Red wine data exploration by Miri Cho

Univariate Plots Section

```
##   X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 1           7.4             0.70        0.00            1.9     0.076
## 2 2           7.8             0.88        0.00            2.6     0.098
## 3 3           7.8             0.76        0.04            2.3     0.092
## 4 4          11.2             0.28        0.56            1.9     0.075
## 5 5           7.4             0.70        0.00            1.9     0.076
## 6 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
## 4                  17                   60  0.9980 3.16      0.58     9.8
## 5                  11                   34  0.9978 3.51      0.56     9.4
## 6                  13                   40  0.9978 3.51      0.56     9.4
##   quality
## 1       5
## 2       5
## 3       5
## 4       6
## 5       5
## 6       5
```

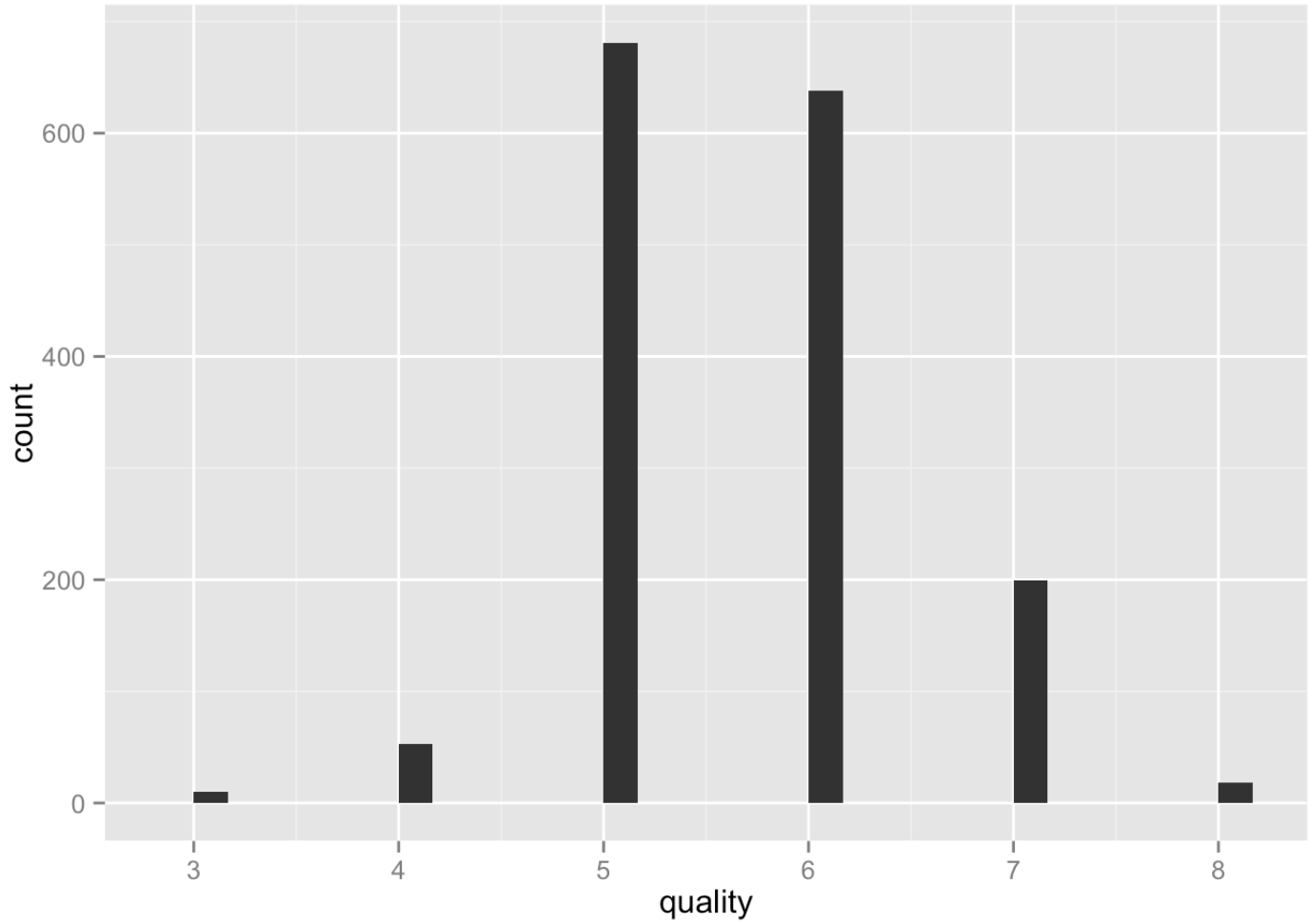
First few lines of the dataset.

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

Data variable types. All of them are numbers and integers (X, quality).

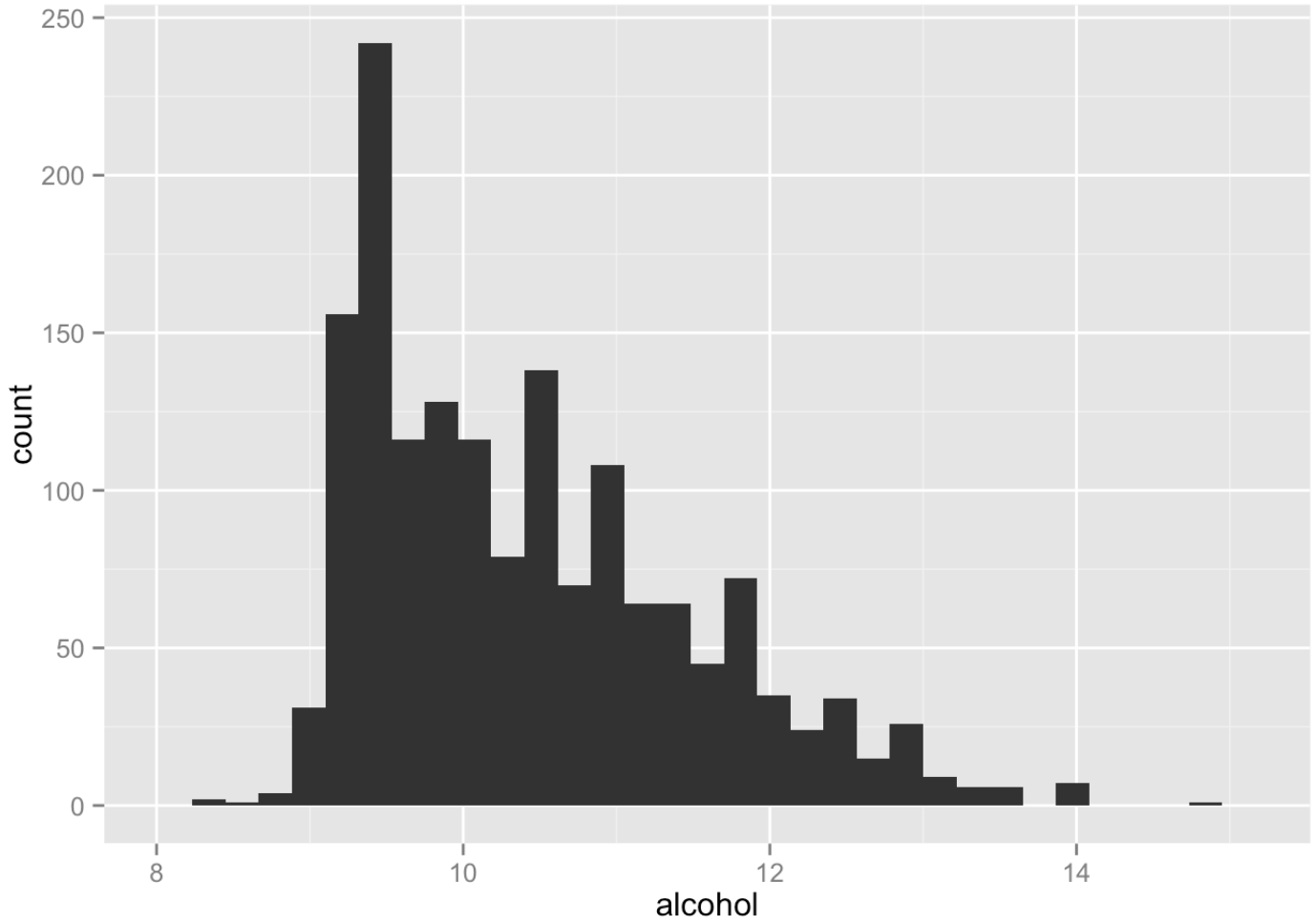
```
##        X           fixed.acidity   volatile.acidity  citric.acid
##  Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000
##  1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090
##  Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260
##  Mean   : 800.0   Mean   : 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
```

Summary statistics of all the variables in the wine data. Just glancing at the statistics shows that some variables (such as total.sulfur.dioxide) have a wider variance than others (such as chlorides).

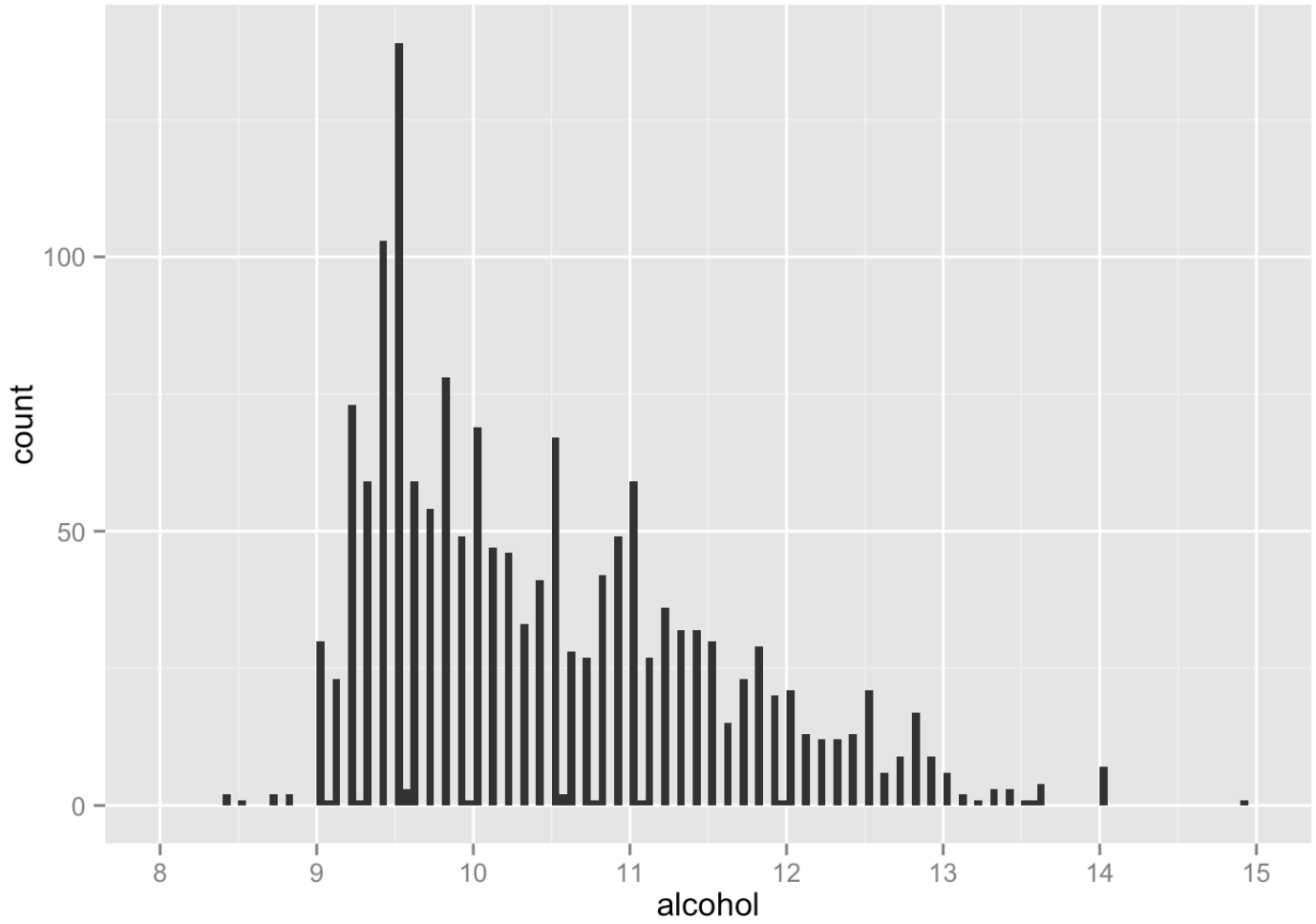



```
##
##   3   4   5   6   7   8
##  10  53 681 638 199  18
```

A histogram of wine quality. Quality is integers with values ranging from 3 to 8. Most of them are 5 or 6.



A historam of alcohol levels. Alcohol doesn’t seem to be all integers. Reducing the binwidth shows this below.

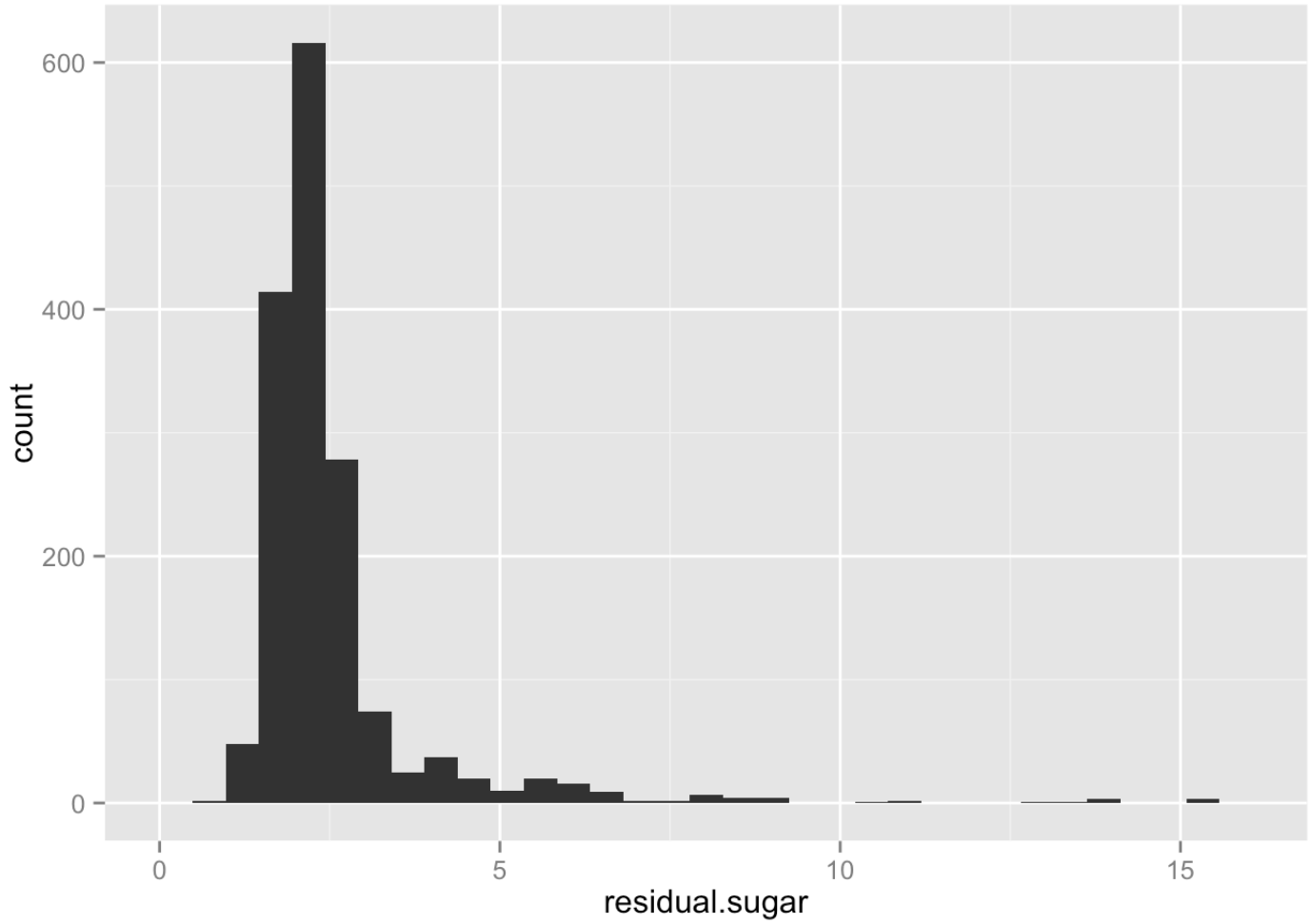



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```

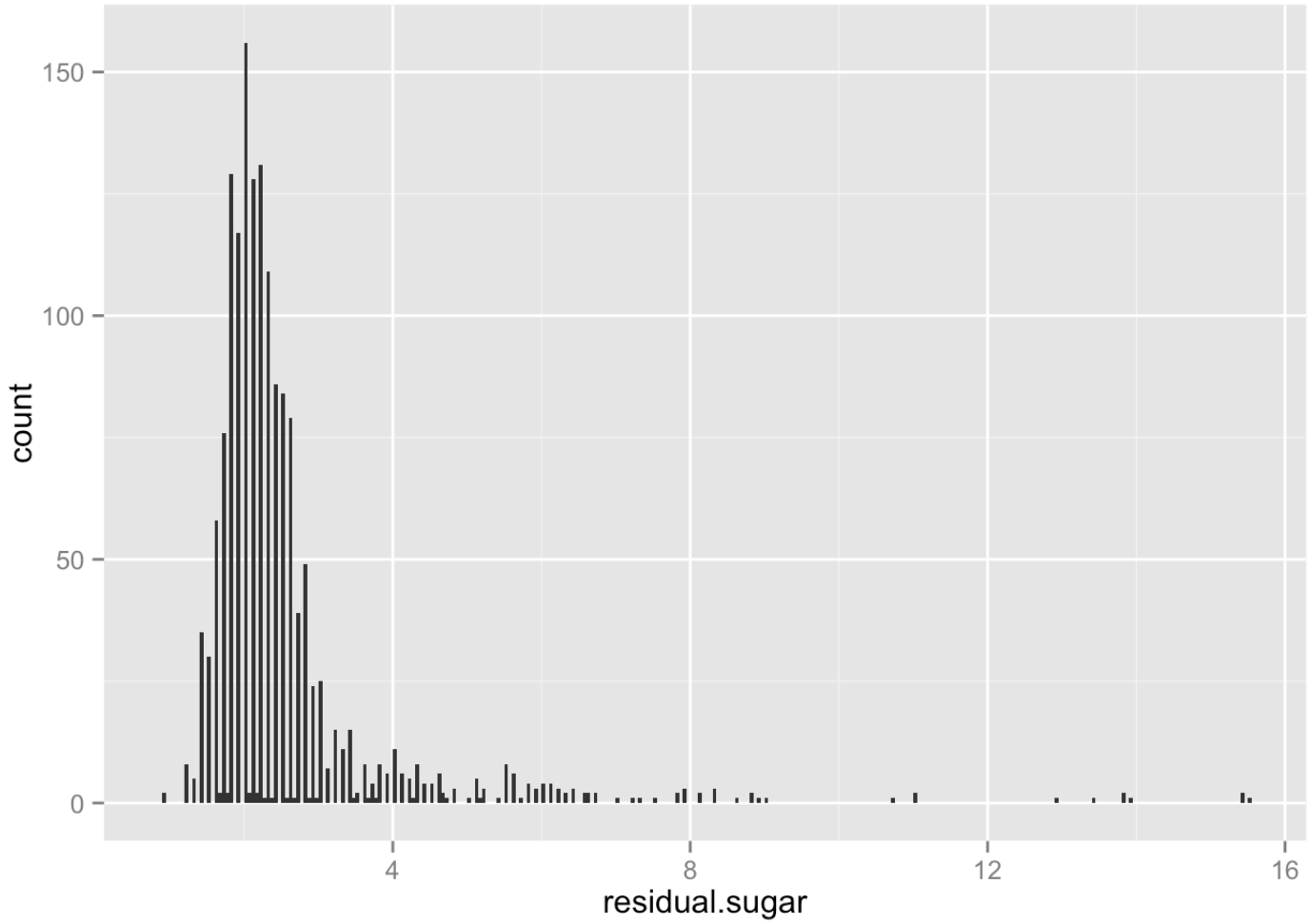
```
##
##               8.4               8.5               8.7               8.8
##                 2                 1                 2                 2
##                 9              9.05               9.1               9.2
##                30                 1                23                72
## 9.23333333333333              9.25               9.3               9.4
##                 1                 1                59               103
##               9.5              9.55 9.56666666666667               9.6
##               139                 2                 1                59
##               9.7               9.8               9.9              9.95
##                54                78                49                 1
##                10 10.0333333333333              10.1              10.2
##                67                 2                47                46
##              10.3              10.4              10.5             10.55
##                33                41                67                 2
##              10.6              10.7             10.75              10.8
##                28                27                 1                42
##              10.9                11 11.0666666666667              11.1
##                49                59                 1                27
##              11.2              11.3              11.4              11.5
##                36                32                32                30
##              11.6              11.7              11.8              11.9
##                15                23                29                20
##             11.95                12              12.1              12.2
##                 1                21                13                12
##              12.3              12.4              12.5              12.6
##                12                13                21                 6
##              12.7              12.8              12.9                13
##                 9                17                 9                 6
##              13.1              13.2              13.3              13.4
##                 2                 1                 3                 3
##              13.5 13.5666666666667              13.6                14
##                 1                 1                 4                 7
##              14.9
##                 1
```

Summary statistics and table of factor levels of alcohol. The most frequent alcohol levels are between 9 and 10. A half of them are 10.20 and below.

