RED WINE ANALYSIS

by Egor Ermilov

Let's take a look at the structure of our dataset:

```
## 'data.frame':    1599 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071
 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

Here is a short description of the variables to understand the dataset better:

1 - fixed acidity (tartaric acid - g / dm^3): most acids involved with wine or fixed or nonvolatile (do not evaporate readily)

2 - volatile acidity (acetic acid - g / dm^3): the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste

3 - citric acid (g / dm^3): found in small quantities, citric acid can add 'freshness' and flavor to wines

4 - residual sugar (g / dm^3): the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet

5 - chlorides (sodium chloride - g / dm^3): the amount of salt in the wine

6 - free sulfur dioxide (mg / dm^3): the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine

7 - total sulfur dioxide (mg / dm^3): amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine

8 - density (g / cm^3): the density of water is close to that of water depending on the percent alcohol and sugar content

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

10 - sulphates (potassium sulphate - g / dm3): a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant

11 - alcohol (% by volume): the percent alcohol content of the wine

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

We don't need the X variable in our dataset

Let's look at the summaries of every variable:

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00
##  Median :0.07900   Median :14.00       Median : 38.00
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00
##     density             pH          sulphates         alcohol
##  Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40
##  1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50
##  Median :0.9968   Median :3.310   Median :0.6200   Median :10.20
##  Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42
##  3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10
##  Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.636
##  3rd Qu.:6.000
##  Max.   :8.000
```

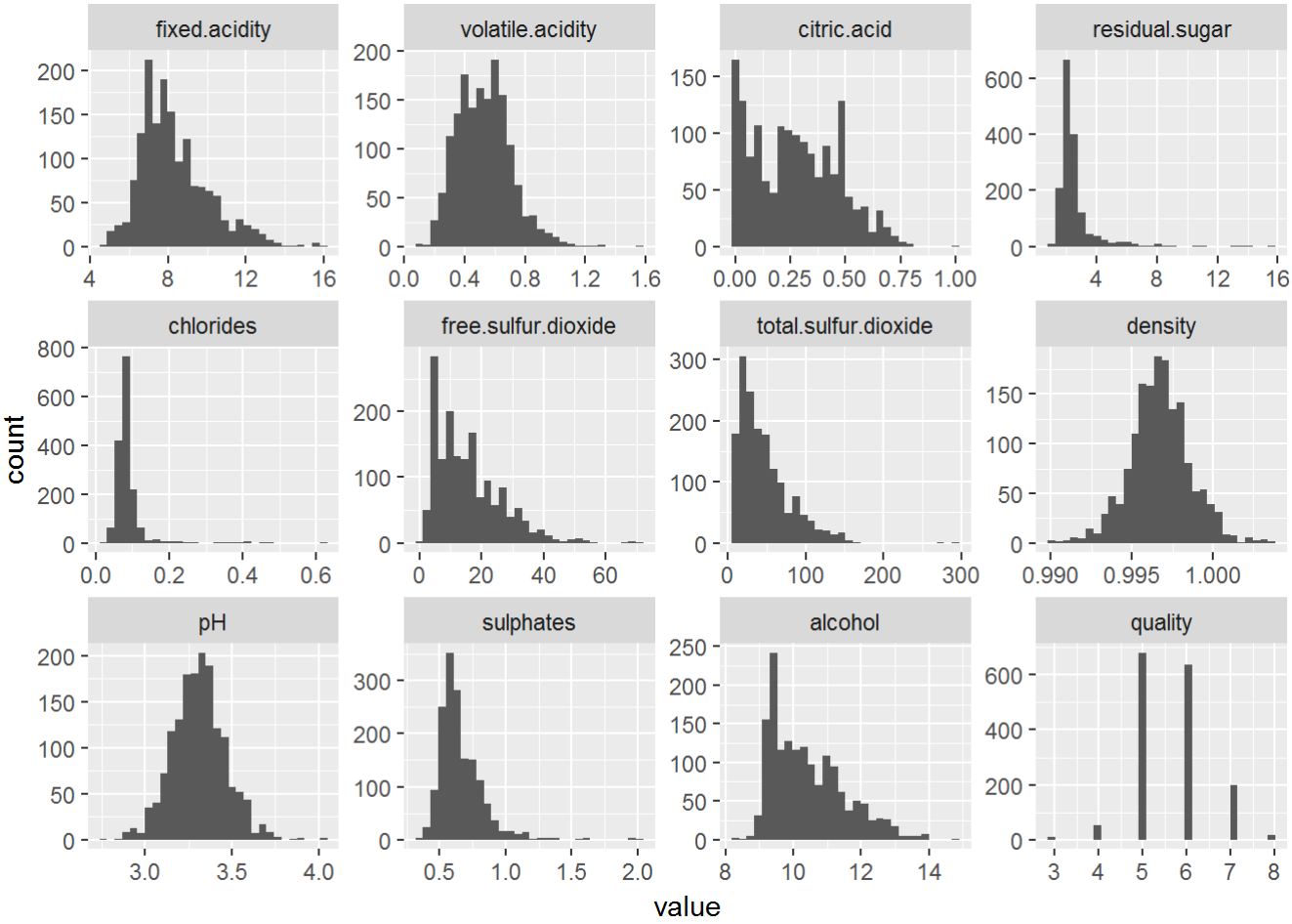
Univariate Plots Section

In addition to the summaries, let's take a look at the distribution of our variables. To show several similar histograms in one plot we need to use facets, therefore we shoud have our dataset molten

How many NA row does the dataset have:

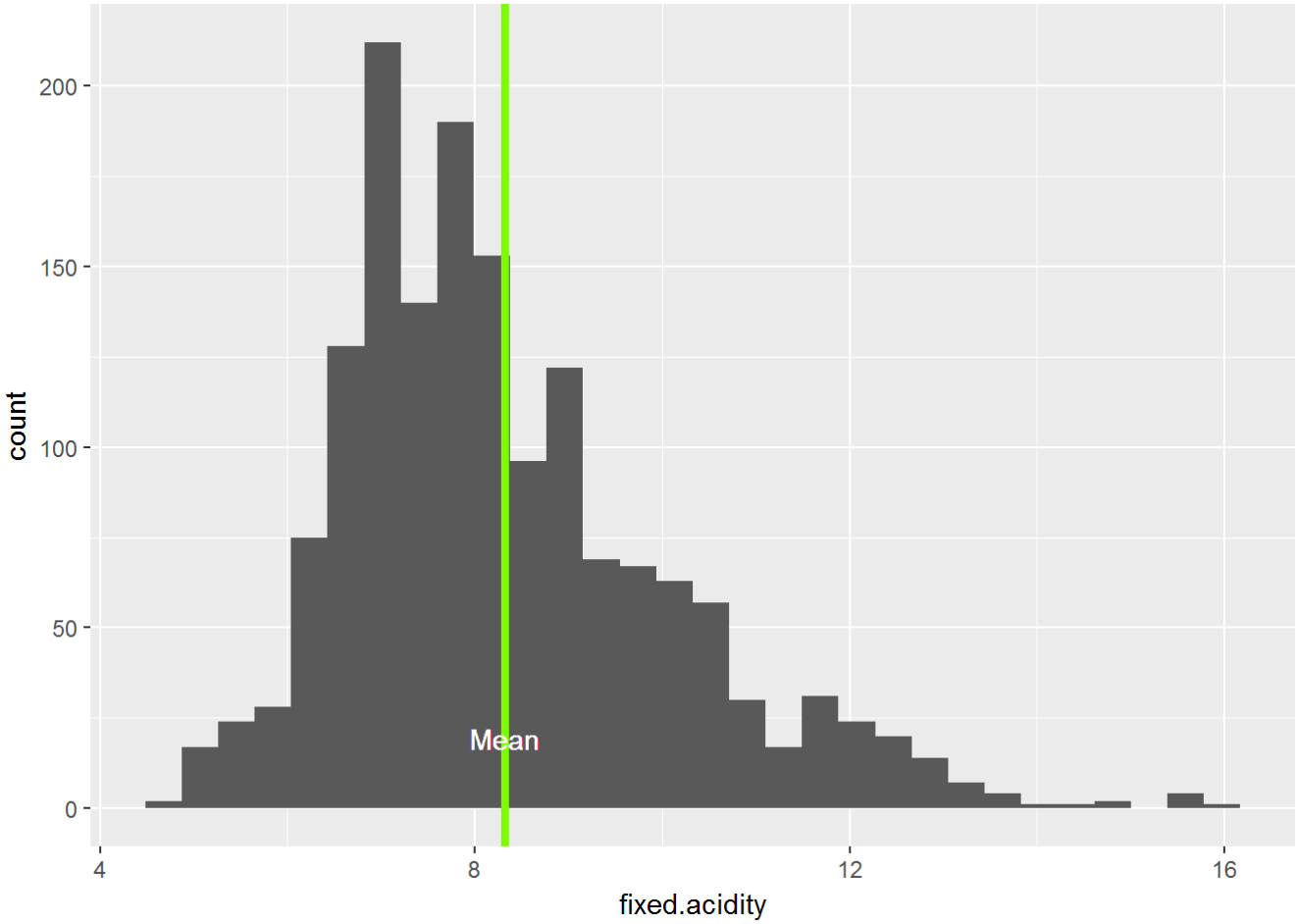
```
## [1] 0
```

There are no NA values, let's look at the histograms:



At first glance, fixed.acidity, residual.sugar, density, pH, sulphates and quality are normally distributed. It's better to create individual plots to take a closer look at them

Fixed Acidity

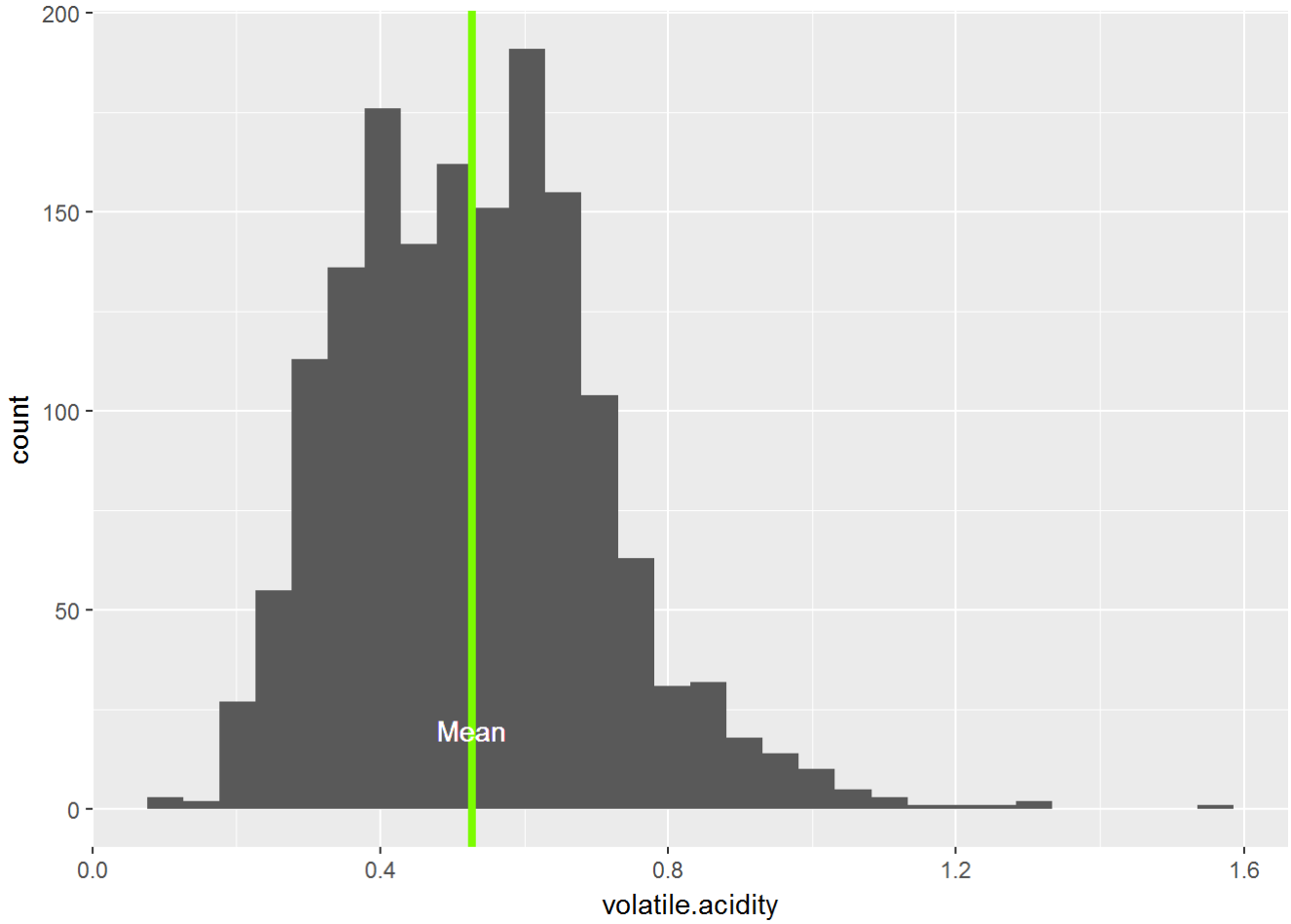


Summary statistics for this variable:

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

This variable refers to the acids that do not evaporate readily. We can see that the most of the wines in the dataset has fixed acidity between 5 and 14. The mean value is 8.32

Volatile Acidity

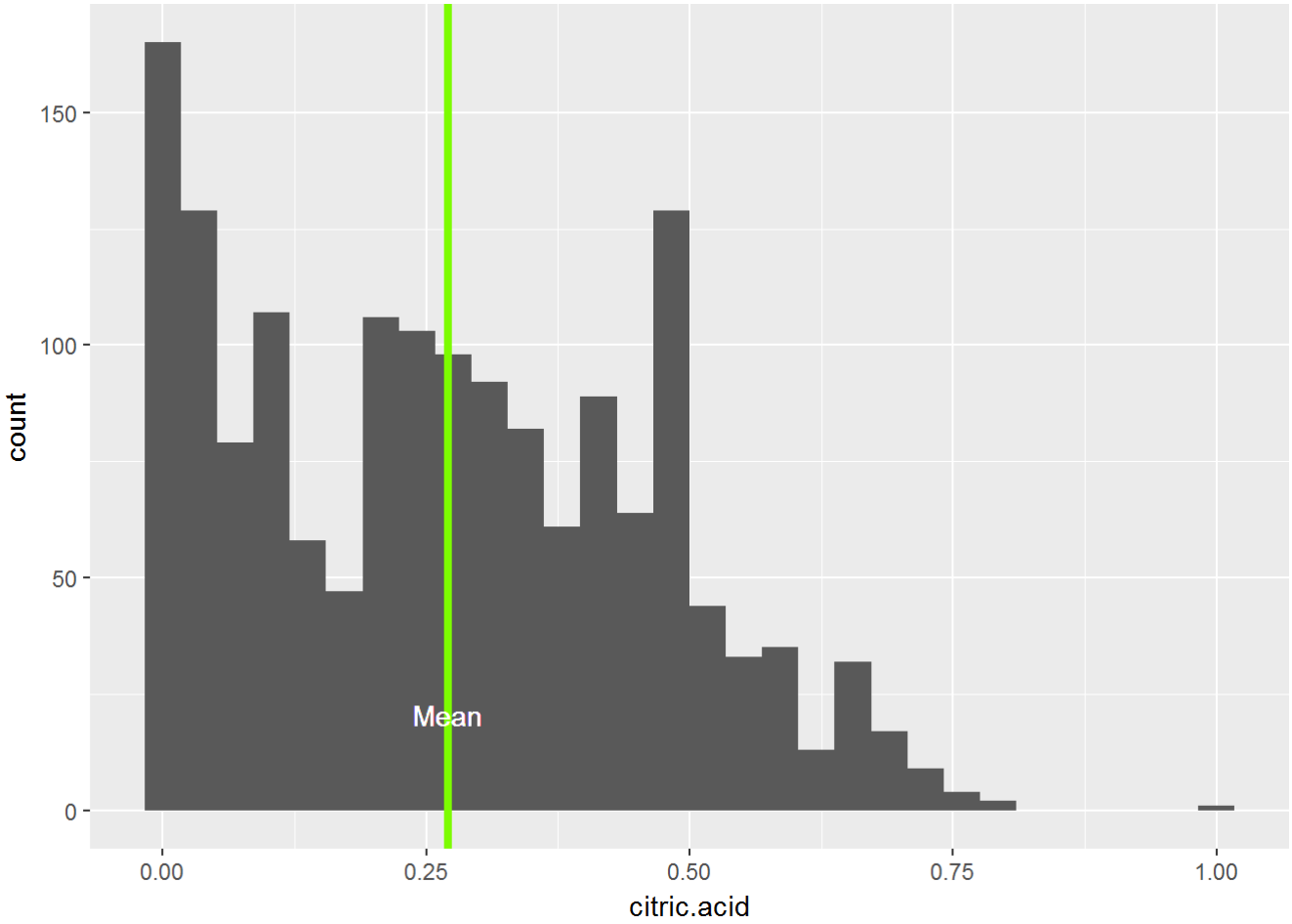


Summary statistics for this variable:

