Red Wine Exploration by Jostin Flake

This report will be used to conduct exploratory data analysis on a dataset containing attributes and characteristics for 1,599 different selections of red wine.

Summary of the Dataframe

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

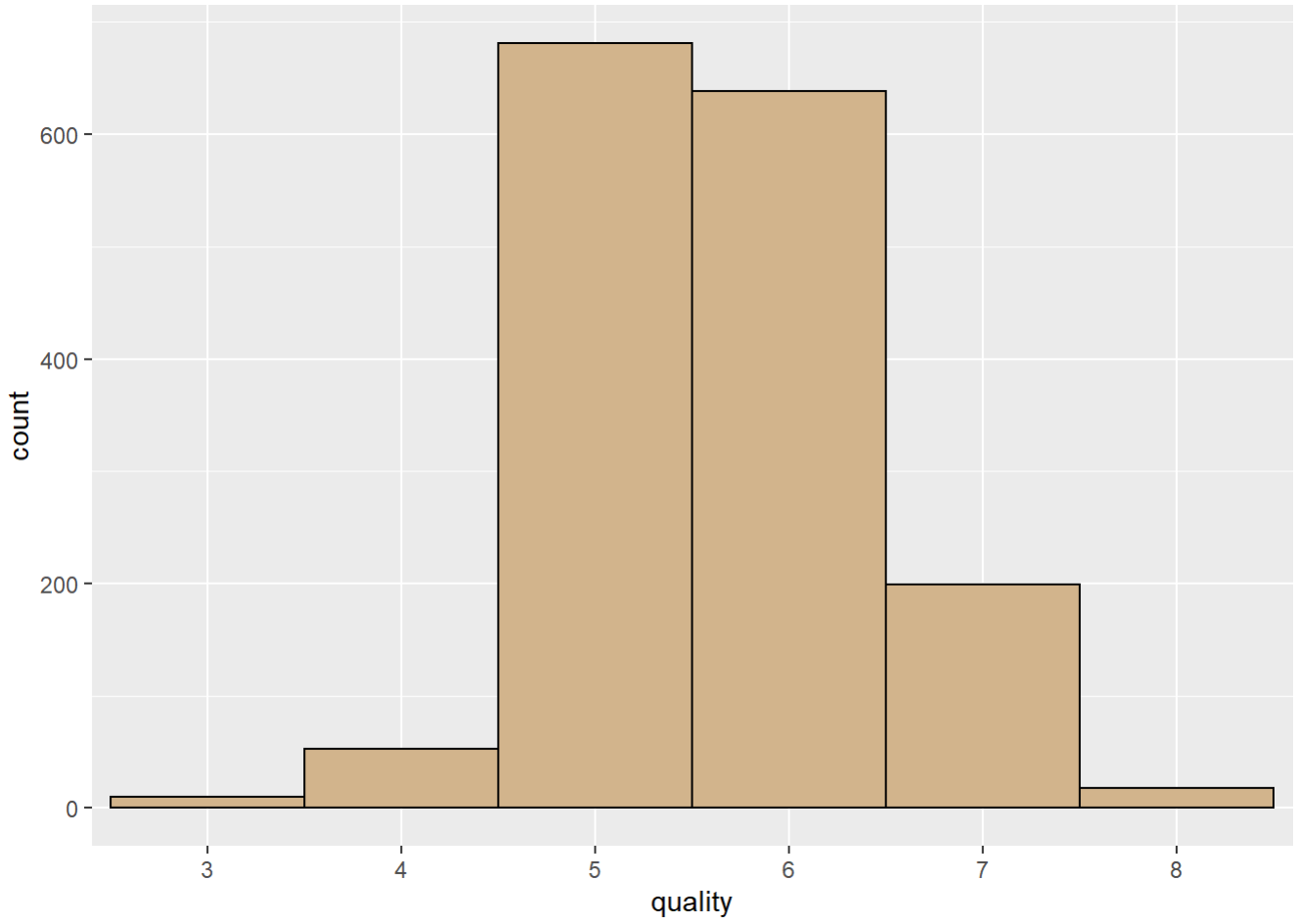
```
## 'data.frame':    1599 obs. of  13 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : 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             : Ord.factor w/ 6 levels "3"<"4"<"5"<"6"<..: 3 3 3 4 3 3 3 5 5 3 ...
##  $ grade               : Ord.factor w/ 3 levels "Mediocre"<"Very Good"<..: 2 2 2 2 2 2 2 3 3
2 ...
```

Here we examine the dimensions and structure of the dataframe.

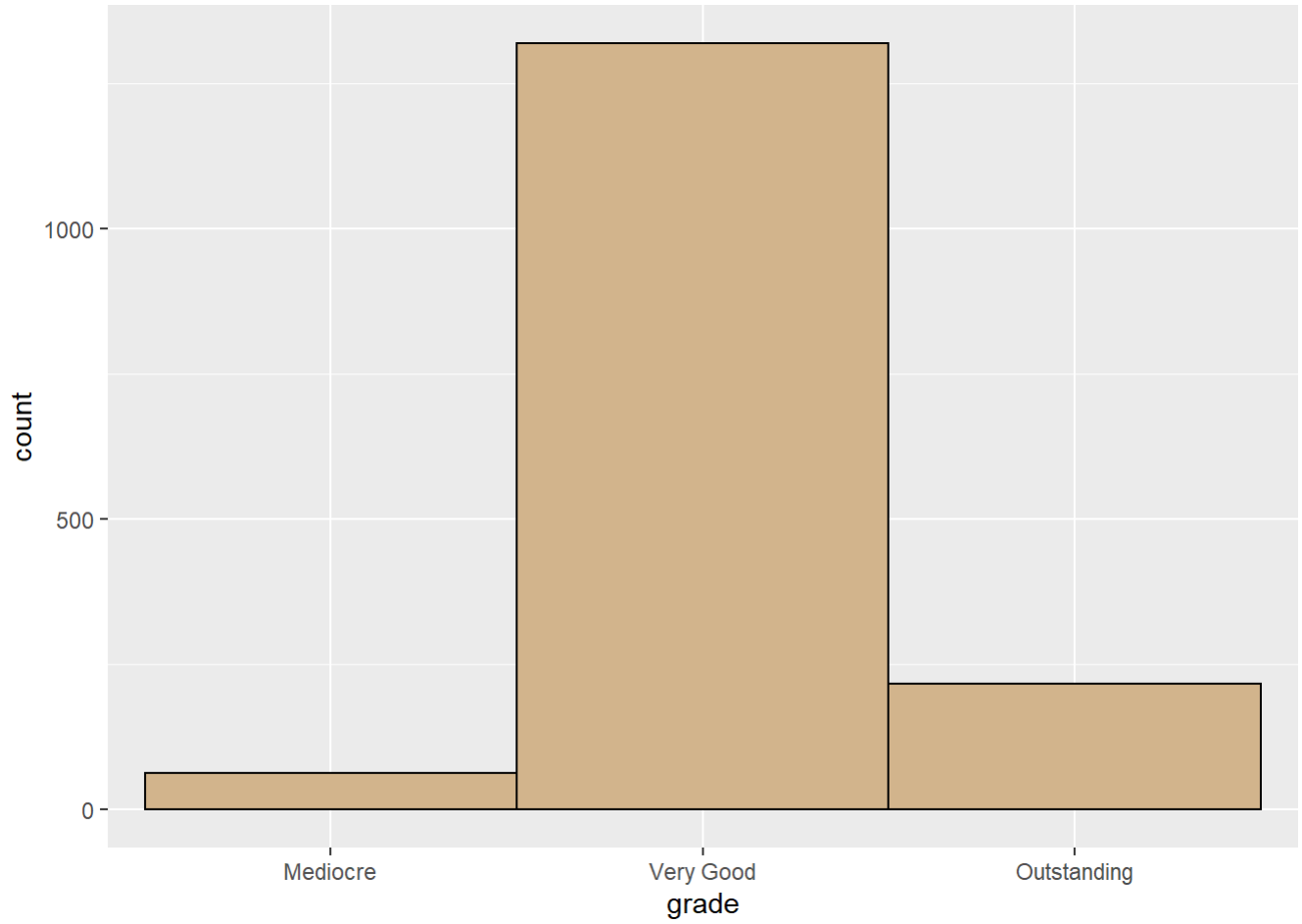
```
##
##
##  fixed.acidity   volatile.acidity    citric.acid    residual.sugar
## --------------- ------------------ --------------- ----------------
##  Min.   : 4.60    Min.   :0.1200    Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10    1st Qu.:0.3900    1st Qu.:0.090   1st Qu.: 1.900
##  Median : 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: statistical summary
##              of the red wine dataframe (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        grade
## --------------- ---------------- --------------- --------- ------------------
##  Min.   :2.740   Min.   :0.3300   Min.   : 8.40    3: 10    Mediocre   :  63
##  1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50    4: 53    Very Good  :1319
##  Median :3.310   Median :0.6200   Median :10.20    5:681    Outstanding: 217
##  Mean   :3.311   Mean   :0.6581   Mean   :10.42    6:638           NA
##  3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10    7:199           NA
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90    8: 18           NA
```

The representation shown directly above is a statistical summary provided for each variable in the dataframe.

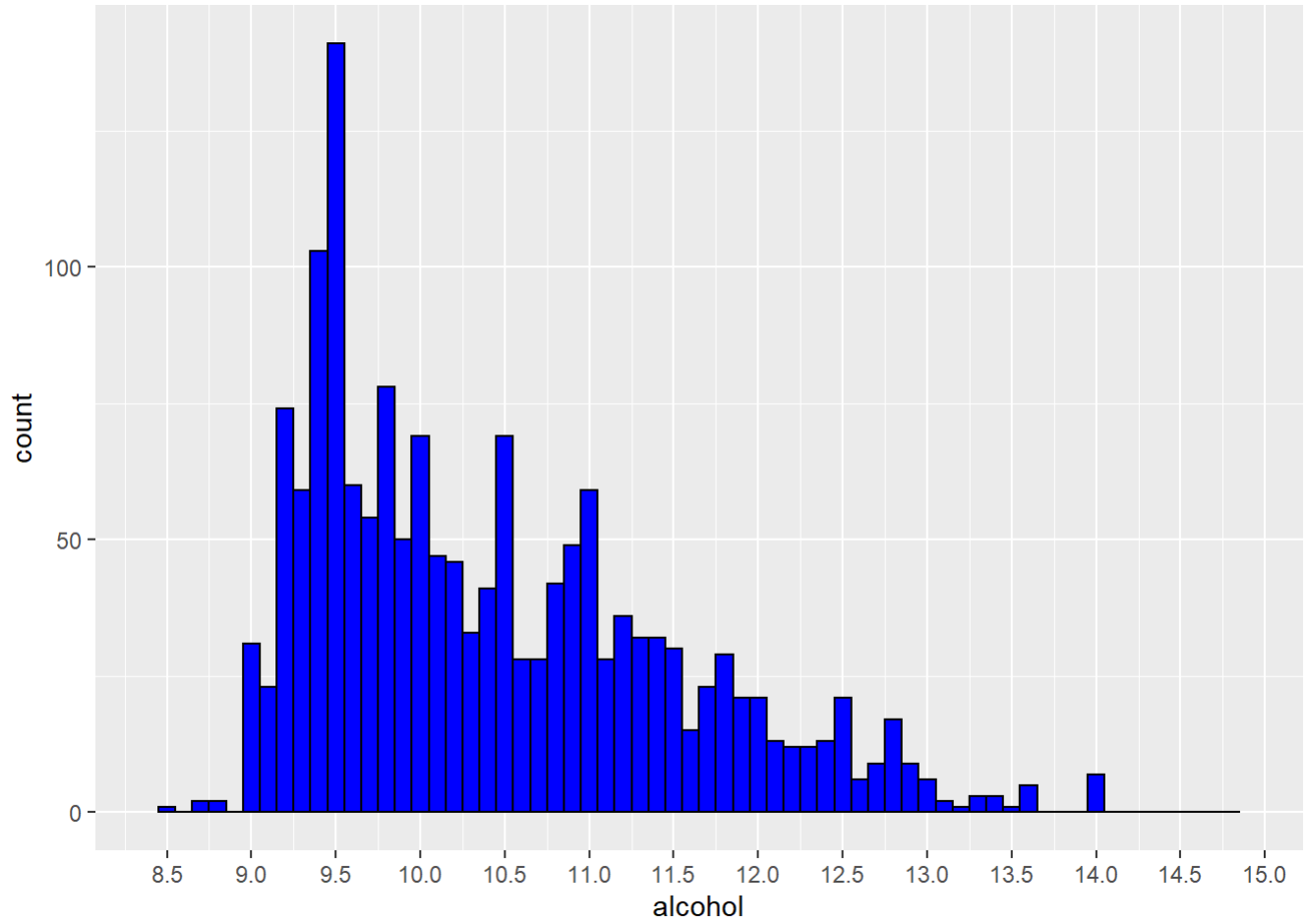
Univariate Plots Section



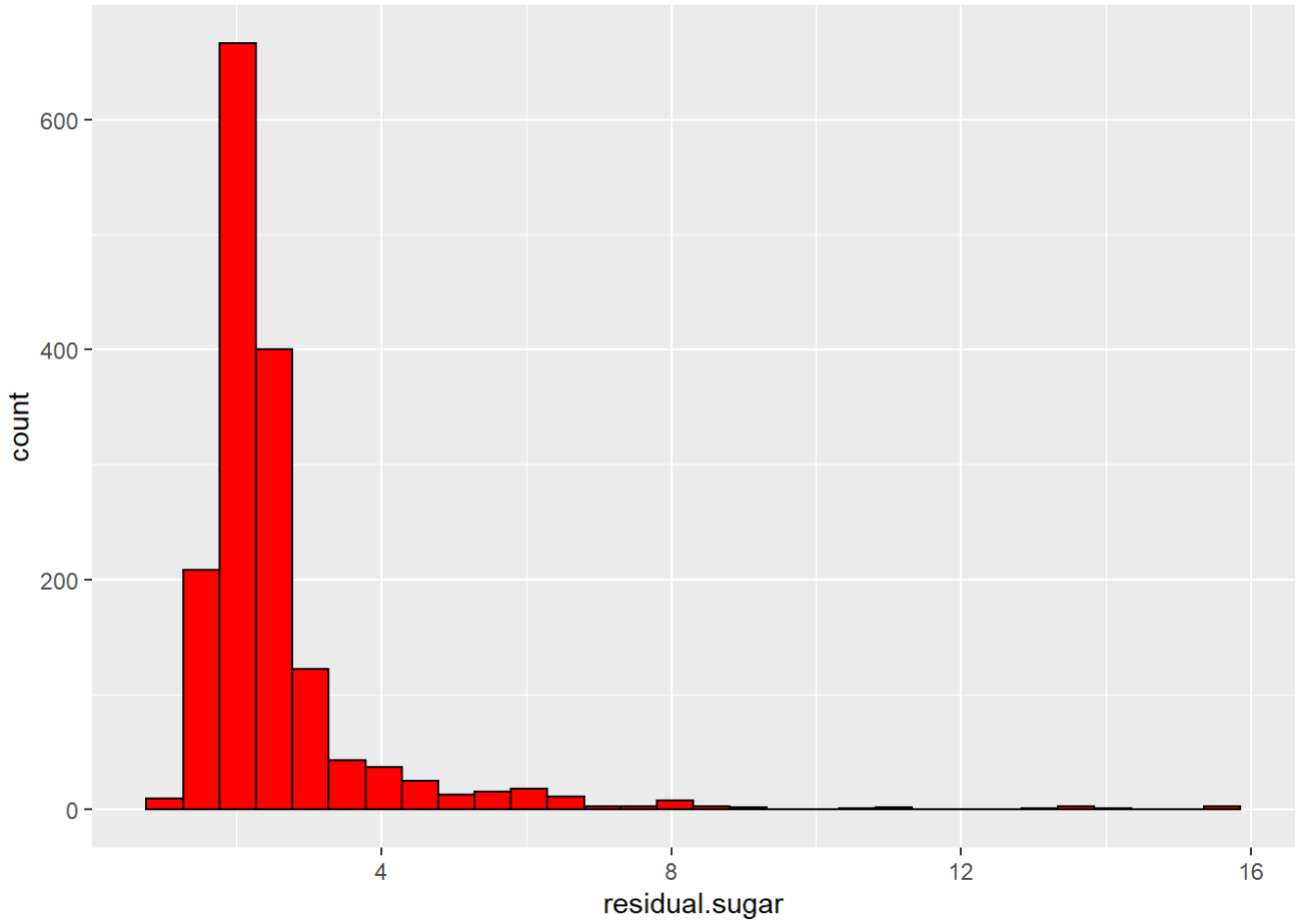
The values of “qualitly” with the highest frequency are 5 and 6.



