White Wine Quality Exploration

by Sarah Aranda

The dataset I'll be looking at is White Wine Quality. It's composed of 4,898 observations with 12 variables, including quality which ranges from 0 to 10 (very bad to excellent). The other variables in the dataset are various chemical properties of the wine. A more detailed description of the dataset can be found here: https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityInfo.txt (https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityInfo.txt)

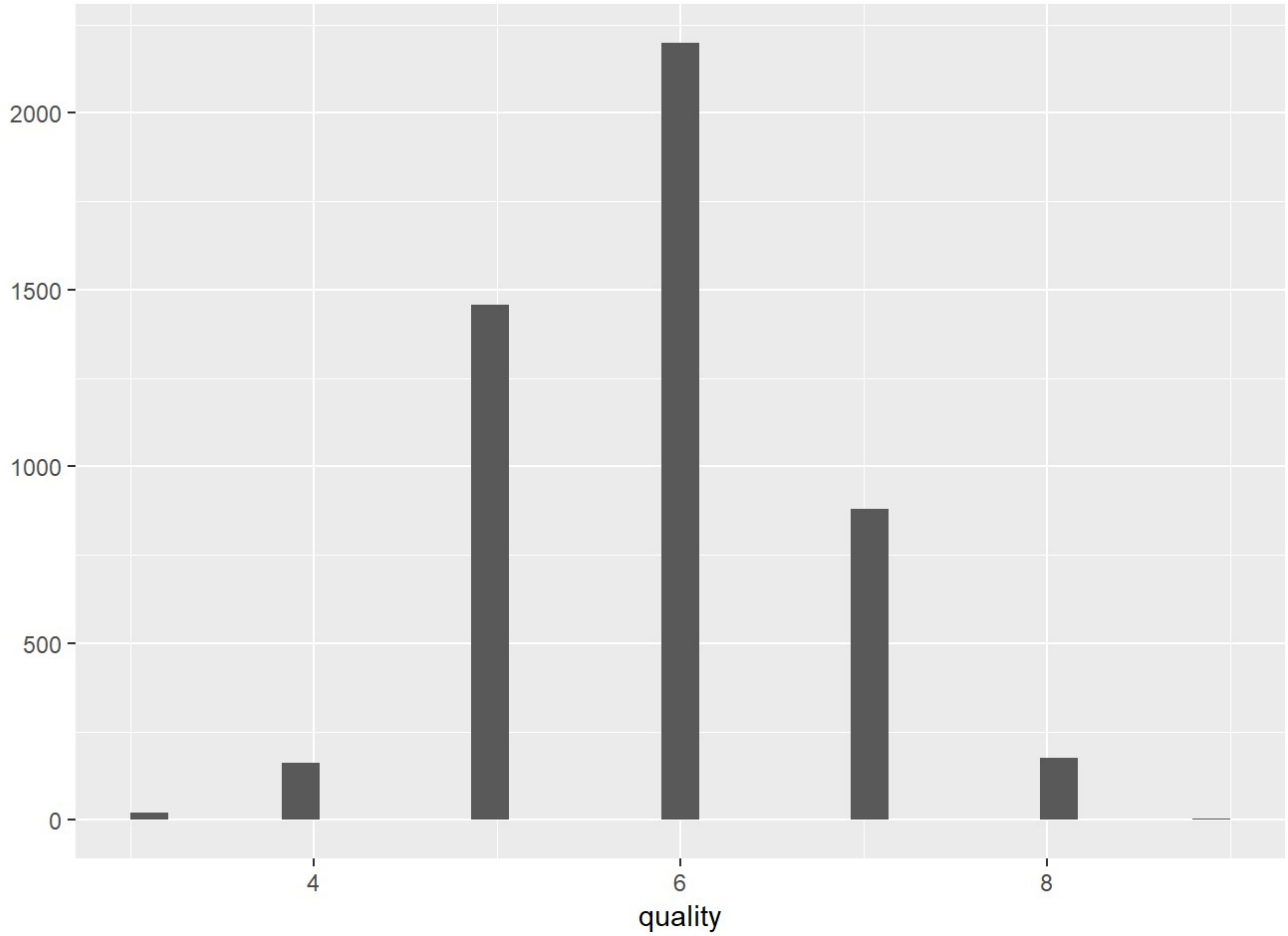
Univariate Plots

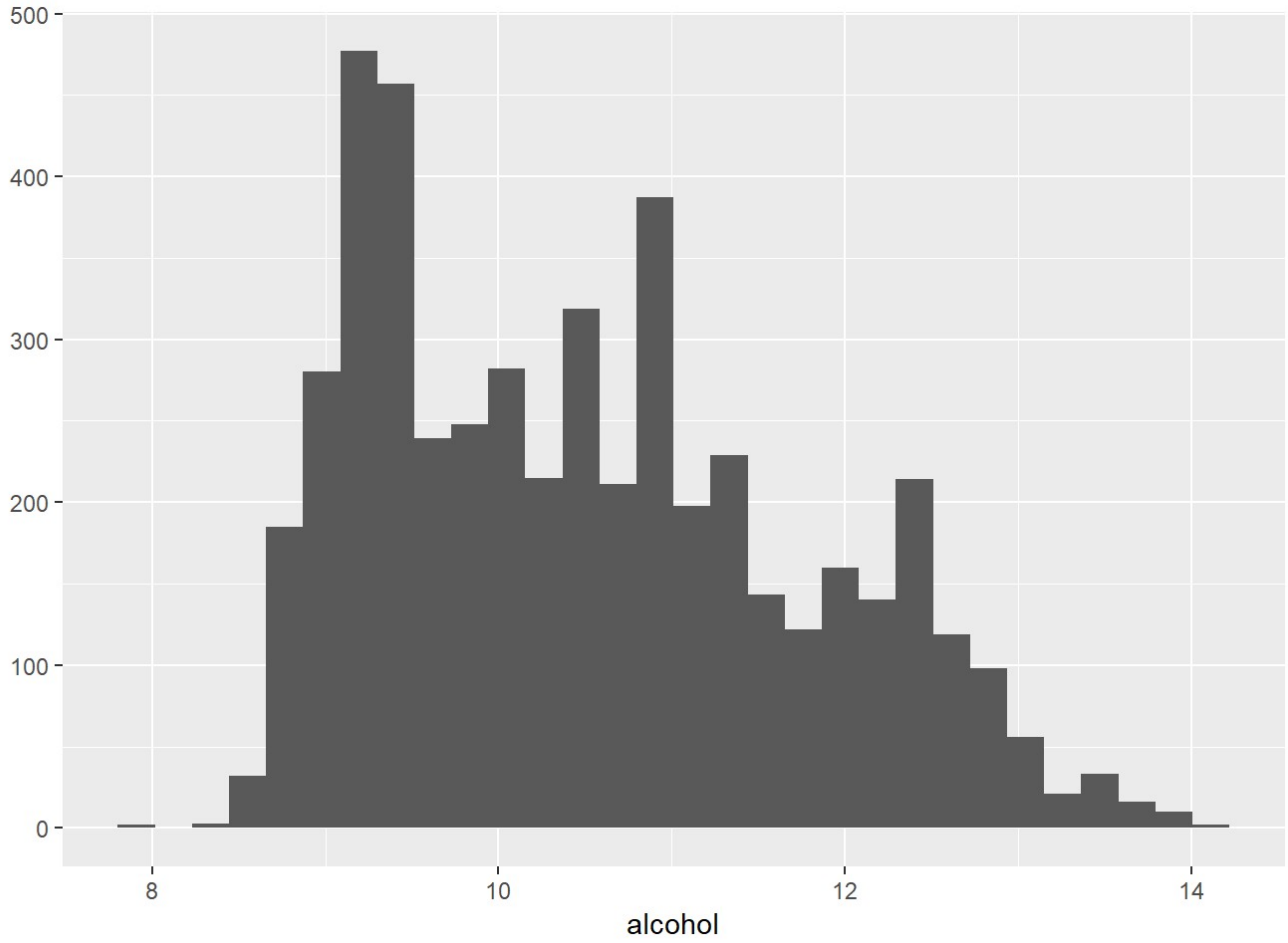
I'll start by getting an idea of the dataset–the variable names, the datatypes, etc.

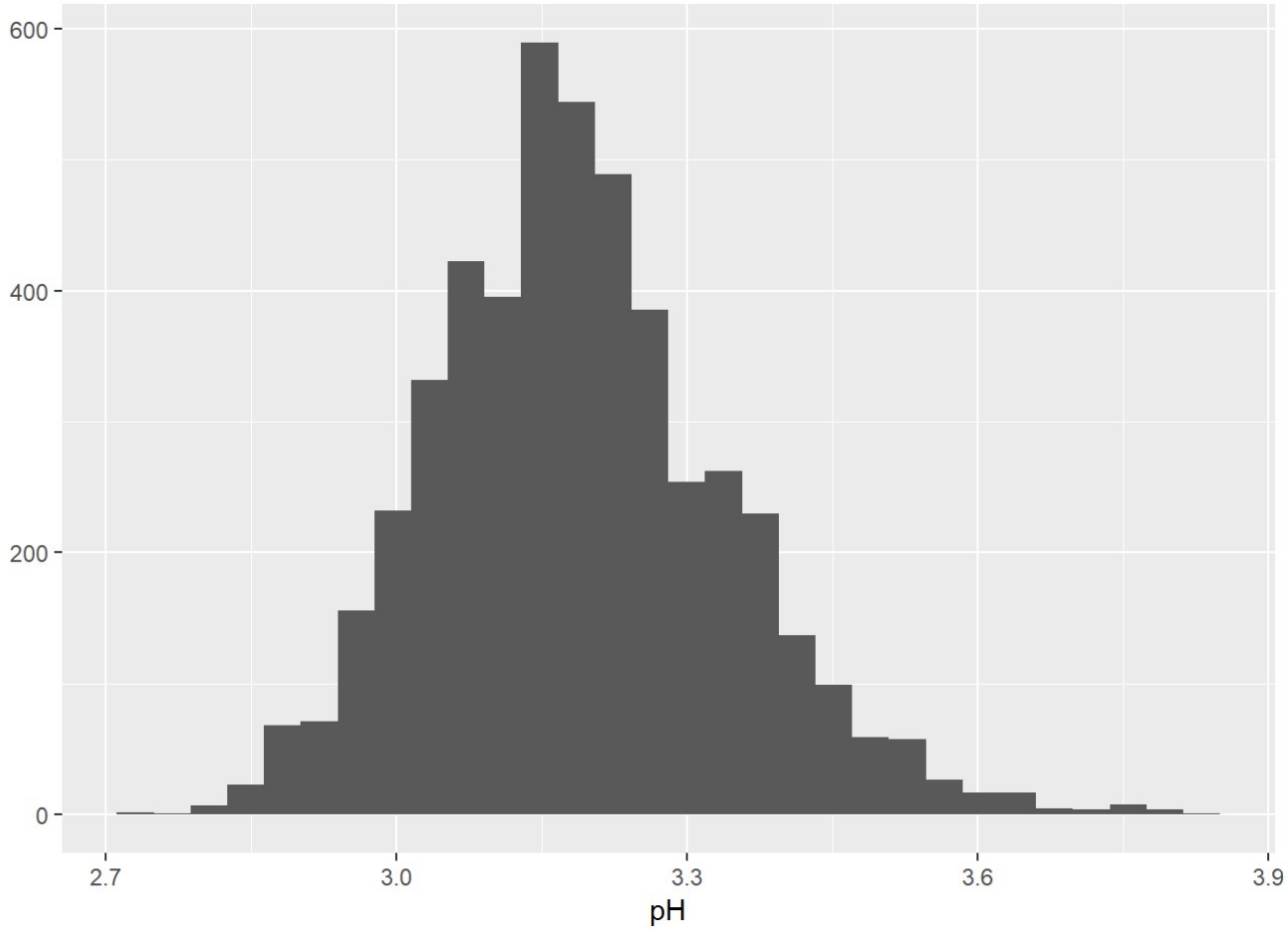
```
## 'data.frame':    4898 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  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
##  Min.   : 0.600   Min.   :0.00900   Min.   :  2.00
##  1st Qu.: 1.700   1st Qu.:0.03600   1st Qu.: 23.00
##  Median : 5.200   Median :0.04300   Median : 34.00
##  Mean   : 6.391   Mean   :0.04577   Mean   : 35.31
##  3rd Qu.: 9.900   3rd Qu.:0.05000   3rd Qu.: 46.00
##  Max.   :65.800   Max.   :0.34600   Max.   :289.00
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  9.0        Min.   :0.9871   Min.   :2.720   Min.   :0.2200
##  1st Qu.:108.0        1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100
##  Median :134.0        Median :0.9937   Median :3.180   Median :0.4700
##  Mean   :138.4        Mean   :0.9940   Mean   :3.188   Mean   :0.4898
##  3rd Qu.:167.0        3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500
##  Max.   :440.0        Max.   :1.0390   Max.   :3.820   Max.   :1.0800
##     alcohol         quality
##  Min.   : 8.00   Min.   :3.000
##  1st Qu.: 9.50   1st Qu.:5.000
##  Median :10.40   Median :6.000
##  Mean   :10.51   Mean   :5.878
##  3rd Qu.:11.40   3rd Qu.:6.000
##  Max.   :14.20   Max.   :9.000
```

Of the variables in the dataset, quality, alcohol, and pH are perhaps the most easily identifiable. The histograms below take a look at their counts.


2000
1500
1000
500
0
4
6
8
quality


500
400
300
200
100
0
8
10
12
14
alcohol

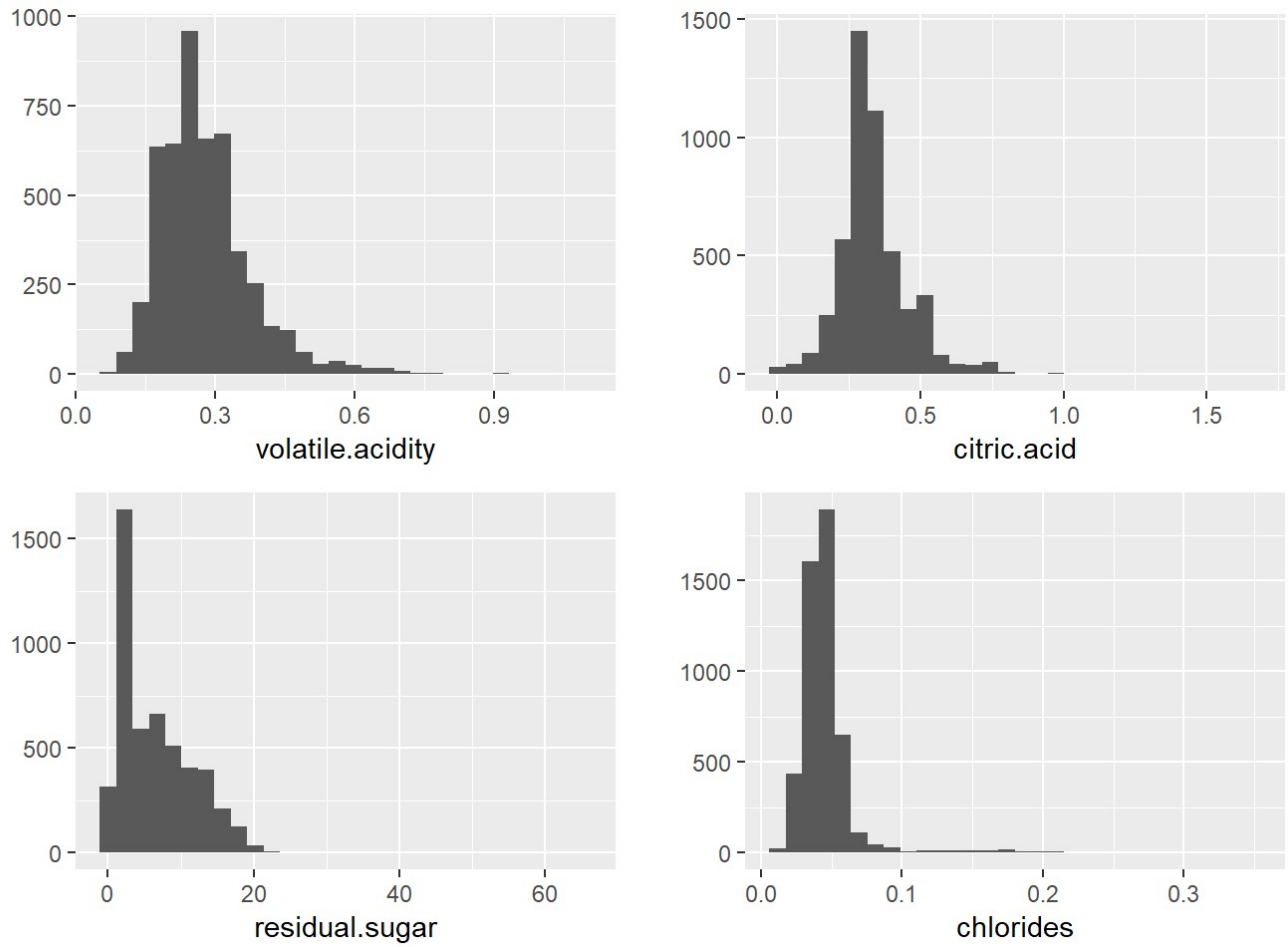


From the plots above, we can see that a quality of 6 is the most common, the most common percent alcohol content is approximately 9-9.5 percent, and that the counts for pH peak between 3.0 and 3.3.

