Red Wine Dataset Analysis

Patrick Flynn

6/3/2019

Introduction

The dataset used in this analysis is the red wine dataset described in Cortez et al., 2009. The red wine dataset contains 1599 observations of red wine samples. For each sample, chemical analysis was completed and the wine was rated by wine experts. The wine experts rated a particular wine on a 1-10 scale represented in this dataset by the *"quality"* variable.

Variables in Dataset (Described in data)

Input variables (based on physicochemical tests):

1. fixed acidity (tartaric acid - g / dmˆ3)
   - most acids involved with wine are fixed or nonvolatile (do not evaporate readily)
2. volatile acidity (acetic acid - g / dmˆ3)
   - the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
3. citric acid (g / dmˆ3)
   - found in small quantities, citric acid can add 'freshness' and flavor to wines
4. residual sugar (g / dmˆ3)
   - the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet
5. chlorides (sodium chloride - g / dmˆ3
   - the amount of salt in the wine
6. free sulfur dioxide (mg / dmˆ3)
   - the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine
7. total sulfur dioxide (mg / dmˆ3)
   - total sulfur dioxide: amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine
8. density (g / cmˆ3)
   - the density of water is close to that of water depending on the percent alcohol and sugar content
9. pH
   - describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
10. sulphates (potassium sulphate - g / dm3)
    - a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant
11. alcohol (% by volume)
    - the percent alcohol content of the wine

Output variable (based on sensory data):

12. quality (score between 0 and 10)

Data Source: P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.

Wine Dataset Summary Statistics:

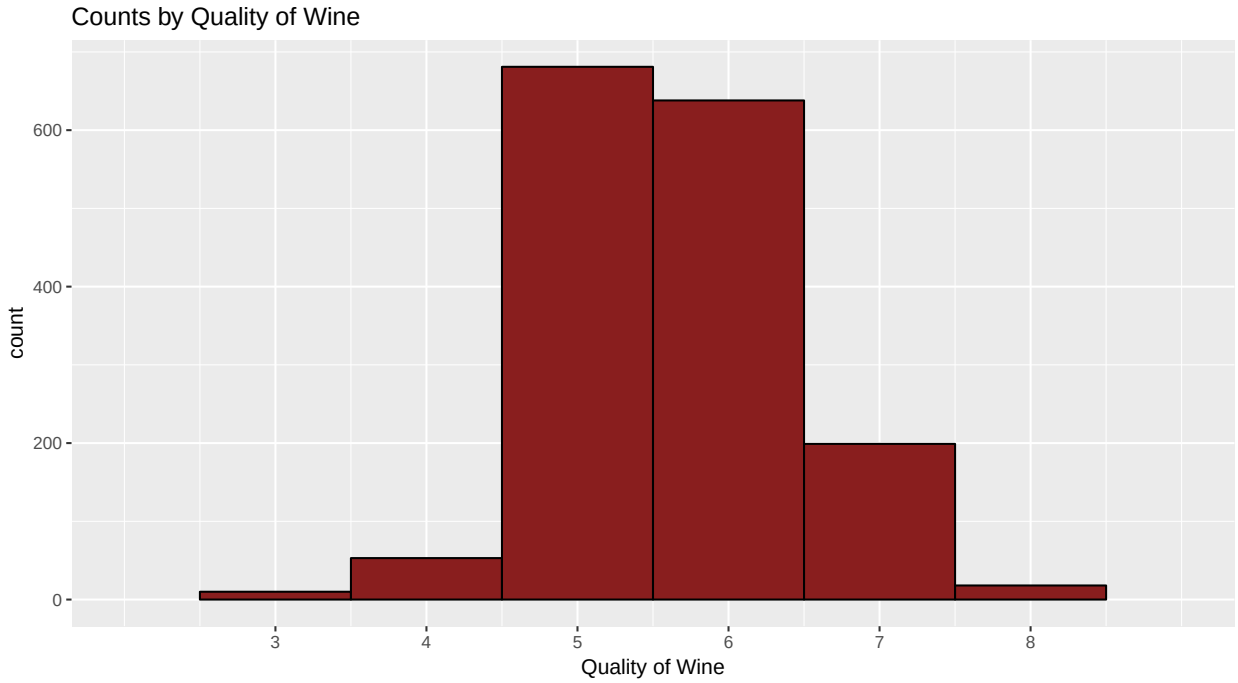
```
##        X          fixed.acidity   volatile.acidity  citric.acid
##  Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000
##  1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090
##  Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260
##  Mean   : 800.0   Mean   : 8.32   Mean   :0.5278   Mean   :0.271
##  3rd Qu.:1199.5   3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420
##  Max.   :1599.0   Max.   :15.90   Max.   :1.5800   Max.   :1.000
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.900   Min.   :0.01200   Min.   : 1.00
##  1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00
##  Median : 2.200   Median :0.07900   Median :14.00
##  Mean   : 2.539   Mean   :0.08747   Mean   :15.87
##  3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00
##  Max.   :15.500   Max.   :0.61100   Max.   :72.00
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  6.00       Min.   :0.9901   Min.   :2.740   Min.   :0.3300
##  1st Qu.: 22.00       1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500
##  Median : 38.00       Median :0.9968   Median :3.310   Median :0.6200
##  Mean   : 46.47       Mean   :0.9967   Mean   :3.311   Mean   :0.6581
##  3rd Qu.: 62.00       3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300
##  Max.   :289.00       Max.   :1.0037   Max.   :4.010   Max.   :2.0000
##     alcohol         quality
##  Min.   : 8.40   Min.   :3.000
##  1st Qu.: 9.50   1st Qu.:5.000
##  Median :10.20   Median :6.000
##  Mean   :10.42   Mean   :5.636
##  3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :14.90   Max.   :8.000
```

Univariate Plots Section

Various single measures of the wine will be expalained in this portion of the analysis. Based on findings, analysis will be done on the relationship between various sets of measures.

Wine Quality

How well were the wines rated?



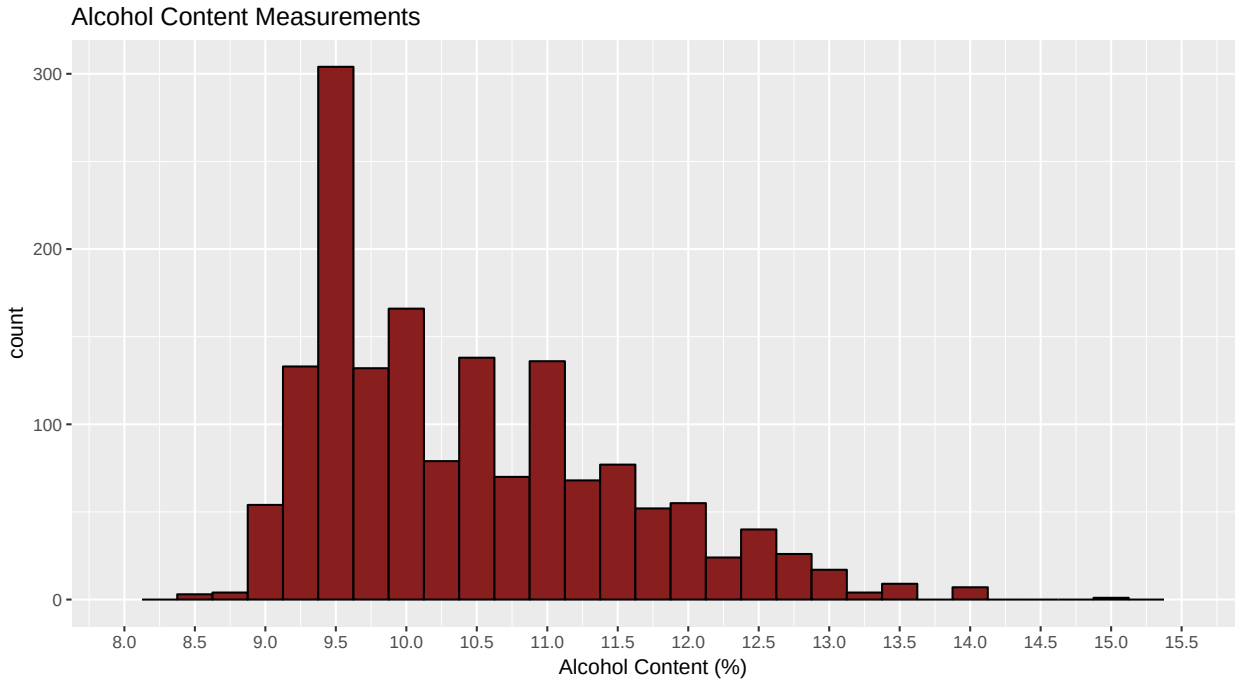
The above chart describes the distribution of quality ratings amongst the wine samples. The wines are very heavily distributed in the 5 and 6 quality ratings. On the high quality end, There are no wines that scored a 9 or 10. Likewise, no wines scored a 1 or 2. These wines were mediocre based on the qualities rated by the wine experts.

The summary statistics are as follows for the quality ratings of the wine:

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

Wine Alcohol Content

How high/low is the alcohol content?



