RedWine by Ameen Syed

Dataset Information

This dataset is about the wine quality of the red variant of the Portuguese "Vinho Verde" wine. There are 1599 samples of red wine. The attributes of the dataset are fixed acidity (tartaric acid g/dmˆ3), volatile acidity (acetic acid g/dmˆ3), citric acid (g/dmˆ3), residual sugar (g/dmˆ3), chlorides (sodium chloride g/dmˆ3), free sulfur dioxide (mg/dmˆ3), total sulfur dioxide (mg/dmˆ3), density (g/cmˆ3), pH, sulphates (potassium sulfate g/dmˆ3), alcohol (% by volume), and quality (Output variable. Score from 0 to 10).

The 11 input variables were taken from physicochemical tests while the output variable, Quality, was determined by 3 wine experts.

Guiding Question:

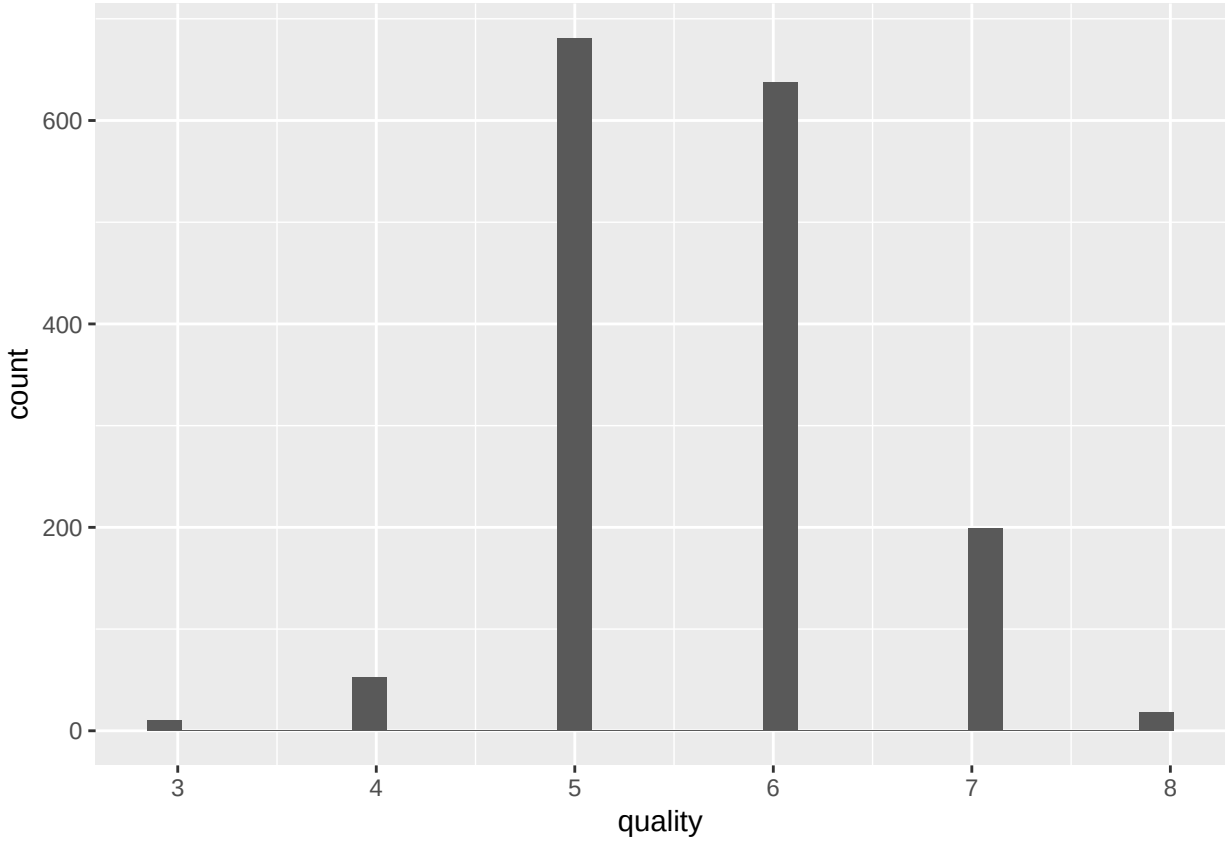
> Which chemical properties influence the quality of red wines?

Univariate Plots Section

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

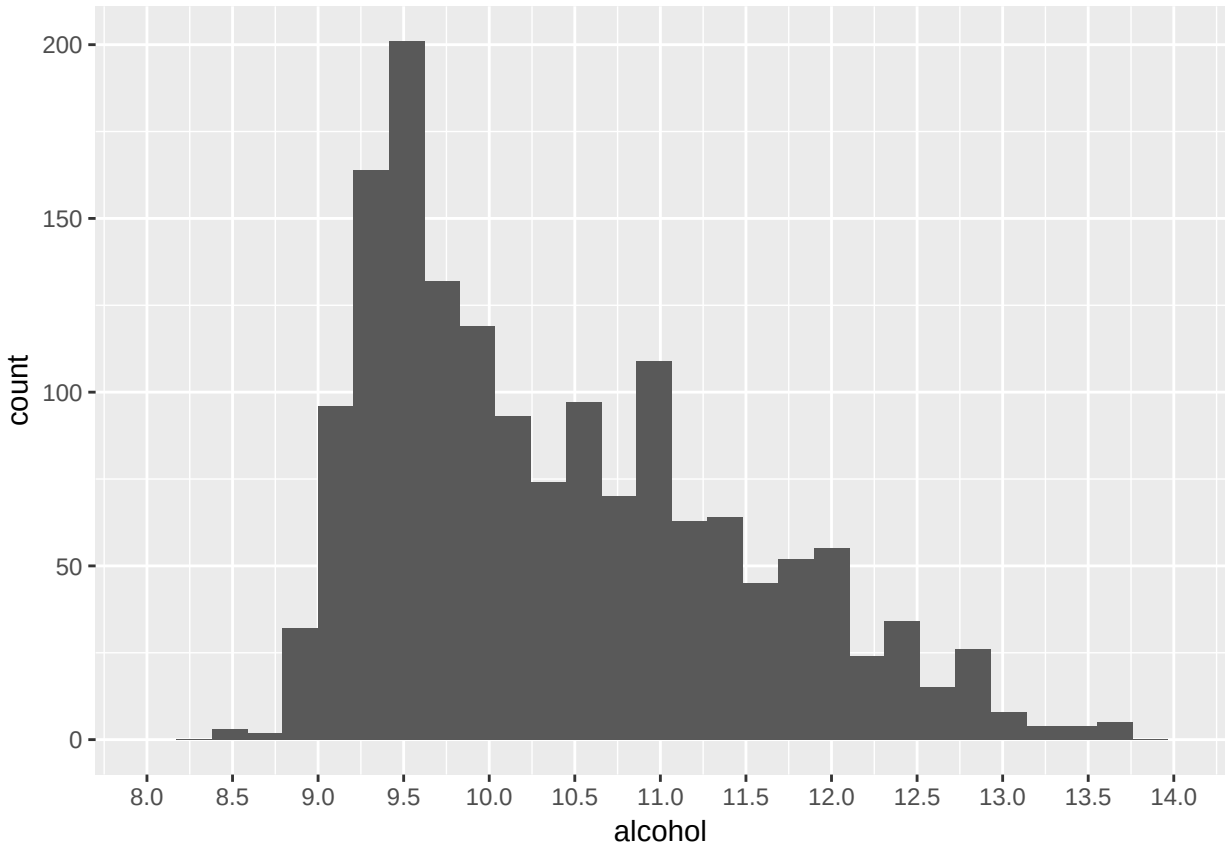
```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :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
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00
##  Median :0.07900   Median :14.00       Median : 38.00
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00
##     density             pH          sulphates         alcohol
##  Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40
##  1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50
##  Median :0.9968   Median :3.310   Median :0.6200   Median :10.20
##  Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42
##  3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10
##  Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.636
##  3rd Qu.:6.000
##  Max.   :8.000
```

There are 13 variables and 1599 observations. Since the variable X is just an ID for each sample, it was dropped when calculating the summary of the dataset.



