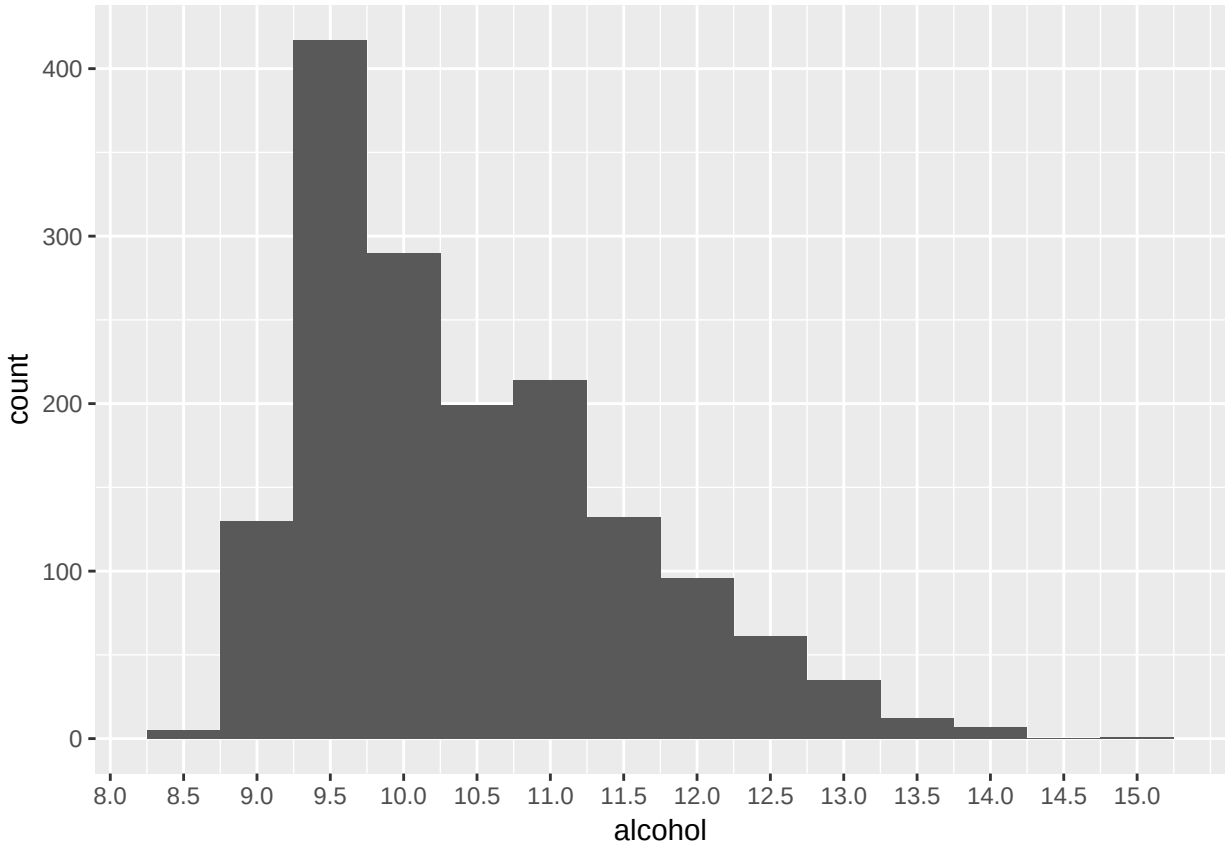
Red Wine Quality by Andrew Phan

```
## [1] "/Users/andrewphan/Downloads"
```

I am going to be exploring a data set about 1599 red wines and 11 of of their chemical properties. To go along with the properties, overall quality ratings, which were medians of ratings from 0-10 from at least three wine experts, were included for each red wine.

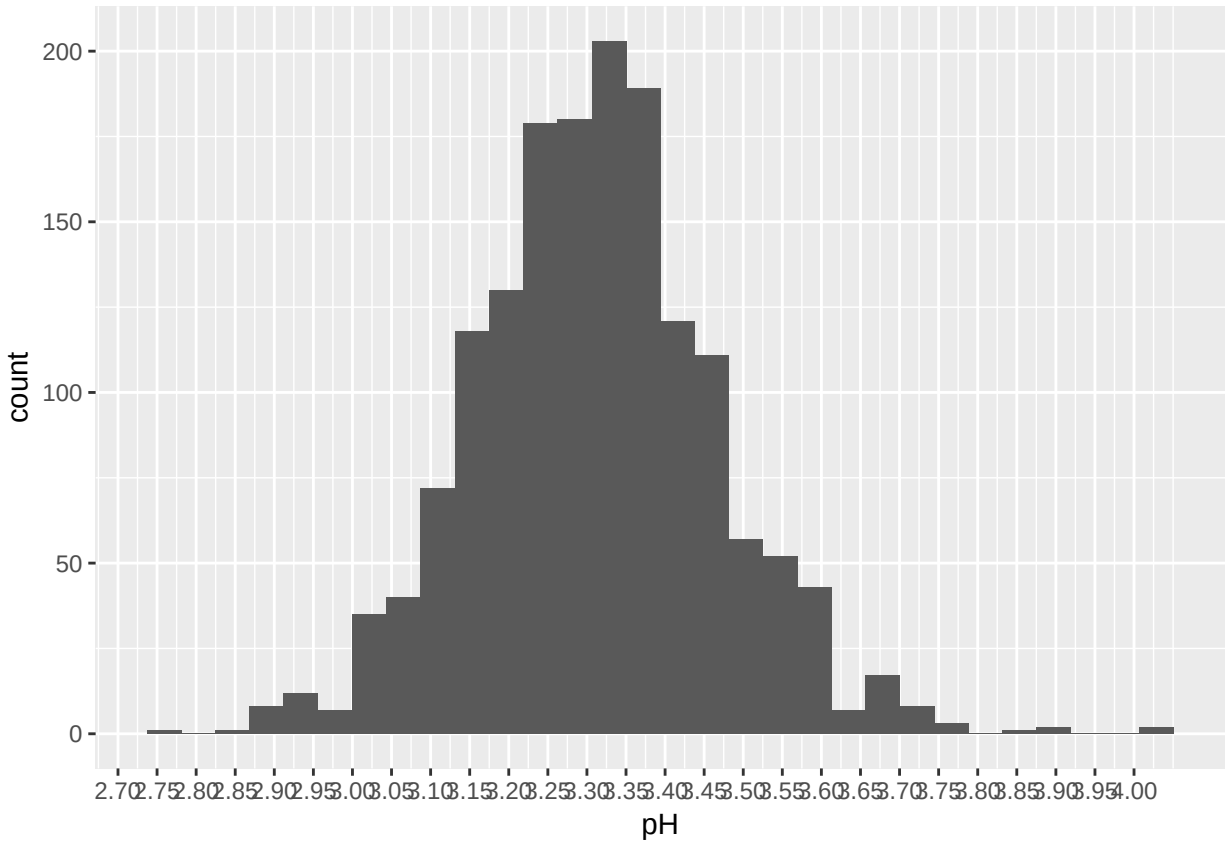
Univariate Plots Section




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

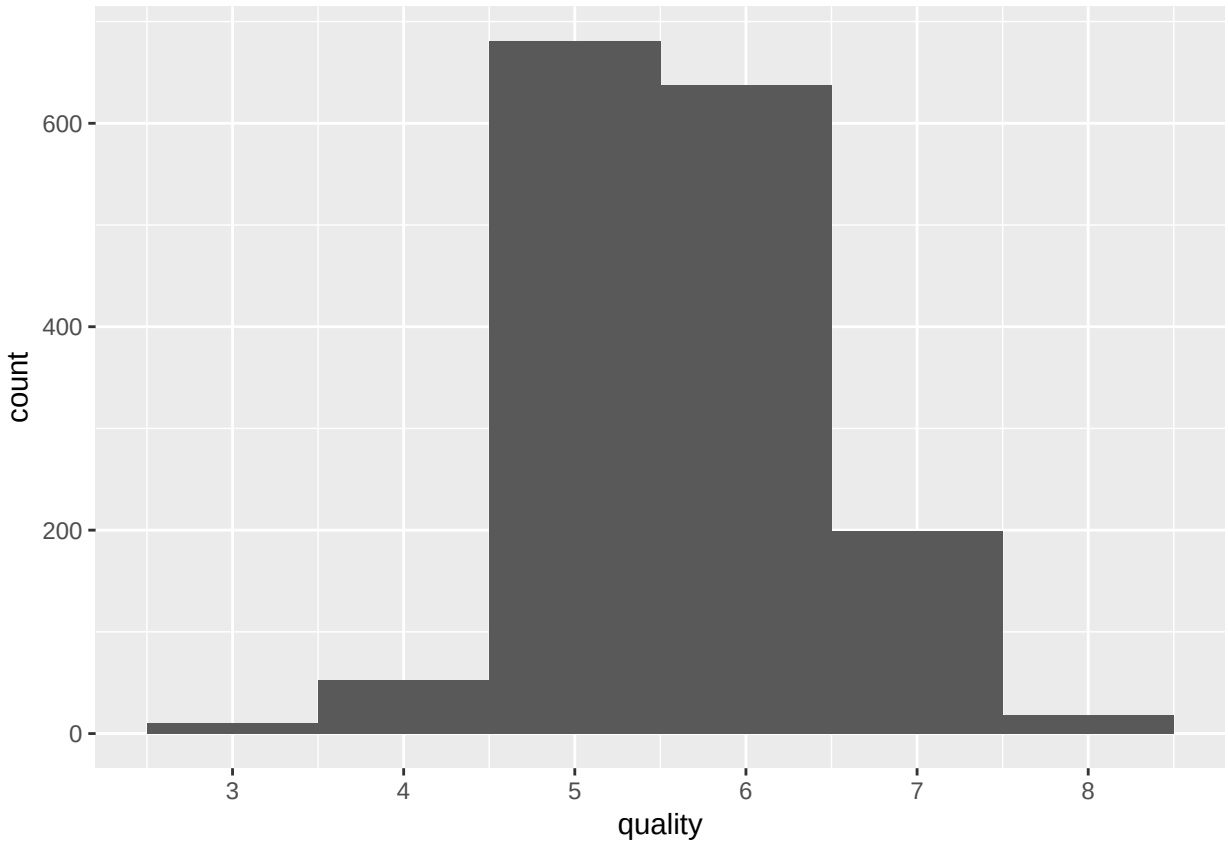
The lowest alcohol percentage a red wine had was 8.4%, and the highest was 14.9%. It seems that the distribution for alcohol percentage, by volume, is skewed right. There seems to be a very high count of red wines which are approximately 9.5% alcohol.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

