WHITE WINE INVESTIGATION by Franz Hastrup-Nielsen
date: May 15, 2018
output: html_document
==========

This report is going through the steps of Exploratory Data Analysis (EDA) using a dataset that contains information about white wines and their associated quality.

Univariate Plots Section

Input variables and short explanation:
1 - fixed.acidity fixed acidity (tartaric acid - g / dm^3)
2 - volatile.acidity volatile acidity (acetic acid - g / dm^3)
3 - citric.acid citric acid (g / dm^3)
4 - res.sugar residual sugar (g / dm^3)
5 - chlorides chlorides (sodium chloride - g / dm^3
6 - free.so2 free sulfur dioxide (mg / dm^3) 7 - total.so2 total sulfur dioxide (mg / dm^3)
8 - density density (g / cm^3)
9 - pH pH
10 - sulphates sulphates (potassium sulphate - g / dm3)
11 - alcohol alcohol (% by volume)
Output variable (based on sensory data):
12 - quality quality (score between 0 and 10)

```
## [1] 4898   13
```

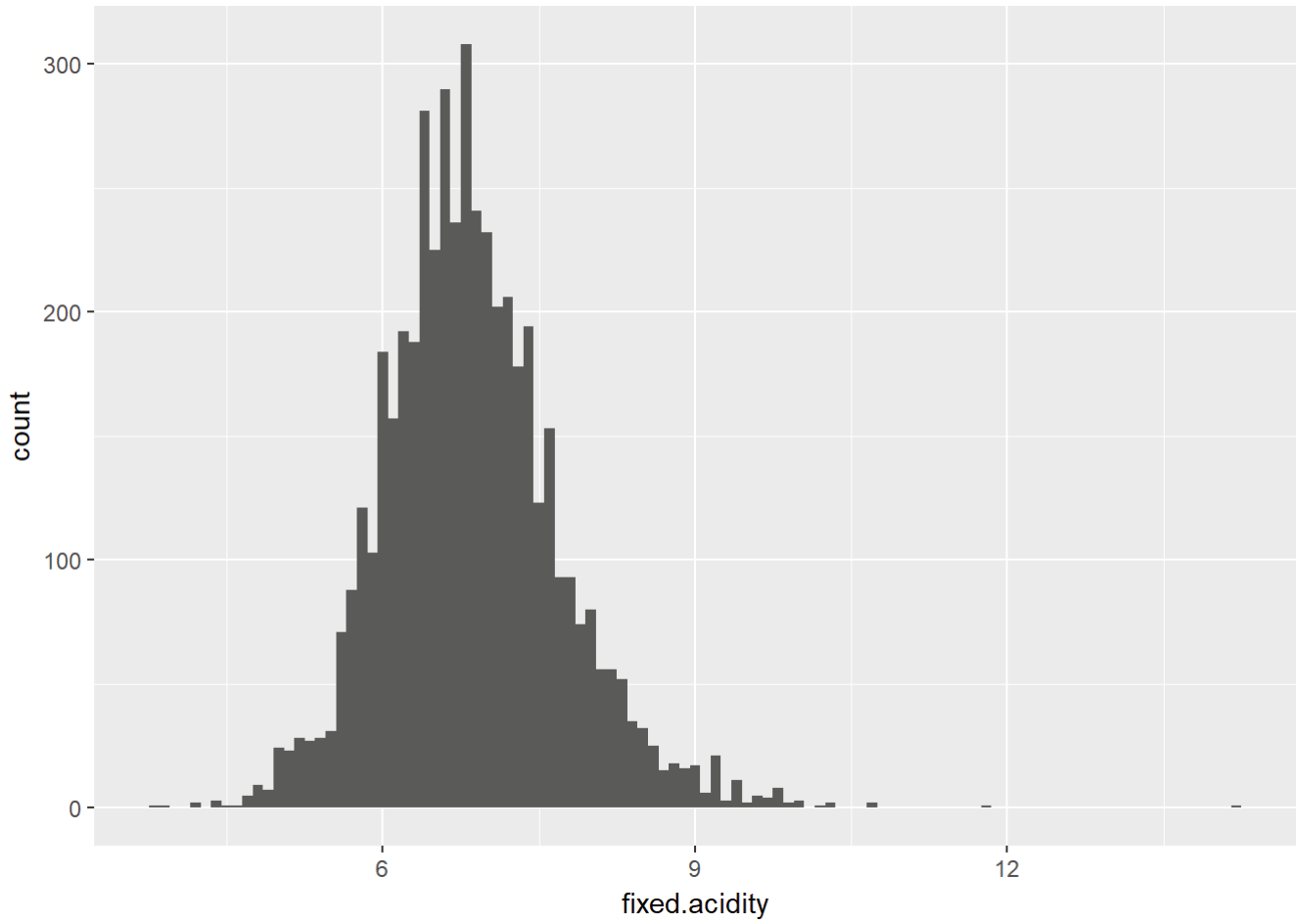
The dataset includes 4898 rows of wines and 12 columns assigned to each wine.

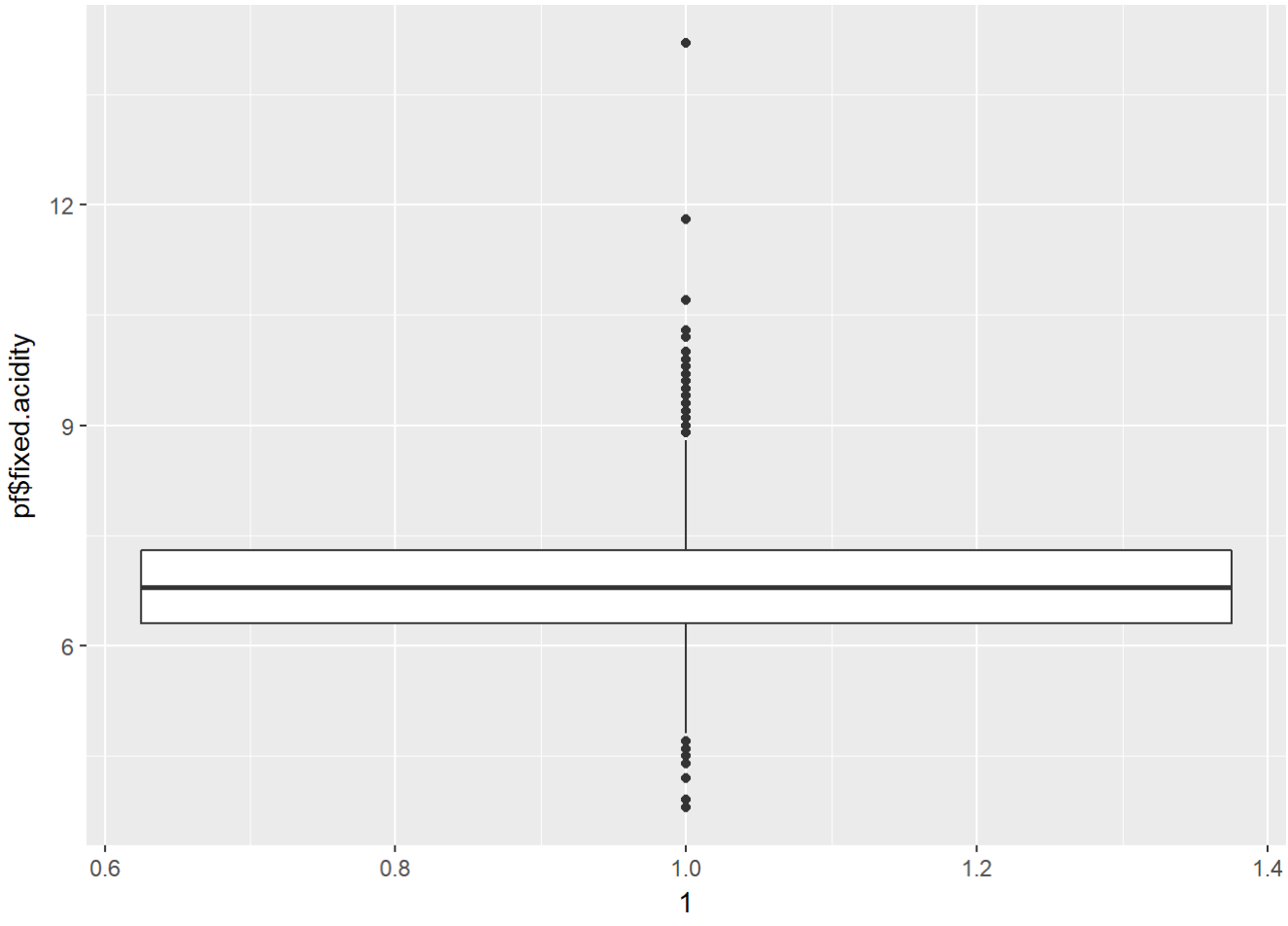
Here are the variables and a summary of the dataset to be investigated:

```
##  [1] "X"              "fixed.acidity"    "volatile.acidity"
##  [4] "citric.acid"    "res.sugar"        "chlorides"
##  [7] "free.so2"       "total.so2"        "density"
## [10] "pH"             "sulphates"        "alcohol"
## [13] "quality"
```

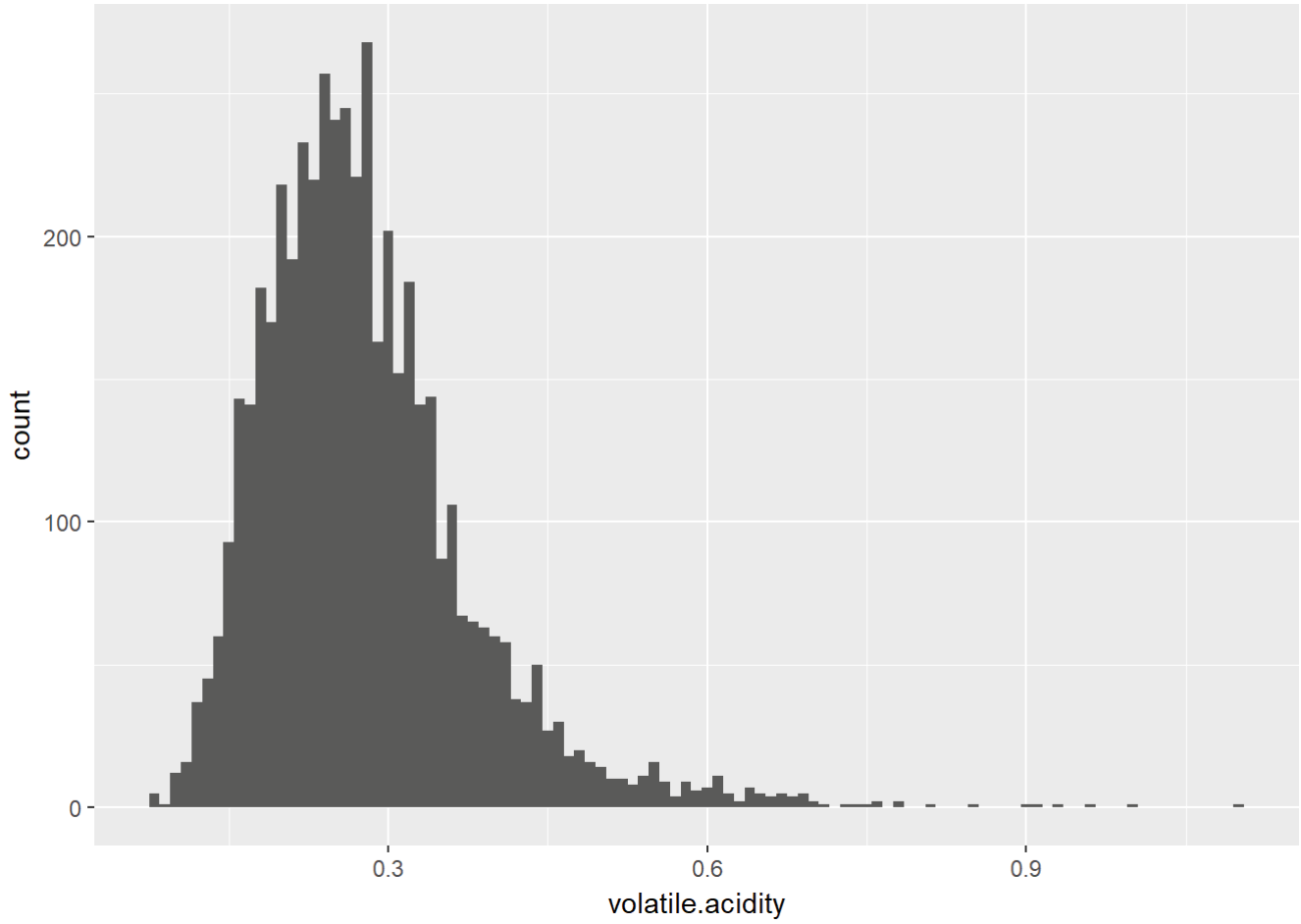
```
##   X fixed.acidity volatile.acidity citric.acid res.sugar chlorides
## 1 1           7.0             0.27        0.36      20.7     0.045
## 2 2           6.3             0.30        0.34       1.6     0.049
## 3 3           8.1             0.28        0.40       6.9     0.050
## 4 4           7.2             0.23        0.32       8.5     0.058
## 5 5           7.2             0.23        0.32       8.5     0.058
## 6 6           8.1             0.28        0.40       6.9     0.050
##   free.so2 total.so2 density   pH sulphates alcohol quality
## 1       45       170  1.0010 3.00      0.45     8.8       6
## 2       14       132  0.9940 3.30      0.49     9.5       6
## 3       30        97  0.9951 3.26      0.44    10.1       6
## 4       47       186  0.9956 3.19      0.40     9.9       6
## 5       47       186  0.9956 3.19      0.40     9.9       6
## 6       30        97  0.9951 3.26      0.44    10.1       6
```

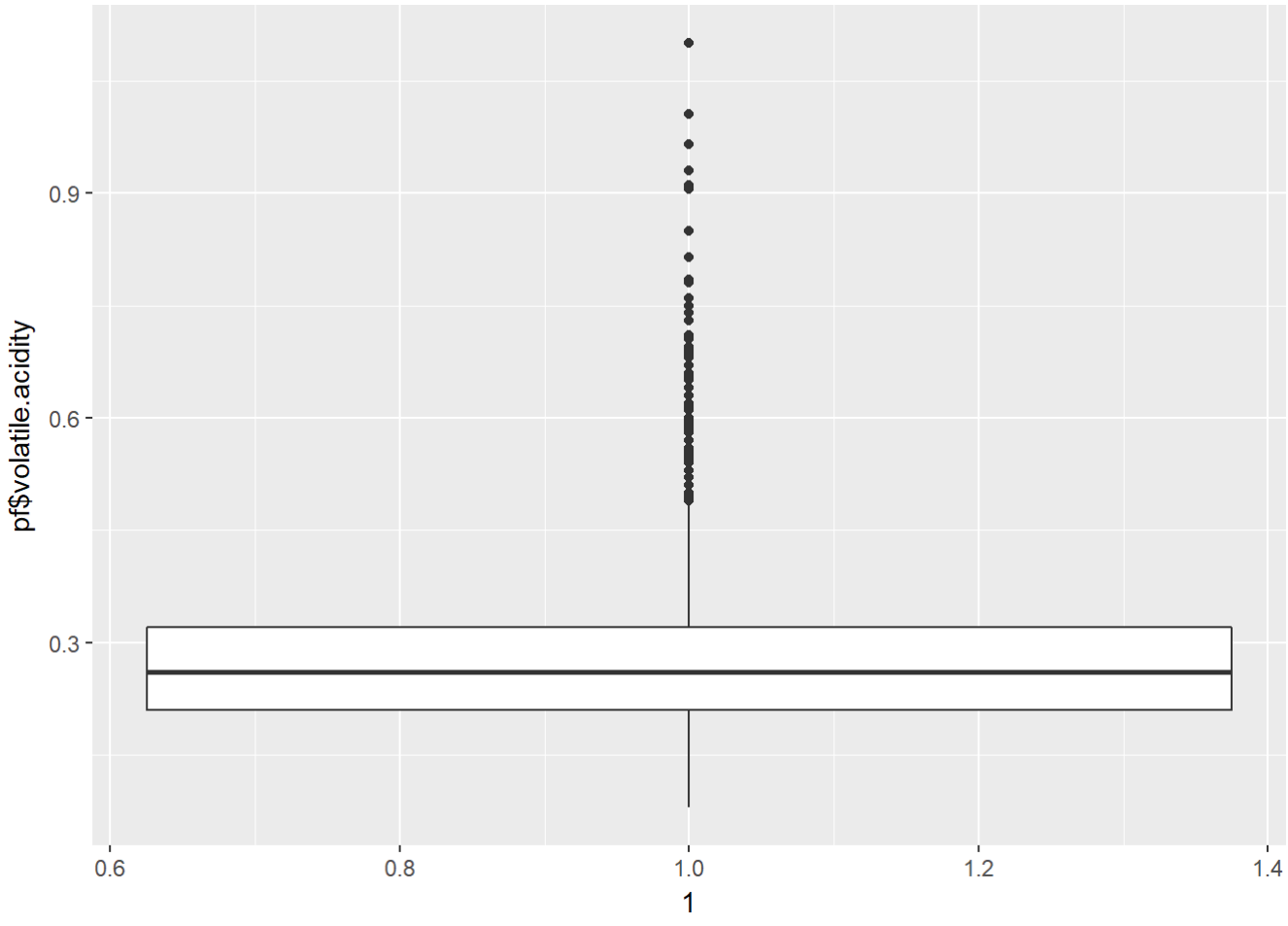
```
##        X        fixed.acidity    volatile.acidity  citric.acid
##  Min.   :   1   Min.   : 3.800   Min.   :0.0800   Min.   :0.0000
##  1st Qu.:1225   1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700
##  Median :2450   Median : 6.800   Median :0.2600   Median :0.3200
##  Mean   :2450   Mean   : 6.855   Mean   :0.2782   Mean   :0.3342
##  3rd Qu.:3674   3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900
##  Max.   :4898   Max.   :14.200   Max.   :1.1000   Max.   :1.6600
##    res.sugar        chlorides          free.so2        total.so2
##  Min.   : 0.600   Min.   :0.00900   Min.   :  2.00   Min.   :  9.0
##  1st Qu.: 1.700   1st Qu.:0.03600   1st Qu.: 23.00   1st Qu.:108.0
##  Median : 5.200   Median :0.04300   Median : 34.00   Median :134.0
##  Mean   : 6.391   Mean   :0.04577   Mean   : 35.31   Mean   :138.4
##  3rd Qu.: 9.900   3rd Qu.:0.05000   3rd Qu.: 46.00   3rd Qu.:167.0
##  Max.   :65.800   Max.   :0.34600   Max.   :289.00   Max.   :440.0
##     density             pH          sulphates         alcohol
##  Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00
##  1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50
##  Median :0.9937   Median :3.180   Median :0.4700   Median :10.40
##  Mean   :0.9940   Mean   :3.188   Mean   :0.4898   Mean   :10.51
##  3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40
##  Max.   :1.0390   Max.   :3.820   Max.   :1.0800   Max.   :14.20
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.878
##  3rd Qu.:6.000
##  Max.   :9.000
```





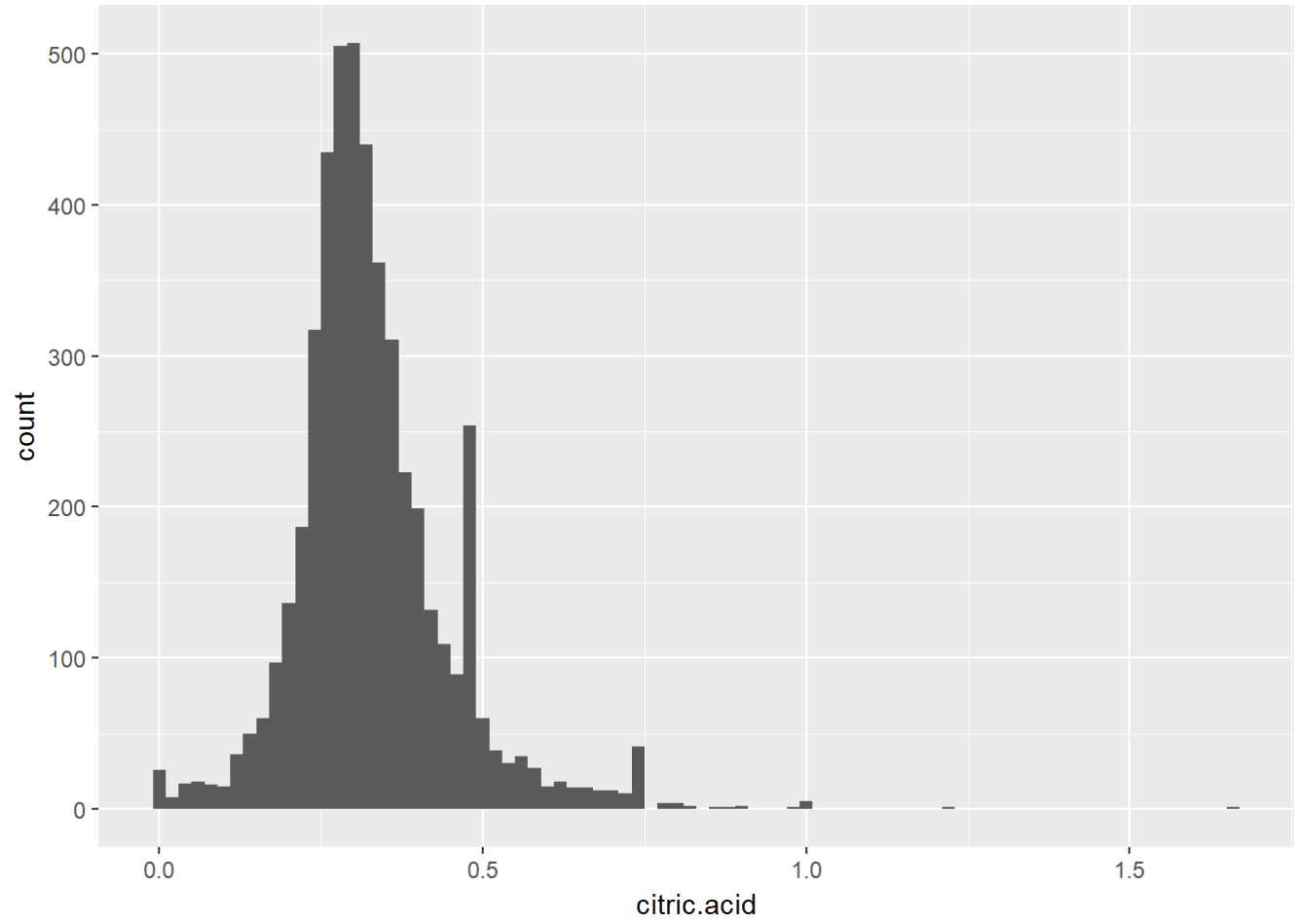

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.800   6.300   6.800   6.855   7.300  14.200
```

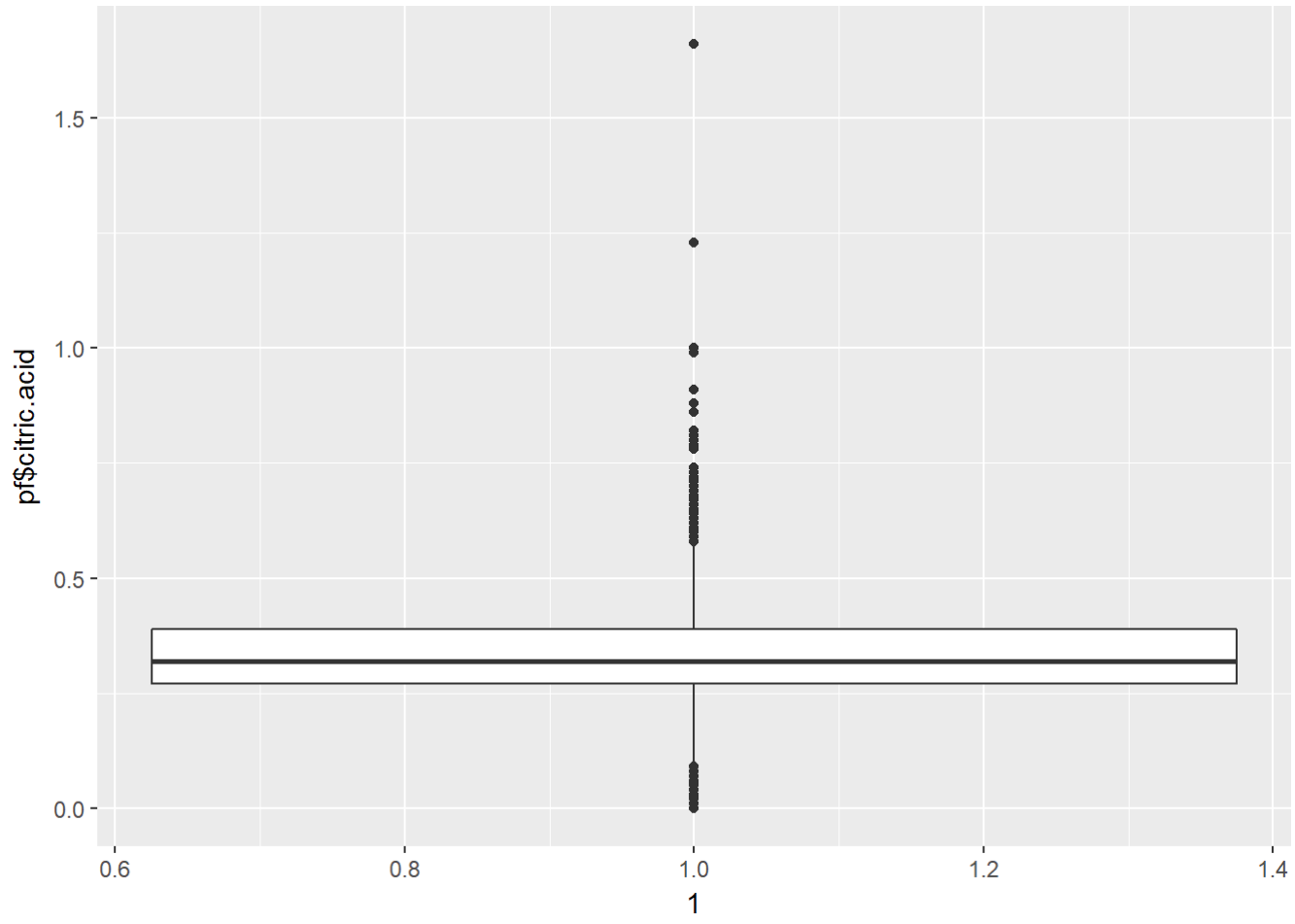





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.2100  0.2600  0.2782  0.3200  1.1000
```

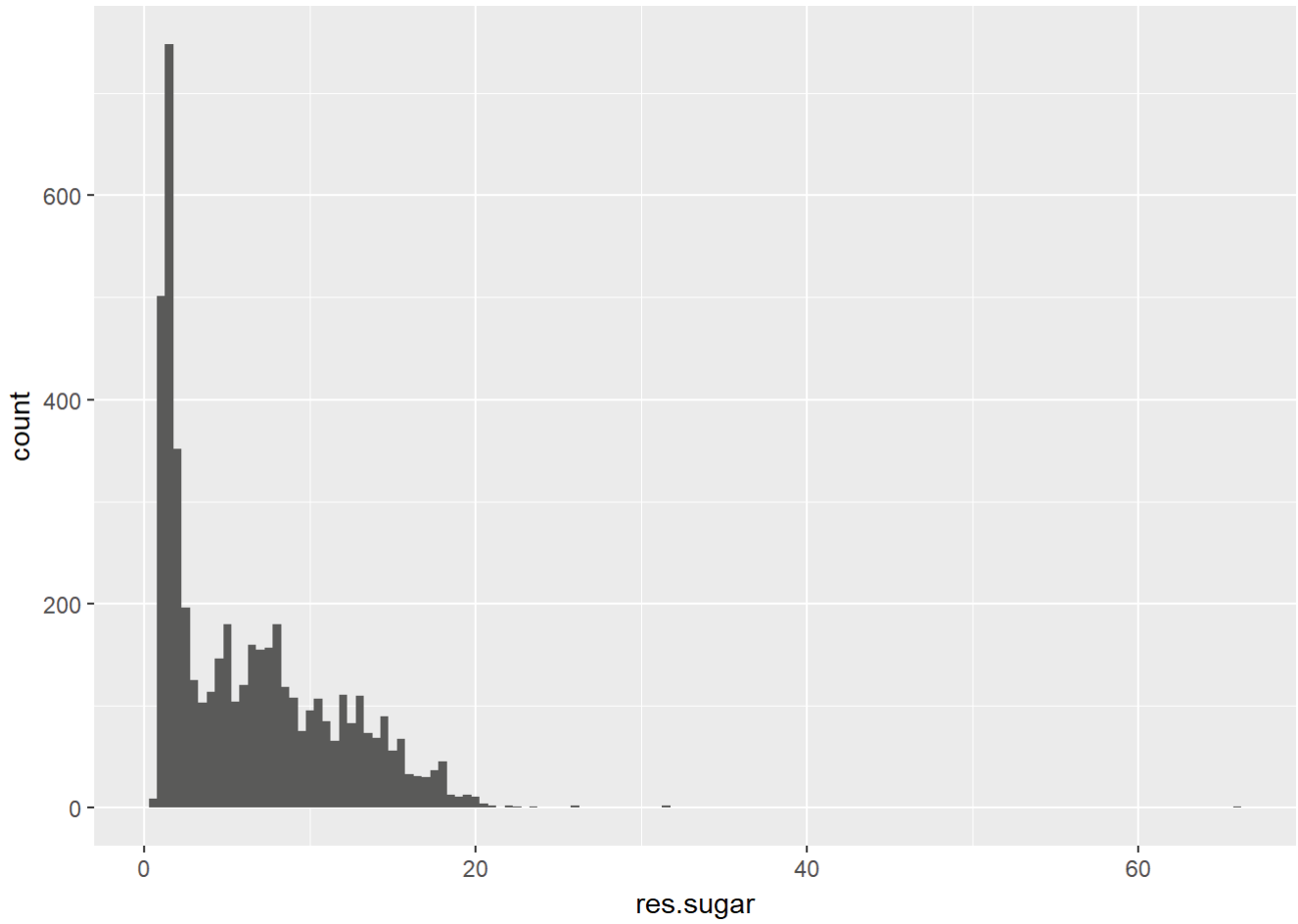
Both the fixed and volatile acidity seems to be normally distributed with a positive skew, as a small tail forms towards the higher values. As the boxplot shows, there seems to be a larger range of high values for both cases.

