Exploratory Analysis on White Wine by Nicholas Pasquinzo

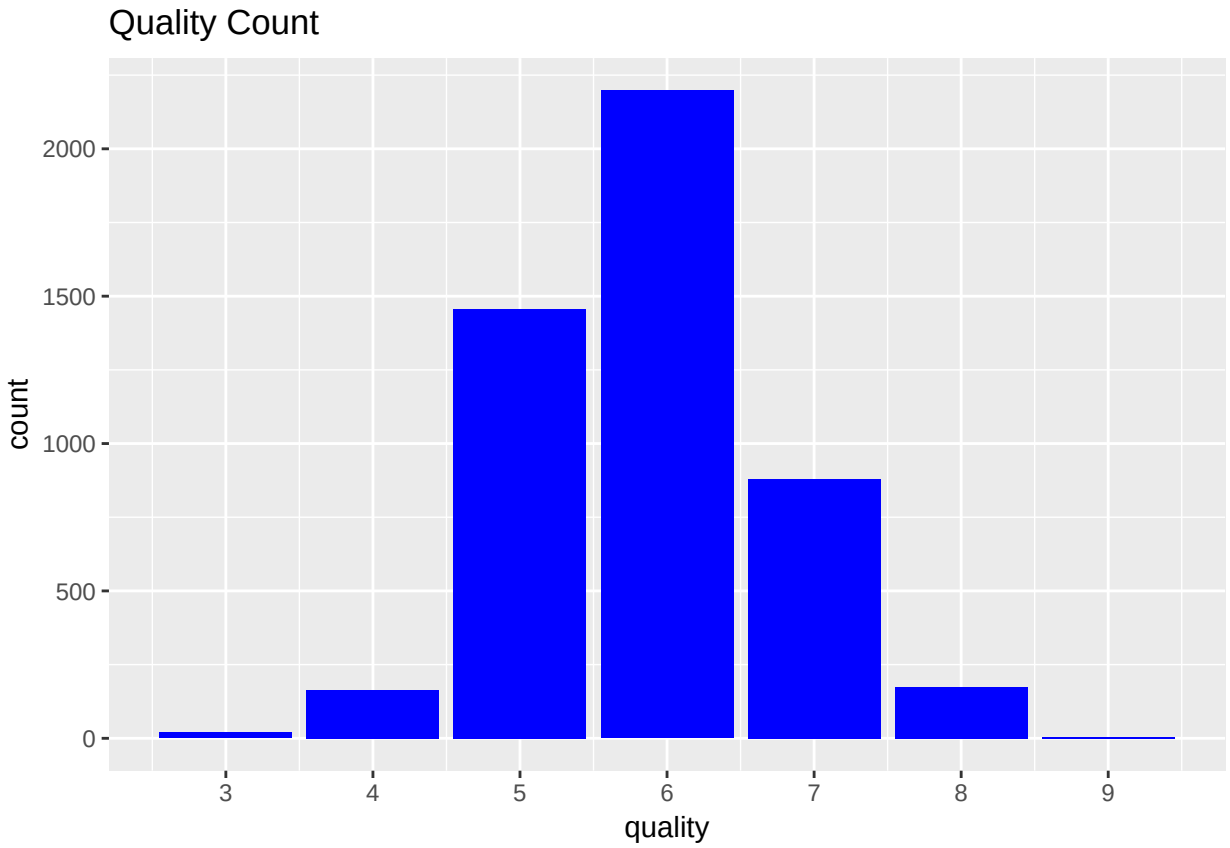
getwd() setwd('~/Documents')

This data set contains variables that describe the attributes of multiple types of white whine. Including chemical attributes of the wine, including values such as citric.acid level, residual sugars and alcohol content. They even include how good the wine tasted based on a 1-10 scale.

Univariate Plots Section

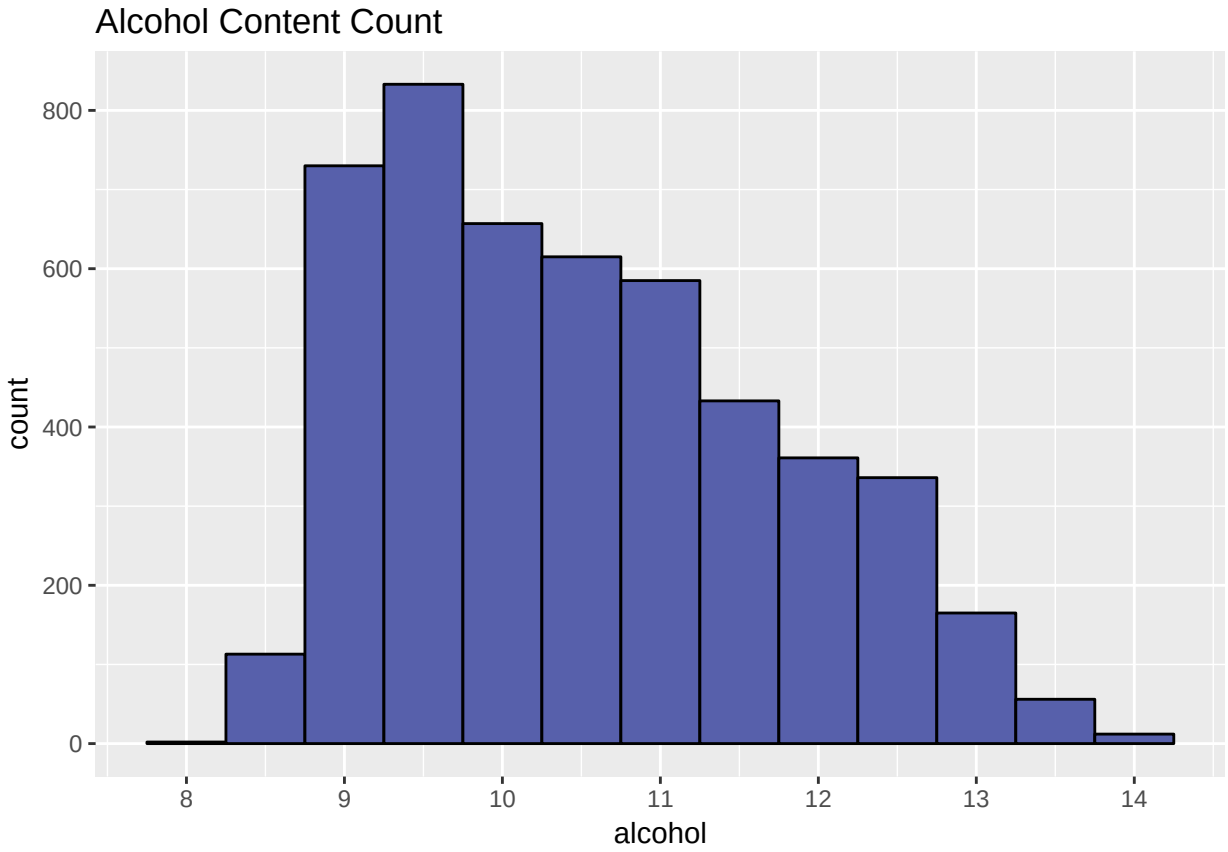
qplot(data = wines, x = alcohol)