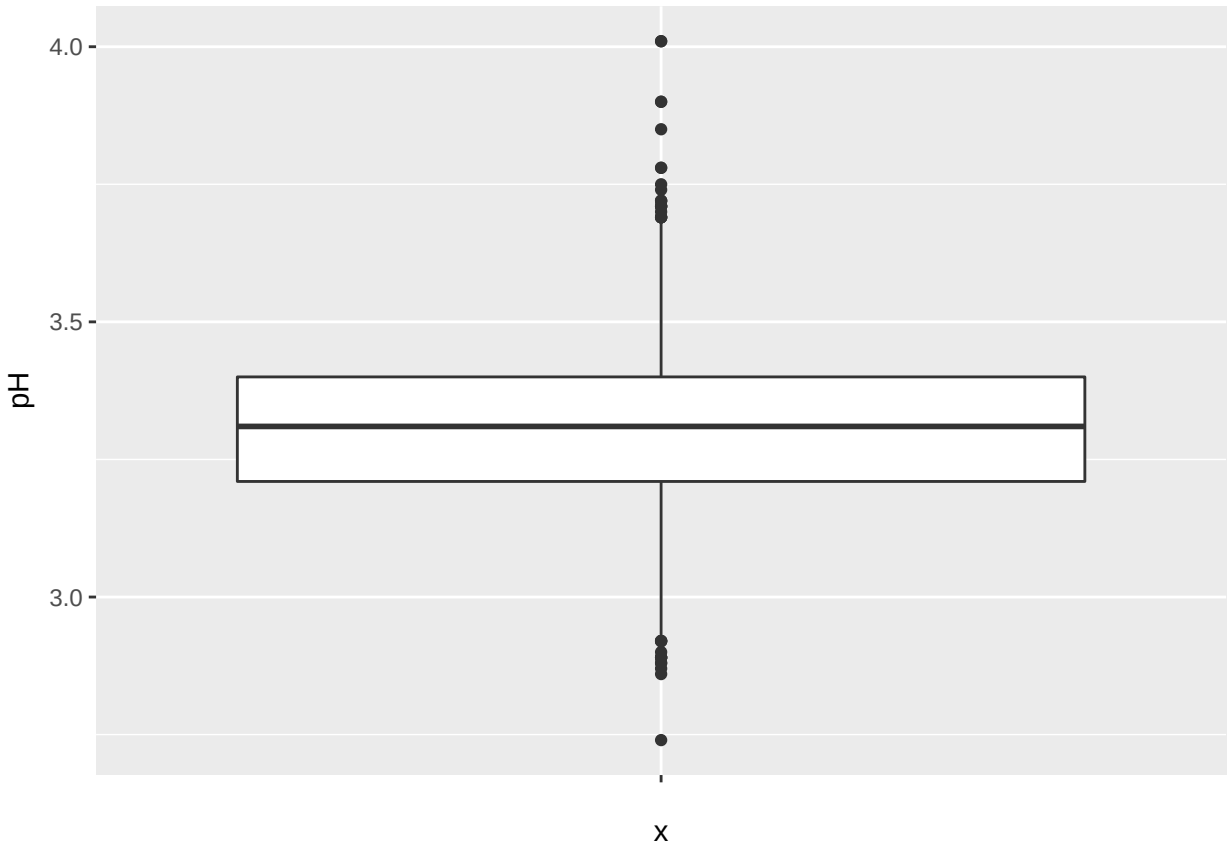



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010
```

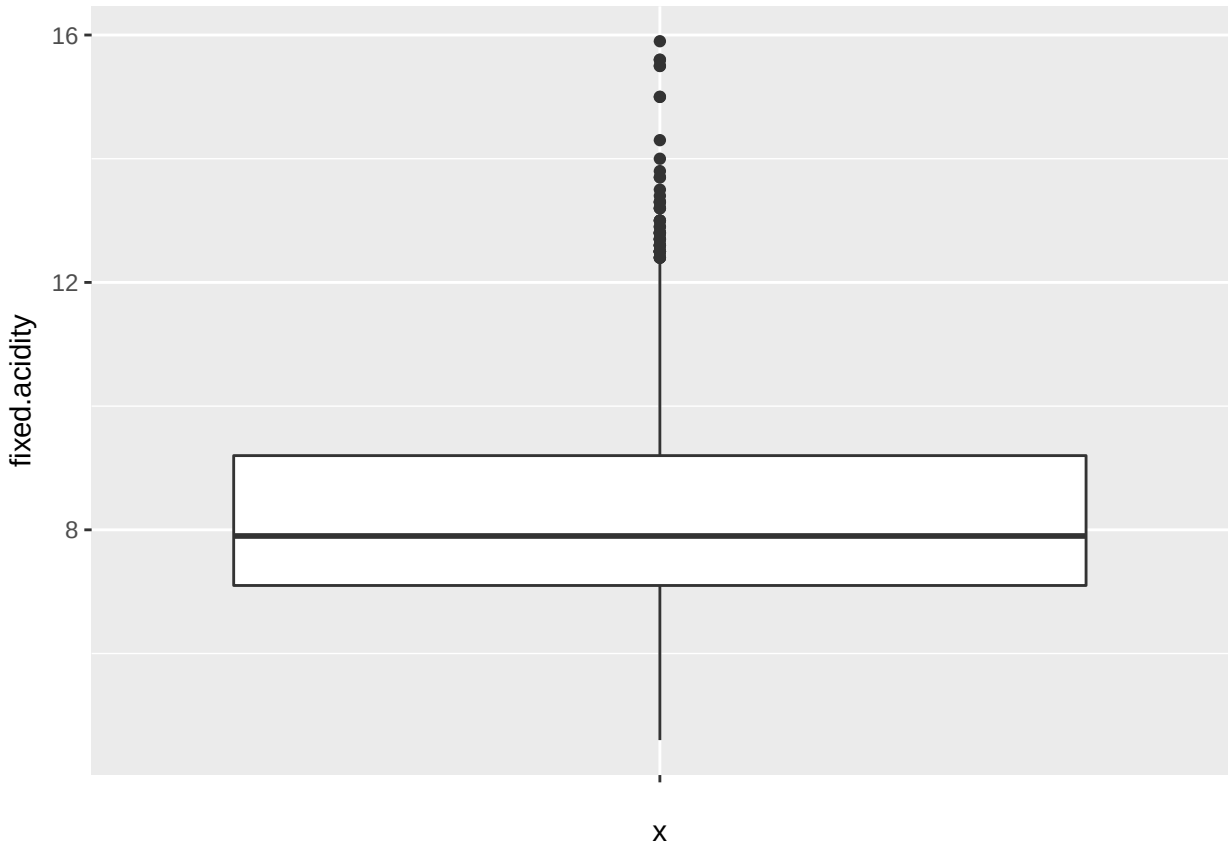
The pH levels for the red wines seem to follow a normal distribution pattern. It seems there are a few possible outliers around 3.85 pH and 2.75 pH. Nothing too out of the ordinary.



I wanted to take a look at the distribution of quality ratings for the red wines to make sure nothing was out of the ordinary. This histogram gives us a good idea of whether the experts were being too harsh, or too easy. The normal distribution for quality ratings with no ratings below 3 and none above 8 gives me the reassurance that the experts were not being overly critical or easygoing.



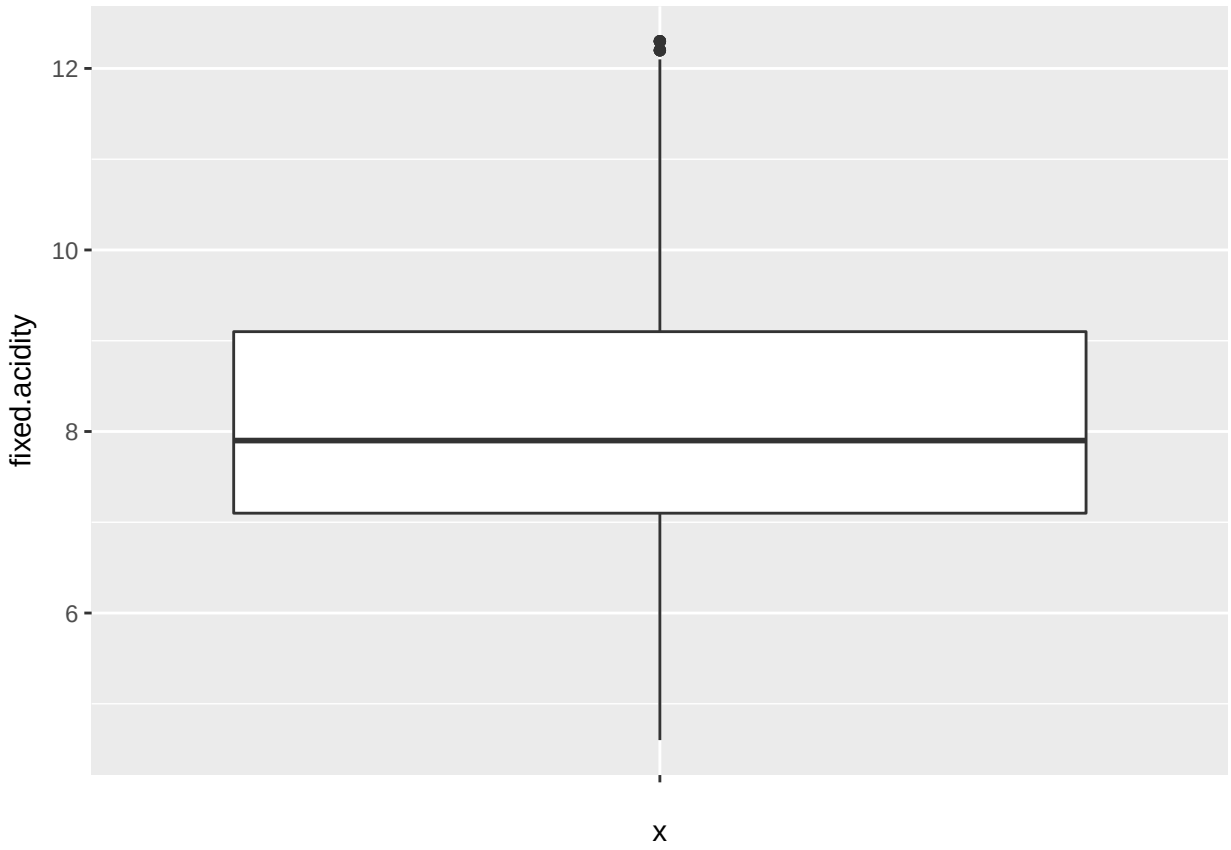
I made a boxplot for pH so I could see if the outliers were having a significant effect on the distribution of the data; they don't seem to be much of a problem based on the boxplot.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
```

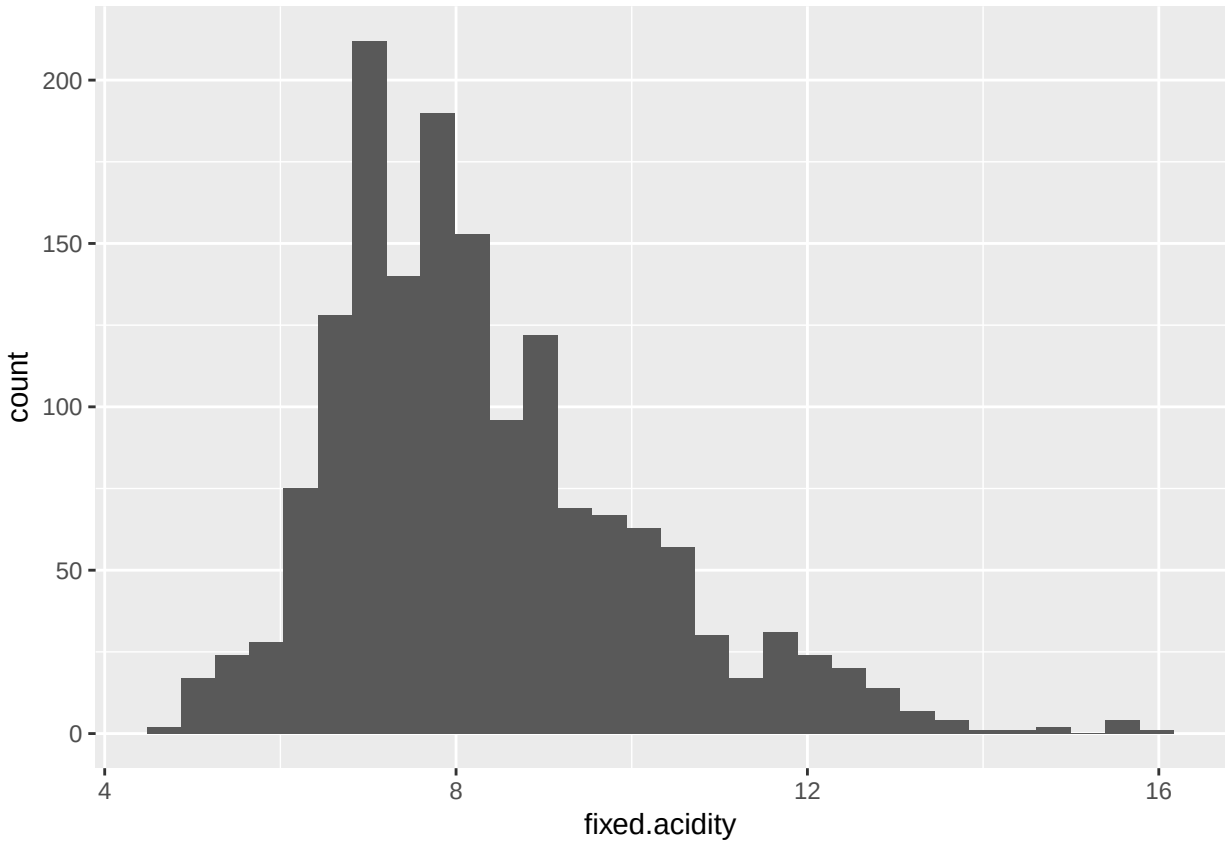
The boxplot for fixed acidity seems to show there are a lot of outliers above 1.5 * IQR. It might be a good idea to see what it looks like with the outliers removed.

Most red wines have fixed acidity levels between 7.10 and 9.20. The IQR is 2.1. 1.5*IQR = 3.15. Removing the cases where the fixed acidity levels are above Q3+3.15 (outliers) might give us a better look at the distribution.



The distribution seems to look more normal now. With the knowledge that there are only outliers above $1.5IQR$, *perhaps we can take a look at the histogram. I would expect the distribution to look slightly skewed right, as it seems the there are a good number of outliers above Q3 + 1.5*IQR.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The prediction was correct!

Univariate Analysis

Structure of my dataset:

The dataset has 13 columns, and 1599 rows. Something to note is one of the columns, X, is just an indexing column. It can likely be deleted. I couldn't find anything in the documentation about the columns about the X column. It is likely the column is redundant.

Main feature of interest:

The main feature of interest in this dataset is the quality column. I'm interested in which of the measured chemical properties have the most influence on the quality rating.

Aside from that, I'm interested in the relationships between some of the chemical properties.