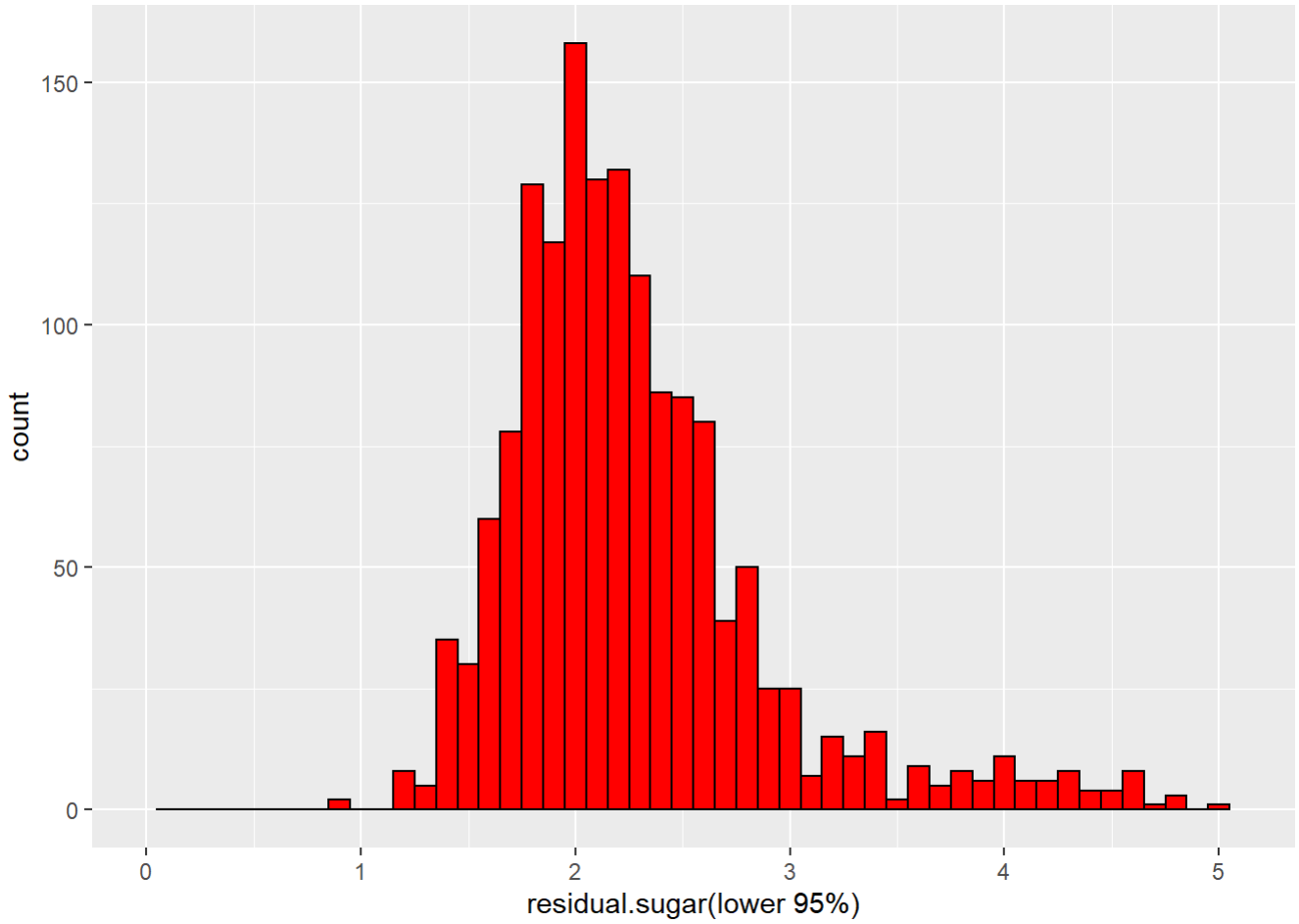
By examining the bar plot above, we can see that the most common “grade” of wine is very good.



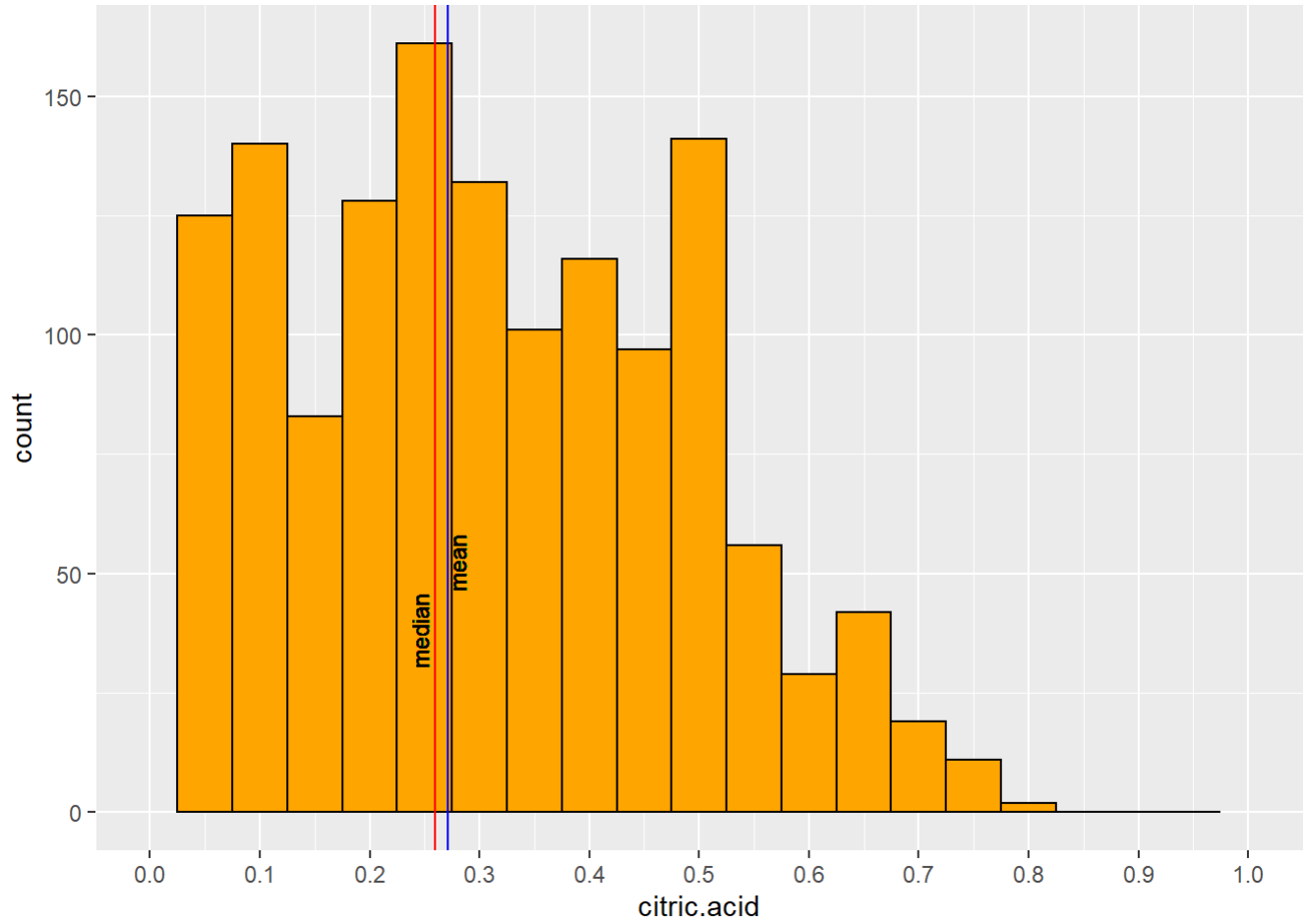
Here we see a negatively skewed histogram for the variable “alcohol”.



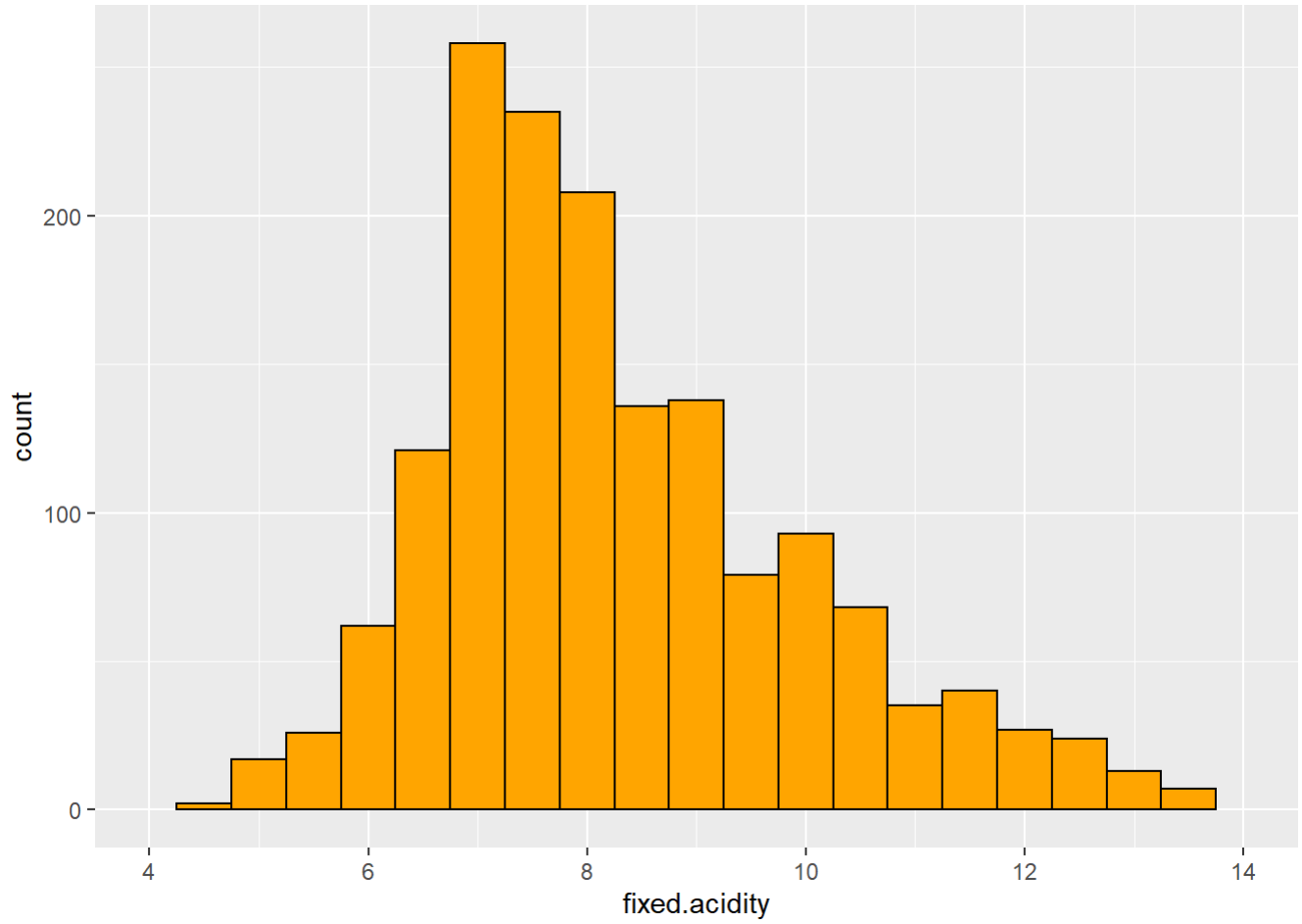
Just like the the plot of alcohol, the “residual.sugar” is negatively skewed. However, this plot represents more of an extreme case due to the large quantity of outliers that are giving the plot its long-tailed distribution. Its also important to notice the steep decline in count.



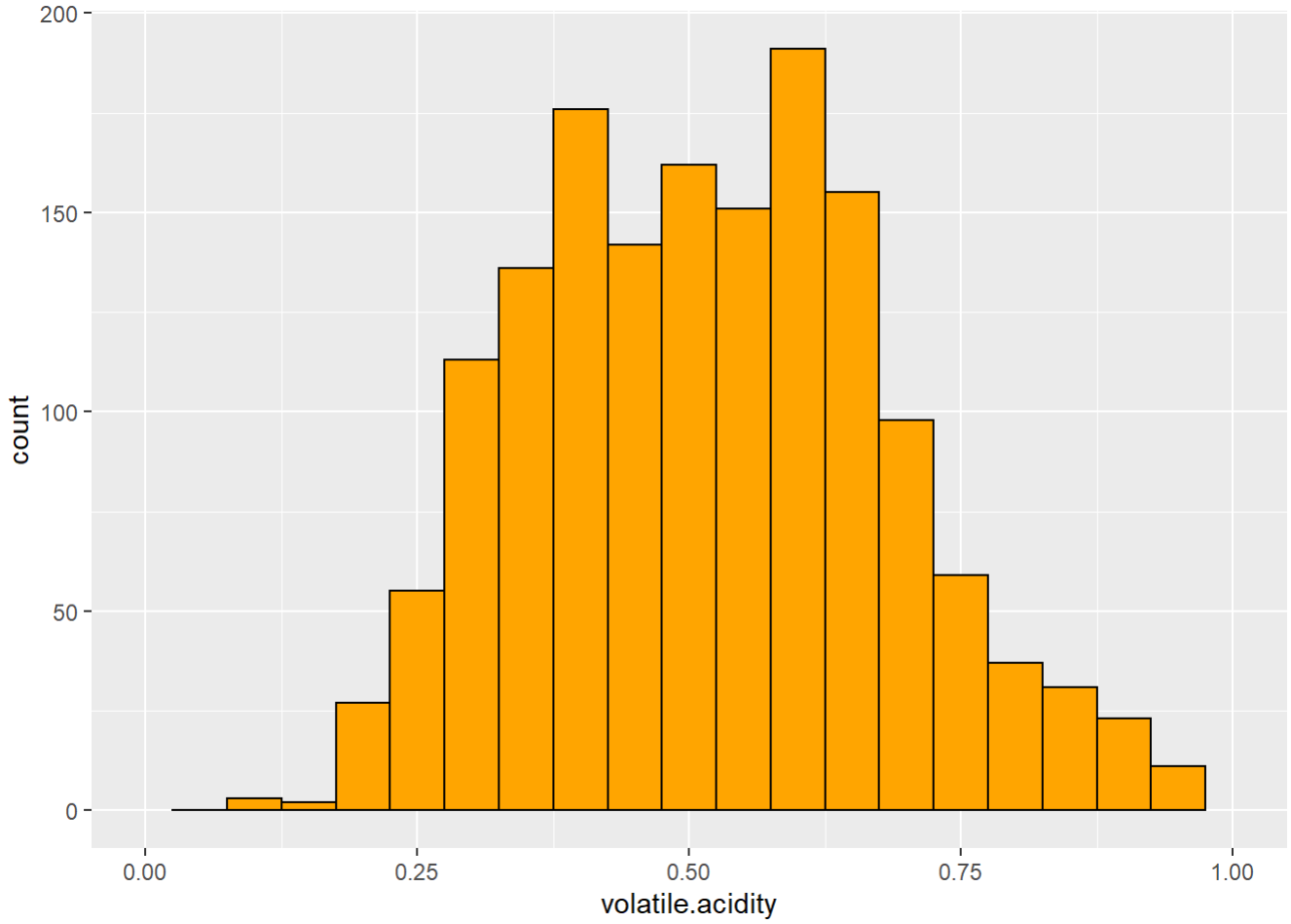
By taking the lower 95% or .95 quantile, we get a more reasonable distribution which allows us to more accurately find trends in the data as it removes any of the unnecessary outliers. Now the plot more closely resembles a normal distribution, which is what we want.



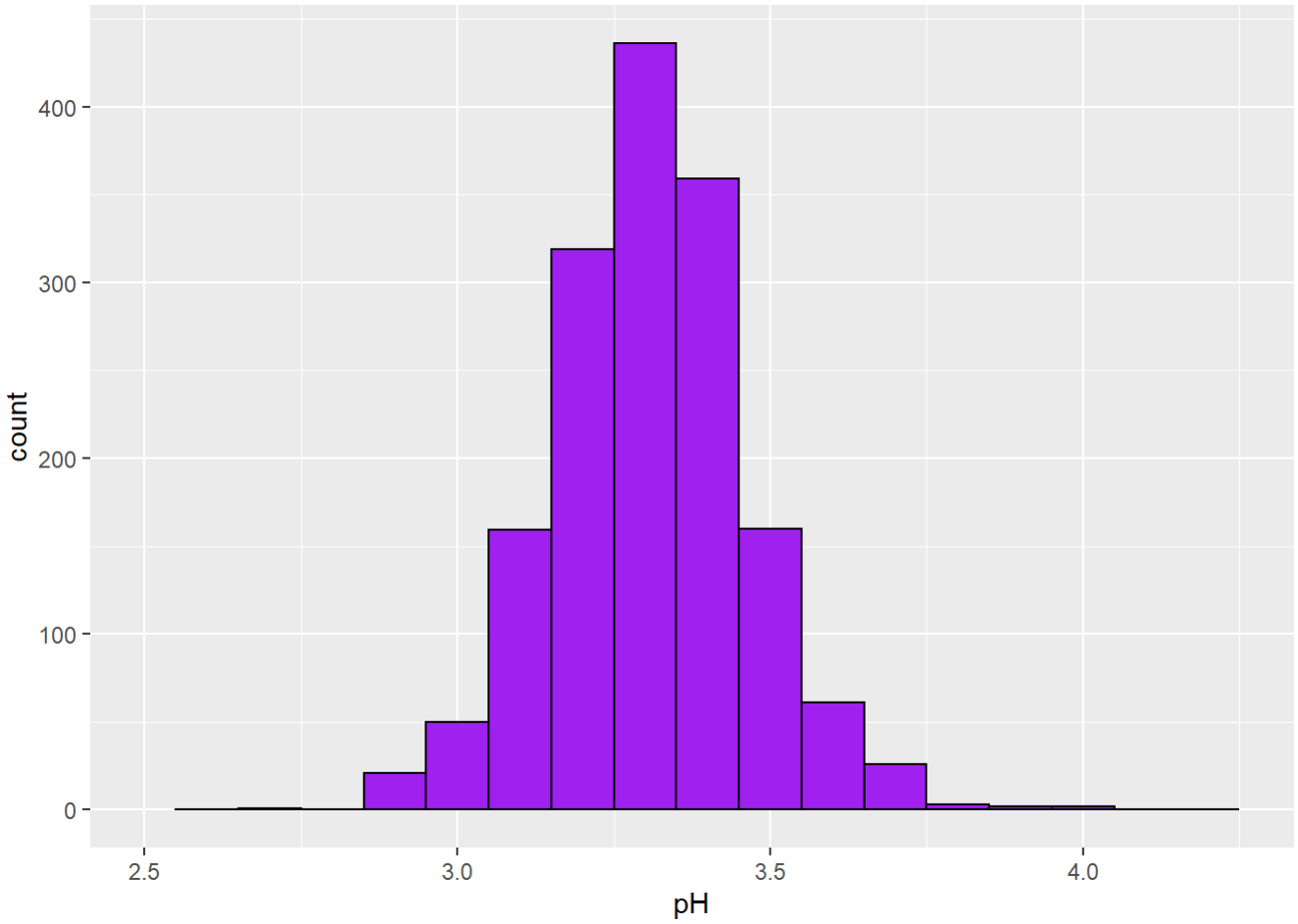
The median and the mean are very close in relation to eachother. This denotes the balancing point or average in the data is near the middle of the values recorded for “citric.acid” in the dataset.



