ANIKO LENGYEL - WHITE WINE DATA SET ANALYSIS

I load the data and get the first seven rows.

```
##   X fixed.acidity volatile.acidity citric.acid residual.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
## 7 7           6.2             0.32        0.16            7.0     0.045
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  45                  170  1.0010 3.00      0.45     8.8
## 2                  14                  132  0.9940 3.30      0.49     9.5
## 3                  30                   97  0.9951 3.26      0.44    10.1
## 4                  47                  186  0.9956 3.19      0.40     9.9
## 5                  47                  186  0.9956 3.19      0.40     9.9
## 6                  30                   97  0.9951 3.26      0.44    10.1
## 7                  30                  136  0.9949 3.18      0.47     9.6
##   quality
## 1       6
## 2       6
## 3       6
## 4       6
## 5       6
## 6       6
## 7       6
```

Looking at the data structure and data types.

```
## 'data.frame':    4898 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
```

Getting the summary of the dataset to get information about the missing rows (if there are any).

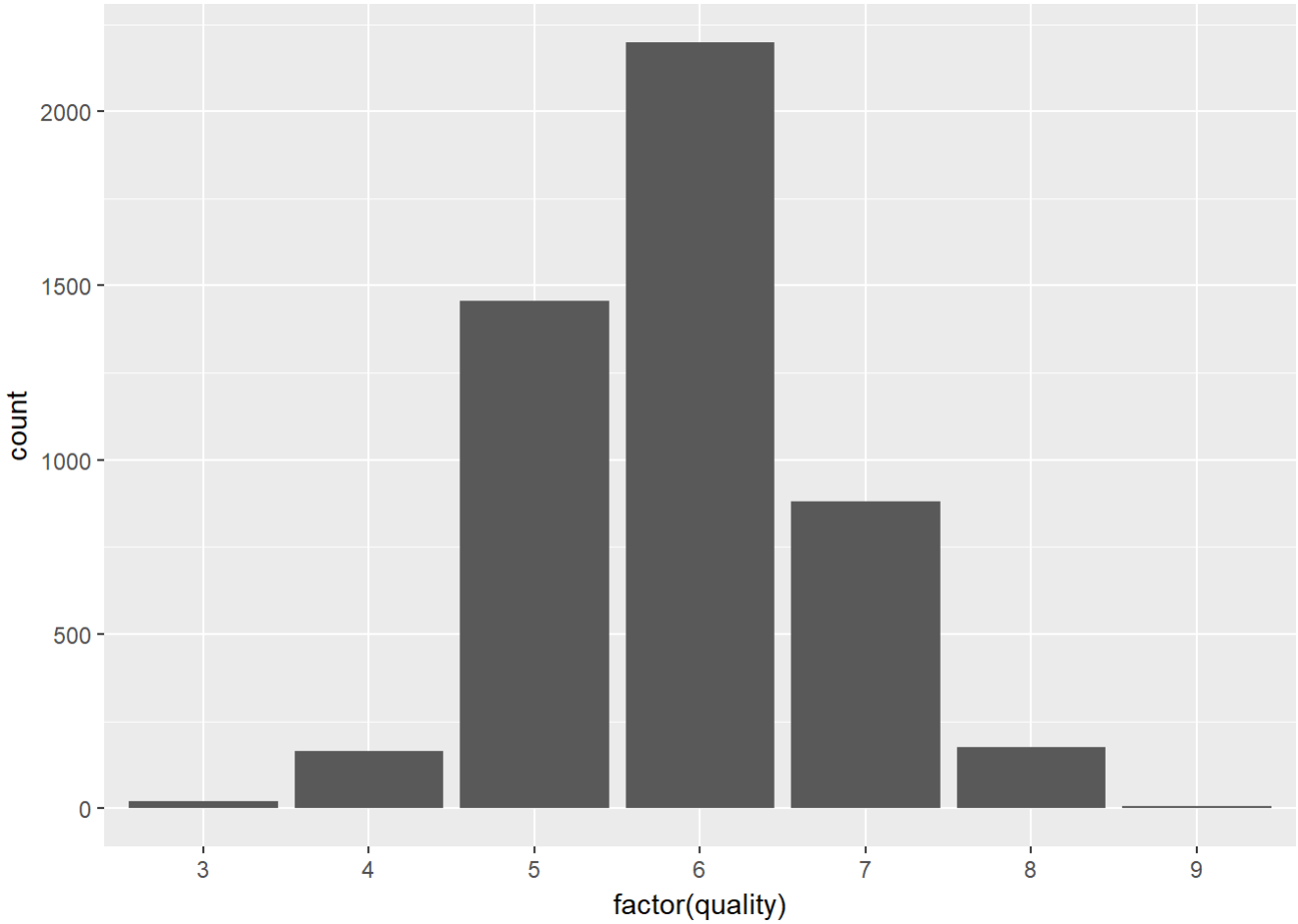
```
## [1] 0
```

Introduction:

I found this public dataset among the Udacity's data set options. It catched my eyes because I am interested in wine making and wanted to know more about its background. The dataset consists of 13 variables and 4898 observations, integer and numeric data types, containing data about different features of white wines, for example acidity, sugar and alcohol contain and quality. I plan to analyze the relation between the different components and wine quality.

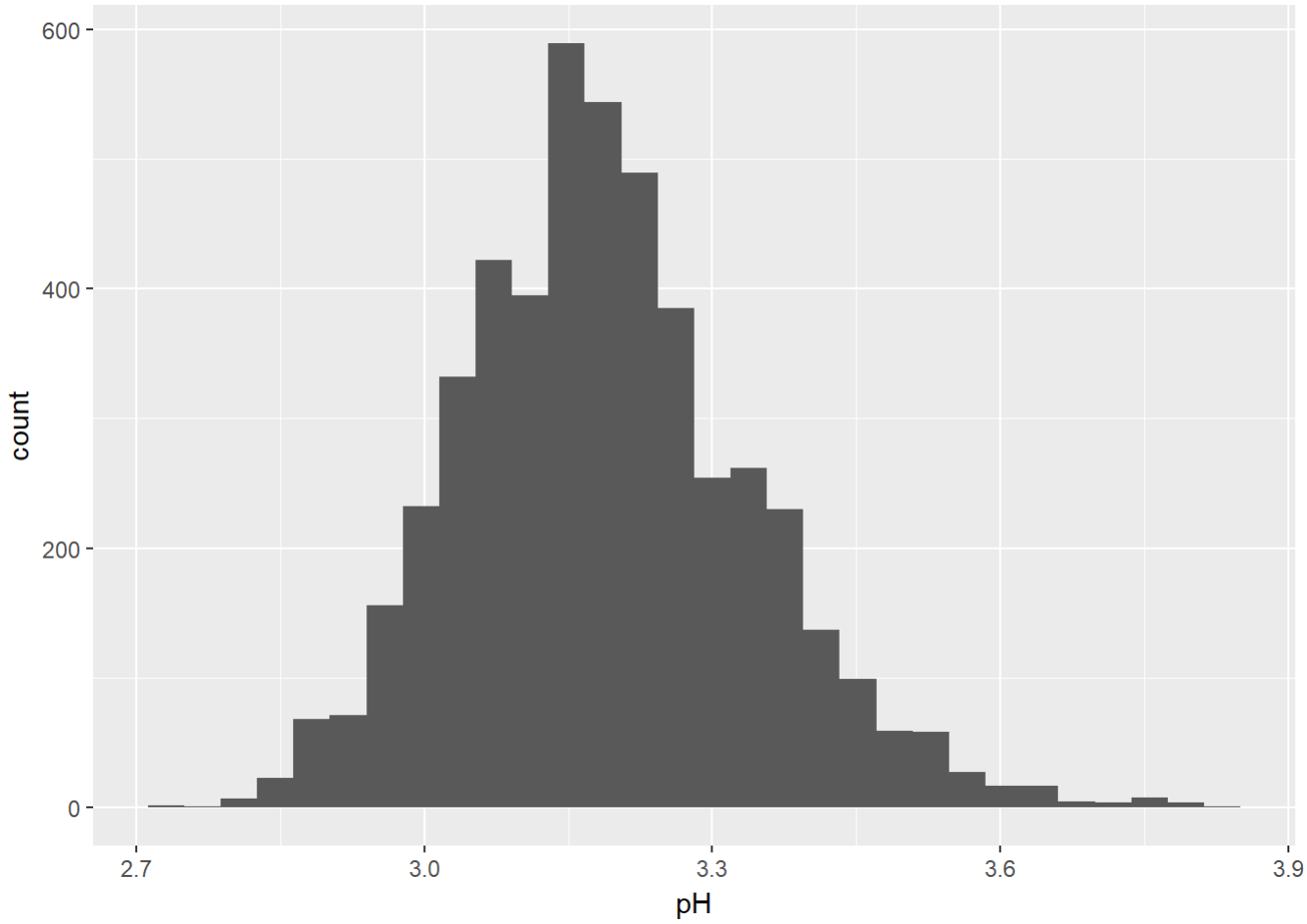
Univariate Plots Section

At first, I create a bar plot to get a quick view about the distribution of the quality rates and I run a summary to generate some summary statistics.



