Analysis of Red Wine Quality by Jennifer Tsou

Chemical Properties:

- fixed acidity: most acids involved with wine or fixed or nonvolatile (do not evaporate readily) (tartaric acid - g / dm^3)
- volatile acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste (acetic acid - g / dm^3)
- citric acid: found in small quantities, citric acid can add 'freshness' and flavor to wines (g / dm^3)
- residual sugar: the amount of sugar remaining after fermentation stops (g / dm^3)
- chlorides: the amount of salt in the wine (sodium chloride - g / dm^3
- free sulfur dioxide: he free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion (mg / dm^3)
- total sulfur dioxide: amount of free and bound forms of S02 (mg / dm^3)
- density: the density of water is close to that of water depending on the percent alcohol and sugar content (g / cm^3)
- pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic)
- sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels (potassium sulphate - g / dm3)
- alcohol: the percent alcohol content of the wine (% by volume)

Output variable (based on sensory data):

- quality (score between 0 and 10)

Univariate Plots Section

The report explores a dataset containing wine quality and attributes for aprroximately 1599 red wines.

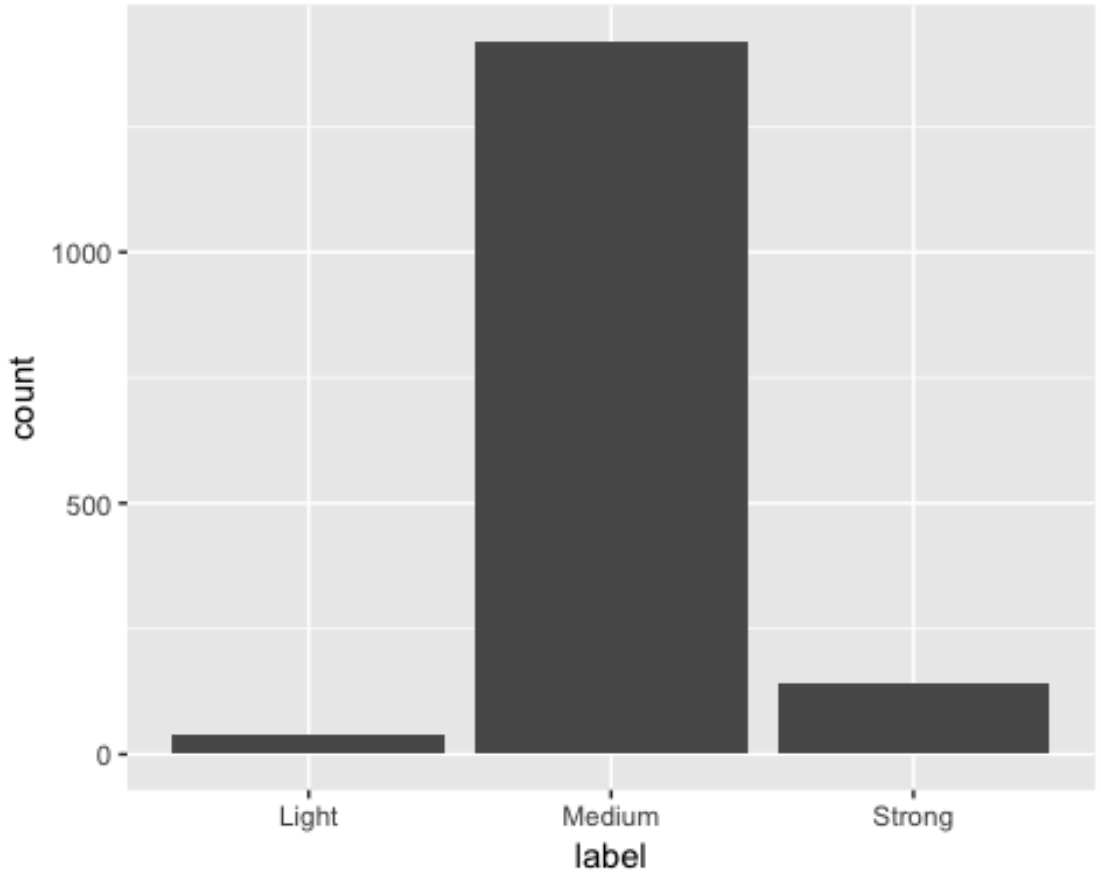
```
I have chosen Red Wine Dataset because I've always been a fan of red wine.

## 'data.frame':    1599 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
```

```
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.
5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.
065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3
.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57
0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

Remove 'X' variable since it is not relevant in the exploration.

```
##
##  Light Medium Strong
##     37   1421    141
```

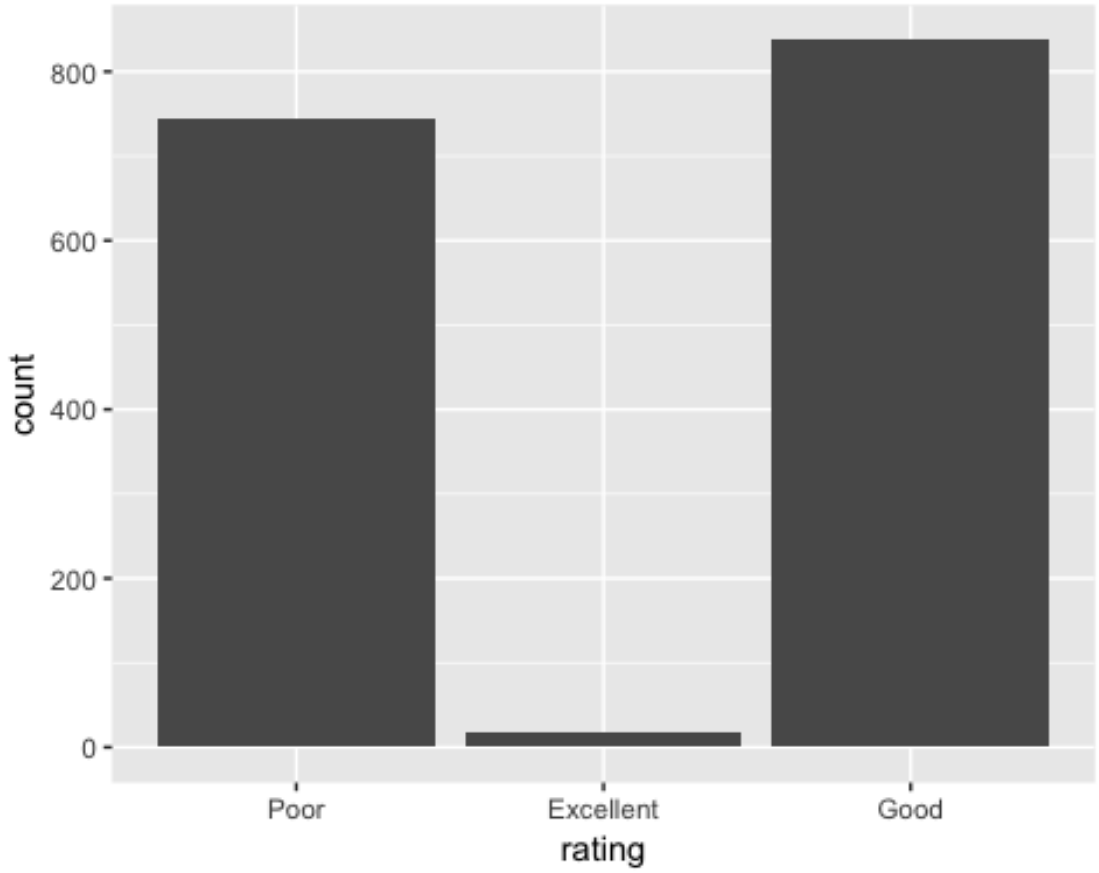
Adding label as alcohol percentage to group the alcohol concentration in the following:
Light = Less than 9 Medium = Between 9 and 12 Strong = Greater than 12



Simple exploration of how many light, medium, strong wine are out there in this sample.

```
##
##      Poor Excellent      Good
##       744        18       837
```

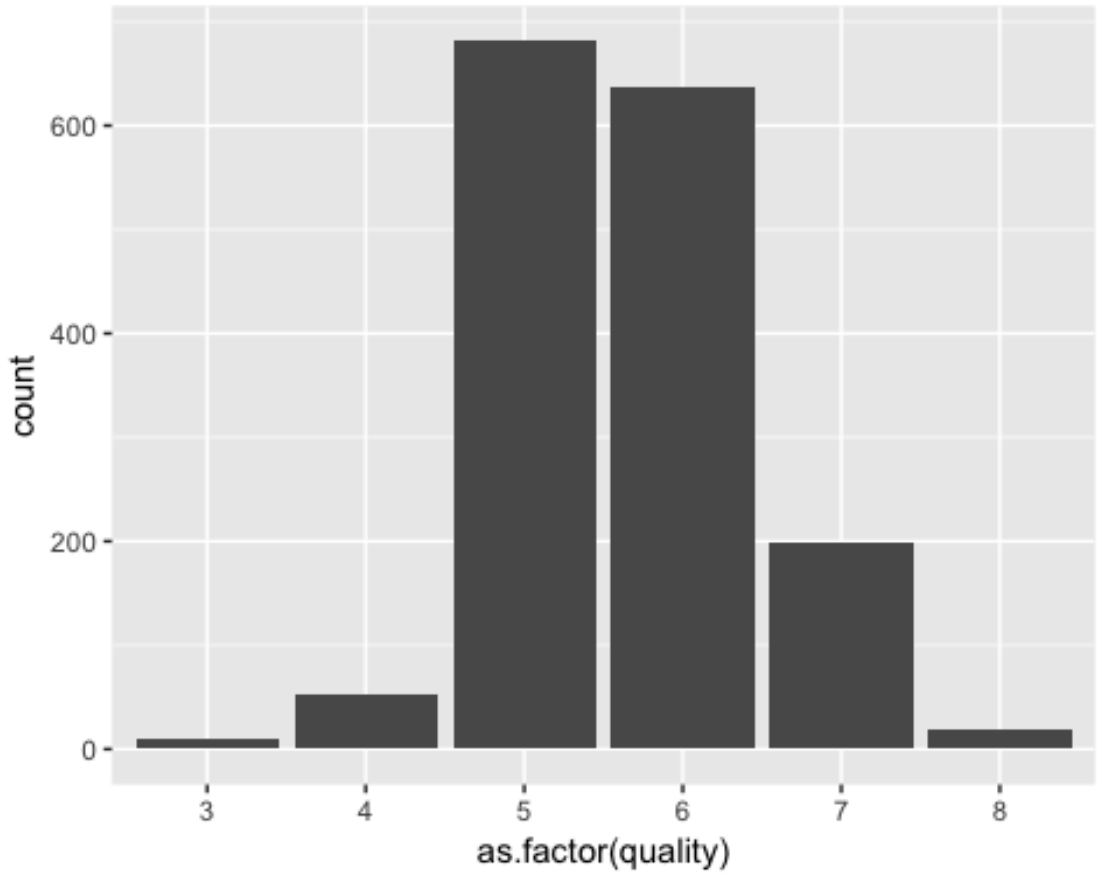
Adding rating as wine quality to group the wine quality in the following: Poor = less than 5 Good = between 5 and 8 Excellent = greater than 8



Simple exploration about how many quality wines are out there.

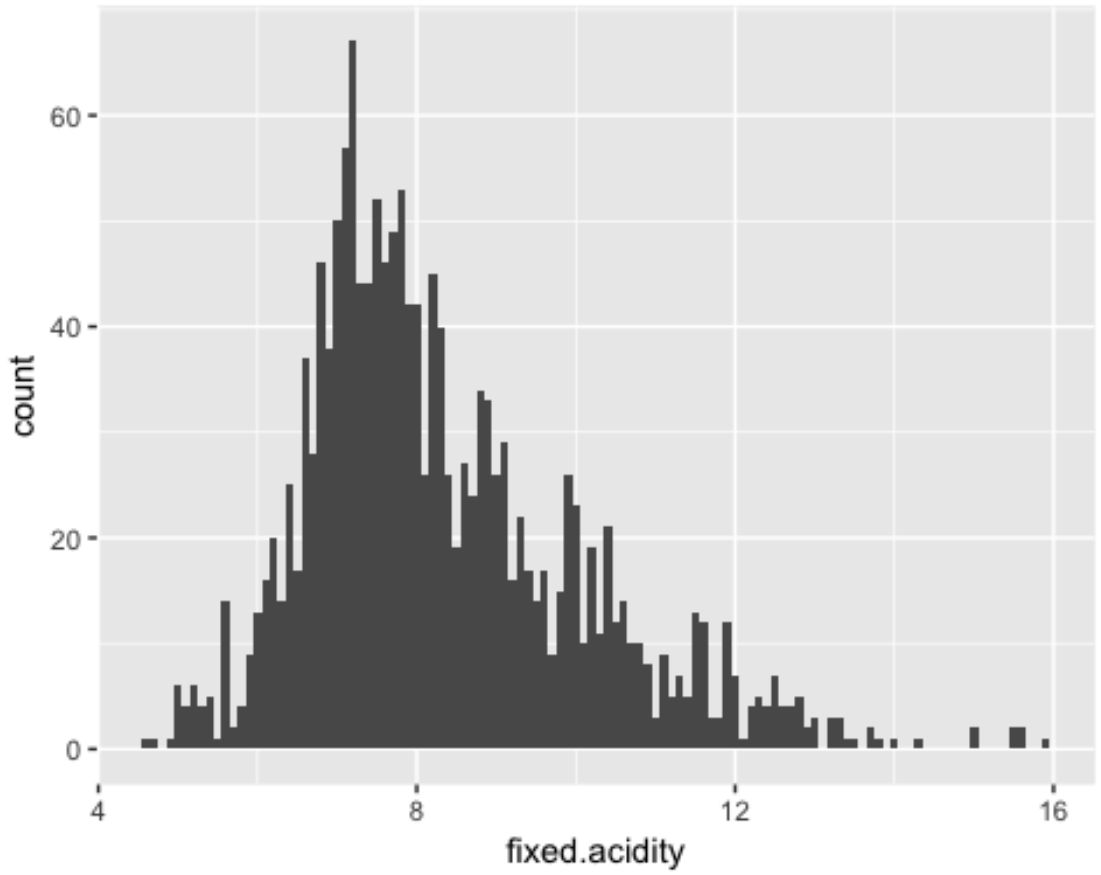
```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean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

