Udacity_EDA/White_Wine

Helen Nguyen

6/7/2018

Wine Exploratory Data Analysis by Helen Nguyen

This Udacity project applies EDA using R to analyze a white wine dataset. The main objective is to understand and determine which variables affect wine quality.

```
#Load the Data
ww <- read.csv('/Users/Helen/Desktop/udacityR/wineQualityWhites.csv')

#Remove unnecessary X column
ww <- ww %>% dplyr::select(-X)
```

Univariate Plots Section

Let's take a look at dimensions, structure, and summary of the dataset

```
## [1] 4898   13

## 'data.frame':    4898 obs. of  13 variables:
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3
0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34
0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045
0.045 0.049 0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22
...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49
0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : Ord.factor w/ 7 levels "3"<"4"<"5"<"6"<..: 4 4 4
4 4 4 4 4 4 4 ...
##  $ rating              : Ord.factor w/ 3 levels "bad"<"average"<..: 2 2 2
2 2 2 2 2 2 2 ...
```

```
##  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
##  Min.   :0.00900   Min.   :  2.00      Min.   :  9.0
##  1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0
##  Median :0.04300   Median : 34.00      Median :134.0
##  Mean   :0.04577   Mean   : 35.31      Mean   :138.4
##  3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0
##  Max.   :0.34600   Max.   :289.00      Max.   :440.0
##
##     density             pH          sulphates         alcohol
##  Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00
##  1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50
##  Median :0.9937   Median :3.180   Median :0.4700   Median :10.40
##  Mean   :0.9940   Mean   :3.188   Mean   :0.4898   Mean   :10.51
##  3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40
##  Max.   :1.0390   Max.   :3.820   Max.   :1.0800   Max.   :14.20
##
##  quality        rating
##  3:  20   bad      : 183
##  4: 163   average  :3655
##  5:1457   excellent:1060
##  6:2198
##  7: 880
##  8: 175
##  9:   5
```

Wine quality has a median value of 6 with a minimum of 3 and a maximum of 9.

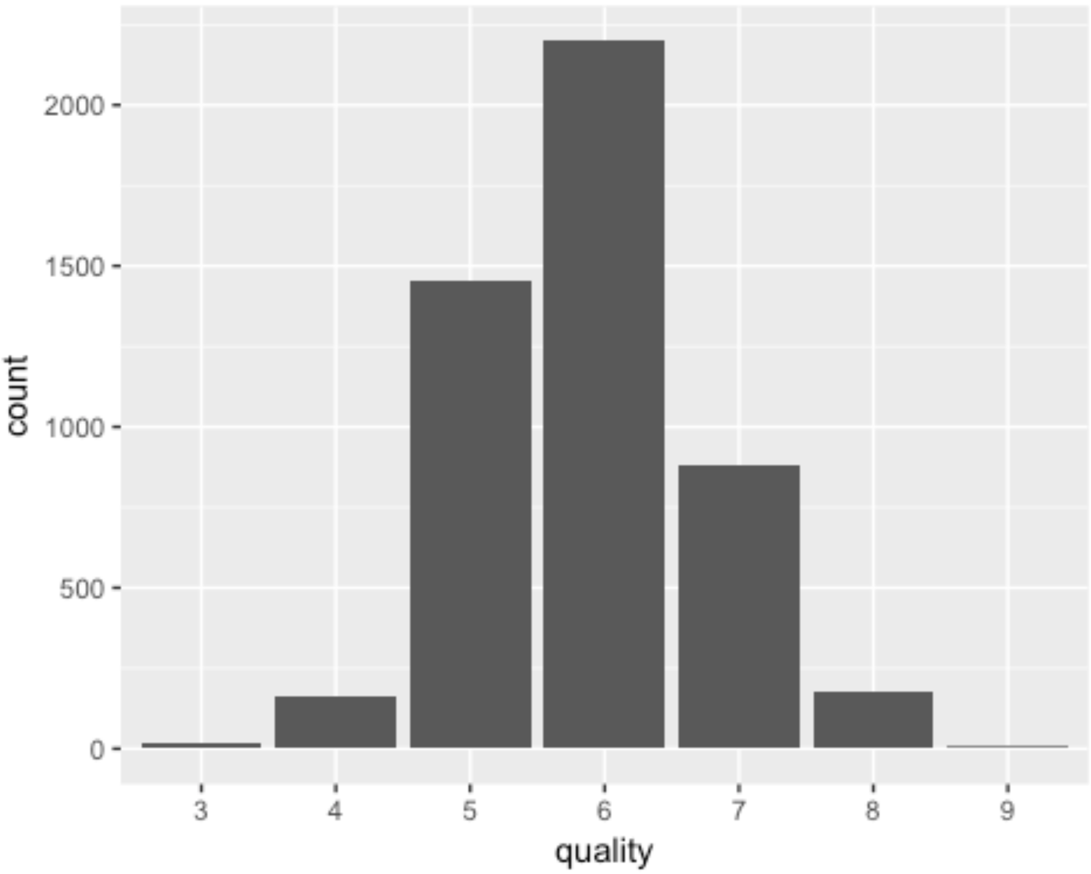
Fixed acidity has a high maximum 14.2 while it’s mean is 8.9 and minimum is 3.8.

Alcohol has a mean of 10.5% with a minimum of 8.0% and a maximum of 14.2%.

Under residual.sugar, there is an unusually high maximum of 65.8 which means that a white wine in the dataset contains a much larger concentration of sugar than the rest. A wine with more than 45 grams/liter of sugar is considered sweet.

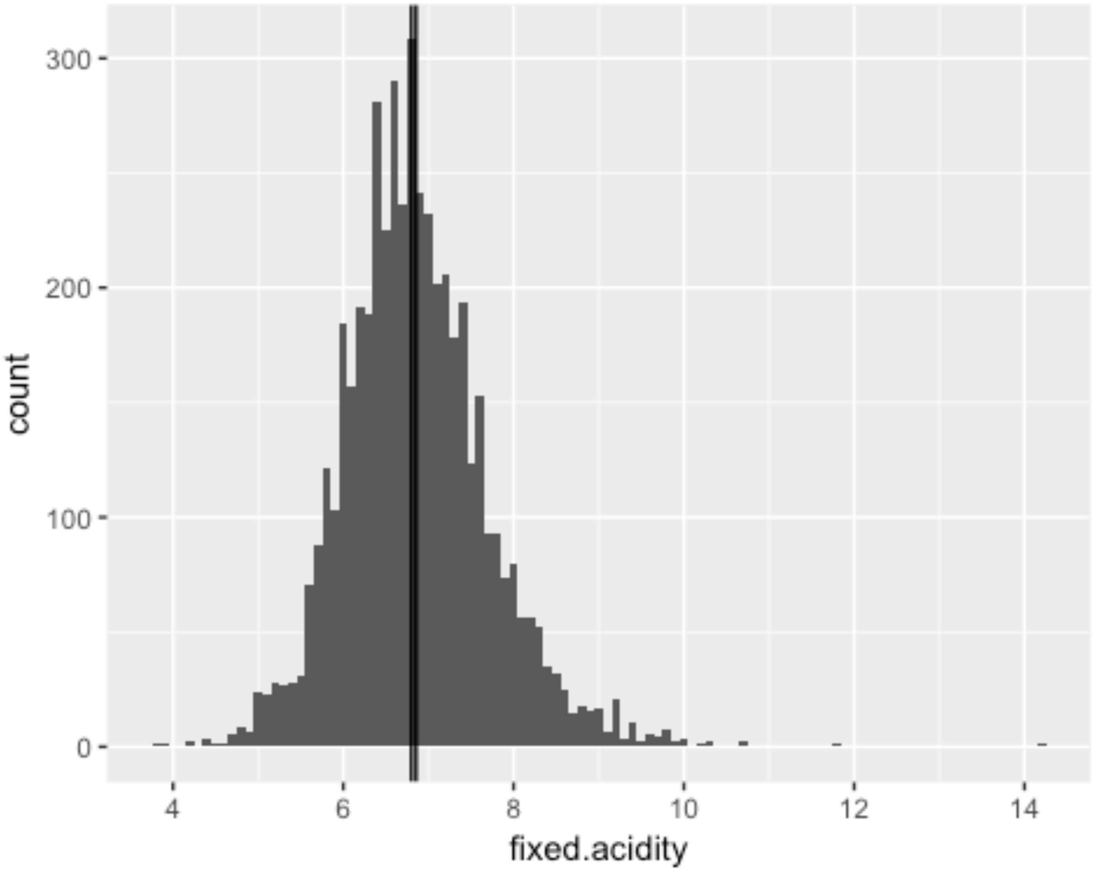
Density values fall between .99 and 1.