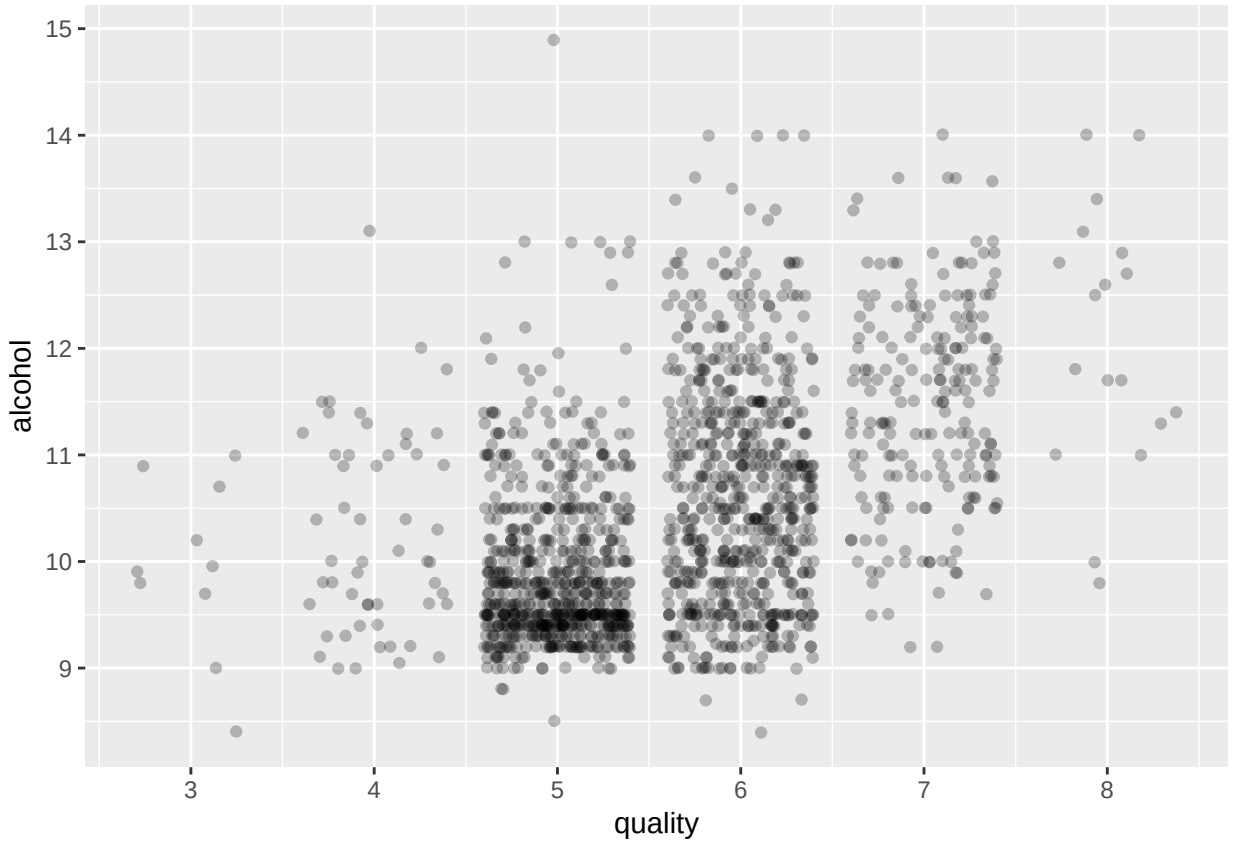
Extra notes about the dataset:

The dataset is fairly clean and tidy. Each variable is a column, and each observation is a row. There doesn't seem to be any data quality issues, either.

Unusual distributions/findings:

As noted in one of my comments, fixed acidity levels had a skewed right distribution. Alcohol percentage by volume is also skewed right, with an extremely high count of red wines with approximately 9.5% percentage alcohol.

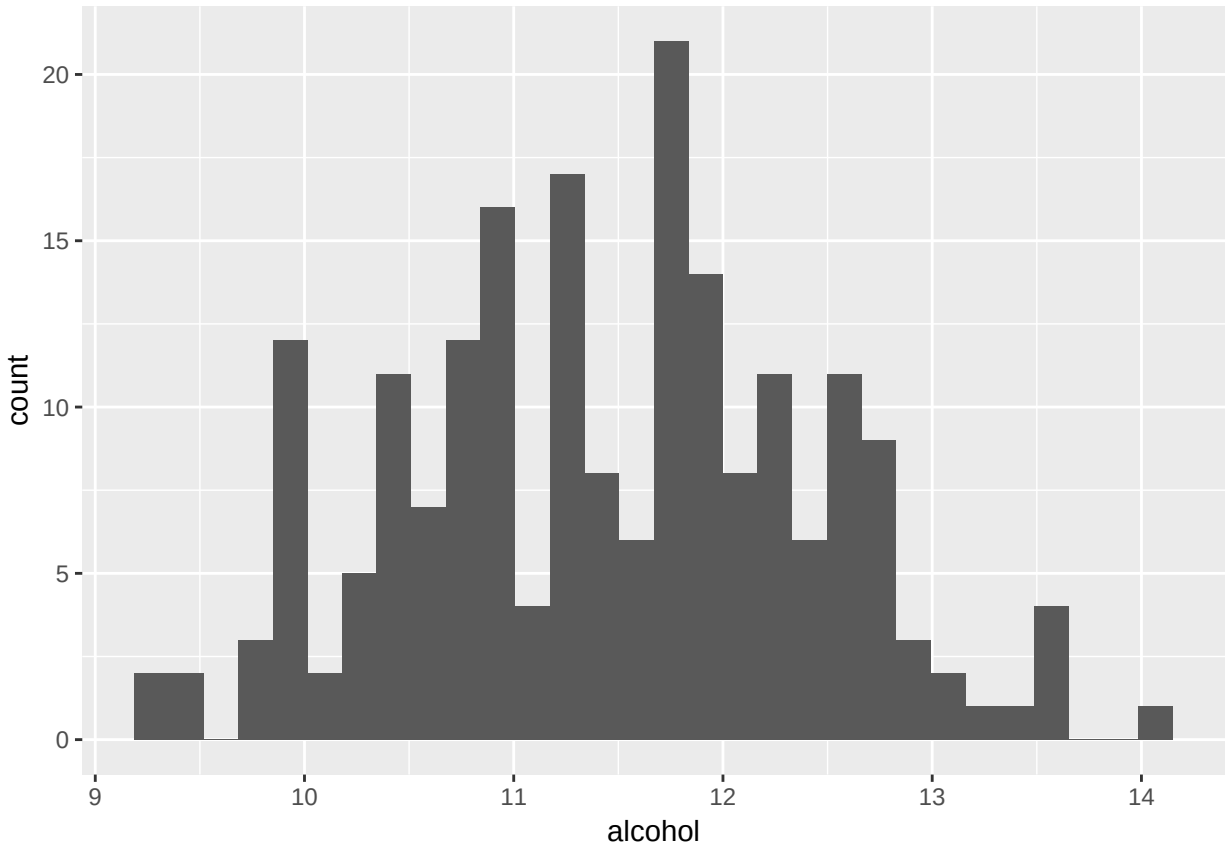
Bivariate Plots Section



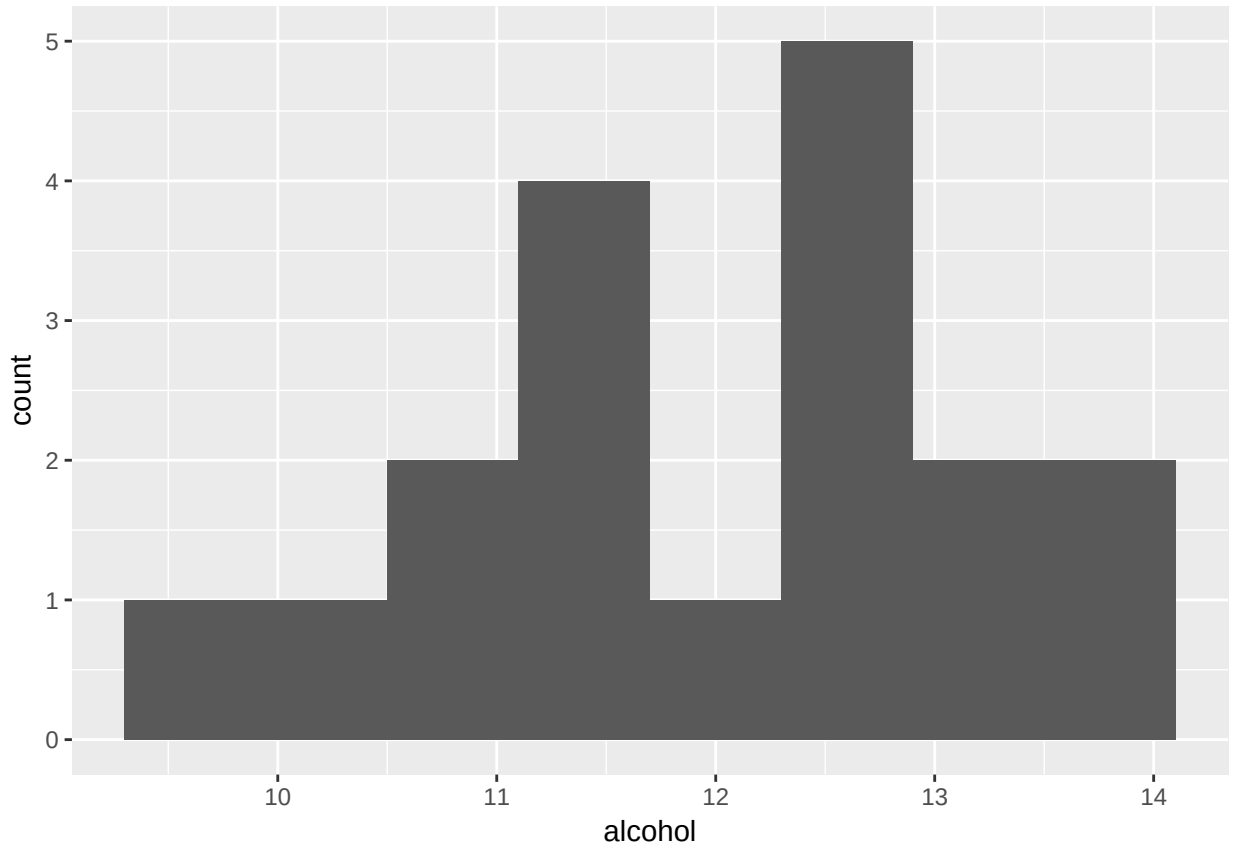

```
##
##  Pearson's product-moment correlation
##
## data:  rw$alcohol and rw$quality
## t = 21.639, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4373540 0.5132081
## sample estimates:
##       cor
## 0.4761663
```

It seems there is a moderate positive relationship between alcohol percentage and quality. I wanted to dig deeper and examine if the higher quality wines were focused around certain alcohol percentage levels.

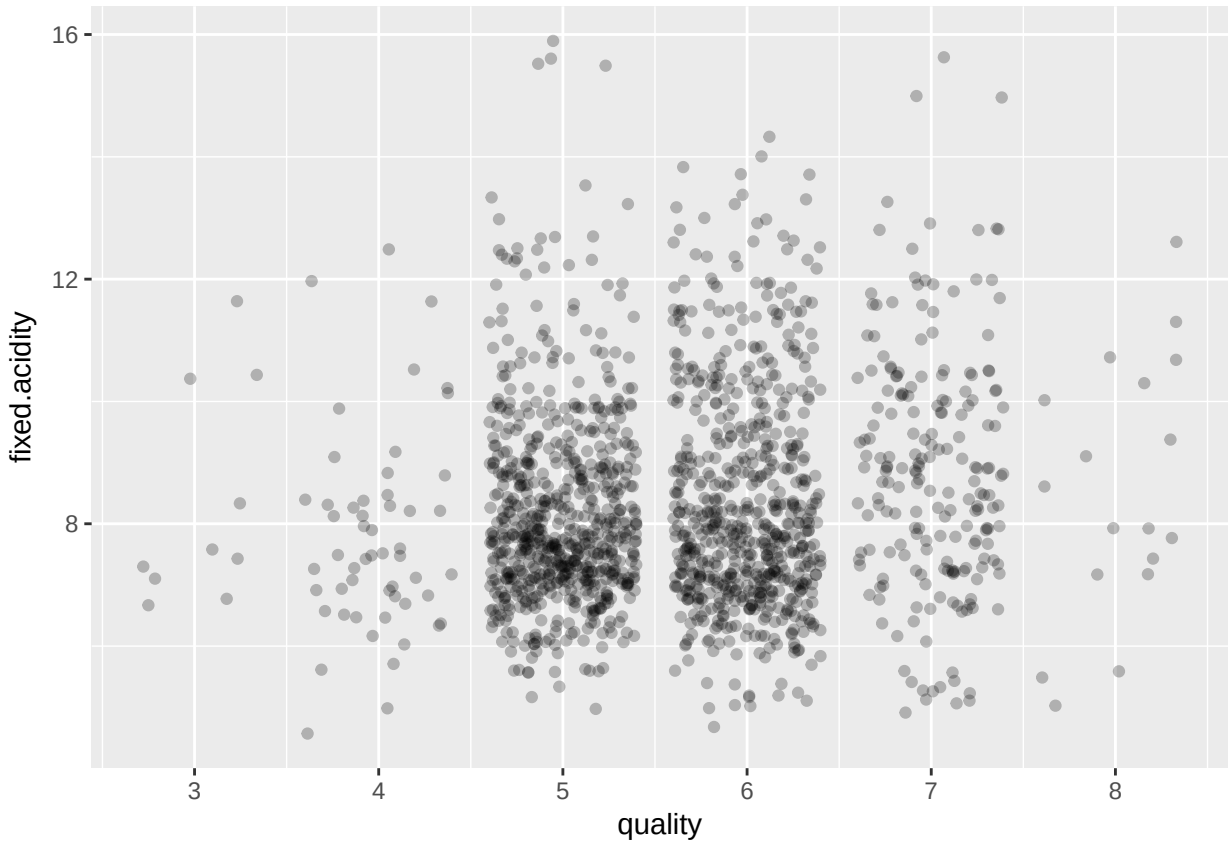
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The distribution of alcohol percentage levels for red wines which were rated 7 is normal.