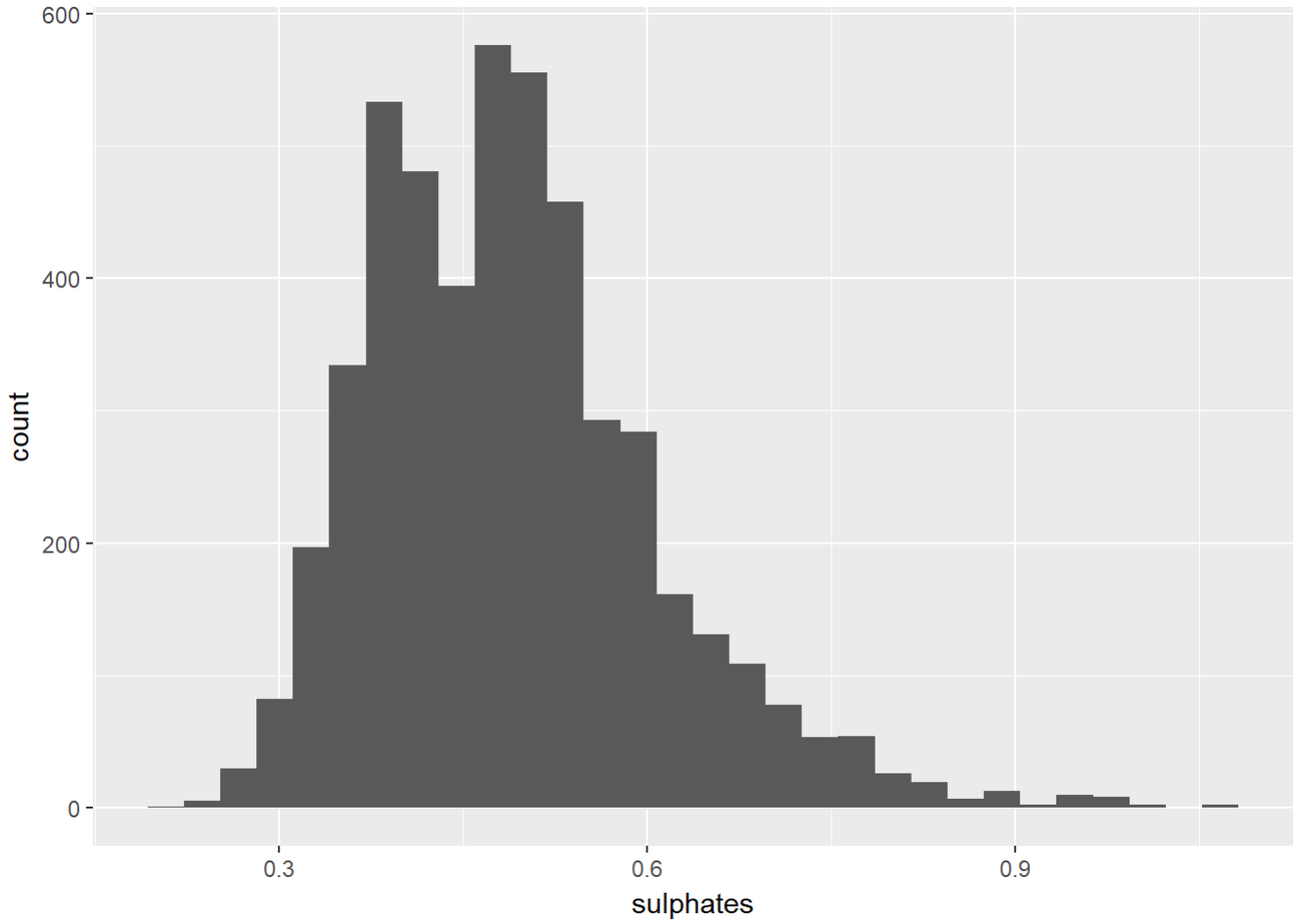

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.878   6.000   9.000
```

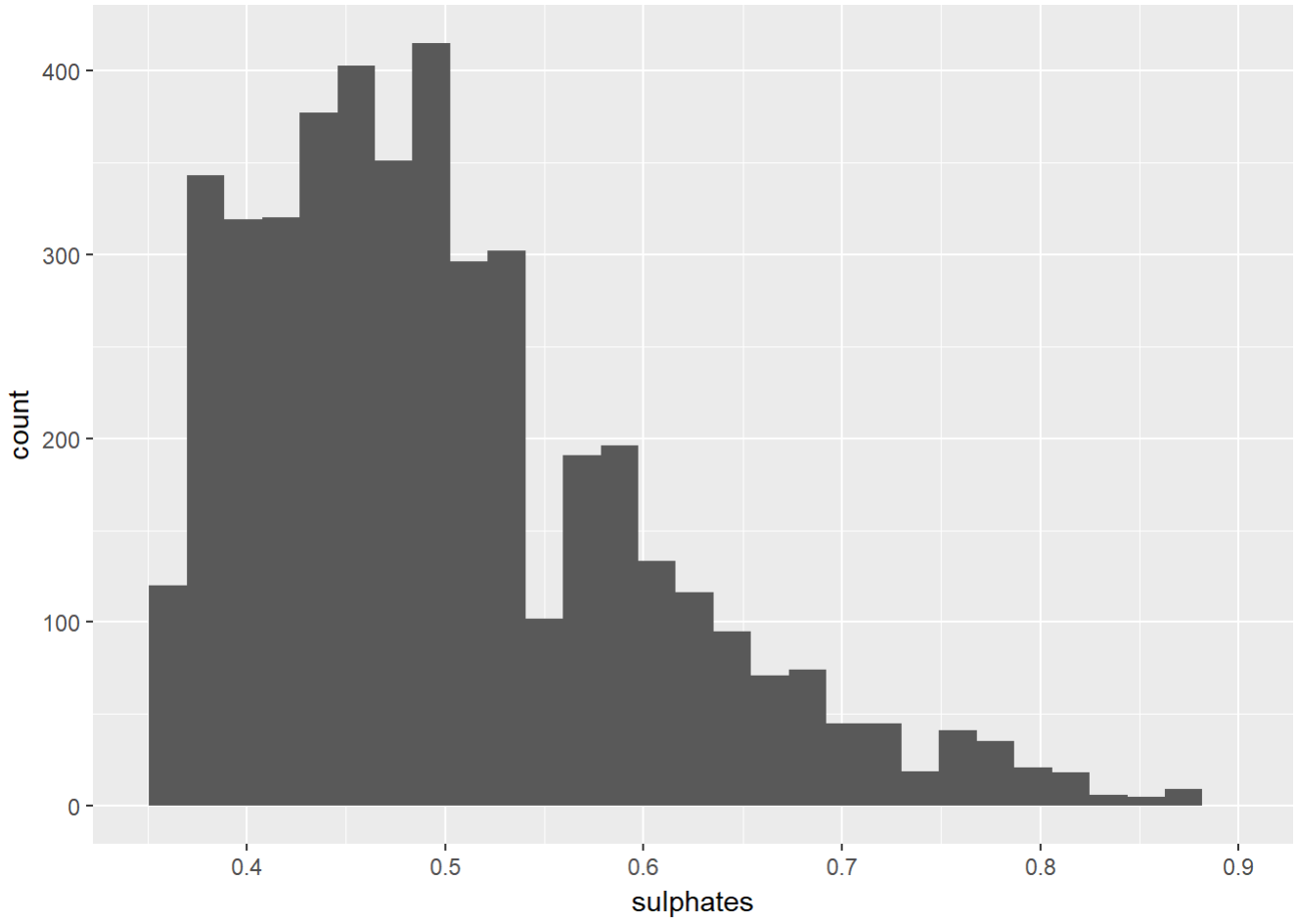
I got a normally distributed histogram, most wine have a quality between 5 and 7, the tallest clusters of bars is 6, representing the most common quality. Based on my summary, the the mean value is 5.878, the min quality is 3 and the max quality is 9.




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

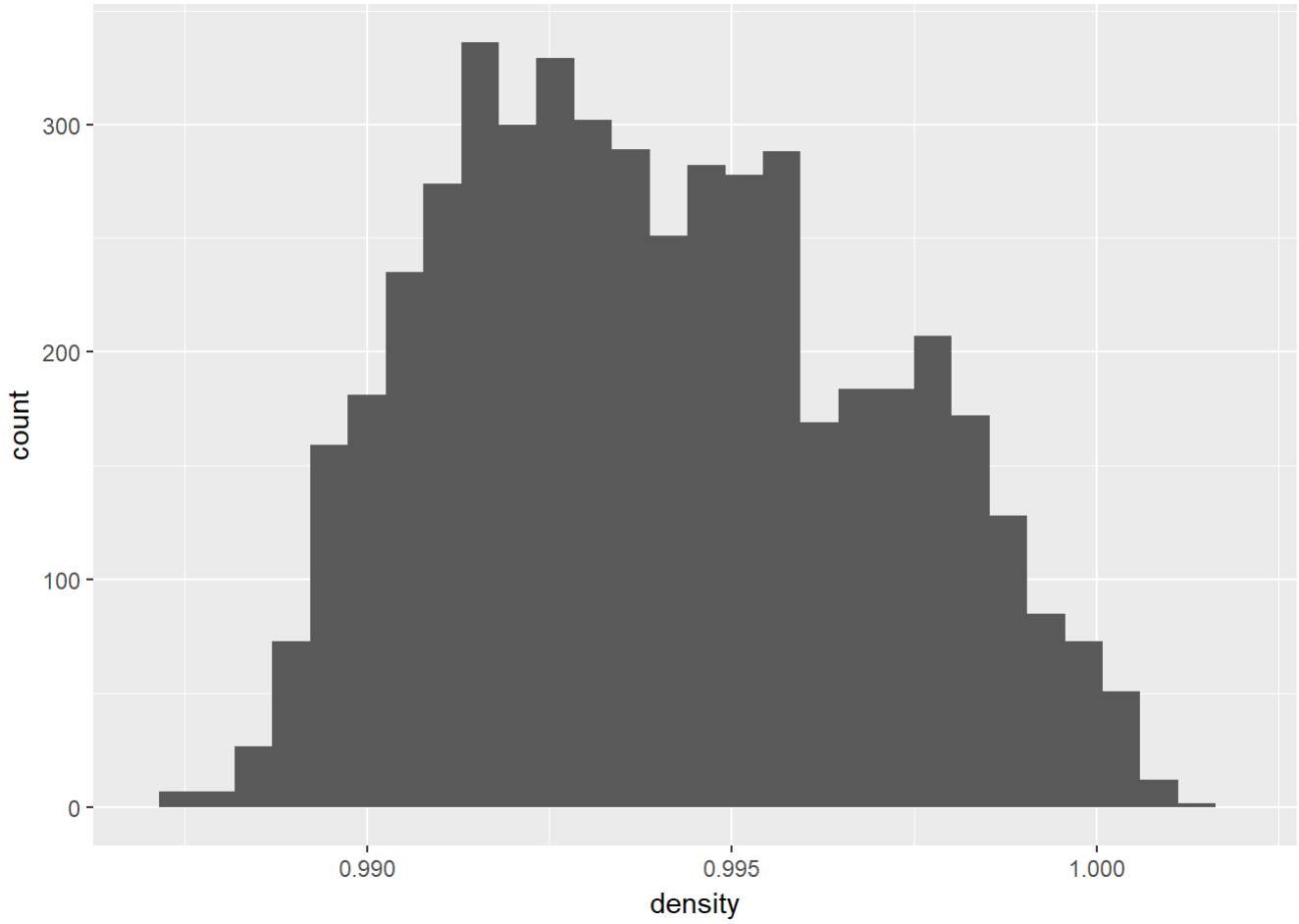
The histgoram of pH is also normally distibuted and concentrated around 3.15. The min and max values are 2.72 and 3.82, the median is 3.18.


600
400
200
0
count
0.3
0.6
0.9
sulphates


400
300
200
100
0
count
0.4
0.5
0.6
0.7
0.8
0.9
sulphates

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

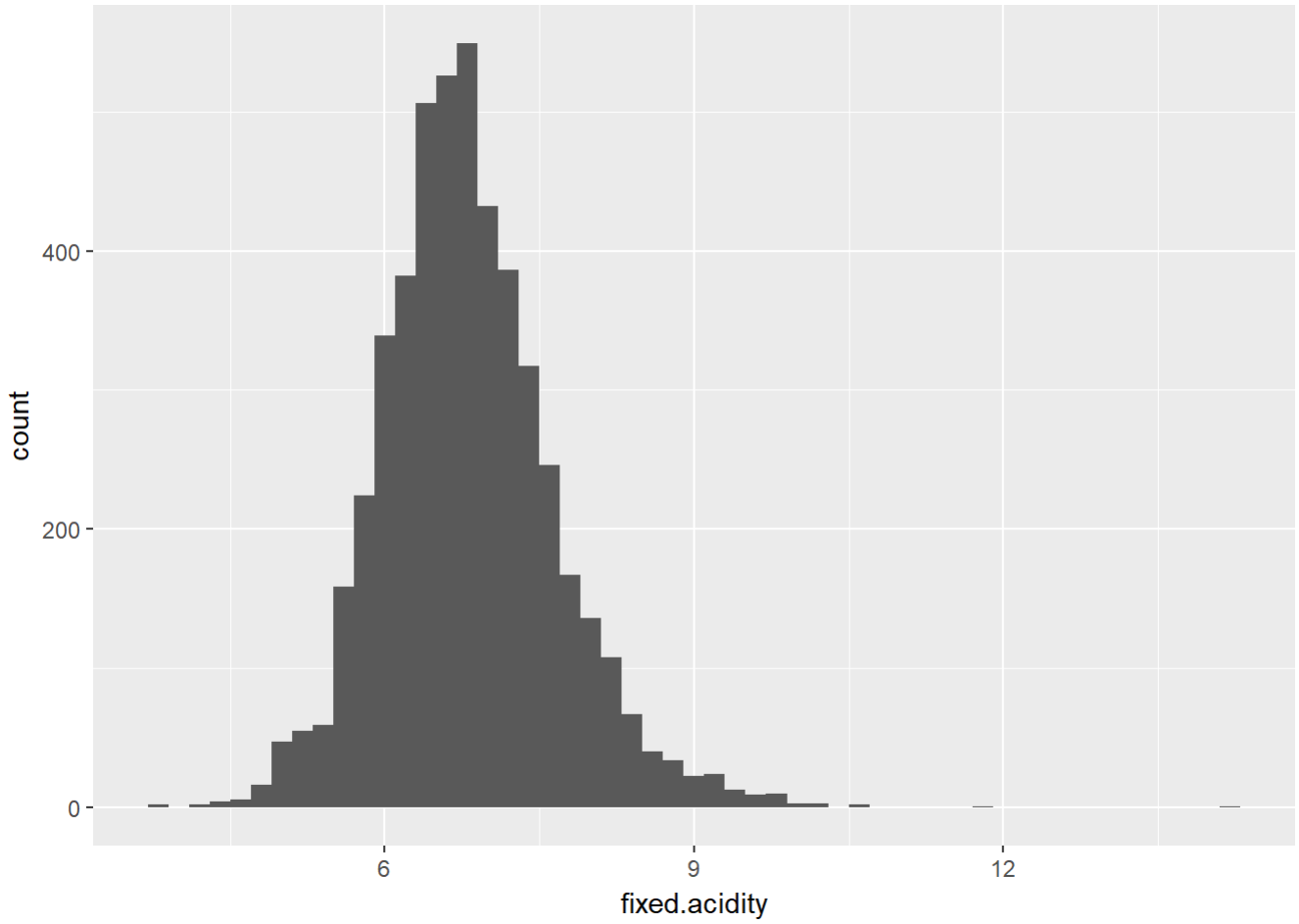
The histogram of suplhate content is right skewed - the most ofttten occuring values are between 0.4 and 0.5, the peak is about 400. I used an xlim function to improve readability and remove outstanding values.

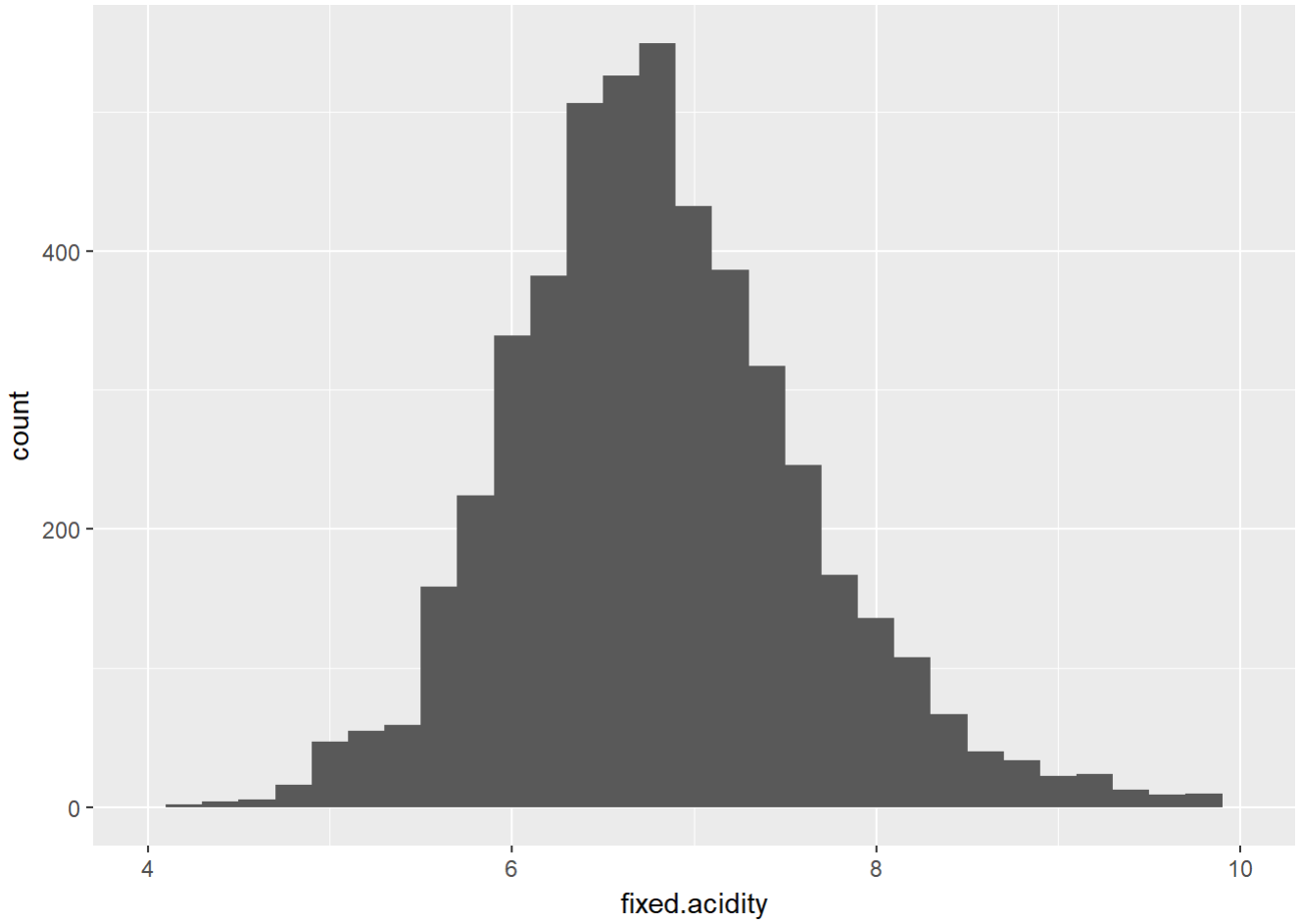



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9917  0.9937  0.9940  0.9961  1.0390
```

The the density hitogram looks normal distributed at the first sight, but it is a little skewed to the left. The difference between min and max values are less than 0.5.

