Red Wines Analysis by Silvia Onofrei

The dataset contains information about red “Vinho Verde”" wine samples, from the north of Portugal, the details are given in [Cortez et al., 2009]. The goal is to determine which physicochemical attributes are relevant to the quality of the wine.

```
## [1] 1599   13
```

There are 1599 samples of wine and 13 attributes for each sample. There are 11 variables based on physicochemical tests and the remaining one is quality.

```
## 'data.frame':    1599 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.
5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.
58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 .
..
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.0
69 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 ...
```

Input variables (based on physicochemical tests):

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 / dm^3), a wine additive which can contribute to sulfur dioxide gas (S02) levels, winch acts as an antimicrobial and antioxidant;
11. **alcohol** (% by volume), the percent alcohol content of the wine.

Output variable (based on sensory data):

12. **quality** (score between 0 and 10).

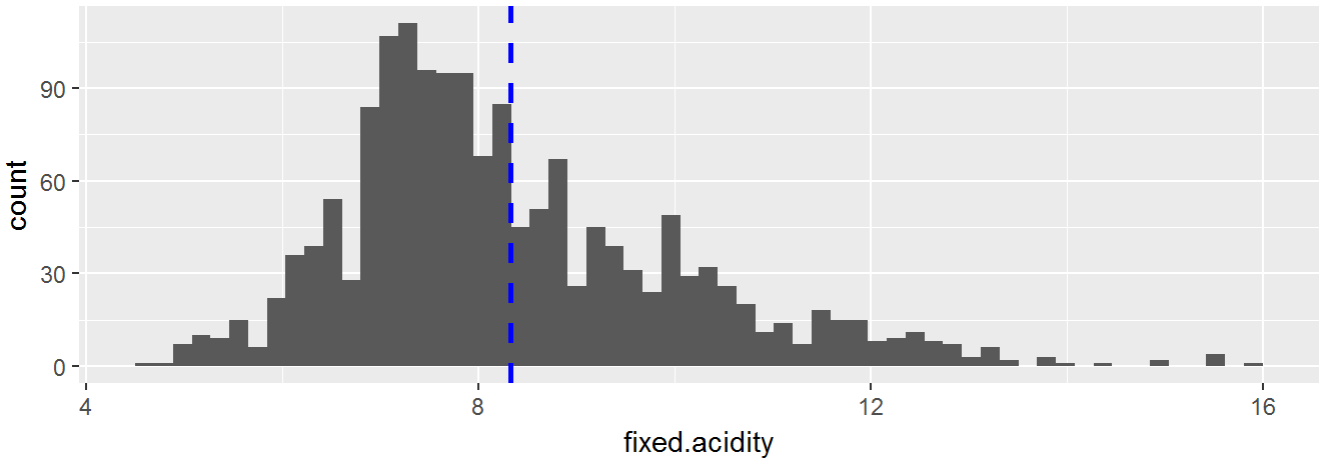
```
##        X          fixed.acidity   volatile.acidity  citric.acid
##  Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000
##  1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090
##  Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260
##  Mean   : 800.0   Mean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

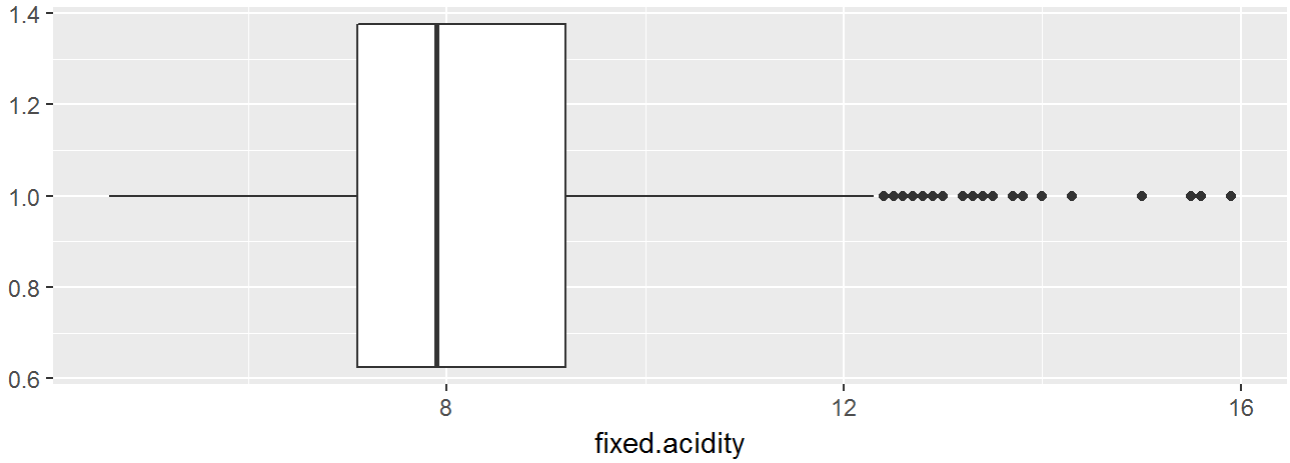
```
##  Min.   :  6.00       Min.   :0.9901   Min.   :2.740   Min.   :0.3300
##  1st Qu.: 22.00       1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500
##  Median : 38.00       Median :0.9968   Median :3.310   Median :0.6200
##  Mean   : 46.47       Mean   :0.9967   Mean   :3.311   Mean   :0.6581
##  3rd 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
```

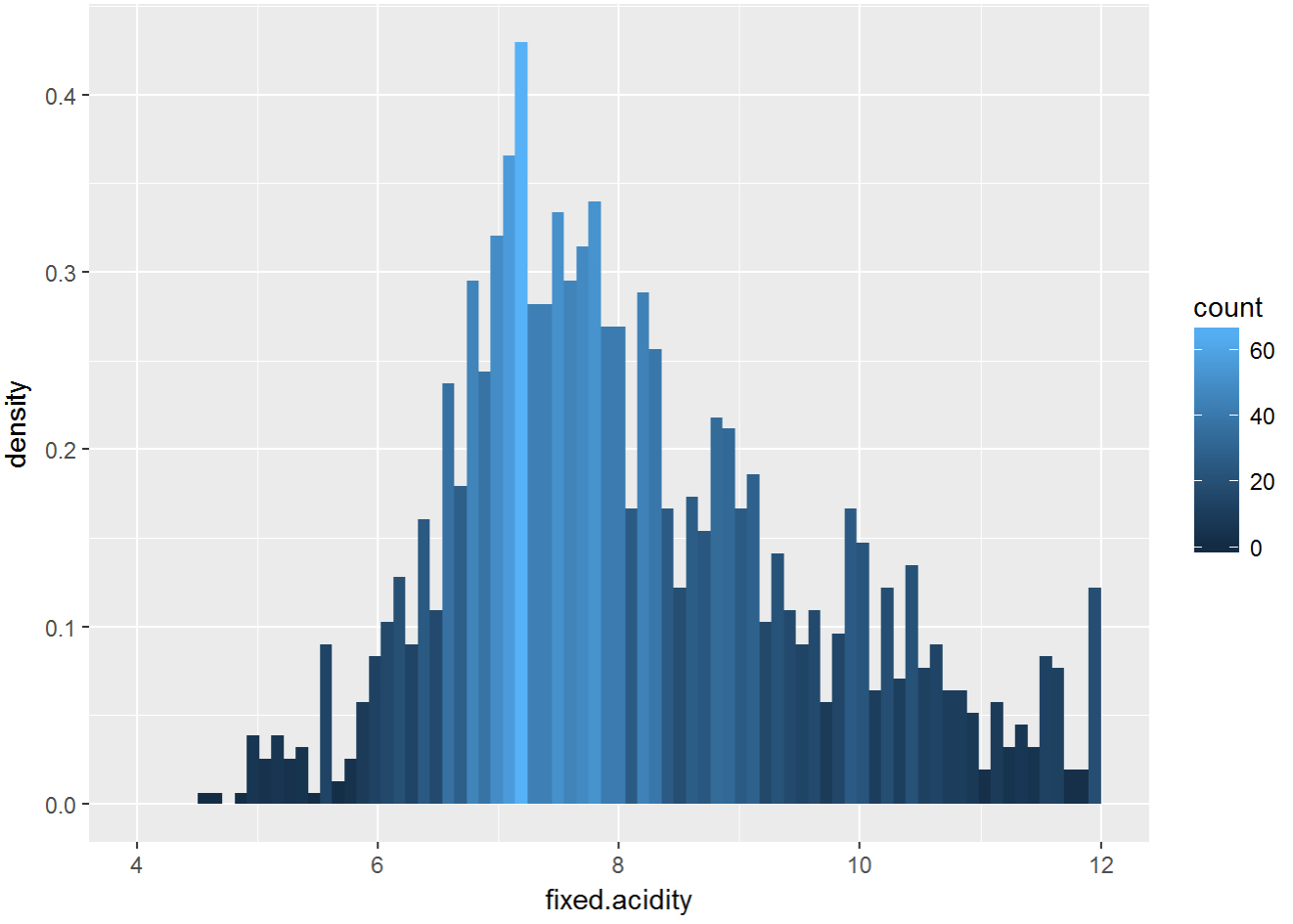
Univariate Plots Section

For each of the 11 numerical attributes I plot a frequency histogram with adjusted binwidth and a vertical line corresponding to the mean value of the dataset. I also include the corresponding boxplot. I investigate the existence of (mild) outliers (values which lie between 1.5 times and 3.0 times the interquartile range below the first quartile or above the third quartile) and extreme outliers (values which lie more than 3.0 times the interquartile range below the first quartile or above the third quartile). I compute the standard deviations for the whole data and for the data without outliers. I extract information about the extreme outliers in each case. To get more insight into and better understanding of the shape of the data I also plot density histograms, which I color by count and from which I exclude the extreme outliers.

Fixed acidity


count
90
60
30
0
4
8
12
16
fixed.acidity


1.4
1.2
1.0
0.8
0.6
8
12
16
fixed.acidity


density
0.4
0.3
0.2
0.1
0.0
4
6
8
10
12
fixed.acidity
count
60
40
20
0

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

```
## Outliers are greater than: 12.35
```

```
## Extreme outliers are greater than: 15.5
```

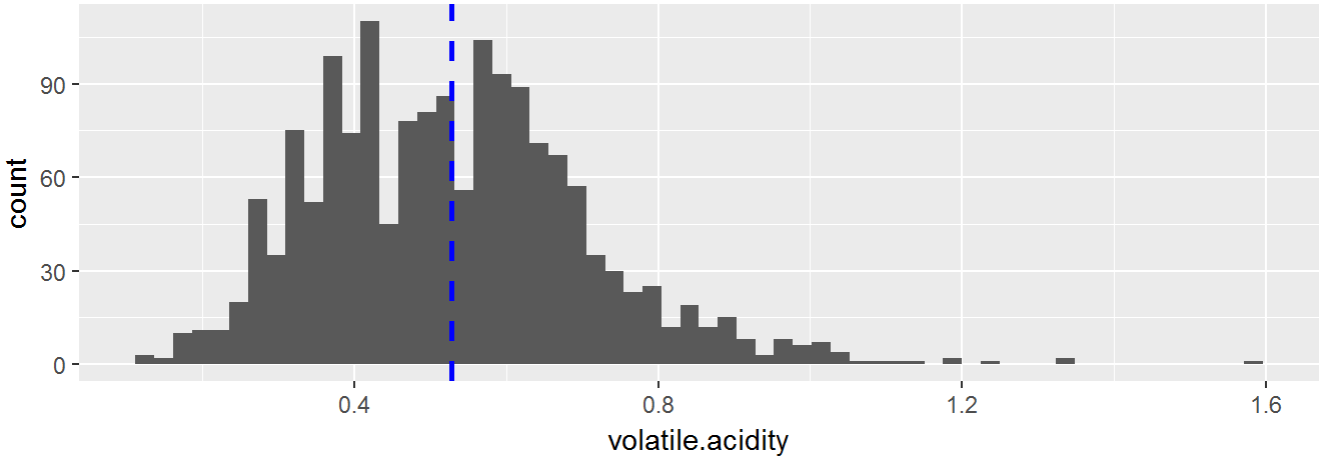
```
## The standard deviation is: 1.741096
```

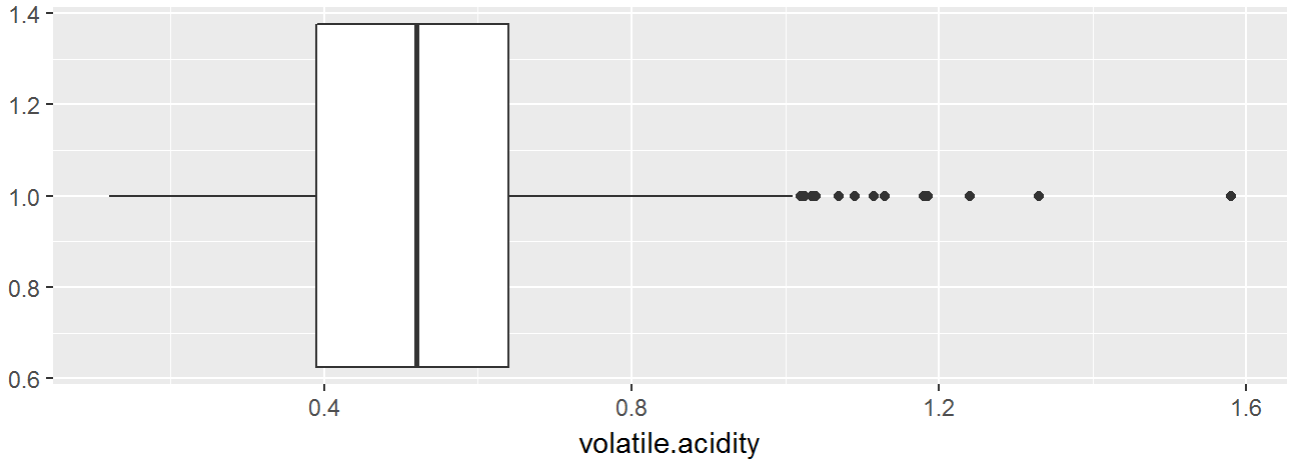
```
## Extreme outliers for fixed.acidity: 5
```

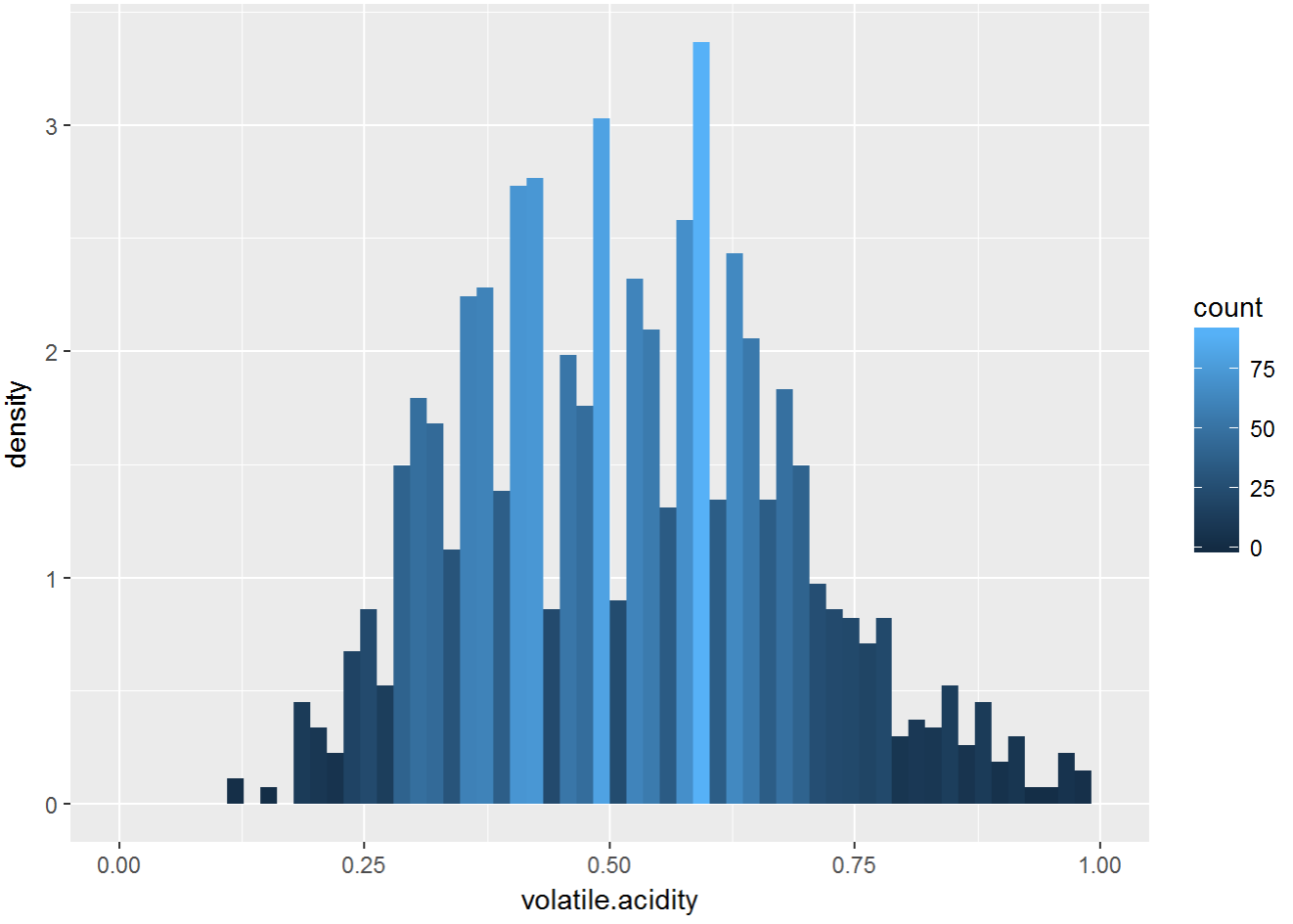
```
## The standard deviation without any outliers: 1.513582
```

The data for fixed.acidity is skewed to the right, with half of the wines having fixed.acidity between 7 and 9. There are several outliers of which five are extreme. It seems that there are certain fixed.acidity levels that are more common, these are suggested by the presence of isolated higher peaks and gaps more noticeable when the binwidth is reduced (as in the density histogram).

Volatile acidity


count
90
60
30
0
0.4
0.8
1.2
1.6
volatile.acidity


1.4
1.2
1.0
0.8
0.6
0.4
0.8
1.2
1.6
volatile.acidity


density
3
2
1
0
0.00
0.25
0.50
0.75
1.00
volatile.acidity
count
75
50
25
0

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

```
## Outliers are greater than: 1.015
```

```
## Extreme outliers are greater than: 1.39
```

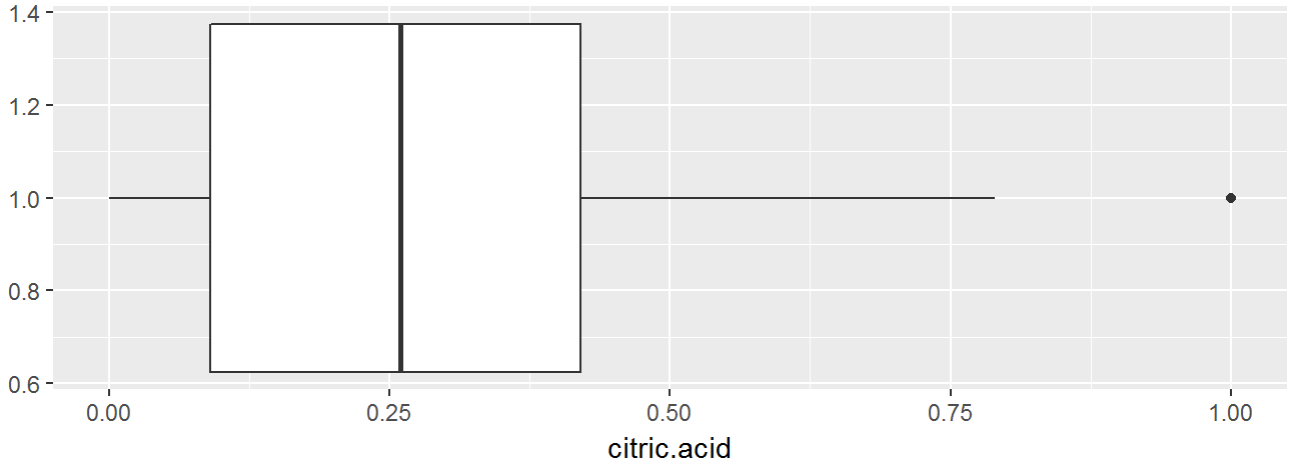
```
## The standard deviation is: 0.1790597
```

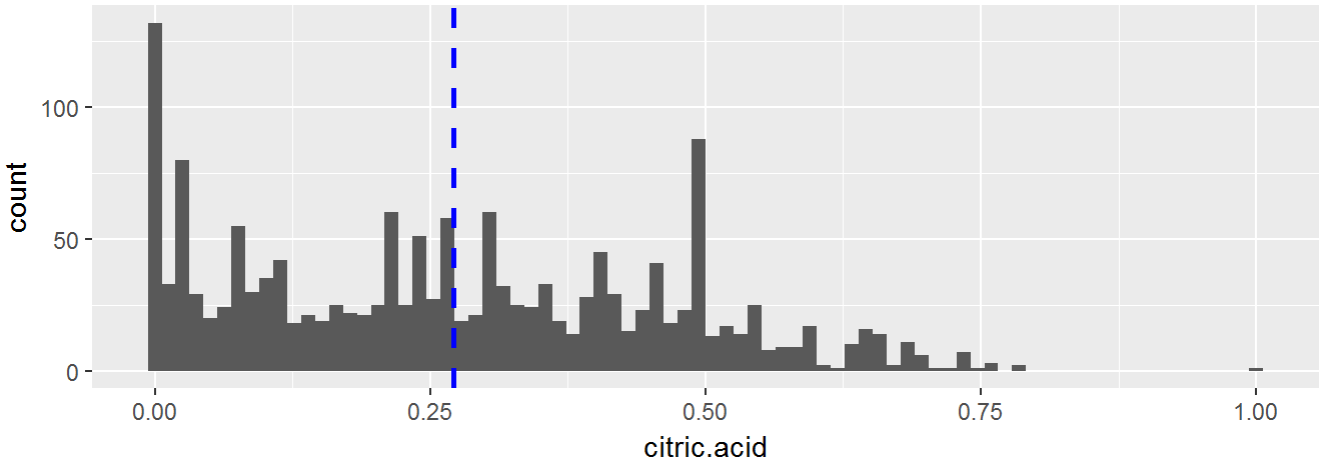
```
## Extreme outliers for volatile.acidity: 1
```

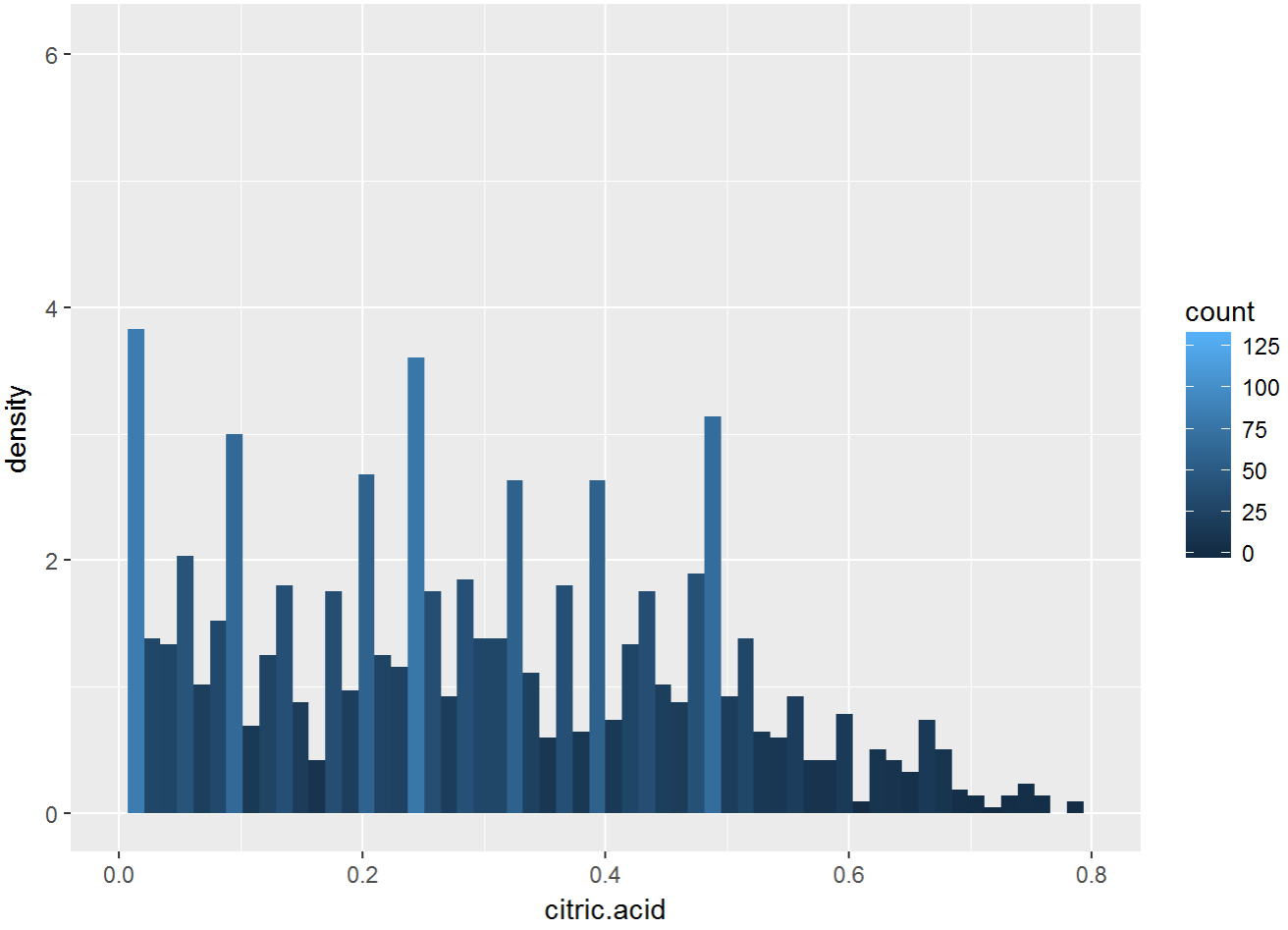
```
## The standard deviation without outliers: 0.1665809
```

The frequency histogram for volatile acidity is multimodal with a slight right skew. About half of the wines have volatile acidity between 0.4 and 0.6. There are three noticeable peaks, these account for almost 30% of the wine samples. As we decrease the bin size and remove the data in the long tail (see the density histogram) the multimodality becomes more obvious. The boxplot indicates the presence of several outliers, of which one is extreme. There are also a couple of isolated data points on the left, not technically outliers.

Citric acid


1.4
1.2
1.0
0.8
0.6
0.00
0.25
0.50
0.75
1.00
citric.acid


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


6
4
2
0
density
0.0
0.2
0.4
0.6
0.8
citric.acid
count
125
100
75
50
25
0

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

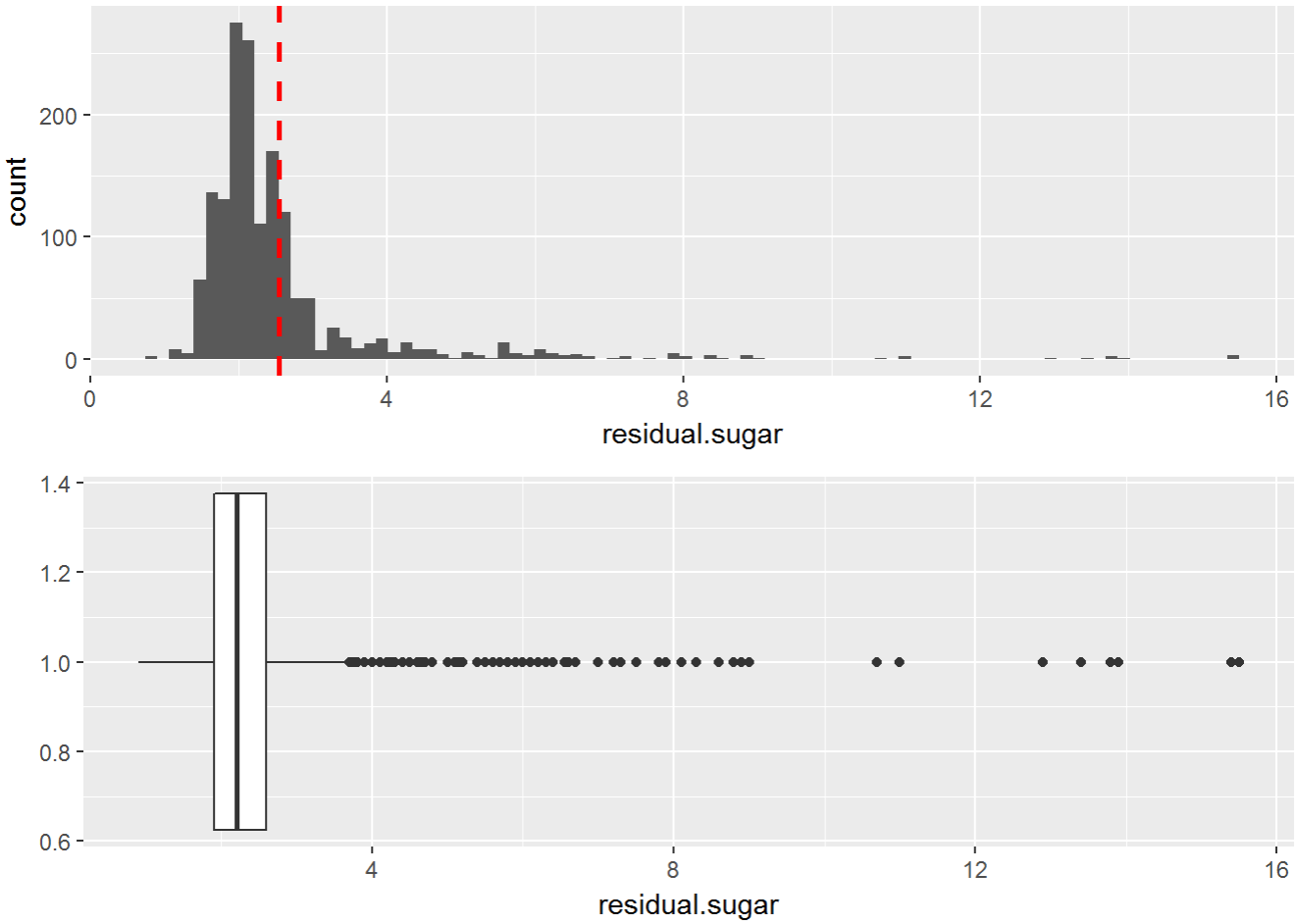
```
## Outliers are greater than: 0.915
```

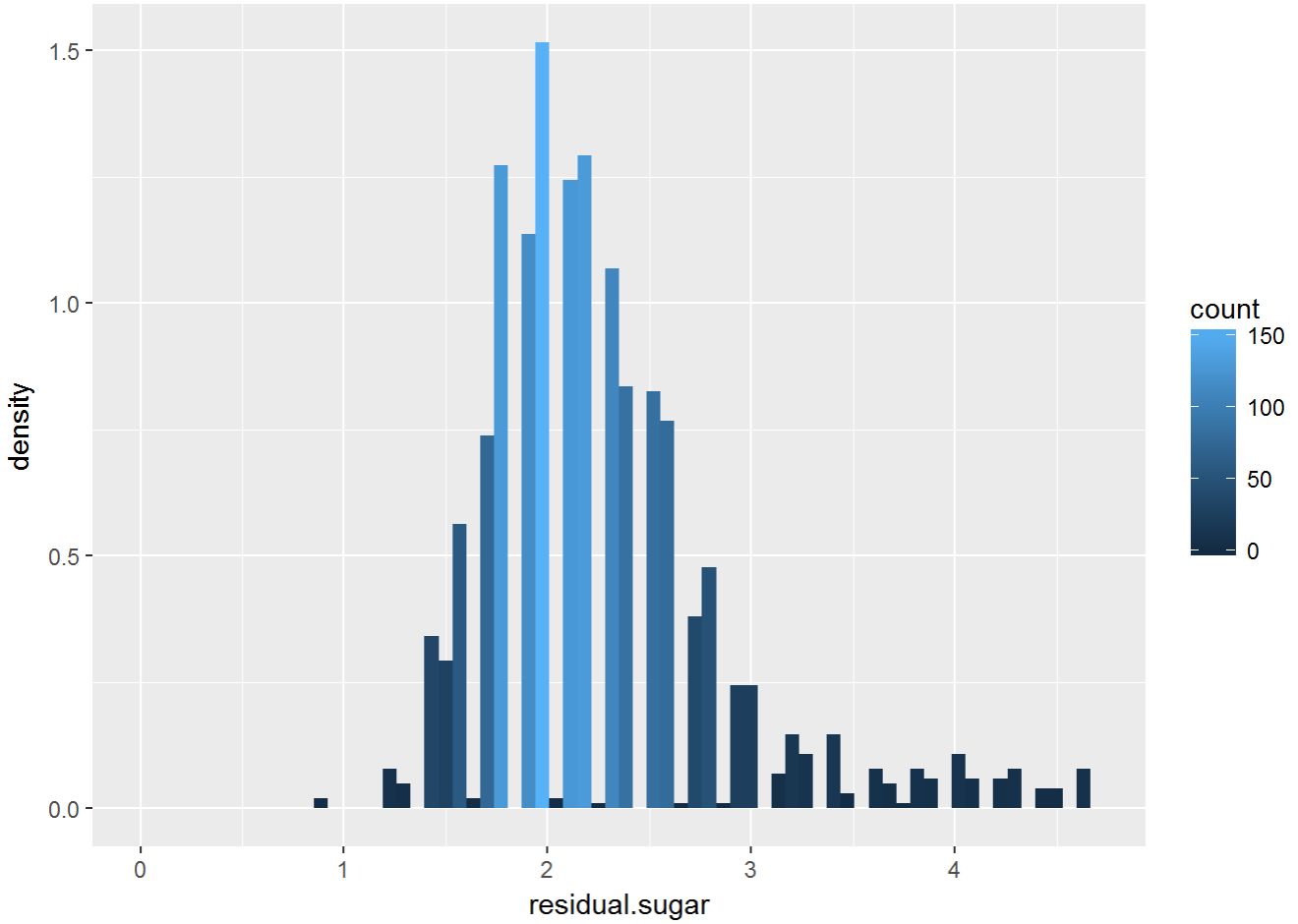
```
## Extreme outliers are greater than: 1.41
```

```
## The standard deviation is: 0.1948011
```

The distribution of the citric.acid content is multimodal. Most of the data has citric.acid less than 0.5. There are several peaks, the highest corresponds to wines with no citric.acid content. There is also an outlier that corresponds to the value 1.

Residual sugar






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