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

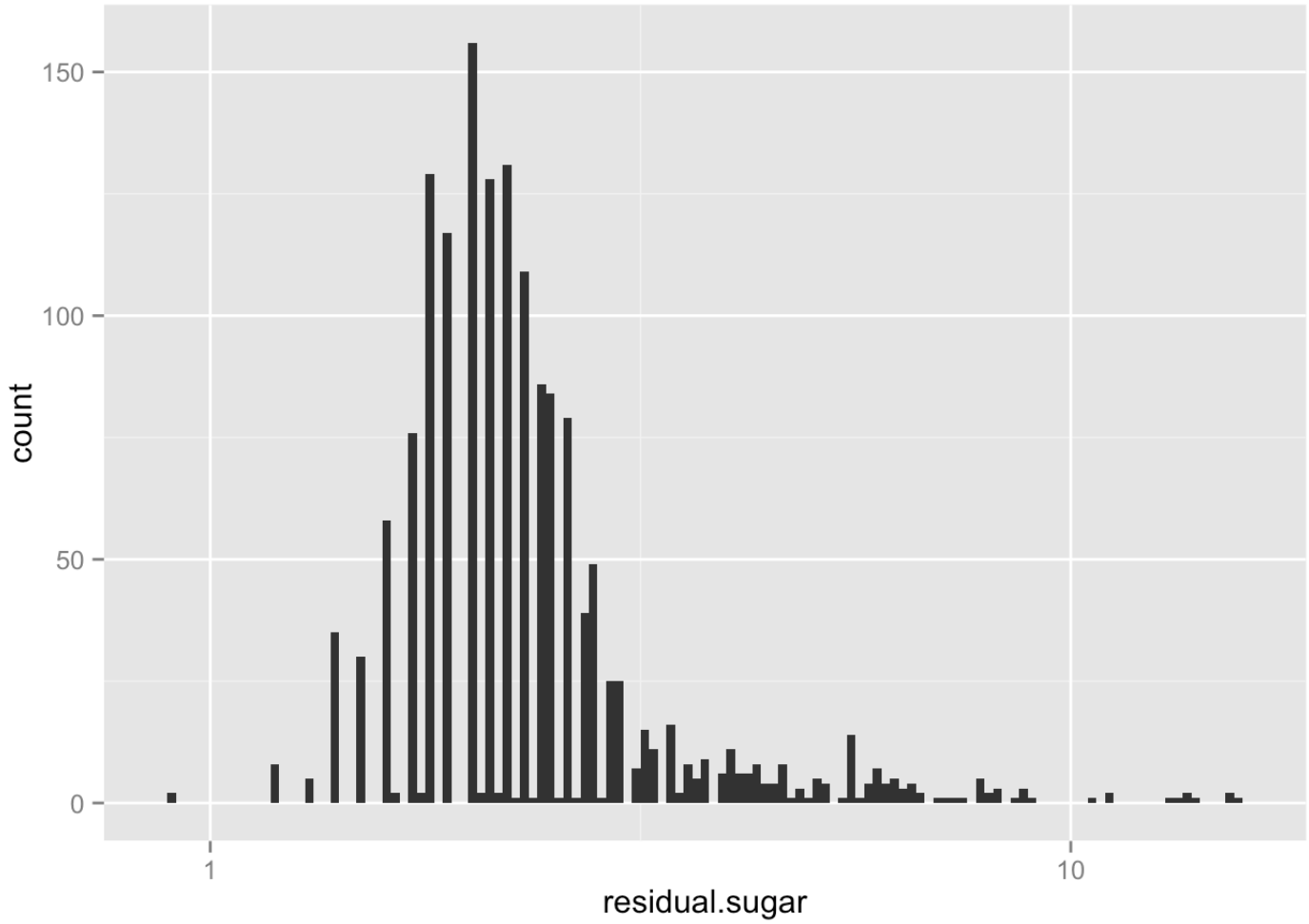


Summary statistics and histogram of residual sugar. 75% of residual sugar is 2.6 or less. The histogram is right skewed a long tail.

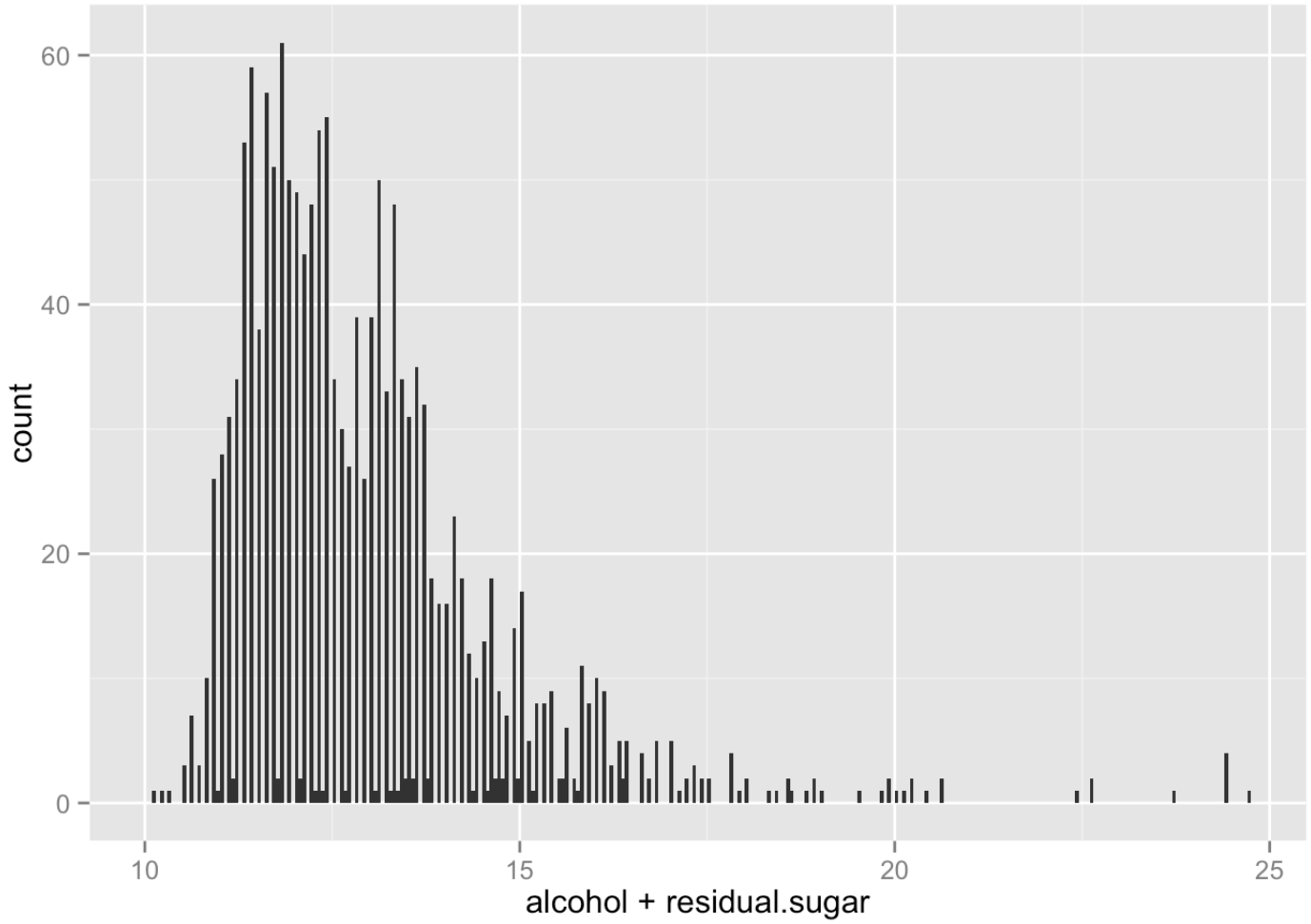


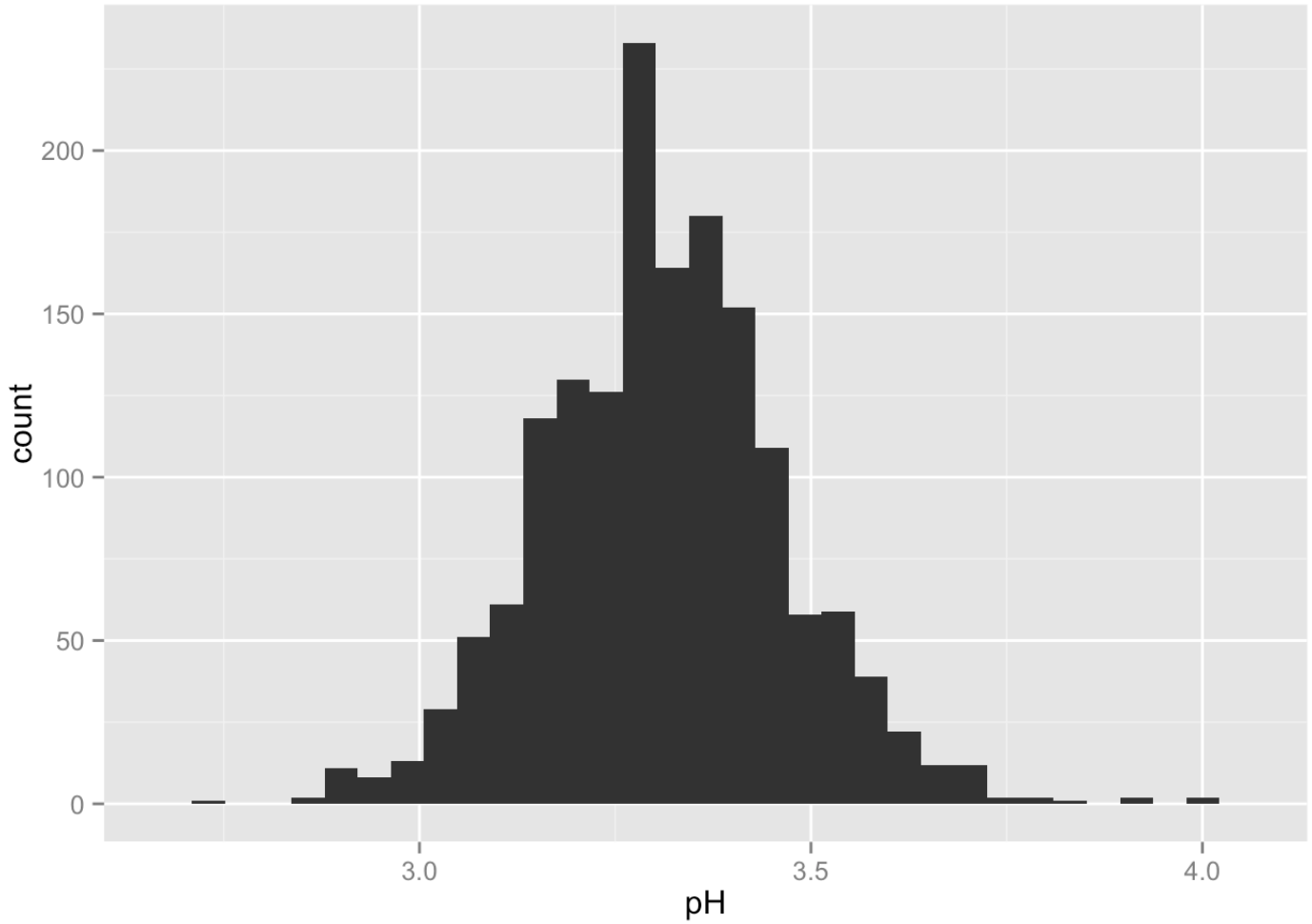
Reducing the bin width of the historam for residual sugar shows the long tail better.

The below is the histogram of log-transformed residual sugar. This makes it easier to visualize to overcome the long tail issue.



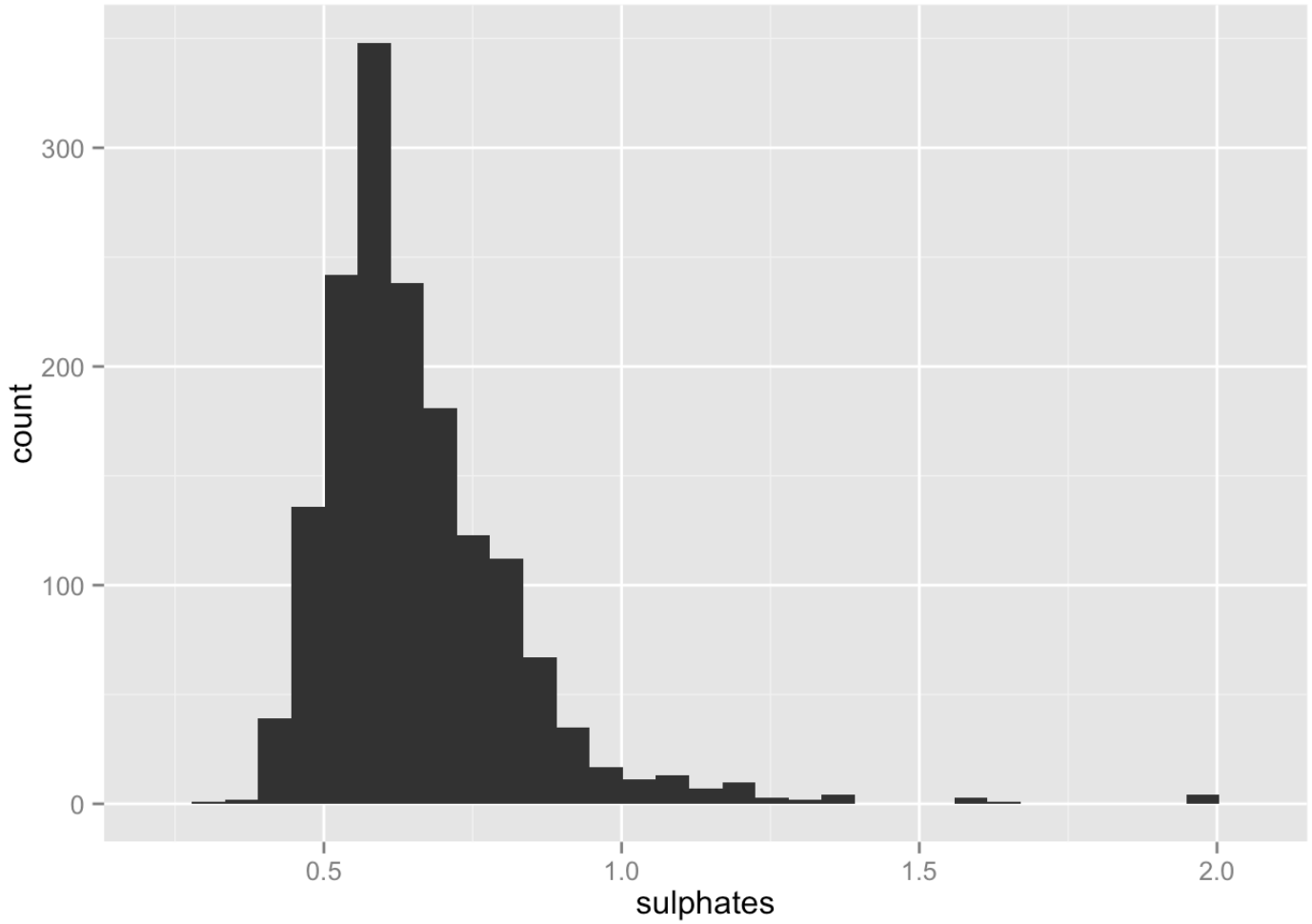
Creating a new variable (alcohol+residual.sugar). The histogram below shows a long tail from residual sugar.


60
40
20
0
count
10
15
20
25
alcohol + residual.sugar



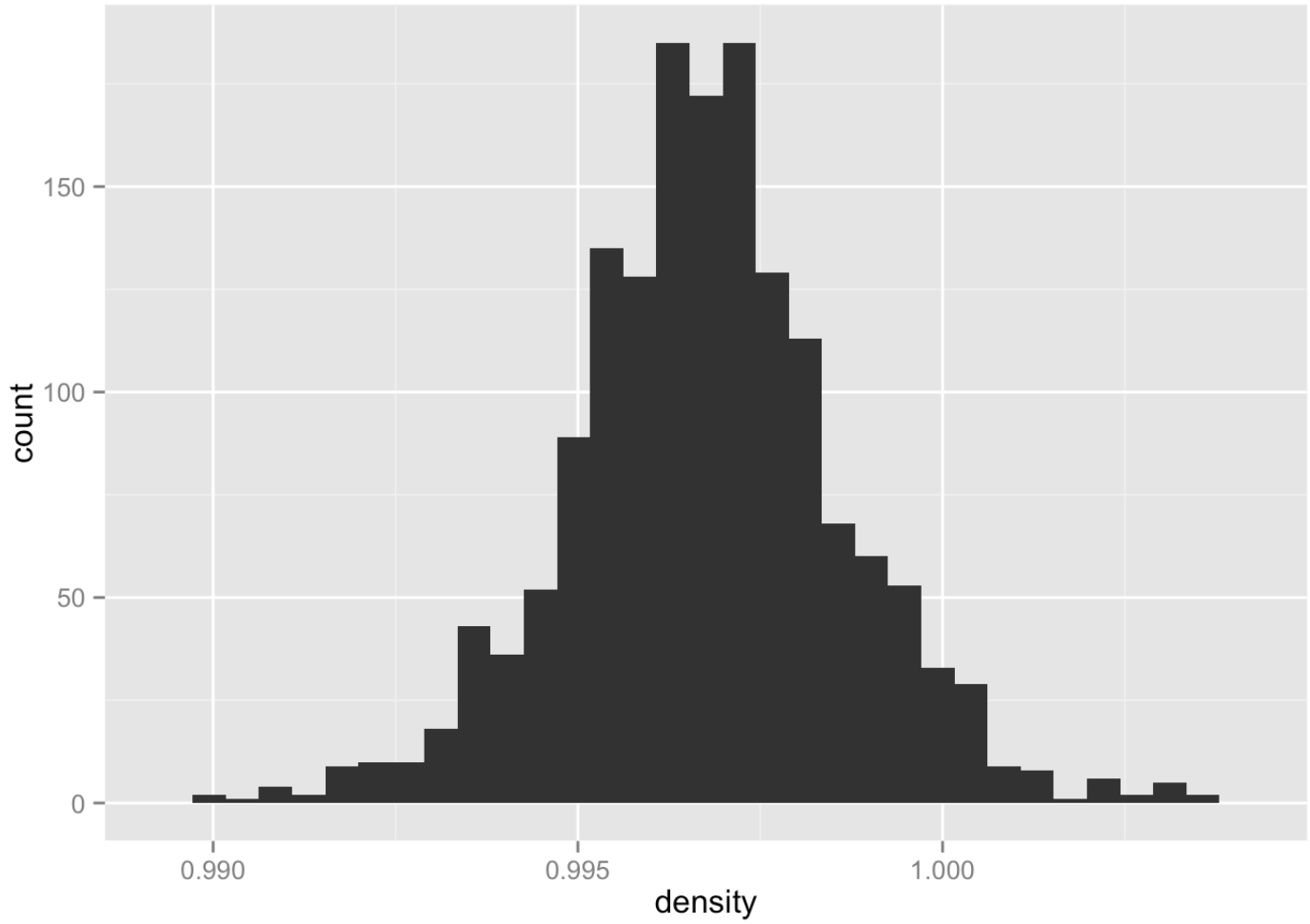

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010
```

A histogram of pH levels and summary statistics. pH level is relatively normally distributed with the middle half (25% to 75%) between 3.210 and 3.4.



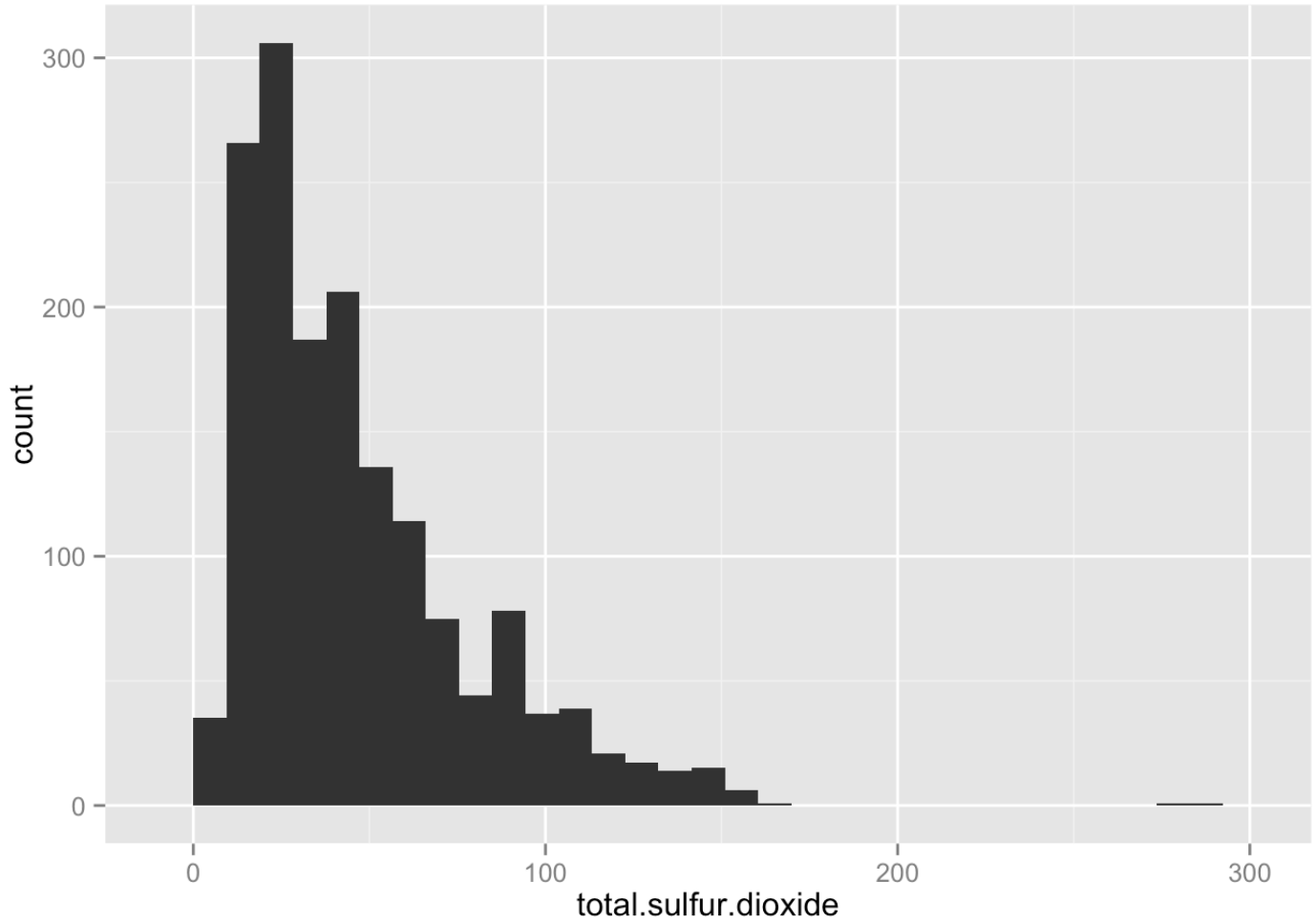

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```

A histogram of sulphates. This is also right skewed a long tail. Summary statistics show that 75% of values are between 0.33 and 0.73.



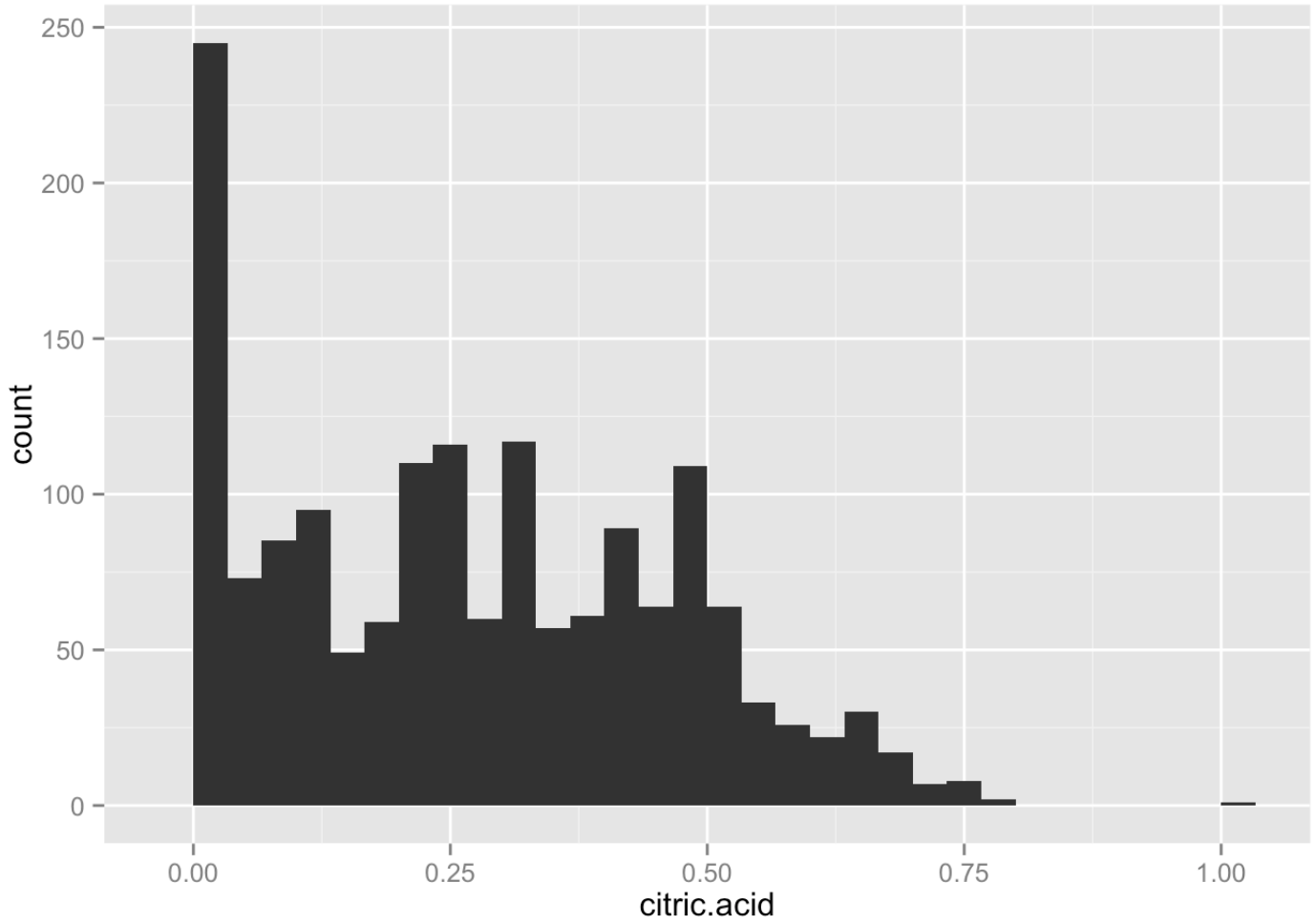

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0040
```

The histogram of density seems normally distributed with the middle half ranging from 0.9956 to 0.9978.