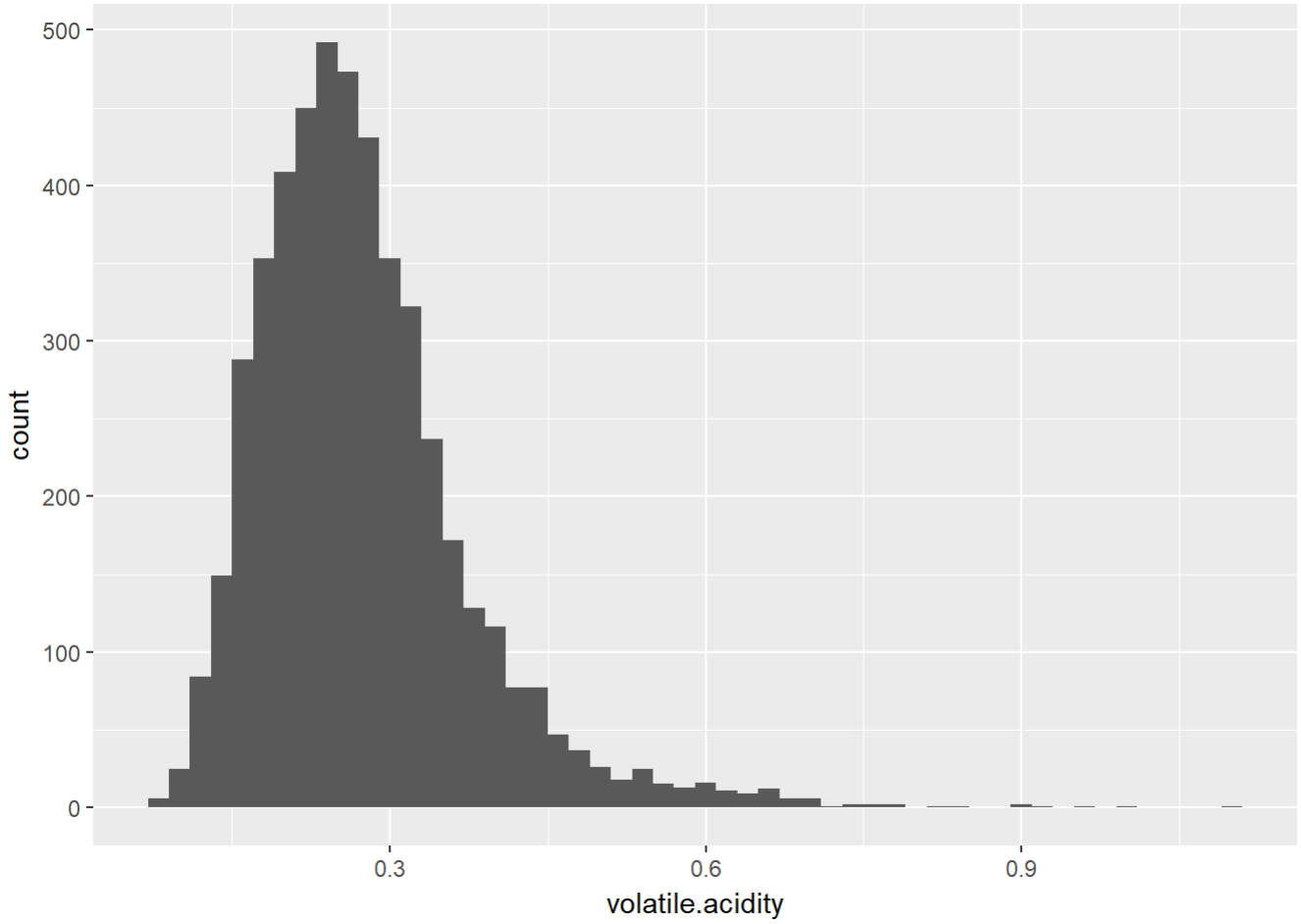
On the second plot, I used the parameter xlim to handle the outliers and make the plot more readable.

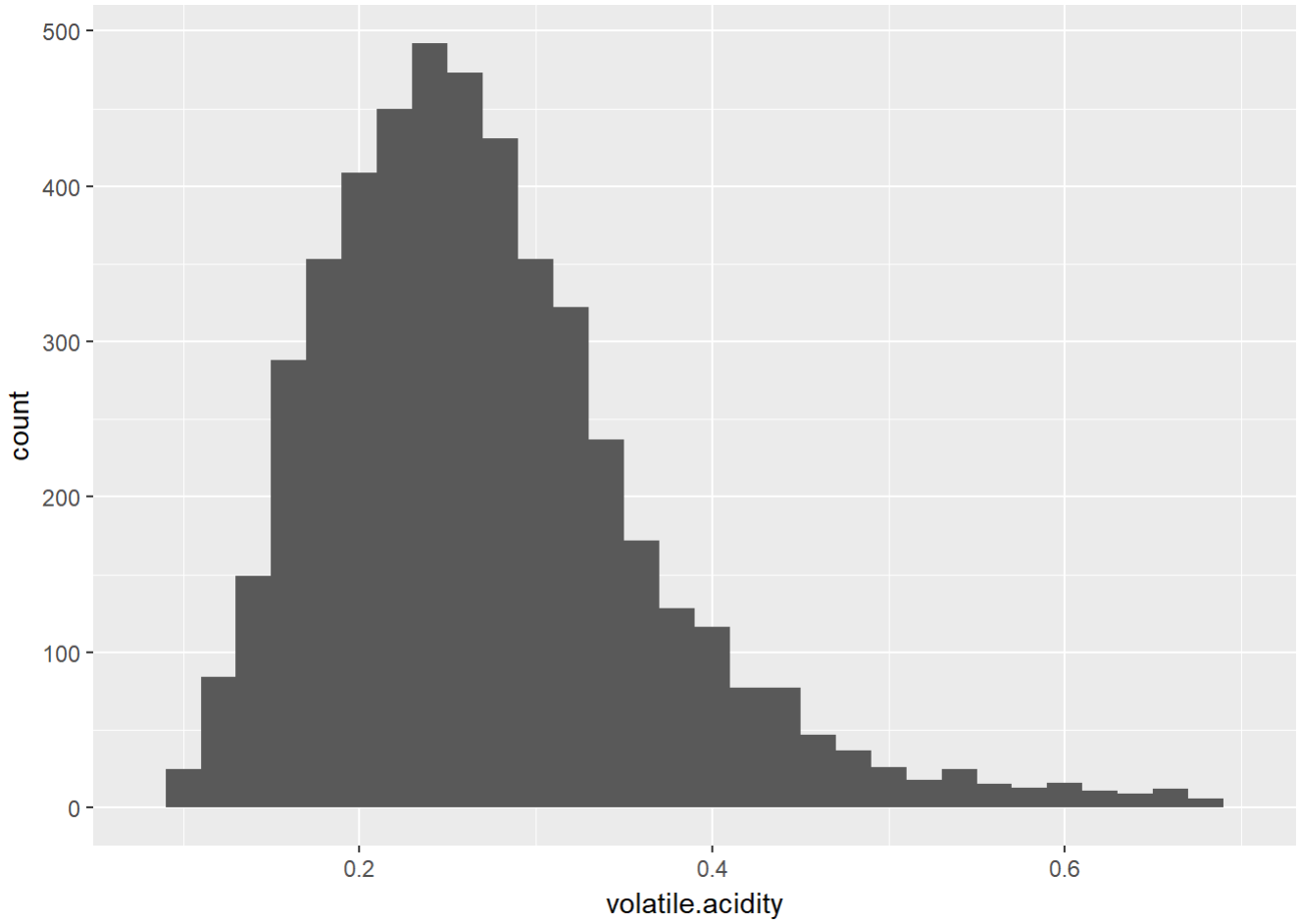

count
400
200
0
6
9
12
fixed.acidity


count
400
200
0
4
6
8
10
fixed.acidity

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

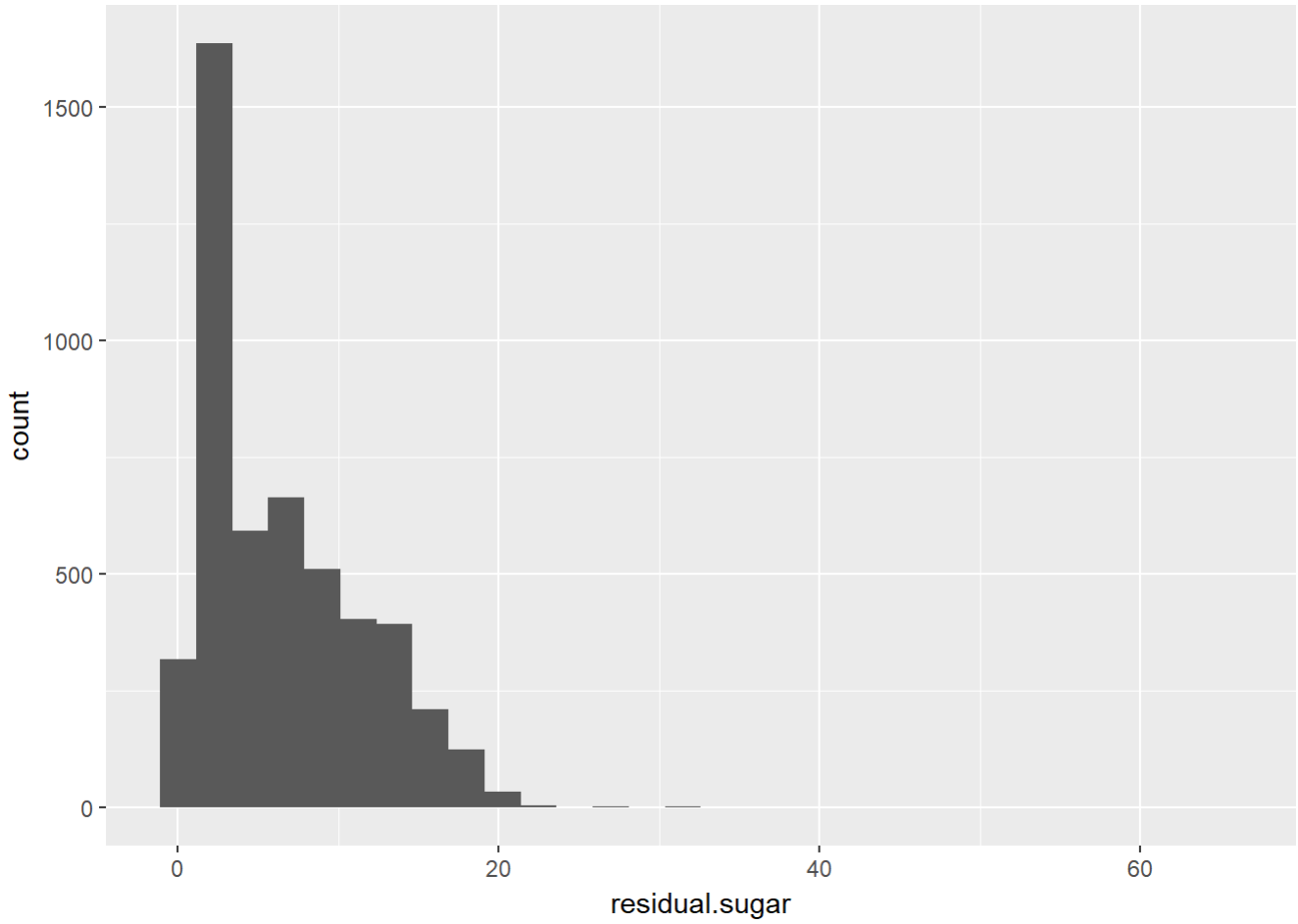
The ditribution of fixed acidity has its peak around 7 and skeewed to the right. I removed the outliers with xlim on the second plot.