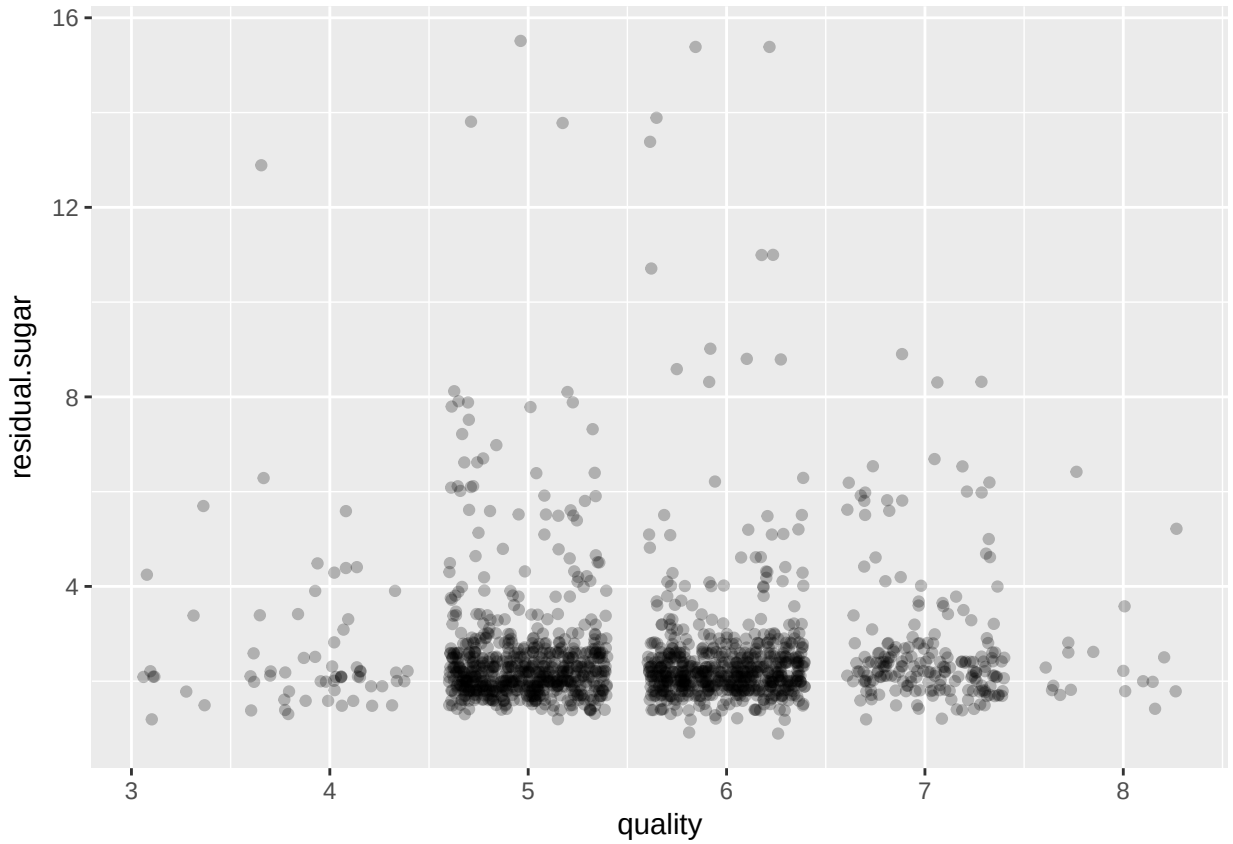
The distribution, for the most part, looks normal based on the small number of observations we have.




```
##
##  Pearson's product-moment correlation
##
## data:  rw$fixed.acidity and rw$quality
## t = 4.996, df = 1597, p-value = 6.496e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.07548957 0.17202667
## sample estimates:
##       cor
## 0.1240516
```

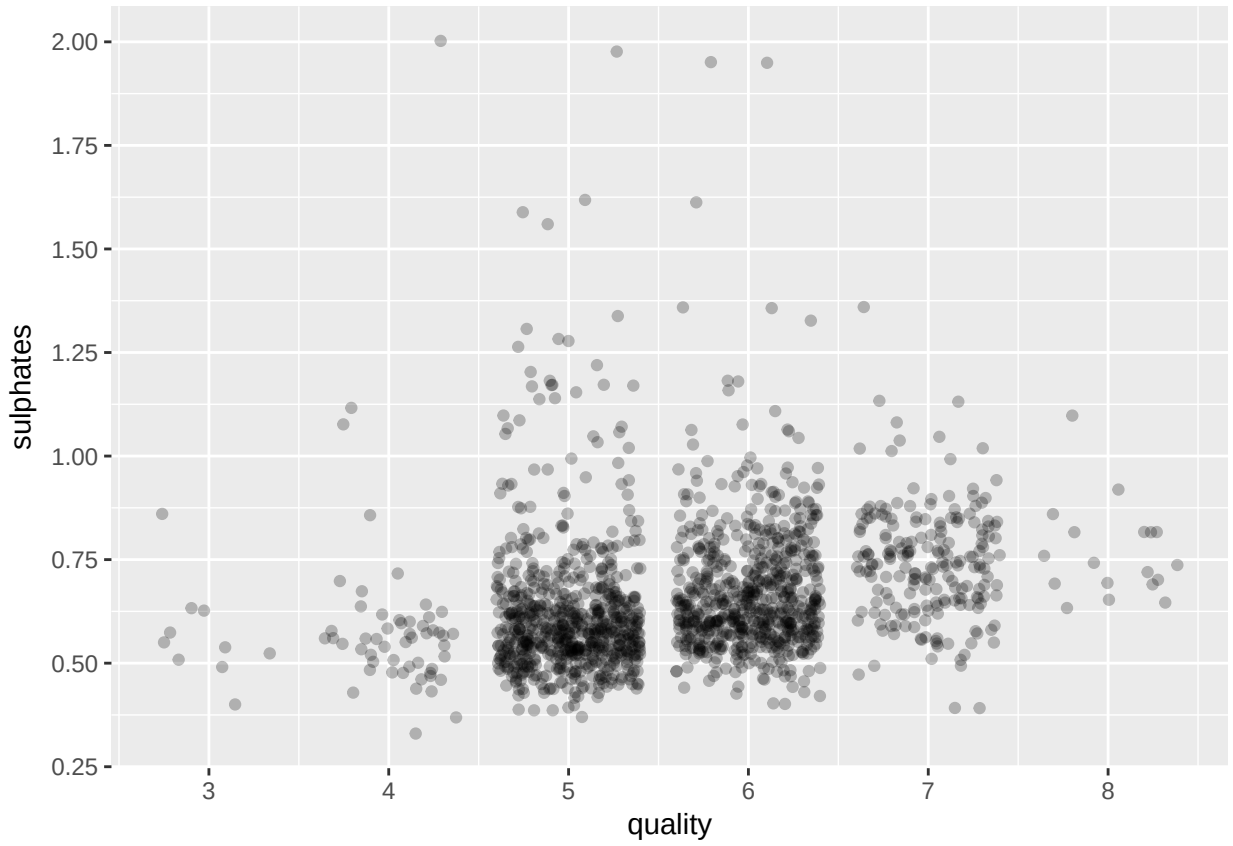
There doesn't seem to any type of significant relationship between fixed acidity levels and quality based on the correlation coefficient of 0.124. The fixed acidity levels for red wines rated an 8 seem to be more focused within a certain range, around 7-10.5 g/dmˆ3. This might not be very telling, since again, there aren't too many red wines with a rating of 8. I'm going to compare the boxplots for fixed acidity for wines of different quality ratings.

It seems that the wines of quality 7 had fixed acidity levels from around 7-10.5 g/dmˆ3. It might be a good idea to now examine the boxplots of lower quality wines next to these.Though 75% of the wines with rating 3 seemed to have acidity levels between 7.5-10 g/dmˆ3, it should be noted that the lower whisker for these wines is very short. The wines with rating 4 actually seemed to be more packed into a certain range: 7-8.5 g/dmˆ3.



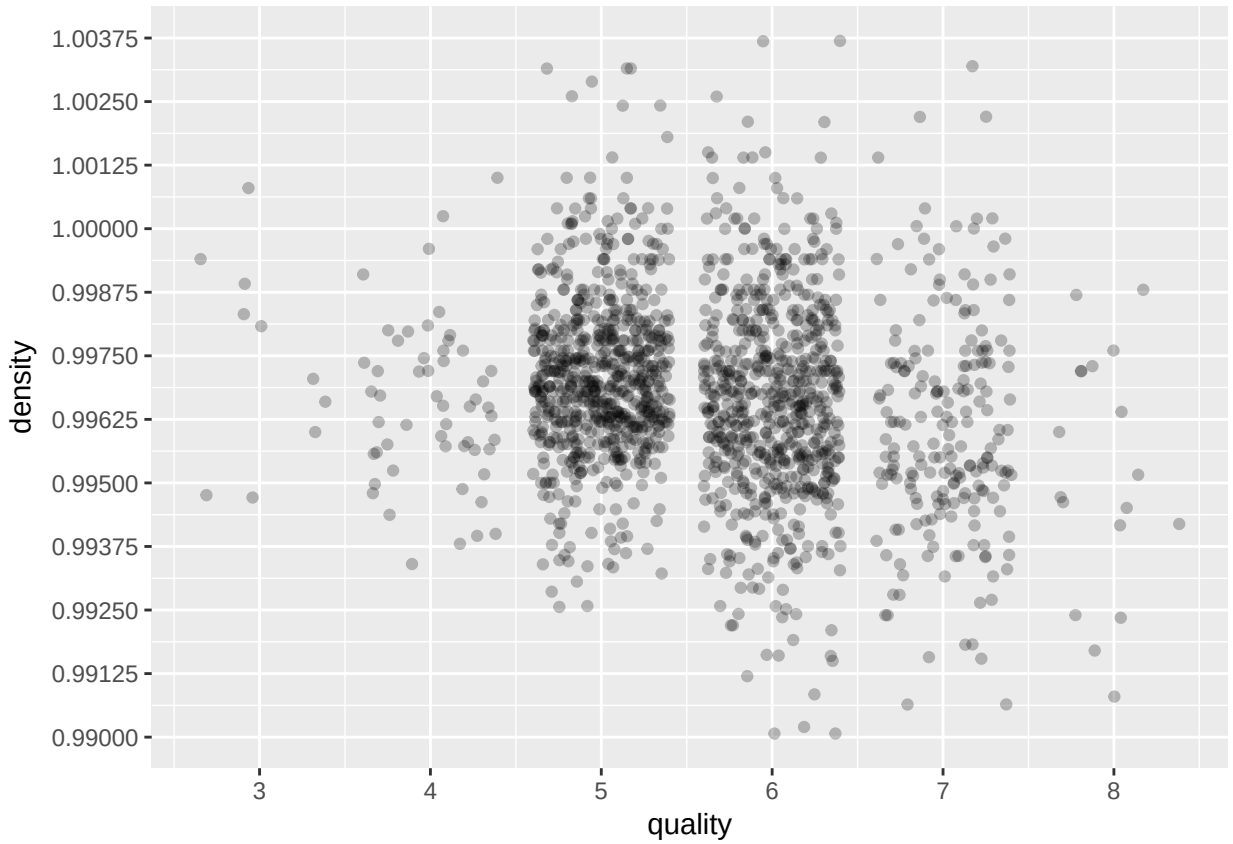

```
##
##  Pearson's product-moment correlation
##
## data:  rw$residual.sugar and rw$quality
## t = 0.5488, df = 1597, p-value = 0.5832
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.03531327  0.06271056
## sample estimates:
##        cor
## 0.01373164
```

Based on this scatterplot, it seems to me that wines with higher quality ratings did not have residual sugar levels above 9. This could be a sign that the fermentation process which results in the best quality red wines also keeps the residual sugar levels fairly and consistently low. However, it doesn't look like there are too many wines with residual sugar levels above 9, even for red wines with quality ratings lower than 7. Interestingly enough, the wines with rating 3 also had no residual sugar levels above 9. This may just be a result of not enough observations for wines of those quality ratings. The correlation coefficient does not suggest a significant linear relationship exists between residual sugar levels and quality ratings.



