EDA Wines

Nicholas Caminiti

August 7, 2018

Exploring Indicators of Wine Quality

This report documents the process of exploring a new dataset for the first time using R. The data being explored actually consists of two related datasets exploring the chemical composition of a selection of red and white wines, respectively. We will consider both datasets to aid in identifying any significant differences between the two when we move into multivariate analyses. (NOTE: I recognize that combining both the red and white wine datasets may mask some bivariate relationships that would have been more evident had either dataset been viewed in isolation, but we will proceed with this analysis for two reasons: 1. Combining the data from the two types of wine will allow us to spot bivariate relationships that are *common* to both red and white wines, and this in itself could be interesting. 2. any bivariate relationships that are masked will be revealed when we use wine color as a grouping factor in multivariate analyses.)

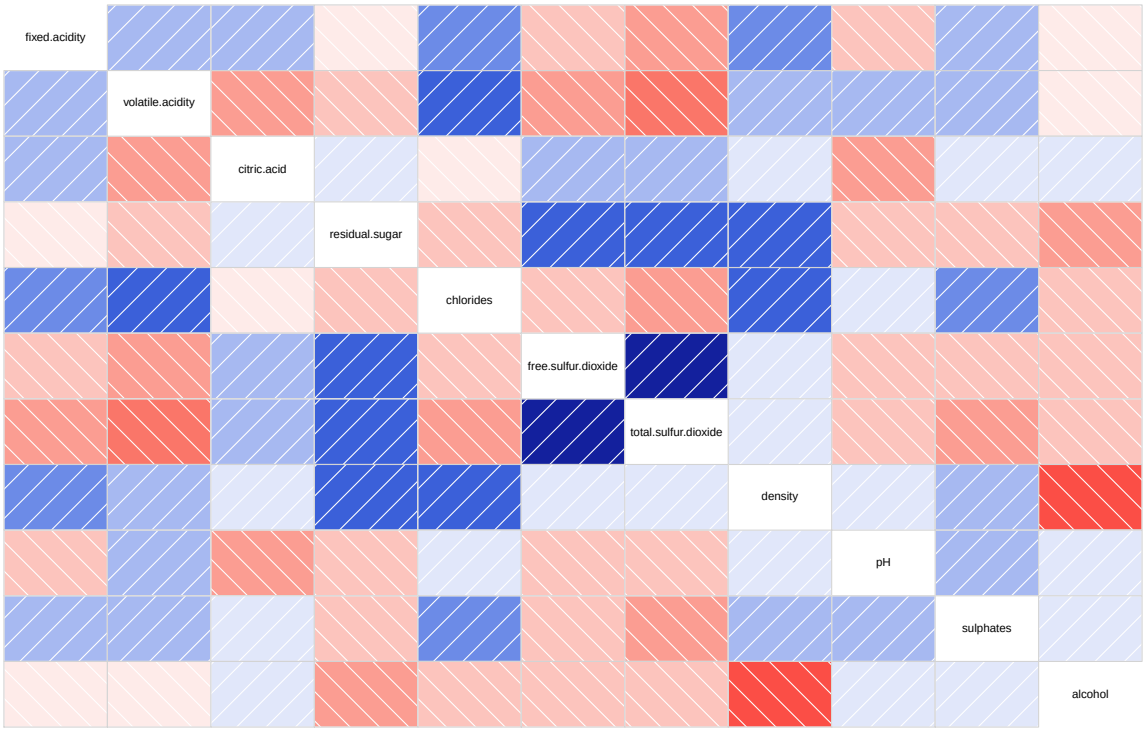
We have access to data on 6497 bottles of "Vinho Verde" in total, 1599 of which are red, and 4898 of which are white. The data include measurements on 11 chemical compounds or properties, plus a subjective rating of "quality" as determined by a panel of wine professionals.

I added a "color" variable to each dataset identifying each entry as "red" or "white", before combining the two sets into a master dataframe. This will allow us to group by color during certain analyses, which in turn will help expose major differences between the characteristics of the two types of wine. (NOTE: for post EDA statistical testing, we would want to code color as dummy integer variables rather than as a categorical string variable.) The "Quality" variable has been recoded as an ordered factor.

NOTE: For all variables except for "quality", individual outlier values (below .01 quantile or above .99 quantile) have been replaced with NA so they do not interfere with observations and calculations

```
##                        mean    sd median   min    max  skew kurtosis
## fixed.acidity          7.18  1.15   7.00  5.10  12.00  1.31     2.37
## volatile.acidity       0.33  0.15   0.29  0.12   0.88  1.18     0.89
## citric.acid            0.32  0.14   0.31  0.00   0.74  0.15     0.64
## residual.sugar         5.32  4.48   3.00  0.90  18.20  1.07     0.01
## chlorides              0.05  0.02   0.05  0.02   0.19  1.83     4.86
## free.sulfur.dioxide   30.16 16.21  29.00  4.00  77.00  0.43    -0.52
## total.sulfur.dioxide 115.44 54.16 118.00 11.00 238.00 -0.13    -0.72
## density                0.99  0.00   0.99  0.99   1.00 -0.06    -0.95
## pH                     3.22  0.15   3.21  2.89   3.64  0.25    -0.35
## sulphates              0.53  0.13   0.51  0.30   0.99  0.81     0.50
## alcohol               10.48  1.15  10.30  8.70  13.40  0.51    -0.72
```

The data documentation specifically states that some of the features may be correllated, and that it might be possible to select only a subset of the most relevant features without losing much information, so before I proceed, I will check for highly correlated features to see if we might be able to trim the amount of data we're looking at.



There are some clear correlations here and there, but nothing so blatant and consistent as to make me feel like features are overly redundant (NOTE: you can think of this step as a very crude preliminary factor analysis). I considered dropping one of the sulfur dioxide measurements because they are highly correlated and almost identical in their correlation profile with the other features, but I am electing to keep them both since they differ in their relationship to the "quality" measurement, which is perhaps the most important feature in the dataset.

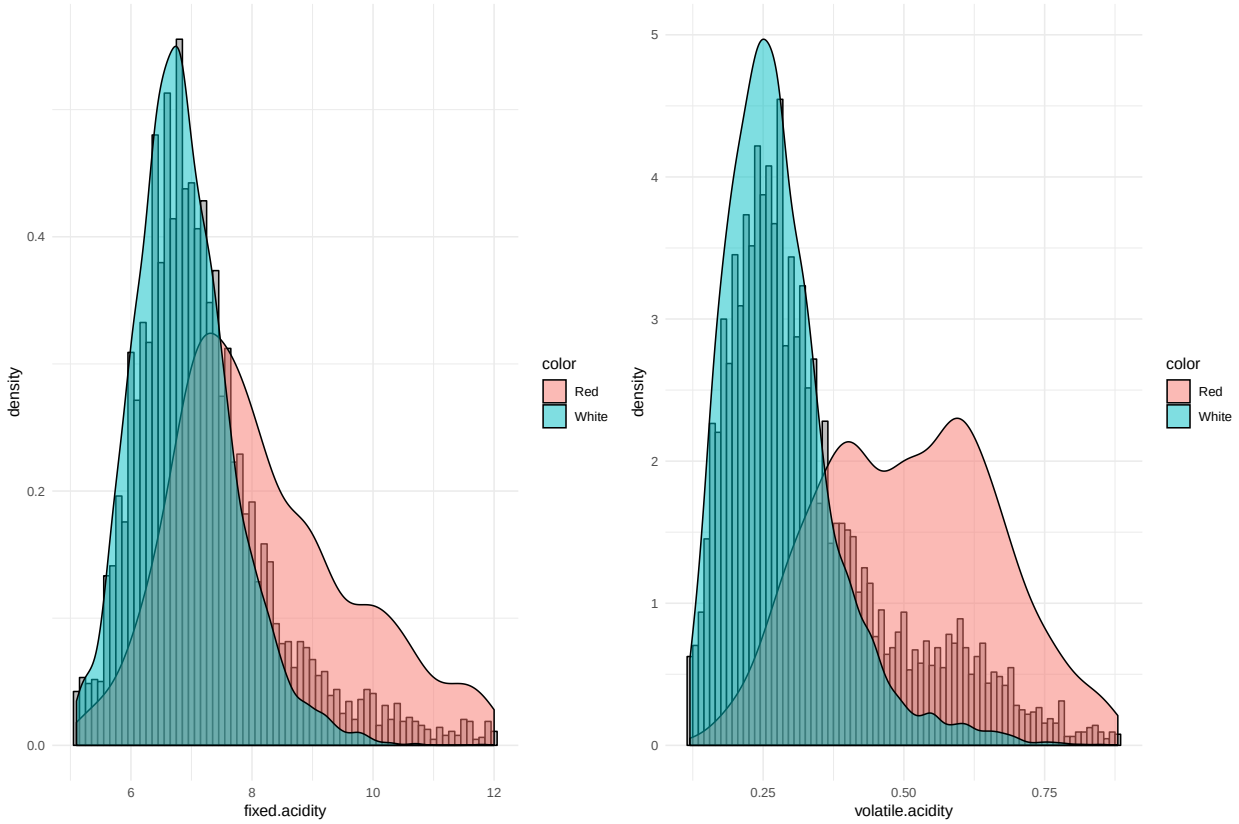
We will revisit some of these correlations in more detail later on, but for now we will proceed with a brief inspection of each feature. Please note that plot ranges have been restricted to make them more visually informative, and that this restriction may in some cases have excluded some outliers. These outliers will be brought back in during bivariate and multivariate analyses, where they may prove more interesting.

Univariate and bivariate (by color) plots and observations

Here we will look at density distributions of each variable. Note: We use density plots here instead of standard histograms because the difference in sample size between the red and white wines render a frequency histogram less informative than we'd like when comparing differences between the two groups.

Superimposing individual density curves for the red and white groups help to illustrate the effect of each subgroup on the distribution of the whole dataset. Any significant differences betweeen the reds and the whites that we discover during this step will help guide some of our multivariate exploration later on.

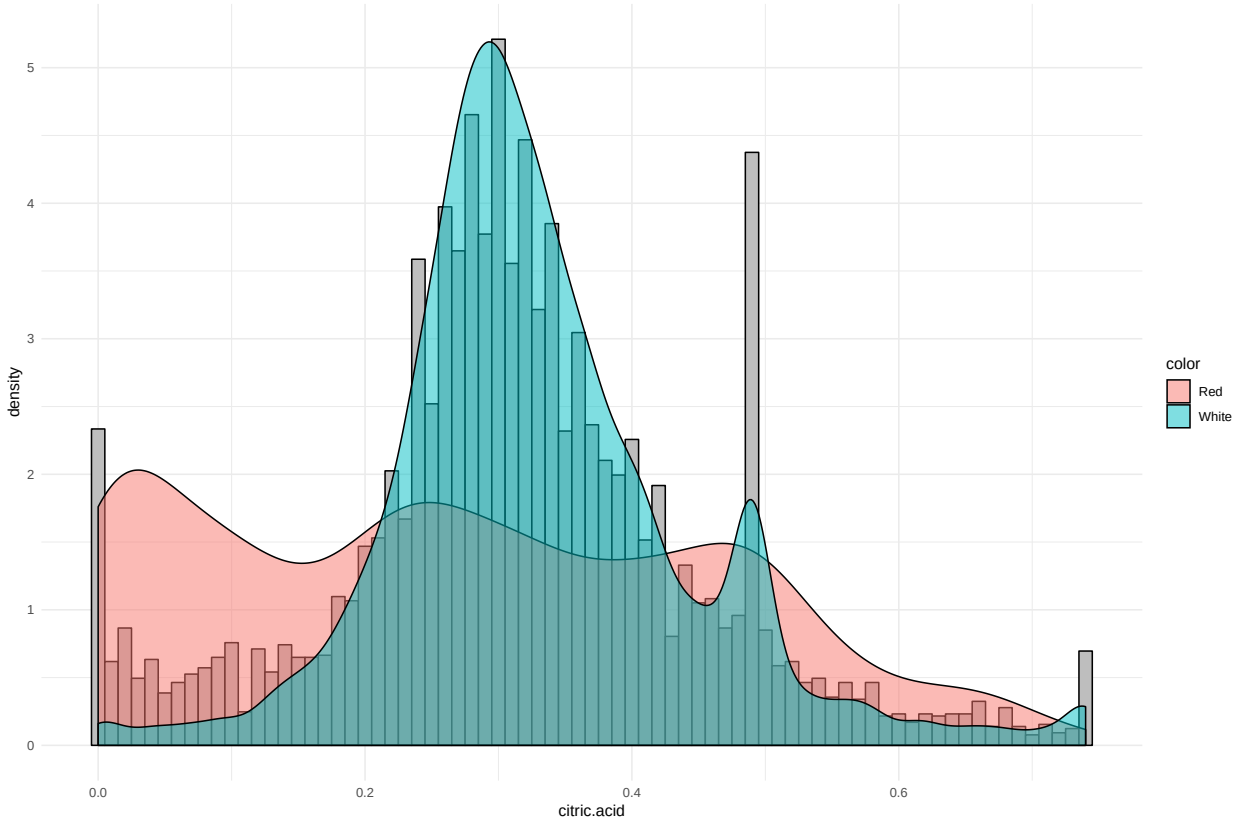
Fixed and Volatile Acidity



The most common fixed acidity value is around 7, but the distribution is clearly right-skewed for both the red and white wines (moreso for the reds). The white wines generally have a lower fixed acidity than the reds.

