Red Wine Analysis by Bindu Paul

```
##  [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"
```

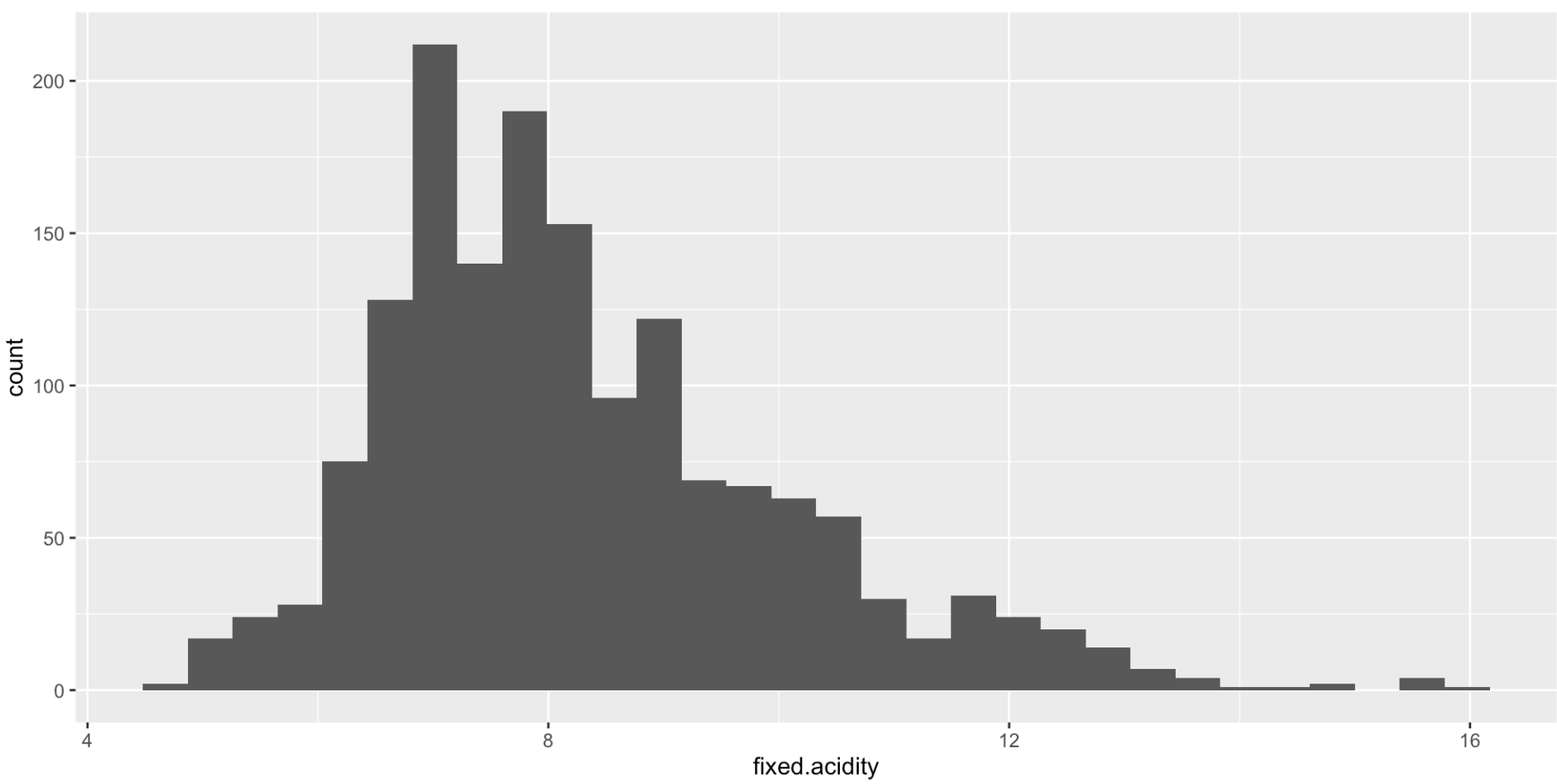
```
##        X          fixed.acidity   volatile.acidity  citric.acid
##  Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000
##  1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090
##  Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260
##  Mean   : 800.0   Mean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

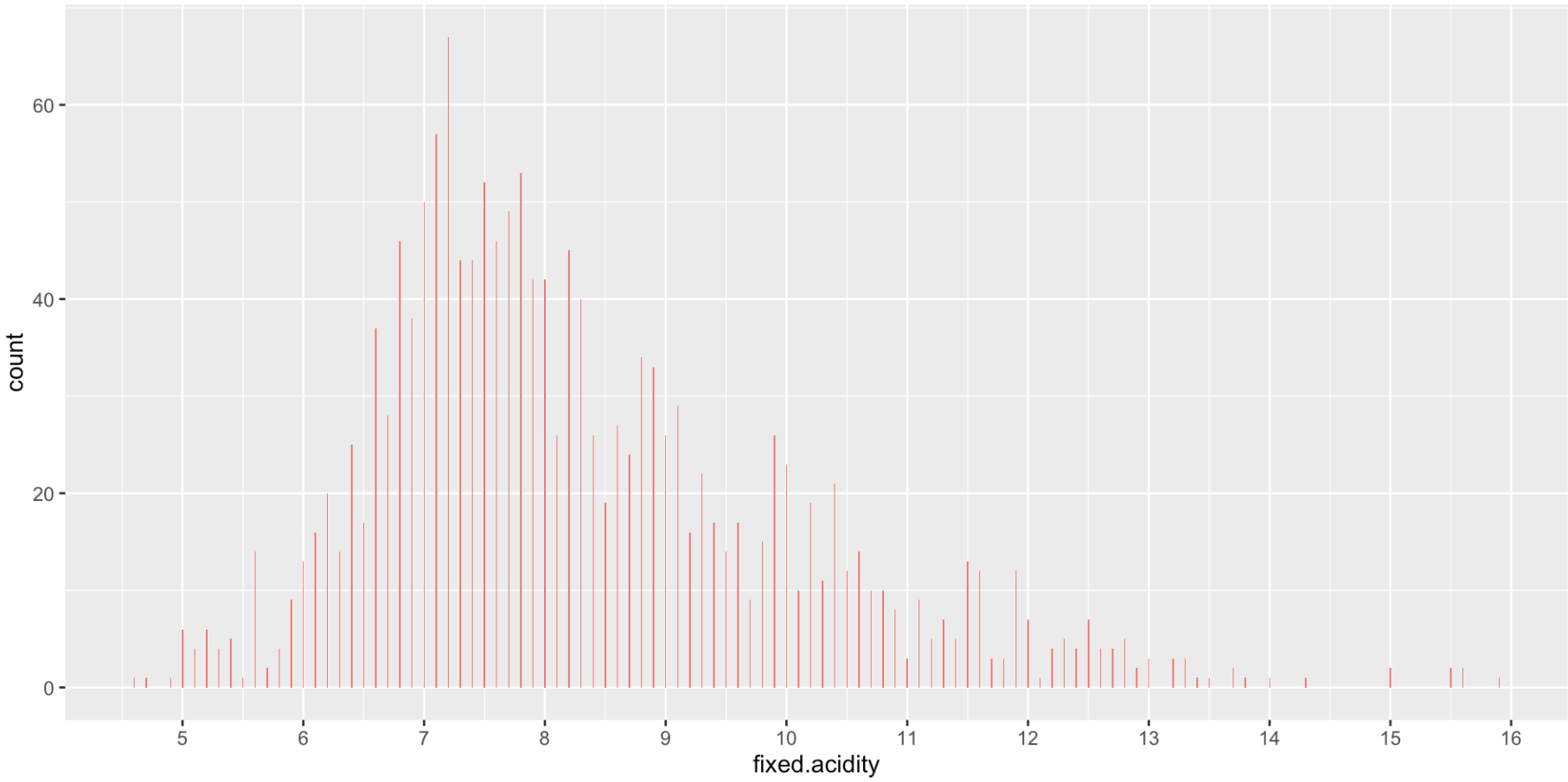
```
## 'data.frame':    1599 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.07
3 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

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

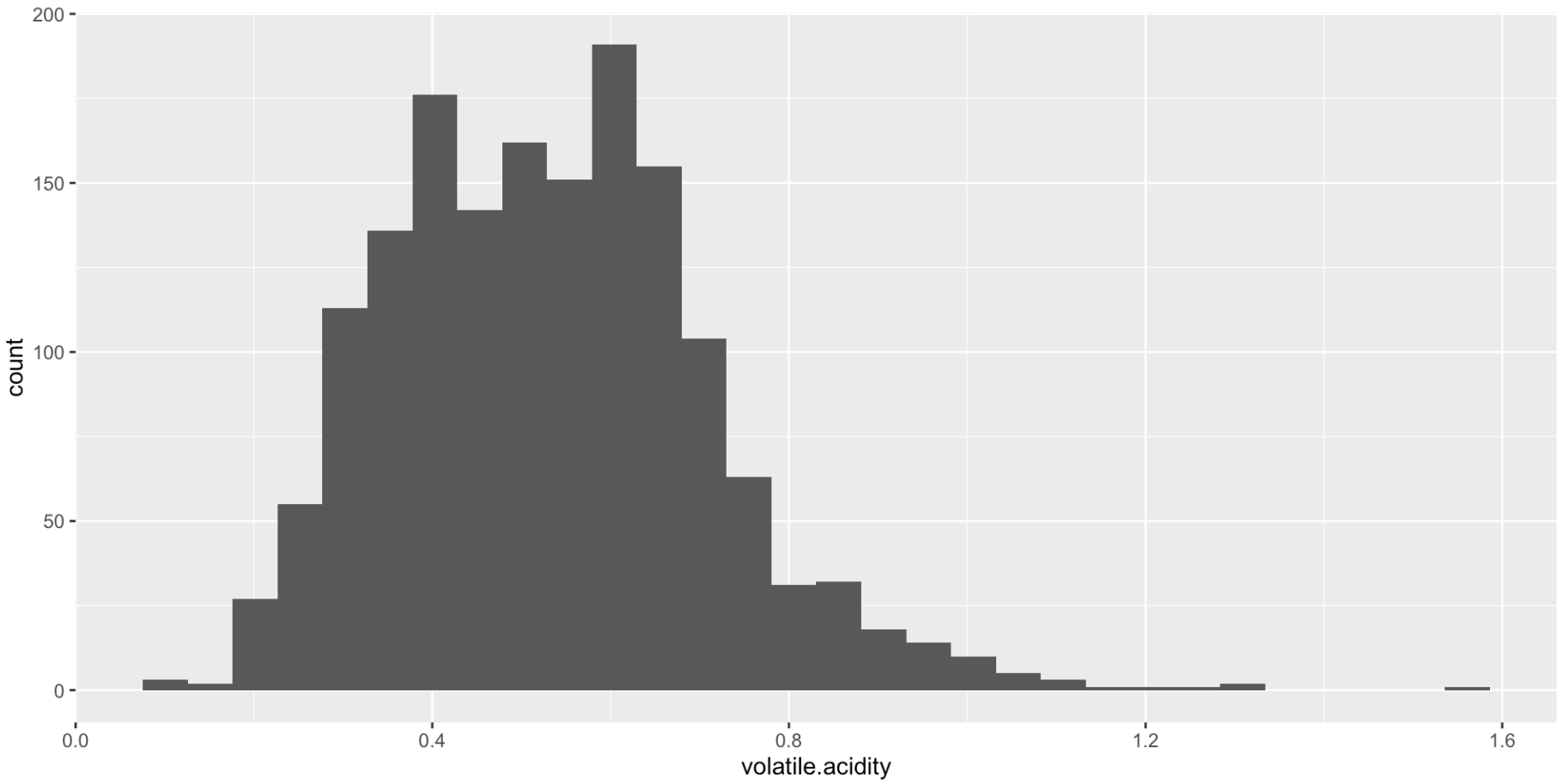
Mean quality is 5.636. Mean alcohol is 10.42. Median total.sulfur.dioxide is 38 while the max is 289, which suggests an outlier. 1599 observations and 13 variables.

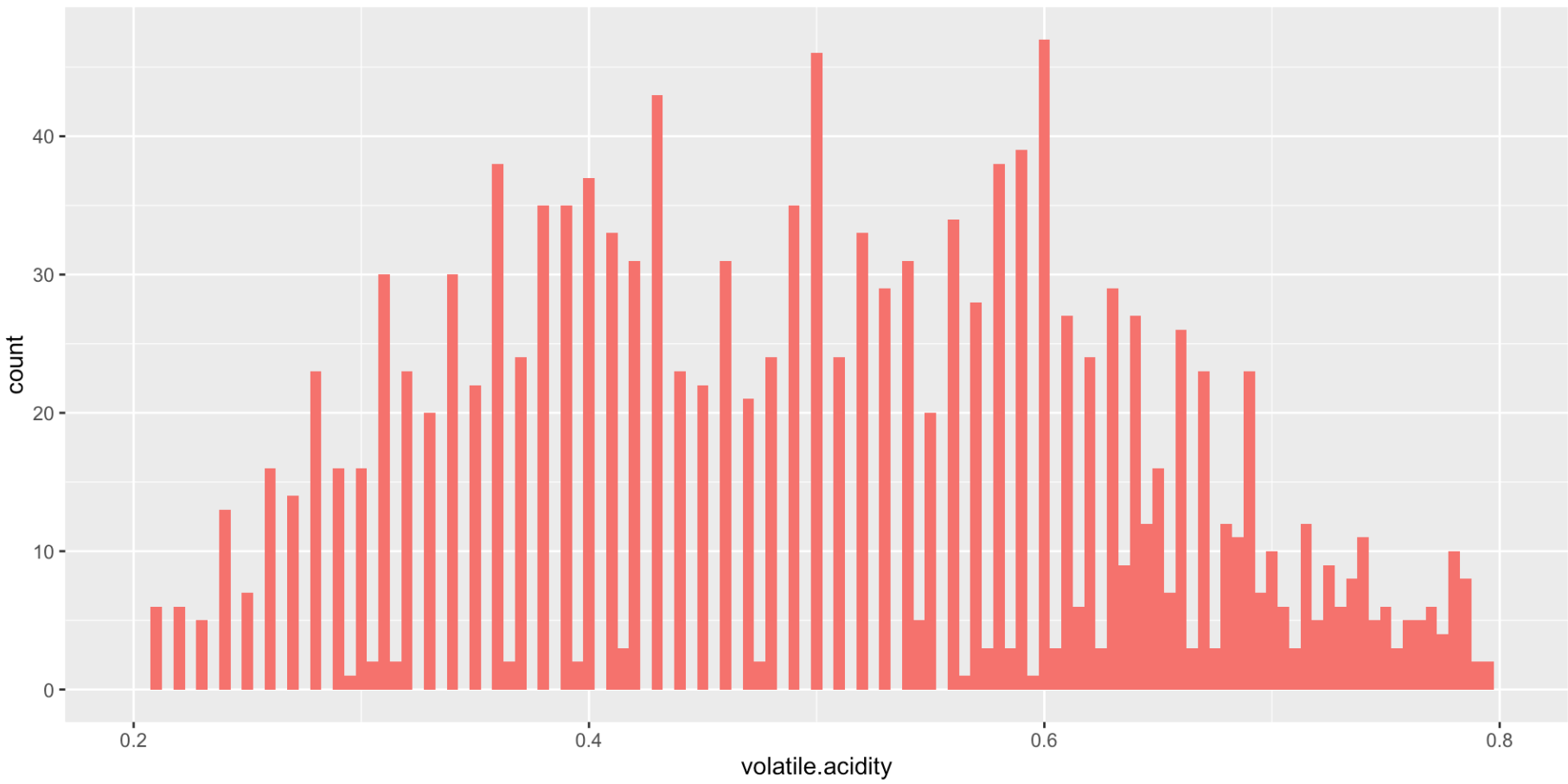
Univariate Plots Section



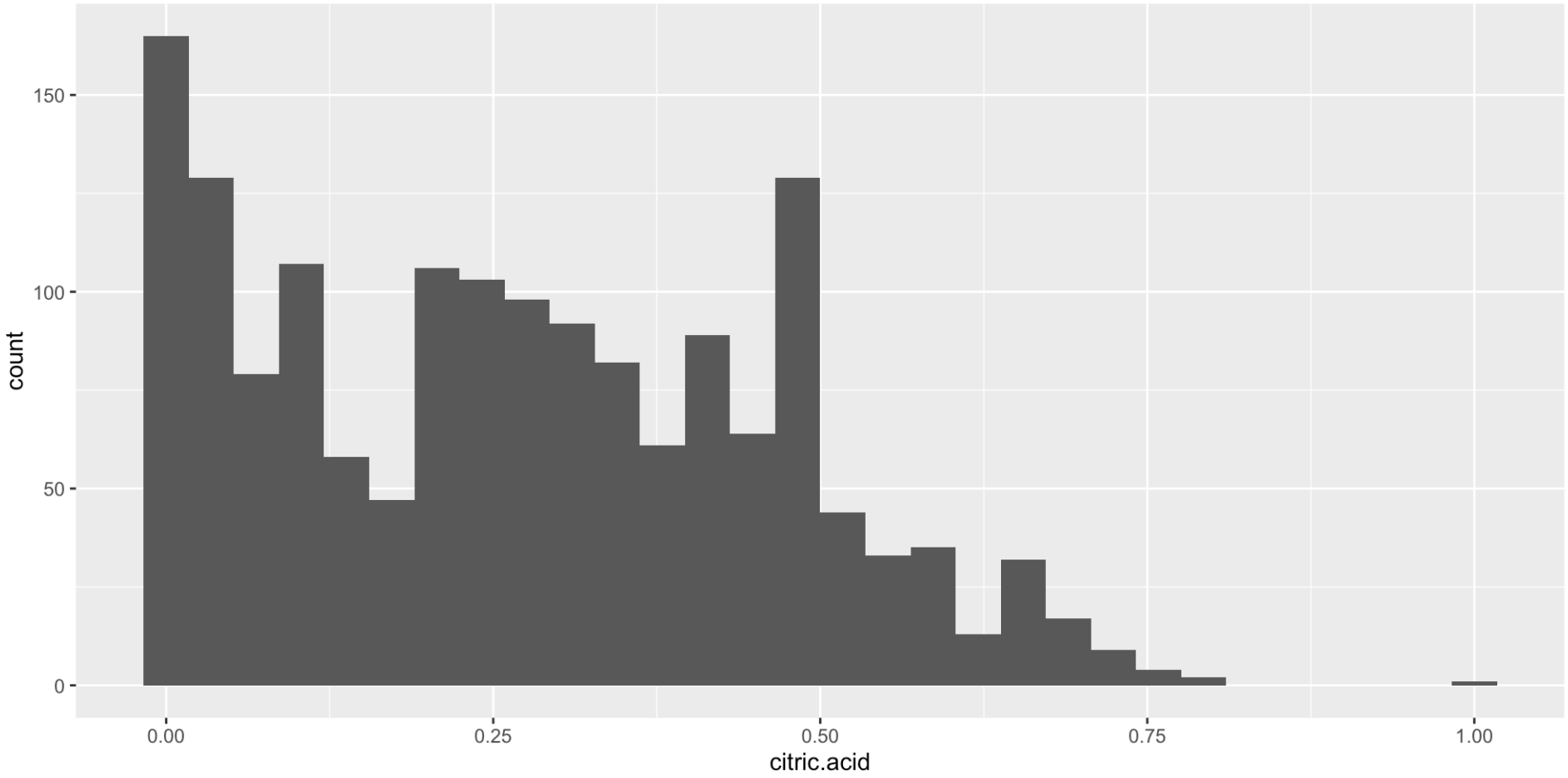


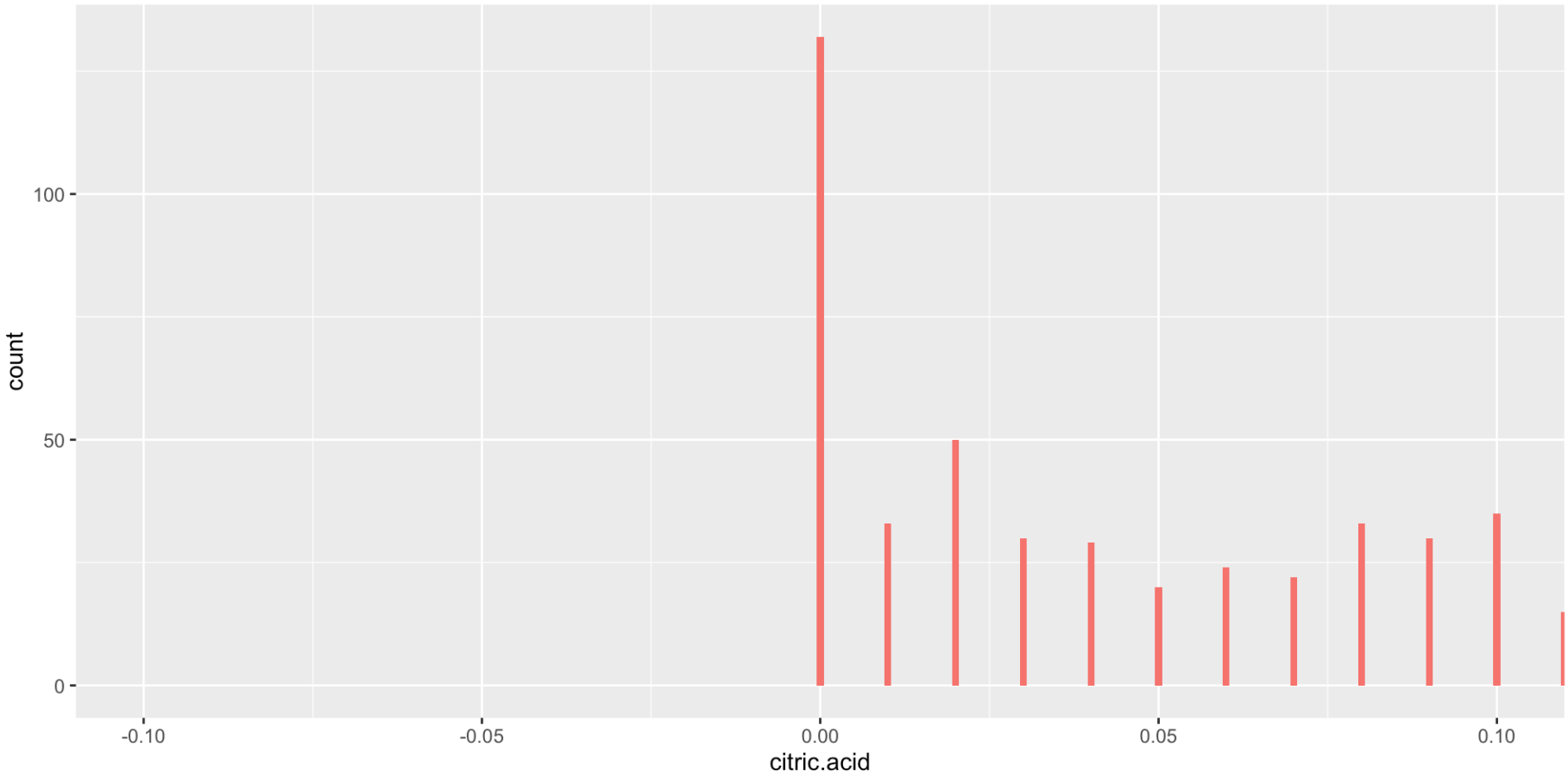
The data was transformed with scale_x_continuous and binwidth of 0.1. Most fixed acidity ranges from x = 7-8.

