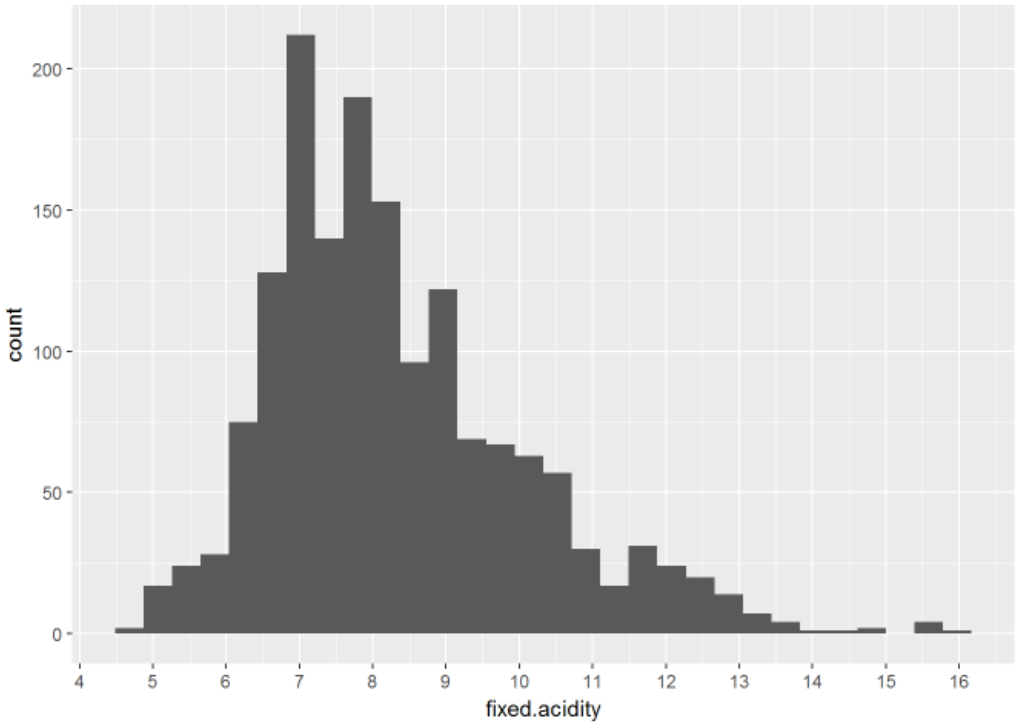
Data Analysis 6 by Eric Braun

This dataset looks at eleven chemical properties of 1,599 bottles of red wine. These wines were rated by experts on a scale of 1 to 10. The basic question I will be asking is- what makes a good wine?

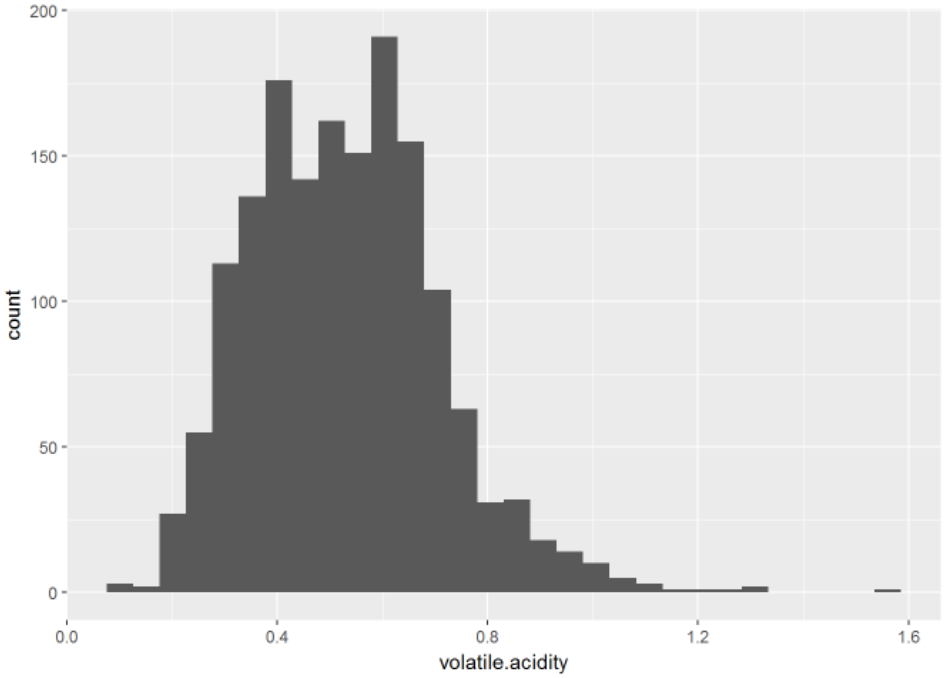
Univariate Plots Section

Before analyzing the relationships between the variables, it is important to understand the individual variables and how they are distributed.



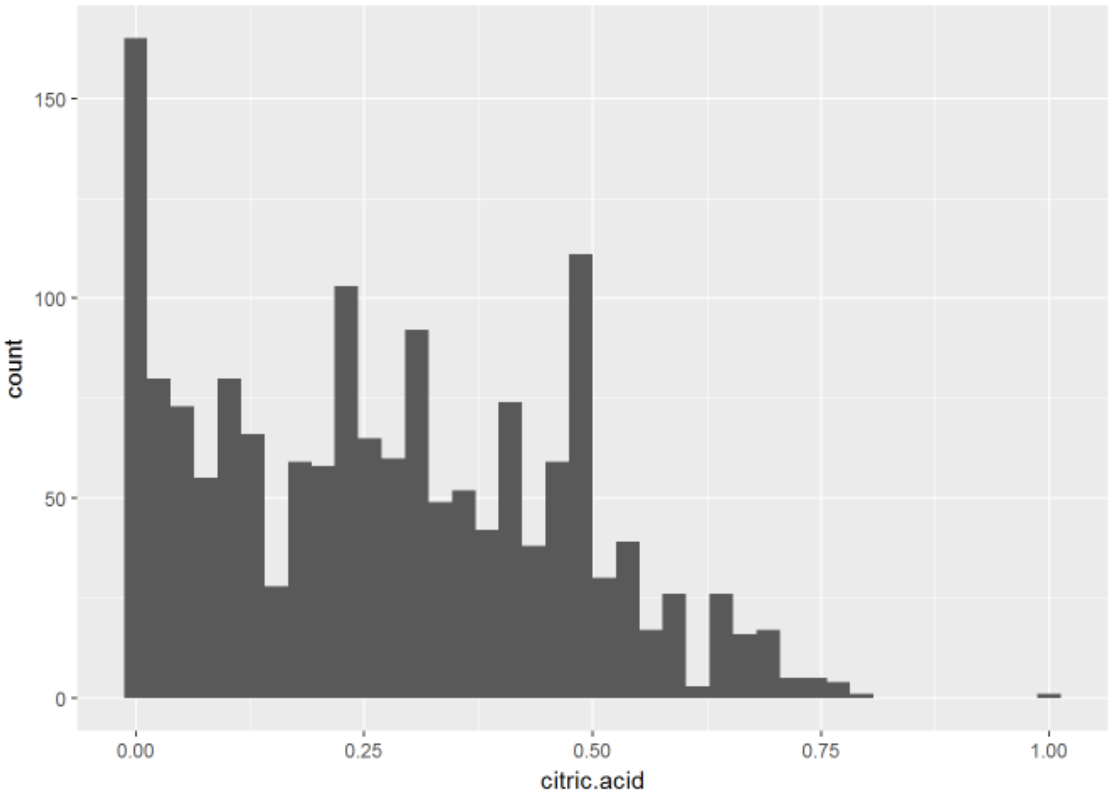

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
```

Fixed acidity traditionally refers to flavorful acids found in wine. A lack of these acids results in a “flat” wine. We can see that most wines have a fixed acidity of around 7 to 8. There is a longer tail of higher content. This makes sense. We would expect more flavor rather than less. Note that citric acid is usually included in fixed acidity counts but this set measures citric acid separately. The minimum value is 4.6, the max is 15.9. The Median is 8.32, the mean is 8.32 and the quartiles are 7.1 and 9.2.



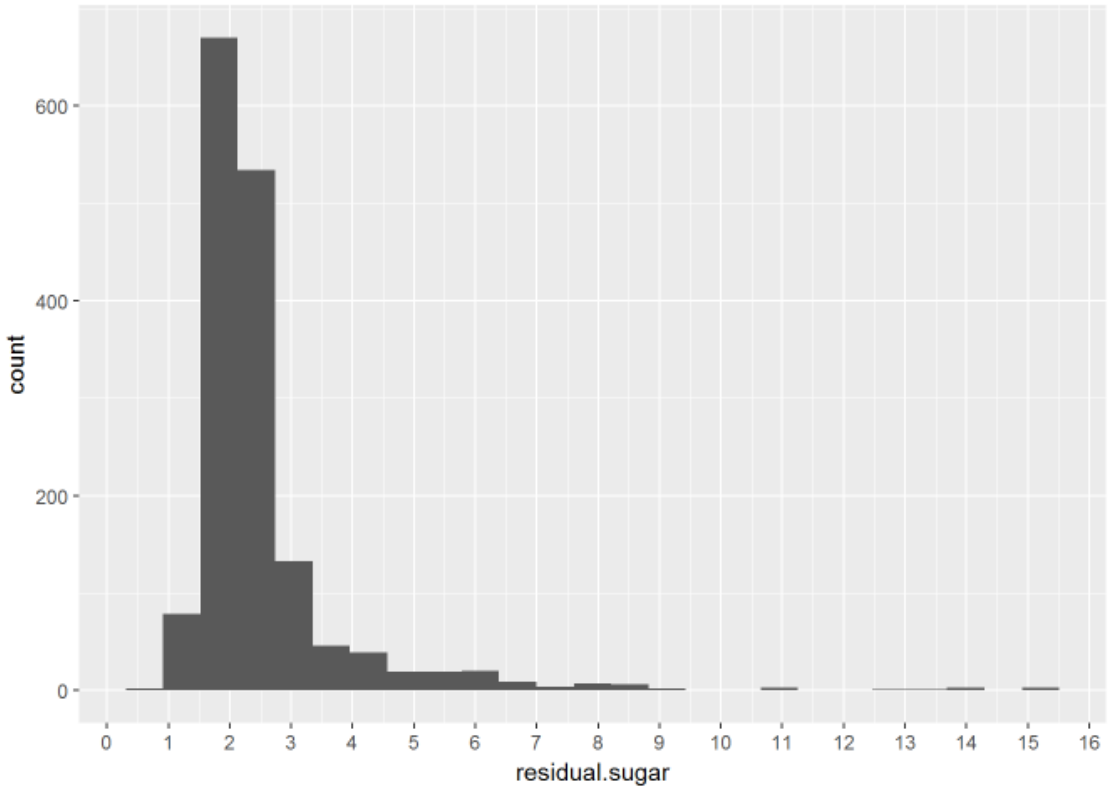

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```

Volatile acidity refers to the acids that are unwelcome in the wine. Primarily this refers to acetic acid (vinegar) but can also include lactic, formic, or butyric acids. We can see in the data that the content peaks at about .6 but in a broader sense, the most common is between .4 and .8. The minimum value is .12 and the max is 1.58. The median is .52 and the mean is .5278. The quartiles are .39 and .64.



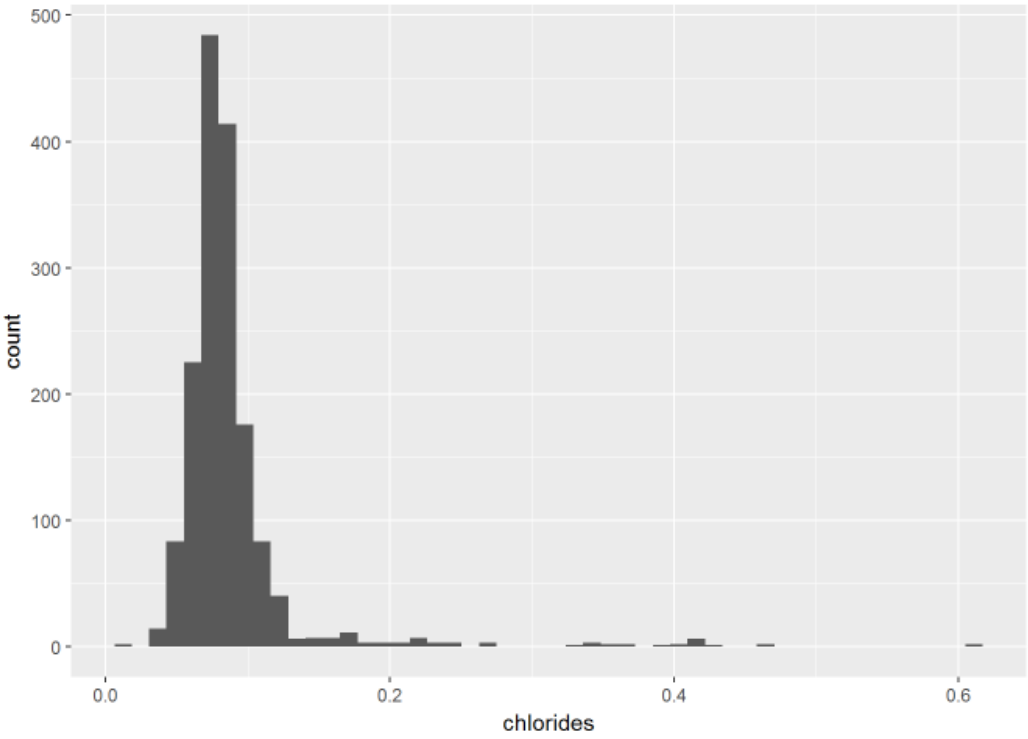

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```

Citric acid is sometimes considered a fixed acid. It is present in trace amounts in grapes, but is often added for flavor. The EU prohibits this usage, but it can be used there to remove minerals from the wine. We don't see a simple bell curve. The basic shape is decreasing quantities with a number of peaks.
The minimum is 0, which is odd and suggestive of incomplete data, the maximum is 1. The median is .26 and the mean is .271. The quartiles are .09 and .42.



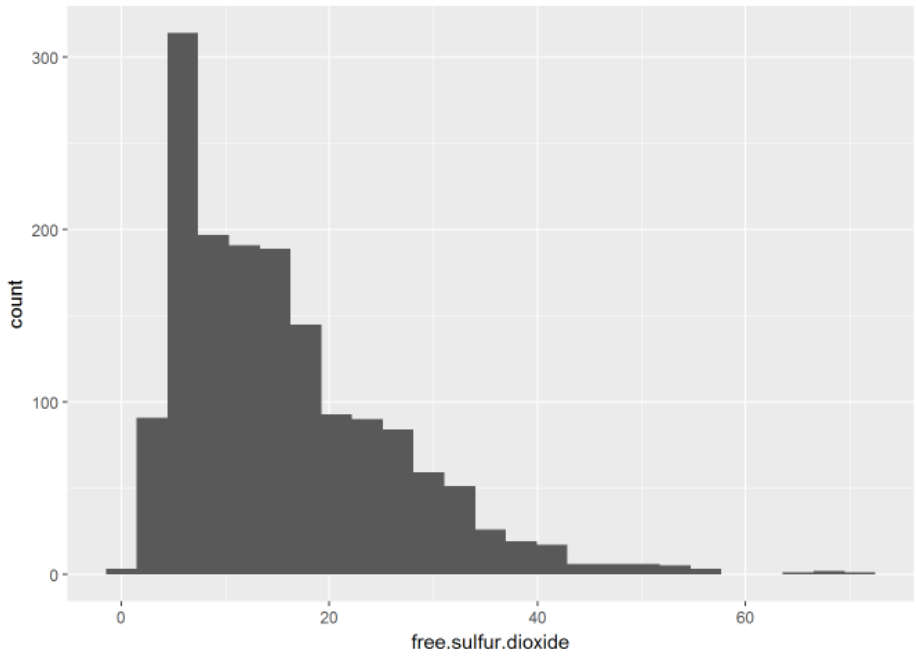

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```

The next variable is residual sugar. Fermentation is the process of converting sugar to alcohol, but not all the sugar is converted. The left over sugar determines the dryness of the wine. Dry reds have 2-3 grams per liter while sweeter reds can have as much as 30. We can see that most of the wine in this set is dry, with the majority of the sample at 2-3.
The minimum value is .9 and the max is 15.5. The median is 2.2 and the mean is 2.539. The quartiles are 1.9 and 2.6.

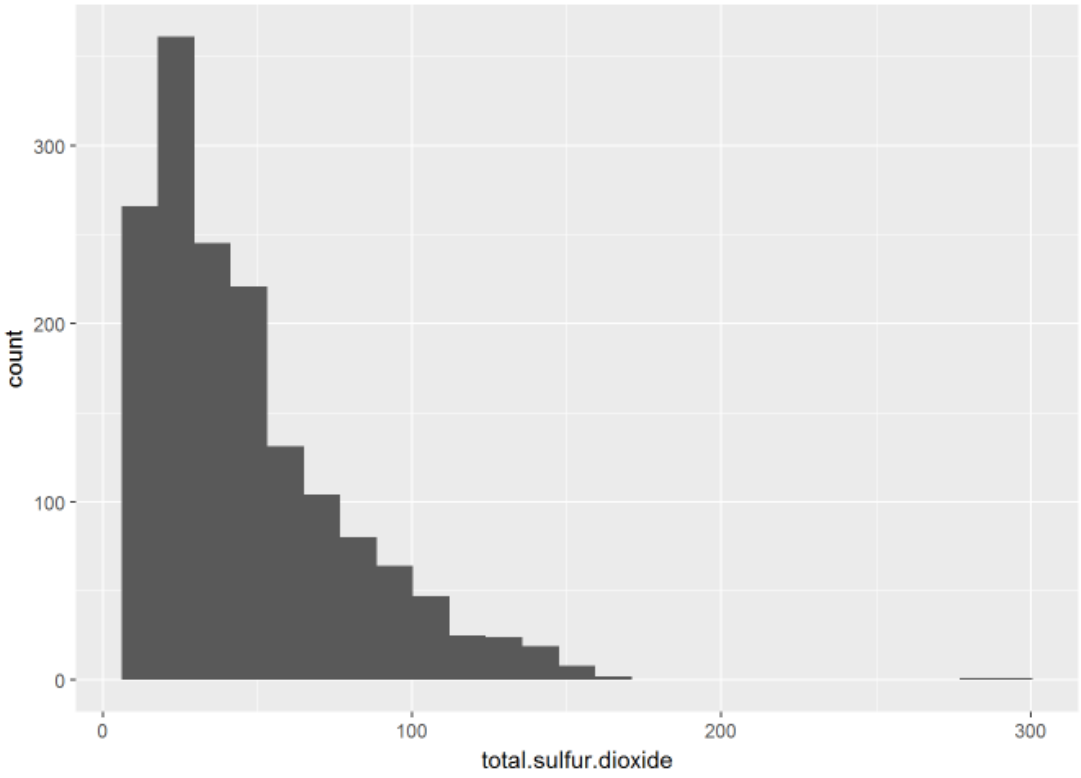



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```

Chlorides are naturally occurring salts such as sodium chloride and potassium chloride. They vary based on region and in large enough quantities are associated with salty, soapy or bitter flavors.Needless to say, these flavors are unwanted There is a clear peak at .1 with a long tail that reaches as far as .6. This is what we'd expect. The minimum value is .012 and the max is .611. The median is .079 and the mean is .08747. The quartiles are .07 and .09)



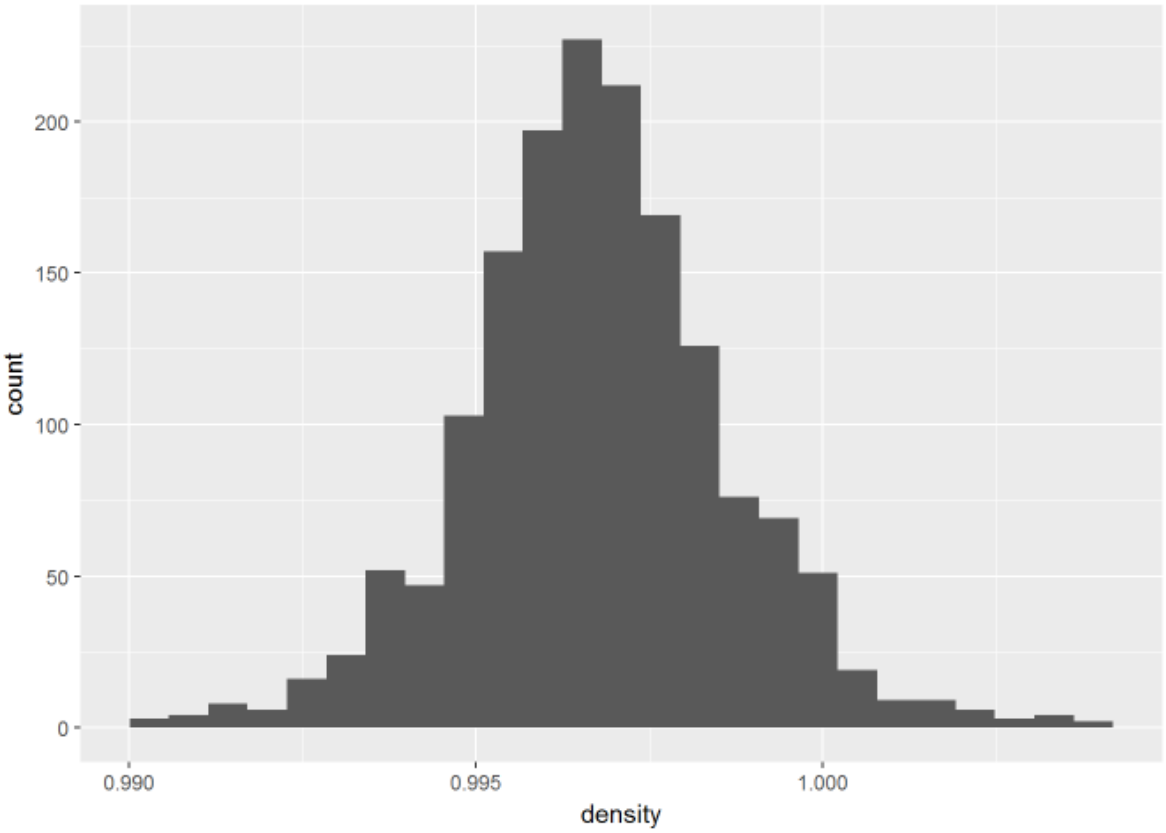

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
```