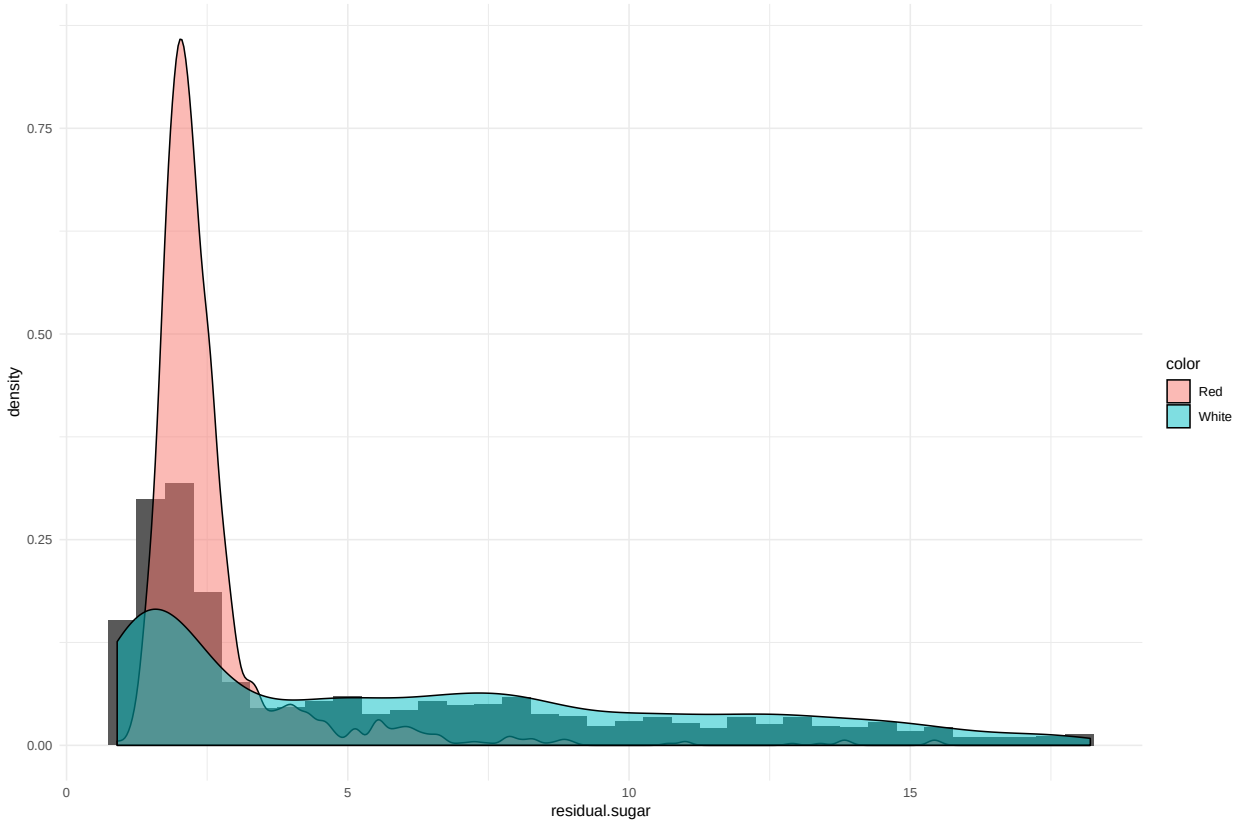
In our volatile acidity histogram, we see What appears to be a simple right-skewedness in volatile acidity similar to that seen in the plots for fixed acidity, but with a little more of a "hump" in the higher values. Our density plot reveals that hump is caused by the clearly bifurcated distribution of this variable. Like with fixed acidity, the reds have higher values than the whites, but here the difference is much more pronounced.

Citric Acid



Unlike with fixed and volatile acidity, the distribution of citric acid concentrations is relatively symetrical, suggesting that there are compounds other than citric acid which ultimately impact the acidity of each wine. The distribution of the values for the red wines does not resemble a normal distribution, and many of the reds have no citric acid at all. Furthermore, there is an unusual spike at CA value of 0.5 attributable to the white wines.

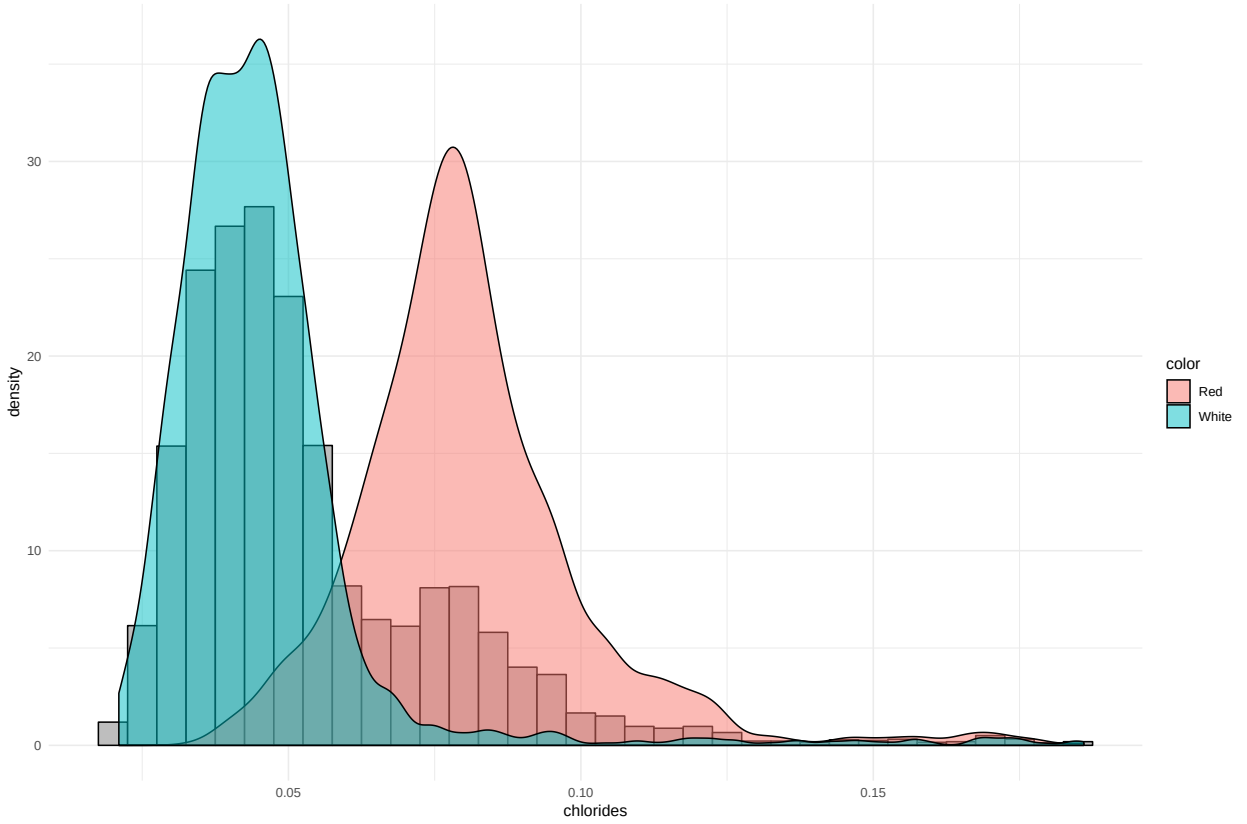
Residual Sugar



Another right-skewed feature. While the majority of the wines have a residual sugar value below 5, there are a handful of sweeter bottles. The sweetest bottle (not shown in plot) actually has a value of 65.8. I wonder what the judges thought of that one...

Our density plot reveals a significant difference between the reds and whites. Whereas the distribution of residual sugars for the reds is very leptokurtic, condensed almost entirely in the lowest values; the whites are platykurtic, with significantly more "sweet" bottles than the reds.

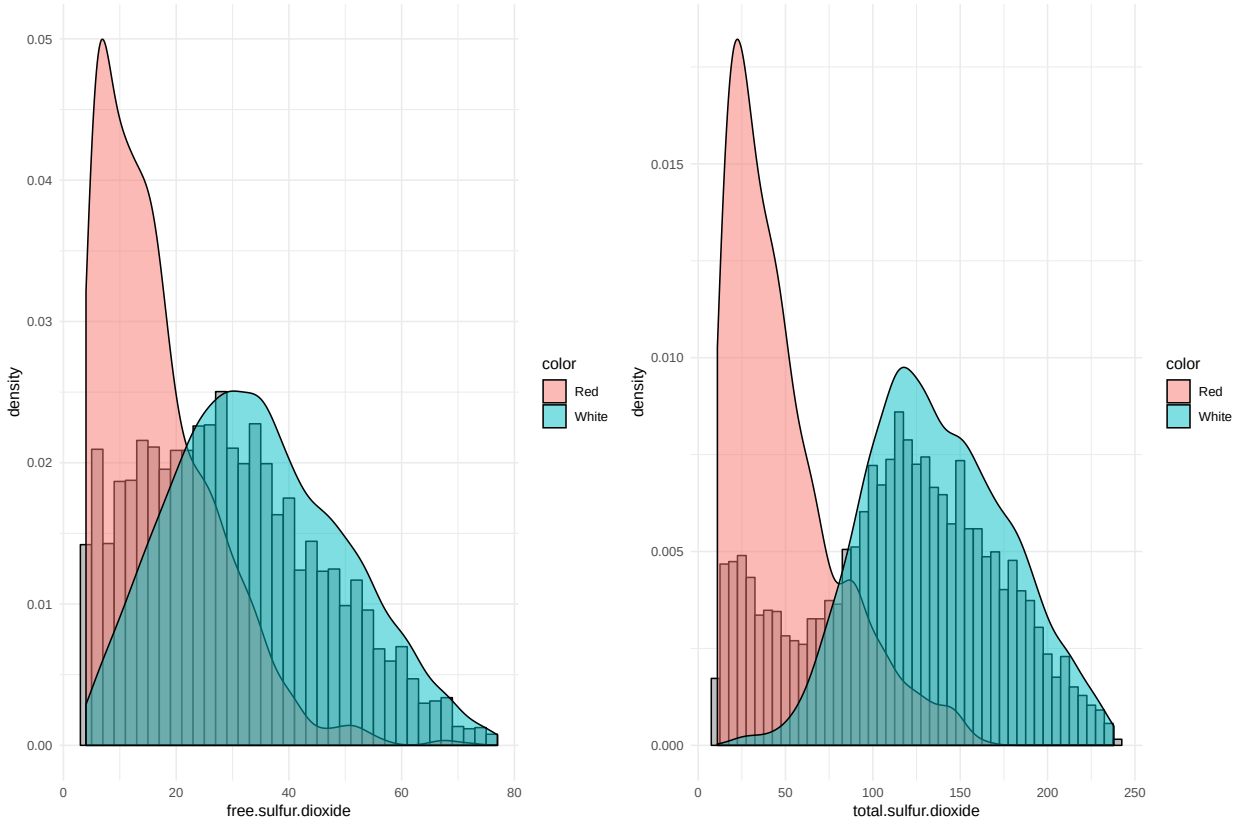
Chlorides



Continuing with our trend of right-skewed data, we have our plot of chloride content. Almost entirely $< .1$. I'm curious about those handful of bottles that have significantly higher chloride values.

Our density plot reveals what we could not see very well in our histogram: that the choloride distribution is heavily bifurcated, with the reds having higher chloride content than the whites.

