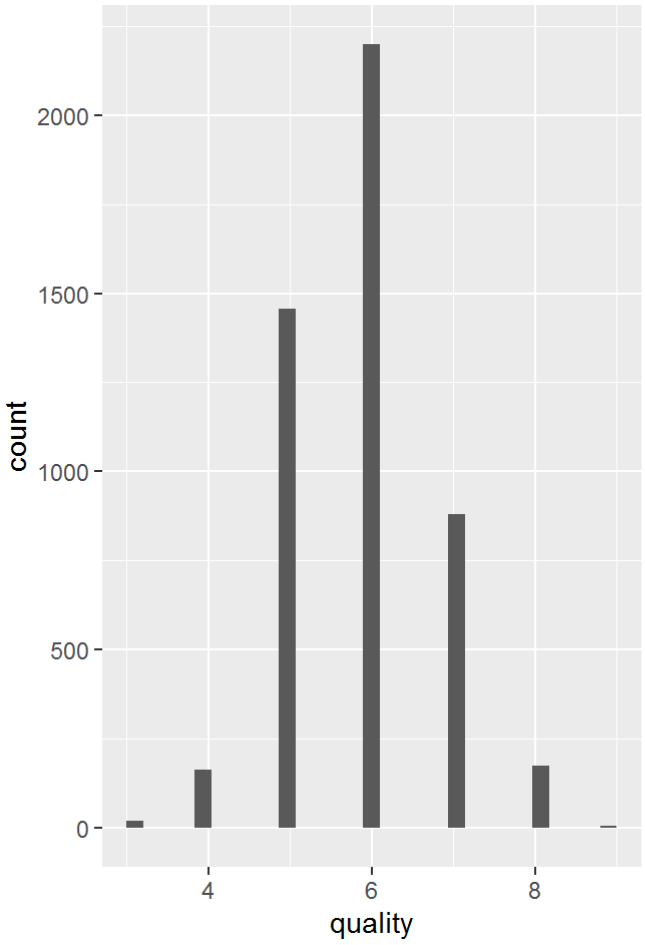
White Wine Quality Analysis by Jeffrey Solis

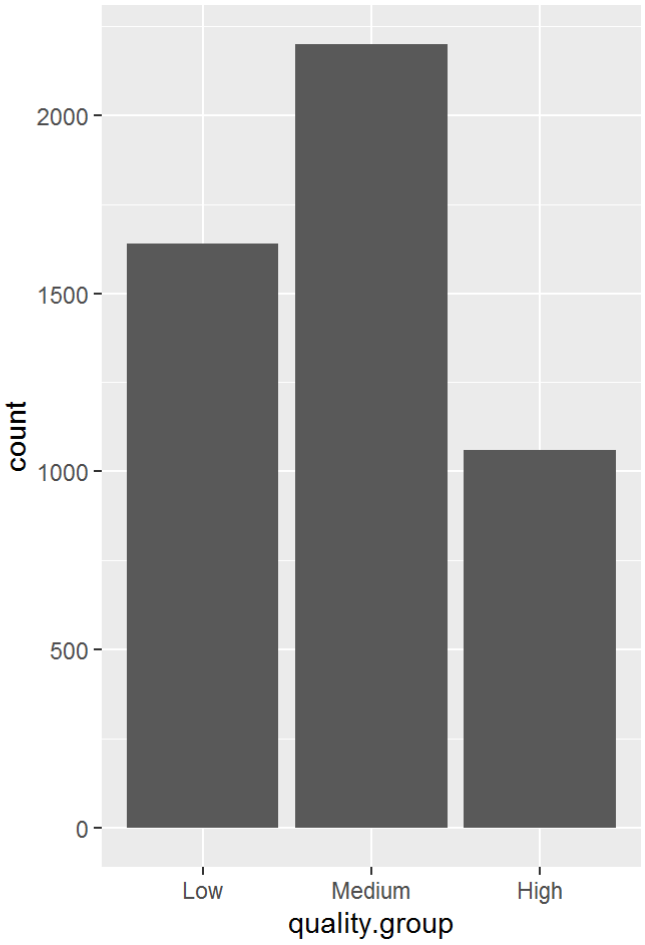
Univariate Plots Section

```
## 'data.frame':    4898 obs. of  12 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             : int  6 6 6 6 6 6 6 6 6 6 ...
```

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

Our dataset contains 12 variables and 4898 observations.



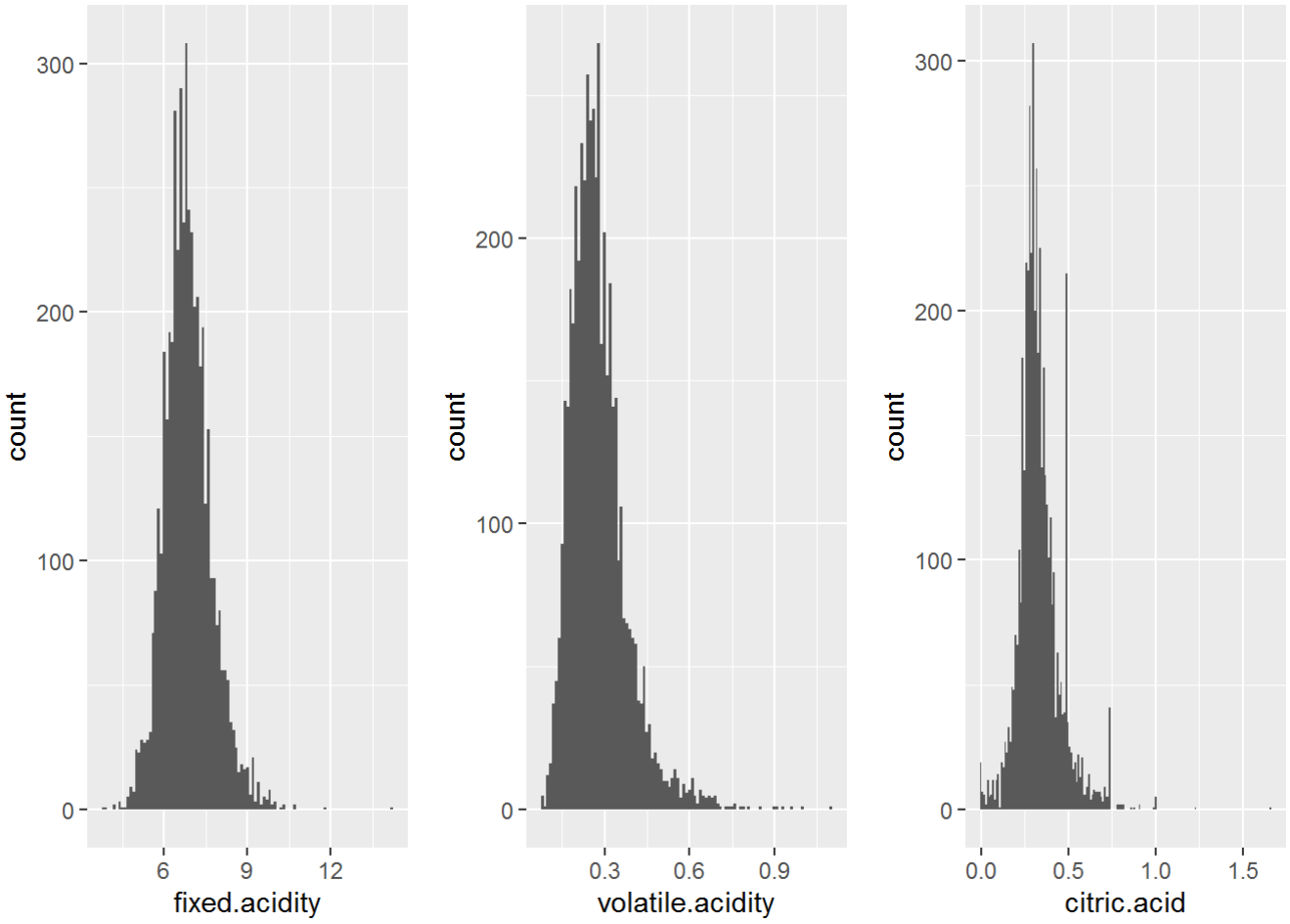


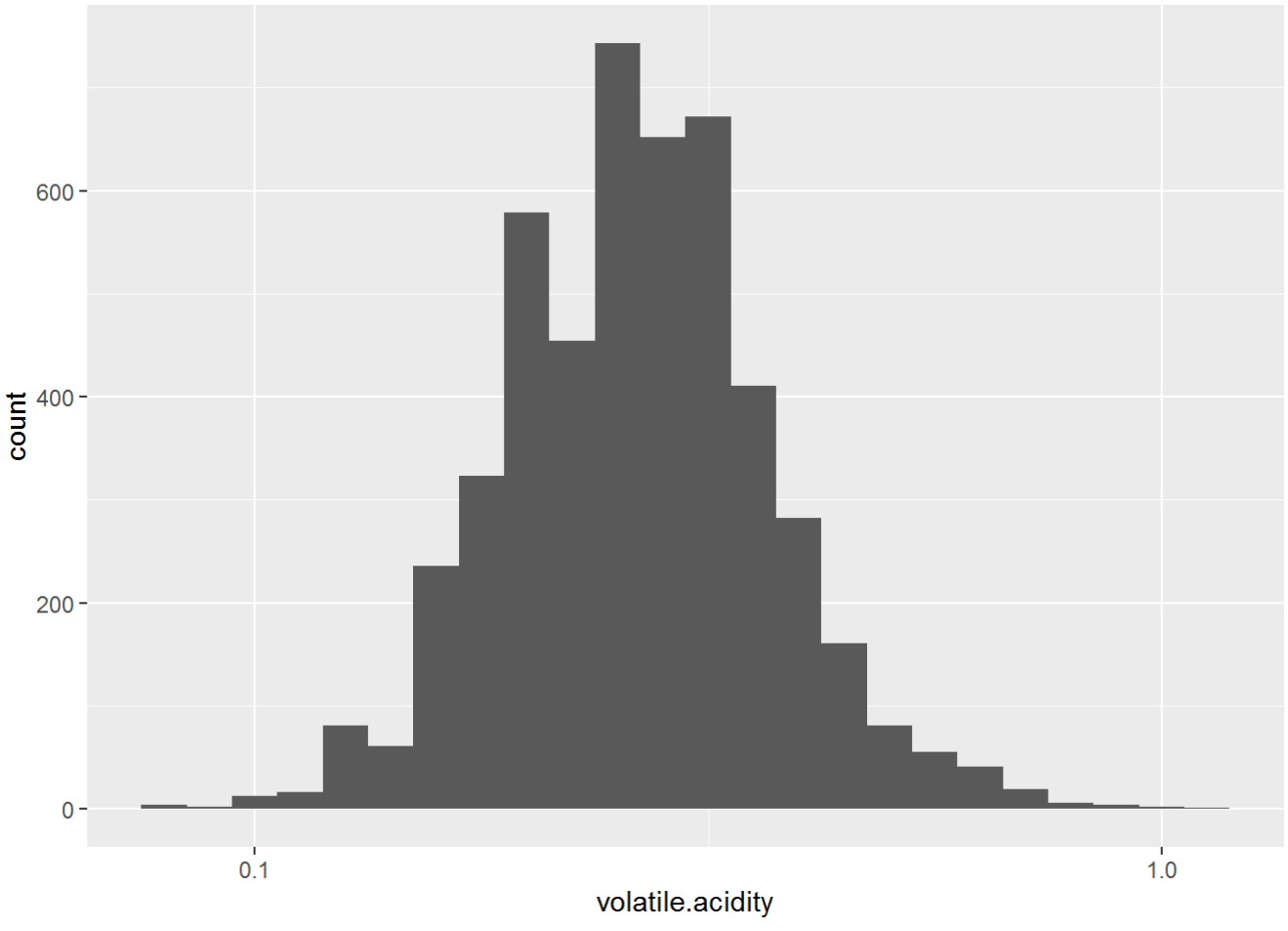

```
##    3    4    5    6    7    8    9
##   20  163 1457 2198  880  175    5
```

```
##    Low Medium   High
##   1640   2198   1060
```