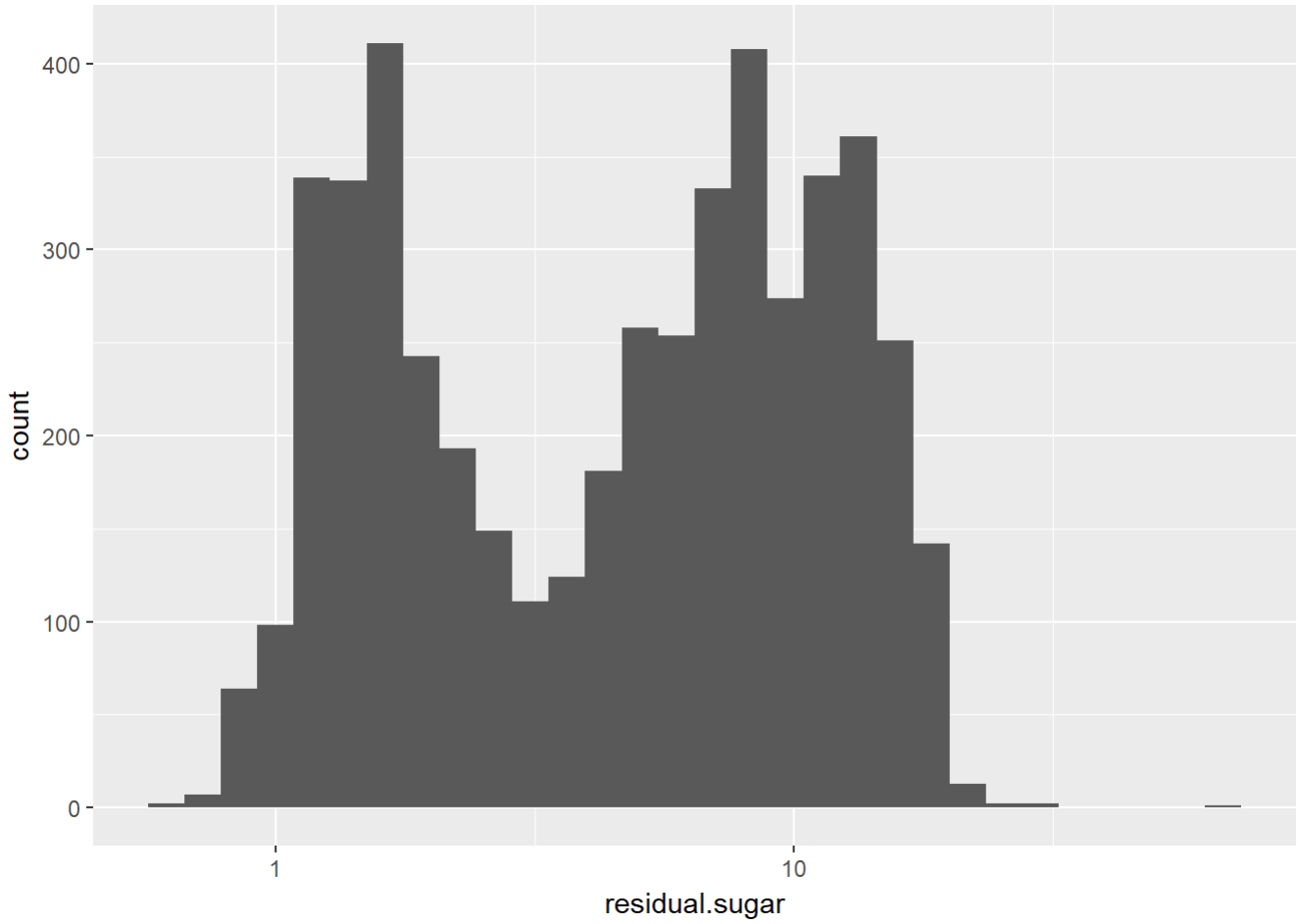





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.2100  0.2600  0.2782  0.3200  1.1000
```

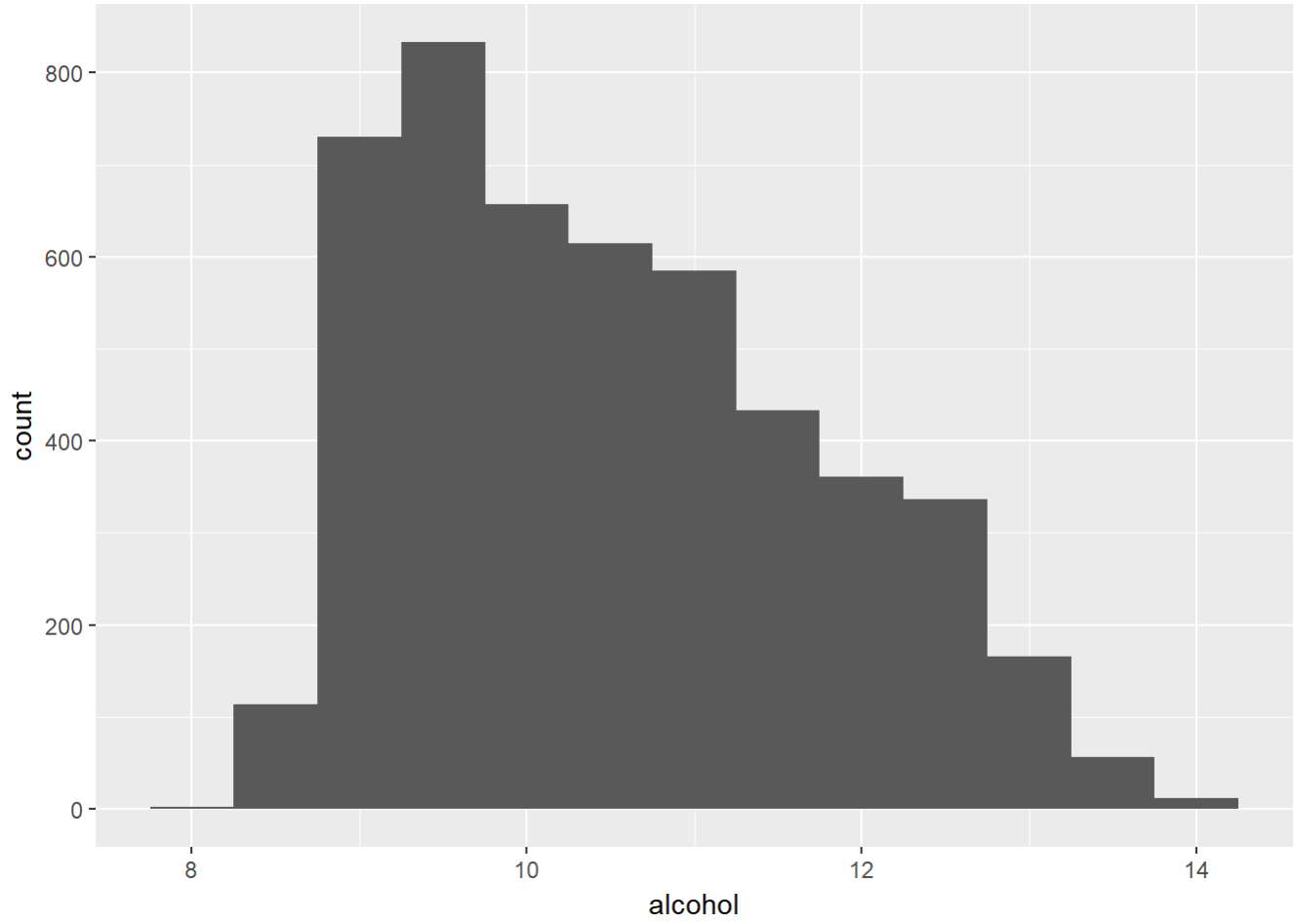
The histogram of volatile acidity is concentrated around 0.2 and skewed to the right. The most values are lacated between 0.15 and 0.325. I transformed the second plot with xlim to limit the outliers.


1500
1000
500
0
count
0
20
40
60
residual.sugar


400
300
200
100
0
count
1
10
residual.sugar

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

The distribution of residual sugar is skewed to the right. Based on the summary and the histogram, the distribution of the sugar content is relatively divided: the min value is 0.6 and the max value is 65 - ten times the min value! The values first three quarter are less than 9.9, indicating that there is only a few wines which contains less sugar. This also corresponds to the histogram: I can hardly see any value above 20. There were some outliers on the first residual sugar plot, so I decided to use scale_x_log_10() to handle them.




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

The distribution of alcohol is right skewed, most wine have their alcohol content between 8.5 and 11. There is no wine with lower alcohol content that 8, and the maximum alcohol content is a little above 14. I setted the bidwidt to 0.5 for a better look.

Univariate Analysis

What is the structure of your dataset?

The dataset consists of 13 variables and 4898 observations and contains data about different features of white wines, including rates of acidity, pH value, sugar contain and quality rates. The variables are all quantitative values, represented by integer and numeric data types. According to my calculation, there are no any missing

values.

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

I would like to determine which features are the most important in predicting the wine quality, so I plan to examine the possible correlaton between wine quality and other components, like residual sugar, alcohol content, acidity and density.

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

As far I know, sugar content can reasonably influence the quality of the wine. The high quality wines have lower sugar content while the lower quality often comes with a higher sugar content. I will also analyze the correlation between alcohol content, density and quality.

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

I did not create any new variables in the dataset. However, I was wondering about creating three bins based on wine quality (like bad-average-good) but I decided to make a more detailed analysis on quality.

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

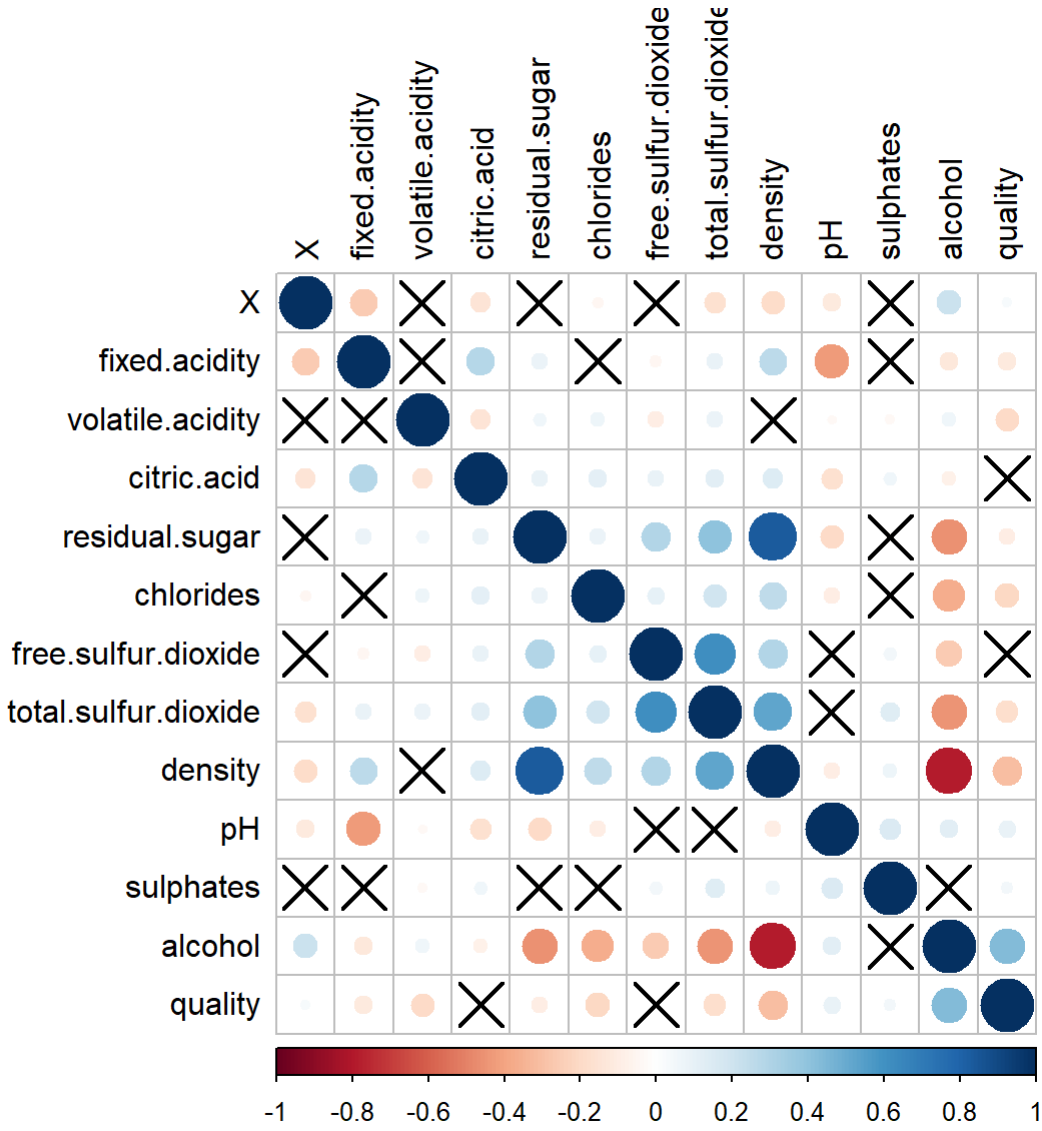
The residual sugar and volatile acidity data have some unusual distributions - both of the histograms are strongly right skewed. I wonder if these components have something to do with wine quality.

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?

Since I got a tidy dataset, I did not need to perform any cleaning operation or other kind of adjusting.

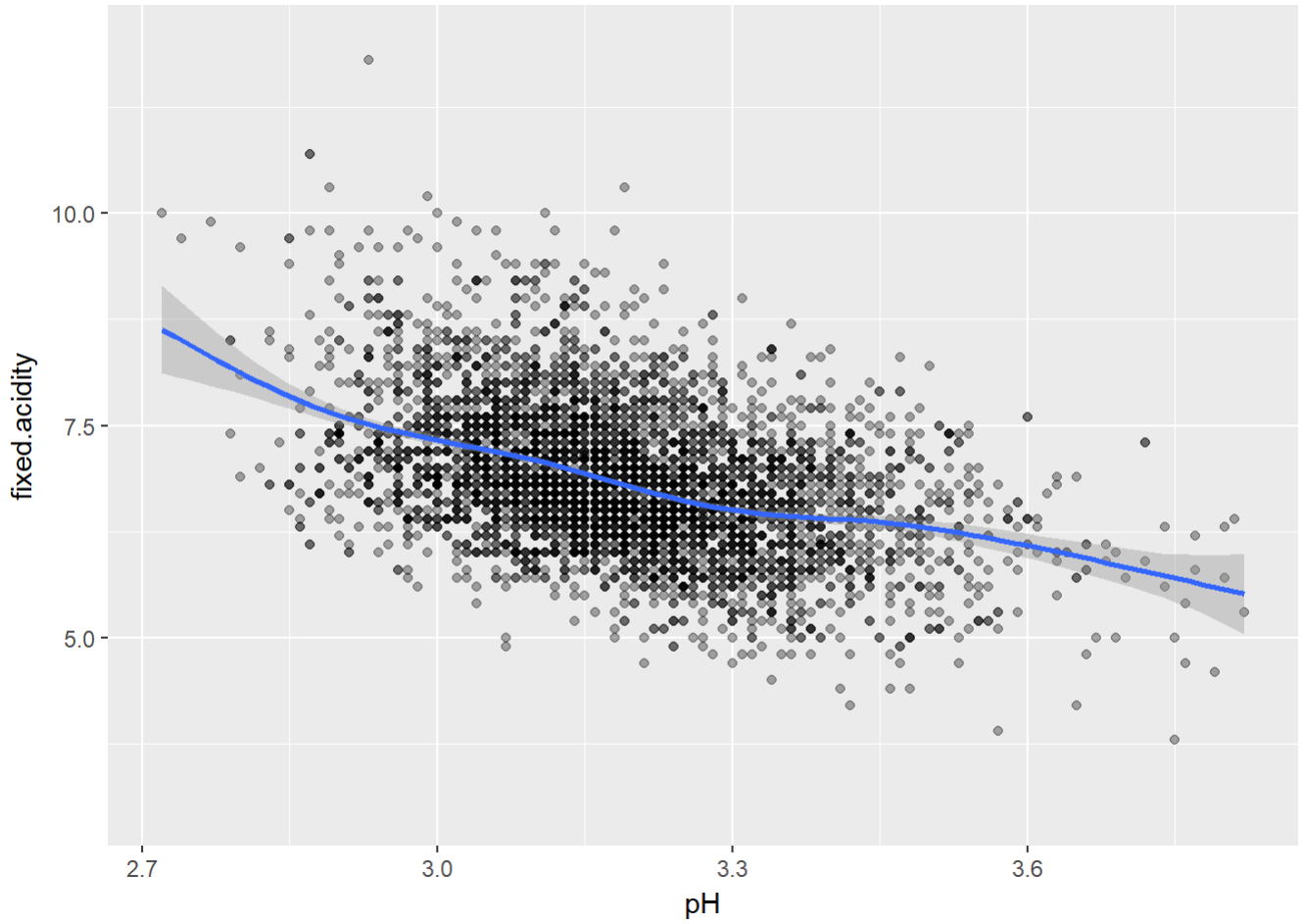
Bivariate Plots Section

I plotted the variables of the dataset to get a quick visualization about which values are relevant in predictiong wine quality. Based on the matrix, presume alcohol content to correlate with quality. Moreover, I also found some other interesing connections worth looking at it. I sat up a 0.95 confidence level and a 0.05 significance and marked the corresponding variables with a black X.



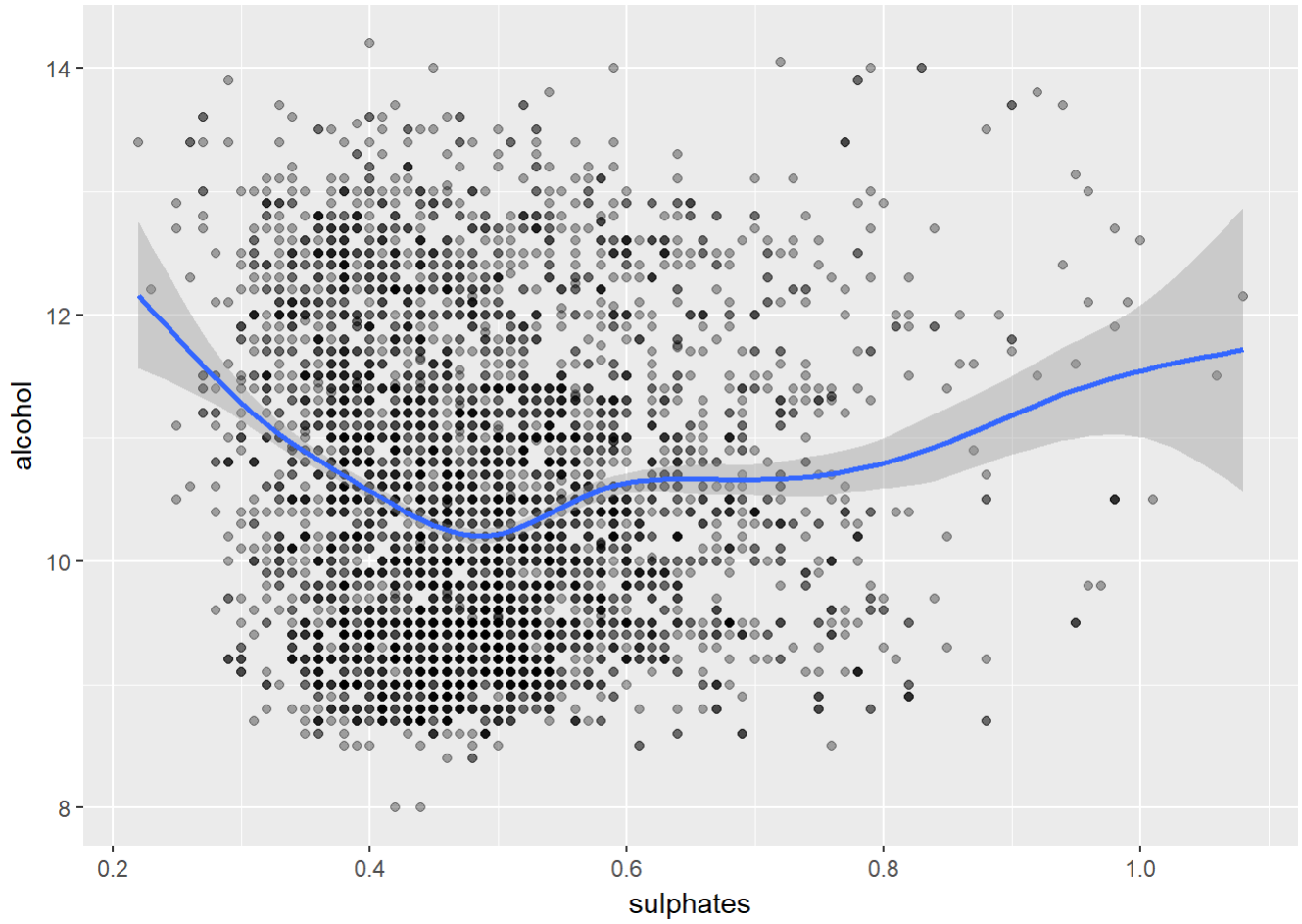
I made a scatter plot diagram about the relationship between fixed acidity and pH. According to the visualization, the lower the fixed acidity, the higher the pH value, what absolutely makes sense. I also counted the correlation between the two variables and found a moderate negative relationship (-0.426) and used stat_smooth() function to represent it.

An article about pH and wine acidity which I used to confirm and better understand my findings: http://winefolly.com/review/understanding-acidity-in-wine/ (http://winefolly.com/review/understanding-acidity-in-wine/)



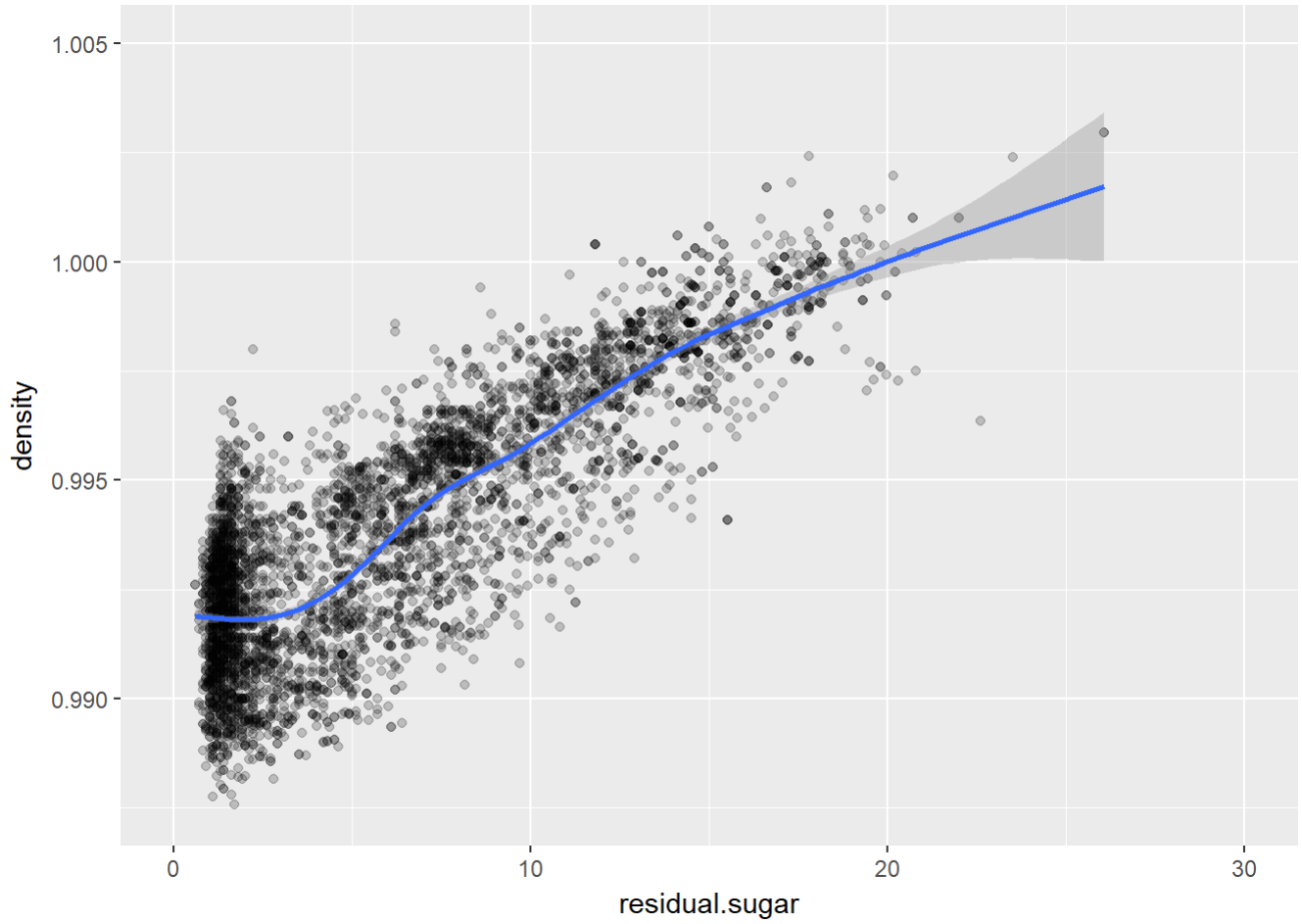

