Red Wine Quality Exploration By Mustafa Bilal Tahir

Univariate Plots Section

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

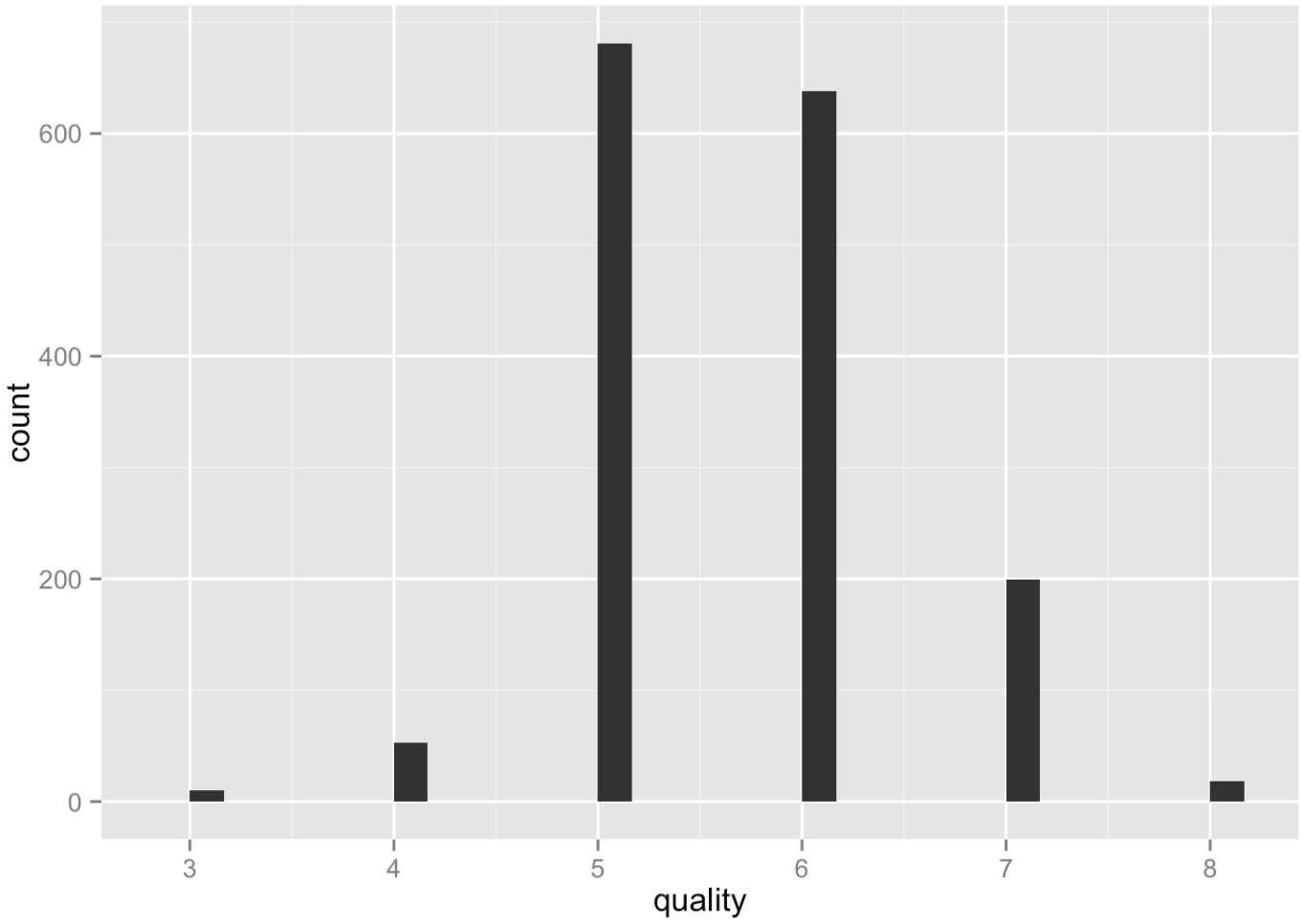
```
## 'data.frame':    1599 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num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

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

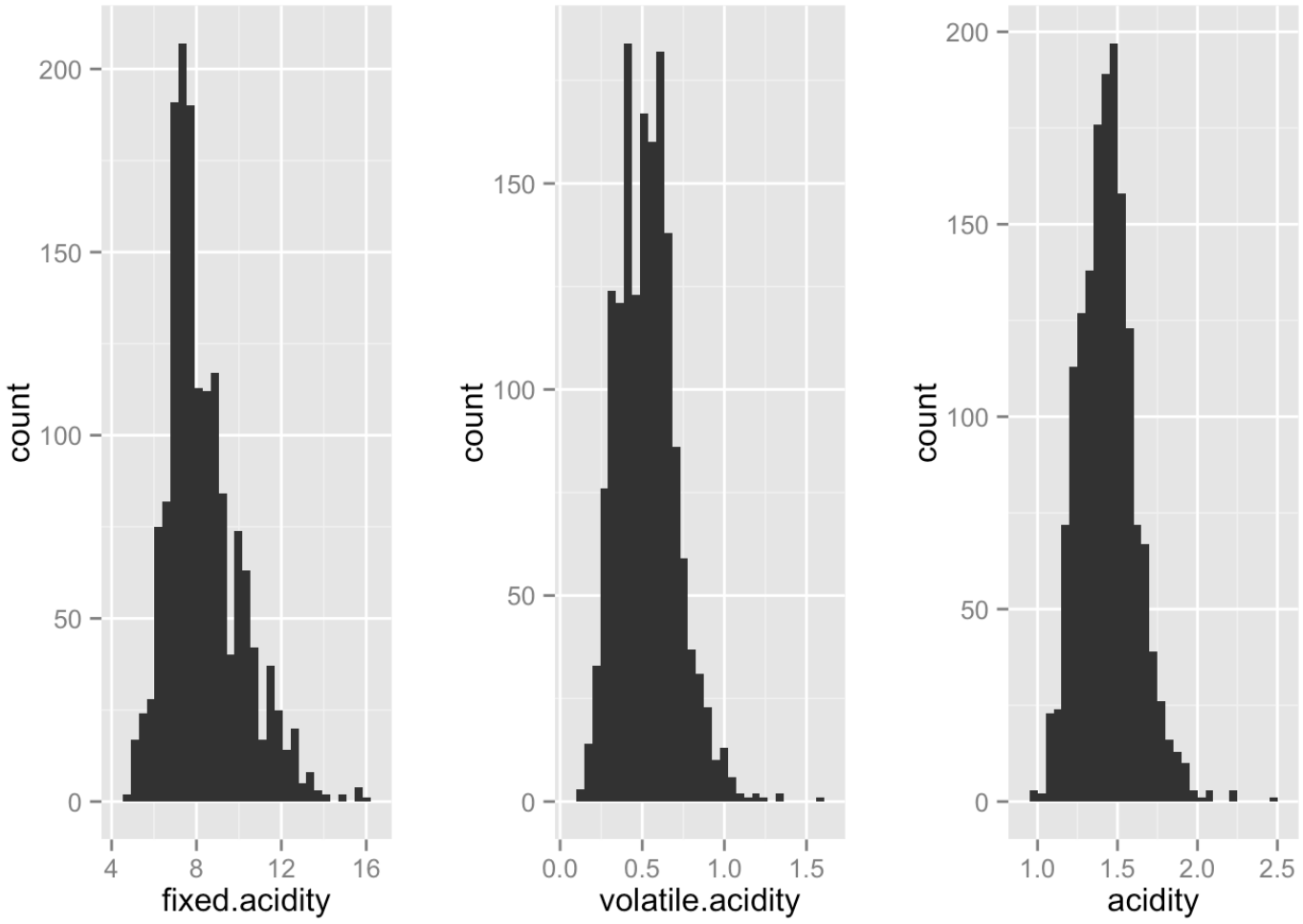
```
##        X           fixed.acidity   volatile.acidity  citric.acid
##  Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000
##  1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090
##  Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260
##  Mean   : 800.0   Mean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

We can see the majority of the wines are of quality 5-6 (average).