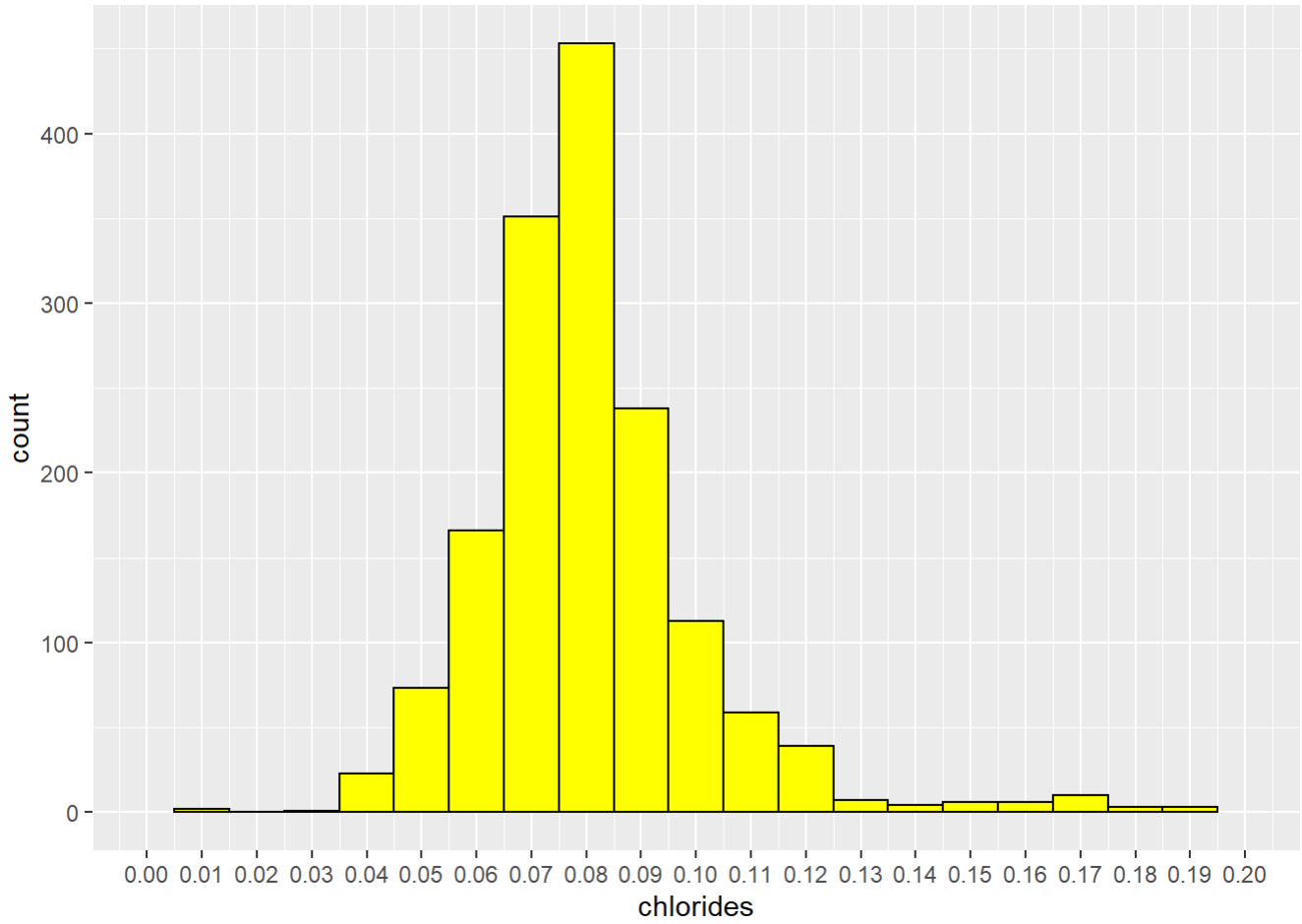
The peak of “fixed.acidity” is around 7 and the plot is very close to a normal distribution.



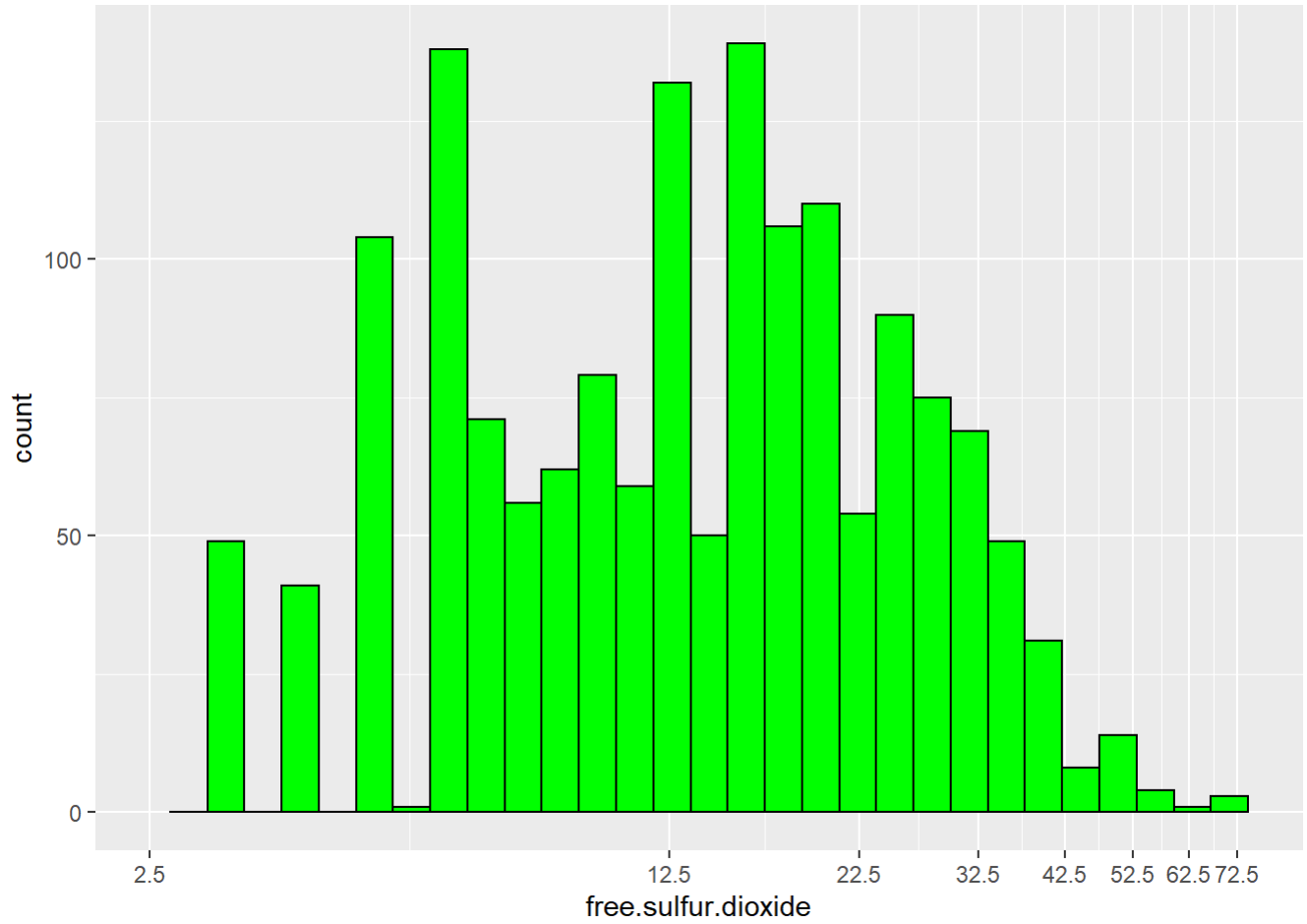
The histogram shows a normal distribution for volatile acidity.



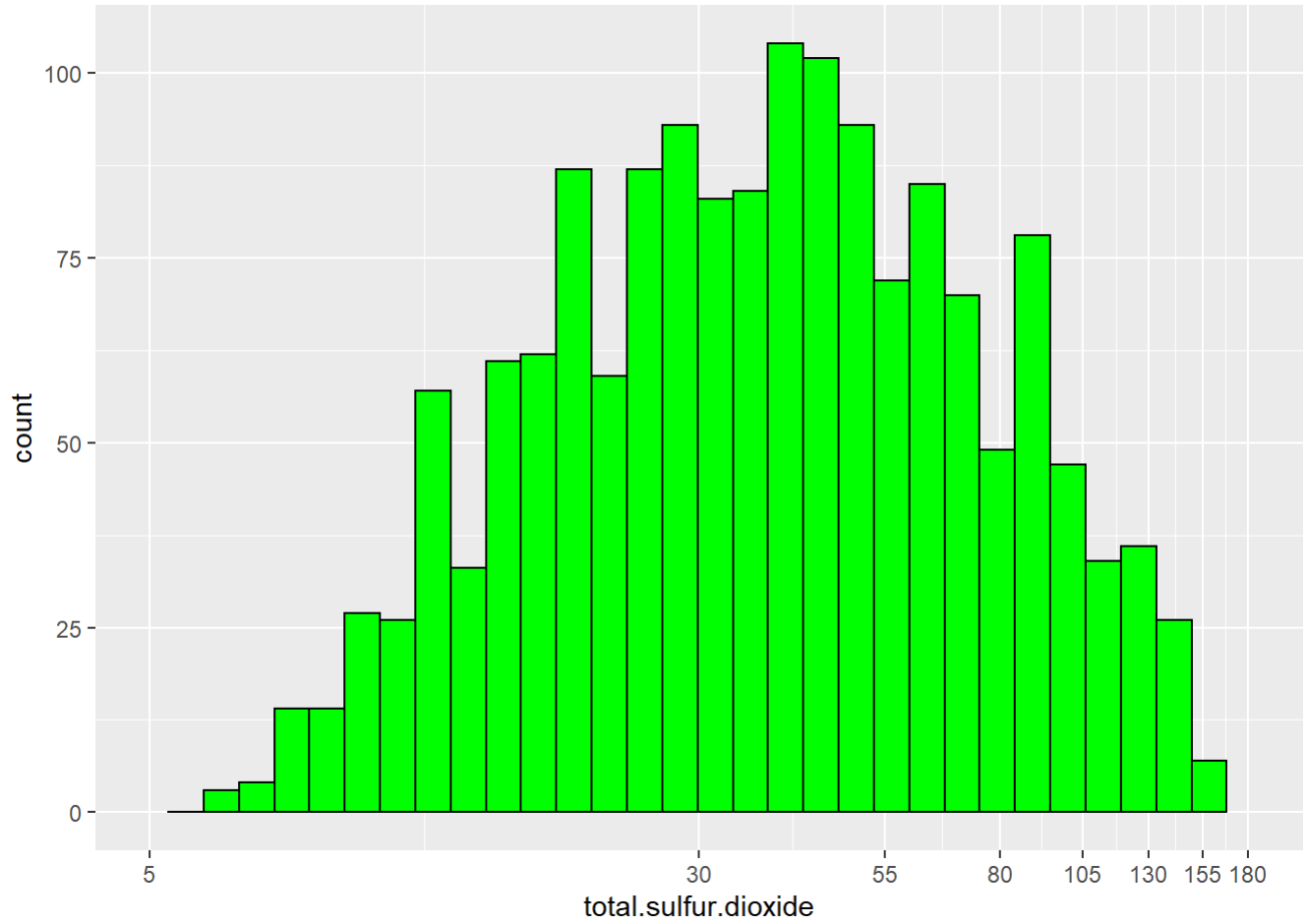
“pH” has a very normal distribution and it is interesting to note that no red wine observed falls into the “basic” category on the pH scale.



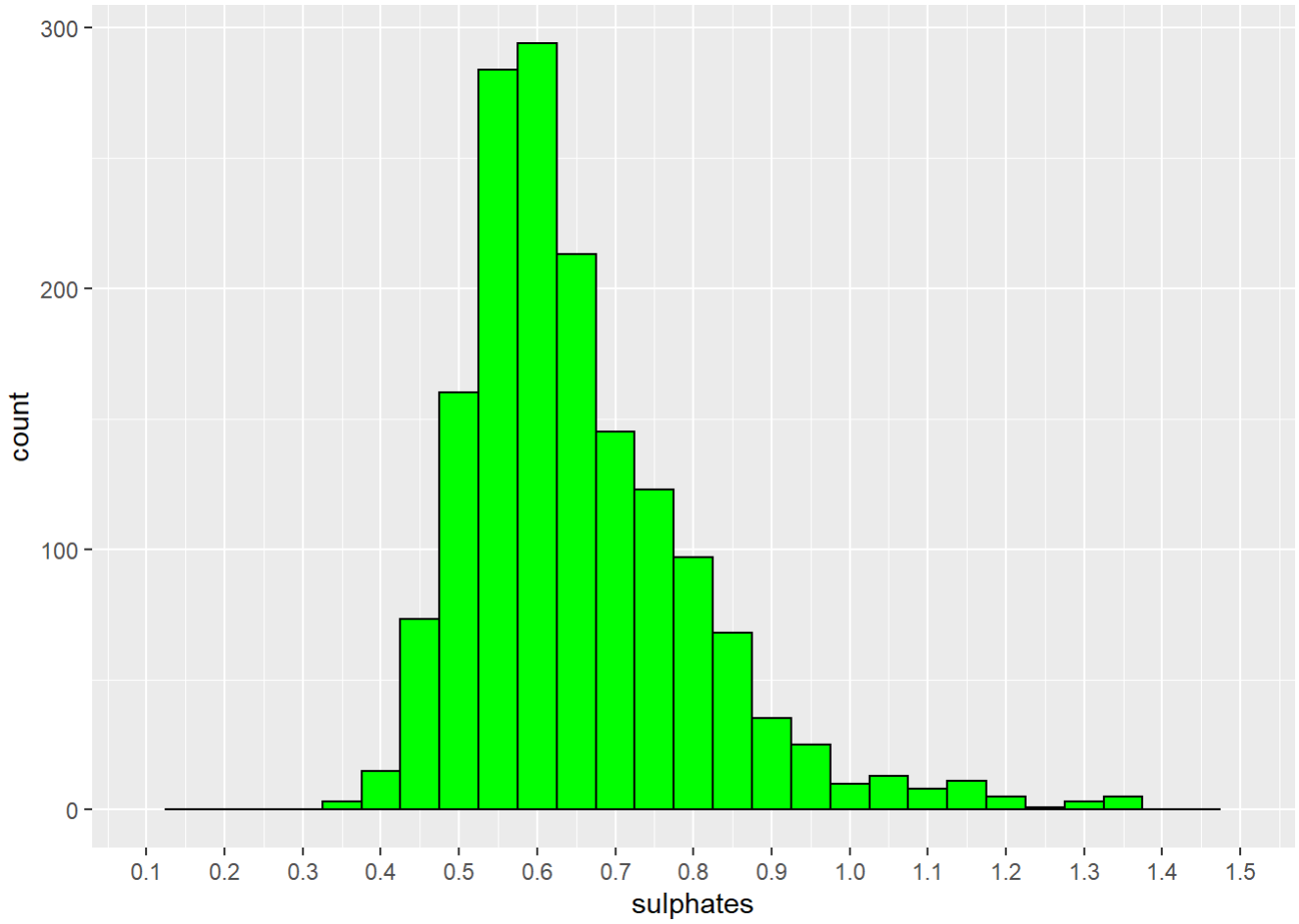
The histogram above uses a different method to eliminate long-tailed data and outliers, than was previously used for the residual.sugar plot.