Now that we have a summary of these variables, creating plots will allow us to view their distribution.



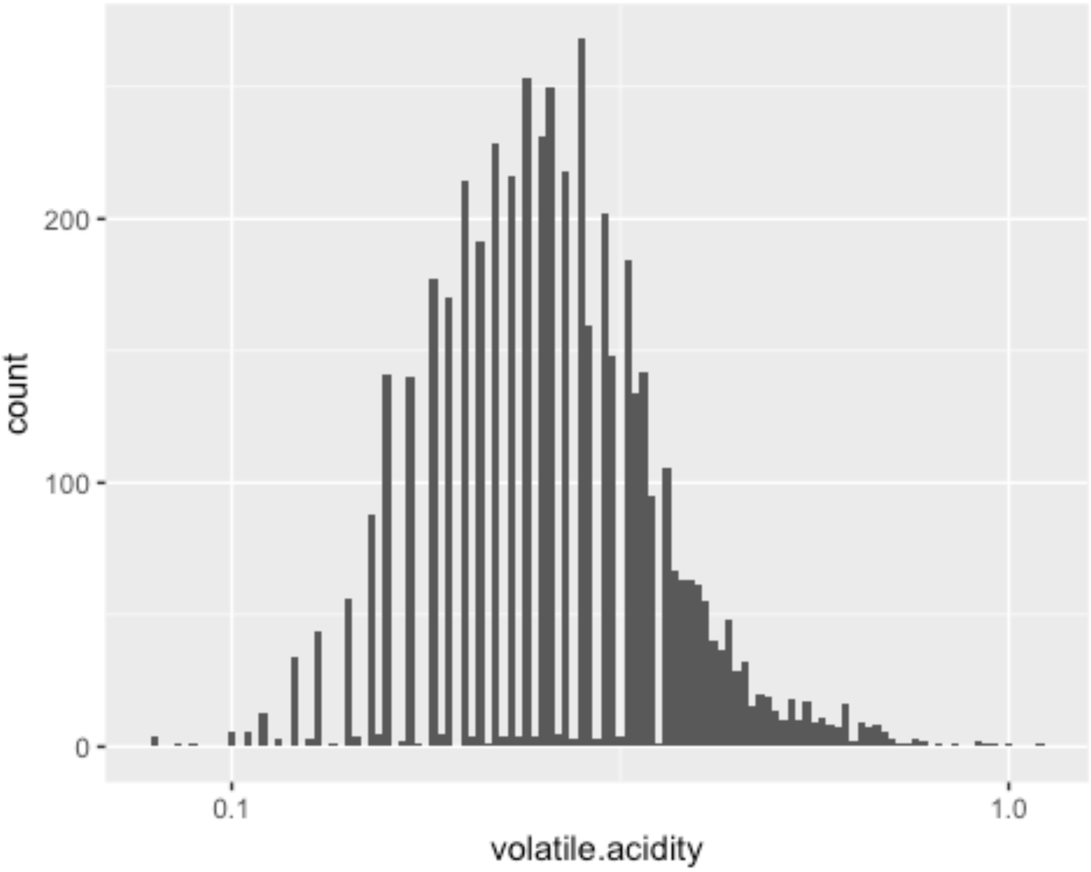
6 is the most common value for wine quality followed by 5 and 7. This tells us that most wines are average. Only a few wines were rated at the opposite ends of the quality scale.

Due to most wines being average, it may affect the accuracy of our model.



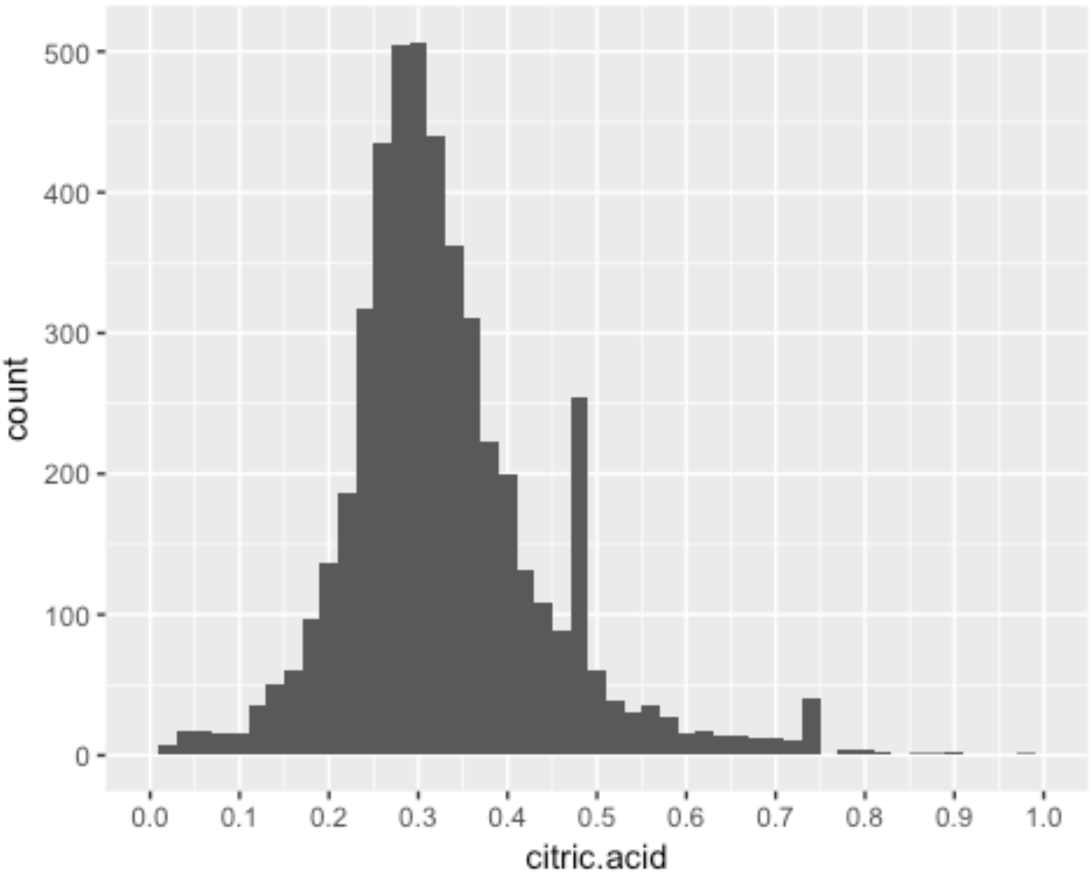
Fixed acidity shows a normal distrubution with mean (6.86) and median(6.80) values that are very close.

```
ggplot(aes(x = volatile.acidity), data = ww) +
  geom_histogram(binwidth = 0.01) +
  scale_x_log10()
```



