Red Wine Quality Data Analysis by Suzanne Shoesmith

```
##
## The downloaded binary packages are in
##  /var/folders/r9/tpbhnq6x2hv75bqyh3zg1y6r0000gp/T//RtmpD5XGFk/downloaded_packages
```

```
##
## The downloaded binary packages are in
##  /var/folders/r9/tpbhnq6x2hv75bqyh3zg1y6r0000gp/T//RtmpD5XGFk/downloaded_packages
```

```
##
## The downloaded binary packages are in
##  /var/folders/r9/tpbhnq6x2hv75bqyh3zg1y6r0000gp/T//RtmpD5XGFk/downloaded_packages
```

```
##
## The downloaded binary packages are in
##  /var/folders/r9/tpbhnq6x2hv75bqyh3zg1y6r0000gp/T//RtmpD5XGFk/downloaded_packages
```

```
##
## The downloaded binary packages are in
##  /var/folders/r9/tpbhnq6x2hv75bqyh3zg1y6r0000gp/T//RtmpD5XGFk/downloaded_packages
```

```
##
## The downloaded binary packages are in
##  /var/folders/r9/tpbhnq6x2hv75bqyh3zg1y6r0000gp/T//RtmpD5XGFk/downloaded_packages
```

For my project I chose to analyze the data set containing wine quality ratings for red wine. This data set was one of the tidy data sets in the suggested data sets in the Udacity link. The data set lists different phsyicochemical properties and corresponding wine quality ratings. The physicochemical properties include acidity, citric acid, sugar, alcohol content, etc. The main variable of interest is quality. It will be interesting to determine if any of the other physicochemical properties have an impact on the quality of the wine.

Univariate Plots Section

In the next few sections, I performed some basic exploration of the data (number of variables, number of columns and rows, and whether or not there are any null values).

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

```
##        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
```

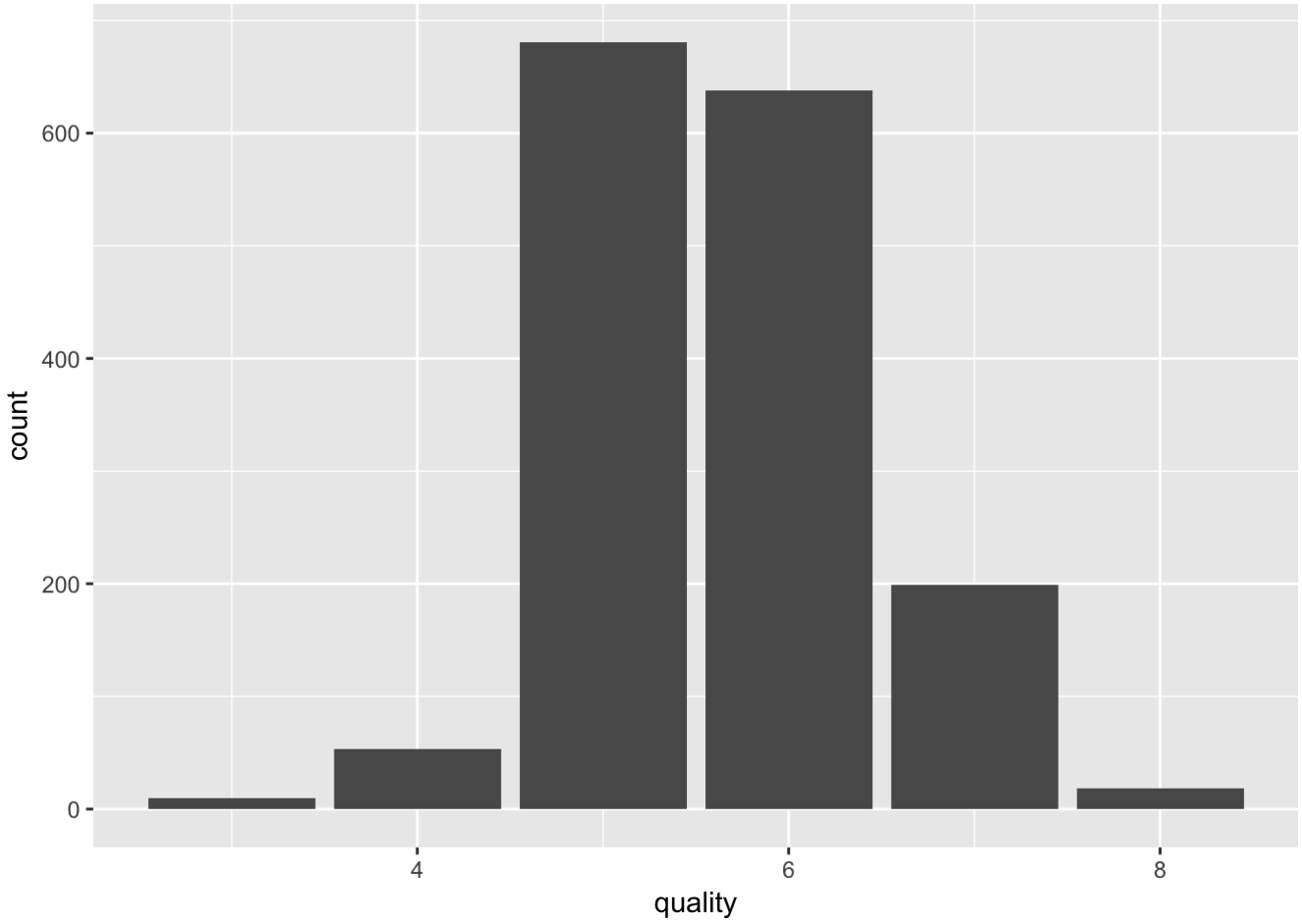
```
## [1] 13
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.636   6.000   8.000
```

```
##  [1] X                   fixed.acidity       volatile.acidity
##  [4] citric.acid         residual.sugar      chlorides
##  [7] free.sulfur.dioxide total.sulfur.dioxide density
## [10] pH                  sulphates           alcohol
## [13] quality
## <0 rows> (or 0-length row.names)
```

```
## 'data.frame':    1599 obs. of  13 variables:
##  $ X                  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity      : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity   : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid        : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar     : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides          : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 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 ...
```

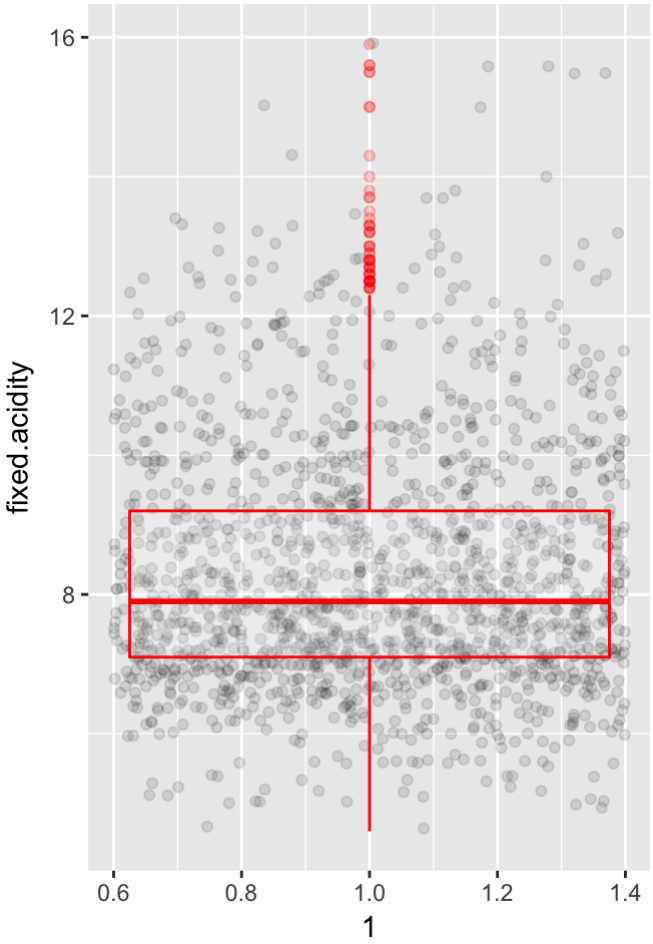


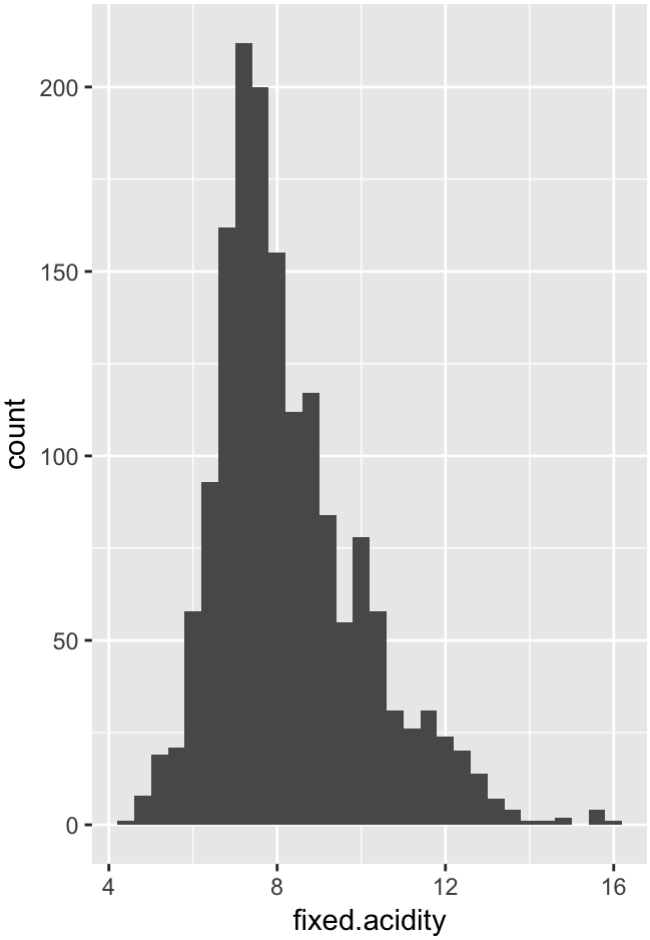
This value appears to be discrete, I'll determine the unique values in the variable price to determine the discrete values:

```
##
##   3   4   5   6   7   8
##  10  53 681 638 199  18
```

Based on the values obtained from the table command there are no ratings that are decimals, they are all whole numbers between 3 and 8. Also, based on the graph, it appears that the majority of wines are rated either a 5 or a 6. Also, there are no wines that were rated on either extreme of quality.

The next variable that I will examine is Fixed Acidity.