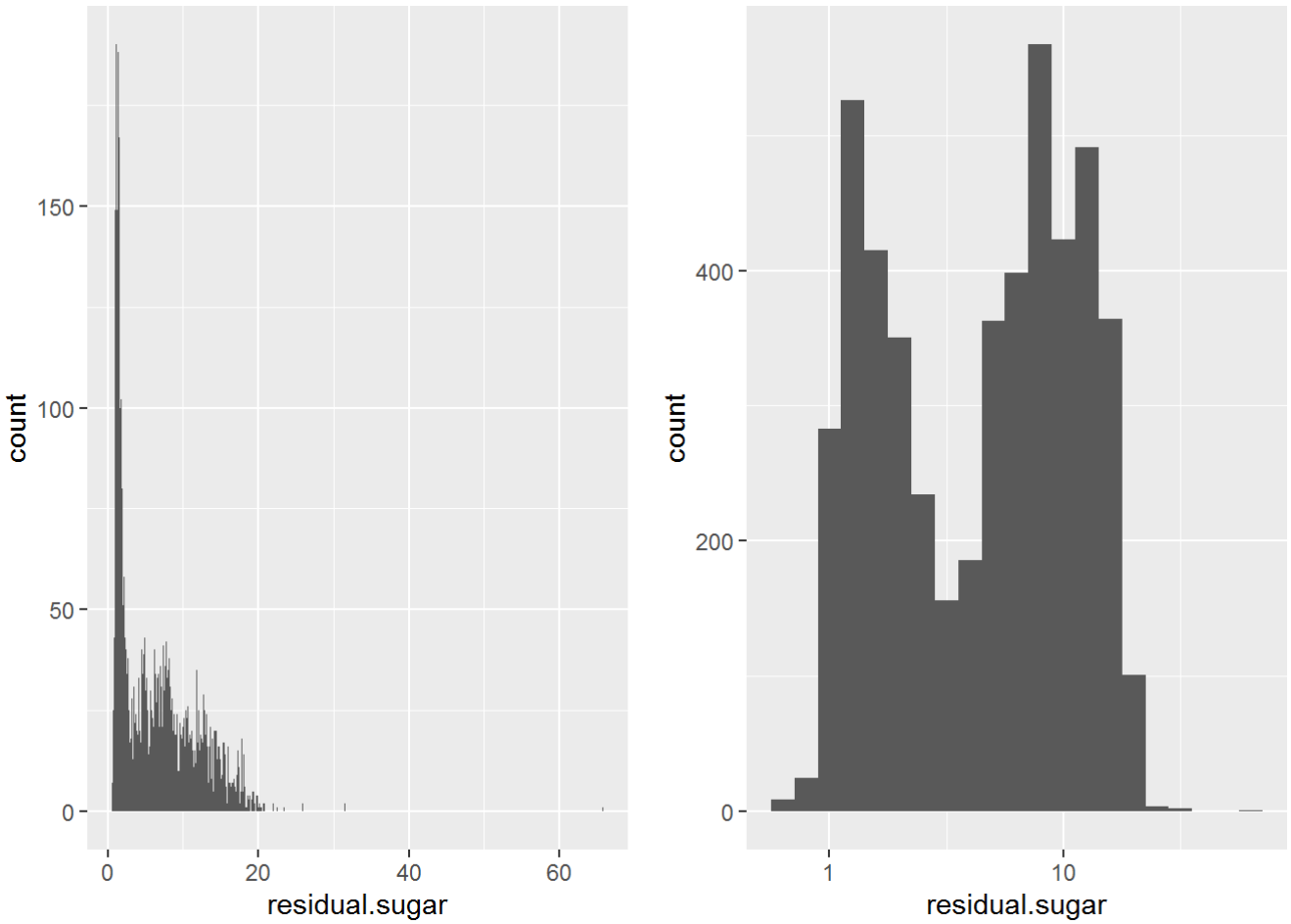
Most wines in this data had a quality score between 5-7. I decided to categorize this value into 3 groups:

- Low - Quality scores between 1-5
- Medium - Quality score of 6
- High - Quality scores between 7-10