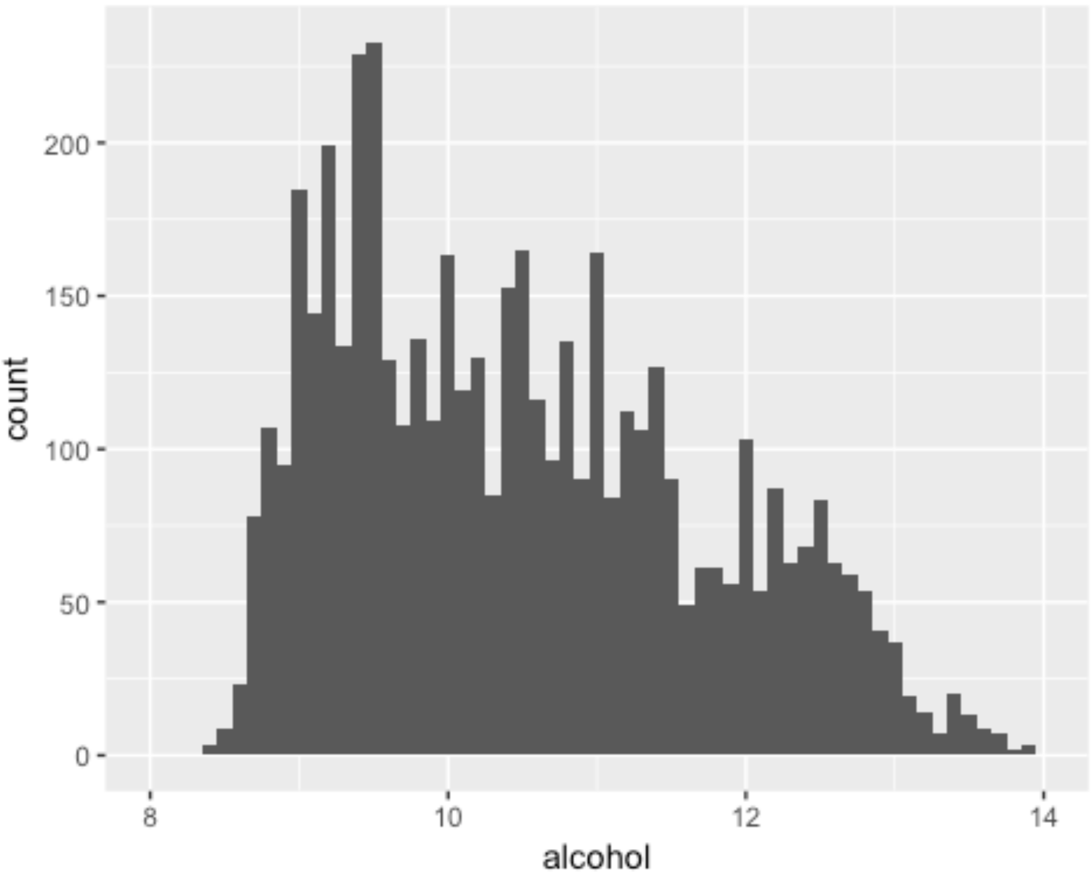
Volatile acidity has a normal distribution after applying the log function and removing outliers.

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```



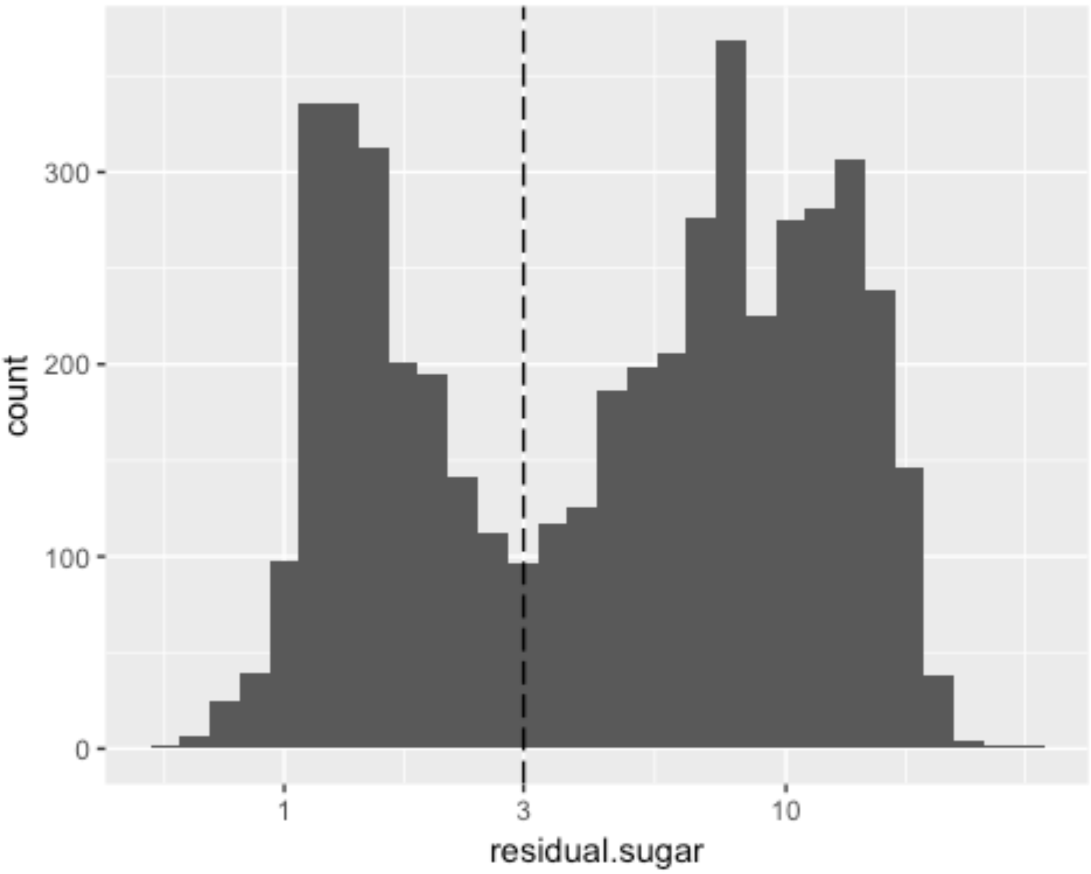
Citric acid appears to also have a normal distribution with many outliers.

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```

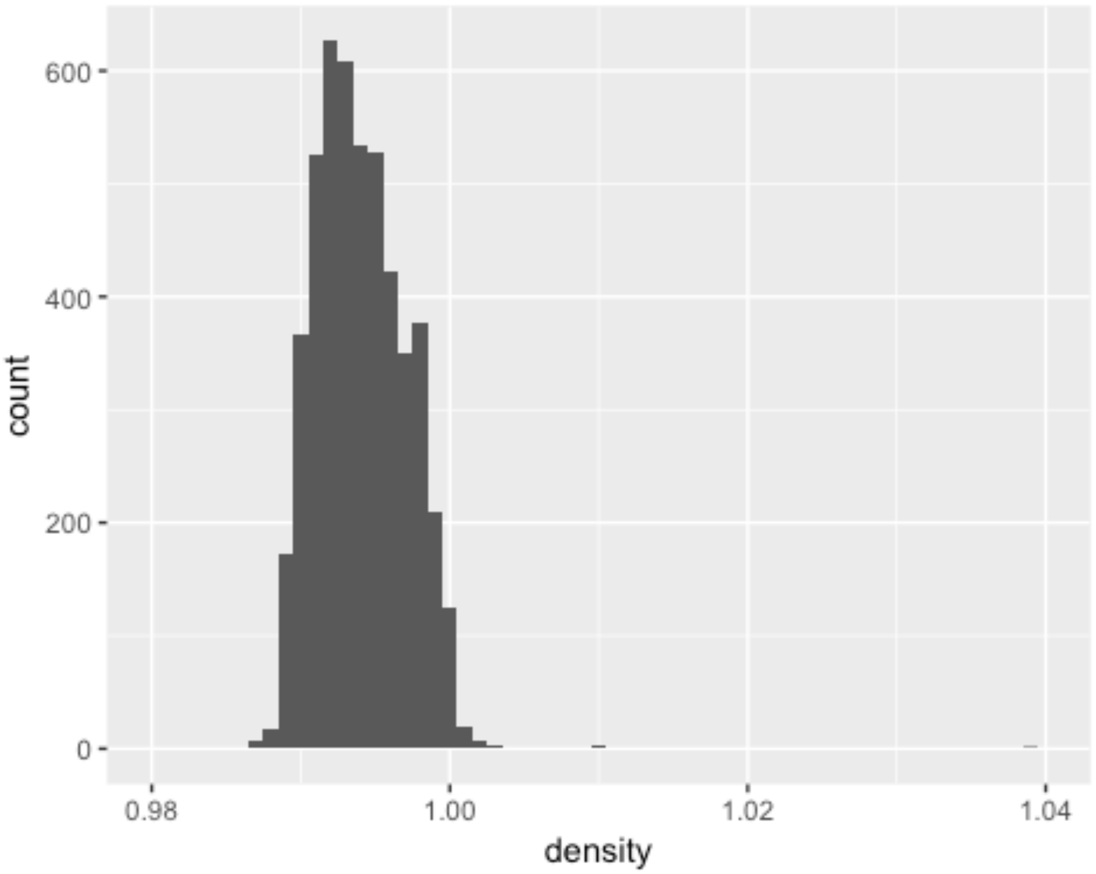


Alcohol does not have a normal distribution with multiple peaks between 8.0% and 14.2%.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Residual sugar has a bimodal distribution with peaks before and after 3.