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

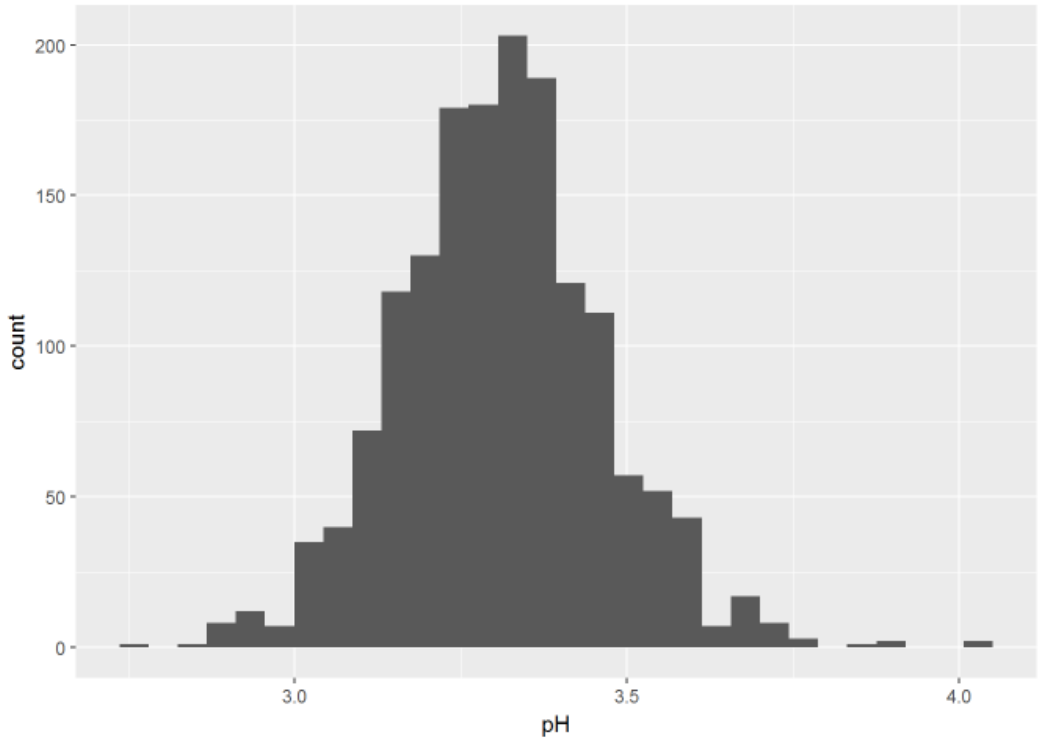
```
##    6.00   22.00   38.00   46.47   62.00  289.00
```

Sulfur dioxide is a naturally occurring chemical in the fermentation process. It is often added to wine for its antioxidant and antimicrobial properties. This is a somewhat controversial process as many drinkers don't like the idea of drinking added chemicals. Free refers to sulfur dioxide not bound to sugar compounds. Both free and total counts are more common in low amounts, steadily decreasing as the proportions grow higher. The minimum value for free sulfur dioxide is 1, the max is 72. The median is 14 and the mean is 15.87. The quartiles are 7 and 72. The minimum value for total sulfur dioxide is 6 and the max is 289. The median is 22 and the mean is 46.47. The quartiles are 22 and 62.



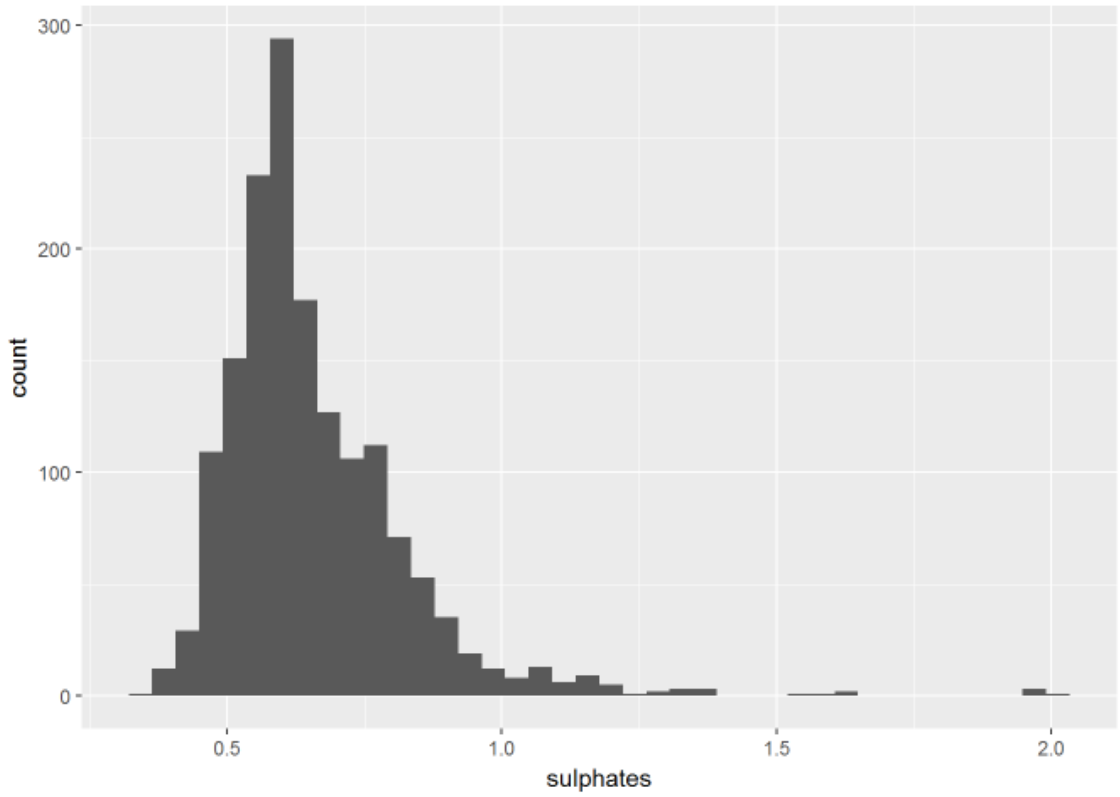

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0037
```

Density refers to the mass of the wine. The basic unit is grams per mL. We can see that there is a very low range of masses with a clear bell curve distribution, peaking at a little under one. The minimum value is .9901 and the max is 1.0037. The median is .9968 and the mean is .9967. The quartiles are .9956 and .9978.



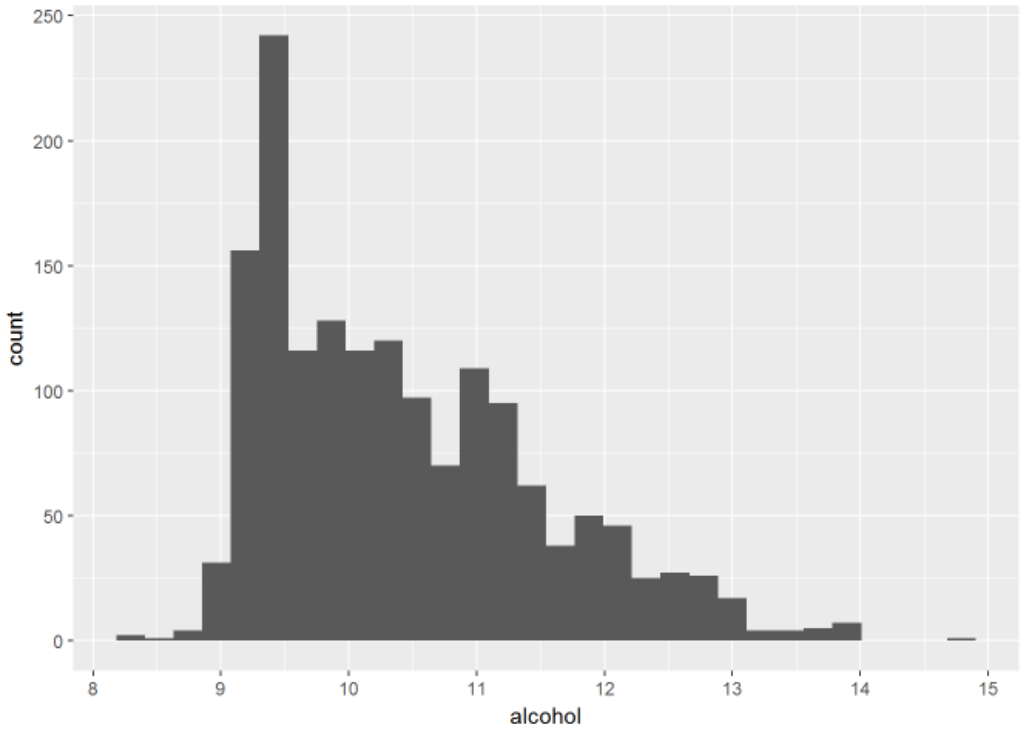

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010
```

pH is the measure of acidity. A 3 rating is about as acidic as vinegar. A 4 is less acidic. The data follows a bell curve with the bulk of the data at about 3.4. The minimum value is 2.74 and the max is 4.01, the median is 3.31 and the mean is 3.311. The quartiles are 3.21 and 3.4.



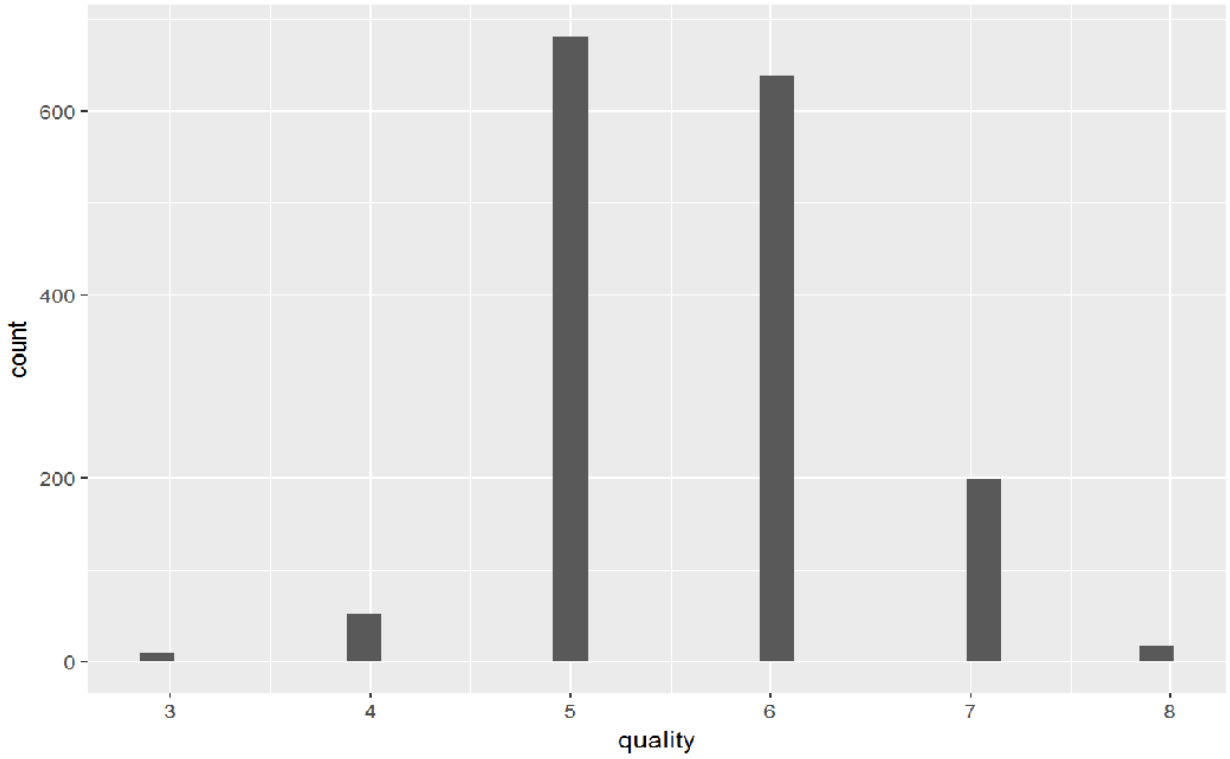

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```

Sulphates are a group of sulfur compounds (including sulfur dioxide which we saw earlier) which are sometimes added to stabilize wine. This is frequently seen as undesirable. Sulphates can also occur naturally. We see a peak at .6 with a long tail reaching to 2. The minimum value is .33 and the max is 2. The median is .62 and the mean is .6581. The quartiles are .55 and .73.



