Wine White Quanlity Analysis by Zhenning Tan 4/6/2016

Basic statistics of the dataset

```
##        X        fixed.acidity    volatile.acidity  citric.acid
##  Min.   :   1   Min.   : 3.800   Min.   :0.0800   Min.   :0.0000
##  1st Qu.:1225   1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700
##  Median :2450   Median : 6.800   Median :0.2600   Median :0.3200
##  Mean   :2450   Mean   : 6.855   Mean   :0.2782   Mean   :0.3342
##  3rd Qu.:3674   3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900
##  Max.   :4898   Max.   :14.200   Max.   :1.1000   Max.   :1.6600
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.600   Min.   :0.00900   Min.   :  2.00
##  1st Qu.: 1.700   1st Qu.:0.03600   1st Qu.: 23.00
##  Median : 5.200   Median :0.04300   Median : 34.00
##  Mean   : 6.391   Mean   :0.04577   Mean   : 35.31
##  3rd Qu.: 9.900   3rd Qu.:0.05000   3rd Qu.: 46.00
##  Max.   :65.800   Max.   :0.34600   Max.   :289.00
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  9.0        Min.   :0.9871   Min.   :2.720   Min.   :0.2200
##  1st Qu.:108.0        1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100
##  Median :134.0        Median :0.9937   Median :3.180   Median :0.4700
##  Mean   :138.4        Mean   :0.9940   Mean   :3.188   Mean   :0.4898
##  3rd Qu.:167.0        3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500
##  Max.   :440.0        Max.   :1.0390   Max.   :3.820   Max.   :1.0800
##     alcohol         quality
##  Min.   : 8.00   Min.   :3.000
##  1st Qu.: 9.50   1st Qu.:5.000
##  Median :10.40   Median :6.000
##  Mean   :10.51   Mean   :5.878
##  3rd Qu.:11.40   3rd Qu.:6.000
##  Max.   :14.20   Max.   :9.000
```

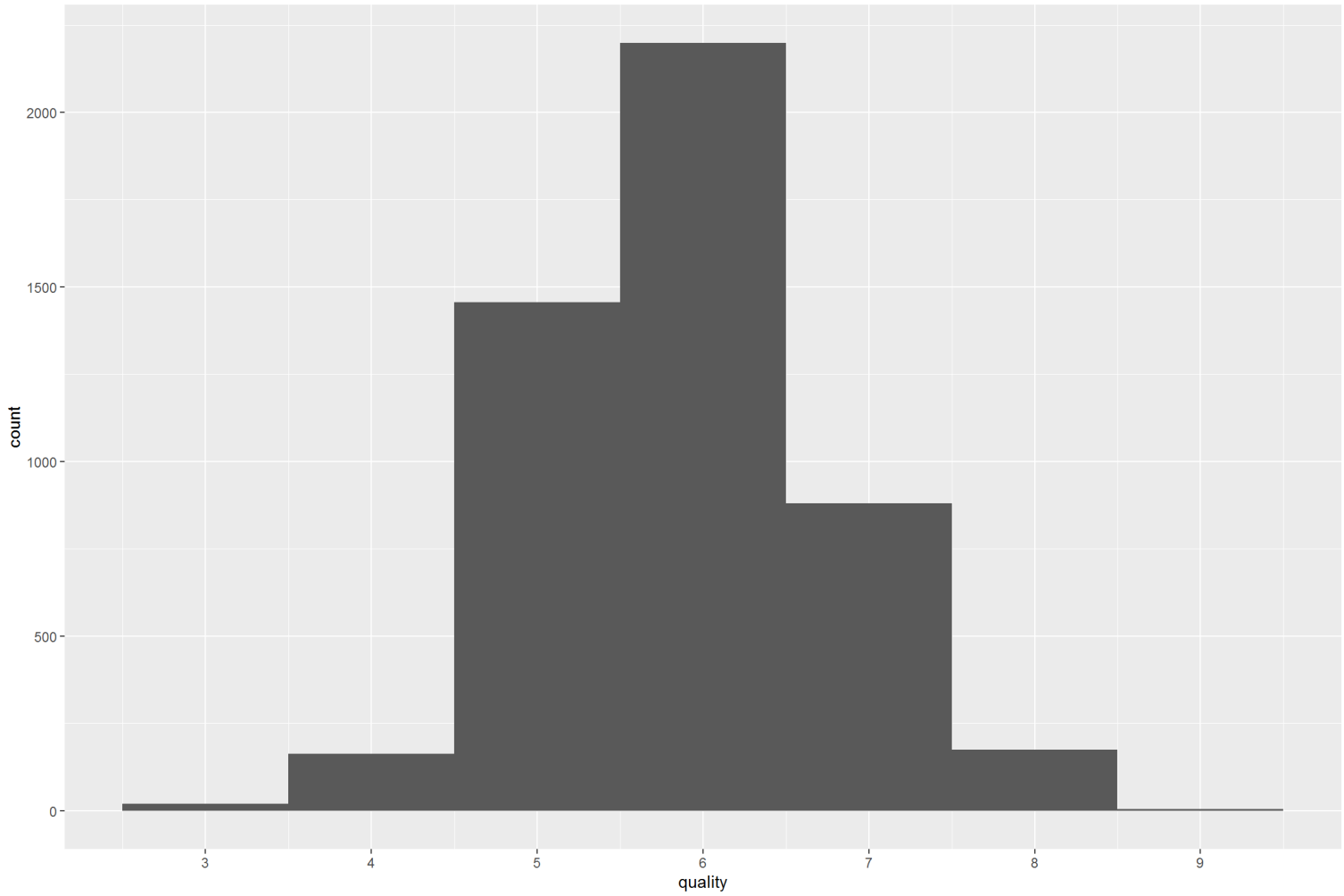
```
## 'data.frame':    4898 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  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 ...
```

This dataset contains 4898 observations of 13 variables.The quality variable is integer. All other factors of the wine chemical properties are numeric type.

Univariate Plots Section

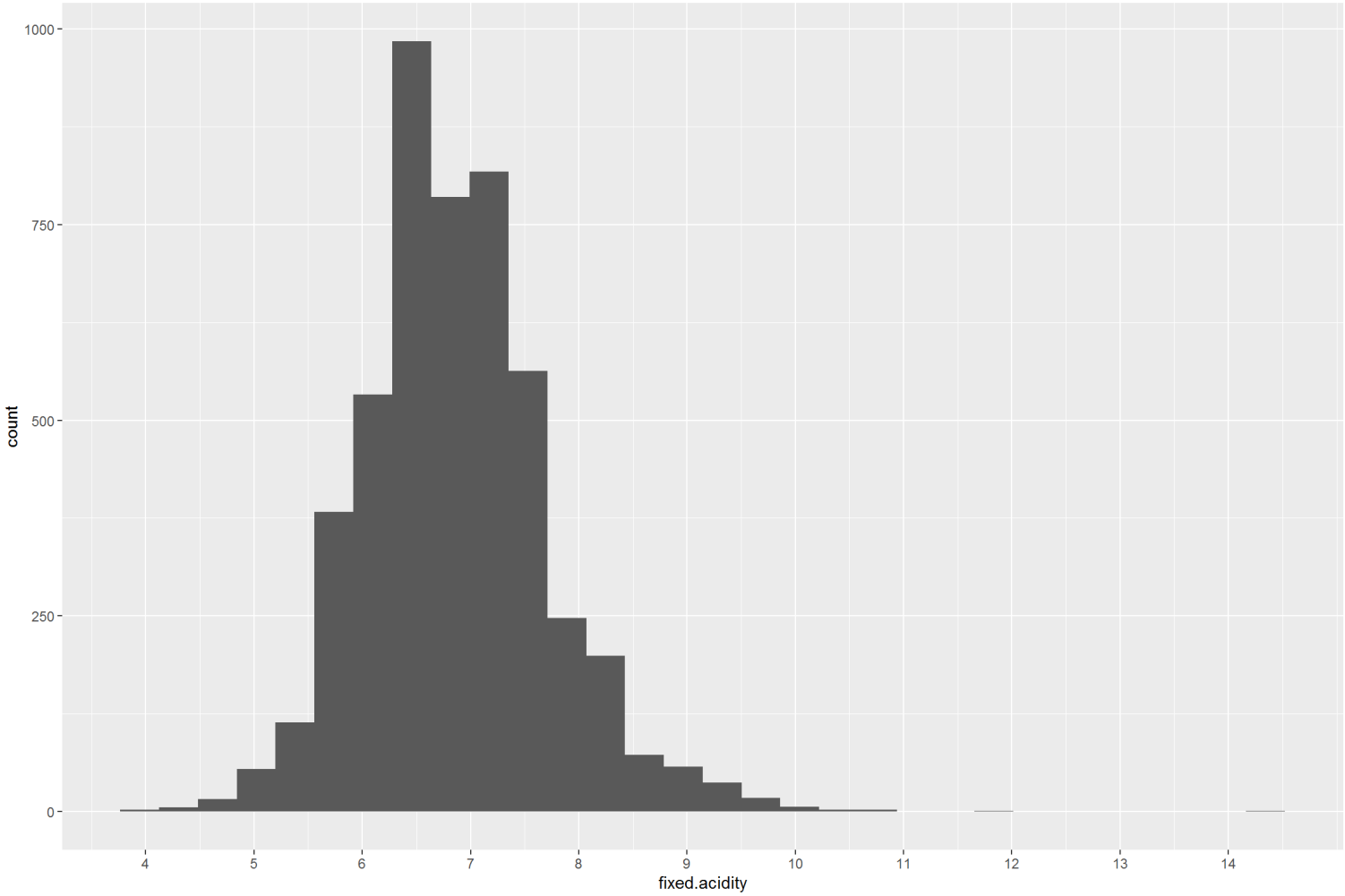
Wine quality should be the dependent variable in this dataset. My goal is to investigate how each chemicals in wine affect its quality. First, I'd like to explore the quality distribution.




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

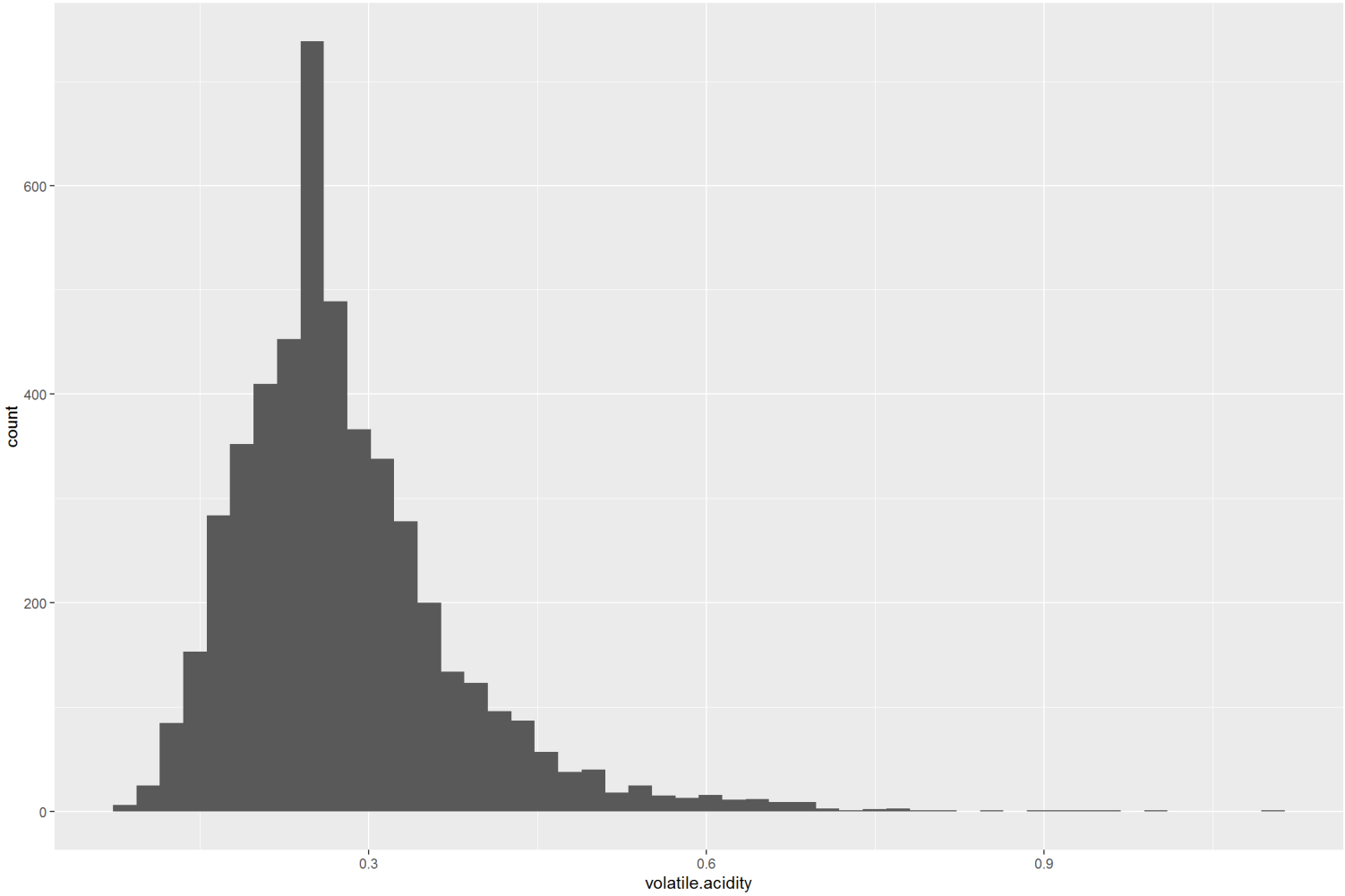
The quality distribution appears normal. The median quality is 6. The mean quality is 5.878.

