WineQualityReds by Ariane Broquet

Introduction

This clean dataset contains 1,599 red wines with 12 variables on the chemical properties of the wine. At least 3 wine experts rated the quality of each wine, providing a rating between 0 (very bad) and 10 (very excellent). Wines contain several chemicals, such as citric acid, chlorides, sulfur and so on. This project will focus on the following question: which chemical properties influence the quality of red wines?

Univariate Plots Section

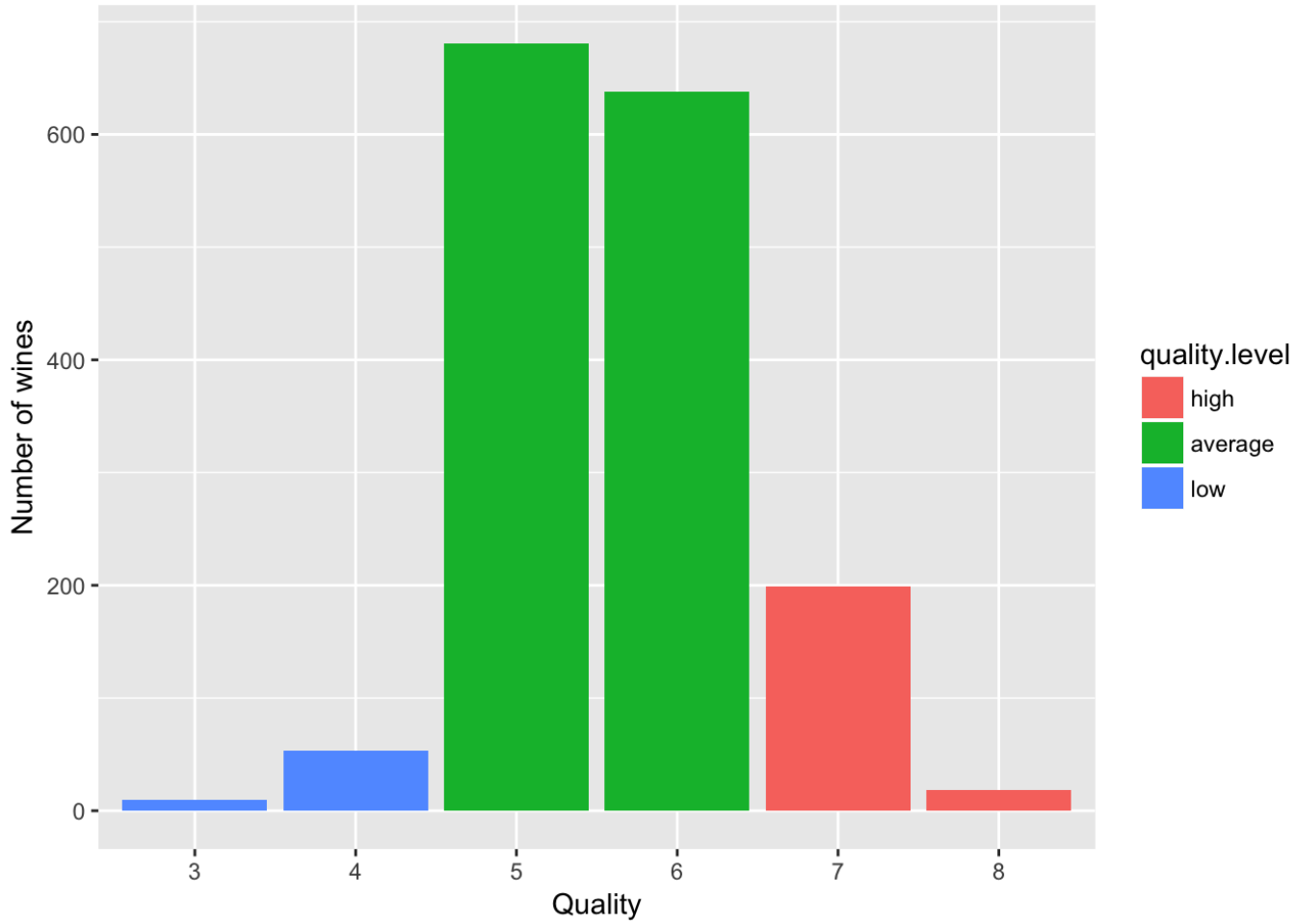
For a short overview of the dataset:

```
##        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
```

```
## 'data.frame':    1599 obs. of  12 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0
.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 .
..
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8
...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

The table above shows the dataset of 1599 observations divided by 12 variables. Each variable has its individual measurement; they show the chemical content of wines, their density, PH, as well as their quality. The variables have the following impacts on wine:

- Fixed acidity: fundamental property of wine, imparting sourness and resistance to microbial infection.
- Volatile acidity: natural part of the fermenting process and can ruin a wine. It’s basically the process of wine turning into vinegar.
- Citric acid: very common in citrus fruits, such as limes, citric acid is found only in very minute quantities in wine grapes.
- Residual sugar: any natural grape sugars that are leftover after fermentation ceases. The juice of wine grapes starts out intensely sweet, and fermentation uses up that sugar as the yeasts feast upon it.
- Chlorides: might give the wine a salty flavor, which may turn way potential consumers.
- Free sulfur dioxide: preventing microbial growth and the oxidation of wine.
- Total sulfur dioxide: works as an antioxidant
- Density: volume to weight.
- PH: scale that measures the concentration of free hydrogen ions floating around in your wine. It is a measurement of how strong an acid is.
- Sulphates: preservative that’s widely used in winemaking
- Alcohol: alcohol directly correlates to the ripeness (i.e., sugar content) of grapes.
- Quality: experts tasted 1,599 wines and rated them from 0 (very bad) to 10 (very excellent).

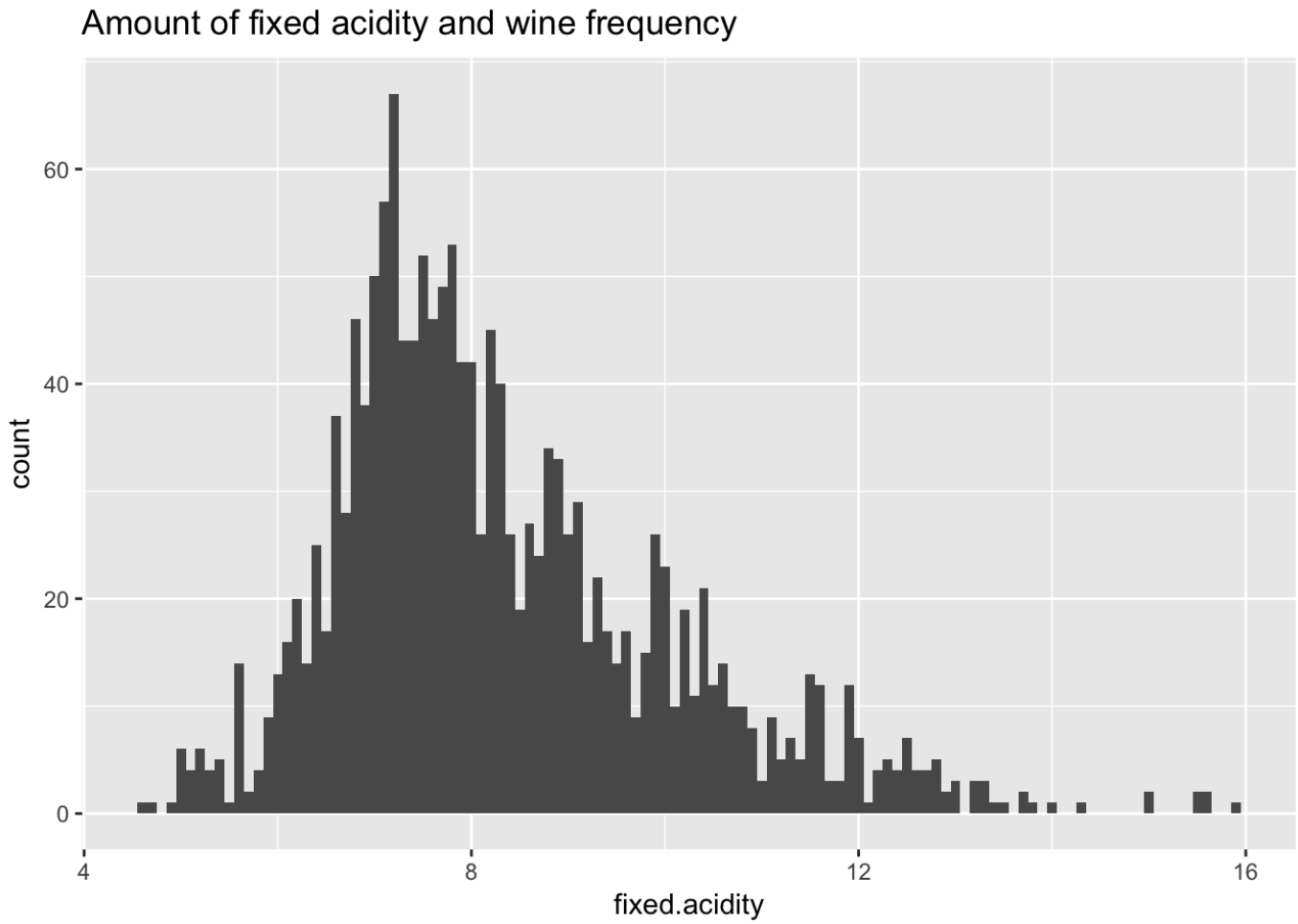


The graph above shows that red wine quality is normally distributed and concentrated around 5 and 6. The mode is 5.

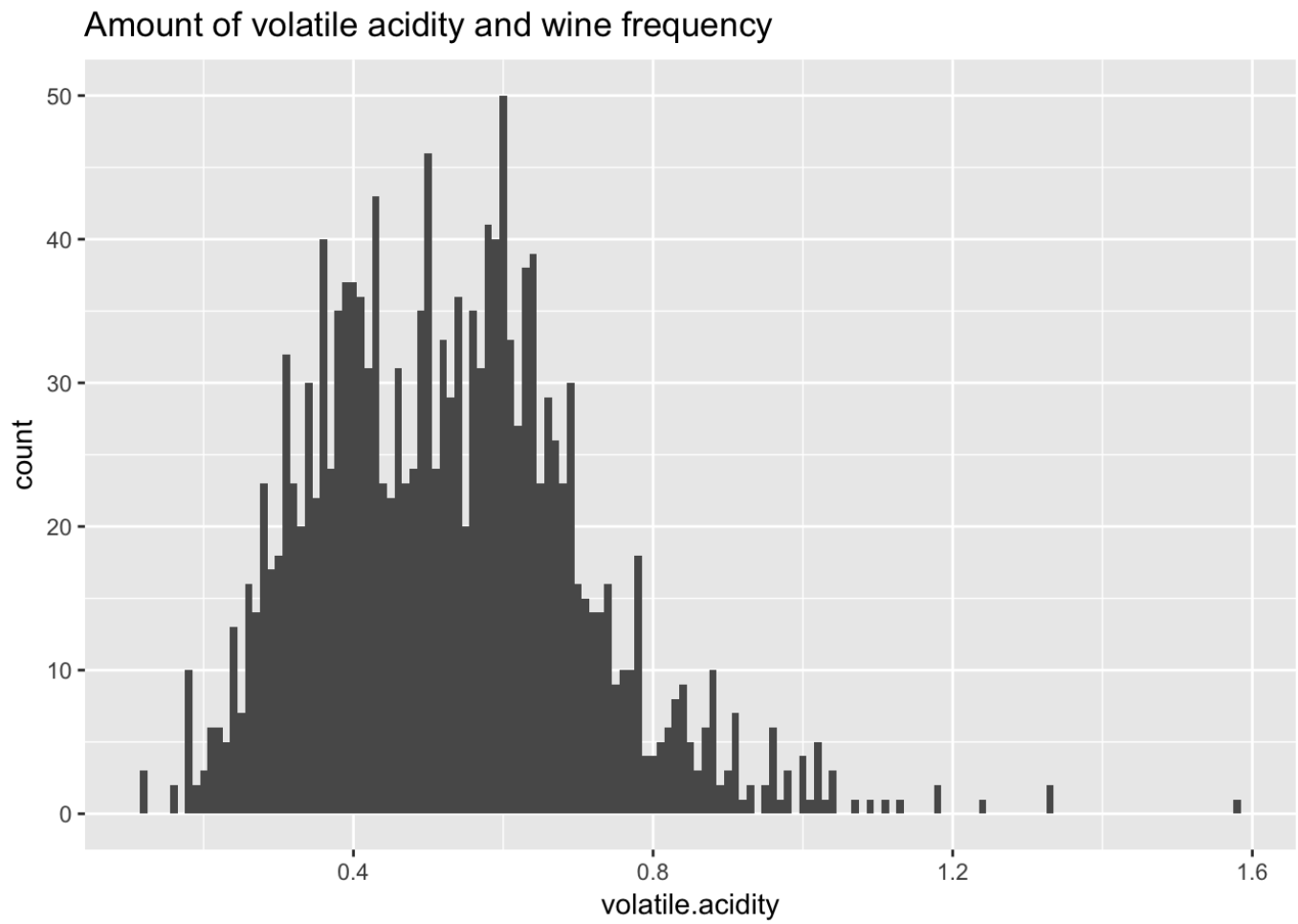
```
##
## ------------------------------------------------------------------
##  fixed.acidity   volatile.acidity    citric.acid    residual.sugar
## --------------- ------------------ --------------- ----------------
##  Min.   : 4.60     Min.   :0.1200     Min.   :0.000    Min.   : 0.900
##
##  1st Qu.: 7.10    1st Qu.:0.3900    1st Qu.:0.090   1st Qu.: 1.900
##
##  Median : 7.90    Median :0.5200    Median :0.260   Median : 2.200
##
##   Mean : 8.32      Mean :0.5278      Mean :0.271     Mean : 2.539
##
##  3rd Qu.: 9.20    3rd Qu.:0.6400    3rd Qu.:0.420   3rd Qu.: 2.600
##
##  Max.   :15.90     Max.   :1.5800     Max.   :1.000    Max.   :15.500
## ------------------------------------------------------------------
##
## Table: Dataset summary (continued below)
##
##
## ------------------------------------------------------------------------
-
##     chlorides      free.sulfur.dioxide   total.sulfur.dioxide      density
## ----------------- --------------------- ---------------------- ---------------
-
##  Min.   :0.01200       Min.   : 1.00          Min.   : 6.00       Min.   :0.9901
```

```
##
##  1st Qu.:0.07000      1st Qu.: 7.00         1st Qu.: 22.00      1st Qu.:0.9956
##
##  Median :0.07900      Median :14.00         Median : 38.00      Median :0.9968
##
##   Mean :0.08747        Mean :15.87           Mean : 46.47        Mean :0.9967
##
##  3rd Qu.:0.09000      3rd Qu.:21.00         3rd Qu.: 62.00      3rd Qu.:0.9978
##
##  Max.  :0.61100       Max.  :72.00          Max.  :289.00       Max.  :1.0037
## ------------------------------------------------------------------------------
-
##
## Table: Table continues below
##
##
## ------------------------------------------------------------------------------
--
##       pH           sulphates         alcohol         quality      quality.level

## --------------- ---------------- --------------- --------------- -------------
--
##  Min.  :2.740    Min.  :0.3300    Min.  : 8.40    Min.  :3.000     high : 217