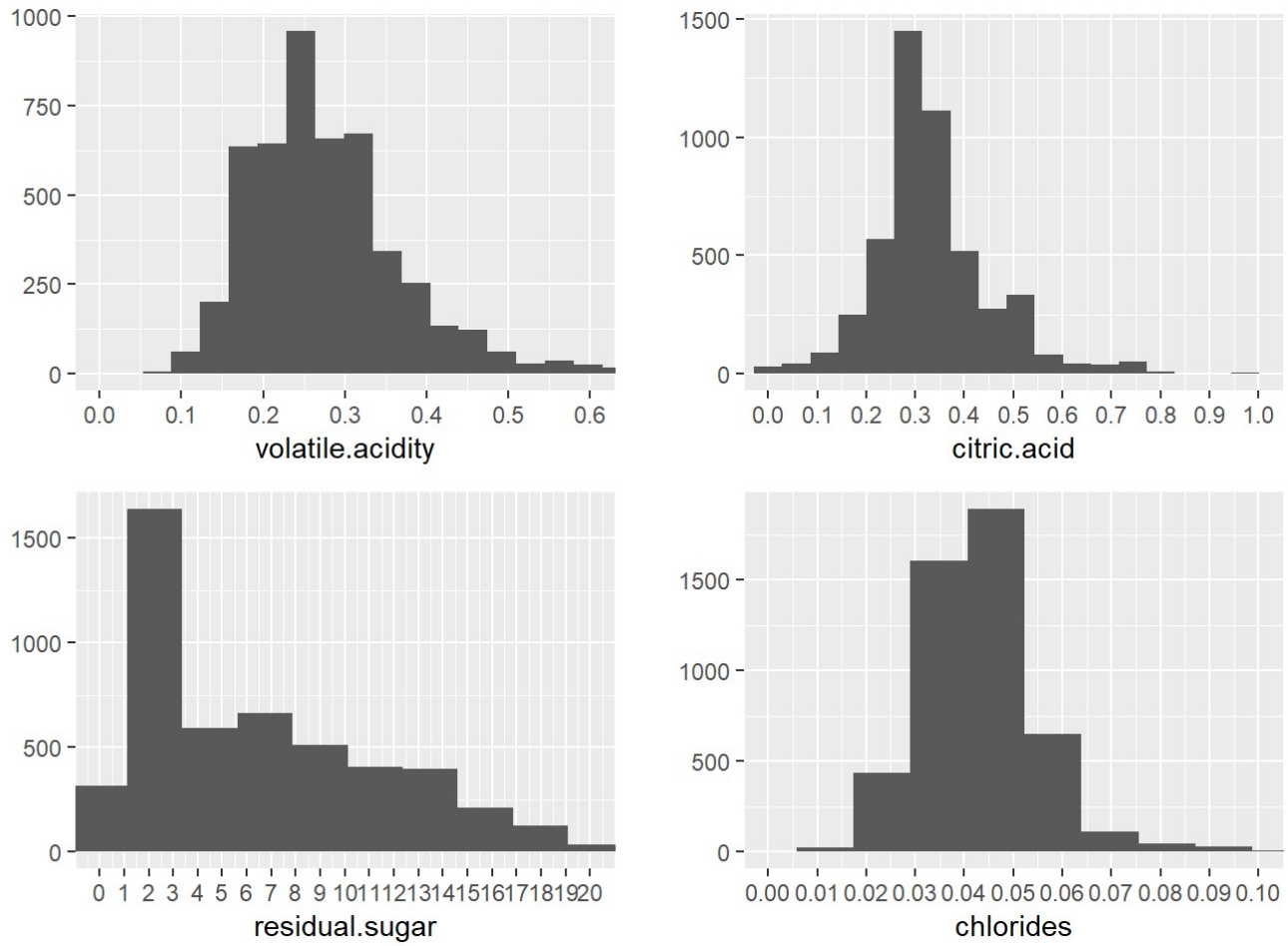
A few other variables which may be interesting to look at and which impact the taste or flavor of the wine include:

1. volatile acidity - too high levels can lead to an unpleasant vinegar taste
2. citric acid - adds ‘freshness’ to the wine’s flavor
3. residual sugar - wines with greater than 45 grams/liter of residual sugar are considered sweet
4. chlorides - the amount of salt in the wine

Let’s take a look at those counts below:

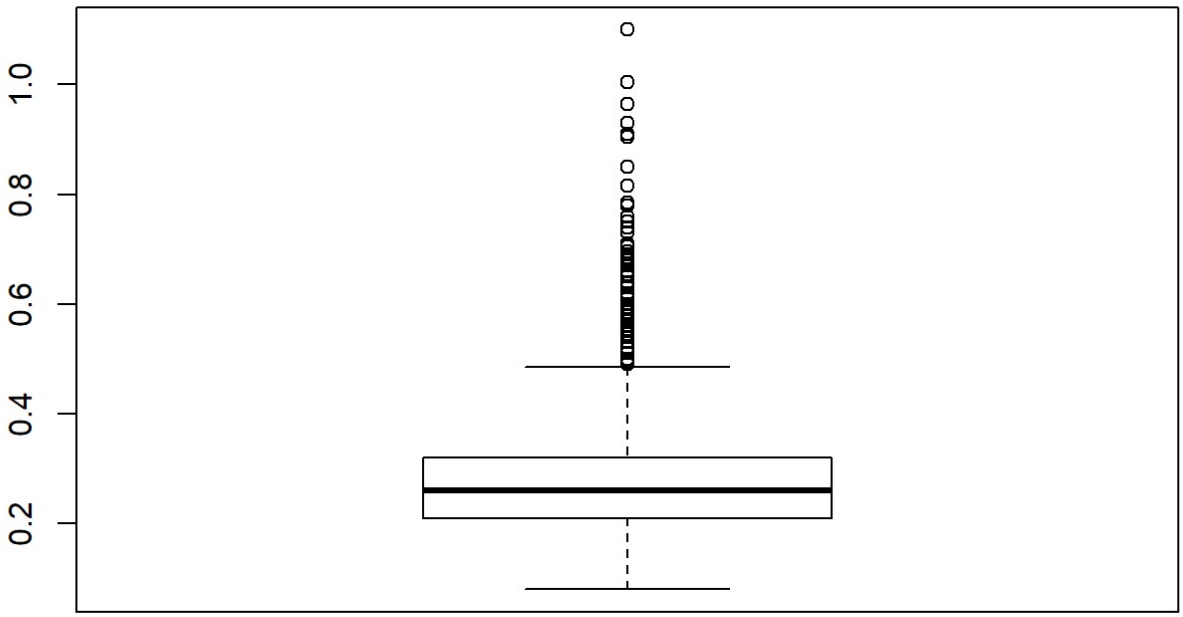


The histograms above are right-skewed. Let’s zoom in on them and focus on the peaks.

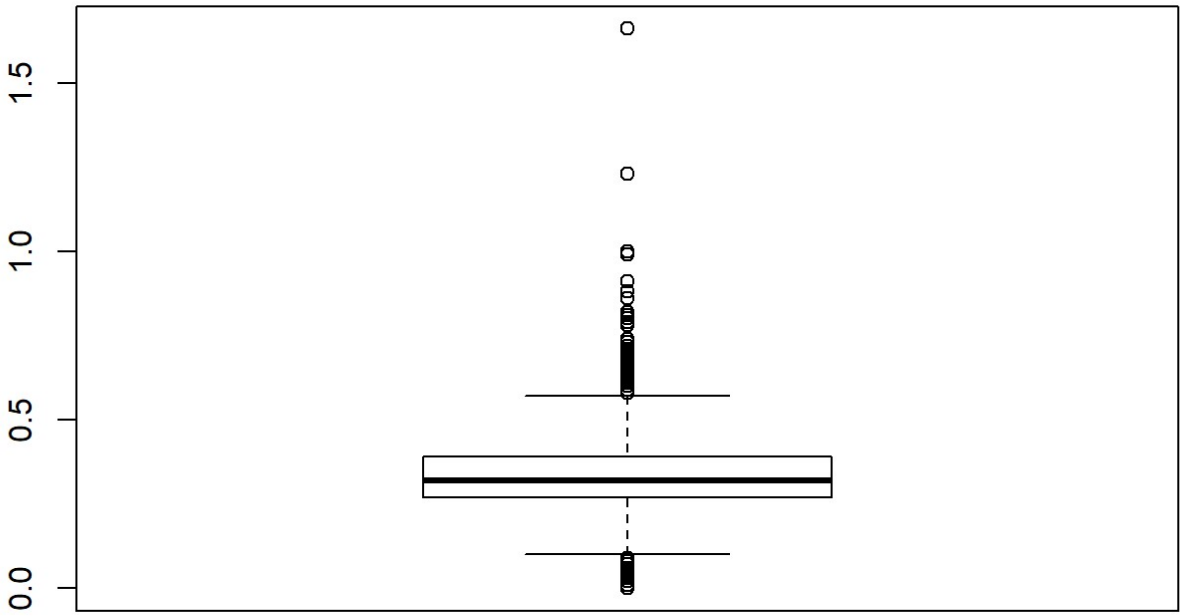


The counts for volatile acidity sharply peaks at approximiately 0.25. For citric acid, it's about 0.3. For residual sugar, it's between 1 and 3.5. And for chlorides, it's between 0.03 and 0.05.

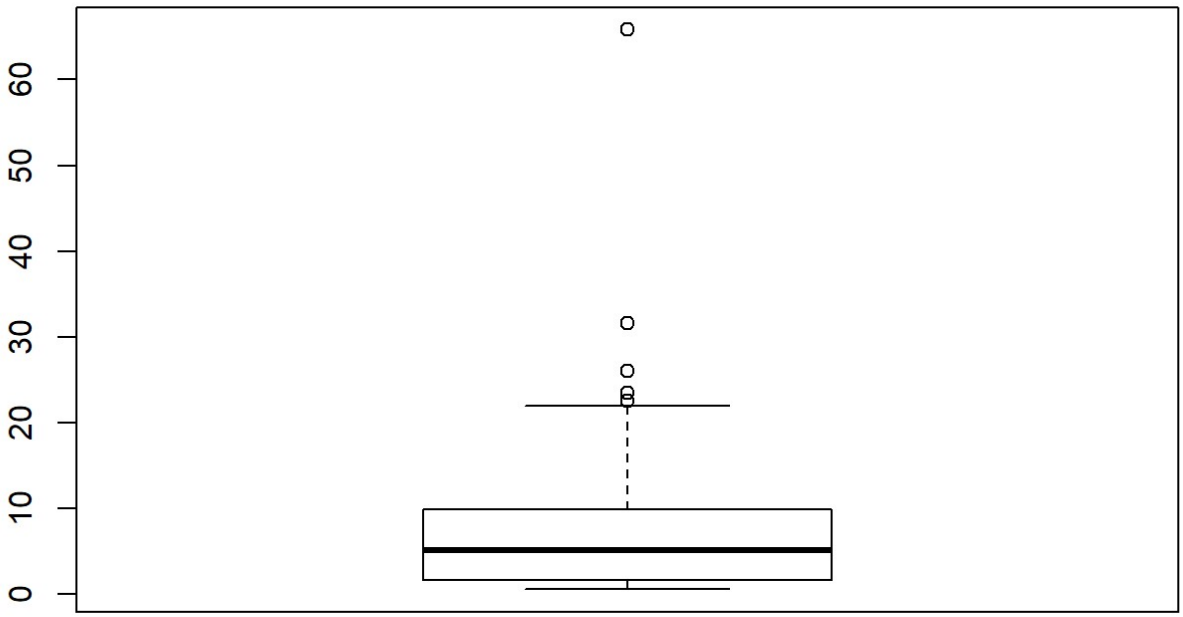
In creating boxplots for them (below), we can see that there are many outliers for these variables.



