Data Driven Wine

By: Jonathan Grays

28 July, 2020

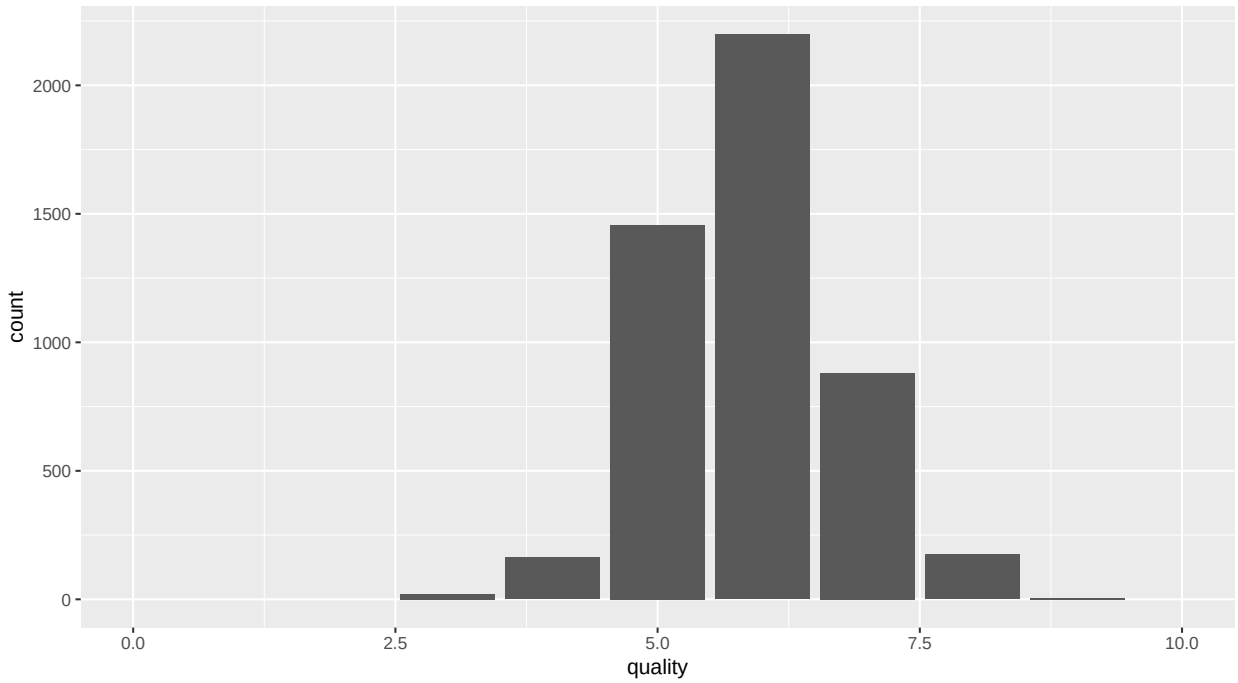
Contents

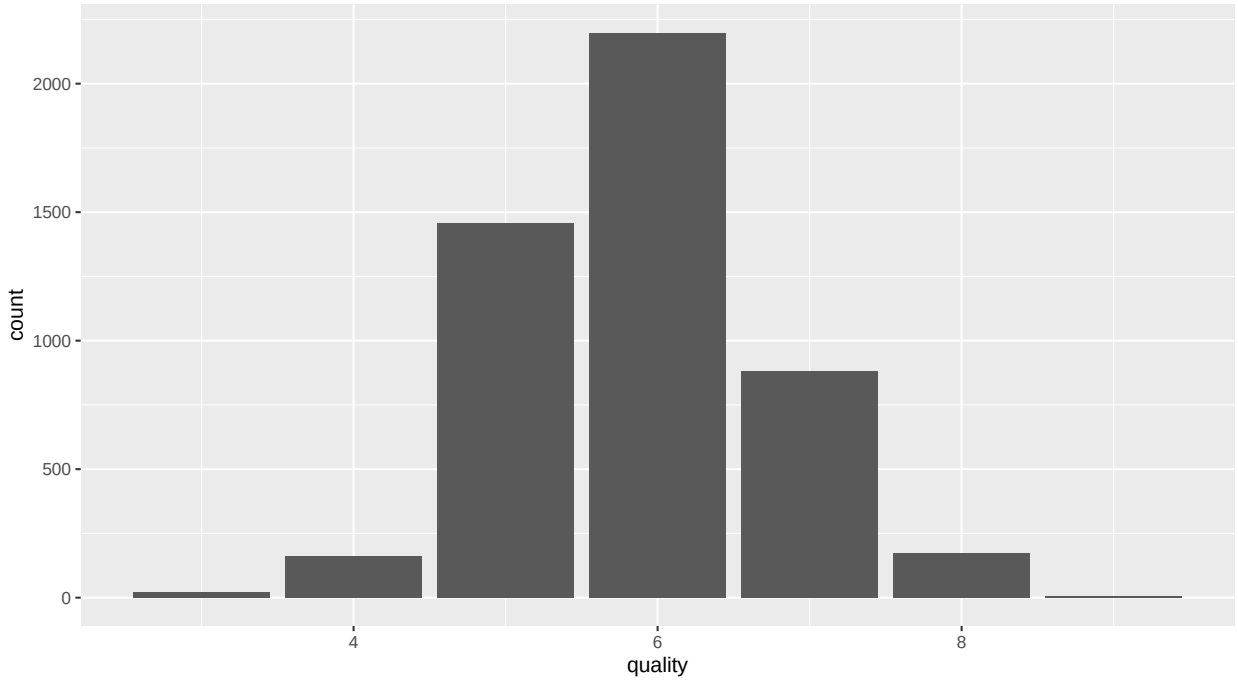
This report will explore a dataset containing chemical attributes for approximately 4,900 white wines.