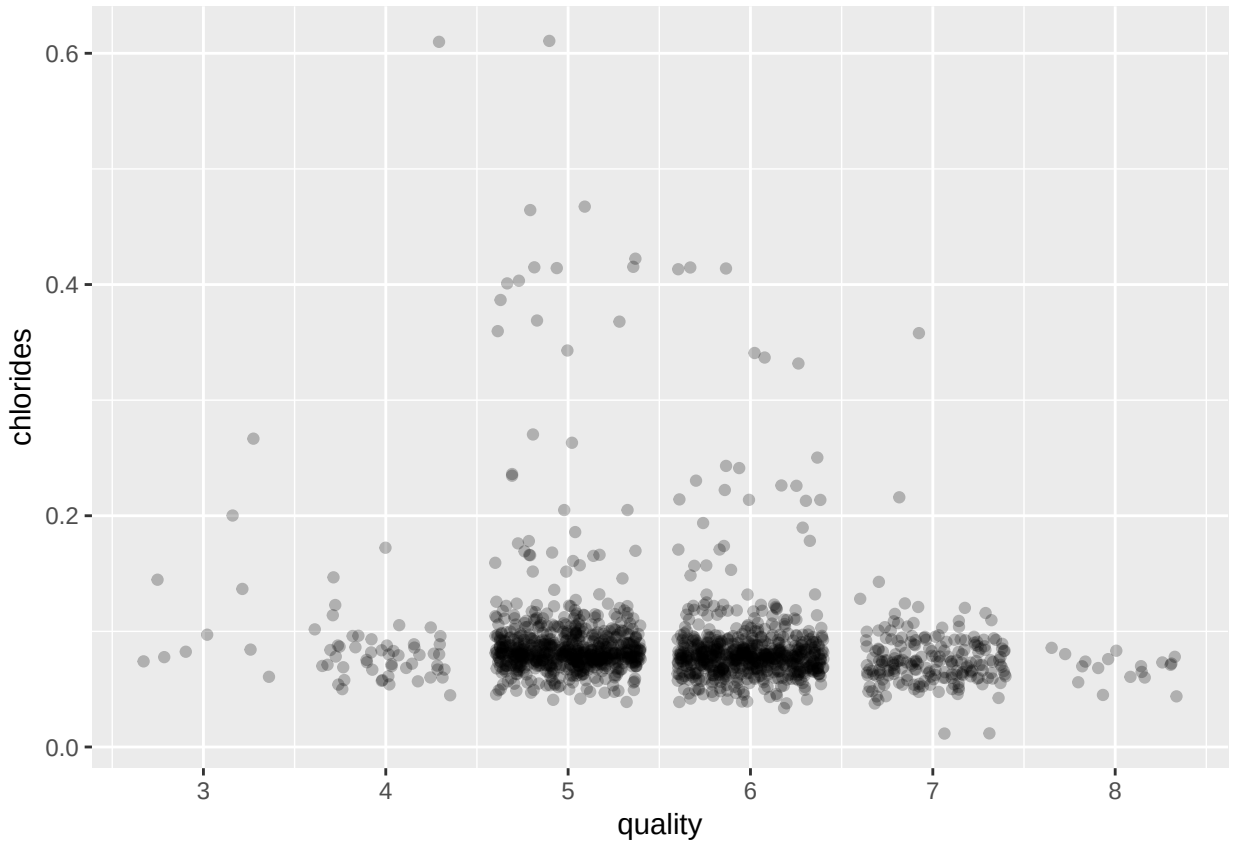

```
##
##  Pearson's product-moment correlation
##
## data:  rw$sulphates and rw$quality
## t = 10.38, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2049011 0.2967610
## sample estimates:
##       cor
## 0.2513971
```

Again, the wines of highest quality ratings, 7 and 8, seem to have more consistency with their sulphate levels. Though it doesn't seem like we have found any one property that will give us a good idea of the quality ratings, it might be true that higher quality wines tend to have more consistency. This makes sense, as you would expect the process of creating the highest quality wines to be very meticulous and careful: a craft of sorts. It seems reasonable to assume that the entire process for creating the higher quality wines leads to more consistency. Perhaps this is what I should be looking for in the next few plots. Being less careful with the process likely leads to lower quality wines and less consistency. It should be noted, however, that wines of fairly low quality with ratings 3 and 4 also seemed to have more consistency than wines with ratings 5 or 6. This, again, points to there actually just not being enough wines with ratings 3, 4, 7, or 8. The correlation coefficient is not significant.




```
##
##  Pearson's product-moment correlation
##
## data:  rw$density and rw$quality
## t = -7.0997, df = 1597, p-value = 1.875e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2220365 -0.1269870
## sample estimates:
##        cor
## -0.1749192
```

It seems the lower quality wines, wines of rating 3, 4, and 5 tended to stay above certain density levels. They never strayed below 0.99250 g/cm^3. Wines of ratings 6, 7, and 8, however, all had instances where the density strayed below 0.99250. Again, the correlation coefficient did not point to a significant relationship between the variables.




```
##
##  Pearson's product-moment correlation
##
## data:  rw$chlorides and rw$quality
## t = -5.1948, df = 1597, p-value = 2.313e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.17681041 -0.08039344
## sample estimates:
##        cor
## -0.1289066
```

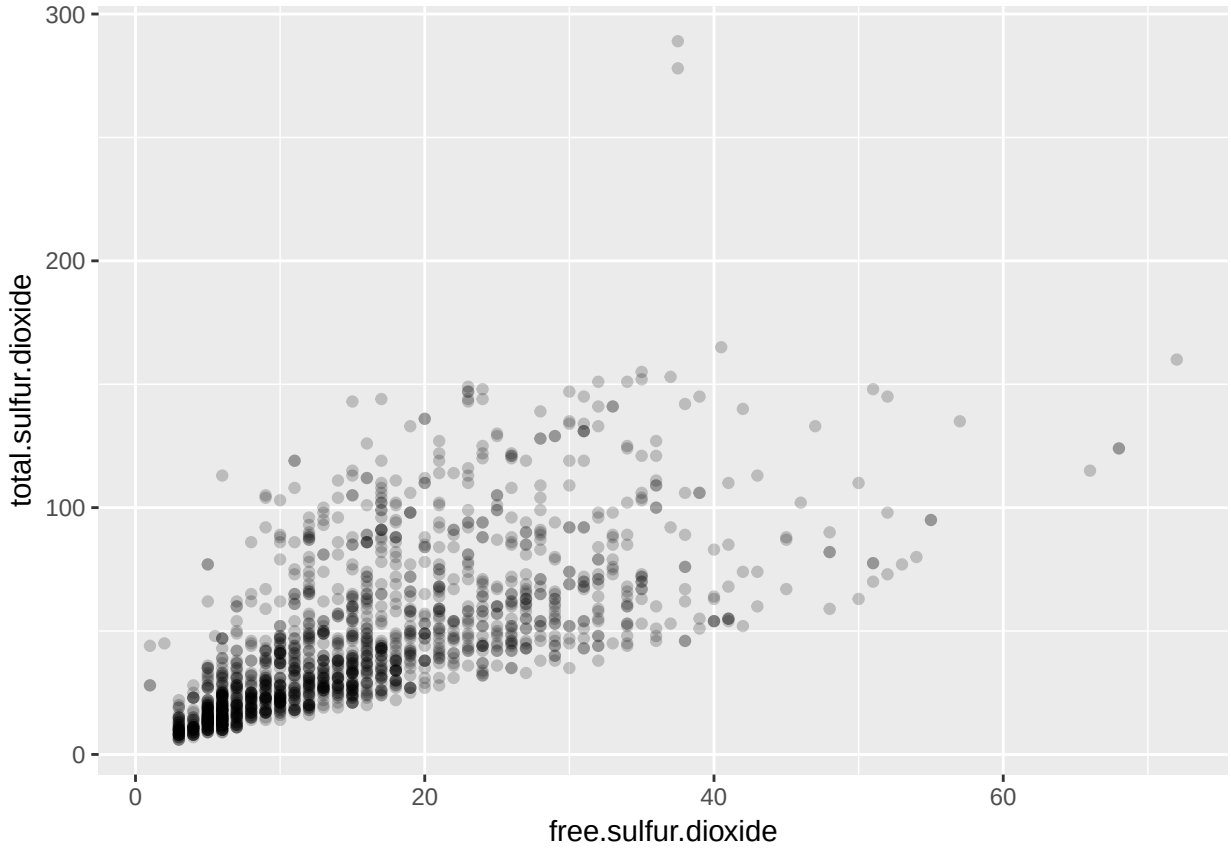
Here, the red wines of higher quality again seemed to be more consistent with their chloride levels (especially those with a rating of 8). Also, it does LOOK LIKE higher quality wines tended to have lower chloride levels.

However, the correlation coefficient does not indicate a significant relationship.

It shoud be noted, again, that our findings about consistency might be a result of there simply being more wines with ratings 5 and 6. We can do a quick check if this is the case with the table command.

```
##
##   3   4   5   6   7   8
##  10  53 681 638 199  18
```

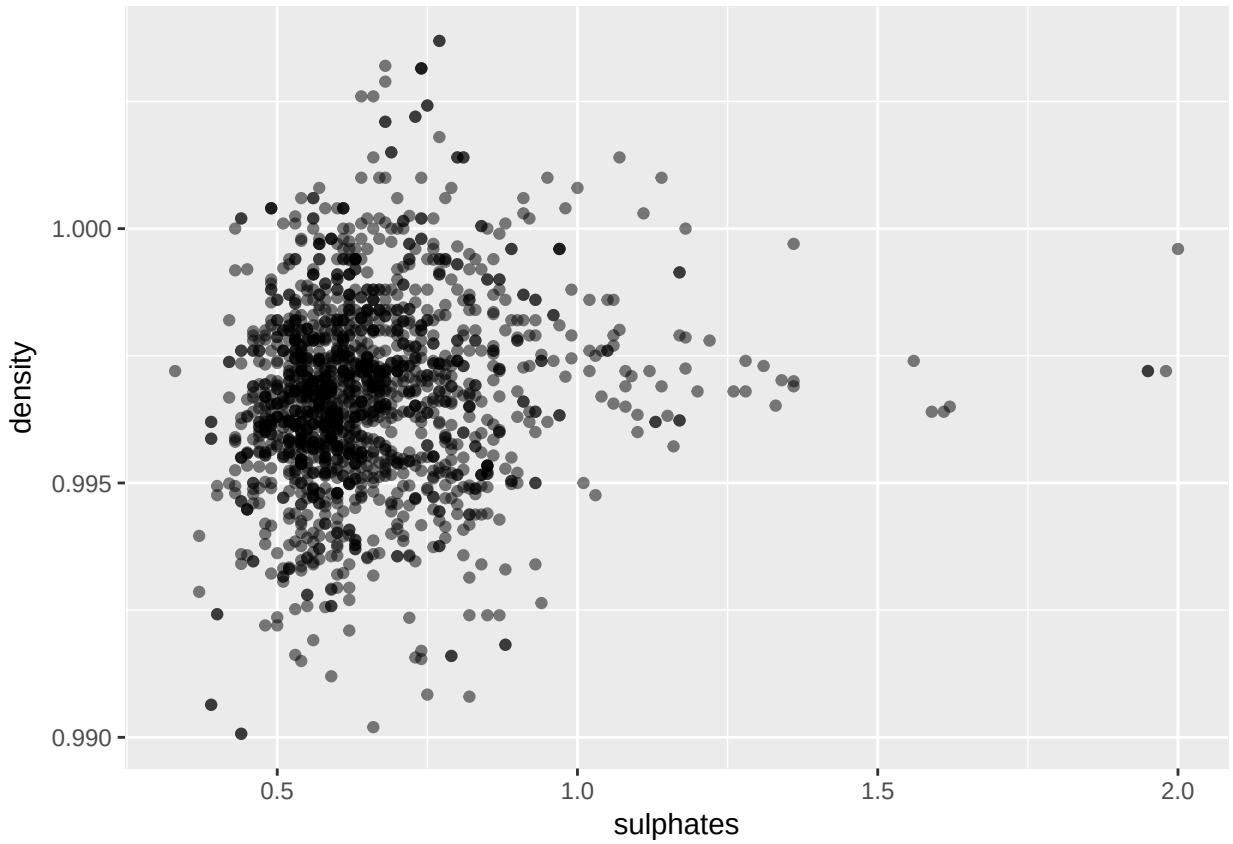
Indeed, it seems that just was a lot more wines with ratings 5 or 6 versus wines rated 3, 4, 7, or 8, which is why it appeared that there was more consistency among the more highly rated wines.