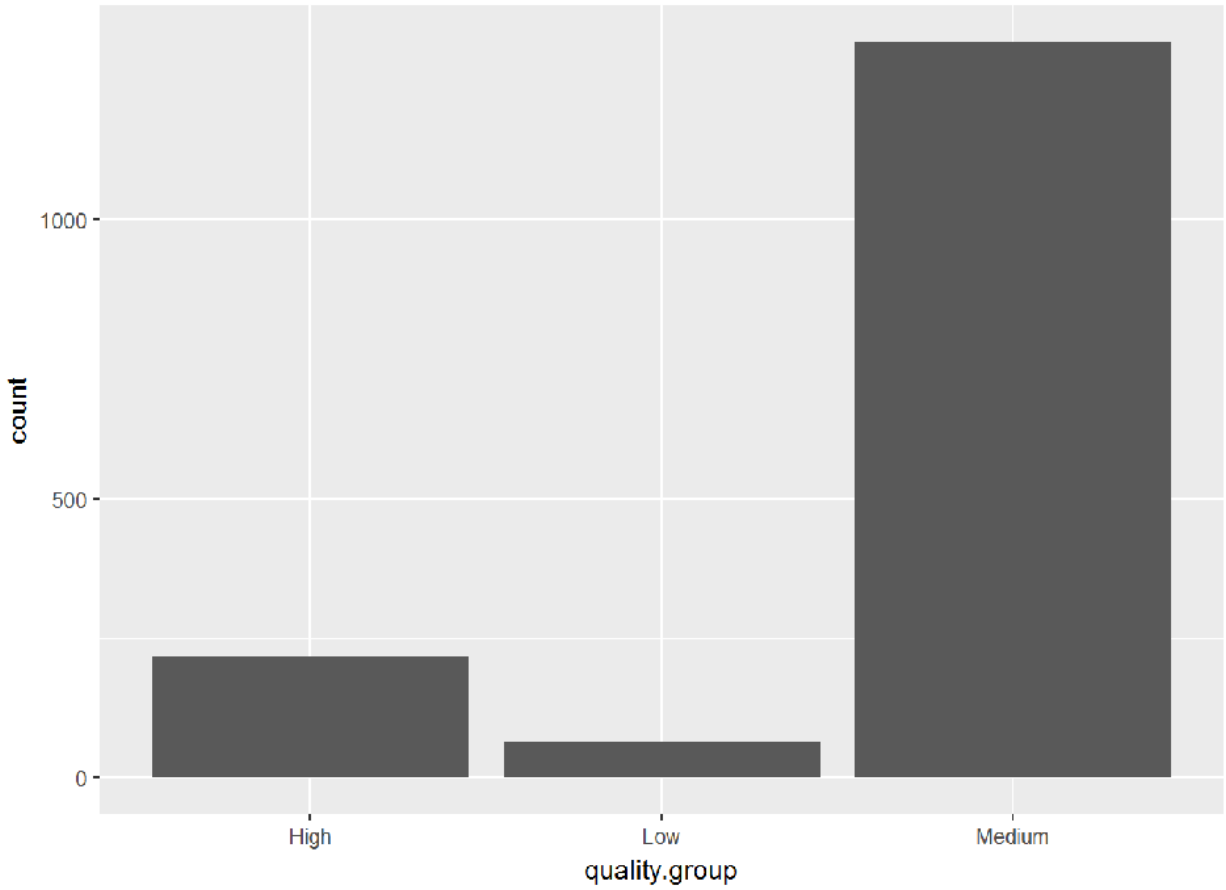

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```

Alcohol content tells us what percentage of the wine is alcohol. We can see that there is more wine with less alcohol with a peaks at 9.5 and a relatively stable drop-off after that.
The minimum value is 8.4 and the max is 14.9. The median is 10.2 and the mean is 10.42. The quartiles are 9.5 and 11.1.

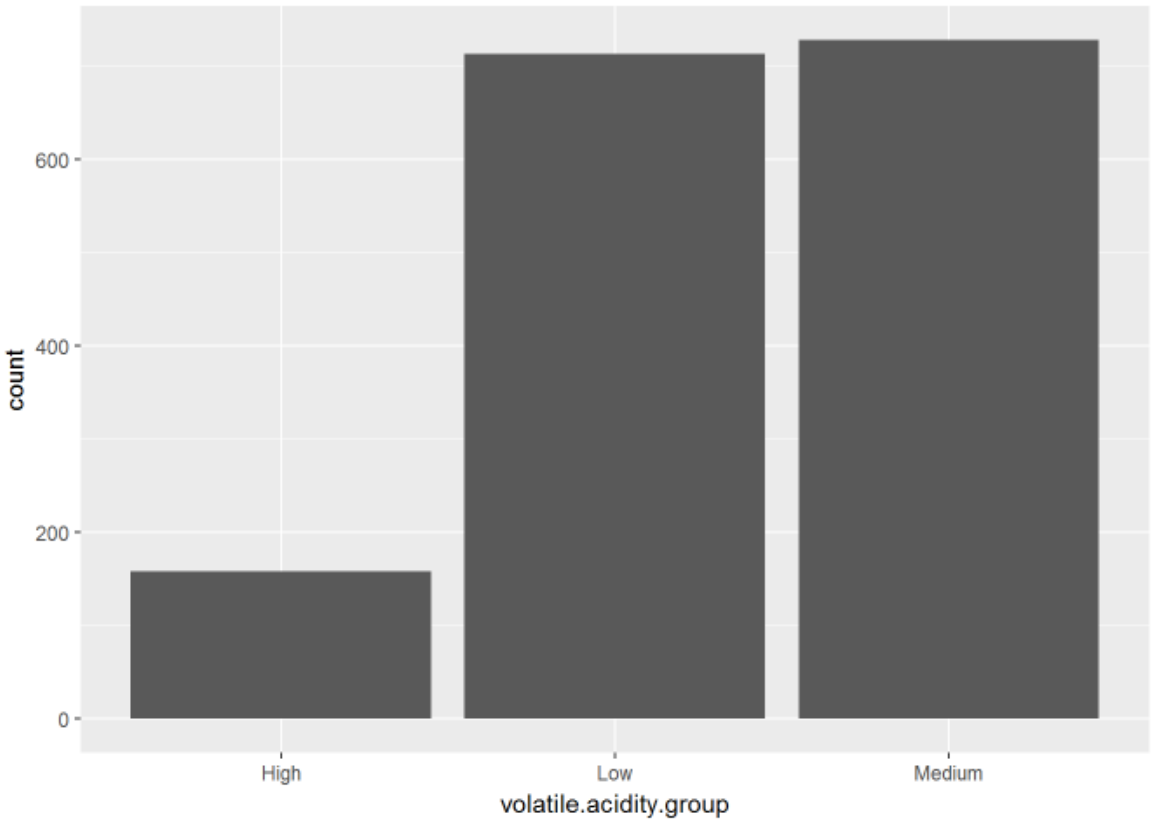


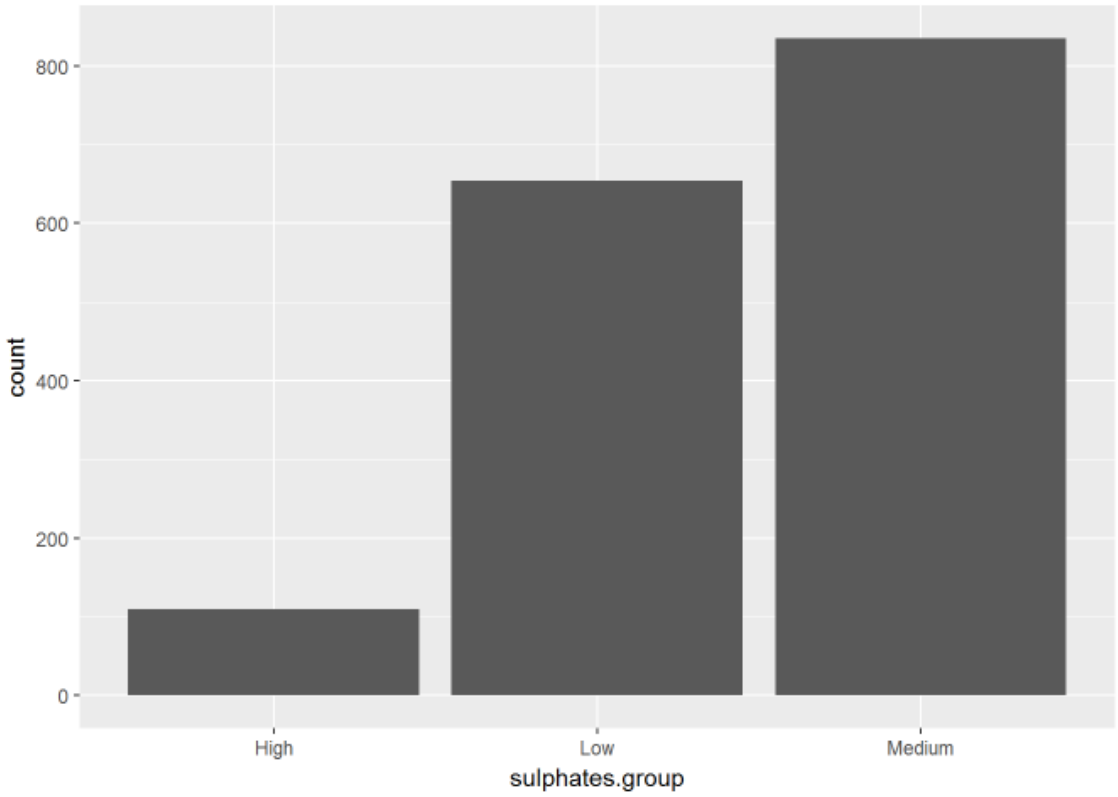

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.636   6.000   8.000
```

Now we come to quality, the rankings of a panel of at least experts. We can see that the lowest rating was 3. Most wines were rated a 5 or 6 with a few exceptional wines rated at 7 and even fewer at 8. The minimum is 3 and the max is 8. The median is 6 the mean is 5.636 and the quartiles are 5 and 6.



For ease of analysis I created a few new variables. For quality, I put the bottom two wines in low quality, the middle two in medium and the upper two in high. It is interesting to not that there are far more medium quality wines than either high or low and low has the fewest.


600
400
200
0
count
High
Low
Medium
volatile.acidity.group


800
600
400
200
0
count
High
Low
Medium
sulphates.group

I created similar variables for sulphates and voltile acids to be used in multivariate analysis. I divided the data into thirds. In both cases, there are more in the medium group and less in the high group.

Univariate Analysis

This data set looks at 1,599 observations of twelves variables related to red wine. I added several variables for clarity in multivariate analysis.(see below) Most of the varibales are realted to the chemical properties of the wine. Quality is a rating given to the wine by at least three experts on tasting

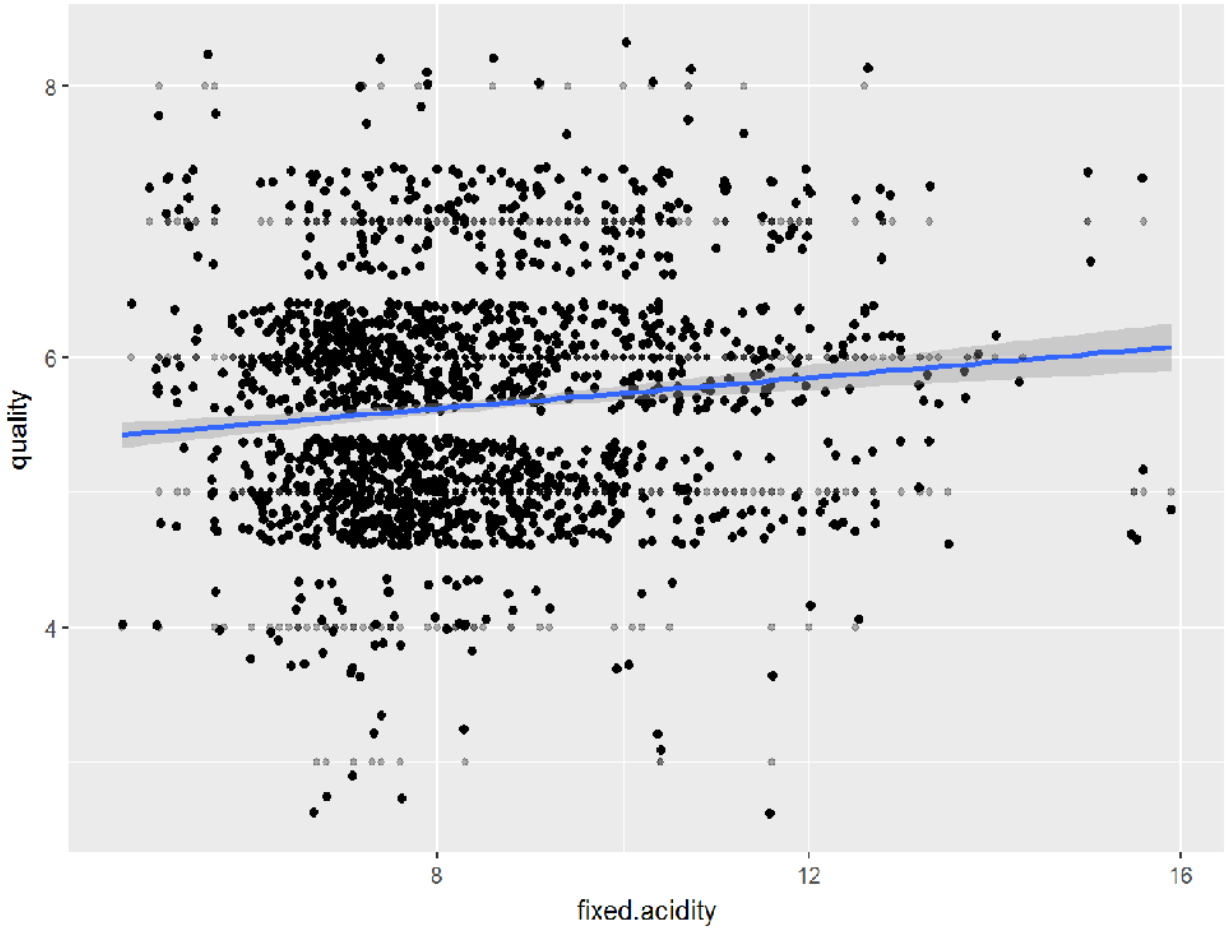
The obvious question to ask with this data is how the various properties of the wine correspong to the quality. In a broad way this provides an answer to the question- what makes a good bottle of red wine?

There are other questions to look at, though. Three variables involve acid, do those values correlate in some way? Similarly, many variables result as natural byproducts of fermentation. Do they relate to each other in a meaningful way?

It was not necessary to create any new variables in this dat set.

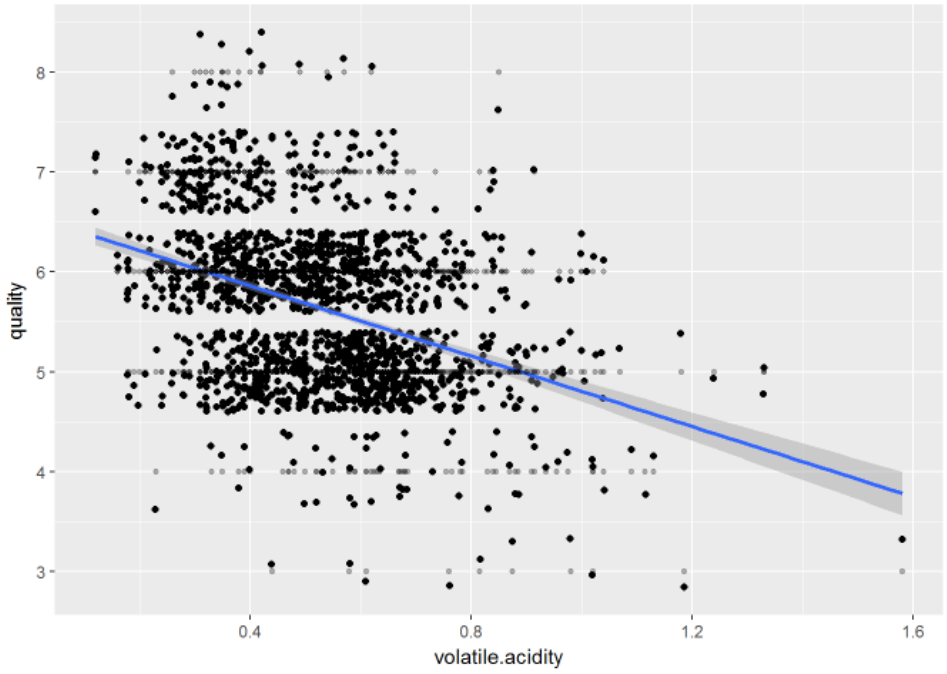
Most of the data is distributed in a relatively smooth way. Many are bell curves others have tails. One odd distribution though is found in amount of citric acid. This is generally a pattern of decrease, there are more bottles with less citric acid. However, there are many peaks that appear within this bigger trend. There is still a clear pattern, but there is more noise than other variables.

Bivariate Plots Section



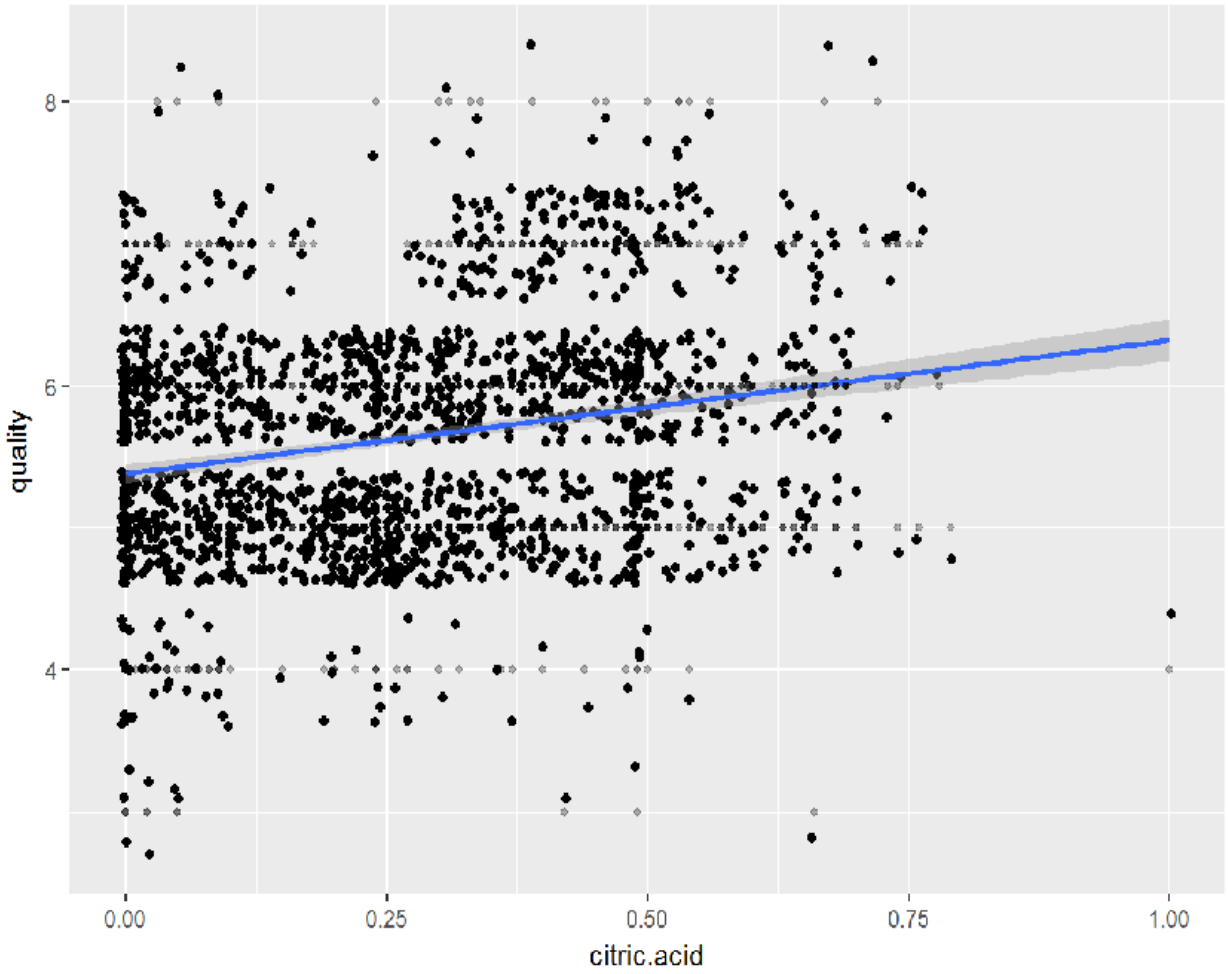