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

This variable refers to the acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste. The distribution seems to be normal with the mean value 0.5278

Citric Acid

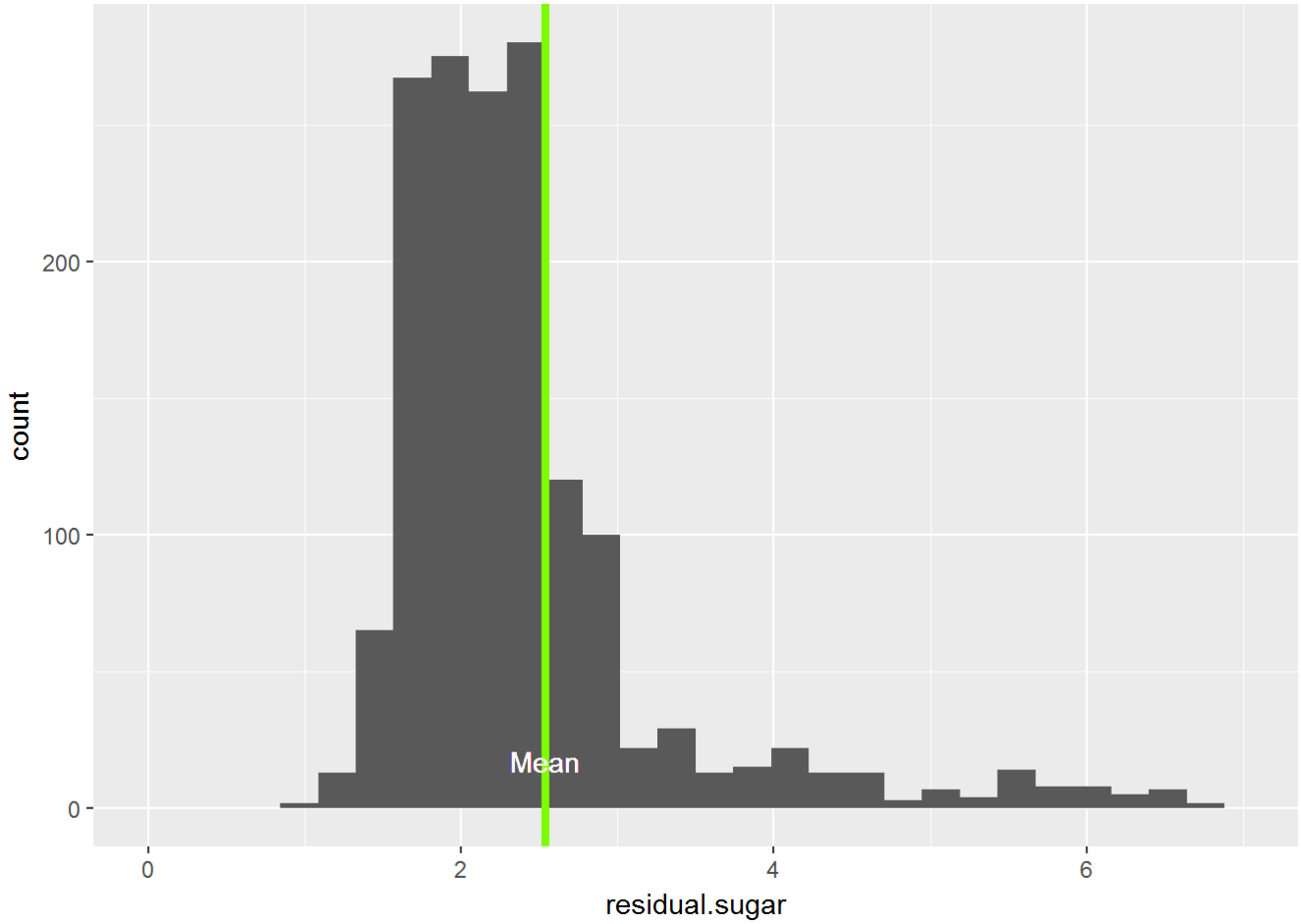


Summary statistics for this variable:

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

Citric acid is a component that can add ‘freshness’ and flavor to wines. The distribution seems to be exponential, with the mean value 0.271

Residual Sugar

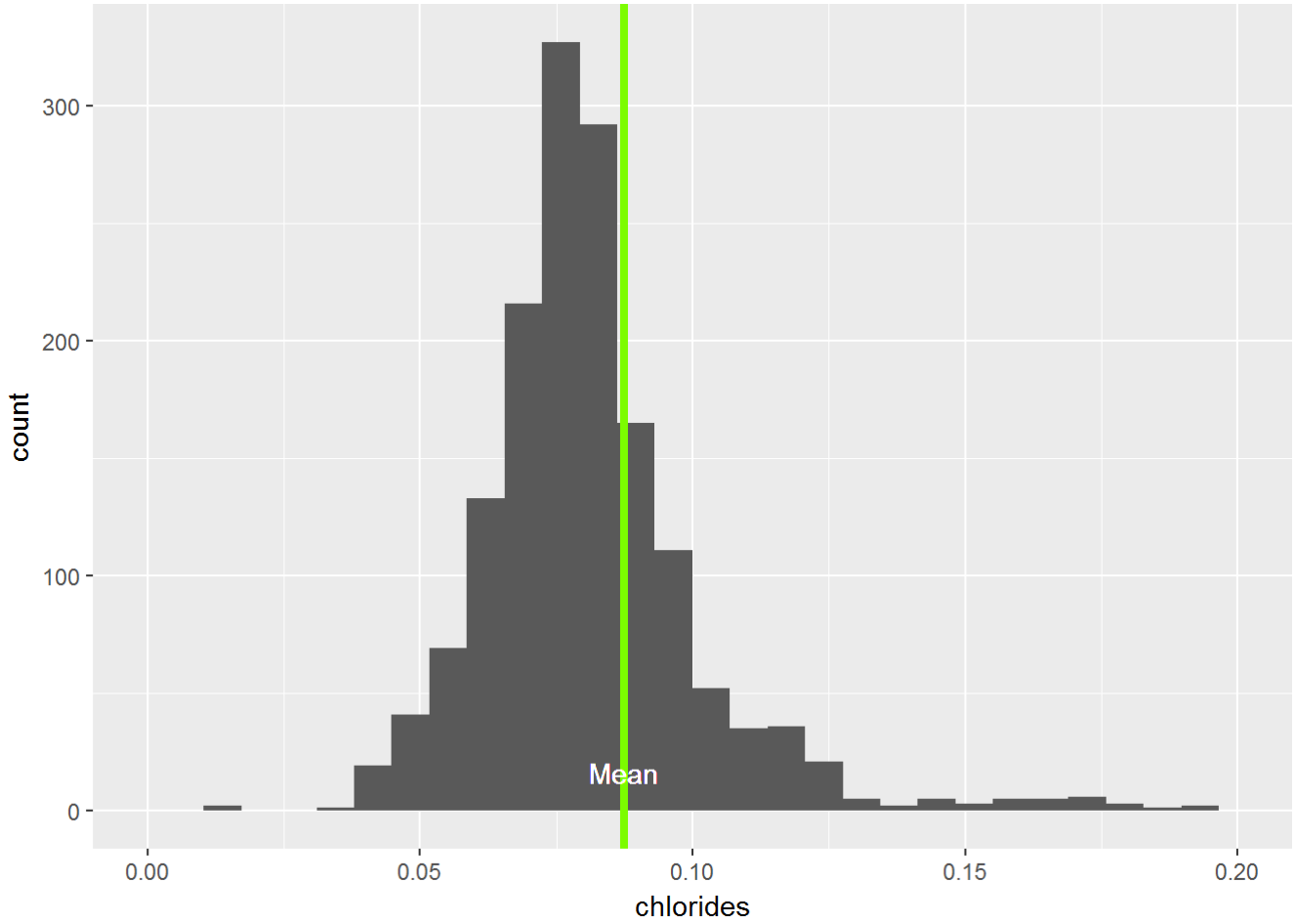


Summary statistics for this variable:

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

Residual sugar is the amount of sugar remaining after fermentation stops. Most of the values is between 0 and 6, there are several outliers that have residual sugar between 12 and 16. The original data has some outliers, which makes the plot less descriptive. Therefore I limited the x axis to avoid showing the outliers.

Chlorides

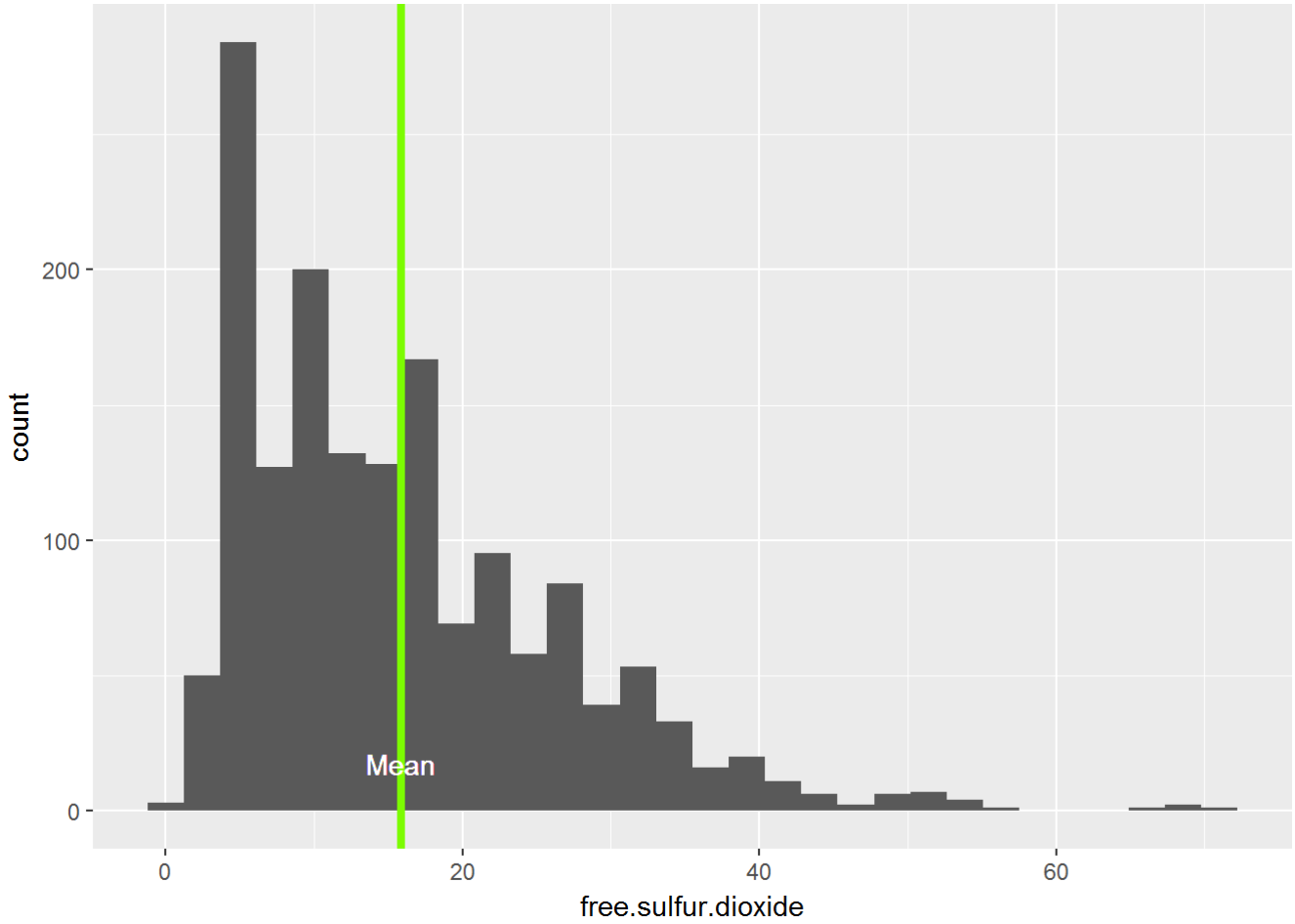


Summary statistics for this variable:

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

This means the amount of salt in the wine. Most of the samples have chlorides between 0.4 and 1.4, but the dataset has many outliers. I also limited the x axis to exclude the outliers.