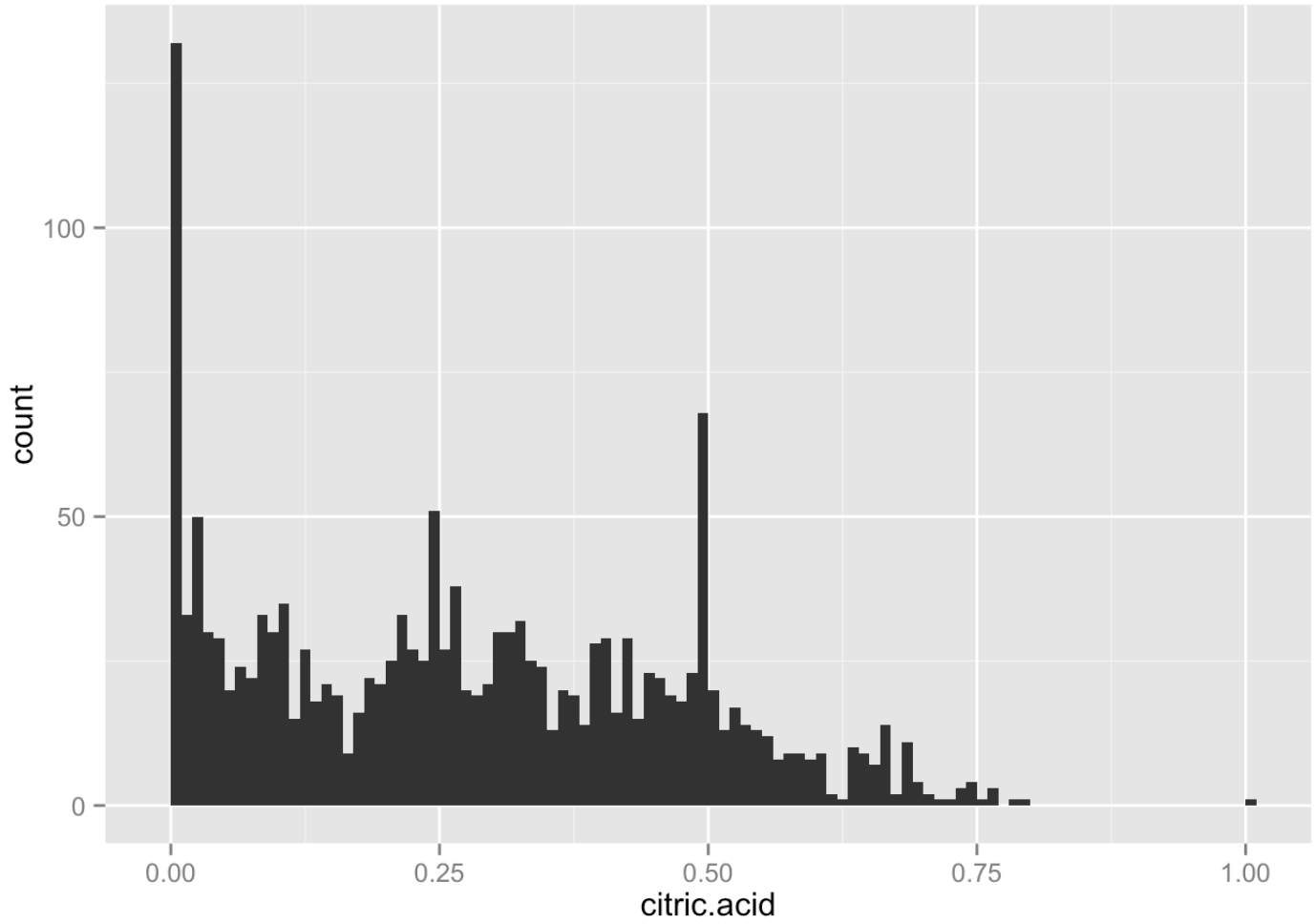



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00
```

The histogram of total sulfur dioxide is right skewed. Summary statistics show that the first half are between 6 and 38. The max is 289.



The histogram of citric acid shows a non-parametric distribution. Reducing the bin width shows this better.There is also a large number of 0’s.




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

Summary statistics of citric acid.

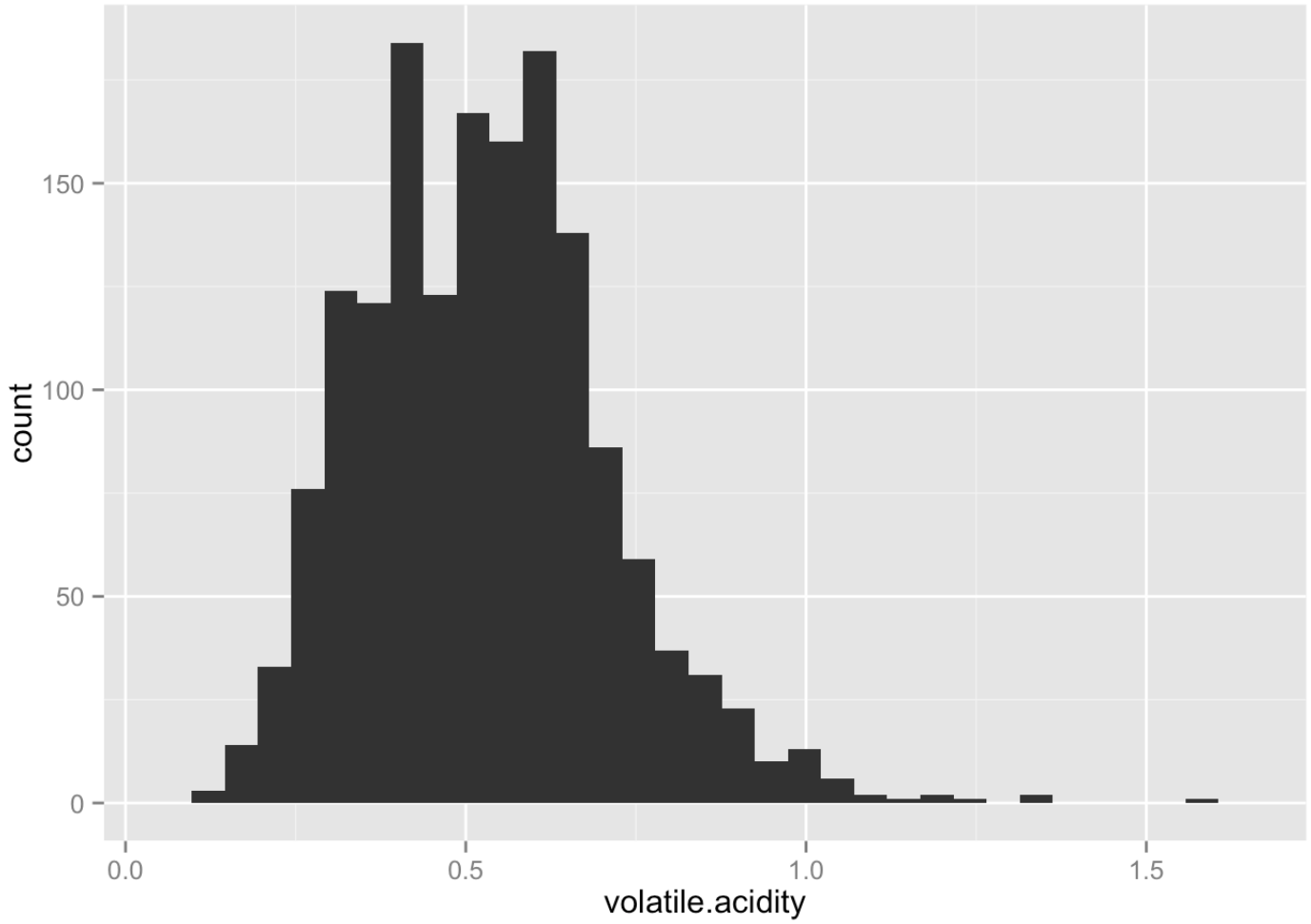
```
##    Mode   FALSE    TRUE    NA's
## logical    1467     132       0
```

Summary statistics of citric acid that is equal to 1.

```
##    Mode   FALSE    TRUE    NA's
## logical     132    1467       0
```

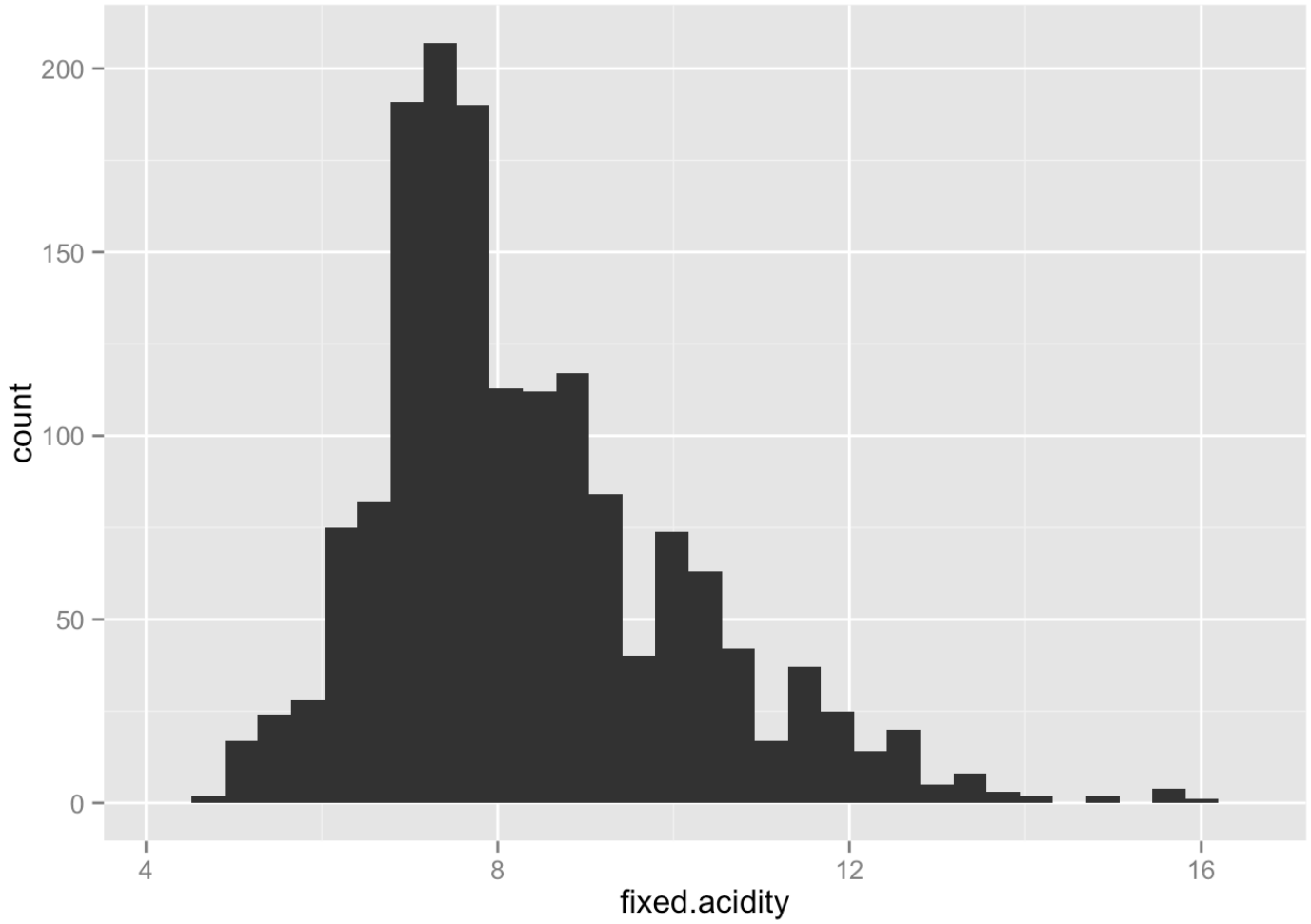
Summary statistics of citric acid that is NOT equal to 1.

About 8% of the wines do not contain any citric acid. Only one of them has the max value of 1.



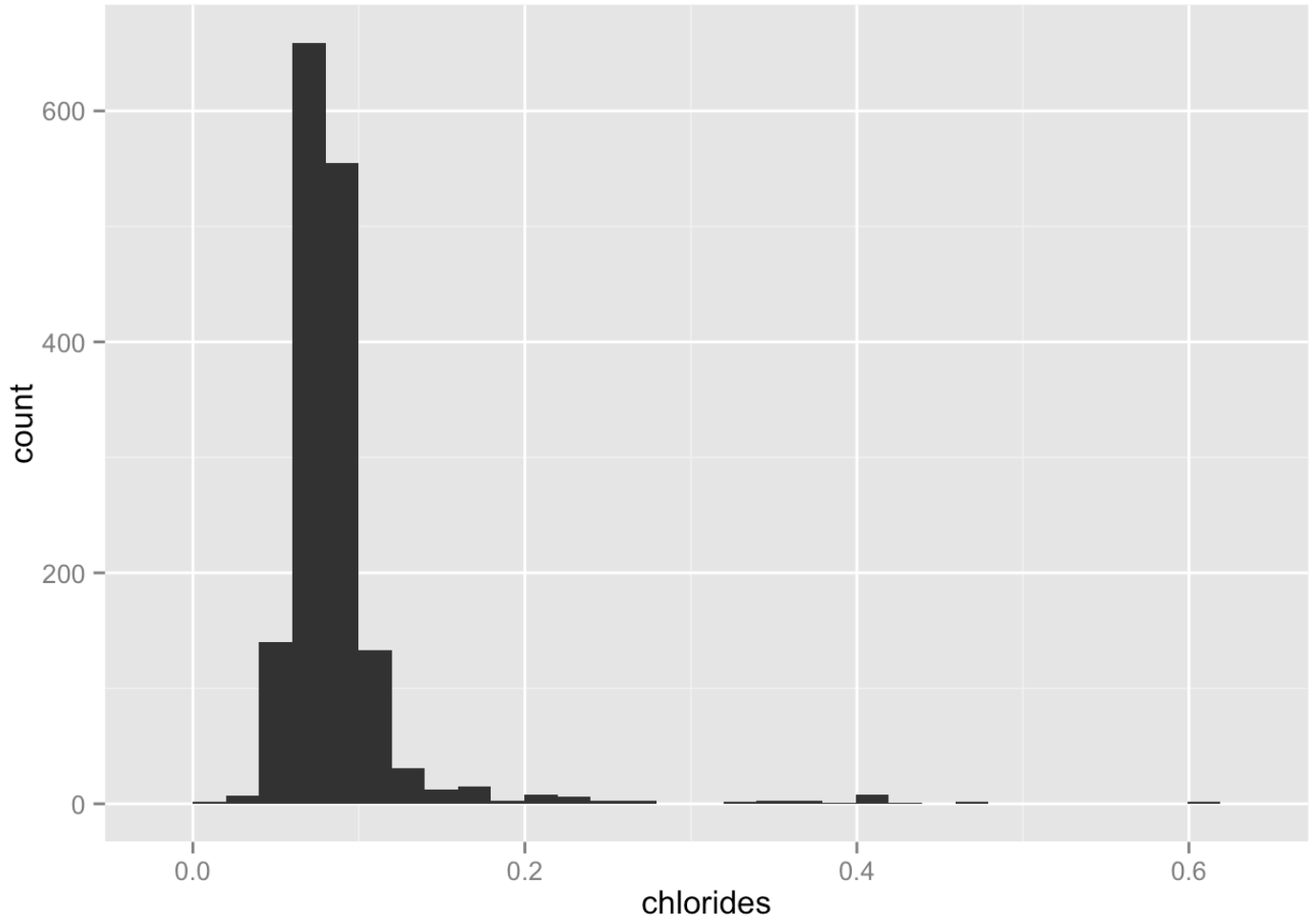

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```

Histogram and summary statistics of volatile acidity. It’s right skewed with 75% of wines having 0.64 or less.



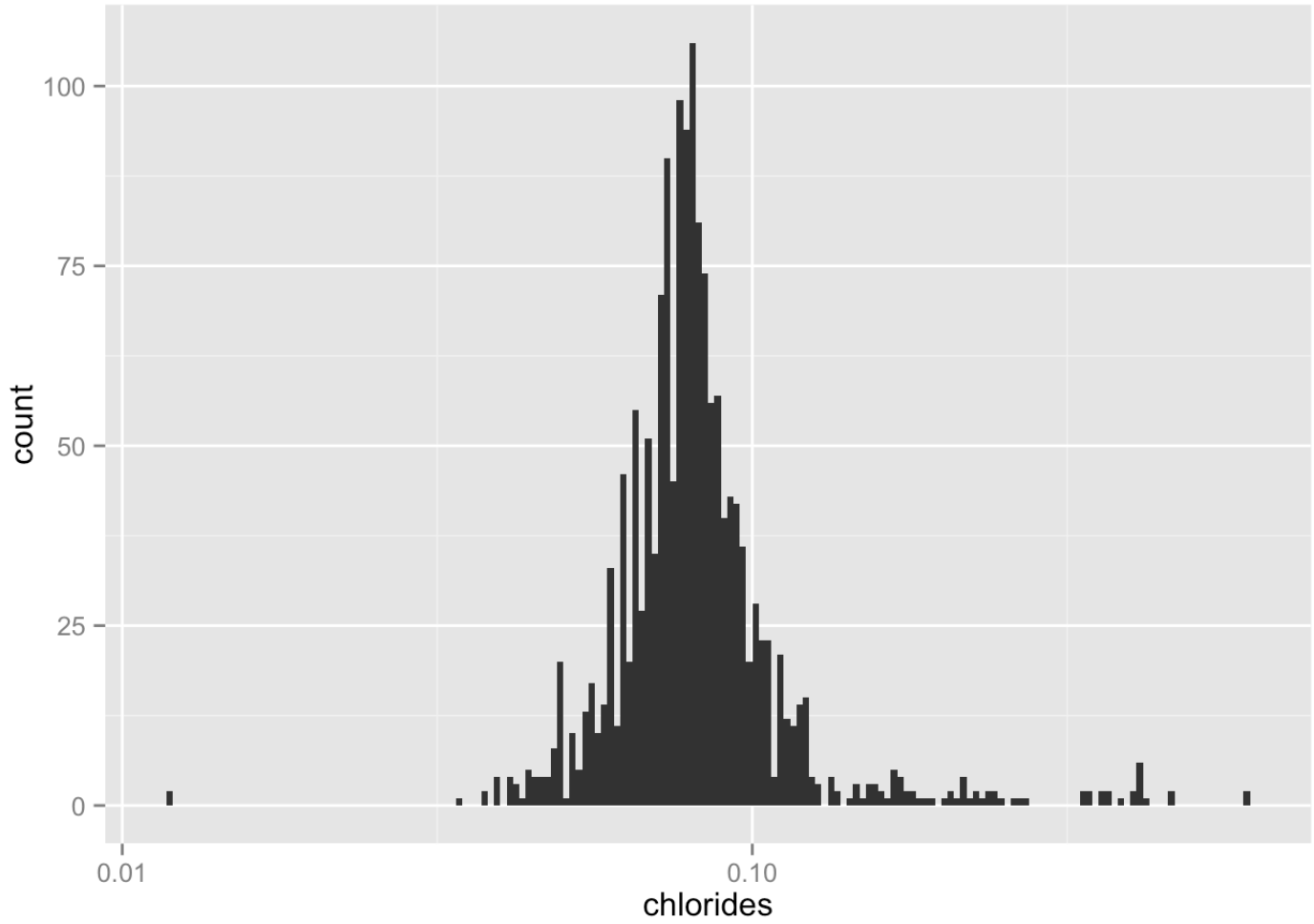

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
```

Histogram and summary statistics of fixed acidity.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```

Histogram and summary statistics of chlorides. This is right-skewed and a long tail.The middle half are between 0.07 and 0.09. The log-transformed chlorides look more normal below.



Univariate Analysis

What is the structure of your dataset?

There are 1599 red wine variants of the Portuguese “Vinho Verde” wine in the dataset with 12 features (the 13th feature(‘x’) is just unique id’s). All of the features are numerical variables and except for the quality feature that is discrete (between 3 and 8), they are all continous.