```
## [1] 0.1240516
```

Here we see the relationship between fixed.acidity and quality. There seems to be a slight positive correlation of .124 between the two.



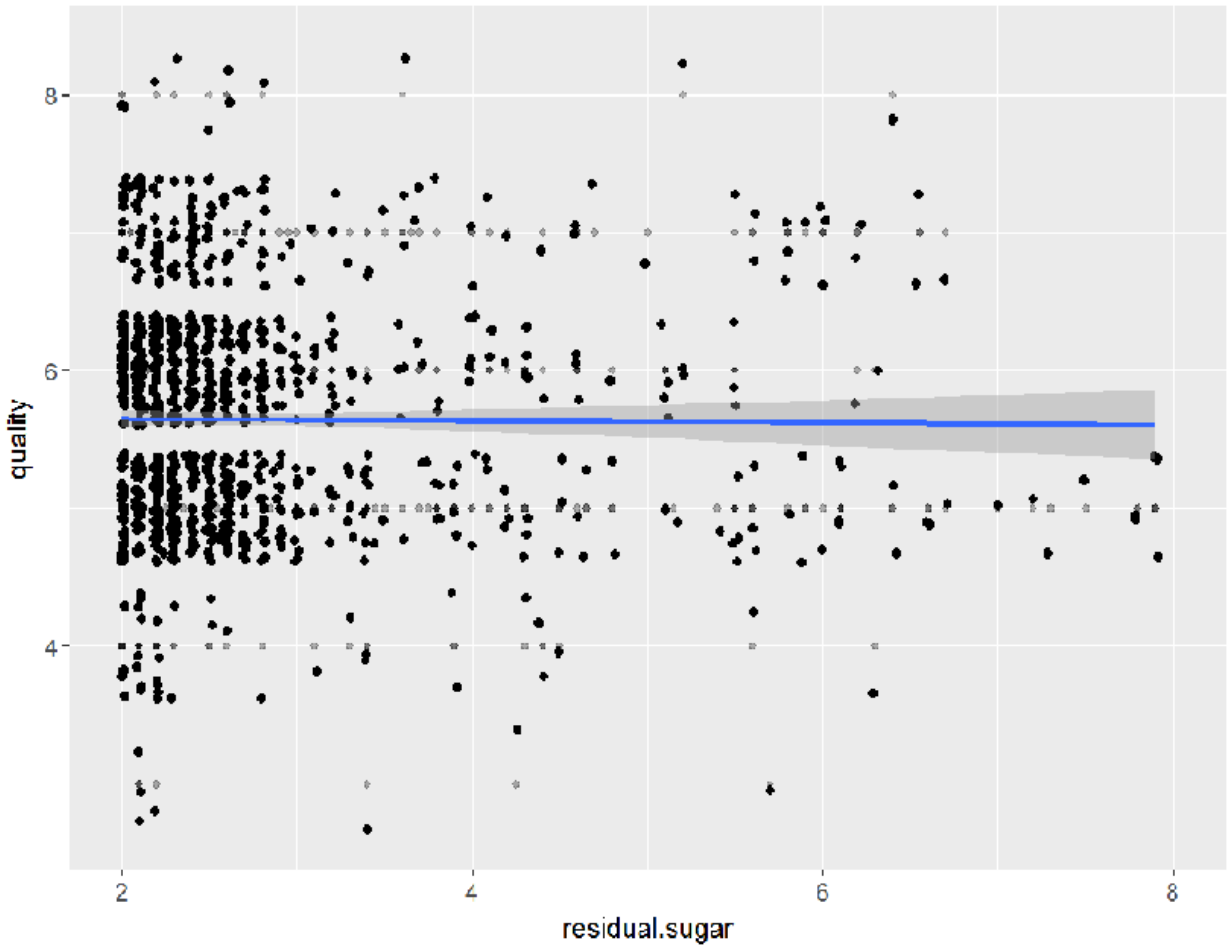

```
## [1] -0.3905578
```

There is a stronger negative relationship of -.39 in this graph comparing volatile acidity and quality.



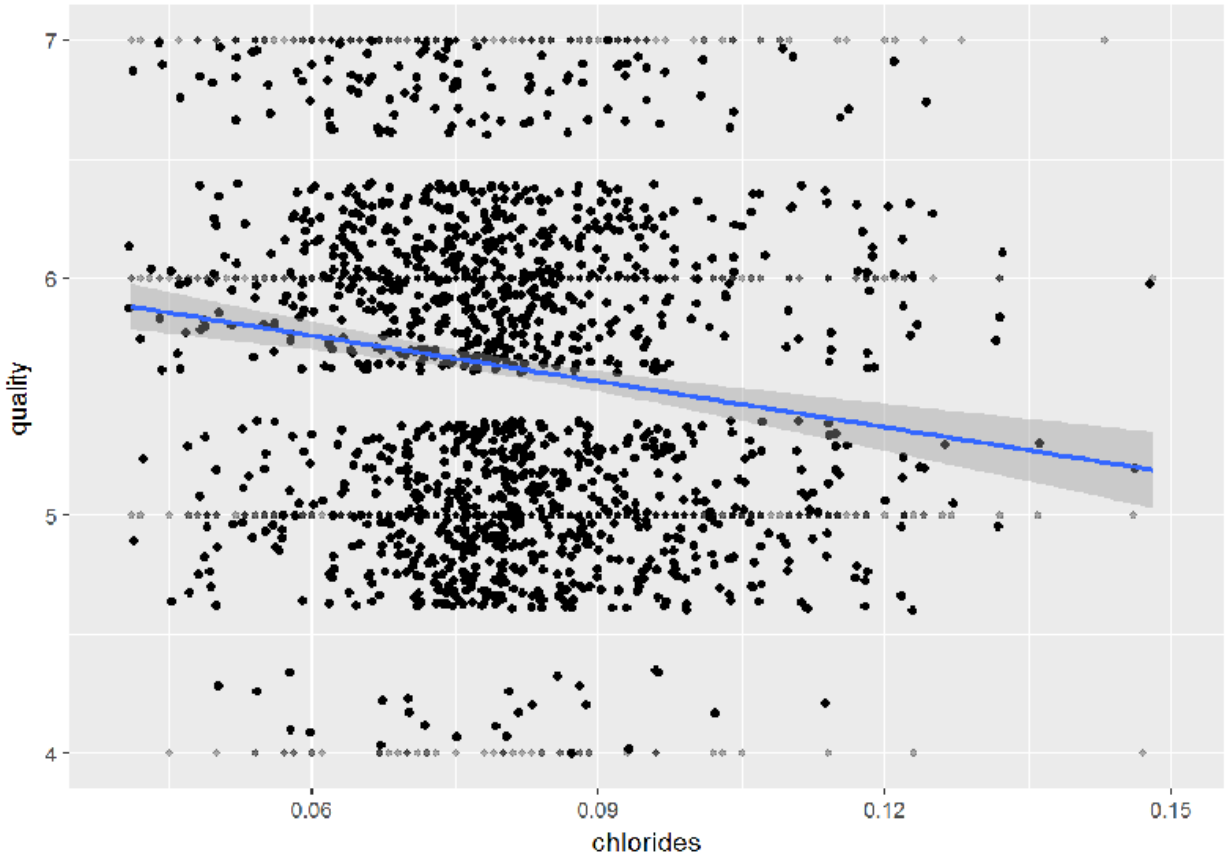

```
## [1] 0.2263725
```

There appears to be a slight positive correlation of .226 between citric acid and quality.



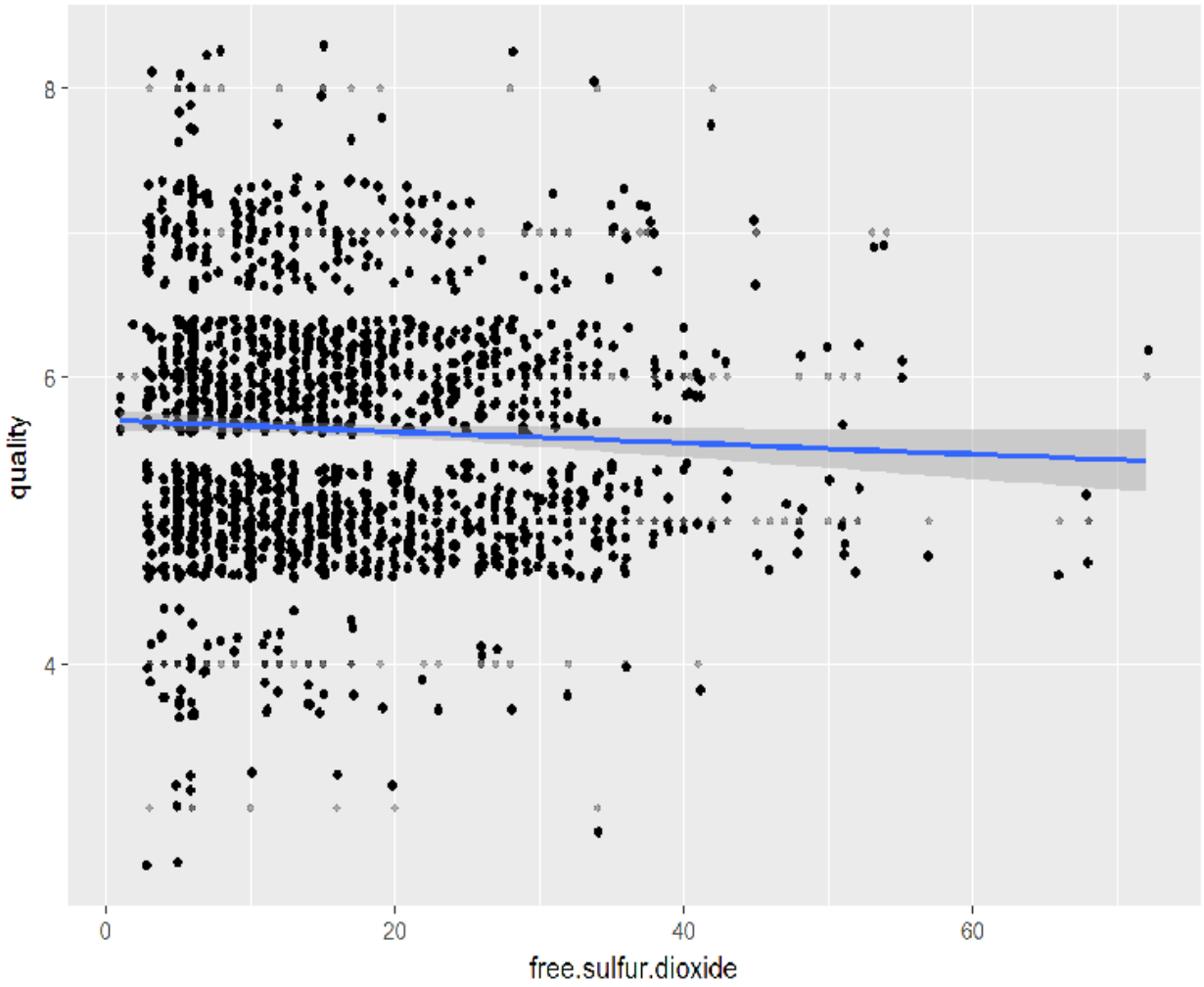

```
## [1] 0.01373164
```

There is a slight positive correlation of .0137, essentially nil between residual sugar and quality.



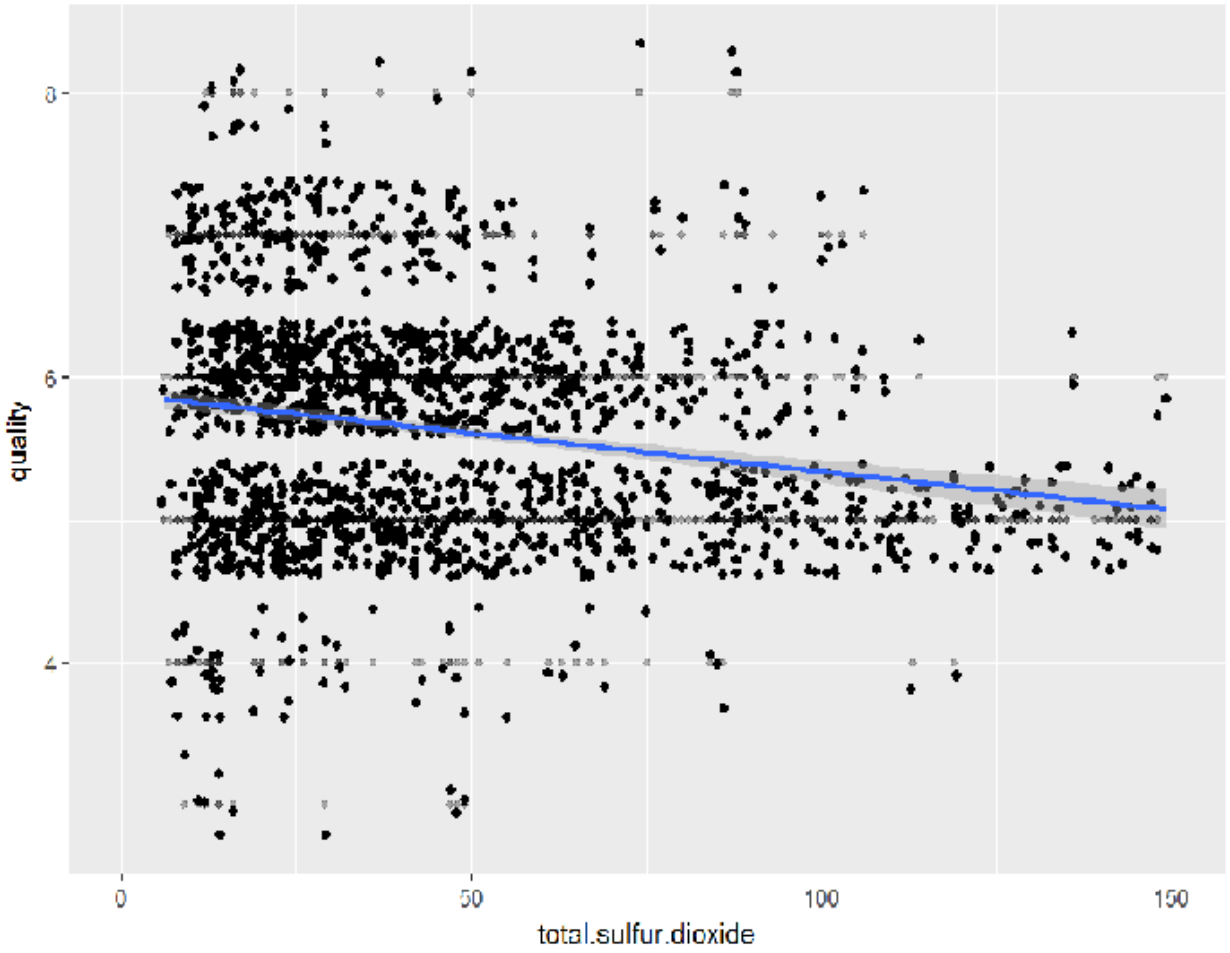

```
## [1] -0.1289066
```

There is a moderate negative correlation of -.12 here between chlorides and quality.



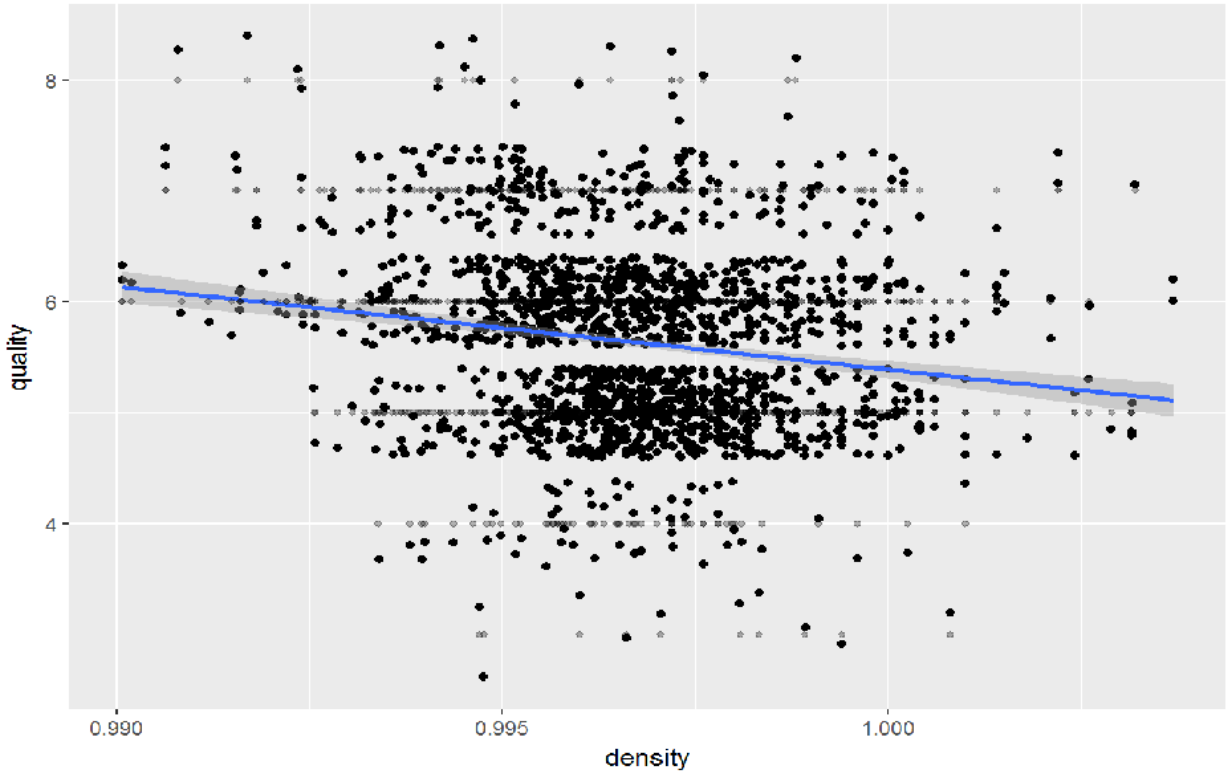