##
##  1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000   average:1319
##
##  Median :3.310   Median :0.6200   Median :10.20   Median :6.000     low : 63

##
##   Mean :3.311     Mean :0.6581     Mean :10.42     Mean :5.636         NA

##
##  3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000        NA

##
##  Max.  :4.010    Max.  :2.0000    Max.  :14.90    Max.  :8.000         NA

## ------------------------------------------------------------------------------
--
```

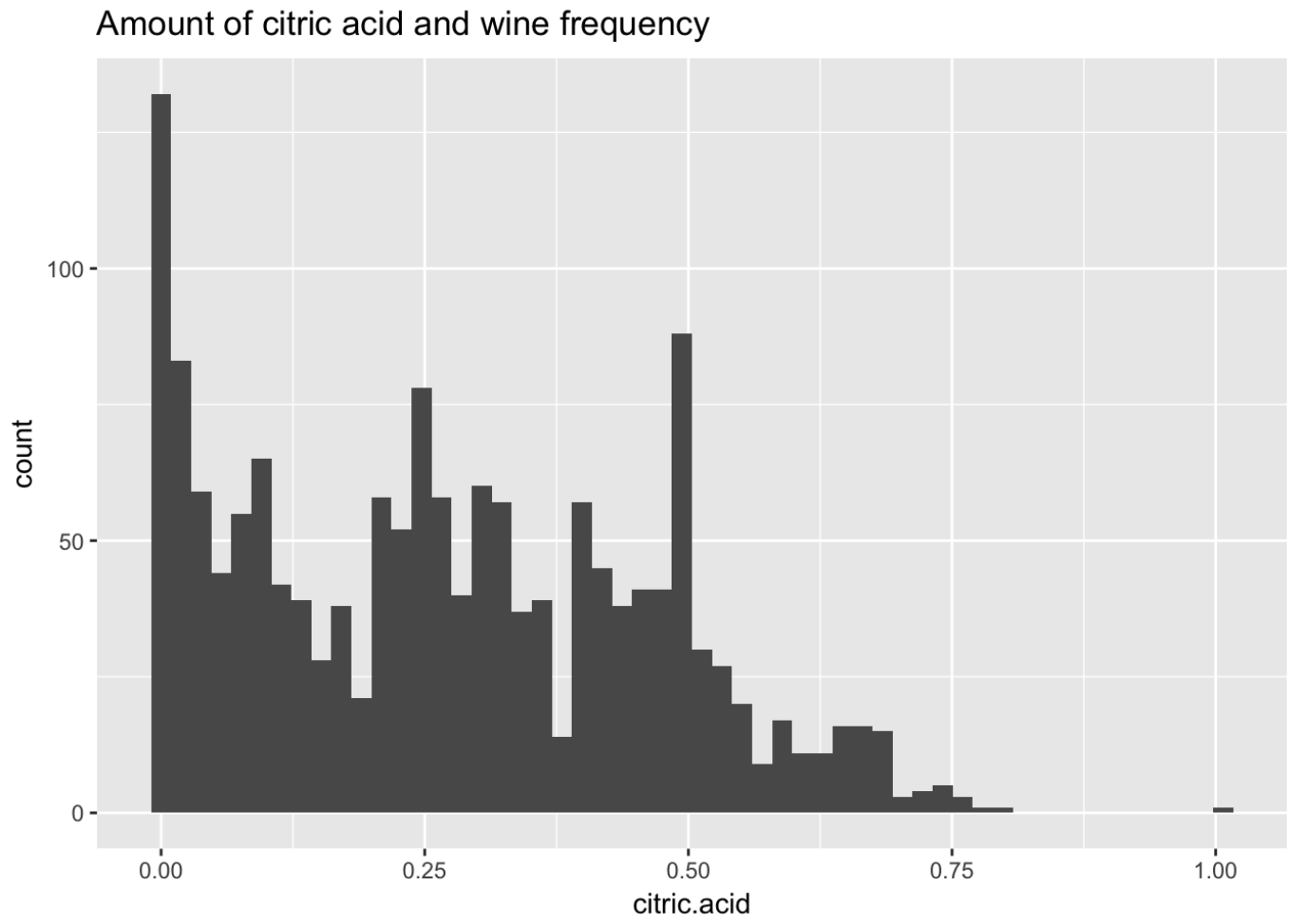
The tables above show deeper information for each variable, such as their median, mean, minimal and maximal value. It can be observed that 217 wines are considered to have a high quality, 1319 to be average and 63 to be low. Individual graphs will be created to observe the distribution shape of each variable.



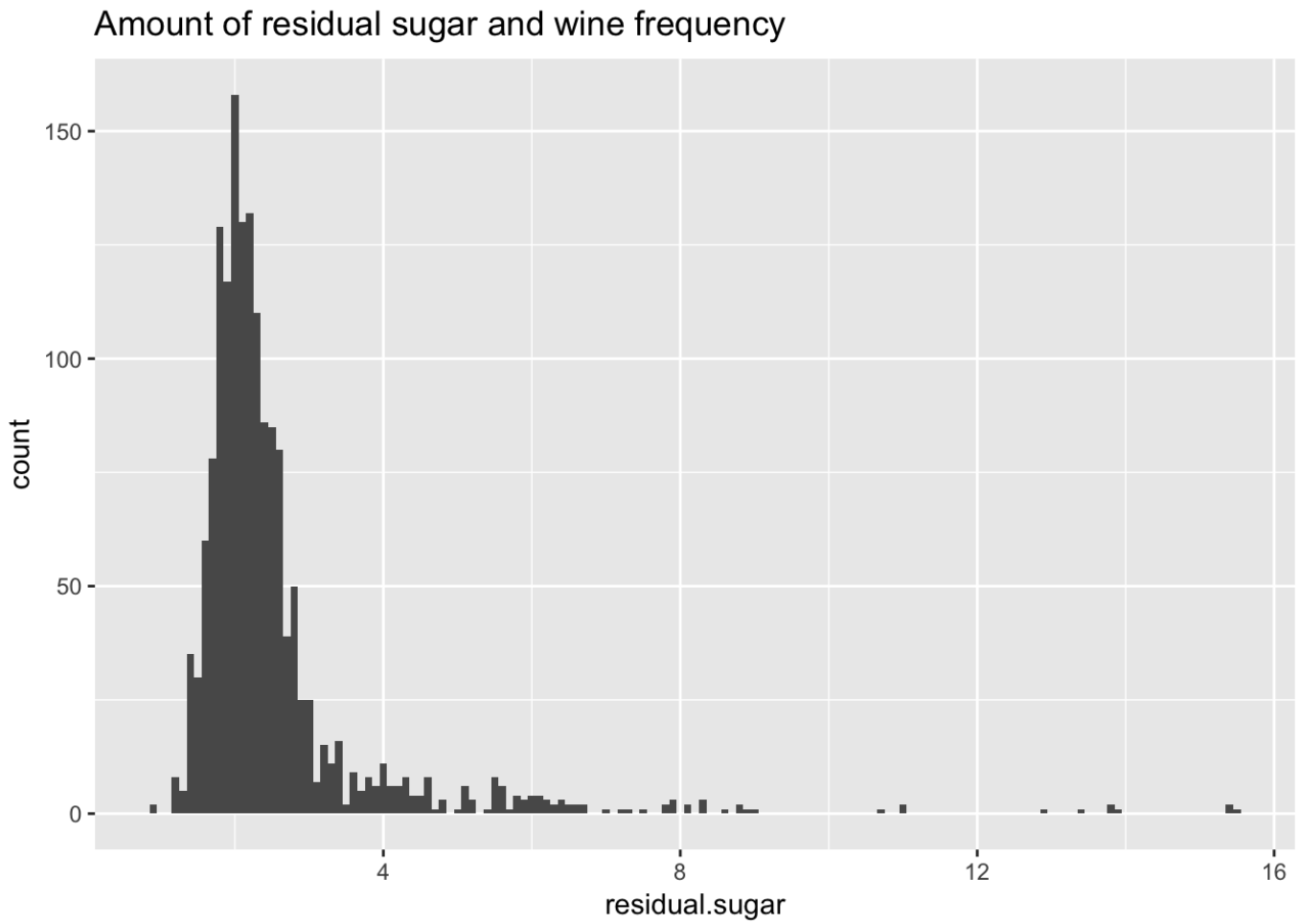
There is a high concentration of wines with fixed acidity at 7.9 (the median). The mean is 8.32, which is higher than the median because there are outliers around the 16th value. The fixed acidity shows a normal distribution. The distribution is a bit skewed, with extreme values upside.



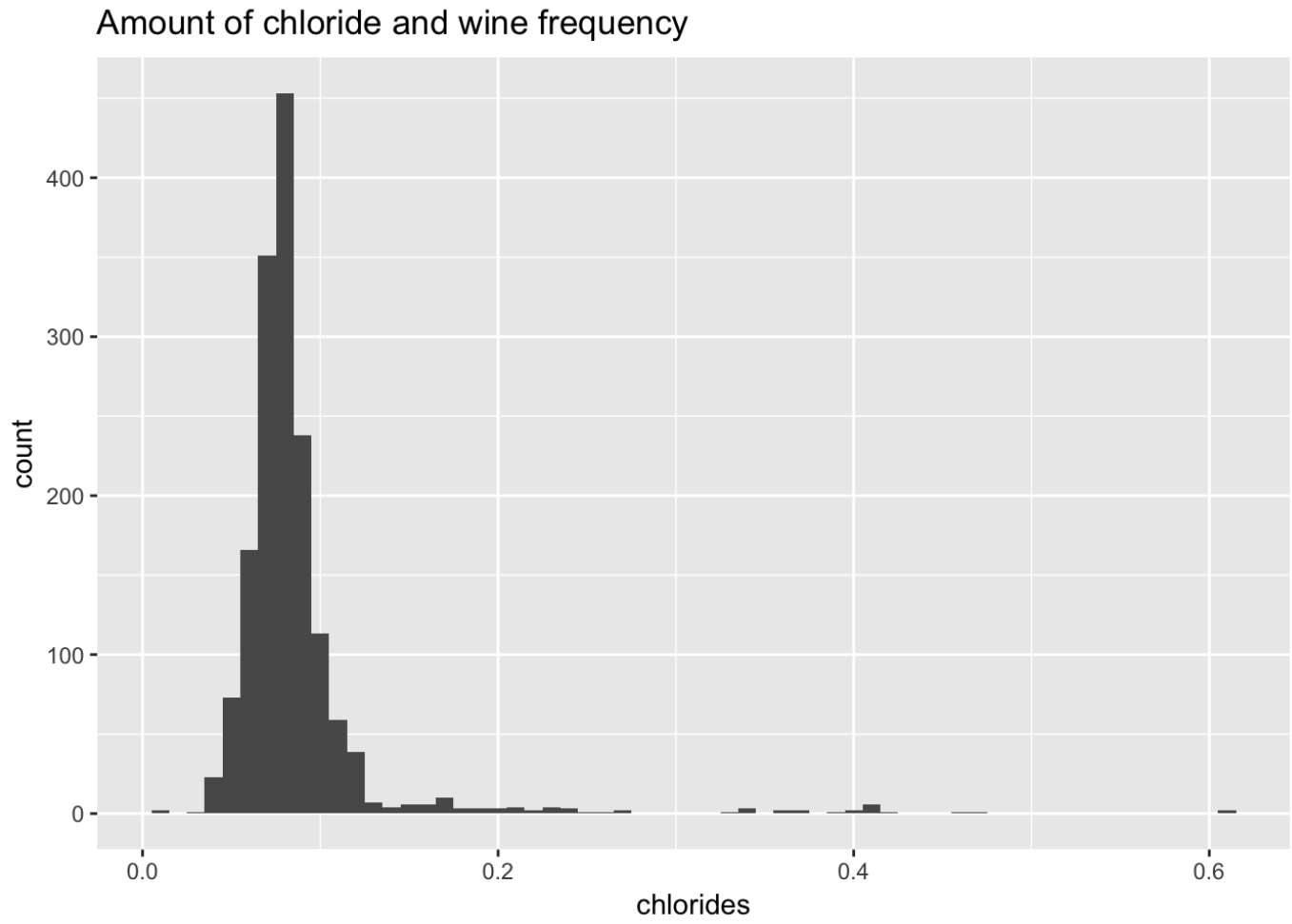
The distribution appears to be bimodal around 0.4 and 0.7 with some outliers around 1.6.



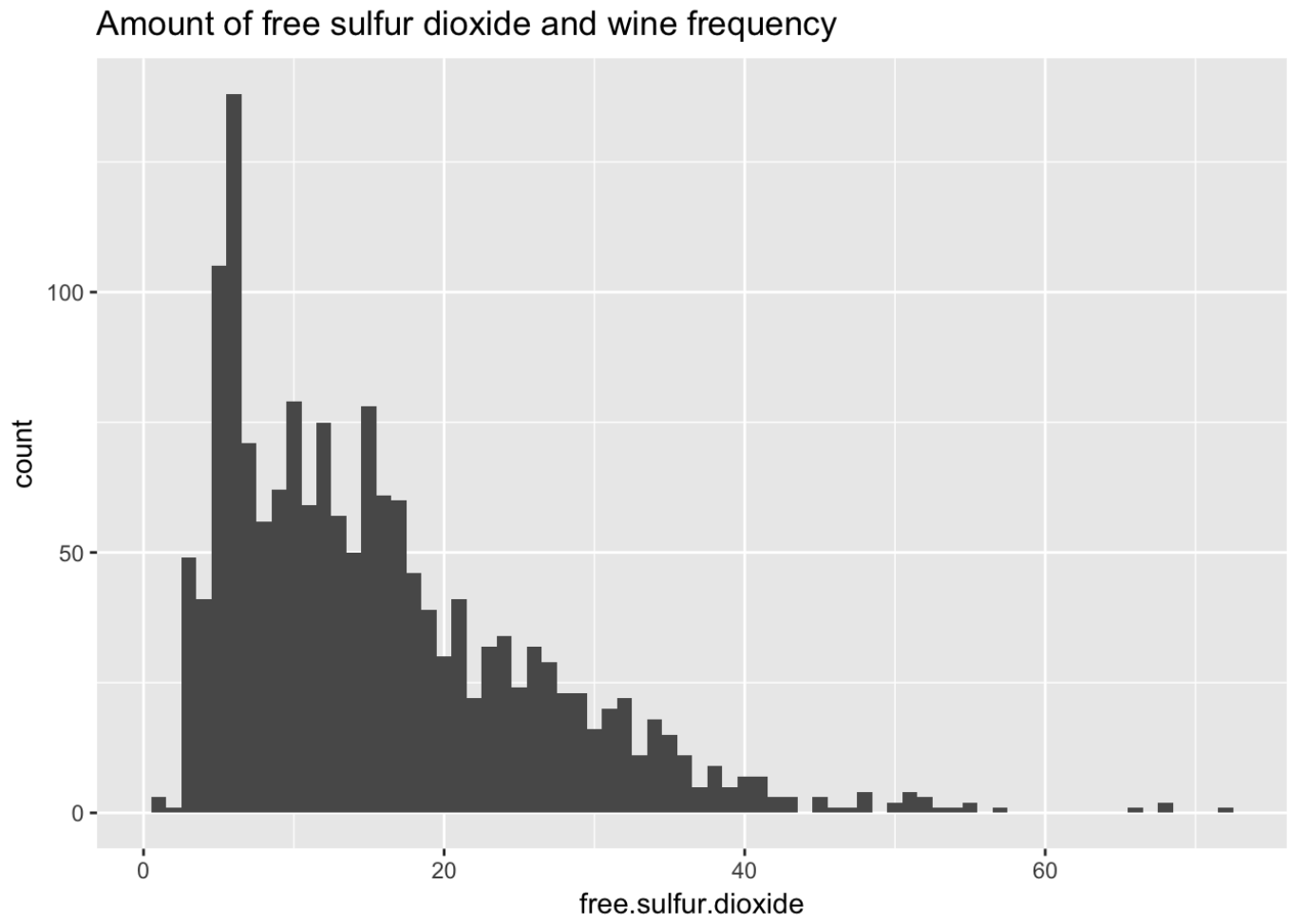
The median is 0.26 and the mean is 0.271. Most wines do not contain citric acid.



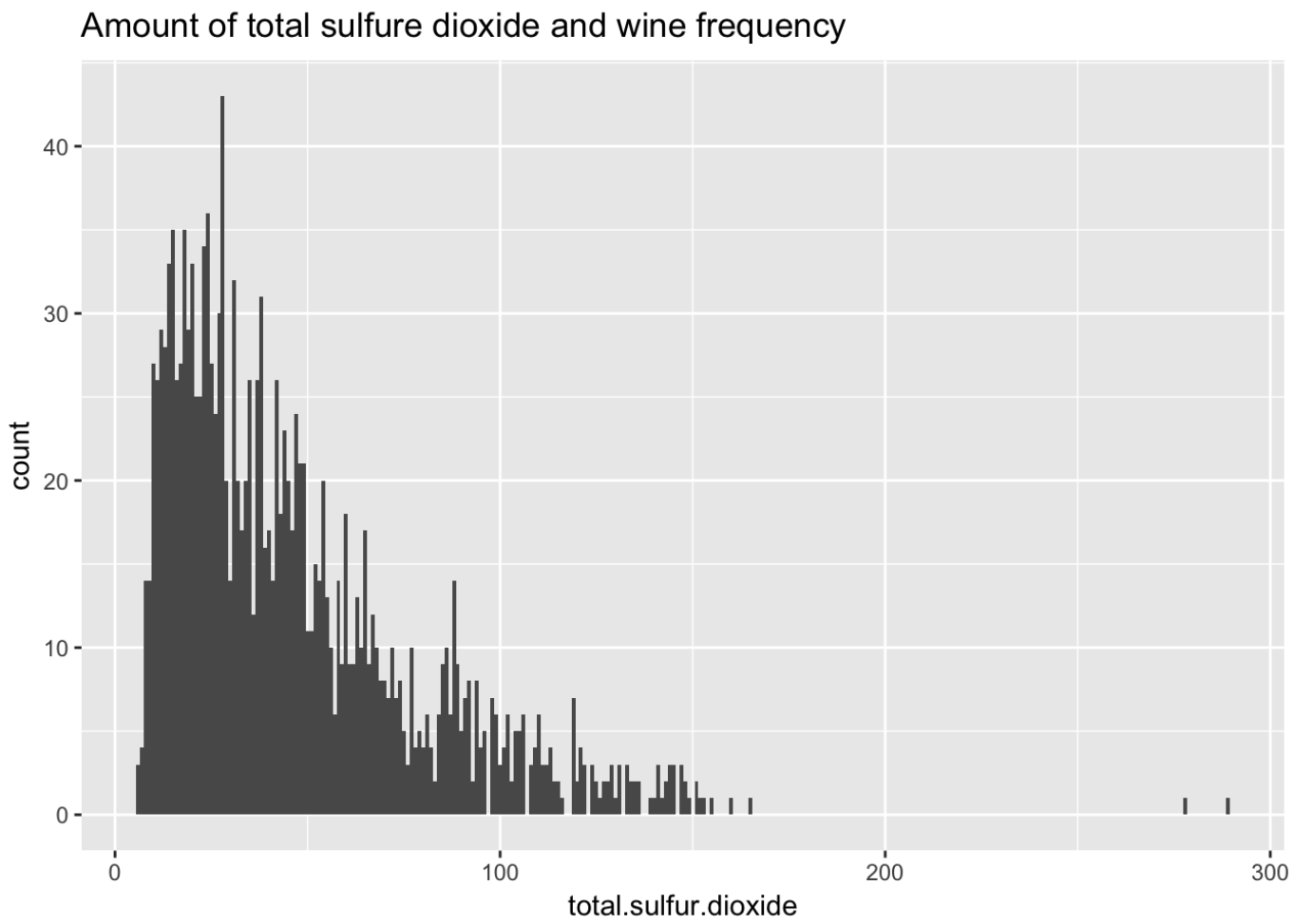
A high concentration of wines contain around 2.2 of residual sugar. Several outliers are present along the higher ranges. The distribution is normally distributed.



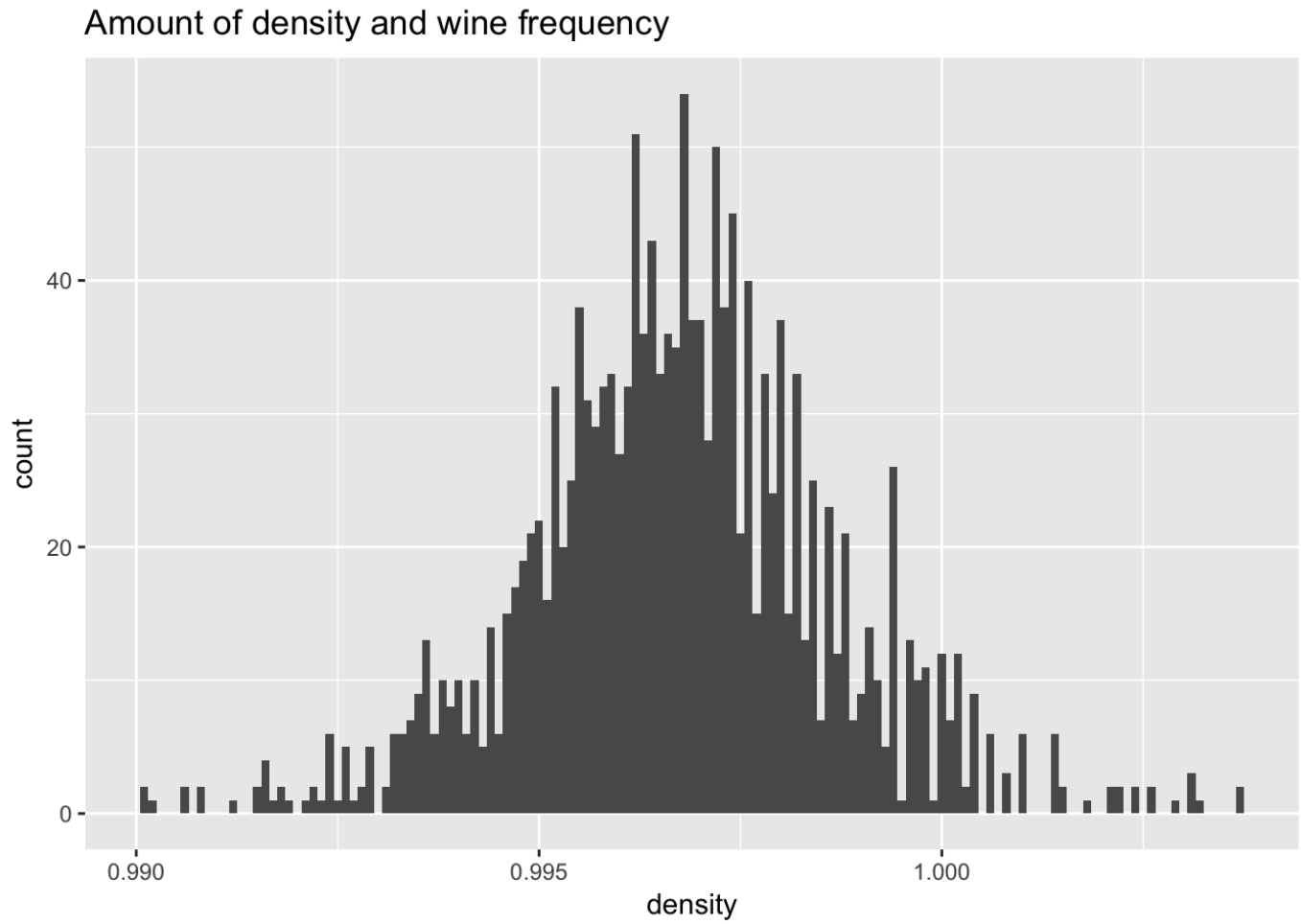