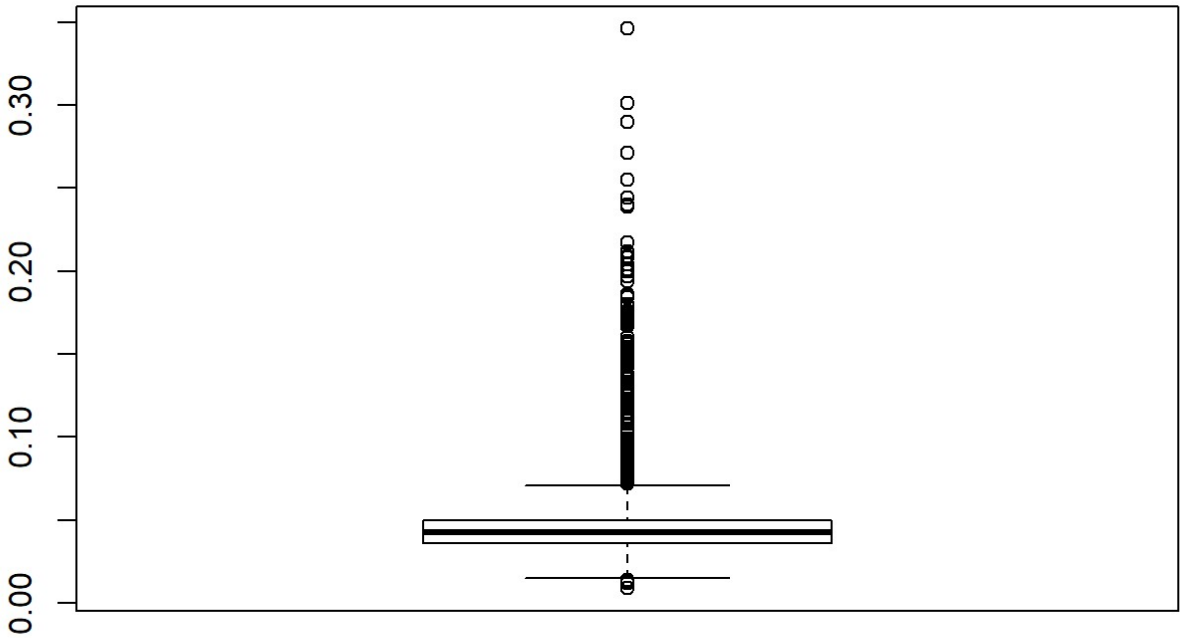
volatile acidity



citric acid



residual sugar



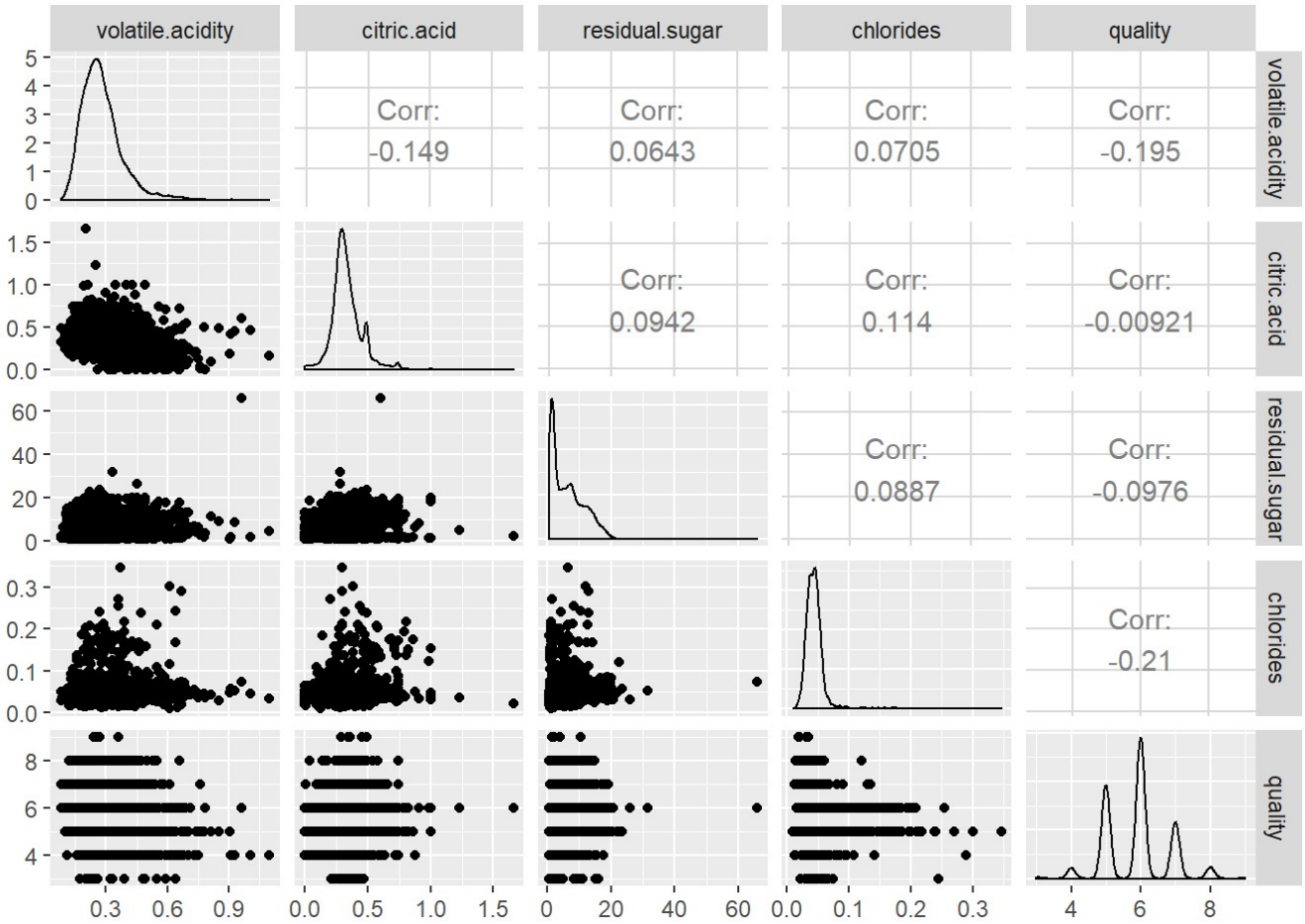
chlorides

Through the boxplots, it appears that the values for variables related to taste and flavor are widely spread and that the interquartile range is relatively narrow in comparison. This is especially true for chlorides.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00900 0.03600 0.04300 0.04577 0.05000 0.34600
```

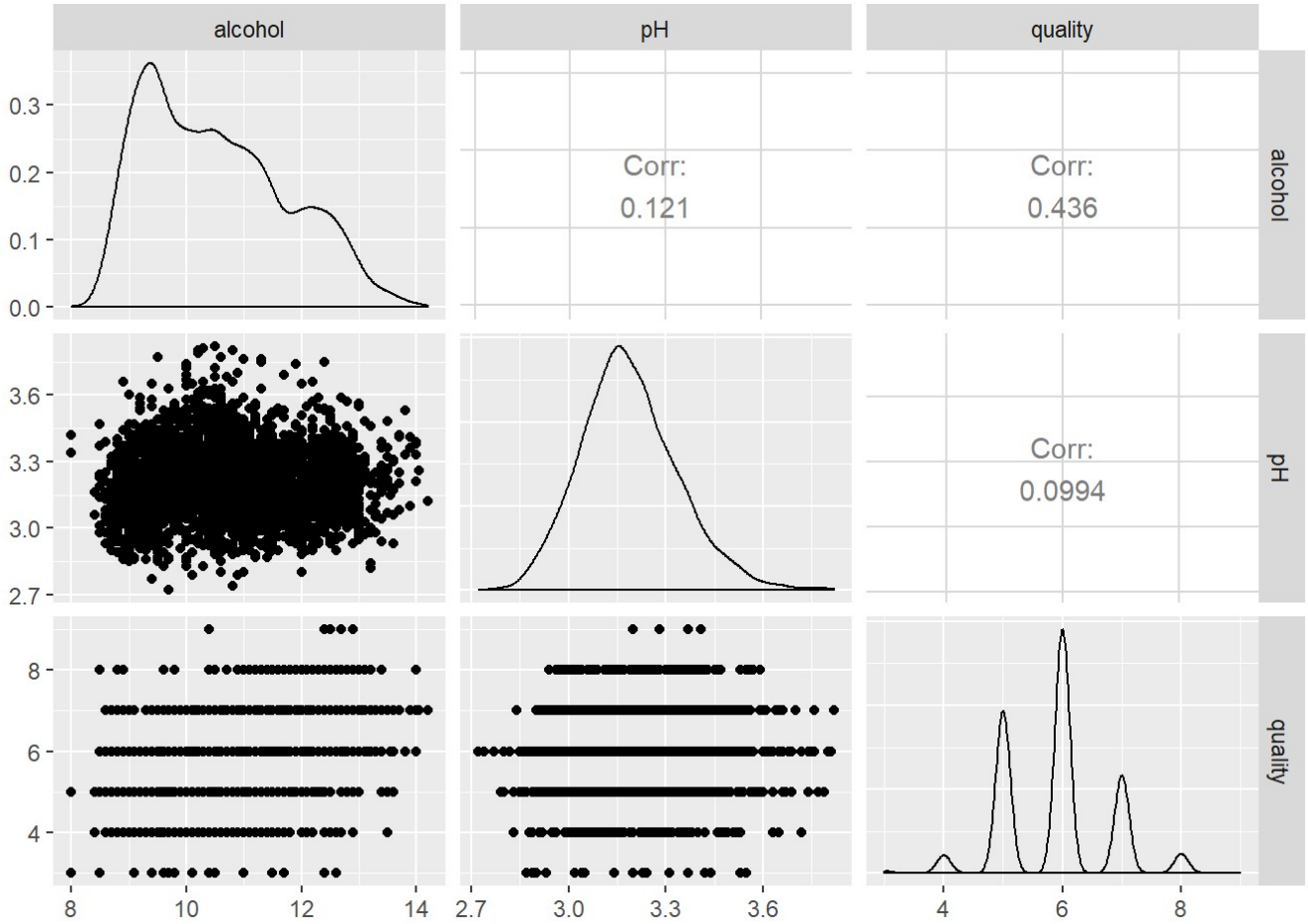
Bivariate Plots

In this section, I'll be looking at bivariate plots. The final set of plots in the previous section were looking specifically at variables that impact the flavor and taste of the wine which I assumed would also impact wine quality. Below is a scatterplot correlation matrix of those variables and also quality.



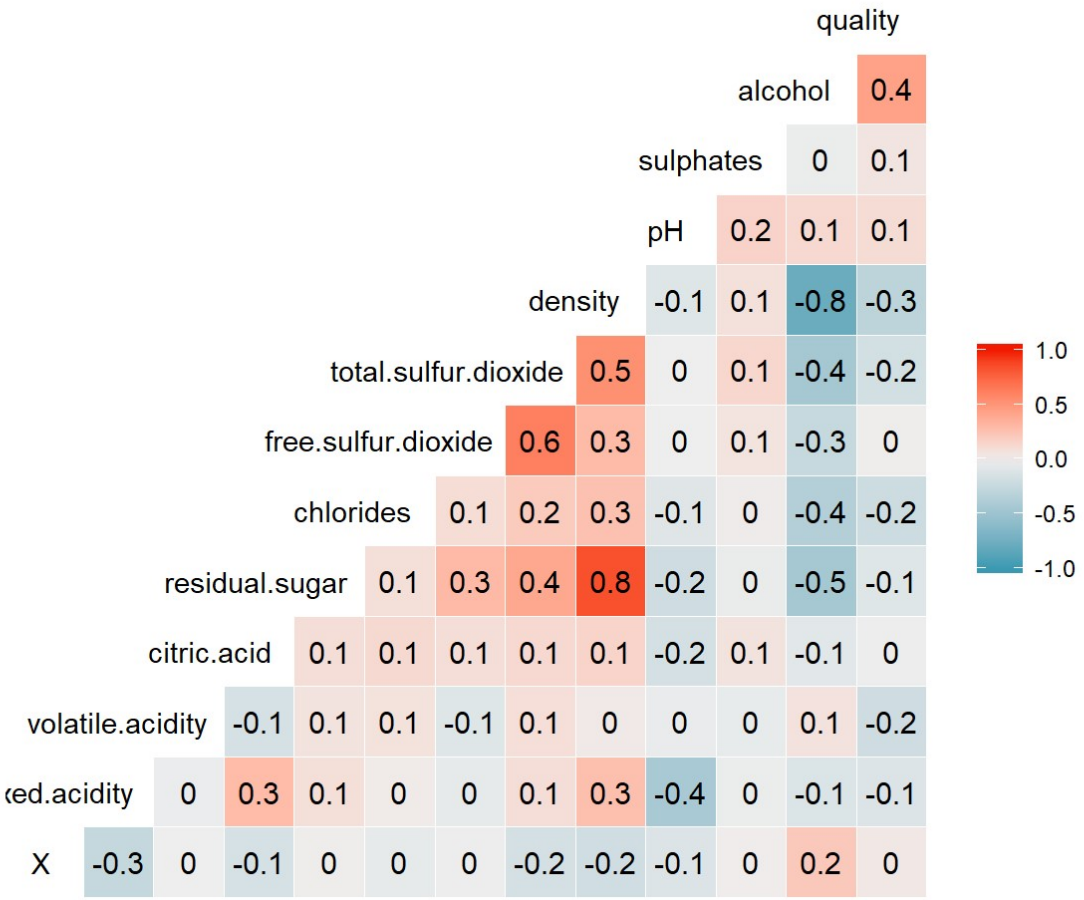
Above, the strongest correlation is a negative one between quality and chlorides which is only a -0.21. In fact, surprisingly, the variables all have weak correlations with quality.

In the univariate section, I had also looked closely at alcohol and pH. A scatterplot correlation matrix is below.



Most of the correlations are weak but there is a moderate correlation between quality and alcohol at 0.43.

To check for correlations across all variables, let's try the correlation matrix below.



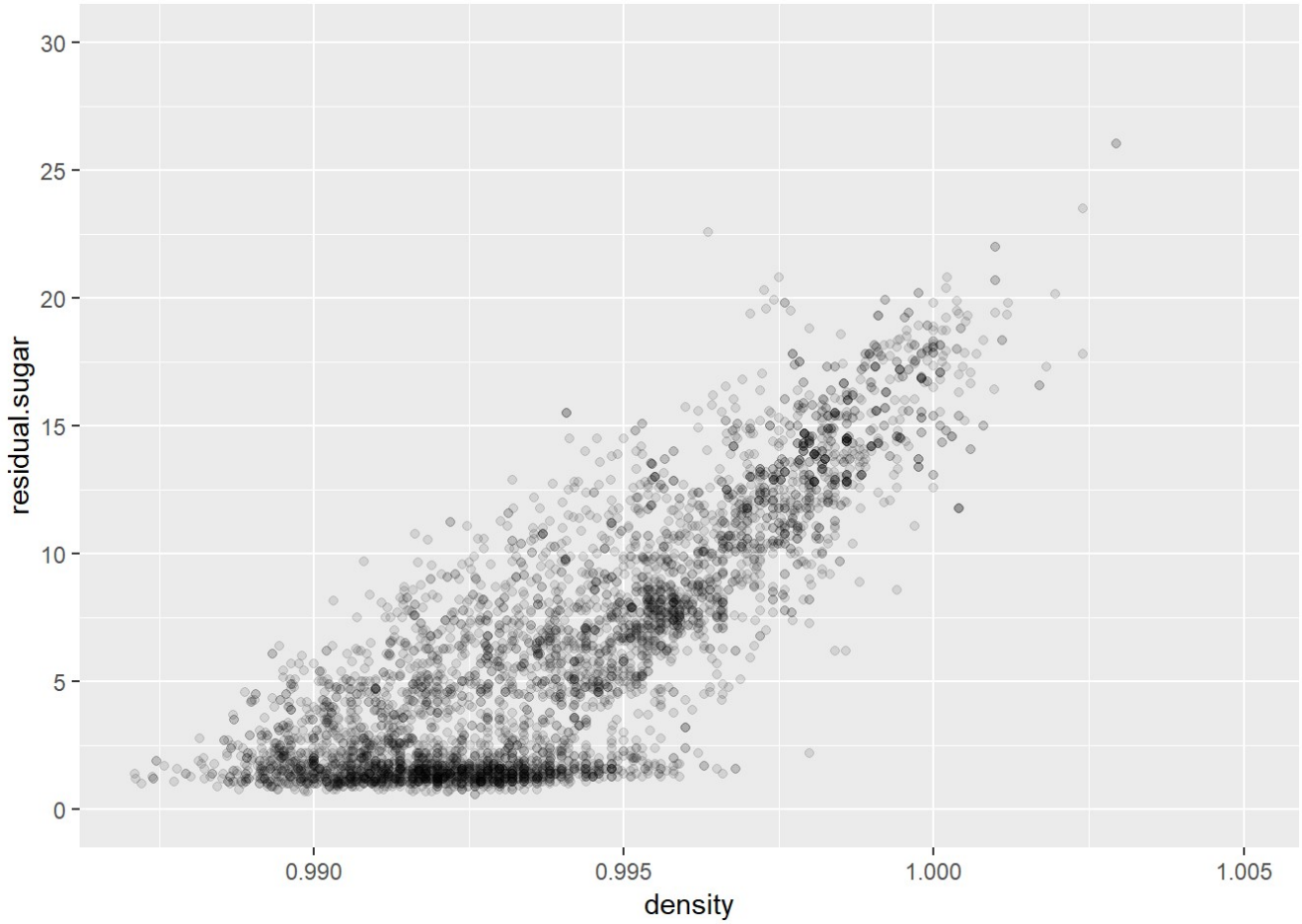
The strongest correlation between quality and another variable is alcohol with a 0.4 (which was discussed earlier). As far as other correlations, the three strongest positive correlations are between residual sugar and density (0.8), free sulfur dioxide and total sulfur dioxide (0.6), and total sulfur dioxide and density (0.5).

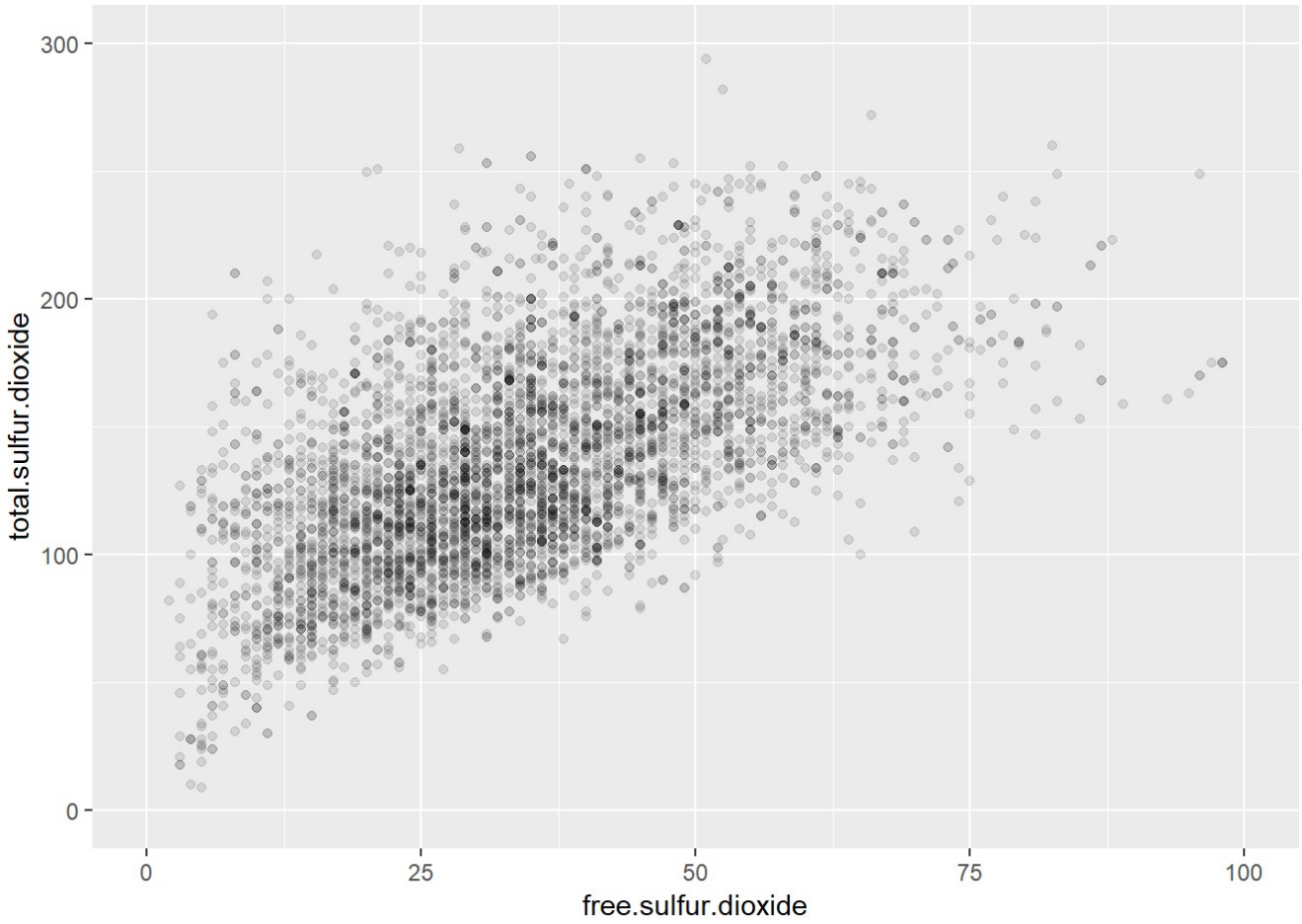
According to the description of the data set, free sulfur dioxide prevents microbial growth and the oxidation of wine. A certain level of oxidation is good for wine but too much leads to its degradation (https://vinepair.com /articles/what-is-oxidation-in-wine/ (https://vinepair.com/articles/what-is-oxidation-in-wine/)). As for the total sulfur dioxide, this measures the free and bound forms of SO2. SO2 is mostly undectable but free SO2 concentrations over 50ppm make it more discernable in the nose and taste of wine.

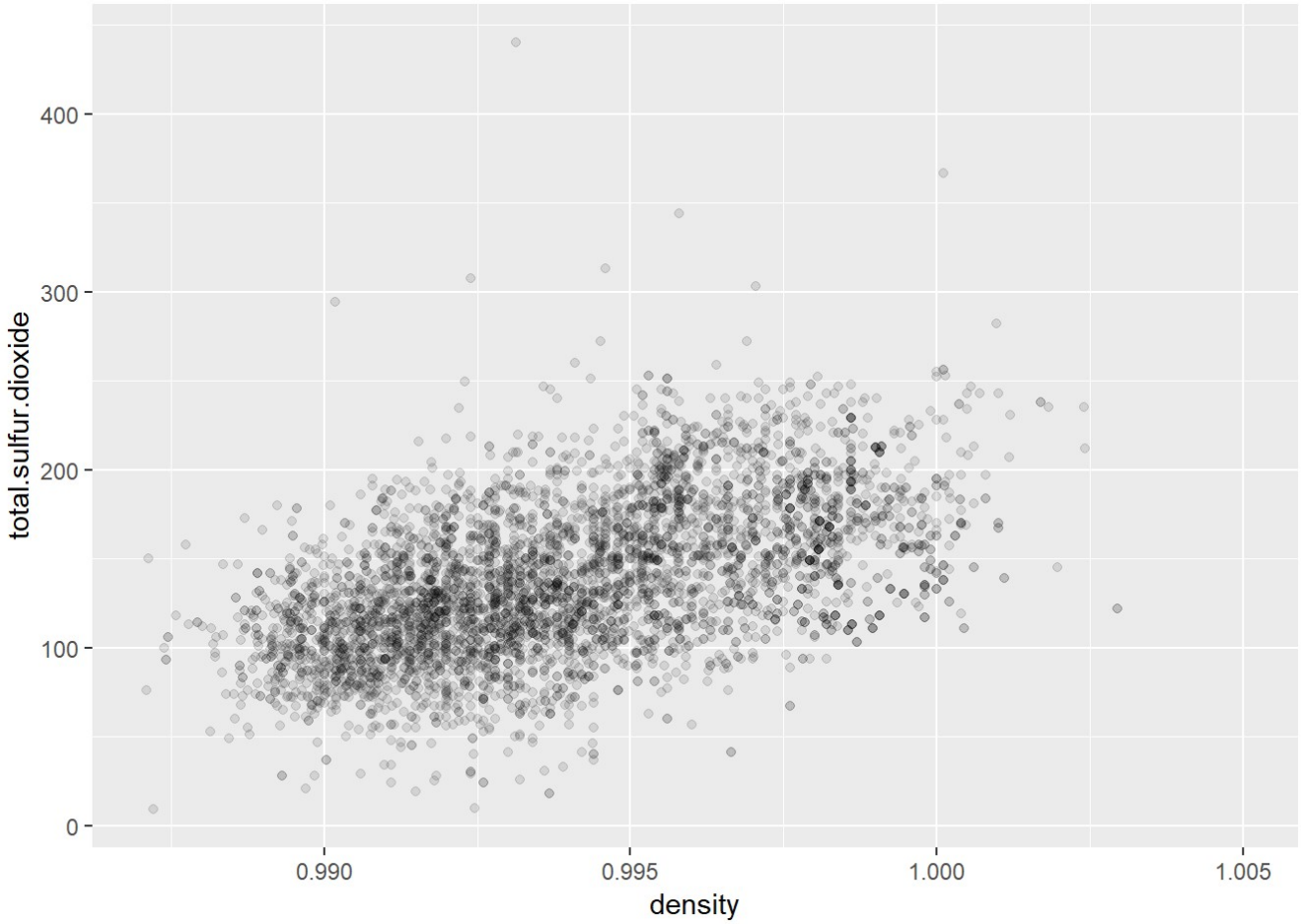
The two strongest negative correlations are between density and alcohol (-0.8) and alcohol and residual sugar (-0.5).

Let's look first at the three strongest positive correlations:

- density and residual sugar
- free sulfur dioxide and total sulfur dioxide
- density and total sulfur dioxide


residual.sugar
30
25
20
15
10
5
0
0.990
0.995
1.000
1.005
density

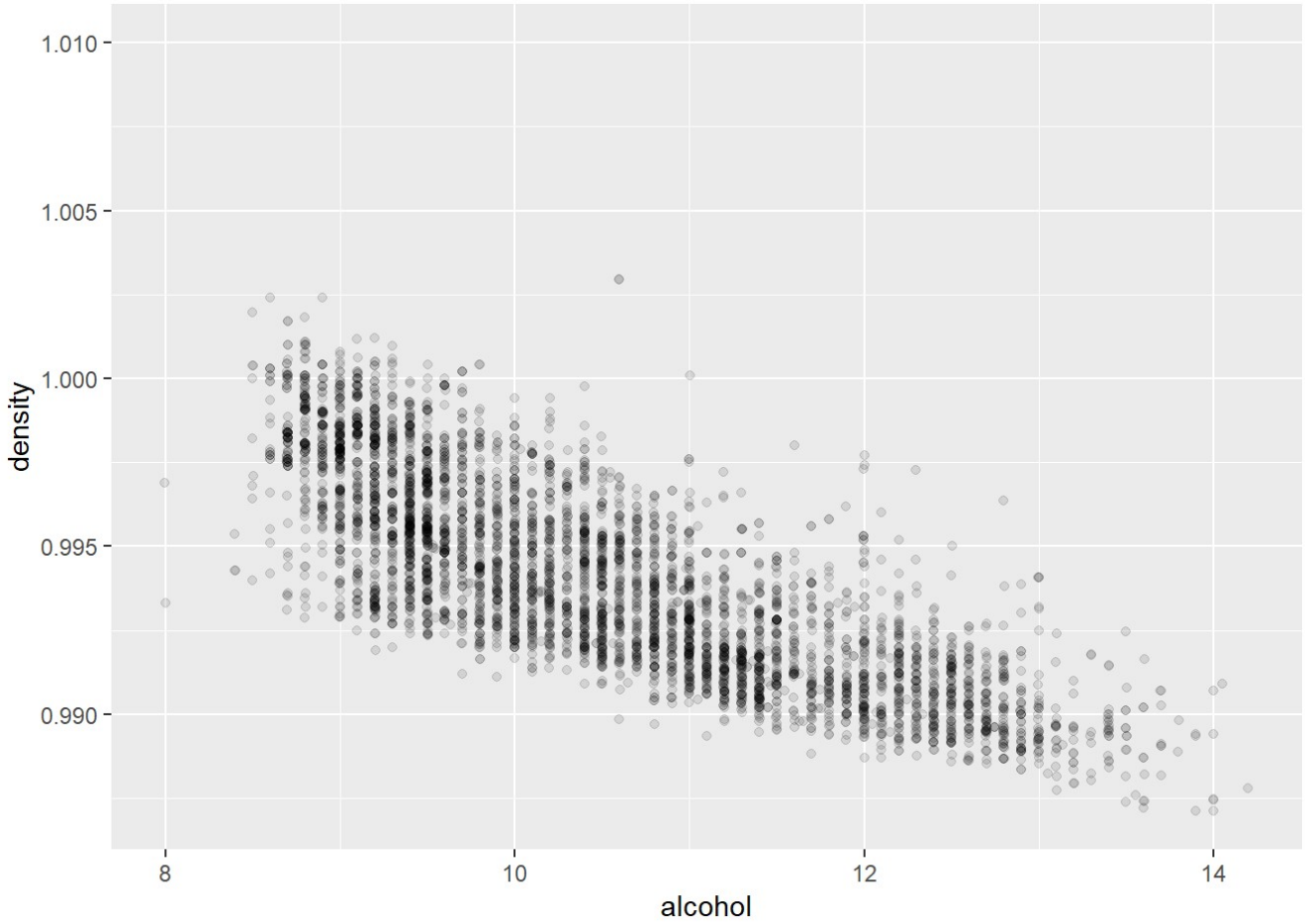

total.sulfur.dioxide
300
200
100
0
0
25
50
75
100
free.sulfur.dioxide

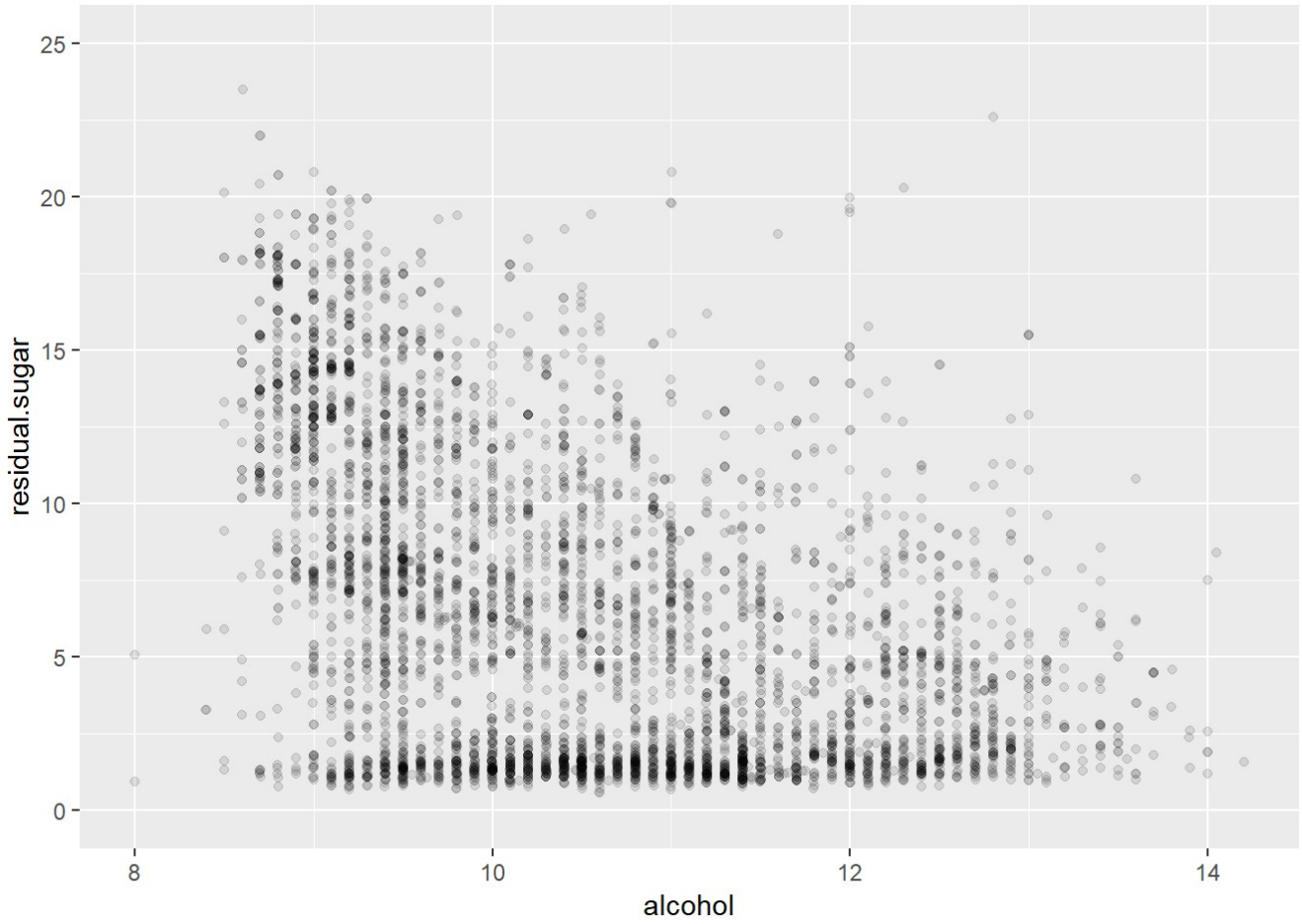


The three plots above confirm what we observed in our correlation matrix. In the density and residual sugar plot, we can also observe a darker horizontal band toward the bottom of the plot where a lot of the observations lie between approximately 1-2.5 residual sugar and a density of between 0.9875 and 0.995.

Now, let's look at the two strongest negative correlations.

- alcohol and density
- alcohol and residual sugar


1.010
1.005
1.000
0.995
0.990
density
8
10
12
14
alcohol


25
20
15
10
5
0
residual.sugar
8
10
12
14
alcohol

The scatterplots confirm the observations from the correlation matrix. Of the two, the alcohol and density plot has a stronger negative correlation. In the alcohol and residual sugar plot, there seems to be a high concentration of observations below a residual sugar of 2.5 and then what looks a little bit like a horizontal break or gap in the plot where few observations fall (approximately where residual sugar is 2.5).

Additionally, the scatterplots show that the dots are arranged almost into columns which suggests that the values in alcohol may be somewhat discrete. We can use the table function to verify.