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



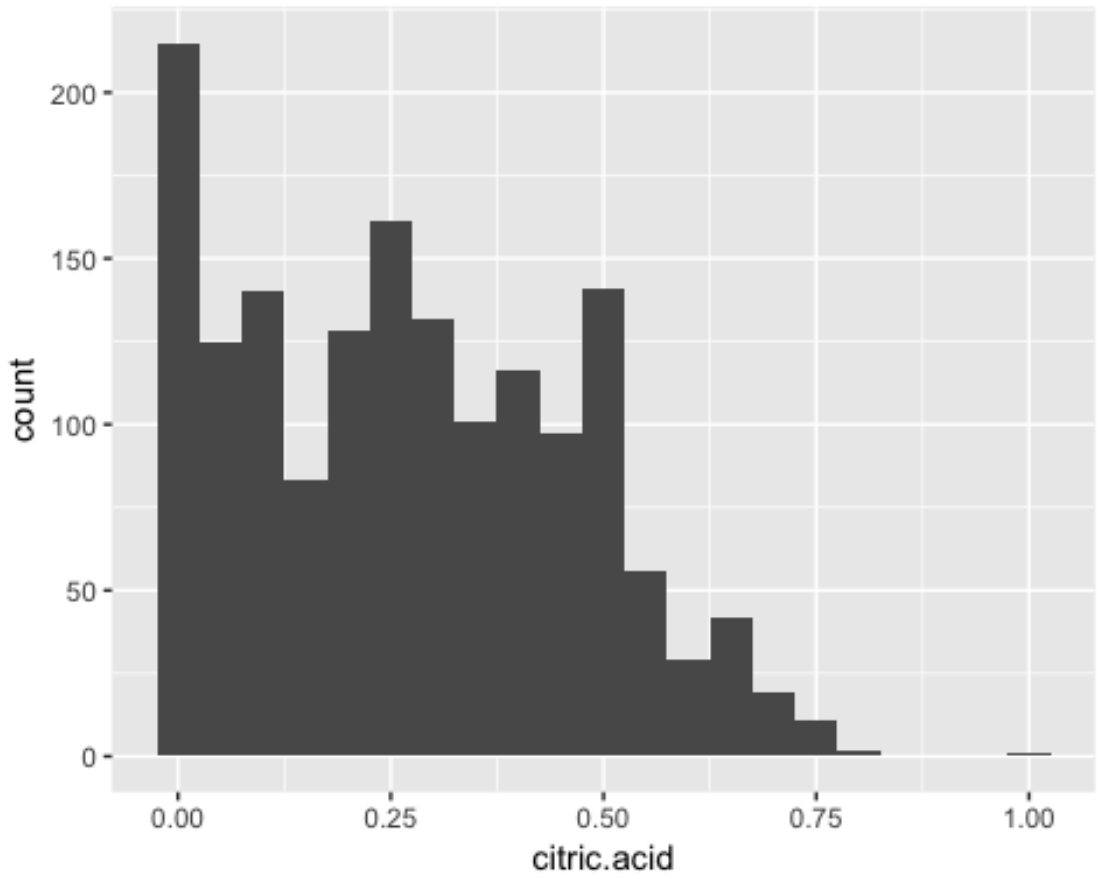
I created a histogram of wine quality vs count to have an idea of how the wine quality distributes. We will further investigate into what affect wine quality.

I will now created simple univariate plots using the variables.

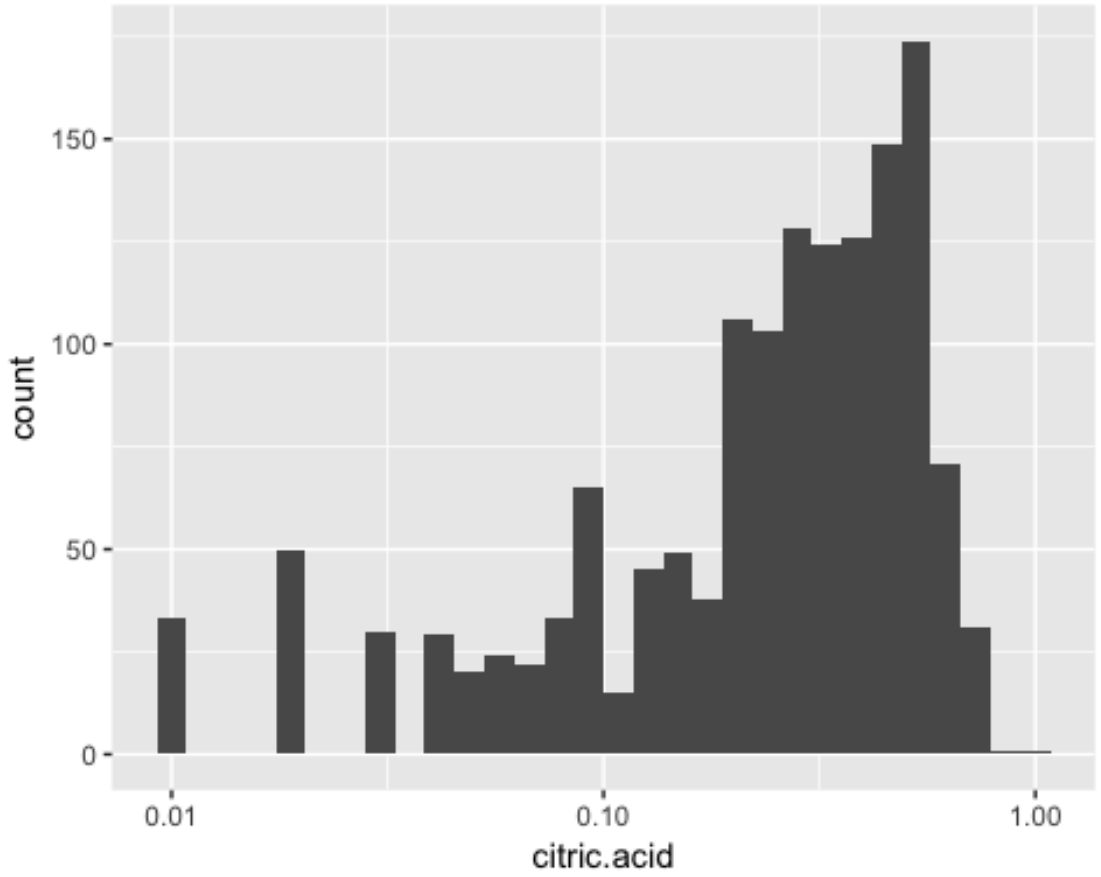


The graph is slightly skewed to the right.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
```

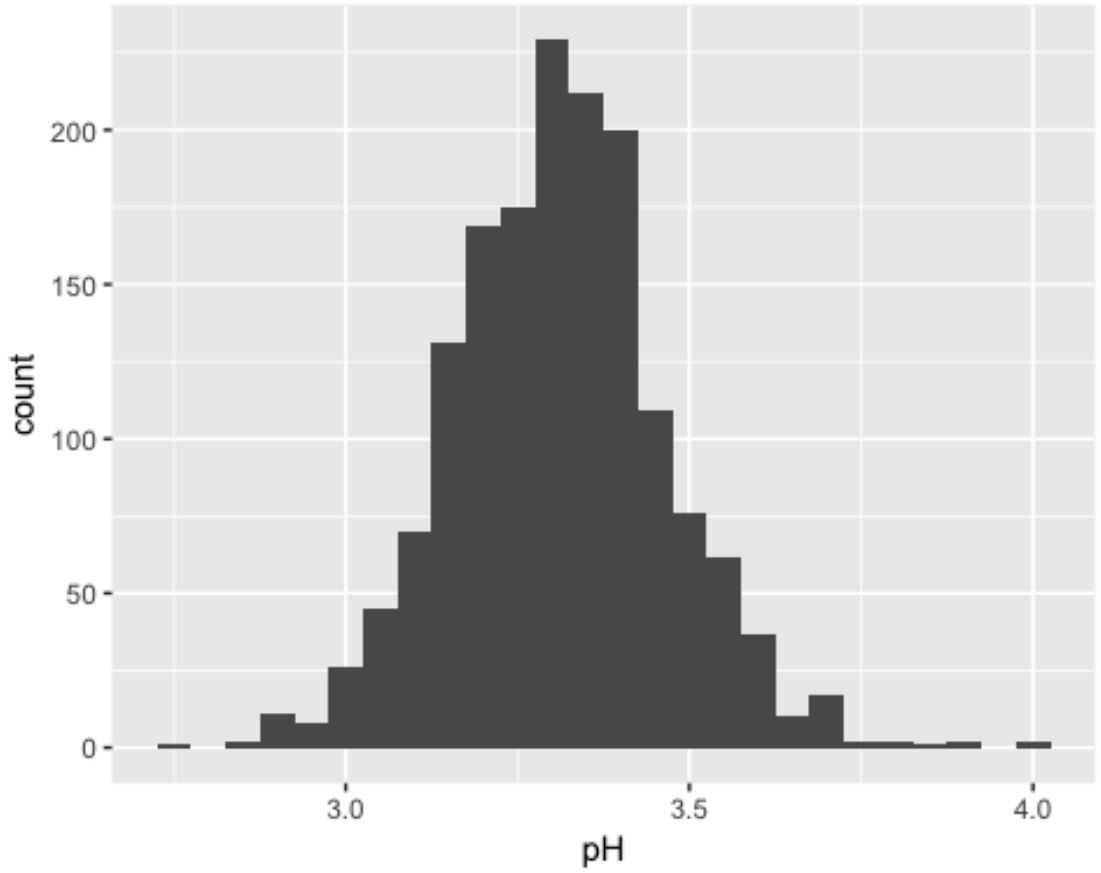


Citric acid has a skewed to the right distribution. Attempt to transform the data did not work.



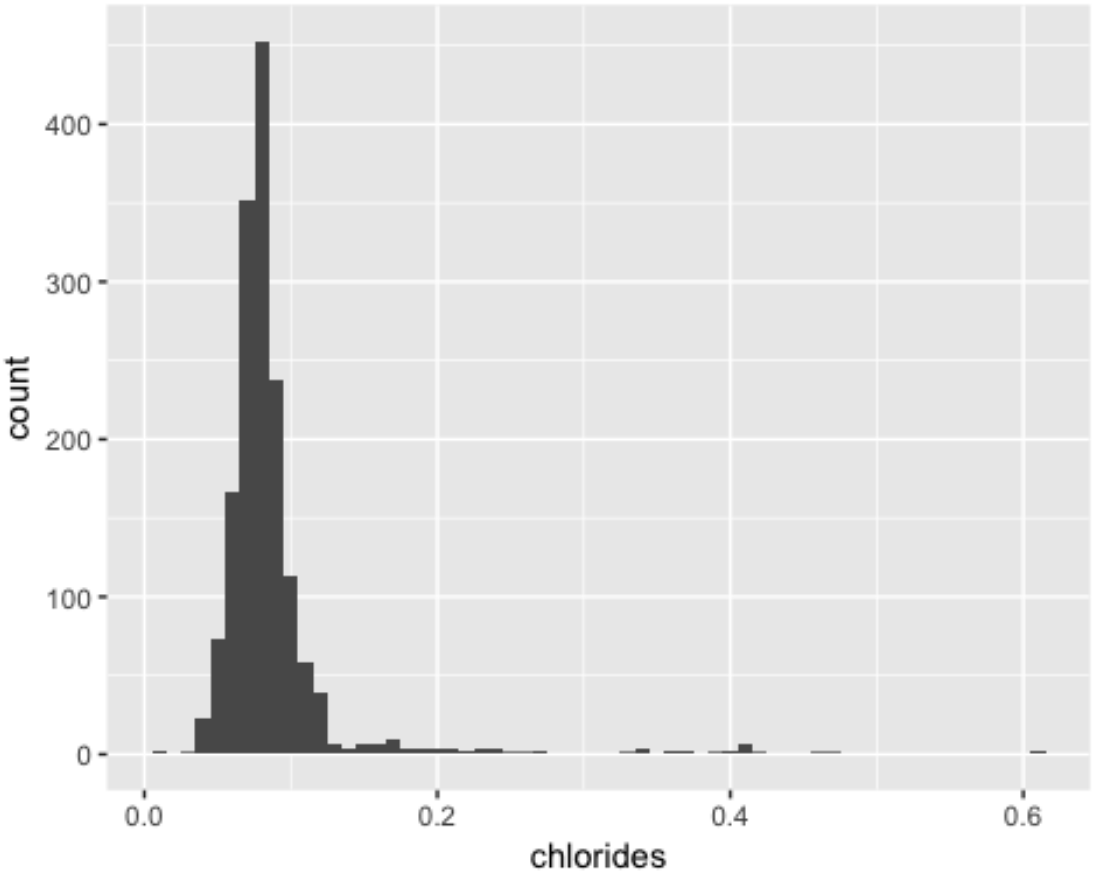
As the graph suggests, not only did the transformation did not work, it is now skewed to the left.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```

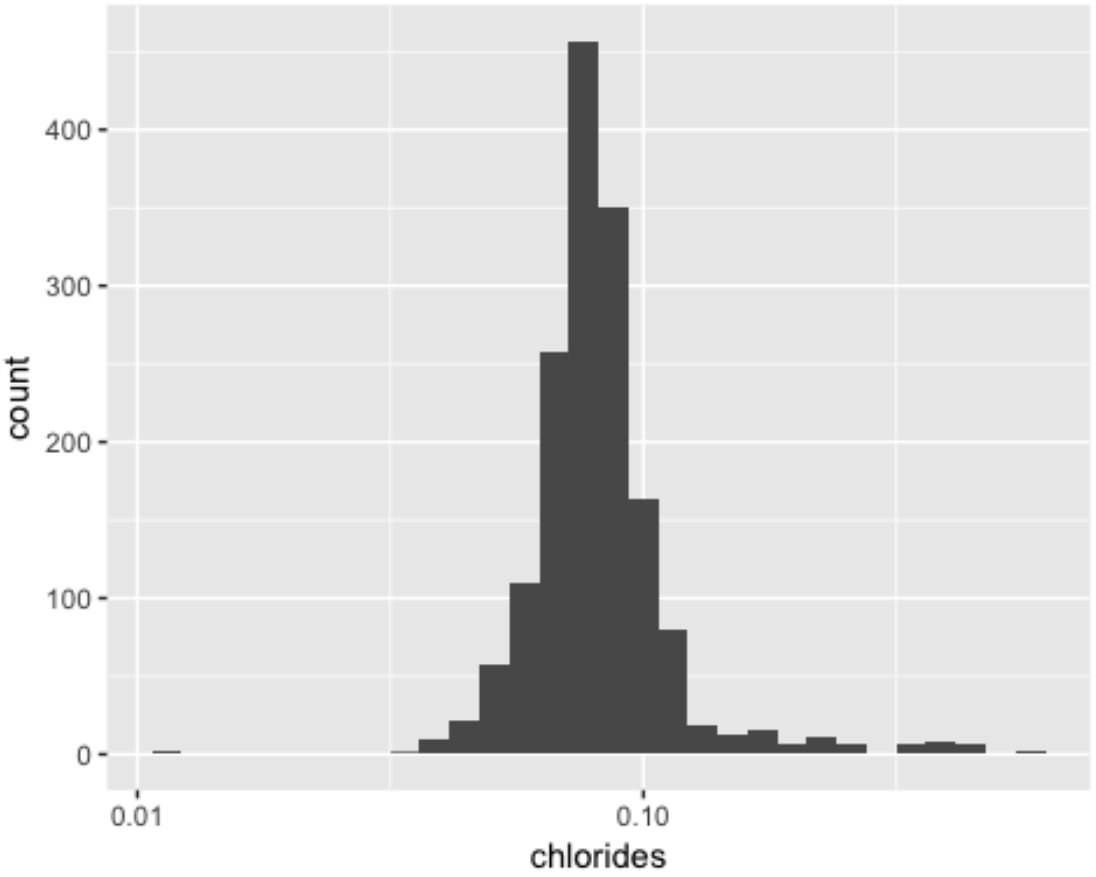


The pH distribution appear to be normal with summary as followed:

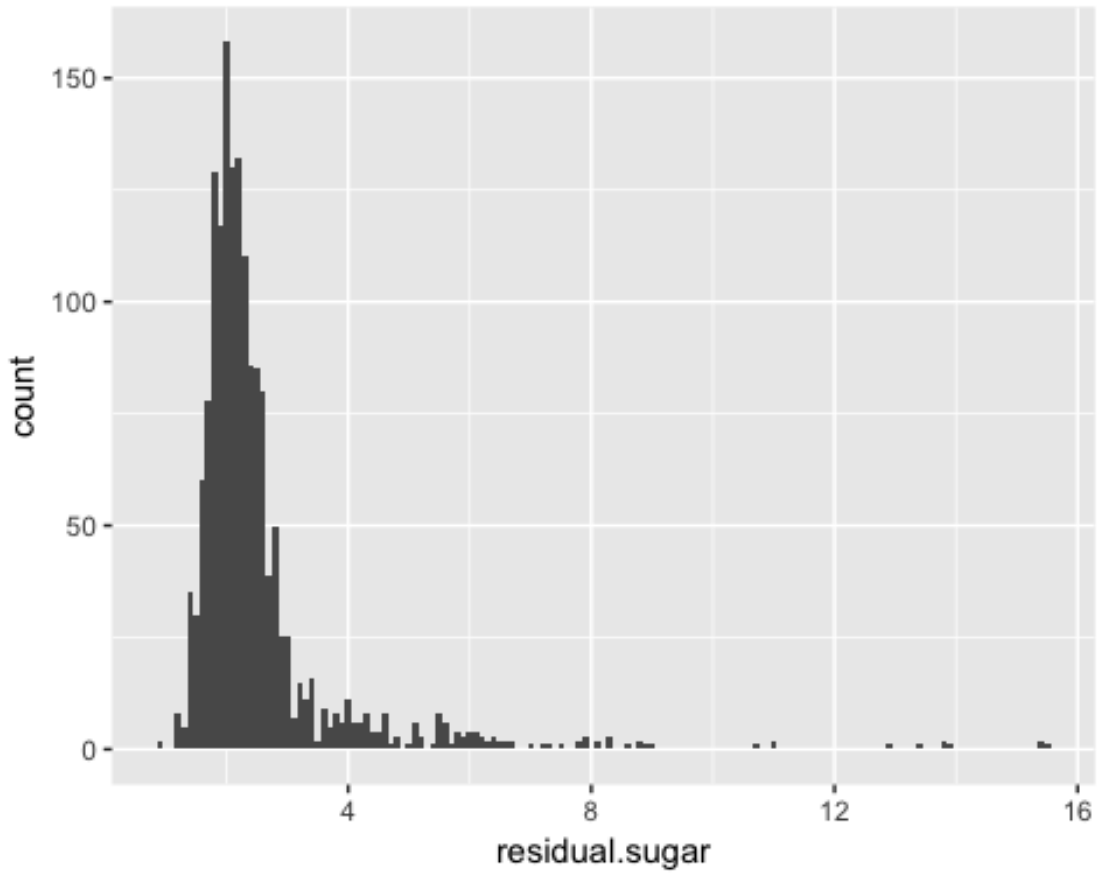
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010
```