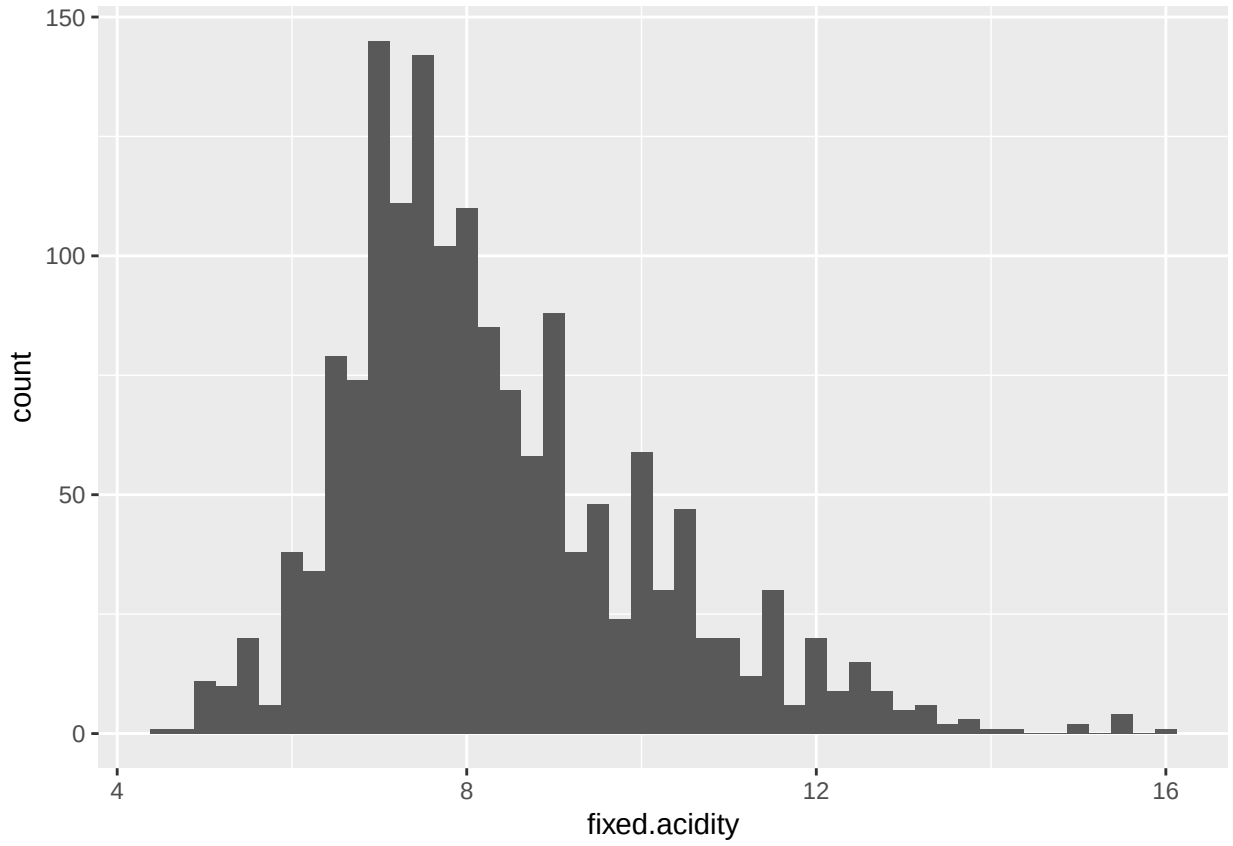

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.636   6.000   8.000
```

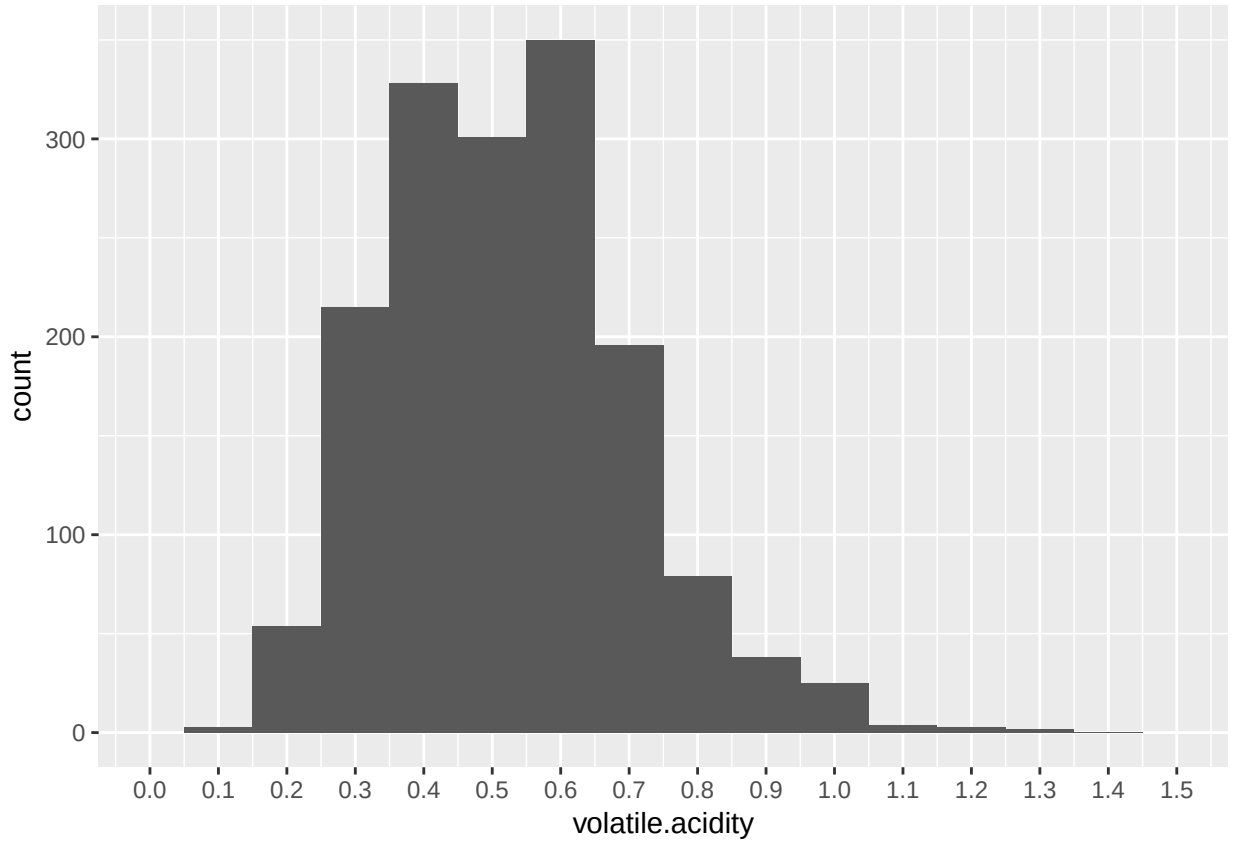
First off, let's look at the distribution of the quality of the samples. It is clear that most wines were of a quality of either 5,6, or 7. I think it will be important to see what is the difference between wines of quality 7 vs wines of quality 5, because there's an adequate amount of samples for the two quality levels.

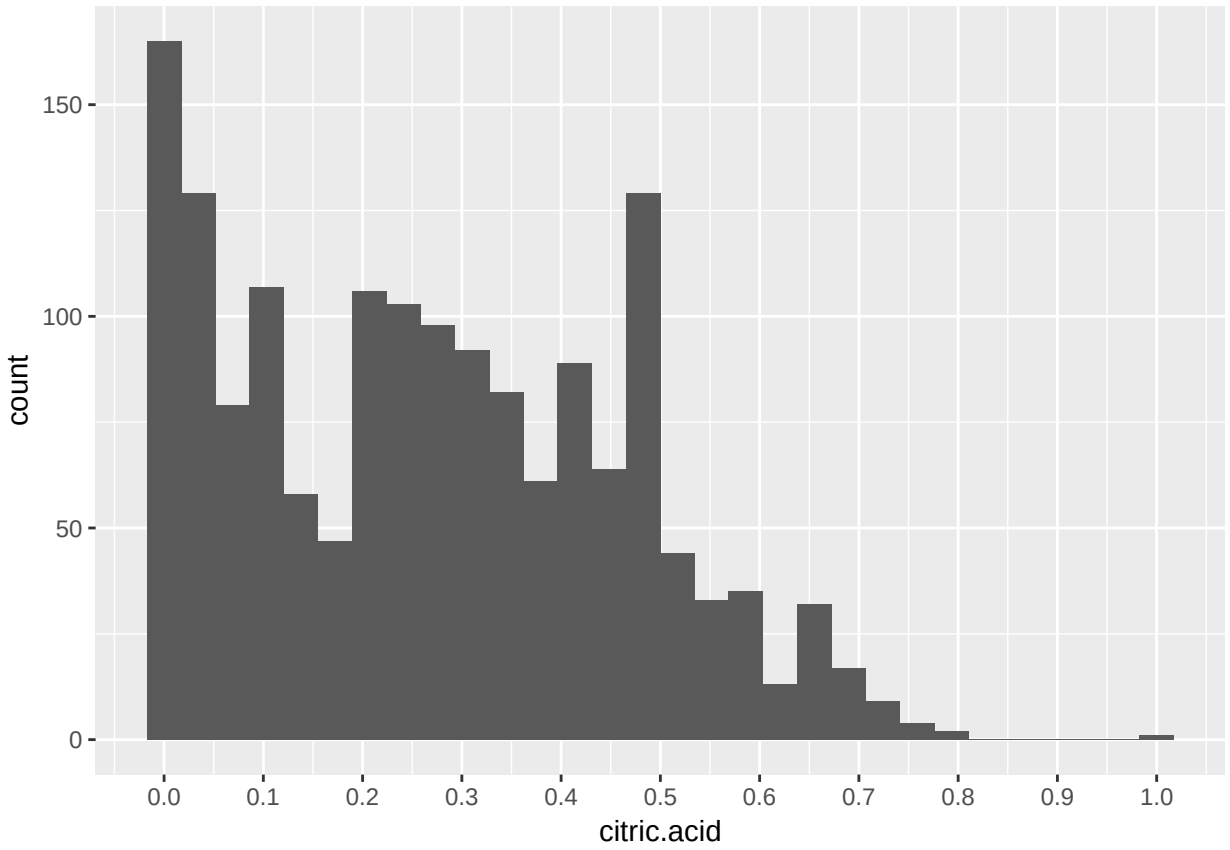



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```

Most of the alcohol distribution was between 9.25 and 9.75.These values are relatively far from the max.