```
## Outliers are greater than: 3.65
```

```
## Extreme outliers are greater than: 4.7
```

```
## The standard deviation is: 1.409928
```

```
## Extreme outliers for residual.sugar: 88
```

```
## The standard deviation without any outliers: 0.4491415
```

The data for residual sugar has a very long right tail. Outside this tail, the data shows a bimodal distribution in the frequency histogram. Most of the wines have residual sugar of up to 2.6 but the maximum value is 15.5, so we see a large spread of values. To plot the density histogram I adjusted the x-axis limits to exclude the extreme outliers. There are numerous gaps and the

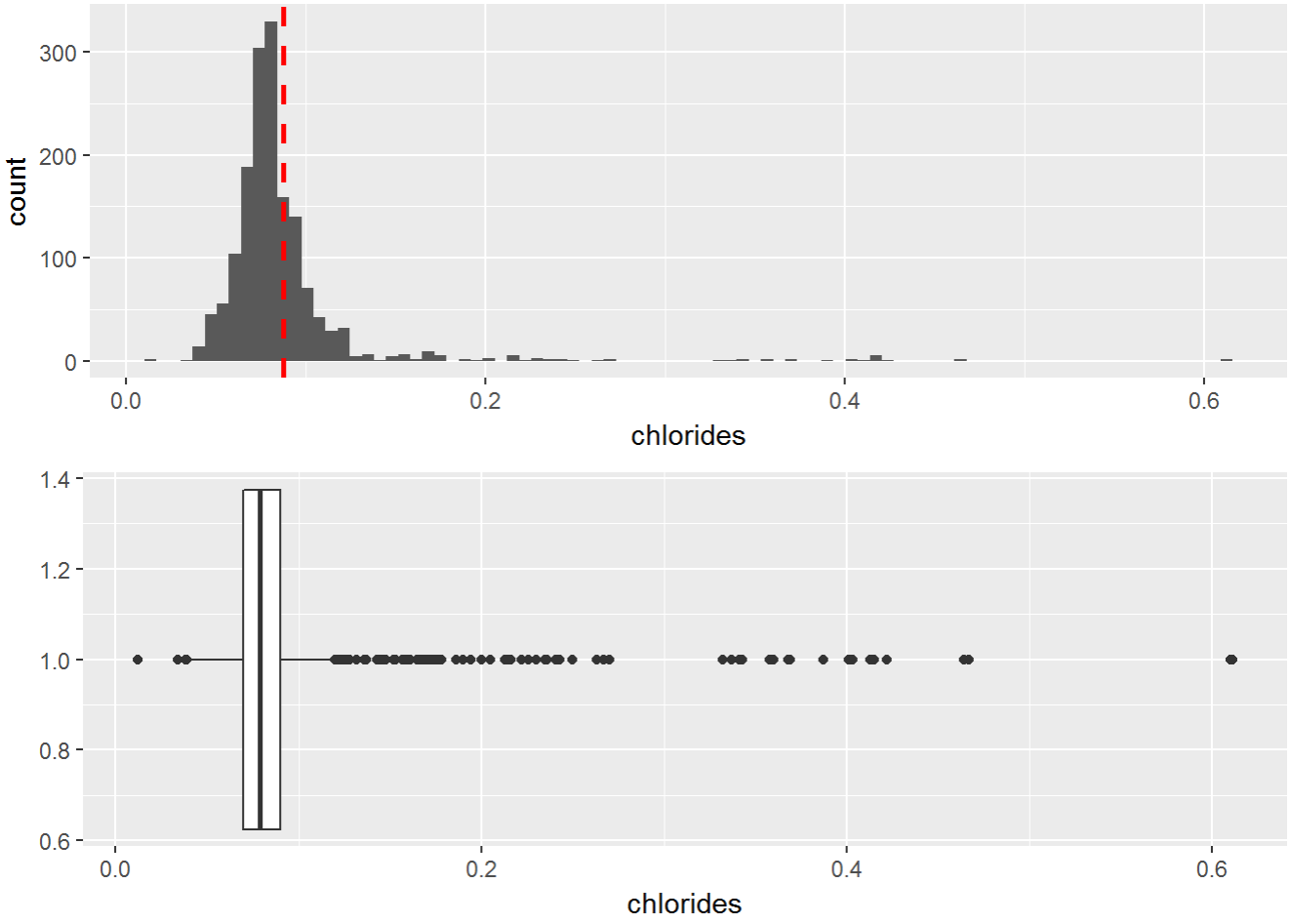
distribution is still right skewed. There are 88 extreme outliers in this case. The main data, without counting the outliers, has a relatively small spread.

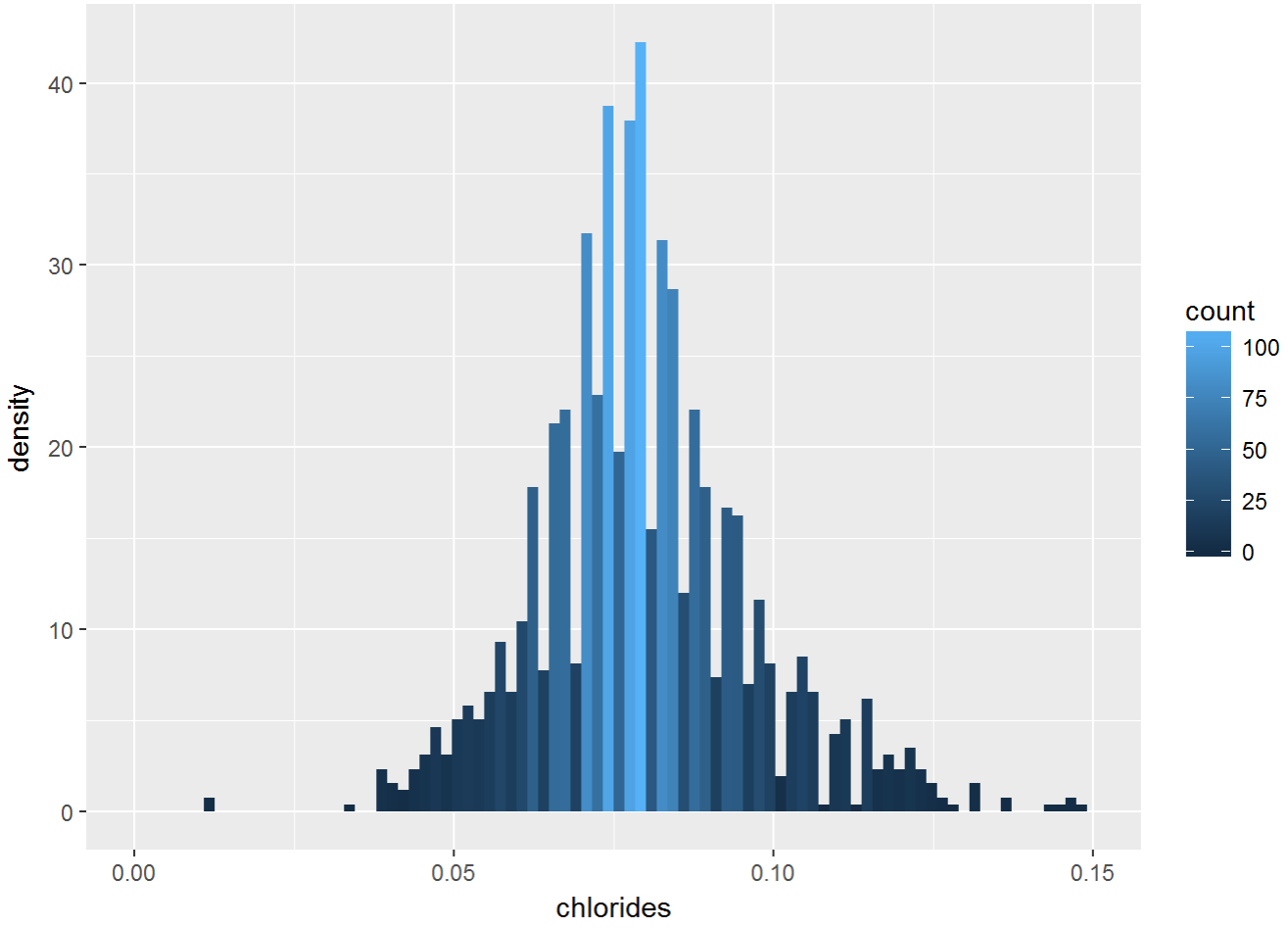
Sugar levels

```
Create a factor variable sugar_levels:

    low_sugar       residual.sugar <= median
    high_sugar      median < residual.sugar < mild_outlier_value
    extra_sugar     mild_outlier_value < residual.sugar
```

```
##
## extra_sugar  high_sugar   low_sugar
##         155         561         883
```

Chlorides






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

```
## Outliers are smaller than: 0.04 and greater than: 0.12
```

```
## Extreme outliers are greater than: 0.15
```

```
## The standard deviation is: 0.0470653
```

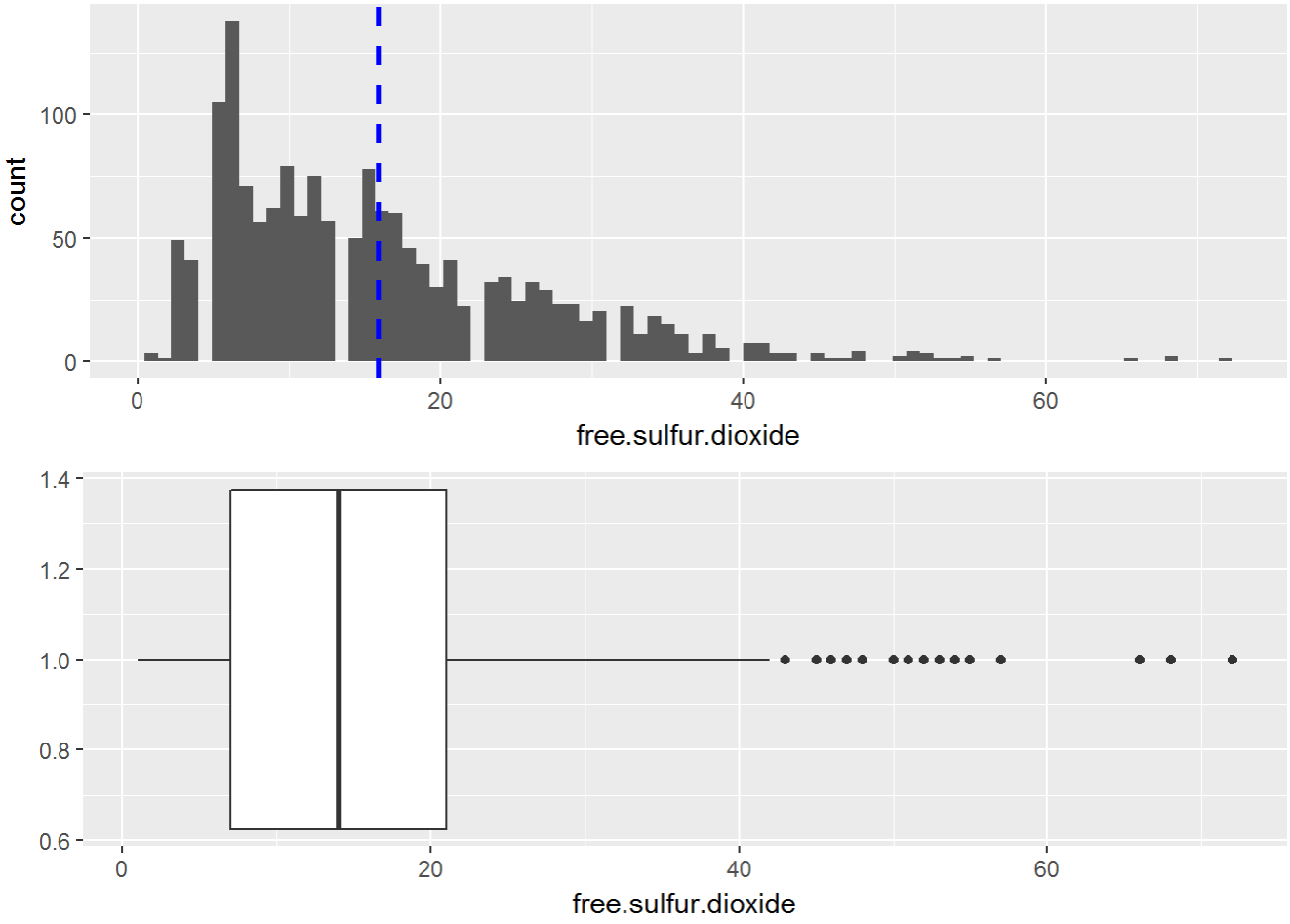
```
## Extreme outliers for chlorides: 67
```

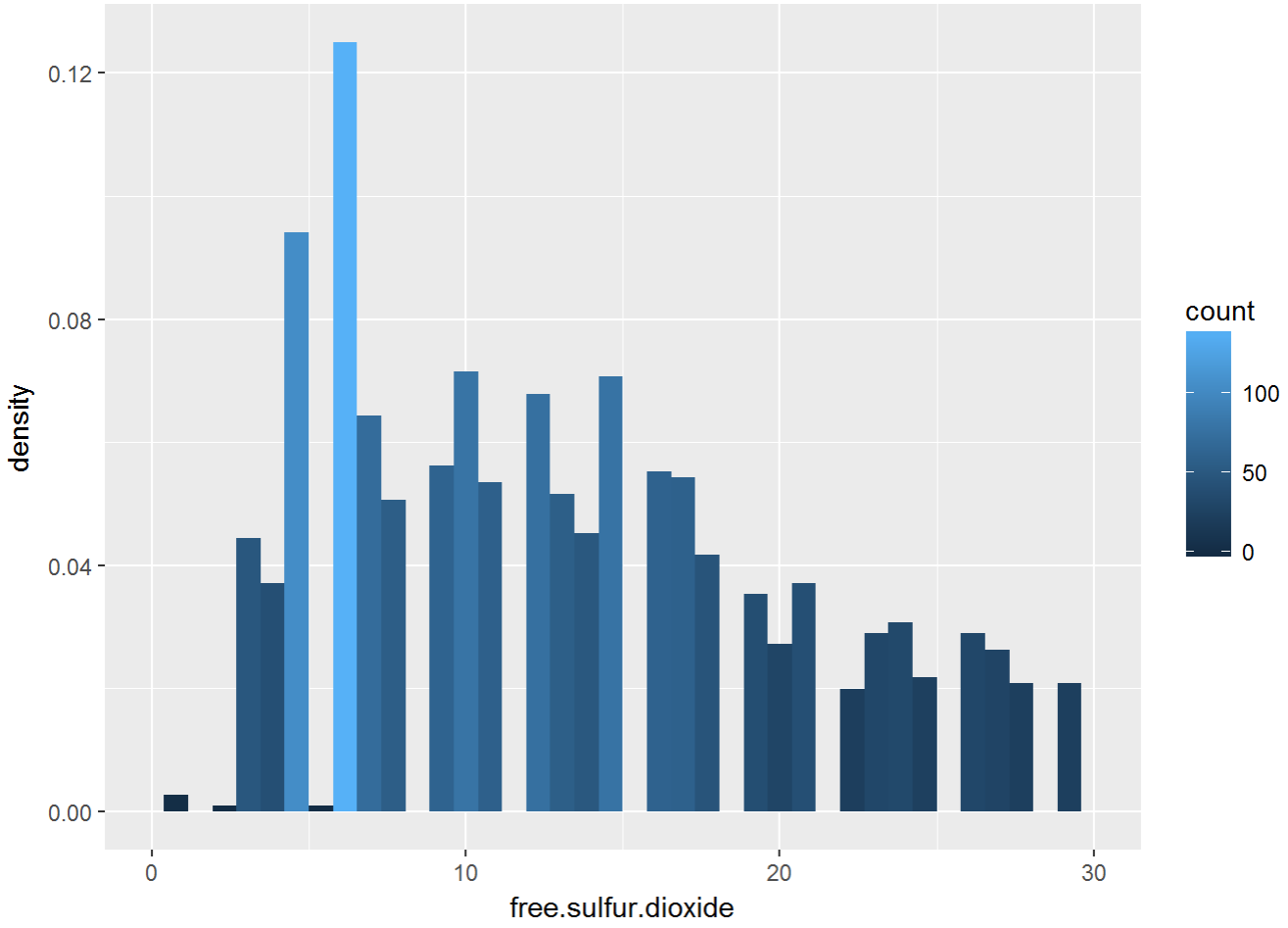
```
## The standard deviation without any outliers: 0.01527472
```

The frequency histogram for chlorides has a long tail to the right. The majority of the data is comprised between 0.07 and 0.09. We see a large spread of values, up to 0.6. There are many outliers, some of them to the left of the mean. Specifically, there are 67 extreme outliers. For the density histogram, I zoomed into the main data, the distribution is almost normal, slightly right

skewed.

Free sulfur dioxide






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

```
## Outliers are greater than: 42
```

```
## Extreme outliers are greater than: 63
```

```
## The standard deviation is: 10.46016
```

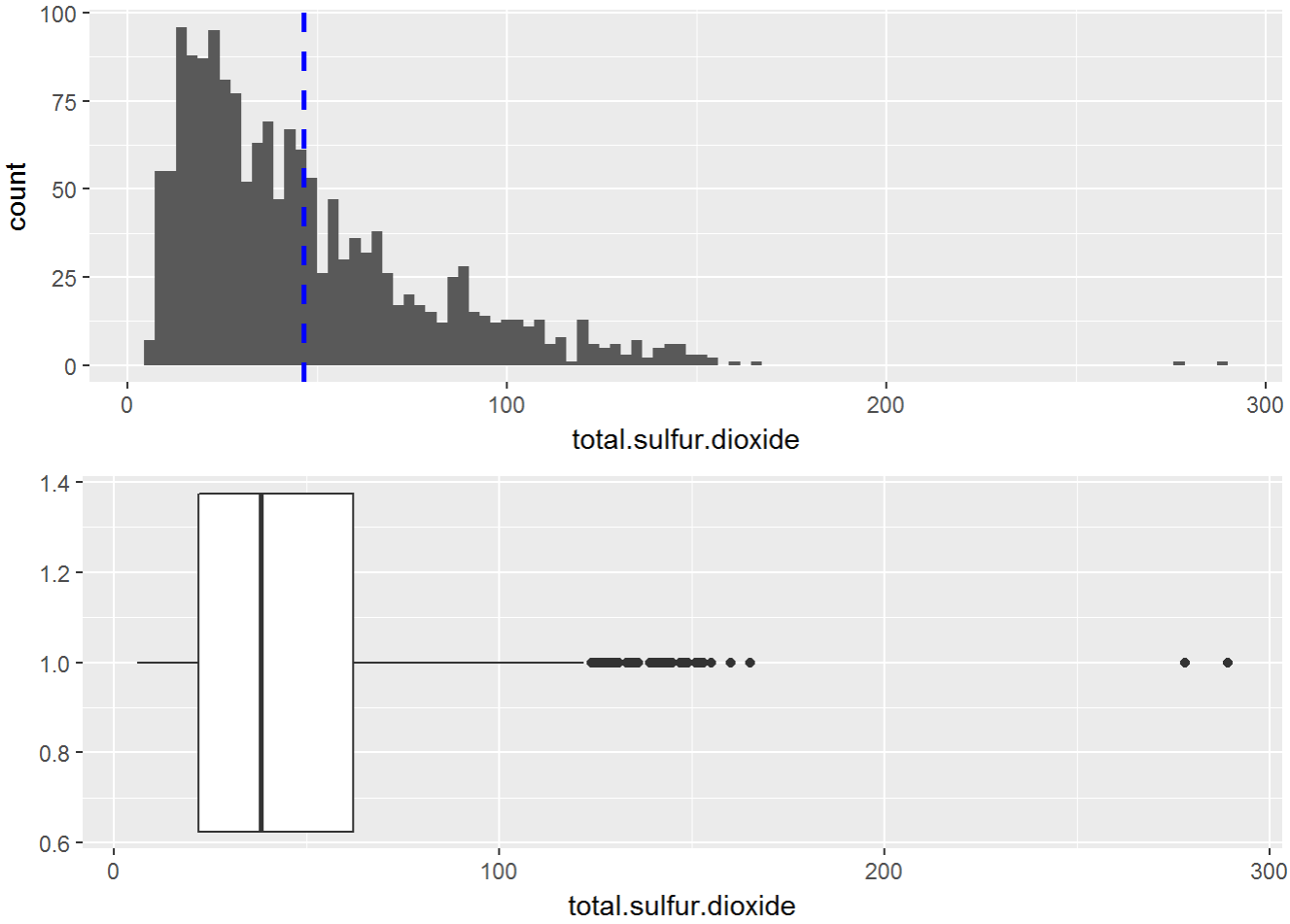
```
## Extreme outliers for free sulfur dioxide: 4
```

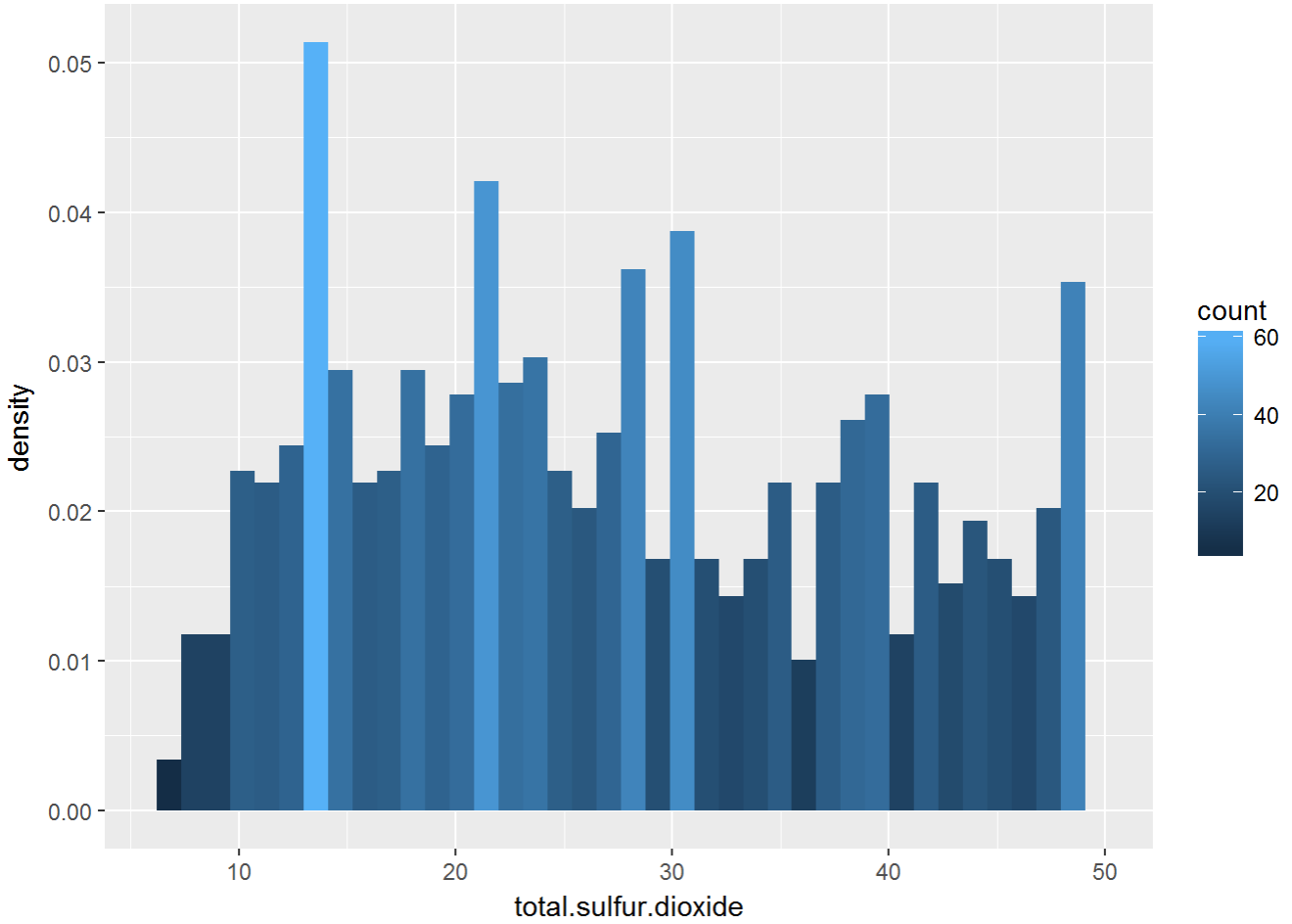
```
## The standard deviation without any outliers: 9.226586
```

The frequency diagram for free sulfur dioxide is right skewed with a long tail, several gaps and two very high peaks. Half of the wines have free sulfur dioxide between 7 and 21. There are several outliers, four of which are extreme. The highest amount of 72 and a large standard deviation indicate a wide spread of values. In the density histogram, I take a second look at the

wines with free sulfur dioxide less than 30. The two high peaks are noticeable and the gaps suggest that there are certain values that are more common. A small number of wines have very low amounts of free sulfur dioxide.

Total sulfur dioxide






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

```
## Outliers are greater than: 122
```

```
## Extreme outliers are greater than: 182
```

```
## The standard deviation is: 32.89532
```

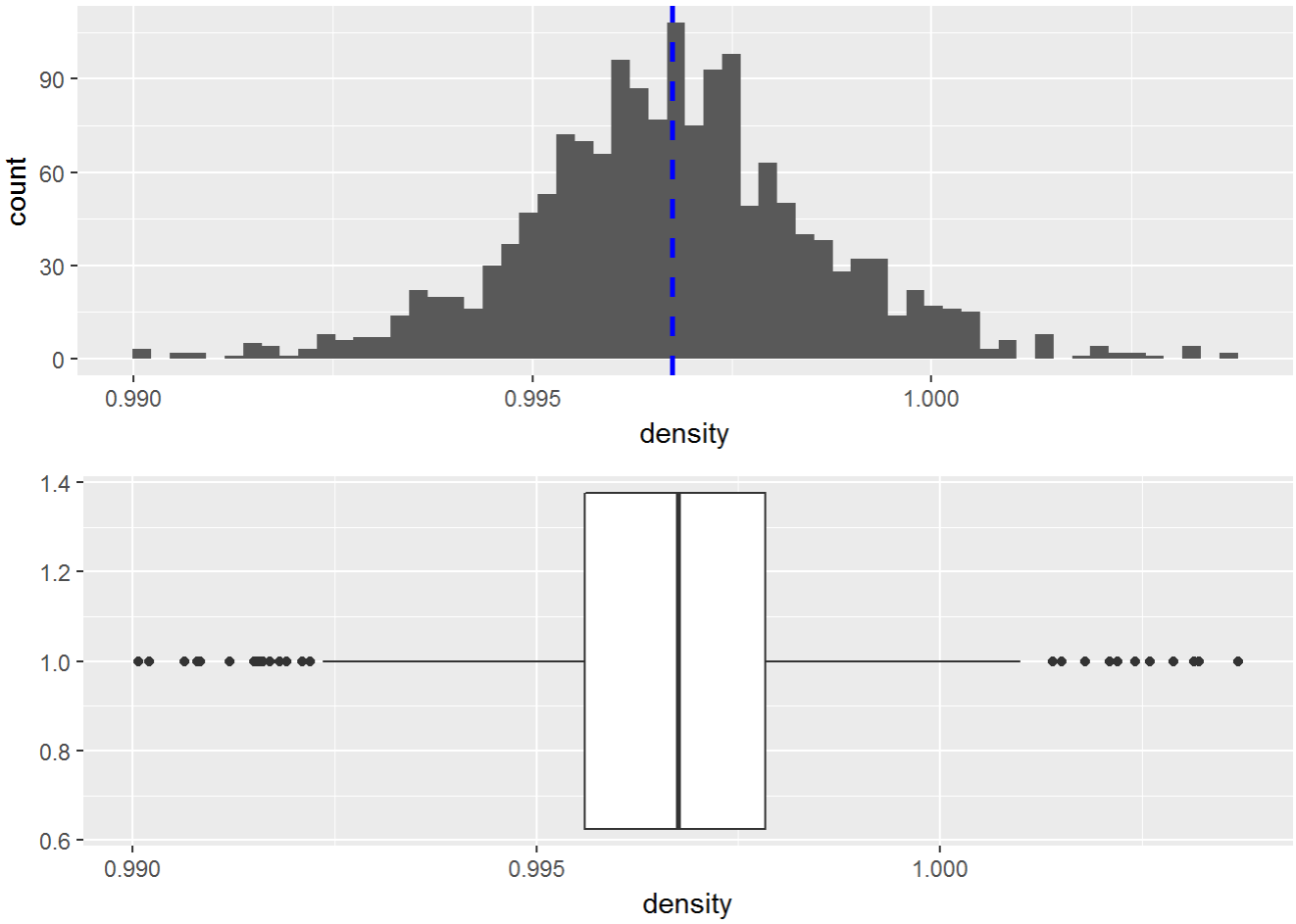
```
## Extreme outliers for total sulfur dioxide: 2
```

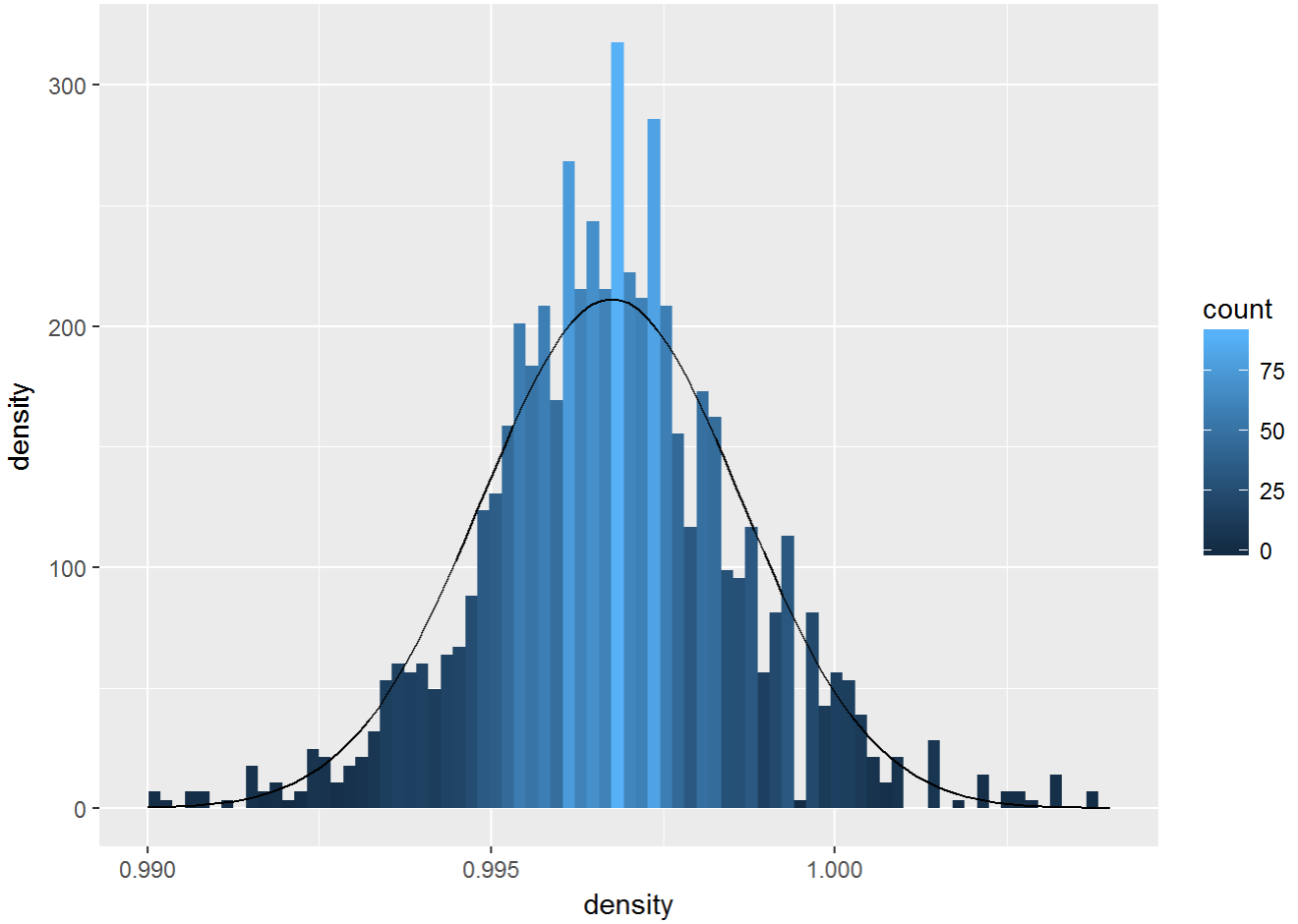
```
## The standard deviation without any outliers: 27.2148
```

The shape of the frequency histogram for the total sulfur dioxide is right skewed, with a rather steep left hand side. Most of the wines have total sulfur dioxide between 6 and 62, but there is a wide spread up to values of almost 300. There are lots of outliers, the two extreme outliers have very large values and I wonder if those records are accurate. In the density histogram, I take a

closer look at the distribution that lies to the left of the mean. This subset of the data has a multimodal distribution with several quite high peaks.

Density






```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0037
```

```
## Outliers are smaller than: 0.9922475 and greater than: 1.001187
```

```
## Extreme outliers are greater than: 1.00454
```

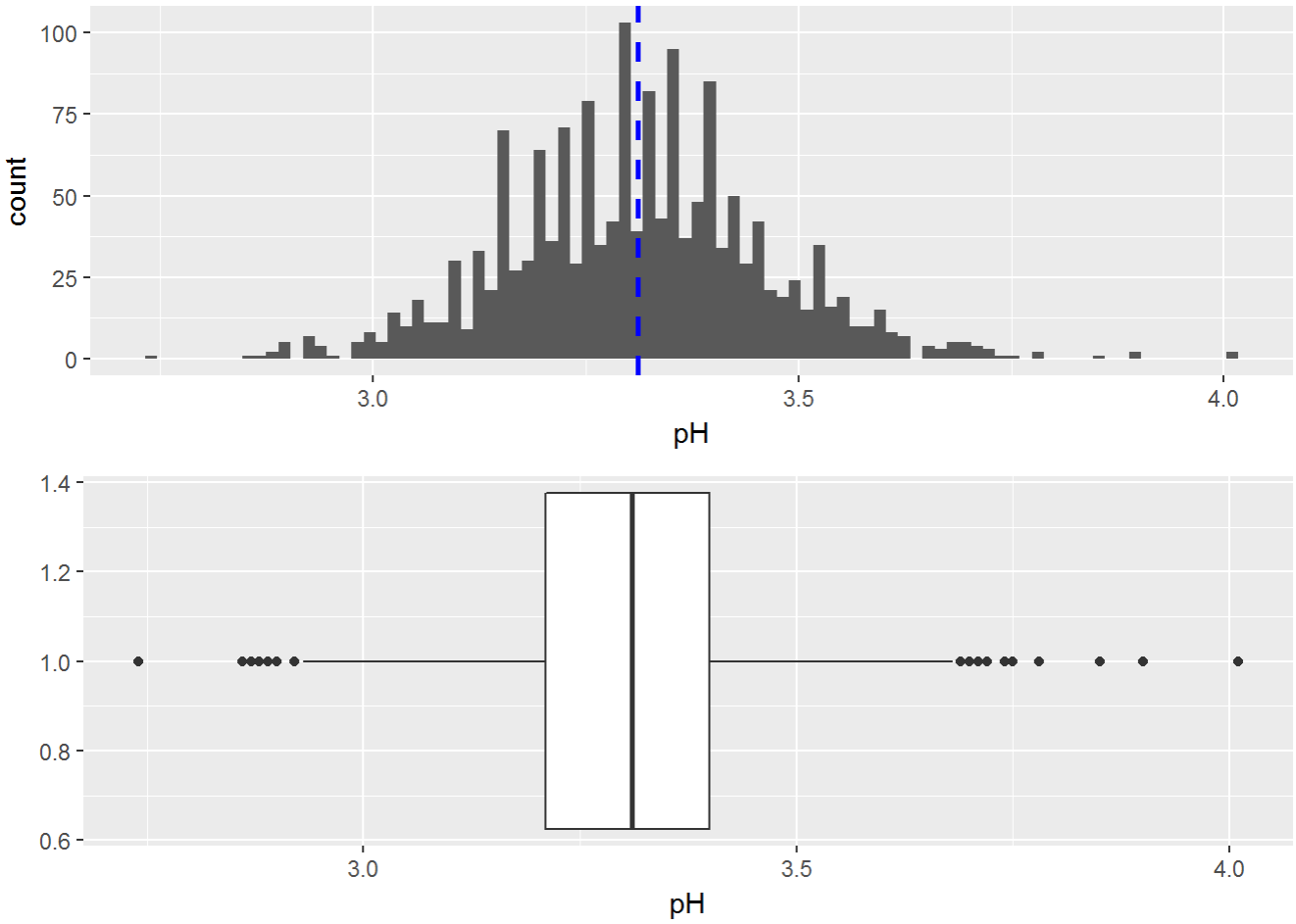
```
## The standard deviation is: 0.001887334
```

```
## Extreme outliers for density: 0
```

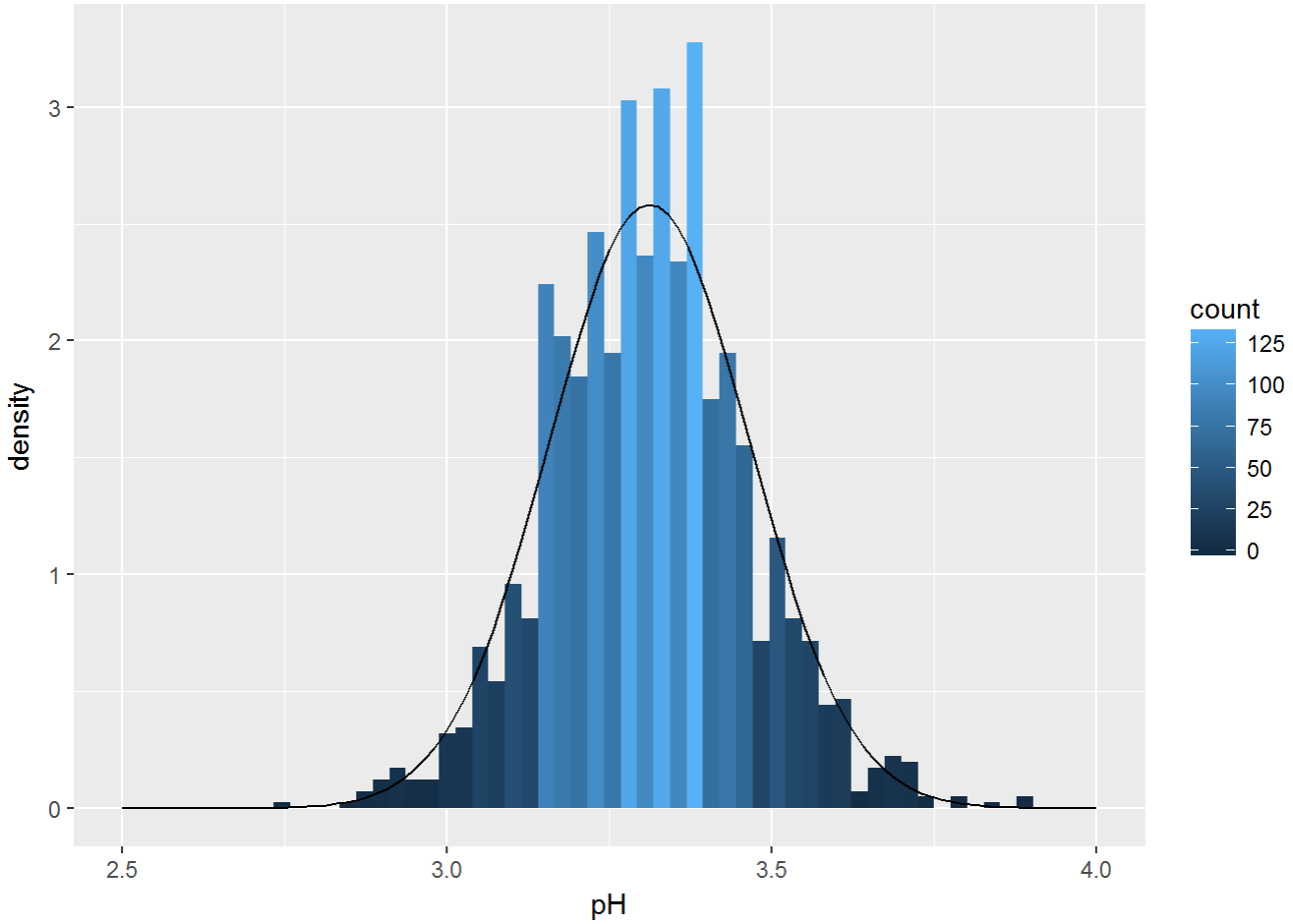
The data for density is interesting, its distribution is normal, with values ranging in a small interval from 0.9901 to 1.0037. There are outliers on both sides, as it can be easily noticed from the boxplot, but none of these outliers are extreme. A normal distribution curve overlaps nicely with the density histogram. There are several peaks and gaps.

---

pH


100
75
50
25
0
count
3.0
3.5
4.0
pH

1.4
1.2
1.0
0.8
0.6
3.0
3.5
4.0
pH




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