The distribution is similar with chlorides. There are also outliers in the higher ranges.



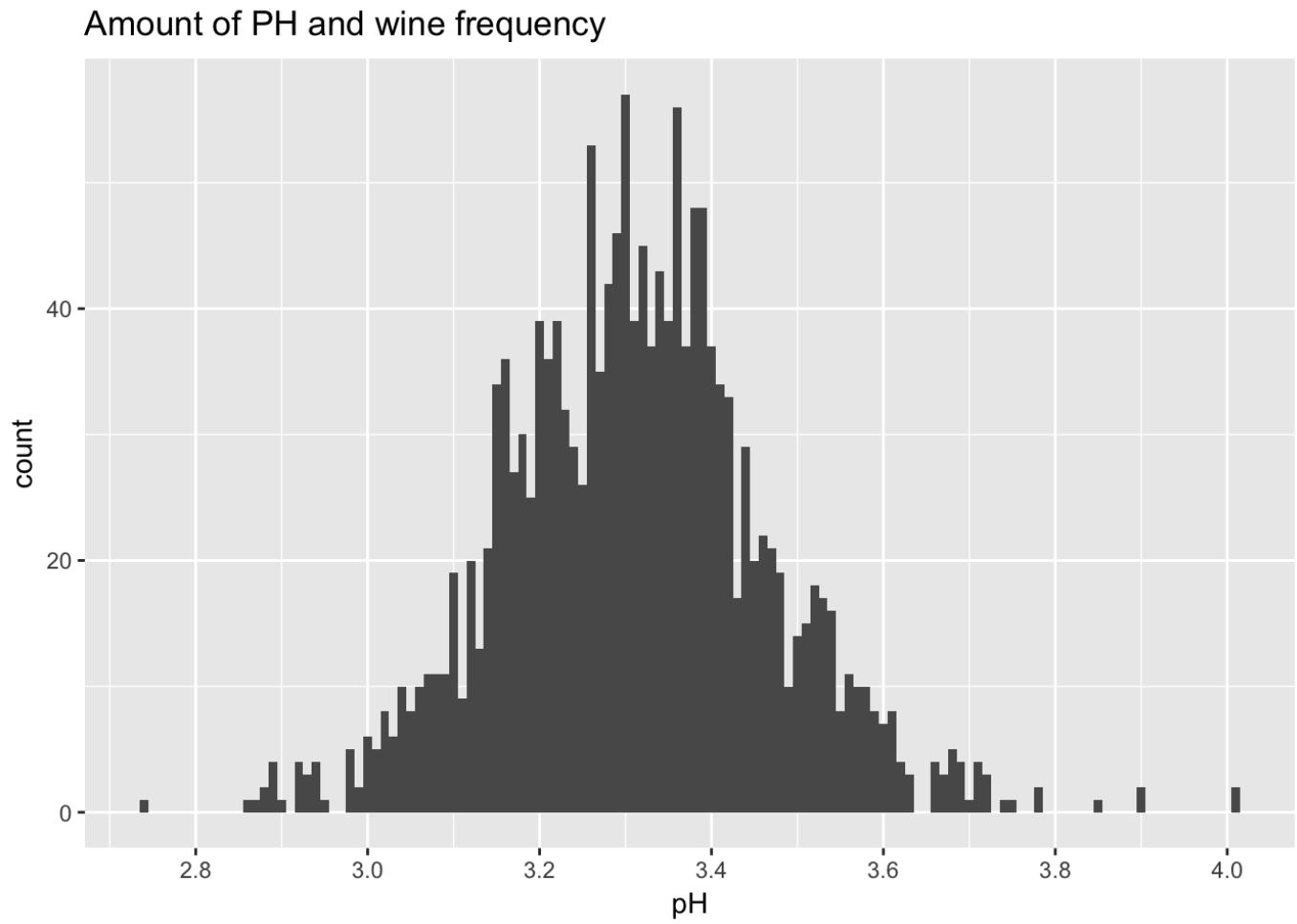
The distributions peaks at around 7 and then few wines contain over 60 of free sulfur dioxide.



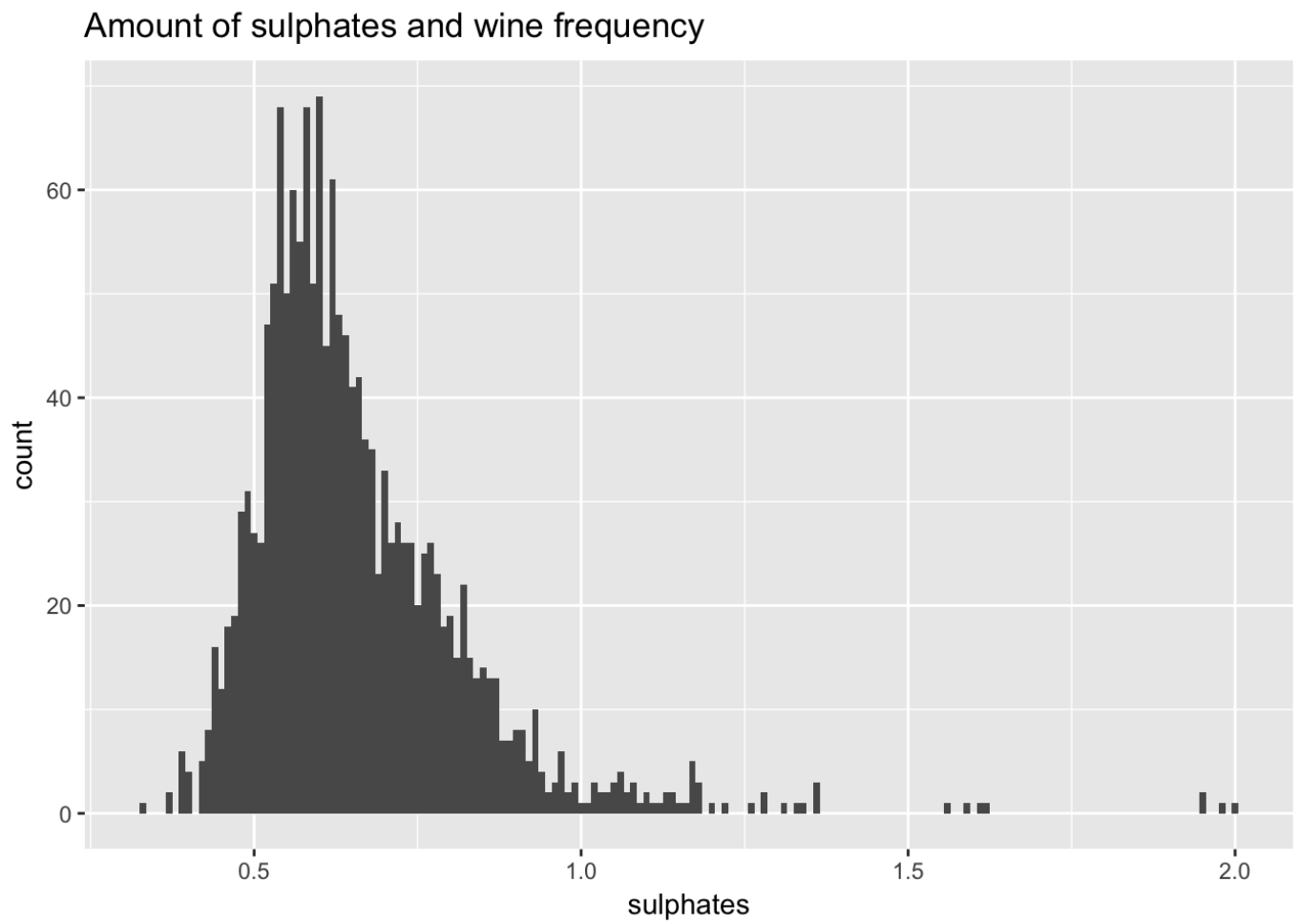
This distribution has a shape quite similar to the previous one, with a long tailed distribution. Outliers appear around 275.



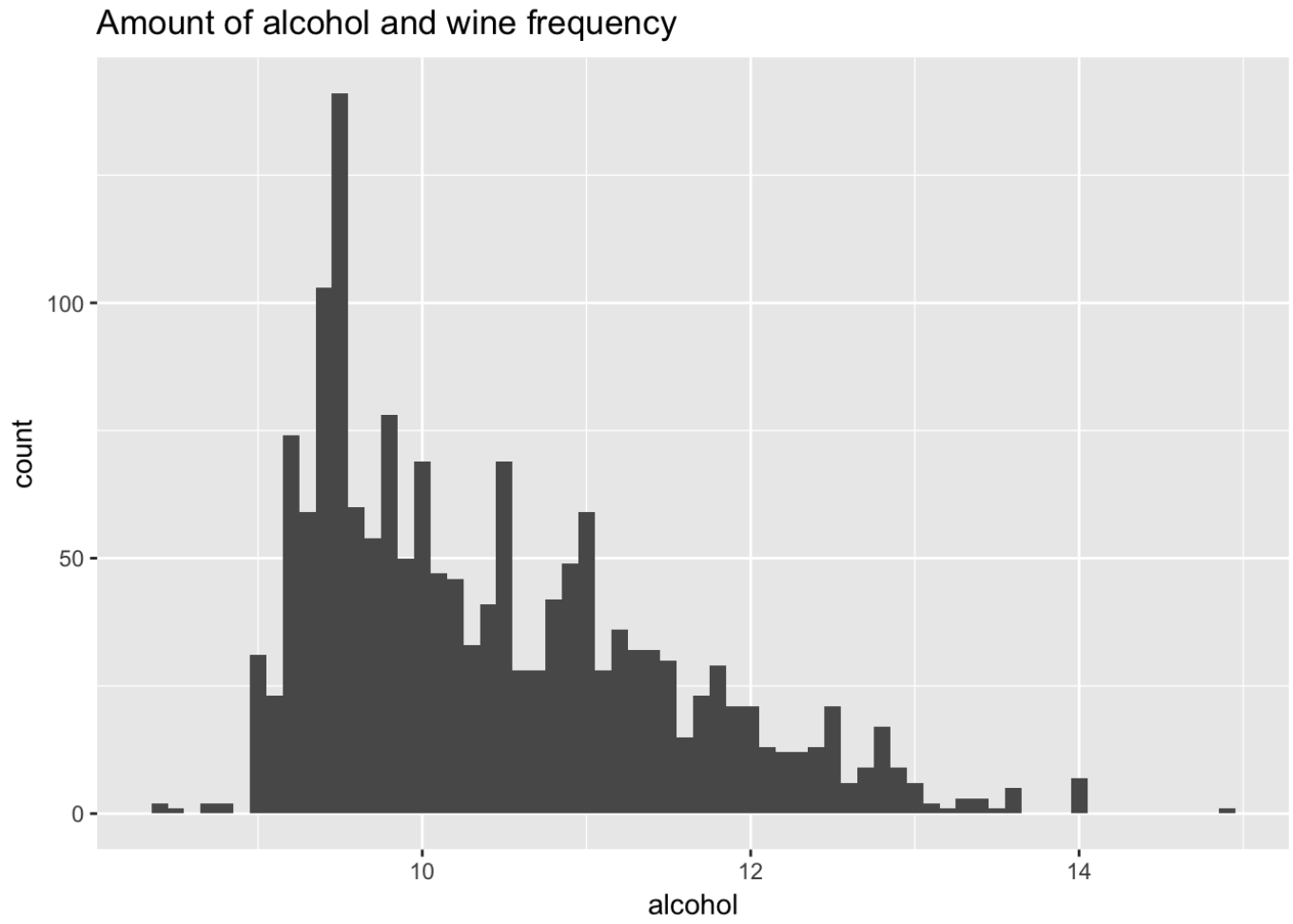
The distribution for density has a normal distribution.



pH is as well normally distributed, with outliers in the higher ranges.



Concerning sulphates, the distribution is similar to the ones of residual.sugar and chlorides.



The distribution is unimodal at around 7.5. It is skewed to the right.

Univariate Analysis

What is the structure of your dataset?

The dataset is composed of 1,599 red wines with 12 features of the chemical properties of wine (fixed.acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol and quality).

The median quality is 5 and about 75% of wine have a quality lower than 6. 50% of wines have a pH of 3.31 or higher. The median percent alcohol content is 10.20 and the maximum percentage of alcohol content is 14.90.

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

The main features in the data set are pH and quality. The aim of this project is to determine which variables have the highest impact on quality.

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

Volatile acidity, citric acid, and alcohol likely contribute to the quality of a wine. I think volatile acidity (the amount of acetic acid in wine) and alcohol (the percent alcohol content of the wine) probably contribute most to the quality after researching information on wine quality.