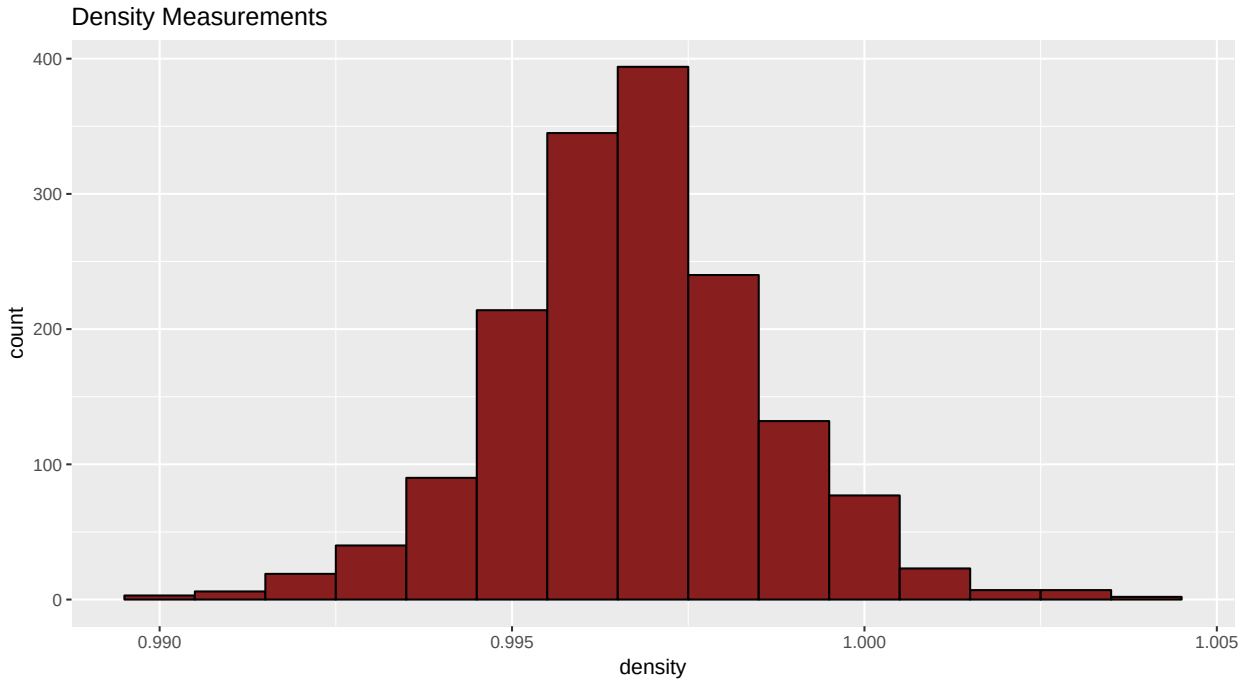
The majority of the wine samples in our data fall between 8% and 11% alcohol content by volume. A large number of the wine samples are concentrated around 9.5% alcohol by volume. Only a small amount of the wine samples are above 12%. According to Alchol.Org, most red wines are typically between 12-15% Source. Perhaps this is why our quality was lower? We will explore this relationship further in our analysis.

The summary statistics are as follows for the alcohol content of the wine:

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

Density Analysis

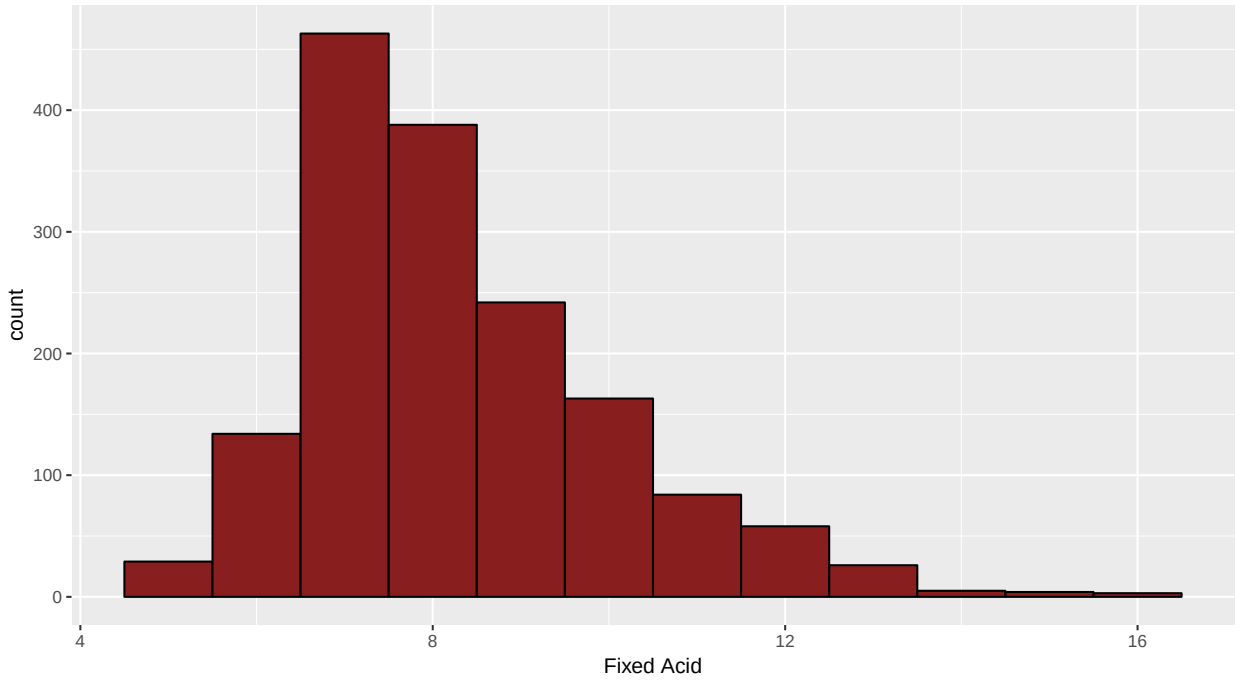
How high/low is the density of the wine?



The density measurements of the wine follow a normal distribution. What will be interesting to analyze will be what variables possibly have an effect on the density? My hypothesis based on the author's notes will be the residual sugar having the most effect on the density.

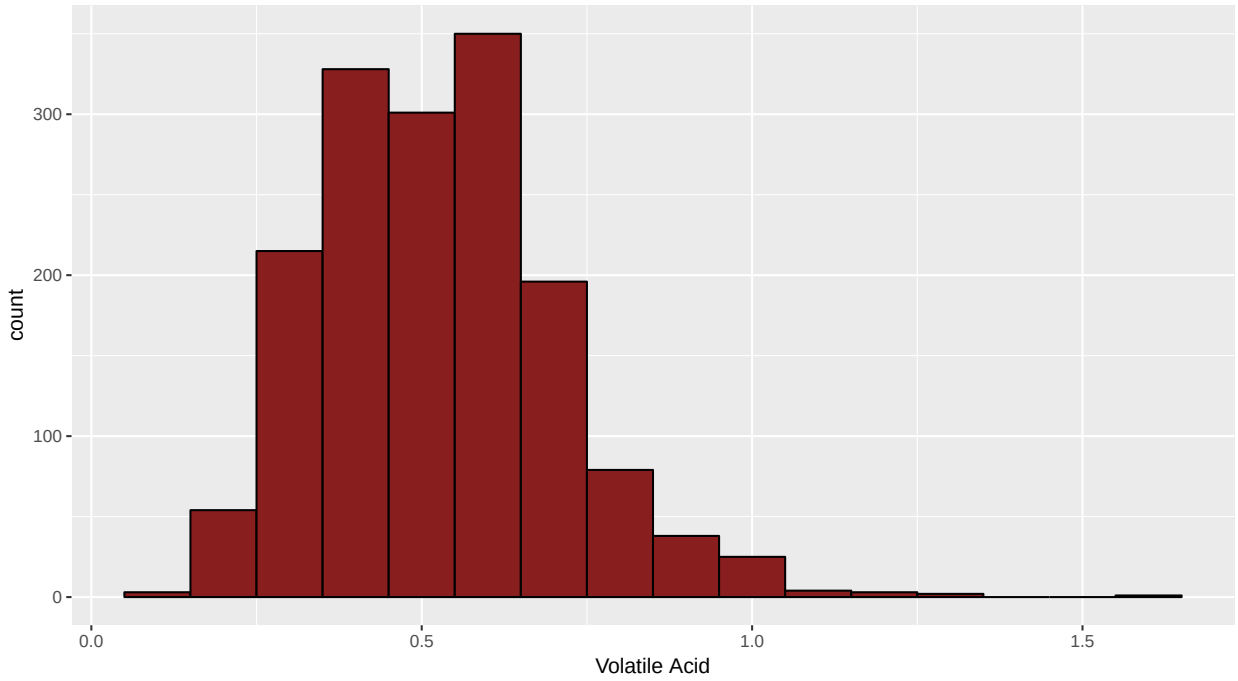
Acidity Analysis

Analysis of the acidity involved four separate measurements - fixed, volitle, citric acidity, and pH.

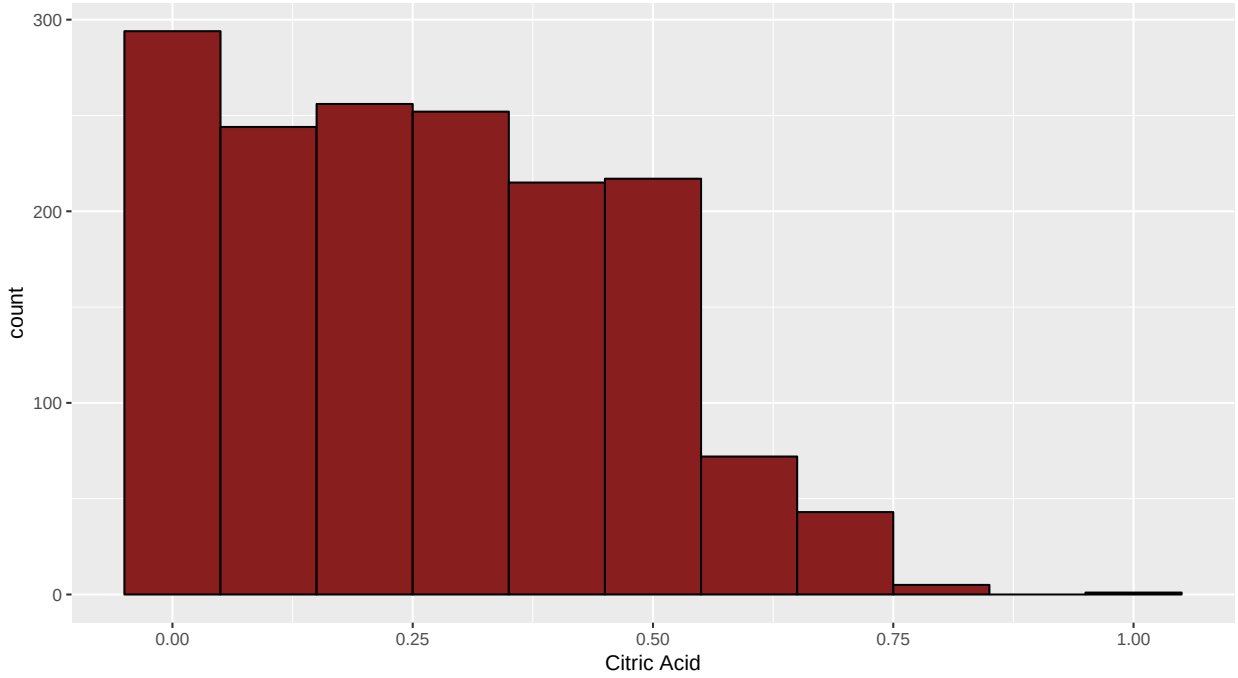


Fixed acidity is a component of all wines and has a normal distribution. There are very few samples that

