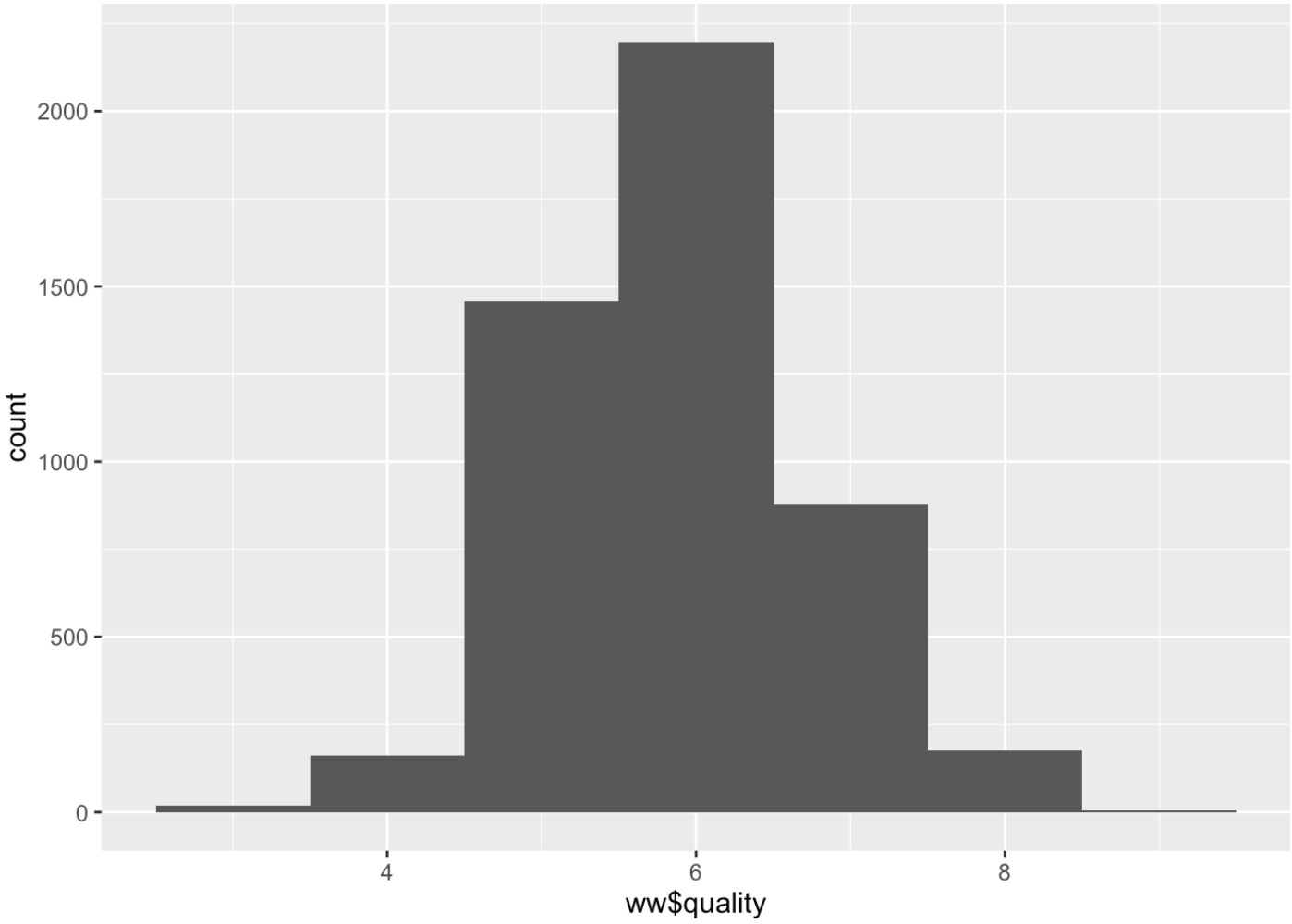
Exploratory Data Analysis and Wine Quality Prediction - by Christian Ramsey

```
##  [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"
```

Univariate Plots Section

```
## [1] 63674
```

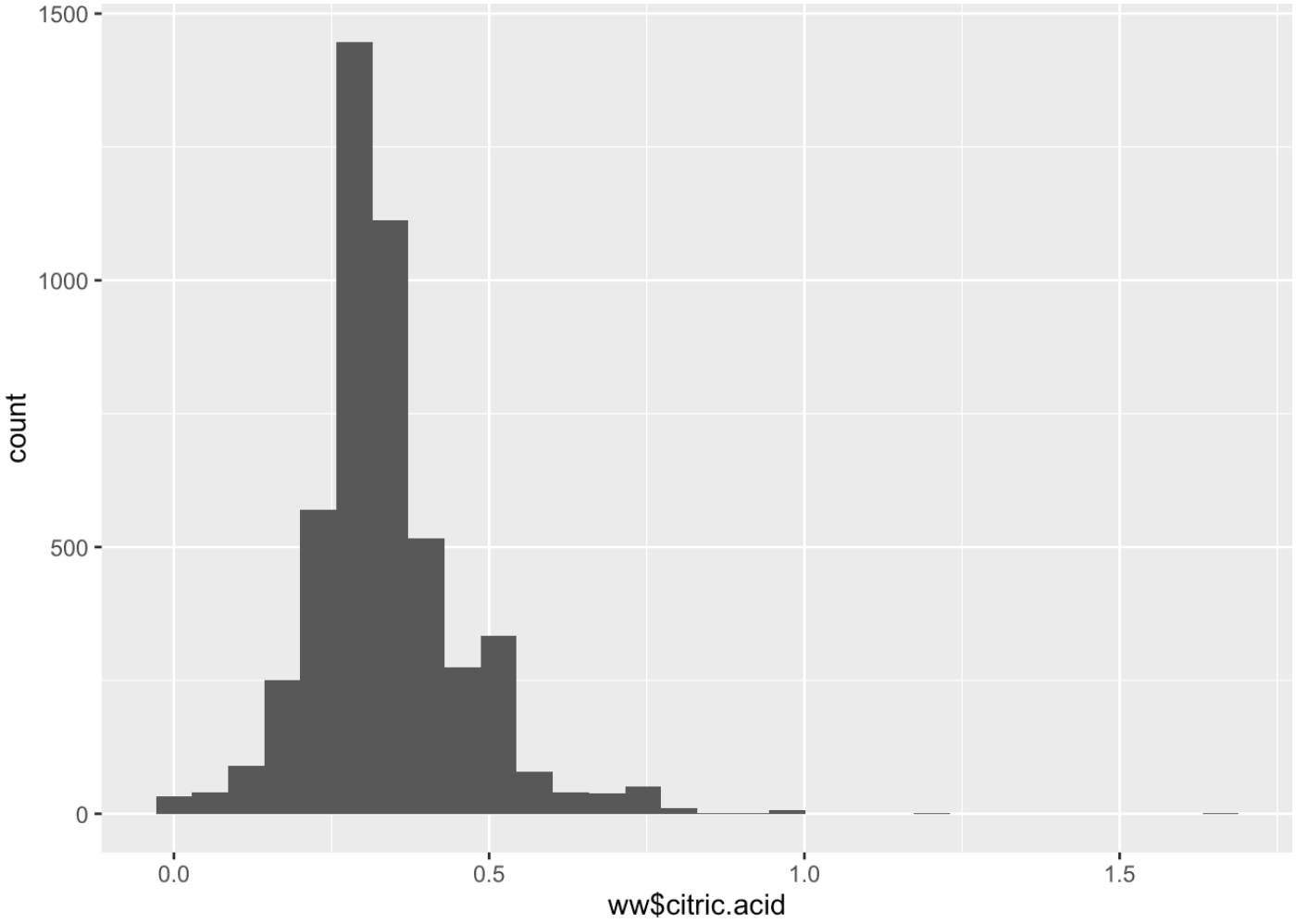
```
## [1] 3 9
```



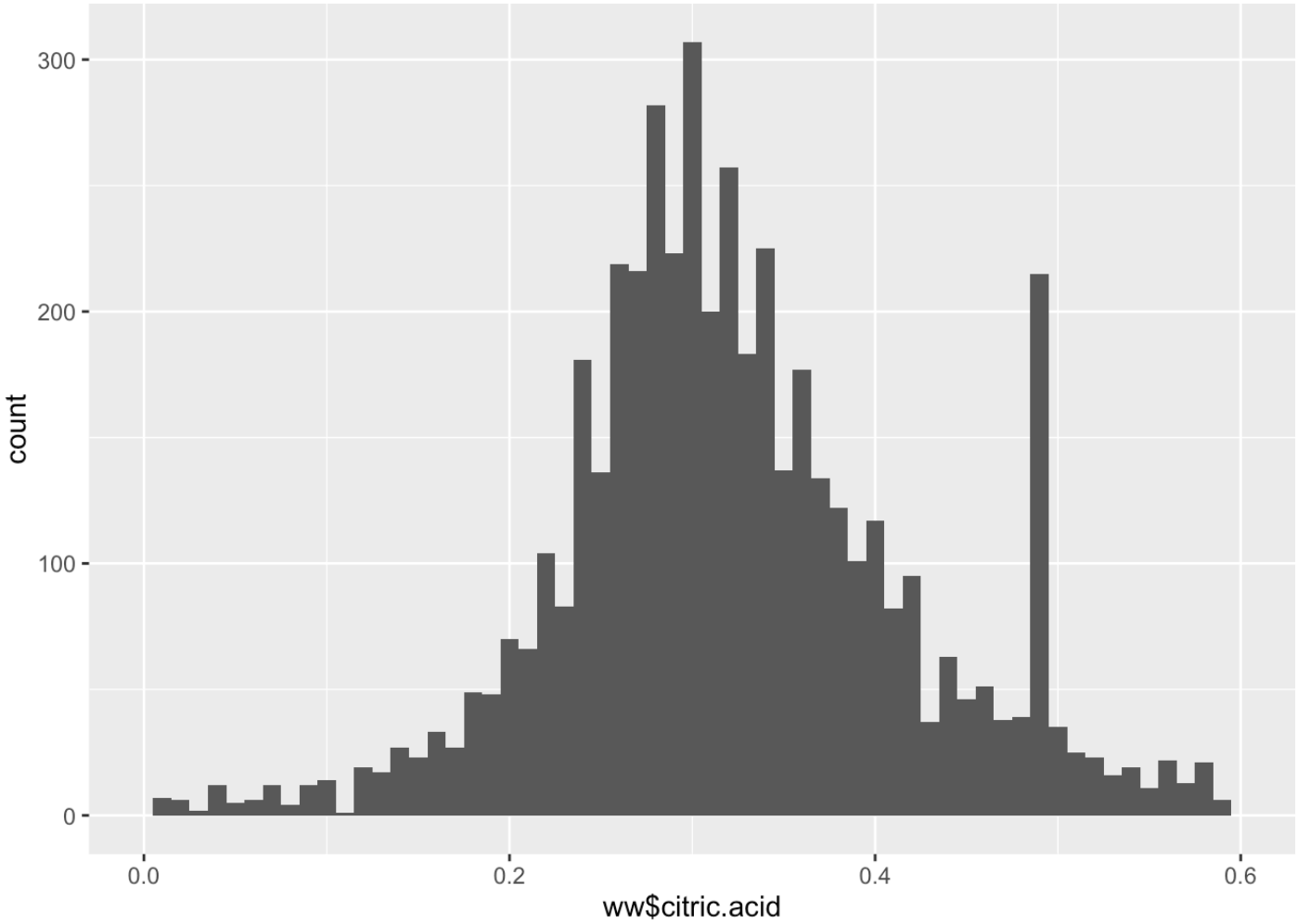

```
## [1] 0.00 1.66
```

```
## [1] 0.36 0.34 0.40 0.32 0.32 0.40
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

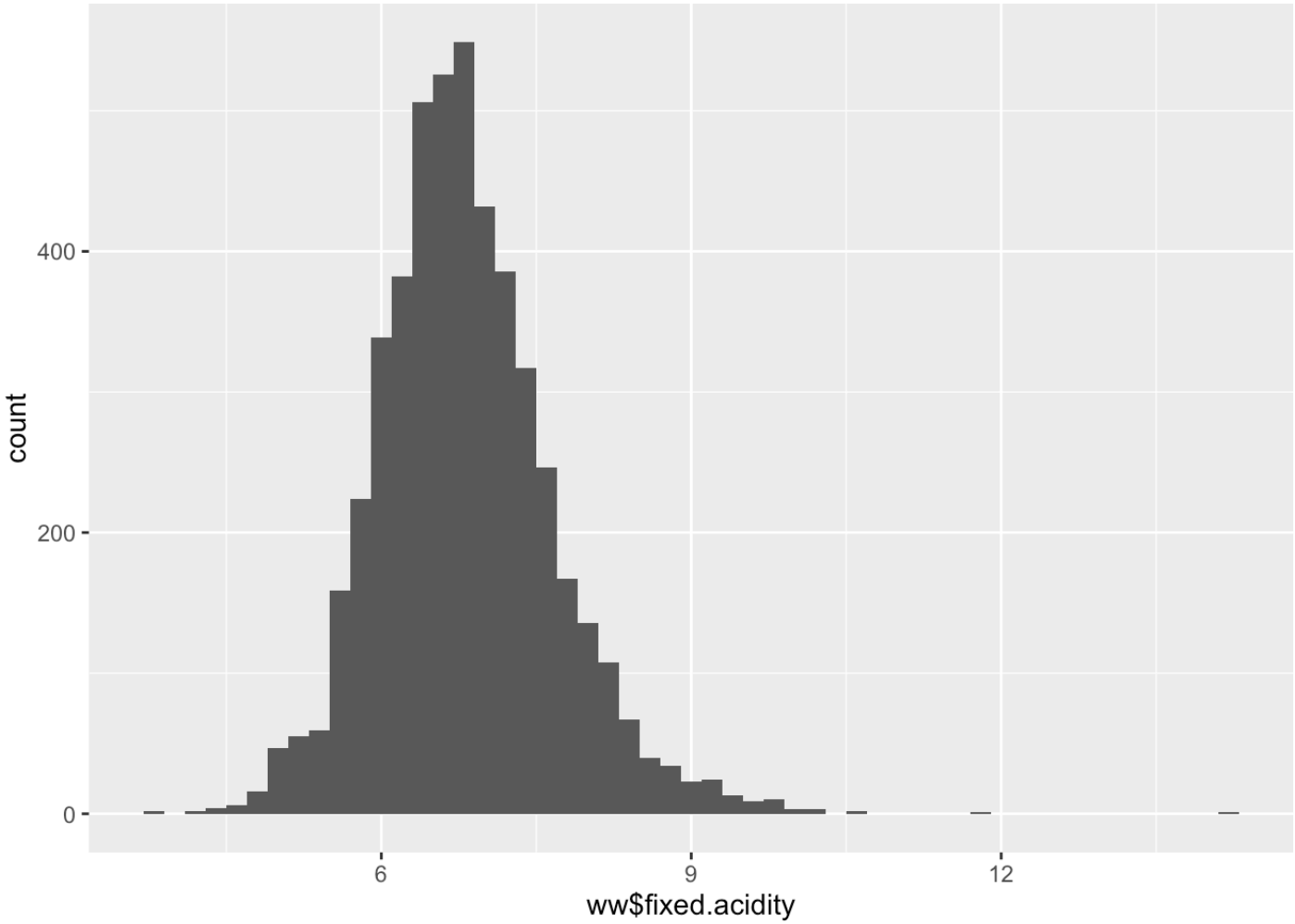



```
## Warning: Removed 152 rows containing non-finite values (stat_bin).
```

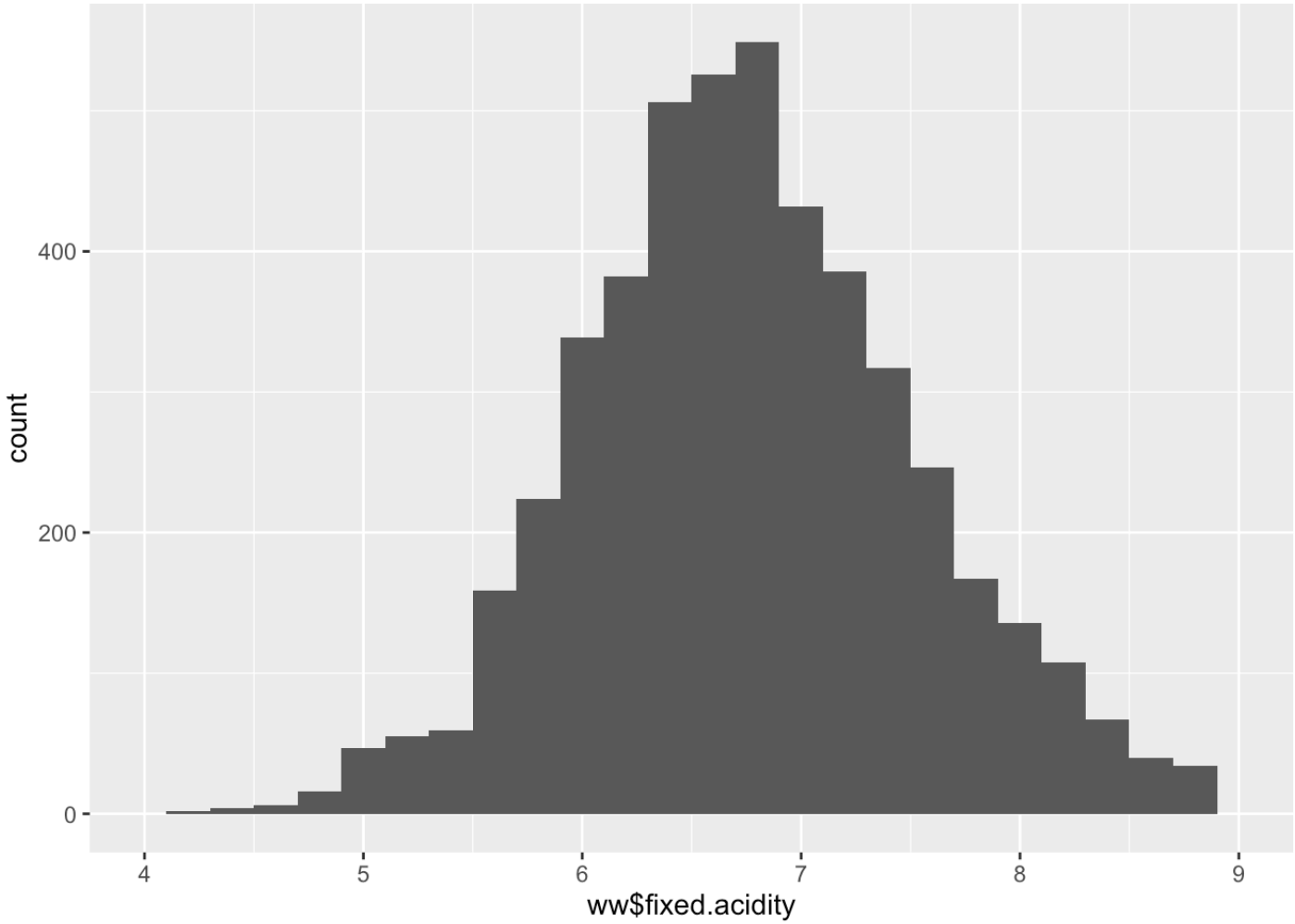



```
## [1]  3.8 14.2
```

```
## [1] 7.0 6.3 8.1 7.2 7.2 8.1
```



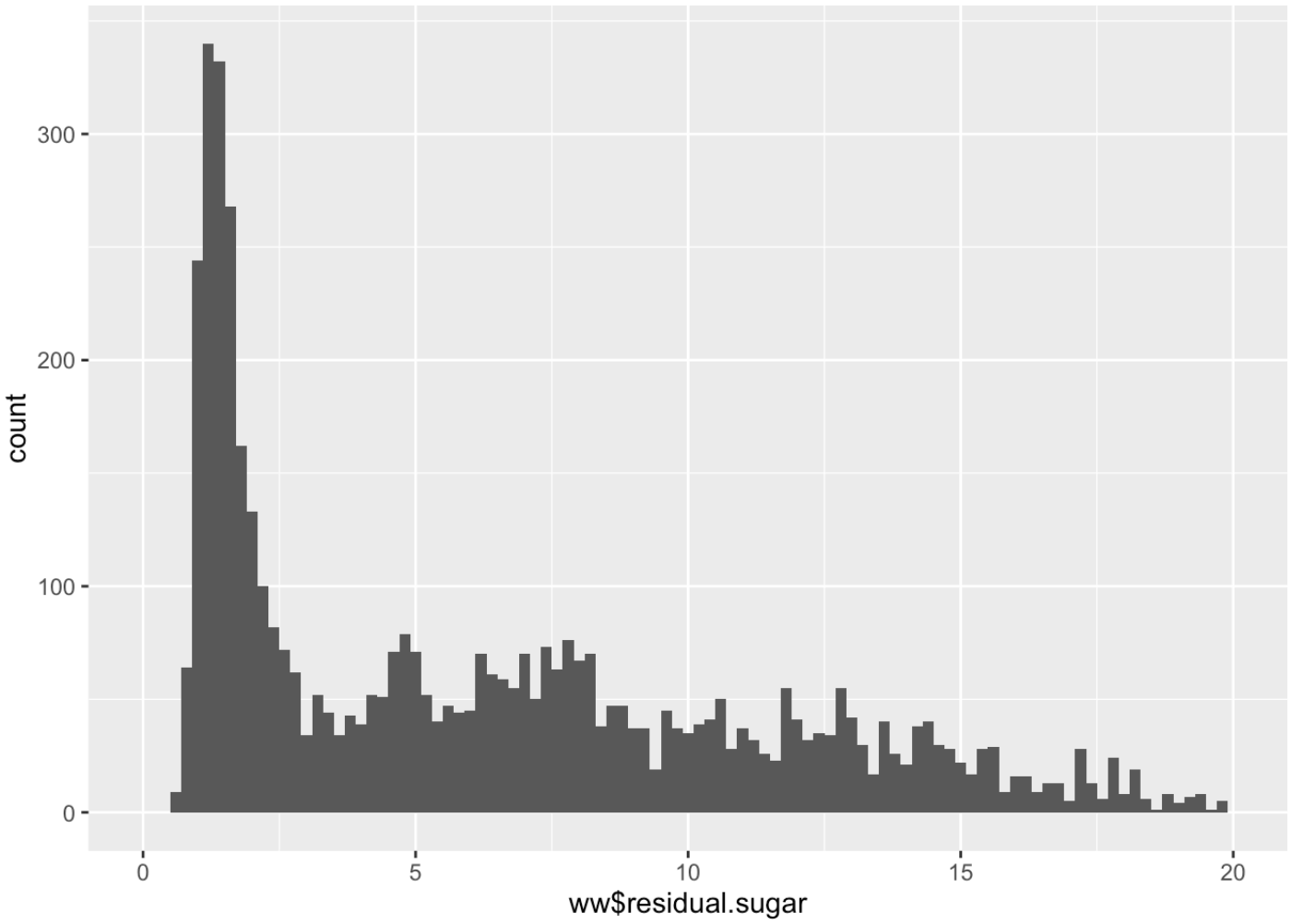

```
## Warning: Removed 74 rows containing non-finite values (stat_bin).
```

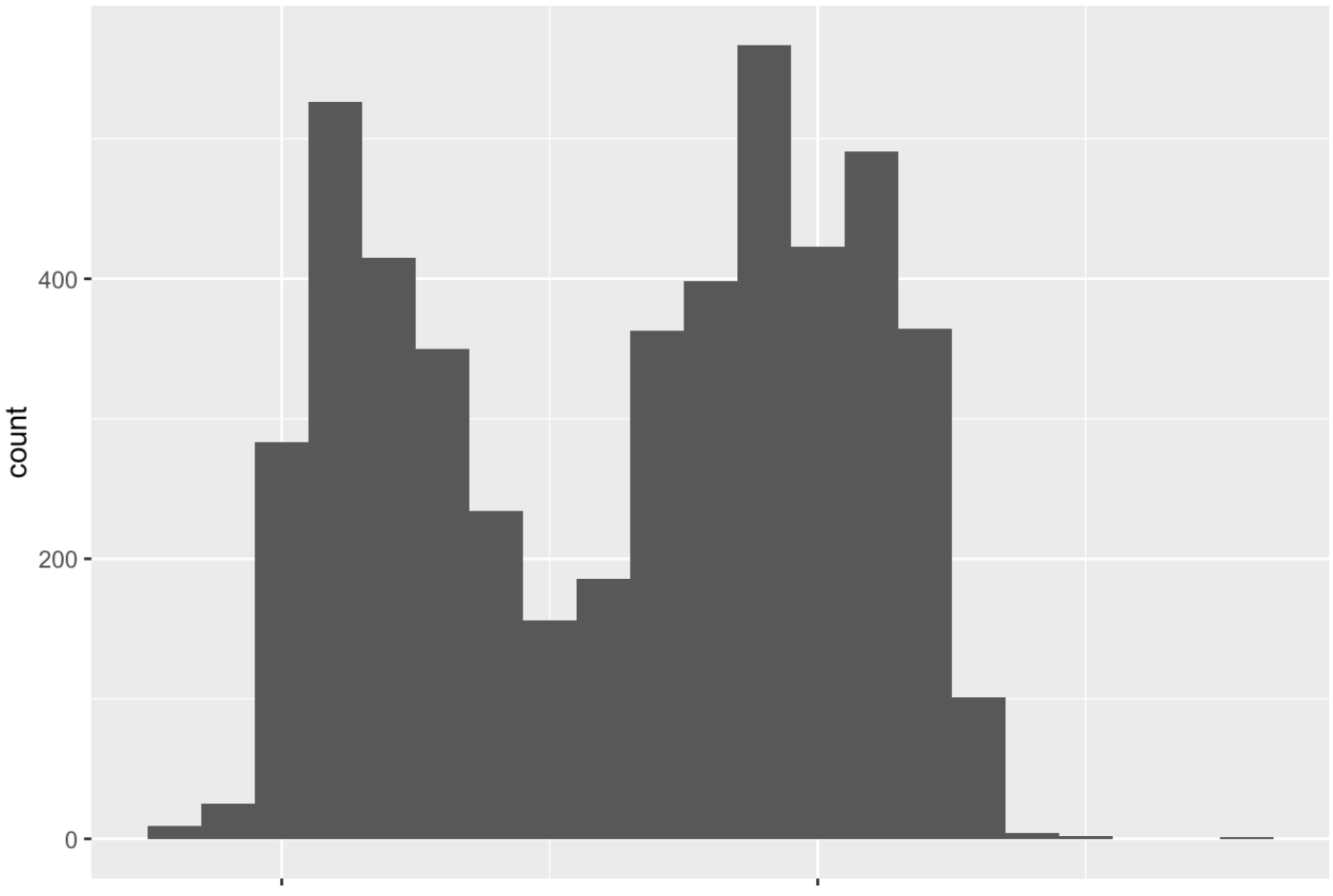



```
## [1]  0.6 65.8
```

```
## [1] 20.7  1.6  6.9  8.5  8.5  6.9
```

```
## Warning: Removed 18 rows containing non-finite values (stat_bin).
```

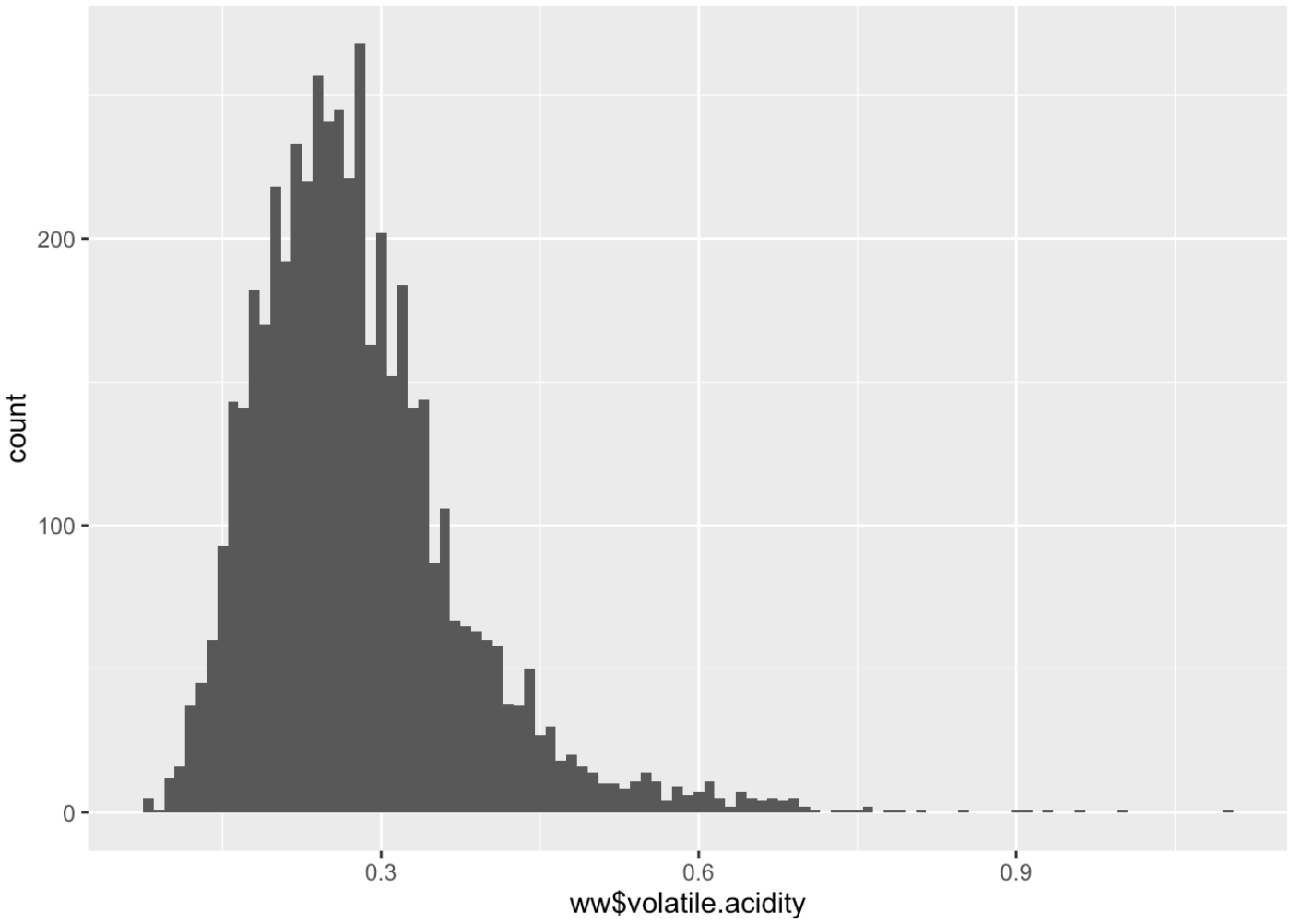

300
200
100
0
count
0
5
10
15
20
ww$residual.sugar

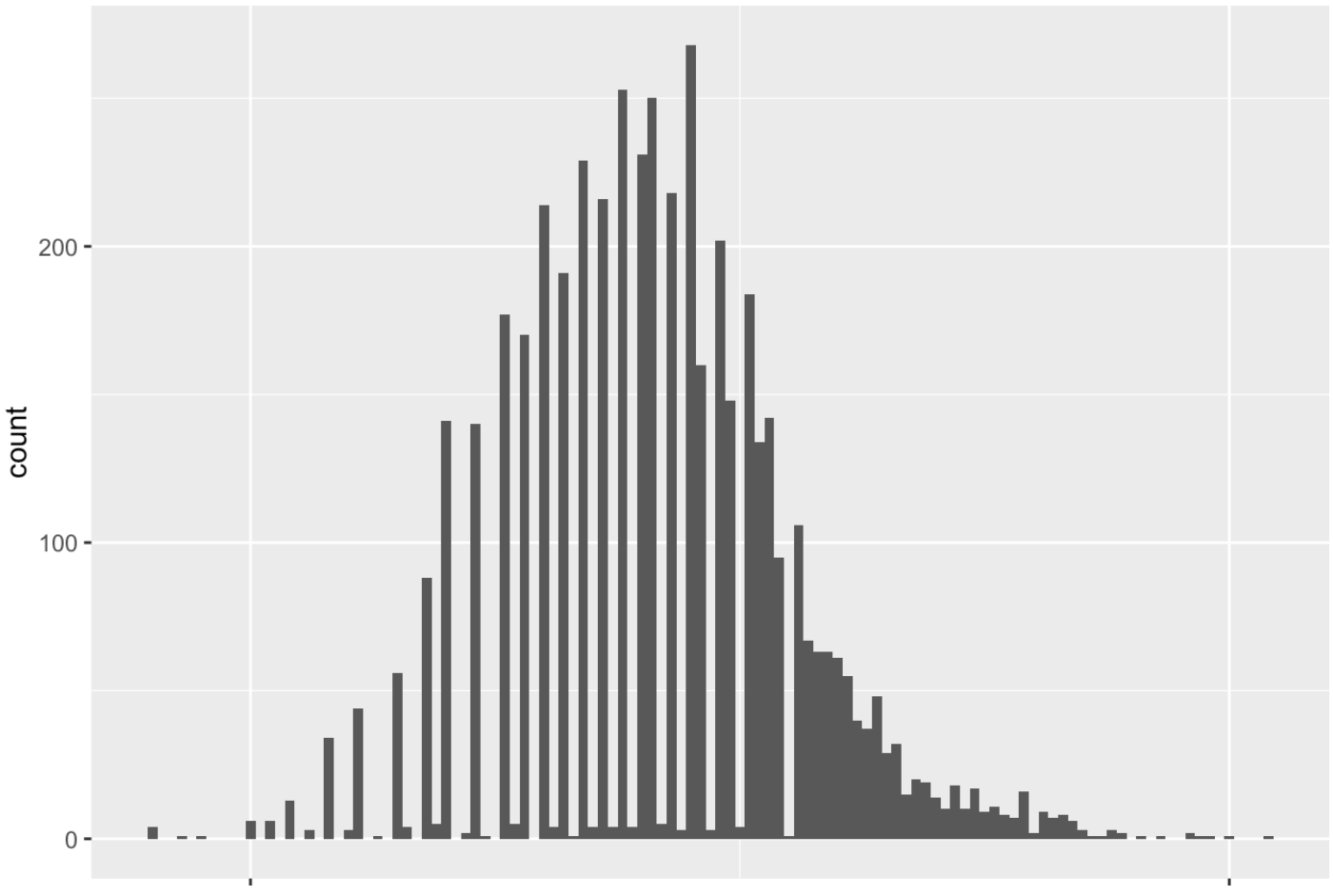

400
200
0
count

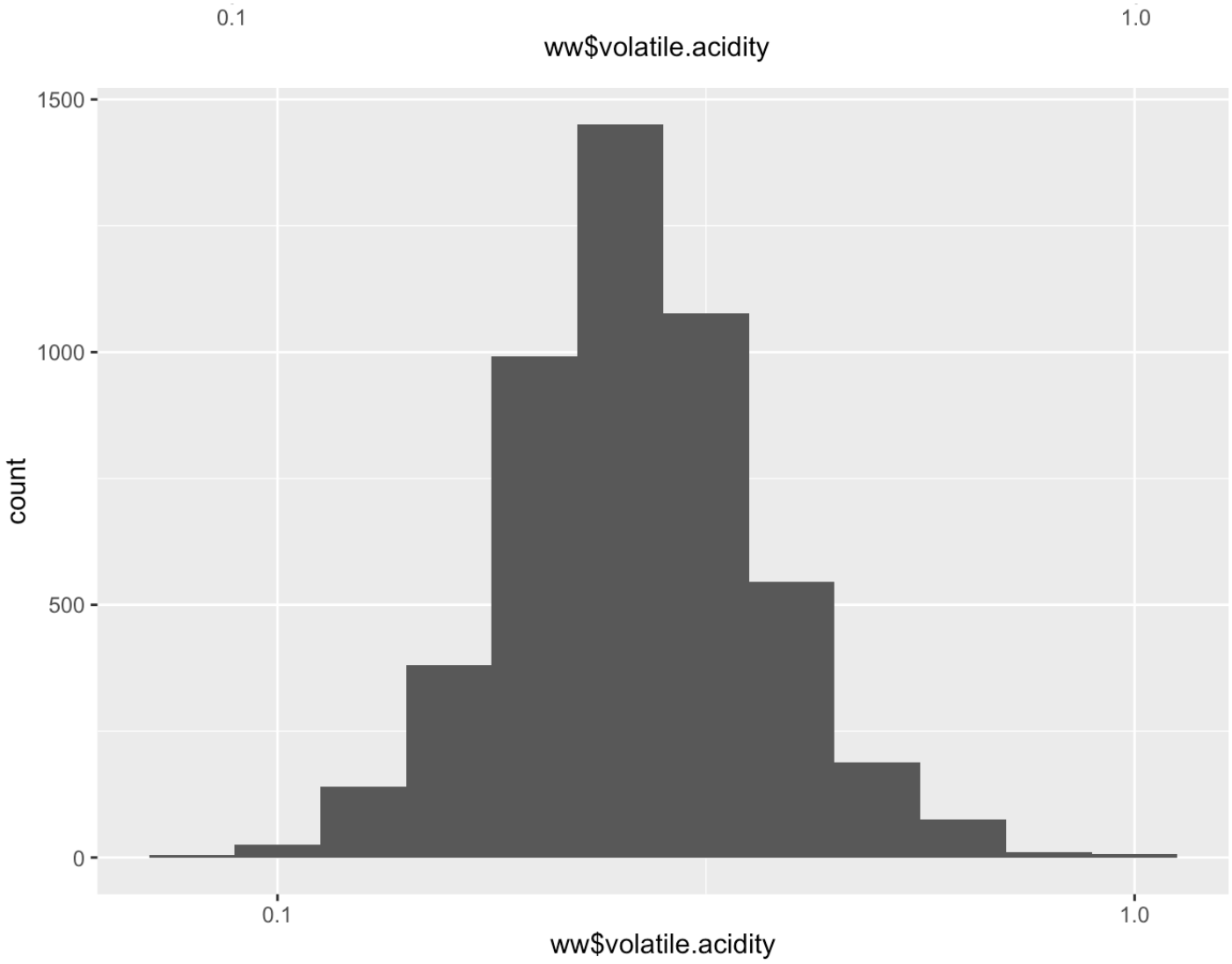

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.600   1.700   5.200   6.391   9.900  65.800
```

```
## [1] 0.08 1.10
```

```
## [1] 0.27 0.30 0.28 0.23 0.23 0.28
```


count
200
100
0
0.3
0.6
0.9
ww$volatile.acidity

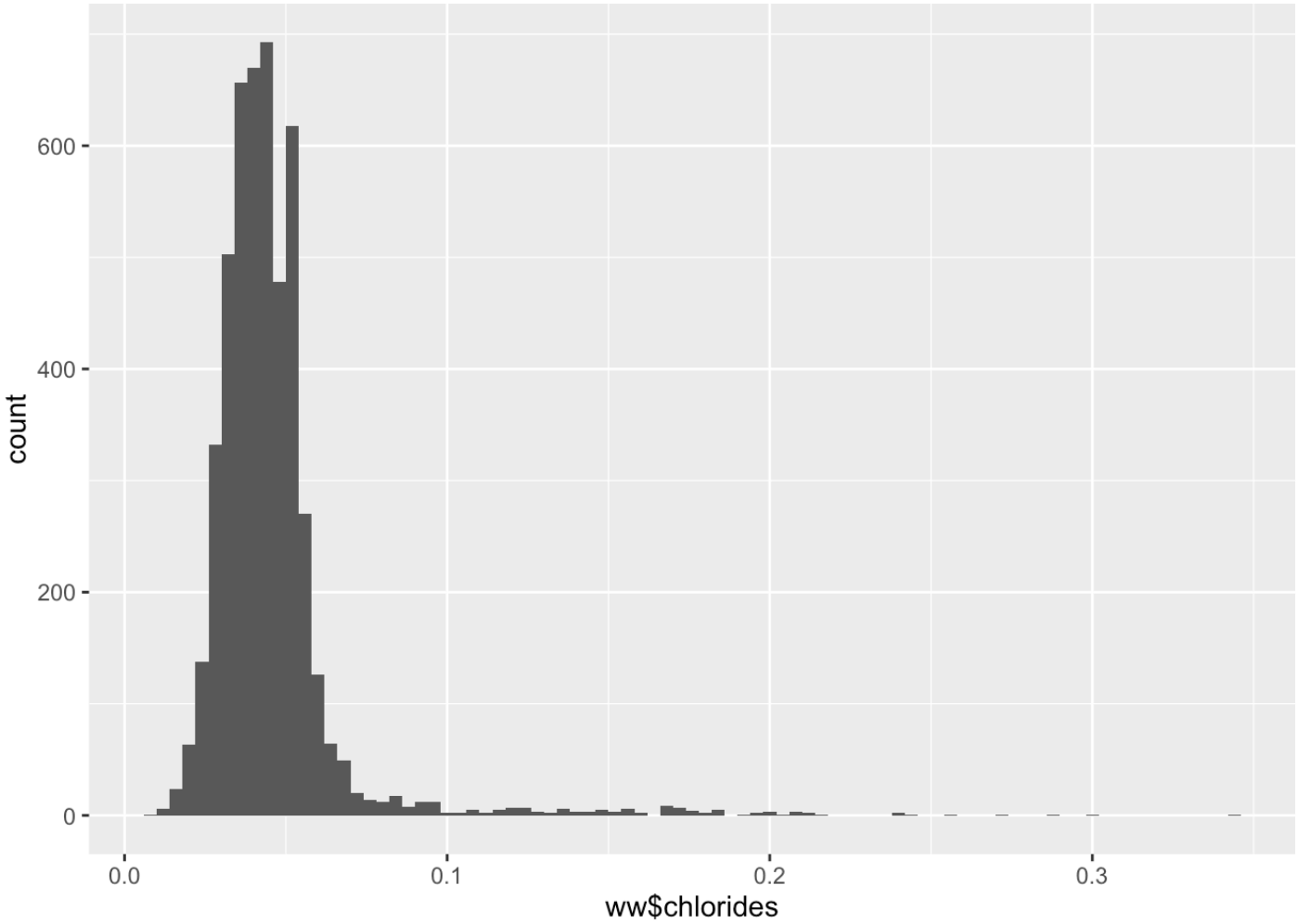

count
200
100
0




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

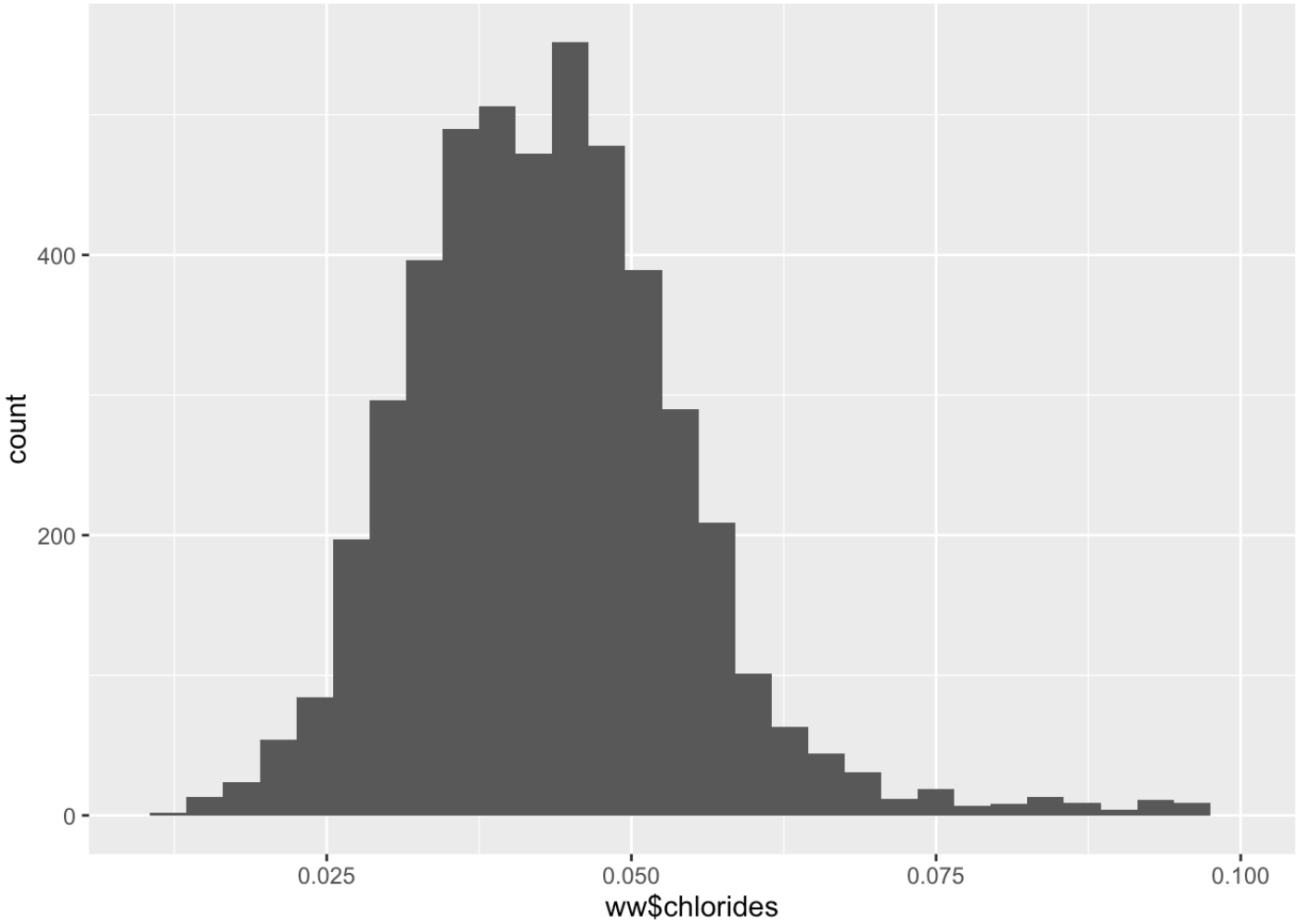
```
## [1] 0.009 0.346
```

```
## [1] 0.045 0.049 0.050 0.058 0.058 0.050
```




```
## Warning: Removed 111 rows containing non-finite values (stat_bin).
```

