Sofia Godovykh

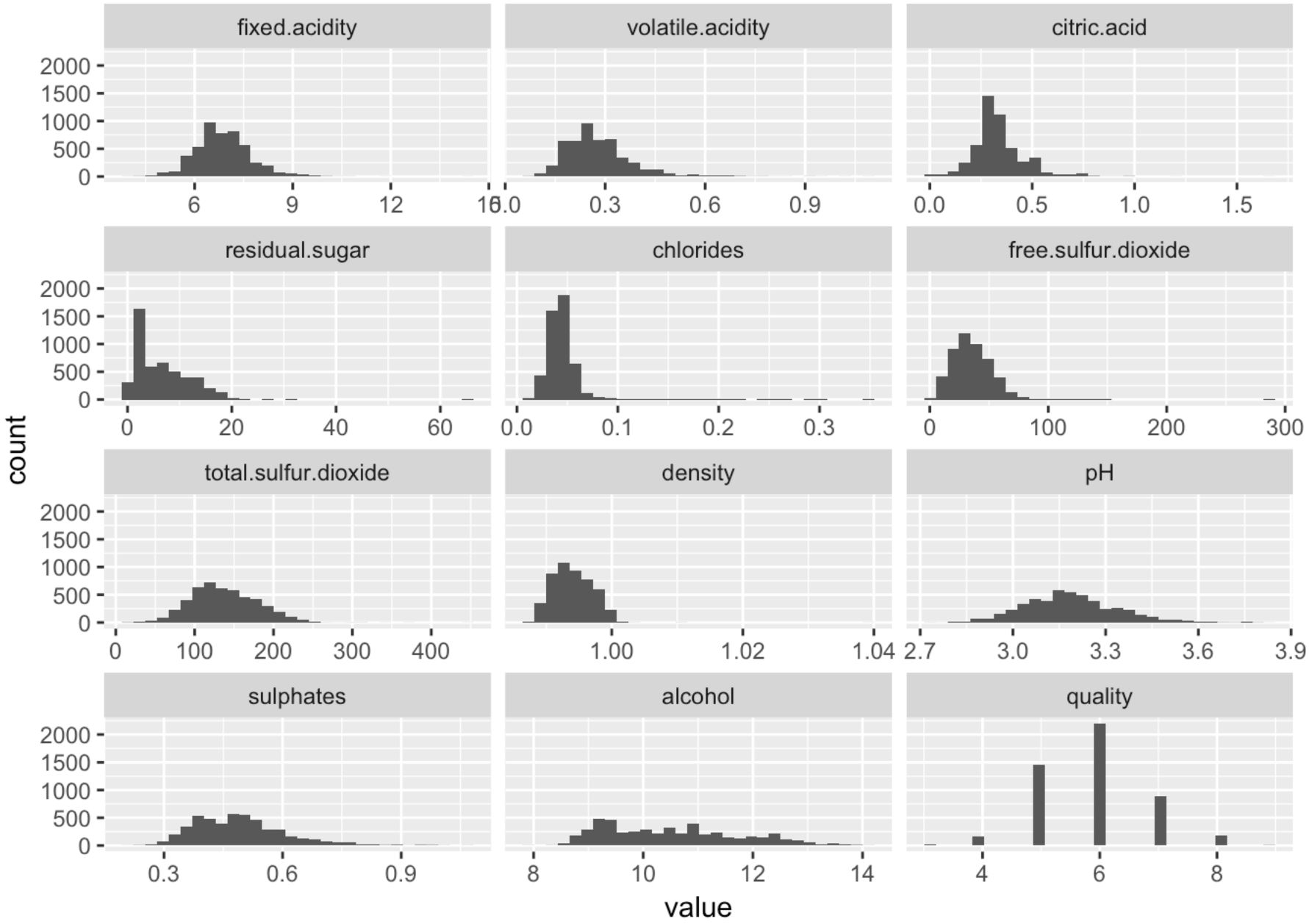
The white wine dataset contains information of acidity, sugar, pH level, and other chemical parameters. Every wine is graded by critics according to its quality. I want to find factors which contribute to wine quality and, specifically, find a difference in parameters between good and poor wines.

Univariate Plots Section

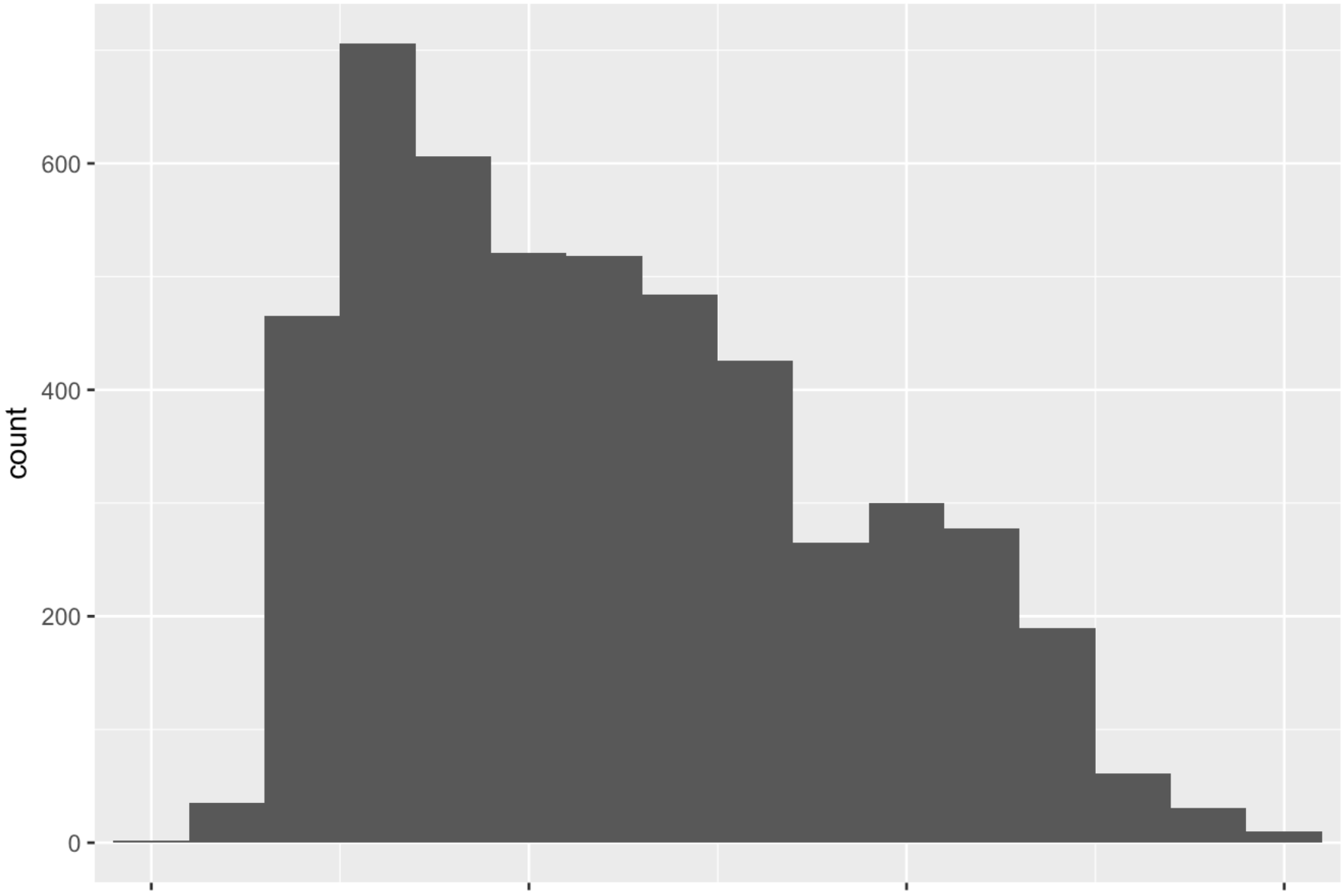
There is a general information about the dataset

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
##  Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
##  Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391
##  3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900
##  Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.00900   Min.   :  2.00      Min.   :  9.0
##  1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0
##  Median :0.04300   Median : 34.00      Median :134.0
##  Mean   :0.04577   Mean   : 35.31      Mean   :138.4
##  3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0
##  Max.   :0.34600   Max.   :289.00      Max.   :440.0
##     density             pH          sulphates         alcohol
##  Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00
##  1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50
##  Median :0.9937   Median :3.180   Median :0.4700   Median :10.40
##  Mean   :0.9940   Mean   :3.188   Mean   :0.4898   Mean   :10.51
##  3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40
##  Max.   :1.0390   Max.   :3.820   Max.   :1.0800   Max.   :14.20
##     quality      quality_factor
##  Min.   :3.000   Poor   :1640
##  1st Qu.:5.000   Average:2198
##  Median :6.000   Good   :1060
##  Mean   :5.878
##  3rd Qu.:6.000
##  Max.   :9.000
```

There are histograms of all parameters in a dateset.

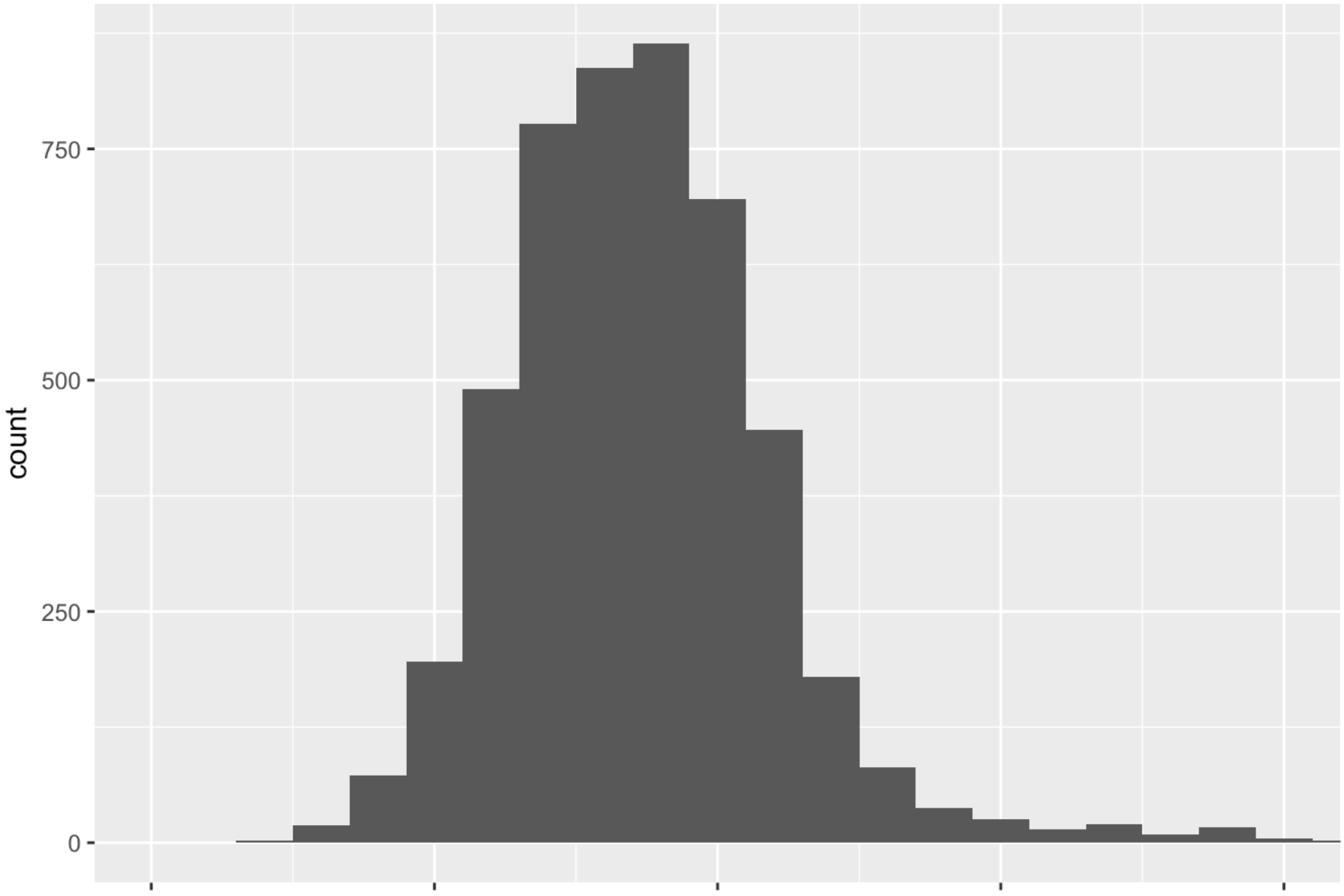


I will explore every feature in detail.



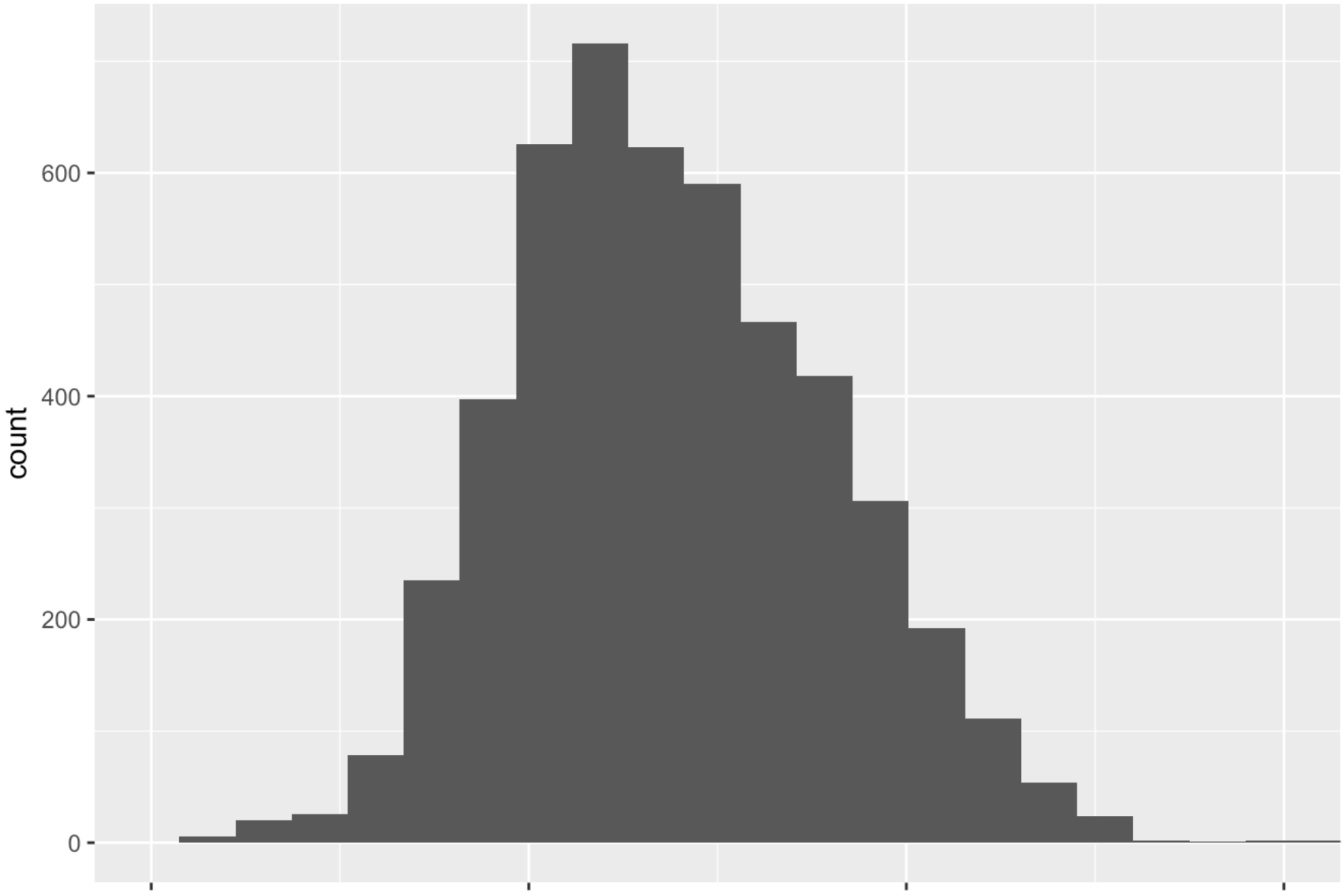

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.50   10.40   10.51   11.40   14.20
```