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

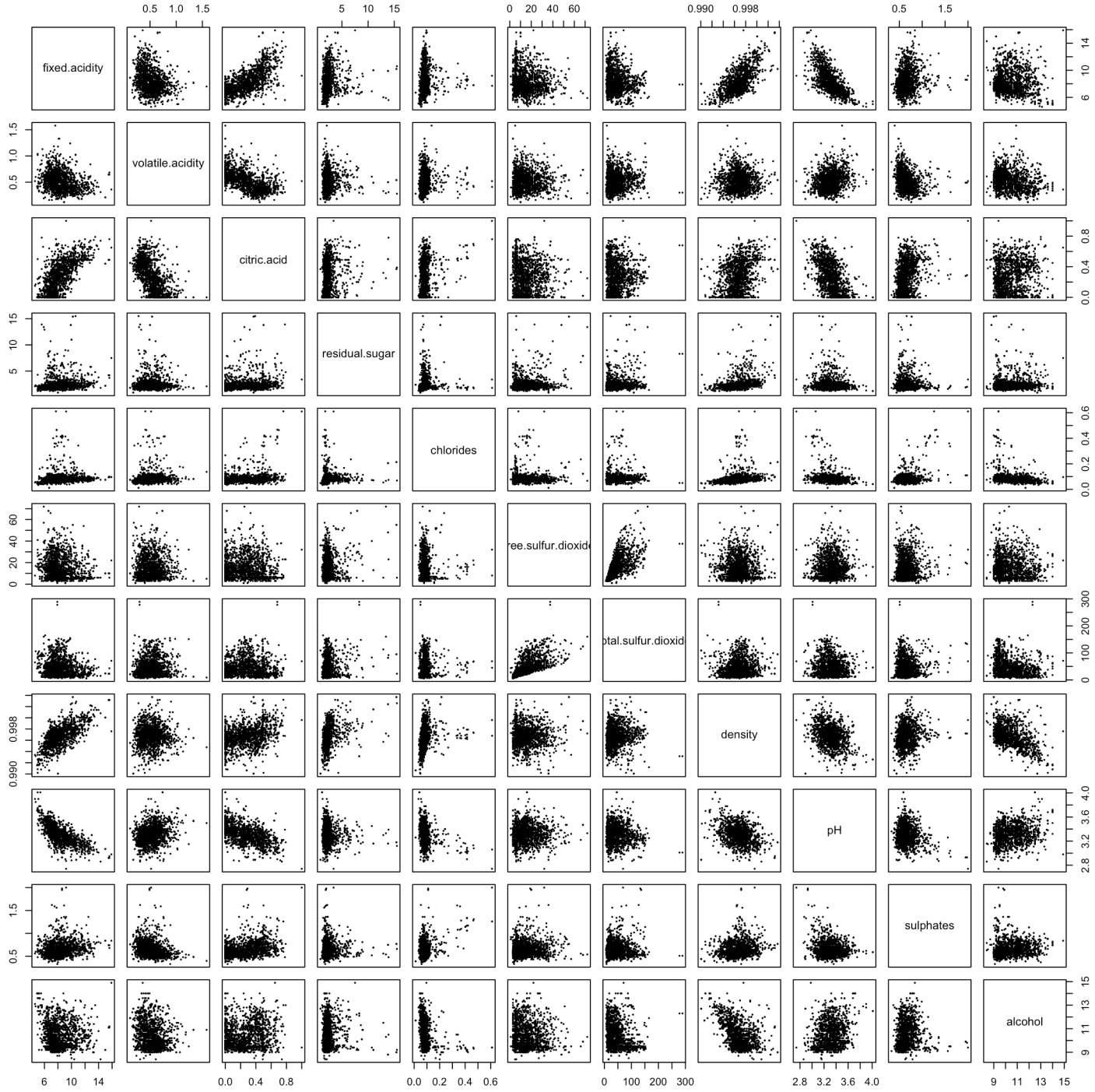
Quality is the output variable and the rest of the 11 features are chemical properties of each wine. I’m interested in finding out which chemical properties influence the quality of red wines. I suspect that combinations of these features would affect the quality more so than individual features.

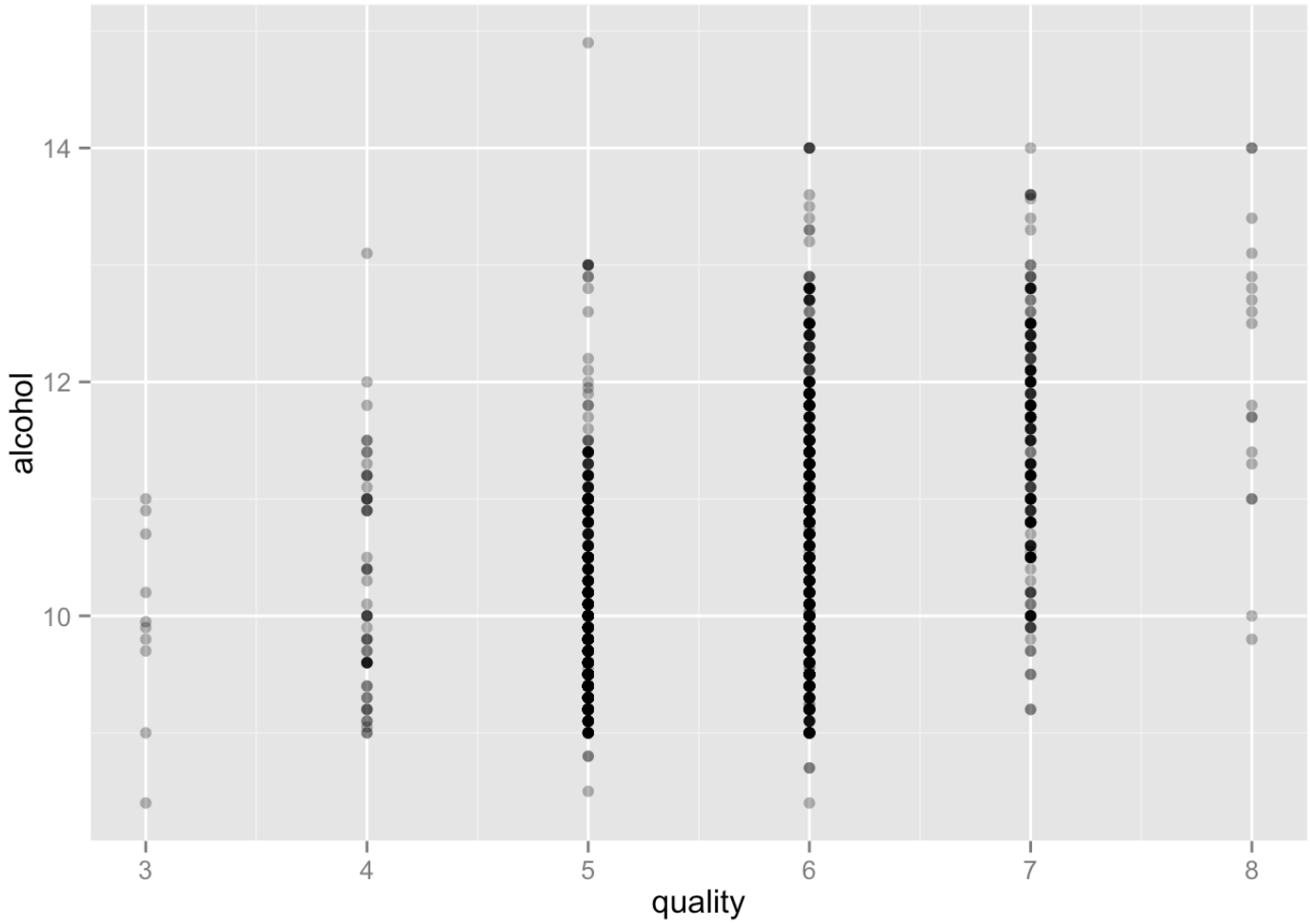
Of the features you investigated, were there any unusual distributions?

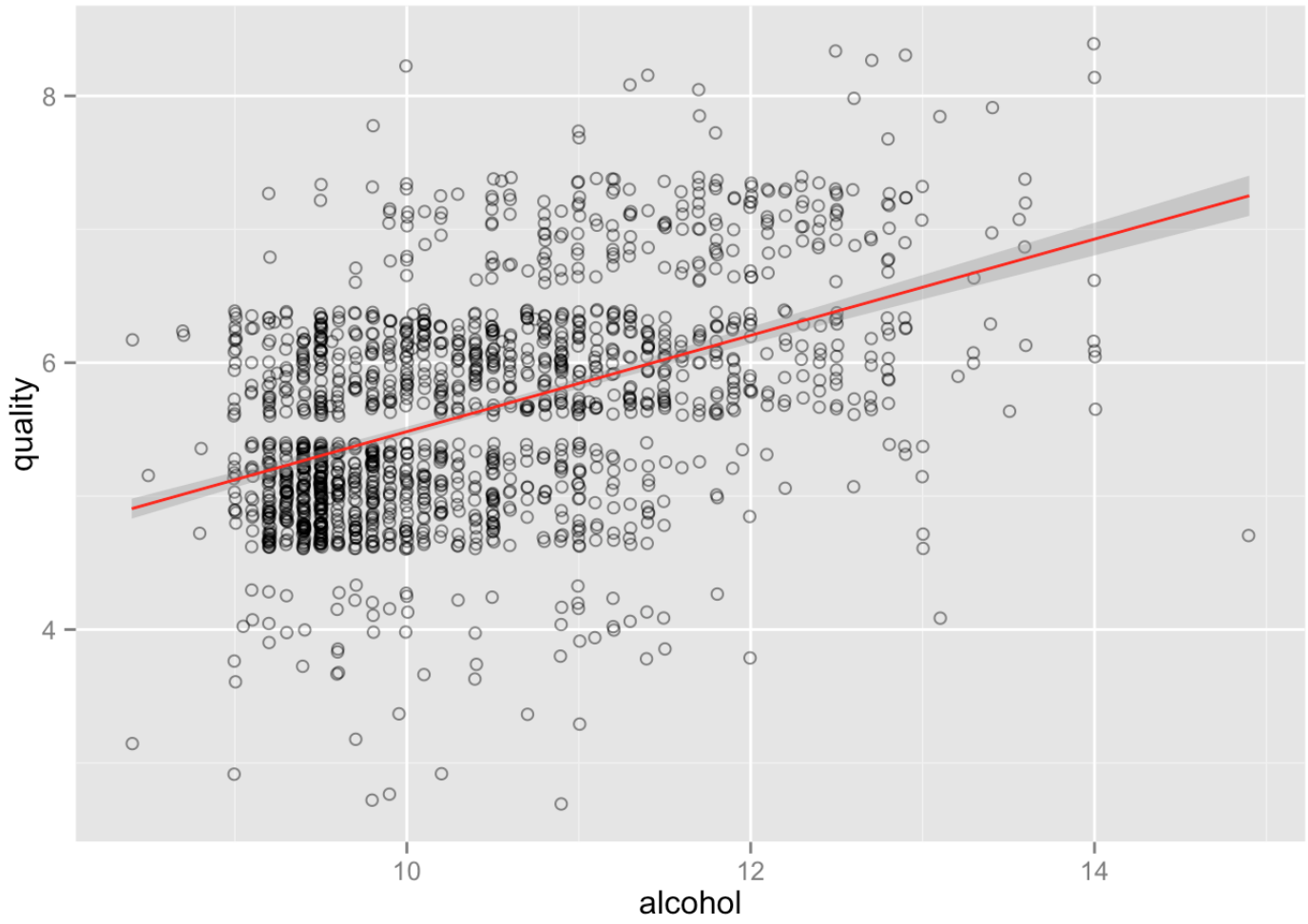
From my initial univariate analysis alone, alcohol, sugar and dioxide are interesting in that the distributions are right skewed or non-parametric in case of citric acid.

Bivariate Plots Section

The correlation matrix shows that alcohol has the highest positive correlation with quality and volatile acidity has the highest negative correlation with quality. Higher correlations between features seem to arise from chemical properties (ie. higher acidity and lower pH)




alcohol
14
12
10
3
4
5
6
7
8
quality

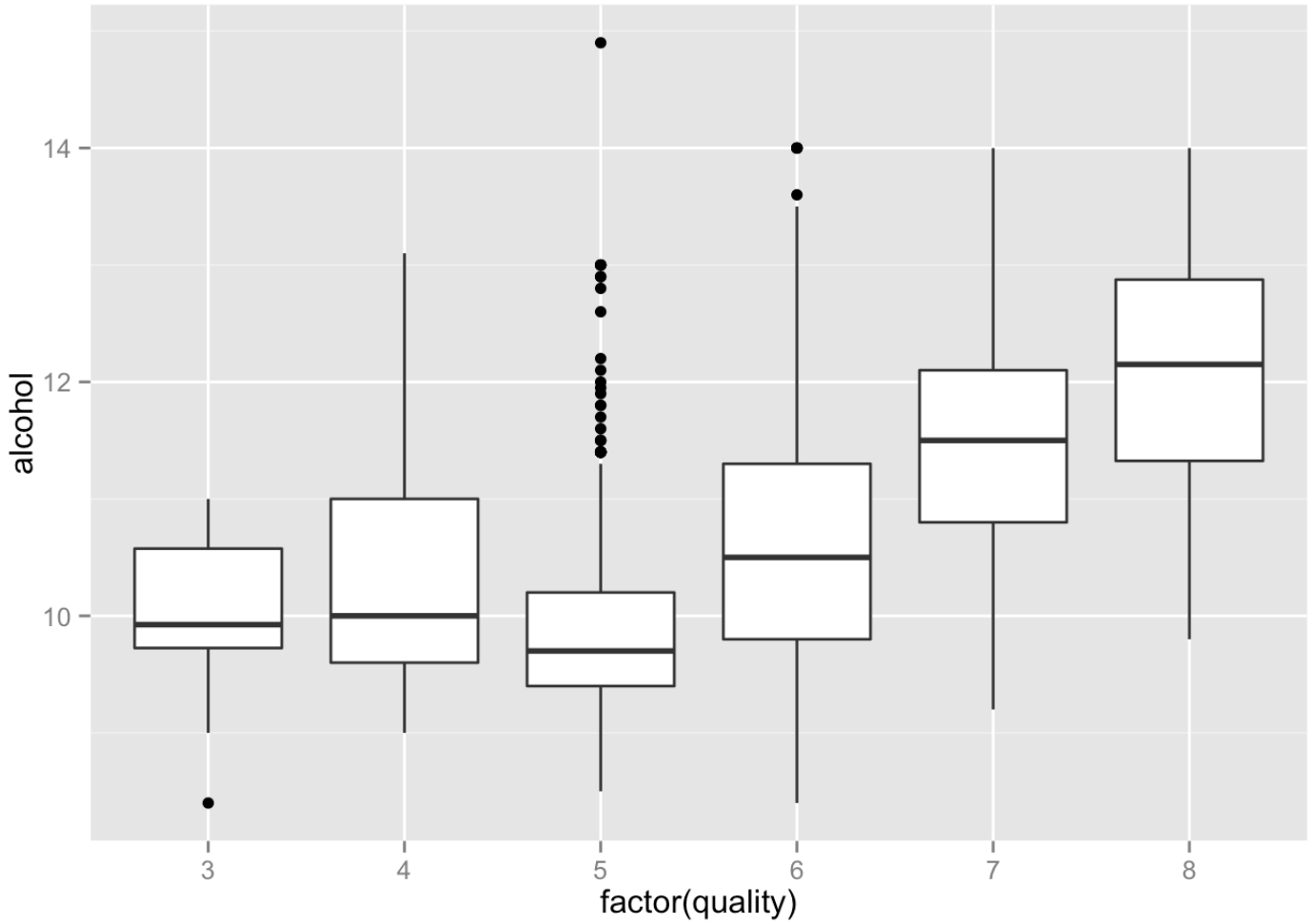

quality
8
6
4
10
12
14
alcohol

Because the plotting quality suffers from overplotting, the above scatterplot added jitter with alpha 1/2. We can see that a majority of alcohol level is less than 12 with quality of 5 and 7. While there is an overal trend of upward slope between alcohol and quality, at lower levels of alcohol, quality has a wide variation.

```
##
## Call:
## lm(formula = quality ~ alcohol, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8442 -0.4112 -0.1690  0.5166  2.5888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.87497    0.17471   10.73   <2e-16 ***
## alcohol      0.36084    0.01668   21.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7104 on 1597 degrees of freedom
## Multiple R-squared:  0.2267, Adjusted R-squared:  0.2263
## F-statistic: 468.3 on 1 and 1597 DF,  p-value: < 2.2e-16
```

Running a linear regression on alcohol and quality shows that alcohol explains 22.6% of variation in quality.

```
##     (0,9.5]  (9.5,10.2] (10.2,11.1] (11.1,14.9]
##         436         406         377         380
```

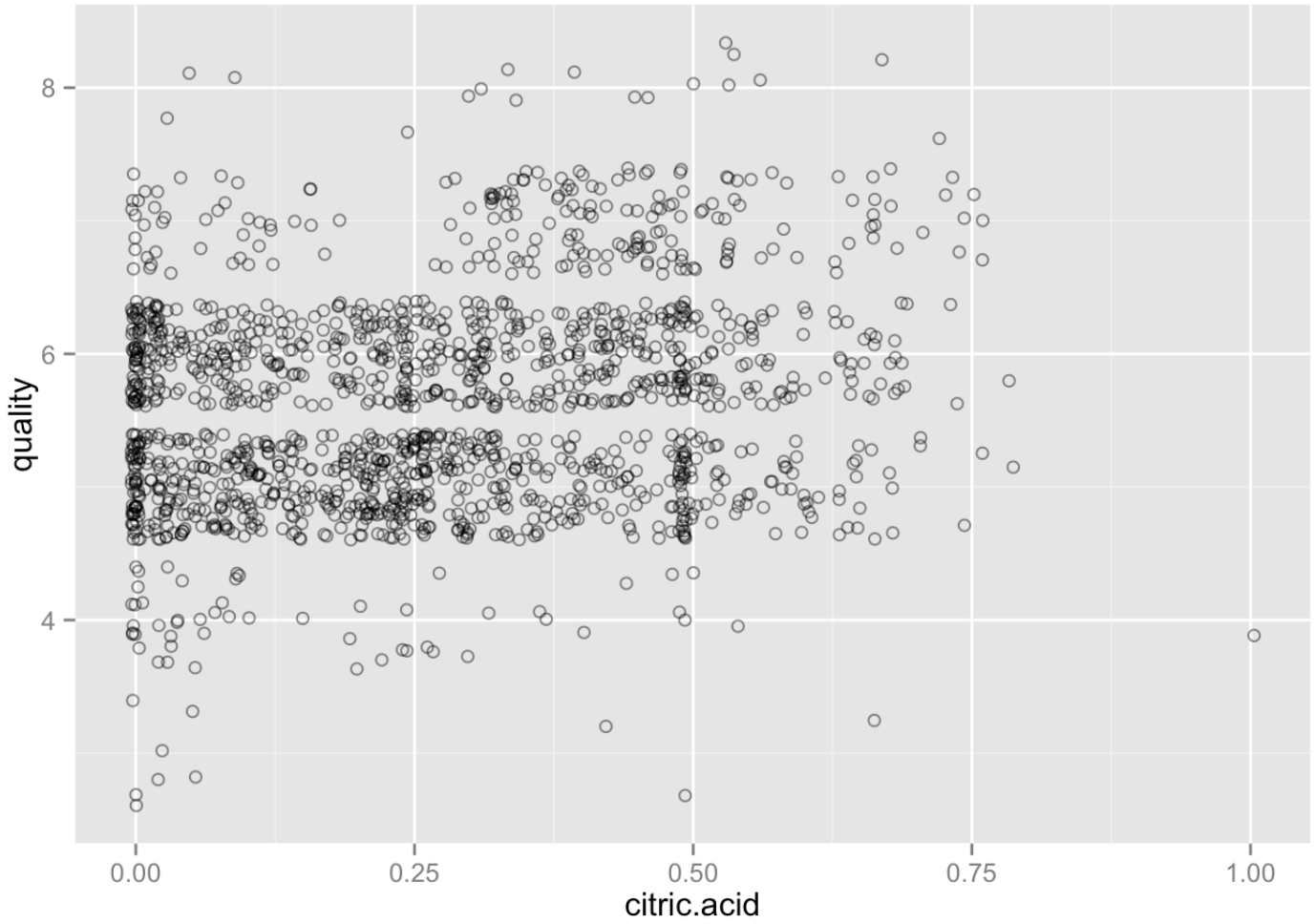


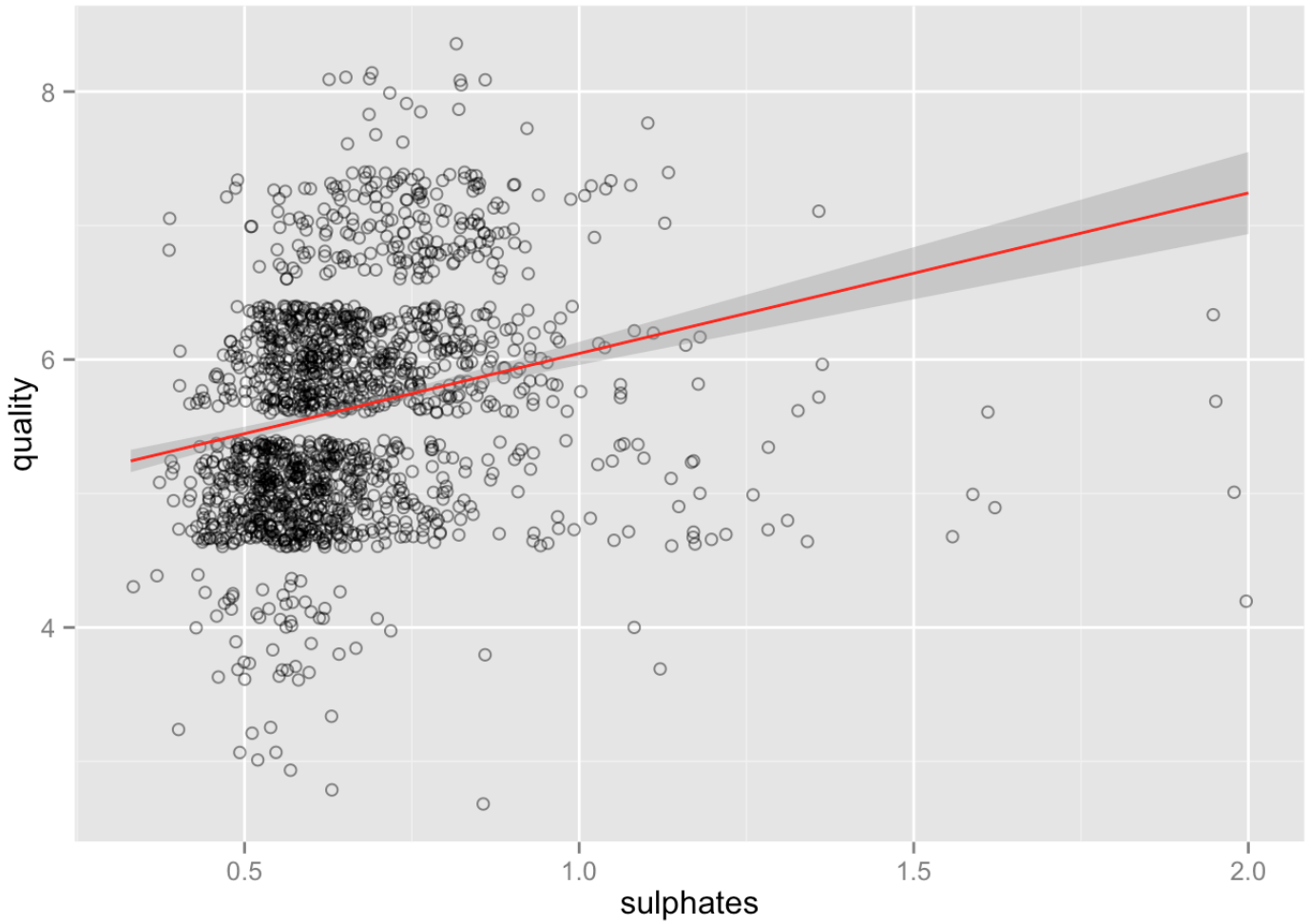
Summary of alcohol group shows that alcohol values are somewhat equally distributed. The boxplots of alcohol levels show that the highest quality (8) have the highest alcohol in general. Also, while as quality increases, alcohol levels seem to increase, there is a dip in quality level of 5. Quality level 5 seems to have a few outliers.