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.878   6.000   9.000
```

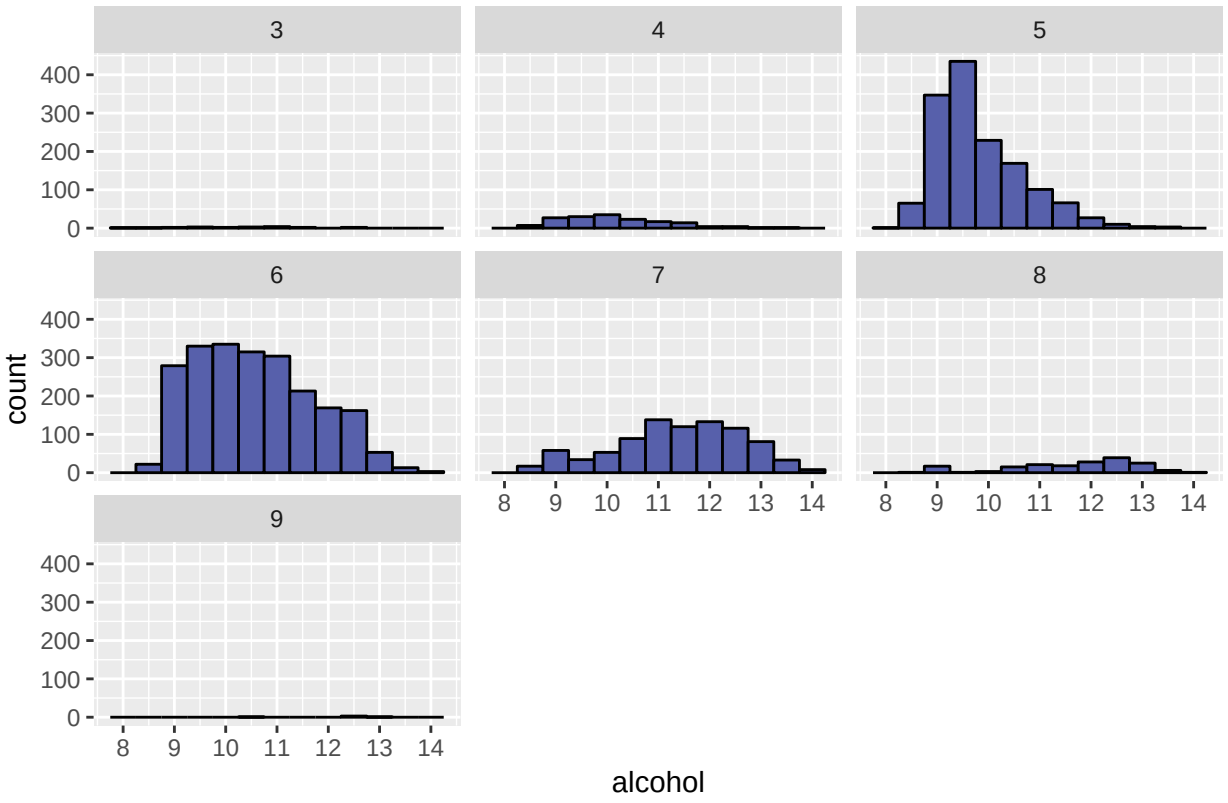


The plot shows the distribution of quality count. As we can see it is a nice even distribution with the most ocurring during the middle. An interesting observation is that there is no values for quality below 3 and above nine. This means that there is no perfect wine, or truly horrible tasting wine in the dataset. I will omit these values in the future for further analysis.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.50   10.40   10.51   11.40   14.20
```



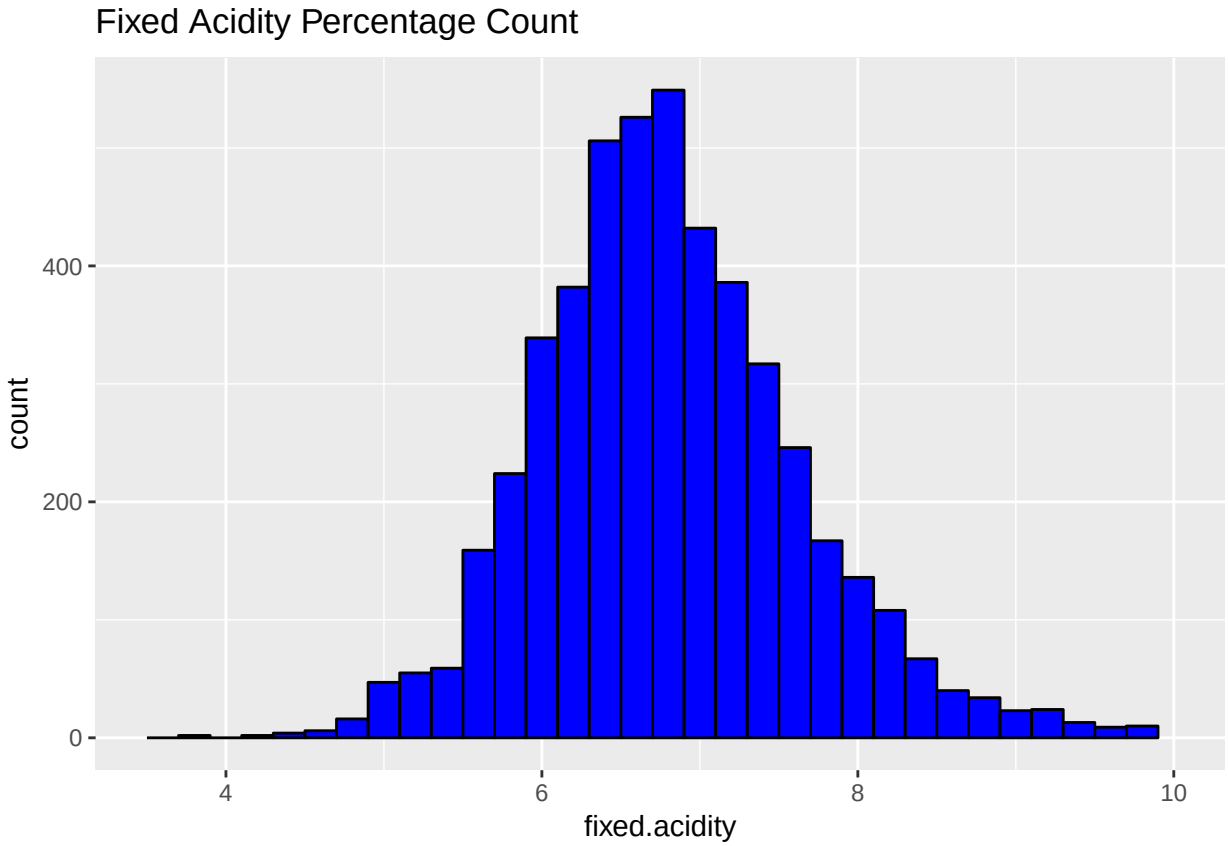
The distribution of alcohol content in the variable. Here we can see that distribution of alohol is a bit skewed to the left, thus showing that there is a higher amount of lower alcohol content wine in the dataset. This will help context when we analyze its rellationship with other variables in later parts of this project.