Free and Total Sulfur Dioxide

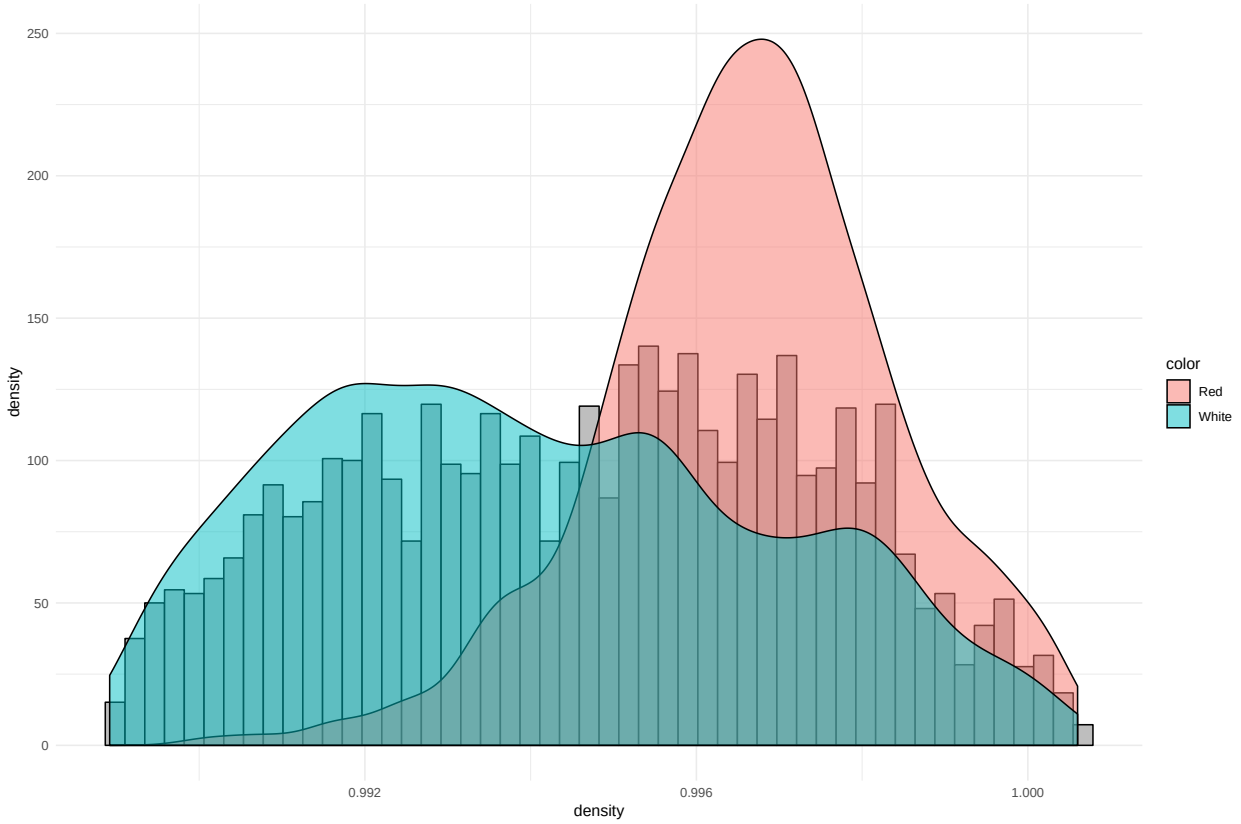


Good thing we didn't drop either variable!

The seemingly innocuous (but right skewed) distribution of the free.sulfur.dioxide histogram masked the bifurcated nature of this variable, but the bifurcation was so strong for total.sulfur.dioxide that the bimodality showed through even in the histogram.

Values are higher for white wines in both sulfur dioxide measurements, but there seems to be something about white wines that almost guarantees that total sulfur dioxide levels will be above a certain level, as there are very few white wines below 50. (NOTE: another thing to consider here is that total sulfur dioxide INCLUDES free sulfur dioxide. Might it be worthwile to create a new variable which subtracts free SO2 from from total SO2, giving us a measurement for "non-free SO2"? I don't know enough about wine to say whether or not this could be informative, but it would be worth looking into.)

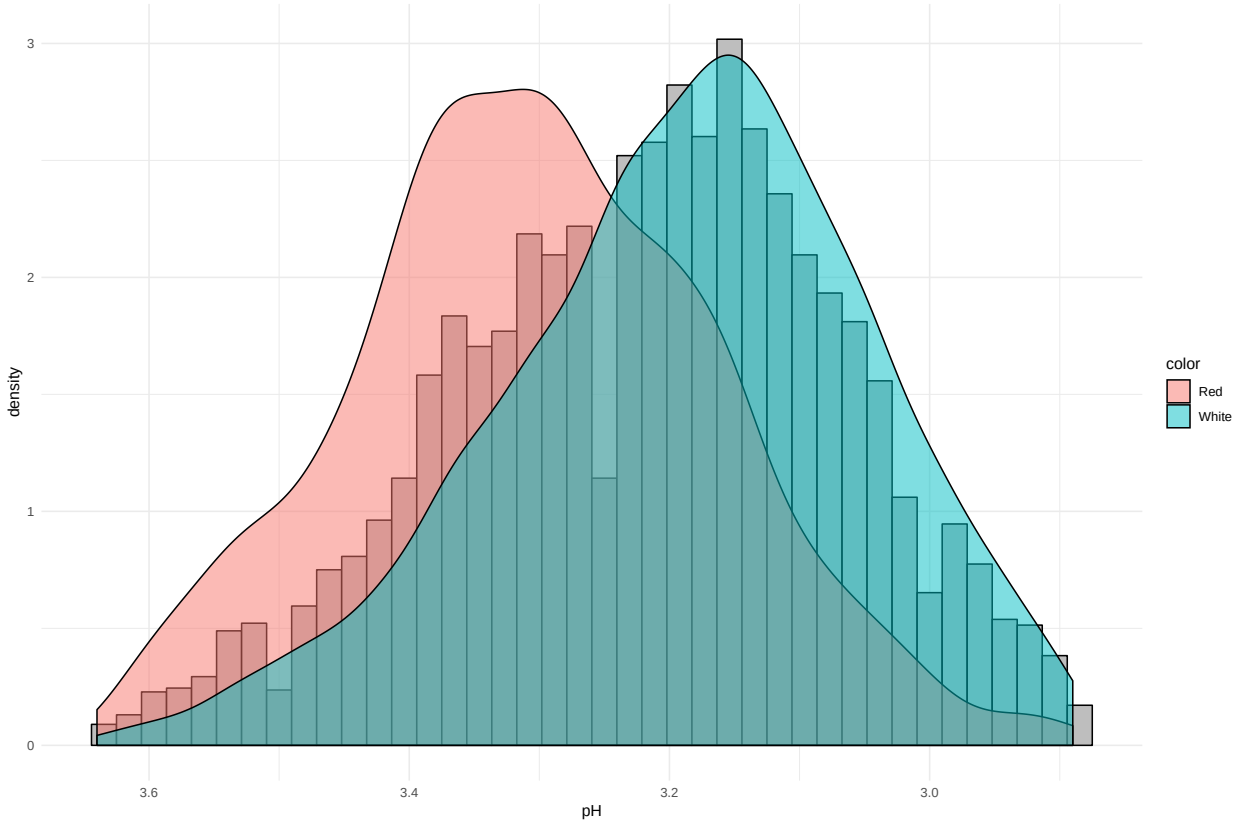
Density



Very little variation here (note the condensed x scale), and a relatively smooth distributions for both reds and white. This may still be interesting though. Our correlation graph showed a strong positive correllation between density and residual sugar (which makes sense because sugar is more dense than water), and a strong negative correlation between density and alcohol content (which makes sense because alcohol is less dense than water). The most dense wines are thus probably those that are the sweetest, and those which have the lowest alcohol content.

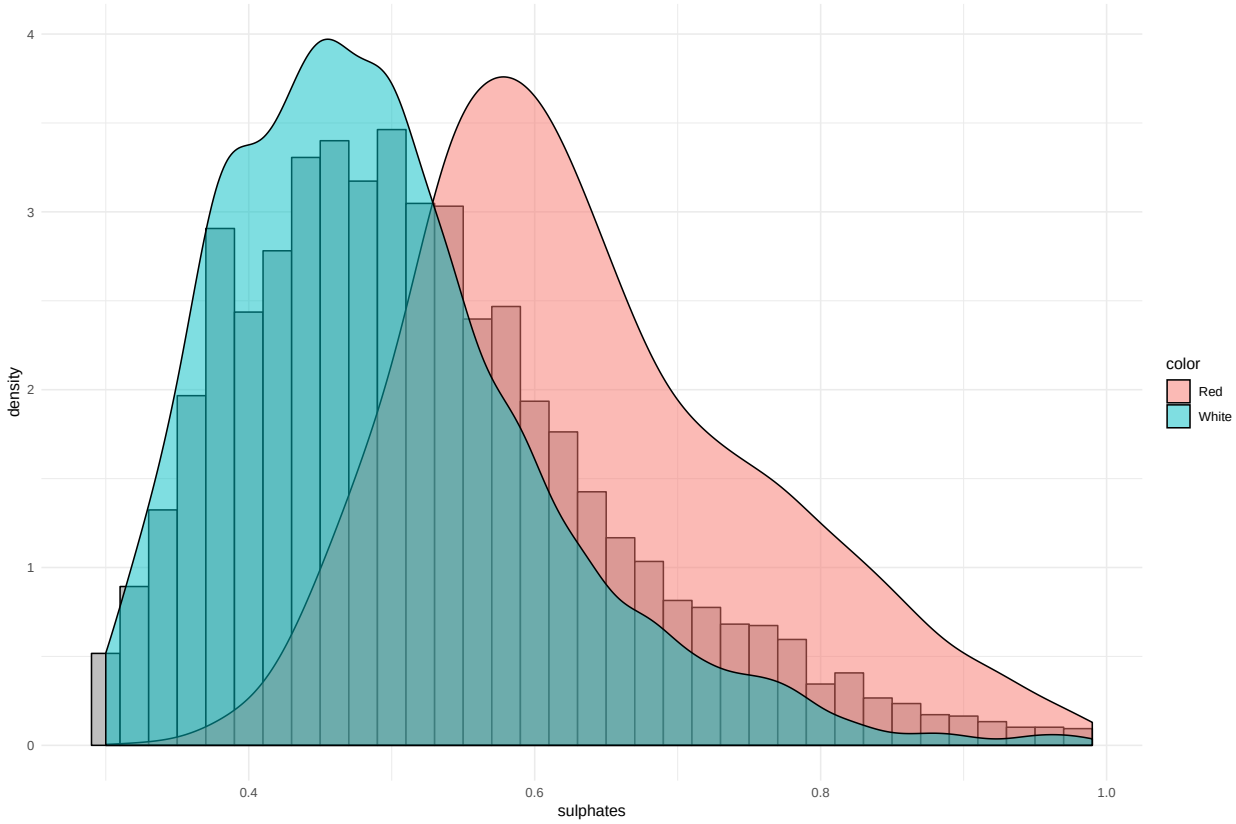
We see more bimodality here. Perhaps in our next steps we will be able to determine why the reds are generally more dense. Could it be that they have more sugar? Less alcohol? Higher concentrations of other solutes?