```
## Outliers are greater than: 3.685
```

```
## Extreme outliers are smaller than 2.64 and greater than: 3.97
```

```
## The standard deviation is: 0.1543865
```

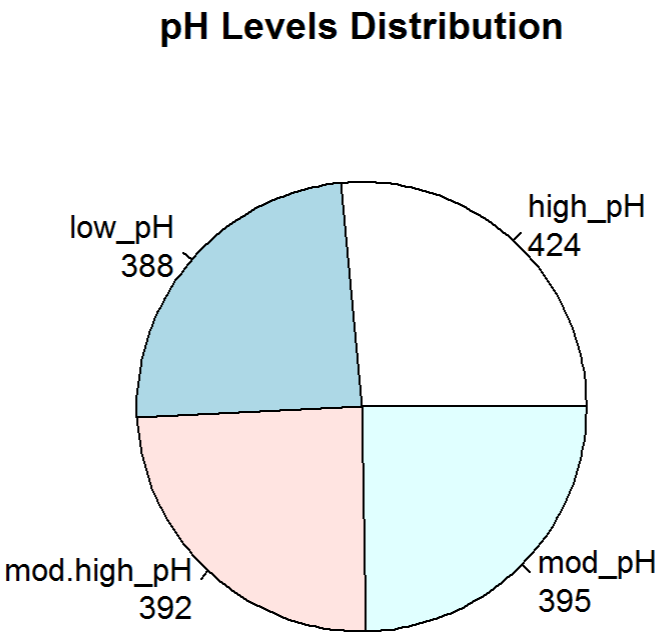
```
## Extreme outliers for pH: 2
```

The distribution for pH is normal, although there are several high peaks that are almost equidistantly distributed. The values range from 2.74 to about 4, with half of the data having pH of about 3.2 to 3.4. The boxplot indicates the presence of outliers, two of which are extreme. The normal distribution curve fits well with the density diagram.

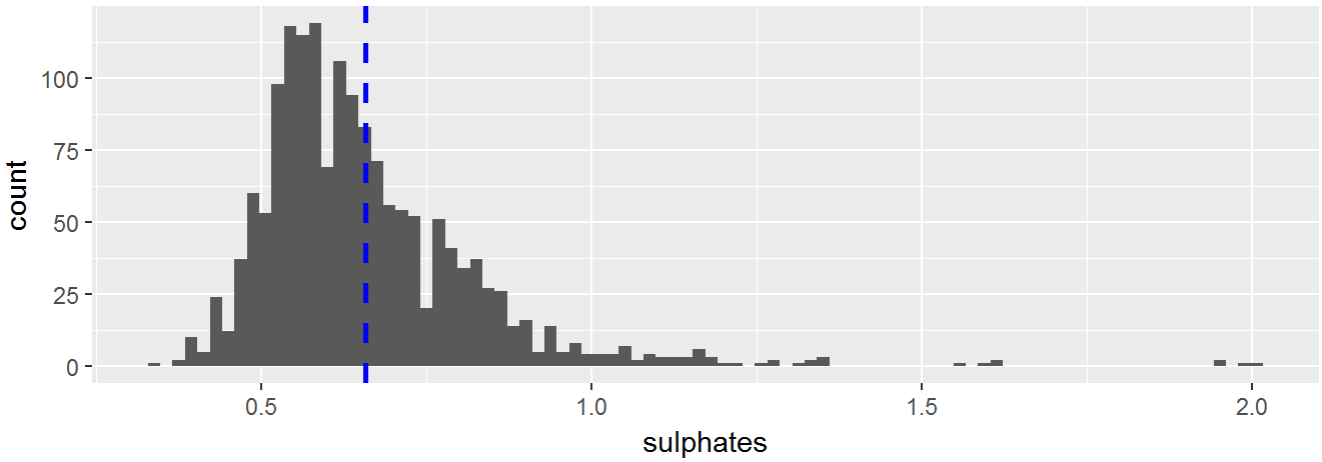
---

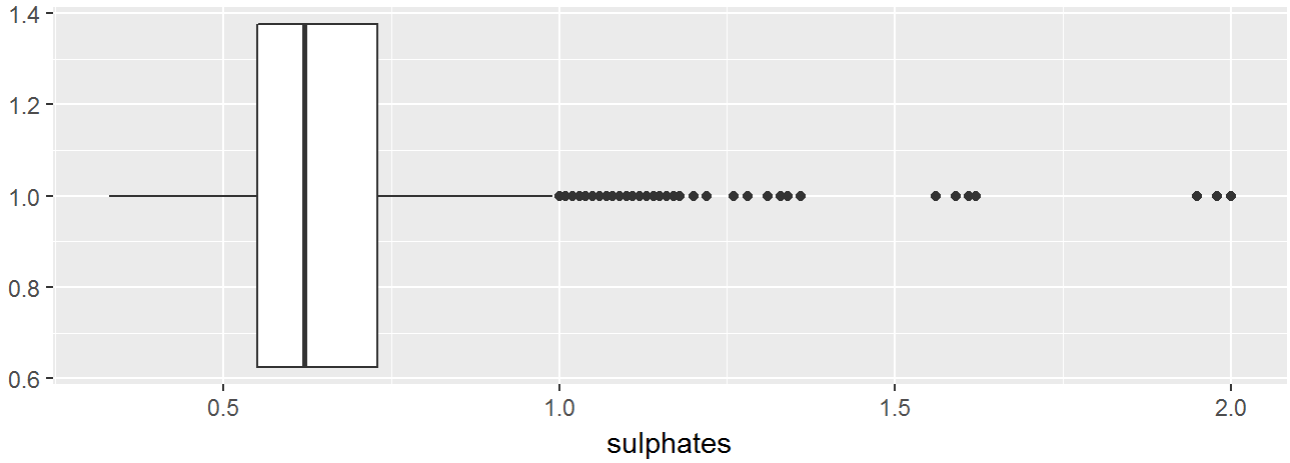
pH levels

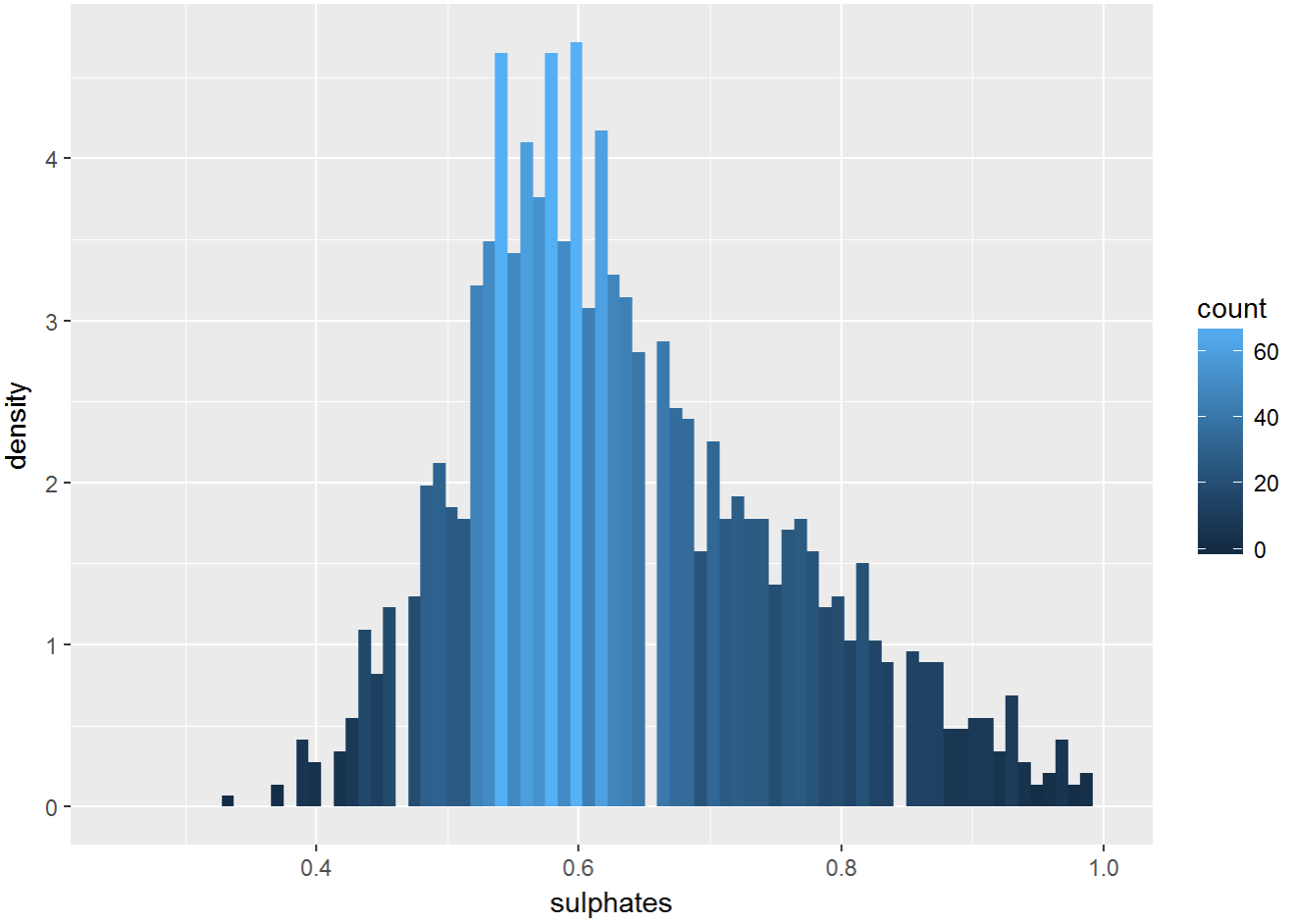
```
Create a factor variable for the pH levels, based on quartiles:

    low_pH           lowest 25% of pH values
    mod_pH           25%-50% of pH values
    mod.high_pH      50%-75% of pH values
    high_pH          highest 25% of pH values
```



Sulphates


count
100
75
50
25
0
0.5
1.0
1.5
2.0
sulphates


1.4
1.2
1.0
0.8
0.6
0.5
1.0
1.5
2.0
sulphates


density
4
3
2
1
0
0.4
0.6
0.8
1.0
sulphates
count
60
40
20
0

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

```
## Outliers are greater than: 1
```

```
## Extreme outliers are greater than: 1.27
```

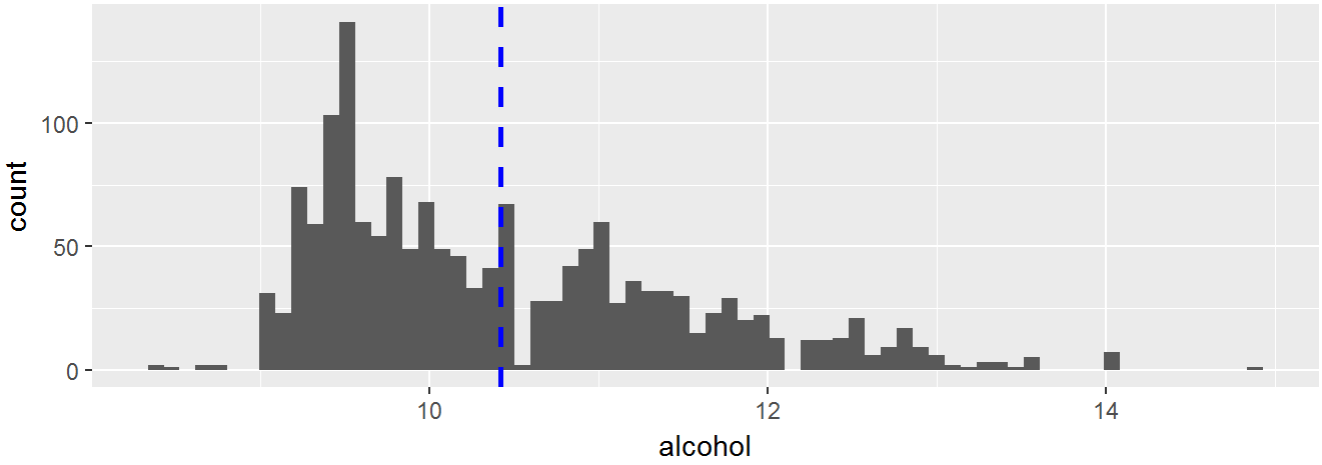
```
## The standard deviation is: 0.169507
```

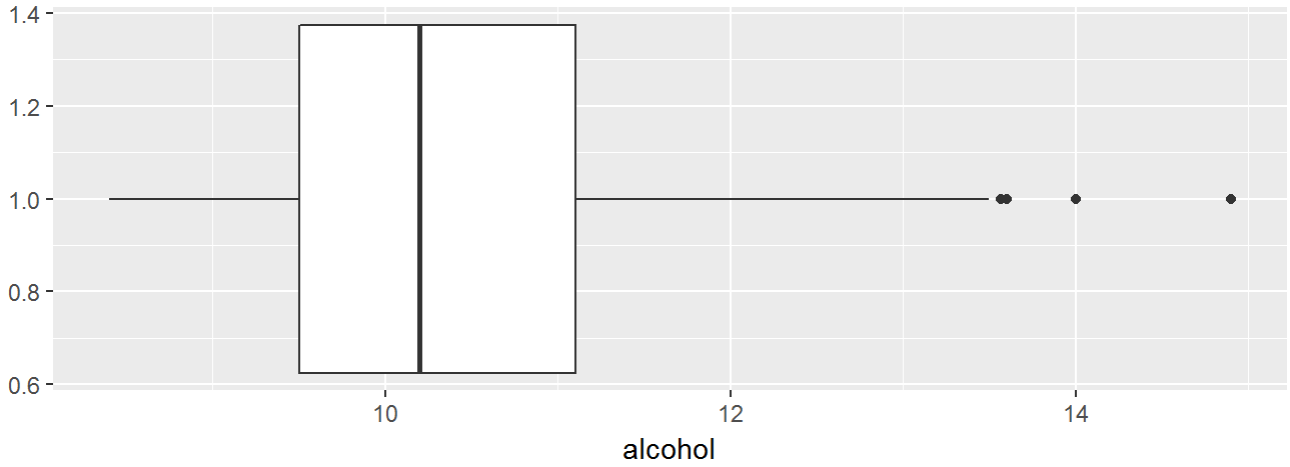
```
## Extreme outliers for sulphates: 16
```

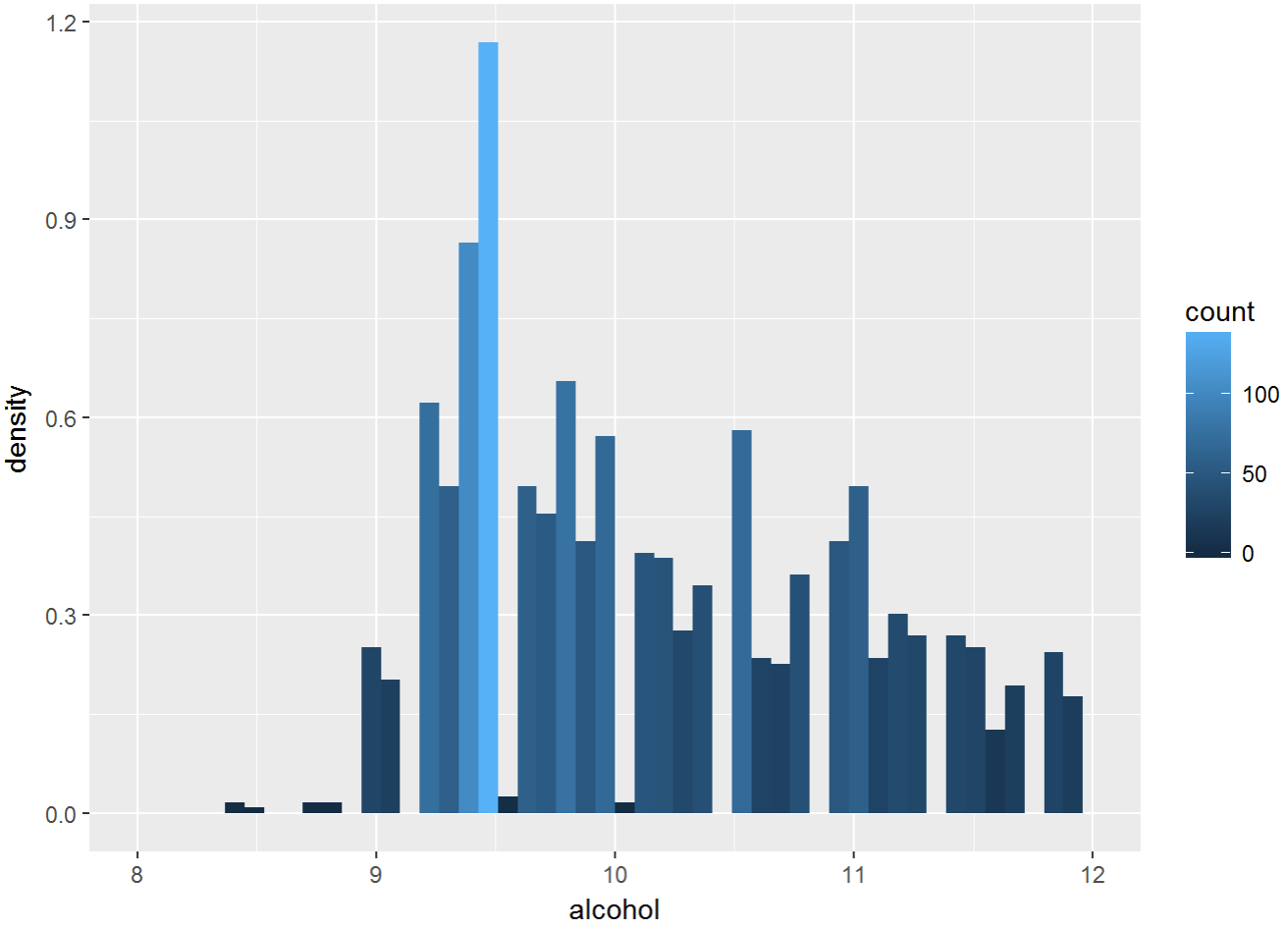
```
## The standard deviation without any outliers: 0.1209633
```

The sulphates have a skewed distribution, with a very long tail to the right due to the numerous outliers. Most values fall between 0.55 and 0.73. There are 16 extreme outliers, that correspond to values as large as 2. In the density histogram, I look at the main bulk of data, and the distribution is skewed to the right. There are several high peaks and a few gaps.

Alcohol


count
100
50
0
10
12
14
alcohol


1.4
1.2
1.0
0.8
0.6
10
12
14
alcohol


density
1.2
0.9
0.6
0.3
0.0
8
9
10
11
12
alcohol
count
100
50
0

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

```
## Outliers are greater than: 13.5
```

```
## Extreme outliers are greater than: 15.9
```

```
## The standard deviation is: 1.065668
```

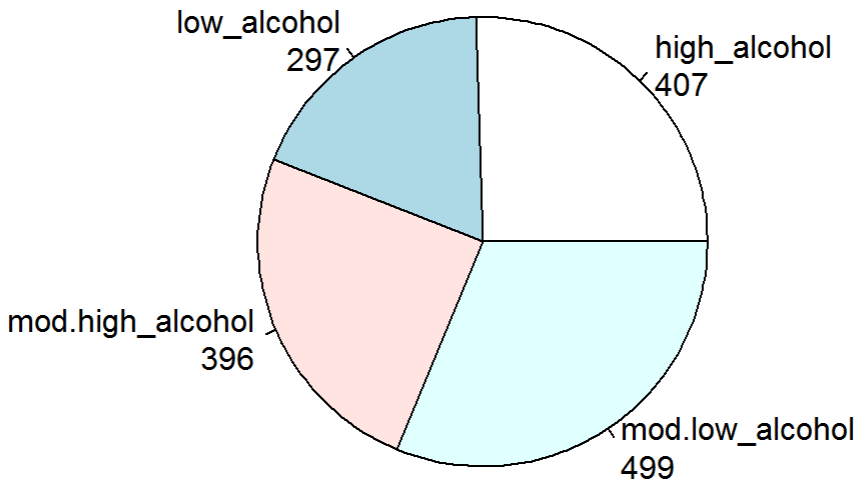
```
## Extreme outliers for alcohol: 0
```

```
## The standard deviation without any outliers: 1.021412
```

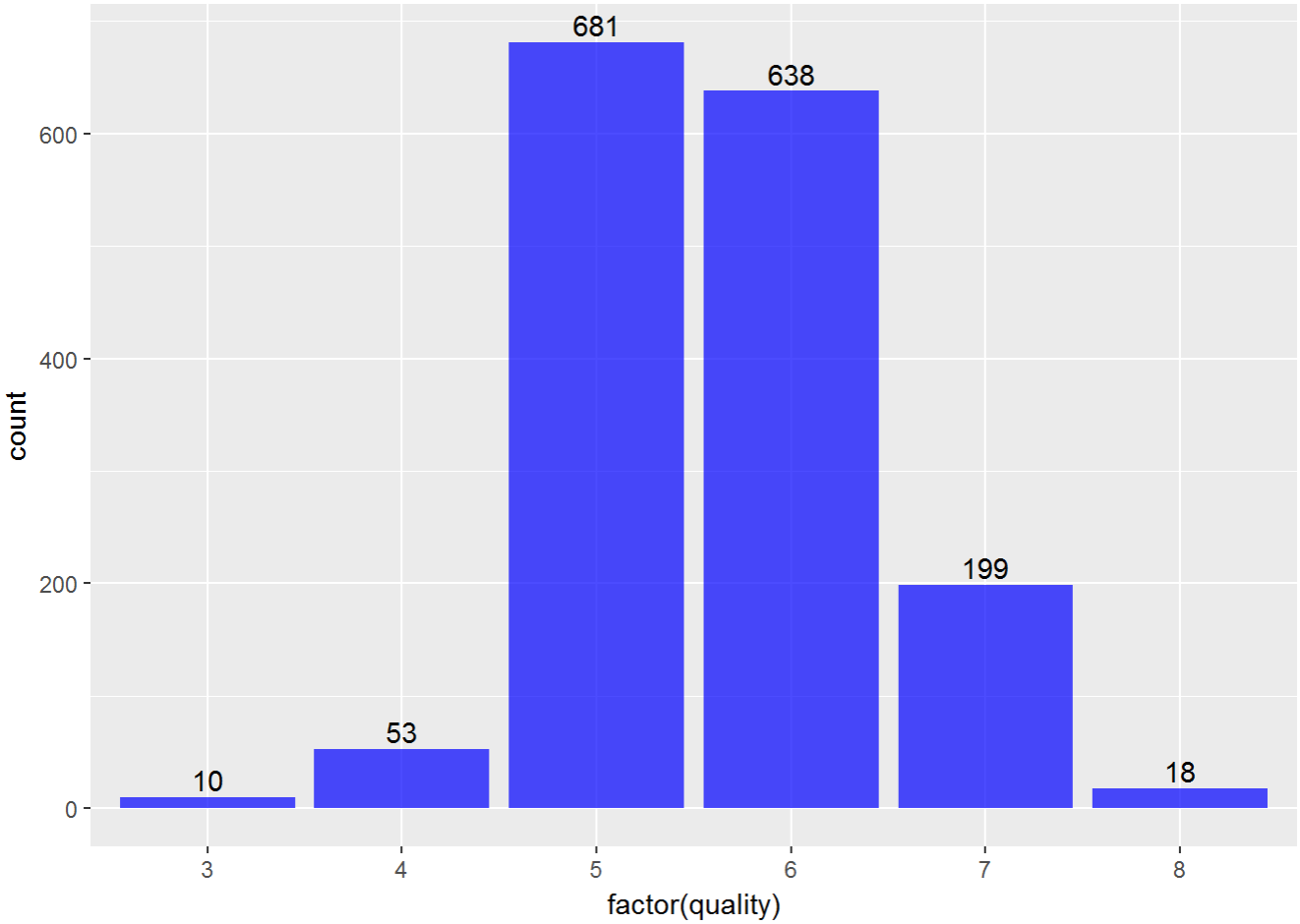
The alcohol content does not vary too much, it takes values in the interval from 8.4 to 14.9, with most of the wines having at alcohol of at most 11.10. The distribution is right skewed. Not too many outliers in this case, no extreme outliers. Both the frequency and the density histograms indicate the presence of a couple of very high peaks for wines with lower alcohol content. Several gaps are also present.

Alcohol levels

```
Create a new factor variable for alcohol levels, using quartiles:

        low_alcohol         lowest 25% of alcohol levels
        mod.low_alcohol     25%-50% of alcohol levels
        mod.high_alcohol    50%-75% of alcohol levels
        high_alcohol        75% to max alcohol levels
```


Alcohol Levels Distribution
low_alcohol
297
high_alcohol
407
mod.high_alcohol
396
mod.low_alcohol
499

Quality



The quality is the only categorical variable we have analyzed so far. The wines in this dataset have quality ranging from 3 to 8, in increments of 1. From the bar chart we see that the vast majority (1319 samples out of 1599 samples) were assigned quality grades of 5 or 6, thus more or less average quality.

Univariate Analysis

What is the structure of your dataset?

The dataset consists of 1599 samples of a certain variety of red wine, 'Vinho Verde', from northern Portugal. Each sample is denoted by a label X, and it is described by 11 physicochemical properties (numerical variables). A quality that ranges from 3 to 8 is also associated to the sample. The dataset is tidy and has no missing values.

Some of my initial observations:

- some of the attributes (such as residual sugar, chlorides, free and total sulfur dioxide) have

wide ranges of values;

- the density and the pH do not vary too much among these samples;
- the mean alcohol level of 10.42% is lower than the usual alcohol levels found in red wines;
- the quality varies between 0 and 10, but in this dataset it takes only values between 3 and 8, with less than 20 wines with quality rating of 8, most of the wines have average quality ratings.

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

The main feature of interest is the quality, in particular how the pysicochemical attributes influence the quality of the sample. I am also interested in related questions, such as if sweeter wines or if higher alcohol percentage wines receive better ratings. Also, I think it would be informative to see what other relations could be found among the physicochemical properties of the wine.

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

At this point, without looking at specific statistics I think that the alcohol, the residual sugar and the pH would be significant in determining the quality of the wine.

In [Cortez et al., 2009] a model to predict the quality of the wine is built. I find it intriguing that this model gives sulphates as the main attribute that would influence the ratings of the wine. I plan to look into the relation between sulphates and quality.

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

Yes I created three factor variables: sugar.levels, pH.levels and alcohol.levels.

The sugar.levels variable uses the median to divide the wines in two groups (low_sugar and high_sugar). The numerous extreme outliers for residual sugar are given the extra.sugar level.

The other two categorical variables, pH.levels and alcohol.levels, divide the wines into four groups each, based on the 5 numbers statistics for pH and alcohol respectively.

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?

First I will discuss the following attributes: fixed.acidity, volatile.acidity, residual.sugar, chlorides, free sulfur.dioxide, total sulfur.dioxide, sulphates and alcohol. All these distributions are skewed to the right, with numerous outliers. I find it interesting that in most cases (except for chlorides) most of the outliers are located in the tail of the distribution, thus they tend to have larger values than average of the analyzed chemical. Although each distribution shows the presence of the outliers, the number of extreme outliers is usually small (less than 5). However, there are three exceptions: residual.sugar (88 extreme outliers), chlorides (67 extreme outliers) and sulphates (16 extreme outliers).

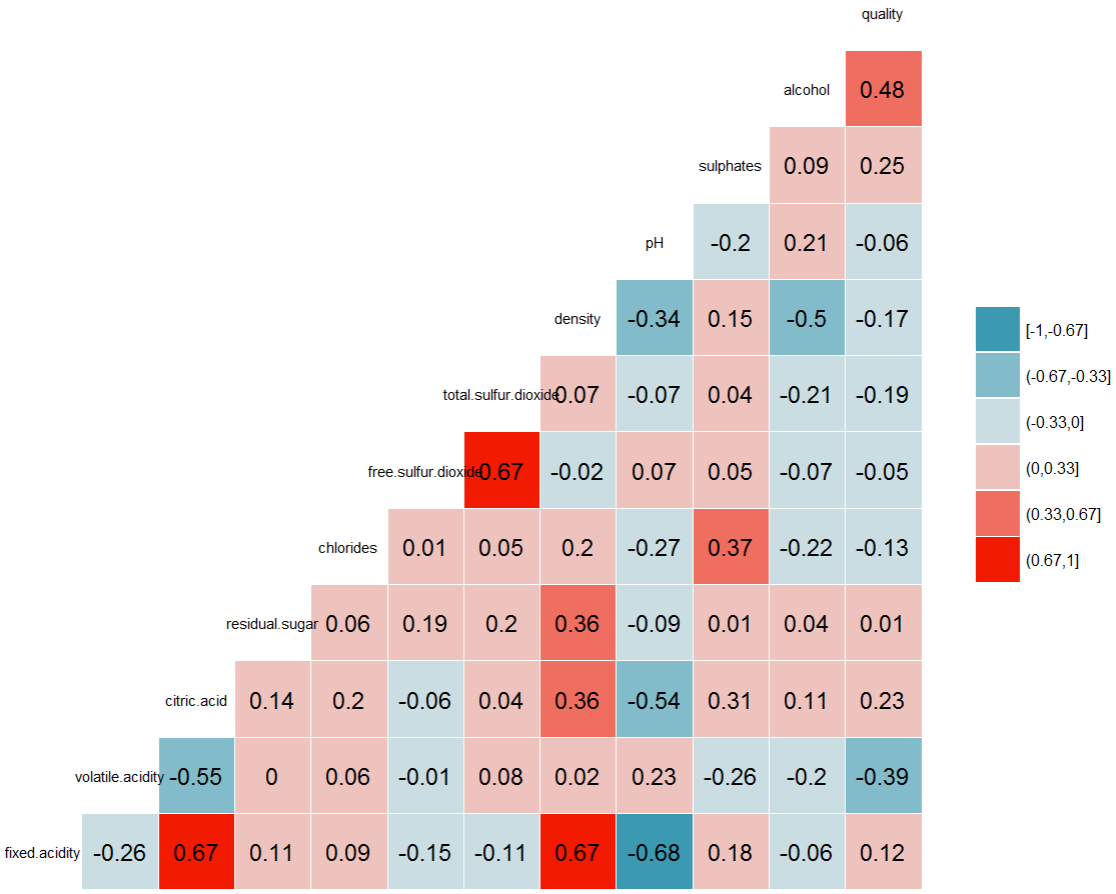
Among the remaining attributes, two of the distributions are normal, density and pH. This is quite unusual for real sets of data.

Regarding the quality. I notice that very few wines receive a grading of 8 (about 18) and none are rated as 9 or 10. At the other end there are only 63 samples with quality 3 or 4. This leaves us with the vast majority of wines having an average quality. Thus, I think that it will be harder to distinguish among the attributes which ones have a significant impact on the quality.

The dataset is already tidy. I did form subsets of extreme outliers for residual.sugar, chlorides and sulfates. I also created subsets of the data that do not contain these extreme outliers. I plan to investigate further these sets of extreme outliers, to determine if there are unusually good or unusually bad wines among them.

I created density plots for each of the numerical variables, in which I also adjusted the x-axis limits, in order to eliminate some of the tails and get a better look at the main data.

Bivariate Plots Section

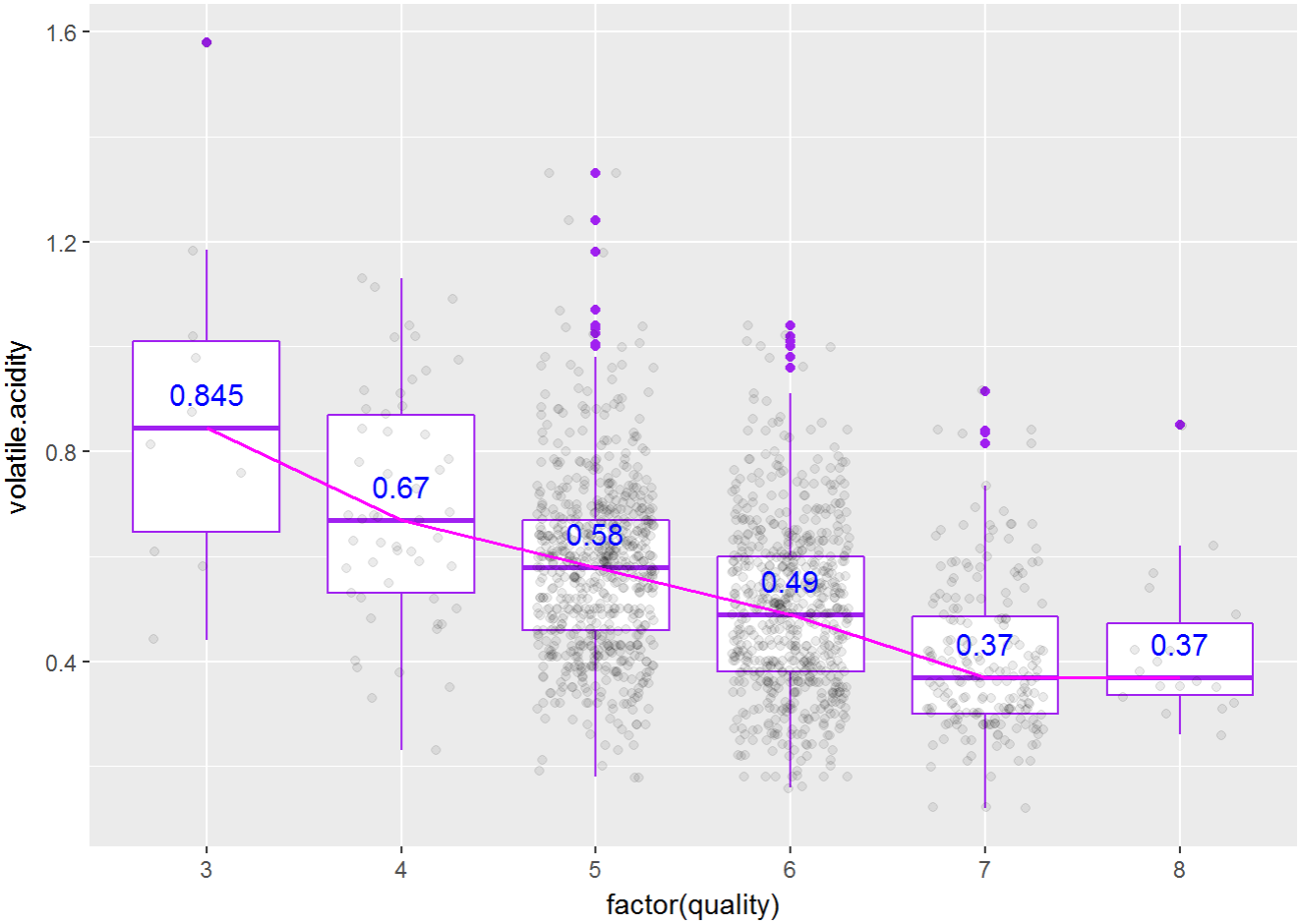


The main feature of interest is the quality. From the correlation matrix I notice that alcohol has strongest correlation of 0.48 with quality, followed by sulphates (correlation 0.25) and citric acid (correlation of 0.23). Also volatile.acidity has negative stronger correlation coefficient (of -0.39) with quality. I am surprised to learn that the residual sugar and the pH have very small influence on quality.

Quality and Various Attributes

The boxplots below show how various attributes behave when divided into subsets based on quality. A line connecting the medians is added and the numerical values of these median values are included in most graphs.

Quality and Volatile Acidity




```
##
##  Pearson's product-moment correlation
##
## data:  rwine$quality and rwine$volatile.acidity
## t = -16.954, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4313210 -0.3482032
## sample estimates:
##       cor
## -0.3905578
```

```
## Volatile acidity descriptive statistics for quality = 3
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4400  0.6475  0.8450  0.8845  1.0100  1.5800
```

```
## Volatile descriptive statistics for quality = 4
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##   0.230   0.530   0.670   0.694   0.870   1.130
```

```
## Volatile acidity descriptive statistics for quality = 5
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.180   0.460   0.580   0.577   0.670   1.330
```

```
## Volatile acidity descriptive statistics for quality = 6
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1600  0.3800  0.4900  0.4975  0.6000  1.0400
```

```
## Volatile acidity descriptive statistics for quality = 7
```

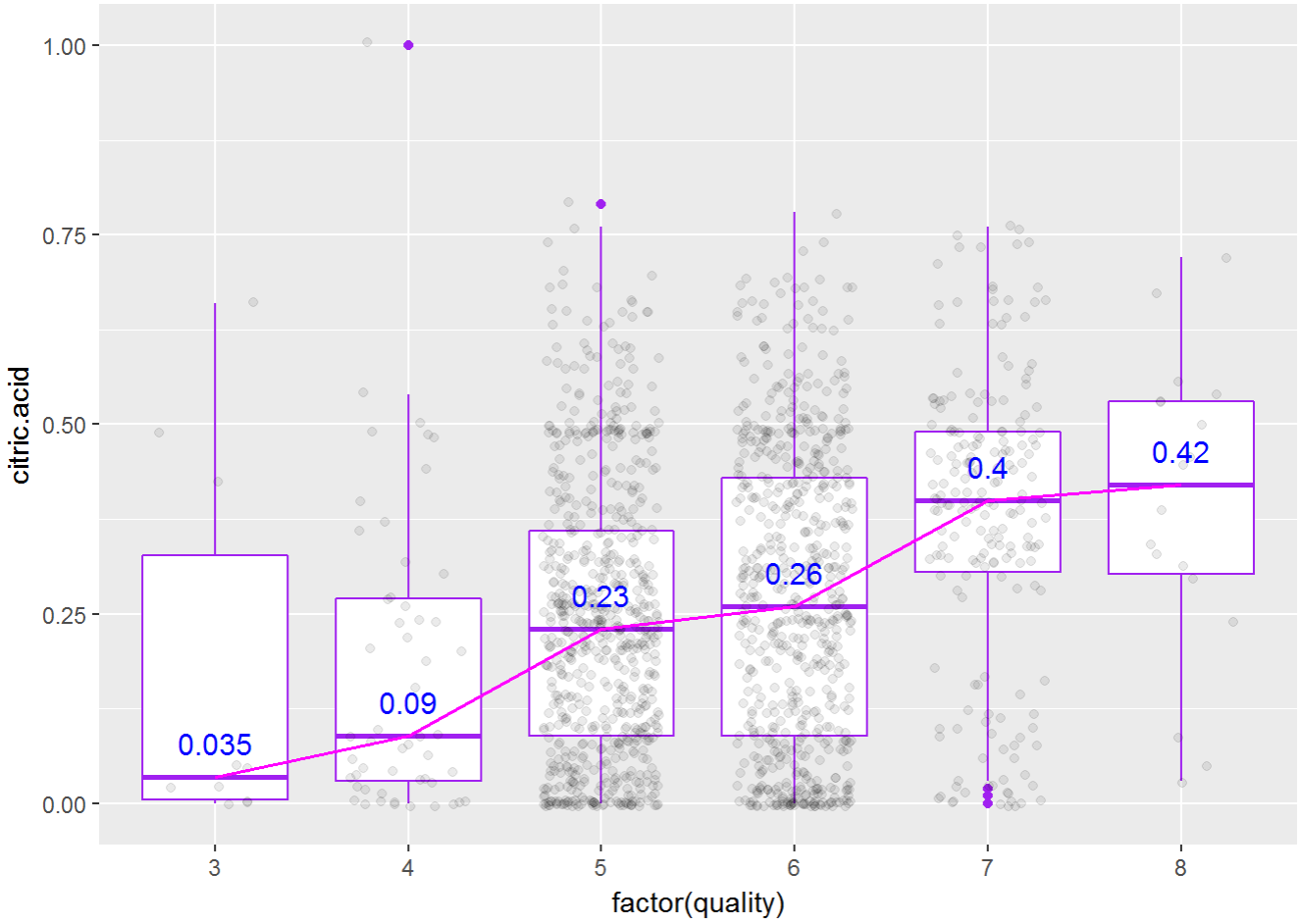
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3000  0.3700  0.4039  0.4850  0.9150
```

```
## Volatile acidity descriptive statistics for quality = 8
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2600  0.3350  0.3700  0.4233  0.4725  0.8500
```

The trend is clear from both the graphical description and the statistics performed on the data. The volatile acidity decreases with the quality, from quality 3 to 7. The better wines with quality 7 and 8 have similar levels of volatile acidity. Most of the outliers are concentrated at levels 5 and 6, where the majority of the samples are found.

