Exploratory Analysis of Red Wine Quality by Venkat Ramesh

The Red Wine Quality dataset contains quality ratings on a scale of 0 to 10, accompanied by differetnt attributes of red wine.

Below is a snapshot of the various columns in the dataset and their datatypes:

```
## [1] 1599   13
```

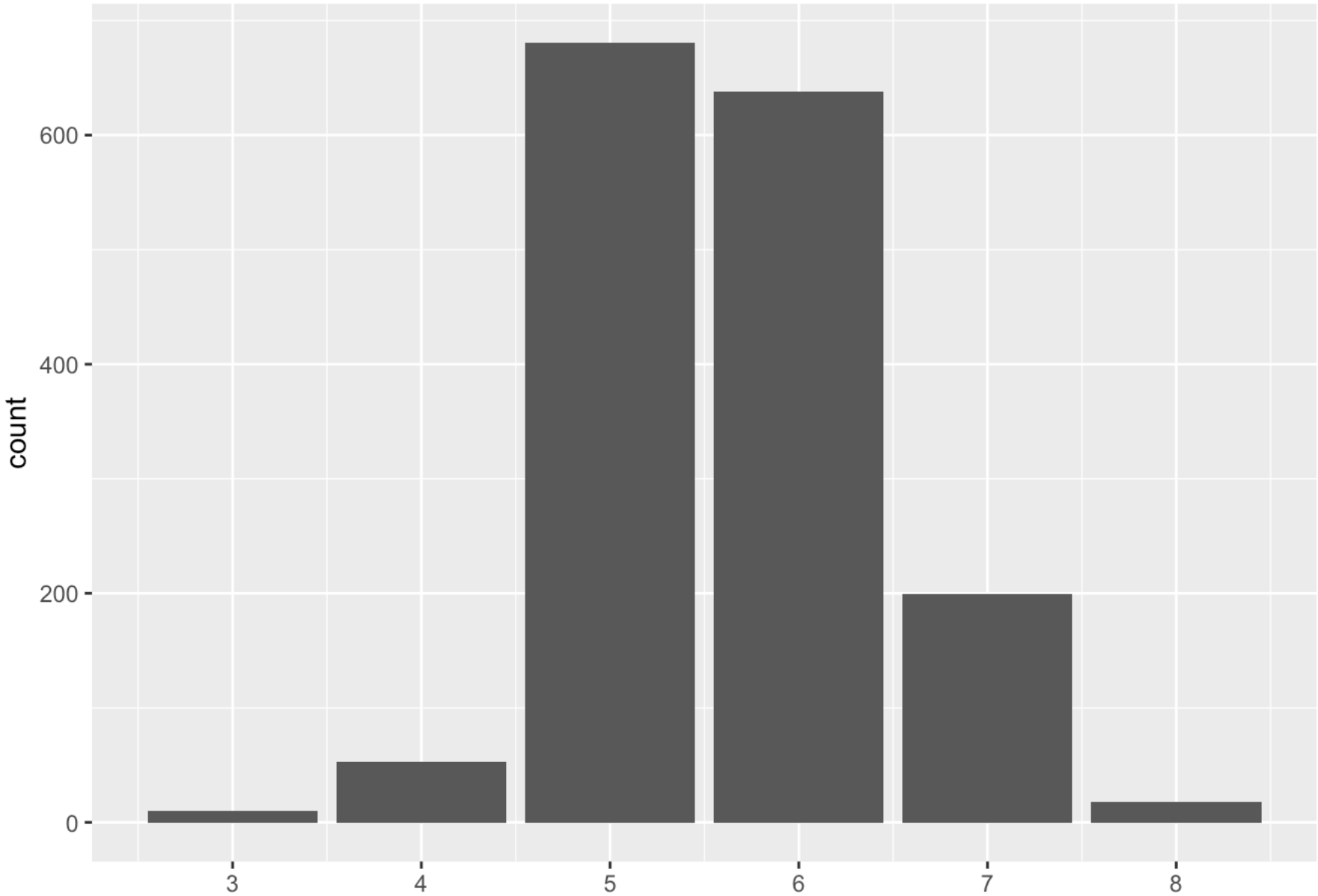
```
## 'data.frame':    1599 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  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 0.07
3 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 ...
```

Univariate Plots Section

Let’s explore each of the columns individually, to begin with.

Quality

Quality is the label column of interest that contains the integer score between 0 and 10. Here’s the histogram of quality ratings in the dataset.



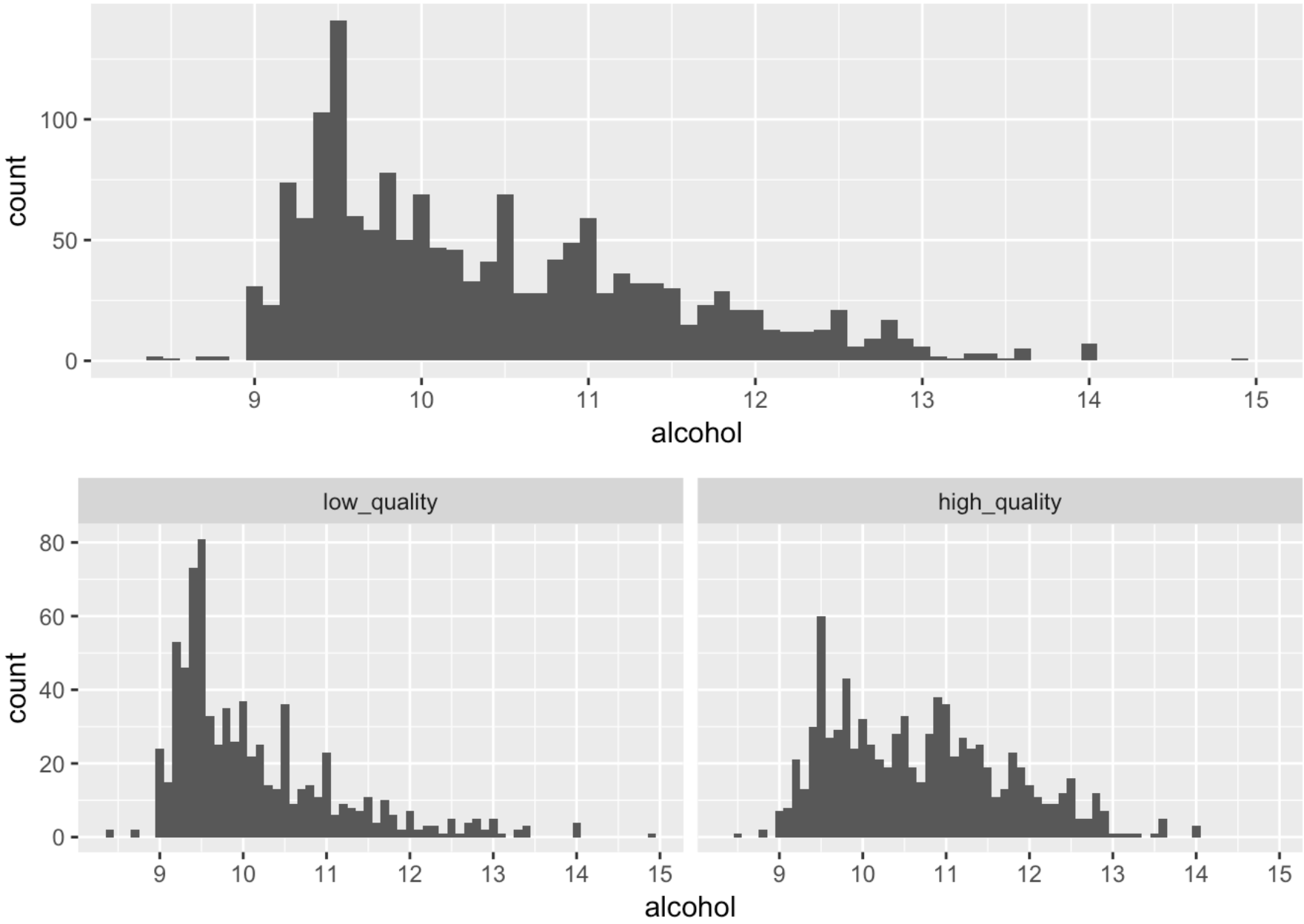
We see a nice normal distribution with most of the quality scores centered around 5. Subsetting the data into those with quality less than 5 and greater than 5 will help us see trends, since it also makes sense intuitively to place the pass score for quality at, greater than 5. We also notice that this splits our data into roughly two equal halves.

```
##
##  low_quality high_quality
##          744           855
```

When we look at our attributes, we will also generate a faceted version of the attribute histogram to get a peek ahead on the interaction of this attribute to th quality. Since all of the columns are of decimal datatype, we will pick 0.1 as our binwidth to be able to capture with enough granularity, unless we encounter a case where the range is too big or two small for this value.

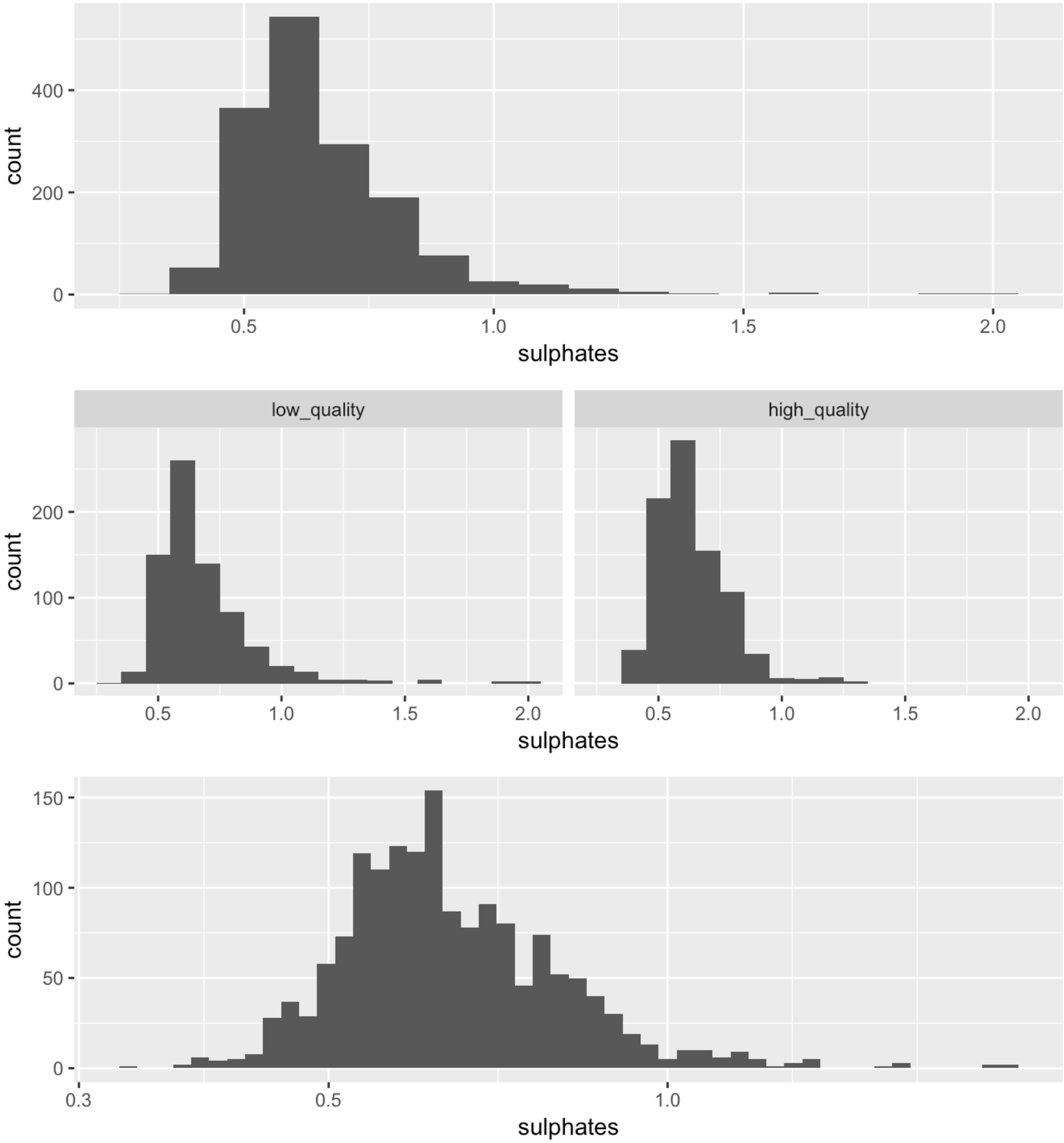
If the distribution deviates significantly from normal distribution and is long tailed, we will apply log10 transformation to the x axis to get a better picture.

Alcohol



We see that low-quality group has long-tailed distribution for alcohol levels, whereas the high-quality group appears to be more uniform. We will revisit this to investigate the correlation between quality and alcohol in our bivariate analysis.

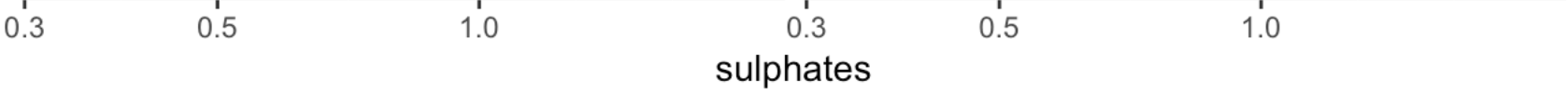
Sulphates


count
400
200
0
0.5
1.0
1.5
2.0
sulphates

low_quality
high_quality
count
200
100
0
0.5
1.0
1.5
2.0
sulphates

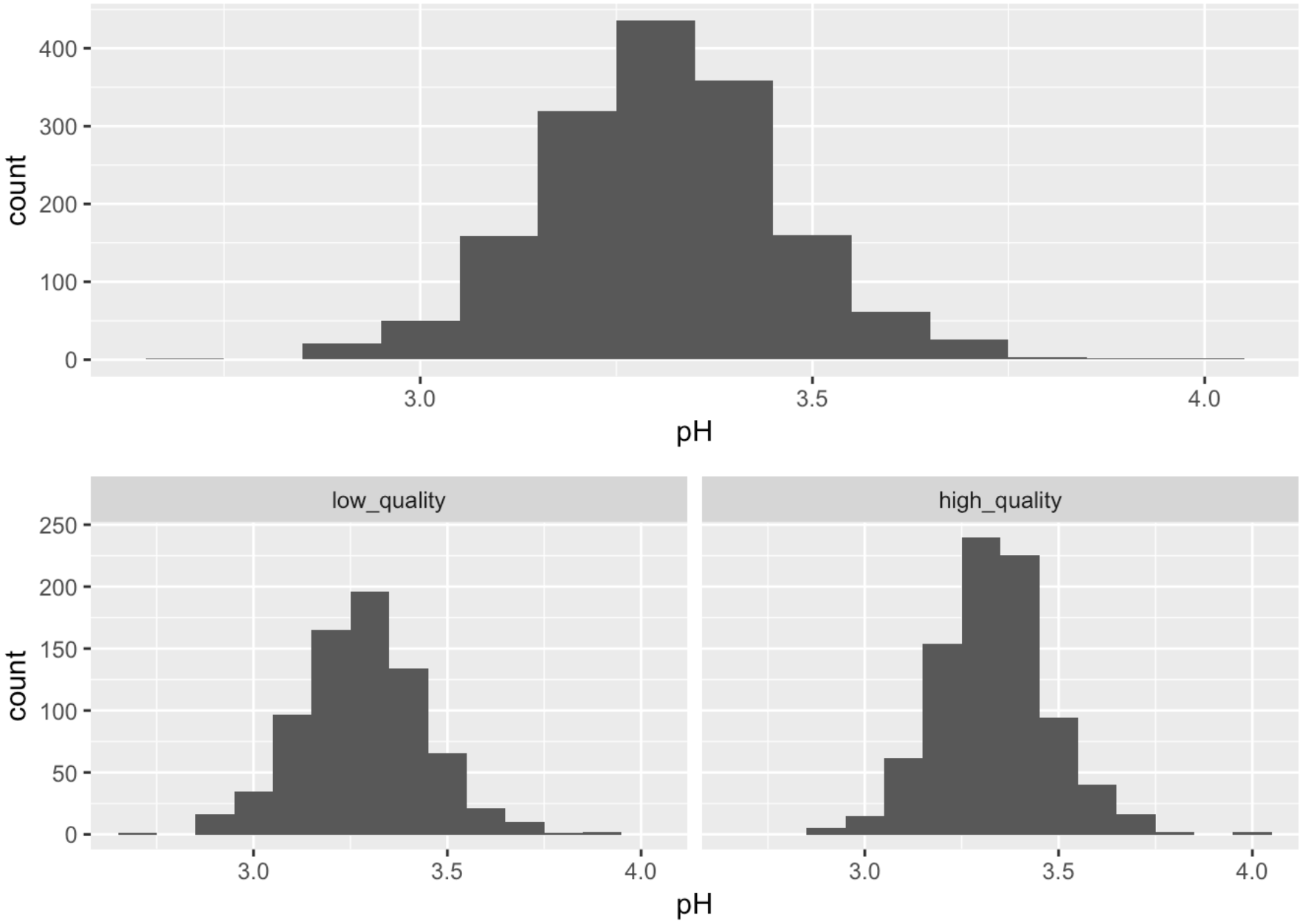
count
150
100
50
0
0.3
0.5
1.0
sulphates

low_quality
high_quality
count
80
60
40
20
0



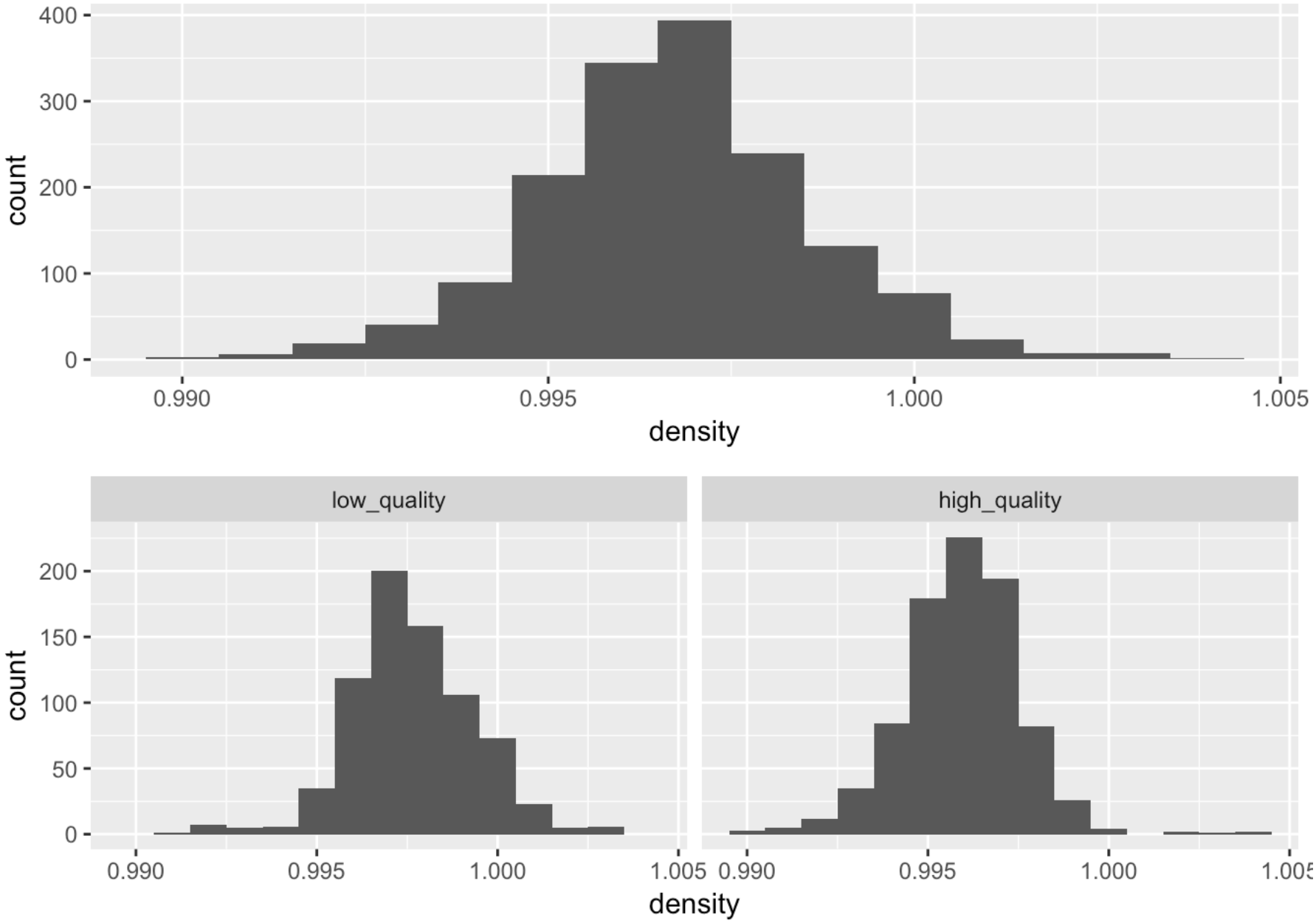
The distribution seems right skewed, but log10 transformation shows nothing different in the high quality vs low quality plots.

pH



The distribution looks normal so no transformation is needed.