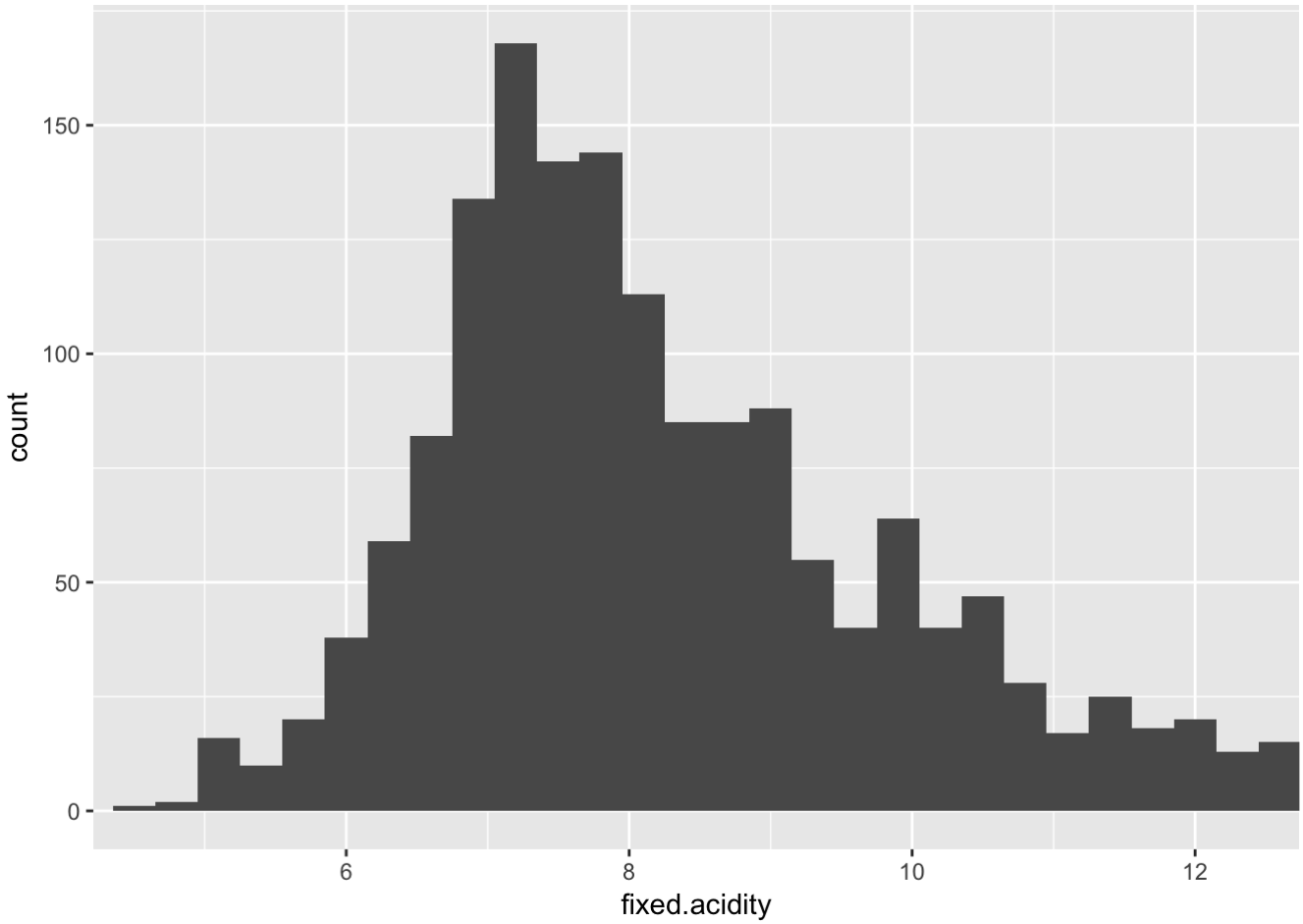

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
```

I initially created a histogram with all the data, and then created a scatterplot with a box plot in order to visually determine if there were any outliers that I could remove the data set to get a better understanding of the variable fixed.acidity.

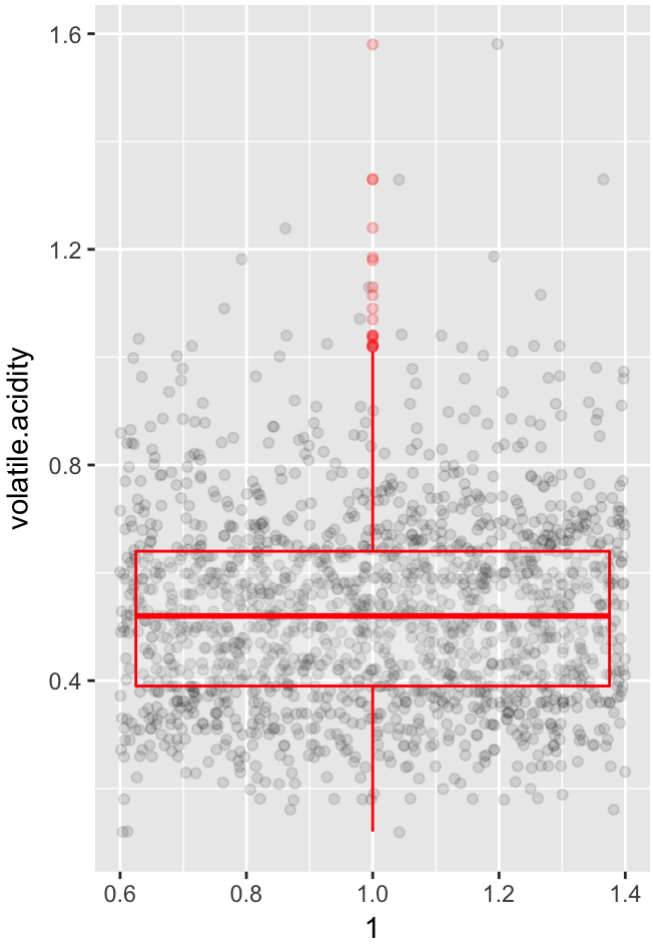
I also summarized this variable Mean, Median, and quartile ranges in order to more accurately remove outliers. The interquartile range for this variable is: 9.20 - 7.10 = 2.1. An outlier would be 1.5 times this number in either direction. 1.5 * 2.1 = 3.15 lower outliers: Q1 - 3.15 upper outliers: Q3 + 3.15

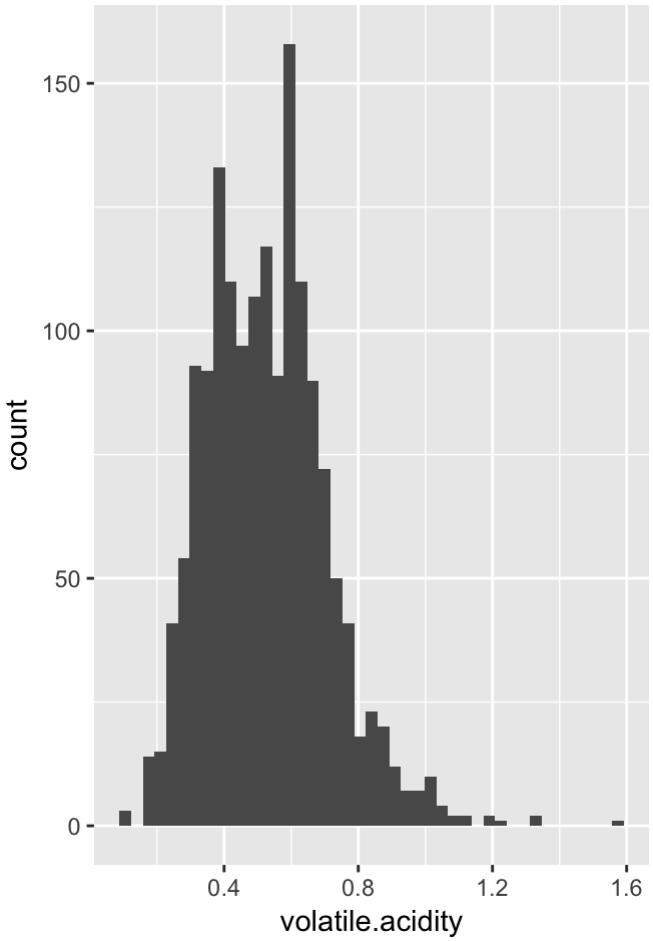
lower outlier limit: 7.10 - 3.15 = 3.95 (no lower outliers) upper outlier limit: 9.20 + 3.15 = 12.35 (outliers are values greater than 12.35, shown as the dots on the box plot)

I'll create another histogram below with the outliers removed. This will provide a more accurate representation of this variable.



With the new histogram with cartesian limits it is possible to see that although the data appears skewed to the right, it is less skewed than orginally depicted in the histogram containing the outliers.



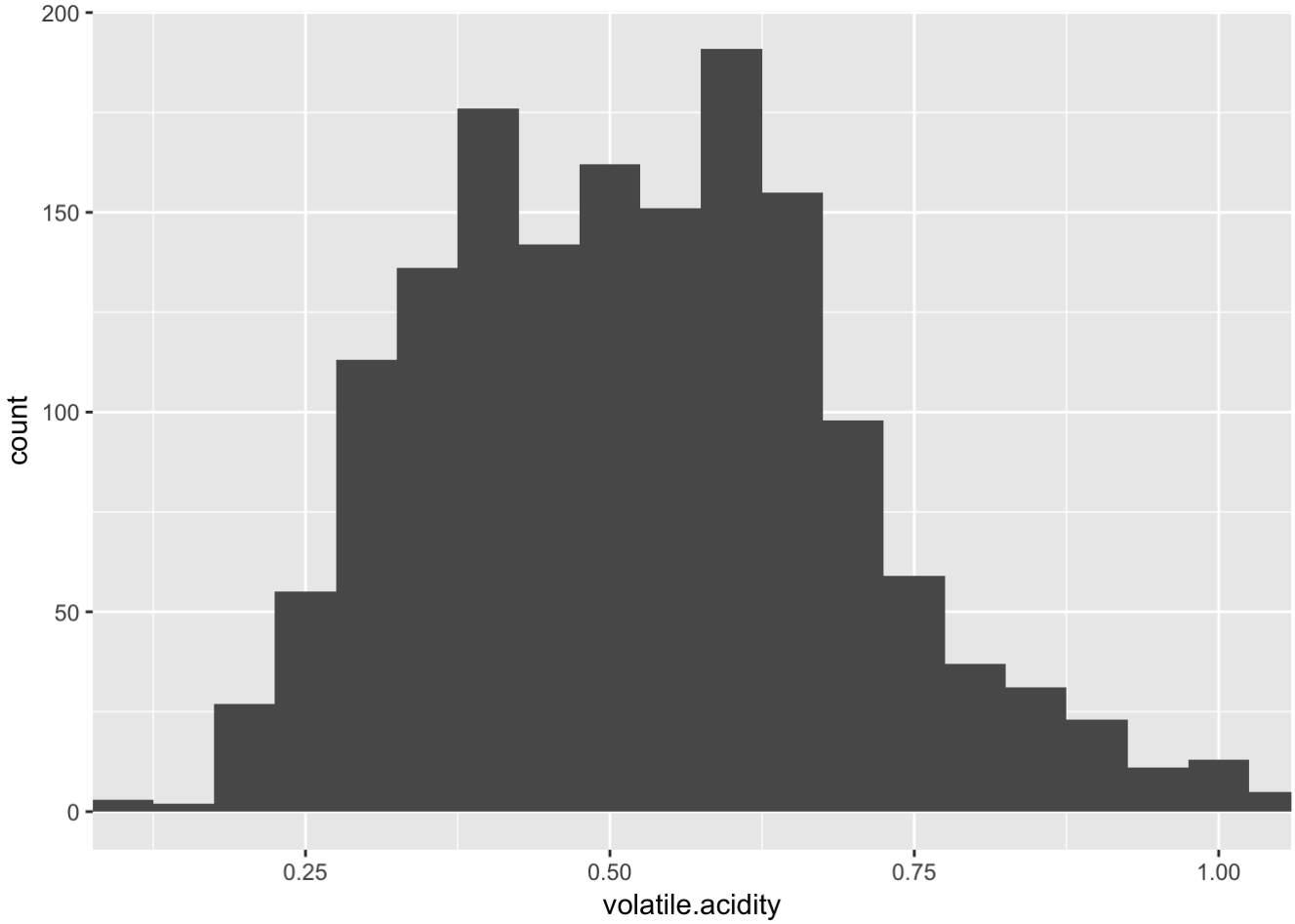



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```

When looking at the plots for volatile acidity, there appear to be outliers as inidicated by the red dots on the scatterplot. The summary of this variable is: Min. 1st Qu. Median Mean 3rd Qu. Max. 0.1200 0.3900 0.5200 0.5278 0.6400 1.5800

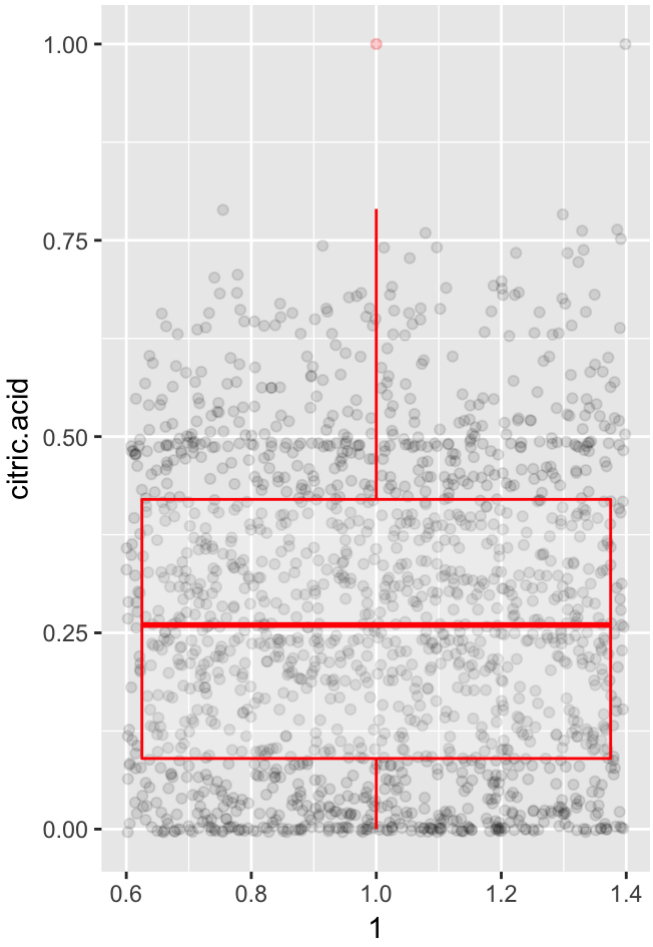
The interquartile range is 0.25, and 1.5x this number is 0.375 Since from the box plot it is obvious there are no lower outliers, I will use cartesian limits to remove upper outliers from the distribution. upper limit: 0.64 + 0.375 = 1.015

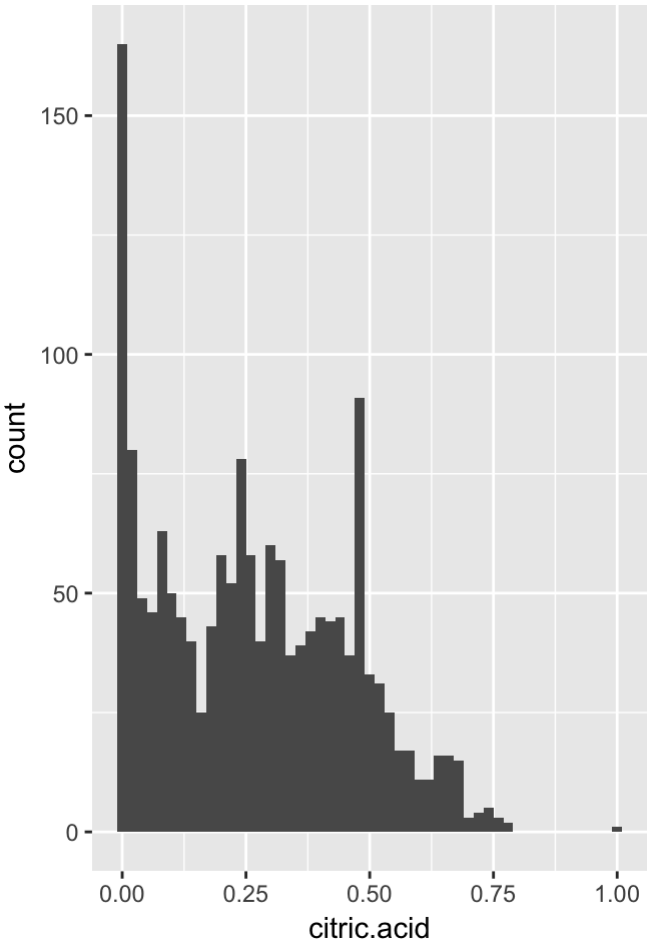
The updated histogram is shown below:



After removing the outliers from the graph and adjusting the binwidth, the distribution of volatile acidity looks more like a normal distribution.

The next variable I'll analyze is citric acid content:





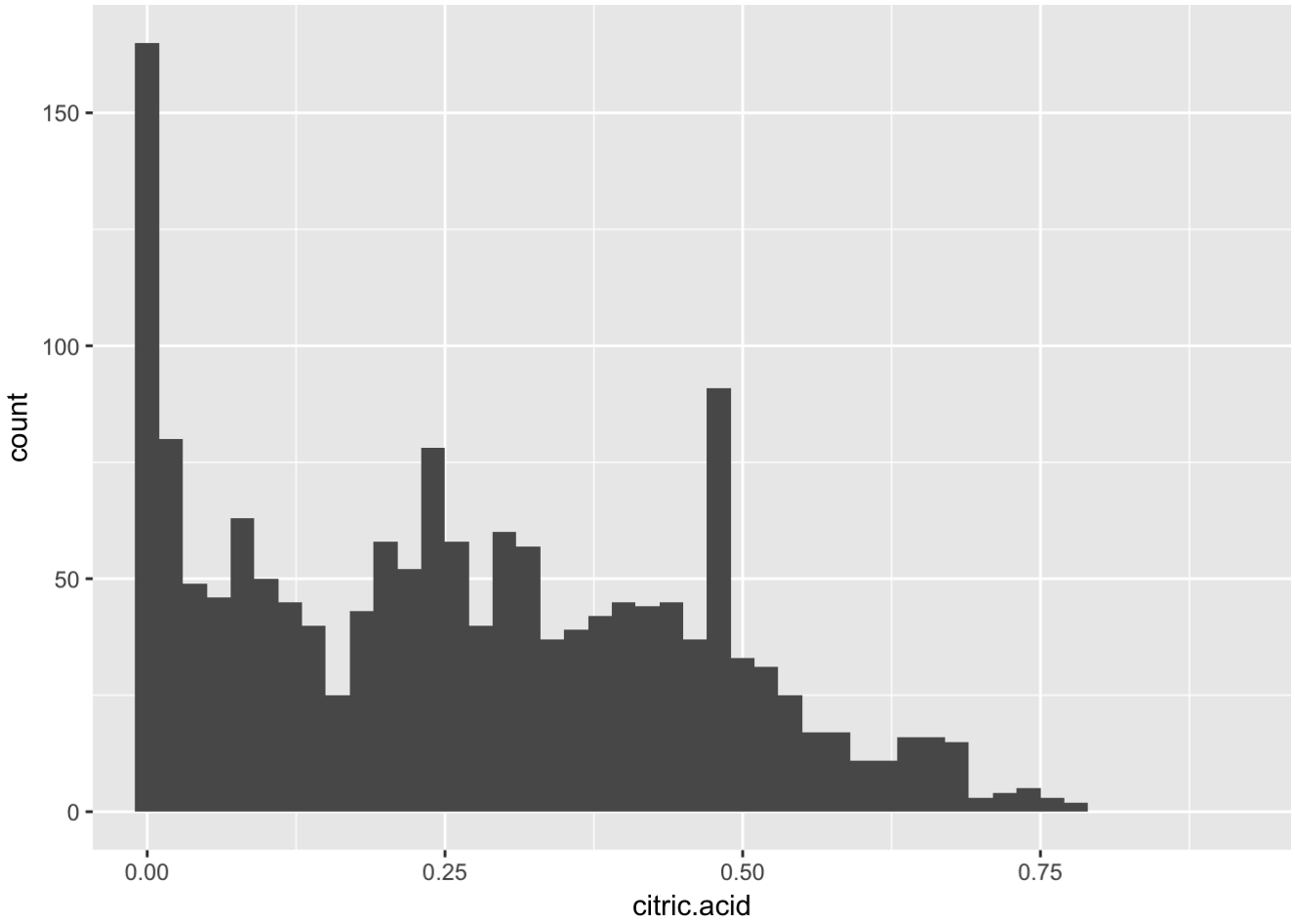

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```

Based on the above box plot, there appears to be only 1 outlier in this variable. The summary statistics are as follows:

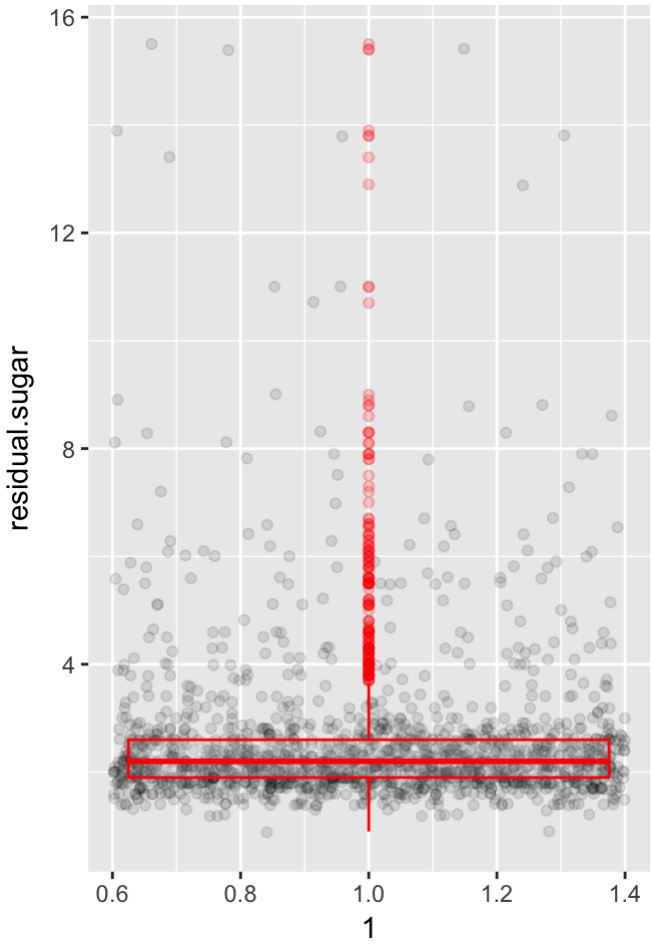
Min. 1st Qu. Median Mean 3rd Qu. Max. 0.000 0.090 0.260 0.271 0.420 1.000

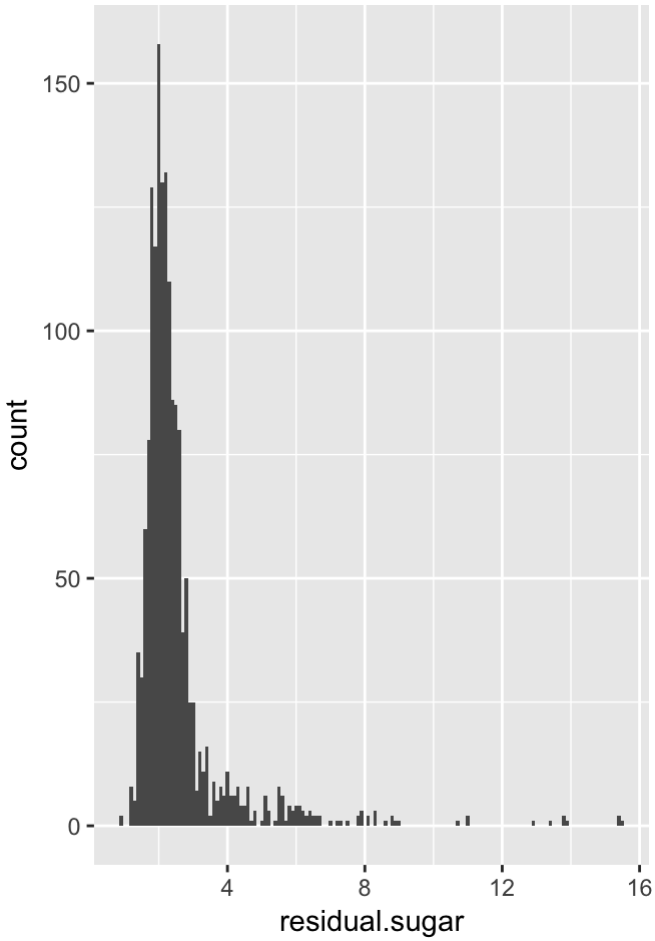
The interquartile range is 0.33, and 1.5x this number is 0.495 Since from the box plot it is obvious there are no lower outliers, I will use cartesian limits to remove upper outliers from the distribution. upper limit: 0.42 + 0.495 = 0.915

The updated histogram is shown below:



Even after removing the outlier, this distribution appears to be somewhat random. The next variable I will analyze is residual sugar:





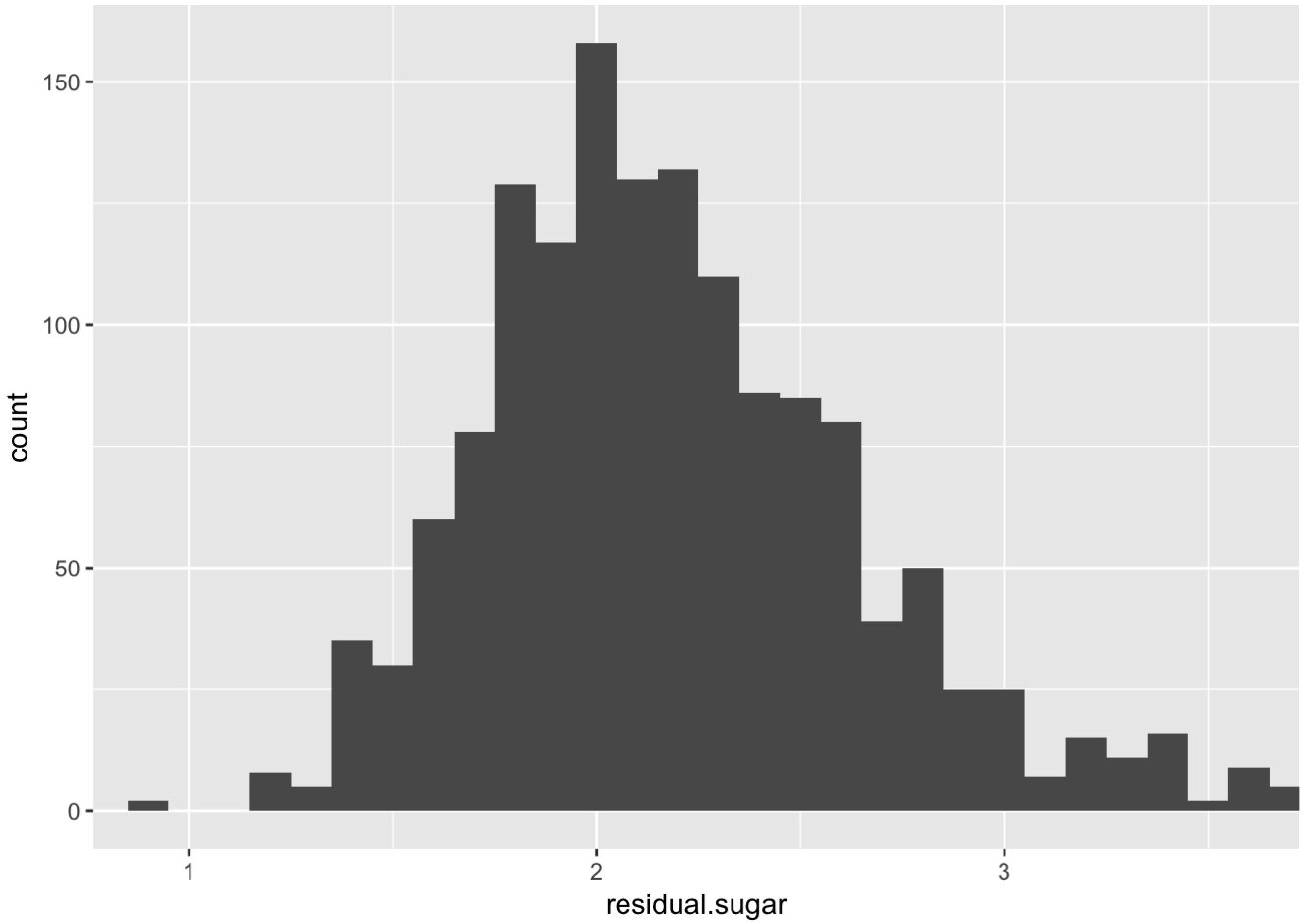

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```

Based on the scatterplot and the initial histogram, there appear to be a large number of high outliers, and the distribution is heavily skewed to the right with these outliers. The summary statistics are as follows:

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

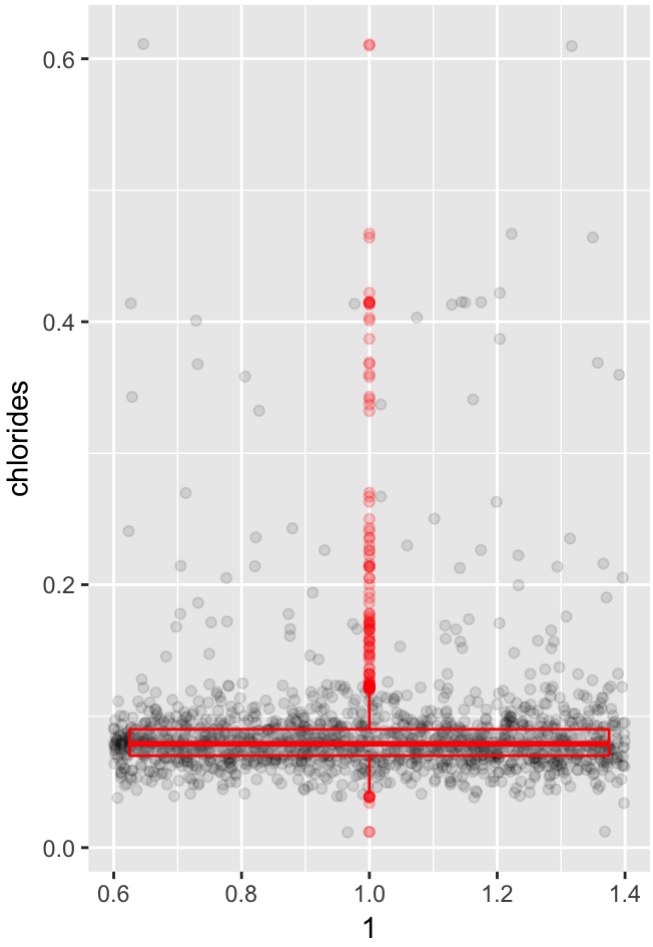
Removing the outliers and replotting:

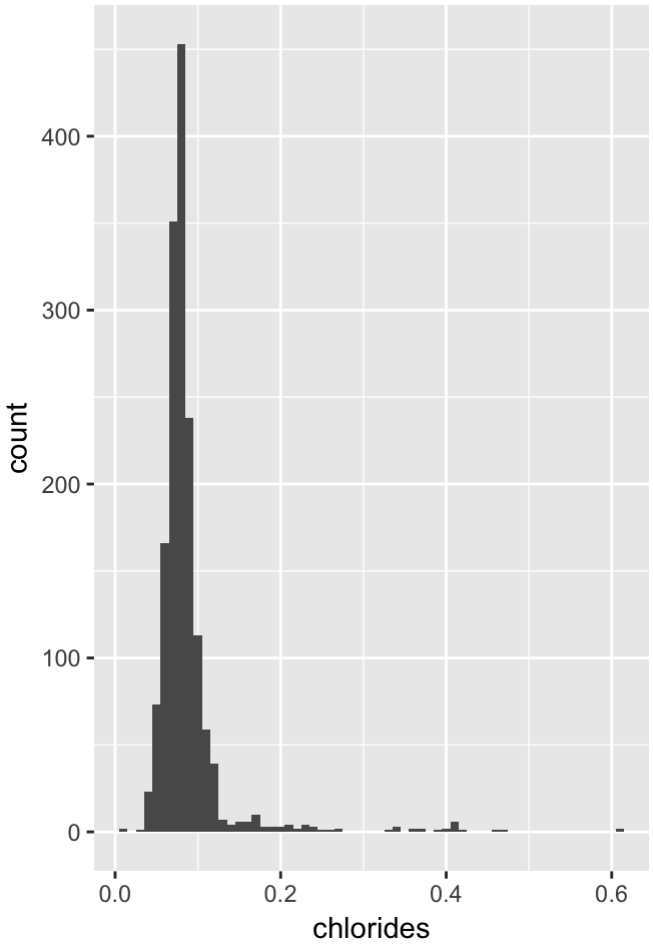
Interquartile range: 2.6 - 1.9 = 0.7 Upper outlier limit: 2.539 + (0.7*1.5) = 3.589 Lower outlier limit: 1.9 - (0.7*1.5) = 0.85



After removing the outliers, the histogram of residual sugar appears to be more normally distributed.

The next variable to be analyzed is chlorides:





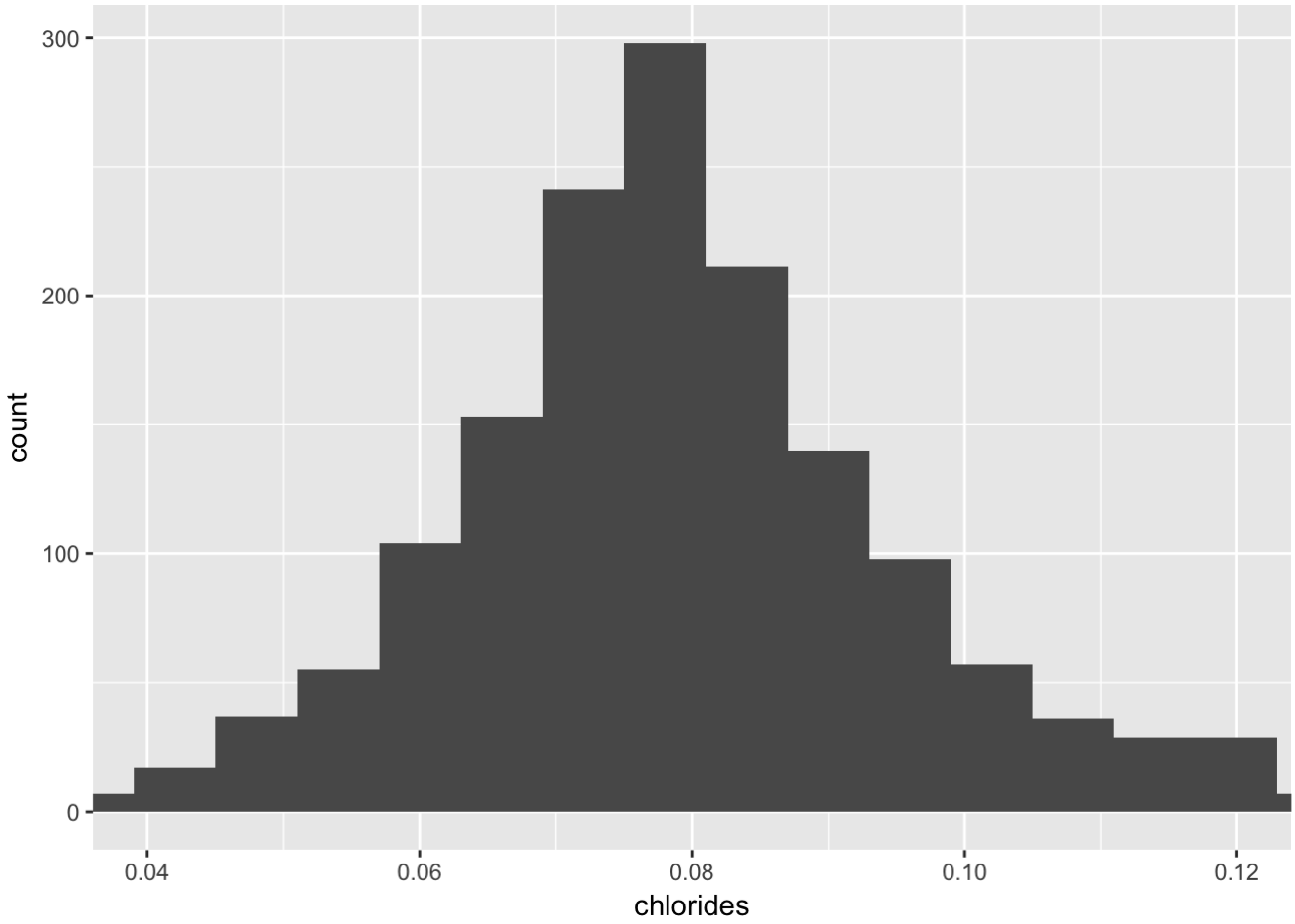

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```

Looking at the box plot, there appear to be outliers on both the upper and lower edges of the distribution. The summary of the variable chlorides is as follows:

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

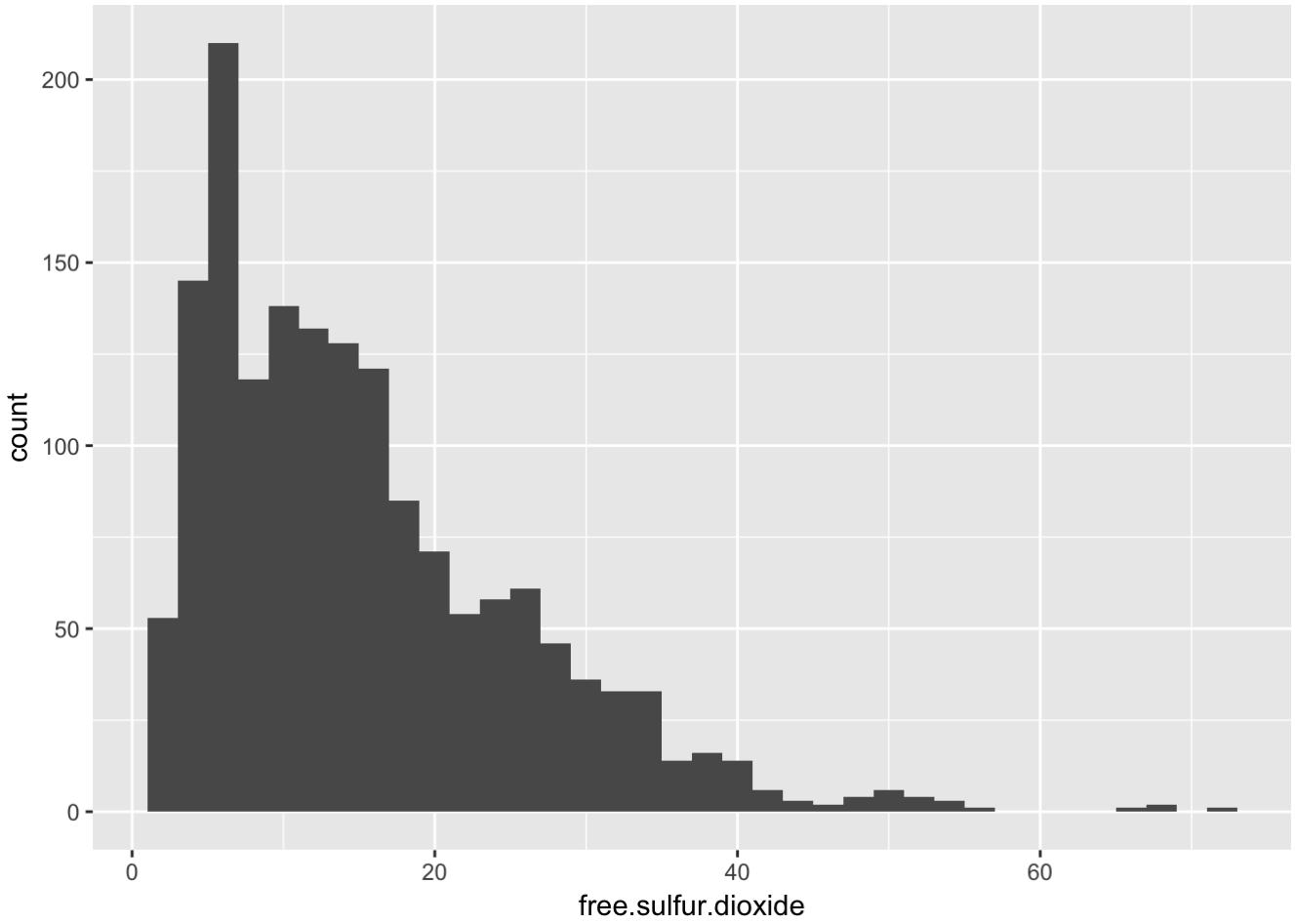
The interqurtile range and outliers are: Interquartile range: 0.09 - 0.07 = 0.02 Upper outlier limit: 0.09 + (0.02*1.5)* = *0.12 Lower outlier limit: 0.07 - (0.02*1.5) = 0.04

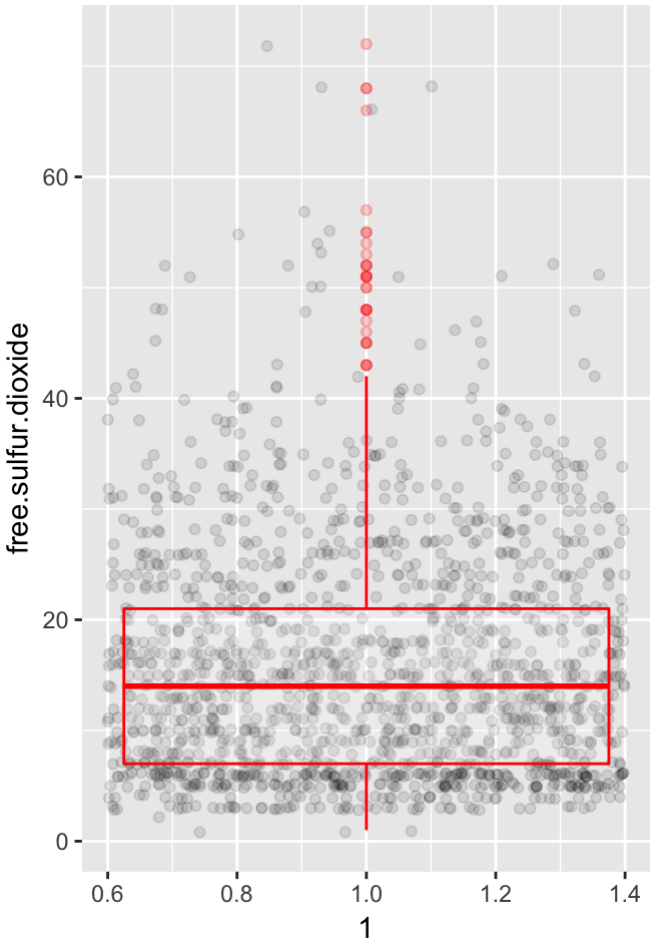
Replotting after removing the outliers from the plot:

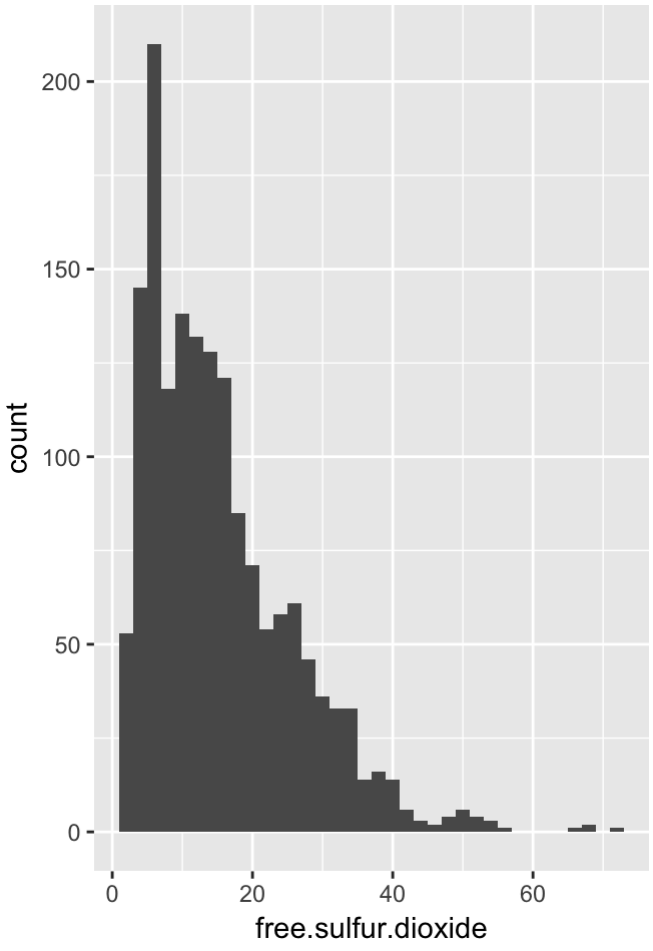


After adjusting the limits and the binwidth, the chlorides variable appears to be normally distributed.

The next variable to be analyzed is free sulfur dioxide:


count
200
150
100
50
0
0
20
40
60
free.sulfur.dioxide


free.sulfur.dioxide
60
40
20
0
0.6
0.8
1.0
1.2
1.4
1


count
200
150
100
50
0
0
20
40
60
free.sulfur.dioxide

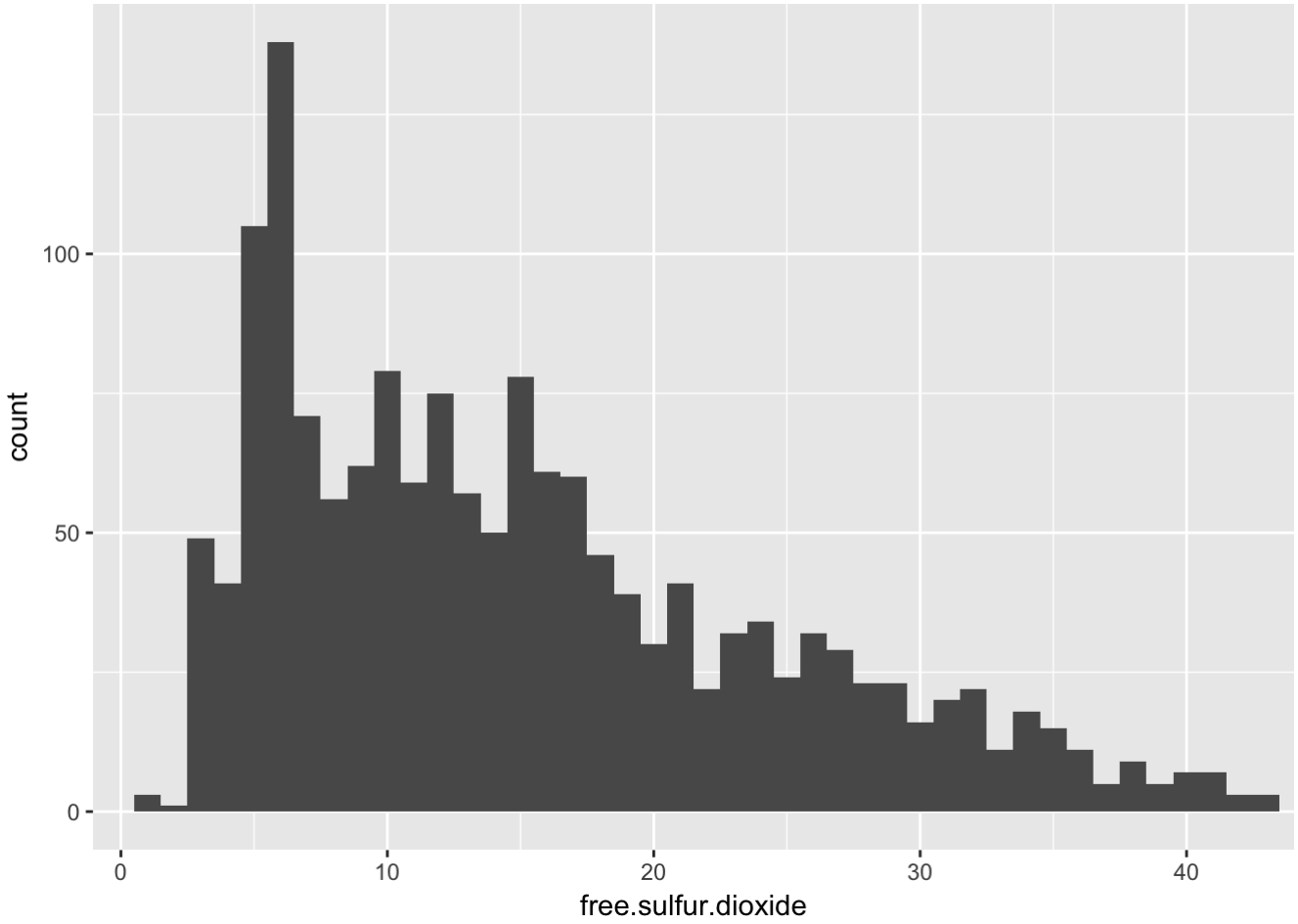
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
```