```
## [1] -0.05065606
```

The correlation in this example is -.05, practically non-existent .



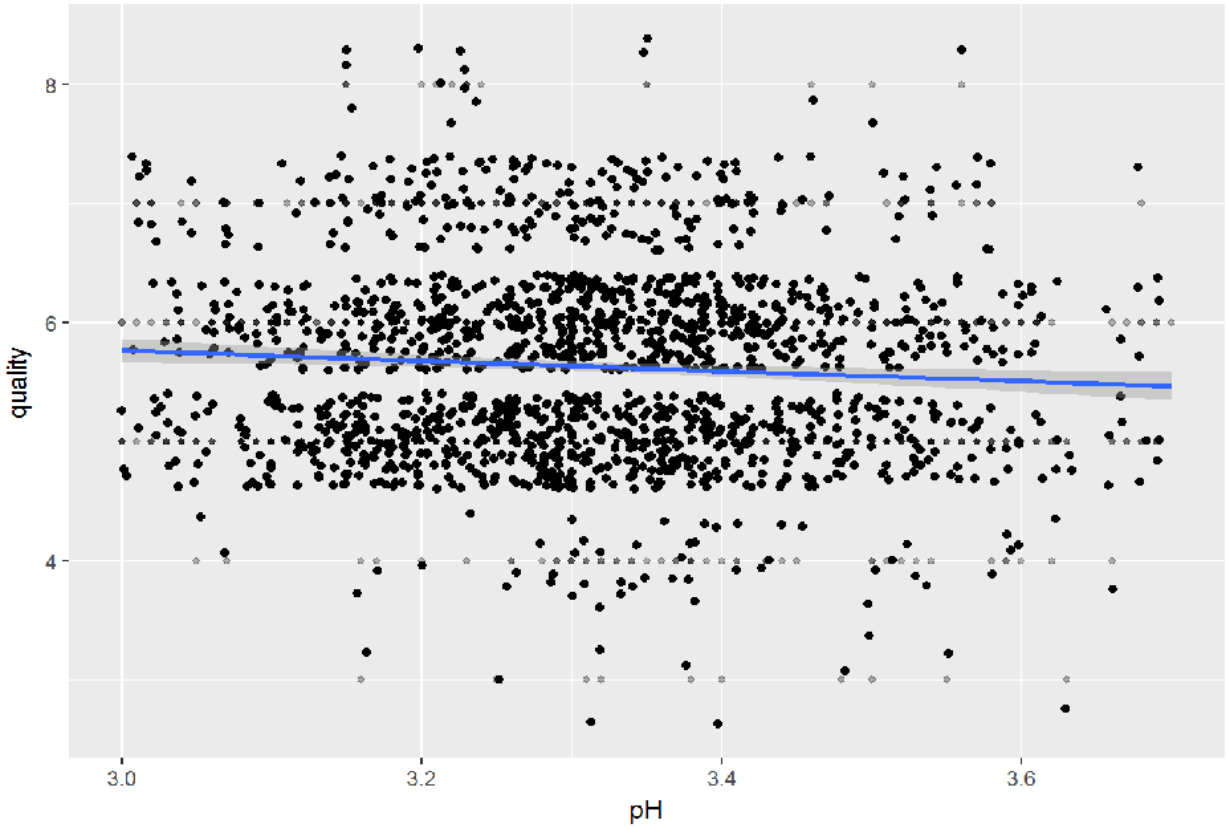

```
## [1] -0.1851003
```

Here there is a clearer relationship than in the last example. The correlation is -.18.



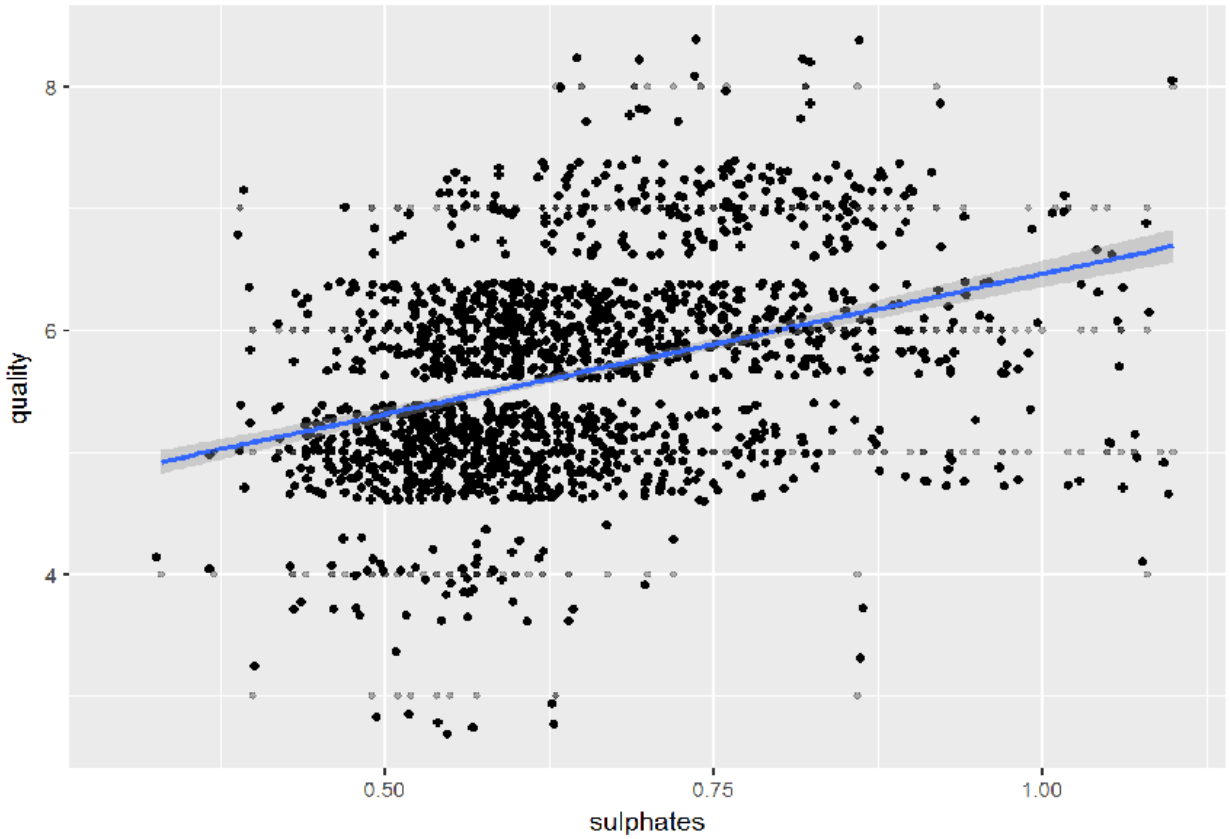

```
## [1] -0.1749192
```

Here there is a slight negative correlation of -.175 between density and quality.



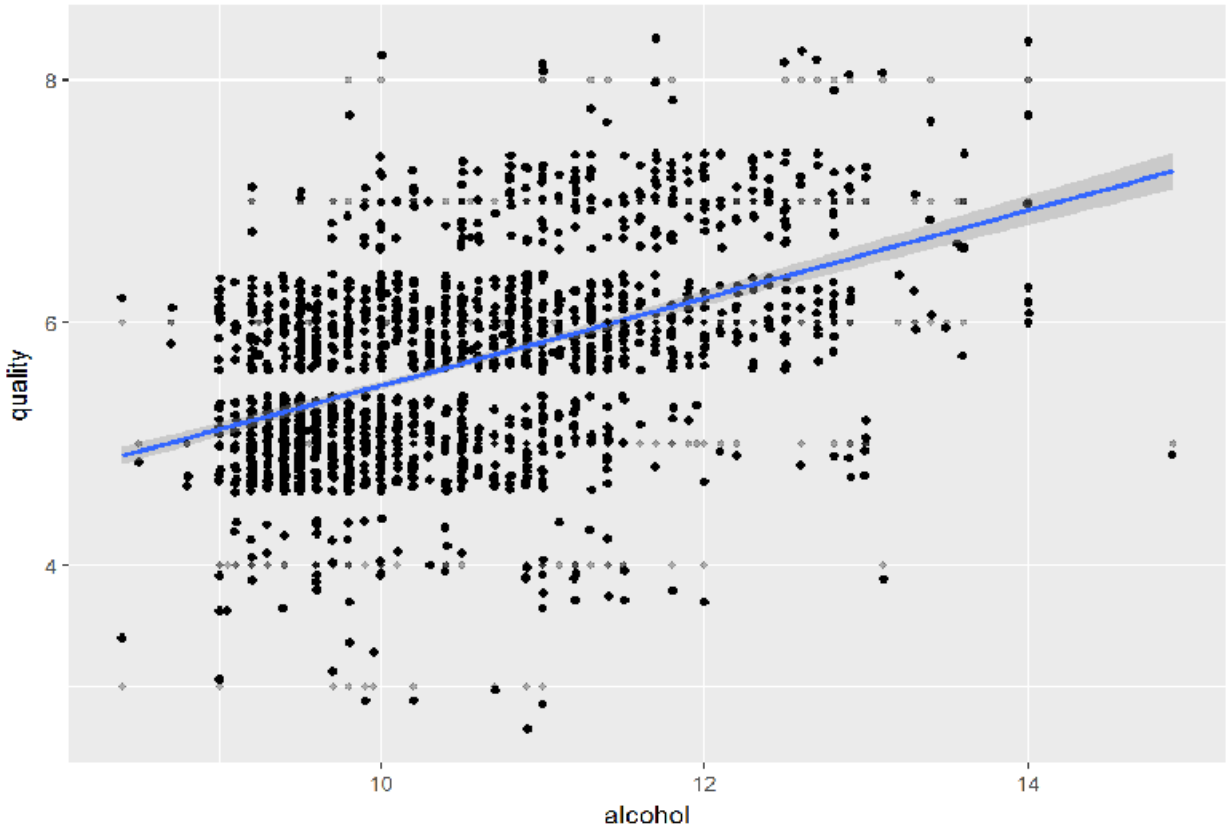

```
## [1] -0.05773139
```

There is not much of a correlation between pH and quality, -.0577.




```
## [1] 0.2513971
```

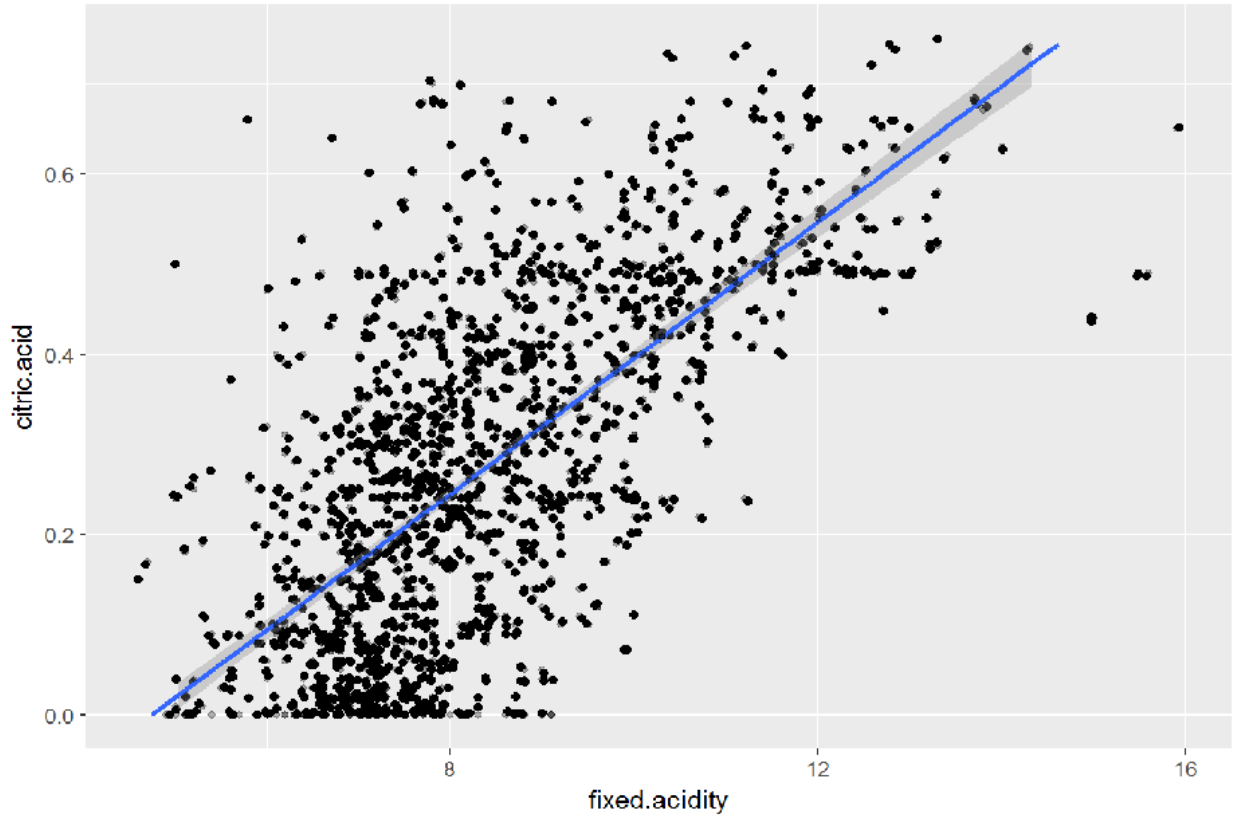
There appears to be a relatively strong positive correlation of .251 between sulphates and quality.




```
## [1] 0.4761663
```

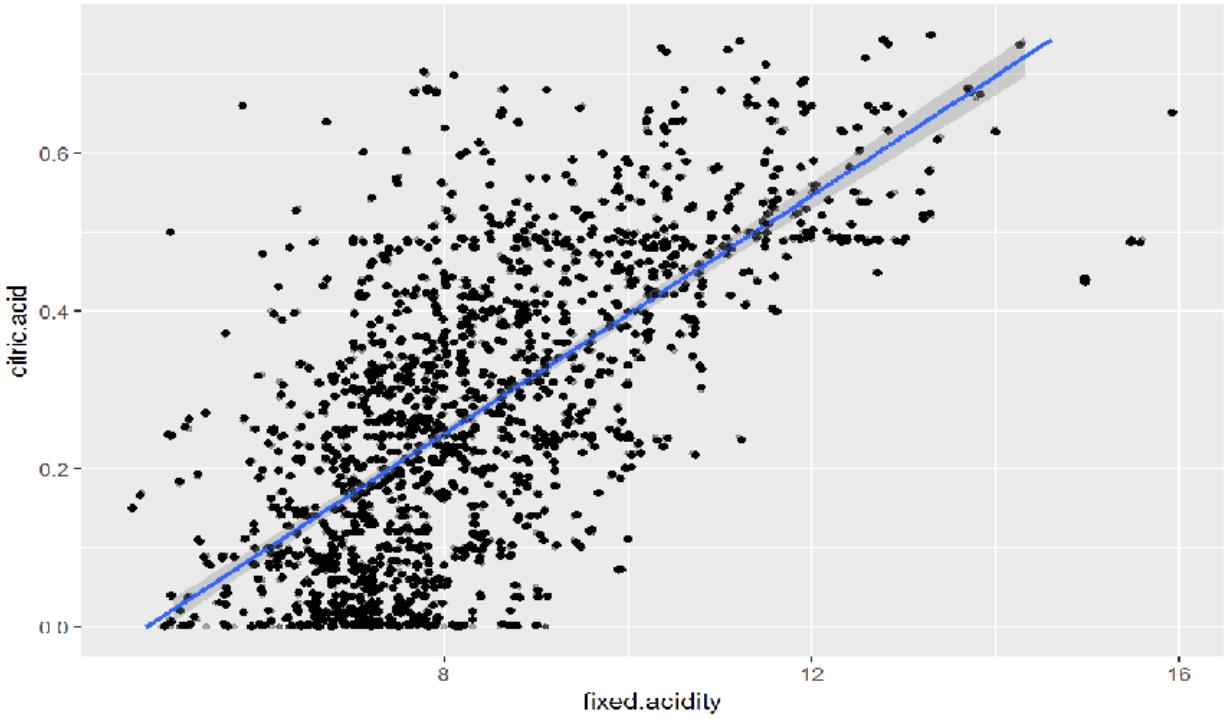
There is a fairly strong correlation between alcohol content and quality, .476.

Other Bivariate Plots



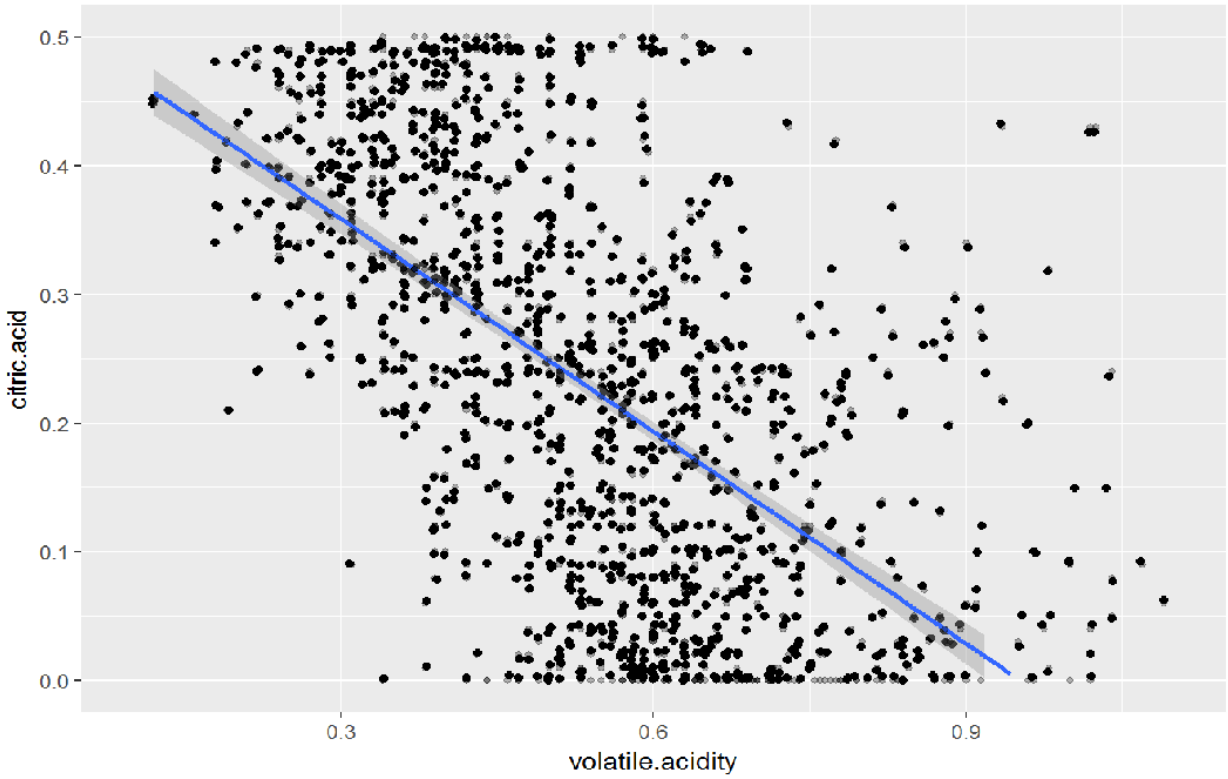