The distribution of alcohol shows that there are two peaks, a major one around 9% and smaller one around 12%.



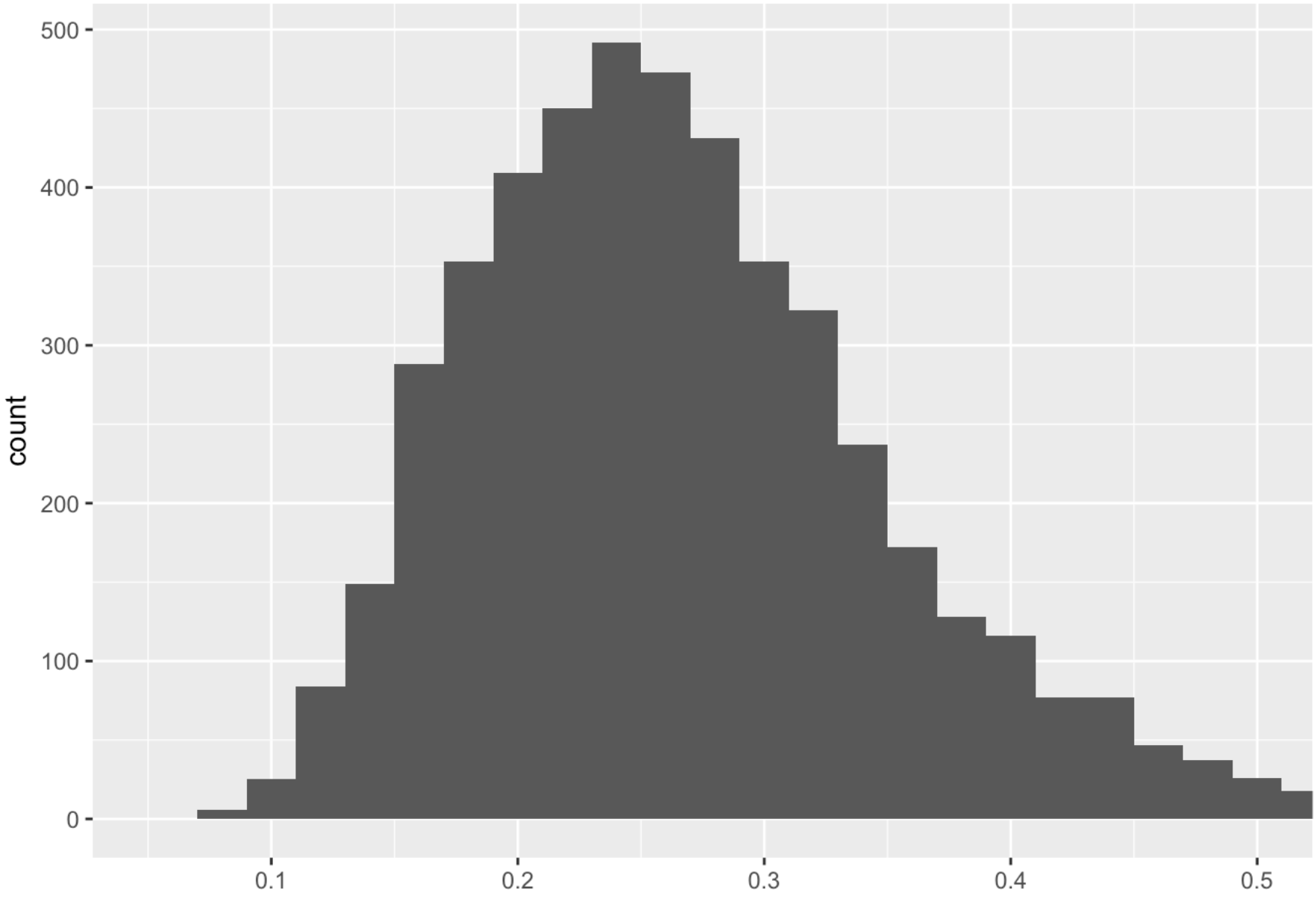

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00900 0.03600 0.04300 0.04577 0.05000 0.34600
```

Chlorides form a bell-curved distribution with long right tail. It would be interesting to know if different wine qualities distributed normally and which wines form the tail. It would be interesting to have a look at the outliers, because maximum value is way higher than median value and 3rd quantile.



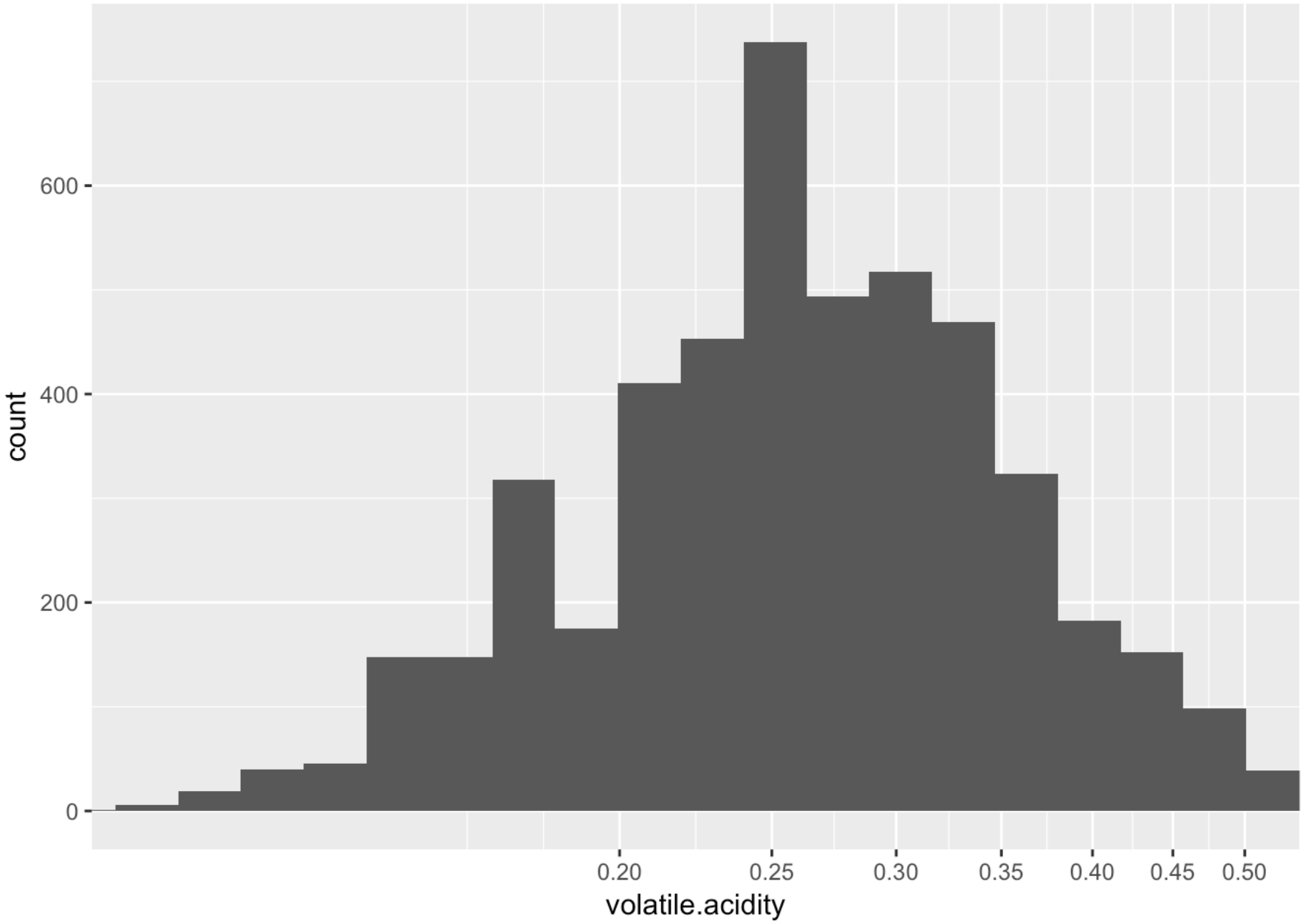

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.0   108.0   134.0   138.4   167.0   440.0
```

Total sulfur dioxide histogram also resembles normal distribution.