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

```
##       X fixed.acidity volatile.acidity citric.acid residual.sugar
## 653 653          15.9             0.36        0.65            7.5
##     chlorides free.sulfur.dioxide total.sulfur.dioxide density   pH
## 653     0.096                  22                   71  0.9976 2.98
##     sulphates alcohol quality alcohol.group
## 653      0.84    14.9       5   (11.1,14.9]
```

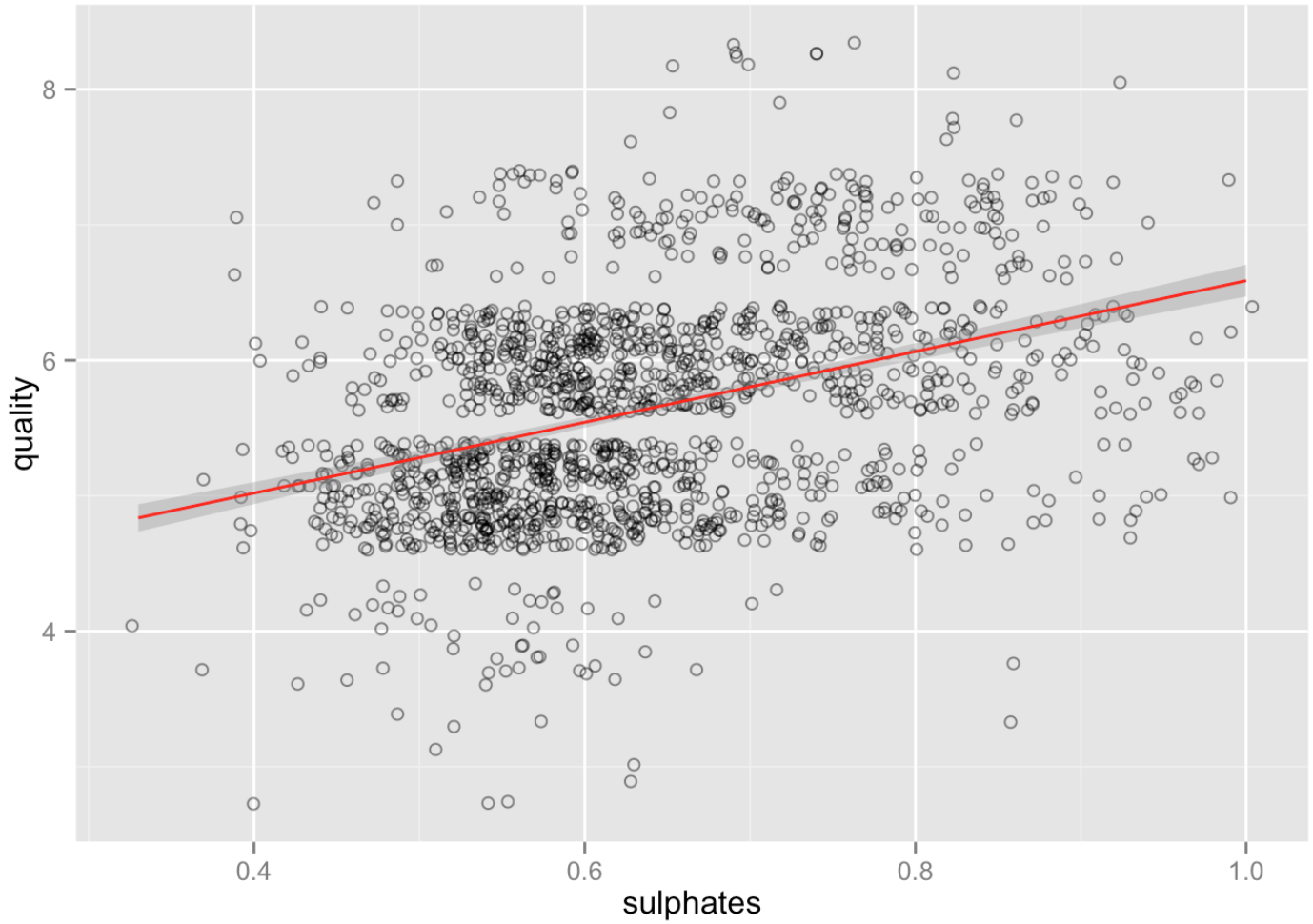
Looking at the largest outlier in quality == 5, it seems that this is the max. alcohol level (x=653)


quality
8
6
4
0.00
0.25
0.50
0.75
1.00
citric.acid

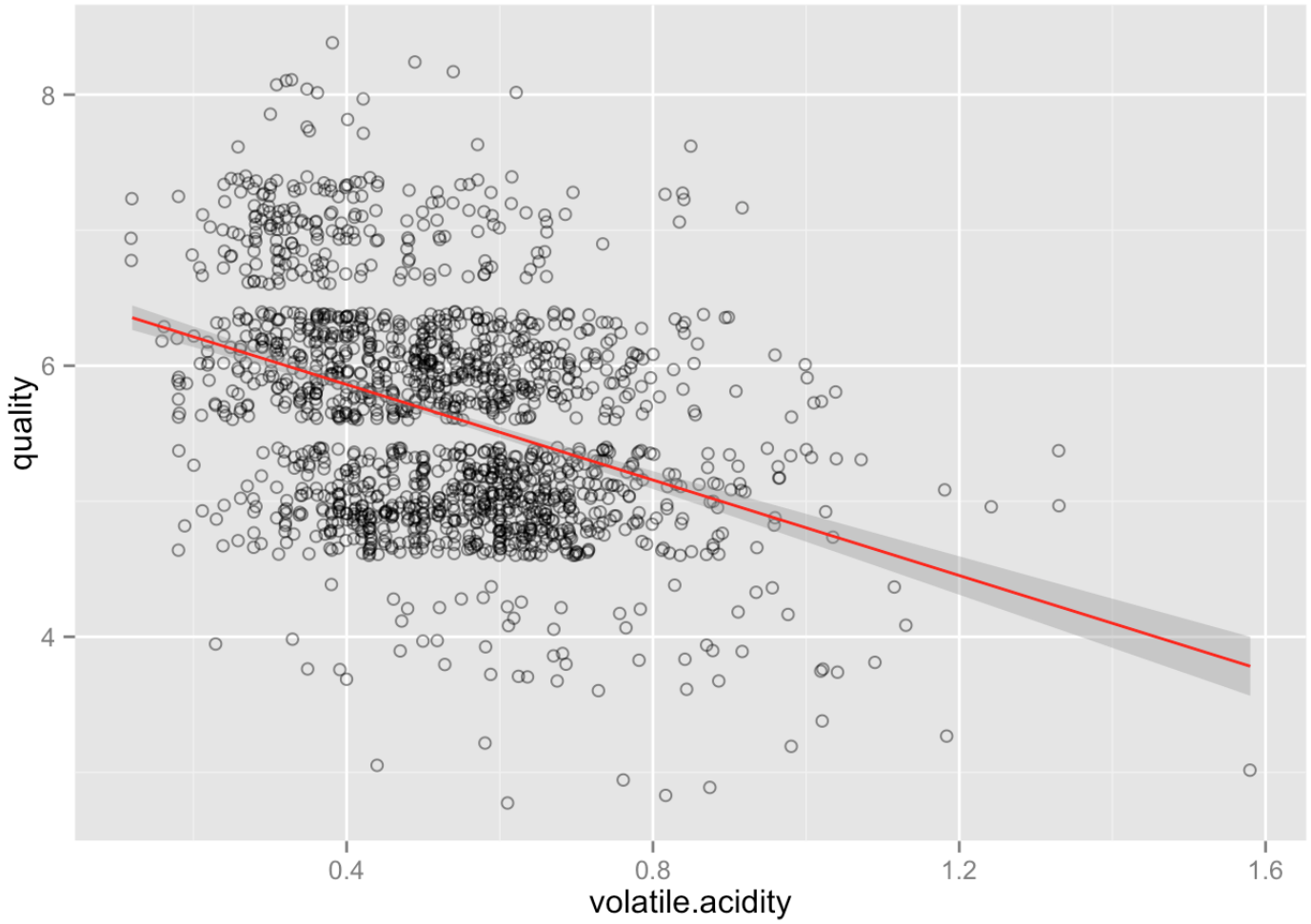

quality
8
6
4
0.5
1.0
1.5
2.0
sulphates

```
##
## Call:
## lm(formula = quality ~ sulphates, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2432 -0.5424  0.1102  0.4456  2.3977
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.84775    0.07842   61.82   <2e-16 ***
## sulphates    1.19771    0.11539   10.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7819 on 1597 degrees of freedom
## Multiple R-squared:  0.0632, Adjusted R-squared:  0.06261
## F-statistic: 107.7 on 1 and 1597 DF,  p-value: < 2.2e-16
```

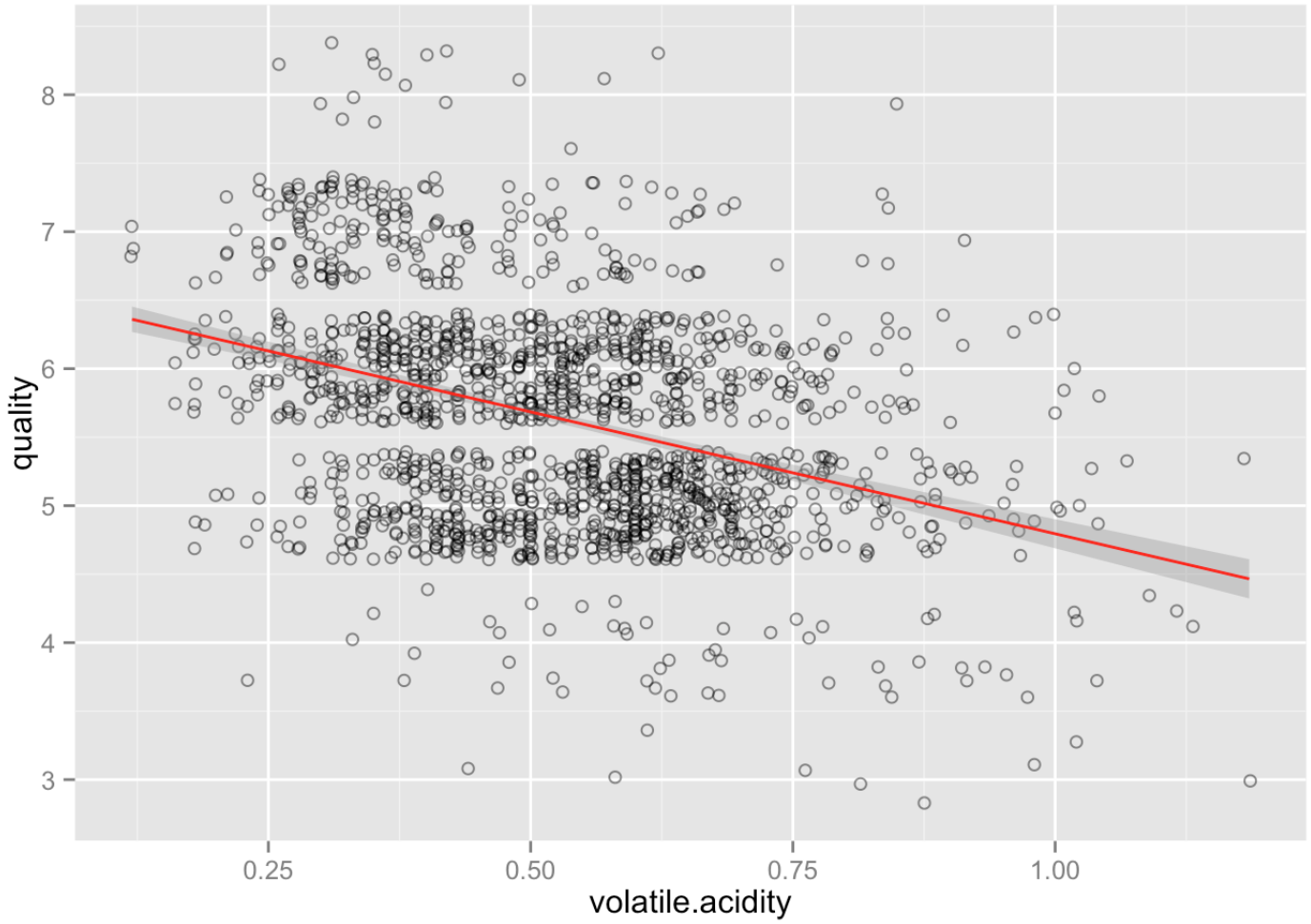
We can see that sulphates over 1.0 are more dispersed. So filtering data that has suphates of 1.0 or below shows a more clear positive correlation with quality. While total sulphates explain 6% of variation in quality, sulphates equal to or less than 1.0 explain 15% of variation in quality.




```
##
## Call:
## lm(formula = quality ~ sulphates, data = subset(data, sulphates <=
##     1))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2217 -0.4898 -0.1171  0.4840  2.3795
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.9739     0.1008   39.42   <2e-16 ***
## sulphates     2.6136     0.1555   16.80   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7402 on 1539 degrees of freedom
## Multiple R-squared:  0.1551, Adjusted R-squared:  0.1545
## F-statistic: 282.4 on 1 and 1539 DF,  p-value: < 2.2e-16
```

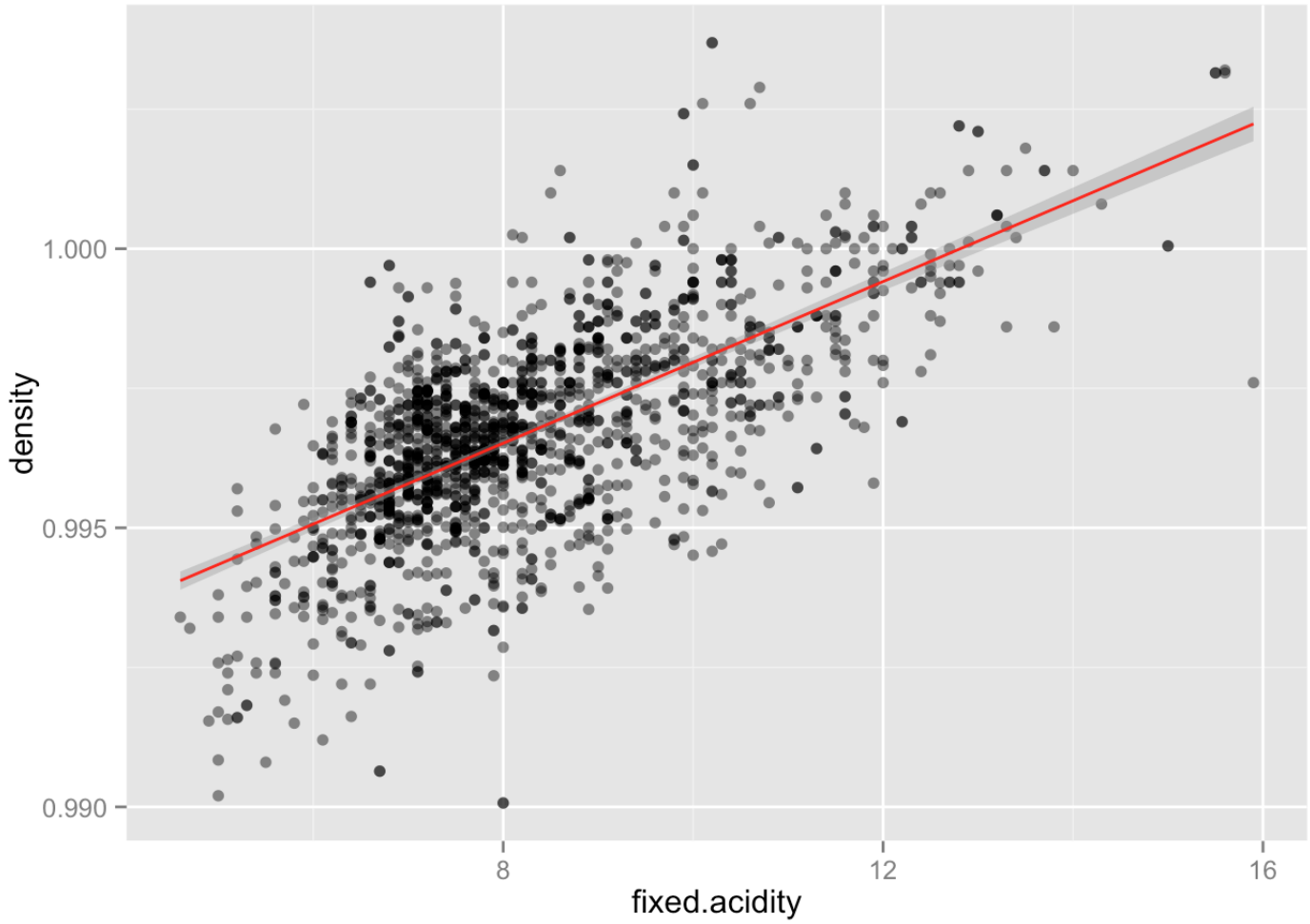


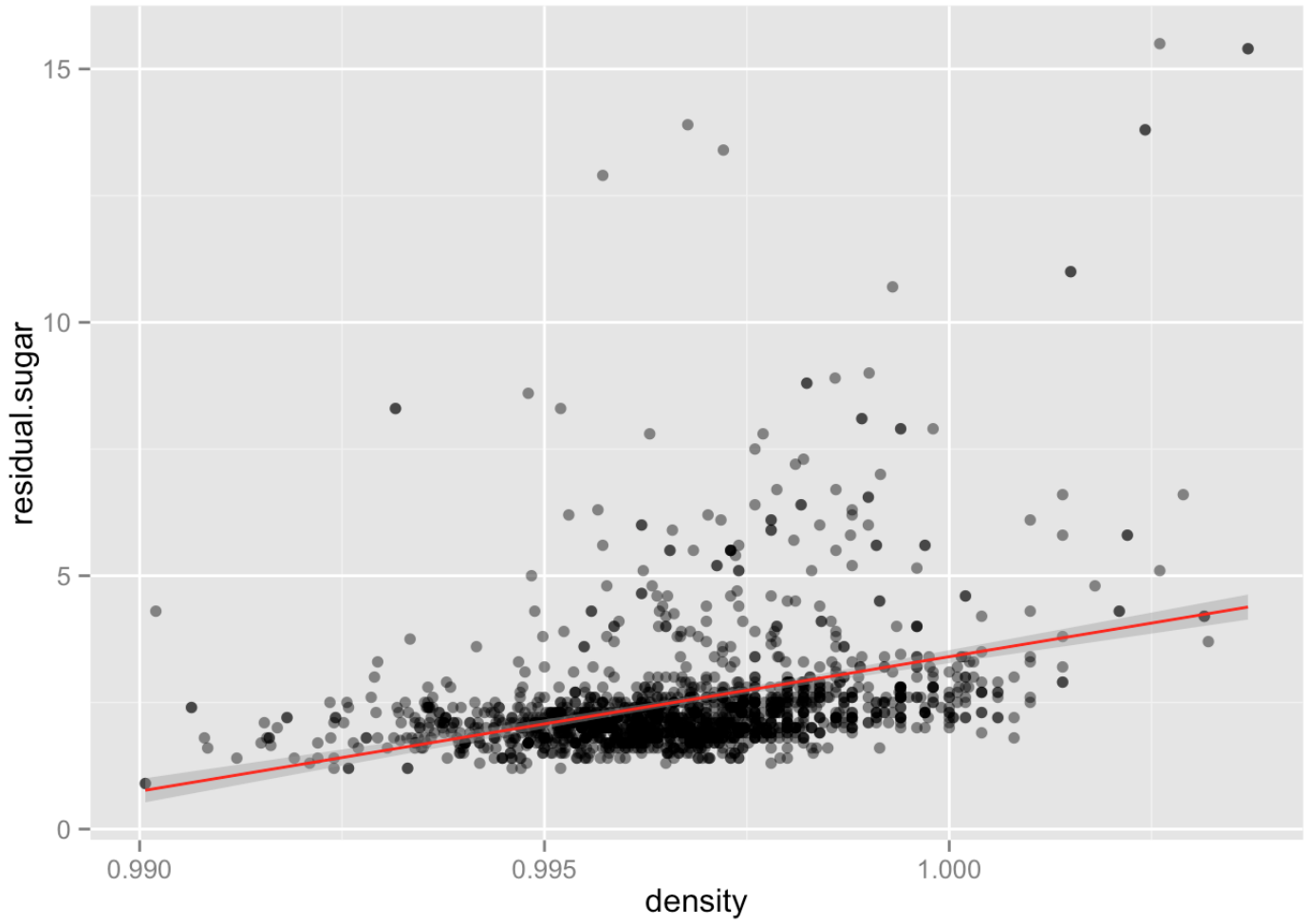
There seems to be an outlier with volatile acidity level of 1.6. Removing volatile acidity levels over 1.2 shows a stronger trend.




```
##
## Call:
## lm(formula = quality ~ volatile.acidity, data = subset(data,
##     volatile.acidity <= 1.2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.79138 -0.54227 -0.00846  0.47108  2.93816
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       6.57430    0.05924  110.98   <2e-16 ***
## volatile.acidity -1.77937    0.10697  -16.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7437 on 1593 degrees of freedom
## Multiple R-squared:  0.148,  Adjusted R-squared:  0.1474
## F-statistic: 276.7 on 1 and 1593 DF,  p-value: < 2.2e-16
```

I added jitter to all plots that include quality to avoid overplotting. Aside from one measure of acidity being correlated with another measure of acidity, I also noticed that measures of acidity and sugar are correlated with density.


density
1.000
0.995
0.990
8
12
16
fixed.acidity


residual.sugar
15
10
5
0
0.990
0.995
1.000
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?

Strongest correlation to quality is alcohol with a linear regression explaining 22% of variation in quality. Citric acid and sulphates are others that have stronger positive correlation with quality. I've also noticed with these variables that 82% of quality is 5 and 6. Jittering the plots to avoid overplotting revealed that while there is an overall upward sloping trend with these variables, concentrated values in quality in 5 and 6 is affecting their ability to explain variation in quality.