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```




```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

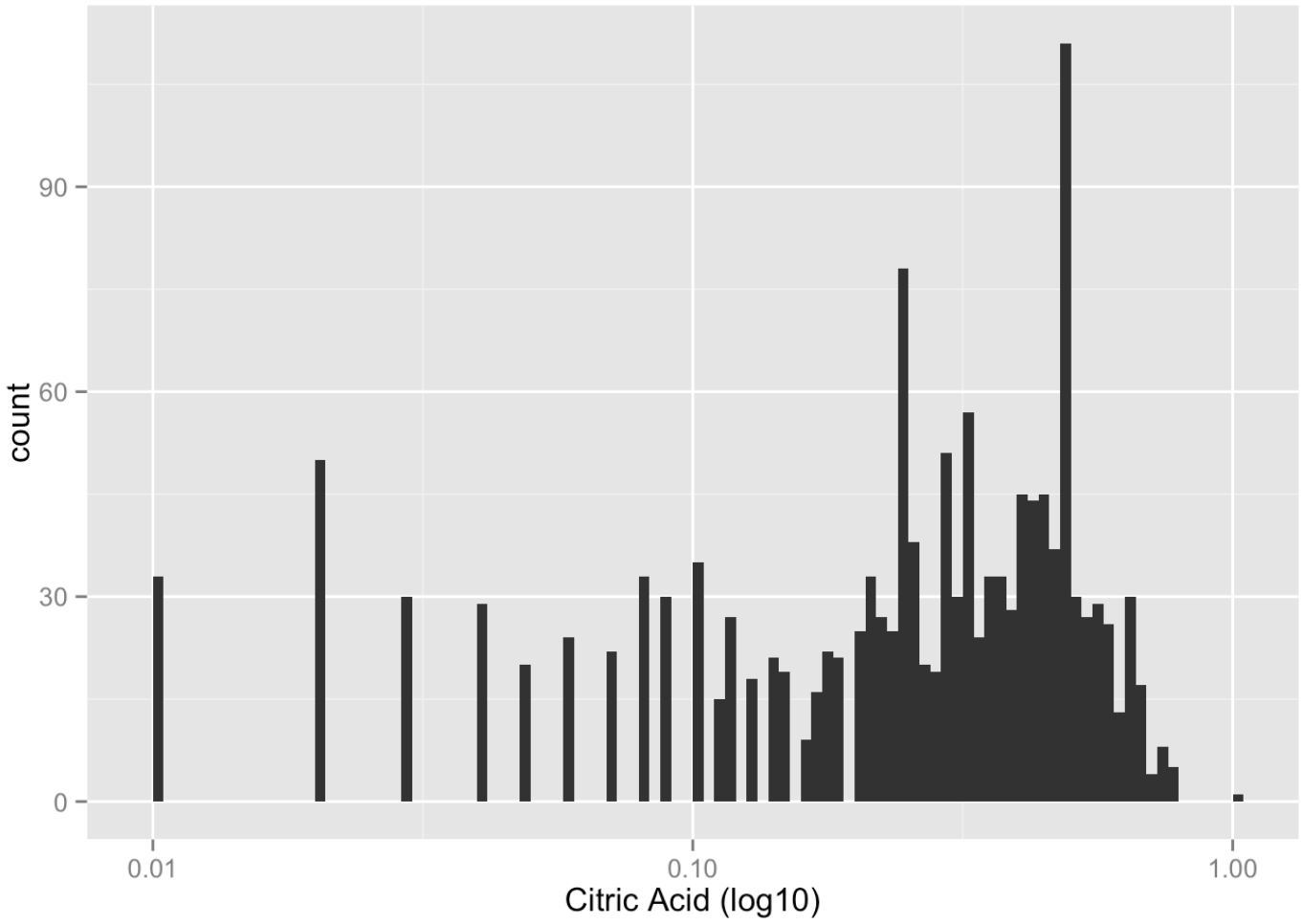


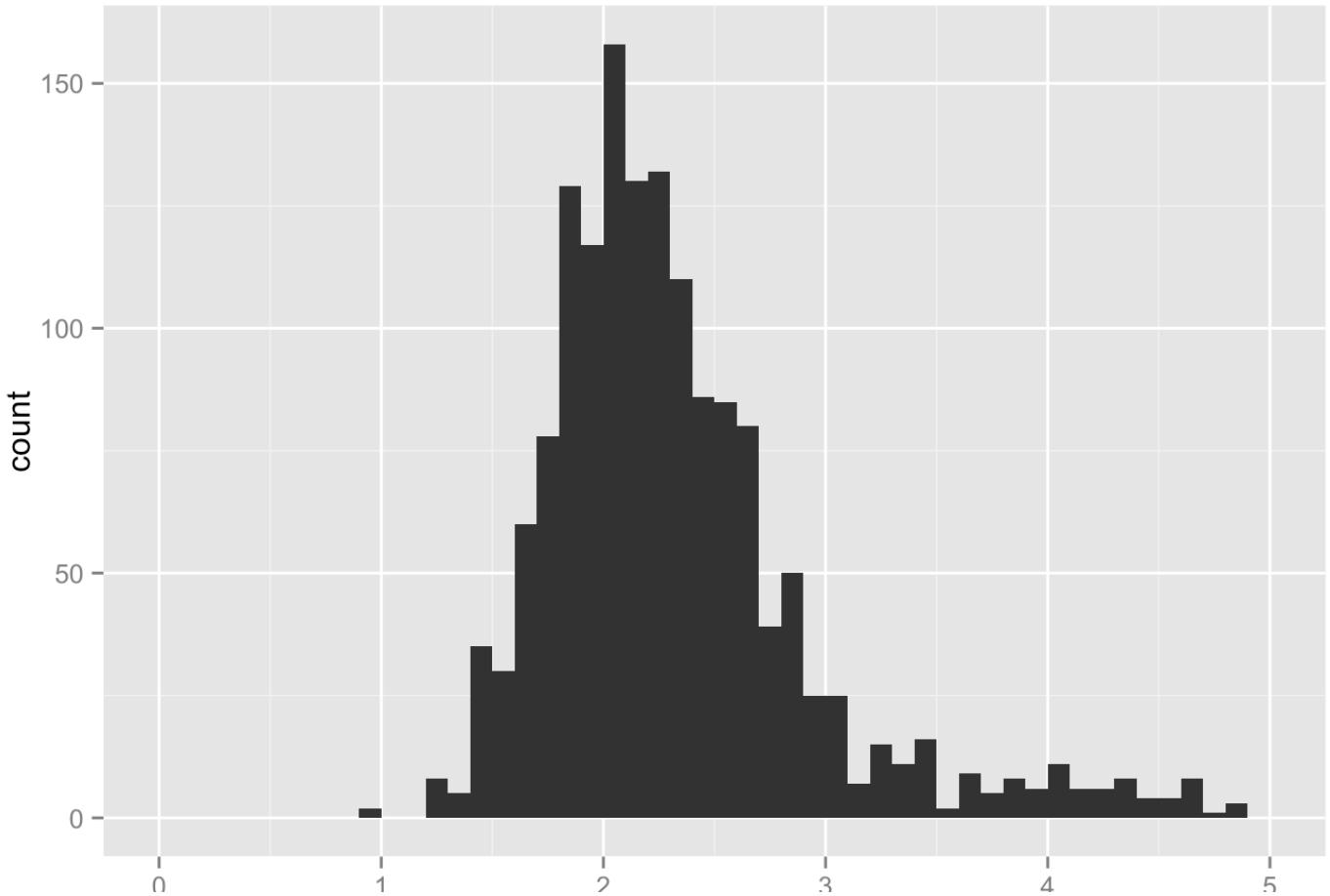
I created a new variable called acidity which successfully mimics the two original acidity variables (fixed and volatile acidity).

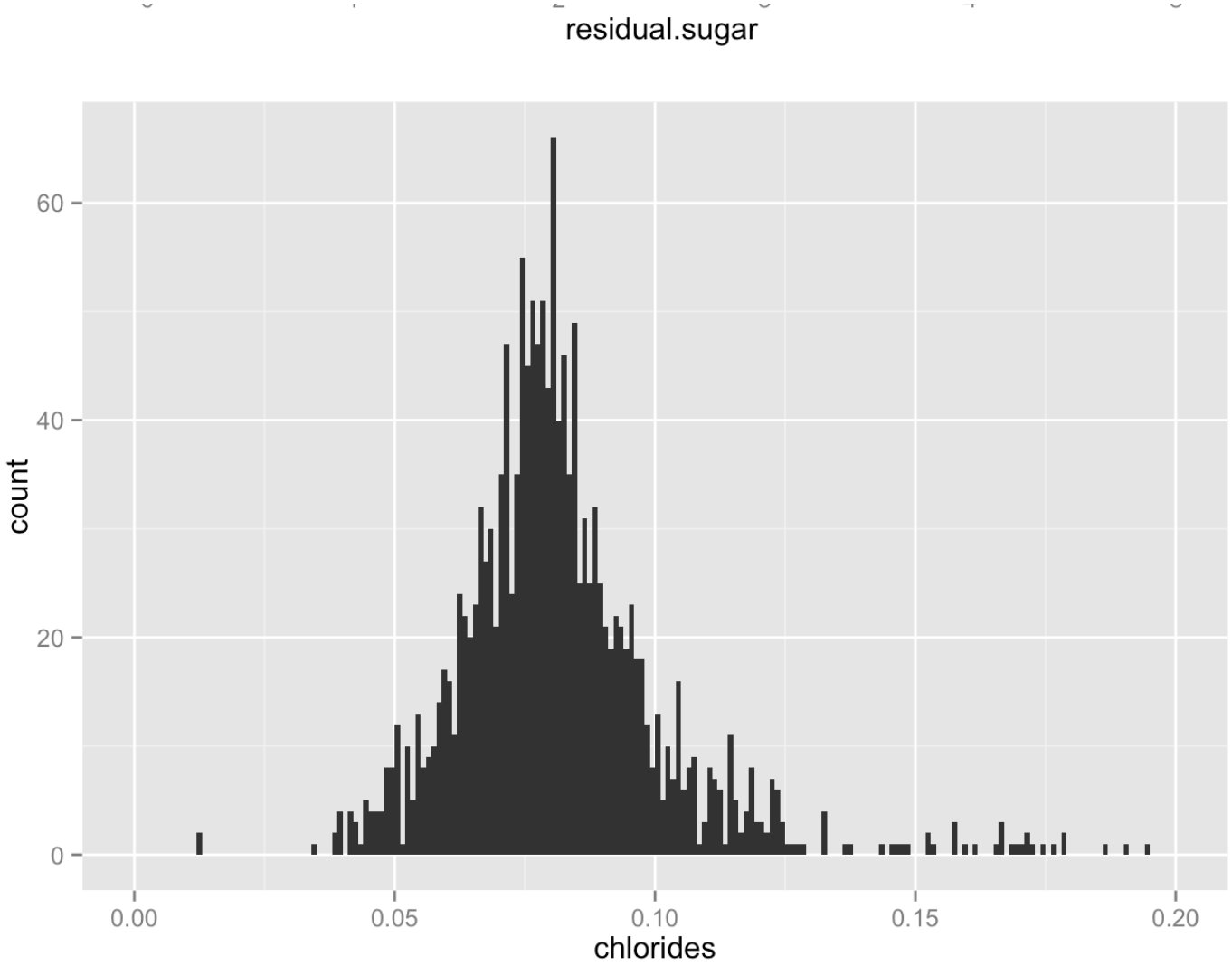
Acidity = log(Fixed Acidity) + Volatile Acidity.

I took the log of fixed acidity since the values are much larger compared to volatile acidity.

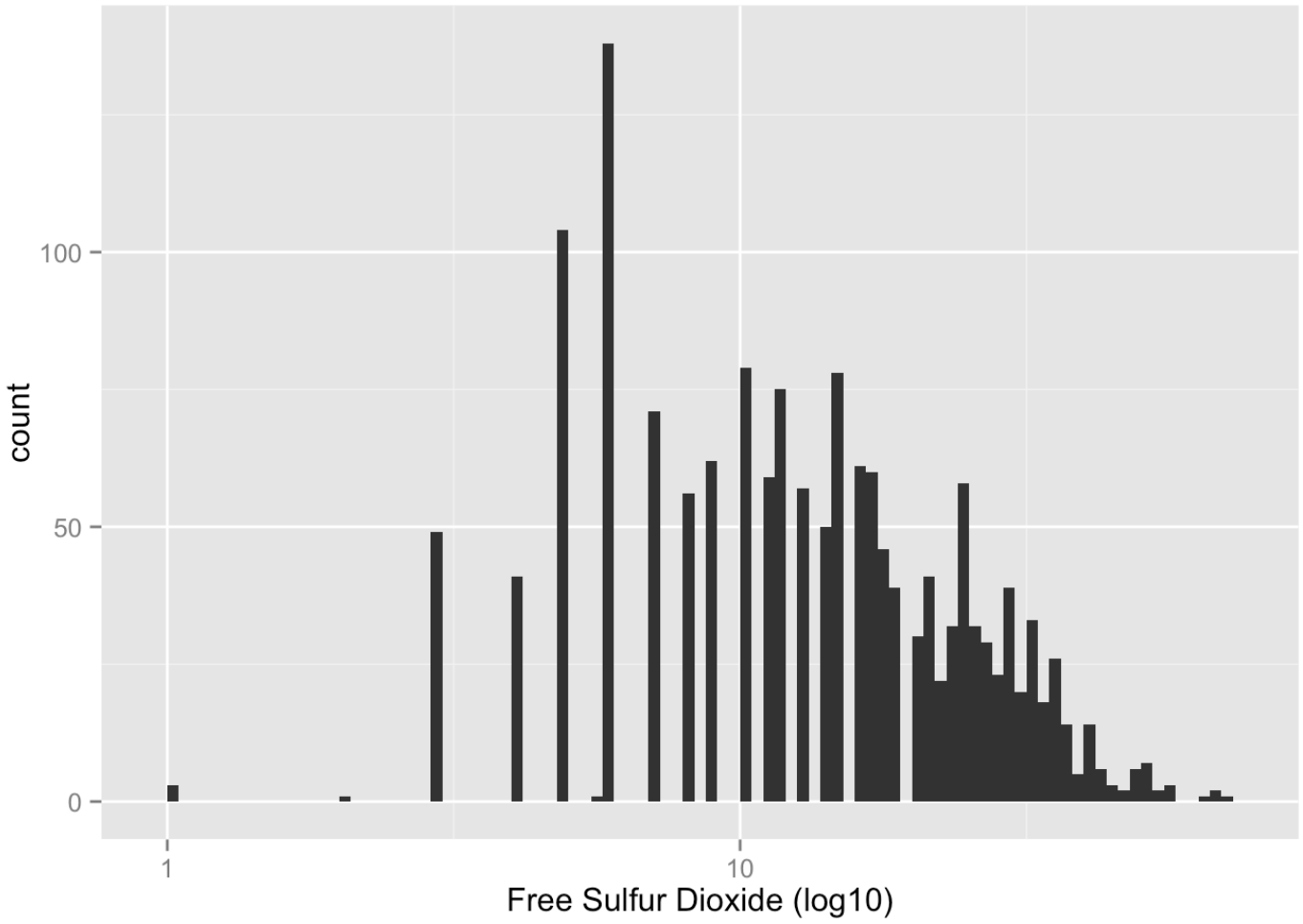
```
## Warning: position_stack requires constant width: output may be incorrect
```


90
60
30
0
count
0.01
0.10
1.00
Citric Acid (log10)

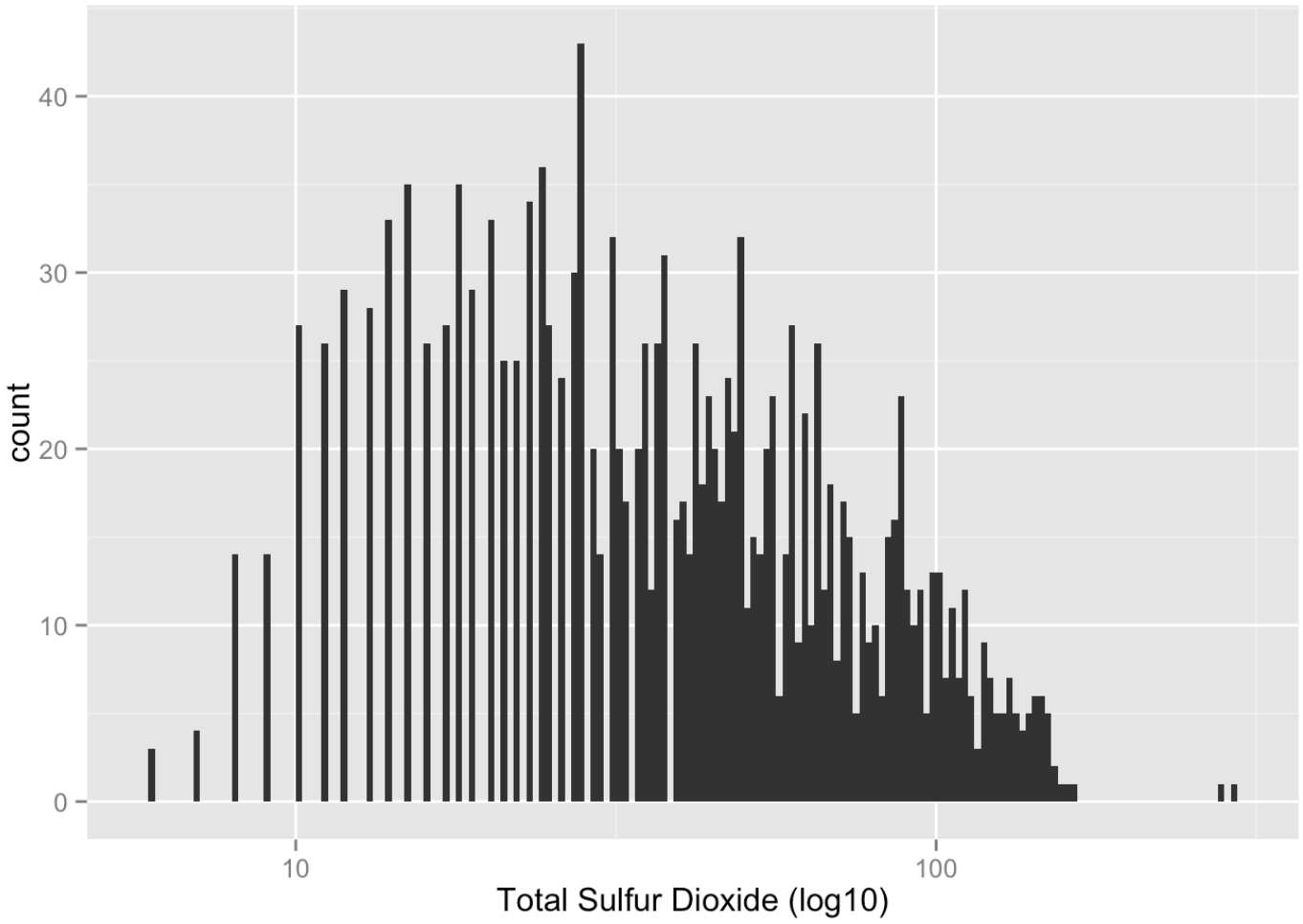

150
100
50
0
count
0
1
2
3
4
5



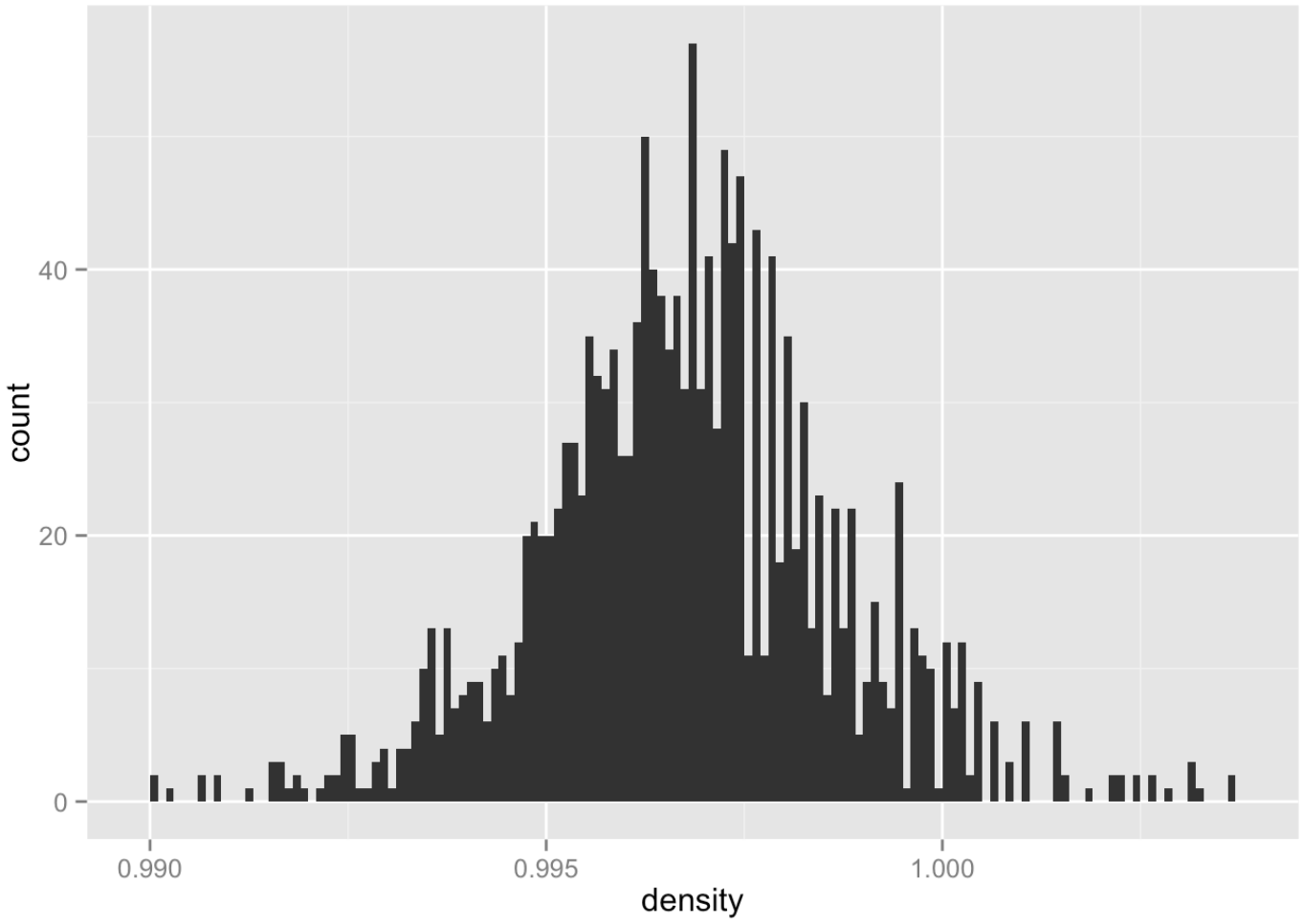

```
## Warning: position_stack requires constant width: output may be incorrect
```

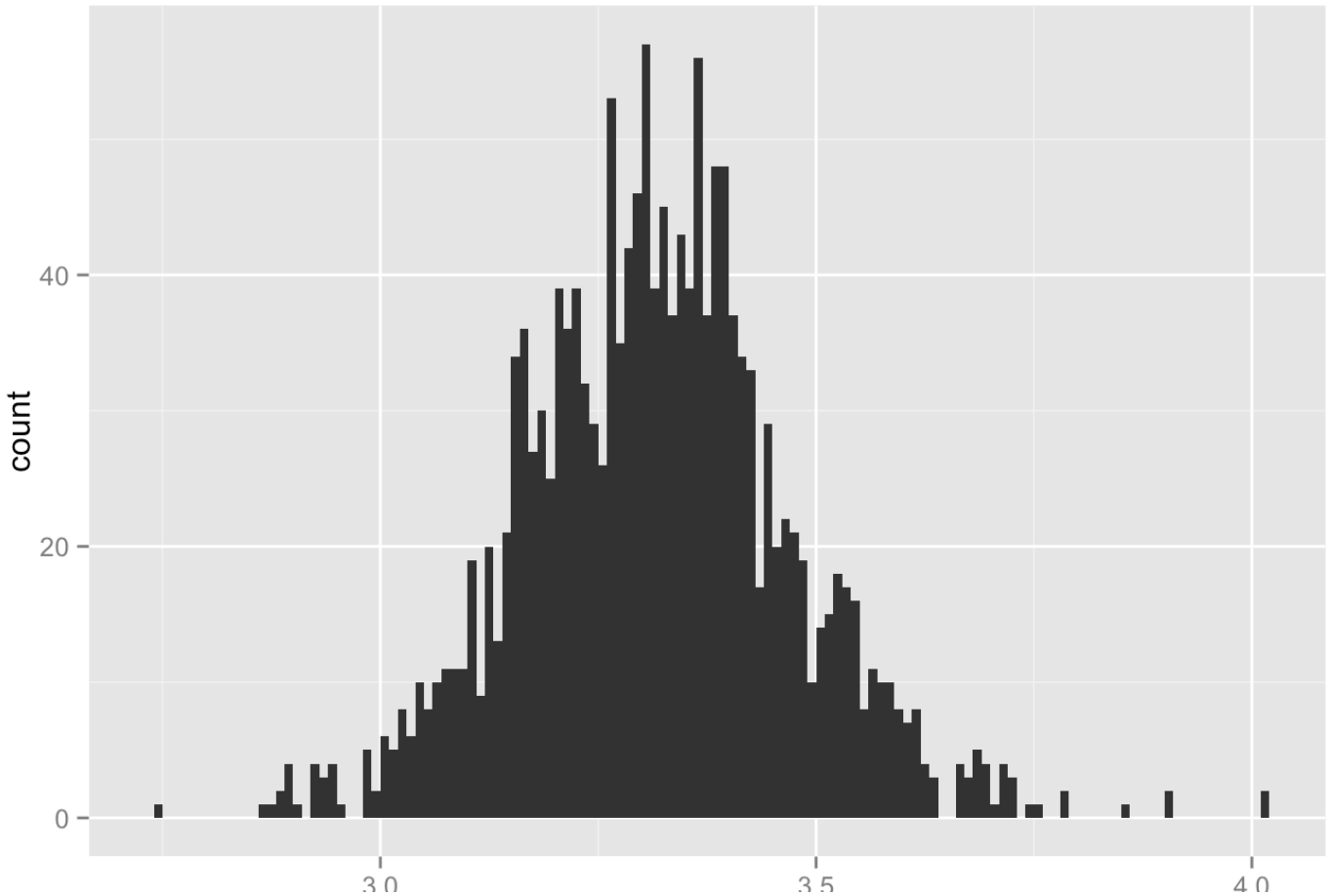