Density appears to have a normal distribution and very few outliers.

Univariate Analysis

Dataset Structure

The dataset contains 4,898 wines along with 11 quantitative variables. In addition to these variables, the dataset also includes quality and rating (created above). At least 3 wine experts rated the quality of each wine with a rating between 0 (very bad) and 10 (very excellent).

Most wines appear to be average quality with a few being very bad or excellent. Due to this finding, it may be difficult to create a predictive model since there's not enough data on excellent and bad wines.

Main Feature in Dataset

The main feature I'm interested in is quality. The goal is to determine which variables affect wine quality.

Hypothesis

I think alcohol and residual sugar have an impact on wine quality.

New Variable

I created an additional variable called “rating” to label wines as either bad, average, or excellent.

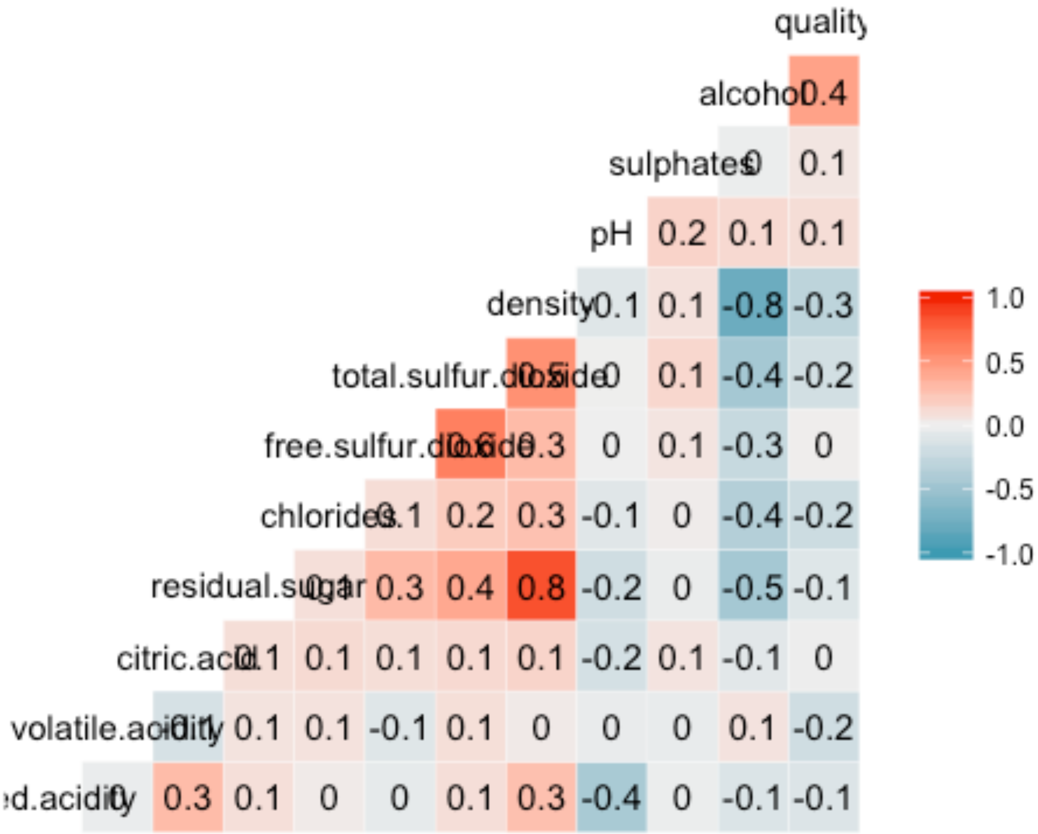
Unusual Distributions

Volatile acidity and citric acid have a significant number of outliers. I had to apply a log function to remove the outliers for volatile acidity to see that it has a normal distribution.

Bivariate Plots Section

Now that we’ve looked at these variables individually, let’s take a look at the correlation between them.

```
## Warning in ggcorr(ww, label = TRUE): data in column(s) 'rating',
## 'numQuality' are not numeric and were ignored
```



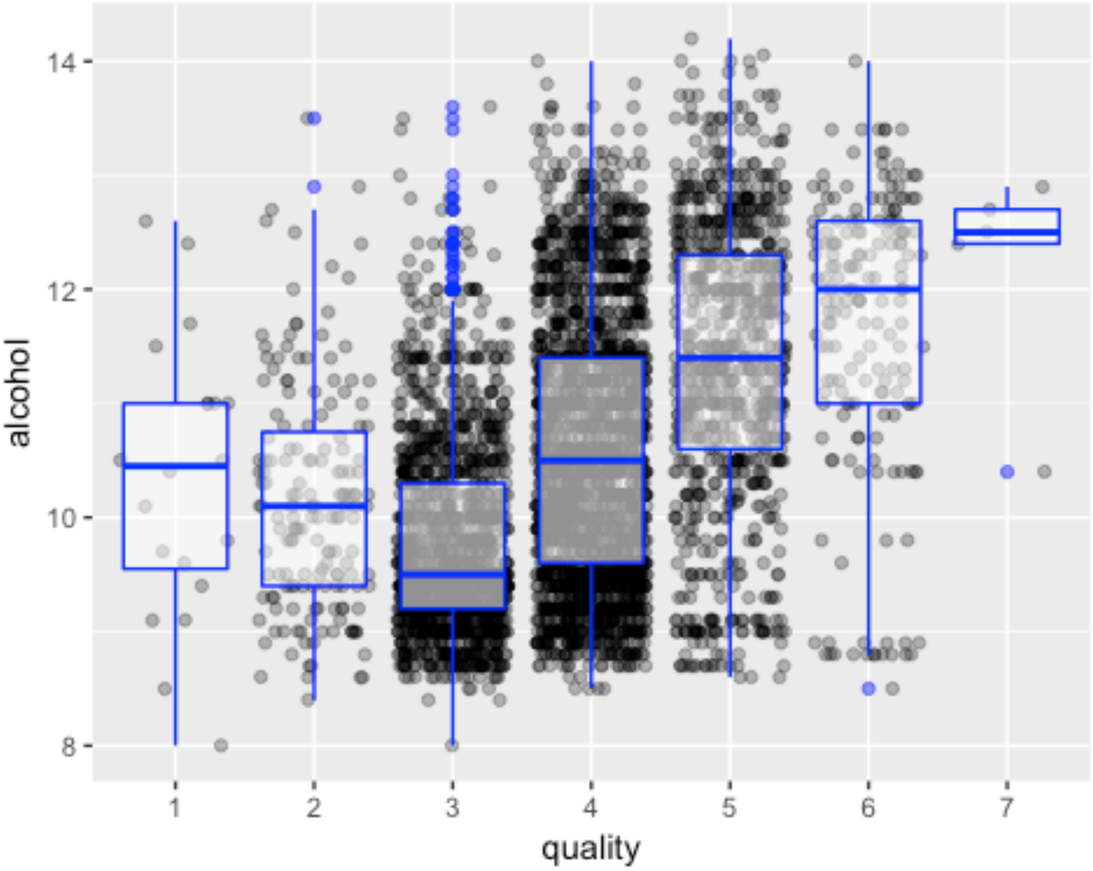
Correlation with Quality

• Alcohol is the most positively correlated with quality (.4) • Density and quality have the strongest negative correlation (-.3) • Residual sugar and quality have a sliglyly negative correlation (-.1) • Volatile acidity and fixed acidity have a slightly negative correlation (-.2) with quality