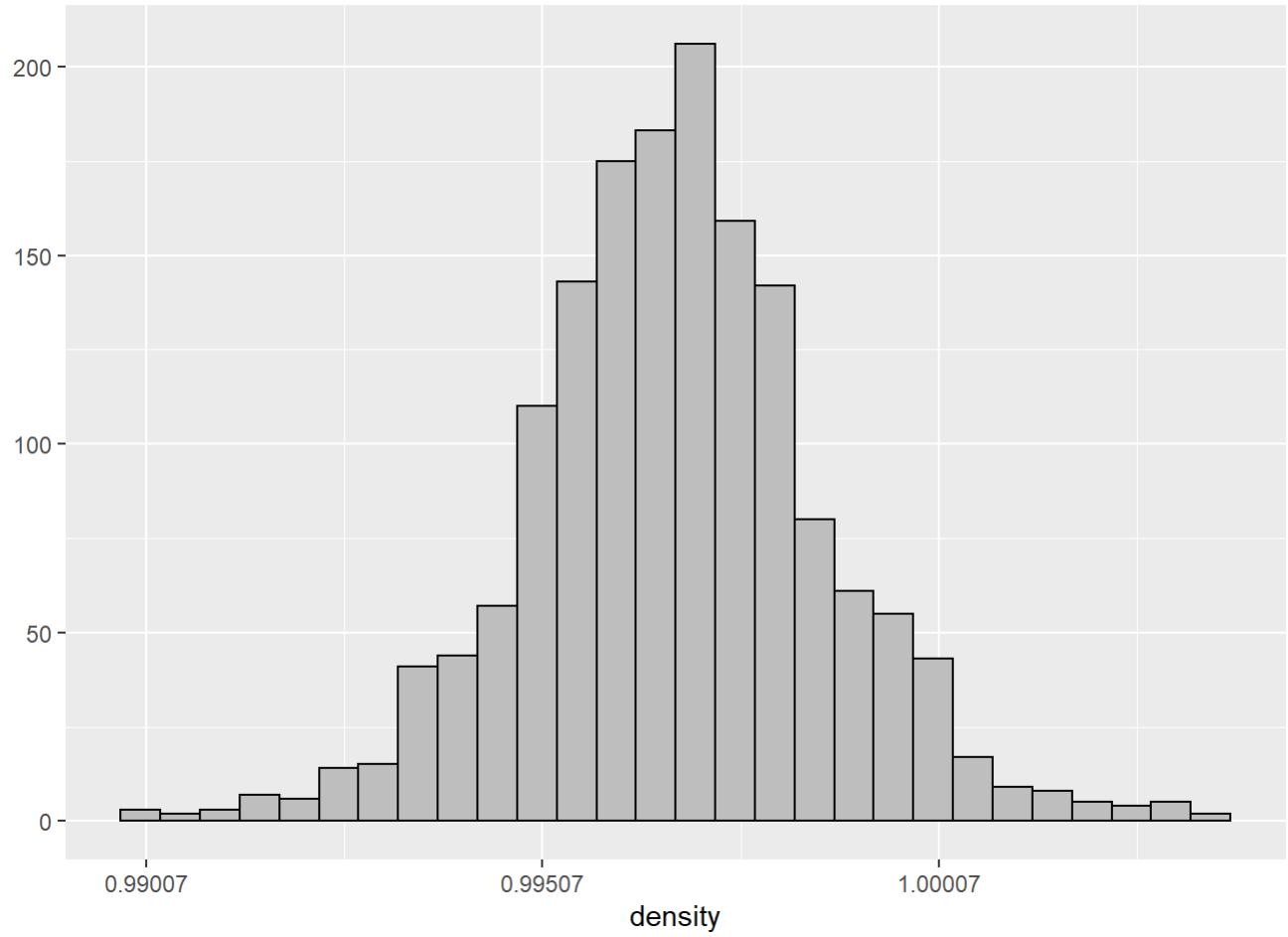
“Free.sulfur.dioxide” was negatively skewed and contained some long-tailed data. The scale_x_log10() is used to give us a better look at the data and uses the logarithm to transform the data closer to a normal distribution.



Similar to “free.sulfur.dioxide”, the histogram for “total.sulfur.dioxide” is tranformed by scale_x_log10 from a negative skew to a normal distribution.



“Sulphates” have a negative distribution with some outliers greater than 1.5, which were excluded from the histogram.



The distribution for “density” is very normal and the range of values found is not large. The values recorded for the “density” don’t differ in any significant manner from one red wine sample to the next.

Univariate Analysis

What is the structure of your dataset?

The dataset consists of 1599 observations of red wine. Each observation provides details regarding a particular red wine’s attributes. There are 13 variables that make up these attributes, they are “fixed.acidity”, “volatile.acidity”, “citric.acid”, “pH”, “residual.sugar”, “chlorides”, “free.sulfur.dioxide”, “total.sulfur.dioxide”, “sulphates”, “alcohol”, “density”, “quality”, and “grade”.

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

I am interested to discover which variables in the dataset contribute most to the “quality” of red wine. In other words, which variables have the strongest correlation to “quality”. I would also like to compare some of the variables to eachother, as to determine their correlation to eachother. For example, I hypothesize that the correlation between the level of acidity(“citric.acid”,“fixed.acidity”,“volatile.acidity”) and “pH” found for an observation would be strong.

What other features will help support the investigation into your interest?

Exploring the differences in the attibutes for the variables linked to outstanding red wine observations vs the rest. For example, if we compare the “residual.sugar” for all the very good red wine against the “residual.sugar” for all the outstanding red wine, what differences would we find? This could provide insight into what sets that “quality” of wine apart from the categories.

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

Yes, I created a new variable called “grade” which would give each observation a grade of mediocre, very good, or outstanding based on the “quality” value for that specific observation. I chose “grade” instead of rating because after some reading I discovered this is the actual terminology and scores that real world wine experts will use when judging and scoring red wines.

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

Yes, I noticed there were some variables such as “residual.sugar” and “chlorides”, that contained an concerning amount of long-tailed data. A large number of values of zero were found for the variable “citric.acid”. Also, there were some variables that included outlier data(“sulphates”, “residual.sugar”), leaving me wondering about the possibilities of inaccurate data entry or external factors that could perhaps contribute to such anomalies.

Did you perform any operations on the data to modify the form of the data?

If so, why did you do this?

Yes, most operations performed were notated directly preceding the visualizations themselves. However, we did perform some adjustments, coloring, transformations, and quantile functions on the data during the Univariate analysis to provide a better look at the data itself. Such operations are performed to get the distributions as close to a normal distrubution as possible. The reason is so that any statistical ananlysis performed on a normal distribution will be much more feasible to calculate, making the results easier to digest.

Bivariate Plots Section

Correlation Coefficient Table

First thing we created a table that would display the correlation of each variable with every other variable in the dataframe. We did this be using the correlation function along with the dplyr package to group the data together.

```
##
##
##                        fixed.acidity   volatile.acidity   citric.acid
## -------------------------- --------------- ------------------ -------------
##     **fixed.acidity**             1              -0.2561         **0.6717**
##    **volatile.acidity**         -0.2561             1            **-0.5525**
##      **citric.acid**           **0.6717**       **-0.5525**          1
##     **residual.sugar**           0.1148            0.001918          0.1436
##       **chlorides**              0.09371            0.0613           0.2038
##  **free.sulfur.dioxide**         -0.1538           -0.0105          -0.06098
##  **total.sulfur.dioxide**        -0.1132           0.07647           0.03553
##        **density**              **0.668**          0.02203         **0.3649**
##           **pH**                **-0.683**          0.2349         **-0.5419**
##       **sulphates**               0.183             -0.261         **0.3128**
##        **alcohol**              -0.06167           -0.2023           0.1099
##        **quality**               0.1241          **-0.3906**         0.2264
##
## Table: Correlation Coefficents accross all variables (continued below)
##
##
##
##                        residual.sugar   chlorides    free.sulfur.dioxide
## -------------------------- ---------------- ------------ ---------------------
##     **fixed.acidity**            0.1148         0.09371           -0.1538
##    **volatile.acidity**         0.001918         0.0613           -0.0105
##      **citric.acid**             0.1436          0.2038          -0.06098
##     **residual.sugar**             1            0.05561            0.187
##       **chlorides**             0.05561            1             0.005562
##  **free.sulfur.dioxide**         0.187          0.005562             1
##  **total.sulfur.dioxide**        0.203           0.0474          **0.6677**
##        **density**             **0.3553**        0.2006           -0.02195
##           **pH**                -0.08565         -0.265            0.07038
##       **sulphates**             0.005527       **0.3713**          0.05166
##        **alcohol**              0.04208          -0.2211          -0.06941
##        **quality**              0.01373          -0.1289          -0.05066
##
## Table: Table continues below
##
##
##
##                        total.sulfur.dioxide     density          pH
## -------------------------- ---------------------- ------------- -------------
##     **fixed.acidity**             -0.1132            **0.668**     **-0.683**
##    **volatile.acidity**           0.07647             0.02203        0.2349
##      **citric.acid**              0.03553           **0.3649**     **-0.5419**
##     **residual.sugar**             0.203            **0.3553**      -0.08565
##       **chlorides**                0.0474             0.2006         -0.265
##  **free.sulfur.dioxide**         **0.6677**          -0.02195        0.07038
##  **total.sulfur.dioxide**            1                0.07127       -0.06649
##        **density**                0.07127                1         **-0.3417**
##           **pH**                  -0.06649           **-0.3417**         1
##       **sulphates**               0.04295              0.1485         -0.1966
```

```
##        **alcohol**                -0.2057          **-0.4962**     0.2056
##        **quality**                -0.1851             -0.1749      -0.05773
##
## Table: Table continues below
##
##
##
##                         sulphates      alcohol       quality
## ------------------------- ------------ ------------ -------------
##     **fixed.acidity**          0.183       -0.06167       0.1241
##    **volatile.acidity**        -0.261       -0.2023     **-0.3906**
##      **citric.acid**         **0.3128**     0.1099        0.2264
##     **residual.sugar**        0.005527      0.04208       0.01373
##       **chlorides**          **0.3713**     -0.2211       -0.1289
##  **free.sulfur.dioxide**      0.05166      -0.06941      -0.05066
##  **total.sulfur.dioxide**     0.04295       -0.2057       -0.1851
##        **density**             0.1485     **-0.4962**     -0.1749
##           **pH**              -0.1966       0.2056       -0.05773
##       **sulphates**              1           0.09359       0.2514
##        **alcohol**            0.09359          1         **0.4762**
##        **quality**             0.2514     **0.4762**          1
```

Now that we have our table, we use some functions inside of the Pander library that allows the tweaking and stylization of a markdown table. We use this to emphasize the variables with a strong correlation(double-asterisk). Obviously a variable is correlated to itself perfectly and so these values were ignored in the criteria when emphasizing strong correlation coefficients.

Below is a break down of the noteworthy coefficents with their correlation type and value.

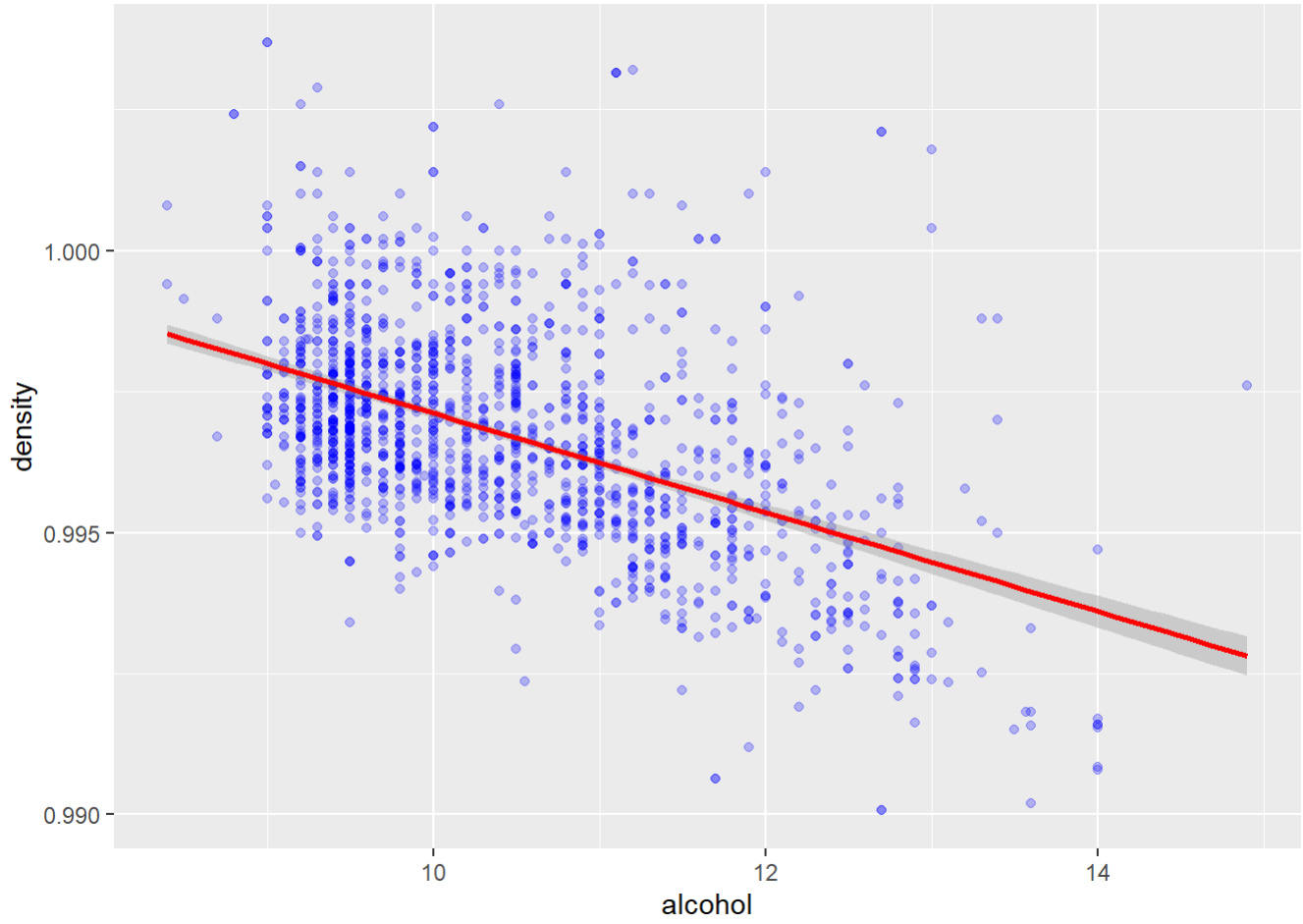
- **Positive Correlation**(For every positive increase in one variable, there is a positive increase of a fixed amount in the other.):
    - “fixed.acidity” & “density” (0.668)
    - “free.sulfur.dioxide” & “total.sulfur.dioxide” (0.6677)
    - “fixed.acidity” & “citric.acid” (0.6717)
    - “alcohol” & “quality” (0.4762)
    - “chlorides” & “sulphates” (0.3713)
    - “citric.acid” & “density” (0.3649)
    - “residual.sugar” & “density” (0.3553)
    - “citric.acid” & “sulphates” (0.3128)
- **Negative Correlation**(For every positive increase in one variable, there is a negative decrease of a fixed amount in the other.):
    - “fixed.acidity” & “pH” (-0.683)
    - “volatile.acidity” & “citric.acid” (-0.5525)
    - “citric.acid” & “pH” (-0.5419)
    - “density” & “alcohol” (-0.4962)
    - “volatile.acidity” & “quality” (-0.3906)
    - “pH” & “density” (-0.3417)

Coefficients closest to zero(not greater than 0.3) were not included in the list due the lack of correlation.

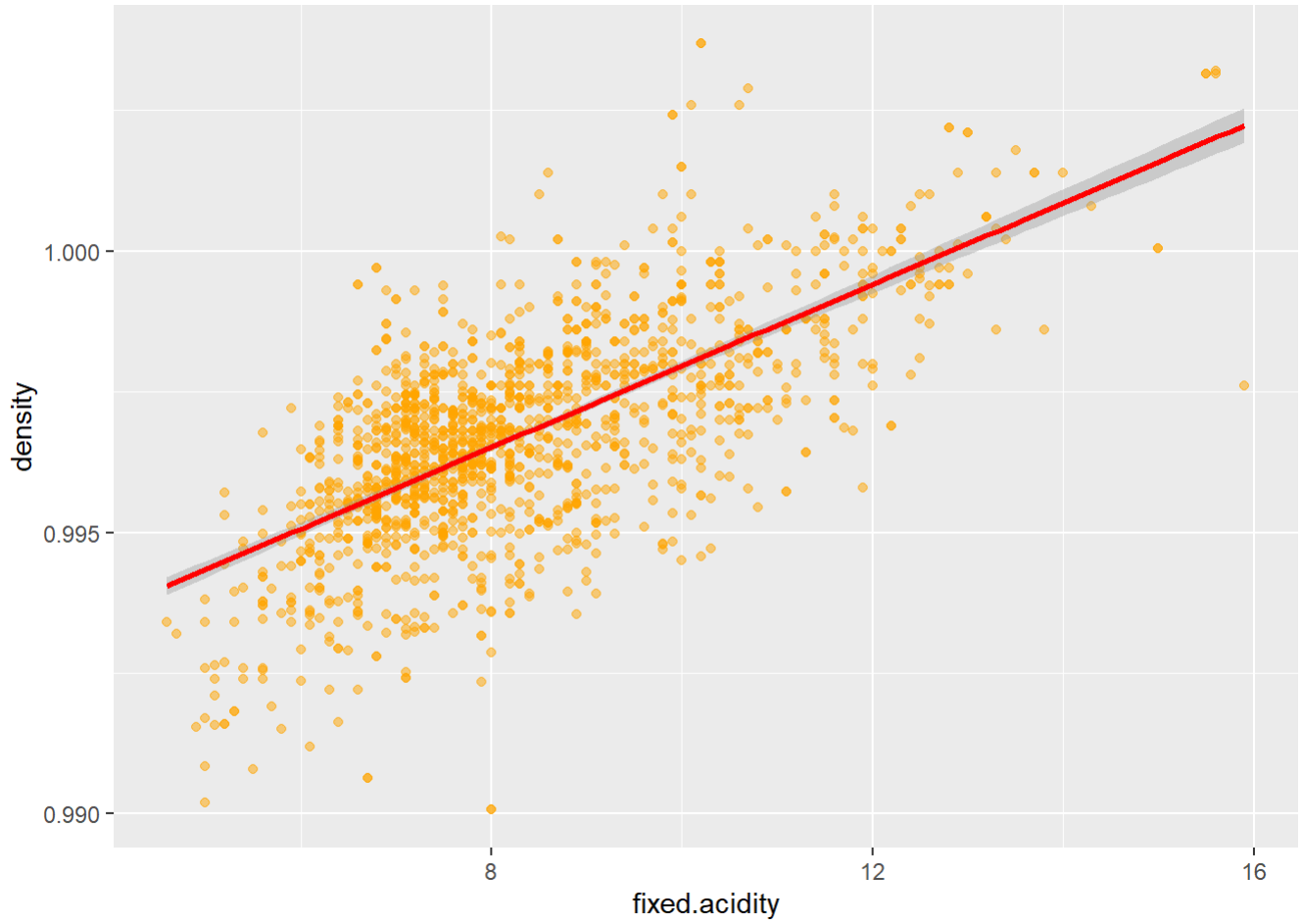
Scatterplots

```
## [1] 0.001887334
```