150
100
50
0
count
4
8
12
16
fixed.acidity

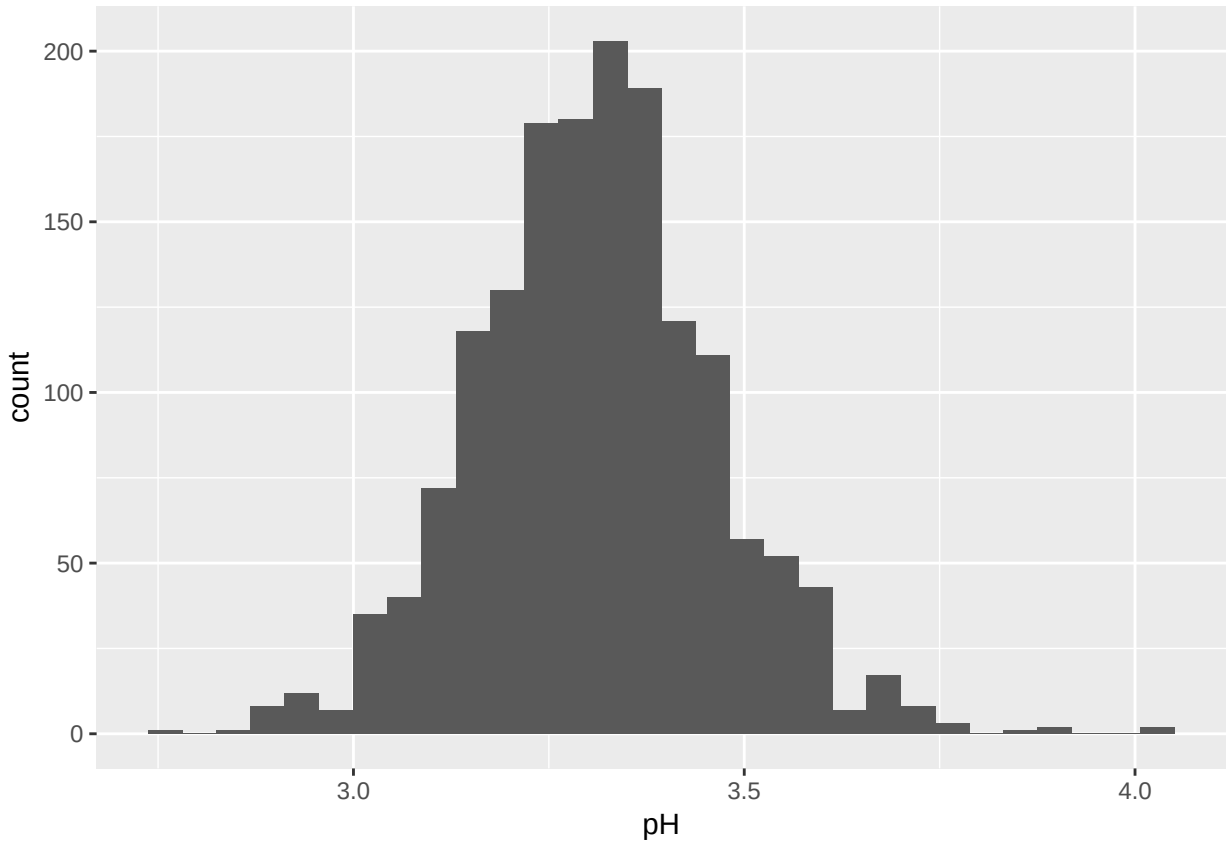

300
200
100
0
count
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
volatile.acidity



From the histograms of the acids, it appears as though in general, there are very few samples with extremely high concentrations of acid. Furthermore, fixed acidity and volatile acidity have a similar distribution where the largest frequencies occur in the lower-middle level of fixed acidity or volatile acidity. The distribution of citric acid is more visibly different than the other two distributions. There are many more samples with low levels of citric acid (0-0.1 out of 1) than there are