pH



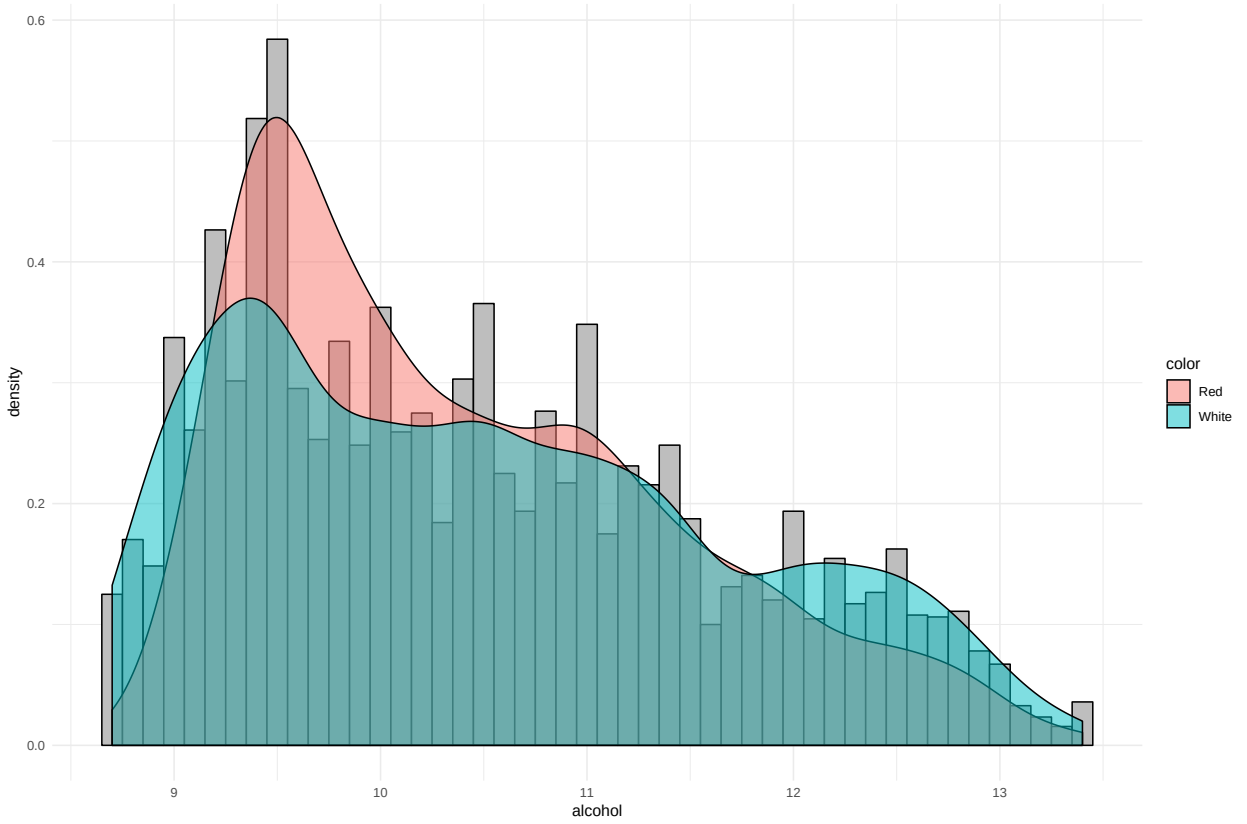
Now this strikes me as unusual. pH is a measurement of acidity, and since both of our "acidity" plots were right-skewed, I would have expected this to be right-skewed as well (higher acidity = lower pH, so the x-scale has been reversed), but this is not the case. If anything, this histogram is slightly left-skewed. This suggests that pH can not be conceptualized as a simple derivation from the "fixed acidity" and "volatile acidity" values. This may be interesting, but further exploration of this will be beyond the scope of this exercise.

Sulphates



Another right-skewed distribution. Levels generally higher in the reds. Not a whole lot to see here.

Alcohol



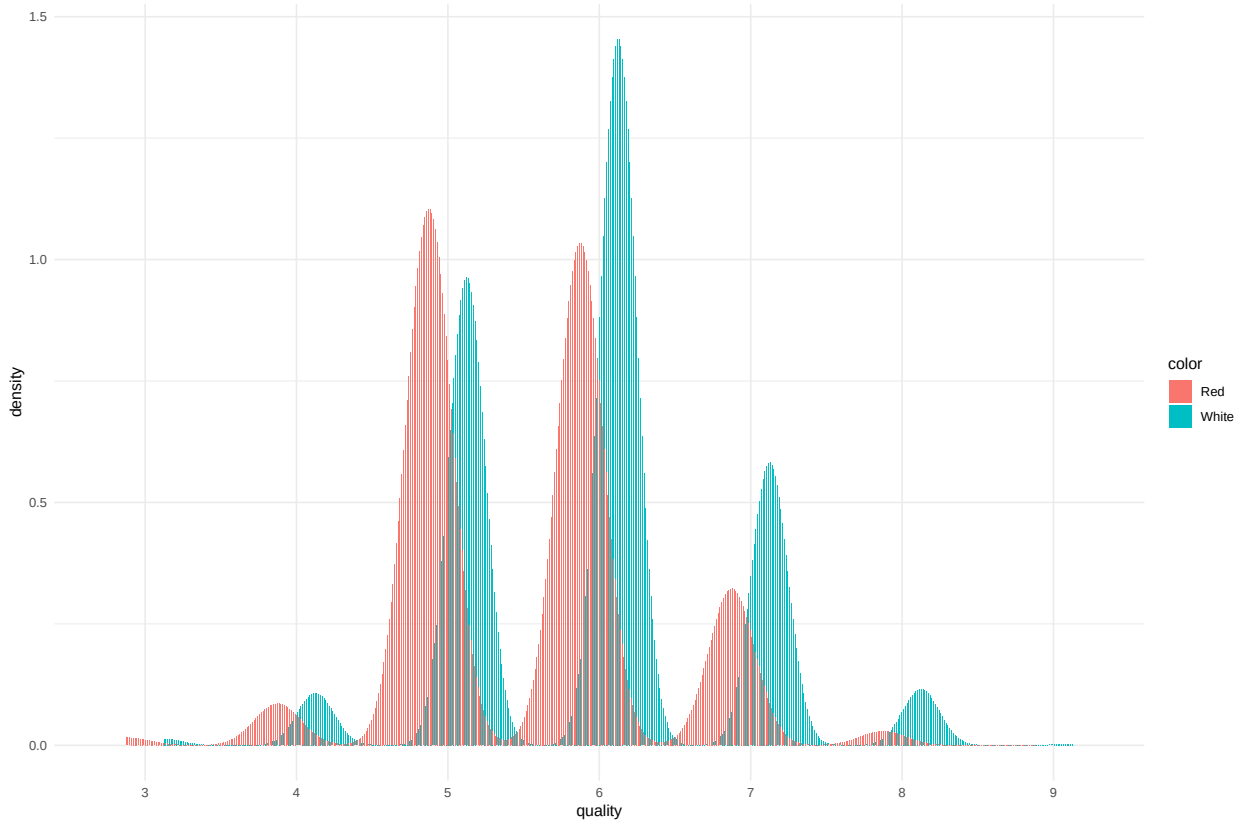
Mean alcohol content is around 10.5, but the most common value was around 9.5. There is quite a bit of variation, with the most alcoholic bottle being almost twice as strong as the least alcoholic bottle. This is the only variable aside from total sulfur dioxide (which was bimodal) to have a distribution with a negative kurtosis value.

Density curves for the reds and the whites are actually quite similar here. Of note however is the hump at 12.25, suggesting that there may be a cluster of high-alcohol content white wines.

While we will not be doing this during our exercise, it might be interesting to break alcohol content into "buckets", thus turning it into an ordered factor rather than a continuous variable. Based on just a crude look at what data we have here, I might consider dividing the bottles into low, mid, and high alcohol content with break-points at 10.5 and 11.75

One last note: There are some odd 'peaks', which seem to be at .0 and .5 values. I wonder why this is.

Quality

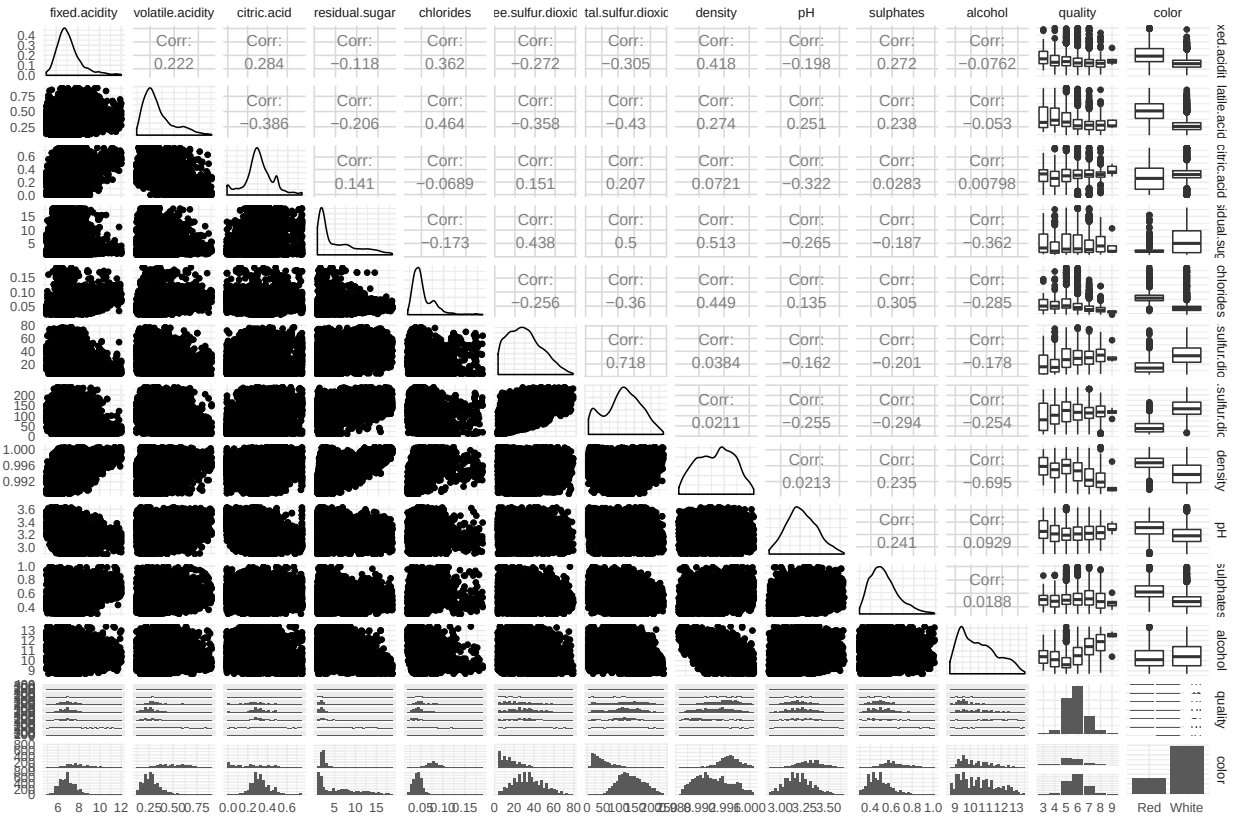


This plot of our final value, quality, shows us that the judges seemed to like this batch of wines generally speaking! They rated nothing below a 3/10, and the most common rating was 6/10 for whites and 5/10 for reds. The white wines were generally perceived to be of a higher quality. No telling why though! It's entirely possible that the judges contributing to this project simply had a personal preference for white wine!

We've now just about exhausted what we can get out of univariate plots. Creating and looking through these plots did help in guiding the development of some inquiries which we'll be able to address as we move into bivariate plots and analysis. Our most relevant analyses at this stage will probably regard how each feature affected the judges "quality" scores.

More Bivariate (other than color) and Multivariate Plots and Observations

Let's go back to our correlation matrix. This time we are interested in details about any particular correlations, so we will use GGally's "ggpairs" function rather than corrgram. This will provide us with bivariate plots as well as correlation coefficients, whereas corrgram only provided us with a visual indicator of where there might be strong correlations.



Along the diagonal, we have essentially the same frequency plots that we looked at individually during our univariate exploration, but including the outliers we previously removed.

Along the bottom and right hand sides of this matrix, we have histograms and boxplots illustrating the variation between red and white wines on each variable.

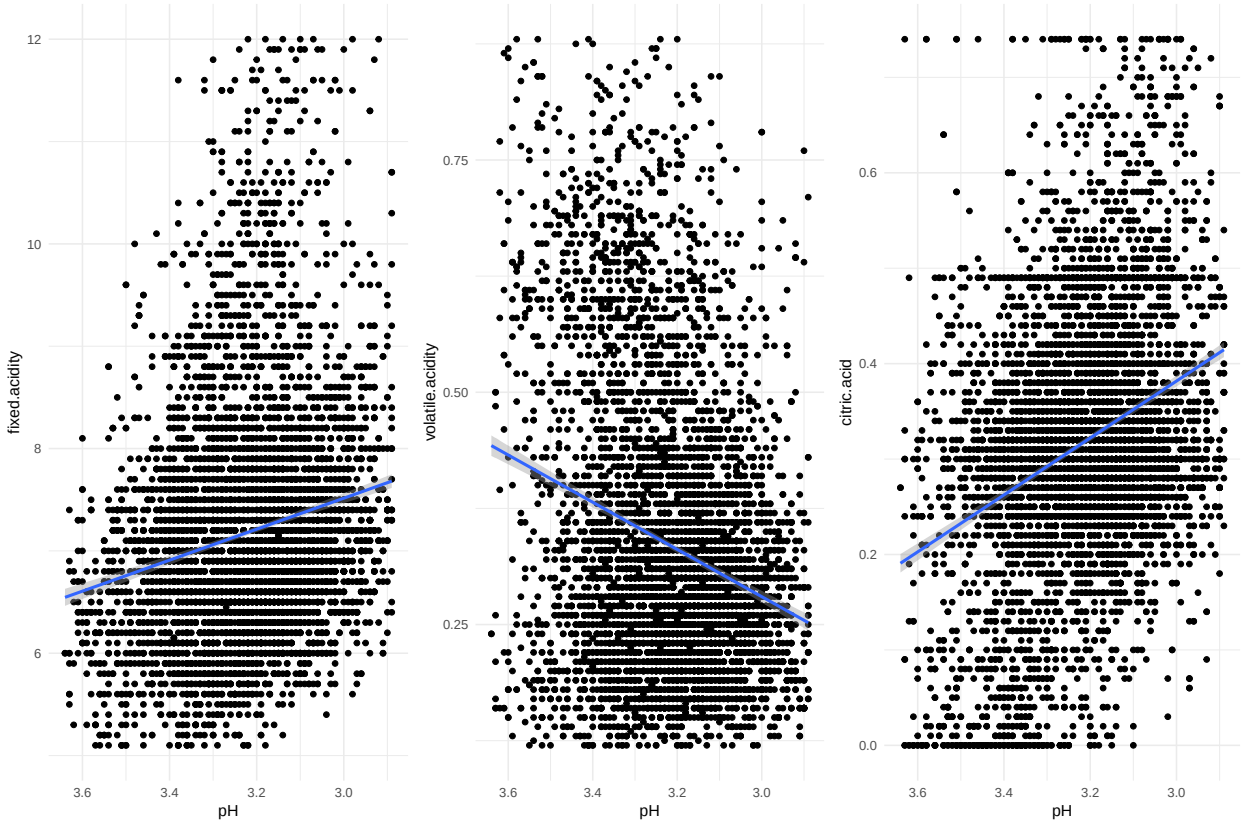
The rest of this plot matrix consists of bivariate scatterplots and their associated correlation coefficients. The scatterplots help us visualize where features may be correlated, and the coefficients provide a nice numeric indicator of just how strong those correlations are.