```
## [1] -0.2561309
```

There is a moderate negative correlation of -.256 between fixed and volatile acidity.



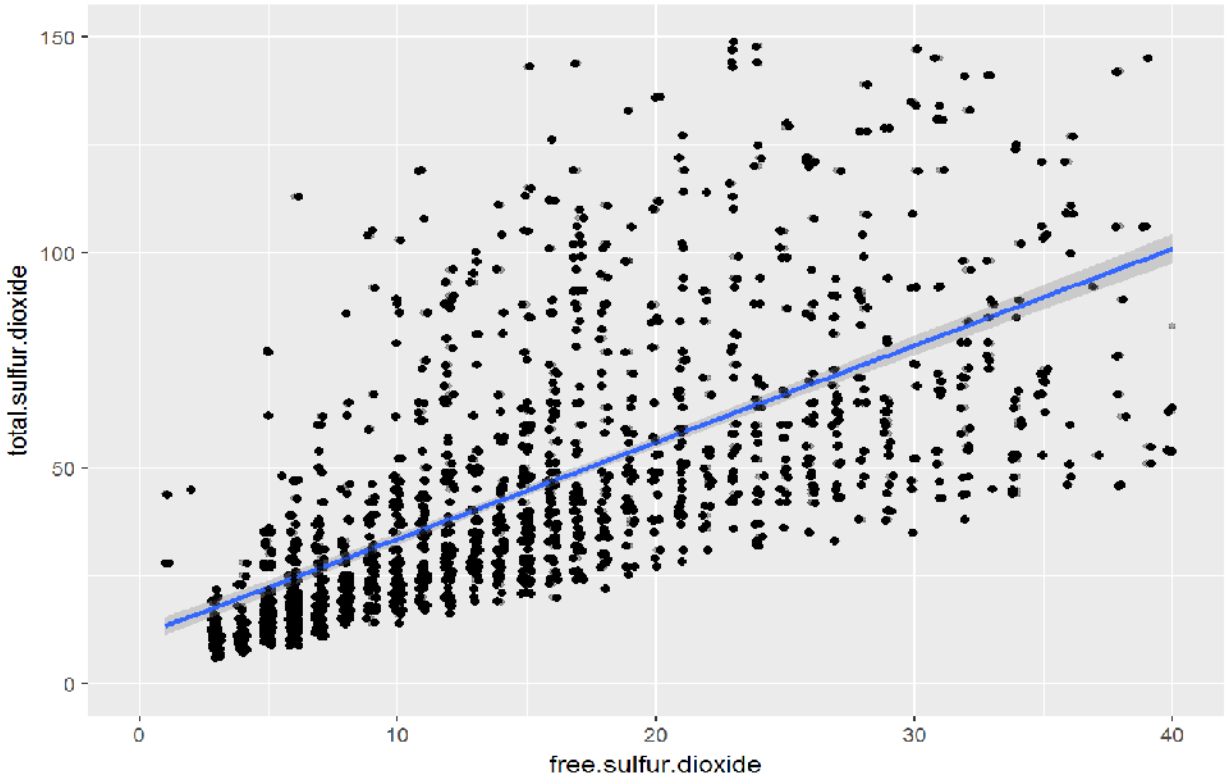

```
## [1] 0.6717034
```

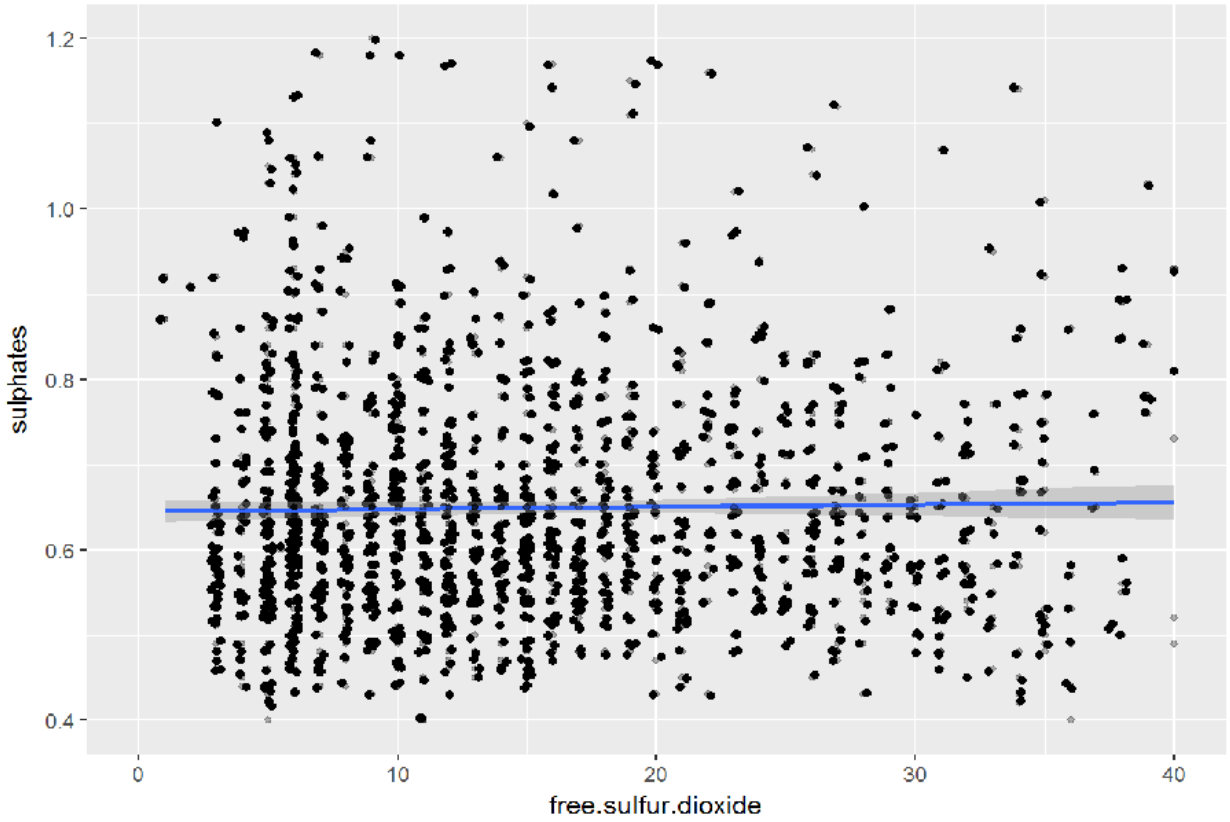
There is a strong positive relationship of .671 between fixed acidity and citric acid.

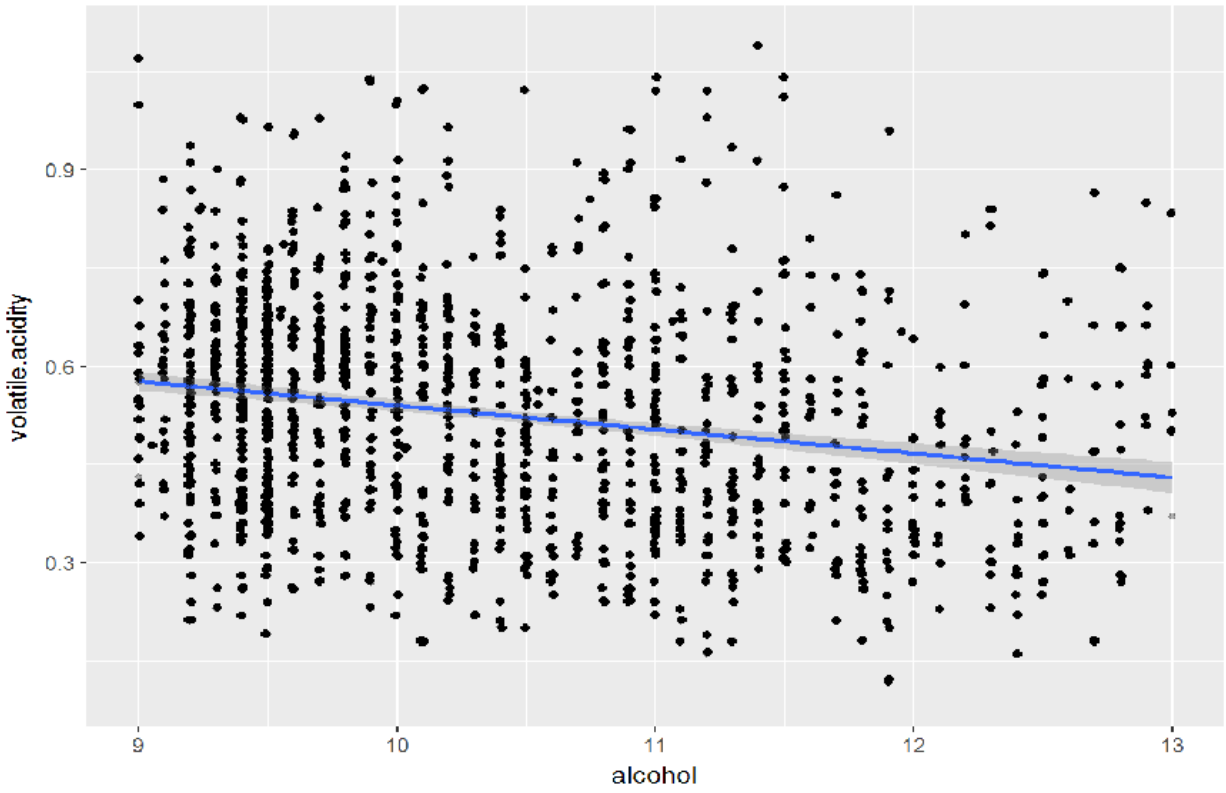


```
## [1] -0.5524957
```

There is a strong negative correlation of .552 between volatile acidity and citric acid.

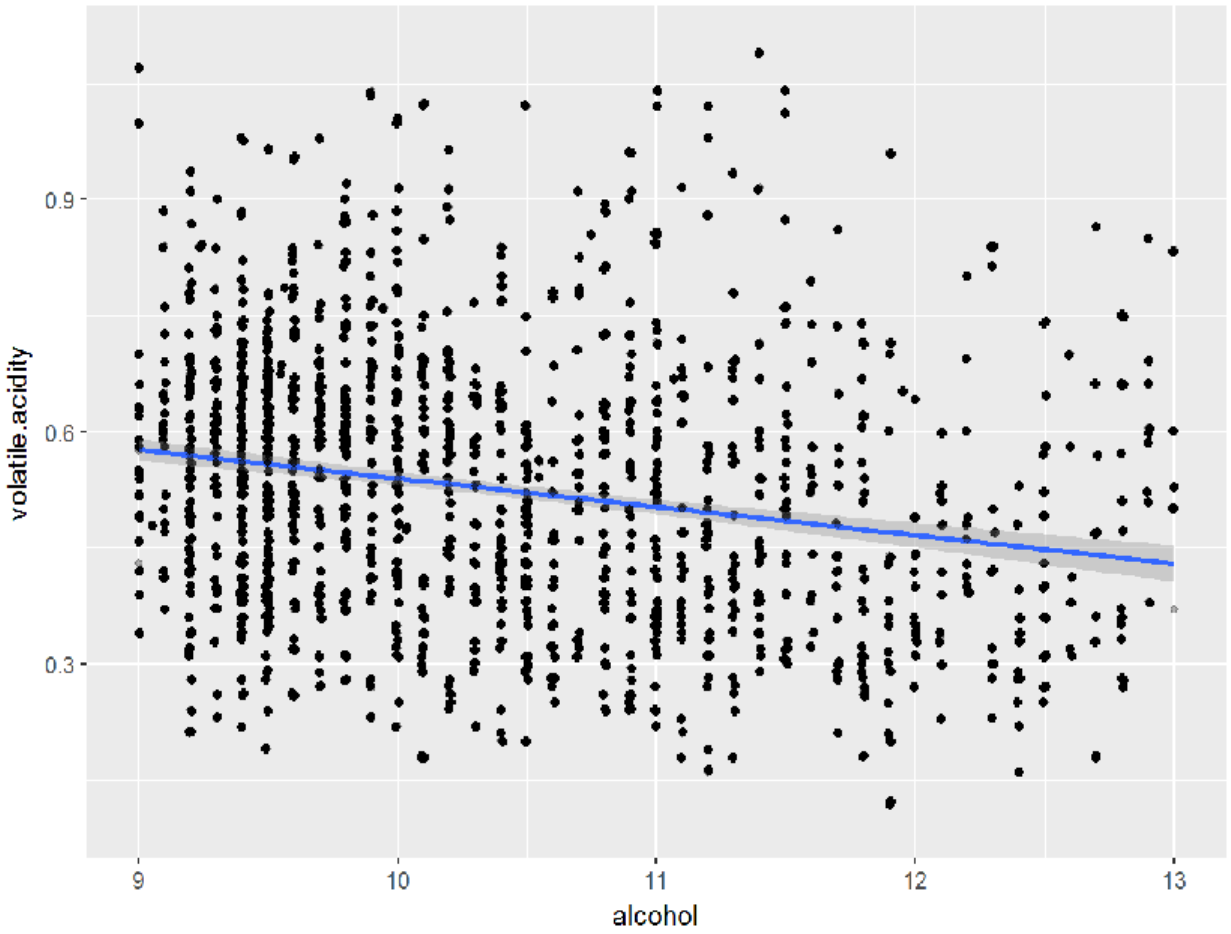



sulphates
1.2
1.0
0.8
0.6
0.4
0
10
20
30
40
free.sulfur.dioxide


volatile.acidity
0.9
0.6
0.3
9
10
11
12
13
alcohol

```
## [1] 0.6676665
## [1] 0.05165757
## [1] 0.04294684
```

There is a fairly strong relationship of .668 between free and total sulfur dioxides. Oddly there is only relationship of .051 between free sulfur dioxides and sulphates and an even lower correlation of .042 between total sulfur dioxides and sulphates. This is unusual as sulfur dioxide is generally considered a kind of sulphate.