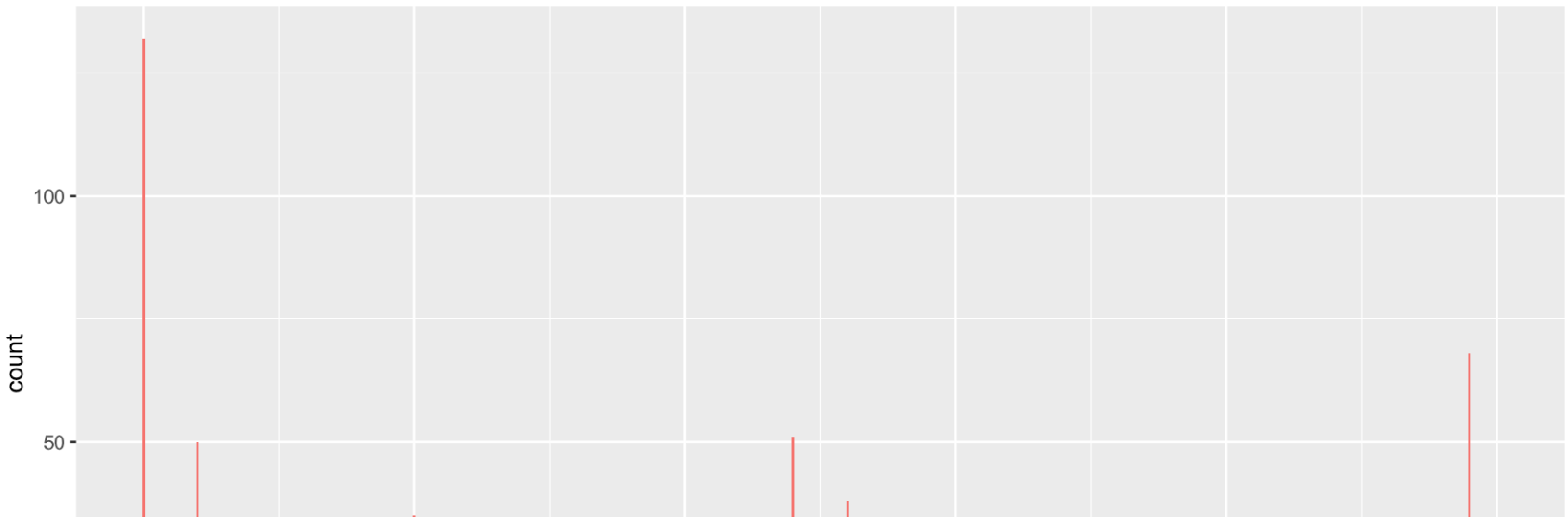


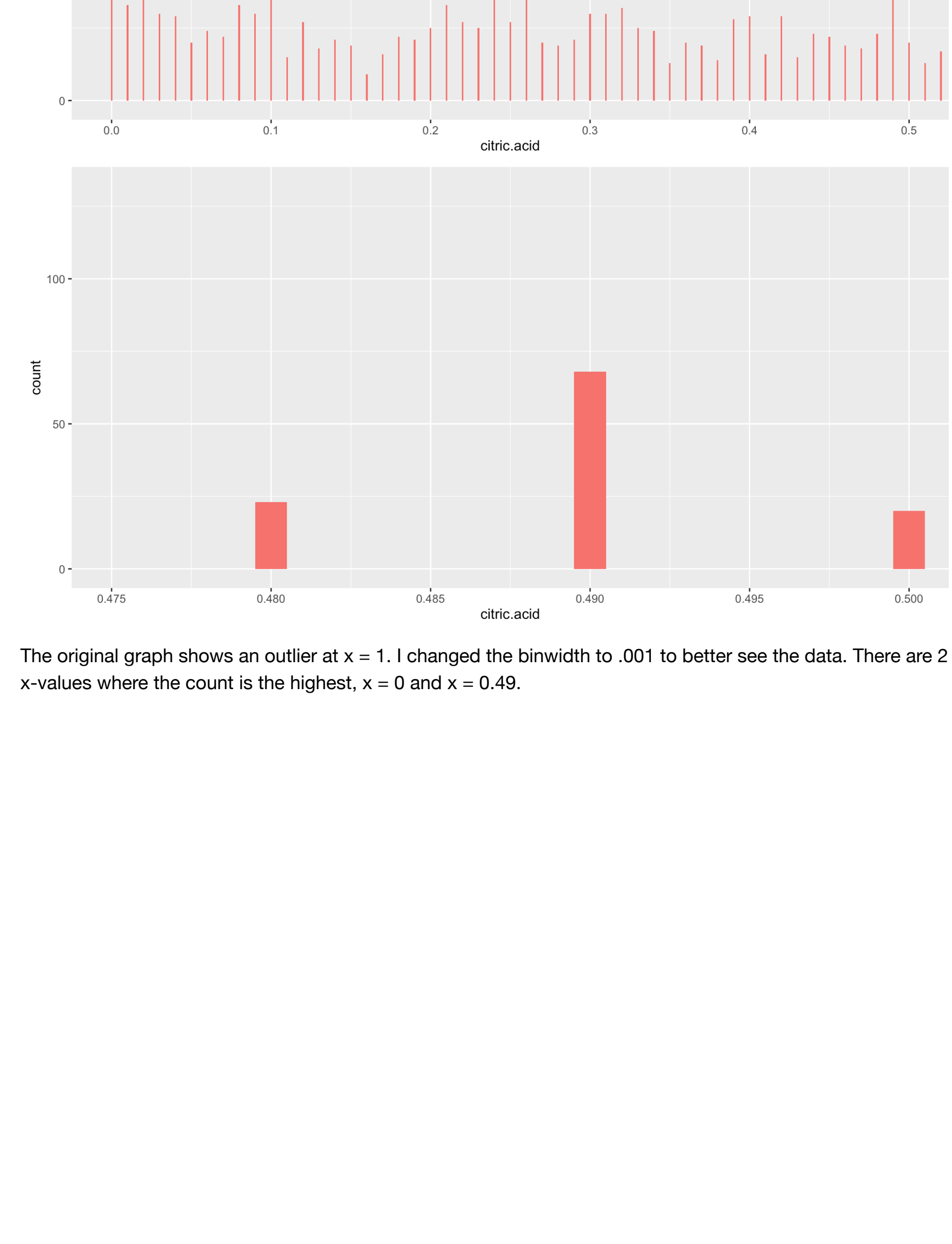


Prior to transformation we notice an outlier of x = 1.6. After changing the binwidth to 0.005 and limiting the x-axis, the graph shows a very loose bimodal peak. I wonder how this graph relates to fixed.acitiy.

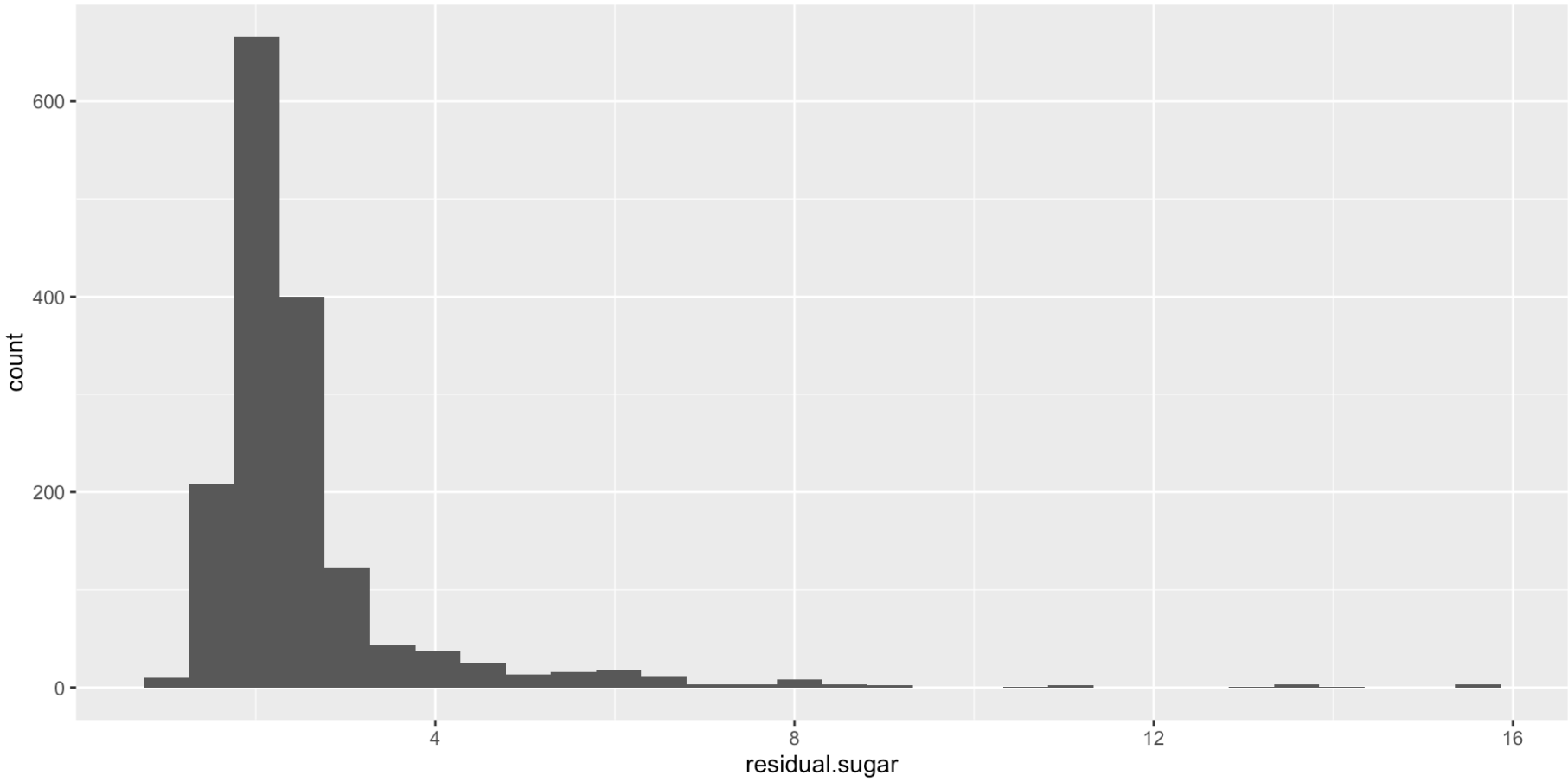

150
100
50
0
count
0.00
0.25
0.50
0.75
1.00
citric.acid

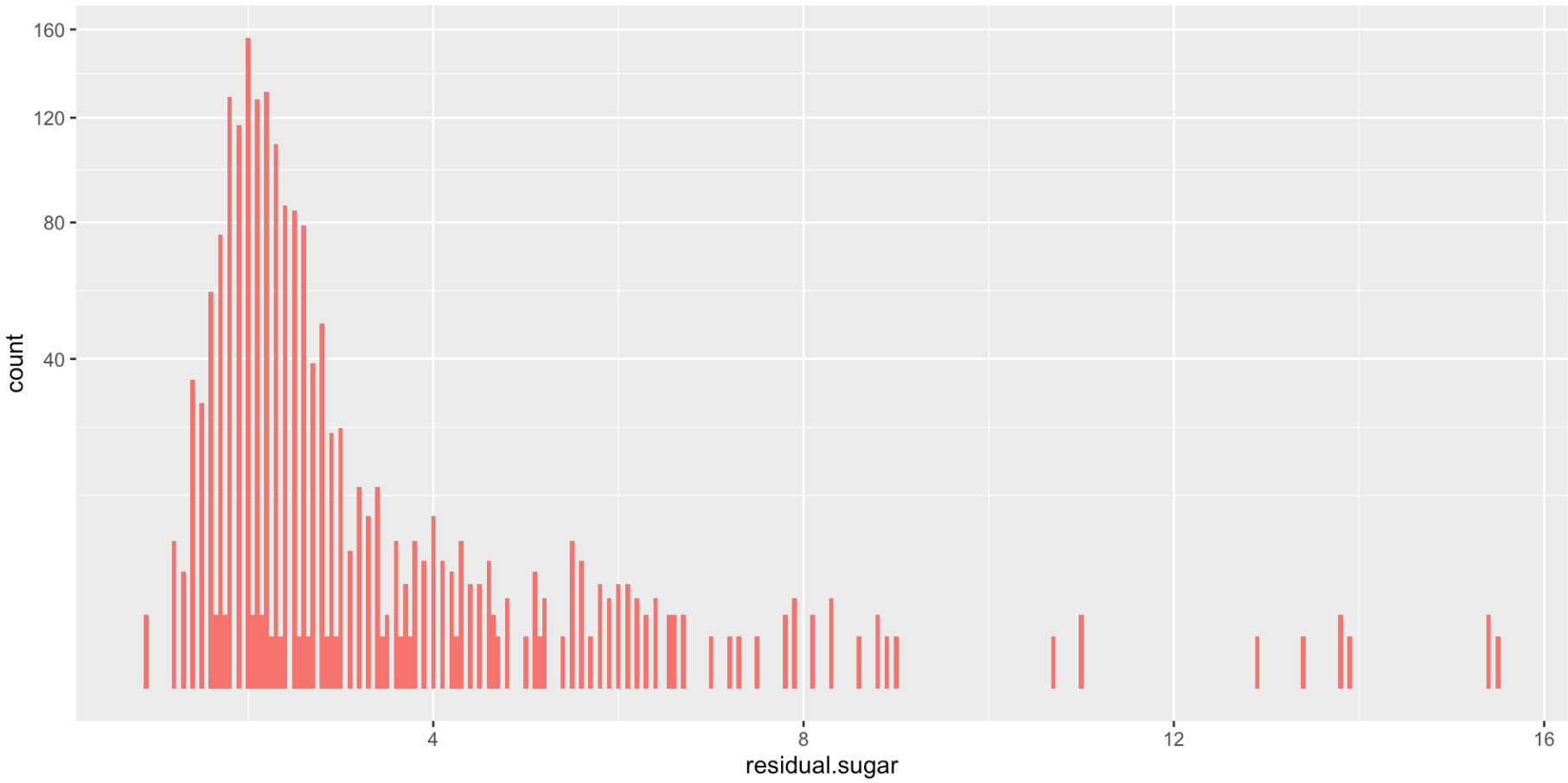

100
50
0
count
-0.10
-0.05
0.00
0.05
0.10
citric.acid

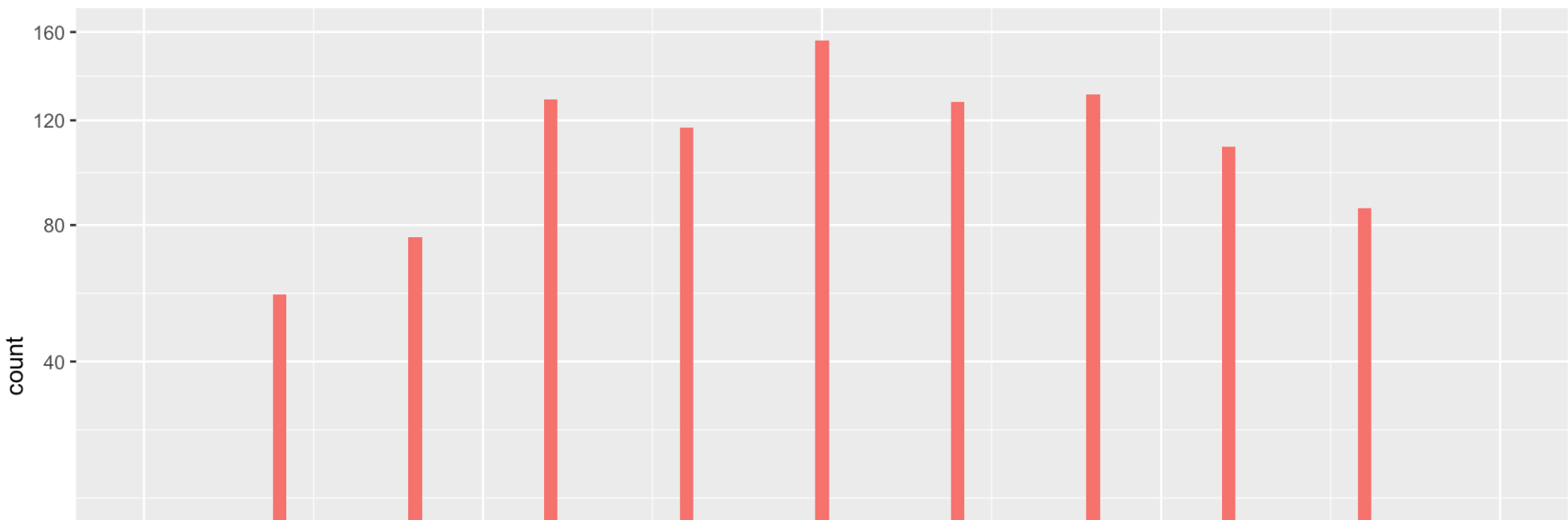

100
50
count



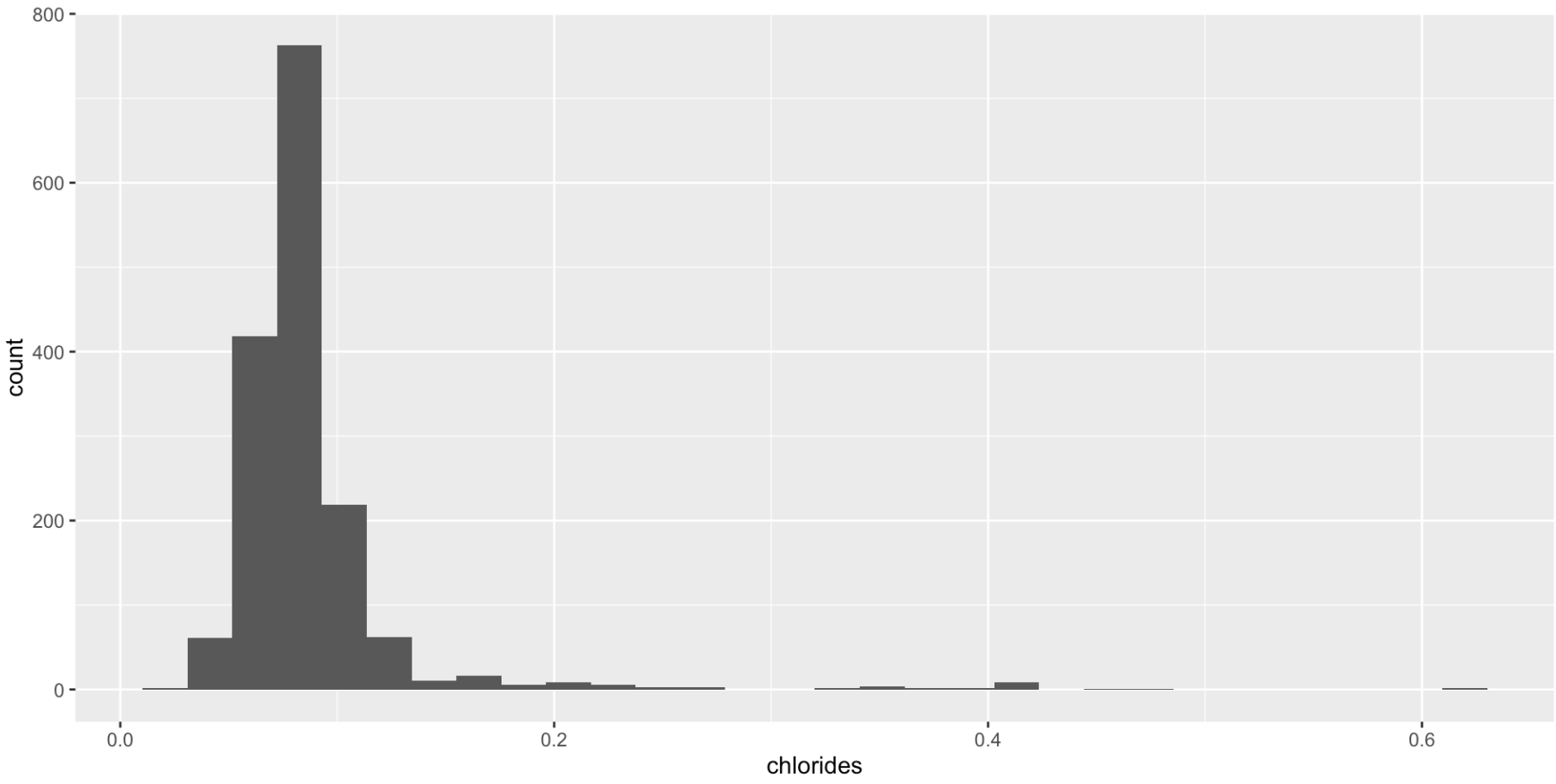
The original graph shows an outlier at x = 1. I changed the binwidth to .001 to better see the data. There are 2 x-values where the count is the highest, x = 0 and x = 0.49.