Correllated features

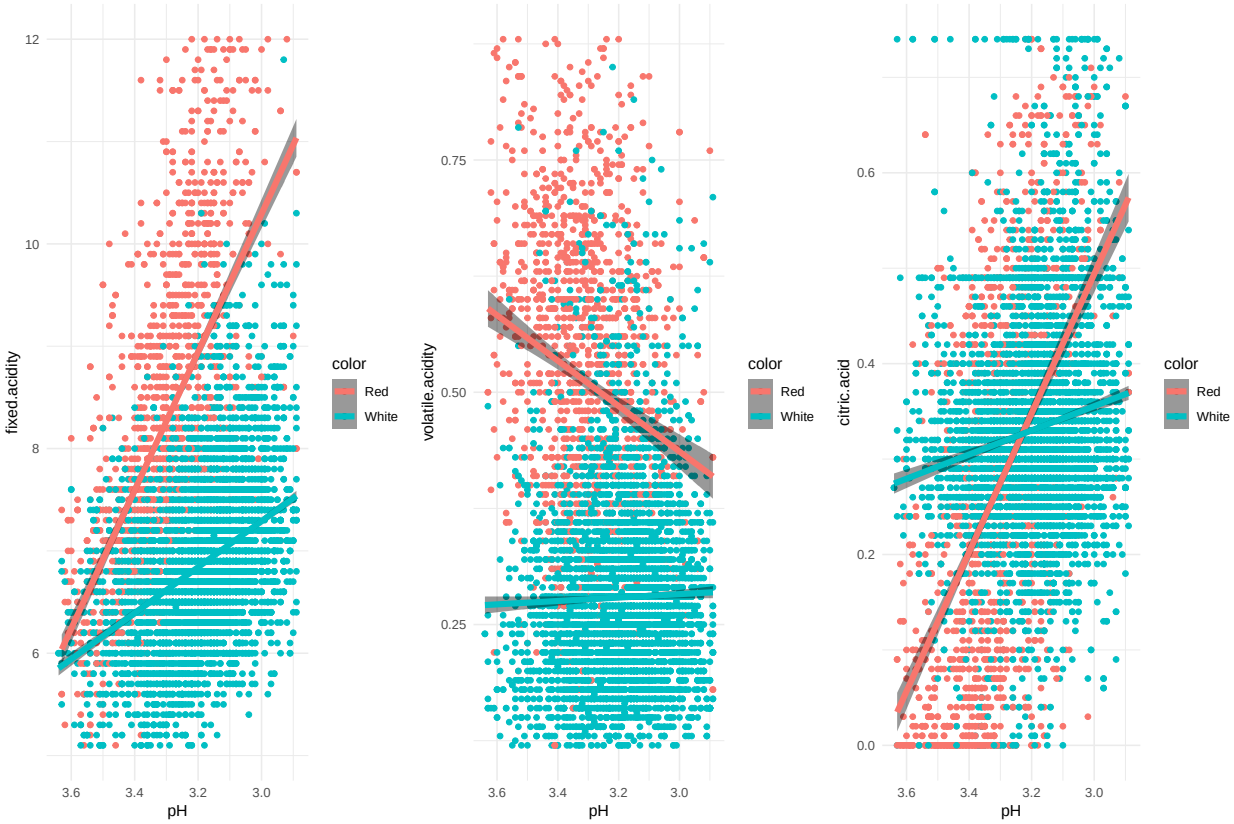
I want to first explore some "clusters" of seemingly associated features before turning my attention to how each feature affects perceived quality. There are three sets of features which seem to lend themselves to a grouped analysis:

1) those associated with acidity (total acidity, volatile acidity, and pH); 2) those which have a strong impact on density (residual sugar, chlorides, sulphates, and alcohol); and *3) the two sulfur dioxide measurements that I noted were highly correlated earlier on.*

Acidity

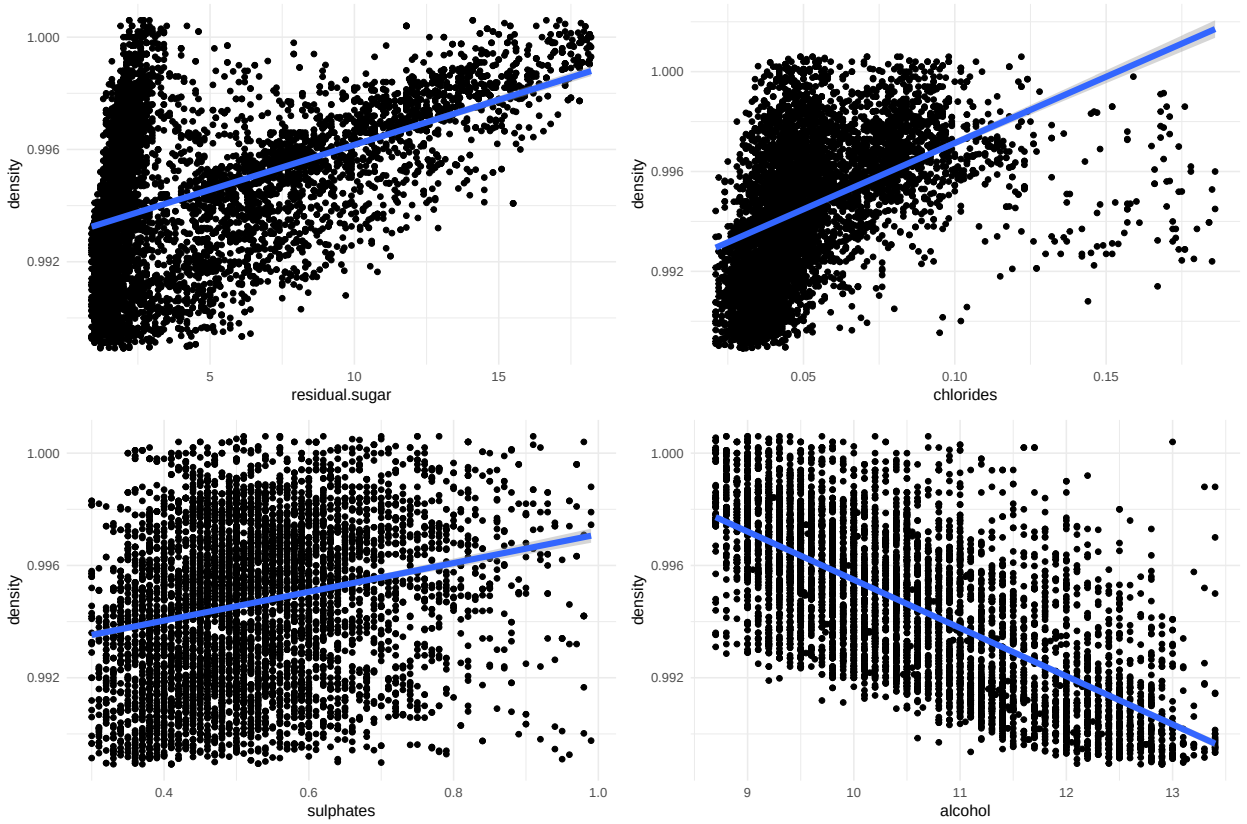


Clearly, the terms "fixed acidity" and "volatile acidity" must have very different meanings. As shown in the top two plots, "fixed acidity" (and citric acid levels) seem to be associated with the classic concept of "acidity" (as measured by pH), whereas "volatile acidity" is negatively correlated with the classic concept of acidity. Lower pH (more acidic) wines actually have *lower* measurements of volatile acidity. We would have to do more research outside of our dataset to learn more about what volatile acidity is actually referring to when it comes to wines.



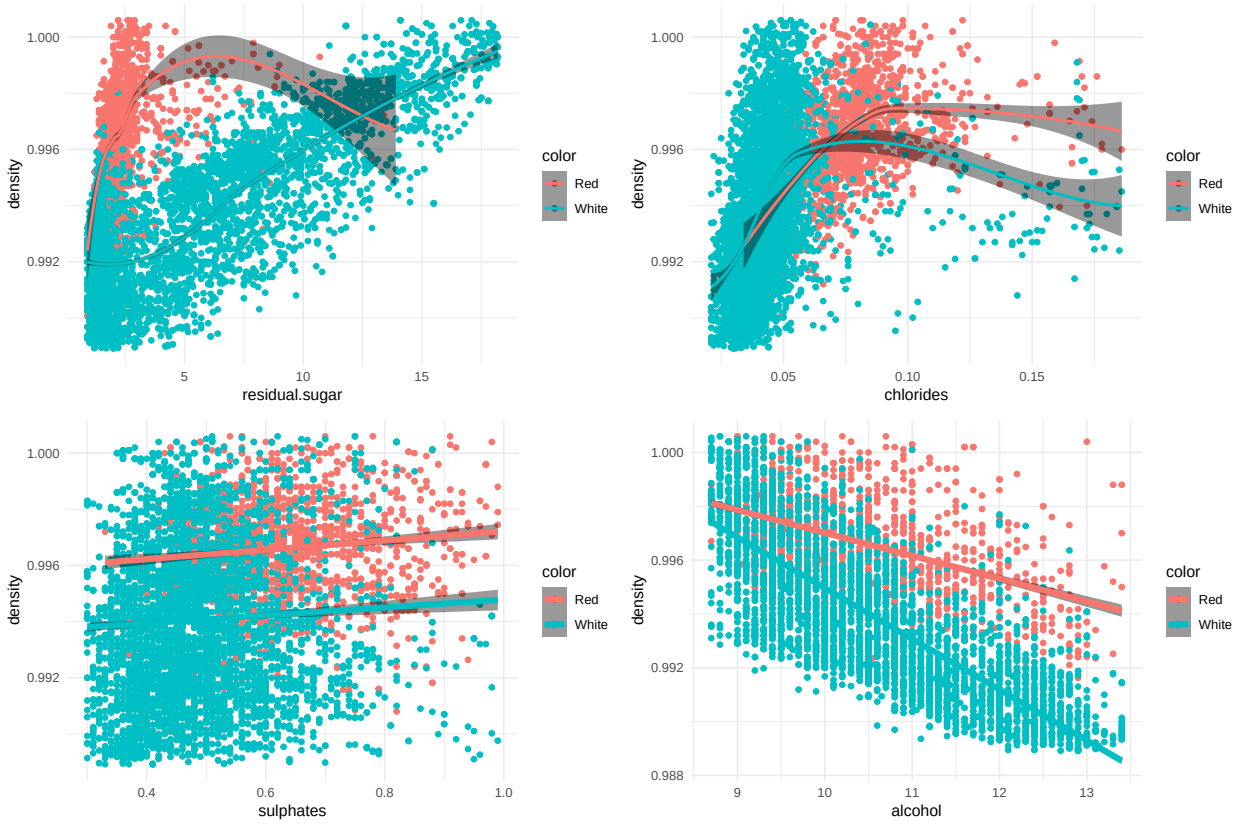
This demonstrates that the links between fixed.acidity and total acidity (as measured by pH), and citric acid and total acidity are stronger in the red wines than in the whites. We can also see that there isn't much of a link at all between volatile acidity and pH in the white wines group.

