Factors Influencing Perception of Red Wine Quality by Steven Hansen

Data Description

This dataset is related to variants of the Portuguese red wine “Vinho Verde”. For more details, consult Vinho or reference [Cortez et al., 2009].

The dataset contains almost 1600 samples, 11 input- and one output variables.

The input variables are test data on the following physiochemical properties:

1. fixed acidity (tartaric acid, g/l)
2. volatile acidity (acetic acid, g/l)
3. citric acid (g/l)
4. residual sugar (g/l)
5. chlorides (sodium chloride, g/l)
6. free sulfur dioxide (mg/l)
7. total sulfur dioxide (mg/l)
8. density (kg/l)
9. pH
10. sulphates (potassium sulphate, g/l)
11. alcohol (percent by volume)

The output variable is a factor called a quality score with levels from 0 (very bad) to 10 (excellent). Each score is a median of at least three evaluations made by wine experts.

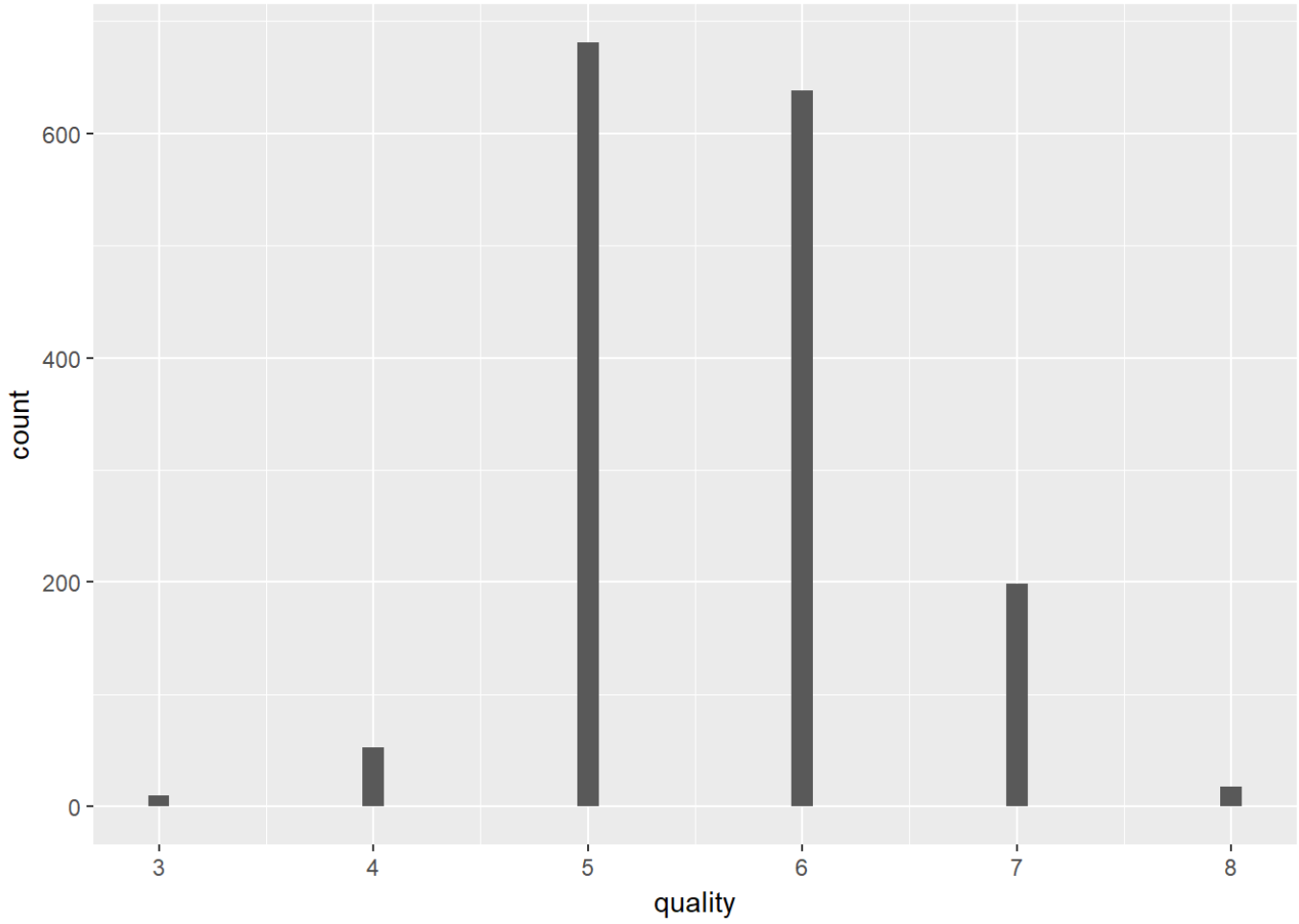
Univariate Plots Section

```
## '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 ...
```

```
##        X          fixed.acidity   volatile.acidity  citric.acid
##  Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000
##  1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090
##  Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260
##  Mean   : 800.0   Mean   : 8.32   Mean   :0.5278   Mean   :0.271
##  3rd Qu.:1199.5   3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420
##  Max.   :1599.0   Max.   :15.90   Max.   :1.5800   Max.   :1.000
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.900   Min.   :0.01200   Min.   : 1.00
##  1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00
##  Median : 2.200   Median :0.07900   Median :14.00
##  Mean   : 2.539   Mean   :0.08747   Mean   :15.87
##  3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00
##  Max.   :15.500   Max.   :0.61100   Max.   :72.00
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  6.00       Min.   :0.9901   Min.   :2.740   Min.   :0.3300
##  1st Qu.: 22.00       1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500
##  Median : 38.00       Median :0.9968   Median :3.310   Median :0.6200
##  Mean   : 46.47       Mean   :0.9967   Mean   :3.311   Mean   :0.6581
##  3rd Qu.: 62.00       3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300
##  Max.   :289.00       Max.   :1.0037   Max.   :4.010   Max.   :2.0000
##     alcohol         quality
##  Min.   : 8.40   Min.   :3.000
##  1st Qu.: 9.50   1st Qu.:5.000
##  Median :10.20   Median :6.000
##  Mean   :10.42   Mean   :5.636
##  3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :14.90   Max.   :8.000
```

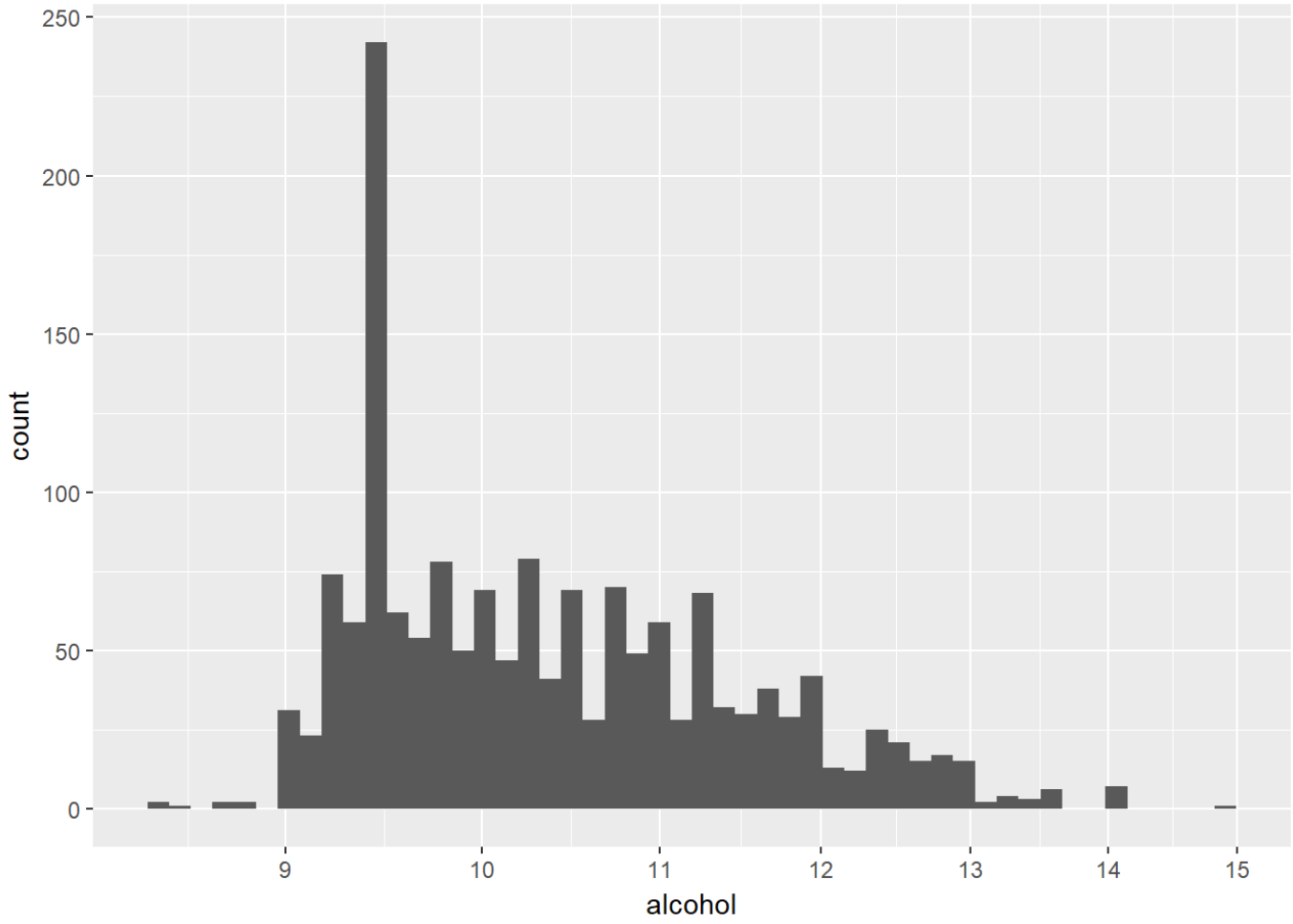
Many of the input variables pertain to concentration ratios, so they have an absolute lower bound of zero. Although there may also be hard upper bounds, these may lie outside the acceptable for a marketeable wine. I expect these variables to have some positive skew and/or more right-hand outliers than left-hand. The mean values generally *are* a bit larger than the median values.

Free sulfur dioxide and a few other variables might have quantized levels.

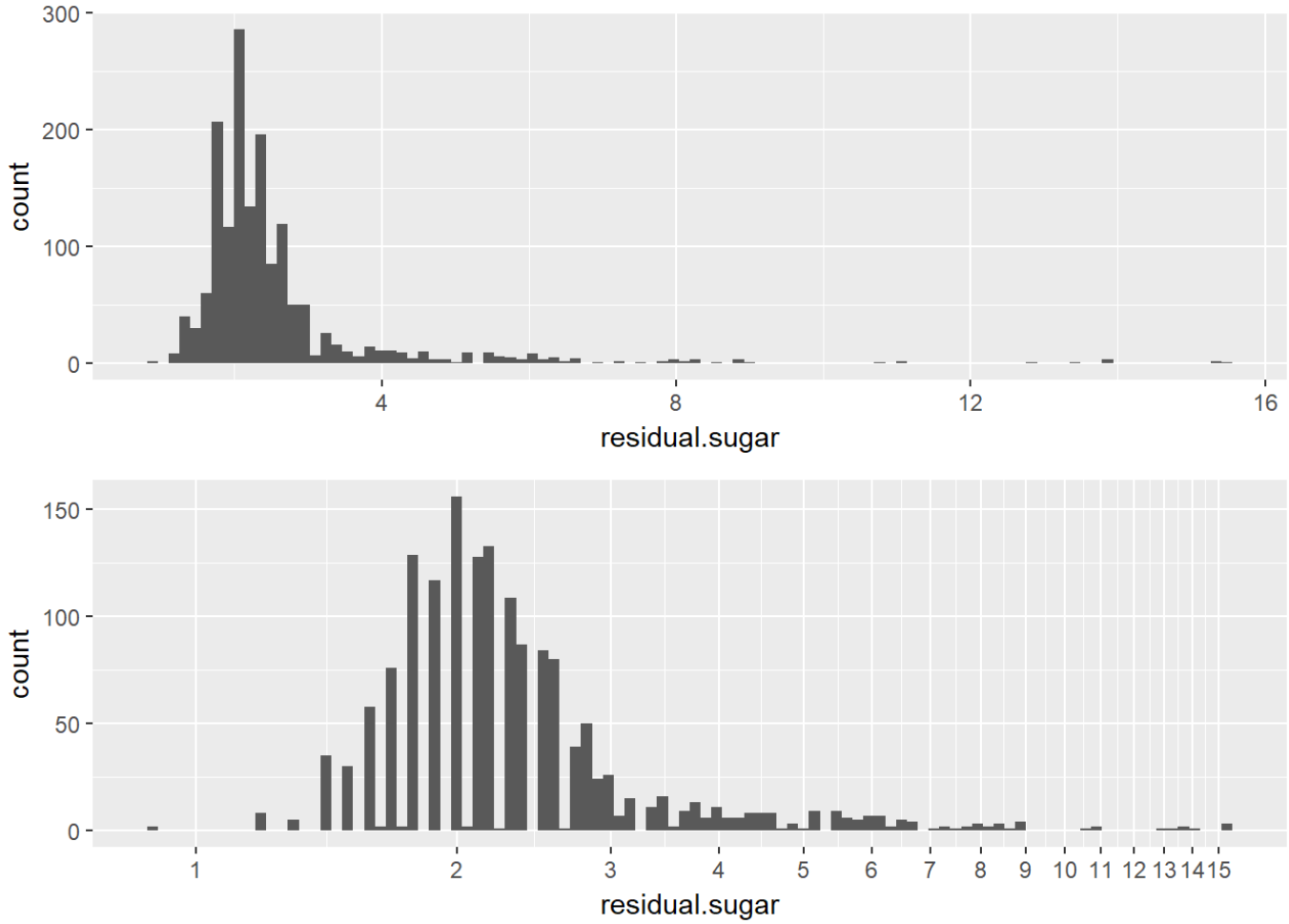


This binwidth illustrates that the quality scores are discrete-valued, so I will add a factor version to the dataset.