Free Sulfur Dioxide

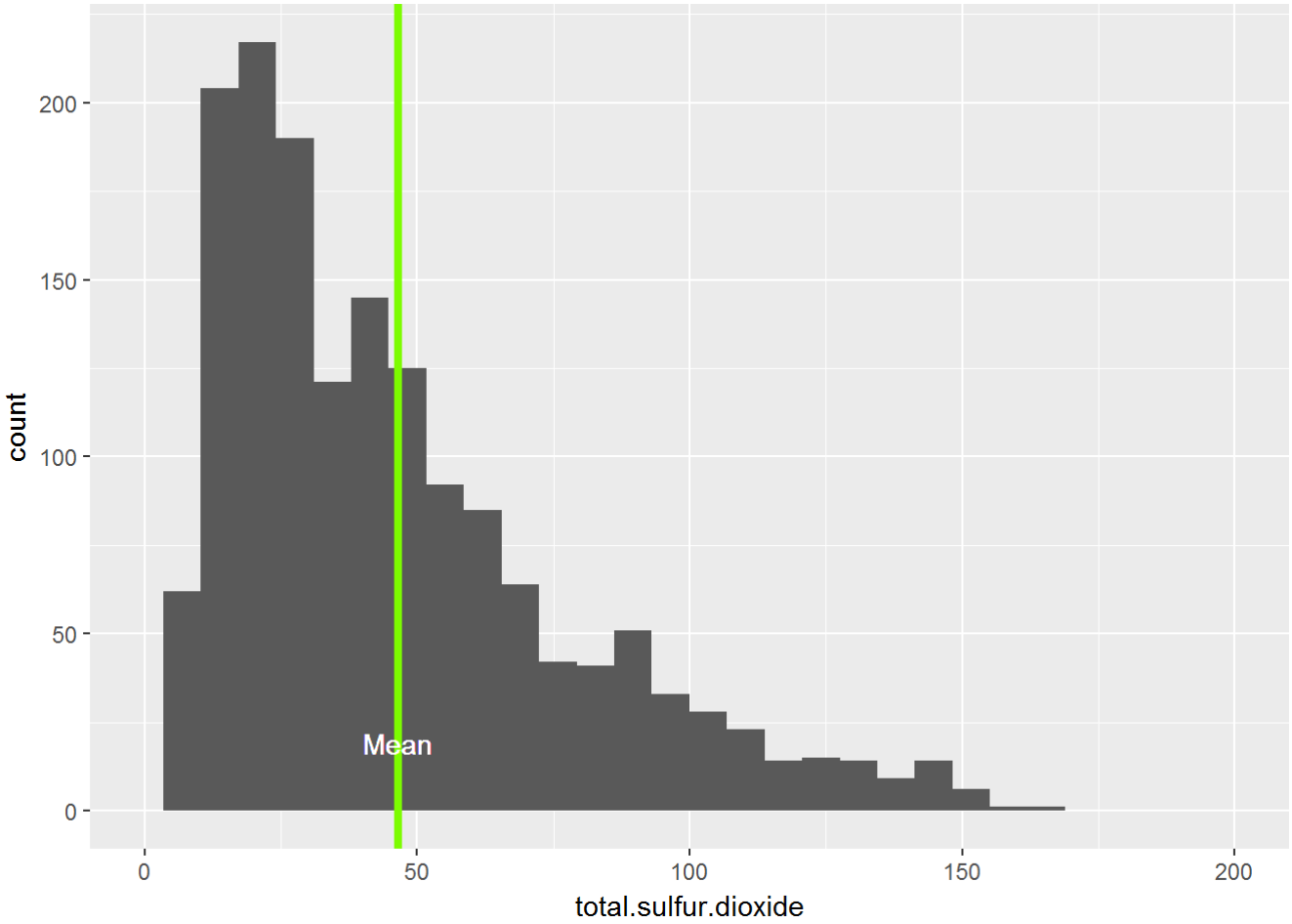


Summary statistics for this variable:

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

The free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine. Most of wine in the dataset has free sulfur dioxide in between 1 to 57 with several outliers.

Total Sulfur Dioxide

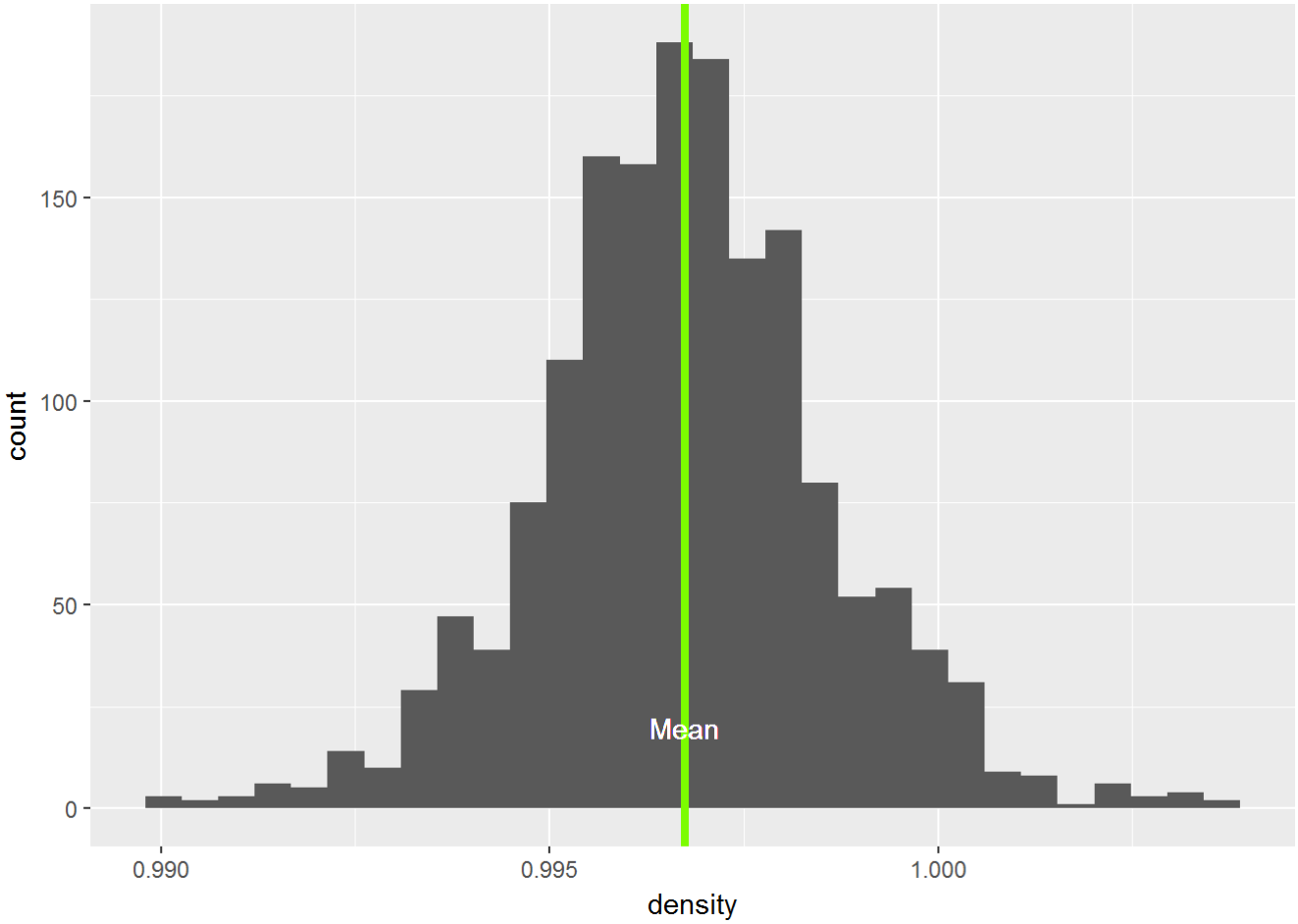


Summary statistics for this variable:

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

The Amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine. Most of the wine fall in the range 16 to 166, and it's clear from the histogram that there are more wines with less total sulfur dioxide and less wines with high value of total sulfur dioxide. I have also excluded the outliers here.

Density

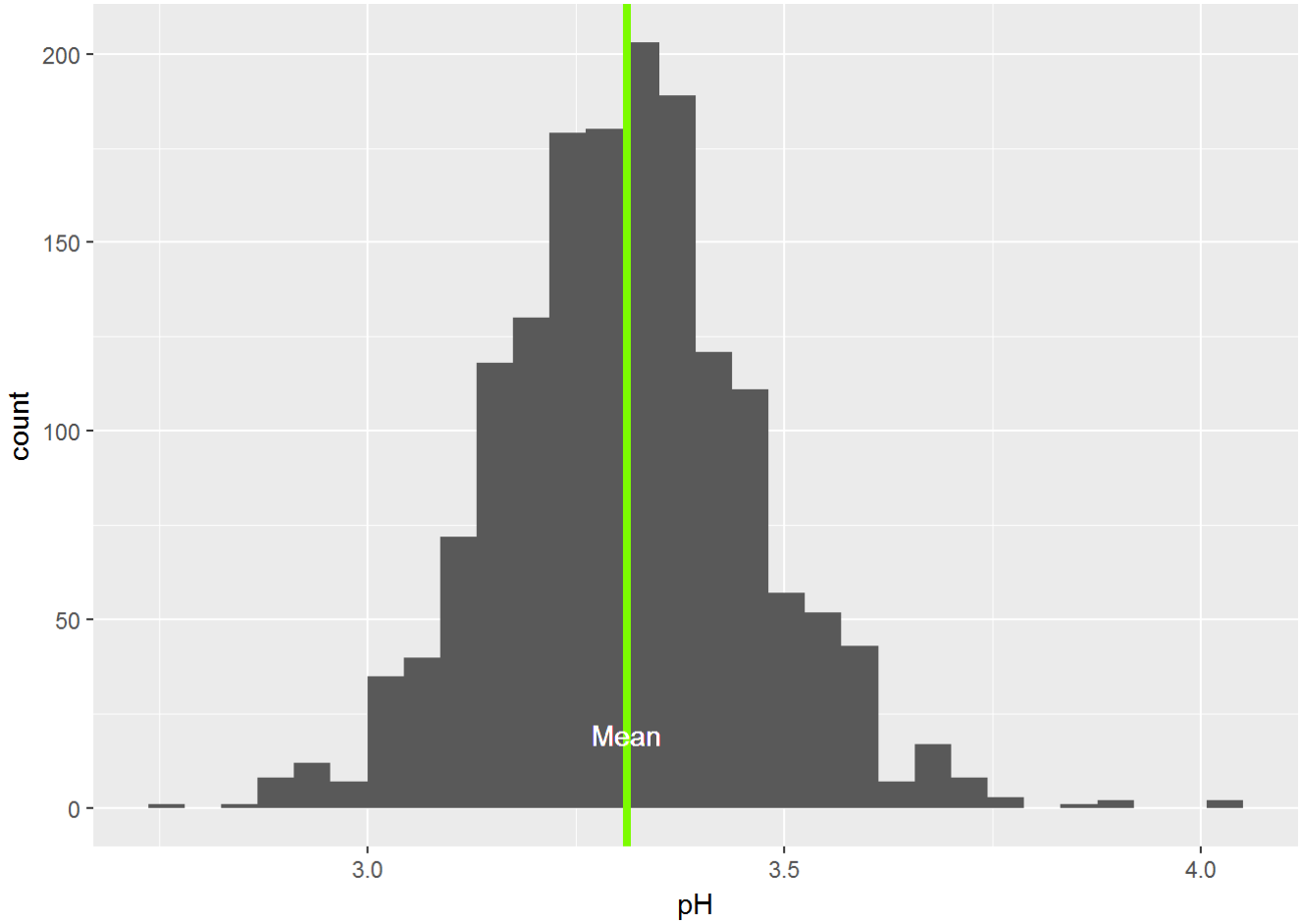


Summary statistics for this variable:

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

The density of water is close to that of water depending on the percent alcohol and sugar content. We can see from the histogram that there is not much variation in this variable.

pH

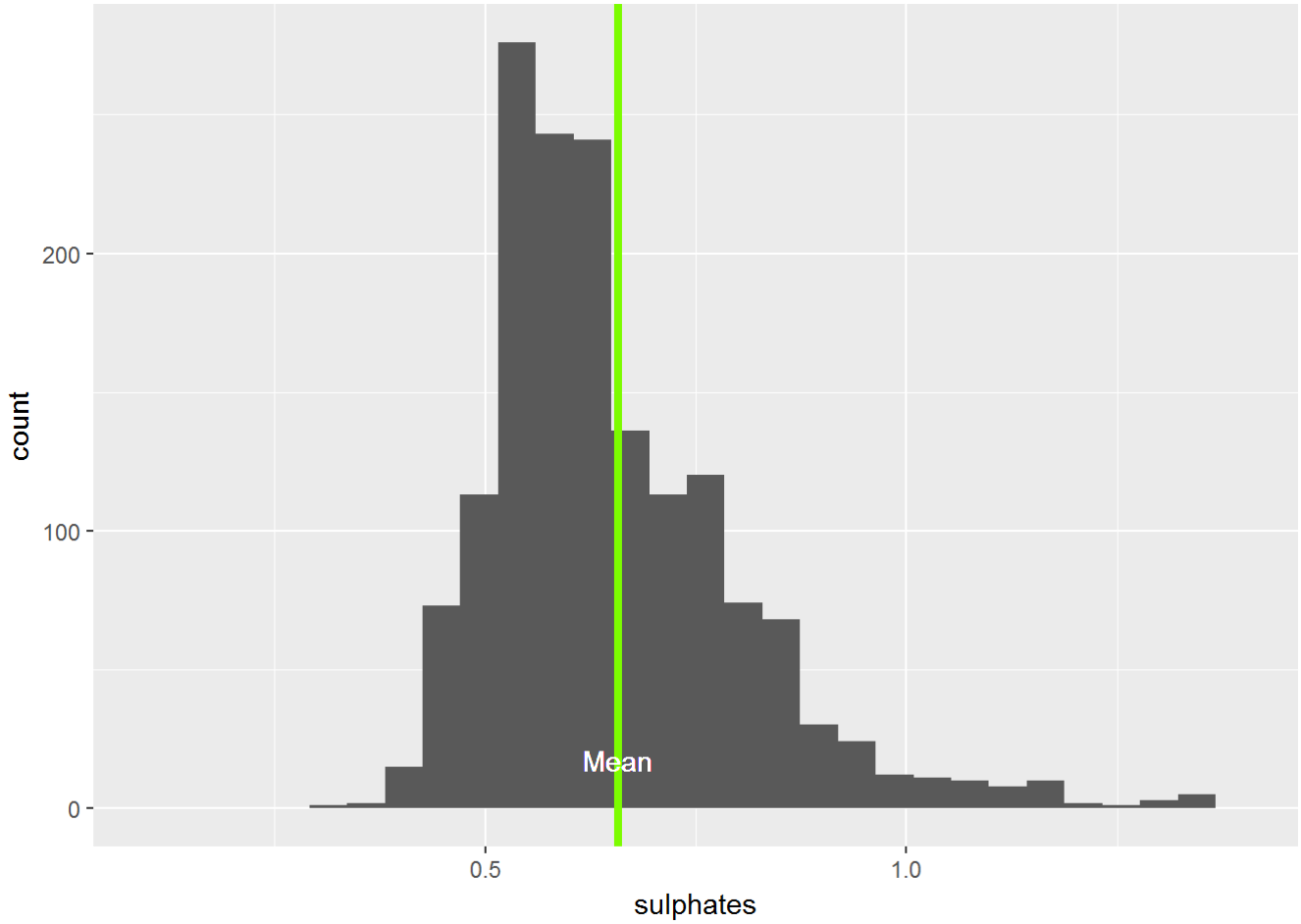


Summary statistics for this variable:

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

This variable describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale. Most wine samples are between 2.8 - 3.7