Density



These observations are fairly straightforward, but interesting nonetheless. Wine is mostly water, and these features all have to do with the ingredients in wine that are *not* water. An increased concentration of solutes (sugar, chlorides, and sulphates) has a predictably positive effect on liquid density, whereas increased alcohol content has a predictably negative impact on the same. (alcohol is less dense than water) The correlation between alcohol and density is much stronger than the correlation between density and any of the solutes. It is actually the second strongest correlation in the dataset, second only to the correlation between the two measurements of sulfur dioxide.

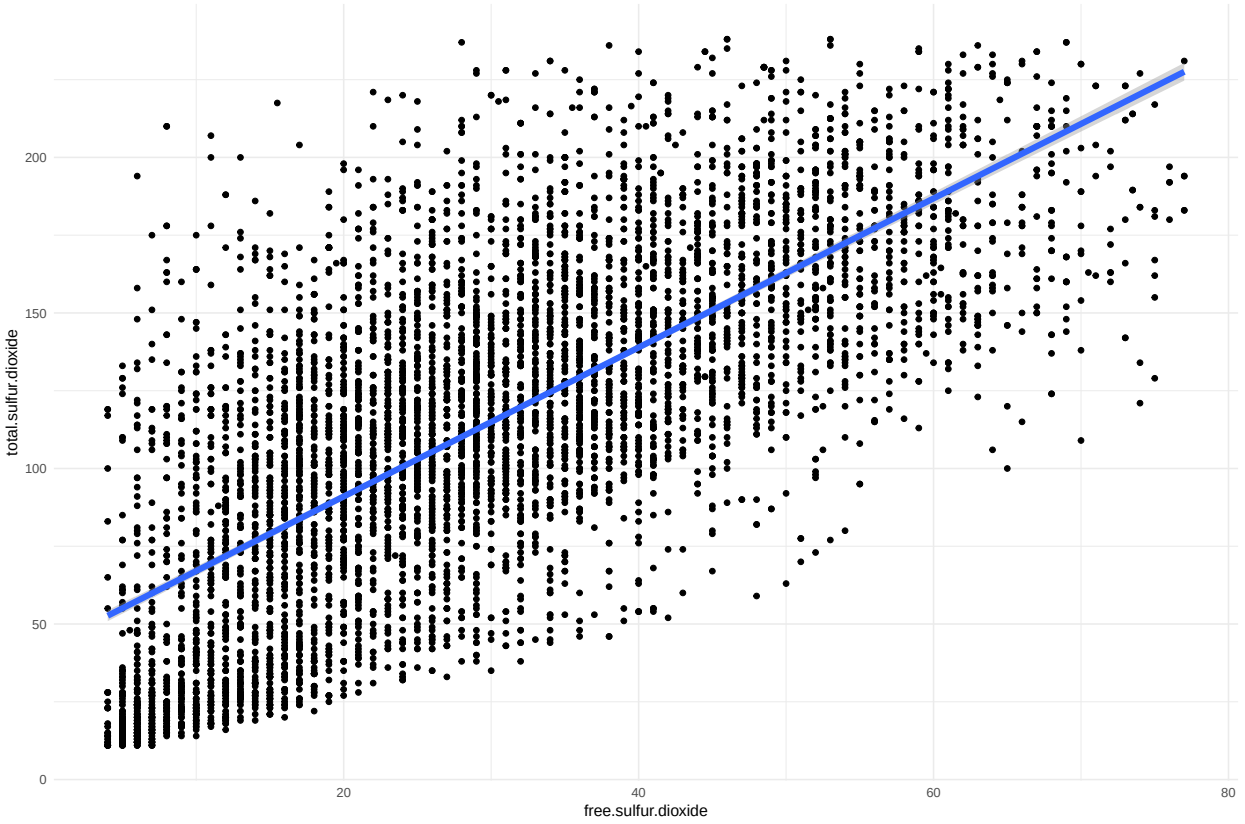
The residual sugar (and to a lesser extent the chlorides) chart is oddly shaped. There almost seems to be two distinct groupings of points, one with a gentle slope spanning the whole x range, and one with a very steep slope in the low x range (0:2.5). I recall seeing significant variation in both residual sugar and chloride levels between the reds and the whites. Let's see if that explains the odd shapes of these distributions.



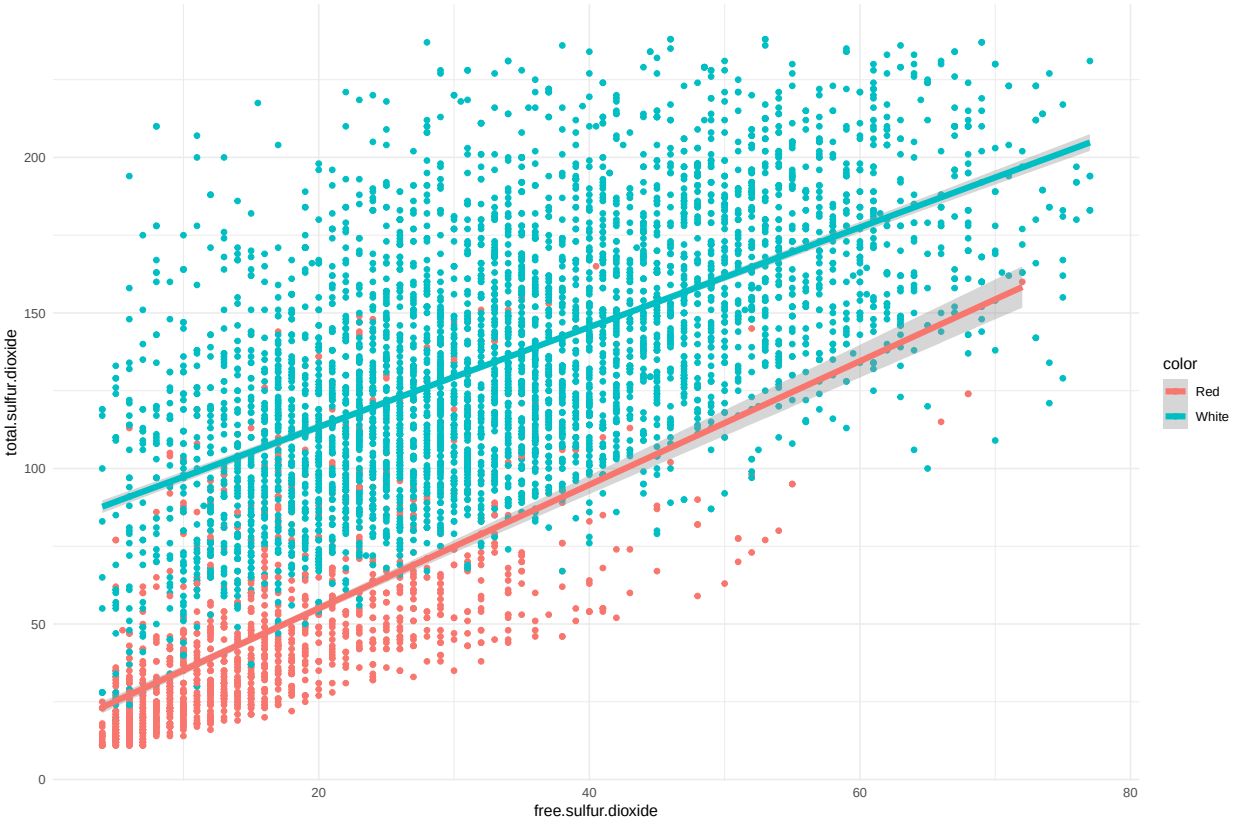
My hunch was correct. The "2-groups" appearance of the residual sugar plot is indeed attributable to differences between the red and white wines. We see much more variation of residual sugar in the whites, whereas the reds are almost entirely low-sugar bottles. We see similar separation in the chlorides plot, but less pronounced.

The bottom plots don't expose much new information, but they do reinforce our earlier findings that sulphate levels and density are both generally higher in red wines.

Sulfur Dioxide



Our dataset's strongest correlation was between free and total sulfur dioxide levels (.72). This seems logical, as "free" sulfur dioxide is a subset of "total" sulfur dioxide. I mentioned earlier that the two may have divergent effects on the subjective quality of wine, but our correllation matrix shows that any such differences are negligible.



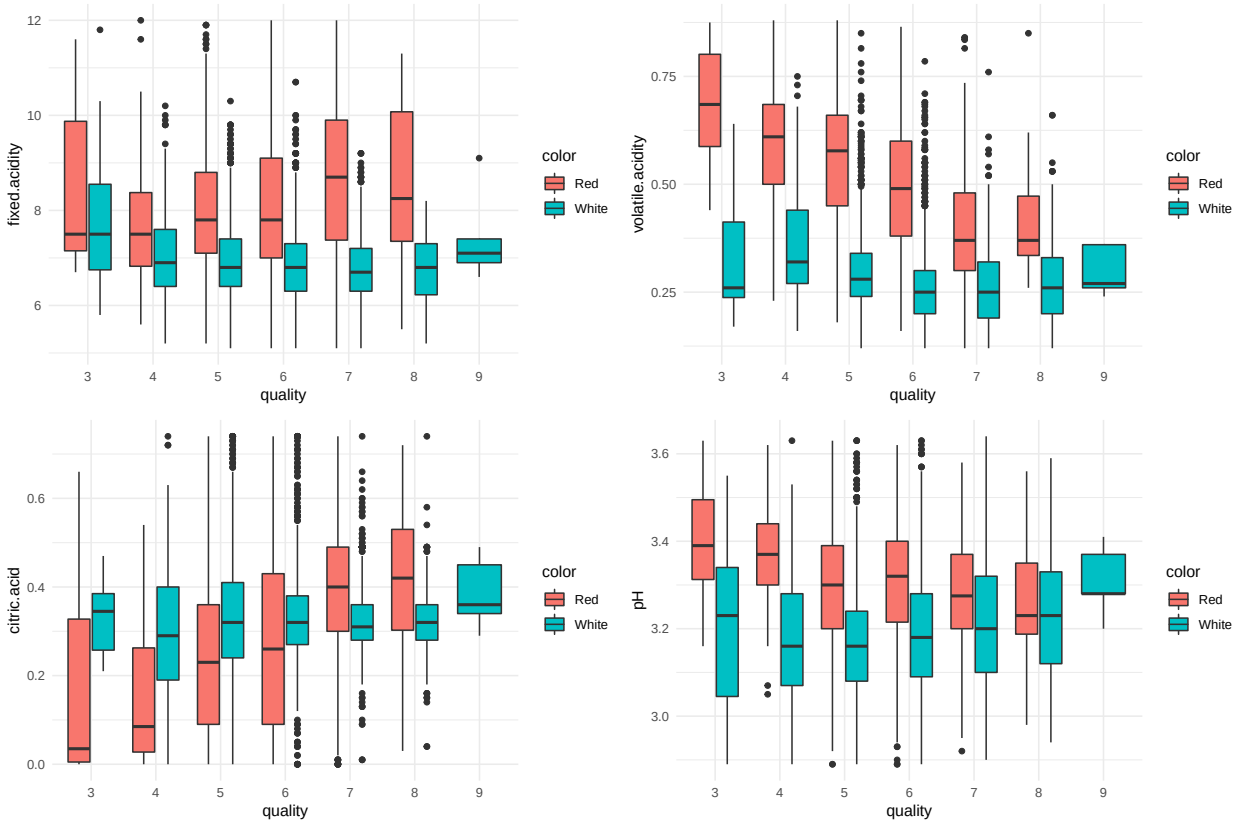
Not too much to see here other than the high concentration of the red wines in the lower left portion of the plot. We already concluded earlier that the reds generally have low levels of both free and total sulfur dioxide relative to the whites.

The impact of select features on the subjective quality of wine.