Quality and Citric Acid




```
##
##  Pearson's product-moment correlation
##
## data:  rwine$quality and rwine$citric.acid
## t = 9.2875, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1793415 0.2723711
## sample estimates:
##       cor
## 0.2263725
```

```
## Citric acid descriptive statistics for quality = 3
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0050  0.0350  0.1710  0.3275  0.6600
```

```
## Citric acid descriptive statistics for quality = 4
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##  0.0000  0.0300  0.0900  0.1742  0.2700  1.0000
```

```
## Citric acid descriptive statistics for quality = 5
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0900  0.2300  0.2437  0.3600  0.7900
```

```
## Citric acid descriptive statistics for quality = 6
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0900  0.2600  0.2738  0.4300  0.7800
```

```
## Citric acid descriptive statistics for quality = 7
```

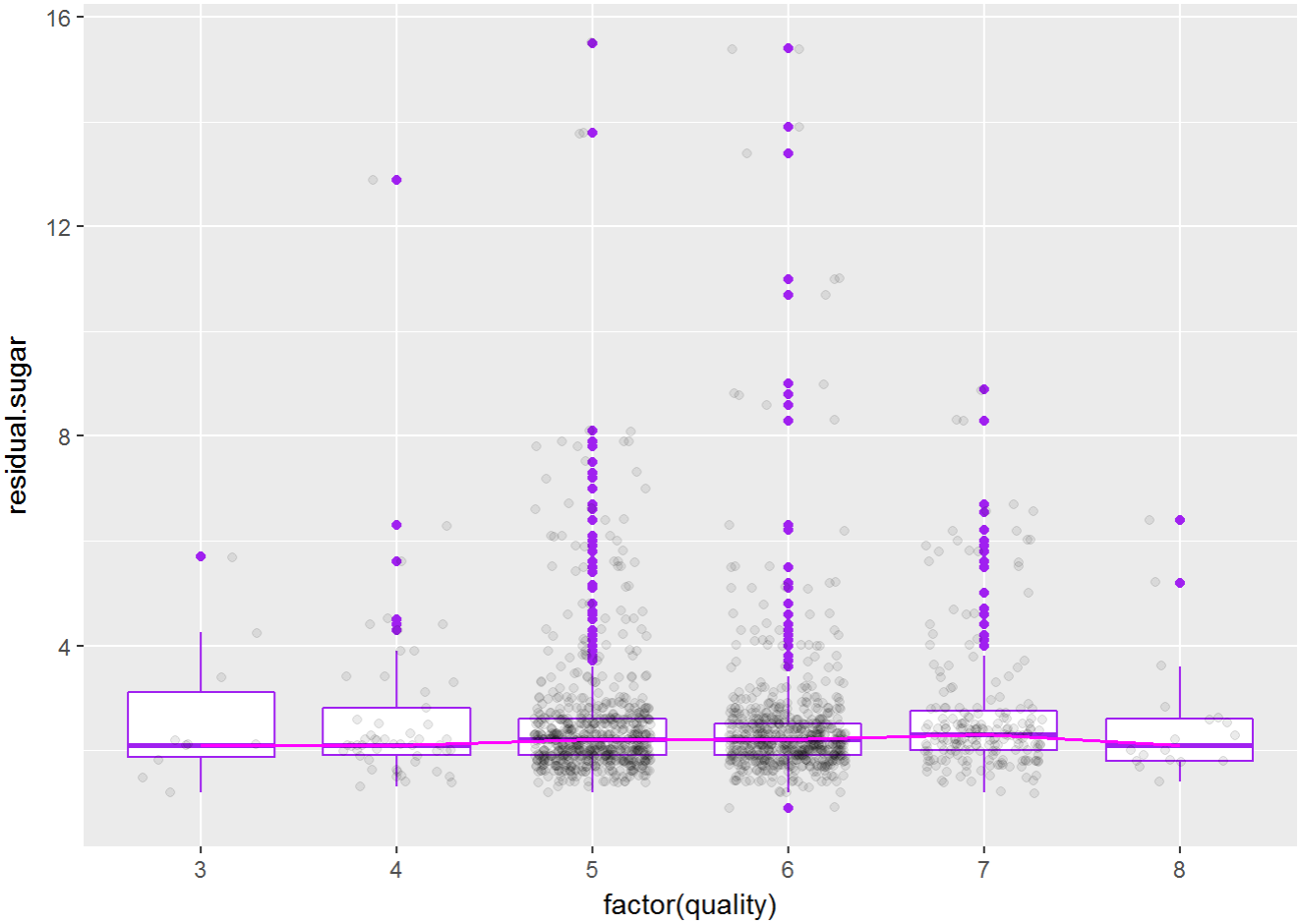
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.3050  0.4000  0.3752  0.4900  0.7600
```

```
## Citric acid descriptive statistics for quality = 8
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0300  0.3025  0.4200  0.3911  0.5300  0.7200
```

The citric acid increases from wines of quality 3 to the wines of quality 8. Furthermore we see a sharp increase in the median value from 0.26 (qualities 5 and 6) to 0.40 (qualities 7 and 8). The better wines have citric acid values that vary less than those in the average quality wines.

Quality and Residual Sugar




```
##
##  Pearson's product-moment correlation
##
## data:  rwine$quality and rwine$residual.sugar
## t = 0.5488, df = 1597, p-value = 0.5832
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.03531327  0.06271056
## sample estimates:
##        cor
## 0.01373164
```

```
## Residual sugar descriptive statistics for quality = 3
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.200   1.875   2.100   2.635   3.100   5.700
```

```
## Residual sugar descriptive statistics for quality = 4
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##   1.300   1.900   2.100   2.694   2.800  12.900
```

```
## Residual sugar descriptive statistics for quality = 5
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.200   1.900   2.200   2.529   2.600  15.500
```

```
## Residual sugar descriptive statistics for quality = 6
```

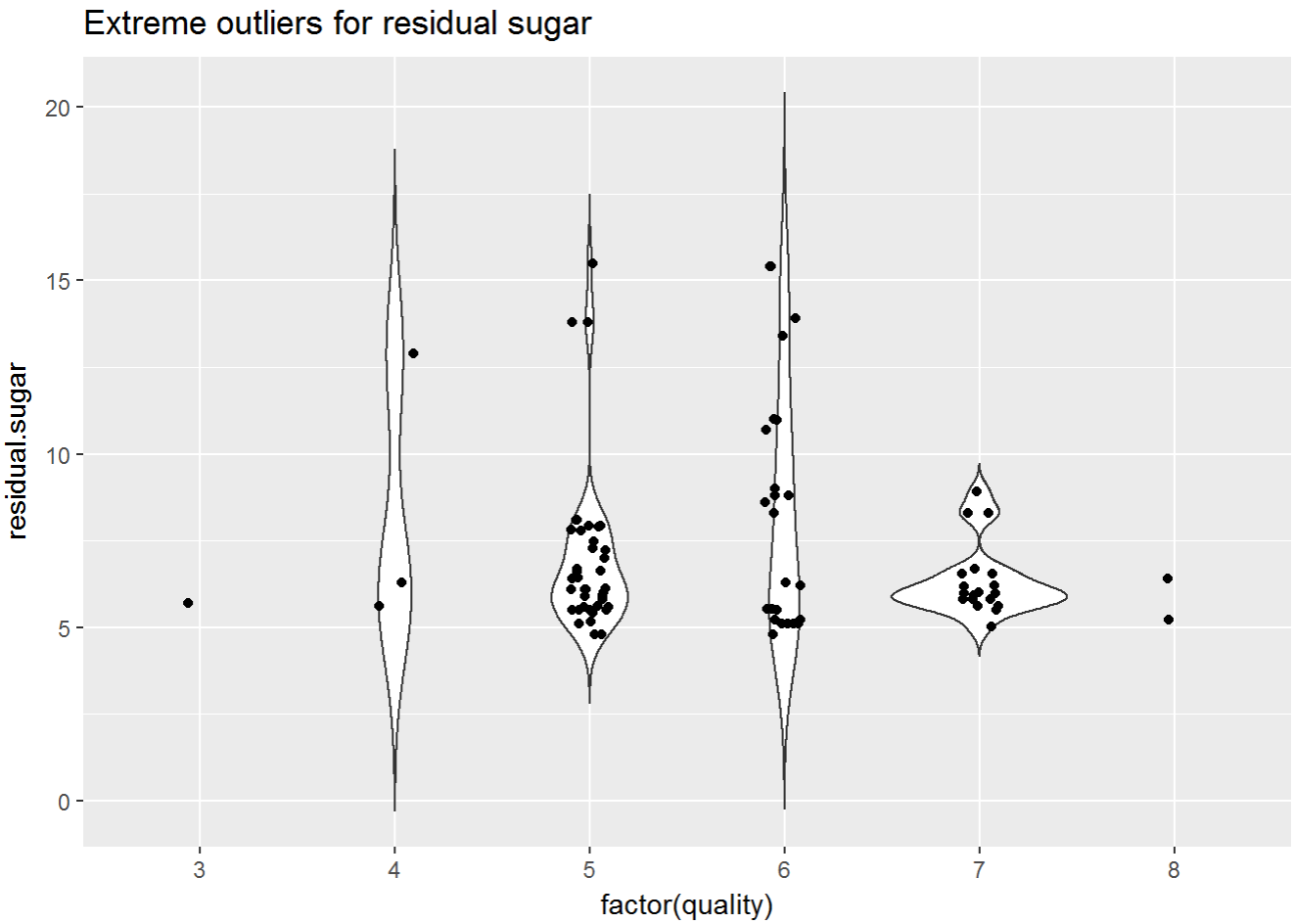
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.477   2.500  15.400
```

```
## Residual sugar descriptive statistics for quality = 7
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.200   2.000   2.300   2.721   2.750   8.900
```

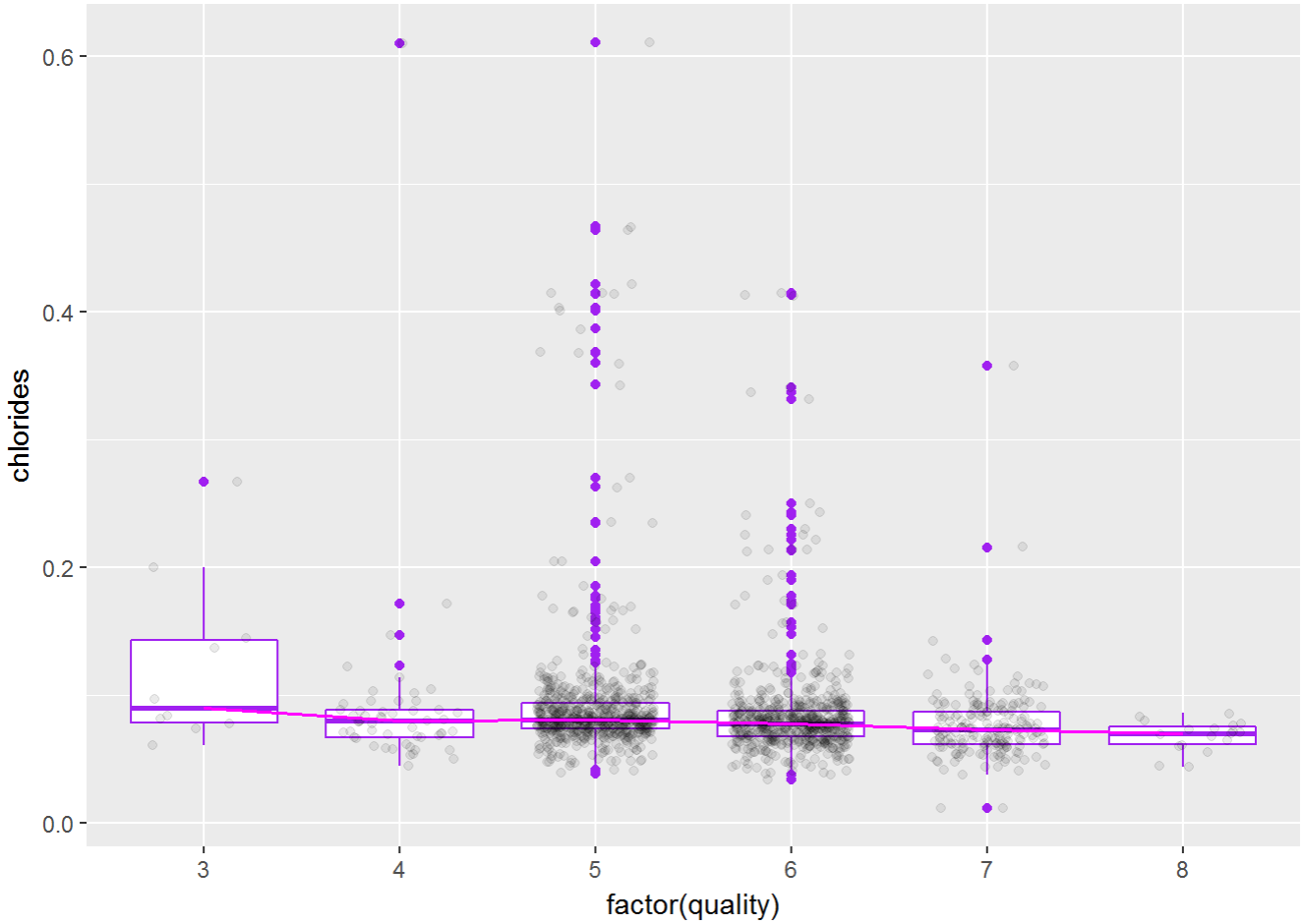
```
## Residual sugar descriptive statistics for quality = 8
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.400   1.800   2.100   2.578   2.600   6.400
```



The residual sugar values seem to be quite evenly distributed among the quality levels, as long as we do not take into account the outliers. Most of these outliers lie in the groups of quality 5, 6 and 7, as can be easily observed in the violin plot. The distribution of the extreme outliers does not indicate that the extra sugar is a deciding factor in determining the wine quality. The concentration of outliers in the average quality categories might just be due to the fact that these groups are more numerous.

Quality and Chlorides




```
##
##  Pearson's product-moment correlation
##
## data:  rwine$quality and rwine$chlorides
## t = -5.1948, df = 1597, p-value = 2.313e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.17681041 -0.08039344
## sample estimates:
##        cor
## -0.1289066
```

```
## Chlorides descriptive statistics for quality = 3
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0610  0.0790  0.0905  0.1225  0.1430  0.2670
```

```
## Chlorides descriptive statistics for quality = 4
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
## 0.04500 0.06700 0.08000 0.09068 0.08900 0.61000
```

```
## Chlorides descriptive statistics for quality = 5
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.03900 0.07400 0.08100 0.09274 0.09400 0.61100
```

```
## Chlorides descriptive statistics for quality = 6
```

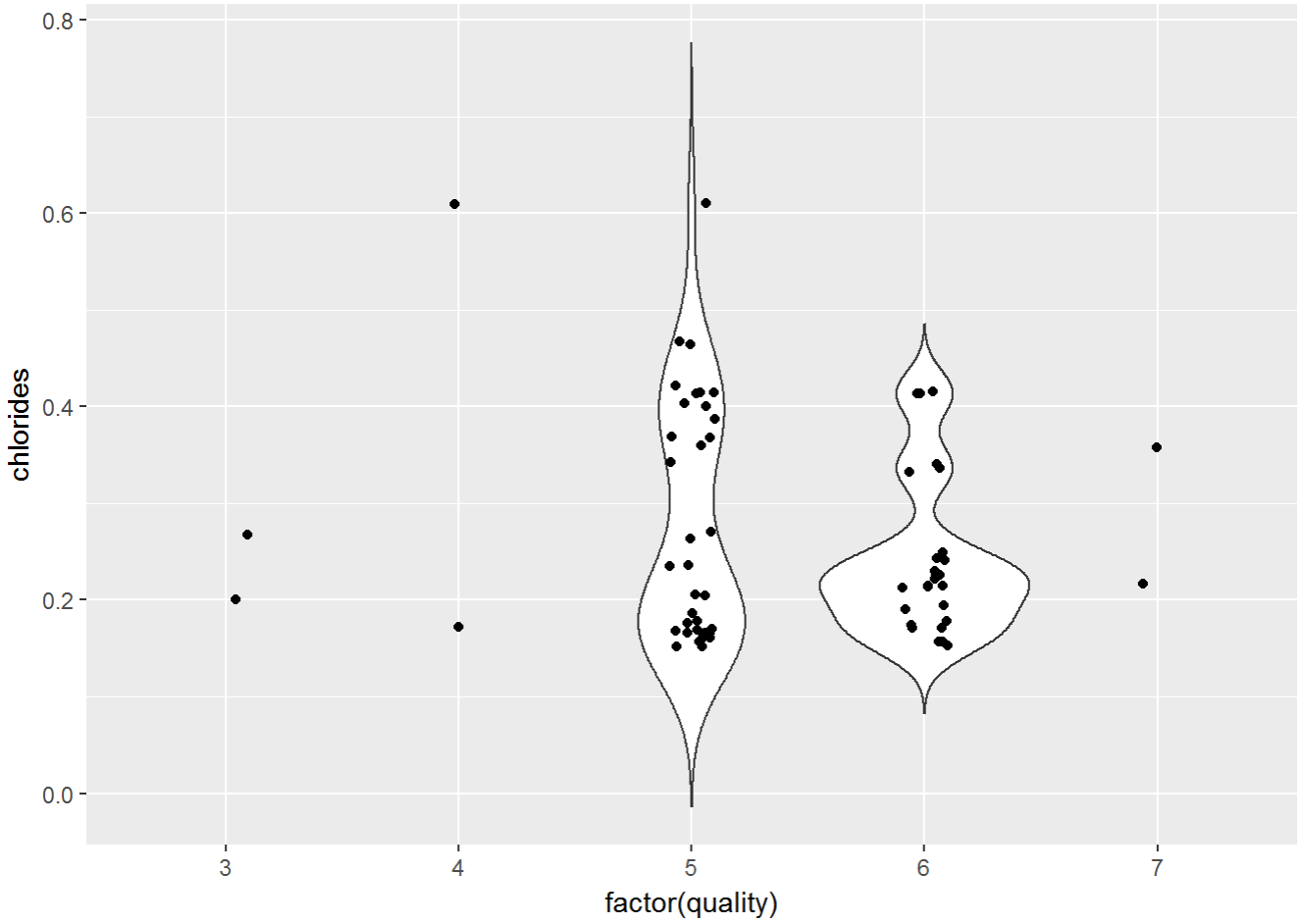
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.03400 0.06825 0.07800 0.08496 0.08800 0.41500
```

```
## Chlorides descriptive statistics for quality = 7
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.06200 0.07300 0.07659 0.08700 0.35800
```

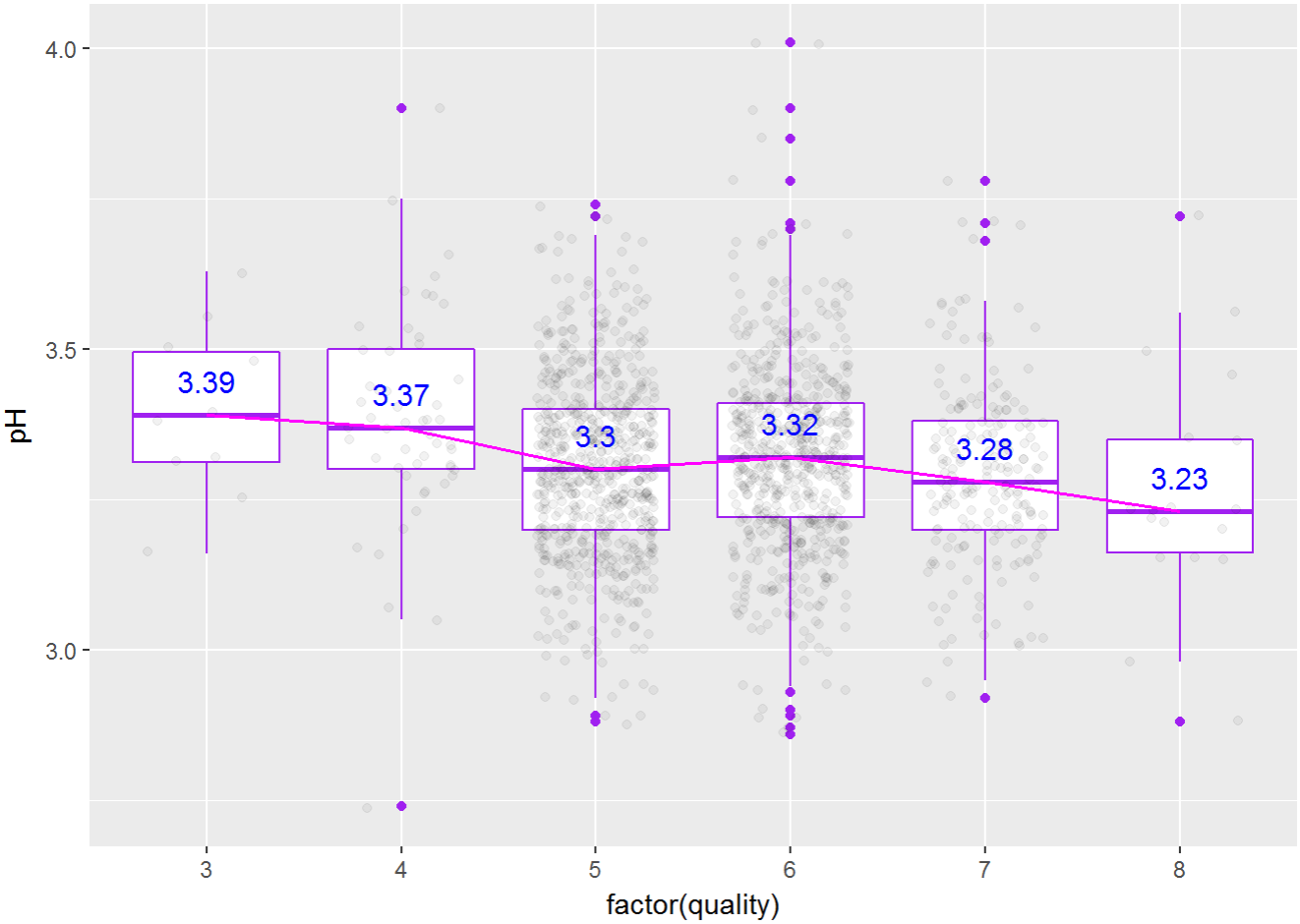
```
## Chlorides descriptive statistics for quality = 8
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.04400 0.06200 0.07050 0.06844 0.07550 0.08600
```



The chlorides level decreases with the quality. The interquantile range decreases significantly with quality. Chlorides is one of the attributes with a large number of extreme outliers, which are mostly distributed among the wines of qualities 5 and 6. Again, it is possible that the concentration of outliers in the categories 5 and 6 might be just due to the fact that these groups are larger.

Quality and pH




```
##
##  Pearson's product-moment correlation
##
## data:  rwine$quality and rwine$pH
## t = -2.3109, df = 1597, p-value = 0.02096
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.106451268 -0.008734972
## sample estimates:
##         cor
## -0.05773139
```

```
## pH descriptive statistics for quality = 3
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.160   3.312   3.390   3.398   3.495   3.630
```

```
## pH descriptive statistics for quality = 4
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##   2.740   3.300   3.370   3.382   3.500   3.900
```

```
## pH descriptive statistics for quality = 5
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.880   3.200   3.300   3.305   3.400   3.740
```

```
## pH descriptive statistics for quality = 6
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.860   3.220   3.320   3.318   3.410   4.010
```

```
## pH descriptive statistics for quality = 7
```

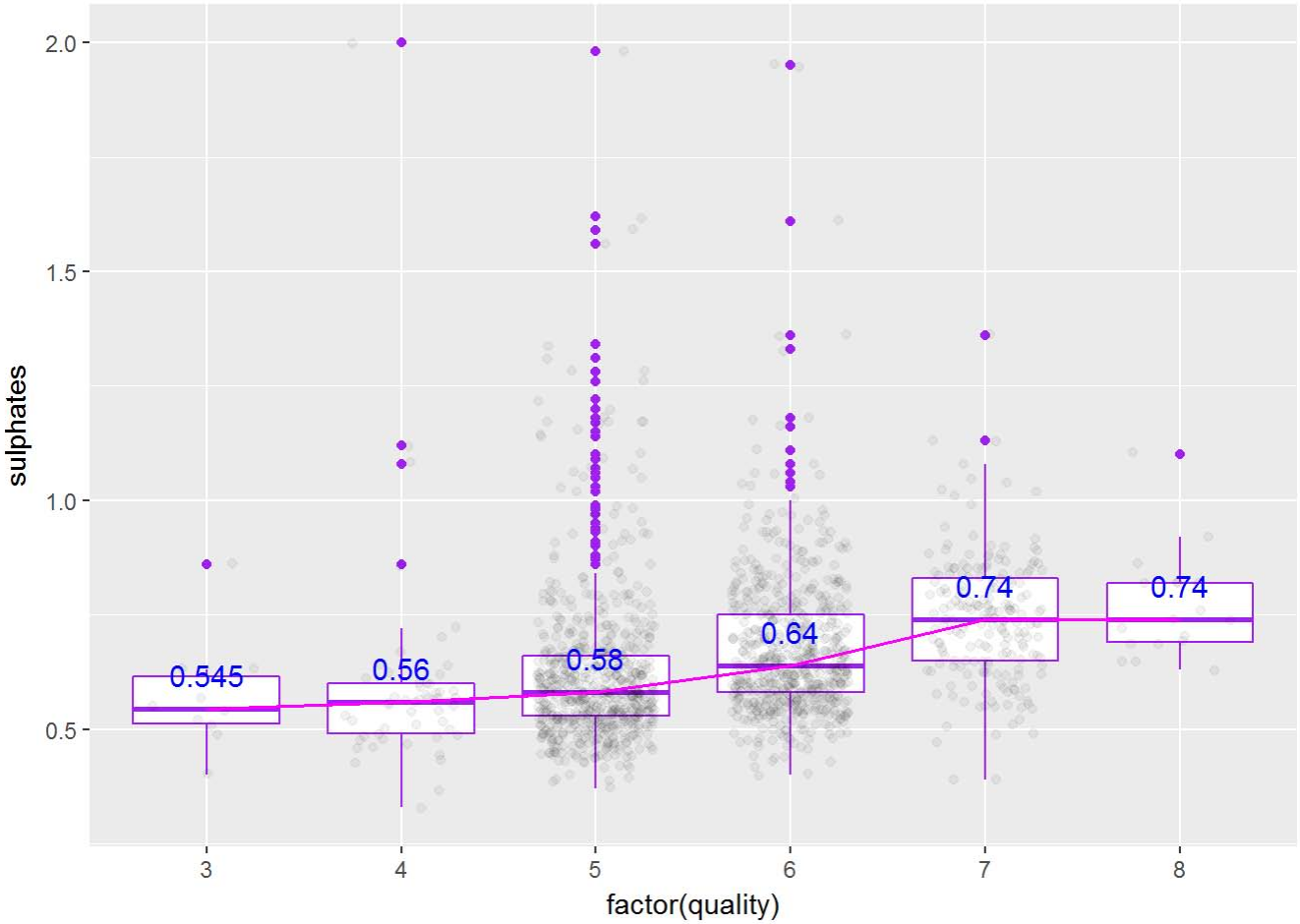
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.920   3.200   3.280   3.291   3.380   3.780
```

```
## pH descriptive statistics for quality = 8
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.880   3.163   3.230   3.267   3.350   3.720
```

The pH has an interesting distribution across quality. The wines with quality 3-4 have similar median pH levels; the same is true for qualities 5-6 and 7-8. The pH decreases with quality, the better wines (qualities 7 and 8) have median pH of 3.28 and 3.23 respectively.

Quality and Sulphates




```
##
##  Pearson's product-moment correlation
##
## data:  rwine$quality and rwine$sulphates
## t = 10.38, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2049011 0.2967610
## sample estimates:
##       cor
## 0.2513971
```

```
## Sulphates descriptive statistics for quality = 3
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4000  0.5125  0.5450  0.5700  0.6150  0.8600
```

```
## Sulphates descriptive statistics for quality = 4
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##  0.3300  0.4900  0.5600  0.5964  0.6000  2.0000
```

```
## Sulphates descriptive statistics for quality = 5
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.370   0.530   0.580   0.621   0.660   1.980
```

```
## Sulphates descriptive statistics for quality = 6
```

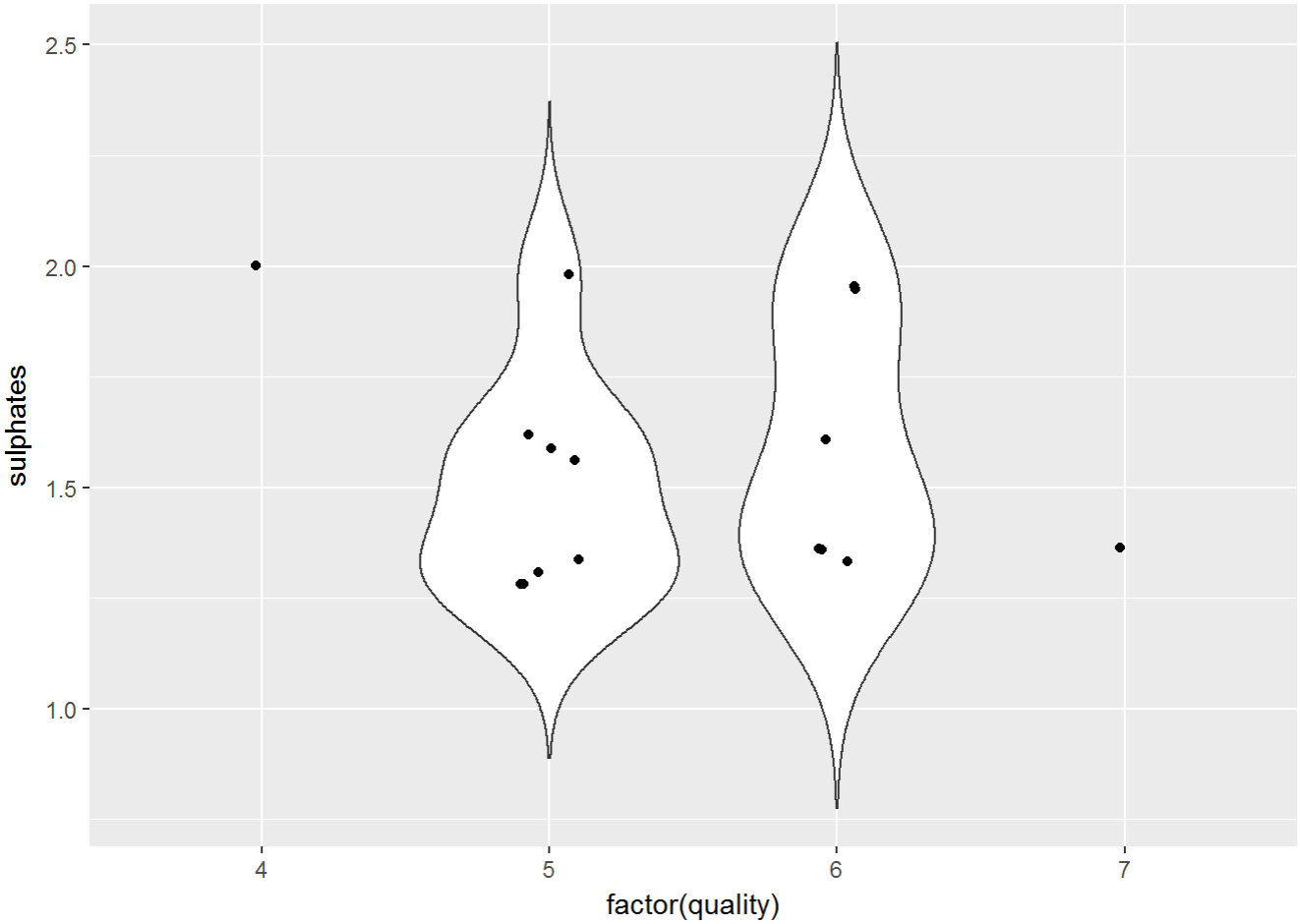
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4000  0.5800  0.6400  0.6753  0.7500  1.9500
```

```
## Sulphates descriptive statistics for quality = 7
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3900  0.6500  0.7400  0.7413  0.8300  1.3600
```

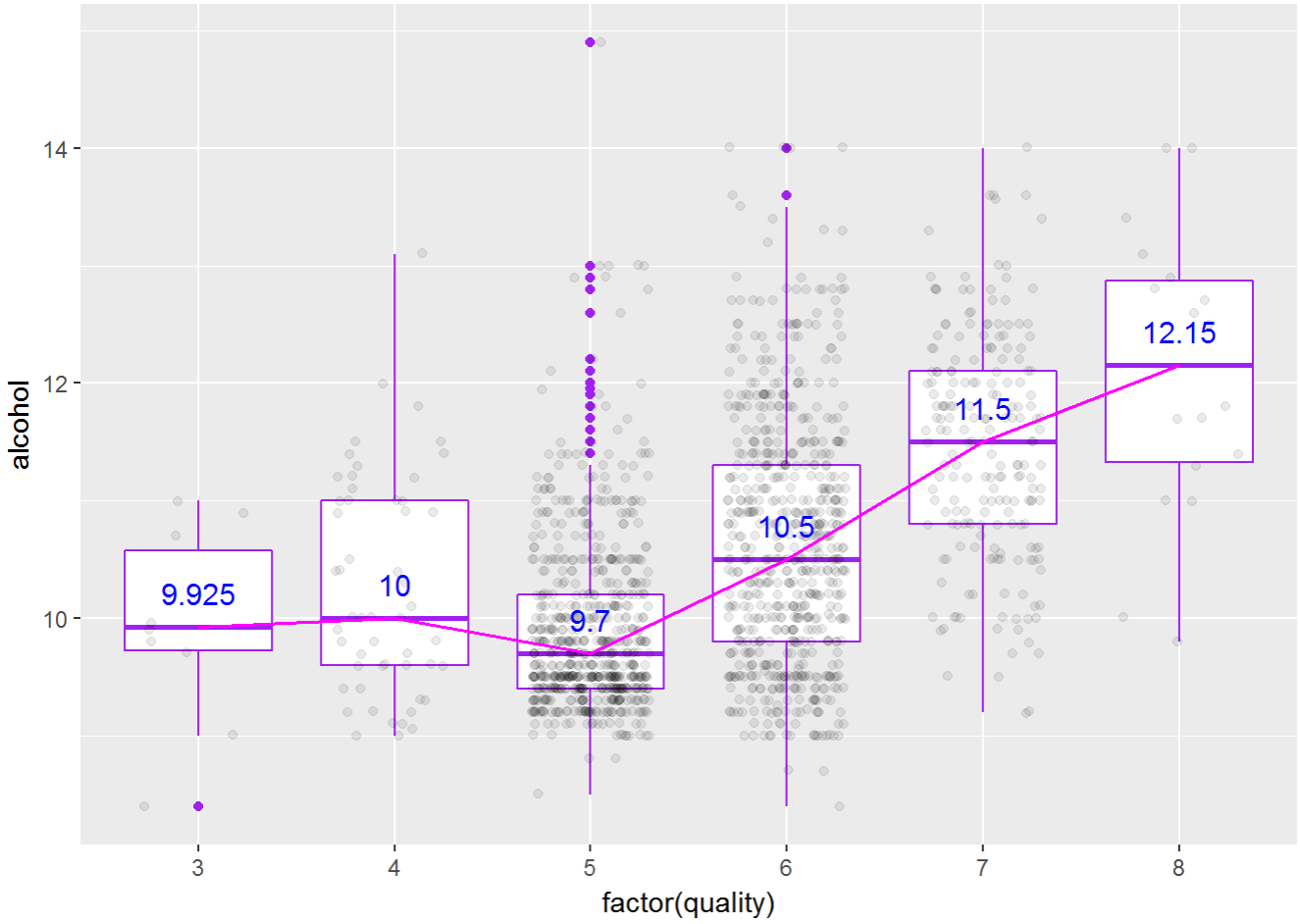
```
## Sulphates descriptive statistics for quality = 8
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.6300  0.6900  0.7400  0.7678  0.8200  1.1000
```



Sulphates increase with quality; while lower quality wines (3 and 4) have similar amounts of sulphates in average, the wines of quality 6 have in average more sulphates than those of quality 5. A large increment is observed in the median sulphates levels as we move from quality 6 to quality 7. The better wines (quality 7 and 8) have the same median levels of sulphates. Most of the outliers can be found among the wines of quality 5. There are 16 extreme outliers, which are also depicted in a violin plot. Except for two of them, the remaining 14 are equally distributed among the wines of qualities 5 and 6.

Quality and Alcohol