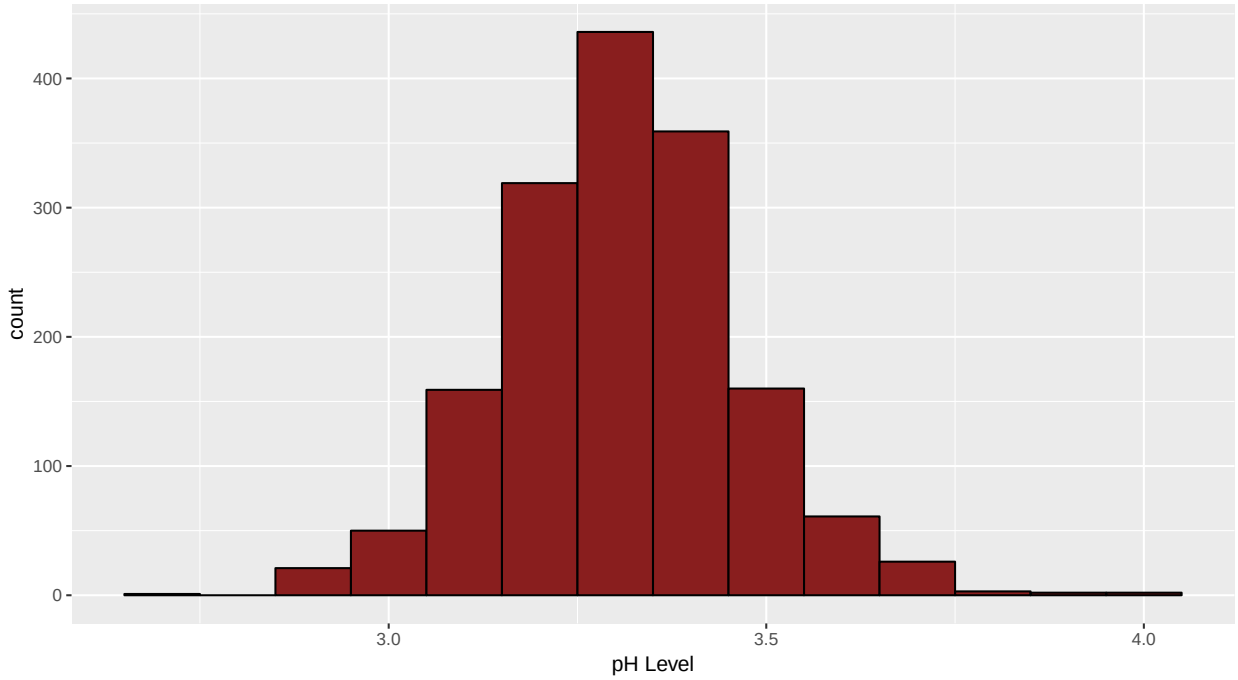
have 4g/dm^3 and similarly very few samples that have 10+ g/dm^3.



Volatile acid is an unwanted component of wine and higher amounts can lead to unpleasant tastes, similarly to fixed acidity, volatile acid is normally distributed. The distribution follows a similar pattern to fixed acidity, however the g/dm^3 level is lower than fixed acidity.



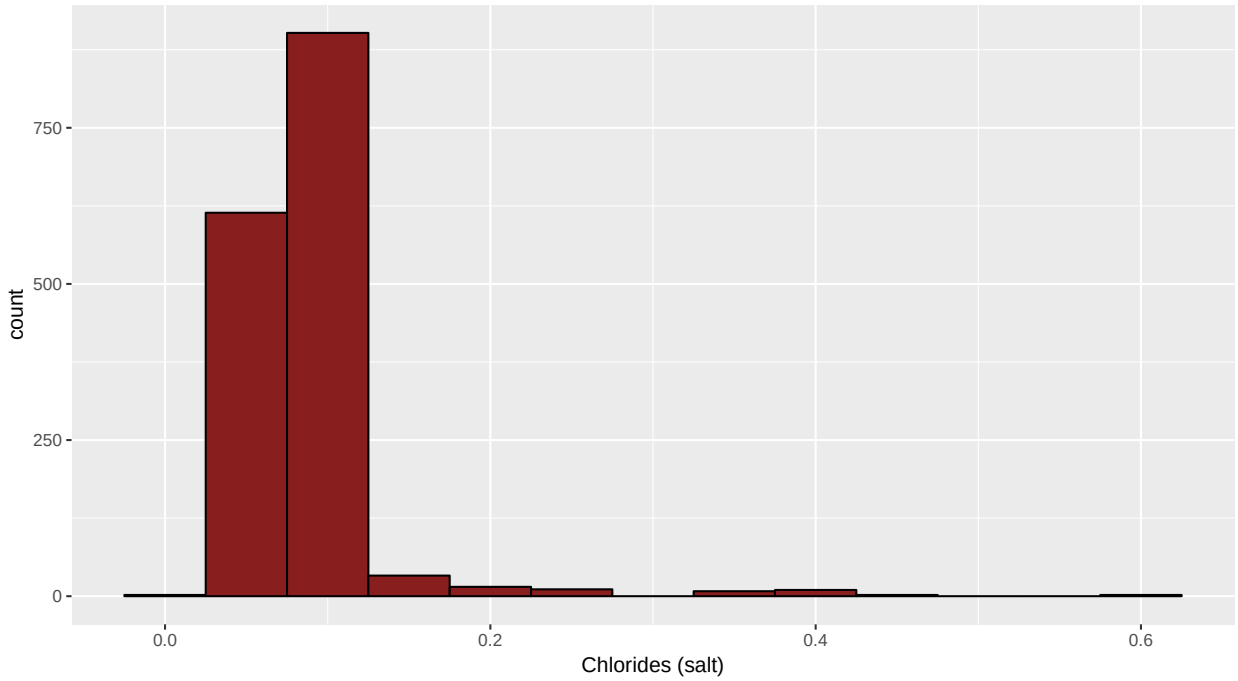
Citric acid has a right-skewed distribution with most wines having smaller amounts. The amount of citric acid does not really crest over 0.75 g/dm^3.



pH level has a near perfect distribution - we would expect to see a possible negative correlation between pH and various measurements of acidity. Further analysis will be done to see if acidity has an impact on quality.

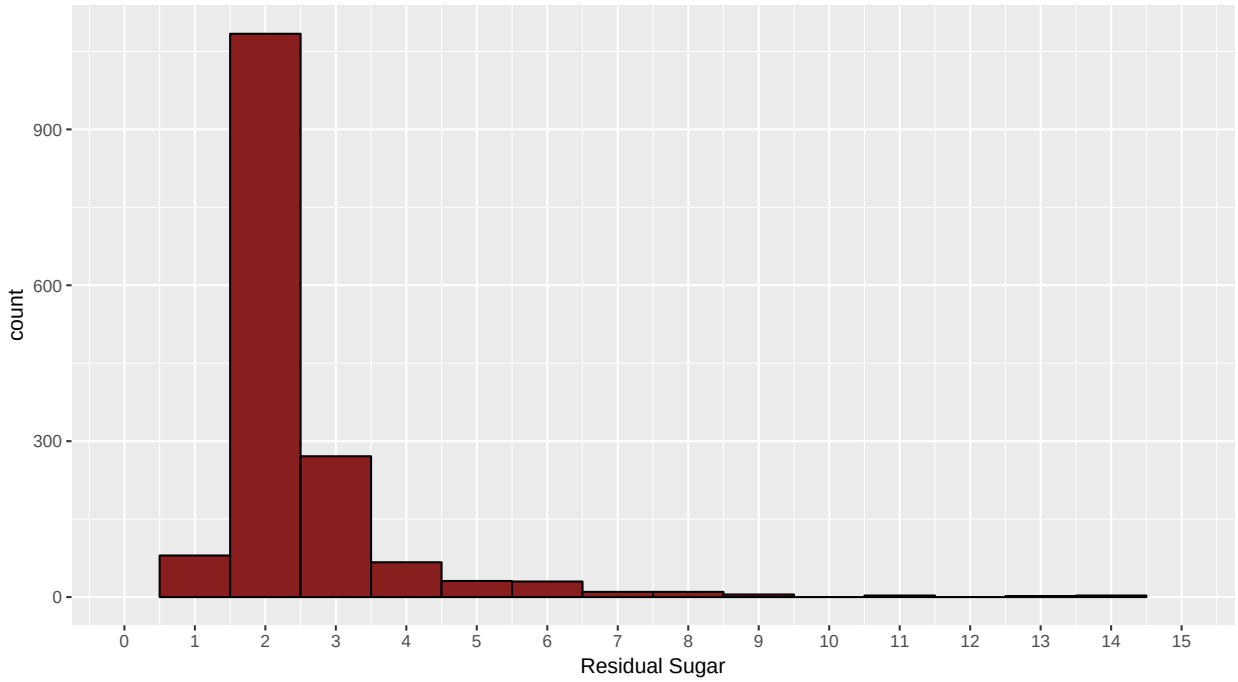
Salt/Sugar Analysis

An analysis was performed on chlorides/sugars to see how they are distributed in the dataset.