Next, I'd like to explore the distribution of all the chemicals in white wine



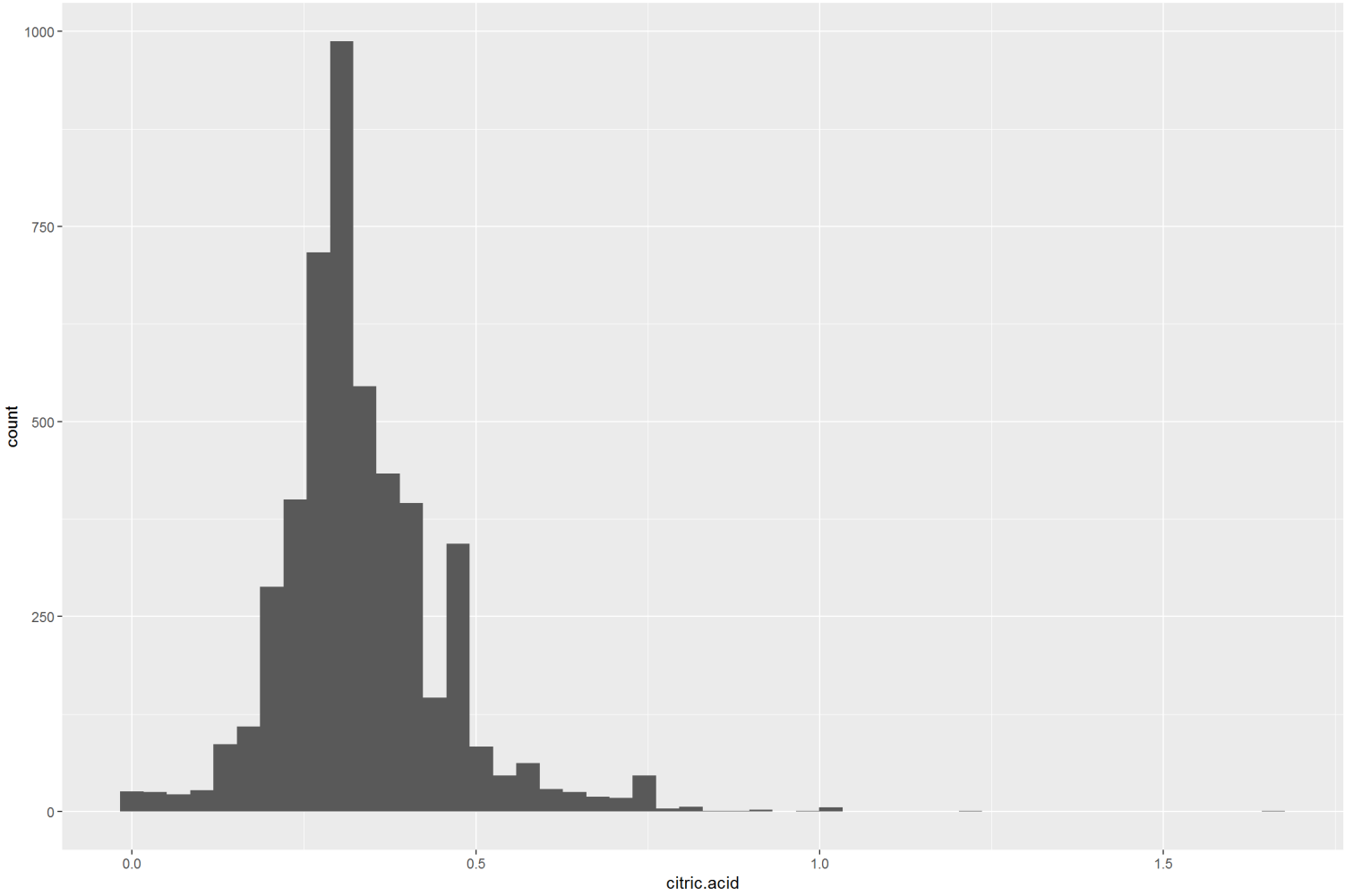

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.800   6.300   6.800   6.855   7.300  14.200
```

Acidity is an important factor in wine quality and contributes to wine taste [1]. It comes from the type of grapes and fermentation process. However, there's no direct correlation of fixed acidity to wine quality. In this plot, fixed acidity appears normal distribution. The median is 6.8. The mean is 6.855. However, there are some outliers with fixed.acidity more than 10.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.2100  0.2600  0.2782  0.3200  1.1000
```

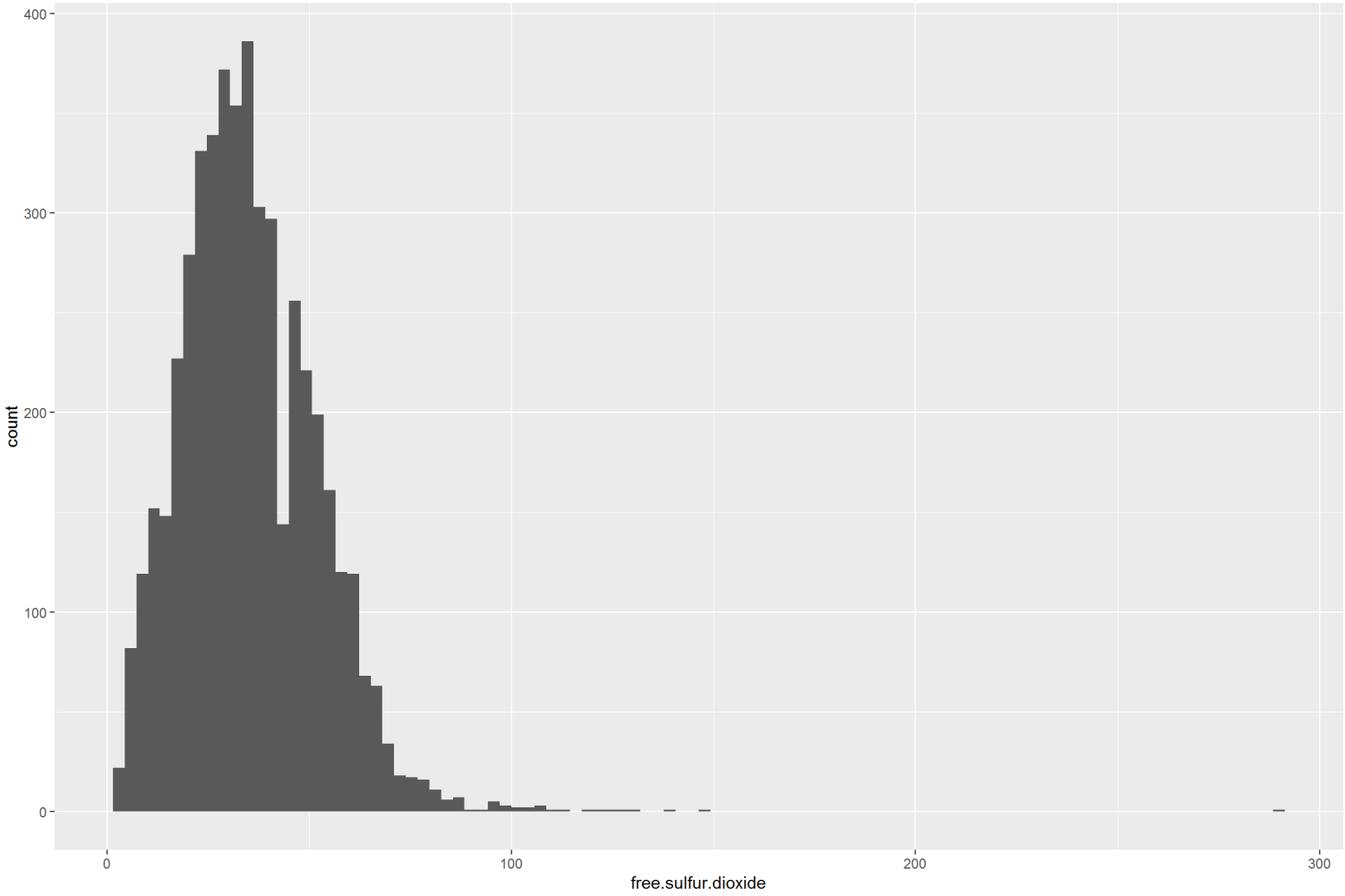
The volatitle acidity comes from acetic acid in wine, a byproduct of bacterial metabolism. The U.S. legal limits of volatile acidity for white table wine is 1.1 g/L [2]. In our dataset, most wine have a volatile acidity between 0.21 and 0.32 g/L. Low volatile acidity can reduce formation of other concomitant, sometimes unpleasant, aroma compounds.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2700  0.3200  0.3342  0.3900  1.6600
```

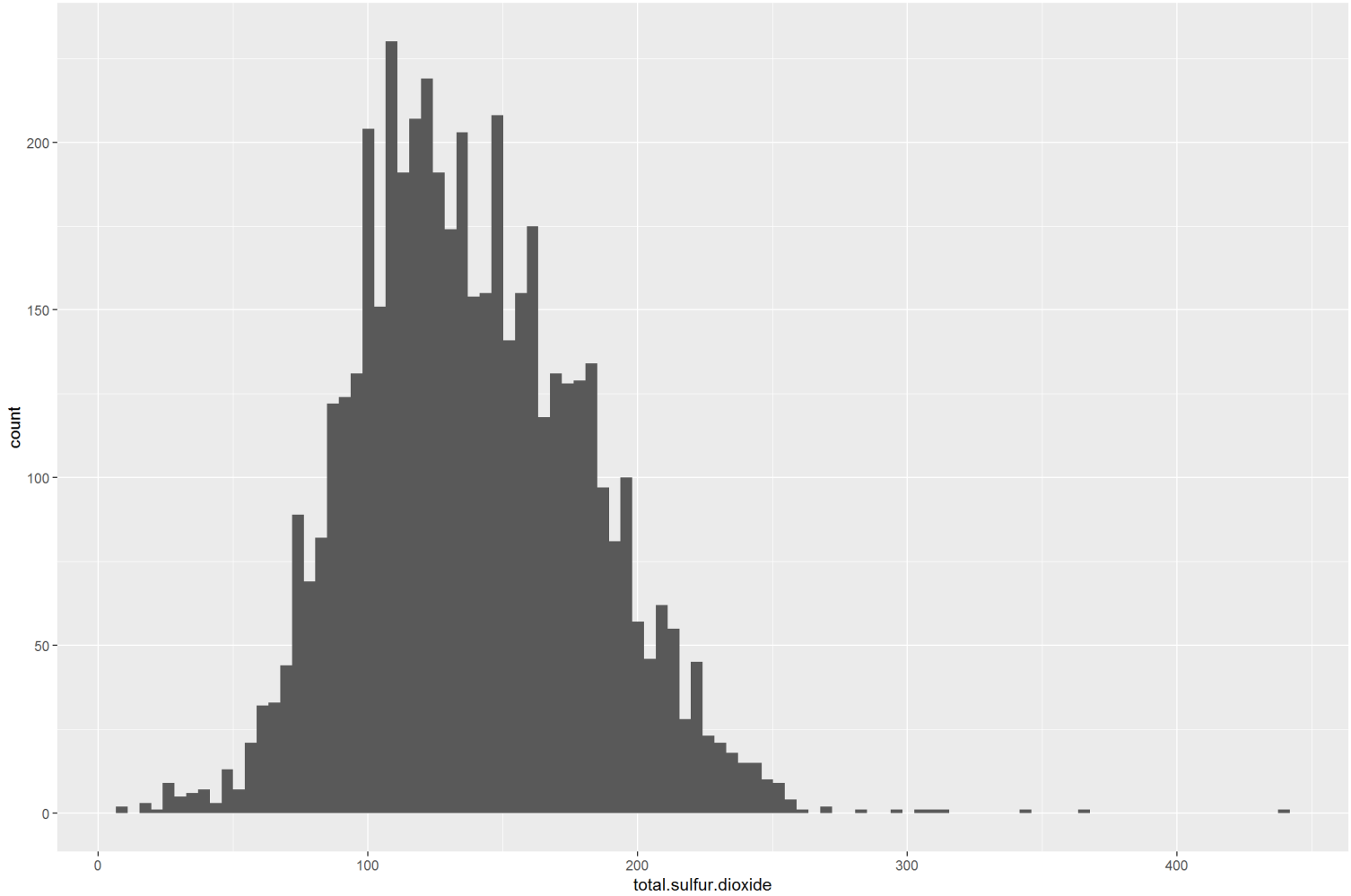
```
##
## FALSE  TRUE
##  4876    22
```

Citric acid contributes to the acidity of wine, adding “freshness” taste to wine. However, as a molecule in energy production reaction, citric acid can also lead to growth of microbes [3]. In our dataset, most wine have a citric acid amount between 0.27 and 0.39. There are 22 outliers with citric acid above 0.75 g/L.



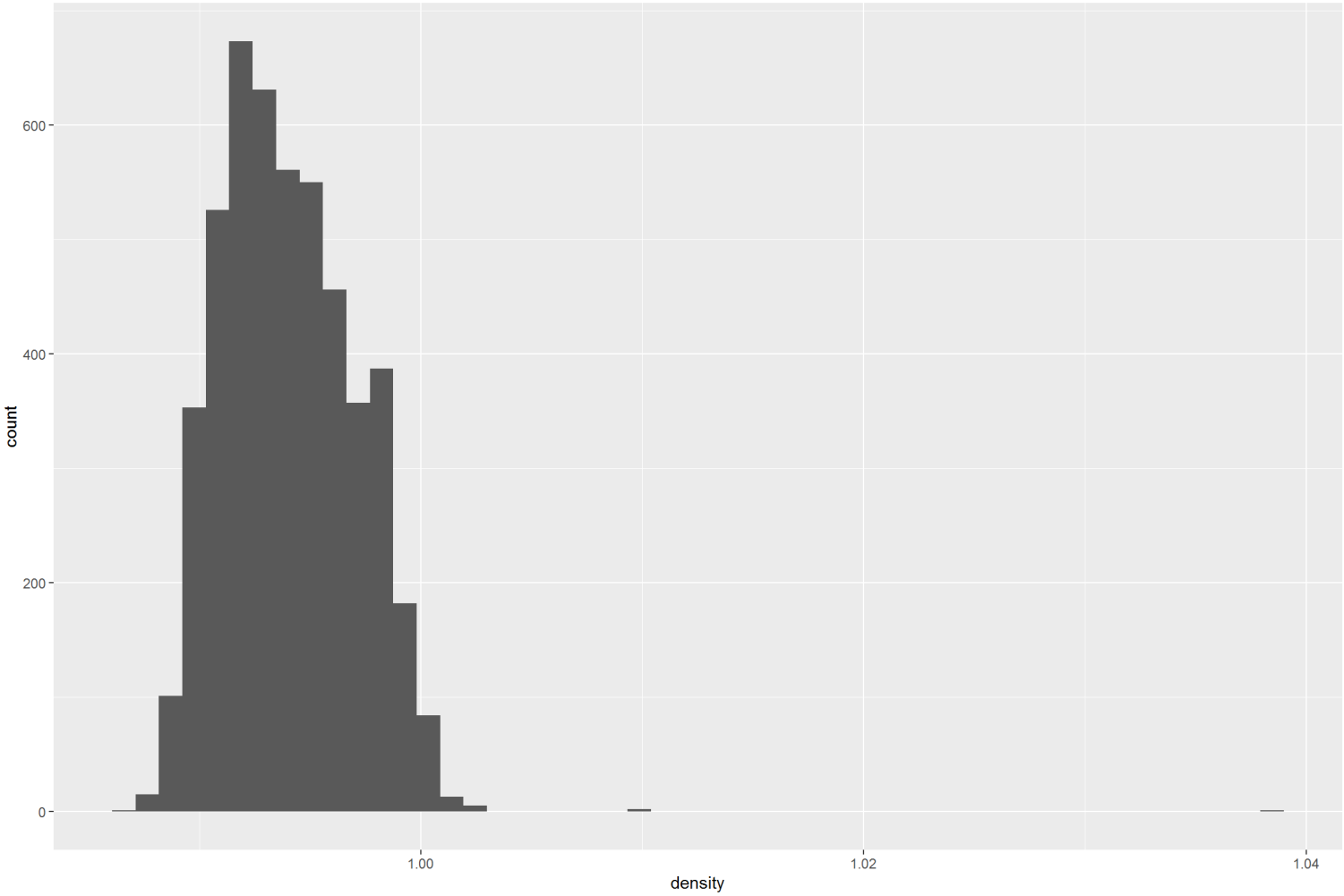

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   23.00   34.00   35.31   46.00  289.00
```

