00900   Min.   :  2.00      Min.   :  9.0
##  1st Qu.: 1.700   1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0
##  Median : 5.200   Median :0.04300   Median : 34.00      Median :134.0
##  Mean   : 6.391   Mean   :0.04577   Mean   : 35.31      Mean   :138.4
##  3rd Qu.: 9.900   3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0
##  Max.   :65.800   Max.   :0.34600   Max.   :289.00      Max.   :440.0
##     density             pH          sulphates         alcohol
##  Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00
##  1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50
##  Median :0.9937   Median :3.180   Median :0.4700   Median :10.40
##  Mean   :0.9940   Mean   :3.188   Mean   :0.4898   Mean   :10.51
##  3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40
##  Max.   :1.0390   Max.   :3.820   Max.   :1.0800   Max.   :14.20
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.878
##  3rd Qu.:6.000
##  Max.   :9.000
```

The first thing I notice when looking at the structure of the dataset is that all variables are continuous. It would be ideal to have a categorical variable so that different categories of wine can be grouped so that common characteristics for each grouping can be further dissected. Next I am going to take a look at a histogram to see the distribution of the quality scores variable.

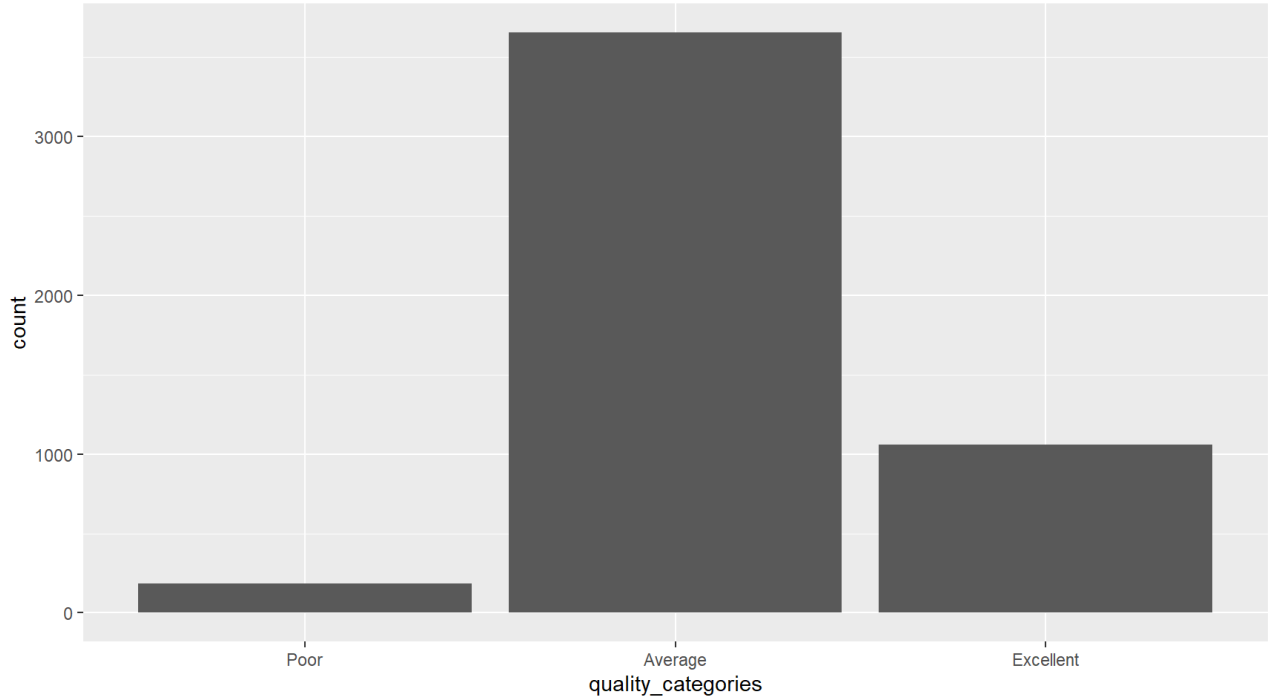



| quality <int> | n <int> |
|---|---|
| 3 | 20 |
| 4 | 163 |
| 5 | 1457 |
| 6 | 2198 |
| 7 | 880 |
| 8 | 175 |
| 9 | 5 |

7 rows

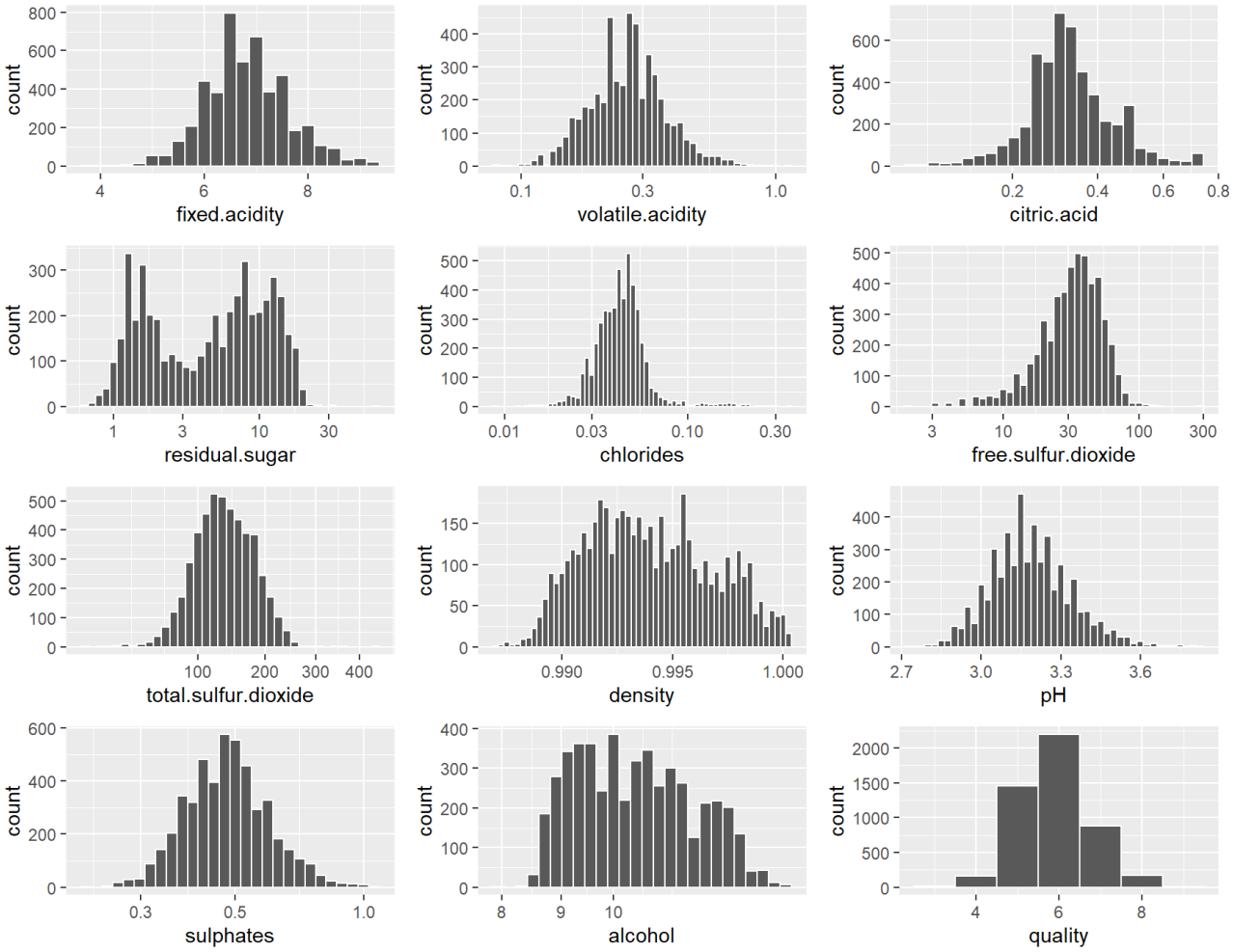
We can see that the mode of the dataset has a quality score of 6. Most of our white wines are in the 5 and 6 quality score range. While this is interesting to know, I am more interested in what causes a wine to score on the higher end range of the scale, the 7, 8, and 9 quality scores. I am also interested in the characteristics that lead to a poor wine quality rating. I will divide the wines into 3 separate categories: the bottom 25% being 'Poor' quality ( below a 5 on the quality scale), the middle 50% being 'Average' (the 5 and 6 ratings), and 'Excellent' being the top 25% (7 and above). This will give me three distinct categories of wines to further analyze the chemical compositions of each.

Now to view the frequencies for each of these categories.




Next, I wanted to view a normalized version of each variable to see any noticeable trends or outliers. Some of the variables I used logarithmic and square root transformations on, and others I excluded outliers to help normalize all the variables. These are displayed below.

# 3 Univariate Plots Section




# 4 Univariate Analysis

## 4.0.1 What is the structure of your dataset?

There are 4898 observations in the dataset and 13 variables.