for the other two acids. In other words, the median value seems to be further from the max value in citric acid than the other two acids. I wonder how the box plots will look like in the next section. For now, I've computed the summary statistics for citric acid.

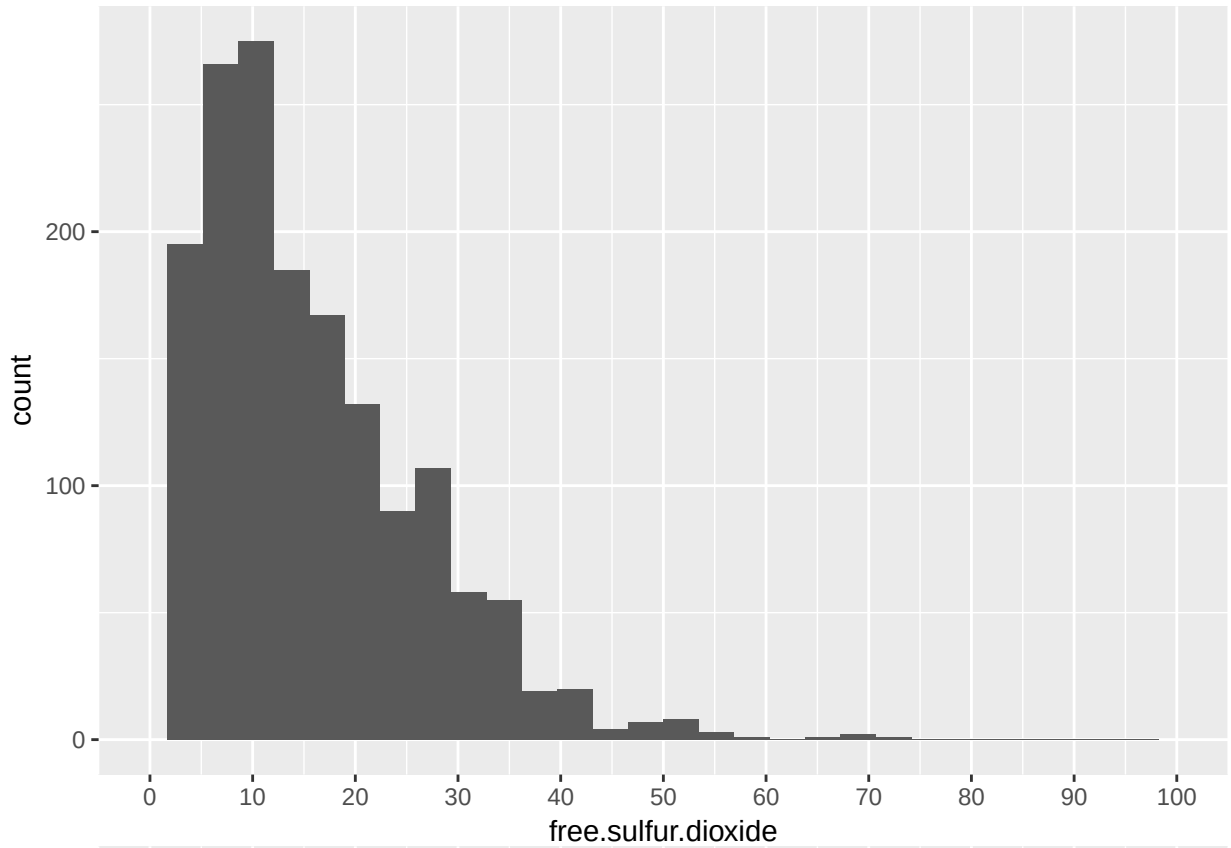
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```

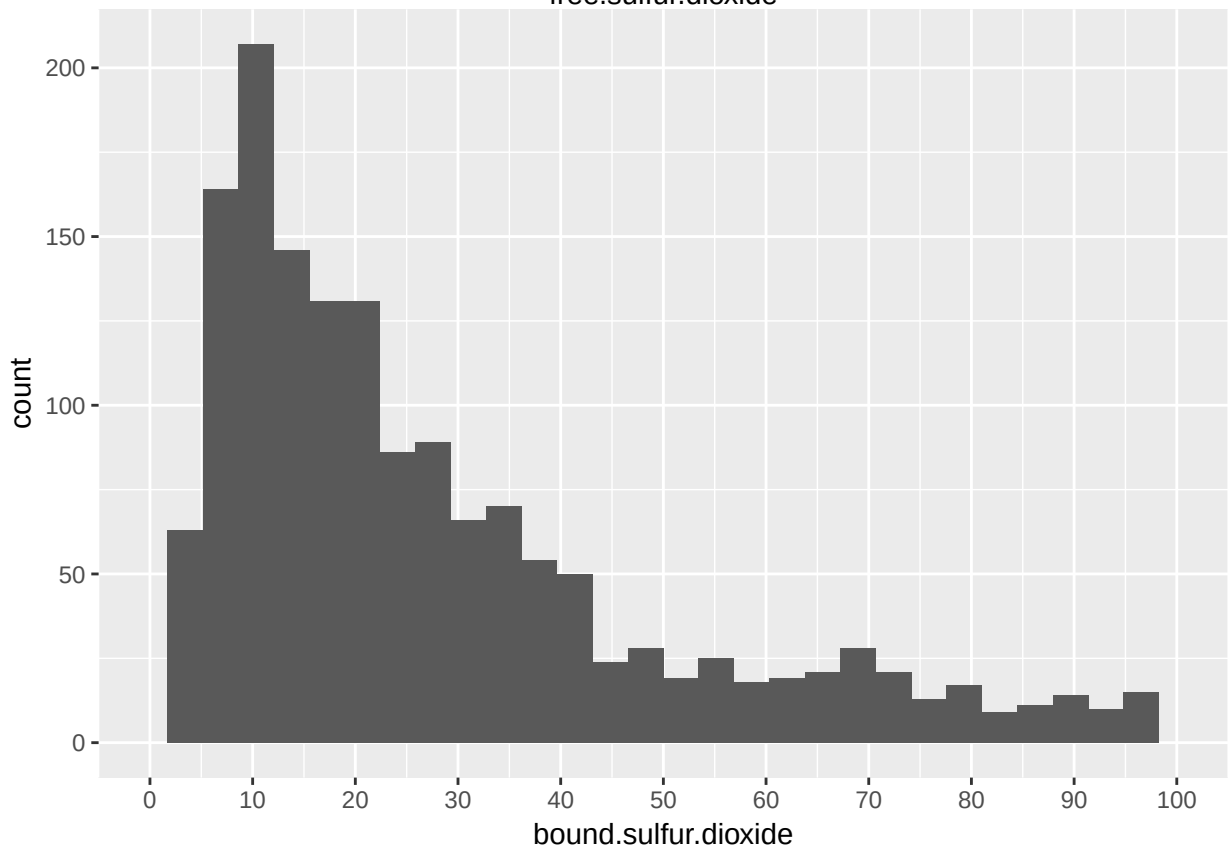
Most samples have a citric acid level of less than 0.5. This statistic confirms our observation than the distribution of citric acid levels is concentrated on the lower levels.

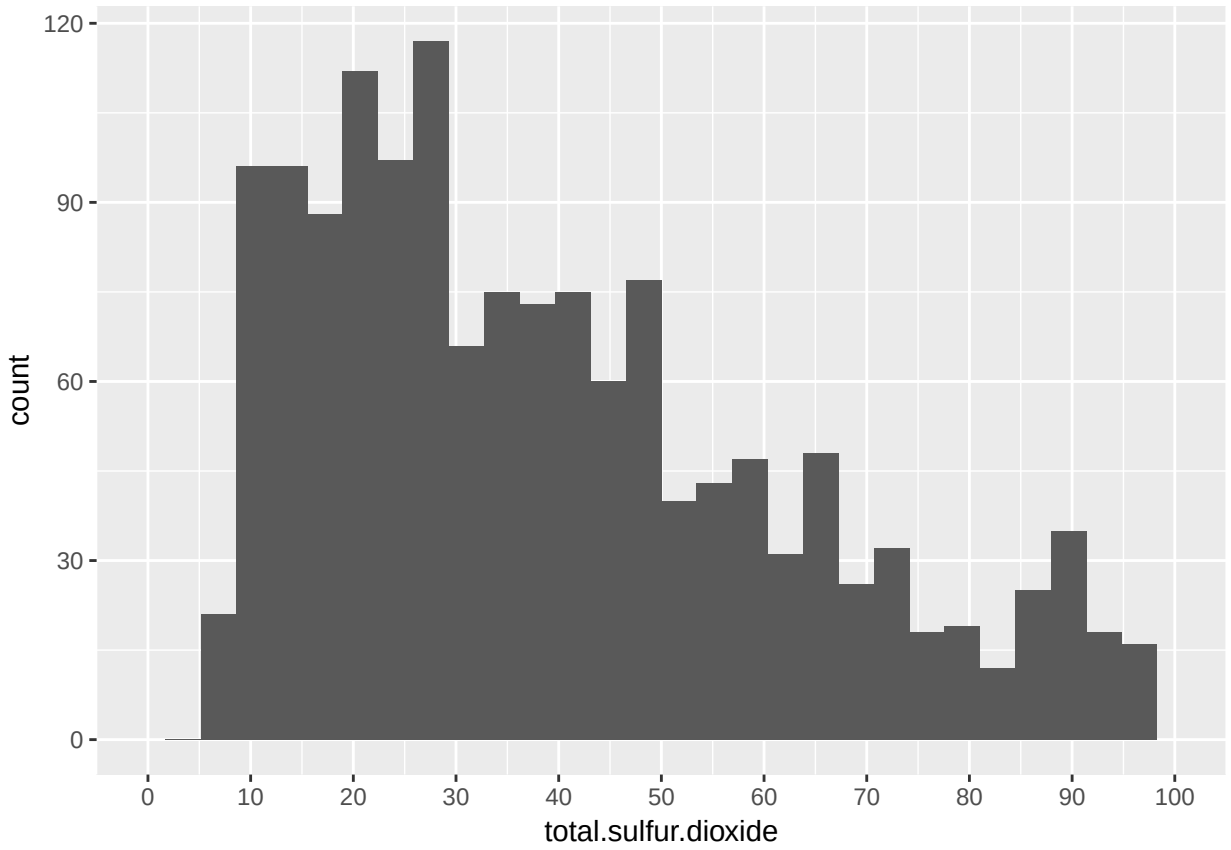



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010
```