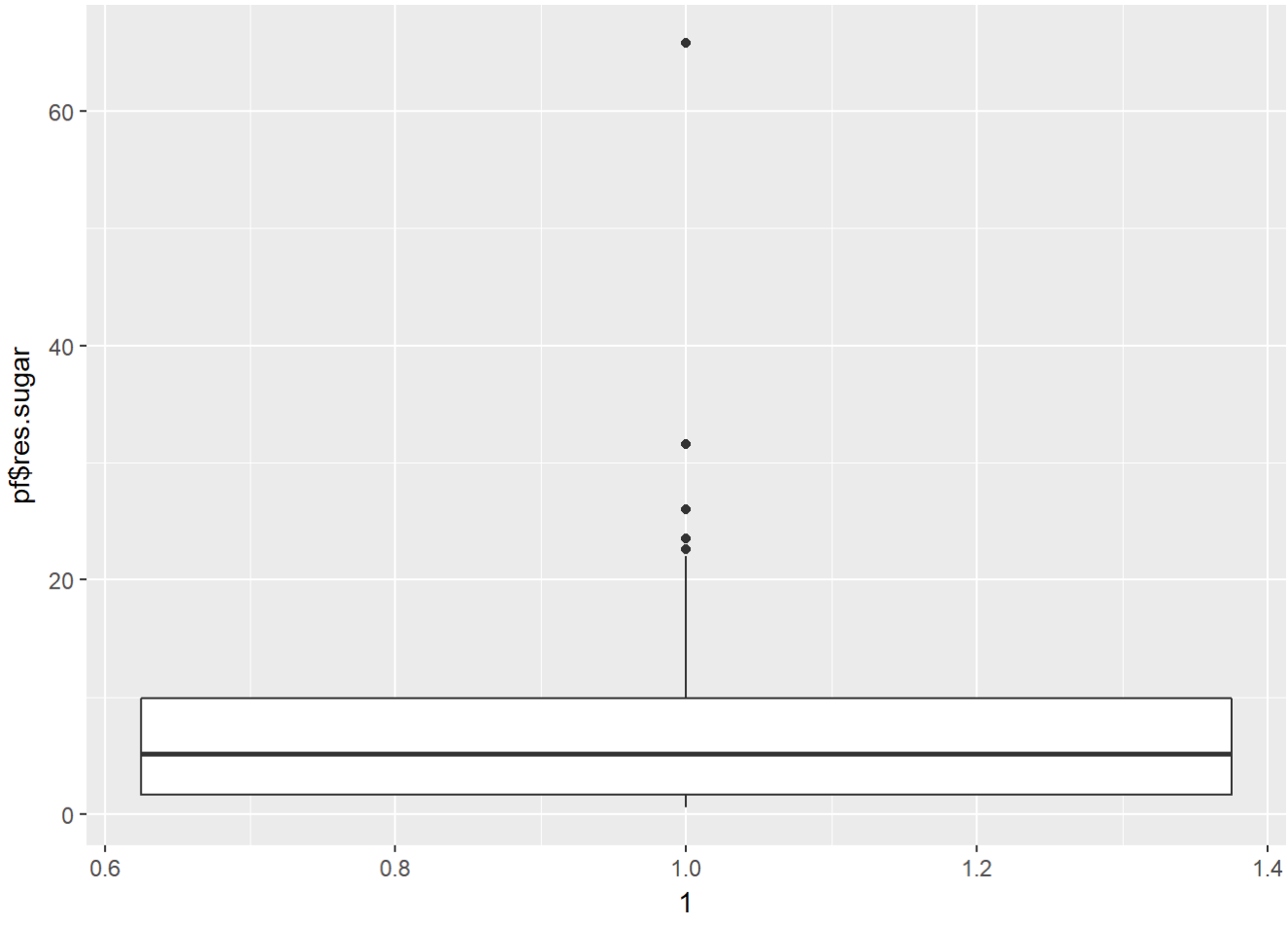




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2700  0.3200  0.3342  0.3900  1.6600
```

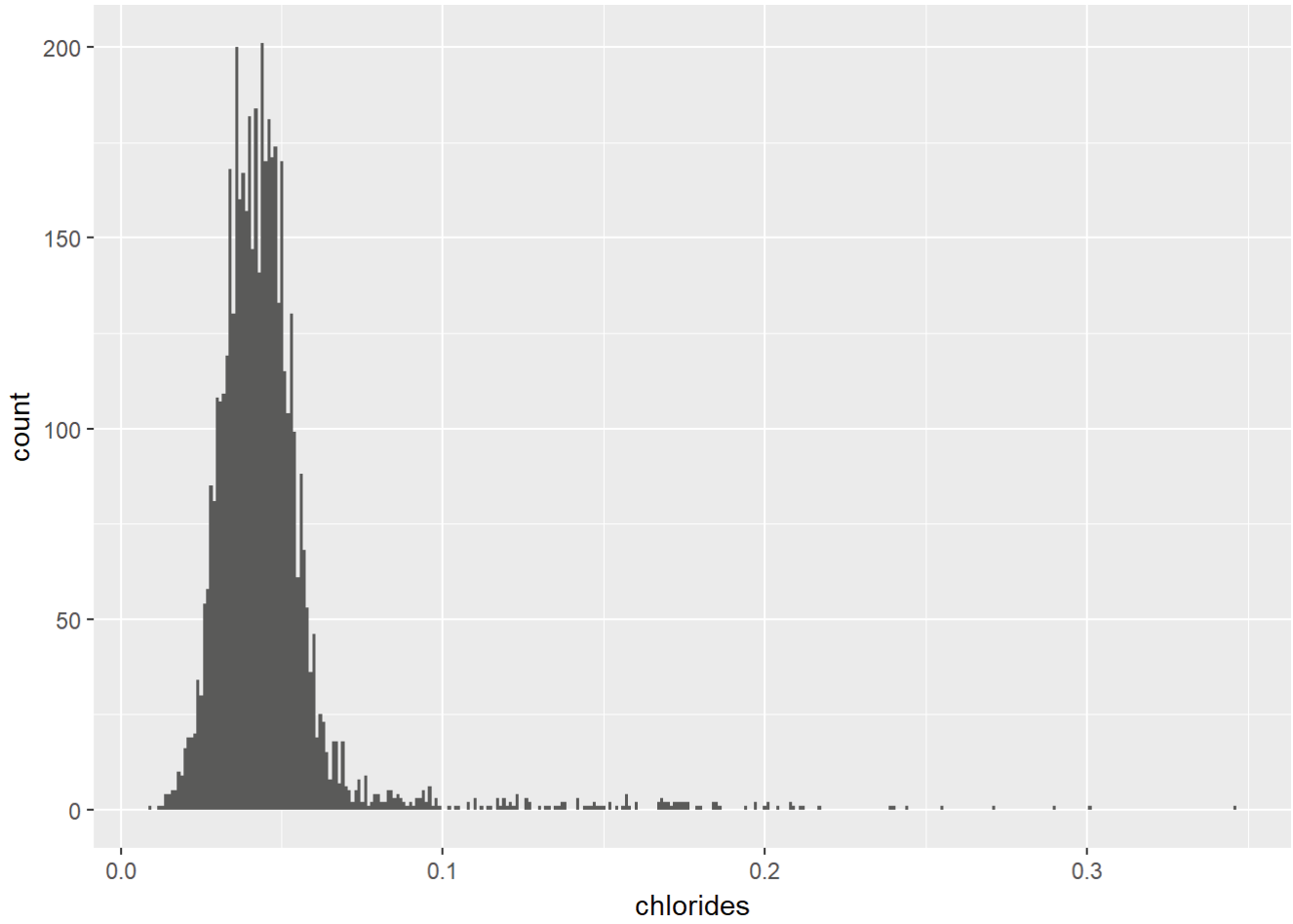
Citric acid is normally distributed with a positive skew due to a longer tail towards the higher values. Furthermore, it seems that two values on the right side is more prevalent close to 0.5 and 0.75. This could need further investigation. The boxplot shows the positive skew as the median is closer to the 25% quantile.

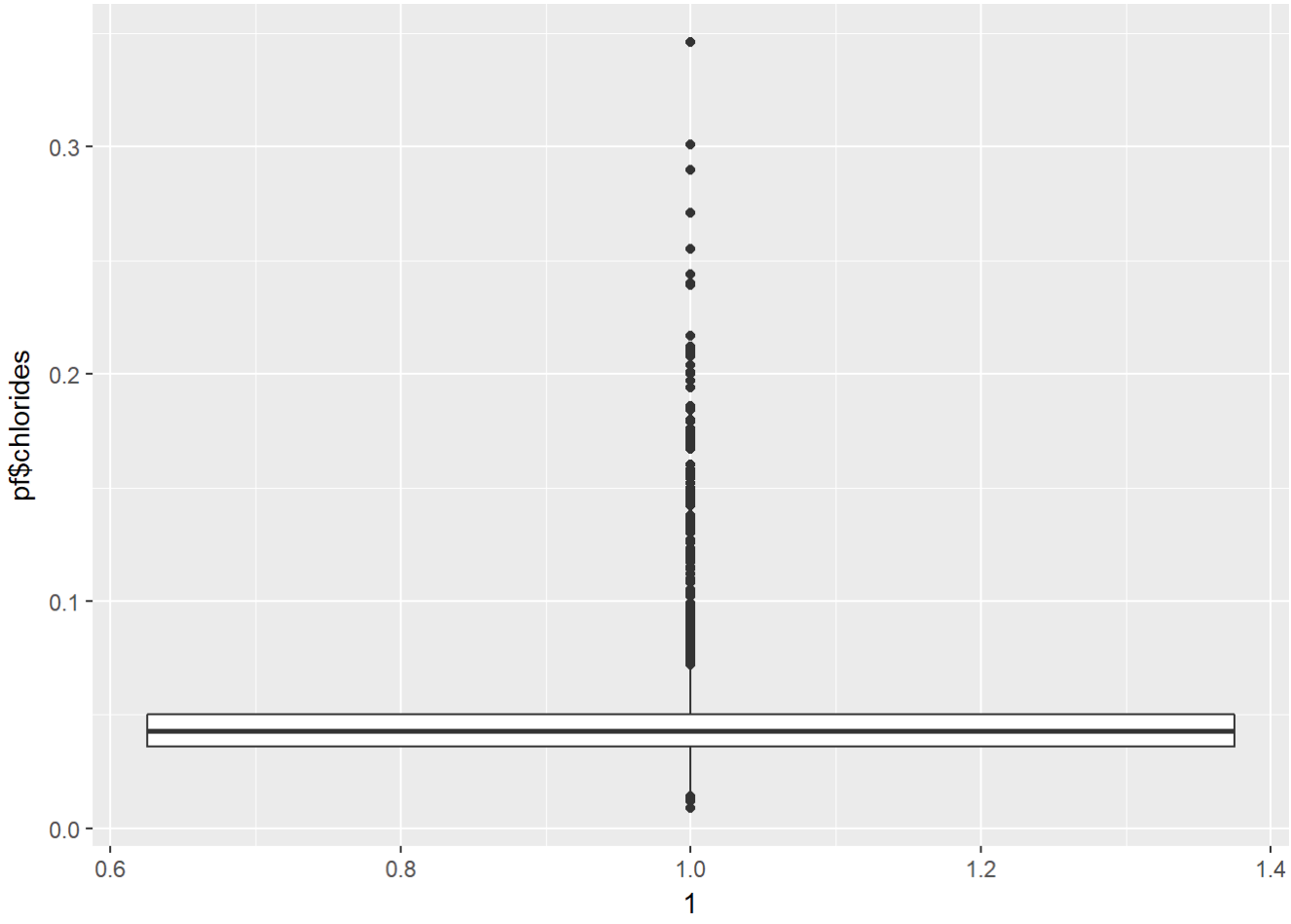





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.600   1.700   5.200   6.391   9.900  65.800
```

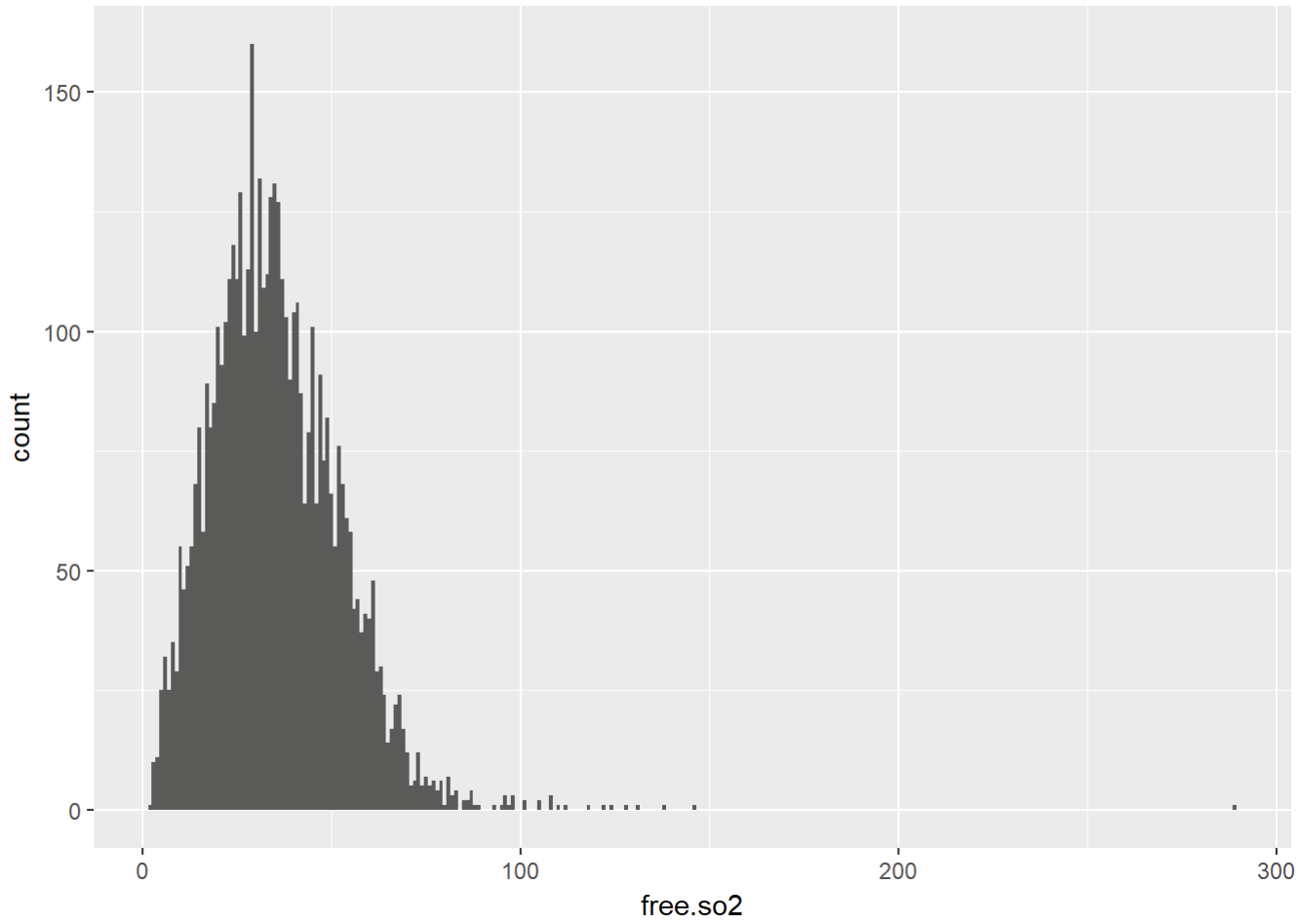
Residual sugar has a normal distribution with positive skew and some outliers at very high values as indicated by the boxplot.