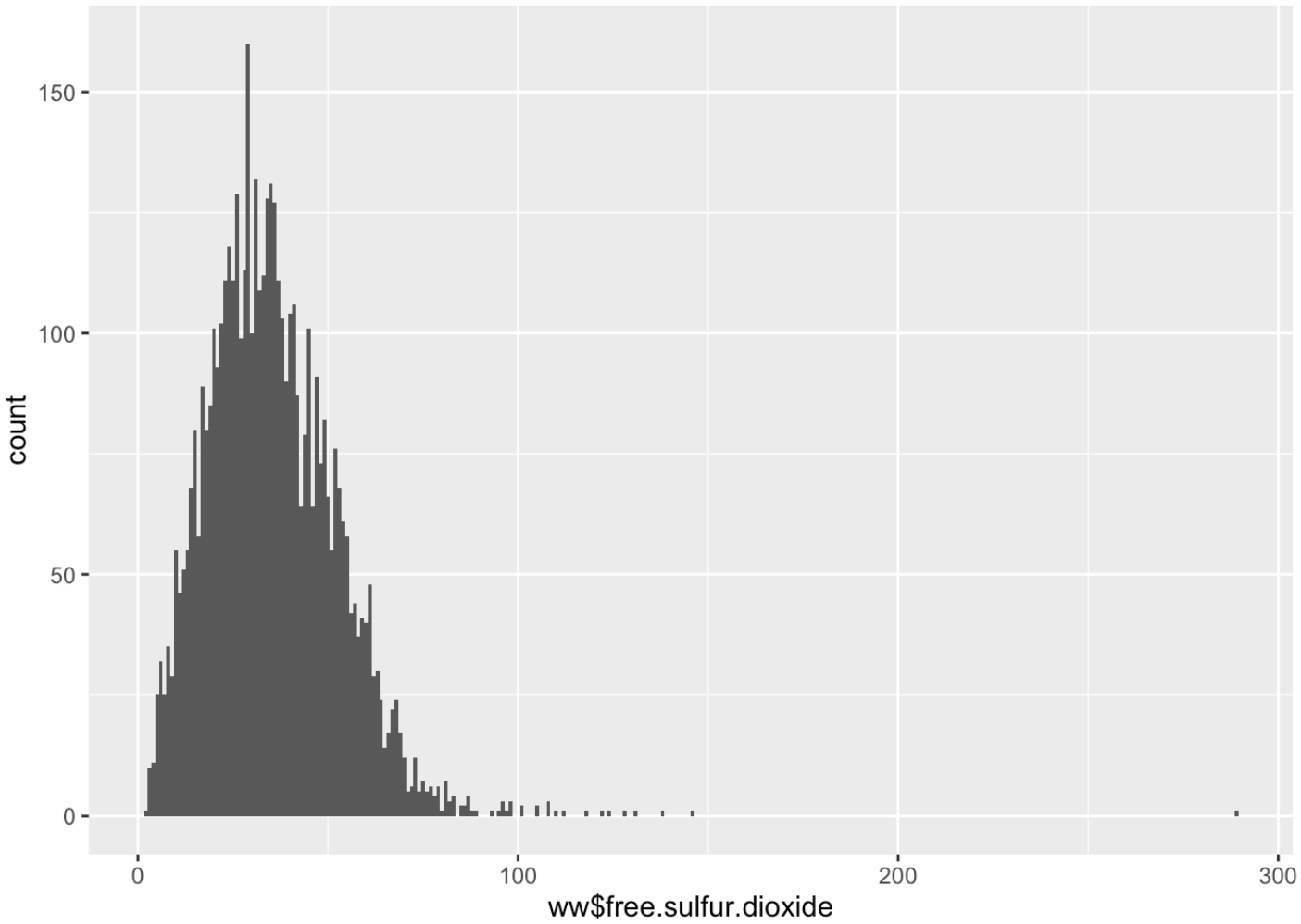
```
## Warning: Removed 1 rows containing missing values (geom_bar).
```



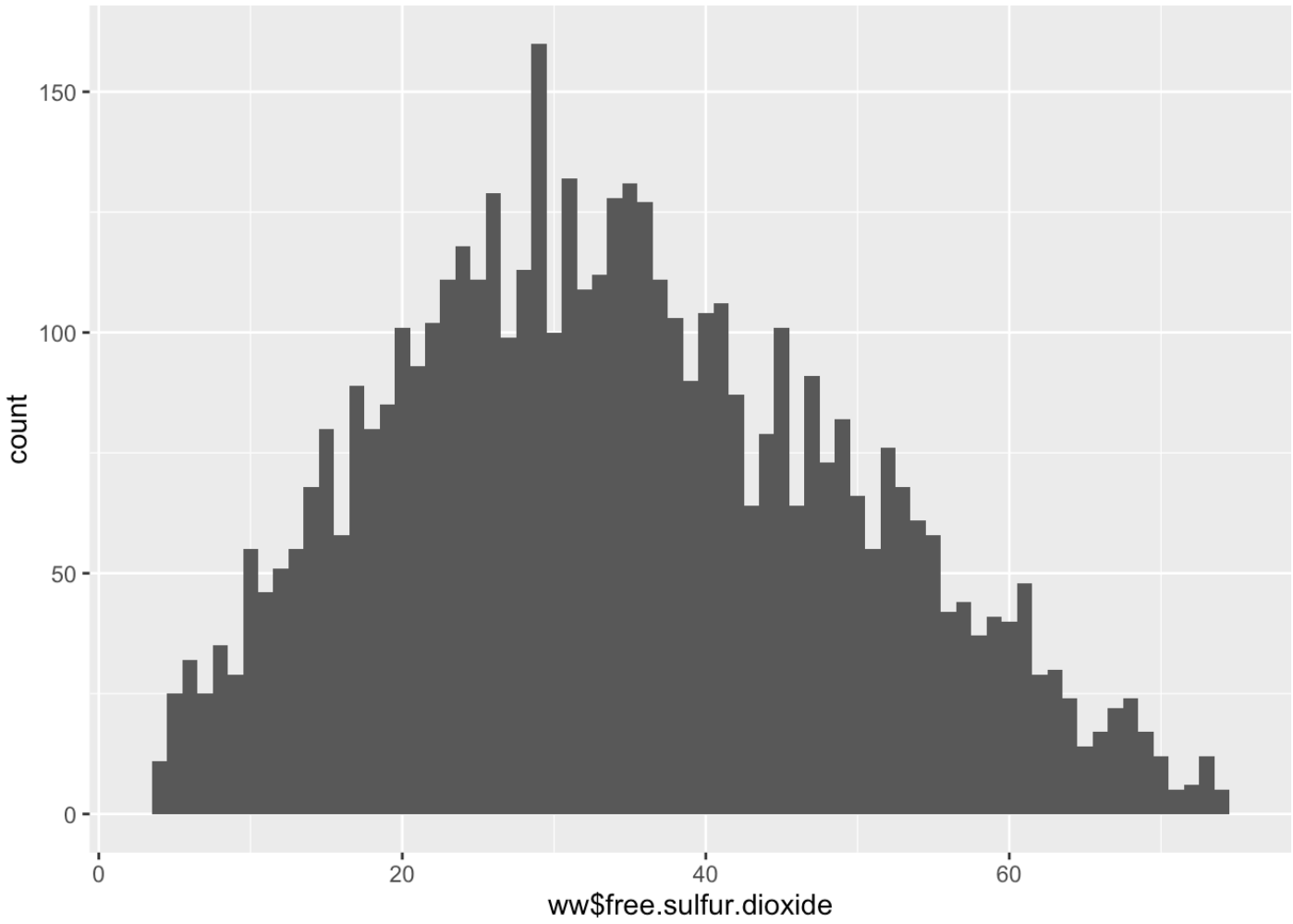

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00900 0.03600 0.04300 0.04577 0.05000 0.34600
```

```
## [1]   2 289
```

```
## [1] 45 14 30 47 47 30
```



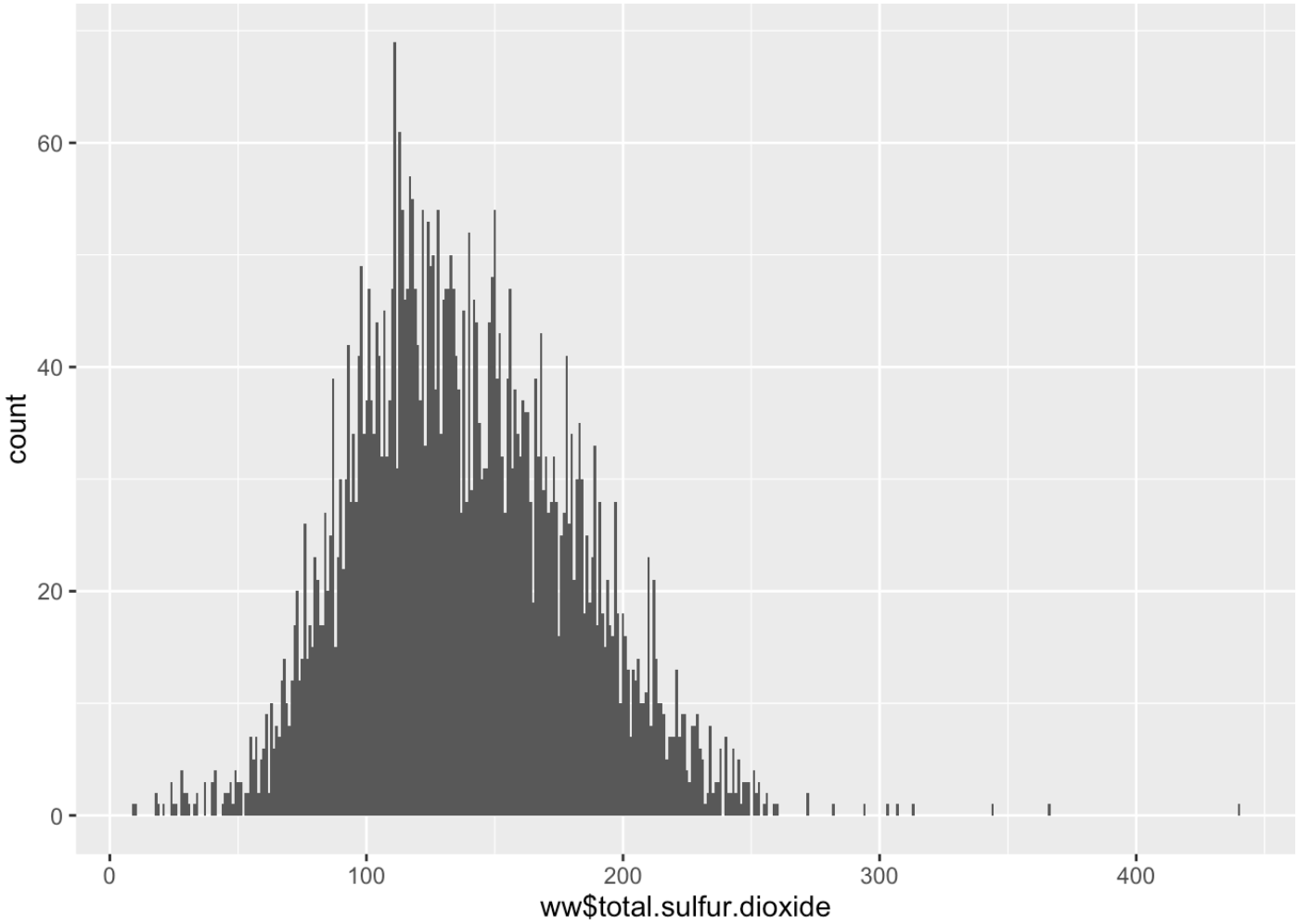

```
## Warning: Removed 73 rows containing non-finite values (stat_bin).
```



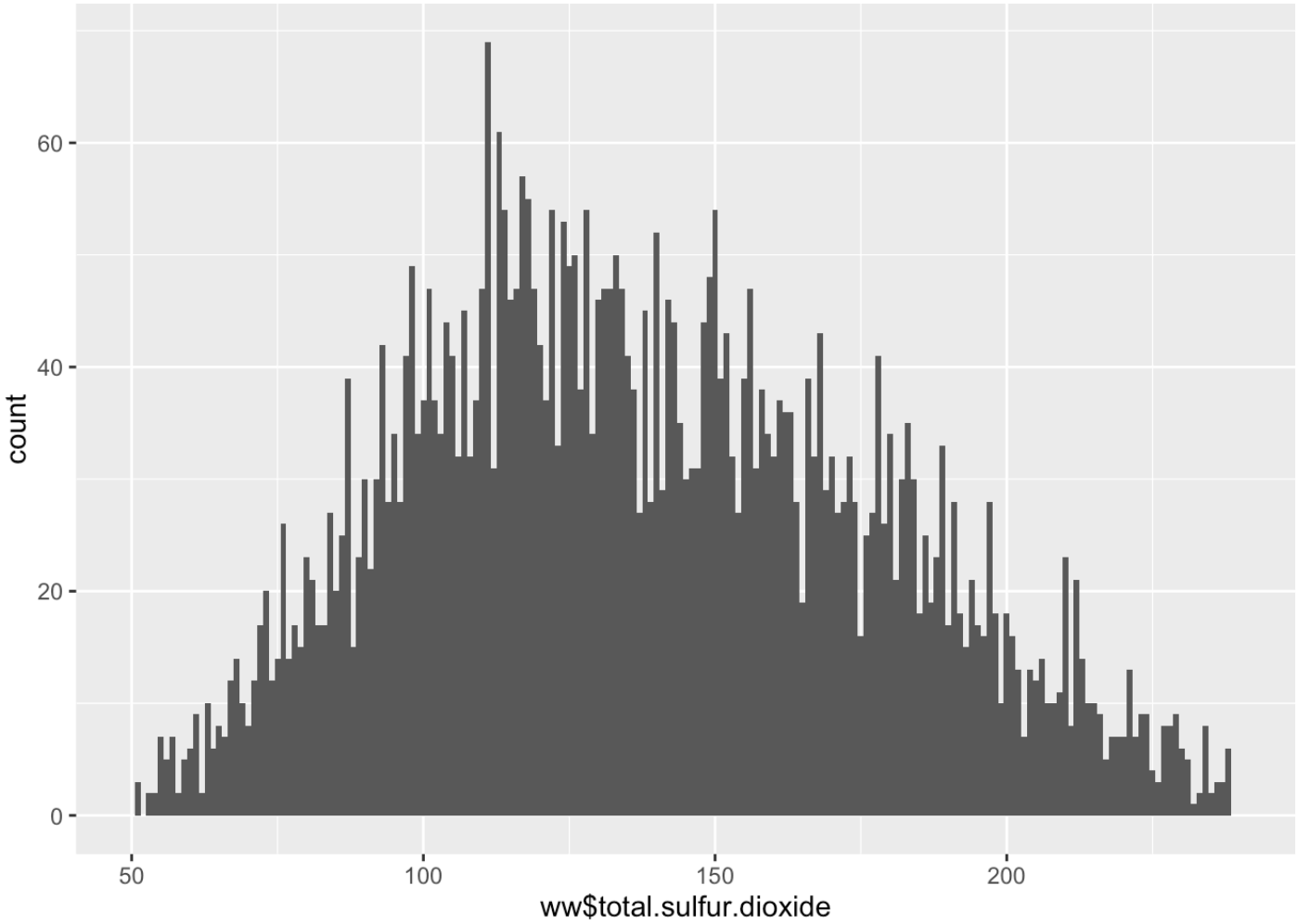

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   23.00   34.00   35.31   46.00  289.00
```

```
## [1]   9 440
```

```
## [1] 170 132  97 186 186  97
```



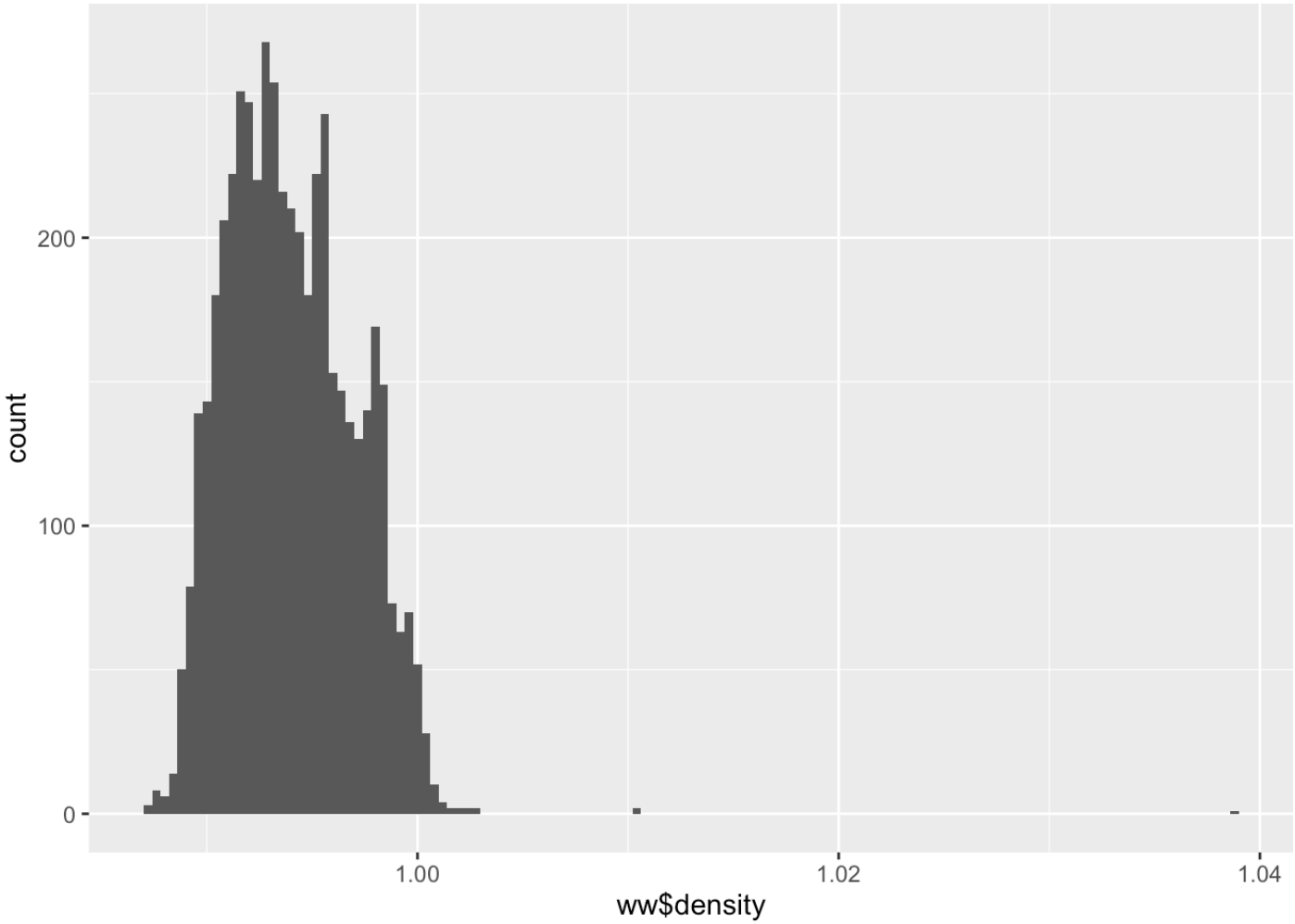

```
## Warning: Removed 97 rows containing non-finite values (stat_bin).
```




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

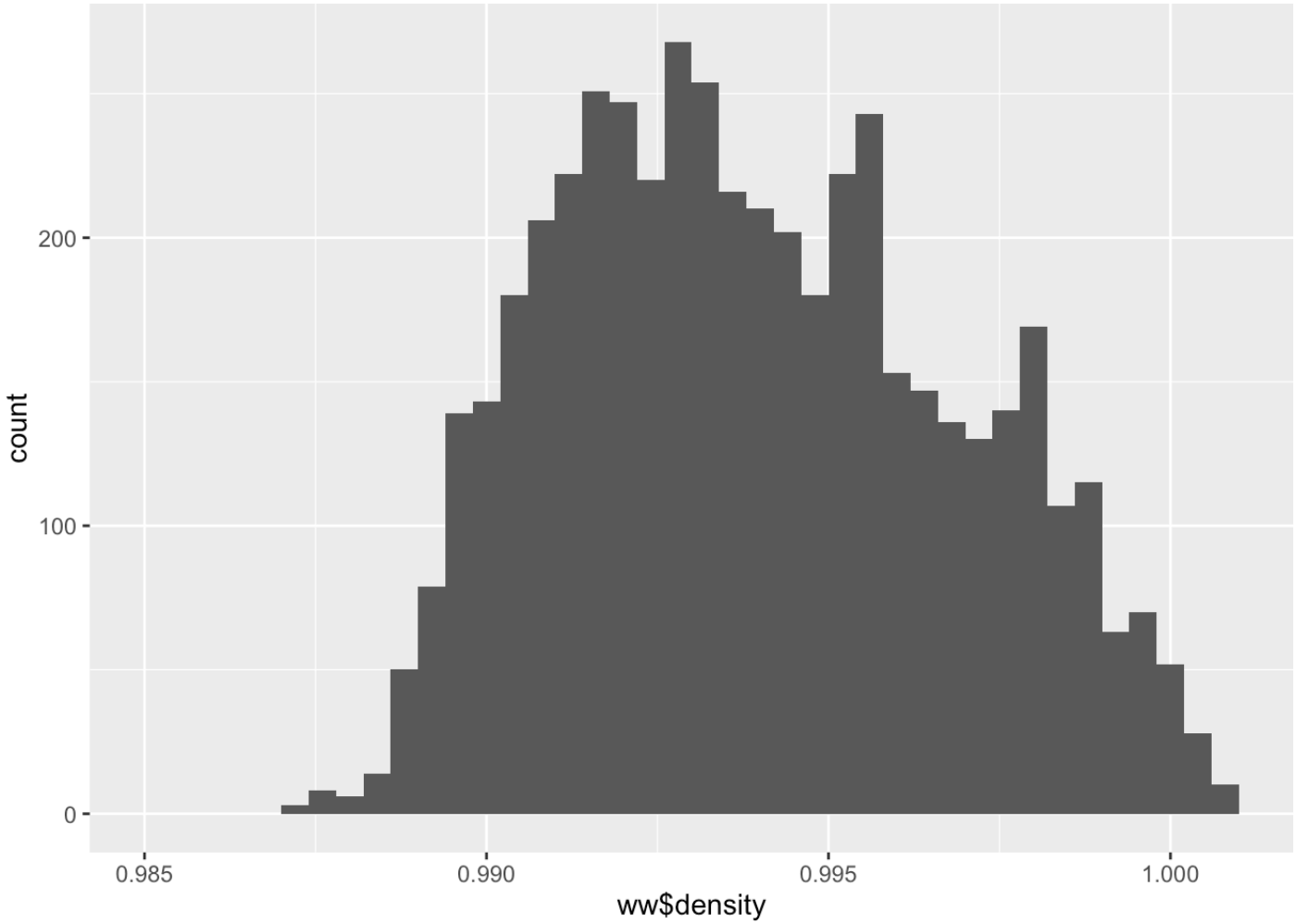
```
## [1] 0.98711 1.03898
```

```
## [1] 1.0010 0.9940 0.9951 0.9956 0.9956 0.9951
```




```
##   vars    n mean sd median trimmed mad  min  max range skew kurtosis se
## 1    1 4898 0.99  0   0.99    0.99   0 0.99 1.04  0.05 0.98     9.78  0
```

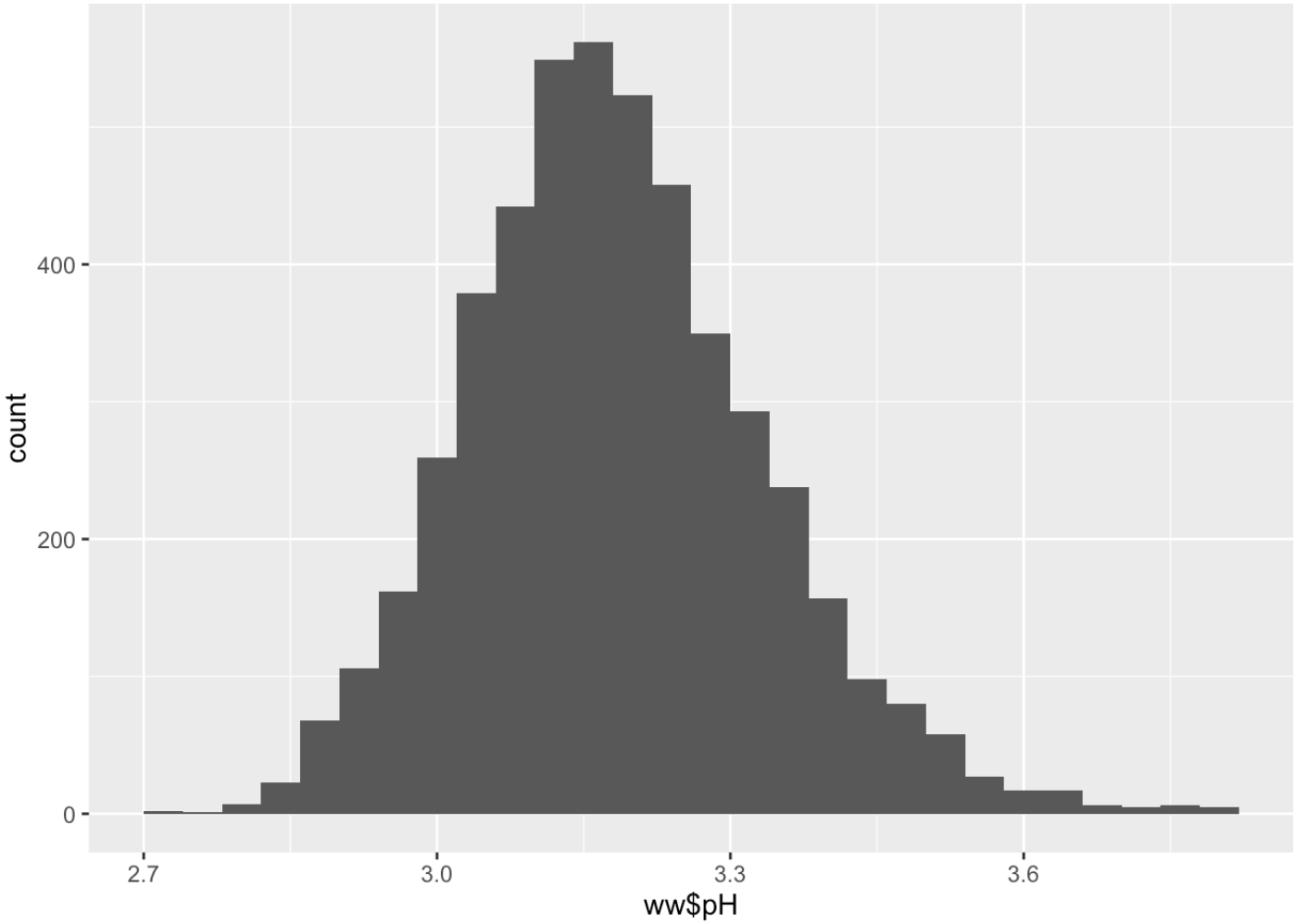
```
## Warning: Removed 15 rows containing non-finite values (stat_bin).
```




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

```
## [1] 2.72 3.82
```

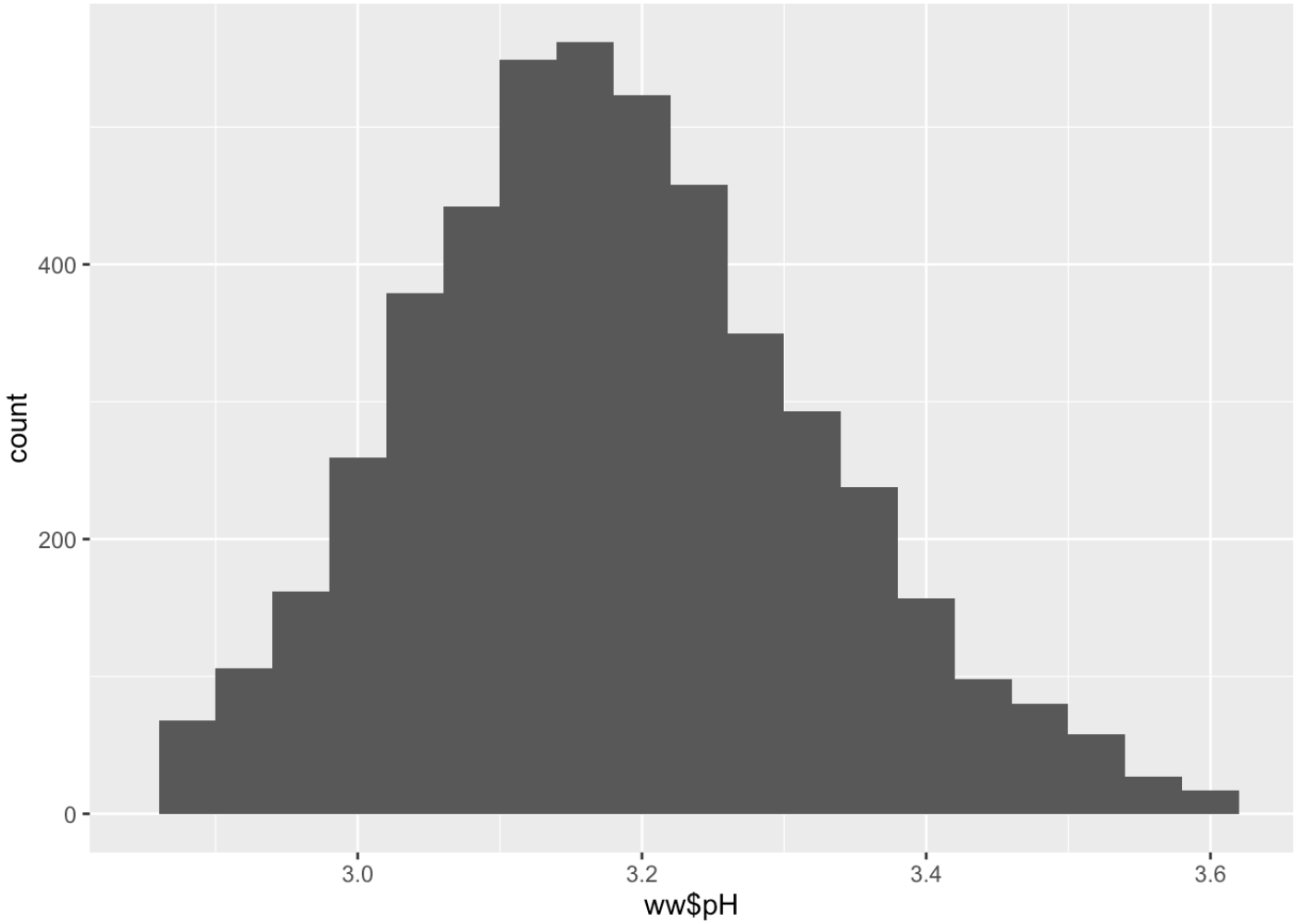
```
## [1] 3.00 3.30 3.26 3.19 3.19 3.26
```




```
##   vars    n mean   sd median trimmed  mad  min  max range skew kurtosis se
## 1    1 4898 3.19 0.15   3.18    3.18 0.15 2.72 3.82   1.1 0.46     0.53  0
```

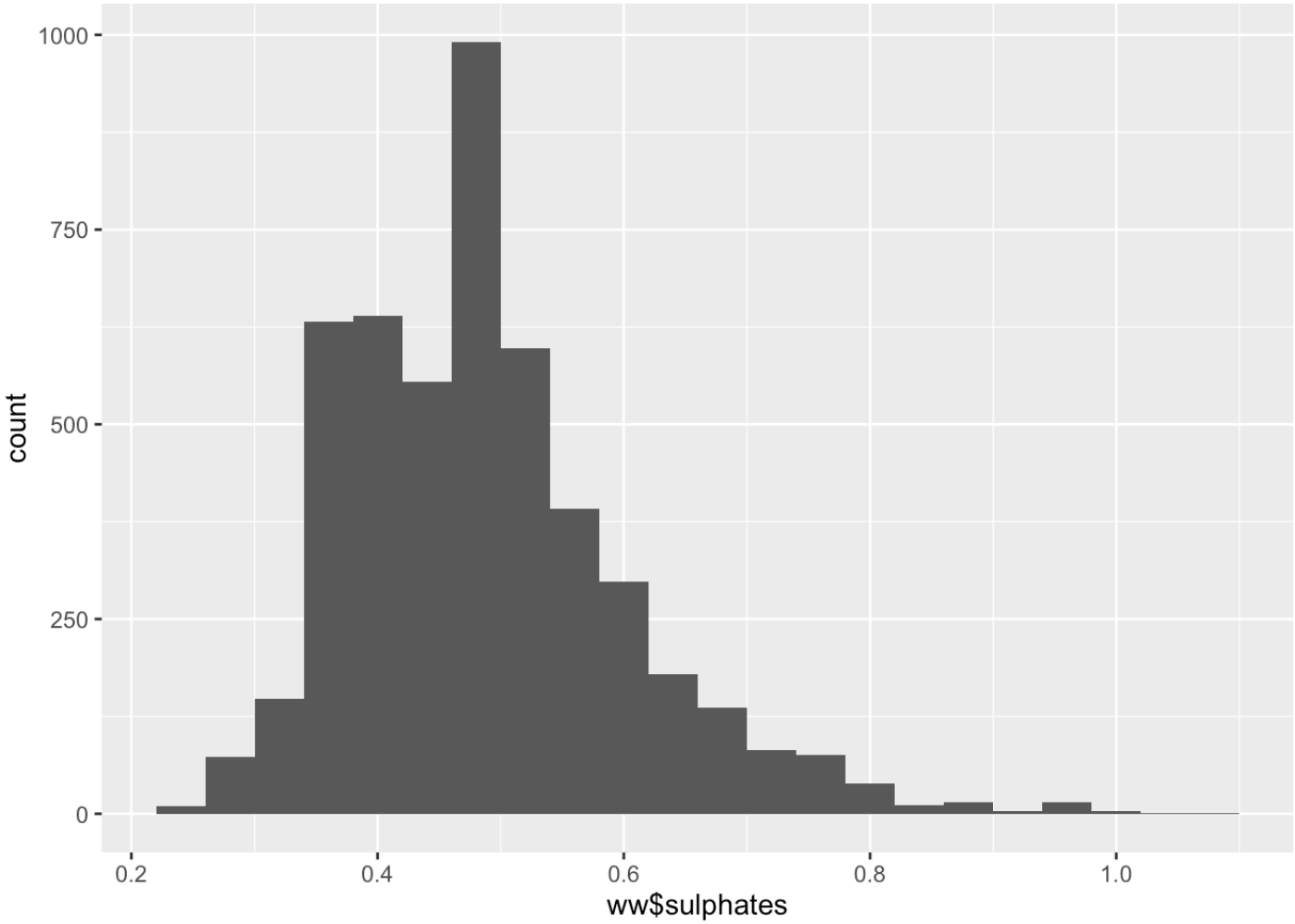
```
## Warning: Removed 54 rows containing non-finite values (stat_bin).

## Warning: Removed 1 rows containing missing values (geom_bar).
```



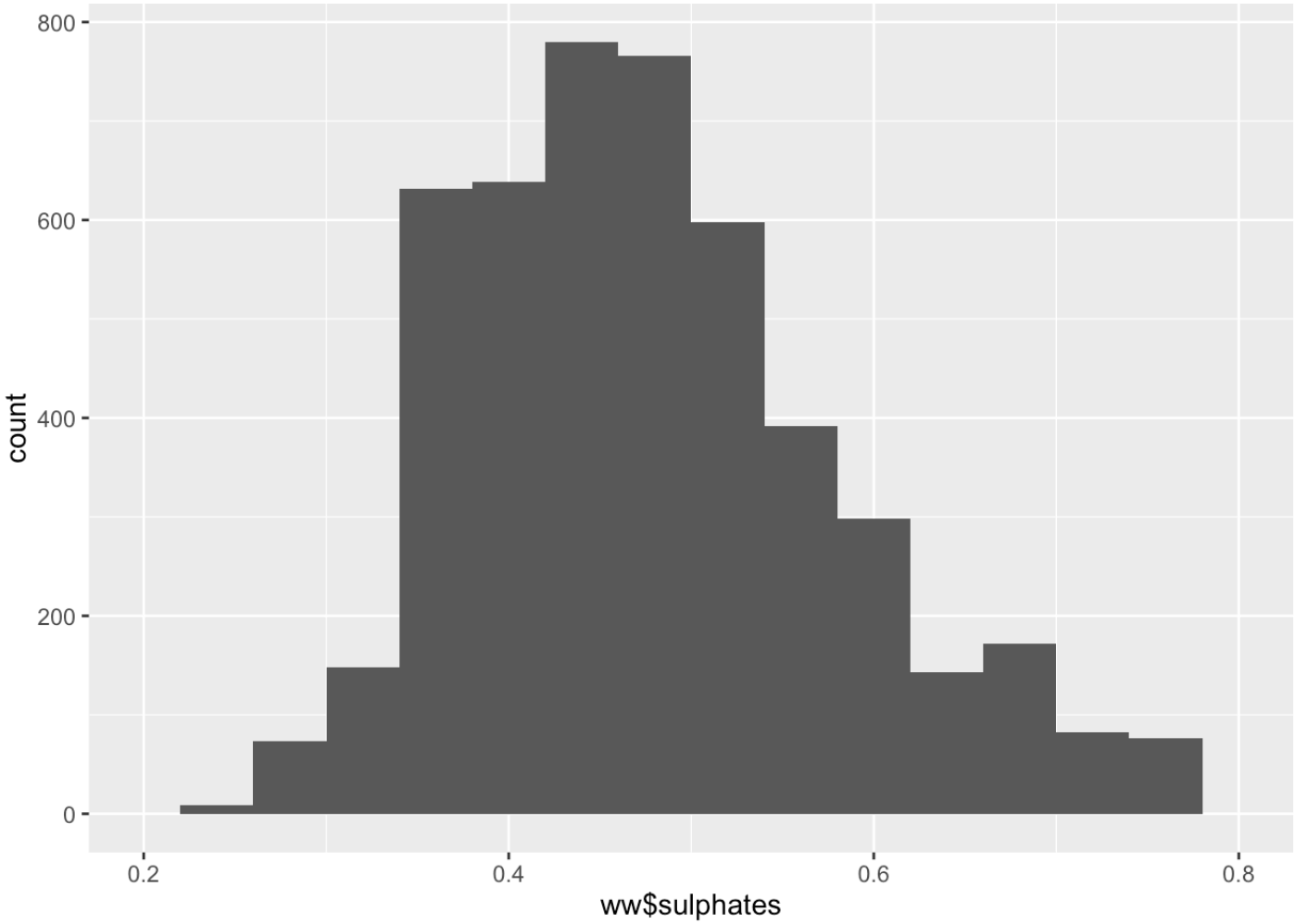

```
## [1] 0.22 1.08
```

```
## [1] 0.45 0.49 0.44 0.40 0.40 0.44
```




```
##   vars    n mean   sd median trimmed mad  min  max range skew kurtosis se
## 1    1 4898 0.49 0.11   0.47    0.48 0.1 0.22 1.08  0.86 0.98     1.59  0
```