```
##
##  Pearson's product-moment correlation
##
## data:  rwine$quality and rwine$alcohol
## t = 21.639, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4373540 0.5132081
## sample estimates:
##       cor
## 0.4761663
```

```
## Alcohol descriptive statistics for quality = 3
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.400   9.725   9.925   9.955  10.575  11.000
```

```
## Alcohol descriptive statistics for quality = 4
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##    9.00    9.60   10.00   10.27   11.00   13.10
```

```
## Alcohol descriptive statistics for quality = 5
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     8.5     9.4     9.7     9.9    10.2    14.9
```

```
## Alcohol descriptive statistics for quality = 6
```

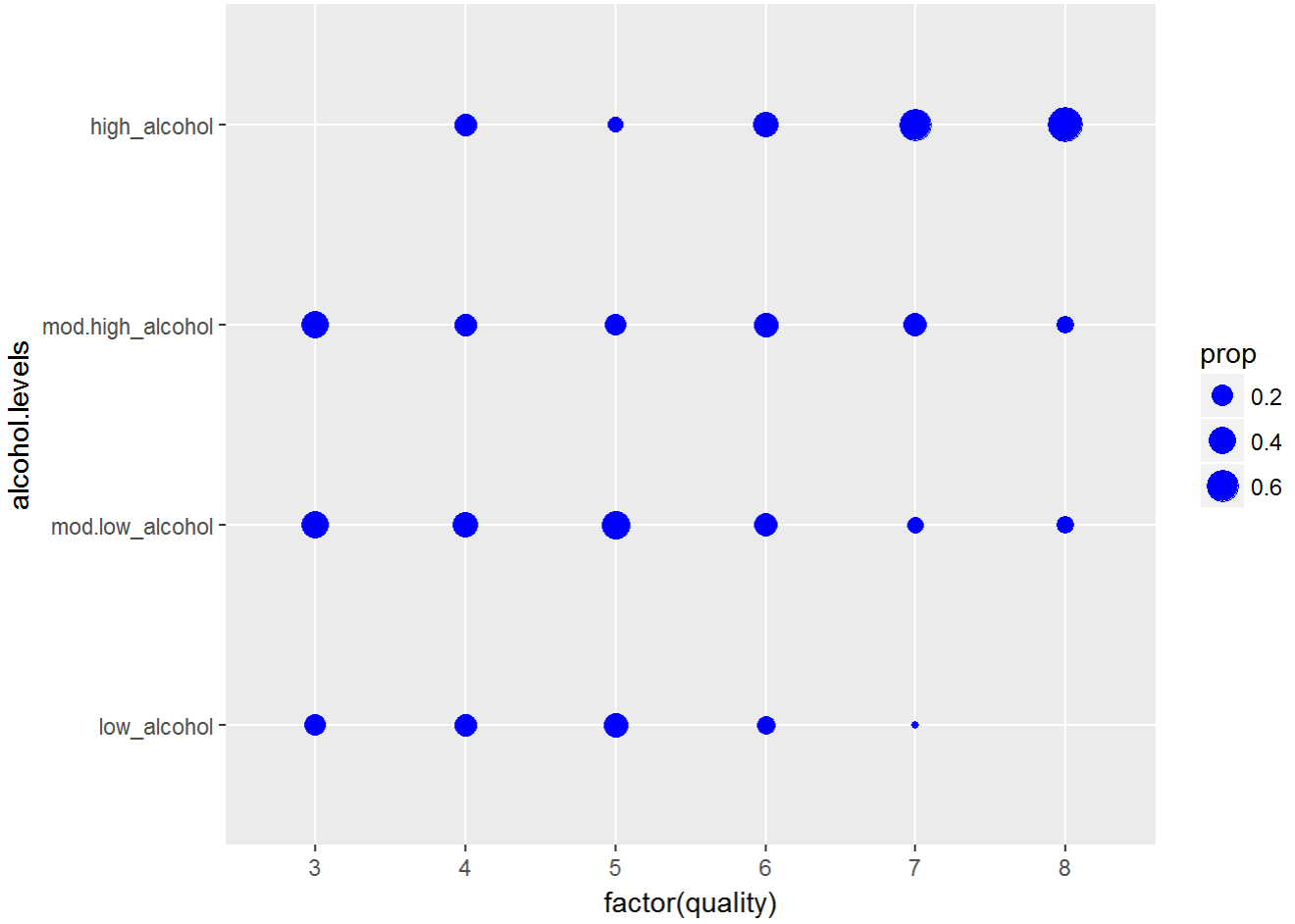
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.80   10.50   10.63   11.30   14.00
```

```
## Alcohol descriptive statistics for quality = 7
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.20   10.80   11.50   11.47   12.10   14.00
```

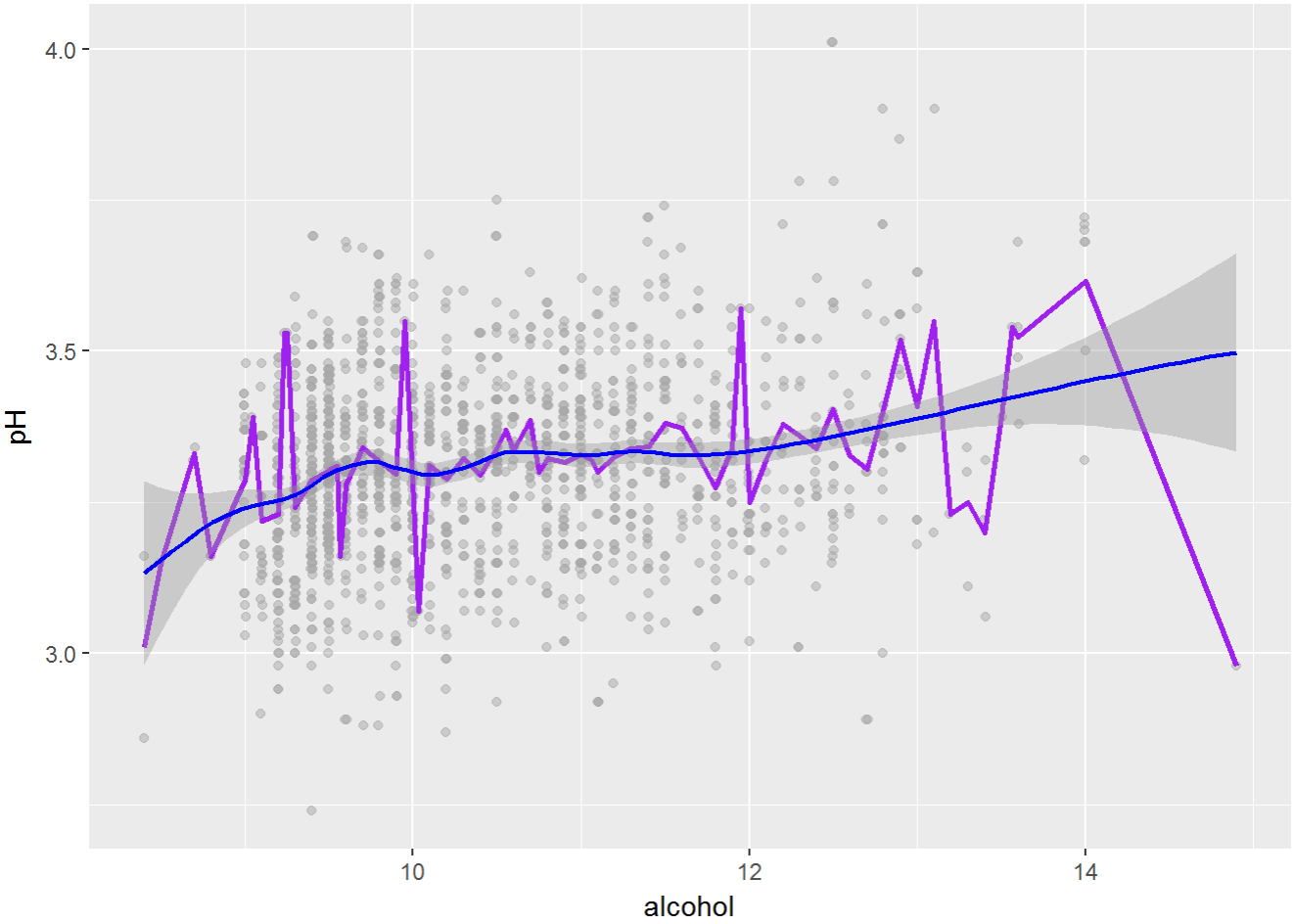
```
## Alcohol descriptive statistics for quality = 8
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.80   11.32   12.15   12.09   12.88   14.00
```

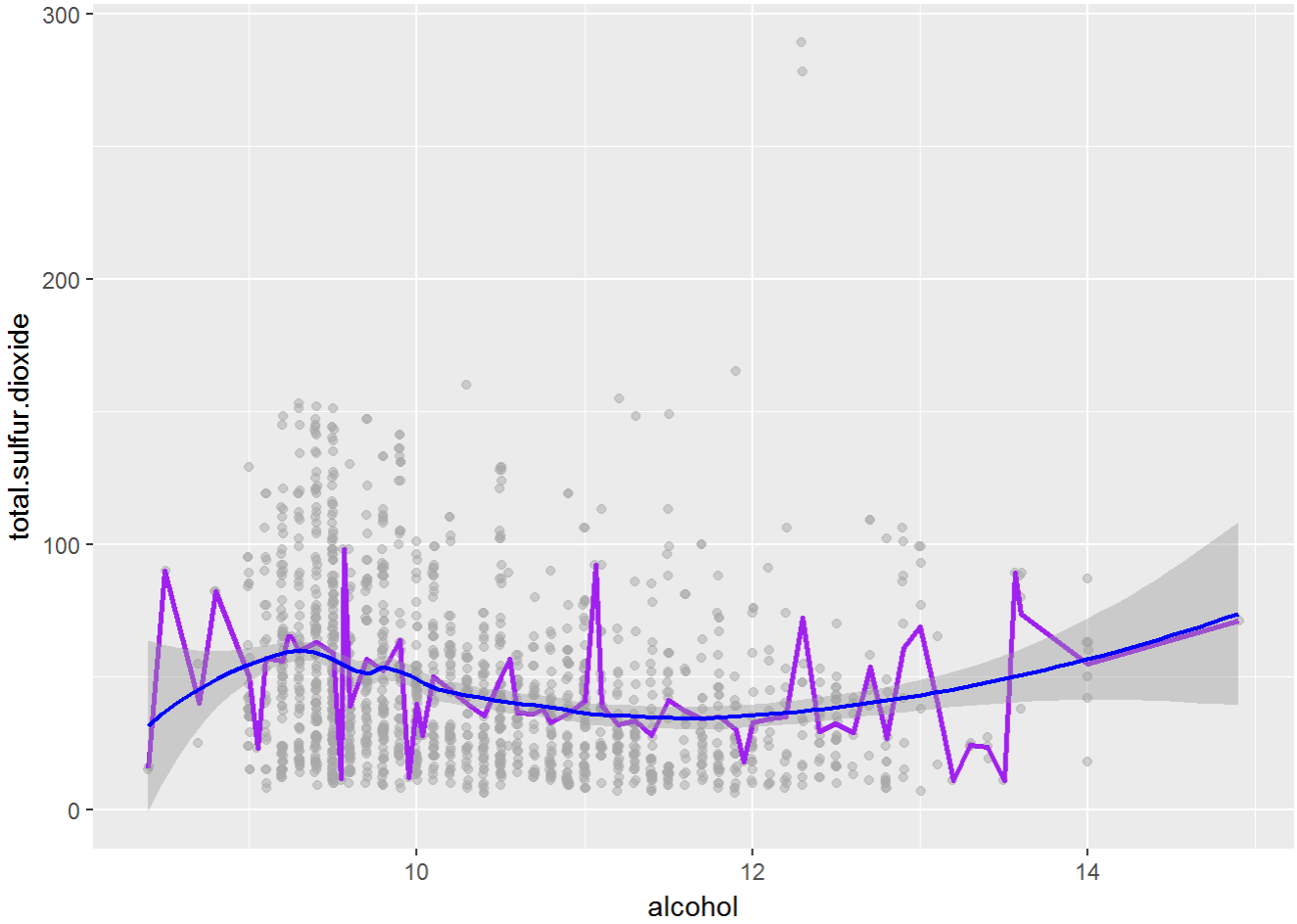


The median alcohol levels have an interesting pattern. Levels 3 and 4 have similar alcohol levels. Clearly the level of alcohol sharply increases from quality 5 to quality 8. The unexpected value is the median level for wines of quality 5, which is much lower than the median alcohol level for quality 4 wines. Most of the outlier values can be found at this quality level and the interquantile range is the smallest among the corresponding ranges. Also, if we look at the proportions of high alcohol wines among the quality levels, we notice that these are larger for the better wines.

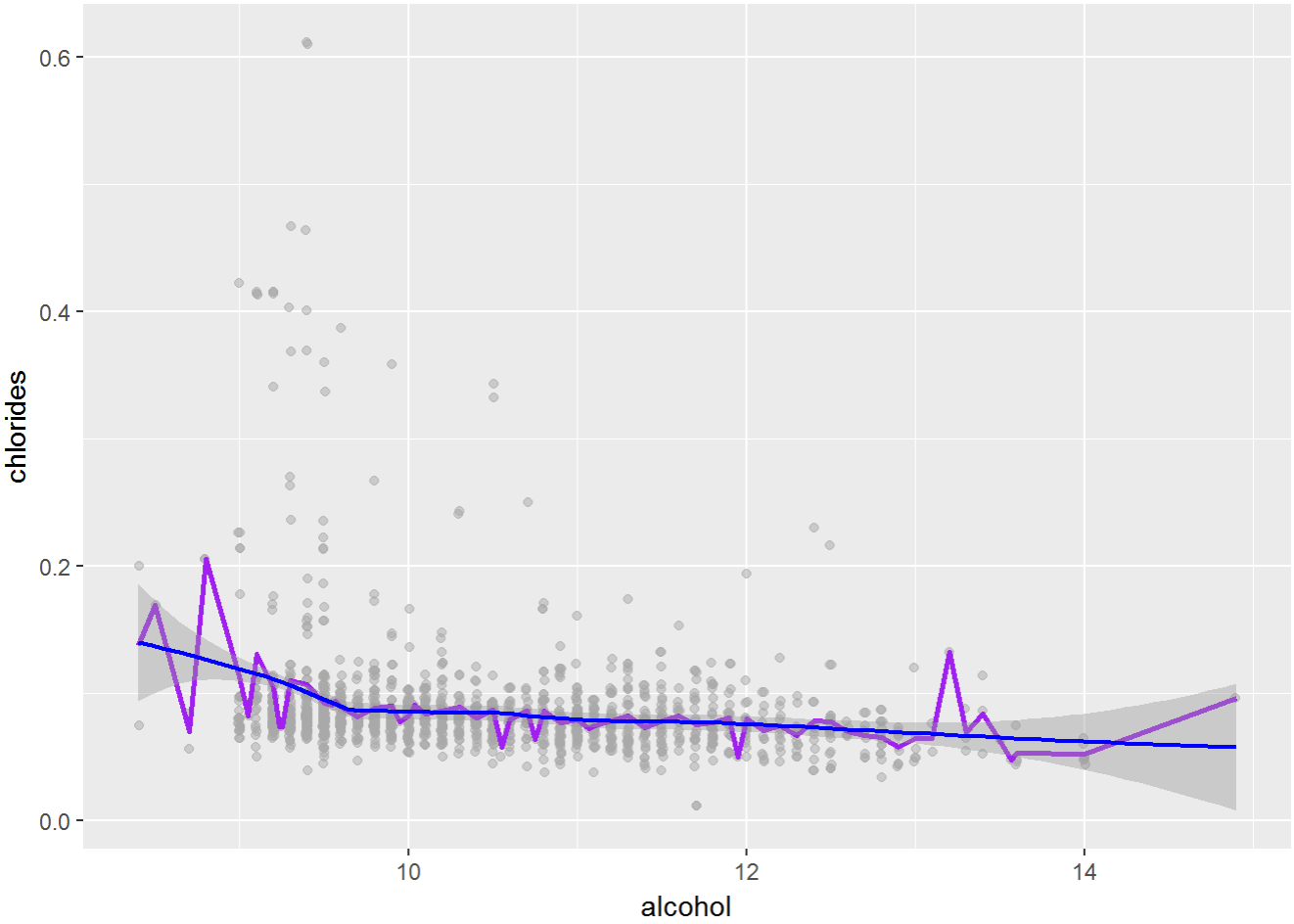
Alcohol and Other Attributes



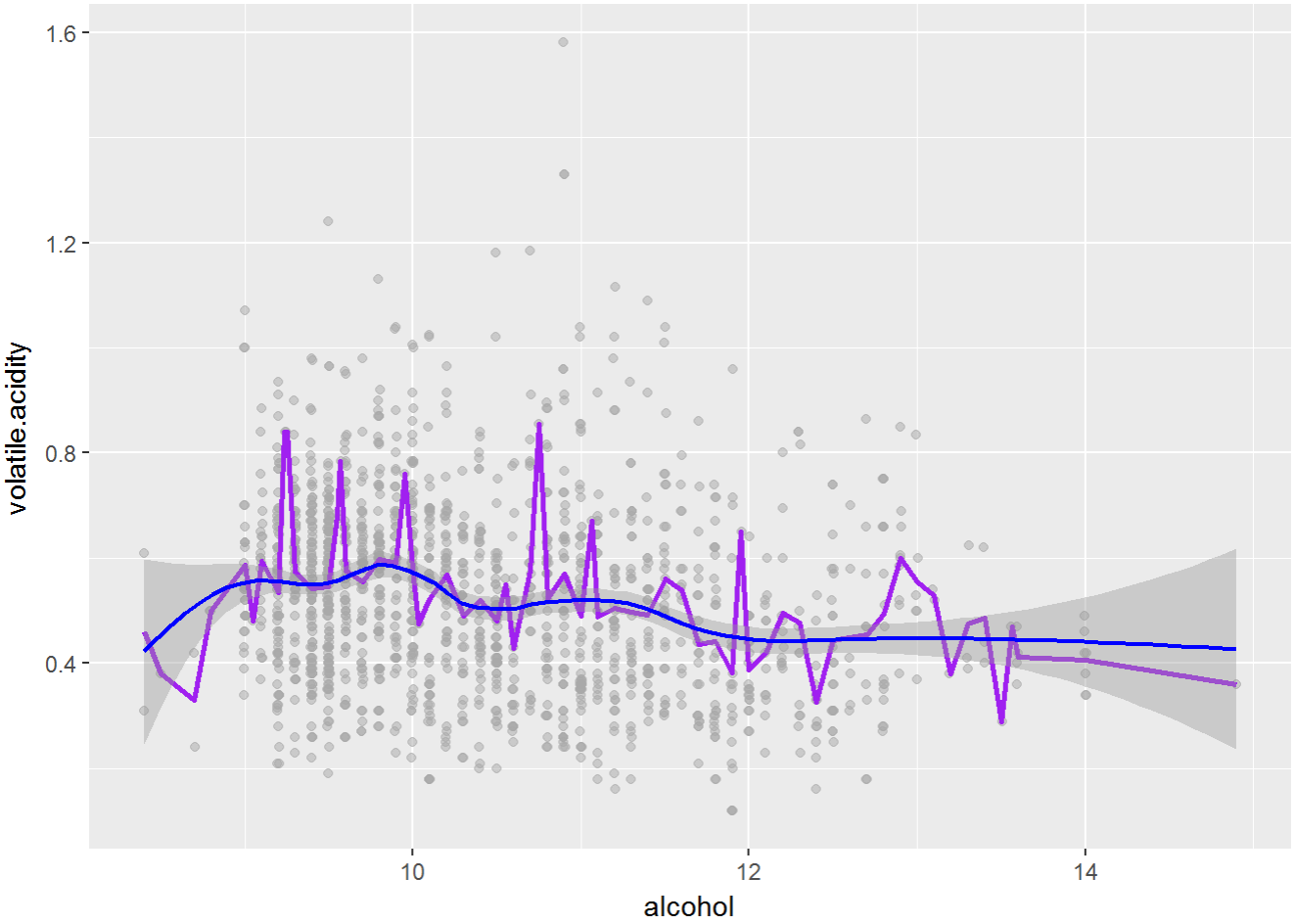

```
##
##  Pearson's product-moment correlation
##
## data:  rwine$alcohol and rwine$pH
## t = 8.397, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1582061 0.2521123
## sample estimates:
##       cor
## 0.2056325
```



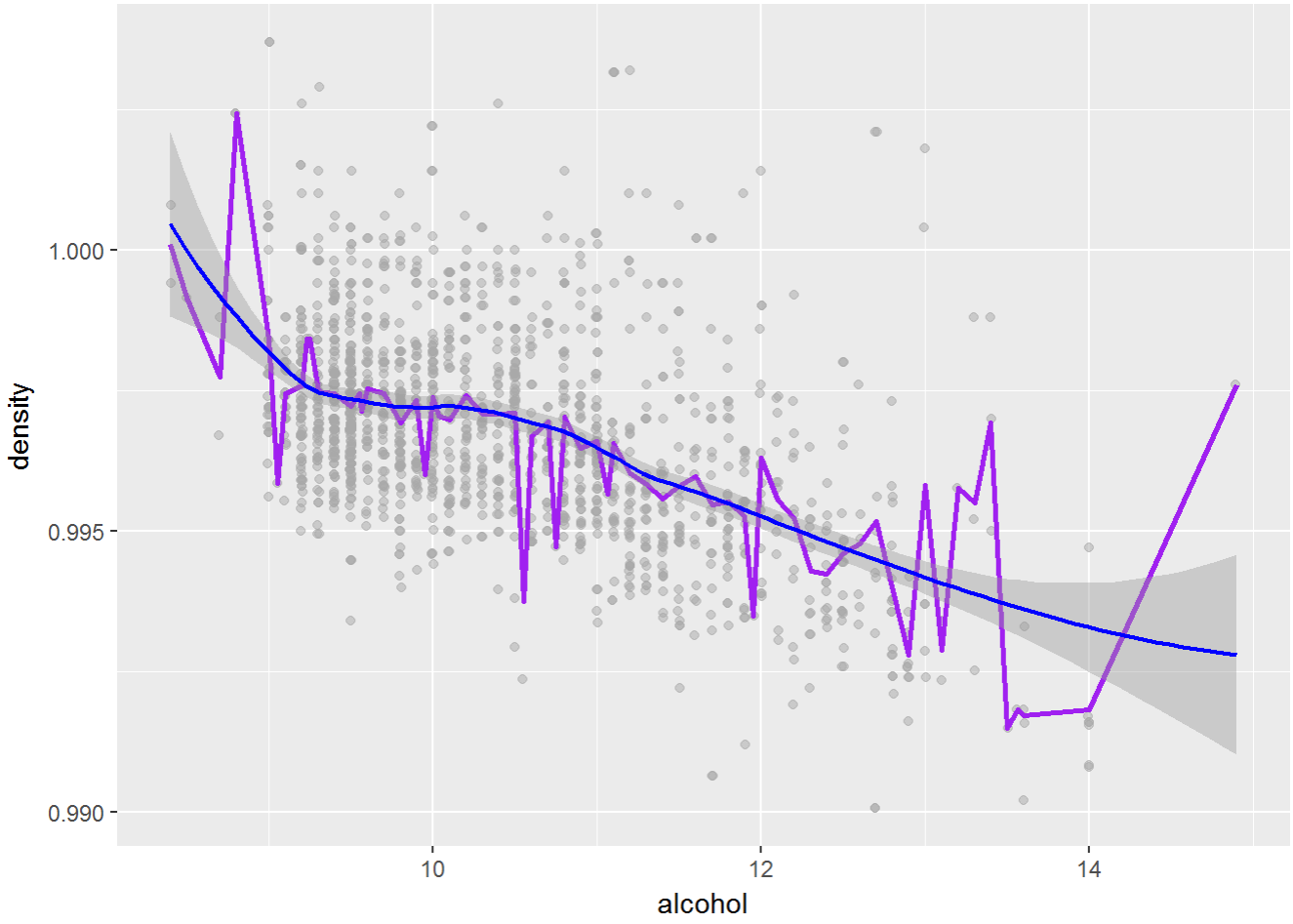

```
##
##  Pearson's product-moment correlation
##
## data:  rwine$alcohol and rwine$total.sulfur.dioxide
## t = -8.398, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2521332 -0.1582280
## sample estimates:
##       cor
## -0.2056539
```



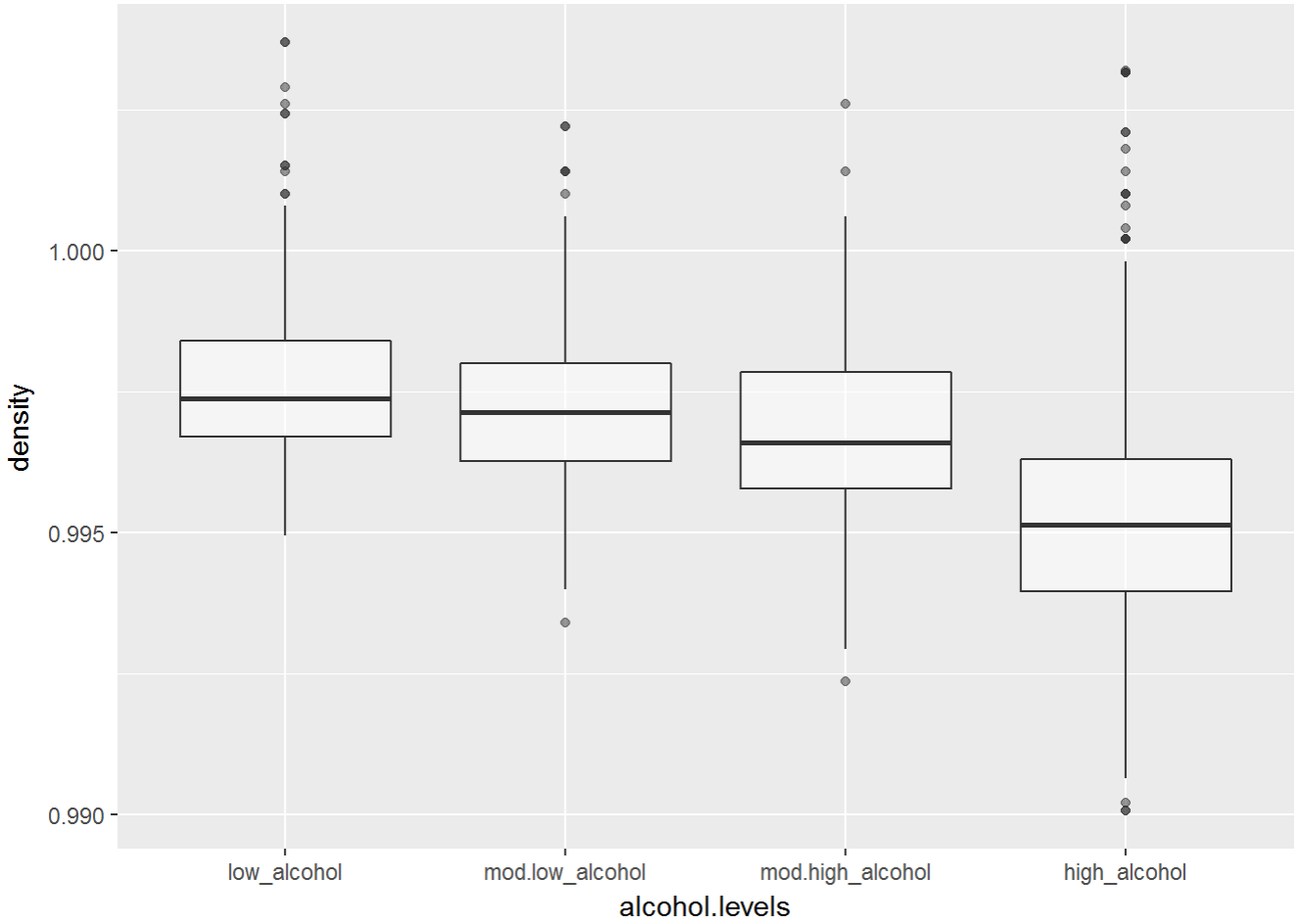

```
##
##  Pearson's product-moment correlation
##
## data:  rwine$alcohol and rwine$chlorides
## t = -9.0617, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2672644 -0.1740057
## sample estimates:
##       cor
## -0.2211405
```




```
##
##  Pearson's product-moment correlation
##
## data:  rwine$alcohol and rwine$volatile.acidity
## t = -8.2546, df = 1597, p-value = 3.155e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2488416 -0.1548020
## sample estimates:
##       cor
## -0.202288
```




```
##
##  Pearson's product-moment correlation
##
## data:  rwine$alcohol and rwine$density
## t = -22.838, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5322547 -0.4583061
## sample estimates:
##       cor
## -0.4961798
```



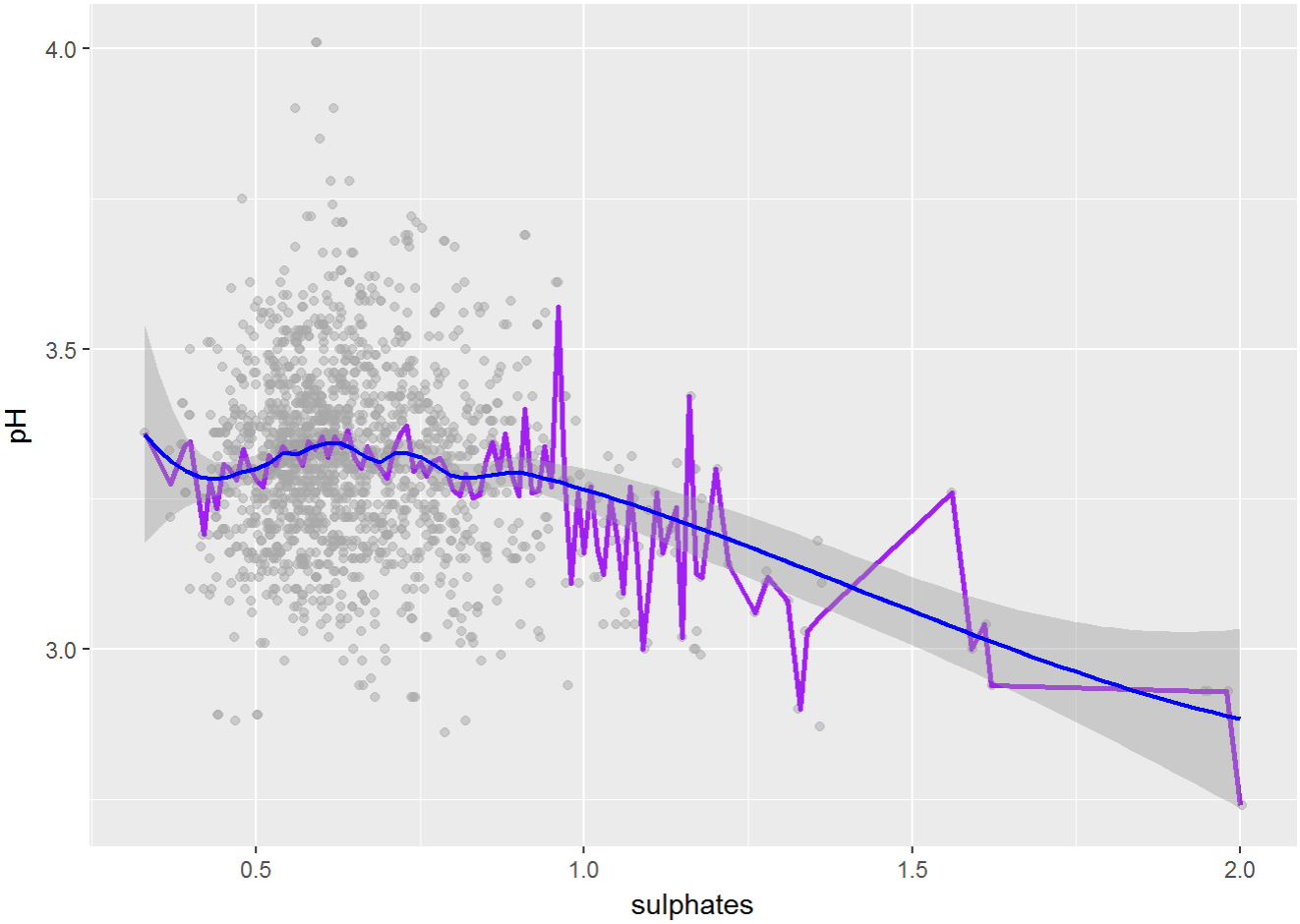
The pH slightly increases with the alcohol content. The total sulfur dioxide, the chlorides and the volatile acidity are not too much influenced by the alcohol levels.

The density decreases with the alcohol content. The loess curve fits better for the main data in both cases. At the two ends (small values and large values) the spread of the data indicates the presence of outliers, and the mean values line indicates large variations. It is interesting to see that three of the four alcohol levels contain wines with very similar density distributions. For high alcohol wines, the density values are lower than in the other cases.

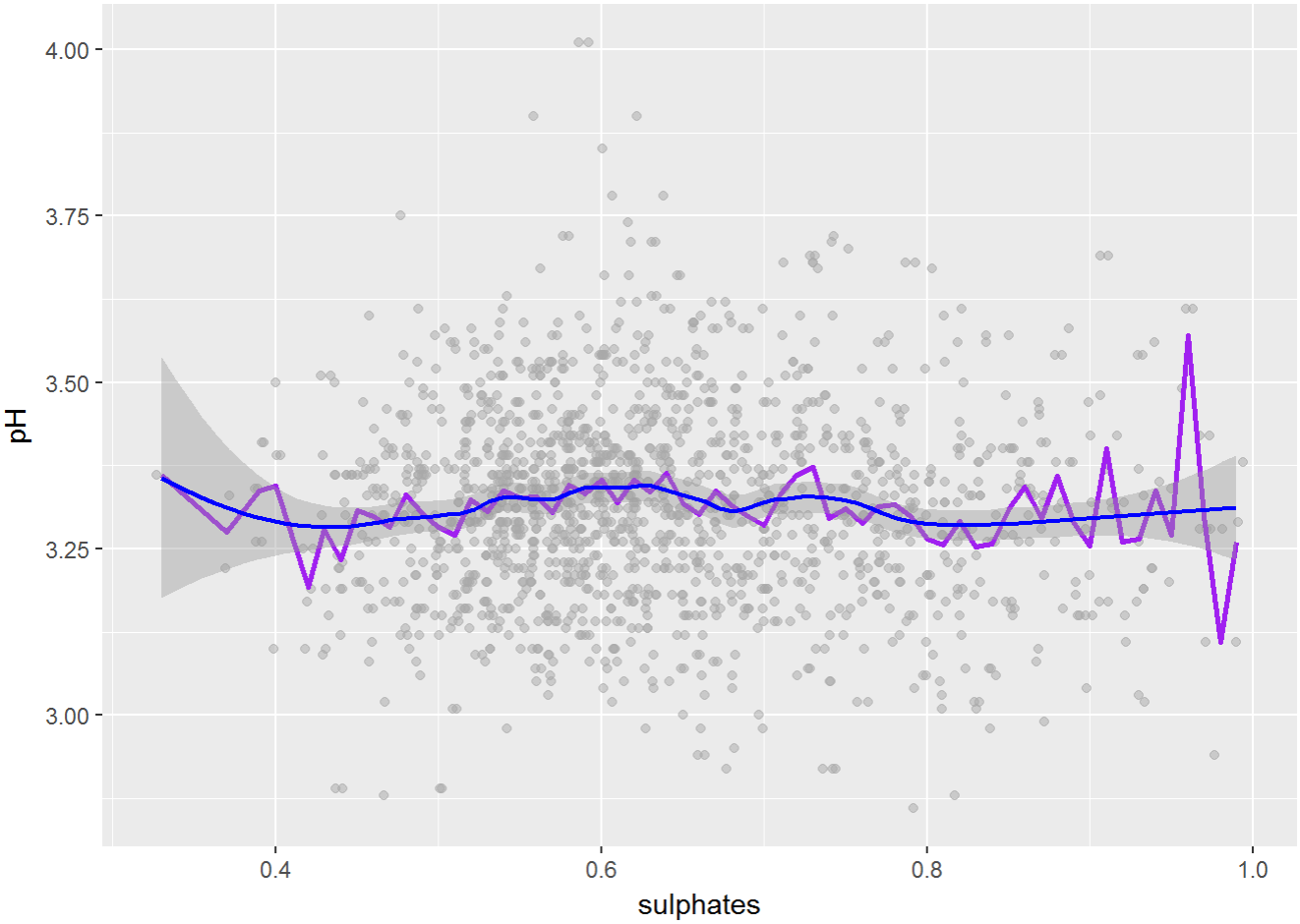
Sulphates and Other Attributes

Sulphates is one of the attributes with more than average extreme outliers. I decided to plot the graphs for the entire data and also for the data without the sulphates outliers (where outliers are defined as values that are 1.5 interquantile range away from the third quartile).