```
## Warning: position_stack requires constant width: output may be incorrect
```



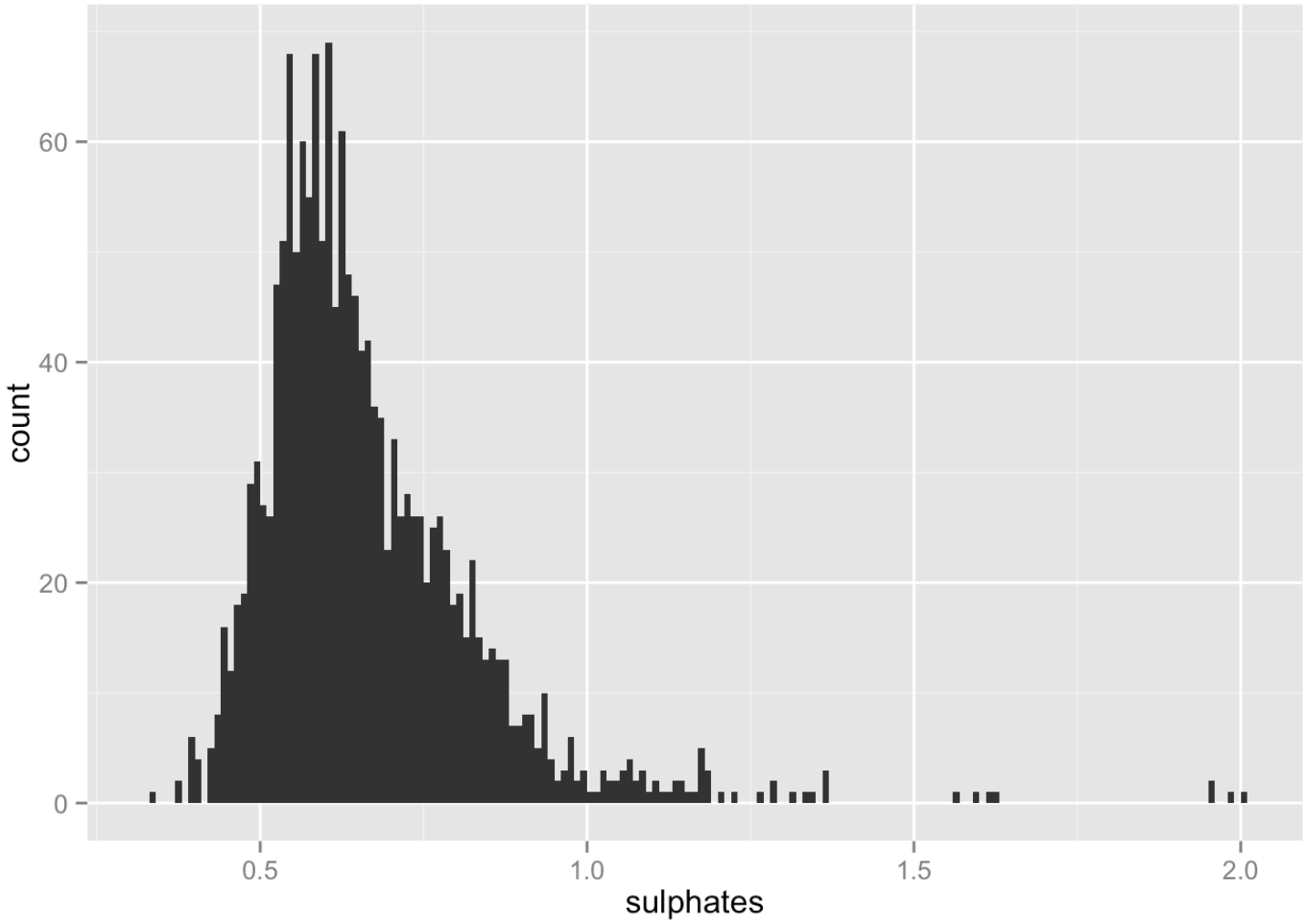

```
## Warning: position_stack requires constant width: output may be incorrect
```


count
40
20
0
0.990
0.995
1.000
density

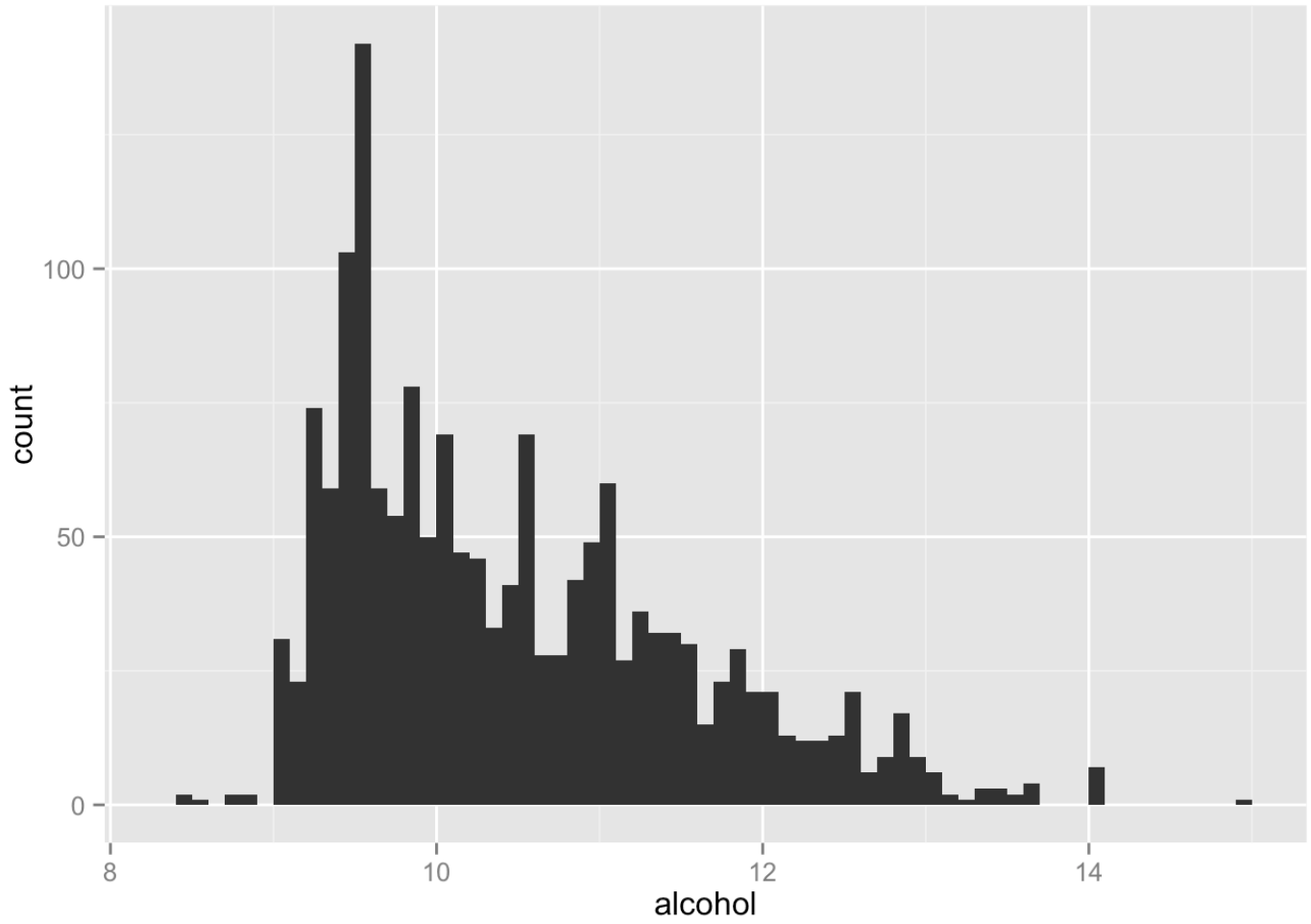

count
40
20
0
3.0
3.5
4.0

pH

```
## Warning: position_stack requires constant width: output may be incorrect
```