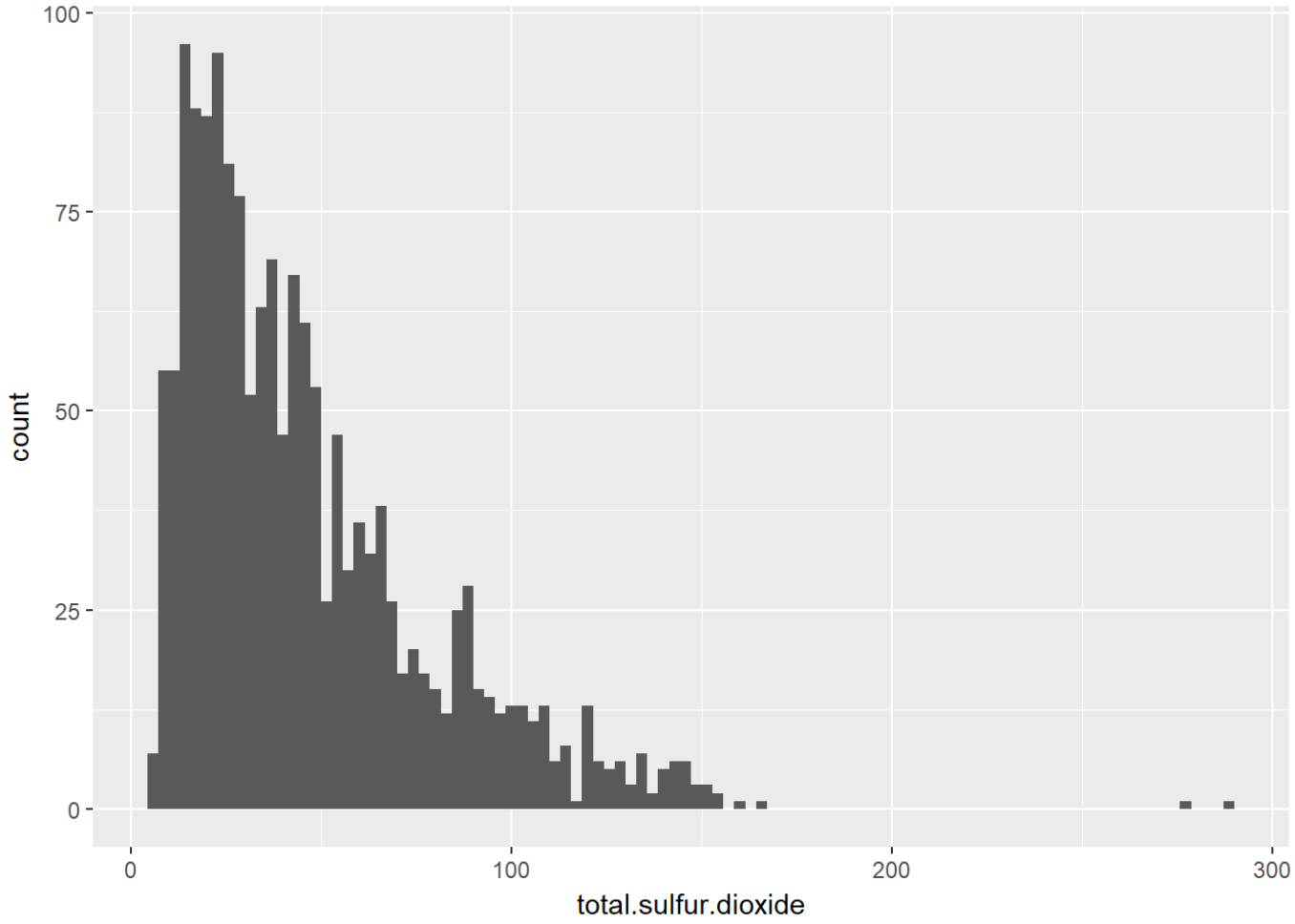
Although quality scores of 0 to 10 are allowed, all of these samples graded between 3 and 8, and most rated 5 to 7. This tight grouping may indicate consistent quality-control at Vinho Verde. A web search shows that most of their wines are priced affordably, so the mid-range quality scores are not necessarily disappointing. The distribution appears roughly binomial without applying any transformation.



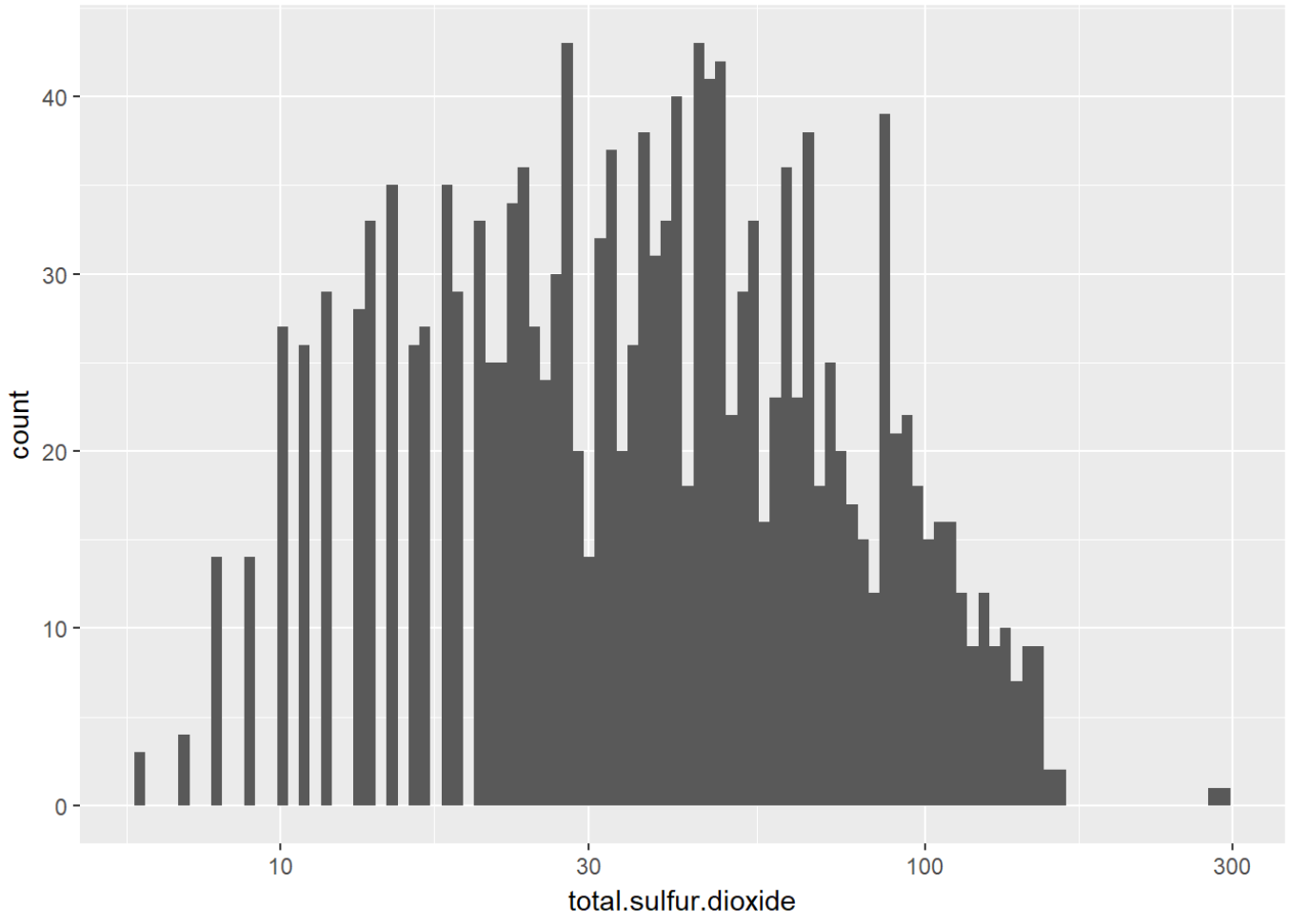
Alcohol is skewed right slightly, so I have plotted it on a log_10 axis. There is a strong modal tendency around 9.4 to 9.5 percent, which might be the ABV for a popular wine. I am surprised to see so many red wines below 10.5 percent.



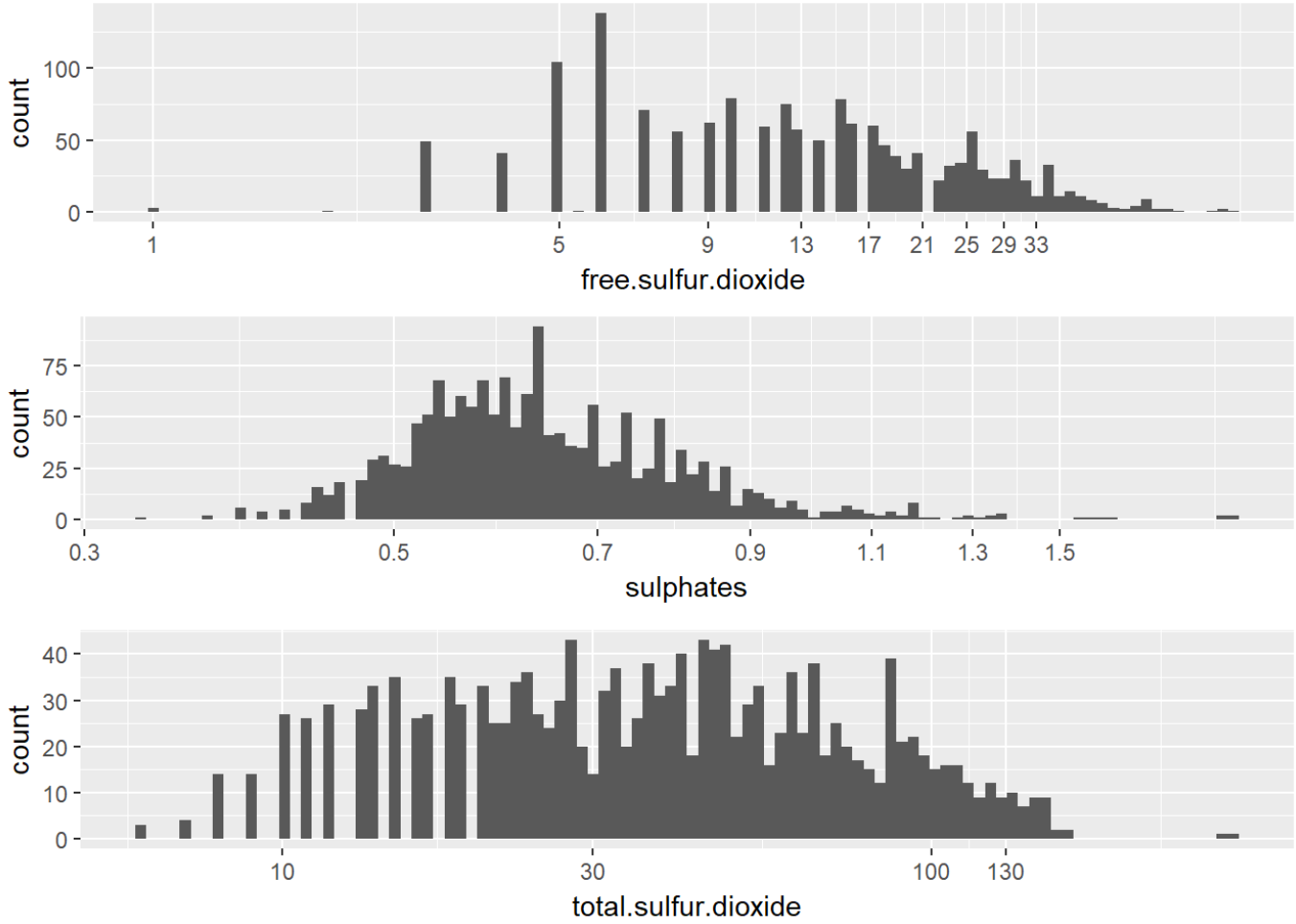
Most of the samples have about 2 percent residual sugar, but a few range as high as 15.5. These could be a small number of dessert wine or Port varietals. Using a log_10 transform helps reveal some tail details, since the peak bin count is reduced by almost half.



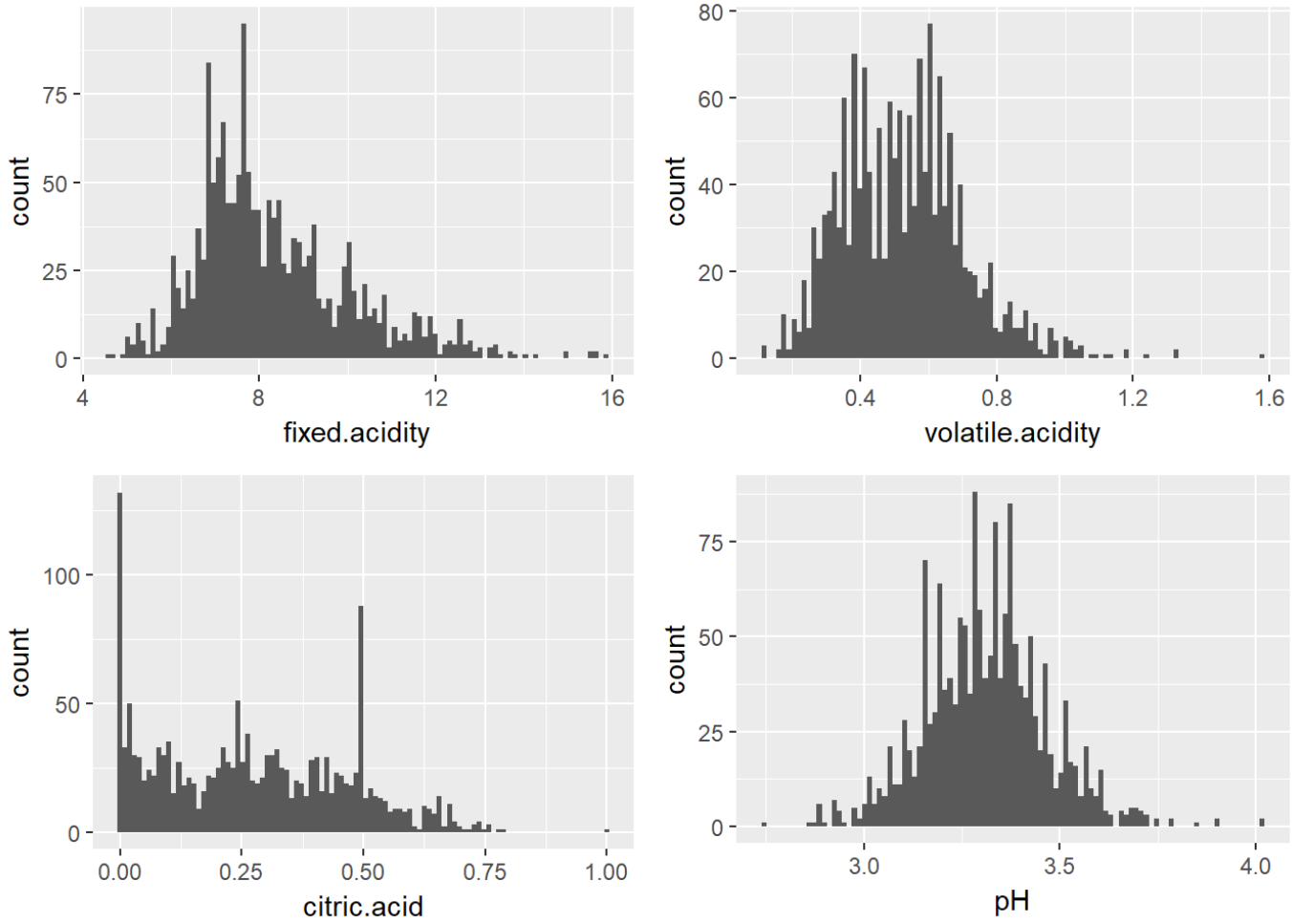
Total sulfur dioxide is definitely skew-right, and there are a few possible outliers.