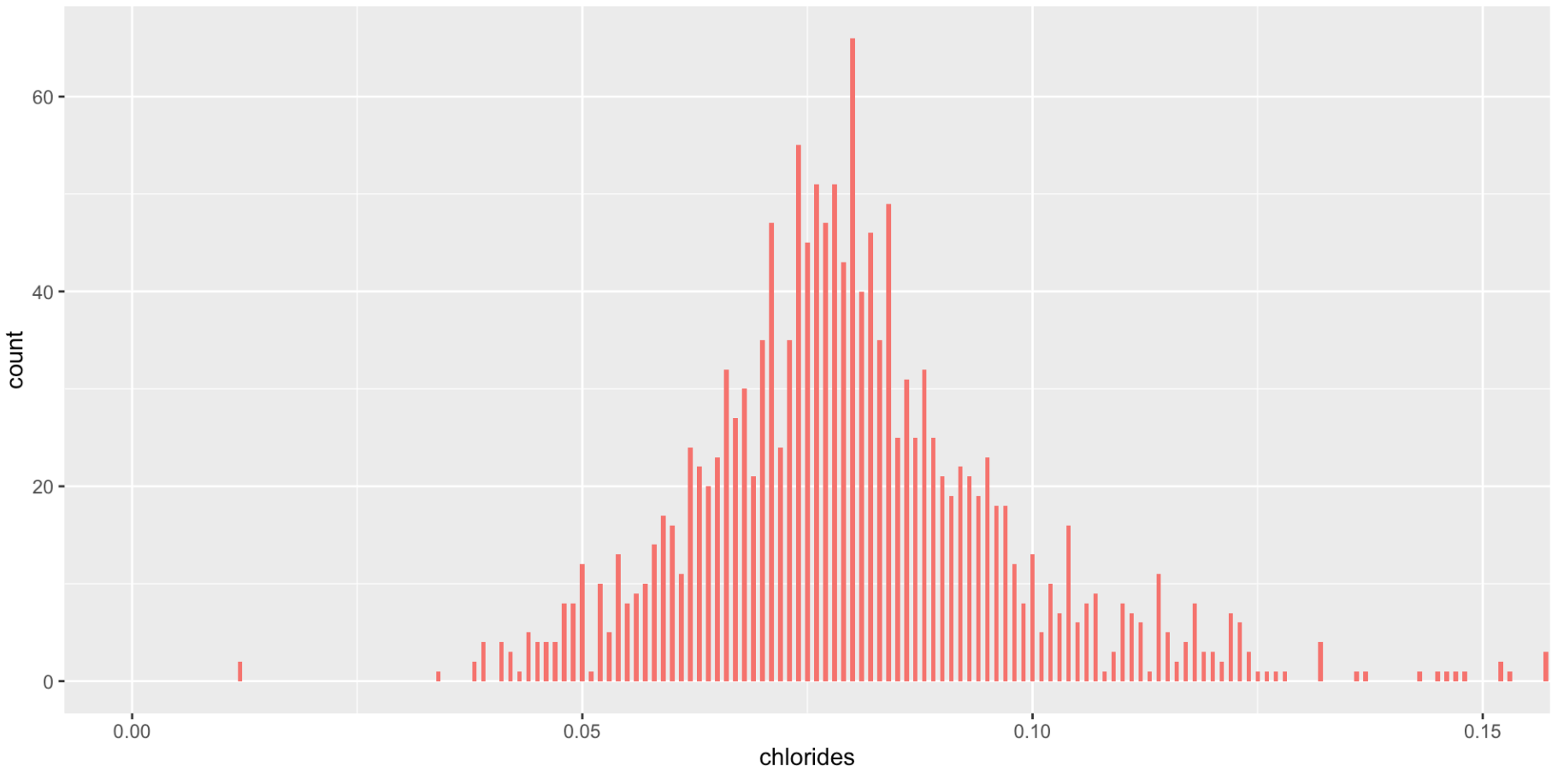

600
400
200
0
count
4
8
12
16
residual.sugar


160
120
80
40
count
4
8
12
16
residual.sugar

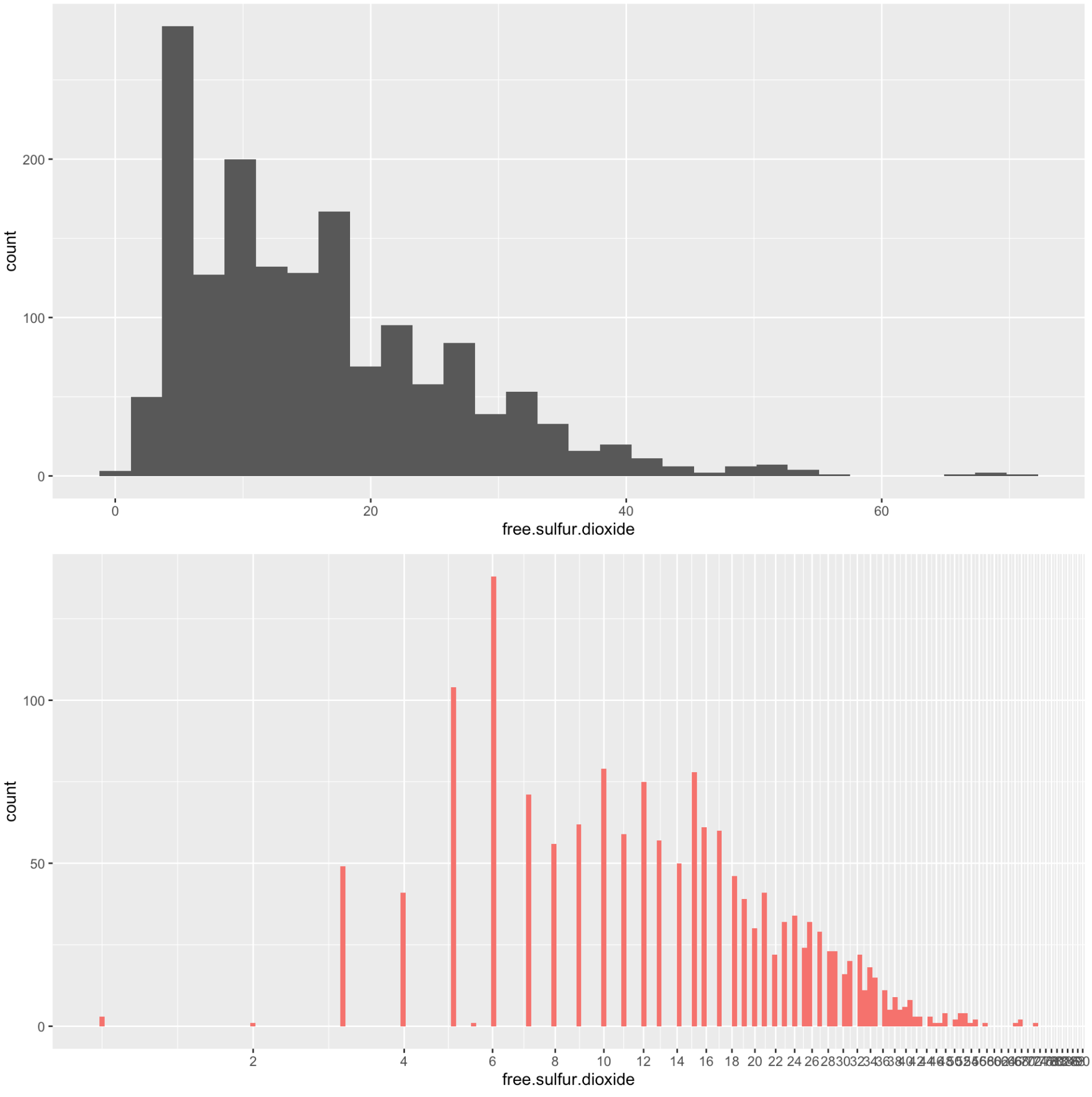

160
120
80
40
count

Majority of residual sugar is less than x = 4. I transformed the y-axis by sqrt and binwidth to .05 to better see the bars. Data does not appear in some ranges such as x = 10, 12, 14, I wonder why. I limited the data to get a closer look at the peak. Most of the data falls in the range of x = (1.5:2.5) with x = 2 being the most.



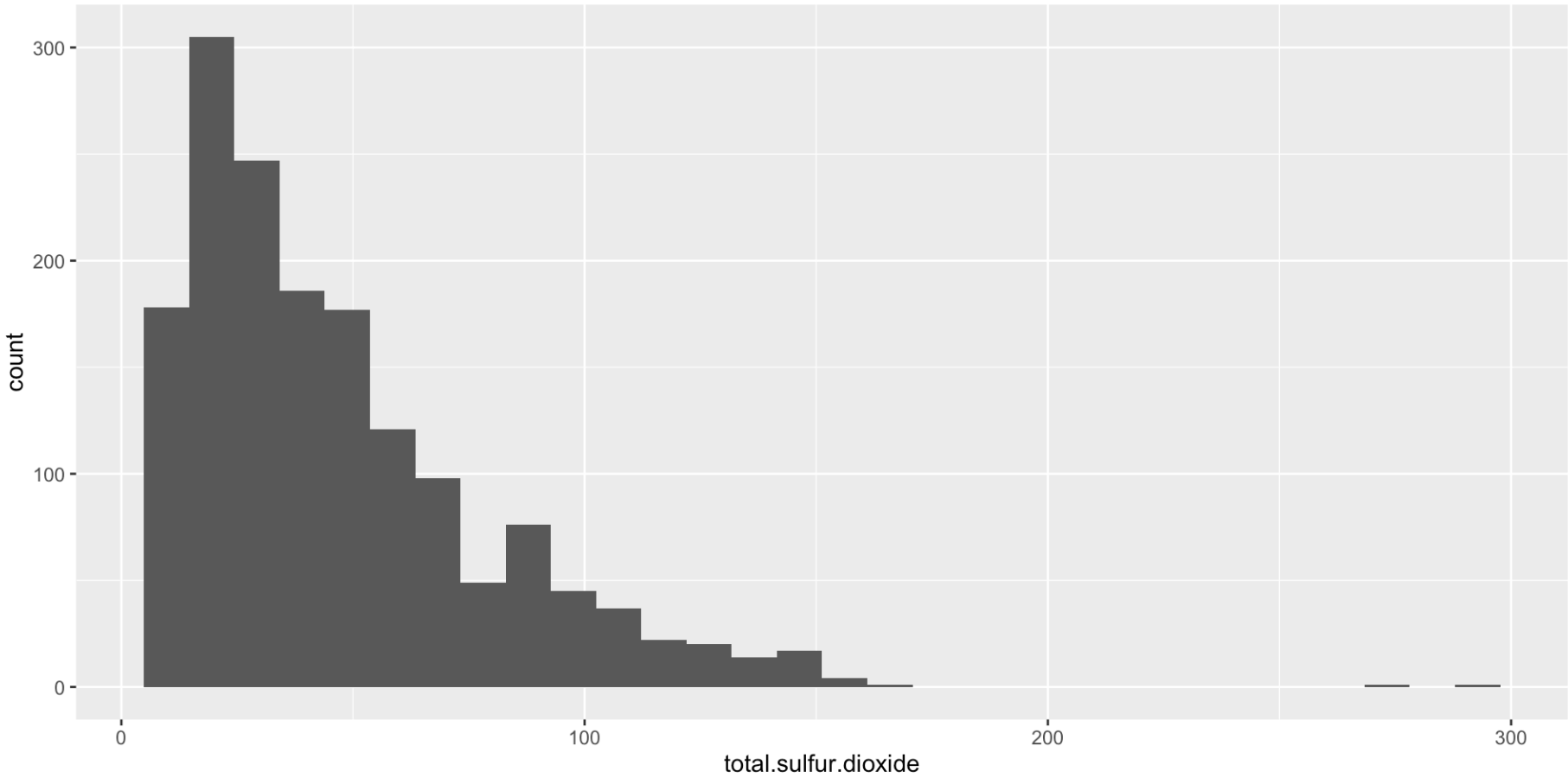


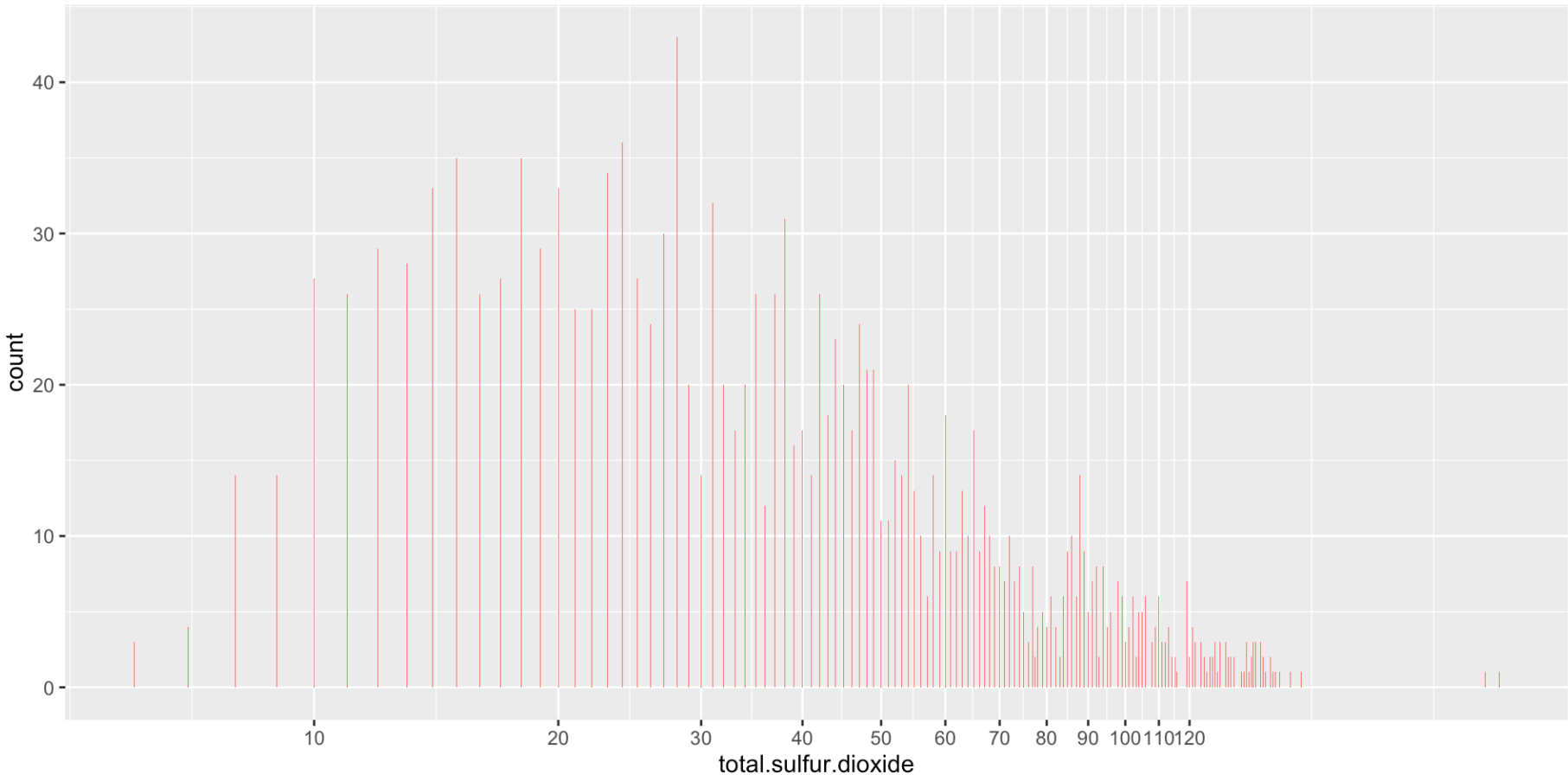
An overwhelming amout of chlorides are below 0.2. Changing the binwidth to .0005 and limiting the X-axis between (0, .15) we see that the range for chlorides is 0.05:0.1 with a normal distribution.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
```

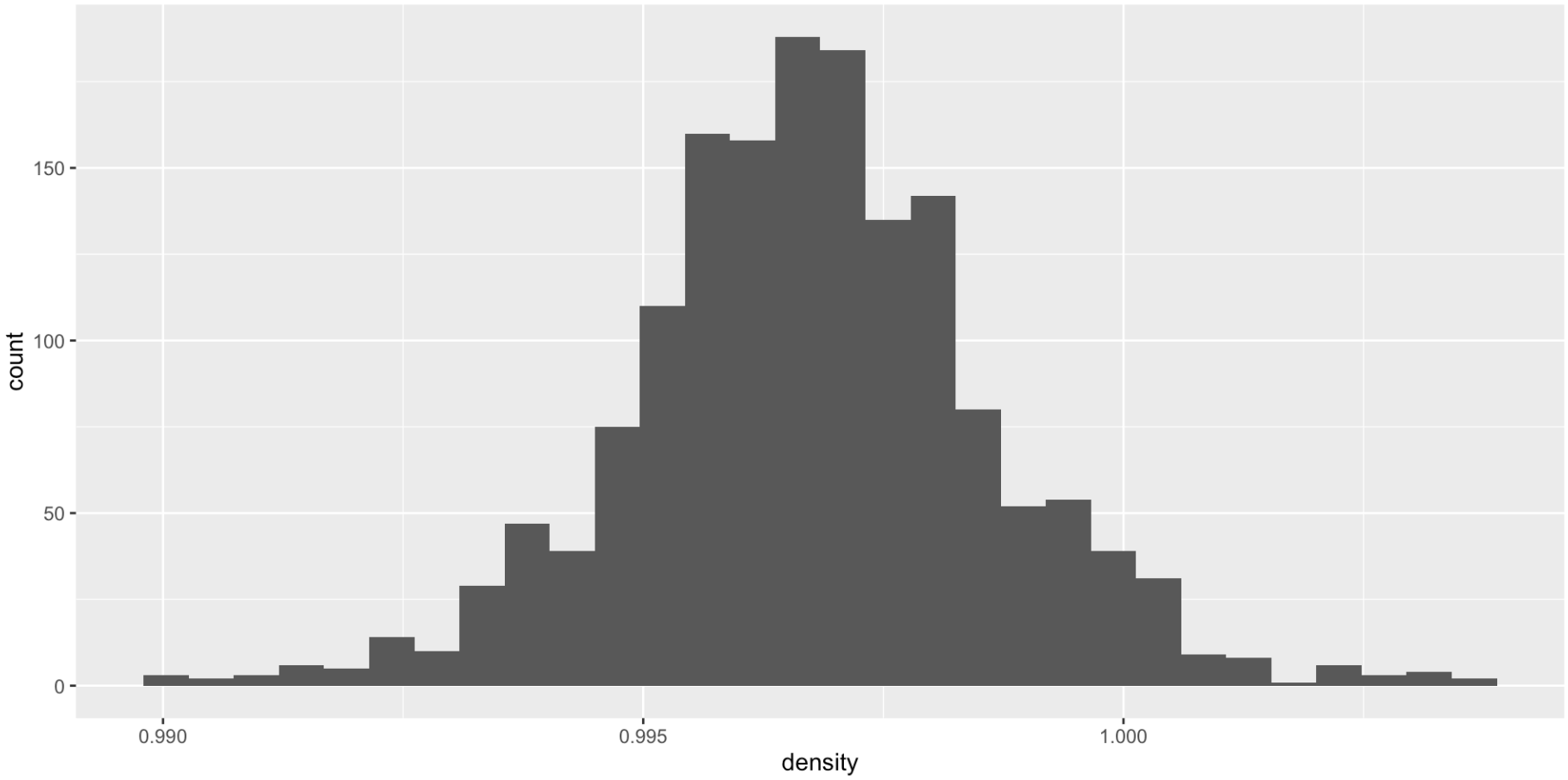
The graph is skewed to the right so I transformed it using log10(). This shows the value of free.sulfur.dioxide is prominent between 5-15. The max count is where x = 6 and the median x-value is 14.

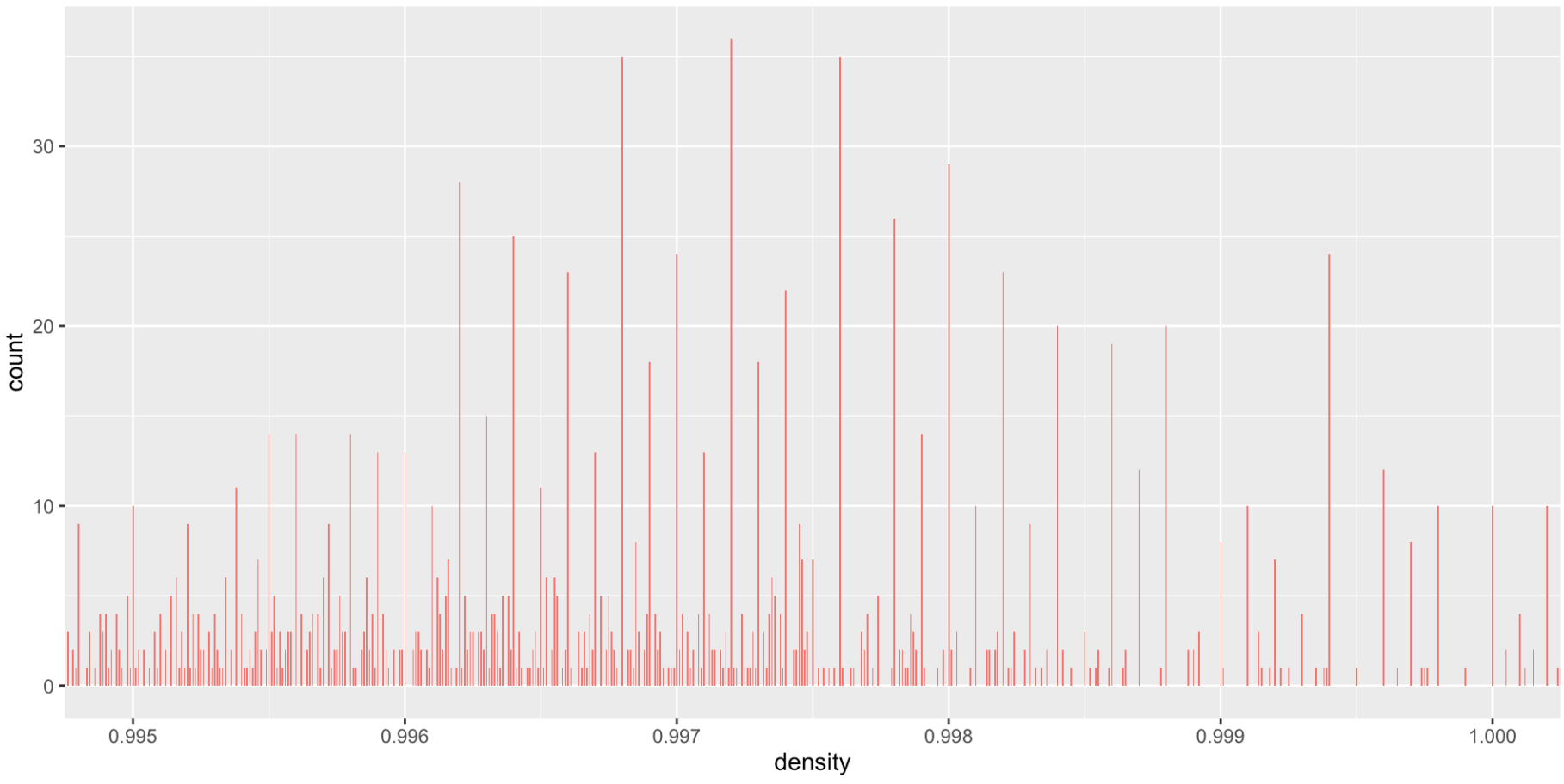




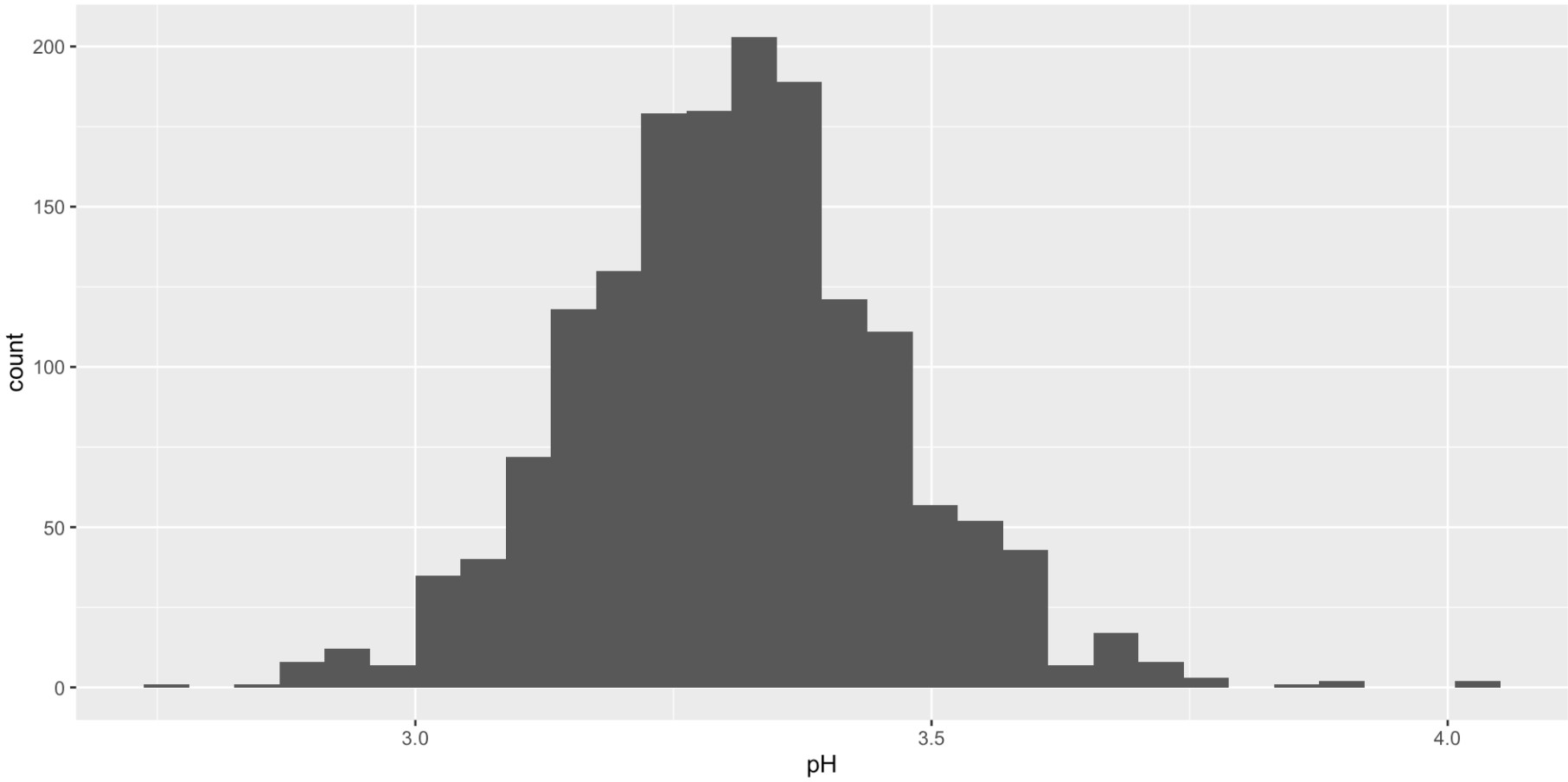

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00
```