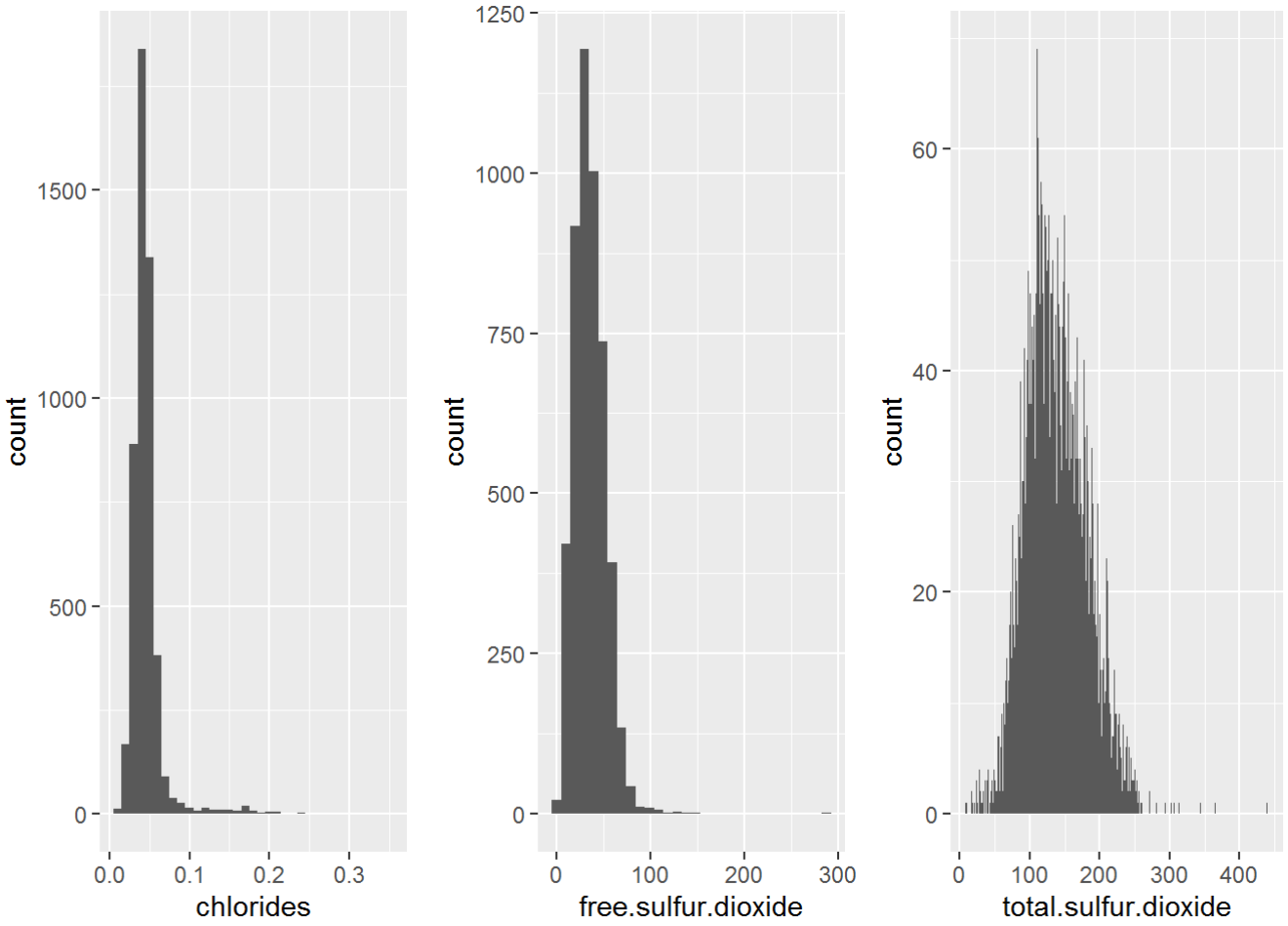

count
fixed.acidity
volatile.acidity
citric.acid


count
volatile.acidity

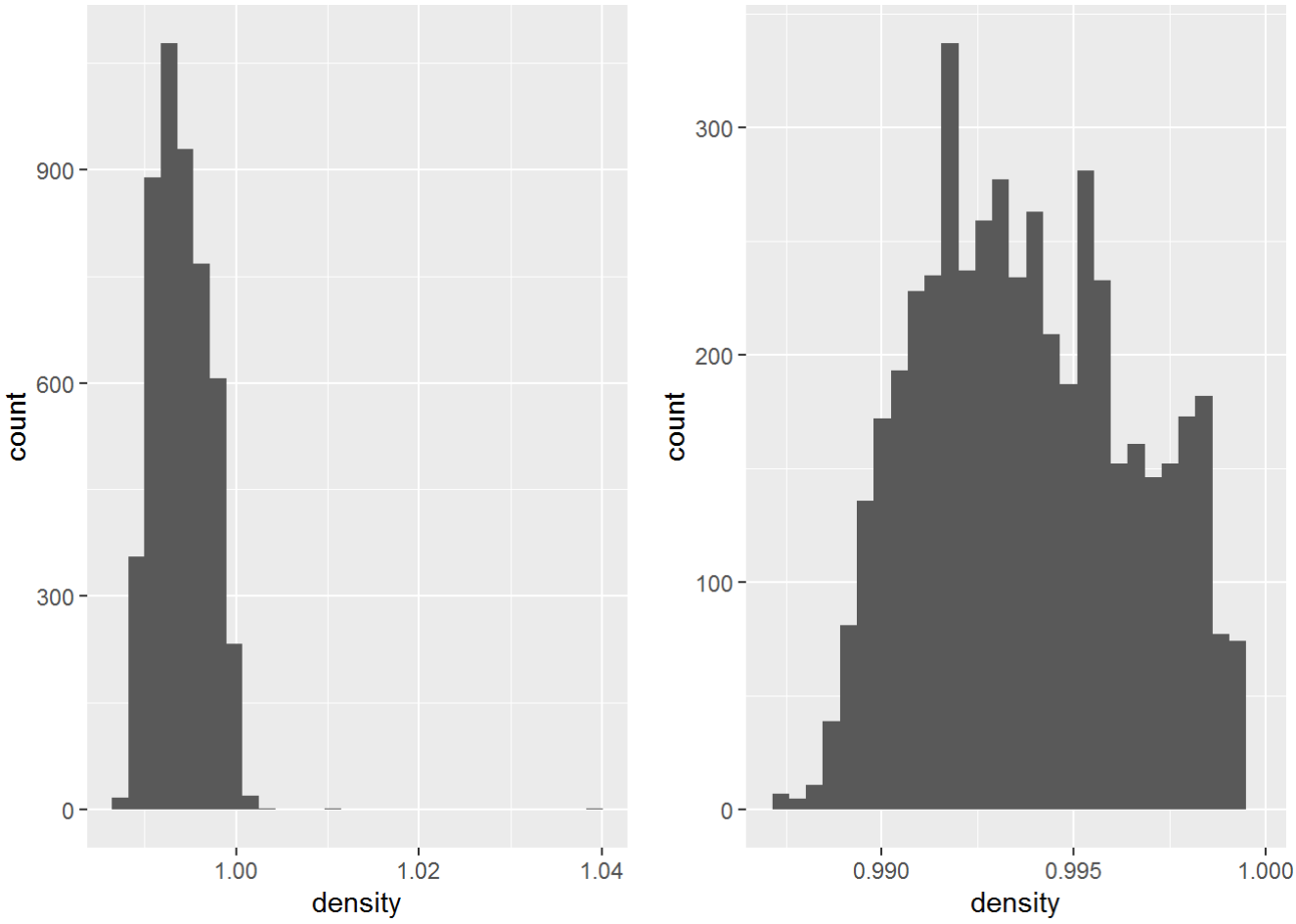
Volatile acidity is skewed right. A log10 transformation was performed on volatile acidity, and the distribution is now normal. Both fixed acidity and citric acid are normally distributed. Citric acid has some interesting peaks around 0.5 and 0.75.



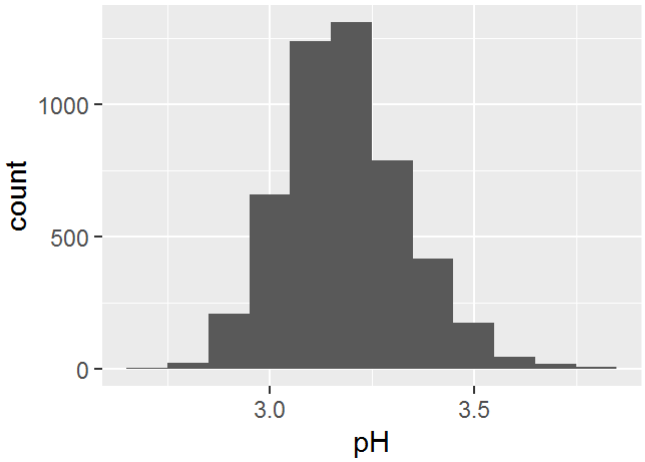
Residual sugar is also skewed right. After performing a log10 transformation, the distribution is bimodal. It looks like the peaks are around 3 and 10 g/dm^3. It is possible that most wines in this dataset are dry and off-dry wines. Also, there is a noticeable outlier around 65 g/dm^3.

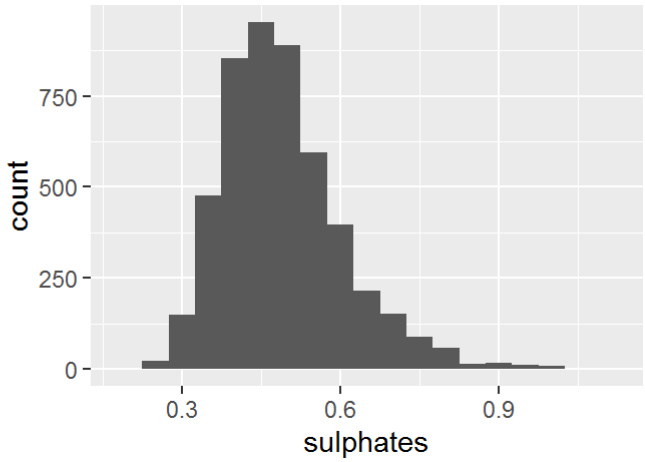


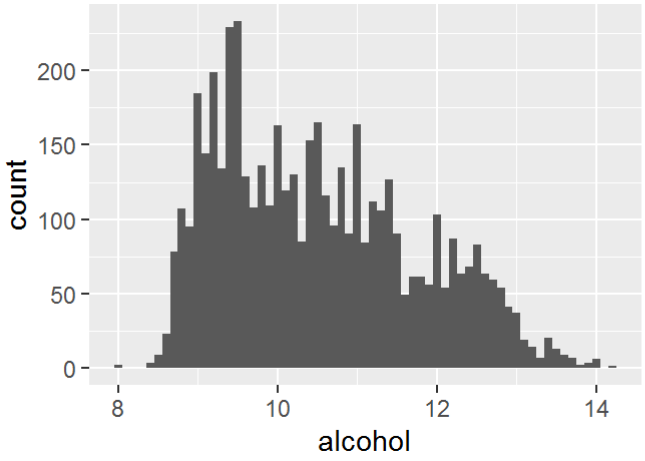
Chlorides and free sulfur dioxide is skewed right with most values at 0.43 g/dm^3 and 34 mg/dm^3 respectively. Total sulfur dioxide is normally distributed, with most values at 134 mg/dm^3.



Density is skewed right with values at 0.9937 g/cm^3. Also, there is an outlier at 1.039. Limiting the x axis to include the 95th % quantile, the density now looks normally distributed.







pH is normally distributed with most values at 3.2. Alcohol and sulphates are both skewed right. Also, alcohol looks biomodal with a strong peak around 9.3 and a small peak around 12.6.

Univariate Analysis

What is the structure of your dataset?

There are 4898 observations and 13 variables. Most variables are numerical values. Since quality is subjective, I thought it was best to cast this variable as a factor.

Most wines in this data set were assigned a quality between 5-7.

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

The main feature of this data set is quality. I want to see which chemical property influences the quality of wine.

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

I think volatile acidity is an interesting property because at high levels, volatile acidity may leave a vinegar taste that may influence quality. Residual sugar and alcohol can also influence quality because wine drinkers will have a preference on how strong or sweet their wines should be.

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

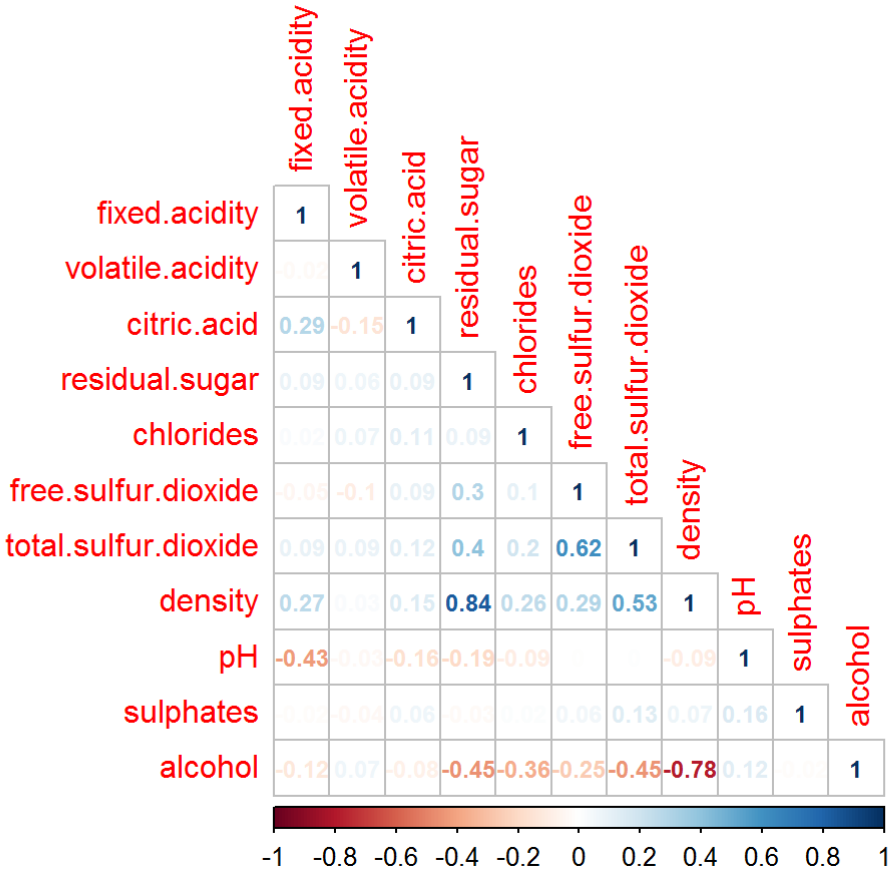
Instead of looking at quality according to a score, I decided to group the values into 3 categories. 1-5 are poor quality, 6 is normal quality, and 7-10 are high quality.

Other possible variables that can be created are total acidity (sum of fixed.acidity, volatile.acidity, and citric.acid) and ratio of free.sulfur.dioxide (free.sulfur.dioxide/total.sulfur.dioxed).

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?