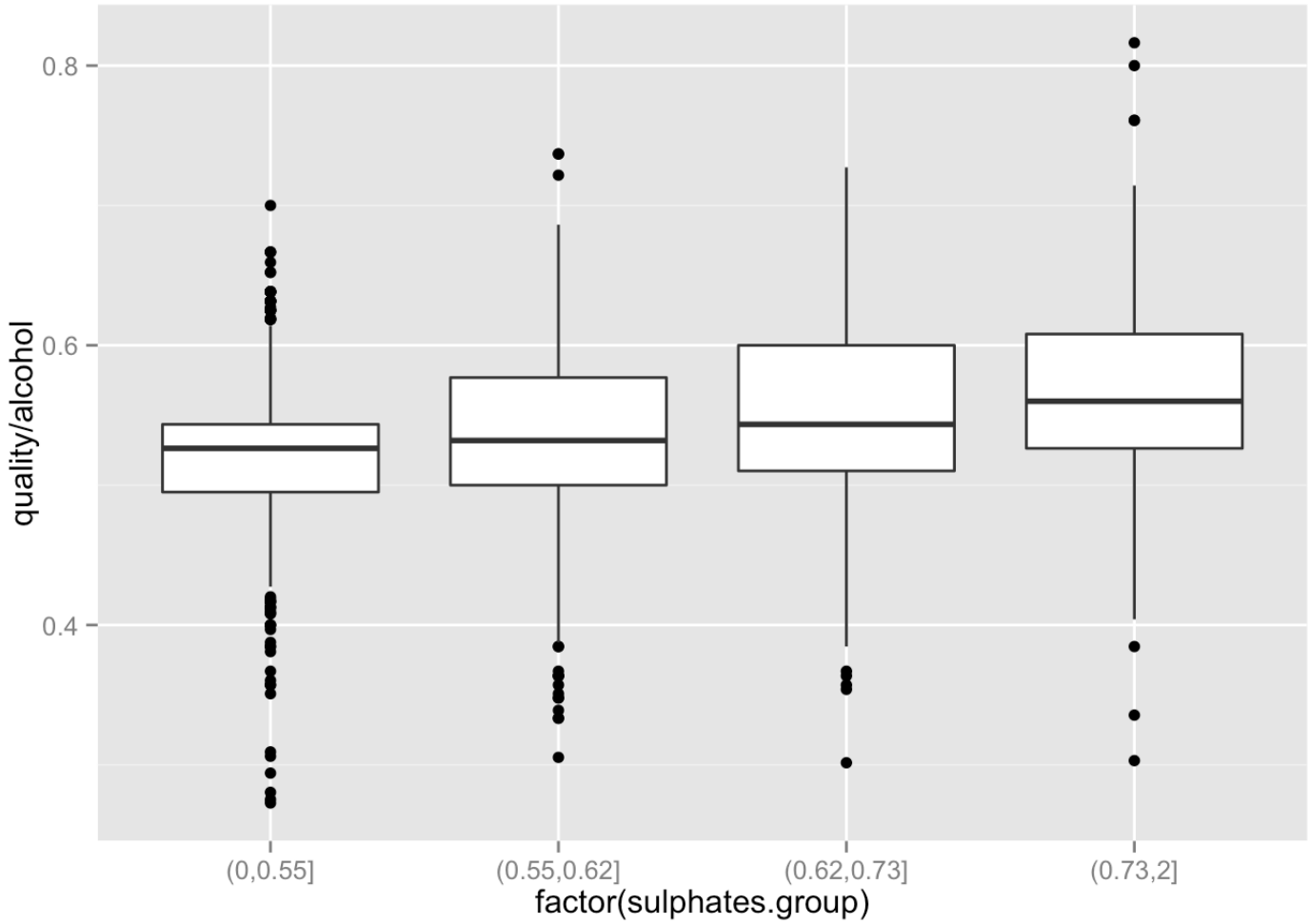
In some cases such as sulphates and volatile acidity, however, removing outliers helped increase R^2.

Other things to keep in mind include correlations between different chemical properties (different measures of acidity, density vs. alcohol and sugar). These will have to be controlled for in estimating quality.

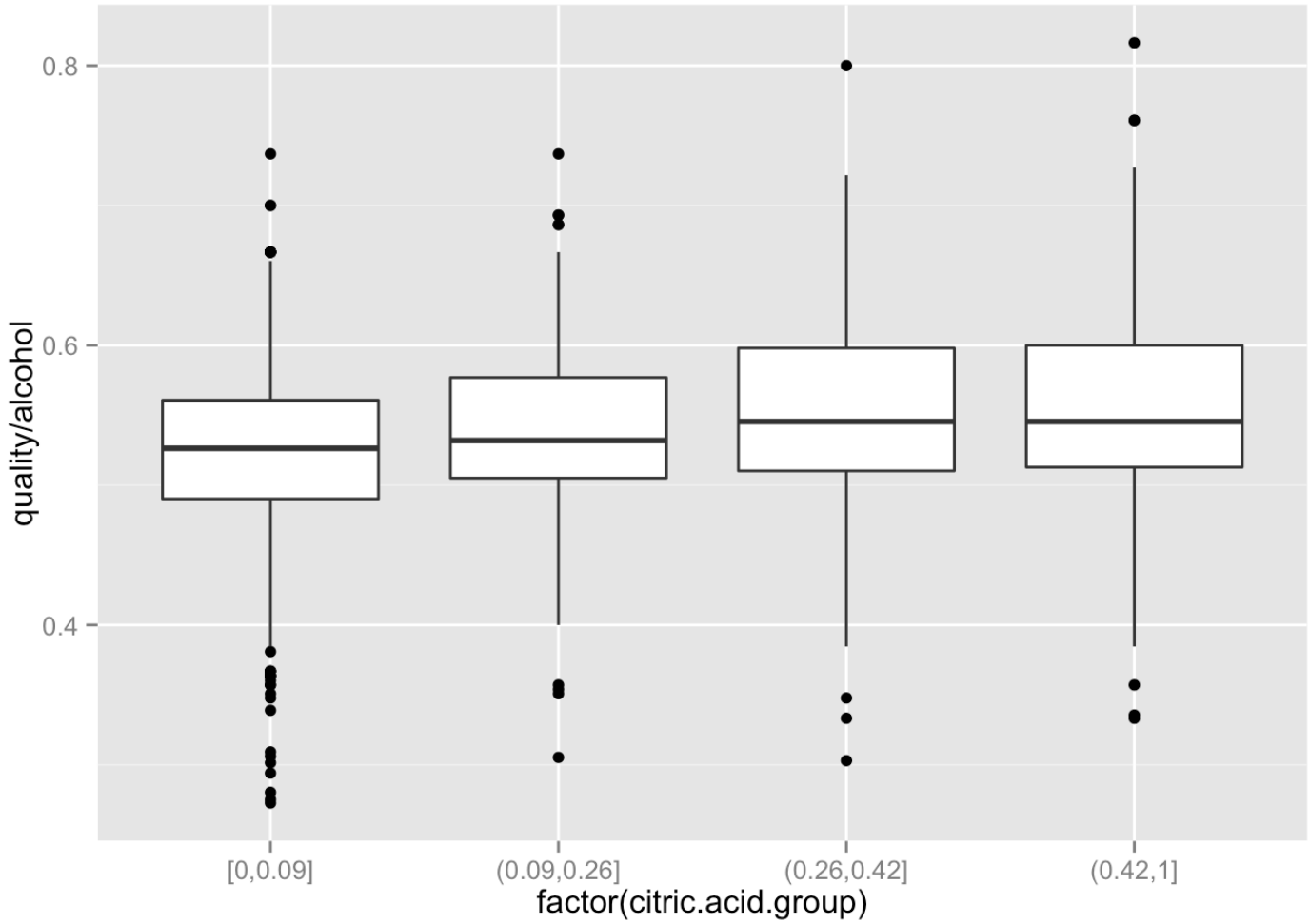
What was the strongest relationship you found?

Quality of wine is positively correlated with alcohol and sulphate. Volatile acidity is negatively correlated with quality. Citric acid is also more positively correlated than others, but less so than alcohol and sulphate.

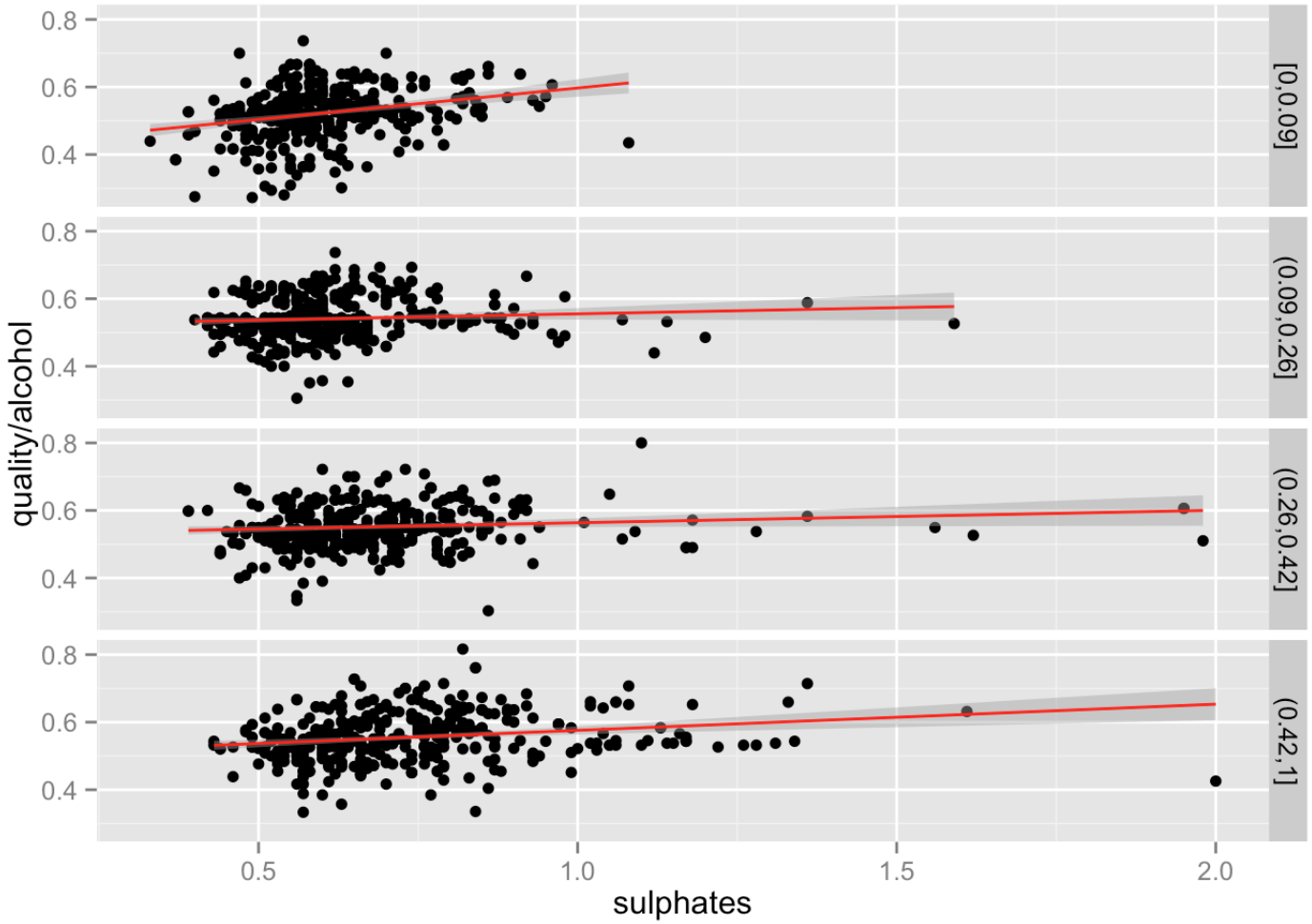
Multivariate Plots Section



The histogram shows that there tend to be more wines with higher quality The boxplot above shows that quality/alcohol is higher among higher sulphate levels. There are also more outliers in the lowest sulphate level.



This boxplot shows that higher citric acid groups have higher quality/alcohol levels but it doesn’t increase at the same rate. Citric acid levels higher than median value result in higher quality/alcohol than Citric acid levels lower than median value.



The scatterplots above show the relationship between sulphates and quality/alcohol in different quartile groups of citric acid. It shows that all groups of citric acid have a positive linear slope.

```
## [1] "[0,0.09]"    "(0.09,0.26]" "(0.26,0.42]" "(0.42,1]"
```

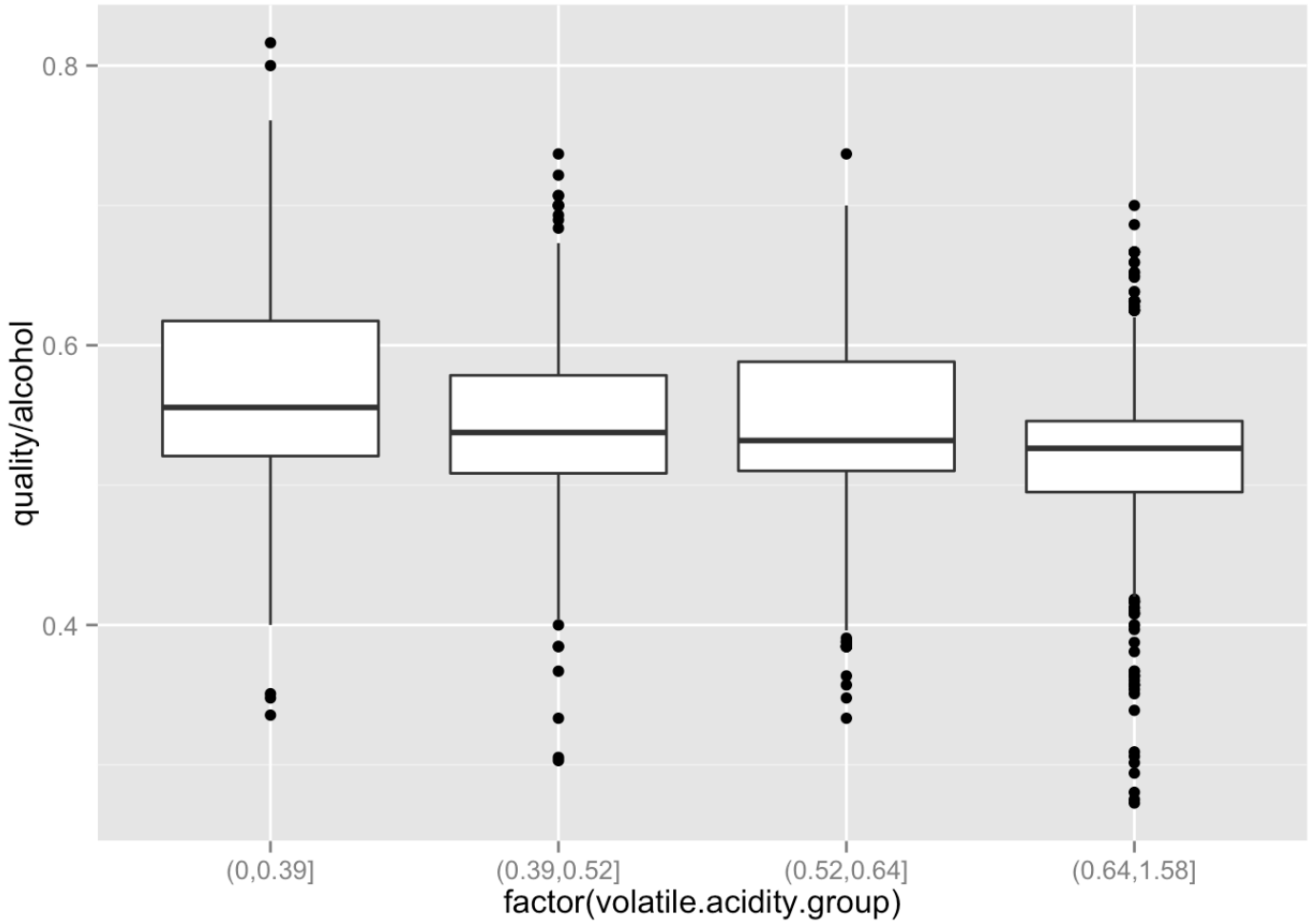
```
##
## Call:
## lm(formula = quality/alcohol ~ sulphates, data = subset(data,
##     data$citric.acid.group == "[0,0.09]"))
##
## Coefficients:
## (Intercept)    sulphates
##      0.4106       0.1864
```

```
##
## Call:
## lm(formula = quality/alcohol ~ sulphates, data = subset(data,
##     data$citric.acid.group == "(0.09,0.26]"))
##
## Coefficients:
## (Intercept)    sulphates
##     0.51836      0.03681
```

```
##
## Call:
## lm(formula = quality/alcohol ~ sulphates, data = subset(data,
##     data$citric.acid.group == "(0.26,0.42]"))
##
## Coefficients:
## (Intercept)    sulphates
##     0.52662      0.03693
```

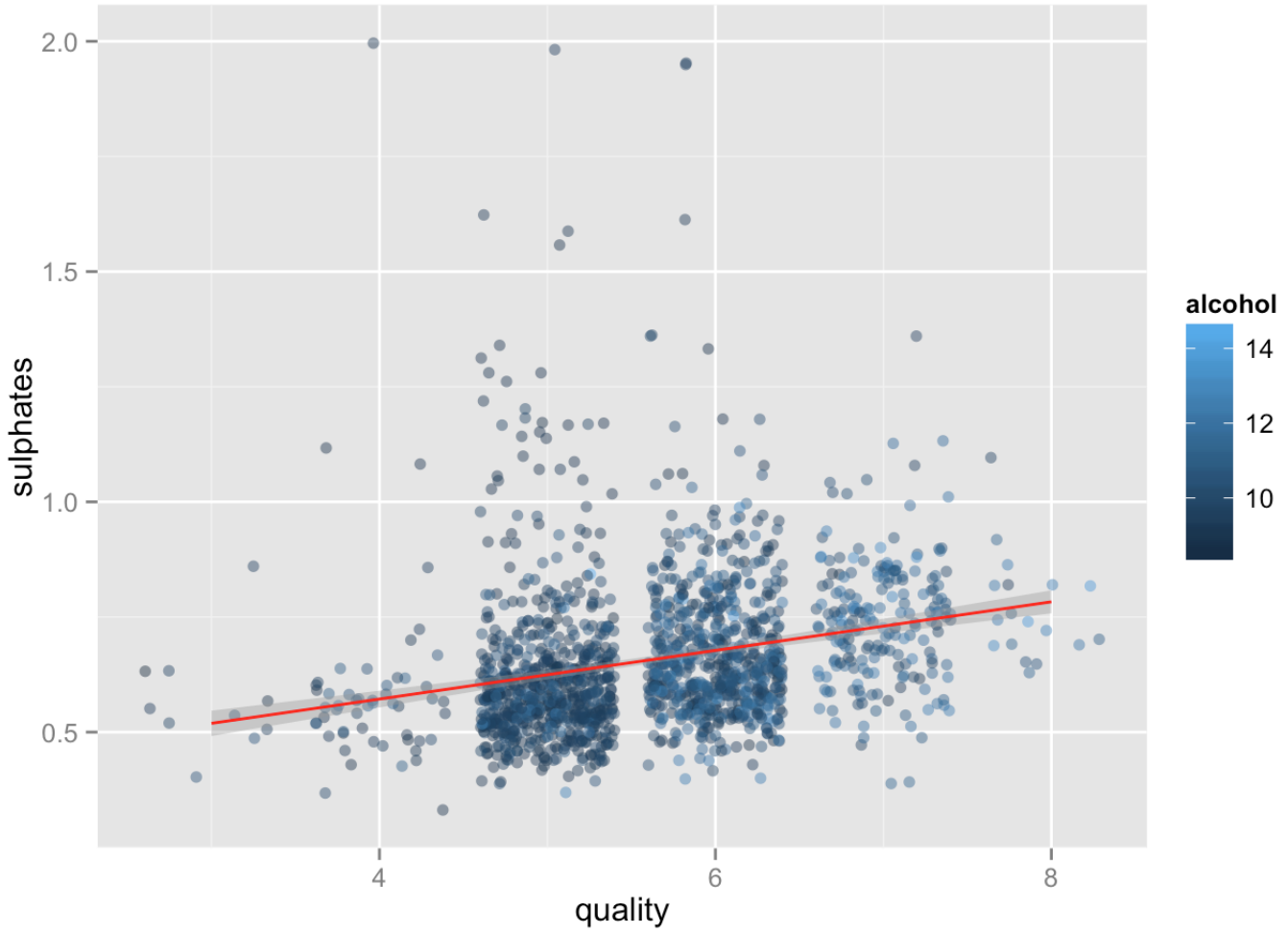
```
##
## Call:
## lm(formula = quality/alcohol ~ sulphates, data = subset(data,
##     data$citric.acid.group == "(0.42,1]"))
##
## Coefficients:
## (Intercept)    sulphates
##     0.49813      0.07759
```

Running linear regressions on each quartile group of citric acid shows that the 1st quartile of citric acid has the highest coefficient of sulphates in estimating quality/alcohol.



The boxplot shows that while median quality/alcohol values go down as it moves up on the volatile acidity scale, there are a lot of outliers in the highest volatile acidity group. Also, the maximum value of the 3rd quartile group is larger than the maximum value of the 2nd quartile group.