```
##
##                8              8.4              8.5              8.6
##                2                3                9               23
##              8.7              8.8              8.9                9
##               78              107               95              185
##              9.1              9.2              9.3              9.4
##              144              199              134              229
##              9.5 9.53333333333333             9.55              9.6
##              228                3                2              128
## 9.63333333333333              9.7 9.73333333333333             9.75
##                1              105                2                1
##              9.8              9.9               10 10.0333333333333
##              136              109              162                1
##             10.1 10.1333333333333            10.15             10.2
##              114                2                3              130
##             10.3             10.4 10.4666666666667             10.5
##               85              153                2              160
## 10.5333333333333            10.55 10.5666666666667             10.6
##                1                2                1              114
##            10.65             10.7             10.8             10.9
##                1               96              135               88
## 10.9333333333333 10.9666666666667            10.98               11
##                2                3                1              158
##            11.05 11.0666666666667             11.1             11.2
##                2                1               83              112
## 11.2666666666667             11.3 11.3333333333333            11.35
##                1              101                3                1
## 11.3666666666667             11.4 11.4333333333333            11.45
##                1              121                1                4
## 11.4666666666667             11.5            11.55             11.6
##                1               88                1               46
## 11.6333333333333            11.65             11.7 11.7333333333333
##                2                1               58                1
##            11.75             11.8            11.85             11.9
##                2               60                1               53
##            11.94            11.95               12            12.05
##                2                1              102                1
## 12.0666666666667             12.1            12.15             12.2
##                1               51                2               86
##            12.25             12.3 12.3333333333333             12.4
##                1               62                1               68
##             12.5             12.6             12.7            12.75
##               83               63               56                3
##             12.8 12.8933333333333             12.9               13
##               54                2               39               36
##            13.05             13.1 13.1333333333333             13.2
##                1               18                1               14
##             13.3             13.4             13.5            13.55
##                7               20               12                1
##             13.6             13.7             13.8             13.9
```

```
##                9                7                2                3
##               14            14.05             14.2
##                5                1                1
```

Though there are some odd values, we see that the values for alcohol are relatively discrete.

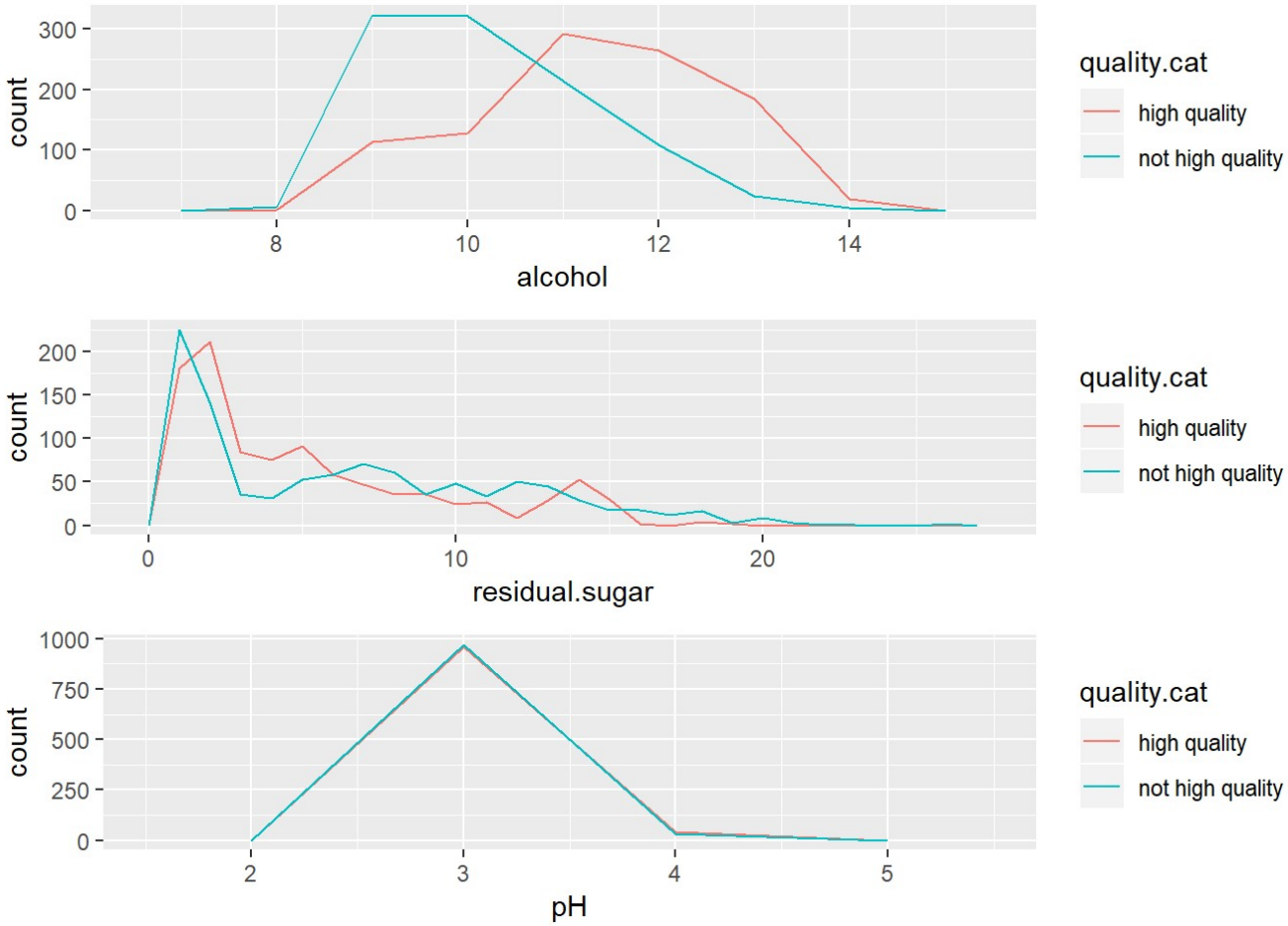
In the beginning of this analysis the first variables I was most interested in looking at were alcohol, pH, and residual sugar. I'm curious whether higher quality wines would more likely fall within a certain range of these variables.

To carry out this analysis, I categorized the wines into high quality (greater than or equal to 7) or not high quality (less than 7) by introducing a new variable (quality.cat) into the dataset. To make the counts even for both categories, I took a random sample without replacement for both categories (1000 each) and created a new dataset with them.

Frequency polygons for the aforementioned variables and from my sample dataset are below.

```
##         X fixed.acidity volatile.acidity citric.acid residual.sugar
## 3100 3100           6.7             0.34        0.40            2.1
## 856   856           7.4             0.20        0.35            2.1
## 1009 1009           6.6             0.22        0.37            1.2
## 3288 3288           6.8             0.21        0.31            2.9
## 53     53           6.2             0.16        0.33            1.1
## 3645 3645           6.4             0.24        0.23            2.0
##      chlorides free.sulfur.dioxide total.sulfur.dioxide density   pH
## 3100     0.033                  34                  111 0.98924 2.97
## 856      0.038                  30                  116 0.99490 3.49
## 1009     0.059                  45                  199 0.99300 3.37
## 3288     0.046                  40                  121 0.99130 3.07
## 53       0.057                  21                   82 0.99100 3.32
## 3645     0.046                  30                  133 0.99080 3.12
##      sulphates alcohol quality  quality.cat
## 3100      0.48    12.2       7 high quality
## 856       0.77    10.3       7 high quality
## 1009      0.55    10.3       7 high quality
## 3288      0.65    10.9       7 high quality
## 53        0.46    10.9       7 high quality
## 3645      0.54    11.4       7 high quality
```

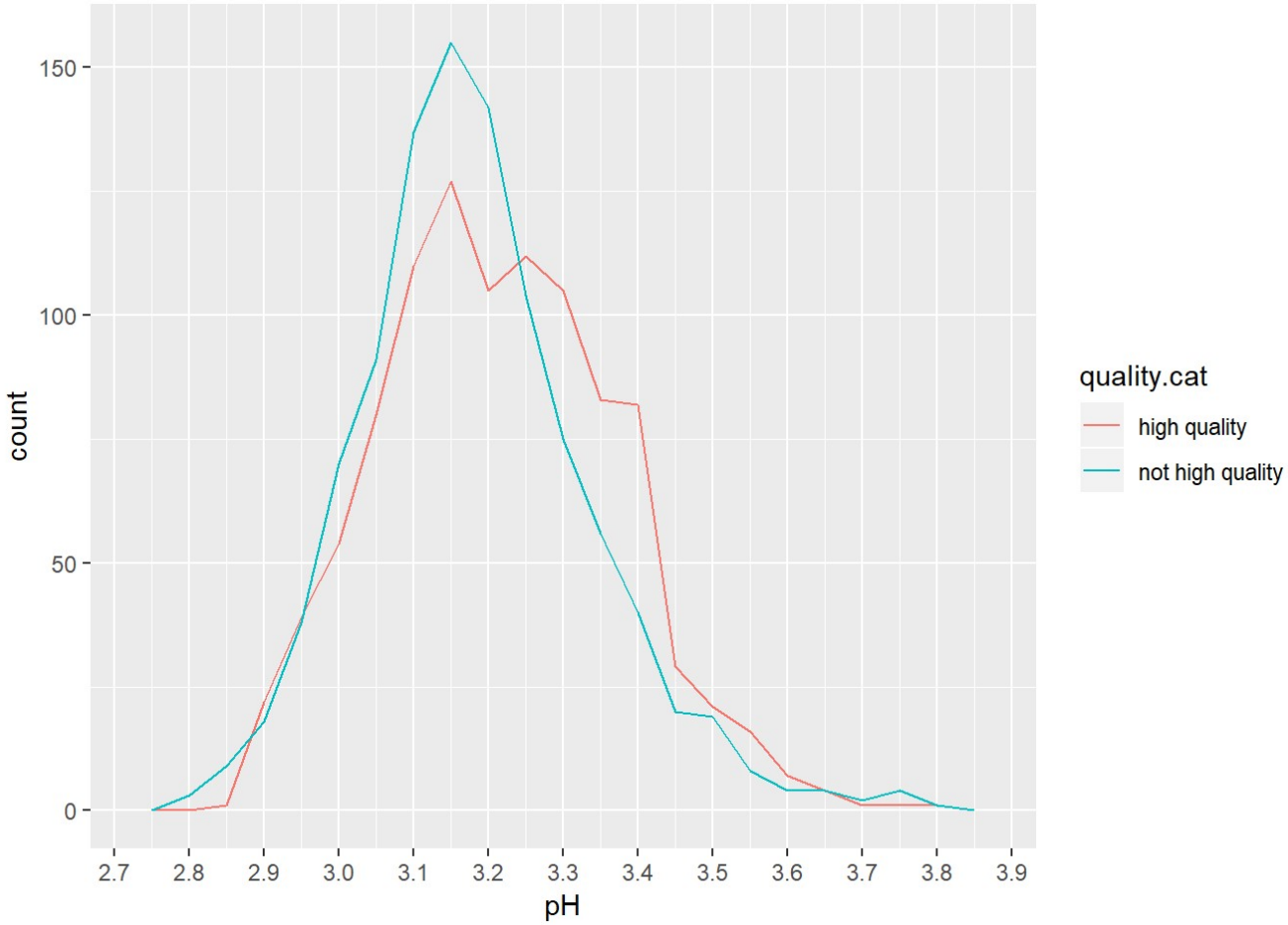
```
##         X fixed.acidity volatile.acidity citric.acid residual.sugar
## 3222 3222           6.6             0.24        0.38          12.75
## 2563 2563           6.9             0.32        0.26           2.30
## 3181 3181           6.5             0.24        0.38           1.00
## 2030 2030           7.6             0.34        0.39           7.60
## 4781 4781           5.8             0.30        0.09           6.30
## 3515 3515           7.7             0.38        0.23          10.80
##      chlorides free.sulfur.dioxide total.sulfur.dioxide density   pH
## 3222     0.034                   8                   74 0.99386 3.10
## 2563     0.030                  11                  103 0.99106 3.06
## 3181     0.027                  31                   90 0.98926 3.24
## 2030     0.040                  45                  215 0.99650 3.11
## 4781     0.042                  36                  138 0.99382 3.15
## 3515     0.030                  28                   95 0.99164 2.93
##      sulphates alcohol quality      quality.cat
## 3222      0.57    12.9       6 not high quality
## 2563      0.42    11.1       6 not high quality
## 3181      0.36    12.3       6 not high quality
## 2030      0.53     9.2       6 not high quality
## 4781      0.48     9.7       5 not high quality
## 3515      0.41    13.6       6 not high quality
```



For the plots above, both the high quality and not high quality wines have about the same counts for residual sugar and look to overlap perfectly with pH. These suggest that these factors don't help to differentiate between

high quality and low quality wines. However, for alcohol, we can see a significant difference in counts and where they peak. For high quality wines, the counts peak about 10.5 to 13 alcohol and those of lower quality peak at about 8.5 to 10.5. This corroborates our previous finding that alcohol and quality are correlated.

Let's take a closer look at the pH and see whether there really is a perfect overlap between the two wine qualities.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.840   3.100   3.200   3.214   3.320   3.820
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.790   3.080   3.170   3.181   3.260   3.800
```

We can see from this zoomed-in frequency polygon that the counts for pH for higher and lower quality wines don't overlap perfectly The peak for higher quality wines is shifted slightly more to the right (more basic) than the lower quality wines.