When we divide the plot by quality given to these specific alcohol content wines we can see that the distribution that we found in the original plot for quality is apparent in 5 and 6, but the distribution for the rest changes. We can see that in plot for quality 8 that there is a slight influx of values with higher alcohol content. This will be explored further in the Bivariate section of the project.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.800   6.300   6.800   6.855   7.300  14.200
```

```
## Warning: Removed 7 rows containing non-finite values (stat_bin).
```



The graph displying the amount showing the amount of fixed acidity in any given wine is a normal distribution. Luckily this makes it easy for use and no scaling needs to be used in order to make this into a reasonable distribution.

Univariate Analysis

What is the structure of your dataset?

The dataset is almost entriely consisted of numbers with only one value being composed of integers.

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

The main feature that interested me when I first looked at this dataset was what factors led to a better tasting wine? If the wine had more acid or more alcohol content, did that often lead to a better tasting wine? The main aspect that I will be focusing on is the relationship between variables in the dataset and the quality rating that the wine was given by experts. As no other variables really count if the wine isn't delicious!

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

The feature that will most likely be helpful to me conducting successful analysis with this data is the fact that they are all numbers. This allows me to easily find correlations between the multiple values. I will be able to perform analysis to see if more acidic win leads to better tasting better tasting wine, or other correlations of this nature.

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

No, I did not see the need to create any new variables yet, perhaps later duirng my analysis I will develop a quesiton that requires me to do so.

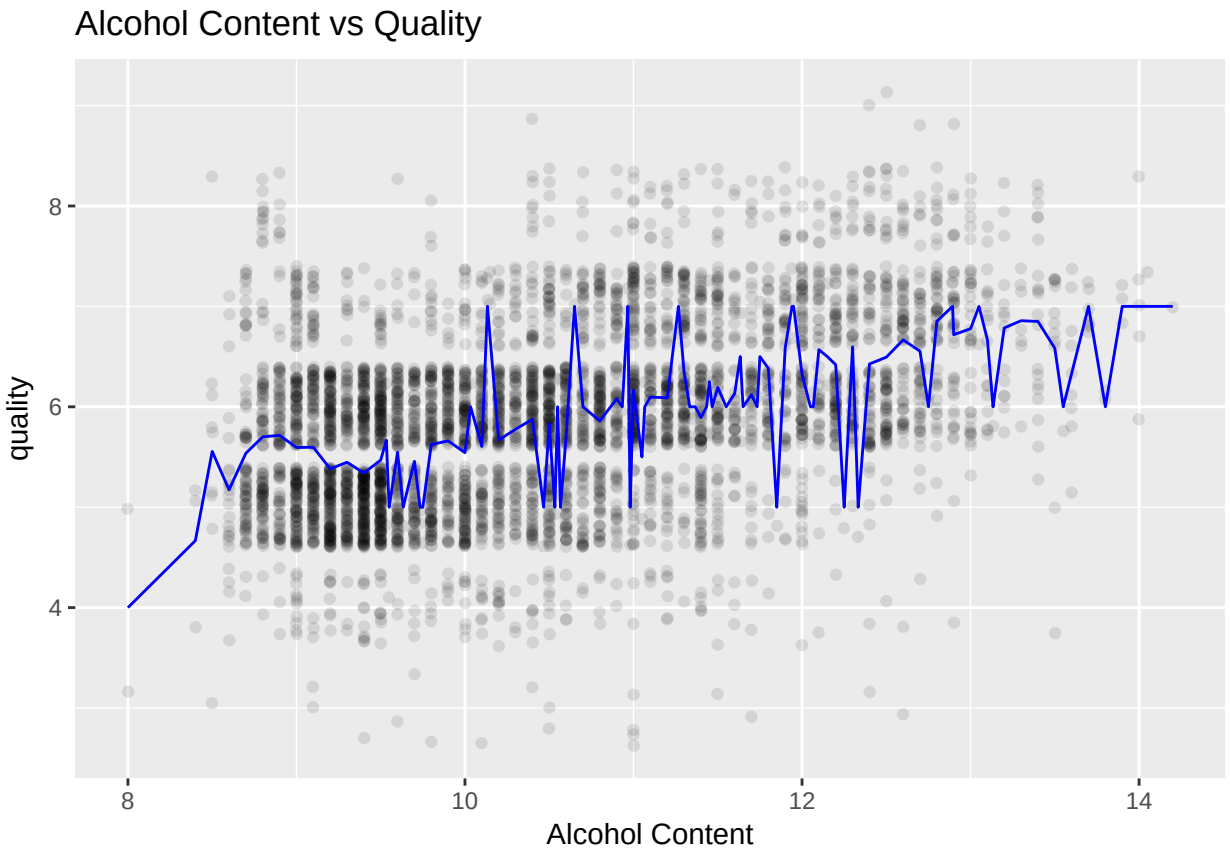
**Of the features you investigated, were there any unusual distributions?
Did you perform any operations on the data to tidy, adjust, or change the form
of the data? If so, why did you do this?**

Bivariate Plots Section

ggplot(aes(x = residual.sugar, y = chlorides), data = wines) + geom_point()