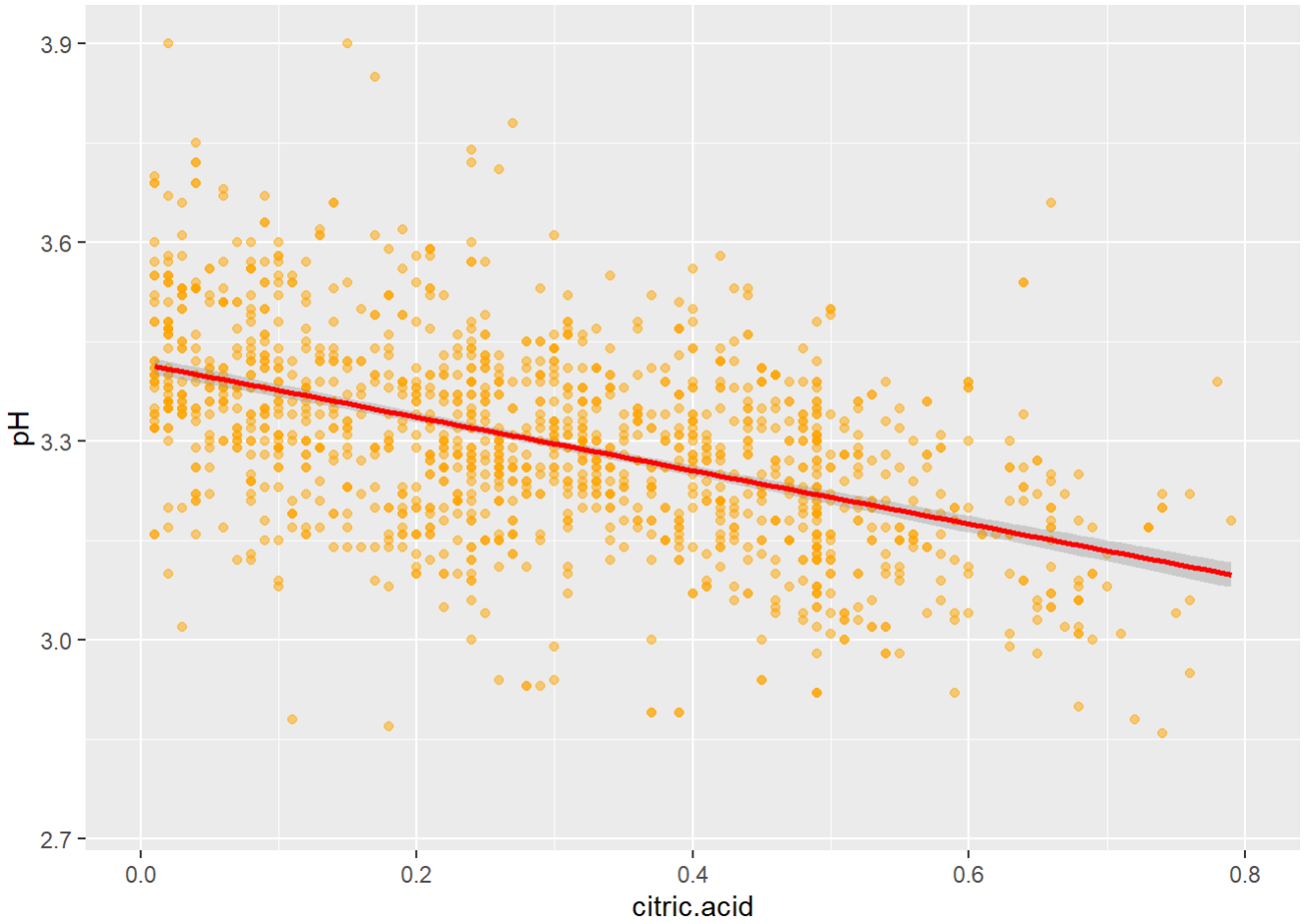
The decimal above represents the standard deviation for the values of “density”, to me I found this very interesting that even though the standard deviation is so small for “density”, a strong correlation bewteen it and “fixed.acidity” was still found.



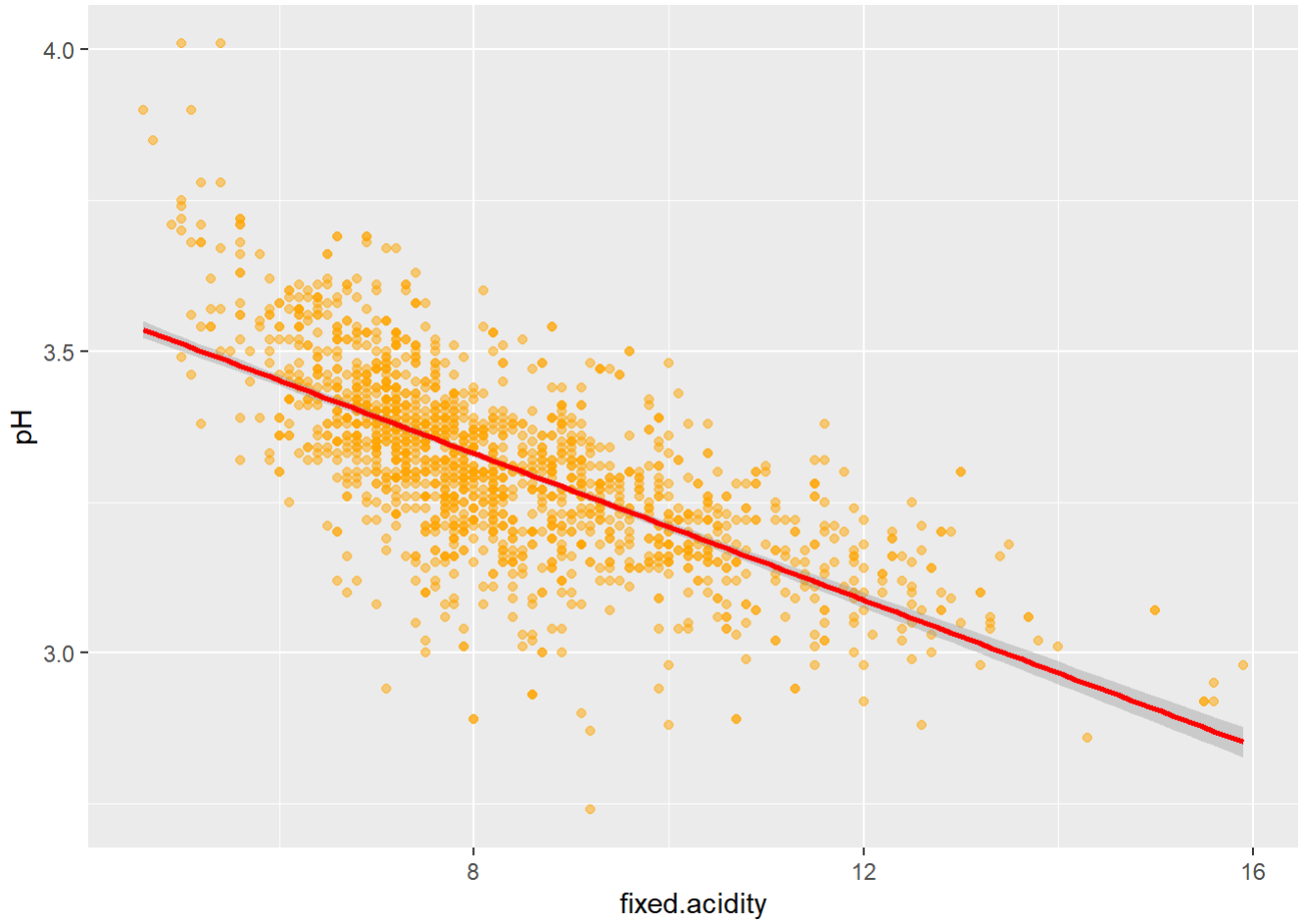
The negative correlation between “density” and “alcohol” is shown in the scatterplot above.



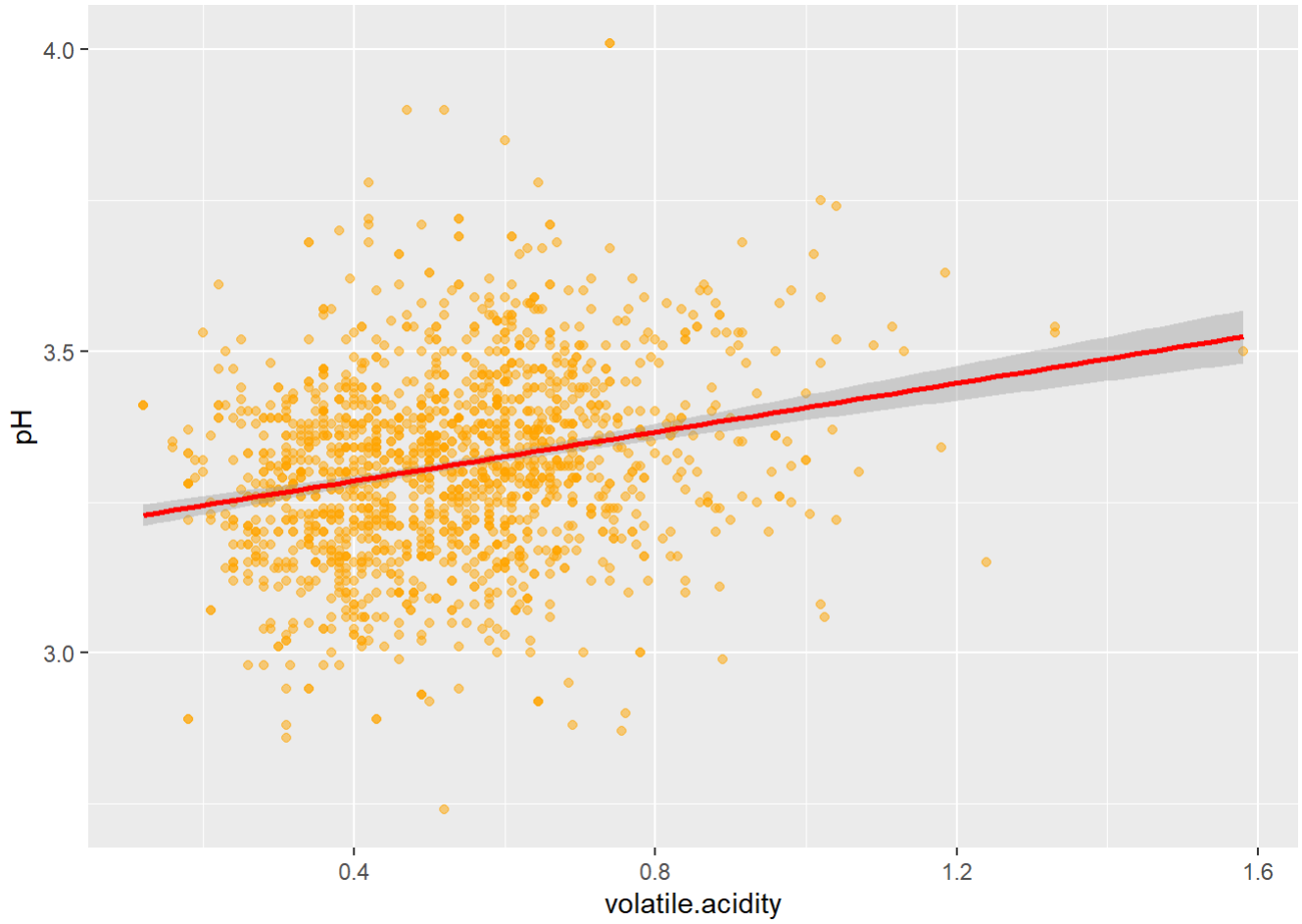
There is a very strong correlation between “fixed.acidity” and “density”.



The scatterplot above provides a visualization of the relationship between “pH” and “citric.acid”, which is a strong negative correlation.



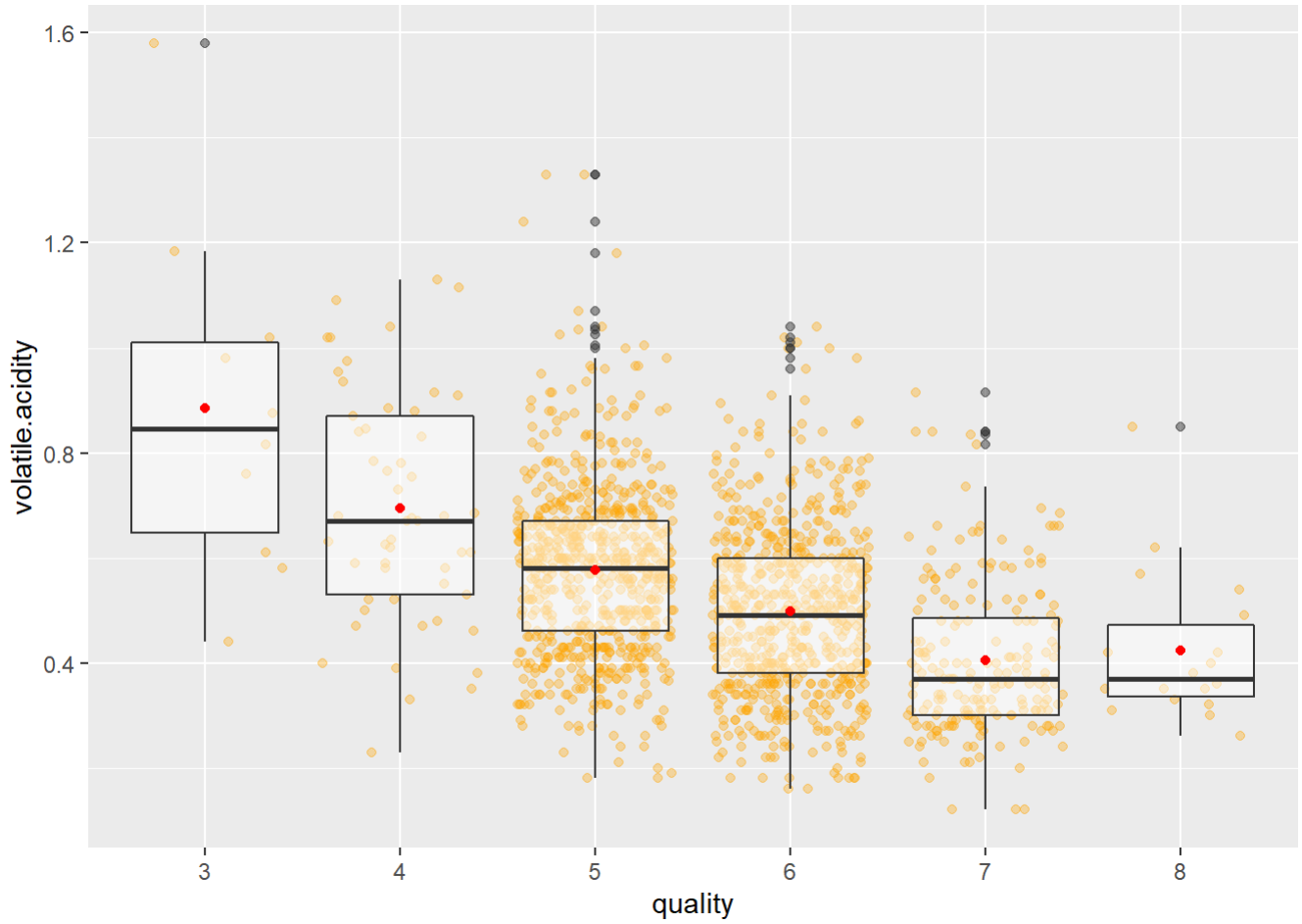
The scatterplot above provides a visualization of the relationship between “pH” and “fixed.acidity”, which is a strong negative correlation.