Looking at the box plot it is clear that there are outliers in the variable free sulfur dioxide. The summary of the variable is as follows:

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

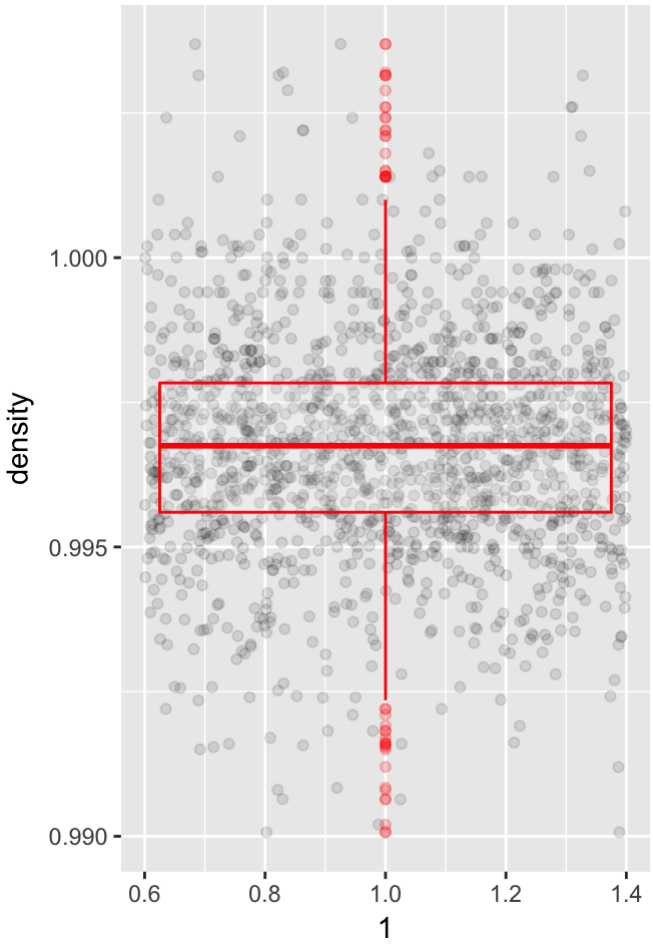
The interqurtile range and outliers are: Interquartile range: 21.00 - 7.00 = 14 Upper outlier limit: 21 + (14*1.5) = 42* *Lower outlier limit: 7-(14*1.5) = -14