Univariate Plots Section

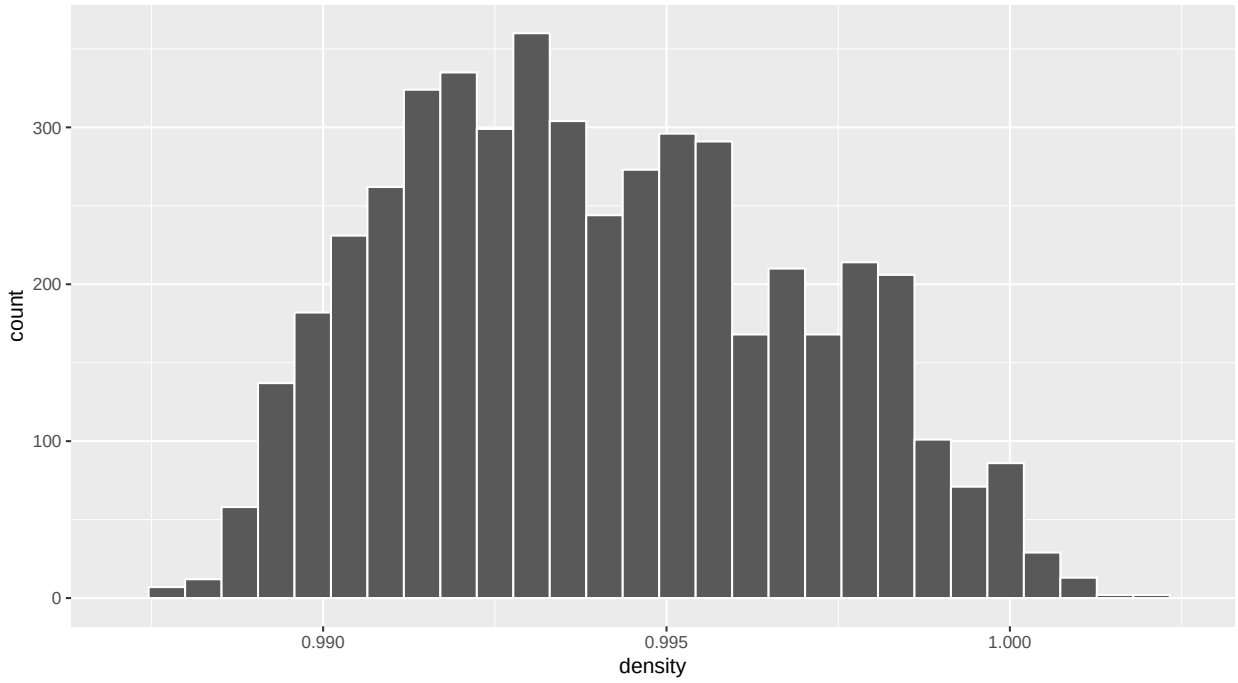
```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
##  Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
##  Mean   : 6.855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

Our dataset consists of 12 variables, with almost 4,900 observations.





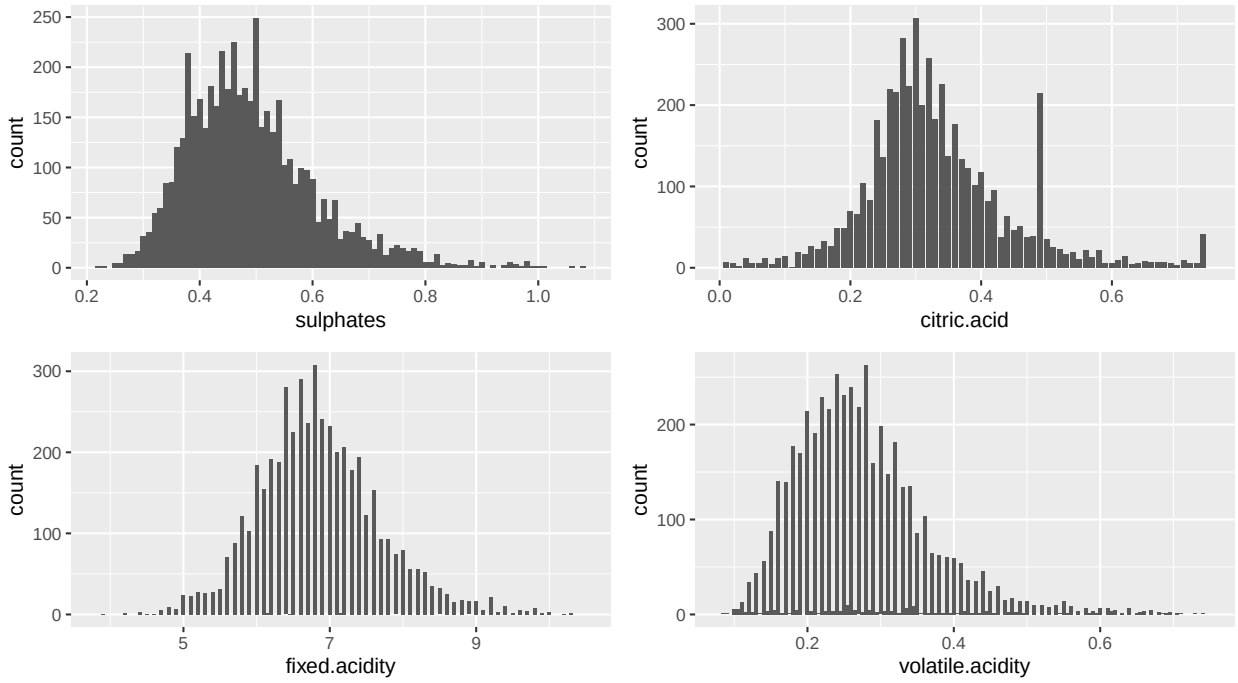
The first variable that caught my attention was “quality”. This looks to be performed on a scale of 1 to 10. Although it is worth noting that the range of this dataset for quality only goes from 3 to 9. My initial thinking pointed to its possible correlations with other variables.