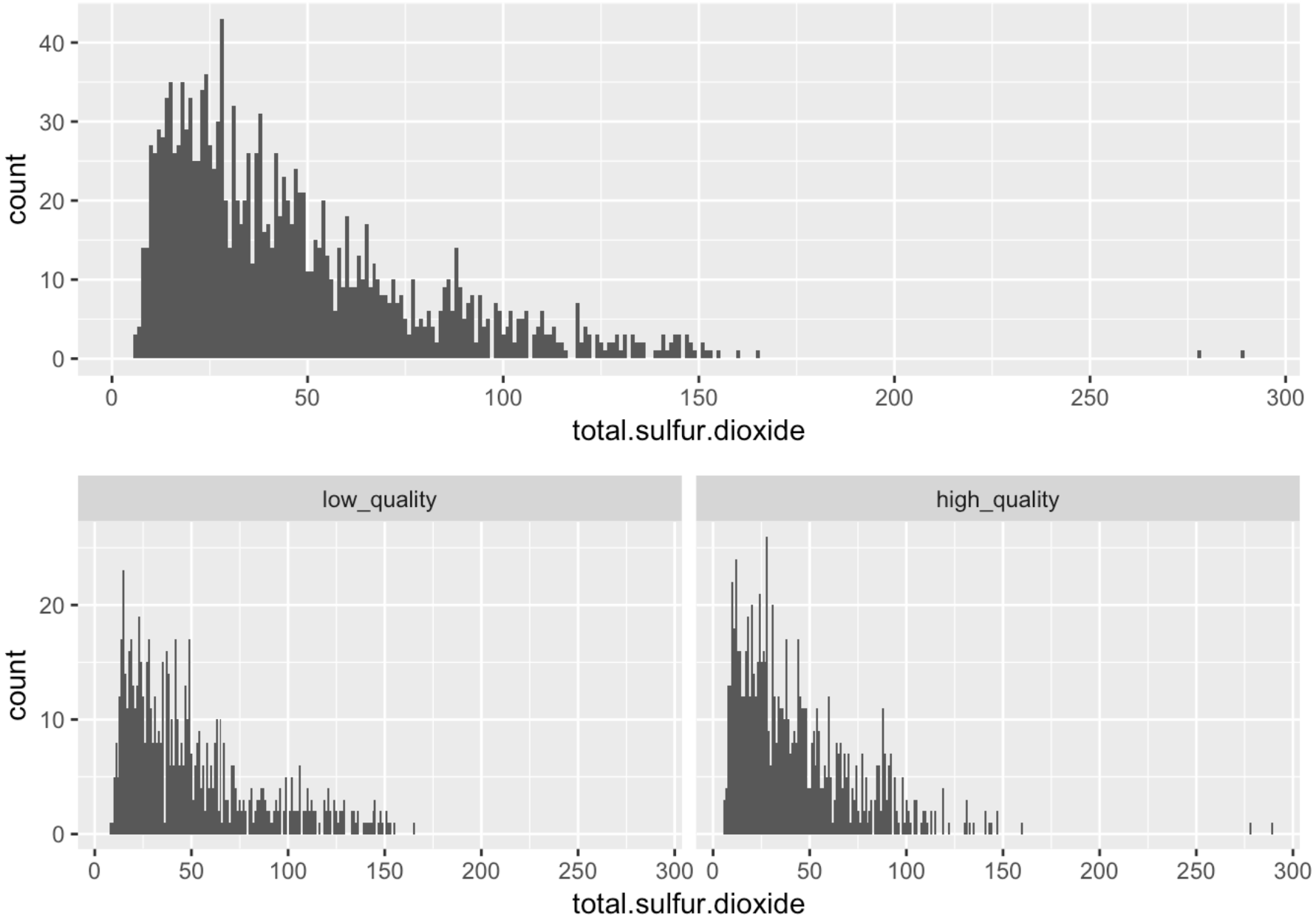
Density

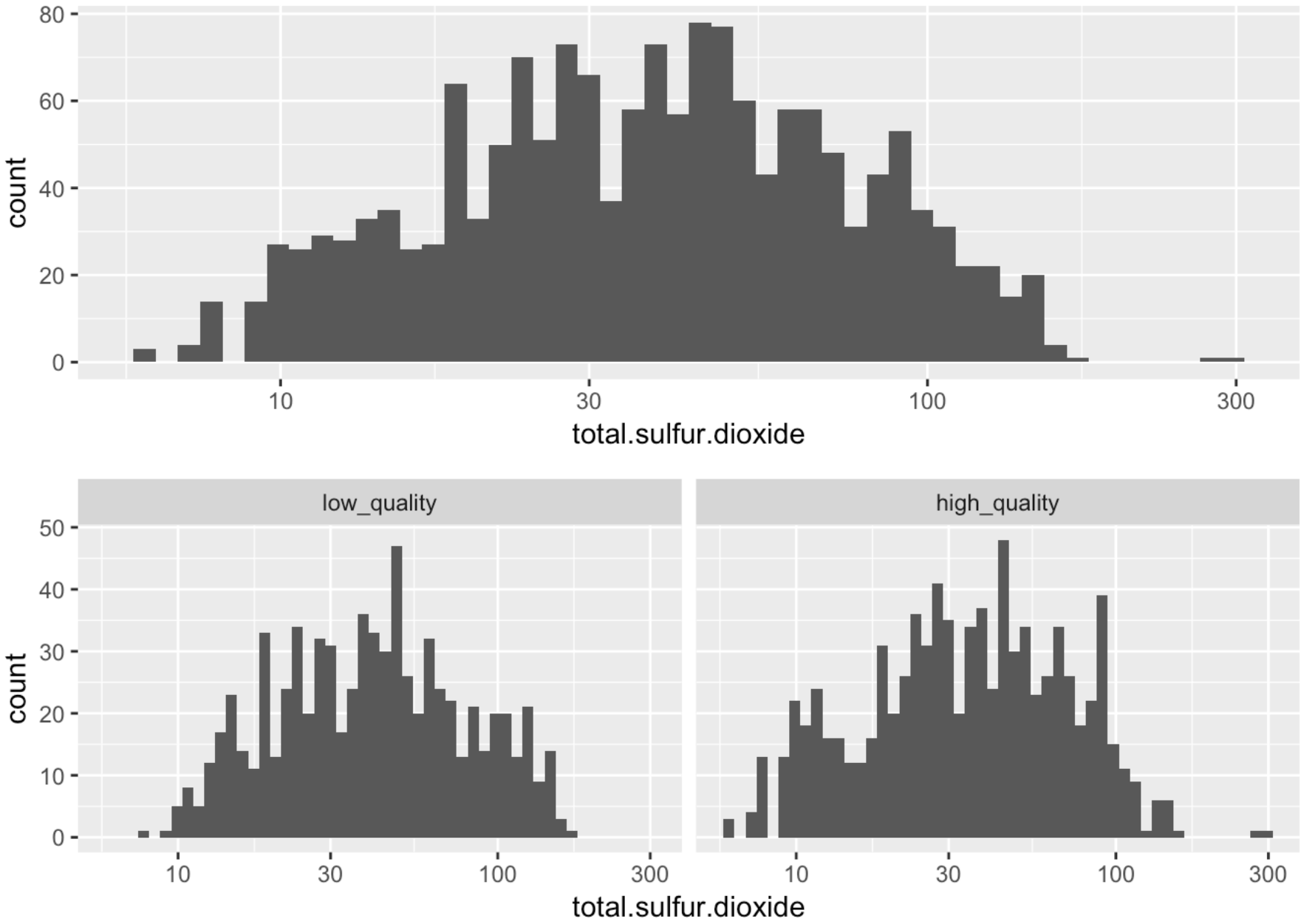


Apart from the fact that the long tail outliers on the right of the distribution belong to the high quality dataset, this is a again pretty normal.

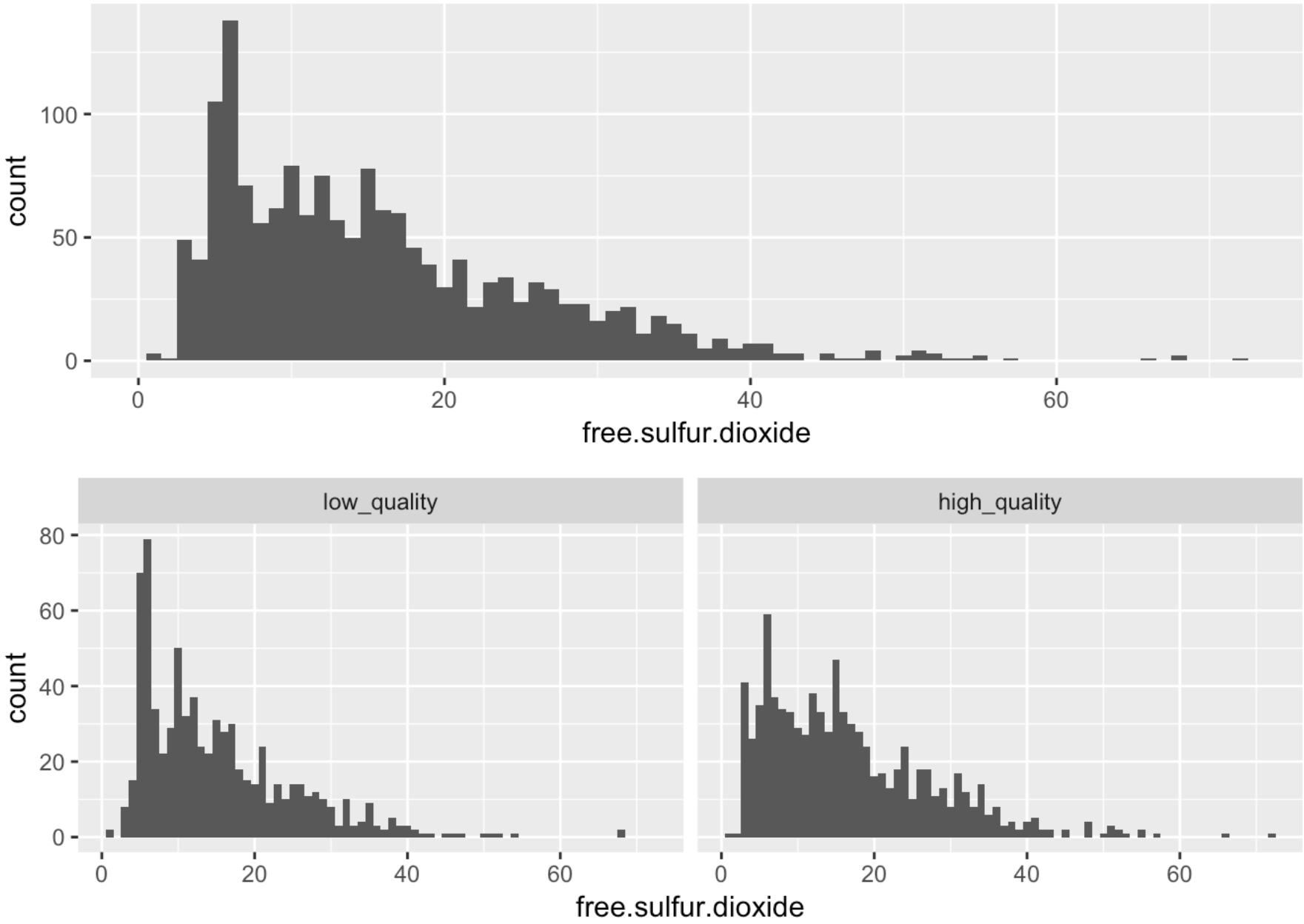
Total Sulfur Dioxide

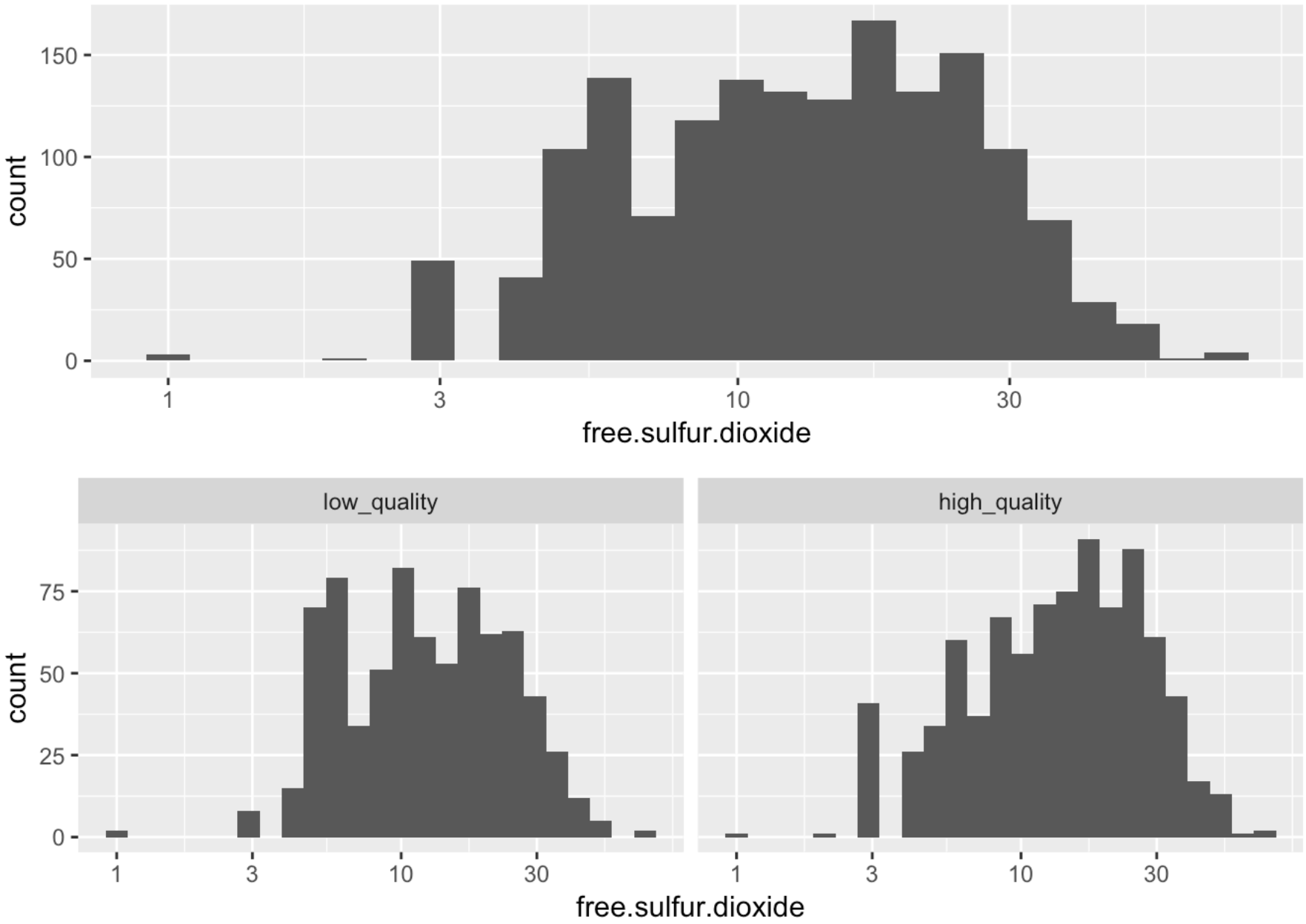
We see from the data description that total sulfur dioxide of greather than 50ppm, becomes evident in taste. We need to pay attention to see if this has an effect on perceived quality.


count
40
30
20
10
0
0
50
100
150
200
250
300
total.sulfur.dioxide
low_quality
high_quality
count
20
10
0
0
50
100
150
200
250
300
0
50
100
150
200
250
300
total.sulfur.dioxide