```
## [1] -0.202288
```

There is a medium negative correlation of -.202 between alcohol and volatile acidity.

Bivariate Analysis

The first question I had involved what variables affected quality in what way. There strongest positive correlations were found in alcohol and sulphate content. It is worth noting that there are many who would rather not drink wine with added sulphates. But based on the data they seem to be doing their job of stabilizing the wine and improving its quality.

The strongest negative factor in the wines is volatile acidity. This makes sense after all volatile acidity is associated with vinegar and other unwanted acids.I will return to volatile acids later in this analysis.

Citric and fixed acids both seem to have some positive correlation with quality with citric acid having a stronger relationship.

pH, free sulfur dioxide and residual sugar both seem to have a mostly neutral relationship with quality. pH and free sulfur dioxide is slightly negative and residual sugar is slightly positive but none seem as important as other factors. Interestingly, residual sugar showsa large amount of clumping on the left side of the graph.This would suggest that wines with less sugar are judged on other factors while wines with more sugar begin to benefit, again ever so slightly, from being a little sweeter.

Density seems to have a low negative relationship with quality. This could suggest that less desirable chemicals may be slightly more dense or that more desirable chemicals may be less dense. Either way, this is a secondary consideration.

Fianlly, chlorides and total sulfur dioxide have a moderate negative correlation with quality. Chlorides are associated with bitter, salty andf soapy flavors, which are hardly popular in wine. Total sulfur dioxide is an interesting case. Total sulfur dioxide is the sum of both free and chemically bound sulfur dioxide compounds. We have data on free sulfur dioxide and can infer something about bound sulfur dioxide from the graphs. Since total sulfur dioxide has a stronger negative correlation than free sulfur dioxide, this would suggest that bound sulfur compounds are more detrimental to the overall quality than free sulfur compounds.

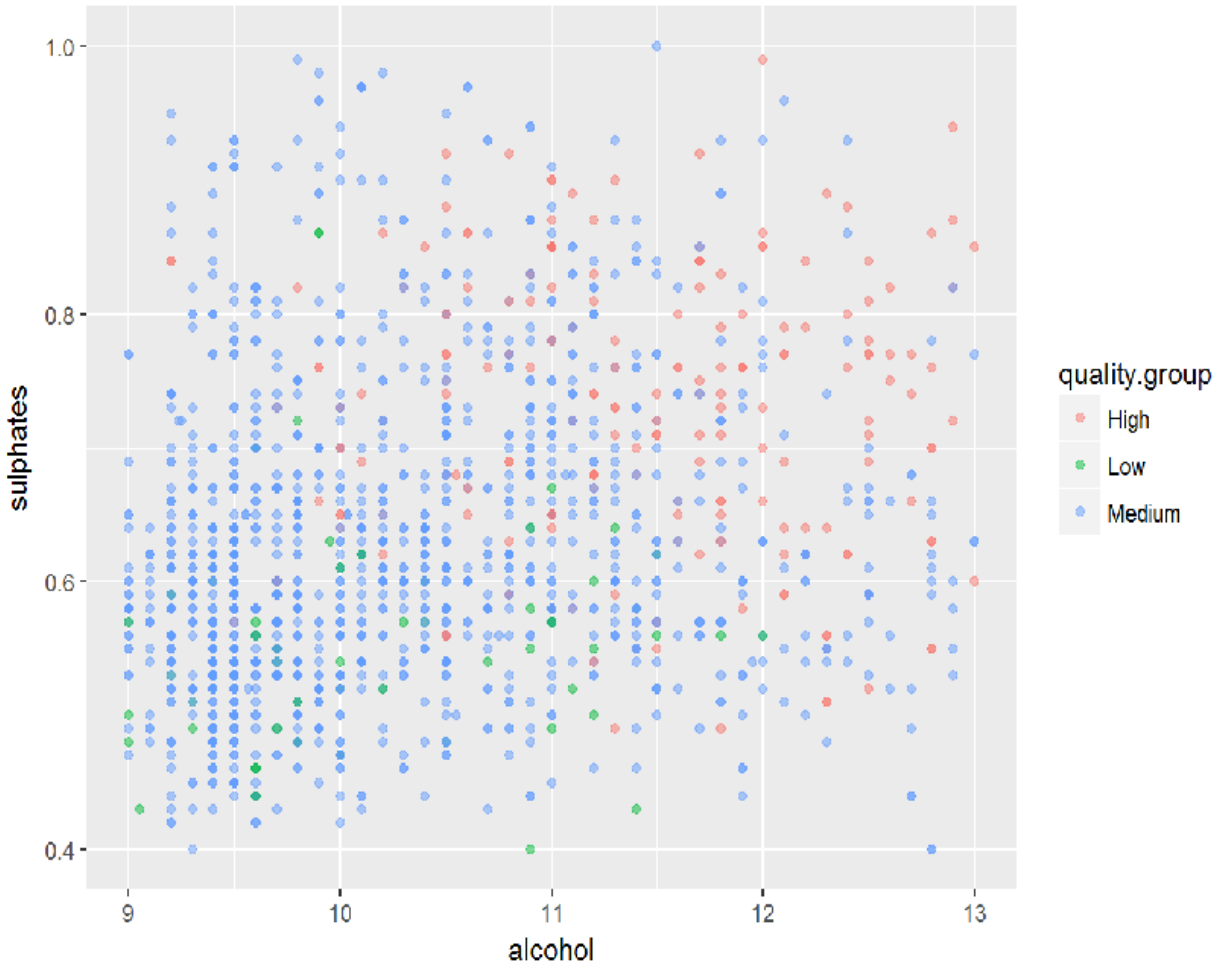
Putting this all together, the best red wines have more alcohol and use sulphates, but not sulfur dioxide, to stabilize the wine. There are more fixed and citric acid, but less volatile acids. Density and pH may vary and there is a slight chance it is a little sweet.

But that is not the end of the analysis. I was also curious at how some of the other features related to each other. I began by exploring the various kinds of acids. Citric acid is often considered a kind of fixed acid, so I was curious how those variables would compare in the dataset. Interestingly there are three strong correlations. Citric acid and fixed acid seem to be very strongly positively correlated, while volatile acidity is negatively correlated with the other two kinds of acid.

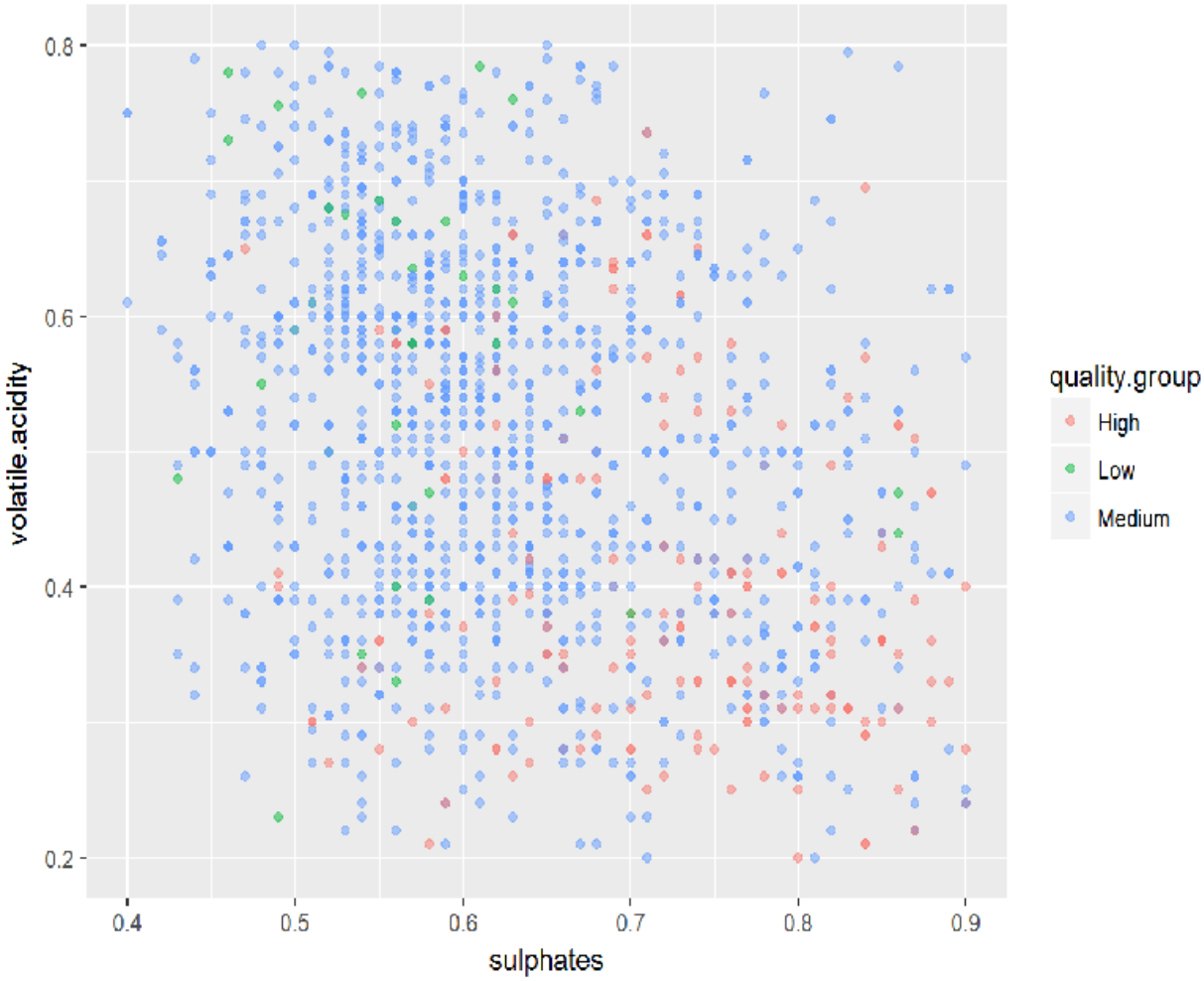
Another set of variables that were related were the sulfure dioxide variables and the sulphate content. Free sulfur dioxide is a part of total sulfur dioxide and both are considered sulphates. Interestingly, sulphates do not correlate strongly with either free or total sulfur dioxide, although both relationships are positive. Total and fixed sulfur dioxide have a medium positive correlation, which, again, was less than expected.

Finally, I waned to see how volatile acidity and alcohol compared. Volatile acids are often formed when something goes wrong in the fermentation process producing vinegars instead of alcohol. The expected result would be less volatile acidity as alcohol increases. We do see this effect, although it is a medium correlation.

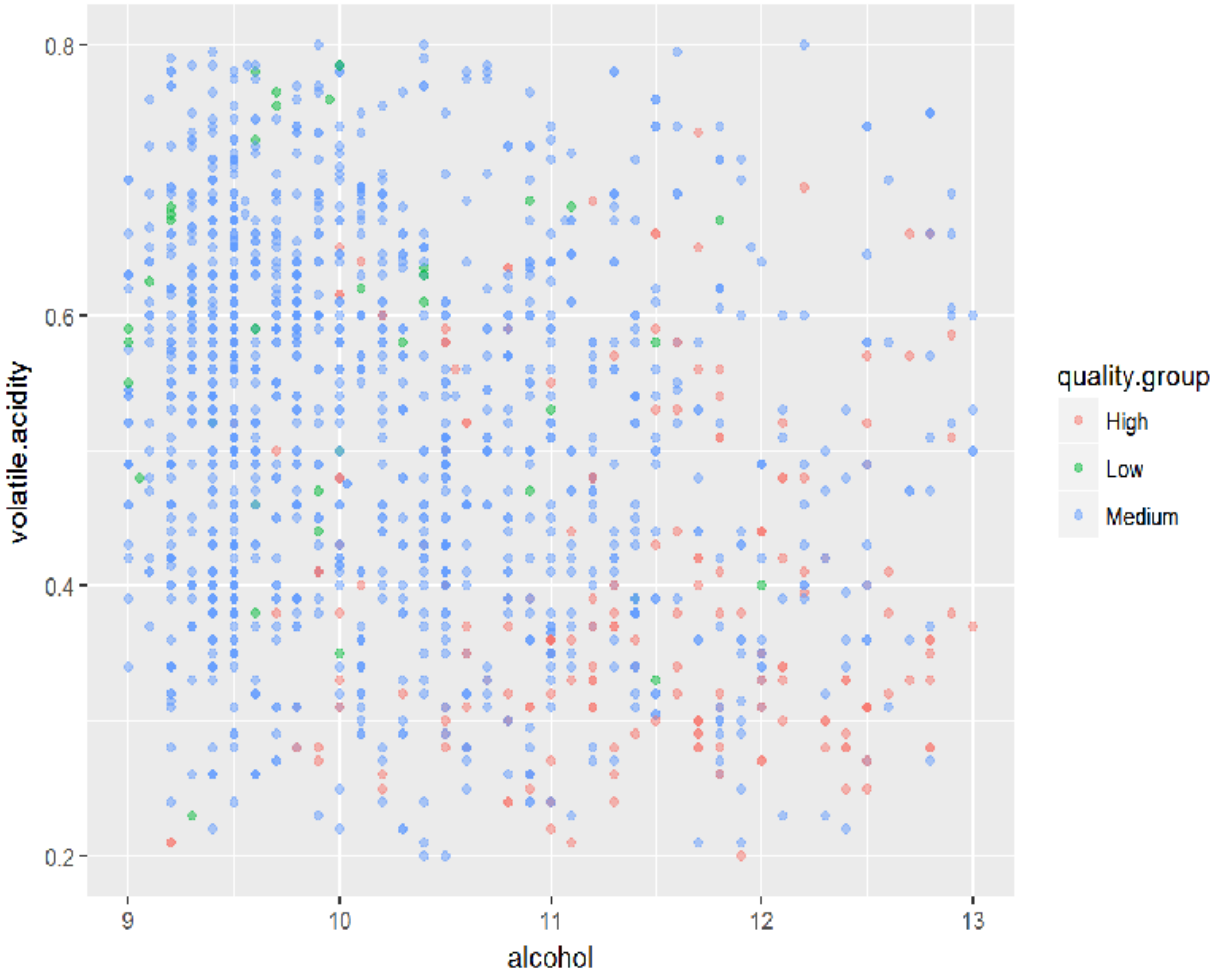
Multivariate Plots Section



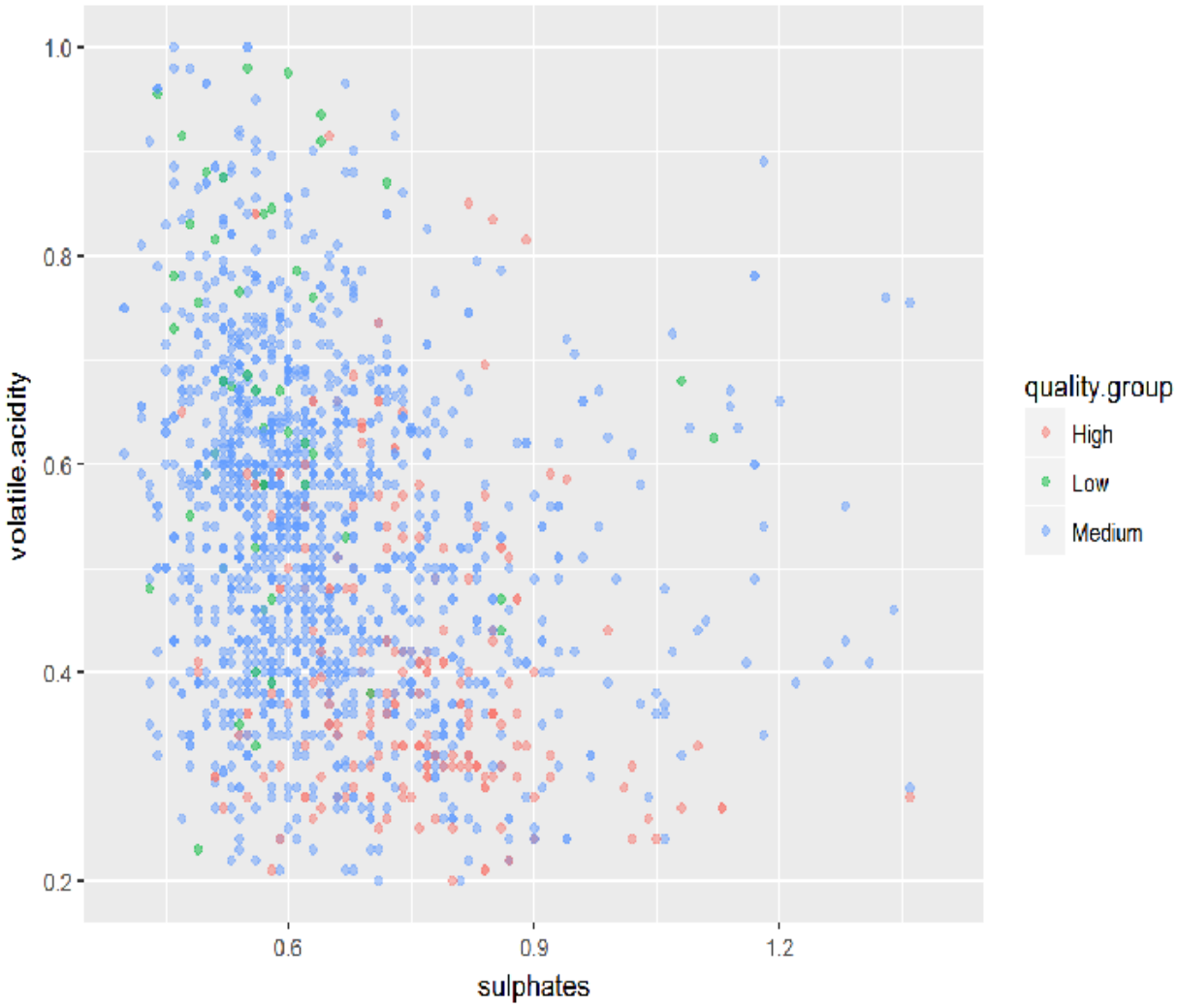
Here we can see the relationship between alcohol and sulppates with a third variable added for quality. I chose these two based on their high correlation with quality. As expected, the more alcohol and sulphates rise together, the better the quality.



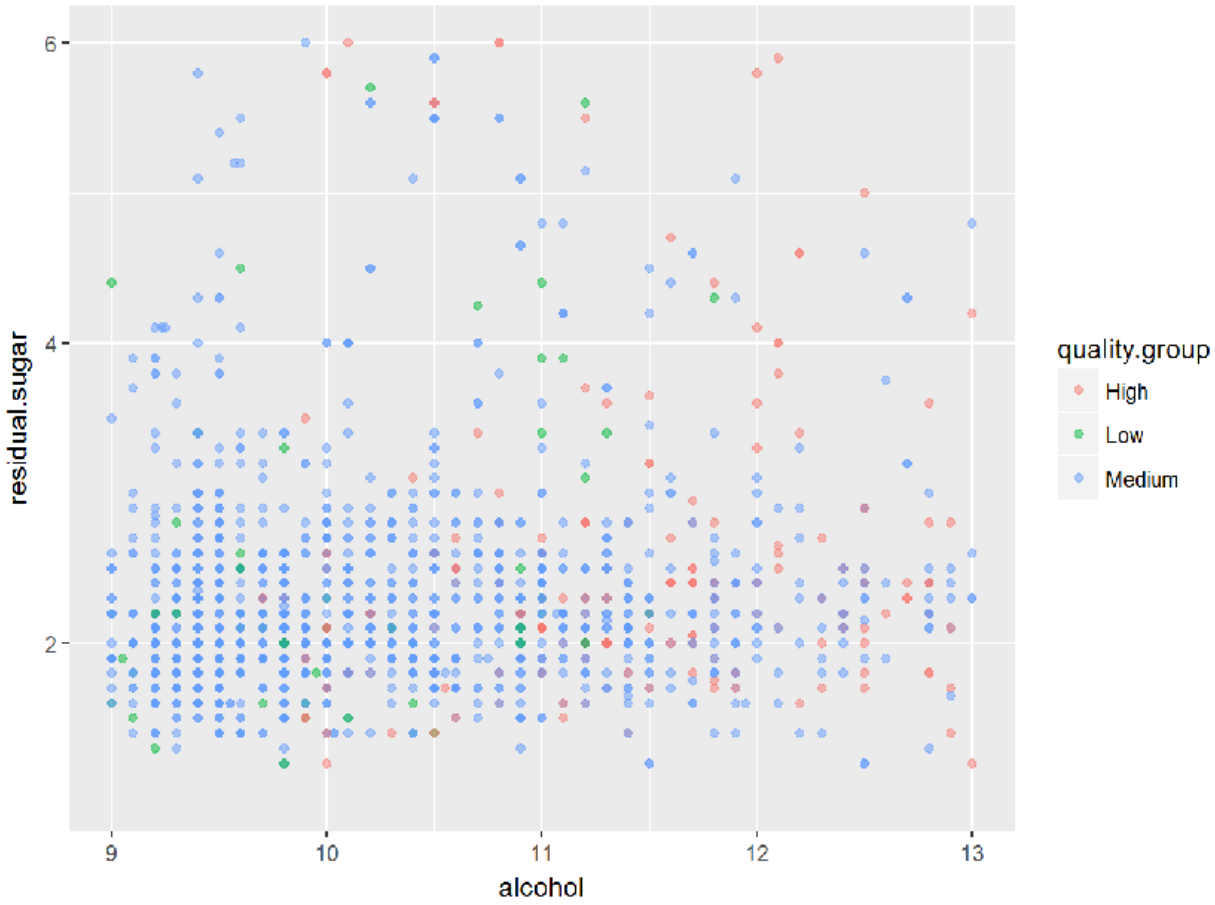
Next I looked at the strongest positive and negative correlations with quality, sulphates and volatile acidity. As expected, the more sulphates with less volatile acids, the better the quality.



Alcohol and volatile acidity show a similar trend. The more alcohol and less volatile acidity, the better the quality.

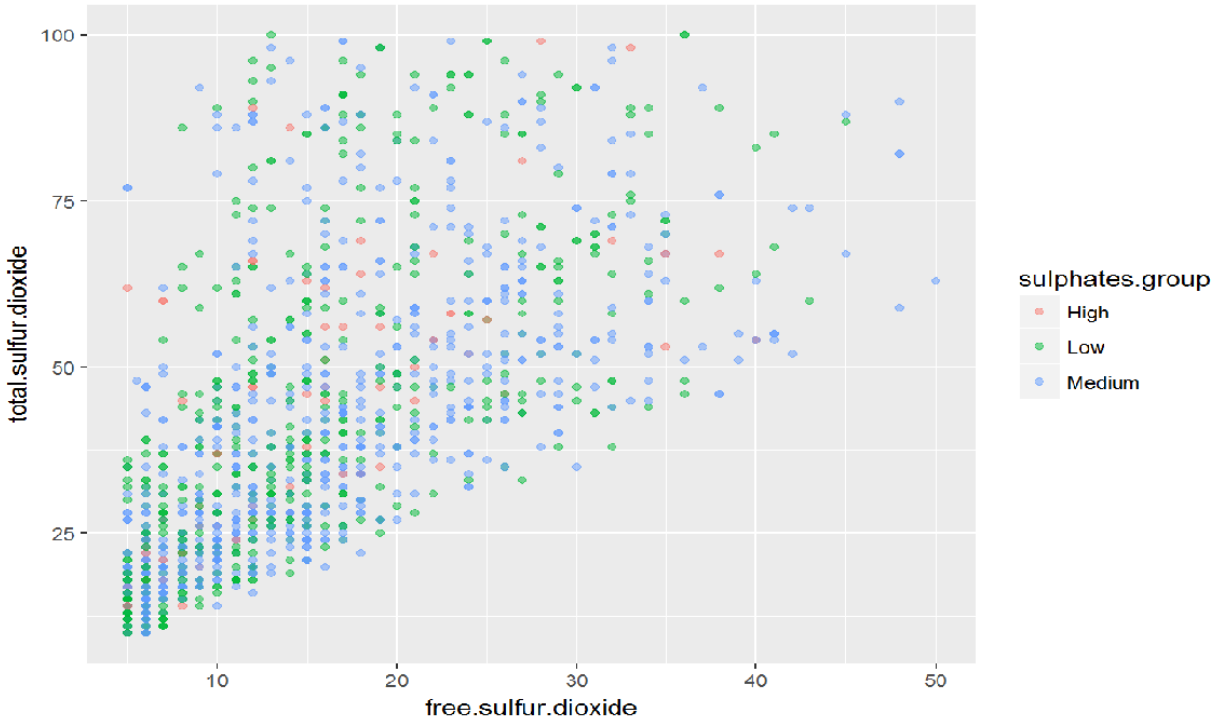


Here we see the relationship between volatile acids and sulphates. More sulphates means better quality, but the relationship seems strongest when volatile acids are low.



At this point, I checked my intuitions by comparing a variable with high correlation with ine that had very little. As expected residual sugar had little correlation with quality before and we can see the same results here.

Here I analyzed the relationship between the various kinds of acids. We can see that as fixed and citric acids increase, there is a lower amount of volatile acids.