Plotting the log_10-transformed total sulfur dioxide gives more insight into the distribution. Although there are fewer samples below about 20 mg/l, the distribution appears rougly normal. The gaps between low-valued samples might have to do with limitations of precision.



Here I have plotted the distributions of all three variable related to sulfur or sulphates on long_10 scales. The unit sizes are are different, but I am wondering if some of these variables are correlated.

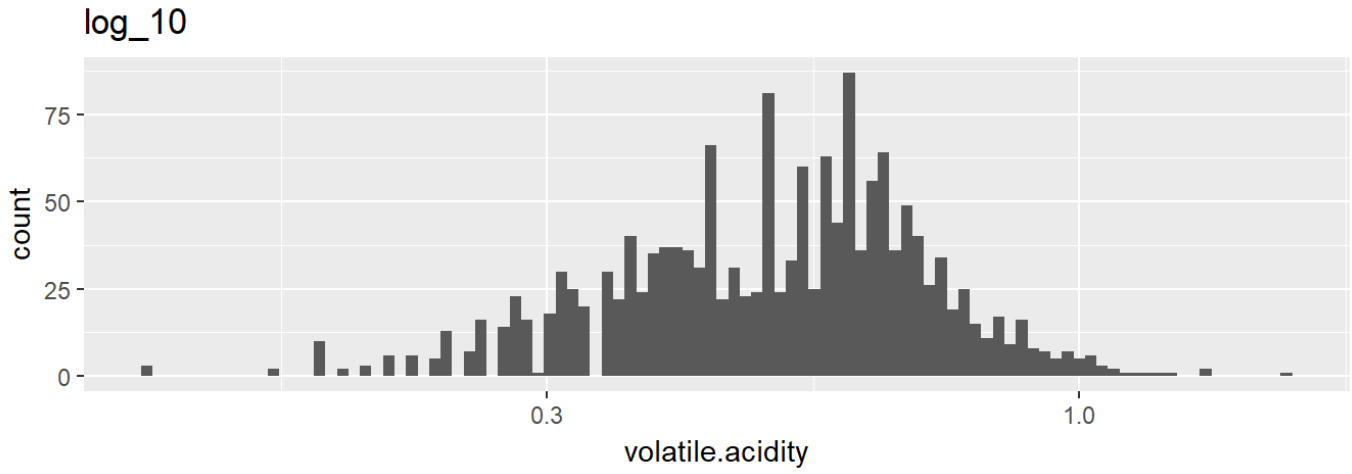


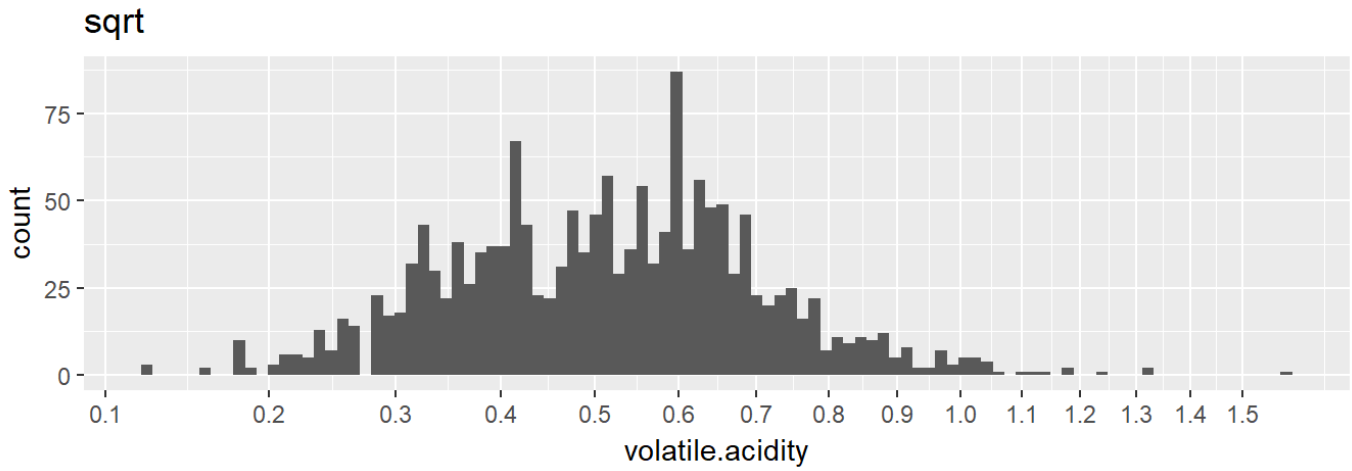

```
##
##    0 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09  0.1 0.11 0.12 0.13 0.14
##  132   33   50   30   29   20   24   22   33   30   35   15   27   18   21
## 0.15 0.16 0.17 0.18 0.19  0.2 0.21 0.22 0.23 0.24 0.25 0.26 0.27 0.28 0.29
##   19    9   16   22   21   25   33   27   25   51   27   38   20   19   21
##  0.3 0.31 0.32 0.33 0.34 0.35 0.36 0.37 0.38 0.39  0.4 0.41 0.42 0.43 0.44
##   30   30   32   25   24   13   20   19   14   28   29   16   29   15   23
## 0.45 0.46 0.47 0.48 0.49  0.5 0.51 0.52 0.53 0.54 0.55 0.56 0.57 0.58 0.59
##   22   19   18   23   68   20   13   17   14   13   12    8    9    9    8
##  0.6 0.61 0.62 0.63 0.64 0.65 0.66 0.67 0.68 0.69  0.7 0.71 0.72 0.73 0.74
##    9    2    1   10    9    7   14    2   11    4    2    1    1    3    4
## 0.75 0.76 0.78 0.79    1
##    1    3    1    1    1
```

Wine pH seems to have a very normal distribution.

Citric acid has two bins at 0 and 0.49 with unusually high frequencies. There is also a possible outlier at 1.00. I may want to exclude those bins when looking at boxplots or scatter plots.

I also notice that volatile acidity *could* be bimodal, so I want to investigate that a little further.





It looks like a square-root transformation gives a bit more symmetric dist- ibution for volatile acidity. It also illustrates a little gap in the data around 0.28 g/l, and arguably, some slight bimodality with a dip in frequency at 0.45. Maybe two distinct categories of wine could be present in the data, but its hard to say.

Univariate Analysis

What is the structure of your dataset?

There are 1599 wine samples in the dataset with 12 features. Eleven are measured physiochemical properties and the twelfth is a quality score of the wine from the median of at least three expert evaluations. All of the properties are continuous variables, although there is some quantization present. The quality score is factor from 0 to 10, with levels of 3 to 8 in this dataset.

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

Since price is not included, the main feature of interest is the quality score. I suspect that some of the chemical properties can be used to build a predictive model to judge wine quality.

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

I doubt that density does much to independently predict wine quality. It should be correlated to alcohol and residual sugar, though. I suspect that high sulfur content will correlate with low quality scores, although I'm not sure yet which of the sulfur properties will be the most predictive. I know from research that sulphates are added to lower quality wines to protect against oxidation and bacterial growth, but at the price point of

these wines, I doubt it will be a strong predictor of quality. Lastly, I think higher citric acid might positively relate to quality, since it makes the wine taste fresh or bright, and many of these samples are below 10 percent ABV, where a refreshing taste may be considered desirable.

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

I created a factor variable, quality.factor, by factoring the quality score.

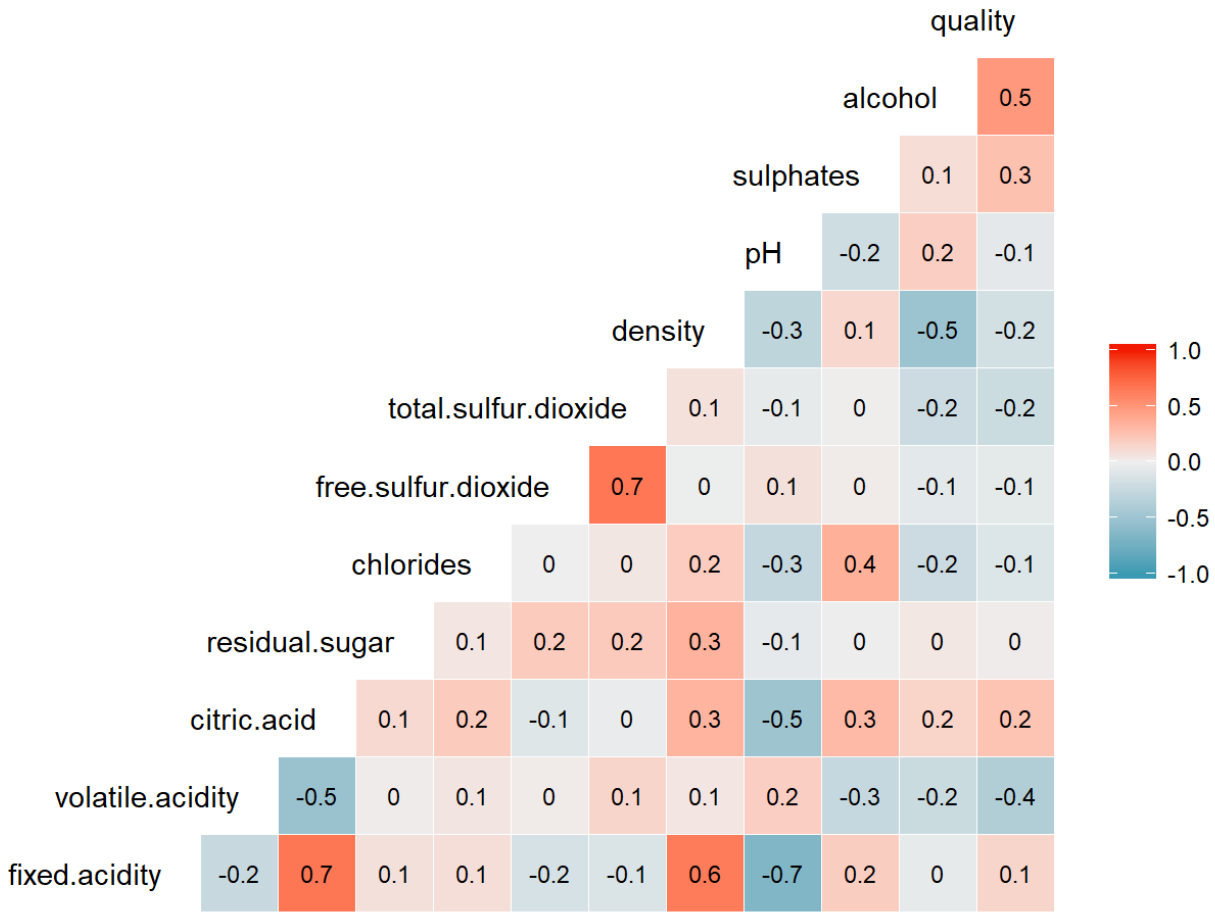
Of the features you investigated, were there any unusual distributions?

I log-transformed residual sugar, and the sulfur/sulphate distributions. I square-root transformed the volatile acidity distribution. Of all the distributions I looked at so far, pH was the most inherently normal.

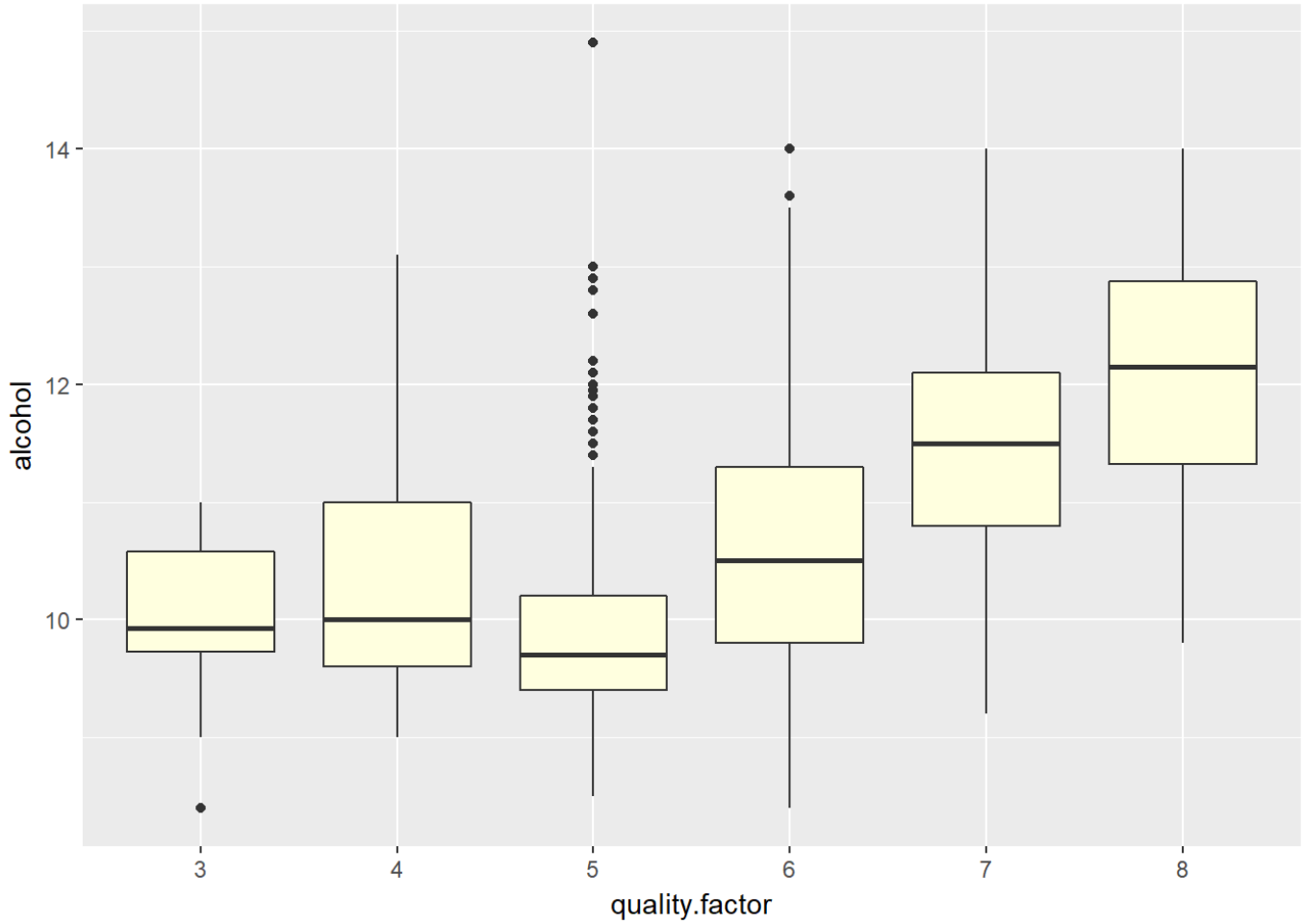
This dataset was already tidy, with no ill-defined values, but I will investigate citric acid with the high-count bins at 0.0 and 0.49 and the outlier at 1.00 removed.