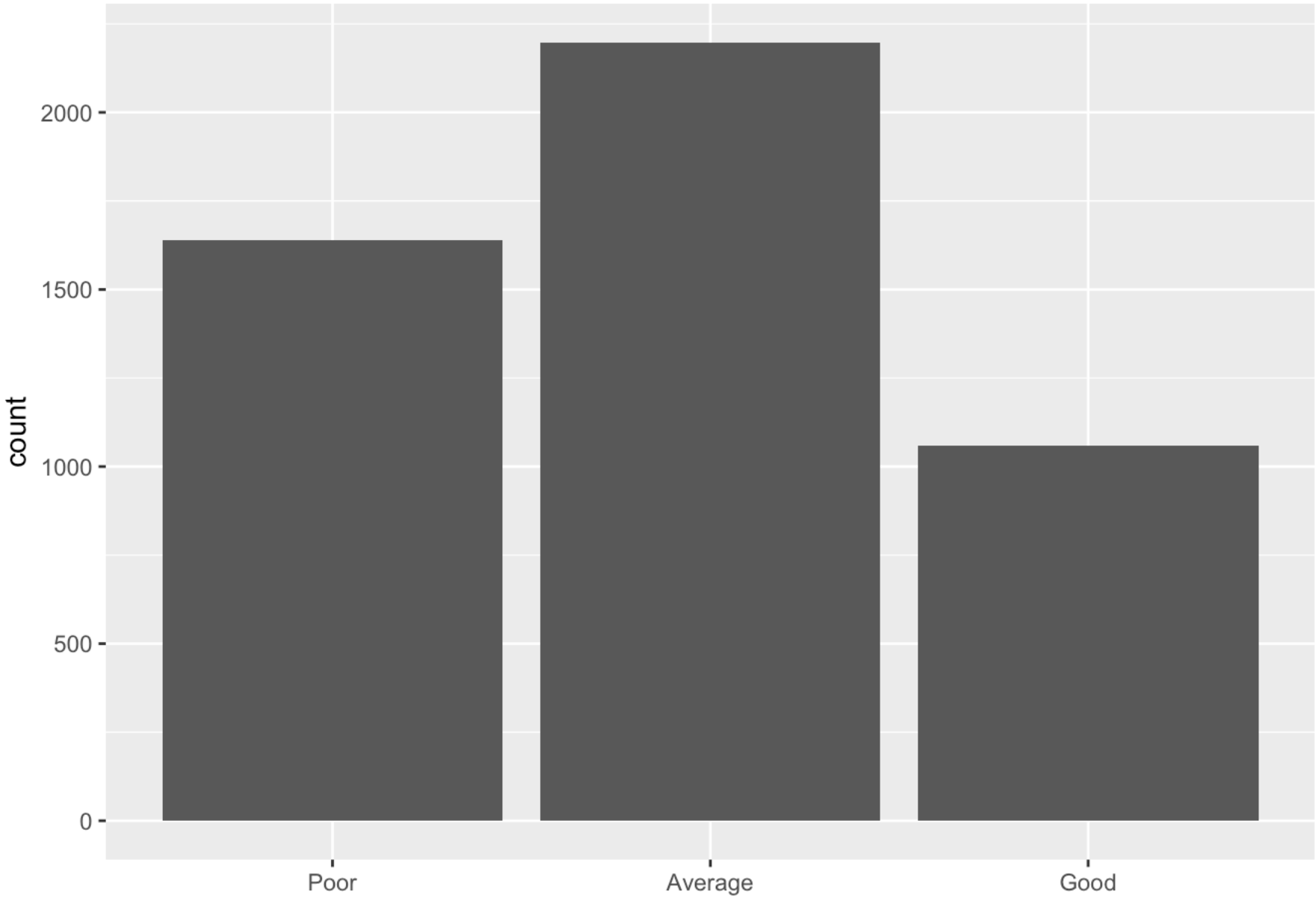



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.2100  0.2600  0.2782  0.3200  1.1000
```

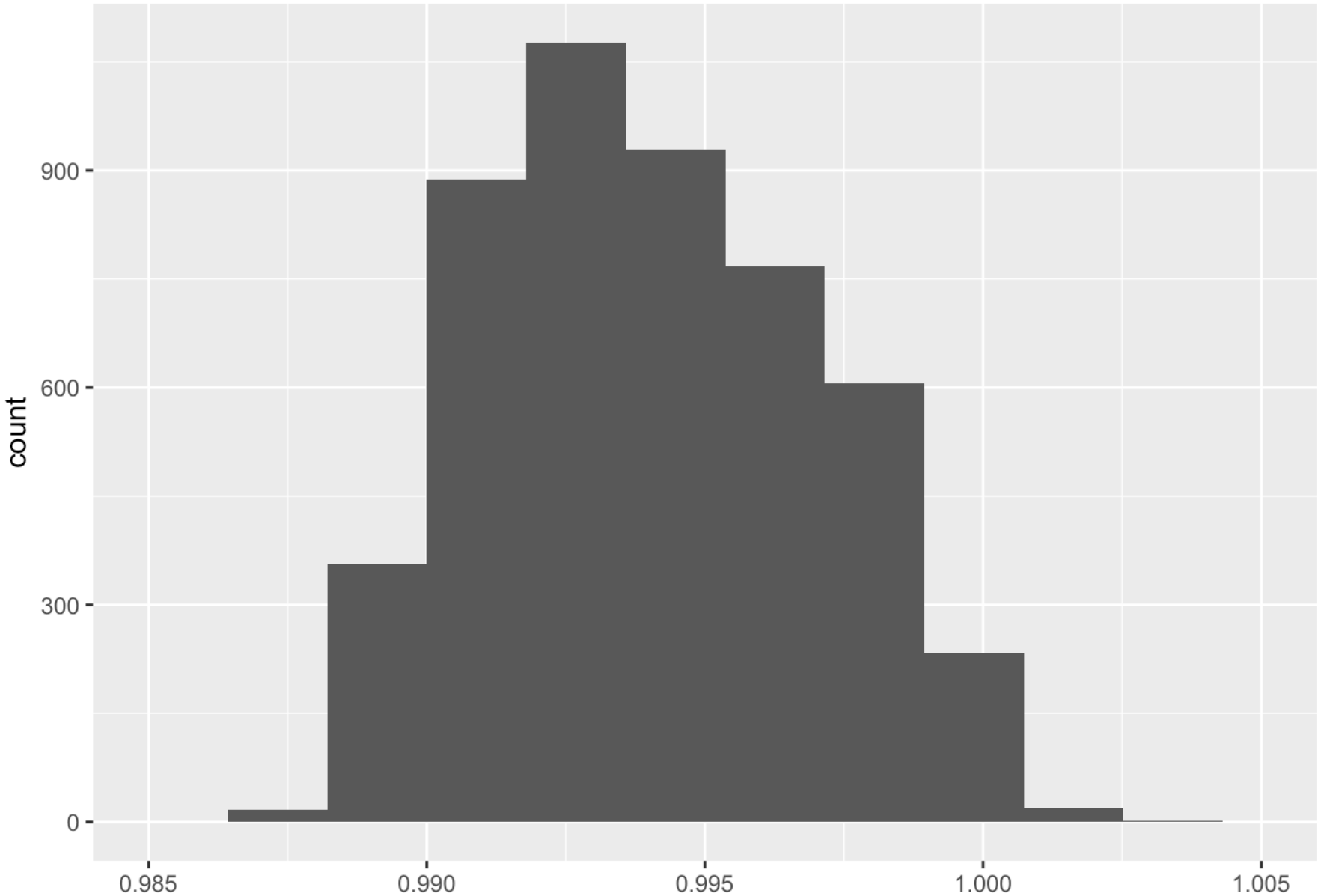
This plot resembles normal distribution, but it has a long tail. Logarithmic scale will help us to get rid of it.



This plot with log scale looks better. We can observe a peak around 0.25 g/L.

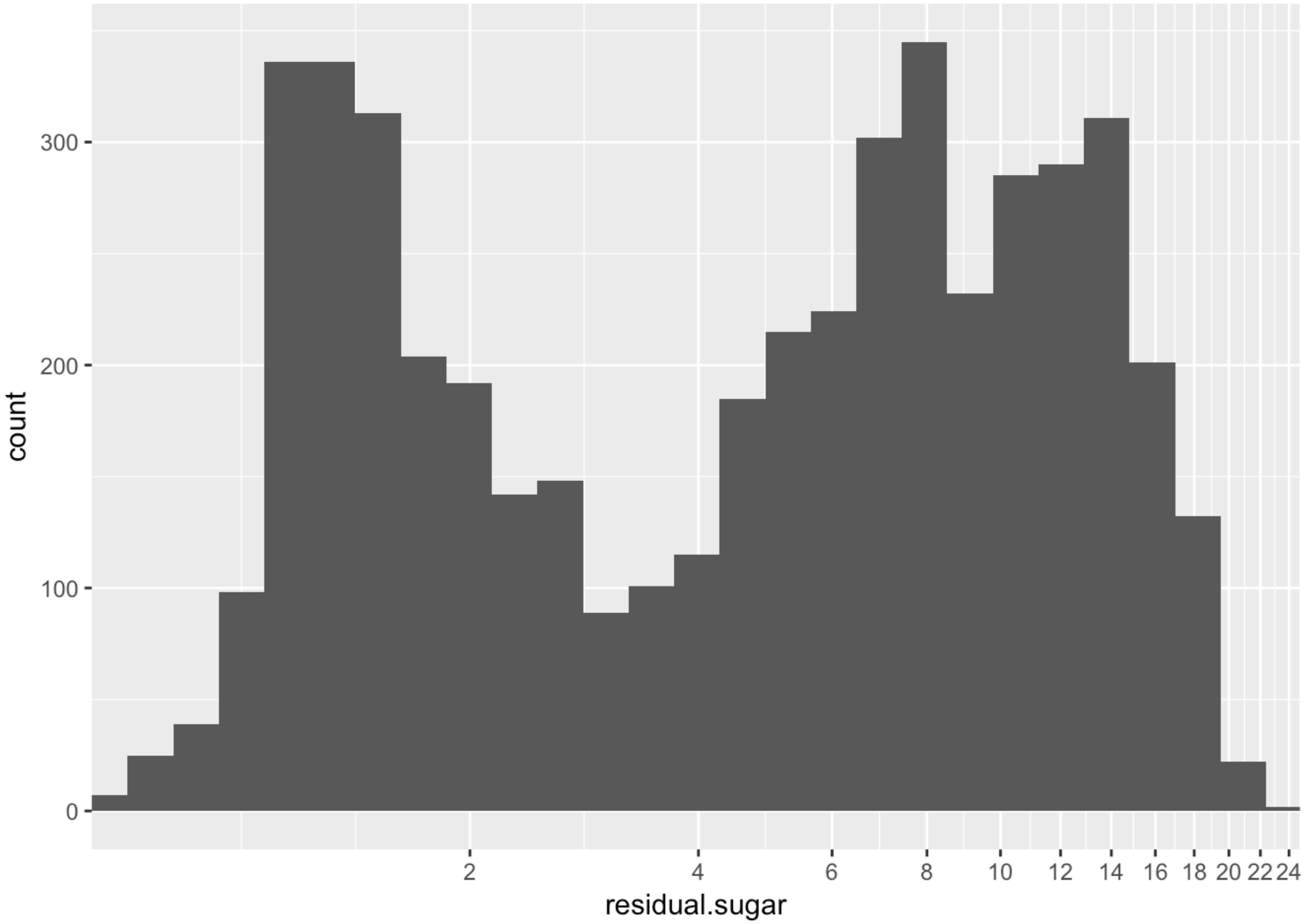


There is a plot of number of wines in each quality category. I am interested how histogram of alcohol grouped by quality would look like. I explore it later, in bivariate analysis section.

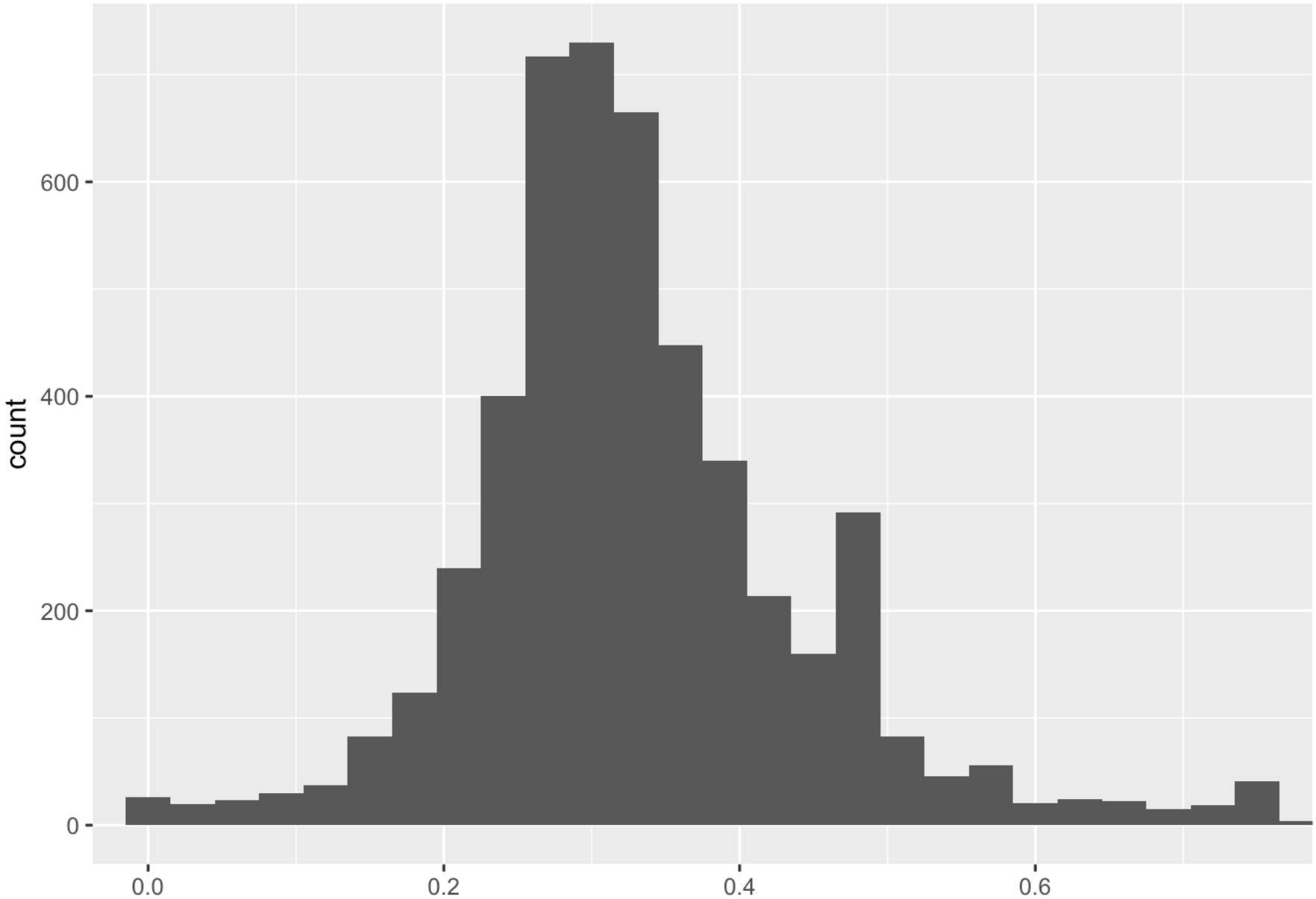



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9917  0.9937  0.9940  0.9961  1.0390
```

Density forms a normal distribution.



It is distribution with two peaks: at 1.5 and 9.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2700  0.3200  0.3342  0.3900  1.6600
```

There is a distribution with one major, 0.3; and minor peak, 0.5. Maximum value is way higher median: 1.66, so I need to pay attention to outliers.

Univariate Analysis

What is the structure of your dataset?

The dataset consists of numerical variables: fixed.acidity, volatile.acidity, citric.acid, residual.acid, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates, alcohol and quality. Quality is an integer parameter, but it would be reasonable to treat it as a categorical one.

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

The main feature I want to explore is quality. Here I have found correlations between features to find out which ones would be useful.