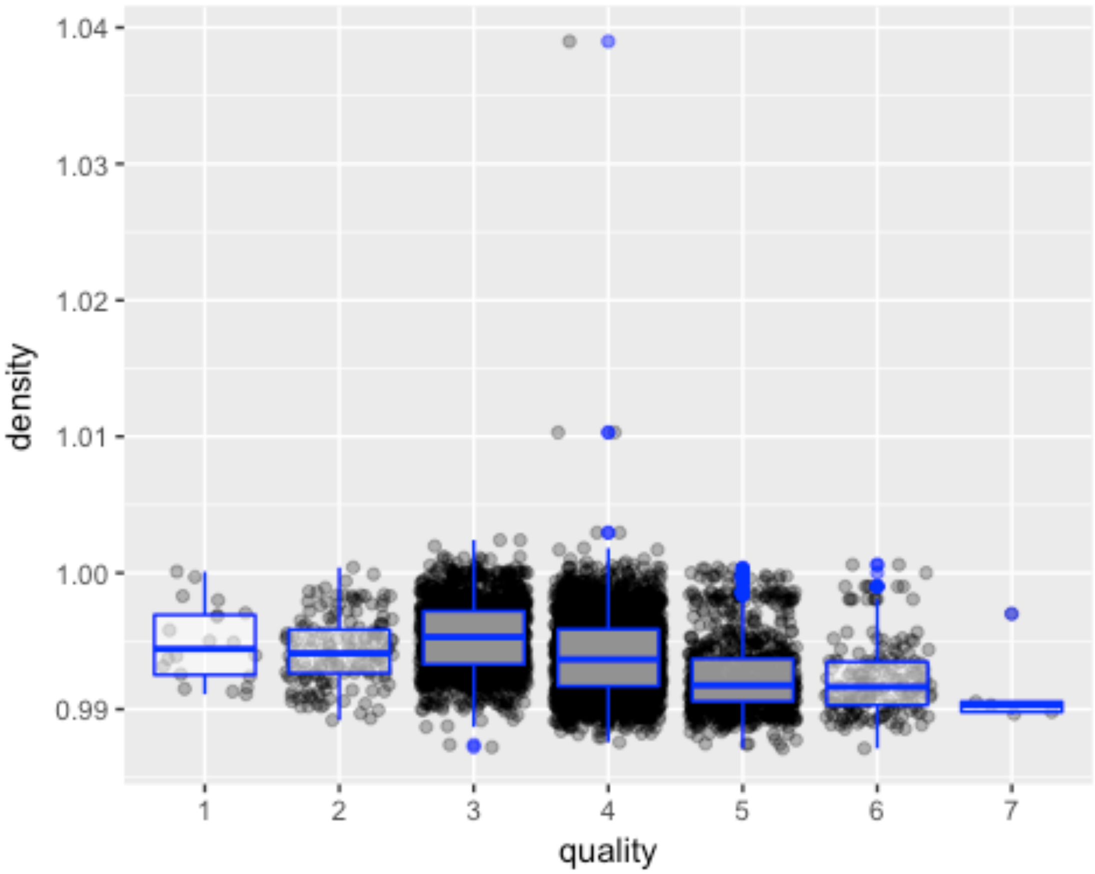
Correlation among Variables

• Density and residual sugar have a strong positive correlation (.8) • Density and alcohol have a strong negative correlation (-.8) • Alcohol and residual sugar have a negative correlation (-.5)

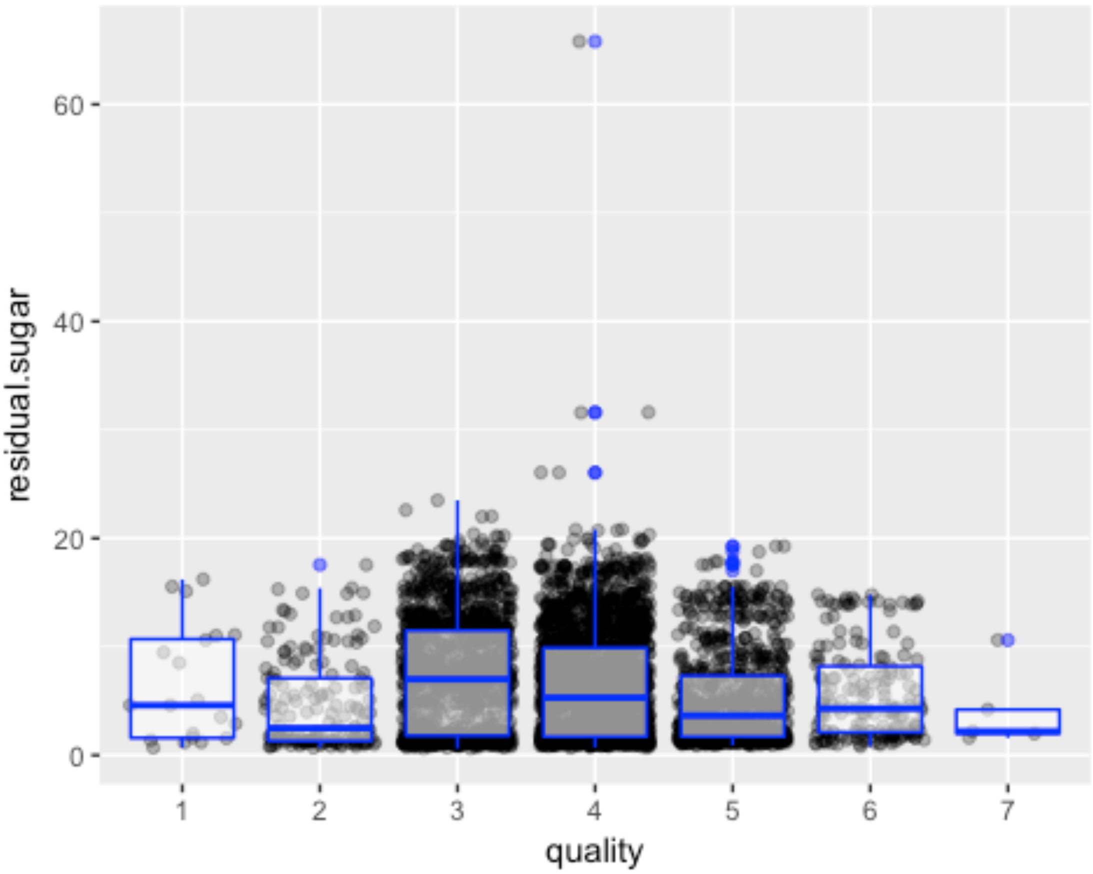
In addition to the correlation table, boxplots will help us explore the relationships among these variables.



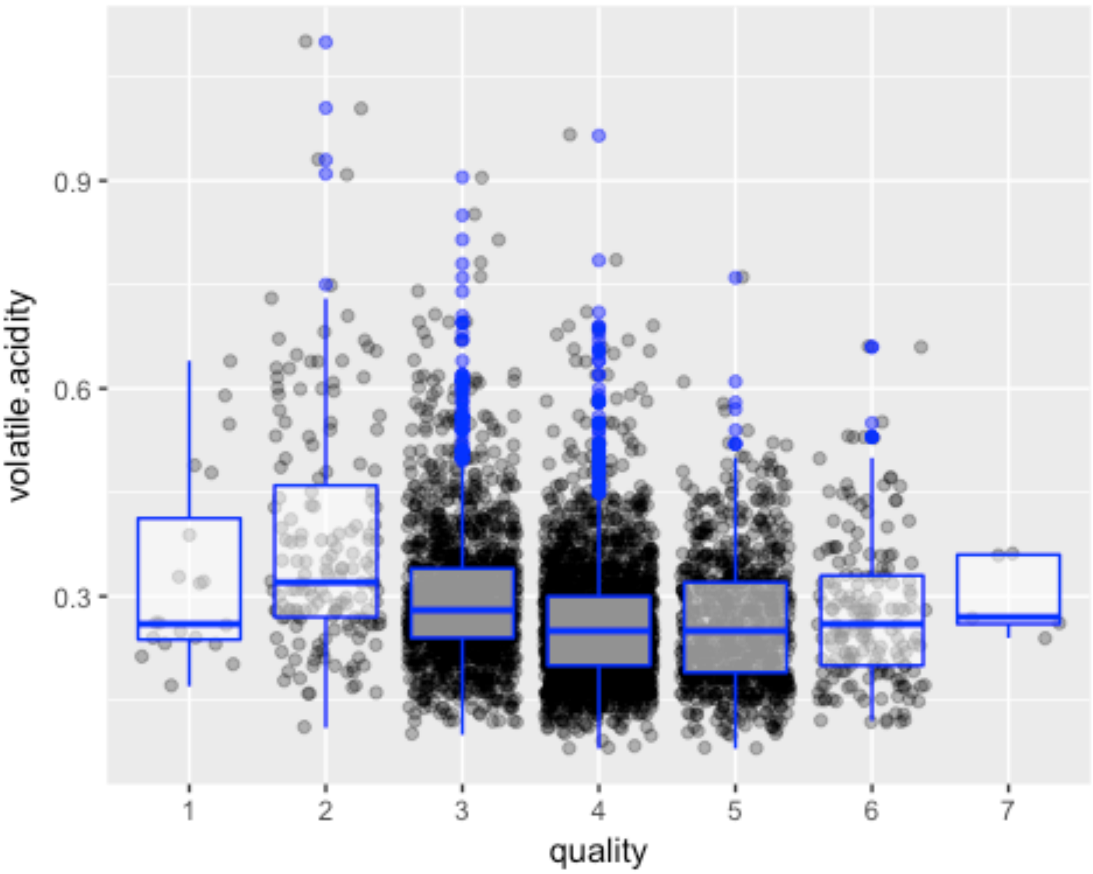
As we saw in the correlation table, alcohol and quality have a strong positive relationship. This is even more apparent when looking at the box plot. As alcohol content goes up, wine quality increases.