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

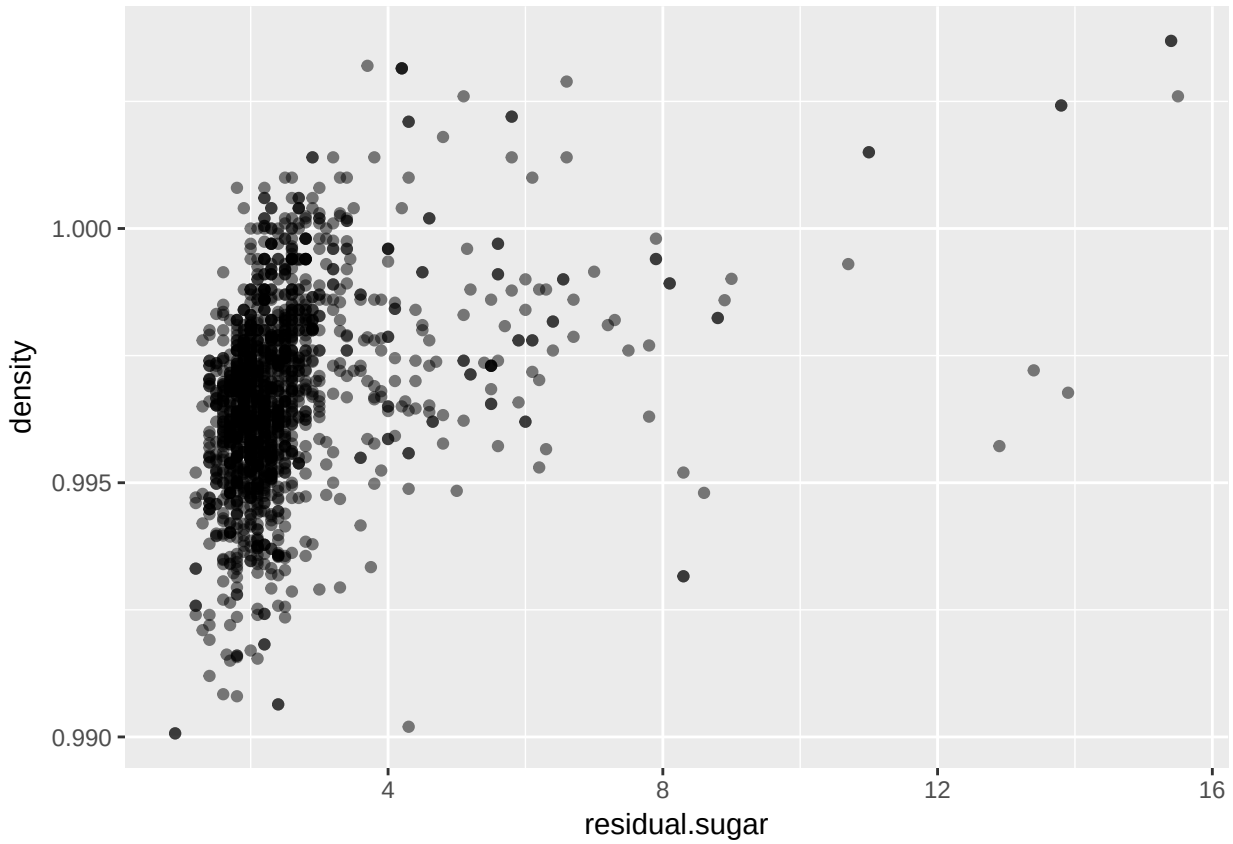
Not a very interesting relationship, but it does seem like the higher free sulfure dioxide levels is positively correlated with higher total sulfur dioxide levels. However, it does seem that a great proportion of the red wines have fairly low free and total sulfur dioxide levels, indicated by the dark blobs still present towards the bottom left of the plot.

It might be interesting to check for relationships between density and variables that might contribute to higher density, like presence of sulphates, residual sugars, chlorides, etc.



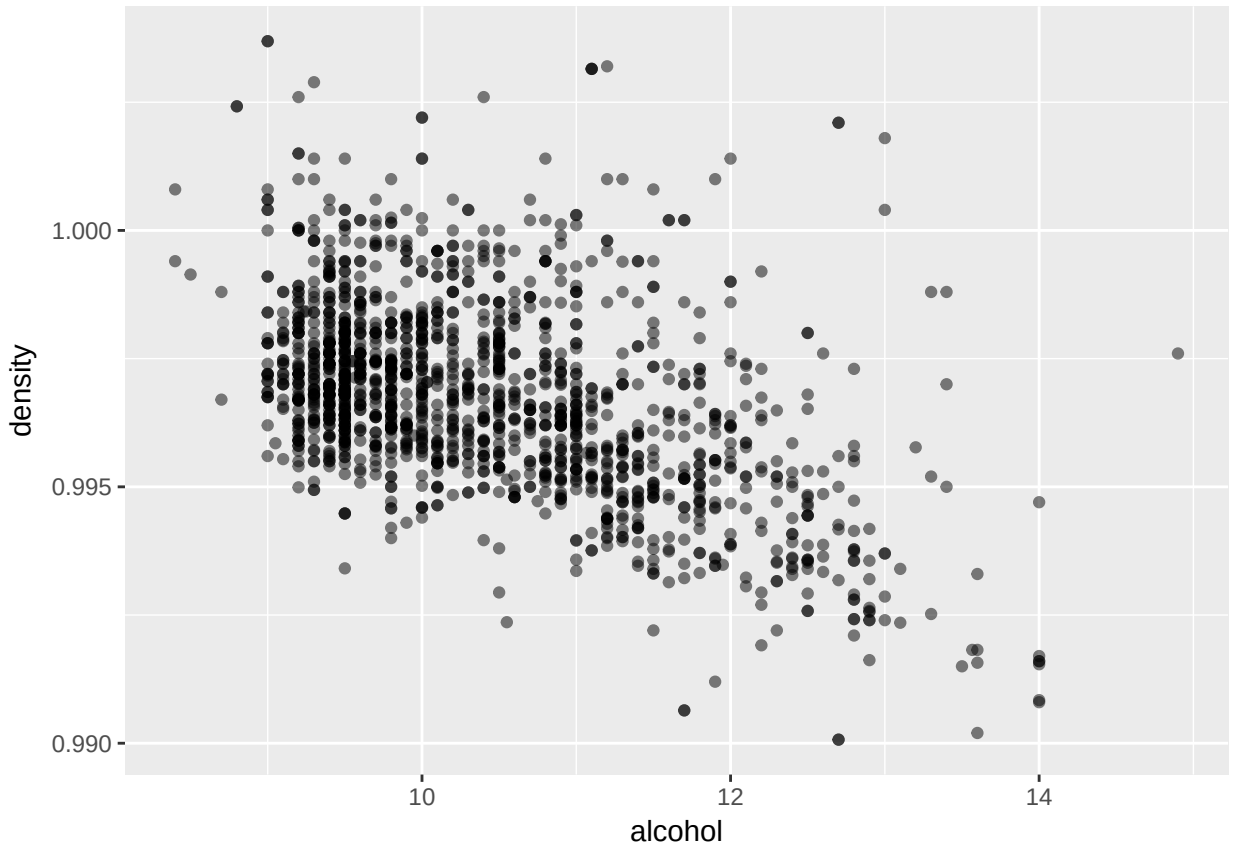

```
##
##  Pearson's product-moment correlation
##
## data:  rw$sulphates and rw$density
## t = 6.0012, df = 1597, p-value = 2.418e-09
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1002148 0.1961000
## sample estimates:
##       cor
## 0.1485064
```

It does not seem like the presence of higher sulphate levels contributed to higher density.




```
##
##  Pearson's product-moment correlation
##
## data:  rw$residual.sugar and rw$density
## t = 15.189, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3116908 0.3973835
## sample estimates:
##       cor
## 0.3552834
```

Based only on the plot, it didn't seem to me like the presence of higher residual sugar levels had much of an effect on density. Actually, it turns out there is a moderate positive relationship between residual sugar levels and density, based on the returned correlation coefficient: 0.355.




```
##
##  Pearson's product-moment correlation
##
## data:  rw$alcohol and rw$density
## t = -22.838, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5322547 -0.4583061
## sample estimates:
##        cor
## -0.4961798
```

Interestingly, there does seem to be a relationship between alcohol percentage levels and density. It seems that as alcohol percentage levels increase, density decreases: these variables are negatively correlated. It might be a good idea to ask why this makes sense in the analysis.

Bivariate Analysis

Relationships observed with main feature of interest:

My main feature of interest was quality. It didn't really seem to me that any one chemical property other than alcohol could be said to be associated with higher/lower quality ratings. However, something I did notice was that wines with higher ratings tended to have more consistency with their measured properties than wines ofquality 5 or 6. To reiterate, I thought this would make sense because you would expect the winemaking process for the highest quality wines to be very precise and painstaking. This would leave little room for inconsistency in the process, and consequently, little room for much inconsistency with the chemical

properties. Unfortunately, this is probably just because there were not as many wines with ratings 7 or 8 versus wines rated 5 or 6.

Relationships observed between other features:

I observed a positive relationship between free sulfur dioxide and total sulfur dioxide levels.

Though I didn't immediately notice from the plot, there was a moderate positive relationship between residual sugar level and density. This makes sense; I expect higher amounts of leftover sugar (denser than water) to contribute to higher density levels.

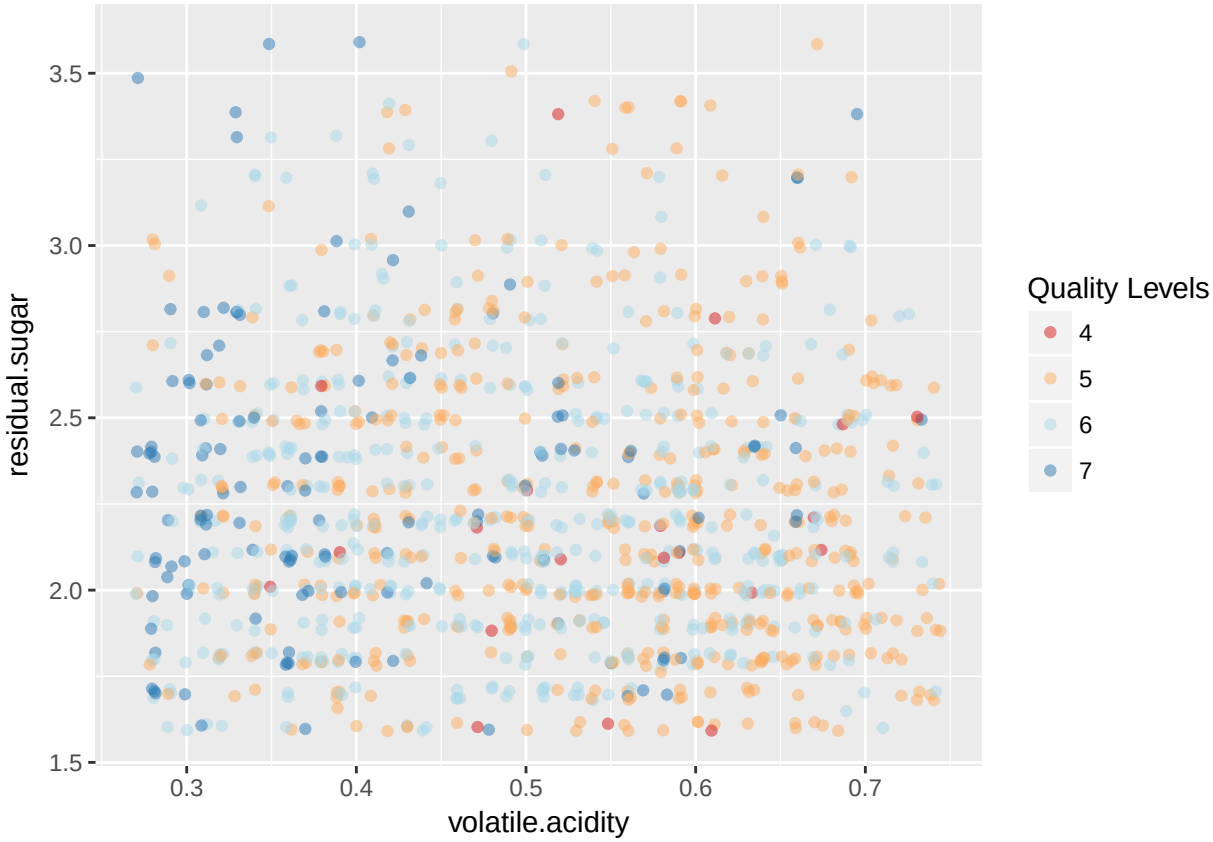
Also, there was a negative correlation between density and alcohol levels. After a quick google search, I found that alcohol is less dense than water. This may serve to explain why wines with higher alcohol percentage levels tended to have lower density.