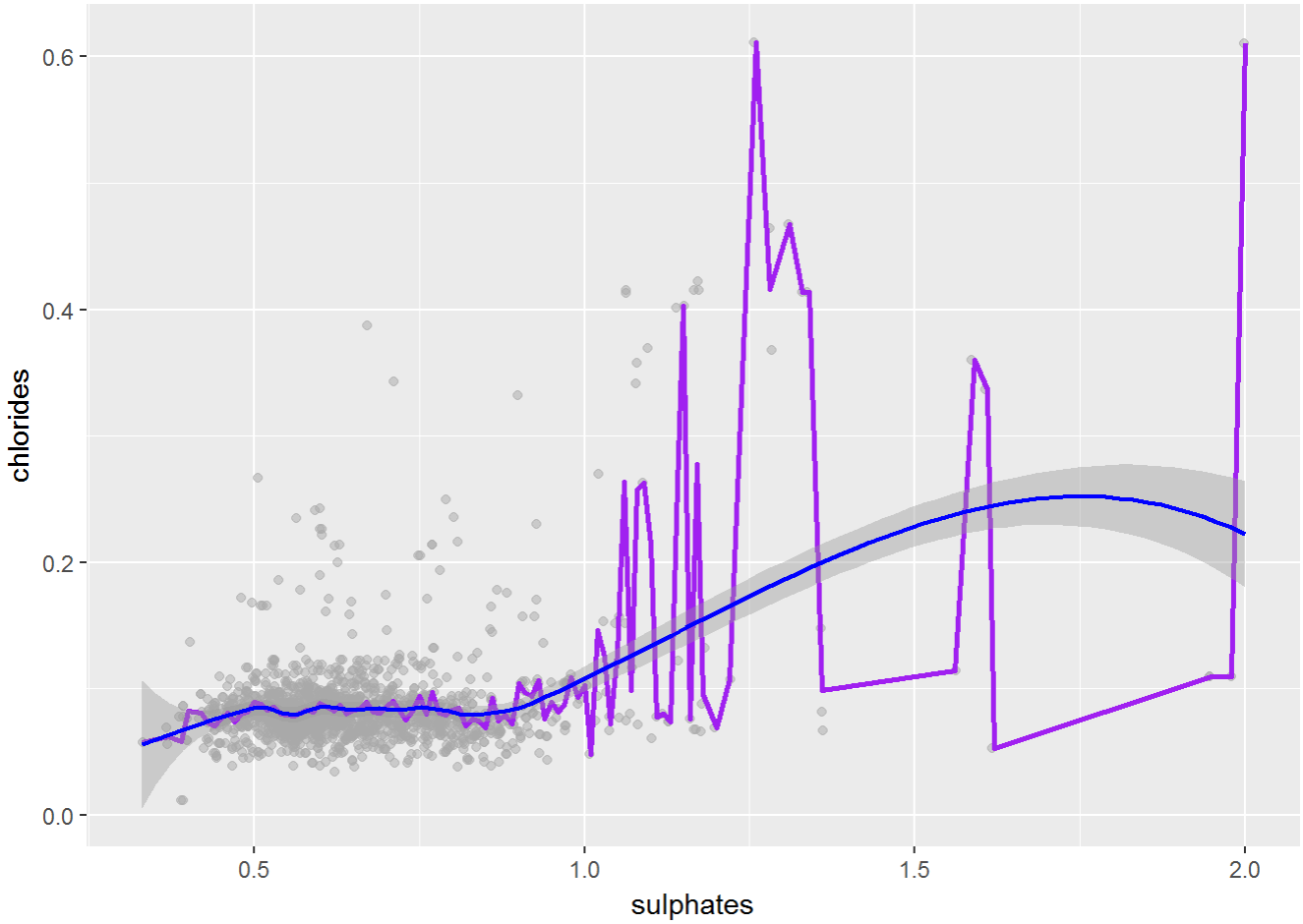
```
## Data without sulphates extreme outliers: s_rwine
```



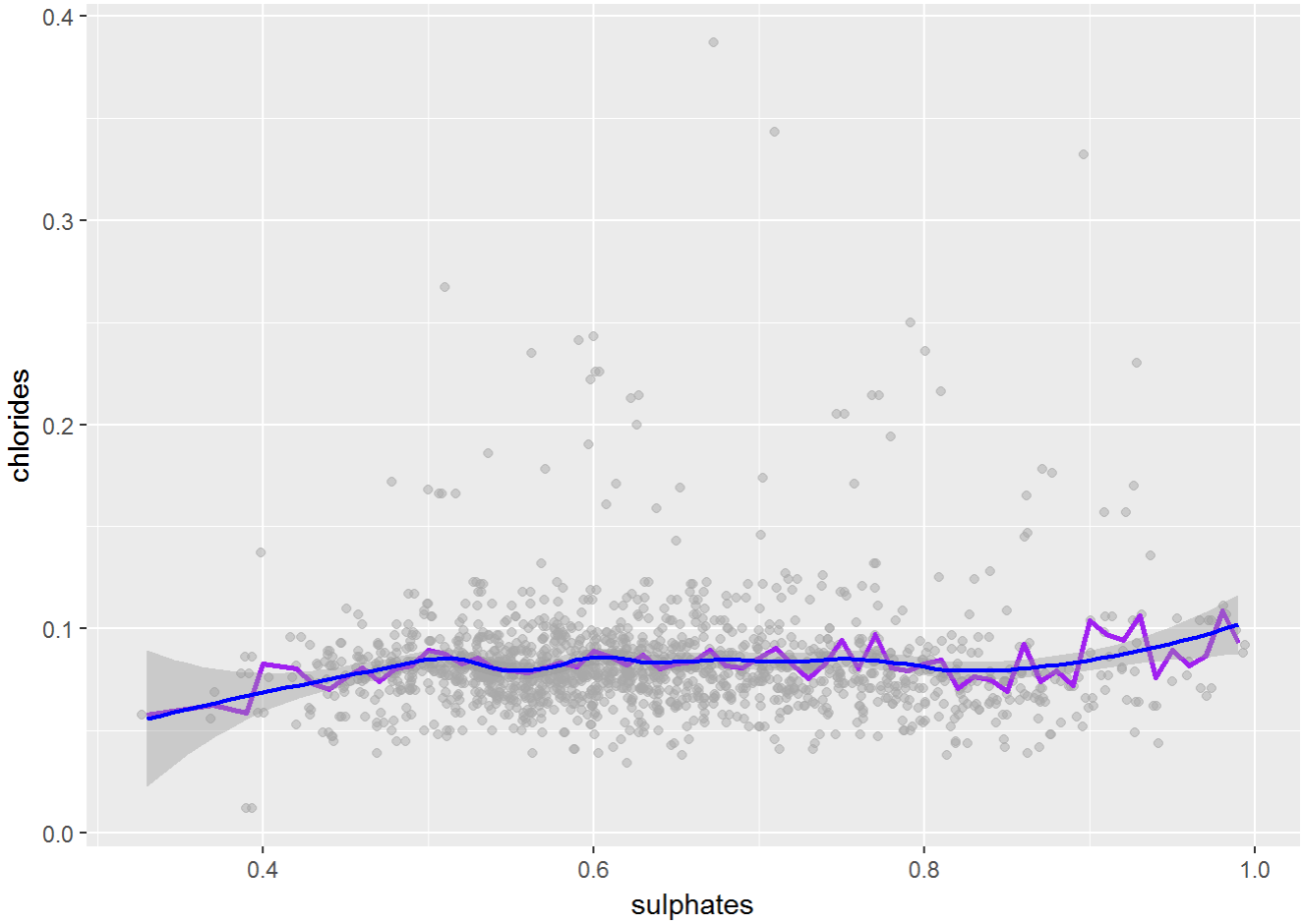

```
##
##  Pearson's product-moment correlation
##
## data:  rwine$sulphates and rwine$pH
## t = -8.015, df = 1597, p-value = 2.107e-15
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2433231 -0.1490634
## sample estimates:
##       cor
## -0.1966476
```



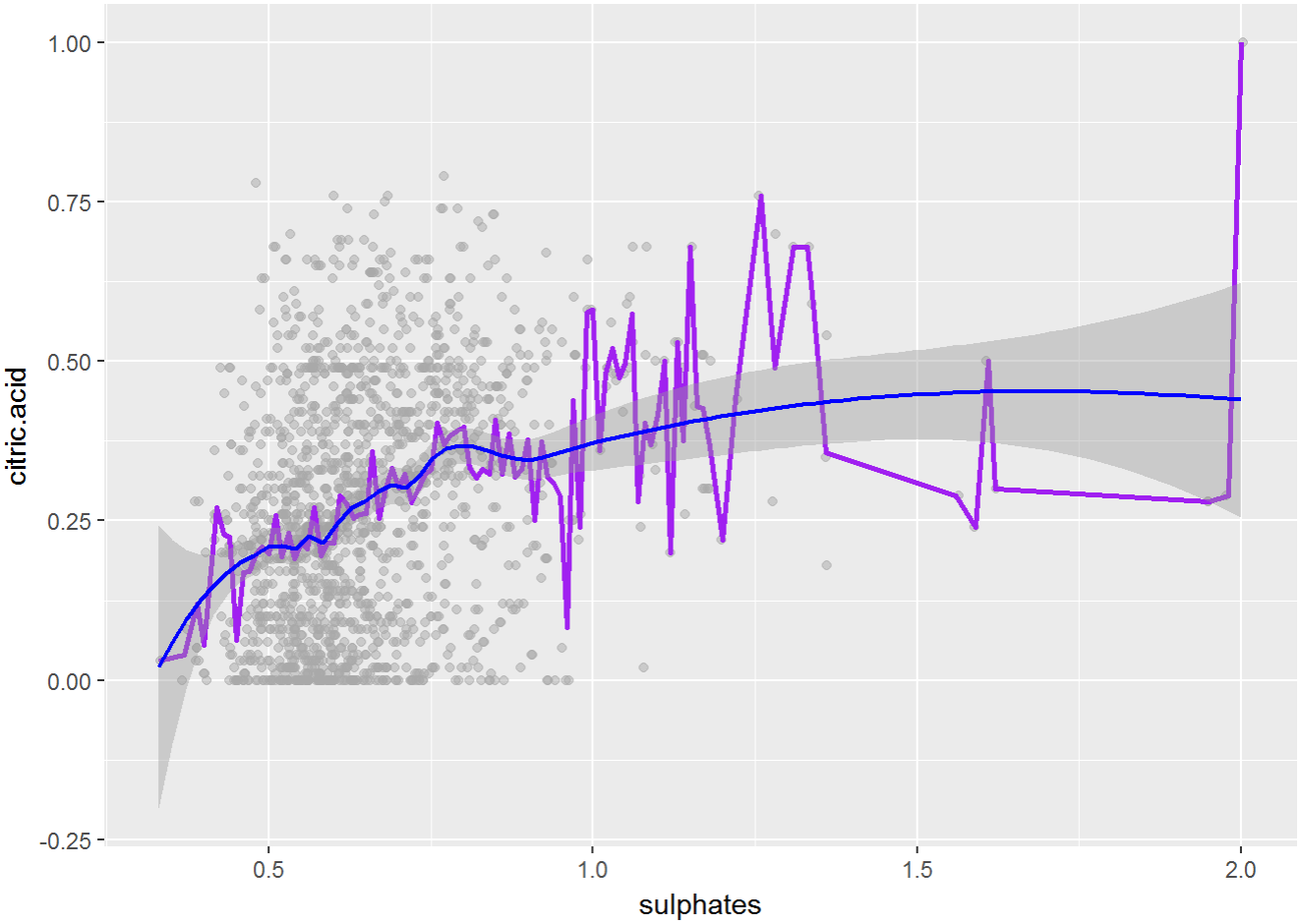
After removing the outliers it is more evident that the pH is not influenced too much by sulphates. Higher levels of sulphates determine a decrease in the pH level.



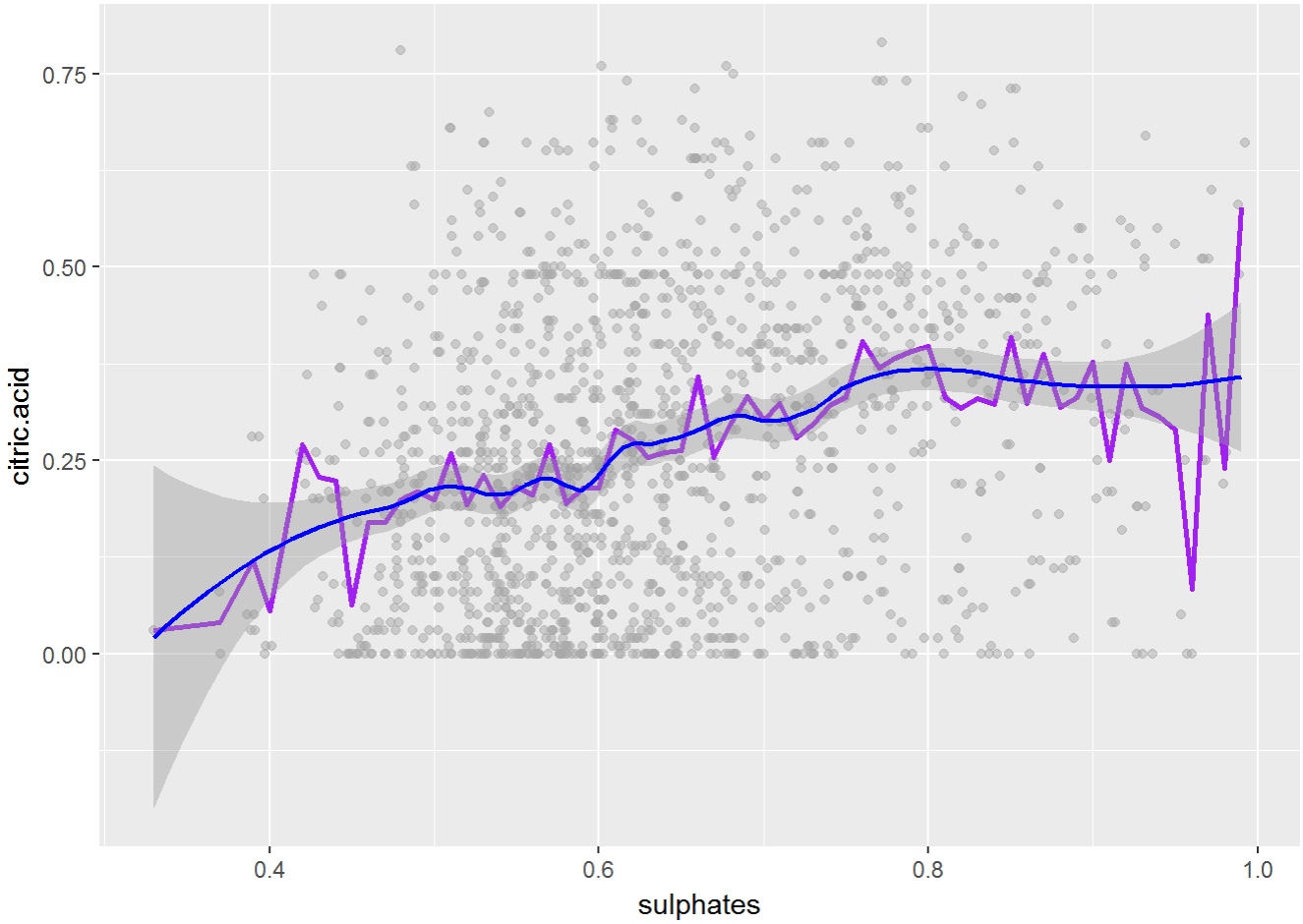

```
##
##  Pearson's product-moment correlation
##
## data:  rwine$sulphates and rwine$chlorides
## t = 15.978, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3282127 0.4127694
## sample estimates:
##       cor
## 0.3712605
```



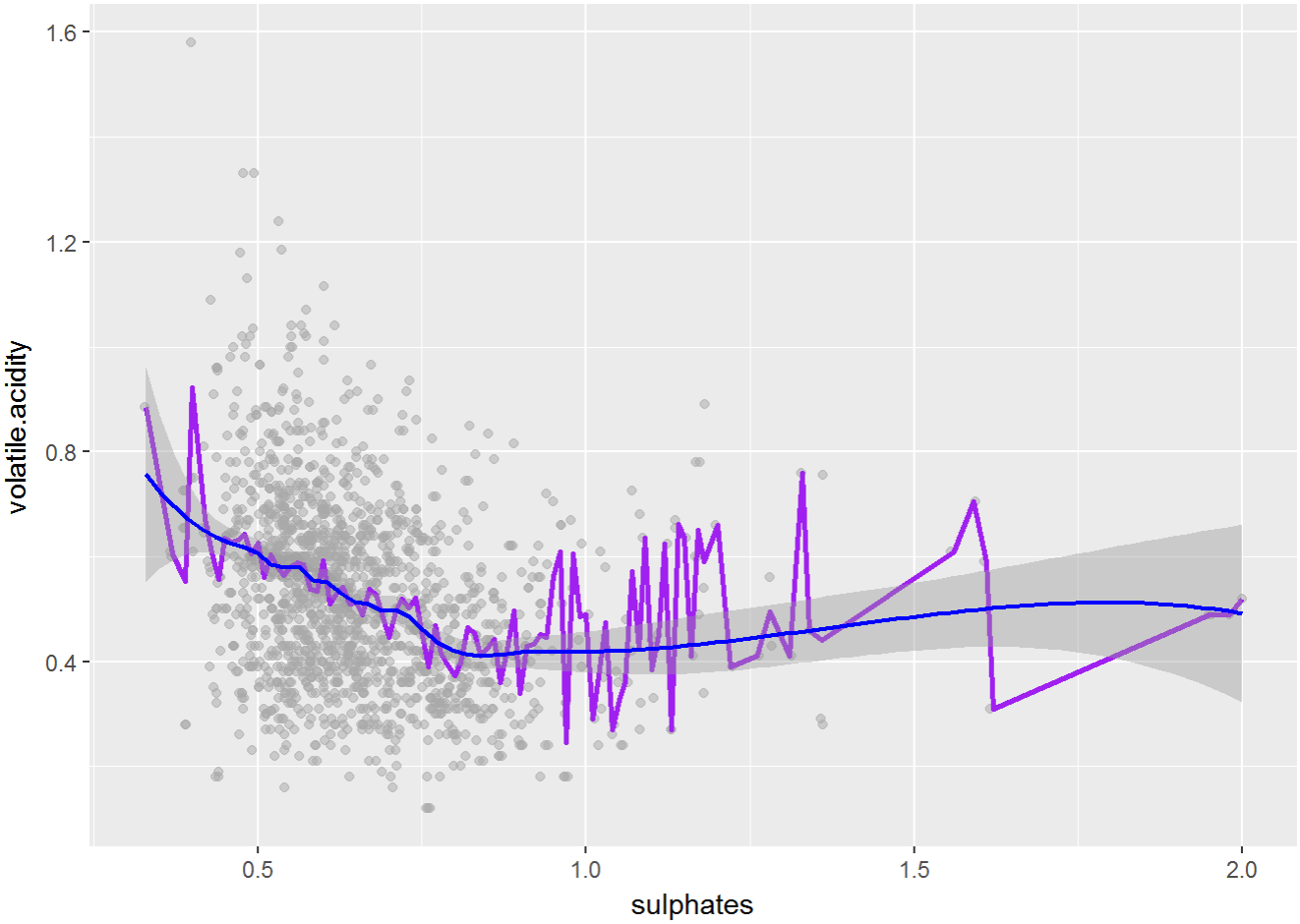
The chlorides are not changing too much with the increase of sulphates for the trimmed data. At large concentrations of sulphates we observe a sharp increase in the chlorides also.



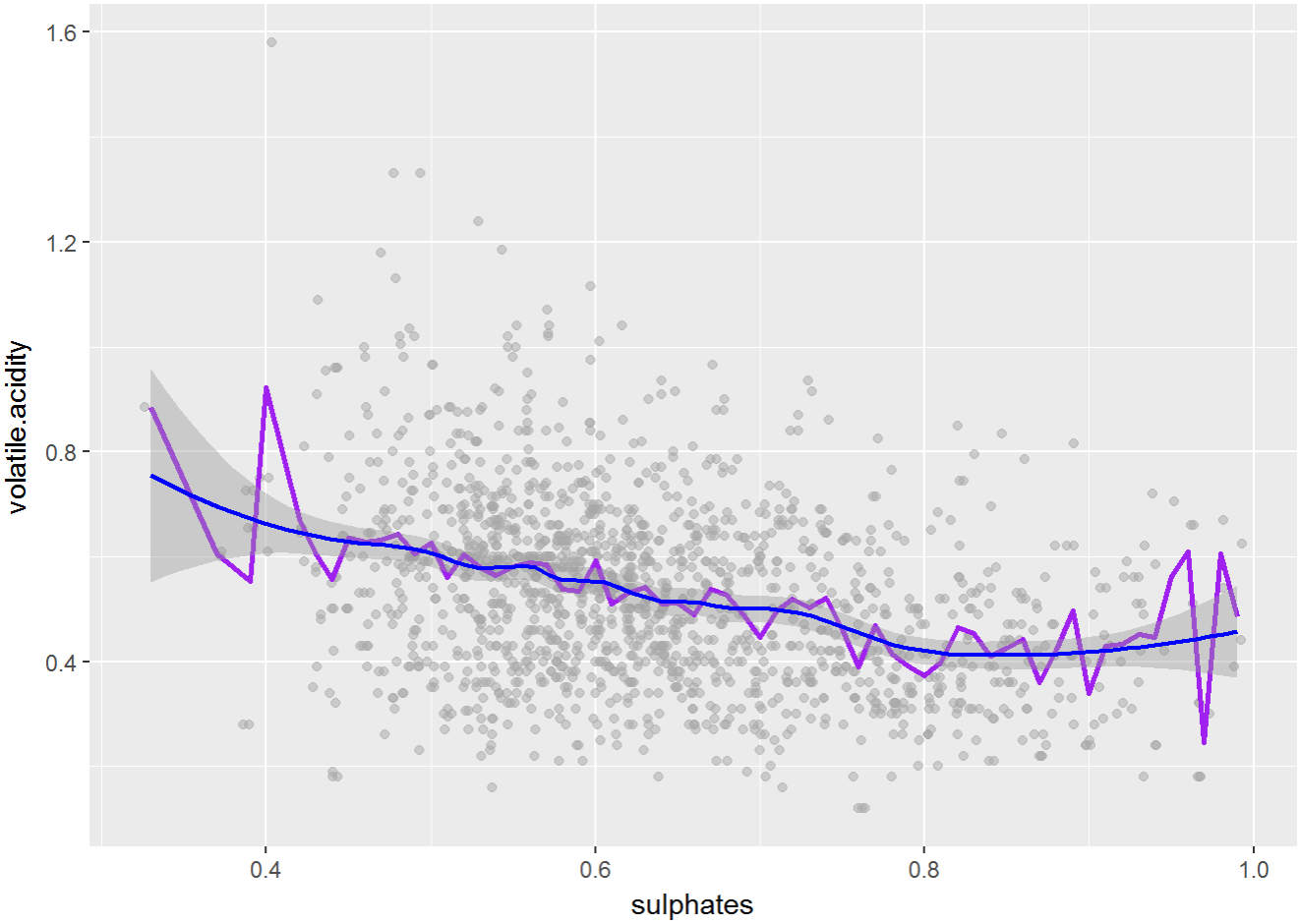

```
##
##  Pearson's product-moment correlation
##
## data:  rwine$sulphates and rwine$citric.acid
## t = 13.159, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2678558 0.3563278
## sample estimates:
##     cor
## 0.31277
```



In the trimmed data, the citric acid slightly increases with the sulphates, but then it stabilizes after the sulphates exceed the outliers threshold (for large values of sulphates in the sample).

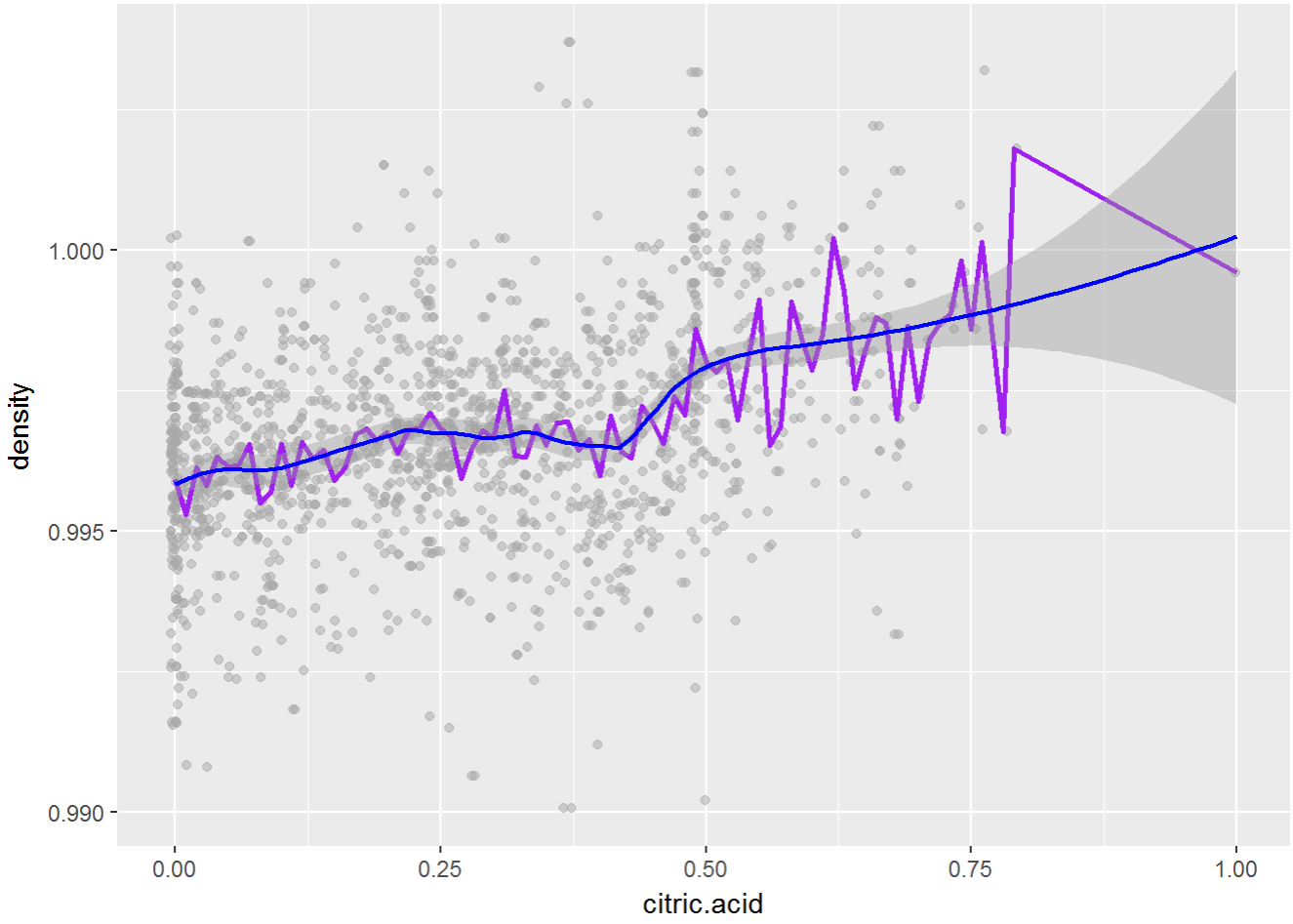



```
##
##  Pearson's product-moment correlation
##
## data:  rwine$sulphates and rwine$volatile.acidity
## t = -10.804, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3060917 -0.2147125
## sample estimates:
##       cor
## -0.2609867
```

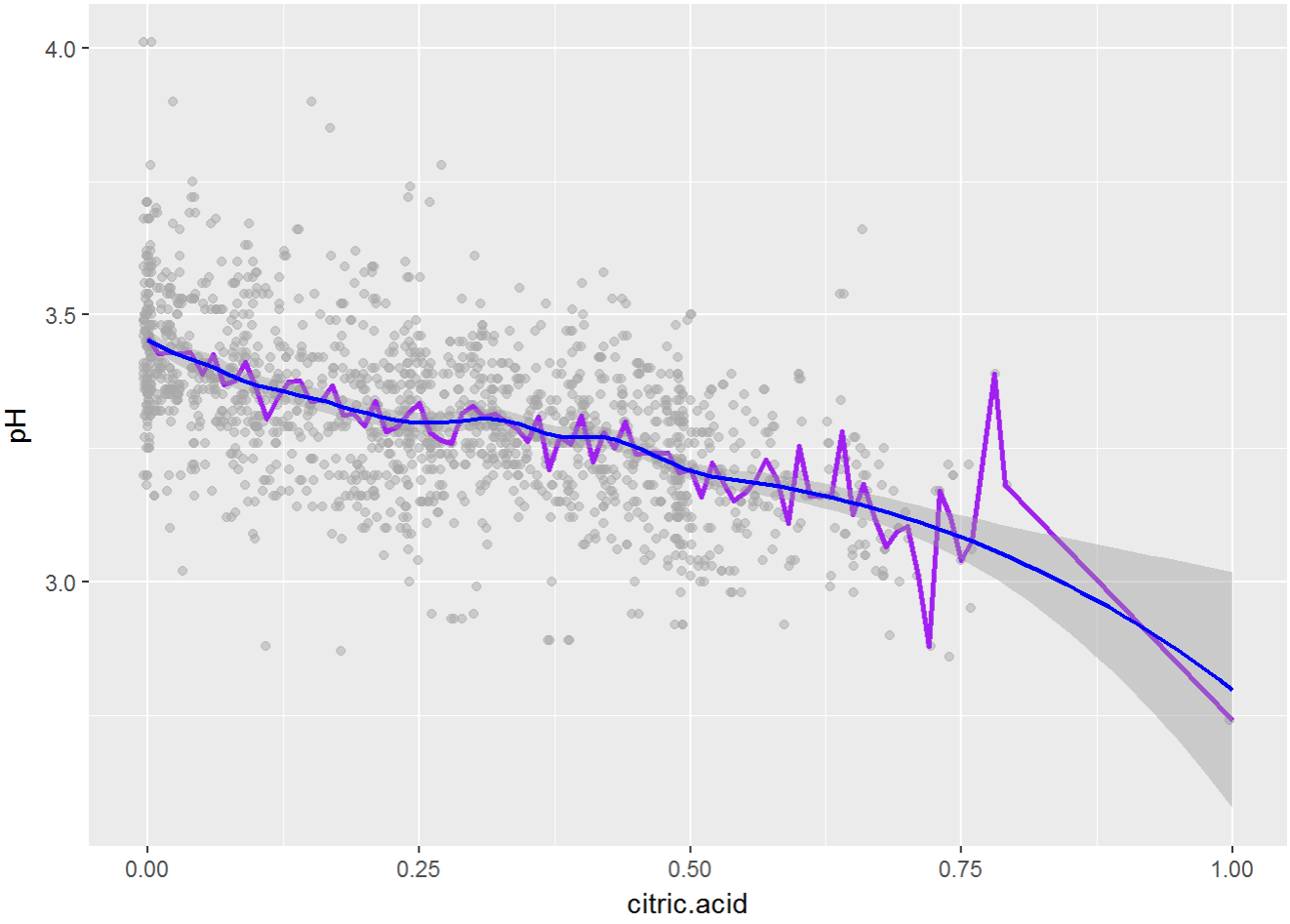


The volatile acidity decreases as sulphates increase in the trimmed data. There seems to be a slight increase in volatile acidity as the sulphates values are in the outliers range.

Citric Acid and Other Attributes

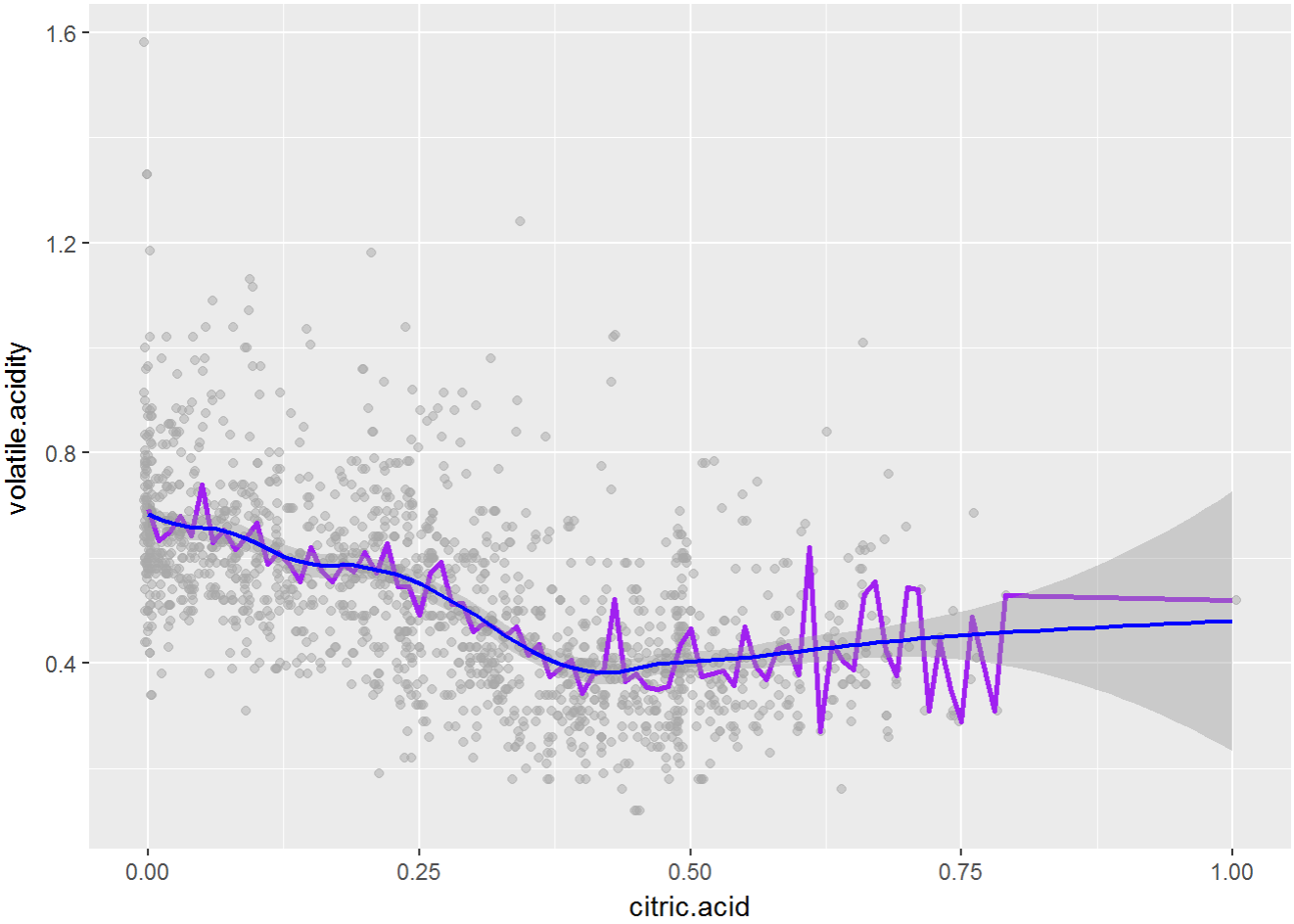


The density increases with citric acid. The mean values line shows a lot of variations in density for larger levels of citric acid. There is also a drop in density for citric acid levels of about 0.40 which is unexpected, given the rest of the pattern.

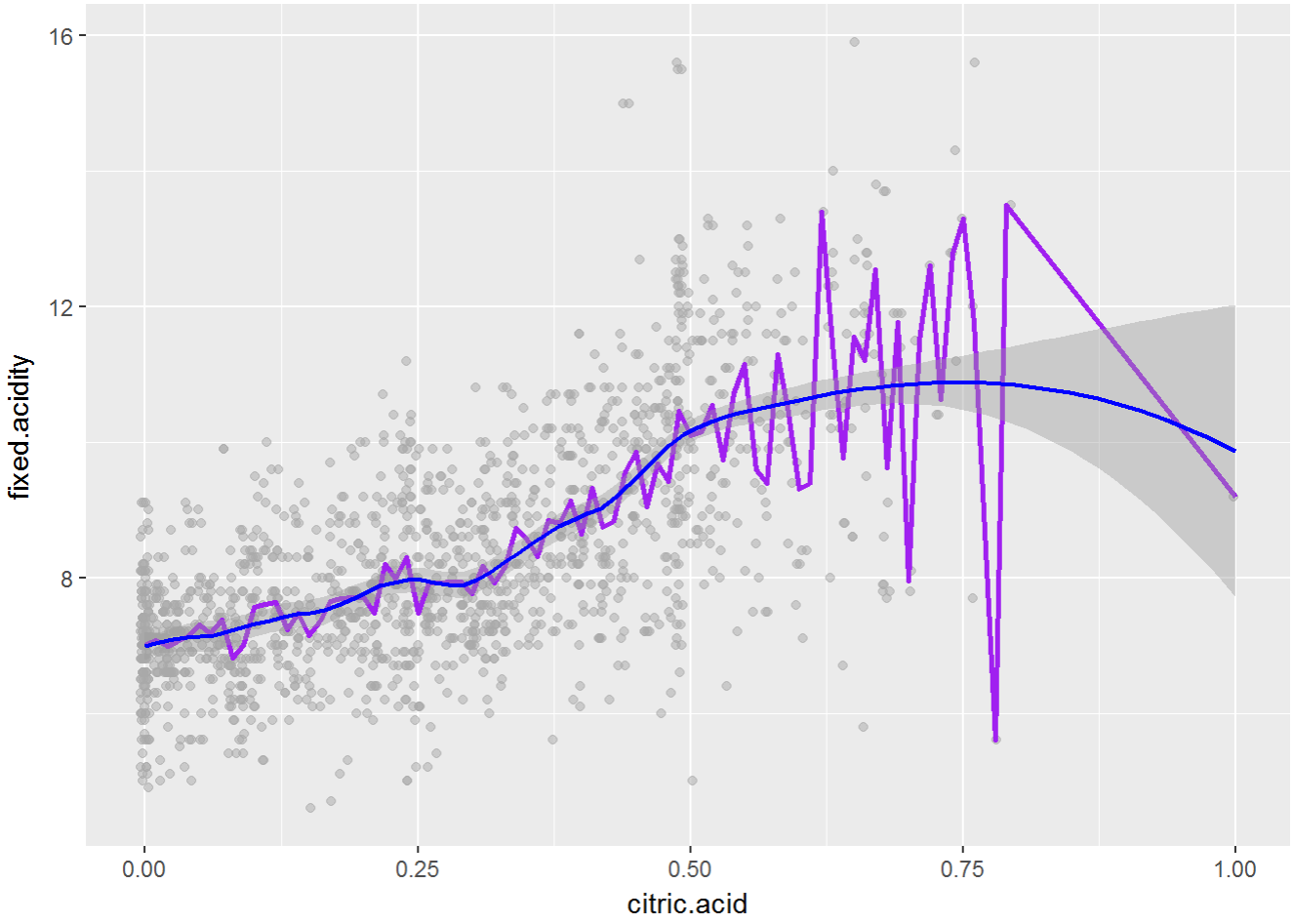



```
##
##  Pearson's product-moment correlation
##
## data:  rwine$citric.acid and rwine$pH
## t = -25.767, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5756337 -0.5063336
## sample estimates:
##       cor
## -0.5419041
```

The pH decreases with the amount of citric acid, and this is what we are expecting to see, the pH and the acidity have a negative relationship, in the sense that higher pH means less acidity.