Adjusted the X axis to remove outliers

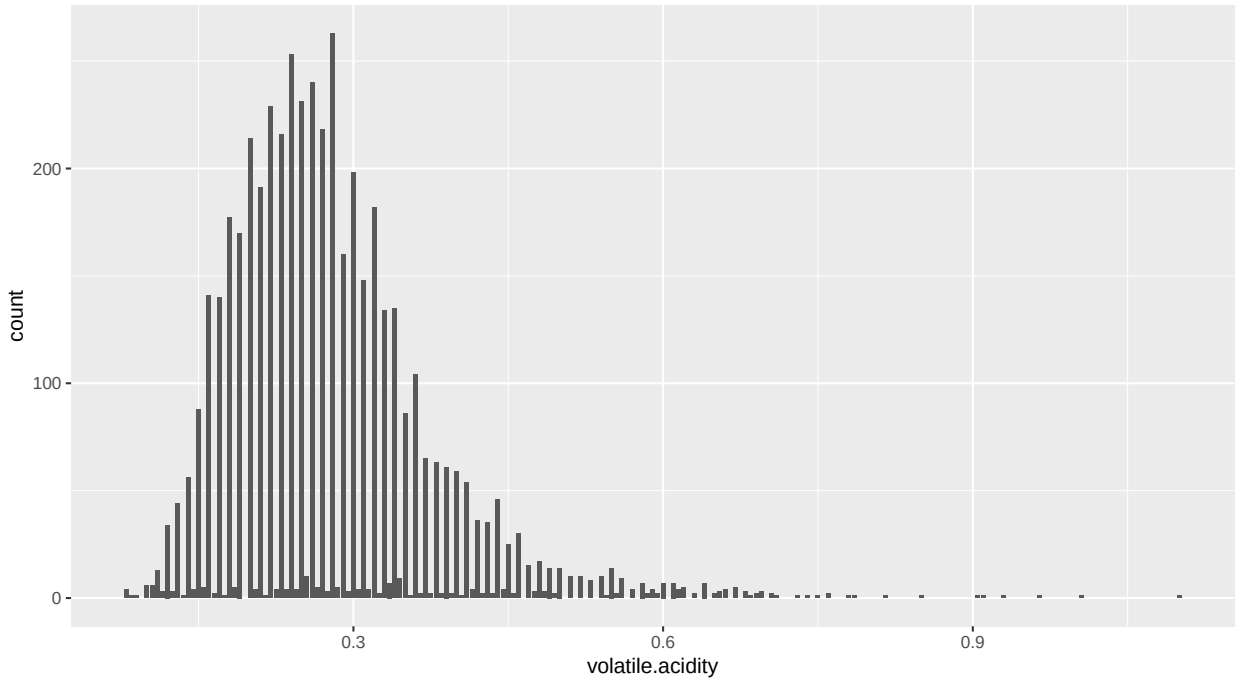
The next variable that seemed interesting was “density”. Measured in grams per cubic centimeter (g / cmˆ3), density shows a fairly normal distribution when you account for outliers. The range for Density in this dataset only goes from .987 g/cmˆ3 to 1.039 g/cmˆ3 but the minute differences in this metric could mean the

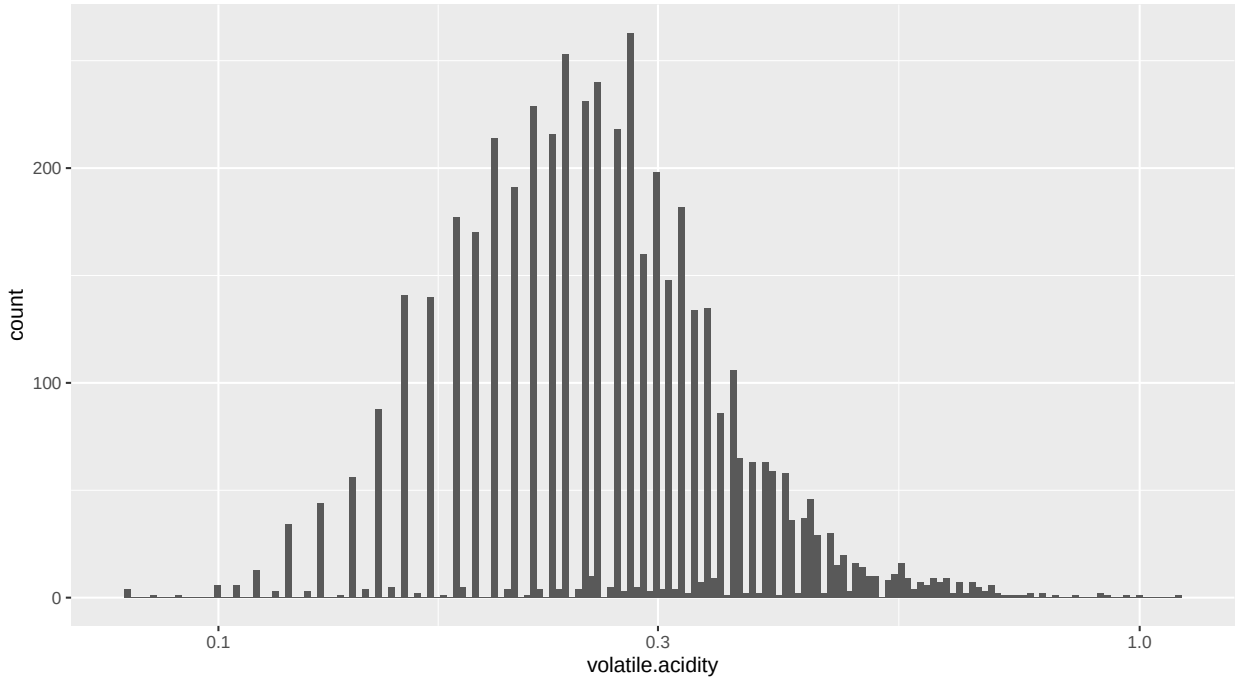
difference between a thicker seeming wine and a thinner wine, depending on the other chemicals included.



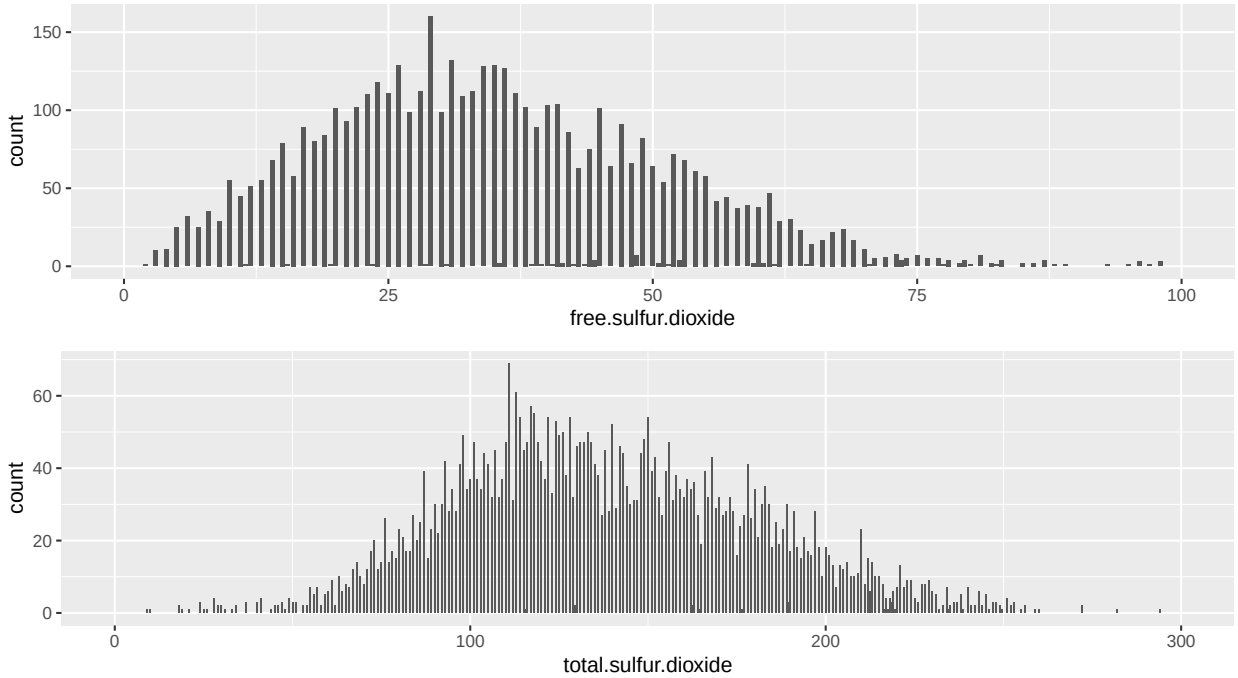
X-axes are adjusted to remove outliers

Above we have a grouping of the differing types of acidity. Since volatile.acidity seems a bit right tailed, I wanted to transform the graph and get a another persective on the data. This is shown below.