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

The main feature I am interested in investigating is what chemical properties of white wine result in Excellent quality score ratings? On the opposite end of the spectrum, what chemical properties are present in Poor quality wines? Investigating these two different groupings of wine qualities will be the focus of my analysis and hopefully lead to some insight as to what chemical properties help determine Poor and Excellent quality wines.

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

I would suspect that residual sugar found in the wine will play a role in a wines' quality rating since sweetness is directly related to how something tastes. Following that same line of thought, I would suspect alcohol content plays a factor as well; generally speaking, the higher alcohol content something has, the more the alcohol taste is noticeable. pH might be another factor since it correlates to the acidity level found in wines and I would think higher acidity levels would have a distinct taste. The density of the wine might also play a factor, differences in viscosity can affect a persons wine preferences.

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

I created the variable quality_categories to group the quality variable into different tiers. The wines with a quality score under 5 are considered 'Poor', those that are 5 and 6 are considered 'Average', and those 7 and above are considered 'Excellent'. Establishing this categorical variable will help identify characteristics within a group and between groups.

### 4.0.5 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?

There is a variable 'x' that is essentially a primary id column for each observation in the dataset. This is irrelevant for my analysis so I will not include this variable.

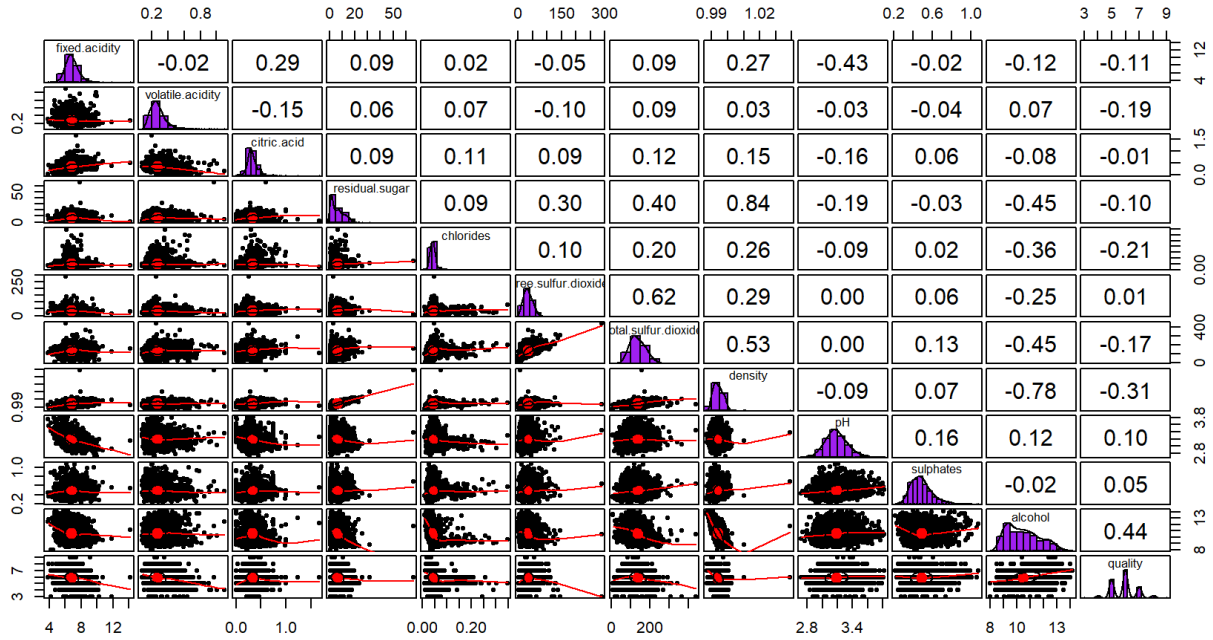
The citric acid variable has 19 values that are 0. The data source indicated this chemical is added for freshness, so it seems 19 wines did not have this additive. To normalize this variable, I excluded the bottom and top 1% and transformed it using the square root function.

The residual sugar variable seems to have a bimodal distribution with two peaks of sugar content.

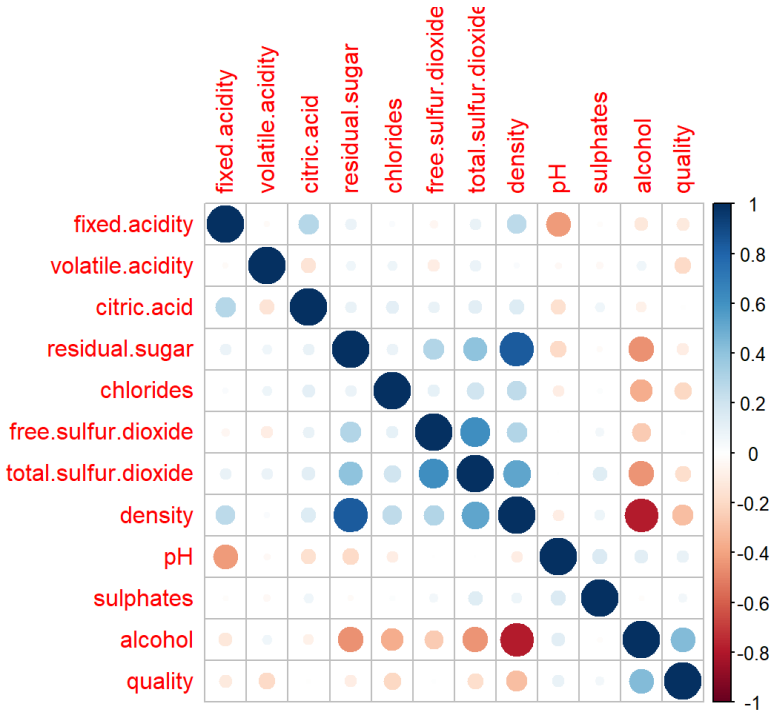
To normalize the variables, I excluded the top 1% of outliers for fixed acidity and density. I also used a logarithmic transformation to normalize the dataset for the following variables that had long tail distributions: volatile acidity, residual sugar, chlorides, free sulfur dioxide, sulphates, and alcohol. Additionally, I used the square root transformation for the total sulfur dioxide variable to normalize its distribution.

# 5 Bivariate Plots Section

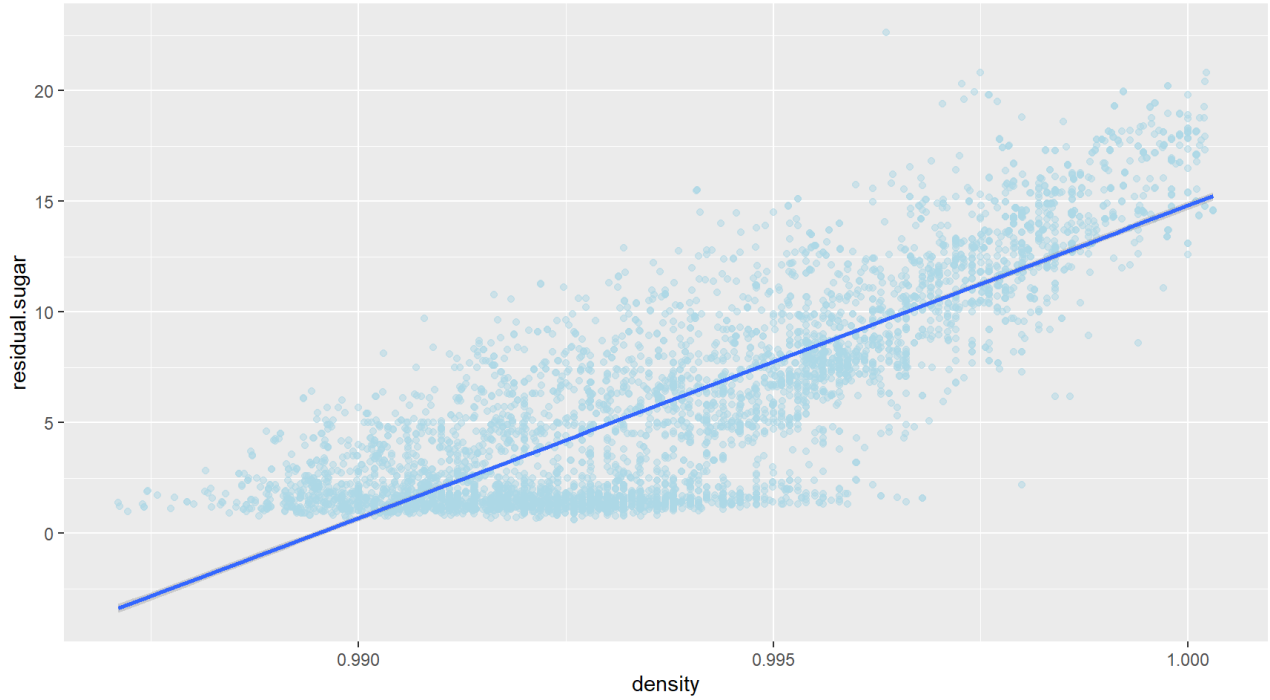
First I wanted to see a scatterplot matrix of all the variables.