```
##                       fixed.acidity volatile.acidity citric.acid
```

```
##                       fixed.acidity volatile.acidity citric.acid
## fixed.acidity                  1.00            -0.02        0.29
## volatile.acidity              -0.02             1.00       -0.15
## citric.acid                    0.29            -0.15        1.00
## residual.sugar                 0.09             0.06        0.09
## chlorides                      0.02             0.07        0.11
## free.sulfur.dioxide           -0.05            -0.10        0.09
## total.sulfur.dioxide           0.09             0.09        0.12
## density                        0.27             0.03        0.15
## pH                            -0.43            -0.03       -0.16
## sulphates                     -0.02            -0.04        0.06
## alcohol                       -0.12             0.07       -0.08
## quality                       -0.11            -0.19       -0.01
##                       residual.sugar chlorides free.sulfur.dioxide
## fixed.acidity                   0.09      0.02                -0.05
## volatile.acidity                0.06      0.07                -0.10
## citric.acid                     0.09      0.11                 0.09
## residual.sugar                  1.00      0.09                 0.30
## chlorides                       0.09      1.00                 0.10
## free.sulfur.dioxide             0.30      0.10                 1.00
## total.sulfur.dioxide            0.40      0.20                 0.62
## density                         0.84      0.26                 0.29
## pH                             -0.19     -0.09                 0.00
## sulphates                      -0.03      0.02                 0.06
## alcohol                        -0.45     -0.36                -0.25
## quality                        -0.10     -0.21                 0.01
##                       total.sulfur.dioxide density    pH sulphates alcohol
## fixed.acidity                         0.09    0.27 -0.43     -0.02   -0.12
## volatile.acidity                      0.09    0.03 -0.03     -0.04    0.07
## citric.acid                           0.12    0.15 -0.16      0.06   -0.08
## residual.sugar                        0.40    0.84 -0.19     -0.03   -0.45
## chlorides                             0.20    0.26 -0.09      0.02   -0.36
## free.sulfur.dioxide                   0.62    0.29  0.00      0.06   -0.25
## total.sulfur.dioxide                  1.00    0.53  0.00      0.13   -0.45
## density                               0.53    1.00 -0.09      0.07   -0.78
## pH                                    0.00   -0.09  1.00      0.16    0.12
## sulphates                             0.13    0.07  0.16      1.00   -0.02
## alcohol                              -0.45   -0.78  0.12     -0.02    1.00
## quality                              -0.17   -0.31  0.10      0.05    0.44
##                       quality
## fixed.acidity           -0.11
## volatile.acidity        -0.19
## citric.acid             -0.01
## residual.sugar          -0.10
## chlorides               -0.21
## free.sulfur.dioxide      0.01
## total.sulfur.dioxide    -0.17
## density                 -0.31
## pH                       0.10
## sulphates                0.05
## alcohol                  0.44
```

```
## quality                1.00
```

What other features in the dataset do you think will help support your

investigation into your feature(s) of interest?

There are negatively correlated features I want to pay attention to: volatile.acidity, chlorides, density, total.sulfur.dioxide and positevely correlated: alcohol. However, it is possible that I will find out something interesting about citric acid and residual sugar.

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

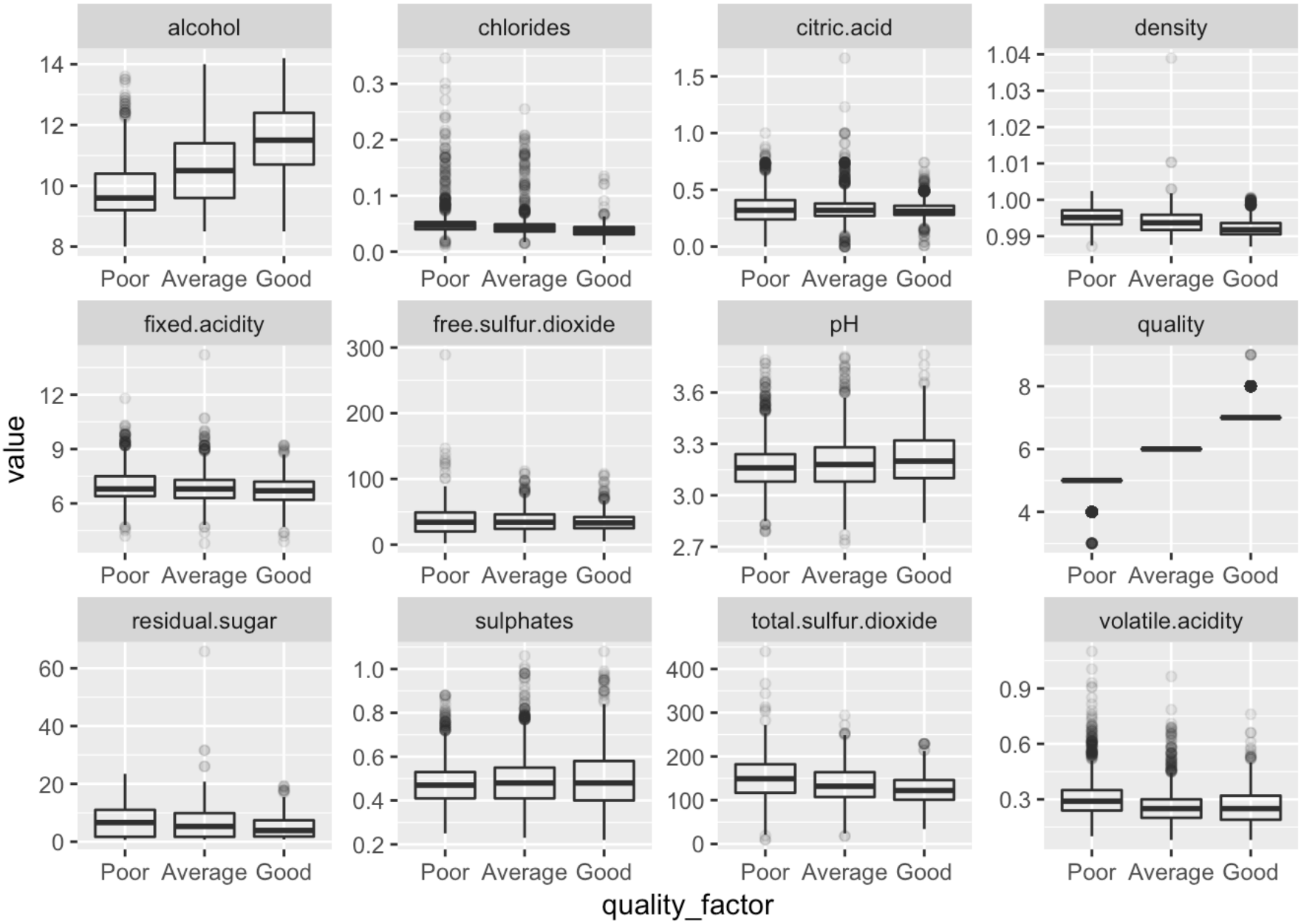
I created a categorical variable for wine quality: “Poor” for wines graded lower than 6, “Average” for 6 and “Good” for 7 and higher.

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

Most of the distributions are usual stadard distributions apart from residual.sugar, and chlorides and volatile acidity with a long right tail.

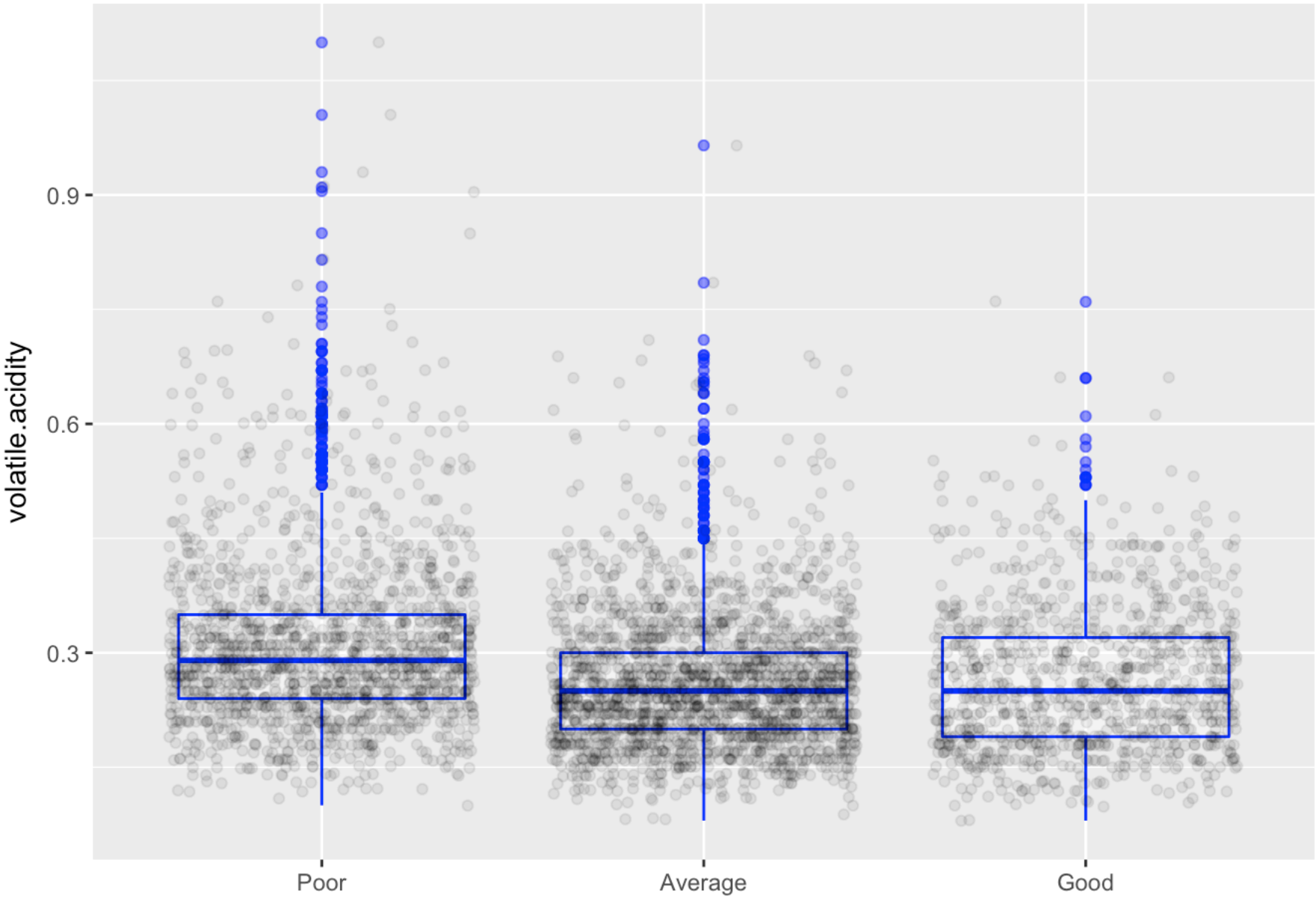
Bivariate Plots Section

I have already found which features are highly correlated with the target feature, quality. For more insights I built a matrix of boxplots



There are some interesting insights: the range of alcohol in good wines is as wide as the range of all wines, while low-rated wines almost never have alcohol higher than 12. The lowest graded wines have less residual sugar. For more detailed analysis, I will build larger, separate plots for every interesting variable vs quality.

Scatterplot and boxplot of volatile.acidity and quality



There were two peaks found in a previous section: at 0.15 and 0.25. This plot explains these peaks: there is a lower boundary around 0.15 and median around 0.25. Poor-quality wines in general have more volatile acidity. Good ones have less, and it has higher standard deviation. There is a main statistical information above.

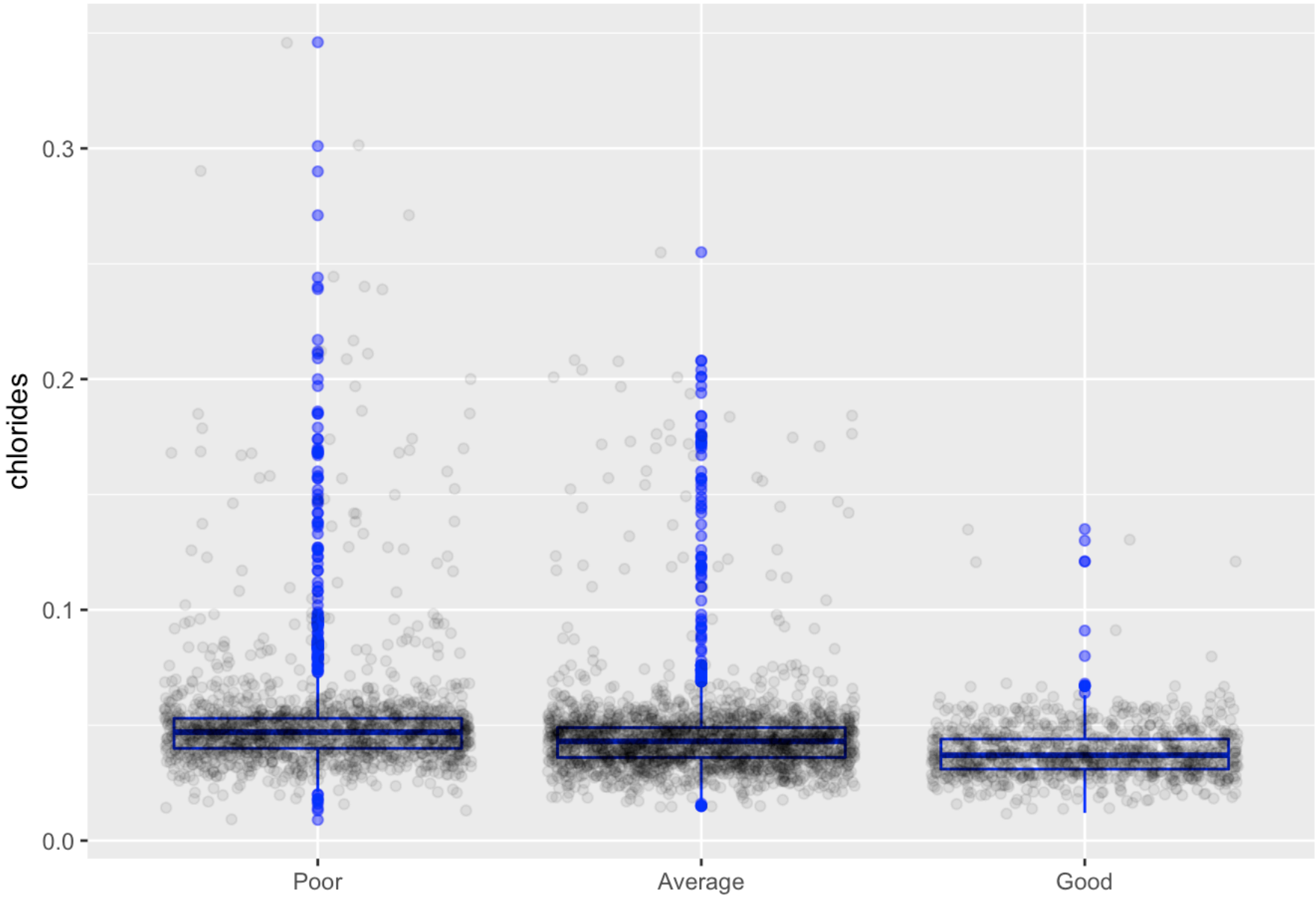
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.2100  0.2600  0.2782  0.3200  1.1000
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1000  0.2400  0.2900  0.3103  0.3500  1.1000
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.2000  0.2500  0.2606  0.3000  0.9650
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.1900  0.2500  0.2653  0.3200  0.7600
```