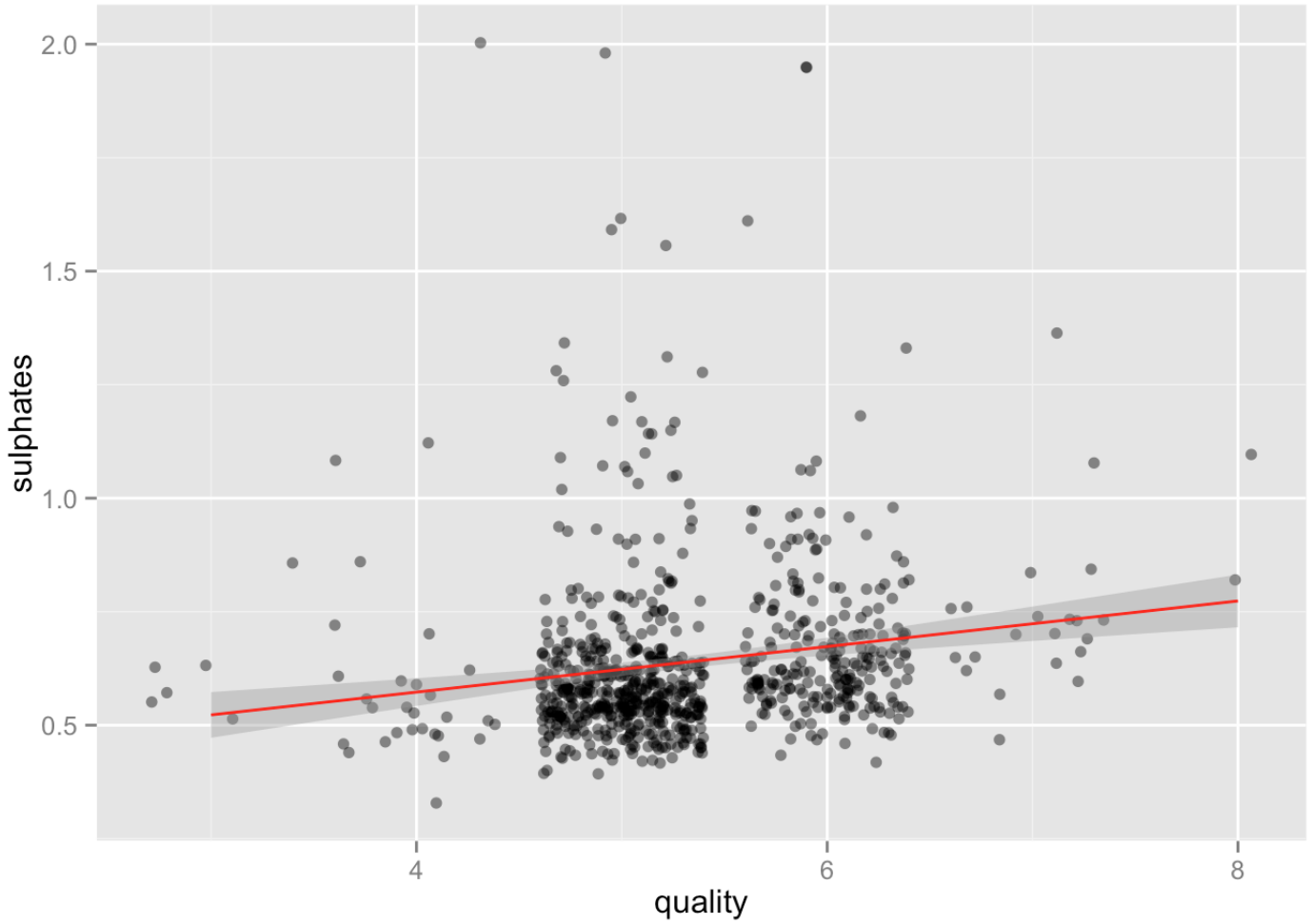
quality and alcohol by each sulphates group



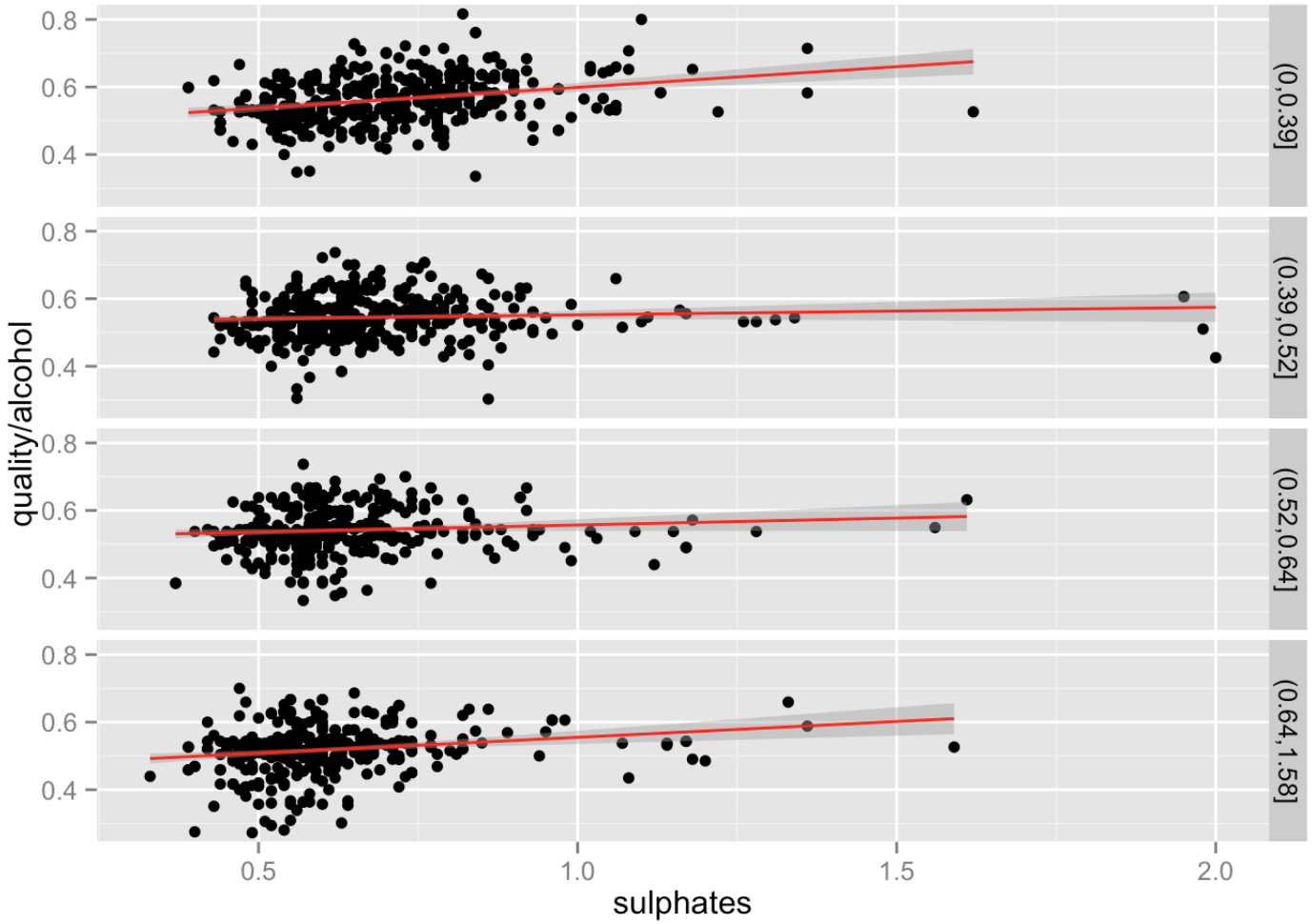
This scatterplot has geom_jitter added to avoid overplotting. It shows that the overall trend shown by the linear smoothing line indicates a positive relationship between quality and sulphates. Also, the darker (lower) alcohol levels seem more concentrated in quality level 5.

To examine this further, the histogram below shows the relationship between quality and sulphates where alcohol is less than the median of alcohol (10.20). It shows that quality = 5 has the highest concentration of sulphates at this alcohol level.

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



quality/alcohol and sulphates by each volatile.acidity group



The facet grid above shows the relationship between quality/alcohol with sulphates at different groups of volatile.acidity. It shows that at all levels of volatile.acidity, there is a positive linear trend between sulphates and quality/alcohol. I also noticed the outliers in 2nd quartile of the volatile.acidity group. Removing these outliers show that the 1st quartile of the volatile.acidity group has the highest linear slope (0.1225) between sulphates and quality/alcohol.

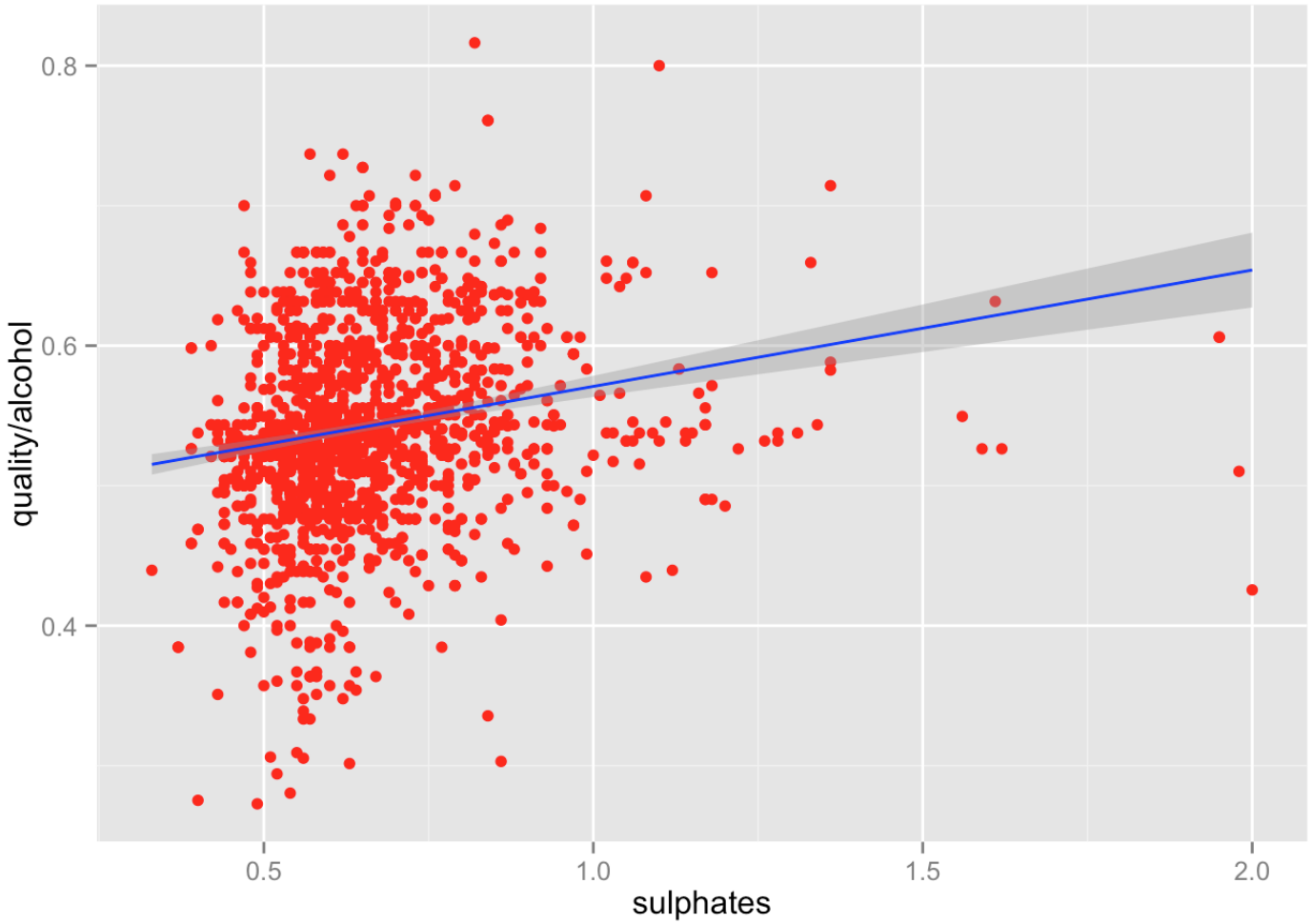
```
##
## Call:
## lm(formula = quality/alcohol ~ sulphates, data = subset(data,
##     data$volatile.acidity.group == "(0,0.39]"))
##
## Coefficients:
## (Intercept)    sulphates
##      0.4764       0.1225
```

```
##
## Call:
## lm(formula = quality/alcohol ~ sulphates, data = subset(VA2Q,
##     VA2Q$sulphates < 1.5))
##
## Coefficients:
## (Intercept)    sulphates
##     0.51403      0.04638
```

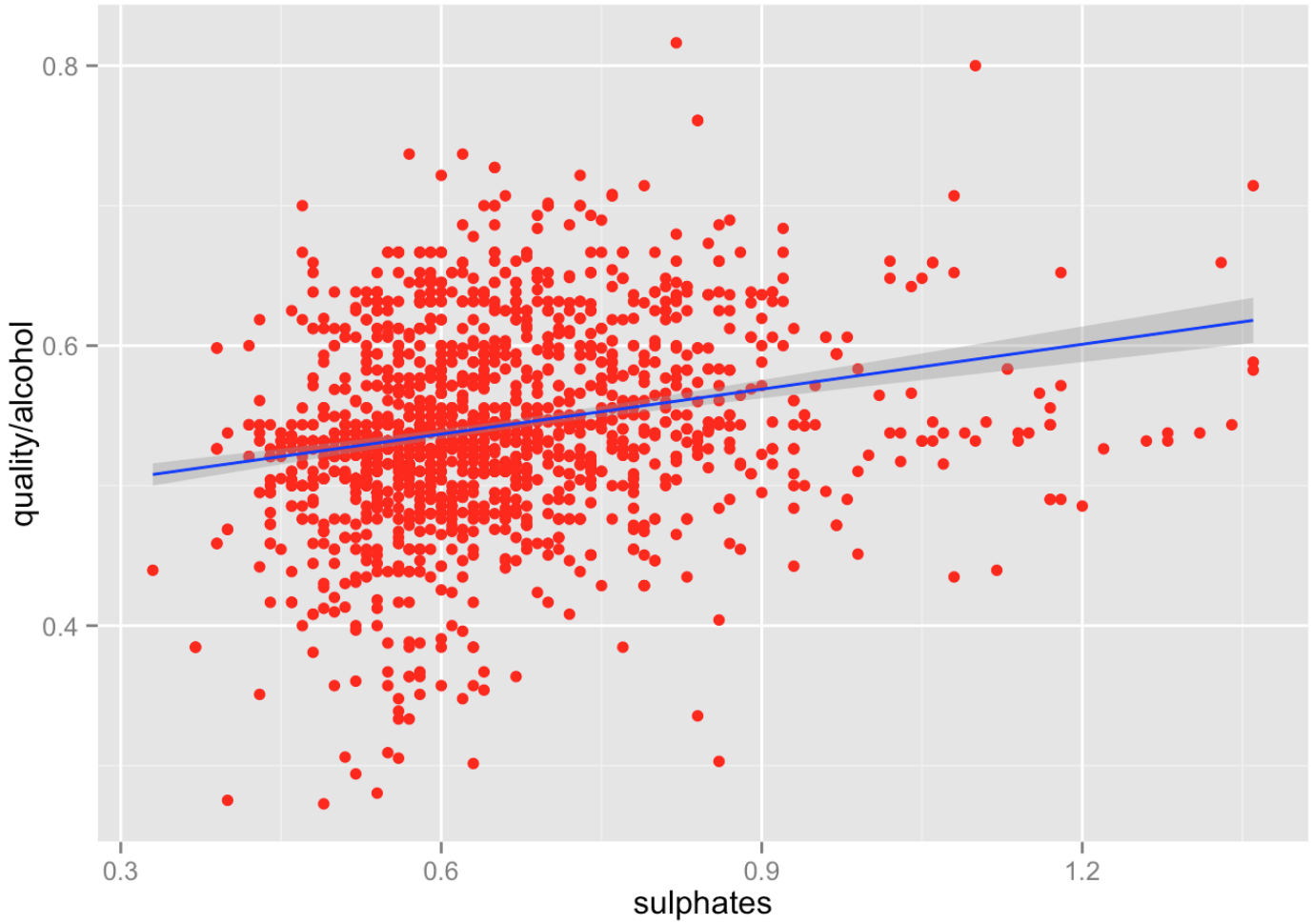
```
##
## Call:
## lm(formula = quality/alcohol ~ sulphates, data = subset(data,
##     data$volatile.acidity.group == "(0.52,0.64]"))
##
## Coefficients:
## (Intercept)    sulphates
##     0.51547      0.04134
```

```
##
## Call:
## lm(formula = quality/alcohol ~ sulphates, data = subset(data,
##     data$volatile.acidity.group == "(0.64,1.58]"))
##
## Coefficients:
## (Intercept)    sulphates
##     0.46137      0.09373
```

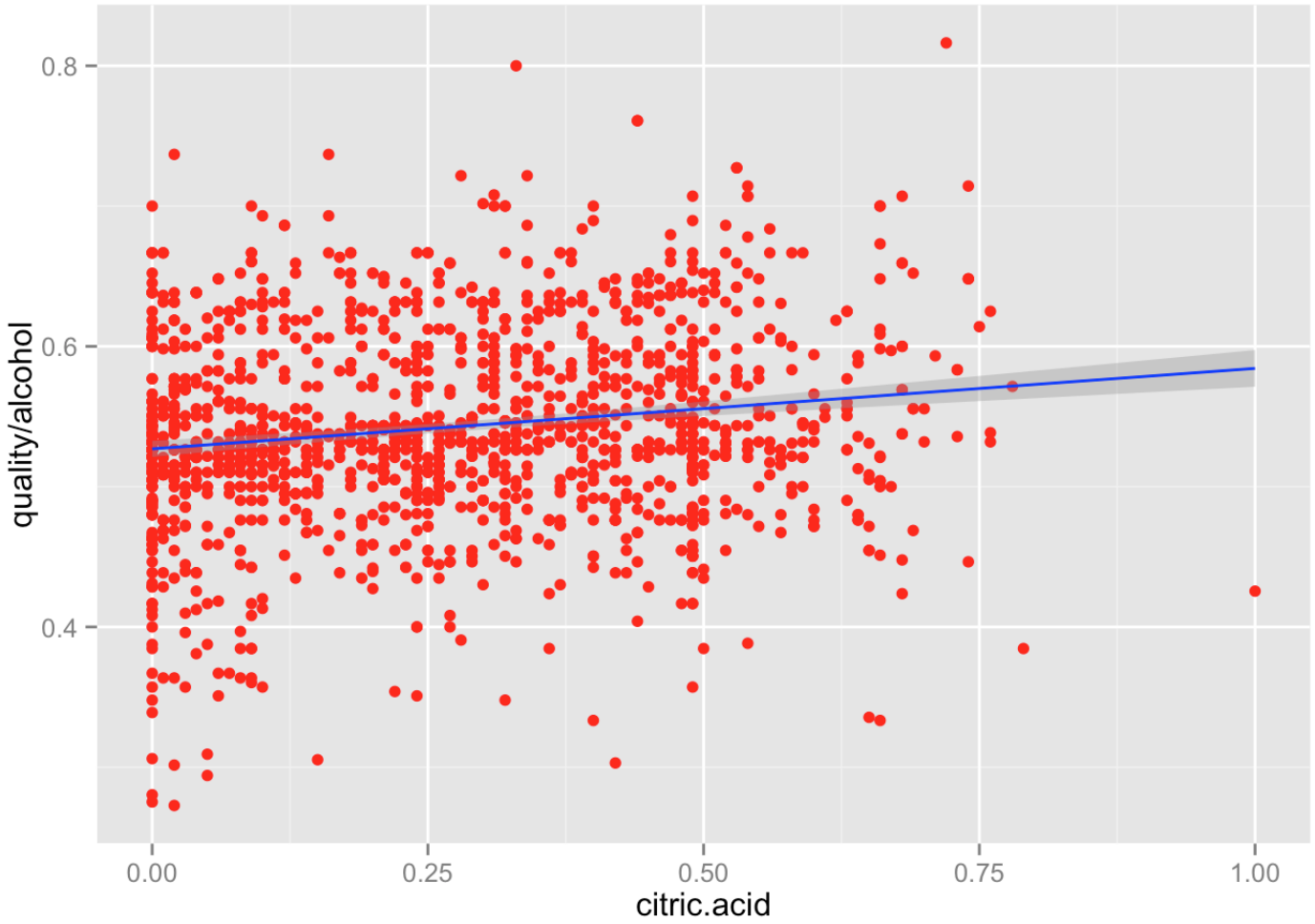
The below scatterplot shows the overall scatterplot of sulphates and quality/alcohol.

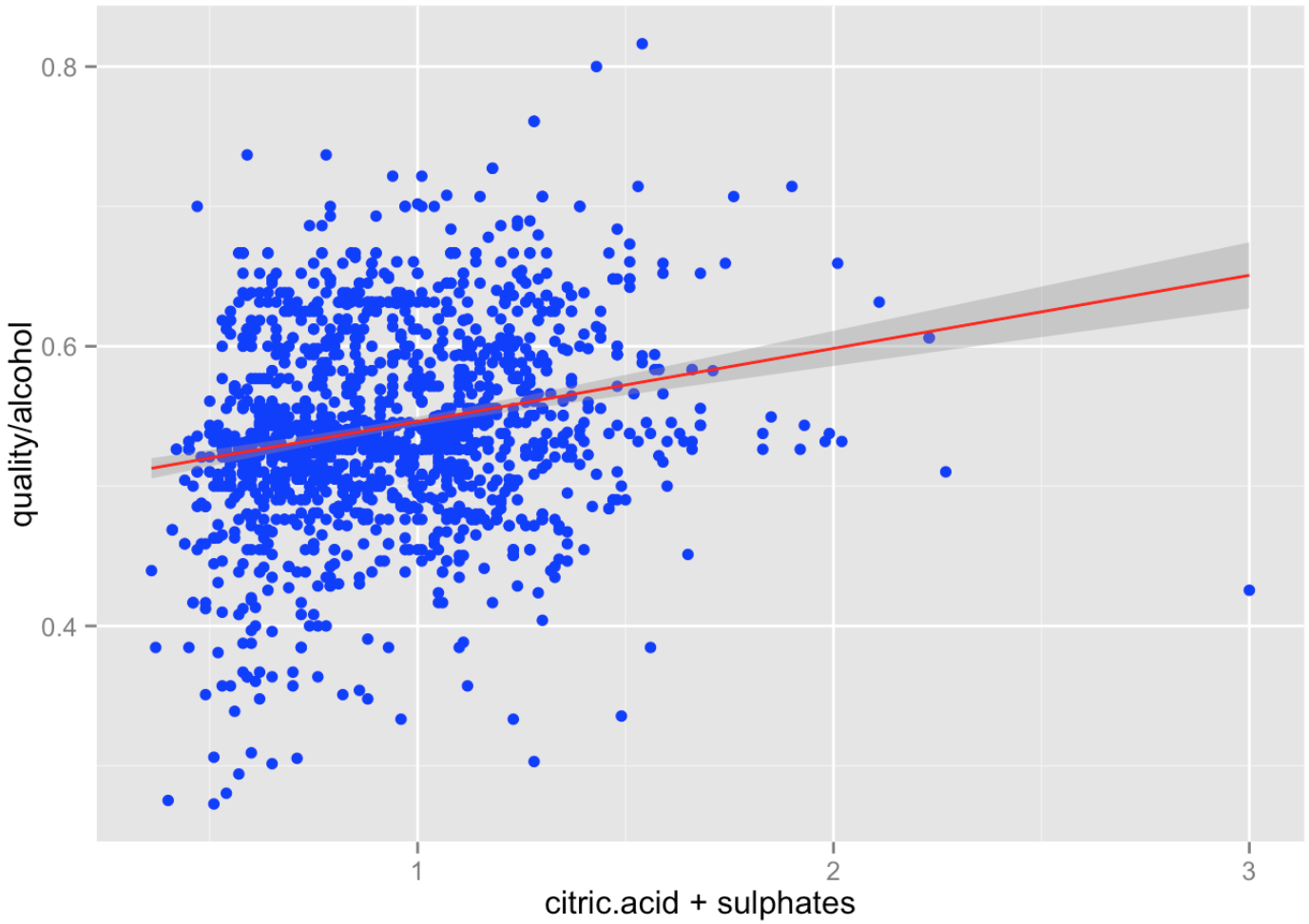


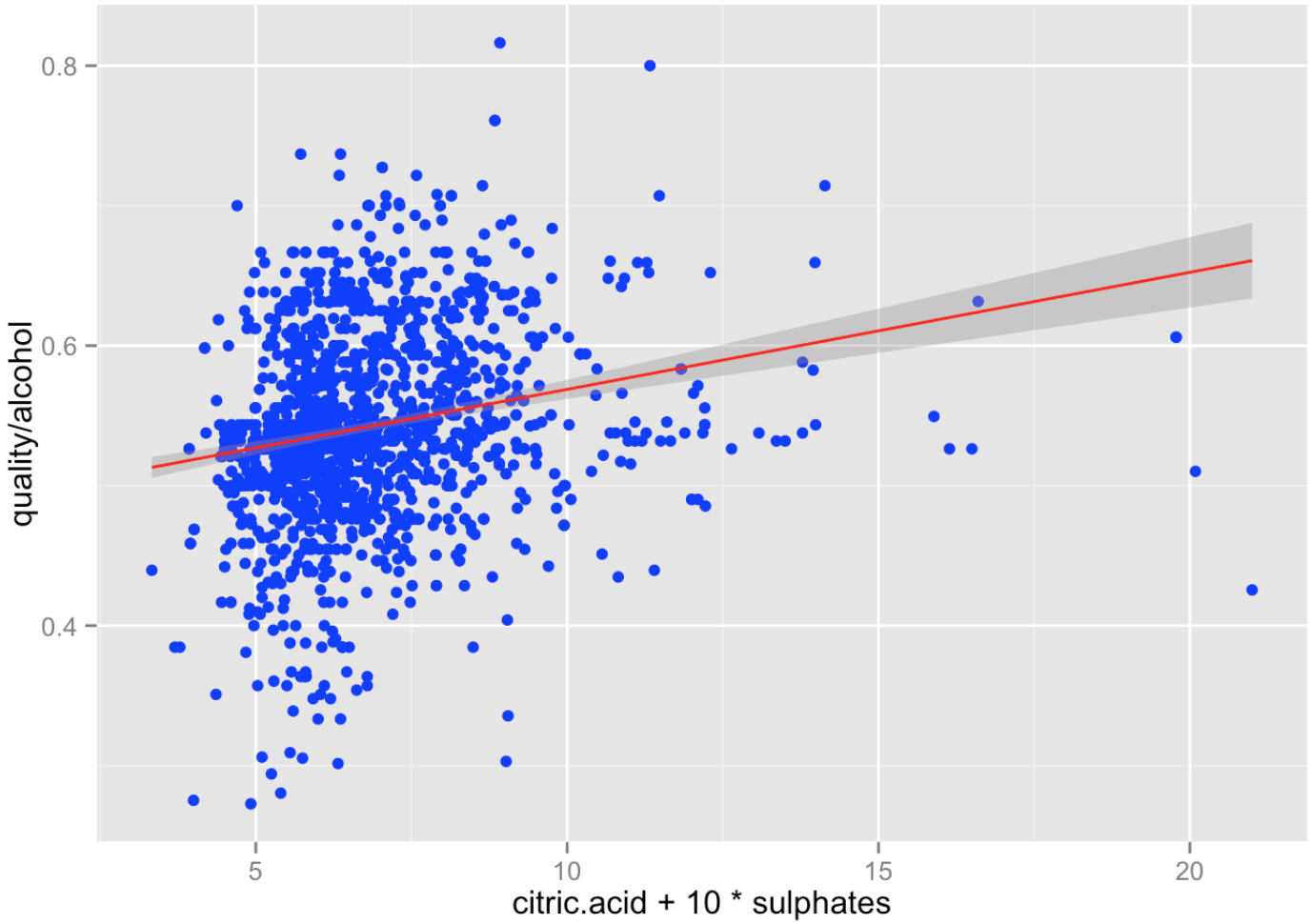
Thhe histogram shows the smoothing line after removing the outliers that are 1.5 and higher in sulphates.