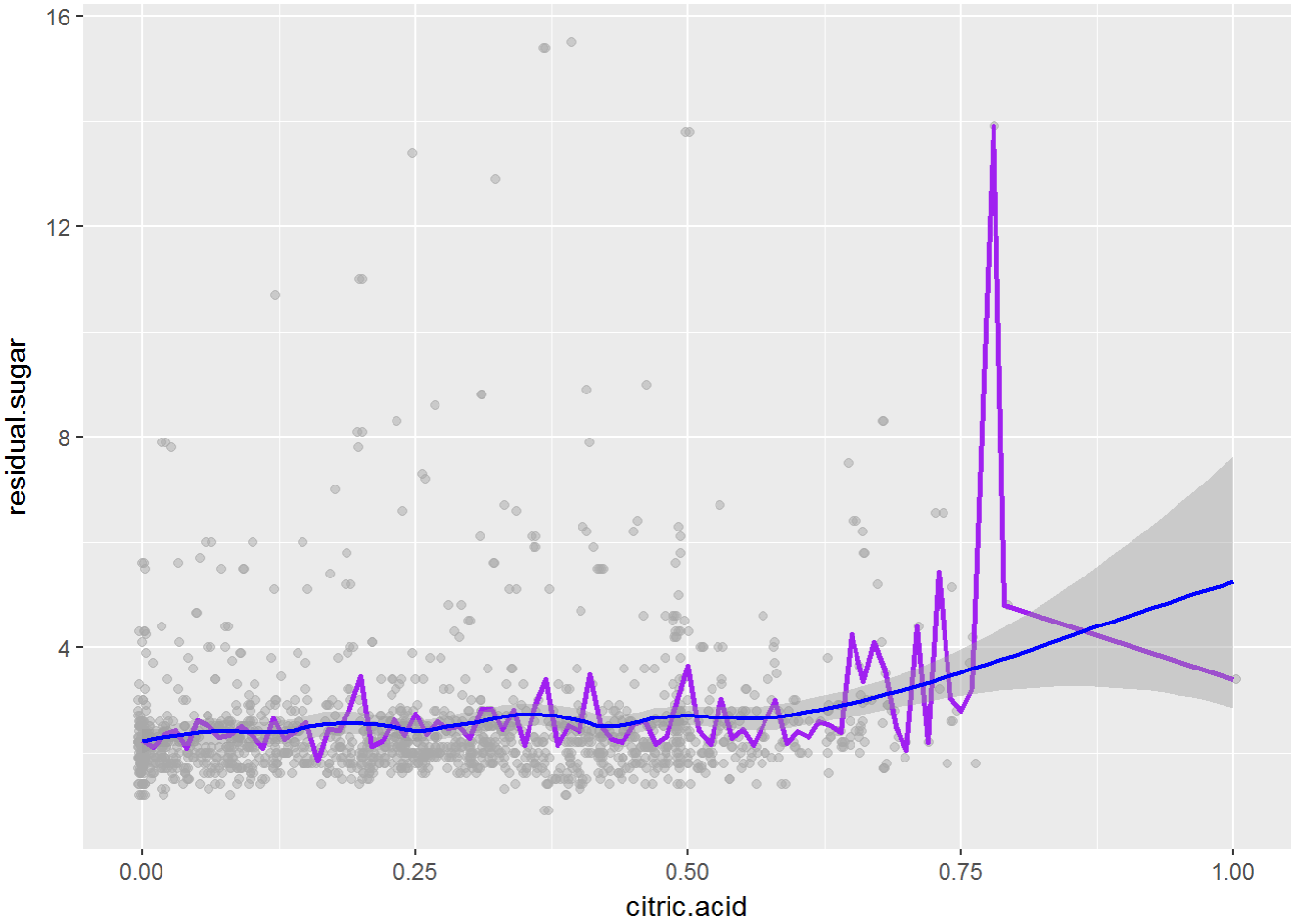



```
##
##  Pearson's product-moment correlation
##
## data:  rwine$citric.acid and rwine$volatile.acidity
## t = -26.489, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5856550 -0.5174902
## sample estimates:
##       cor
## -0.5524957
```

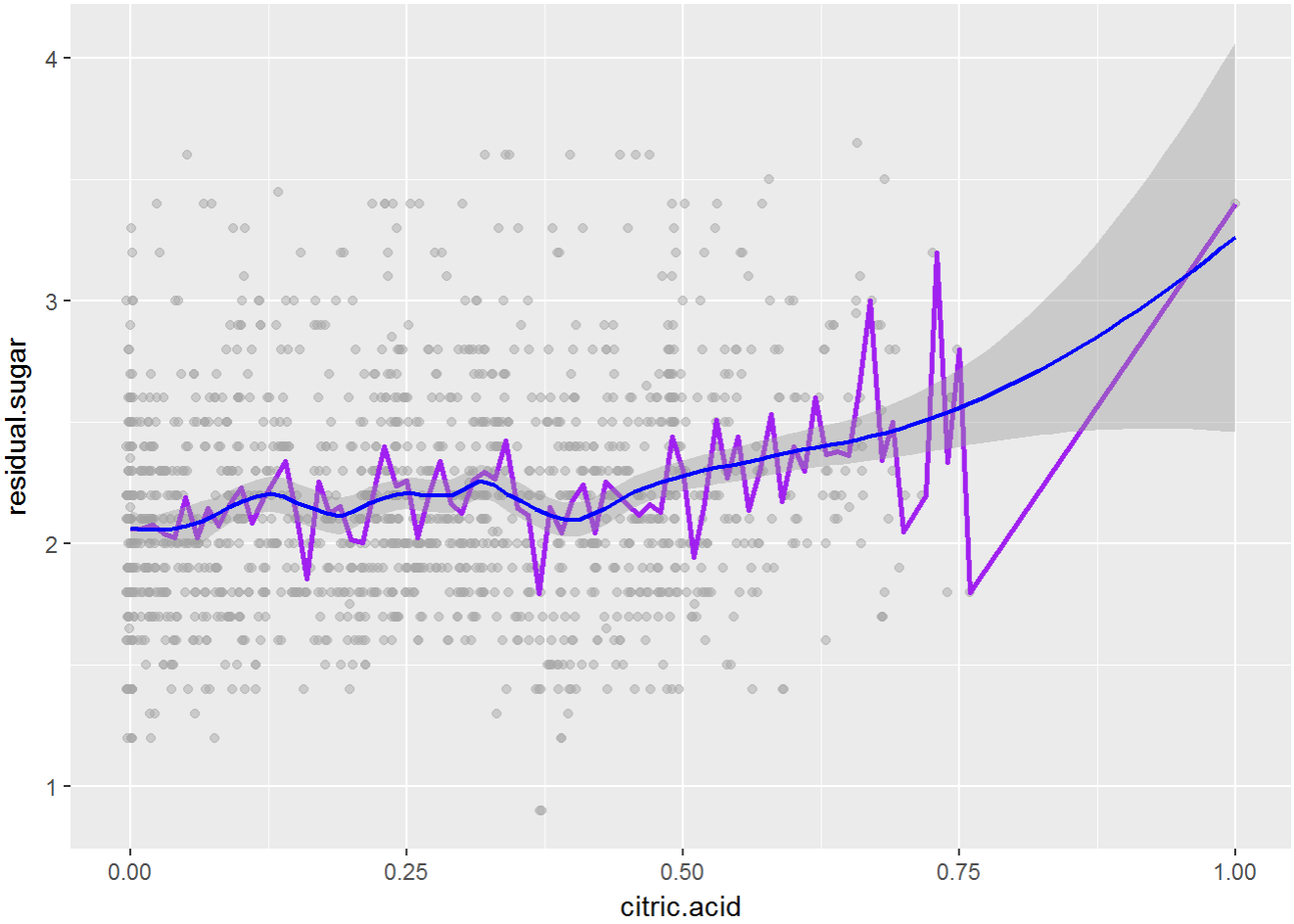



```
##
##  Pearson's product-moment correlation
##
## data:  rwine$citric.acid and rwine$fixed.acidity
## t = 36.234, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6438839 0.6977493
## sample estimates:
##       cor
## 0.6717034
```

It is interesting to see that the volatile acidity decreases with citric acid, for smaller amounts of citric acid. As expected the fixed acidity increases with the citric acid, as the fixed acidity measurement includes the citric acid.



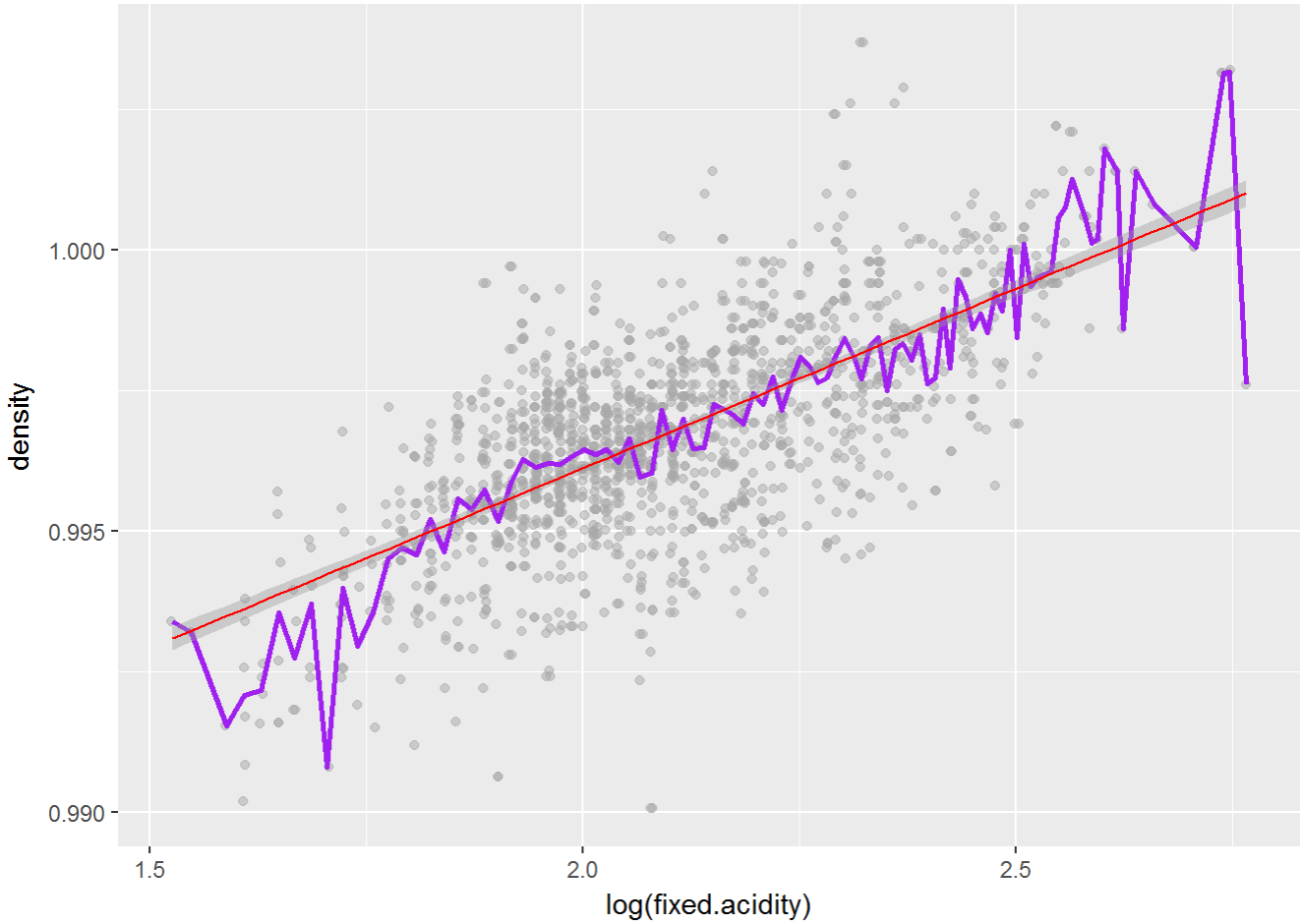

```
##
##  Pearson's product-moment correlation
##
## data:  rwine$citric.acid and rwine$residual.sugar
## t = 5.7978, df = 1597, p-value = 8.084e-09
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.09522625 0.19125221
## sample estimates:
##       cor
## 0.1435772
```



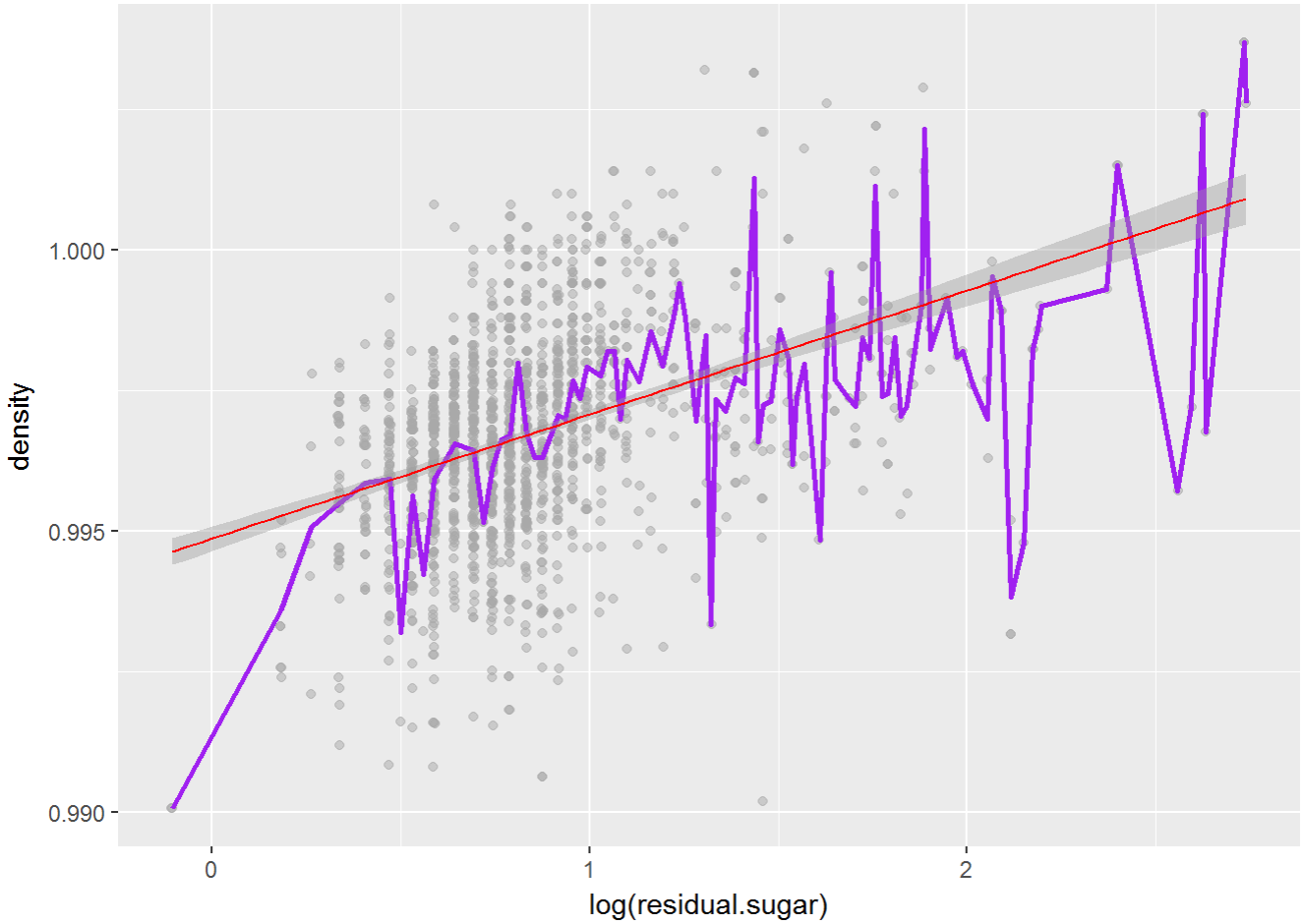
The first plot shows a slight increase of sugar with the citric acid. Given that the residual sugar has numerous outliers, in the second plot we removed these outliers. There is a positive trend, with a drop in the sugar level at about 0.40 citric acid. This drop is observed in the density plot also, and it makes sense to see that the density is lower if there is less sugar in the sample.

Bivariate Plots for Various Attributes

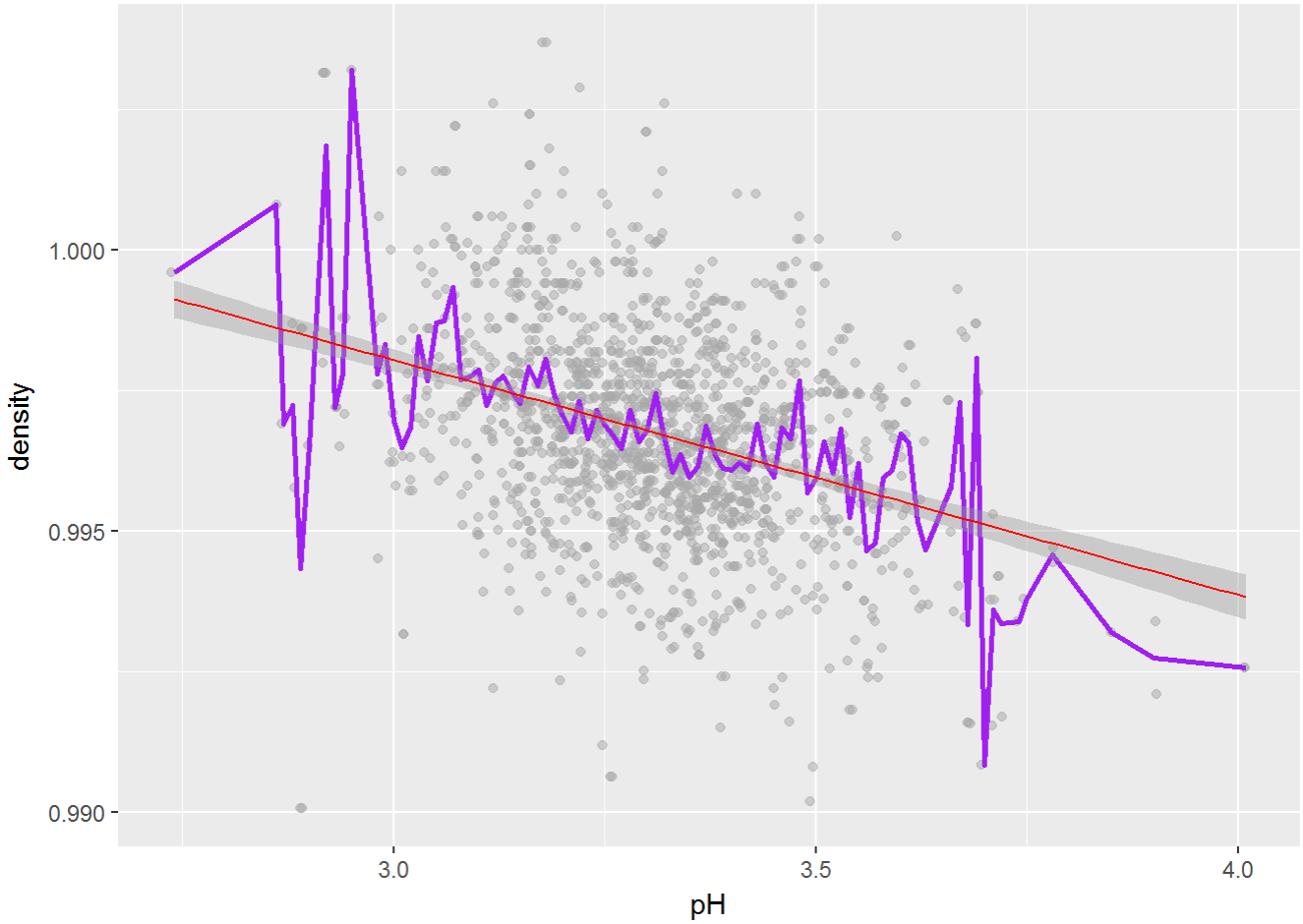
I plot several pairs of attributes. To make the plots more clear I choose in some cases a logarithmic scale along the x-axis. The purple line represents the mean values of the y-variable. The red line is a linear regression line.

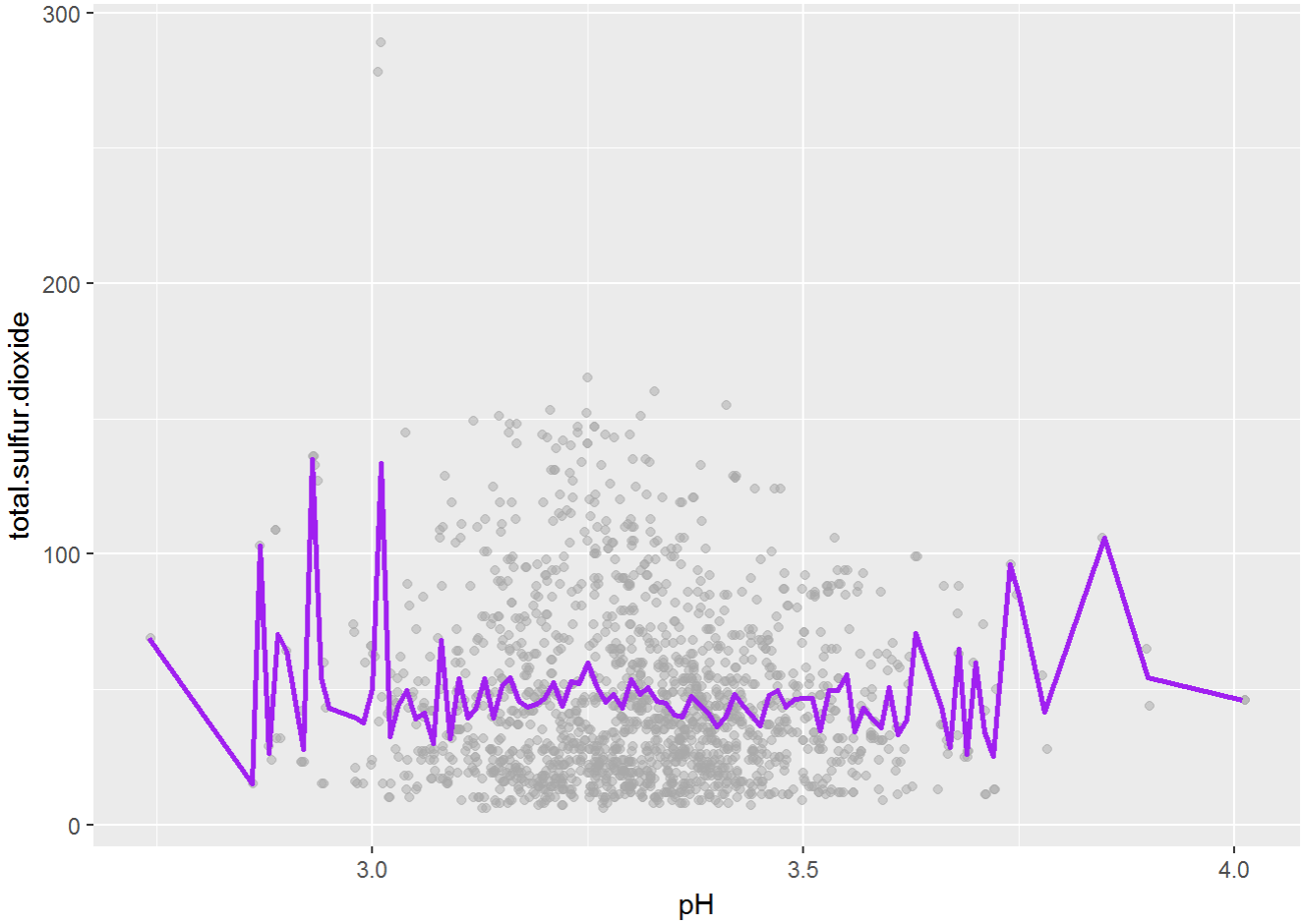


There is a clear pattern here, the density increases with the fixed acidity.



The density mean values have large variations when plotted against the logarithm of residual sugar. The linear model (the red line) indicates that the density increases with the residual sugar level. The linear regression works better for the set of wines with smaller concentration of residual sugar.


density
1.000
0.995
0.990
3.0
3.5
4.0
pH


total.sulfur.dioxide
300
200
100
0
3.0
3.5
4.0
pH

The density decreases with pH. For the main data the total sulfur dioxide is not changing with the pH. There are some variations in the mean values of both density and total sulfur dioxide for pH less than 3 and pH larger than 3.5.

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?

Some of the first relations I noticed are that the citric acid increases with quality, while volatile acidity is smaller in better wines. It is interesting to see that the citric acid mean levels are similar in wines of quality 3 and 4, same for qualities 5 and 6, and again similar for qualities 7 and 8. The values almost double between the three pairs of quality levels.

The residual sugar and the chlorides seem to be quite consistent among all the quality levels. I am slightly intrigued that the sugar levels have almost no relevance on determining the quality of the sample.

Th pH levels slightly decrease with quality, the plots do not indicate any strong relationship nor unusual patterns.

Both the sulphates and the alcohol have stronger correlations with quality. The sulfates median values are equal for the wines of qualities 7 and 8. In the case of the alcohol we see the only sharp increase as the quality of the wines is better. I find it interesting that, in average, the wines of quality 5 have much less alcohol than the wines of quality 4.

The extreme outliers which we found for residual sugar, chlorides and sulfates are distributed mostly in categories 5 and 6 (for quality). I do not think that they bear any relevance in determining the quality of the wine.

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

I noticed that the density increases with the levels of fixed acids (including citric acid) and residual sugar, and it decreases with pH and alcohol. These are all expected behaviors that are based on the chemical and physical properties of these attributes.

I noticed that the density of the wines with high alcohol content (more than 75% of the

maximum alcohol levels) is significantly lower than of the other wines.

The sulfates increase with citric acid, but decrease with volatile acidity.

I was surprised to see that there is a positive correlation coefficient between pH and volatile acidity. The fixed acidity increases with citric acid. The volatile acidity decreases with citric acid (up to values of 0.40 for citric acid) after which there is a slight increase in the values of volatile acidity.

Another unexpected feature is that when the citric acid is about 0.40, the residual sugar and consequently the density of the wines have smaller values than those in the vicinity. It seems that around 0.40 g/dm^3 citric acid there are some changes in the physicochemical properties of these types of wine. Of course, this can be just a local anomaly with no real relevance.

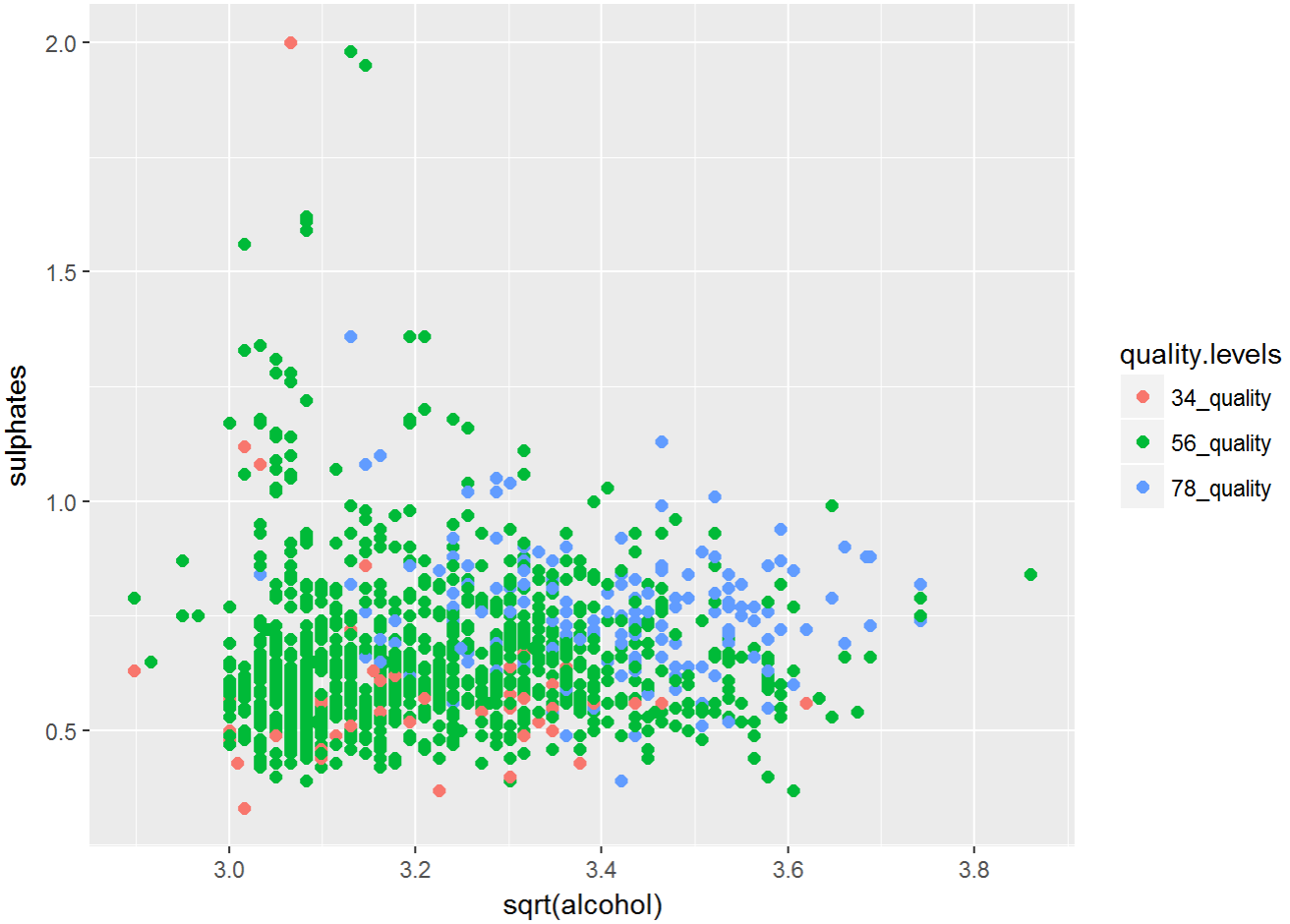
What was the strongest relationship you found?

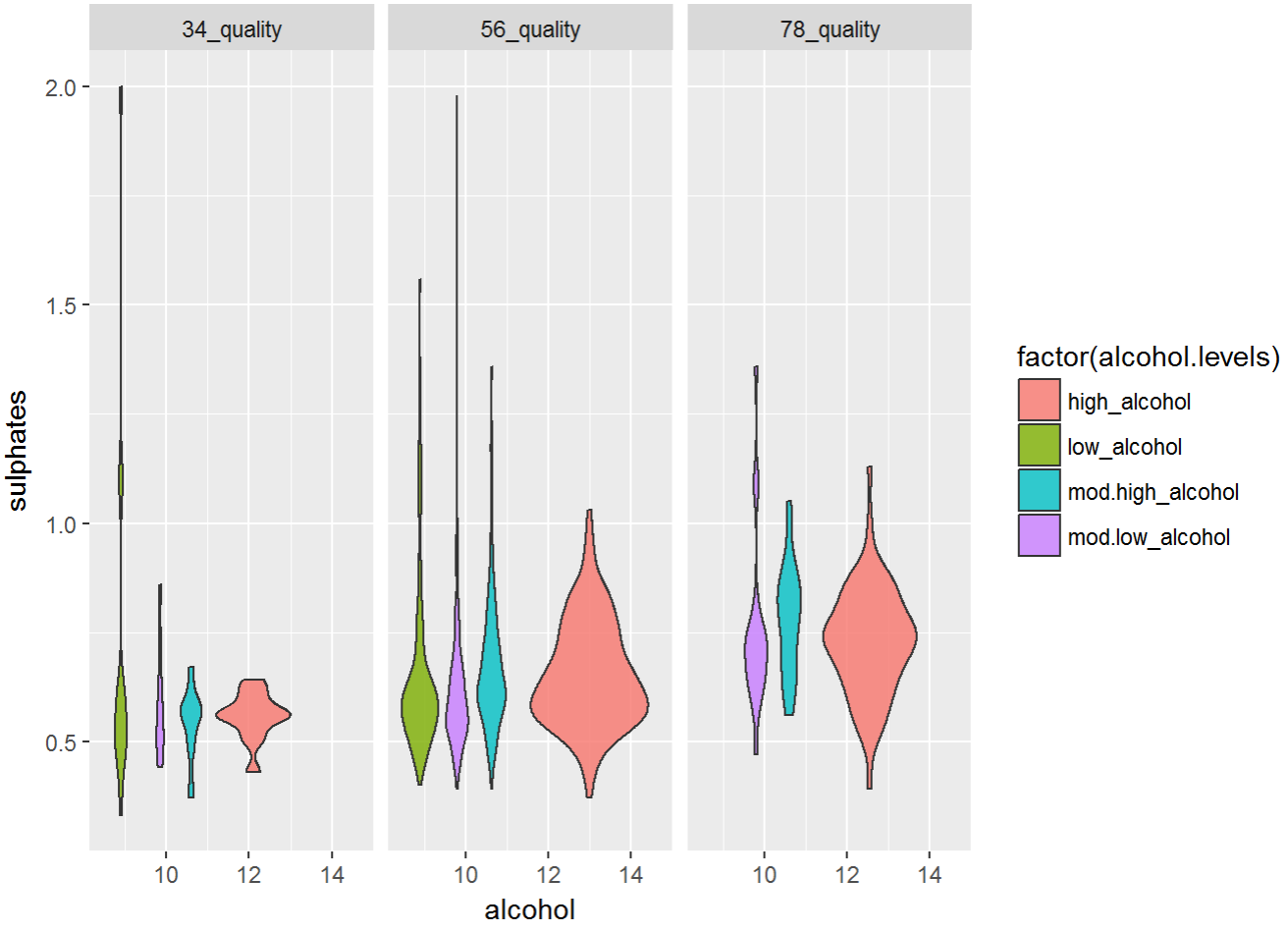
The strongest relationship of quality is with alcohol. Next the quality is influenced by sulphates and citric acid, and in a negative manner by volatile acidity.