```
##
##  Pearson's product-moment correlation
##
## data:  wines$alcohol and wines$quality
## t = 33.858, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4126015 0.4579941
## sample estimates:
##       cor
## 0.4355747
```

First I tried to determine if there was a correlation between what I believe to be the most importnat values in the dataset; Alcohol and Quality. After running the correlation test I was able to determine that there is fairly large correlation between the two which allowed me to conduct further analysis on the topic with confidence.



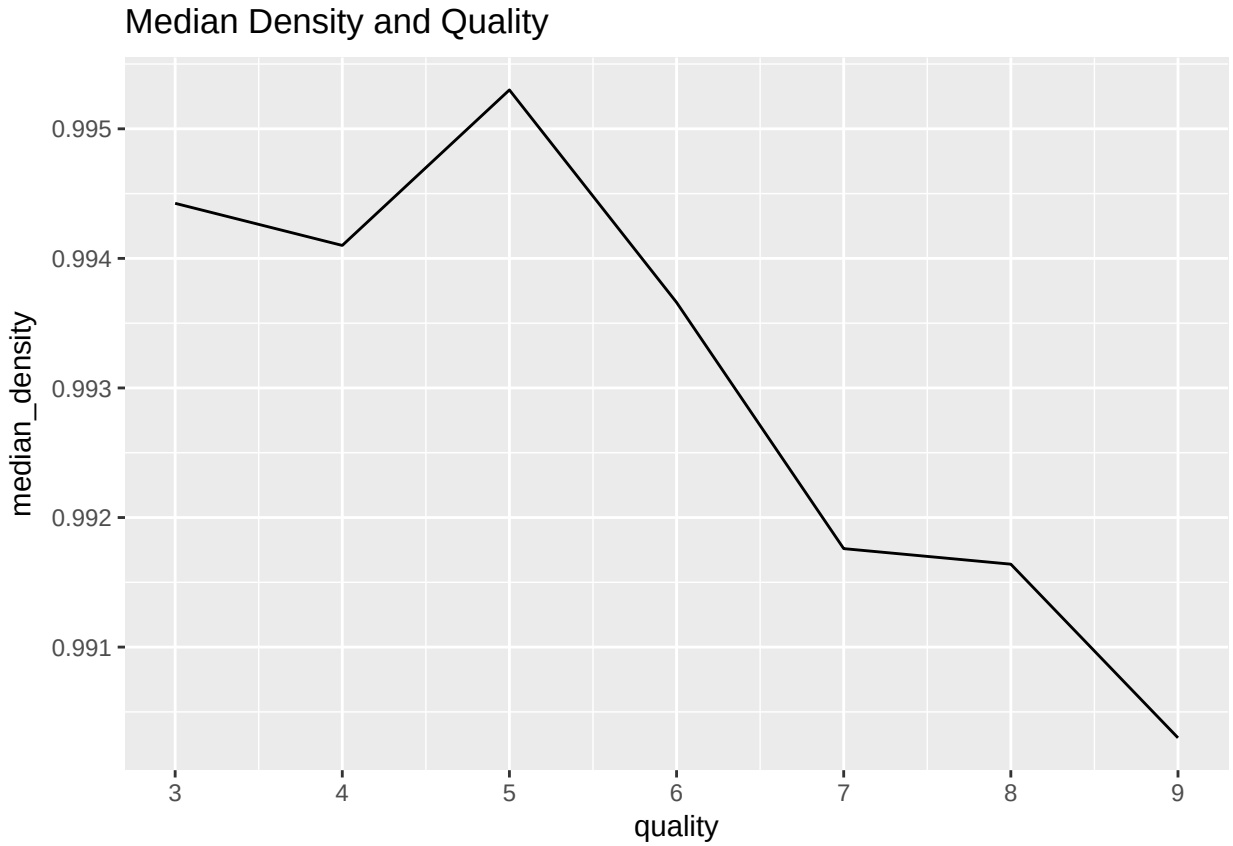
Here I created a scatterplot that graphs alohol and its relationship to quality. The blue line respresents the mean of quality at any any given alcohol level. By looking at the graph we can see that as alcohol content goes up there is an increase in the number of plots with higher quality gradings. The mean of qaulity can also be visibly be seen increasing as alcohol content increases. Now that we know there is a correlation between these two variables let's invistigate further.

```
##
##  Pearson's product-moment correlation
##
## data:  wines$density and wines$quality
## t = -22.581, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3322718 -0.2815385
## sample estimates:
##        cor
## -0.3071233
```

This plot shows the relationship between density and quality. We can kind of see that that there is a higher amount of lower density values at quality levels 8 and 9, however, the plot is pretty noisy and hard to draw a conclusion. Let's try to remedy this.

```
## # A tibble: 6 × 6
##   quality mean_density median_density max_density min_density     n
##     <int>        <dbl>          <dbl>       <dbl>       <dbl> <int>
## 1       3    0.9948840       0.994425     1.00010     0.99110    20
## 2       4    0.9942767       0.994100     1.00040     0.98920   163
## 3       5    0.9952626       0.995300     1.00241     0.98722  1457
## 4       6    0.9939613       0.993660     1.03898     0.98758  2198
## 5       7    0.9924524       0.991760     1.00040     0.98711   880
## 6       8    0.9922359       0.991640     1.00060     0.98713   175
```

Due to the fact that there was a minor correlation between density and quality of the wine, I decided to create a new dataframe that displayed the summary statistics for quality and its relationship with density.



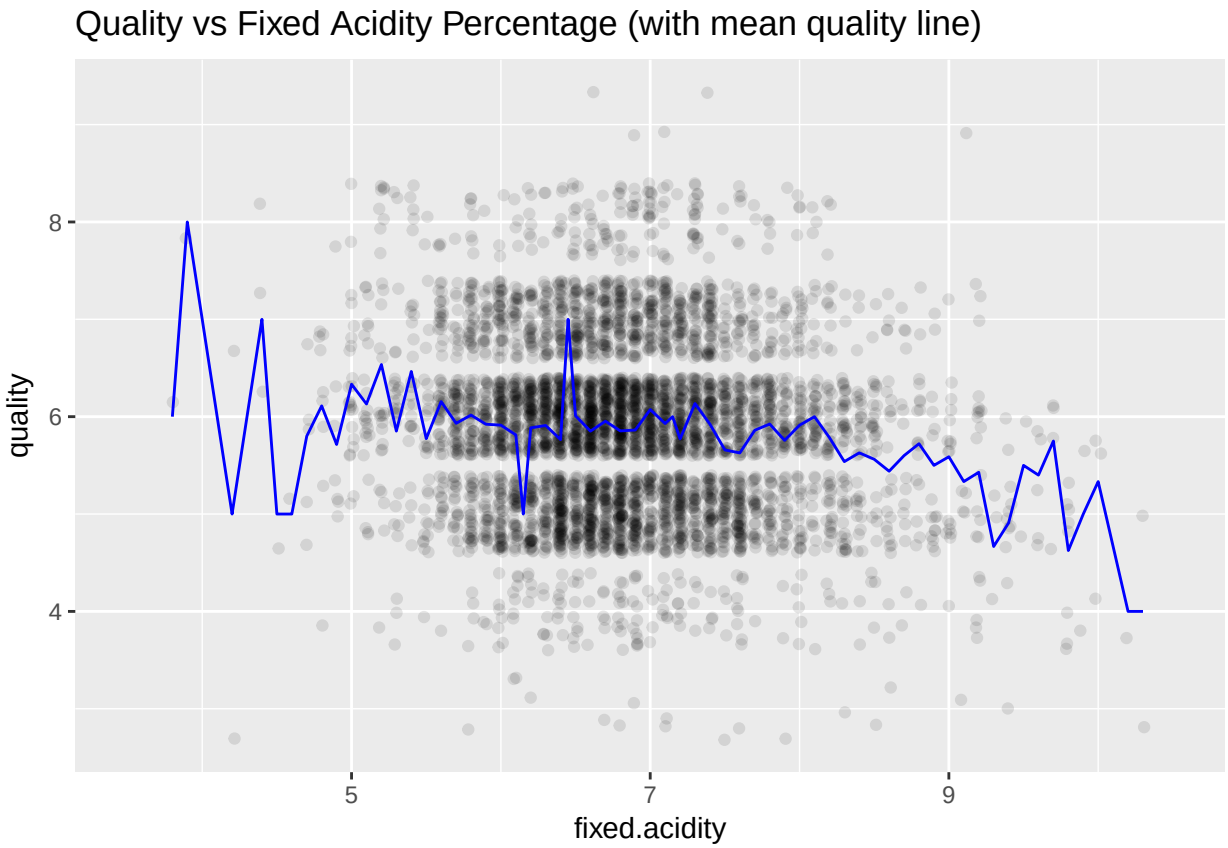
Here we can see through plotting the median density of the wines that it has a direct inverse relationship with teh quality of wine. As the density of the wine decreases we can see a clear relationship with quality increasing as well. Hence, making it likely that density is a direct factor in contributing to the quality and taste of the wine. This conveys it much clearer than the previous plot and eliminated much of the noise.