Bivariate Plots Section

Most of the variables have a non-uniform distribution, including alcohol, volatile acidity, sulphates, citric acid and density. I am curious to plot quality versus some of these quantities and plot them against one another, to see if their distributions are related.

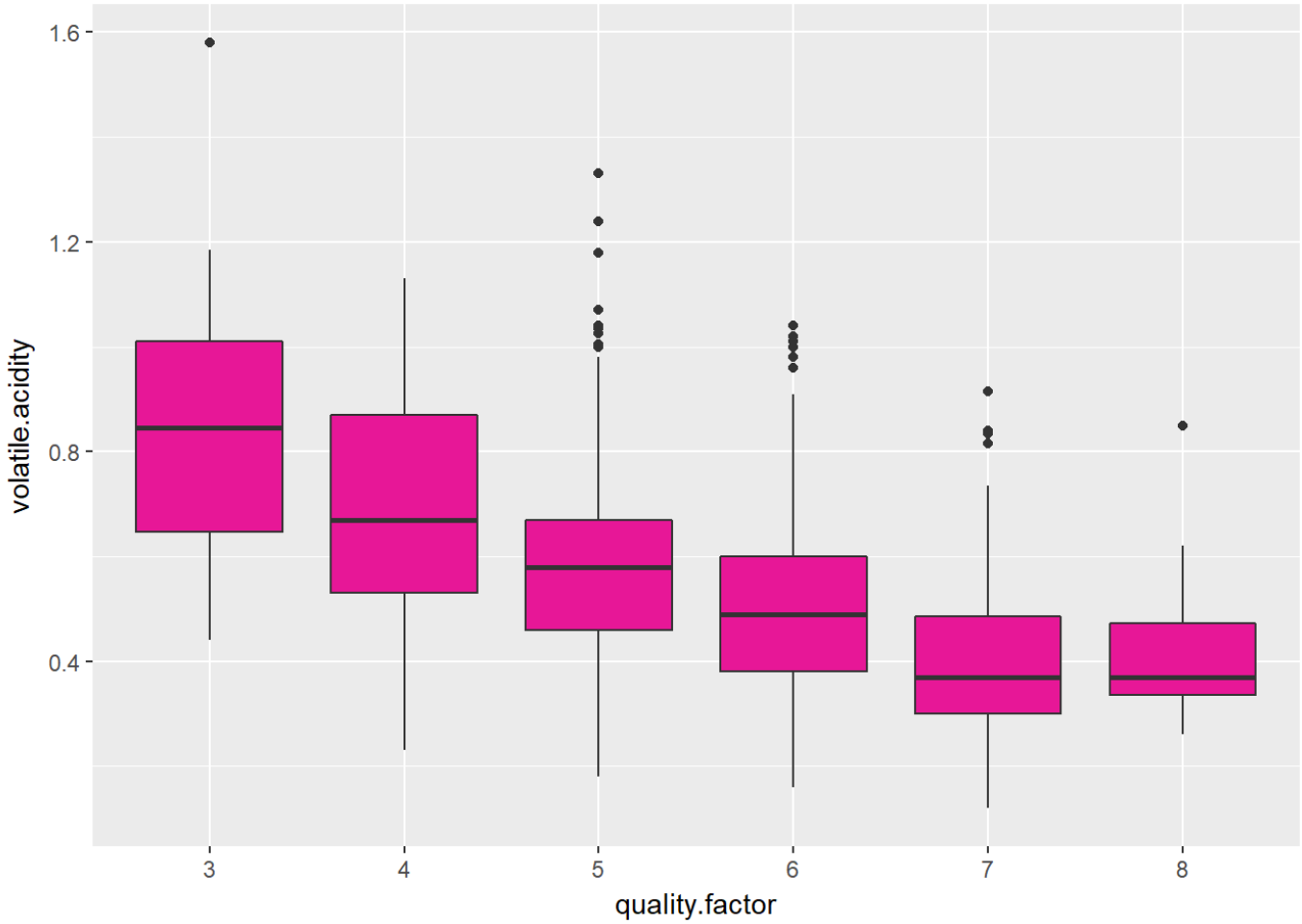


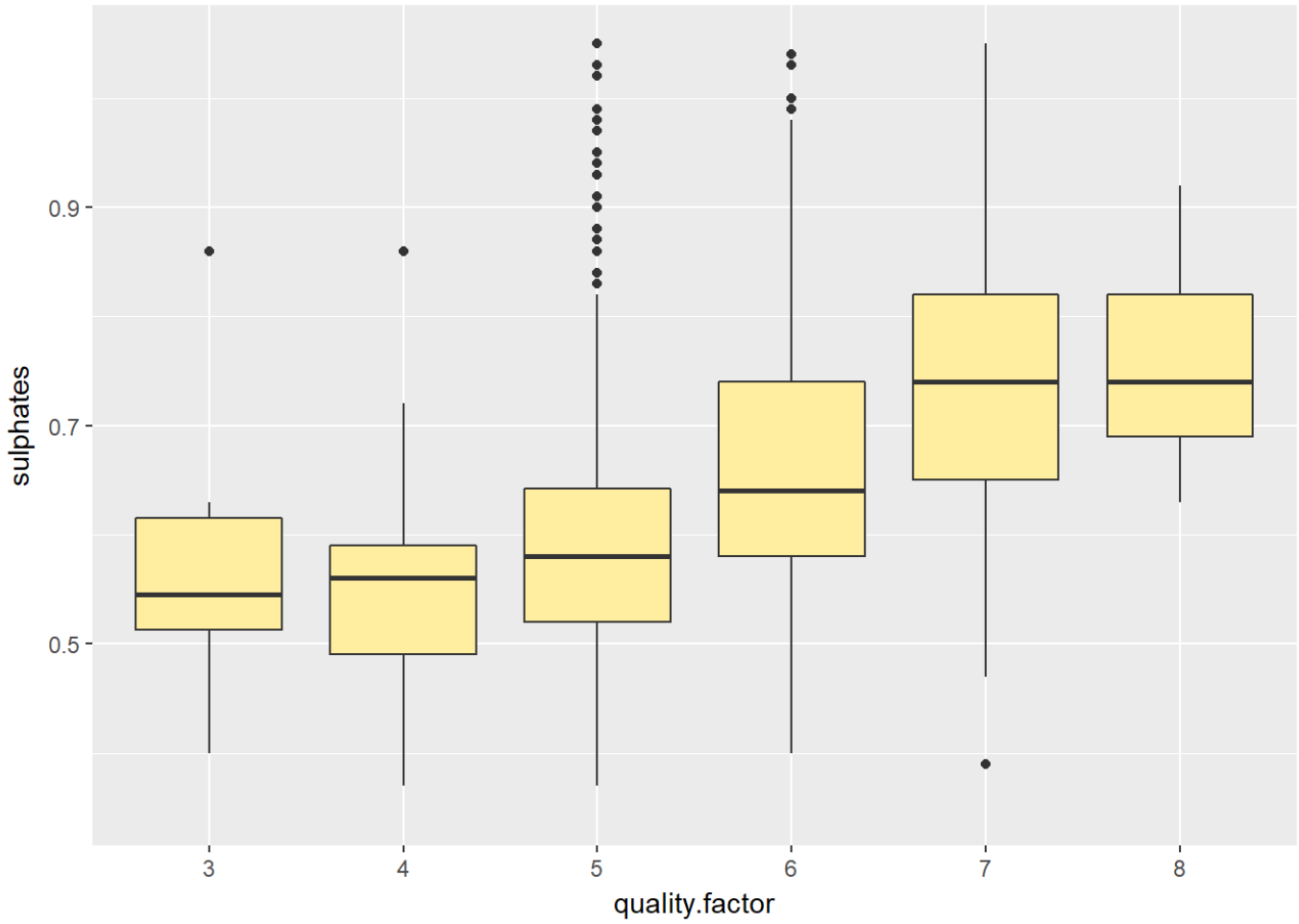
There are several interesting weak to moderatley strong correlations in the scatterplot matrix. I'll start by examining variables that correlate to quality.

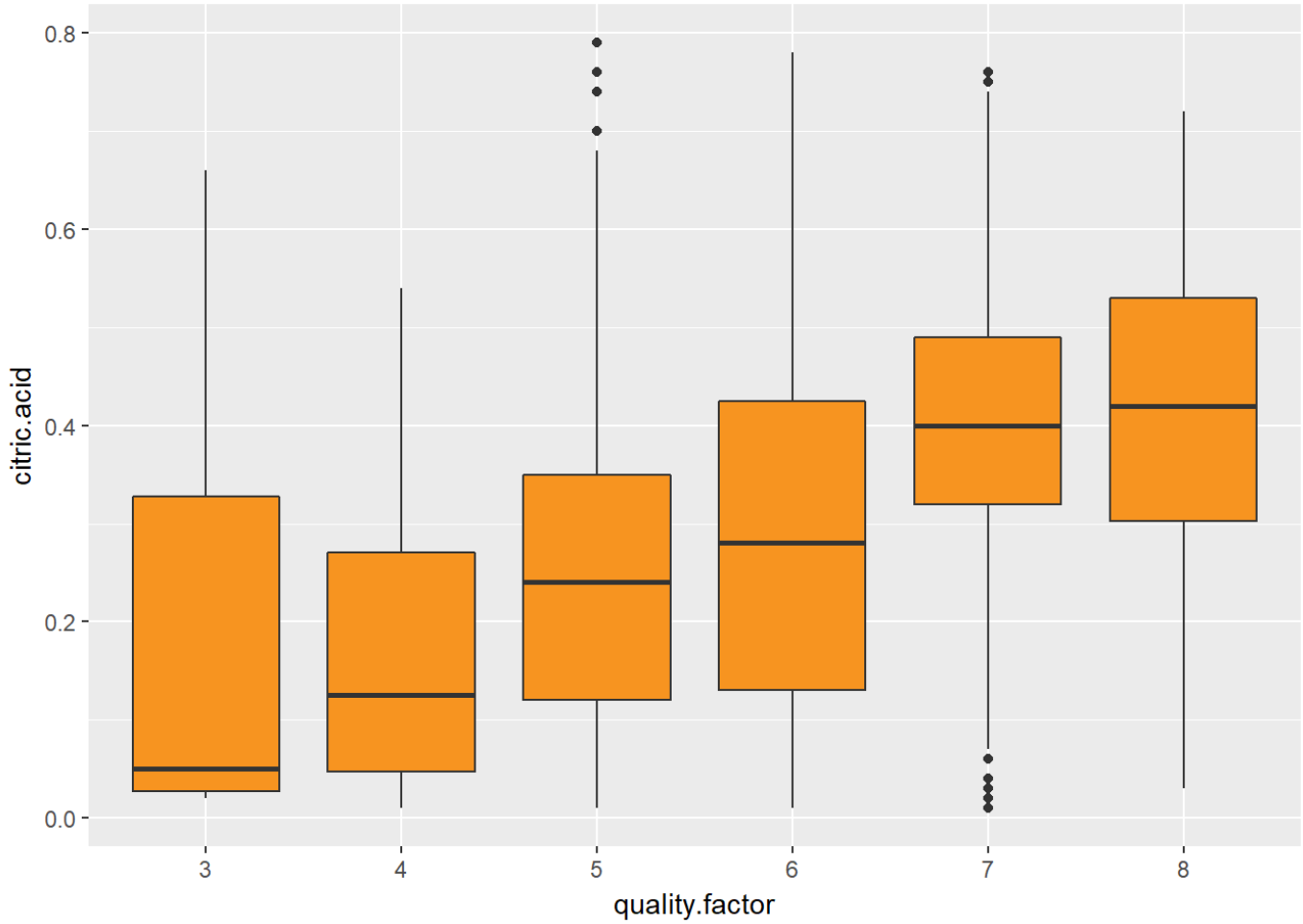


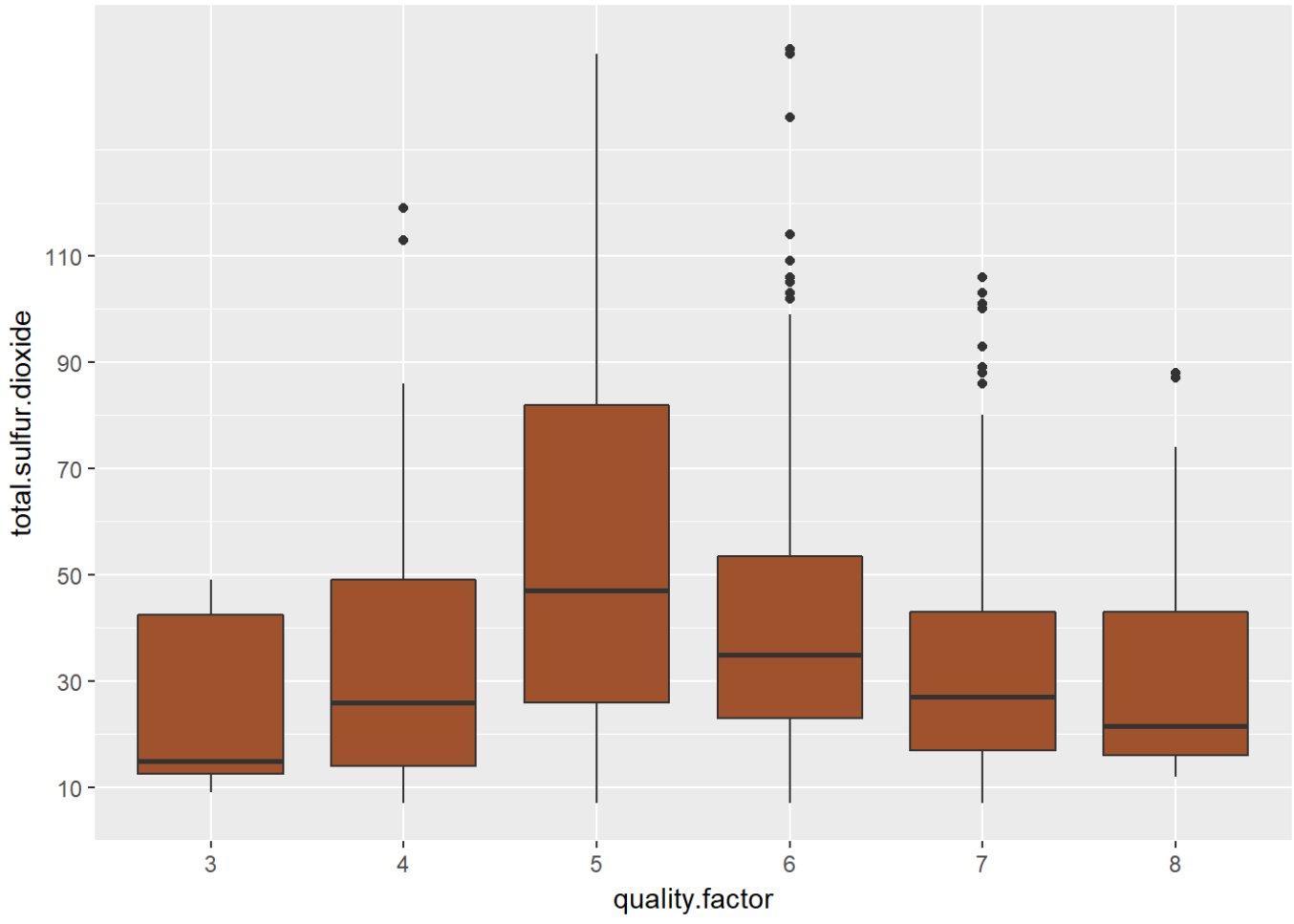
This boxplot shows a moderately strong relationship between quality and alcohol content. I don't know whether alcohol itself imputes quality, or possibly low alcohol is a marker for a problem in the wine-making process. There are a few higher-alcohol wines that rated middling, but no wines rated as high as 8 with an ABV below 9.8 percent.