Multivariate Plots Section

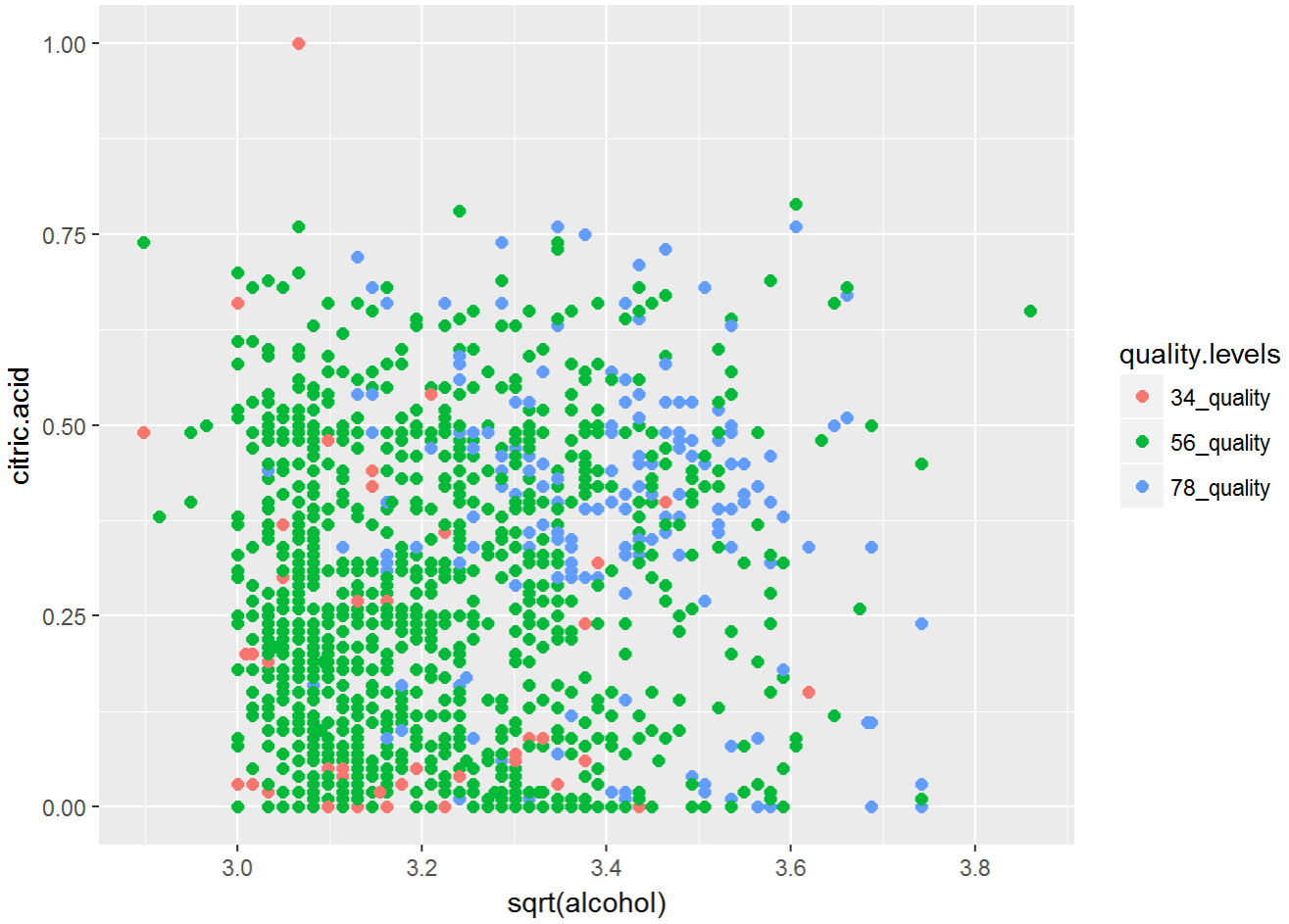
In the bivariate plots section I noticed that the wines of quality 3 and 4 have similar properties. The same is true for the wines with qualities 5 and 6, and also for the better wines, qualities 7 and 8. I will treat each of these pairs together. In order to do so, I create the quality.levels factor variable that combine the quality in three pairs.

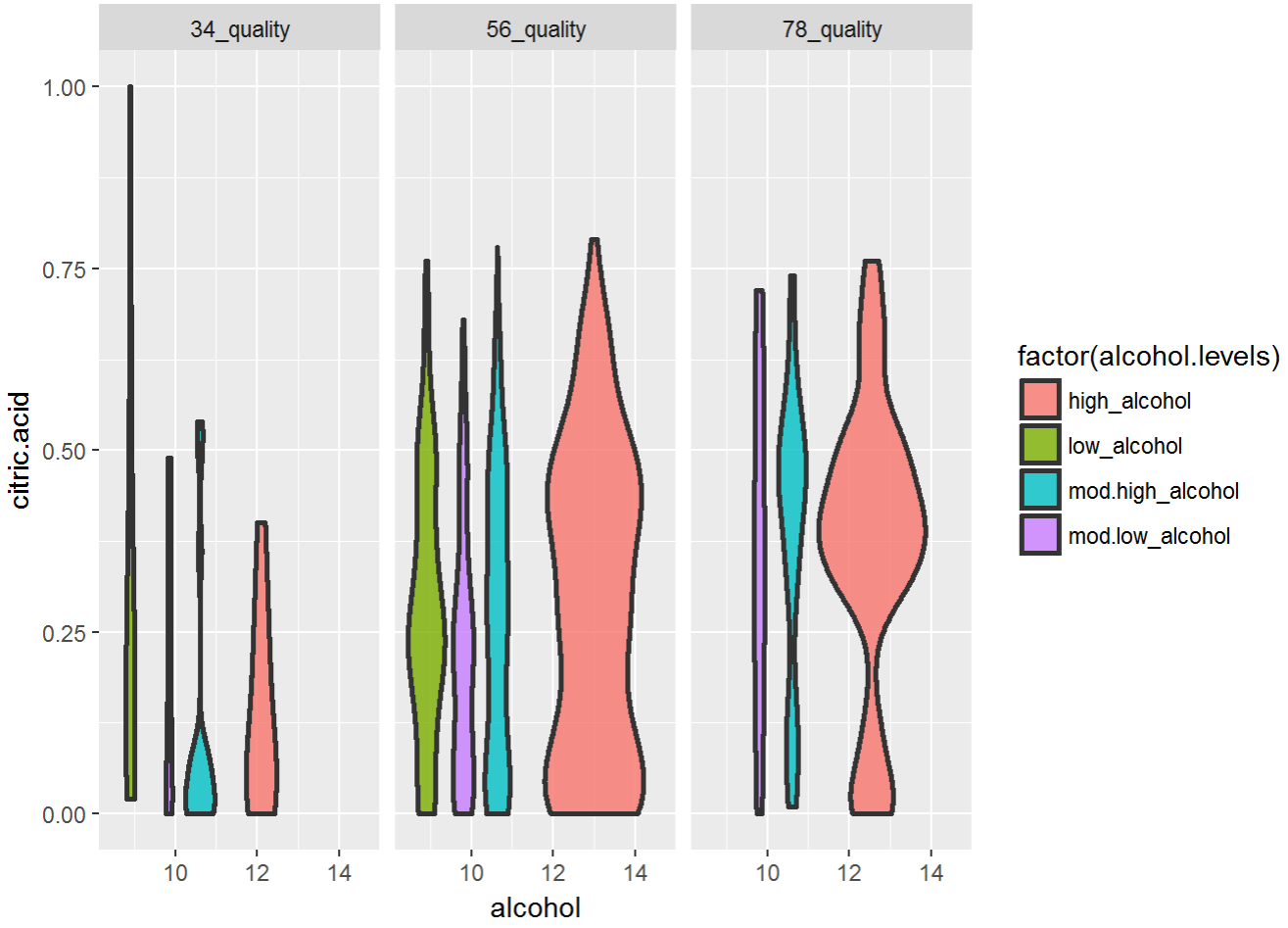
```
##
## 34_quality 56_quality 78_quality
##         63       1319        217
```


sulphates
2.0
1.5
1.0
0.5
3.0
3.2
3.4
3.6
3.8
sqrt(alcohol)
quality.levels
34_quality
56_quality
78_quality

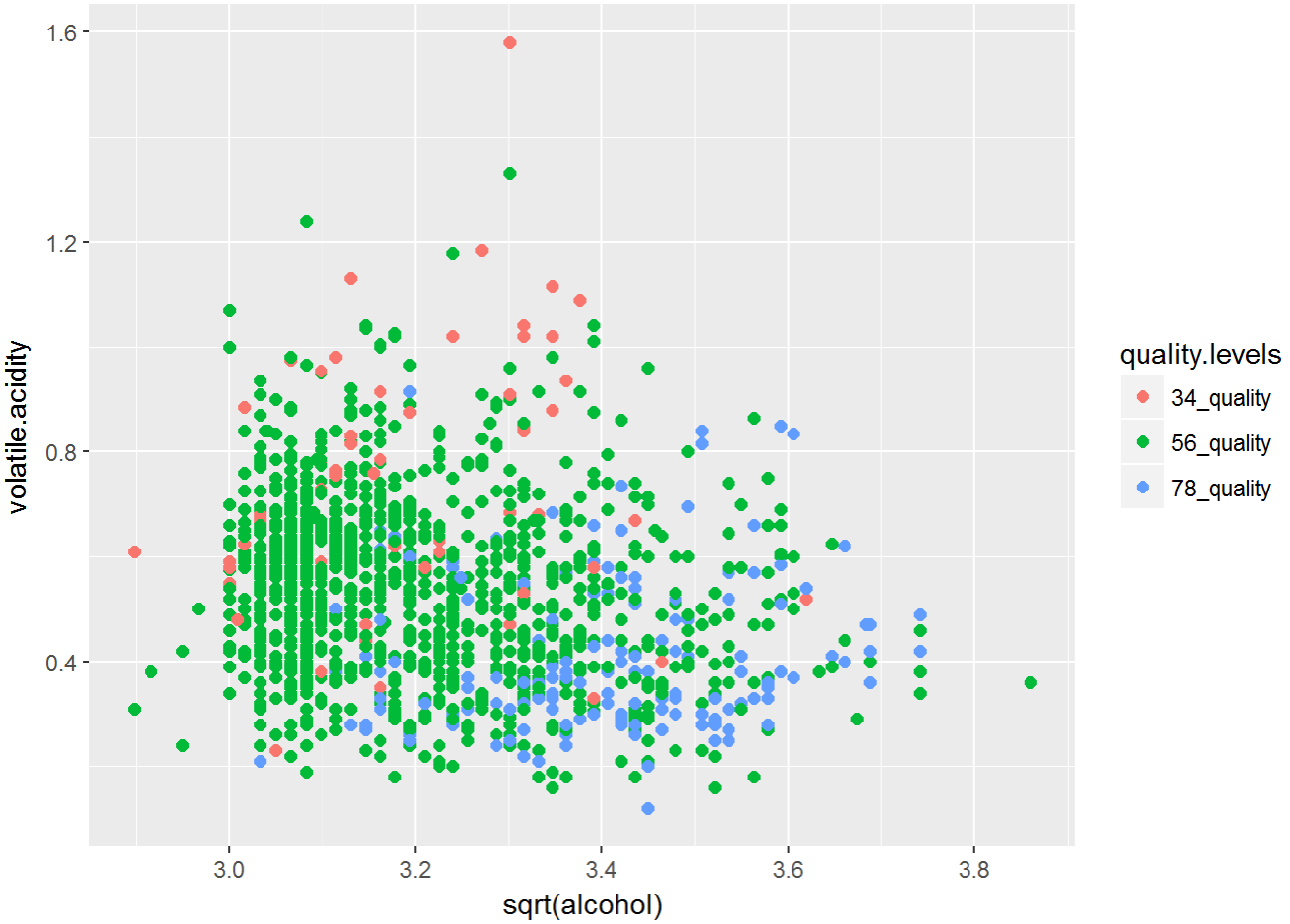

34_quality
56_quality
78_quality
sulphates
2.0
1.5
1.0
0.5
10
12
14
alcohol
factor(alcohol.levels)
high_alcohol
low_alcohol
mod.high_alcohol
mod.low_alcohol

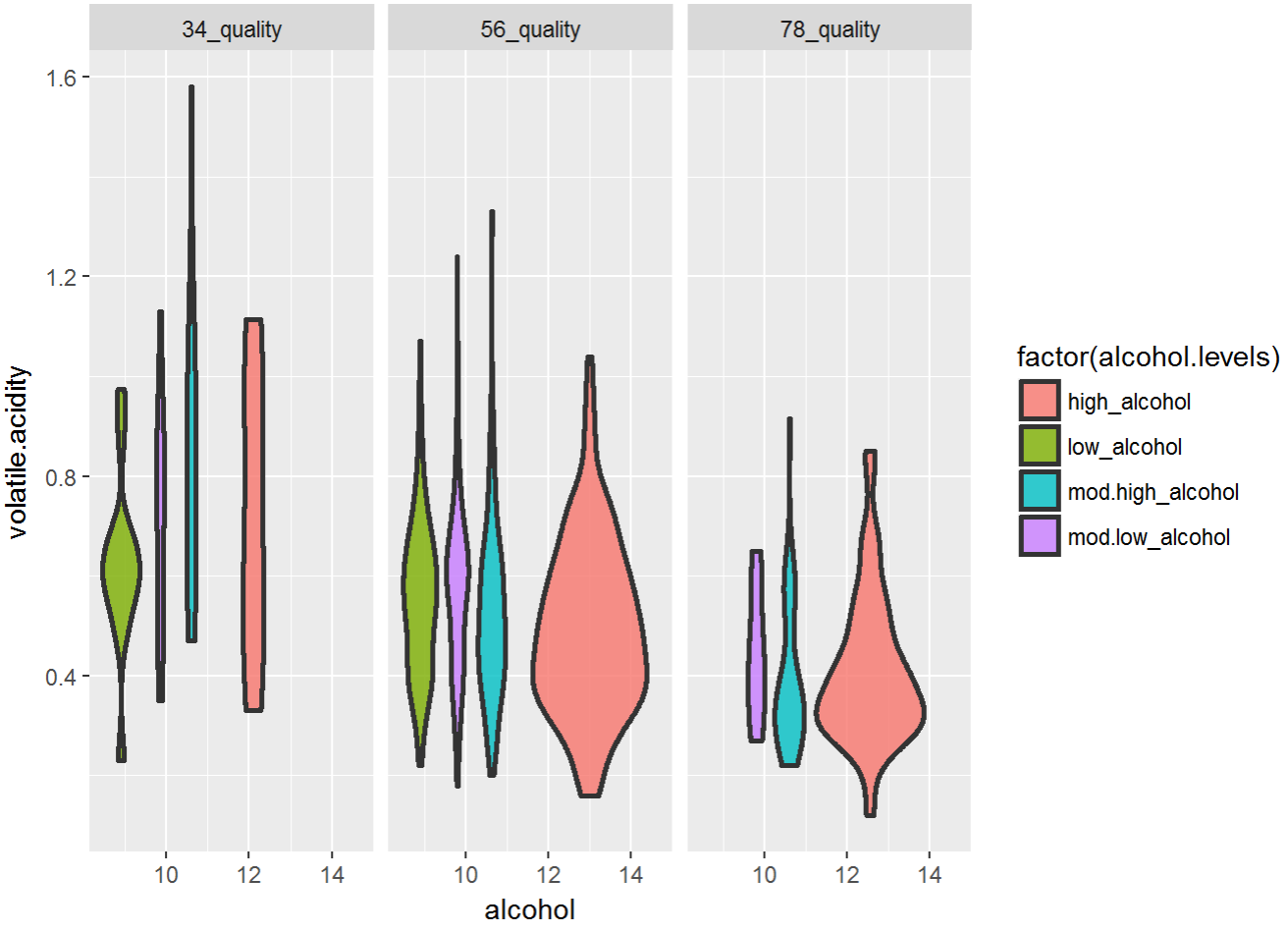
I notice that the 34_quality wines tend to have fewer sulphates and lower levels of alcohol. The average wines, of quality 5 and 6, are more numerous. Their sulphates levels show outliers (which I also noticed in the bivariate plots) and although they have representatives in all four alcohol levels, they tend to be concentrated to the left of the scatterplot (of smaller alcohol contents). The better wines, 78_quality, are higher in both sulphates and alcohol. In fact, there are no low_alcohol wines in the 78_quality group.





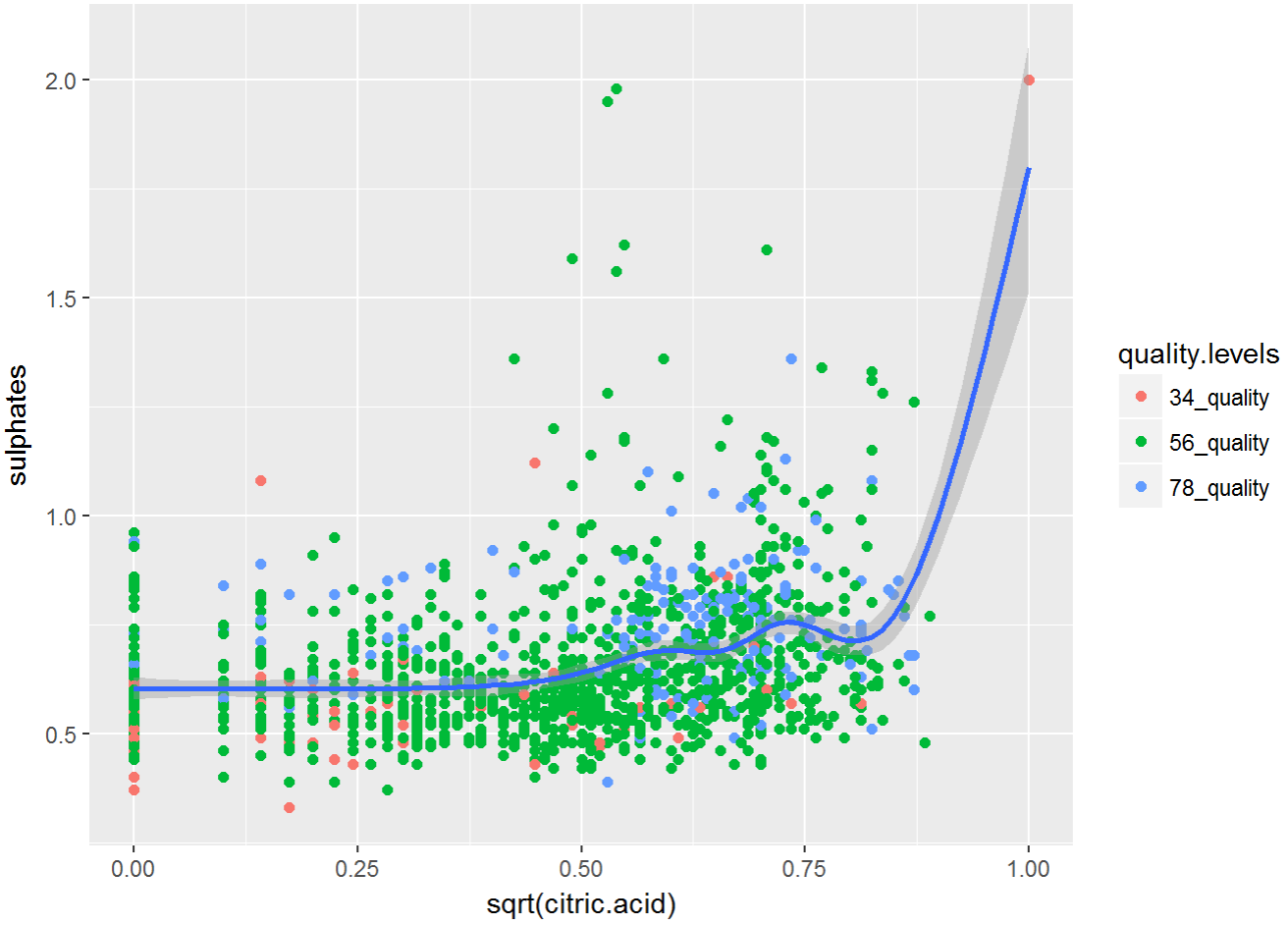
The lower quality wines are widely spread along the y-axis. Again I notice that the good wines (those of quality 7 and 8) mostly cluster in the upper right corner, thus having higher concentrations of citric acid. From the violin plot, we see that the citric acid is quite evenly distributed among the alcohol levels of 78_quality wines.


volatile.acidity
1.6
1.2
0.8
0.4
3.0
3.2
3.4
3.6
3.8
sqrt(alcohol)
quality.levels
34_quality
56_quality
78_quality


34_quality
56_quality
78_quality
volatile.acidity
1.6
1.2
0.8
0.4
10
12
14
alcohol
factor(alcohol.levels)
high_alcohol
low_alcohol
mod.high_alcohol
mod.low_alcohol

It is important to recall that volatile acidity and quality have a negative correlation coefficient. This can be noticed by analyzing the two graphs, the 34_quality wines have higher volatile acidity and less alcohol.

```
## `geom_smooth()` using method = 'gam'
```



From the plot citric.acid versus sulphates, I notice that in the 78_quality wines the quantity of sulphates is more consistent. The high peaks in sulphates are seen in the other two groups of wines, the lower value are more prevalent in the 34_quality wines.

Multilinear regression models

```
##
## Call:
## lm(formula = rwine$quality ~ rwine$alcohol + rwine$sulphates +
##     rwine$citric.acid)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7565 -0.3535 -0.1007  0.5067  2.2125
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.43392    0.17615   8.140 7.86e-16 ***
## rwine$alcohol      0.33841    0.01619  20.903  < 2e-16 ***
## rwine$sulphates    0.81403    0.10651   7.643 3.65e-14 ***
## rwine$citric.acid  0.51345    0.09284   5.531 3.72e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6842 on 1595 degrees of freedom
## Multiple R-squared:  0.2836, Adjusted R-squared:  0.2823
## F-statistic: 210.5 on 3 and 1595 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = rwine$quality ~ rwine$alcohol + rwine$sulphates +
##     rwine$citric.acid + rwine$volatile.acidity)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.71408 -0.38590 -0.06402  0.46657  2.20393
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              2.64592    0.20106  13.160  < 2e-16 ***
## rwine$alcohol            0.30908    0.01581  19.553  < 2e-16 ***
## rwine$sulphates          0.69552    0.10311   6.746 2.12e-11 ***
## rwine$citric.acid       -0.07913    0.10381  -0.762    0.446
## rwine$volatile.acidity -1.26506    0.11266 -11.229  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6588 on 1594 degrees of freedom
## Multiple R-squared:  0.3361, Adjusted R-squared:  0.3345
## F-statistic: 201.8 on 4 and 1594 DF,  p-value: < 2.2e-16
```

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?

I considered the relations among five attributes: alcohol, sulphates, citric acid, volatile acidity and quality. In order to avoid overlapping and redundancy, I grouped the quality in three different levels.

The 78_quality wines (the best wines) have higher levels of alcohol. The volatile acidity is low in these wines. At the other end, the 34_quality wines have lower sulphates and higher volatile acidity. I also notice that this group of wines contain samples with extremely high sulphates, citric acid and volatile acidity.

It seems more evident now that while the good wines have a high alcohol level, higher levels of sulphates and low volatile acidity, the poor quality wines (those of quality 3 and 4) are in particular characterized by high volatile acidity.

If the alcohol is low (less than 9.4%) the sulphates and the citric acid do not have a significant effect on the quality of the wine, most of these wines are of average 5 and 6 quality.

From the relation between citric acid and sulphates, I can see that the better quality is associated to higher levels of citric acid. Many of the low quality wines (34_quality group) have very low levels of citric acid.

Were there any interesting or surprising interactions between features?

It is interesting to notice that lower level of sulphates have a negative effect on the quality of the wine, regardless of the amounts of citric acid and alcohol. It seems that alcohol makes better wines while not enough sulfates the opposite.

Also, it looks like excessive levels of sulphates or alcohol do not make the wine a bad wine, just an average quality wine. On the other side, high levels of volatile acidity have a clear negative impact on the quality.

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

I created two multiple linear regression models. The first relates quality to sulphates, alcohol and citric acid. The second takes into account the volatile acidity also.

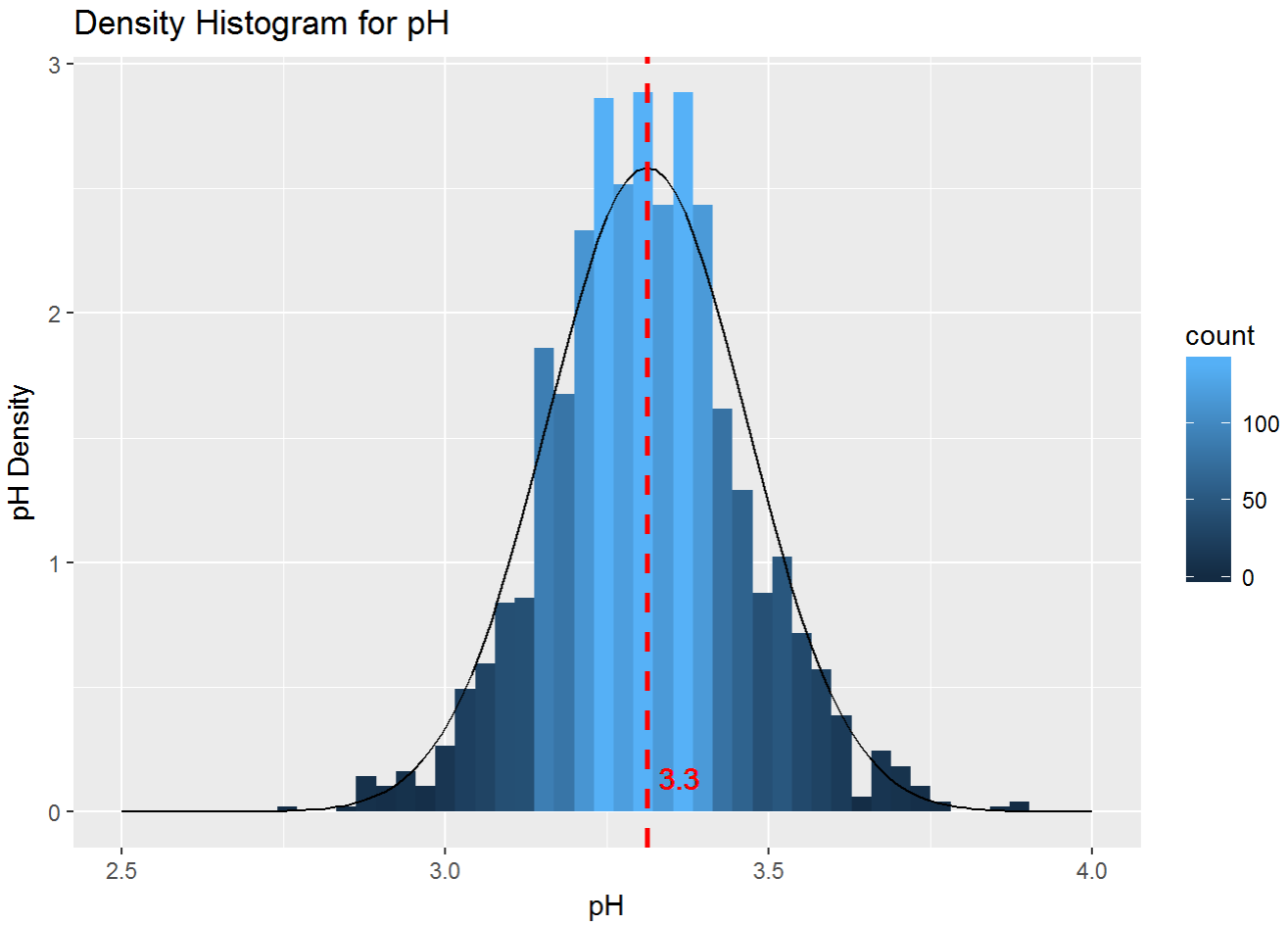
In the first linear model, all three attributes have the same significance when deciding the quality of the wine. The R-squared value of 0.28 does not indicate a very good correlation.

The median residual for the second model is smaller than in the first model. Interestingly, when

the volatile acidity is taken into account, the citric acid significance level drops and the coefficient changes from 0.51 in the first model to -0.08 in the second model. These changes might be due to multicolinearity introduced in the model, by considering two types of acid contents. In the second model, R-squared is higher with a value of 0.34.

Final Plots and Summary

Plot One




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

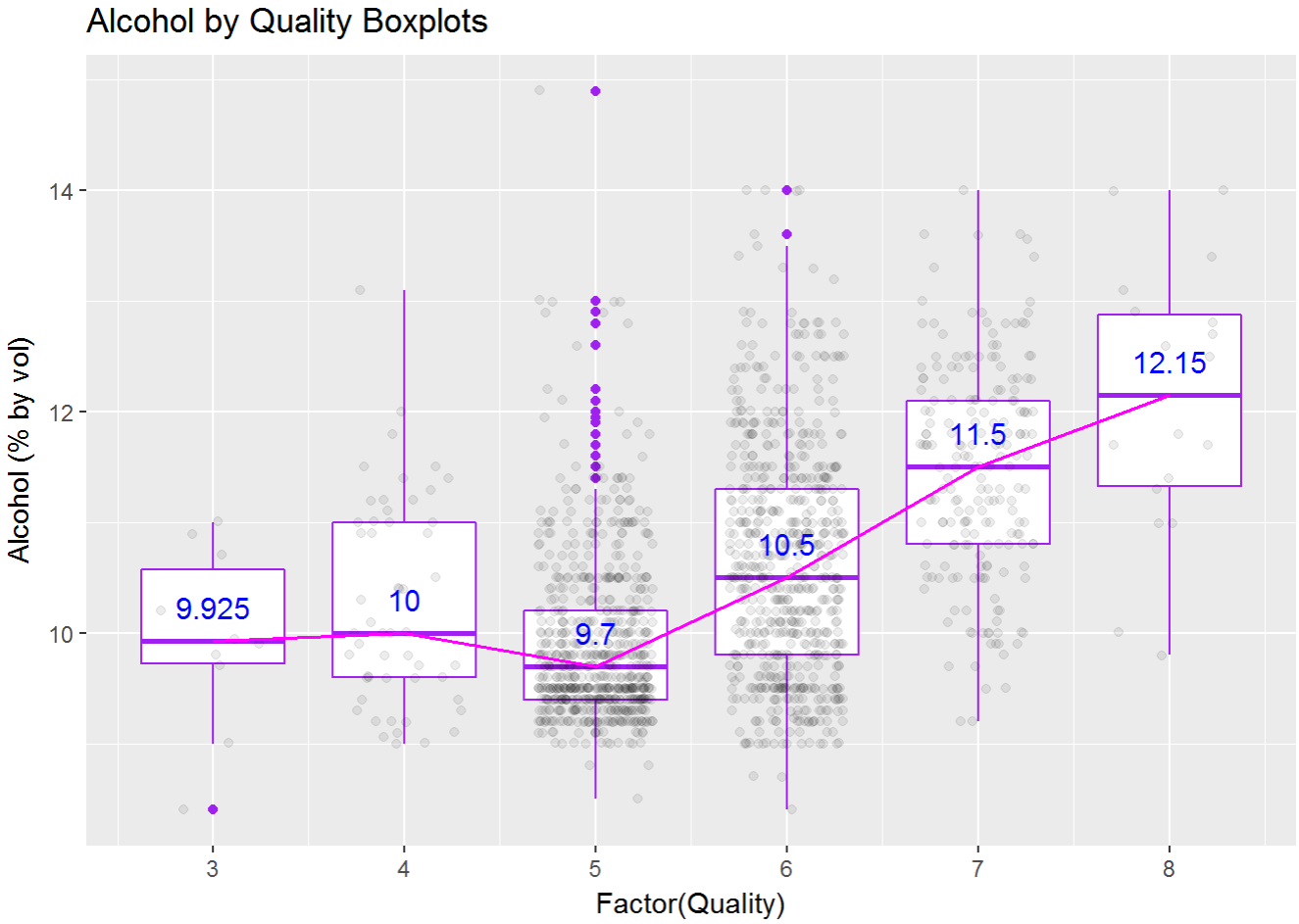
```
## Standard deviation is: 0.1543865
```

Description One

One of the first unexpected findings was that pH does not have too much relevance to the quality of the wines for this dataset. The pH levels have a normal distribution. The normal curve overlaps almost perfectly over the density histogram. The mean level of pH is 3.3 and the

standard deviation is 0.15.

Plot Two



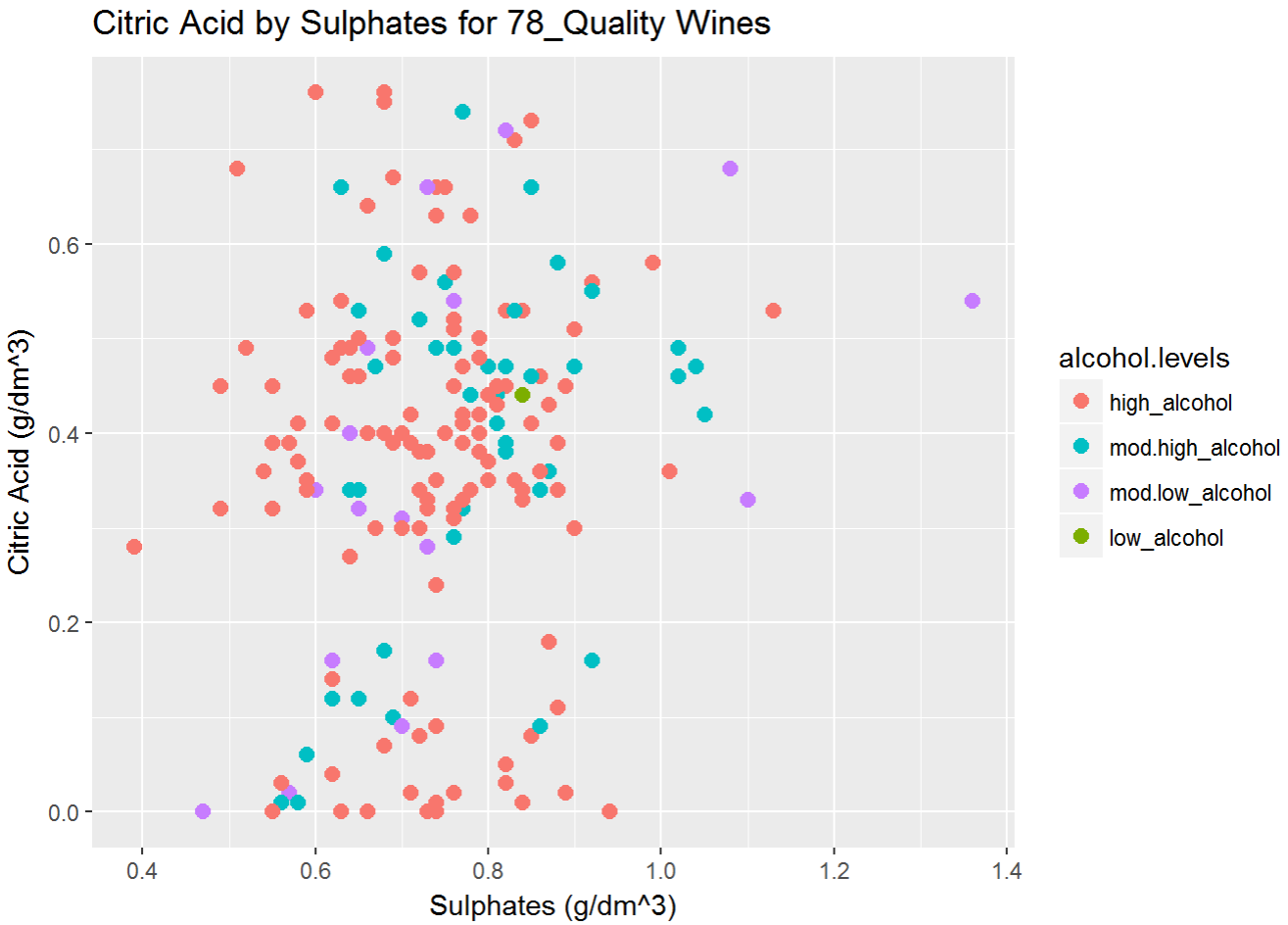

```
## Alcohol descriptive statistics for quality = 5
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     8.5     9.4     9.7     9.9    10.2    14.9
```

Description Two

The attribute that has the highest correlation coefficient with the quality is the alcohol. It became clear after analyzing the data that the best wines have higher alcohol percentages. The set of boxplots give the alcohol distribution for each quality, the median levels are displayed. There is a sharp increase in the alcohol levels from level 5 to level 8. Somehow intriguing is the data for quality 5, which does not exactly align with the general pattern. In this group we also find almost all the outliers.

Plot Three



Description Three

The scatterplot is only for the best wines in the data set, those of qualities 7 and 8. It contains information about the three attributes that have the most direct correlation with quality: sulphates, citric acid and alcohol. Most of these wines have high alcohol content. We are also able to notice a higher density of samples with citric acid around the value 0.40. The sulphates are consistenntly lower than 0.90.

Reflections

The data I analyzed in this project contains 1599 samples of red wines and 11 physicochemical attributes. The main goal is to determine how these attributes affect the quality of the wine.

Some of these attributes are correlated, for example the total acidity and the citric acid, the total sulfur dioxide and the free sulfur dioxide, pH and total acidity.

Many of the attributes (such as fixed acidity, residual sugar, free sulfur dioxide, etc.) have skewed distributions, with long tails to the right only. It seems that it is more likely for these wines to have higher than average quantities of these chemicals, as I did not see too many outliers to the left of the median lines.

The analysis indicates that the alcohol by far has the greatest influence on the quality of the

wine. The next attributes to influence the quality are the sulphates, followed by citric acid. I must admit that I am surprised that sulphates have such relevance on quality. I also noticed in [Cortez et al., 2009] that their regression models predicted sulphates as the factor with highest relevance.

I initially thought that the residual sugar will play a more significant role in deciding the quality. This is probably true more for white wines than for the red ones. In the initial stages of the project I constructed two factor variables for the residual sugar and pH which I did not get the chance to use too much, as these attributes had small correlation with quality.

Analyzing the citric acid versus other attributes, I noticed around the value 0.40 for citric acid there is a slight drop in residual sugar and consequently density. I also noticed that the high quality wines with this level of acidity seem to have sulphates levels with wider distribution than similar wines that have citric acid of 0.20 or 0.60 g/dm^3.

On a scale from 0 to 10, the quality for the wines in the dataset vary from 3 to 8. Also among these wines only 217 have qualities 7 or 8. Most wines are of average quality. It is possible that the average nature of these wines makes the alcohol such a relevant factor, which overcomes other aspects of a good wine such as bouquet, which is influenced by other chemical factors.

My biggest struggle with this analysis relies in the fact that the data does not have too much variation, the results are mostly average and the quality does not significantly change when the properties change. There seem to be many factors to take into account at the beginning, but after looking at the correlation coefficients I was able to narrow them down significantly. I think that using factor variables, for alcohol levels and for quality also helped me to understand the relations between various attributes without the hassle of too many details.

Given the nature of the physicochemical properties, I would look for correlations among these attributes. Also, it would be interesting to further investigate if there is anything unusual with the wines that have citric acid of 0.40 g/dm^3.

Although there are plenty of attributes to choose from already included in the dataset, some standard attributes are not present here, such as tannins and phenols, both with high relevance on the wine bouquet. I would like to see how these attributes influence the quality of the wine.

Another factor that seems to be missing from this dataset is any type of information on the wine's bouquet or aroma. Such addition to the data will make the results more revealing to a wine lover.

The wines analysed here come from the same region, and are of the same type. I think that some information on the climate and terroir where these wines are produced would add valuable insight into the properties of these wines.

References

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

http://waterhouse.ucdavis.edu/whats-in-wine