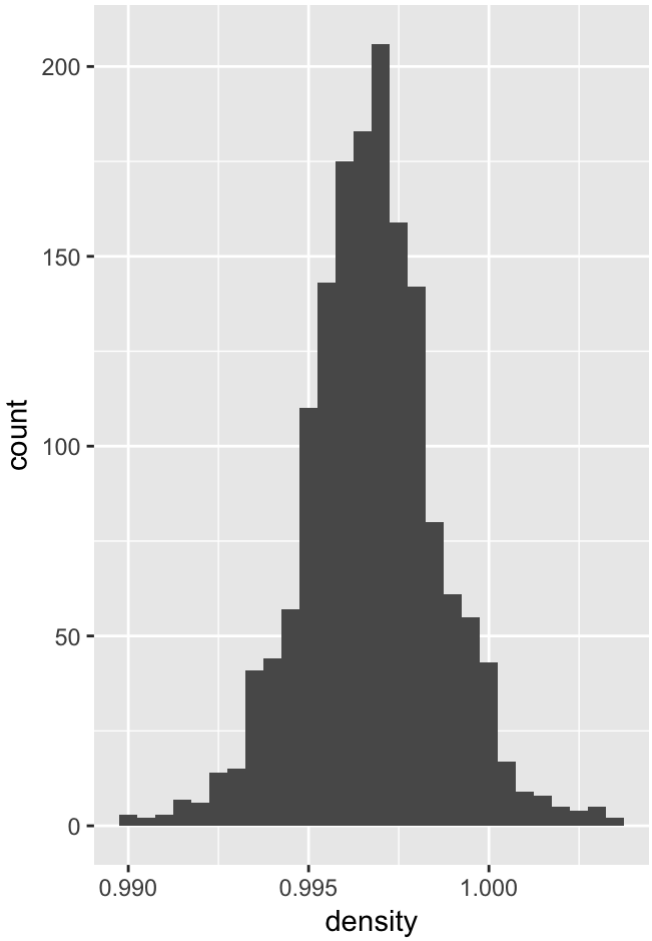
Replotting after removing the outliers from the plot:



After applying cartesian limits and adjusting the binwidth, the distribution still appears to be skewed to the right.

The next variable I“ll analyze is density:





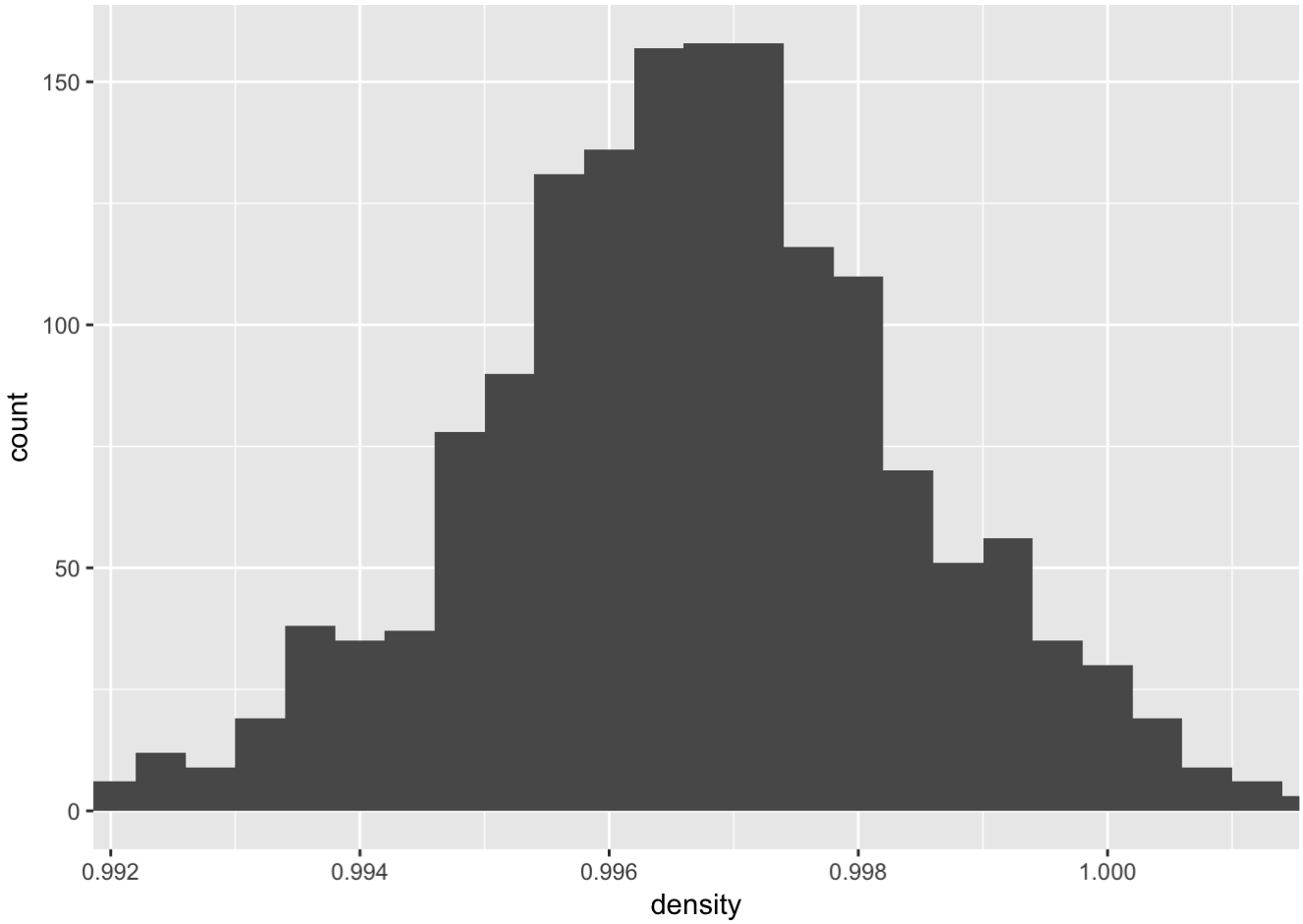

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0037
```

The distribution of density appears normal, but the box plot shows there are some outliers, both on the lower and higher ends. The summary of the variable density is as follows:

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

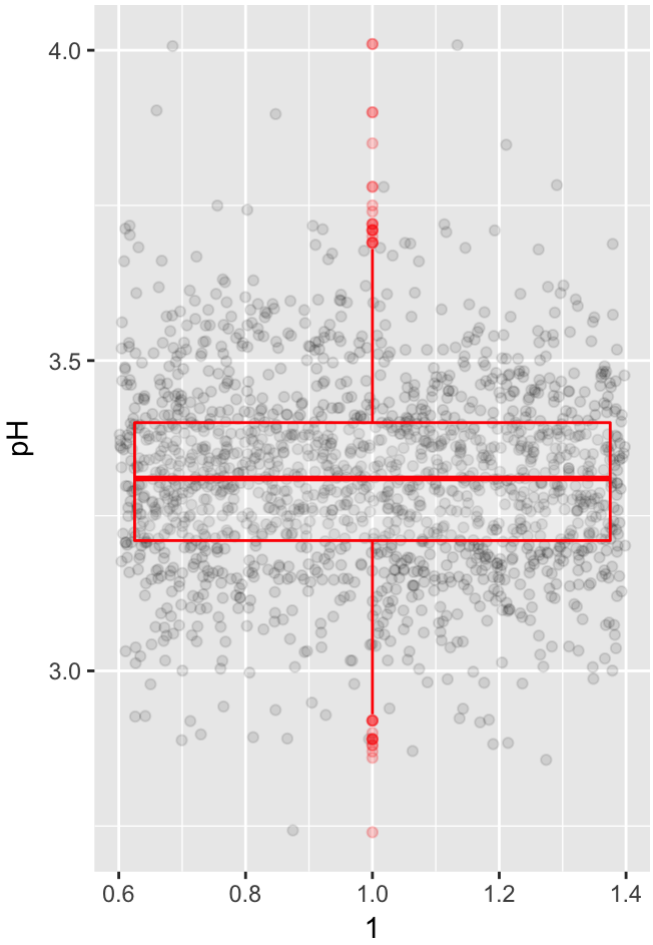
The interqurtile range and outliers are: Interquartile range: 0.9978 - 0.9956 = 0.0022 Upper outlier limit: 0.9978 + 1.5(0.0022) = 1.0011 Lower outlier limit: 0.9956 - 1.5(0.0022) = 0.9923

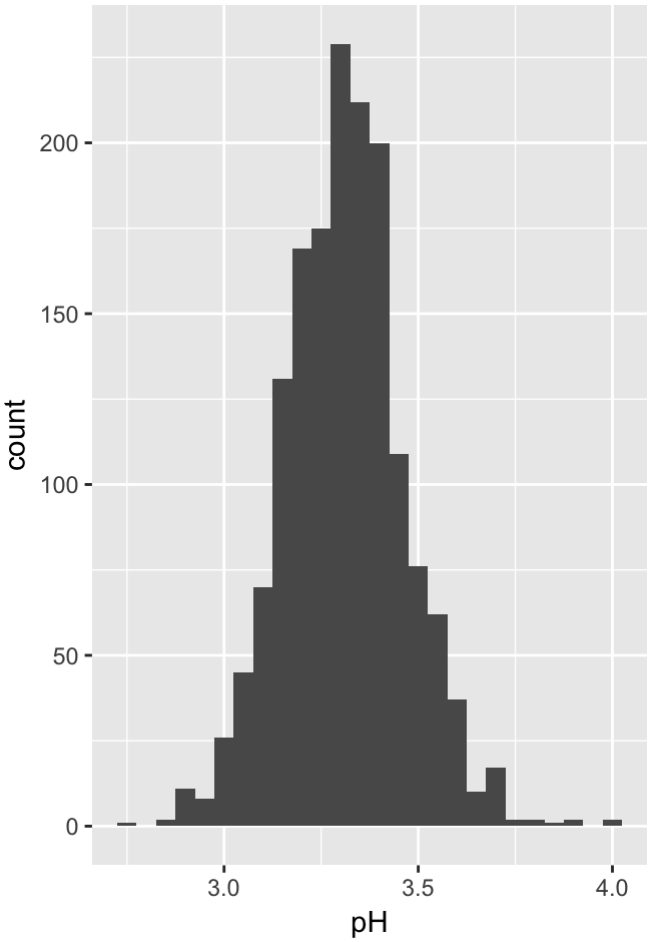
replotting after applying cartesian limits:



I replotted density using the cartesian limits and adjusted with binwidth. The data still appears normally distributed. After removing outliers, the tails on either end are smaller.

The next variable for analysis is pH:





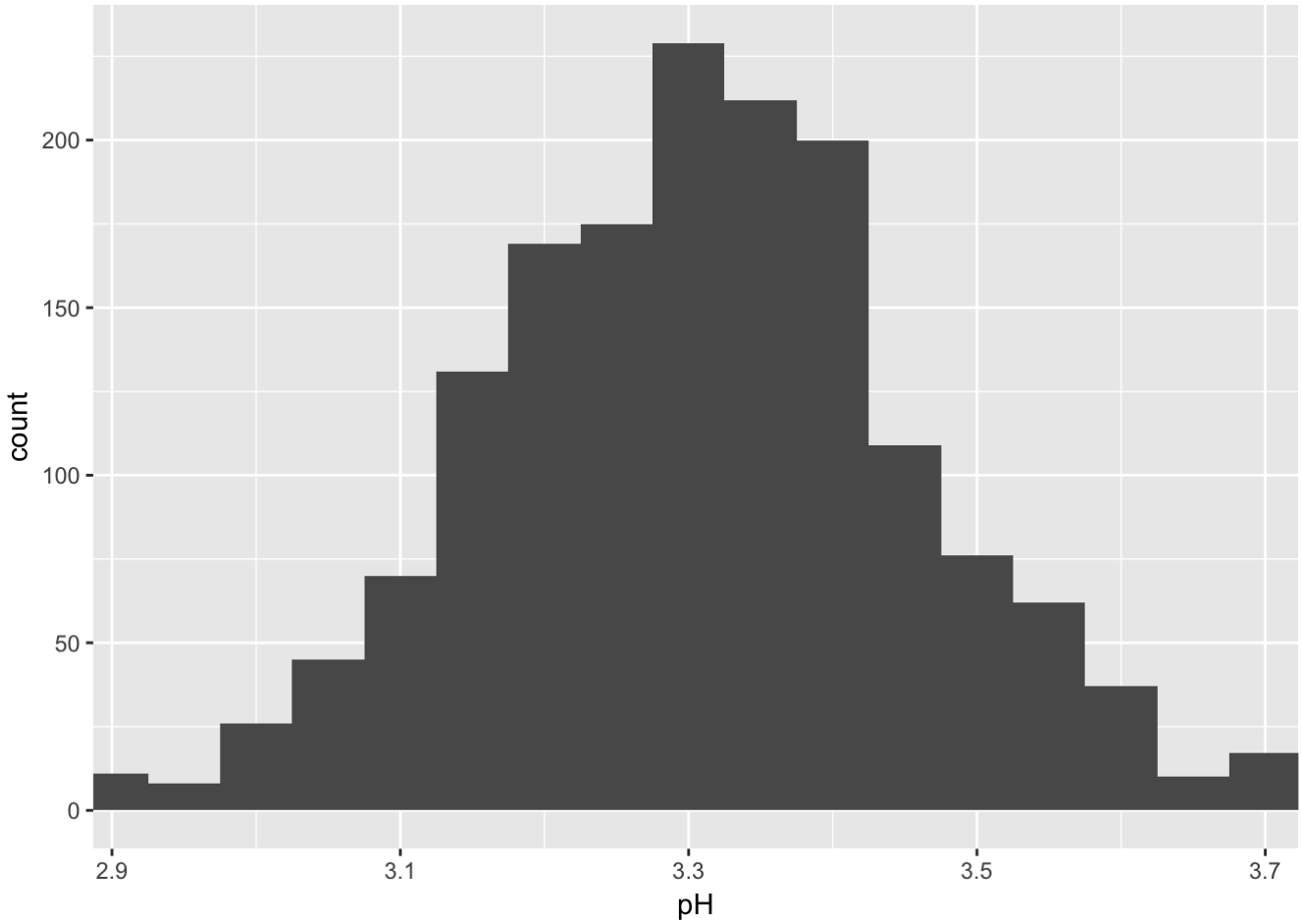

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010
```

The distribution of pH in the dataset appears to be normally distributed with some outliers. The summary of the variable is as follows:

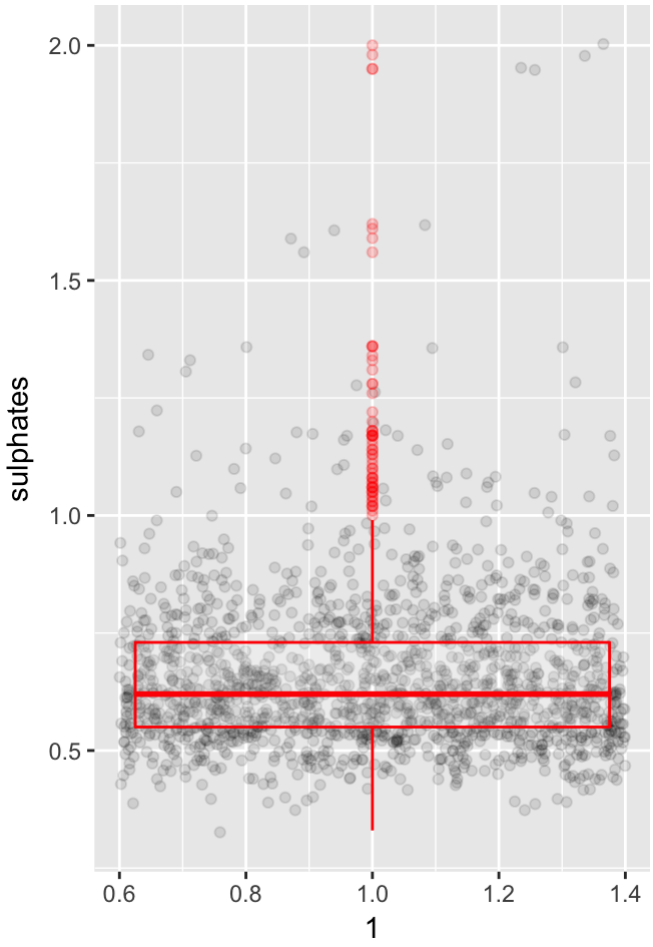
Min. 1st Qu. Median Mean 3rd Qu. Max. 2.740 3.210 3.310 3.311 3.400 4.010

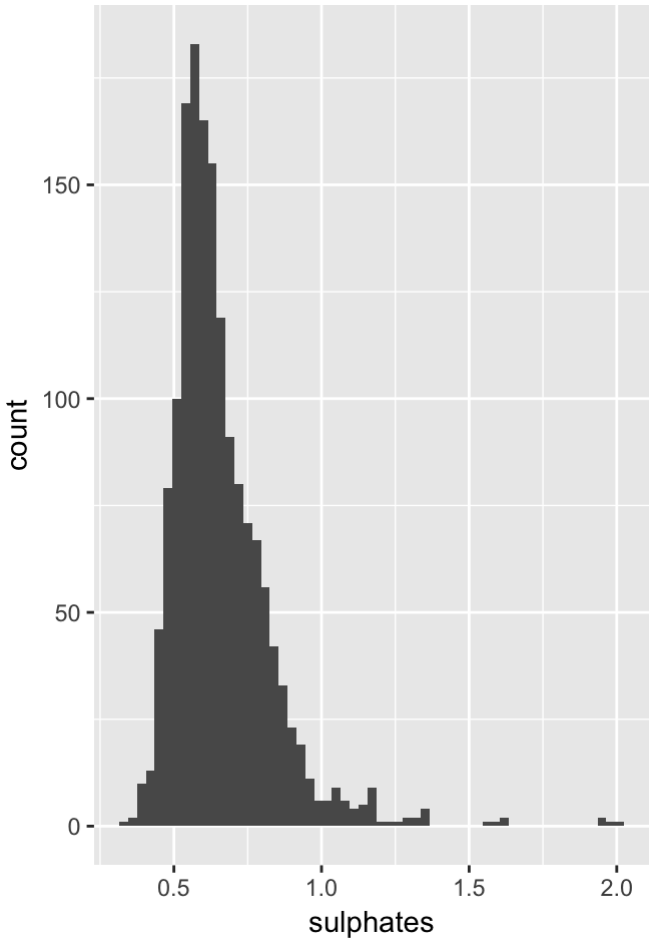
The interqurtile range and outliers are: Interquartile range: 3.4 - 3.210 = 0.19 Upper outlier limit: 3.4 + 1.5(0.19) = 3.685 Lower outlier limit: 3.210 - 1.5(0.19) = 2.925

replotting after applying cartesian limits:



After removing the outliers and applying cartesian limits the pH data still appears normally distributed.