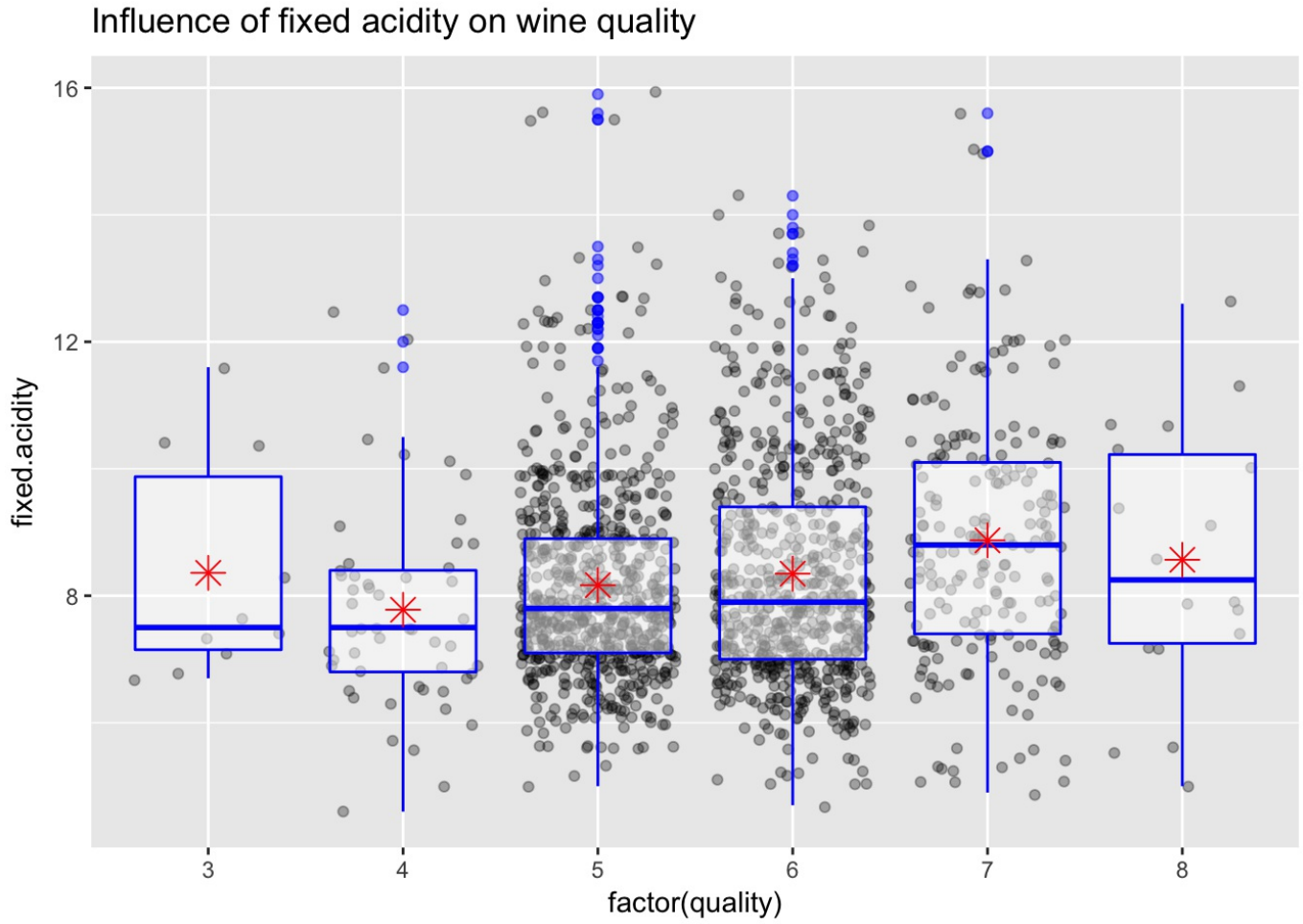
Did you create any new variables from existing variables in the dataset?

No I did not.

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?

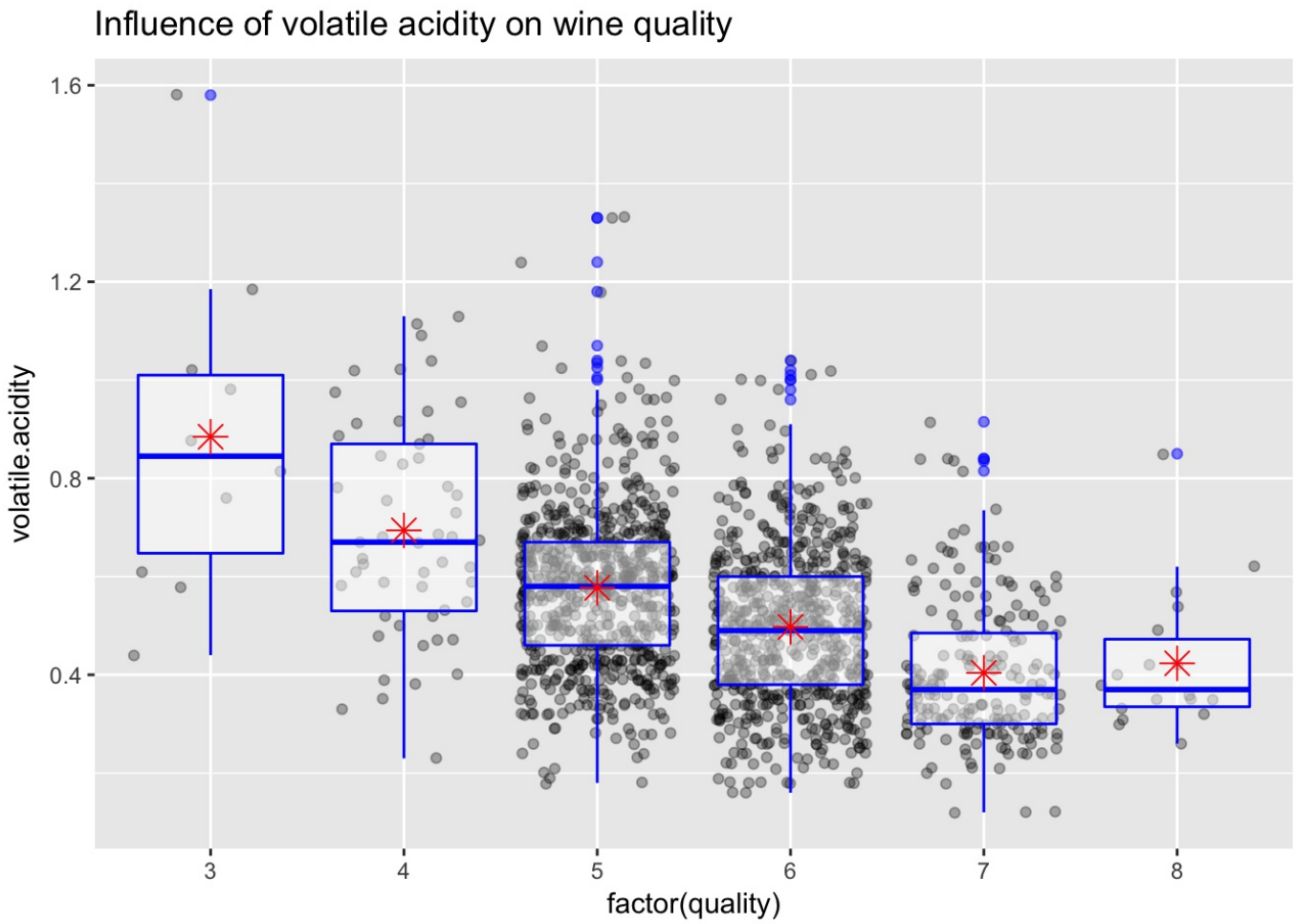
Since the data is clean, I did not perform any cleaning process or modification of the data. I did not observed any unusual distributions.

Bivariate Plots Section



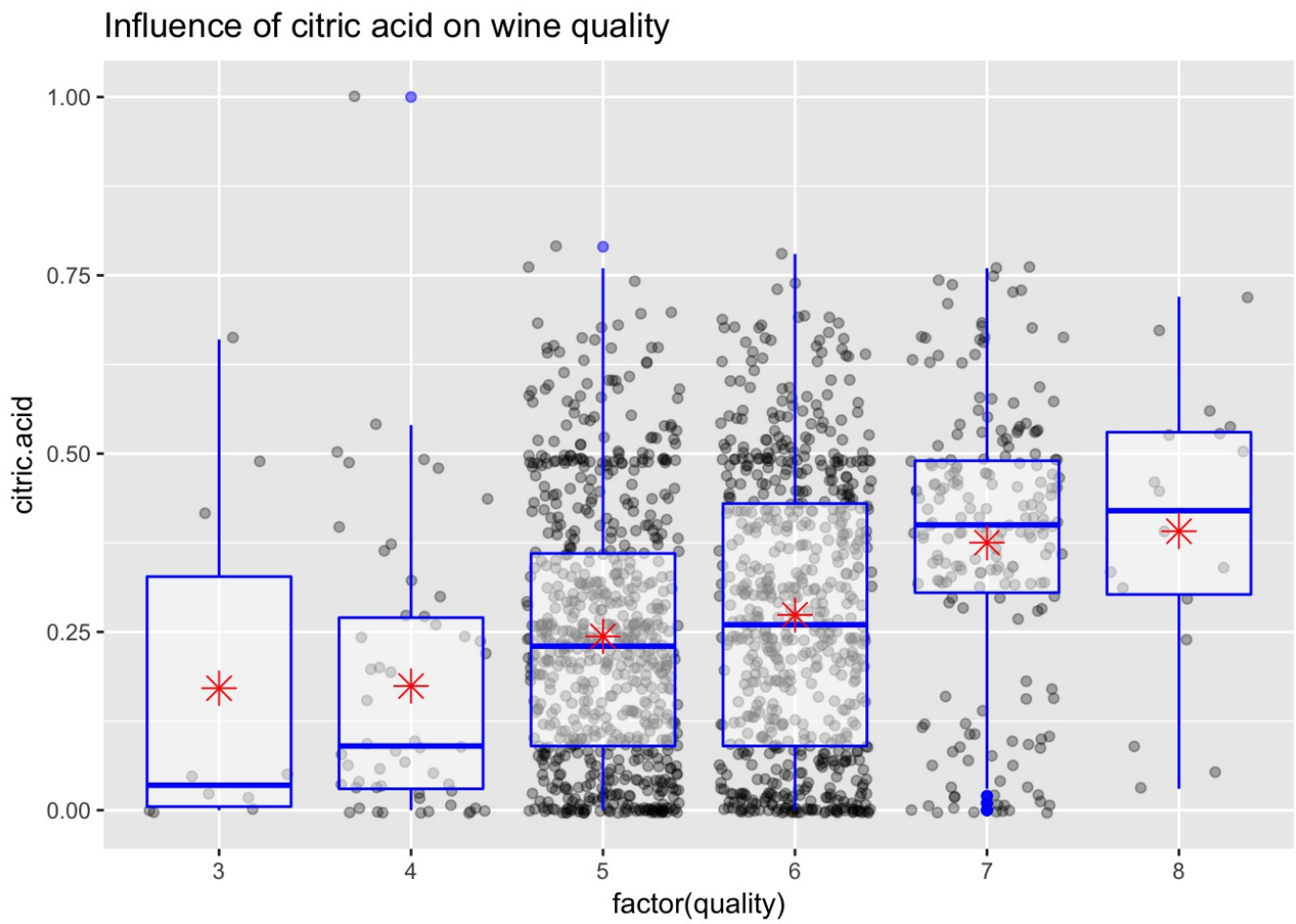
```
##
## --------------------------
##  quality   mean    median
## --------- ------- --------
##     3       8.36     7.5
##
##     4       7.779    7.5
##
##     5       8.167    7.8
##
##     6       8.347    7.9
##
##     7       8.872    8.8
##
##     8       8.567    8.25
## --------------------------
##
## Table: Summaries for fixed.acidity grouped by quality
```

The boxplot above shows that there are several outliers for a quality from 4 to 7. As shown in the table above, fixed acidity seems to have a slightly effect on the quality of wines.



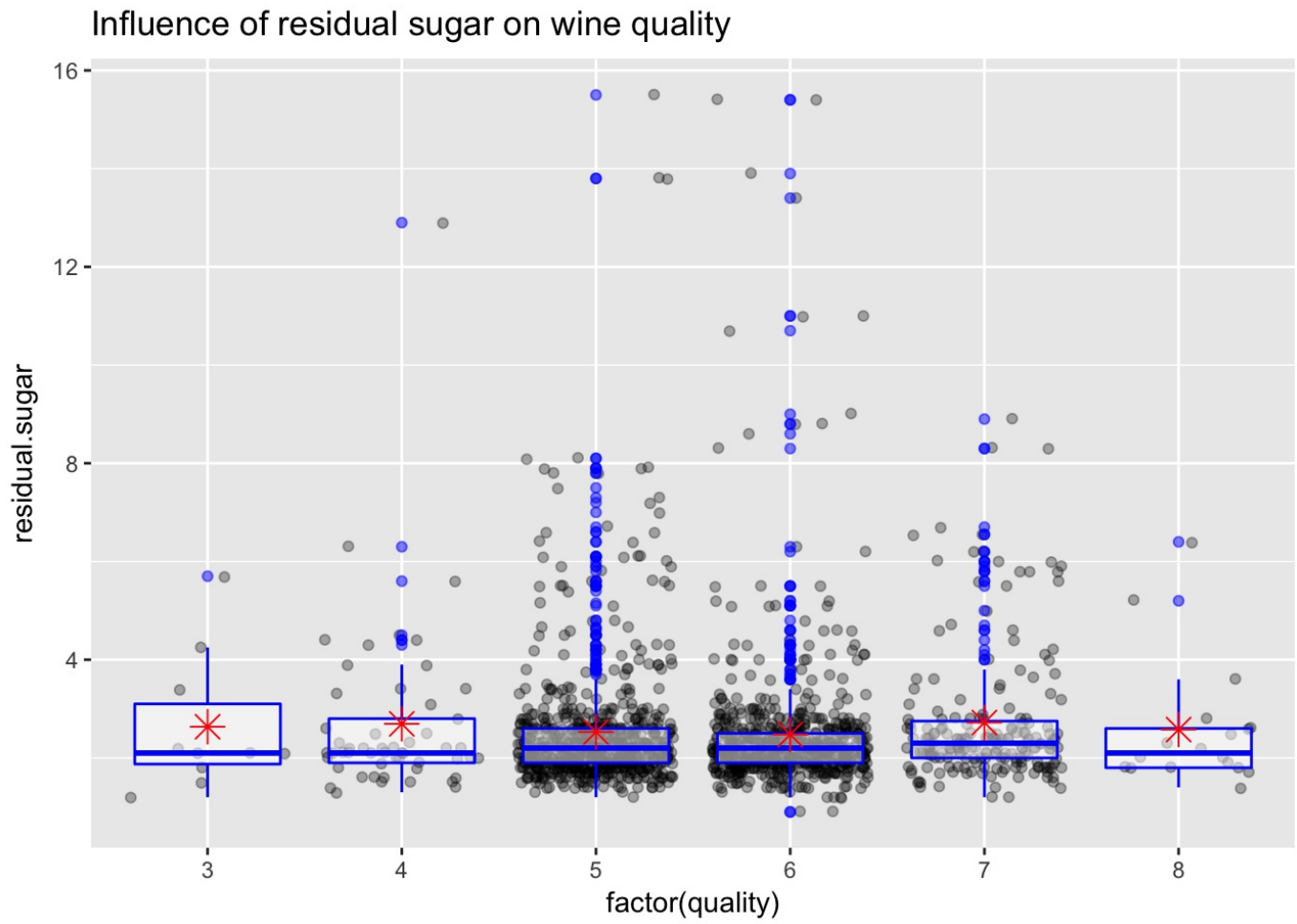
```
##
## ----------------------------
##  quality    mean    median
## --------- -------- --------
##      3       0.8845   0.845
##
##      4       0.694     0.67
##
##      5       0.577     0.58
##
##      6       0.4975    0.49
##
##      7       0.4039    0.37
##
##      8       0.4233    0.37
## ----------------------------
##
## Table: Summaries for volatile.acidity grouped by quality
```

The gaph above shows that the lowest the volatile acidity, the lowest the quality. Therefore, volatile acidity has an impact on the quality of wines.