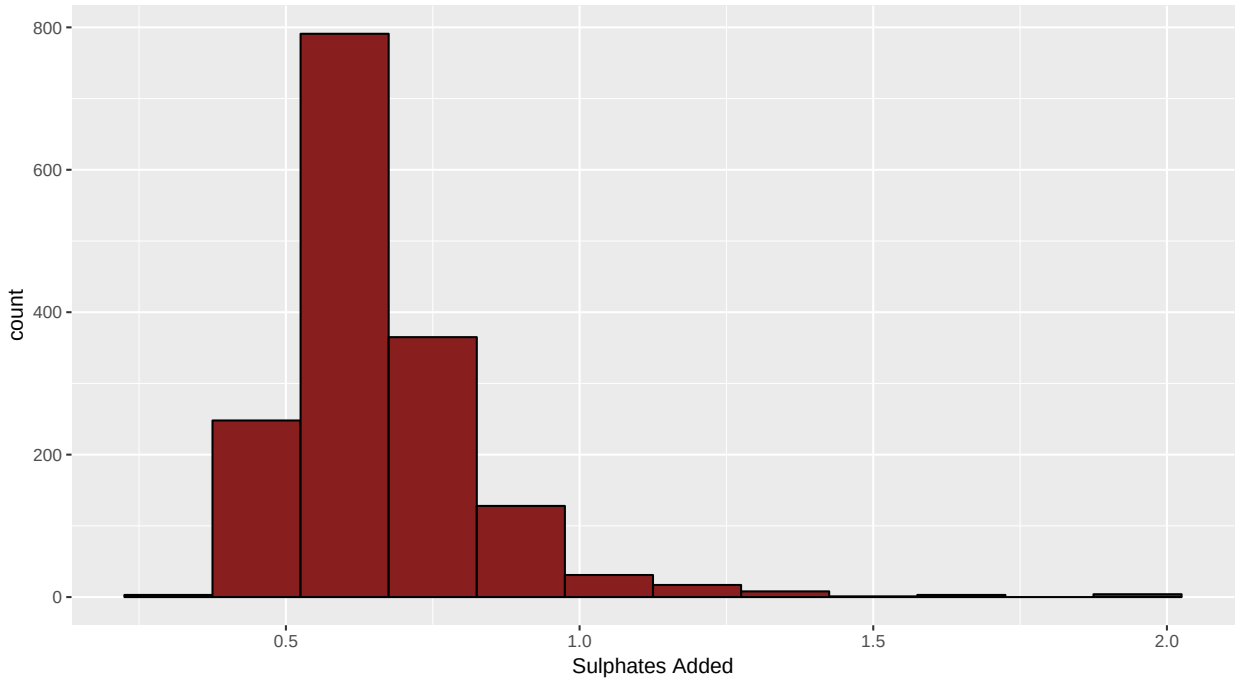
Salt levels are not very evenly distributed and have a signifcant right skewness. It is safe to say in our sample

that there is not a lot of salt content.

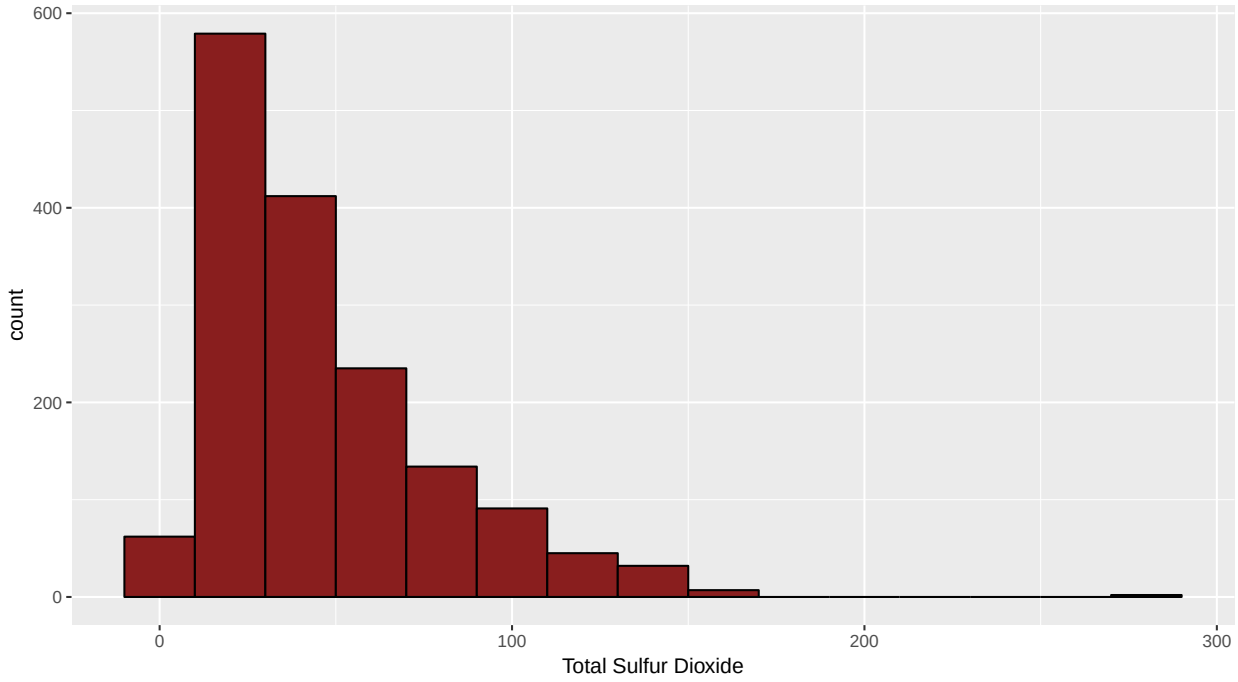


From a conventional standpoint, visualizations of chloride and residual sugar content make sense. Too much salt or sugar would result in an overly salty or sweet wine. Will the wines that have very high residual sugar content also have a high quality rating?

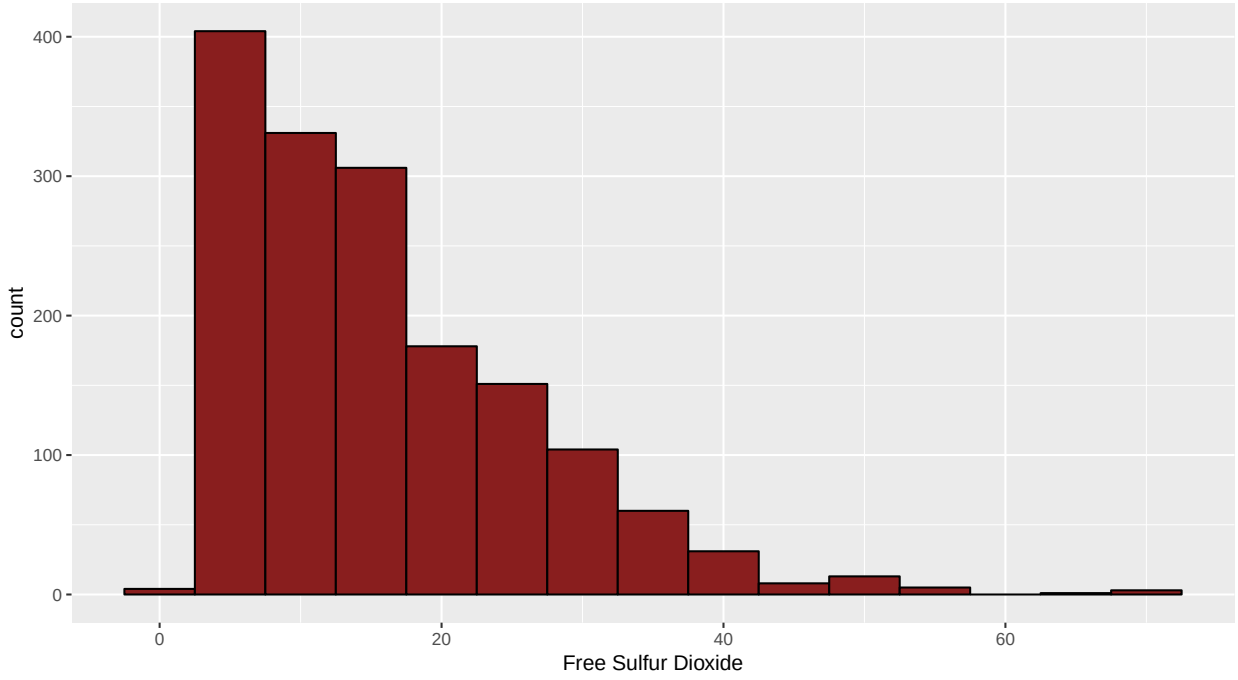
Sulfur Dioxide Analysis



The amount of sulphates added serves to help keep the wine fresh. The number of sulphates added hovers predominantly around 0.6-0.7 g/dmˆ3.



Similar to our sulphate levels, total sulfur dioxide measurements are right skewed leaning almost towards a normal distribution.



Free sulfur dioxide appears to follow an almost identical distribution to total sulfur dioxide. The overall scale is smaller which leads me to believe that perhaps free sulfur dioxide is a possible result of total sulfur dioxide

or sulphates added? Multivariate analysis or correlation could indicate if this is probable.

Sulfur related measurements involve the amount of sulfur dioxide present in the wine. While different scales, the sulphates added, total sulfur dioxide, and free sulfur dioxide appear to follow similar distributions indicating perhaps they are correlated in some way. The author indicates sulfur levels over 50ppm can lead to a noticeable increase. Will the wines with higher sulfur content have lower quality?

Univariate Analysis

What is the structure of your dataset?

The dataset consists of 1599 wine samples with 13 variables (1 of which - "X" is an identifier for the wine and not analyzed in this research). The quality measurement is based on a 1-10 rating scale given to the wine by a wine expert. Every column in the dataset is numerical, there are no factors/categorical variables present in the dataset.