```
## [1] -0.4258583
```

I created a scatter plot about sulphates and alcohol content. I did not found any correlation between the two vales, according to my correlation coefficient number, their correlation is just -0.017.



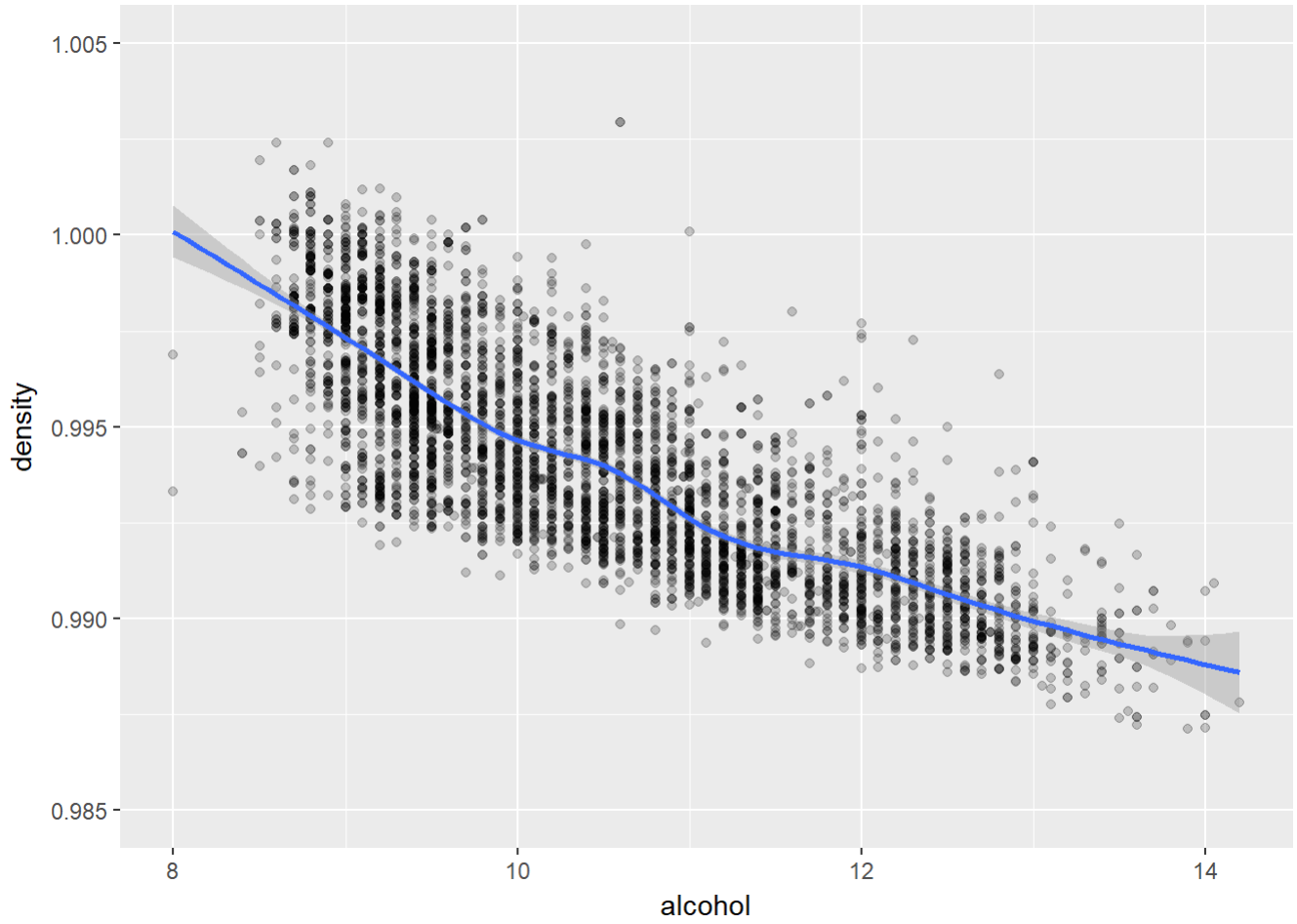