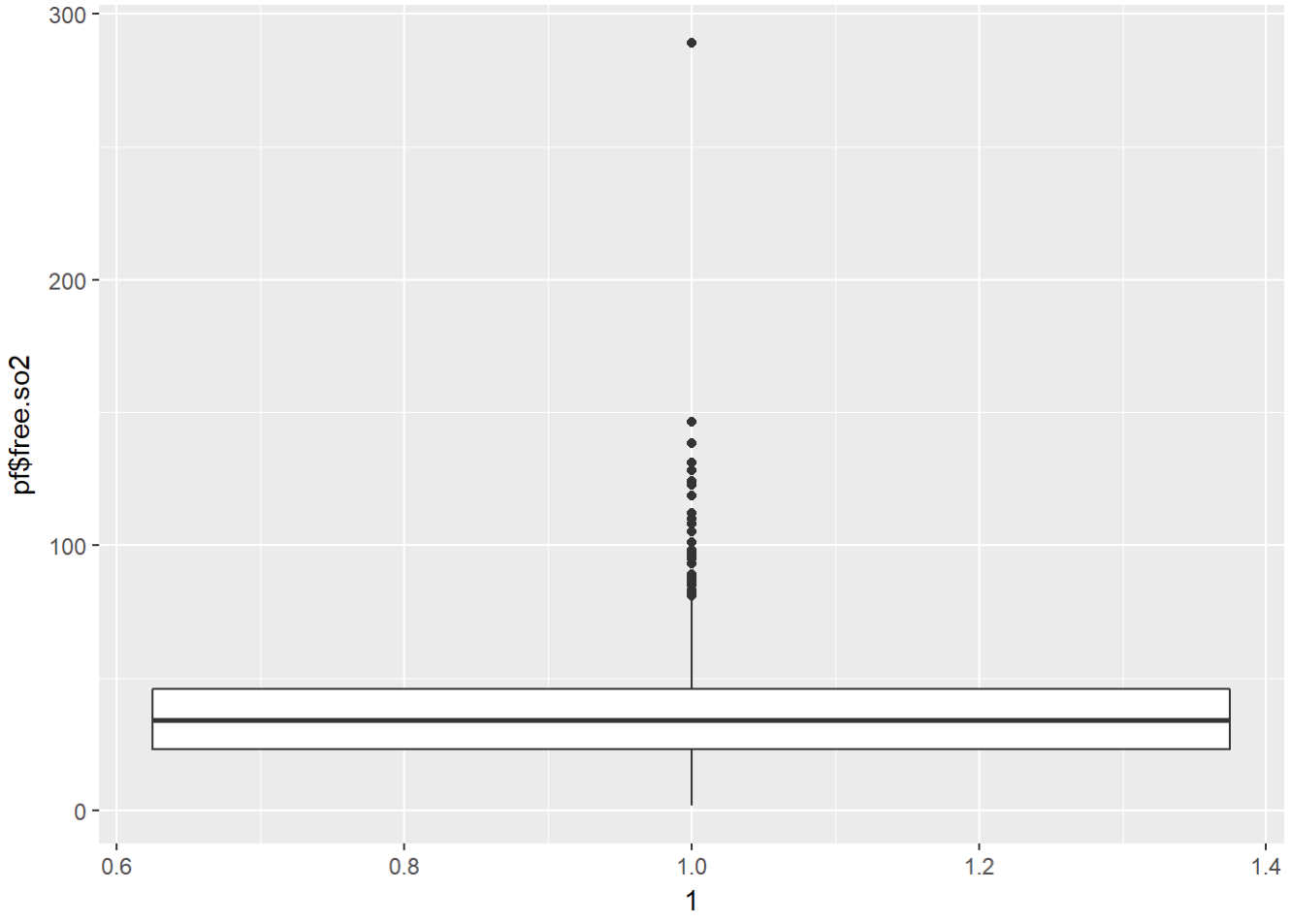




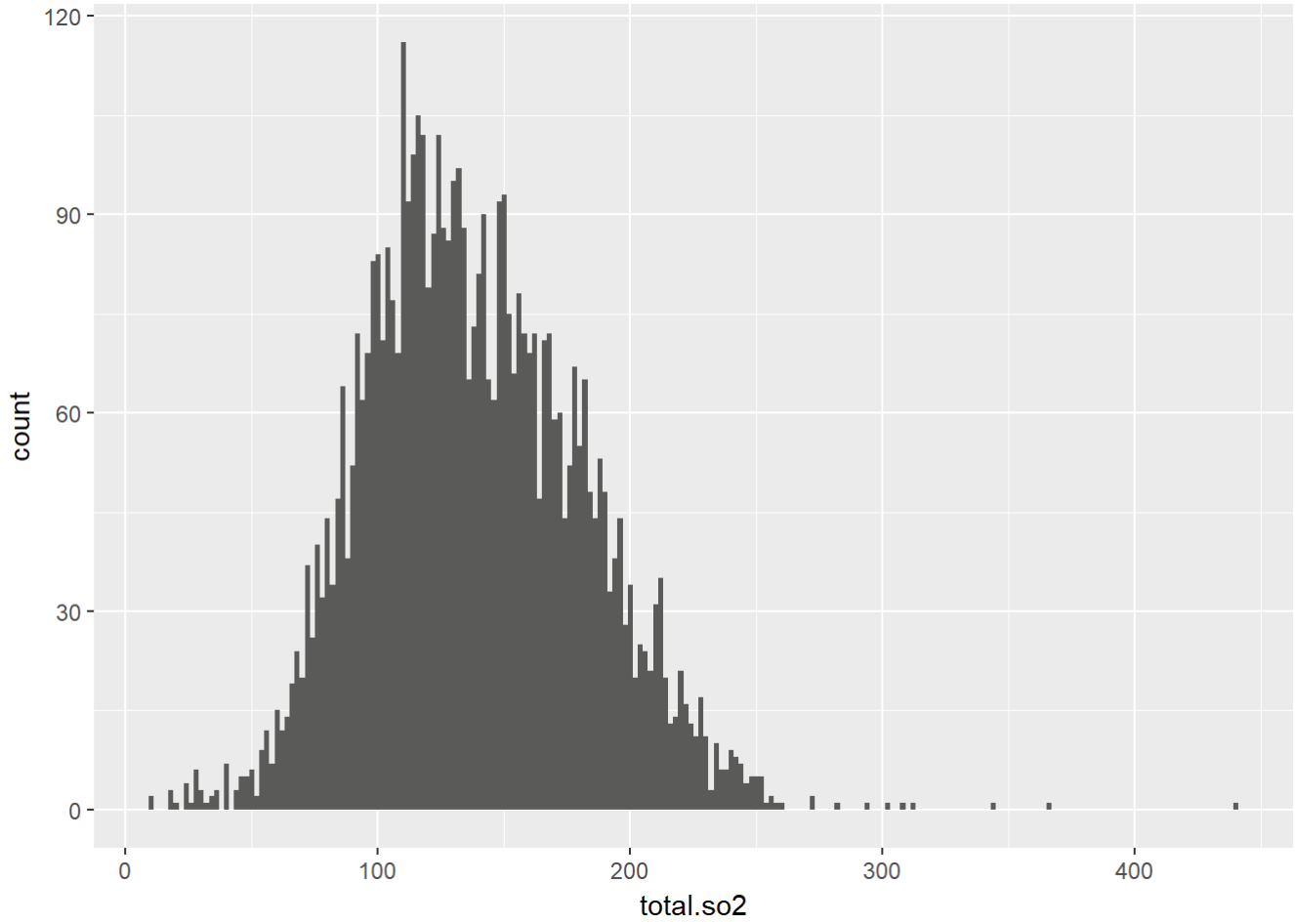
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00900 0.03600 0.04300 0.04577 0.05000 0.34600
```

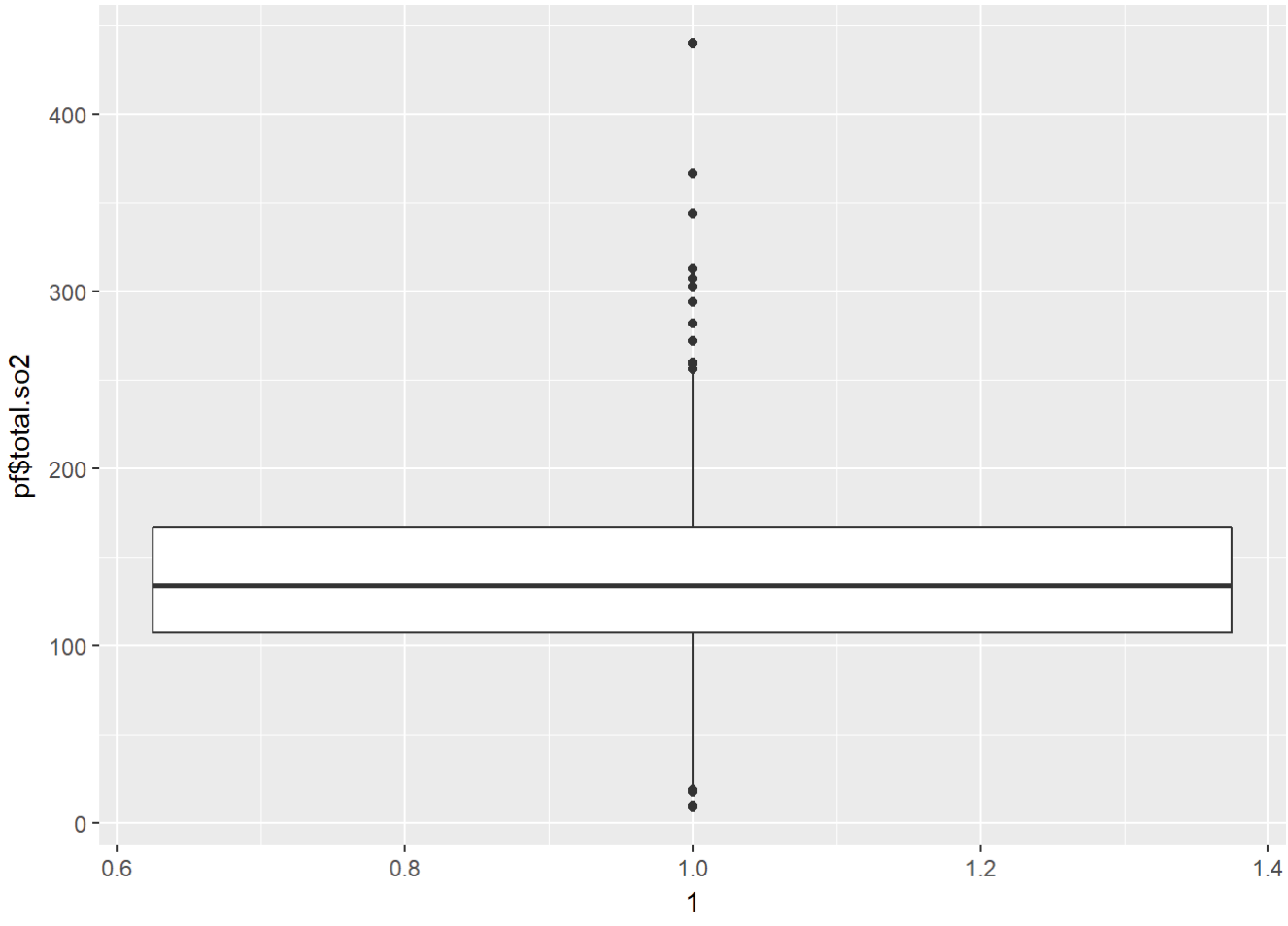
Chlorides behaves like the rest of the data with a normal distribution with a positive skew. With a low binwidth the distribution could be displaying some bi-modal behavior





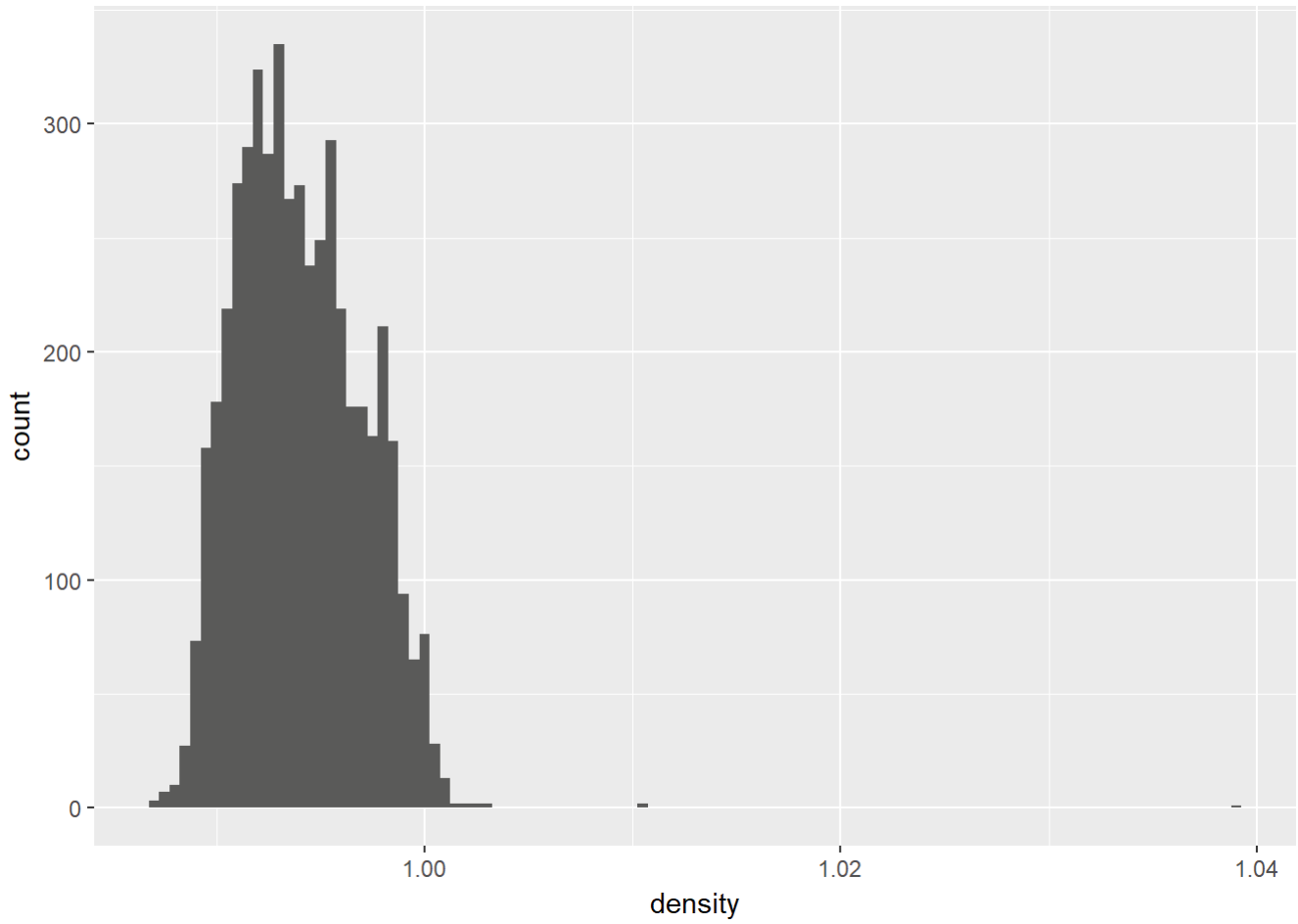

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   23.00   34.00   35.31   46.00  289.00
```

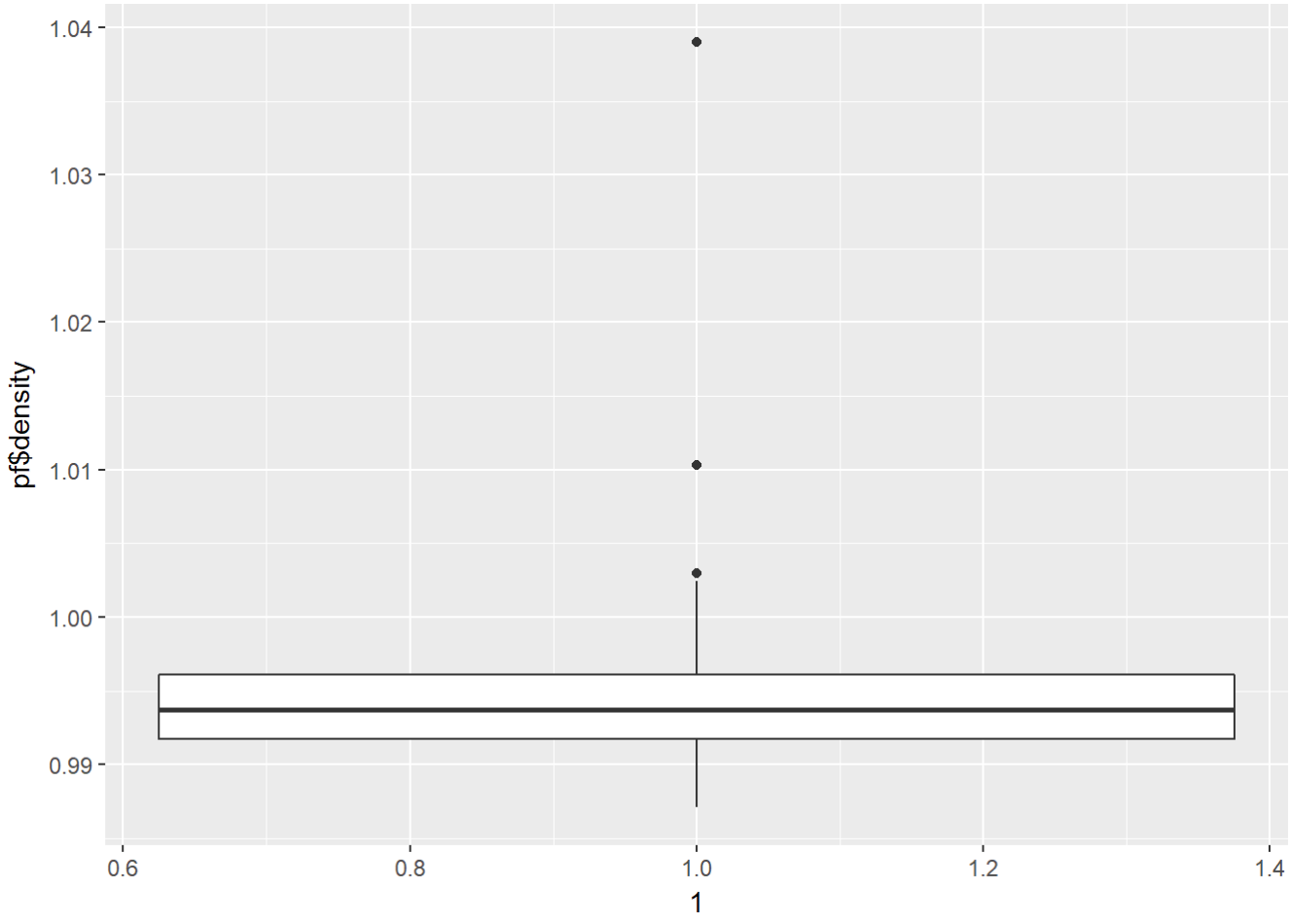




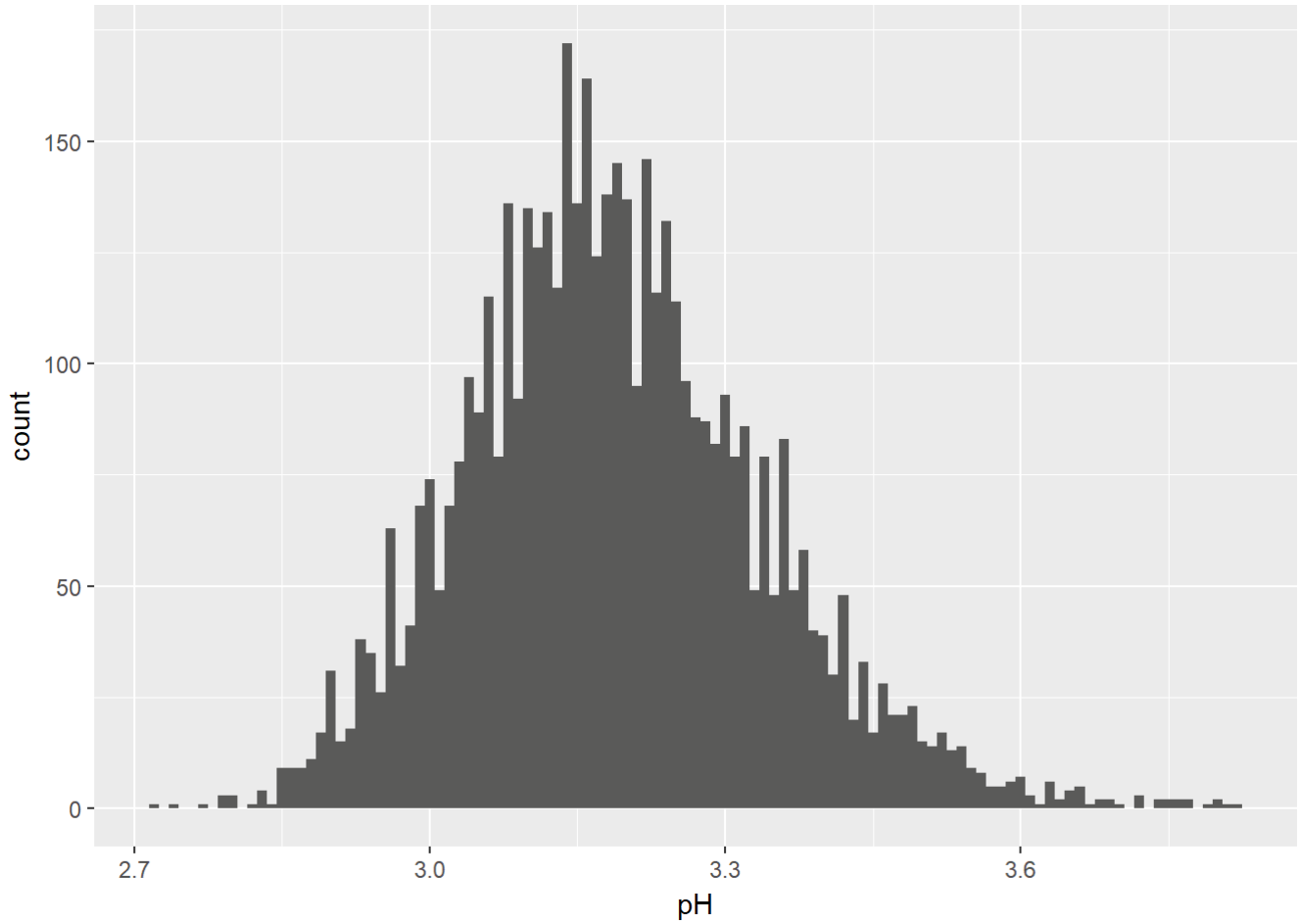

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.0   108.0   134.0   138.4   167.0   440.0
```

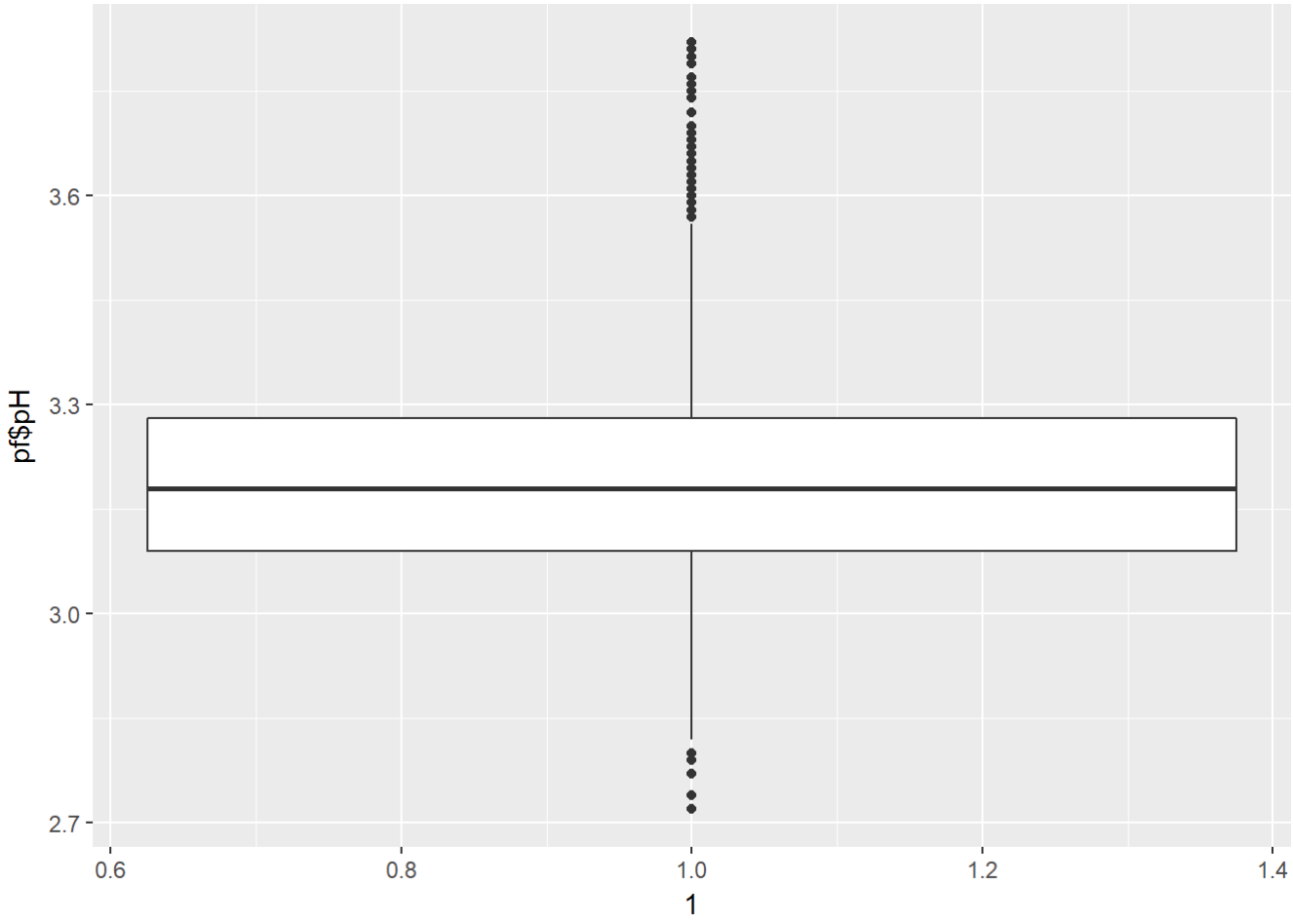
The free and total sulfur dioxide are both normally distributed with a positive skew



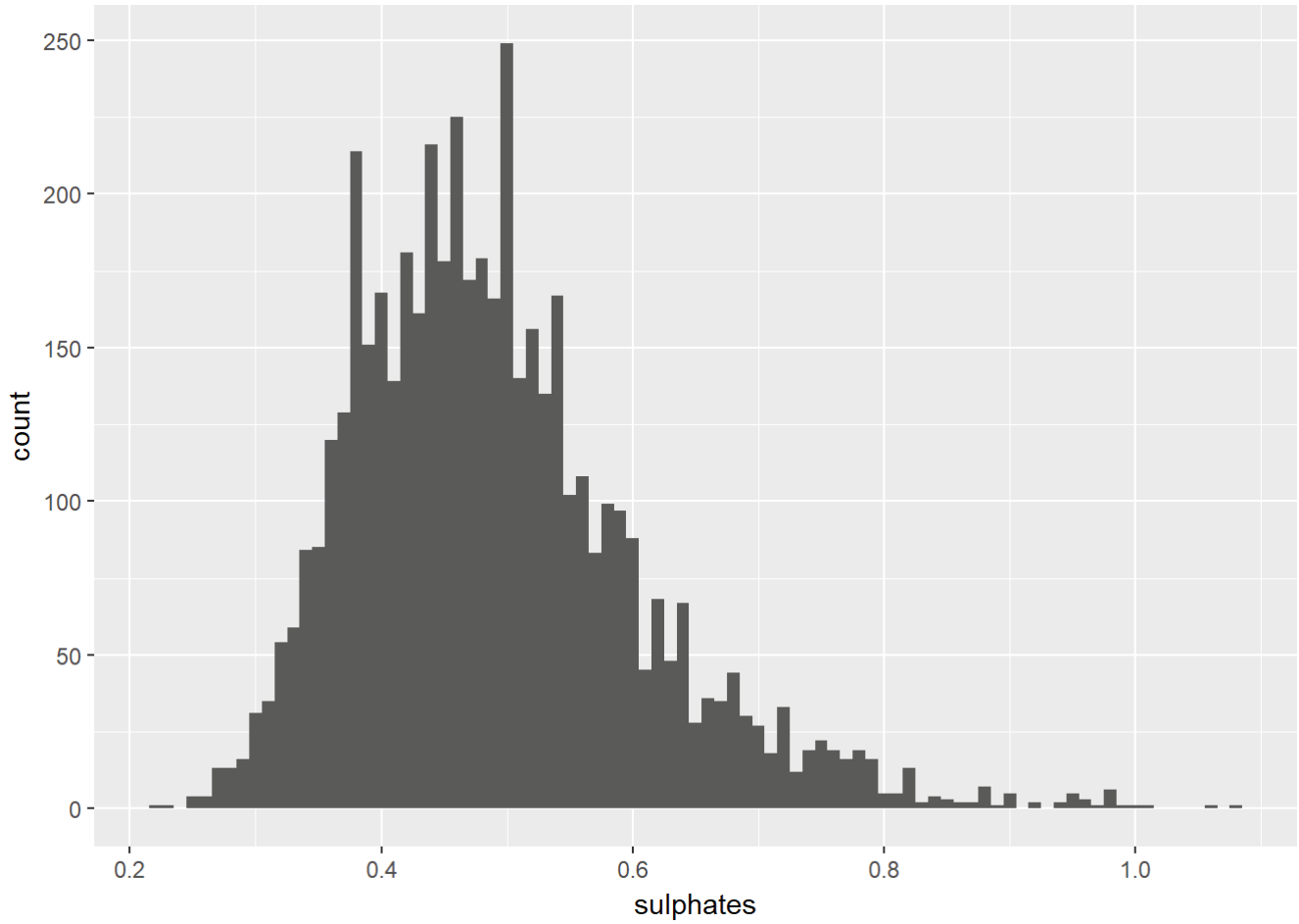


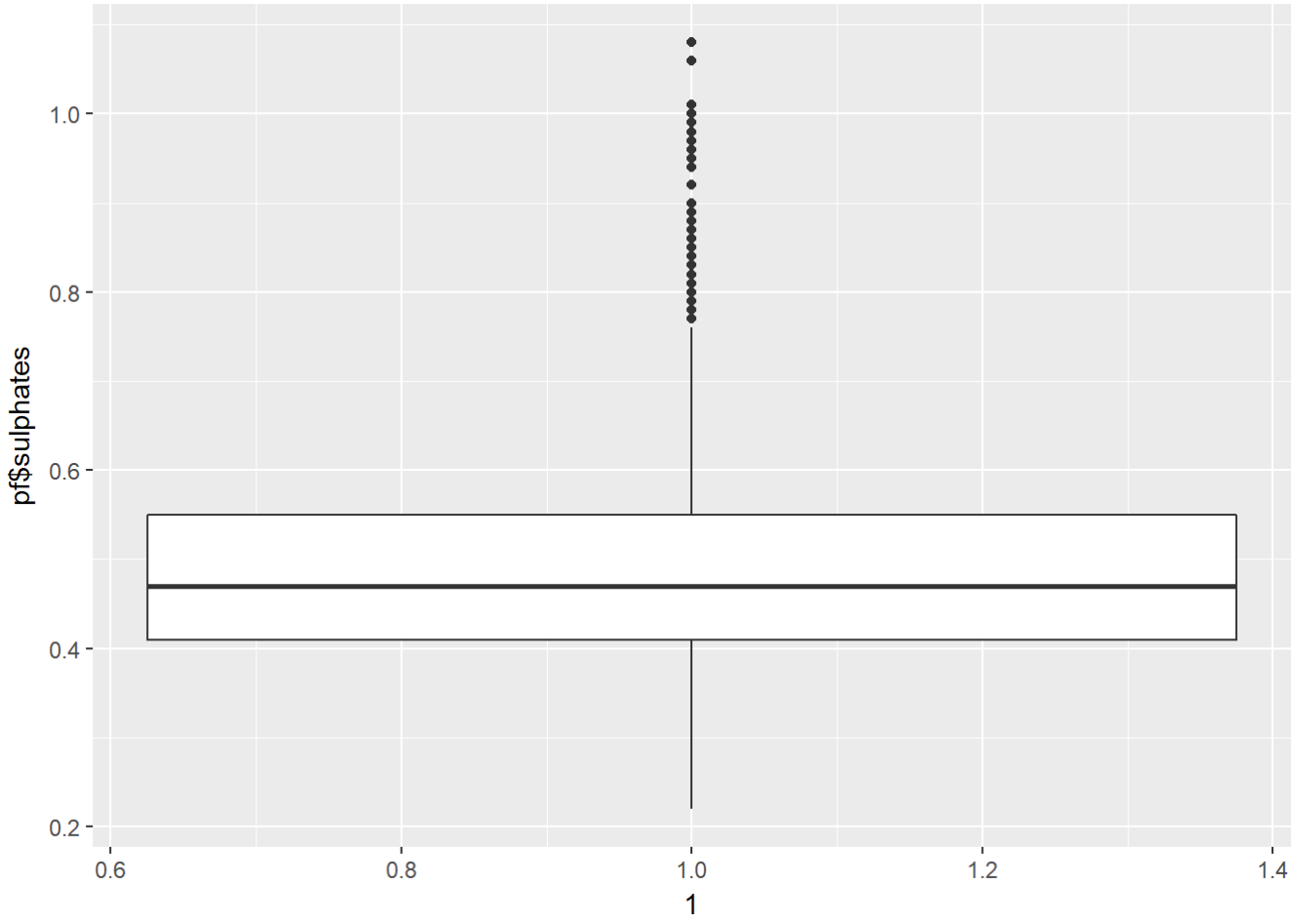
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9917  0.9937  0.9940  0.9961  1.0390
```