```
##
##  Pearson's product-moment correlation
##
## data:  wines$fixed.acidity and wines$quality
## t = -8.005, df = 4896, p-value = 1.48e-15
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.14121974 -0.08592991
## sample estimates:
##        cor
## -0.1136628
```

The correaltion between fixed acidity and quality is relatively shocking. One would believe that as acidity went up there would be a drastic change in how quality would be perceived. However, based on the correlation between the two we can infer that there is little correlation between the two. This is shown further in the plot below.

```
## Warning: Removed 4 rows containing non-finite values (stat_summary).
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```



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?

The main feature that I wanted to delve into during this part of the analysis was the realtionship quality had with other variables that I believed would have the greatest affect on what score it was given. These were quality and its relationship with alcohol content, density and acidity. Density and alcohol content behaved the way I expected, each directly correlating to an increase in quality, however, I did not expect for acidity to have little to no effect on how quality was perceived. I expected an increase in acidity to have a drastic effect on decrease in quality score given, however the effect was nominal at best.

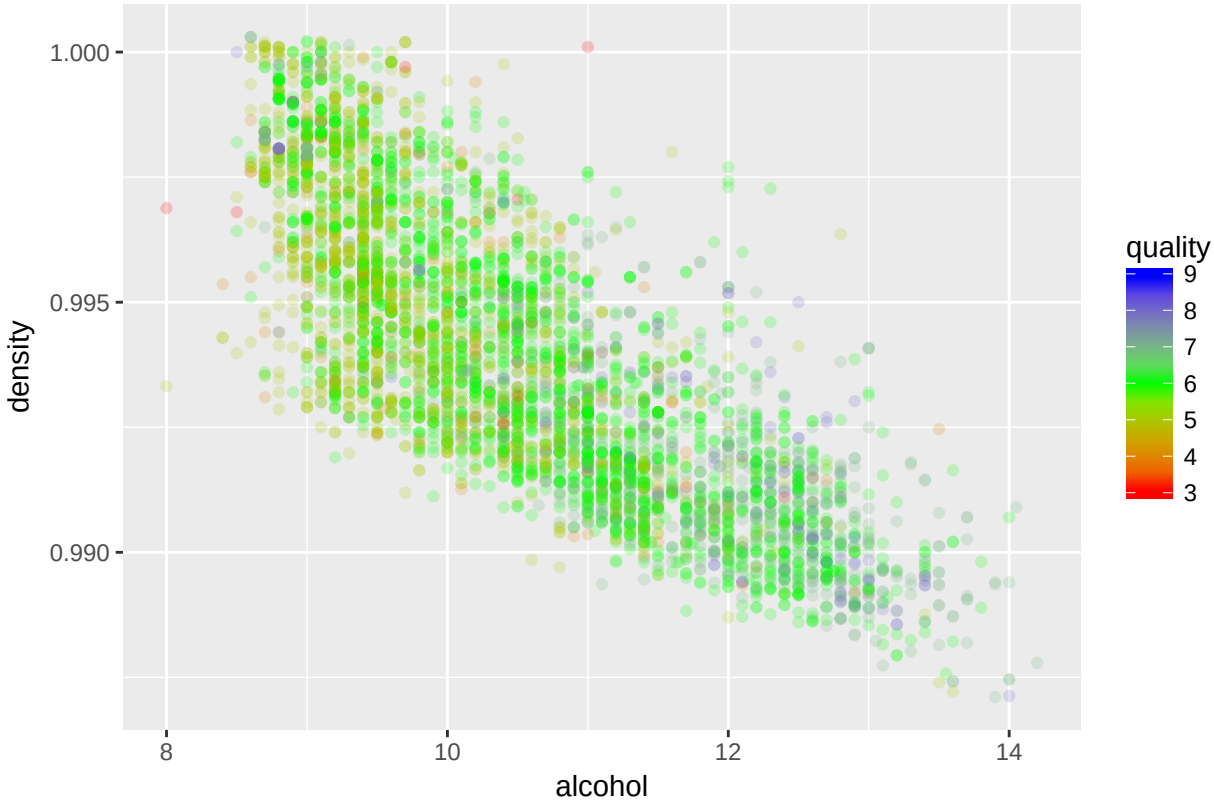
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

What was the strongest relationship you found?