Sulfur dioxide is used to inhibit growth and metabolism of microbes. This can prevent oxidation and preserve the original fruity flavor and freshness taste [4]. In this dataset, most of the free sulfur dioxide is between 23 and 46 ppm. However, there are outliers with



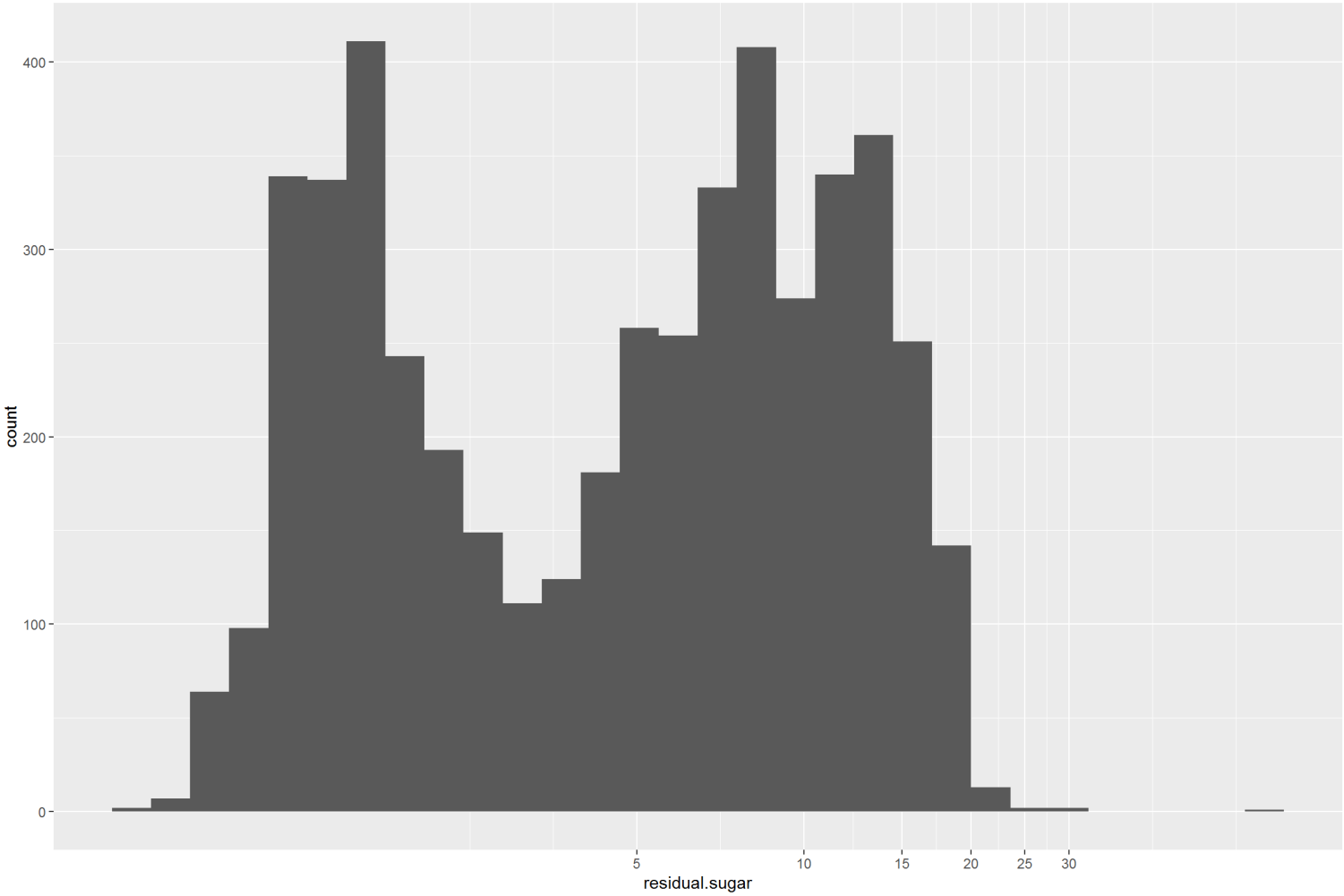

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.0   108.0   134.0   138.4   167.0   440.0
```

Most of the total sulfur dioxide is between 108 and 167 ppm. However, there are outliers with extreme total sulfur dioxide.

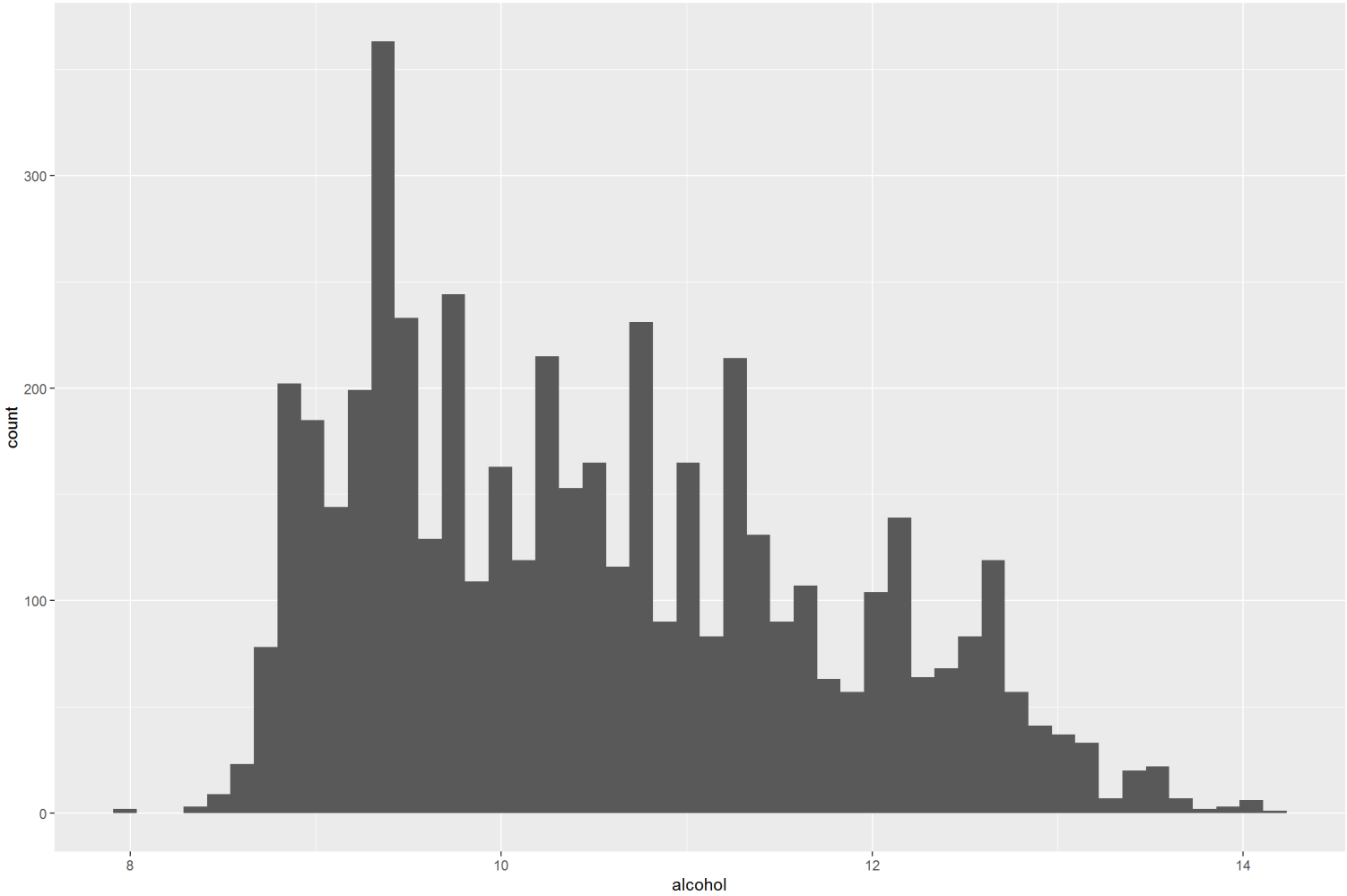



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9917  0.9937  0.9940  0.9961  1.0390
```

Most wines have density between 0.9917 and 0.9961. However, there are outliers with extreme density.



Residual sugar affects the sweetness of wine. Wines have different sweetness depending on the residual sugar. In the plot, I see that the residual sugar appears as bimodal distribution. This two peaks probably correponds to dry wine and sweet wine.



Alcohol content is also a determinant of the wine quality and taste. It has a wide distribution.

Univariate Analysis

What is the structure of your dataset?

This white wine quality dataset contains 4898 observations of 13 variables, which describes chemical properties of the wine, such as acidity, sugar, density, pH, etc.

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

Most of features of the wine in the dataset have similar normal distribution as quality, except residual sugar and alcohol content. I need to run bivariate analysis to determine which features are related to wine quality.

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

At this step, I cannot determine the features that will contribute to the wine quality. Given the small number of variates, I will perform bivariate analysis on all the features and work on the pairs of features with high correlation coefficient.