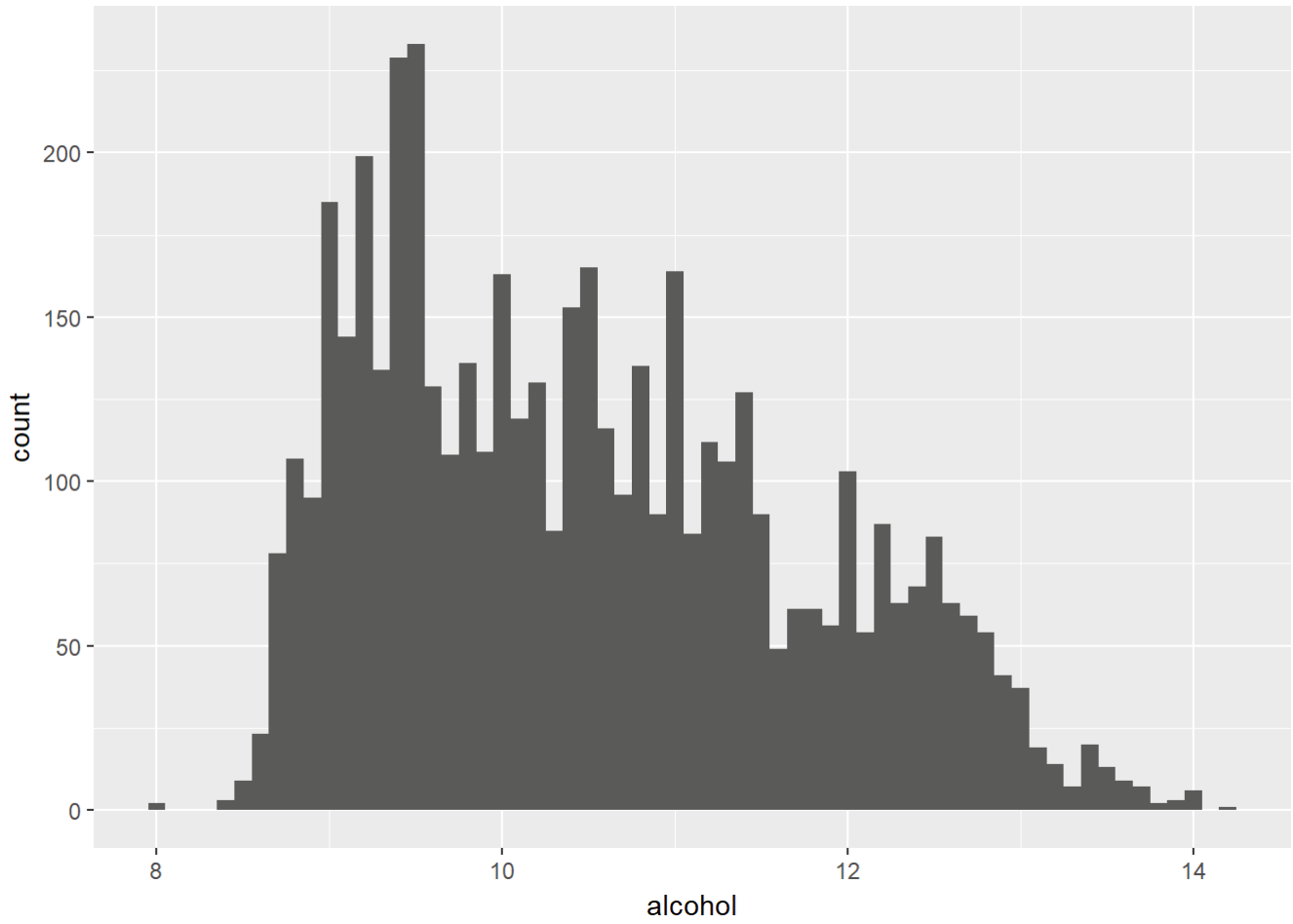
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.720   3.090   3.180   3.188   3.280   3.820
```

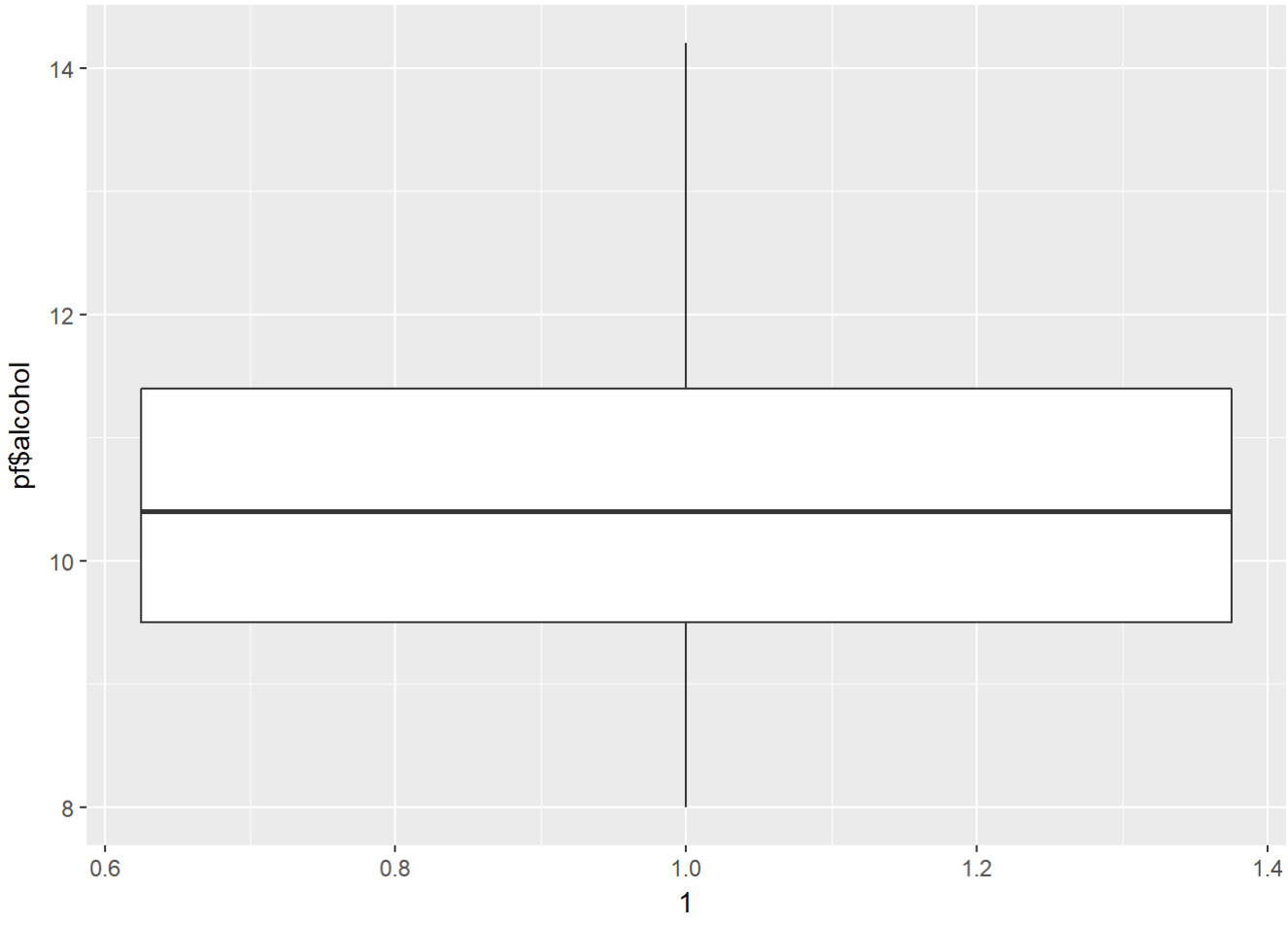




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2200  0.4100  0.4700  0.4898  0.5500  1.0800
```

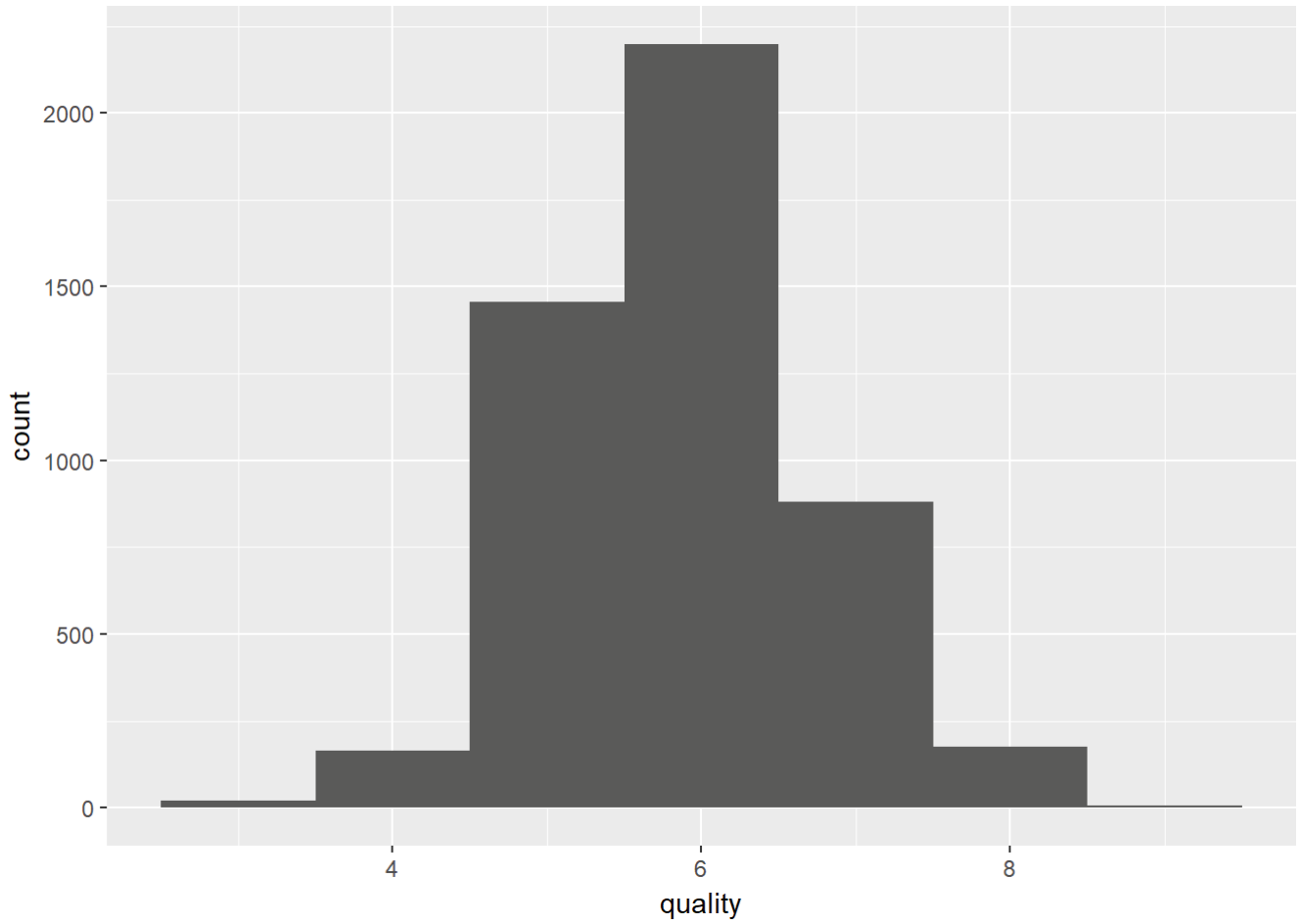
Density, pH and sulphates all shows a normal distribution with a positive skew

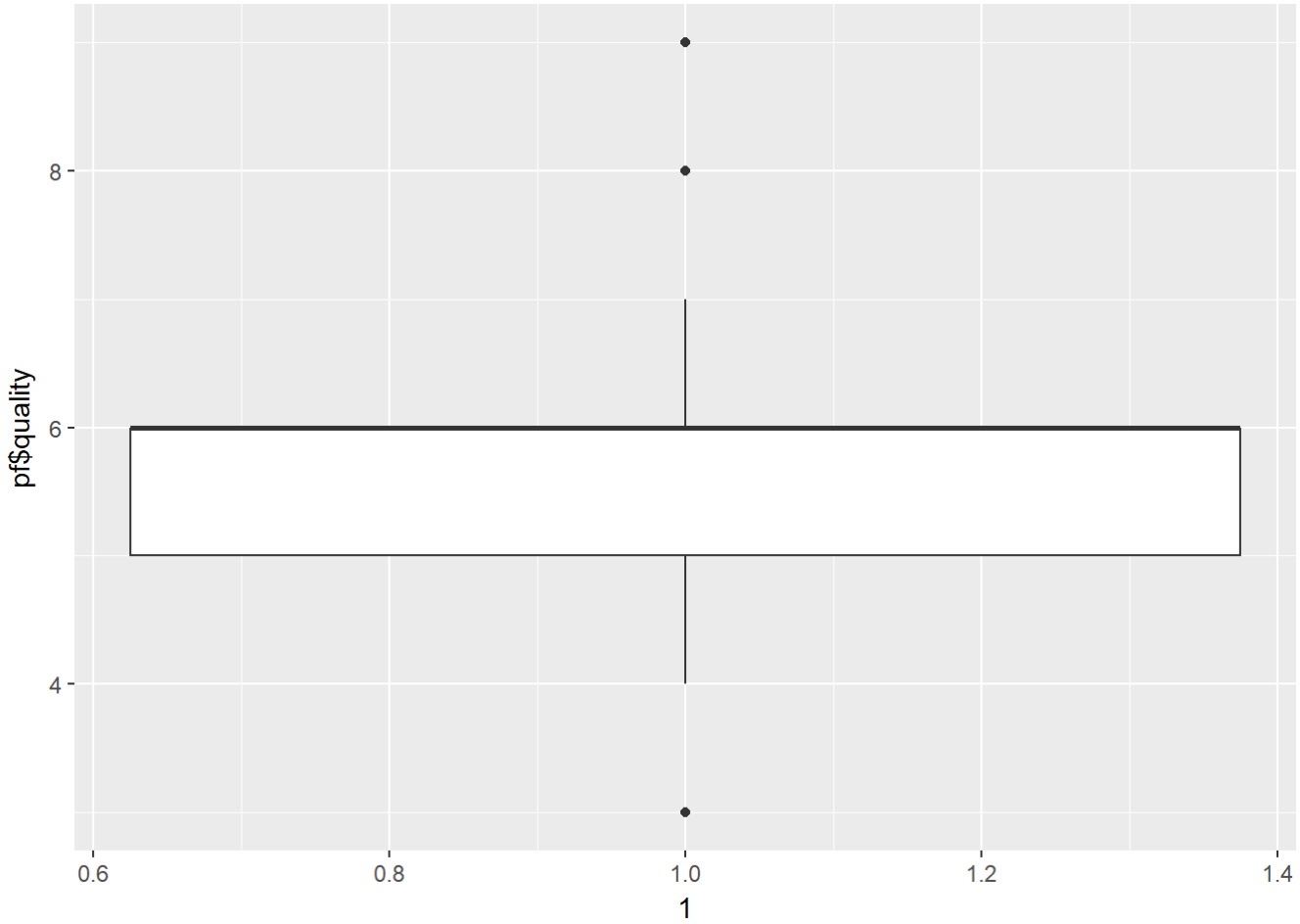





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.50   10.40   10.51   11.40   14.20
```

The alcohol is normally distributed with a very positive skewed distribution.





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

The quality is normally distributed with a negative skew

Univariate Analysis

What is the structure of your dataset?

The dataset contains information about the quality of white wine. It has 4898 observations with 12 different variables.

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

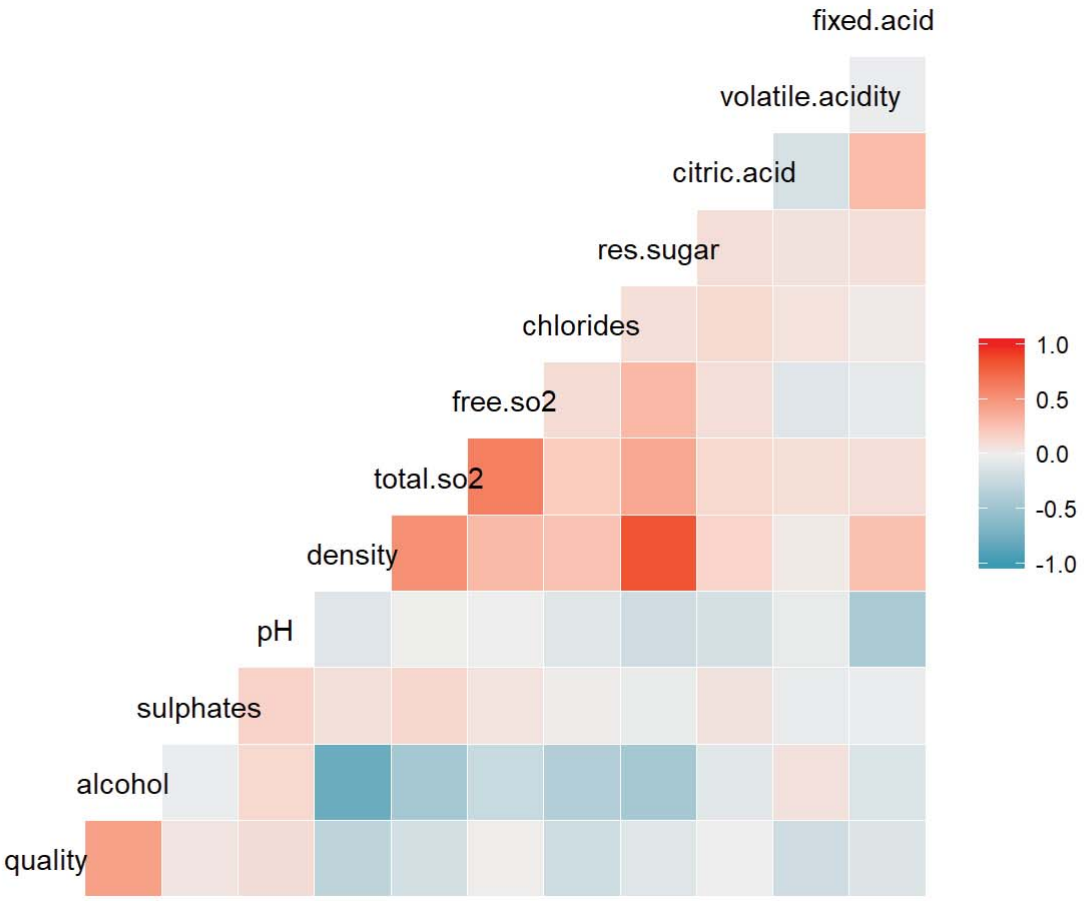
The main features that will be investigated is how the different variables affect the quality, e.g. acidity, sugar, chlorides, sulphates and alcohol level.

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

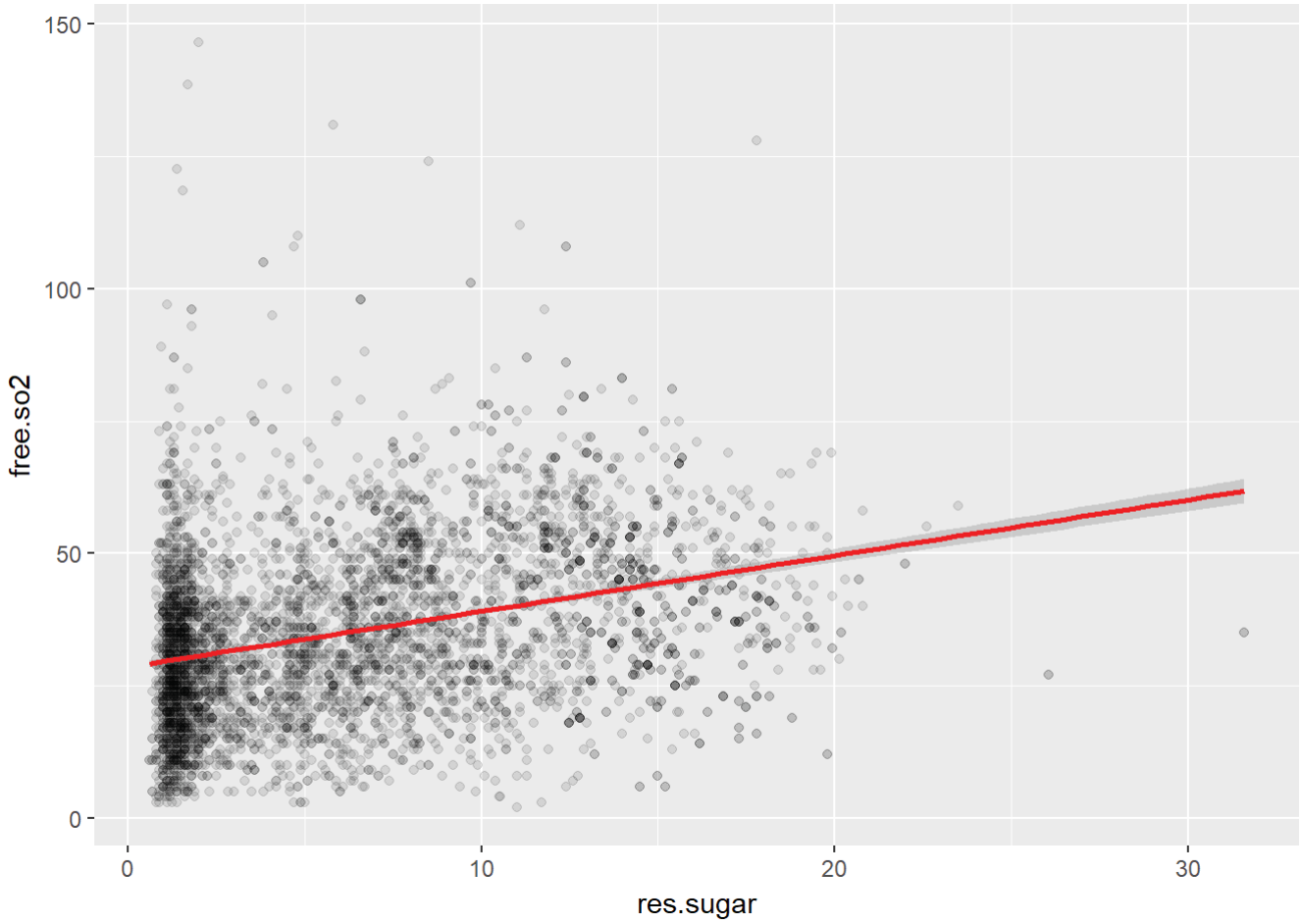
Most of the data is found to have a normal distribution with a positive skew. For several of the variables it seems that they include some outliers. These could be removed if deemed necessary for this investigation.

Bivariate Plots Section

An overview of all the variables and their correlations are created:

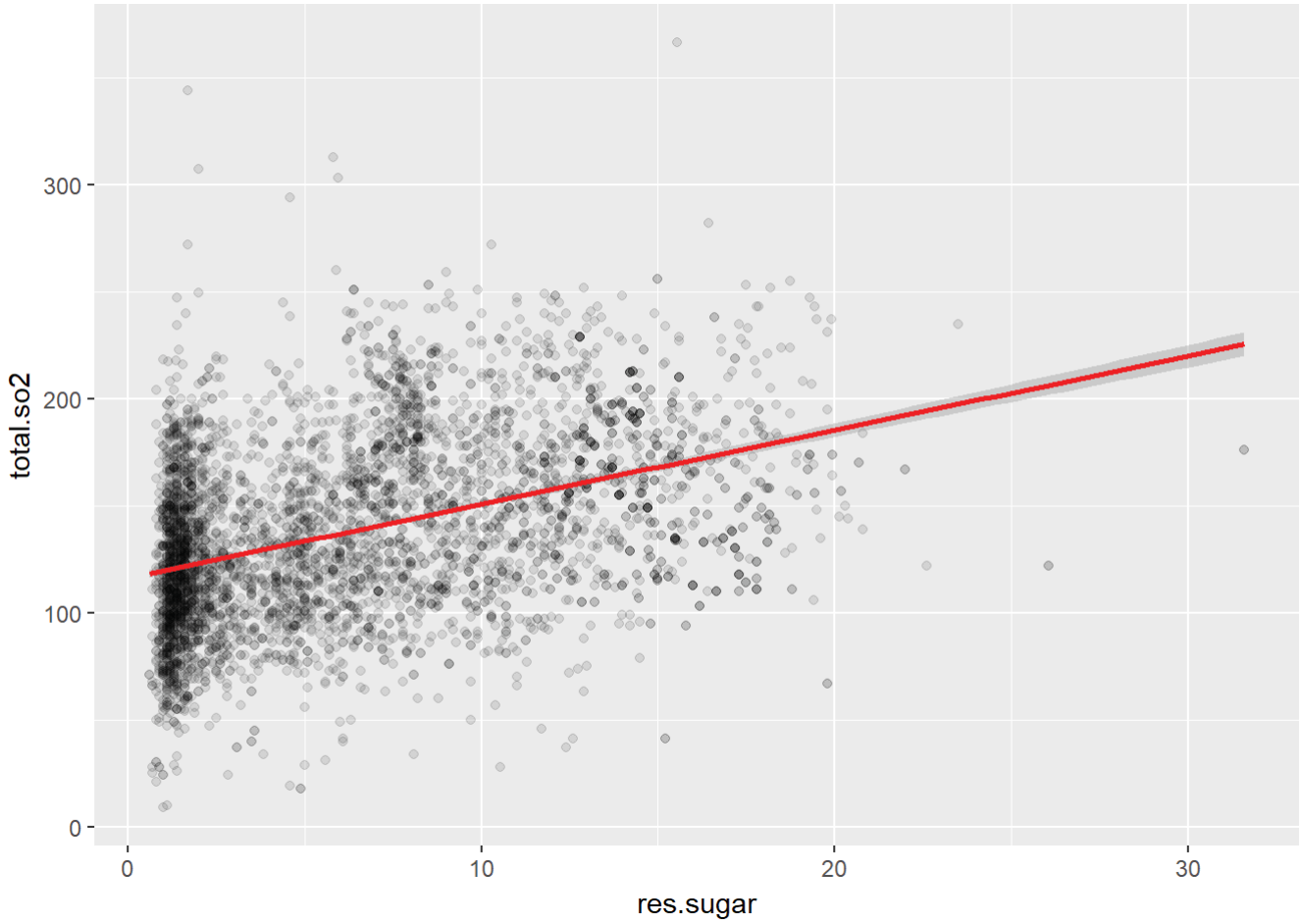


Correlations that seems significant from this overview will be investigated further below. The residual sugar seems to be correlated to many parameters and will be investigated first.
From previous analysis above, it was evident that residual sugar has one big outlier at 65.8, which will be removed from here on. Furthermore, free sulfur dioxide has one outlier at 289, which also will be removed.



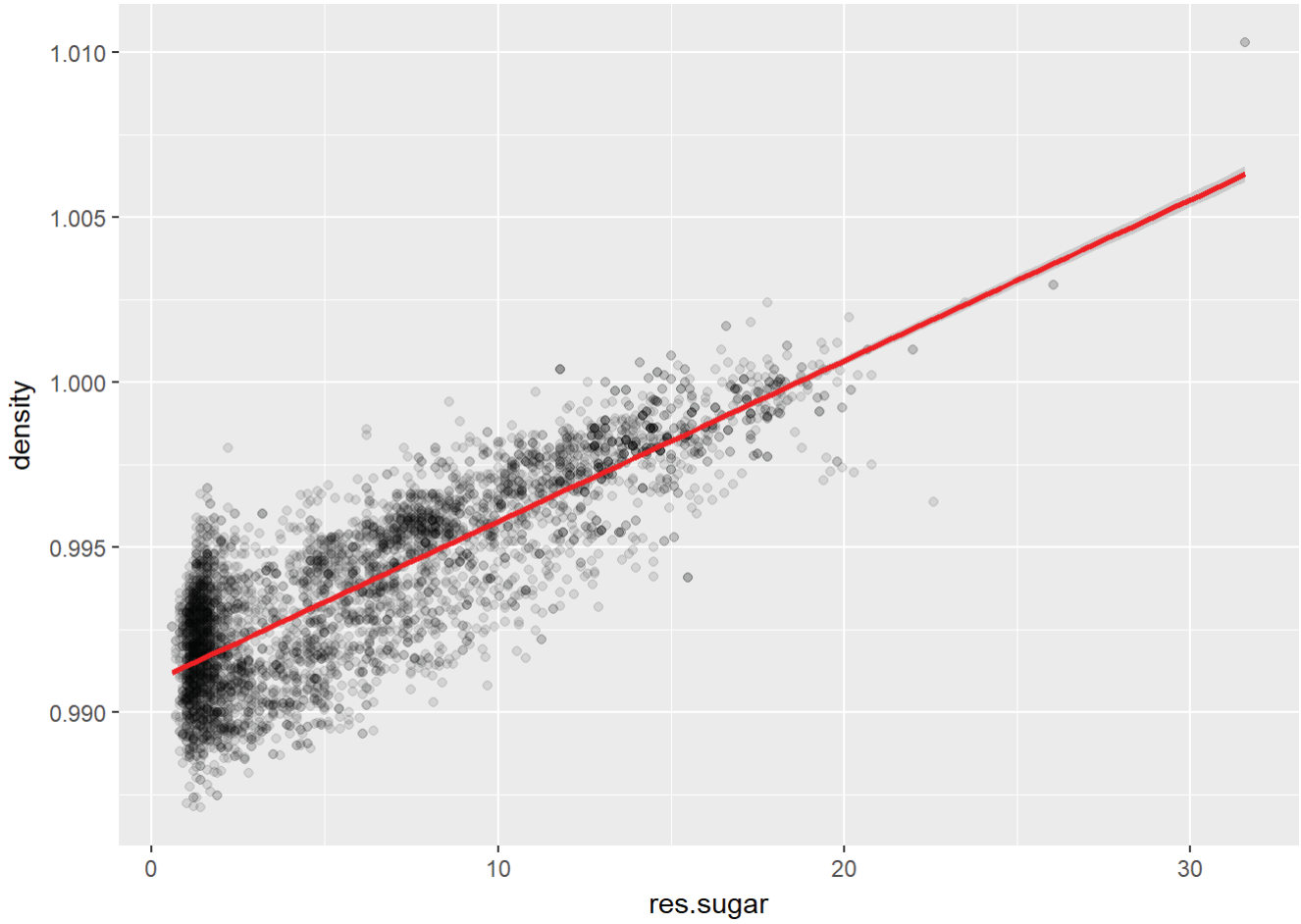