```
## Warning: position_stack requires constant width: output may be incorrect
```



The variables seem to be approximately normally distributed.

Univariate Analysis

What is the structure of your dataset?

There are 1,599 wines in the dataset with 11 features(fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates, alcohol. The main indicator of Red wine quality is the ordered factor variable called 'quality' which has the following levels:

worst ——————-> best

quality: 3 - 8

Here is a description of the variables from the data source: 1 - fixed acidity: most acids involved with wine or fixed or nonvolatile (do not evaporate readily)

2 - volatile acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste

3 - citric acid: found in small quantities, citric acid can add 'freshness' and flavor to wines

4 - residual sugar: the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet

5 - chlorides: the amount of salt in the wine

6 - free sulfur dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine

7 - total sulfur dioxide: amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine

8 - density: the density of water is close to that of water depending on the percent alcohol and sugar content

9 - pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale

10 - sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant

11 - alcohol: the percent alcohol content of the wine

Output variable (based on sensory data): 12 - quality (score between 0 and 10)

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

The main feature is the quality variable. I'd like to see what features drive this indicator and can be used as a predictor of the quality of the wine. We can see at first glance, that the majority of wines tend to be in the 5-6 quality level category, with very few wines being in the extreme (3 or 8) buckets.

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

Alcohol, pH, density, acidity can all be potential contributors. I also suspect some of the indicators are pointing to the same underlying characteristics e.g. free.sulfur.dioxide v.s. total.sulfur.dioxide, fixed.acidity v.s. volatile.acidity.

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

I created the variable acidity which is basically the sum of fixed acidity and volatile acidity to see if there was an aggregate affect in line with the individual features. I took the log (base 10) of fixed acicidty before summing because the values were much higher in comparison to volatile acidity.

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

did you do this?

I log transformed citric acid, free sulfur dioxide and total sulfur dioxide. The reason was that the parameters were really small values and I wanted to see a distribution which would exaggerate the differences. The values were also getting skewed to the right because of some outliers the transformation helped in seeing everything together.

Additionally, I also excluded some really large values (outliers) from the histogram plots for residual sugar and chlorides. I could have used a log transformation here as well but the bulk of the data was already normally distributed.

Overall, it looks like all the variables follow an aproximately normal distribution.

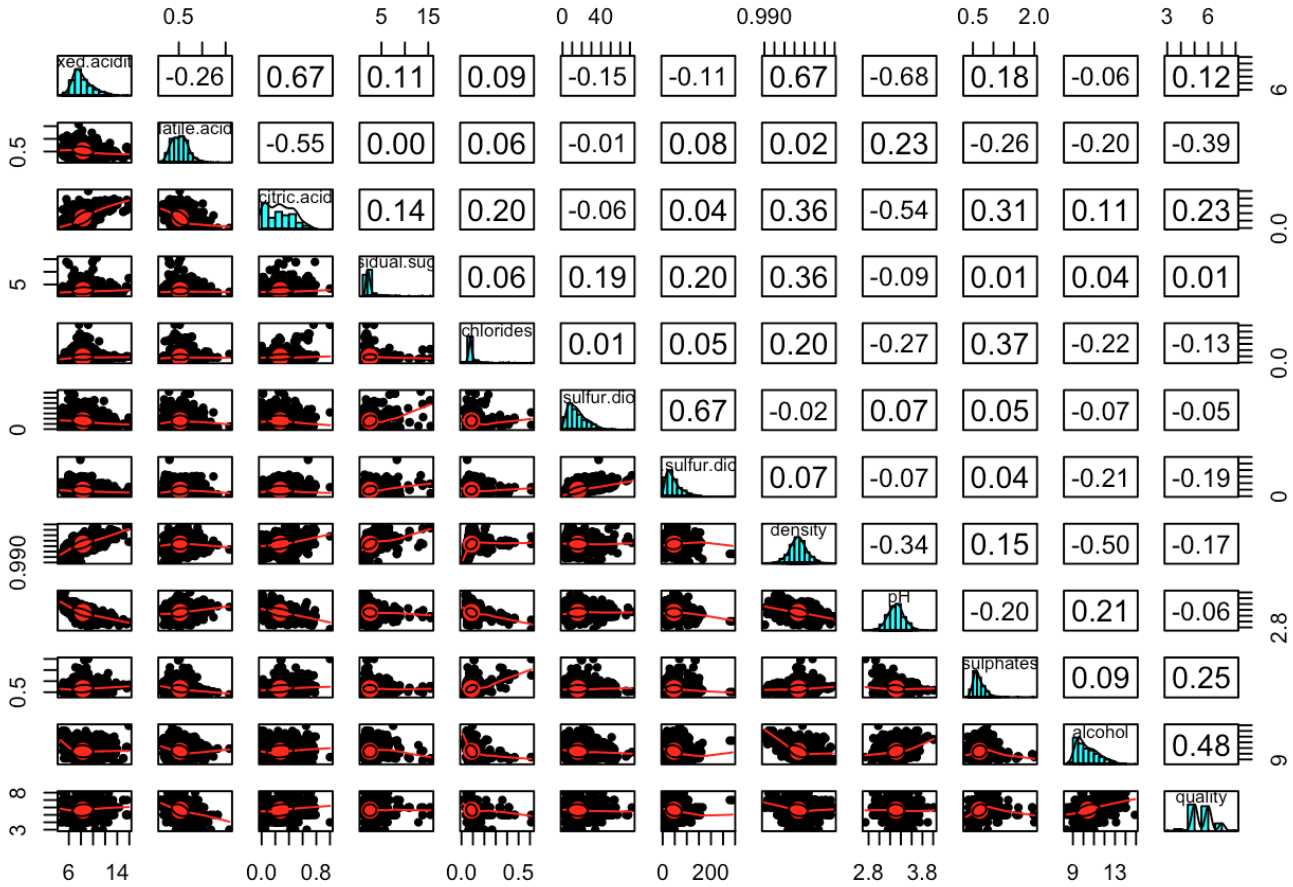
Bivariate Plots Section

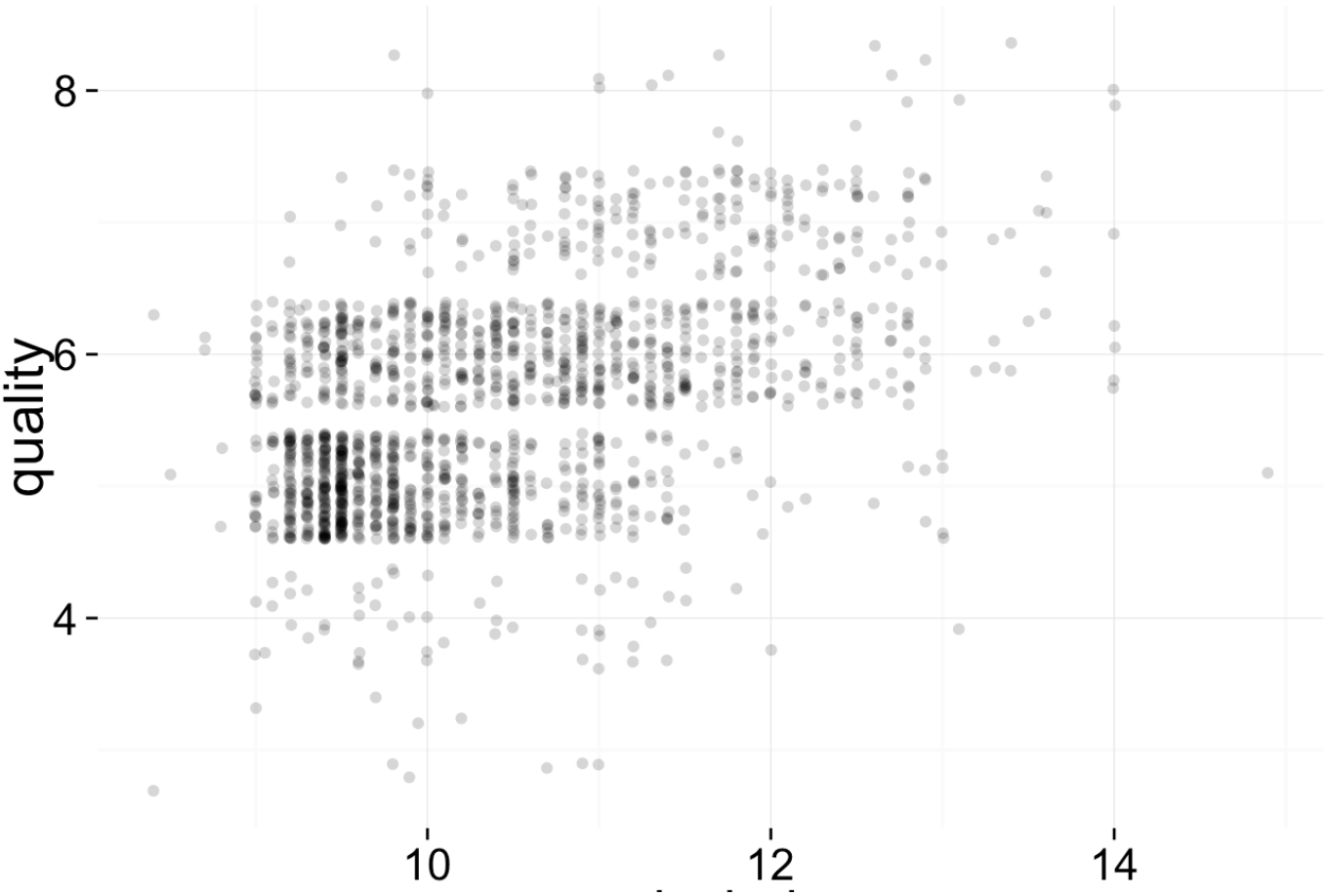
```
##                              X fixed.acidity volatile.acidity
## X                1.000000000   -0.26848392     -0.008815099
## fixed.acidity   -0.268483920    1.00000000     -0.256130895
## volatile.acidity -0.008815099   -0.25613089      1.000000000
## citric.acid     -0.153551355    0.67170343     -0.552495685
## residual.sugar  -0.031260835    0.11477672      0.001917882
## chlorides       -0.119868519    0.09370519      0.061297772
## free.sulfur.dioxide  0.090479643   -0.15379419     -0.010503827
## total.sulfur.dioxide -0.117849669   -0.11318144      0.076470005
## density         -0.368372087    0.66804729      0.022026232
## pH               0.136005328   -0.68297819      0.234937294
## sulphates       -0.125306999    0.18300566     -0.260986685
## alcohol          0.245122841   -0.06166827     -0.202288027
## quality          0.066452608    0.12405165     -0.390557780
## acidity         -0.141582717    0.22410387      0.882421368
##                 citric.acid residual.sugar    chlorides
## X               -0.15355136   -0.031260835 -0.119868519
## fixed.acidity    0.67170343    0.114776724  0.093705186
## volatile.acidity -0.55249568    0.001917882  0.061297772
## citric.acid      1.00000000    0.143577162  0.203822914
## residual.sugar   0.14357716    1.000000000  0.055609535
## chlorides        0.20382291    0.055609535  1.000000000
## free.sulfur.dioxide  -0.06097813    0.187048995  0.005562147
## total.sulfur.dioxide  0.03553302    0.203027882  0.047400468
## density          0.36494718    0.355283371  0.200632327
## pH              -0.54190414   -0.085652422 -0.265026131
## sulphates        0.31277004    0.005527121  0.371260481
## alcohol          0.10990325    0.042075437 -0.221140545
## quality          0.22637251    0.013731637 -0.128906560
## acidity         -0.23105679    0.055066311  0.115602970
##                 free.sulfur.dioxide total.sulfur.dioxide     density
## X                      0.090479643          -0.11784967 -0.36837209
## fixed.acidity         -0.153794193          -0.11318144  0.66804729
## volatile.acidity      -0.010503827           0.07647000  0.02202623
```