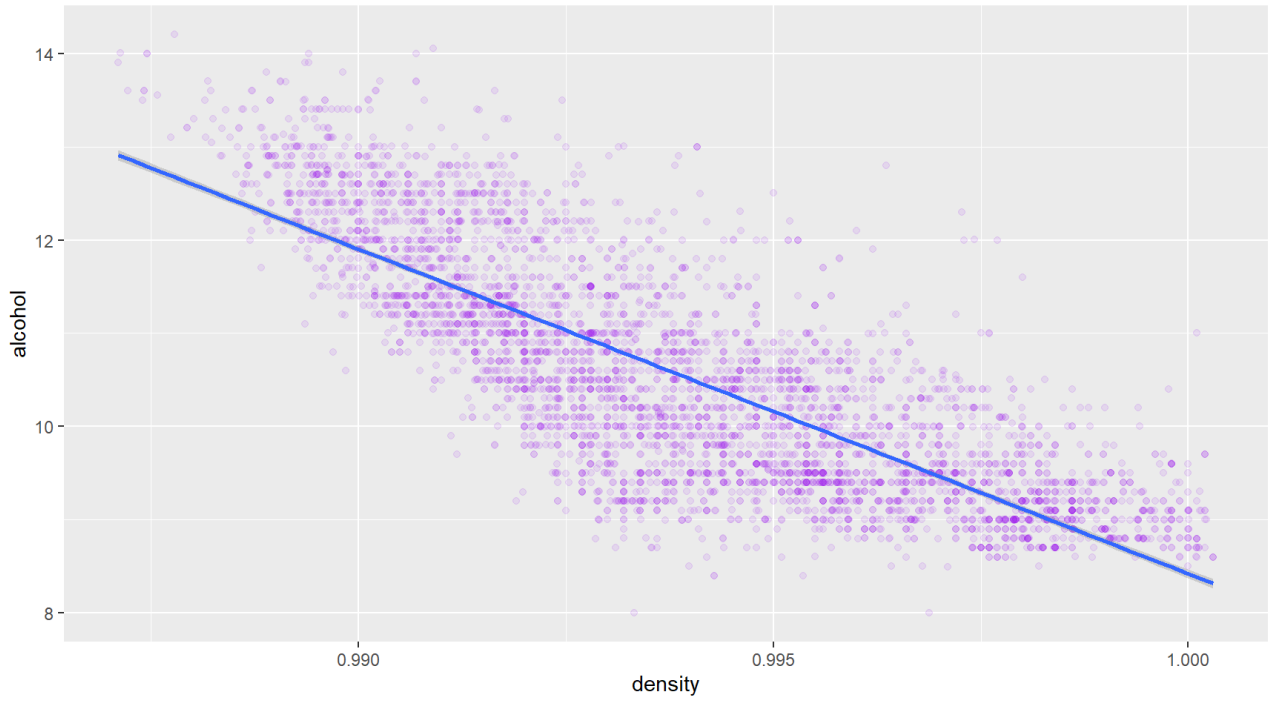

I then decided to also do a correlation matrix to better visualize the correlations between variables.




This visual more easily displays the relationships between variables. The strongest correlations are a strong positive correlation between density and residual sugar as well as a strong negative relationship between density and alcohol content. I will take a closer look at these next.

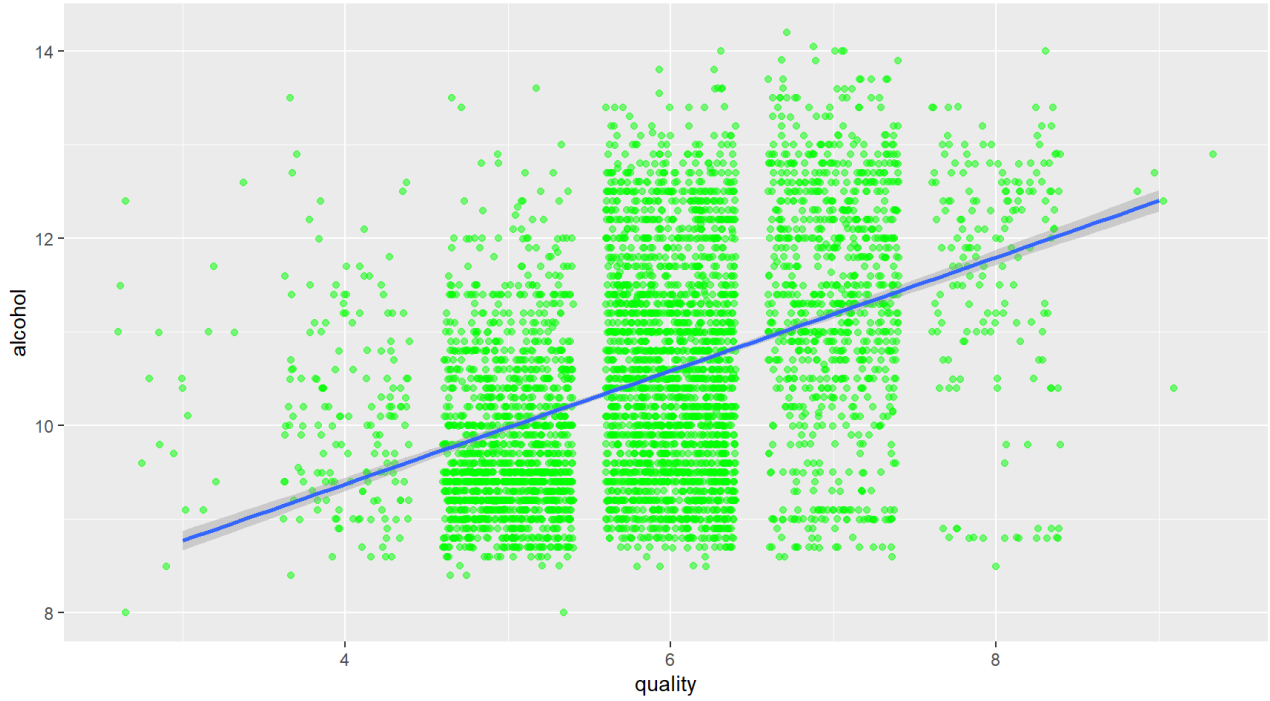



```
##
##  Pearson's product-moment correlation
##
## data:  ww$density and ww$residual.sugar
## t = 107.87, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8304732 0.8470698
## sample estimates:
##       cor
## 0.8389665
```

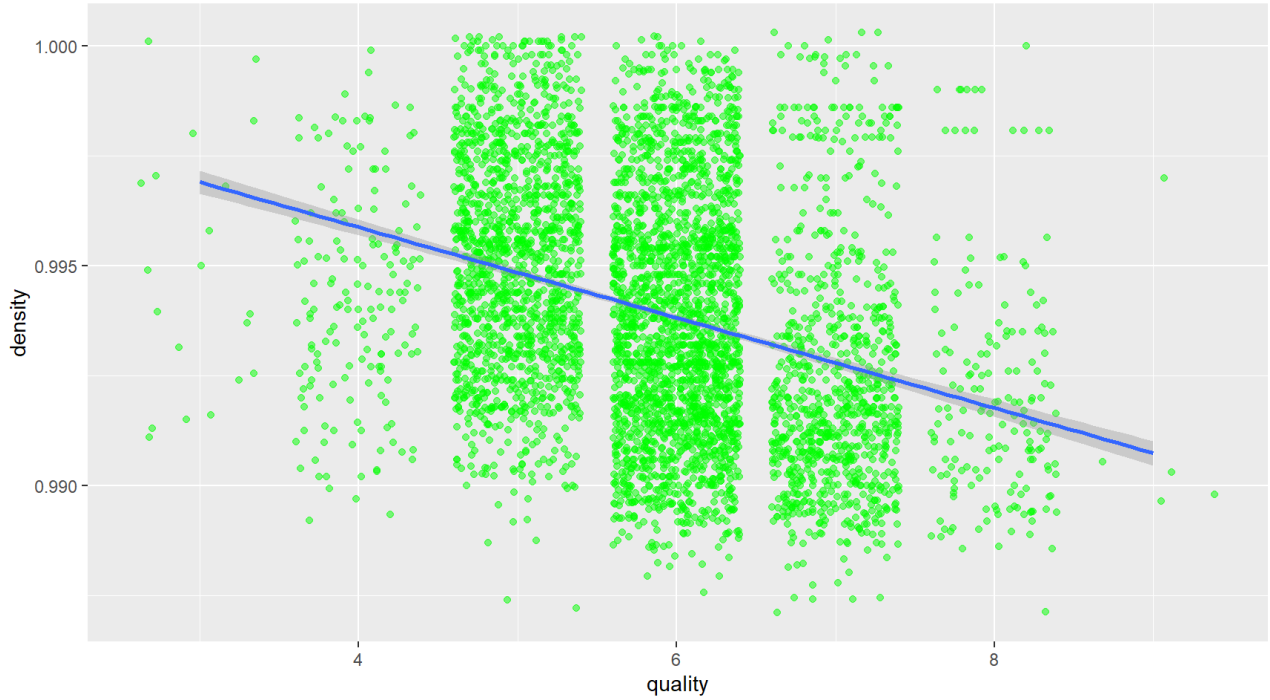


```
##
##  Pearson's product-moment correlation
##
## data:  ww$density and ww$alcohol
## t = -87.255, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7908646 -0.7689315
## sample estimates:
##       cor
## -0.7801376
```

This makes sense because more sugar in a wine would make it denser and we should expect a positive correlation. Also, the more alcohol in a wine would result in a lower density since alcohol is less dense than water. I am also interested in the correlations that exist between quality rating and other variables. I will explore this further.