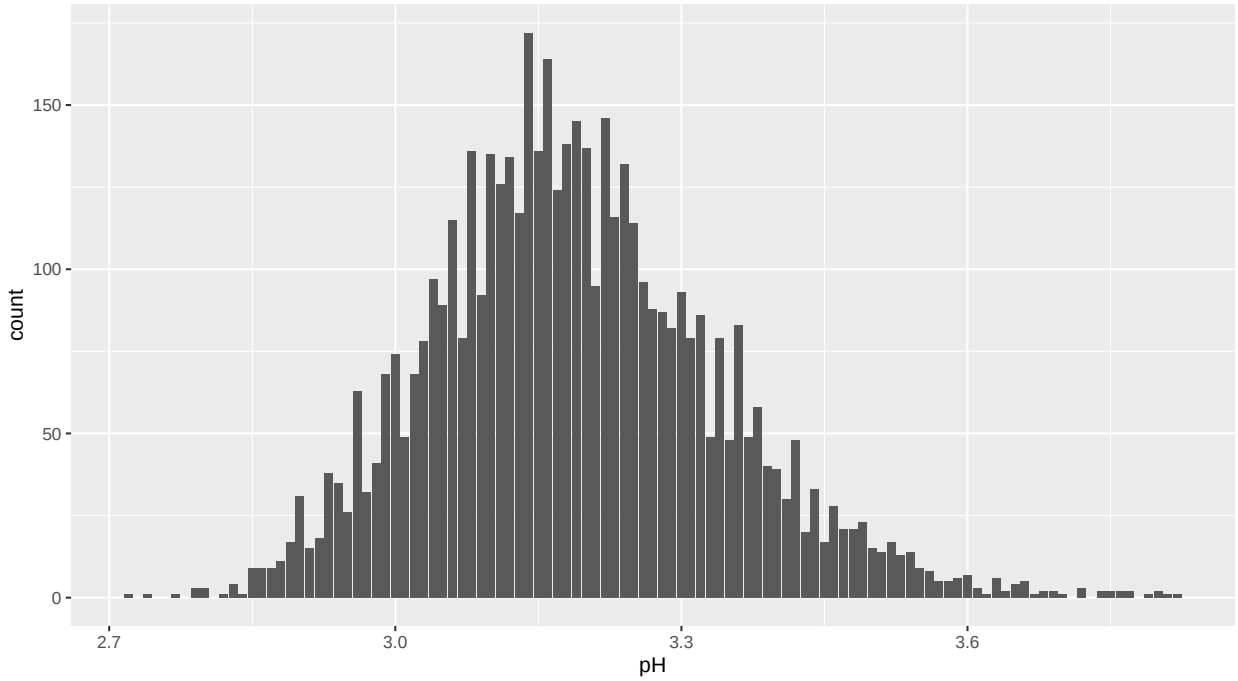


Transforming the data with a log10 function didn't provide much more information aside from more normally distributing the data.



X-axes are adjusted to remove outliers

Here we have a grouping of free sulfur dioxide and total sulfur dioxide graphs. These are normally distributed for the most part, sulphates and volatile.acidity are slightly right-tail skewed.



The plot above shows the distribution of pH values for the wines in this dataset.

Acidity is a very important part of constructing a wine as it can give wine a "fresher" taste when used in proper proportions. Acidity is measured on the pH scale (1-14). Wines on this listed have a pH level ranging from 2.72 to 3.82, this is on the acidic side of the scale.

Univariate Analysis

What is the structure of your dataset?

Our dataset consists of 12 variables, with almost 5,000 observations.

The 12 variables are listed as follows:

- fixed acidity
- volatile acidity
- citric acid
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- density
- pH level
- sulphates
- alcohol
- quality

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

The most likely scale for quality in this dataset is 1 to 10, although it is worth noting that the range of this variable was only 3 through 9. Assuming this scale to be true, most of the wines in this dataset have a quality at or just above average (5-7). Comparatively, very few scored above or below this range.

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

I believe all of the other variables will play at least some part in determining the quality of the wines. Density seems to be another interesting variable as it is not something you generally think of when considering a wine.

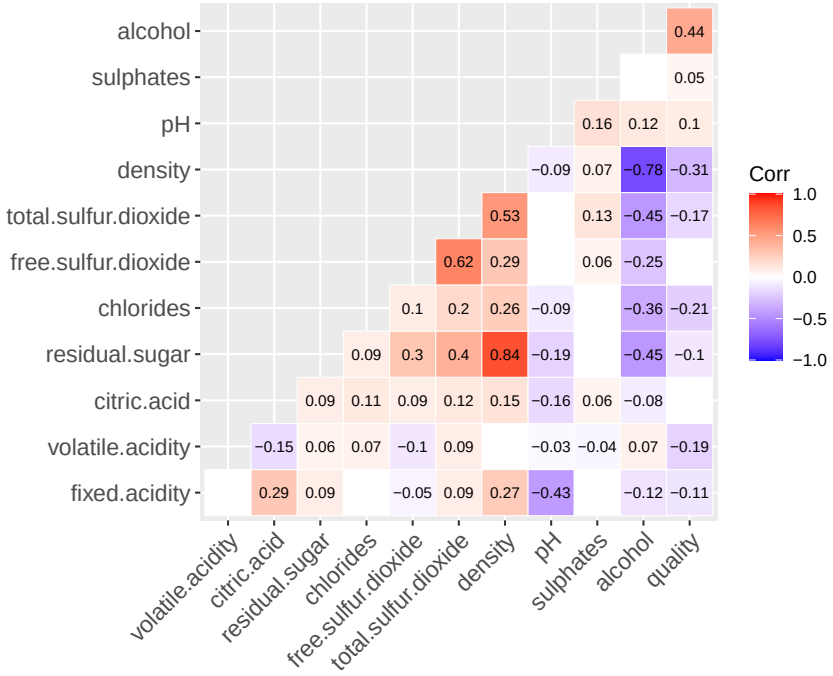
Did you create any new variables from existing variables in the dataset?