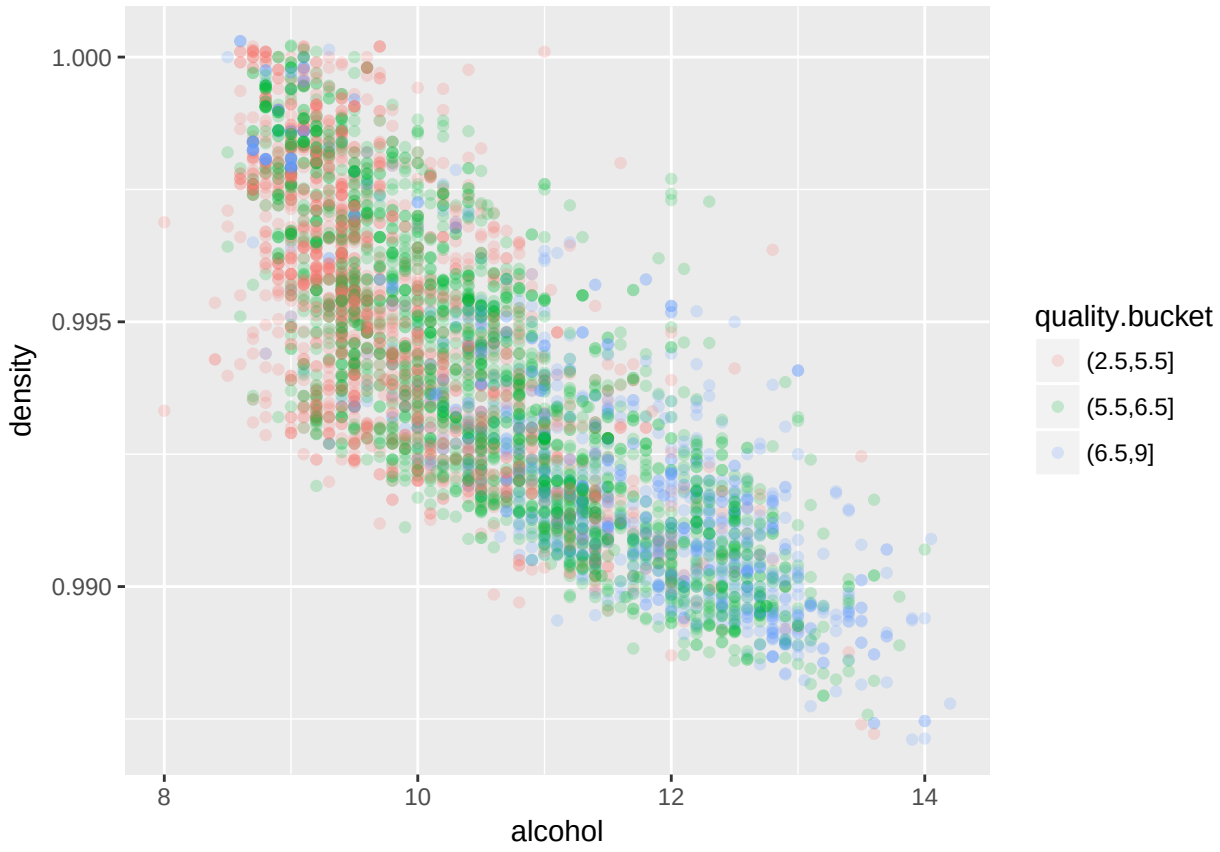
Multivariate Plots Section

Tip: Now it's time to put everything together. Based on what you found in the bivariate plots section, create a few multivariate plots to investigate more complex interactions between variables. Make sure that the plots that you create here are justified by the plots you explored in the

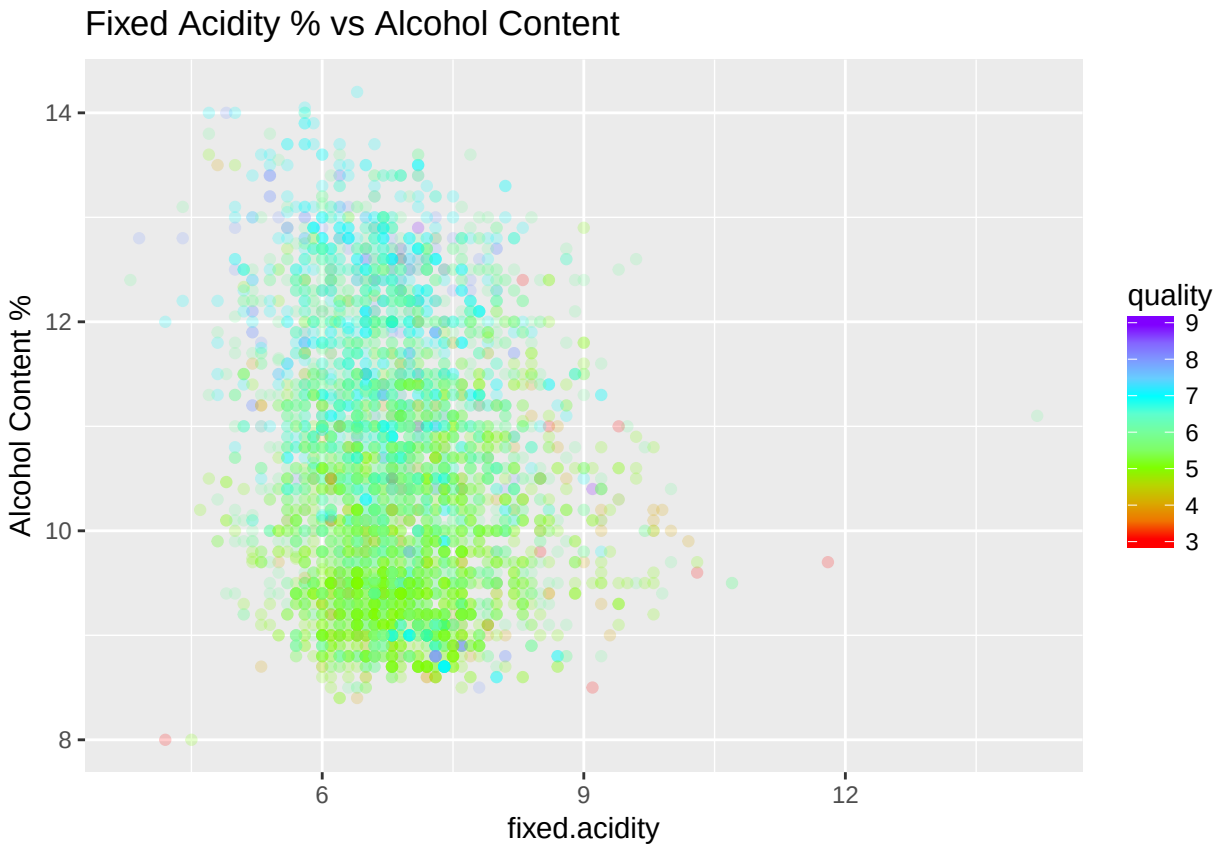
previous section. If you plan on creating any mathematical models, this is the section where you will do that.



The plot shows that there is a correaltion between the 3 variables. We can observe that the higher quality values are where we though they would be due to the previous graphs. They are in the location where alcohol content is higher and density is near its lowest. However, the plot is relatively noisy, so I will attempt to clean it up.



Here we graphed the corelation between densisty and alcohol content with the color differentiated by the quality bin that the specific wine was given. We can clearly see that there is a correlation between alcohol content/density of the wine with the quality that it is given. The majority of the wines that were given the highest ratings (between 7 & 9) were in the lower part of the graph where there is the highest alcohol content and the lowest density of the wines. Basically just showing a clearer relationship with density, alcohol and quality.



Here in this graph we can durther see the relationsionships that we found earlier in the bivariate plot section. We can see that fixed acidity doesn't really play a large part in the quality that the wine is given, but alcohol content still remained a pivitol factor.

```
##
## Calls:
## m1: lm(formula = quality ~ alcohol, data = wines)
## m2: lm(formula = quality ~ alcohol + density, data = wines)
## m3: lm(formula = quality ~ alcohol + density + fixed.acidity, data = wines)
##
## ============================================================
##                        m1           m2           m3
## ------------------------------------------------------------
##   (Intercept)       2.582***   -22.492***   -32.669***
##                    (0.098)      (6.165)      (6.360)
##   alcohol           0.313***     0.360***     0.373***
##                    (0.009)      (0.015)      (0.015)
##   density                       24.728***    35.427***
##                                 (6.079)      (6.300)
##   fixed.acidity                              -0.087***
##                                              (0.014)
## ------------------------------------------------------------
##   R-squared             0.2          0.2          0.2
##   adj. R-squared        0.2          0.2          0.2
##   sigma                 0.8          0.8          0.8
##   F                  1146.4       583.3        404.5
##   p                     0.0          0.0          0.0
##   Log-likelihood    -5839.4      -5831.1      -5812.2
```

```
##   Deviance          3112.3     3101.8      3077.9
##   AIC              11684.8    11670.3     11634.4
##   BIC              11704.3    11696.2     11666.8
##   N                 4898       4898        4898
## ====================================================
```

Here we have created a linear model with the goal of being able to predict what quality that the judges would give the wine based off of the main variables that we delved into during our exploration. Unfortunately the R squared values are extremely low for each of the models so we cannot be certain, or have very high confidence at all that these models would accurately predict the quality at any given point.