Scatterplot and boxplot of chlorides and quality



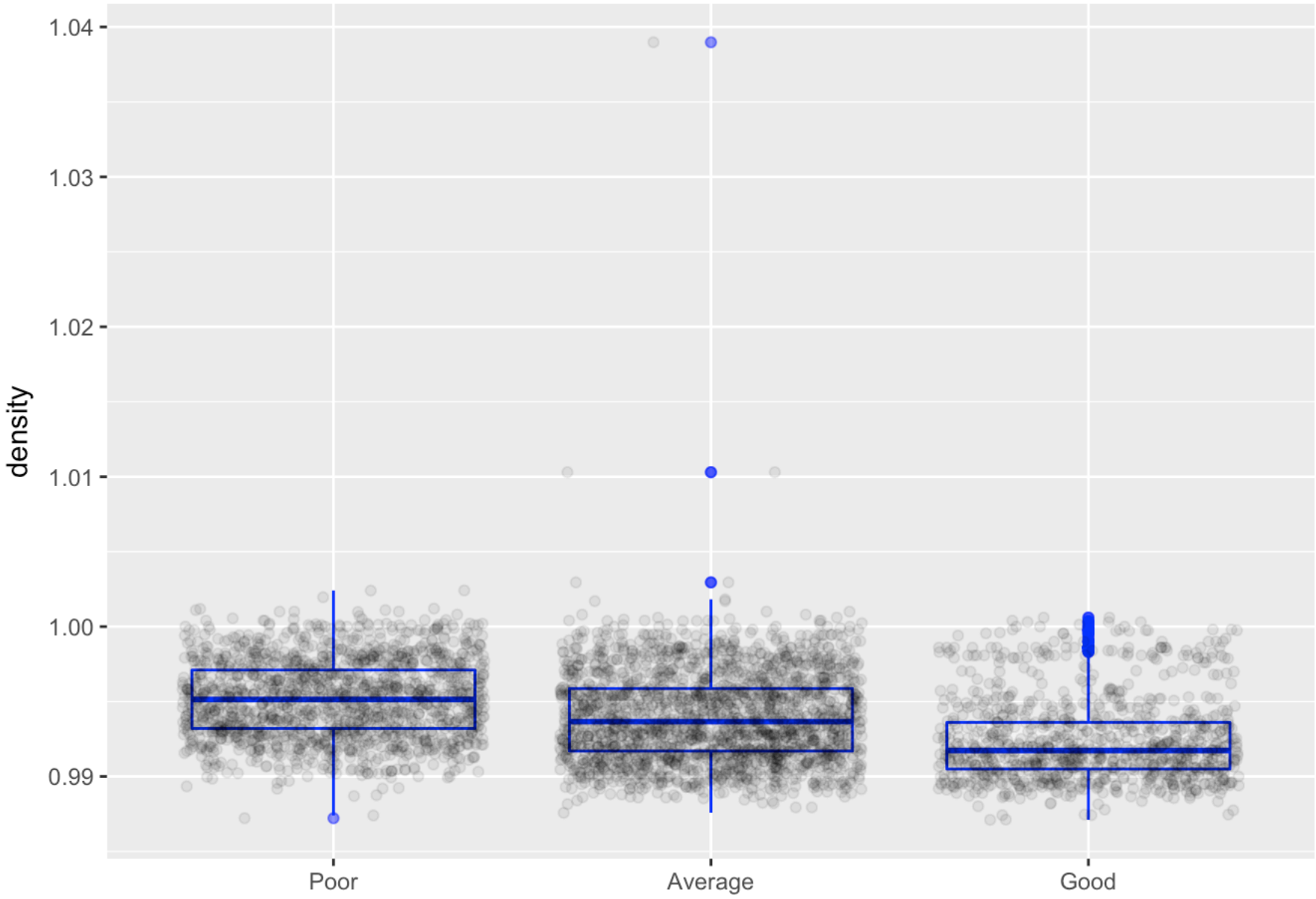
Now it is clear that poor and average wines with extremely high chlorides level contribute to the long tail found in a previous section. There is a tendency of lowering the chlorides level with raising of quality.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00900 0.04000 0.04700 0.05144 0.05300 0.34600
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01500 0.03600 0.04300 0.04522 0.04900 0.25500
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.03100 0.03700 0.03816 0.04400 0.13500
```

Scatterplot and boxplot of density and quality



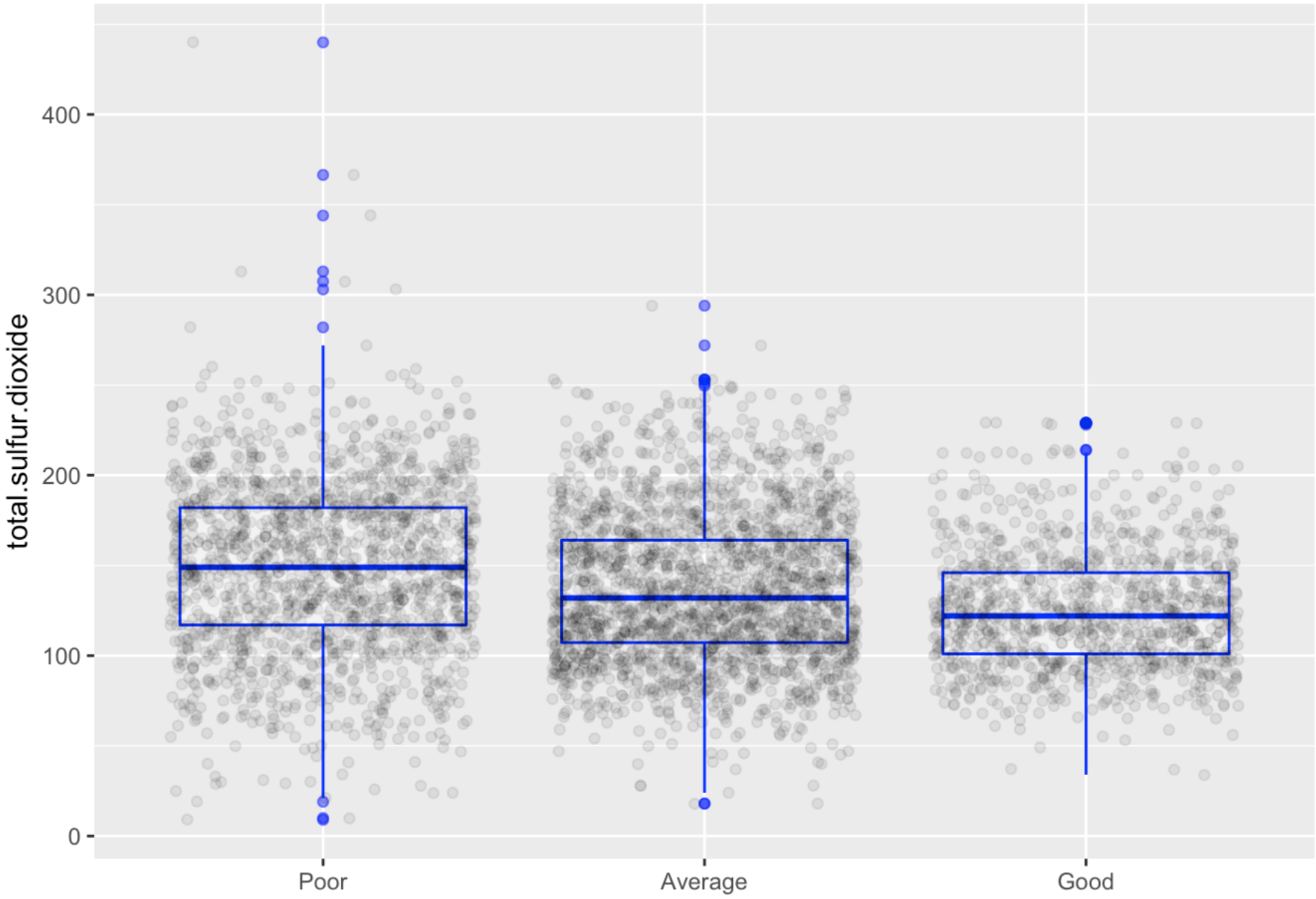
There are a few outliers among average wines. There is a tendency of lowering density level with raising of quality.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9872  0.9932  0.9951  0.9952  0.9971  1.0024
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9876  0.9917  0.9937  0.9940  0.9959  1.0390
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9905  0.9917  0.9924  0.9936  1.0006
```