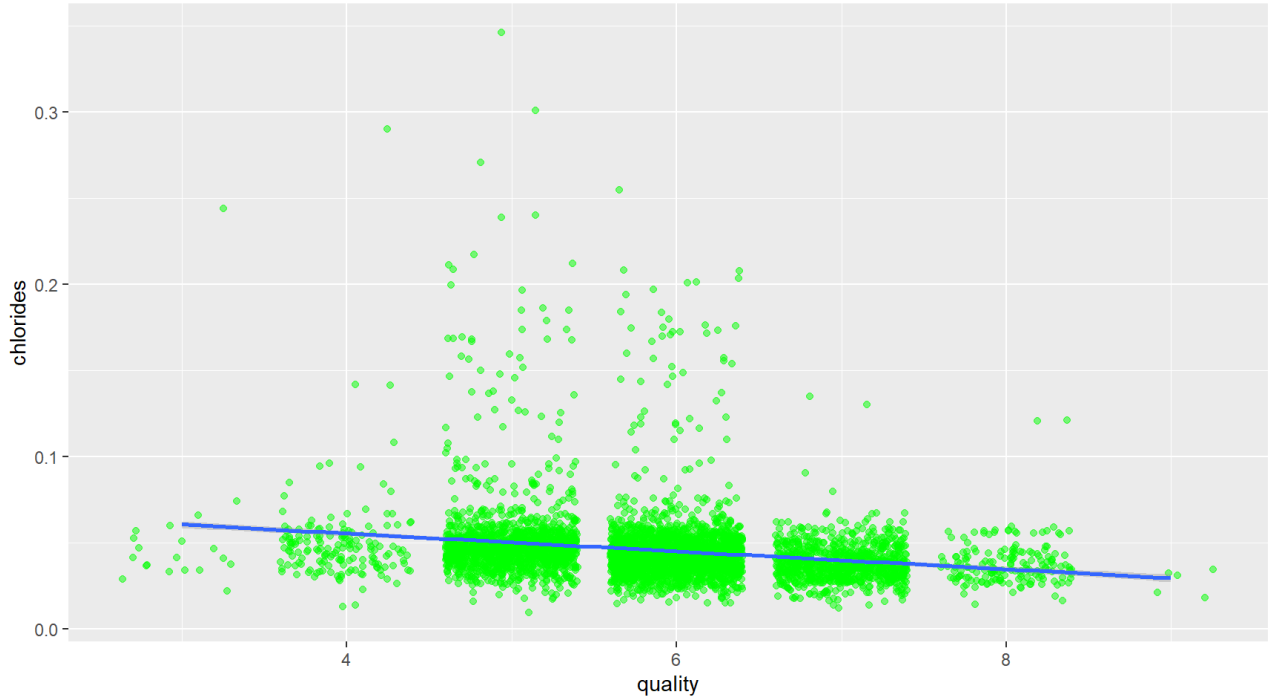



```
##
##  Pearson's product-moment correlation
##
## data:  ww$quality and ww$alcohol
## t = 33.858, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4126015 0.4579941
## sample estimates:
##       cor
## 0.4355747
```



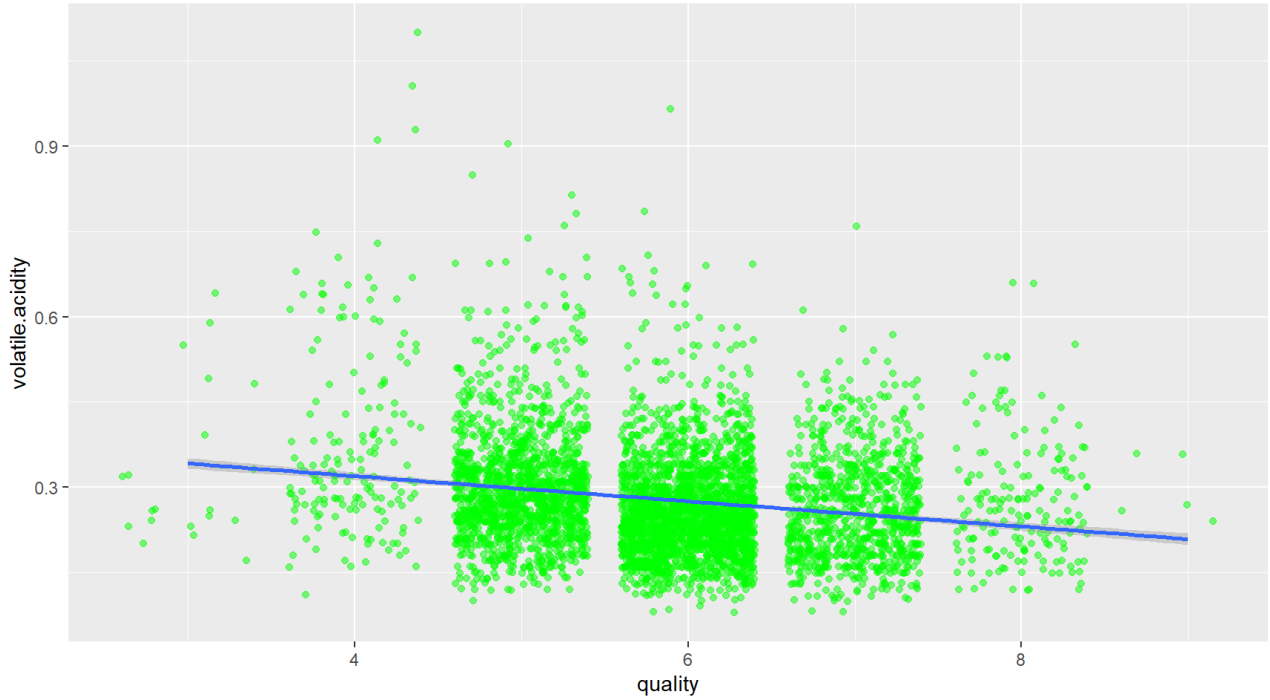

```
##
##  Pearson's product-moment correlation
##
## data:  ww$quality and ww$density
## t = -22.581, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3322718 -0.2815385
## sample estimates:
##        cor
## -0.3071233
```

We can see that we have a medium correlation between alcohol content and the quality of wine at 0.44. We also have a medium correlation of -0.31 between wine quality and density. This is not to say having a higher alcohol content leads to a better wine, or that having lower density means a better wine, but it is worth further exploring. We previously noted a strong negative correlation between density and alcohol content, so these are related. I suspect that there could be additional variables in play that happen to correspond with high quality wines having higher alcohol content and lower densities, not that these variables are directly responsible for wine quality themselves. Let us take a closer look at some other chemical characteristics.