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?

In this part of the analysis I was able to further create a correlation with the variables that had the greatest affect on quality, density and alcohol content. Here we can see that it is very unlikely that a wine would get one of the highest ratings if it did not have a density below .993 and an alcohol content above 12%. The buckets allow us to see that more clearly.

Were there any interesting or surprising interactions between features?

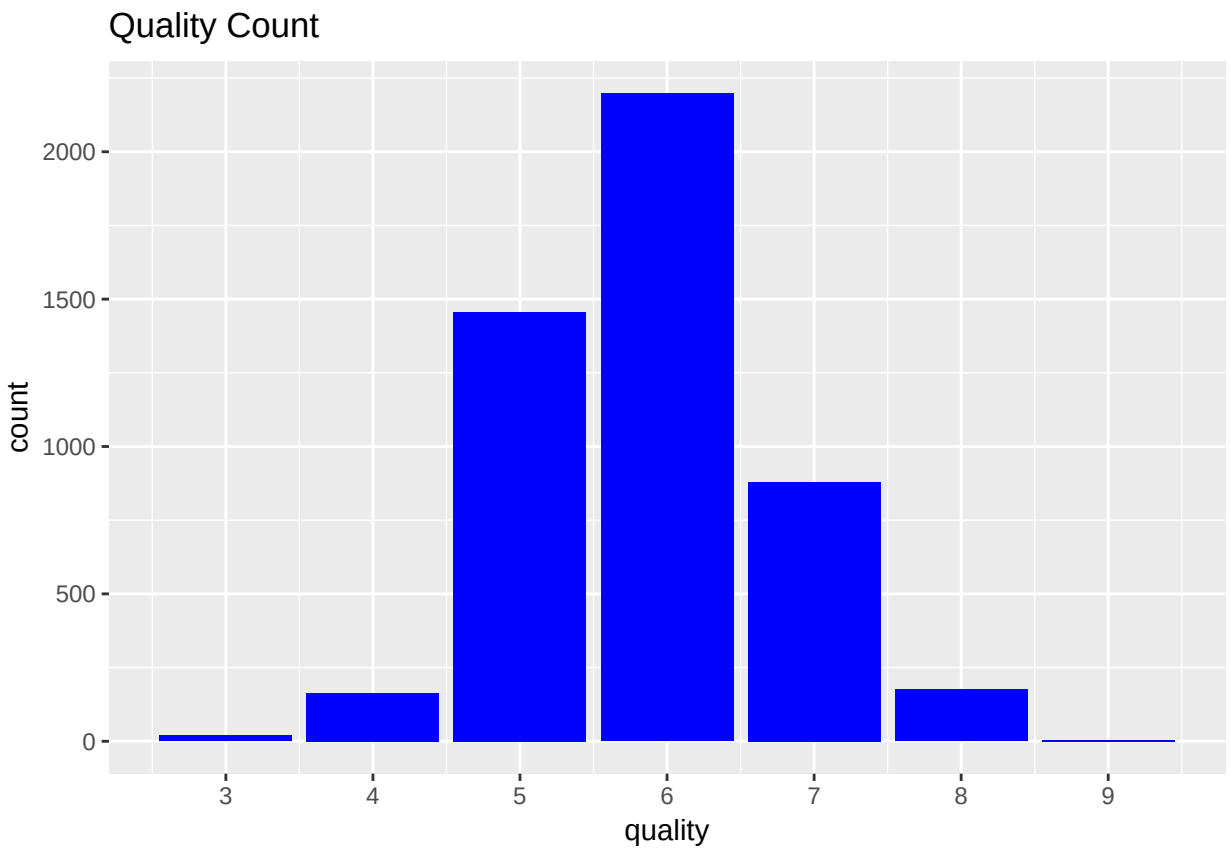
Another observation that I made, was the relationship between citric acid and quality. Before I though that there was a greater relationship with citric acid and the quality rating given. But based on the plots we can see that there isn't a great difference between high acidic content and low acidic qualities. In fact it was mostly up to the other variables that determined what quality rating it was given, thus ruining my initial hypothesis that fixed acidity would be a major component in scores.

OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model.

The linear model that I created has some limitations. This is mostly due to the R squared value being so low. This is most likely due to the fact that the dependent variable, quality, being determined with extreme bias by human judges. The bias that they have in their ratings is prevalent in the R squared and makes it limited on how much we can predict. In addition, we don't know if it is the same judges ranking the wines everytime, so this could create inconsistencies if a new person/group is reviewing the wines at every given time.