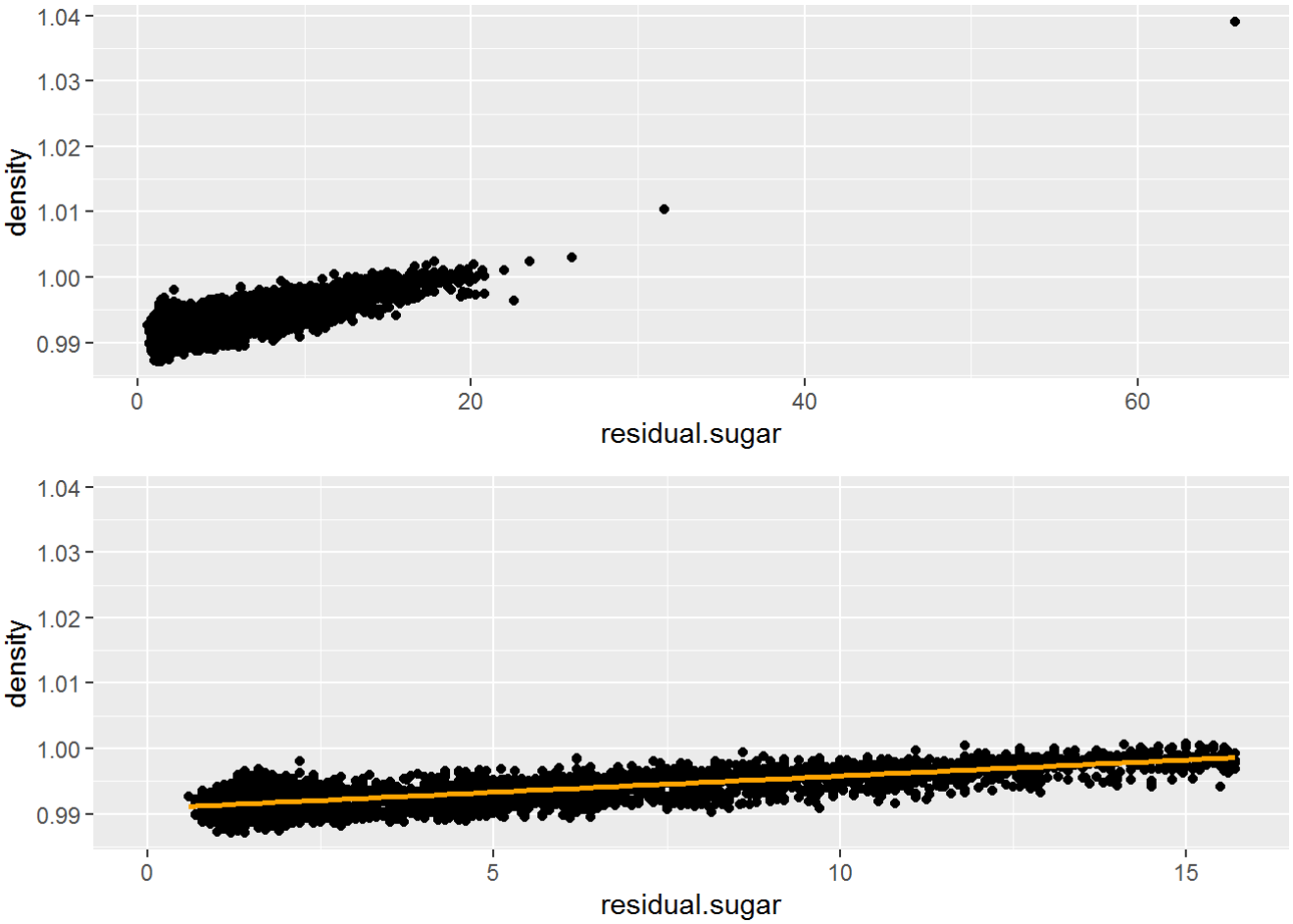
Volatile Acidity, residual sugar, and something were all skewed right. To get a better sense of the data, a log10 transformation was performed to getting a better understanding of the distribution.

Bivariate Plots Section

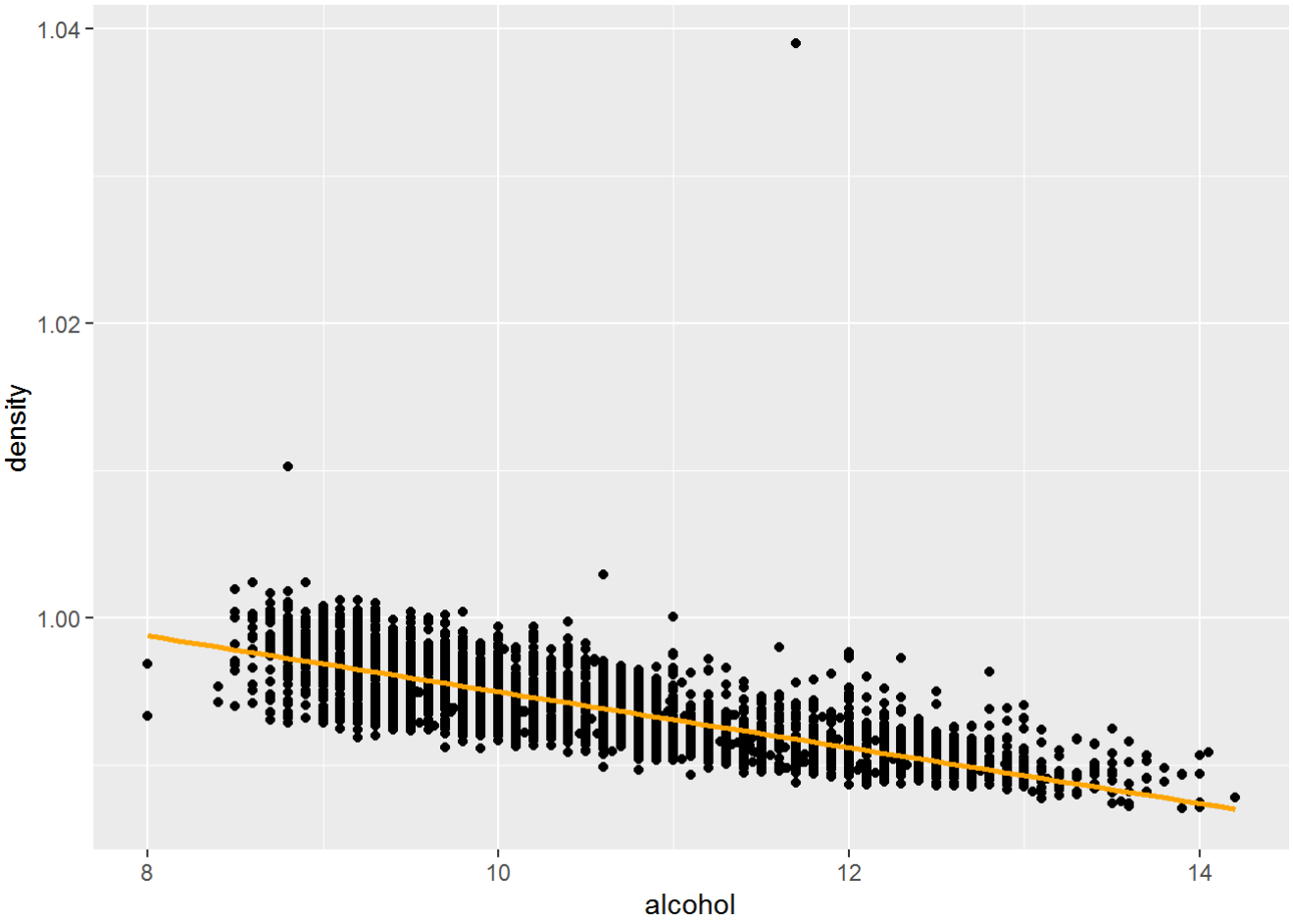


The correlation plot above shows the relationship between each chemical property in the dataset. The highest correlation were between residual sugar/density and density/alcohol. Other interesting correlations are between alcohol/total sulfur dioxide and alcohol/residual sugar.

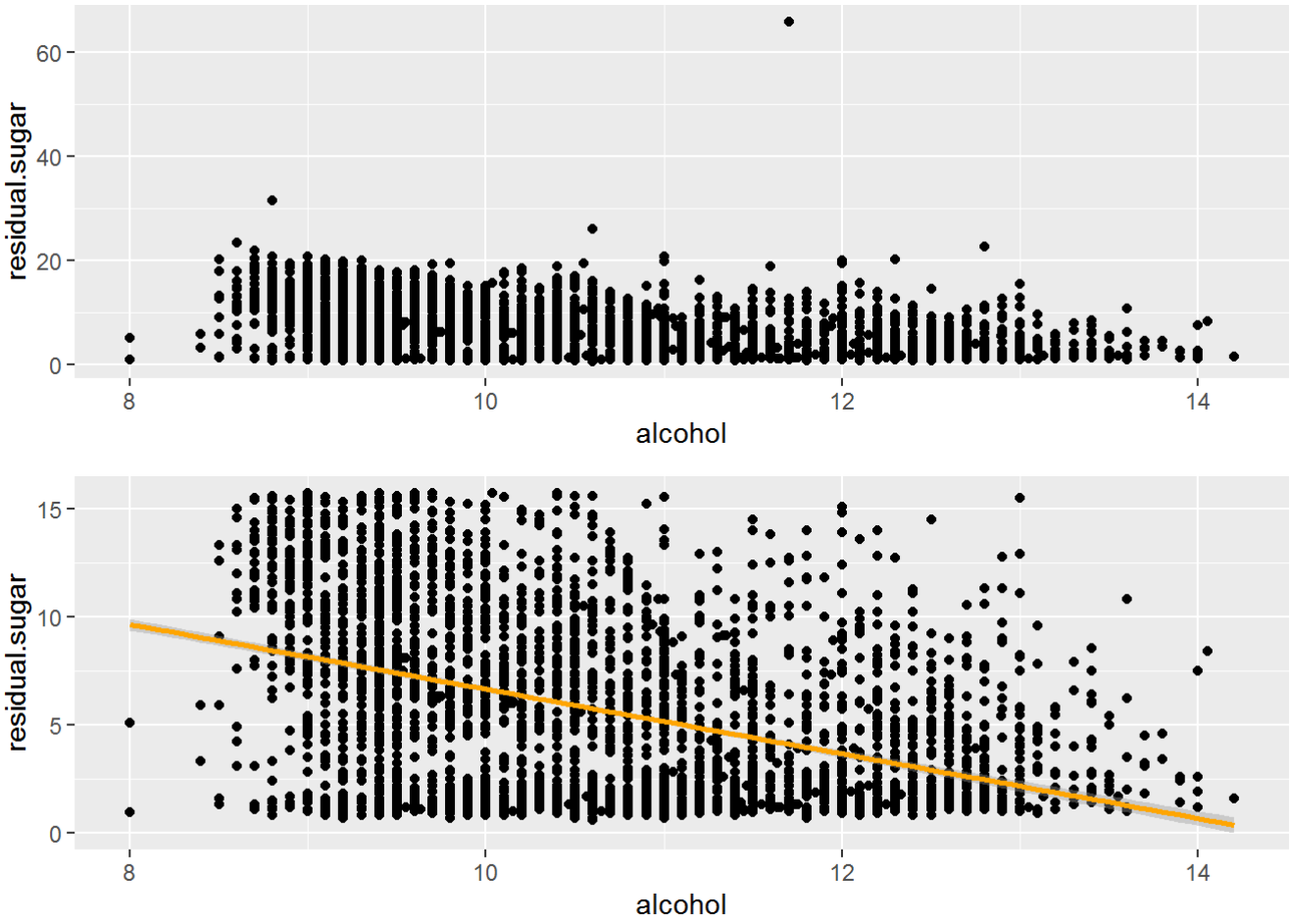
One surprising correlation is between fixed acidity and pH. Since pH is used to measure how acidic or basic an aqueous solution is, one would assume the correlation to be higher.