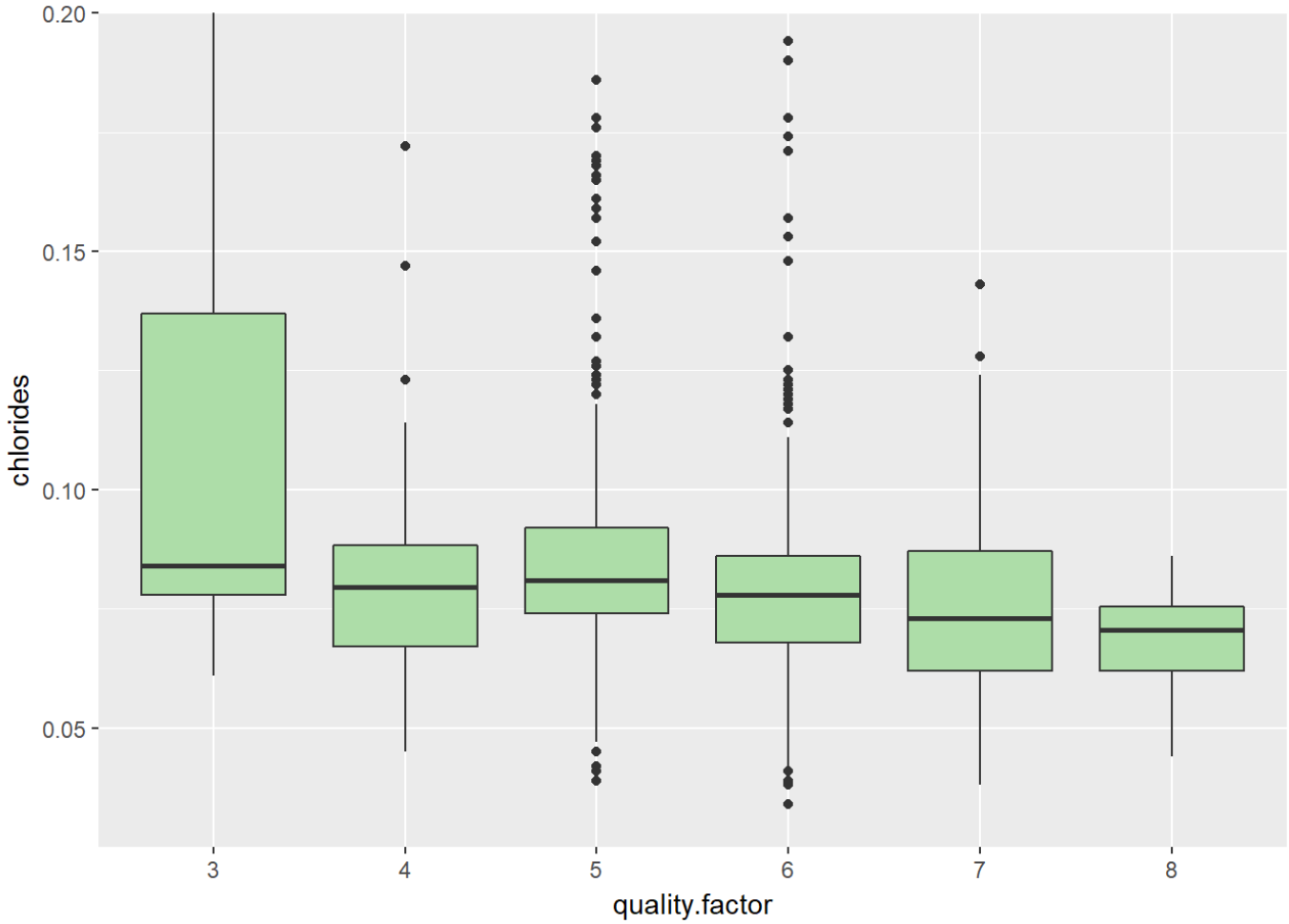
The next-highest correlations to quality were volatile acidity, sulphates, citric acid and total sulfur dioxide. I will look each one. I will zoom in a little bit on volatile acidity because has some distant outliers. For citric acid, I will omit the strange values at 0.00, 0.49 and 1.00.


1.6
1.2
0.8
0.4
volatile.acidity
3
4
5
6
7
8
quality.factor


0.9
0.7
0.5
sulphates
3
4
5
6
7
8
quality.factor

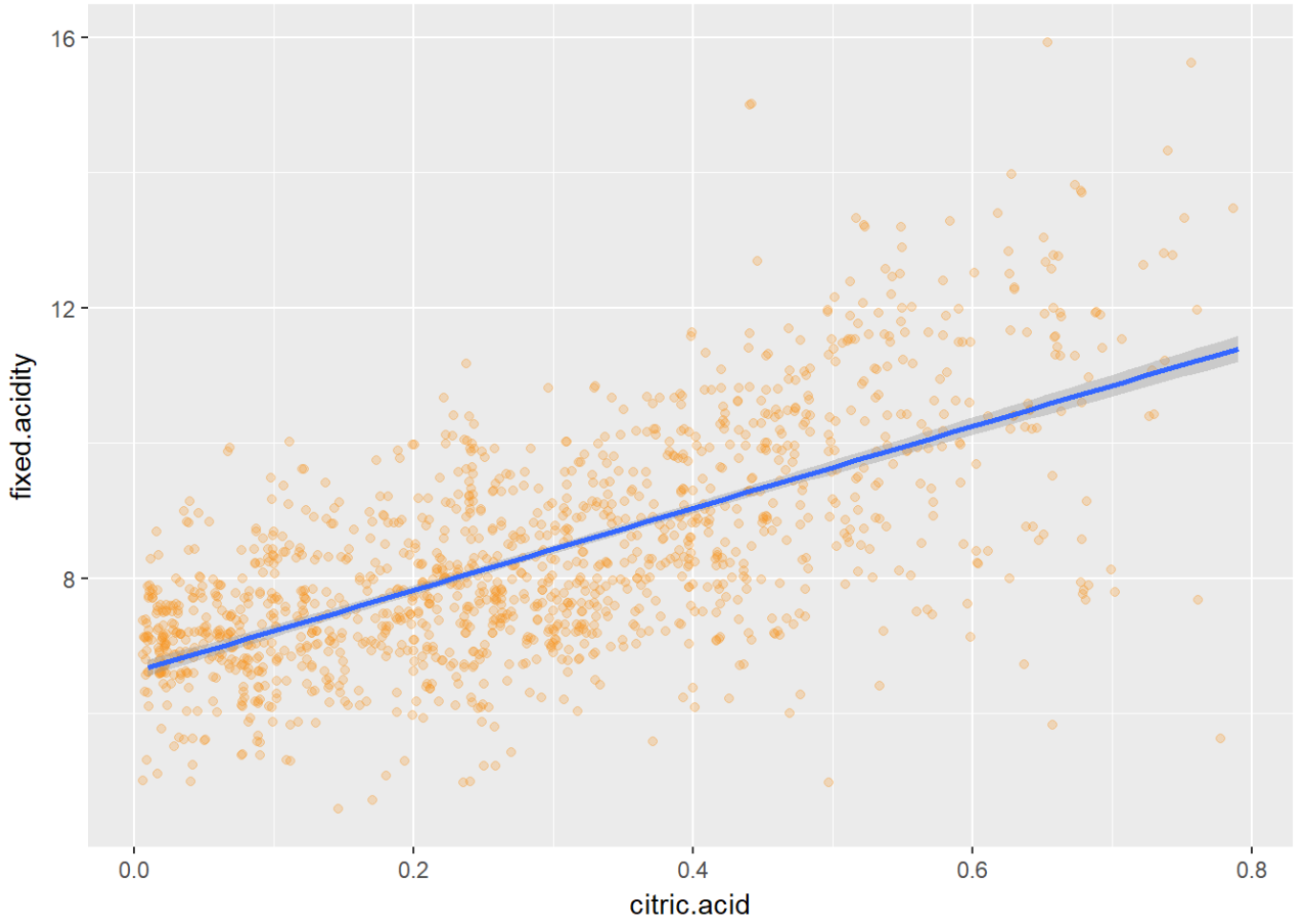

0.8
0.6
0.4
0.2
0.0
citric.acid
3
4
5
6
7
8
quality.factor

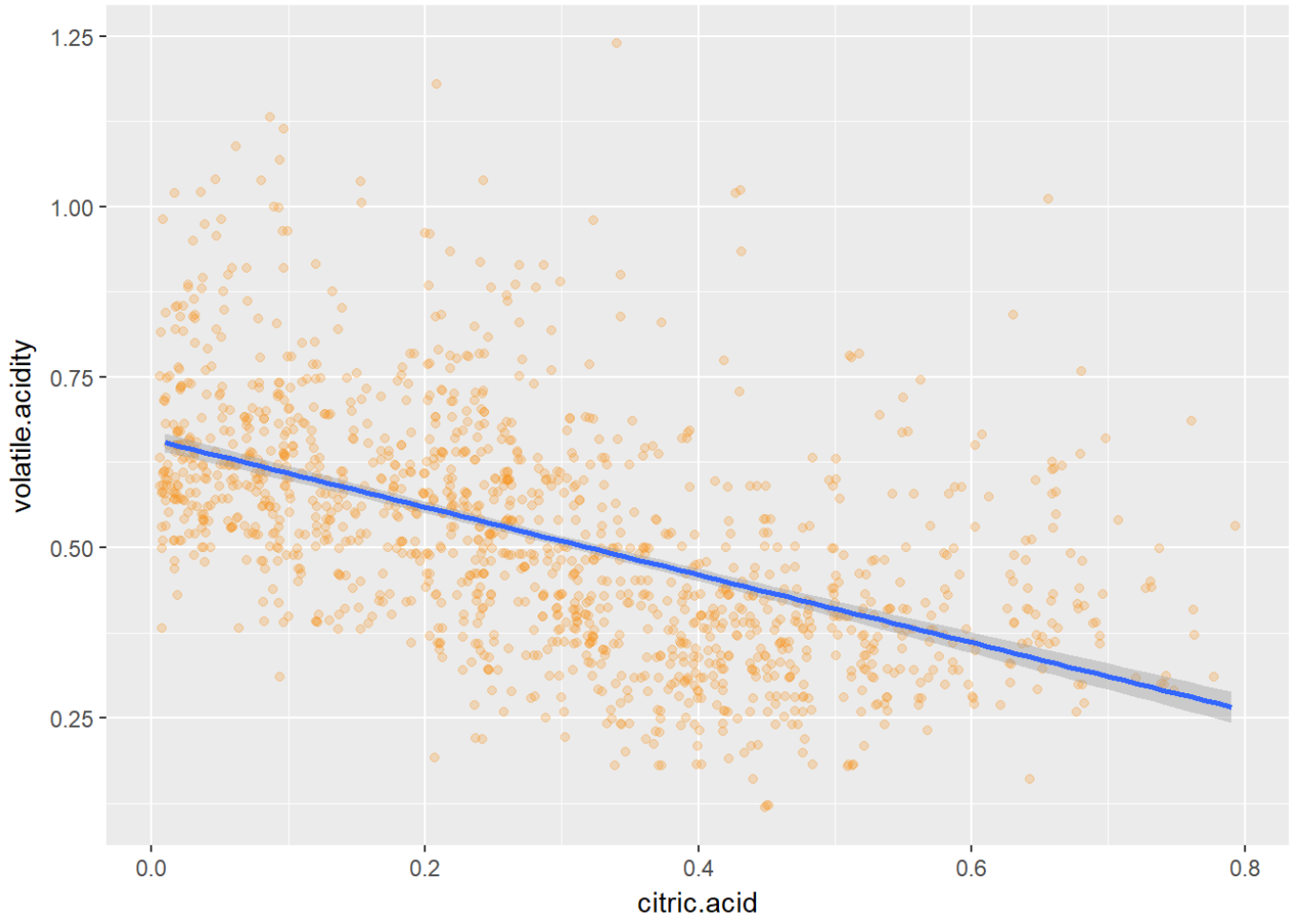

110
90
70
50
30
10
total.sulfur.dioxide
3
4
5
6
7
8
quality.factor



Volatile acidity correlates negatively to quality. This type of acidity could be related to vinegary flavors and smells. Sulphates actually correlate positively. Although we typically associate sulphates with inexpensive wine, they do serve a preservative function that could positively affect quality for a certain segment of the wine market. There is a weak positive correlation between citric acid and quality, which I expected. The lack of relation between total sulfur dioxide and quality is a bit surprising. I think of sulfur as bad trait in wine, but it actually peaks out at mid qualities rather than the poorest qualitiy of samples in this data set. I can almost imagine that sulfur dioxide being a by-product of wine-making, that is undesirable in itself but correlated to some other, desirable characteristic. Or, this distribution could just be related to the particular wines that Vinho Verde sells.