Final Plots and Summary

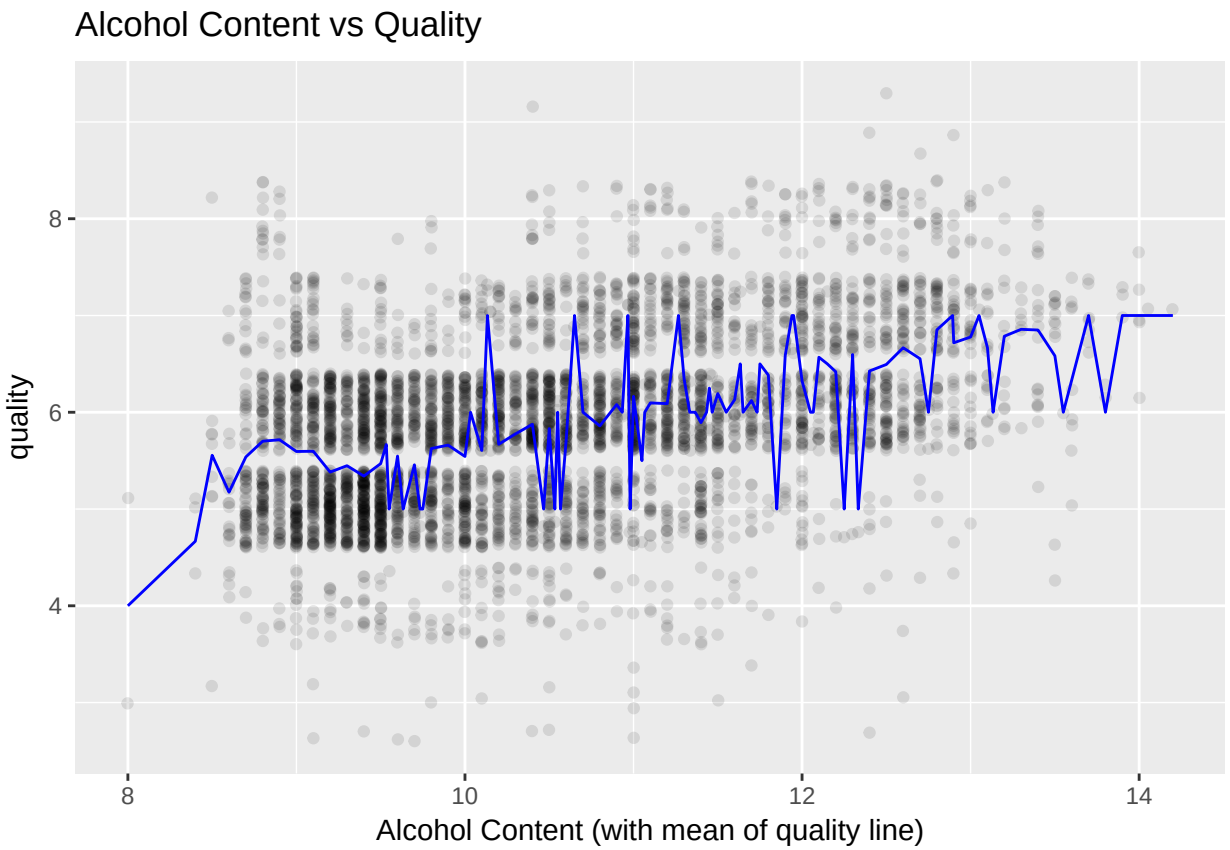
Plot One



Description One

The first plot that I chose was a histogram of the quality levels. Here we can observe that there is uniform distribution of quality levels assigned to the wines that the judges were given to rate. This distribution is most likely to be a result of human tendacy to usually give ratings like this due to their innate bias to not have to many high or low ratings. This plot gives us a general idea of how to pursue our future analysis since they will be mostly based on how variables affect the quality of wines.

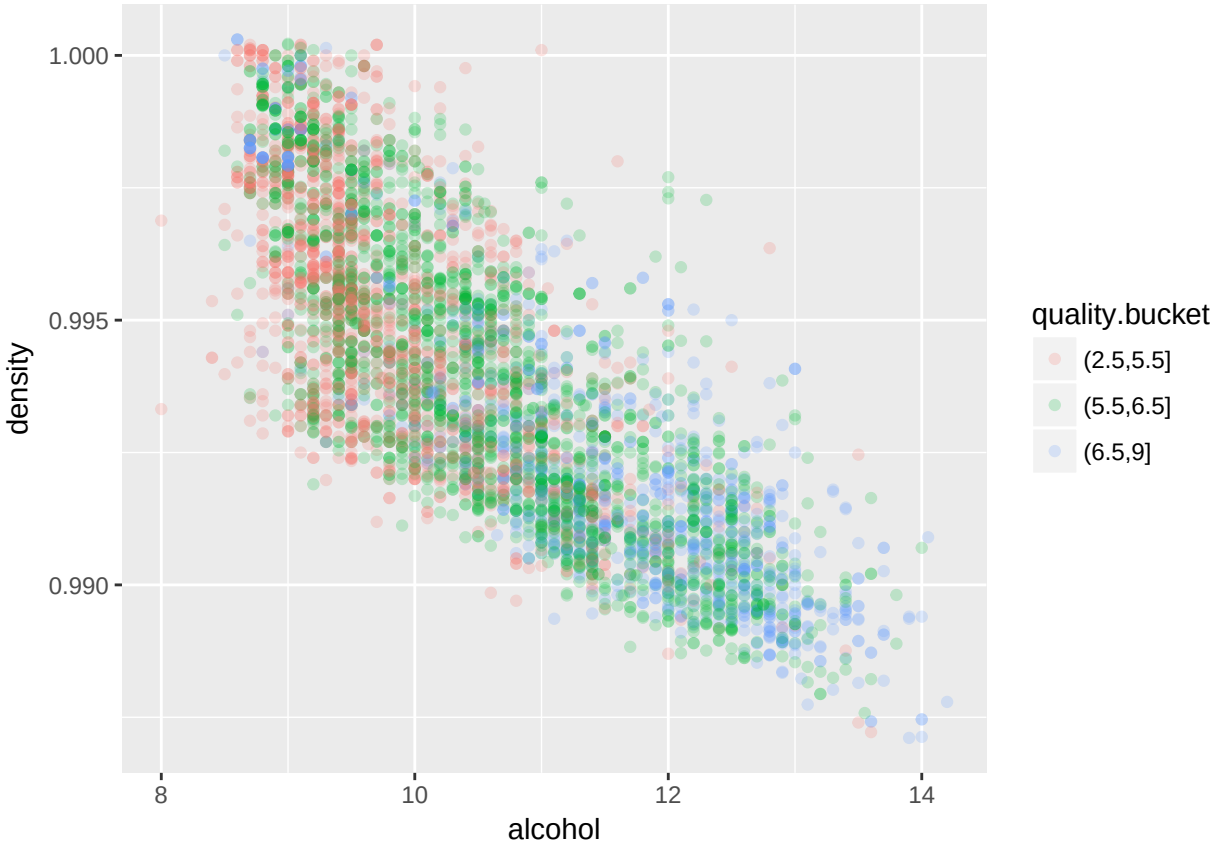
Plot Two



Description Two

Here I created a scatterplot that graphs alohol and its relationship to quality. The blue line respresents the mean of quality at any any given alcohol level. By looking at the graph we can see that as alcohol content goes up there is an increase in the number of plots with higher quality gradings. The mean of qaulity can also be visibly be seen increasing as alcohol content increases. This a more clear way of seeing that there is a innate positive relationship between the two variables.

Plot Three



Description Three

In this third plot we are able to see the direct correlation that density and alcohol content has with the quality that they are assigned. The highest concentration of high quality bucket variables (blue) are found in the lower right. This coincides with the idea that lower density and higher alcohol content is a more desirable trait when looking for better tasting wine. This gives further credence to the idea tht alcohol content is a necessary aspect of wine grading and displays the affect low density has as well.

Reflection

Coming into the dataset I had many preconceived notions on what created a great white wine. I already came in with the belief that a lighter, more smooth wine would be rated higher in the quality taste tests. My belief was validated. Through the use of univariate analysis I was able to create histograms that gave a brief, yet informative viewpoint on how the wines would regulated and broken up into. After that the bivariate analysis, which resulted in me creating scatterplots showed me that there was a greater relationship with wine attributes and quality than I could've imagined. I initially thought it would be a fairly normal distribution, with greater alcohol content always receiving higher marks (but that could just be the inner college student in me), but I was surprised to see that a myriad of things went into the quality rating given to the wine. I created a new dataset that had the summary stats for quality and used that in order to more accurately portray what relationships that variables had with it. Later on in the multivariate portion I found this plots that had quality divided by color to be better visual represenations of what went into a quality

wine than the ones with only two variables. I believe this to be the case that even though they were saying mostly things we had already visited in previous plots, the ability to see multiple relationships at once (such as density and alcohol content on quality) allows the reader to better visualize the wine that they would be tasting in that instance. One thing I was disappointed with was that the linear model created did not have a high R squared, at which point I realized it was most likley due to the high bias in the judges that gave them the score. Since there was no rubric they were following the bias made it nearly impossible to accurately predict what scores would be given out to any given wine.