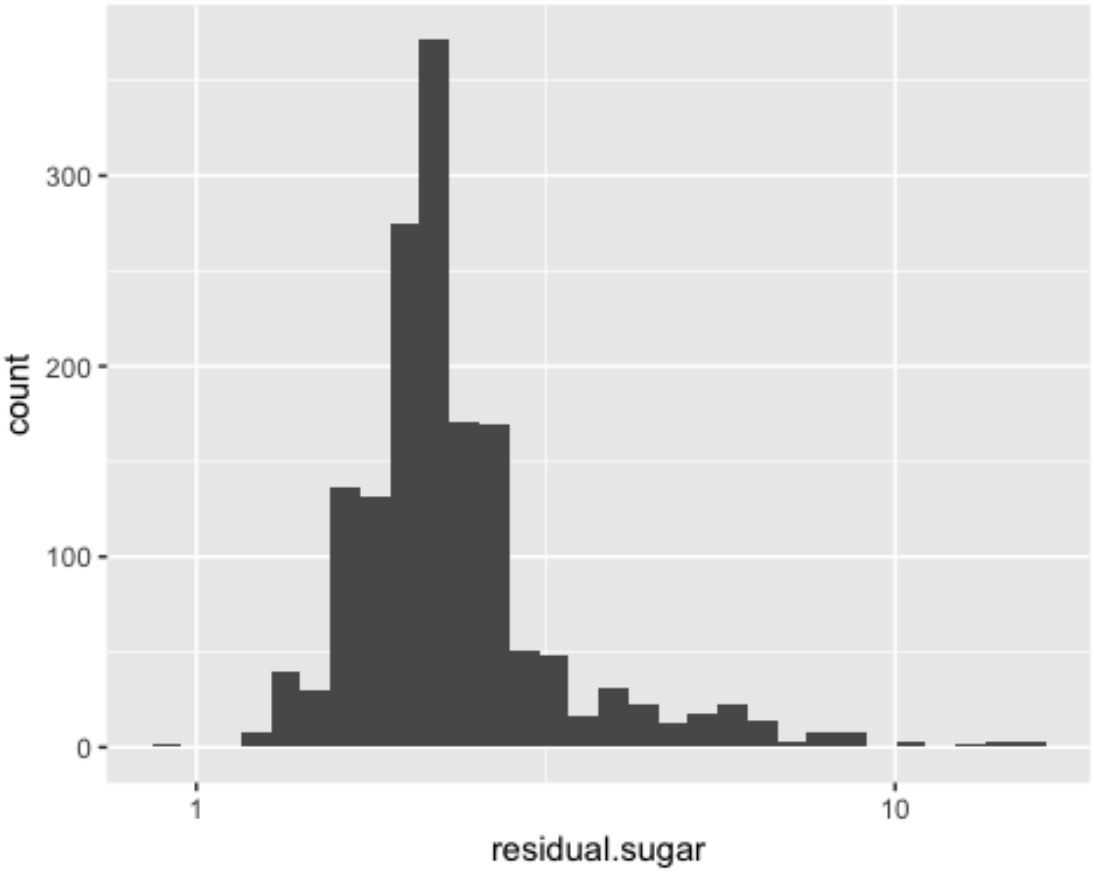
The distribution of chlorides is skewed to the right. Applying log10 transformation to the graph to reduce the variability.



After applying tranformation to the graph, we can see that it is normal now.

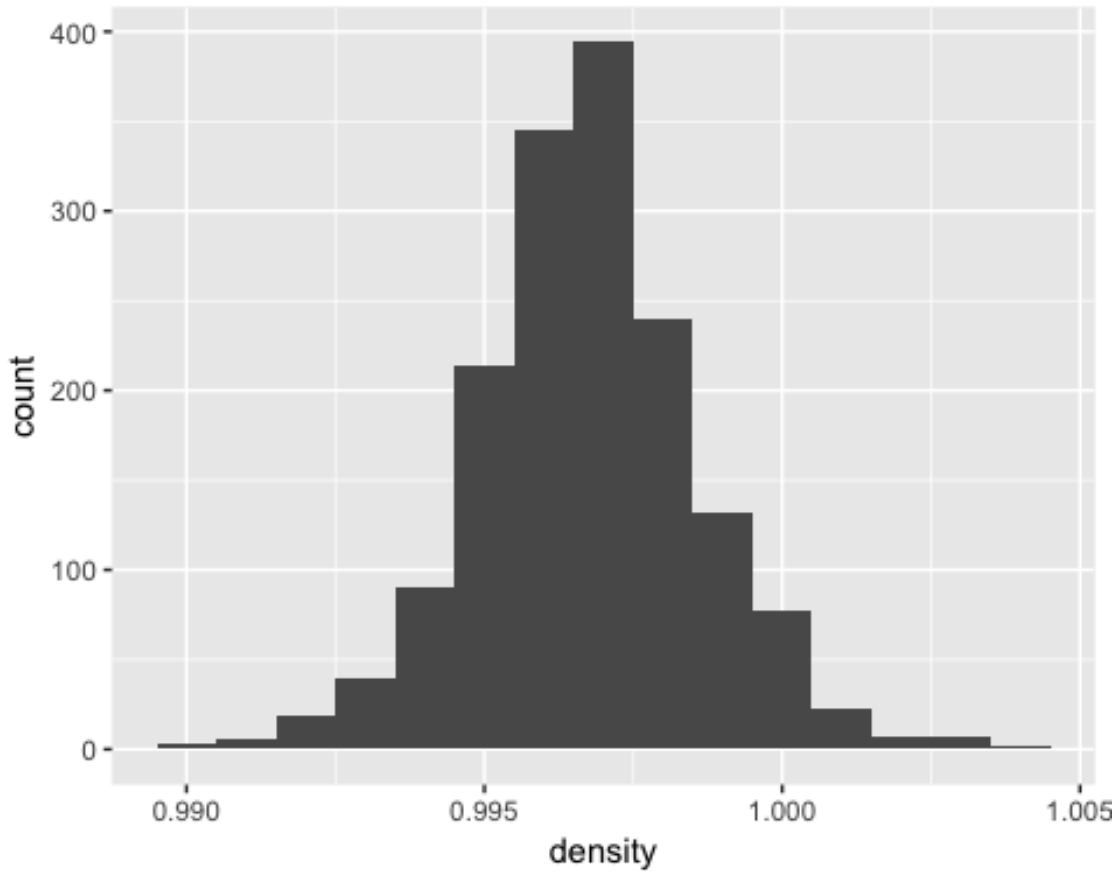


The distribution for residual sugar is skewed to the right. I will attempt to apply log10 transformation.