```
## [1] "Pearsons Correlation Coefficient: 0.316778"
```

An alpha value of 0.1 has been chosen to show the distribution of the data points. The free sulfur dioxide has a positive correlation of 0.32 with the residual sugar. There is a lot of wines with a very low amount of residual sugar.



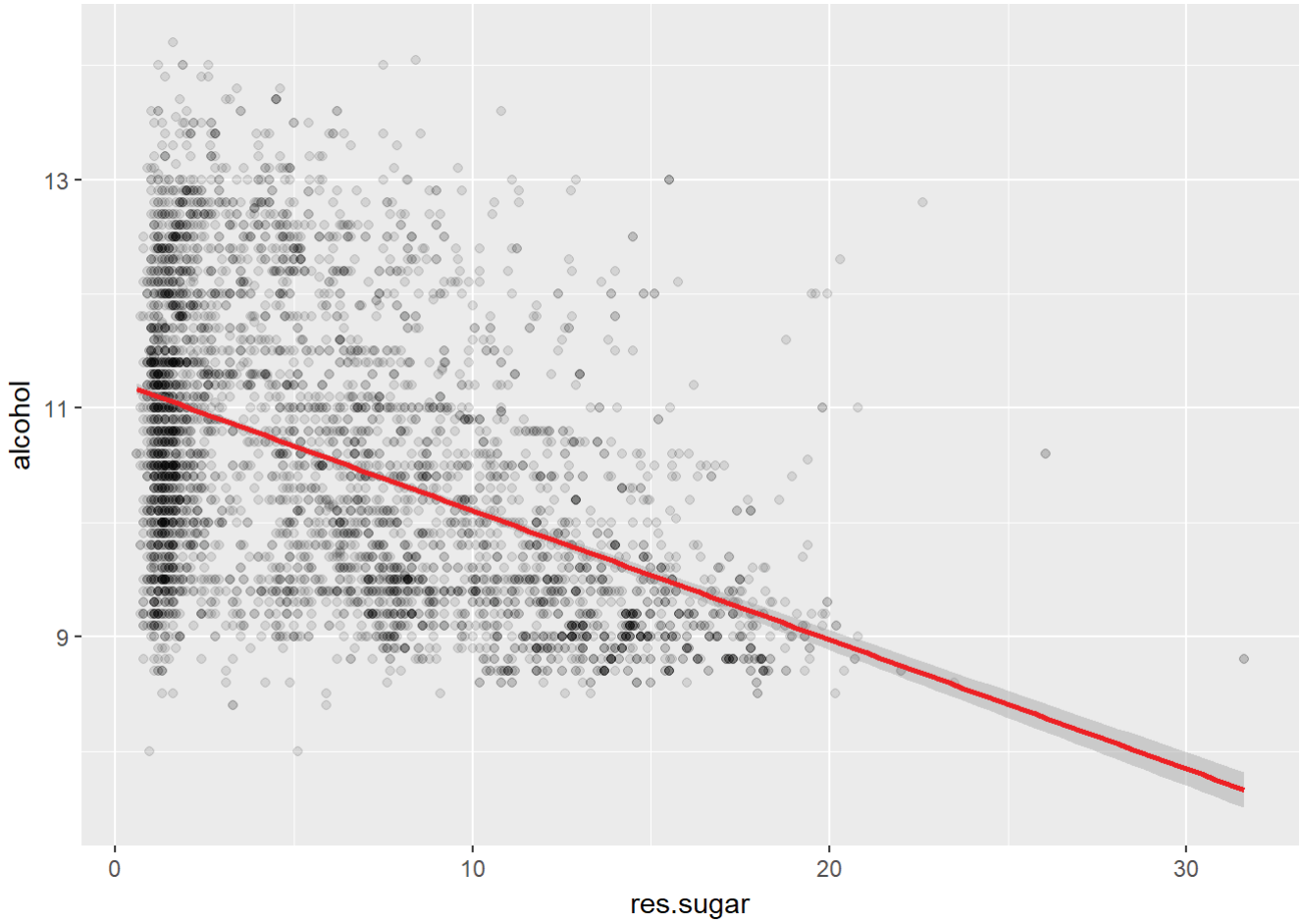

```
## [1] "Pearsons Correlation Coefficient: 0.409099"
```

The free and total sulfur dioxide behaves similarly when plotted against the residual sugar. For the total sulfur dioxide there is a positive correlation of 0.41 with the residual sugar.




```
## [1] "Pearsons Correlation Coefficient: 0.833975"
```

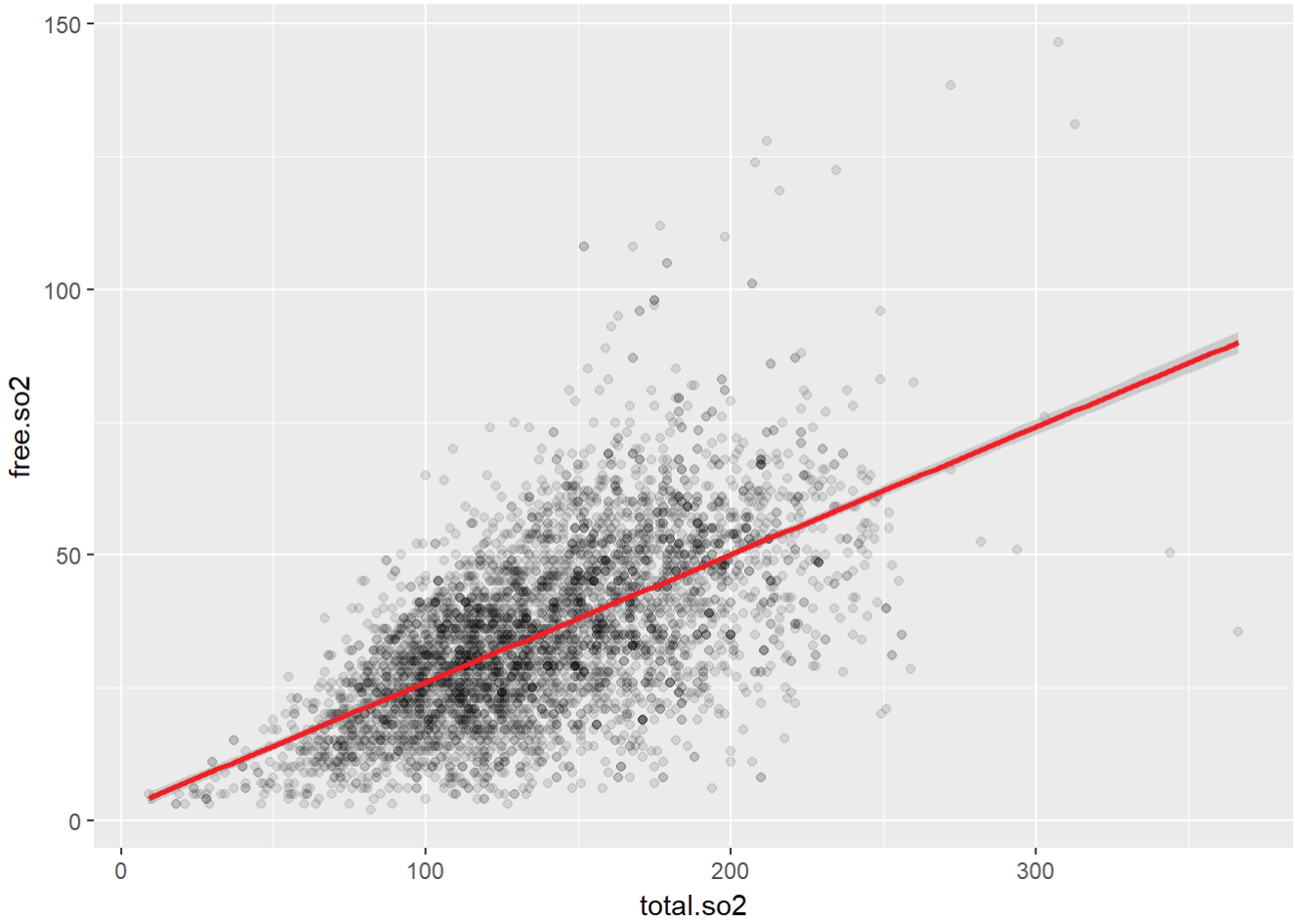
The correlation between density and residual sugar is positive at 0.83, which is a very high value and indicates a strong correlation. This is also evident from the plot. The datapoints for a low amount of residual sugar is less spread out for the density when compared to the sulfur dioxide.




```
## [1] "Pearsons Correlation Coefficient: -0.459487"
```

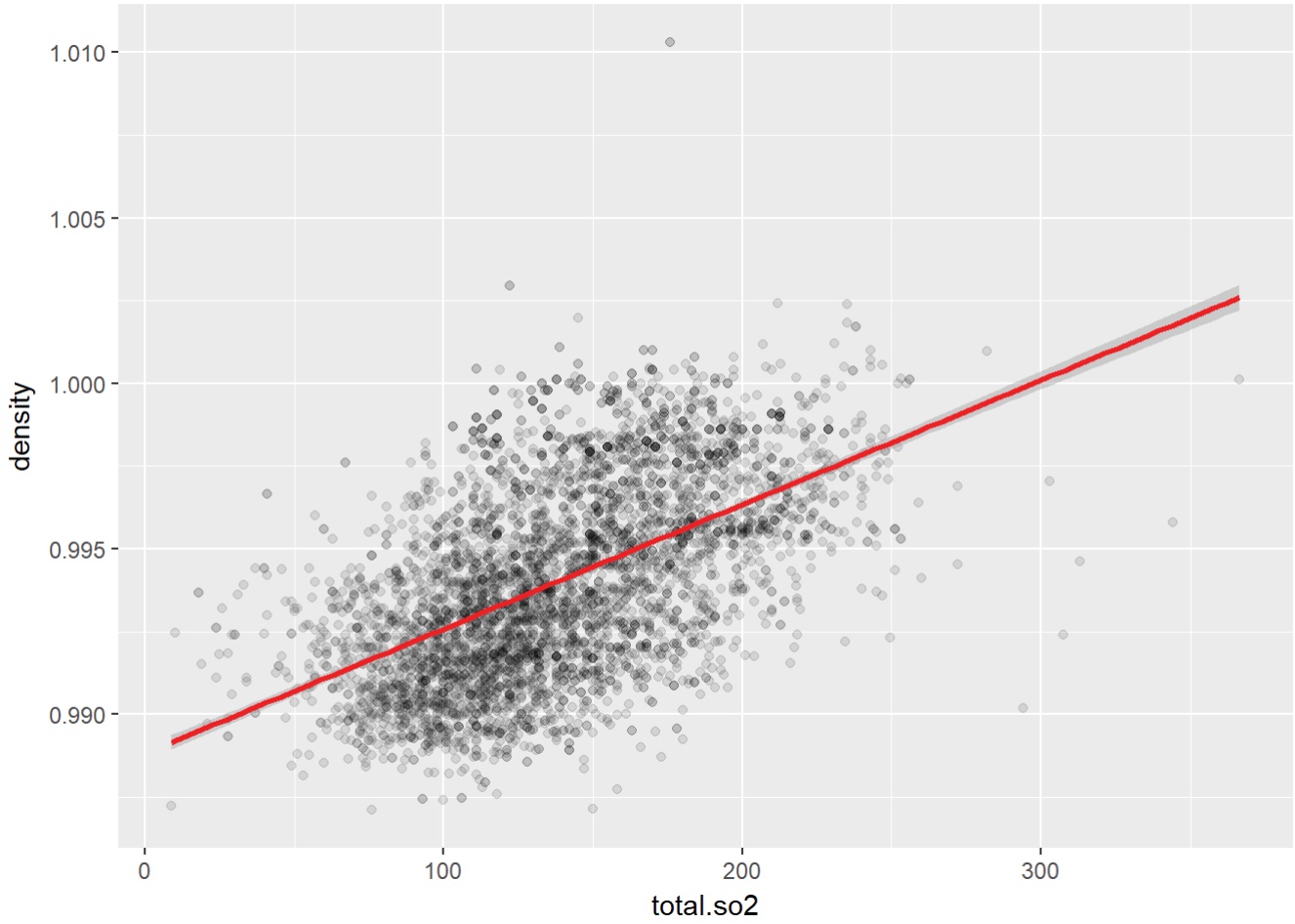
There is a negative correlation of -0.46 between alcohol and residual sugar. There is a clear declining trend to the upper value of alcohol with increasing residual sugar.

All the plots above here, investigates the residual sugar and the significant correlations found. The two first plots are the free and the total sulfur dioxide. Persons correlation yields 0.32 and 0.41 for these two plots, which could indicate a moderate correlation between these and residual sugar. Both have a lot of datapoints with very little residual sugar, but otherwise there seems to be a positive correlation.
The residual sugar and density seems to be well correlated, with a positive Pearson correlation coefficient of 0.83.
Residual sugar and alcohol also seems to show some correlation with a negative Pearson correlation coefficient of -0.46. All of these datasets shows that a lot of wines have very little residual sugar, but a correlation seems to be prevalent for an increasing amount of residual sugar.



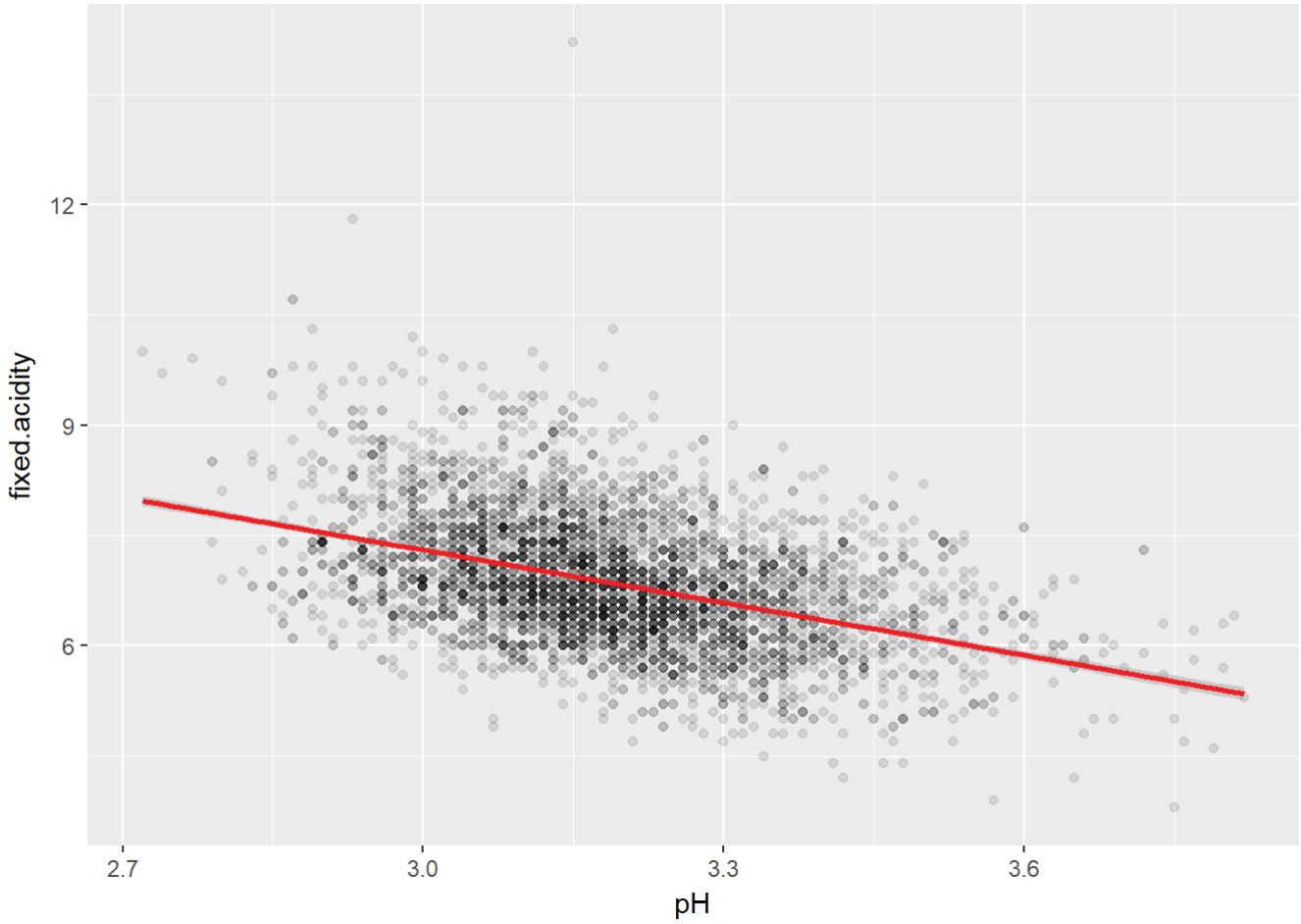

```
## [1] "Pearsons Correlation Coefficient: 0.611357"
```

There is a moderate to strong relationship between the free sulfur dioxide and the total sulfur dioxide with a positive correlation coefficient of 0.61.



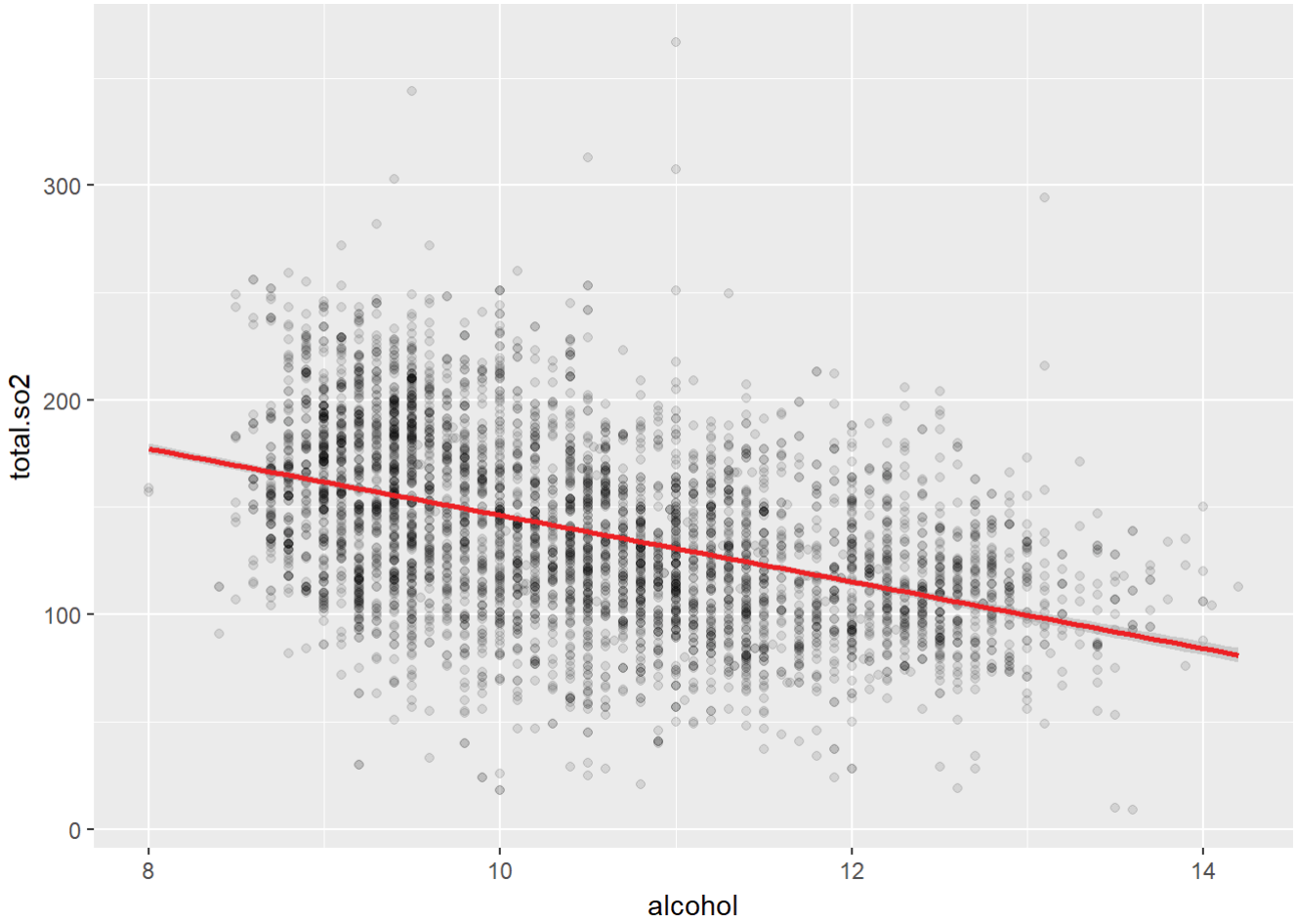

```
## [1] "Pearsons Correlation Coefficient: 0.544207"
```

The total sulfur dioxide seems to be correlated with the density. The Pearsons correlation coefficient is 0.54, which indicates a moderate to good positive correlation.



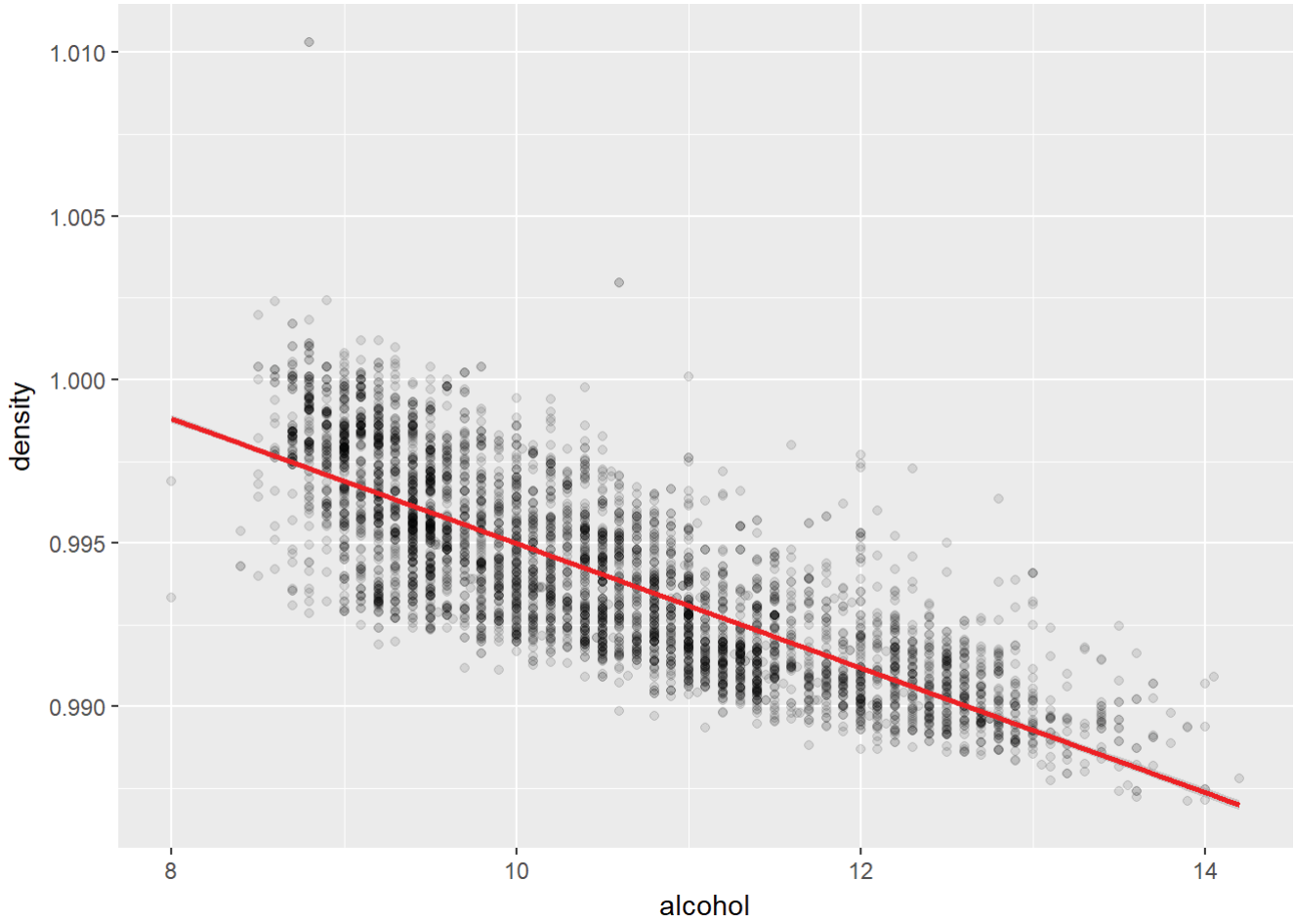

```
## [1] "Pearsons Correlation Coefficient: -0.426147"
```

pH and fixed acidity seems to be negatively correlated with a Pearsons correlation coefficient of -0.43. It makes sense to find a correlation between these, as a low pH is a sign of high acidic content.