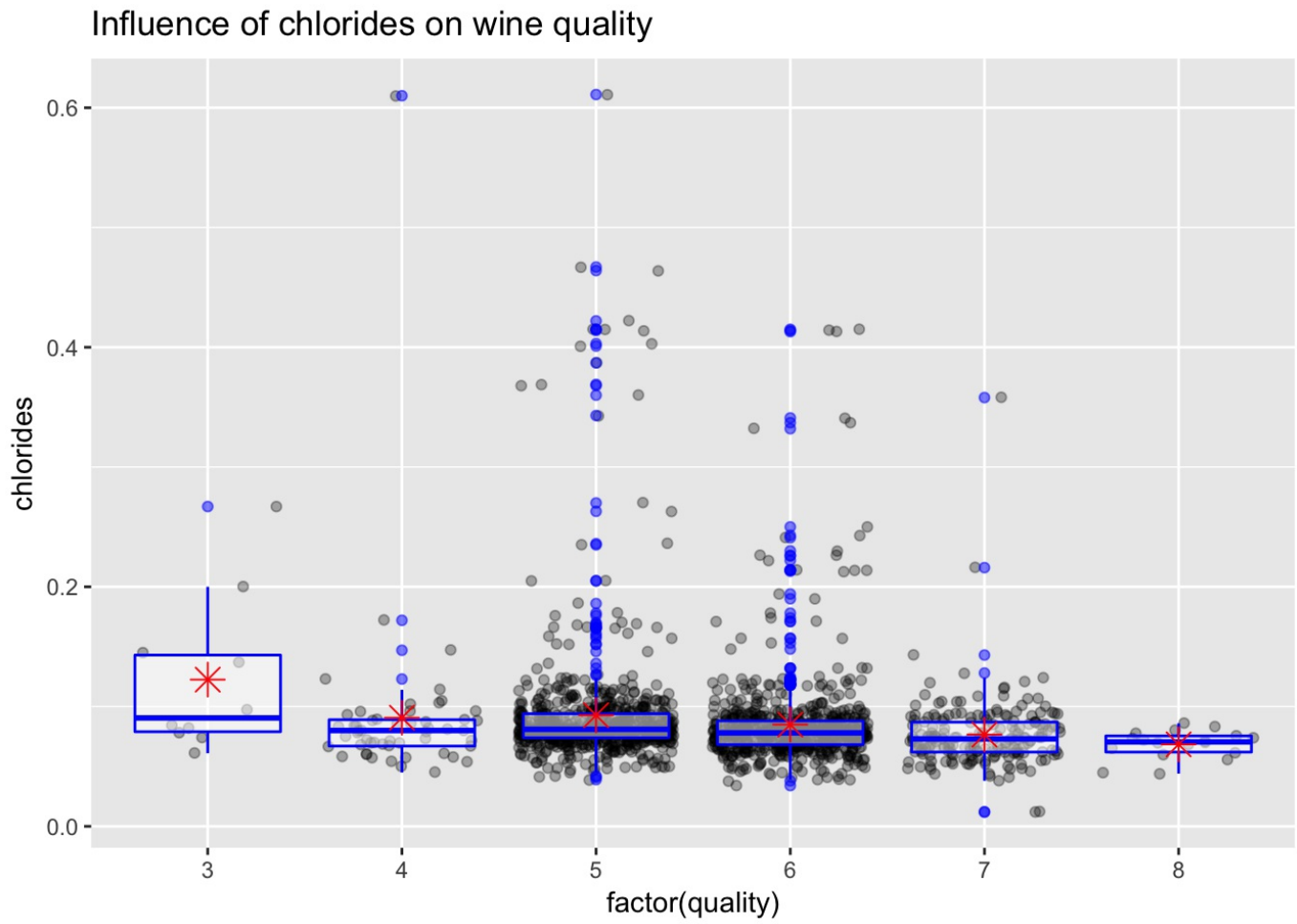
```
##
## ---------------------------
##  quality    mean    median
## --------- -------- --------
##     3       0.171    0.035
##
##     4       0.1742    0.09
##
##     5       0.2437    0.23
##
##     6       0.2738    0.26
##
##     7       0.3752    0.4
##
##     8       0.3911    0.42
## ---------------------------
##
## Table: Summaries for citric.acid grouped by quality
```

Better wines tend to have higher concentration of citric acid.



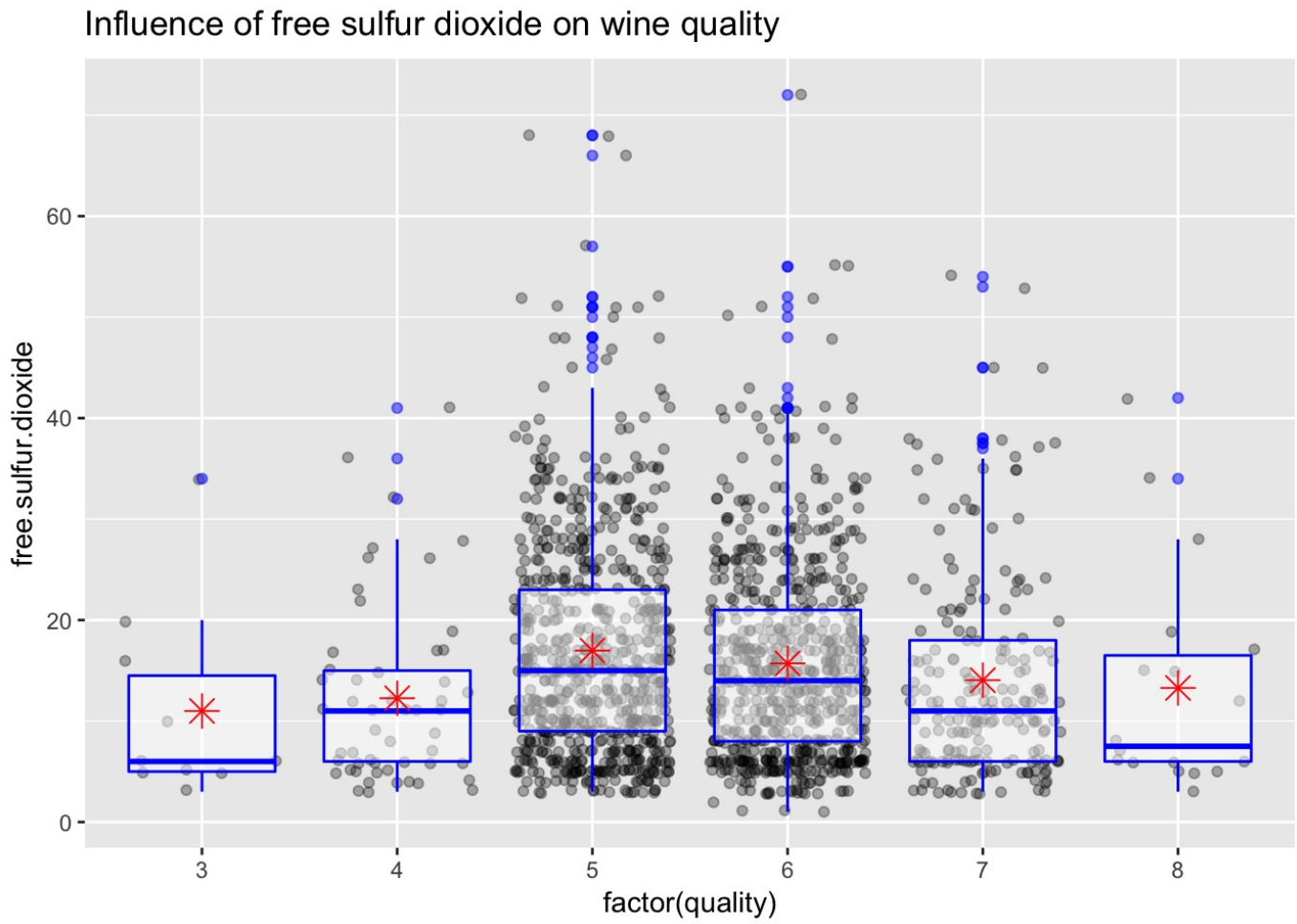
```
##
## --------------------------
##  quality   mean    median
## --------- ------- --------
##     3      2.635    2.1
##
##     4      2.694    2.1
##
##     5      2.529    2.2
##
##     6      2.477    2.2
##
##     7      2.721    2.3
##
##     8      2.578    2.1
## --------------------------
##
## Table: Summaries for residual.sugar grouped by quality
```

Contrary to expectations, residual sugar has little effect on the perceived quality of experts.



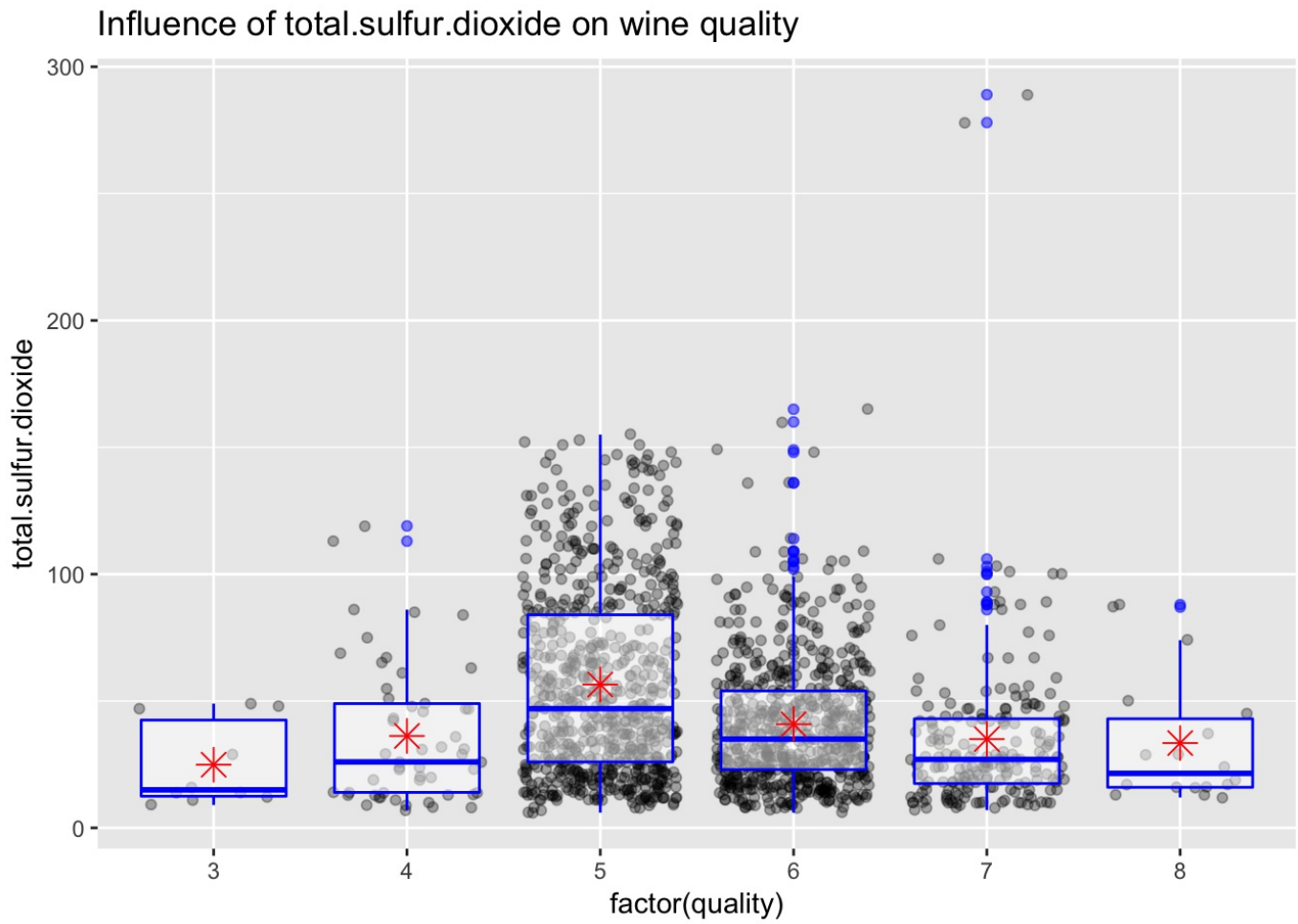
```
##
## ----------------------------
##  quality    mean     median
## --------- --------- --------
##     3       0.1225    0.0905
##
##     4       0.09068    0.08
##
##     5       0.09274   0.081
##
##     6       0.08496   0.078
##
##     7       0.07659   0.073
##
##     8       0.06844   0.0705
## ----------------------------
##
## Table: Summaries for chlorides grouped by quality
```

As for residual sugar, chlorides has seems to have a little impact on quality.



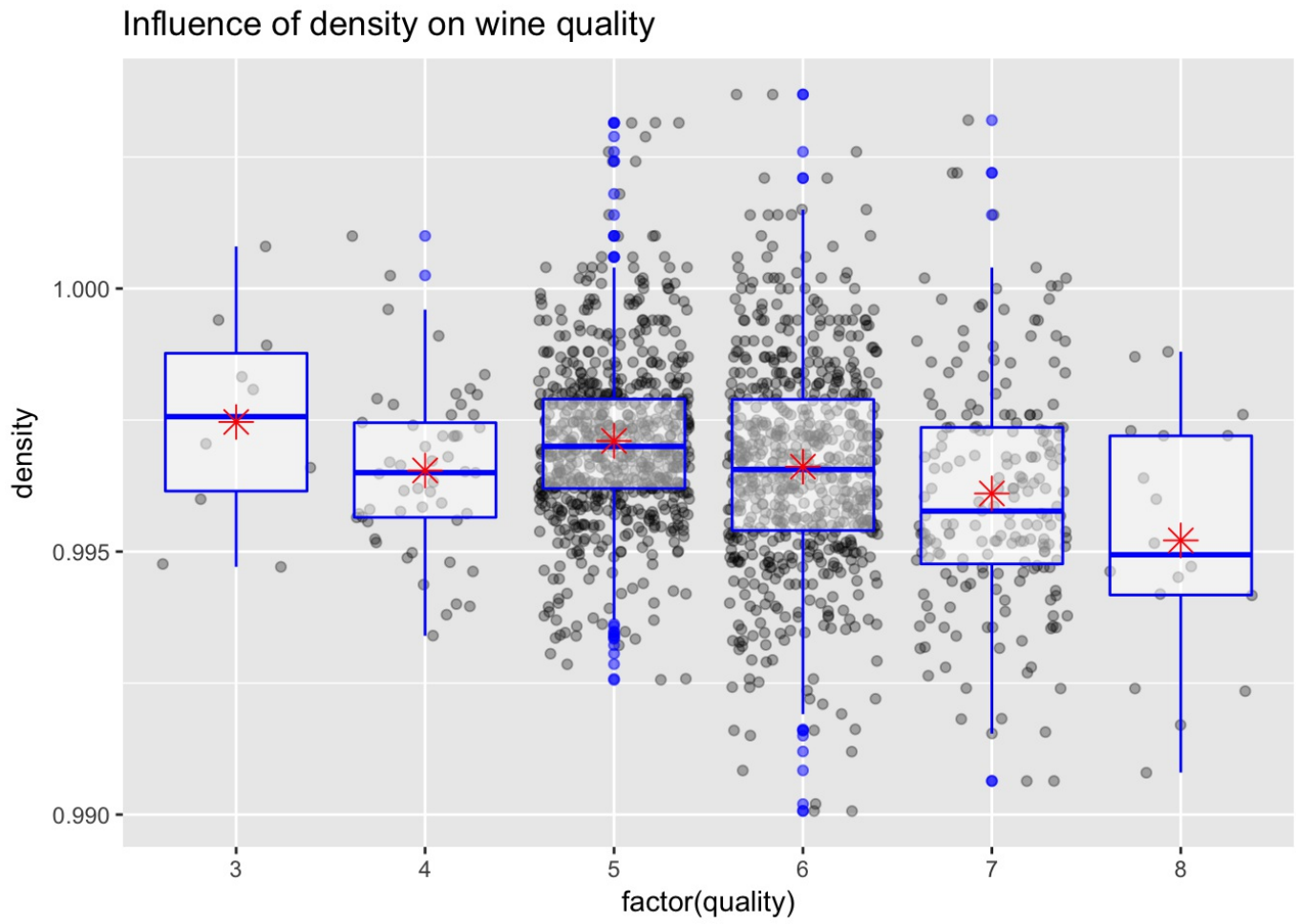
```