I did not find a reason to create any new variables. The existing variables were capable of being self-explanatory.

Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

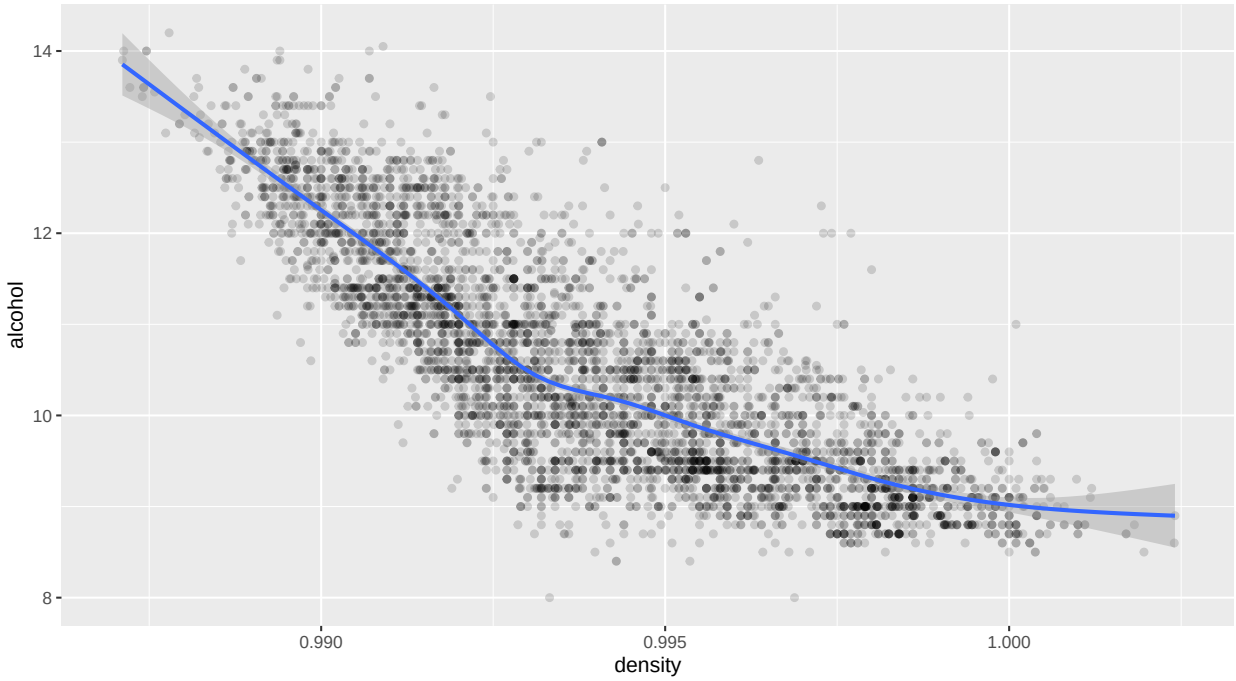
The data itself was tidy enough to work with and did not require any adjusting/changing. The only variable that had at least one zero value was citric.acid, which could make sense.

Bivariate Plots Section

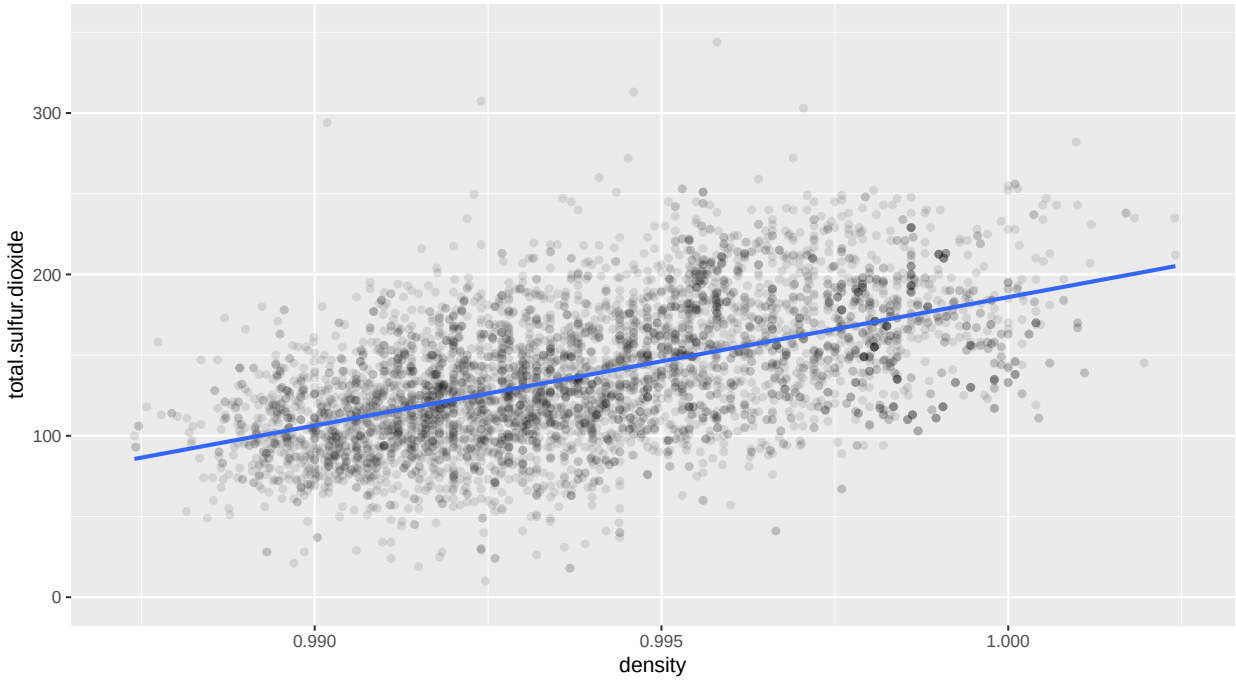


Some bivariate relationships that I would like to explore:

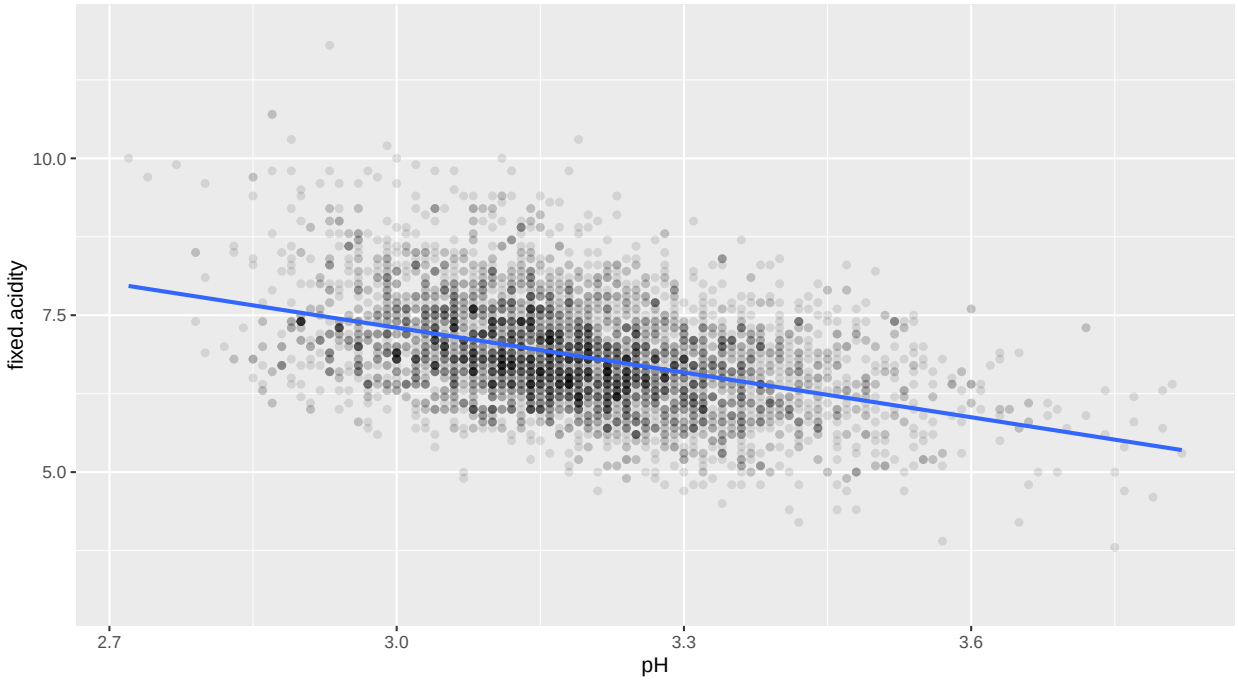
- density vs alcohol content
- density vs total sulfur dioxide content
- pH vs fixed acidity



I used the Loess regression curve in the plot above to illustrate that the data curtails as you approach an alcohol content of 8%.