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

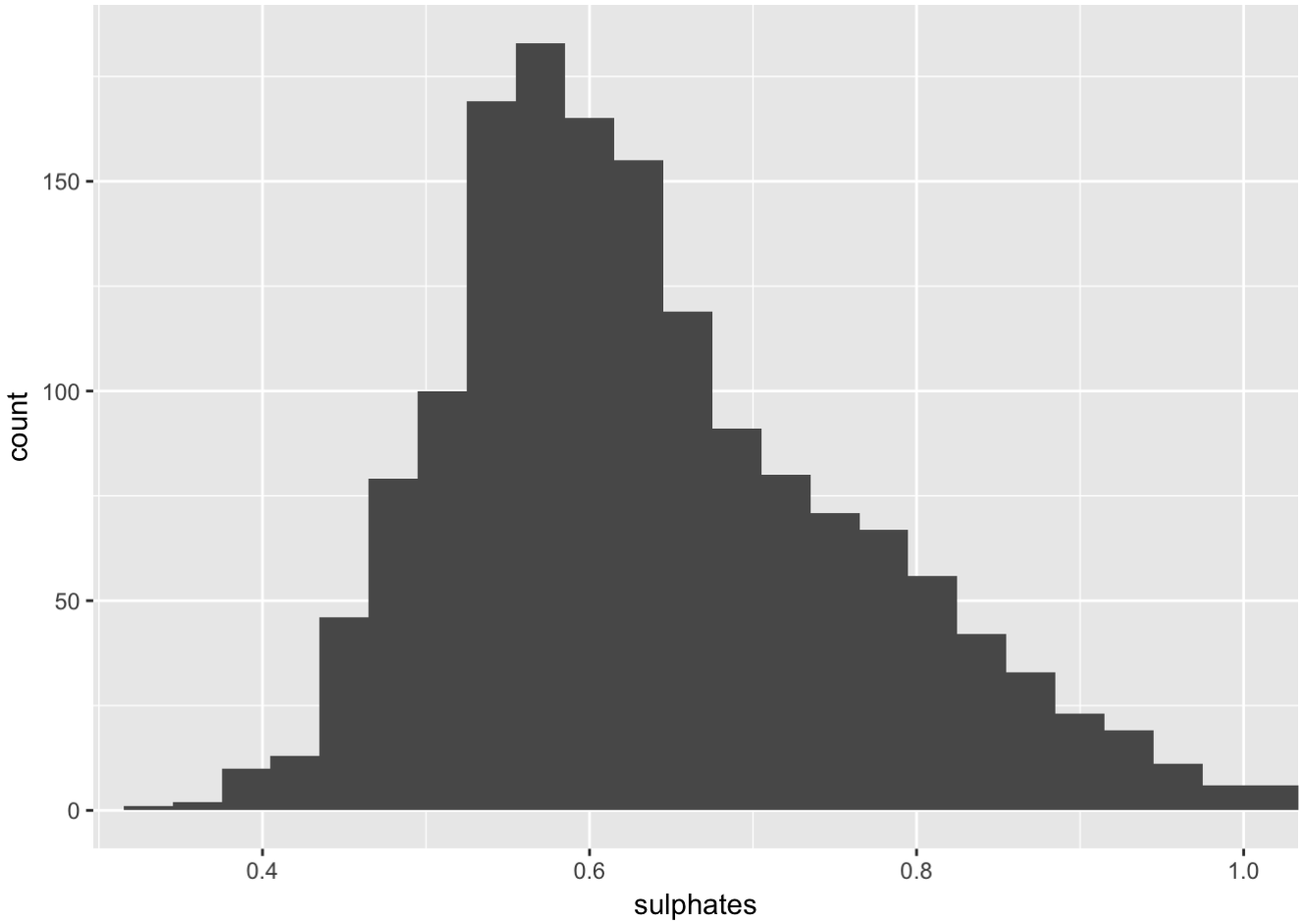
The data for sulphates appears to be skewed to the right. Looking at the box plot there are outliers on the upper end of the distribution.

The summary of the variable is as follows:

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

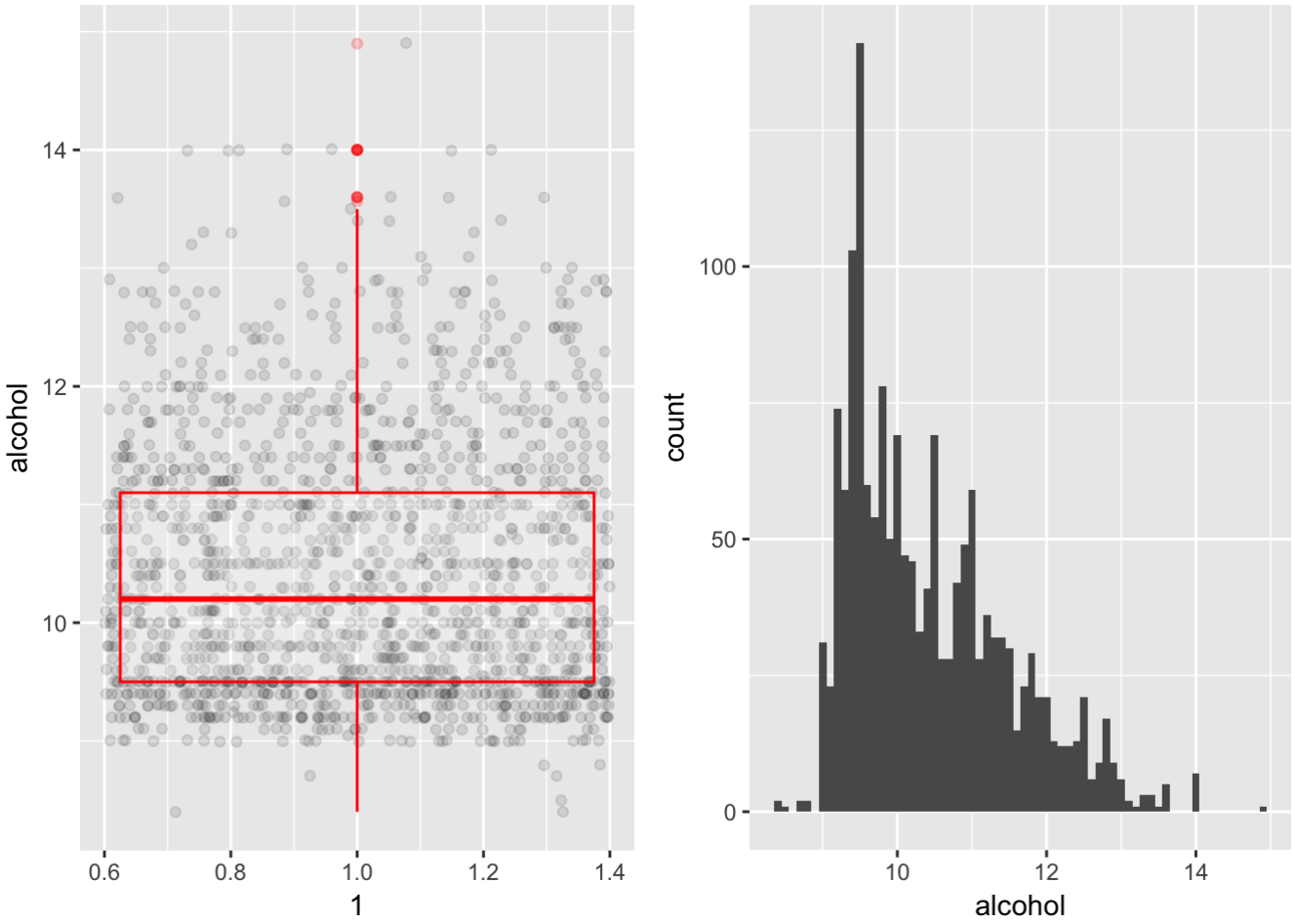
The interqurtile range and outliers are: Interquartile range: 0.73 - 0.55 = 0.18 Upper outlier limit: 0.73 + 1.5(0.18) = 1 Lower outlier limit: 0.55 - 1.5(0.18) = 0.28

replotting after applying cartesian limits:



In this case, even after removing the upper outliers, the distribution is still skewed to the right.

The next variable to analyze is alcohol:



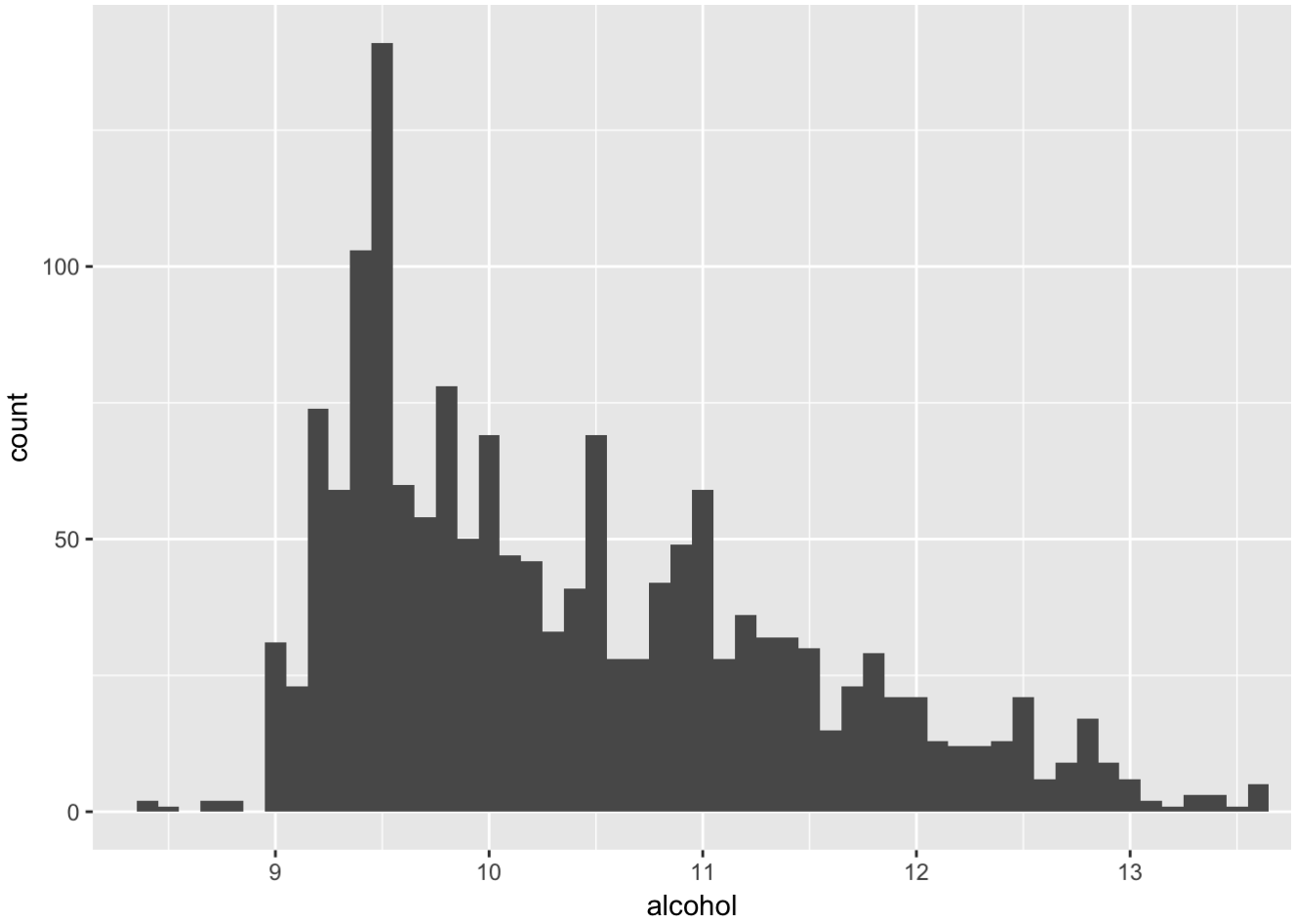

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```

The distribution of alcohol is skewed to the right. The summary of the variable is as follows:

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

The interqurtile range and outliers are: Interquartile range: 11.1 - 9.5 = 1.6 Upper outlier limit: 11.1 + 1.5(1.6) = 13.5 Lower outlier limit: 9.5 - 1.5(1.6) = 7.1

replotting after applying cartesian limits to remove the upper outliers:



This variable still appears heavily skewed to the right.

For all of the above plots, I chose binwidths that didn't overly smooth out the distribution, but not so small that it wasn't possible to observe the shape of the distribution.

There are a few observations I made about these single variable plots:

- Most variables appear to be normally distributed or skewed to the right. I did not notice any variables that were skewed to the left.
- Variables that appear to be slightly skewed to the right are fixed acidity, volatile acidity, free sulfur dioxide, sulphates, and alcohol. That is most wines in the data set have low values for these variables.
- Variables showing a long tail are alcohol and sulphates.
- The histogram for citric acid appeared to be somewhat random, with most values at zero.

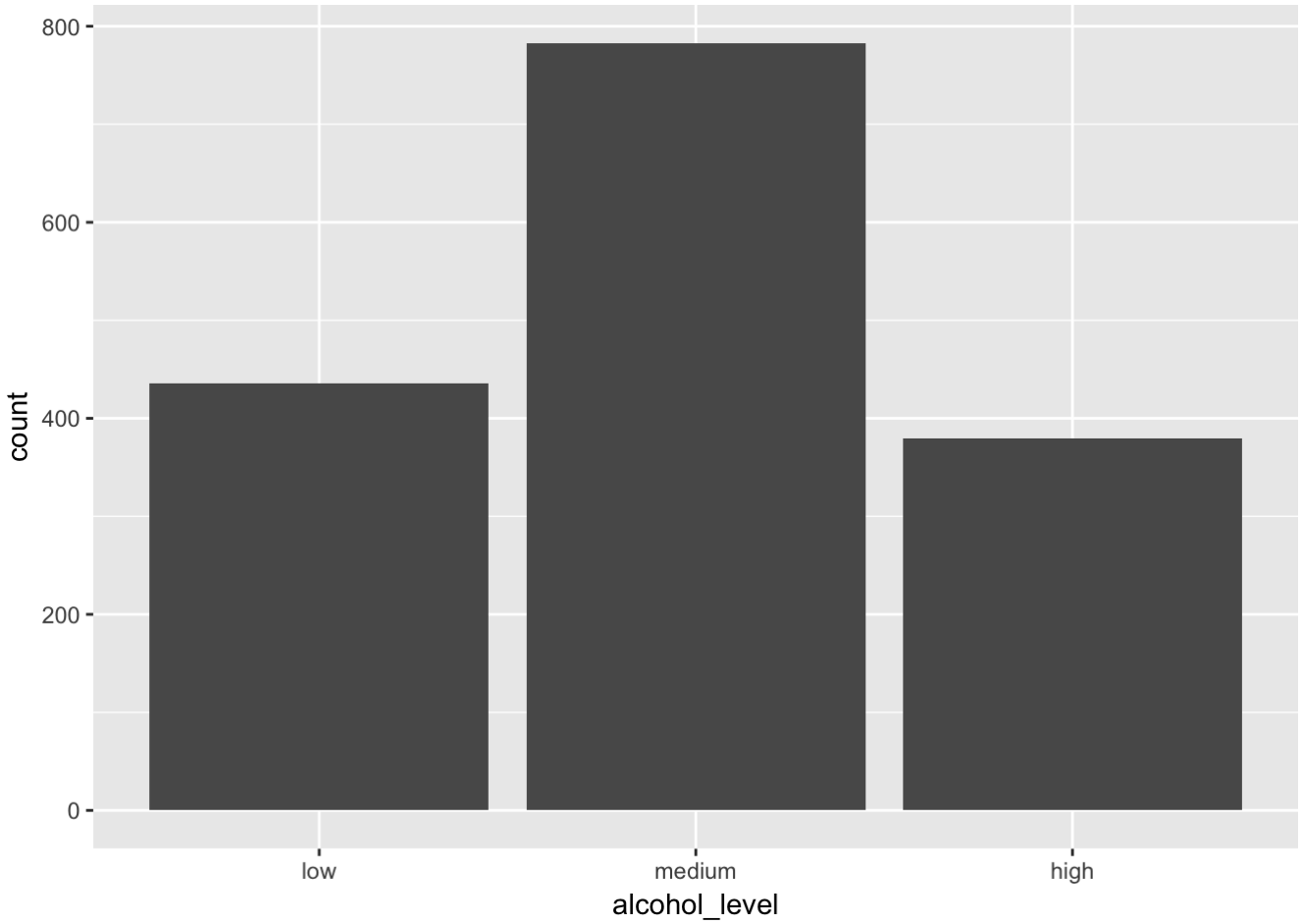
Finally, looking at the distribution for alcohol content the distribution appears to be somwhat skewed to the right, with the largest population around 9.5%.

In addition to the univariate plots shown above, I created another variable, called alcohol_level, which is a categorical variable for the alcohol level. I divided the alcohol level into "low", "medium", and "high", using the levels:

Alcohol percentage of 7.4 - 9.5: low Alcohol percenrage of 9.5 - 11.1: medium Alcohol percentage of 11.1 - 15: high

The plot below shows the distribution of this variable:

```
##   X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 1           7.4             0.70        0.00            1.9     0.076
## 2 2           7.8             0.88        0.00            2.6     0.098
## 3 3           7.8             0.76        0.04            2.3     0.092
## 4 4          11.2             0.28        0.56            1.9     0.075
## 5 5           7.4             0.70        0.00            1.9     0.076
## 6 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
## 4                  17                   60  0.9980 3.16      0.58     9.8
## 5                  11                   34  0.9978 3.51      0.56     9.4
## 6                  13                   40  0.9978 3.51      0.56     9.4
##   quality alcohol_level
## 1       5           low
## 2       5        medium
## 3       5        medium
## 4       6        medium
## 5       5           low
## 6       5           low
```



Based on the distribution of the alcohol level, most wines in the dataset fall into the medium alcohol level (that is, an alcohol % of 9.5 - 11.1).

What is the structure of your dataset?

The dataset contains 1599 rows and 13 variables. There are no null values. I also determined that the X column and the quality variables are integers while all other variables are numbers.

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

The main feature of interest in the dataset is quality rating.

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

The features in the dataset that would support the investigation are the physicochemical properties in the dataset. To name a few, these are alcohol, residual sugar, sulphates, etc.

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

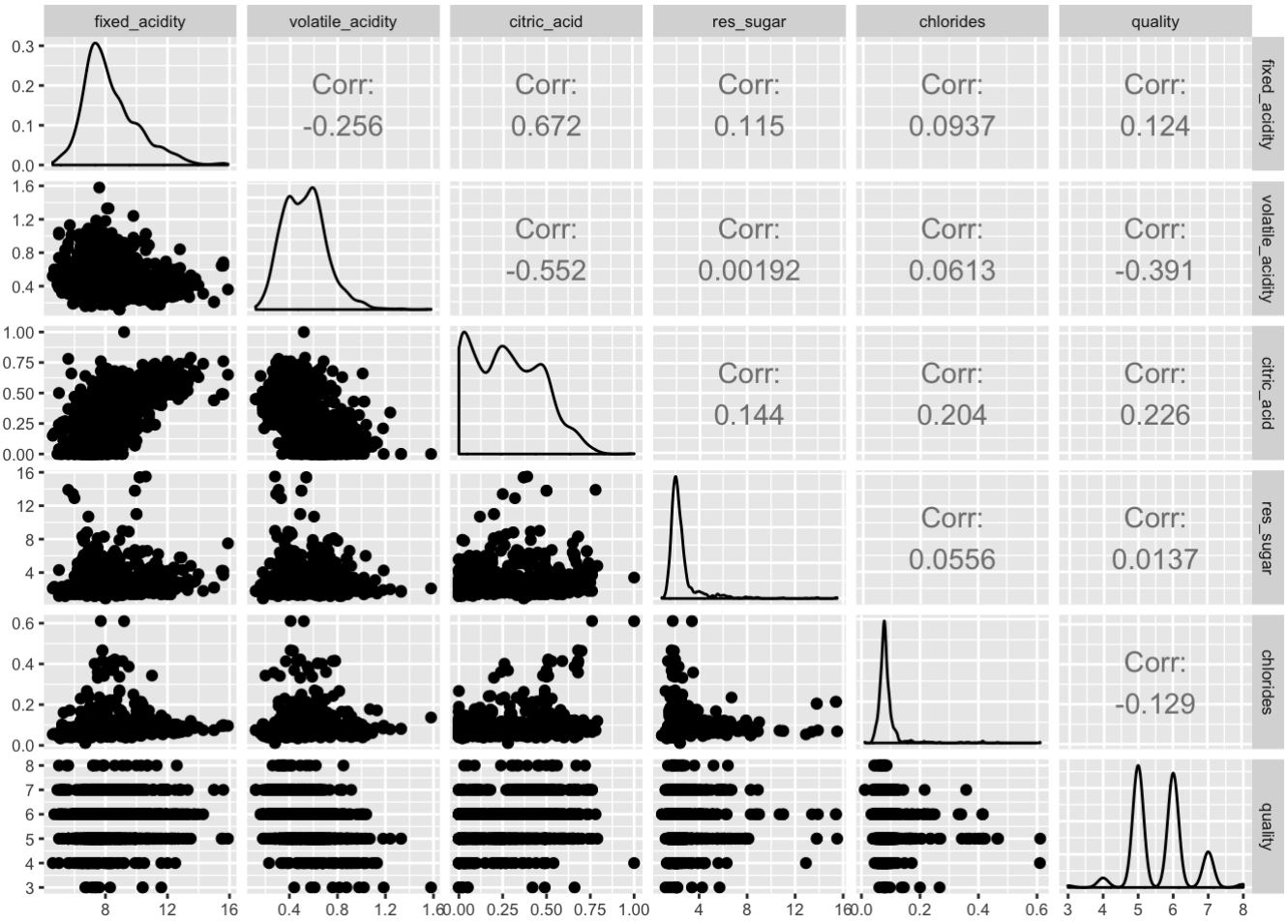
The additional variables were alcohol_level (based on the quartile ranges).