The distribution of pH is similar to volatile acidity and fixed acidity, except it is more symmetrical. Citric acid's distribution, however, is different because it is more concentrated on the lower section while pH is more symmetric, being most concentrated in the center.

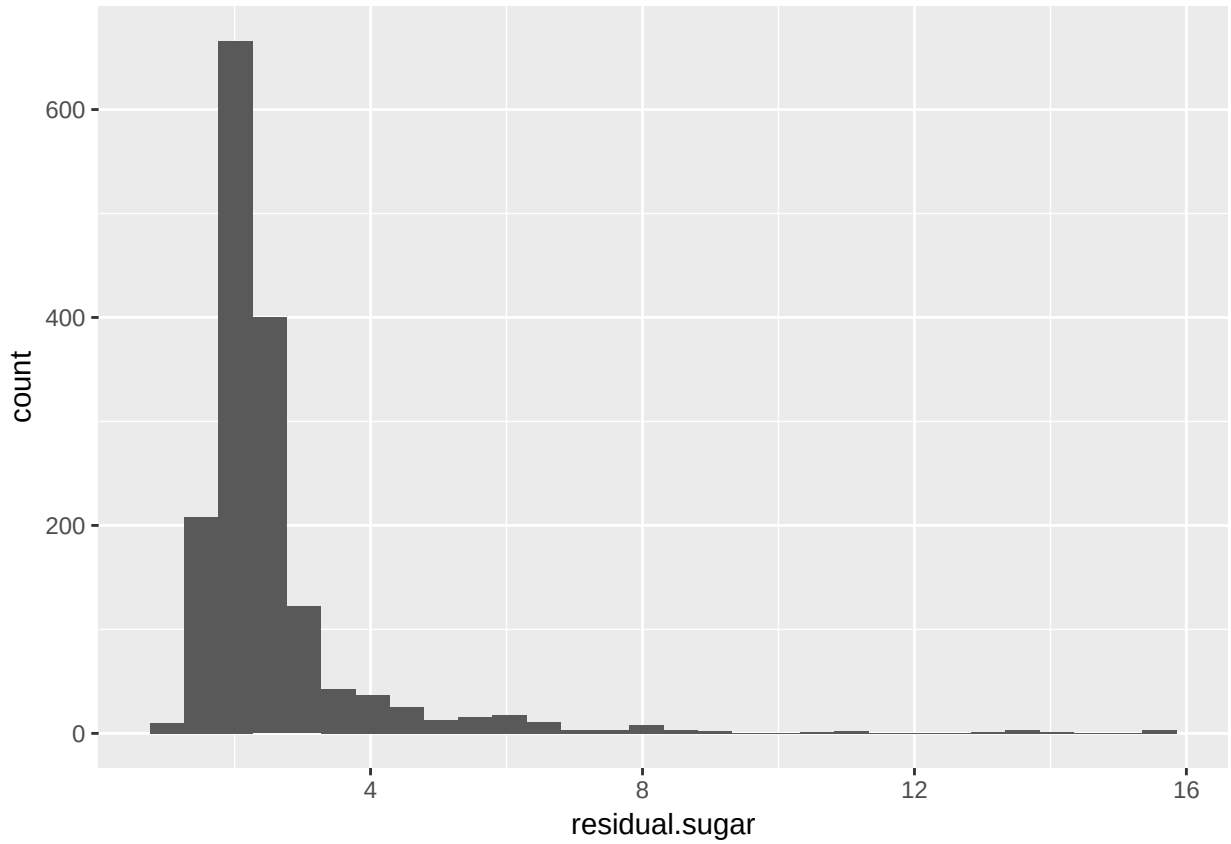

count
200
100
0
0
10
20
30
40
50
60
70
80
90
100
free.sulfur.dioxide

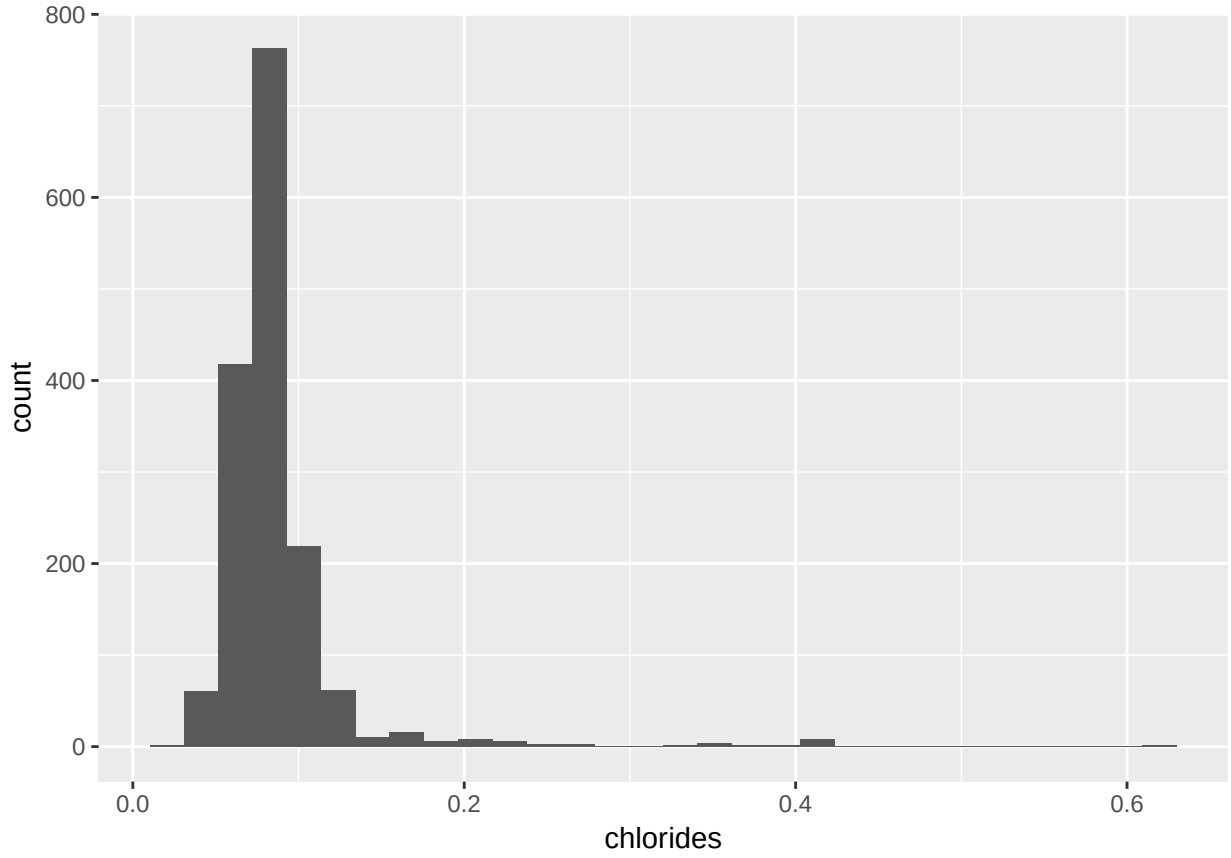

count
200
150
100
50
0
0
10
20
30
40
50
60
70
80
90
100
bound.sulfur.dioxide




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00
```

The total sulfur dioxide seems to be distributed in the lower section primarily also. The highest counts of total sulfur dioxide occur from 10 - 30, and they keep getting lower in frequency as the level of total sulfur dioxide increases. In general, there's more bound sulfur dioxide than free sulfur dioxide.


600
400
200
0
count
4
8
12
16
residual.sugar


800
600
400
200
0
count
0.0
0.2
0.4
0.6
chlorides

Both these histograms have a similar distribution. That's very peculiar, unless there is a relationship between them. Let's see if they are the same samples or different ones, so we know whether to expect a relationship between the 2 variables or not. It is also worth noting that they are both concentrated on low side with some samples having medium, high, and very high levels (of residual sugar and chlorides respectively). Let's look more at the high-value samples.

```
##         X fixed.acidity volatile.acidity citric.acid residual.sugar
## 34     34           6.9            0.605        0.12           10.7
## 325   325          10.0            0.490        0.20           11.0
## 326   326          10.0            0.490        0.20           11.0
## 481   481          10.6            0.280        0.39           15.5
## 495   495           6.5            0.390        0.23            8.3
## 650   650           6.7            0.420        0.27            8.6
## 912   912           9.1            0.280        0.46            9.0
## 918   918           6.8            0.410        0.31            8.8
## 924   924           6.8            0.410        0.31            8.8
## 1044 1044           9.5            0.390        0.41            8.9
## 1072 1072           7.5            0.770        0.20            8.1
## 1075 1075           7.5            0.770        0.20            8.1
## 1080 1080           7.9            0.300        0.68            8.3
## 1082 1082           7.9            0.300        0.68            8.3
## 1236 1236           6.0            0.330        0.32           12.9
## 1245 1245           5.9            0.290        0.25           13.4
## 1435 1435          10.2            0.540        0.37           15.4
## 1436 1436          10.2            0.540        0.37           15.4
## 1475 1475           9.9            0.500        0.50           13.8
## 1477 1477           9.9            0.500        0.50           13.8
## 1575 1575           5.6            0.310        0.78           13.9
##      chlorides free.sulfur.dioxide total.sulfur.dioxide density   pH
## 34       0.073                40.0                   83 0.99930 3.45
## 325      0.071                13.0                   50 1.00150 3.16
## 326      0.071                13.0                   50 1.00150 3.16
## 481      0.069                 6.0                   23 1.00260 3.12
## 495      0.051                28.0                   91 0.99520 3.44
## 650      0.068                24.0                  148 0.99480 3.16
## 912      0.114                 3.0                    9 0.99901 3.18
## 918      0.084                26.0                   45 0.99824 3.38
## 924      0.084                26.0                   45 0.99824 3.38
## 1044     0.069                18.0                   39 0.99859 3.29
## 1072     0.098                30.0                   92 0.99892 3.20
## 1075     0.098                30.0                   92 0.99892 3.20
## 1080     0.050                37.5                  278 0.99316 3.01
## 1082     0.050                37.5                  289 0.99316 3.01
## 1236     0.054                 6.0                  113 0.99572 3.30
## 1245     0.067                72.0                  160 0.99721 3.33
## 1435     0.214                55.0                   95 1.00369 3.18
## 1436     0.214                55.0                   95 1.00369 3.18
## 1475     0.205                48.0                   82 1.00242 3.16
## 1477     0.205                48.0                   82 1.00242 3.16
## 1575     0.074                23.0                   92 0.99677 3.39
##      sulphates alcohol quality bound.sulfur.dioxide
## 34        0.52     9.4       6                43.0
## 325       0.69     9.2       6                37.0
## 326       0.69     9.2       6                37.0
```

```
## 481       0.66     9.2       5                17.0
## 495       0.55    12.1       6                63.0
## 650       0.57    11.3       6               124.0
## 912       0.60    10.9       6                 6.0
## 918       0.64    10.1       6                19.0
## 924       0.64    10.1       6                19.0
## 1044      0.81    10.9       7                21.0
## 1072      0.58     9.2       5                62.0
## 1075      0.58     9.2       5                62.0
## 1080      0.51    12.3       7               240.5
## 1082      0.51    12.3       7               251.5
## 1236      0.56    11.5       4               107.0
## 1245      0.54    10.3       6                88.0
## 1435      0.77     9.0       6                40.0
## 1436      0.77     9.0       6                40.0
## 1475      0.75     8.8       5                34.0
## 1477      0.75     8.8       5                34.0
## 1575      0.48    10.5       6                69.0
```

```
##         X fixed.acidity volatile.acidity citric.acid residual.sugar
## 18     18           8.1            0.560        0.28            1.7
## 20     20           7.9            0.320        0.51            1.8
## 43     43           7.5            0.490        0.20            2.6
## 82     82           7.8            0.430        0.70            1.9
## 84     84           7.3            0.670        0.26            1.8
## 107   107           7.8            0.410        0.68            1.7
## 152   152           9.2            0.520        1.00            3.4
## 170   170           7.5            0.705        0.24            1.8
## 227   227           8.9            0.590        0.50            2.0
## 259   259           7.7            0.410        0.76            1.8
## 282   282           7.7            0.270        0.68            3.5
## 292   292          11.0            0.200        0.48            2.0
## 452   452           8.4            0.370        0.53            1.8
## 693   693           8.6            0.490        0.51            2.0
## 731   731           9.5            0.550        0.66            2.3
## 755   755           7.8            0.480        0.68            1.7
## 1052 1052           8.5            0.460        0.59            1.4
## 1166 1166           8.5            0.440        0.50            1.9
## 1261 1261           8.6            0.635        0.68            1.8
## 1320 1320           9.1            0.760        0.68            1.7
## 1371 1371           8.7            0.780        0.51            1.7
## 1373 1373           8.7            0.780        0.51            1.7
##      chlorides free.sulfur.dioxide total.sulfur.dioxide density   pH
## 18       0.368                 16                   56 0.99680 3.11
## 20       0.341                 17                   56 0.99690 3.04
## 43       0.332                  8                   14 0.99680 3.21
## 82       0.464                 22                   67 0.99740 3.13
## 84       0.401                 16                   51 0.99690 3.16
## 107      0.467                 18                   69 0.99730 3.08
## 152      0.610                 32                   69 0.99960 2.74
## 170      0.360                 15                   63 0.99640 3.00
## 227      0.337                 27                   81 0.99640 3.04
## 259      0.611                  8                   45 0.99680 3.06
## 282      0.358                  5                   10 0.99720 3.25
```