```
## [1] "Pearsons Correlation Coefficient: -0.451359"
```

There is a negative relationship between alcohol and total sulfur dioxide of -0.45. There seems to be a decreasing scatter of total sulfur dioxide with increasing alcohol content.

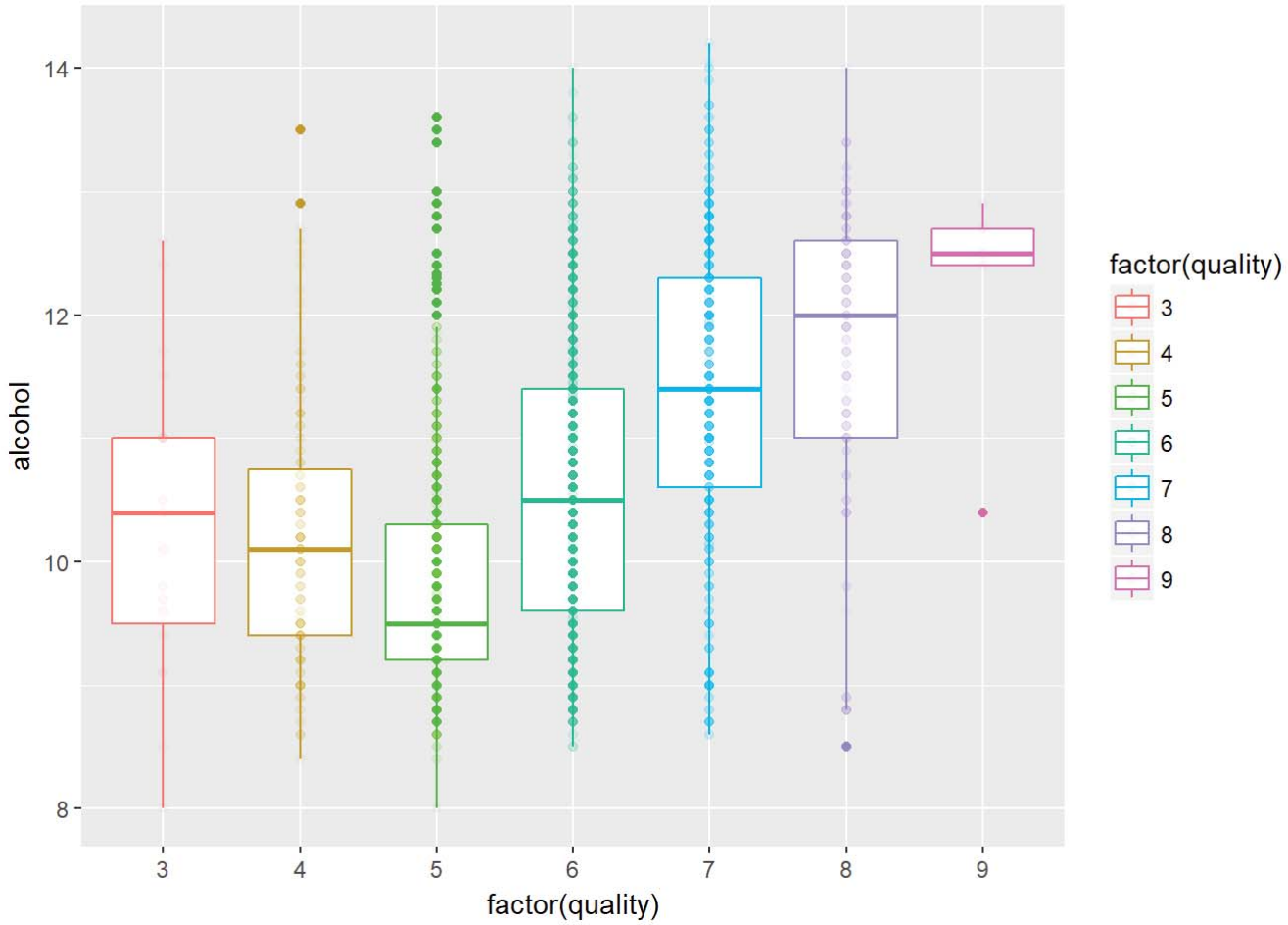