Comparing the various forms of sulfur dioxide to the sulphate groups we can see that as total sulfur dioxide and free sulfur dioxide rise, there is a less consistent pattern, although there is less in the medium and low range.

Multivariate Analysis

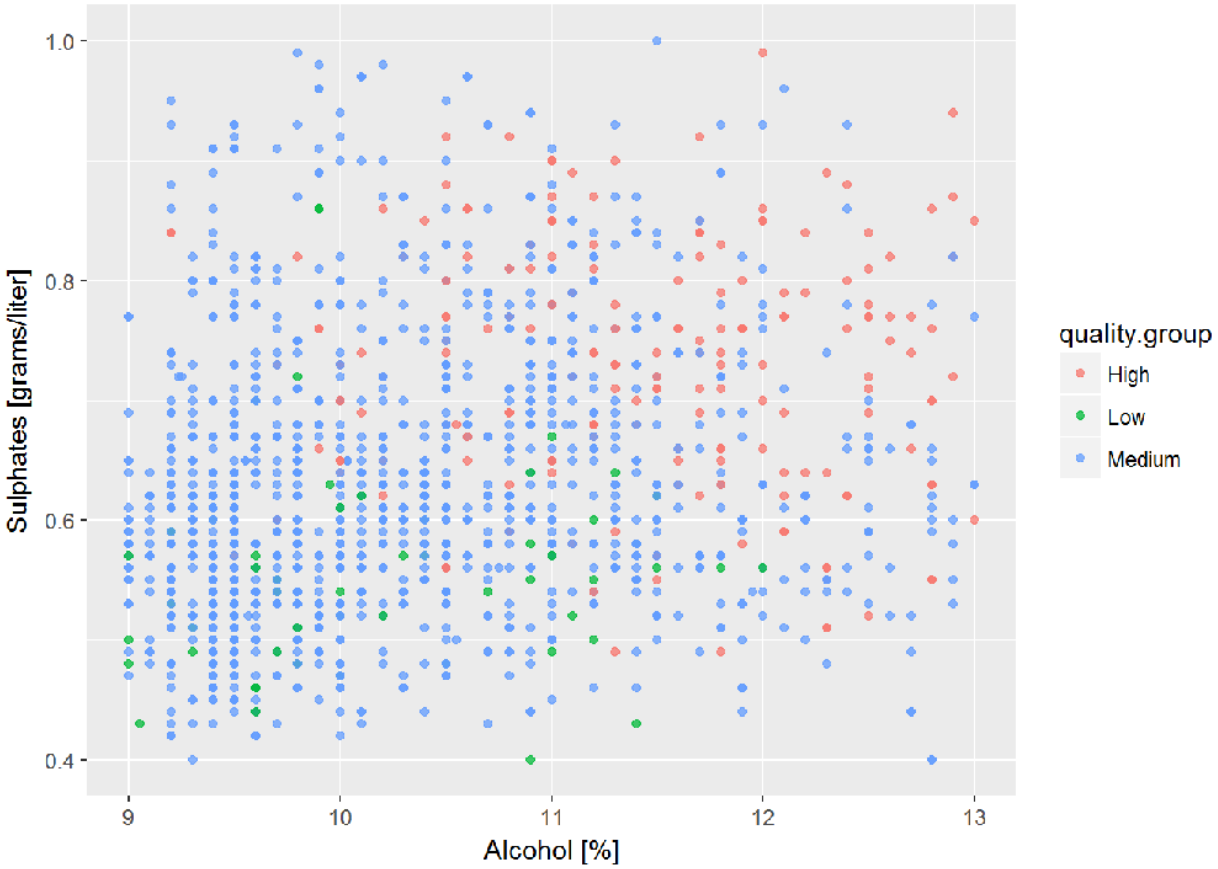
I tested my intuitions from the bivariate plots and found they held up. First, I created three new variables where I divided a previous variable into three boxes- High, Medium and Low. I did this for quality,volatile acidity and sulphates tomake them easier to see on their respective plots.

We can see that as alcohol and sulphates increase, quality also increases, that fixed and citric acidity are negatively correlated with volatile acidity. Volatile acidity and alcohol have the opposite relationsip. We can see that as fixed and citric acids rise together as volatile acids drop.

Once again there is less of a clear cut relationship between sulphates and sulfur dioxide than would be expected given that sulfur dioxide is usually classified as a sulphate.

Final Plots and Summary

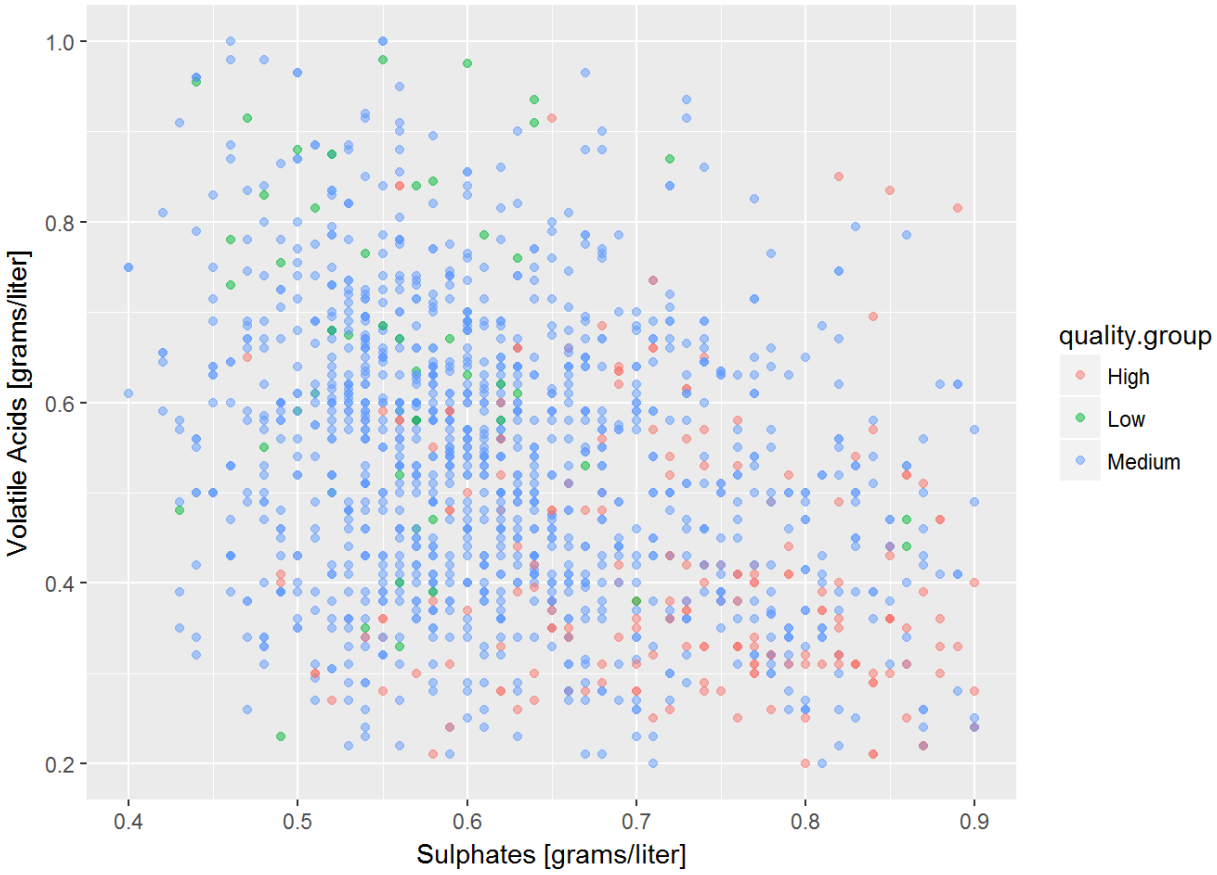
Plot One



Description One

This graph shows the relationship between the two variables most strongly tied to an increase in quality. It shows that both alcohol content and sulphates are important for a high quality wine. A closer look also reveals that there are more low quality wine high in alcohol and low in sulphates than low in alcohol and high in sulphates. Given the bad reputation added sulphates has, this shows that it can be the crucial difference in quality.

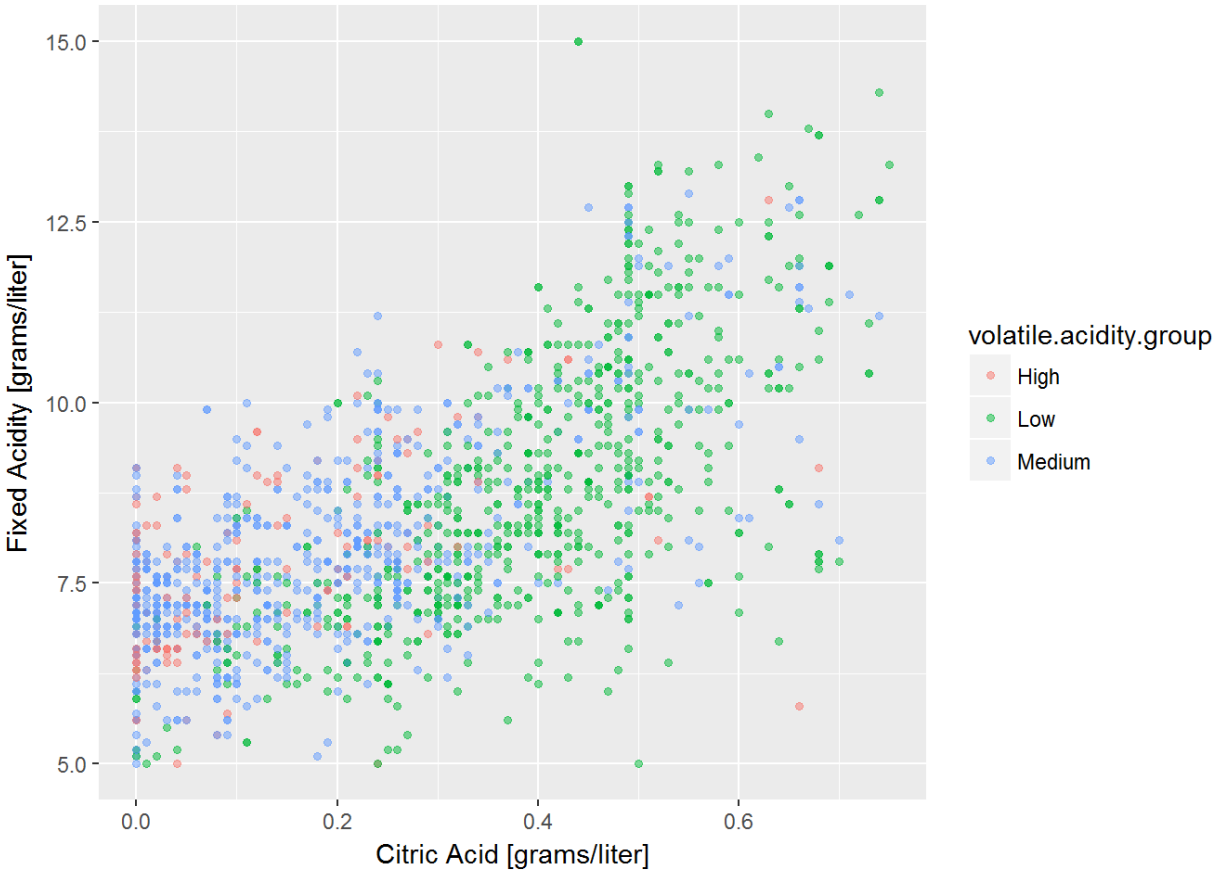
Plot Two



Description Two

Here we see sulphates again, but this time weighed against the strongest negative predictor for quality. Volatile acids can occur when wine doesn't properly ferment and sulphates are added to stabilize the wine. We would expect to see the relationship we see above, with more sulphates meaning better wine AND lower levels of volatile acidity.

Plot Three



Description Three

This time I am looking at the relationship between volatile and other kinds of acidity. There is a very strong relationship in this graph, showing that as citric and fixed acidity levels rise, volatile acids drop significantly. ——

Reflection

There were some interesting patterns to the data. Sulphates and alcohol content corresponded quite closely with quality, which as mentioned above, was a surprise as sulphates are generally seen negatively in the wine world.

Acid levels are also important, with fixed and citric acids being positively correlated with quality, but more weakly and volatile acid negatively correlated but quite strongly. The acids seem to occur in clusters, which would be interesting to explore in a future study.

Similarly, for future analyses it would be interesting to get into the differences between sulfur dioxides and sulphates as they are both sulfur compounds and, while one is closely correlated with quality the other is not and they don't correlate with each other despite sulfur dioxide being a type of sulphate. Whether this is due to sulfur dioxide and sulphates being seperate from each other or because other sulphates tip the scale in the opposite direction is unclear from this data set.

In a similar vein, citric acid is usually considered to be a kind of fixed acid. This raises some questions about those two acids in the data set. This could also explain the large number of zero values in the citric acid variable. A future study could do more to disambiguate these two.

Chlorides appear to be bad for quality, but not tied to any particular variable. A further exploration of that and density, another variable weakly correlated negatively to quality could be warranted.

It would be interesting to explore the role of chlorides more extensively in a future study.

In my analysis, I found it difficult to visualize some of the trends relating multiple variables. I fixed this problem by clustering the variables into different “buckets” and using those as the color factor on the scatterplot.

As mentioned above, there was some ambiguity about the exact meaning and relationship of the variables. Sulfur dioxide does not seem to be related to sulphates- is this because they are categorized seperately in the data set or because other sulphates make a big difference? Citric acid and fixed acid correlate strongly- is this because citric acid is included as a fixed acid or do those variables co-occur in the data for another reason? Does this relate to the high number of zero values in the citric acid column? For now these questions remain unanswered.

For future study it would be nice to have a larger data set to play with, and perhaps include more variables for analysis, such as price, location of the vineyard, age of the wine etc. There have been interesting studies showing that even experts can rate a cheap wine more highly if it is put in a fancy bottle.

Overall the data suggests that the optimal wine had low volatile acids, low chlorides, higher levels of sulphates and alcohol as well as more citric and fixed acids.