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

The main feature of interest in this dataset is quality. The rating given by the wine expert. This analysis could provide with wine makers, etc. the ability to rate a wine without a subjective tasting and instead rely on data. A predictive model could even be built given a set of features (such as sugar, etc.) and sucessfully predict the quality.

What other features in the dataset do you think will help support your

While present across different measurements, three collections of measurements interest me greatly: sulfur dioxide content, sugar/salt content, and acidity levels of the wine. The remainder of this analysis will focus on determining which of these measurements have an impact on quality.Additionally, does alcohol content improve/diminish quality in the wine?

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

Because the variables (measurements) are all numeric data types, no new variables were created. Instead analysis (means, regression lines, etc.) will be created in an adhoc manner.

Of the features you investigated, were there any unusual distributions?

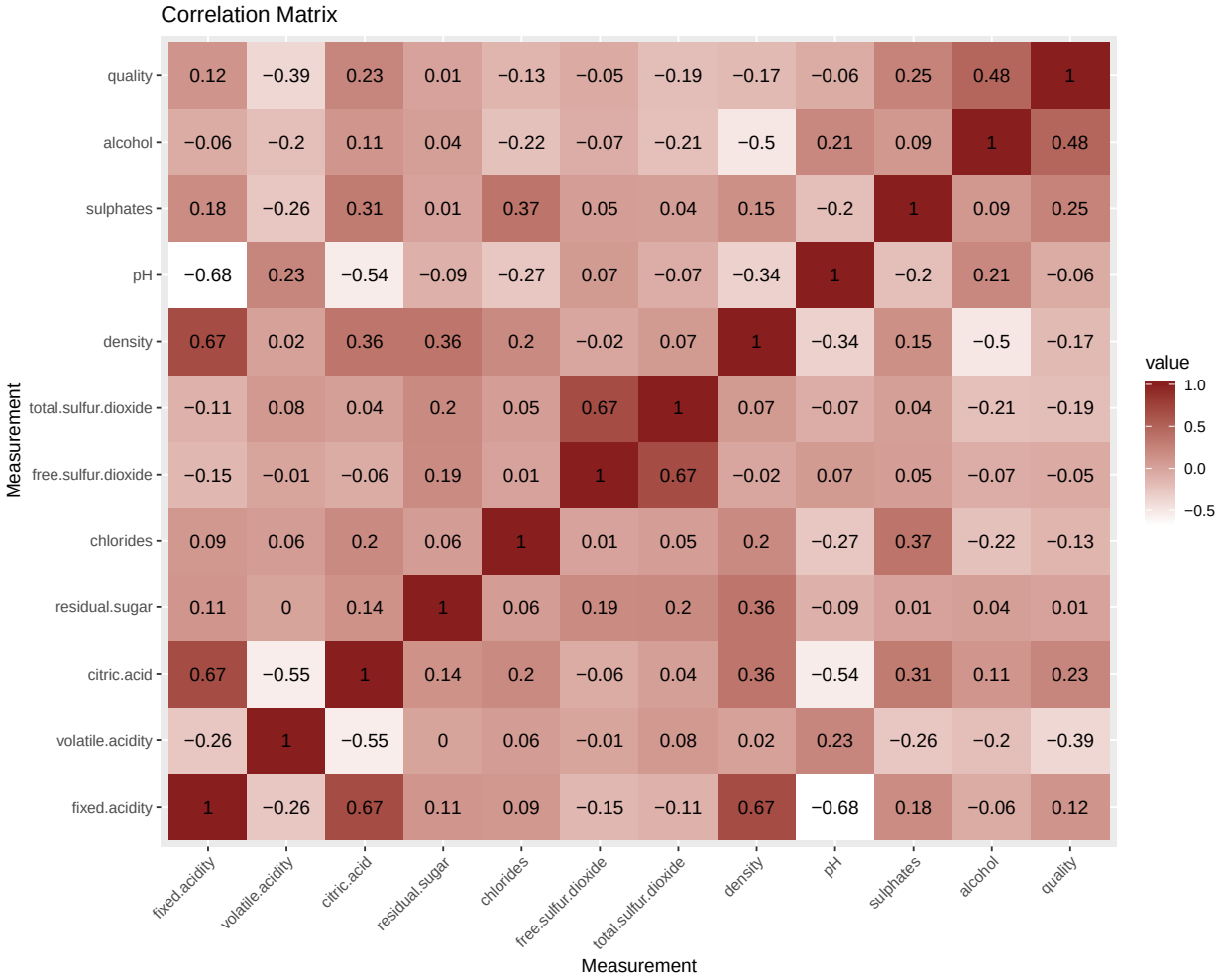
Yes, while the quality of the wine followed a normal distribution - I found it very unusual that there were so many of the wines rated at a 5/6 and no wines that scored in the 9/10 or 1/2 range. In addition to the quality measurements, there were a few samples taken of the wine that had (compartively) much higher measurements in terms of chlorides and residual sugar levels.

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?

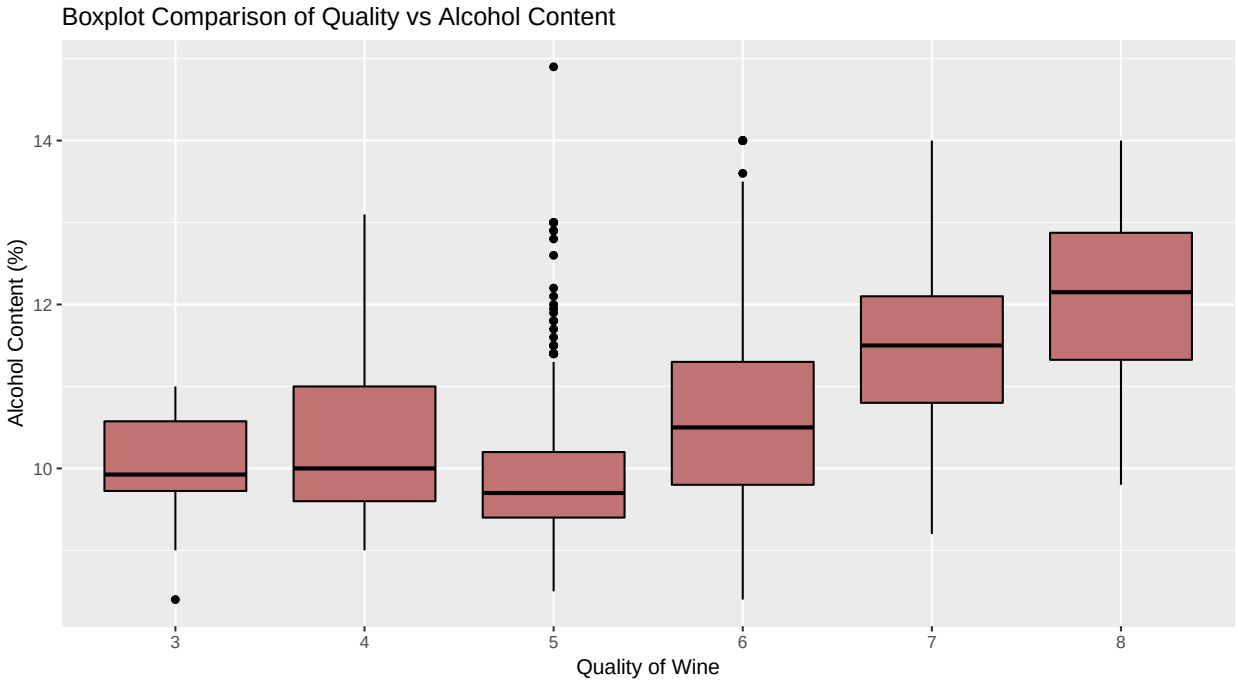
Yes - the bin distributions were modified for almost every graph. Due to the very shallow or very wide distribution for some of the measurements, the bins needed to account as such. I considered log transforming the X axis in the residual sugar visualization, but the resulting log10 or log2 transformations were not much different than the final product.

Bivariate Plots Section

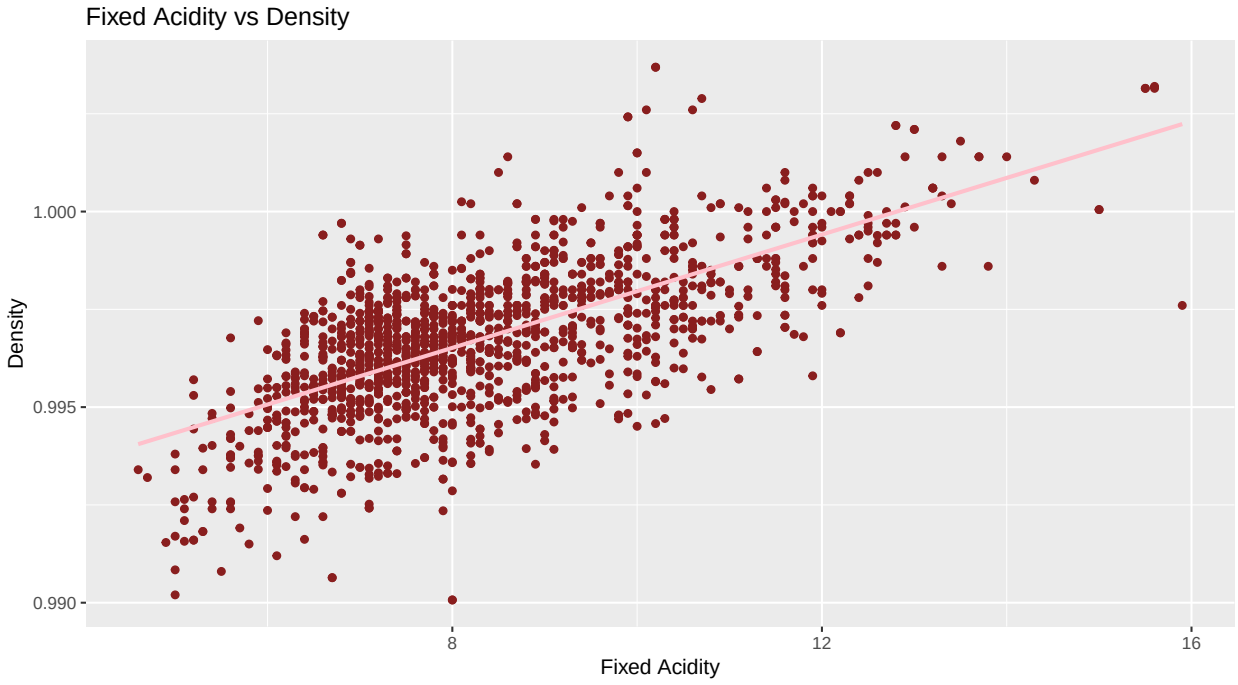
Now with measurements of interest in mind, variable values will be compared/measured against other variables. Of first importance is determining which variables are correlated:



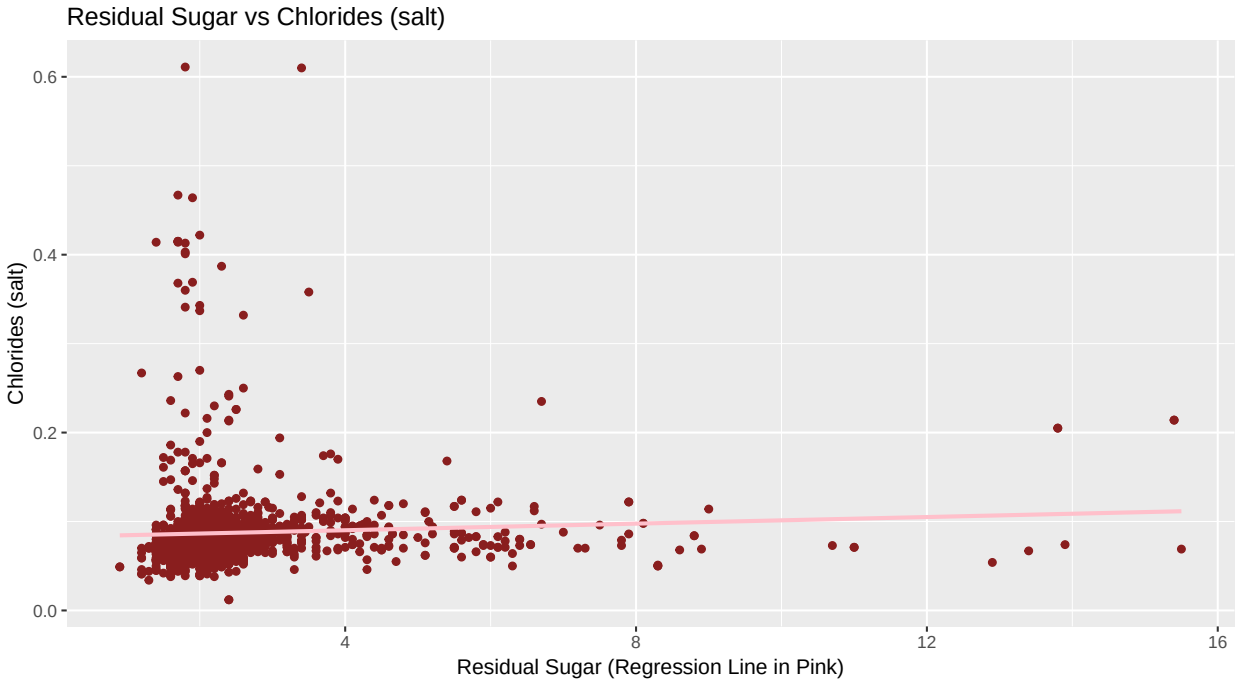
The above correlation matrix shows some incredibly interesting relationships in the data. The pH/fixed acidity/volatile acidity/citric acidity all appear to be negatively correlated. This makes sense based on how pH works. As pH increases, wine becomes more basic and less acidic. Based on this correlation matrix, as acidity of any kind increases, pH decreases. There appears to be a positive correlation between quality and alcohol content, indicating that as alcohol content increases, the quality rises. Alcohol content also appears to be positively affected by sulphates added and fixed/citric acidity. However, the opposite is true of volatile acidity, which increases has a negative impact on quality.



The above boxplot demonstrates a very clear positive correlation between alcohol content and the quality of the wine. However, due to the presence of so many outliers in the "5" quality rating, additional features will be considered for their impact on quality in further multivariate analysis.



Based on our correlation matrix, fixed acidity and density were important measurements for me to consider the relationship between. The two measurements had one of the strongest correlations of all the variables. This visualization and regression line shows a very strong positive correlation between the fixed acidity levels and the density of the wine.



The above chart appears to show that as residual sugar levels increase in a wine, a decrease is seen in chloride levels. However, fitting a regression line to the data displays that while there are a few samples that have this relationship, the overall relationship is just barely positively correlated.

Bivariate Analysis

Talk about some of the relationships you observed in this part of the

The most interesting and revealing relationship observed here is between quality and alcohol content. Almost like steps on a staircase, as alcohol content increases, so too does the quality of the wine. The one item of interest however is the significant amount of outliers in the "5" quality rating category. Several samples that receieved a rating of "5" had alcohol content above 11%. This will be an important metric to try and unravel.

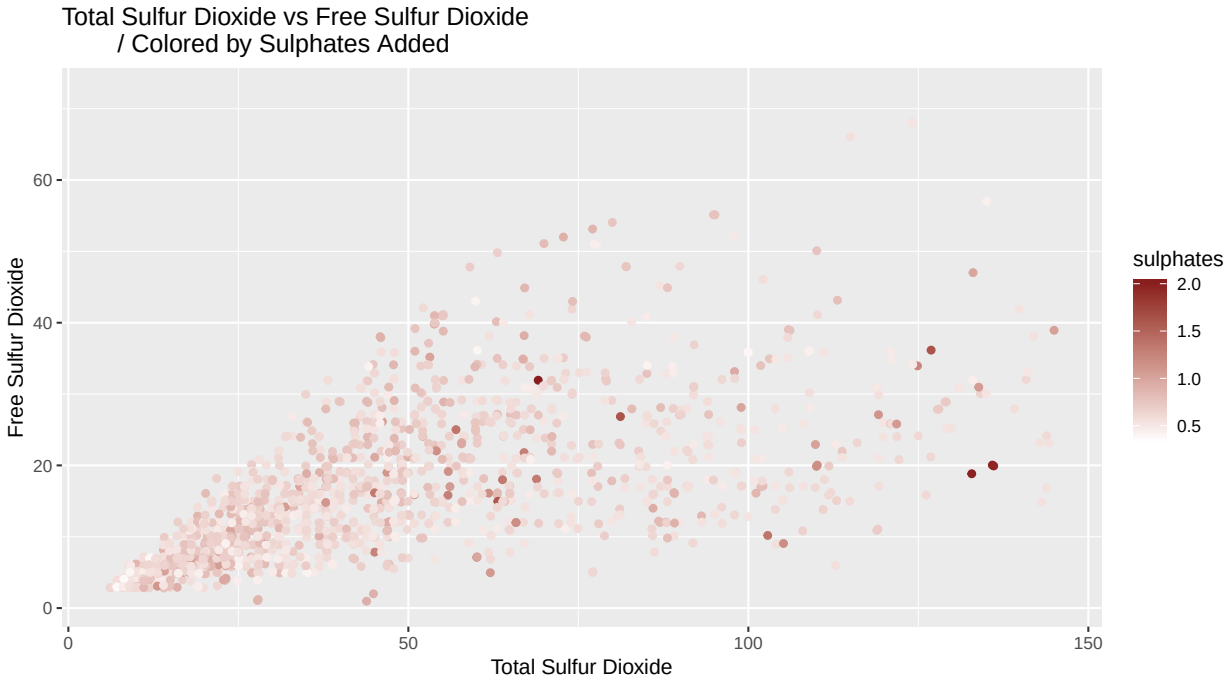
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

Interestingly, an initial inspection of sugar vs salt would indicate that an intuitive relationship exists between the two measures (i.e. as sugar increases, salt decreases). However, fitting a model revealed just a slight trend upward. A multivariate plot will be analyzed to discover if quality increases more so with residual sugar, clorides, or a increase/combination of both.

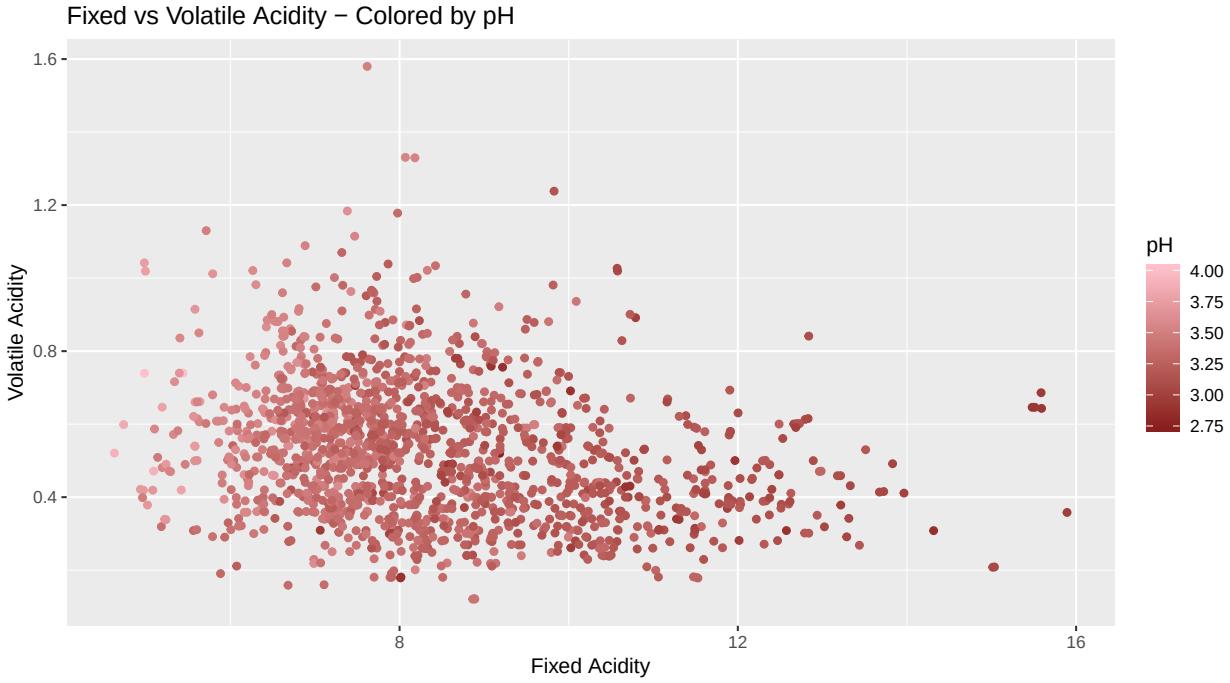
What was the strongest relationship you found?