```
## [1] "Pearsons Correlation Coefficient: -0.801895"
```

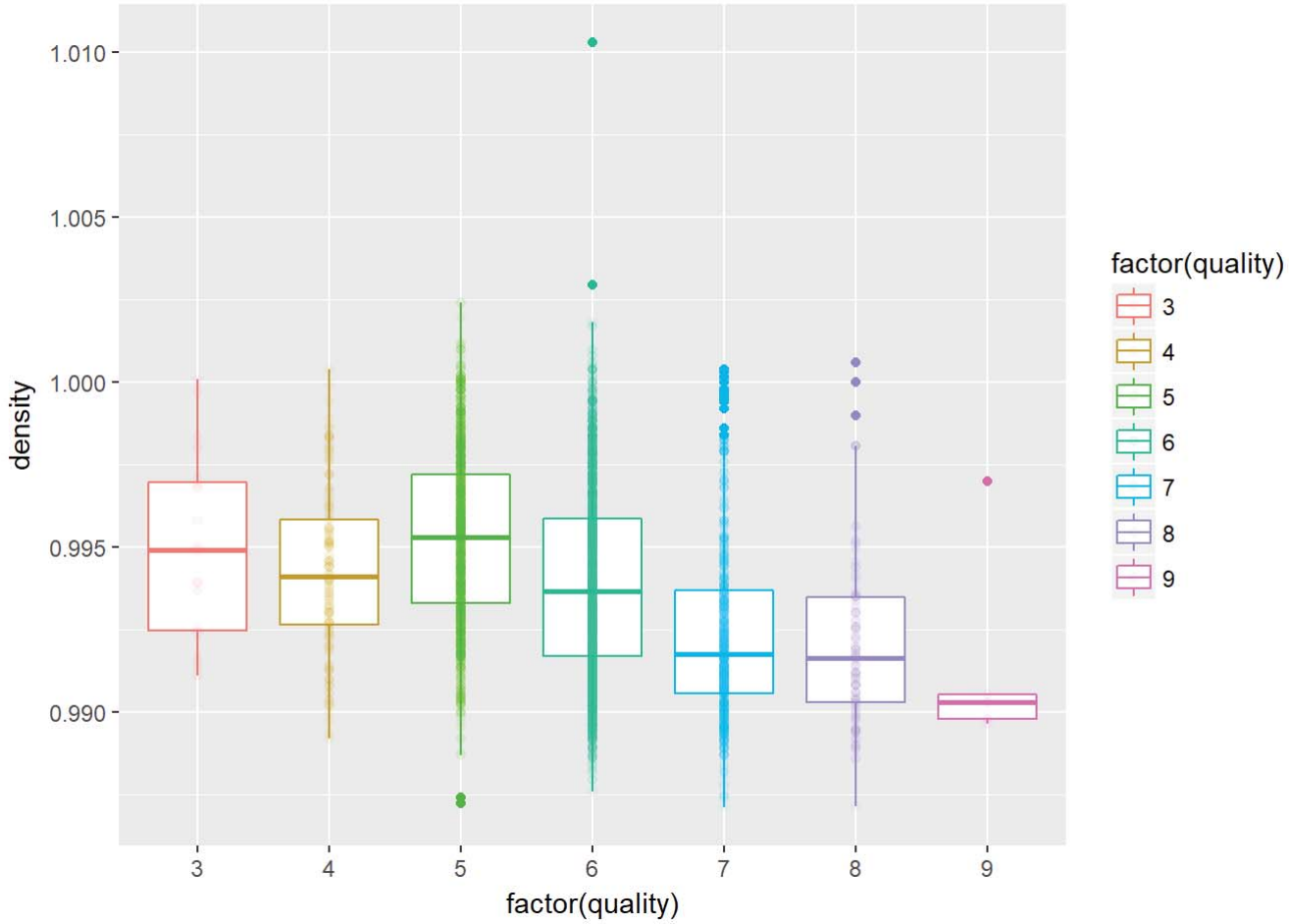
These plots investigate the correlations with alcohol. The residual sugar correlation was already investigated earlier with a Pearsons correlation coefficient of -0.46. Here the total sulfur dioxide has a negative Pearsons correlation coefficient of -0.45. Looking at the plot, there is some scatter but overall the correlation seems to be evident.

Alcohol and density seems to be well correlated with a Pearsons correlation coefficient of -0.8. This was expected as the density of alcohol is less than water, such that an increase in alcohol will lower the overall density of the wine.

From the correlation overview made in the start of this section, only alcohol seemed to have some significant correlation with quality and to some extent density as well. These will be investigated here.

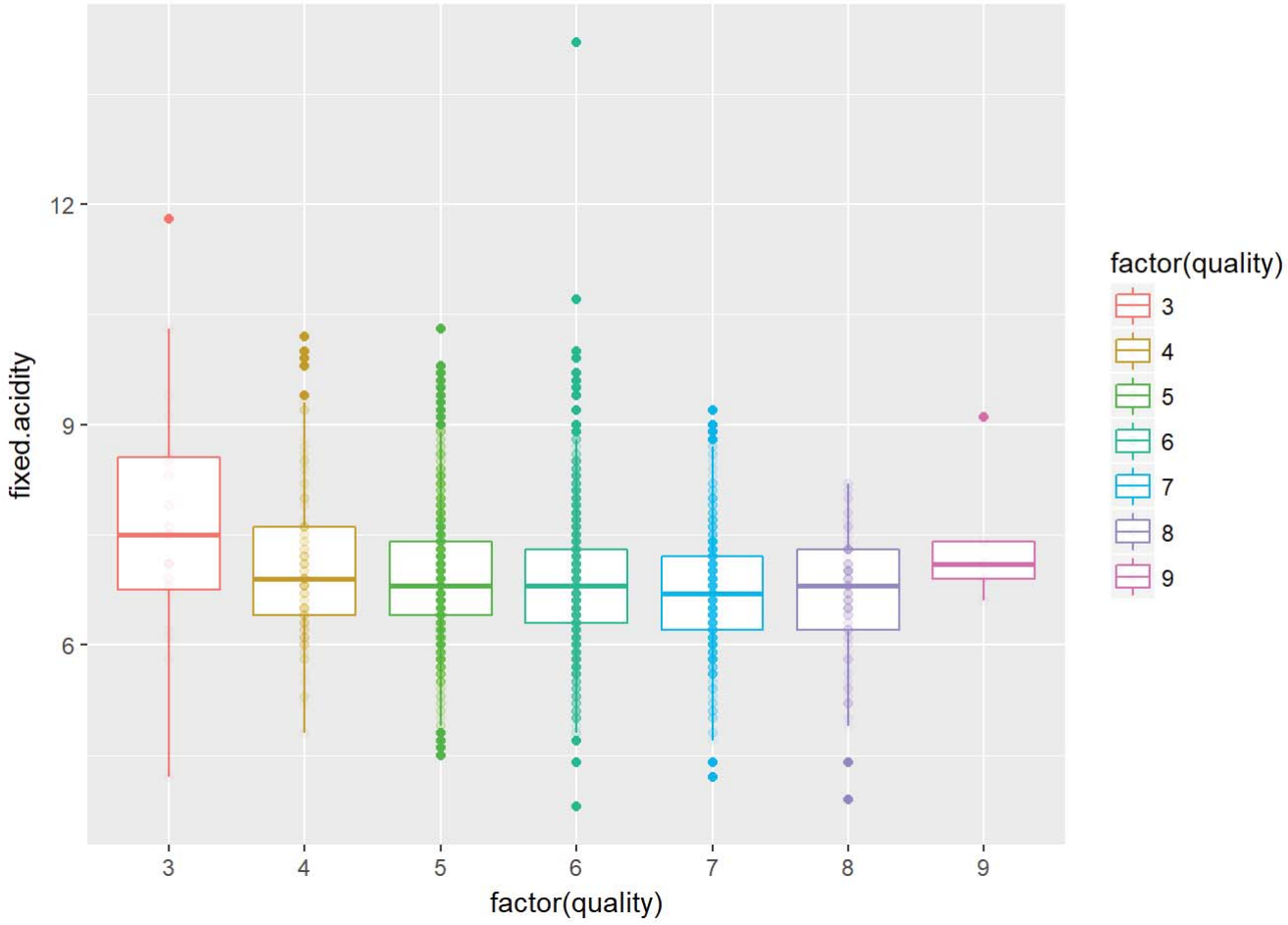


There is a tendency that the higher alcohol content wines have a higher quality. However, for the lowest quality, the alcohol content seems to be increasing a bit as well. Wine with a quality of 5 seems to be skewed.

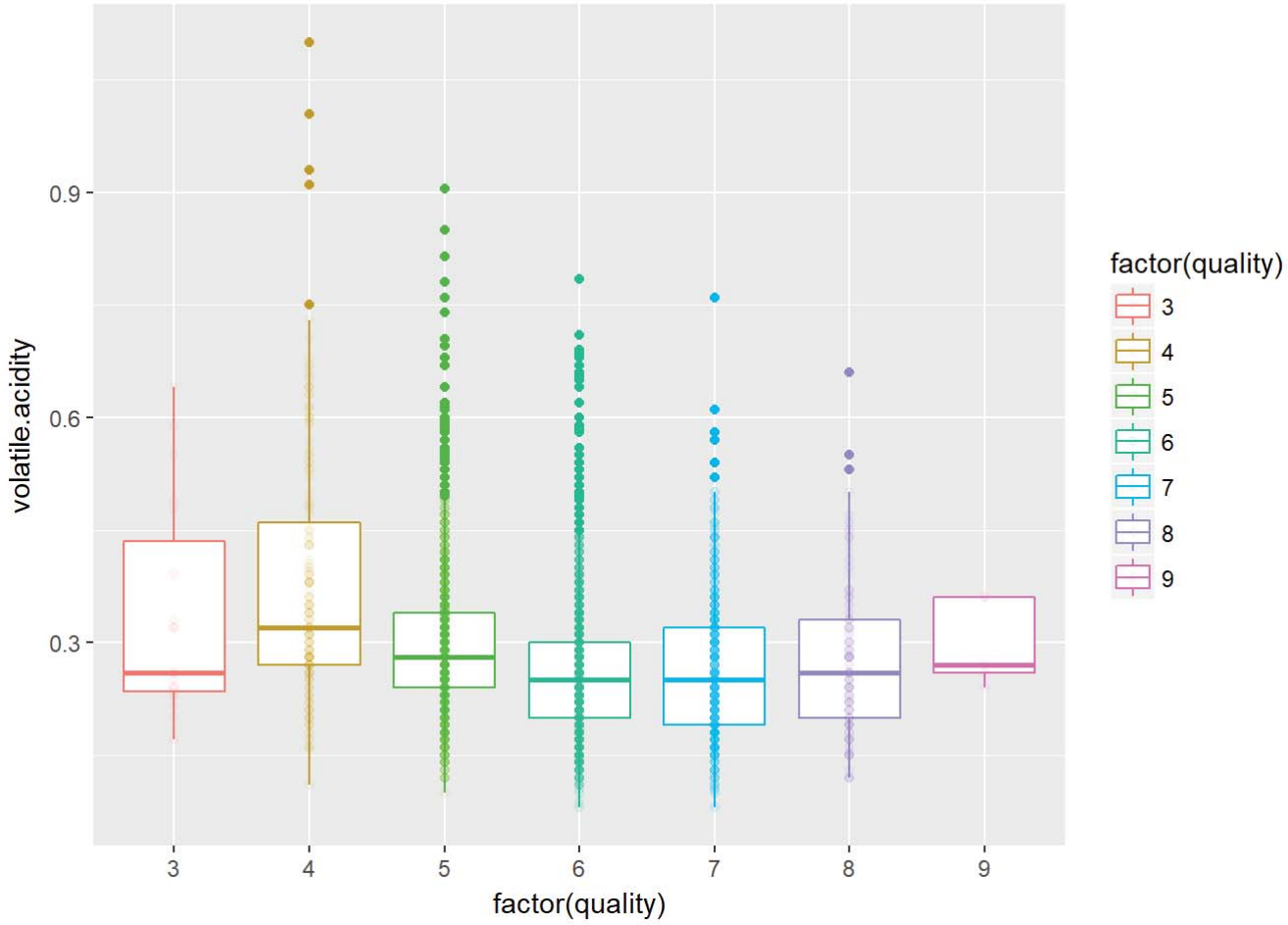


The density and quality shows a reverse trend to what was seen in the previous plot with alcohol. As alcohol and density are so closely related, this trend is given due to the physical relationship between alcohol content and density.

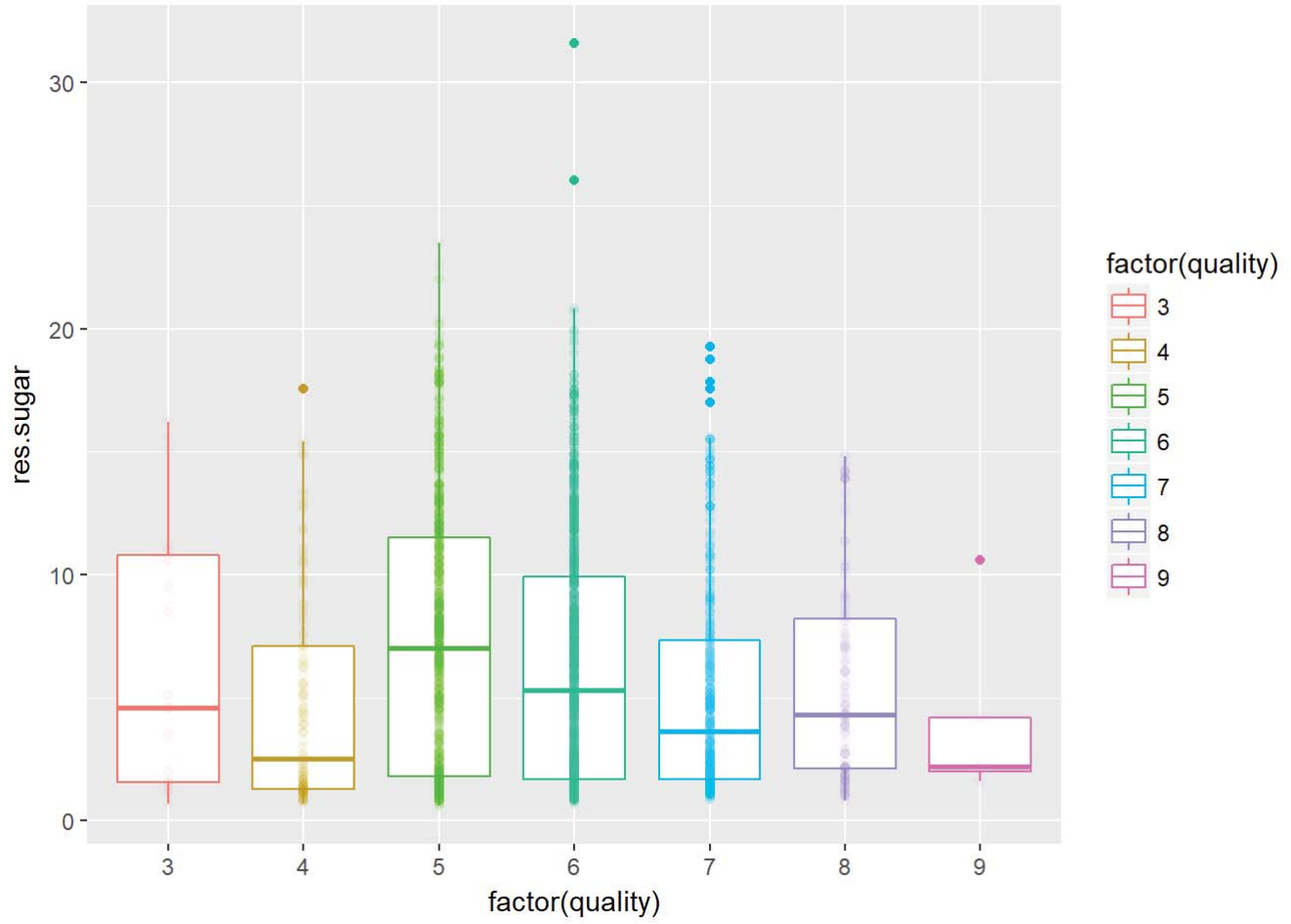
Below here it will be investigated if other variables could have a significant influence on the quality.



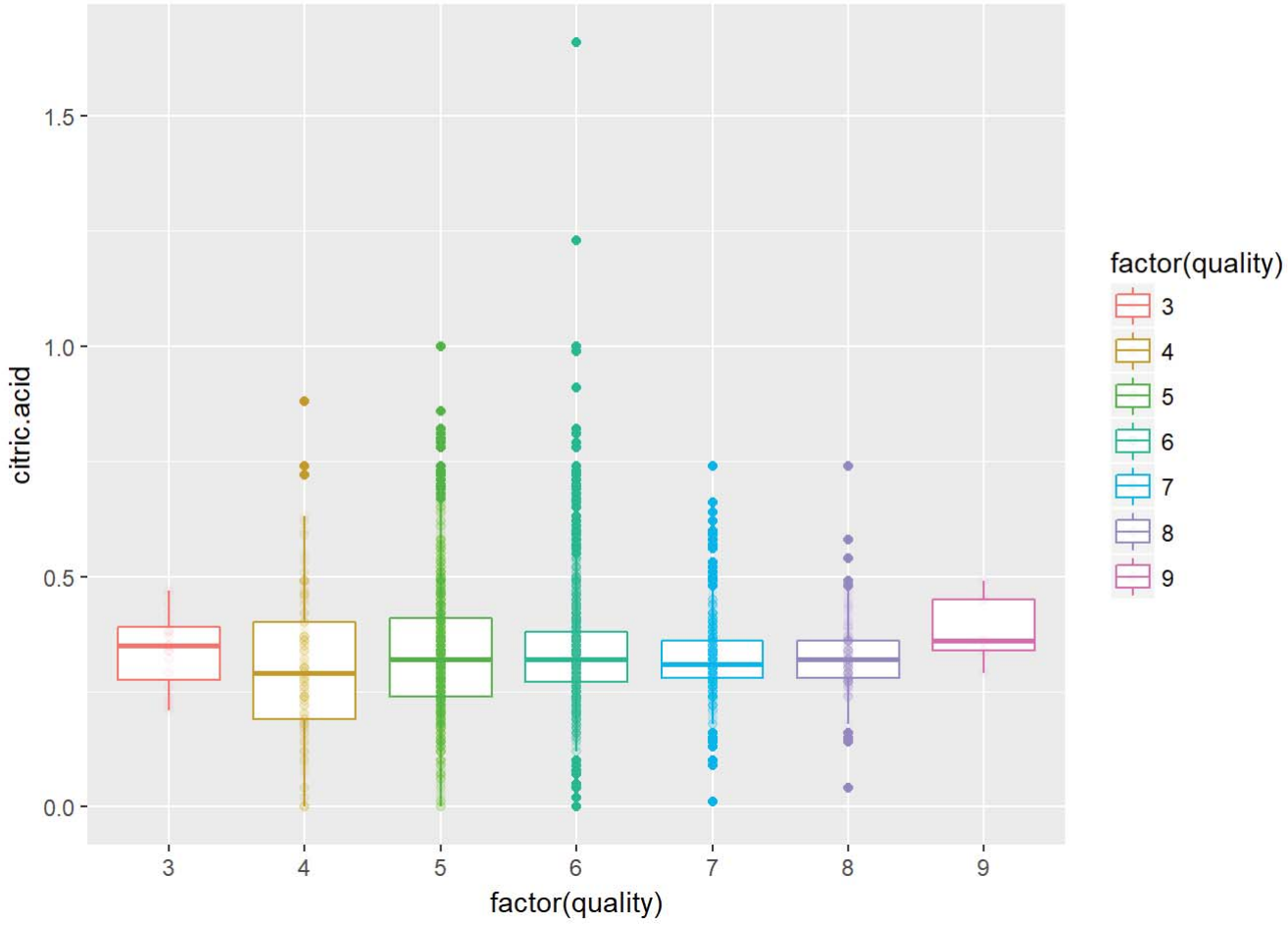
It seems that for medium to high quality the fixed acidity is close to constant. Only the lowe quality wines have an increase in fixed acidity and a large spread of data.



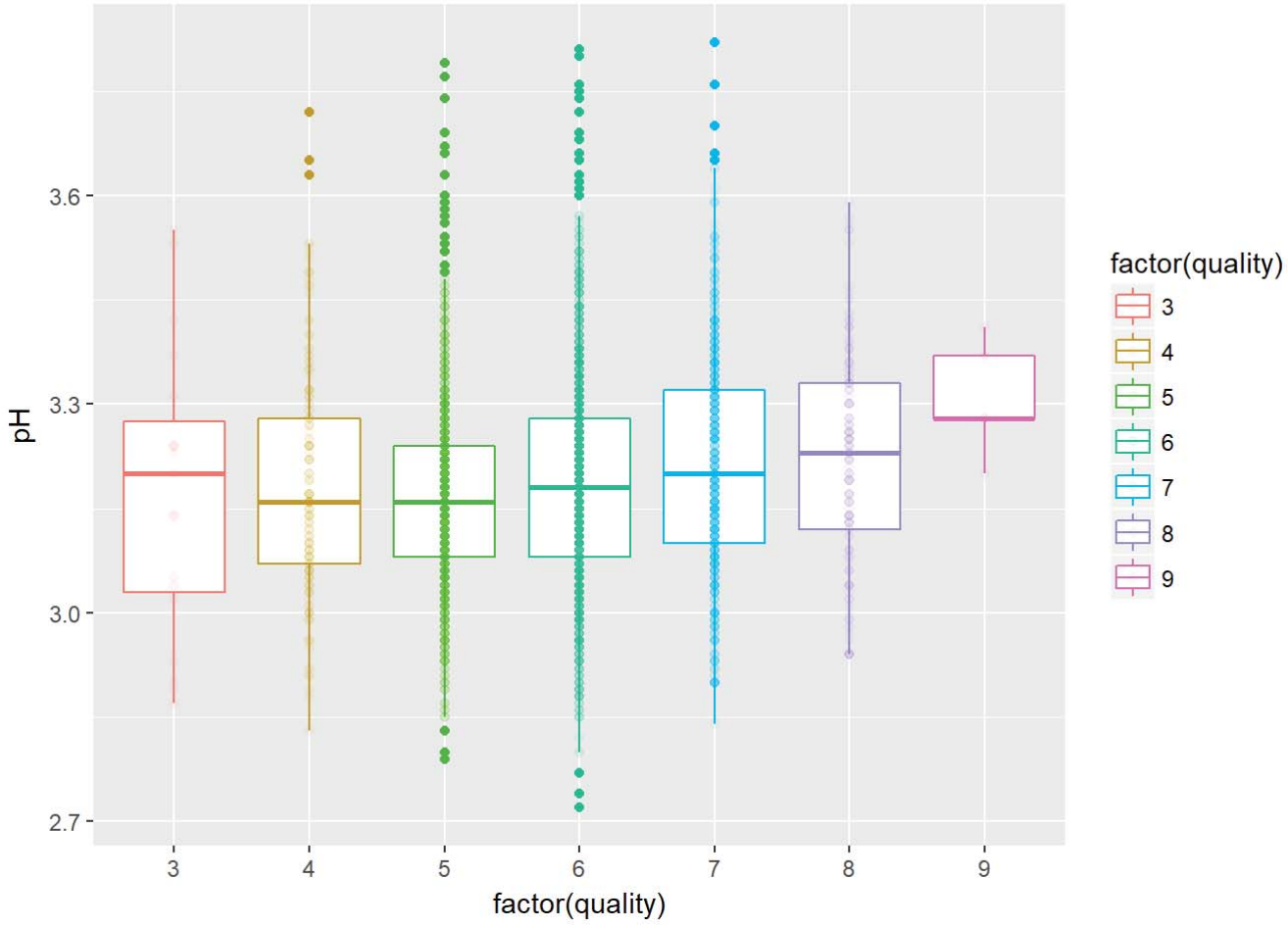
The volatile acidity is high for low and high quality wines. The lowest volatice acidity is found for the medium quality white wines.



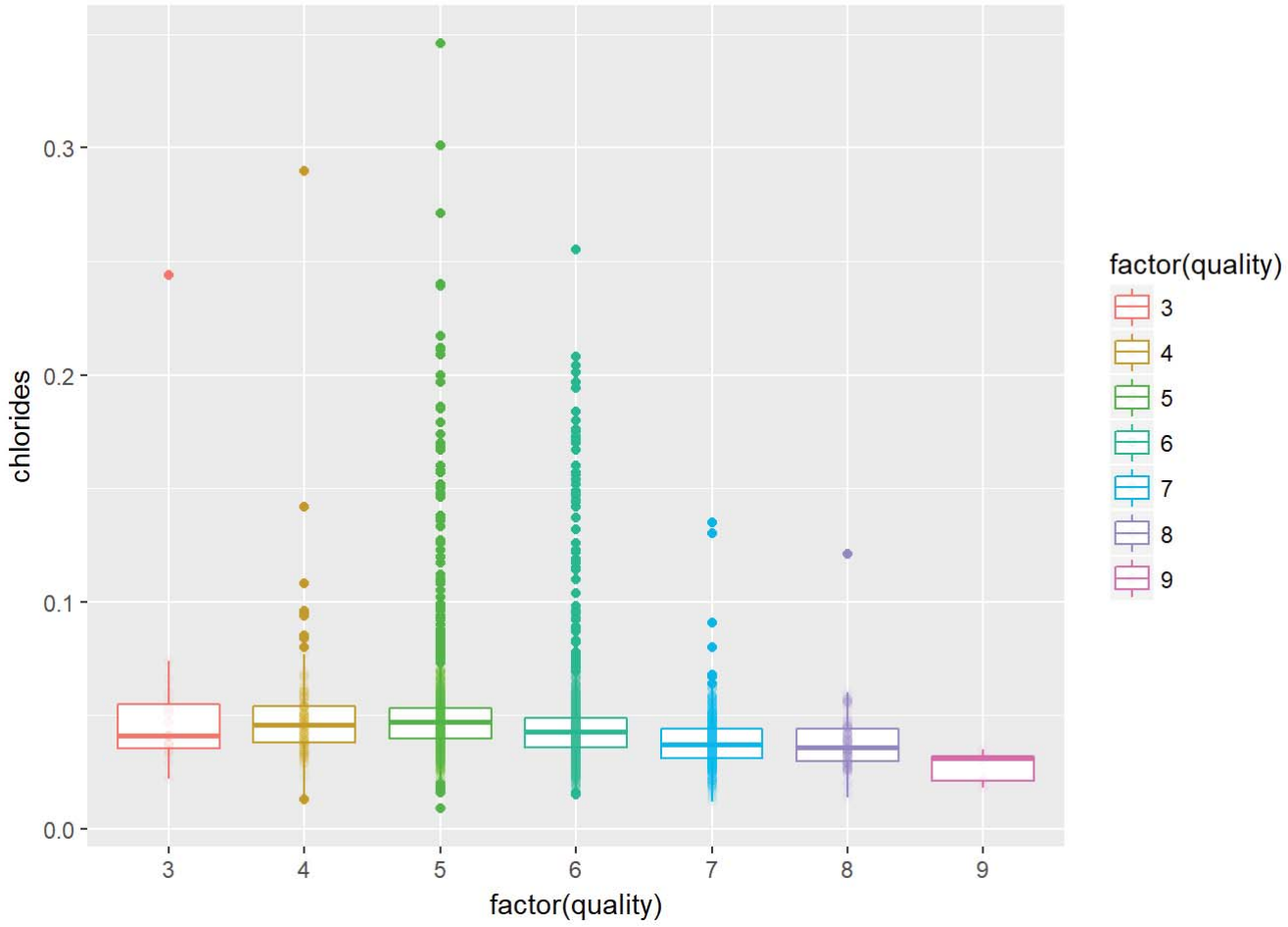
The lowest value of residual sugar seems to be rather constant for all qualities but the highest values varies with quality. Most of the datasets are skewed.



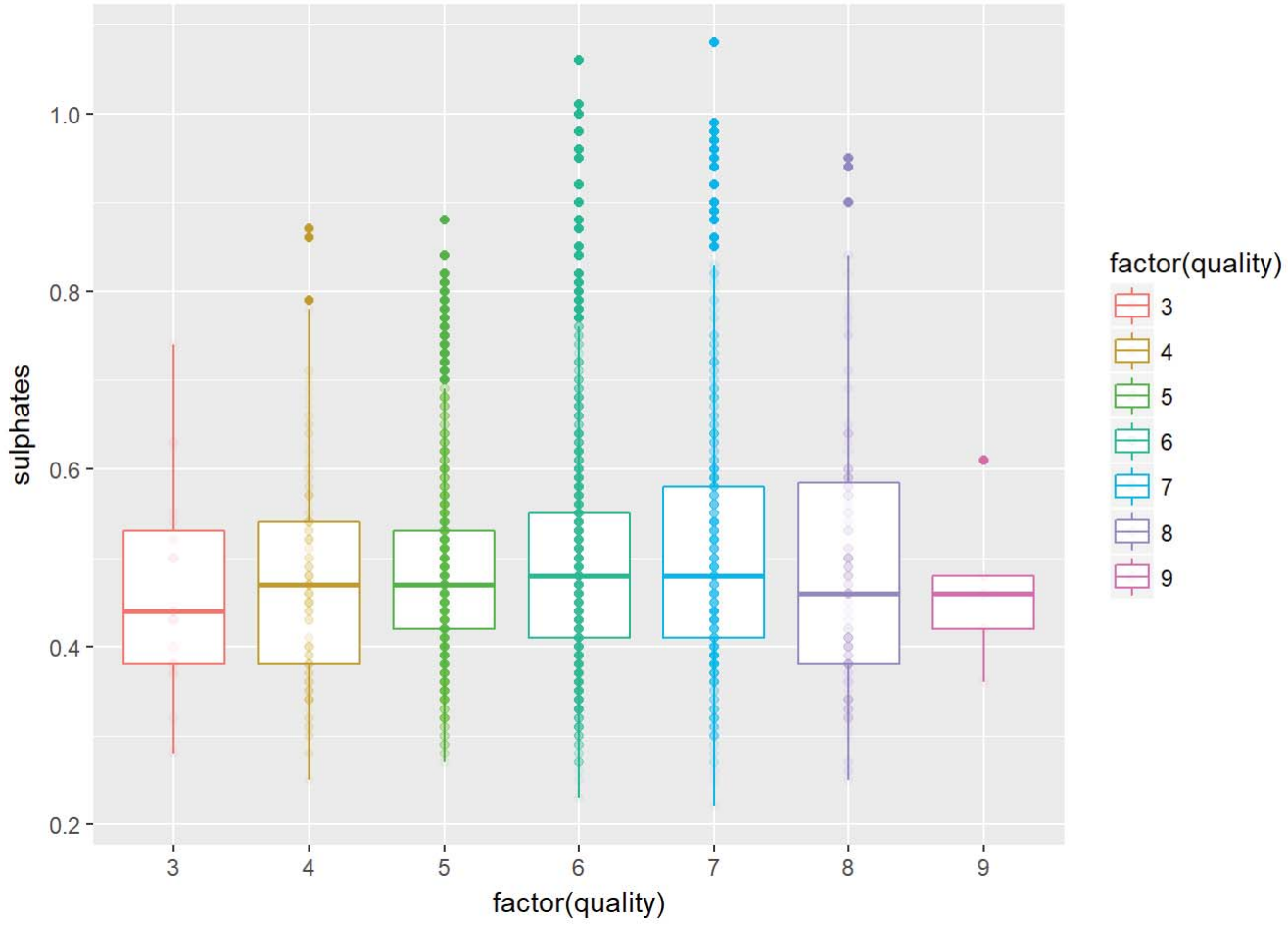
The quality does not seem to vary a lot with the citric acidic values. The highest quality wines seems to have more citric acid.



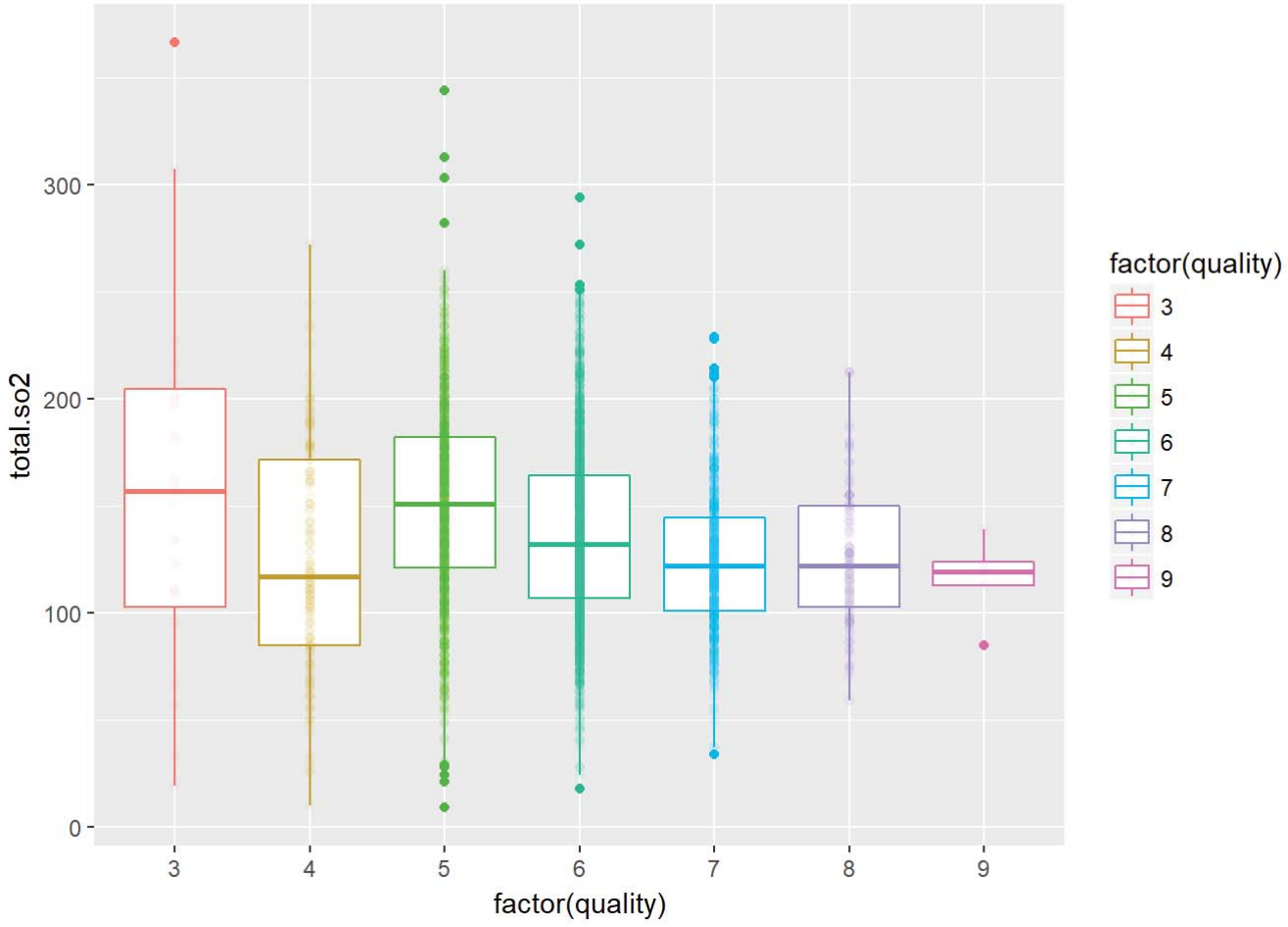
It could seem that an increasing pH could mean a higher quality wine, however the low quality wines also has a high pH. pH alone is not an indicator of quality.



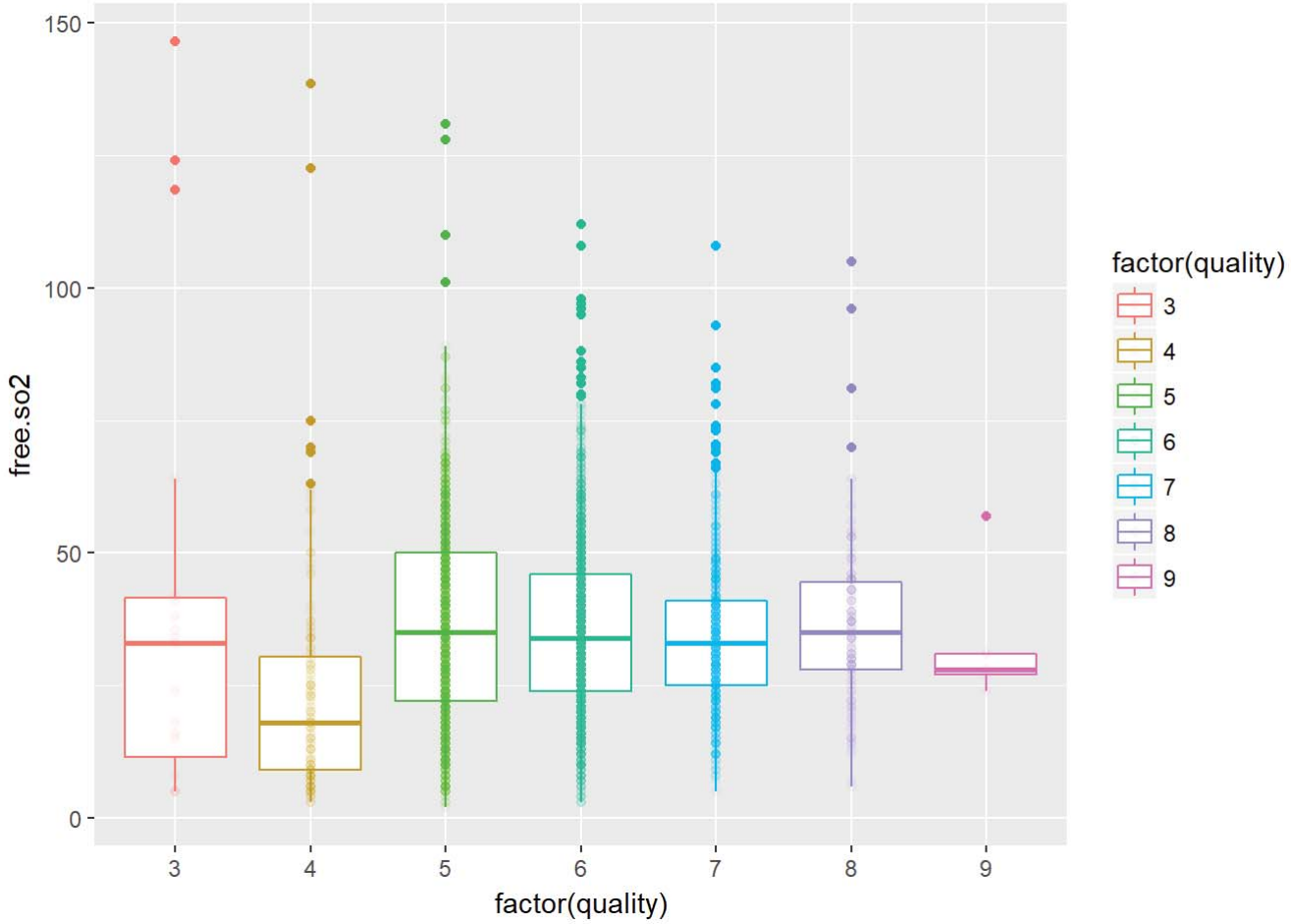
For the cholride content the quartiles in the boxplot are very narrow in range with a large range of outliers. The median trend is for high quality wines to have a low content of chlorides and the outliers seems to narrow in as well.



The median for sulphatesshows a nearly constant trend and does not seem to have a major impact on the quality of the wine.



There seems to be a trend of decreasing total sulfur dioxide for higher quality wines.



Both low and high quality wines have a low amount of free sulfur dioxide. This makes it difficult to identify the wine quality from this parameter.

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?

This dataset seems to only contain a limited amount of very high and very low quality wines but many more in the middle. This results in a lower spread of the variables when investigating the low and high quality wines, whereas a larger spread in the data is seen for the ratings in between.
For some of the variables, it seems that they are converging on a very limited range for a high quality wine, e.g. when looking at the pH. This small range of high quality might not represent the truth due to the limited amount of high quality wines in the dataset.

Most of the datasets plotted against quality, showed either an increasing or decreasing trend to the best quality. However in most of them, the lowest quality wines usually also showed the same trend as the high quality wines, which makes it difficult to properly identify the best wines from these physical parameters.

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

There seems to be a strong relationship between alcohol and both residual sugar and density. The density was expected, as alcohol has a lower density than water, hence increasing the alcohol will decrease the wines density.
Another expected correlation was between citric acid and the pH level, as pH will decrease when more acid is present.

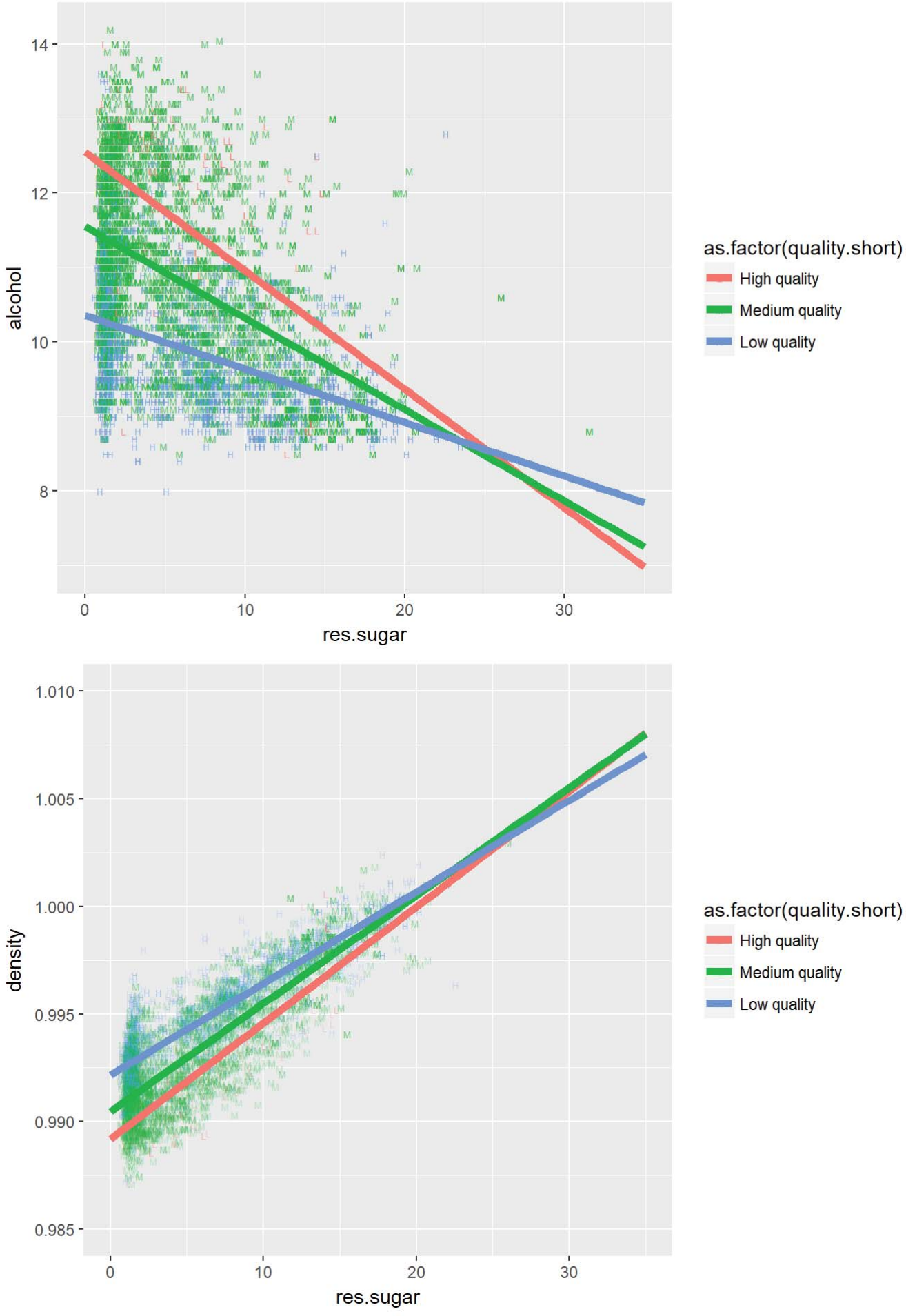
What was the strongest relationship you found?

Two strong relationships were found. The density and alcohol content has a Pearson correlation coefficient of -0.80 and residual sugar and density has a Pearson correlation coefficient of 0.83. Density is clearly influenced by both the alcohol content and residual sugar. Increasing the alcohol content will lower the density, as it is lighter than water. Increasing the residual sugar increases the density.

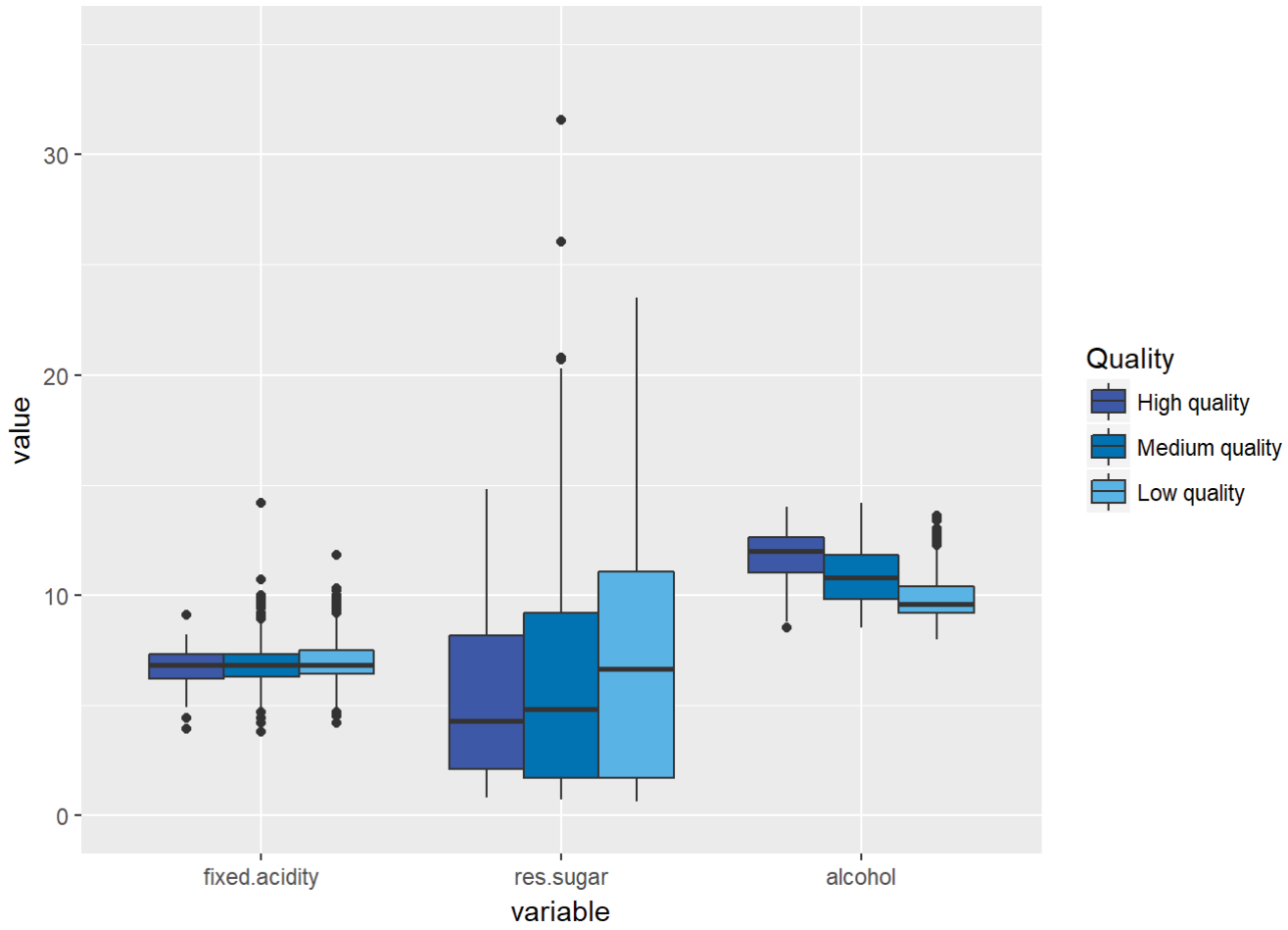
Multivariate Plots Section

To ease the comparison between the qualities of wine, they will be collected into three groups:
Low quality: <= 5
Medium quality: 6-7

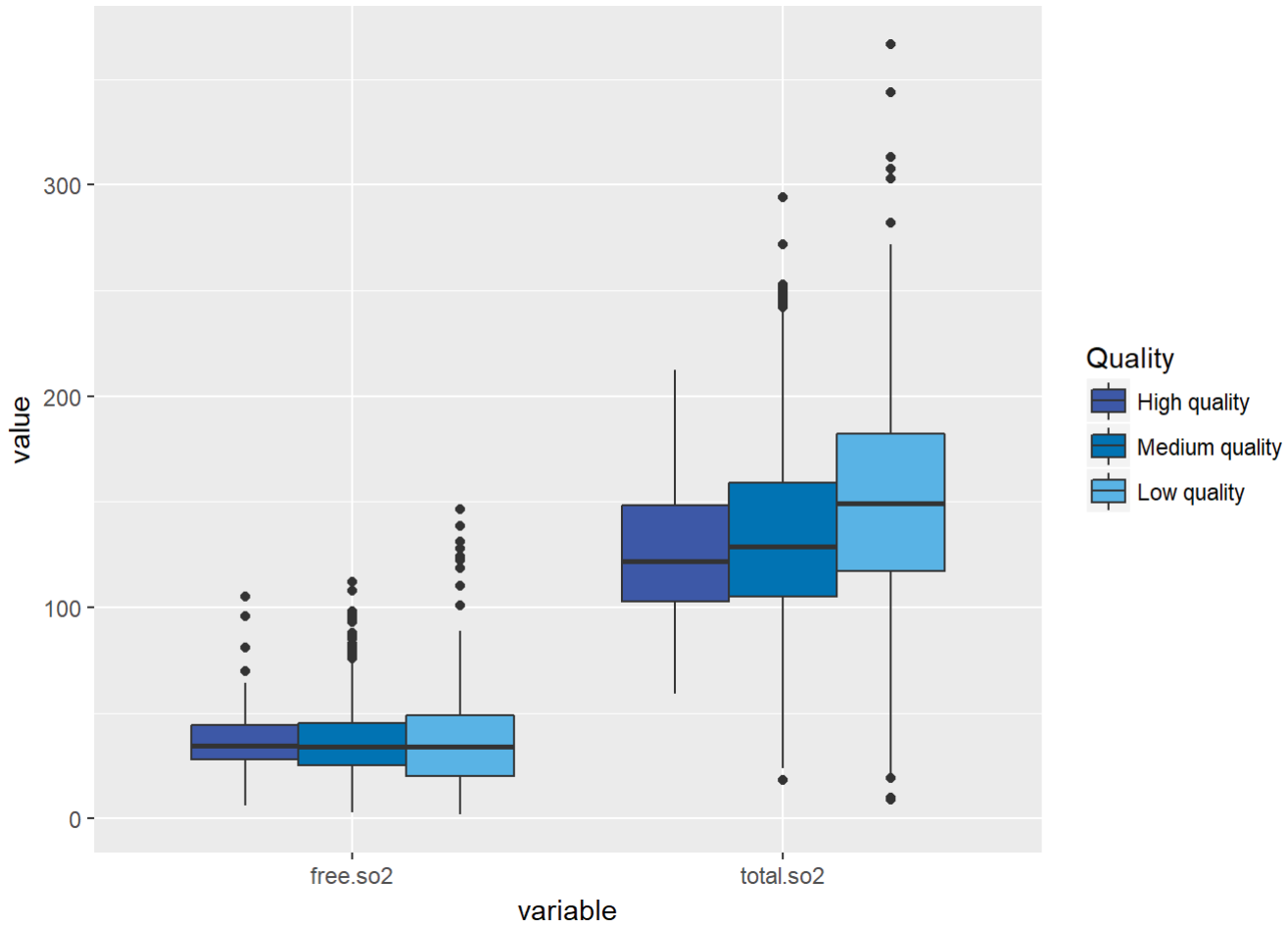
High quality: => 8



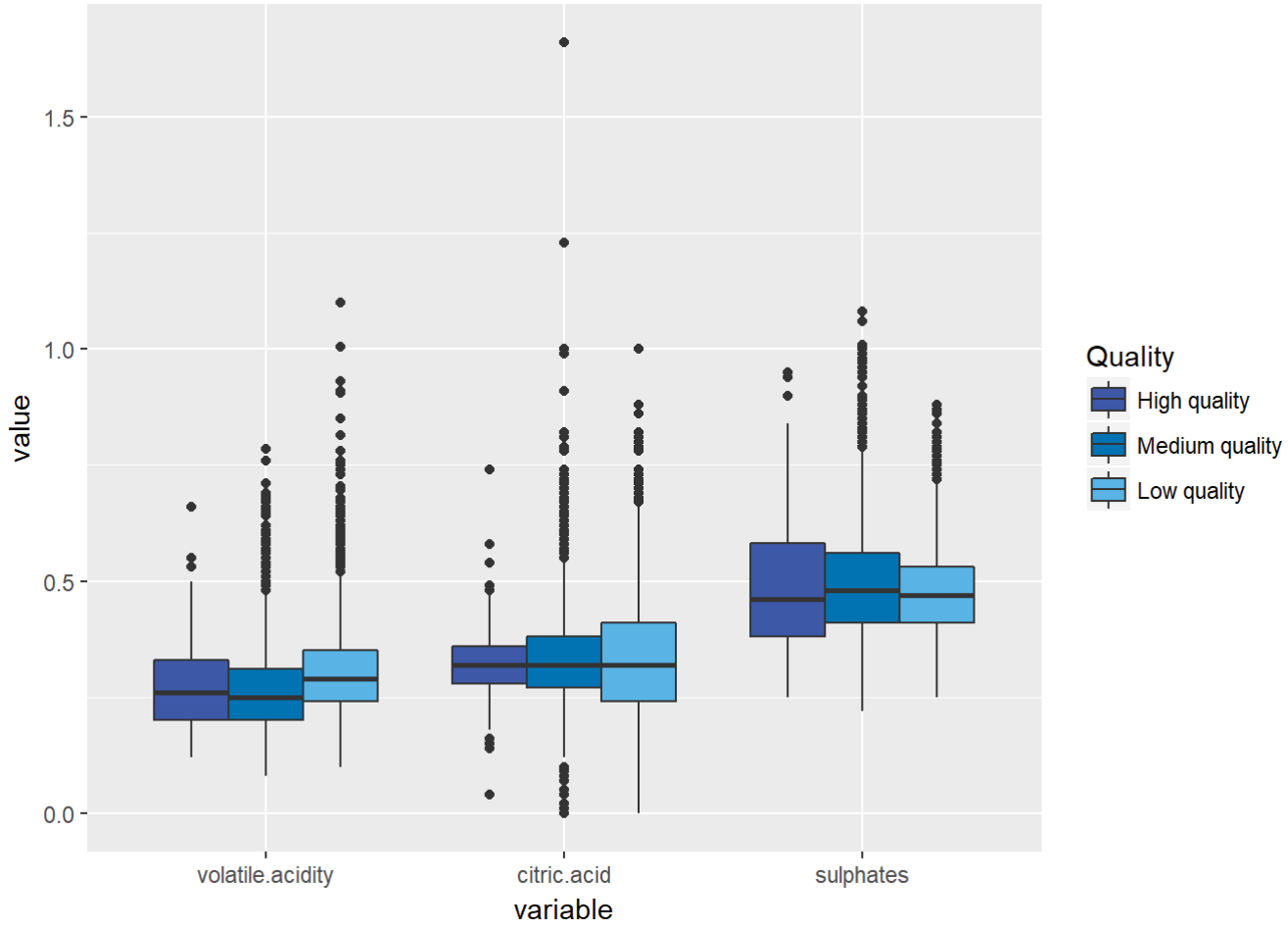
The first plot, shows the relationship between residual sugar and alcohol. The three lines are showing the low, medium and high quality wines. It is quite evident, that the higher alcohol content, the higher quality. All 3 lines almost intersects in the same place, but this is outside the data range and does not seem to be valid.
The second plot, shows the relationship between residual sugar and density The three lines are showing the low, medium and high quality wines. For this plot, the order of the lines have been reversed. High quality wines have a low density and low quality wines have a high density - this is again related to alcohol.



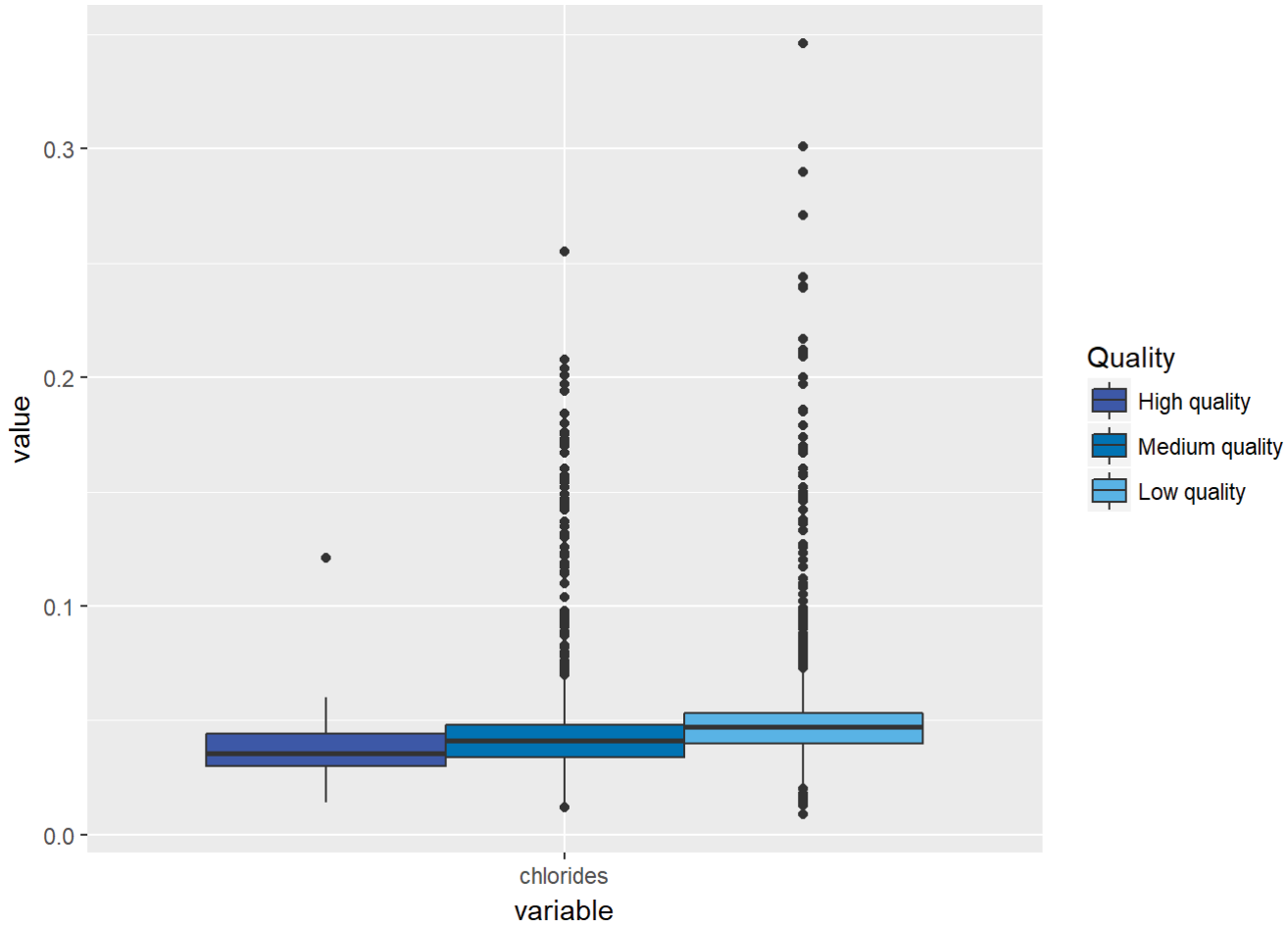
The median for ficed acidity seems to increase slightly for decreasing quality, whereas it is opposite for alcohol. For residual sugar, the level is almost the same for high and medium quality wines, where low quality wines have a much higher median.



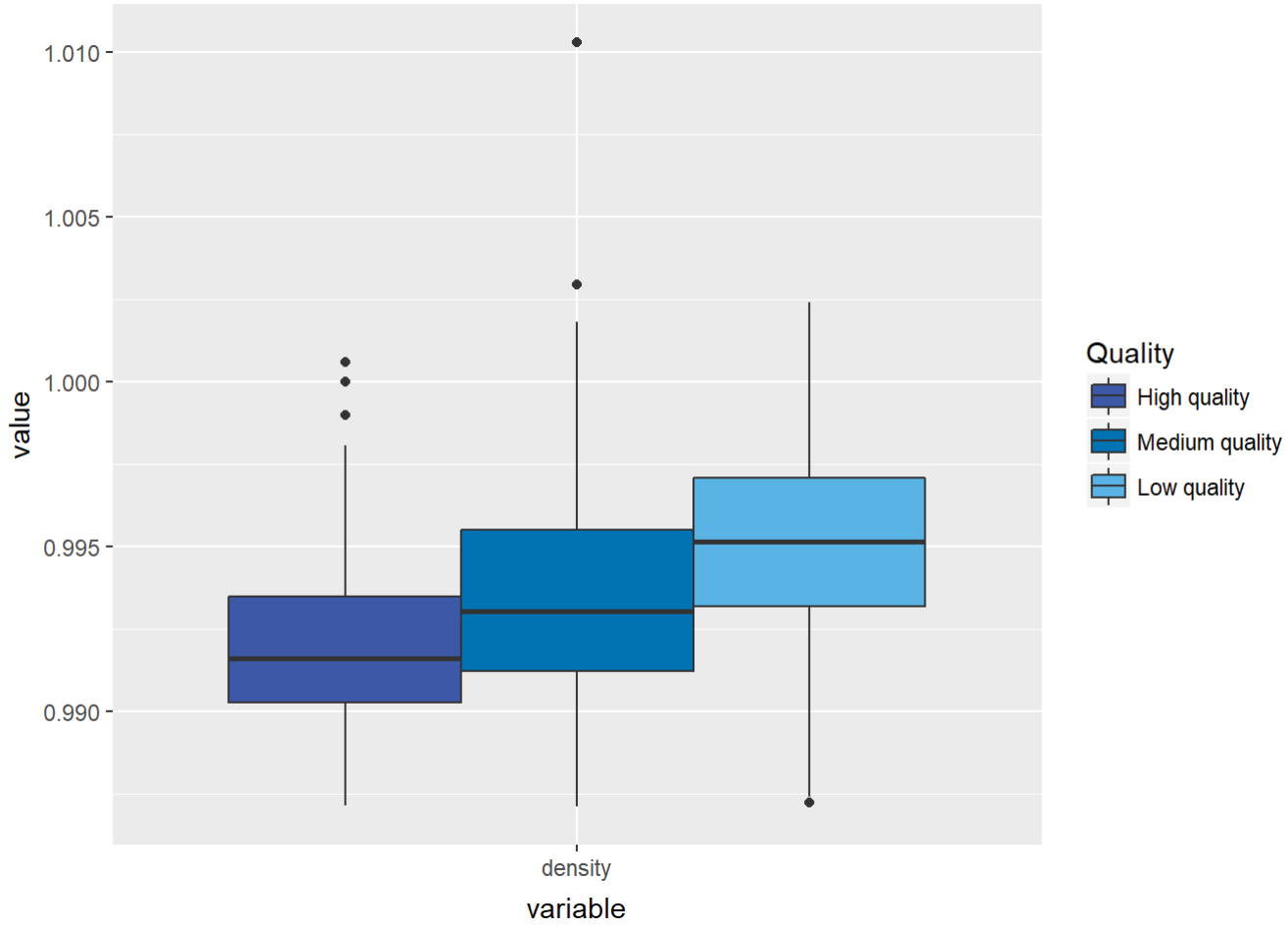
The free sulfur dioxide content does not seem to influence the quality, but the lower total sulfur dioxide the better quality.



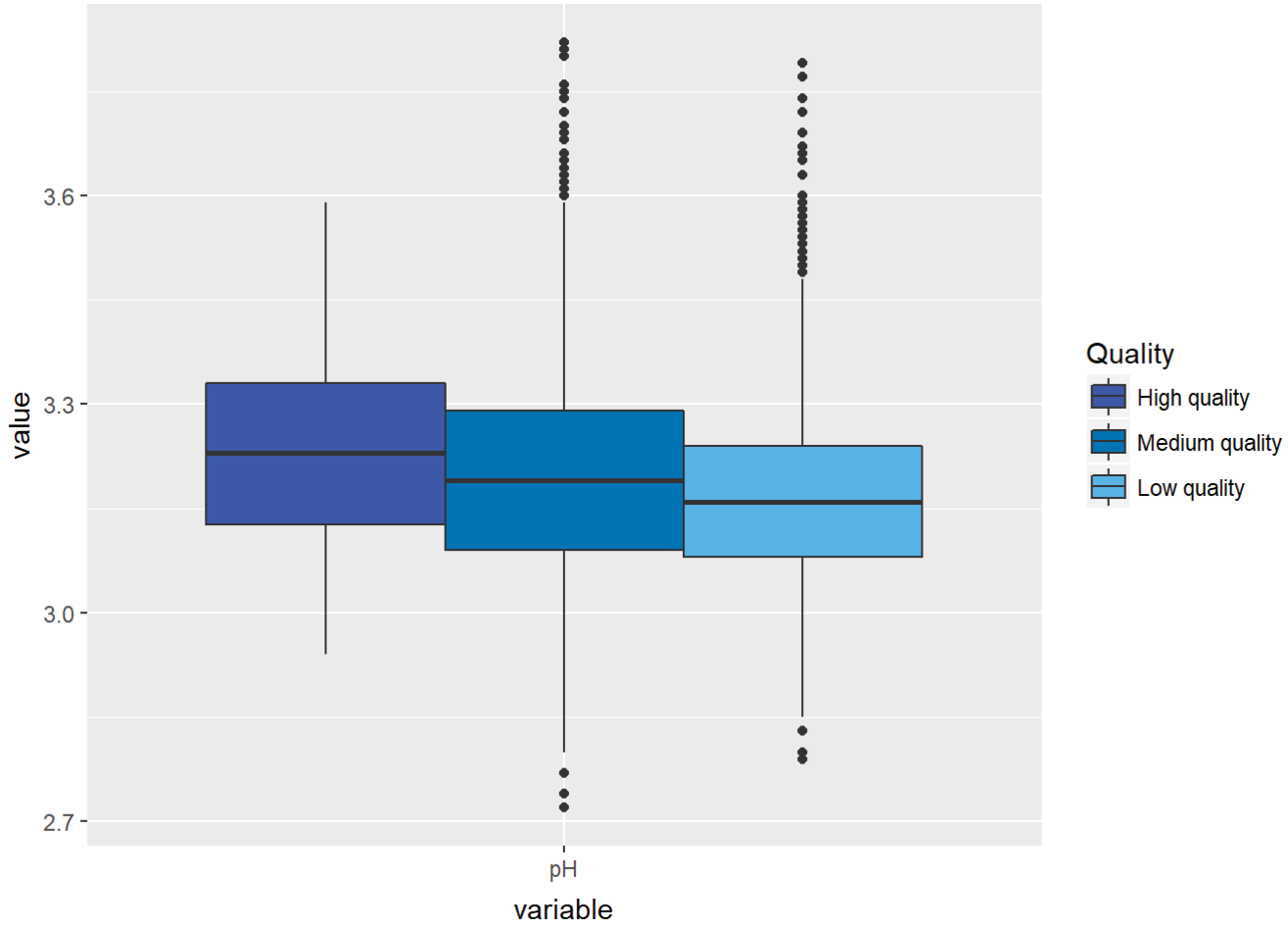
Acidity, citric acid and sulphates does not seem to have a large influence on the quality of the wines as the medians are almost constant across the qualities.



It could seem that the for lower chlorides you get a the better quality wine.



The lower density the higher quality. This is the reverse from alcohol as seen earlier and is expected due to the physical dependency between the two.



This plot indicated that for a higher pH you get a better quality white wine.

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?

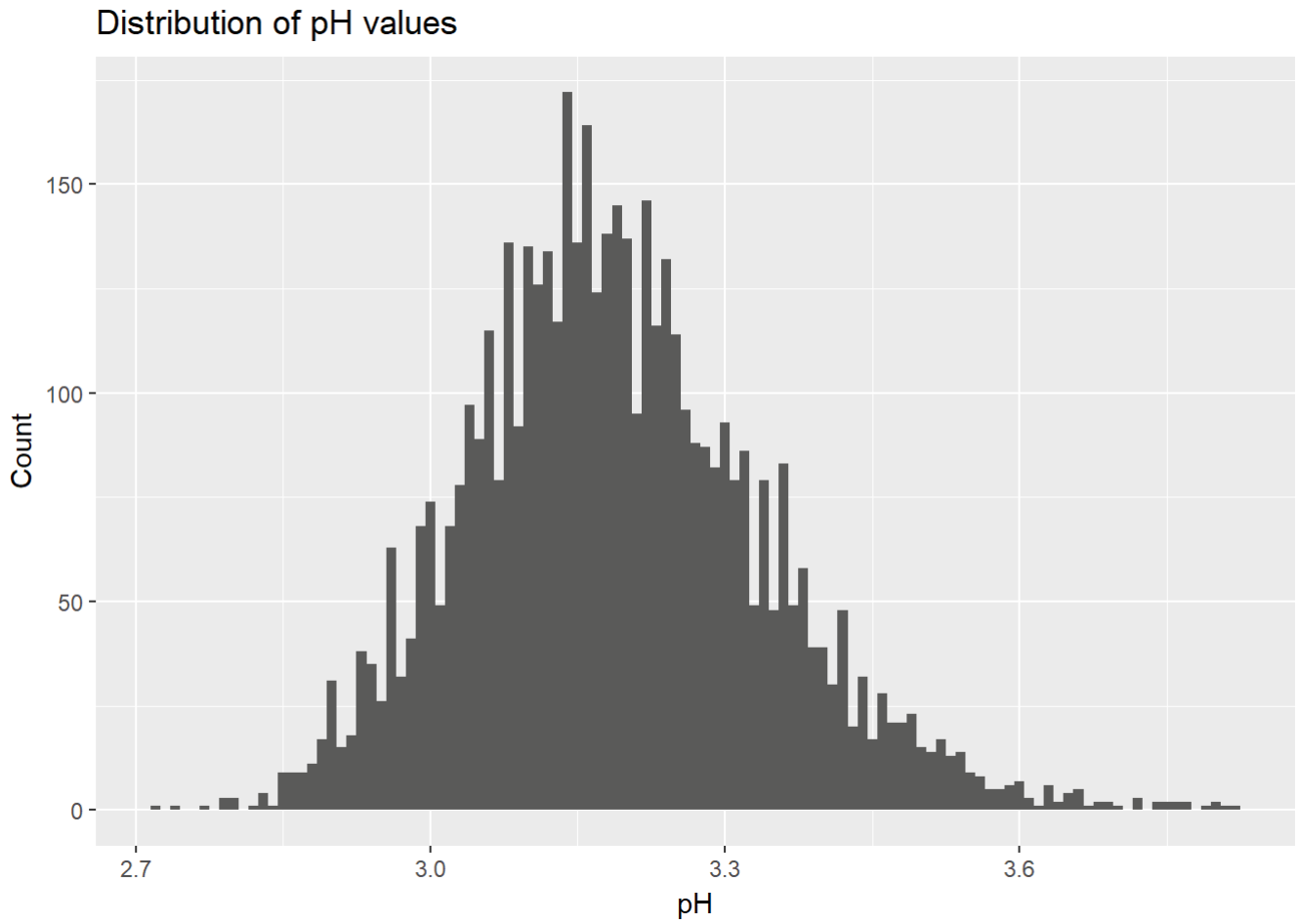
Some of the relationships between the different variables and quality, that was investigated in the bivariate section, has become more clear. E.g. the alcohol vs sectioned quality only indicates higher quality with increasing alcohol, although the plot in the bivariate section showed a slight increase in alcohol for low quality wines. This is due to the sectioning of quality into low, medium and high quality. This sectioning could remove some trends, e.g. the tendency that was seen in bivariate analysis with high and low quality wines showing the same behaviors.

Were there any interesting or surprising interactions between features?

I would have assumed that citric acidity, volatile acidity and sulphates would have had a larger influence on the perception of the quality of white wine. From this investigation, it does not seem that these variables have a significant influence and other parameters are of much more interest.

Final Plots and Summary

Plot One

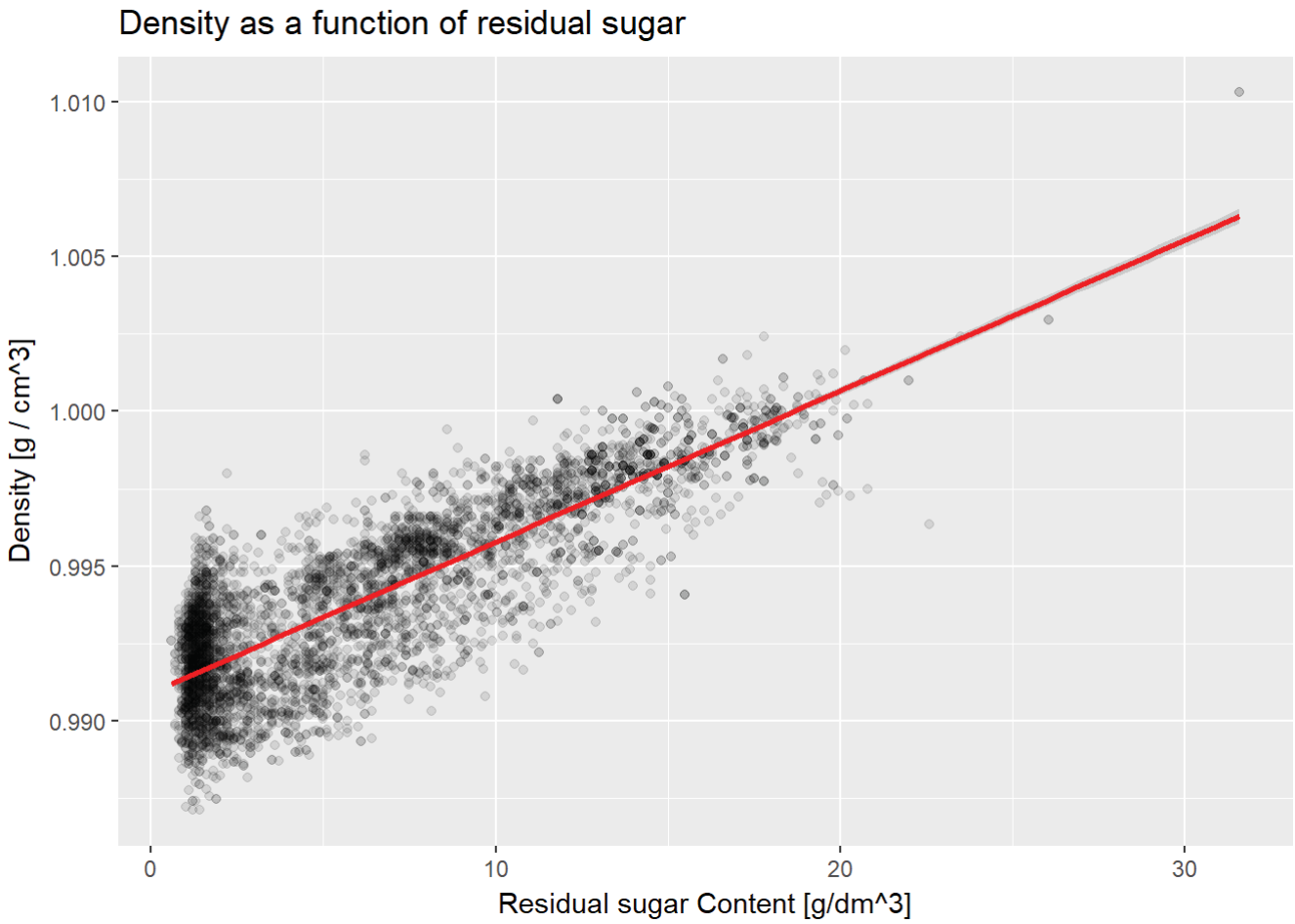