```
## 292      0.343                6                   18 0.99790 3.30
## 452      0.413                9                   26 0.99790 3.06
## 693      0.422               16                   62 0.99790 3.03
## 731      0.387               12                   37 0.99820 3.17
## 755      0.415               14                   32 0.99656 3.09
## 1052     0.414               16                   45 0.99702 3.03
## 1166     0.369               15                   38 0.99634 3.01
## 1261     0.403               19                   56 0.99632 3.02
## 1320     0.414               18                   64 0.99652 2.90
## 1371     0.415               12                   66 0.99623 3.00
## 1373     0.415               12                   66 0.99623 3.00
##      sulphates alcohol quality bound.sulfur.dioxide
## 18        1.28     9.3       5                   40
## 20        1.08     9.2       6                   39
## 43        0.90    10.5       6                    6
## 82        1.28     9.4       5                   45
## 84        1.14     9.4       5                   35
## 107       1.31     9.3       5                   51
## 152       2.00     9.4       4                   37
## 170       1.59     9.5       5                   48
## 227       1.61     9.5       6                   54
## 259       1.26     9.4       5                   37
## 282       1.08     9.9       7                    5
## 292       0.71    10.5       5                   12
## 452       1.06     9.1       6                   17
## 693       1.17     9.0       5                   46
## 731       0.67     9.6       5                   25
## 755       1.06     9.1       6                   18
## 1052      1.34     9.2       5                   29
## 1166      1.10     9.4       5                   23
## 1261      1.15     9.3       5                   37
## 1320      1.33     9.1       6                   46
## 1371      1.17     9.2       5                   54
## 1373      1.17     9.2       5                   54
```

It appears that they are different samples. So, it is just a coincidence that the distribution of the histograms look similar.