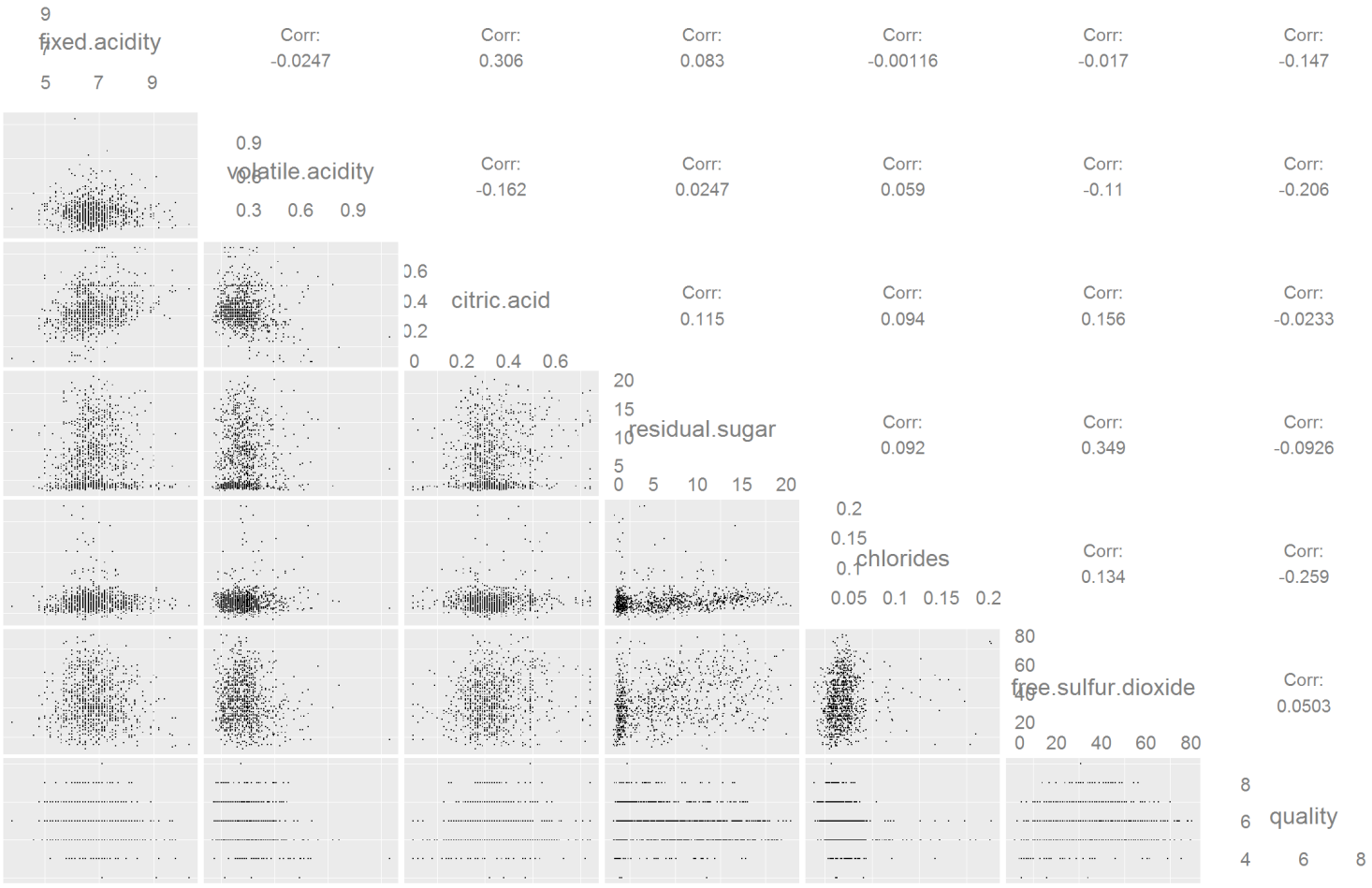
Did you create any new variables from existing variables in the dataset?

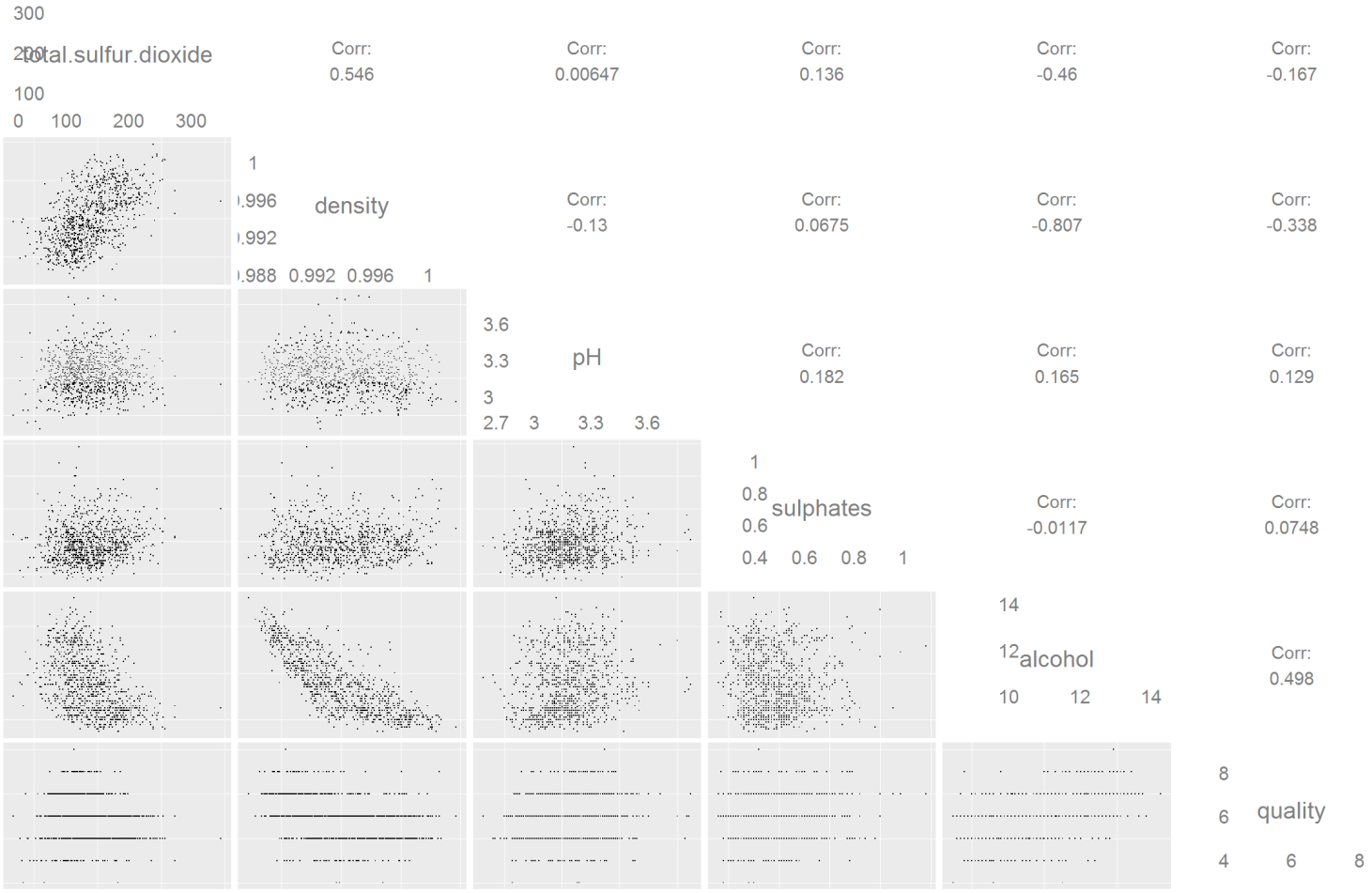
Later in the analysis, I transformed the numeric data type of quality to categorical type.

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?

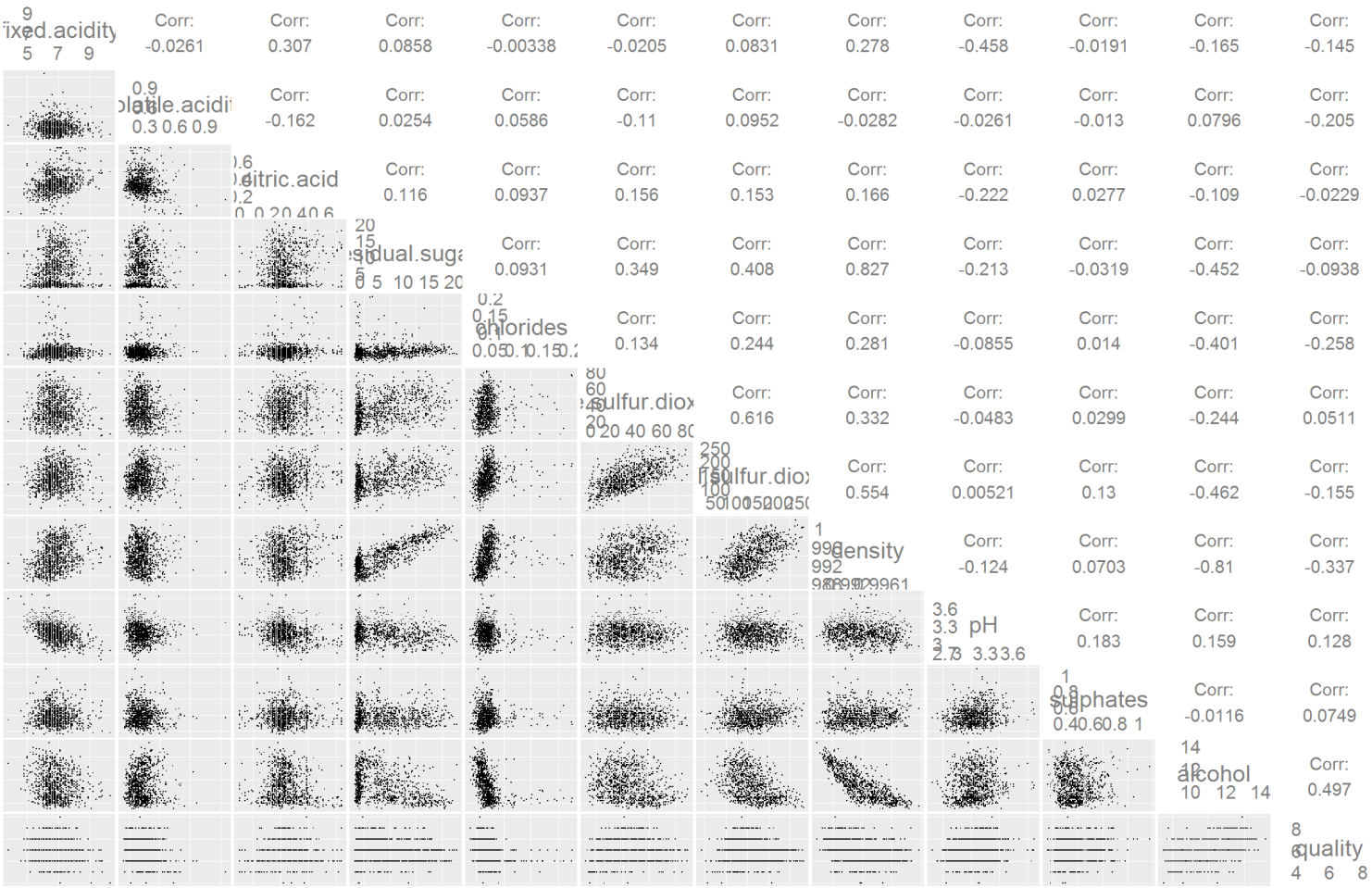
Since the wine quality appears as normal distribution, I would guess that the factors that determines of wine quality will have similar normal distribution. I noticed that residual sugar distribution has two peaks. This is interesting and I will pay more attention in later analysis.