Scatterplot and boxplot of total.sulfur.dioxide and quality



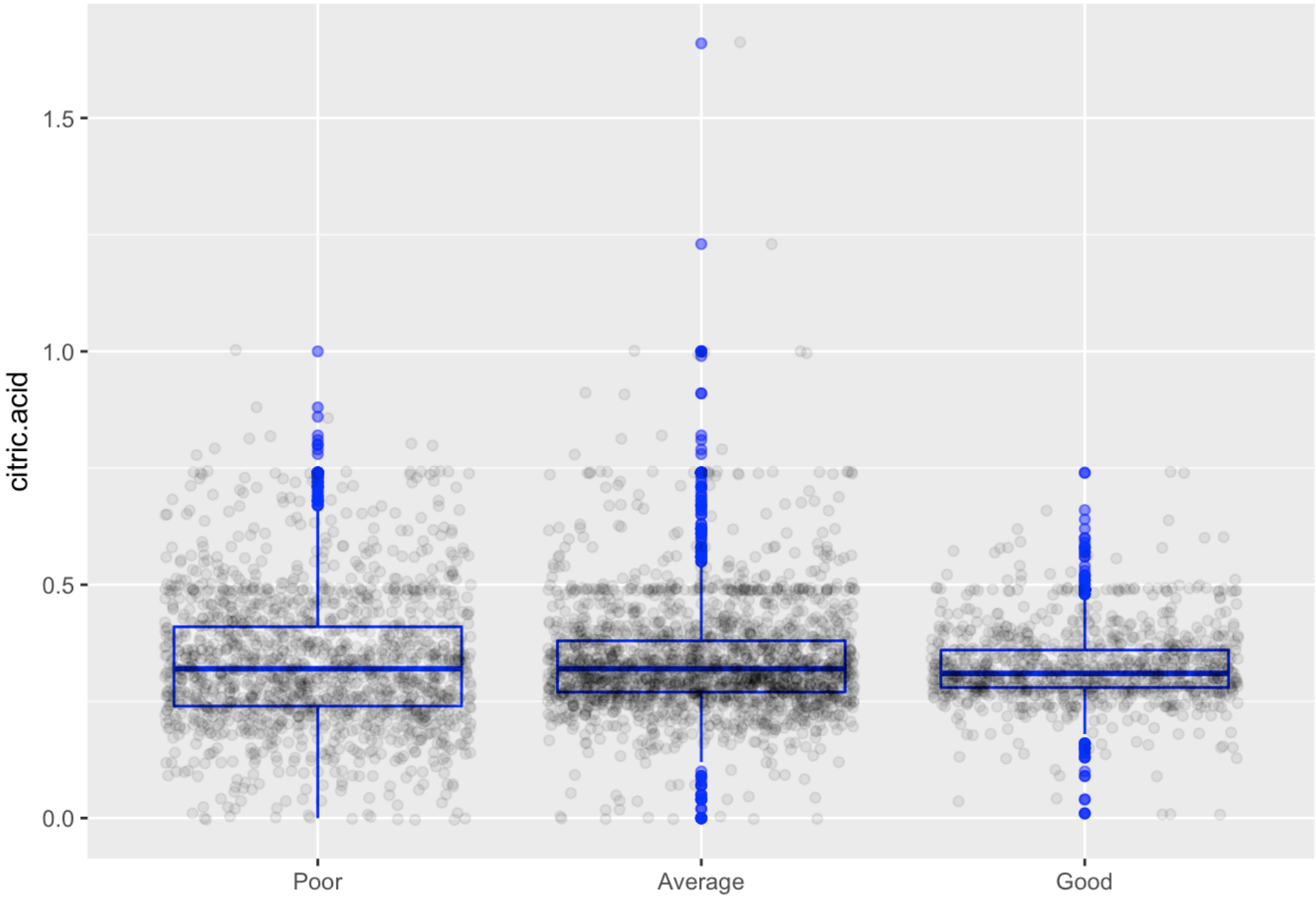
There is a tendency of lowering sulfur dioxide with raising of quality. Standard deviation is also becoming lower.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.0   117.0   149.0   148.6   182.0   440.0
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    18.0   107.2   132.0   137.0   164.0   294.0
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    34.0   101.0   122.0   125.2   146.0   229.0
```

Scatterplot and boxplot of citric.acid and quality



I noticed an unusual peak around 0.5 in citric acid histogram. This plots shows that there are lots of wines that have exactly 0.5 g/L of citric acid. Median of good wines is slightly lower compared to other categories. Standard deviation is getting lower with raising of quality.

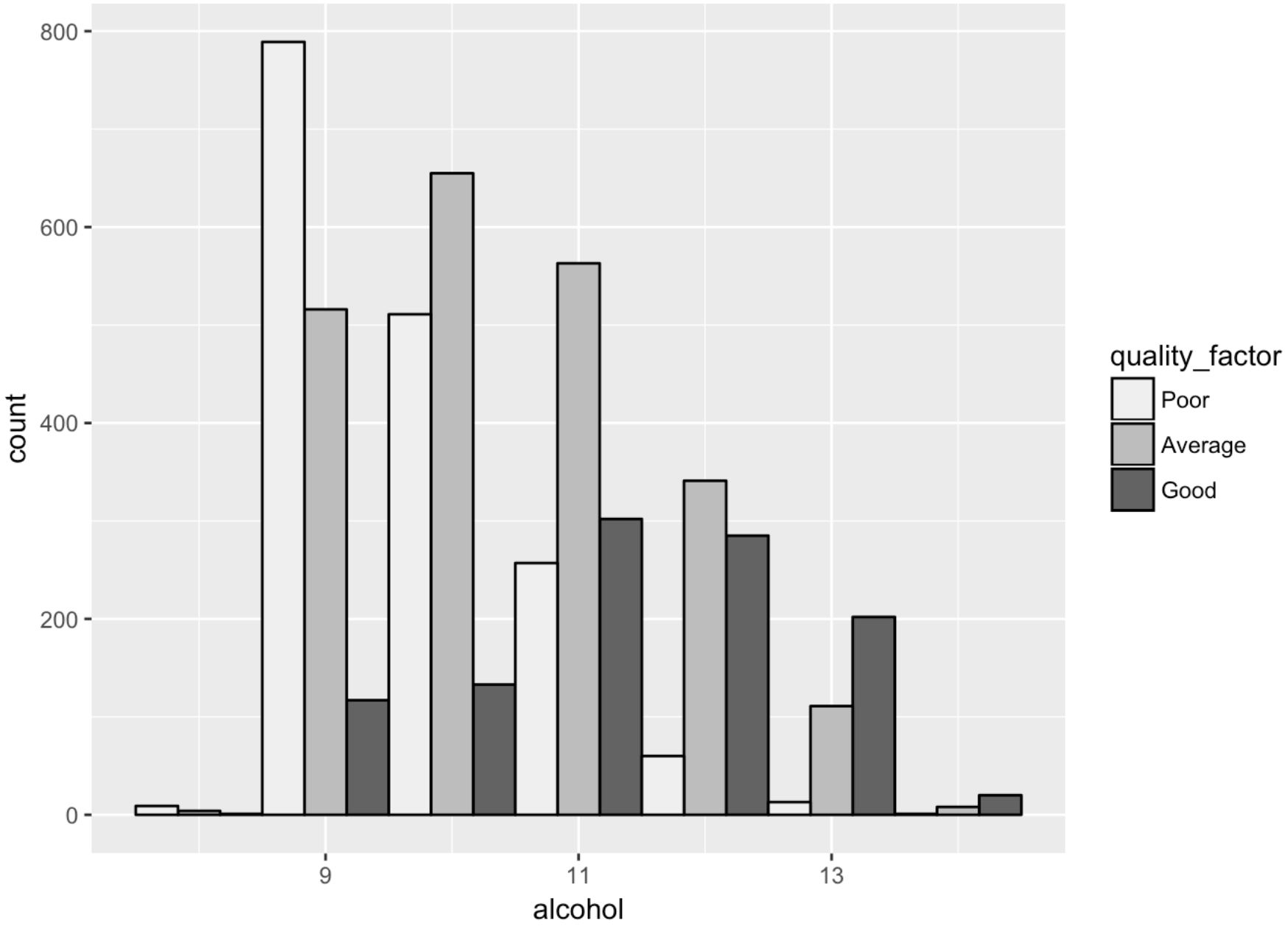
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2400  0.3200  0.3343  0.4100  1.0000
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.270   0.320   0.338   0.380   1.660
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0100  0.2800  0.3100  0.3261  0.3600  0.7400
```

Scatterplot and boxplot of alcohol and quality

Good wines tend to contain more alcohol. It would be interesting to have a look to a histogram of alcohol level separated by quality.



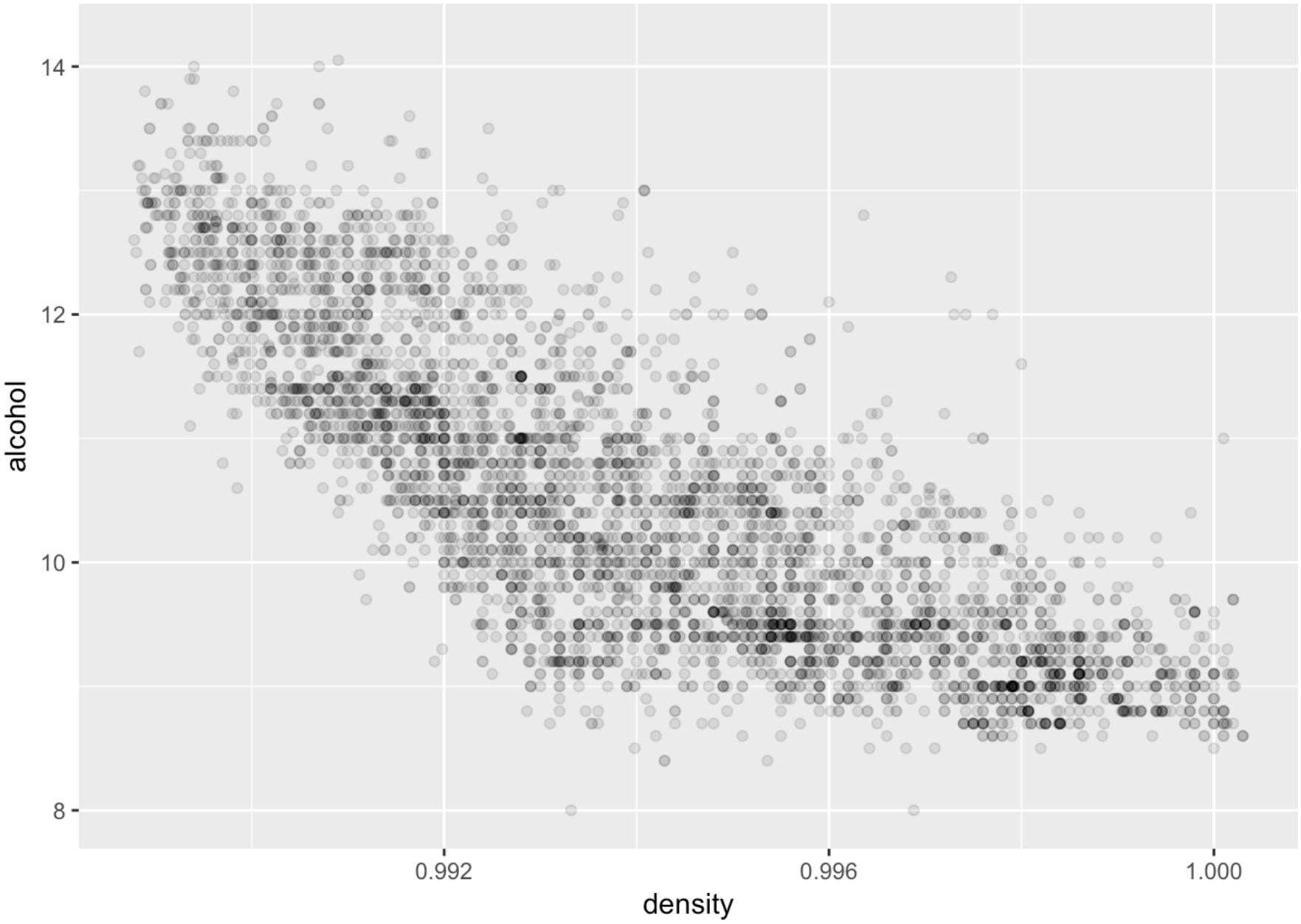
It is interesting how wines are distributed. Good wines frequently have higher alcohol level. Low quality wines have the lowest alcohol level.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.20    9.60    9.85   10.40   13.60
```

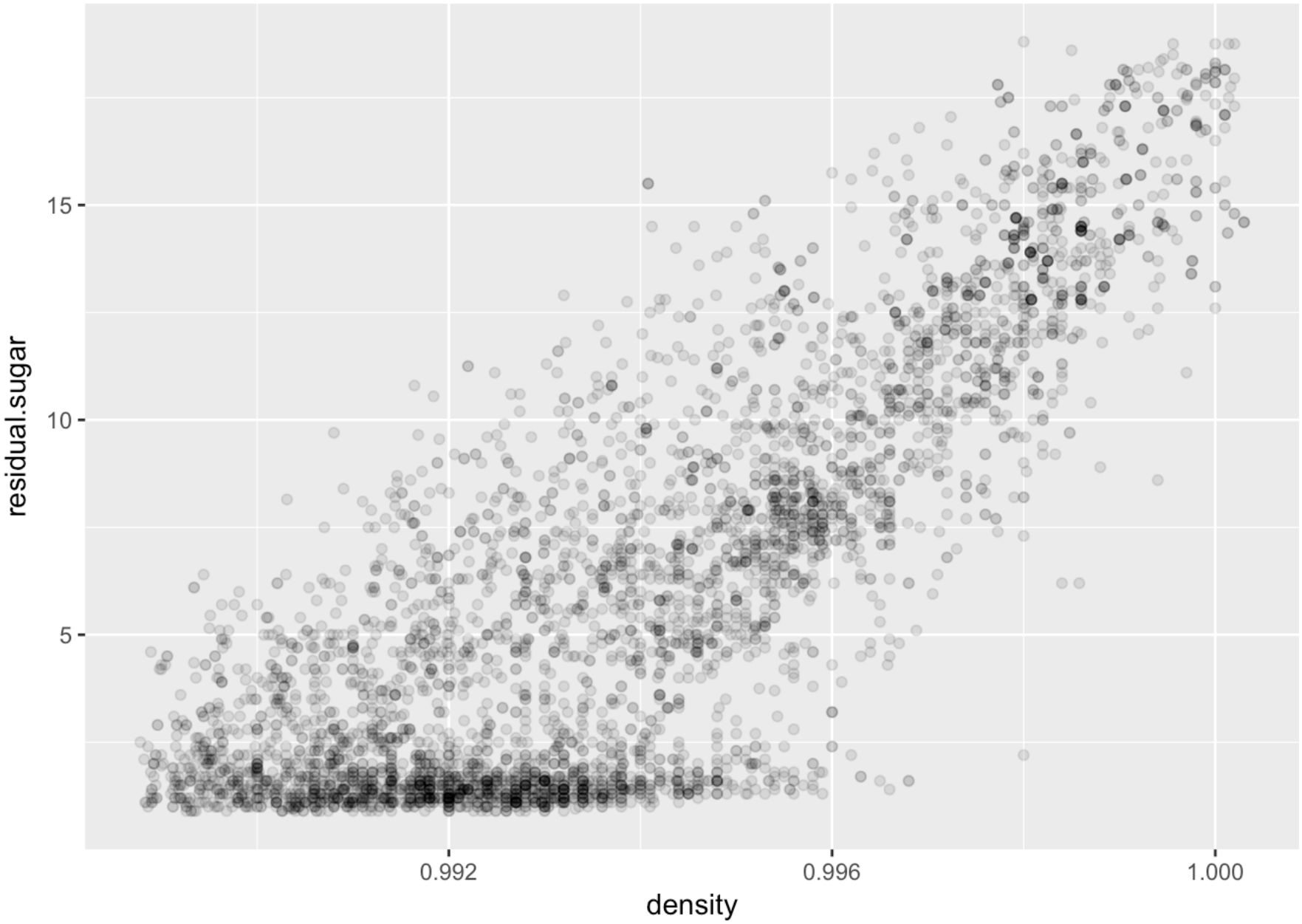
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50    9.60   10.50   10.58   11.40   14.00
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50   10.70   11.50   11.42   12.40   14.20
```

The correlation between density and alcohol is very high, as well as for density and residual.sugar. I will draw plot these plots.



Alcohol vs density plot shows that there are almost no wines with both high alcohol and high density, and no wines with less than average alcohol combined with less than average density.



Interesting things with density vs residual sugar: there are almost no wines with both high density and low residual sugar, as well as there are no wines with low density and high sugar.