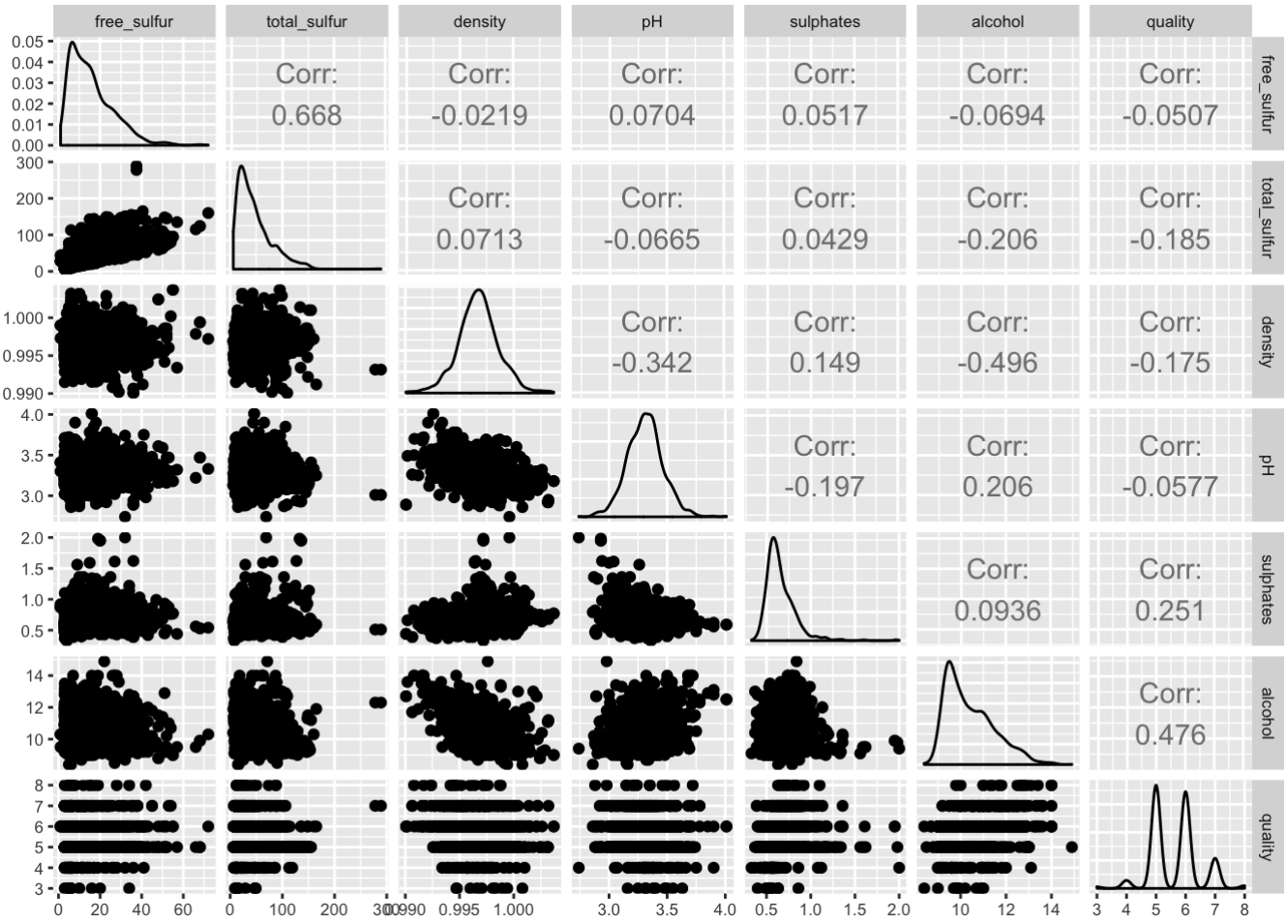
Of the features you investigated, were there any unusual distributions?

The distributions are descibed above. 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? The data was already tidy, I didn't need to perform any additional operations.

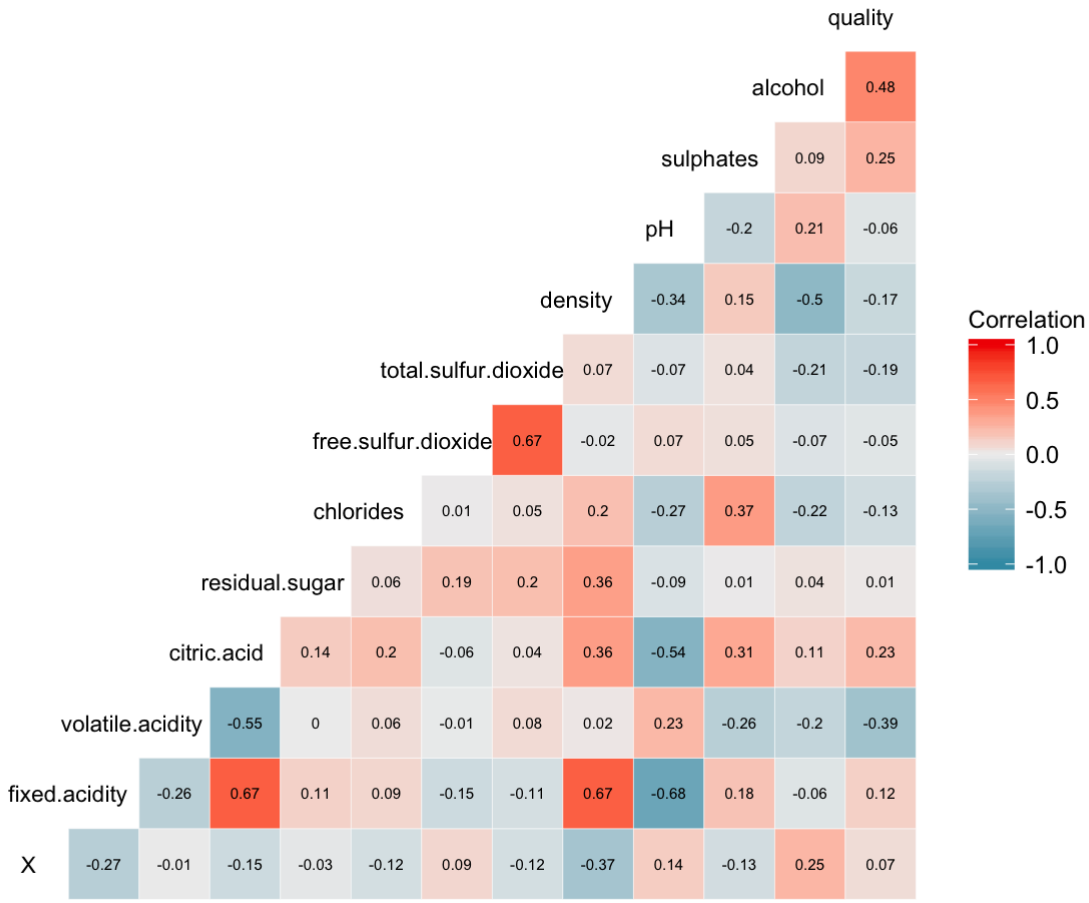
Bivariate Plots Section

To start to gain an understanding of a possible correlation between the physicochemical properties and quality, I plotted a scatterplot matrix of all the variables in the dataset. I made a few interesting observations.


fixed_acidity
volatile_acidity
citric_acid
res_sugar
chlorides
quality
Corr: -0.256
Corr: 0.672
Corr: 0.115
Corr: 0.0937
Corr: 0.124
Corr: -0.552
Corr: 0.00192
Corr: 0.0613
Corr: -0.391
Corr: 0.144
Corr: 0.204
Corr: 0.226
Corr: 0.0556
Corr: 0.0137
Corr: -0.129


free_sulfur
total_sulfur
density
pH
sulphates
alcohol
quality
Corr: 0.668
Corr: -0.0219
Corr: 0.0704
Corr: 0.0517
Corr: -0.0694
Corr: -0.0507
Corr: 0.0713
Corr: -0.0665
Corr: 0.0429
Corr: -0.206
Corr: -0.185
Corr: -0.342
Corr: 0.149
Corr: -0.496
Corr: -0.175
Corr: -0.197
Corr: 0.206
Corr: -0.0577
Corr: 0.0936
Corr: 0.251
Corr: 0.476

Another way to summarize the correlation coefficients is using ggcorr:



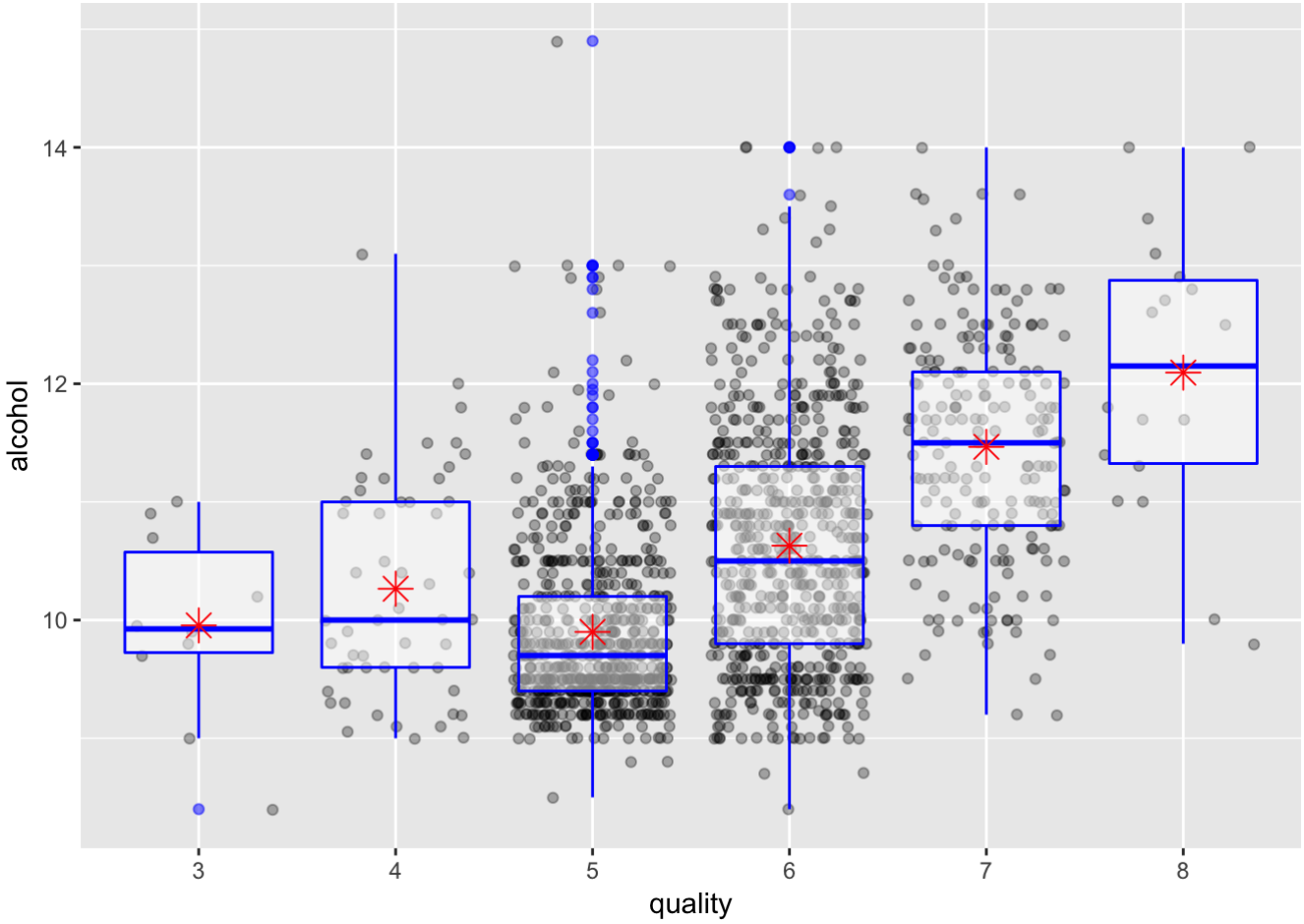
The above ggcorr visualization more clearly summarizes the correlation coefficients for all the variables in the dataset.

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?

Using the scatterplot matrix, I looked at the variable quality and looked at the correlation coefficients to determine where if at all there is a correlation, specifically with quality. The greatest correlation is a value of 0.476, which is the correlation coefficient between the two variables alcohol (which refers to alcohol content) and quality. To further investigate this, I created a boxplot of alcohol content and quality, and use color to differentiate between the different alcohol levels.

To further clarify, I created a box plot summarizing the alcohol content of different quality ratings. Alcohol content does seem to impact quality rating, but only when the alcohol content is above 10. It's also interesting to look at the median of the alcohol % for each quality level.




```
##   X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 1           7.4             0.70        0.00            1.9     0.076
## 2 2           7.8             0.88        0.00            2.6     0.098
## 3 3           7.8             0.76        0.04            2.3     0.092
## 4 4          11.2             0.28        0.56            1.9     0.075
## 5 5           7.4             0.70        0.00            1.9     0.076
## 6 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
## 4                  17                   60  0.9980 3.16      0.58     9.8
## 5                  11                   34  0.9978 3.51      0.56     9.4
## 6                  13                   40  0.9978 3.51      0.56     9.4
##   quality alcohol_sulphates
## 1       5              9.96
## 2       5             10.48
## 3       5             10.45
## 4       6             10.38
## 5       5              9.96
## 6       5              9.96
```

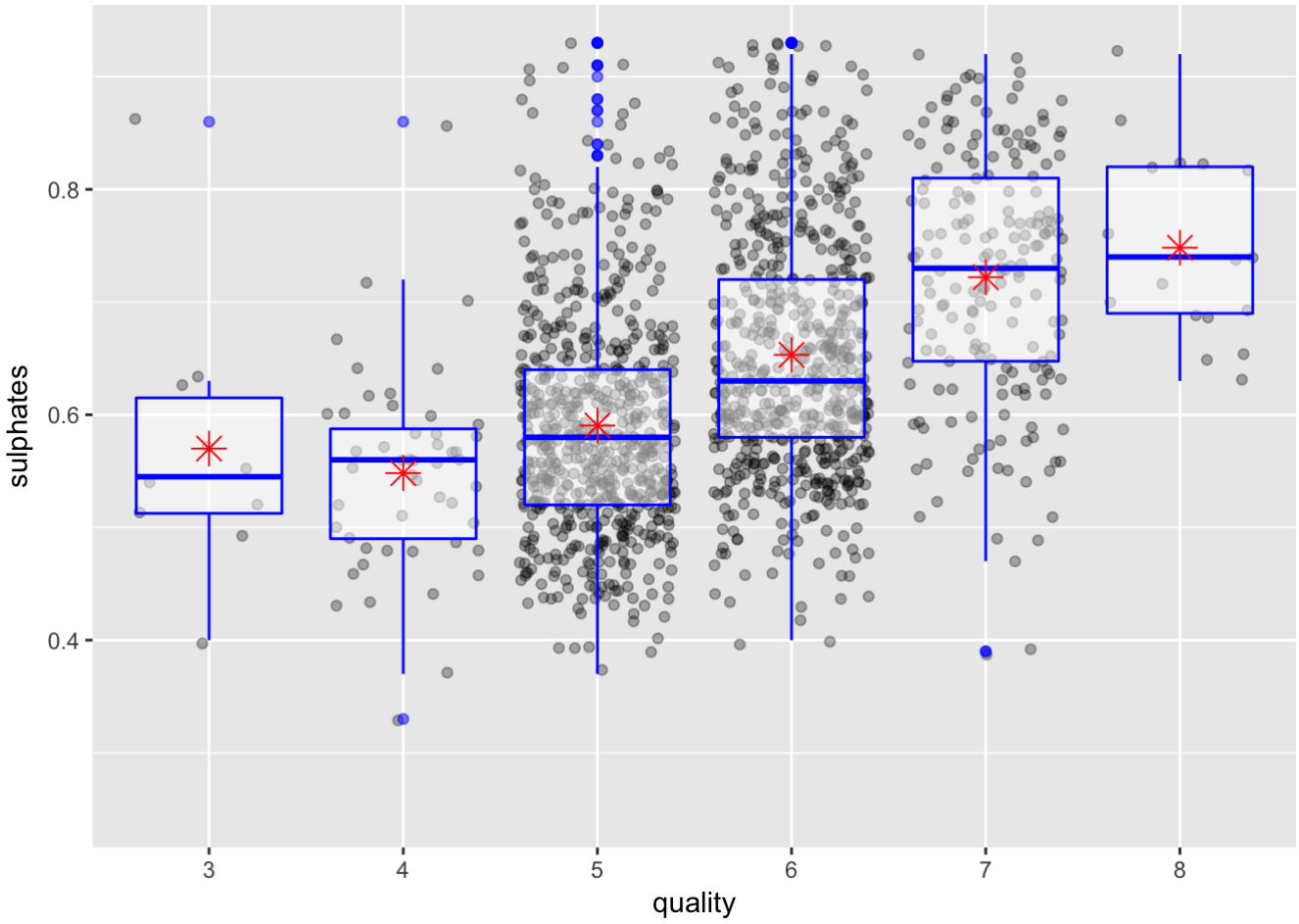
As seen from the box plot, the quality rating seems to be increasing with the increase in the alcohol_median, but only above an alcohol media of 10 % and a quality rating of 5.