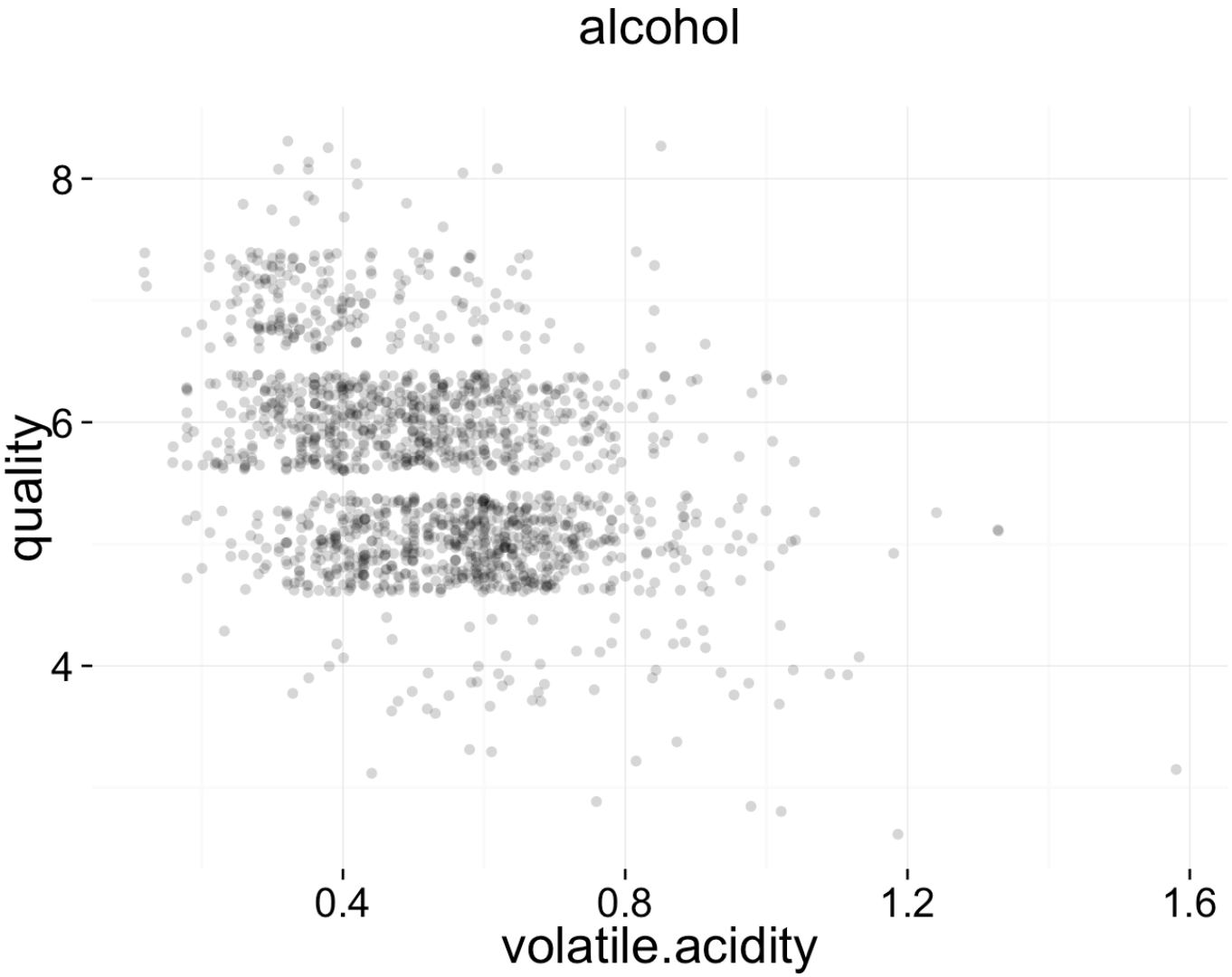
```
## citric.acid                -0.060978129           0.03553302  0.36494718
## residual.sugar              0.187048995           0.20302788  0.35528337
## chlorides                   0.005562147           0.04740047  0.20063233
## free.sulfur.dioxide         1.000000000           0.66766645 -0.02194583
## total.sulfur.dioxide        0.667666450           1.00000000  0.07126948
## density                    -0.021945831           0.07126948  1.00000000
## pH                          0.070377499          -0.06649456 -0.34169933
## sulphates                   0.051657572           0.04294684  0.14850641
## alcohol                    -0.069408354          -0.20565394 -0.49617977
## quality                    -0.050656057          -0.18510029 -0.17491923
## acidity                    -0.083947079           0.02601640  0.35026277
##                                  pH    sulphates     alcohol     quality
## X                       0.13600533 -0.125306999  0.24512284  0.06645261
## fixed.acidity          -0.68297819  0.183005664 -0.06166827  0.12405165
## volatile.acidity        0.23493729 -0.260986685 -0.20228803 -0.39055778
## citric.acid            -0.54190414  0.312770044  0.10990325  0.22637251
## residual.sugar         -0.08565242  0.005527121  0.04207544  0.01373164
## chlorides              -0.26502613  0.371260481 -0.22114054 -0.12890656
## free.sulfur.dioxide     0.07037750  0.051657572 -0.06940835 -0.05065606
## total.sulfur.dioxide -0.06649456  0.042946836 -0.20565394 -0.18510029
## density                -0.34169933  0.148506412 -0.49617977 -0.17491923
## pH                      1.00000000 -0.196647602  0.20563251 -0.05773139
## sulphates              -0.19664760  1.000000000  0.09359475  0.25139708
## alcohol                 0.20563251  0.093594750  1.00000000  0.47616632
## quality                -0.05773139  0.251397079  0.47616632  1.00000000
## acidity                -0.10727457 -0.174937840 -0.25144718 -0.33711431
##                            acidity
## X                      -0.14158272
## fixed.acidity           0.22410387
## volatile.acidity        0.88242137
## citric.acid            -0.23105679
## residual.sugar          0.05506631
## chlorides               0.11560297
## free.sulfur.dioxide    -0.08394708
## total.sulfur.dioxide    0.02601640
## density                 0.35026277
## pH                     -0.10727457
## sulphates              -0.17493784
## alcohol                -0.25144718
## quality                -0.33711431
## acidity                 1.00000000
```

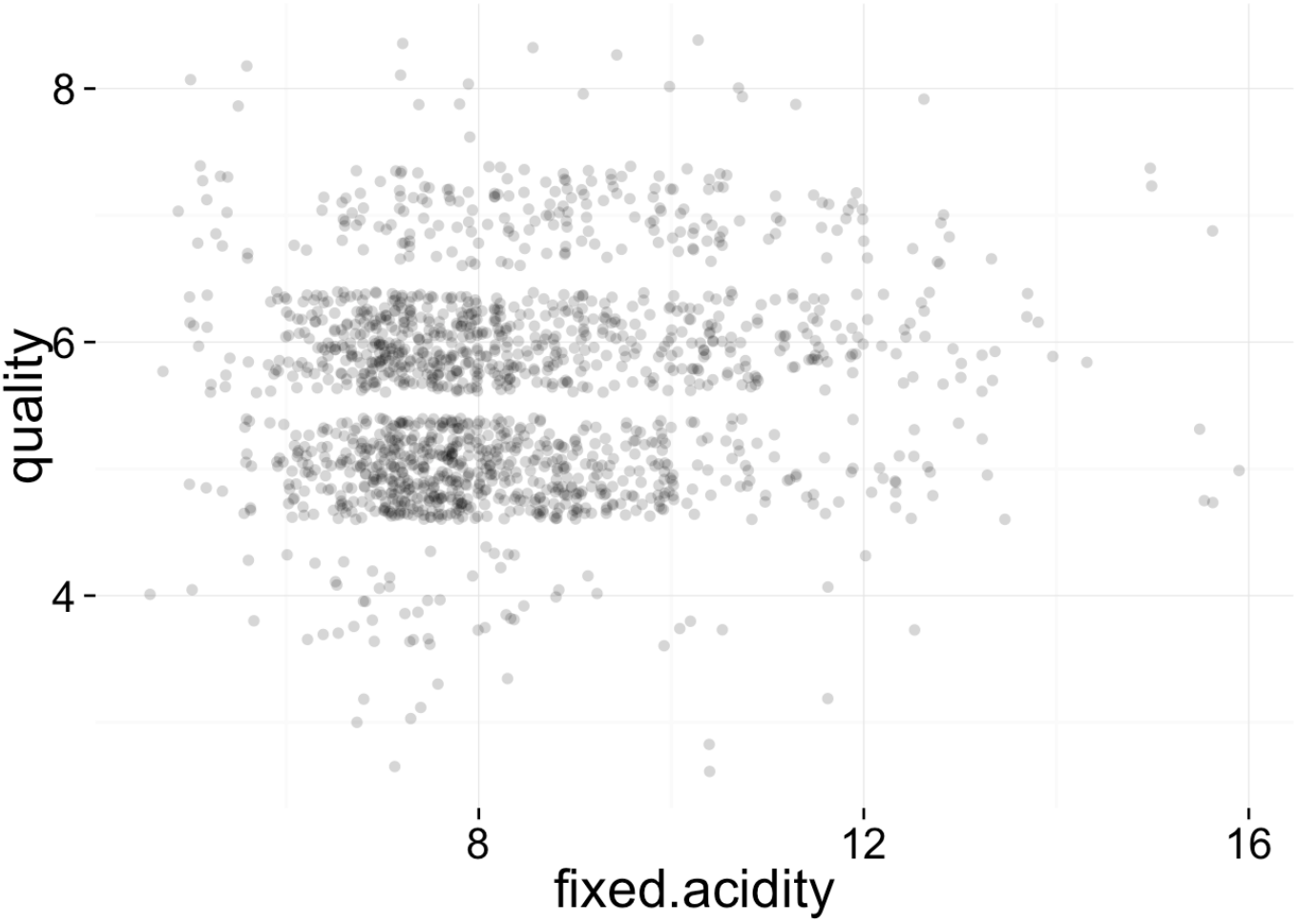
```
##              X fixed.acidity volatile.acidity citric.acid residual.sugar
## [1,] 0.06645261     0.1240516       -0.3905578   0.2263725     0.01373164
##       chlorides free.sulfur.dioxide total.sulfur.dioxide    density
## [1,] -0.1289066         -0.05065606           -0.1851003 -0.1749192
##               pH sulphates   alcohol quality    acidity
## [1,] -0.05773139 0.2513971 0.4761663       1 -0.3371143
```

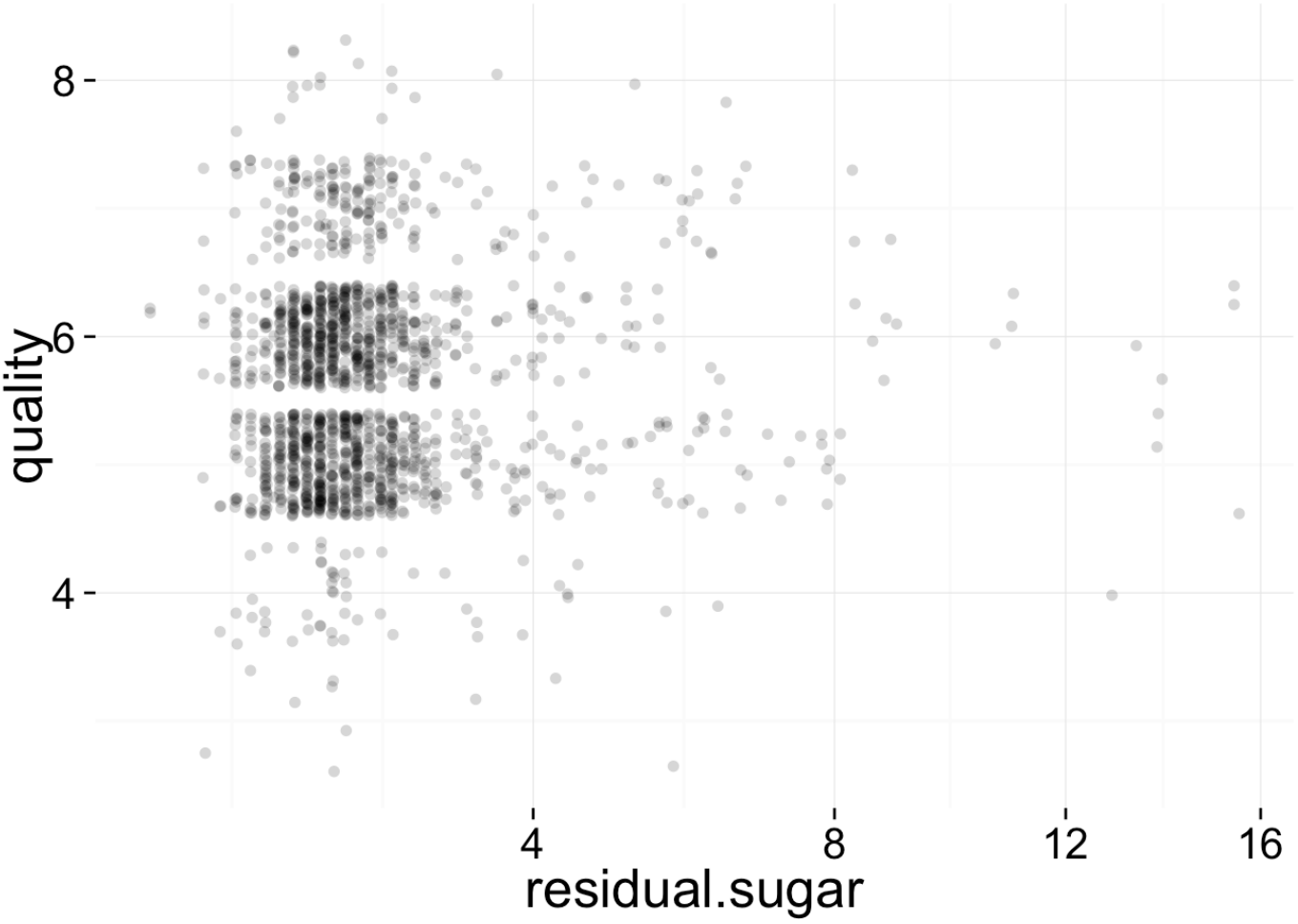
There does not seem to be any high correlations for any of the variables. The highest seems to be alcohol content but even that is below 0.5. On the other side, the volatile acidity seems to be almost equally negatively correlated with quality.

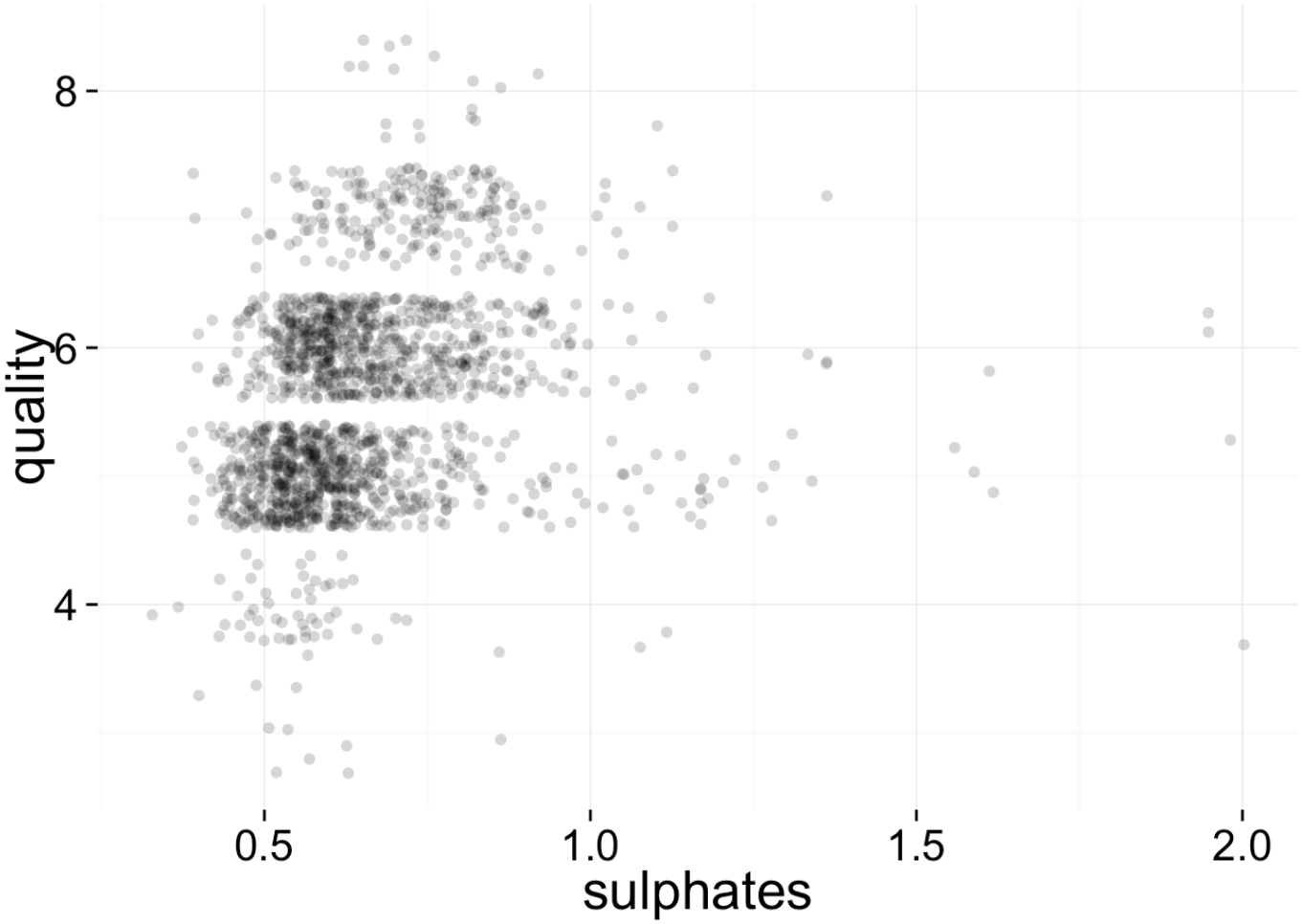

fixed.acidit
-0.26
0.67
0.11
0.09
-0.15
-0.11
0.67
-0.68
0.18
-0.06
0.12
latile.acid
-0.55
0.00
0.06
-0.01
0.08
0.02
0.23
-0.26
-0.20
-0.39
citric.acid
0.14
0.20
-0.06
0.04
0.36
-0.54
0.31
0.11
0.23
sidual.sug
0.06
0.19
0.20
0.36
-0.09
0.01
0.04
0.01
chlorides
0.01
0.05
0.20
-0.27
0.37
-0.22
-0.13
sulfur.dio
0.67
-0.02
0.07
0.05
-0.07
-0.05
sulfur.dio
0.07
-0.07
0.04
-0.21
-0.19
density
-0.34
0.15
-0.50
-0.17
pH
-0.20
0.21
-0.06
sulphates
0.09
0.25
alcohol
0.48
quality

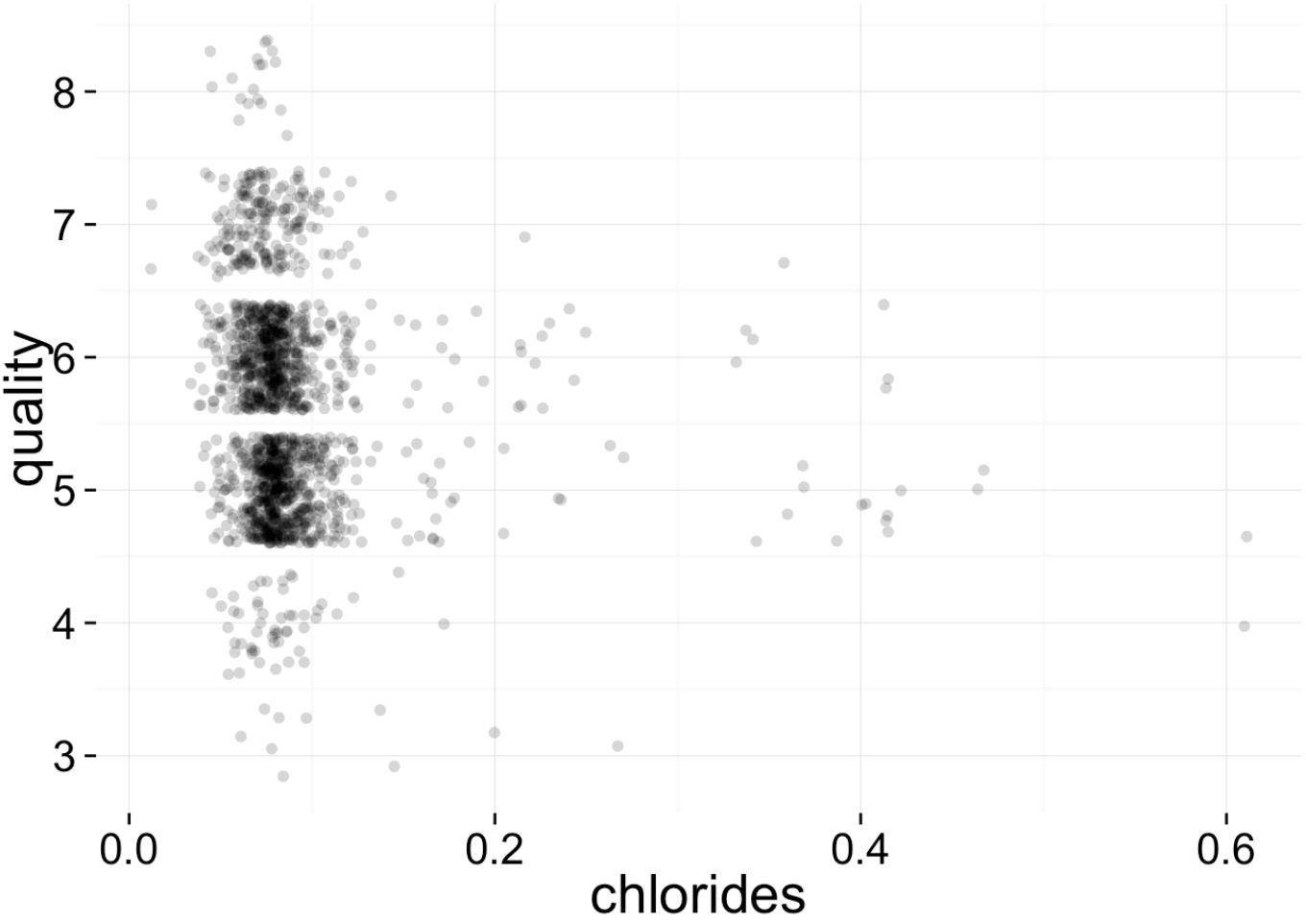

quality
8
6
4
10
12
14


alcohol
quality
8
6
4
0.4
0.8
1.2
1.6
volatile.acidity


quality
8
6
4
8
12
16
fixed.acidity

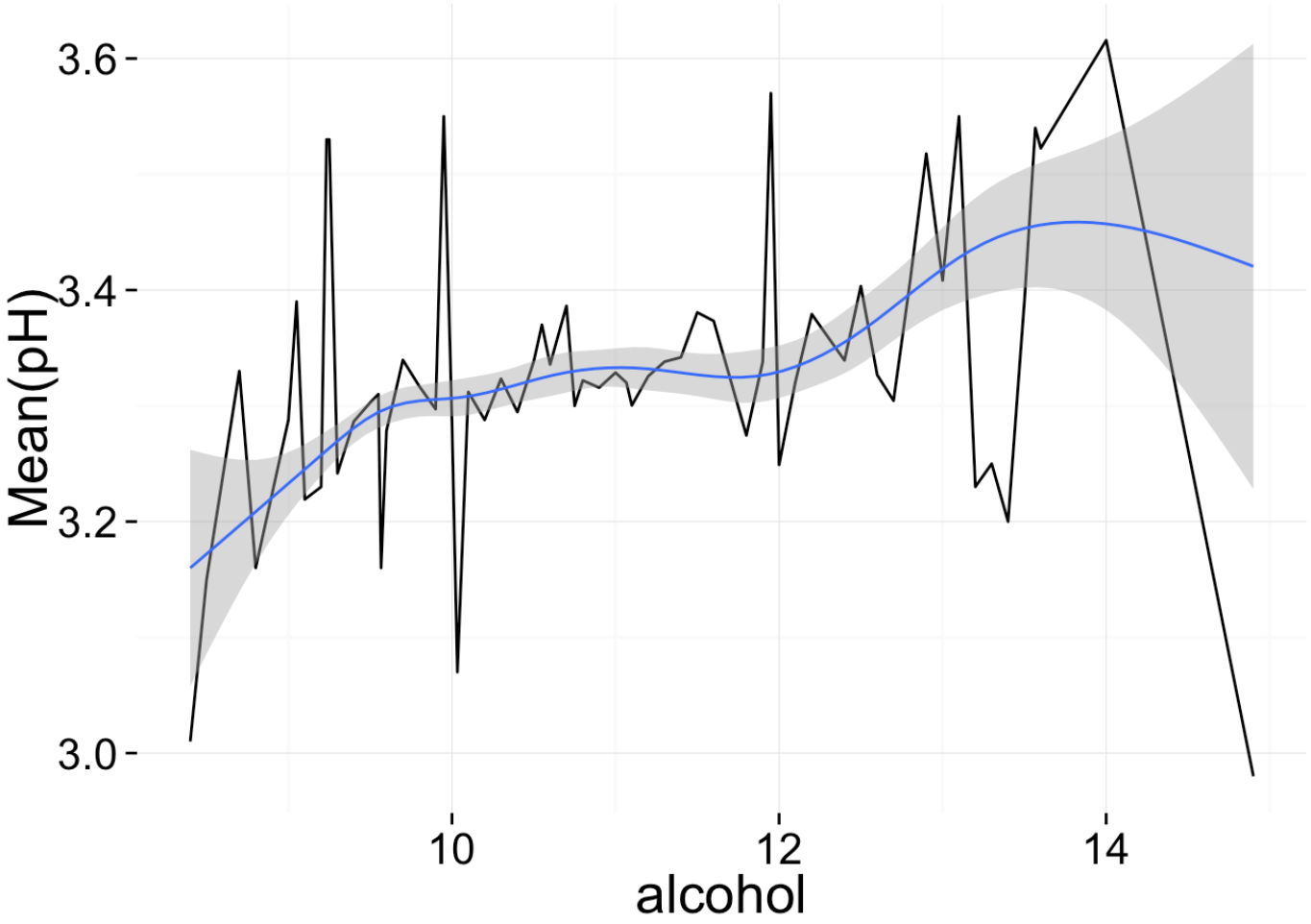

8
6
4
quality
4
8
12
16
residual.sugar


quality
8
6
4
0.5
1.0
1.5
2.0
sulphates

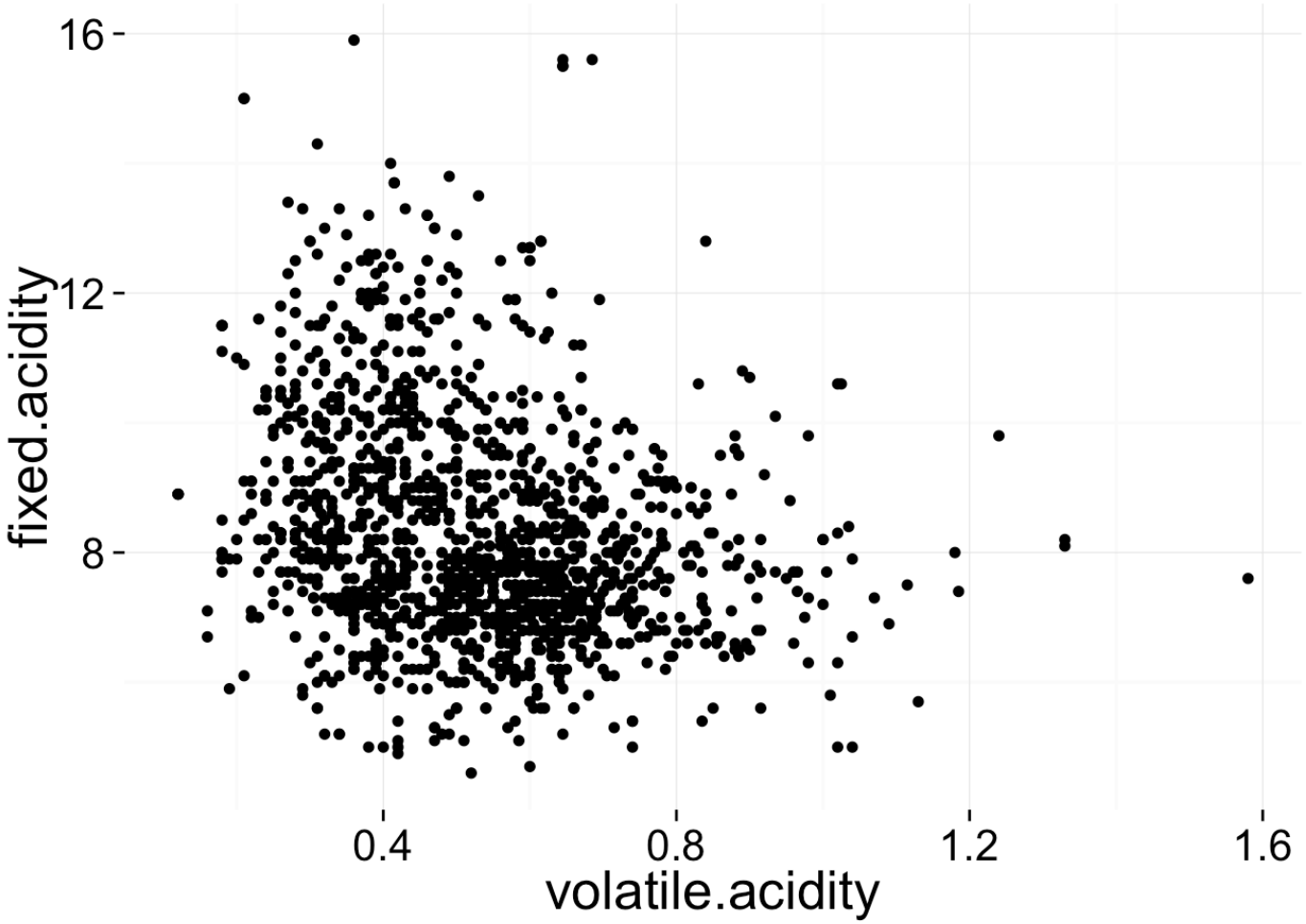


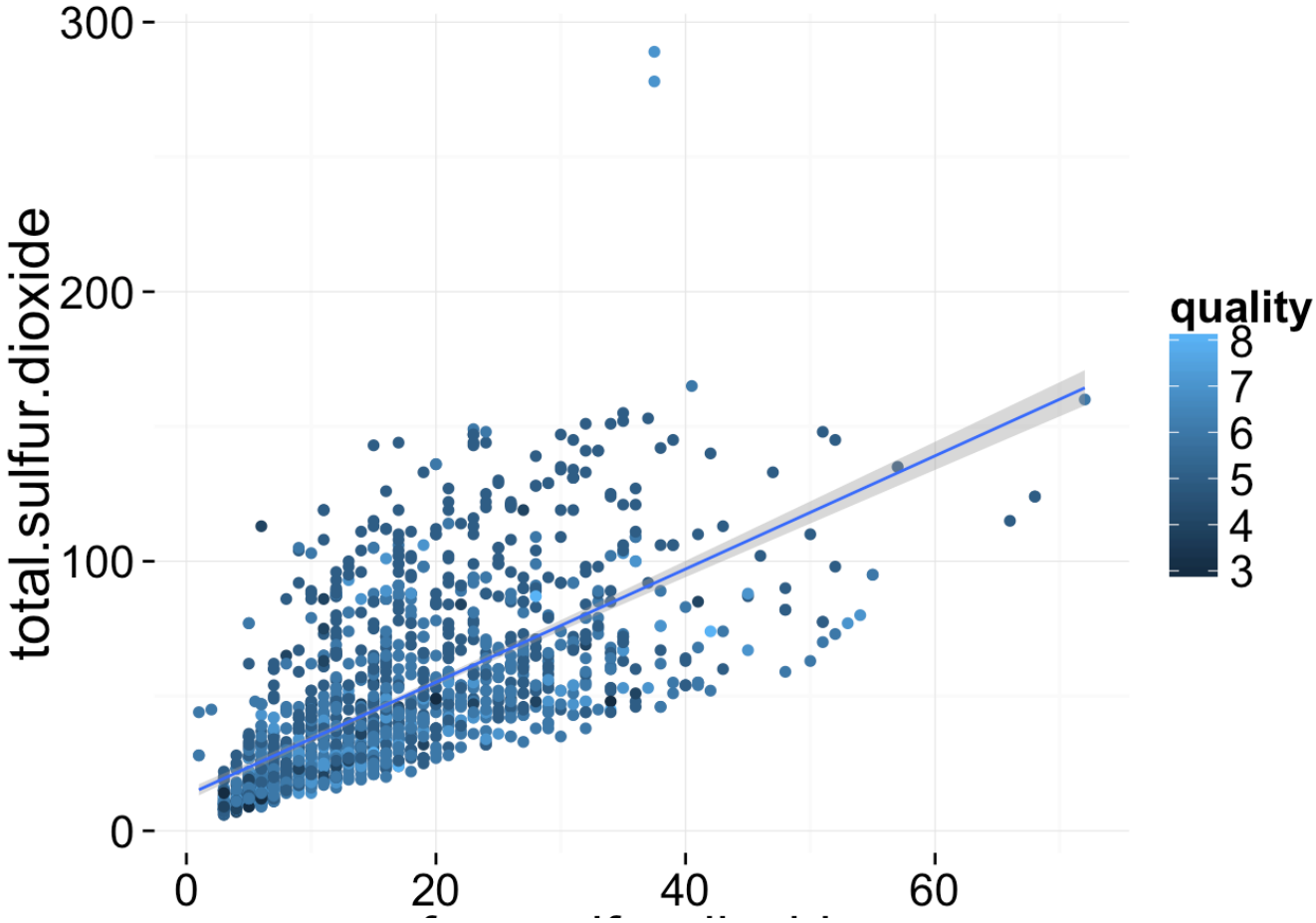
Looking at all the scatterplots of quality against the various variables, we see there is no one single variable that seems to have a strong relationship with quality.