```
## [1] "The pH median is 3.2 and the mean is 3.2"
```

Description One

The pH variable is a nice example of a normal distribution with the mean equal to the median indicating no skew.

Plot Two

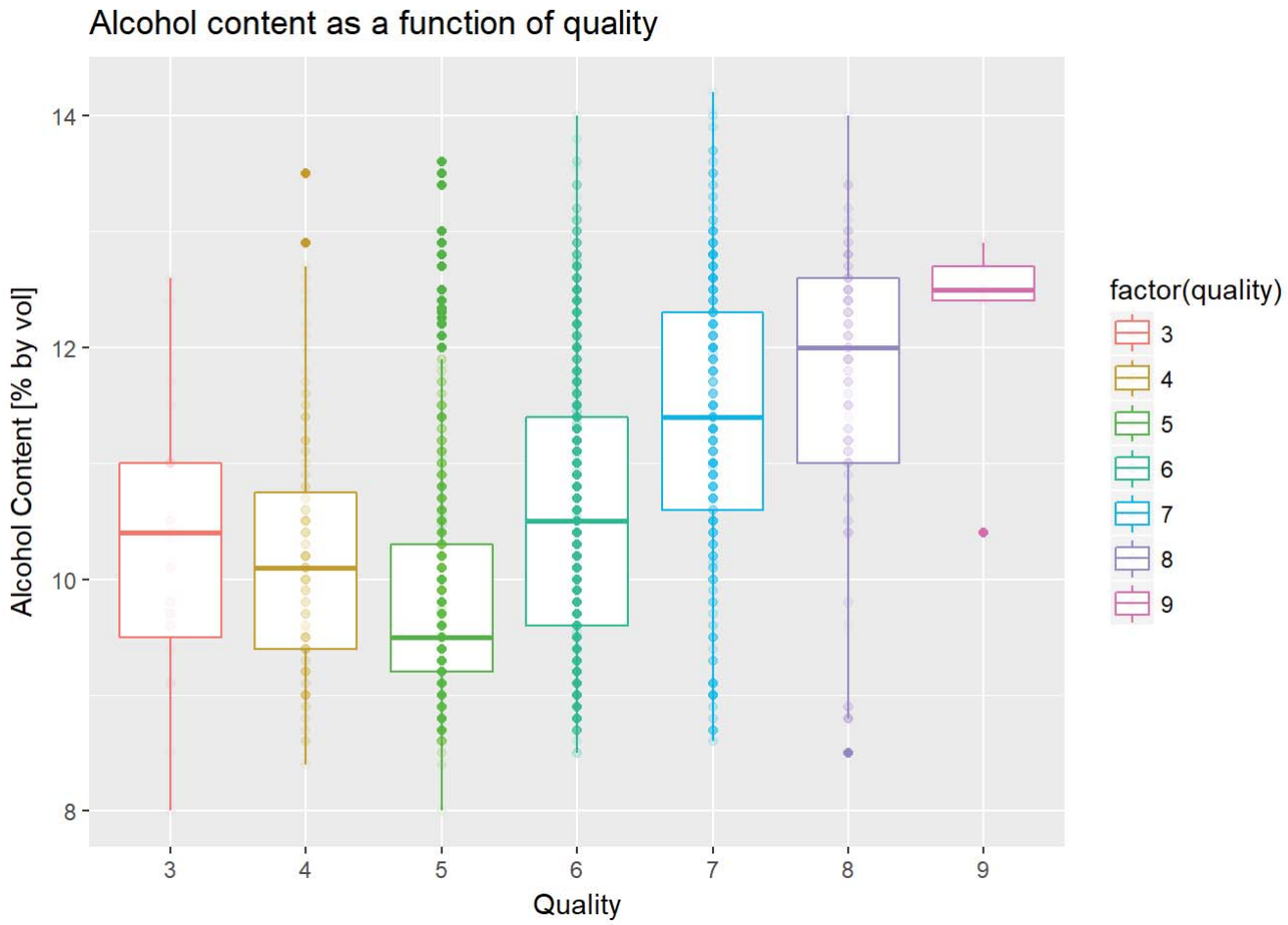



```
## [1] "Pearsons Correlation Coefficient: 0.833975"
```

Description Two

There is a strong positive correlation between density and the residual sugar. This makes sense as increasing the sugar content of the wine will increase the density.

Plot Three



Description Three

The data seems to indicate that you should look for a high alcohol content and preferably above 12% in order to get a nice white wine. Interestingly you could be tricked when choosing a white wine with an alcohol content of ~10.5%, as it could have a quality rating of both 3 and 6.

Reflection

It is surprising that citric acid, sulphates and volatile acidity does not seem to have a big influence on the quality of white wines, whereas alcohol and density seems to be the most important factors. Density is closely correlated to alcohol content and residual sugar as these two impacts the density of the wine. From all the parameters a high alcohol content is the key indicator for a good quality white wine.
This evaluation builds on physical parameters of white wine and some quality value. According to the dataset, the quality parameter has been established using "sensory data". Wine taste and quality are very subjective and will most likely vary from person to person, which is why this quality parameter can be hard to trust. To test some of the conclusions from this dataset a blind tasting with a large group of people could be used to verify the conclusions.

One of the struggles with this dataset has been to identify the correlations between the variables and finding the greatest indicators of a high quality white wine. After identifying relationships and correlations I have spent a great deal of time contemplating how to show the data and which plots and R functions that could be utilized.