Bivariate Analysis

The most interesting thing I discovered is a relationship of alcohol vs quality. There is a tendency of increasing an alcohol level with increasing its quality. Density is the highest in wines with the lowest grade.

Did you observe any interesting relationships between the other features

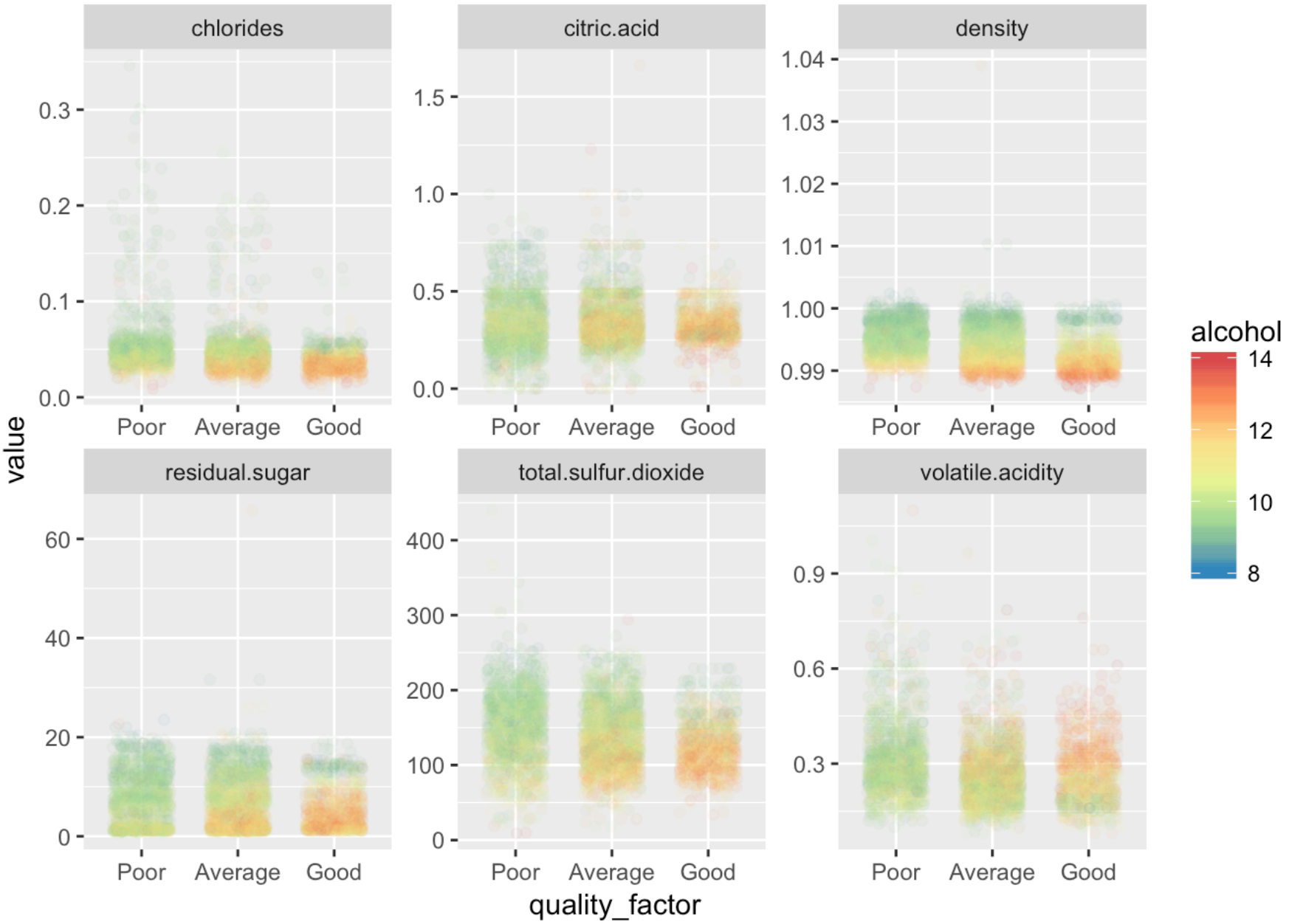
There are two clusters in a plot with residual.sugar and density.

What was the strongest relationship you found?

There is a very strong relationship between alcohol and quality: the highest rated wines have higher alcohol level.

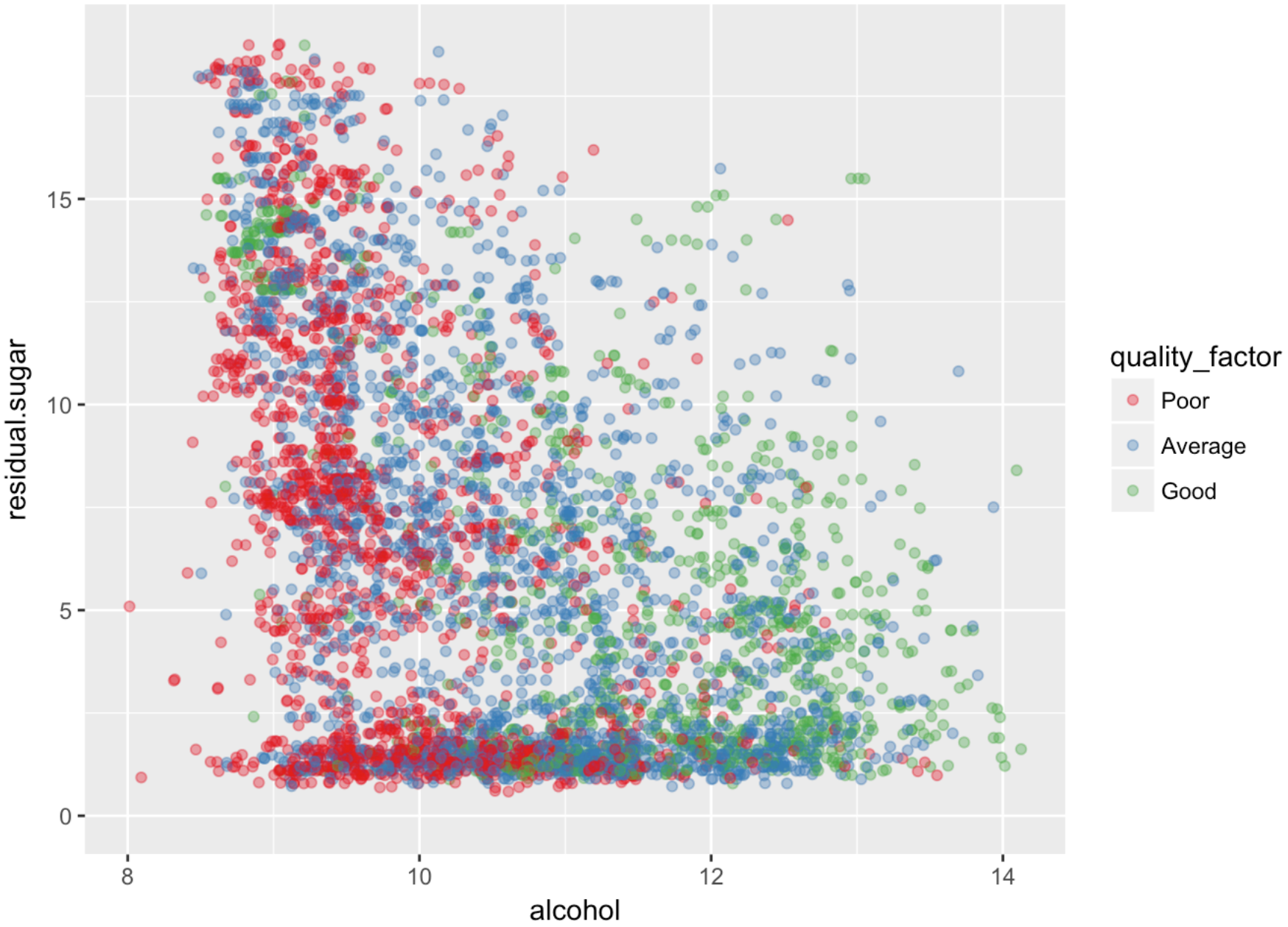
Multivariate Plots Section

I will build scatterplots with quality, alcohol and other features which could be interesting.



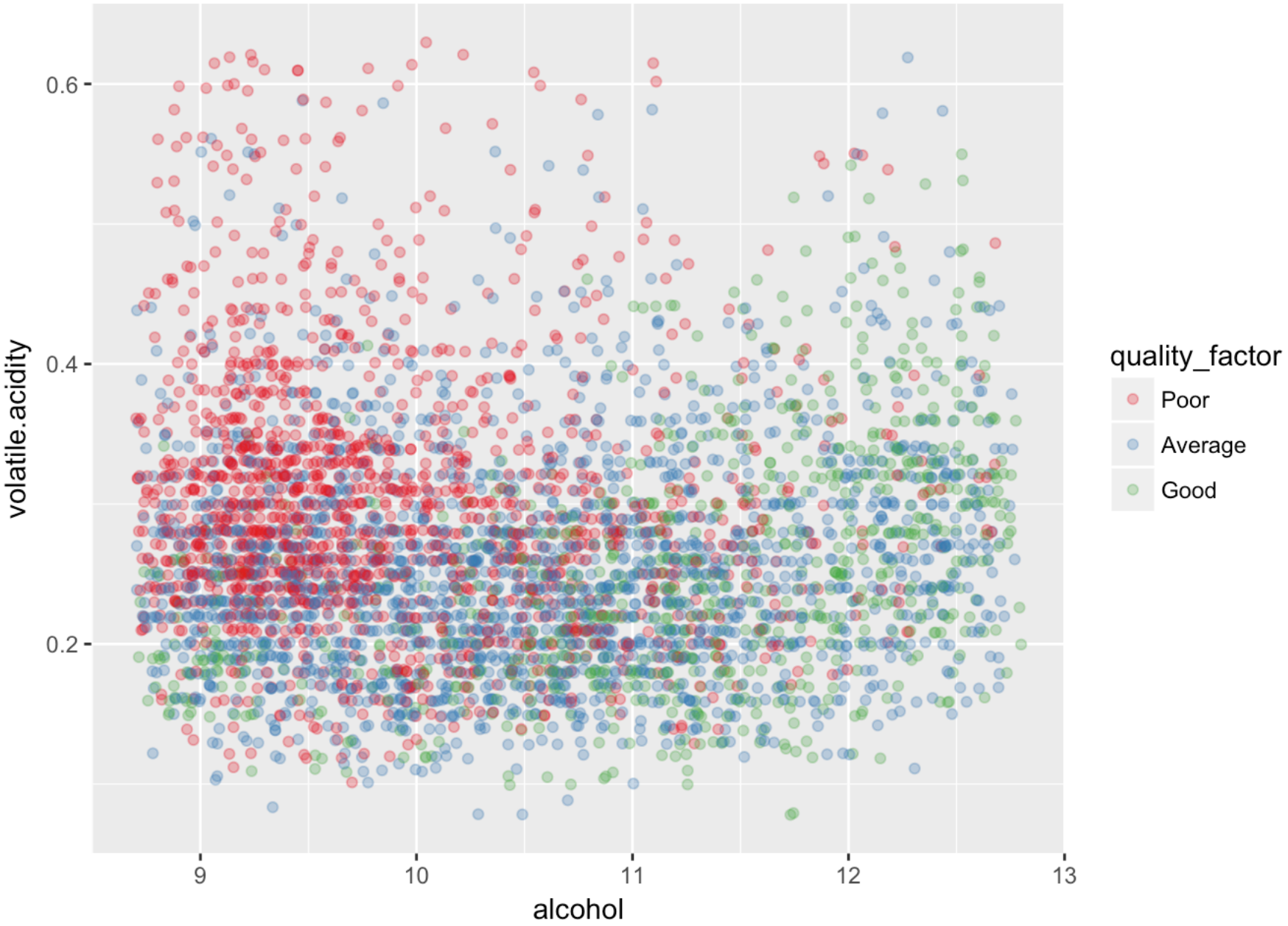
Then I want to have a closer look to features' relations.

There is a scatterplot with alcohol, sugar and quality. High quality wines tend to have from 10 ABV to 14 ABV and, in most cases, no more than 10 gram per litre of residual sugar. Most of low quality wines have a range of alcohol from 9 ABV to 10.5 ABV and sugar from 1 g/L to 18 g/L.