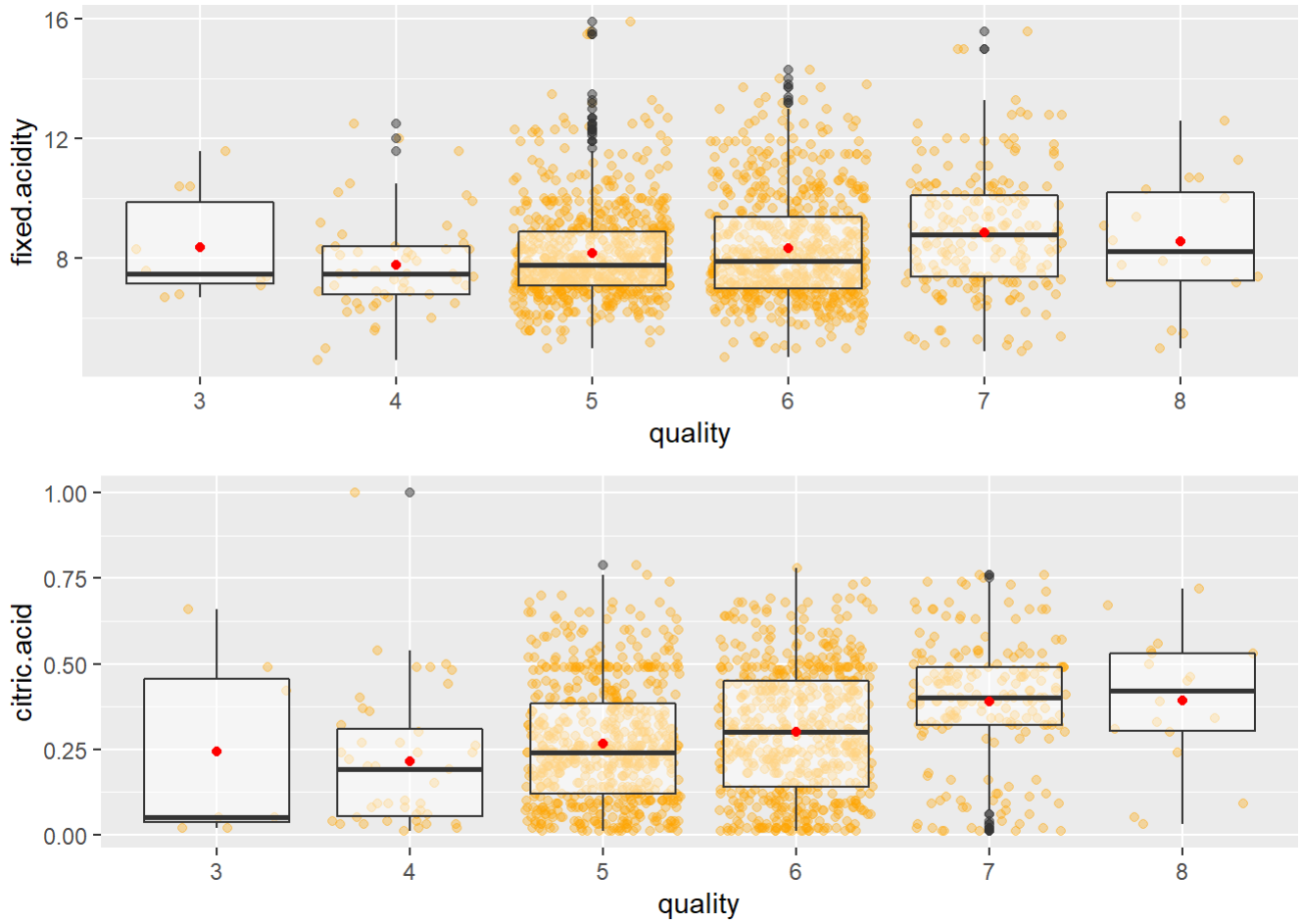
The scatterplot above provides a visualization of the relationship between “pH” and “volatile.acidity”. Based on the scatterplot above, the correlation shows an exact opposite of the previous two plots, which is a strong positive correlation. Why?

Box plots

Now we can take a look into the main investigation of finding out which variables most influence the quality of wine. Based on the correlation table above we find that the two variables with the strongest correlation coefficient with “quality” are “volatile.acidity” and “alcohol”. Let’s use some box plox to further investigate.



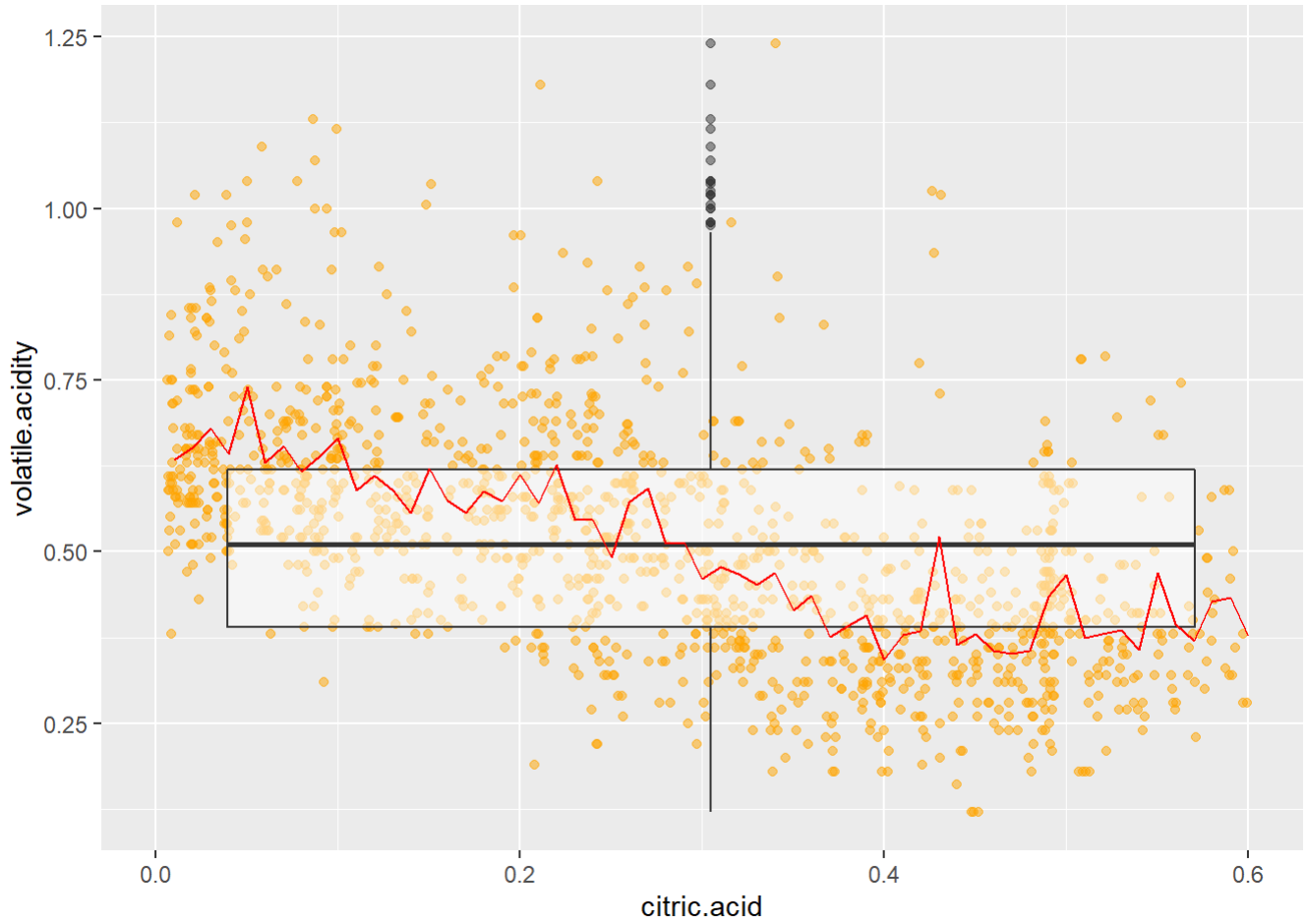
The boxplot aboves shows the negative correlation between “volatile.acidity” and “quality”. The quality of wine declines as the “volatile.acidity” increases, but why? Let’s use more box plots to take a look at the other acidic components, and investigate how they affect “quality”.



Above we see two seperate box plotsf for “fixed.acidity” and “citric.acid” correlating to “quality”.

In the first boxplot we can see that there is not a significant correlation between “fixed.acidity” and “quality”.

However, in the second boxplot we can see a positive impact on the quality of red wine’s when they contain more “citric.acid”. Due to these results, it’d be beneficial to look at the relationship with “citric.acid” & “volatile.acid”, as one positively affects the quality of red wine while the other negatively affects it.



The scatterplot above shows that the mean for “volatile.acidity” declines as “citric.acid” increases.

To better understand “volatile.acidity”, let’s backtrack to the scatterplot of this variable and “pH” to determine whether or not Simpson’s Paradox is the cause of the increase of “pH” with the increase of “volatile.acidity”.

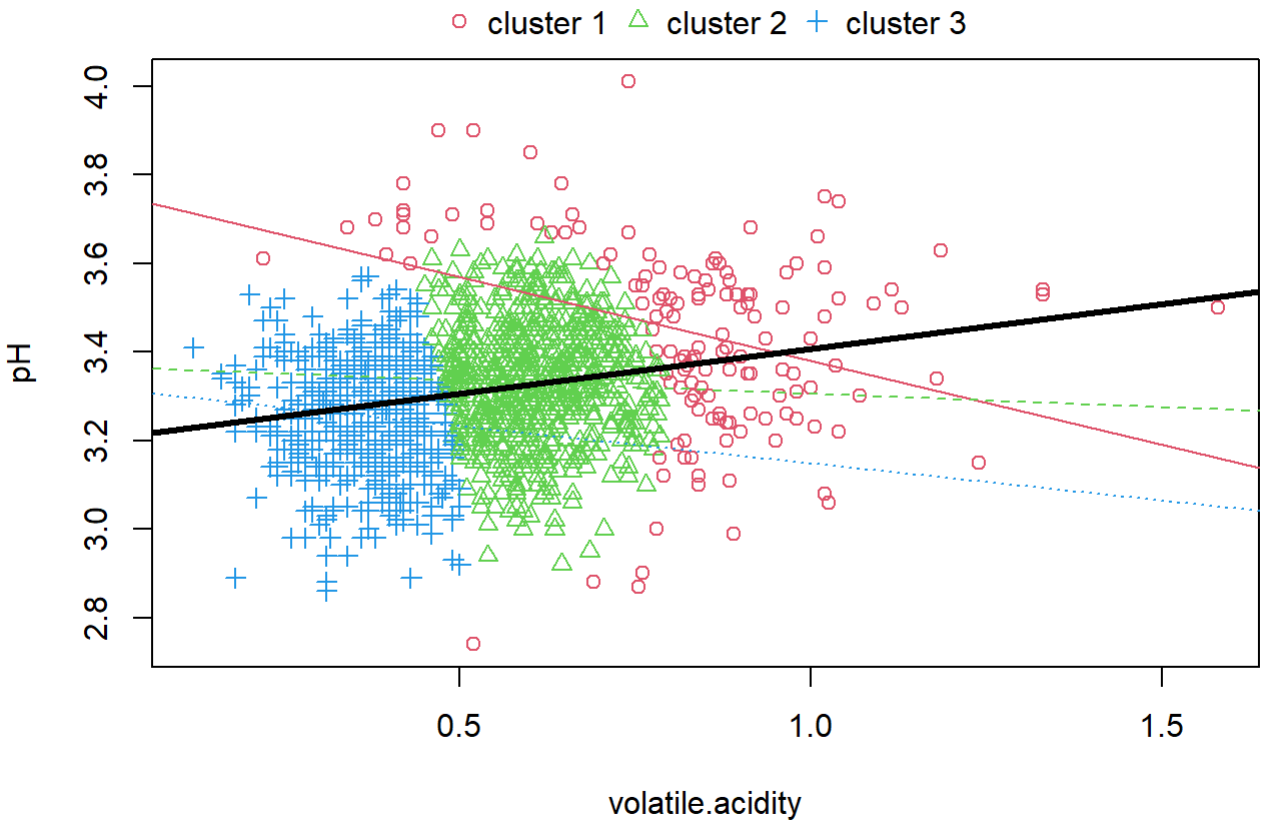


fig. 1.1

The solid black line represents the regression coeffecient of “volatile.acidity” and “pH” found at the level of the whole dataset, while the solid red, dashed green, and dotted blue lines represent the regression found at the level of the three subpopulations or clusters.

The regression signs are significantly different and in the opposite direction compared to the group(dataset), Simpson’s paradox has been successfully detected in this instance.

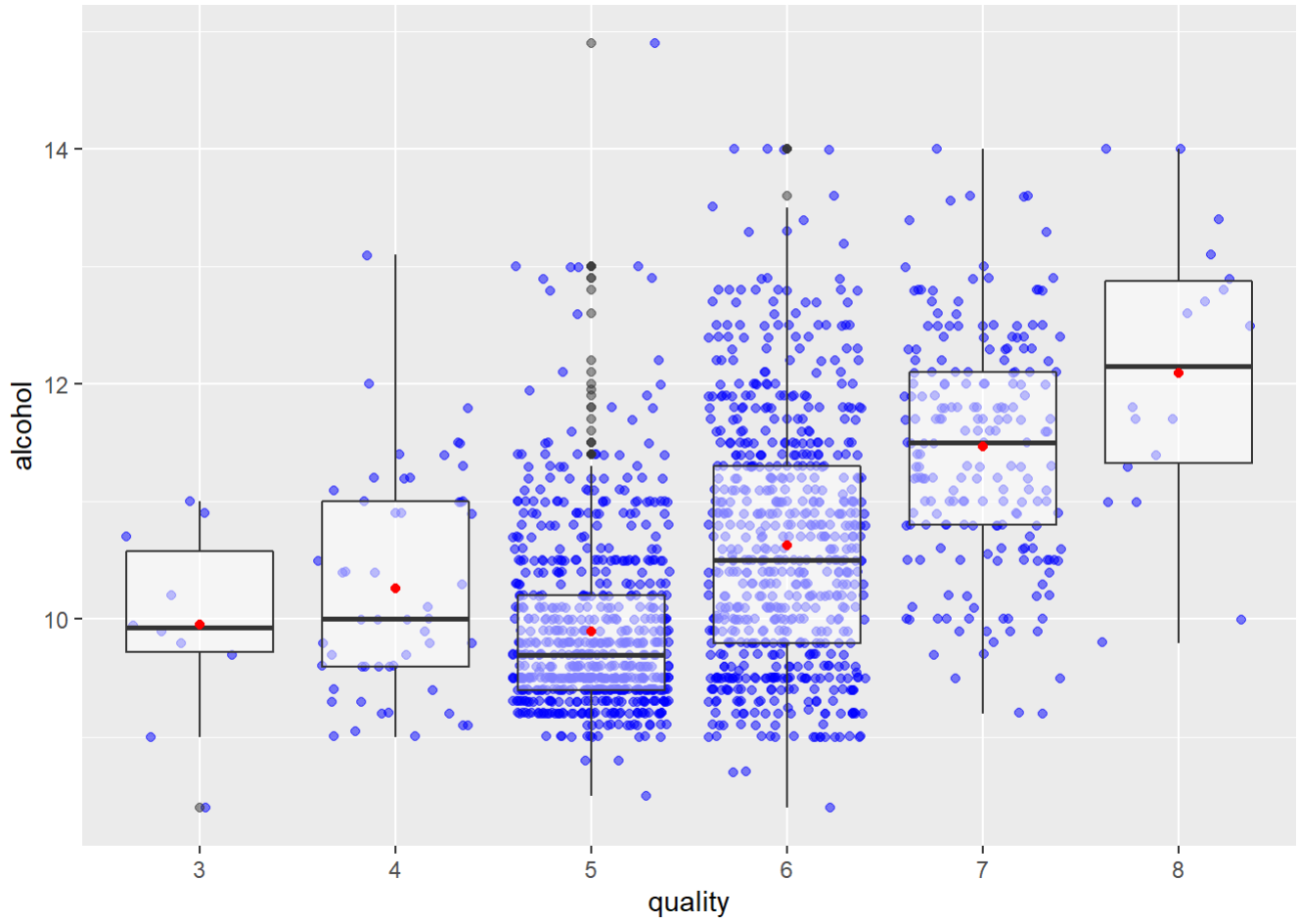


fig. 1.2

The boxplot demonstrates that there is a strong positive correlation between the “quality” of red wine and the amount of “alcohol” that it contains. However, I notice that there is a lot of overplotting which suggests that perhaps alcohol only contributes a portion to the overall quality in red wine.

Bivariate Analysis

Talk about the relationships you observed in this part of the investigation.

How did the feature(s) of interest vary with other features in the dataset?

The relationships observered in the bivariate analysis were the variables that correlated most to the “quality” of red wine. First we conducted correlations tests to analyze each variables relation to the main feature of interest. This led to the following discoveries:

- Observered that “fixed.acidity” has almost no correlation to “quality”.
- Observered that “quality” of red wine improves when larger amounts of “citric.acid” are found.
- Observered a strong negative correlation between “volatile.acidity” and “quality”.
- Observered a strong positive correlation between “alcohol” and “quality”.
- Observered that “residual.sugar” held hardly any significant affect on the “quality” of red wine.

After determining the two variables with the strongest positive and negative correlation coefficient, more analysis was performed to better understand these variables.

Throughout the exploration of the analysis, discoveries were made that proved certain presumptions about the variables true. I had a presumption that the “density” of the red wine would increase the less amount of “alcohol” it contained. This is due to my prior knowledge that certain compounds are less dense than water(ice cubes, vegetable oil, and alcohol).

However, presumptions about other variables relationships to eachother or their impact on the “quality” were found to be false. I was shocked and surprised to find out that my presumption that “volatile.acidity” would correlate with “pH” and that as one variable would increase the other would decrease.

Did you observe any interesting relationships between the other features

besides the main feature of interest?

Yes, other interesting relationships were observed. For example, during the bivariate plotting of “volatile.acidity” and “pH” I was expecting specific results and became confused when the data showed the contrary. To better understand this relationship a test was performed to evaluate whether or not Simpson’s Paradox could be detected as a cause.

Above our Simpsons Paradox’s plot(see fig. 1.1) generated three clusters to measure the regression for “volatile.acidity” and “pH” and determine if the regression line was in direct opposition to the regression line across the entire dataset.

The solid black line represents the regression found at the level of the whole dataset, while the solid red, dashed green, and dotted blue lines represent the regression found at the level of the three subpopulations or clusters.