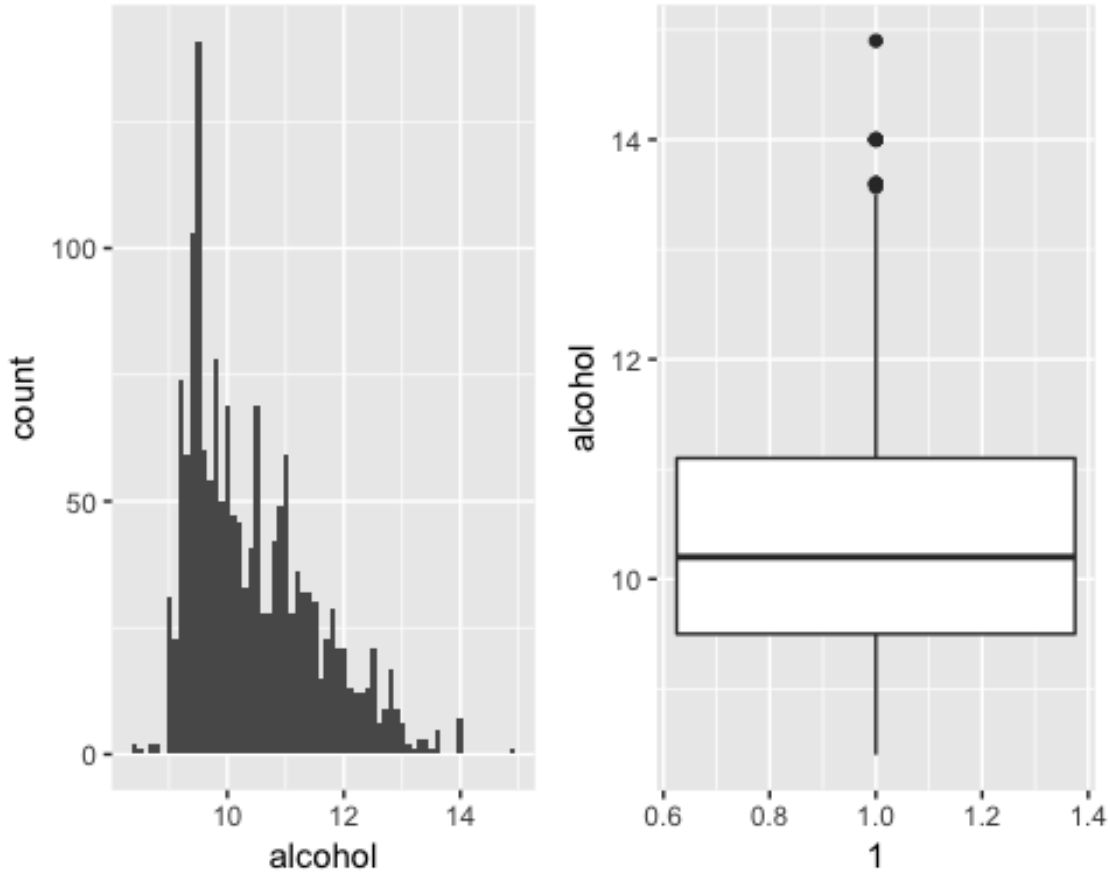


As the graph show, it is still skewed to the right. There is an outlier of 15.5 that caused the issue.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.900   1.900   2.200   2.539   2.600  15.500
```

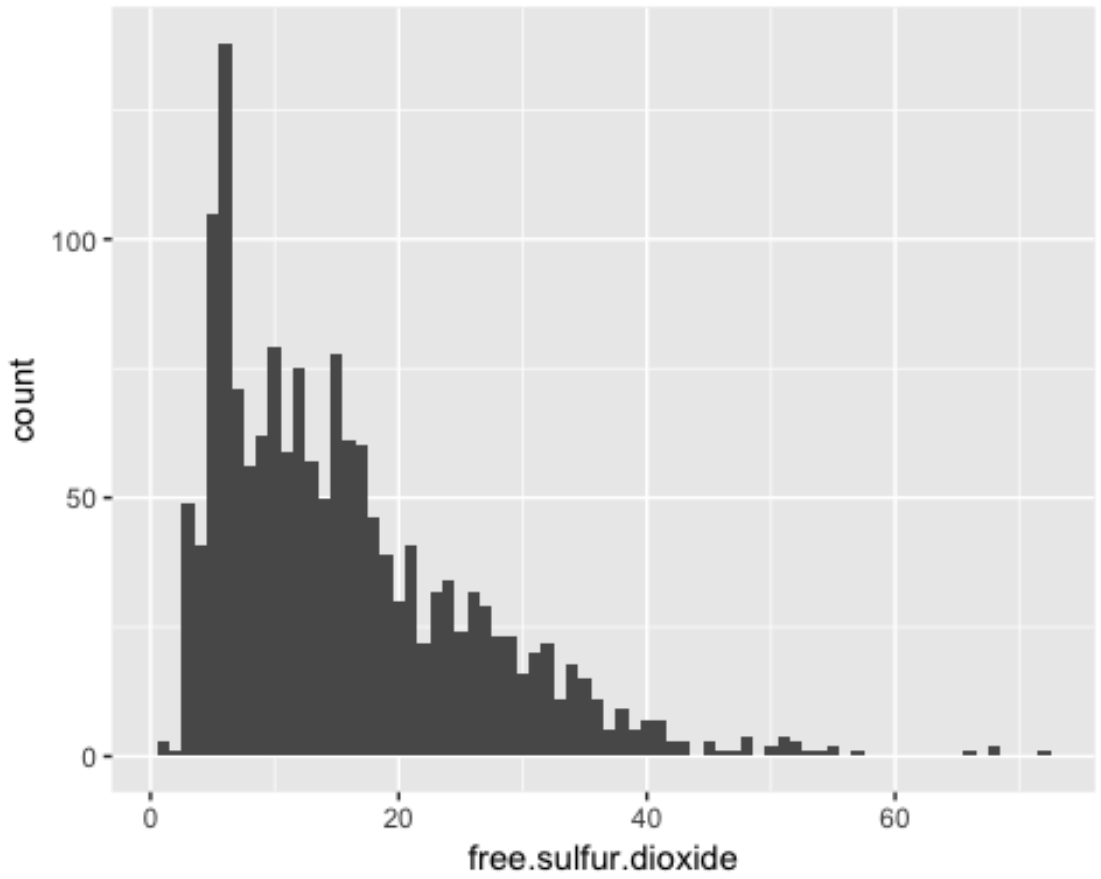


Density has a normal distribution.

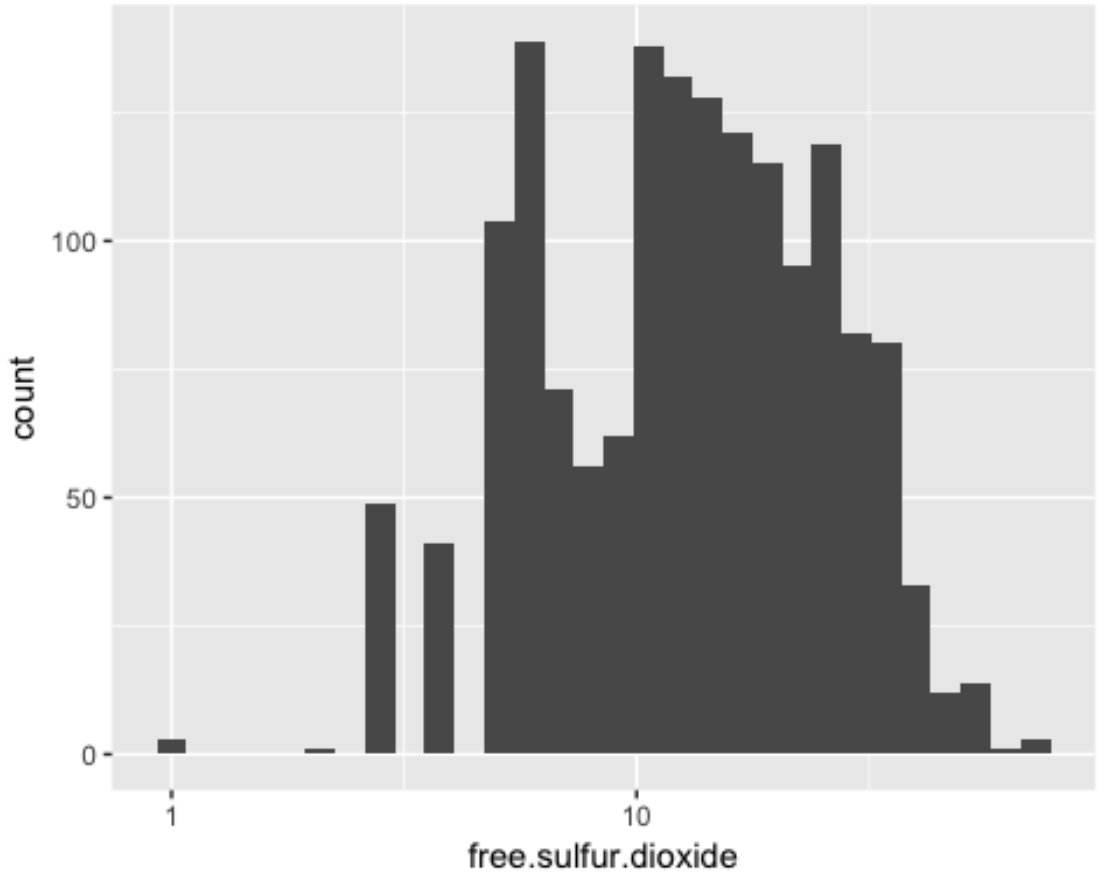


From the distribution, we can see that most alcohol percentage on concetrated from 9 to 11.

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

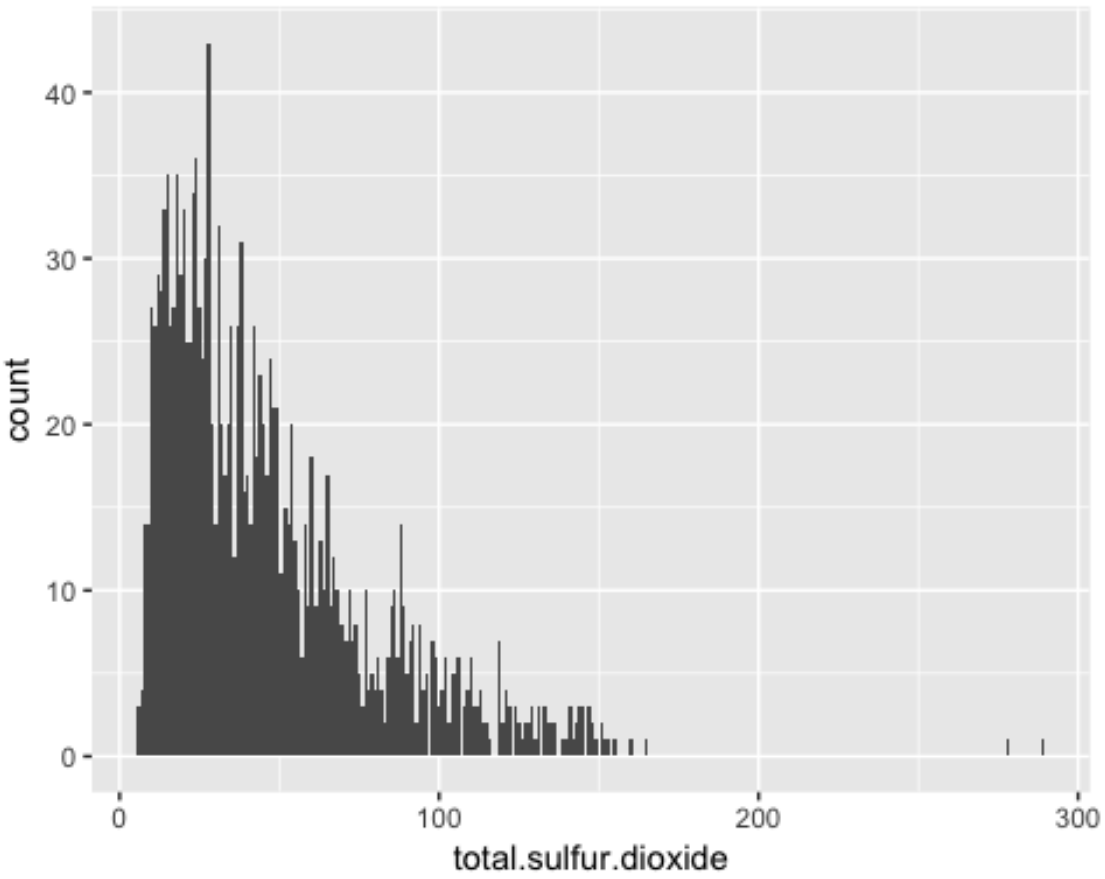


The distribution is skewed to the right.

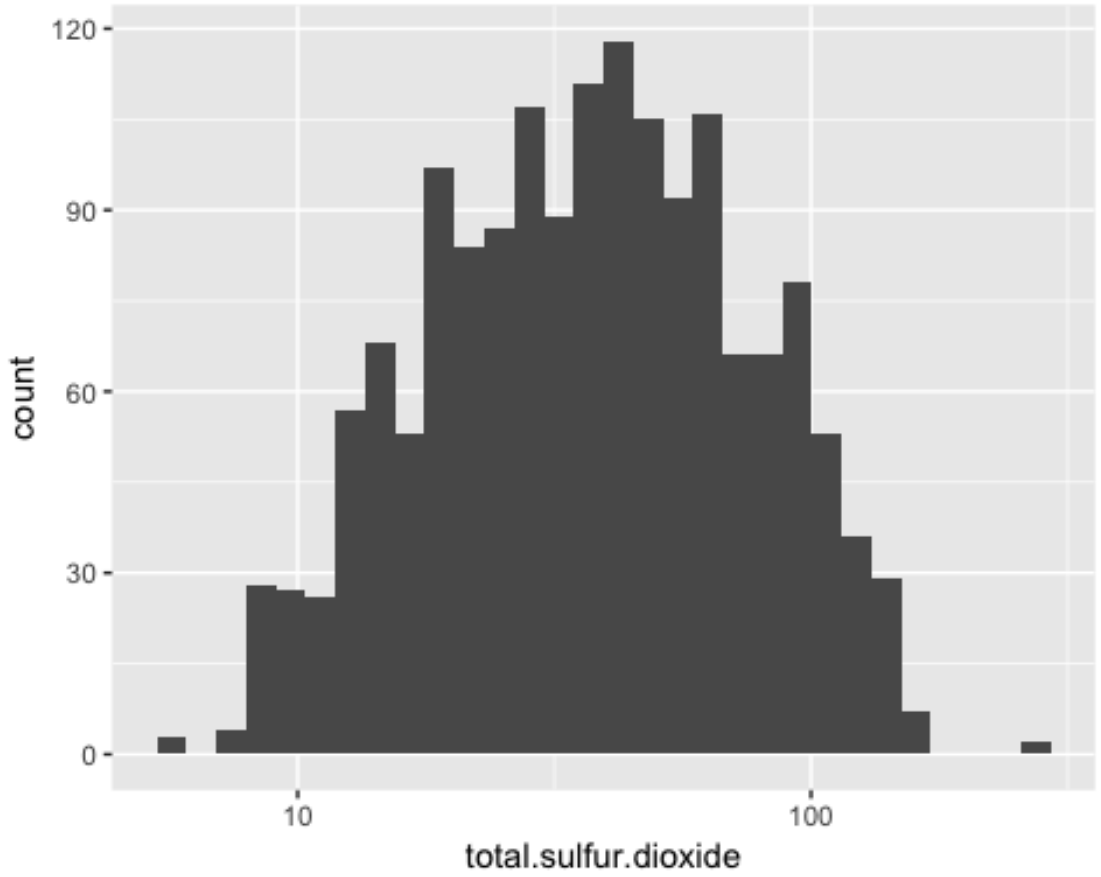


After log10 transformation is applied, I see that free almost has a bimodol distribution.

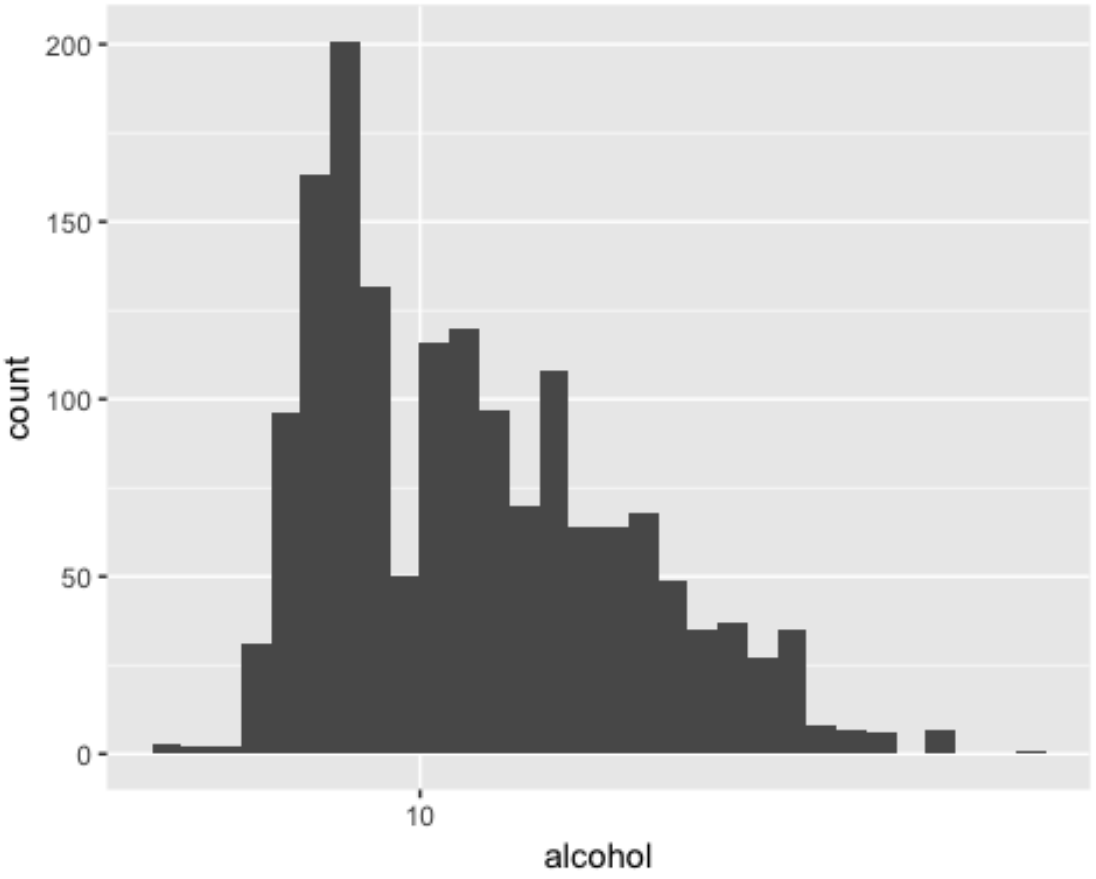
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
```