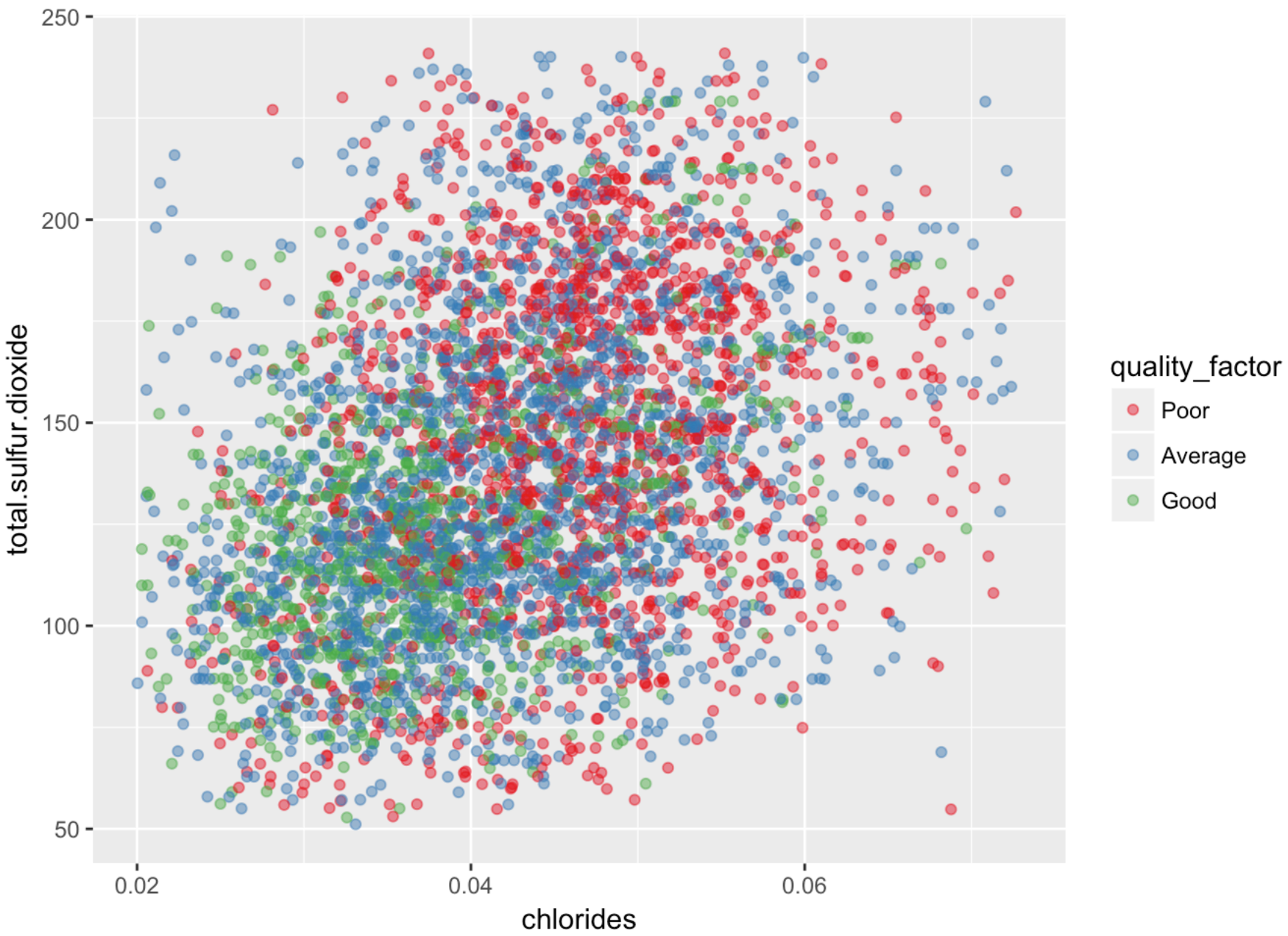


Then there is a plot with citric acid, density and quality. High quality wines tend to have less density. The interesting thing is that there is a visibile peak of citric acid at 0.49 g/Ml.

There is a scatterplot with alcohol, volatile.acidity and quality. While a majority of the highest quality wines have at least 11.5 ABV, there are some wines with less alcohol. These wines have quite a low volatile acidity. The least graded wines have a high acidity.



Two last features I have not analyzed in deatils yet: total.sulfur.dioxide and chlorides. High-graded wines have fewer sulfur dioxide and chlorides.



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?

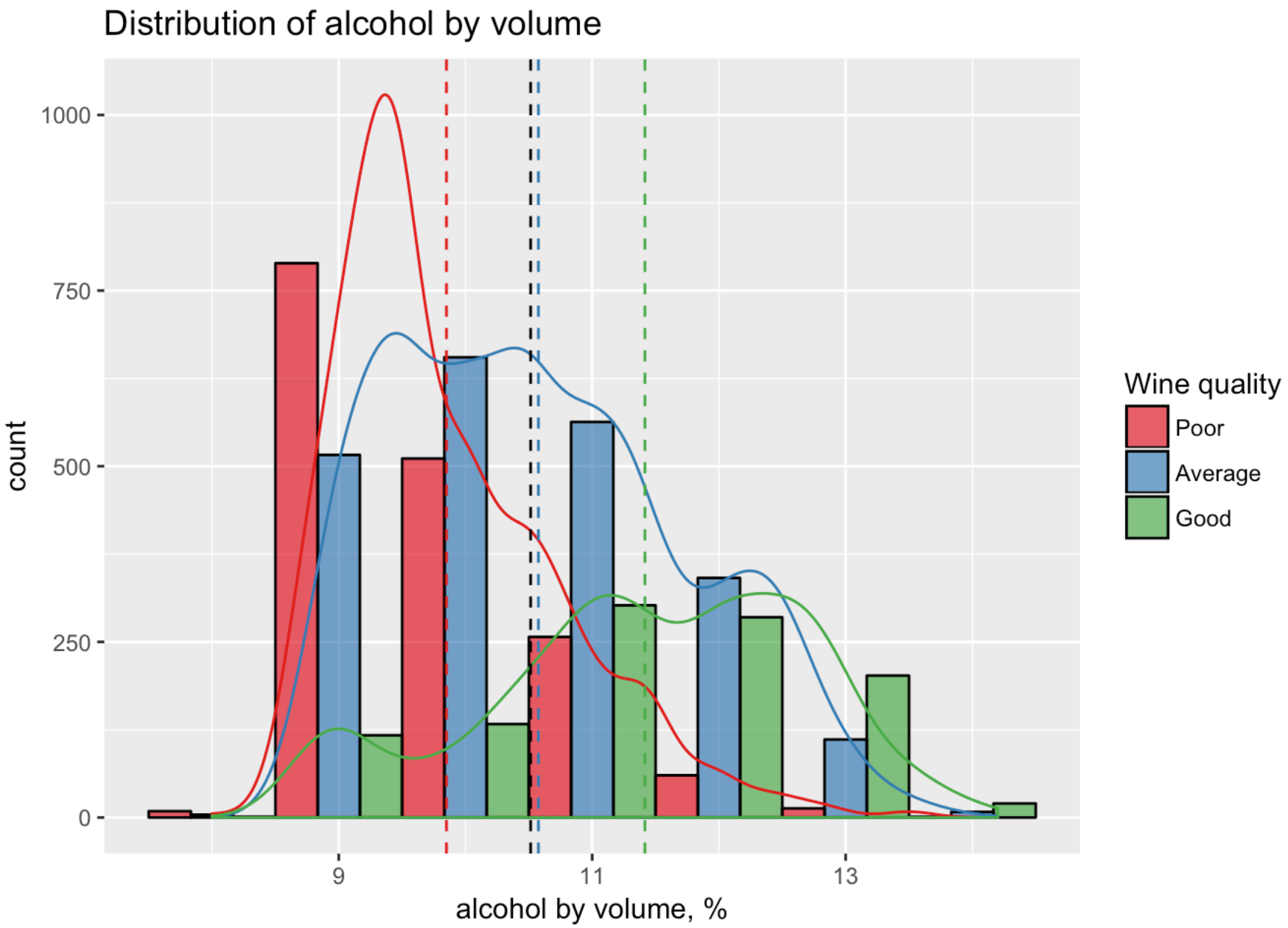
Goods wines have more alcohol and less than average residual sugar. Low-ranked wines have less alcohol and all range of residual sugar. Amount of sugar decreases with growth of quality. Good wines have less than average density. High rated wines have high level of alcohol and all range of volatile acidity. Howewer, good wines with less alcohol have low acidity. High-graded wines have fewer sulfur dioxide and chlorides, while low-graded wines are high with both. Turns out that alcohol level has the highest influence on wine's grade. There are some exceptions from this observation: wines with high residual sugar and low alcohol could be really good, as well as wines with low alcohol and low volatile acidity.

Were there any interesting or surprising interactions between features?

Cholrides and total sulfur dioxide both negatively affect wine grade. The interesting thing is that all the good wines have less than average of sulfur dioxide and cholrides at the same time. Good wines tend to have more alcohol with one exception: combination of low ABV and high residual sugar.

Final Plots and Summary

Plot One



Description One

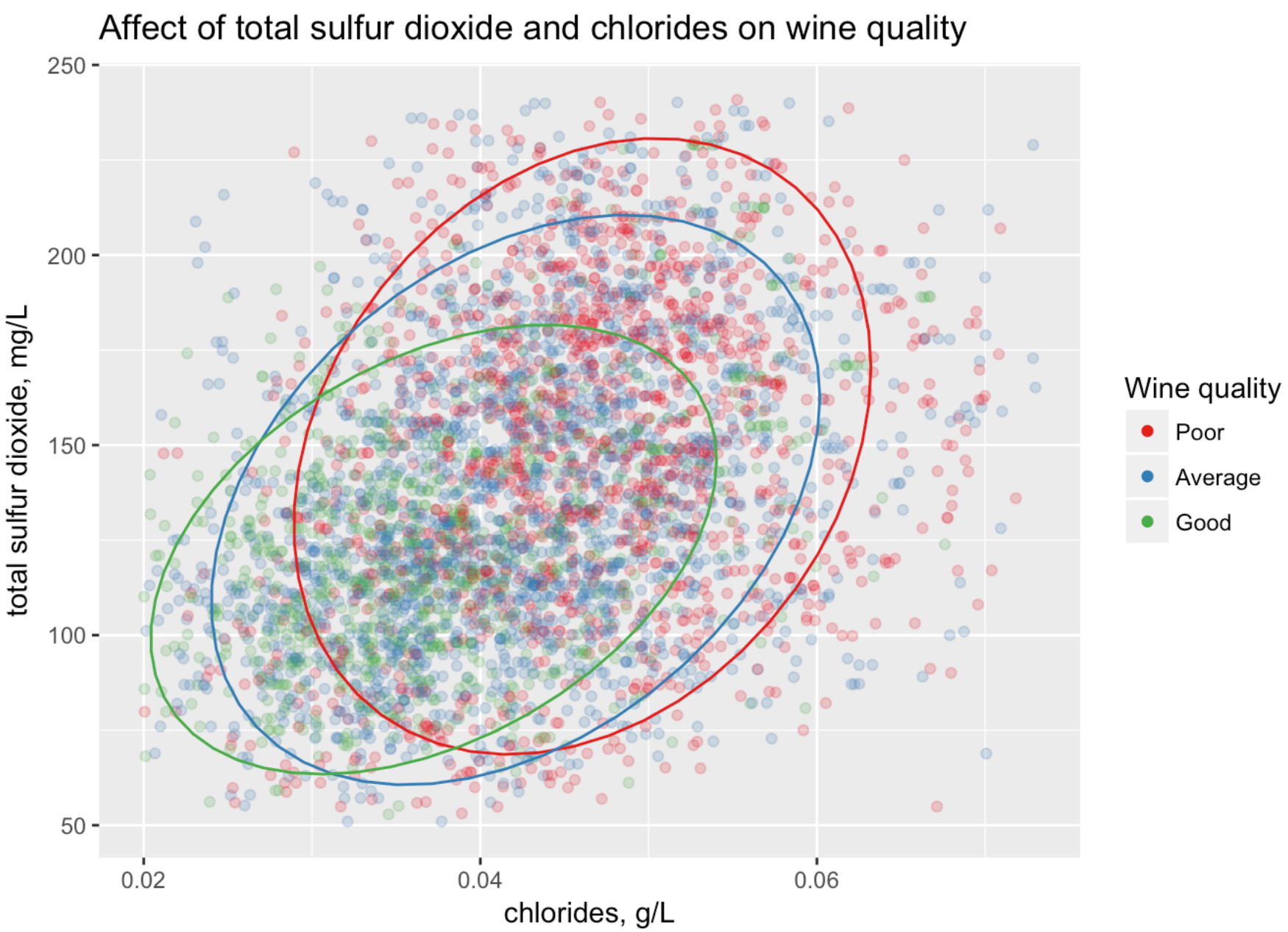
There is a histogram of alcohol by volume, split by wine quality. I noticed a strong correlation between quality and ABV and decided to look at it closer. I made three histograms with corresponding density plots and marked three means along with the general mean. Low quality wines have a peak at 9.5% and a mean around 9.75%. Average wines have a peak at 10.5%, almost like all the wines combined. Good wines distribution does not have an obvious peak. The mean is at 11.4%. It is interesting that there are just a few low-graded wines with AVB higher than this value.

Plot Two

Description Two

This plot describes volatile acidity, alcohol and quality. Ellipses cover 80% of samples. According to this plot, poor wines form a cluster with alcohol from 8.5% to 11% and acidity from 0.15 g/L to 0.43 g/L. Average wines ellipse is wider: it has alcohol range 8.5%-12% and acidity from 0.15 g/L to 0.35 g/L. Good wines stay apart from others: the center of the cluster is significantly on the right at the alcohol scale. Alcohol varies from 9.5% to 13%, acidity varies from 0.12 to 0.375.

Plot Three



Description Three

This plot describes affect of chlorides and sulfur dioxide on wine quality. Ellipses represent 90% of samples. Good wines have lower chlorides and sulfur dioxide, poor wines contains more chlorides and sulfur dioxide. Despite of this trend, the ellipses share quite a wide common area.

Reflection

I explored how different parameters are associated with wine quality. I split wine quality for three categories: poor, average and good and perform an analysis. In general, good wines contain more alcohol and volatile acidity, less chlorides, sulfur dioxide, density and sugar. I was challenged with building a linear model for quality classification. I used alcohol, volatile acidity, chlorides, sulfur dioxide, density and sugar as features to predict a quality class. The accuracy of the model was only 60%. For better results it would be useful to tune the model better or use more advanced method.

Resources

https://stackoverflow.com (https://stackoverflow.com)
https://www.statmethods.net (https://www.statmethods.net)
http://www.sthda.com/ (http://www.sthda.com/)
http://ggplot2.tidyverse.org/index.html (http://ggplot2.tidyverse.org/index.html)