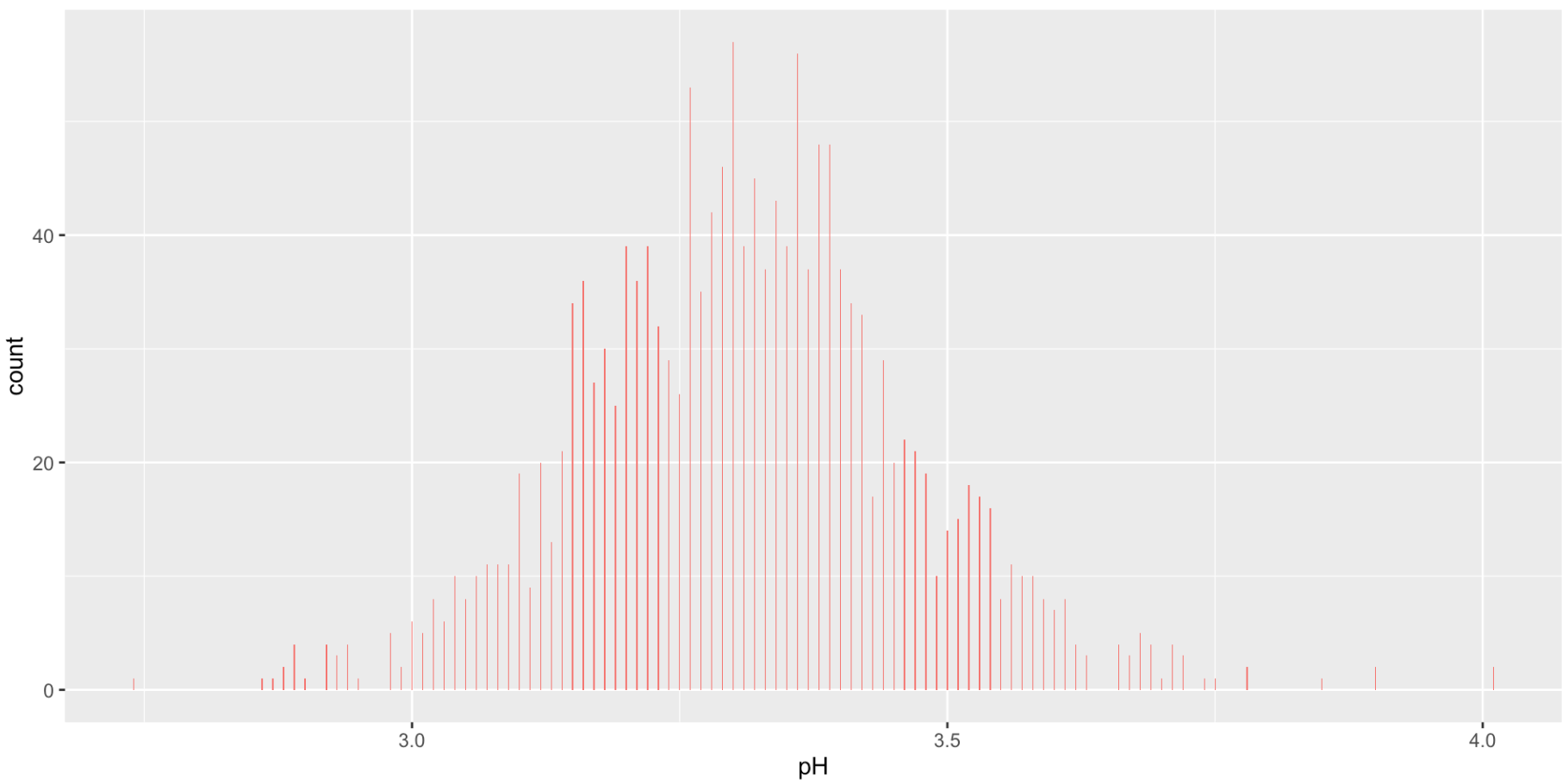
I log transformed the x-axis by log10 and binwidth to .001 since it was also scewed to the right. Compared to free sulfur dioxide, there is a better range of data. We have a max value of 289 but median of 38. The outliers are probably affecting the data.



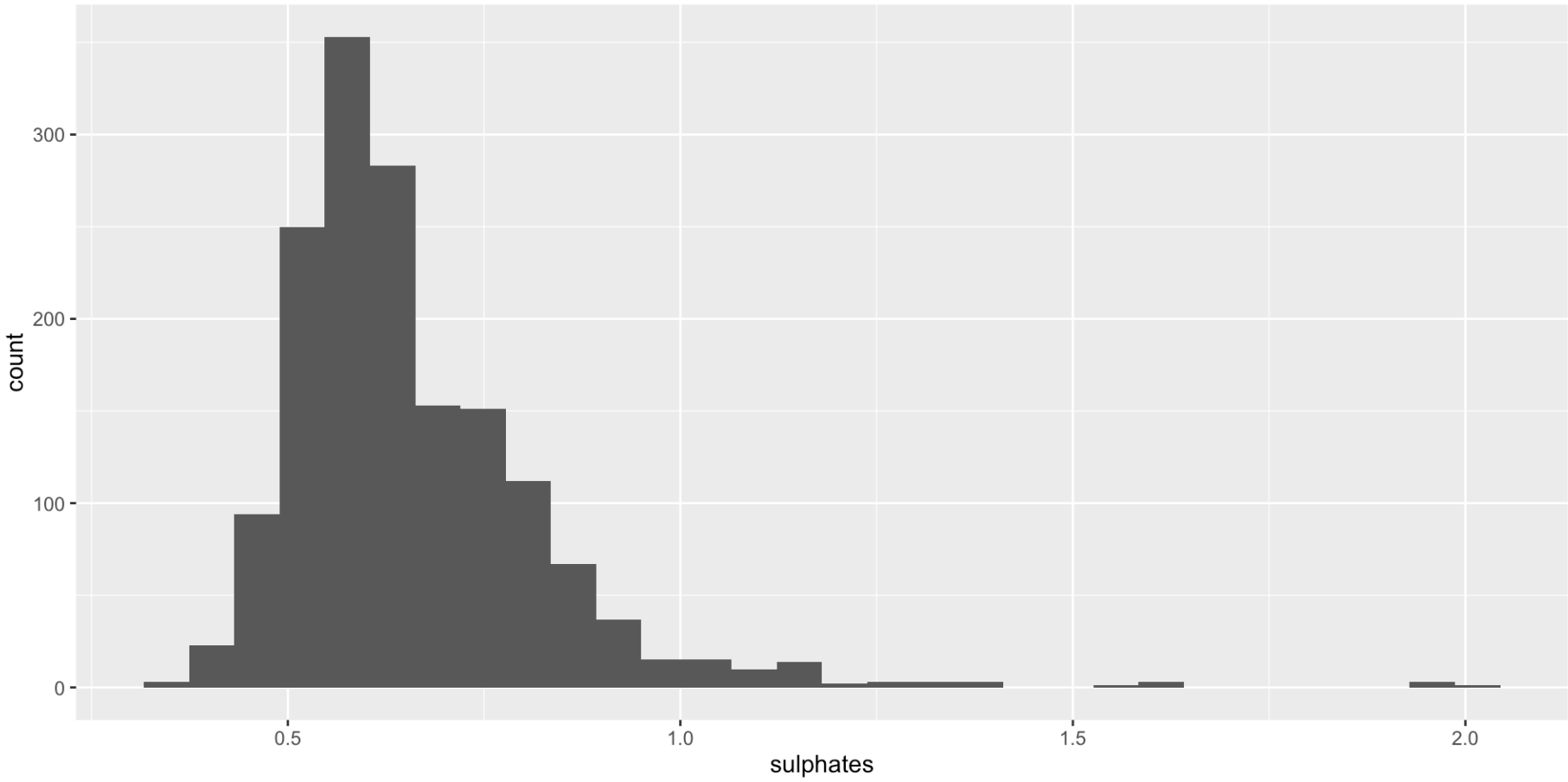


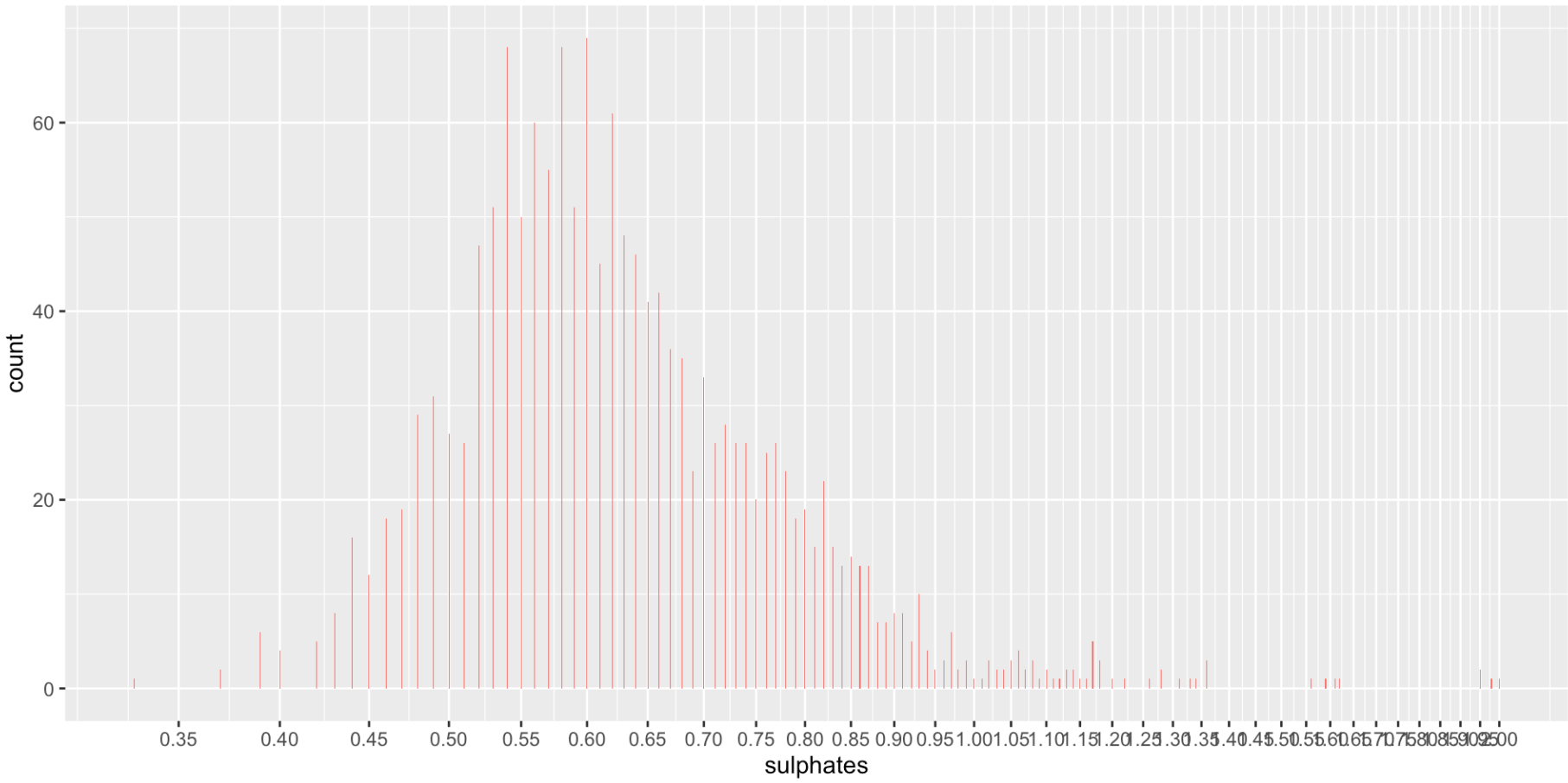
Normal distribution. I transformed the graph by changing the binwidth to .000005 and limiting the x-axis to (.995:1). The density values are very different from one another since the max count is about 37.





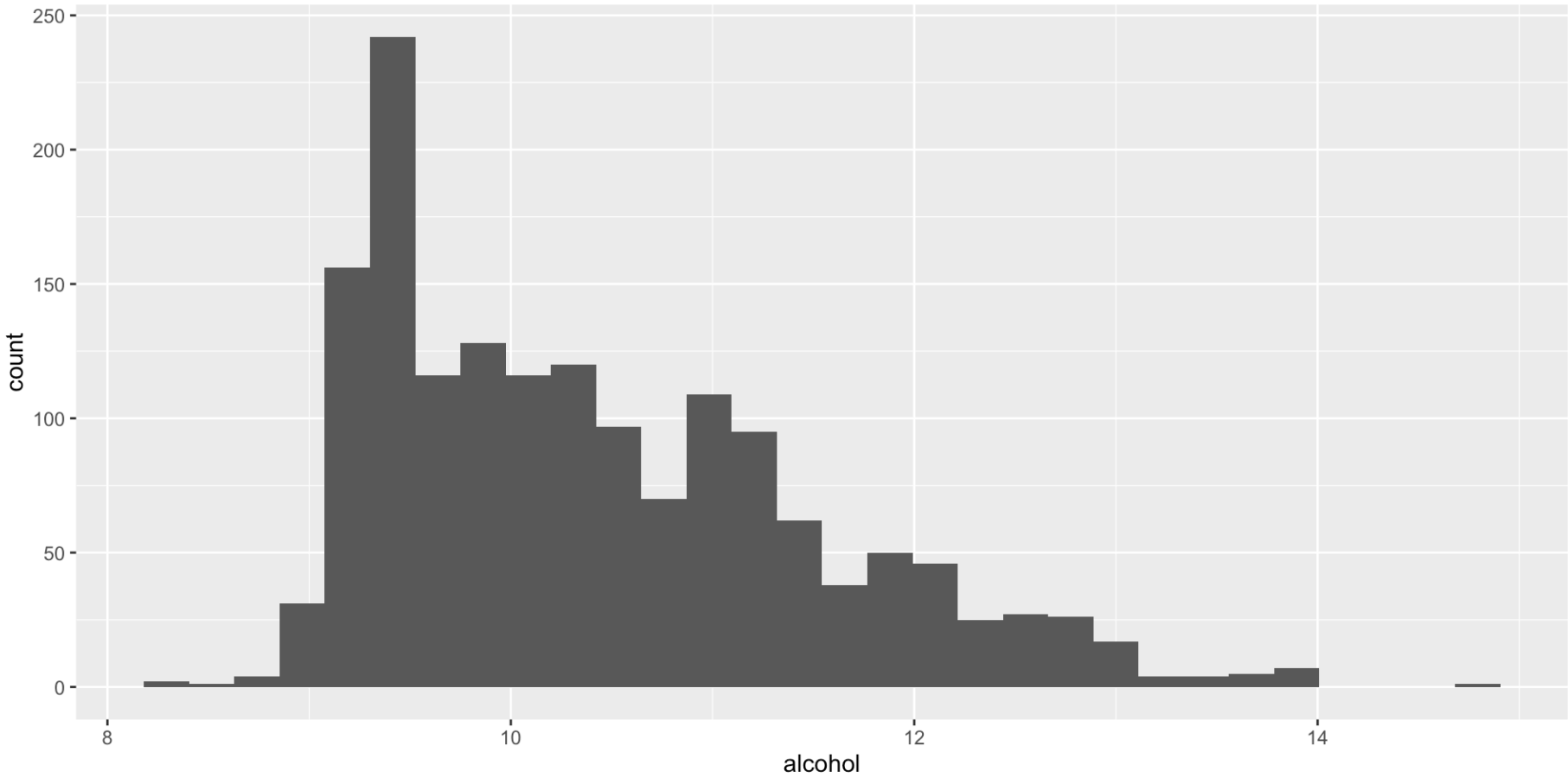
Normal distribution. Changed the binwidth to .001 to see the bars better. The prime pH range is 3:3.5. This makes sense since the average for red wine is between 3.3 to 3.6.