Adjusted the X and Y axes to eliminate outliers



Adjusted the Y axis to eliminate outliers

Bivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

Exploring different chemicals as they help increase or decrease the density of a wine is interesting. Alcohol, being a very light chemical, generally helps to decrease the density of wine. This is true until you approach being 8% alcohol. This is an important data point as there must be at least 8% alcohol in the mixture for it to be classified as a wine.

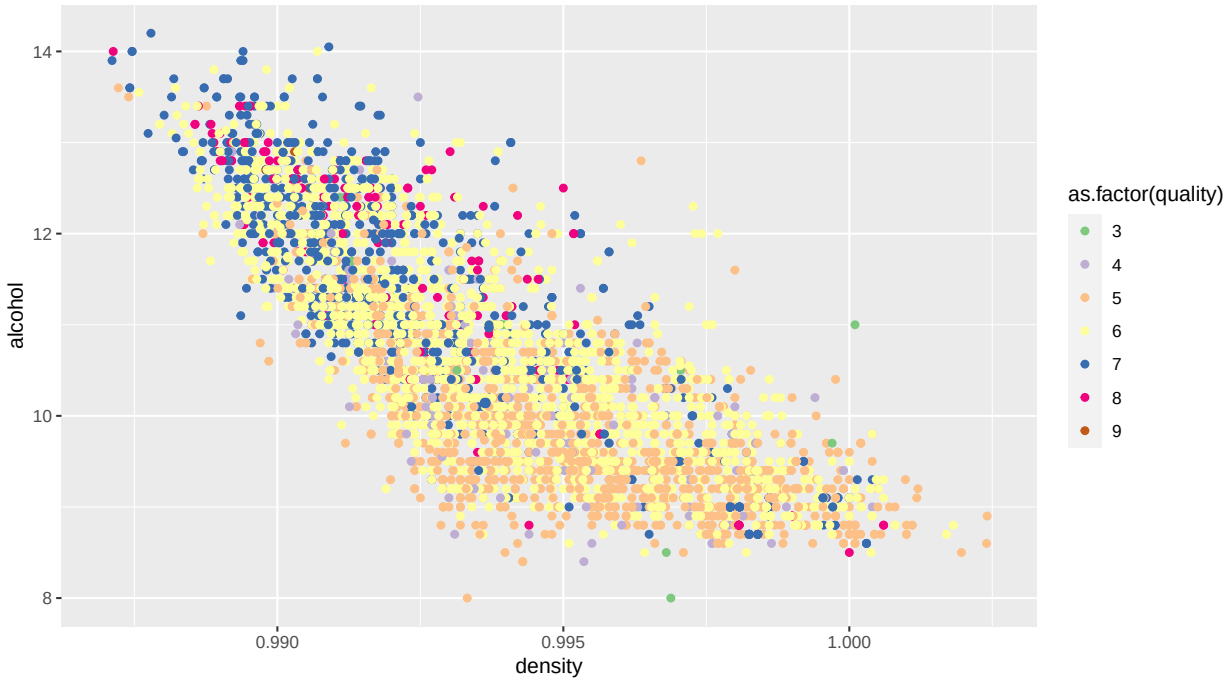
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

Sulfur Dioxide content was also an interesting relationship as it has a moderate/strong correlation with density.

What was the strongest relationship you found?