```
## geom_smooth: method="auto" and size of largest group is >=1000, so using gam wi
th formula: y ~ s(x, bs = "cs"). Use 'method = x' to change the smoothing method.
```



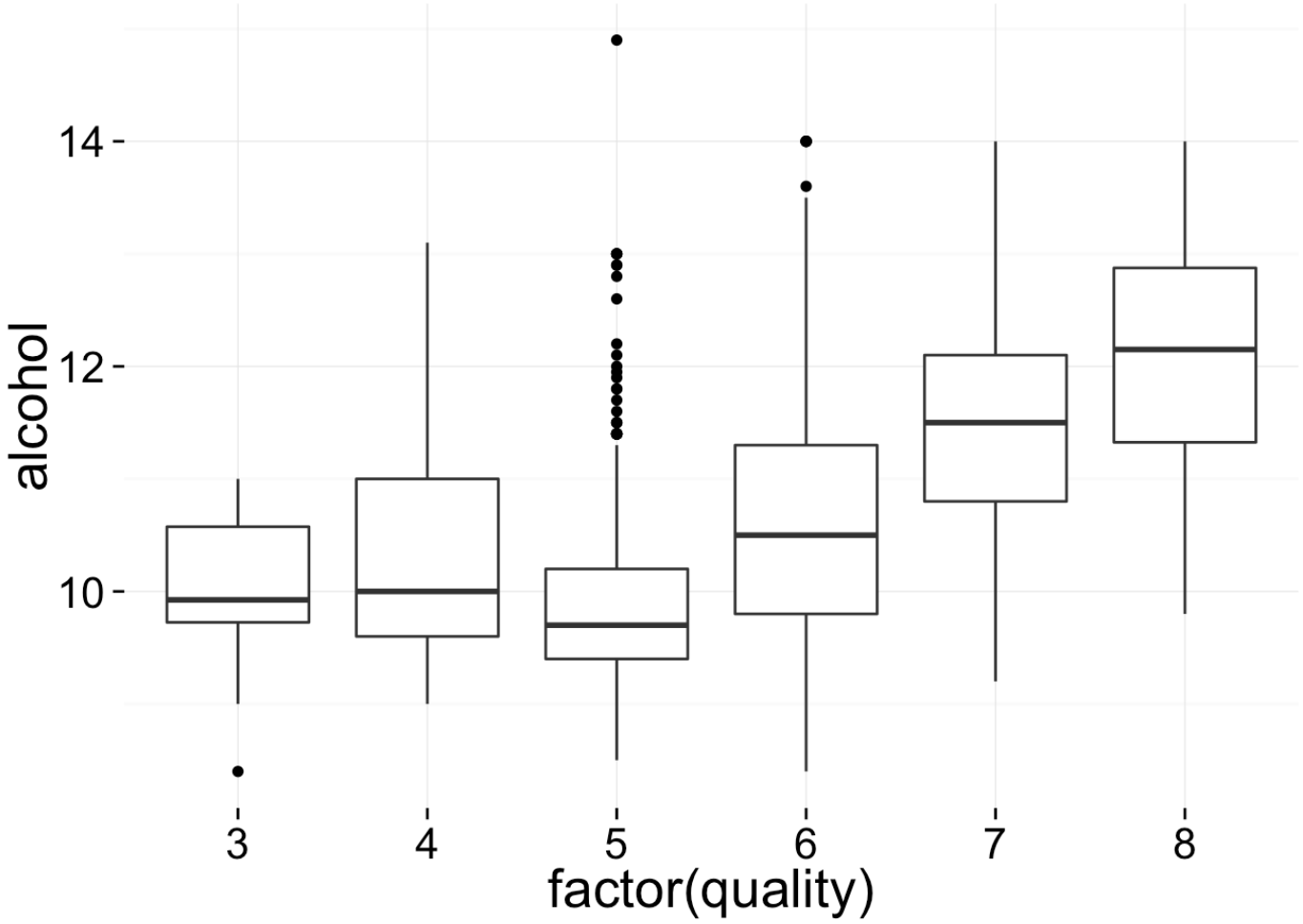
There does not seem to be any strong relationship between pH and alcohol.


16
12
8
fixed.acidity
0.4
0.8
1.2
1.6
volatile.acidity


300
200
100
0
total.sulfur.dioxide
0
20
40
60
quality
8
7
6
5
4
3

free.sulfur.dioxide

No strong relationship between volatile.acidity and fixed.acidity but there does seem to be one for free.sulfur.dioxide and total.sulfur.dioxide. This might be justification to take one of these varibales out of the final model.



We can see that higher alcohol content corresponds to higher quality.

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?

There are no strong relationships between the variables which is good because we have a less likelihood of having issues with collinearity in a regression model i.e. we can be confident in assuming the variables are independent of each other.

We can see some trends however: Higher alcohol and sulfates correspond to higher quality while lower volatile acidity, pH, and residual sugars correspond to lower quality red wines.

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

As expected, total sulfur dioxide and free sulfur dioxide are strongly correlated and have a linear relationship. I thought the correlation would be stronger than what it is (0.68) since I initially thought one variable was a subset of the other but that does not appear to be completely true. However, the relationship is strong enough for us to perhaps discard one of the variables in our prediction model.

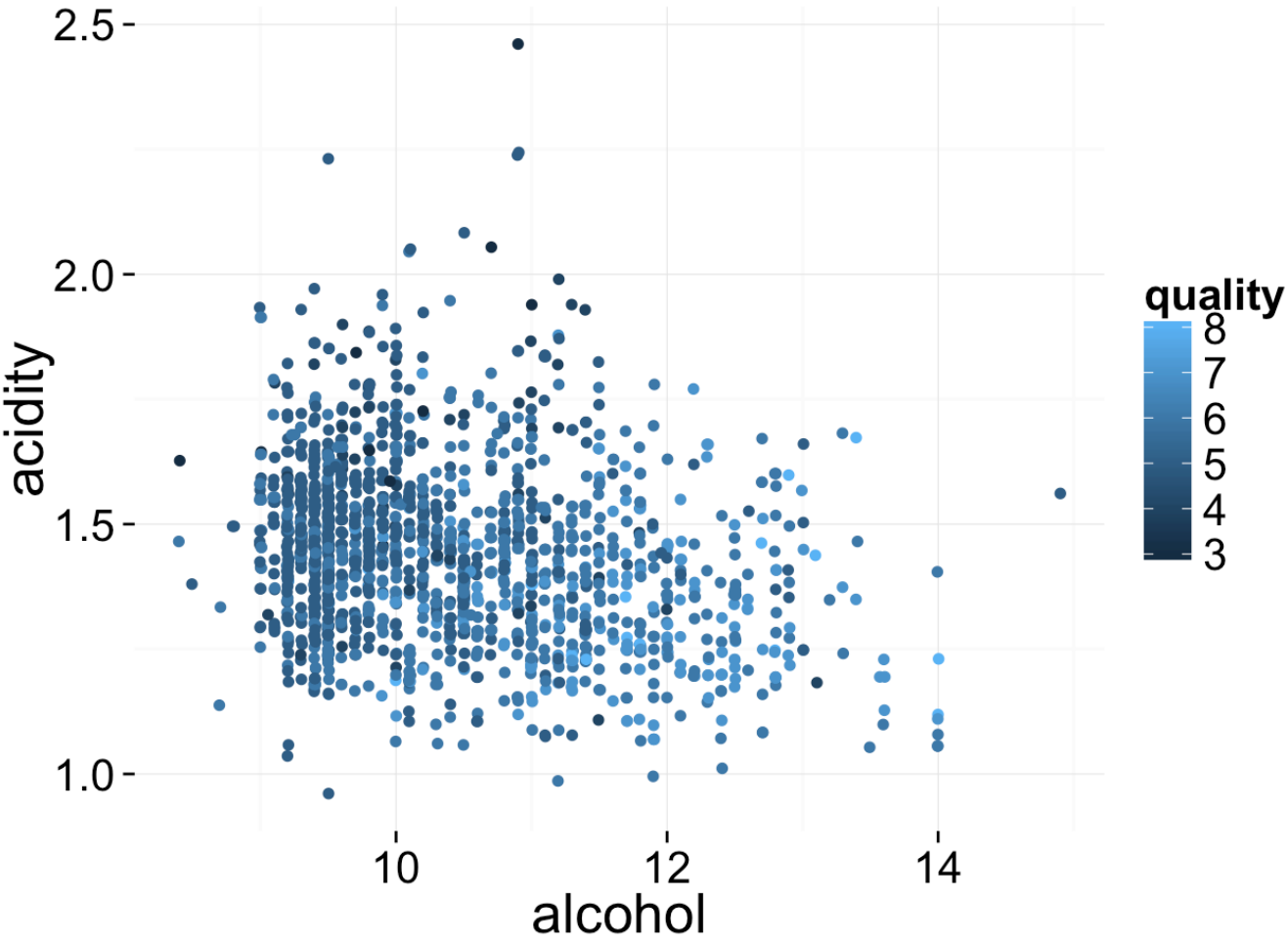
Surprisingly, fixed acidity and volatile acidity do not seem to have a strong correlation. Applying any log transformation to the fixed acidity variable does not change this either.

The only other (slightly) linear relationship is between fixed acidity and density (correlation of 0.67).

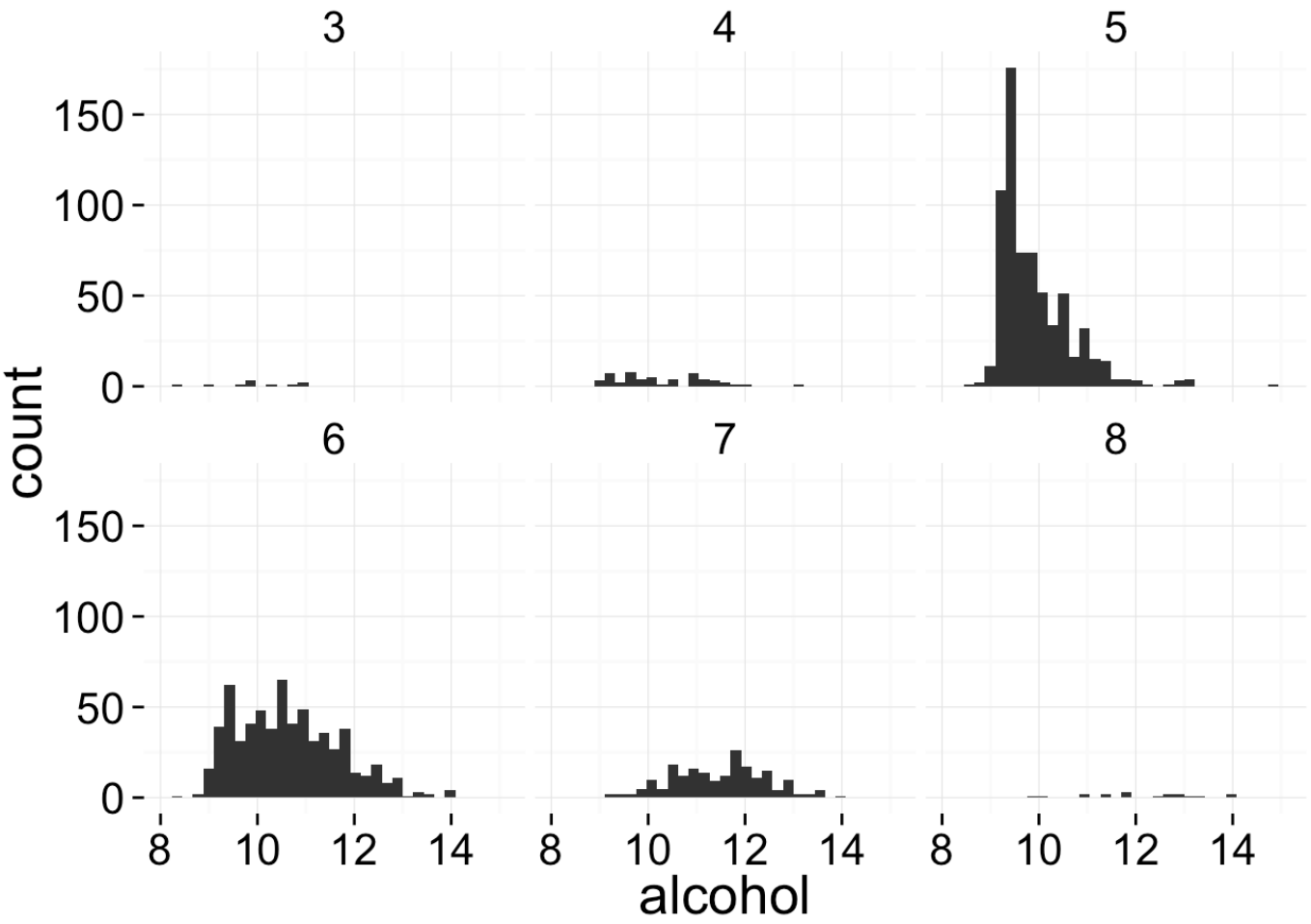
What was the strongest relationship you found?

There does not seem to be any high correlations for any of the variables towards quality. The highest seems to be alcohol content but even that is below 0.5. On the other side, the volatile acidity seems to be almost equally negatively correlated with quality.

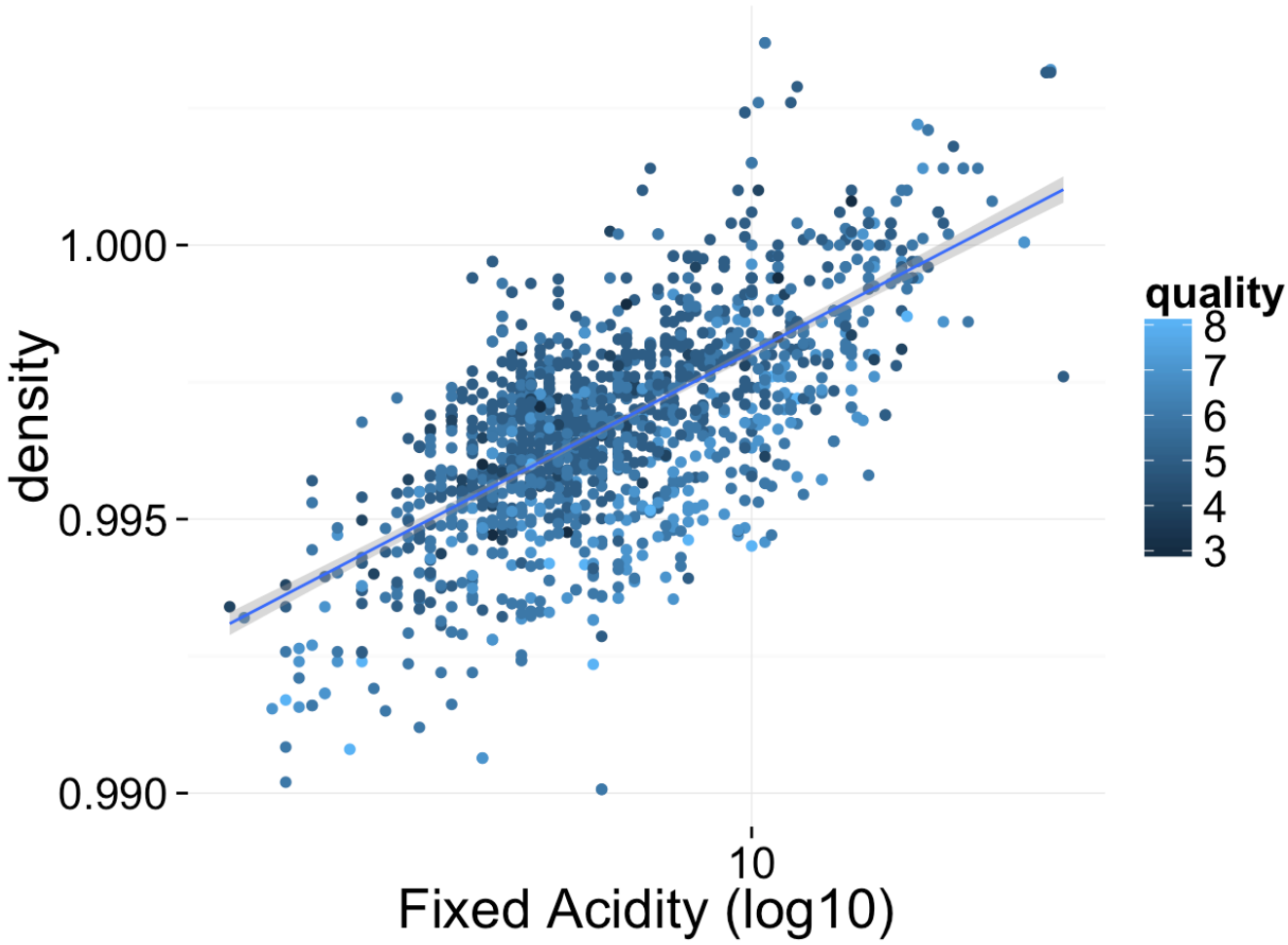
Multivariate Plots Section



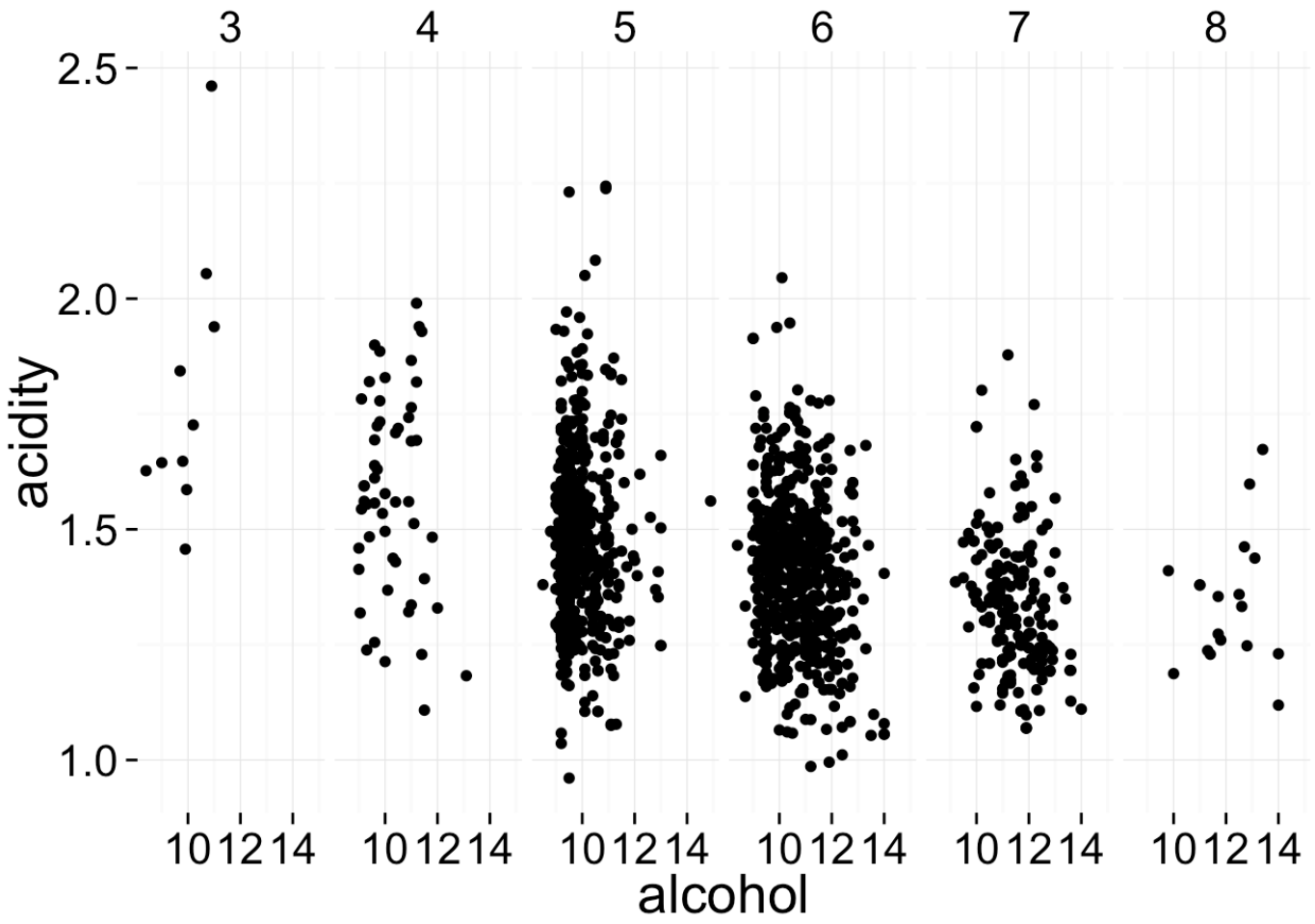
```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