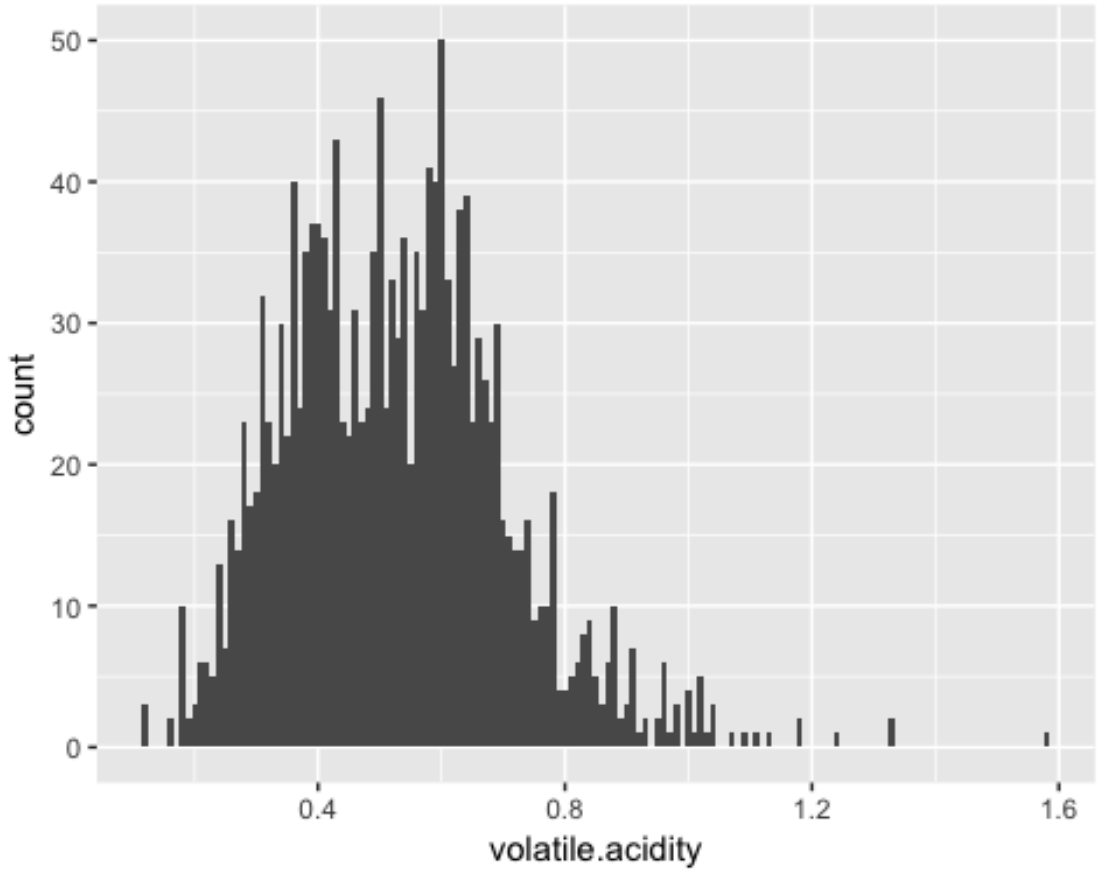


Total sulfur is heavily skewed to the right.



After applying the transformation, it is now a normal distribution.


200
150
100
50
0
count
10
alcohol



Univariate Analysis

What is the structure of your dataset?

There are 1599 red wine in the dataset with 14 variables as listed below.

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

```
##  [1] "fixed.acidity"        "volatile.acidity"     "citric.acid"
##  [4] "residual.sugar"       "chlorides"            "free.sulfur.dioxide"
##  [7] "total.sulfur.dioxide" "density"              "pH"
## [10] "sulphates"            "alcohol"              "quality"
## [13] "label"                "rating"
```

Out of the 16 variables, which included the newly added 2, I am mainly interested in chlorides, density, alcohol percentage, and quality.

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

The main features in the dataset are rating, label and residual sugar. I would like to determine what feature are best for predicting the quality of red wine. I believe that alcohol concentration along with some combinations of other variables can be used to make such a prediction.

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

I created rating and labels as my new variables from wine quality and alcohol percentage. I grouped those two variables to help me perform analysis later on.

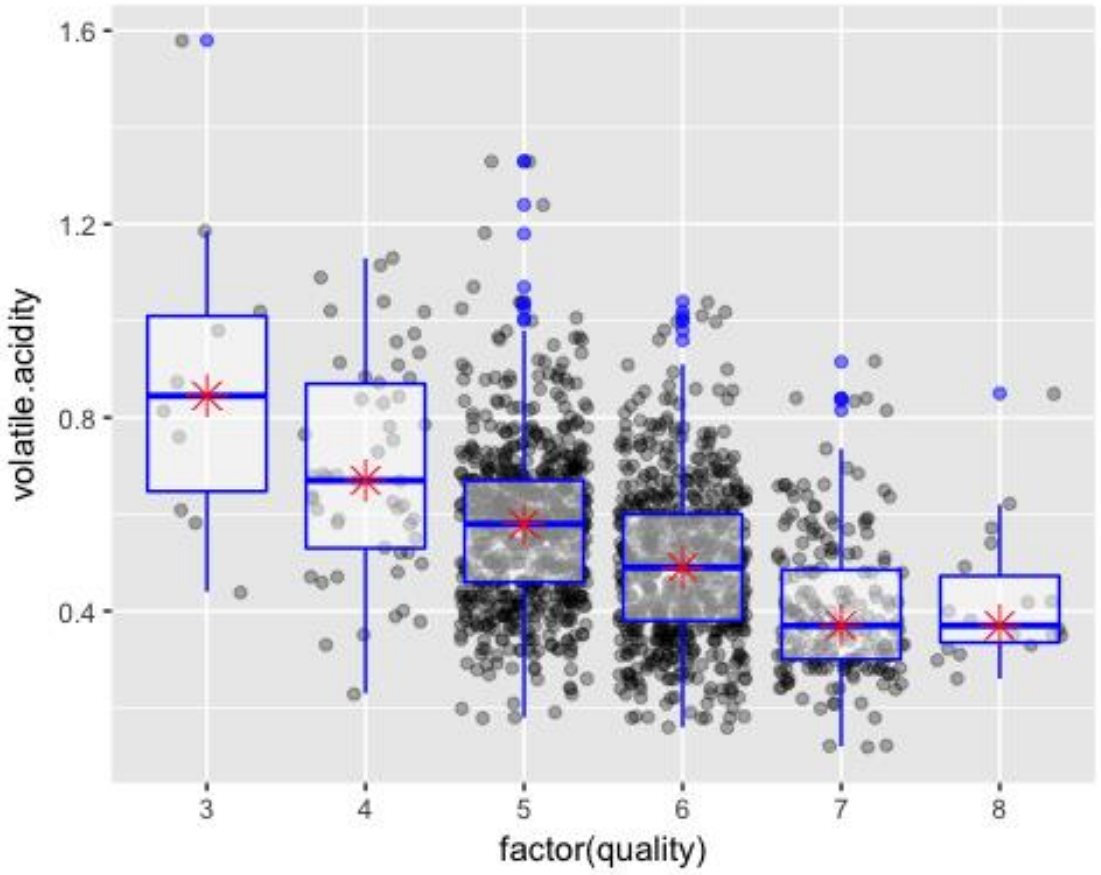
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

I noticed that alcohol concentration and residual sugar are not normally distributed, so I tried to make it normal by log-transformed the right skewed distribution. However, it remained right skewed. Therefore I have reason to believe that alcohol concentration and residual sugar have some influence on the quality of wine.

Bivariate Plots Section

```
## # A tibble: 6 × 4
##   quality mean_volatile.acidity median_volatile.acidity     n
##     <int>                 <dbl>                   <dbl> <int>
## 1       3             0.8845000                   0.845    10
## 2       4             0.6939623                   0.670    53
## 3       5             0.5770411                   0.580   681
## 4       6             0.4974843                   0.490   638
## 5       7             0.4039196                   0.370   199
## 6       8             0.4233333                   0.370    18
```

```
##     quality     mean_volatile.acidity median_volatile.acidity
##  Min.   :3.00   Min.   :0.4039        Min.   :0.3700
##  1st Qu.:4.25   1st Qu.:0.4419        1st Qu.:0.4000
##  Median :5.50   Median :0.5373        Median :0.5350
##  Mean   :5.50   Mean   :0.5800        Mean   :0.5542
##  3rd Qu.:6.75   3rd Qu.:0.6647        3rd Qu.:0.6475
##  Max.   :8.00   Max.   :0.8845        Max.   :0.8450
##        n
##  Min.   : 10.00
##  1st Qu.: 26.75
##  Median :126.00
##  Mean   :266.50
##  3rd Qu.:528.25
##  Max.   :681.00
```



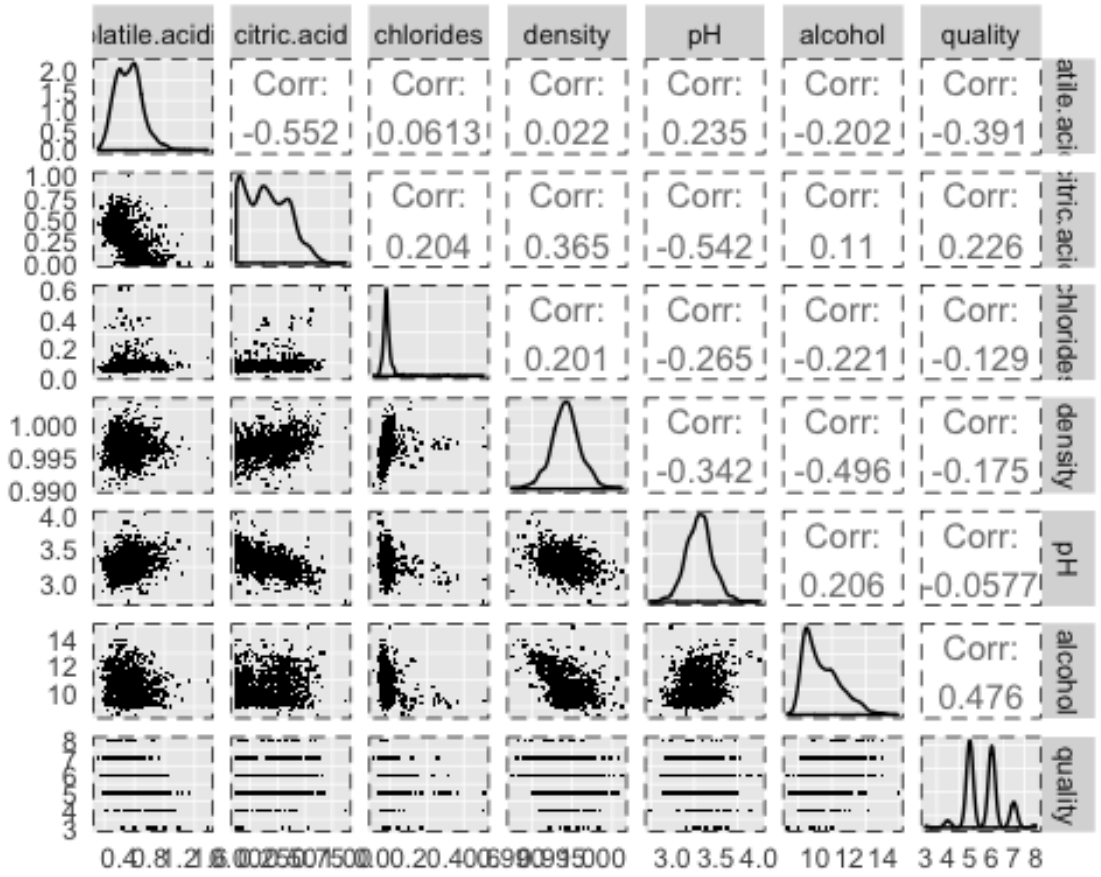
We can see here the trend of boxplot with the higher the quality leading to the lower the volatile acidity. We look at the median value here for fair comparison, to avoid any outlier.