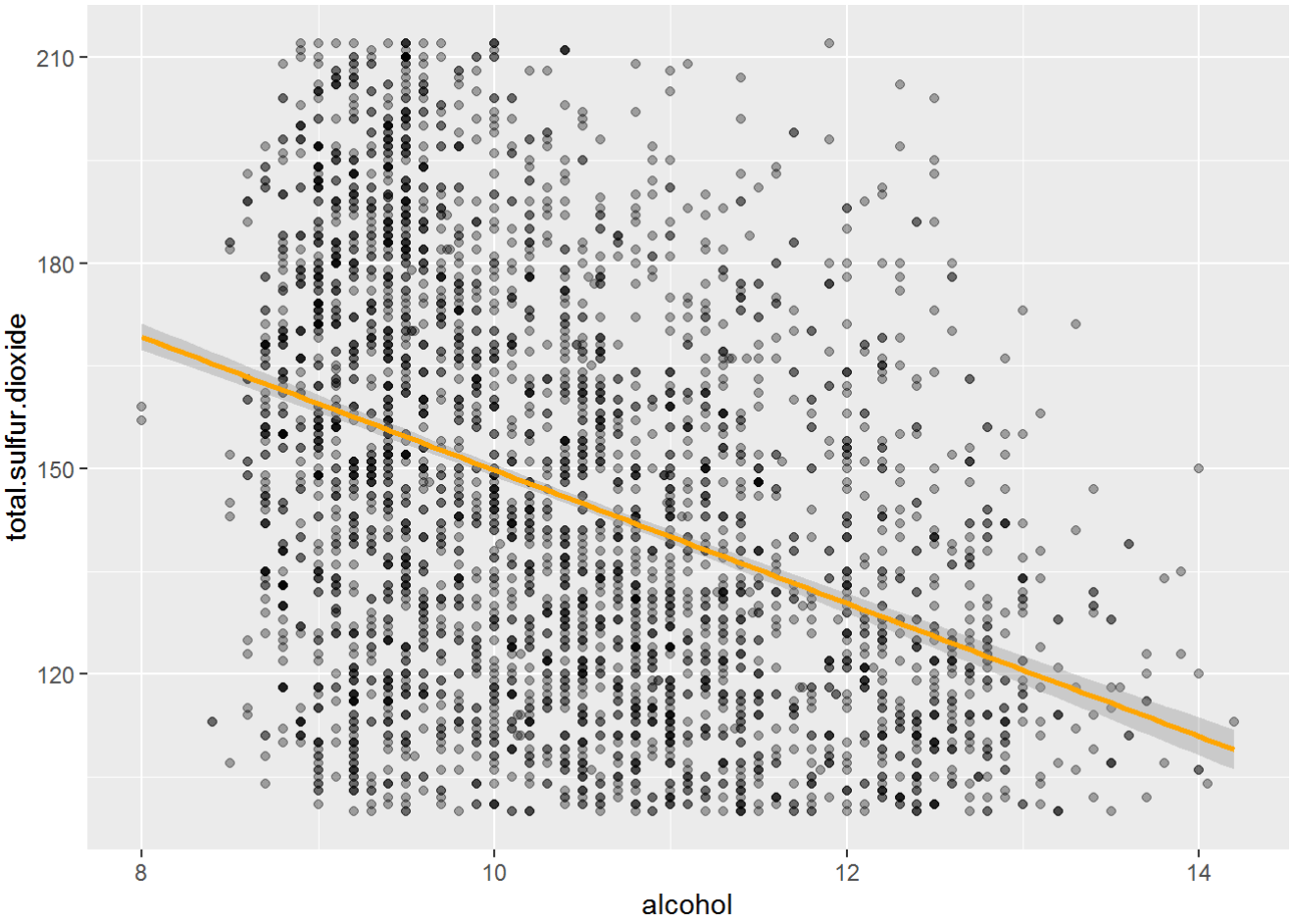
It looks like the data contains a couple outliers. I limited the y axis where 0 is the lower bound and the 95 % quantile is the upper bound, and applied a log10 transformation to residual sugar. It looks like a positive linear relationship between density and residual sugar. The higher the sugar content in the wine, higher the density. This explains the high correlation between the two variables.



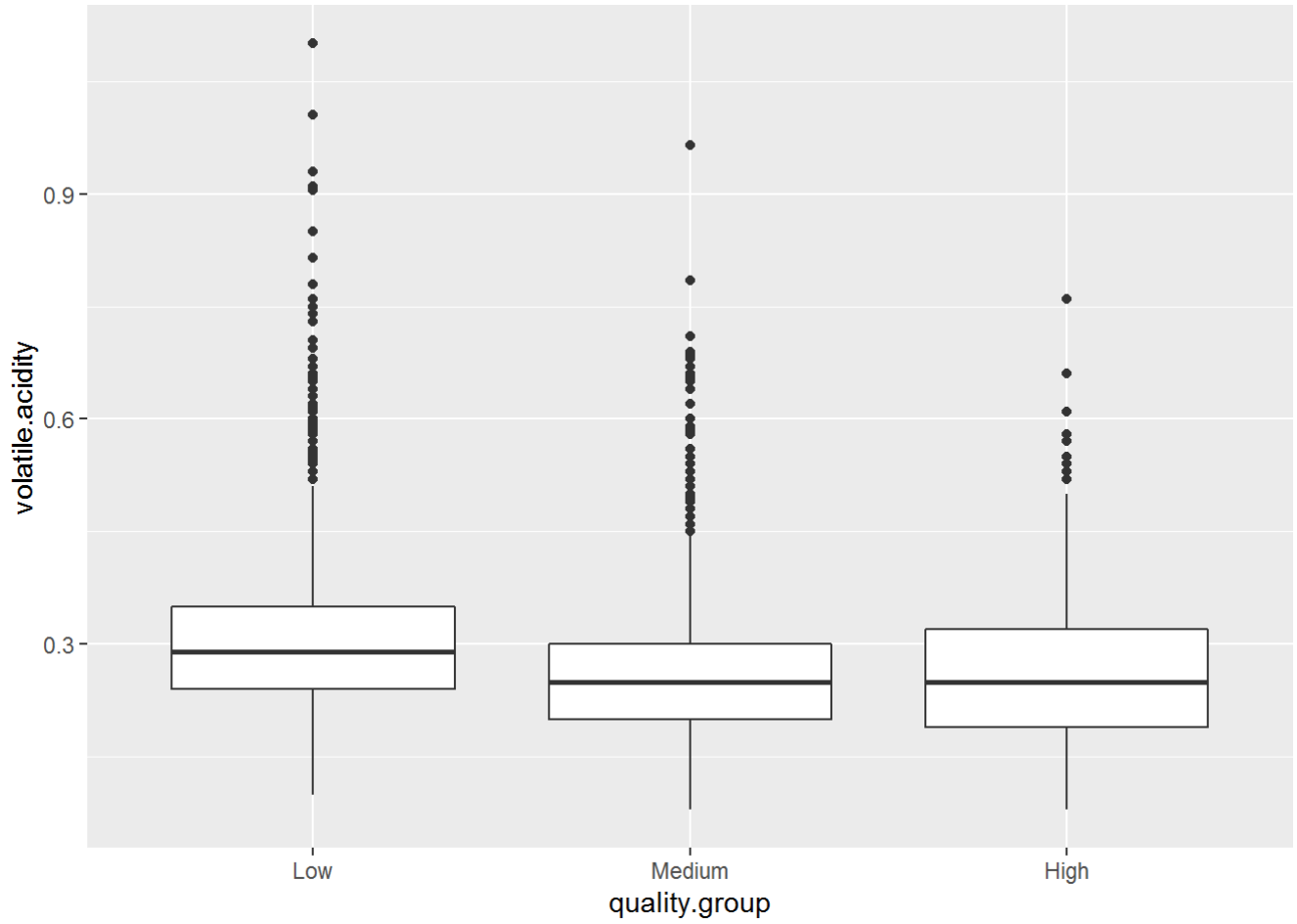
This plot shows that as alcohol levels increase the density will decrease. This supports the relationship between residual sugar and denisty and how the fermentation process affects the density of the wine.



According the the correlation matrix above, residual sugar and alcohol have a -0.45 correlation coefficient. I expected to see a stronger negative trend in the scatter plot since alcohol is a by-product of yeast when its used during fermentation.

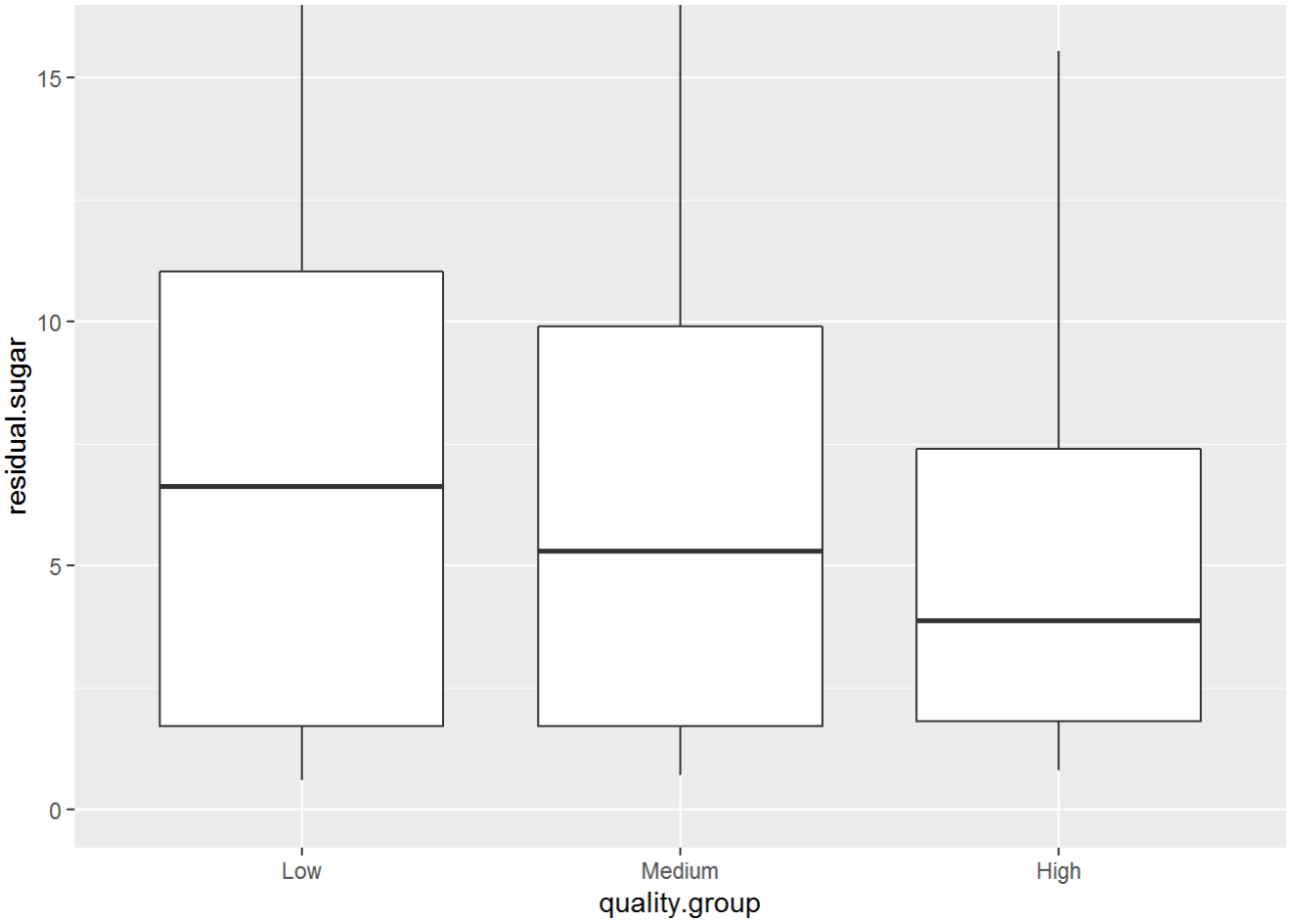


Alcohol and total sulfur dioxide have a negative relationship. Though the correlation coefficient is similar to that of alcohol and residual sugar, I do not understand the relationsip between the two variables.