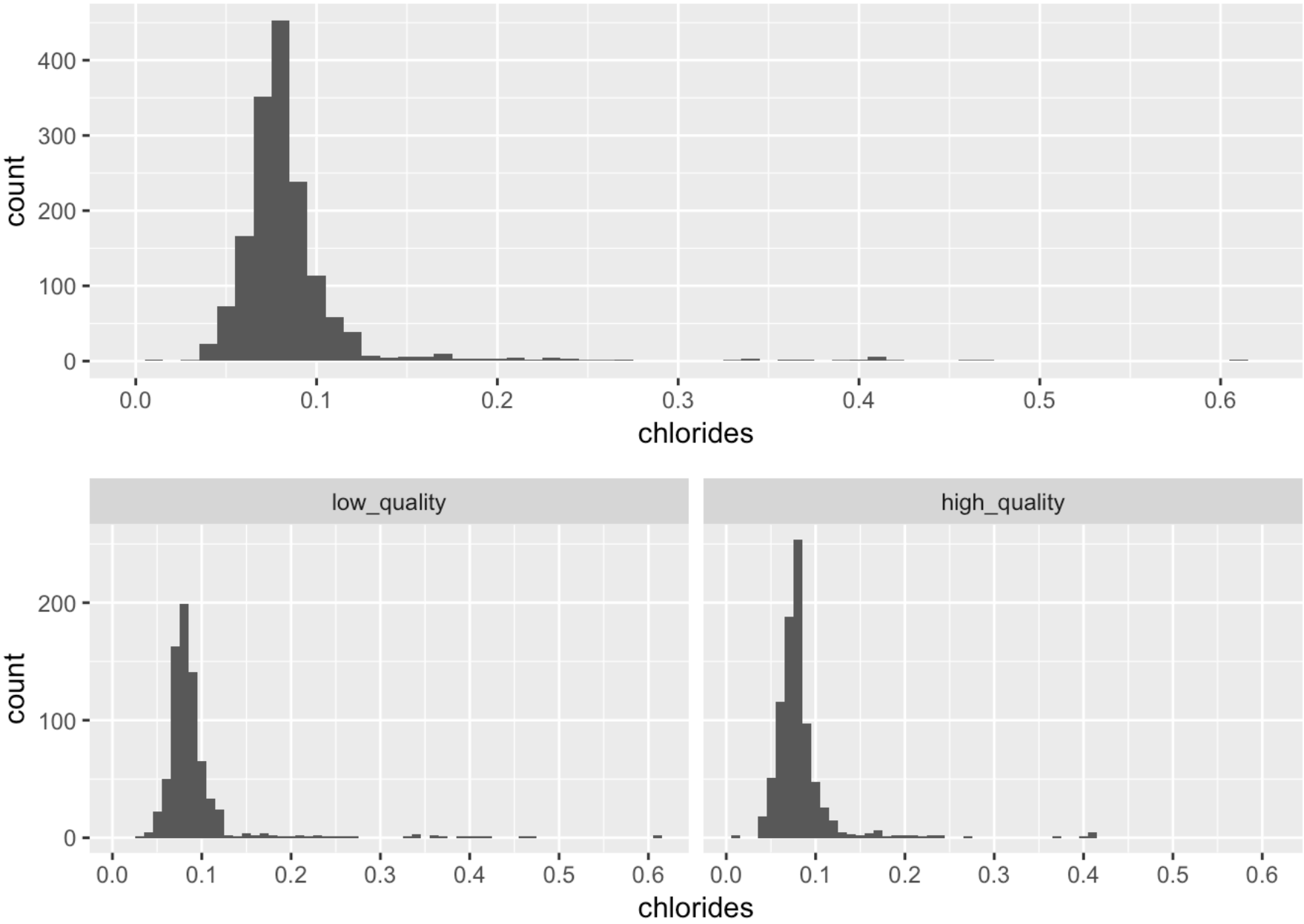
Free Sulfur Dioxide

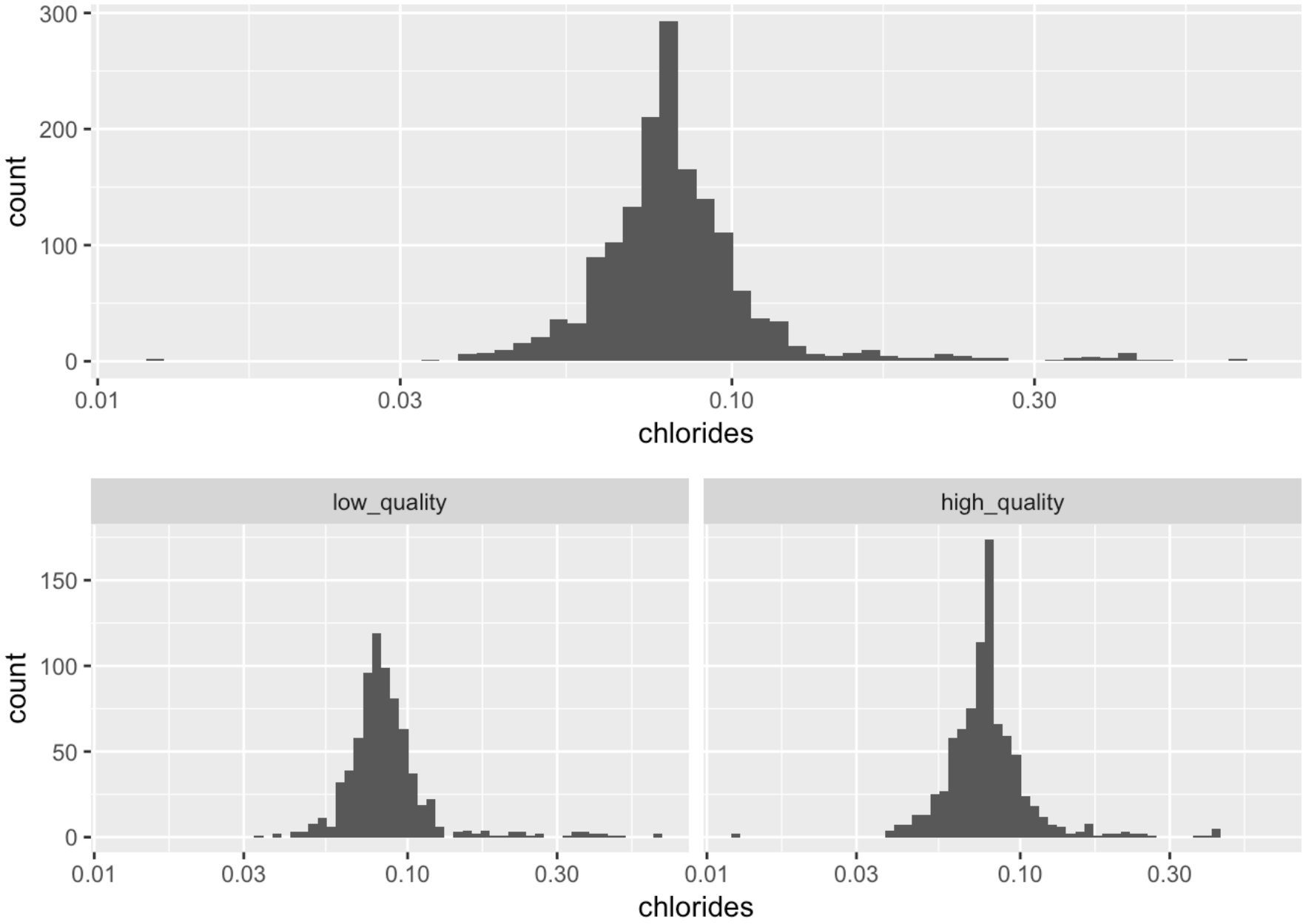

count
100
50
0
0
20
40
60
free.sulfur.dioxide
low_quality
high_quality
count
80
60
40
20
0
0
20
40
60
0
20
40
60
free.sulfur.dioxide



We would expect free sulfur dioxide to be correlated with total sulfur dioxide. We will examine this in the bi-variate section.

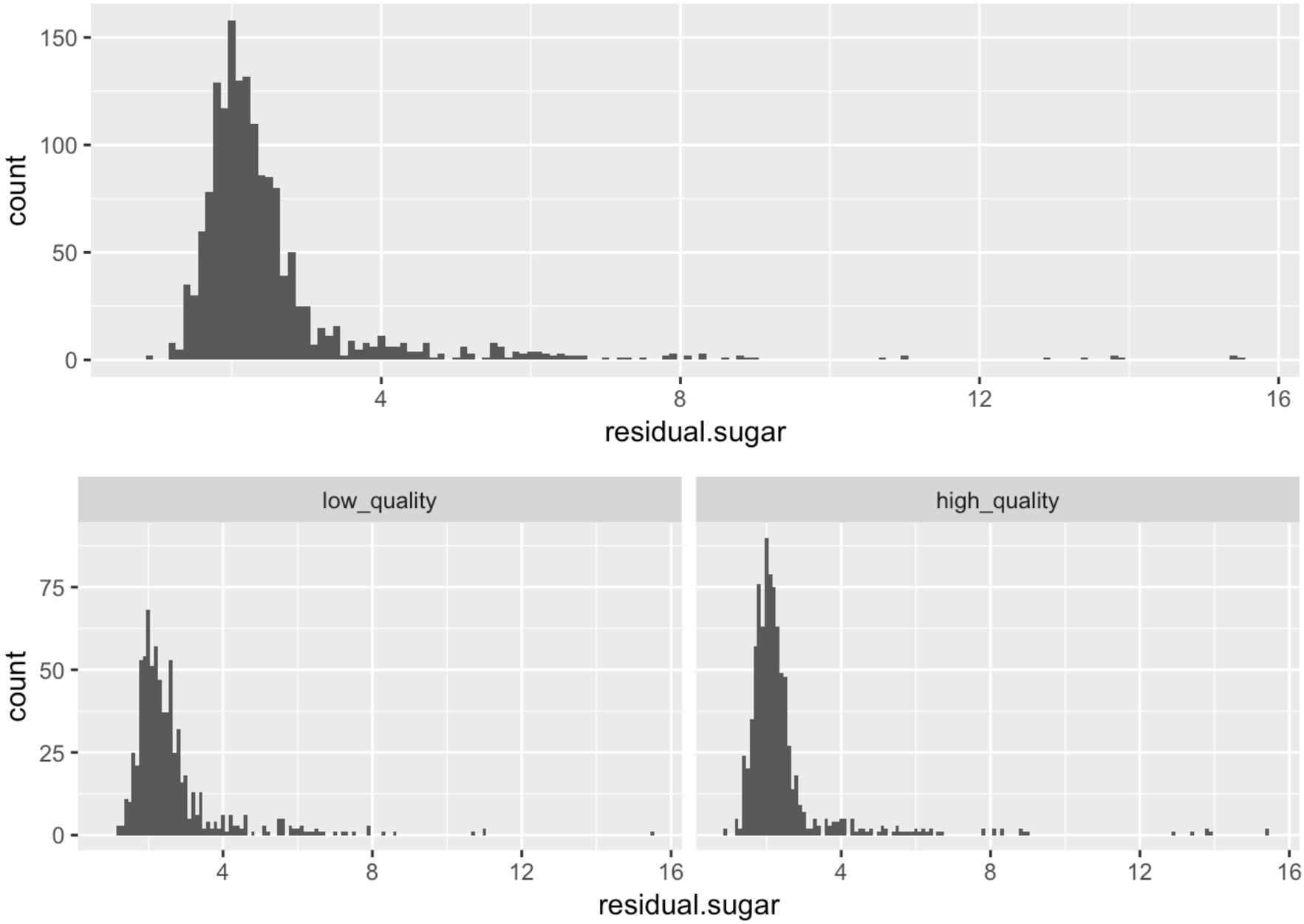
Chlorides

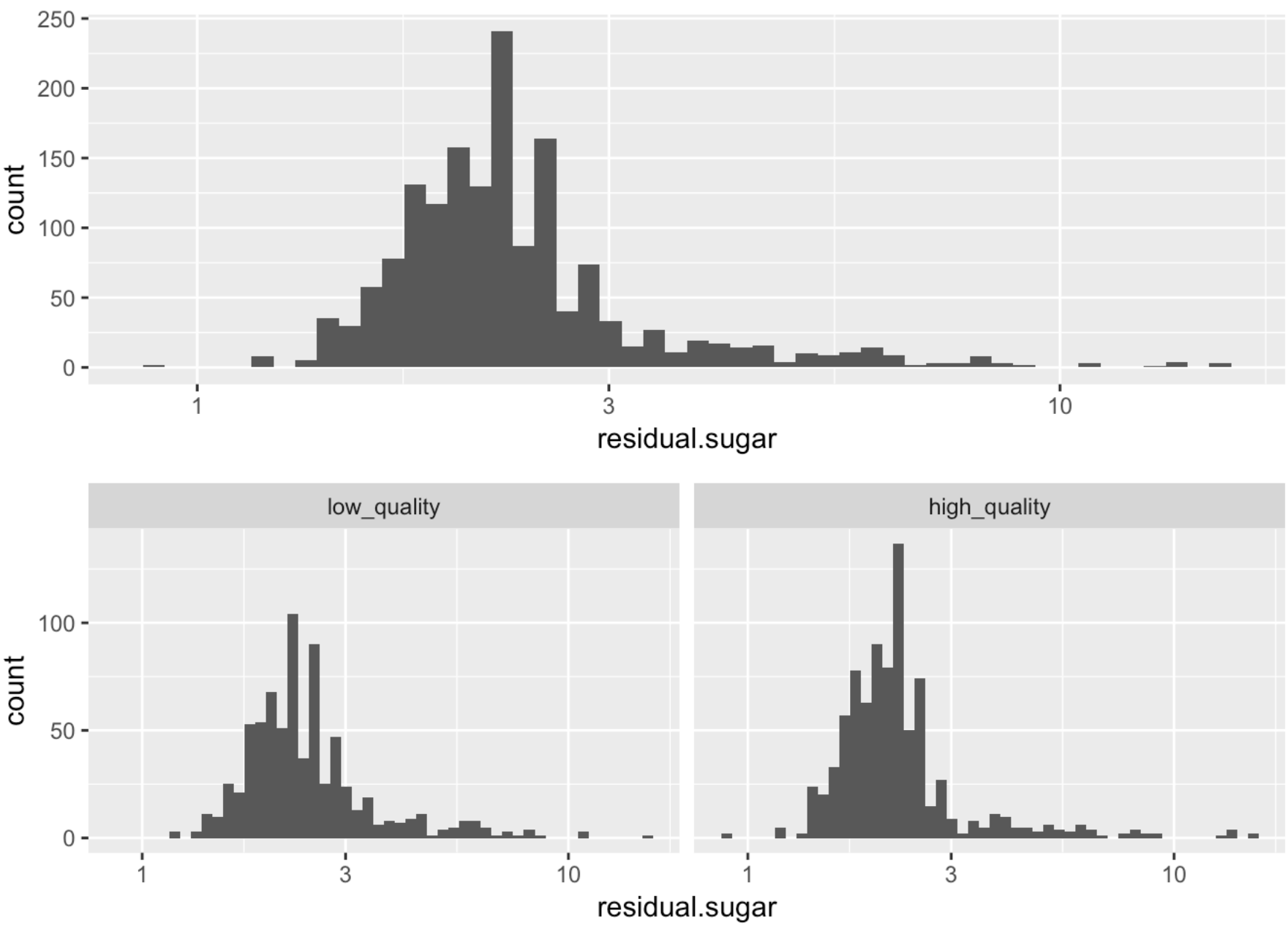

count
400
300
200
100
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
chlorides
low_quality
high_quality
count
200
100
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
chlorides



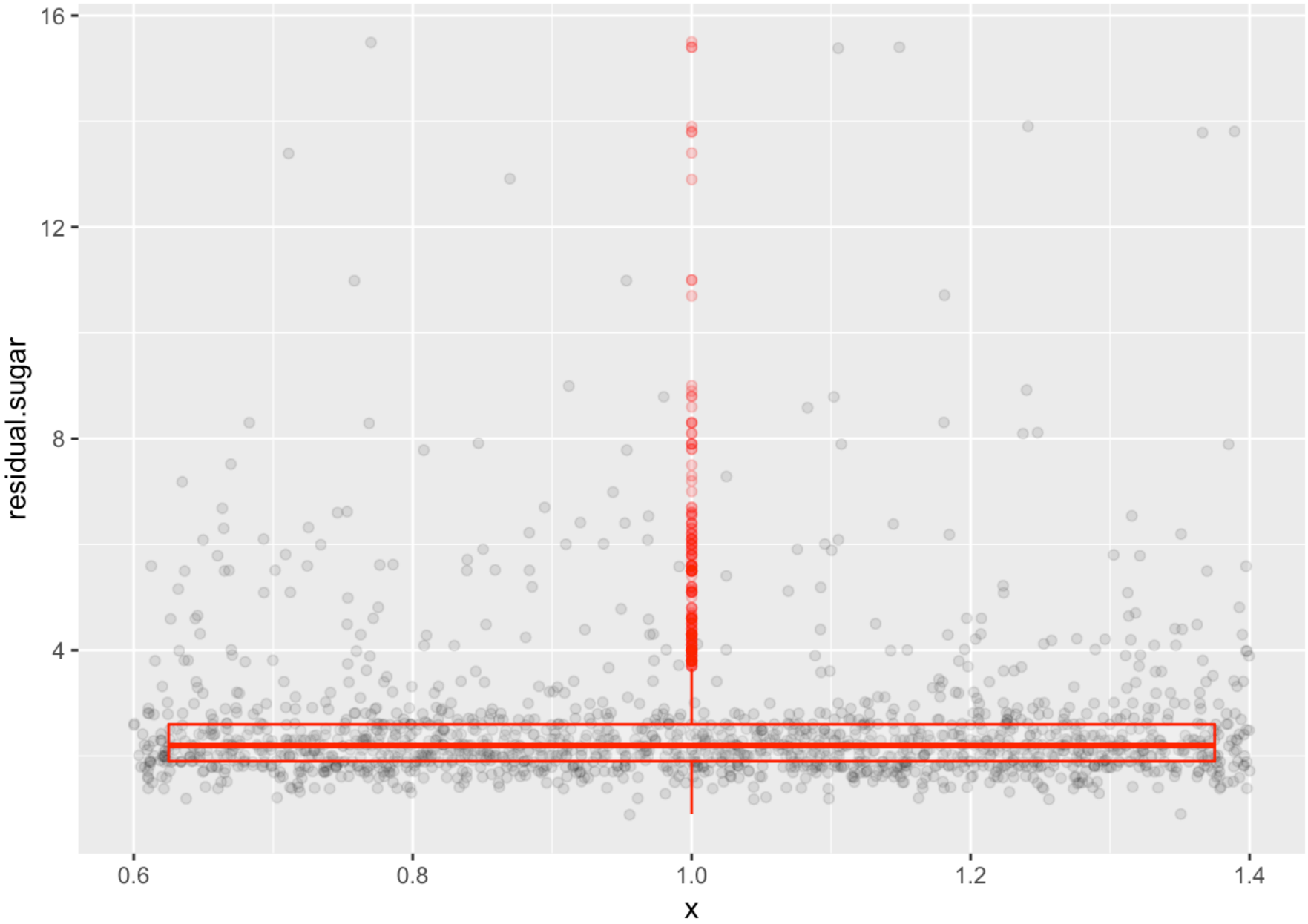
The distributions for cholrides are identical in both quality classes except for the fact that the peak around mean is higher for high quality subset.