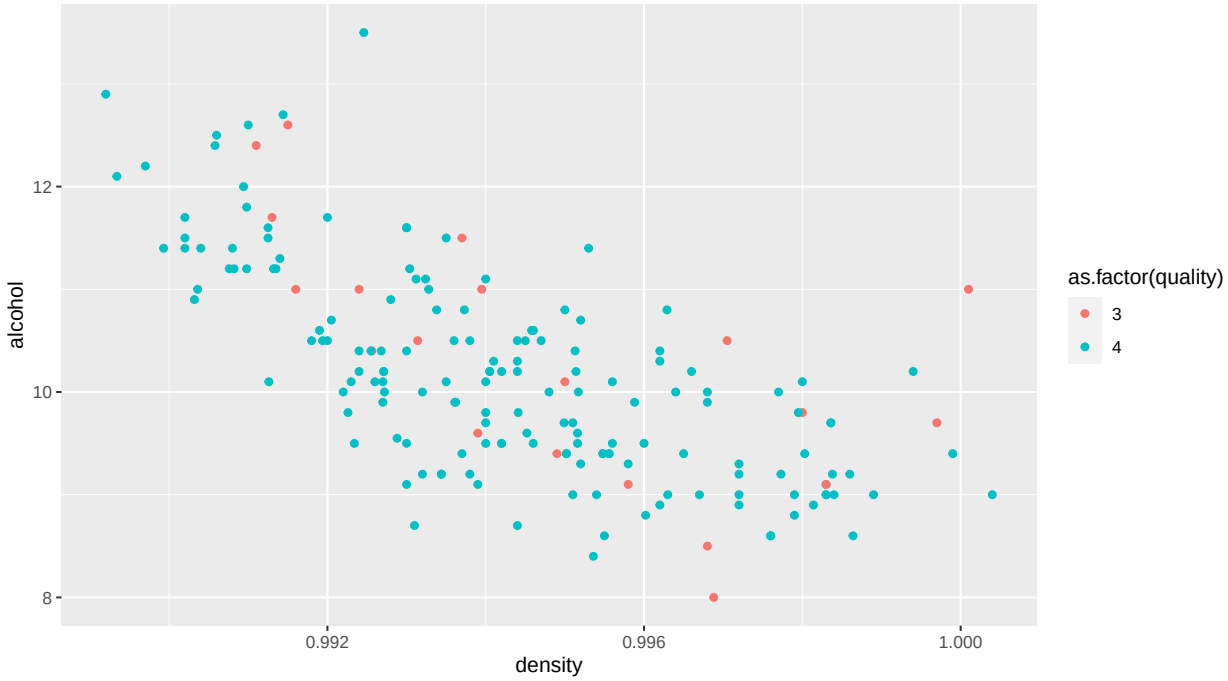
The strongest variable relationship I found was between the white wine density and residual sugar content. This positive and strong relationship makes sense as sugar is a very dense substance. So, adding more sugar should raise the overall density of the compound.

Multivariate Plots Section



Adjusted the Y axis to eliminate outliers

Here we have a scatterplot showing the trends among density and alcohol, broken out in color by the differing qualities.



This plot shows the same information from the previous plot except it is only showing the lowest quality

wines, quality 3 and 4. Notice the large variance for these quality of wines. This, compared to the previous plot, could indicate that higher quality wines tend to stay around a certain (higher) level of alcohol per volume. Whereas, the lower quality wines could be of any alcohol by volume.

Multivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

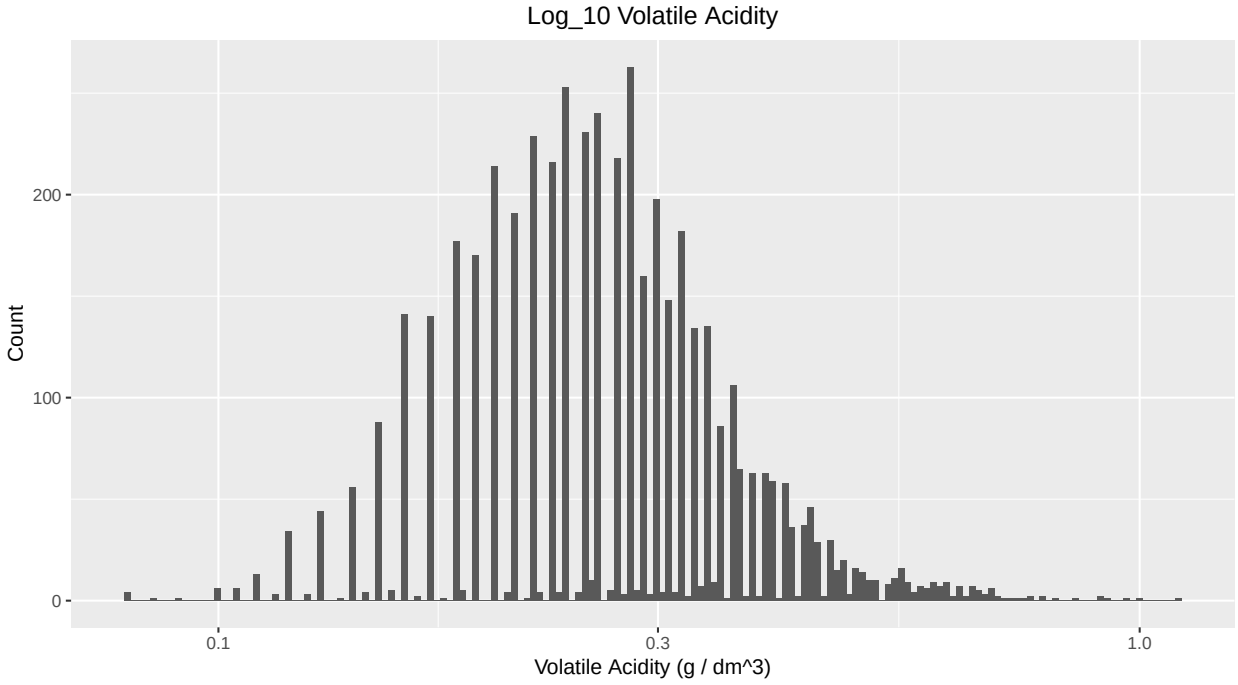
Exploring the scatterplot of Density vs alcohol (as broken out by quality) showed us that higher quality wines (7, 8 and 9's) tend to have higher alcohol content and lower density than lower quality wines.

Were there any interesting or surprising interactions between features?

One thing of interest is that the lowest quality wines (3 and 4), albeit there were only a few observations of each quality, didn't seem to trend at a specific area of the plot. There was quite a lot of variance at these qualities. Alcohol by volume may not be a good way to judge the quality of a wine, unless you are dealing with higher quality wines (by this standard, quality 5 and above).