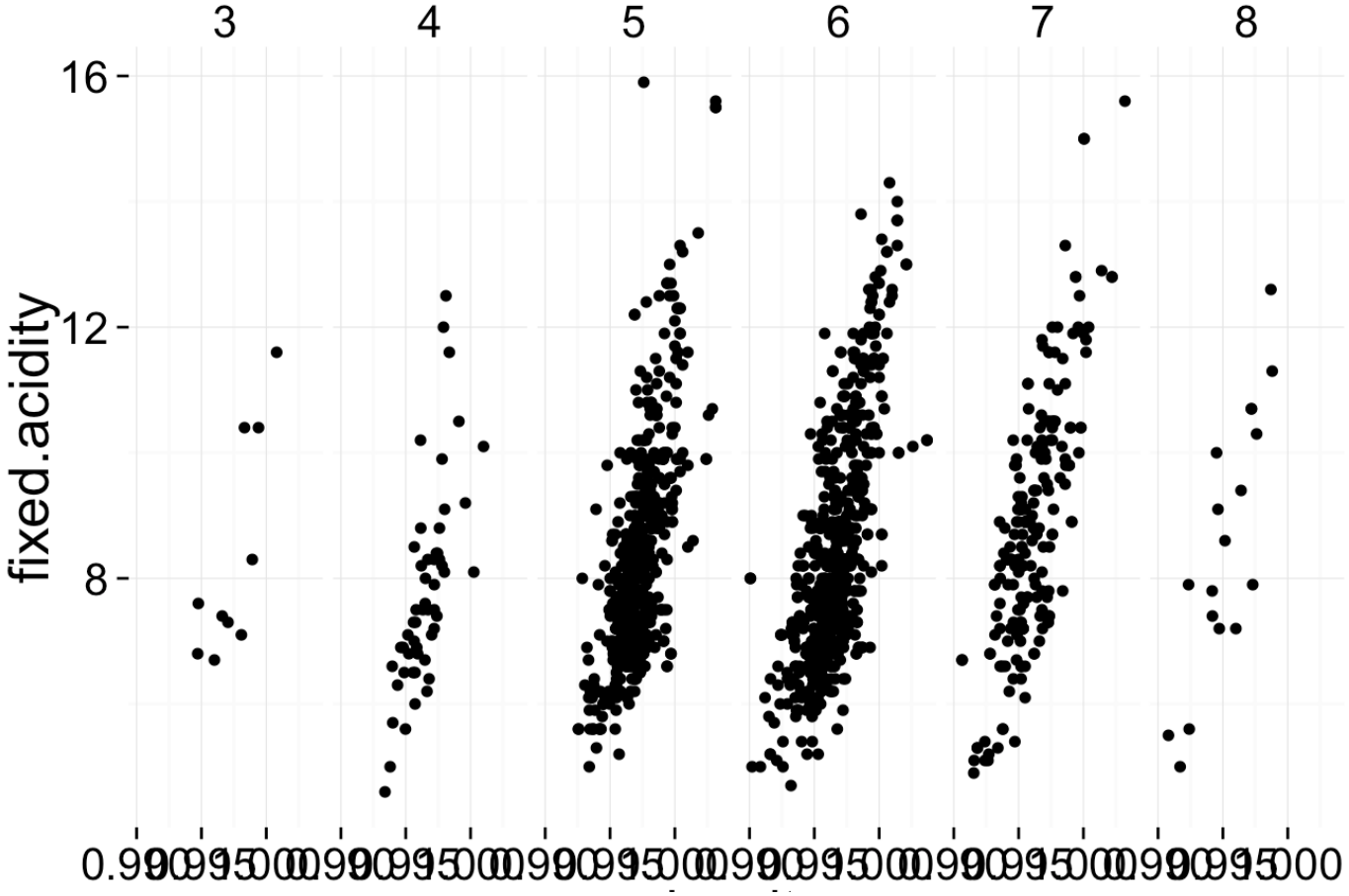


There does not seem to be any strong relationship between acidity and alcohol but we confirm again that higher quality corresponds to higher alcohol content.



Taking log10 of Fixed acidity, we see it has a linear relationship with density.

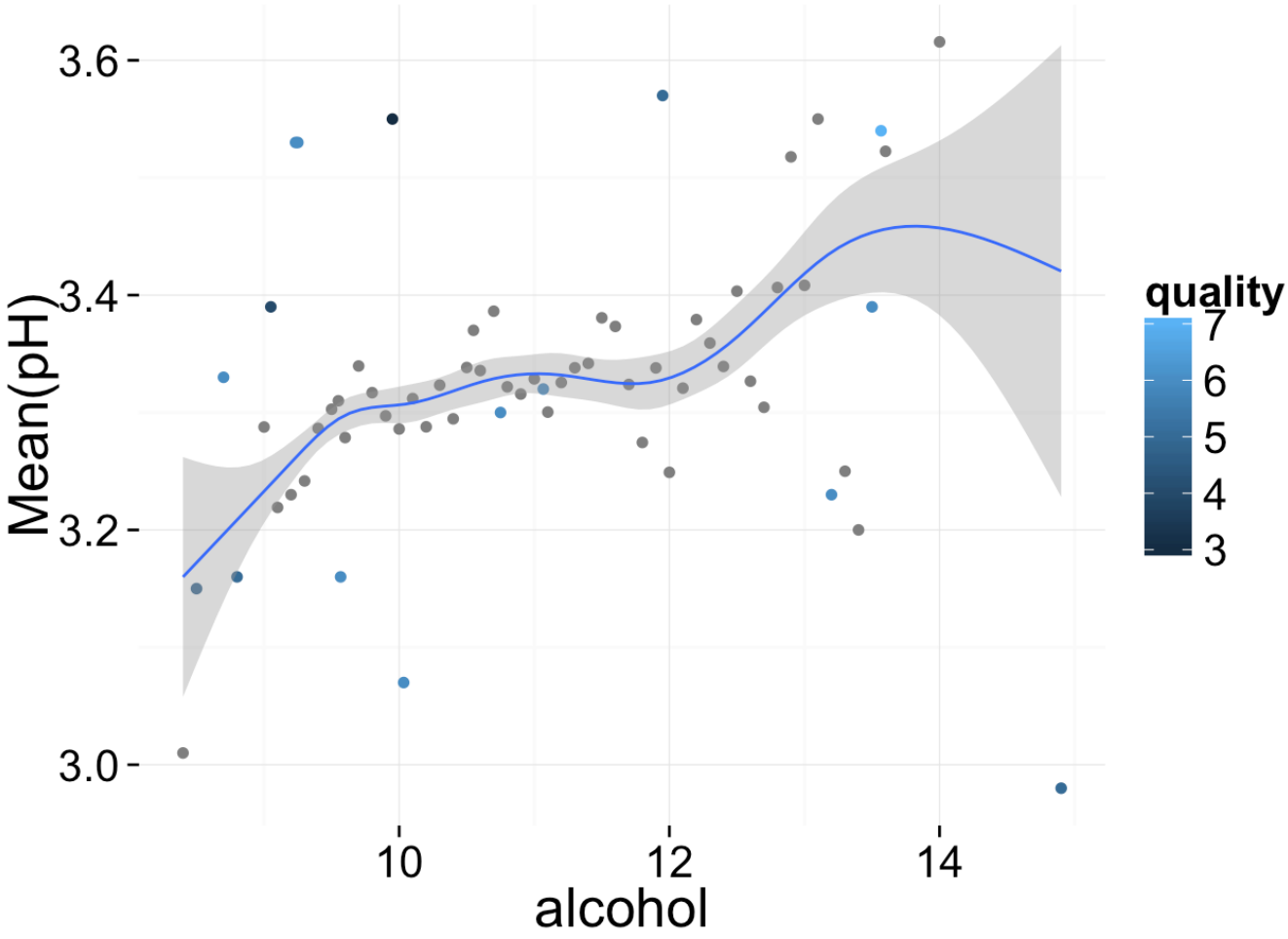

3
4
5
6
7
8
2.5
2.0
1.5
1.0
acidity
10 12 14
alcohol


3
4
5
6
7
8
16
12
8
fixed.acidity

density

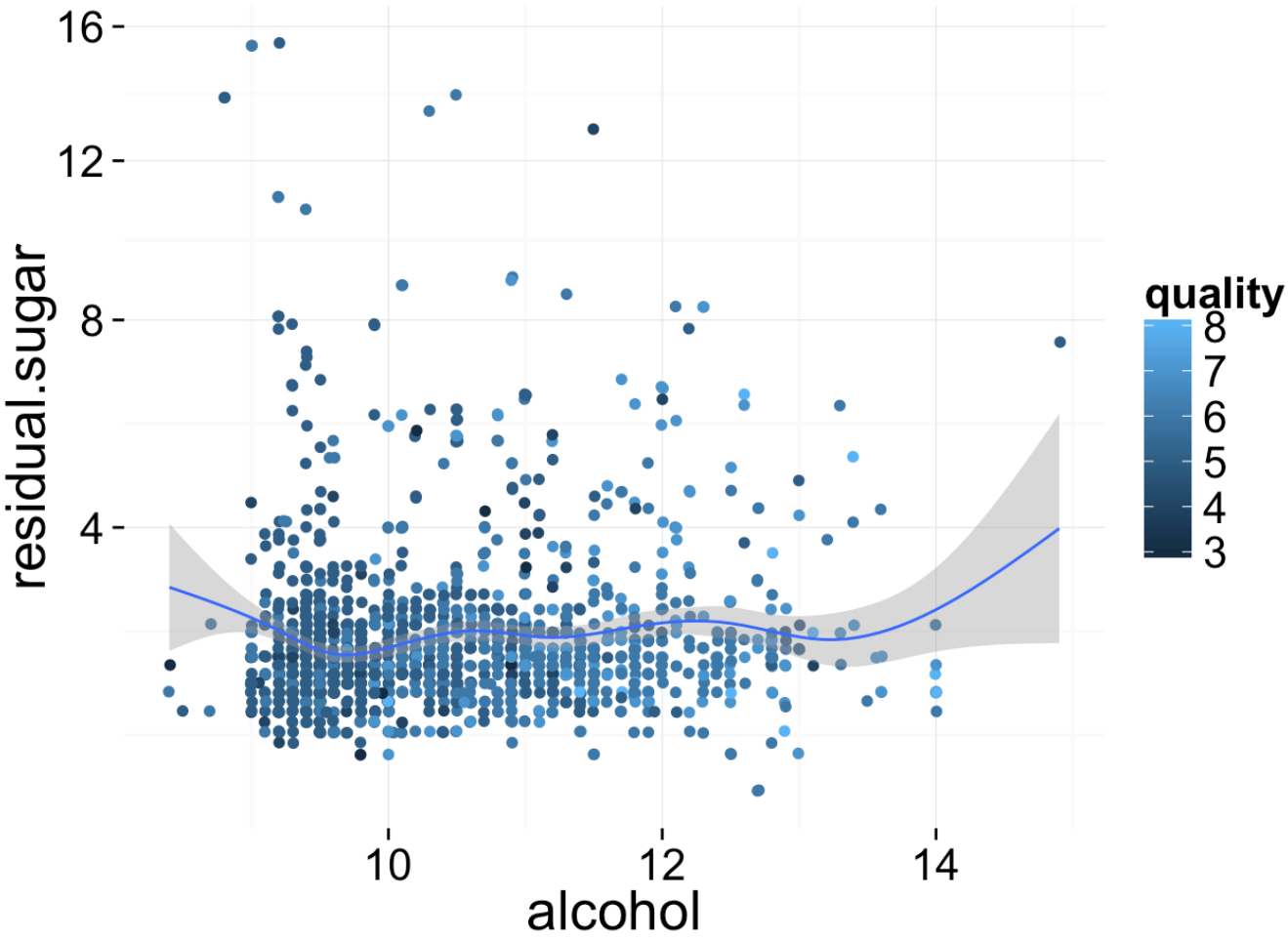
Lower acidity seems to correspond to higher alcohol content and quality. We can also see a somewhat linear trend with lower acidity corresponding to lower density - both of whom correspond to higher quality.

```
## geom_smooth: method="auto" and size of largest group is >=1000, so using gam wi
th formula: y ~ s(x, bs = "cs"). Use 'method = x' to change the smoothing method.
```

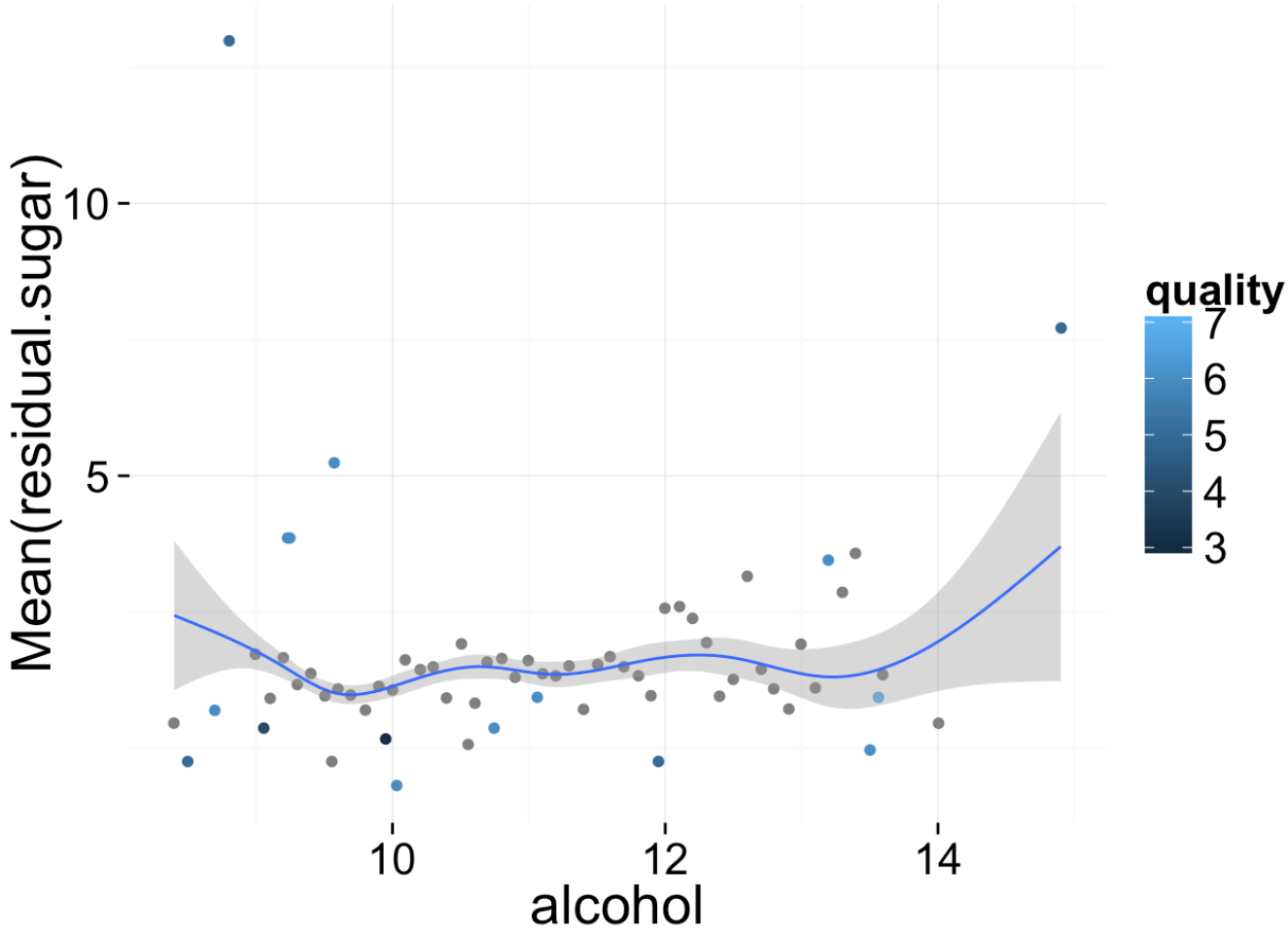


There is a trend towards lower pH as quality increases, however there is no clear relationship with alcohol.

```
## geom_smooth: method="auto" and size of largest group is >=1000, so using gam wi
th formula: y ~ s(x, bs = "cs"). Use 'method = x' to change the smoothing method.
```




```
## geom_smooth: method="auto" and size of largest group is >=1000, so using gam wi
th formula: y ~ s(x, bs = "cs"). Use 'method = x' to change the smoothing method.
```