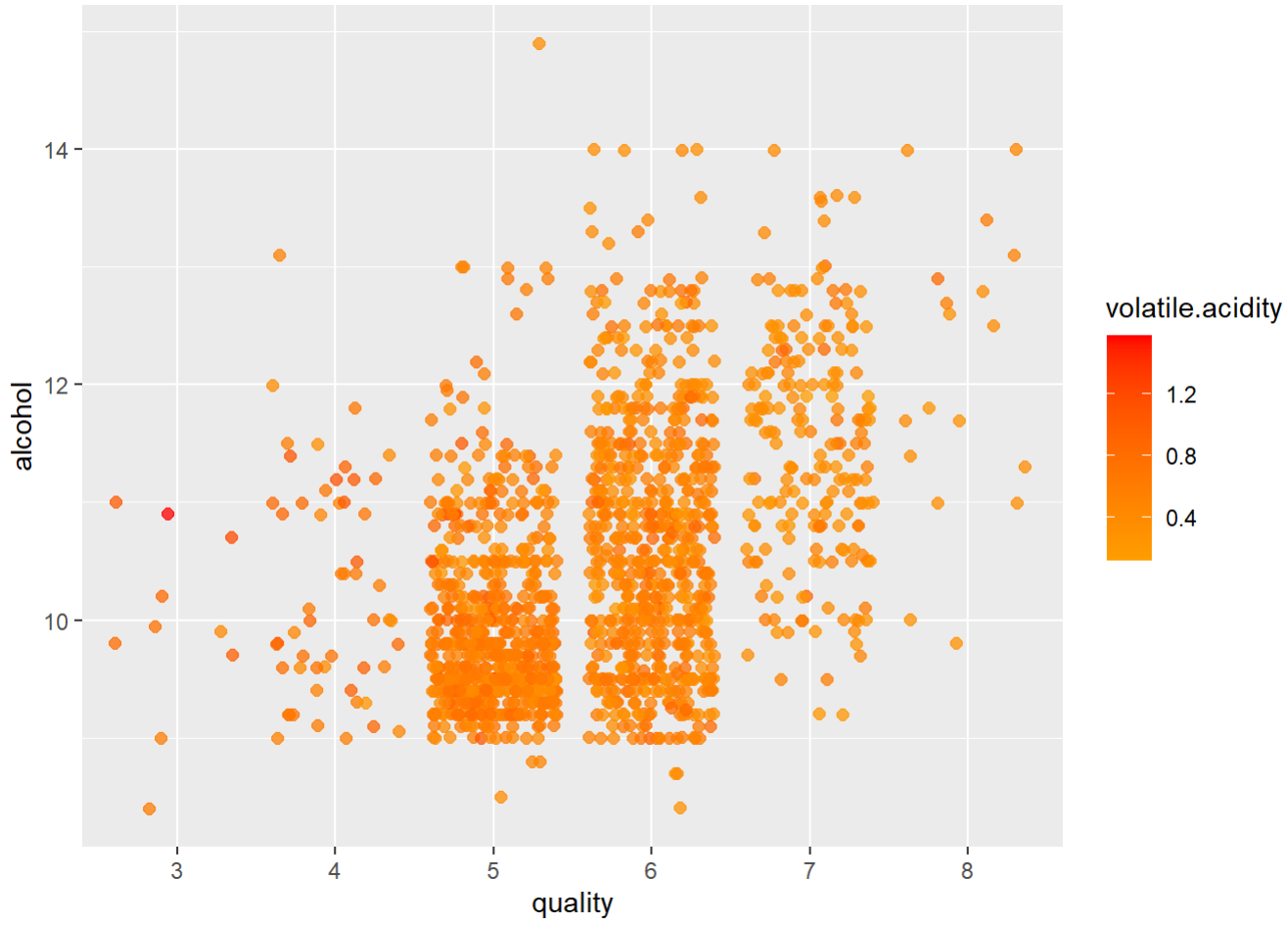
Because the regression signs are significantly different and in the opposite direction compared to the group(dataset), we know that “volatile.acidity” is being influenced by a lurking variable, successfuly detecting Simpson’s Paradox as the cause.

The correlation with “citric.acid” and “volatile.acidity” was also observered and it was discovered that there was a strong negative correlations between the two variables.

What was the strongest relationship with quality observered?

The strongest positive correlation with “quality” observered was “alcohol”. The strongest negative correlation with “quality” observered was “volatile.acidity”.

Multivariate Plots Section



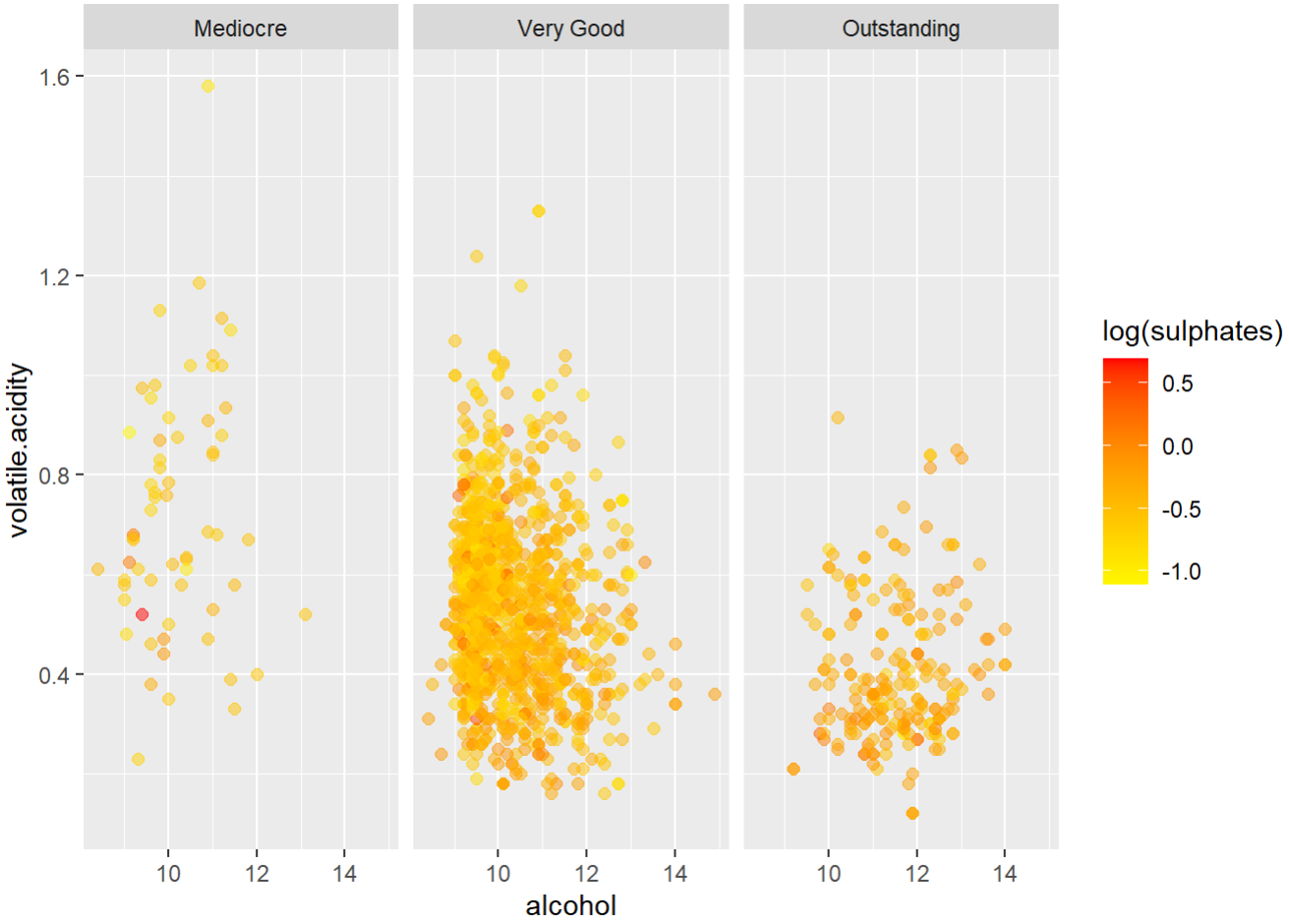
According to thehe multivariate scatterplot above we can see that the “quality” of red wine will be graded higher with a higher level “alcohol” while decreasing the levels of “volatile.acidity” as much as possible.

```
##
## Call:
## lm(formula = as.numeric(quality) ~ alcohol, data = red_wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8442 -0.4112 -0.1690  0.5166  2.5888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.12503    0.17471  -0.716    0.474
## alcohol      0.36084    0.01668  21.639   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7104 on 1597 degrees of freedom
## Multiple R-squared:  0.2267, Adjusted R-squared:  0.2263
## F-statistic: 468.3 on 1 and 1597 DF,  p-value: < 2.2e-16
```

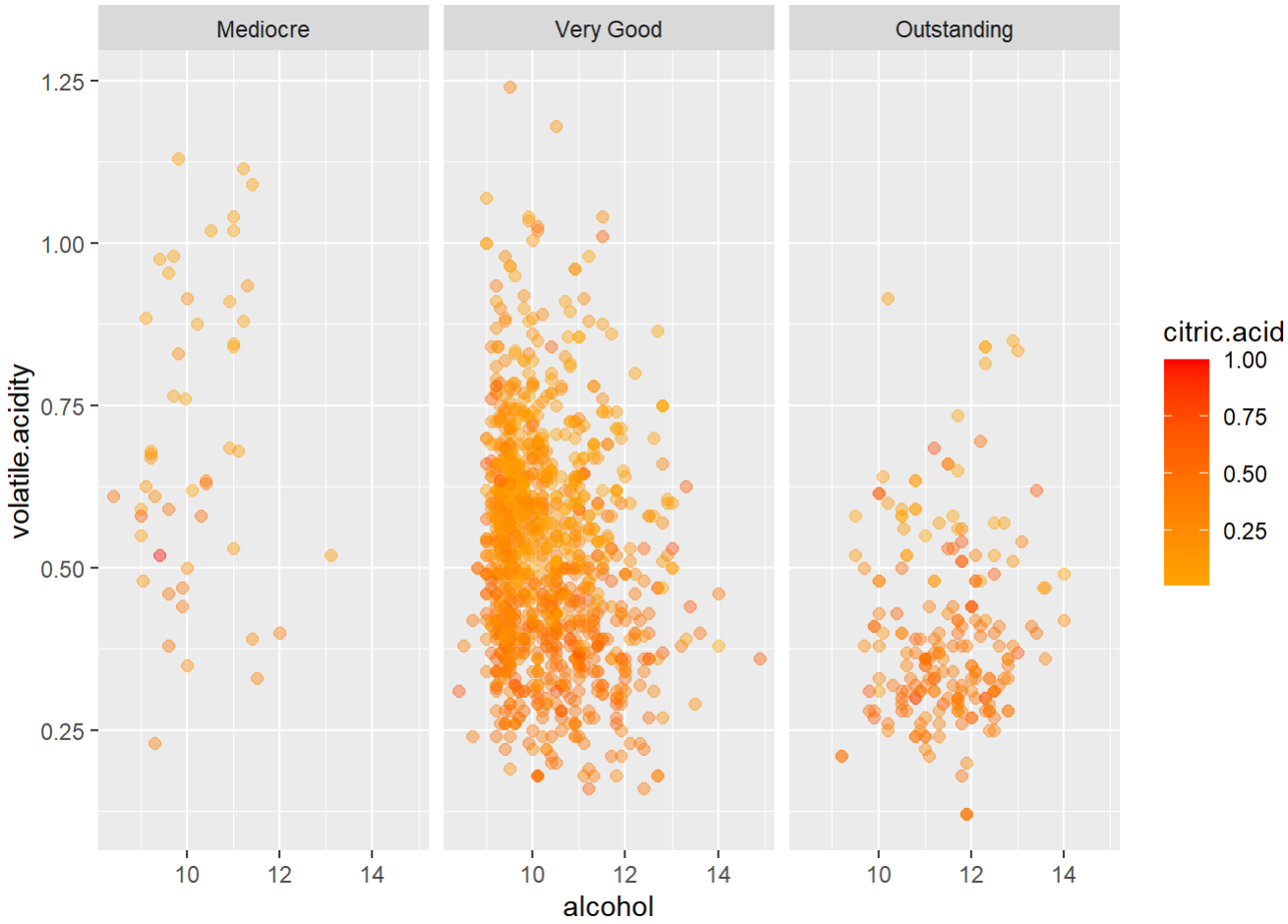
The summary above provides statistical information of a linear model based on the “quality” and “alcohol” variables. During the bivariate section a good amount of outlier data was prevalent within their boxplot(see fig. 1.2).

In order to determine the portion contributed by alcohol to the overall quality in red wine, we create a linear module of the two variables, and then take the value for multiple R-squared which is 0.2267.

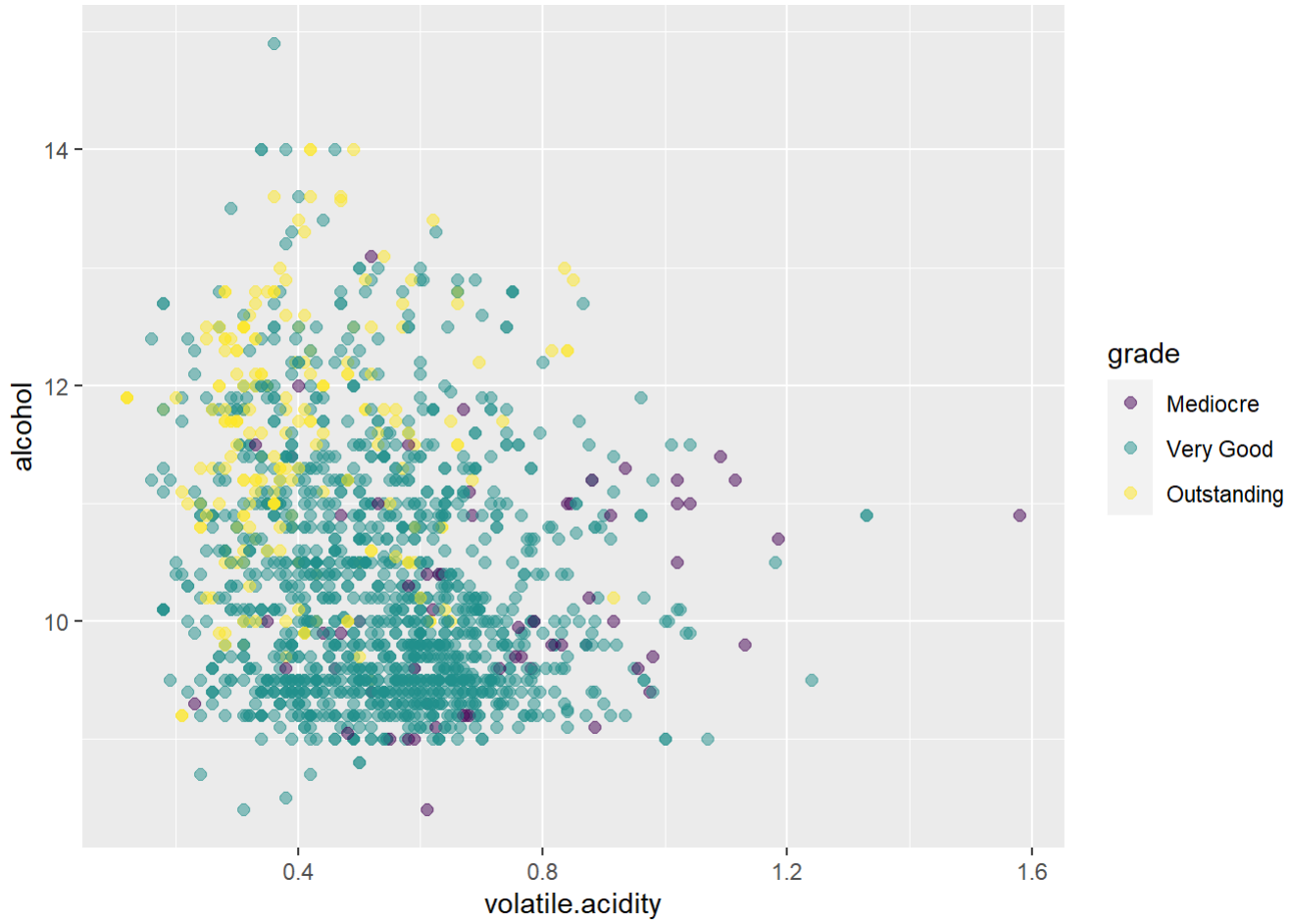
It looks like alcohol supplies only 22% to the overall quality of red wine. We should probably investigate other variables in combination with the two major correlation coefficients we found in the previous section.



By taking the next strongest positive correlation coeffecient for “quality” from our correlation table. We add “sulphates” and can see that a higher “grade” of red wine will contain lower “volatile.acidity”, and a higher amount of “alcohol” and sulphates.



The fourth strongest correlated coefficient with “quality” was “citric.acid”. The scatterplot shows that a higher “grade” of red wine will contain lower “volatile.acidity”, and a higher amount of “alcohol” and citric acid.



This multivariate scatterplot shows two of the main variables, “alcohol” and “volatile.acidity” and their affect on the “grade” of the red wine. Higher alcohol and lower volatile acidity result in a higher grade.

This multivariate scatterplot shows two of the main variables, “sulphates” and “citric.acid” and their affect on the “grade” of the red wine. Higher sulphates and higher citric acid result in a higher grade.