Residual Sugar

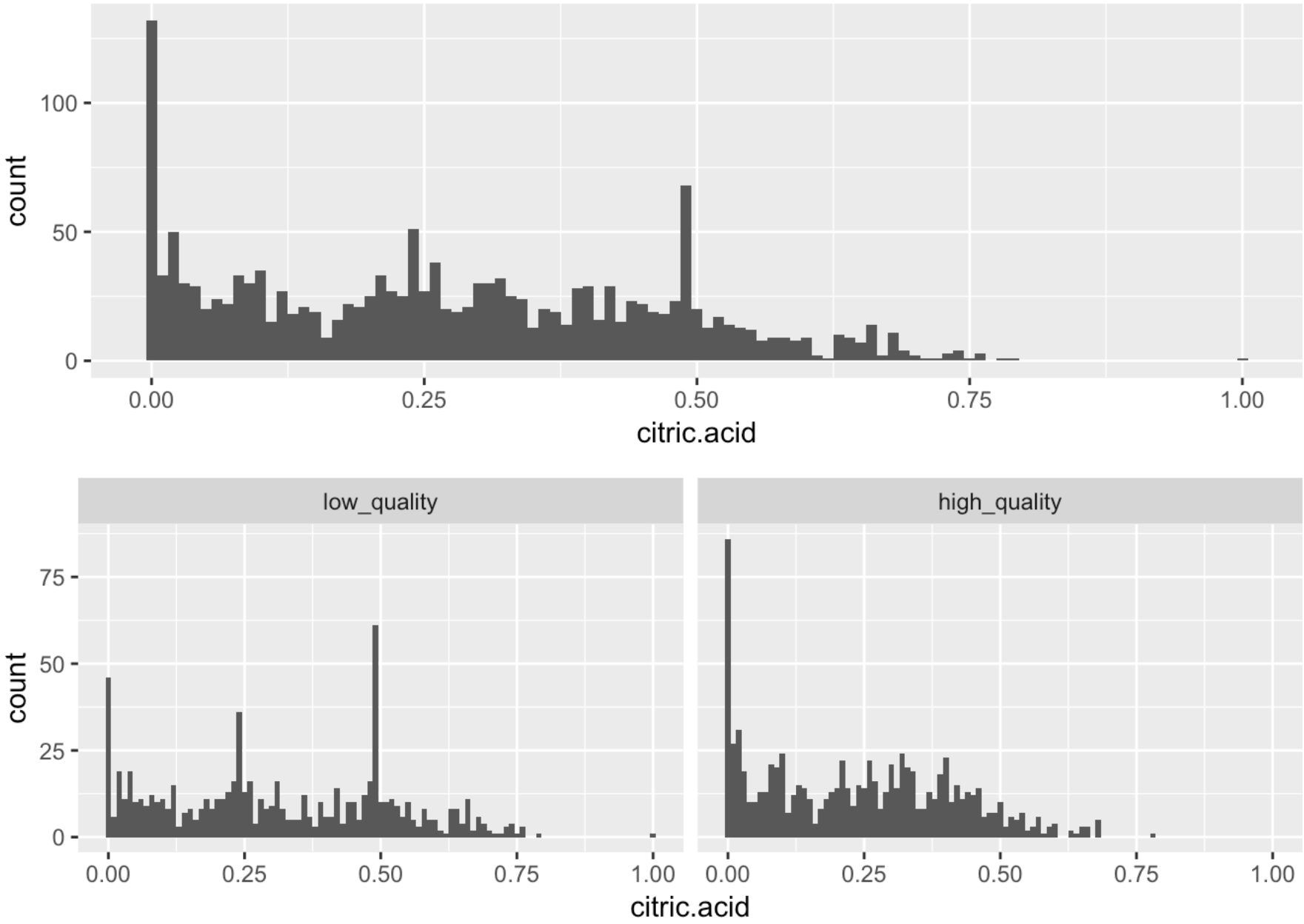

150
100
50
0
count
4
8
12
16
residual.sugar
low_quality
high_quality
75
50
25
0
count
4
8
12
16
4
8
12
16
residual.sugar



We notice that the long tail on the right of high quality subset, and would like to investigate if residual sugar has an effect on perceived quality. Also, in general, the distribution of residual sugar has considerably far-off outliers as evidenced in the box plot below.

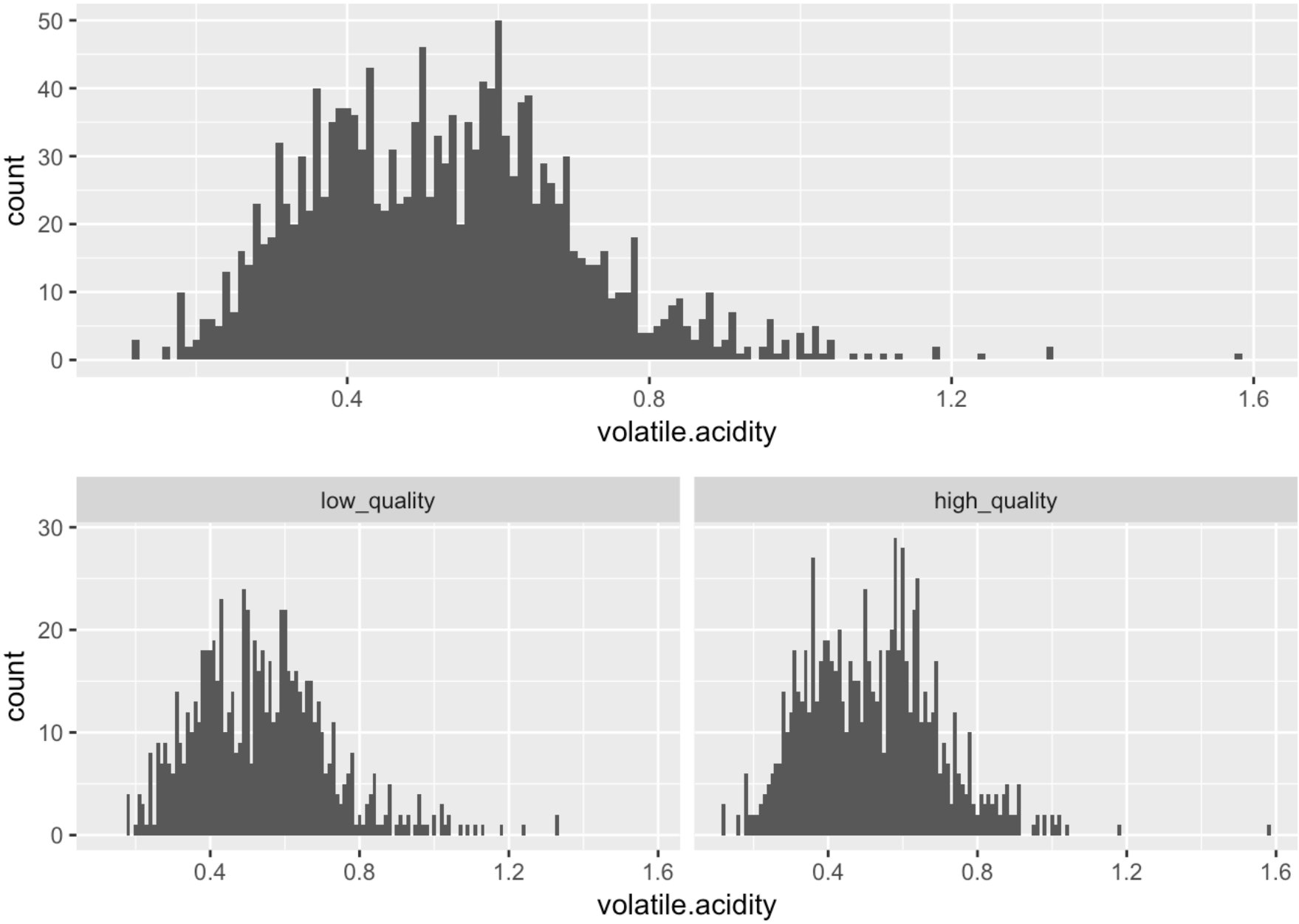


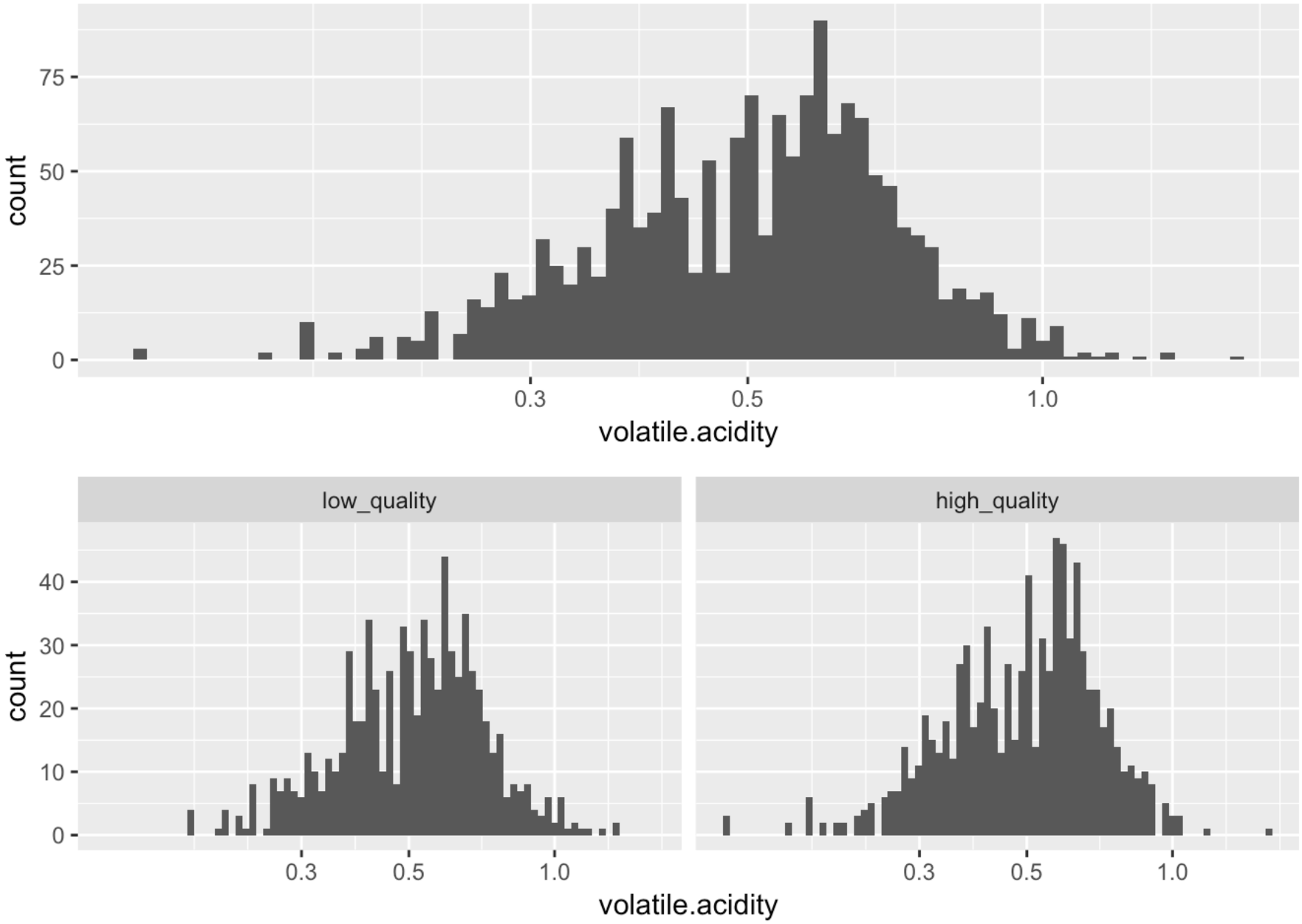
Citric Acid



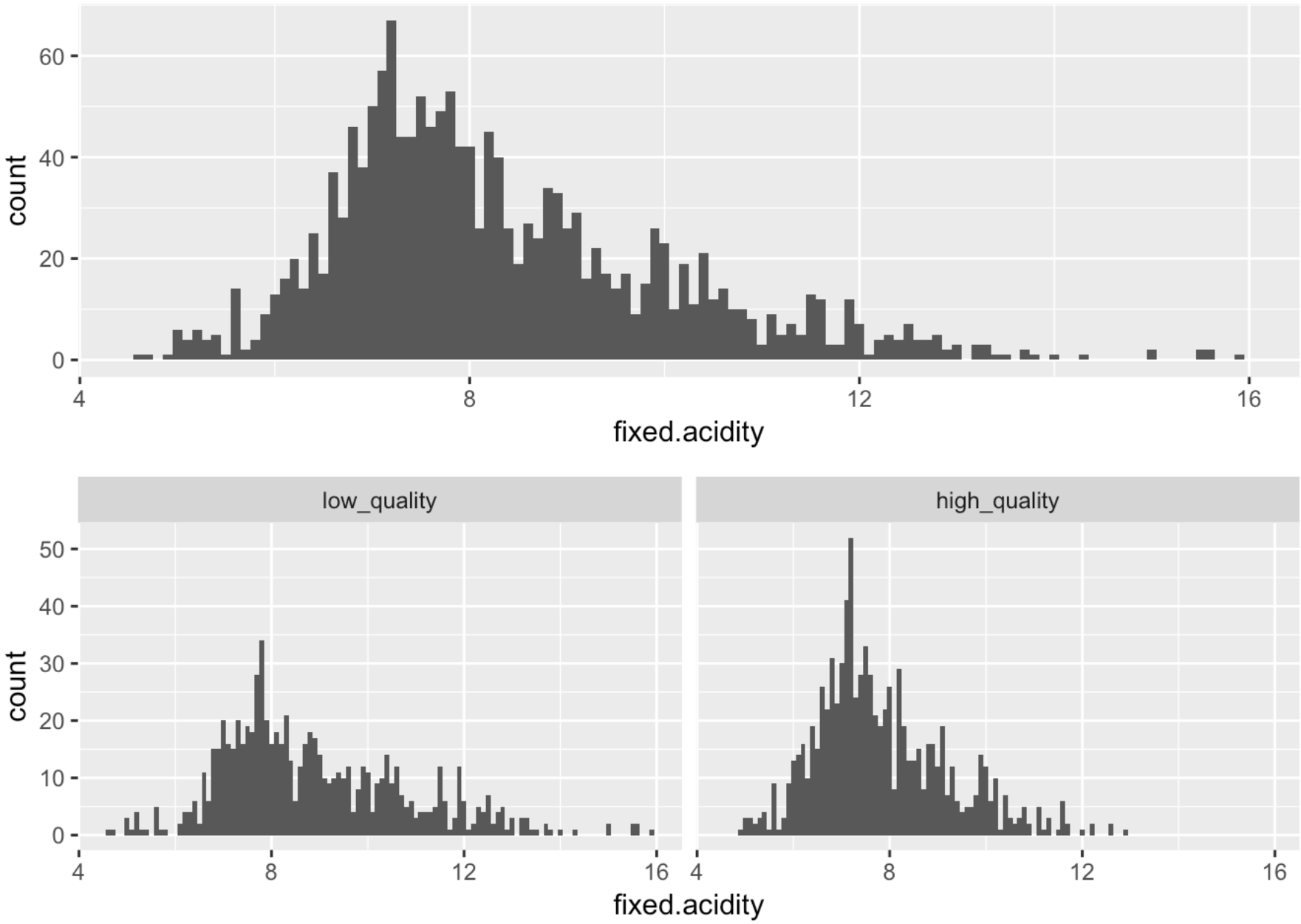
We notice an almost trimodal distribution for citric acid in the low quality case, and they occur around means of 0.0, 0.25, 0.5 etc.

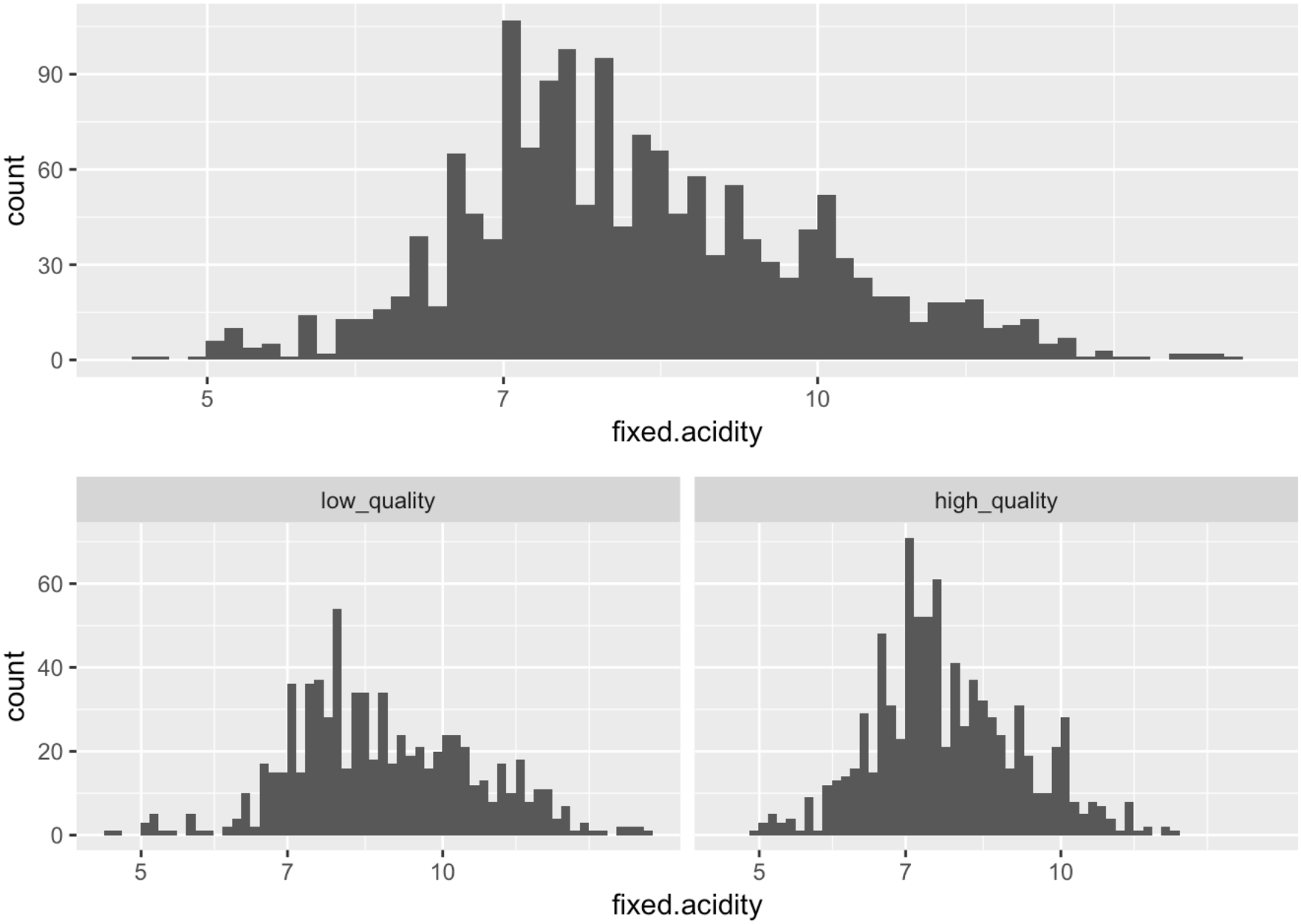
Volatile acidity


count
50
40
30
20
10
0
0.4
0.8
1.2
1.6
volatile.acidity
low_quality
high_quality
count
30
20
10
0
0.4
0.8
1.2
1.6
0.4
0.8
1.2
1.6
volatile.acidity



Fixed Acidity


count
60
40
20
0
4
8
12
16
fixed.acidity
low_quality
high_quality
count
50
40
30
20
10
0
4
8
12
16
4
8
12
16
fixed.acidity



Univariate Analysis

What is the structure of your dataset?