Final Plots and Summary

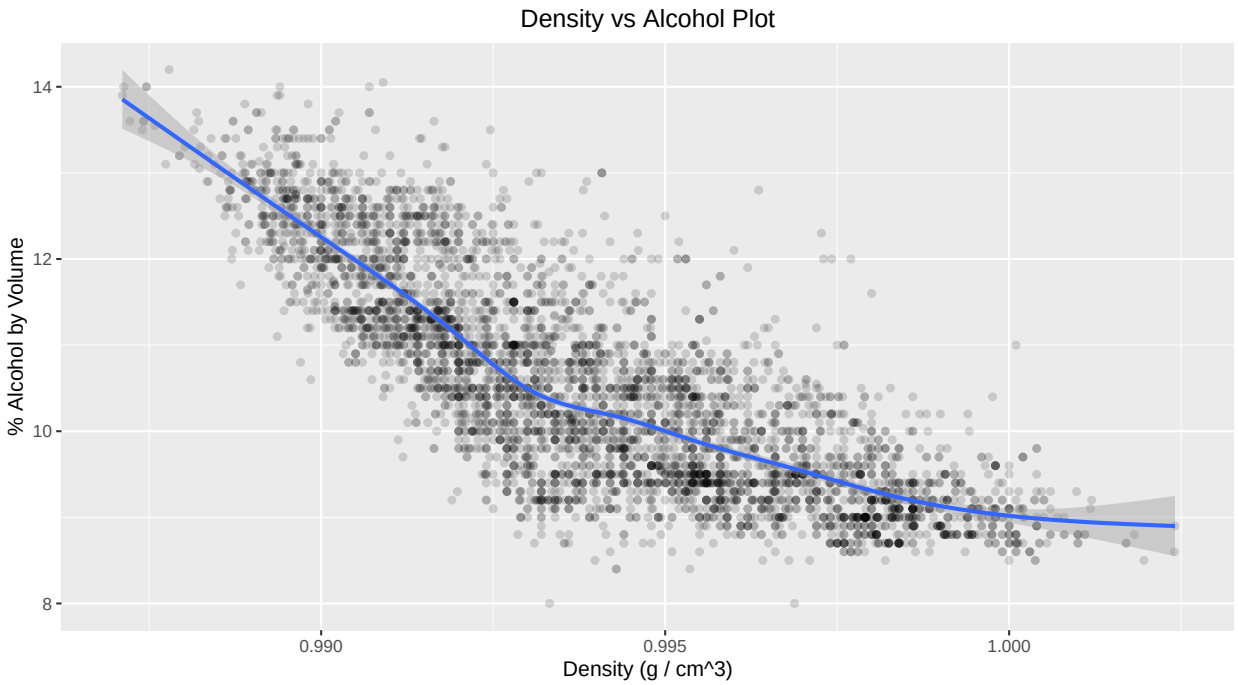
Plot One



Description One

I wanted to explore the distribution of the Volatile Acidity in white wines a bit closer as the original distrubtion was right-tailed. A square root function did not seem to anything useful so I went with a Log_10 function. This showed a more normally distributed section of data. The majority of volatile acid counts fall between .2 and .35 it seems.

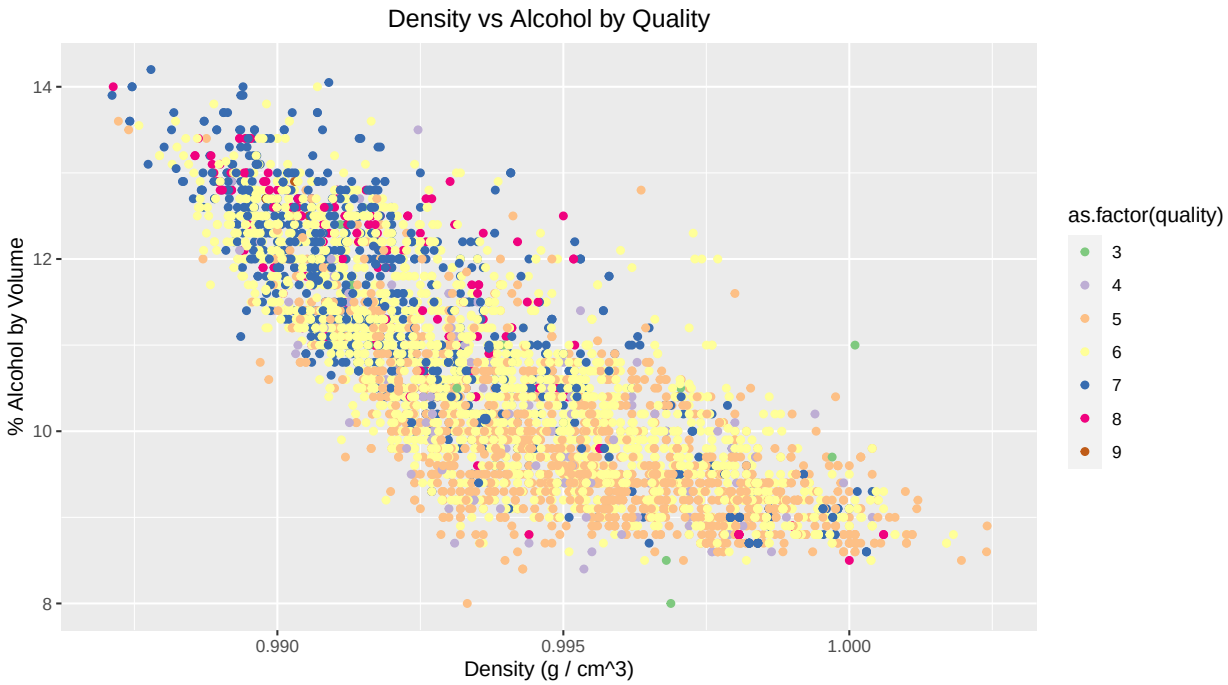
Plot Two



Description Two

One of the plots that I found the most descriptive was the scatterplot used for show the alcohol content by volume vs it's corresponding density. The plot showed a linear negative correlation at relatively lower densities but lessened the rate of correlation as the alcohol content approached 8%. This makes sense as one of the requirements of wine is that it has 8% or more of alcohol by volume.

Plot Three



Description Three

The multivariate scatterplot exporing alcohol and density distributions by quality showed a very interesting set of relationships. The higher quality wines tended to have a higher alcohol content by volume and a lower density (going back to other plots does tend to make sense). Whereas, the lower quality wines did not have a tendency. Of course, it is important to remember that our previous findings determined that variables alcohol by volume and density have a negative linear relationship.

Reflection

By and far the most interesting information taken away from this analysis is how the different chemical additives can influence attributes of wine, such as its density or quality. Also, there doesn't seem to be a clear and quick way to determine the quality of a wine as I had origianlly thought there would be going into this project. Quality seems to stem from several factors working together in a balanced chemical formula that can be modified to suit the consumer's taste.

For future dives into this dataset, I would say including the price of wines would be helpful. Pricing information as it pertains to variables such as quality, alcohol and sugar contents could be quite interesting to look into later.