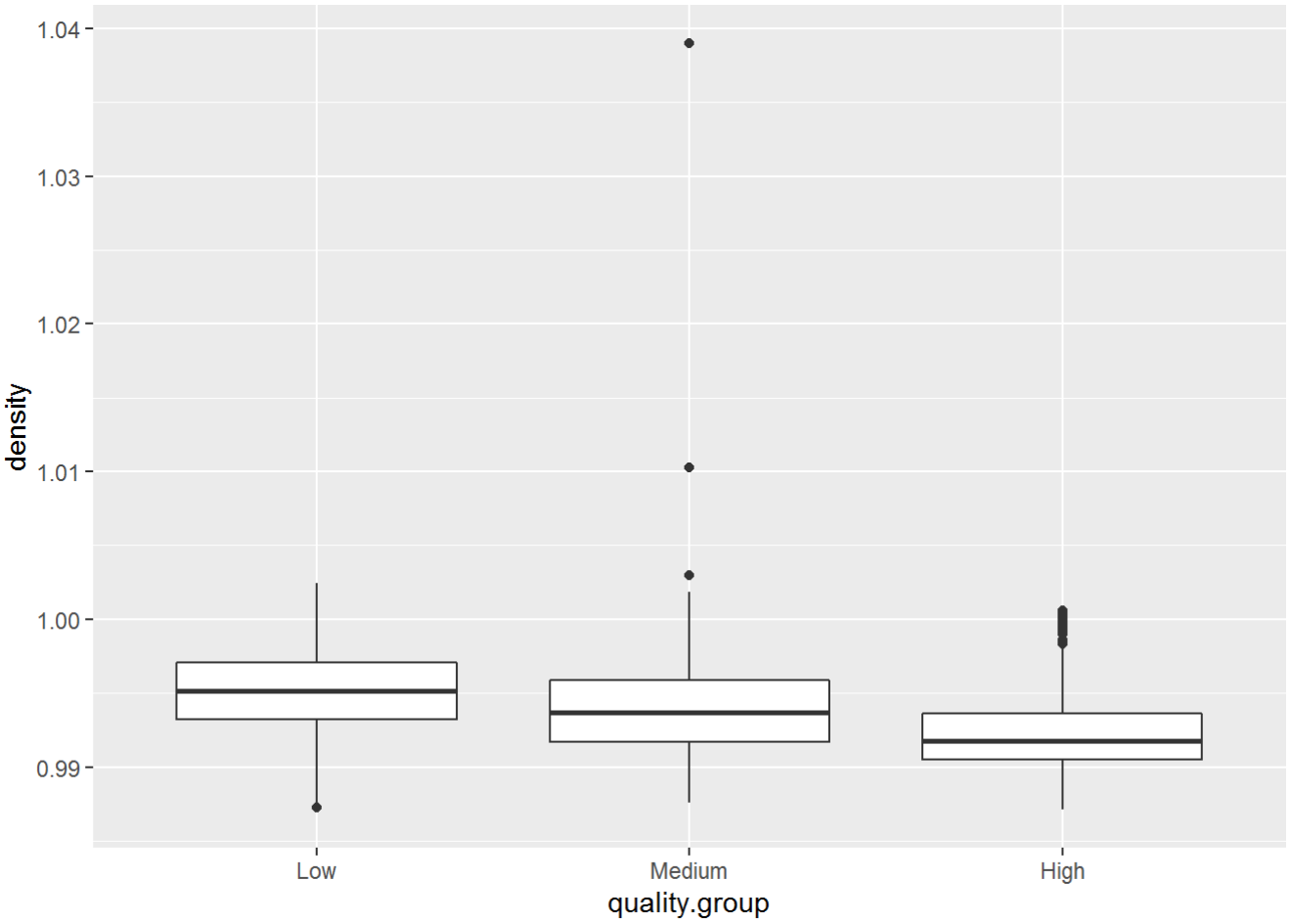

```
## wine$quality.group: Low
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1000  0.2400  0.2900  0.3103  0.3500  1.1000
## ------------------------------------------------------------
## wine$quality.group: Medium
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.2000  0.2500  0.2606  0.3000  0.9650
## ------------------------------------------------------------
## wine$quality.group: High
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.1900  0.2500  0.2653  0.3200  0.7600
```

According to the box plots, low quality wines have slightly higher volatile acidity compared to medium and high quality wines. At high levels, volatile acidity may produce a vinegar like taste, and I thought this could affect the quality of the wine. However, this plot shows that most wines in this dataset have low levels of volatile acidity.



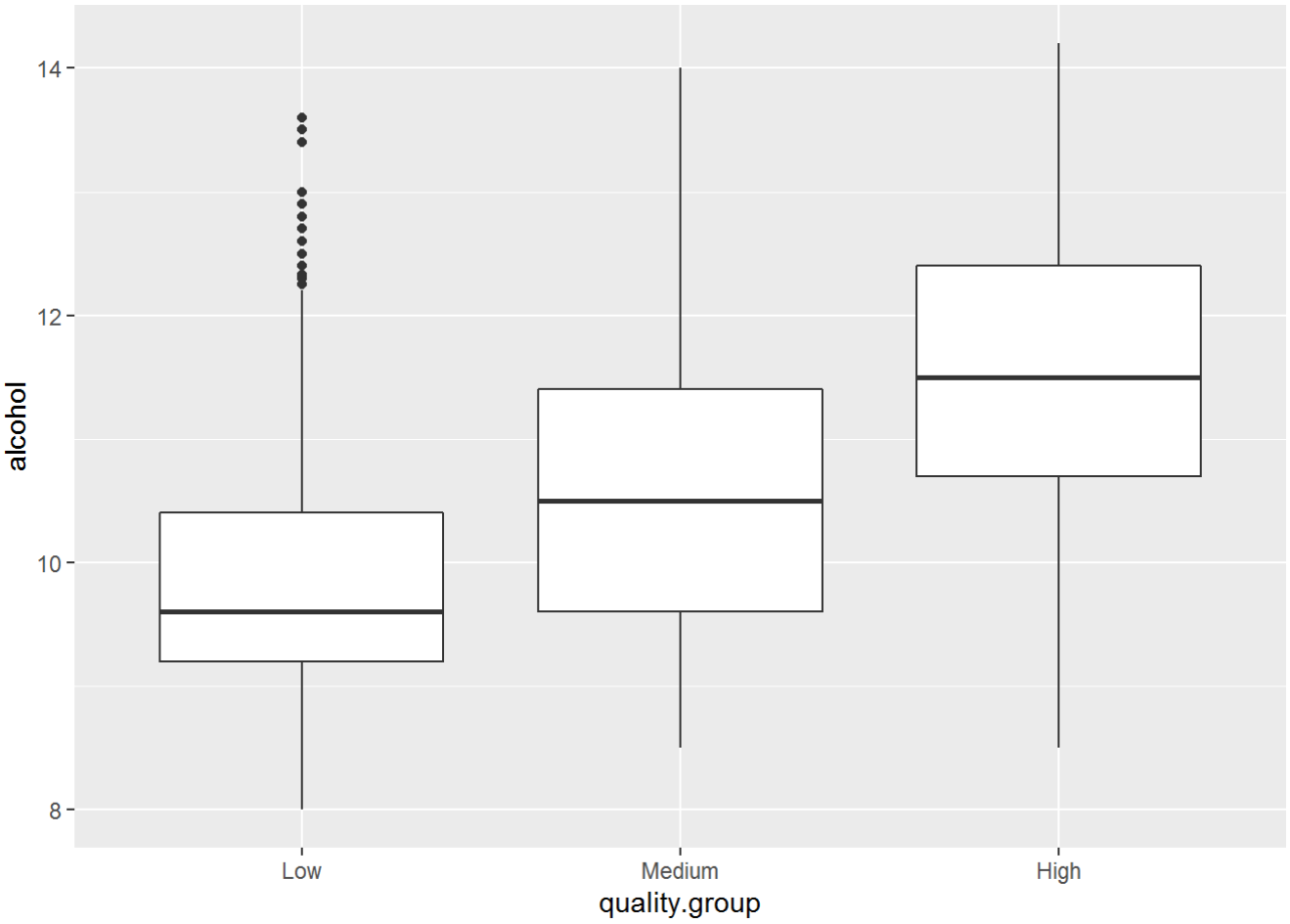

```
## wine$quality.group: Low
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.600   1.700   6.625   7.054  11.020  23.500
## ------------------------------------------------------------
## wine$quality.group: Medium
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.700   1.700   5.300   6.442   9.900  65.800
## ------------------------------------------------------------
## wine$quality.group: High
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.800   1.800   3.875   5.262   7.400  19.250
```

As you go up a quality group, the median residual sugar value decreases. I find it interesting that the distribution of residual sugar becomes closer as you go up a quality group.




```
## wine$quality.group: Low
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9872  0.9932  0.9951  0.9952  0.9971  1.0020
## ------------------------------------------------------------
## wine$quality.group: Medium
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9876  0.9917  0.9937  0.9940  0.9959  1.0390
## ------------------------------------------------------------
## wine$quality.group: High
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9905  0.9917  0.9924  0.9936  1.0010
```

As the quality of wine improves, the median density of wine decreases by 0.002 g/cm^3. This change in density does not look large and could go unnoticed by a wine drinker.




```
## wine$quality.group: Low
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.20    9.60    9.85   10.40   13.60
## ------------------------------------------------------------
## wine$quality.group: Medium
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50    9.60   10.50   10.58   11.40   14.00
## ------------------------------------------------------------
## wine$quality.group: High
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50   10.70   11.50   11.42   12.40   14.20
```

The median alcohol % by volume is higher by 1% every quality group. What I find most interesting about this chart is the 1st quartile range is higher in comparison to the other groups.