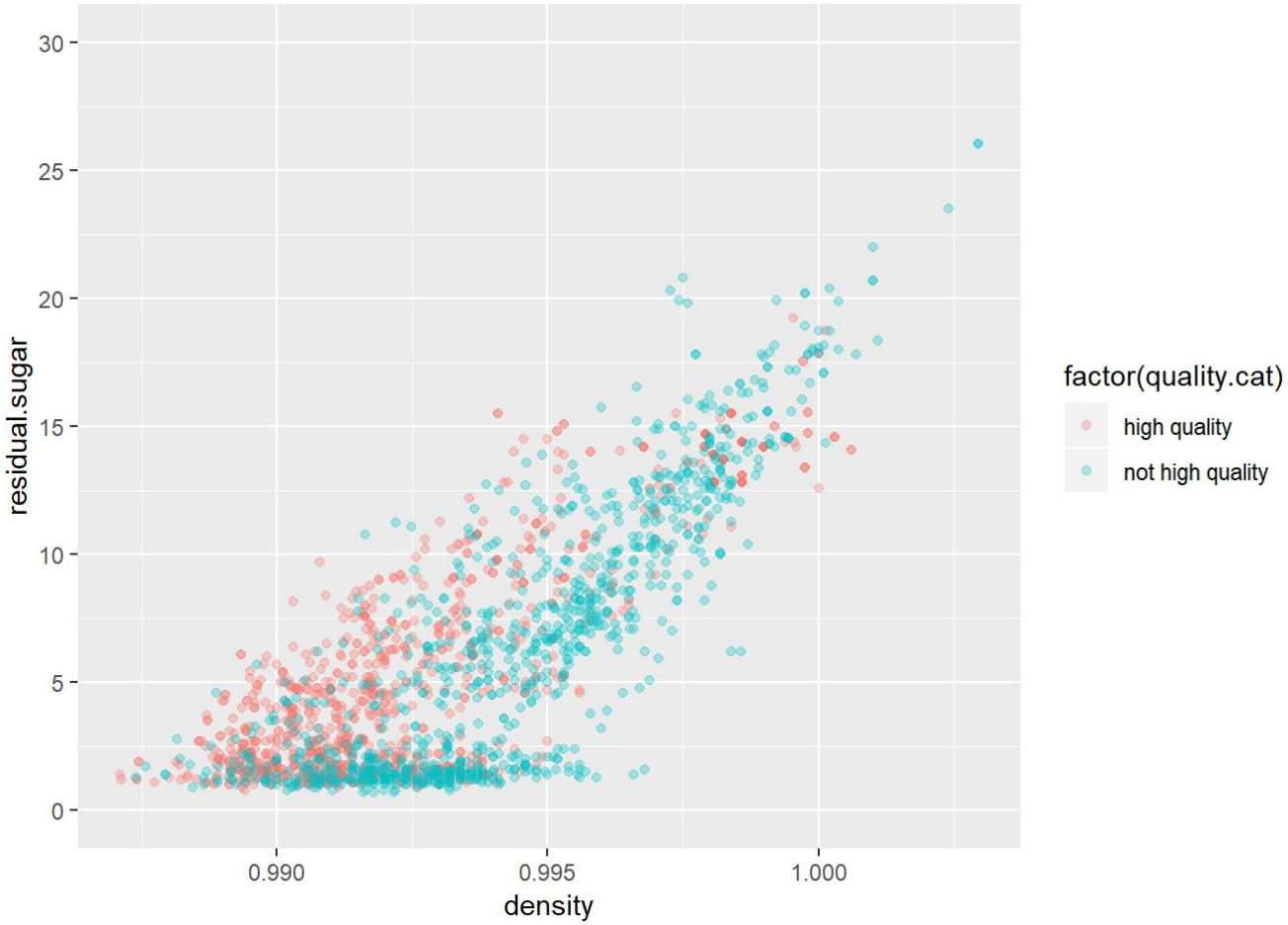
Multivariate Plots

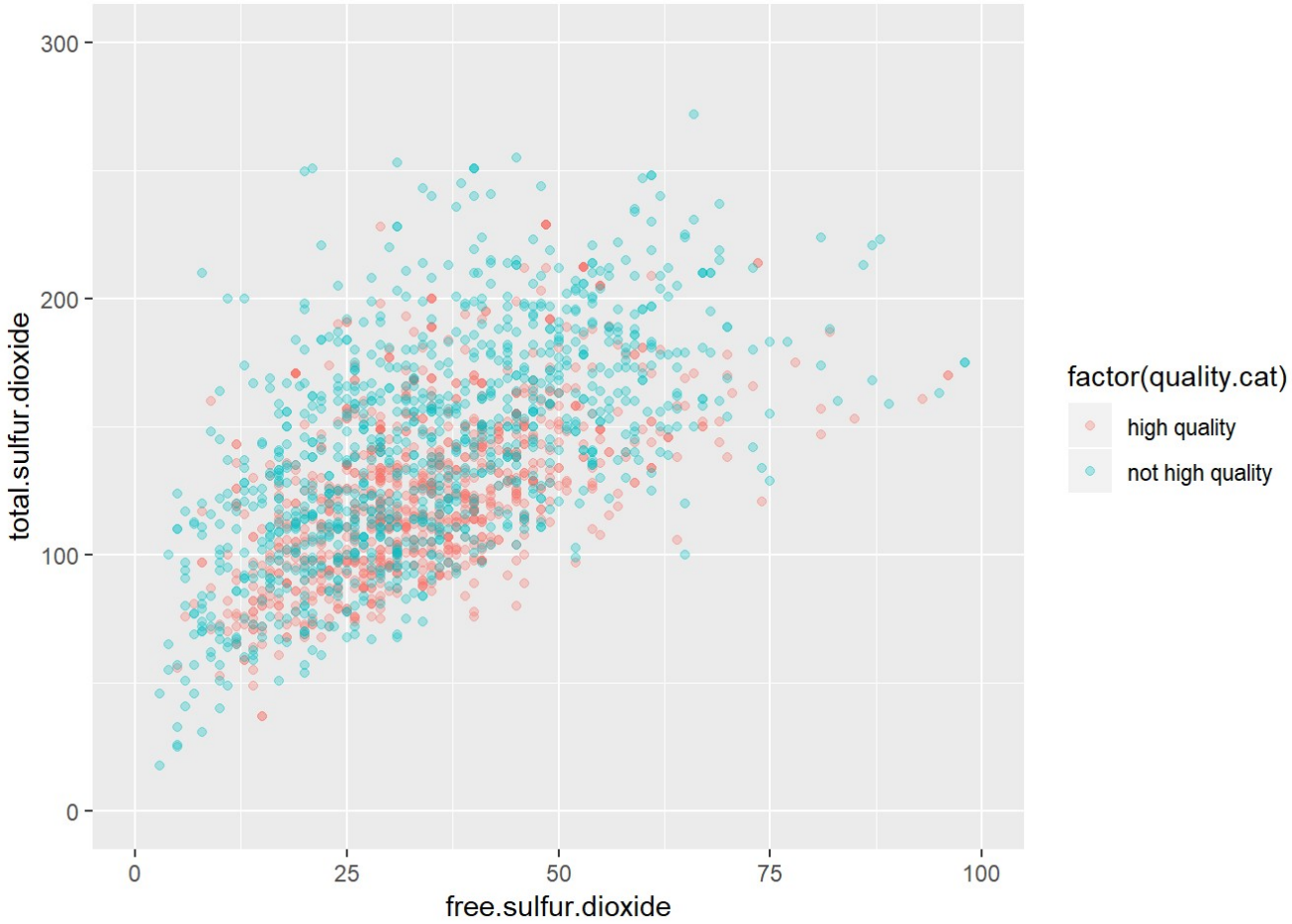
From the Bivarite Plots in the previous section, I assumed that a lot of the variables related to taste and flavor would impact the quality of the wine (i.e. have a moderate to strong correlation). However, the analyses above suggest otherwise.

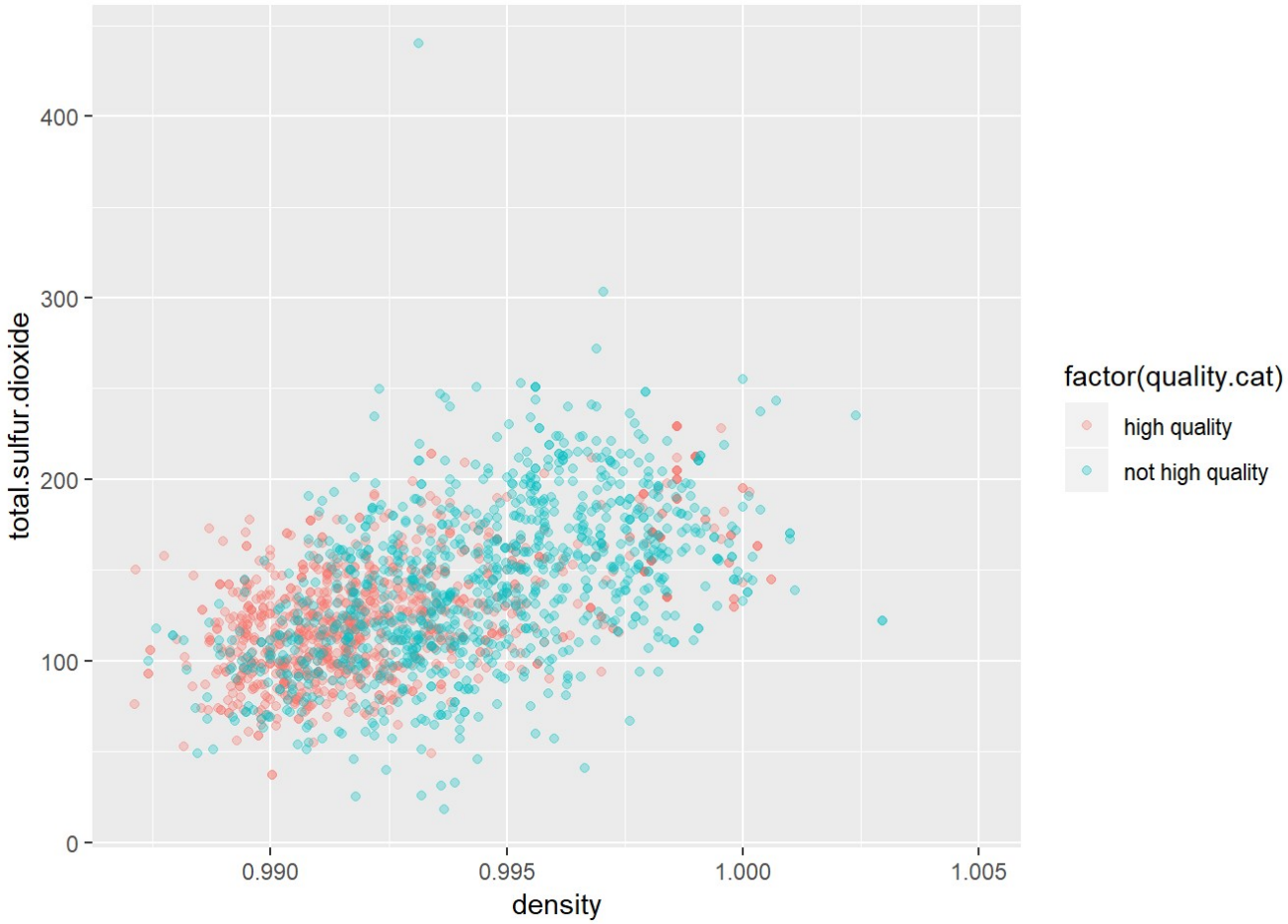
Let's revisit the plots for our three strongest positive correlations and include quality since that criteria is what

I'm most interested in understanding. The quality variable I'll be using is "high quality" versus "not high quality" (from quality.cat which I created) which will plot an equal number of observations from both categories.

- density and residual sugar
- free sulfur dioxide and total sulfur dioxide
- density and total sulfur dioxide




300
200
100
0
total.sulfur.dioxide
0
25
50
75
100
free.sulfur.dioxide
factor(quality.cat)
high quality
not high quality

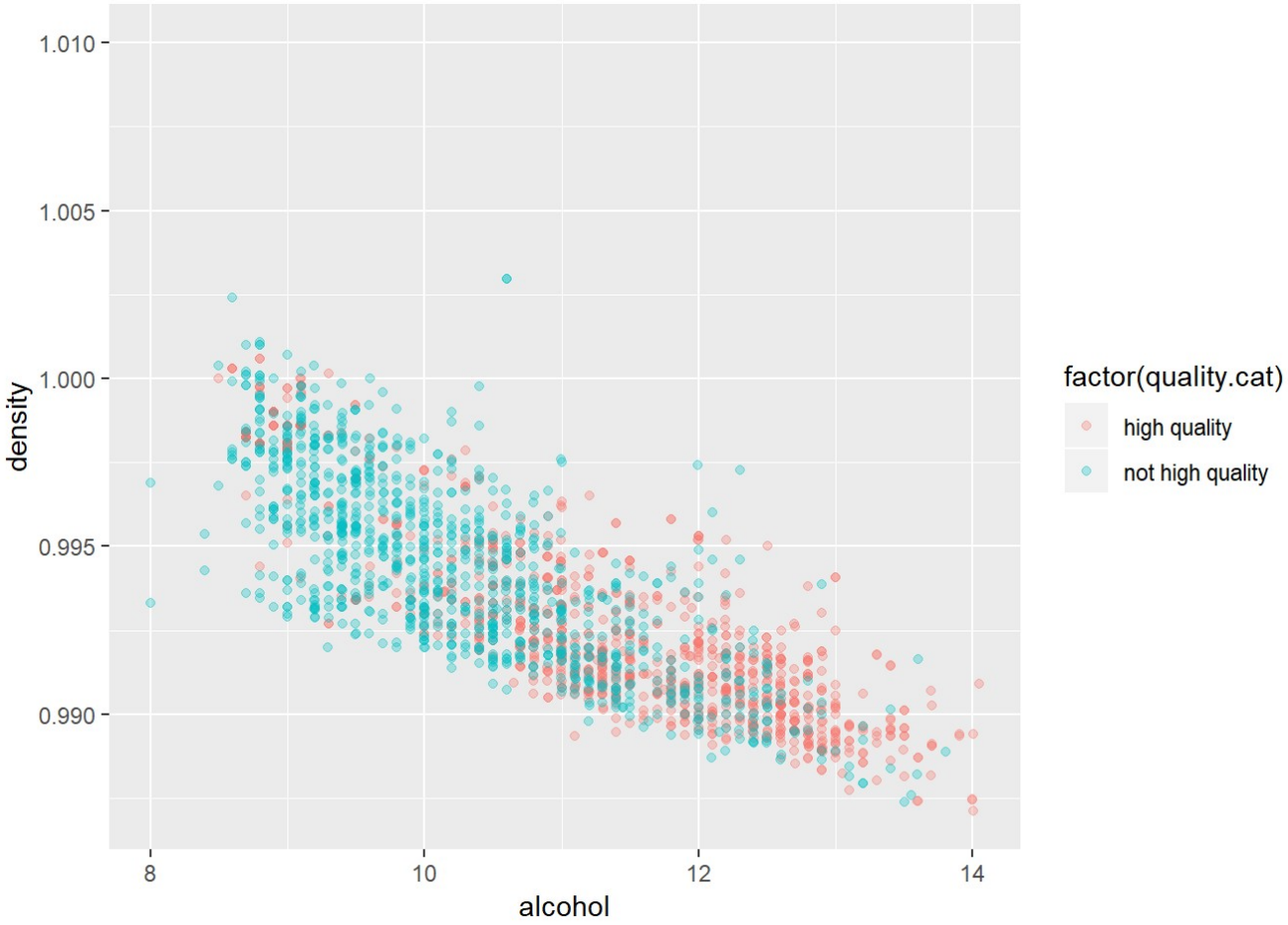

400
300
200
100
0
total.sulfur.dioxide
0.990
0.995
1.000
1.005
density
factor(quality.cat)
high quality
not high quality

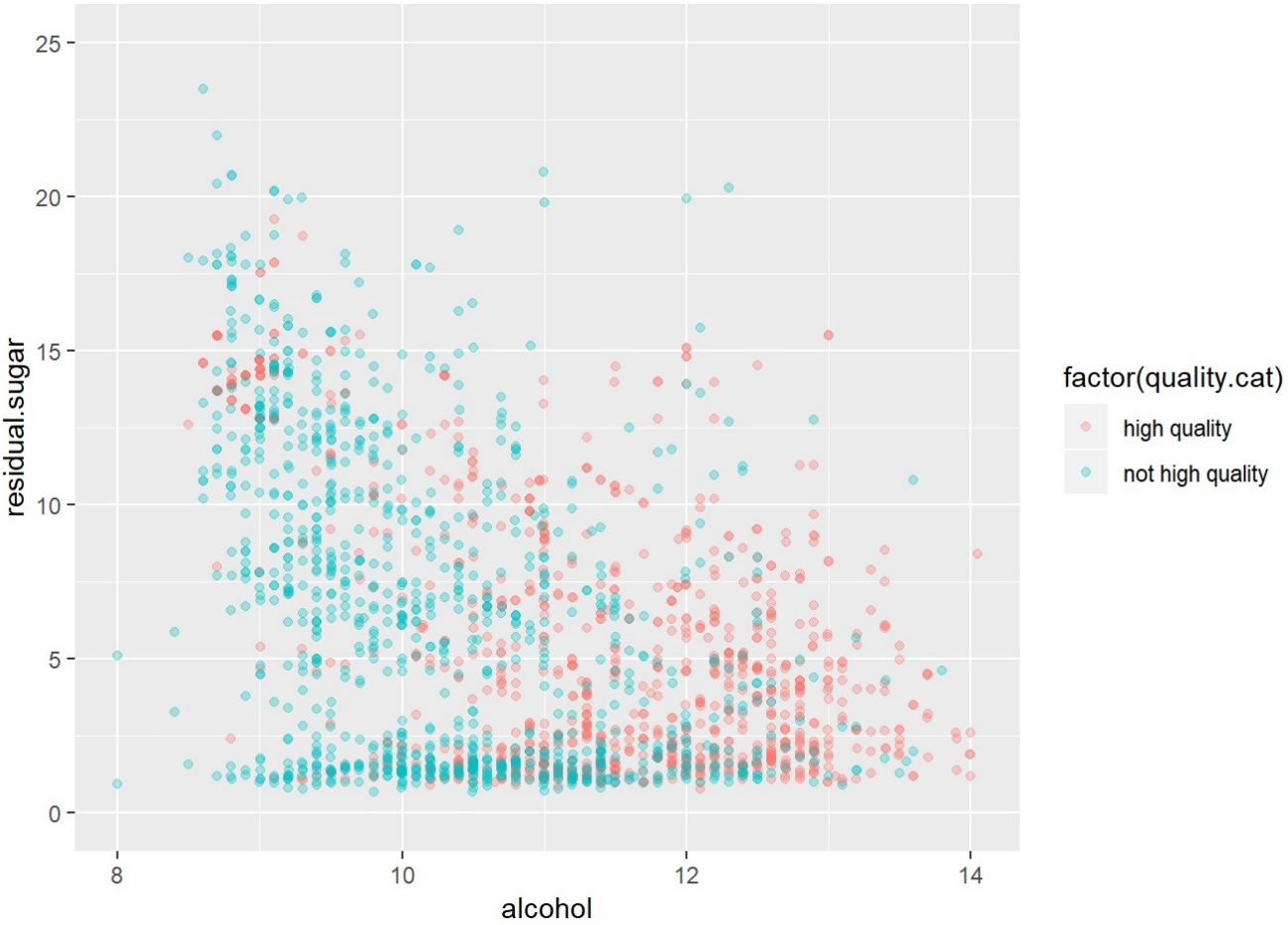
By overlaying the quality categories over the scatterplots we looked at earlier, it becomes clear that there's not only a relationship between the x and y variables of the three plots above, there also appears to be a relationship between those variables and wine quality.