##
## --------------------------
##  quality   mean    median
## --------- ------- --------
##     3        11        6
##
##     4       12.26     11
##
##     5       16.98     15
##
##     6       15.71     14
##
##     7       14.05     11
##
##     8       13.28    7.5
## --------------------------
##
## Table: Summaries for free.sulfur.dioxide grouped by quality
```

The ranges are really close to each other but it seems too little sulfur dioxide and we get a poor wine, too much and we get an average wine.



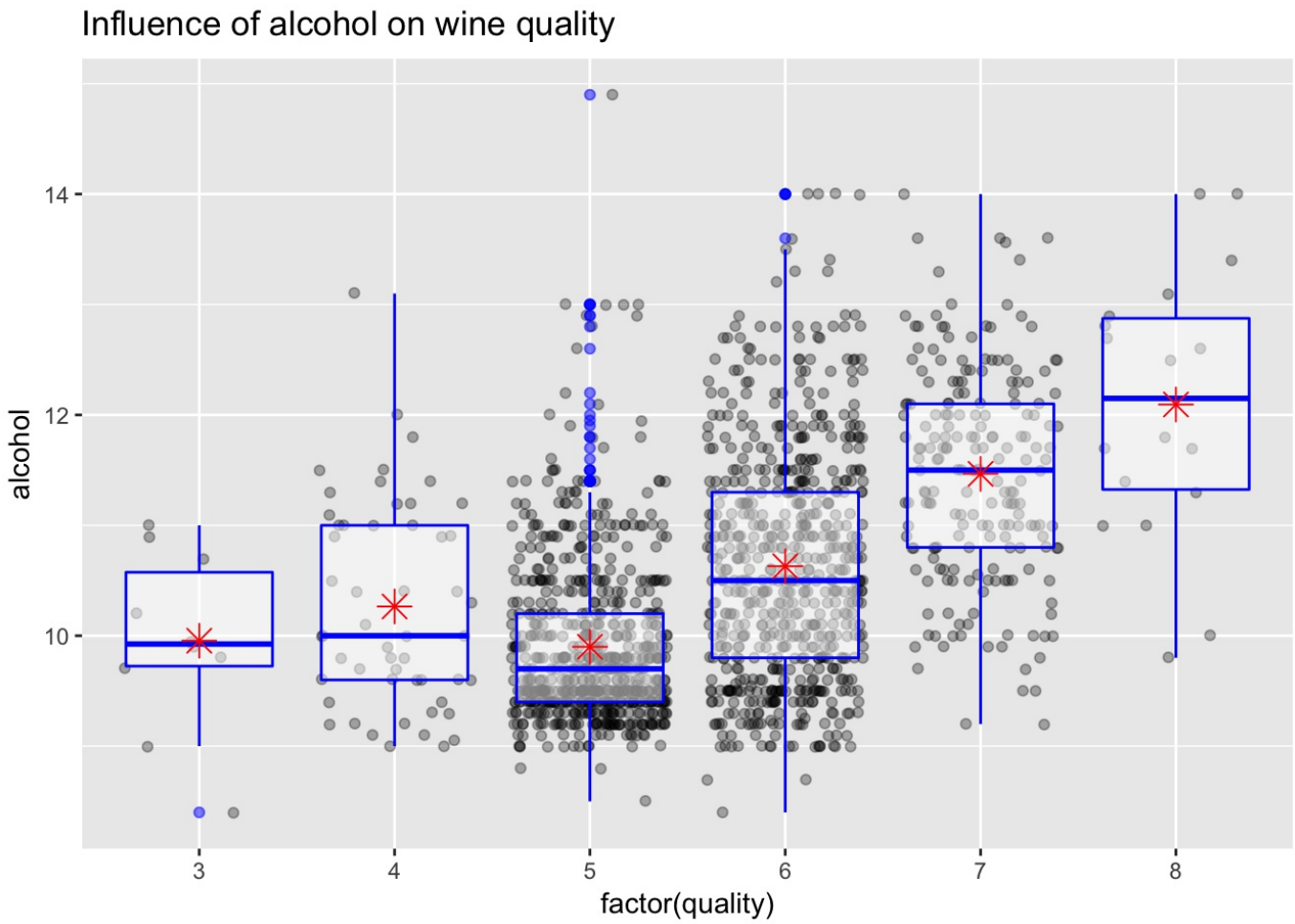
```
##
## --------------------------
##  quality   mean    median
## --------- ------- --------
##     3       24.9       15
##
##     4       36.25      26
##
##     5       56.51      47
##
##     6       40.87      35
##
##     7       35.02      27
##
##     8       33.44    21.5
## --------------------------
##
## Table: Summaries for total.sulfur.dioxide grouped by quality
```

The total sulfur dioxide did not show any pattern concerning the quality of wine. The best wines have a sulfur dioxyde of 21.5, average wines of from 35 to 47 and the worst wines of 15.



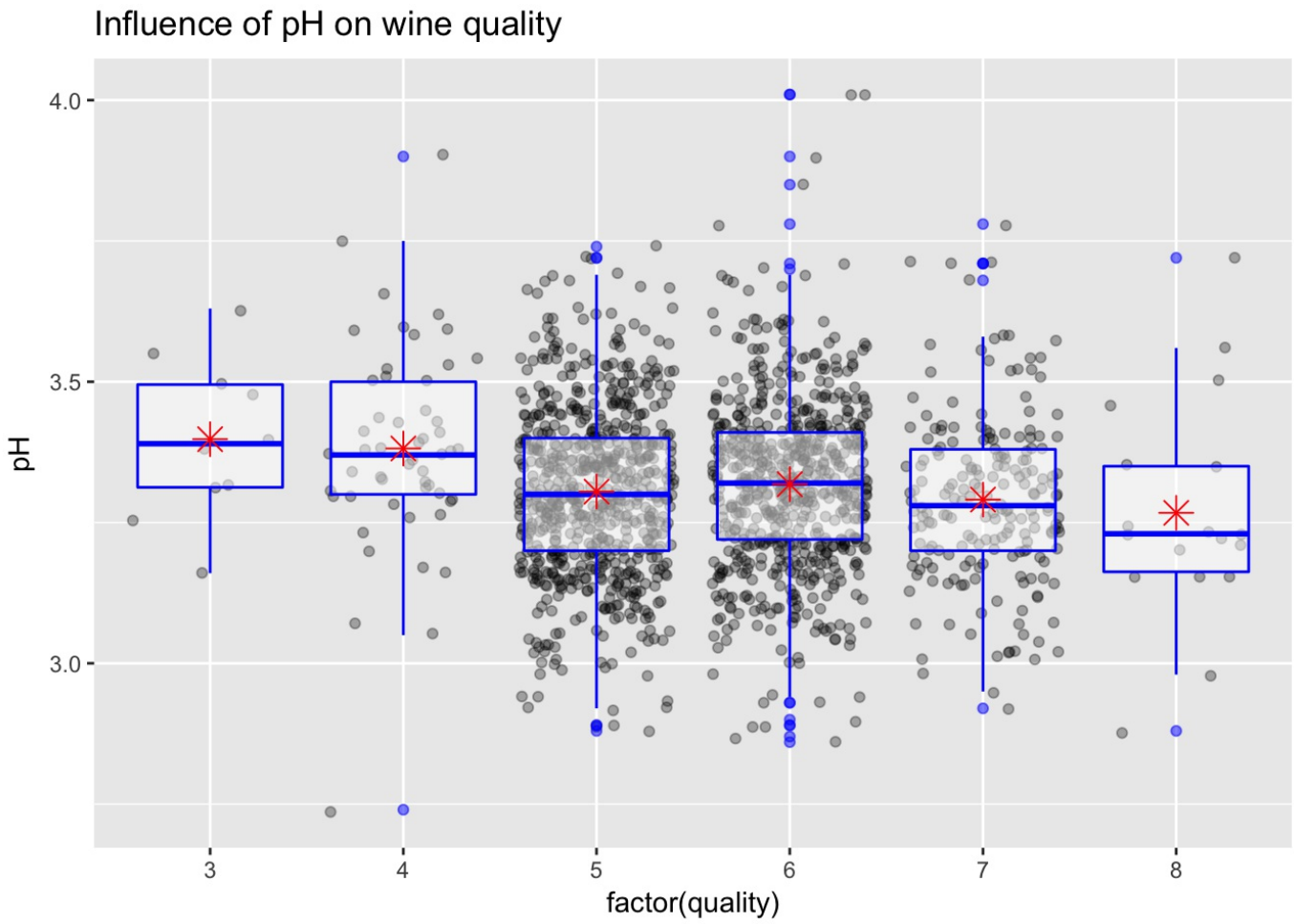
```
##
## ---------------------------
##  quality    mean    median
## --------- -------- --------
##     3       0.9975   0.9976
##
##     4       0.9965   0.9965
##
##     5       0.9971   0.997
##
##     6       0.9966   0.9966
##
##     7       0.9961   0.9958
##
##     8       0.9952   0.9949
## ---------------------------
##
## Table: Summaries for density grouped by quality
```

Better wines tend to have a lower density, which is probably connected to the amount of alcohol in the wines. Water is a little bit denser than alcohol, the higher alcohol, the least density.