There are 1599 rows of wine quality data. Each row has 12 columns of numerical attributes describing aspects of the red wine and a column of integers that corresponds to a quality score between 0 and 10.

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

Upon looking at Univariate plots, the following column appear to be interesting in understanding the main feature of interest which is the quality of wine. - Higher alcohol seems to be associated with high quality wine subset. - Suphur dioxide dependence and correlation between total and free levels needs to be looked at. - Acidity related columns such as pH, Citric Acid, fixed and volatile acidity need to be looked at overall acidic wines with lower pH seem to be higher quality .

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

N/A

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

We generated a quality class based on whether the quality score is greater than 5 or not. We may have to create additional factor variables from numericals for trivariate analysis.

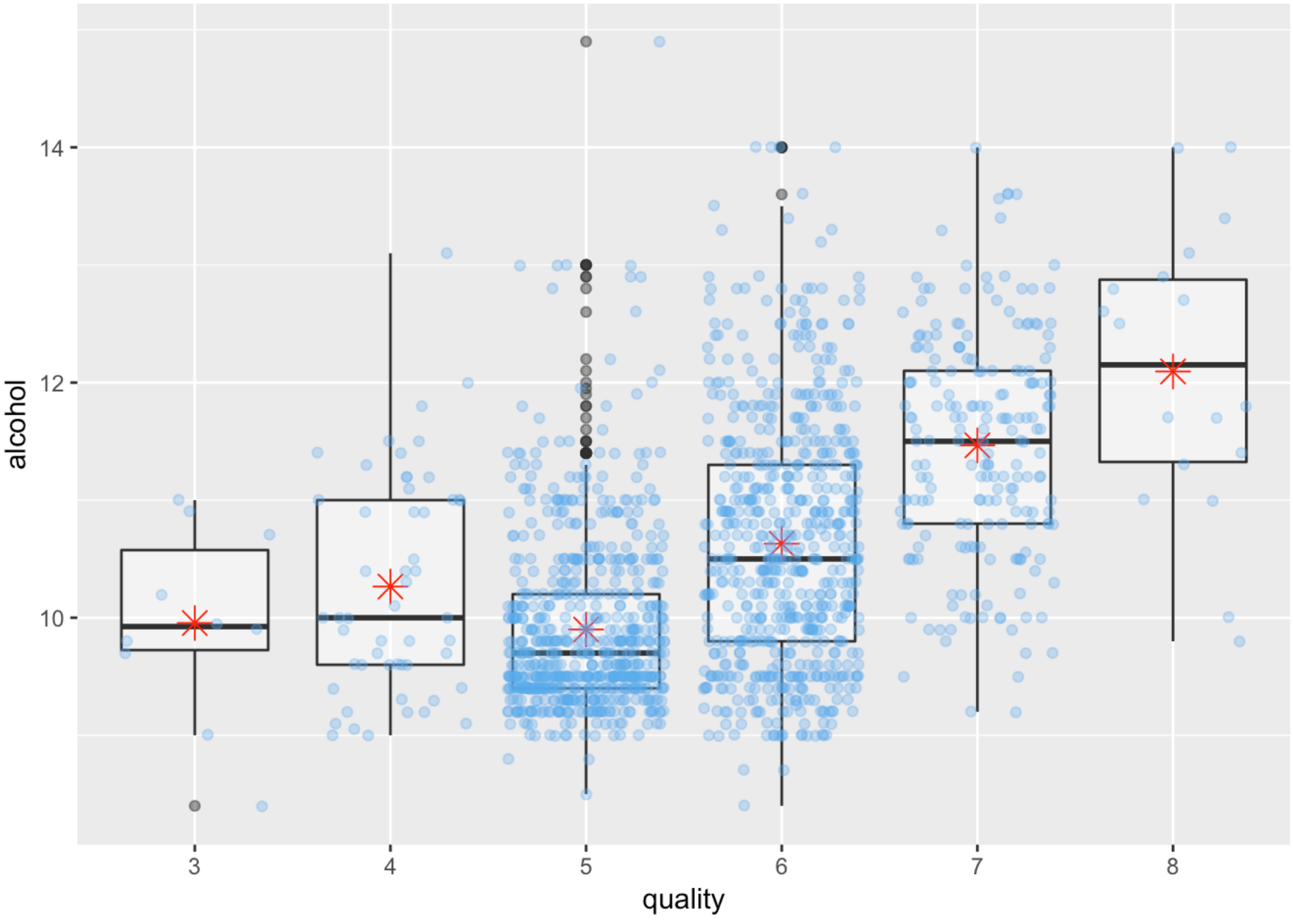
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

The distributions that were right skewed were transformed using log10 scaling.

Bivariate Plots Section

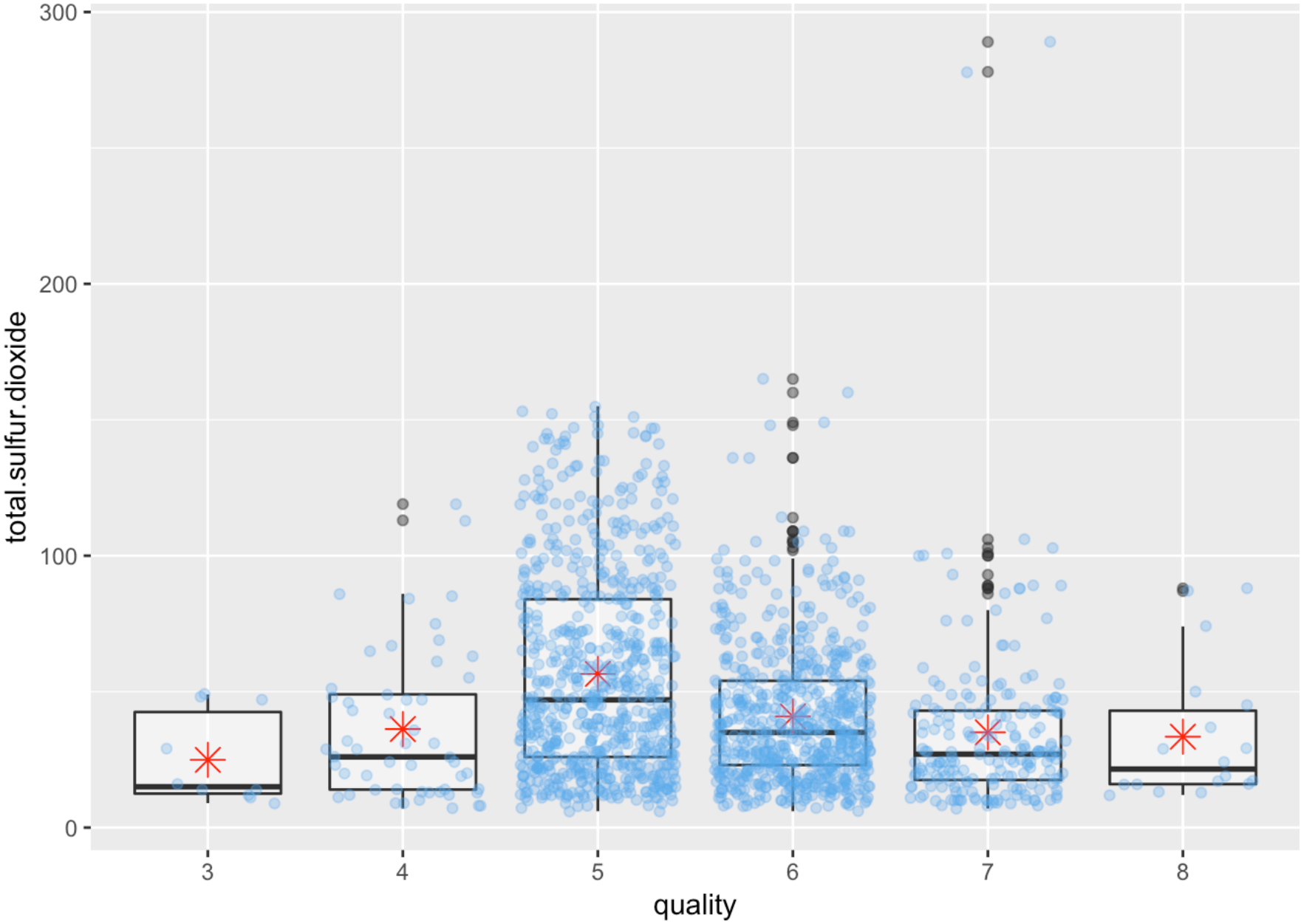
Let's start our bivariate analysis by analyzing key columns identified earlier for their effects on quality.

Quality vs Alcohol



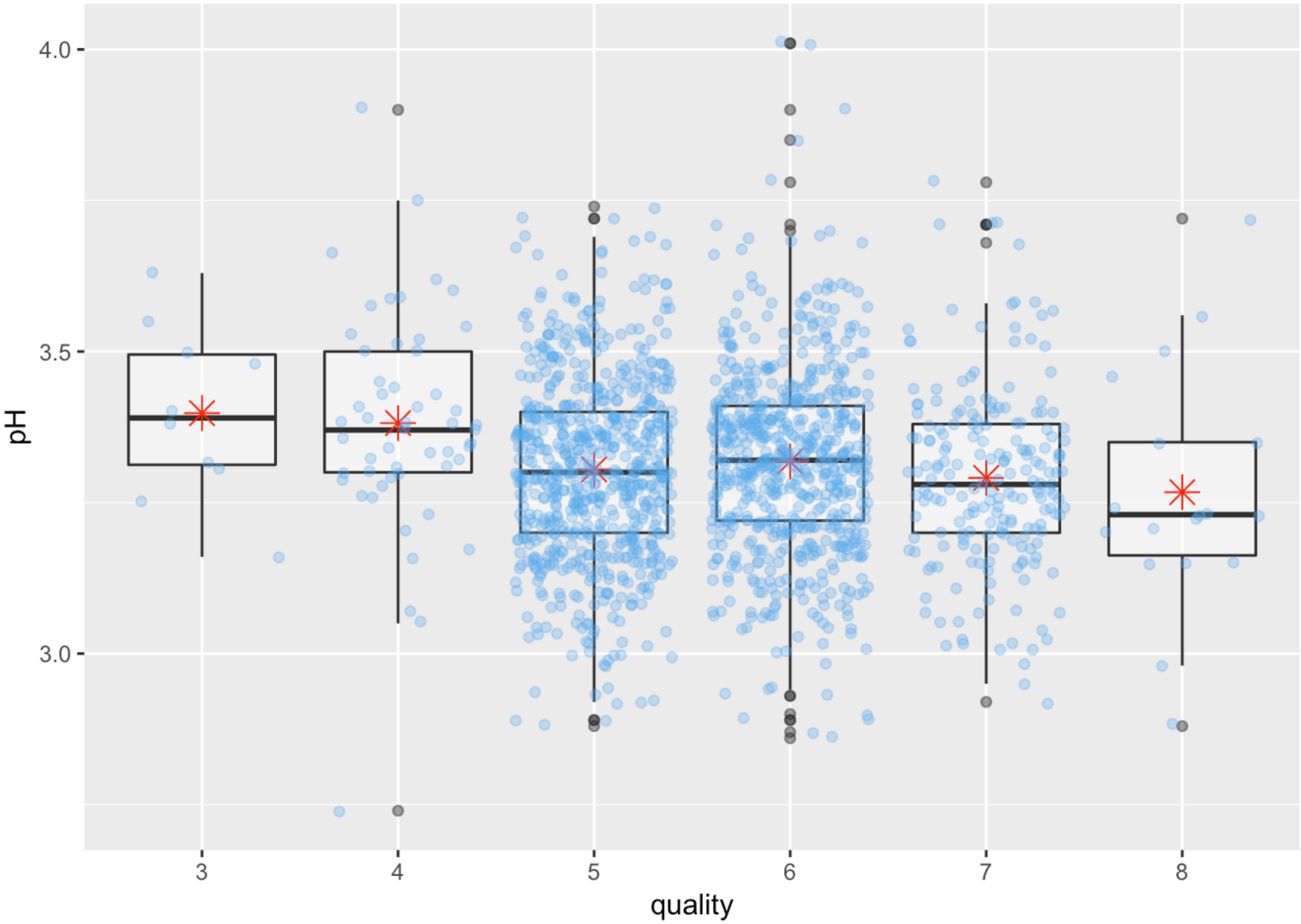
The mean alcohol level at each quality level is denoted by 'x'. We see that overall higher alcohol levels are asssociated with higher quality, and at the highest quality all the datapoints are above 7 % alcohol.

Quality vs Total Sulfur Dioxide



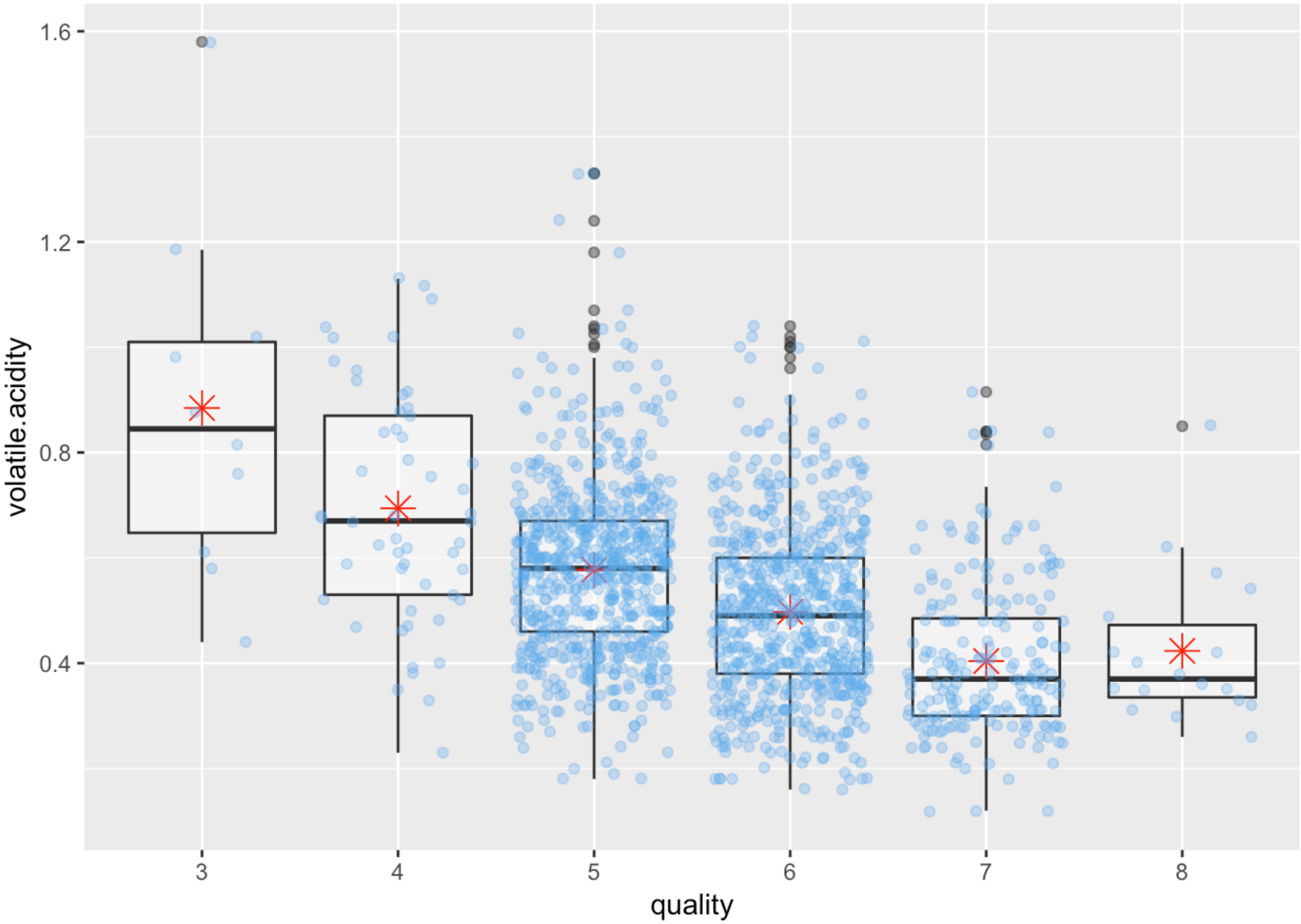
The total sulfur dioxide relation is more complicated. At low levels of less than 50ppm total sulfur dioxide, the datapoints are distributed across all quality levels whereas the datapoints with greater than 120ppm or so, are invariably in the 5-6 range always, indicating that at higher levels of sulfur dioxide, its negative impact on perceived quality is more pronounced.