```
## [1] -0.01743277
```

I wondered if residual sugar can influence the density of a wine, so, I plotted the correlation between density and residual sugar and counted the correlation coefficient. I found a strong positive correlation between the two variables (0.839) - as sugar content increases, the density increases.



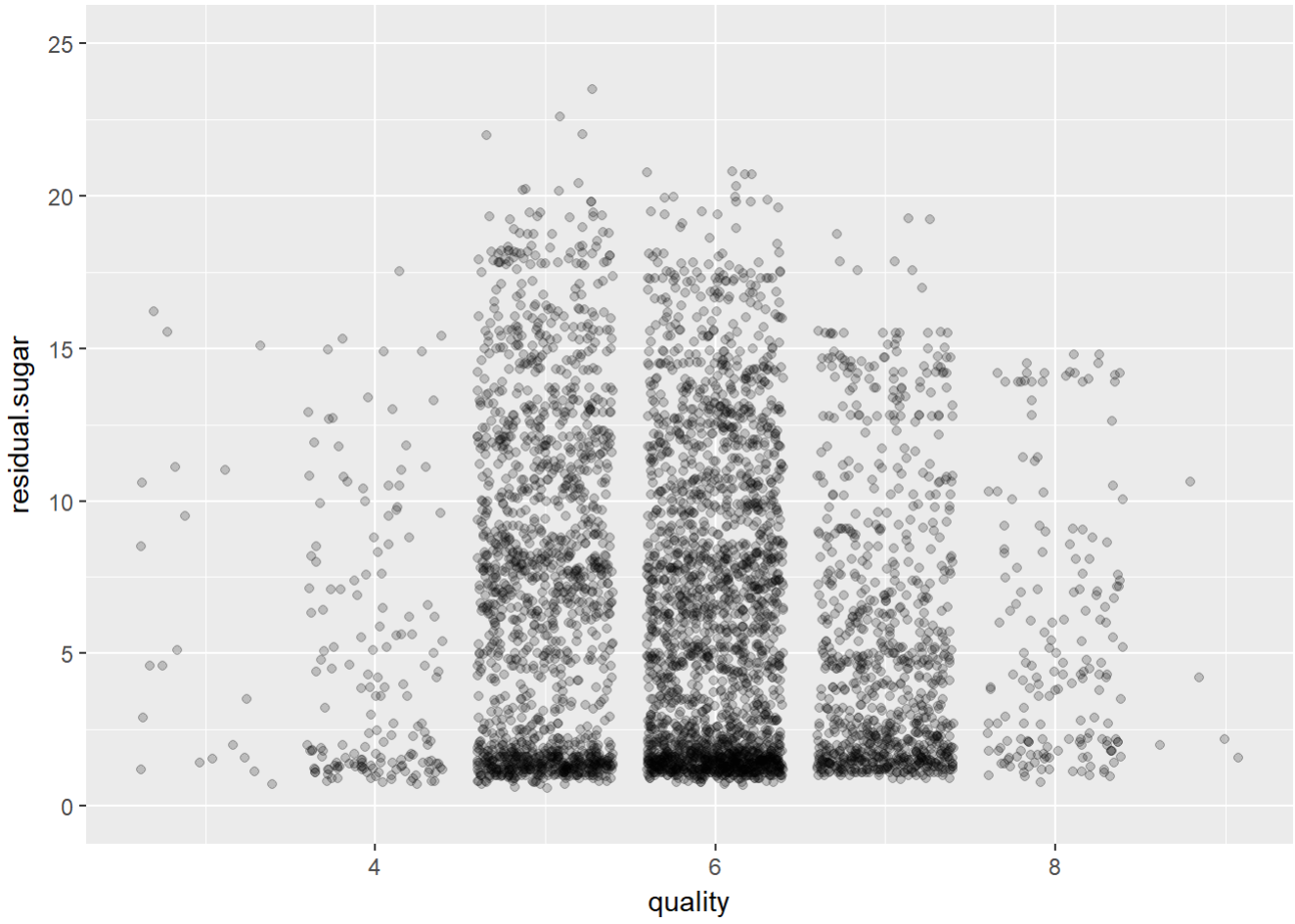

```
## [1] 0.8389665
```

A created a scatter plot about the alcohol content and density and counted the correlation coefficient. According to my results, there is a strong negative correlation between the two component (-0.7801) - the higher the alcohol content is, the lower the density.




```
## [1] -0.7801376
```

I created a jittering scatter plot about the correlation between residual sugar and quality. I did not find any correaltion between the two and the correlation coefficient was also quite week, only -0.0976. According to my results, I can state that there is no evidence that residual sugar would influence quality. This outcome does not match up with my expectations, I thought the best and most prestigous wines are dry and have a lower sugar content on the average.



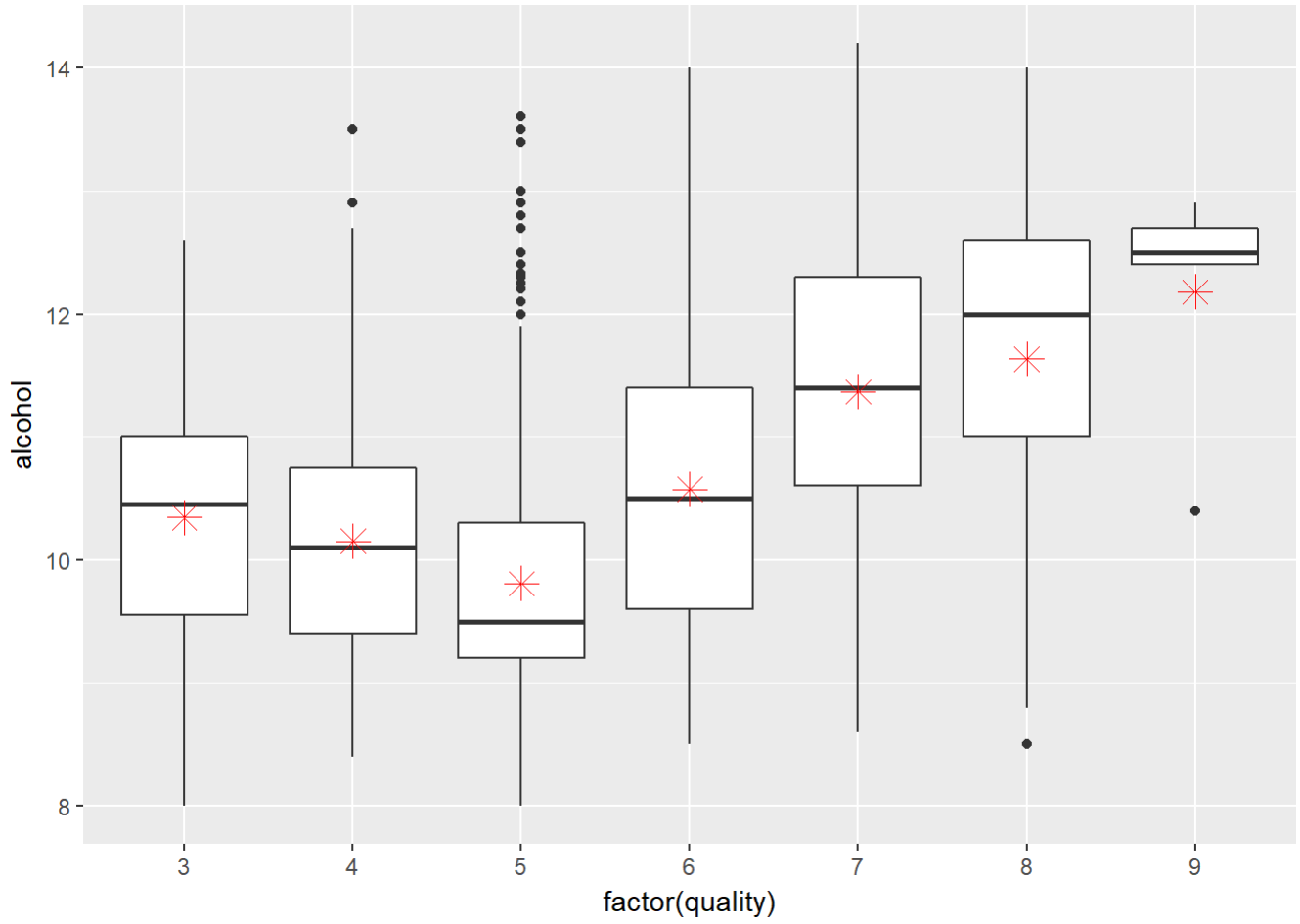

```
## [1] -0.09757683
```

I used the summaryBy function from doBy package to summarize the residual sugar content by quality as I was curious about the more precise mean values.

Source: https://www.rdocumentation.org/packages/doBy/versions/4.5-15/topics/summaryBy (https://www.rdocumentation.org/packages/doBy/versions/4.5-15/topics/summaryBy)

```
##   quality residual.sugar.mean
## 1       3            6.392500
## 2       4            4.628221
## 3       5            7.334969
## 4       6            6.441606
## 5       7            5.186477
## 6       8            5.671429
## 7       9            4.120000
```

I used boxplots to visualize the correlation between quality and alcohol content. I counted the correlation coefficient and got a positive moderate correlation (0.435). The boxplots also show that higher quality wines have higher median alcohol content (above 11 and even 12), while lower quality wines have their median alcohol content around 10. The minimum and maximum values also follow this pattern: the worse wines are their maximum around 12-13 while the betters around 14.



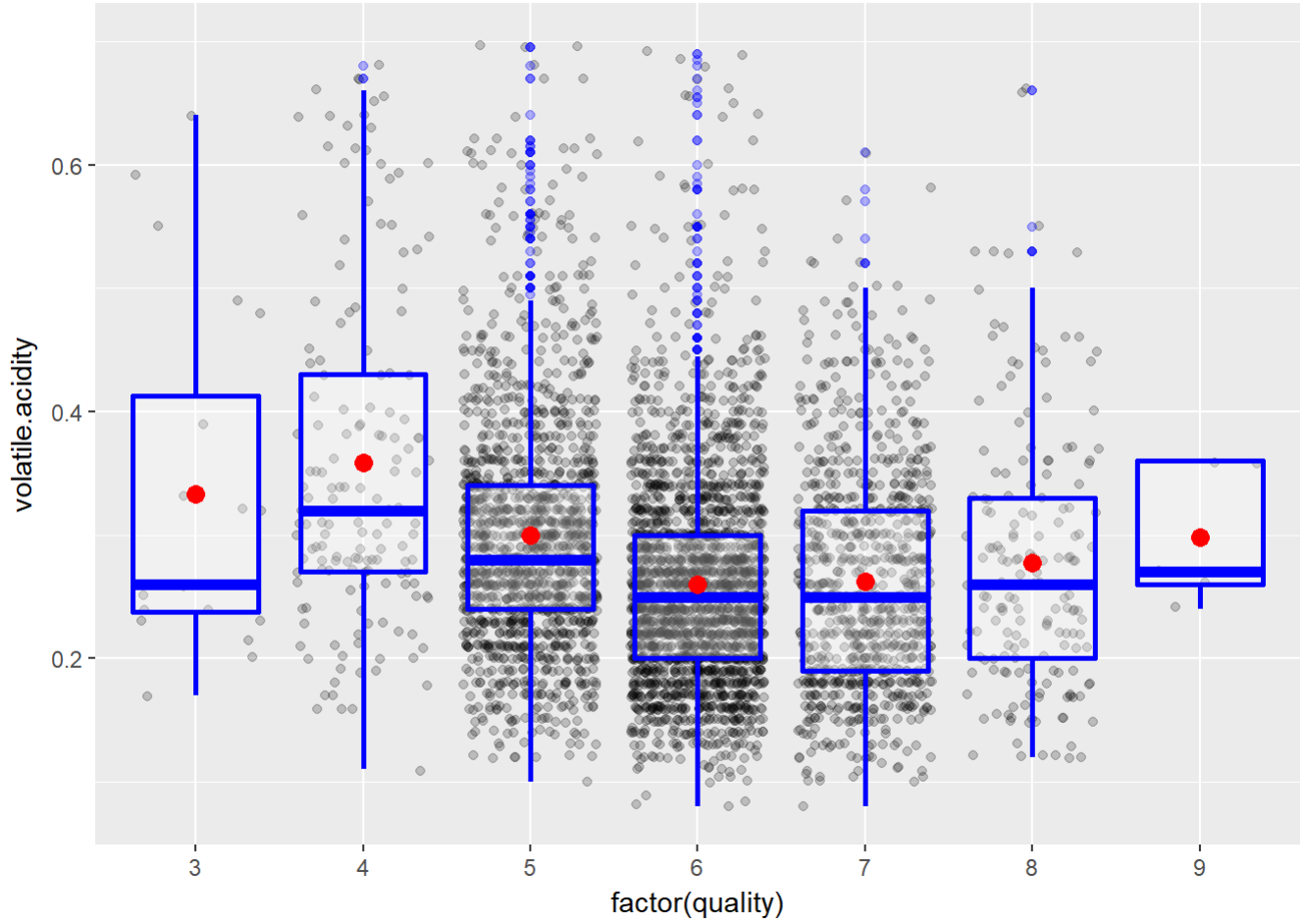

```
##   quality alcohol.mean alcohol.min alcohol.max
## 1       3     10.34500         8.0        12.6
## 2       4     10.15245         8.4        13.5
## 3       5      9.80884         8.0        13.6
## 4       6     10.57537         8.5        14.0
## 5       7     11.36794         8.6        14.2
## 6       8     11.63600         8.5        14.0
## 7       9     12.18000        10.4        12.9
```