```
##
## --------------------------
##  quality   mean    median
## --------- ------- --------
##     3      9.955   9.925
##
##     4      10.27     10
##
##     5       9.9      9.7
##
##     6      10.63    10.5
##
##     7      11.47    11.5
##
##     8      12.09   12.15
## --------------------------
##
## Table: Summaries for alcohol grouped by quality
```

As expected, the wines with the higher amount of alcohol have the best quality.



```
##
## --------------------------
##  quality   mean    median
## --------- ------- --------
##     3      3.398    3.39
##
##     4      3.382    3.37
##
##     5      3.305    3.3
##
##     6      3.318    3.32
##
##     7      3.291    3.28
##
##     8      3.267    3.23
## --------------------------
##
## Table: Summaries for pH grouped by quality
```

There is definitely a trend that suggests that better wines have more acid. The lower the pH, the higher the quality of the wine.

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?

I observed the relationship between quality and variables.

Fixed acidity seems to have little to no effect on quality. The less the volatile acidity, the less the quality, indeed the higher ranges seem to produce more average and poor wines. Better wines tend to have higher concentration of citric acid.As for residual sugar, it seems to have little to no effect on perceived quality of red wines. Altough weakly correlated, a lower concentration of chlorides seem to produce better wines. Better wines tend to have lower densities and consequently a higher amount of alcohol. In terms of pH, it seems that better wines are more acid but there were many outliers.

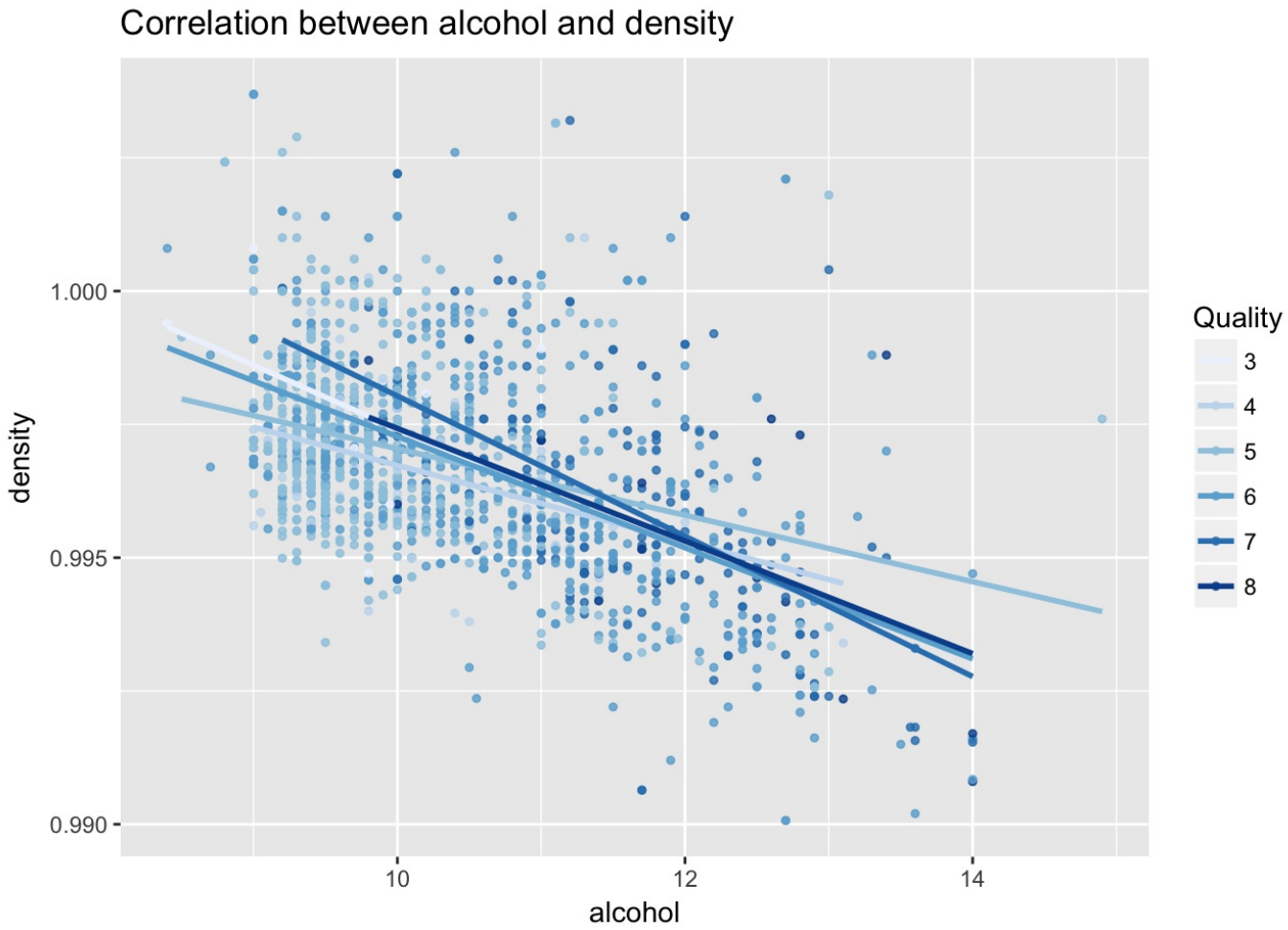
Did you observe any interesting relationships between the other features
(not the main feature(s) of interest)?

I would like to dig the is a correlation between density and alcohol, as well as pH and citric acid. The relation between citric acide and volatile acid is as well relevant for futher analysis.

What was the strongest relationship you found?

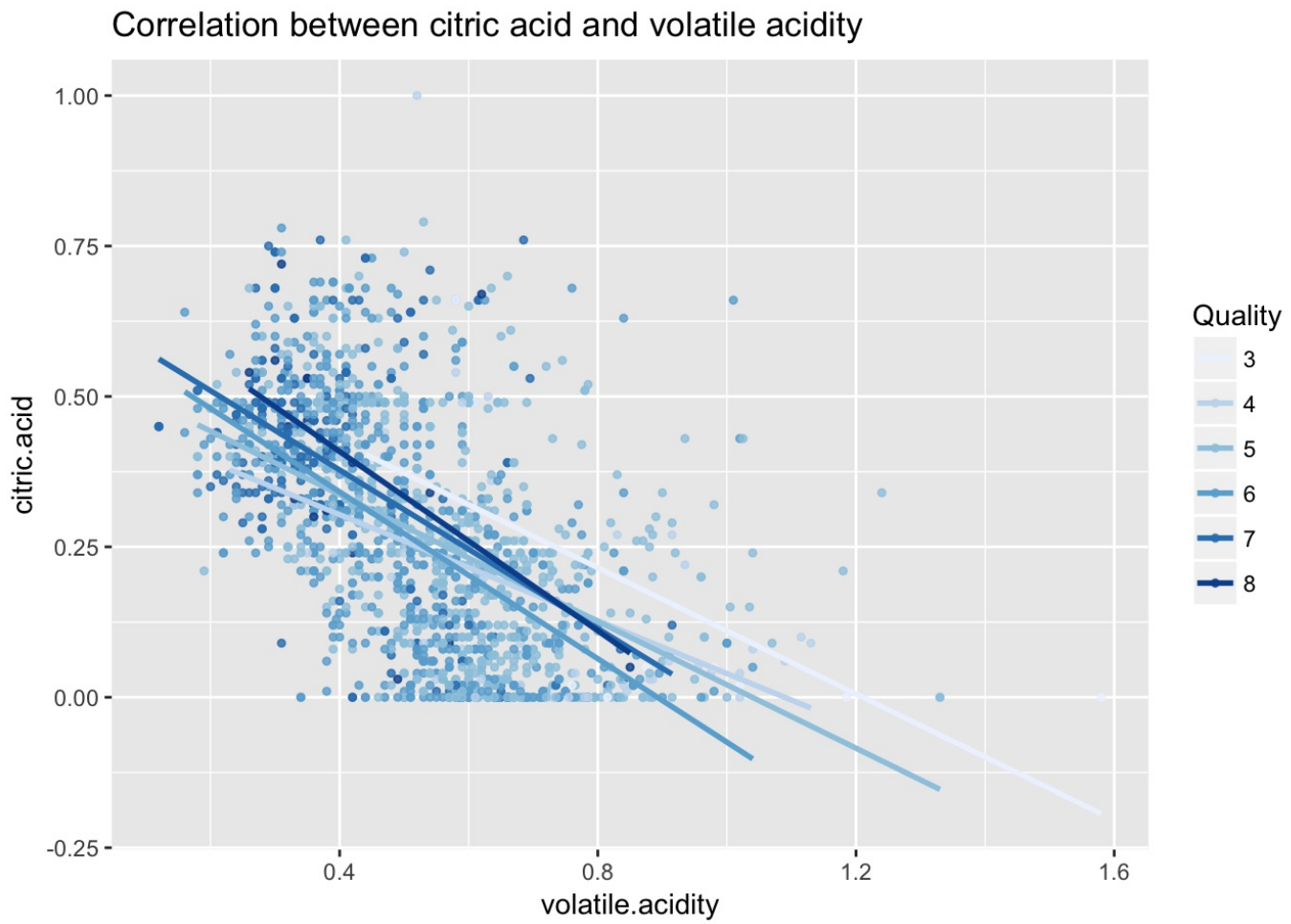
The relationship between the variables quality and citric acid, as well as volatile acidity and quality.

Multivariate Plots Section



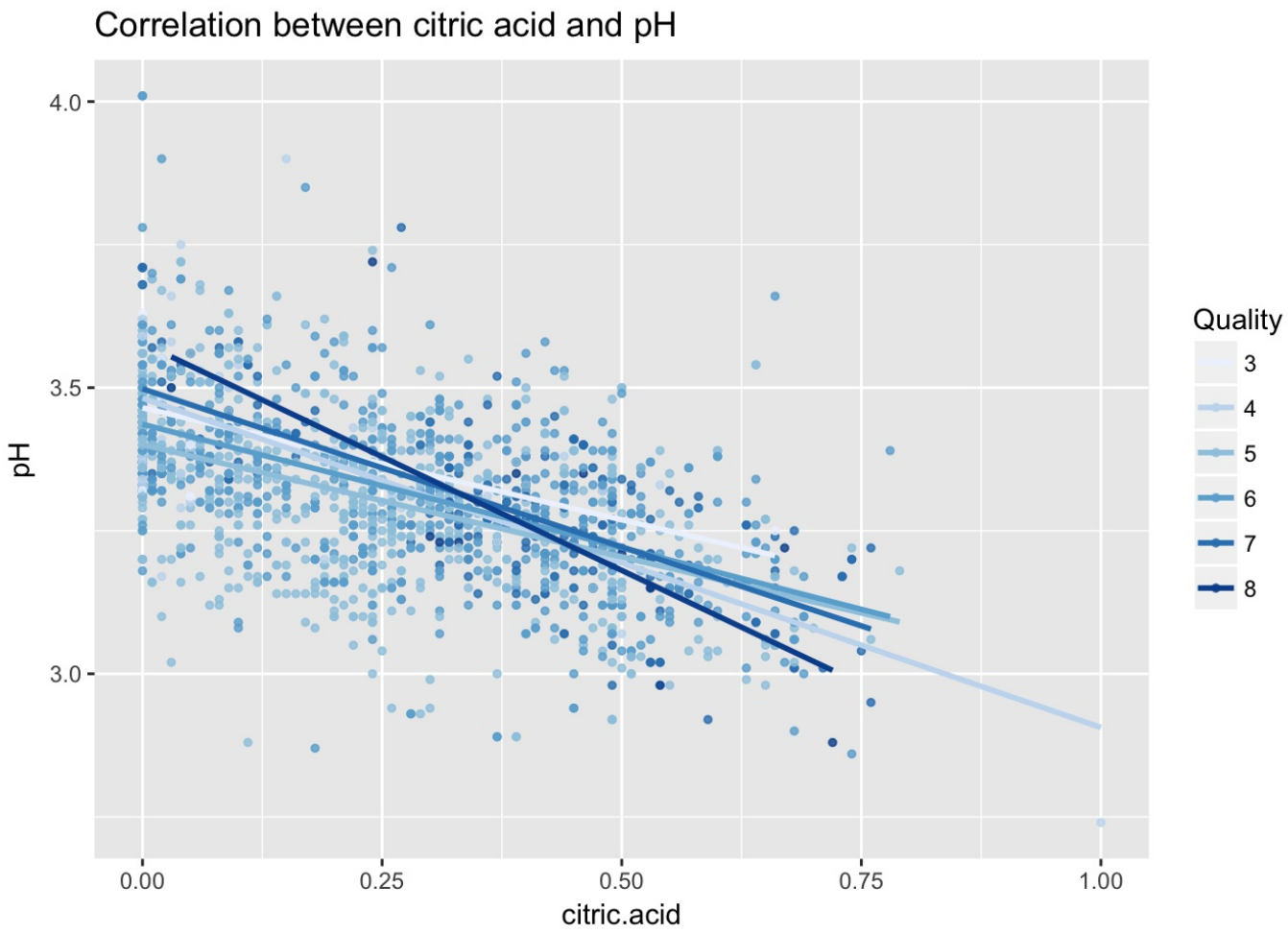