Bivariate Plots Section

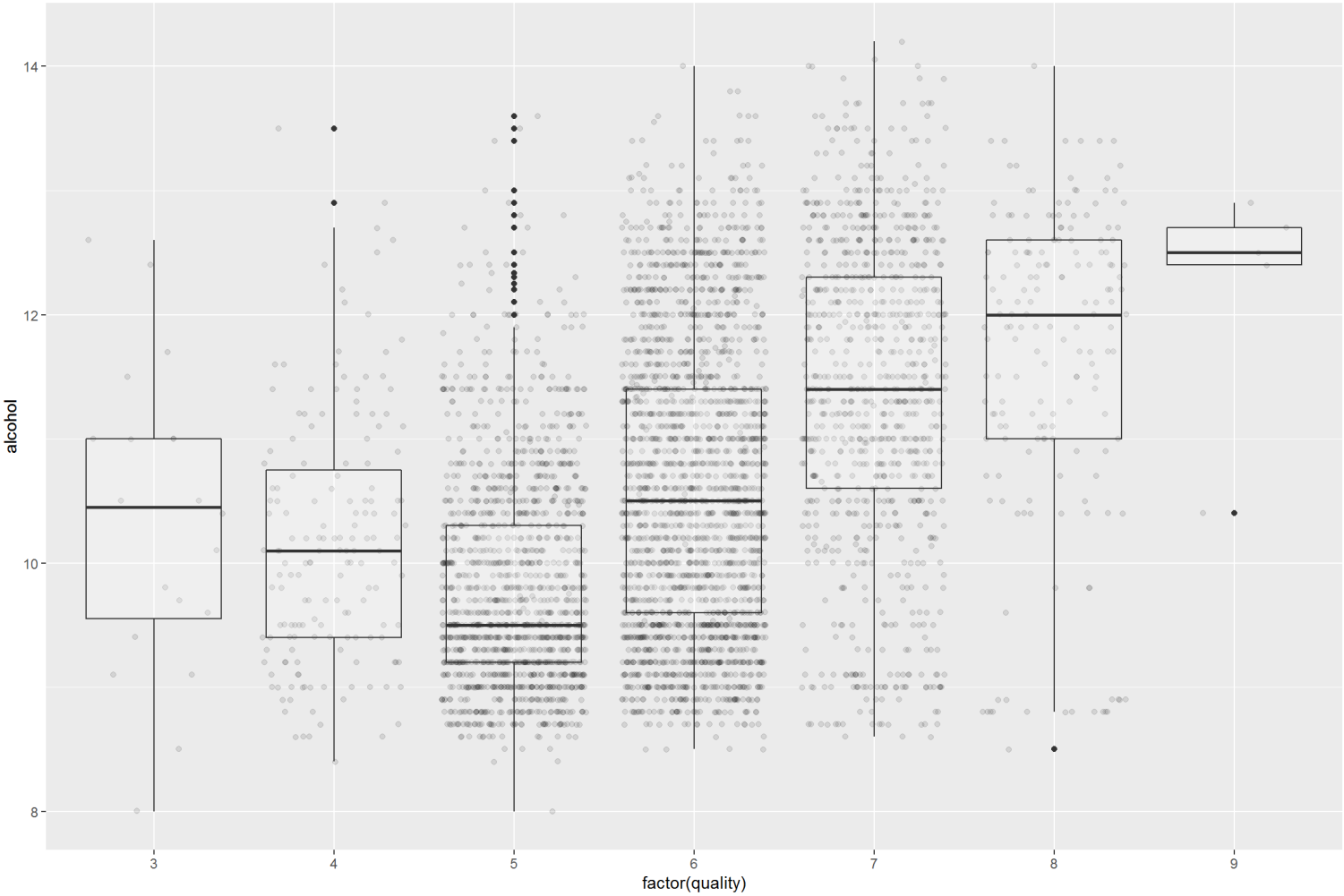




The above ggpair plot give clear visulation of labels by separating the features into two parts. However, in this case, I may miss interesting relationships between features. So I did the ggpair analysis on the whole dataset again to explore possible correlations between features.

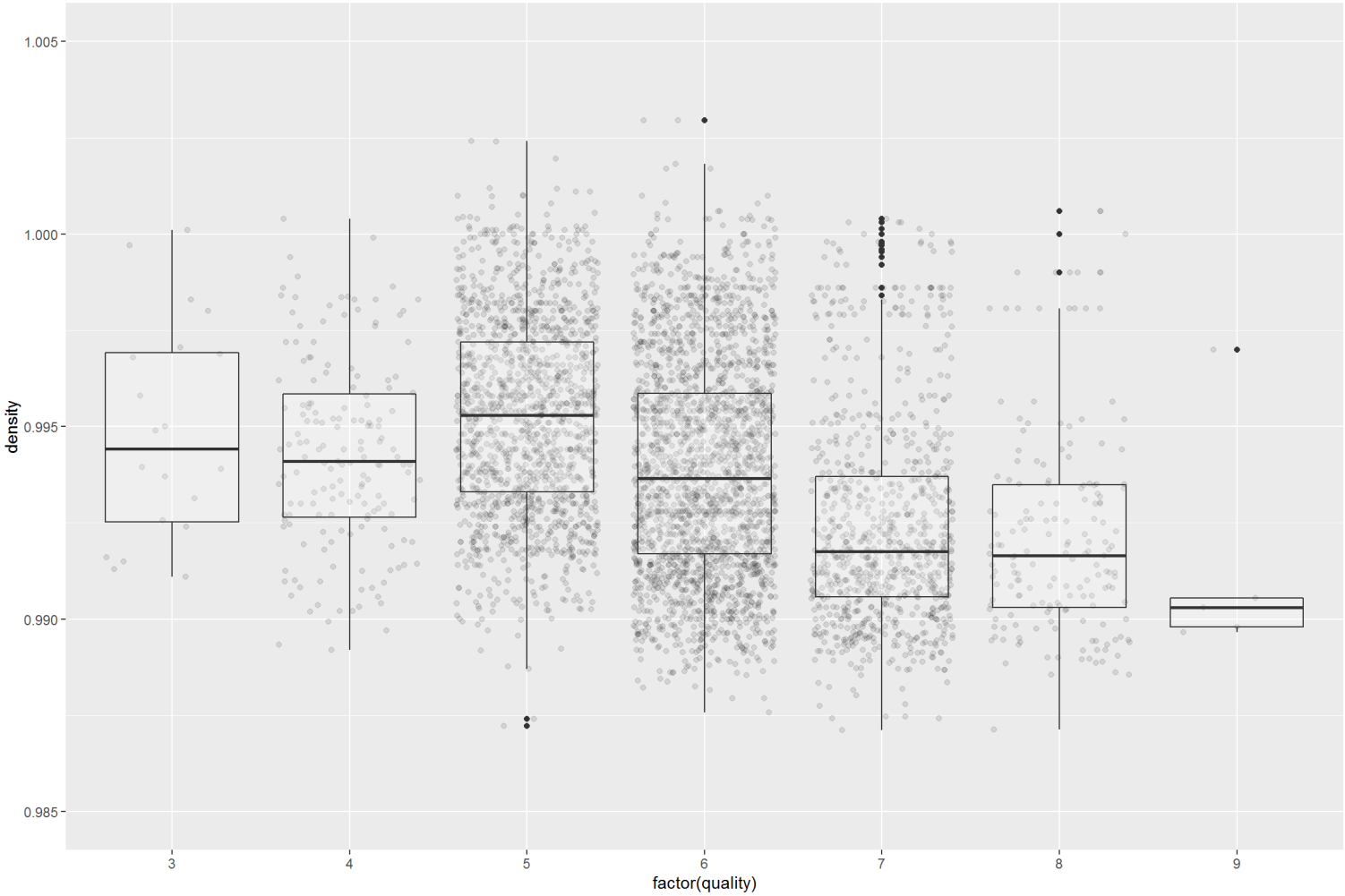


Based on the bivariate plots, several pairs of factors are with relatively high correlation coefficient. Next, I examined each pair of features with correlation coefficients (above 0.3)



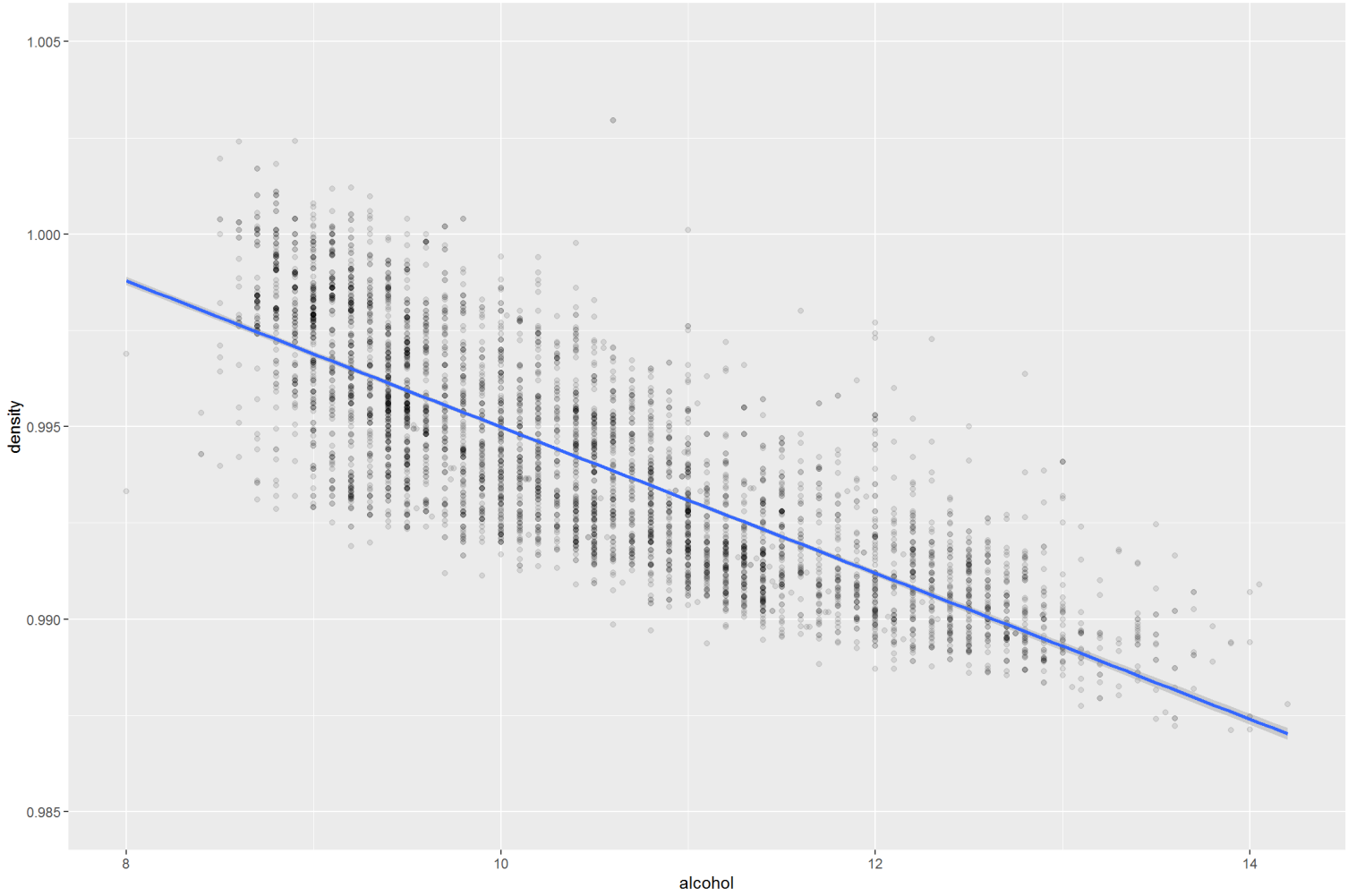

```
## [1] 0.4355747
```

The overlay of boxplot and scatter plot clearly demonstrate the alcohol content for each quality. Generally higher quality wine has higher alcohol content.



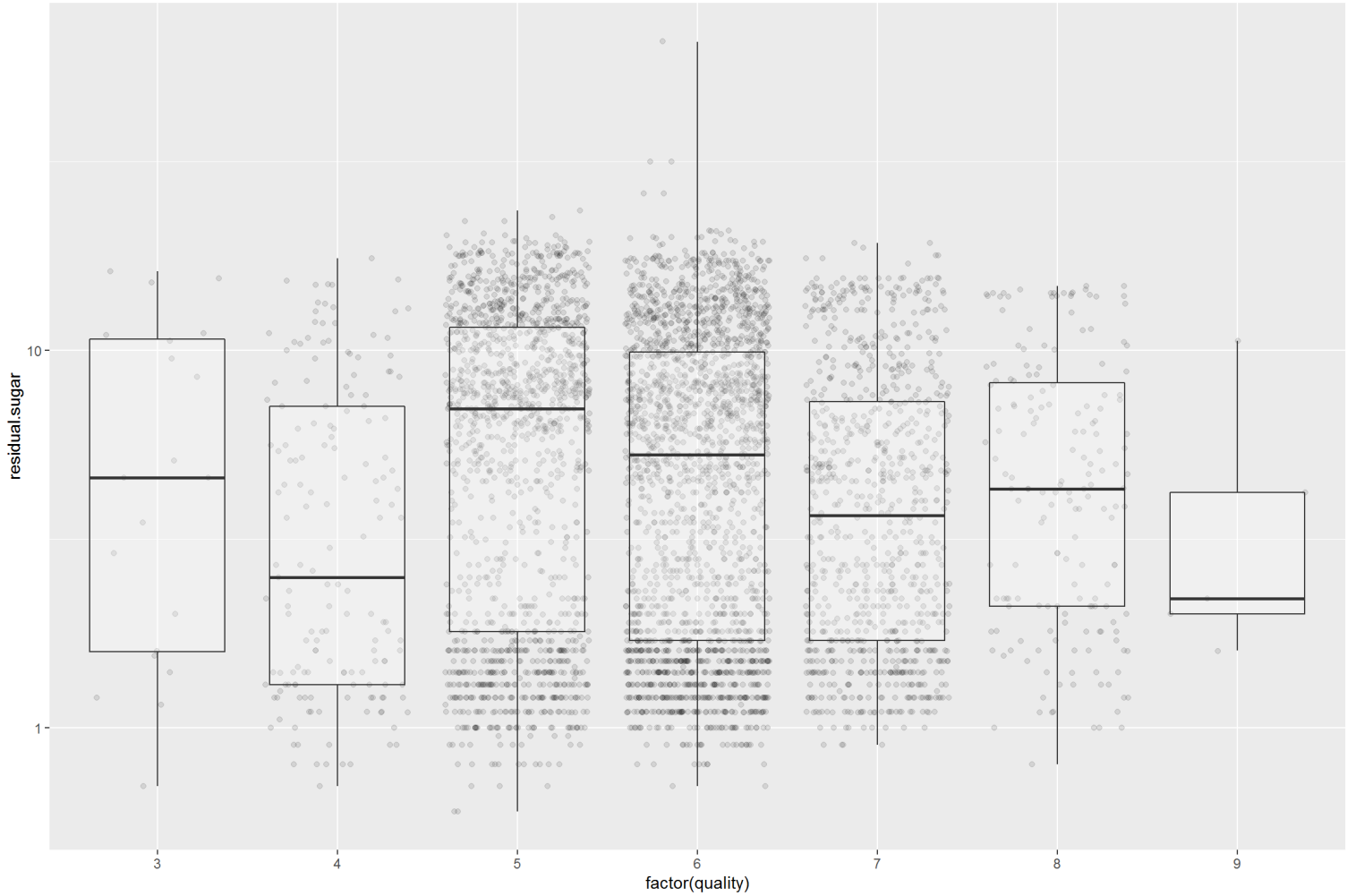

```
## [1] -0.3071233
```

The overlay of boxplot and scatter plot show that wine with higher quality usually has lower density. This is consistent with alcohol content since alcohol density is lower than water.