Sulphates

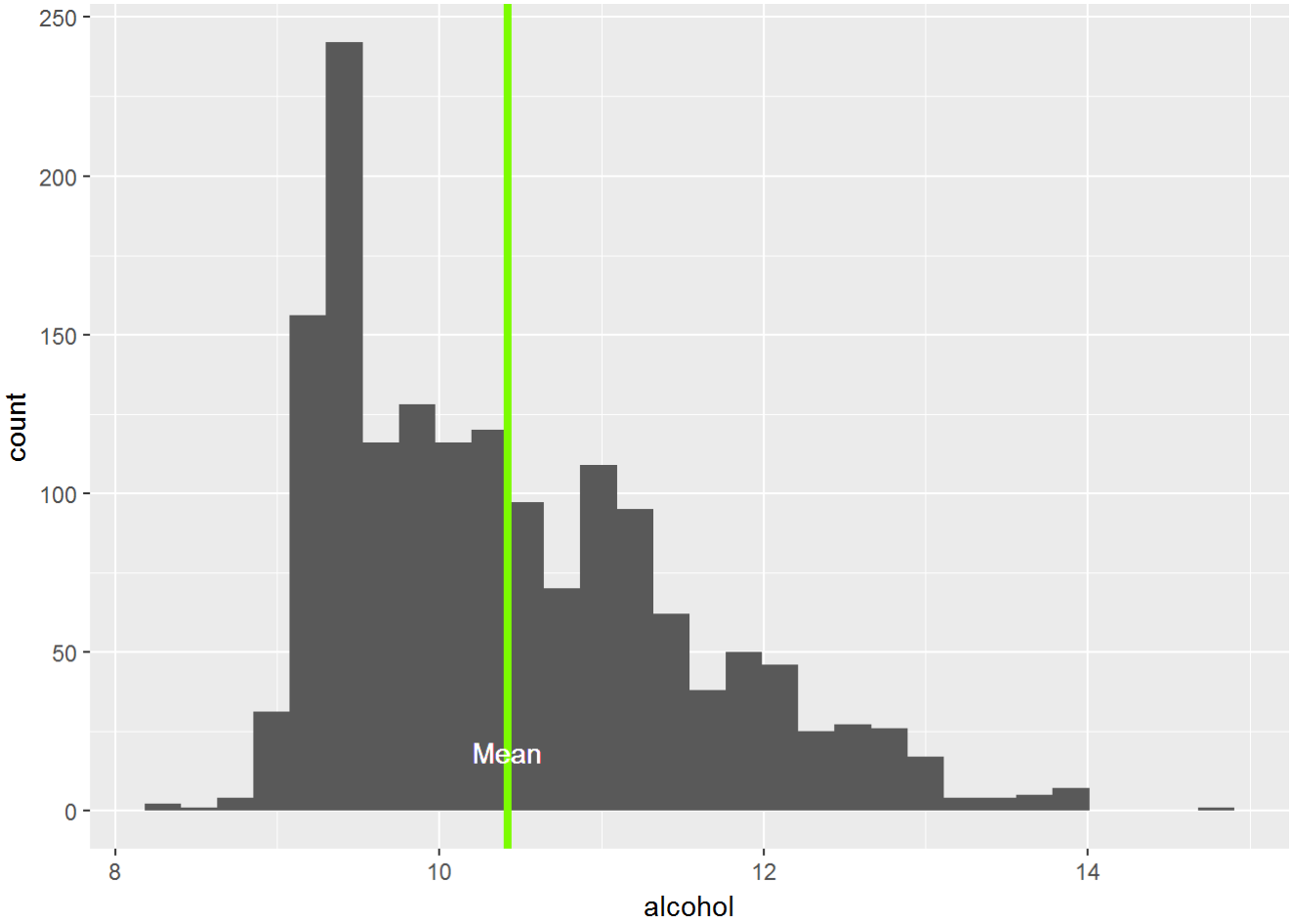


Summary statistics for this variable:

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

This describes a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant. The mean value of this variable is 0.66. I have limited the x axis to hide the outliers.

Alcohol

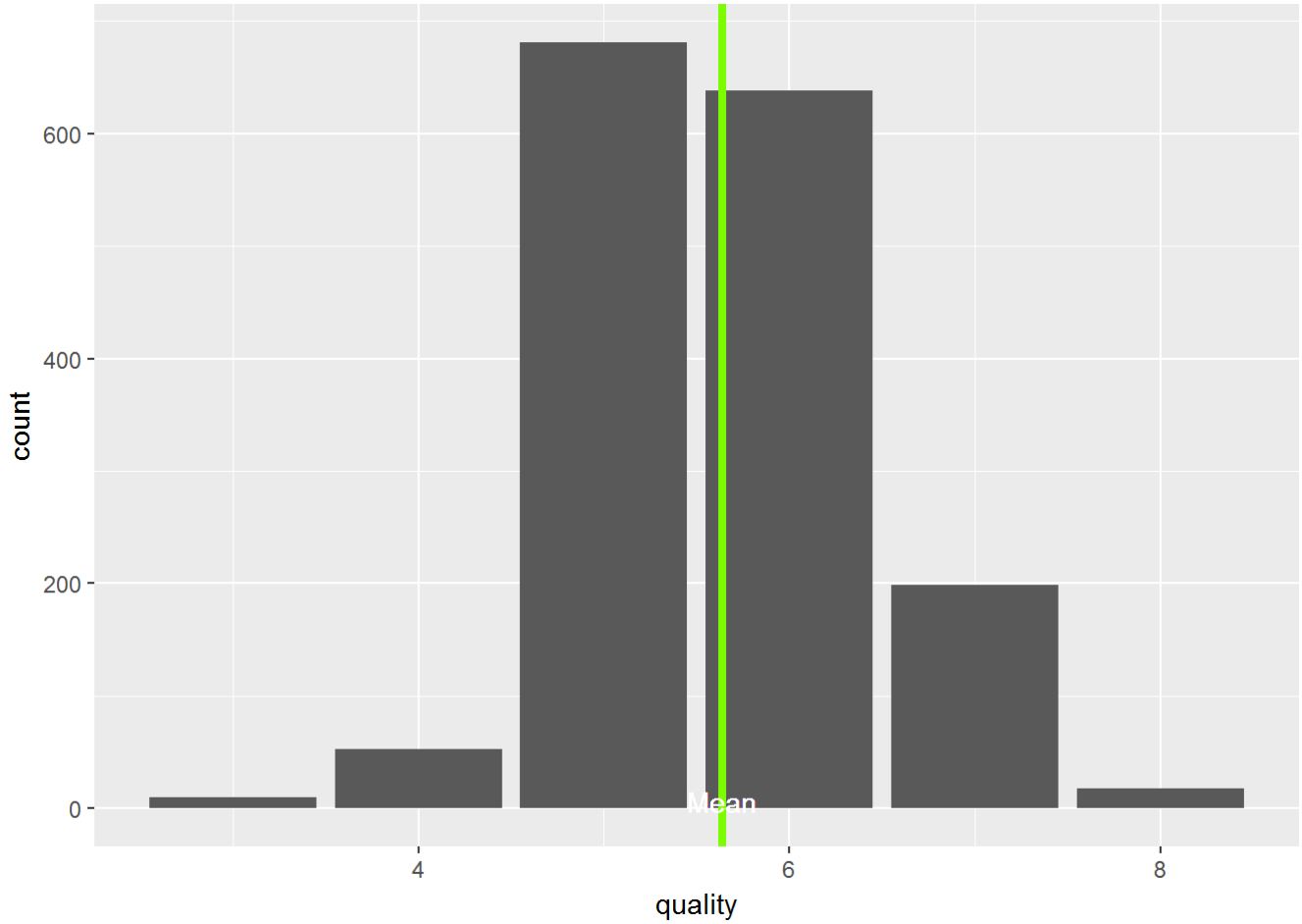


Summary statistics for this variable:

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

Most of wines have the amount of alcohol between 8.9 to 13.4 with the mean value 10.42

Quality



Summary statistics for this variable:

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

This is the most important variable in this dataset. In our data set the quality of wine ranges between 3 and 8 with the mean value 5.6

Univariate Analysis

What is the structure of your dataset?

The dataset contains 1599 observations of 12 variables (X variable was removed from the dataset because it's just an index). The only variable with the integer data type is quality, other variables are num. The dataset has no NA variables.

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

I think the main feature is the wine quality, it's very interesting to know what people think about wine samples, and how it depends on different parameters.

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

It would be great to look at the variables that everyone can easily detect by drinking a glass of wine, for example residual sugar, citric acid, chlorides and alcohol

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

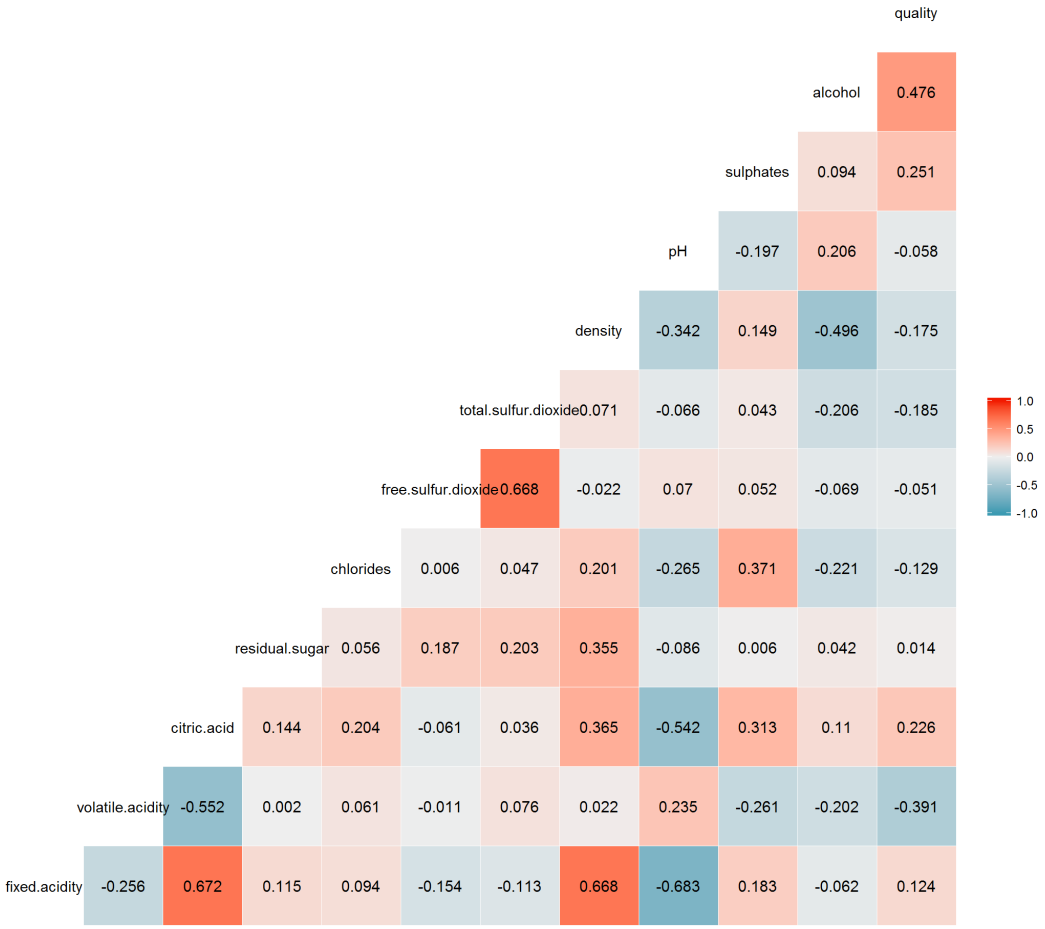
No, I didn't

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

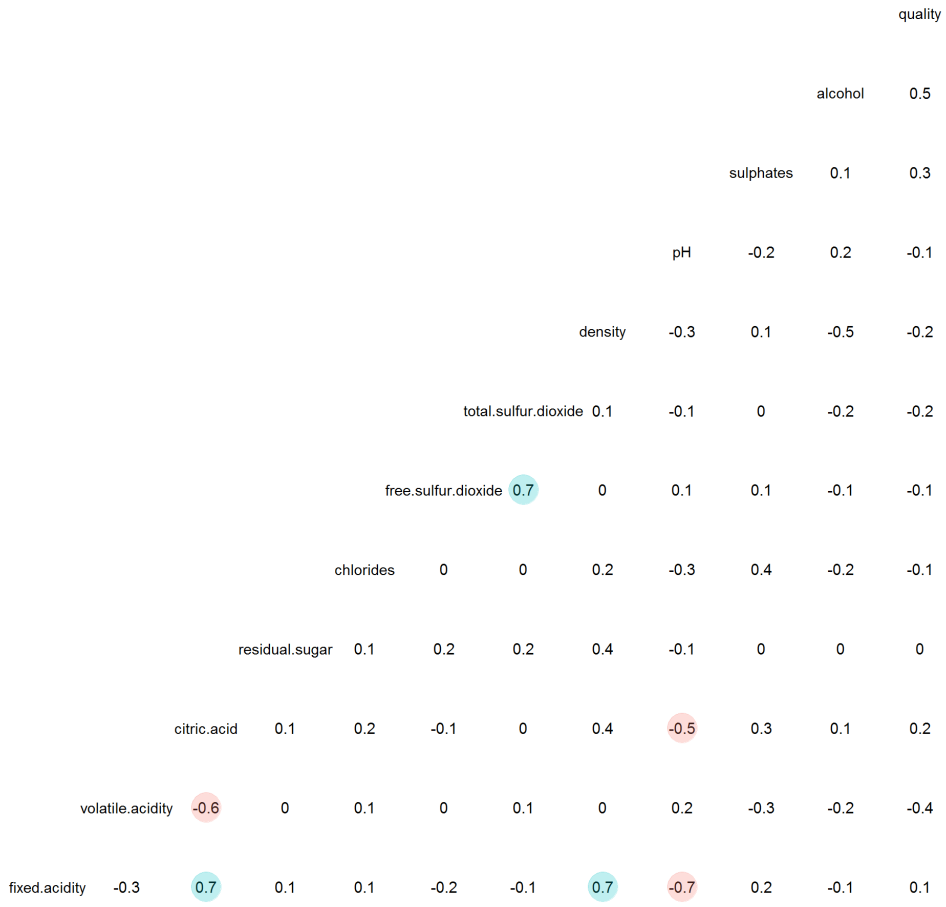
I think this dataset has no unusual distribution, every variable is distributed either normally or exponentially. And yes, I did a tidy dataset to be able to to show several similar histograms in one plot with facets.