```
##
##  Pearson's product-moment correlation
##
## data:  rw$density and rw$alcohol
## t = -22.838, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5322547 -0.4583061
## sample estimates:
##       cor
## -0.4961798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

The scatterplot above shows that the more alcohol, the lower the density. There is a negative correlation of -0.5. Good wines have in average a range from 11 to 13 of concentration in acohol. Wines with a high density tends to have a poor quality.



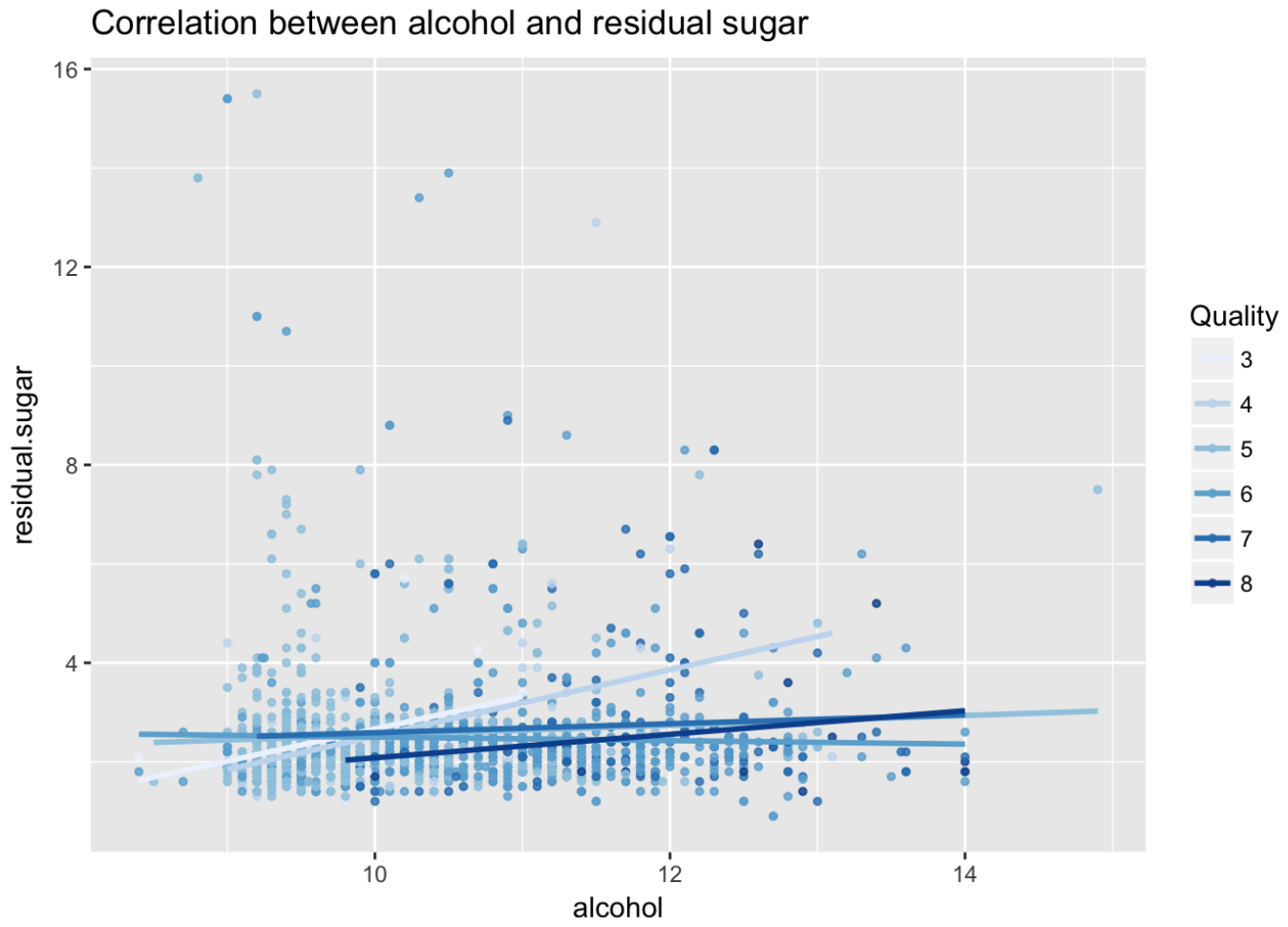

```
##
##  Pearson's product-moment correlation
##
## data:  rw$density and rw$alcohol
## t = -22.838, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5322547 -0.4583061
## sample estimates:
##        cor
## -0.4961798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

The scatterplot above shows a downhill pattern, there is a negative correlation of -0.55. As the amount of citric acid increases, volatile acidity diminishes. Wines of quality tend to have a low amount of volatile acidity and a high concentration of citric acid.




```
##
##  Pearson's product-moment correlation
##
## data:  rw$density and rw$alcohol
## 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 scatterplot above shows a dowhill pattern, with a concentration of good wines that have the highest content of citric acid. There is a negative correlation, the more citric acid, the lower the pH.




```
##
##  Pearson's product-moment correlation
##
## data:  rw$density and rw$alcohol
## 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
```

This scatteprlot shows no correlation. Residual sugar does not have any impact on the amount of alcohol in wines.

Multivariate Analysis

Talk about some of the relationships you observed in this part of the

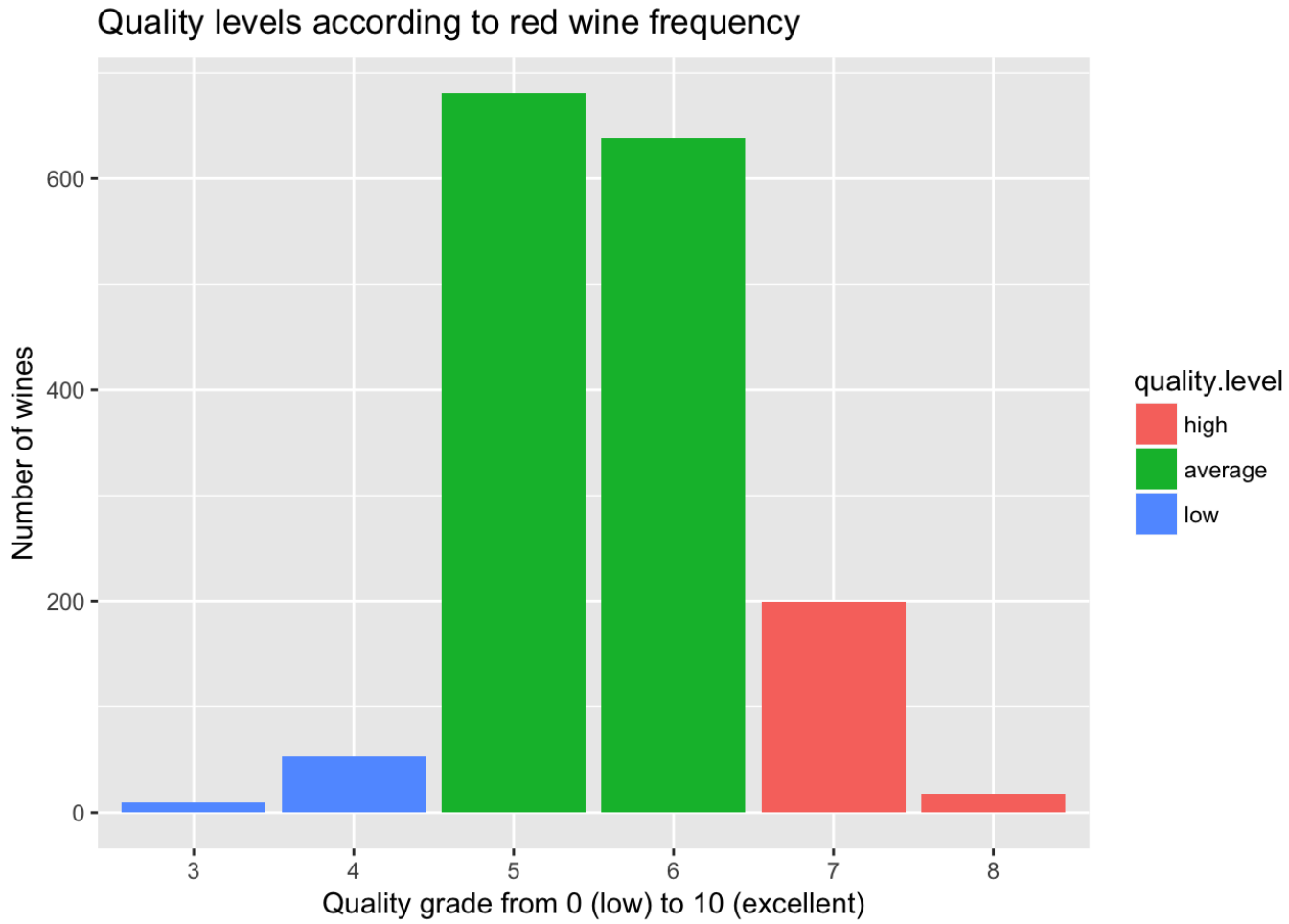
The more alcohol, the better the wine. This fact is similar to citric acid. There is no relationship between alcohol and residual sugar. There is a negative correlation between citric acid and pH, which is logical as pH measure the amount of acid in a solution.

Were there any interesting or surprising interactions between features?

I did not expect to observe that wines with more alcohol tend to be better perceived. I expected citric acid and volatile acidity to have no correlation at all. Also, I am surprise that residual sugar and acohol do not have a correlation, as the fermenting of wine transforms sugar into alcohol and the residual sugar is the amount of sugar left that has not been transformed into alcohol.

Final Plots and Summary

Plot One



Description One

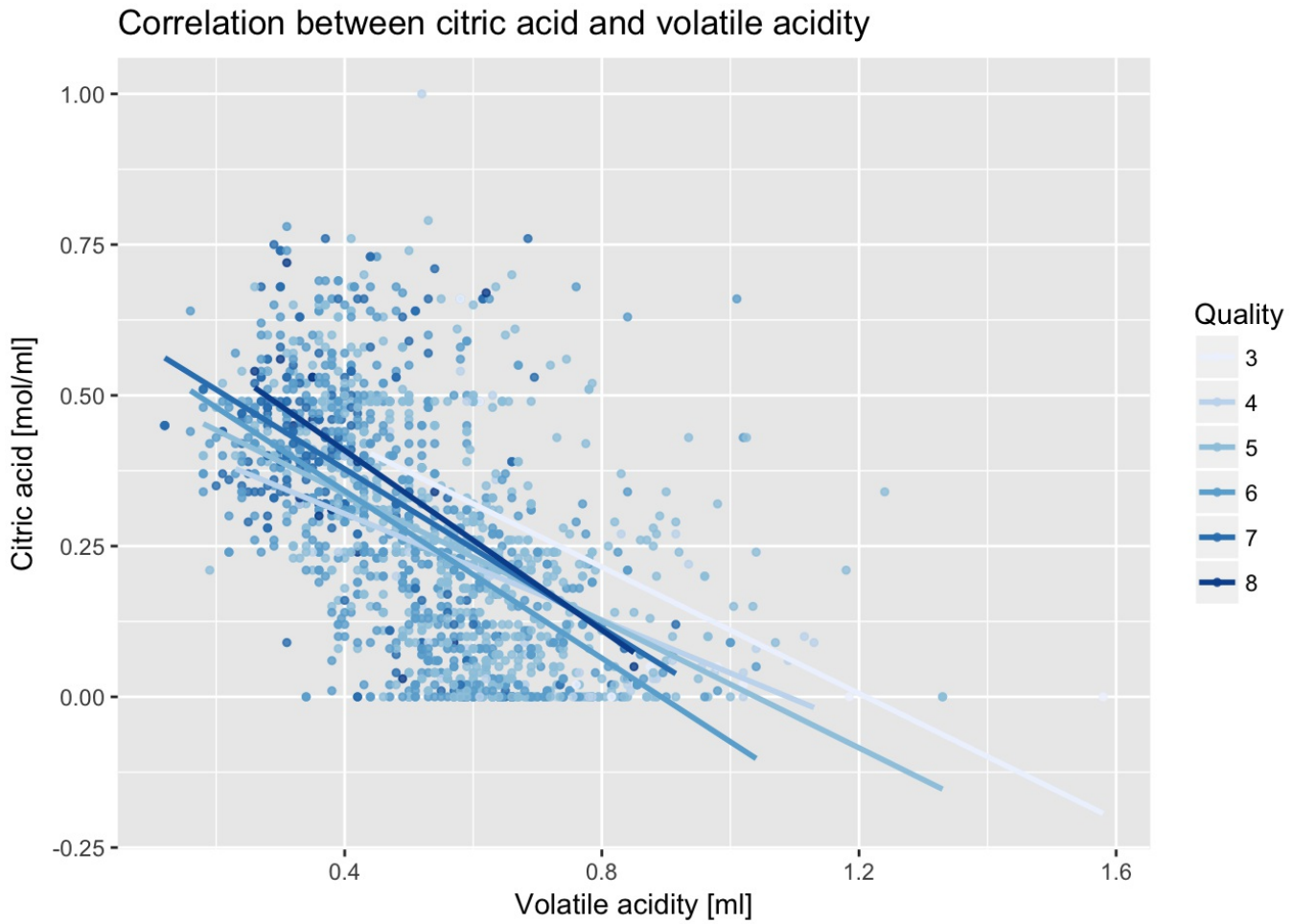
The graph above shows the frequency of wines per grade of quality. The distribution of red wine quality appears to be normal. About 80% of wines are rated from 5 to 6 (average quality) and although the rating scale is between 0 and 10, there are no wine with a rate of 1, 2, 9 or 10.

Plot Two

Description Two

The boxplot above shows that the more alcohol, the higher the quality of the wine. There are several outliers for the grade quality of 5 and 6.

Plot Three



Description Three

The red wines with the highest concentration of citric acid have the highest grade. Indeed, citric acid is an important component in the quality of wine. Most high quality wines have a relatively high citric acid, whereas average and low quality wines have lower amount of citric acid. Concerning volatile acidity, lower quality wines tend to have higher amount of this chemical, while better wines have a lower concentration of volatile acidity. The scatterplot above shows a negative correlation between citric acid and volatile acidity. Therefore, the more citric acid in a wine results in less concentration in volatile acidity.

Reflection

The dataset provides information on 1,599 wines with twelve variables. I started to analyze the shape of the distribution of each variable. Then, I explored the relation of each variable with the quality of wines. Finally, I created scatterplots to observe the correlations between variable, combined with the quality of wines.

There was a trend between the volatile acidity of a wine and its quality. There was also a trend between the alcohol and its quality.

There were very few wines that were rated with low or high quality. In addition, it was difficult to compare the results of the best wines with the other grades, as the number of wines per grade varied. There were way more wines that obtained an average grade than wines that got the highest grades. The quality of this analysis could have been improved by collecting more data, and creating more variables that might have contributed to the quality of wine. This would have certainly improved the accuracy of the conclusions of this analysis.

All in all, features that impacted the most the quality of red wines were the pH, as well as the density and the

amounts of alcohol, volatile assidity and citric acid. Indeed, acids are major wine constituents and contribute greatly to its taste. In fact, acids impart the sourness or tartness that is a fundamental feature in wine taste. Wines lacking in acid are considered as “flat.”

References:

- http://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity
- https://winemakermag.com/676-the-perils-of-volatile-acidity
- https://github.com/pcasaretto/udacity-eda-project/blob/master/wine.Rmd
- http://drinks.seriouseats.com/2014/04/best-tequilas-under-25-budget-spirit-best-affordable-brand-jimador-lunazul-olmeca-tapatio-espolon.html