```
##
##  Pearson's product-moment correlation
##
## data:  ww$quality and ww$chlorides
## t = -15.024, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2365501 -0.1830039
## sample estimates:
##        cor
## -0.2099344
```

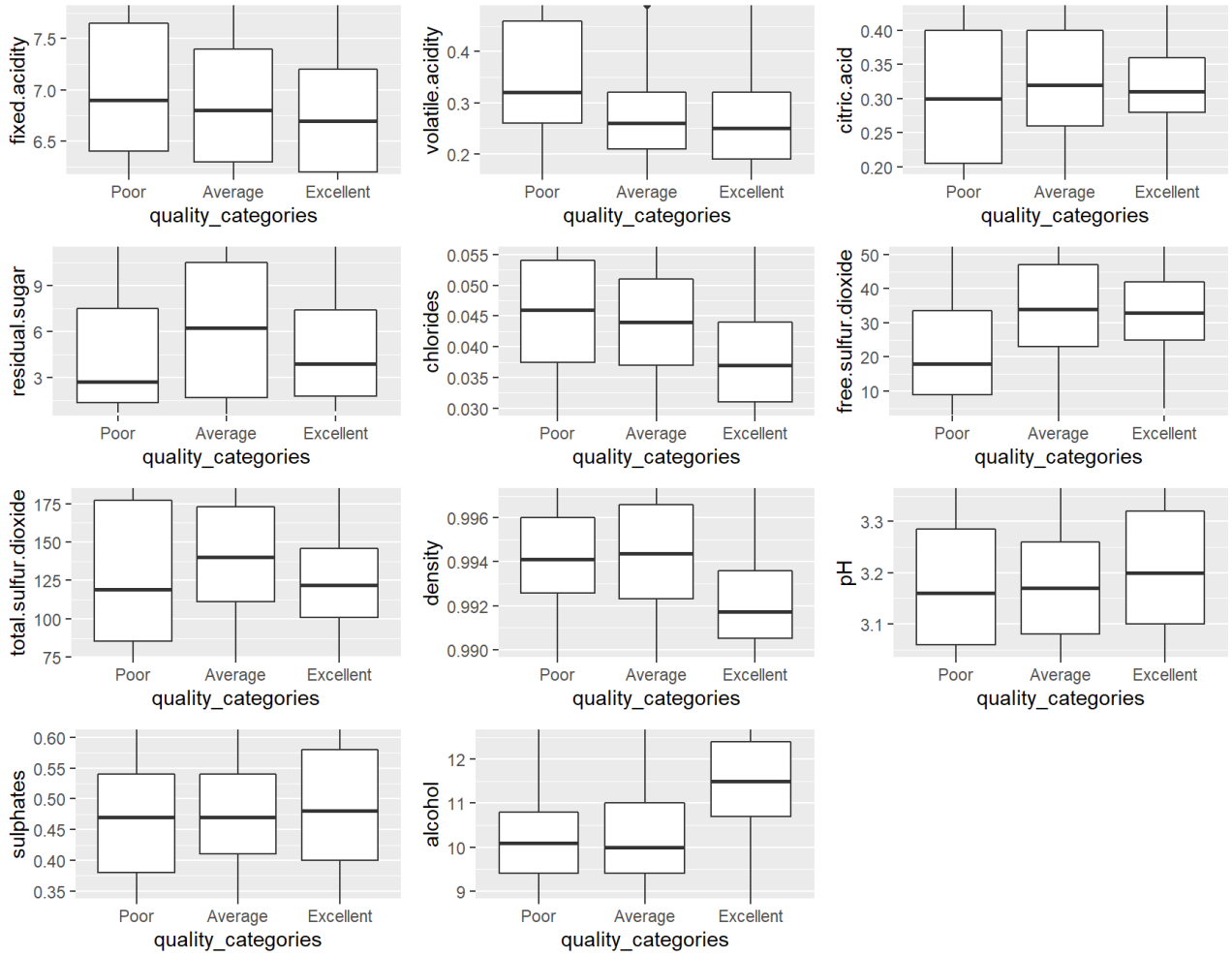
There is a small negative correlation between wine quality and chlorides at -0.21. Chlorides are the amount of salt in the wine and it appears having less is somewhat indicative of better wine quality and this is worth further exploring.




```
##
##  Pearson's product-moment correlation
##
## data:  ww$quality and ww$volatile.acidity
## t = -13.891, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2215214 -0.1676307
## sample estimates:
##       cor
## -0.194723
```

There is also a small negative correlation between wine quality and volatile acidity at -0.19. Dataset documentation indicated that this shows the amount of acetic acid in a wine, and too much can lead to an unpleasant vinegar taste. This negative correlation would indicate that lower levels of volatile acidity (acetic acid) would improve wine quality, and this is also worth further exploring.

I am curious to see how all the variables compare to the quality_categories variable I created. Will there be obvious differences between the different types of wine qualities?




Looking at the mean value of the boxplots for the different variables and how they compare across categories of quality, it is apparent there are some characteristics that stand out. The amount of fixed acidity and volatile acidity in a wine are inversely related to wine quality. Higher amounts of both acidity variables are present in Poor quality wines and are lower in Excellent quality wines.

The same relationship holds true for chloride levels found in our wine samples; more chlorides are found in Poor quality wines and less in Excellent quality wines.

Free sulfur dioxide levels present a relationship that was not apparent when looking at our correlations. Having lower free sulfur dioxide levels are more common in Poor quality wines, while Excellent quality wines tend to have higher levels. According to information from the source dataset, sulfur dioxide prevents microbial growth and oxidation of the wine. When it is also found to be in concentration levels higher than 50ppm, it becomes evident in the nose and taste of the wine. This seems to be a chemical characteristic that improves wine quality.

Density and alcohol content have already been addressed when looking at the correlation between variables. Lower density levels are more common in Excellent quality wines and higher density levels are more common in Poor quality. A higher alcohol content is found in wines of Excellent quality and lower levels are found in those that are Poor quality.

# 6 Bivariate Analysis

### 6.0.1 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 residual sugar variable has a negative relationship with alcohol content but a positive relationship with density levels. It also shows a medium strength correlation with free and total sulfur dioxide levels which might be worth further investigating.

The pH variable shows a negative correlation with fixed acidity and citric acid, which is understandable considering an acid would make the pH levels become lower. It also has a small negative relationship with residual sugar, which is of itself neutral on the pH scale and would explain how it would make an acidic solution more basic.

There are several variables that appear to have a negative correlation with alcohol content: fixed acidity, residual sugar, chlorides, total & free sulfur dioxide levels, as well as density. The same negative relationship exists between quality and fixed acidity, chlorides, total sulfur dioxide levels, and density. There is a medium strength correlation between quality and alcohol content, but I think it has more to do with similar relationships to other variables than just simply higher alcohol content leading to higher quality scores.

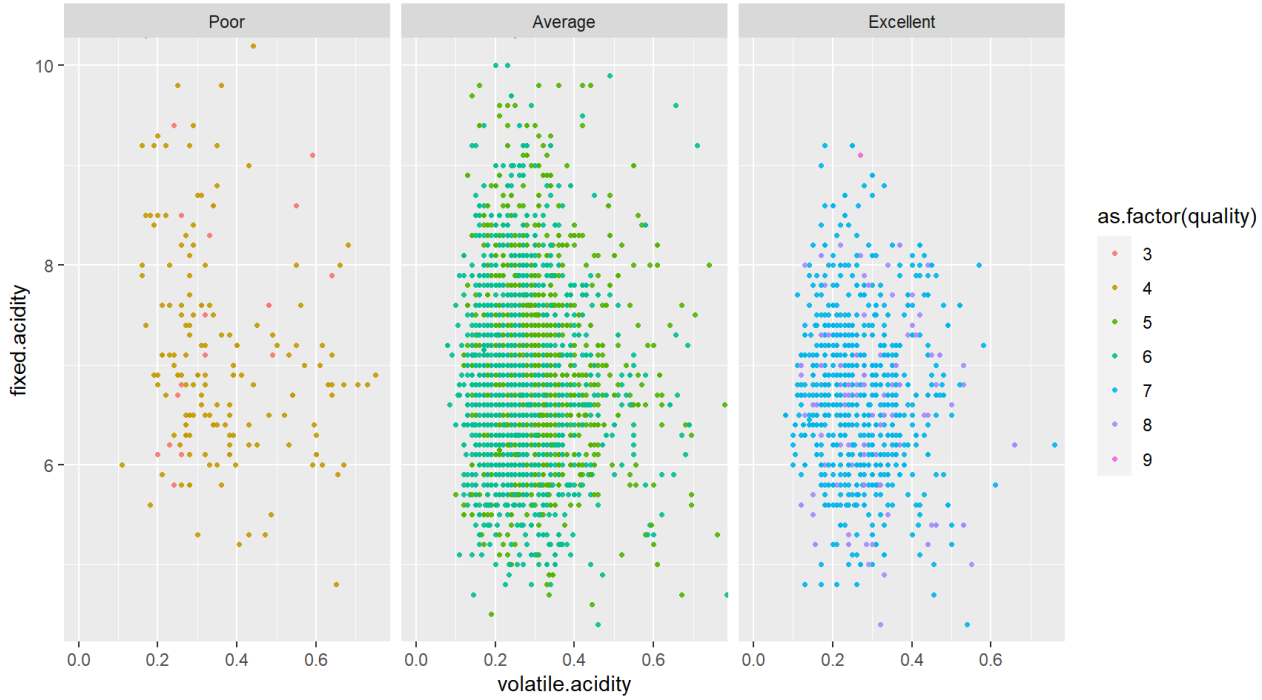
### 6.0.2 Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

When comparing the variables free sulfur dioxide and quality, our correlation shows virtually no relationship with a correlation coefficient of 0.01. When looking at the box plot of free sulfur dioxide levels and their average amounts in each quality category, a different story is told. Poor quality wines on average have less that half the amount of free sulfur dioxide levels as Excellent quality wines. I would have expected a positive correlation to have existed between free sulfur dioxide levels and quality, but this is not the case.

### 6.0.3 What was the strongest relationship you found?

The strongest relationship exists between density and residual sugar with a correlation coefficient of 0.84. This is not a surprise because having a higher concentration of sugar in a liquid would increase its density.