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 first relationship I was curious about was quality and volatile acidity. If there is a high level of volatile acidity, there is a possibility of the wine smelling like vinegar. However, I was surprised to see that quality of wine did not have a relationship with that chemical.

Another relationship I was curious about was quality and total sulfur dioxide. Sulfur dioxide prevents microbial growth and oxidation. The higher the quality, lower the sulfur dioxide.

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

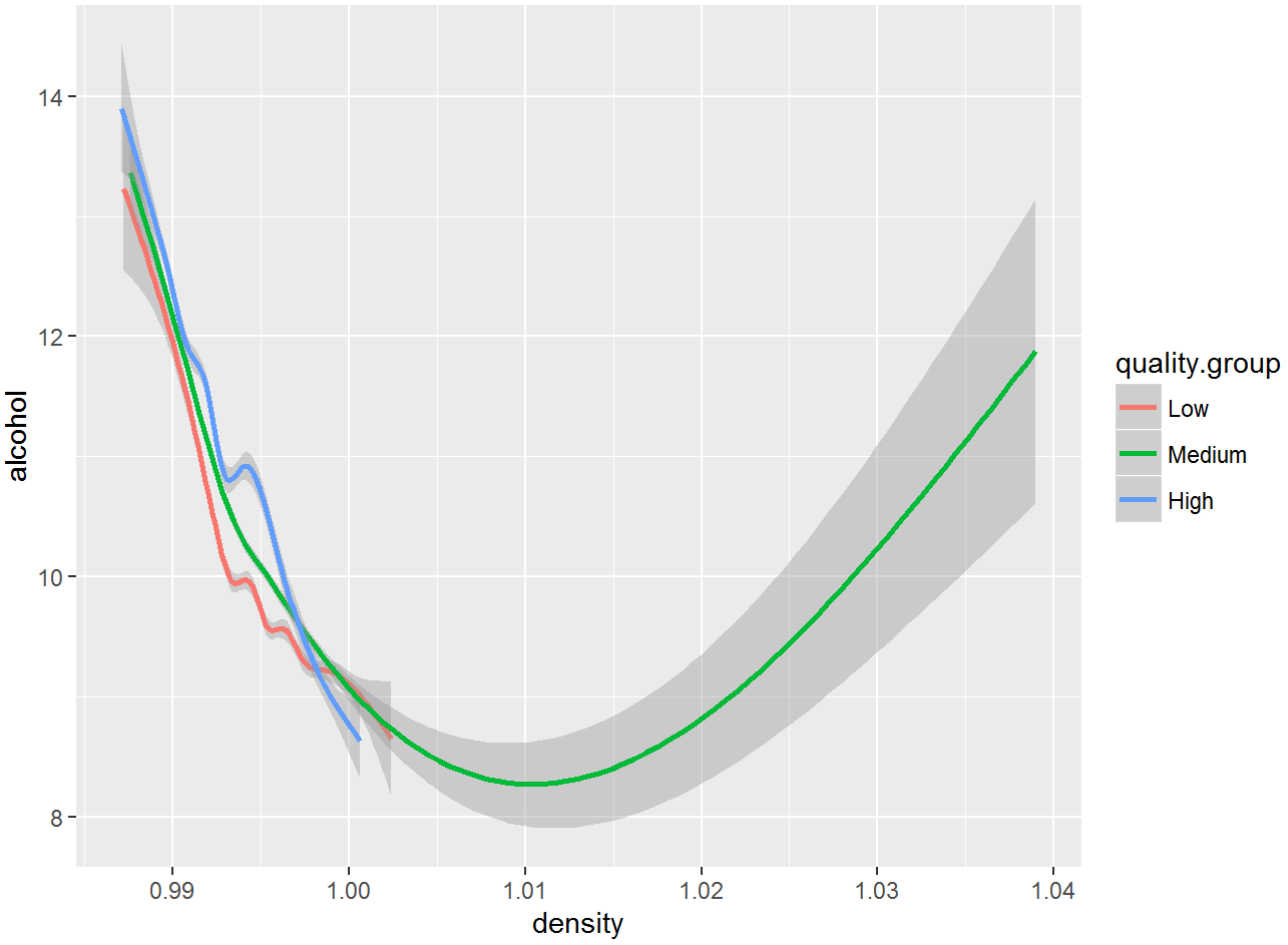
Density and residual sugar had the strongest correlation. Residual sugar has the most mass compared to the other chemical properties. As the yeast converts the sugars to alcohol, during fermentation, it makes sense that the density of the wine will decrease.

What was the strongest relationship you found?

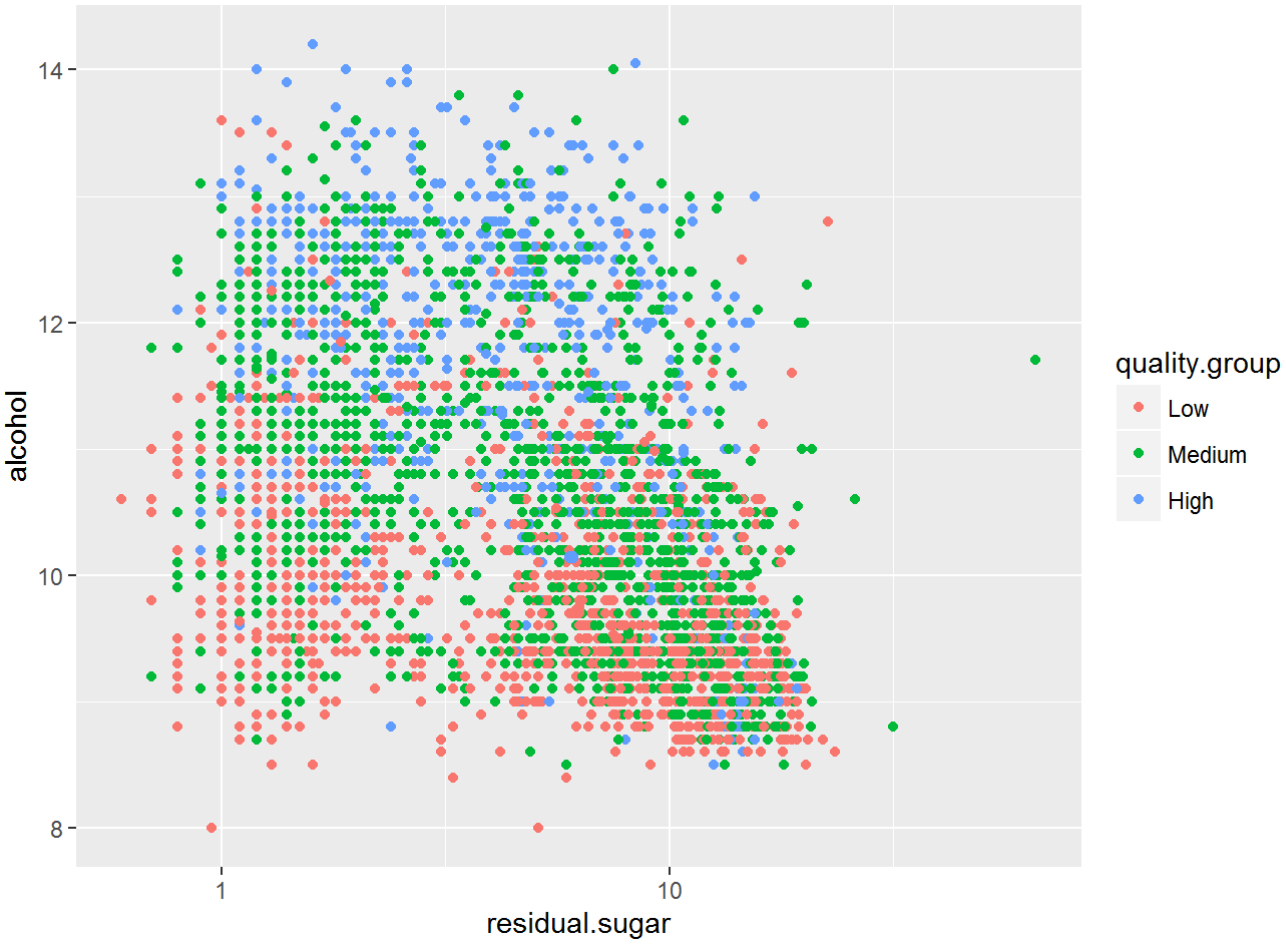
The strongest relationship with quality was with alcohol. As the quality of the wine increases, the median alcohol % increased.

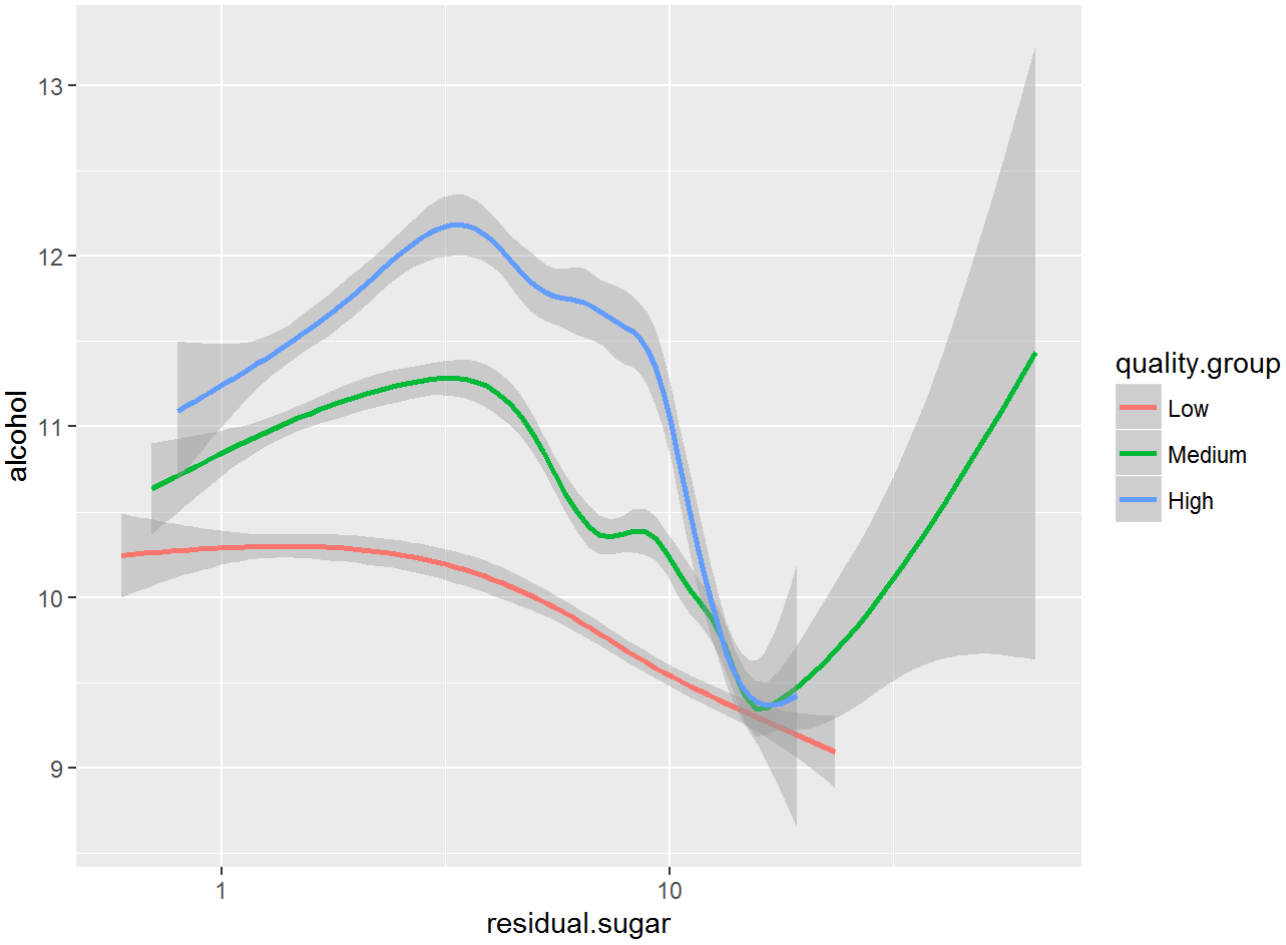
Multivariate Plots Section


alcohol
density
quality.group
Low
Medium
High


alcohol
density
quality.group
Low
Medium
High

As the scatter plot shows above, as alcohol in wine increases, the density decreases. However, wines of any quality has about the same alcohol content with the same density.