In the first of the three plots–density and residual sugar–we can see the dots for higher quality wines shifted slightly to the left and not as spread out along the y-axis. For the last of the three plots, we notice that the higher quality values tend to fall in lower ranges for both density and total sulfur dioxide.

Let's try the plots for the two strongest negative correlations.

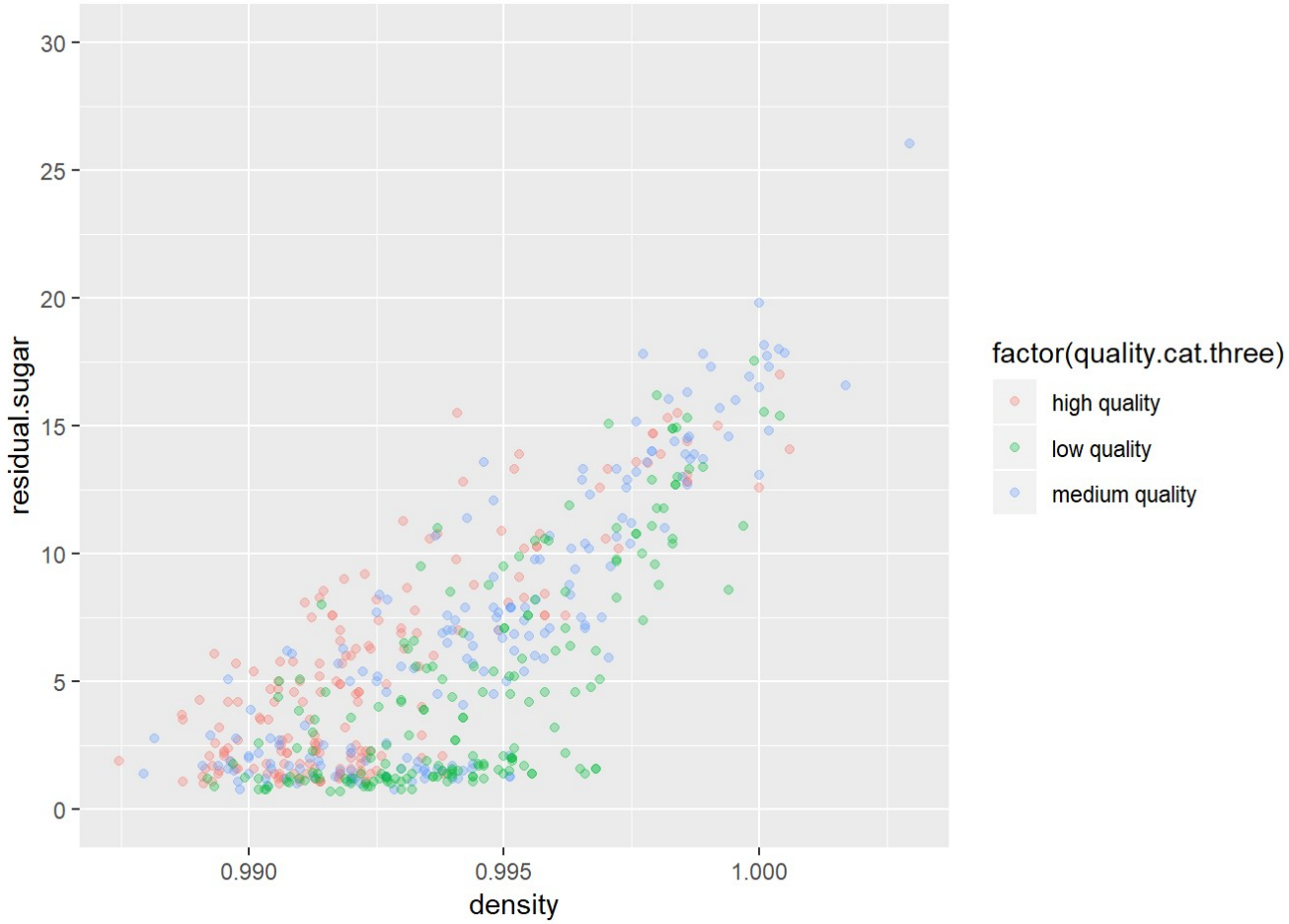
- alcohol and density
- alcohol and residual sugar

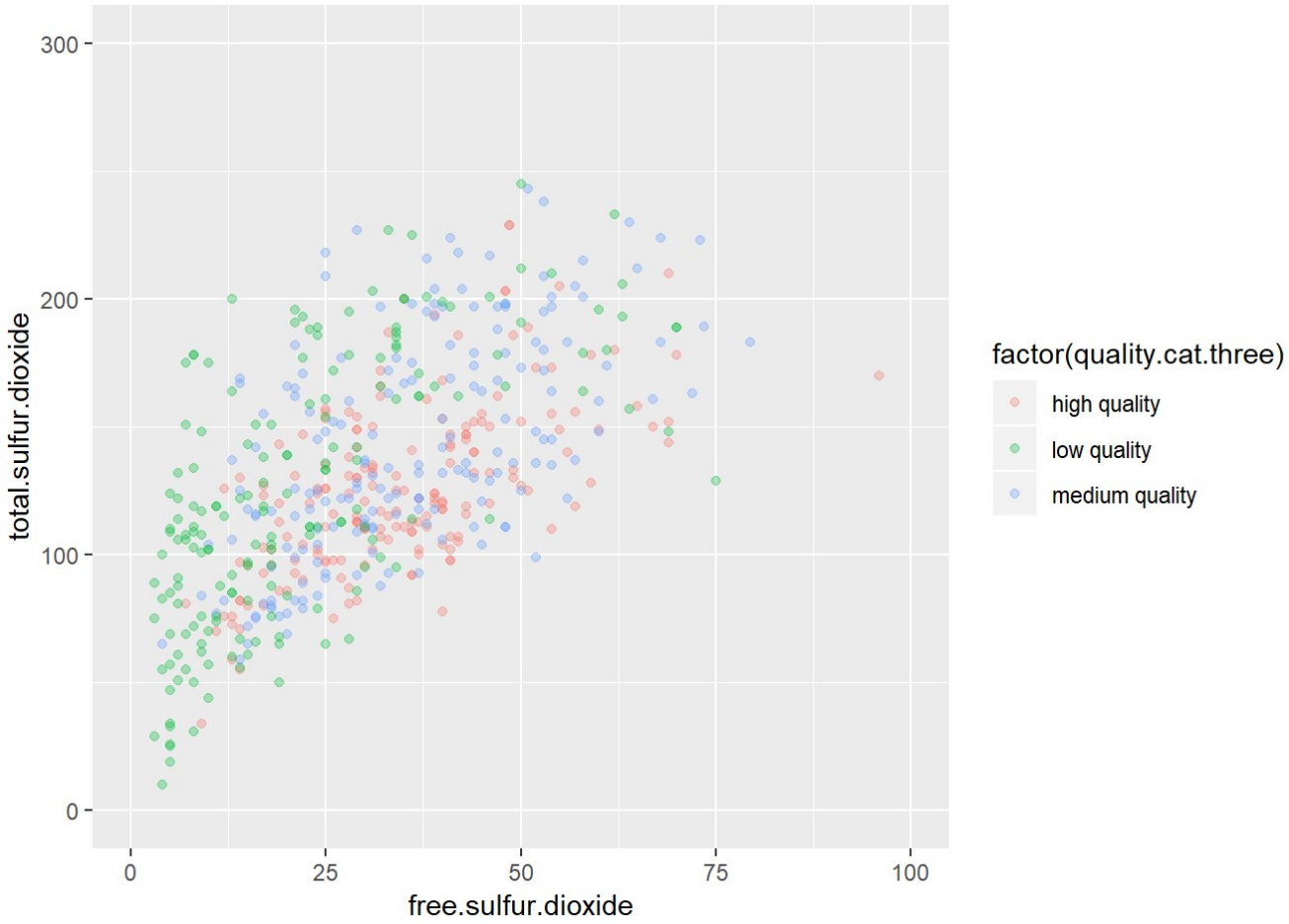


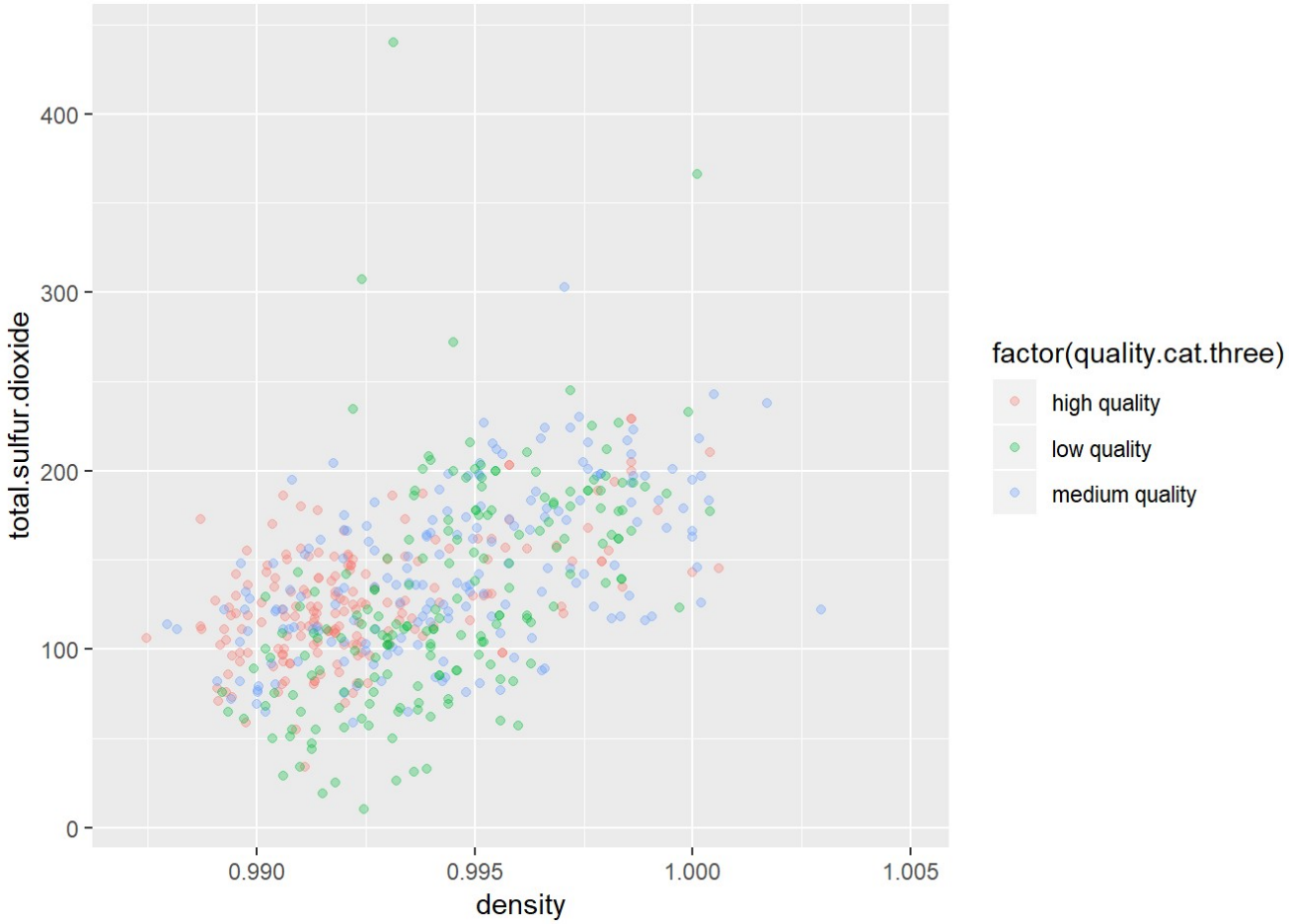


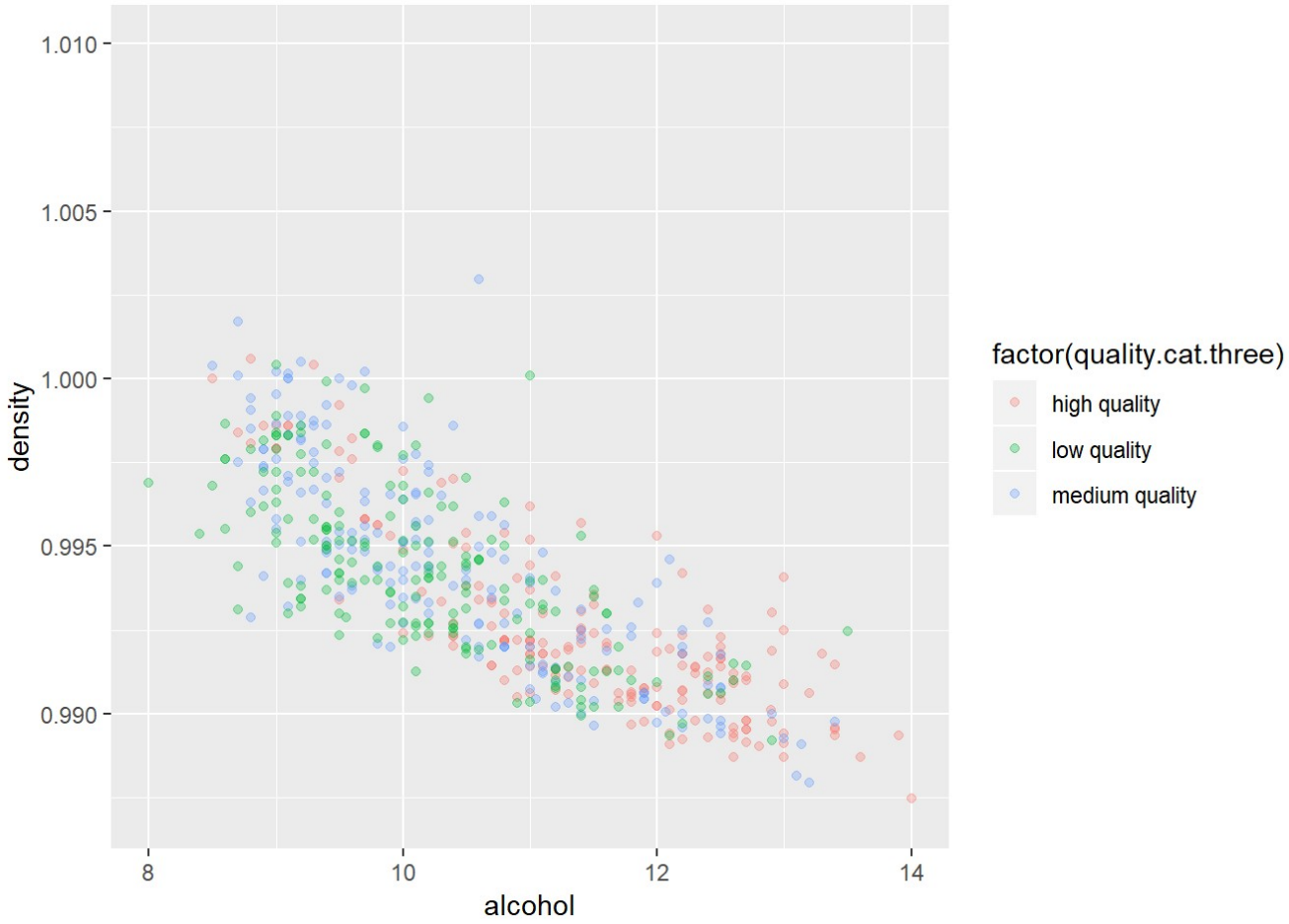
Overlaying the quality categories over the alcohol vs density and alcohol vs residual sugars plots again reveal that differences between higher and lower quality wines within the context of these variables.