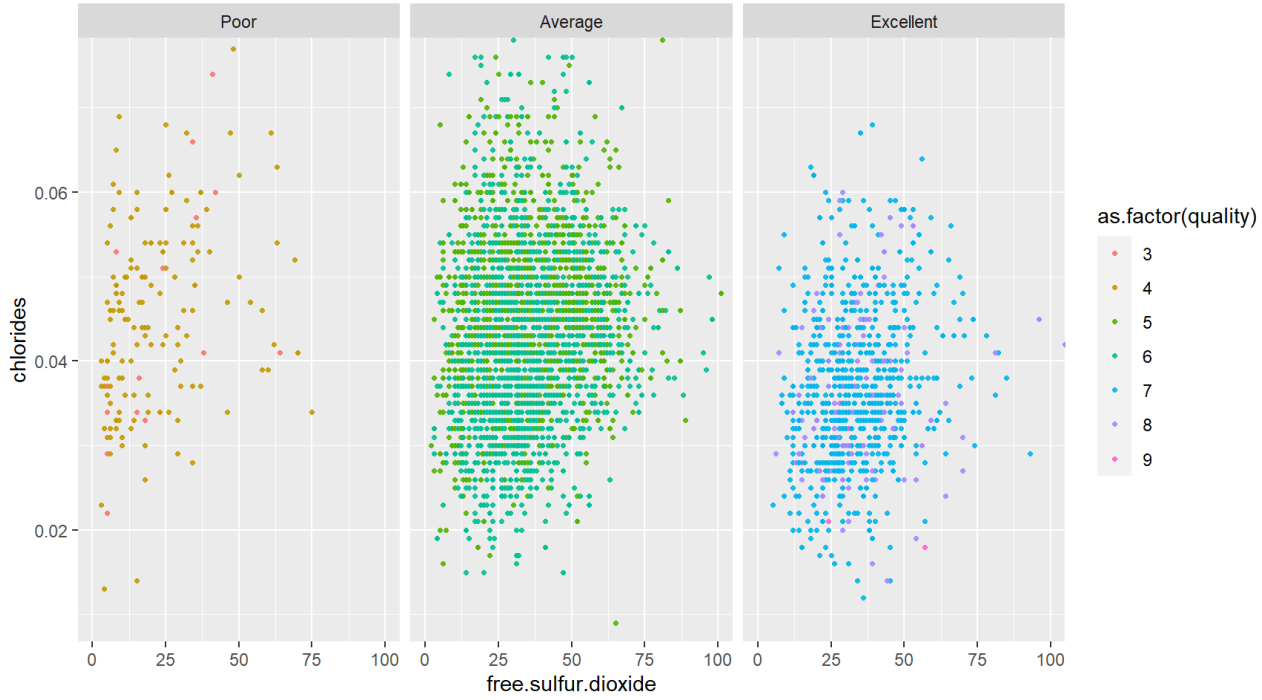
# 7 Multivariate Plots Section




```
## ww$quality_categories: Poor
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.200   6.400   6.900   7.181   7.650  11.800
## ------------------------------------------------------------
## ww$quality_categories: Average
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.800   6.300   6.800   6.876   7.400  14.200
## ------------------------------------------------------------
## ww$quality_categories: Excellent
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.900   6.200   6.700   6.725   7.200   9.200
```

```
## ww$quality_categories: Poor
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.110   0.260   0.320   0.376   0.460   1.100
## ------------------------------------------------------------
## ww$quality_categories: Average
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.2100  0.2600  0.2771  0.3200  0.9650
## ------------------------------------------------------------
## ww$quality_categories: Excellent
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.1900  0.2500  0.2653  0.3200  0.7600
```

I chose to start with a comparison of fixed acidity vs volatile acidity against quality. From the boxplots it revealed that both variables are present in lower levels as wine quality increases. From Poor to Excellent quality wines our mean and median values of both variables go down. Average and Excellent wine qualities are similar in values, but Excellent is still lower and both are quite a bit lower than Poor quality wines. The scatterplot shows this trend as the cluster moves down and toward the left as wine quality improves.

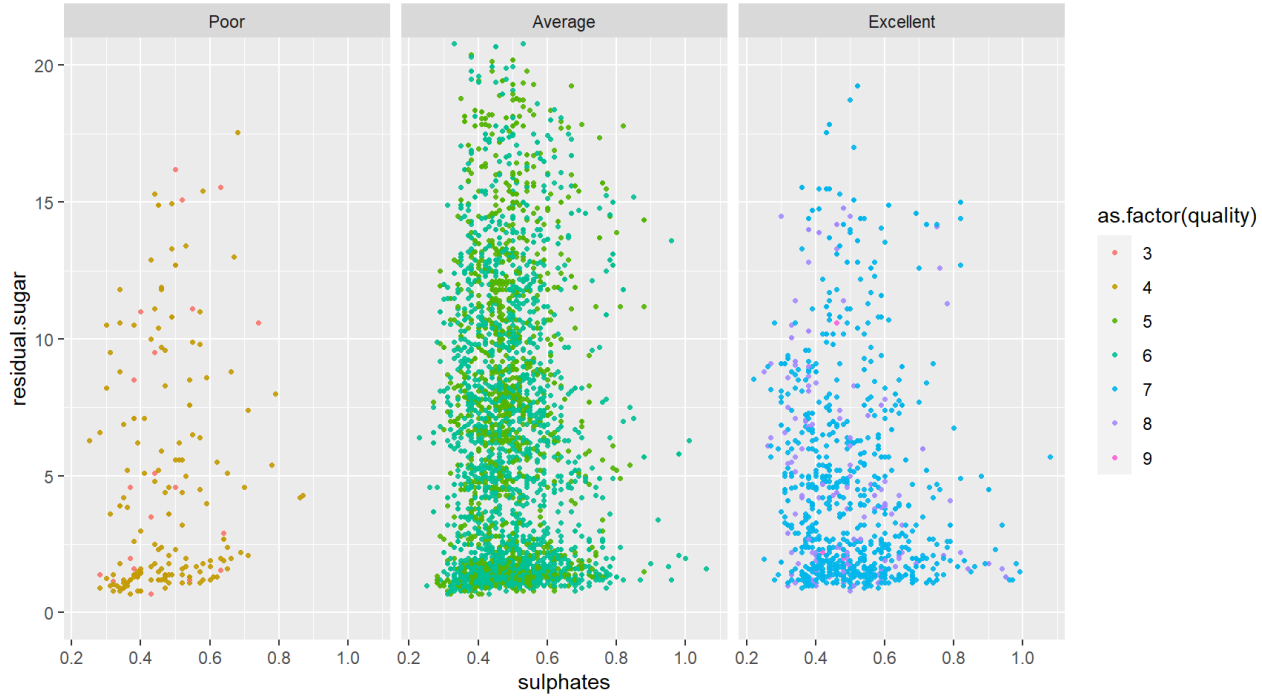



```
## ww$quality_categories: Poor
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01300 0.03750 0.04600 0.05056 0.05400 0.29000
## ------------------------------------------------------------
## ww$quality_categories: Average
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00900 0.03700 0.04400 0.04774 0.05100 0.34600
## ------------------------------------------------------------
## ww$quality_categories: Excellent
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.03100 0.03700 0.03816 0.04400 0.13500
```

```
## ww$quality_categories: Poor
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    9.00   18.00   26.63   33.50  289.00
## ------------------------------------------------------------
## ww$quality_categories: Average
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   23.00   34.00   35.96   47.00  131.00
## ------------------------------------------------------------
## ww$quality_categories: Excellent
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00   25.00   33.00   34.55   42.00  108.00
```

As chloride levels decrease and as free sulfur dioxide levels increase, the quality of wine improves. This is evident in the shift of our cluster down and to the right. The mean and median values for chloride levels decrease as quality levels improve. When looking at the free sulfur dioxide mean and median, we can see that the Average quality wines are slightly higher in free sulfur dioxide levels than the Excellent quality wines. However, both Average and Excellent quality wines have significantly more free sulfur dioxide levels than Poor

quality wines. The fact that Average quality wines were slightly higher is why no obvious correlation was present when looking at correlation coefficients. Both chloride levels and free sulfur dioxide levels affect wine quality.