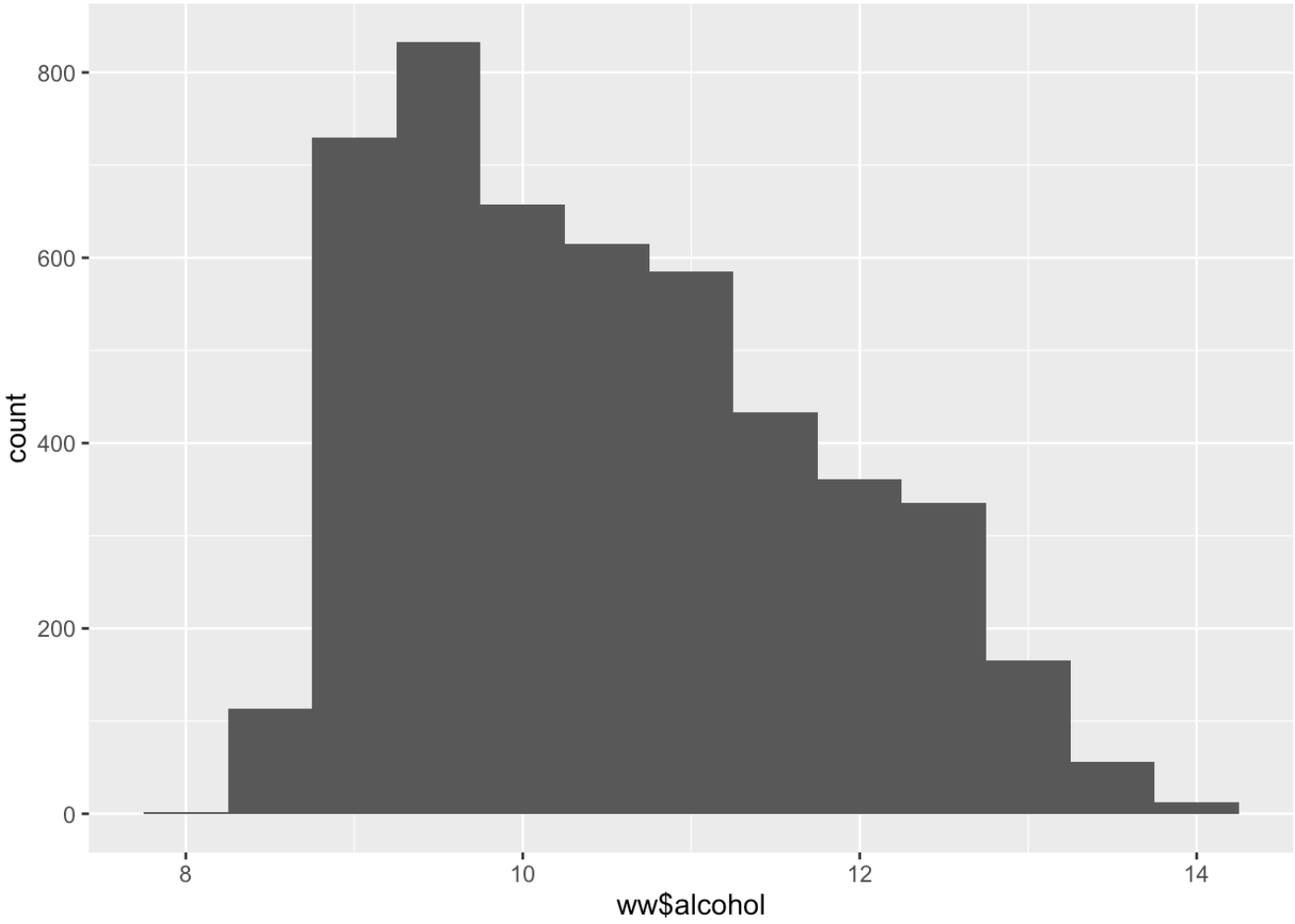
```
## Warning: Removed 68 rows containing non-finite values (stat_bin).
```




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2200  0.4100  0.4700  0.4898  0.5500  1.0800
```

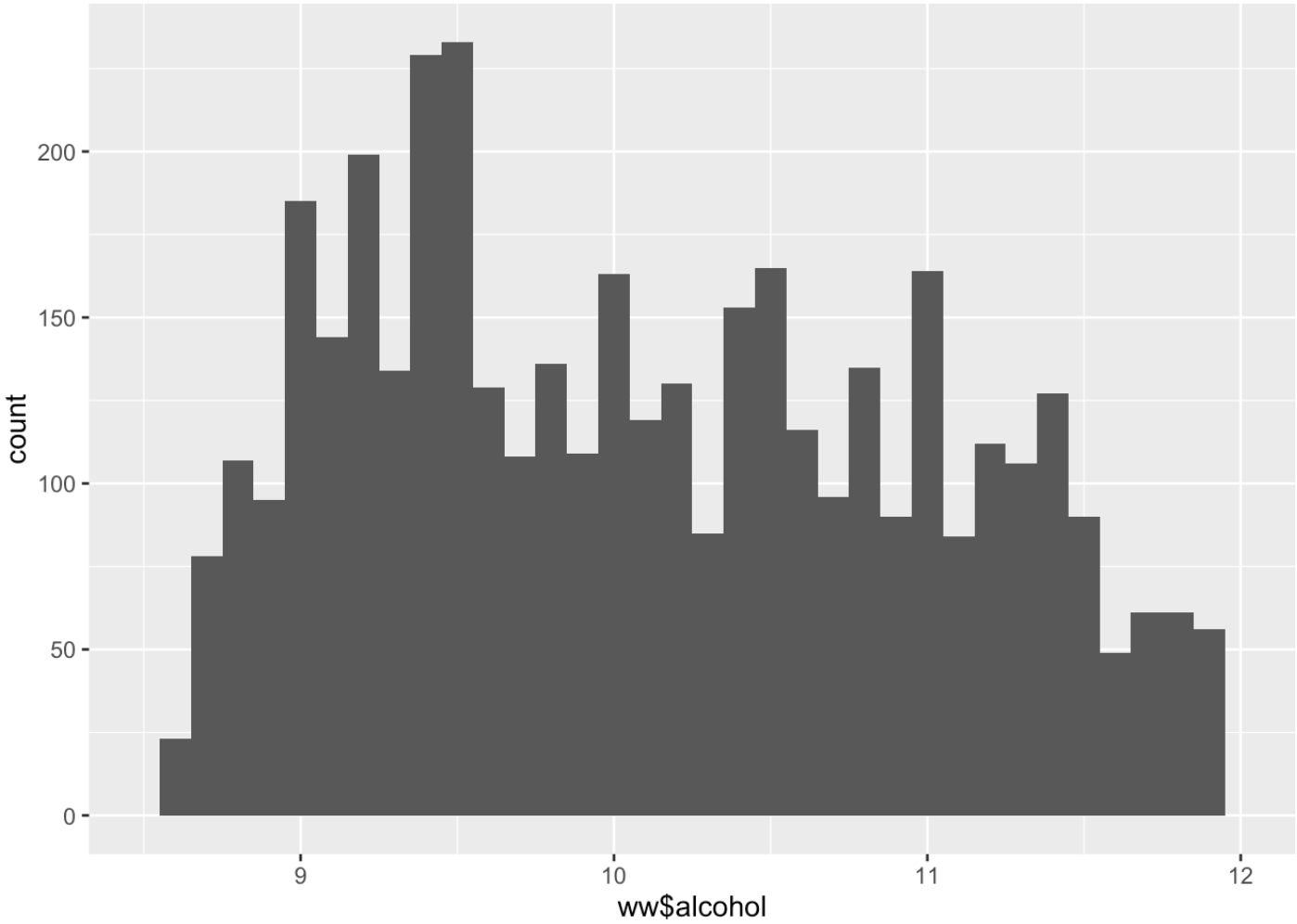
```
## [1]  8.0 14.2
```

```
## [1]  8.8  9.5 10.1  9.9  9.9 10.1
```




```
##   vars    n  mean   sd median trimmed  mad min  max range skew kurtosis
## 1    1 4898 10.51 1.23   10.4   10.43 1.48   8 14.2   6.2 0.49     -0.7
##     se
## 1 0.02
```

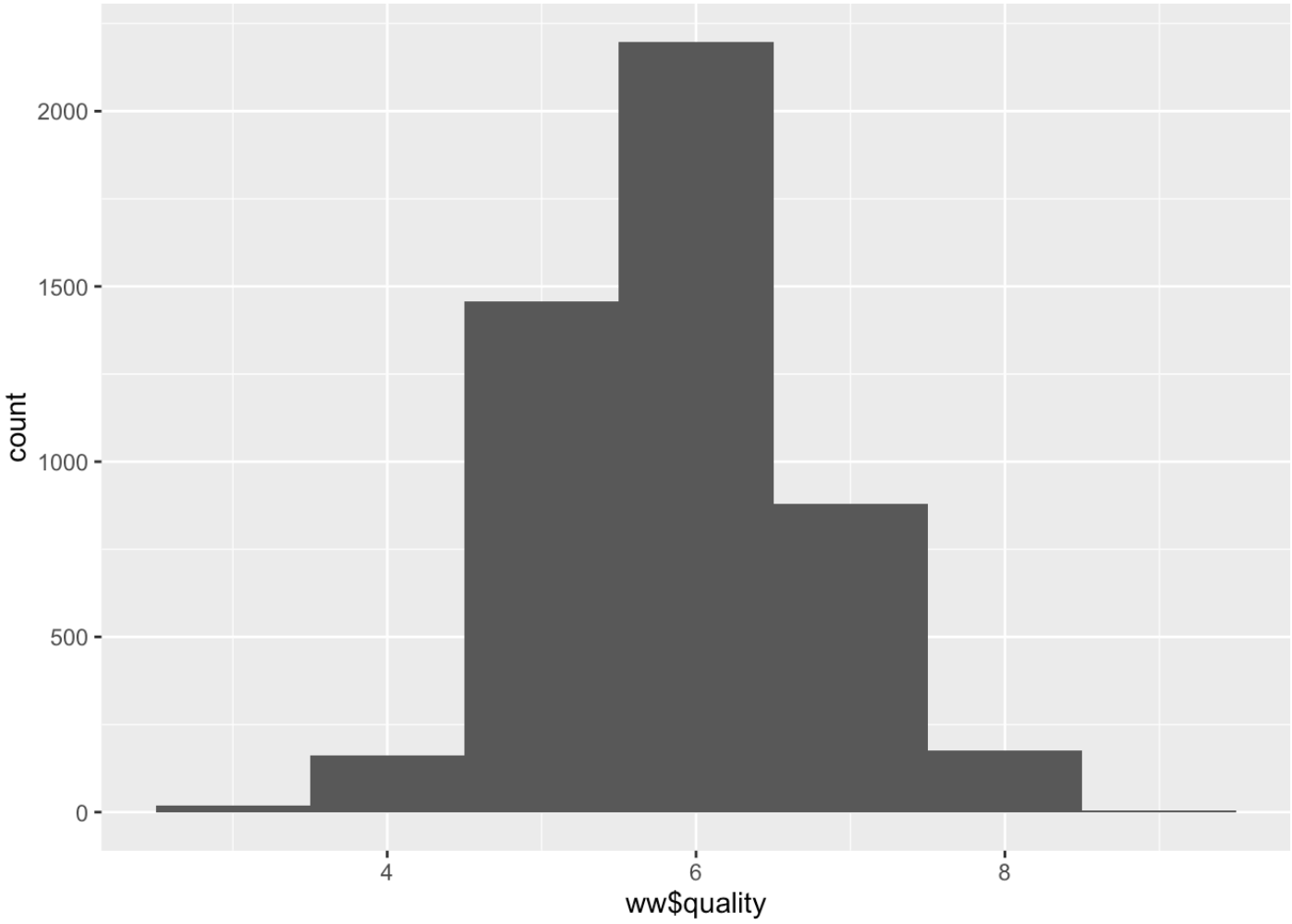
```
## Warning: Removed 716 rows containing non-finite values (stat_bin).
```



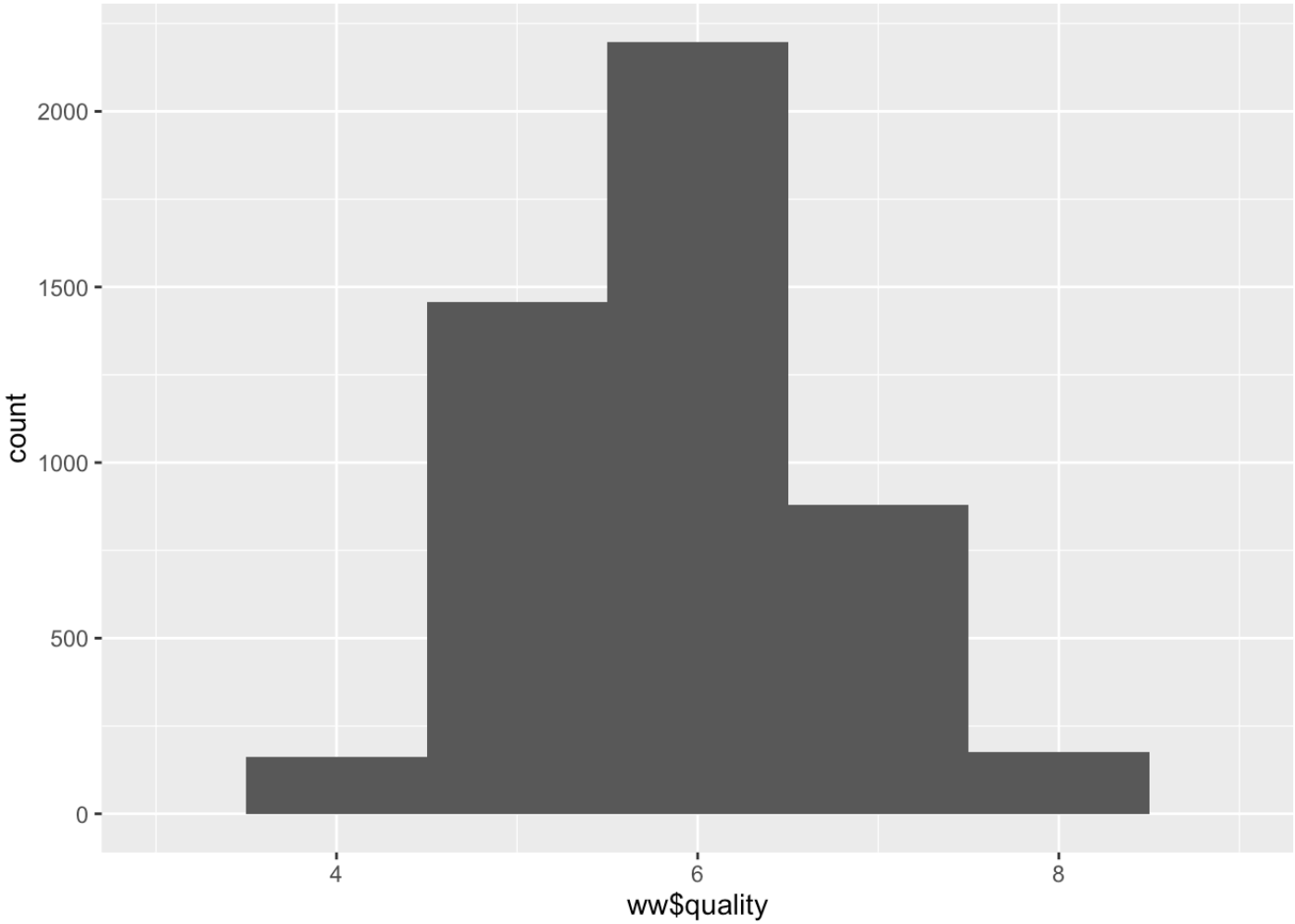

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.50   10.40   10.51   11.40   14.20
```

```
## [1] 3 9
```

```
## [1] 6 6 6 6 6 6
```




```
##   vars    n mean   sd median trimmed  mad min max range skew kurtosis   se
## 1    1 4898 5.88 0.89      6    5.85 1.48   3   9     6 0.16     0.21 0.01
```




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

```
##  [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"
```

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

Univariate Analysis

What is the structure of your dataset?

There are 4890 wine observations in the dataset with 13 features (X, fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide,total.sulfur.dioxide, density,pH, sulphates, alcohol). The variable quality is the only ordered factor variables with the following levels.

(worst) ———-- (best) quality: 3, 4, 5, 6, 7, 8

Other observations:

Most wines have a quality rating 6 Because of the normal dist. it is nearly equally likely to get a quality rating of 4 as it is 8. The median pH level 3.180. Acidity seems to be broken down into three variables (fixed, volatile, and citric)

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

I'm interested in the chemical makeup of each wine and how it contributes to quality. The main features i'm interested in are acidity and residual.sugar as they contribute to quality.

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

I believe pH levels, alcohol will be helpful predictors as well.

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

I've added a ratio between the absolute value alcohol and relative value residual.sugar under the assumption that all observations have the exact same volume.

I also added bound_s02, which is total.sulfur.dioxide - free.sulfur.dioxide, this gives us the bounded S02 by leaving out the free dioxide.

I've also put volatily and fixed acids together to understand the overall acid levels. Notice that I did not add citric acid because in early tests I couldn't find any relation between them and citric acid may also be an additive that is very minimal.

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?

Log Transforms: Residual.sugar and volatile.acidity

Residual sugar I've chosen to do a log transform here because the residual sugar distribution is right skewed with a very long tail. By doing so it has been transformed to be closer to a normal distribution with signs of a bi-modal distribution.

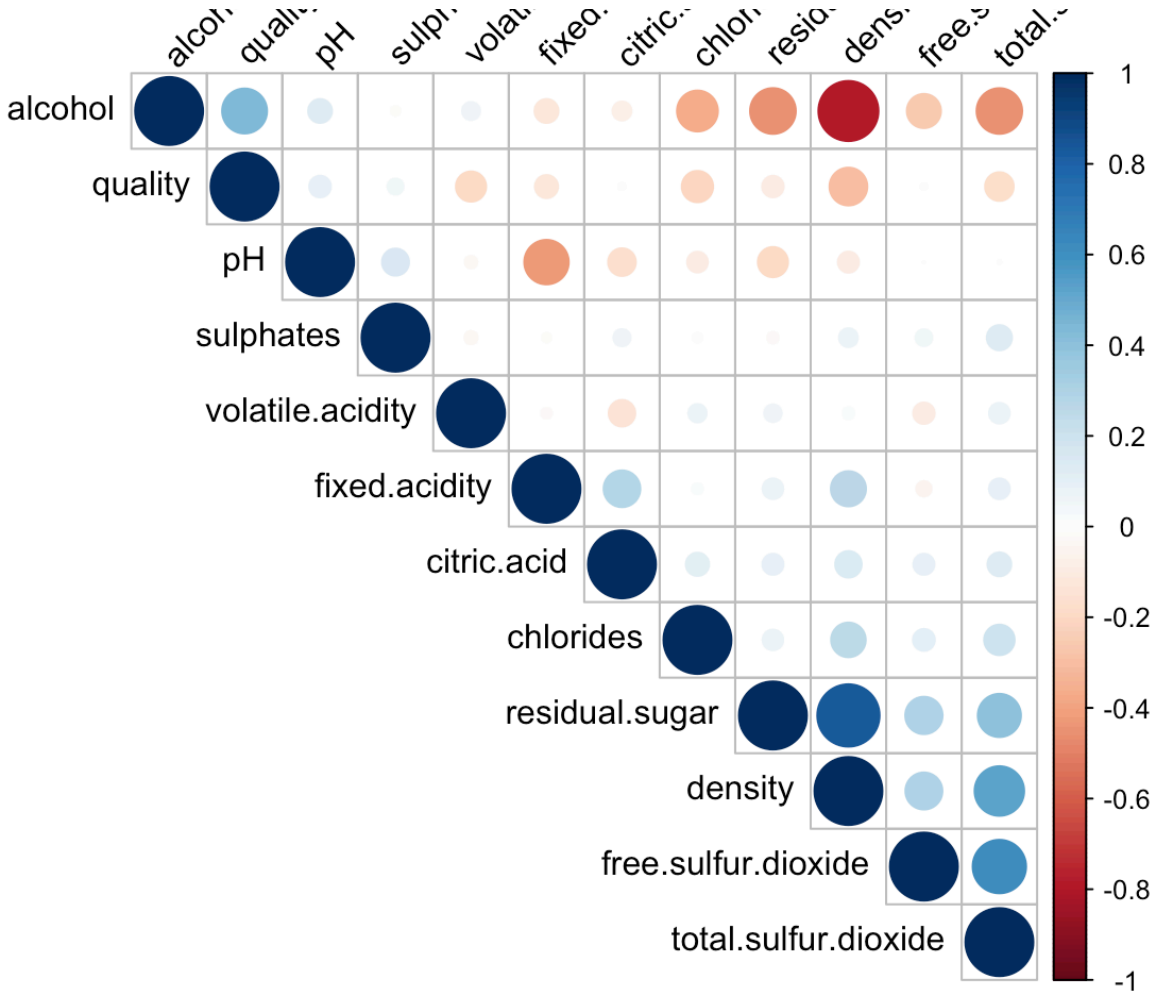
Volatile.Acidity Since volatile acidity is right skewed I've transformed the data using log10 and then changed the binwidth accordingly.

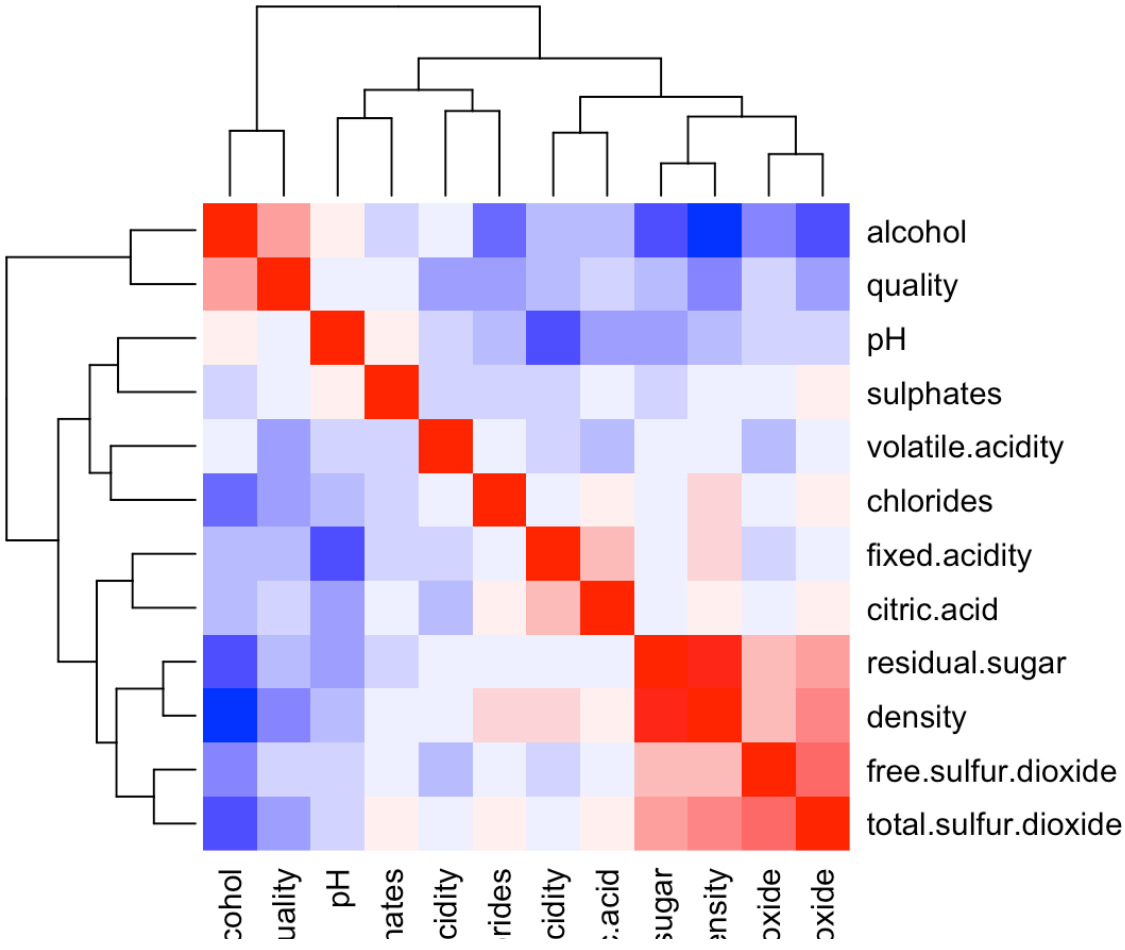
FORM CHANGE possiblities - Grouping - Acidity (citric/fixed/volatile) - Dioxide (total/free) Further research needed

Bivariateplots

```
##  [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"
```

```
##                       fx. v c. r ch f.. t d p s a q
## fixed.acidity         1
## volatile.acidity          1
## citric.acid                 1
## residual.sugar                 1
## chlorides                        1
## free.sulfur.dioxide                 1
## total.sulfur.dioxide           .    ,   1
## density                        +        . 1
## pH                     .                     1
## sulphates                                      1
## alcohol                        . .      . ,     1
## quality                                   .     . 1
## attr(,"legend")
## [1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '*' 0.95 'B' 1
```

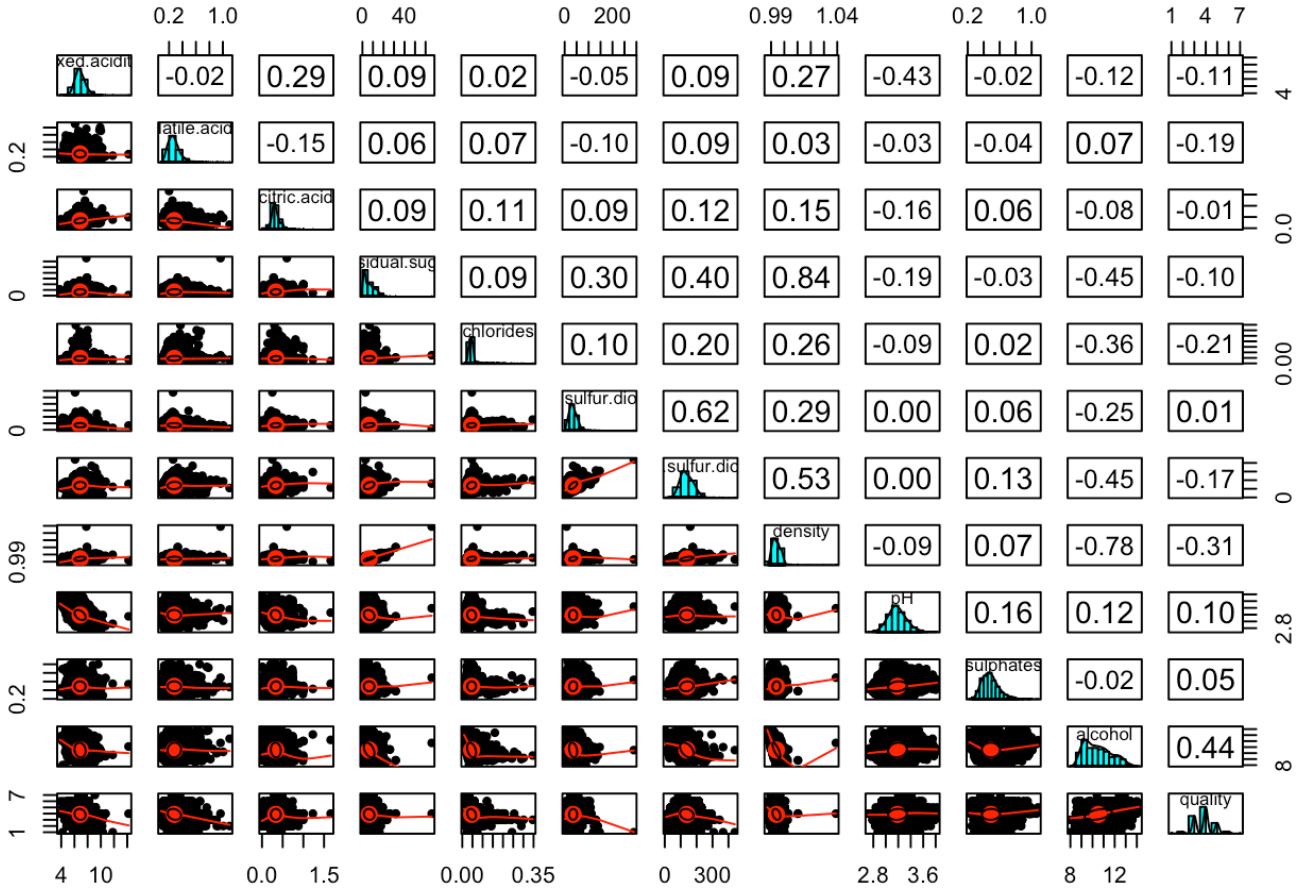

alcohol
quality
pH
sulphates
volatile.acidity
fixed.acidity
citric.acid
chlorides
residual.sugar
density
free.sulfur.dioxide
total.sulfur.dioxide
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

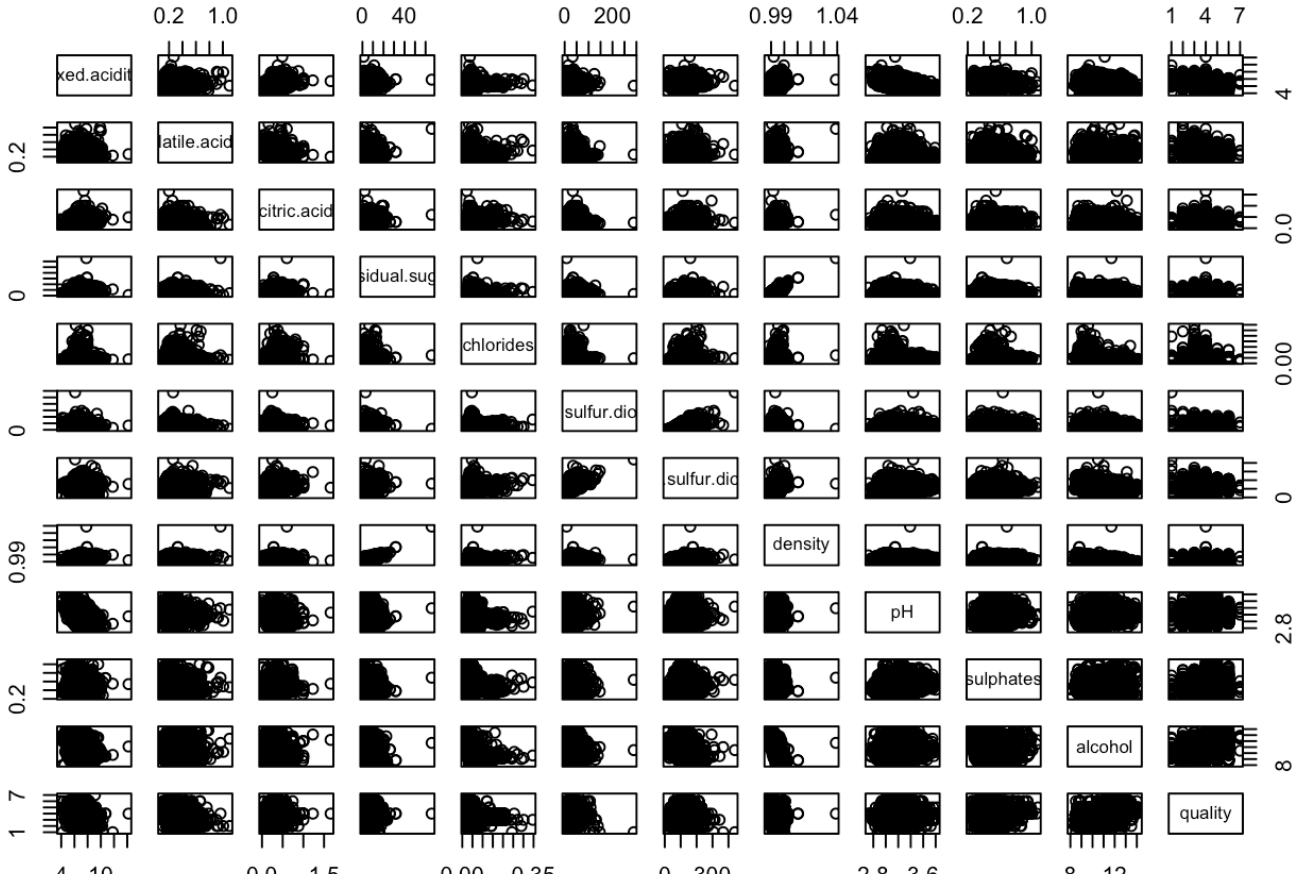

alcohol
quality
pH
sulphates
volatile.acidity
chlorides
fixed.acidity
citric.acid
residual.sugar
density
free.sulfur.dioxide
total.sulfur.dioxide

alc
q
sulpl
atile.a
chlo
xed.a
citric
idual.s
de
lfur.dic
lfur.dic

```
## [1] FALSE
```

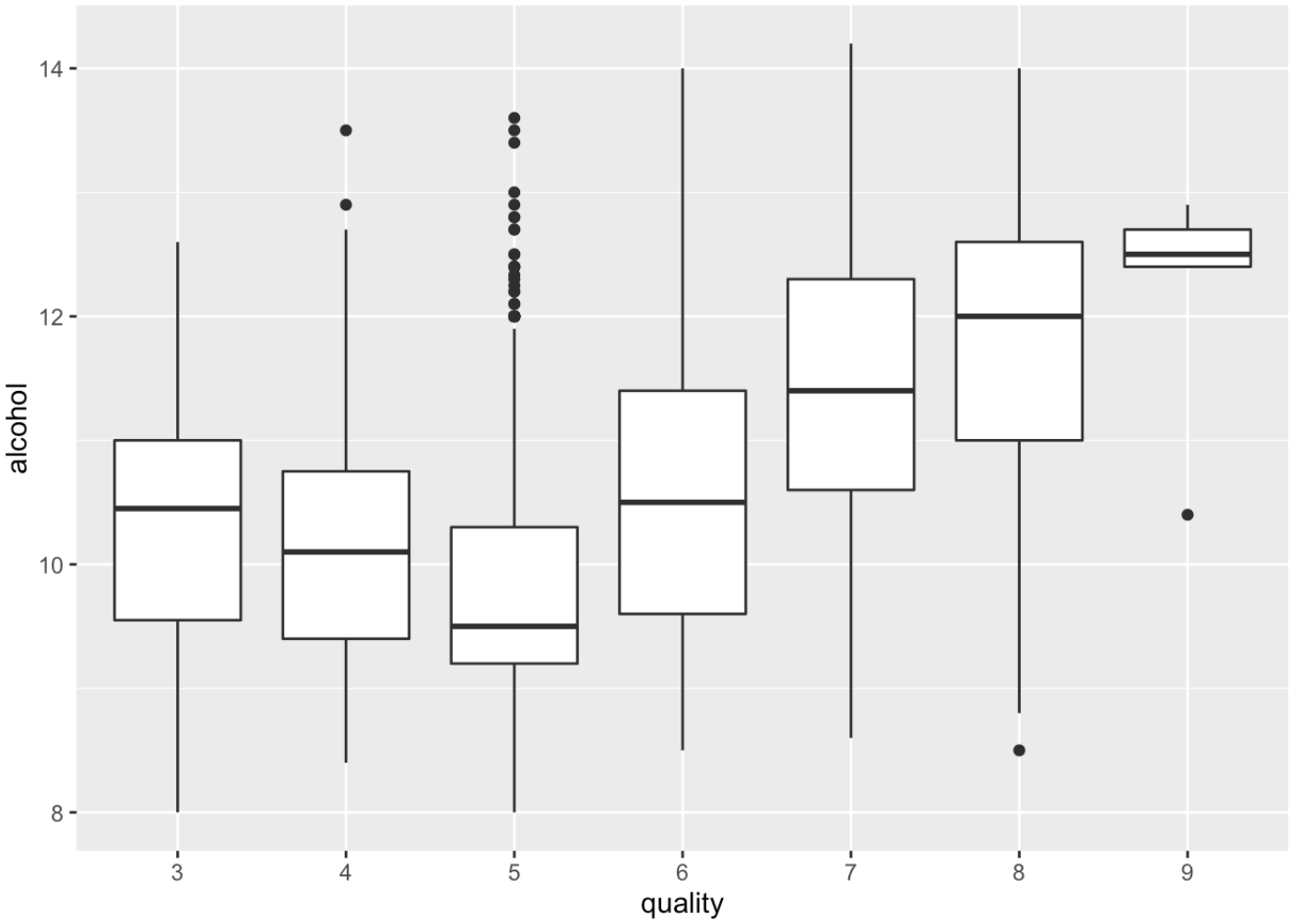
```
## [1] TRUE
```

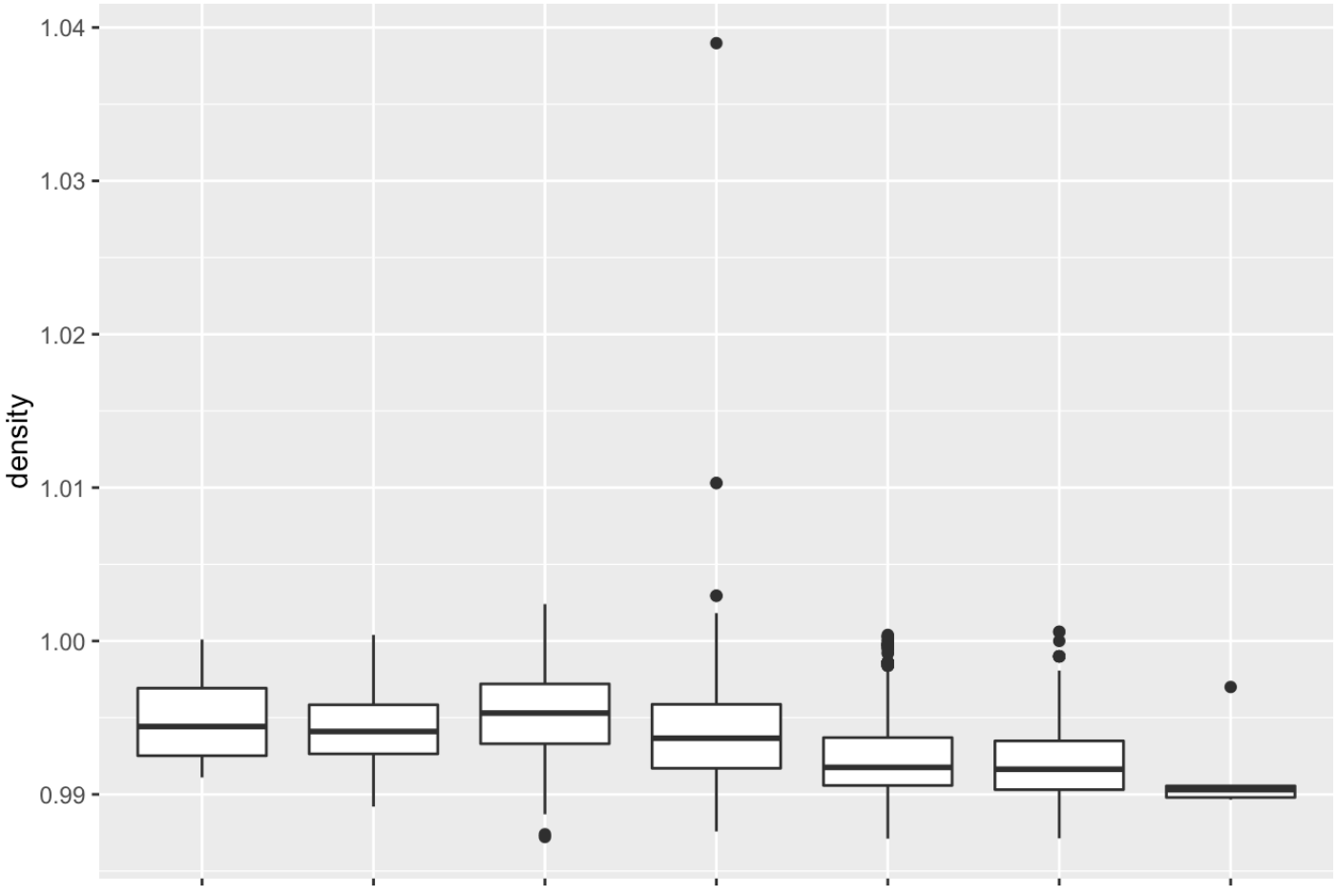

xed.acidit
-0.02 0.29 0.09 0.02 -0.05 0.09 0.27 -0.43 -0.02 -0.12 -0.11
latile.acid
-0.15 0.06 0.07 -0.10 0.09 0.03 -0.03 -0.04 0.07 -0.19
citric.acid
0.09 0.11 0.09 0.12 0.15 -0.16 0.06 -0.08 -0.01
sidual.sug
0.09 0.30 0.40 0.84 -0.19 -0.03 -0.45 -0.10
chlorides
0.10 0.20 0.26 -0.09 0.02 -0.36 -0.21
sulfur.dio
0.62 0.29 0.00 0.06 -0.25 0.01
sulfur.dio
0.53 0.00 0.13 -0.45 -0.17
density
-0.09 0.07 -0.78 -0.31
pH
0.16 0.12 0.10
sulphates
-0.02 0.05
alcohol
0.44
quality
0.2 1.0
0 40
0 200
0.99 1.04
0.2 1.0
1 4 7
4 10
0.0 1.5
0.00 0.35
0 300
2.8 3.6
8 12