Bivariate Plots Section

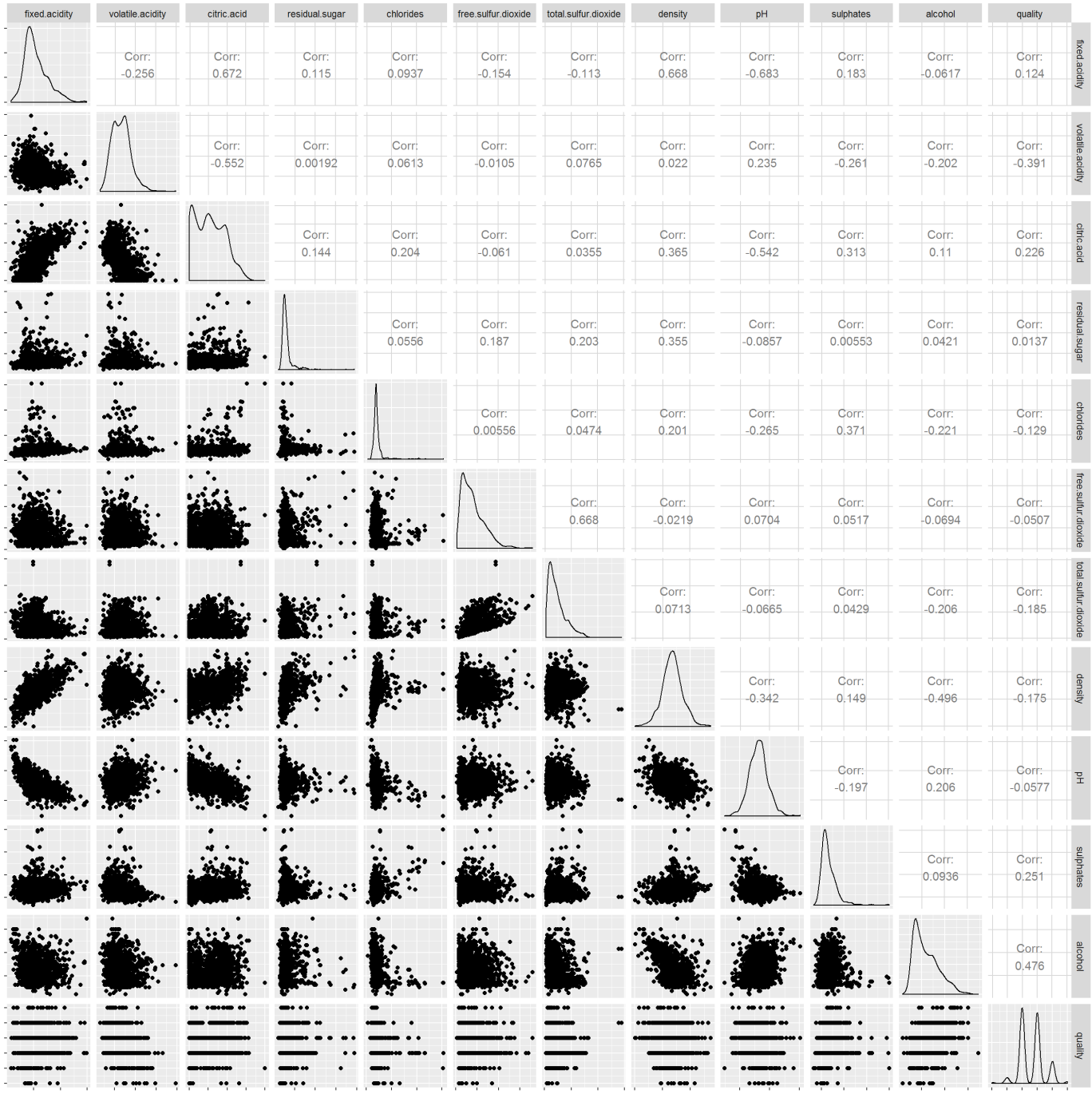
Now let's take a look at relations between our variables. The best way to do it – ggplot2 correlation matrixes. This matrix shows us correlation coefficients between the variables:



Here is a little improvement to show the most significant relations (with the absolute value of r > 0.5):



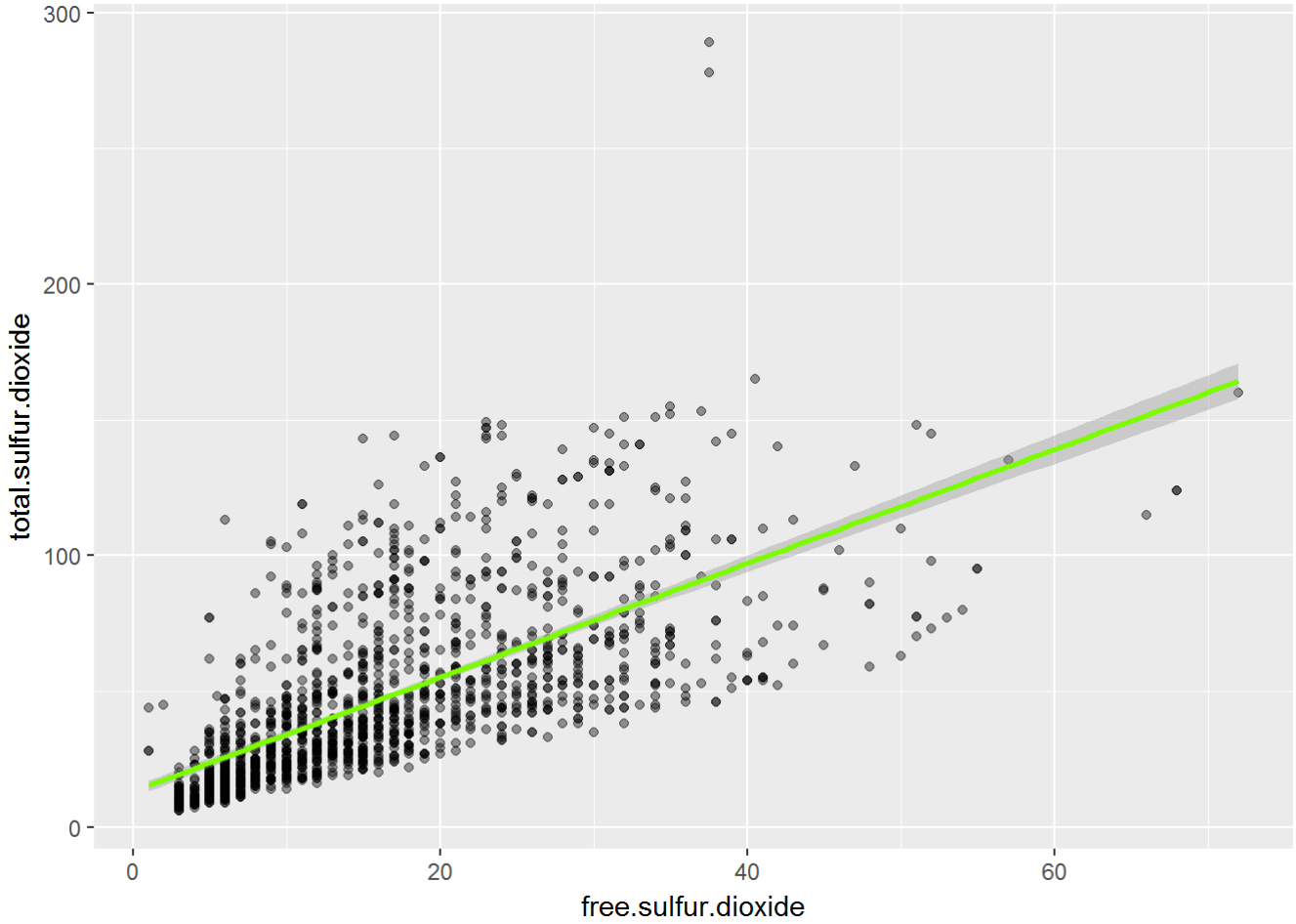
This matrix shows us how relations between the variables look like on the plot (on a series of small plots in fact)



It seems interesting to check the relations between the following variables:

- free.sulfur.dioxide and total.sulfur.dioxide
- citric.acid and pH
- volatile.acidity and citric.acid
- fixed.acidity and citric.acid
- fixed.acidity and density
- fixed.acidity and pH
- alcohol and quality
- citric acid and quality

Free Sulfur Dioxide / Total Sulfur Dioxide

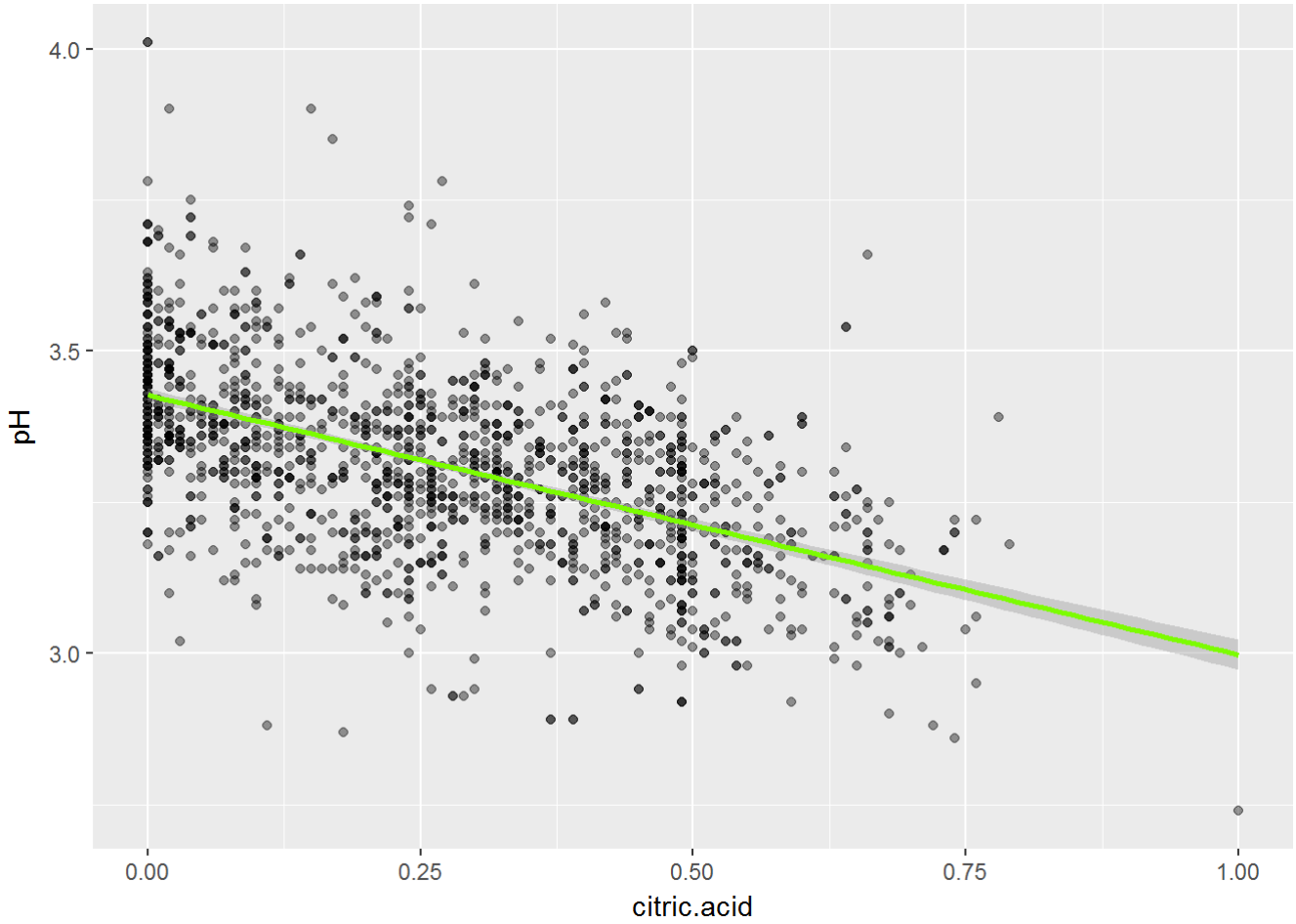


Correlation coefficient:

```
## [1] 0.6676665
```

We can see a strong positive correlation between Free Sulfur Dioxide and Total Sulfur Dioxide – if one increases other also increases. This relation is quite logical.

Citric Acid / pH

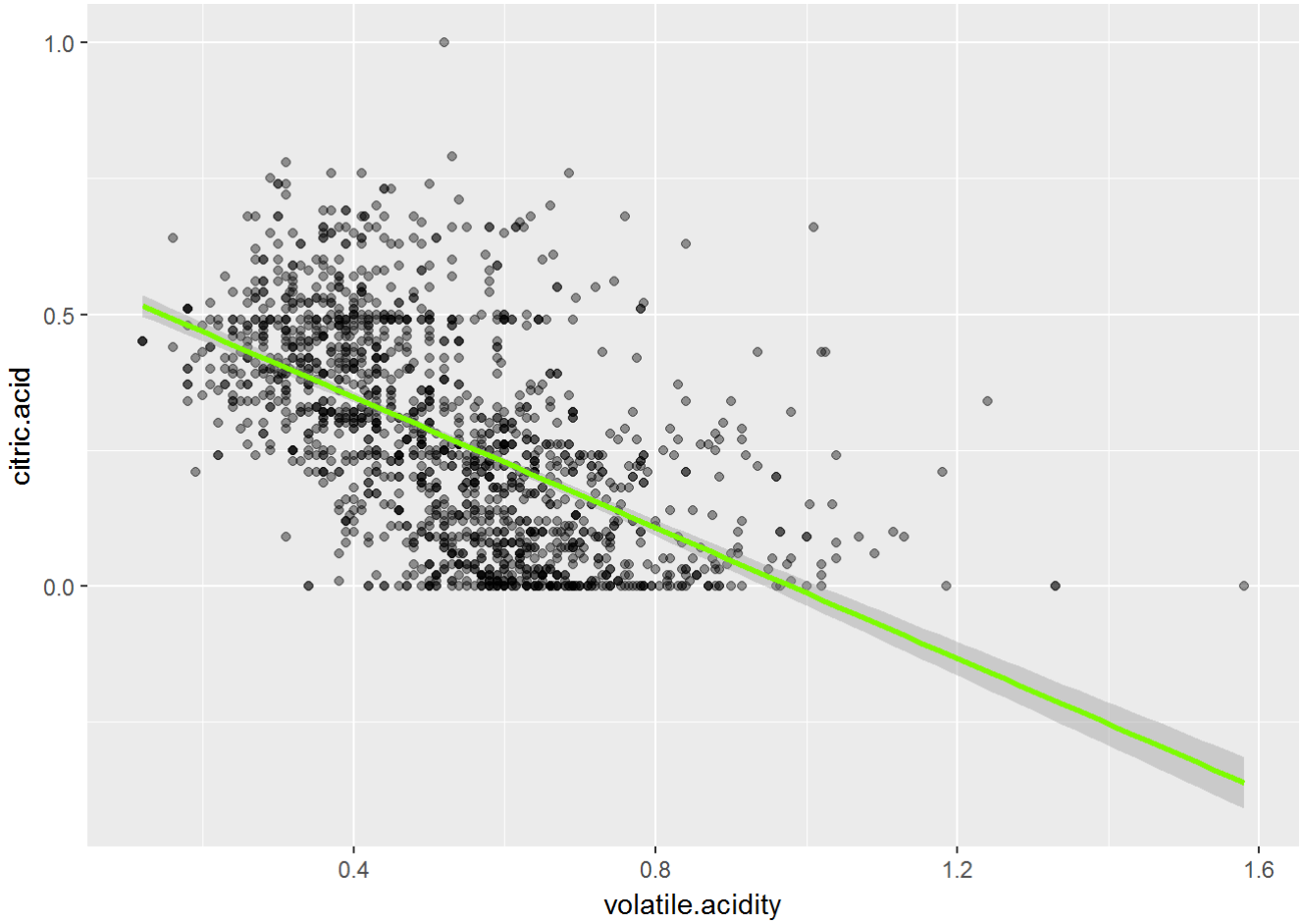


Correlation coefficient:

```
## [1] -0.5419041
```

These two variables also show a strong correlation, which is expected: the more acid – the more acidic solution – the lower pH.