Quality vs pH



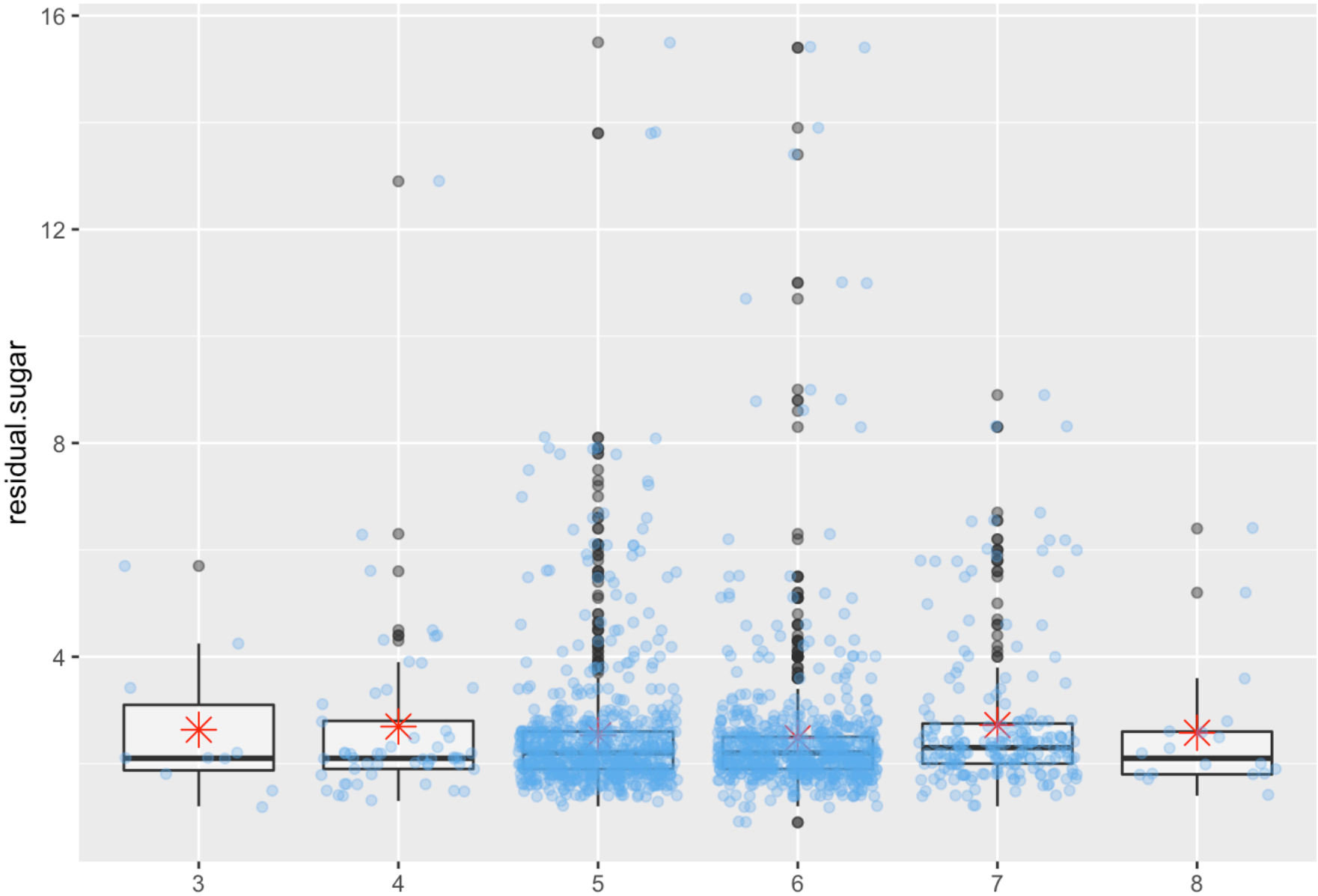
Lower pH seems to be associated with higher quality indicating acidic wines are better tasting.

Quality vs Volatile Acidity



However, volatile acidity is not favored since higher volatile acidity leads to lower wine quality. Beyond volatile acidity value of 1, there are no more high quality wine datapoints.

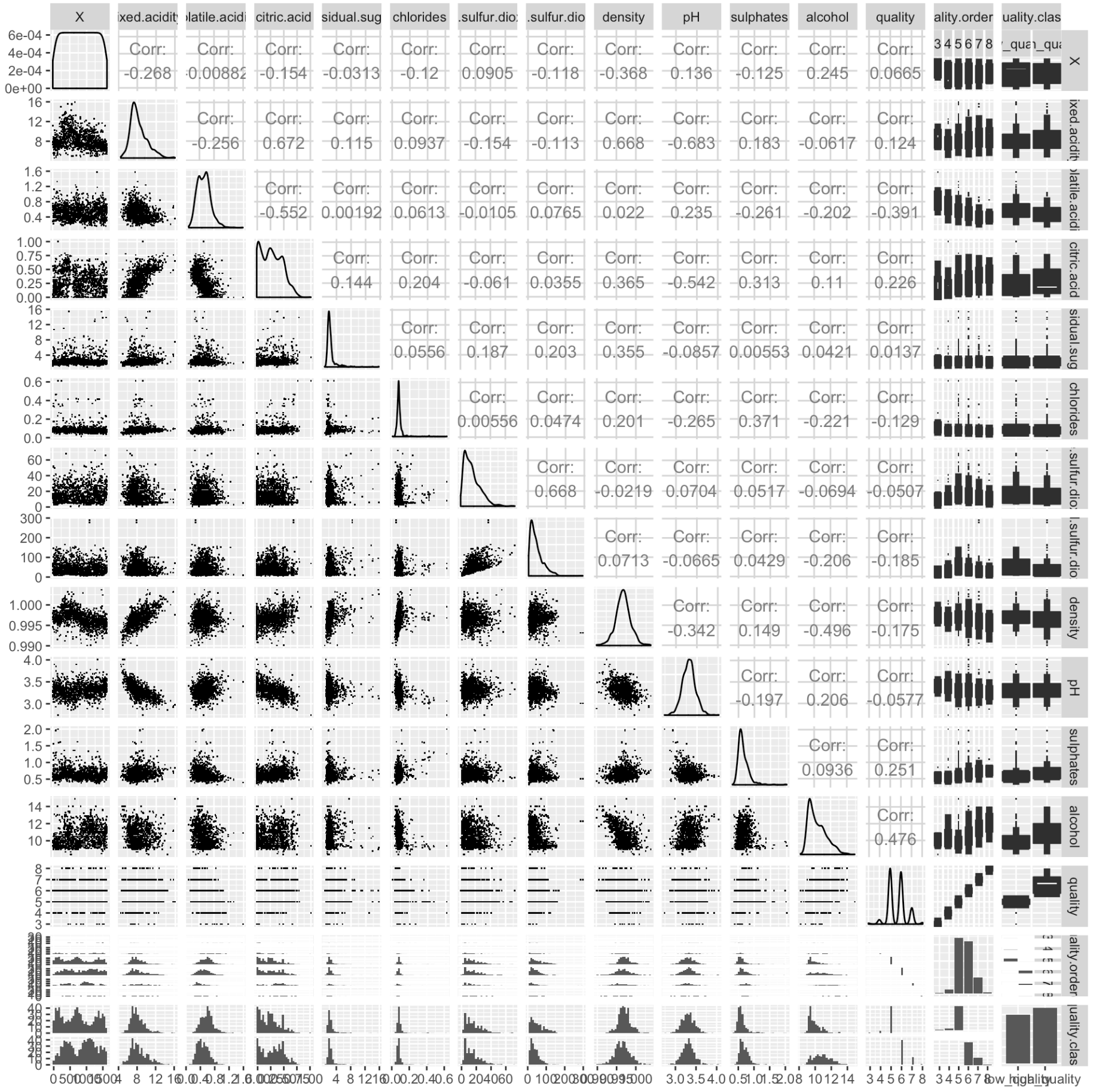
Quality vs Residual Sugar



Residual sugar doesn’t seem to be indicative of quality, which is contrary to intuition. Also, the IQR of residual sugar in the dataset is small, with some far off outliers which seem associated with low and mid-quality wines mostly.

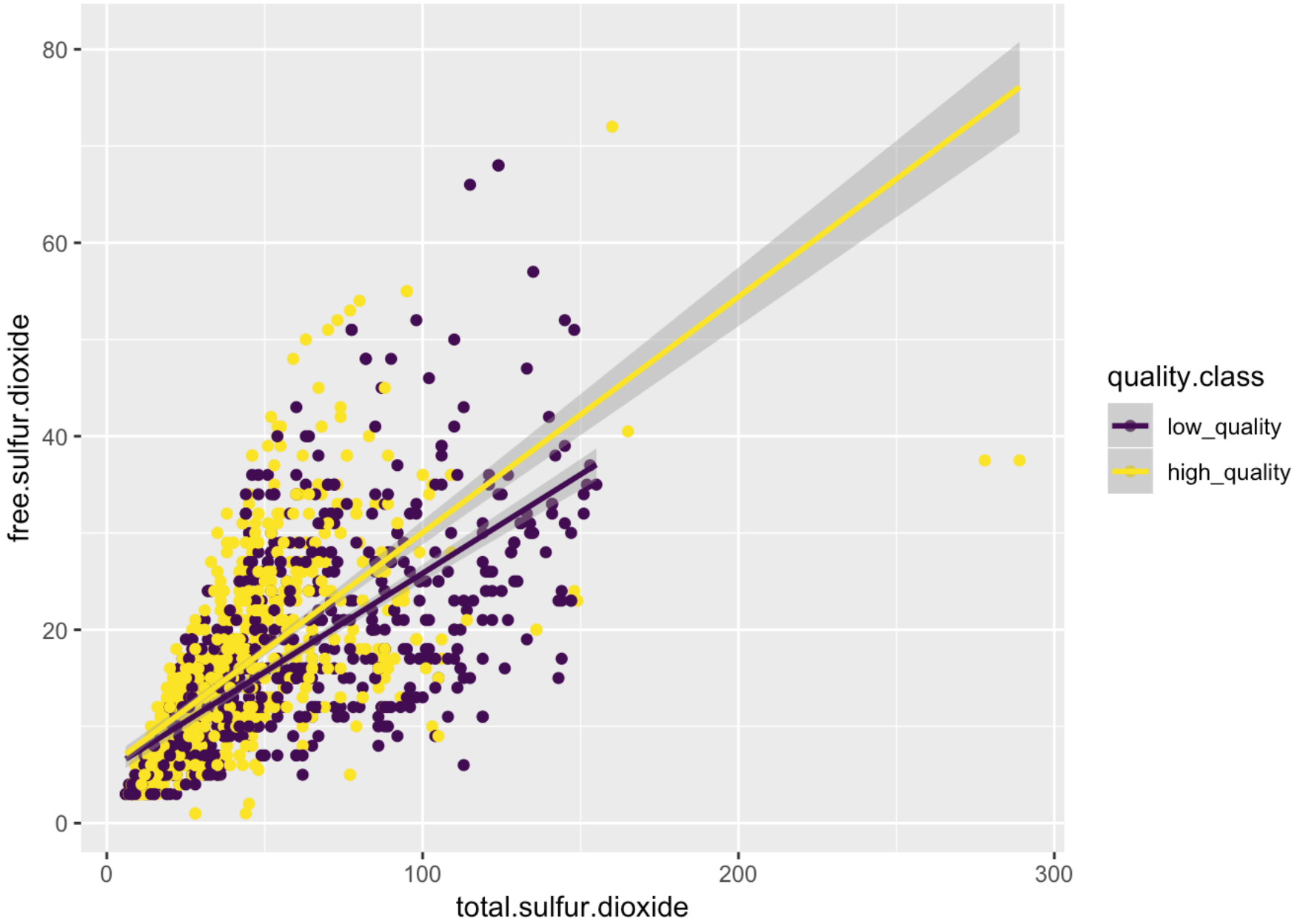
Scatter matrix

Now, let’s look at the scatter matrix to identify correlation among attributes.



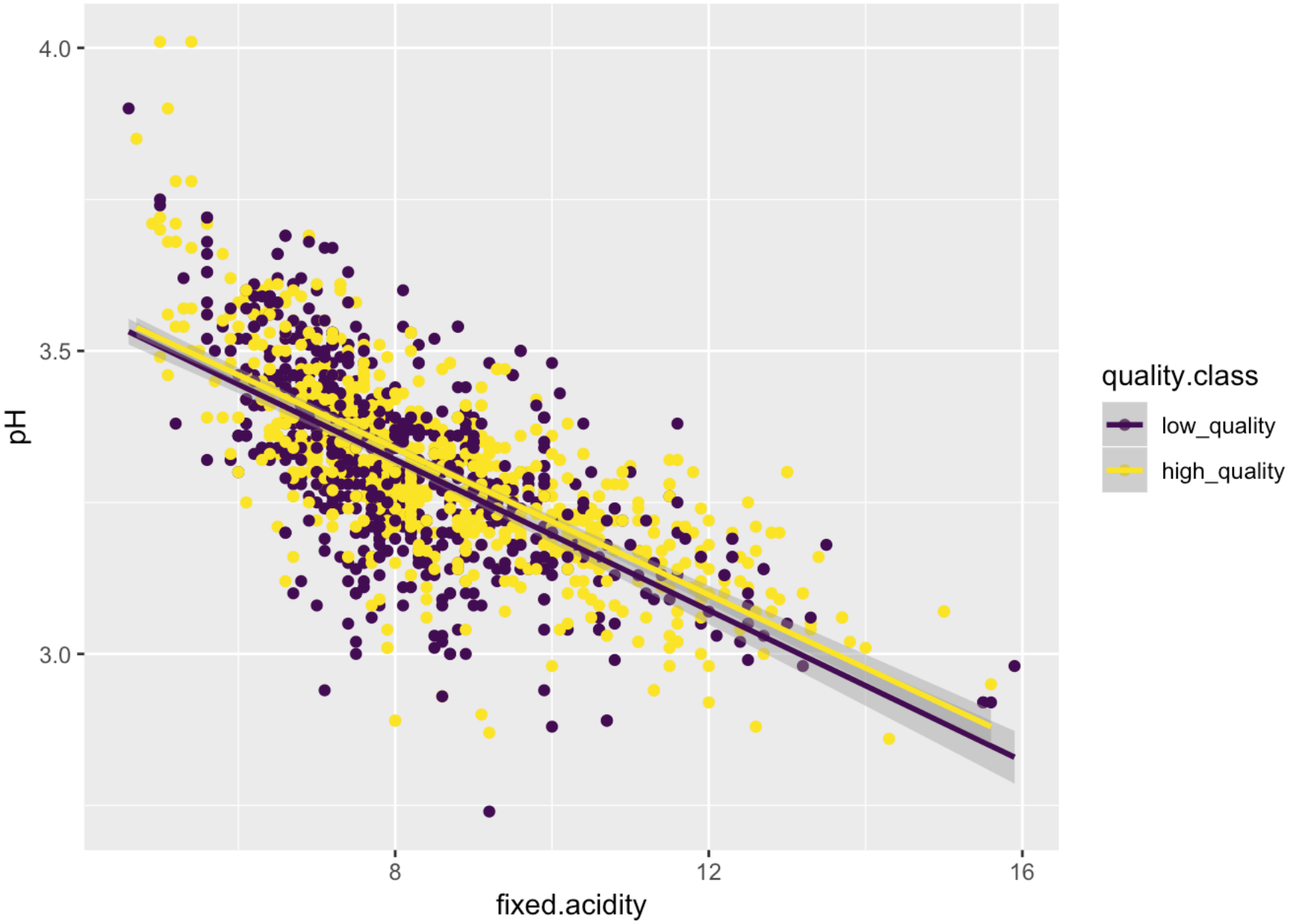
As expected, the acidity metrics are highly correlated with one another. We also see correlations between acidity metrics and density, and between density and alcohol content. As expected, the sulfur dioxide levels also correlate well for free vs total levels.

Free vs Total Sulfur Dioxide



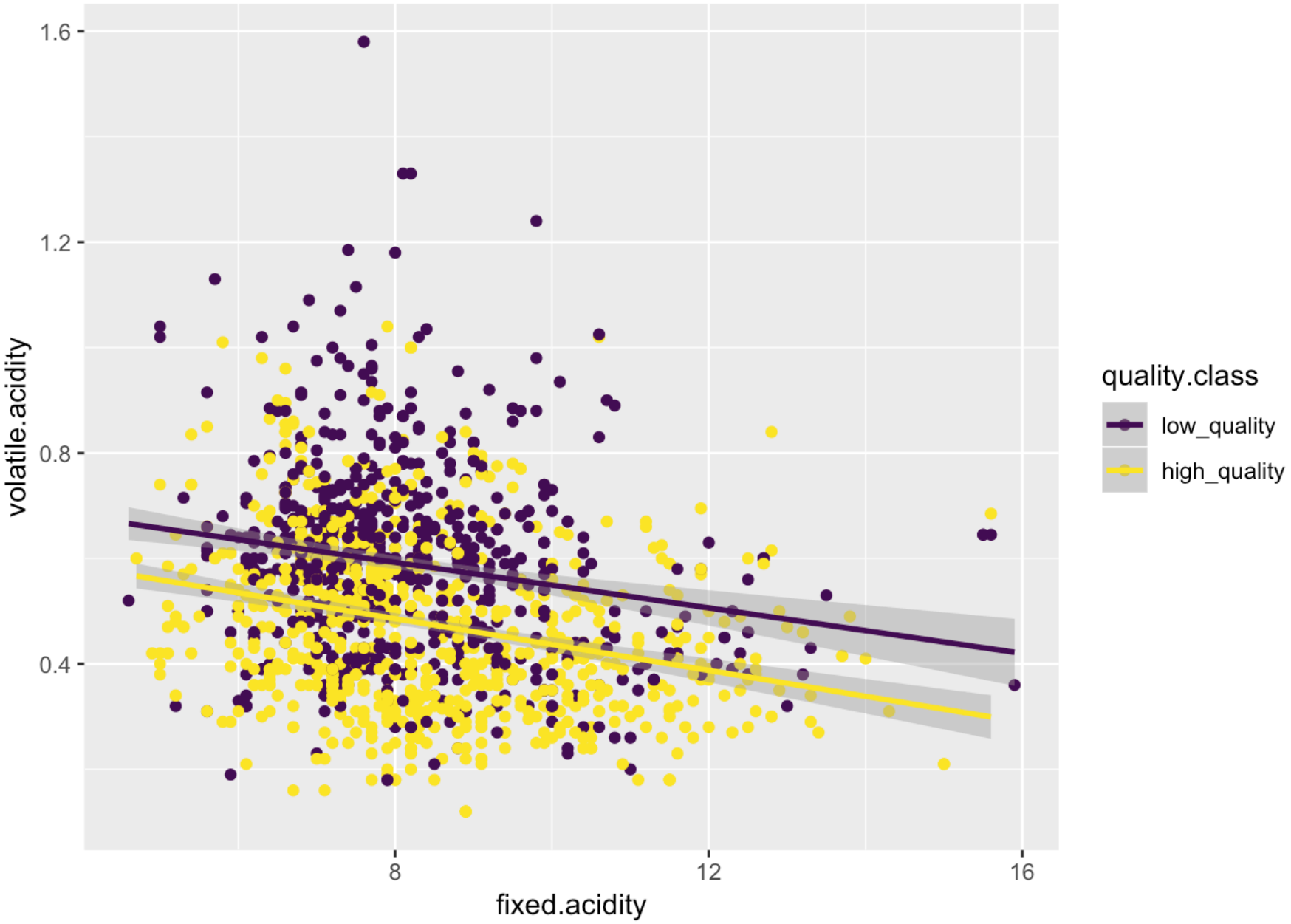
The points spread out more at higher total levels and overall it appears as though more free to total ratio might result in higher quality.