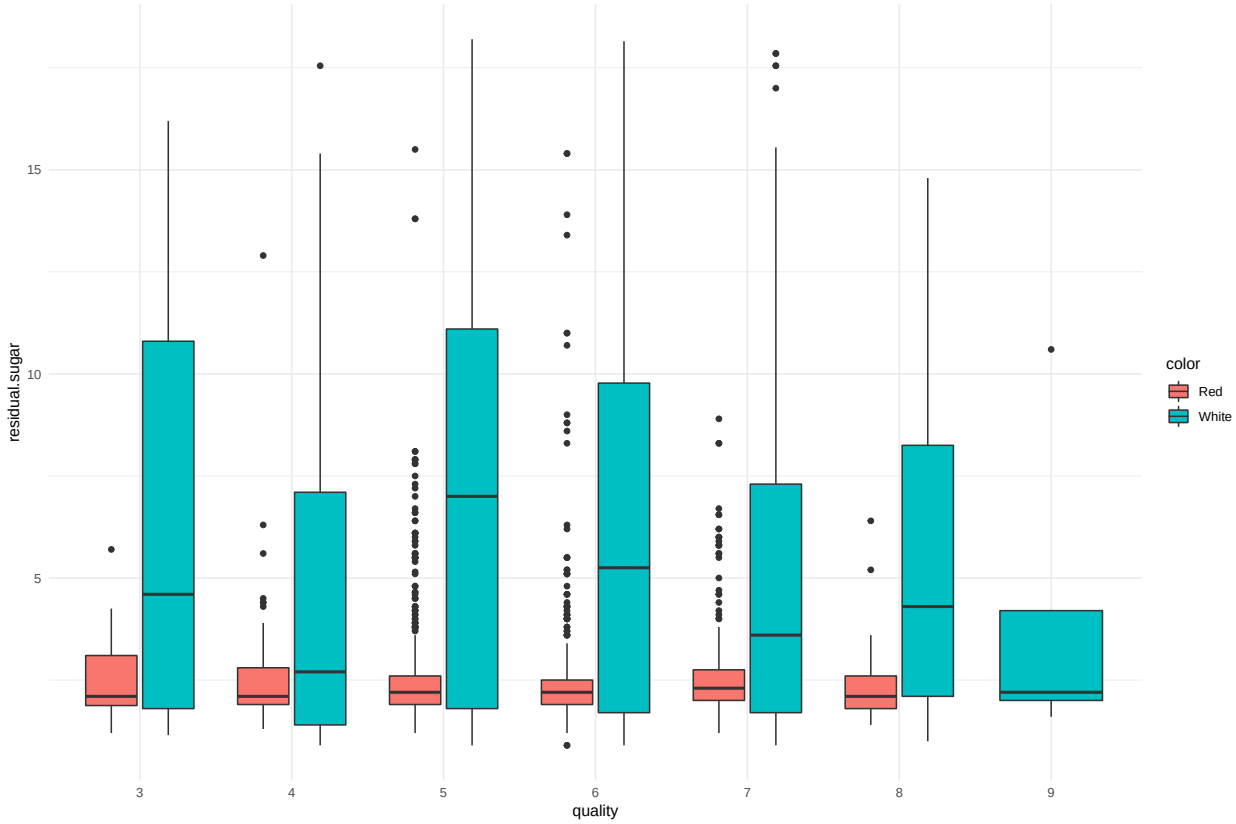
Now on to the most important correlations! Which measurable characteristics of wine are strongly correllated with perceived quality?

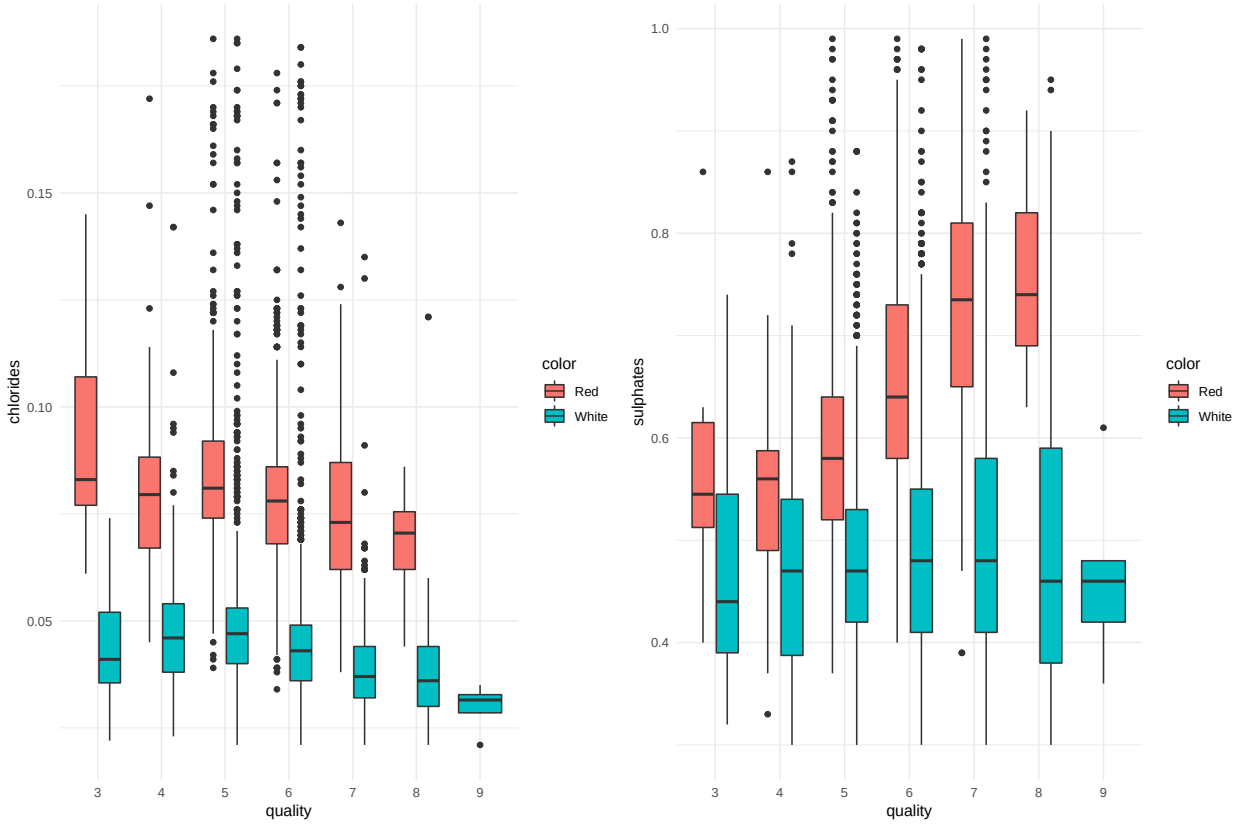
Well... not many it seems. According to our ggpairs correlation matrix, the variables that are most strongly related with differences in the perceived quality of wines are, in descending order: alcohol content, density, chloride content, and volatile acidity (spearman's rho of .439, -.316, -.288, and -.242 respectively).

We must note, however, that the correlations we just identified are from the whole data set (red and white combined). Since we have already discovered so many significant differences between the reds and the whites, we should probably take the time to revisit every variable in this step to make sure we don't overlook a significant relationship because it got masked by our decision to combine both datasets.

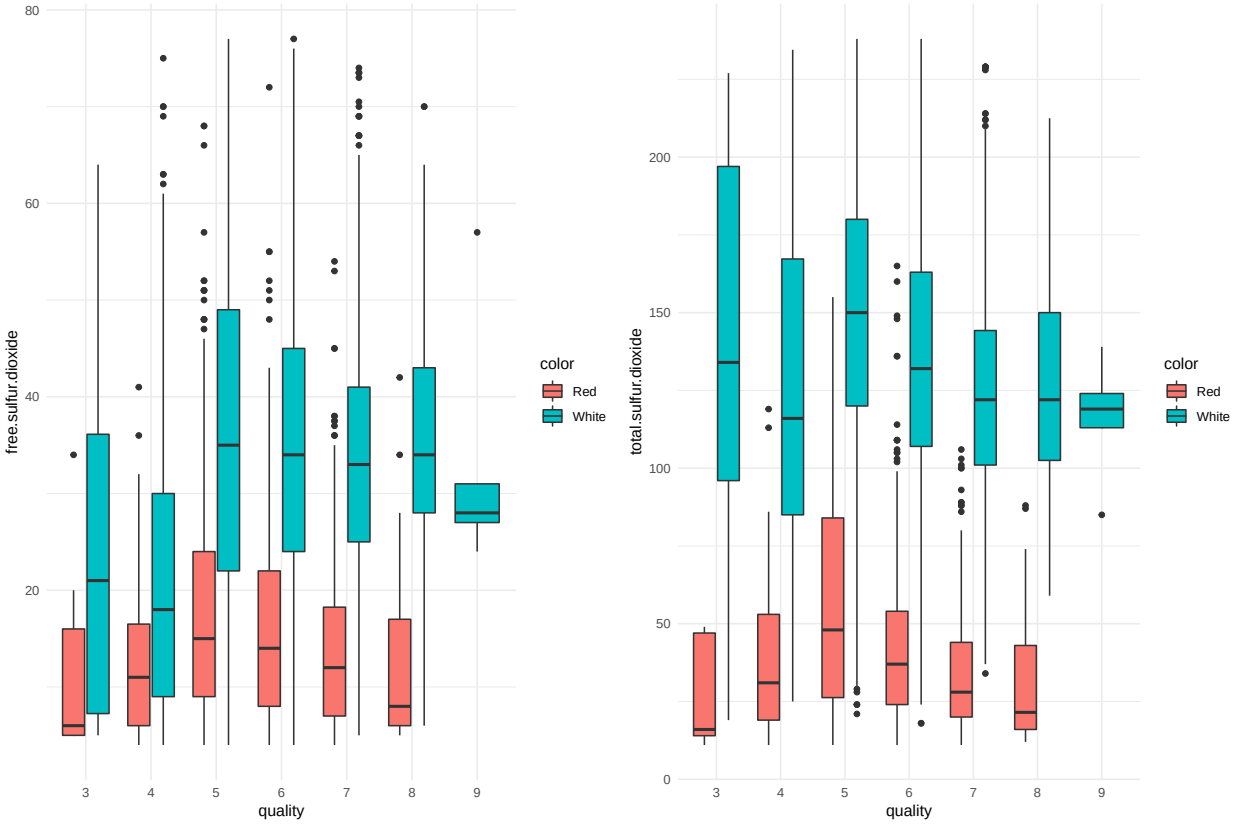


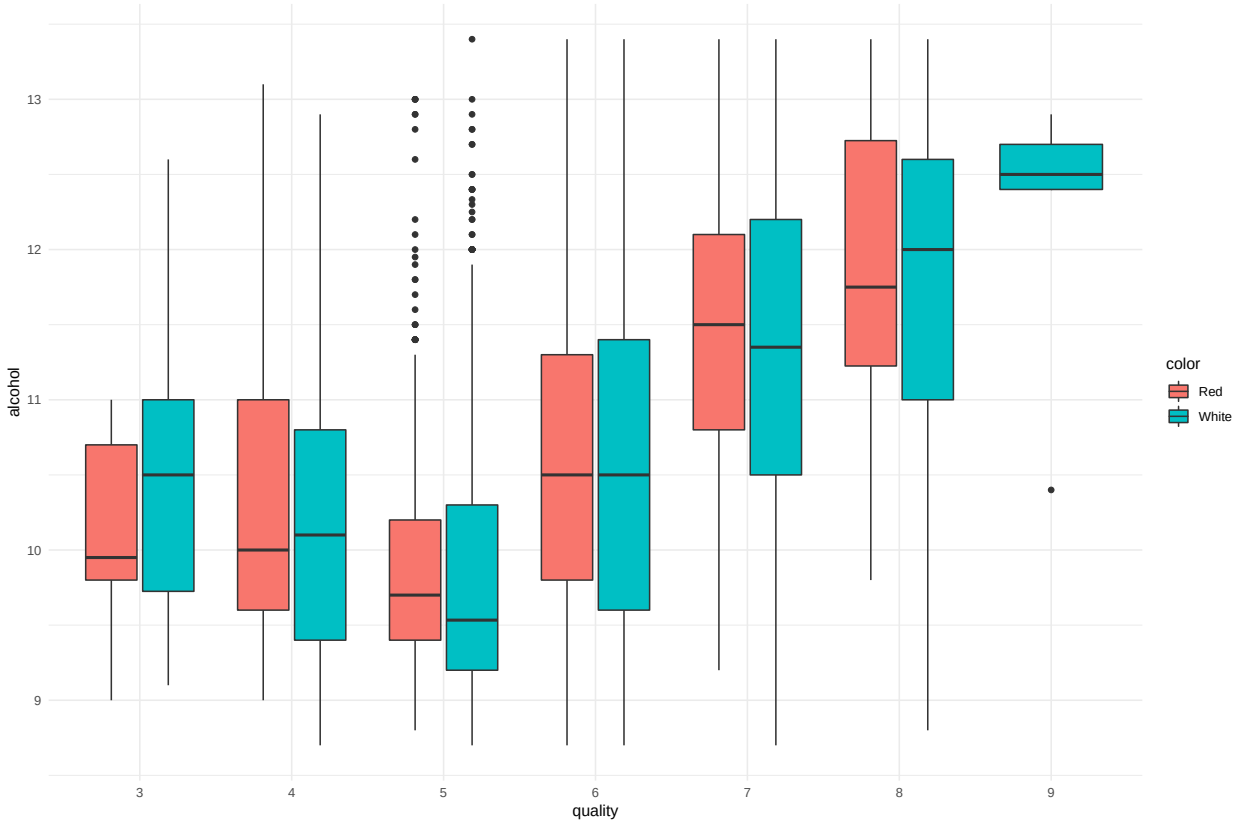
There was less variation in volatile acidity amongst the high quality wines, but there is significant overlap in the IQR at all quality levels. We may need to revisit this in multivariate analysis.


residual.sugar
15
10
5
3
4
5
6
7
8
9
quality
color
Red
White


chlorides
0.15
0.10
0.05
3
4
5
6
7
8
9
quality
color
Red
White
sulphates
1.0
0.8
0.6
0.4
3
4
5
6
7
8
9
quality
color
Red
White

Higher chloride count did not render wines catagorically unpleasant (see high values at quality ratings of 6, 7, and 8), the highest quality wines generally had very low chloride levels.