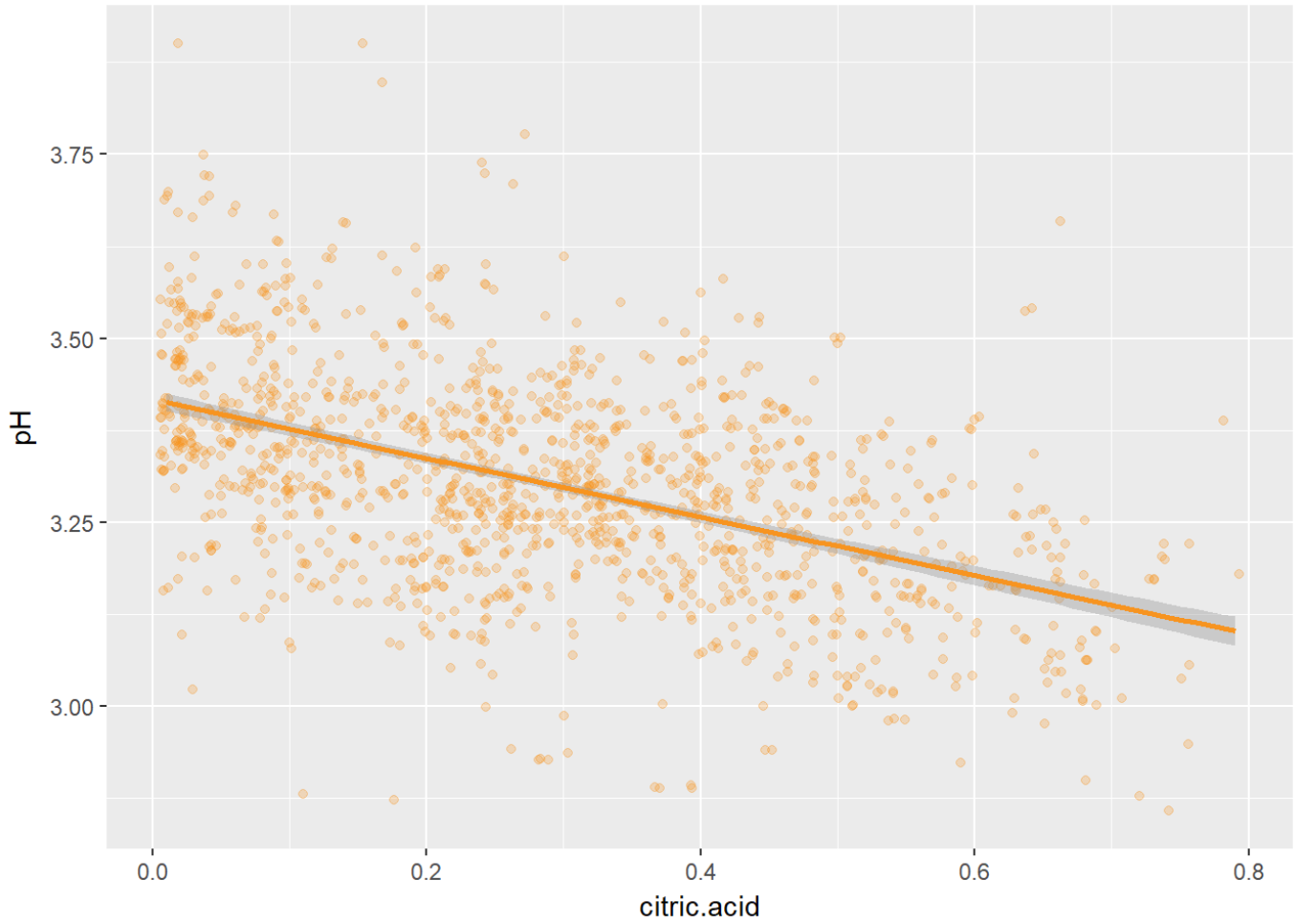
Next, I want to investigate the correlations between some of the physiochemical variablese themselves. It's possible some of these variables are related. If so, a combination or subset of them might provide the simplest model for wine quality.





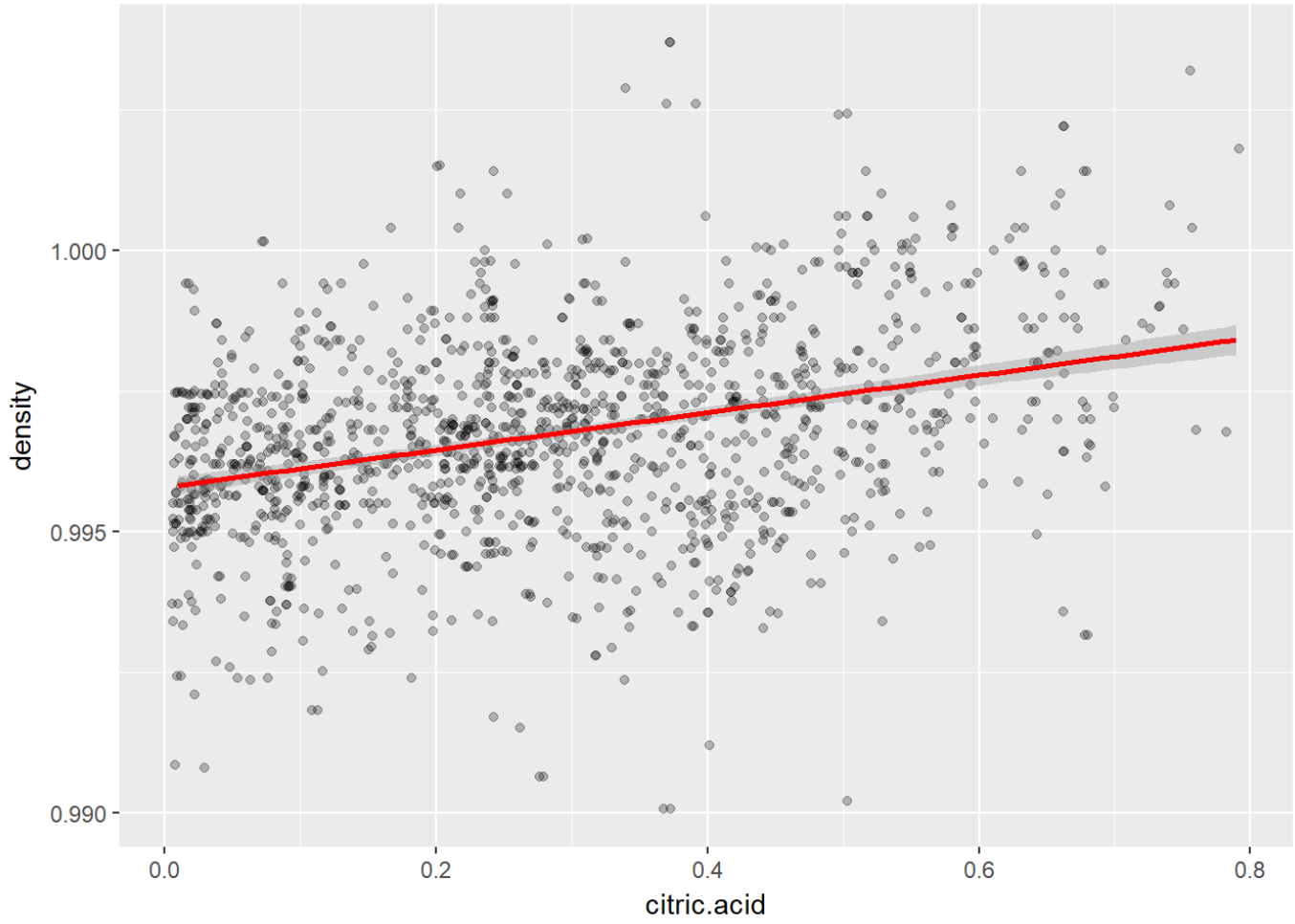
Fixed and volatile acidity have opposite-signed correlations to citric acid. I suspect citric acid contributes to overall acidity but is not highly volatile. Both of these correlations are only moderatly strong, so there may be

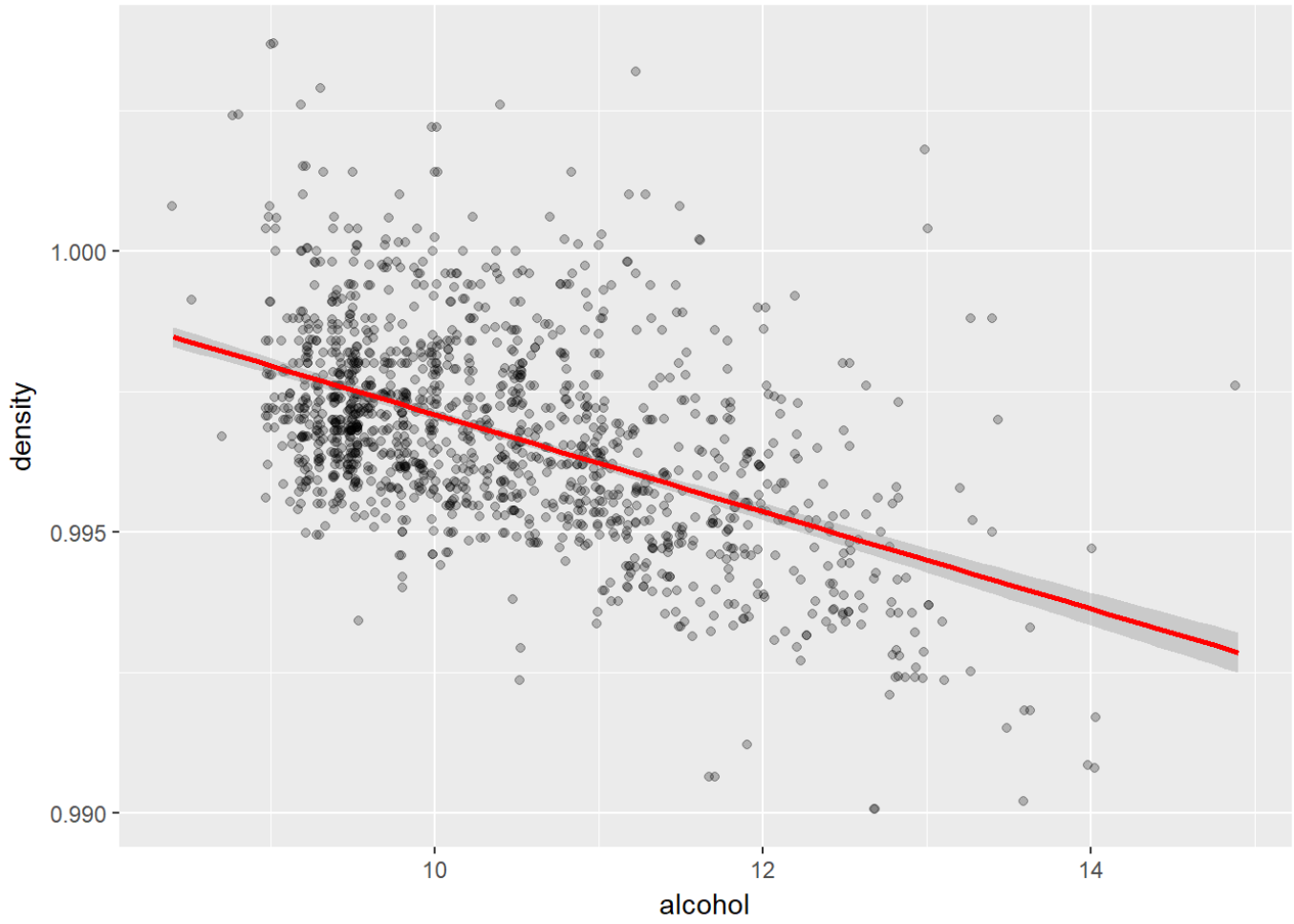
lurking variables at play who aren't in this dataset.



This correlation is unsurprising. Higher citric acid levels result in lower (more acidic) pH, but again there could be other factors at play.

Next, I want to look at density.

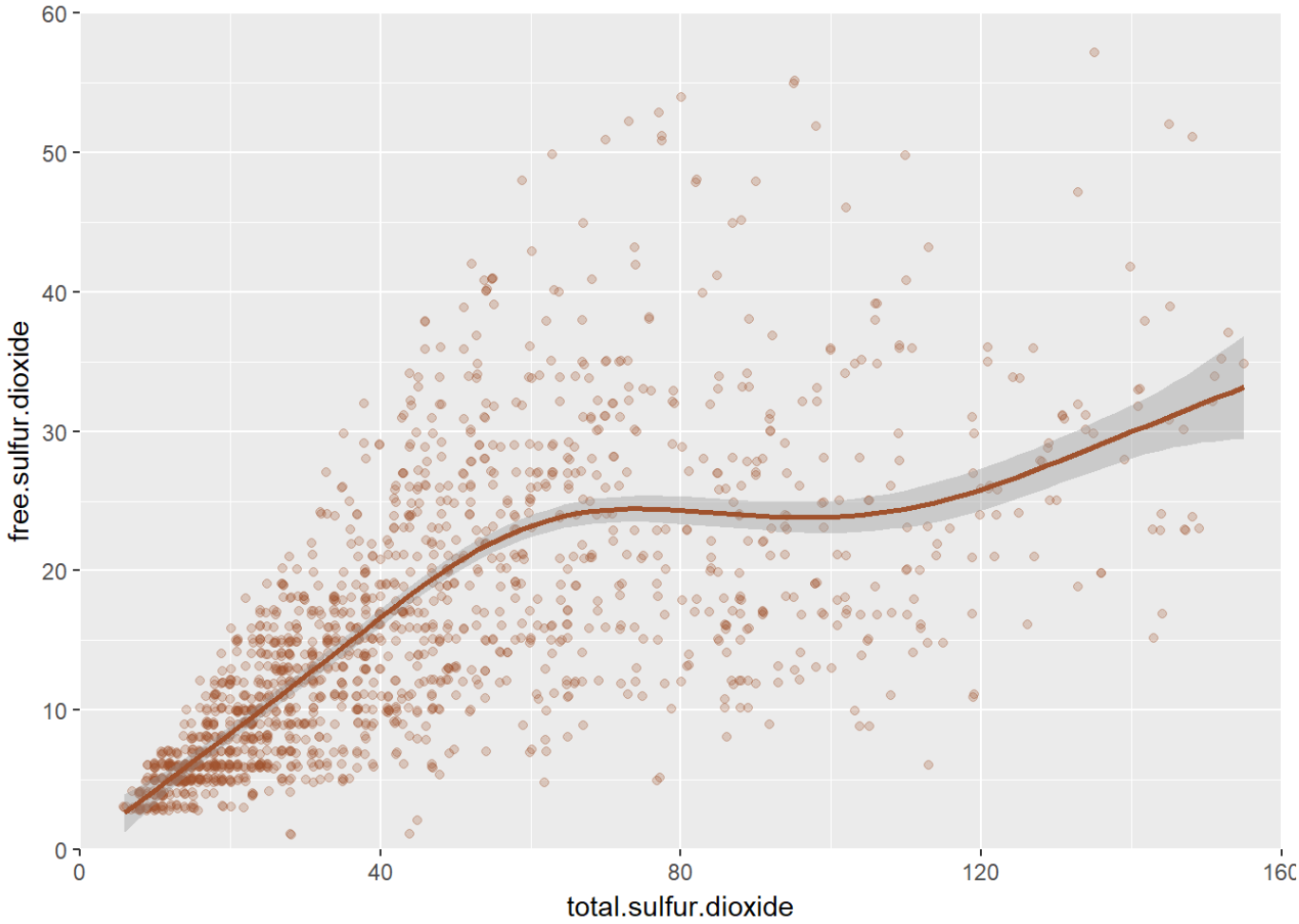




Density is moderately correlated with citric acid. This surprised me initially, but after looking it up I see that citric acid has a density of 1.66 g/cm^3, which is much higher than the typical overall density of wine, so a

little bit in solution can raise the overall density. I expected density to mainly be a proxy for alcohol content, since it is lighter than water. In any event, I doubt that the small overall range in density affects quality much directly.

Looking back at the scatterplot matrix, the strongest correlation was total sulfur dioxide to free sulfur dioxide, which I plot now.