Volatile Acidity / Citric Acid

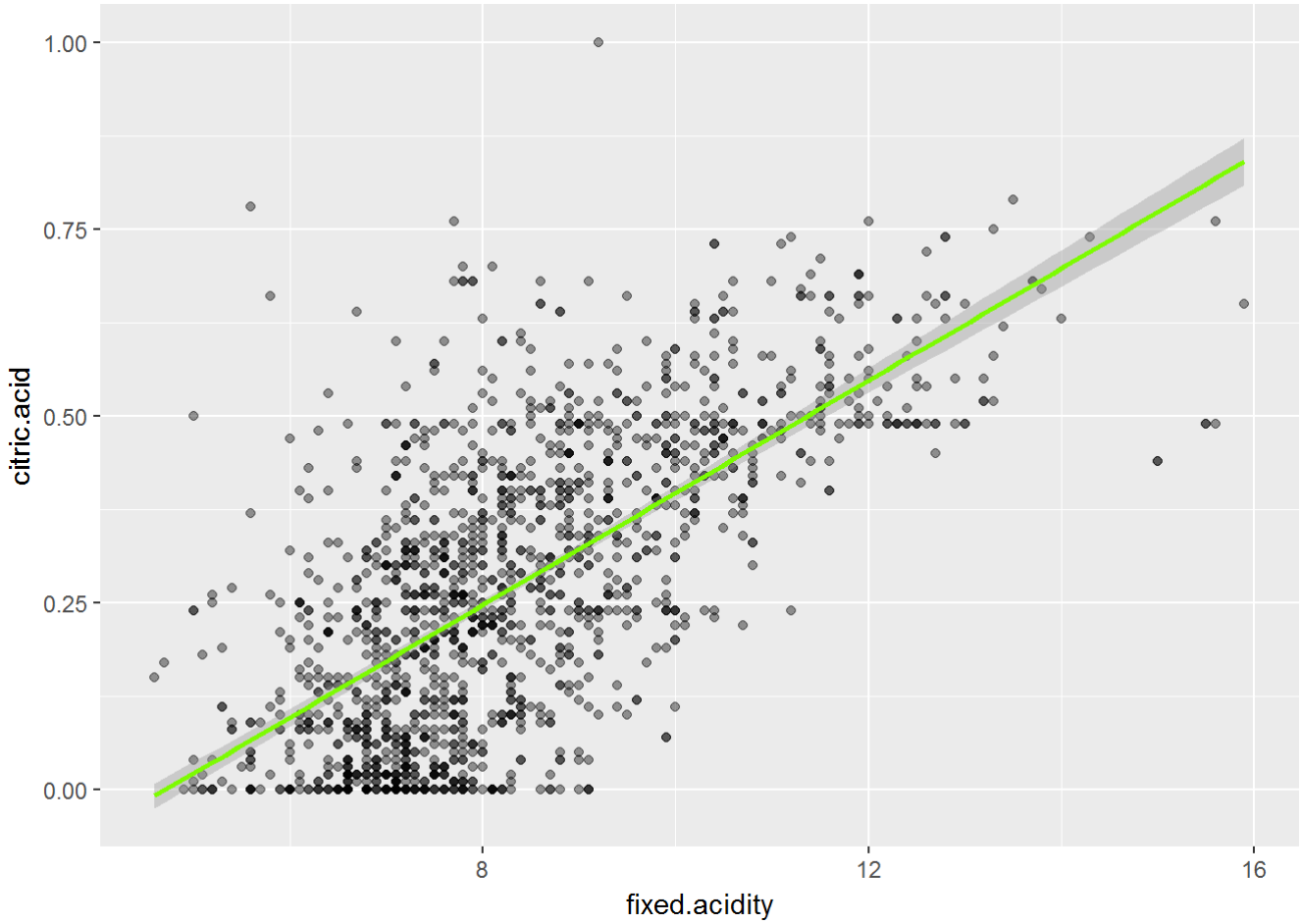


Correlation coefficient:

```
## [1] -0.5524957
```

These two variables are negatively correlated, we can see it on the plot. It's hard to explain this correlation withoud good knowledges of chemistry.

Fixed Acidity / Citric Acid

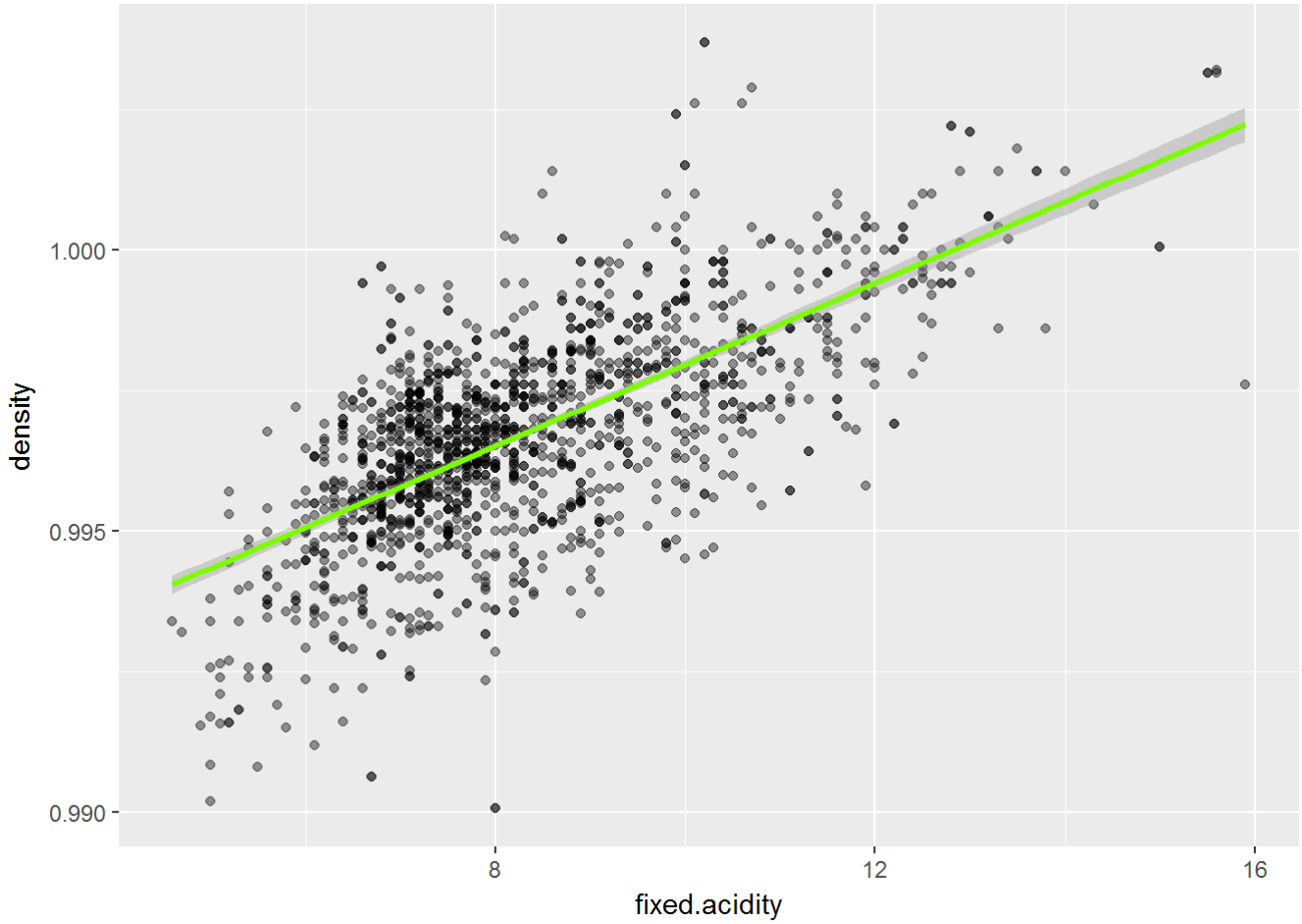


Correlation coefficient:

```
## [1] 0.6717034
```

This positive correlation, in contrast to the previous one, is more understandable: the more citric acid – the higher acidity.

Fixed Acidity / Density

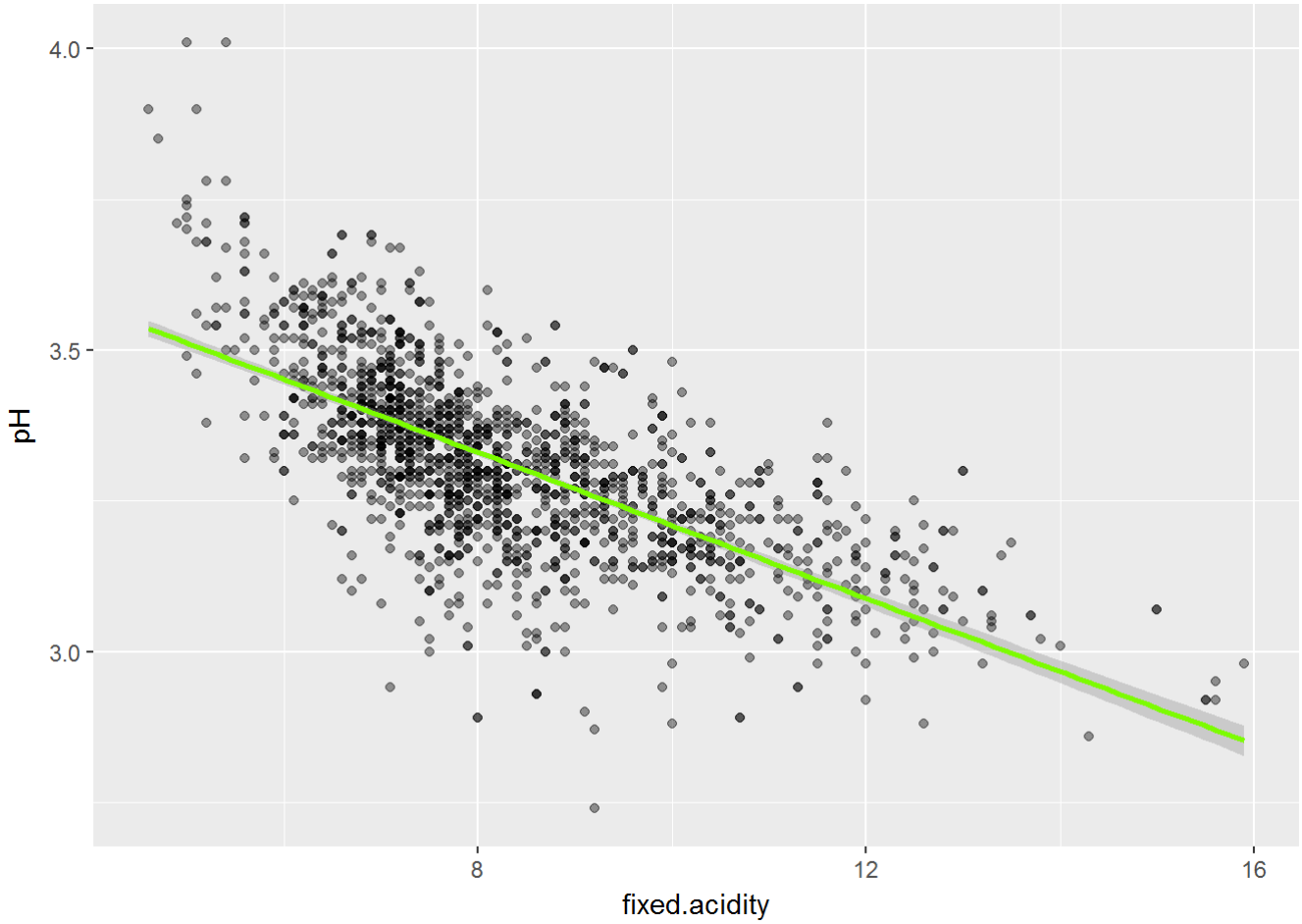


Correlation coefficient:

```
## [1] 0.6680473
```

We can see on the plot that this is a strong positive correlation. The higher fixed acidity means the higher density.

Fixed Acidity / pH



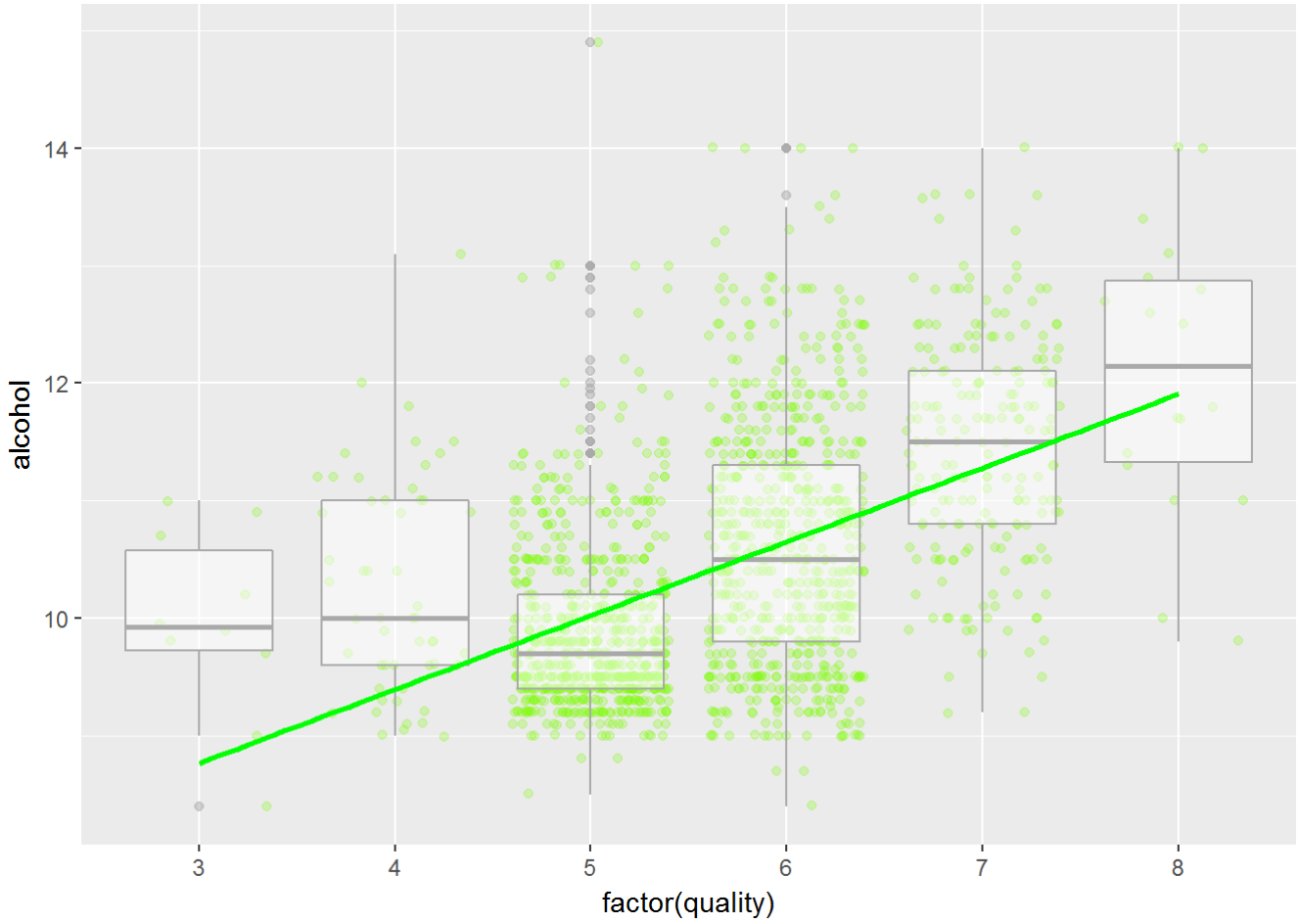
Correlation coefficient:

```
## [1] -0.6829782
```

We can see that fixed acidity among the citric acidity affects the pH level. We should keep in mind that fixed acidity and citric acidity are also strongly correlated.