Strongest relationship:

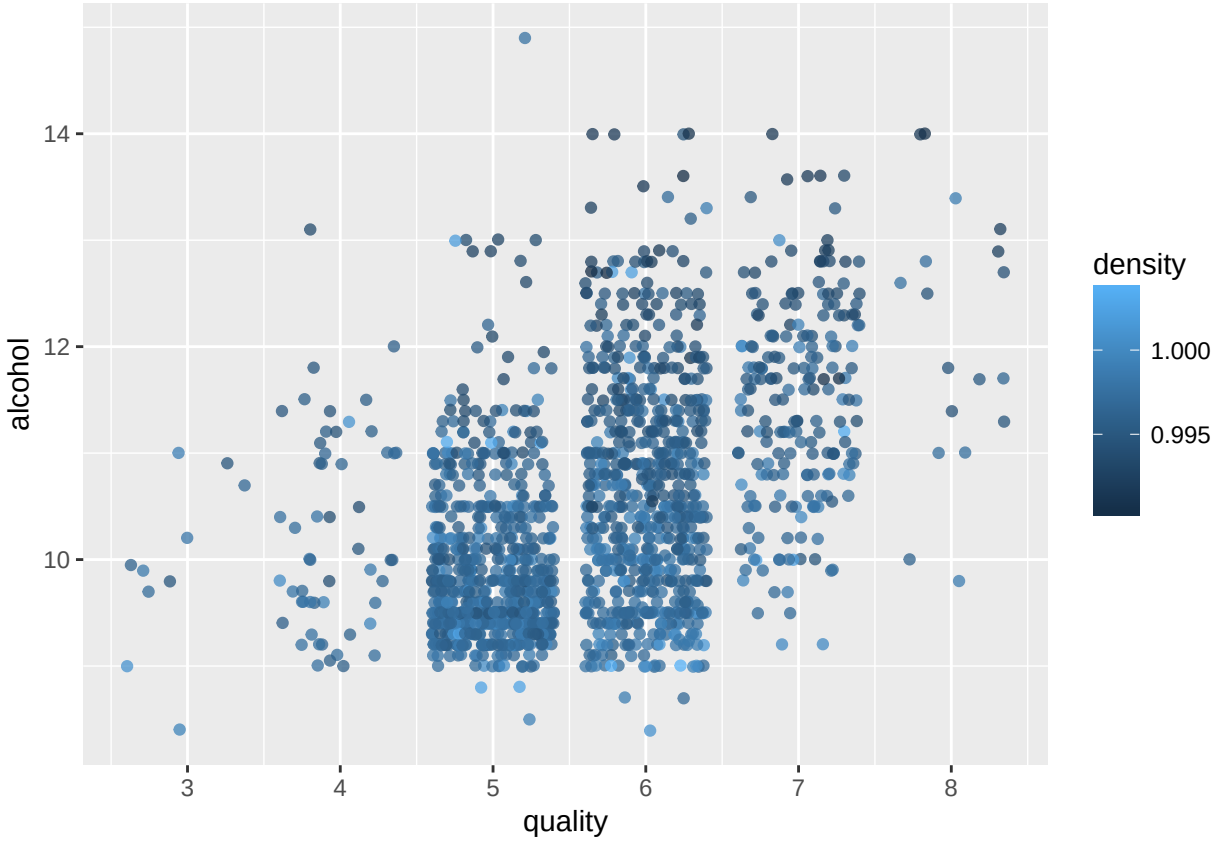
The strongest relationship I found was between free sulfur dioxide and total sulfur dioxide. The correlation coefficient returned was 0.667.

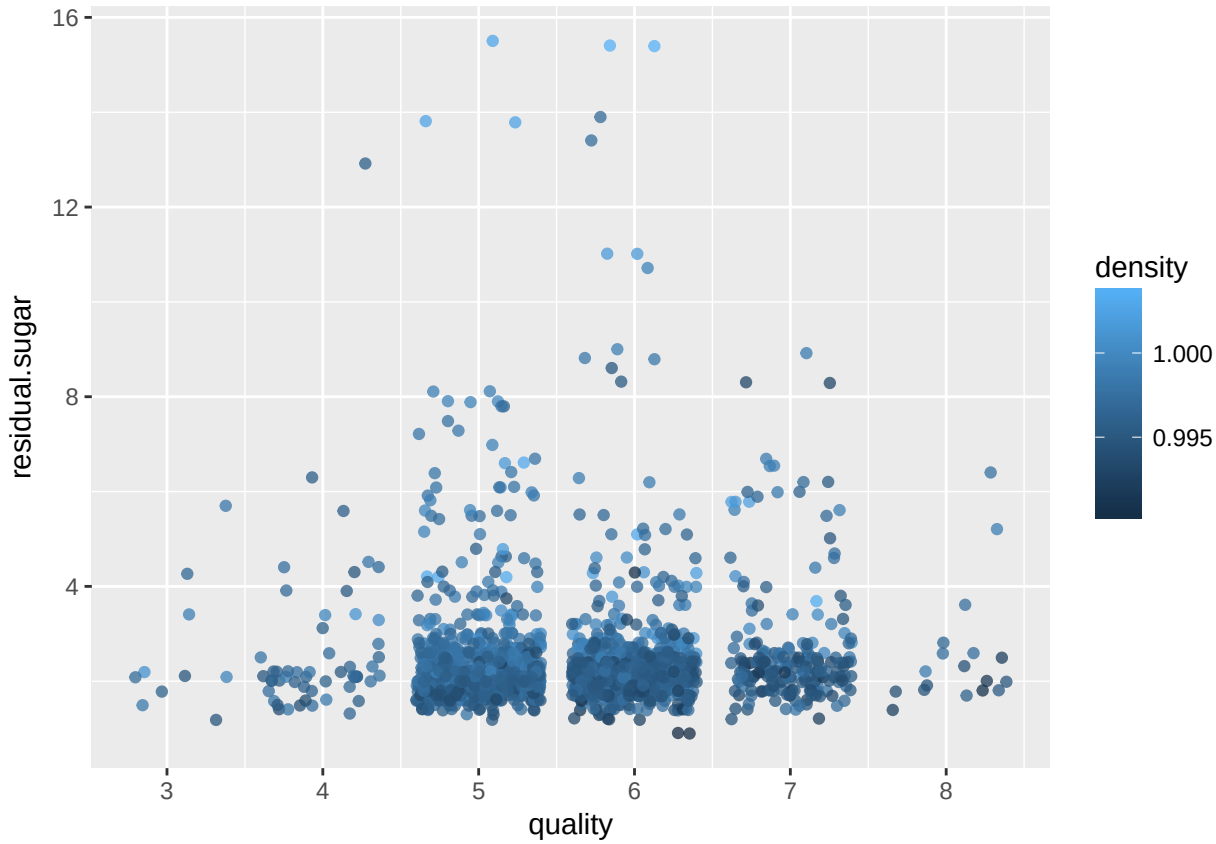
Multivariate Plots Section

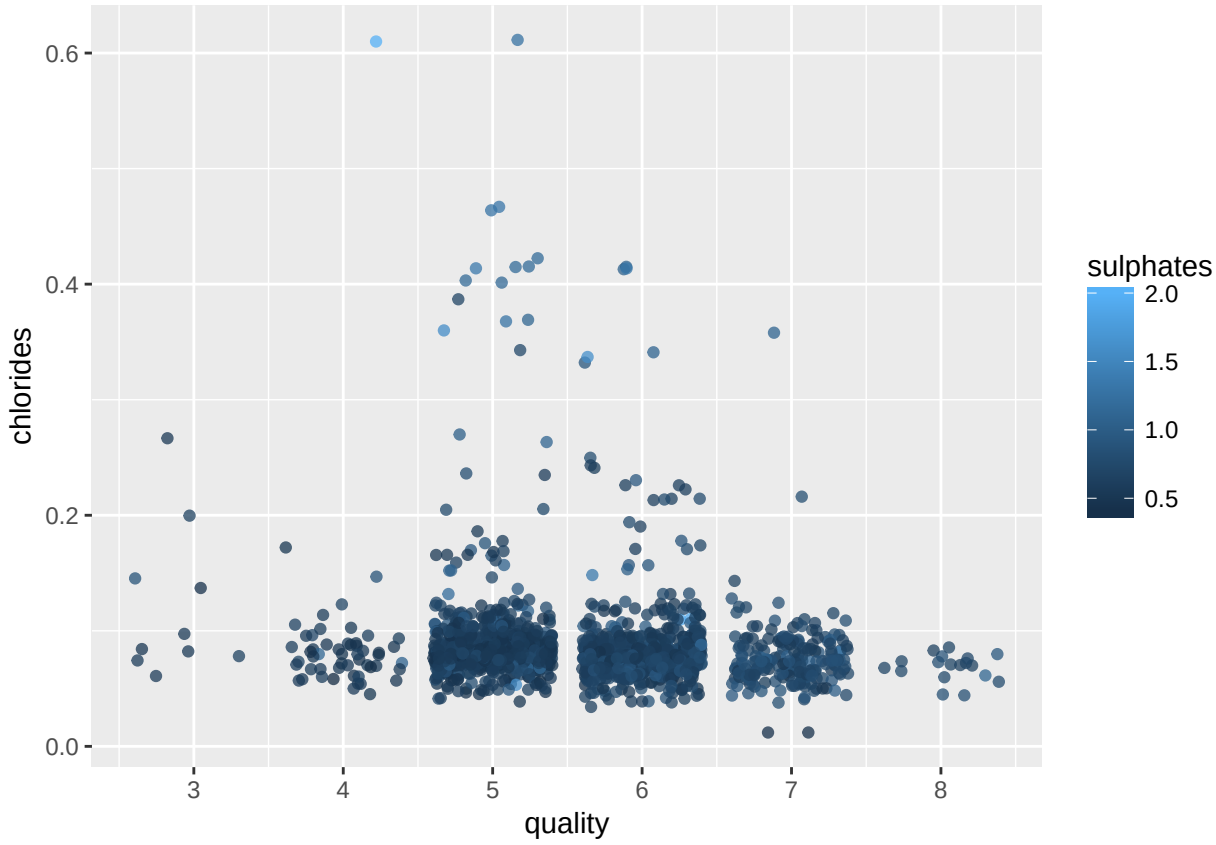
```
## Warning: Removed 439 rows containing missing values (geom_point).
```



It looks like red wines of quality ratings 5 and 6 were more spread out in terms of volatile acidity and residual sugar levels, while red wines of quality rating 7 seemed to be focused around the lower volatile acidity levels. Again, it seems these are the results of there being more red wines rated 5 or 6 than 7.

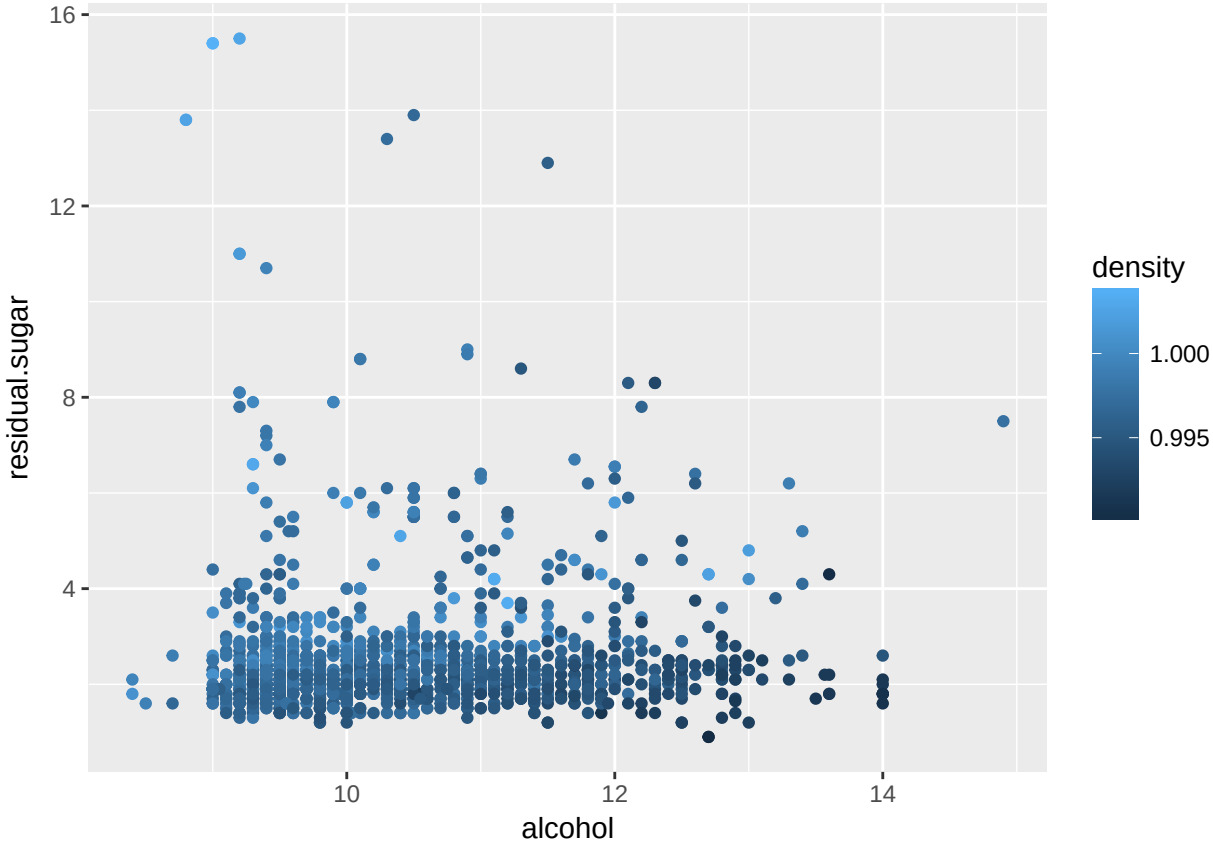






I didn't really find the above three plots very educational. I don't notice any clear trends or interesting

relationships which weren't already discussed in my bivariate plot discussion.