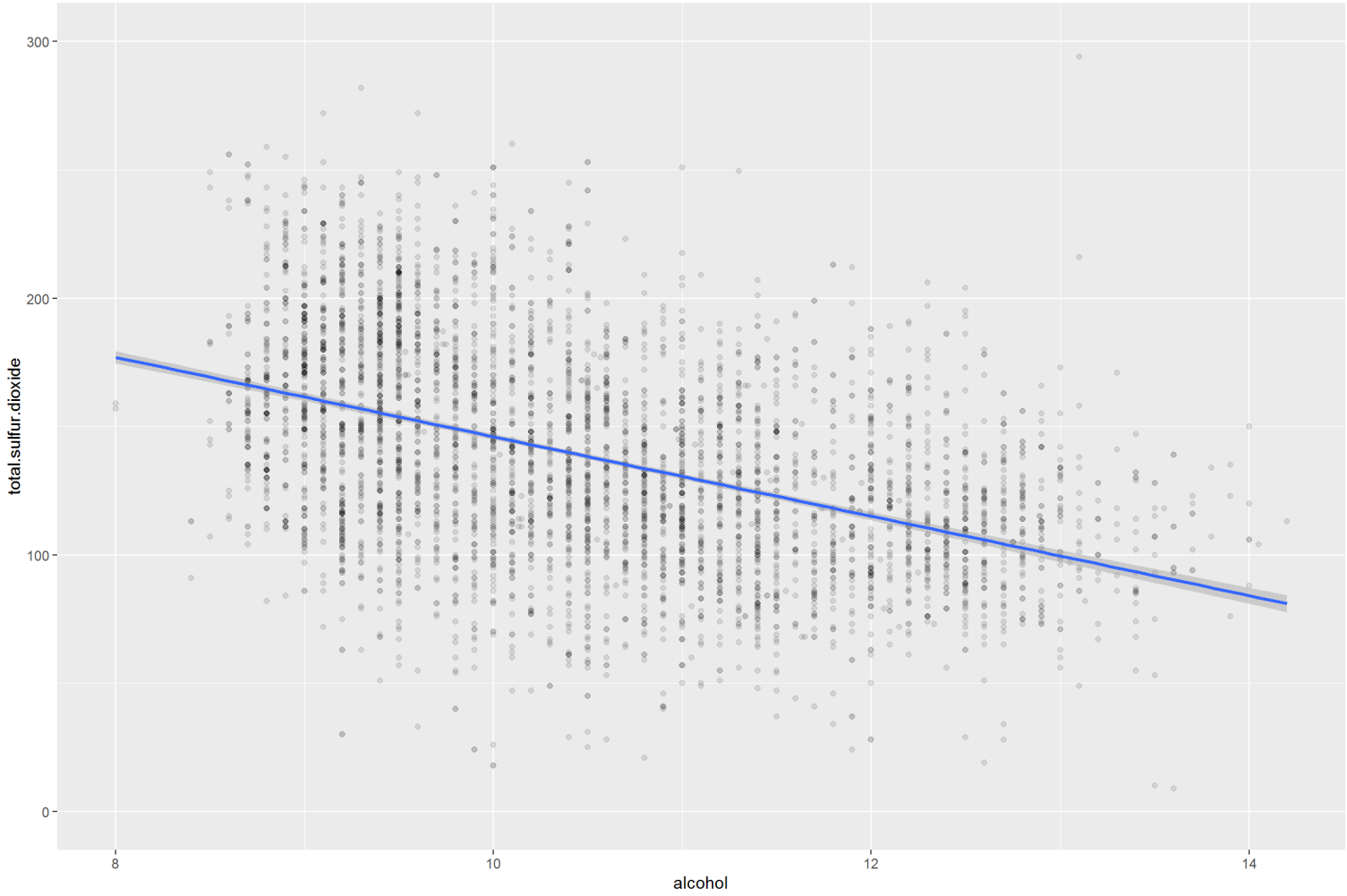

```
## [1] -0.7801376
```

The above scatter plot shows the correlation between alcohol content and density. The trend line shows that with increasing alcohol content, wine density decreases.



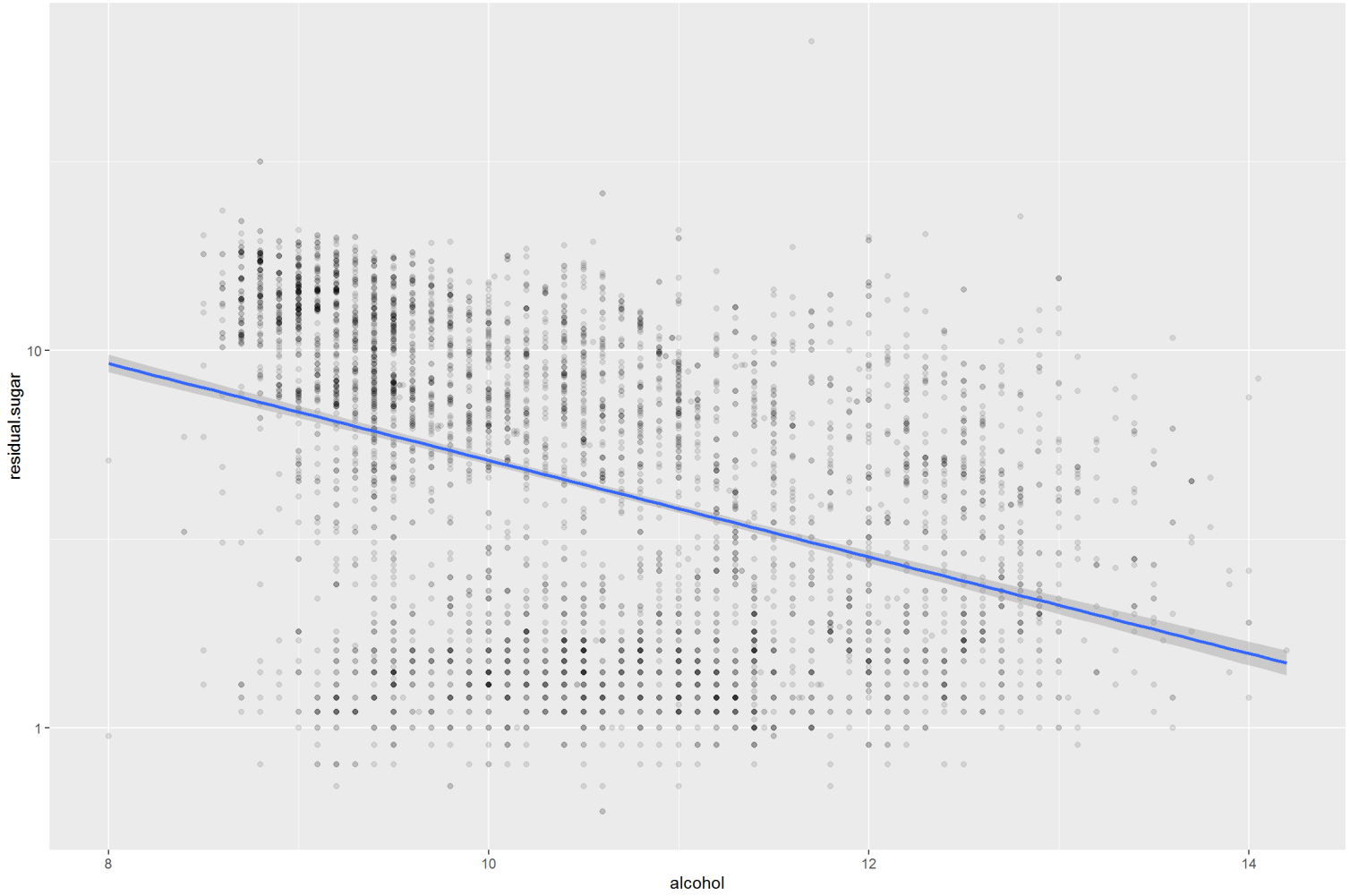

```
## [1] -0.09757683
```

This scatter plot shows that for wine with quality 4 to 8, residual sugar appears to have bimodal distribution. For wine quality 3 and 9, the data points are too few to see the bimodal distribution. There's no strong correlation between wine quality and residual sugar



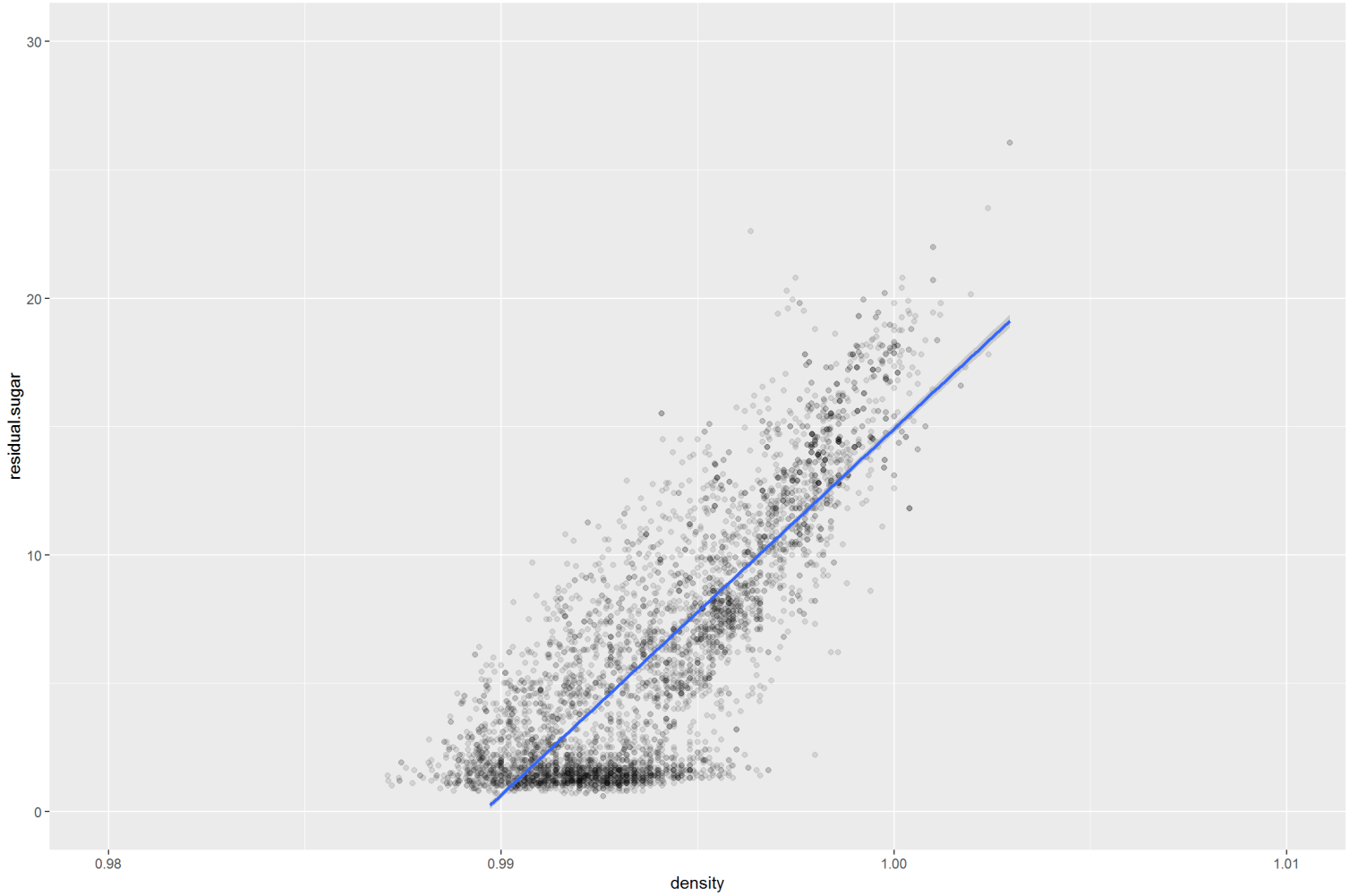

```
## [1] -0.4488921
```

This scatter plot shows that higher alcohol content negatively relates to total sulfur dioxide.




```
## [1] -0.4506312
```

This graph shows the negative correlation between alcohol and residual sugar. With increasing alcohol content, residual sugar decreases.




```
## [1] 0.8389665
```

This graph shows strong correlation between density and residual sugar. Higher density wine has higher residual sugar amount.