This visualization reinforces the negative relationship between density and alcohol. Wines with a higher quality have lower densities.



There is a weak correlation between residual sugar and quality due to low levels of sugar across all wines. Only a few wines have higher sugar content.



From the plot, as volatile acidity decreases, wine quality increases.

```
##
## Call:
## lm(formula = as.numeric(quality) ~ alcohol, data = ww)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5317 -0.5286  0.0012  0.4996  3.1579
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.582009   0.098008   5.938 3.08e-09 ***
## alcohol     0.313469   0.009258  33.858  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7973 on 4896 degrees of freedom
## Multiple R-squared:  0.1897, Adjusted R-squared:  0.1896
## F-statistic:  1146 on 1 and 4896 DF,  p-value: < 2.2e-16
```

Bivariate Analysis

Observations

The plots created in this section support my hypothesis that alcohol plays a role in wine quality. However, my hypothesis that residual sugar affects wine quality was proven wrong. Our findings reveal a strong positive relationship between alcohol and quality but a weak relationship between residual sugar and quality.

Quality correlates negatively with density and volatile acidity. Wine quality increases as density and volatile acidity decrease. Out of the two variables, density has a stronger correlation with quality.

Interesting Relationships

I find it interesting that alcohol is negatively correlated with all the variables we plotted (density, residual sugar, volatile acidity).

Strongest Relationship

Density and residual sugar formed the strongest positive relationship (.8), while density and alcohol had the strongest negative relationship (-.8).

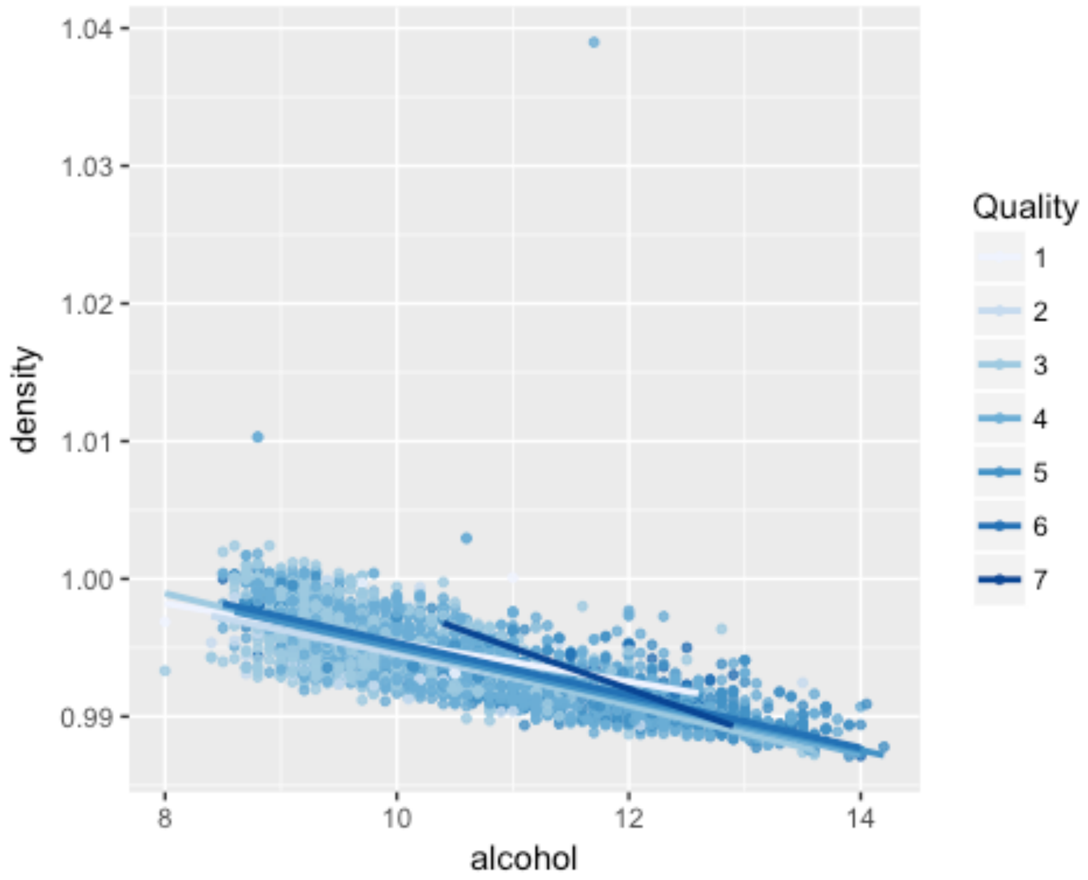
Multivariate Plots Section

As seen in the previous section, alcohol plays a significant role in wine quality. However alcohol alone can't be the only factor. This leads us to dig deeper to determine how much of an impact alcohol has on quality.

```
## 
## Call:
## lm(formula = as.numeric(quality) ~ alcohol, data = ww)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.5317 -0.5286  0.0012  0.4996  3.1579 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 0.582009   0.098008   5.938 3.08e-09 ***
## alcohol     0.313469   0.009258  33.858  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.7973 on 4896 degrees of freedom
## Multiple R-squared:  0.1897, Adjusted R-squared:  0.1896 
## F-statistic:  1146 on 1 and 4896 DF,  p-value: < 2.2e-16
```

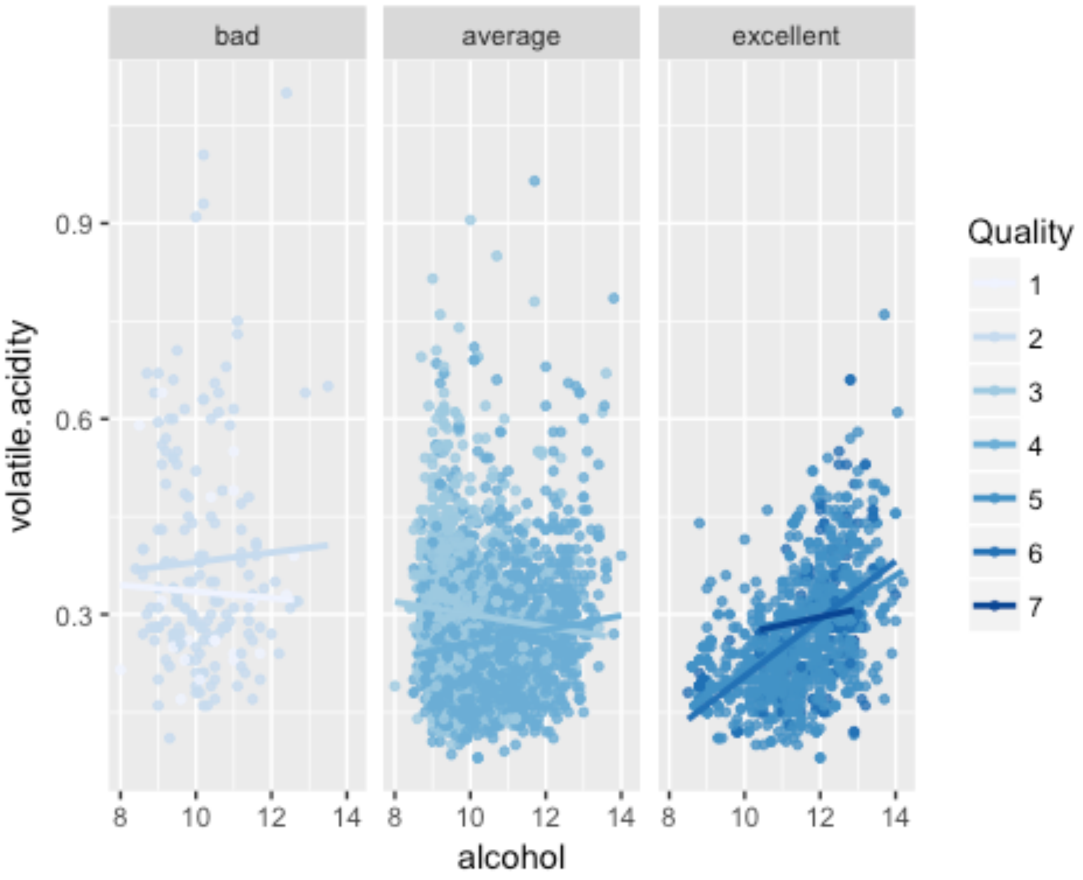
According to this model, alcohol contributes to 19% of the total factors affecting quality. We need to further investigate the relationships among variables to determine other factors.

```
ggplot(data = ww, aes(y = density, x = alcohol, color = quality)) +
  geom_point(alpha = 0.8, size = 1) +
  geom_smooth(method = "lm", se = FALSE,size=1)  +
  scale_color_brewer(type='seq', guide=guide_legend(title='Quality'))
```



Density doesn’t appear to significantly change the quality of alcohol. This means the negative correlation between density and quality is due to the presence of alcohol.