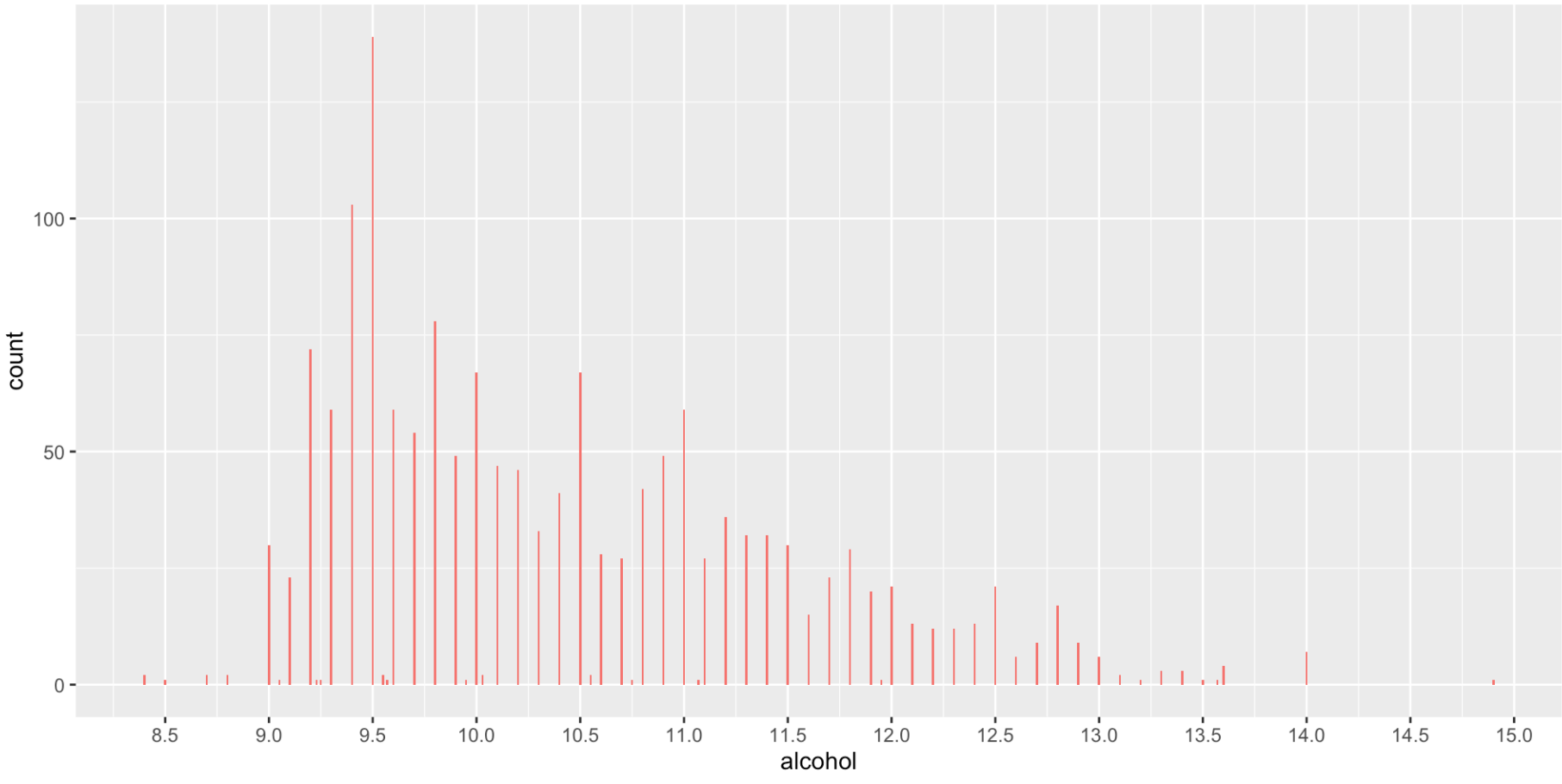




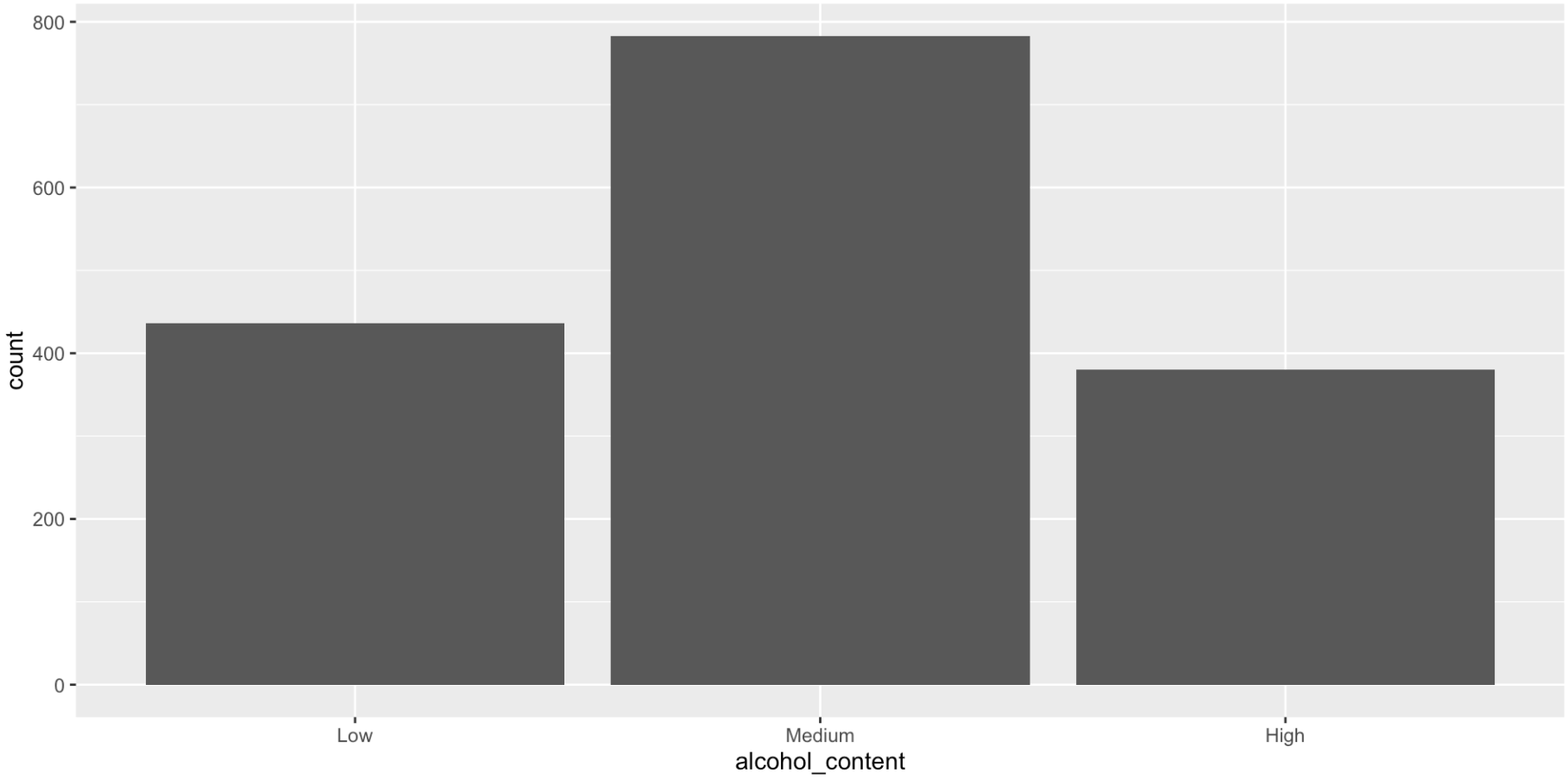

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```

We have outliers around x = 2. This graph is skewed to the the right so I transformed x-axis by log10() and binwidth to .0005. The values are very small and are prominent from 0.5:0.6, the Mean is x = 0.658 and median = 0.62. We can see the outliers don't affect the data as much.