Mathematical Modeling using Linear Models

```
# useing the linear module function and the update function we can incrementaly
# create a linear model and then use the model tabling function we can generate
# a multivariate linear model table.

m1 <- lm(as.numeric(quality) ~ volatile.acidity, data = subset(red_wine,
                                                red_wine$citric.acid > 0))
m2 <- update(m1, ~. + alcohol)
m3 <- update(m2, ~. + sulphates)
m4 <- update(m3, ~. + citric.acid)

mtable(m1,m2,m3,m4, summary.stats = c('sigma','R-squared', 'F', 'p', 'N'))
```

```
##
## Calls:
## m1: lm(formula = as.numeric(quality) ~ volatile.acidity, data = subset(red_wine,
##     red_wine$citric.acid > 0))
## m2: lm(formula = as.numeric(quality) ~ volatile.acidity + alcohol,
##     data = subset(red_wine, red_wine$citric.acid > 0))
## m3: lm(formula = as.numeric(quality) ~ volatile.acidity + alcohol +
##     sulphates, data = subset(red_wine, red_wine$citric.acid >
##     0))
## m4: lm(formula = as.numeric(quality) ~ volatile.acidity + alcohol +
##     sulphates + citric.acid, data = subset(red_wine, red_wine$citric.acid >
##     0))
##
## ================================================================================
##                          m1           m2           m3           m4
## --------------------------------------------------------------------------------
##   (Intercept)          4.562***     0.942***     0.479*       0.511*
##                       (0.061)      (0.196)      (0.207)      (0.211)
##   volatile.acidity    -1.765***    -1.322***    -1.170***    -1.213***
##                       (0.113)      (0.104)      (0.105)      (0.120)
##   alcohol                           0.326***     0.322***     0.322***
##                                    (0.017)      (0.017)      (0.017)
##   sulphates                                      0.644***     0.662***
##                                                 (0.103)      (0.105)
##   citric.acid                                                -0.084
##                                                              (0.112)
## --------------------------------------------------------------------------------
##   sigma                0.745        0.666        0.658        0.658
##   R-squared            0.142        0.315        0.333        0.333
##   F                  242.959      337.056      243.729      182.882
##   p                    0.000        0.000        0.000        0.000
##   N                 1467         1467         1467         1467
## ================================================================================
##   Significance: *** = p < 0.001; ** = p < 0.01; * = p < 0.05
```

The table above produces semantics for the linear regression model created for the top four correlated variables that contribute to the overall “quality” of red wine.

The table produces the “Coefficient of determination” or “R-Sqaured” which essentially tells use the proportions contributed by each of these variables.

They are as follows:

- “volatile.acidity” = 0.142
- “alcohol” = 0.315
- “sulphates” = 0.333
- “citric.acid” = 0.333

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?

During this part of the investigation I focused on the four variables that most correlated with the dependent variable “quality”.

There were variables that strengthened eachother to result in a higher quality of red wine. Specifically, “sulphates” and “alcohol”, when both were found at a higher concentration within the red wine, it greatly influenced its quality.

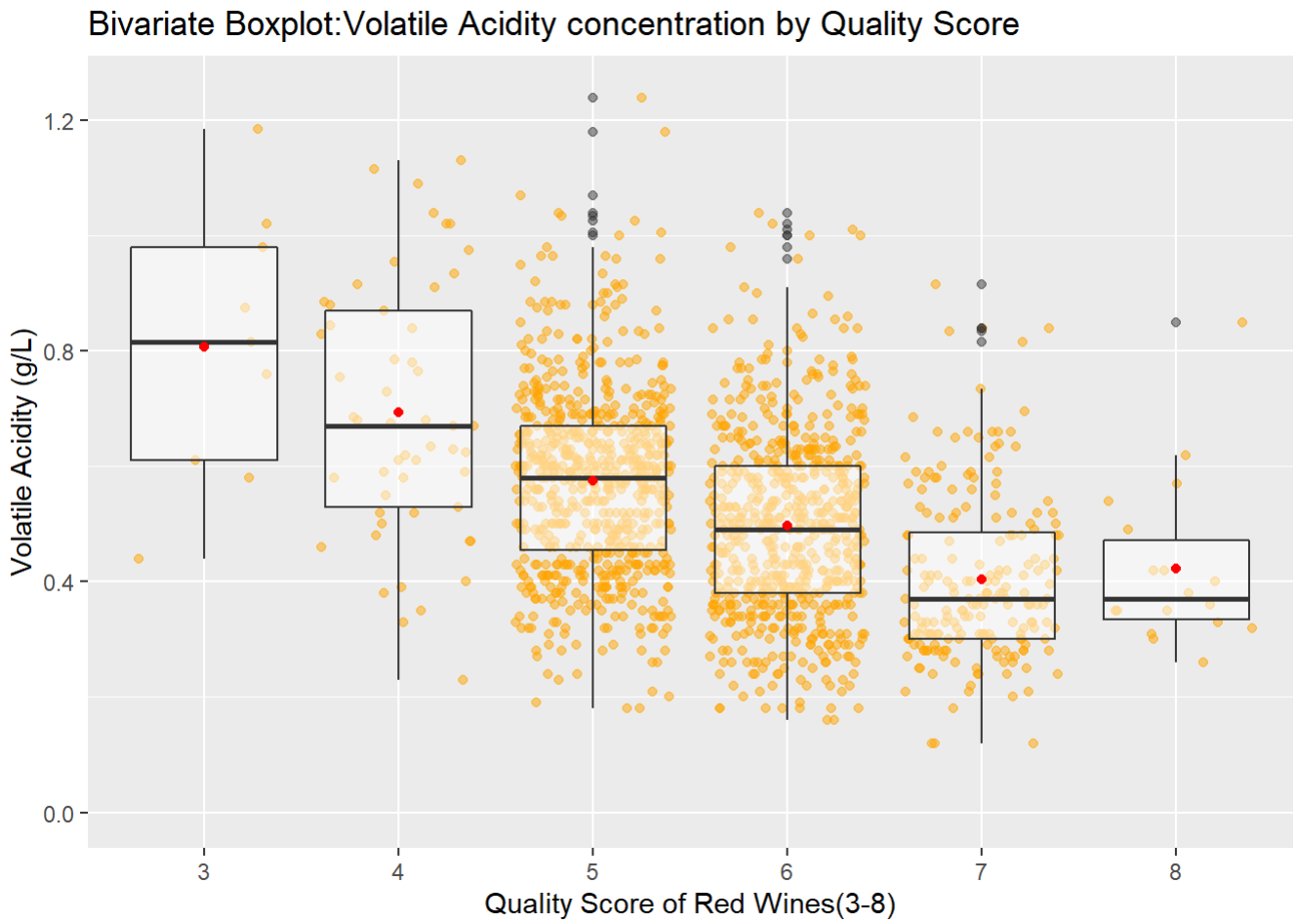
Were there any interesting or surprising interactions between features?

It was interesting that the lower amount of “volatile.acidity” and higher amount of “alcohol” were not the only contributing factors to the “quality” of red wine.

Final Plots and Summary

> **Tip**: You’ve done a lot of exploration and have built up an understanding of the structure of and relationships between the variables in your dataset. Here, you will select three plots from all of your previous exploration to present here as a summary of some of your most interesting findings. Make sure that you have refined your selected plots for good titling, axis labels (with units), and good aesthetic choices (e.g. color, transparency). After each plot, make sure you justify why you chose each plot by describing what it shows.

Plot One

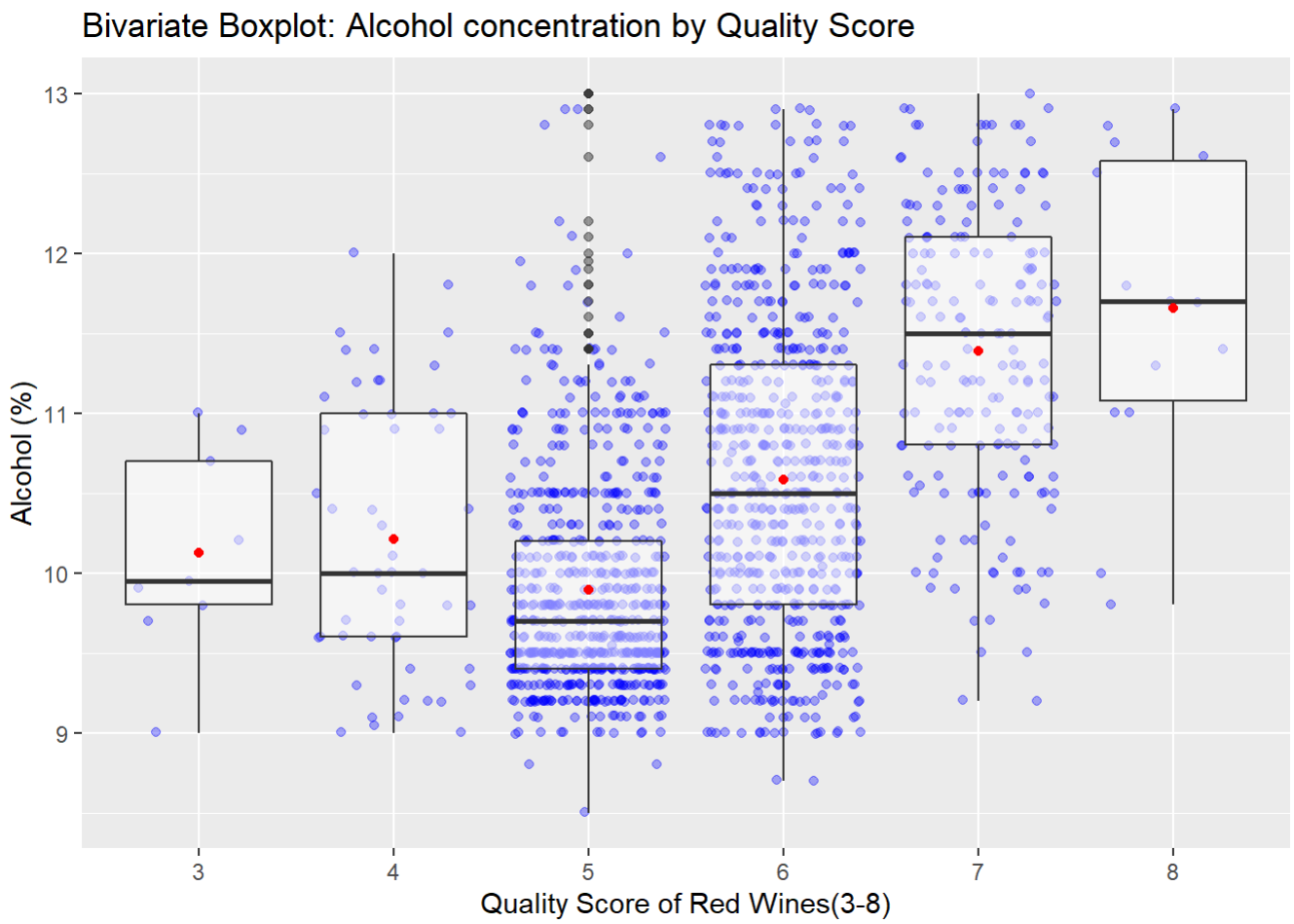


Description One

This bivariate boxplot was one of the most influencial plots I discovered during my analysis. It helped me validate that “volatile.acidity” was found in lower concentrations as the “quality” of the red wine increased.

This discovery was in direct conflict with my presumptions of the variable and led to further investigation into other variable’s relationship with volatile acidity.

Plot Two



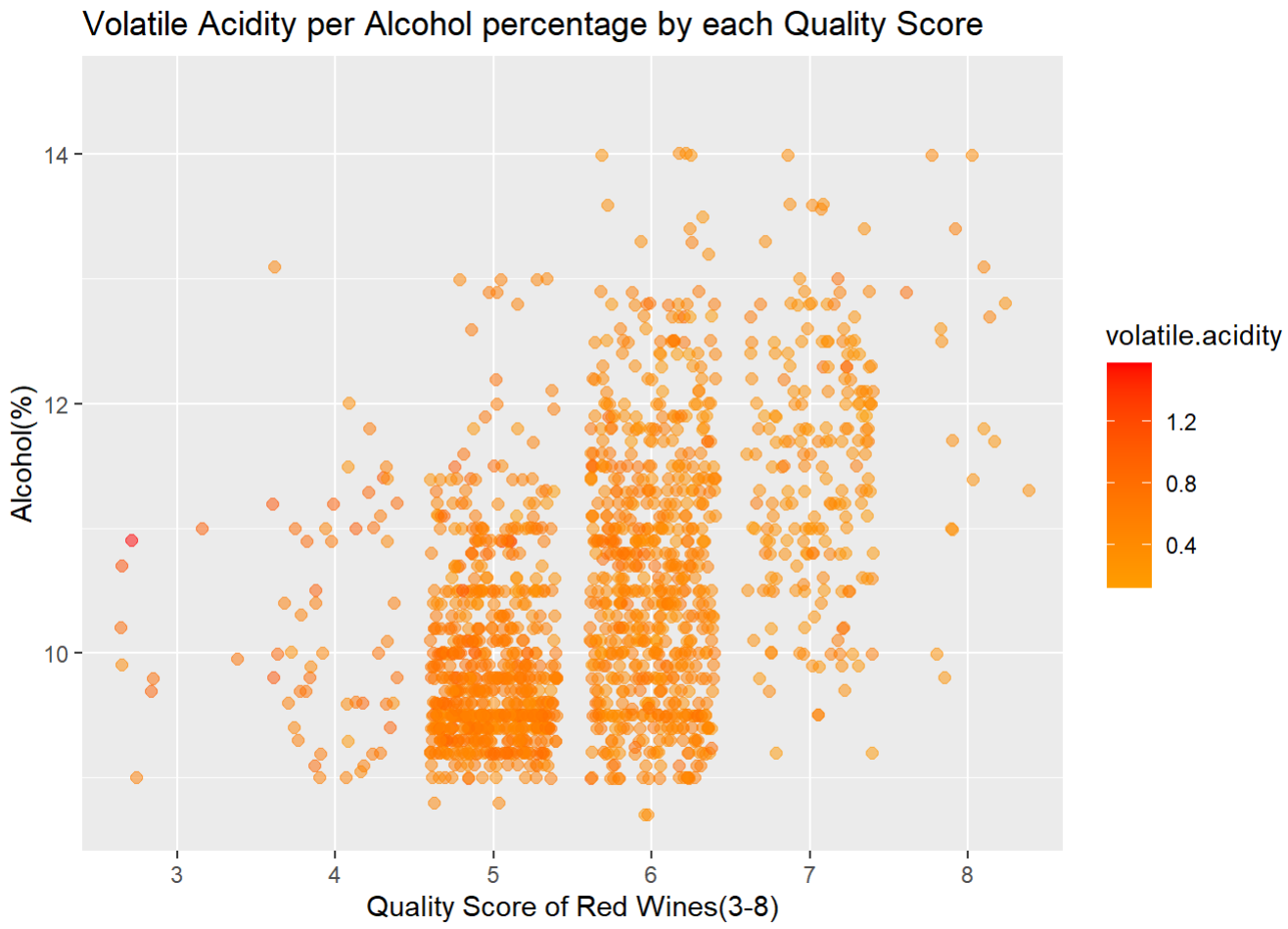
Description Two

The second plot I felt was very impactful during my analysis of the red wine dataset was this bivariate boxplot on “alcohol” percentage by “quality”.

This confirmed that the average percentage of alcohol was increasing as the quality of wine got better. However, there was another very important insight gained through the boxplot. Due to the amount of outlier data present within the plot, it was theorized that there could possibly be other variables that exihibited a strong positive correlation with the quality of red wine.

Thus leading to the discovery of the positive impact on quality being contributed by “sulphates” and “citric.acid”.

Plot Three



Description Three

The last plot that was crucial in visualizing the top correlating factors that made up the “quality” of red wine was the multivariate scatterplot above.

To me this scatterplot did a wonderful job to represent the strong positive correlation between “alcohol” and quality, while at the same time demonstrating that the concentration of “volatile.acidity” would also decrease as the regression was followed.

Reflection

As I look back on some of the struggles I had during this analysis of the data, I thought for the most part I found the data to be easy to understand and that the data types correctly corresponded with the variable assigned.

However, as I descended into the exploration of the data, one of the early issues I had was my confidence in the how the data was retrieved or wrangled. Because the dataset was provided to me, I felt my analysis was hindered or limited by many unknown variables. For example, the dataset contained a relatively small amount of observations, and there seemed to be an extreme bias or high volume of average graded red wines. This led me to doubt in the integrity of the dataset as a whole. I found myself asking questions such as, “were these 1599 observations truly a random selection from a larger population?”, and, “could there be a possibility that bad or incorrect values may have been entered for some of the variables?”.

This uncertainty only compounded when I discovered that a lot of values of zero, were entered in for citric acid for a good number of the observations. Due to my lack of subject-matter I presumed that because red wine came from fruit that citric acid must be found in all samples of red wine. If this presumption was incorrect then a lot of the

visualizations that used a subset function to remove values of zero before plotting, would be tainted.

Also, along the same lines of subject matter, I found it difficult that units of measurements for each variable were excluded from the dataset. I often asked myself if whether or not the decimal I was looking at was larger or smaller than I was presuming. By including a unit of measurement such as mg. or g. next to the variable name, I felt it would allow for perhaps a more indepth and accurate analysis.

I believe it is important that before analysis is begun having a good understanding of the subject matter will not only benefit the data analysis process but also can help during the process of data wrangling.

After proceeding past the initial hurdles the rest of the analysis went very well. I found the bivariate and multivariate plots were formulated a lot easier than I thought they were going to be initially. I presumed that more conversions or calculations to alter data types in order to perform certain functions would need to be performed. However, this was not the case and the large majority of the data not only played nice with eachother, I also found it to plot out quite naturally, leading to less time taken on optimizing plot parameters.

The biggest surprise to me was that residual sugar didn't play that large of a factor in the quality of red wine. I correlated sugar with taste and then decided that the quality of wine would only improve with the amount of sugar found inside of the wine. This was obviously an ignorant presumption and due to the chemistry and variety of substances found within wine, I discovered that their relationship with eachother was the most important factor in understanding what made great red wine.

If future analysis were to be performed against this dataset, it would be awesome to figure out the price, brand, and country belonging to each sample. Then we could use analysis to determine how much of an impact these variables influenced the quality of wine. One would presume that the more expensive the wine the better the quality, and so it would be intriguing to discover those results. Also, if there were a specific brand or country that controlled the market for highest quality of red wine.

I enjoyed this project a lot and learned a ton about the benefits and use case for exploratory data analysis.