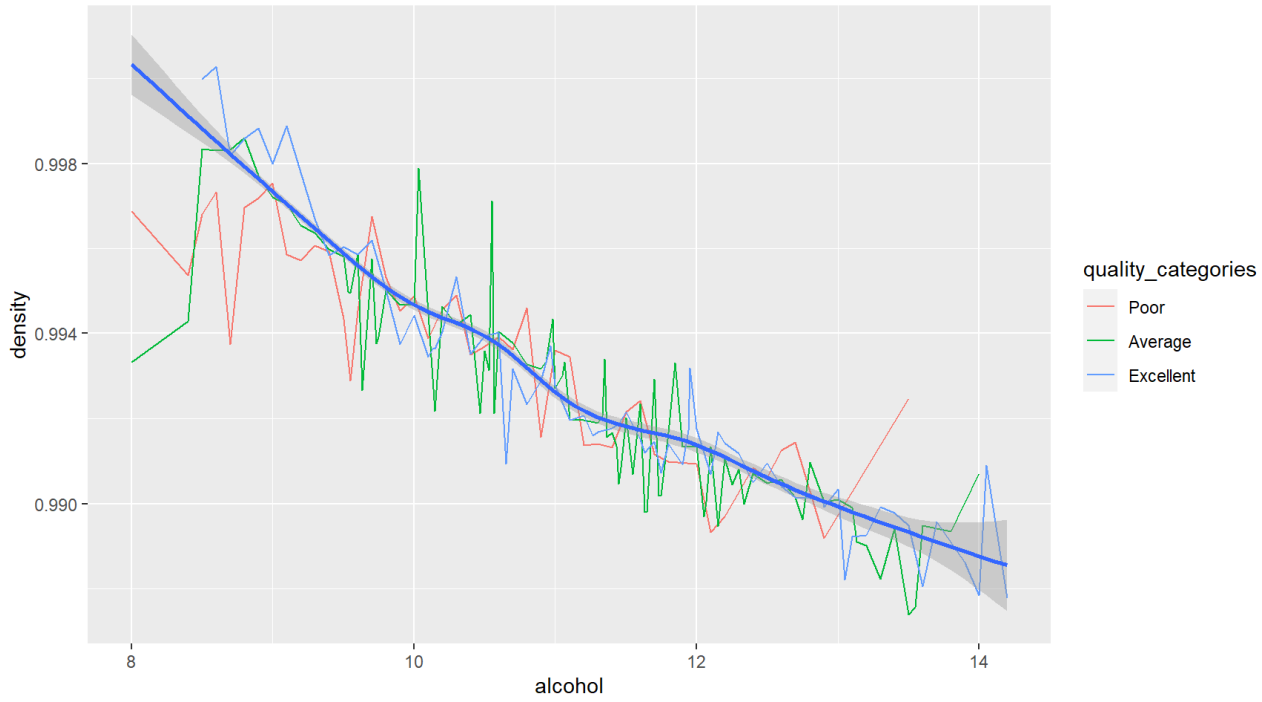



```
## ww$quality_categories: Poor
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.700   1.350   2.700   4.821   7.500  17.550
## ------------------------------------------------------------
## ww$quality_categories: Average
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.600   1.700   6.200   6.798  10.500  65.800
## ------------------------------------------------------------
## ww$quality_categories: Excellent
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.800   1.800   3.875   5.262   7.400  19.250
```

```
## ww$quality_categories: Poor
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.250   0.380   0.470   0.476   0.540   0.870
## ------------------------------------------------------------
## ww$quality_categories: Average
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2300  0.4100  0.4700  0.4876  0.5400  1.0600
## ------------------------------------------------------------
## ww$quality_categories: Excellent
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2200  0.4000  0.4800  0.5001  0.5800  1.0800
```

We can see that the residual sugar levels are lowest in Poor quality wines and highest in Average quality wines, while Excellent wines tend to fall somewhere in the middle. When looking at the median values of each of the quality categories, Poor quality wines have a median of 2.7, Average have a median of 6.2, and Excellent have a median of 3.9. There seems to be a sweet spot for ideal sugar levels, but not enough sugar is most common with Poor quality wines.

Our sulphate levels are all very similar, but there is a very small correlation between sulphate levels and wine quality. Our plot shows better quality wines have slightly more sulphates and our mean and median values reflect this increase in sulphate content as wine quality improves.




I just wanted to reiterate that regardless of quality category, density decreases with an increase of alcohol content.

# 8 Multivariate Analysis

## 8.0.1 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?

Using a scatterplot, the relationship between variables became even more apparent. Some of the stronger relationships I investigated were between fixed acidity levels, volatile acidity levels, chloride levels, free sulfur dioxide levels, and alcohol content. These all had an impact on quality.

Fixed acidity and volatile acidity both decreased in amounts as wine quality improved. The amount of acetic and tartaric acid in wines has a clear impact on quality ratings and less of each resulted in improved taste.

As chloride levels in wine decrease, the quality increased. Chloride levels were the amount of sodium chloride, salt, that is present in a wine. Lower salt levels seemed to correspond to better quality wines and had a clear impact on quality ratings.

Free sulfur dioxide levels tended to increase as our wine quality increased. Levels seemed to plateau between Average and Excellent quality wines, but when compared to Poor quality wines, there is a significantly higher chloride content. This additive clearly has an impact on quality ratings.

Sulphates are another SO2 level indicator, and while not as strong of a correlation as free sulfur dioxide levels, the same relationship held true. Increase sulphate levels increased quality ratings. Sulphates and free sulfur dioxide levels are both measures of SO2 that help with microbial growth and oxidation of wines, and having higher levels improves quality.

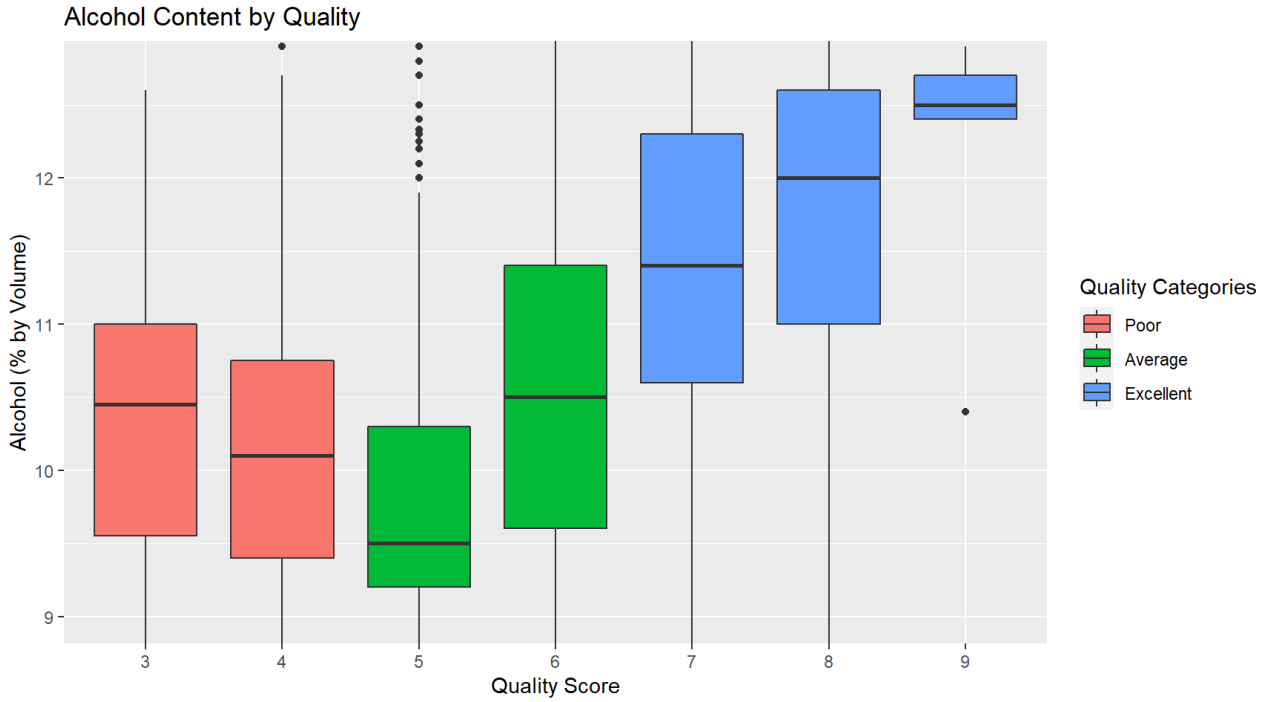
### 8.0.2 Were there any interesting or surprising interactions between features?

I found the residual sugar levels to be an interesting feature that affected wine quality. Not enough sugar levels seem to be common with Poor quality wines while higher sugar content seemed to be common with Average quality wines. A ‘sweet-spot’ existed somewhere between the two that corresponded to an Excellent quality wine.