The plot possibly indicates a nonlinear relationship. It looks like there may be a much lower limit on how much free sulfur dioxide exists in wine. Also since “free” is a subset of “total” sulfur dioxide. There cannot be any points above the slope = 1, intercept = 0 line on this plot.

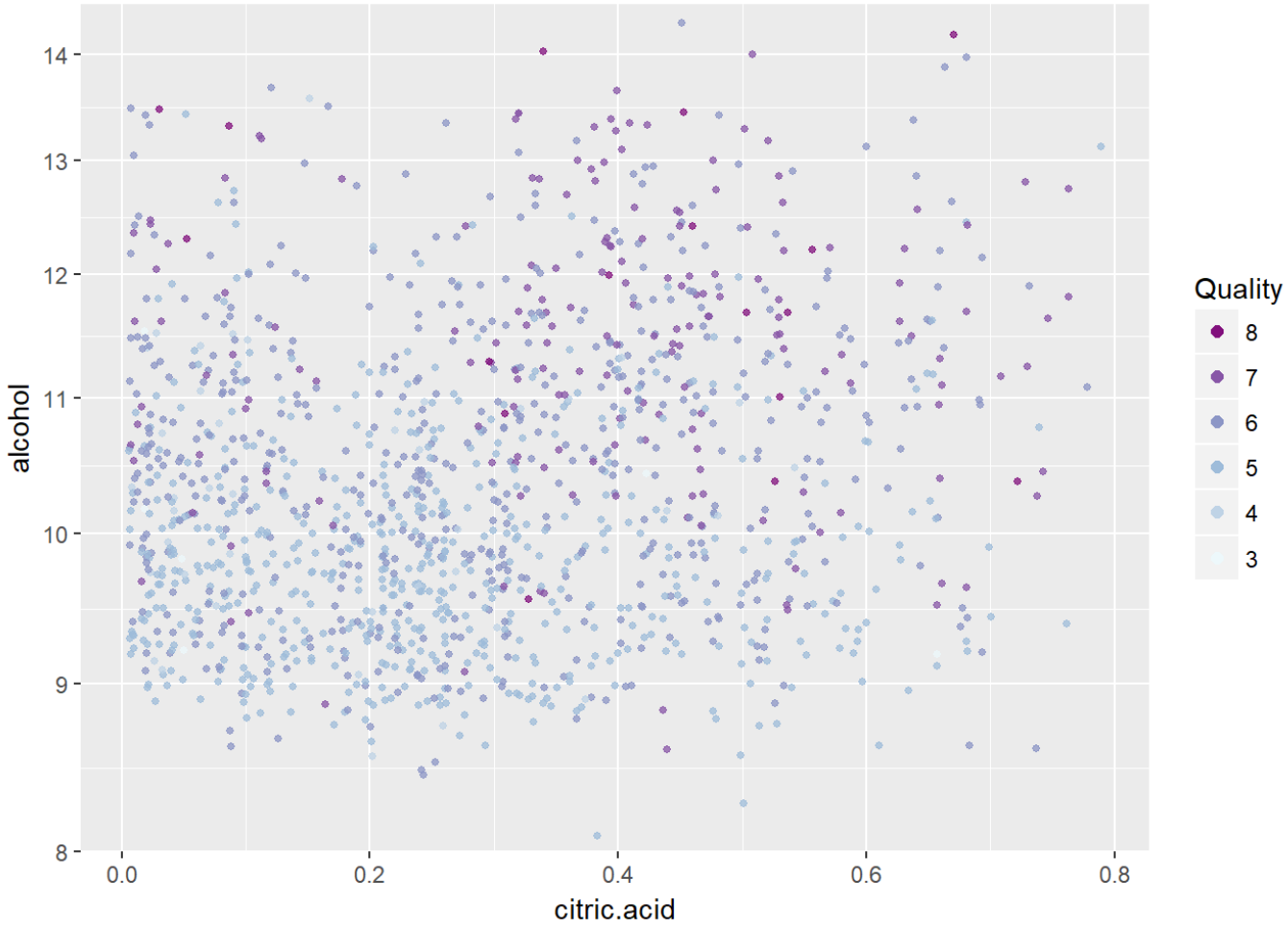
Multivariate Plots Section

Overall, the strongest correlation to wine quality was with alcohol by volume, but there were less strong correlations to volatile acidity, sulphates and citric acid. I want to make some multivariate plots using color scaling to depict quality, to investigate whether a combination of variables can explain quality better than any other single variable.

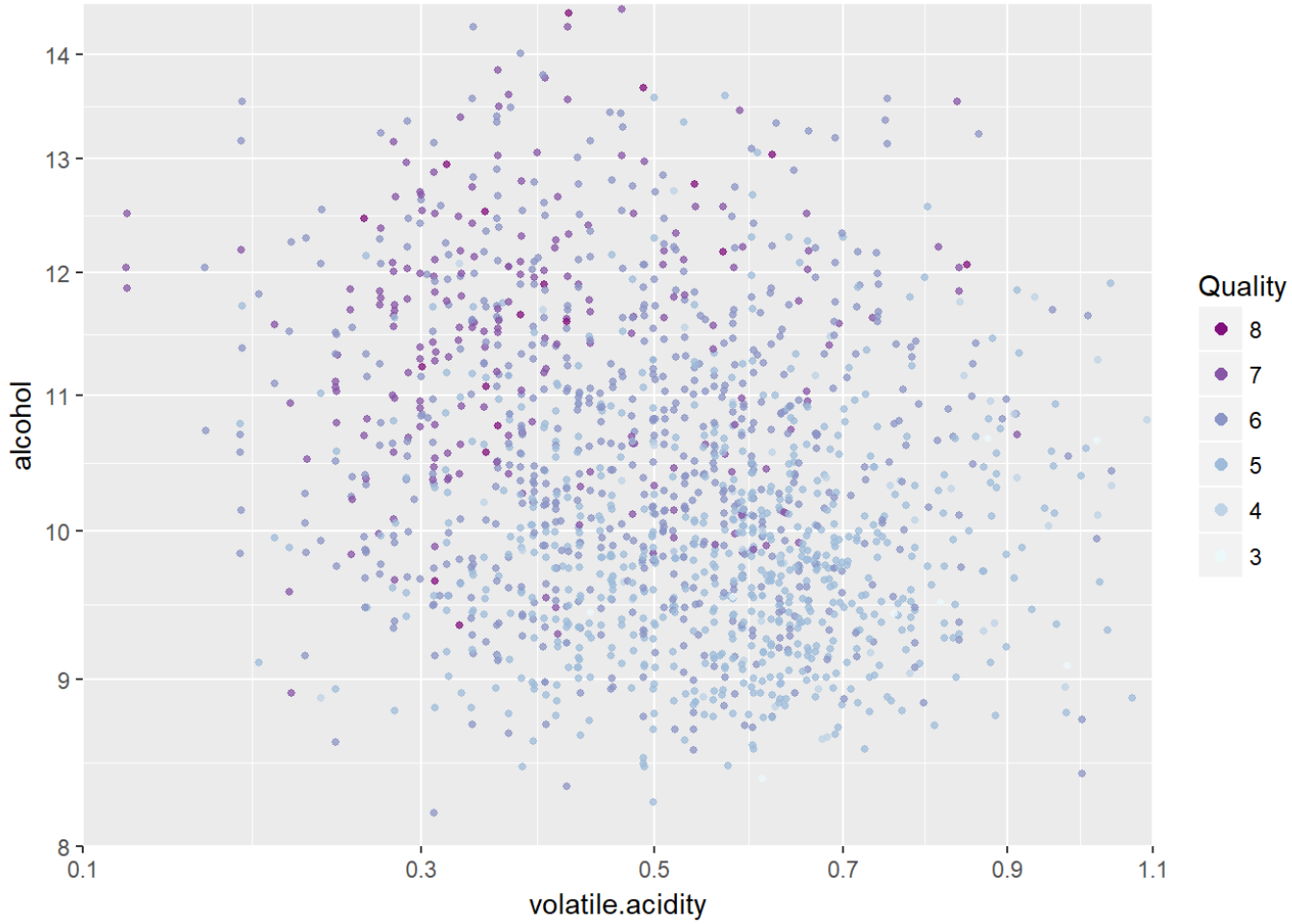
I also want to group quality scores of 3 and 4 together, since there were few.

```
##
##   3   4   5   6   7   8
##  10  53 681 638 199  18
```

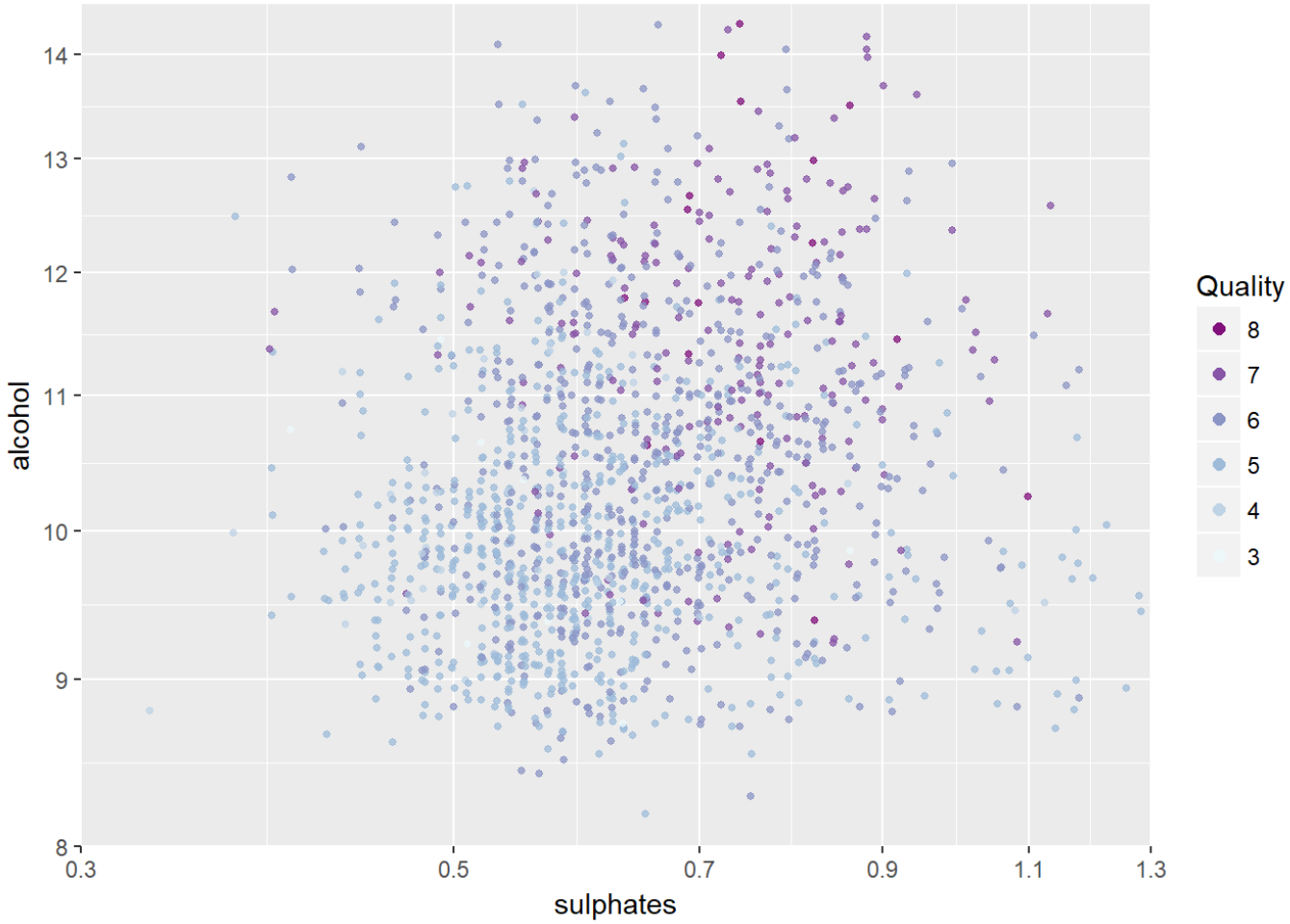
```
##
## (2,4] (4,5] (5,6] (6,7] (7,8]
##    63   681   638   199    18
```



This plot illustrates pretty well that quality is influenced by alcohol on the vertical axis, but citric acid on the horizontal axis seems to have a limited affect. Wines of quality 5, in green, tend to be grouped toward the bottom and somewhat left, while wines of higher quailty 7 and 8 are grouped in toward the top and somewhat righ. Wines with quality of 6 are pretty highly dispersed. There is definitely high variability of quality for a given citric acid content.



In this plot I have swapped out citric acid for volatile acidity on the horiz- ontal axis and changed to square-root axis scaling. We are seeing somewhat better color grouping. Alcohol continues to be a relevant factor, but with volatile acidity, we see the lower quality wines are grouped in the lower middle-to-right versus higher quality wines in the upper left. This is because there was an inverse correlation between volatile acidity and quality.



Now we are looking at quality versus ABV and sulphate concentrations. It seems sulphates also explain some of the variability in wine quailty at a given alcohol content.

Multivariate Analysis

There are a few physiochemical properties that are at least wealky correlated to the quality of red wine from Vinho Verde. Higher concentrations of alcohol and sulphates tend associate with improved quality, whereas higher volatile acidity associates with poorer quality. These findings are mostly consistent with conventional wisdom. The main surprise here was the moderately strong correlation of alcohol content to wine quality.

These plots suggest that a linear model will only very roughly predict the quality of wine. The results of my model are summarized below.

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

Yes, I created a linear model starting from wine quality and log of alcohol content by volume. The variables in the linear model account for 35.4 percent of the variance in quality of Vinho Verde red wines. The addition of square-root of volatile acidity, log10 of sulphates, and total sulfur dioxide concentrations improved the R^2 value by 11.5 percent. Citric acid did not improve the model appreciably.