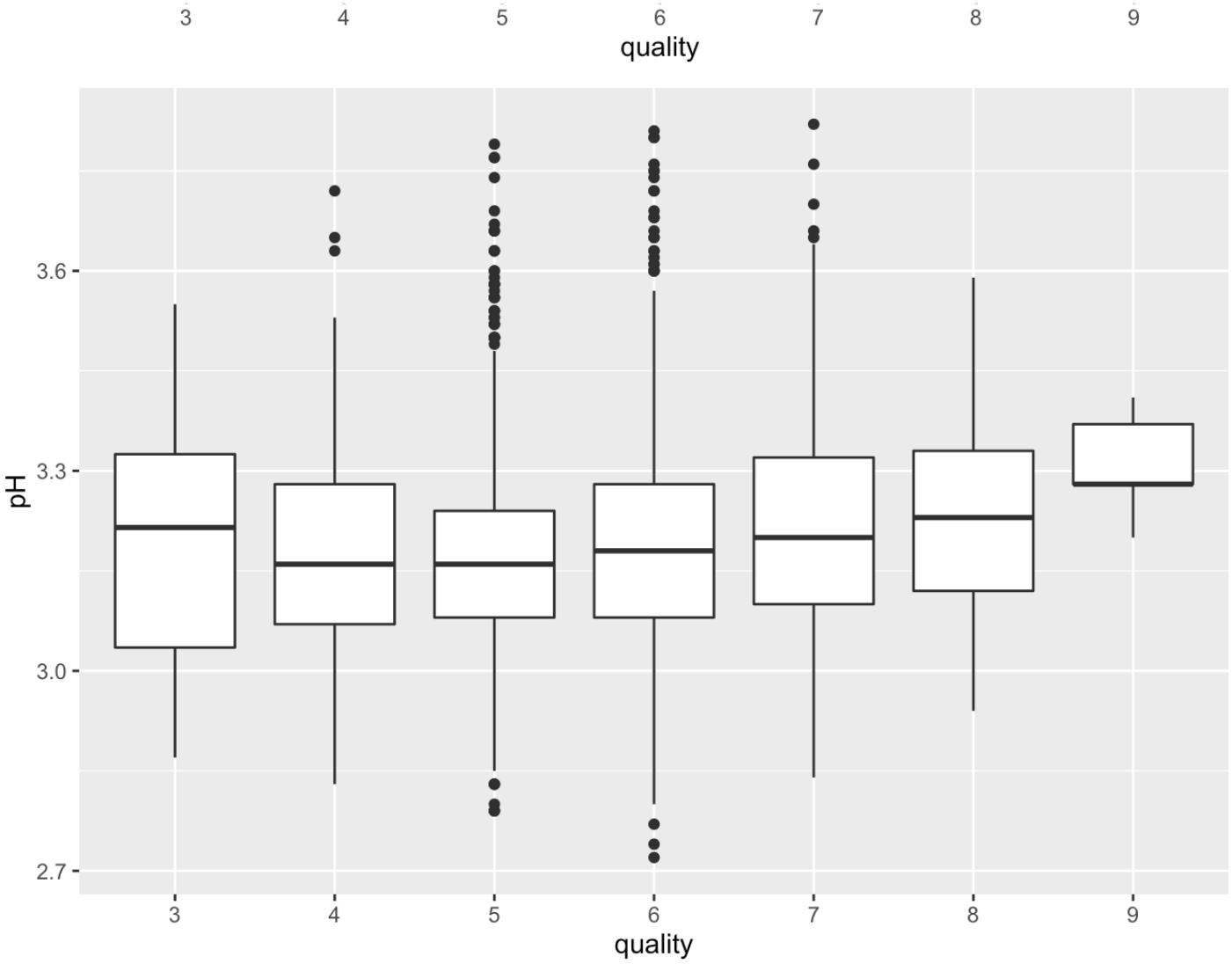

xed.acidit
latile.acid
citric.acid
sidual.sug
chlorides
sulfur.dio
sulfur.dio
density
pH
sulphates
alcohol
quality
0.2 1.0
0 40
0 200
0.99 1.04
0.2 1.0
1 4 7

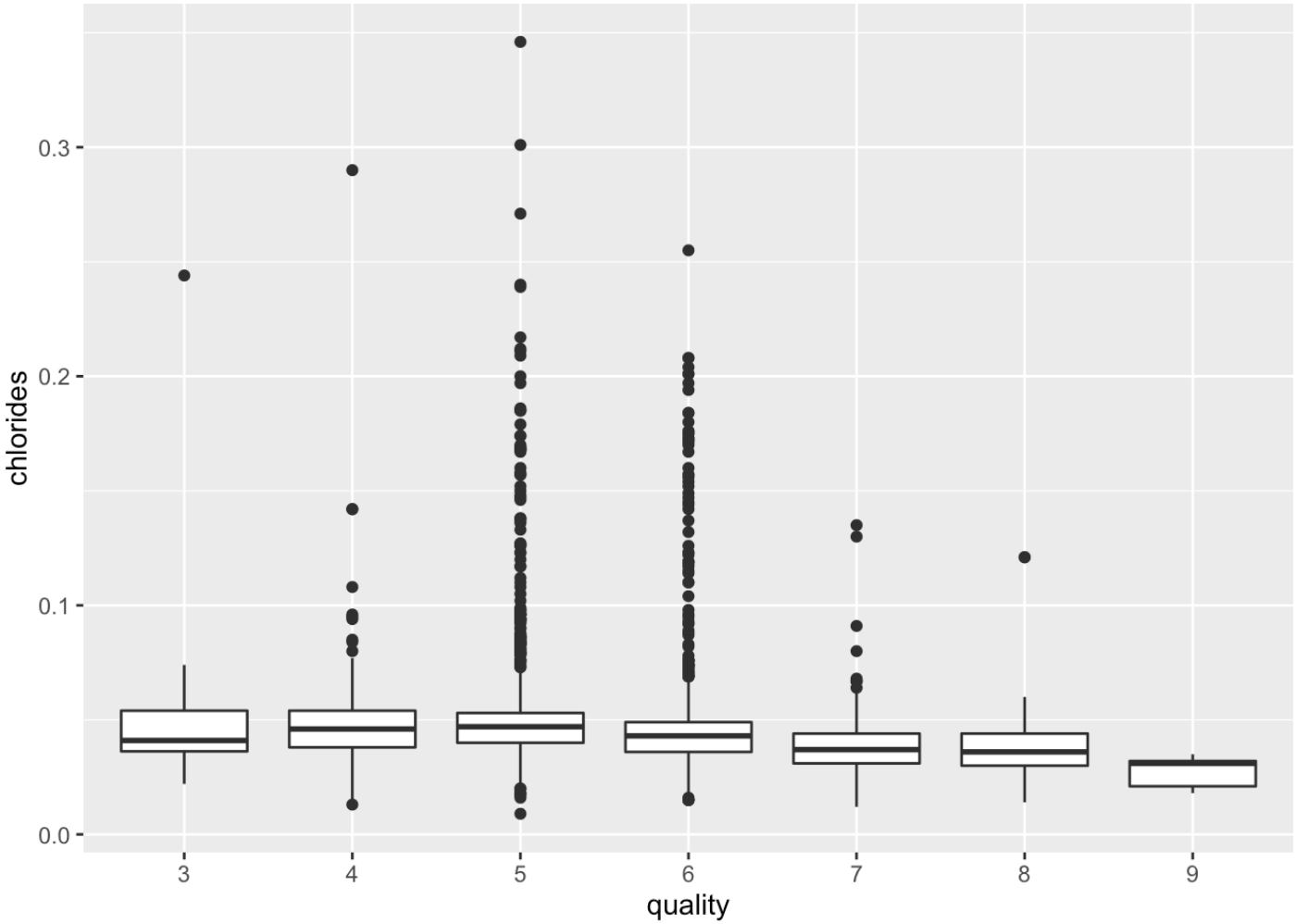
4 10 0.0 1.5 0.00 0.35 0 300 2.8 3.6 8 12

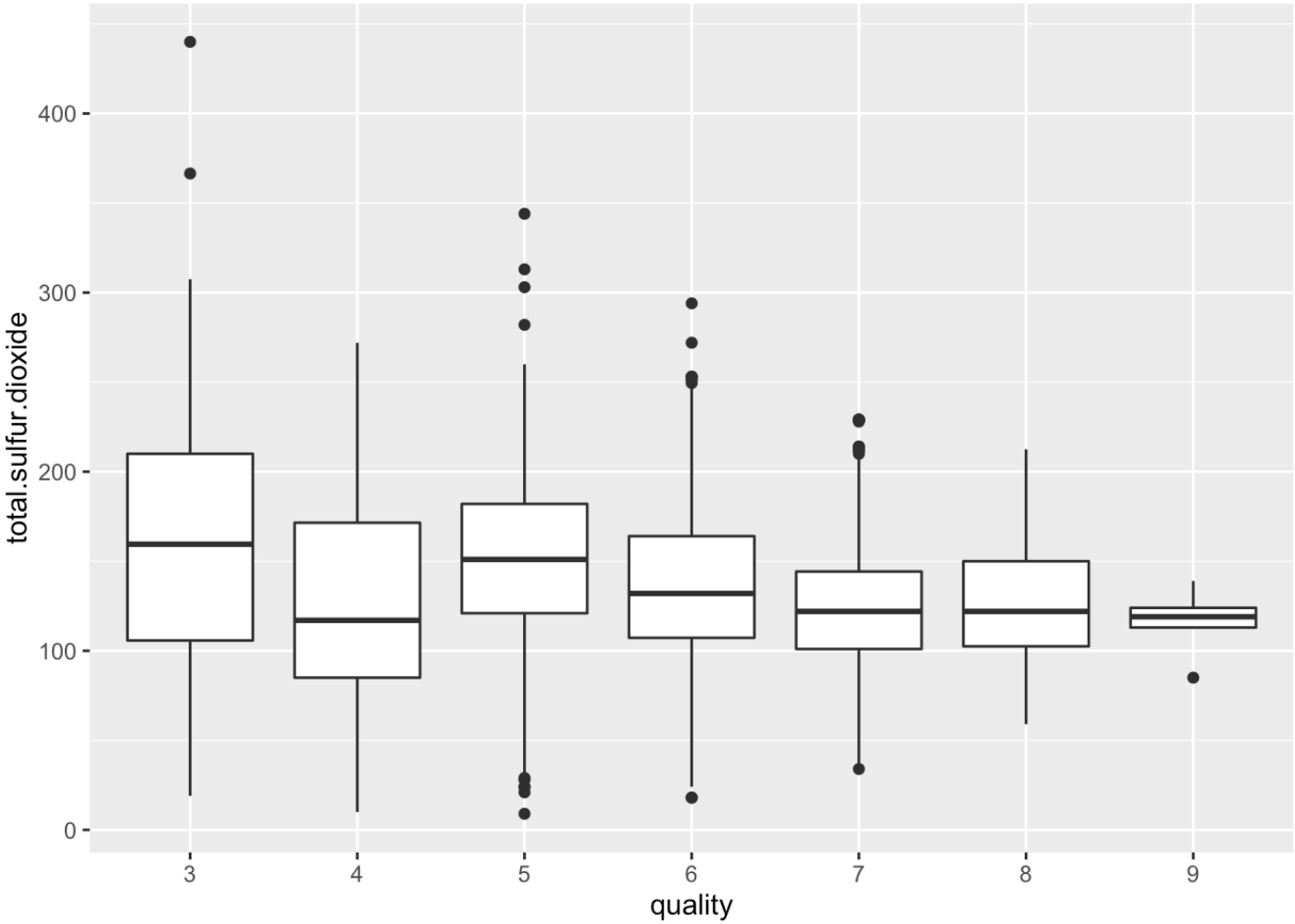
```
##  [1] "fixed.acidity"        "volatile.acidity"     "citric.acid"
##  [4] "residual.sugar"       "chlorides"            "free.sulfur.dioxide"
##  [7] "total.sulfur.dioxide" "density"              "pH"
## [10] "sulphates"            "alcohol"              "quality"
```

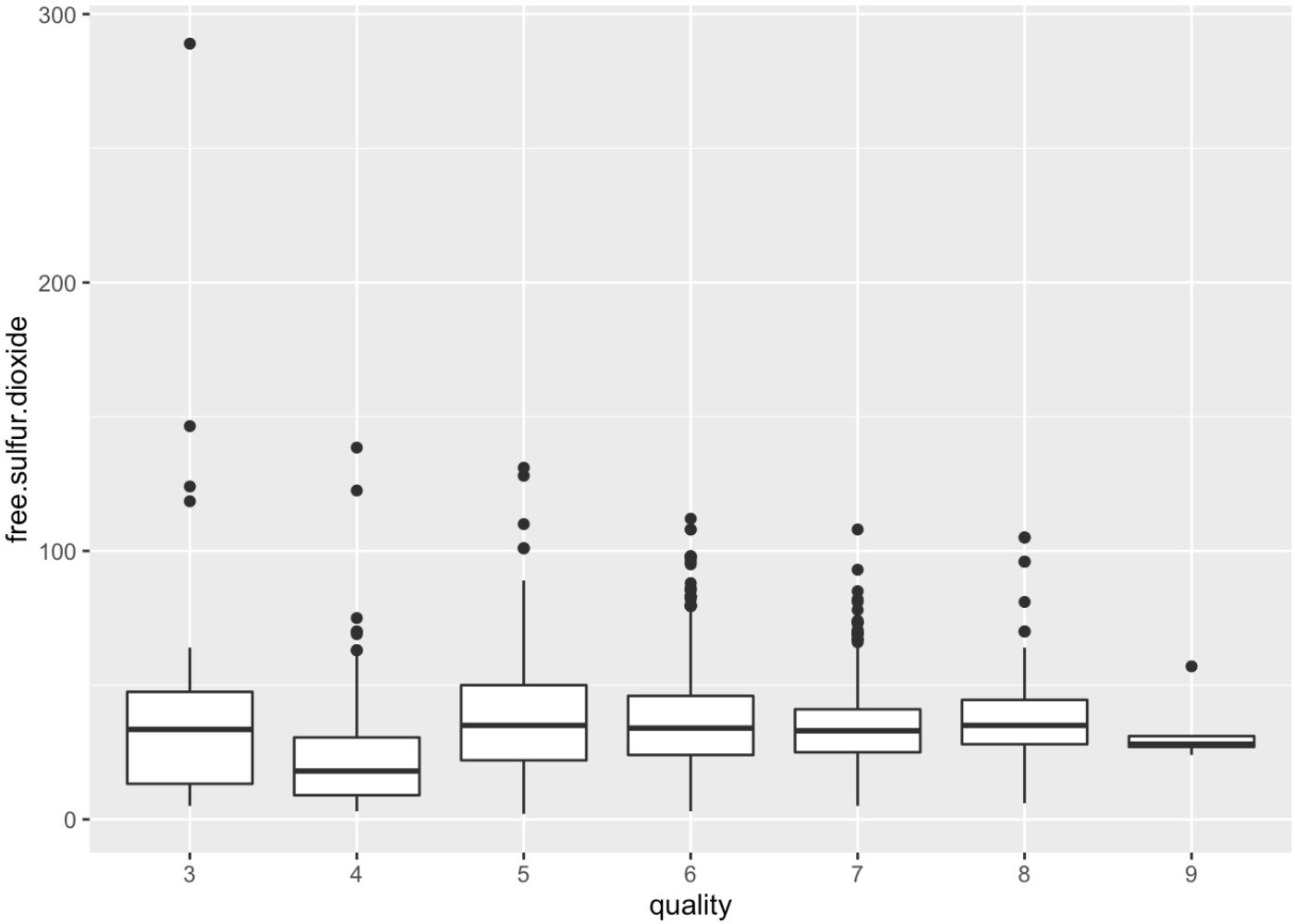

14
12
10
8
alcohol
3
4
5
6
7
8
9
quality

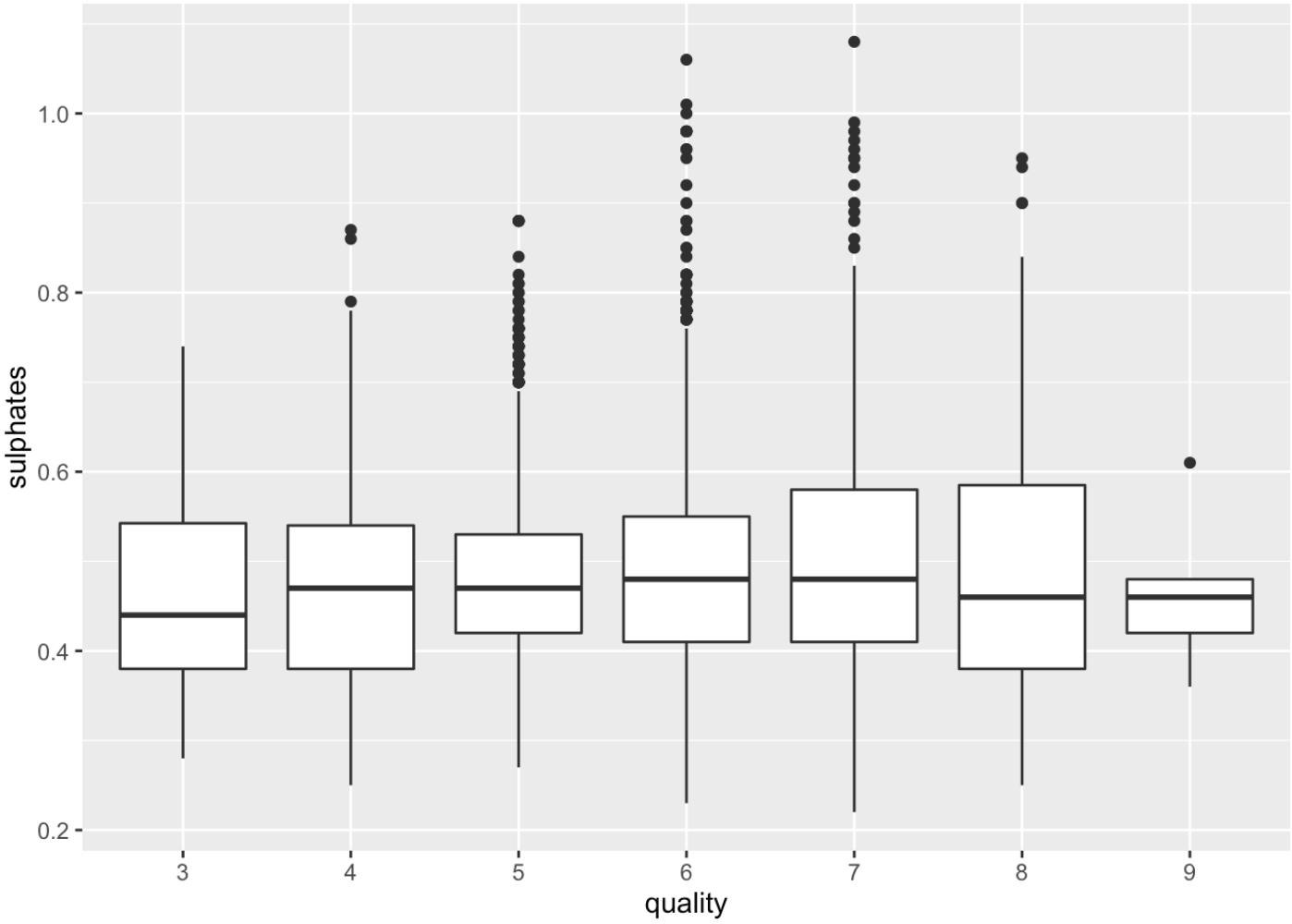

1.04
1.03
1.02
1.01
1.00
0.99
density

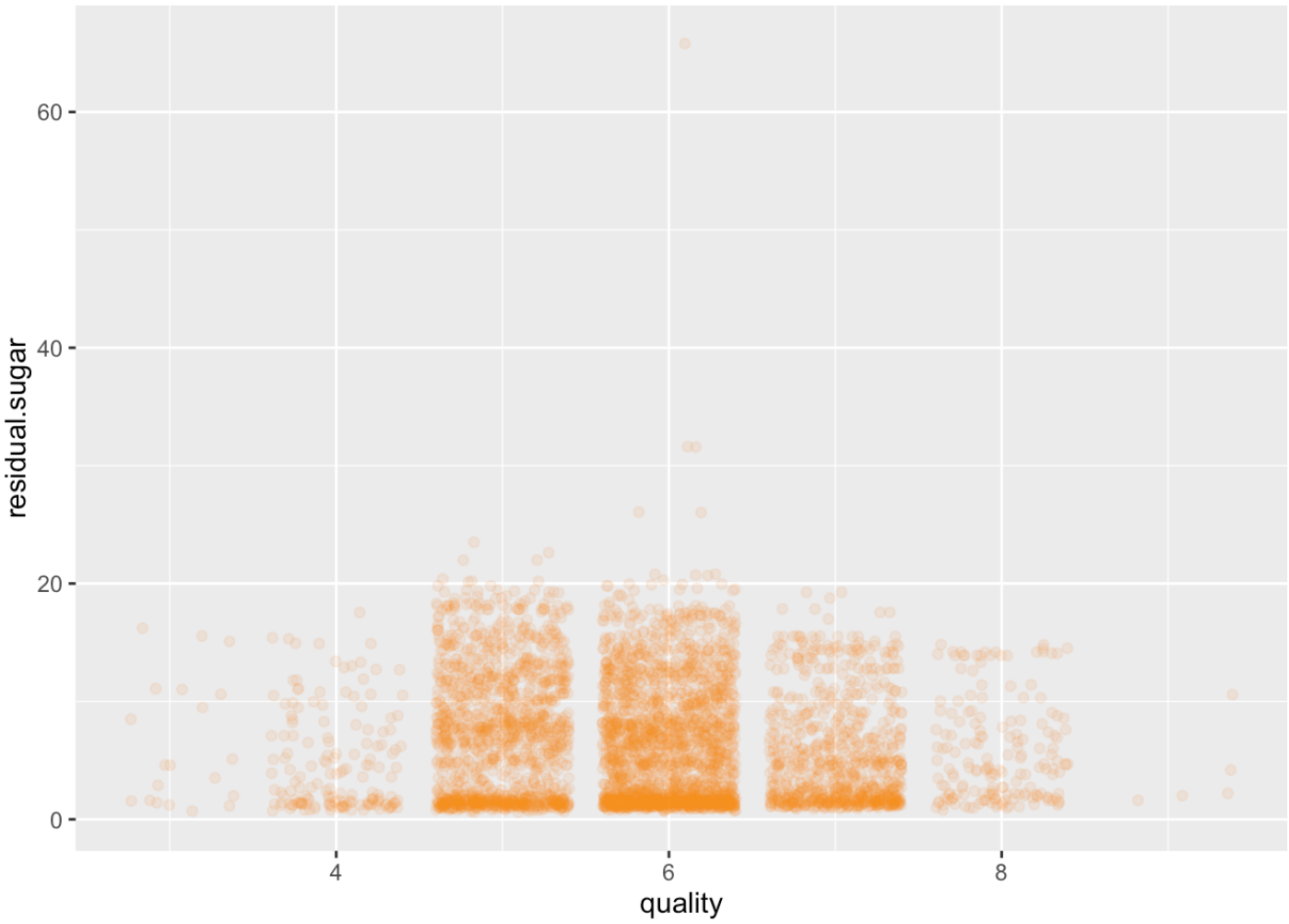

3
4
5
6
7
8
9
quality
3.6
3.3
3.0
2.7
pH
3
4
5
6
7
8
9
quality

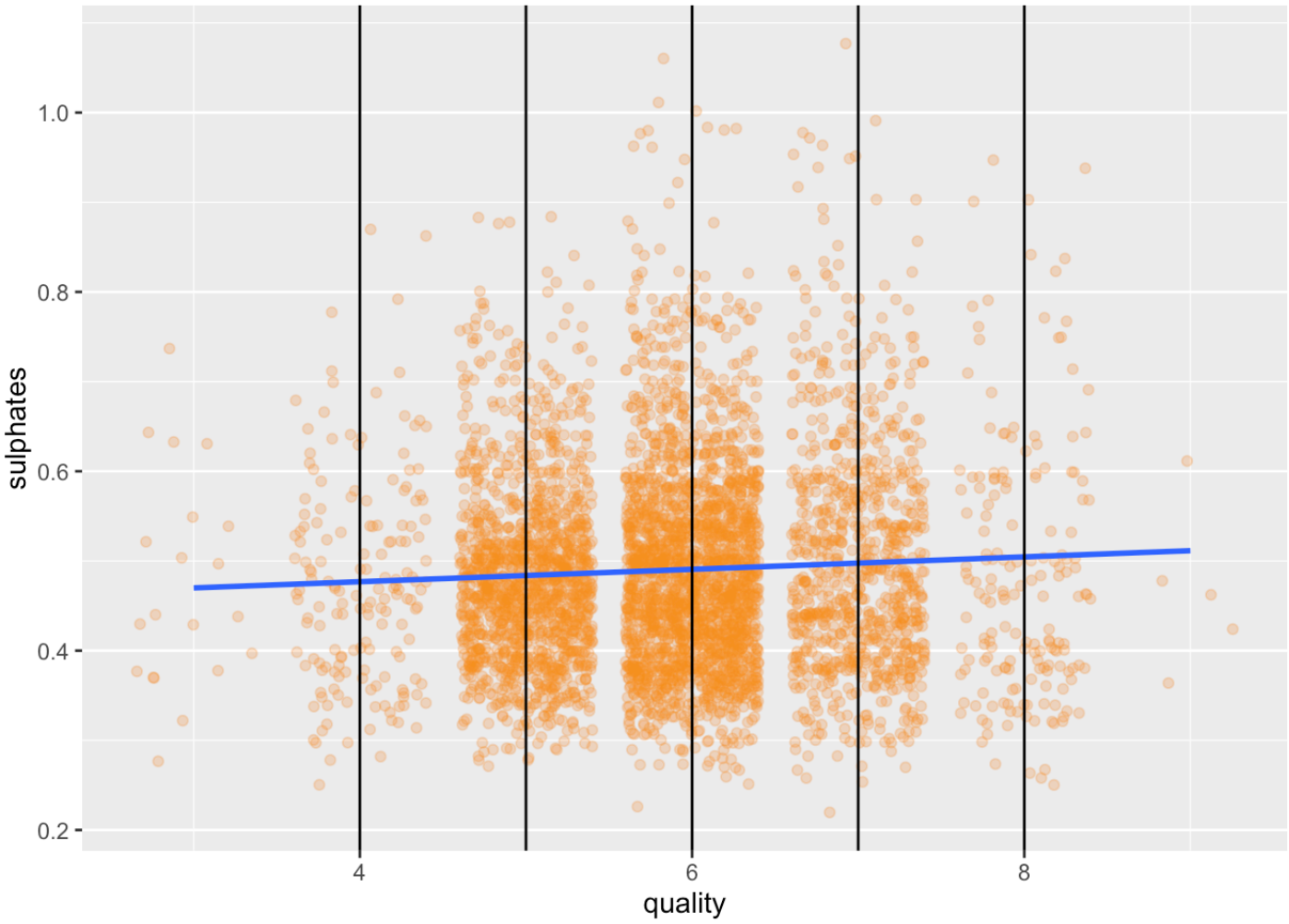

chlorides
0.3
0.2
0.1
0.0
3
4
5
6
7
8
9
quality


total.sulfur.dioxide
400
300
200
100
0
3
4
5
6
7
8
9
quality

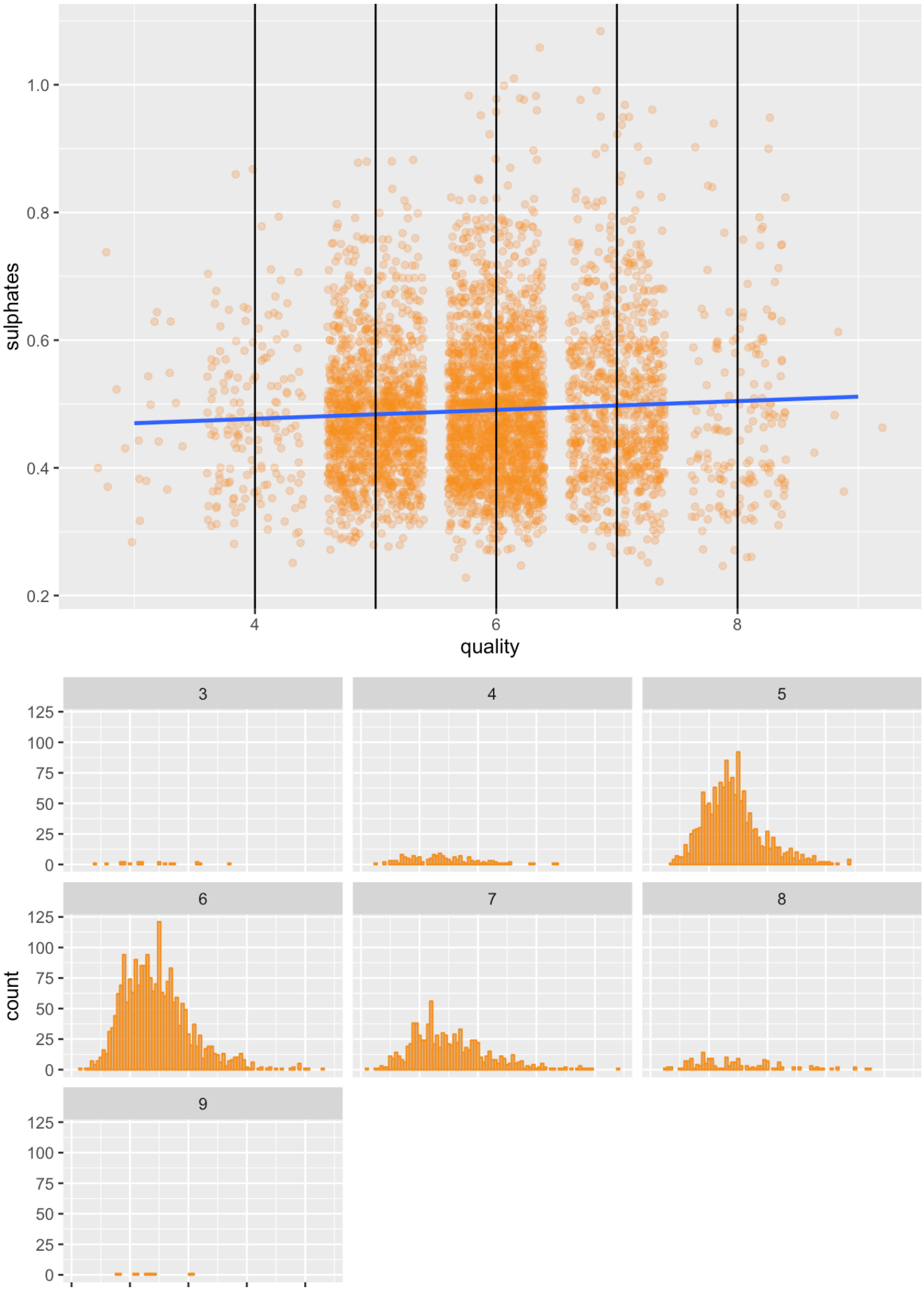

300
200
100
0
free.sulfur.dioxide
3
4
5
6
7
8
9
quality


sulphates
1.0
0.8
0.6
0.4
0.2
3
4
5
6
7
8
9
quality

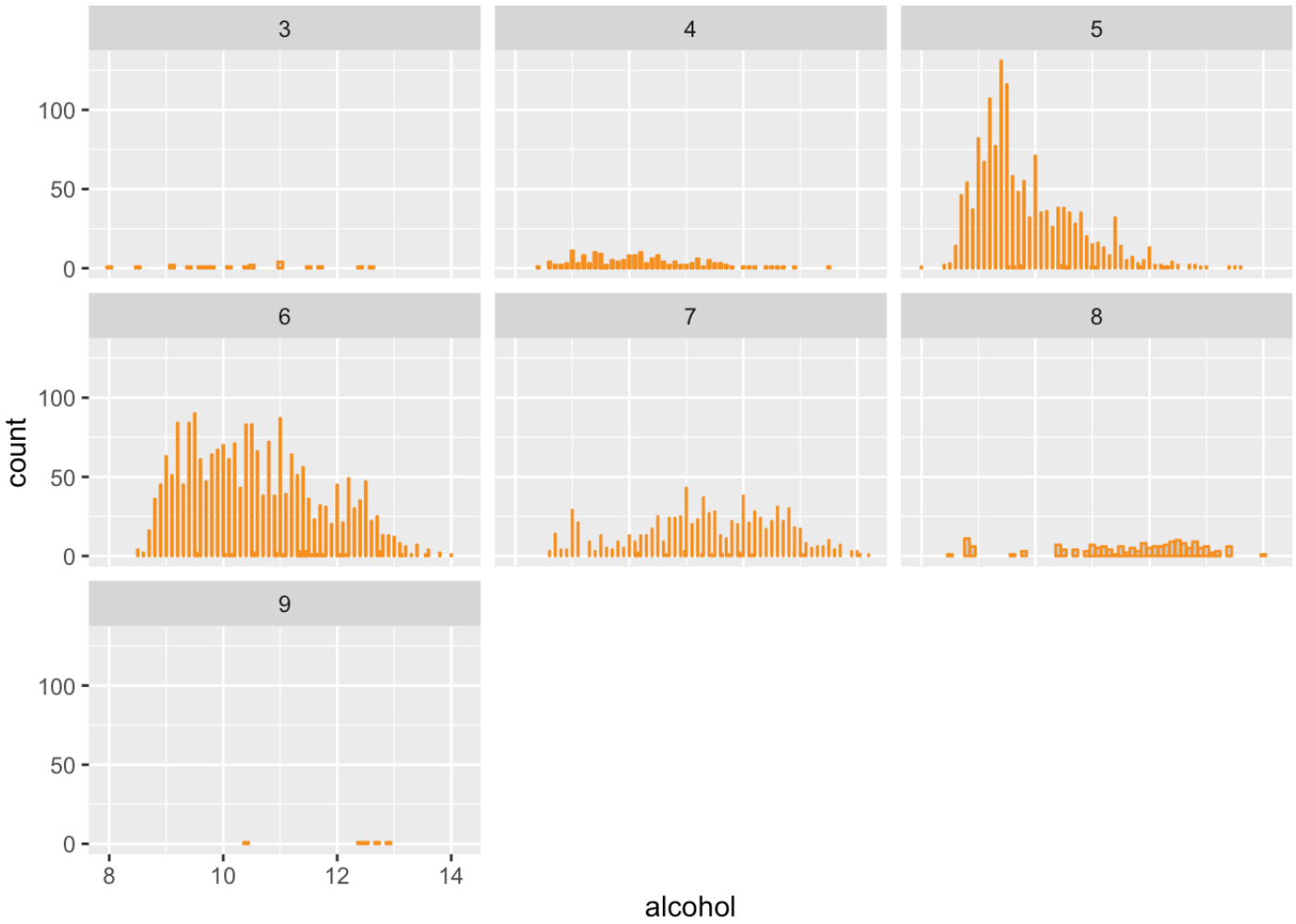

residual.sugar
60
40
20
0
4
6
8
quality

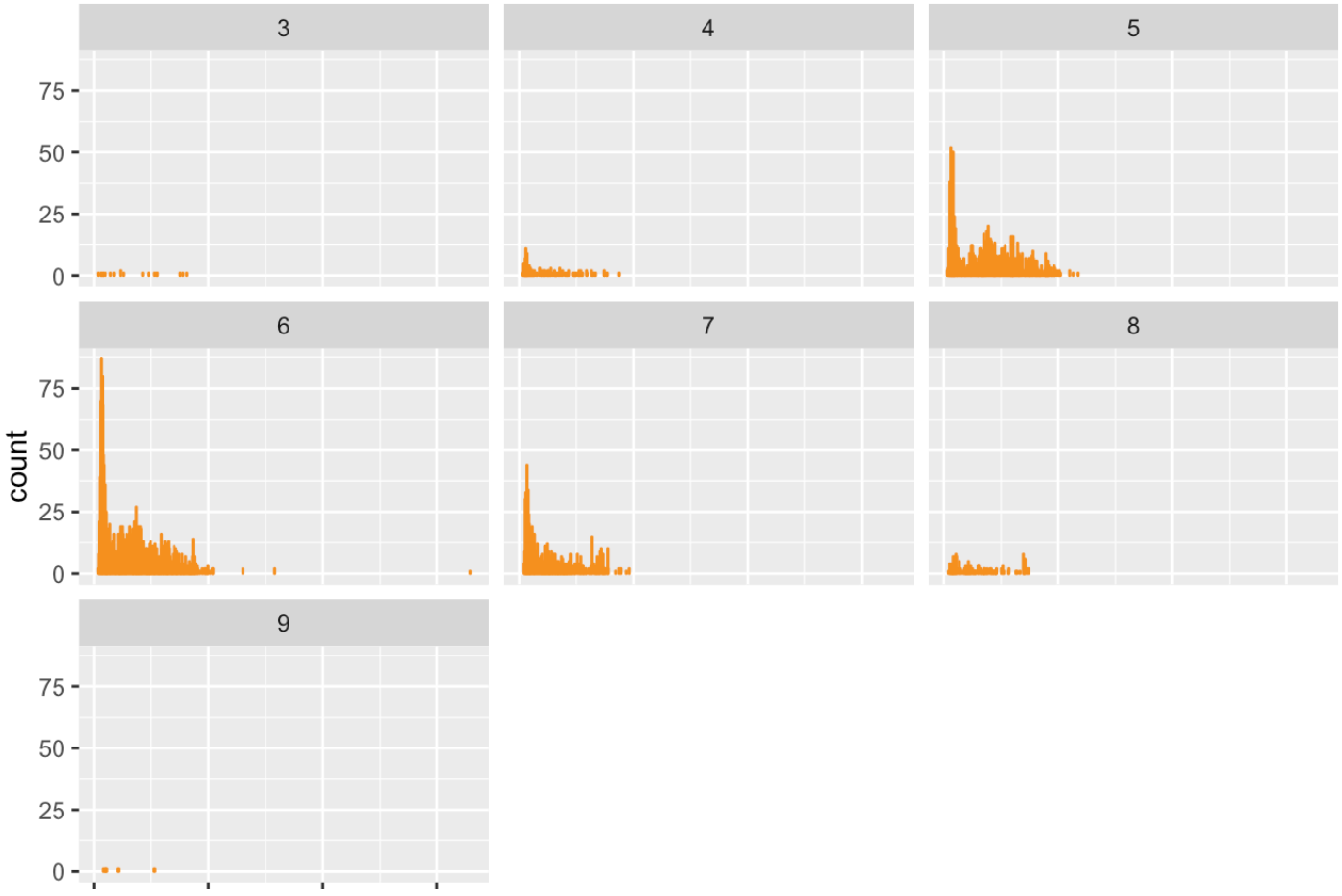


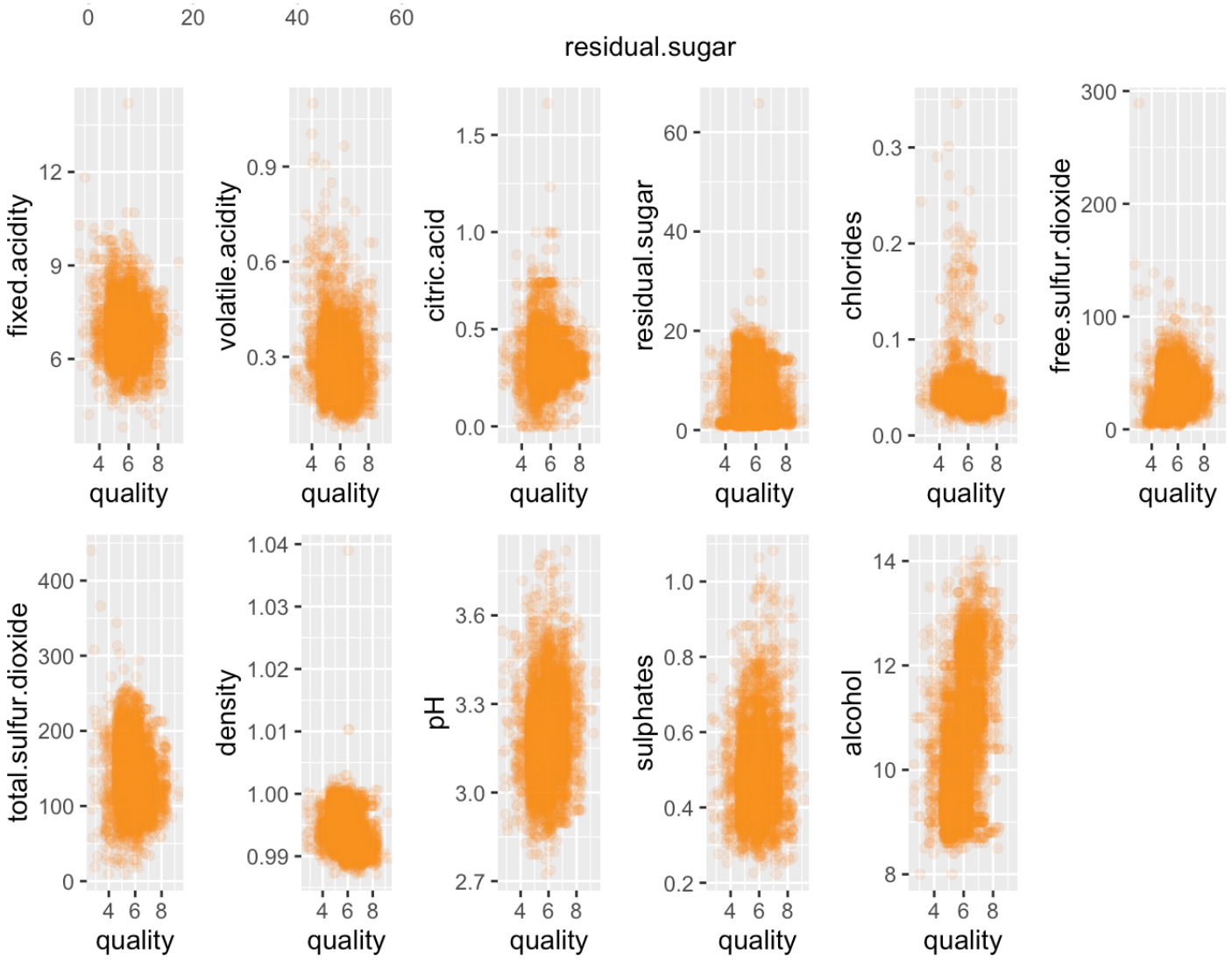

```
## [1] 0.05367788
```


sulphates
1.0
0.8
0.6
0.4
0.2
4
6
8
quality
3
4
5
6
7
8
9
count
125
100
75
50
25
0