From the correlation matrix we could see that the quality mostly depends on the alcohol and Volatile Acidity. Let's find out how

Alcohol / Quality

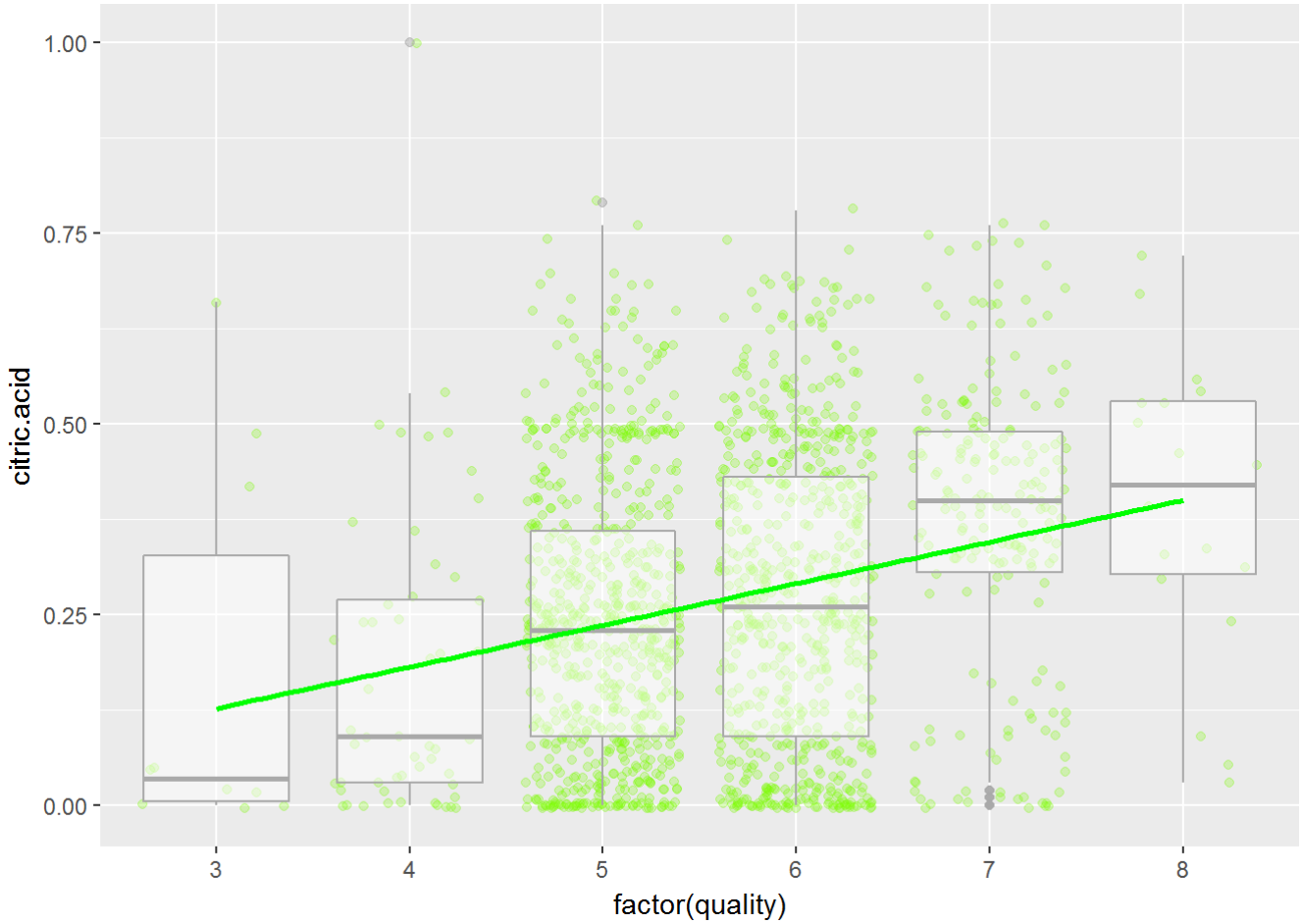


Correlation coefficient:

```
## [1] 0.4761663
```

We can see that the quality increases as the amount of alcohol increases, which is not surprising

Citric Acid / Quality



Correlation coefficient:

```
## [1] -0.3905578
```

Here we can see that more acidic wine samples has the better quality, which is the same for coffee, where an acidic flavour is considered to be better.

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?

We found the strongest correlations between the quality and alcohol, quality and citric acidity. It was clear on the plots that the wine sample with the better quality contain more alcohol and has higher citric acidity.

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

The relationships found in the dataset refreshes the knowledge of chemistry: more citric acid and higher fixed acidity decreases pH, which makes the solution more acidic and less alkaline.

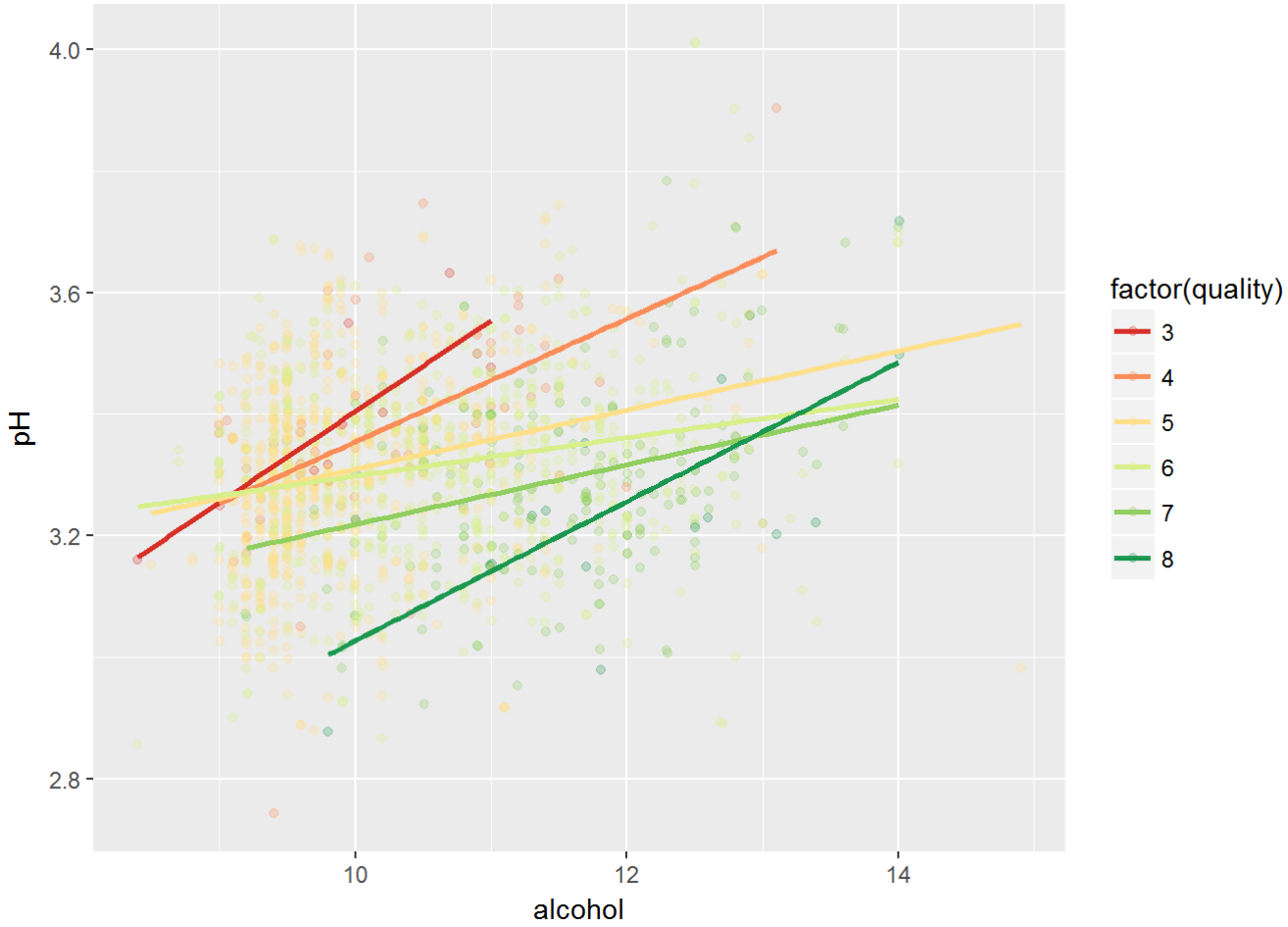
What was the strongest relationship you found?

The strongest relationship in the dataset is between fixed acidity and pH (r = -0.6829782). There is also a more meaningful relationship between alcohol and quality (r = 0.4761663)

Multivariate Plots Section

In this section our task is to find out how the quality changes with the combination of two variables.

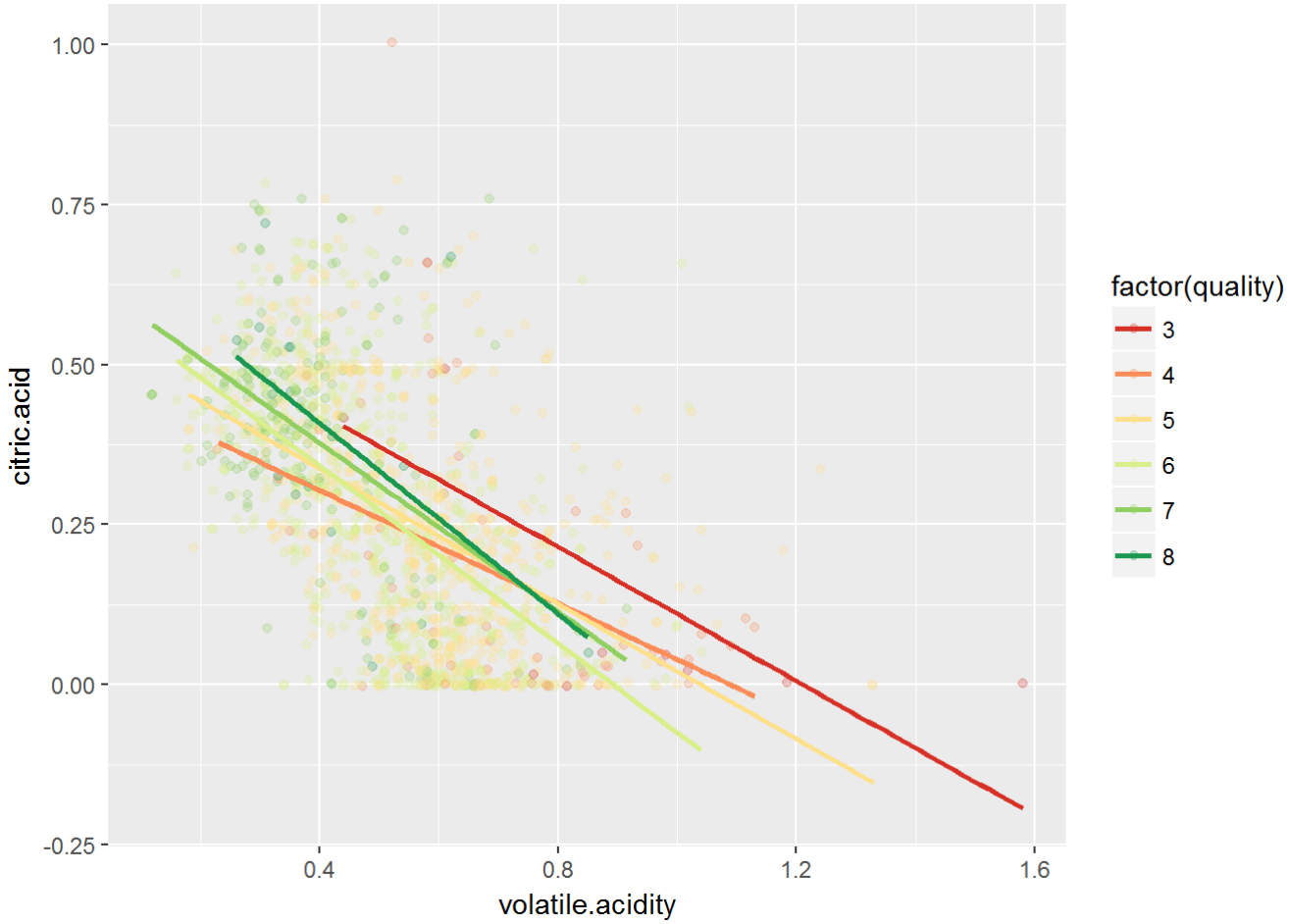
Alcohol / pH / Quality



We can see on the plot that lines with better wind quality are located primarily on the top left side, which means high pH and low alcohol make the wine of bad quality.

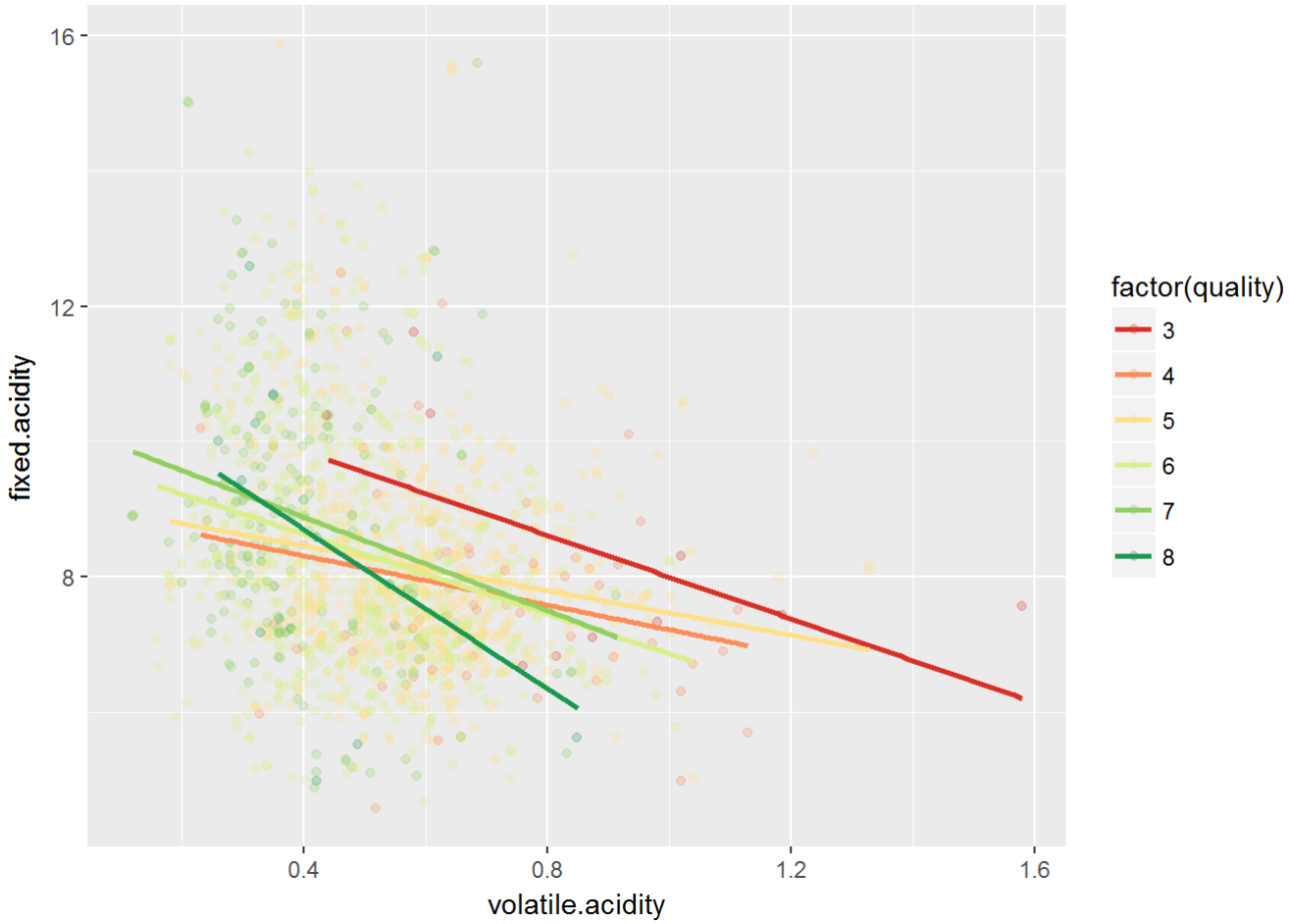
Let's see how both Volatile Acidity and Citric Acid impact the quality

Volatile Acidity / Citric Acid / Quality



It's clear that higher volatile acidity and lower citric acidity makes the wine bad.