Fixed acidity vs pH



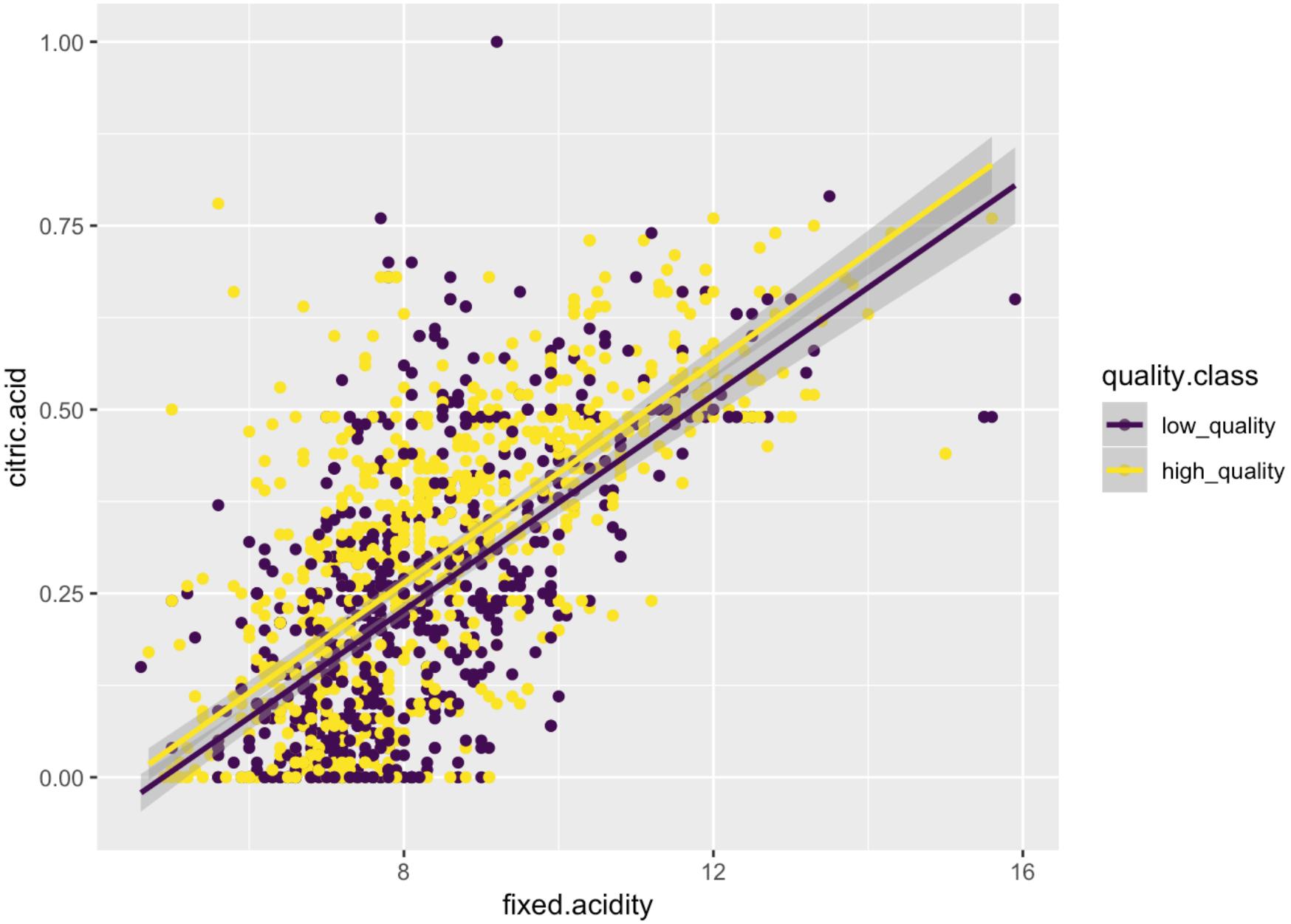
There is a strong negative correlation as lower pH levels mean more acidity.

Fixed vs Volatile Acidity



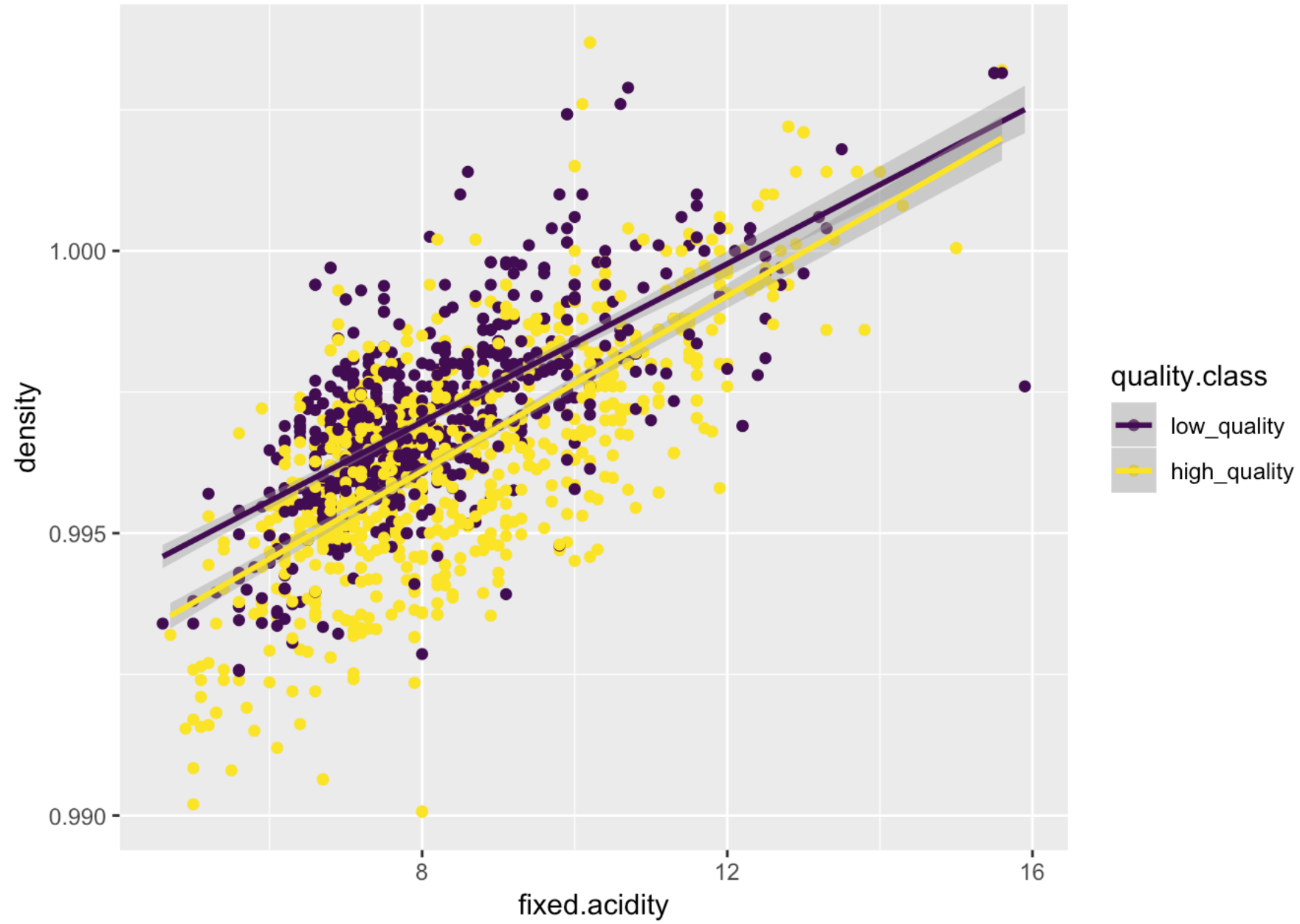
Again, we can see from this plot that high quality wines seem to have lower volatile to fixed acidity ratios.

Fixed acidity vs Citric Acid



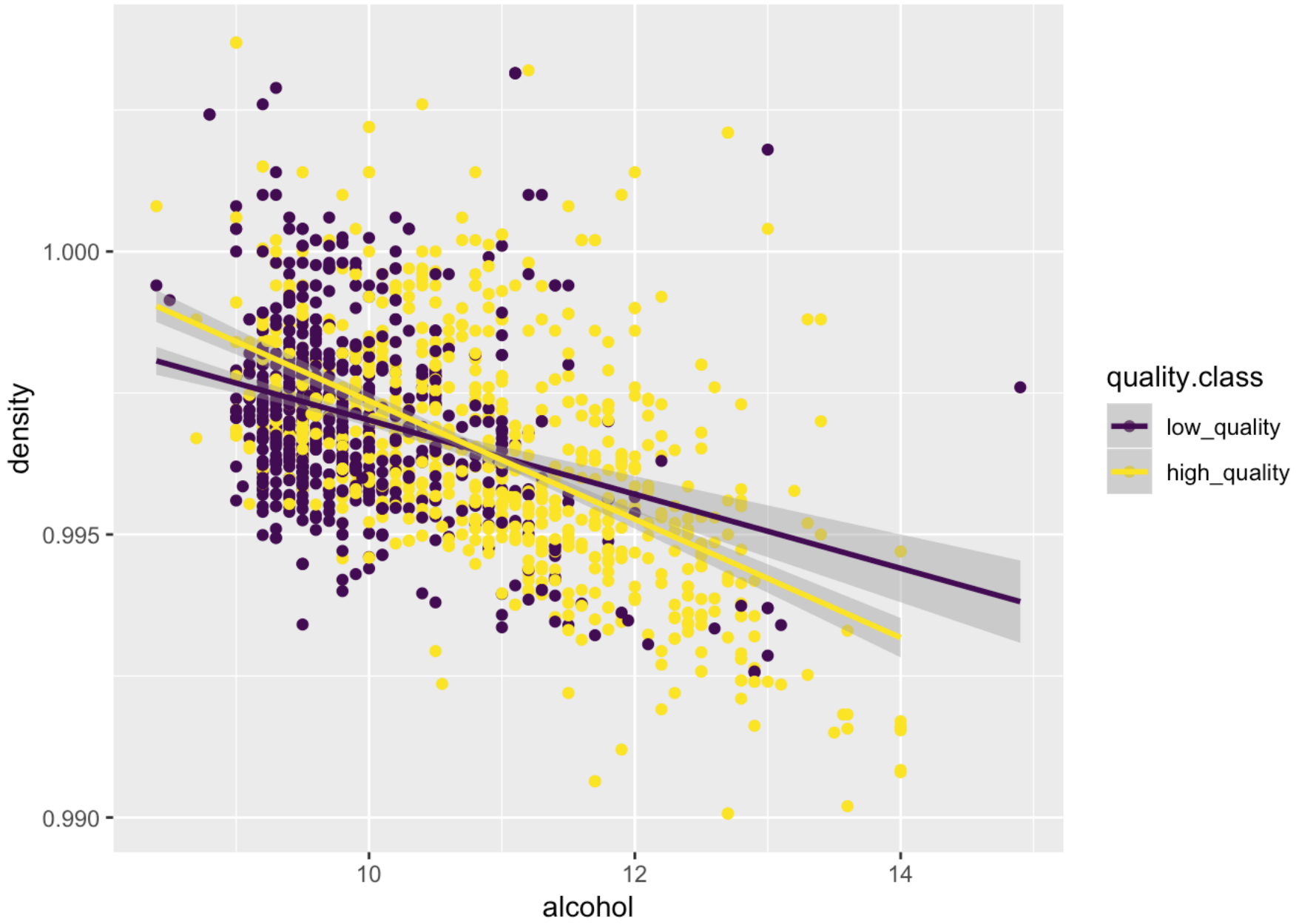
Addition of higher levels of citric acid seem to be associated both with higher quality and also higher fixed acidity.

Density vs Fixed Acidity



High quality wine seems to be associated with lower density for given fixed acidity.

Density vs Alcohol



As expected most of the high quality wines are associated with higher alcohol levels and higher alcohol levels also seem to correlate with lower density.

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?

High quality wine seems to be mosly acidic (lower pH) but with low levels of volatile acidity. Alcohol levels are directly proportional to perceived quality. High quality wines also seem to be less dense.

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

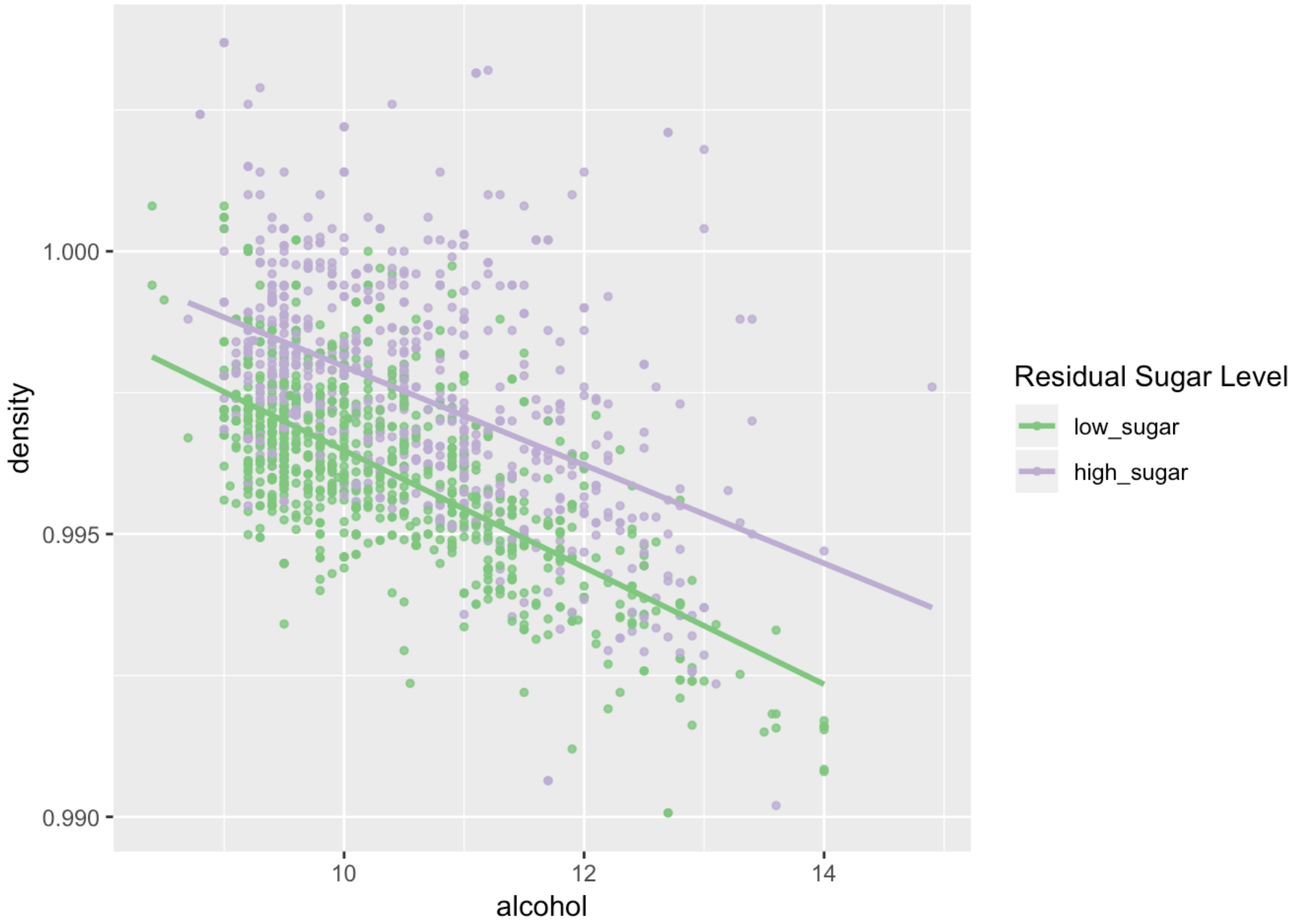
Density and fixed acidity seem to be strongly related in a directly proportional way.Density and alcohol content are inversely proportional.

What was the strongest relationship you found?

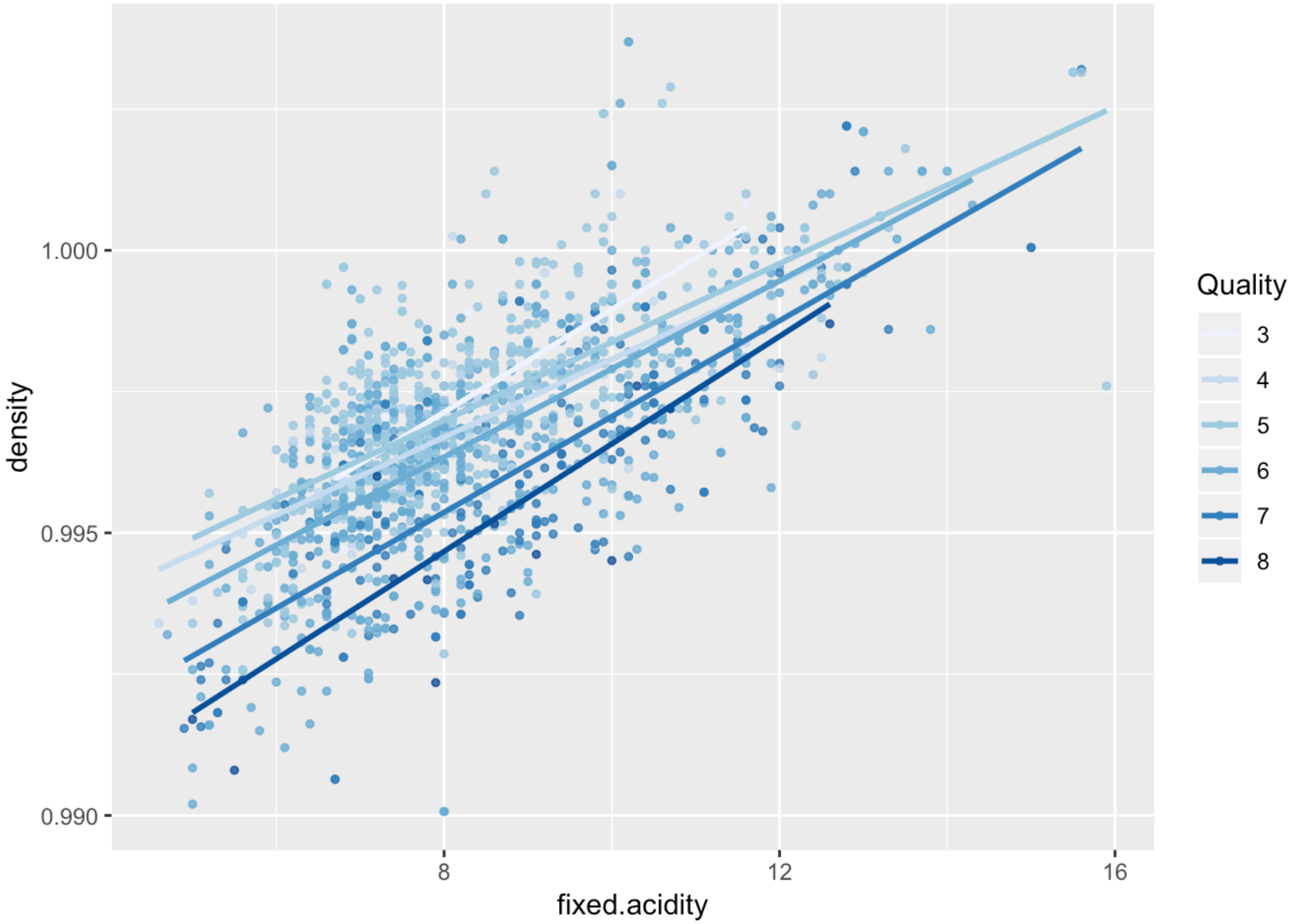
Higher alcohol levels seem to be associated most strongly with higher quality. Similarly, lower pH is also associated with higher quality.