High quality wines tend to have lower residual sugar. This becomes clearer in the second plot where we look at just the mean.

```
##
## Calls:
## m1: lm(formula = alcohol ~ quality, data = rw)
## m2: lm(formula = alcohol ~ quality + sulphates, data = rw)
## m3: lm(formula = alcohol ~ quality + sulphates + fixed.acidity, data = rw)
## m4: lm(formula = alcohol ~ quality + sulphates + fixed.acidity +
##     volatile.acidity, data = rw)
## m5: lm(formula = alcohol ~ quality + sulphates + fixed.acidity +
##     volatile.acidity + log_sulf, data = rw)
## m6: lm(formula = alcohol ~ quality + sulphates + fixed.acidity +
##     volatile.acidity + log_sulf + pH, data = rw)
## m7: lm(formula = alcohol ~ quality + sulphates + fixed.acidity +
##     volatile.acidity + log_sulf + pH + citric.acid, data = rw)
## m8: lm(formula = alcohol ~ quality + sulphates + fixed.acidity +
##     volatile.acidity + log_sulf + pH + citric.acid + density,
##     data = rw)
## m9: lm(formula = alcohol ~ quality + sulphates + fixed.acidity +
##     volatile.acidity + log_sulf + pH + citric.acid + density +
##     chlorides, data = rw)
## m10: lm(formula = alcohol ~ quality + sulphates + fixed.acidity +
```

```
##     volatile.acidity + log_sulf + pH + citric.acid + density +
##     chlorides + residual.sugar, data = rw)
##
##
================================================================================
================================================
##                         m1           m2           m3           m4           m5
m6          m7          m8          m9          m10
## ------------------------------------------------------------------------------
------------------------------------------------------
## (Intercept)          6.882***    6.945***    7.406***    7.797***    8.960***
1.845*      1.533      447.540***  440.218***  559.323***
##                     (0.165)     (0.173)     (0.191)     (0.260)     (0.296)
(0.817)     (0.806)     (13.739)    (13.926)    (13.459)
## quality              0.628***    0.638***    0.651***    0.627***    0.585***
0.565***    0.566***     0.326***    0.316***    0.235***
##                     (0.029)     (0.030)     (0.030)     (0.032)     (0.032)
(0.031)     (0.030)      (0.025)     (0.025)     (0.022)
## sulphates                        -0.175      -0.052      -0.101       0.049
0.156       0.004        0.592***    0.721***    0.949***
##                                 (0.143)     (0.143)     (0.145)     (0.144)
(0.140)     (0.140)      (0.110)     (0.118)     (0.104)
## fixed.acidity                                -0.074***   -0.081***   -0.093***
0.018      -0.058**      0.399***    0.380***    0.493***
##                                             (0.014)     (0.014)     (0.014)
(0.018)     (0.020)      (0.021)     (0.022)     (0.020)
## volatile.acidity                                        -0.325*     -0.319*
-0.440**     0.118        0.619***    0.698***    0.594***
##                                                         (0.146)     (0.144)
(0.141)     (0.158)      (0.124)     (0.126)     (0.111)
## log_sulf                                                            -0.257***
-0.222***   -0.250***    -0.050      -0.058*     -0.085***
##                                                                     (0.033)
(0.033)     (0.032)      (0.026)     (0.026)     (0.023)
## pH
1.865***    2.010***     3.414***    3.269***    3.630***
##
(0.200)     (0.198)      (0.160)     (0.167)     (0.148)
## citric.acid
1.337***     1.133***    1.234***    0.815***
##
(0.184)      (0.143)     (0.146)     (0.130)
## density
-455.905*** -447.874*** -569.623***
##
(14.029)    (14.254)    (13.774)
## chlorides
-1.293**    -0.902*
##
(0.436)     (0.384)
## residual.sugar
```

```
0.256***
##
(0.012)
## ------------------------------------------------------------------------------
-------------------------------------------------------
## R-squared              0.227       0.227       0.242       0.244       0.271
0.309       0.331       0.598       0.600       0.690
## adj. R-squared         0.226       0.226       0.240       0.242       0.269
0.306       0.328       0.596       0.598       0.688
## sigma                  0.937       0.937       0.929       0.928       0.911
0.888       0.874       0.677       0.676       0.595
## F                    468.267     234.959     169.372     128.575     118.560
118.537     112.450     295.649     265.066     353.741
## p                      0.000       0.000       0.000       0.000       0.000
0.000       0.000       0.000       0.000       0.000
## Log-likelihood     -2164.504   -2163.751   -2148.983   -2146.512   -2117.149
-2074.812   -2048.714   -1641.500   -1637.086   -1433.279
## Deviance            1403.295    1401.974    1376.315    1372.069    1322.591
1254.376    1214.091     729.541     725.524     562.266
## AIC                 4335.007    4335.502    4307.966    4305.024    4248.298
4165.624    4115.428    3303.001    3296.172    2890.557
## BIC                 4351.139    4357.010    4334.851    4337.287    4285.938
4208.641    4163.822    3356.772    3355.320    2955.083
## N                   1599        1599        1599        1599        1599
1599        1599        1599        1599        1599
##
================================================================================
=======================================================
```

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?

We can see that lower acidity corresponds to higher alcohol content and both of these correspond to higher quality in wine. Lower pH and residual sugars are associated with higher quality wines.

Were there any interesting or surprising interactions between features?

Free sulfur dioxide and total sulfur dioxide are highly correlated and have a fairly linear relationship. It leads me to belive that there is overlap in terms of what they are measuring.

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

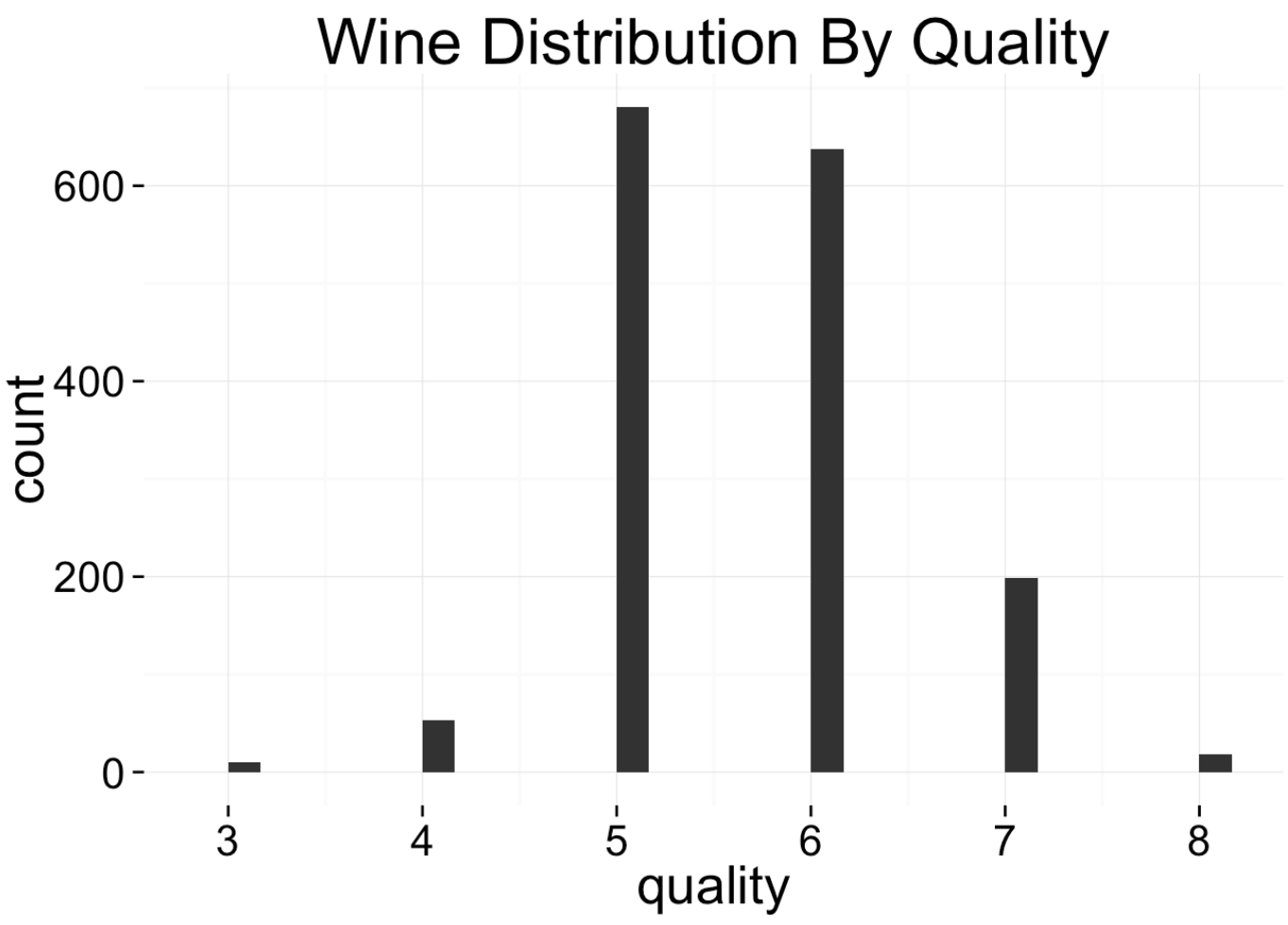
I created a simple linear regression model using all the variables except free sulfur dioxide (and leaving out the derived variable acidity). I left that one out as I felt total sulfur dioxide was highly correlated with it and the R2 value of the model was not significantly impacted because of its omission. I also log tranformed total sulfur dioxide since the log was closer to being normally distributed. I ended up getting around a 69% R2 value which is a reasonable goodness of fit considering there were no variables that were strongly driving the wine quality values.

It is also clear that a linear model is perhaps not the best way to model wine quality since, even though the variables are fairly independent of each other, and are approximately normally distributed, there does not seem to be a linear relationship between wine quality and the variables.

Final Plots and Summary

Plot One

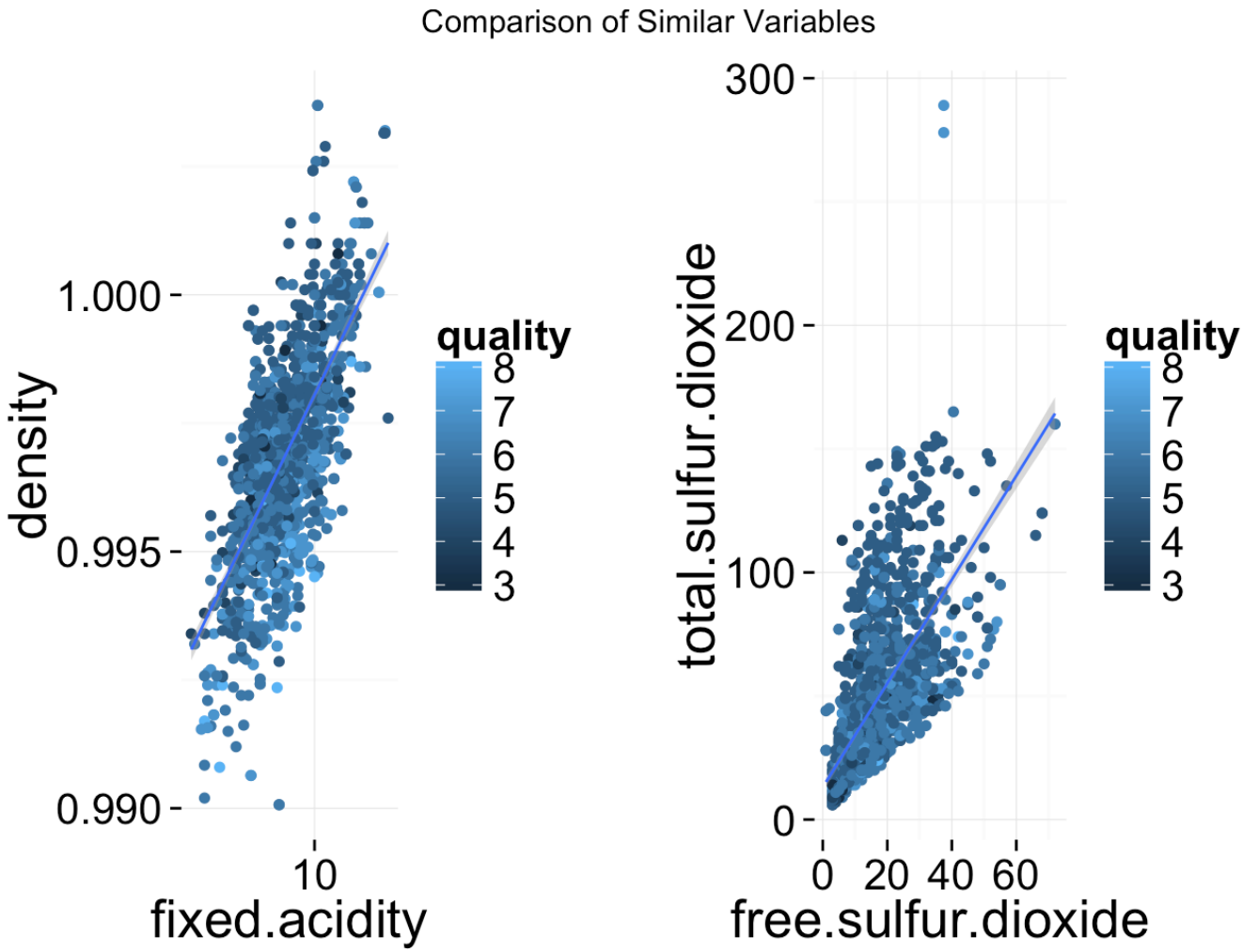
```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



Description One

We can see the majority of the wines have a quality level of 5-6 with very few wines in either extremes (3 or 8).

Plot Two

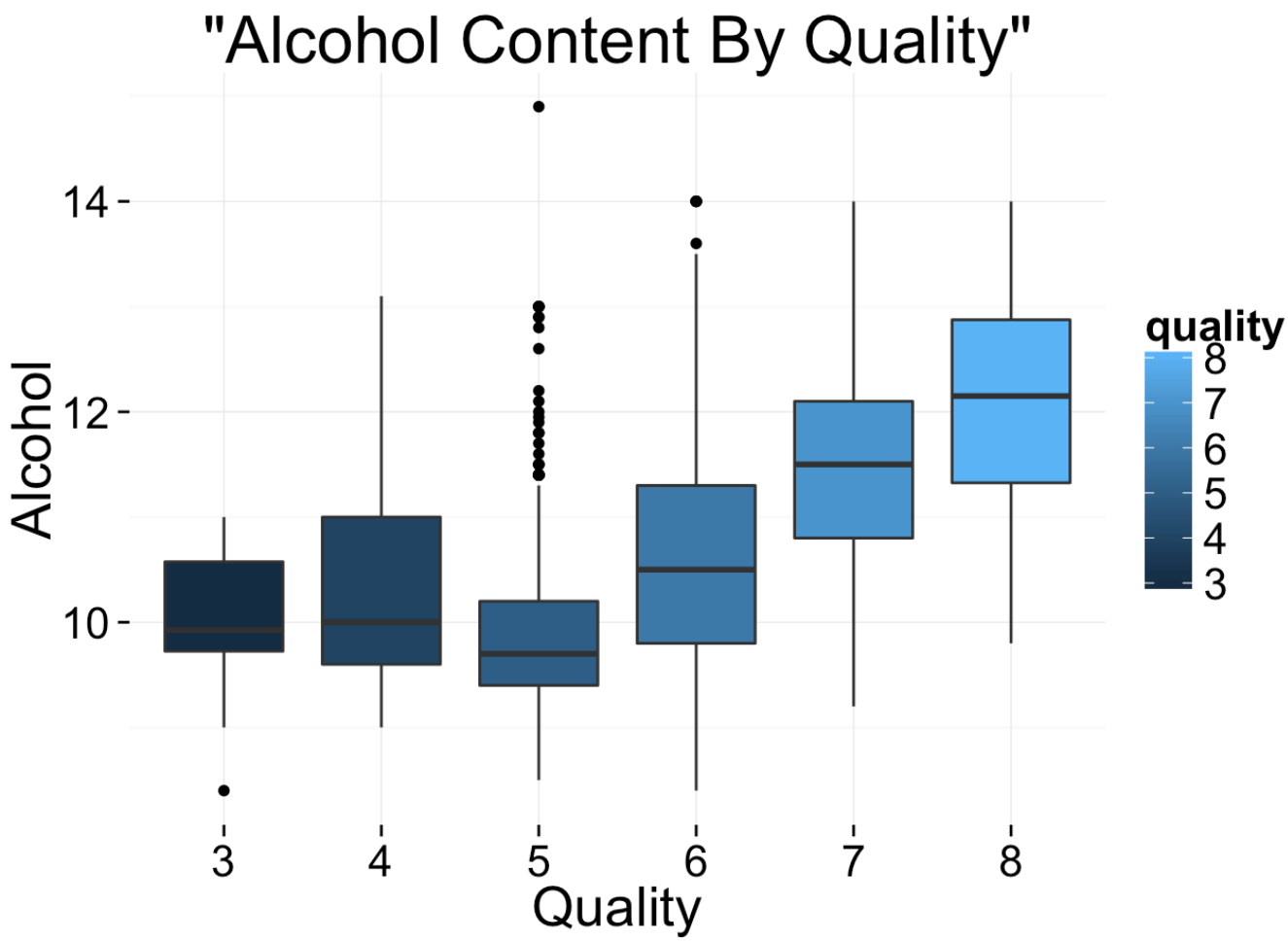


Description Two

I looked closely at the variables that seemed to be correlated to each other and had linear relationships. I used this to exlcude free sulfur dioxide from my model as it was measuring part of the same thing as total sulfur dioxide and was not impacting the goodness of fit of the model significantly.

We can also seee how quality varies with these variables. For the 1st plot, the majority of wines (which are of average quality) are clustered together but we can see the wines of higher quality making up alot of the outliers - particularly towards the bottom left i.e. lower density and acidity. It is harder to deduce any particular trend of wine quality in the second plot because most of the wines are clustered together very tightly sharing a small range of values among them.

Plot Three



Description Three

Higher alcohol content tends to be correlated with higher wine quality. There is a sharp jump in the mean of the alcohol variable from 5 to the higher quality levels 6-8 which probably signifies a threshold manufacturers have in terms of alcohol content when producing higher quality red wines. This is not to say lower quality wines do not have high alcohol content as we can see there are a lot of outliers for quality level 5 (which is also the biggest category in terms of number in our sample).

Reflection

I started my analysis in geting a sense of the data distrbution which was normal for all the variables as well as the wine quality distribution with the majority of the wines being of quality 5-6. I could not pick out any single feature that was a strong driver of wine quality with alcohol amount being the closest followed by fixed acidity in the other direction. Looking at the varaibles and how they were interacting with each other, we could see some of the common traits of high quality red wines (high alcohol content and sulfates, low fixed acidity, pH and residual sugars). With a few exceptions, most of the features were fairly independent of each other and this helped in the linear regression model where we were able to get a 69% goodness fit. For a more thorugh analysis, we can try alternative modelling techniques which are not assuming linearity in the data to see if we can get a better prediciton model.