Volatile Acidity / Fixed Acidity / Quality



The increase in volatile acidity with the increase in the fixed acidity result in lower wine quality.

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?

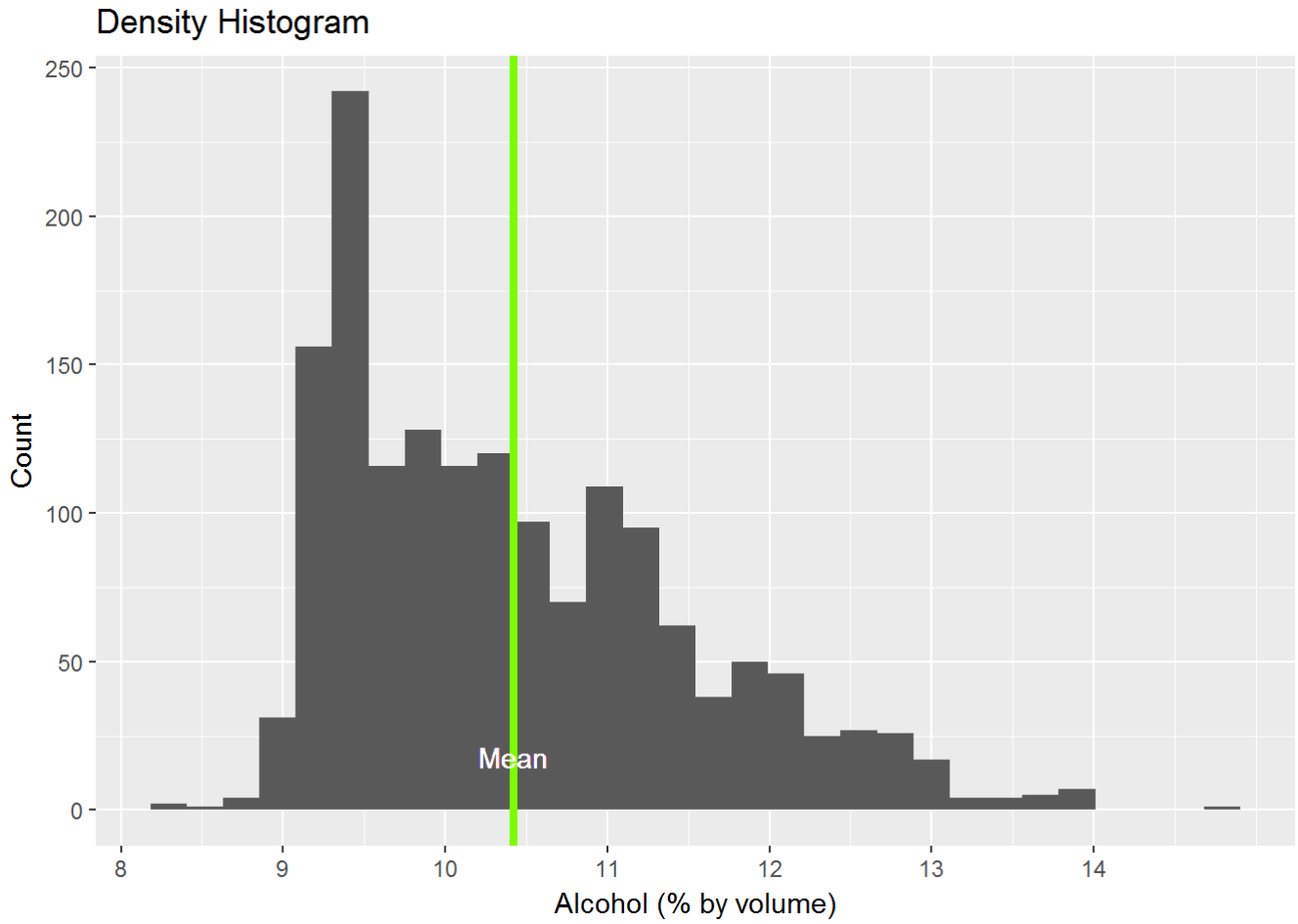
Citric acidity, volatile acidity and fixed acidity has a bad impact on the wine quality due to the fact that they are all connected (correlated in pairs).

Were there any interesting or surprising interactions between features?

The relationship between pH, alcohol and quality is interesting, but not so surprising (we already saw the relationships in pairs alcohol/quality and ph/quality in the previous section)

Final Plots and Summary

Plot One



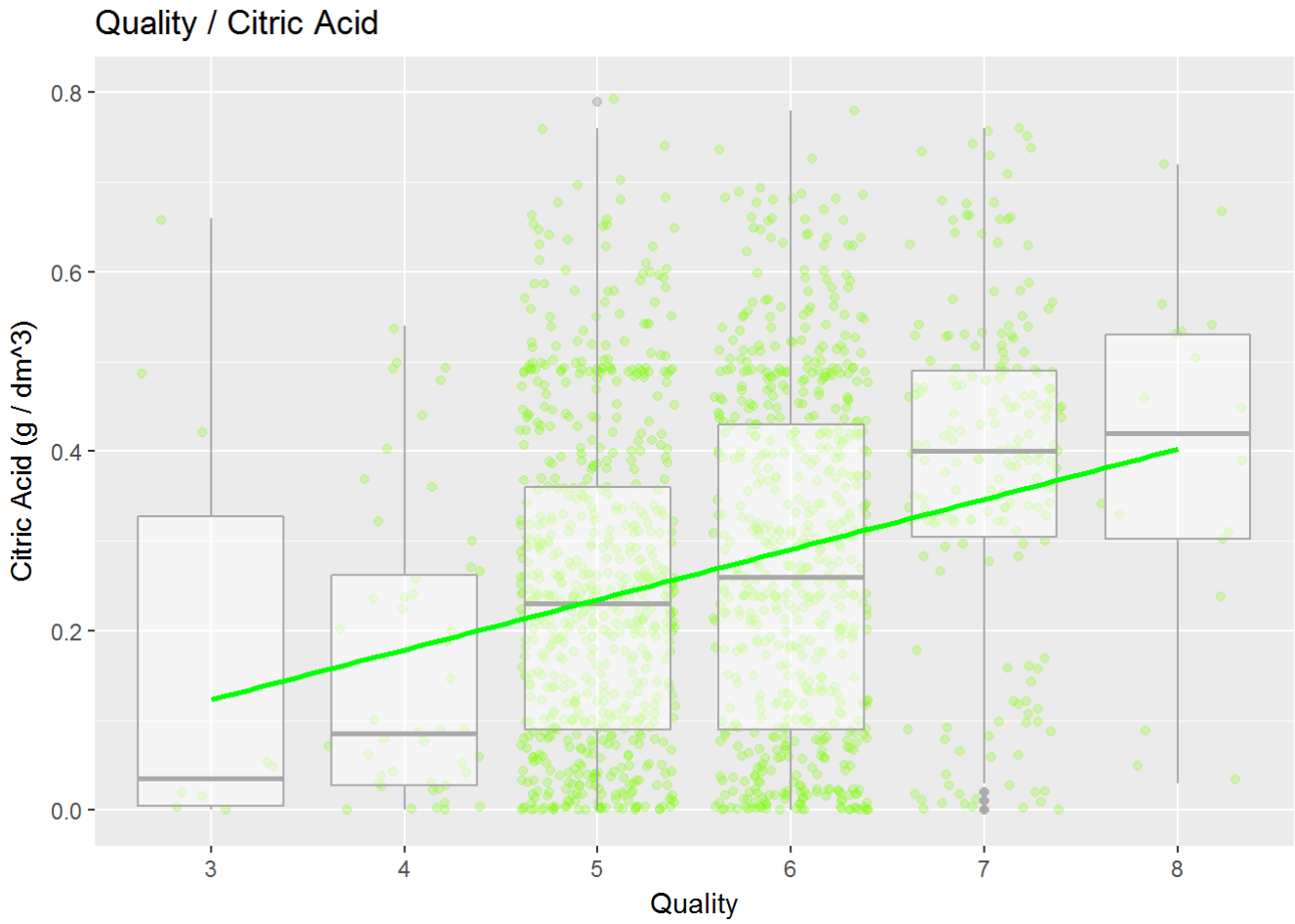
Description One

This is a very important plot showing the disribution of alcohol in the wine samples. The most of samples contain from 9% to 13% of alcohol

Summary statistics for Alcohol:

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

Plot Two



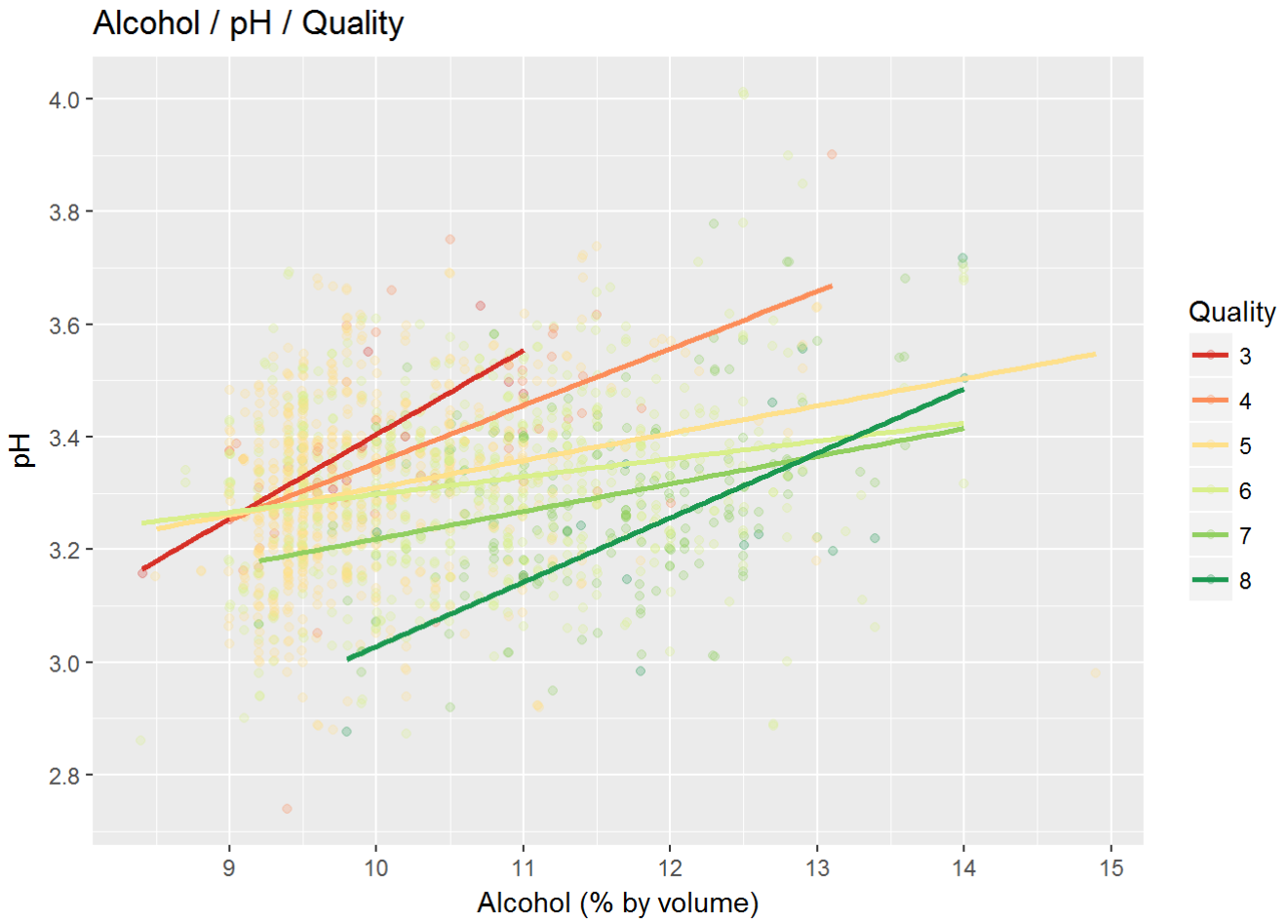
Description Two

I didn't choose the plot alcohol vs quality because it wasn't as surprising as this plot, showing that higher acidity has a posive impact on the wine quality.

Correlation coefficient for these variables equals:

```
## [1] 0.2263725
```

Plot Three



Description Three

This plot shows that both alcohol and acidity impact the quality of the wine samples. All three plots are important for our analysis, because they shows us the distribution of a single variable, the relationships between two and three variables, and most of all show some interesting insights in our data.

Reflection

In this analysis I used the dataset of red wine samples. The initial task was to find interesting trends and patterns in the data.

As a first step I took a look into the dataset by calling a str() function. It showed me that the dataset has 1599 observations of 13 variables. After looking at the structure I decided to remove the X variable, because it was just an index variable.

To conduct the univariate analysis, before making individual plots I plotted the distributions of all variables using facets.

Drawing individual plots shows the distribution of every variable in details.

After that I conducted the bivariate analysis, that allows to peek into relationships between two variables. I made correlation matrixes, that showed me some valuable information about the relationships between the variables in the dataset. In this part of the analysis I found some interesting patterns, for example correlations between citric acid and pH, fixed acidity and pH, fixed acidity and density, citric acid and quality, alcohol and quality.

Next I conducted the multivariate analysis, that showed deeper insights, for example alcohol vs pH vs quality.

During this analysis I learned several new things (such as facets and correlation matrixes) and learned some new facts about red wine.

The most important conclusions are:

- Wine samples with more alcohol are considered better. I don't know why, but maybe people prefer stronger beverages.
- Higher acidity means better quality. People prefer more acidic coffee flavour too. But I personally prefer more bitter flavor.
- Lower pH means higher acidity (school chemistry program refreshments)

For me it was absolutely easy to code and read the documentation, but the communication faze was not that easy, because every time I write a report I think about how obvious what I'm describing is. However, I understand the importance of descriptions and explanations and do my best to thoroughly write it.

If I had more time I would make the plots interactive (using ggvis), then for example the univariate plots would become one uniform plot with a dropdown menu to choose a variable.

I am curious to see more variables in the dataset. For example, country of origin, average summer temperature and price. Then we could answer interesting questions such as "which country has better wine?" or "does higher price mean higher quality?"

References

- https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityInfo.txt (https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityInfo.txt)
- http://stackoverflow.com/questions/14622421/how-to-change-legend-title-in-ggplot-density (http://stackoverflow.com/questions/14622421/how-to-change-legend-title-in-ggplot-density)
- https://www.r-bloggers.com/multiple-regression-lines-in-ggpairs/ (https://www.r-bloggers.com/multiple-regression-lines-in-ggpairs/)
- https://yihui.name/knitr/options/ (https://yihui.name/knitr/options/)
- http://docs.ggplot2.org/current/scale_brewer.html (http://docs.ggplot2.org/current/scale_brewer.html)