The source of dataset came from volatile acidity, I just took the median value.

```
##                       fixed.acidity volatile.acidity citric.acid
## fixed.acidity                 1.000           -0.256       0.672
## volatile.acidity             -0.256            1.000      -0.552
## citric.acid                   0.672           -0.552       1.000
## residual.sugar                0.115            0.002       0.144
## chlorides                     0.094            0.061       0.204
## free.sulfur.dioxide          -0.154           -0.011      -0.061
## total.sulfur.dioxide         -0.113            0.076       0.036
## density                       0.668            0.022       0.365
## pH                           -0.683            0.235      -0.542
## sulphates                     0.183           -0.261       0.313
## alcohol                      -0.062           -0.202       0.110
## quality                       0.124           -0.391       0.226
##                       residual.sugar chlorides free.sulfur.dioxide
## fixed.acidity                  0.115     0.094               -0.154
## volatile.acidity               0.002     0.061               -0.011
## citric.acid                    0.144     0.204               -0.061
## residual.sugar                 1.000     0.056                0.187
```

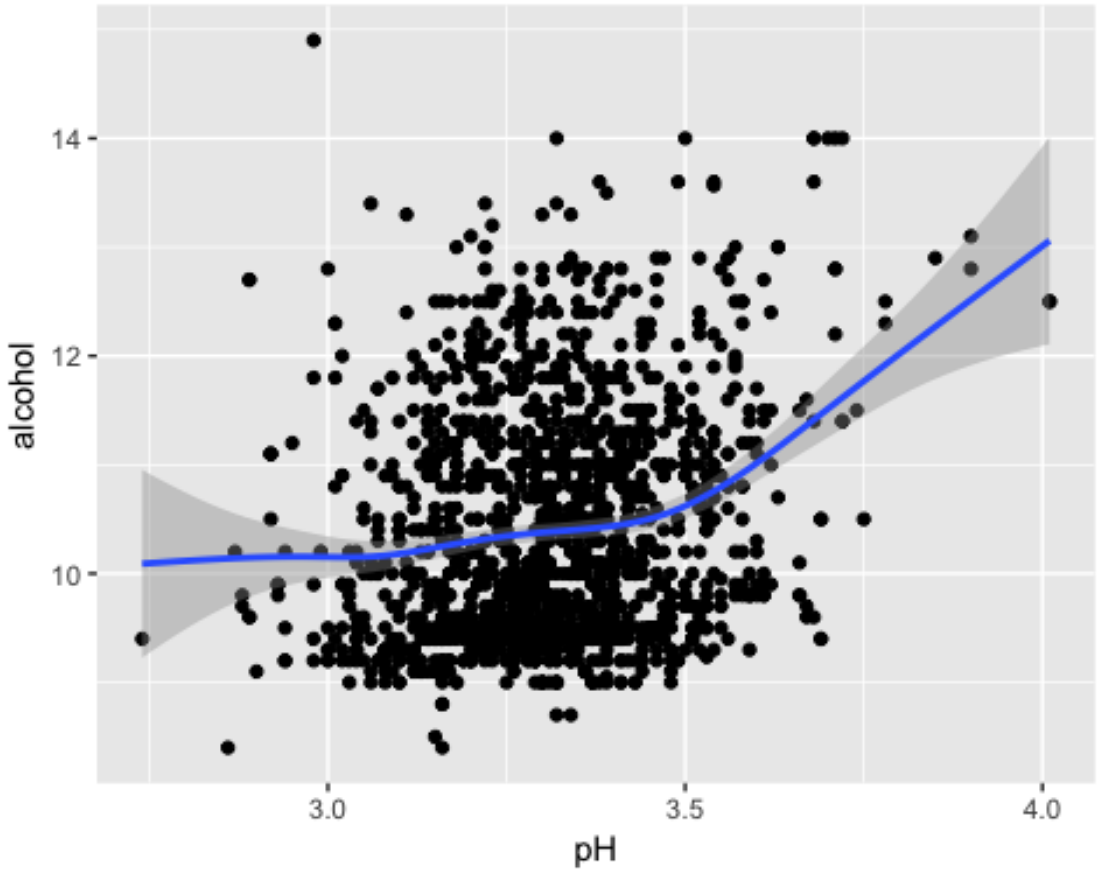
```
## chlorides                     0.056     1.000              0.006
## free.sulfur.dioxide           0.187     0.006              1.000
## total.sulfur.dioxide          0.203     0.047              0.668
## density                       0.355     0.201             -0.022
## pH                           -0.086    -0.265              0.070
## sulphates                     0.006     0.371              0.052
## alcohol                       0.042    -0.221             -0.069
## quality                       0.014    -0.129             -0.051
##                       total.sulfur.dioxide density     pH sulphates alcohol
## fixed.acidity                       -0.113   0.668 -0.683     0.183  -0.062
## volatile.acidity                     0.076   0.022  0.235    -0.261  -0.202
## citric.acid                          0.036   0.365 -0.542     0.313   0.110
## residual.sugar                       0.203   0.355 -0.086     0.006   0.042
## chlorides                            0.047   0.201 -0.265     0.371  -0.221
## free.sulfur.dioxide                  0.668  -0.022  0.070     0.052  -0.069
## total.sulfur.dioxide                 1.000   0.071 -0.066     0.043  -0.206
## density                              0.071   1.000 -0.342     0.149  -0.496
## pH                                  -0.066  -0.342  1.000    -0.197   0.206
## sulphates                            0.043   0.149 -0.197     1.000   0.094
## alcohol                             -0.206  -0.496  0.206     0.094   1.000
## quality                             -0.185  -0.175 -0.058     0.251   0.476
##                       quality
## fixed.acidity           0.124
## volatile.acidity       -0.391
## citric.acid             0.226
## residual.sugar          0.014
## chlorides              -0.129
## free.sulfur.dioxide    -0.051
## total.sulfur.dioxide   -0.185
## density                -0.175
## pH                     -0.058
## sulphates               0.251
## alcohol                 0.476
## quality                 1.000
```

I created the correlation table to have a closer look of any chances of multi collinearity. In addition, I am curious to see if there is one particular var iable that stands out that have strong correlation with the quality. Out of a ll the variables, it seems alcohol percentage has the strongest position corr elation out of all. And density has the strongest negative correlation with a lcohol percentage.



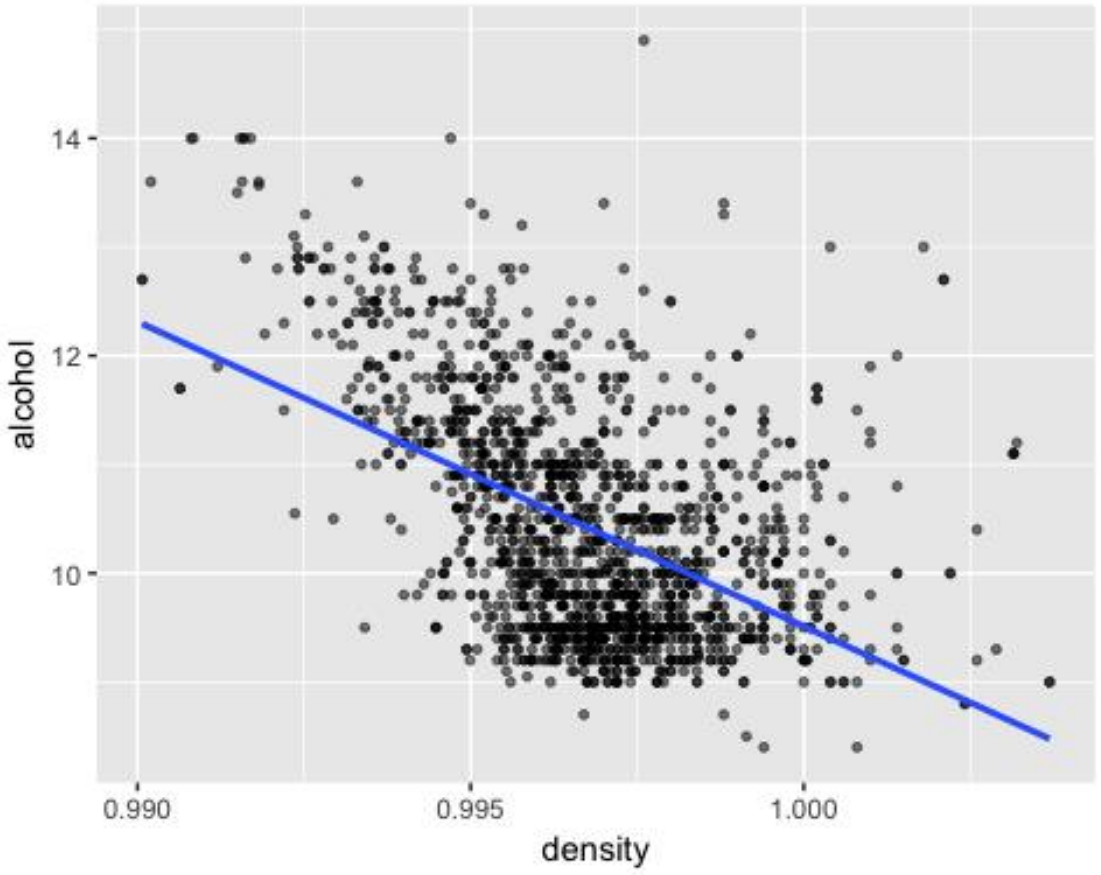
In the ggpair correlation graph, I excluded the facotrs that I am not interested in exploring furture, since the correlation between those variables are not high enough with alcohol percentage nor hte wine quality for me to investigate further. I consider correlation of less than 0.2 as not high enough of correlation.

```
I created a plot version of correlation for visualization purposes.
```



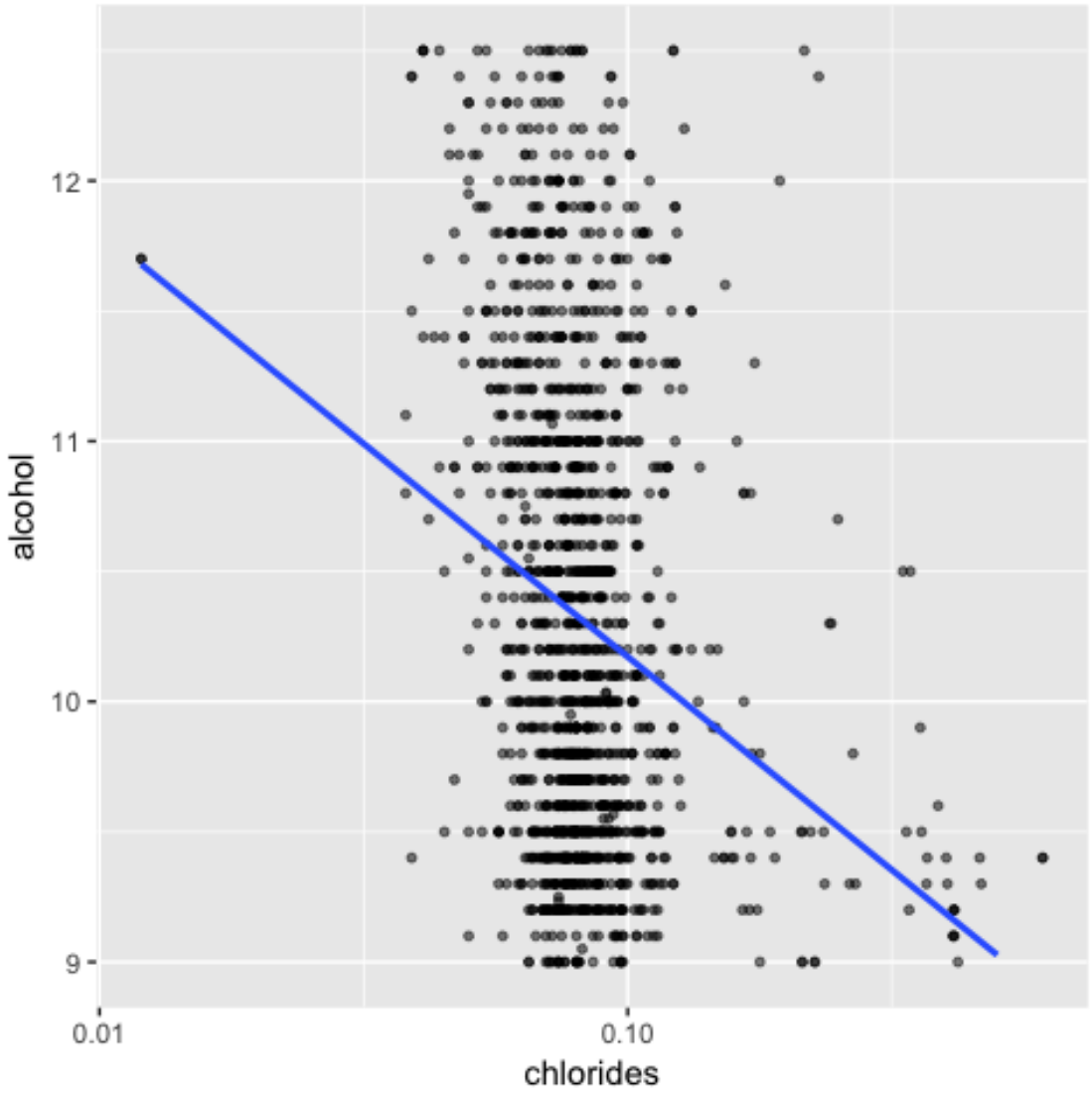

```
##
##  Pearson's product-moment correlation
##
## data:  wine$pH and wine$alcohol
## t = 8.397, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1582061 0.2521123
## sample estimates:
##       cor
## 0.2056325
```

Both the plot and correlation test suggest that pH and alcohol concentrate has little correlation.



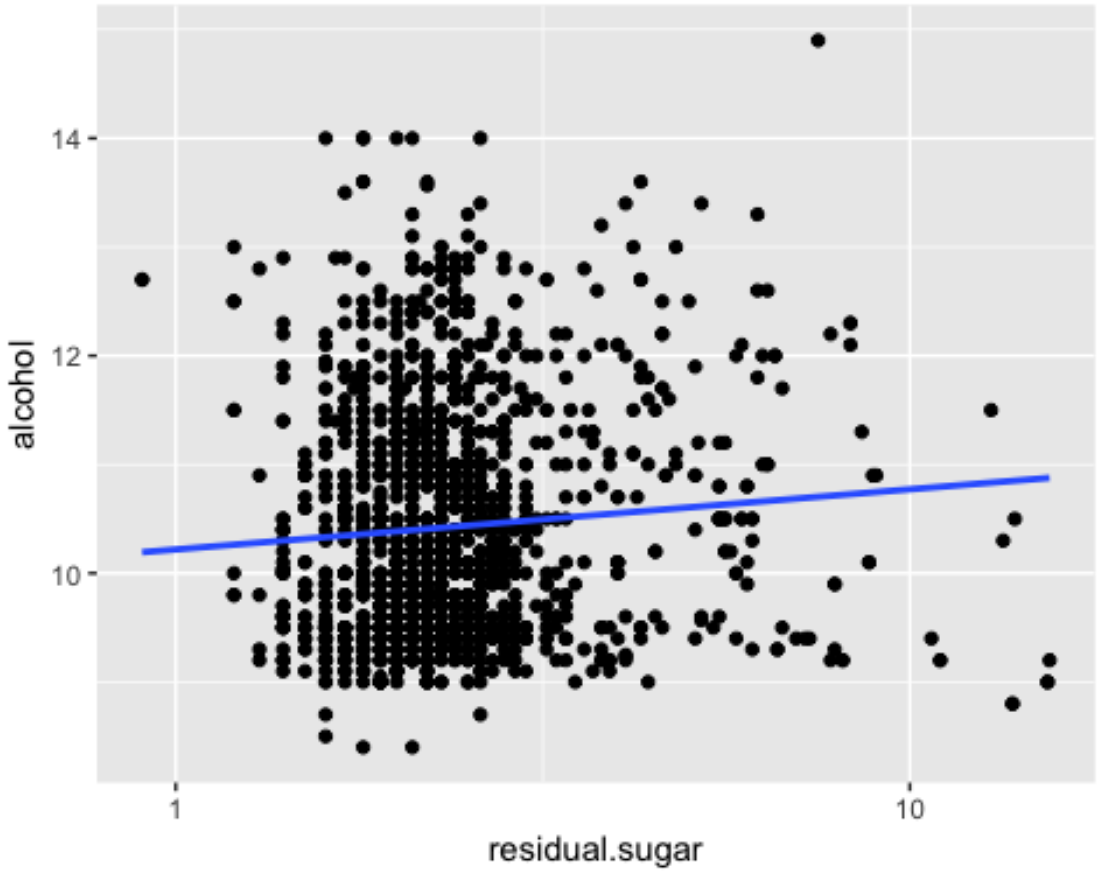

```
##
##  Pearson's product-moment correlation
##
## data:  wine$density and wine$alcohol
## t = -22.838, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5322547 -0.4583061
## sample estimates:
##       cor
## -0.4961798
```

Both the plot and correlation test suggest that there is a moderate correlation between density and alcohol percentage.