Another possible correlation is quality with sulphates. The correlation coefficient is 0.251. I plotted a box plot to look at this relationship. I used limits to get a better view of the data without the outliers. When I decreased the scale of the y aixs to 95% of the values on the upper limit, the trend became a little more obvious. It appears that the between the wine quality values of 4 and 7, the amount of sulphates is higher. and then levels off between the quality ratings of 7 and 8. Just like in the previous set of graphs, I thought it would be interesting to calculate the median sulphate per quality rating and plot that vs. quality rating.

```
## Warning: Removed 79 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 79 rows containing non-finite values (stat_summary).
```

```
## Warning: Removed 83 rows containing missing values (geom_point).
```

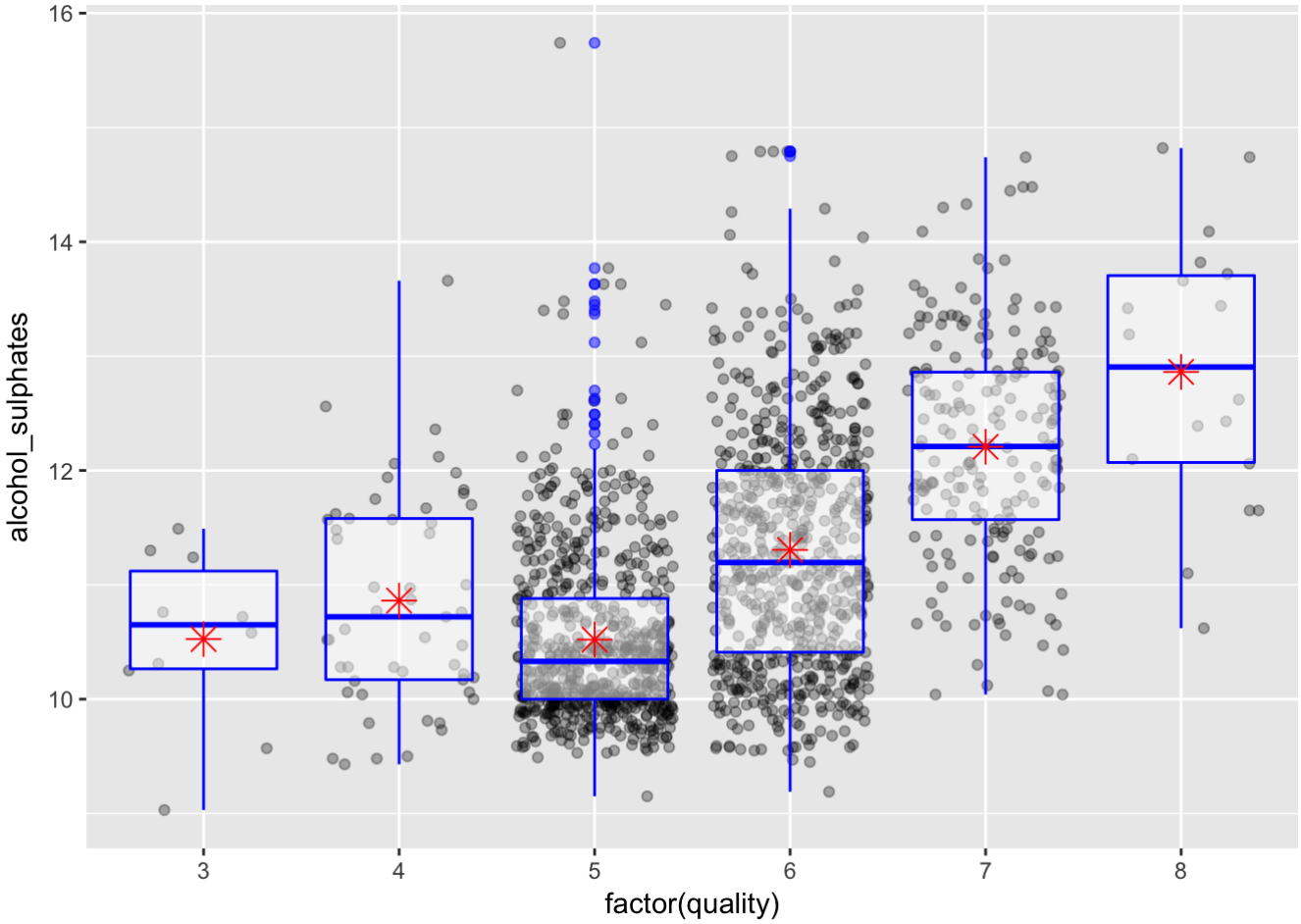


```
##   X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 1           7.4             0.70        0.00            1.9     0.076
## 2 2           7.8             0.88        0.00            2.6     0.098
## 3 3           7.8             0.76        0.04            2.3     0.092
## 4 4          11.2             0.28        0.56            1.9     0.075
## 5 5           7.4             0.70        0.00            1.9     0.076
## 6 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                  34  0.9978 3.51      0.56     9.4
## 2                  25                  67  0.9968 3.20      0.68     9.8
## 3                  15                  54  0.9970 3.26      0.65     9.8
## 4                  17                  60  0.9980 3.16      0.58     9.8
## 5                  11                  34  0.9978 3.51      0.56     9.4
## 6                  13                  40  0.9978 3.51      0.56     9.4
##   quality alcohol_sulphates
## 1       5              9.96
## 2       5             10.48
## 3       5             10.45
## 4       6             10.38
## 5       5              9.96
## 6       5              9.96
```

The box plot of sulphates vs. quality rating appear that as sulphates increase, the quality of the wine increases. Sulphates seem to be increasing with increasing wine quality rating. Unlike the alcohol content plots, this plot shows a more dramatic increase, followed by a levelling off at a quality rating of 7.

Since I noticed that both alcohol and sulphates may be correlated with quality, I created an additional variable, the sum of alcohol content and sulphates, and plotted that against quality. I also plotted a summary of the median alcohol_sulphates variable vs. Quality for each quality rating.

```
##   X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 1           7.4             0.70        0.00            1.9     0.076
## 2 2           7.8             0.88        0.00            2.6     0.098
## 3 3           7.8             0.76        0.04            2.3     0.092
## 4 4          11.2             0.28        0.56            1.9     0.075
## 5 5           7.4             0.70        0.00            1.9     0.076
## 6 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                  34  0.9978 3.51      0.56     9.4
## 2                  25                  67  0.9968 3.20      0.68     9.8
## 3                  15                  54  0.9970 3.26      0.65     9.8
## 4                  17                  60  0.9980 3.16      0.58     9.8
## 5                  11                  34  0.9978 3.51      0.56     9.4
## 6                  13                  40  0.9978 3.51      0.56     9.4
##   quality alcohol_sulphates
## 1       5              9.96
## 2       5             10.48
## 3       5             10.45
## 4       6             10.38
## 5       5              9.96
## 6       5              9.96
```

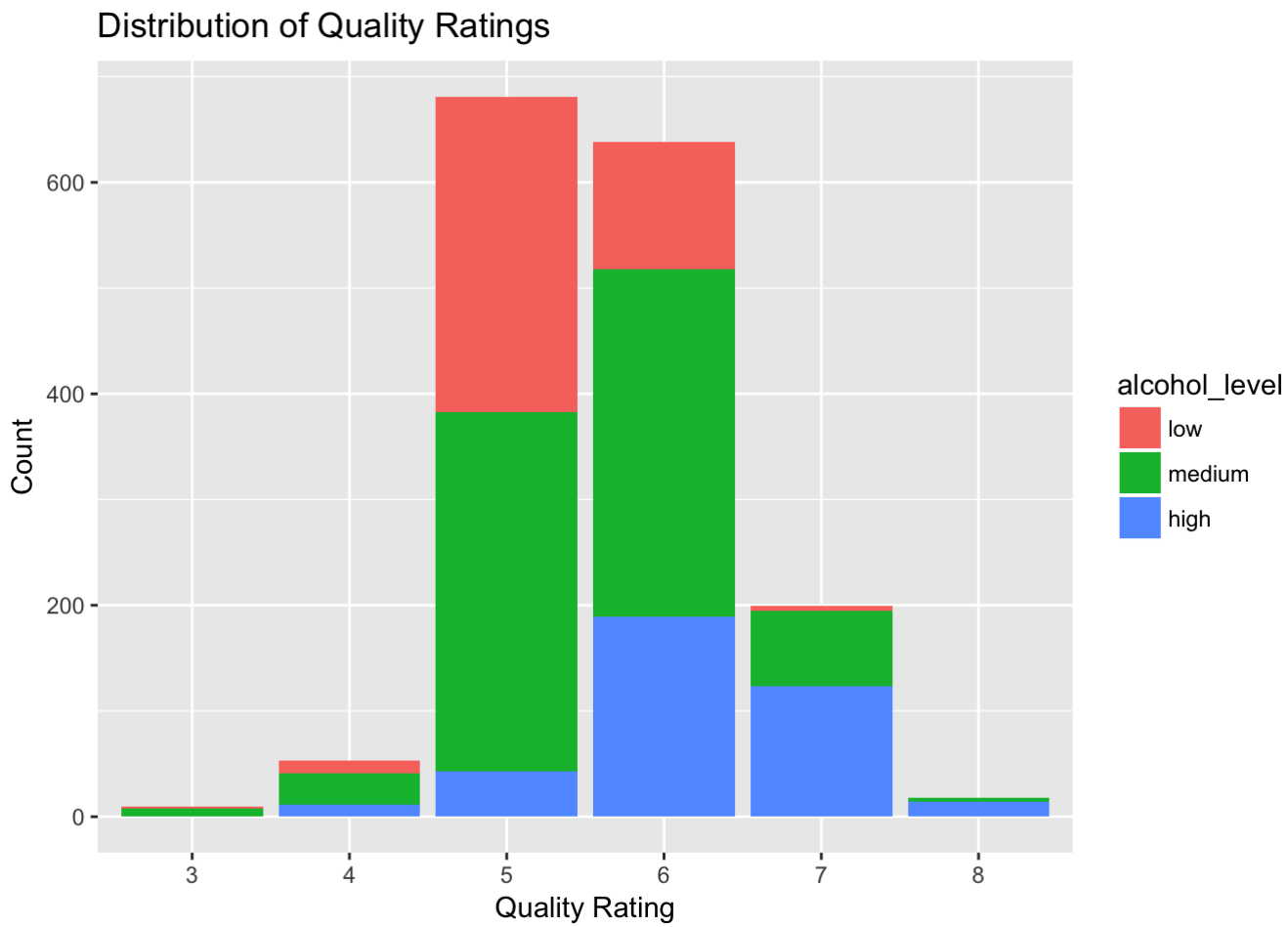



```
## <ScaleContinuousPosition>
##  Range:
##  Limits:  9.5 -- 13.2
```

Looking at the above box plot of alcohol + sulphates vs. quality, the trend appears very similar to alcohol vs. quality and sulphates vs. quality. There is a dip at the quality rating of 5, but overall there appears to be an increase in quality rating with an increase in the variable alcohol + sulphates.

To further explore the relationship between alcohol content and quality, I made a bar graph summarizing the different alcohol levels, shaded by quality:

```
##   X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 1           7.4             0.70        0.00            1.9     0.076
## 2 2           7.8             0.88        0.00            2.6     0.098
## 3 3           7.8             0.76        0.04            2.3     0.092
## 4 4          11.2             0.28        0.56            1.9     0.075
## 5 5           7.4             0.70        0.00            1.9     0.076
## 6 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
## 4                  17                   60  0.9980 3.16      0.58     9.8
## 5                  11                   34  0.9978 3.51      0.56     9.4
## 6                  13                   40  0.9978 3.51      0.56     9.4
##   quality alcohol_sulphates alcohol_level
## 1       5              9.96           low
## 2       5             10.48        medium
## 3       5             10.45        medium
## 4       6             10.38        medium
## 5       5              9.96           low
## 6       5              9.96           low
```

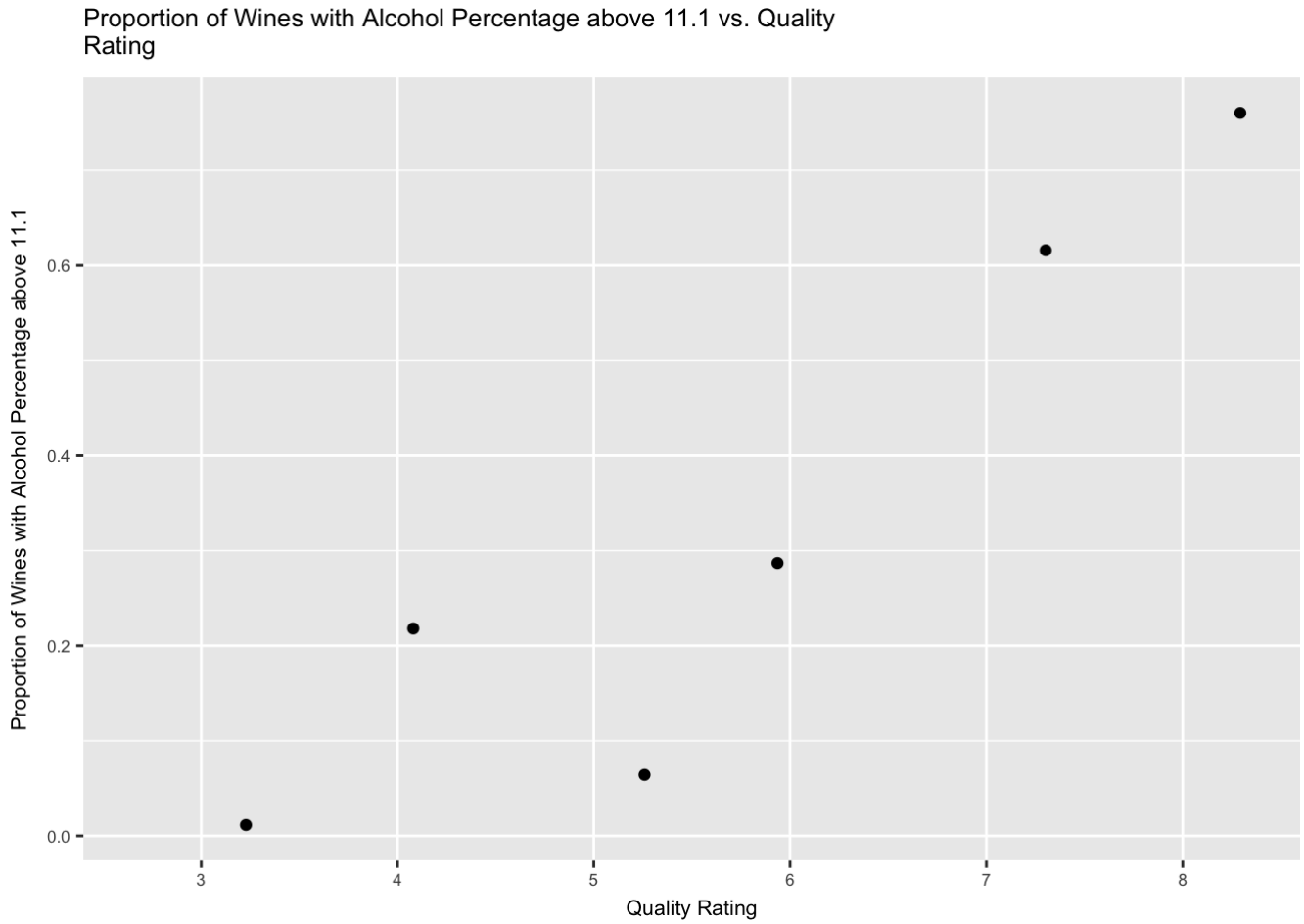


It's interesting to note that the area in blue in the graph above (representing the "high" alcohol level, that is, a percentage greater than 11.10) seems to be increasing with the quality rating and the portion of the bar in green seems to be decreasing. This bar graph suggests that the proportion of wines that are higher in alcohol % is

greater in wines with a higher quality rating when compared to a lower quality rating. To determine if this was the case, I calculated the proporting of high alcohol level wines in each quality group and plotted vs. quality:

```
##   X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 1           7.4             0.70        0.00            1.9     0.076
## 2 2           7.8             0.88        0.00            2.6     0.098
## 3 3           7.8             0.76        0.04            2.3     0.092
## 4 4          11.2             0.28        0.56            1.9     0.075
## 5 5           7.4             0.70        0.00            1.9     0.076
## 6 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
## 4                  17                   60  0.9980 3.16      0.58     9.8
## 5                  11                   34  0.9978 3.51      0.56     9.4
## 6                  13                   40  0.9978 3.51      0.56     9.4
##   quality alcohol_level
## 1       5           low
## 2       5        medium
## 3       5        medium
## 4       6        medium
## 5       5           low
## 6       5           low
```

```
## # A tibble: 6 x 5
##   quality num_high_alc `n()` mean_alcohol proportion_high
##     <int>        <int> <int>        <dbl>           <dbl>
## 1       3            0    10         9.96          0
## 2       4           11    53        10.3           0.208
## 3       5           43   681         9.90          0.0631
## 4       6          189   638        10.6           0.296
## 5       7          123   199        11.5           0.618
## 6       8           14    18        12.1           0.778
```



Interestingly, there does appear to be an upward trend in the proportion of wines with an alcohol % of greater than 11.10 as the quality rating goes up. There is a dip at the quality rating of 5, but in general there appears to be an upward trend. What's interesting also is that at a quality rating of 8, almost 80% of the wines are above an alcohol percentage of 11.10.

Did you observe any interesting relationships between the other features

(not the main feature(s) of interest)?

I found some relationships the can be explained by the physicochemical properties. For example, acetic acid content decreases the pH. There was a correlation coefficient of -0.496 between alcohol content and density. Another relationship that makes sense is the relationship between pH and fixed acidity (pH decreases as fixed acidity increases, correlation -0.683). These variables are due to nature (that is, pH is defined by acidity).

What was the strongest relationship you found?

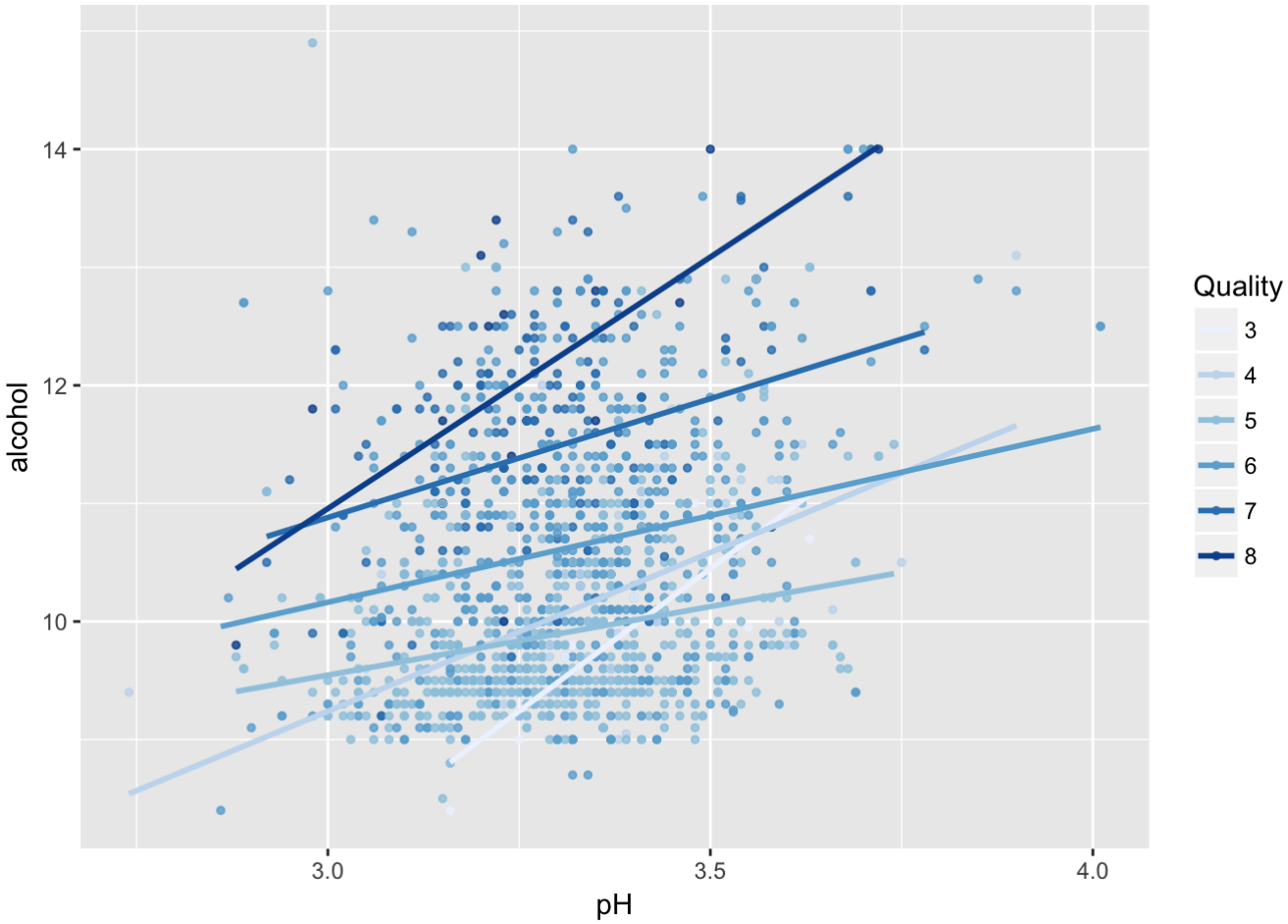
The strongest relationship that I found was between alcohol content and quality.