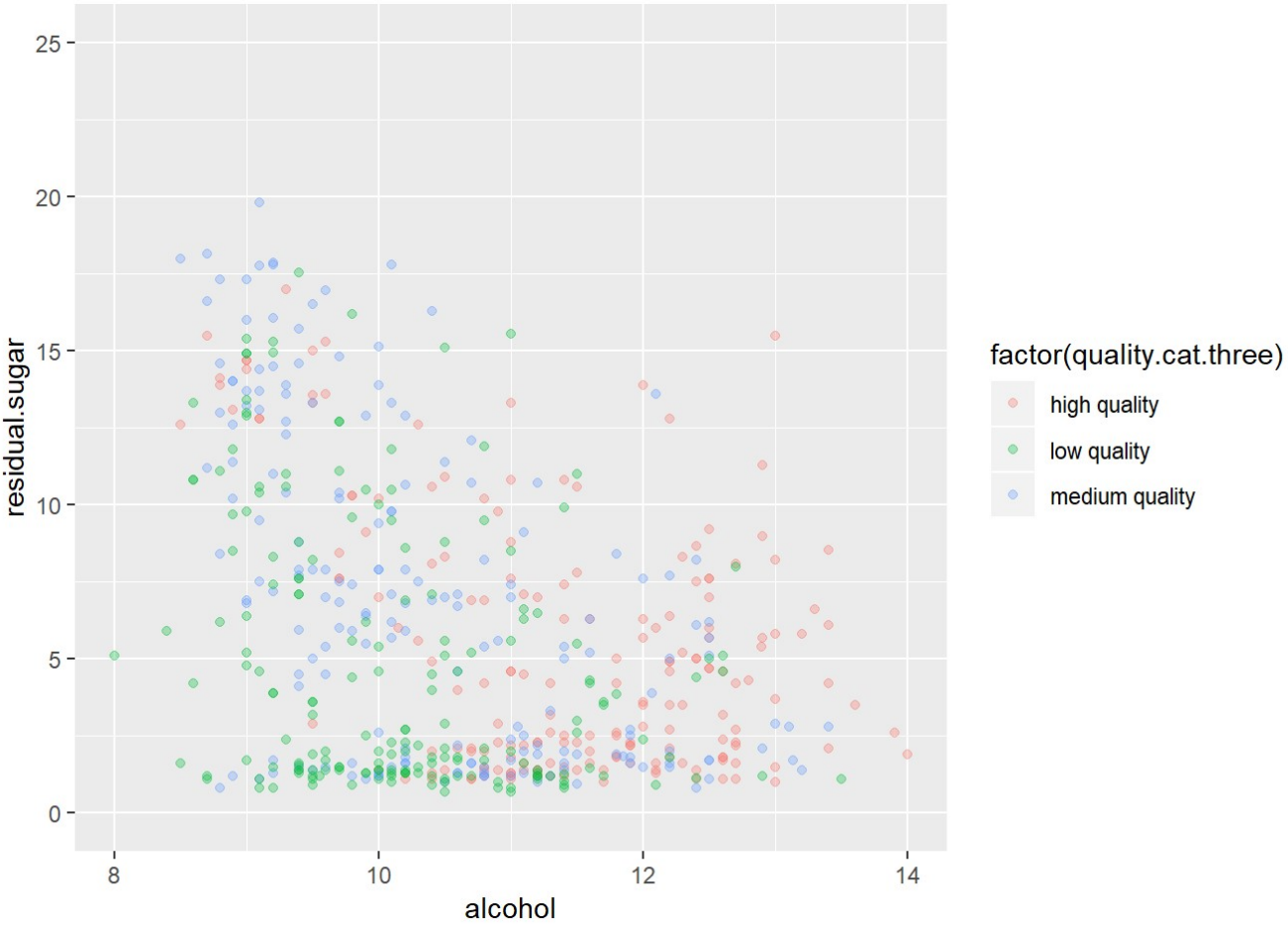
I want to look at these plots even more closely, particularly quality. I'll do this by categorizing the wine qualities into three categories instead of the two we have. The categories will be low (wine qualities 3 and 4), medium (wine qualities 5-6), and high (wine qualities 7-9). In comparing the number of observations across these three categories, low quality wines have the fewest observations (183), therefore, I'll take a random sample of 183 observations from the three categories for a total of 549 observations in this new sample dataset.


residual.sugar
density
0.990
0.995
1.000
0
5
10
15
20
25
30
factor(quality.cat.three)
high quality
low quality
medium quality


total.sulfur.dioxide
free.sulfur.dioxide
0
25
50
75
100
0
100
200
300
factor(quality.cat.three)
high quality
low quality
medium quality

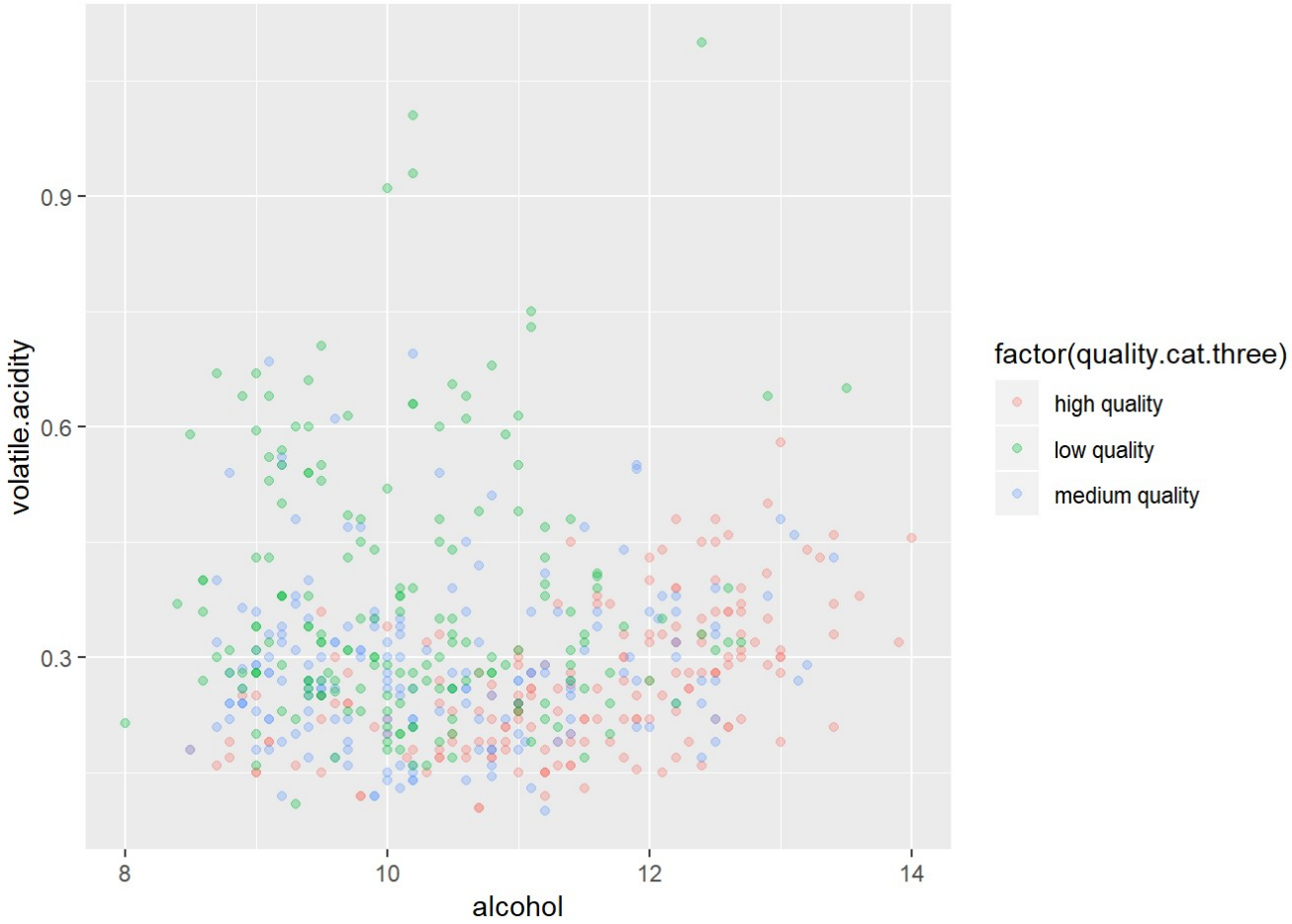

total.sulfur.dioxide
400
300
200
100
0
0.990
0.995
1.000
1.005
density
factor(quality.cat.three)
high quality
low quality
medium quality

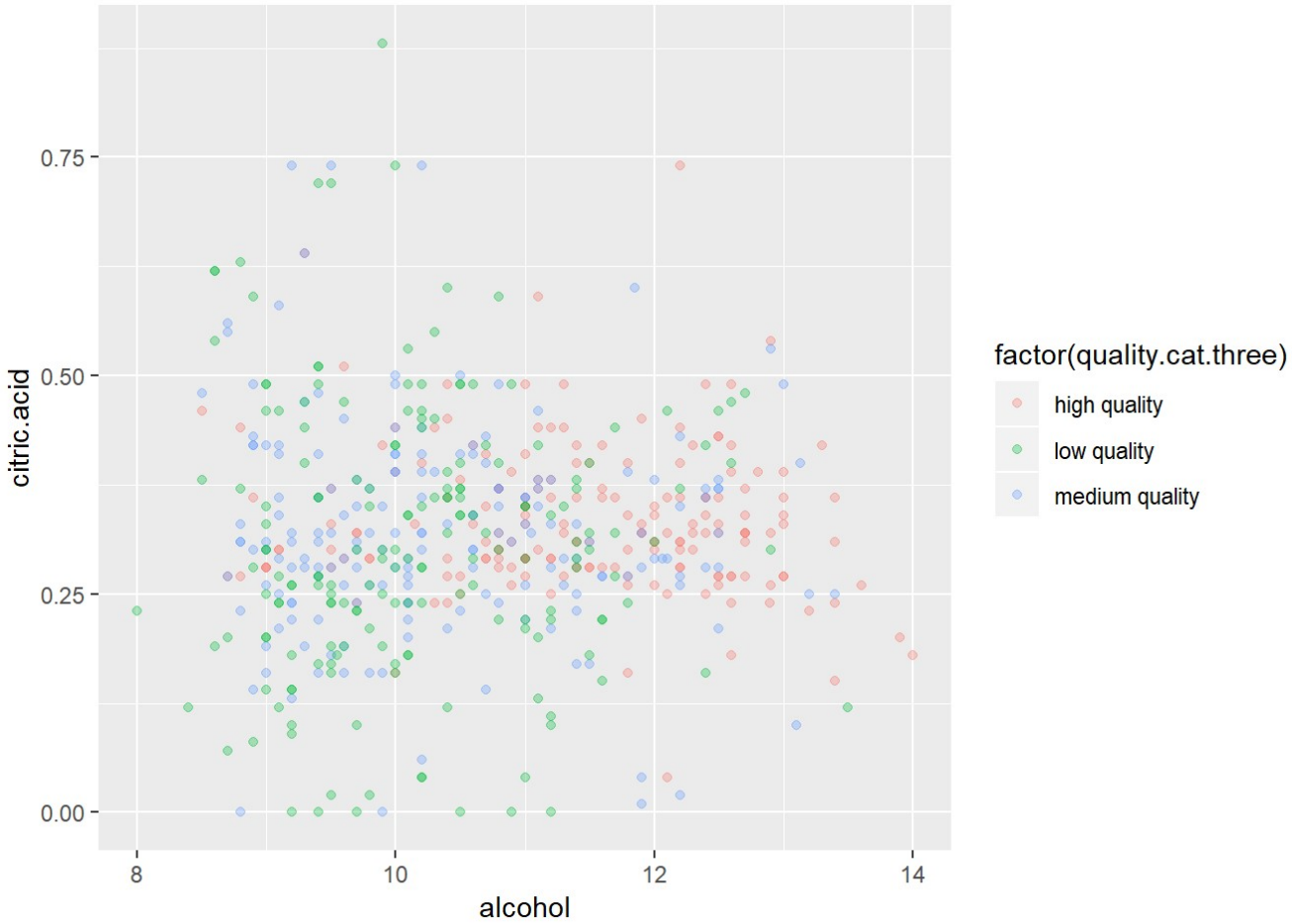

density
1.010
1.005
1.000
0.995
0.990
8
10
12
14
alcohol
factor(quality.cat.three)
high quality
low quality
medium quality

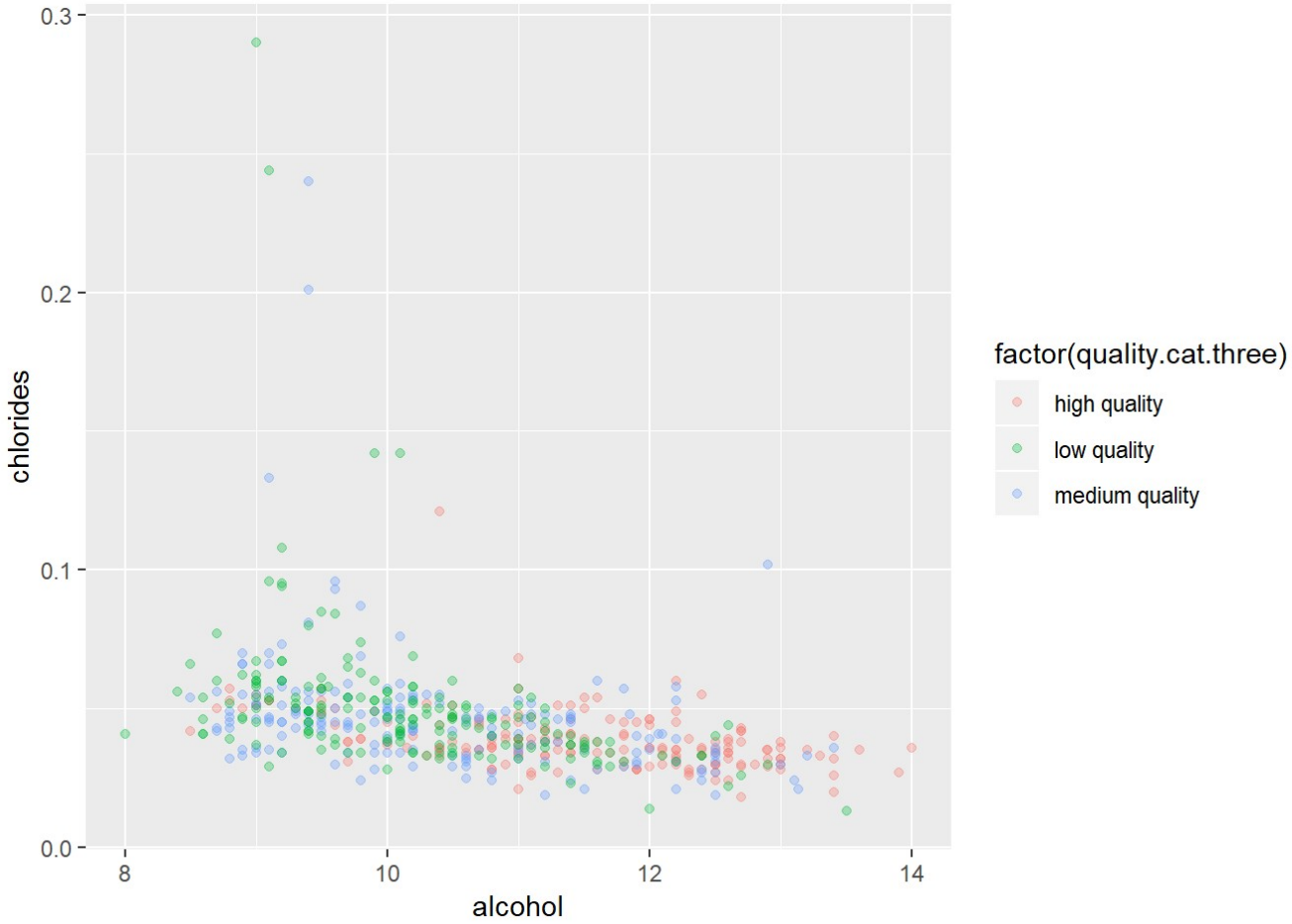


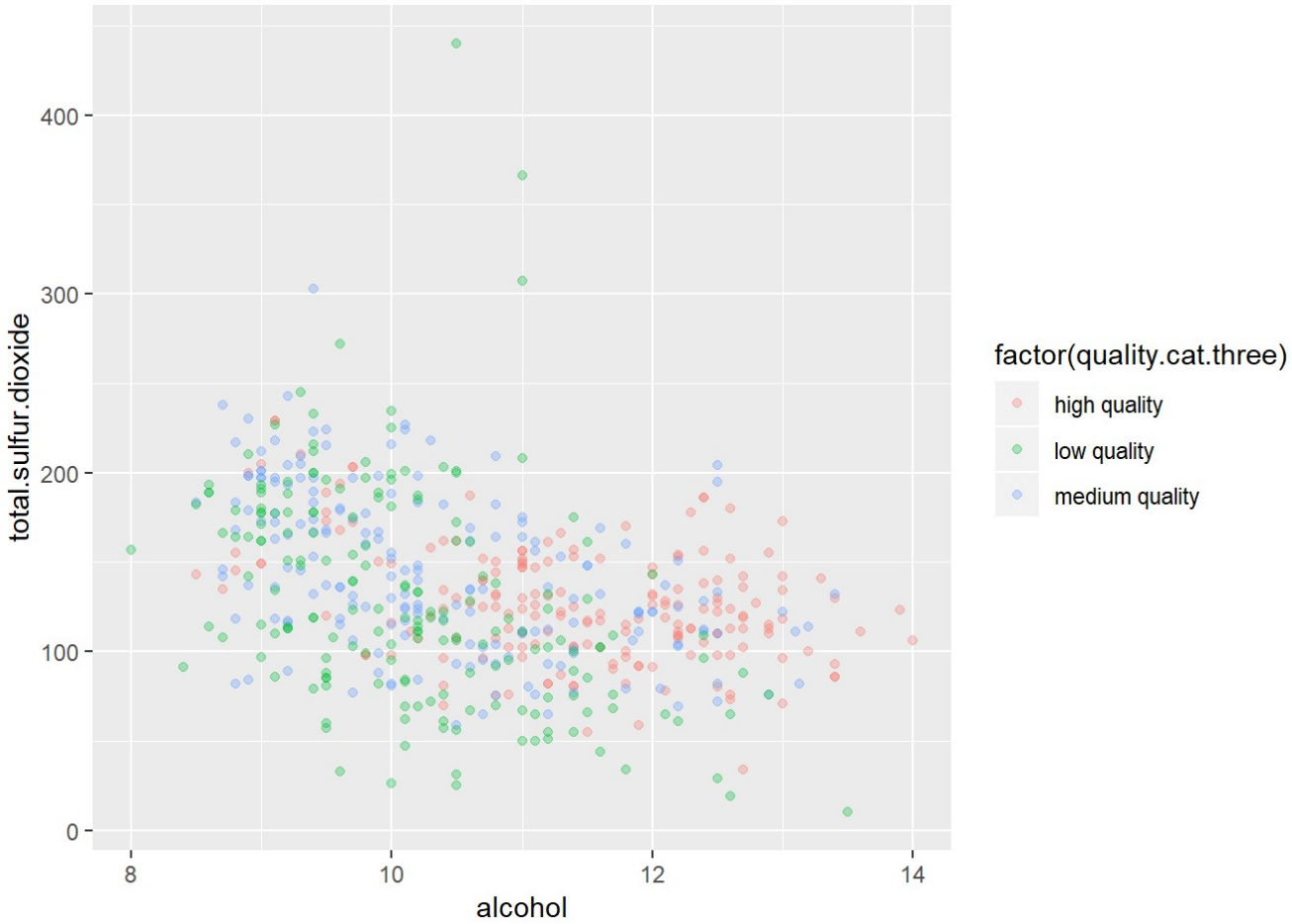
Breaking down the previous scatterplots into three wine qualities, we notice some clear distinctions among the high quality wines (rated 7 and up) versus the low and medium quality wines.