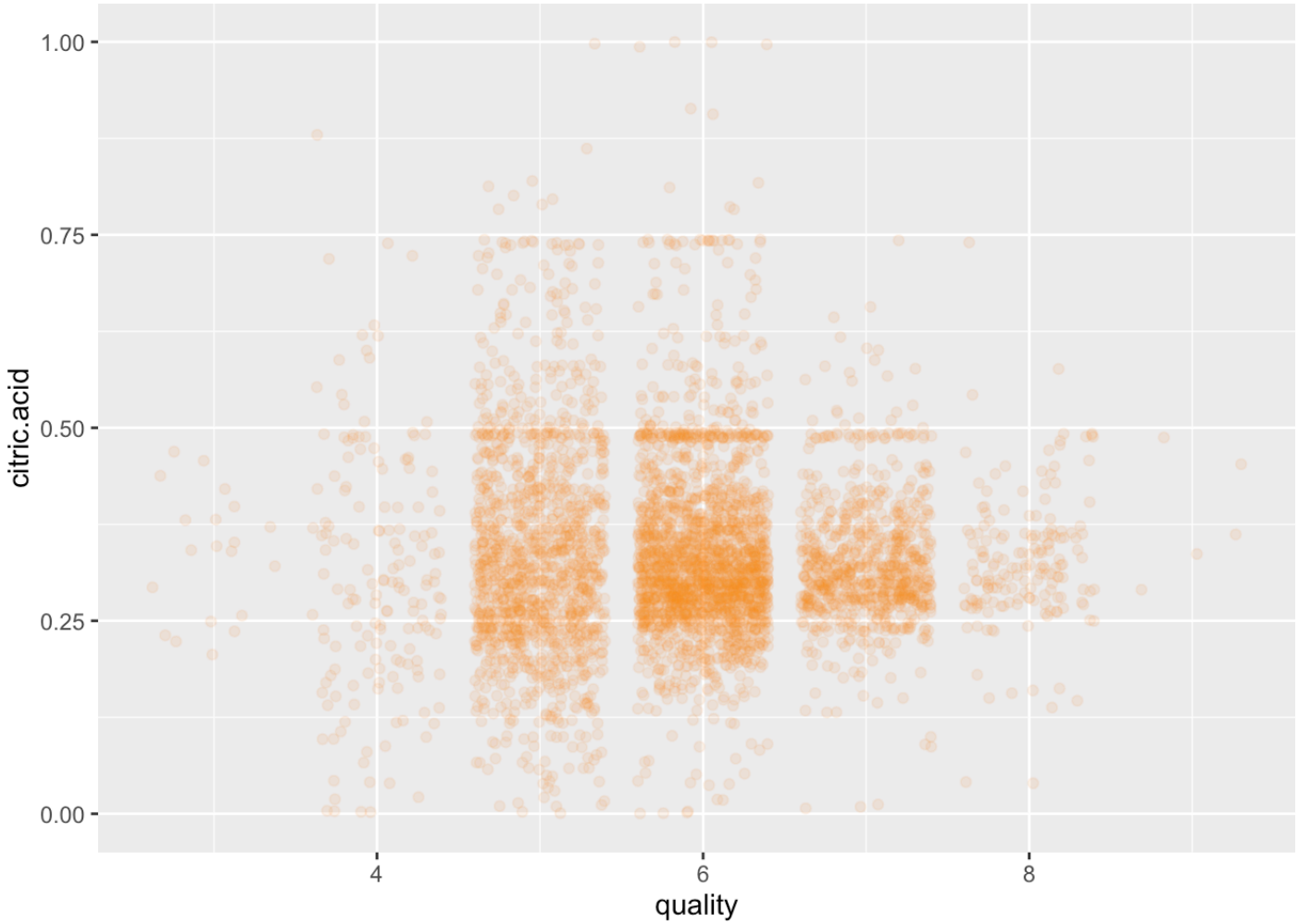

```
## [1] 0.05367788
```


3
4
5
6
7
8
9
100
50
0
count
8
10
12
14
alcohol

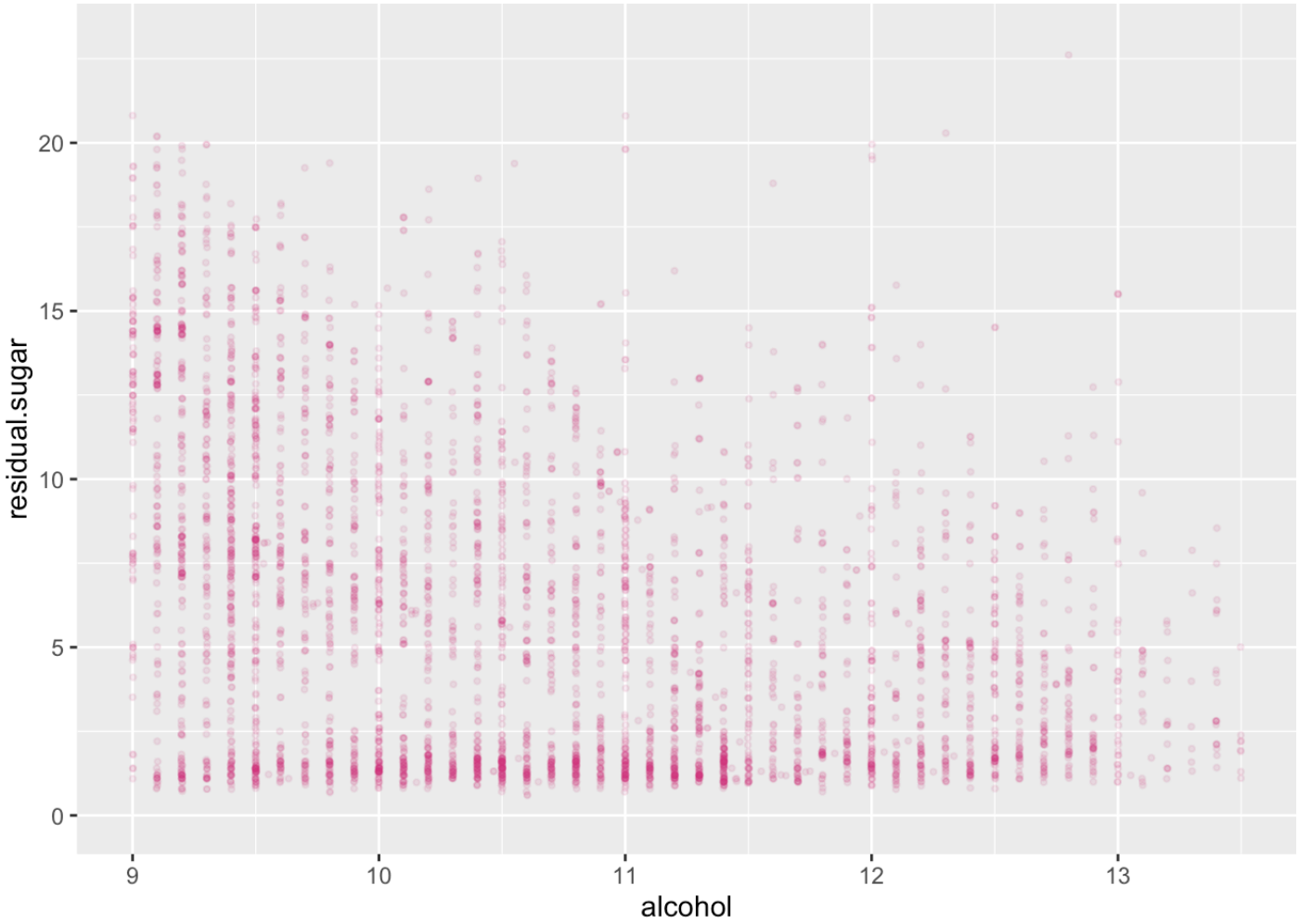

3
4
5
6
7
8
9
75
50
25
0
count



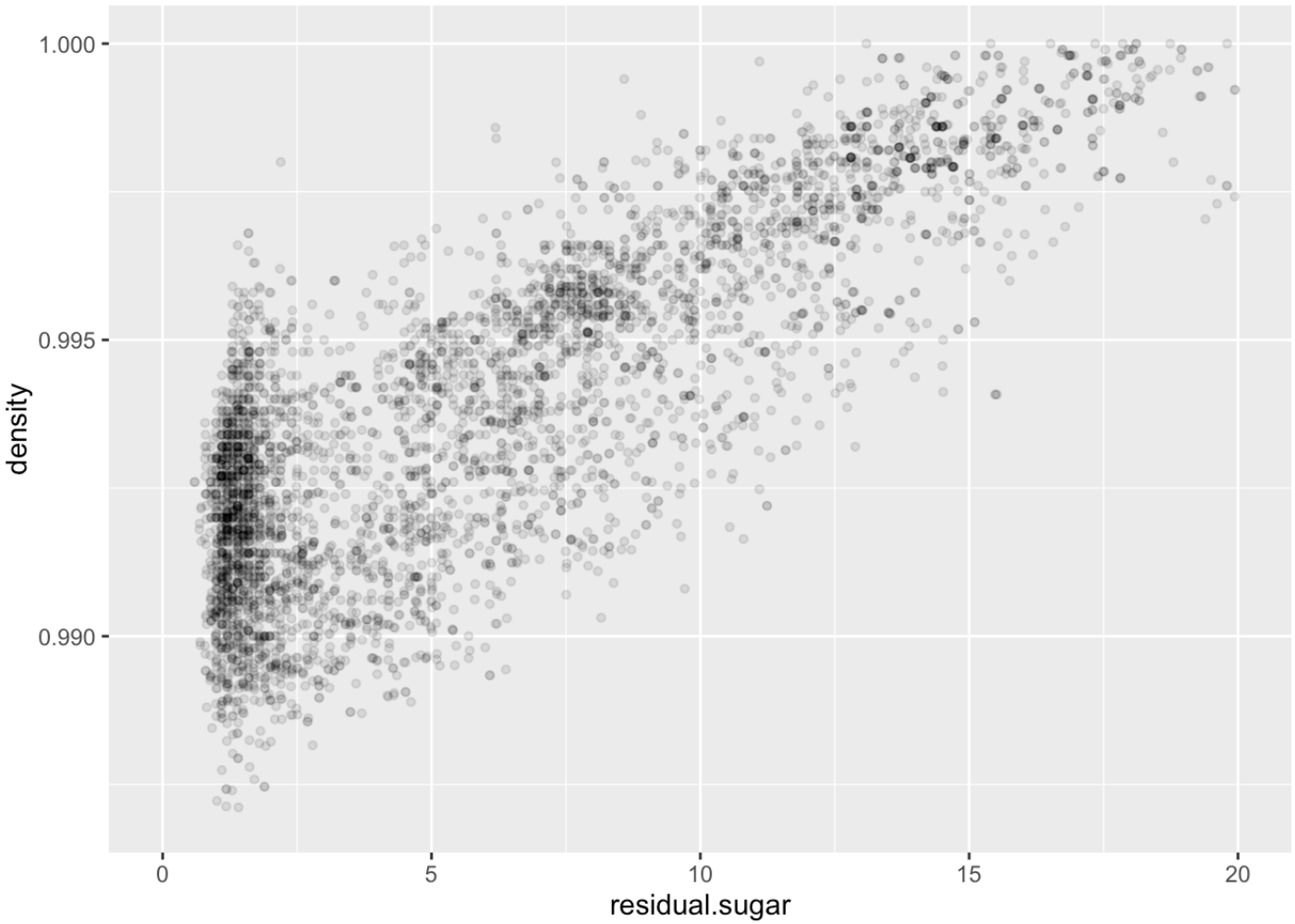

```
## Warning: Removed 12 rows containing missing values (geom_point).
```



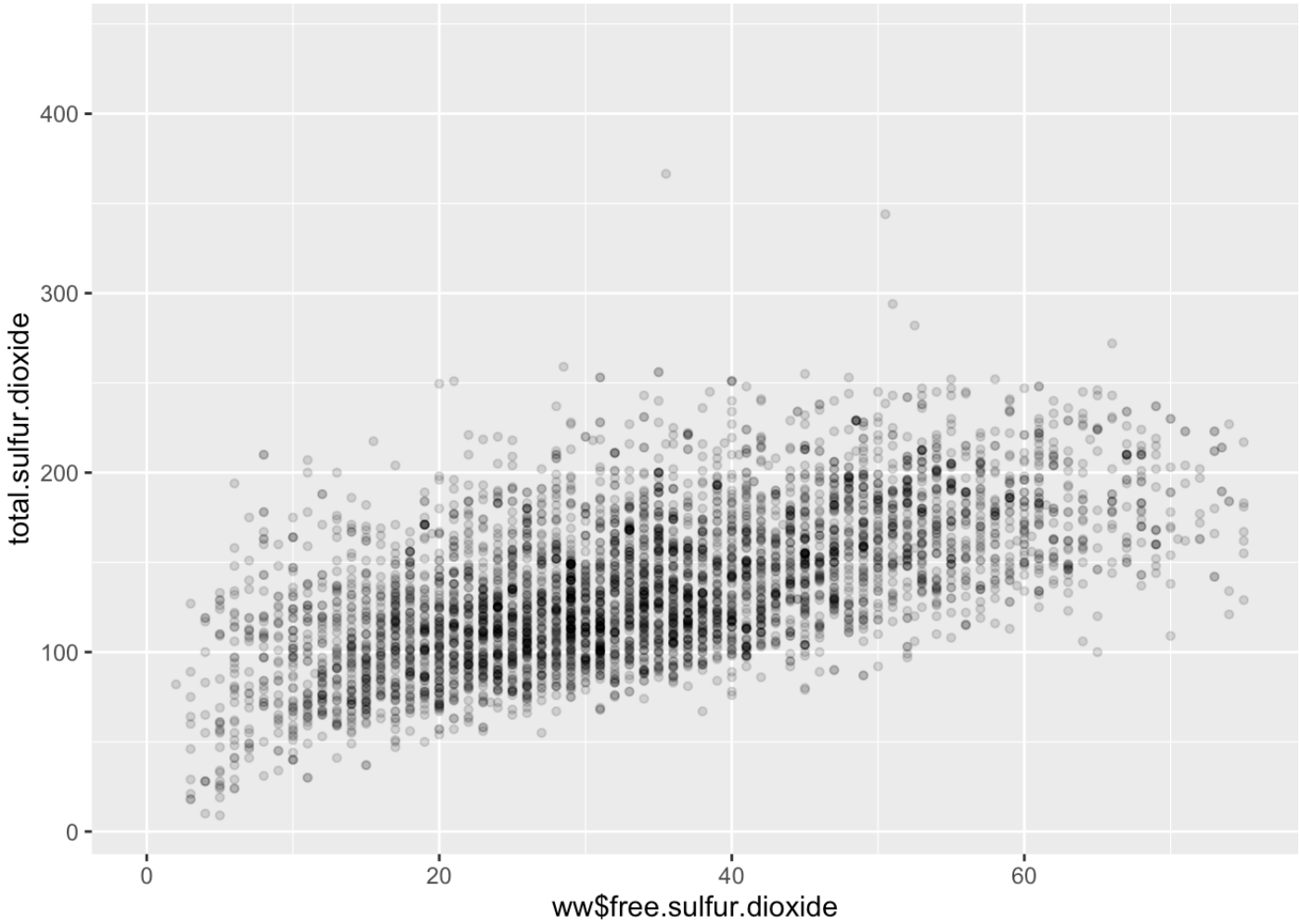

```
## Warning: Removed 443 rows containing missing values (geom_point).
```



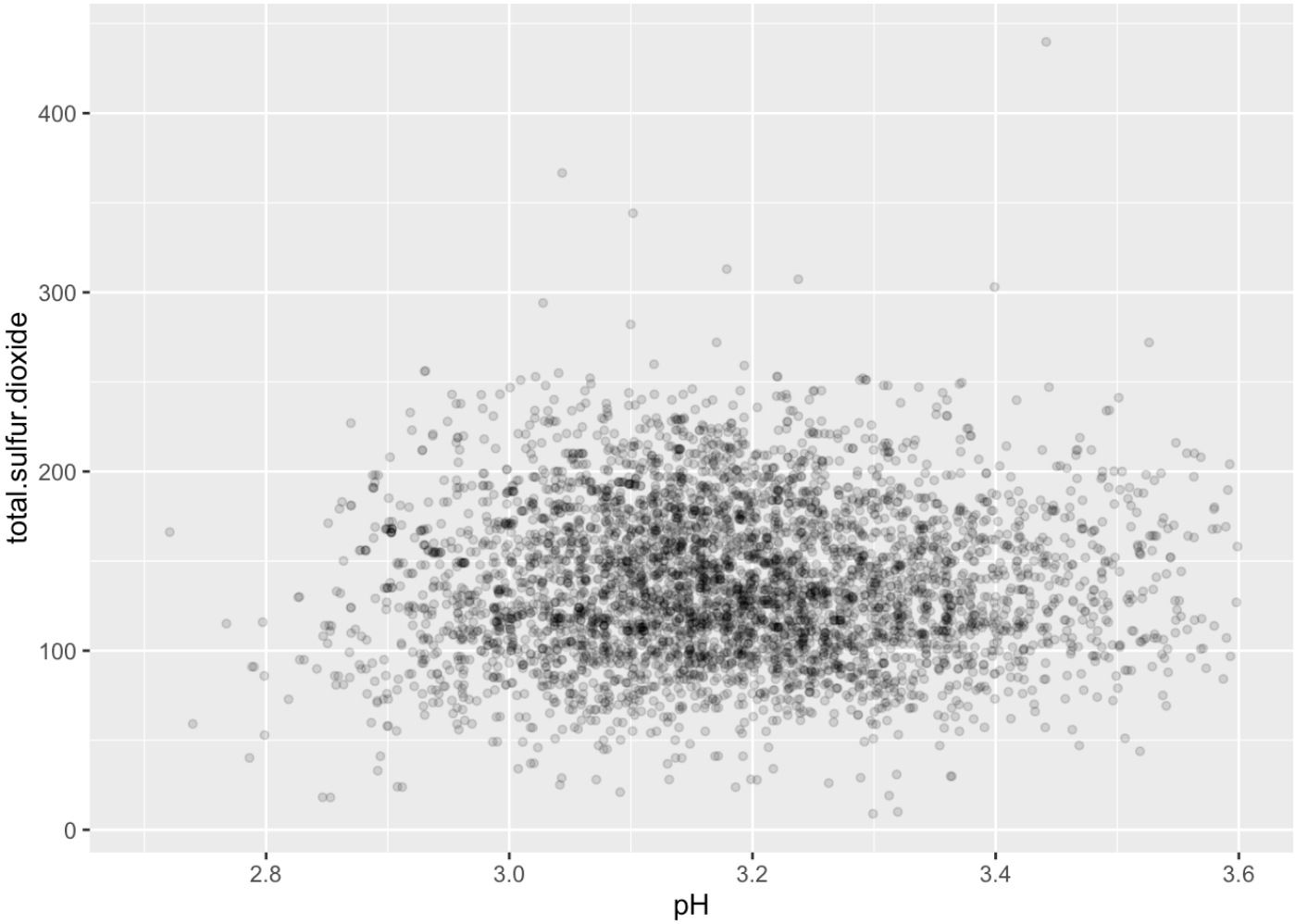

```
## Warning: Removed 91 rows containing missing values (geom_point).
```



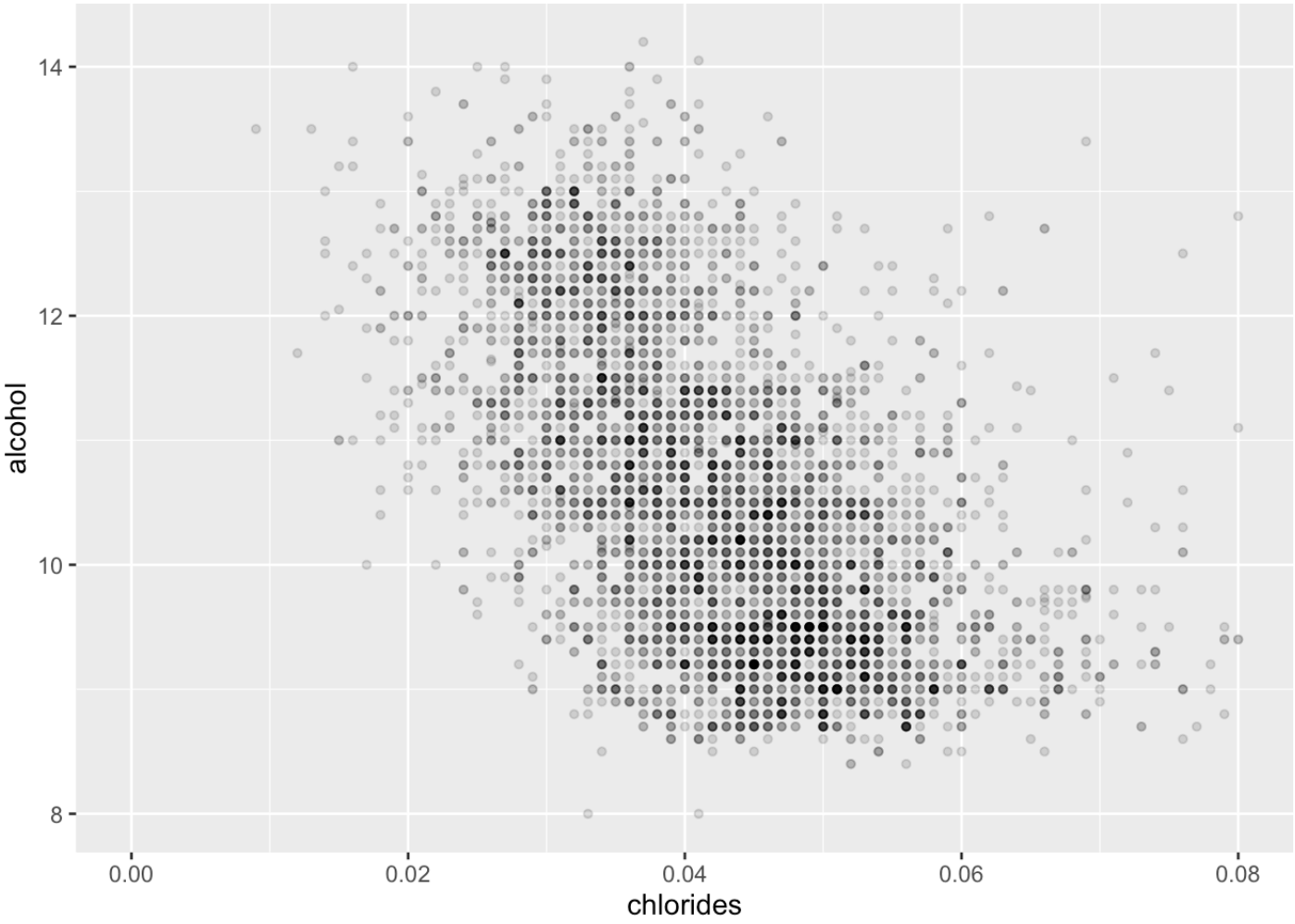

```
## Warning: Removed 72 rows containing missing values (geom_point).
```



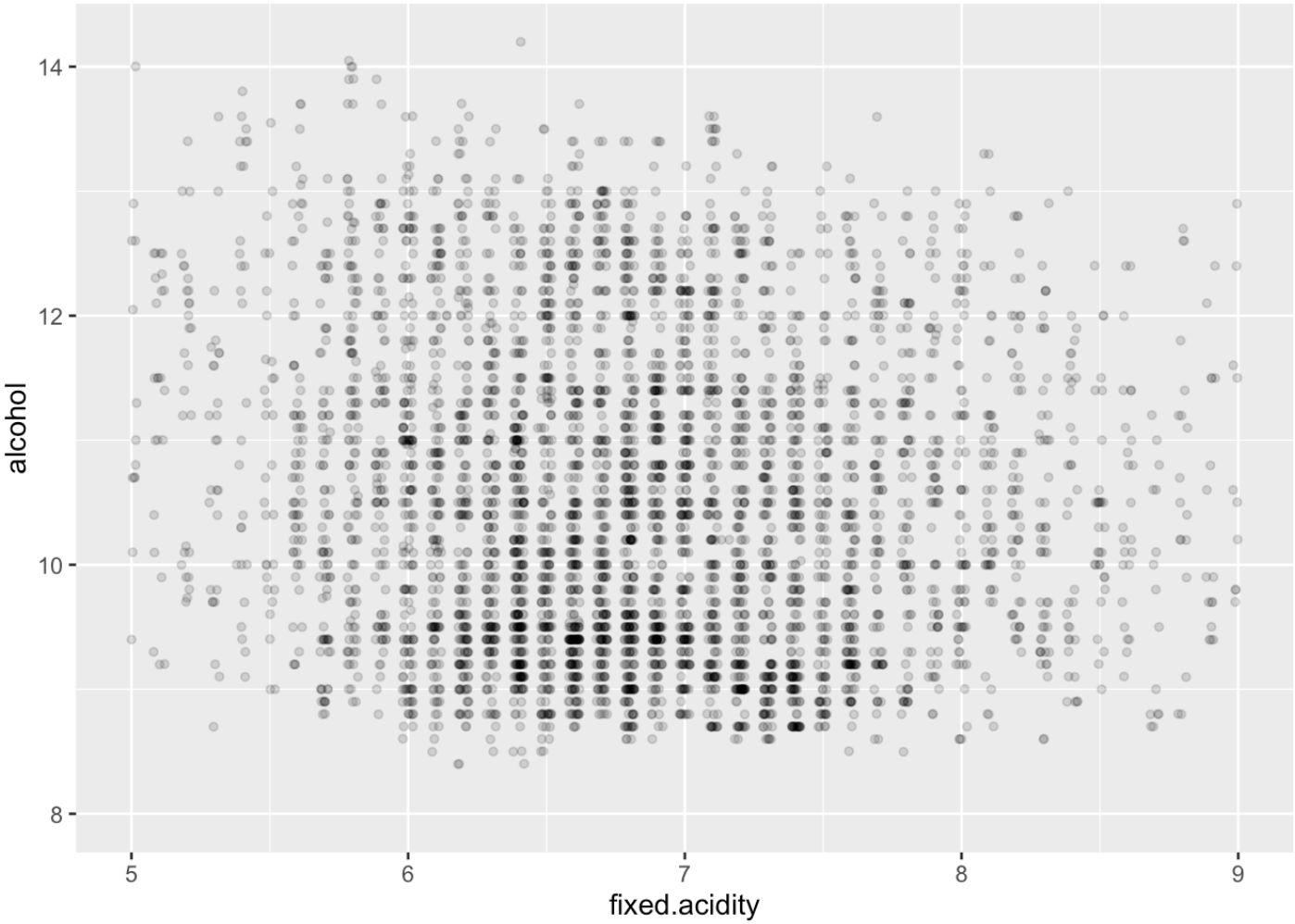

```
## Warning: Removed 48 rows containing missing values (geom_point).
```



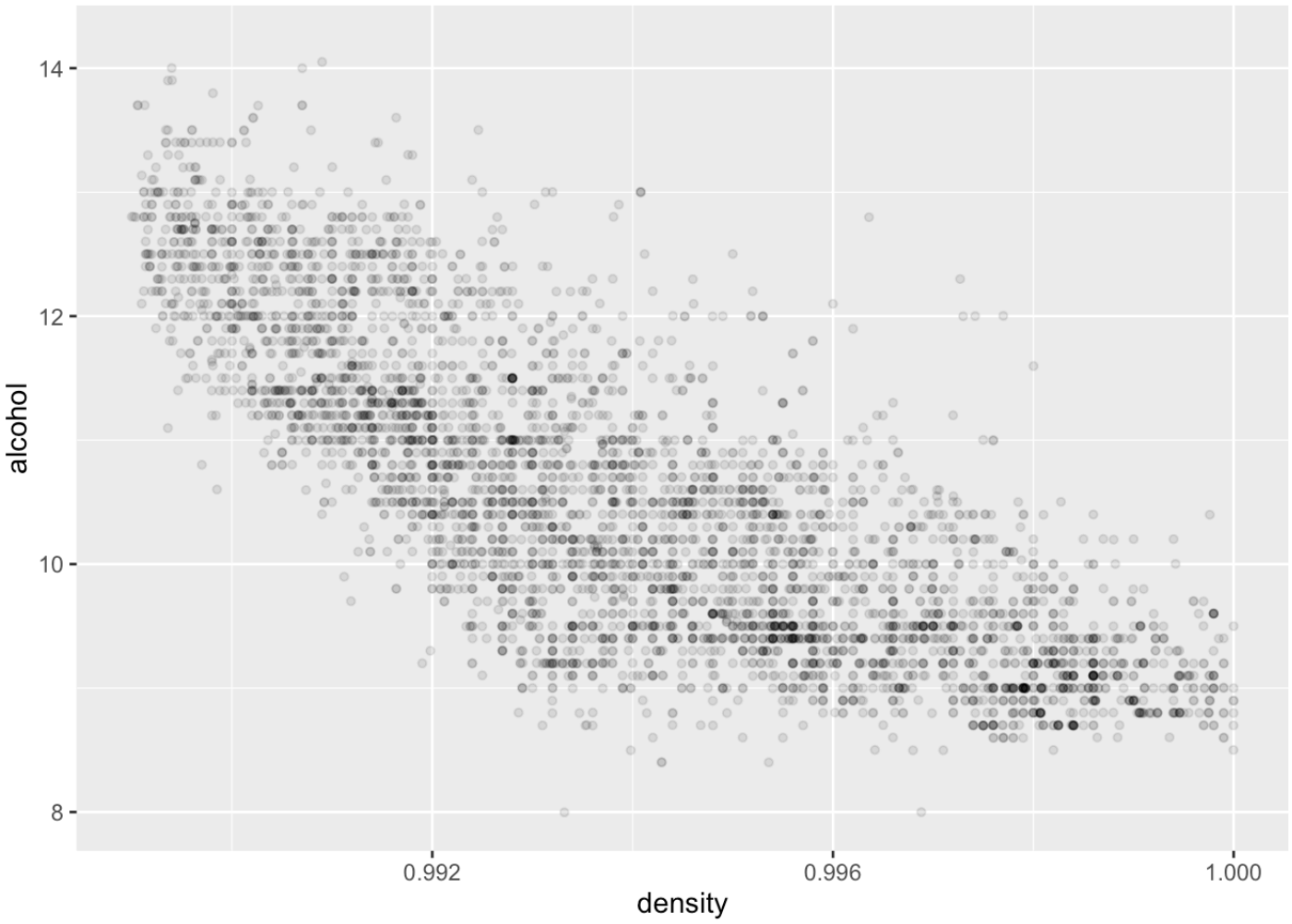

```
## Warning: Removed 164 rows containing missing values (geom_point).
```



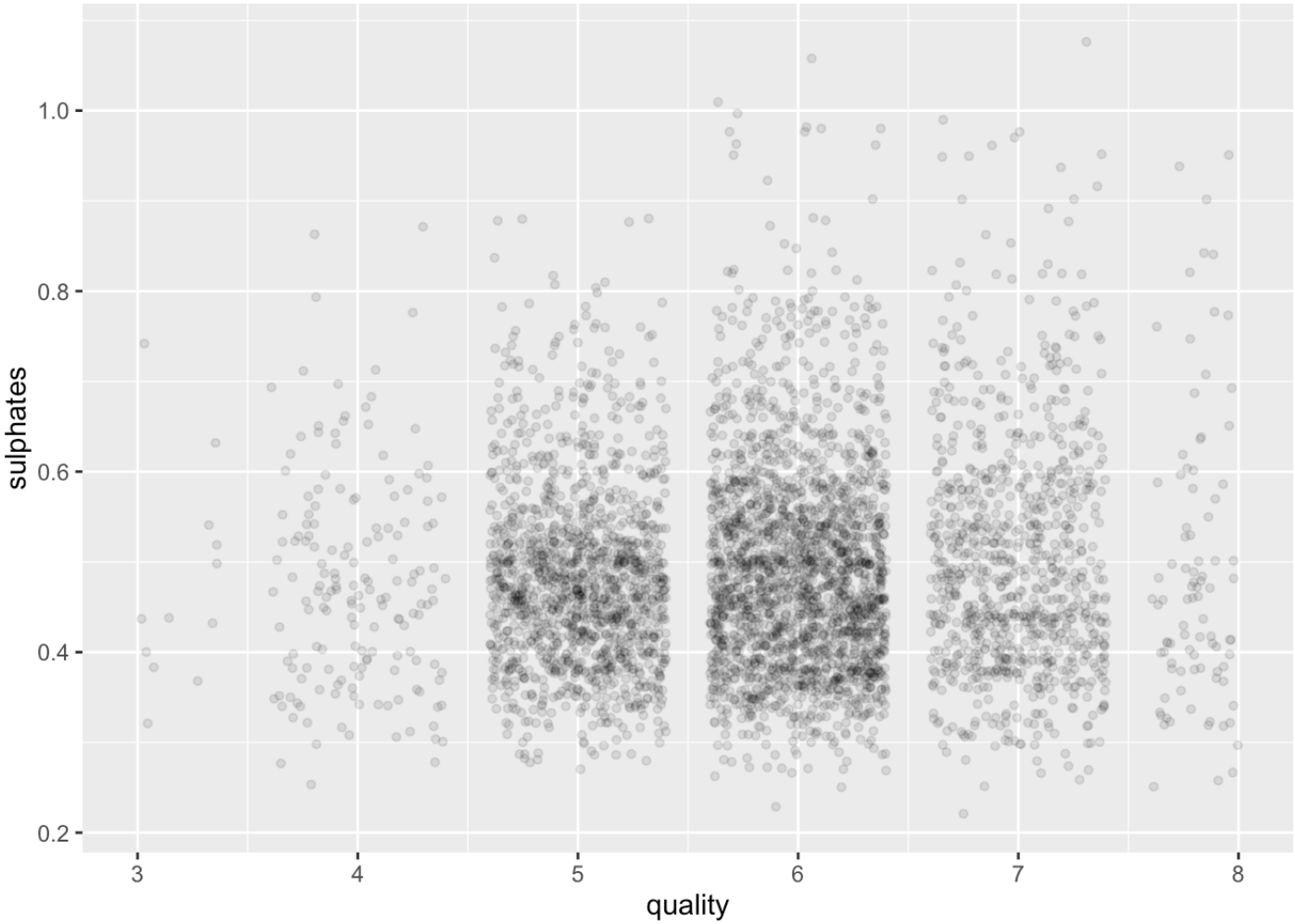

```
## Warning: Removed 120 rows containing missing values (geom_point).
```




```
## Warning: Removed 168 rows containing missing values (geom_point).
```




```
## Warning: Removed 97 rows containing missing values (geom_point).
```



#Cluster Analysis

```
wine_chemicals <- select(ww_without_x, -quality)
head(wine_chemicals)
```

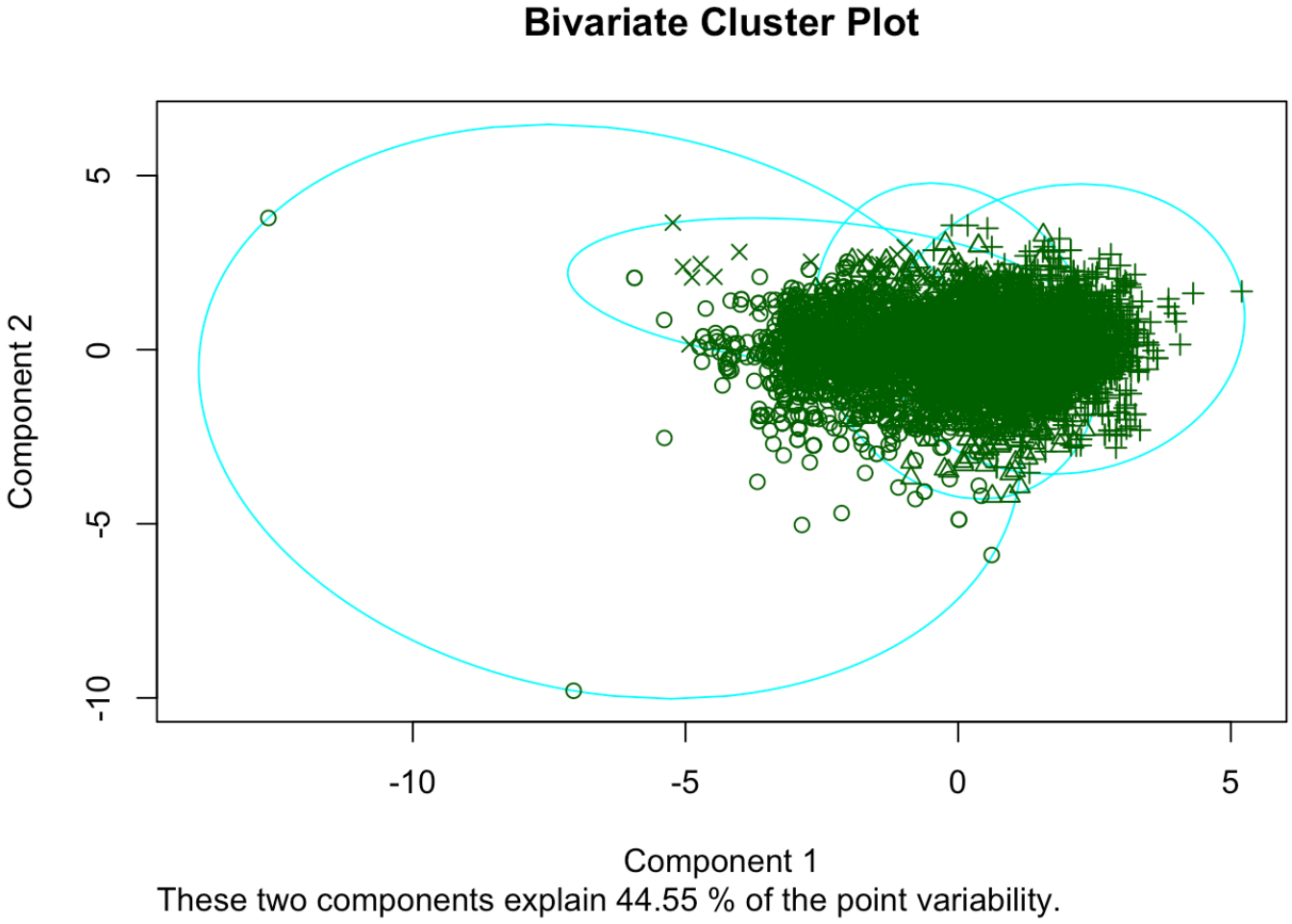
```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.0              0.27        0.36           20.7     0.045
## 2           6.3              0.30        0.34            1.6     0.049
## 3           8.1              0.28        0.40            6.9     0.050
## 4           7.2              0.23        0.32            8.5     0.058
## 5           7.2              0.23        0.32            8.5     0.058
## 6           8.1              0.28        0.40            6.9     0.050
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  45                  170  1.0010 3.00      0.45     8.8
## 2                  14                  132  0.9940 3.30      0.49     9.5
## 3                  30                   97  0.9951 3.26      0.44    10.1
## 4                  47                  186  0.9956 3.19      0.40     9.9
## 5                  47                  186  0.9956 3.19      0.40     9.9
## 6                  30                   97  0.9951 3.26      0.44    10.1
```

```
wine <- wine_chemicals

# pam (partioning around medoids)
set.seed(794)
fit.pam <- pam(wine[-1], k=4, stand=TRUE)
fit.pam$medoids
```

```
##      volatile.acidity citric.acid residual.sugar chlorides
## [1,]             0.24        0.38            8.3     0.045
## [2,]             0.23        0.34            4.0     0.047
## [3,]             0.30        0.33            3.5     0.033
## [4,]             0.29        0.49            1.4     0.142
##      free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates
## [1,]                  50                  185 0.99578 3.15      0.50
## [2,]                  24                  128 0.99440 3.20      0.52
## [3,]                  25                  116 0.99057 3.20      0.44
## [4,]                  52                  148 0.99370 3.08      0.49
##      alcohol
## [1,]     9.5
## [2,]     9.7
## [3,]    11.7
## [4,]     9.0
```

```
clusplot(fit.pam, main="Bivariate Cluster Plot")
```