```
## [1] 0.4355747
```

```
##   quality alcohol.mean
## 1       3     10.34500
## 2       4     10.15245
## 3       5      9.80884
## 4       6     10.57537
## 5       7     11.36794
## 6       8     11.63600
## 7       9     12.18000
```

I read in the dataset description that too high volatile acidity can lead to an unpleasent, vinegar taste and ruin qualityso I made boxplots about the correlation between volatile acidity and quality.The calculated correlation between volatile acidity and quality does not match up with my expectation. However, if we look at the boxplots,

and compare the mins, maxes and look at the values between the first and third quantile, we can find out that lower quality wines still have a higher volatile acidity content. I also used a jittering plots to visualize the distribution of the values and marked the means with a red dot.

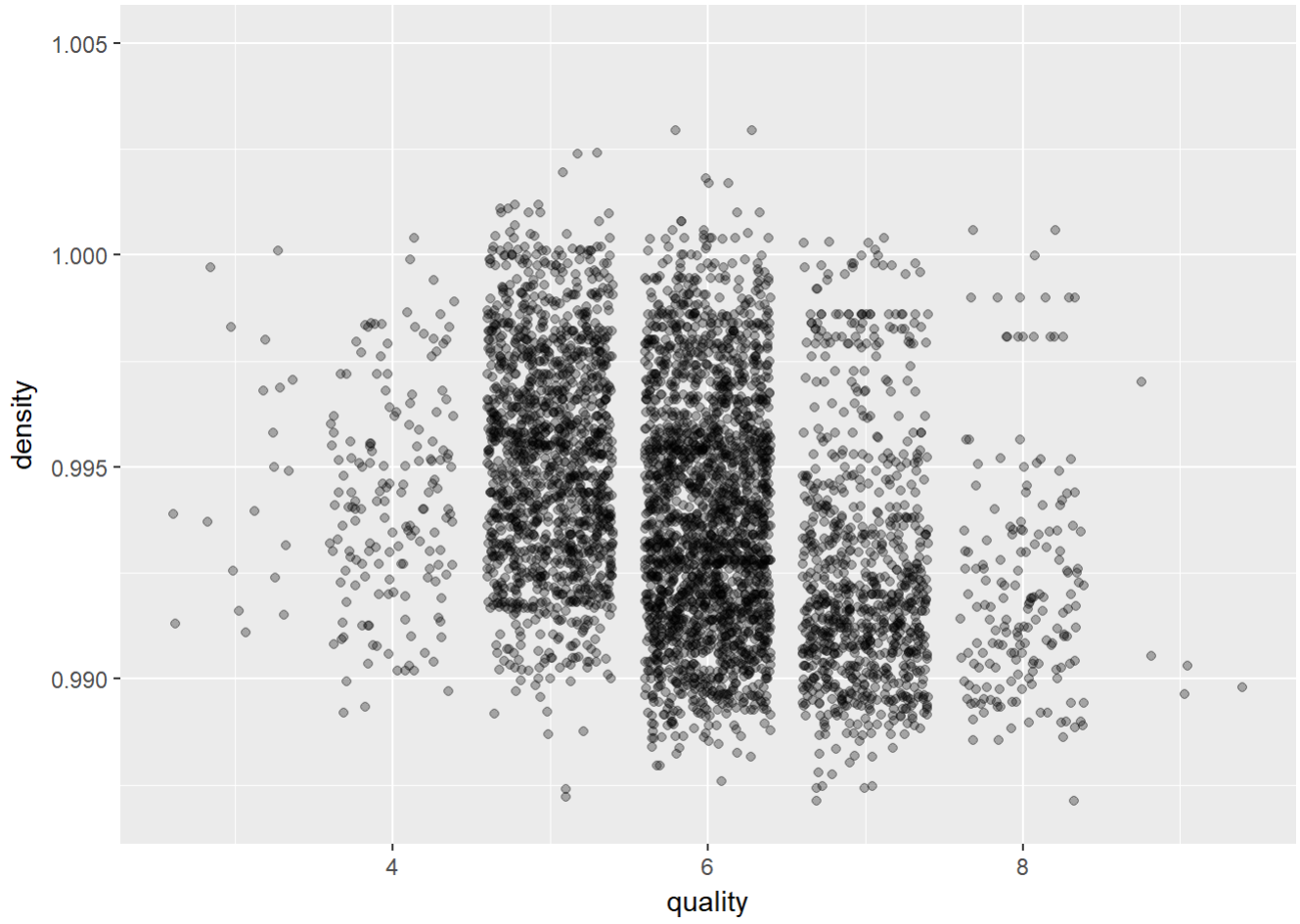



```
## [1] -0.194723
```

```
##   quality volatile.acidity.mean volatile.acidity.min volatile.acidity.max
## 1       3              0.3332500                 0.17                 0.640
## 2       4              0.3812270                 0.11                 1.100
## 3       5              0.3020110                 0.10                 0.905
## 4       6              0.2605641                 0.08                 0.965
## 5       7              0.2627670                 0.08                 0.760
## 6       8              0.2774000                 0.12                 0.660
## 7       9              0.2980000                 0.24                 0.360
```

I plotted the correlation between wine quality and density and found a negative relationship on the edge of weak and moderate categories (-0.3071). However, according to my groupped calculation, the best wines have the lowest mean density.

Source for interpreting correlation coefficient: http://www.dummies.com/education/math/statistics/how-to-interpret-a-correlation-coefficient-r/ (http://www.dummies.com/education/math/statistics/how-to-interpret-a-correlation-coefficient-r/)