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?

1. The alcohol content positively relates to wine quality, with a correlation coefficient of 0.436.
2. Wine quality is negatively correlated with density.
3. The above relationships are also reflected in the correlation between density and alcohol content.

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

1. The white density and residual sugar amount has an interesting relationship. At density above 0.995, there's a strong positive correlation between density and residual sugar amount. However, for density below 0.995, the majority of wines have a similar level of residual sugar.

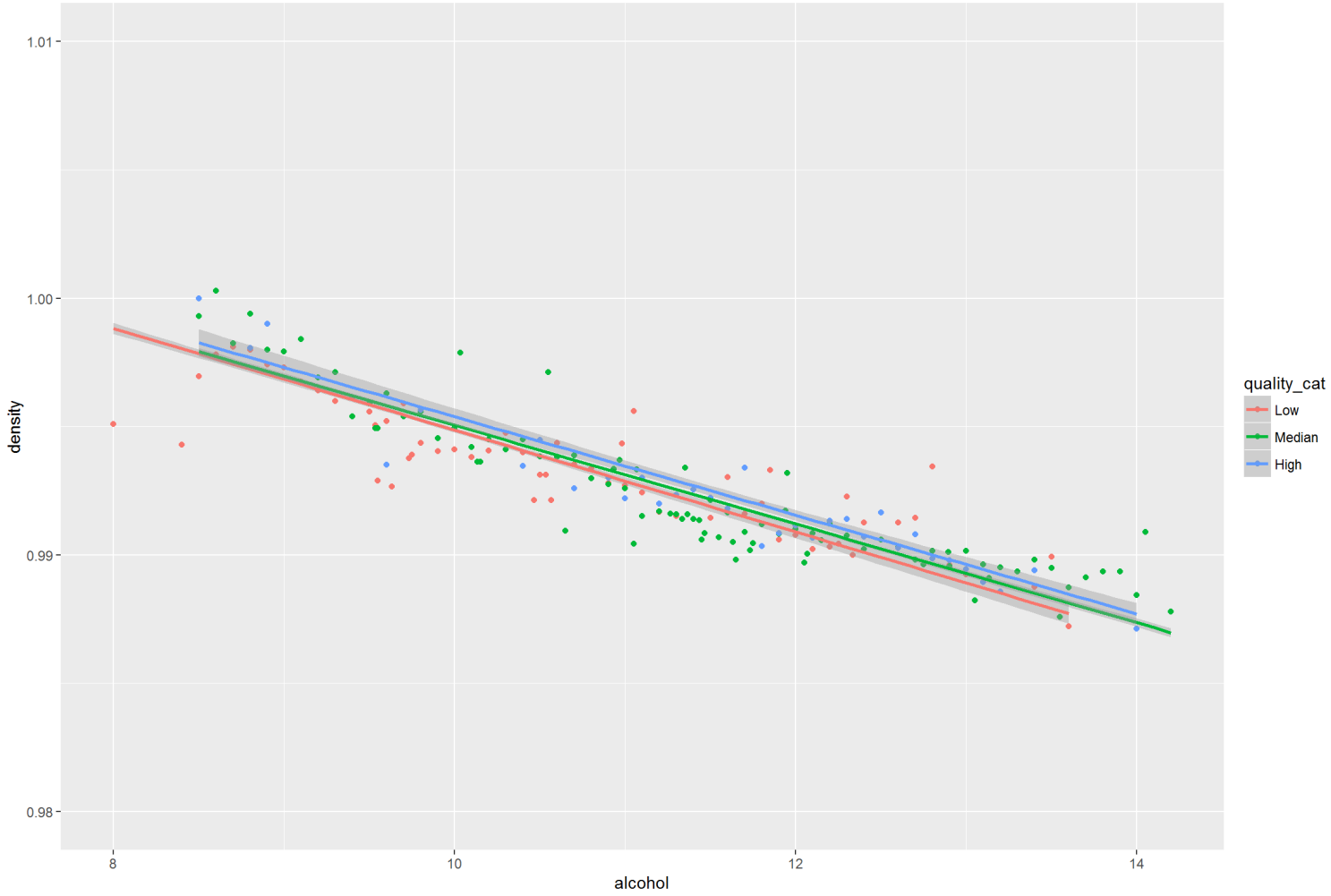
2. Alcohol and density is negatively correlated. This is due to the physical property of alcohol.

What was the strongest relationship you found?

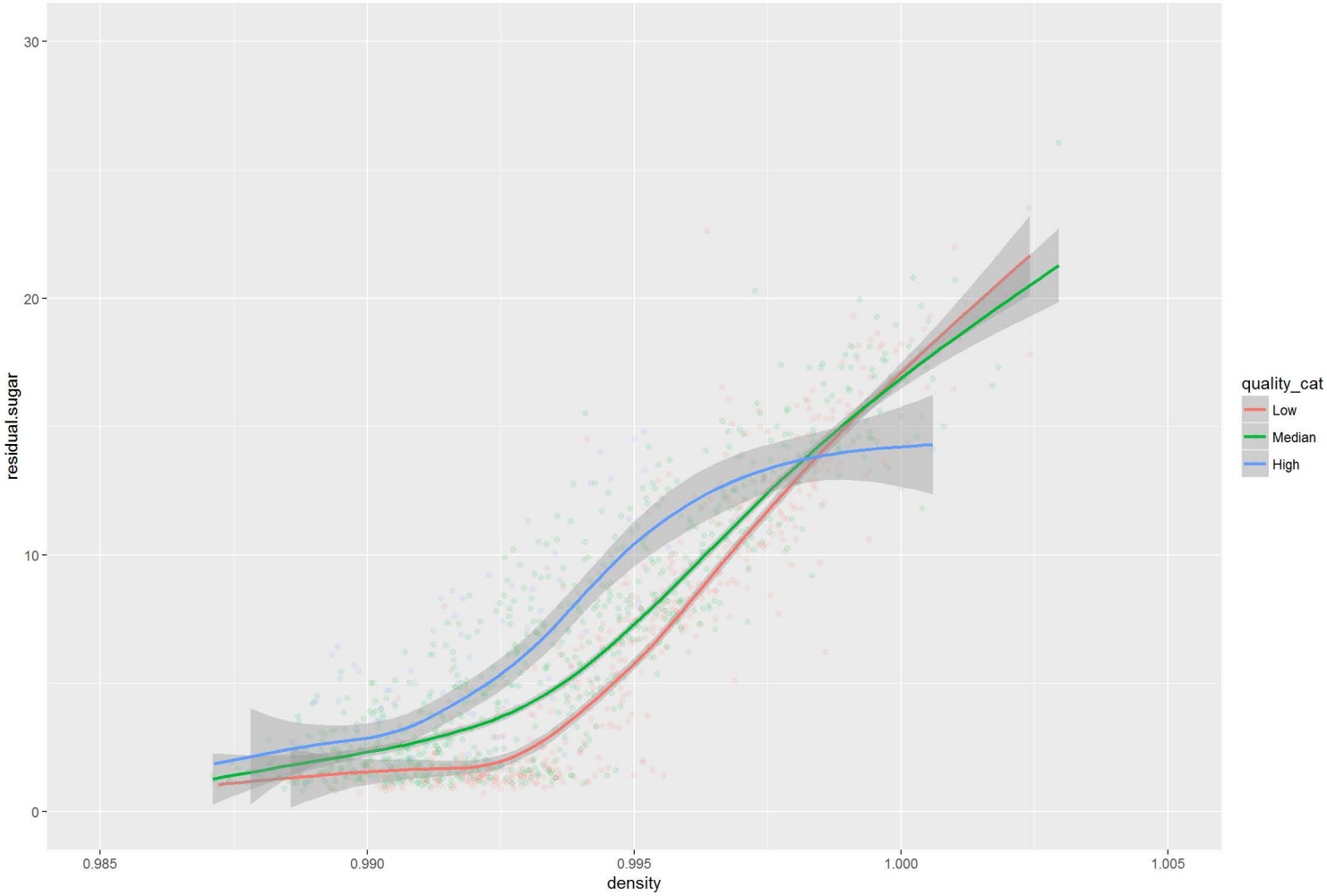
The strongest relationship is the negative correlation between alcohol and density.

Multivariate Plots Section

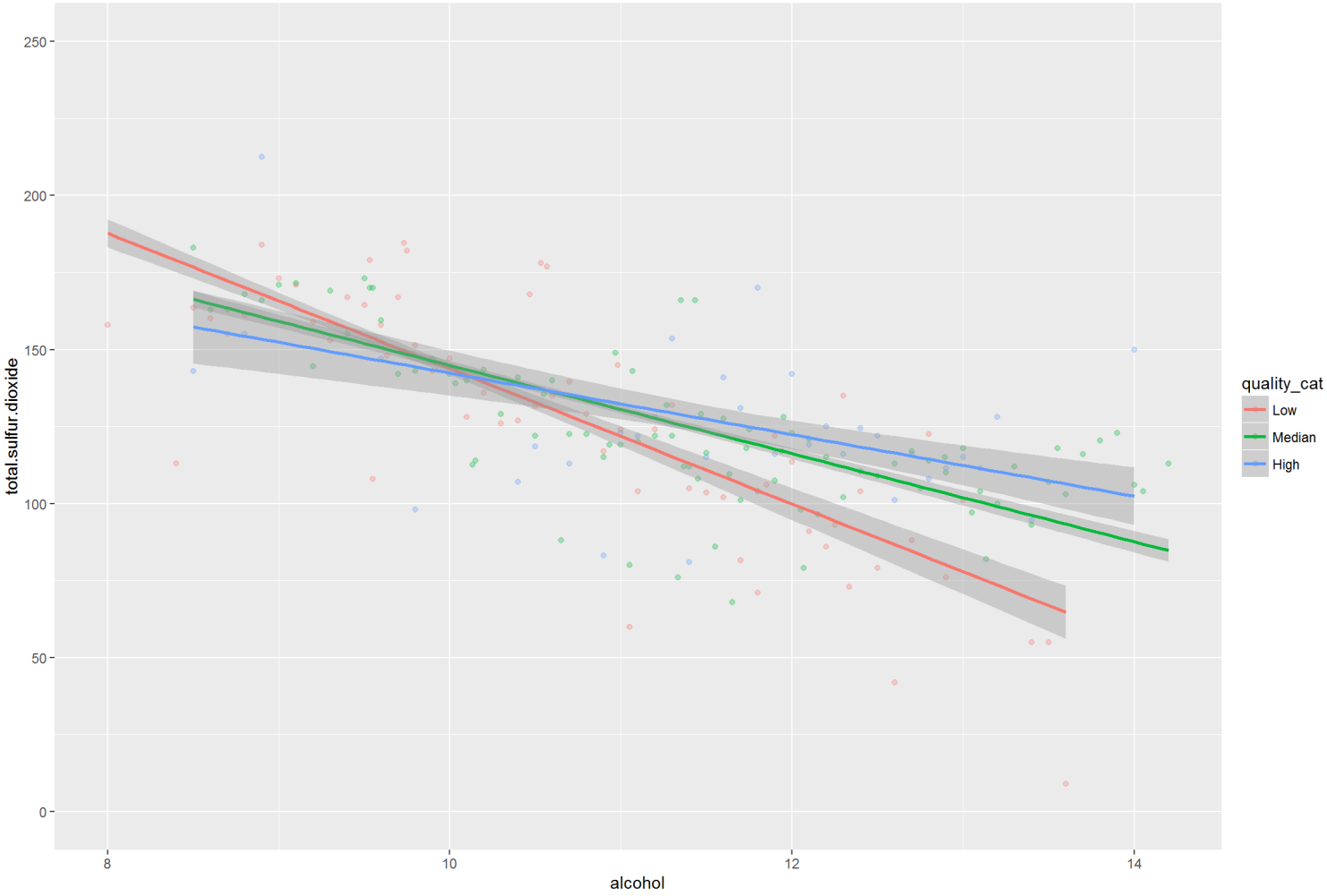
```
##
##    Low Median   High
##   1640   3078    180
```



The above graph shows that median desnity for different alcohol content for each quality of wine. Wine density shows a negative correlation with alcohol content, indepdent of wine quality.



The above graphs shows the trend of residual sugar vs density for different quality of wines. Overall, residual sugar increases with increasing of wine density. The residual sugar of high quality wines reaches plateau eventually. However, low and median quality wines keep having higher residual sugar with increasing of density.




```
## [1] -0.4089919
```

```
## [1] -0.4466632
```

```
## [1] -0.3893316
```

The above graph shows the linear negative correlation of alcohol content and total sulfur dioxide. For all quality of wines, higher alcohol content correlates to lower total sulfur dioxide. Based on statistical analysis, median quality wines have the highest correlation coefficient.

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?

As discussed above, I observed that with increasing content of alcohol, the density decreases. This trend is independent of wine quality.

Different quality of wines have different relationship of residual sugar vs density.

Were there any interesting or surprising interactions between features?