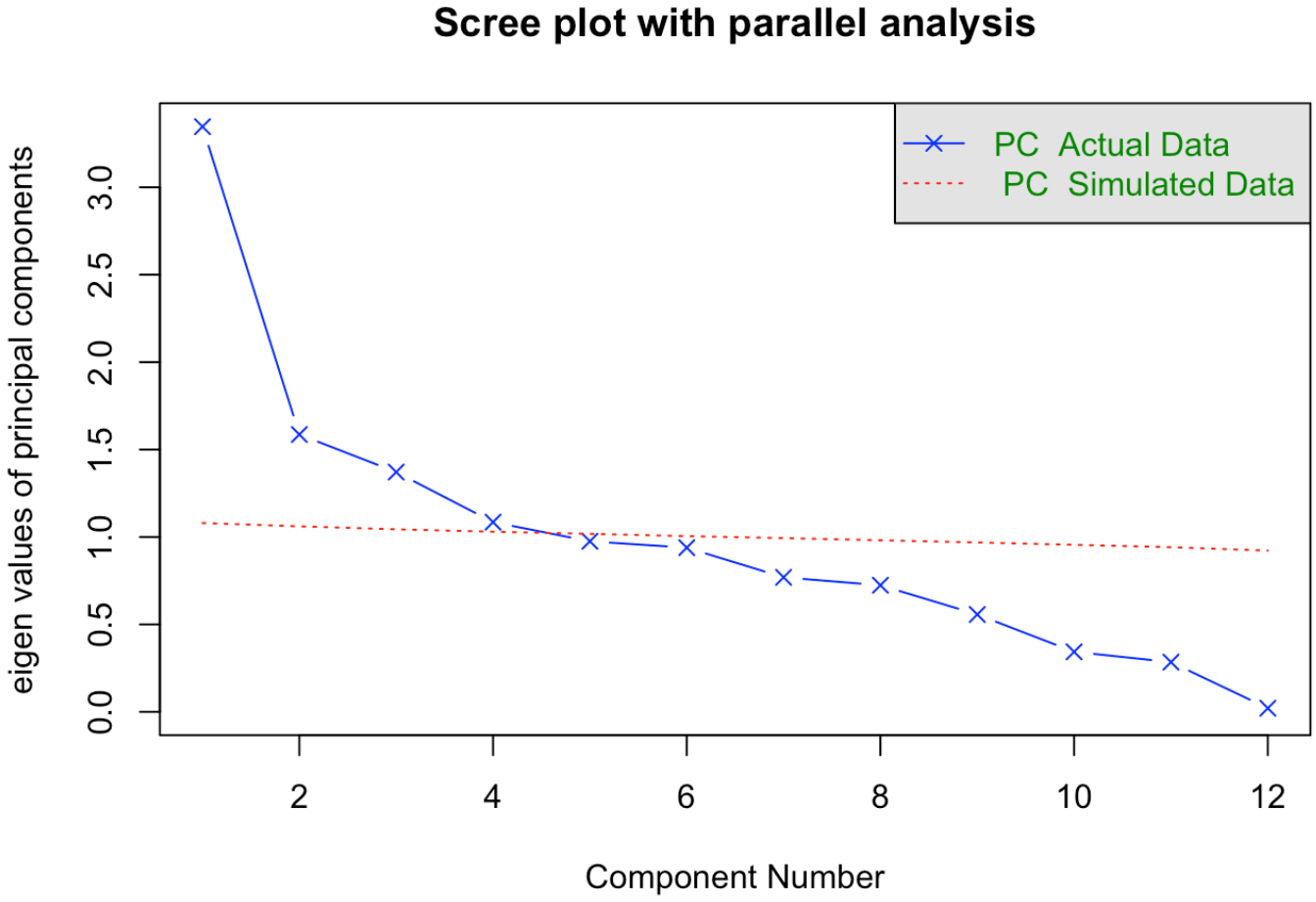


PCA

```
ww_without_x$quality <- as.integer(ww_without_x$quality)

ww_without_x.cor <- cor(ww_without_x)

fa.parallel(ww_without_x.cor, n.obs=4898, fa="pc", n.iter=100,
            show.legend=TRUE, main="Scree plot with parallel analysis")
```




```
## Parallel analysis suggests that the number of factors =  NA  and the number of c
omponents =  4
```

```
pc <- principal(ww_without_x.cor, nfactors=4, rotate="varimax")
pc
```

```
## Principal Components Analysis
## Call: principal(r = ww_without_x.cor, nfactors = 4, rotate = "varimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                        RC1   RC2   RC4   RC3   h2   u2 com
## fixed.acidity         0.04  0.78  0.10  0.14 0.64 0.36 1.1
## volatile.acidity     -0.04 -0.10  0.30 -0.55 0.40 0.60 1.6
## citric.acid           0.07  0.48  0.07  0.58 0.58 0.42 2.0
## residual.sugar        0.81  0.22  0.01 -0.21 0.74 0.26 1.3
## chlorides             0.10  0.01  0.68  0.18 0.51 0.49 1.2
## free.sulfur.dioxide   0.70 -0.16 -0.15  0.24 0.60 0.40 1.5
## total.sulfur.dioxide  0.76 -0.09  0.17  0.17 0.65 0.35 1.2
## density               0.82  0.22  0.36 -0.05 0.85 0.15 1.5
## pH                   -0.05 -0.77 -0.01  0.15 0.62 0.38 1.1
## sulphates             0.03 -0.28  0.18  0.59 0.46 0.54 1.6
## alcohol              -0.62 -0.12 -0.56  0.00 0.71 0.29 2.1
## quality              -0.09 -0.10 -0.74  0.23 0.62 0.38 1.3
##
##                        RC1  RC2  RC4  RC3
## SS loadings           2.81 1.68 1.64 1.25
## Proportion Var        0.23 0.14 0.14 0.10
## Cumulative Var        0.23 0.37 0.51 0.62
## Proportion Explained  0.38 0.23 0.22 0.17
## Cumulative Proportion 0.38 0.61 0.83 1.00
##
## Mean item complexity =  1.5
## Test of the hypothesis that 4 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.11
##
## Fit based upon off diagonal values = 0.82
```

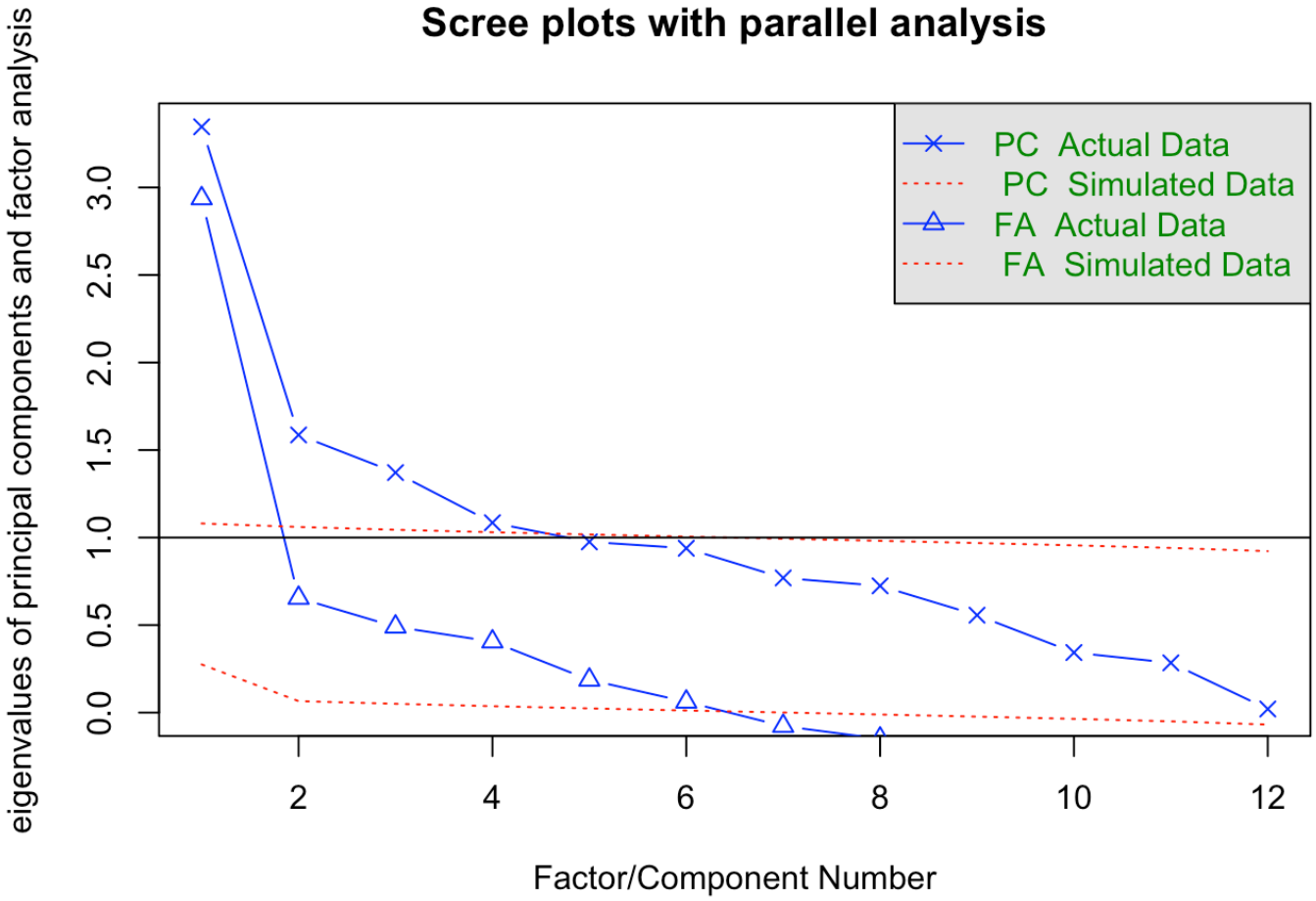
Exploratory factor analysis

```
#correlation matrix
ww_without_x.cor
```

```
##                       fixed.acidity volatile.acidity  citric.acid
## fixed.acidity            1.00000000      -0.02269729  0.289180698
## volatile.acidity        -0.02269729       1.00000000 -0.149471811
## citric.acid              0.28918070      -0.14947181  1.000000000
## residual.sugar           0.08902070       0.06428606  0.094211624
## chlorides                0.02308564       0.07051157  0.114364448
## free.sulfur.dioxide     -0.04939586      -0.09701194  0.094077221
## total.sulfur.dioxide     0.09106976       0.08926050  0.121130798
## density                  0.26533101       0.02711385  0.149502571
## pH                      -0.42585829      -0.03191537 -0.163748211
## sulphates               -0.01714299      -0.03572815  0.062330940
```

```
## alcohol                -0.12088112       0.06771794 -0.075728730
## quality                -0.11366283      -0.19472297 -0.009209091
##                      residual.sugar   chlorides free.sulfur.dioxide
## fixed.acidity            0.08902070  0.02308564       -0.0493958591
## volatile.acidity         0.06428606  0.07051157       -0.0970119393
## citric.acid              0.09421162  0.11436445        0.0940772210
## residual.sugar           1.00000000  0.08868454        0.2990983537
## chlorides                0.08868454  1.00000000        0.1013923521
## free.sulfur.dioxide      0.29909835  0.10139235        1.0000000000
## total.sulfur.dioxide     0.40143931  0.19891030        0.6155009650
## density                  0.83896645  0.25721132        0.2942104109
## pH                      -0.19413345 -0.09043946       -0.0006177961
## sulphates               -0.02666437  0.01676288        0.0592172458
## alcohol                 -0.45063122 -0.36018871       -0.2501039415
## quality                 -0.09757683 -0.20993441        0.0081580671
##                      total.sulfur.dioxide     density            pH
## fixed.acidity                 0.091069756  0.26533101 -0.4258582910
## volatile.acidity              0.089260504  0.02711385 -0.0319153683
## citric.acid                   0.121130798  0.14950257 -0.1637482114
## residual.sugar                0.401439311  0.83896645 -0.1941334540
## chlorides                     0.198910300  0.25721132 -0.0904394560
## free.sulfur.dioxide           0.615500965  0.29421041 -0.0006177961
## total.sulfur.dioxide          1.000000000  0.52988132  0.0023209718
## density                       0.529881324  1.00000000 -0.0935914935
## pH                            0.002320972 -0.09359149  1.0000000000
## sulphates                     0.134562367  0.07449315  0.1559514973
## alcohol                      -0.448892102 -0.78013762  0.1214320987
## quality                      -0.174737218 -0.30712331  0.0994272457
##                        sulphates     alcohol      quality
## fixed.acidity        -0.01714299 -0.12088112 -0.113662831
## volatile.acidity     -0.03572815  0.06771794 -0.194722969
## citric.acid           0.06233094 -0.07572873 -0.009209091
## residual.sugar       -0.02666437 -0.45063122 -0.097576829
## chlorides             0.01676288 -0.36018871 -0.209934411
## free.sulfur.dioxide   0.05921725 -0.25010394  0.008158067
## total.sulfur.dioxide  0.13456237 -0.44889210 -0.174737218
## density               0.07449315 -0.78013762 -0.307123313
## pH                    0.15595150  0.12143210  0.099427246
## sulphates             1.00000000 -0.01743277  0.053677877
## alcohol              -0.01743277  1.00000000  0.435574715
## quality               0.05367788  0.43557472  1.000000000
```

```
fa.parallel(ww_without_x.cor, n.obs=4898, fa="both", n.iter=1000, main="Scree plots
with parallel analysis")
```




```
## Parallel analysis suggests that the number of factors =  6  and the number of co
mponents =  4
```

```
# Parallel analysis suggests that the number of factors =  6  and the number of com
ponents =  4

fa.promax <- fa(ww_without_x, nfactors=6, rotate="promax", fm="minres")
```

```
## Warning in fac(r = r, nfactors = nfactors, n.obs = n.obs, rotate =
## rotate, : A Heywood case was detected. Examine the loadings carefully.
```

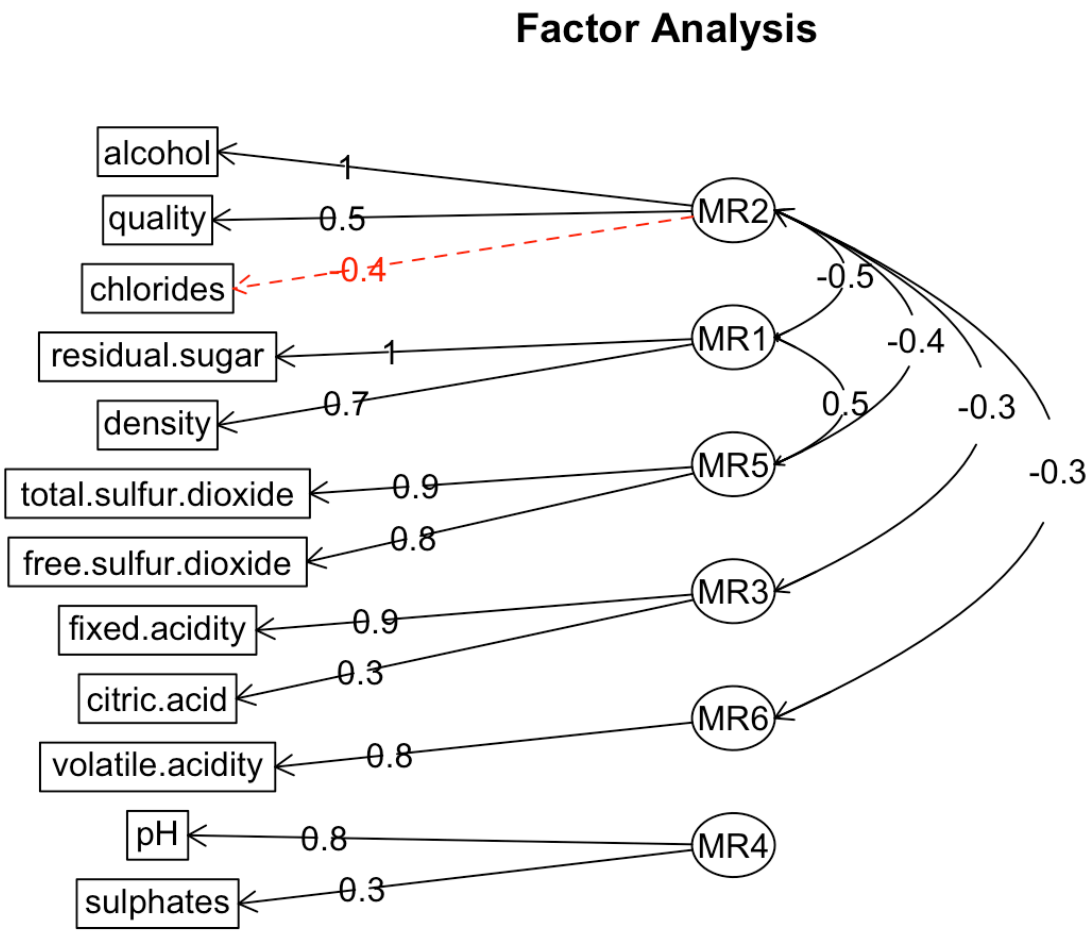
```
fa.promax
```

```
## Factor Analysis using method =  minres
## Call: fa(r = ww_without_x, nfactors = 6, rotate = "promax", fm = "minres")
##
##  Warning: A Heywood case was detected.
## Standardized loadings (pattern matrix) based upon correlation matrix
##                       MR2   MR1   MR5   MR3   MR6   MR4   h2     u2 com
## fixed.acidity        0.10 -0.04 -0.04  0.94  0.00 -0.12 0.87 0.1318 1.1
```

```
## volatile.acidity      0.23  0.06  0.06  0.00  0.80 -0.02 0.57 0.4344 1.2
## citric.acid           0.04  0.02  0.10  0.32 -0.16 -0.03 0.15 0.8520 1.8
## residual.sugar        0.05  1.03  0.01 -0.10  0.06 -0.09 1.00 0.0050 1.0
## chlorides            -0.38 -0.10  0.08  0.00  0.07  0.00 0.16 0.8369 1.3
## free.sulfur.dioxide   0.02  0.00  0.77 -0.13 -0.11 -0.09 0.57 0.4261 1.1
## total.sulfur.dioxide -0.05  0.00  0.85  0.06  0.15  0.04 0.82 0.1819 1.1
## density              -0.40  0.67  0.00  0.15 -0.01  0.14 1.00 0.0050 1.9
## pH                    0.07 -0.03 -0.03 -0.17 -0.02  0.76 0.68 0.3226 1.1
## sulphates             0.06 -0.03  0.11  0.13 -0.02  0.32 0.12 0.8810 1.7
## alcohol               1.04 -0.05 -0.02  0.10  0.20  0.06 0.99 0.0082 1.1
## quality               0.48  0.11  0.04 -0.05 -0.23  0.03 0.32 0.6790 1.6
##
##                        MR2  MR1  MR5  MR3  MR6  MR4
## SS loadings           1.69 1.62 1.37 1.07 0.72 0.76
## Proportion Var        0.14 0.14 0.11 0.09 0.06 0.06
## Cumulative Var        0.14 0.28 0.39 0.48 0.54 0.60
## Proportion Explained  0.23 0.22 0.19 0.15 0.10 0.11
## Cumulative Proportion 0.23 0.46 0.65 0.79 0.89 1.00
##
##  With factor correlations of
##       MR2   MR1   MR5   MR3   MR6   MR4
## MR2  1.00 -0.46 -0.39 -0.34 -0.31 -0.08
## MR1 -0.46  1.00  0.46  0.26  0.04 -0.04
## MR5 -0.39  0.46  1.00  0.17 -0.02  0.14
## MR3 -0.34  0.26  0.17  1.00  0.03 -0.21
## MR6 -0.31  0.04 -0.02  0.03  1.00  0.01
## MR4 -0.08 -0.04  0.14 -0.21  0.01  1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 6 factors are sufficient.
##
## The degrees of freedom for the null model are  66  and the objective function wa
s  5.41 with Chi Square of  26470.7
## The degrees of freedom for the model are 9  and the objective function was  0.08
##
## The root mean square of the residuals (RMSR) is  0.02
## The df corrected root mean square of the residuals is  0.05
##
## The harmonic number of observations is  4898 with the empirical chi square  261.
84  with prob <  3.1e-51
## The total number of observations was  4898  with MLE Chi Square =  406.46  with
prob <  5.7e-82
##
## Tucker Lewis Index of factoring reliability =  0.89
## RMSEA index =  0.095  and the 90 % confidence intervals are  0.087 0.103
## BIC =  329.99
## Fit based upon off diagonal values = 0.99
## Measures of factor score adequacy
##                                                MR2  MR1  MR5  MR3  MR6
## Correlation of scores with factors            0.99 1.00 0.92 0.96 0.79
## Multiple R square of scores with factors      0.98 1.00 0.85 0.92 0.63
```

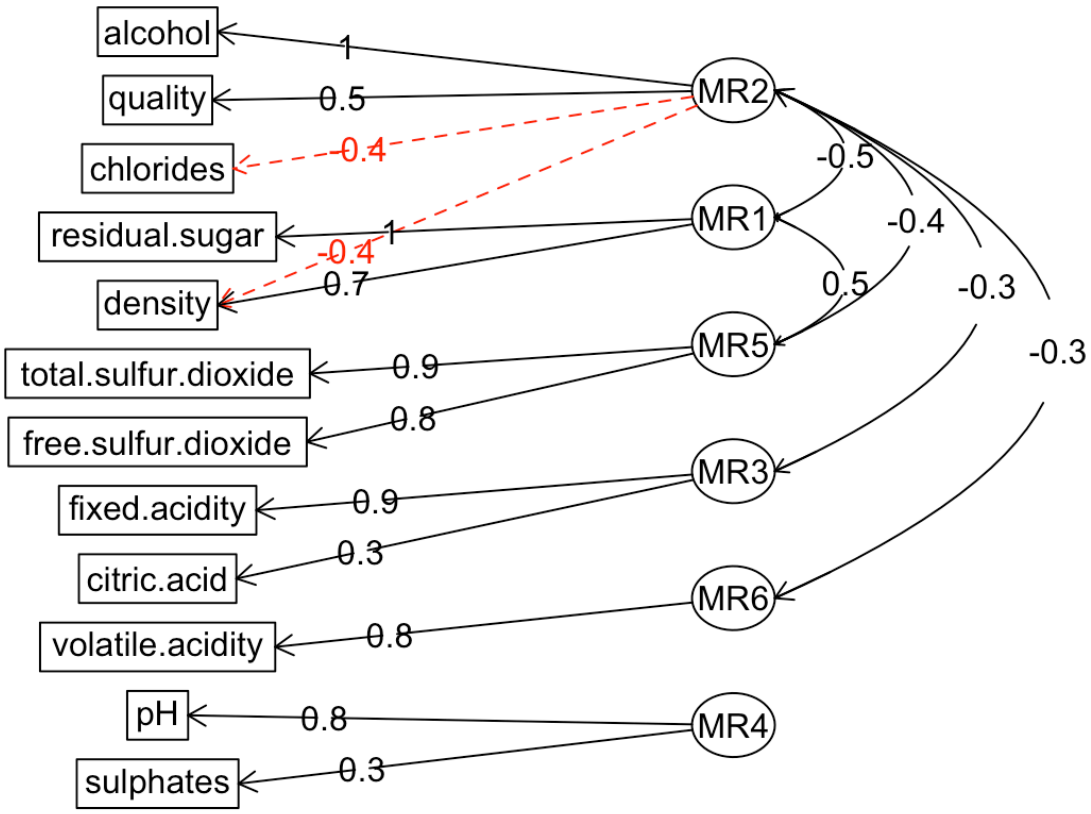
```
## Minimum correlation of possible factor scores  0.96 0.99 0.70 0.83 0.26
##                                                 MR4
## Correlation of scores with factors             0.92
## Multiple R square of scores with factors       0.85
## Minimum correlation of possible factor scores  0.71
```

```
fa.diagram(fa.promax)
```




```
fa.diagram(fa.promax, simple=FALSE)
```

Factor Analysis



Factor Analysis I choose 6 factors because 6 factors explains around 60% of the variance and still keeps the amount of factors sufficiently low. I used EFA because I was curious as to what factors would be composed of and how they might shed light later in the analysis.

Factor Notes: MR1: Notice how residual.sugar and density are loaded here against alcohol (-0.8) this is consistent with the relationship between sugar and density, as well as density and residual sugar. MR2: Notice how quality is paired with alcohol which is consistent with pairs correlation. MR3: Combines citric.acid and fixed.acidity. MR4: pH and sulphates are paired though they have a insignificant correlation coefficient. MR5: Notice how free sulfur dioxide and total sulfur dioxide are loaded at MR5 which is consistent as one is a part of the other. MR6: volatile.acidity has been left alone.

Overall 60% of the variance is explained by these factors.

Top:

Residual.Sugar x Density | 0.84 |

FSD x T.S.D | 0.62 |

T.S.D x Density | 0.53 |

Residual.Sugar x Alcohol | -0.45 |

Quality x Alcohol | 0.44 |

Fixed.Acidity x pH | -0.43 |

Chlorides x Alcohol | -0.36 |

Quality x Density | -0.31 |

Quality x Chlorides | 0.21 |

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?

Surprisingly, in this dataset, alcohol is the best raw predictor and contributor of quality, with a corr. coefficient of 0.44.

Fixed acidity had a positive correlation coef. with the pH level at 0.43, this follows what we know about the pH level's purpose to measure the acidity level. The two other acidity based variables, citric and volatile acidity, also have a insignificantly negative relationship with pH level in much lesser degrees (-0.03 vs -0.16) .

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

There were obvious correlations between variables such as fixed acidity and ph levels and between density and sugar. But I hadn't expected sugar and density levels to be so positively correlated.

Since free sulfur dioxide is the free form of S02 (dissolved gas), it makes up a part of total sulfur dioxide, which would explain ther correlation coefficient.

What was the strongest relationship you found?

Residual Sugar and Density: .84 As noted in the definitions: 'the density of wine is close to that of water depending on the percent alcohol and sugar content'

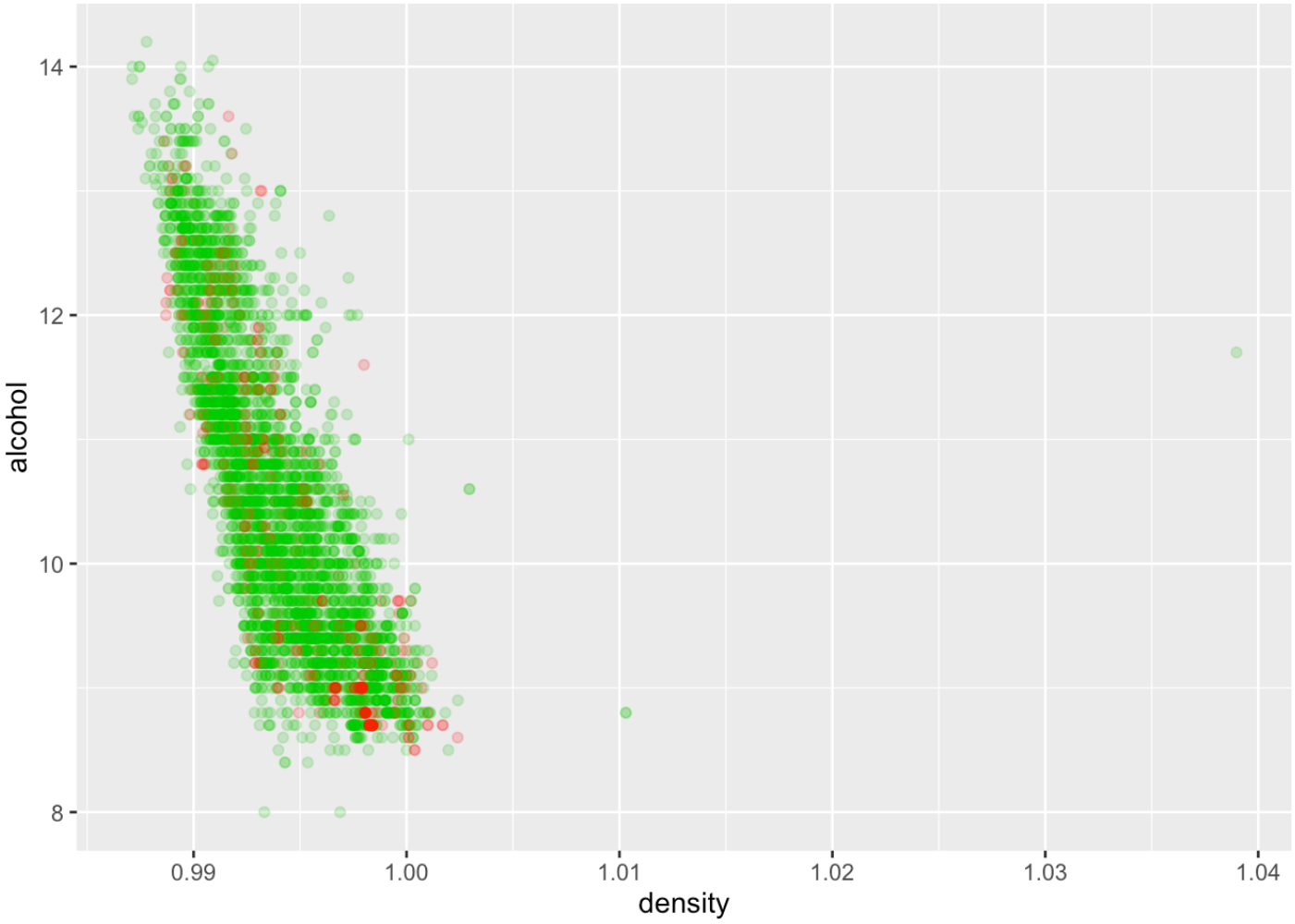
So we can expect density to be highly influenced by sugar content and alcohol levels which is displayed above.

Density has a negative relationship of -0.31 with Quality. After doing some research on red wines, we could assume that sugar in red wine should be much less in contrast with white wines where residual sugar may have a positive relationship with quality. In this particular dataset, quality and residual sugar is slightly negative -.10 but it's too weak for us to assume such a relationship exists.

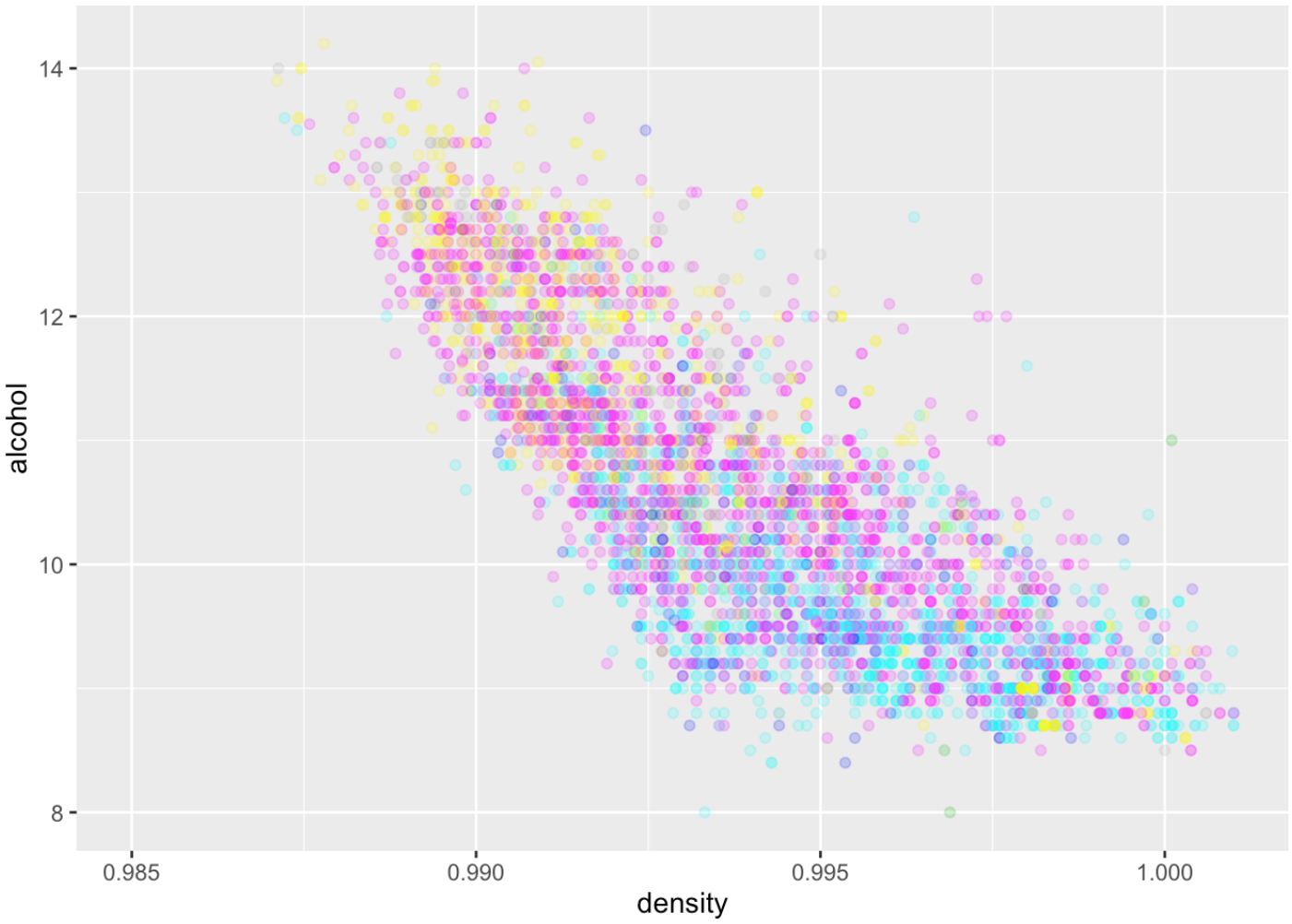
Residual.sugar has a negative correlation with alcohol. -0.45. Where as alcohol seems to have a slightly negative correlation with density.

Multivariate Plots Section

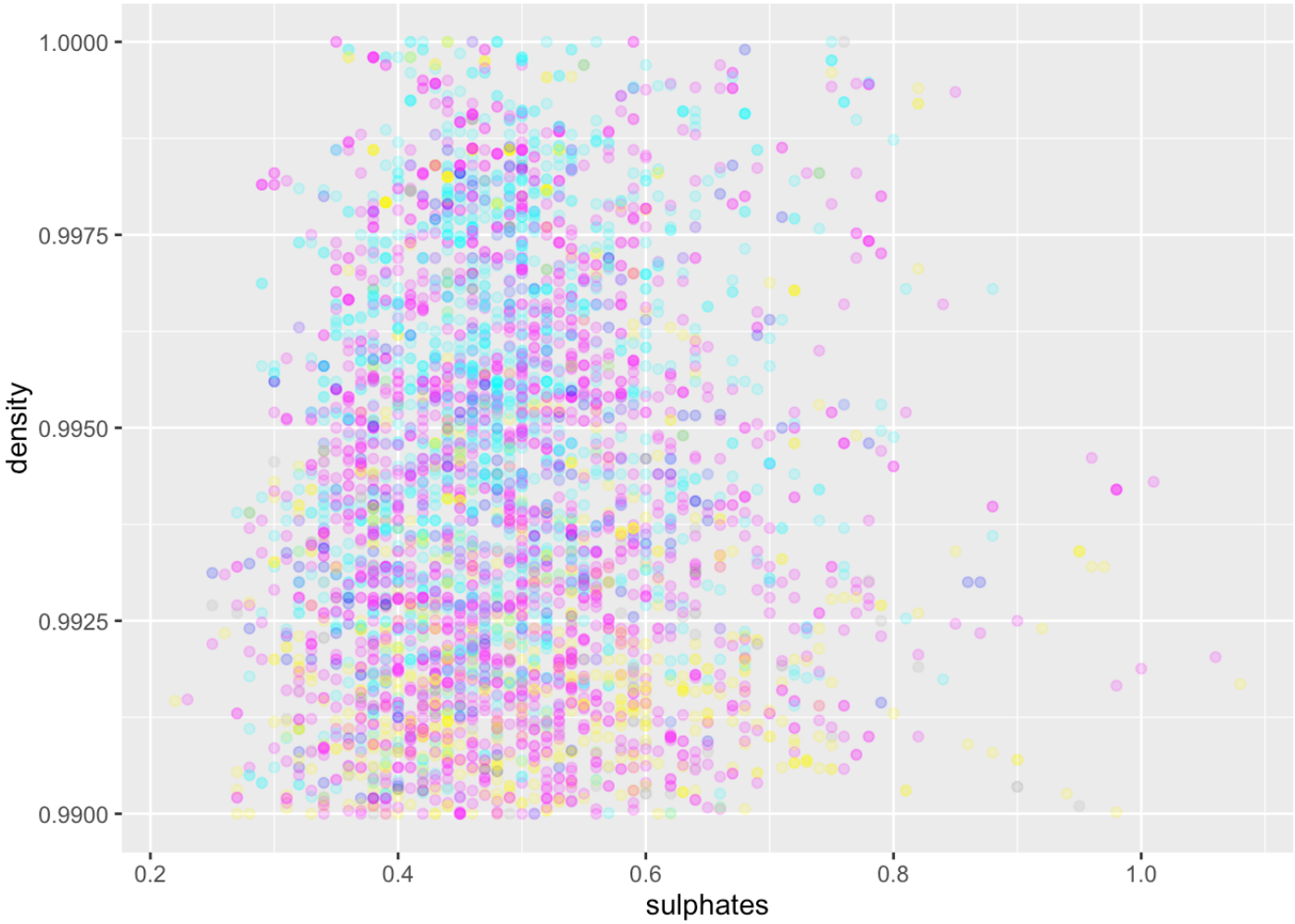
```
## [1] 6 5 7 8 4 3 9
```



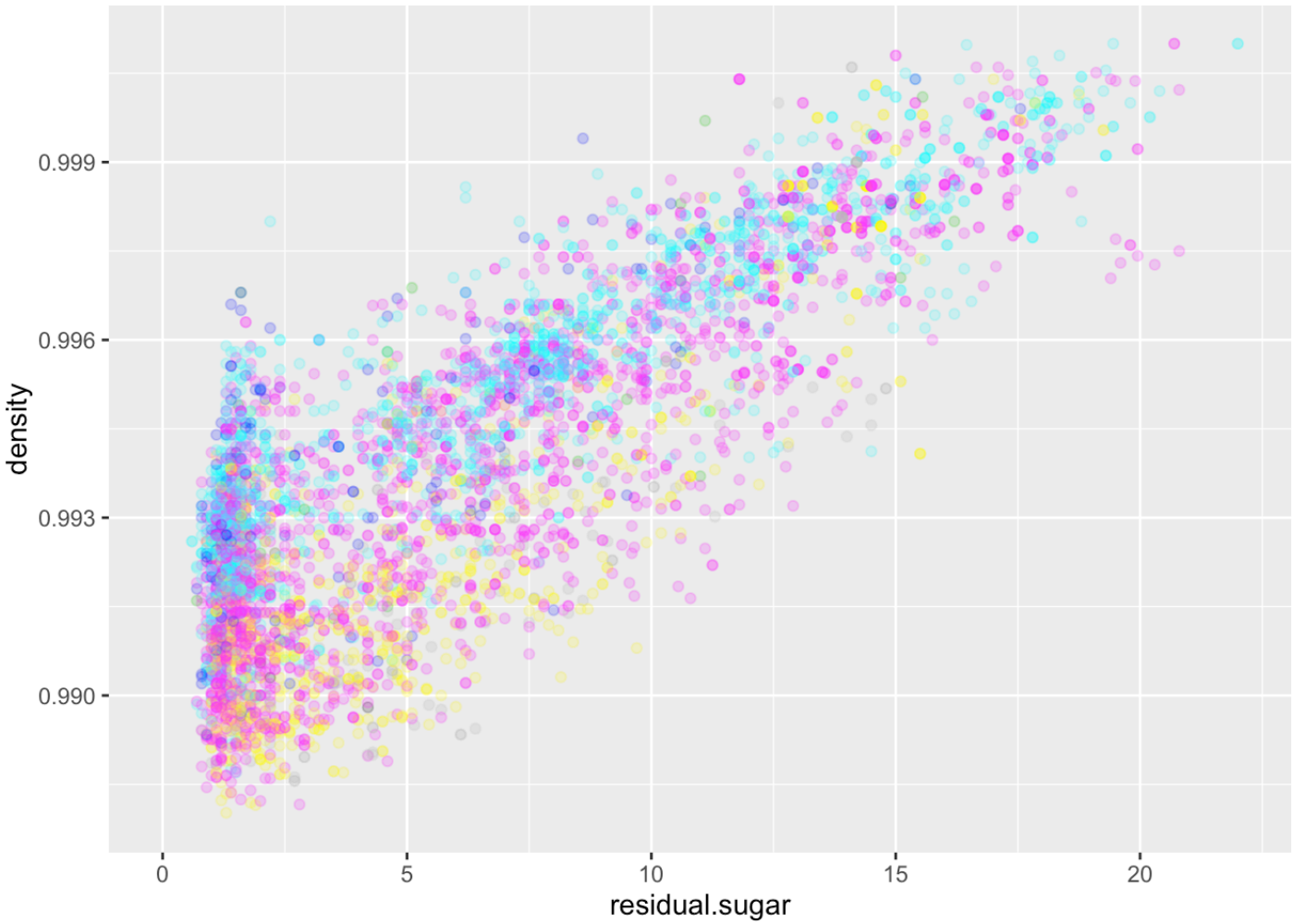

```
## Warning: Removed 15 rows containing missing values (geom_point).
```