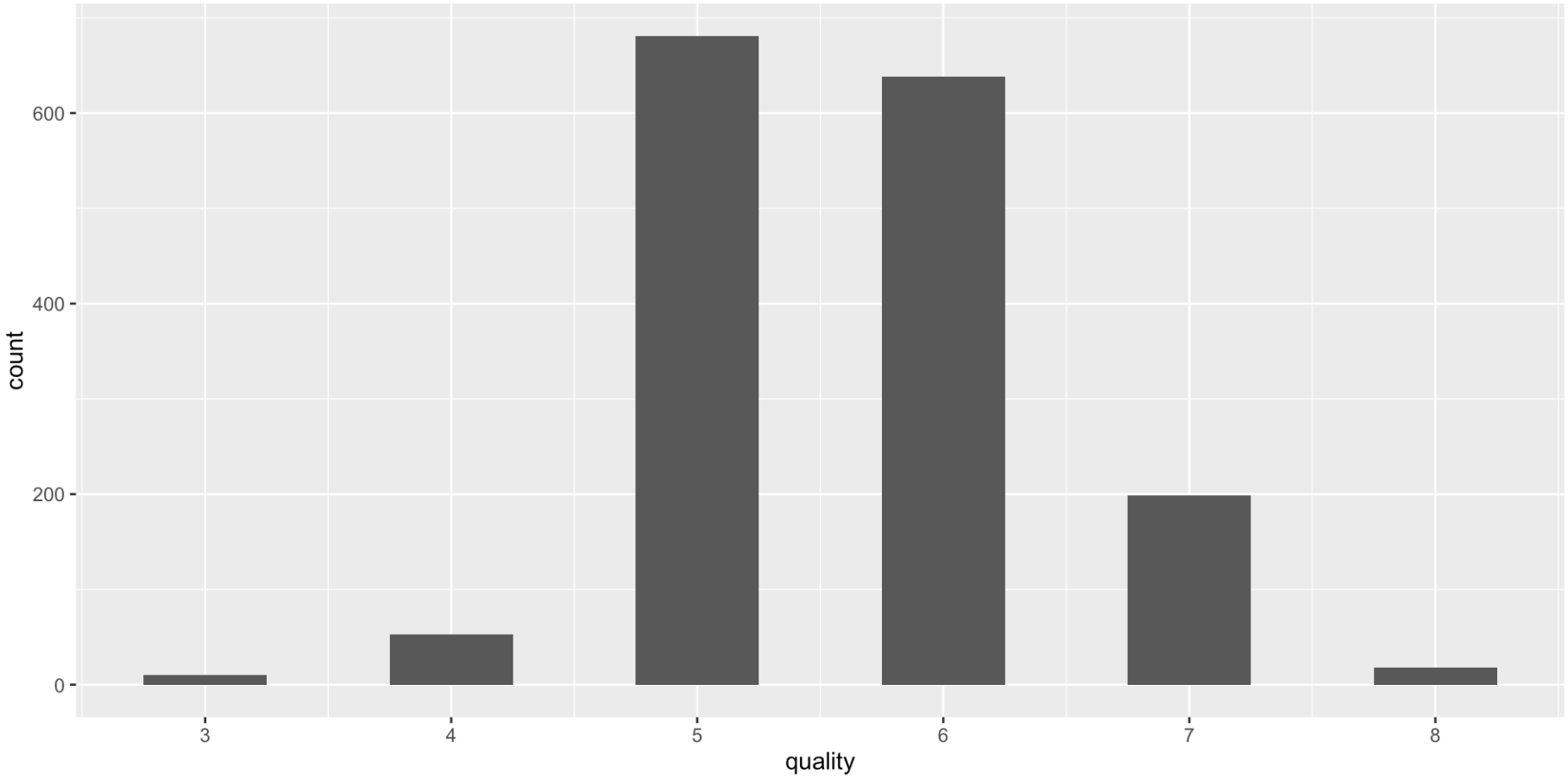





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```



Created a new variable (alcohol_content) with low, medium, high alcohol content. There is a strict value (x=9.5%) where alcohol content is optimally chosen for wine. The count gradually decreases as content increase and immediately drops as content decreases below 9.5%. The average alcohol content is 10.42%.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.636   6.000   8.000
```

The scale ranges from 0:10 with 0 being the worst quality and 10 being the best. There are no quality values for 0,1,2,9, and 10. We can also observe a normal distribution.

Univariate Analysis

What is the structure of your dataset?

1599 observations and 13 variables (x, fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfu.dioxide, total.sulfur.dioxide, density, pH, sulphates, alcohol, quality.)

Other observations: Quality range is 0:10 and mean is 5.6 Max pH is 4 mean and median alcohol is about 10%

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

The quality of the wine.

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

The alcohol content to determine the boldness, the pH to determine acidity/tartness, and residual sugar to determine sweetness. They will all affect the quality of the wine.

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

Yes, I created alcohol_content which consists of 3 buckets (0, 9.5], (9.5, 11.1], and (11.1, 15]. Then I renamed each bucket to Low, Medium, and High to differentiate the amount of alcohol catagorically.

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?

Most graphs where either normally distributed or skewed right. Citric acid was the only graph that was a little different since there was a sharp decrease from 0:0.25, then another surge at 0.50.

Bivariate Plots Section

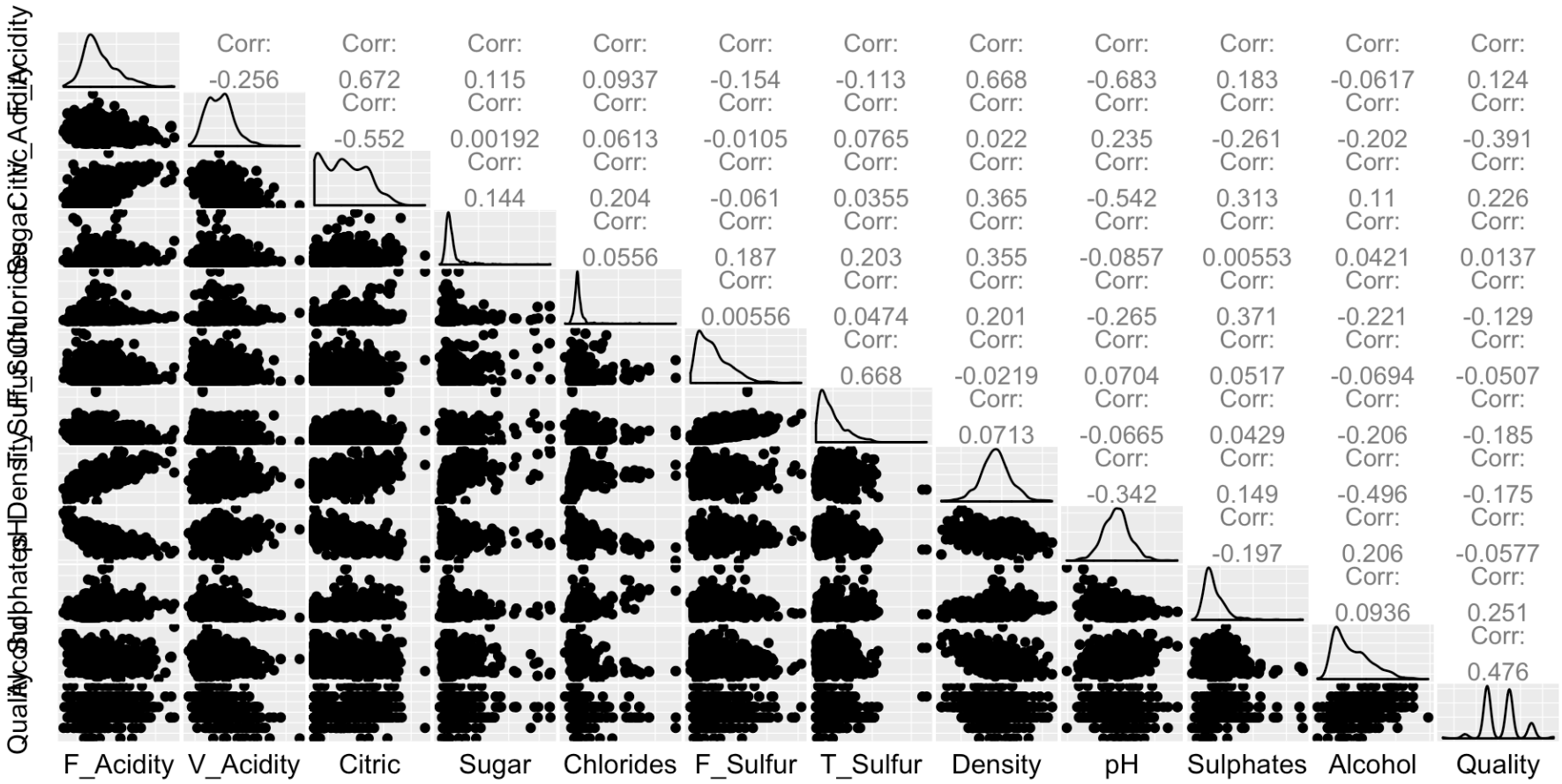
```
##                                X fixed.acidity volatile.acidity
## X                    1.000000000    0.00000000      0.724669575
## fixed.acidity       -0.268483920    1.00000000      0.000000000
## volatile.acidity    -0.008815099   -0.25613089      1.000000000
## citric.acid         -0.153551355    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
##                       citric.acid residual.sugar     chlorides
## X                    6.744481e-10   2.115297e-01  1.533390e-06
## fixed.acidity        0.000000e+00   4.199465e-06  1.751746e-04
## volatile.acidity     0.000000e+00   9.389168e-01  1.422491e-02
## citric.acid          1.000000e+00   8.083723e-09  2.220446e-16
## residual.sugar       1.435772e-01   1.000000e+00  2.617079e-02
## chlorides            2.038229e-01   5.560954e-02  1.000000e+00
## free.sulfur.dioxide -6.097813e-02   1.870490e-01  5.562147e-03
## total.sulfur.dioxide 3.553302e-02   2.030279e-01  4.740047e-02
## density              3.649472e-01   3.552834e-01  2.006323e-01
## pH                  -5.419041e-01  -8.565242e-02 -2.650261e-01
## sulphates            3.127700e-01   5.527121e-03  3.712605e-01
## alcohol              1.099032e-01   4.207544e-02 -2.211405e-01
## quality              2.263725e-01   1.373164e-02 -1.289066e-01
##                     free.sulfur.dioxide total.sulfur.dioxide
## X                          2.915917e-04         2.297726e-06
## fixed.acidity              6.335579e-10         5.709033e-06
## volatile.acidity           6.747011e-01         2.213857e-03
## citric.acid                1.473916e-02         1.555454e-01
## residual.sugar             4.685141e-14         2.220446e-16
## chlorides                  8.241238e-01         5.809120e-02
## free.sulfur.dioxide        1.000000e+00         0.000000e+00
## total.sulfur.dioxide       6.676665e-01         1.000000e+00
## density                   -2.194583e-02         7.126948e-02
## pH                         7.037750e-02        -6.649456e-02
## sulphates                  5.165757e-02         4.294684e-02
## alcohol                   -6.940835e-02        -2.056539e-01
## quality                   -5.065606e-02        -1.851003e-01
##                           density            pH    sulphates      alcohol
## X                    0.000000e+00  4.770847e-08 4.992031e-07 0.000000e+00
## fixed.acidity        0.000000e+00  0.000000e+00 1.648681e-13 1.364868e-02
## volatile.acidity     3.787554e-01  0.000000e+00 0.000000e+00 3.330669e-16
## citric.acid          0.000000e+00  0.000000e+00 0.000000e+00 1.059462e-05
## residual.sugar       0.000000e+00  6.065915e-04 8.252134e-01 9.258425e-02
## chlorides            5.551115e-16  0.000000e+00 0.000000e+00 0.000000e+00
## free.sulfur.dioxide  3.804985e-01  4.869975e-03 3.888321e-02 5.492314e-03
## total.sulfur.dioxide 4.354284e-03  7.818341e-03 8.601835e-02 1.110223e-16
## density              1.000000e+00  0.000000e+00 2.418474e-09 0.000000e+00
## pH                  -3.416993e-01  1.000000e+00 2.109424e-15 1.110223e-16
## sulphates            1.485064e-01 -1.966476e-01 1.000000e+00 1.783053e-04
## alcohol             -4.961798e-01  2.056325e-01 9.359475e-02 1.000000e+00
## quality             -1.749192e-01 -5.773139e-02 2.513971e-01 4.761663e-01
##                          quality
## X                   7.857465e-03
```

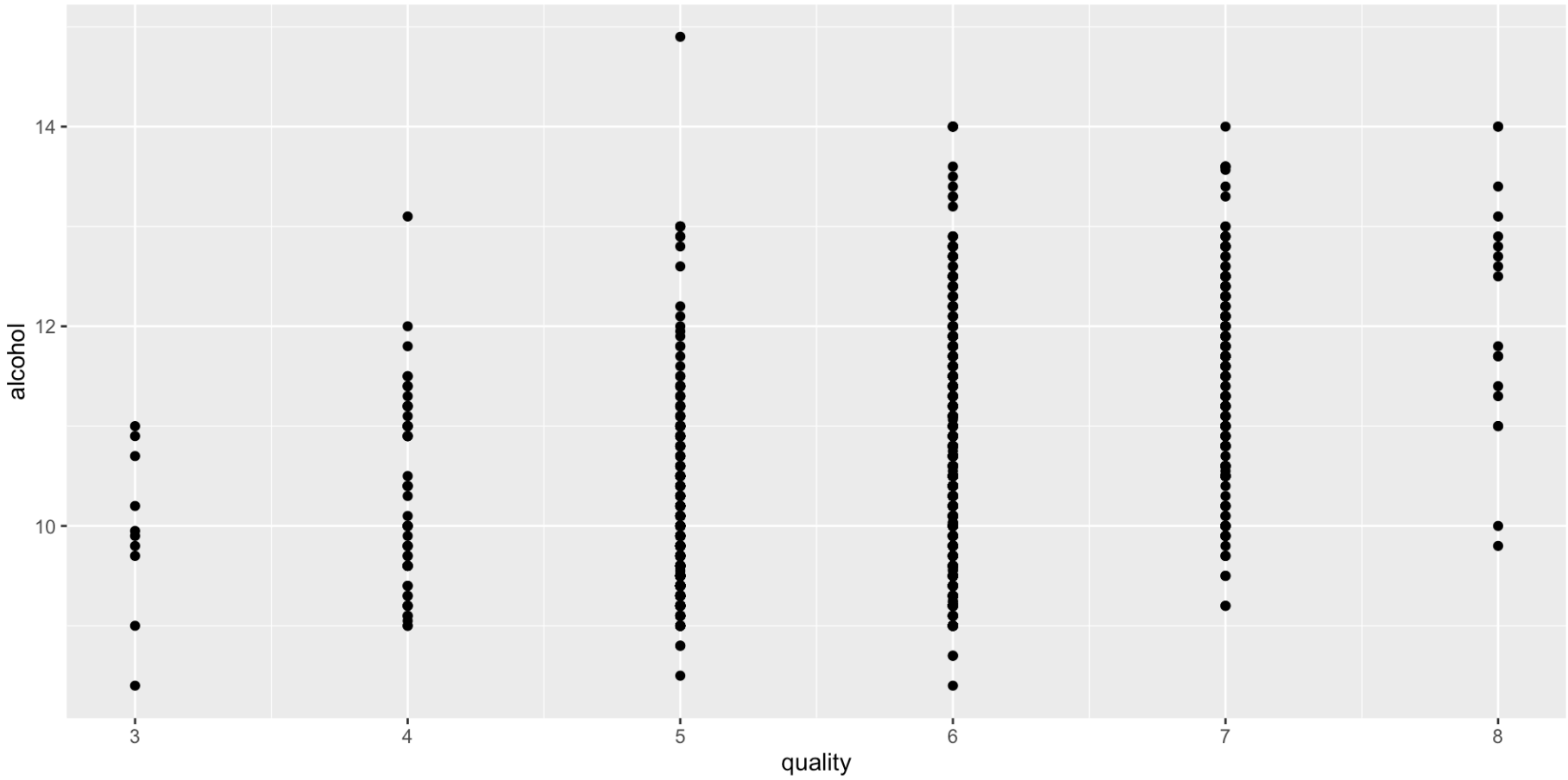
```
## fixed.acidity         6.495635e-07
## volatile.acidity      0.000000e+00
## citric.acid           0.000000e+00
## residual.sugar        5.832180e-01
## chlorides             2.313383e-07
## free.sulfur.dioxide   4.283398e-02
## total.sulfur.dioxide 8.615331e-14
## density               1.874945e-12
## pH                    2.096278e-02
## sulphates             0.000000e+00
## alcohol               0.000000e+00
## quality               1.000000e+00
```

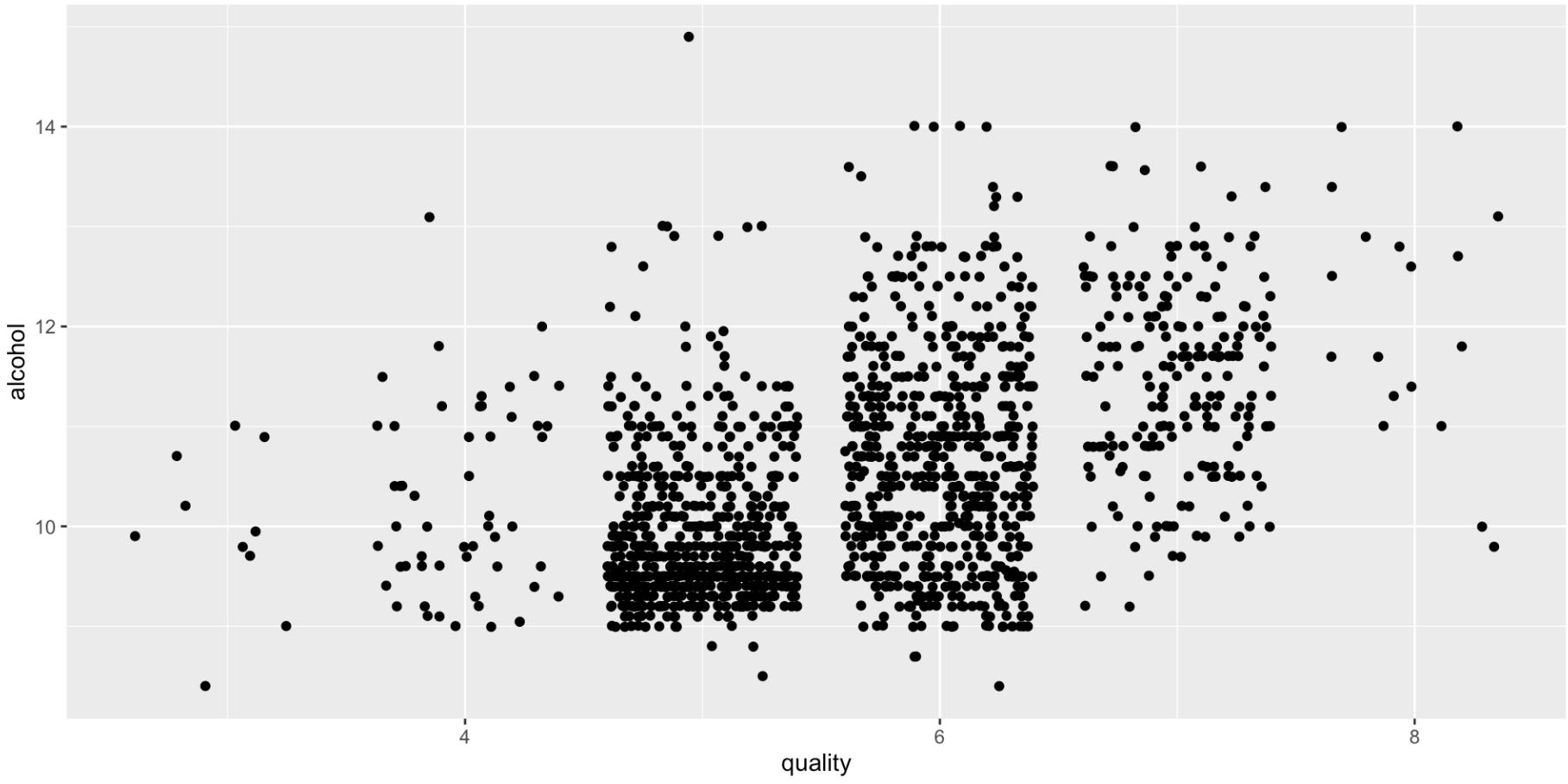
https://stat.ethz.ch/pipermail/r-help/2001-November/016201.html (https://stat.ethz.ch/pipermail/r-help/2001-November/016201.html)

I found the cor.prob function online (listed above) that allows me to compare the correlation for each variable combination in the dataset. The best correlation for quality is residual sugar. I also imported the data as data_original to exclude the alcohol_content column, since I am not able to take the correlation of non integers.

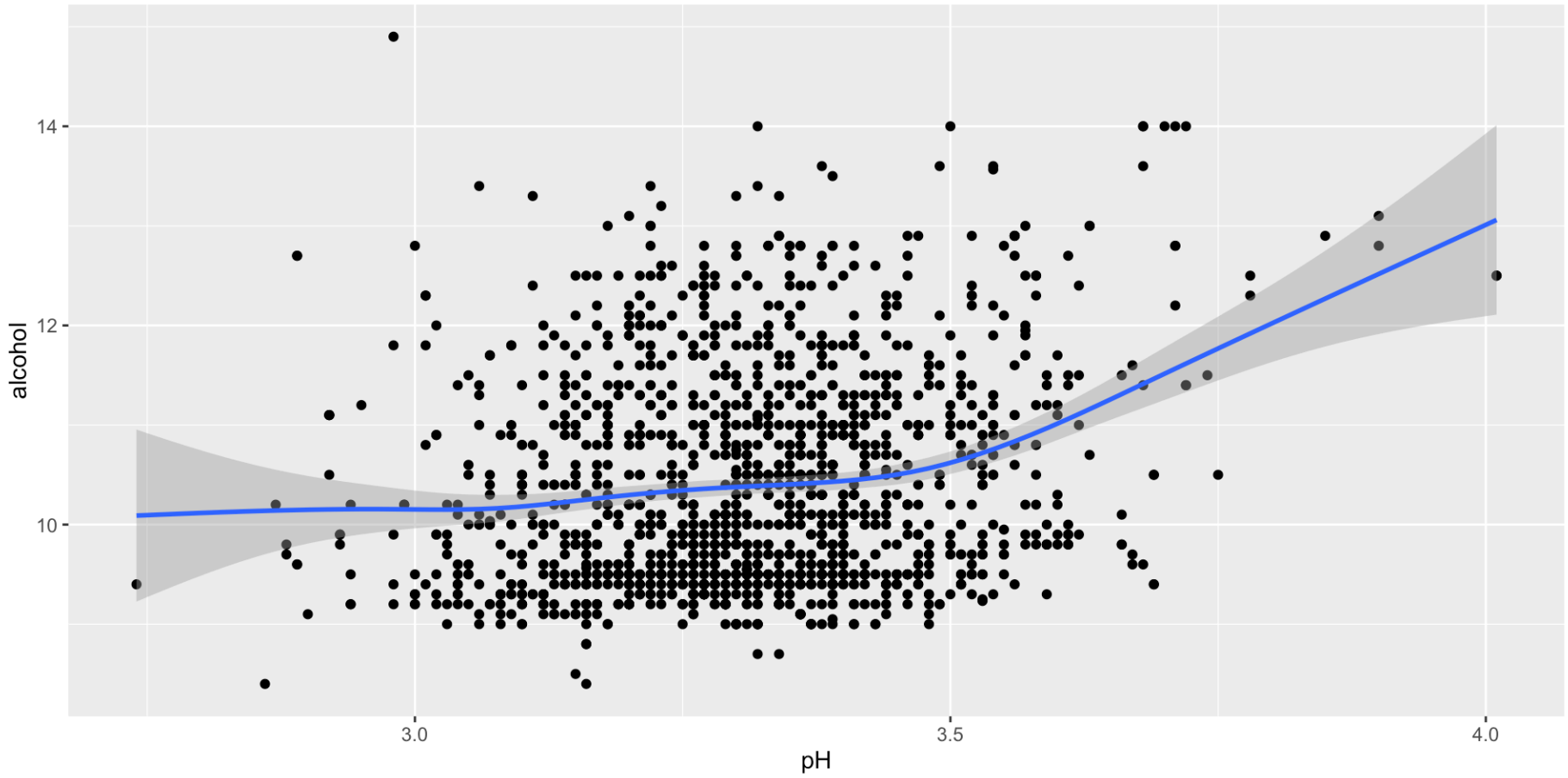


F_Acidity(fixed.acidity), V_Acidity(volatile.acidity), F_Sulfur(free.sulfur.dioxide), T_Sulfur(total.sulfur.dioxide)



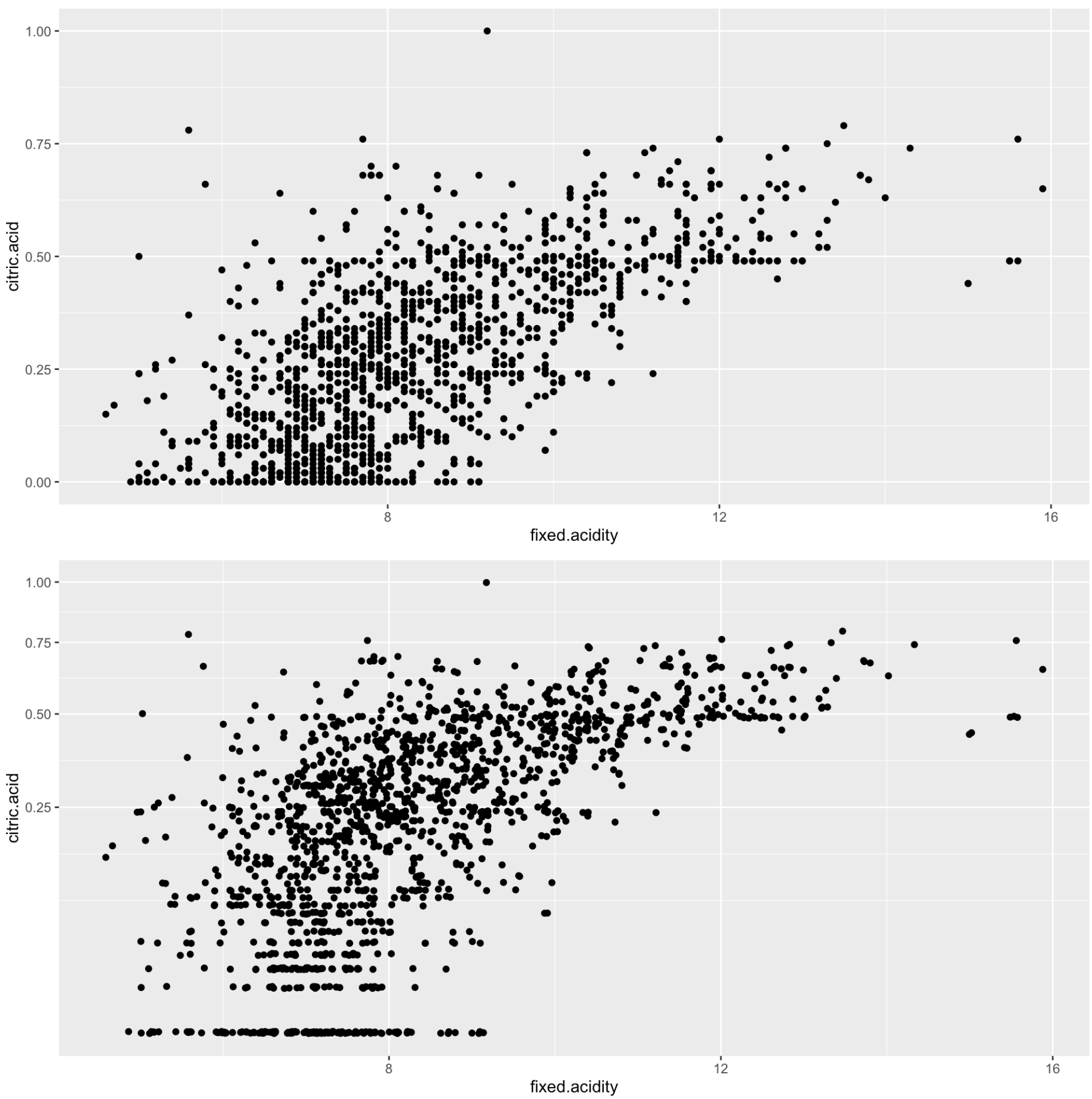


Wanted to look closer at the relationship of alcohol content and quality, but it’s difficult to draw any conclusions besides, low quality generally has low alcohol content. This might be becuase there was no high alcohol, low quality wine sampled. Majority of the quality is 5, 6, or 7.

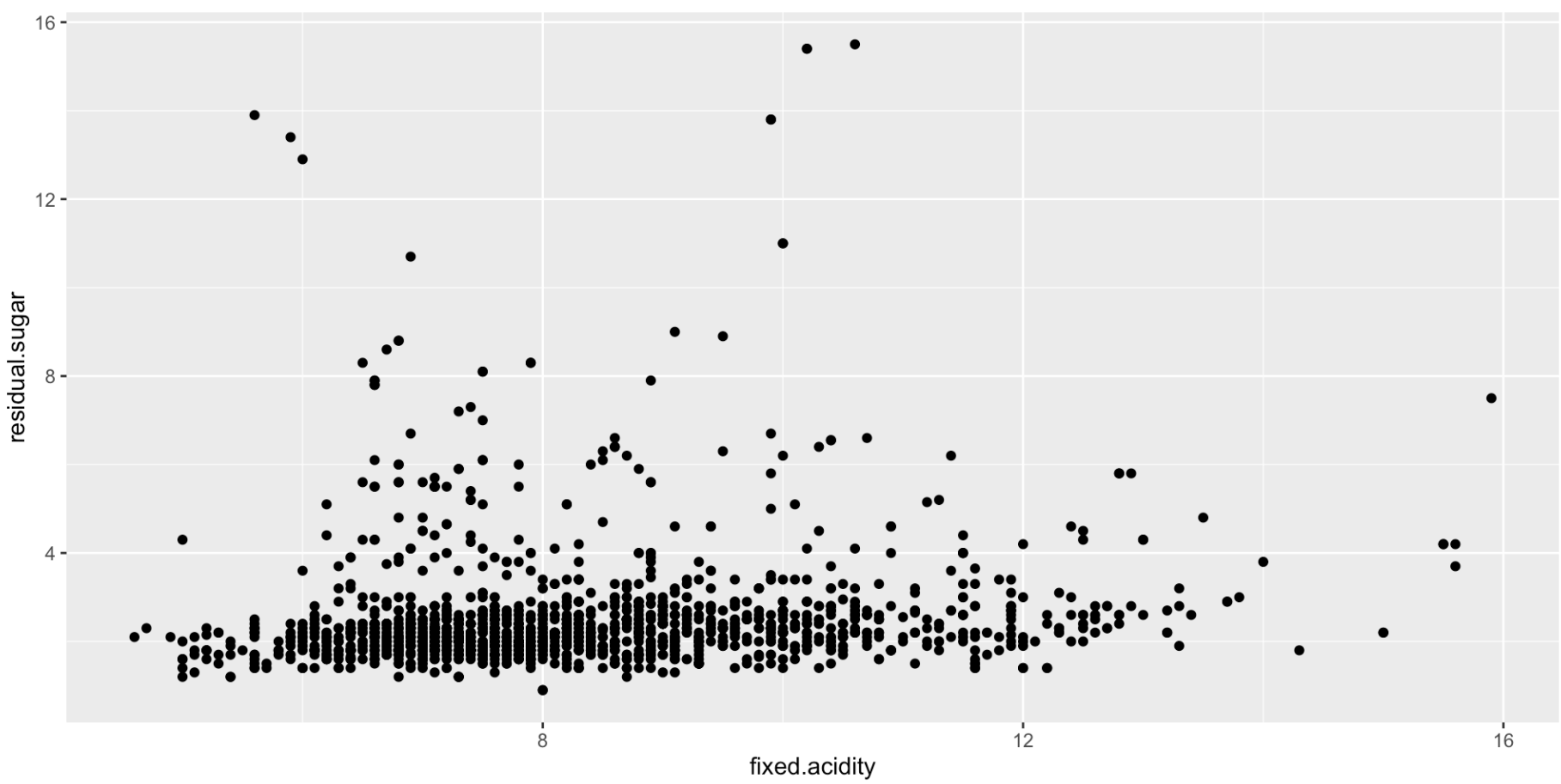



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010
```

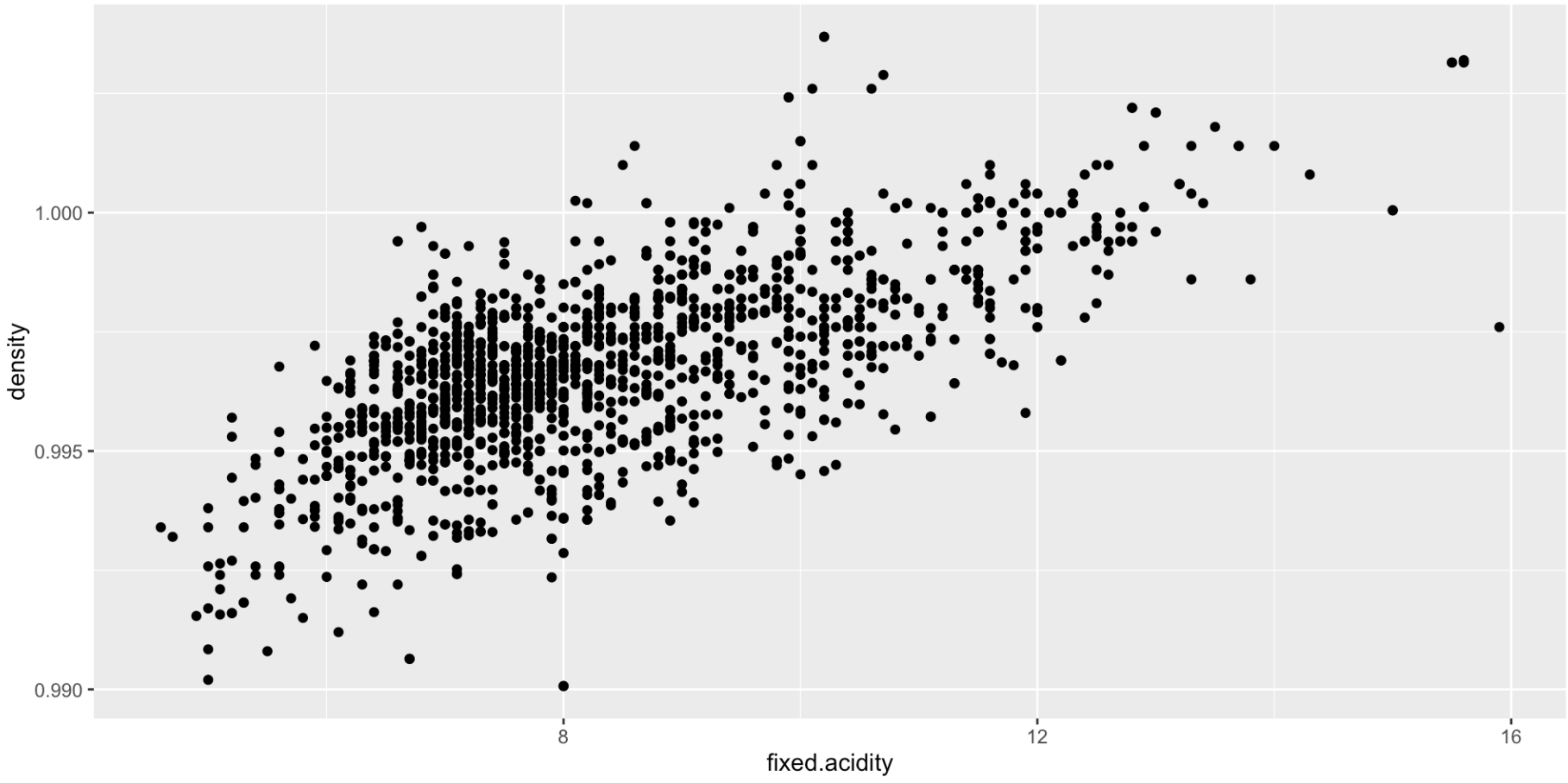
We can see some horizontal ‘lines’ which suggest the alcohol content is staying the same but there is a range of pH between 3-3.5. The mean pH is 3.311.