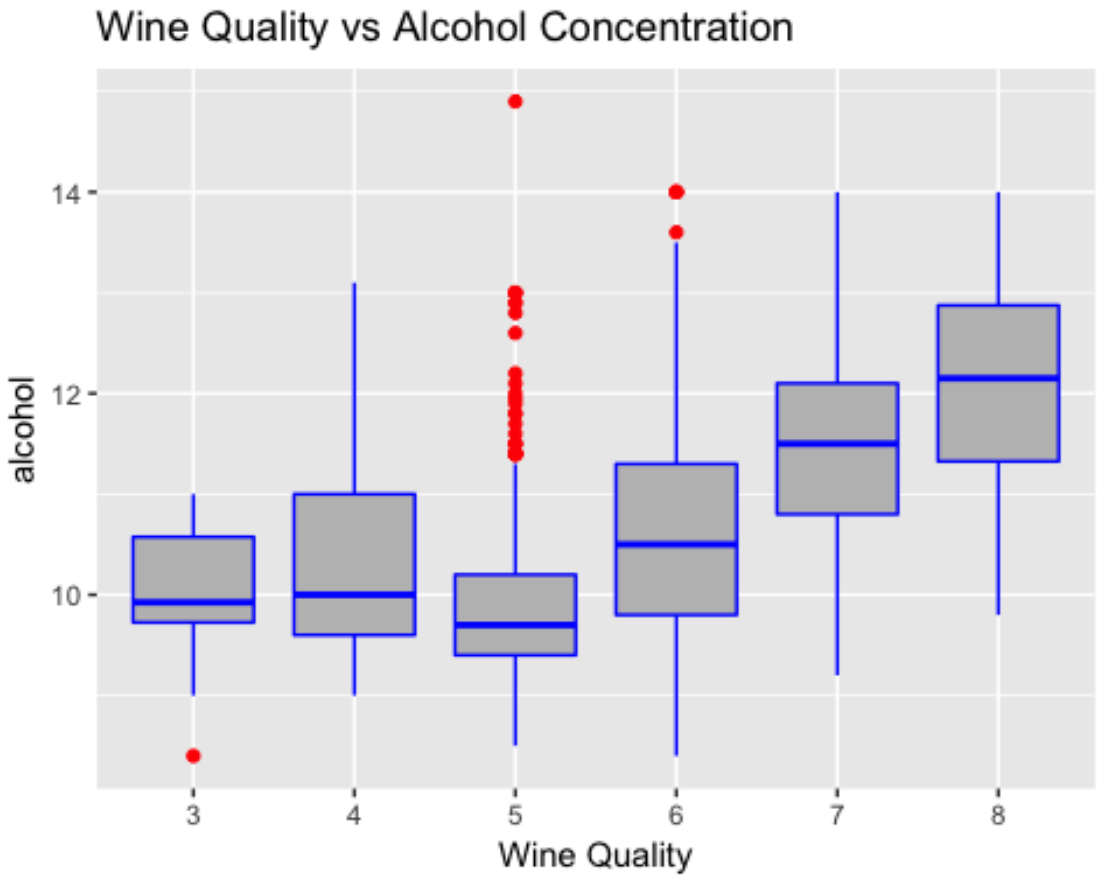


From the graph, we can see one outlier on the left that skewed our result.

```
##
##  Pearson's product-moment correlation
##
## data:  wine$chlorides and wine$alcohol
## t = -9.0617, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2672644 -0.1740057
## sample estimates:
##        cor
## -0.2211405

The plot and correlation test here suggest that chlorides has weak correlatio
n with alcohol concentration.
```



Resudial sugar has little correlation with alcohol concentration.




```
However, there is clearly a relationship between wine quality and alcohol con
centration. This further convinces me that alcohol concentration has strong i
nfluence on wine quality, and exploring what influences alcohol concentration
would lead me to the path of finding real quality wine.

I am going to create a model with just the variables that I believe are criti
cal to the prediction.

##
## Calls:
## m1: lm(formula = I(quality) ~ I(alcohol), data = wine)
## m2: lm(formula = I(quality) ~ I(alcohol) + volatile.acidity, data = wine)
## m3: lm(formula = I(quality) ~ I(alcohol) + volatile.acidity + density,
##     data = wine)
## m4: lm(formula = I(quality) ~ I(alcohol) + volatile.acidity + density +
##     chlorides, data = wine)
## m5: lm(formula = I(quality) ~ I(alcohol) + volatile.acidity + density +
##     chlorides + pH, data = wine)
##
## ============================================================================
====
##                         m1         m2          m3          m4          m5
## ----------------------------------------------------------------------------
```

```
----
##   (Intercept)        1.875***   3.095***  -18.407     -19.637      -8.811
##                     (0.175)    (0.184)    (10.298)    (10.352)    (10.747)
##   I(alcohol)         0.361***   0.314***    0.333***    0.330***    0.336*
**
##                     (0.017)    (0.016)     (0.018)     (0.019)     (0.019)
##   volatile.acidity             -1.384***   -1.365***   -1.362***   -1.260*
**
##                                (0.095)     (0.096)     (0.096)     (0.099)
##   density                                   21.360*     22.660*     13.173
##                                            (10.228)    (10.289)    (10.588)
##   chlorides                                             -0.422      -0.725
##                                                         (0.366)     (0.374)
##   pH                                                                -0.439*
**
##                                                                     (0.122)
## ------------------------------------------------------------------------------
----
##   R-squared             0.2        0.3        0.3         0.3         0.3
##   adj. R-squared        0.2        0.3        0.3         0.3         0.3
##   sigma                 0.7        0.7        0.7         0.7         0.7
##   F                   468.3      370.4      248.9       187.0       153.3
##   p                     0.0        0.0        0.0         0.0         0.0
##   Log-likelihood    -1721.1    -1621.8    -1619.6     -1619.0     -1612.6
##   Deviance            805.9      711.8      709.9       709.3       703.6
##   AIC                3448.1     3251.6     3249.3      3249.9      3239.1
##   BIC                3464.2     3273.1     3276.1      3282.2      3276.7
##   N                  1599       1599       1599        1599        1599
## ==============================================================================
====
```

The formula here for the quality of wine is -8.811 + (0.336)(alcohol) - (1.260)(volatile acidity) + 12.173 (density) -0.725(chlorides) - 0.439 (pH) We have a R-squared of 0.3, which means that the data is only moderate at fitting the regression line. However, the R-squared increased from the first update, meaning the model improved.

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?

```
After experimenting with some variables to see if they have any impact with a
lcohol concentration, I discovered that residual sugar has weak correlation w
ith the concentration of alcohol. On top of that, pH turned out to have weak
correlation but on the plot the points all resides within 3.0 to 3.5, which i
s acidic. However, density comes into play, the higher the alcohol concentrat
ion, the less the density.
```

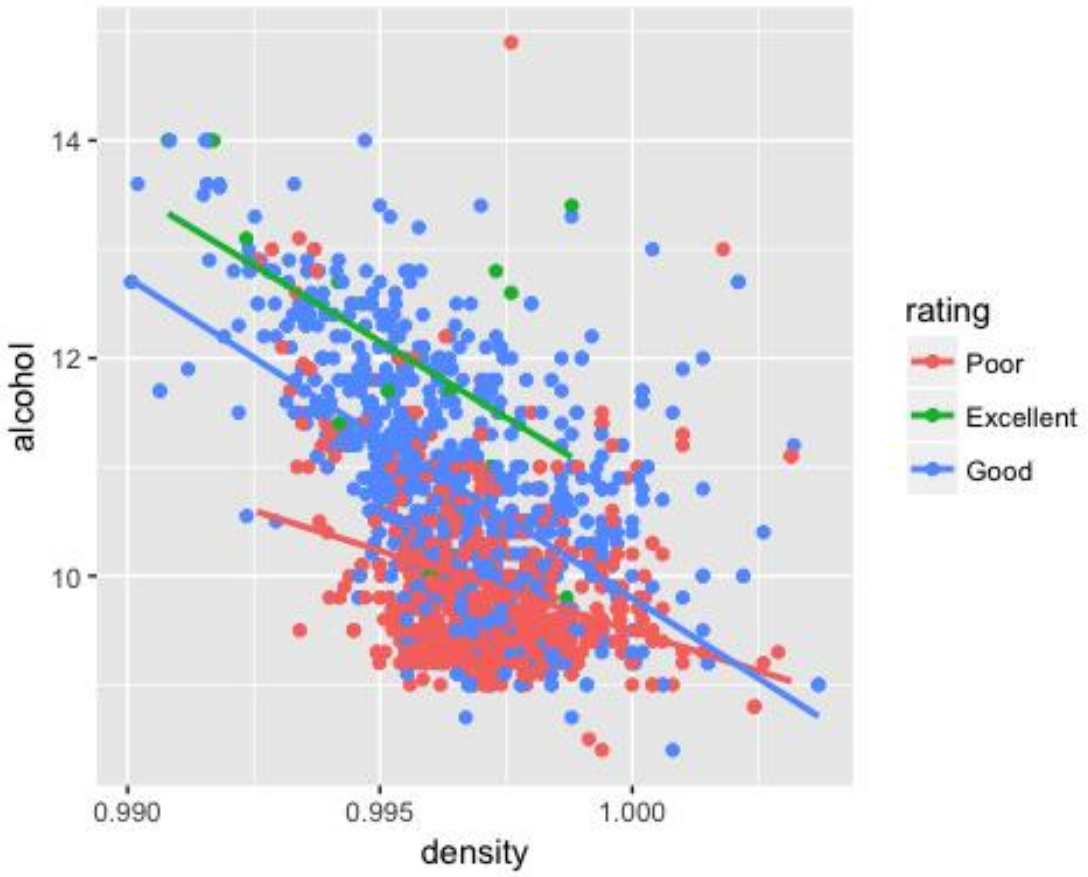
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

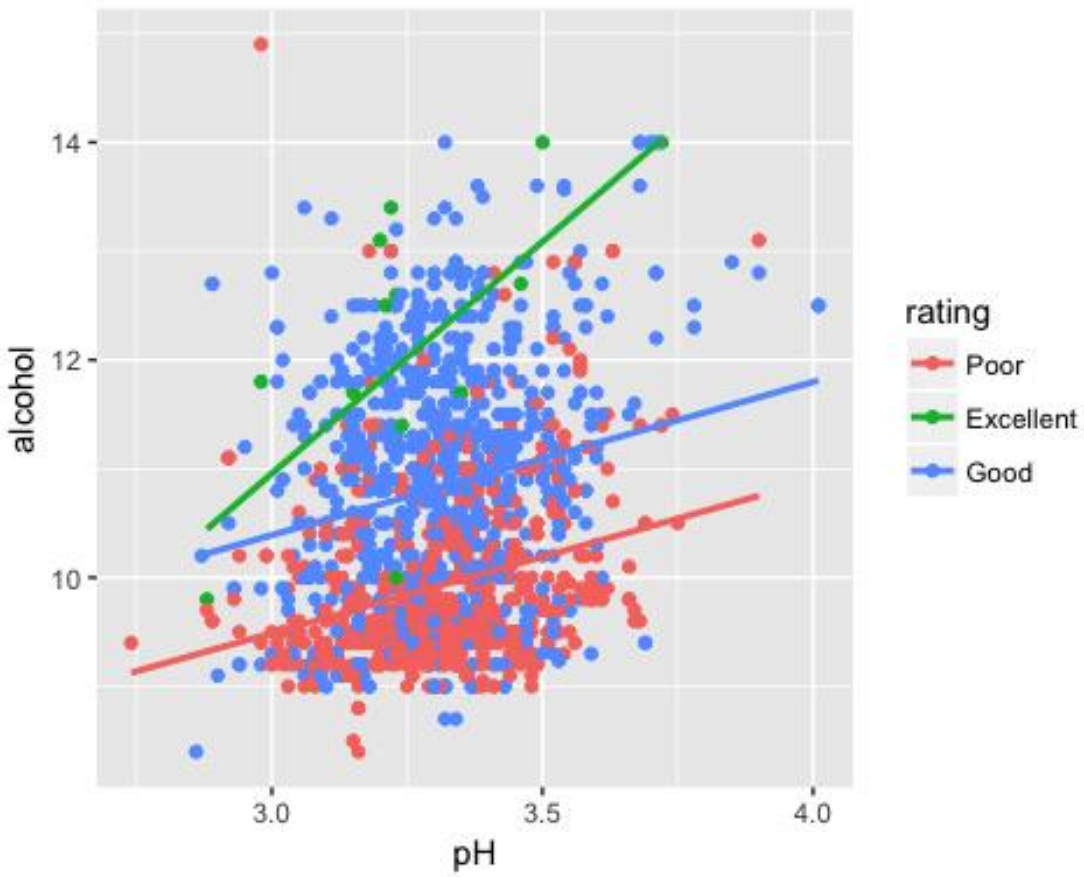
I observed that density have a stronger effect on alcohol concentration.

What was the strongest relationship you found?