```
## [1] -0.3071233
```

```
##   quality density.mean density.min density.max
## 1       3    0.9948840     0.99110     1.00010
## 2       4    0.9942767     0.98920     1.00040
## 3       5    0.9952626     0.98722     1.00241
## 4       6    0.9939613     0.98758     1.03898
## 5       7    0.9924524     0.98711     1.00040
## 6       8    0.9922359     0.98713     1.00060
## 7       9    0.9914600     0.98965     0.99700
```

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?

My main feature of the interest was the quality of the wine, I examined which components can influence the wine quality in the dataset and how strong is the correlation between quality and other features. I found out that alcohol content and density can play a lot more important role in quality.

- Residual sugar does not correlate with wine quality at all. The mean values vary between 4 and 7.5 and the best wines have the lowest mean sugar content (4.12). Hovewer, we can not predict quality based on only the residual sugar.
- The variation of mean alcohol content among the different qualities is about 2%. The best wine have the highest mean alcohol content (12.18). The top three wine have their mean alcohol above 11 while the worst above 11.
- The variation among mean density values is less than 0.01. There is negative 0.307 correlation between quality and density that I would rather interpret as negligible and weak.
- Volatile acidity has the strongest correlation with wine quality among other acidity indicators, the worst wines have a higher volatile acidity proportion.

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

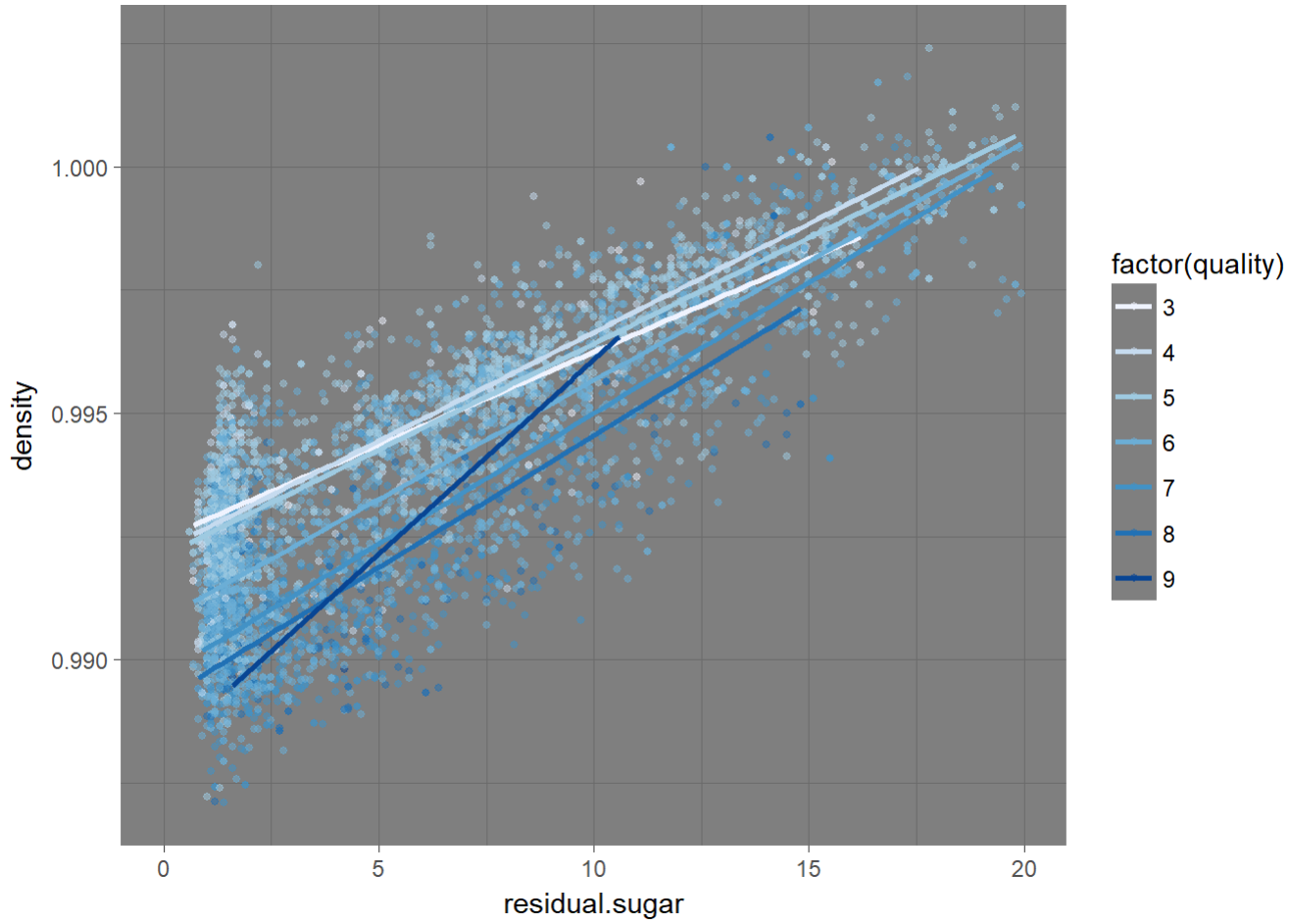
I observed a strong correlation between residual sugar and density, indicating that wines with more density have more residual sugar content. I also found interesting the strong negative correlation between density and alcohol content - the more alcohol the wine contains the lower the density is. As a non-professional, I would expect quite the opposite.

What was the strongest relationship you found?

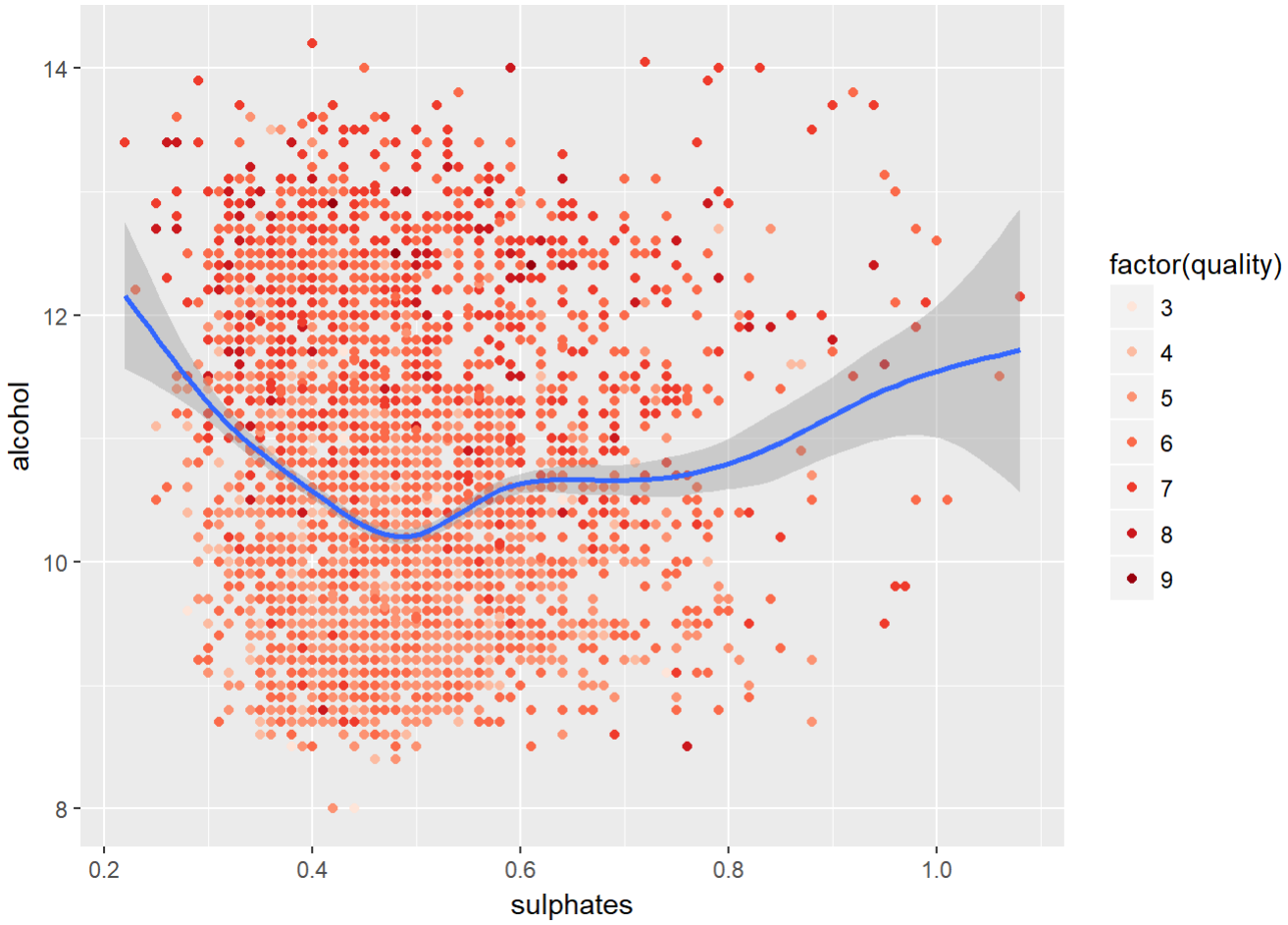
The strongest relationship I found was between alcohol content and residual sugar, it is 0.8389.

Multivariate Plots Section

I created a scatterplot to investigate the interactions between residual sugar, density. I grouped and colored the dots by quality to show the correlation between density and quality, too. As a result, we got a strong positive correlatedscatter plot where dots, representing the better wines can be found on the bottom of the plot. I also used regression lines and dark theme to make the trends easier to understand.



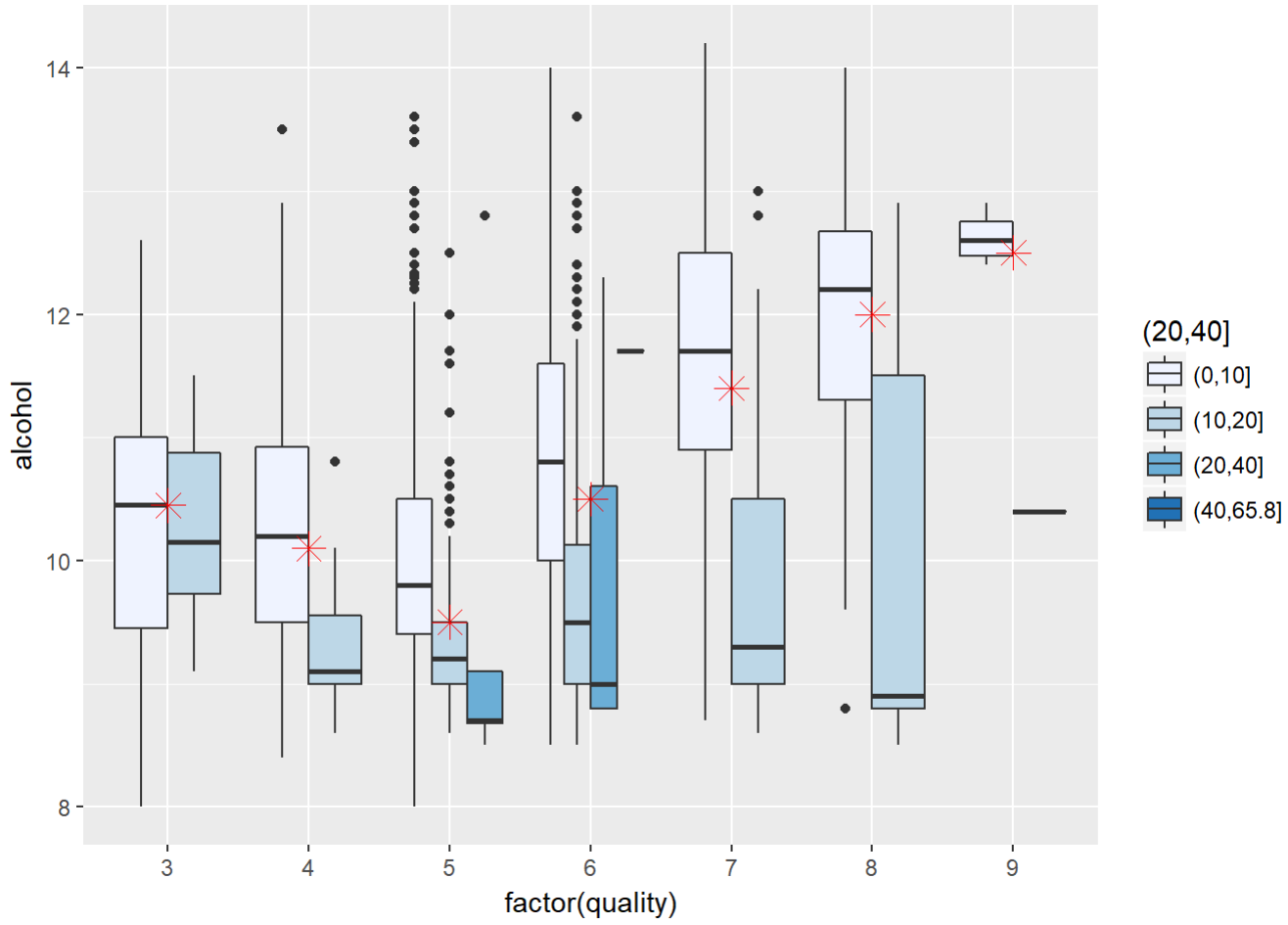
I make a plot about the possible trend between pH and sulphates, colored by quality. I used the stat_smooth() function to stress the interesting trend what can be even interpreted as nonlinear. The lower and higher suplhate content correlates with relative high alcohol, but the average suplhate content indicates lower alcohol proportion. If we look at the colors, we can also notice the strong positive trend of distribution of quality.



I created a box plot to visualize the trend between quality, alcohol and residual sugar. I made buckets to mark and separate the different levels of residual sugar proportions and get a better readable visualization. We can observe the variation on alcohol content by quality: the better the wine the higher the median alcohol content is for each level of quality (the medians are marked by a red star).

The opposite goes for the residual sugar, as wines with higher alcohol content contain less sugar. This tendency also reflects on one important part of wine making, as “residual sugar is the level of glucose and fructose (grape sugars) that are not converted into alcohol during fermentation”.

Source: http://winefolly.com/tutorial/wines-from-dry-to-sweet-chart/ (http://winefolly.com/tutorial/wines-from-dry-to-sweet-chart/)



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?

Alcohol content, residual sugar and density on Plot 1 strengthened each other as I started using color to hightlight some aspects of the data. This is also true in case of Plot 3, where I compared alcohol content and residual sugar to quality.

Were there any interesting or surprising interactions between features?

I found really interesting the interaction between alcohol an residual sugar content and its reason. I never thought about this part of wine making before. I also found an unusual, non linear-looking correlation between alcohol and sulphates. As a non-professional, I made a little research to find out if there are any information about their relationship, but I have not found anything yet, it can also be a random exception. Suplhates are not involved in wine making naturally, but some wine maker use it to correct mineral deficiencies.

Source: http://www.foodsmatter.com/allergy_intolerance/sulphites/articles/sulphates_sulphites.html (http://www.foodsmatter.com/allergy_intolerance/sulphites/articles/sulphates_sulphites.html)

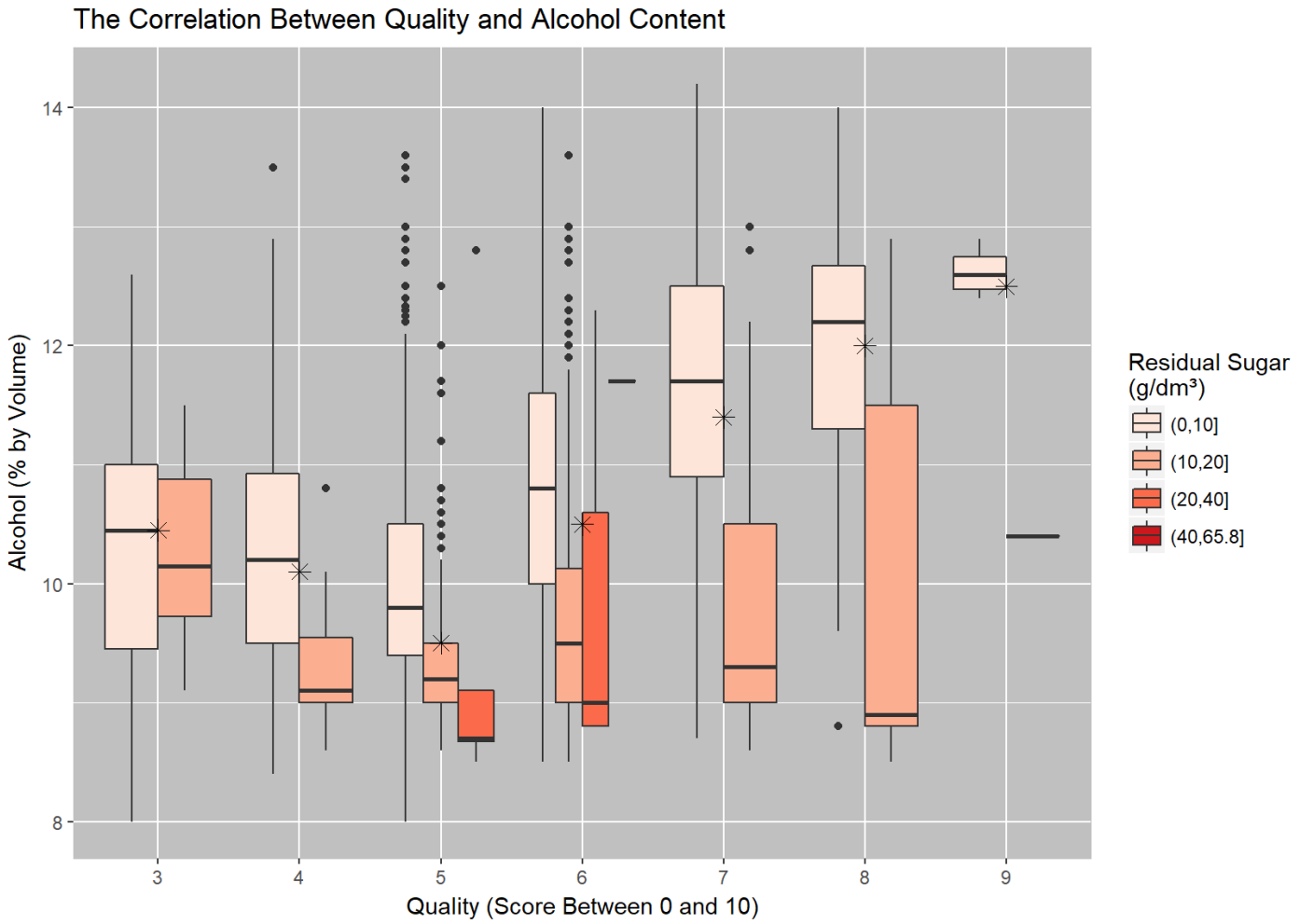
Final Plots and Summary

Plot One

Description I:

The visualisation shows the correlation between volatile acidity and quality. We can not necessarily observe a linear trend among the quality ranges but the worst wines tend to have a higher volatile acidity, what absolutely makes sense: wines with higher volatile acidity could taste bad. If we compare the values between the first and third quantile we can see that the worst wines (quality 3 and 4) have higher volatile acidity while the better ones (quality 5-7) have lower values. This trend does not corresponds to the best wines (quality 8 and 9).

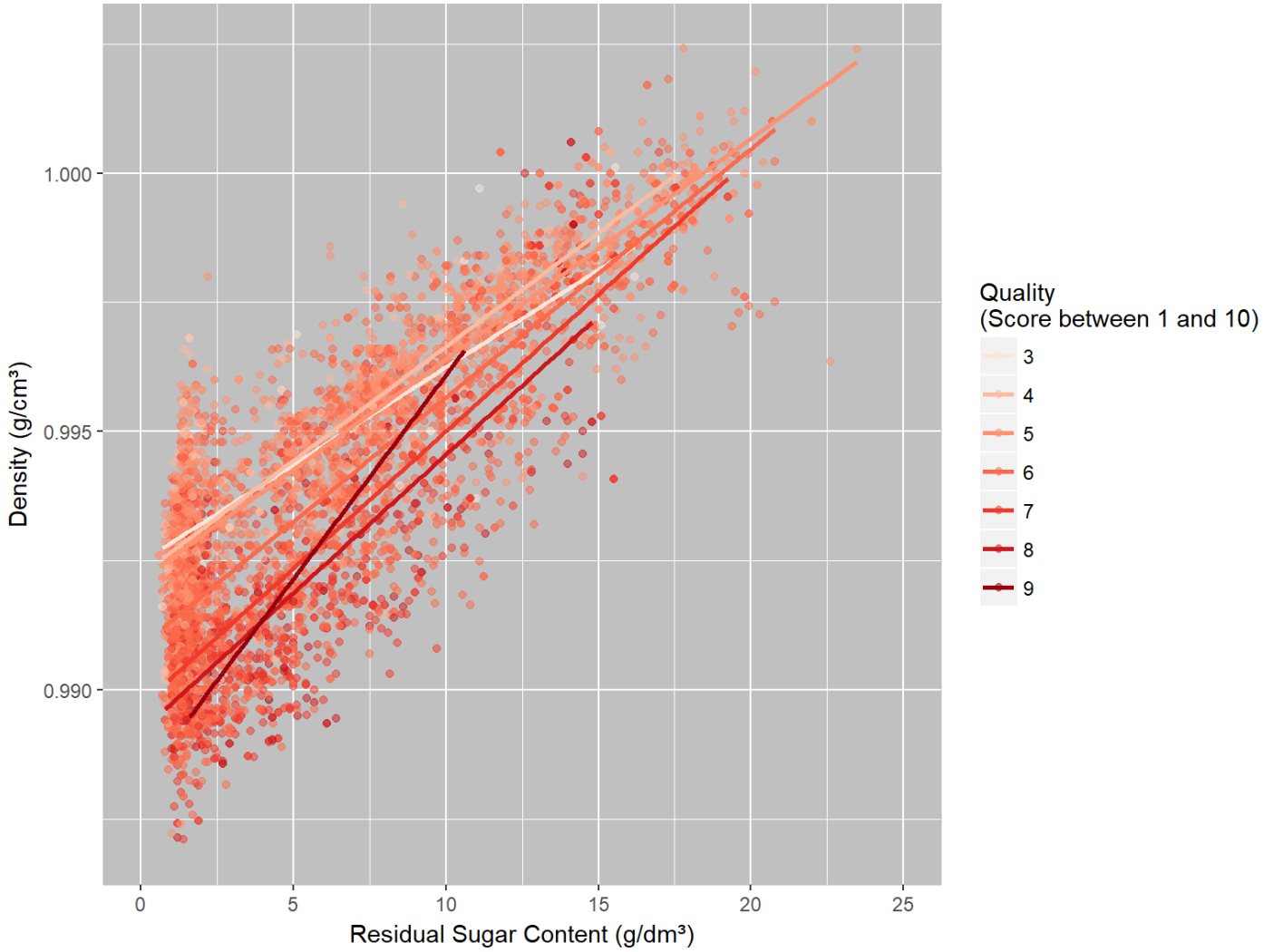
Plot Two



Description II:

I used the previously bucketed values and box plots to show the most important factor which influence wine quality - alcohol. We can also get a more detailed perspective about the correlation, thanks to the categorized correlation lines. The better wines have a higher alcohol content and lower residual sugar. Previously we found that residual sugar does not correlate directly with wine quality, however, sugar and alcohol have a strong correlation and alcohol indicates wine quality very strongly.

Plot Three



Description III:

Residual sugar and density are highly correlated. Quality -marked by color- also tend to correlate with density. The lower the quality is, the higher the density.

Reflection

My main feature of interest was wine quality and I could find some variables which tend to influence this factor, like alcohol content and density. In the first part I made some simple hitograms about the main factors and other variables as well. In the part of bivariate analysis, I tried to get a better knowledge on the factors which influence wine quality. I was suprised that residual sugar does not correlate with wine quality. I also could make some other interesting explorations which bring further questions and could be analyzed in the future, especially the relationship of sulphates and alcohol. In the third part I highlighted some of my main findings, and reflected the main components that influence quality.

The project work took me more time than I previously estimated: one of the reason is that I forgot about an important rule: the devil is in the details. It took time to choose the best corresponding plots and to detect the variables that influence quality. I had to examine each variable separatly while not forgetting the whole picture which I found really challenging and exciting.