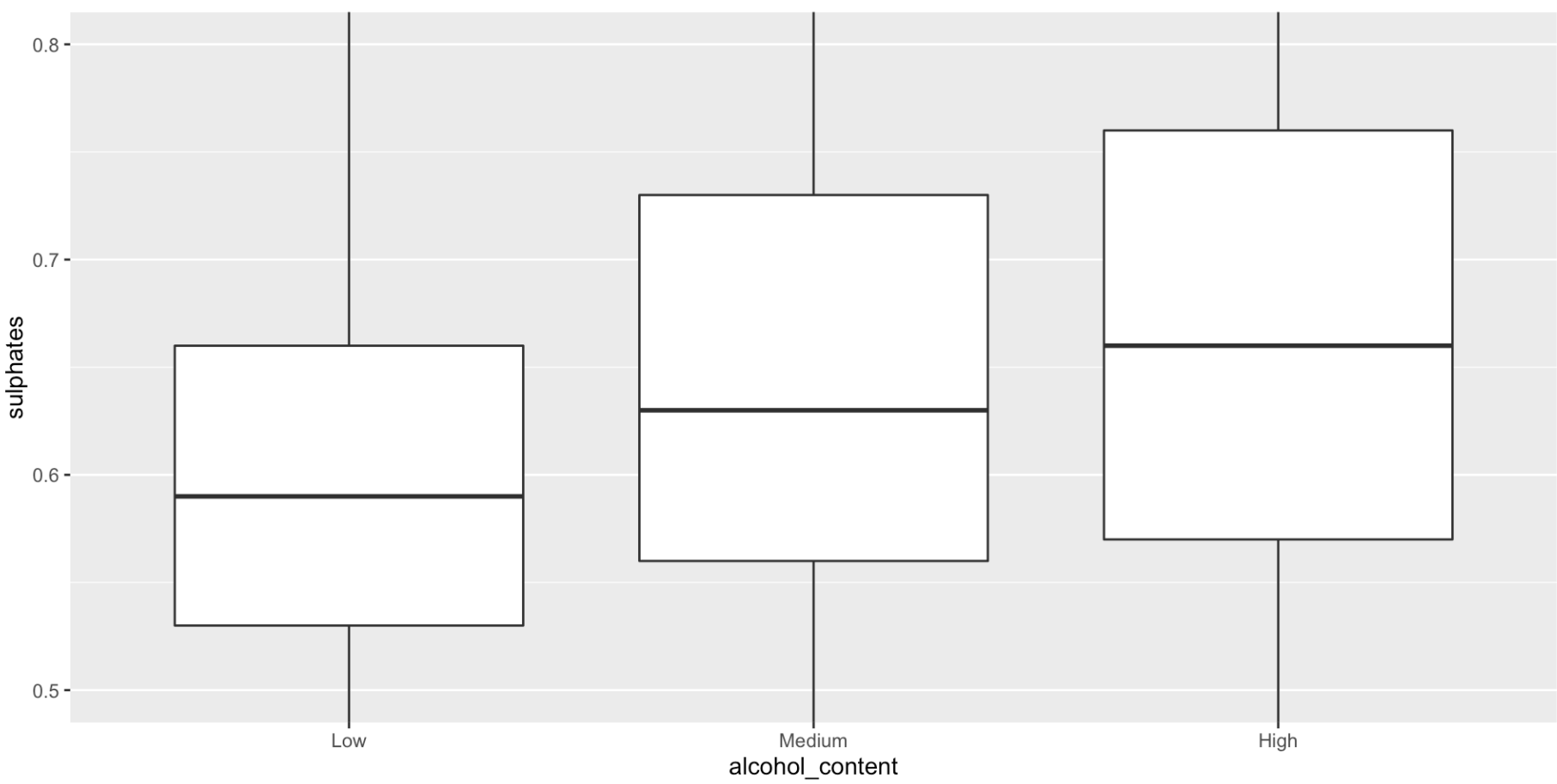
There is somewhat of a direct relationship shown in the first graph. After the transformation of the y-scale by sqrt(), it shows a pool of data points for y = 0. There are many y-value horizontal lines as well leading up to y = 0.25.



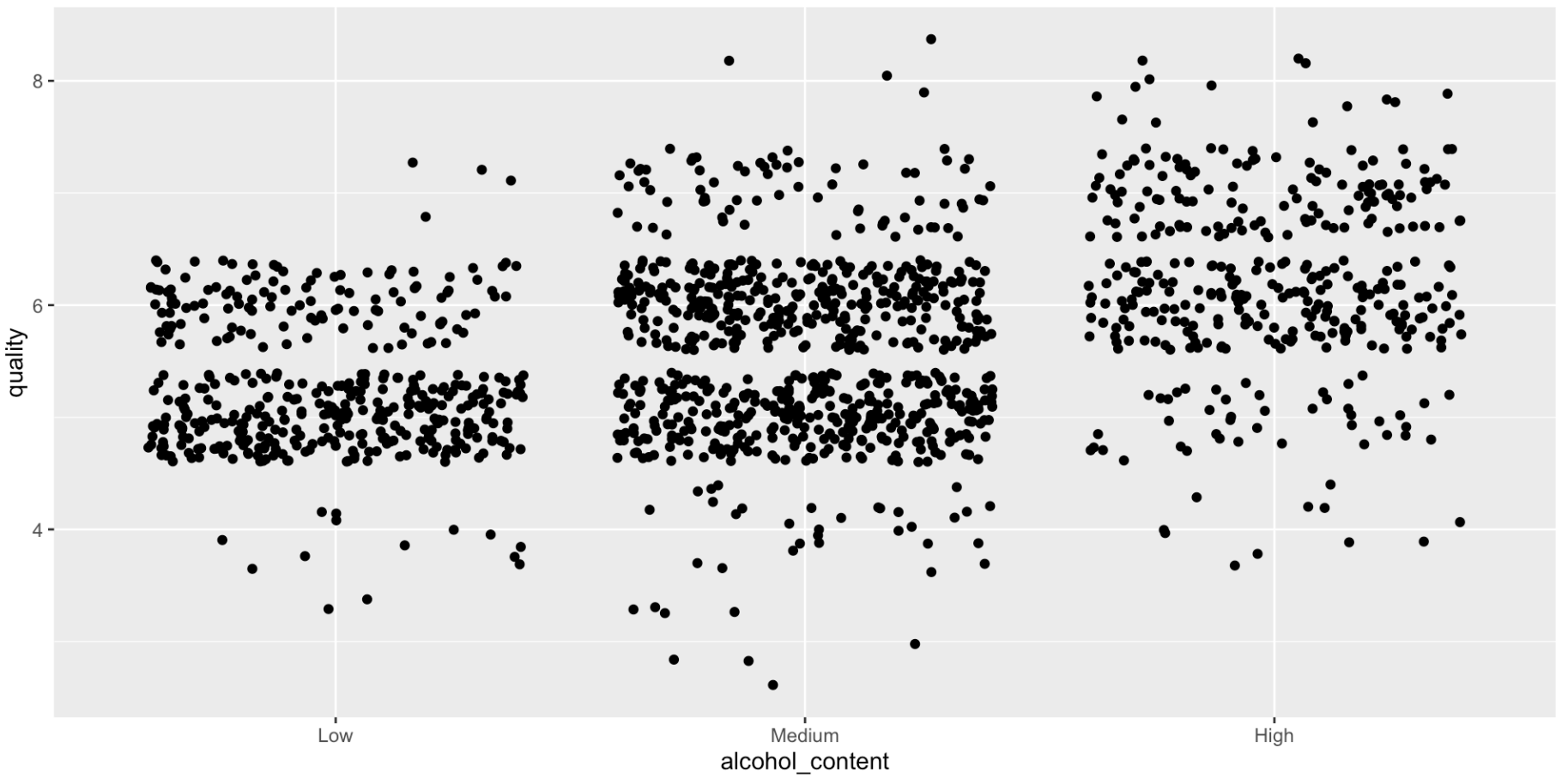
There are some very high values for residual sugar, 12-16 range, (likely more of the sweeter wines). No strict relationship can be determined.



There is a clear direct relationship.



Higher alcohol content wines contain greater average sulphates.



Most quantity of alcohol type are as follows: (alcohol_content, quality) –> (Low, 5), (Medium, 5 & 6), (High, 6 & 7).

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?

Refering to alcohol vs quality graph: Higher quality (8) red wine has higher alcohol content while low quality (3 and 4) has lower content. Quality of 5, 6, and 7 have a bigger range but you can see a general shift. It's hard to determine if there is a relationship though, since there isn't enough data points for low and high quality wines compared to average quality.

Refering to alcohol vs pH: We can see multiple horizontal "lines" which suggests varying pH for the same alcohol content. This is probably becuase companies choose similar alcohol amounts but vary in tartness and crispness.

Refering to residual sugar vs fixed acidity: We can see pooling of data points below sugar of 4. There is no real relationship here since sugar ranges the same as acidity increases. The outlier data ranging from y = 12-16 is interesting but does not show any relationship to acidity.

Referring to quality vs alcohol_content: With the amount of data we have, there is a relationship between alcohol content and quality. Generally the higher content is of better quality since there are more data points. This might change if we introduce further data.

As pH decreases fixed acidity increases, obviously this makes sense since that is the definition of acidic solution.

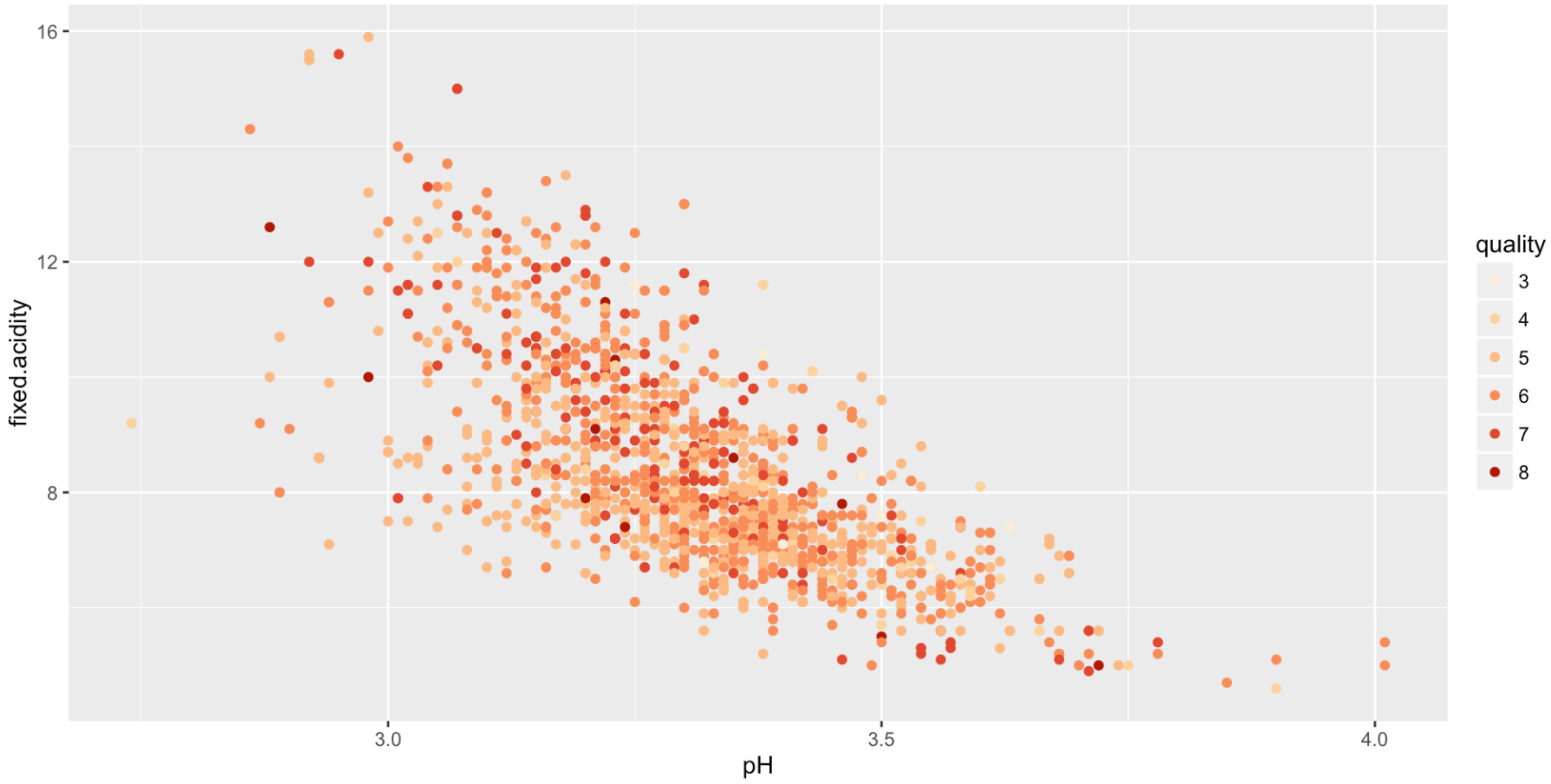
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

I compared fixed acidity vs density and found a strong direct relationship. Quality is also affected by density and acidity as higher quality wines are more acidic and have less density.

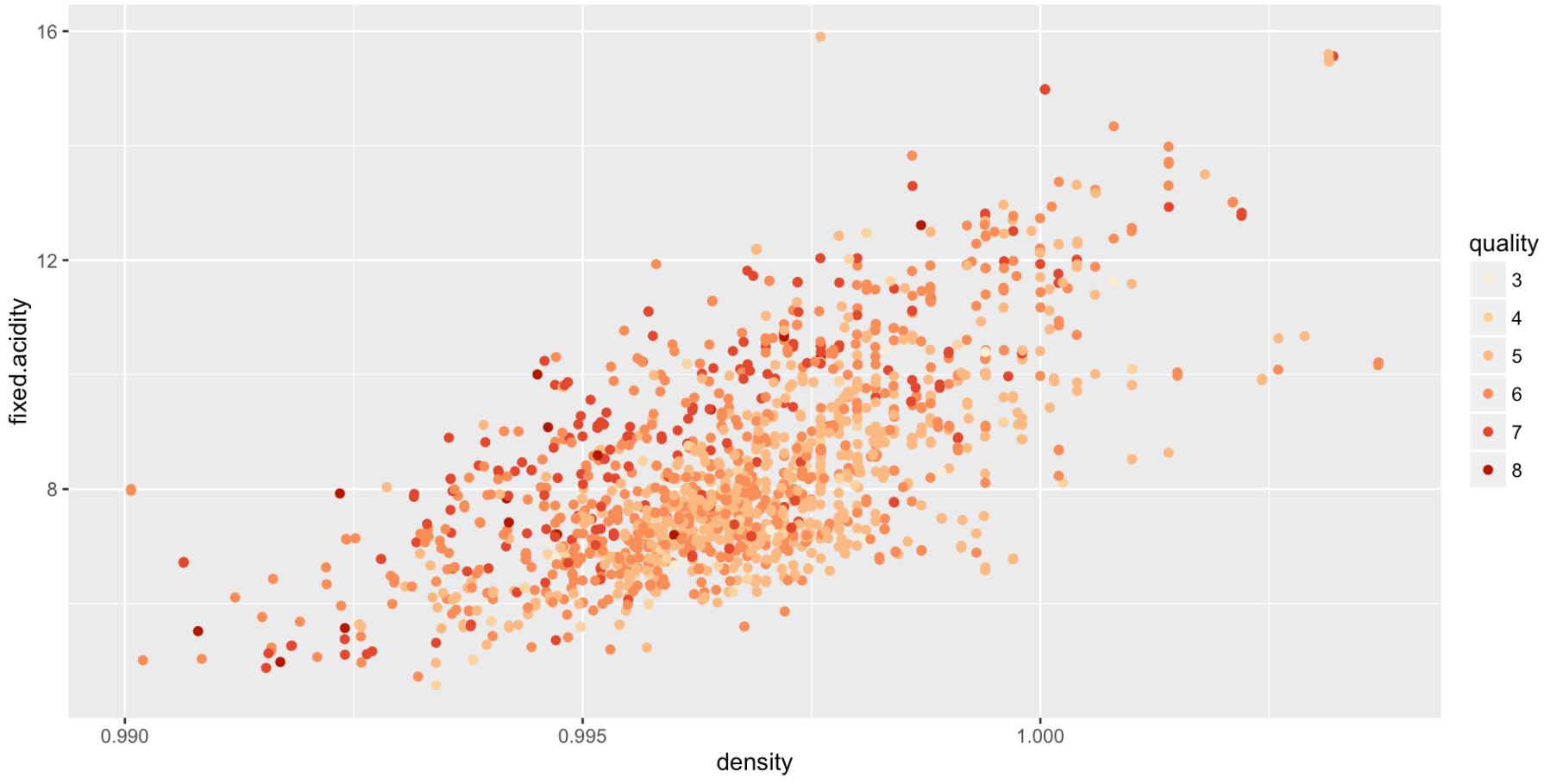
What was the strongest relationship you found?

The fixed acidity vs density graph.