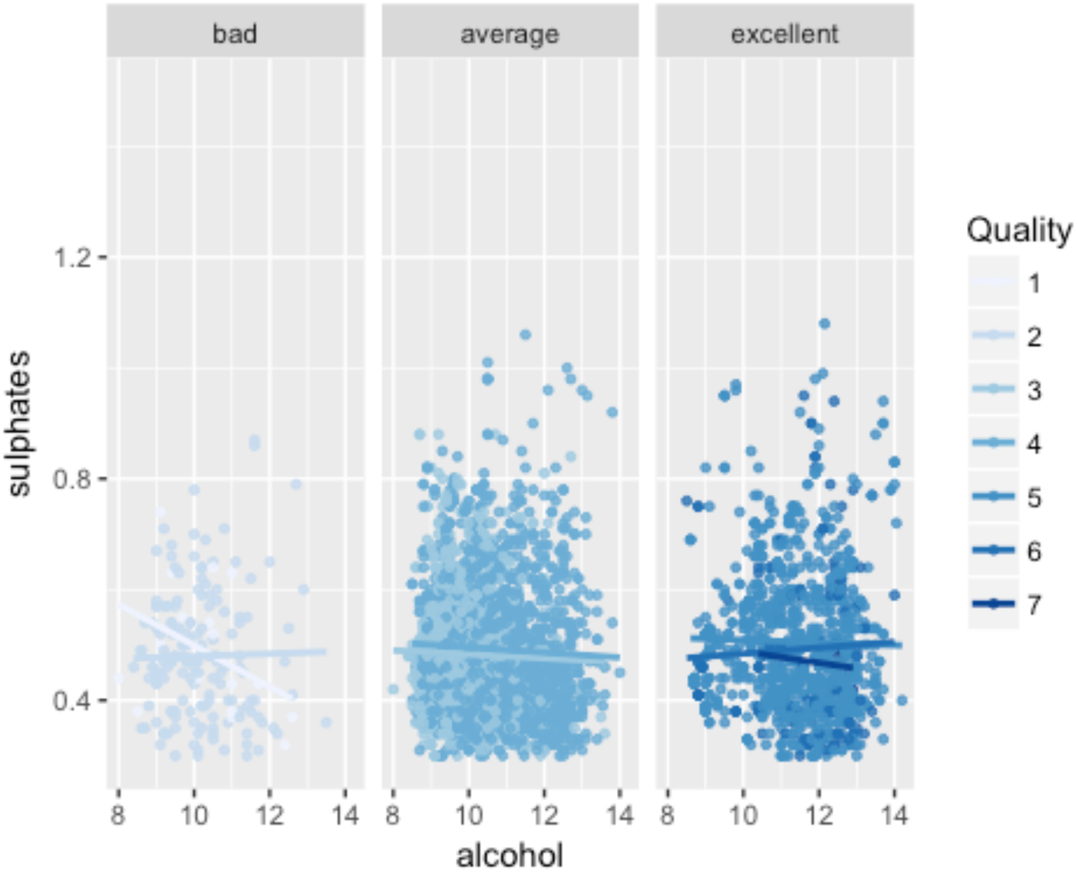
```
ggplot(data = ww, aes(x = alcohol, y = volatile.acidity, color = quality)) +
  geom_point(alpha = 0.8, size = 1) +
  geom_smooth(method = "lm", se = FALSE,size=1)  +
  facet_wrap(~rating) +
  scale_color_brewer(type='seq', guide=guide_legend(title='Quality'))
```



Low concentration of volatile acidity and high alcohol content produce better wines.

```
## Warning: Removed 52 rows containing non-finite values (stat_smooth).

## Warning: Removed 52 rows containing missing values (geom_point).
```



Referring back to the correlation table, we see that sulphates has the second most positive relationship with quality. After plotting sulphates against alcohol, it appears that higher sulphate concentration and alochol content produce better wines.

Multivariate Analysis

Observations

After looking at the relationships between individual variables with quality, I plotted some variables against alcohol along with quality/rating. This allowed me to determine whether or not these variables have an actual impact on quality. I chose three variables to plot against alcohol: density, volatile acidity, and sulphates.

Density doesn't appear to affect quality. Low volatile acidity and high concentration of sulphates combined with high alcohol content produce better wines.

Interesting Relationships

In the Bivariate plot section, we saw that density has a negative relationship with alcohol and quality. Plotting these two variables together allowed me to see the impact of density not just on quality but alcohol as well. This led to an interesting finding: density's correlation with quality is a result of alochol content.

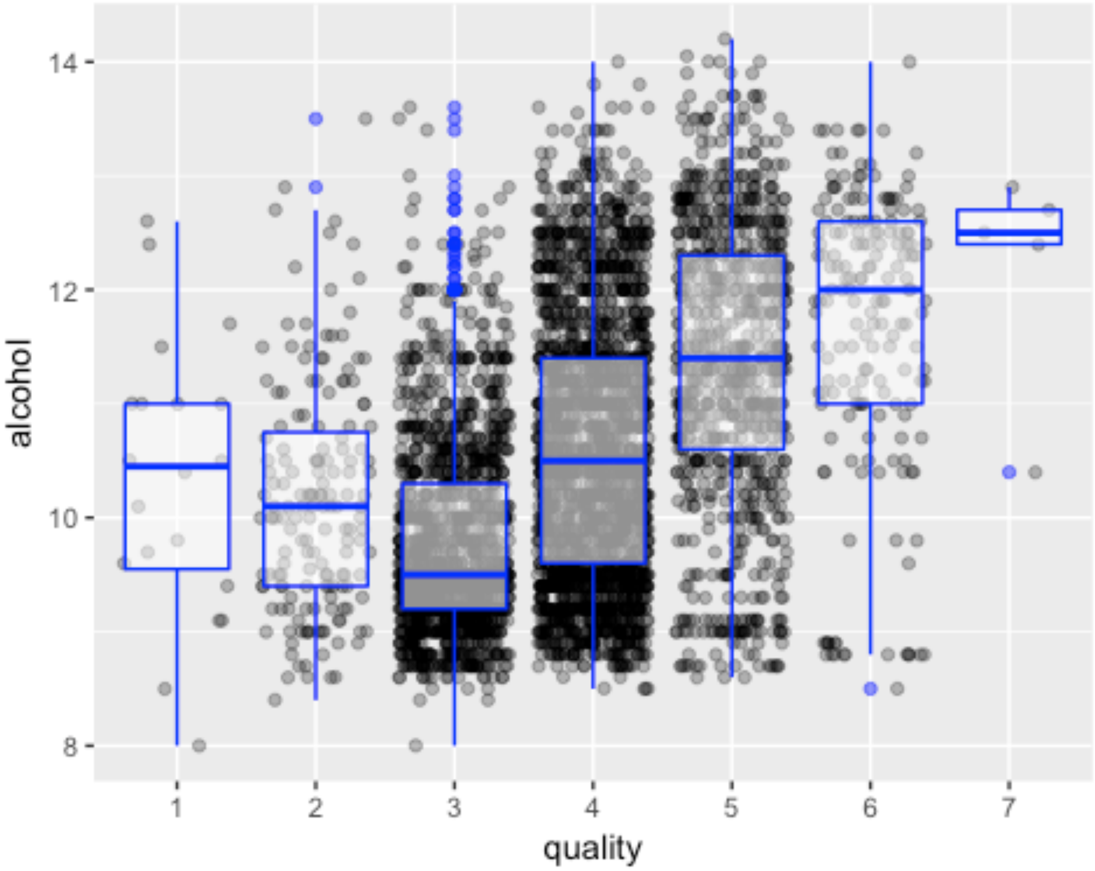
Model

I created a linear model to calculate alcohol's contribution to quality which is 19%.