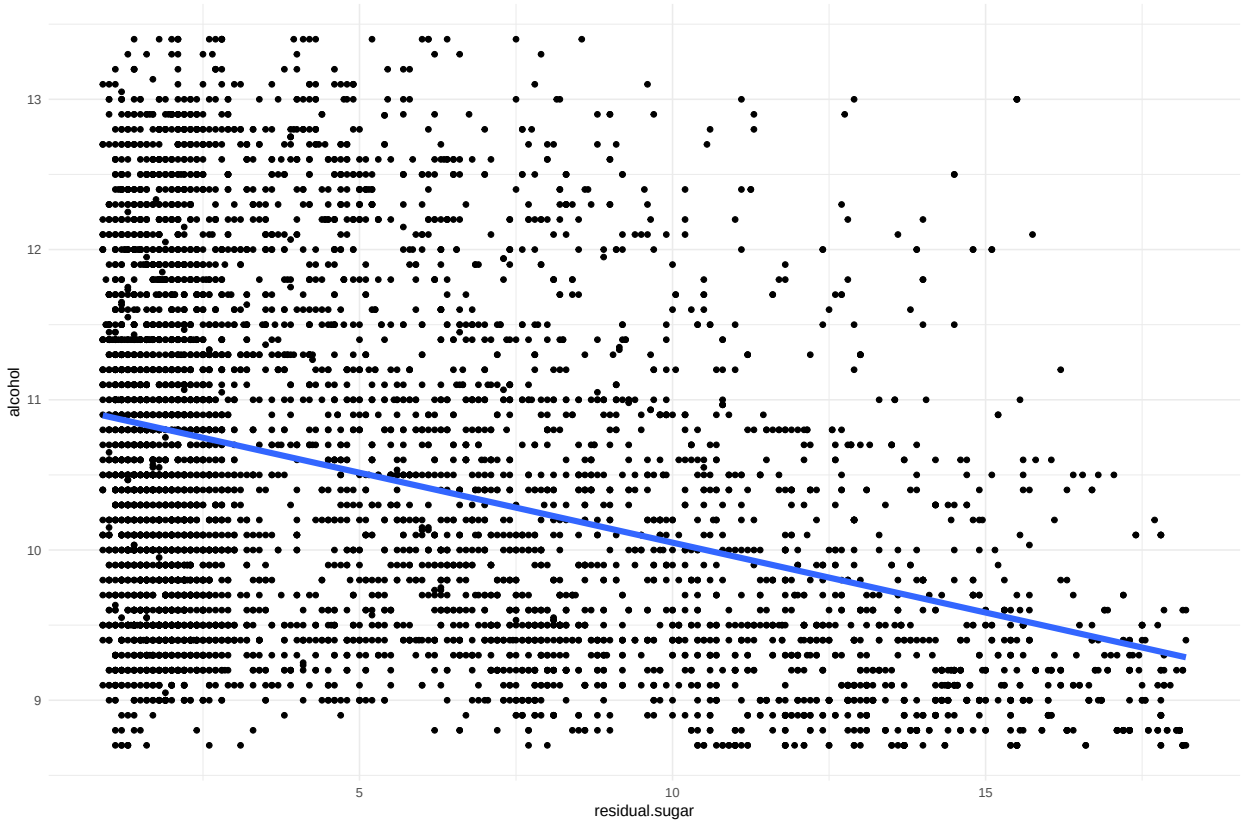


There is a clear relationship between alcohol content and quality, but it is not linear. The alcohol contents of the best wines were almost uniformly higher than lower-rated wines, but wines ranked 3 and 4 had generally higher alcohol content than those ranked 5.

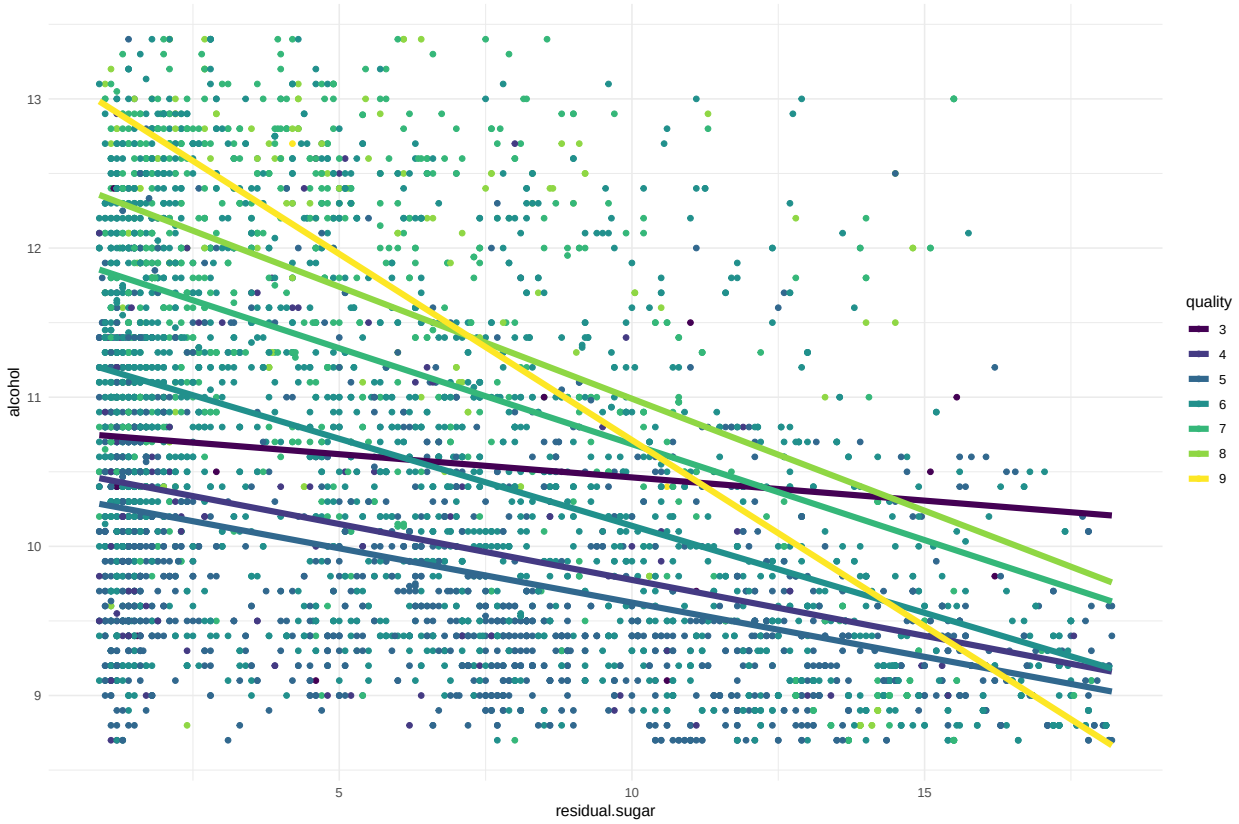
The more we explore this data, the more it seems that there are very few cut-and-dry correlations. To really expose any meaningful trends, we would probably have to employ more complex multivariate analyses and maybe run statistical tests for covariance and interaction effects between different variables.

Before we move on, there is one more thing I'd like to take a look at.

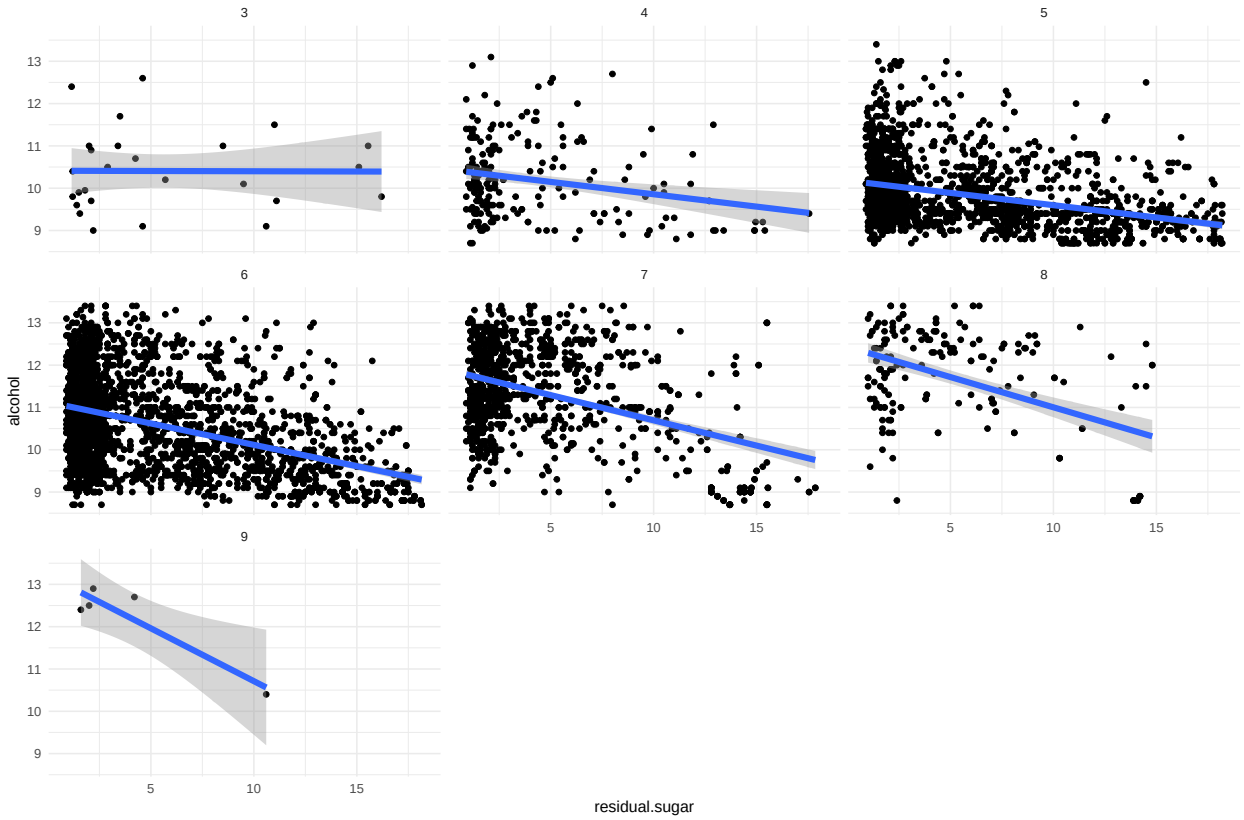
I'm not a wine expert, but one thing that I *do* know is that wine is fermented grape juice, and that fermentation is a natural process by which sugars are converted to alcohol. The longer something ferments, the more of its sugars will become alcohol. So.... it seems to follow that wines with higher sugar content should have lower alcohol content (and vice versa). Let's take a look and see if this is shows through in our data. ()



Well, there seems to be a negative correlation like I predicted, but the data is pretty messy. Let's see if we can do a better job of sorting through this. One thing I know we can probably do right off the bat is to subset out only the white wines (since the reds all have low sugar content). While doing that, let's also see if quality factors into this relationship at all.



There does in fact seem to be some variation between each quality group, but this is pretty noisy so lets get a better look by separating this out into facets by quality.



Well that's interesting! The relationship that I predicted I'd see between sugar and alcohol content seems to be more prevalent in higher quality wines. Now I can't jump to any conclusions from this, but it makes me wonder if some of the lower quality wines might contain added sugar and/or alcohol? I'll jot this down in my notebook of "questions for a wine expert"

The same trend viewed in non-graphical representation of each quality grouping's linear coefficient:

```
## Quality 3 : -0.0012
## Quality 4 : -0.0584
## Quality 5 : -0.0581
## Quality 6 : -0.101
## Quality 7 : -0.1194
## Quality 8 : -0.1429
## Quality 9 : -0.2496
```

Final Plots

One

Two

Three

Reflection

In retrospect, it would have been wiser to keep the red and white wine datasets separate. For stats analysis: 1) dummy variable for color 2) Factor analysis to reduce dimensionality (or PCA, or clustering)