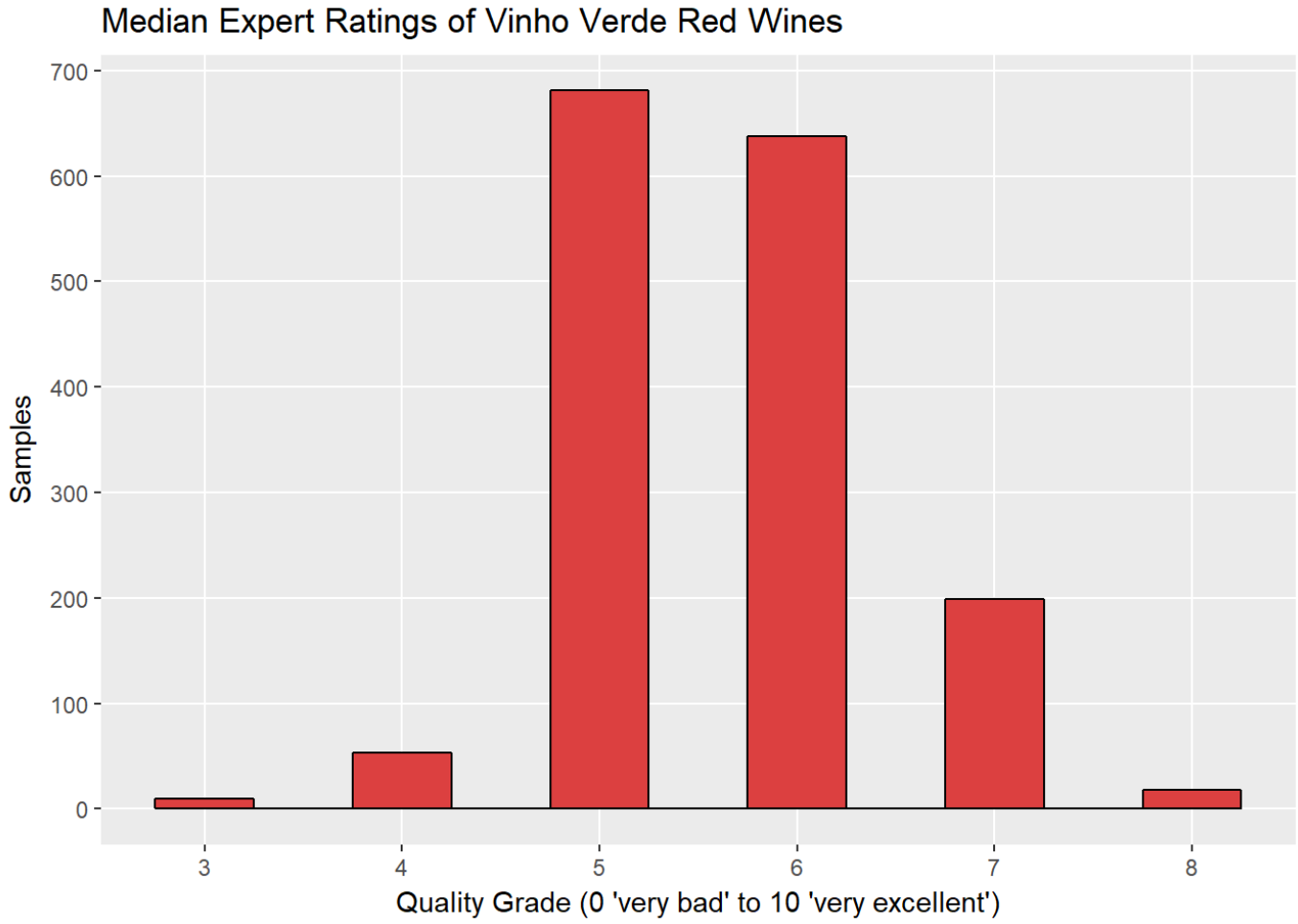
```
##
## Calls:
## m1: lm(formula = I(quality) ~ I(log10(alcohol)), data = wine.tidy)
## m2: lm(formula = I(quality) ~ I(log10(alcohol)) + sqrt(volatile.acidity),
##     data = wine.tidy)
```

```
## m3: lm(formula = I(quality) ~ I(log10(alcohol)) + sqrt(volatile.acidity) +
##     log10(sulphates), data = wine.tidy)
## m4: lm(formula = I(quality) ~ I(log10(alcohol)) + sqrt(volatile.acidity) +
##     log10(sulphates) + citric.acid, data = wine.tidy)
## m5: lm(formula = I(quality) ~ I(log10(alcohol)) + sqrt(volatile.acidity) +
##     log10(sulphates) + citric.acid + total.sulfur.dioxide, data = wine.tidy)
##
## ============================================================================
================
##                                m1            m2            m3            m4
m5
## ----------------------------------------------------------------------------
------------------
##   (Intercept)                -3.696***     -1.070*       -0.737        -0.738
-0.348
##                              (0.448)       (0.475)       (0.466)       (0.476)
(0.483)
##   I(log10(alcohol))           9.217***      7.948***      7.674***      7.674**
*      7.328***
##                              (0.441)       (0.431)       (0.422)       (0.422)
(0.429)
##   sqrt(volatile.acidity)                   -1.888***     -1.550***     -1.549**
*     -1.462***
##                                            (0.152)       (0.154)       (0.176)
(0.176)
##   log10(sulphates)                                        1.538***      1.538**
*      1.553***
##                                                         (0.187)       (0.192)
(0.191)
##   citric.acid                                                            0.001
0.028
##                                                                        (0.115)
(0.115)
##   total.sulfur.dioxide
-0.002***
##
(0.001)
## ----------------------------------------------------------------------------
------------------
##   R-squared                   0.239         0.314         0.346         0.346
0.354
##   adj. R-squared              0.238         0.313         0.345         0.344
0.351
##   sigma                       0.701         0.666         0.651         0.651
0.647
##   F                         437.656       319.862       245.864       184.266
152.298
##   p                           0.000         0.000         0.000         0.000
0.000
##   Log-likelihood          -1487.013     -1413.782     -1380.771     -1380.770
-1372.623
##   Deviance                  686.951       618.624       590.088       590.088
583.251
##   AIC                      2980.025      2835.563      2771.541      2773.541
2759.247
##   BIC                      2995.754      2856.534      2797.755      2804.998
2795.946
```

```
##   N                         1398          1398          1398          1398
1398
## ============================================================================
===============
```

Final Plots and Summary

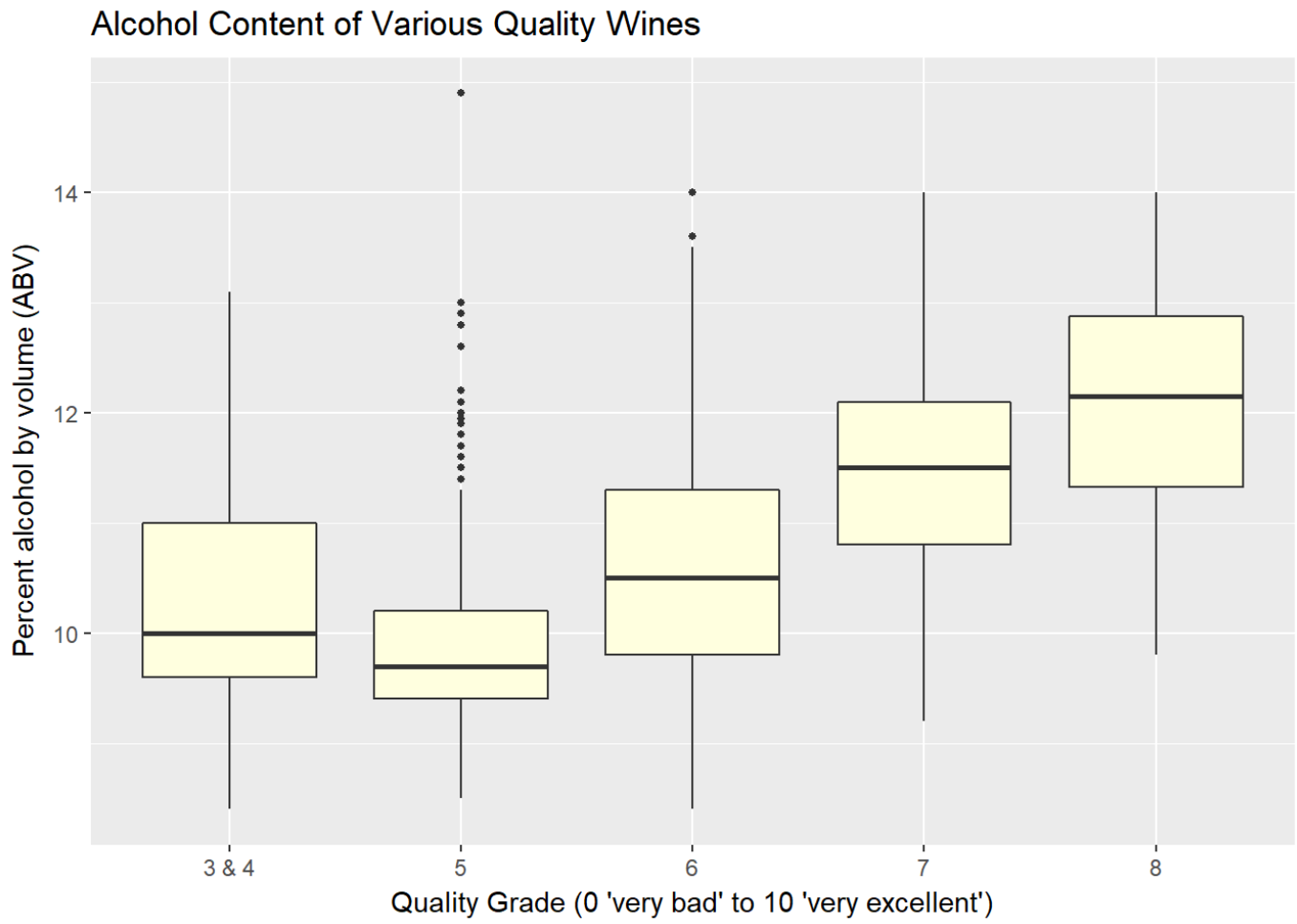
Plot One



Description One

The distribution of Vinho Verde red wine quality is roughly binomial with a strong central tendency at quality at 5 to 6, indicating consistent wine quality. There some higher quality wines in the sampling, which perhaps *could* have been rebranded for higher profit. This could also just be due to a degree of subjectivity in the wine expert grading. To their credit, Vinho Verde produced relatively few (63 out of 1599) lower quality wines in this sample.

Plot Two

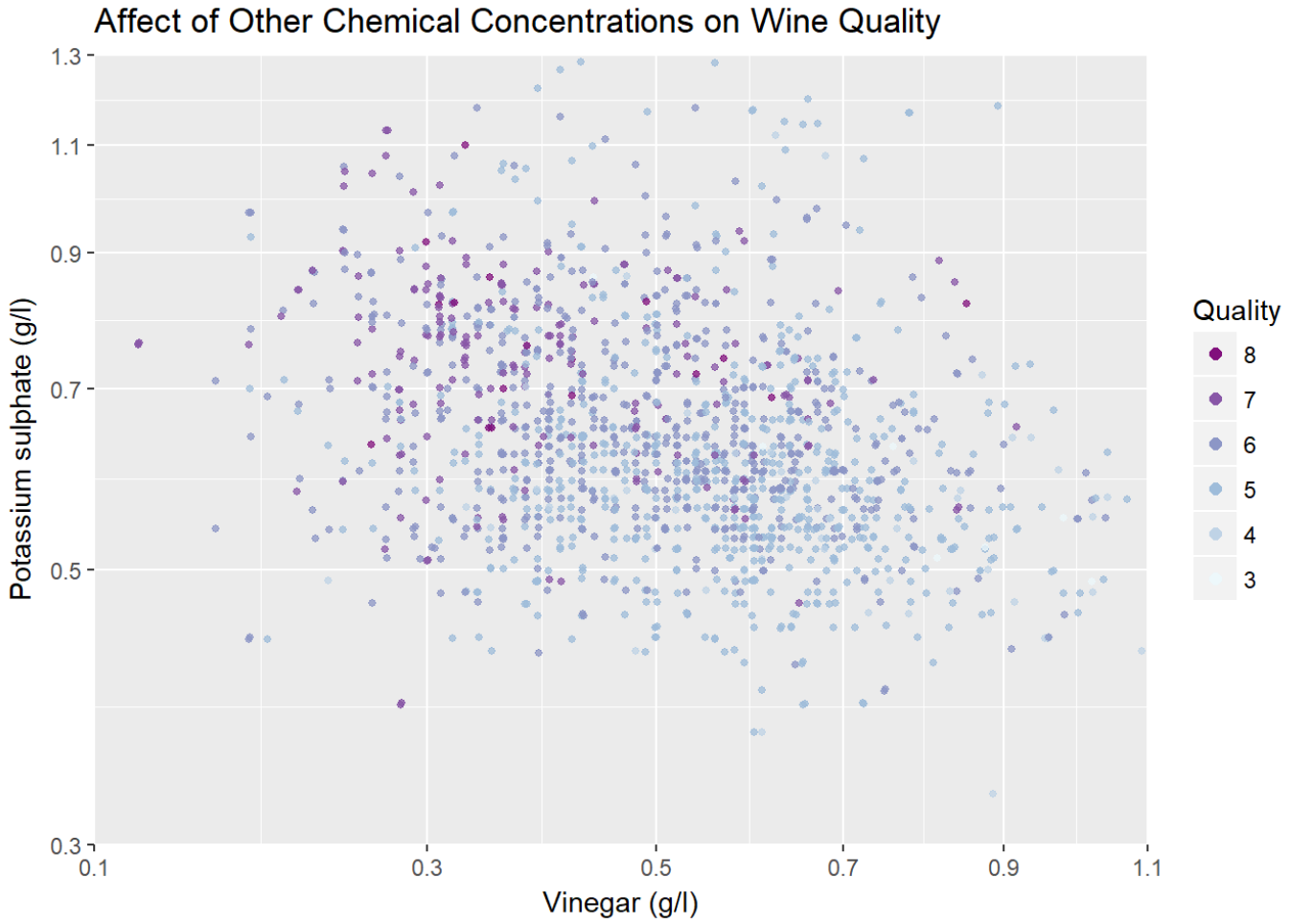


Description Two

These boxplots shows a moderately strong relationship between quality and alcohol content. For wines that graded at least 5, there is a clear march upward in quality as average alcohol content increases. This is somewhat surprising, but it may be a useful indication of market preference for stronger wines. At least, to the extent that wine expert preferences reflect the typical wine consumer.

The relationship to alcohol content breaks-down for wines with poorer qualities of 3 or 4. These wines could be offensive for reasons other than their lack of ‘punch’.

Plot Three



Description Three

At any given alcohol content, wine quality is affected by other chemicals. Here, we see that higher volatile acidity (vinegar concentration) tends to lower the quality of the wine. Although we don't typically think of sulphates in excellent wine, we can see that, for the Vinho Verde market, higher sulphate content can improve the perceived quality of the wine.

Reflection

The red wine data set contains information on almost 1,600 samples across 12 variables. I used the help file to understand the individual variables in the data set, and I researched Vinho Verde online. Then, I explored the data by plotting individual distributions and bivariate correlations. Eventually, I explored the quality of wine across many variables and tested the value of a linear model to predict wine quality.

There was a moderately strong correlation between alcohol content and perceived wine quality. There were weak correlations between quality and a few secondary chemical properties of the wine. After transforming alcohol percent by to log10 values and including the volatile acidity, log10 of sulphates, and total sulfur dioxide concentrations, the linear model account for 35.4 percent of the variance in quality of Vinho Verde red wines. This is not a great result and reinforces the continued value of expert wine evaluations.

To better understand how physiochemical properties influence wine quality, we could conduct a larger sampling from more vendors in the future. If we are primarily interested in the performance of Vinho Verde wines, it would be intersting to collect price data (or pursue an NDA). It may be that the varying quality of wine samples is justified by their relative pricing.