Final Plots and Summary

In this section, I chose 3 plots to summarize my findings.

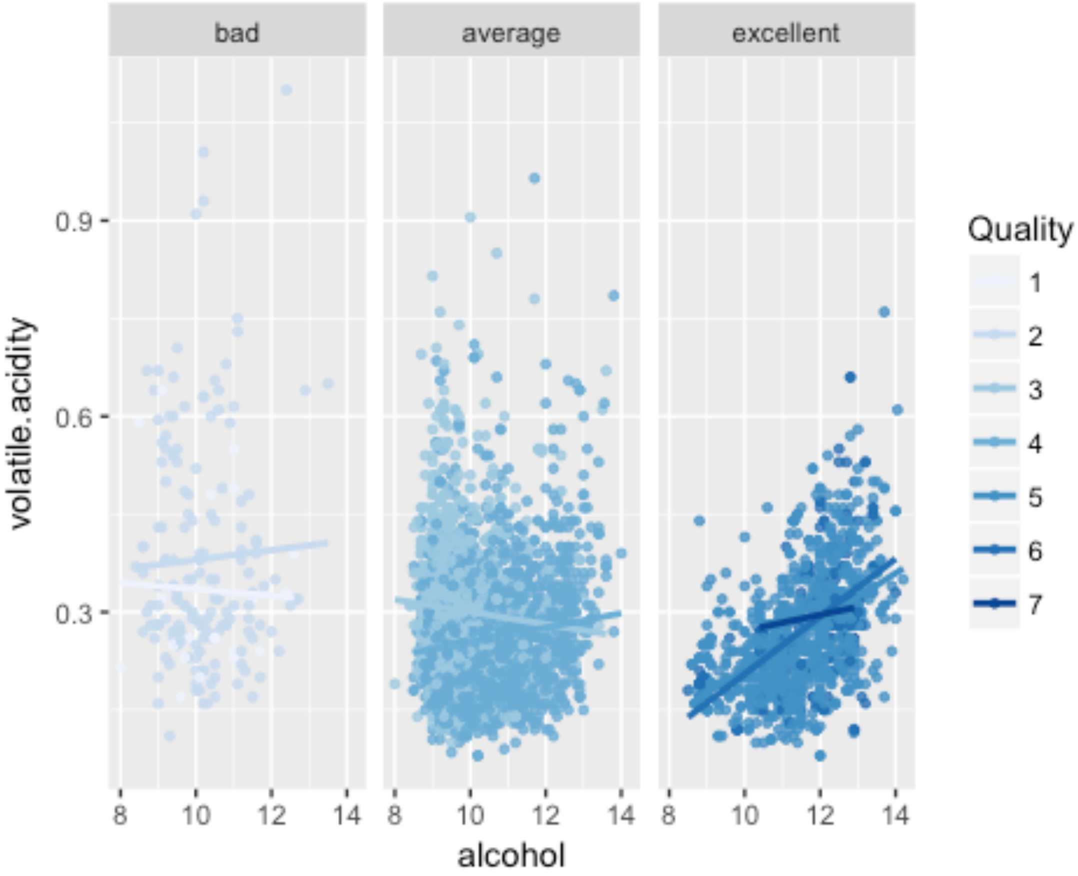
Plot One



Description One

Alcohol has played a significant role in determining wine quality. The correlation table and this plot reflect the strong positive relationship between alcohol and quality. As alcohol goes up, quality increases.

Plot Two



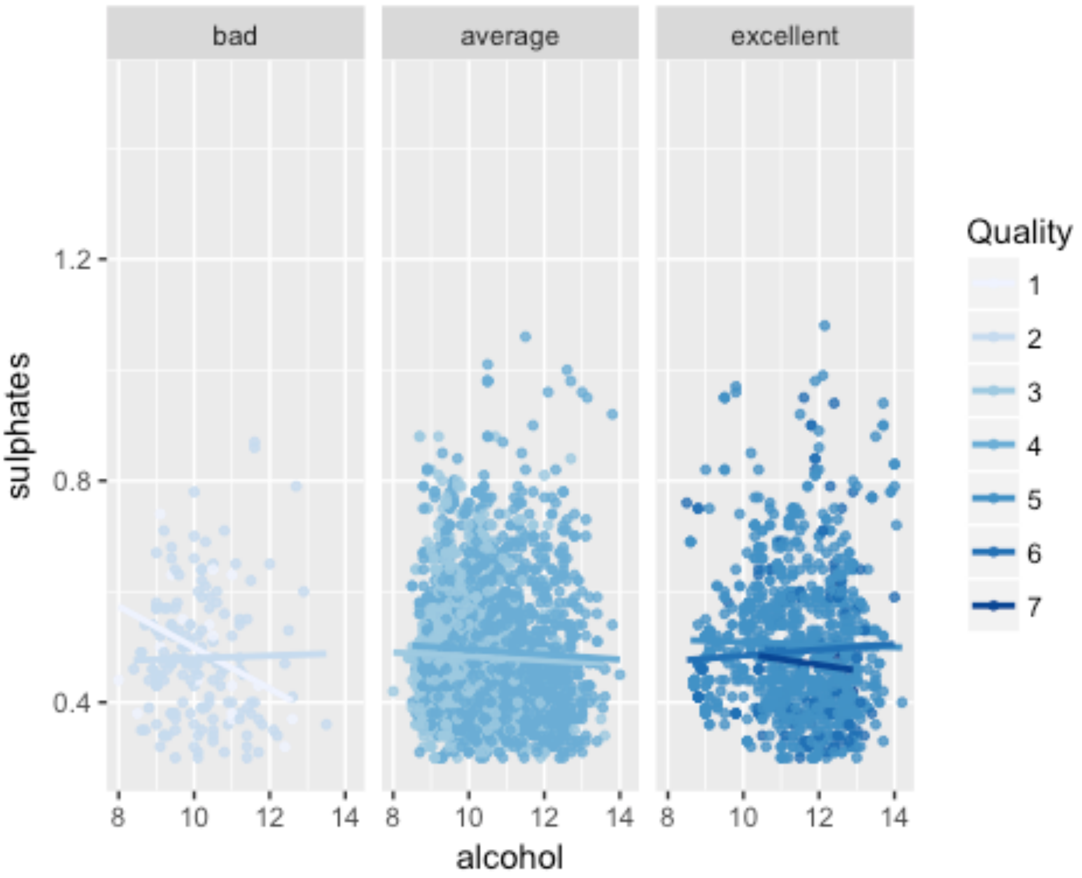
Description Two

Volatile acidity may not have as strong of a relationship to quality as alcohol, but it is a factor of quality. A high concentration of alcohol and low concentration of volatile acicity produce higher quality wines.

Plot Three

```
## Warning: Removed 52 rows containing non-finite values (stat_smooth).

## Warning: Removed 52 rows containing missing values (geom_point).
```



Description Three

Given the fact that density and residual sugar do not play a major role in quality, I decided to refer back to the correlation table to review correlation of other variables with alcohol and quality. Sulphates stood out to me and the visualization tells us that a higher concentration of sulphates and alcohol produces better wines.

Reflection

To recap, the objective of this project was to analyze white wine features and their relationships to each other to determine which factors affect our dependent variable, quality. The approach to investigating the dataset involved using exploratory data analysis. Initially we looked at the variables individually, then we started digging deeper and began looking for insights about relationships among the variables.

This led to some interesting findings:

• Alcohol has the strongest relationship to quality. As alcohol increases so does quality. • Variables that make wines taste better include high concentration of alcohol and sulphates and low volatile acidity.

There are limitations to this analysis due to the sampling of only Portuguese wines. The dataset does not include features such as grapes used or wine age. For future analysis, I'd explore a dataset that includes these features. I'd also explore wines from other countries to see if factors affecting quality vary from country to country.