14
12
10
8
alcohol
1
10
residual.sugar
quality.group
Low
Medium
High


13
12
11
10
9
alcohol
1
10
residual.sugar
quality.group
Low
Medium
High

Unlike the density vs. alcohol chart, wines of higher quality has a higher alcohol content with the same sugar levels.

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?

When looking at wines with the same residual sugar, there is a significant difference in alcohol content when breaking it down by quality.

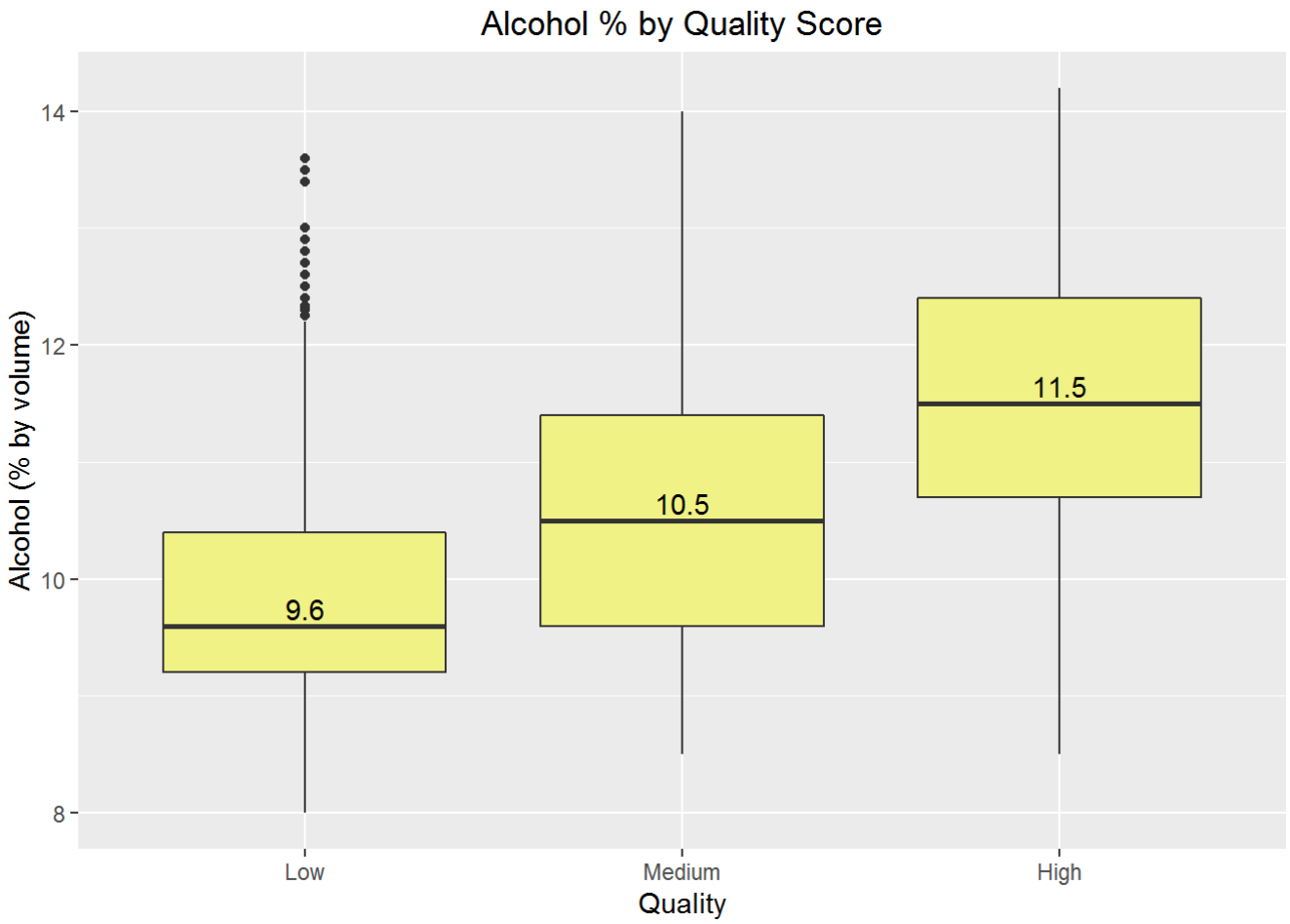
Final Plots and Summary

Plot One

Description One

The distribution of quality appears normally distributed. 92.5% of white wines were assessed a quality score between 5-7.

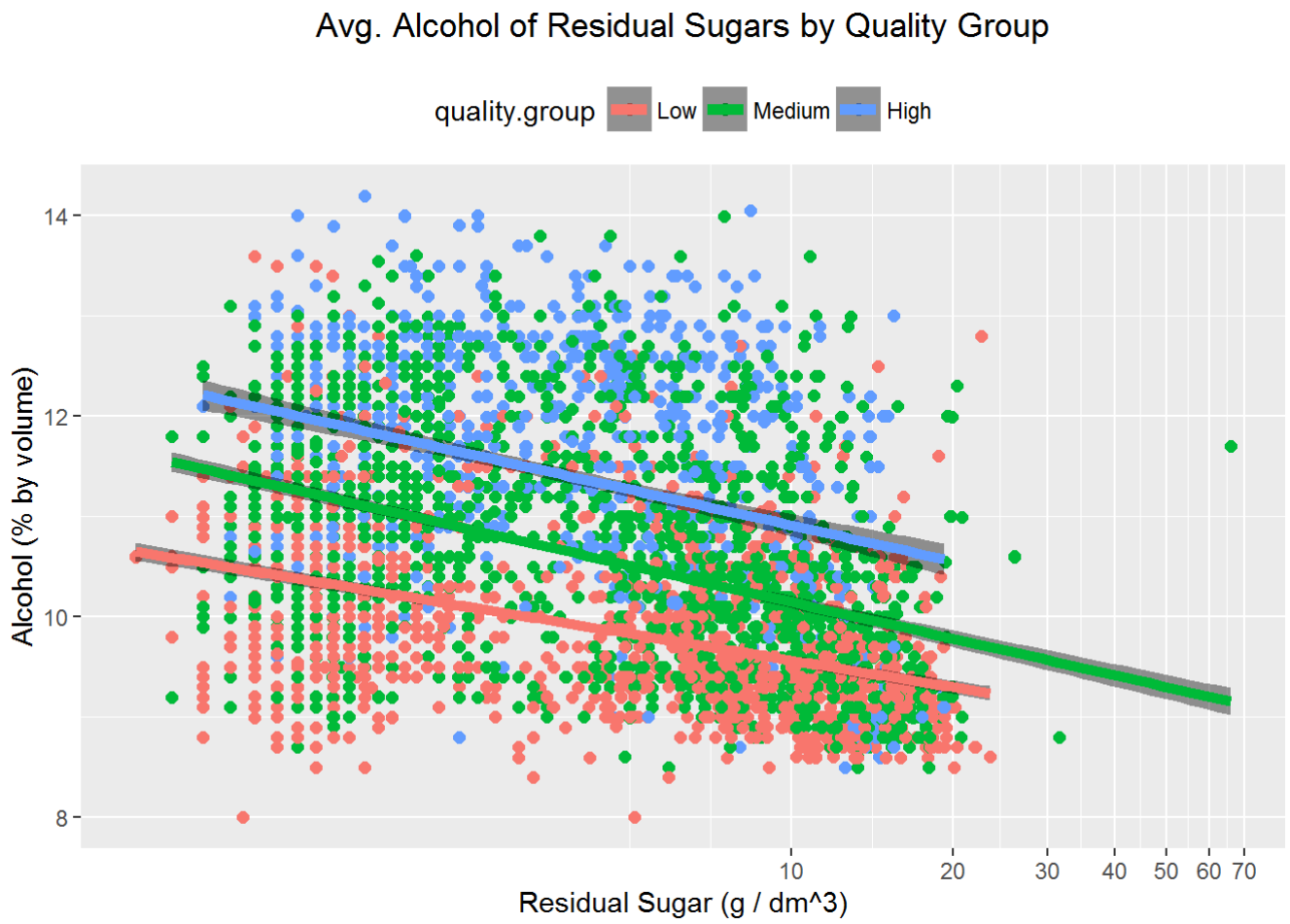
Plot Two



Description Two

Wines with a higher quality score have higher levels of alcohol. As you go up a quality group, the median alcohol level is about 1% higher than the previous group.

Plot Three



Description Three

Holding the residual sugar levels constant, wines of a higher quality will generally contain higher levels of alcohol by volume. There is also an interesting gap around 3 g/dm^3 residual sugars.

Reflection

To get myself familiar with this data set, I checked the distributions of each variable and hoped to find some unusual shapes in the graphs. Residual sugar had the most unusual distribution. Since it was bimodal, it made me think if most wines in the data set were of two types, dry and off-dry wines.

Since I wanted to know which chemical property contributes most to the quality of wine, I began to analyze quality to the properties There was a trend between quality and alcohol. The higher quality wines tend to have more alcohol compared to lower quality wines.

In hopes to develop a predictive model, I started to compare the chemical properties to each other. I could not find any interesting correlations between any variables except density, alcohol, residual sugar, and total sulfur dioxide. I struggled to understand the relationship between total sulfur dioxide, residual sugar, and alcohol.

For future work, I am interested to see how wine quality differs between countries. Using the same wines from the dataset, how would other countries score the wines? Also, can wine quality change depending on weather or season?