Univariate Analysis

What is the structure of your dataset?

There are 1599 samples with 12 features (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and quality).

Some Observations: - Most wine samples have a quality of 5. - The median alcohol content was 10.2. - The distribution of pH matches the distribution of volatile acidity. - The median total sulfur dioxide level is 38. - Most wine samples have residual sugar less than 8 and chlorides less than 0.3.

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

The main feature of interest is the quality and how other factors relate to it. I think there can be a predictive model built from some combination of other features for determining an optimal wine sample.

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

Other features like acidity, sulfur dioxide, alcohol, and sugar all combine in particular amounts to make the ideal wine sample. My goal is to determine which features are best for predicting the quality of wine samples.

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

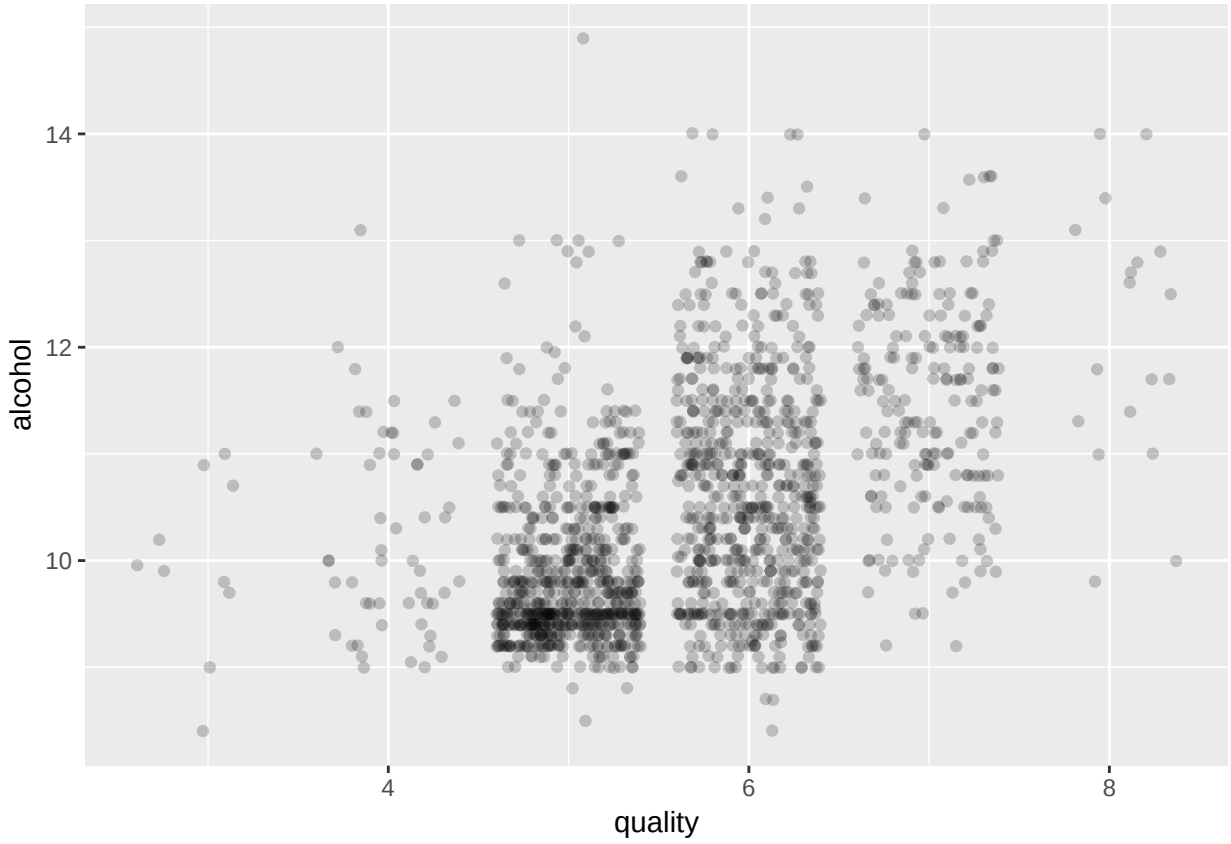
Yes, I created the Bound Sulfur Dioxide variable, based on my research. I did this by subtracting Free Sulfur Dioxide from Total Sulfur Dioxide. I think Bound Sulfur Dioxide will serve as another helpful factor for determining a relationship in the Quality of wine samples.

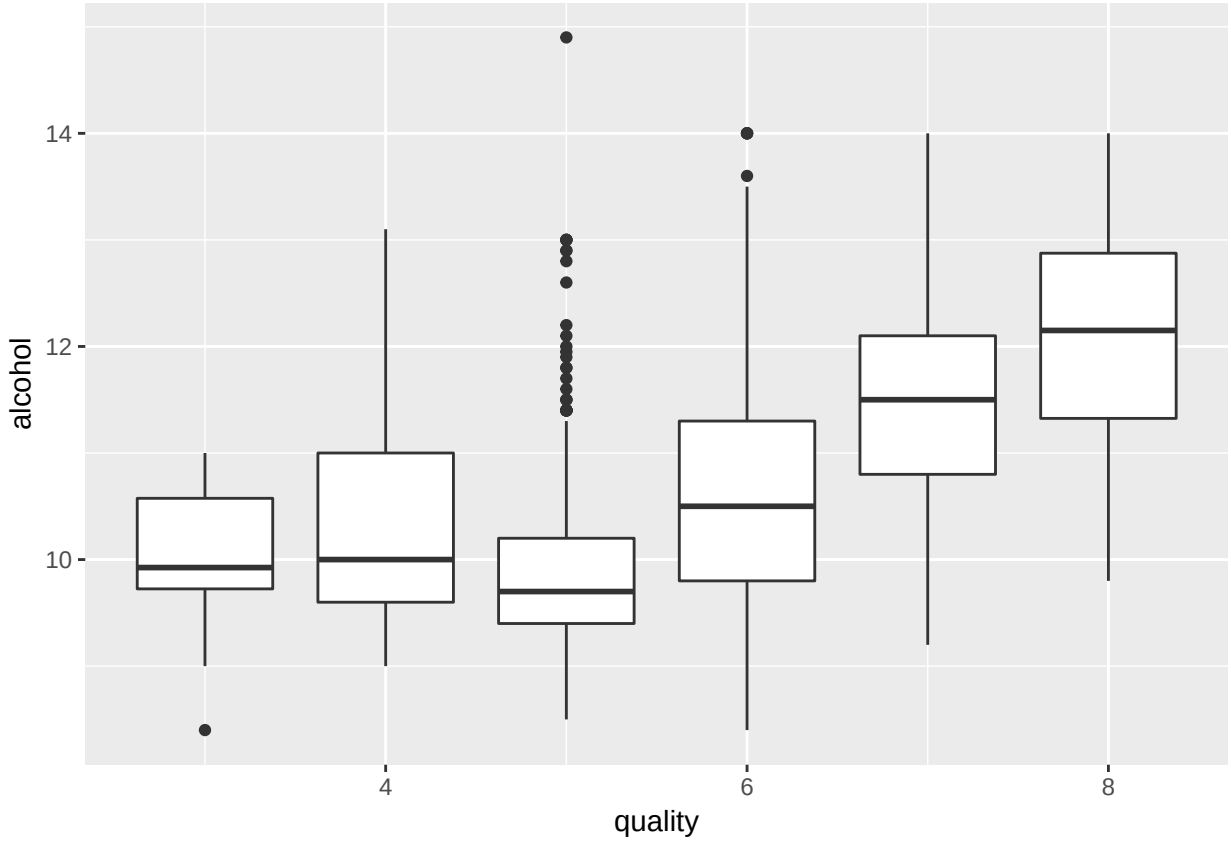
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

Yes, I think it was unusual how the distribution of residual sugar and chlorides was very similar, even though the samples which had the higher values in one had no correlation to the higher values in the other. I think it's a very strange coincidence that this happend in a collection of 1599 samples.

I also think it's interesting how the pH level distribution was nearly symmetrical. This will make it easier to determine the relationship between quality and acidity.

Bivariate Plots Section

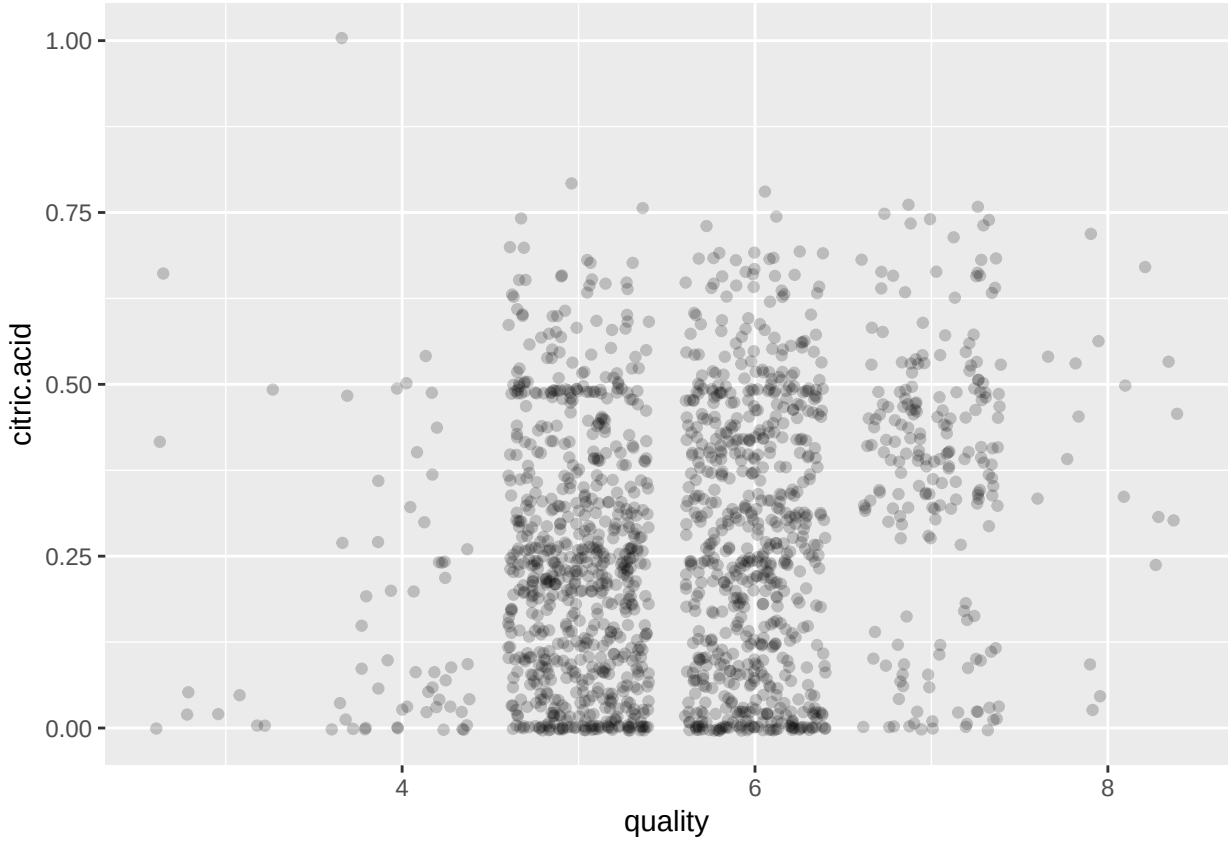


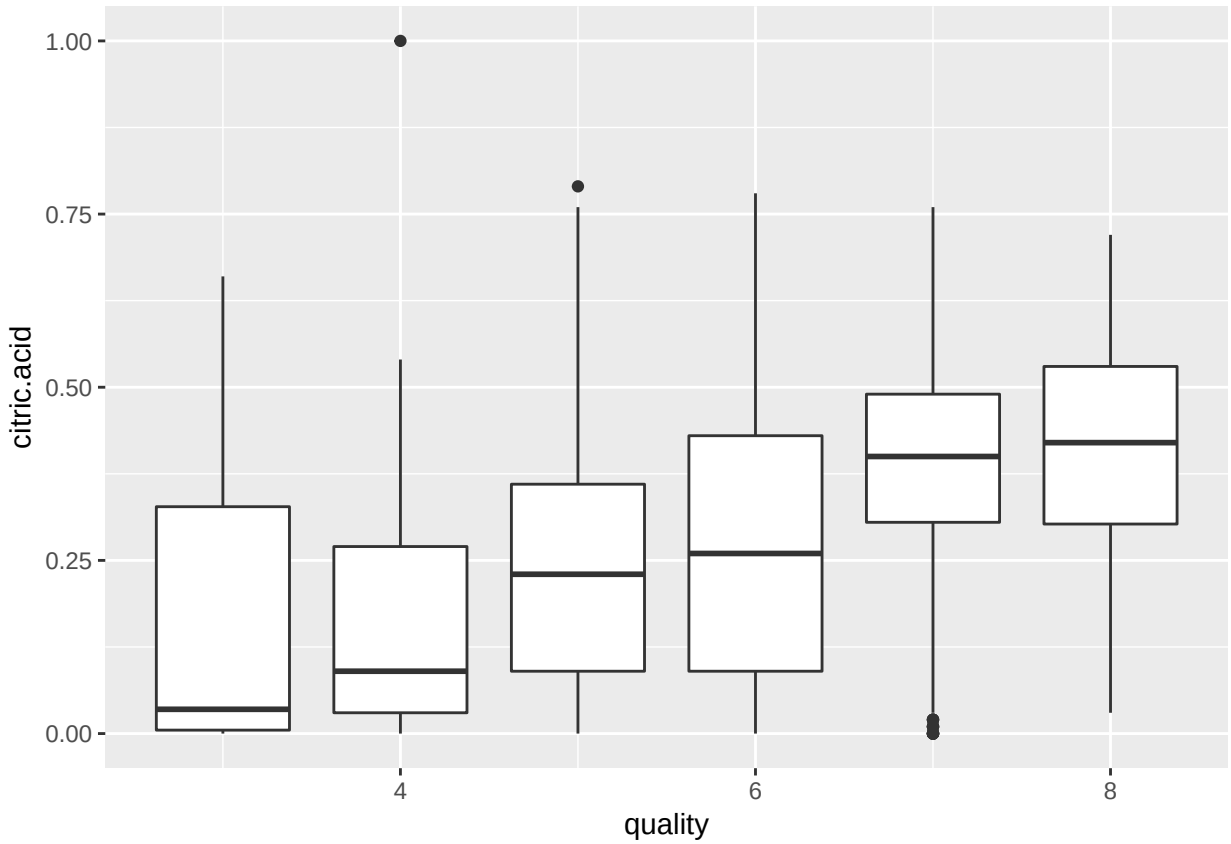