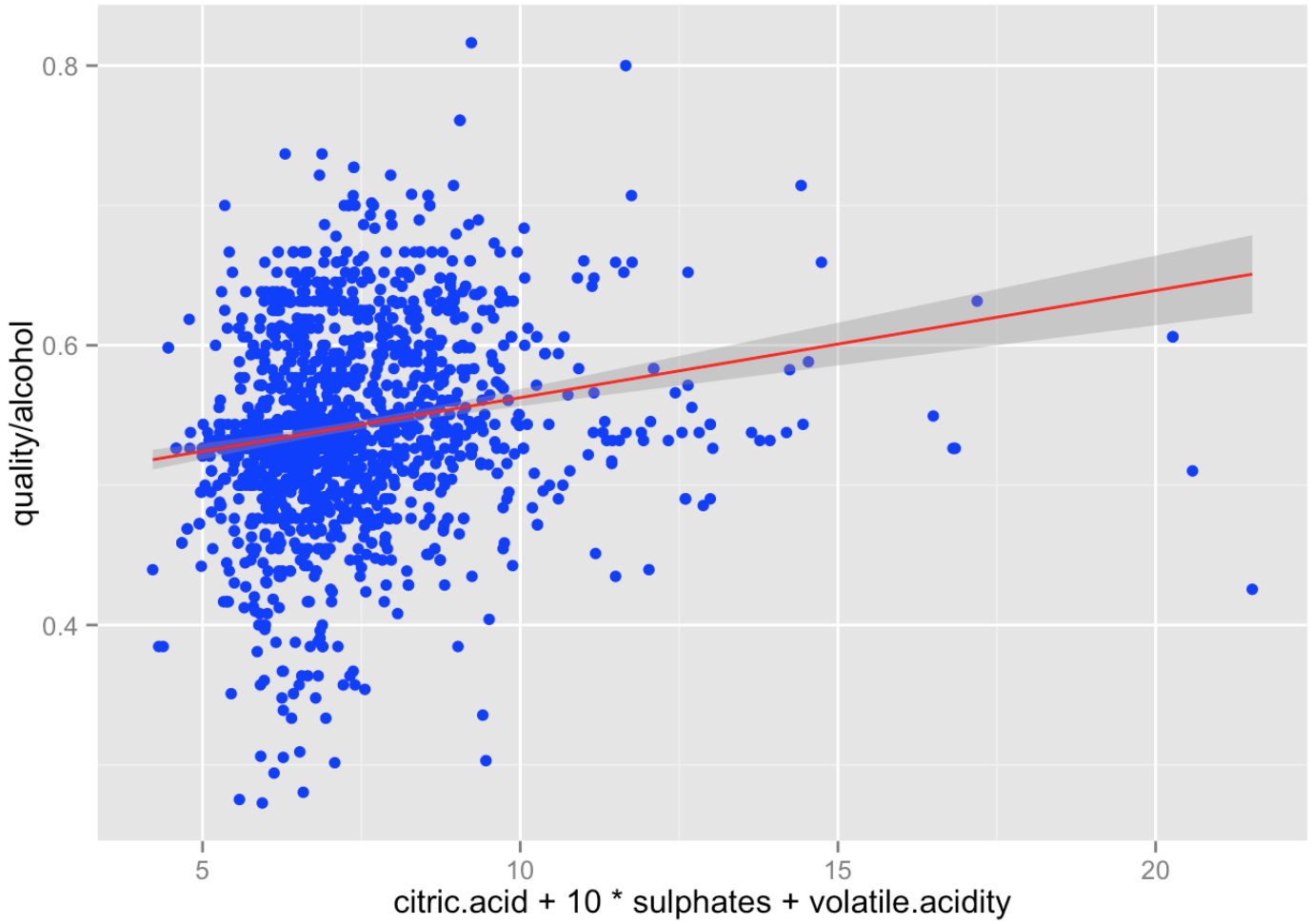


The following scatterplots tweak and create new variables to better model the relationship with quality/alcohol.


0.8
0.6
0.4
quality/alcohol
0.00
0.25
0.50
0.75
1.00
citric.acid


0.8
0.6
0.4
quality/alcohol
1
2
3
citric.acid + sulphates


0.8
0.6
0.4
quality/alcohol
5
10
15
20
citric.acid + 10 * sulphates



Multiplying sulphates by 10 helps to eliminate the variance between values. Creating linear models based on this show that using these 4 variables (alcohol, sulphates, citric acid, volatile acidity) improve the R^2 by over 12%.

```
##
## Call:
## lm(formula = quality ~ alcohol, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8442 -0.4112 -0.1690  0.5166  2.5888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.87497    0.17471   10.73   <2e-16 ***
## alcohol      0.36084    0.01668   21.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7104 on 1597 degrees of freedom
## Multiple R-squared:  0.2267, Adjusted R-squared:  0.2263
## F-statistic: 468.3 on 1 and 1597 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = quality ~ alcohol + sulphates * citric.acid + volatile.acidity,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.71232 -0.38772 -0.06075  0.46631  2.26531
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.33146    0.22553  10.337  < 2e-16 ***
## alcohol                0.30628    0.01579  19.393  < 2e-16 ***
## sulphates              1.18470    0.19066   6.214  6.6e-10 ***
## citric.acid            0.85977    0.32508   2.645  0.00825 **
## volatile.acidity      -1.21410    0.11361 -10.687  < 2e-16 ***
## sulphates:citric.acid -1.37600    0.45161  -3.047  0.00235 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6571 on 1593 degrees of freedom
## Multiple R-squared:   0.34,  Adjusted R-squared:  0.3379
## F-statistic: 164.1 on 5 and 1593 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = quality ~ alcohol * I(sulphates * 10) + citric.acid +
##     volatile.acidity, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.76963 -0.38271 -0.08959  0.47365  2.24285
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 6.46949    0.71146   9.093  < 2e-16 ***
## alcohol                    -0.07310    0.07004  -1.044    0.297
## I(sulphates * 10)          -0.50221    0.10264  -4.893 1.10e-06 ***
## citric.acid                -0.07938    0.10284  -0.772    0.440
## volatile.acidity           -1.22336    0.11186 -10.937  < 2e-16 ***
## alcohol:I(sulphates * 10)  0.05673    0.01013   5.598 2.55e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6526 on 1593 degrees of freedom
## Multiple R-squared:  0.3489, Adjusted R-squared:  0.3469
## F-statistic: 170.8 on 5 and 1593 DF,  p-value: < 2.2e-16
```

The last linear model explain 34.9% of variance in the quality of wines.

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?

Because alcohol was the best indicator in explaining variance in quality of wines, I used other variables to explain quality/alcohol as a dependent variable. Relationships between other variables and quality were largely consistent with their relationships with quality/alcohol.

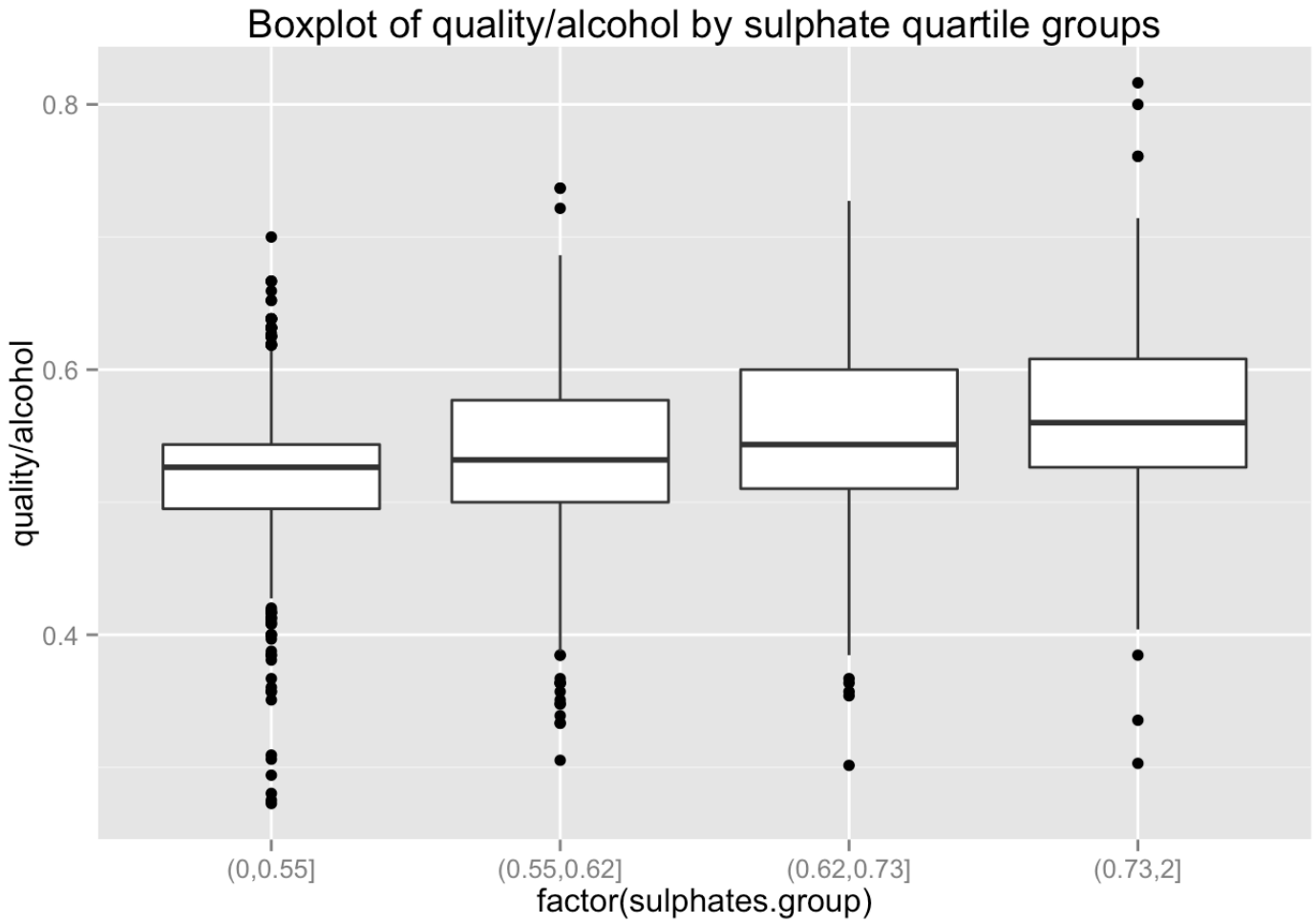
Were there any interesting or surprising interactions between features?

In reducing the wide variance between variables and quality/alcohol, multiplying sulphates by 10 had the best results. On other variables, multiplying numbers only widened the variance.

Final Plots and Summary

Plot one

Boxplots of quality/alcohol and sulphate

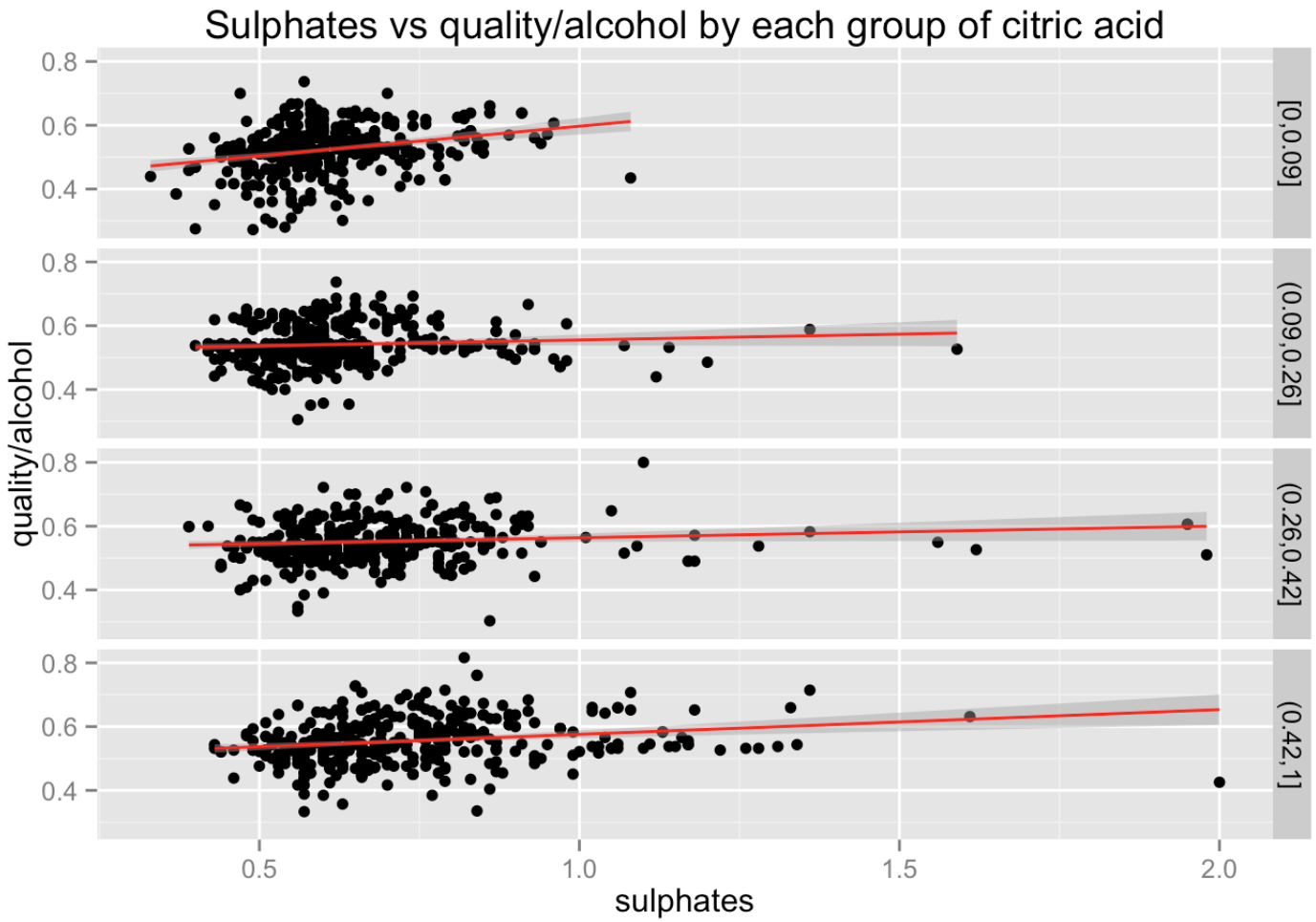


Using quality/alcohol metric helps avoid overplotting and the difficulty of visualizing quality values that are integers. The boxplot shows the trend clearly that higher sulphate groups have higher min, median, max quality/alcohol values. The four sulphate groups represent 1st, 2nd, 3rd and 4th quartile.

This plot supports other findings of the positive relationship between sulphates and quality. Using this finding with others, I use them later to add sulphates in the linear regression model to predict the quality of wines.

Plot two

quality and alcohol by each sulphates group

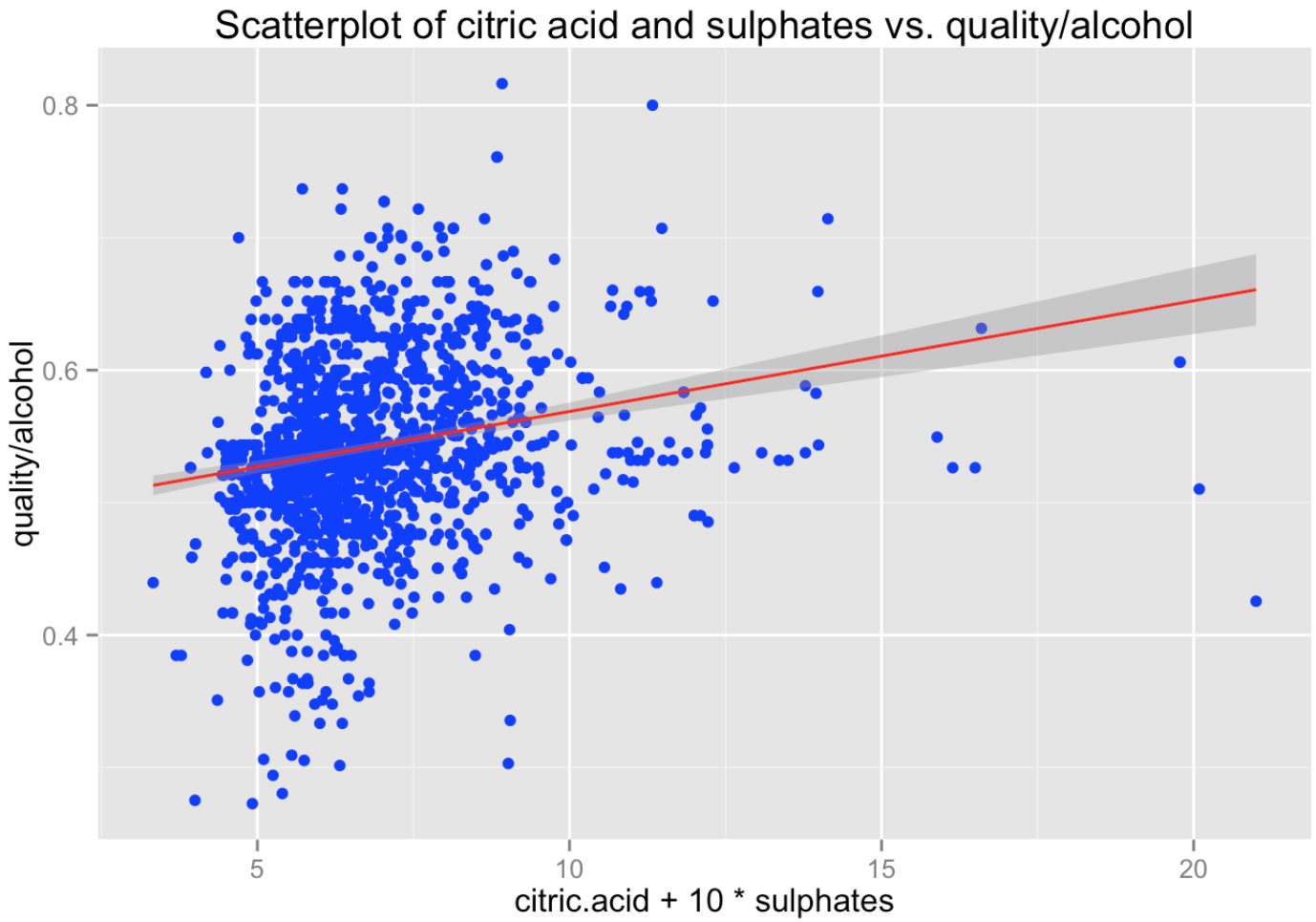


The above scatterplot shows sulphates in the x-axis and quality/alcohol in the y-axis. Each facet represents 1st through 4th quartiles of citric acid. Running linear regressions previously showed that 1st quartile of citric acid has the highest slope (0.1864).

The facet grid helps observe and compare scatterplots and this one supports previous findings that sulphates and quality/alcohol are positively correlated in all levels of citric acid, which is another more strongly correlated variable with quality. Using this finding with others, I add citric acid in my final linear regression model to predict quality.

Plot three

quality/alcohol vs. citric.acid + 10*sulphates



Using quality/alcohol as the dependent variable, this scatterplot reduced the variance with the interaction between citric acid and sulphates. This shows a clear trend that as the interaction between citric acid and sulphates goes up, quality/alcohol goes up.

This scatterplot was derived by creating different variables using citric acid and sulphates to explain quality/alcohol. I used this in my final linear regression model.

Reflection

Possible because the wine quality ratings are based on individuals' subjective tastes, 82% of quality of wines were either 5 or 6. Because of this and the fact that quality variable were integers between 3 and 8, there were difficulties of visualizing and avoiding overplotting. As a way to solve this problem, I used quality/alcohol as the dependent variable instead. To investigate reasons why there were so many 5 and 6 quality ratings, it would be helpful to get data on other types of wines.

There were also correlations between variables rising from chemical properties (between different acidity meatures and between density and alcohol). I tried to account for these to avoid correlations between independent variables.