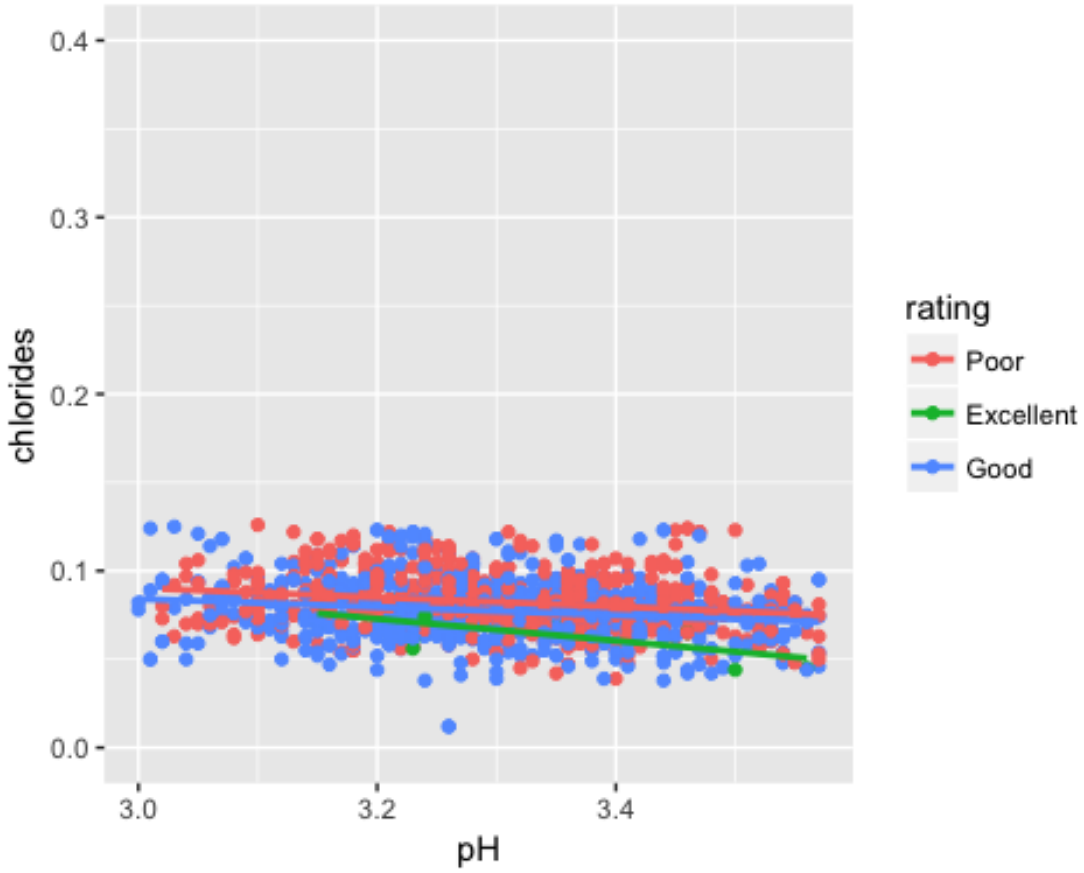
The alcohol concentration is not particularly strongly associated with any of the variables.
Which makes me wonder if wine quality has more of a direct association with alcohol concentration.

Multivariate Plots Section





As the plots above suggest, on top of the relationship between alcohol and ot
her variables, the higher the rating the higher the alcohol concentration.



Despite the almost none existant relationship between pH and chlorides, we can see that the lower the chlorides, the better the wine quality.

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?

Adding ratings into the relationship, I can see that ratings indeed is a major factor in how some factors correlate. For example, previously we did not see much of a strong correlation with density and alcohol concentration, but with the rating, we can see that the relationship in negative.

Were there any interesting or surprising interactions between features?

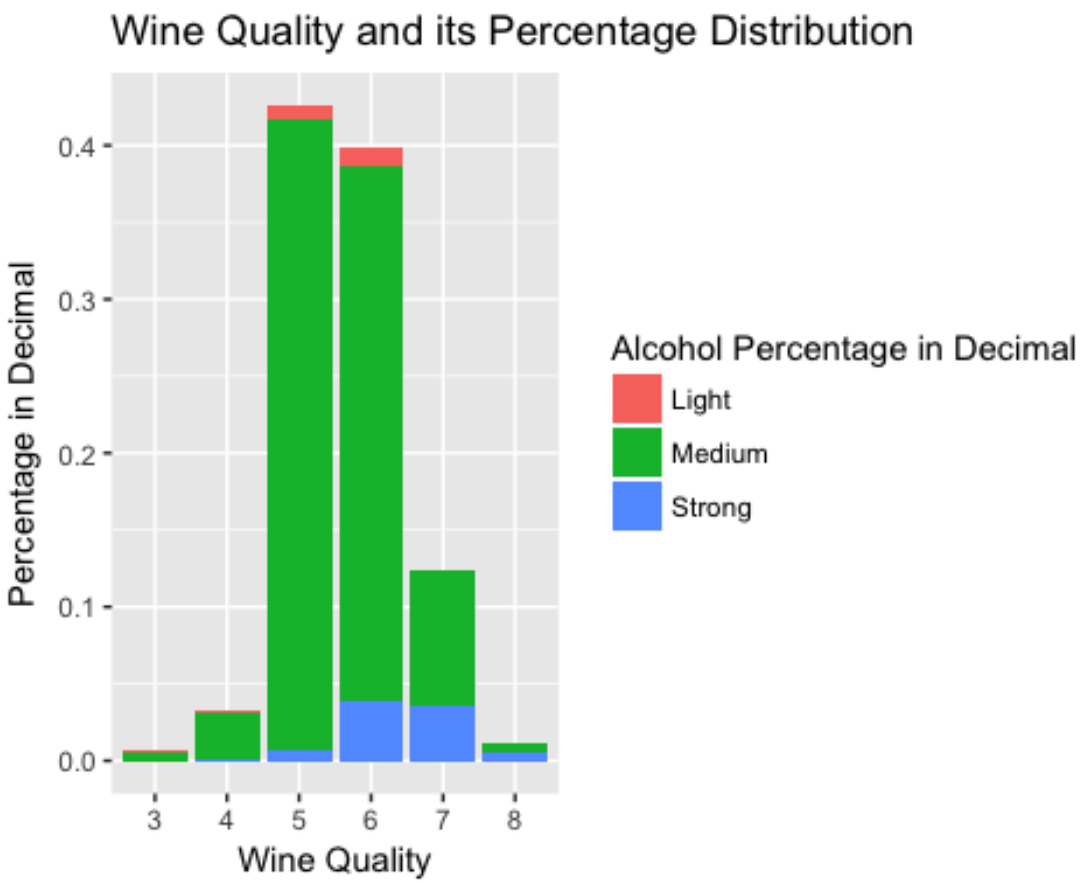
Nothing particular surprising, but if I have to point out, it would be the fact that ratings changes how we view the relationship between the variables.

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

I created model for my dataset. Although at a glance it seems useful, there might be cases of multicollinearity.
The strengh is that we can discover the correlations between the variables that we deem are important.

Final Plots and Summary

Plot One



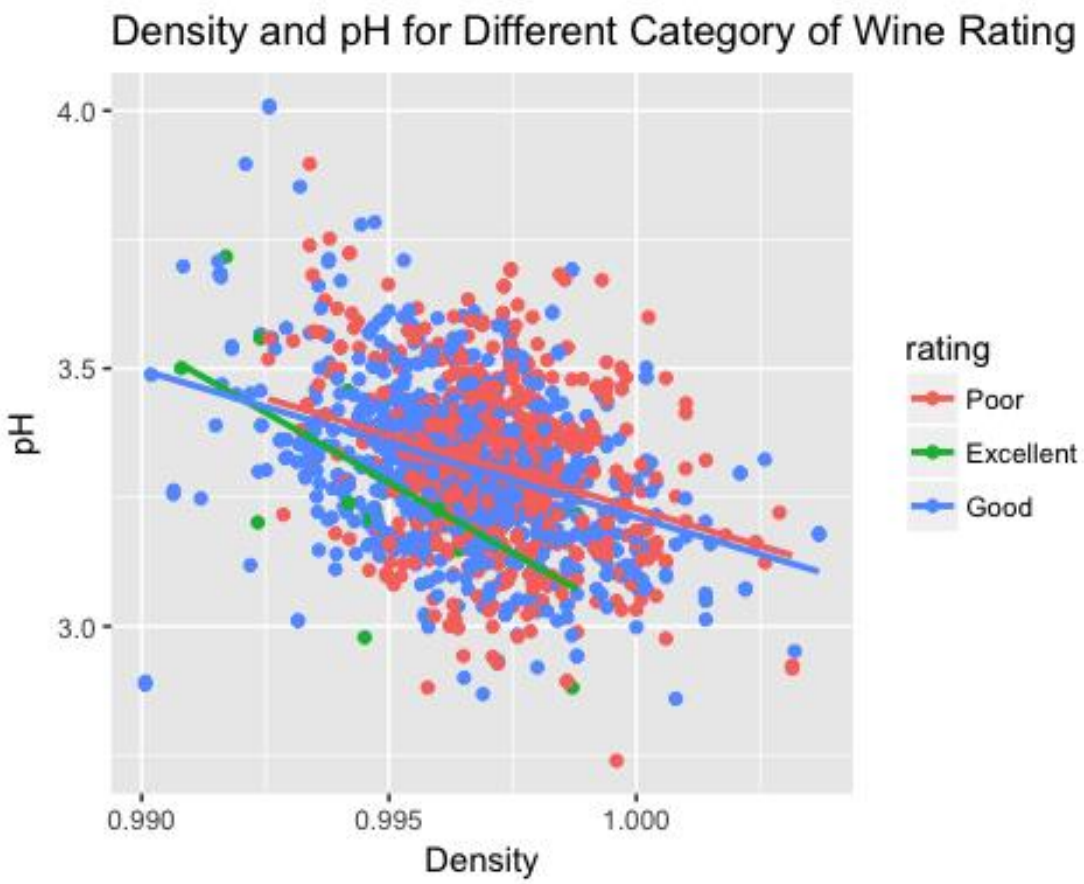

```
## wine$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.400   9.725   9.925   9.955  10.580  11.000
## ------------------------------------------------------------
## wine$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00    9.60   10.00   10.27   11.00   13.10
## ------------------------------------------------------------
```

```
## wine$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     8.5     9.4     9.7     9.9    10.2    14.9
## ------------------------------------------------------------
## wine$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.80   10.50   10.63   11.30   14.00
## ------------------------------------------------------------
## wine$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.20   10.80   11.50   11.47   12.10   14.00
## ------------------------------------------------------------
## wine$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.80   11.32   12.15   12.09   12.88   14.00
```

Description One

```
I created a new variable 'label' to show how alcohol percentage and wine qual
ity vary. We can see that medium alchol is the majority.
```

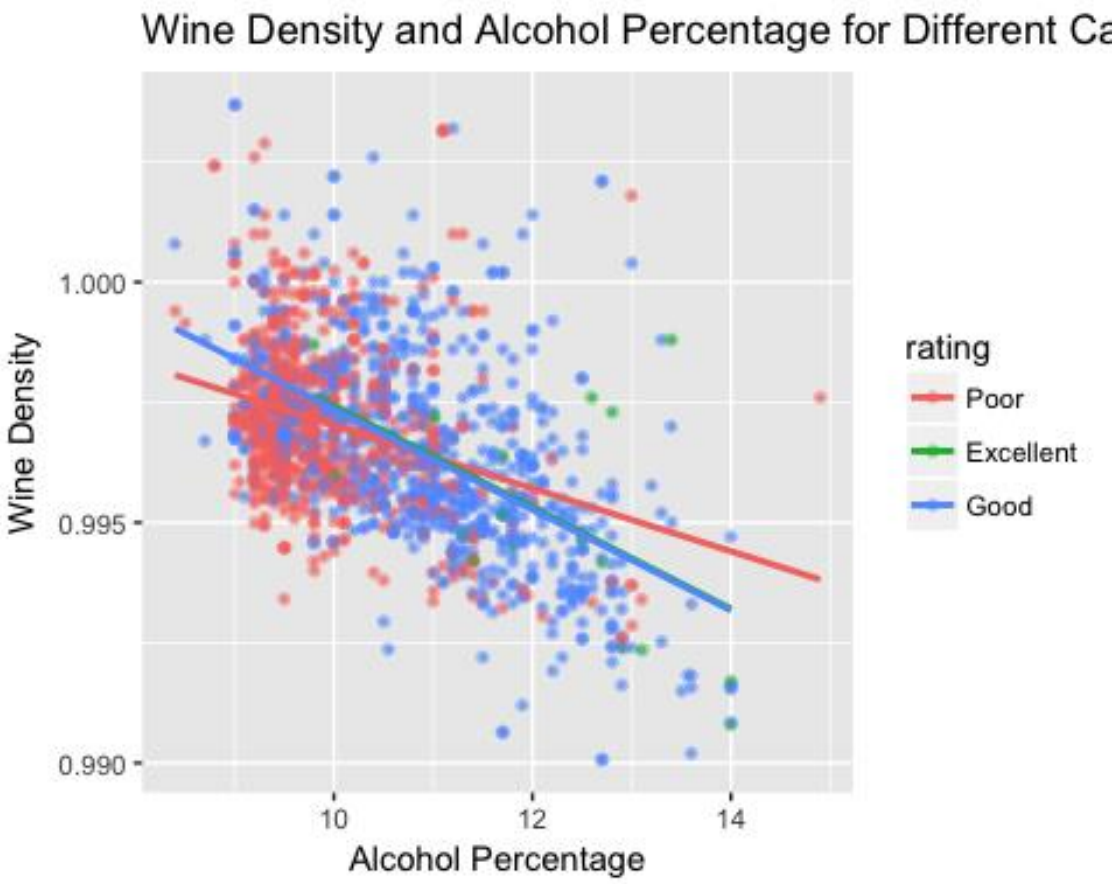
Plot Two



Description Two

From the scatter plot distribution, there is a negative relationship between density and pH. The lines are grown to see the ratings. The correlation we id entified earlier suggested a relatively moderate relationship which is -0.342 . The graph here also suggests that the lower the pH with density be between approximately 0.99 and 1, the rating of the alcohol would remain excellent.

Plot Three



Description Three

The distribution of Alcohol Percentage and Wine Density is strong. The higher the alcohol percenrage, the lower is the density. We can also see in this plo t that stronger wine tend to have higher rating.

Reflection

Based on the analysis I did for the dataset, I am convinced that alcohol conc entrate is the most important factor to deciding the quality of Red Wine woul d be density. The lower the density, the higher the alcohol concentration, an d the higher the alcohol concentration the better the quality of wine. One of

the challenges I encountered however, is that although I like wine, I do not know the chemistry behind it. It was a little tough for me to wrap my mind around what might be the most important component in making a quality wine. After playing around with the variables and creating the plots, the results eventually made sense to me. I wish I could enhence my analysis by knowing which brands that are highly rated by consumers fit my prediction. This analysis serves as a rough idea of what makes a good wine and the audience can just go from there.