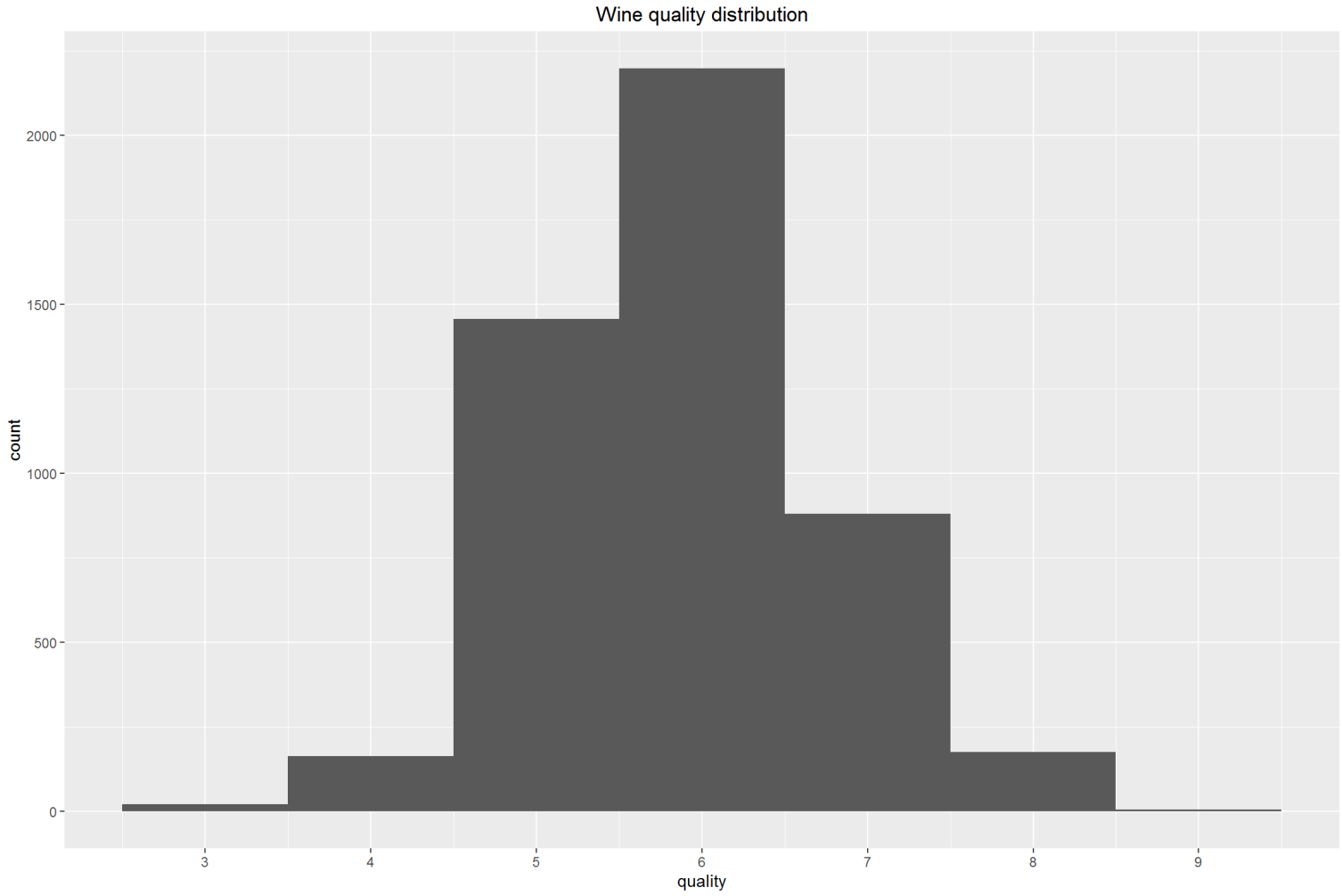
Due to the weak correlations between most of the features, I did not find strong interesting correlations between mutliple variates.However, the different trend of residual sugar vs density for different quality of wines are interesting.

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

I did not create any model.

Final Plots and Summary

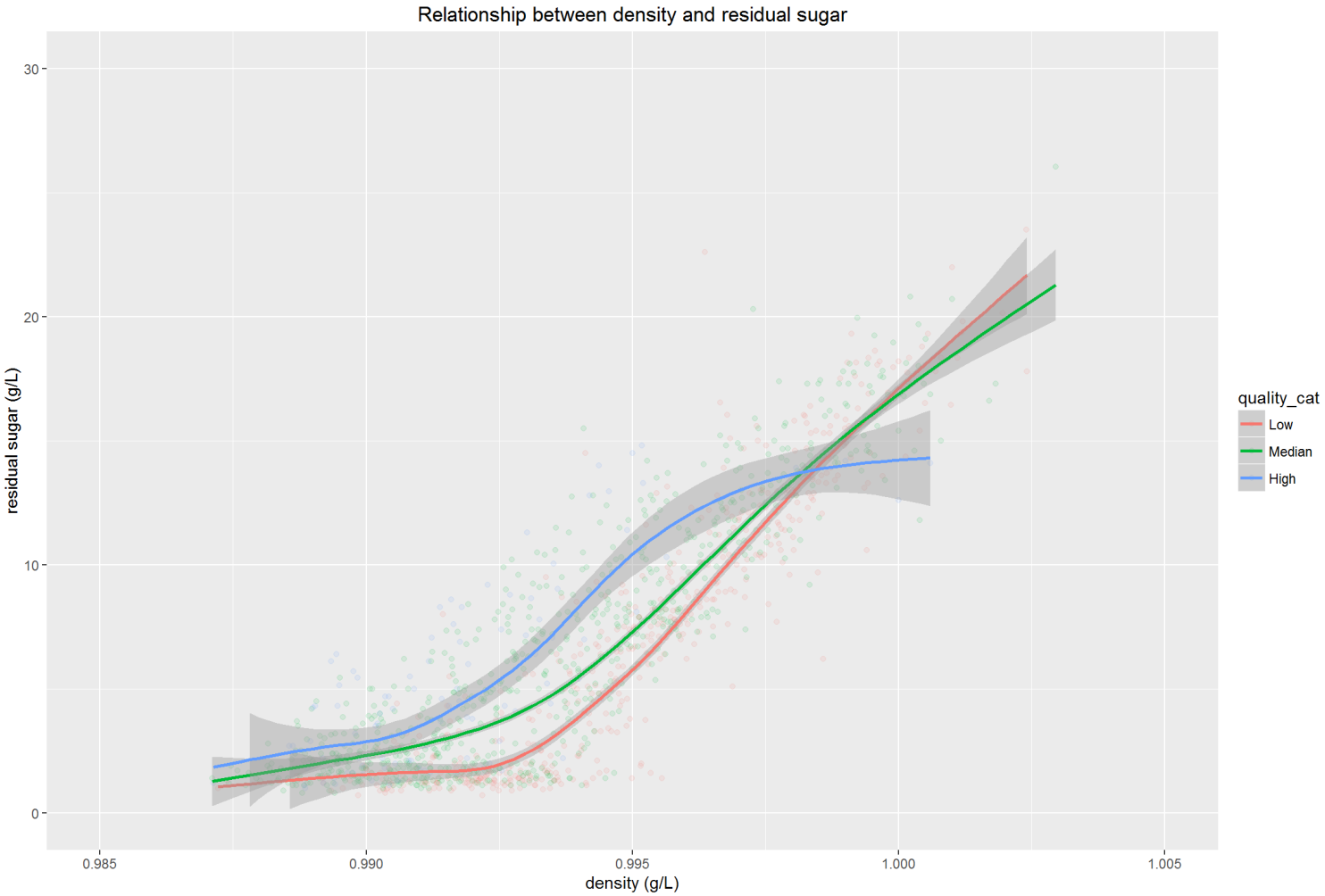
Plot One



Description One

Wine qualit appears to be normal distribution. Based on statistic analysis, the mean quality is 5.878. The median quality is 6.

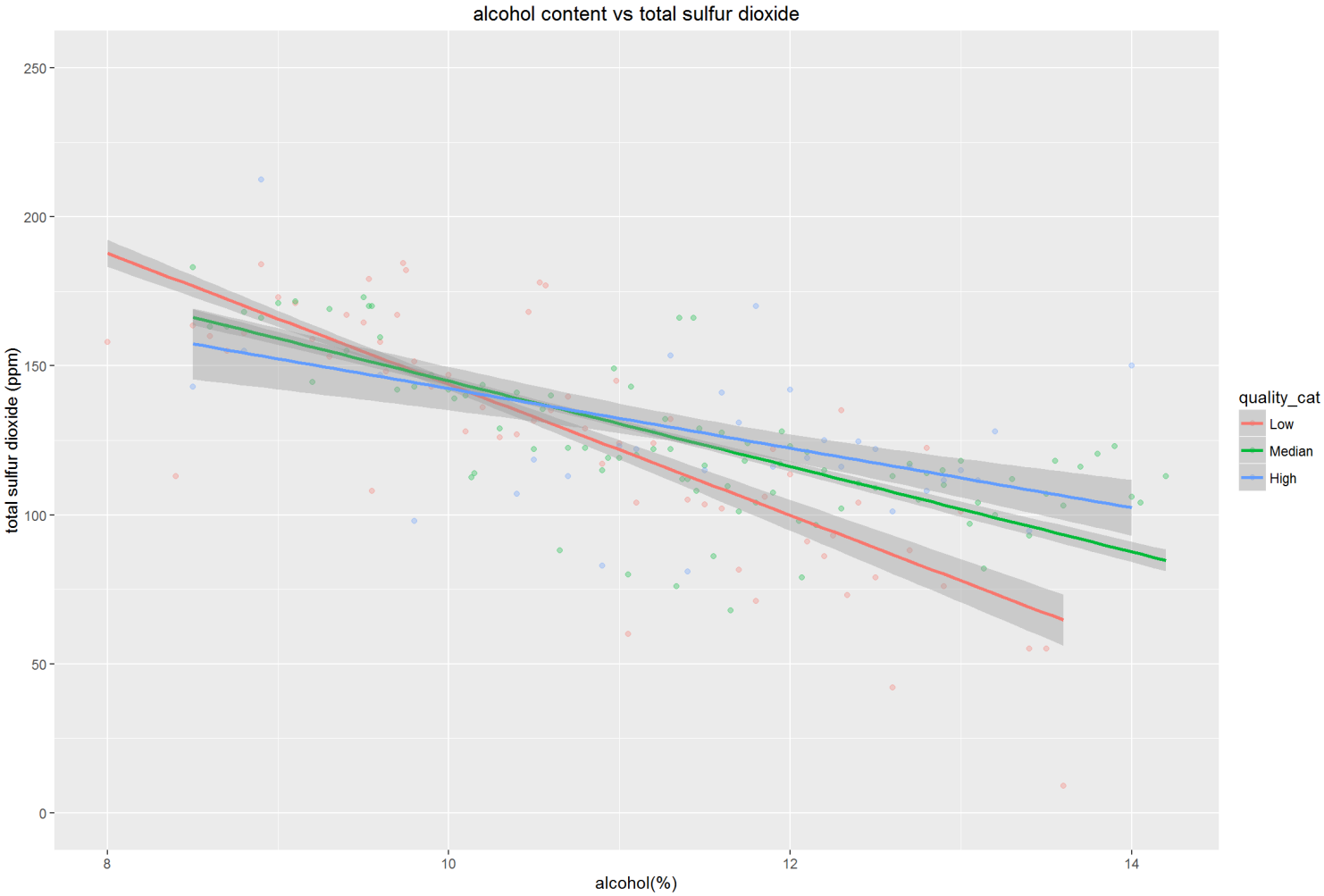
Plot Two



Description Two

The plot of residual sugar vs density shows different trend for different quality of wine. For low (quality between 3 and 5) and median (quality between 6 and 7) quality wines, residual sugar started to increase with density for density above 0.992. For high (quality between 8 and 9) quality wines, residual sugar started to increase at even lower density. However, it reaches plateau around density 0.998.

Plot Three



Description Three

There is a negative correlation between alcohol content and total sulfur dioxide. Different quality wines show slight different correlation coefficient. Based on statistialy analysis shown in the “multivariate plot section”, median quality wine has the strongest correlation, while high quality wine shows a relative weaker correlation. Low quality wine has a steeper slope, indicating that change of alcohol content corresponds to a bigger change of total sulfur dioxide compared to median and high quality wines.

Reflection

This white wine dataset is relatively small, with only 15 variables. Initially, I wanted to explore features that affect wine quality. First, I looked at each individual feature distribution using histograms. Then I performed bivariate analysis to search for features that coorelate with wine quality. I found that acohol content and wine density have the highest correlation coefficient. In addition, I also analyzed other correlated features, including residual sugar content, density and pH. At last, I transformed the quality variable from integer to categorical variable to perform multi-variate anslysis.

One of the struggles I had is how to find features that contribute to wine quality. Unlike the diamond dataset, after running the bivariate analysis, many features correlate with diamond price with correlation coefficient higher than 0.8. In this wine dataset, the correlation coefficient between chemical properties of wine and wine quality is small, less than 0.4. This indicates that it is hard to predict wine quality with current features accurately. It may require feature engineering to discover the deep relationship between features and wine quality. Given my limited knowledge of wine, currently I am not able to perform complicated feature engineering to come up a reliable model for prediction of wine quality based on

chemical properties.

Overall, the features in the dataset did not have strong correlation with wine quality. It is difficult to build up a prediction model with current features. In the future, it would be interesting to know more features about the wine to get better idea of wine quality, such as grape type, incubation time, manufacture and production location.