The correlation matrix displayed a VERY strong positive correlation between citric/fixed acidity, density/fixed acidity, and total sulfur dioxide/free sulfur dioxide.

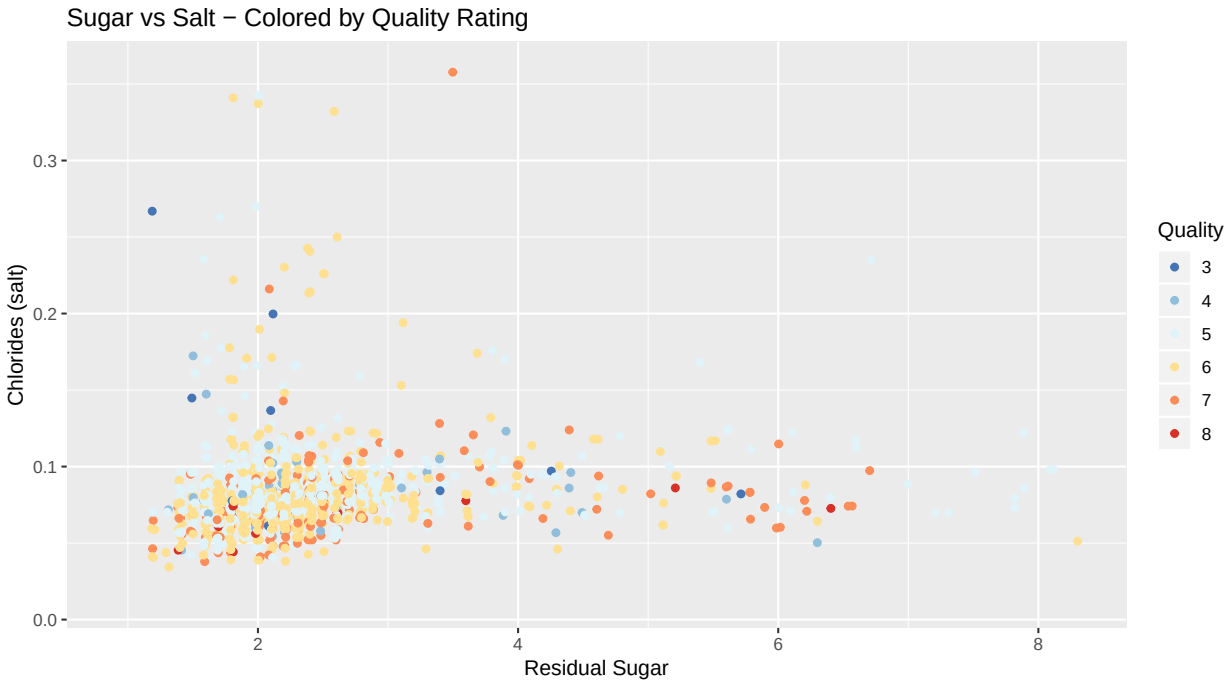
Multivariate Plots Section



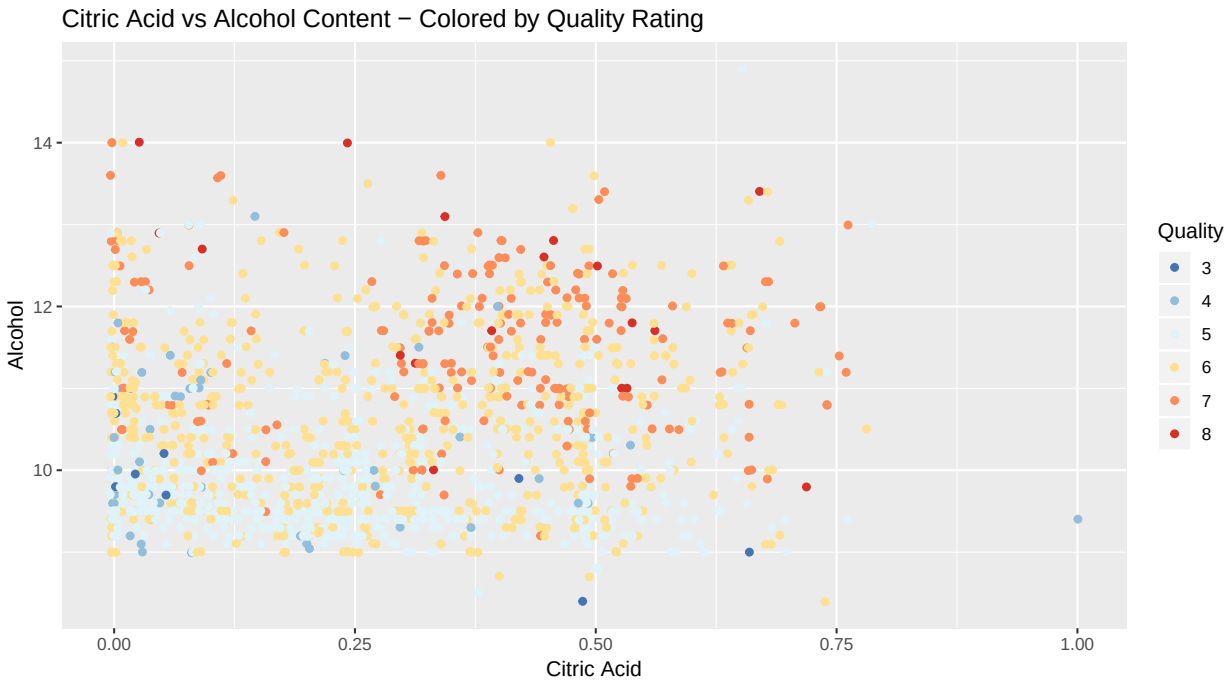
This visualization demonstrates the relationship between sulphates, total sulfur dioxide, and free sulfur dioxide. The vast majority of the dataset is concentrated in the lower levels of sulfur dioxide. There appears to be a strong positive correlation between the three variables as increasing the sulfur dioxide also increases the free sulfur dioxide. Sulphates added impact both the total sulfur dioxide content as well as the free sulfur dioxide.



This scatter plot shows a very interesting relationship between the acidity measurements in the wine data. The left quadrant of the visualization shows that lower pH values result in lower fixed acidity and lower volatile acidity. There aren't really any visible samples that have a high volatile or fixed acidity that have a lower pH. Again, knowing what we know about the pH scale, this makes sense and is what we expect to see!



As a follow up to our previous bivariate line chart, this scatter plot shows us that there does not appear to be any real strong relationship between quality, sugar, and salt. Extreme color choices were used in order to more easily visualize the distinct qualities the wines were given. While some of our higher quality wines have more sugar and less salt - we have a high concentration of low sugar/low salt higher quality wines. This effectively puts to rest any concerns that too low of a sugar/salt measurement will hurt our quality.



At first glance, this visualization does not tell us much - however upon further inspection we can glean some interesting relationships. In our correlation matrix we saw that the overall quality of a wine was greatly impacted by alcohol content. Looking at the alcohol content we can see that it is correlated to an increase in citric acid. This visualization seeks to explore the relationship between those three measures. What we can see is that wines that have a low citric acid AND alcohol content tended to have a lower quality rating. Inversely, wines that have a higher citric acid and alcohol content tended to have reliably higher quality ratings.

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?

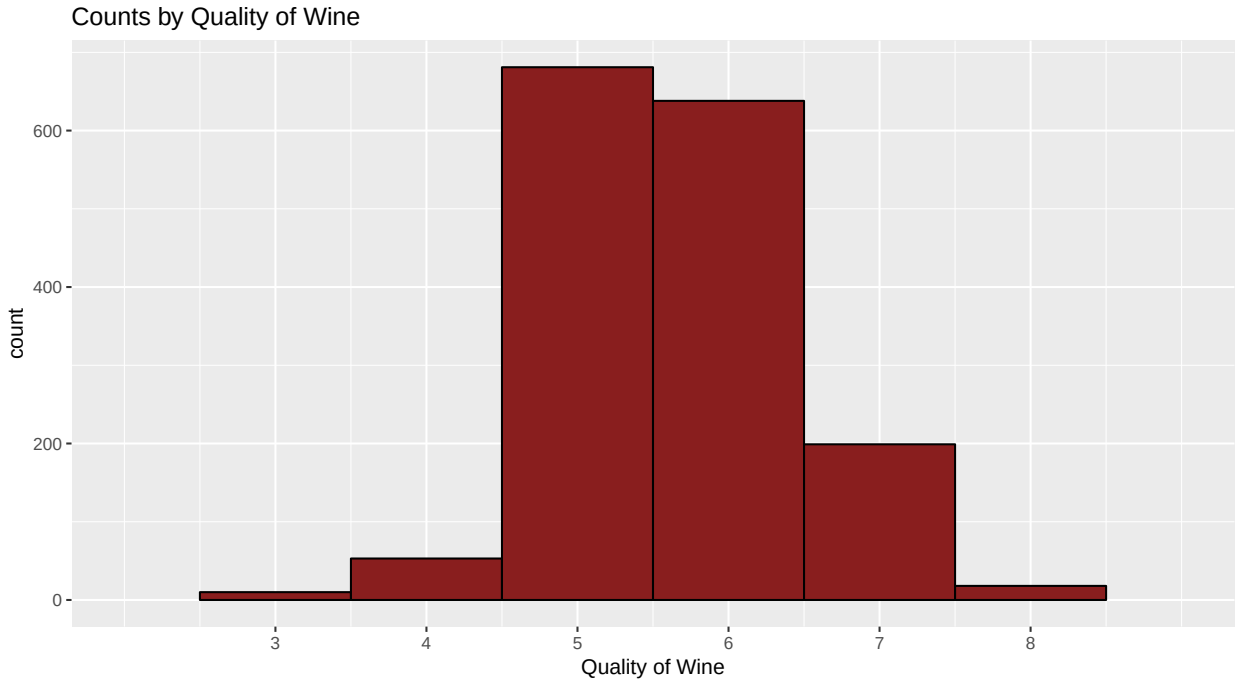
All of the acidity measurements were very dependent on one another. This enforces what we see in our correlation matrix. As pH increases, our acidity goes higher. Becuase there is a moderate positive correlation with quality, it is same to assume that our more acidic wines will likely have higher quality. But experiments/statistical comparisons will be needed for this.