Based on this scatterplot, it seems that higher alcohol levels along with low residual sugar levels were associated with lower density levels. Again, this makes sense based on the bivariate plots of alcohol vs. density and residual sugar levels vs. density.

Multivariate Analysis

Relationships observed:

There seemed to be more consistently low volatile acidity levels for red wines of quality rating 7 compared to red wines rated 5 or 6. Again, this is likely a result of there simply being many more 5's and 6's versus 7's.

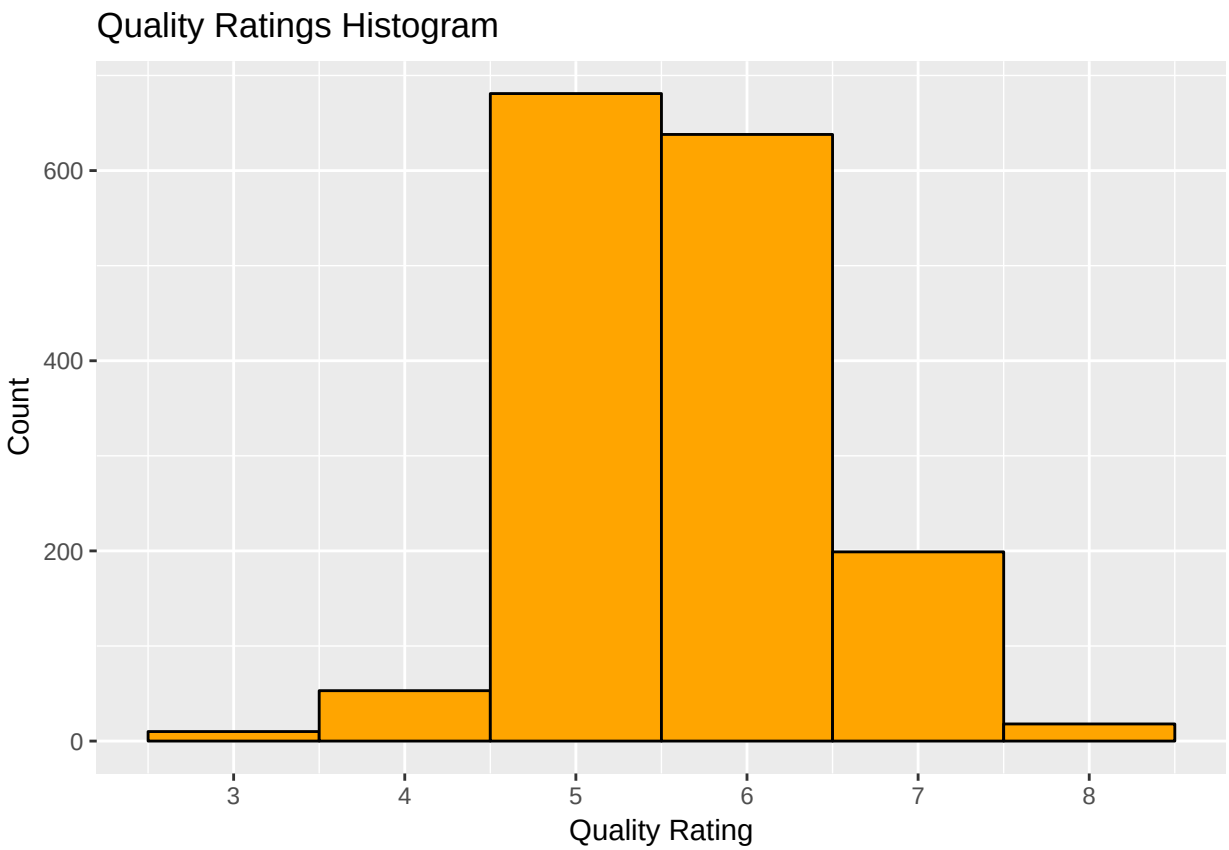
The relationship between residual sugar levels, alcohol, and density were made even more clear with the multivariate plot of the three variables. Lower alcohol levels tended to be associated with higher density. Higher alcohol levels tended to be associated with lower density. The opposite held for residual sugar levels.

Interesting/Surprising Interactions:

The most interesting interaction between features was between residual sugar, alcohol, and density. We did already know that there was a positive relationship between residual sugar levels and density, and a negative relationship between alcohol and density. We examined the multivariate scatter plot of alcohol vs. residual sugar levels, with the colors of the points varying with density. The plot reinforced our findings from our bivariate plots; the red wines with the higher alcohol levels and lower residual sugar levels tended to have lower densities, and vice versa.

Final Plots and Summary

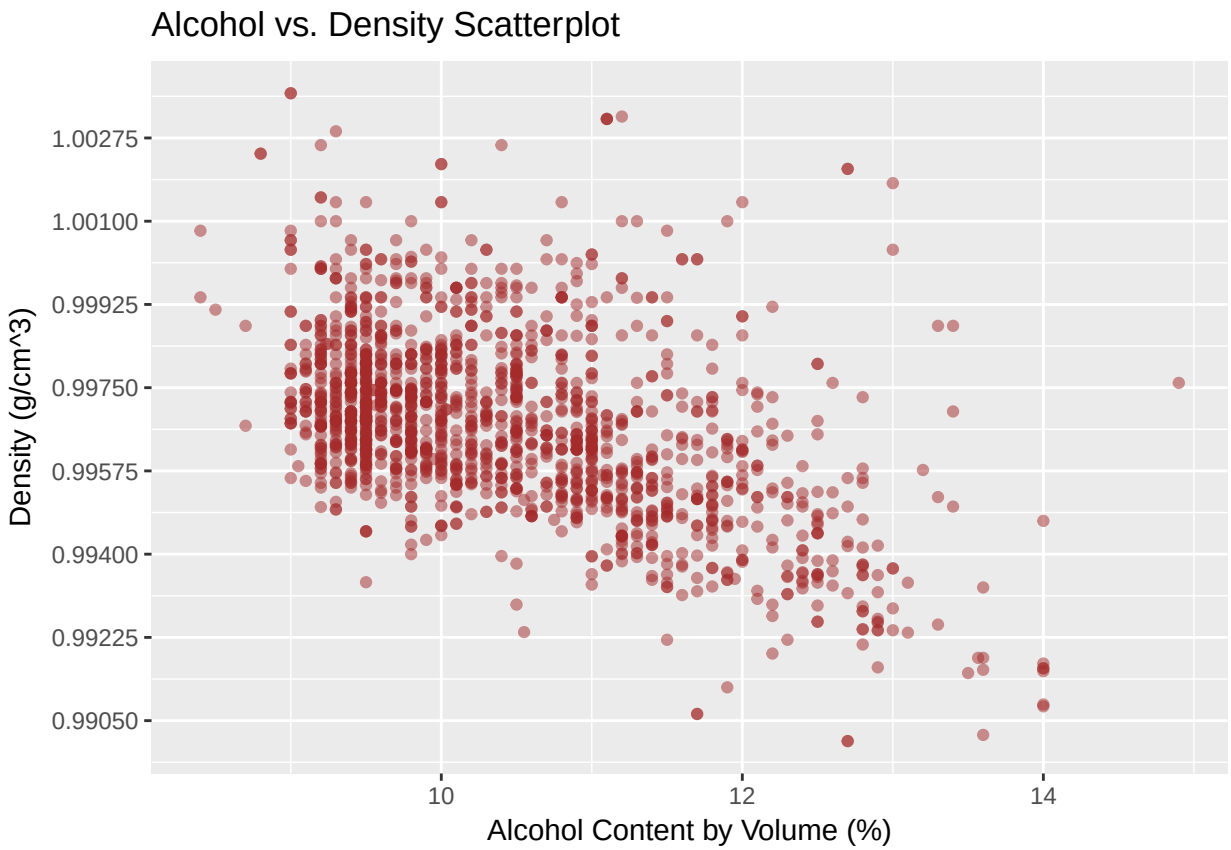
Plot One



Description One

I believed one of the most important distributions to examine was that of the quality ratings. I found it fairly significant that the distribution was normal, as this signaled the wine experts were not being overly critical or easygoing when rating the wines.

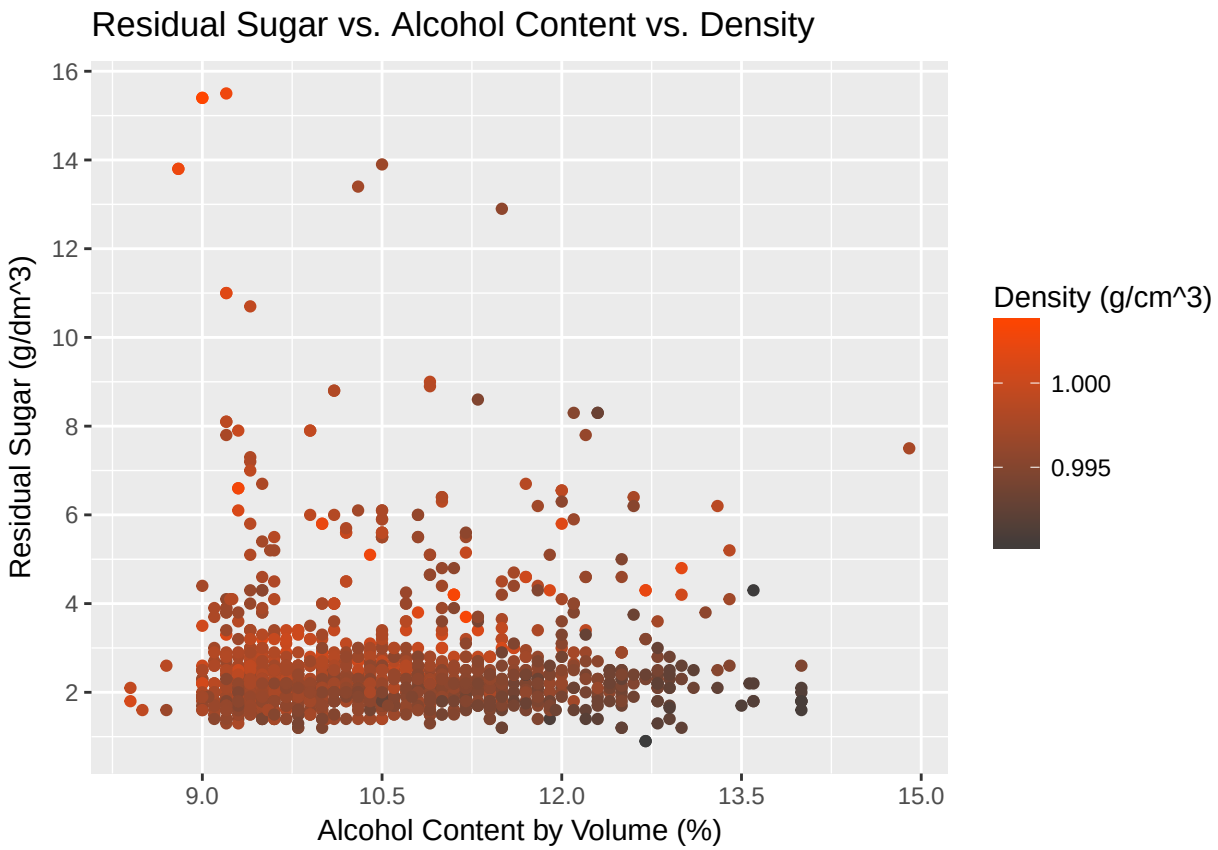
Plot Two



Description Two

I chose this plot because it illustrated a relationship between chemical properties which could be easily explained and understood. As alcohol percentage level increased, density tended to decrease. Again, this is because alcohol is not very dense (less dense than water). Consequently, the higher the alcohol content, the less dense the red wine, and vice versa.

Plot Three



Description Three

I chose this plot because it brought together my findings from two of my bivariate plots. As stated above, I found a negative correlation between alcohol content and density. In the bivariate plots section, I found a positive relationship between residual sugar levels and density. It then made sense to create a multivariate plot of alochol content vs. residual sugar levels with density as color, as it would demonstrate both relationships at once in one plot. Checking the plot, you can see that red wines with high alcohol content and low residual sugar levels tended to have lower densities, and vice versa. Both relationships are demonstrated in one nice plot.

Reflection

Overall, the code required for this project was not too difficult. I would say I had a lot of difficulties in a few areas. First, I had a hard time "exploring" different plots; if I thought that comparing two variables didn't really make sense then I would have difficulty justifying creating the plots. This made it hard to find interactions/relationships that I didn't already know about. Second, I didn't really feel like I found many meaningful relationships for my most important variable: quality. Of course, that could just mean that it is a combination of all the properties that makes a wine "good". However, it was still discouraging to find little to no very interesting plots which measured quality as one of the variables, other than alcohol vs. quality. I also had a lot of trouble figuring out exactly which variables to plot for the multivariate plots, and why. Lastly, I am still having some trouble actually analyzing the meaning of the plots after I generate them. This might

just have to do with the relationships I chose to examine being insignificant. Something that was surprising to me was how easy it was to clean up the plots with labels and colors/outlines. In the future, it might be beneficial to create a model to predict if a red wine will be of high quality based on given chemical properties.