---

# 9 Final Plots and Summary

### 9.0.1 Plot One

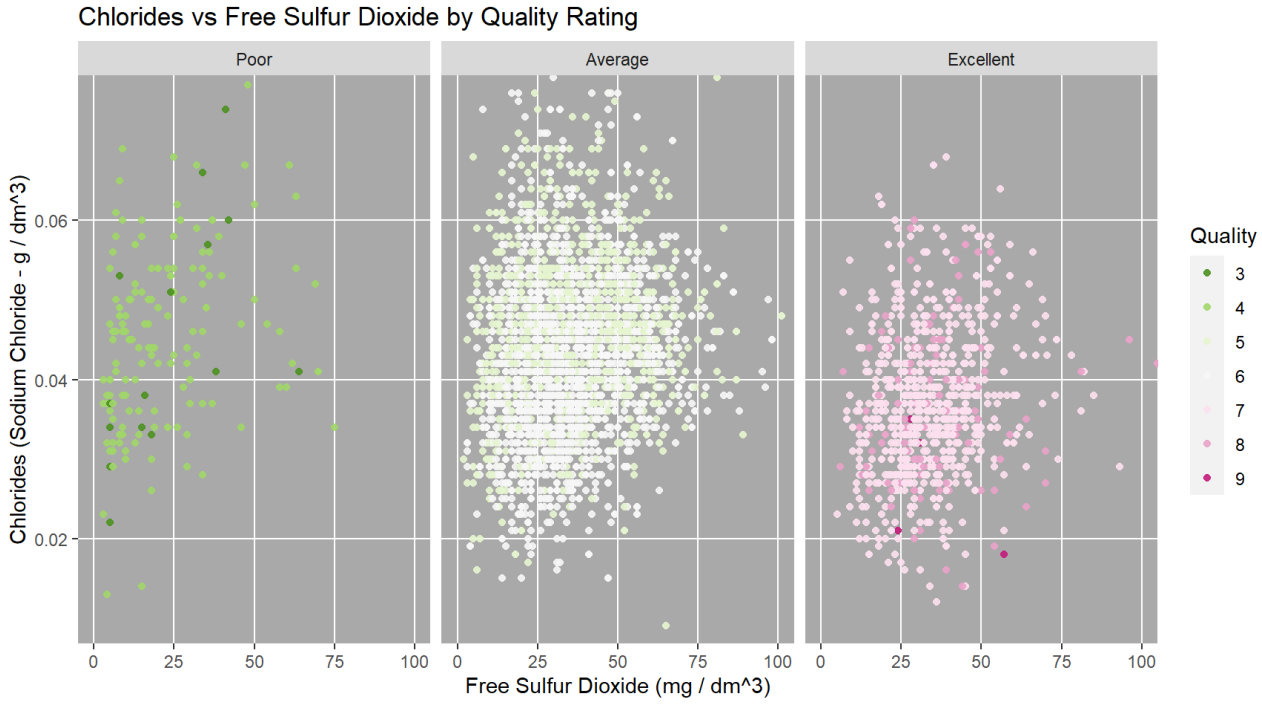



```
## ww$quality_categories: Poor
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.40   10.10   10.17   10.80   13.50
## ------------------------------------------------------------
## ww$quality_categories: Average
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.40   10.00   10.27   11.00   14.00
## ------------------------------------------------------------
## ww$quality_categories: Excellent
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50   10.70   11.50   11.42   12.40   14.20
```

## 9.0.2 Description One

Alcohol content is much higher for Excellent quality wines and is about the same for both Poor and Average quality wines. Excellent quality wines have approximately 1.5 % higher alcohol content than the other two. Poor and Average being around 10%, and Excellent being around 11.5%. While on the surface it might appear that alcohol content correlates to better wine quality, but further investigation showed this just happens to correspond with other chemical components of our wines.

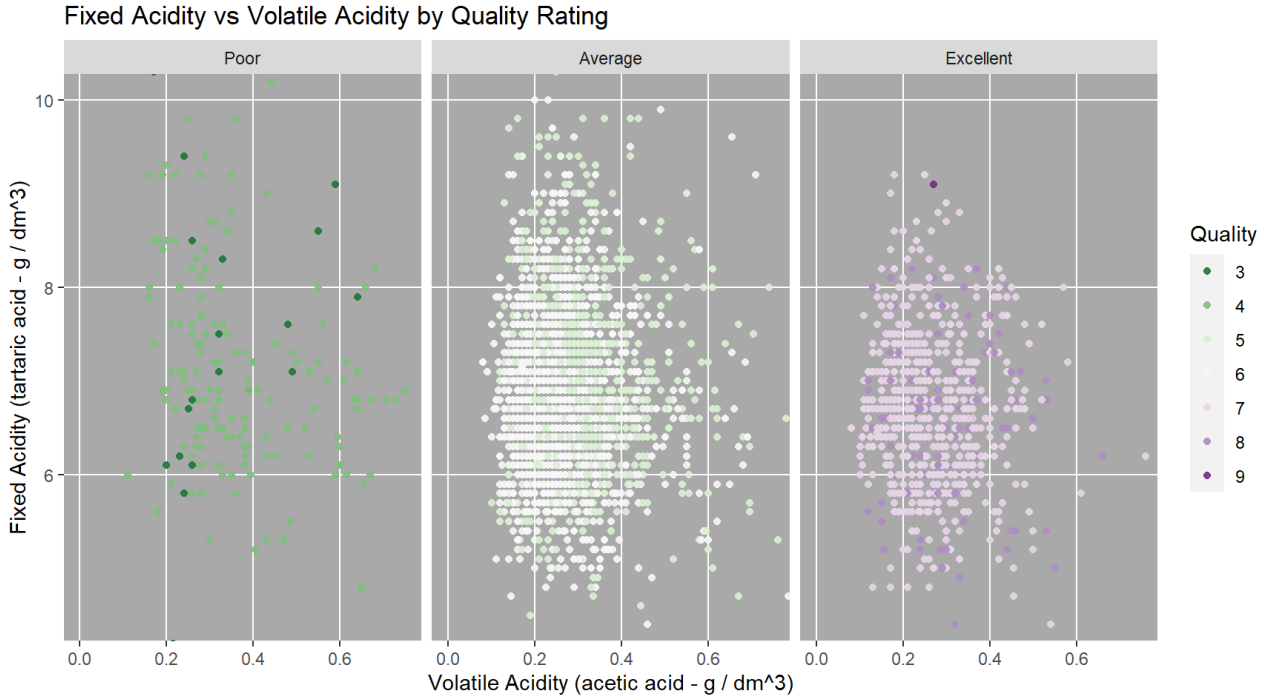
## 9.0.3 Plot Two




## 9.0.4 Description Two

Chloride levels decrease and free sulfur dioxide levels increase as the quality of wine increases. The scatterplot shifts down and to the right to indicate this trend. Both chemicals play a role in the overall flavor profile of wine.

## 9.0.5 Plot Three




## 9.0.6 Description Three

Fixed acidity levels and volatile acidity levels decrease as wine quality increases. The scatterplot shows a shift down and to the left to indicate this trend. Both chemicals play a role in the flavor profile of wine.

---

# 10 Reflection

At first glance it appeared alcohol content played a large role in white wine quality. Excellent quality wines had median and mean alcohol levels much higher than lower quality wines and a correlation coefficient of .44 existed between alcohol content and wine quality, which was the highest correlation between wine quality and any other variables. Upon further inspection, it was apparent this was more of a coincidence than an actual correlation, other chemical factors played a more important role in white wine quality than alcohol content alone. Alcohol content just so happened to have similar relationships with these variables that more significantly impacted wine quality.

Volatile and fixed acidity levels both decreased as wine quality improved. Fixed acidity also negatively correlated with alcohol content. These two acidity variables are responsible for acetic and tartaric acid concentrations found in our wine samples, lower levels of these two acids produced less vinegar taste, which was a preferred quality in wines. Chloride levels also have a negative relationship with both quality and alcohol content. Having less amounts of sodium chloride, salt, in our wines improves the taste quality and happens to correspond with higher alcohol levels.

Free sulfur dioxide levels improved wine quality with increased amounts. Higher levels of this variable improved wine quality. Sulfur dioxide levels are used to prevent microbial growth and oxidation of wine, and it seems the flavor sulfur dioxide adds to a wine is desirable.

My investigation into density and pH variables lead to a dead end as I continued to explore them. pH levels are naturally going to change as alcohol and acidity levels increase. Density will also decrease with higher alcohol levels or increase with more sugar. Investigating these variables did no reveal much about wine preferences.

It seems the perfect wine has a moderate amount of sugar, lower levels of acetic acid, lower levels of chlorides, higher levels of sulfur dioxide and a higher alcohol content. But we need to keep in mind that this analysis is based on 3 wine experts, which does not necessarily reflect the average person's opinion. It would be interesting to do the same type of analysis with average wine drinkers and to see the differences in trends. This could be an idea for a way to further explore wine preferences and characteristics, after all, wine companies are selling to the general population, not just wine experts.

## 10.1 Struggles & Successes

I found it challenging to focus in on what variables had a significant impact on wine quality. Looking at just correlation coefficients did not tell the whole story. Once I decided to group the wine qualities into categories and then look at characteristics of each category, more

information came to light. I also had a hard time deciding how to visually represent some of the graphs. I chose to use the scatterplots divided into each quality category to show the variations between quality groups. I also found boxplots to be quite helpful in my analysis.

# 11 References

http://www.sthda.com/english/wiki/scatter-plot-matrices-r-base-graphs (http://www.sthda.com/english/wiki/scatter-plot-matrices-r-base-graphs)

https://www.rdocumentation.org/packages/psych/versions/1.9.12.31/topics/pairs.panels (https://www.rdocumentation.org/packages/psych/versions/1.9.12.31/topics/pairs.panels)

https://stackoverflow.com/questions/33666935/how-to-understand-which-variables-are-correlated-with-each-other/33667708 (https://stackoverflow.com/questions/33666935/how-to-understand-which-variables-are-correlated-with-each-other/33667708)

https://www.displayr.com/how-to-create-a-correlation-matrix-in-r/ (https://www.displayr.com/how-to-create-a-correlation-matrix-in-r/)