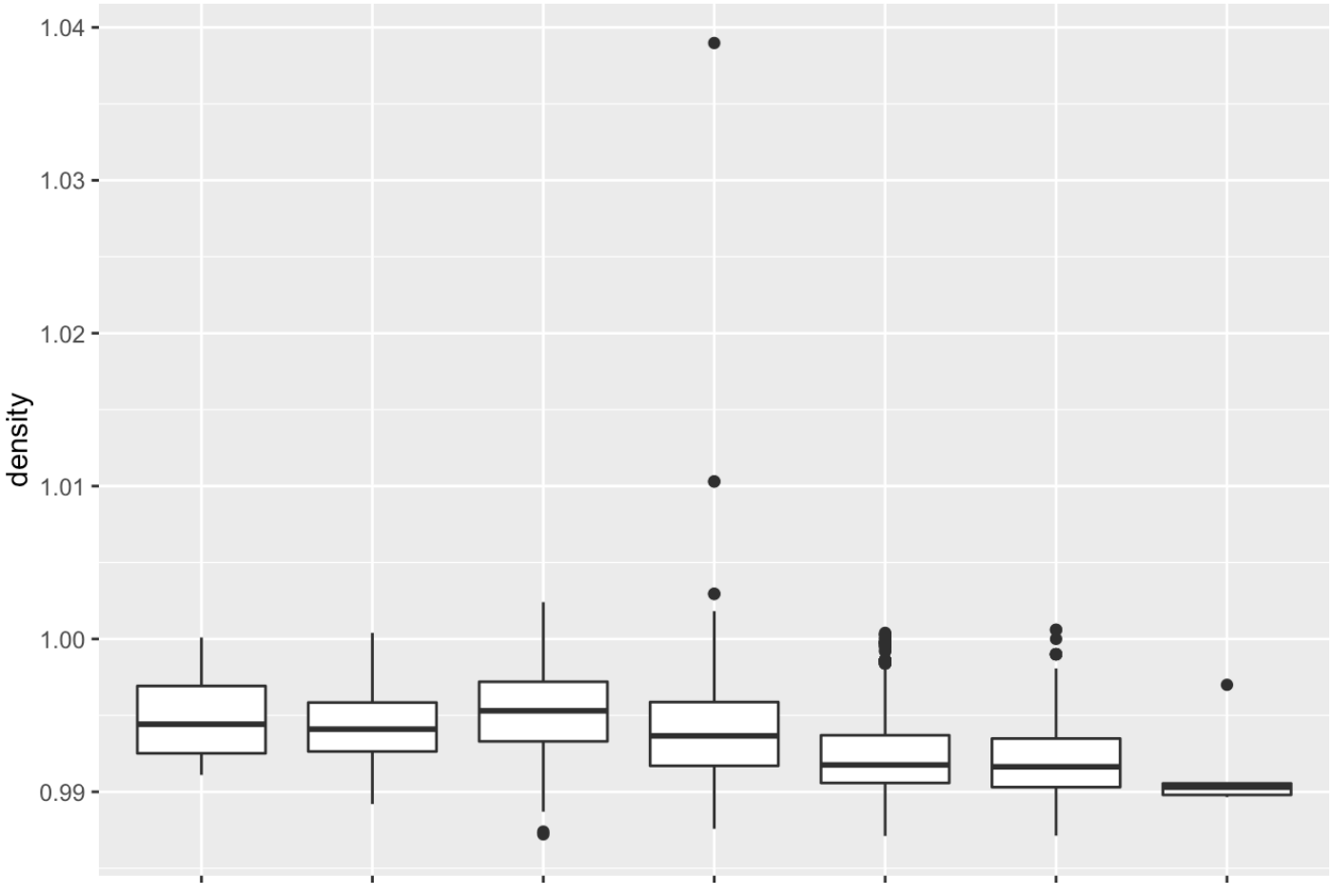


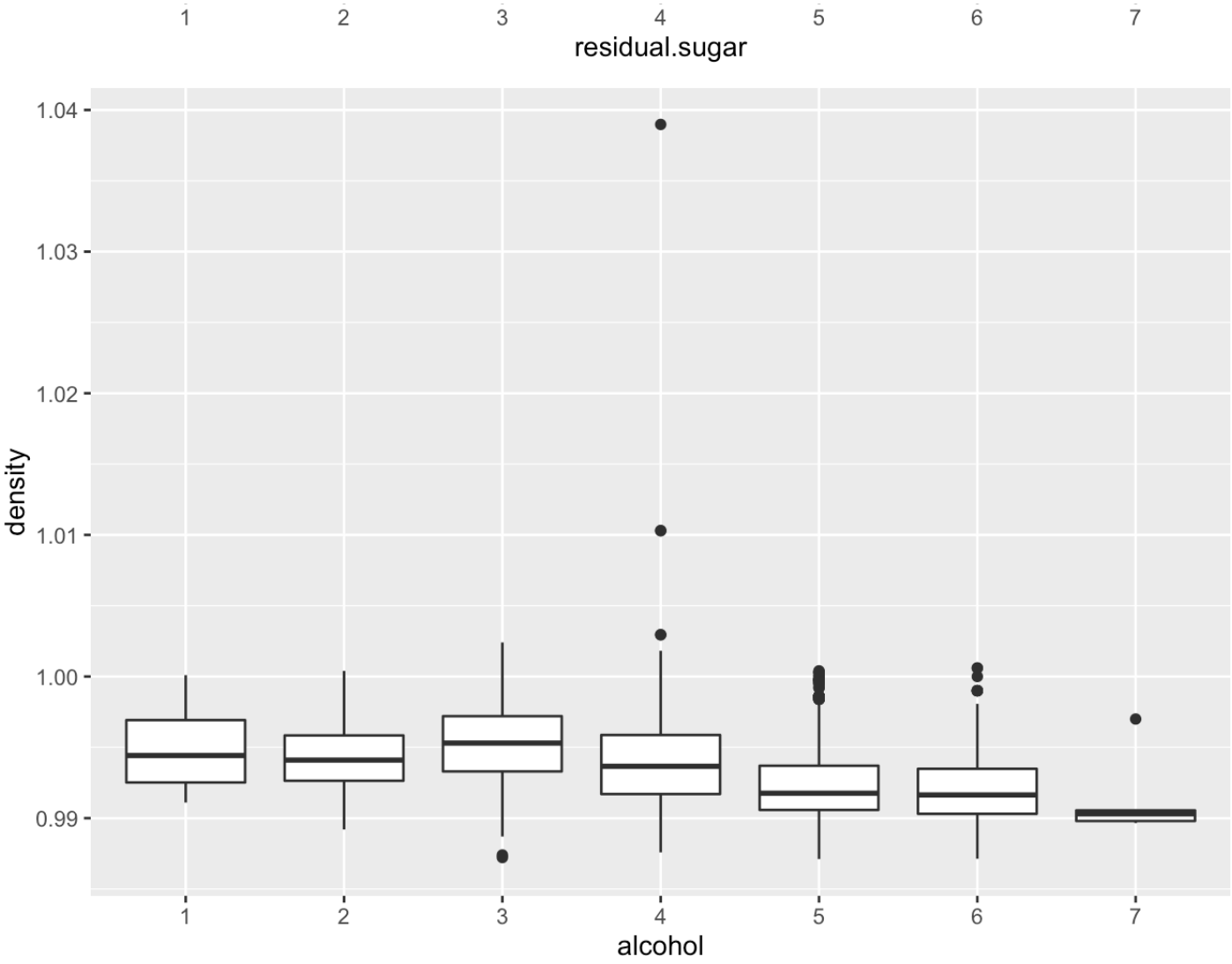

```
## Warning: Removed 421 rows containing missing values (geom_point).
```

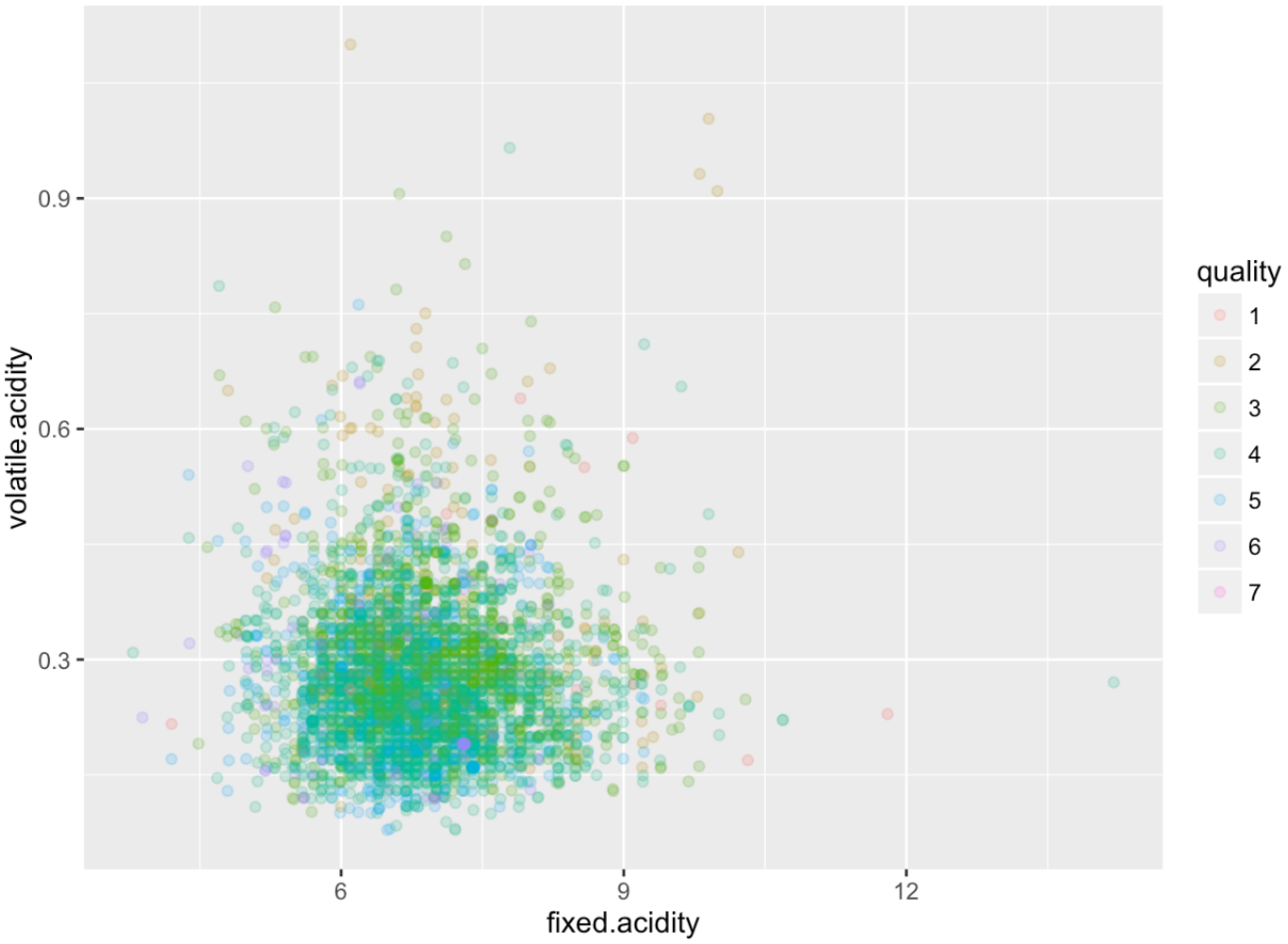


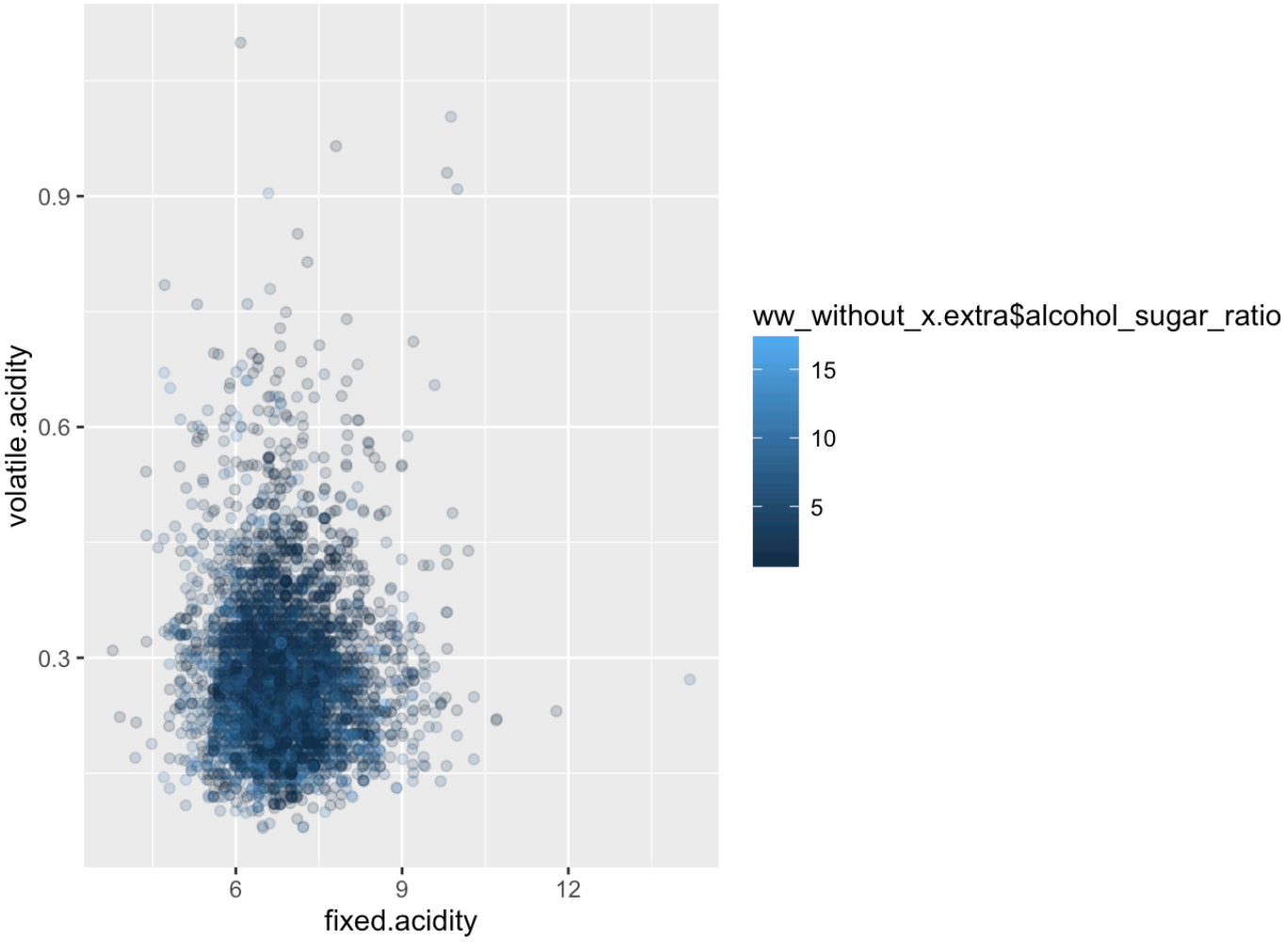

```
## Warning: Removed 29 rows containing missing values (geom_point).
```

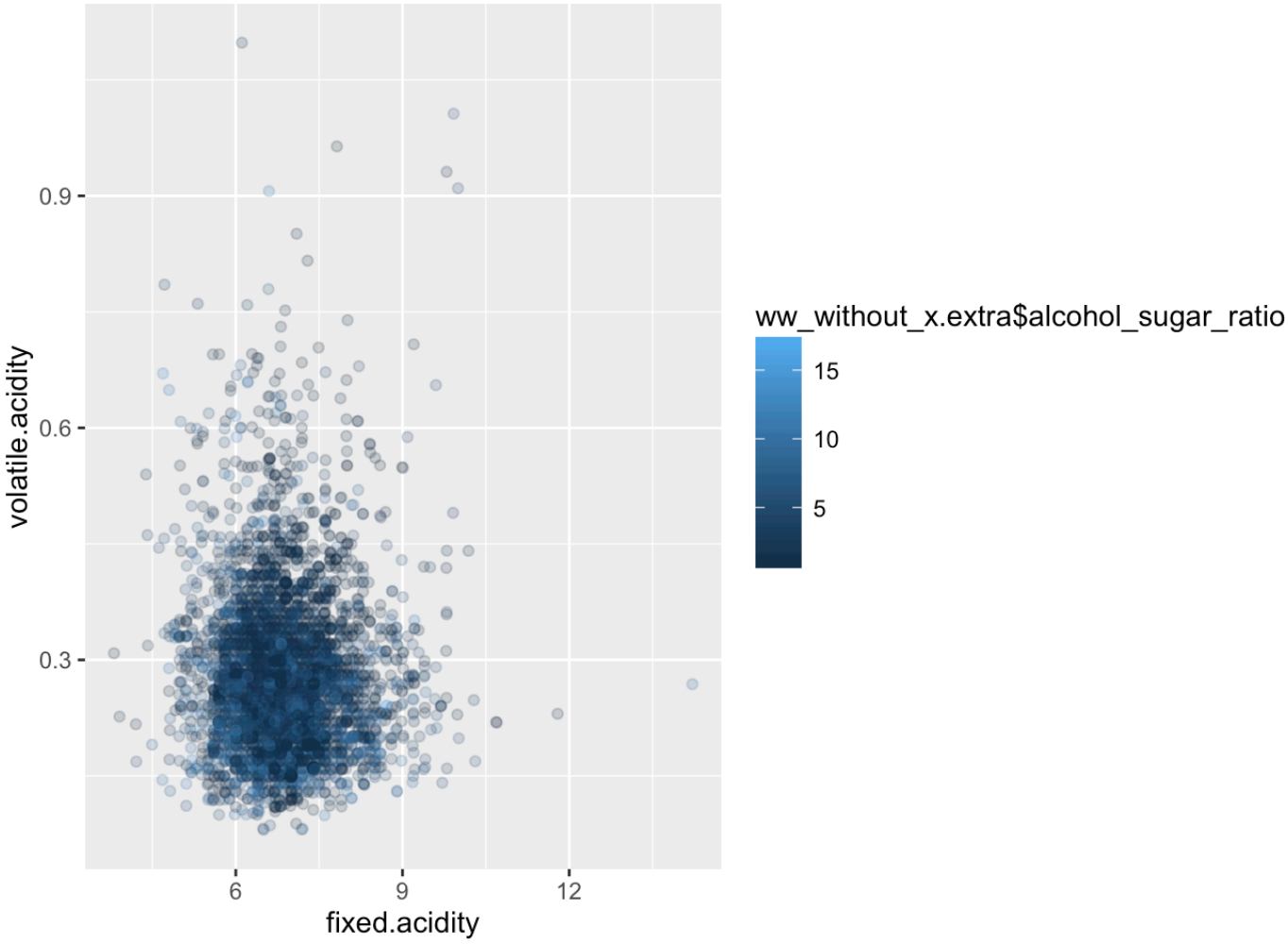

density
0.999
0.996
0.993
0.990
0
5
10
15
20
residual.sugar


density
1.04
1.03
1.02
1.01
1.00
0.99

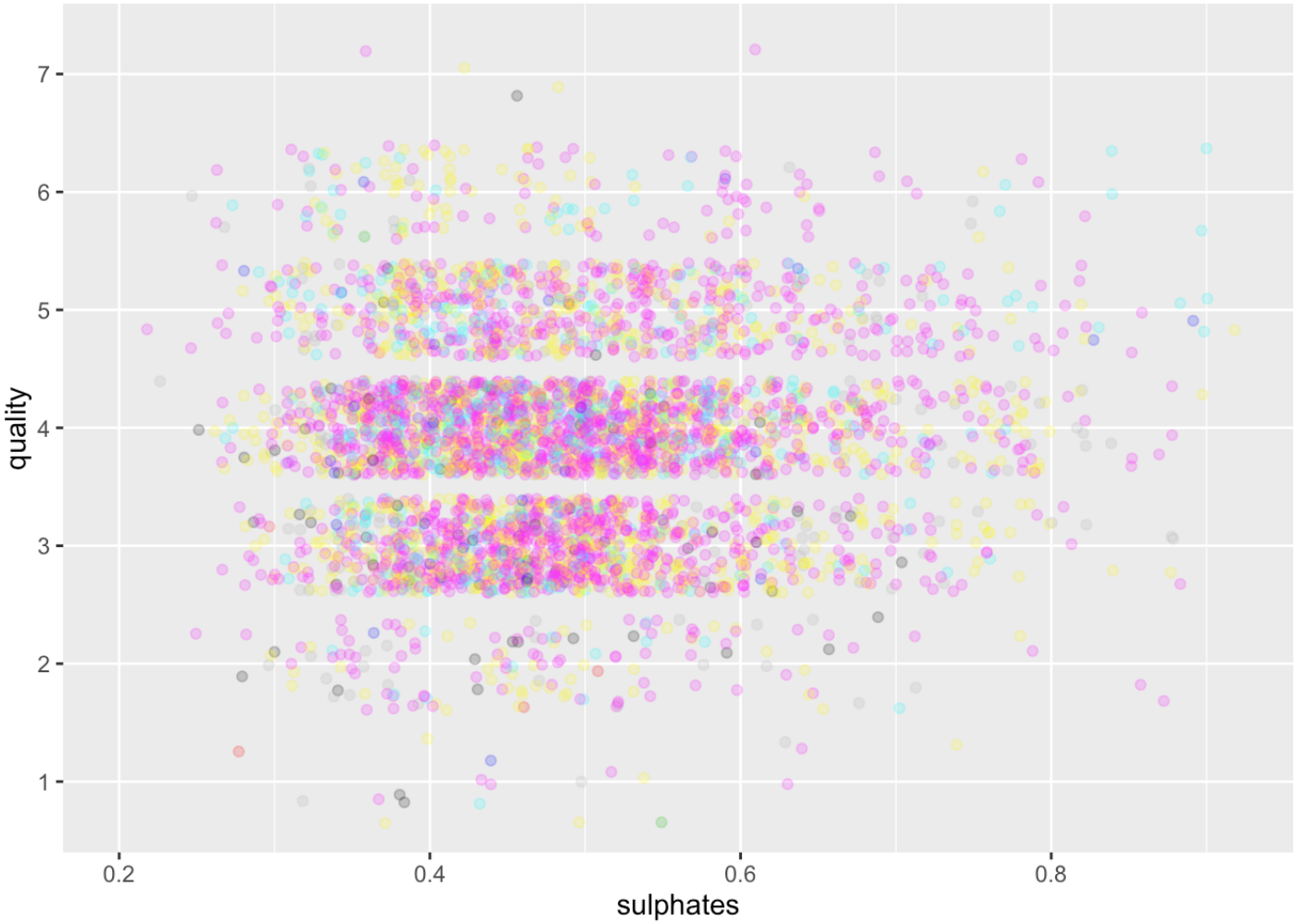

residual.sugar
1
2
3
4
5
6
7
density
1.04
1.03
1.02
1.01
1.00
0.99
alcohol


volatile.acidity
0.9
0.6
0.3
6
9
12
fixed.acidity
quality
1
2
3
4
5
6
7


volatile.acidity
0.9
0.6
0.3
6
9
12
fixed.acidity
ww_without_x.extra$alcohol_sugar_ratio
15
10
5

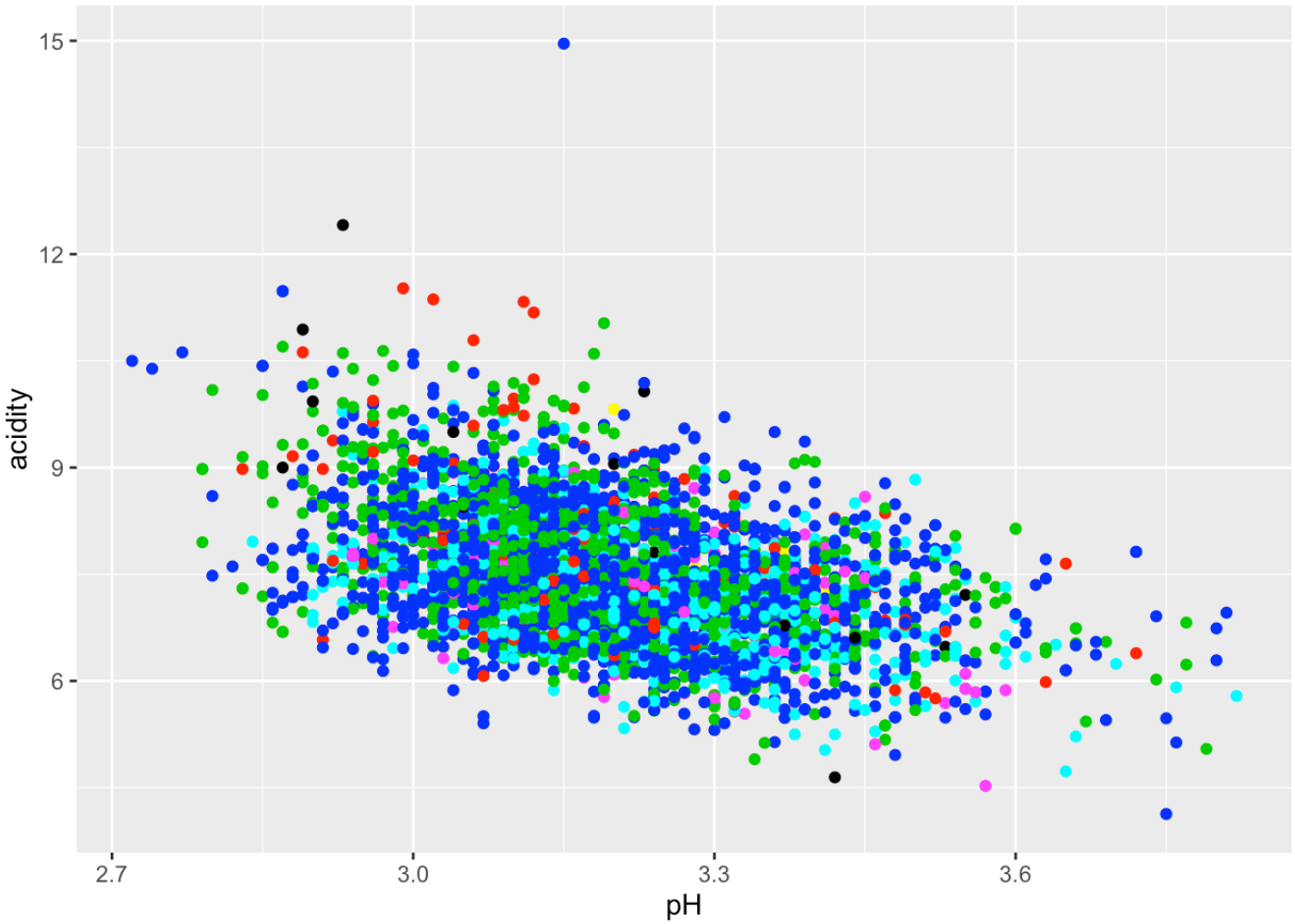


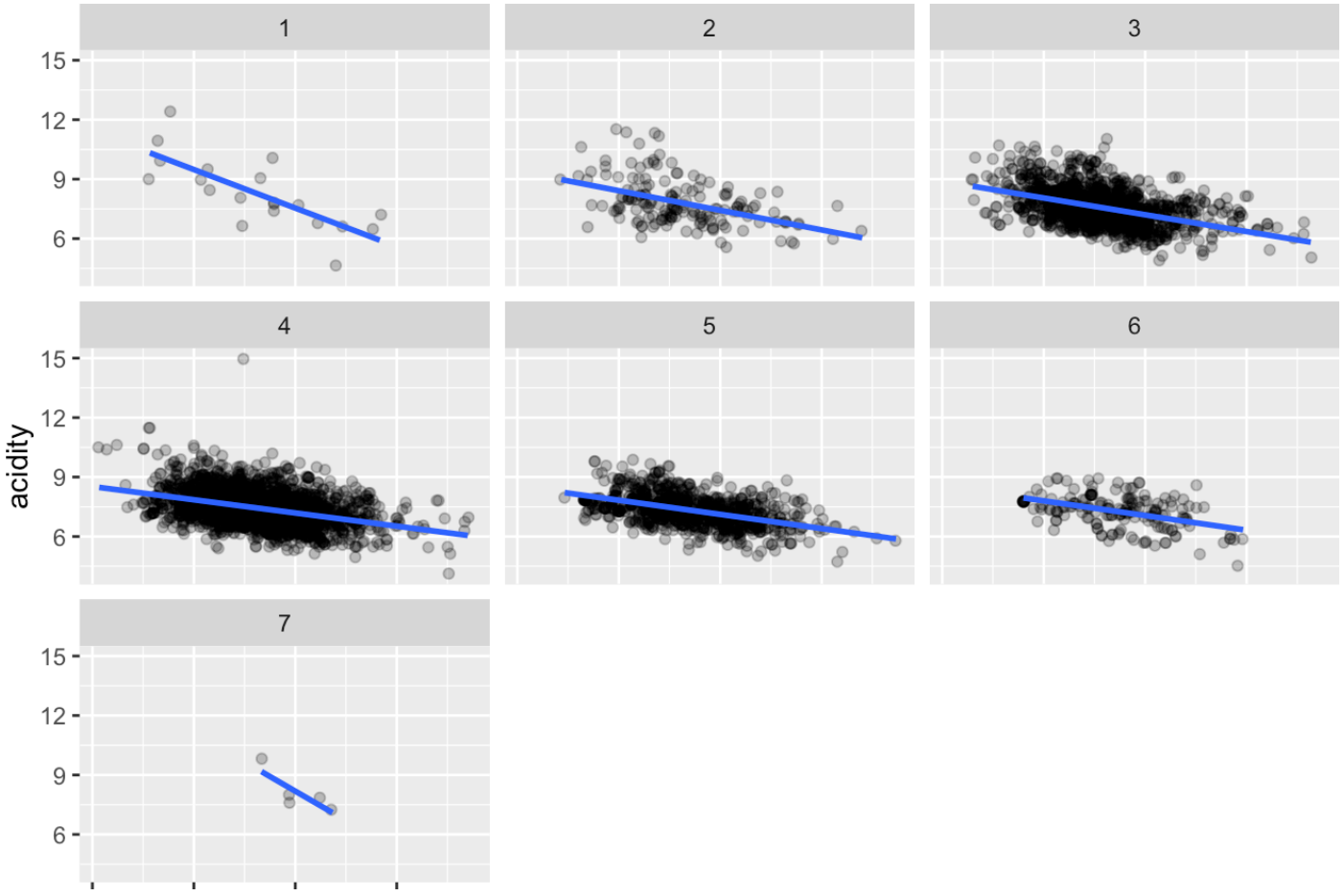

```
## Warning: Removed 23 rows containing missing values (geom_point).
```

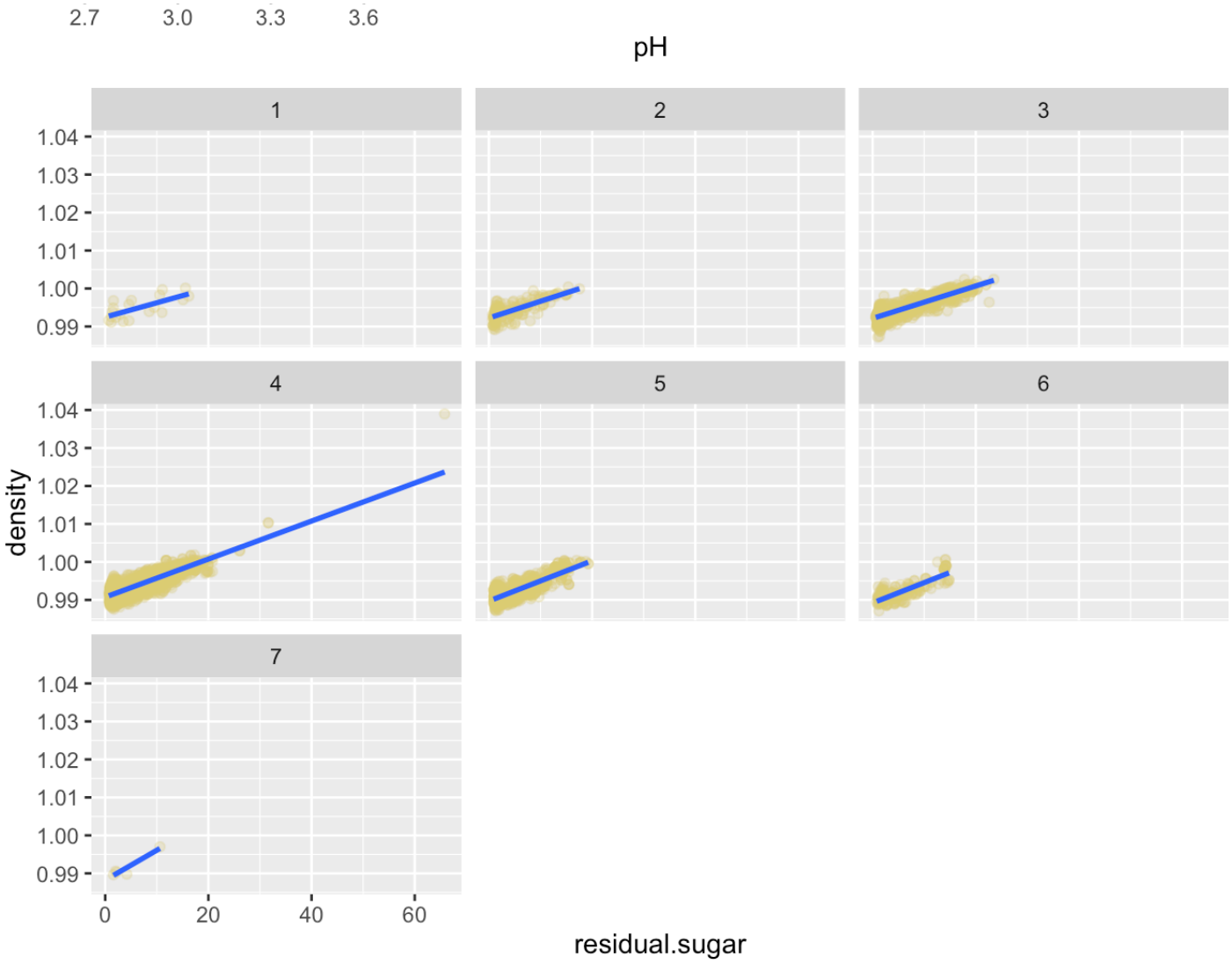


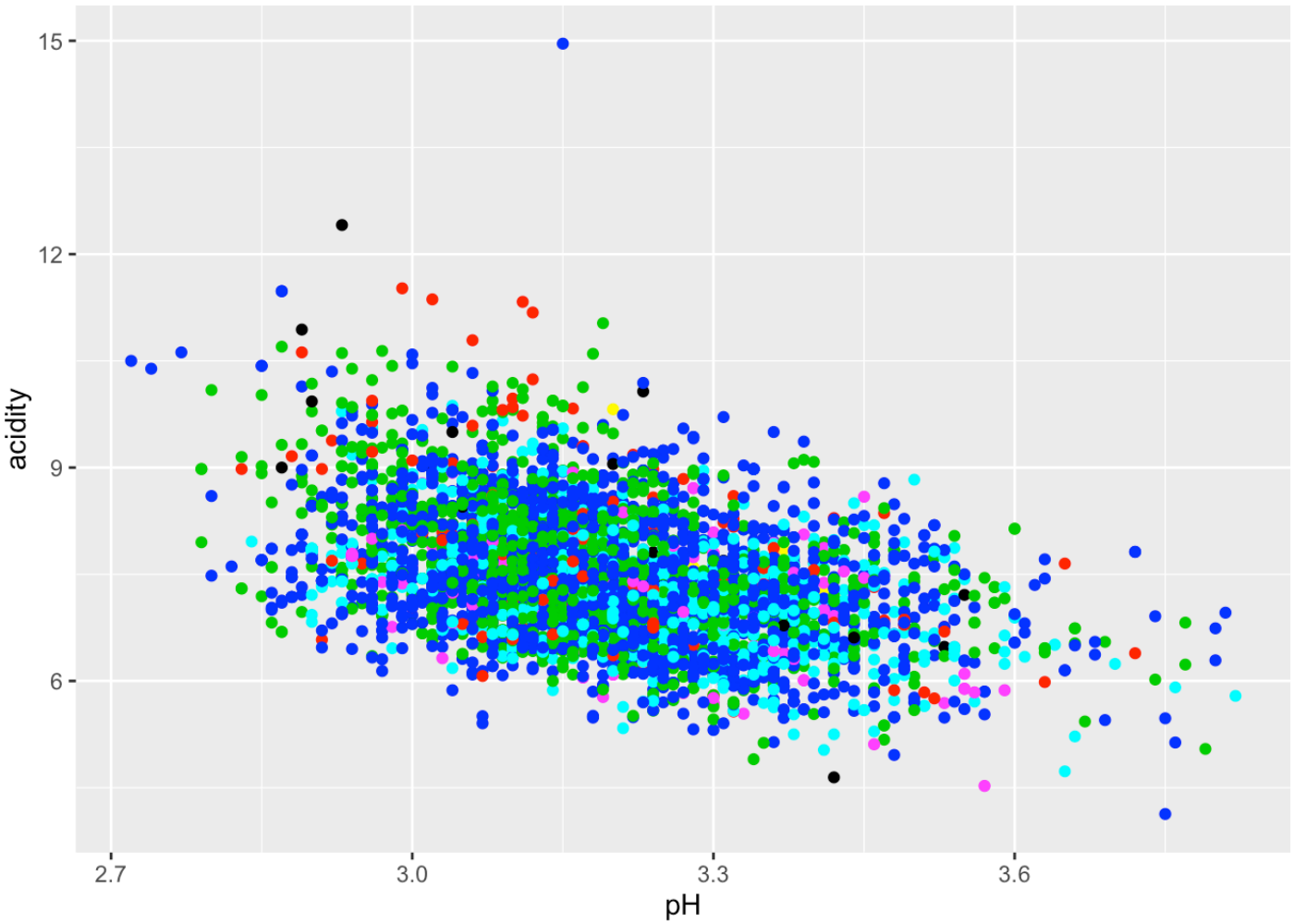

```
## [1] -0.4306513
```

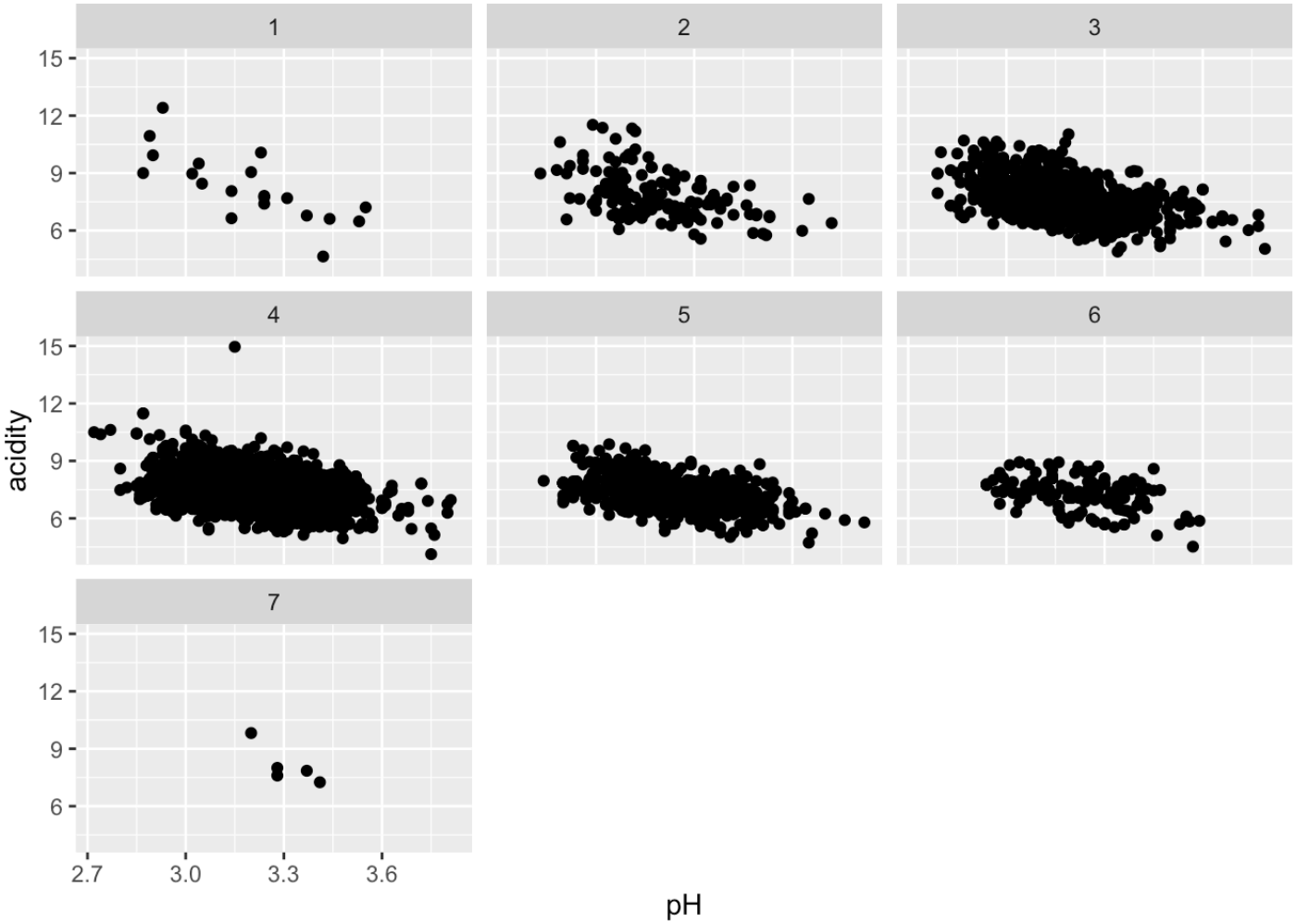
```
## [1] -0.1313772
```


15
12
9
6
acidity
2.7
3.0
3.3
3.6
pH


1
2
3
4
5
6
7
15
12
9
6
acidity


2.7
3.0
3.3
3.6
pH
1
2
3
4
5
6
7
density
1.04
1.03
1.02
1.01
1.00
0.99
0
20
40
60
residual.sugar


15
12
9
6
acidity
2.7
3.0
3.3
3.6
pH




```
##  [1] "fixed.acidity"        "volatile.acidity"     "citric.acid"
##  [4] "residual.sugar"       "chlorides"            "free.sulfur.dioxide"
##  [7] "total.sulfur.dioxide" "density"              "pH"
## [10] "sulphates"            "alcohol"              "quality"
## [13] "bound_s02"            "acidity"              "alcohol_sugar_ratio"
```

```
##
##  Pearson's product-moment correlation
##
## data:  alcohol_sugar_ratio and as.integer(quality)
## t = 4.4271, df = 4896, p-value = 9.757e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.03520027 0.09099002
## sample estimates:
##       cor
## 0.06314448
```

```
names(ww_without_x.extra)
```

```
##  [1] "fixed.acidity"        "volatile.acidity"     "citric.acid"
##  [4] "residual.sugar"       "chlorides"            "free.sulfur.dioxide"
##  [7] "total.sulfur.dioxide" "density"              "pH"
## [10] "sulphates"            "alcohol"              "quality"
## [13] "bound_s02"            "acidity"              "alcohol_sugar_ratio"
```

```
is.factor(ww_without_x.extra$quality)
```

```
## [1] FALSE
```

```
ww_without_x.extra$quality <- as.integer(ww_without_x.extra$quality)
is.factor(ww_without_x.extra$quality)
```

```
## [1] FALSE
```

```
m1 <- lm(quality ~ alcohol, data = ww_without_x.extra)
m2 <- update(m1, ~ .   * acidity)
m3 <- update(m2, ~ .   * sulphates)
m4 <- update(m3, ~ .   * density)
m6 <- update(m4, ~ .   * chlorides)
m7 <- update(m6, ~ .   * pH)
m8 <- update(m7, ~ .   * total.sulfur.dioxide)
m9 <- update(m8, ~ .   * free.sulfur.dioxide)
m10 <- update(m9, ~ .  * residual.sugar)
m11 <- update(m10, ~ . * citric.acid)
m12 <- update(m11, ~ . * acidity)
m13 <- update(m12, ~ . * bound_s02)
m14 <- update(m13, ~ . * alcohol_sugar_ratio)
# mtable(m1, m2, m3, m4, m6, m7, m8, m9, m10, m11)

print("R - Squared: ")
```

```
## [1] "R - Squared: "
```

```
summary(m12)$r.squared
```

```
## [1] 0.5460446
```

```
print("Residual Standard Error (sigma): ")
```

```
## [1] "Residual Standard Error (sigma): "
```

```
summary(m12)$sigma
```

```
## [1] 0.6607302
```

```
print("R - Squared: ")
```

```
## [1] "R - Squared: "
```

```
summary(m14)$r.squared
```

```
## [1] 0.7824696
```

```
print("Residual Standard Error (sigma): ")
```

```
## [1] "Residual Standard Error (sigma): "
```

```
summary(m14)$sigma
```

```
## [1] 0.5628939
```

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?