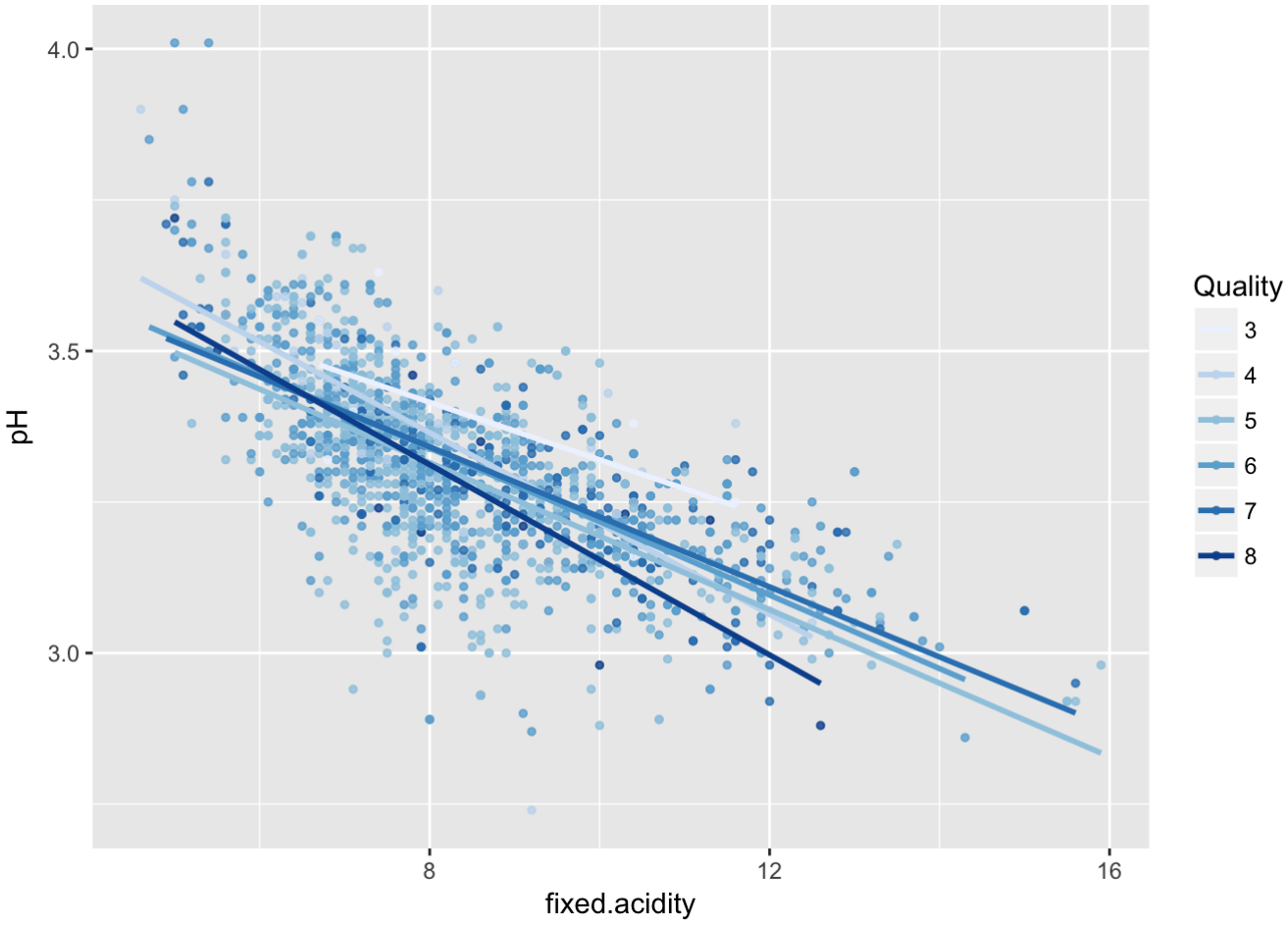
Multivariate Plots Section

Although based on what I've seen so far with the correlation coefficients and the visualizations, it appears that alcohol content has the greatest relationship to quality rating, with sulphates also appearing to be somewhat correlated with quality. However, the multivariate plots below look to further investigate any additional

relationships between variables.

In addition to alcohol content and sulphates, I'll plot variables that are related (based on the scatter plot matrix) such as pH and fixed acidity.


alcohol
14
12
10
3.0
3.5
4.0
pH
Quality
3
4
5
6
7
8


pH
4.0
3.5
3.0
8
12
16
fixed.acidity
Quality
3
4
5
6
7
8

```
## Warning: Removed 14 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 14 rows containing missing values (geom_point).
```

Looking at the first multivariate scatterplots in the above section (alcohol vs pH shaded by quality with a linear regression model added) for quality ratings 5, 6, 7, and 8, the linear model gets steeper with each increasing quality rated wine. In other words, there appears to be a stronger relationship of alcohol % to pH for higher quality wines when compared to lower quality wines.

For the plot of pH vs. fixed acidity, the relationship between pH and fixed acidity makes sense, since increasing acidity naturally decreased the pH. However, it's difficult to make out any patterns with regards to quality in this graph.

The third multivariate plot reflects the same observation I made previously. That is, lower alcohol and lower sulphates appear to be correlated with lower quality, and higher sulphates and higher alcohol % seem to be correlated with higher quality ratings. That is, the darker lines corresponding to higher quality wines are higher on the graph (corresponding to higher sulphates). It's less clear with alcohol content.

As a follow up to the third graph, I am calculating the correlation coefficient of the alcohol+ sulphates variable with quality.

```
##
##  Pearson's product-moment correlation
##
## data:  pf$alcohol_sulphates and pf$quality
## t = 23.227, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4649334 0.5382637
## sample estimates:
##       cor
## 0.5025017
```

```
##
##  Pearson's product-moment correlation
##
## data:  alcohol_sulphates and quality
## t = 23.227, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4649334 0.5382637
## sample estimates:
##       cor
## 0.5025017
```

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

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

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

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

Multivariate Analysis

Adding the sulphates and the alcohol variables together was an attempt to see an additive effect of both variables on quality, since looking at each individually there appeared to be a small correlation. This can also be done using a linear model.

```
##   X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 1           7.4             0.70        0.00            1.9     0.076
## 2 2           7.8             0.88        0.00            2.6     0.098
## 3 3           7.8             0.76        0.04            2.3     0.092
## 4 4          11.2             0.28        0.56            1.9     0.075
## 5 5           7.4             0.70        0.00            1.9     0.076
## 6 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
## 4                  17                   60  0.9980 3.16      0.58     9.8
## 5                  11                   34  0.9978 3.51      0.56     9.4
## 6                  13                   40  0.9978 3.51      0.56     9.4
##   quality alcohol_sulphates
## 1       5              9.96
## 2       5             10.48
## 3       5             10.45
## 4       6             10.38
## 5       5              9.96
## 6       5              9.96
```

```
##
## Calls:
## m1: lm(formula = I(quality) ~ I(alcohol), data = pf)
## m2: lm(formula = I(quality) ~ I(alcohol) + sulphates, data = pf)
##
## ================================================
##                         m1             m2
## ------------------------------------------------
##   (Intercept)          1.875***       1.375***
##                       (0.175)        (0.177)
##   I(alcohol)           0.361***       0.346***
##                       (0.017)        (0.016)
##   sulphates                           0.994***
##                                      (0.102)
## ------------------------------------------------
##   R-squared            0.227          0.270
##   adj. R-squared       0.226          0.269
##   sigma                0.710          0.690
##   F                  468.267        294.988
##   p                    0.000          0.000
##   Log-likelihood   -1721.057      -1675.142
##   Deviance           805.870        760.894
##   AIC               3448.114       3358.284
##   BIC               3464.245       3379.793
##   N                 1599           1599
## ================================================
```

Were there any interesting or surprising interactions between features?

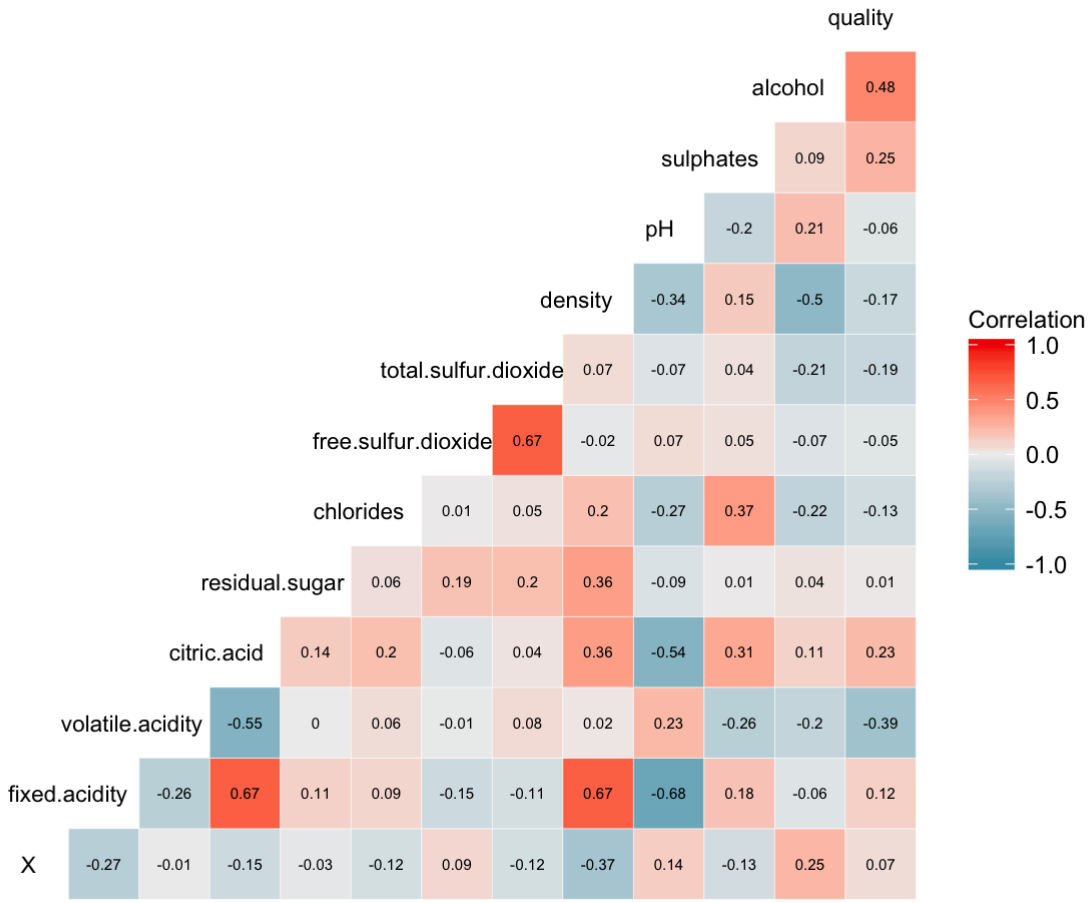
The interesting interaction was between alcohol and sulphates and the impact on quality. When added together these two variables resulted in a pearson's correlation coefficient of 0.5. It's surprising that sulphates would impact the correlation with quality. Without sulphates, the correlation coefficient was 0.476.

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

I created a linear model. The model included the variables alcohol and sulphates and quality. The model appears to be a weak model, with an R squared value of 0.227 without including sulphates and an R squared value of 0.270 after including sulphates.

Final Plots and Summary

Plot One

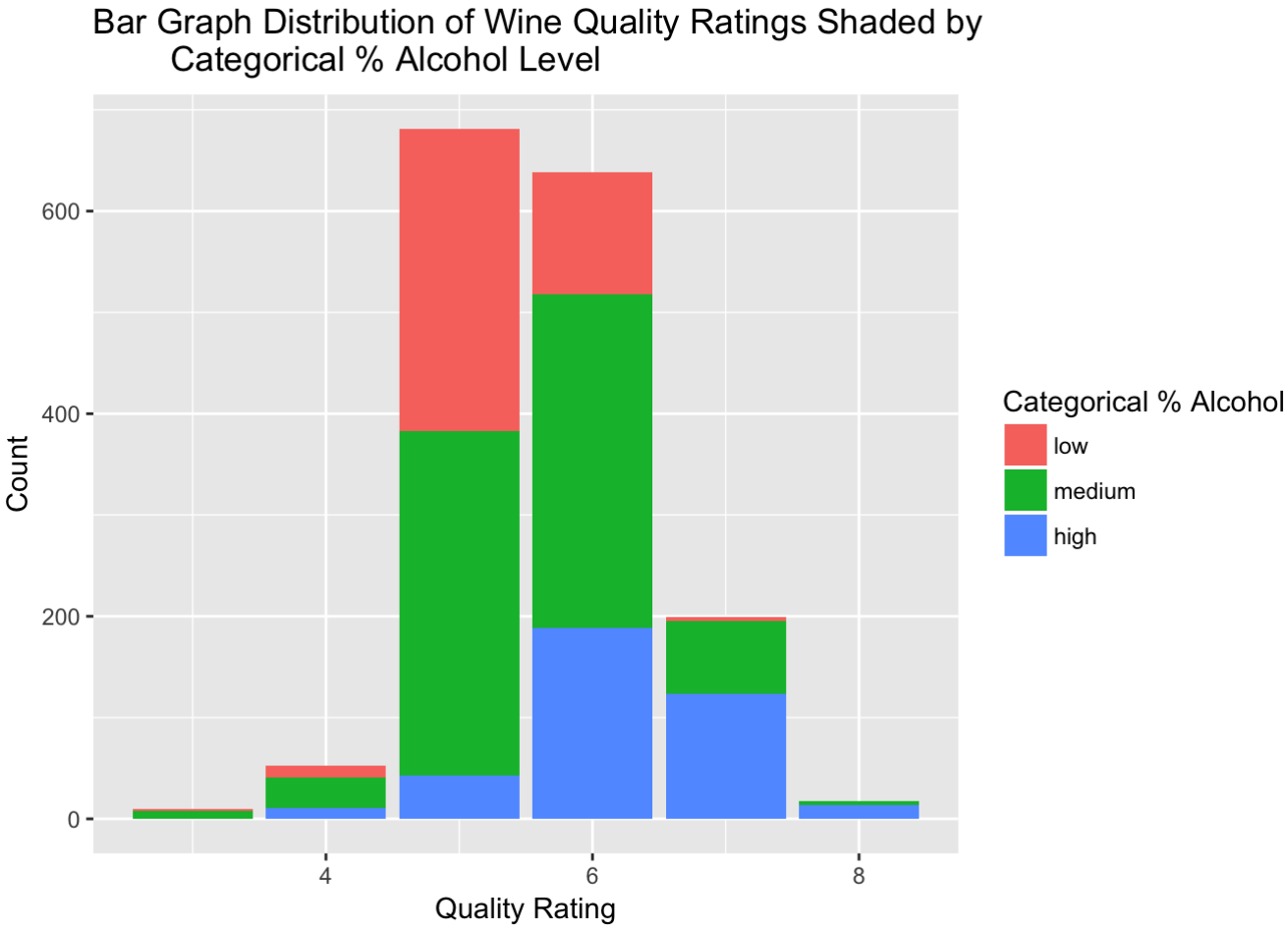


Description One

The first graph that I chose is the ggcorr summary with all of the original variables in the data set. I found this information most useful when starting the analysis. Looking at the correlation coefficients of quality with alcohol and sulphates, since they were the highest, helped in finding a few variables to start the analysis with.

Plot Two

```
##   X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 1           7.4             0.70        0.00            1.9     0.076
## 2 2           7.8             0.88        0.00            2.6     0.098
## 3 3           7.8             0.76        0.04            2.3     0.092
## 4 4          11.2             0.28        0.56            1.9     0.075
## 5 5           7.4             0.70        0.00            1.9     0.076
## 6 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                  34  0.9978 3.51      0.56     9.4
## 2                  25                  67  0.9968 3.20      0.68     9.8
## 3                  15                  54  0.9970 3.26      0.65     9.8
## 4                  17                  60  0.9980 3.16      0.58     9.8
## 5                  11                  34  0.9978 3.51      0.56     9.4
## 6                  13                  40  0.9978 3.51      0.56     9.4
##   quality alcohol_level
## 1       5           low
## 2       5        medium
## 3       5        medium
## 4       6        medium
## 5       5           low
## 6       5           low
```



Description Two

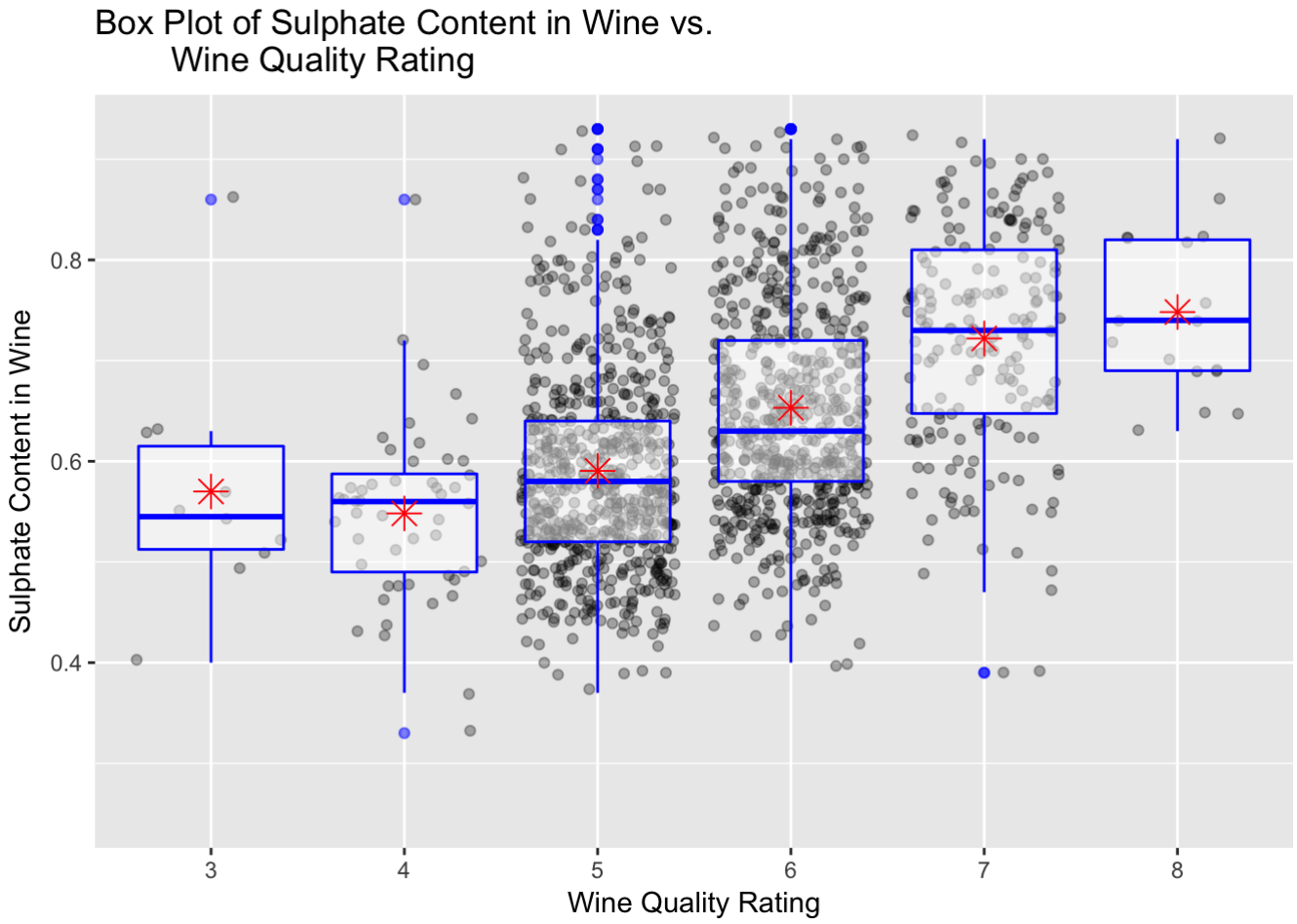
I chose this as my second graph because I thought it shows very nicely how the proportion of blue (A high alcohol level as defined in this analysis) in each graph increases with increasing quality rating. At the same the proportion low alcohol wines seems to be decreasing with increasing quality.

Plot Three

```
## Warning: Removed 79 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 79 rows containing non-finite values (stat_summary).
```

```
## Warning: Removed 87 rows containing missing values (geom_point).
```



Description Three

I chose this graph for my final figure because it summarizes one of the factors that I noticed had the strongest relationship to quality rating (sulphates). From this box plot, it's easy to see a gradual increase in wine quality rating with increasing sulphate concentration. ——

Reflection

I found the project very interesting, and enjoyed working through the analysis and finding which physicochemical proporties could impact the quality rating. I found that alcohol level (and possibly sulphate content) seemed to be related to quality rating. The difficulty came in explaining the dip in the alcohol content at the quality rating of 5. This could be due to several factors, but one that I can think of is the error/subjectiveness of the assignment of the quality rating to the different wines. The process is manual, and based on sensory response from the judges. Also, perhaps it is easier for a wine taster to discern a very good wine from a very bad wine, but the values in the middle may become more difficult to assign.

Another challenge comes from the limited number of ratings. The range of quality ratings span only from 3 to 8. This gives us a smaller space to work in with regards to finding any trends in the data.

The matrix scatterplot helped in terms of finding a place to start in terms of which variables could be related to wine quality. Perhaps I would've started with acidity instead of sulphates. Using the scatterplot I was able to narrow in on a couple of variables. It surprised me that sulphates were mildly related to rating, I wouln't have guessed that.

Future work that can be done with this dataset could entail expanding the dataset to include a wider range of quality ratings. This could possibly clarify any possible relationships between physicochemical proporties and wine quality.