Multivariate Plots Section

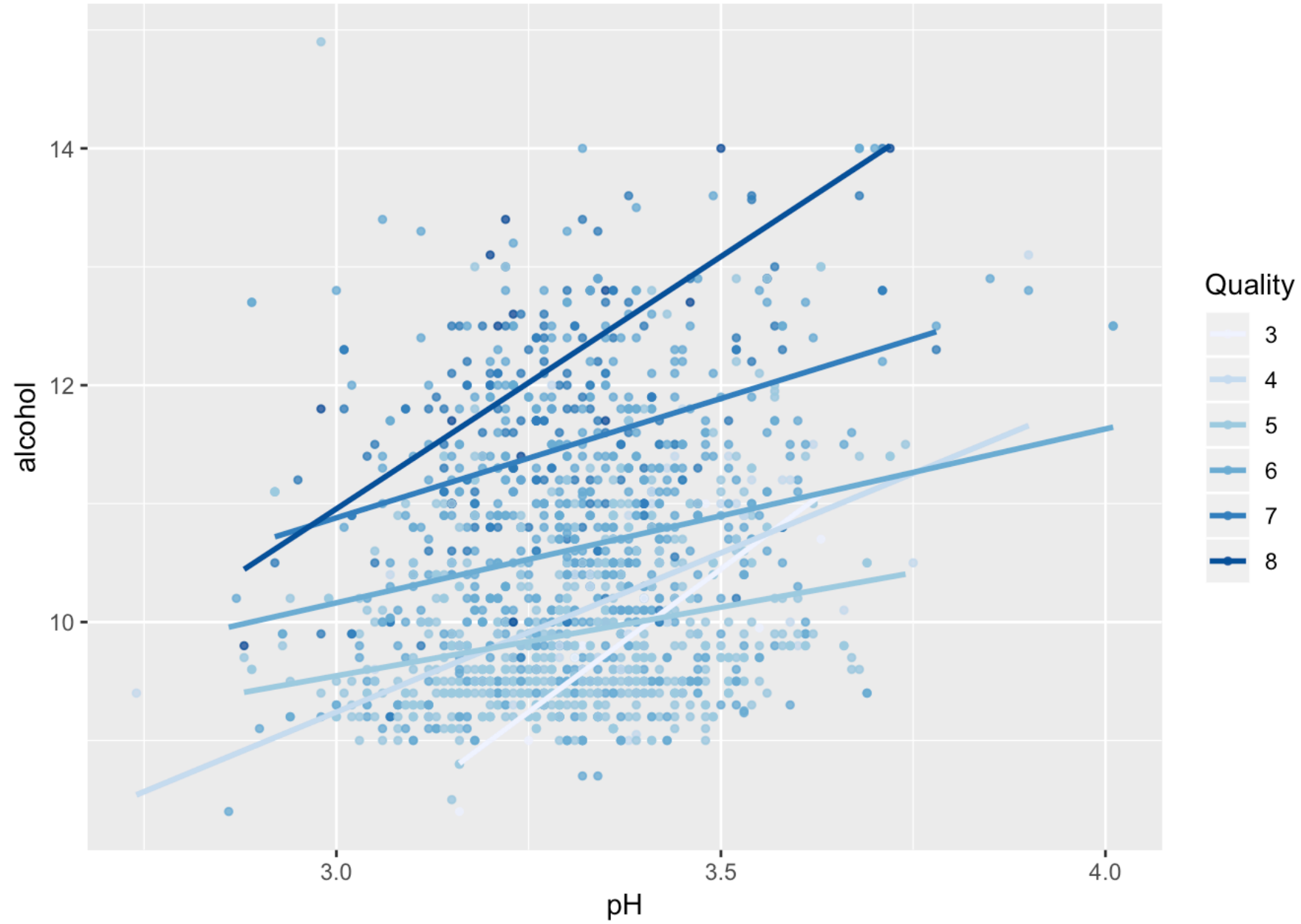
Let's extend the visualization of relation between density and alcohol from previous section also adding a residual sugar class (less than median is low sugar and greater than median is high sugar.)



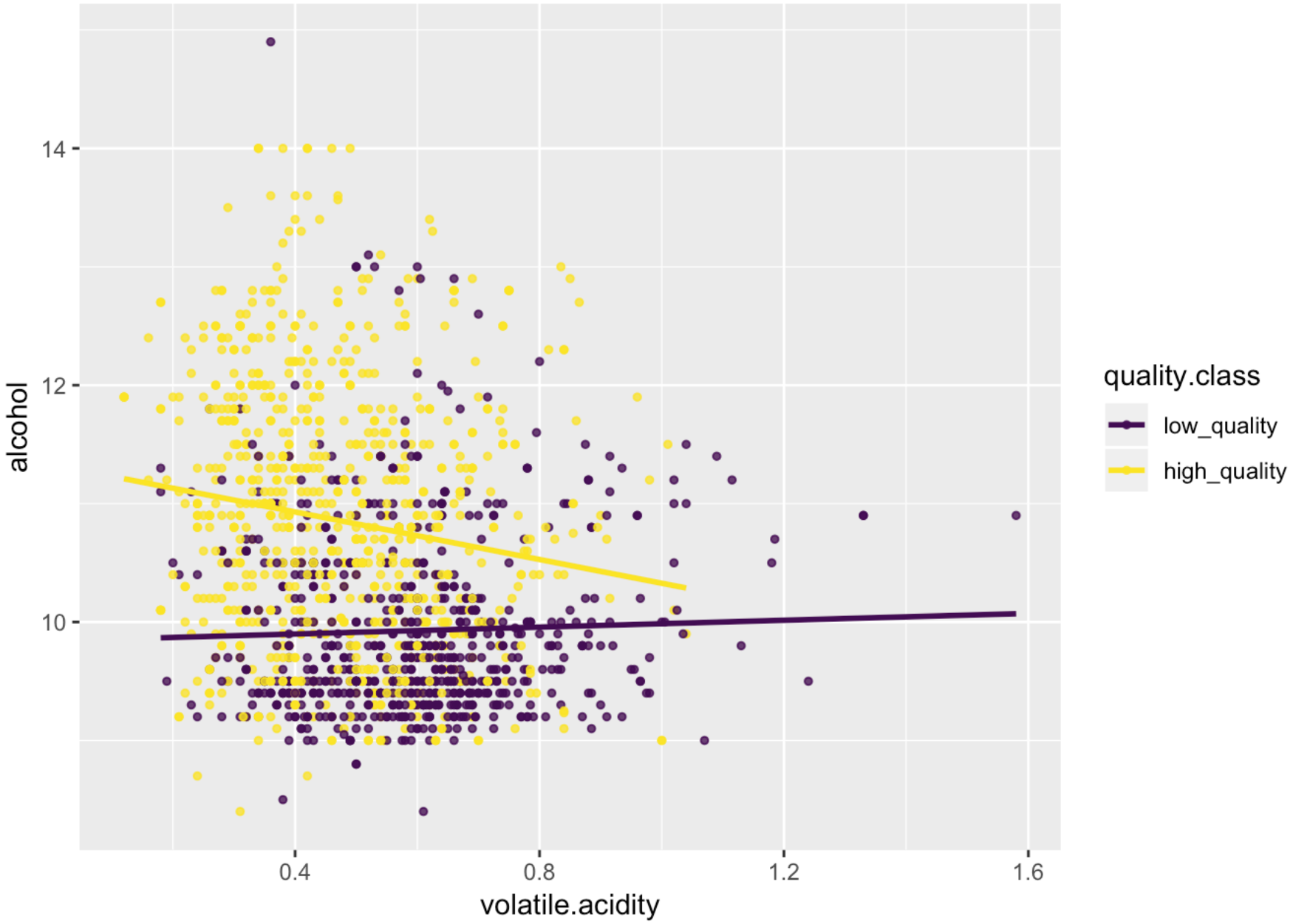
Sweeter wines with higher sugar are also associated with greater density. However, higher alcohol levels are associated with lower density.



This plot show further shows the effect of density and fixed acidity on quality. The darker colors are invariable on the lower envelope of points showing that for a given fixed acidity, lighter density wines are higher quality.



In this plot, the darker points are mostly at the top, showing the effects of two of the strongest indicators of quality that we identified, namely alcohol and acidity levels. This shows that hight quality wines are mostly high on alcohol and occur on lower pH levels.



This shows a positive indicator of quality i.e. alcohol vs negative indicator of quality namely volatile acidity and clearly the high quality wines are high on alcohol and low on volatile acidity.

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?

Lower pH levels, lower volatile acidity and higher alcohol levels strengthened each other in improving wine quality.

Were there any interesting or surprising interactions between features?

Surprisingly residual sugar did not play a role in quality determination even though it increased density thereby potentially lowering quality rating.

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

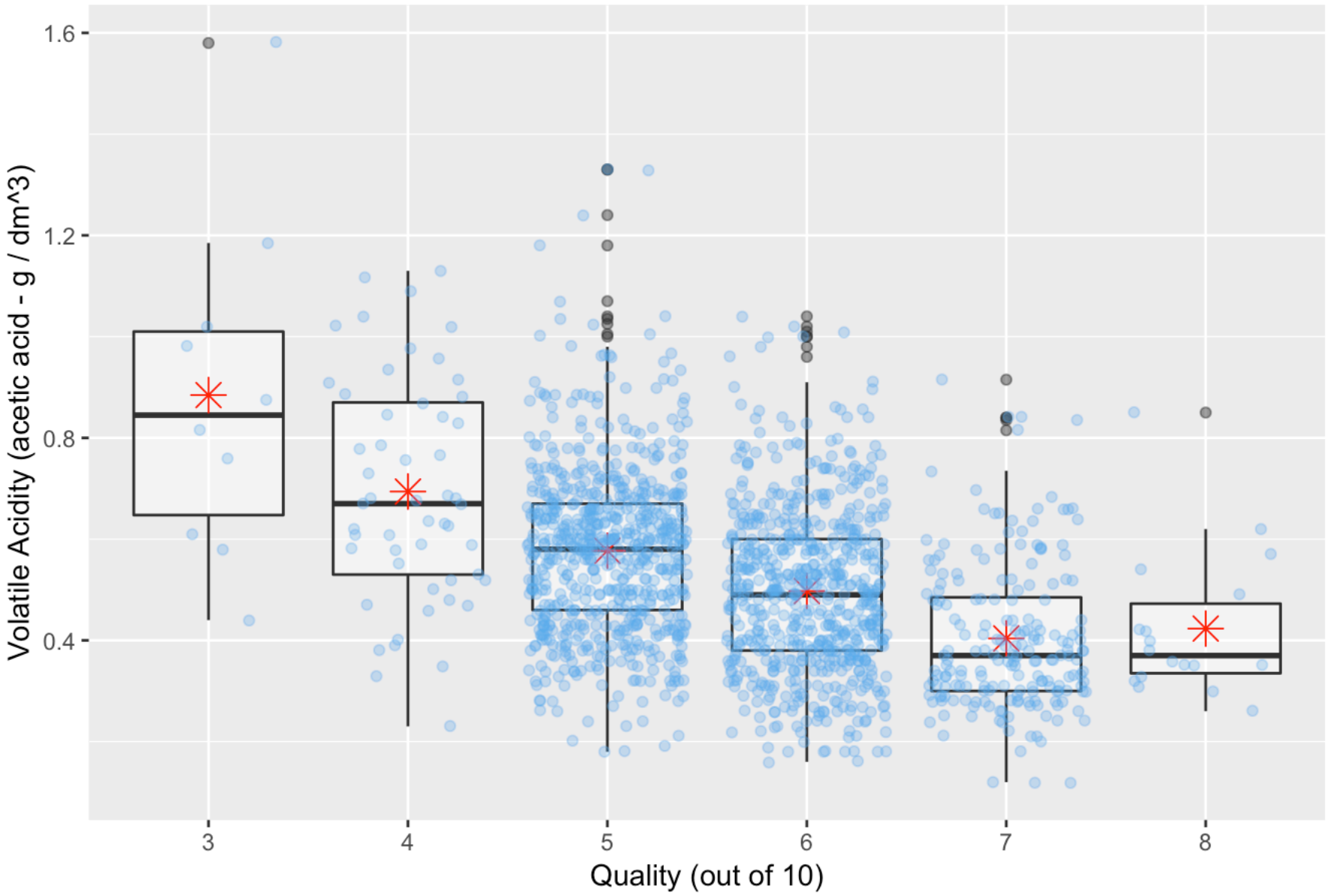
N/A

Final Plots and Summary

Plot One

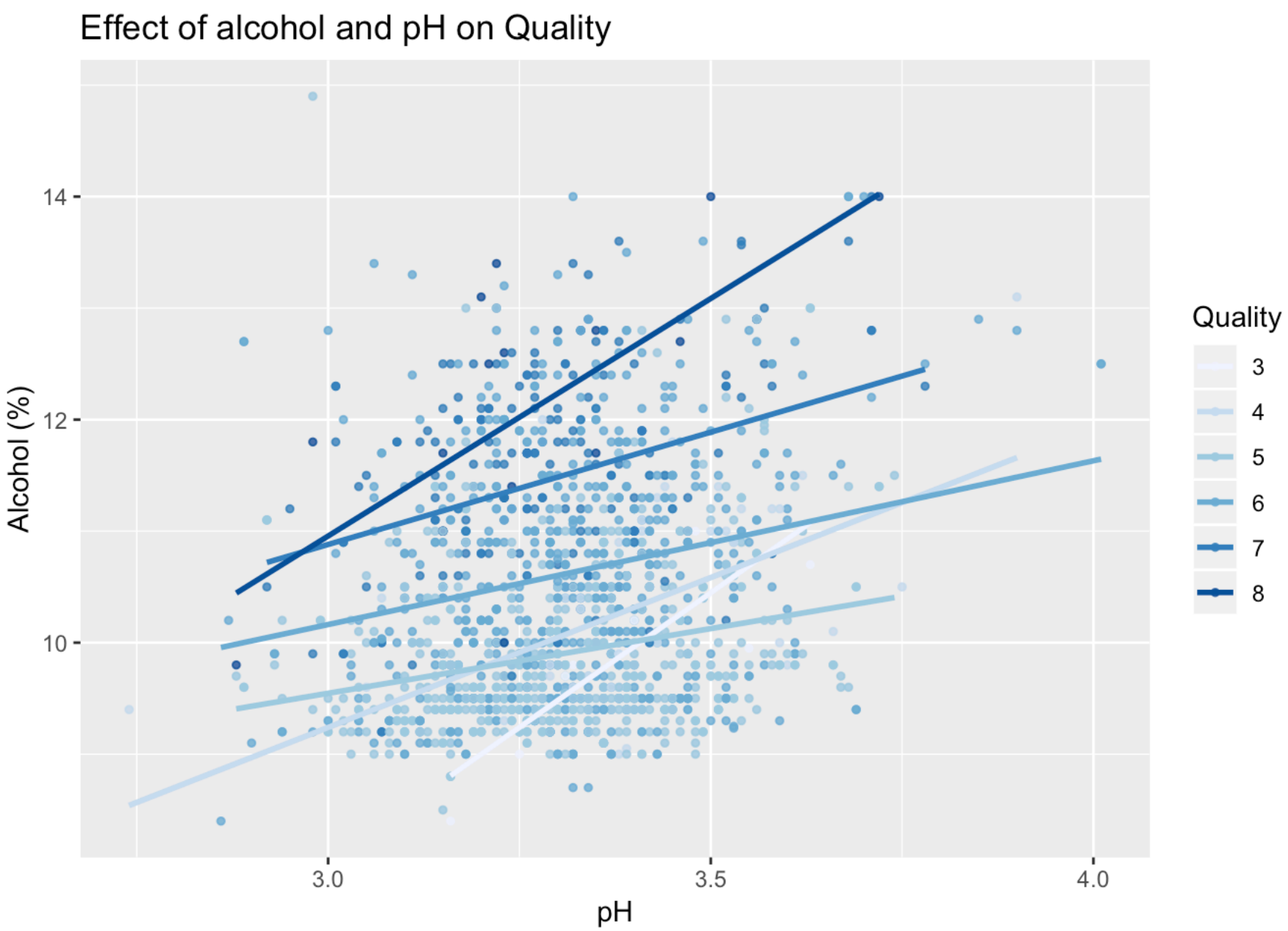
This plot is interesting because it explains the relationship of sugar and alcohol to density which was alluded to in wineQualityInfo.txt. Higher sugar wines seem to have a consistent lead in density compared to wines in lower half in residual sugar range. Alcohol on the other hand, brings down the density at higher levels. This also runs counter to intuition that sweeter wines would be better received, since we see that less dense and higher alcohol level wines are normally preferred in this dataset.

Plot Two



This graph succintly captures the degradation in wine quality with increasing volatile acidity.

Plot Three



This final plot shows the remaining pieces of the trend in high quality wines of having higher alcohol levels and lower pH levels indicating that acidic and more concentrated wines are preferred.

Reflection

The biggest challenge in dealing with this dataset was the lack of domain knowledge on the account of being a teetotaler. However, exploring one variable at a time and faceting by quality class to identify candidates for two variable and then faceting again during two variable analysis to identify three variable interactions, seem to uncover the relationships organically. Scatter matrix was particulary helpful as were summary stats in discovering trend.