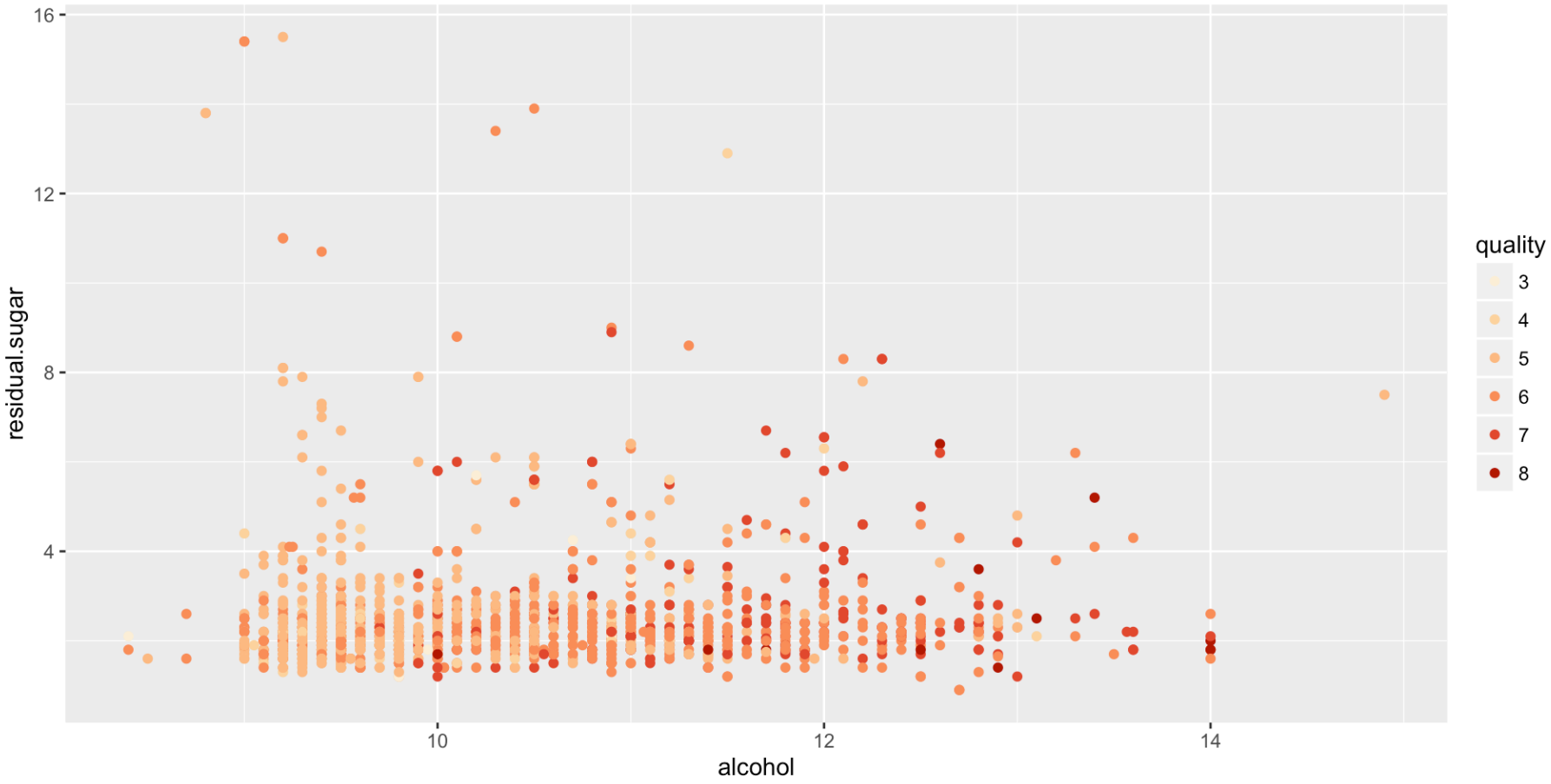
Multivariate Plots Section



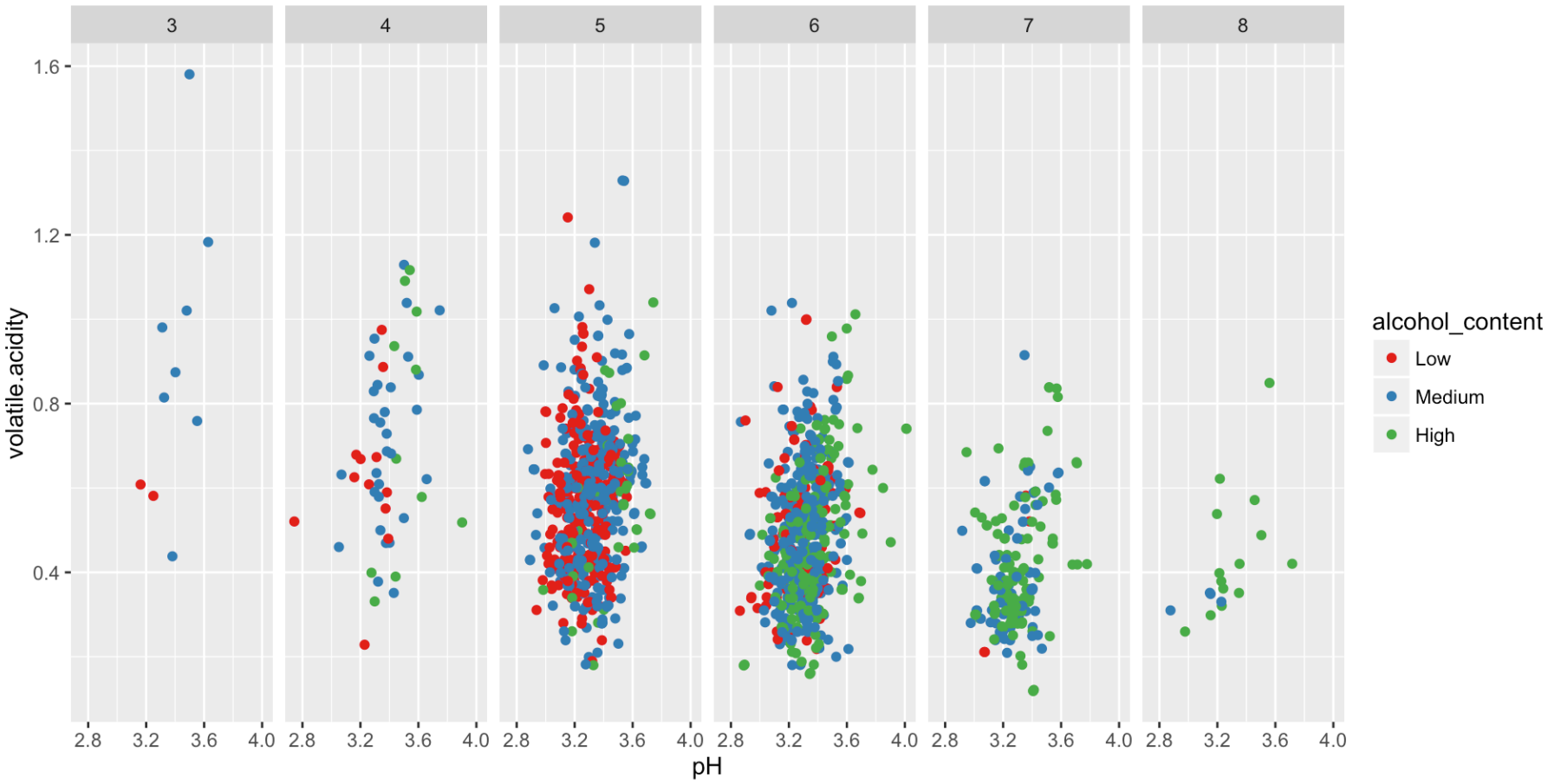
This graph makes sense since higher acidic wines will naturally have a lower pH and vice versa. Coloring the graph by quality does not show any strong relationships.



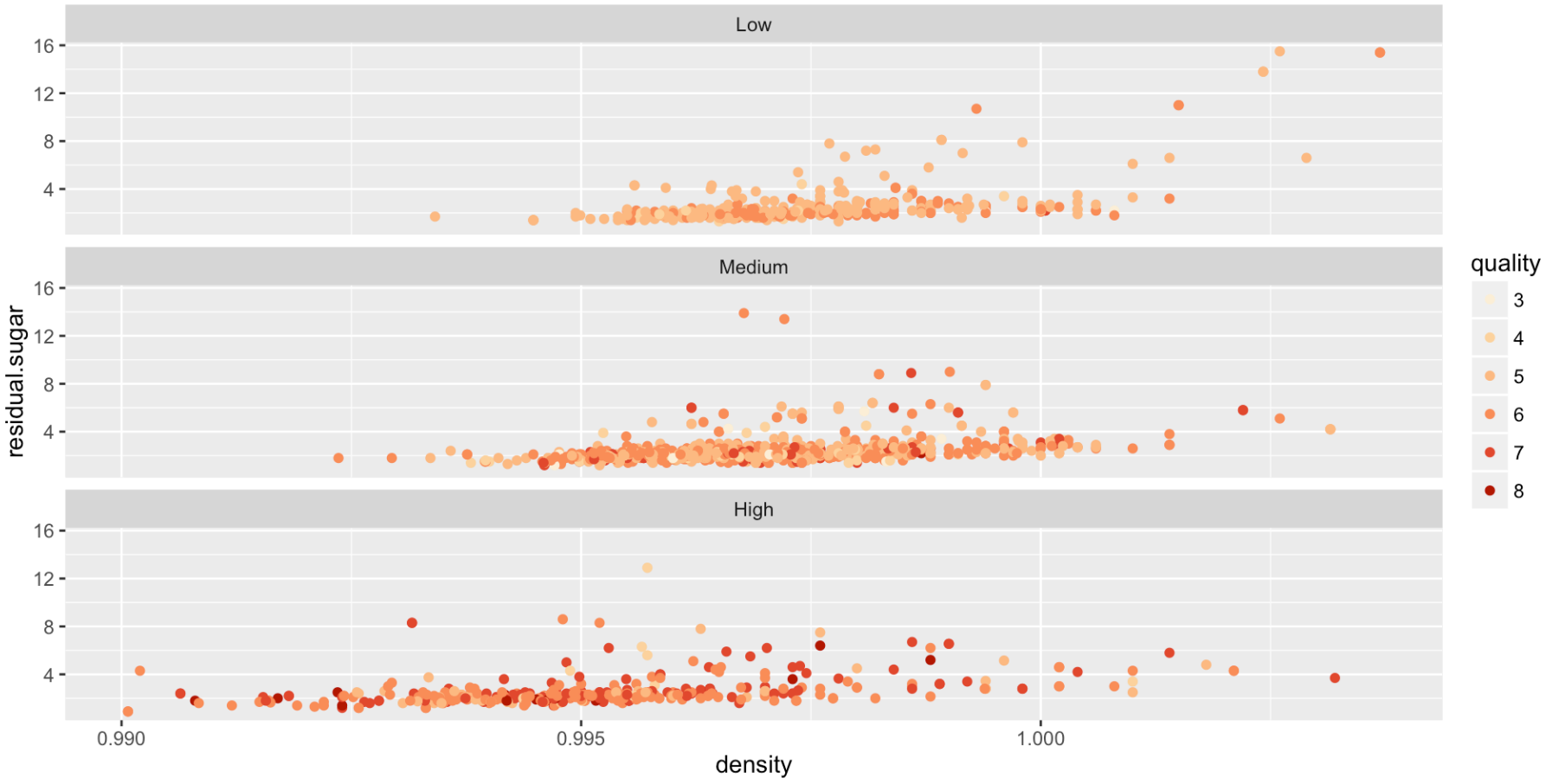
After reviewing the correlation graph I decided to look into density and acidity. There seems to be a strong direct relationship where density increases as acidity increases. Adding quality by color shows lower quality (4 and 5) are high density and low acidity compared to (6, 7, 8).



Comparing the residual sugar and alcohol shows most of the red wine is below 4 which means they would be off-dry wines and range from 9-13% in alcohol content. When differentiating by quality it can be seen that higher quality wines have higher alcohol content (>10%) and generally less sweet.



Each graph is colored by alcohol_content and seperated by quality. We see medium and high alcohol content more prominent in higher quality (7, 8). There is also a small shift down on the y-axis (decrease in volatile acidity) as the quality increases.



Density of wine is closer to water (1.000) when alcohol content is lower and residual sugar ranges from 2 to 4.

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?

There is a relationship between the amount of alcohol and the quality of the wine. Generally higher quality wines will be of higher alcohol content. We can also see a shift towards less volatile acidity. This makes sense since higher levels of acetic acid will cause an unpleasant taste.

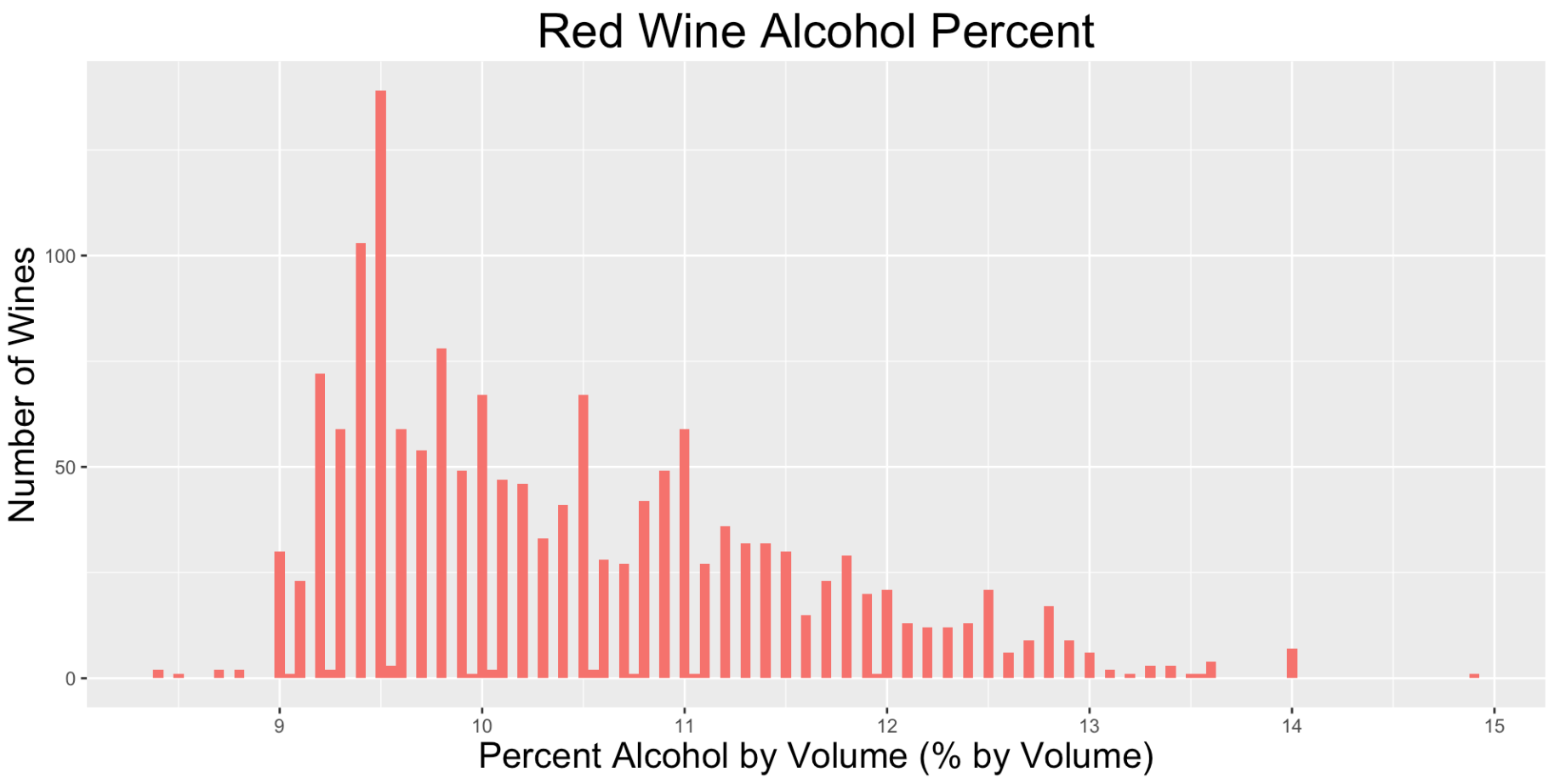
Were there any interesting or surprising interactions between features?

I did not find anything too surprising but after facet wrapping the graph by quality it was easier to see the differences in quality and alcohol content. I was surprised to see how much they related since I was not able to see the relationship clearly in my previous graphs.

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

Final Plots and Summary

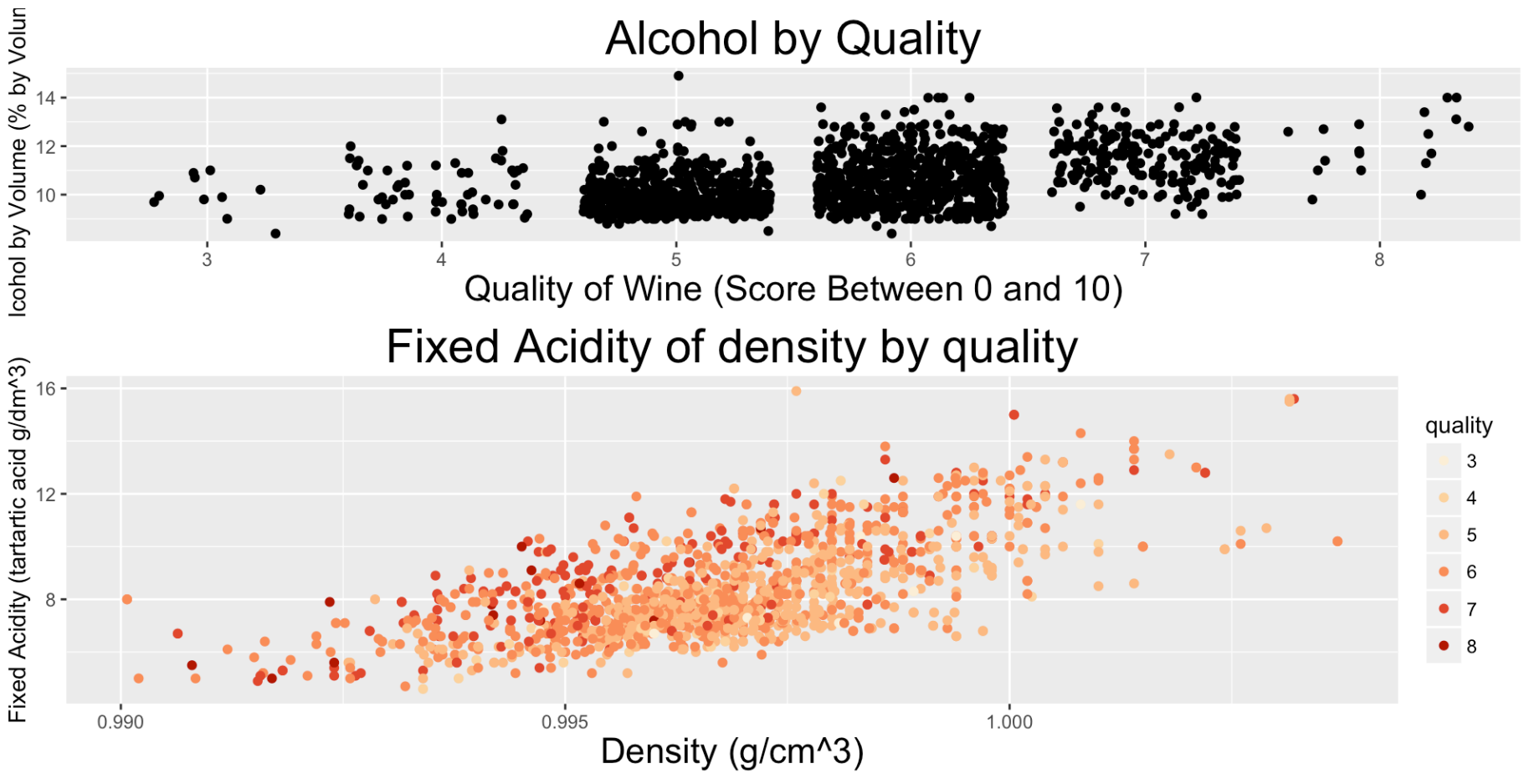
Plot One



Description One

This graph shows the greatest alcohol % to be around 9.5% and we see a gradual decrease as the % increases or decreases. This is probably becuase that's the optimal alcohol % that is time efficient (as in the amount of time needed for fermentation) for wholesale, or the amount most consumers prefer.

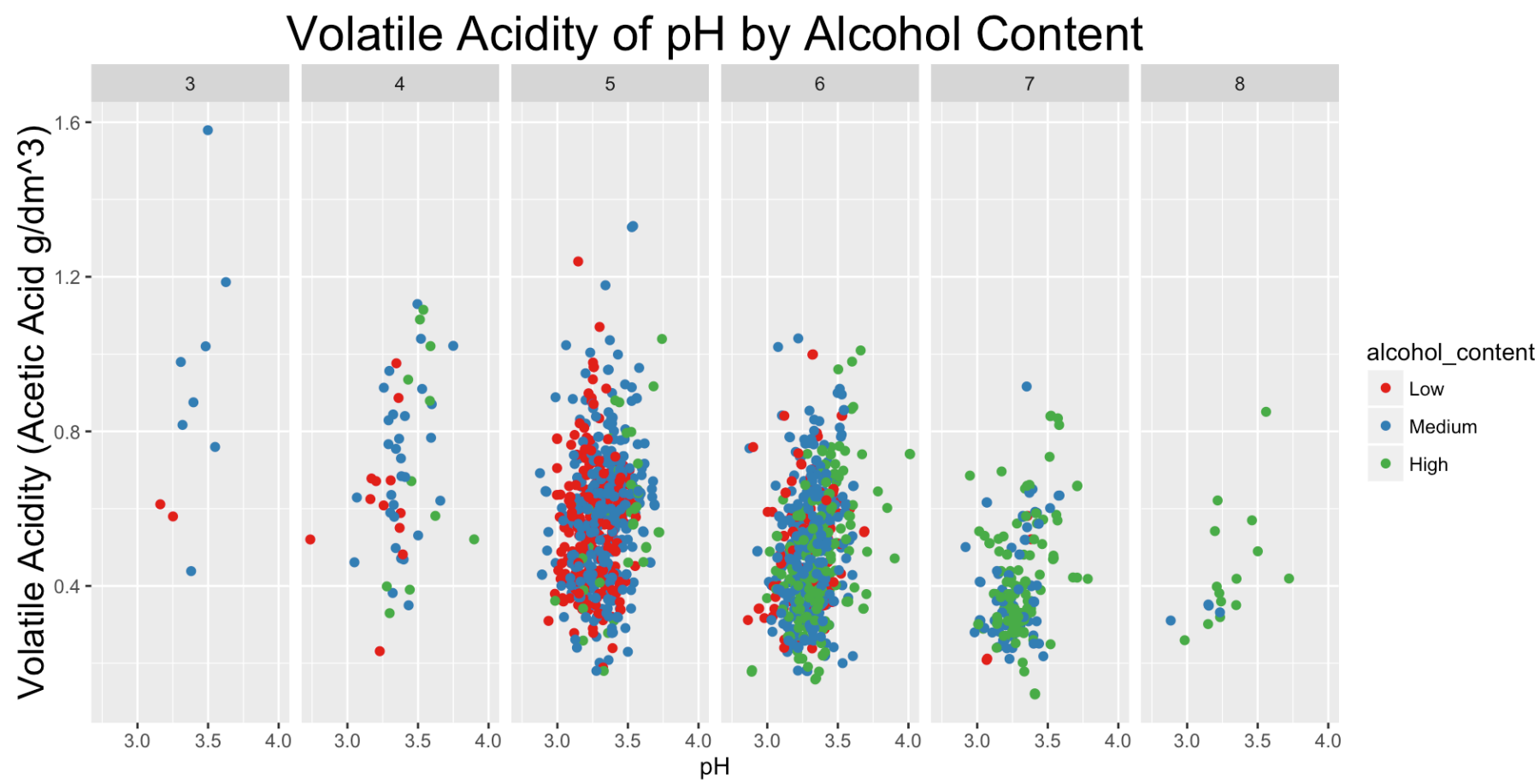
Plot Two



Description Two

The top graph compares the alcohol % by quality of wine to better show which quality has the most data points (5,6,7). The bottom graph shows how acidity increases as density increases. It is also able to determine loosely, how the higher quality is more acidic.

Plot Three



Description Three

This plot shows how the alcohol content changes with quality, mostly high content wines are promenent in quality 8 while the opposite is true for quality 3. There is also a small down shift towards volatile acidity as quality increases, I suspect to lower the vinger taste.

Reflection

This dataset contains various chemicals commonly found in wine and also the quality rating for each wine. Starting off, I knew I wanted to answer the question “what chemicals affect the quality of the wine most?”. Through my intial analysis I found that acidity and alcohol % had pretty good distributions. Through further analysis I found that higher quality wines have a greater alcohol percent. Now, that does not mean higher alcohol percent will necesarily be of higher quality. Examining the coorelation plot shows a value of 0.476 for quality and alcohol which is the highest compared to all other quality paired correlations. I also noticed how acidiy was affected by density and that was a surprise since I did not think density had any strong relationships. I was able to determine that higher quality wines seem to show affinity for lower density and higher acidity, which makes sense since the alcohol % is increasing and moving further away from the density

of water (1). These are most of the relationships that I have observered to influence quality but, quality is very subjective. I feel that a quality of 8 may be a quality of 3 to me or 10 to someone else, since it depends on the subtleties of wine that are not always picked up by everyone.

For future work I think it would be interesting to compare the quality of wine to the price. I would graph the quality on y-axis and price on x-axis to show the general relationship. Then, I would create a seperate graph for each quality and see how the price differs and try to find any correlation to other variables (acidity, sugar, density) that may also affect the price for each quality.