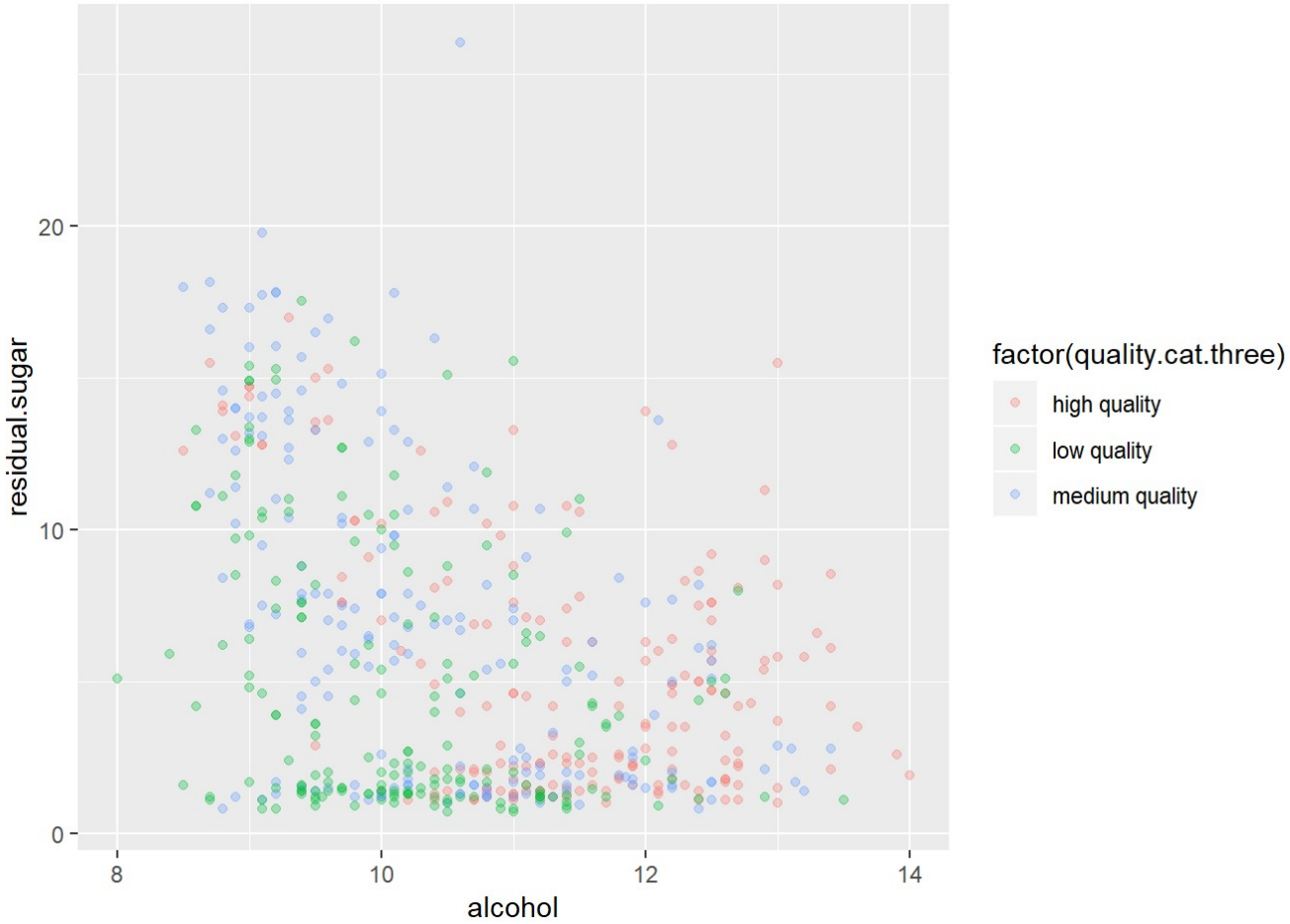
Now, let's return to the variables related to flavor. The plots below consider the relationship between alcohol (which we know is moderately correlated with quality) and the other variables that impact taste and flavor. For this set of plots, I'm interested in looking at three quality categories I created in my second sample data set.


volatile.acidity
0.9
0.6
0.3
8
10
12
14
alcohol
factor(quality.cat.three)
high quality
low quality
medium quality


citric.acid
0.75
0.50
0.25
0.00
8
10
12
14
alcohol
factor(quality.cat.three)
high quality
low quality
medium quality


0.3
0.2
0.1
0.0
chlorides
8
10
12
14
alcohol
factor(quality.cat.three)
high quality
low quality
medium quality

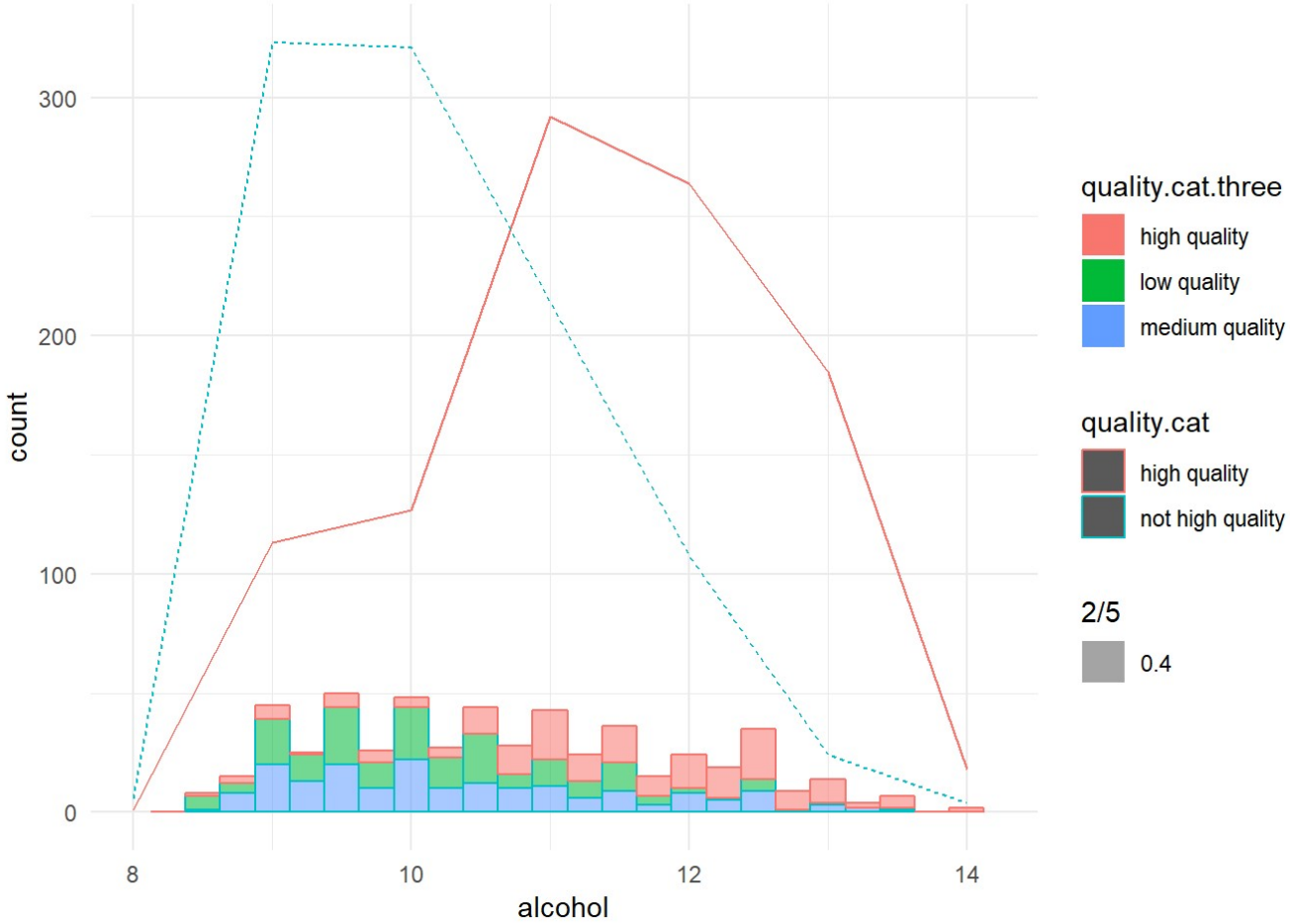

400
300
200
100
0
total.sulfur.dioxide
8
10
12
14
alcohol
factor(quality.cat.three)
high quality
low quality
medium quality



For the plots above, perhaps the two most insightful are residual.sugar vs citric.acid and residual.sugar vs chlorides. For residual.sugar vs citric.acid, the range of values of residual sugar seems to be comparable across all wine qualities; however, high quality wines are concentrated within a smaller range of citric acid values (approximately between 0.25 - 0.375). For residual.sugar vs chlorides, we can see that there is a particularly high concentration of low quality wines when chlorides and residual.sugar are both pretty low.

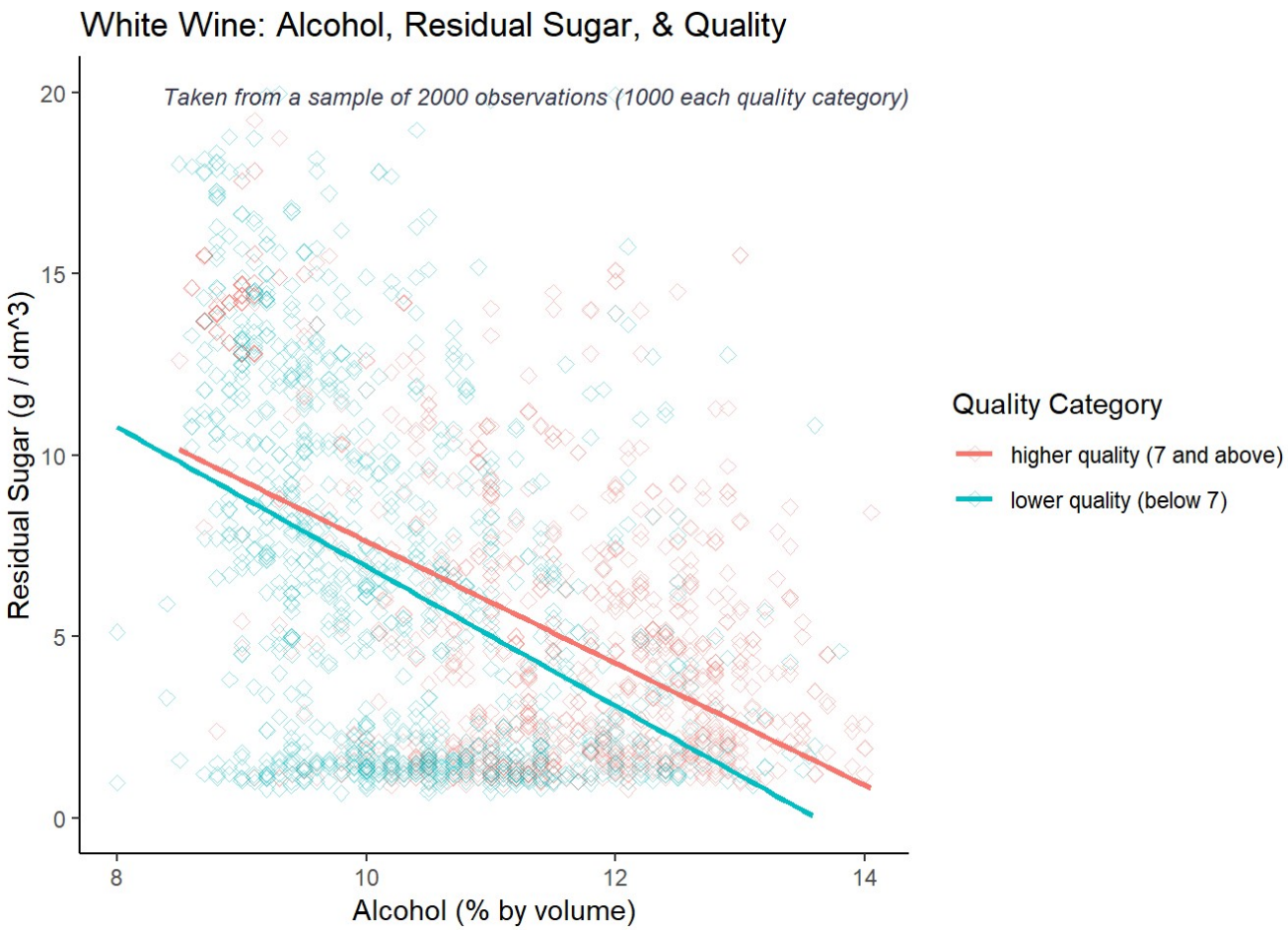
Final Plots

Plot 1 Alcohol and Quality



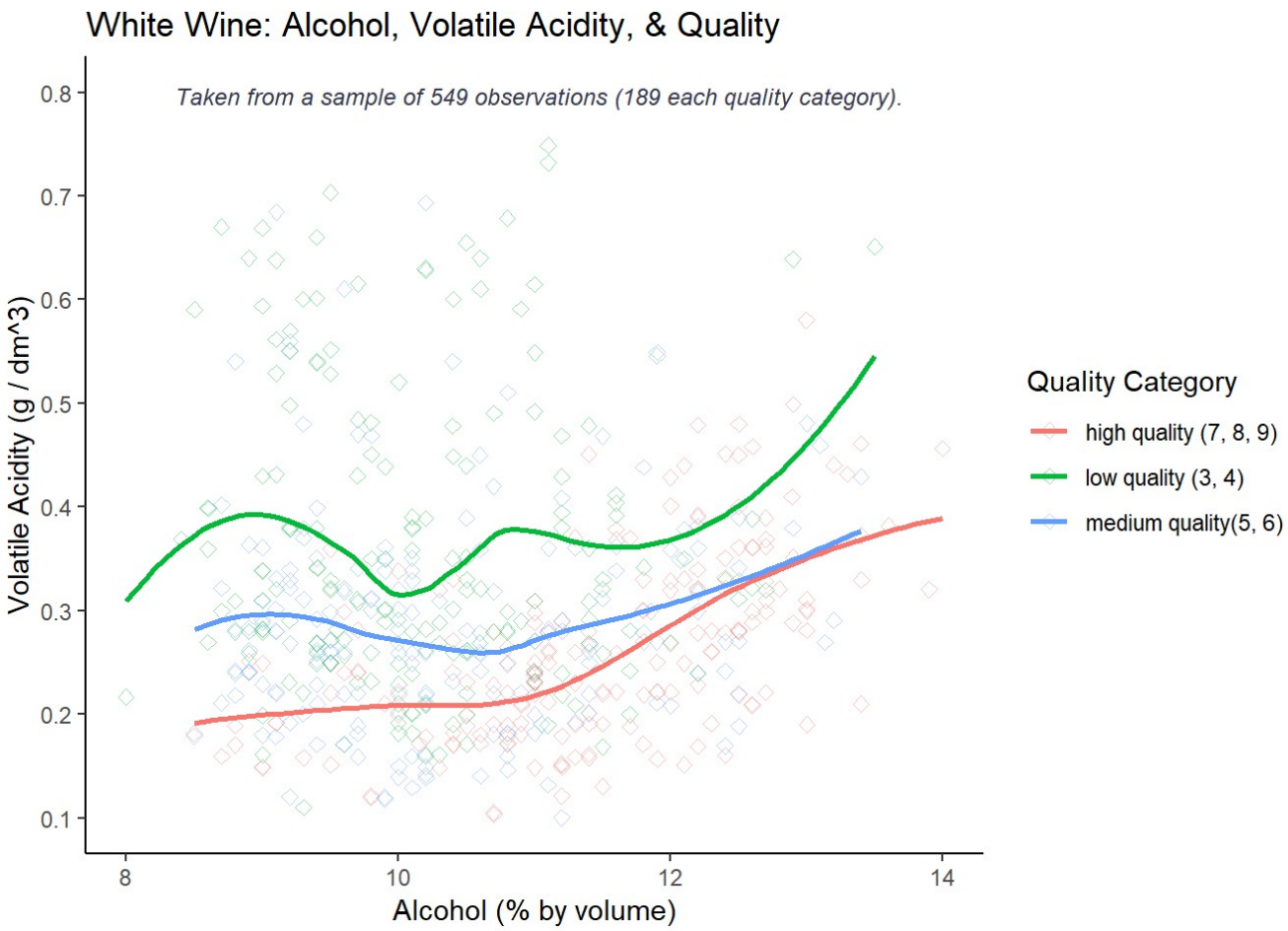
For counts of wine quality in alcohol percentage, generally the higher the alcohol content, the more red wines we find.

Plot 2 Residual Sugar, Alcohol, and Quality



When looking at alcohol and residual sugars there's an inverse relationship. When comparing higher quality wines with lower quality wines, the former has higher residual sugar as percent volume of alcohol increases.

Plot 3 Volatile Acidity, Alcohol, and Quality



Volatile acidity in medium quality wines remains pretty stable as alcohol percentage increase. For high quality wines, volatile acidity increases after about an alcohol percentage 11.5%, but the volatile acidity never reaches 0.4 For low quality wines, volatile acidity also increase at 11.5% but rather dramatically and surpasses an acidity of 0.5.

Summary & Reflections

Summary

My exploration was divided into three parts-univariate, bivariate, and multivariate. In the univariate exploration, I began by looking at the counts for variables that were easily identifiable (alcohol, PH, and quality). Within the same section, I transitioned into variables that-after doing some research-I understood as corresponding to the wine's flavor or taste (volatile acidity, citric acid, residual sugar, and chlorides).

I had made the assumption that the "flavor" variables would impact wine quality. Therefore, for the bivariate plots section, I was searching for any correlation between any of those values and the ratings for wine quality. I also discovered during this section that, of all the variables, alcohol had the strongest correlation to wine quality. In addition, I looked at the three strongest positive correlations between variables and the two strongest negative correlations.
In the multivariate section, I took a more in-depth look into how wine quality related to various variables by grouping the wine qualities using sampling. The analysis was structured to look at wine quality and alcohol and wine quality and the flavor variables.

Struggles & Successes

Surprisingly, my assumption that the flavor variables would have a higher correlation to wine quality was disproved. There was no clear indication that this hypothesis was true which was disappointing to uncover. However, as the plots came to finer detail, a general pattern between lower quality and higher quality wines became discernable-values for variables for lower quality wines were spread out over a wider range and narrower for higher quality wines.

Moving Forward

One idea to carry this preliminary exploration forward is to apply a similar analysis to a dataset of red wine with the same variables. Here is such a one: https://www.kaggle.com/piyushgoyal443/red-wine-dataset (https://www.kaggle.com/piyushgoyal443/red-wine-dataset). I'm curious whether the findings for white wine quality parallels that with red wine quality.

- Do flavor variables also have a weak correlation to red wine quality as they do for white wine quality?
- Do higher-rated red wines have variable values that are less spread out than lower-rated wines?
- Which variables have an inverse relationship to red wine quality compared to white wine quality (if any)?