Looking at citric acid/alcohol content levels was very interesting. These two variables seemed to have a reliable effect on the quality rating given to a wine.

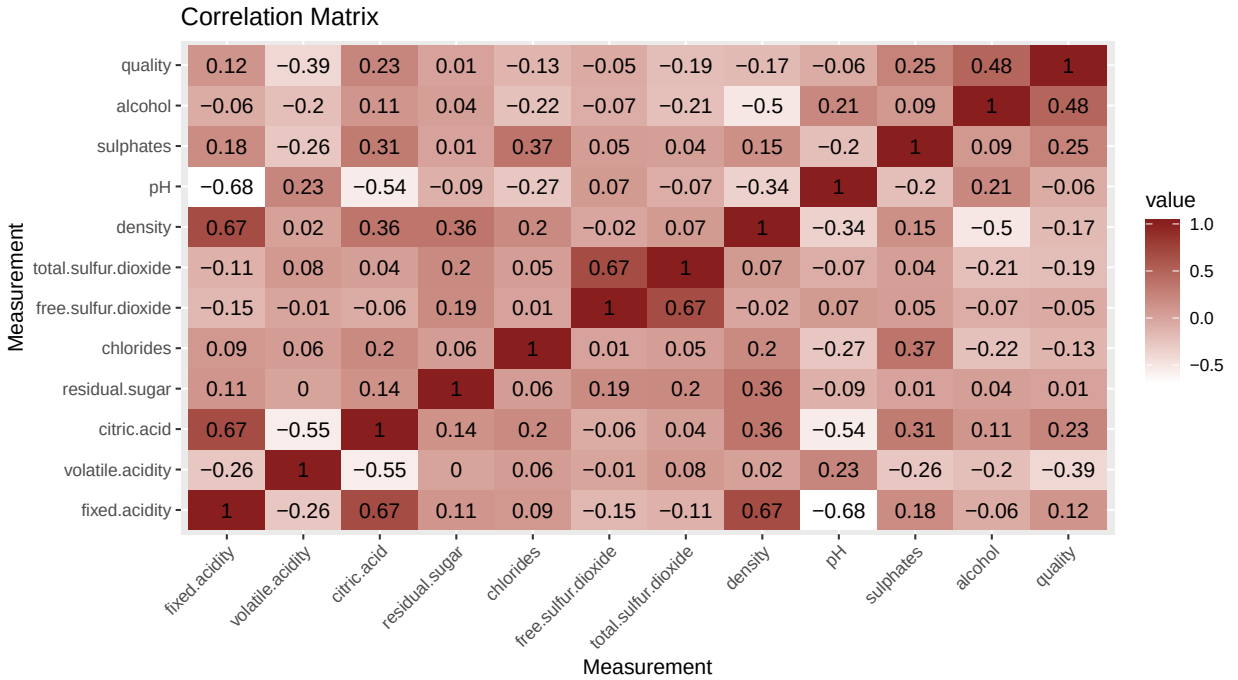
Were there any interesting or surprising interactions between features?

The relationship between quality/salt/sugar was VERY surprising! Upon initial investigation I would have assumed that higher sugar wines would by far have higher residual sugar measurements. This did not really prove to be true, instead there was not a strong relationship between sugar/chlorides and quality ratings.

Top 3 Insightful Visualizations and Summary

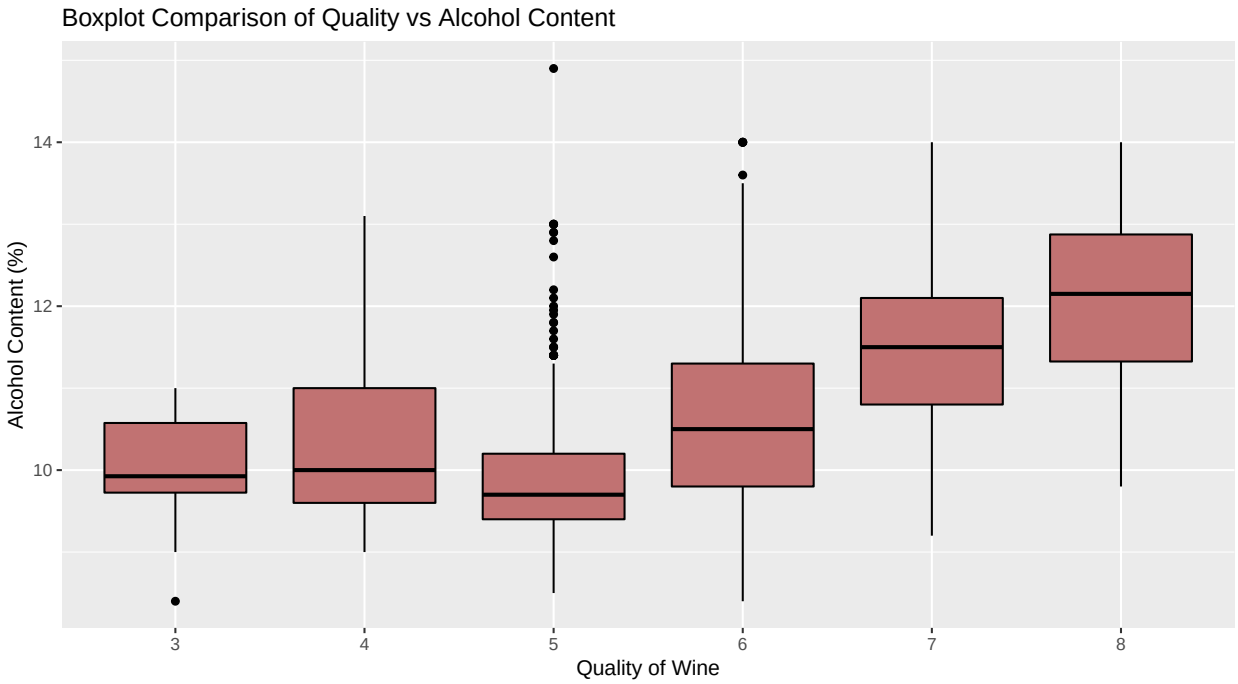


Insights: Because our dataset is very dense with quality rating 5 and 6 wines, using this dataset to help determine what makes an excellent (9/10) wine will be very difficult. Likewise to try and use this dataset to determine what makes a wine terrible (1/2) is also difficult due to our lack of observations of data matching that criteria. Overall our red wine dataset should be called the mediocre red wine dataset.



Insights: Our correlation matrix was very useful in pointing out our acidity and sulfur dioxide related

measurements correlations. What our dataset is missing are any "red flags" (no pun intended) that indicate a sure-fire high/low quality wine. Like mentioned in the insight above, this is likely due to our limited dataset and having no terrible/outstanding samples to analyze. While we see weak/moderate positive relationships, there are no measurements that we can difinitively state are indicitive of a gold medal winning wine.



Insights: Alcohol content's impact on quality was the most apparent impact on quality of wine. While the data collection methods were not clear on if all of this wine sampling occured in one night, there is a noticeable increase in the average alcohol content for each level in the quality scale. Of all of our measures, this variable would likely be our best single predictor variable in our dataset.

Summary:

Wine quality has several chemical measurements that work together to make what we know and love. While this particular dataset did not have any "smoking guns", it is clear that overall factors such as acidity, sulfur dioxide content, and alcohol content mildy/moderatley impact the quality of wine. The two variables that I would say impact our wine quality the most is citric acid and alcohol content. These appeared to be the most impactful measurements related to the quality rating given to the wine.

Moving forward, a more robust dataset containing samples deemed excellent/terrible would be important to a more thorough and complete analysis (especially if estimation/prediction was a goal). In the lower quality range of red wine, alcohol content is typically about 1-1.5% lower than the higher rated wine.

Reflection

This dataset was incredibly interesting. While I am not an avid wine drinker, it was fascinating to see the various chemical properties that impact wine. Like detailed at great length above, I wish the dataset had (ANY) samples that were deemed excellent or terrible (quality 1,2,9,10). Because this dataset was primarily numeric, that also created the challenge of not being able to do a lot of factoring/categorization, the only variable that could really be easily factored was the quality. Because the chemical measurements are not a familiar measurement unless you are a scientist or wine expert, it also makes the data hard to

understand/communicate in laymans terms. Overall the analysis was a sucess and it was very surprising that alcohol content had such a clear relationship with quality!

Future Work: For this dataset, I would like to apply some machine learning models that could take several measurements as inputs (perhaps using density, acidity measurements, alcohol content) and is able to accurately predict a quality as an output. This data would likely lend itself well to a regression model.