Looking at the relationship of density and residual sugar across each rating remained consistent and undeterred by different quality ratings.

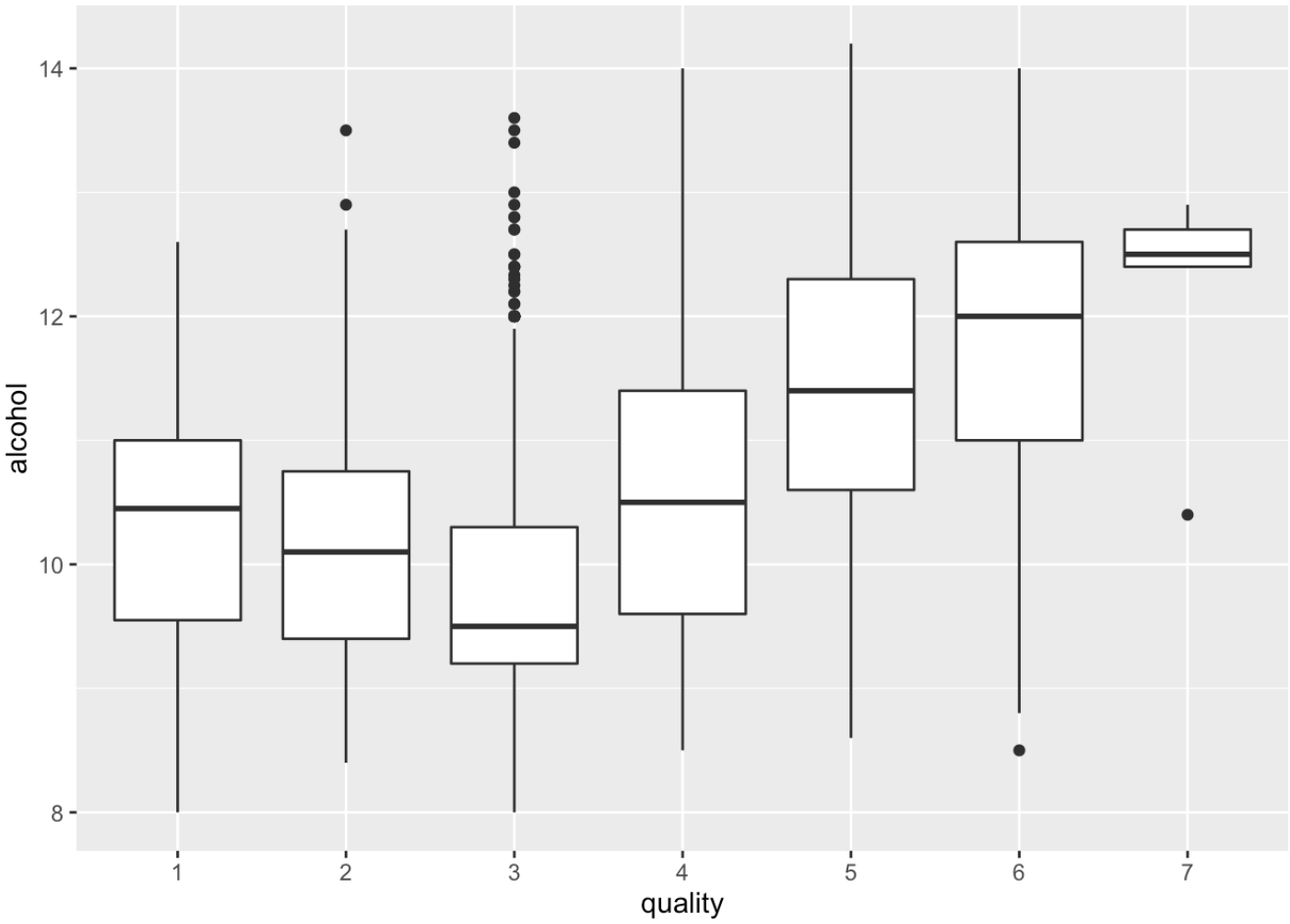
Though, insignificant, there was some quality based patterning when looking at residual sugar and density when coloured by quality. There seemed to be some stratification moving up from yellow to blue.

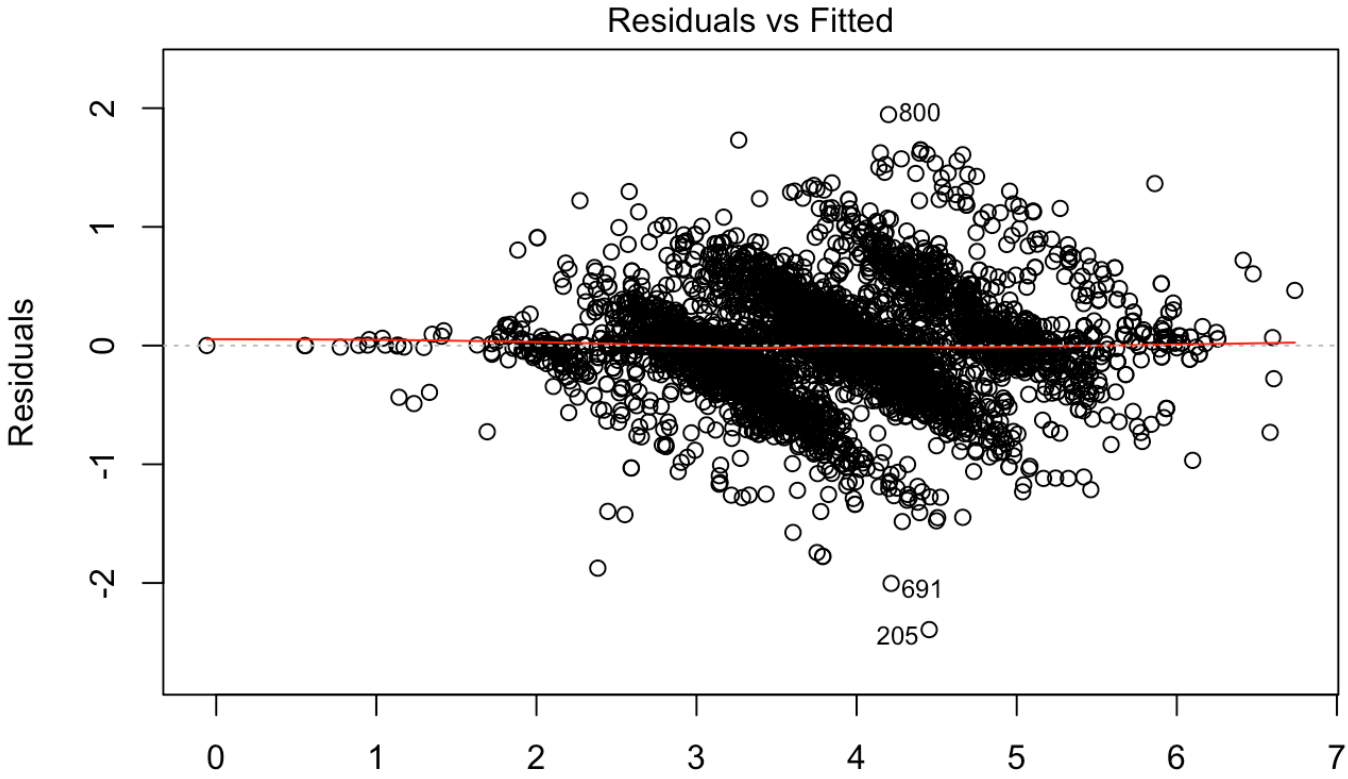
Prediction

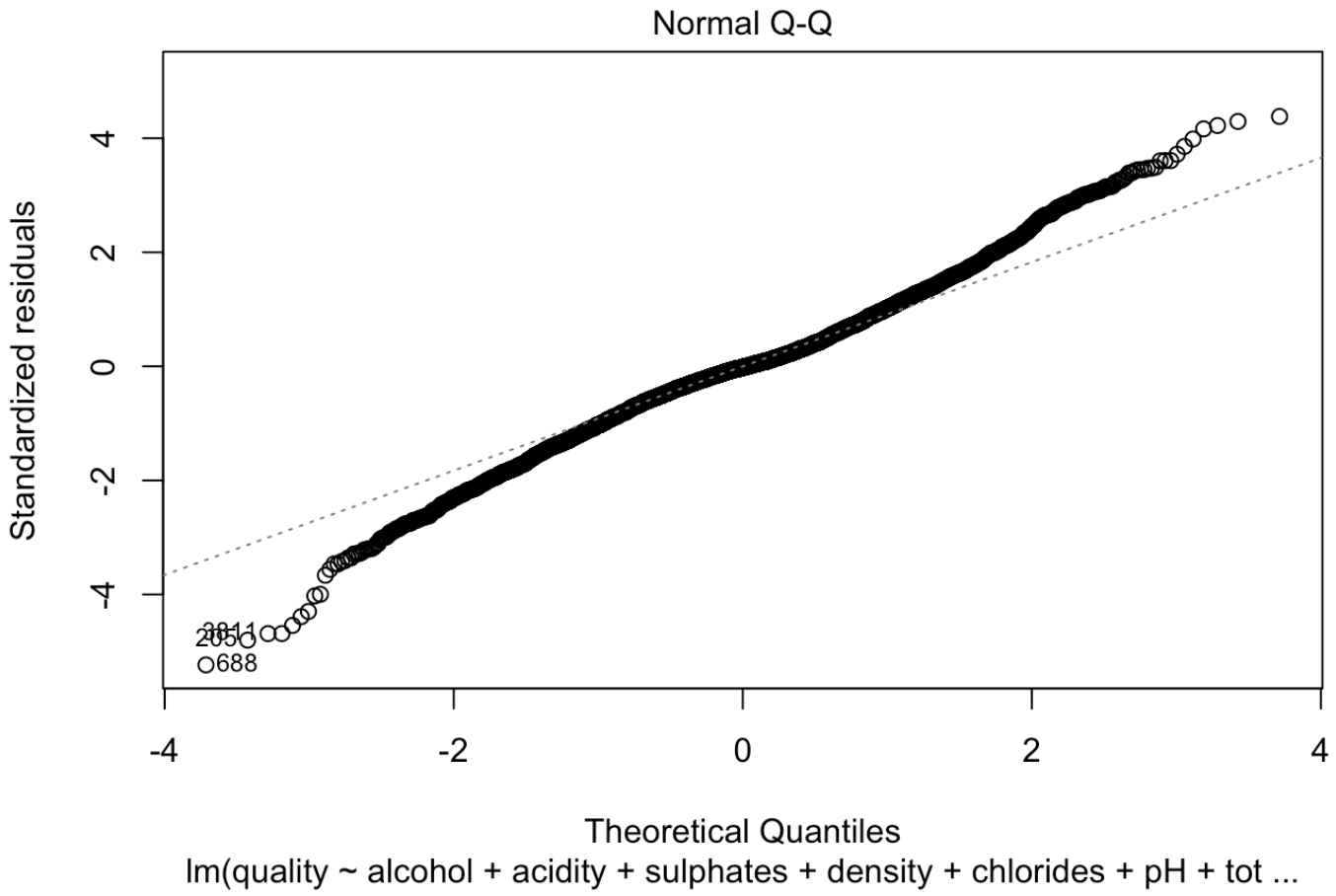
In my prediction moels

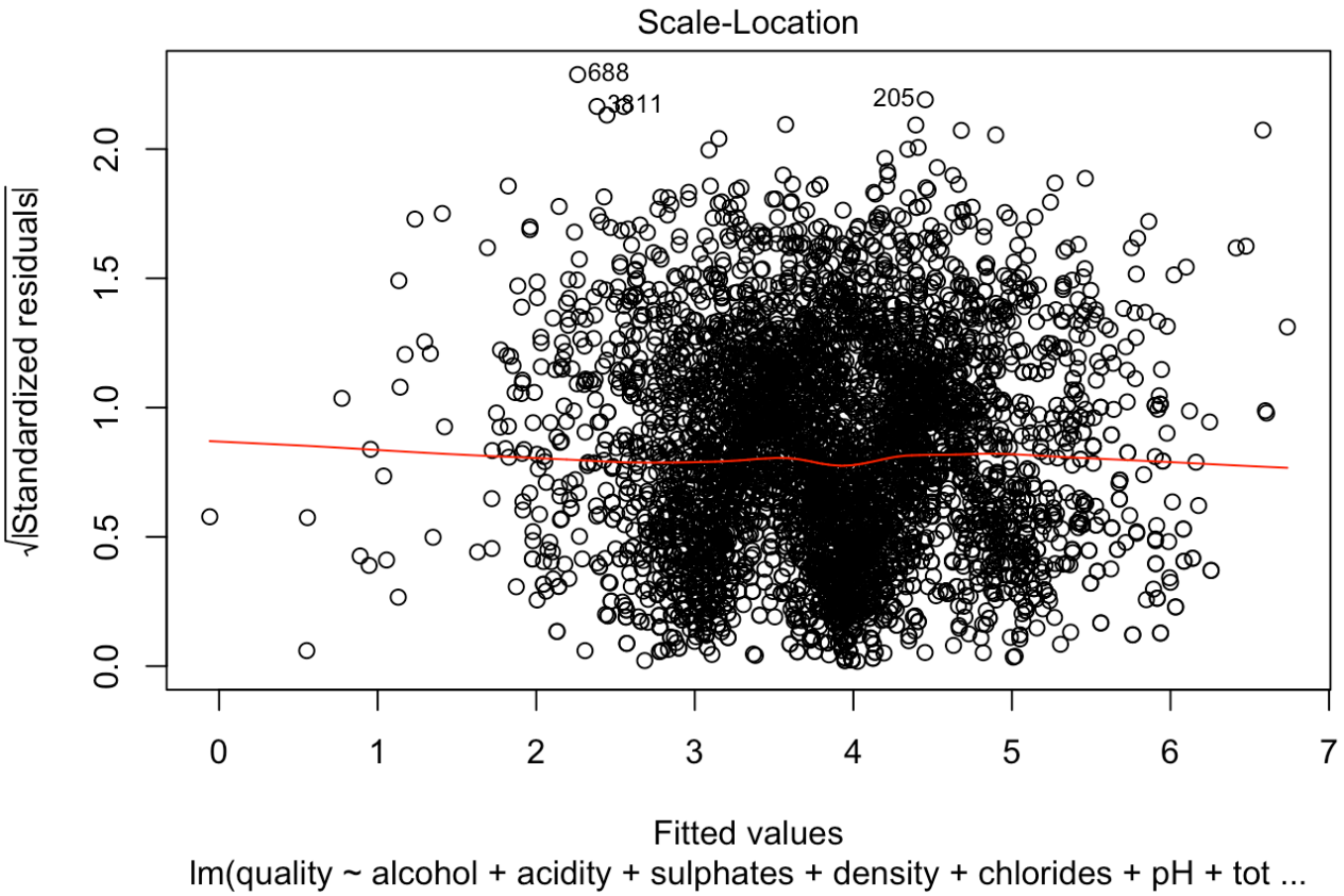
Final Plots and Summary

Plot One


alcohol
quality
14
12
10
8
1
2
3
4
5
6
7

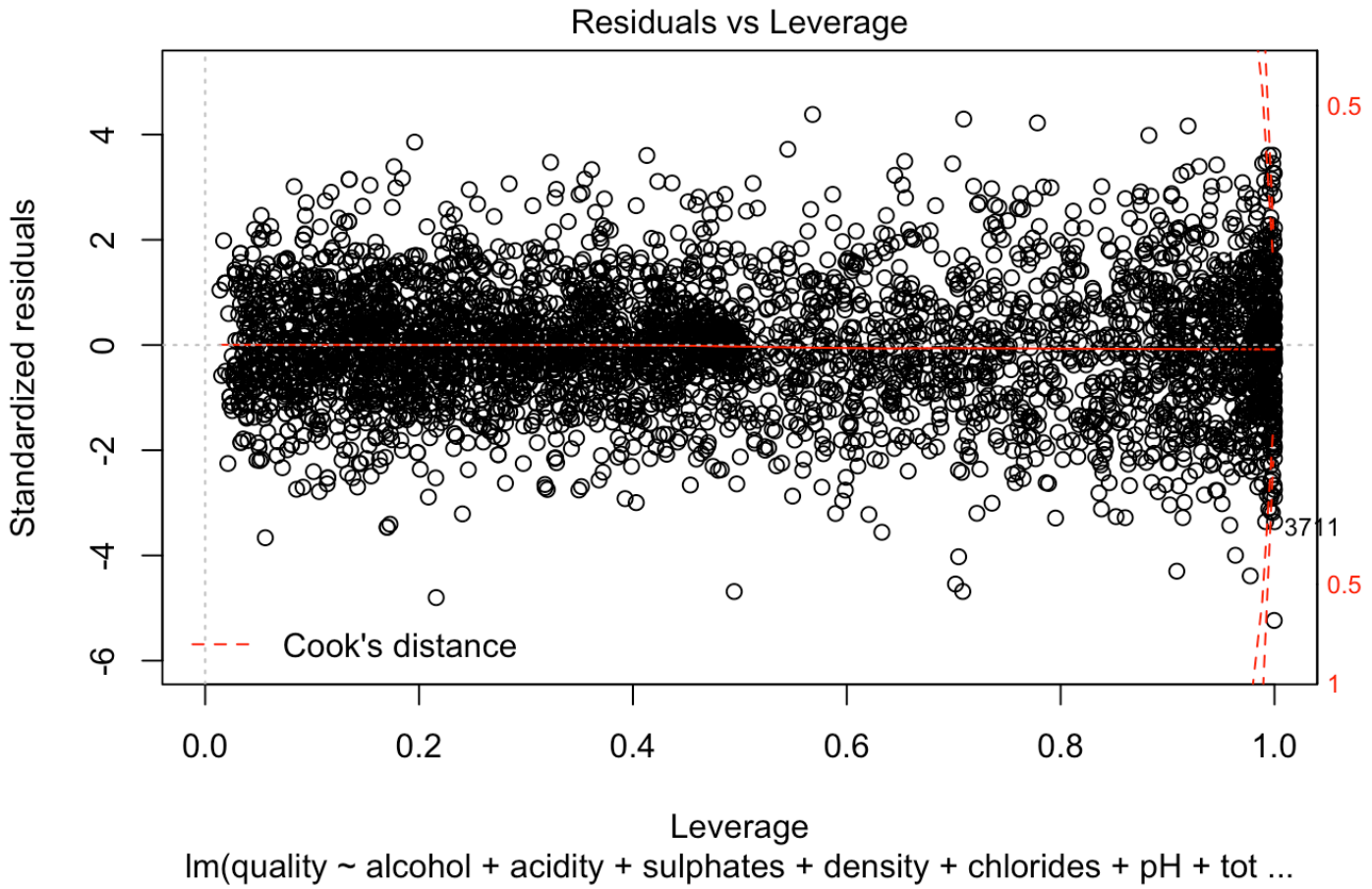

Residuals vs Fitted
Residuals
800
691
205
2
1
0
-1
-2
0
1
2
3
4
5
6
7


Fitted values
lm(quality ~ alcohol + acidity + sulphates + density + chlorides + pH + tot ...
Normal Q-Q
Standardized residuals
4
2
0
-2
-4
2811
205
688
-4
-2
0
2
4
Theoretical Quantiles
lm(quality ~ alcohol + acidity + sulphates + density + chlorides + pH + tot ...




```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



Description One

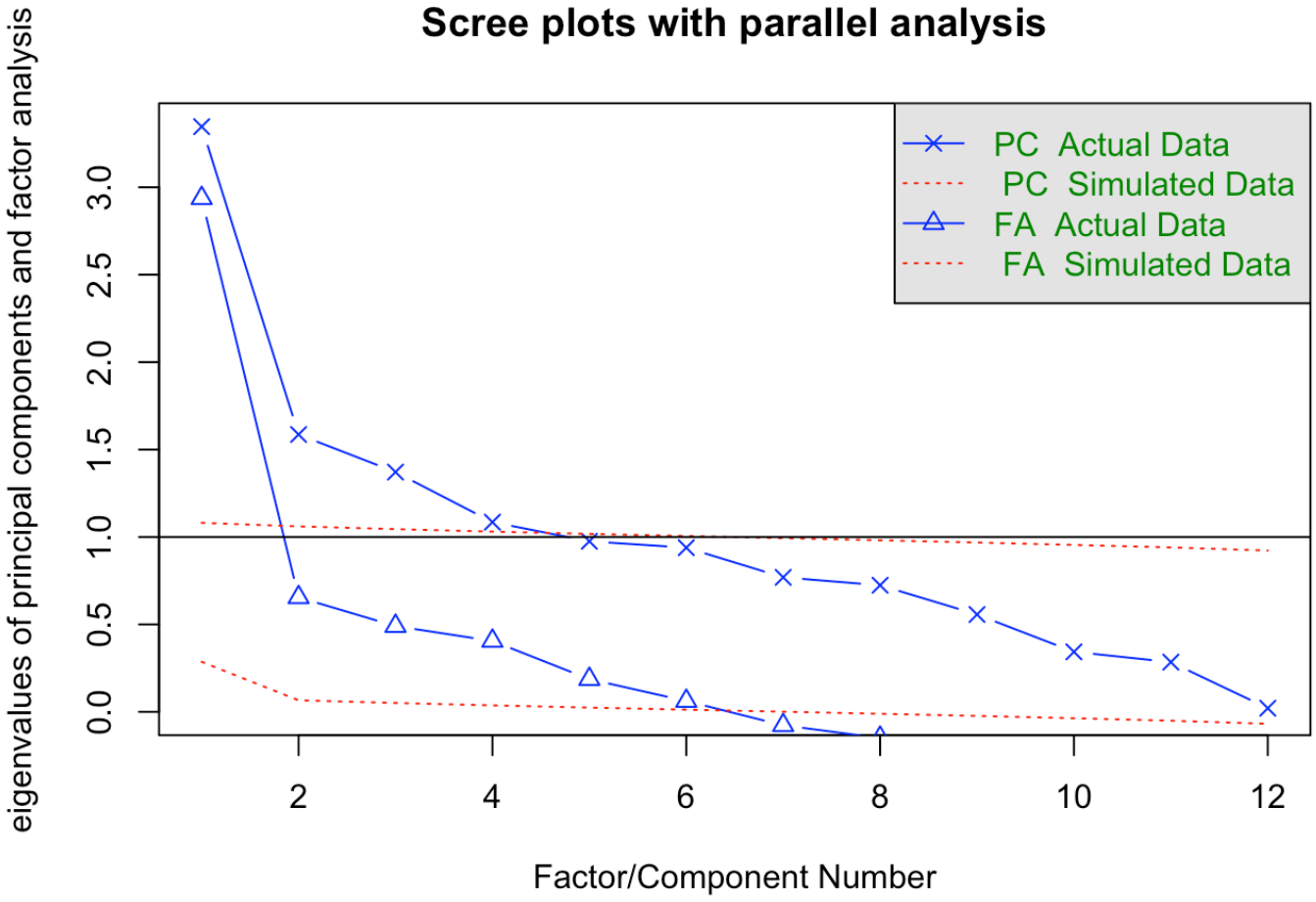
Quality wasn't easily predicted by any one variable, but surprisingly the amount of alcohol could be said to be the most likely single predictor of quality, with it's 44% correlation. Even various pairings with quality resisted any increase in correlation or variance predictability.

In my prediction, using the original variables I was able to achieve (possibly overfitted) a Multiple R-Squared of 0.546, and a residual standard error of 0.6607 consider all 11 physiochemical variables.

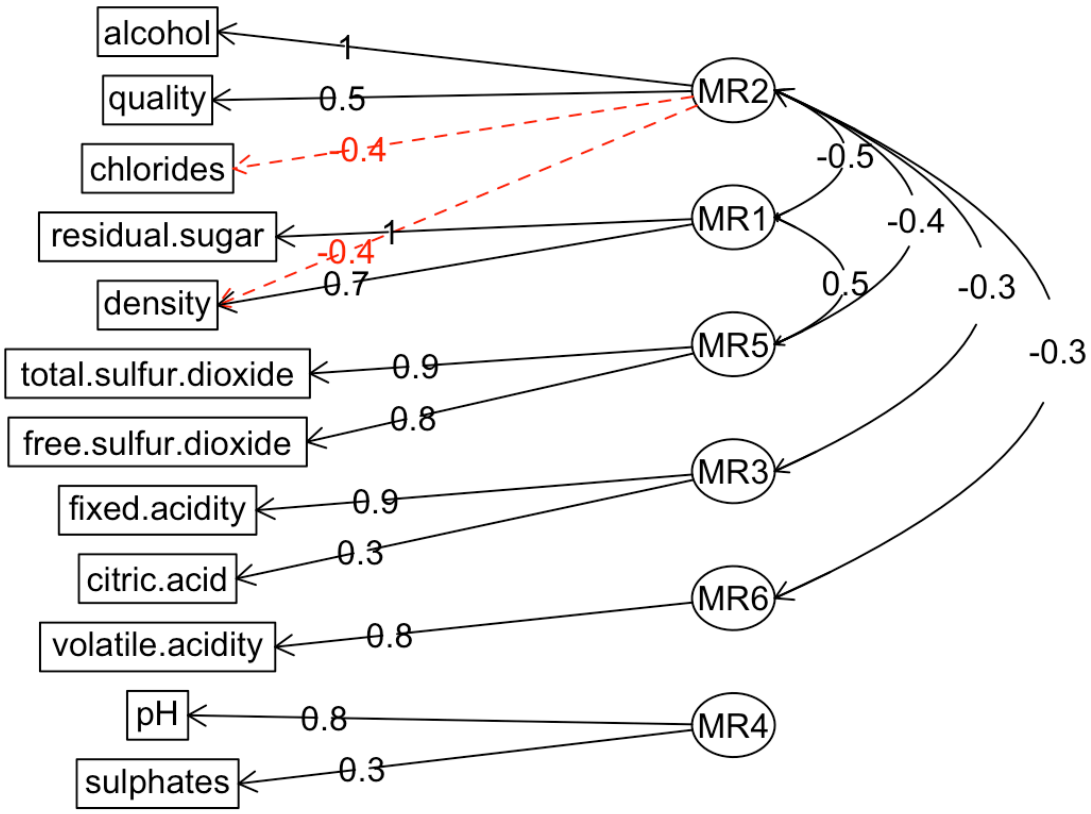
After adjusting the model by adding in 3 variables I created which include acidity (volatile.acidity + fixed.acidity + citric.acid), bound s02 (total.sulfur.dioxide - free.sulfur.dioxide), and alcohol to sugar ratio (alcohol/residual.sugar), I was able to increase the Multiple R-Squared to .7825 and lower the residual standard error to 0.5629.

This analysis surely suffers from overfitting and overlapping variables where they are either redundant or too correlated with each other.

Plot Two




```
## Parallel analysis suggests that the number of factors =  6  and the number of co
mponents =  4
```



I used Exploratory Factor Analysis is usually to explore what latent factors may be visible, I decided to do this at the onset and discovered some pairs that I would look at further later. The above scree plots recommend the optimal number of factors to be 6 and components to be 4. Since I am concerned with what factors are worth investigating, I leave out a thorough PCA analysis. The final result of EFA included pairs like alcohol and quality, which we know to be highly correlated, residual sugar and density, total sulfur dioxide and free sulfur dioxide. There were also some surprises like fixed acidity and citric acid which I combined later in my study. As well as pH level and sulphates. Using this type of analysis I was able to bring new questions and arrangements of data to investigate later.

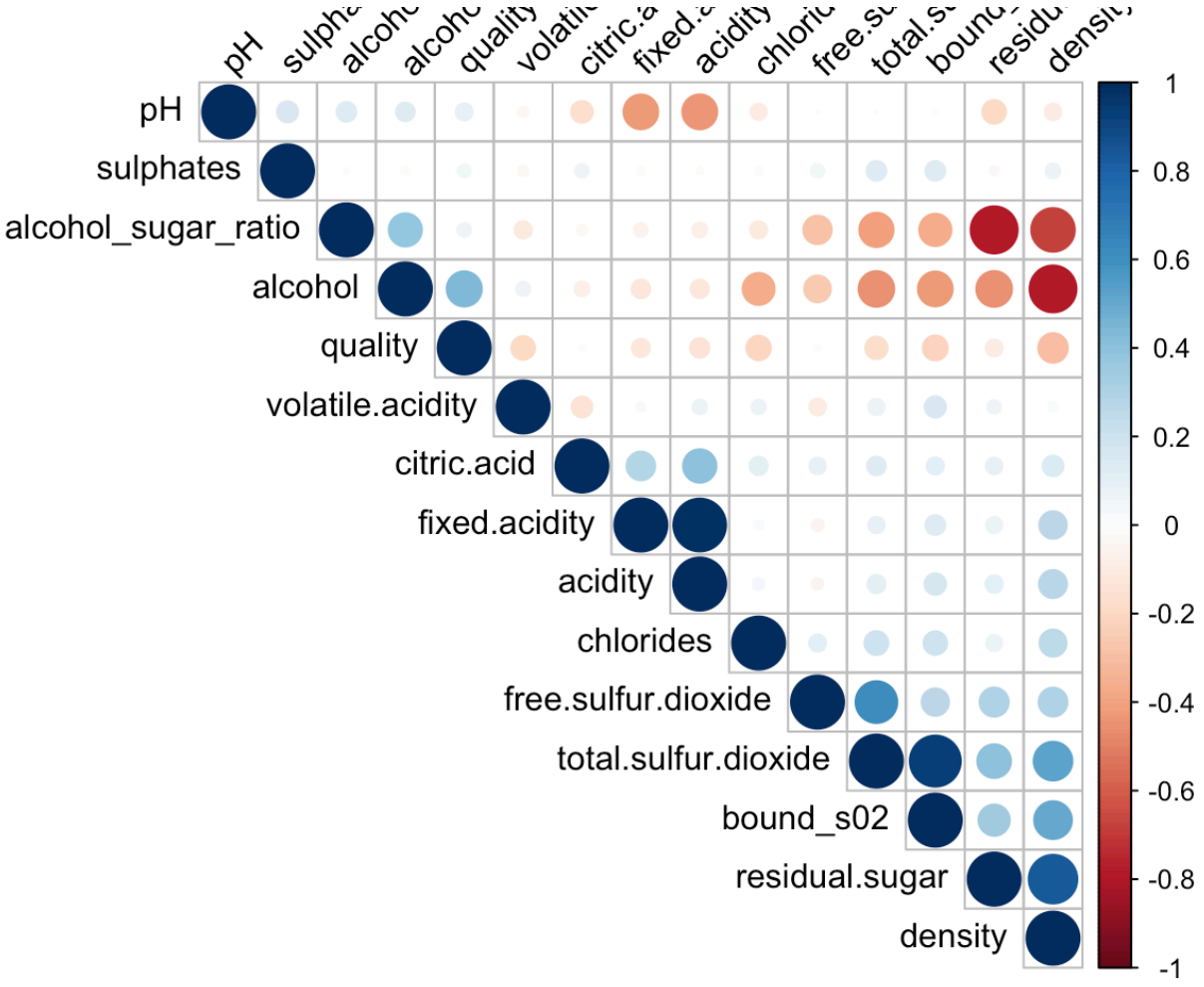
Description Two

Plot Three

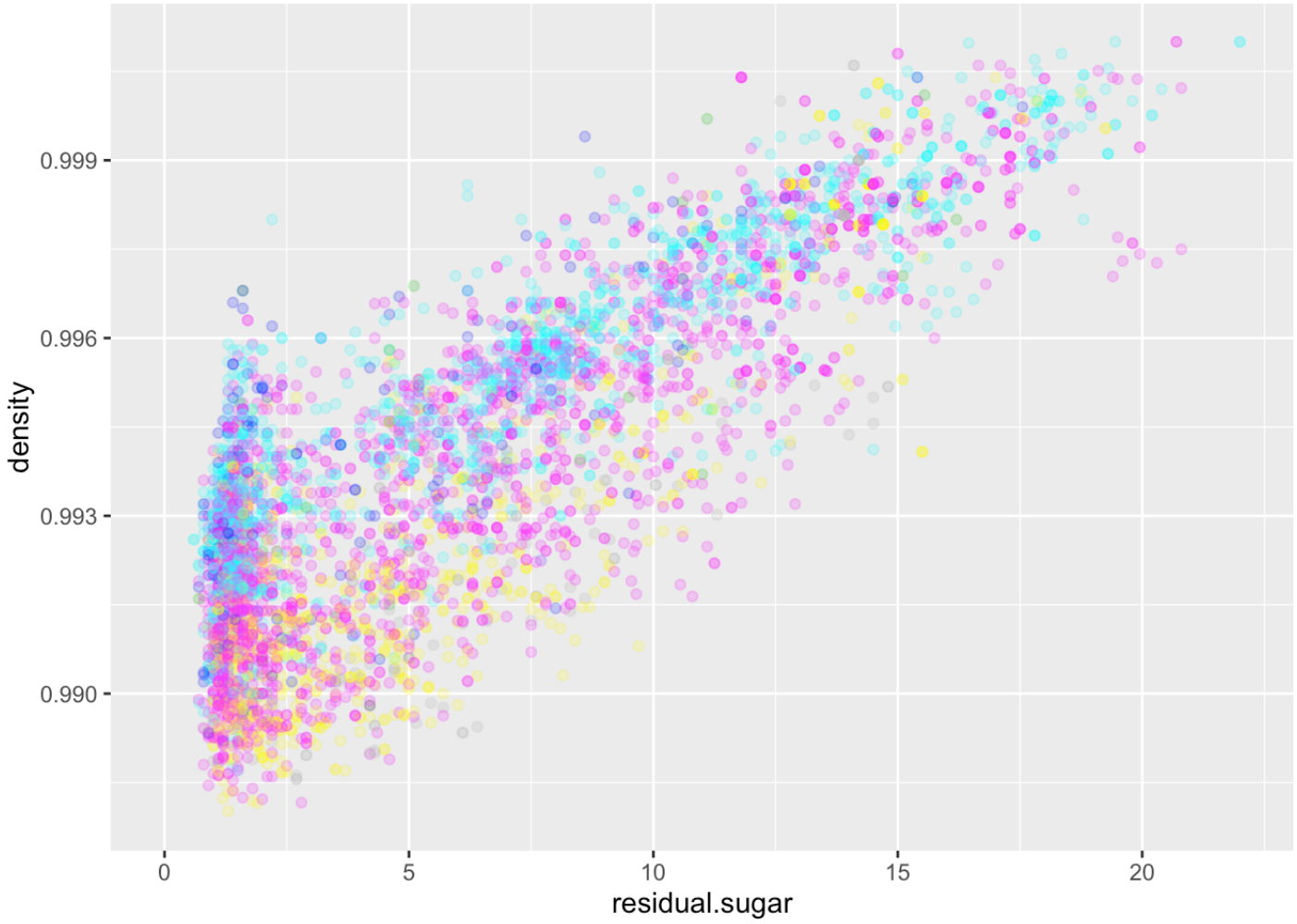
```
##                       fixed.acidity volatile.acidity  citric.acid
## fixed.acidity           1.00000000      -0.02269729  0.289180698
## volatile.acidity       -0.02269729       1.00000000 -0.149471811
## citric.acid             0.28918070      -0.14947181  1.000000000
## residual.sugar          0.08902070       0.06428606  0.094211624
## chlorides               0.02308564       0.07051157  0.114364448
## free.sulfur.dioxide    -0.04939586      -0.09701194  0.094077221
## total.sulfur.dioxide    0.09106976       0.08926050  0.121130798
```

```
## density                0.26533101       0.02711385  0.149502571
## pH                    -0.42585829      -0.03191537 -0.163748211
## sulphates             -0.01714299      -0.03572815  0.062330940
## alcohol               -0.12088112       0.06771794 -0.075728730
## quality               -0.11366283      -0.19472297 -0.009209091
## bound_s02              0.13566071       0.15676923  0.102179337
## acidity                0.98717874       0.07157062  0.394143356
## alcohol_sugar_ratio   -0.06199198      -0.10395048 -0.033957333
##                     residual.sugar   chlorides free.sulfur.dioxide
## fixed.acidity           0.08902070  0.02308564       -0.0493958591
## volatile.acidity        0.06428606  0.07051157       -0.0970119393
## citric.acid             0.09421162  0.11436445        0.0940772210
## residual.sugar          1.00000000  0.08868454        0.2990983537
## chlorides               0.08868454  1.00000000        0.1013923521
## free.sulfur.dioxide     0.29909835  0.10139235        1.0000000000
## total.sulfur.dioxide    0.40143931  0.19891030        0.6155009650
## density                 0.83896645  0.25721132        0.2942104109
## pH                     -0.19413345 -0.09043946       -0.0006177961
## sulphates              -0.02666437  0.01676288        0.0592172458
## alcohol                -0.45063122 -0.36018871       -0.2501039415
## quality                -0.09757683 -0.20993441        0.0081580671
## bound_s02               0.34484449  0.19379550        0.2635372837
## acidity                 0.10473749  0.04552987       -0.0451333172
## alcohol_sugar_ratio    -0.78536189 -0.10217918       -0.2848408090
##                     total.sulfur.dioxide     density            pH
## fixed.acidity              0.091069756  0.26533101 -0.4258582910
## volatile.acidity           0.089260504  0.02711385 -0.0319153683
## citric.acid                0.121130798  0.14950257 -0.1637482114
## residual.sugar             0.401439311  0.83896645 -0.1941334540
## chlorides                  0.198910300  0.25721132 -0.0904394560
## free.sulfur.dioxide        0.615500965  0.29421041 -0.0006177961
## total.sulfur.dioxide       1.000000000  0.52988132  0.0023209718
## density                    0.529881324  1.00000000 -0.0935914935
## pH                         0.002320972 -0.09359149  1.0000000000
## sulphates                  0.134562367  0.07449315  0.1559514973
## alcohol                   -0.448892102 -0.78013762  0.1214320987
## quality                   -0.174737218 -0.30712331  0.0994272457
## bound_s02                  0.922482350  0.50444690  0.0031433874
## acidity                    0.113188502  0.27560881 -0.4306513315
## alcohol_sugar_ratio       -0.413235753 -0.67313846  0.1340508625
##                        sulphates     alcohol      quality    bound_s02
## fixed.acidity       -0.017142985 -0.12088112 -0.113662831  0.135660713
## volatile.acidity    -0.035728147  0.06771794 -0.194722969  0.156769227
## citric.acid          0.062330940 -0.07572873 -0.009209091  0.102179337
## residual.sugar      -0.026664366 -0.45063122 -0.097576829  0.344844495
## chlorides            0.016762884 -0.36018871 -0.209934411  0.193795498
## free.sulfur.dioxide  0.059217246 -0.25010394  0.008158067  0.263537284
## total.sulfur.dioxide 0.134562367 -0.44889210 -0.174737218  0.922482350
## density              0.074493149 -0.78013762 -0.307123313  0.504446902
## pH                   0.155951497  0.12143210  0.099427246  0.003143387
## sulphates            1.000000000 -0.01743277  0.053677877  0.135693943
```

```
## alcohol              -0.017432772  1.00000000  0.435574715 -0.426923036
## quality               0.053677877  0.43557472  1.000000000 -0.217867760
## bound_s02             0.135693943 -0.42692304 -0.217867760  1.000000000
## acidity              -0.011852249 -0.11751272 -0.131377207  0.160645384
## alcohol_sugar_ratio   0.005781016  0.38630792  0.063144479 -0.366266429
##                          acidity alcohol_sugar_ratio
## fixed.acidity         0.98717874        -0.061991985
## volatile.acidity      0.07157062        -0.103950477
## citric.acid           0.39414336        -0.033957333
## residual.sugar        0.10473749        -0.785361890
## chlorides             0.04552987        -0.102179178
## free.sulfur.dioxide  -0.04513332        -0.284840809
## total.sulfur.dioxide  0.11318850        -0.413235753
## density               0.27560881        -0.673138463
## pH                   -0.43065133         0.134050863
## sulphates            -0.01185225         0.005781016
## alcohol              -0.11751272         0.386307923
## quality              -0.13137721         0.063144479
## bound_s02             0.16064538        -0.366266429
## acidity               1.00000000        -0.075341319
## alcohol_sugar_ratio  -0.07534132         1.000000000
```



```
## Warning: Removed 29 rows containing missing values (geom_point).
```



Description Three

As we were able to see, it was helpful to understand the relationships amongst each feature, including the features that I added to the study. From our clusters we come to confirm what we know about chemicals in the following ways.

Density is defined as mass / volume. It is positively correlated with residual sugar because sugar molecules will add to the volume, filling up the space, increasing the density. Given they are mixed within the process, the sugar will mix evenly, thus increasing the overall density of the content.

We know that pH levels are negatively affected by acids so it's not surprising that the pH variable maintains negative relationships with acid.

Because alcohol is less dense than water, (wine density is often close to water) it has a negative effect on the density variable. That being said, chlorides positive correlation with density infers the opposite, since chloride increases the density as well.

Reflection

The dataset represents 4938 white wines of the Portuguese “Vinho Verde” type which had 11 physiochemical variables included along with one sensory output variable ‘quality’. I looked at each variable up close, and then explored variable interactions to learn more about the make up of the dataset. I used techniques like Principal Component Analysis and Exploratory Factor Analysis to look for latent variable combinations and I also used statistical methods to understand the dataset graphically and numerically. In the end I created a linear model to predict the quality of wine.

The amount of alcohol in the white wines, in this dataset, had the highest correlation coefficient with quality and remains the biggest predictor. In my linear model, I was able to achieve an Multiple R-squared of 0.546 but because my model takes in so many of the variables it may suffer from overfitting. After being able to achieve 0.7825 Multiple R-Squared value, I believe that with some work and possibly using a non-linear model that better predictions can be had for that dataset using machine learning techniques.

The limitations of this data came up through my analysis were plenty. I wanted to focus on the substantive chemical formulas but because the data did not give the volume I wasn’t able easily compare the chemical makeup in probabilities. I also did not know if each observation was a wine bottle or a vat of wine of some sort. The data was cross sectional so I wasn’t able to understand how these changes might happen over time and if that has an effect on quality.