```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```

```
## [1] 0.4761663
```

Alcohol content seems to have a relationship with quality. If we look at the scatter plot, there's a heavy concentration of low alcohol content with a quality of 5. The alcohol content in quality 6 samples is more scattered out on top. Quality 7 samples has alcohol content even more scattered out towards the top. And quality 8 samples has mostly alcohol content greater than 12. These findings are further reinforced by the fact that the boxes seem to be going higher and higher starting from 5 going to 8, with 8 having the highest median. And since the correlation coefficient of alcohol and quality is 0.476, we can know for sure that there is somewhat of a relationship with the two variables.

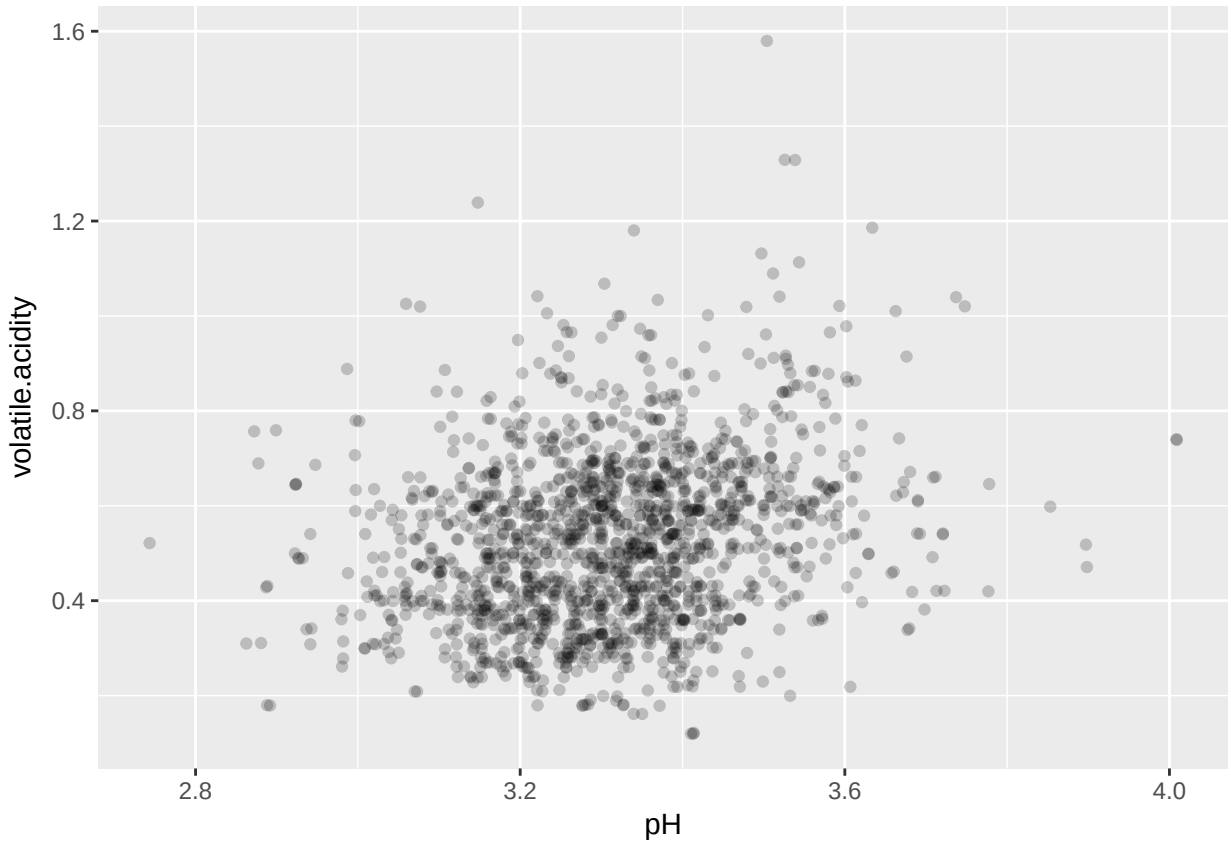


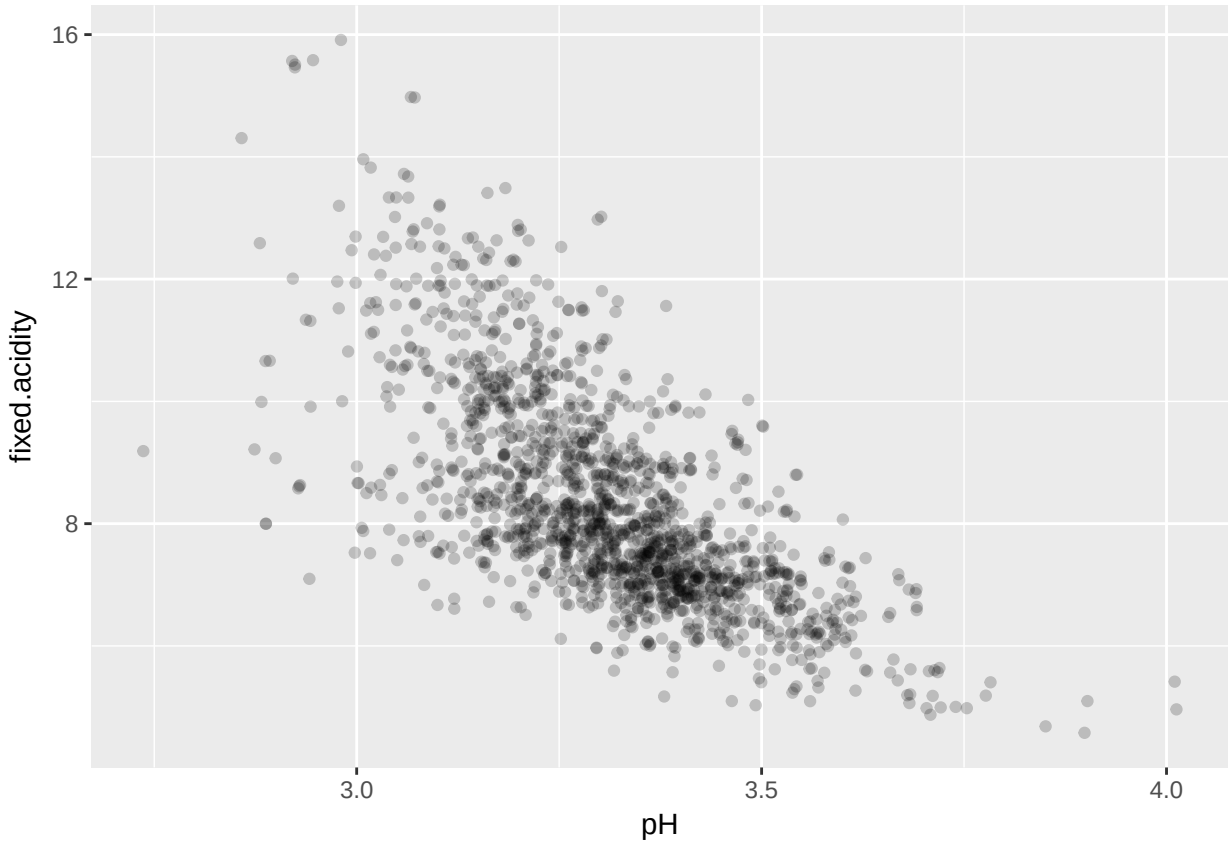



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```

```
## [1] 0.2263725
```

There seems to be less correlation with citric acid and quality than with alcohol and quality. That's because even in the higher qualitie samples like 7 or 8, there are sizable samples that have small amounts of citric acid, according to the scatter plot. However, according to the box plot, the median of the citric acid is increasing as the quality increases, even if by a little amount. This shows that there is a very little correlation, also confirmed by the correlation coefficient which is 0.22.

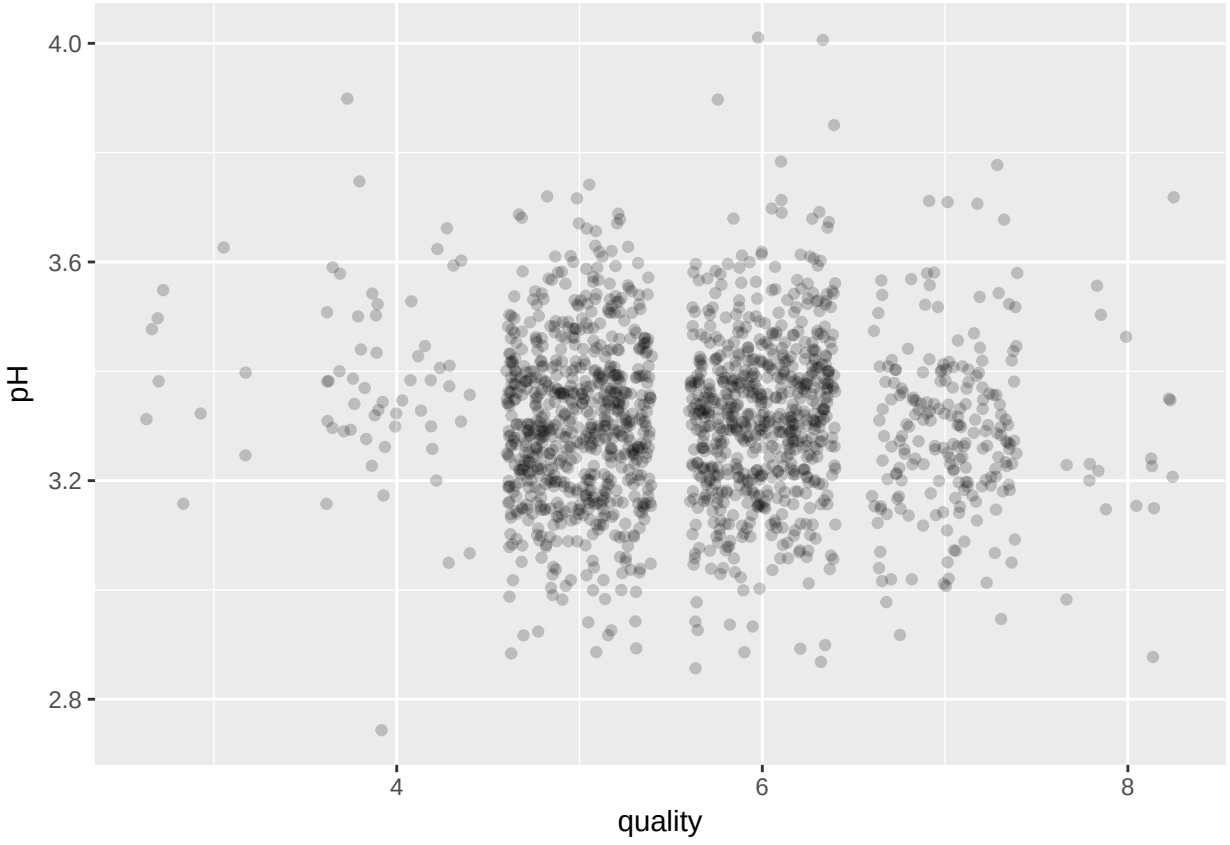

volatile.acidity
1.6
1.2
0.8
0.4
2.8
3.2
3.6
4.0
pH

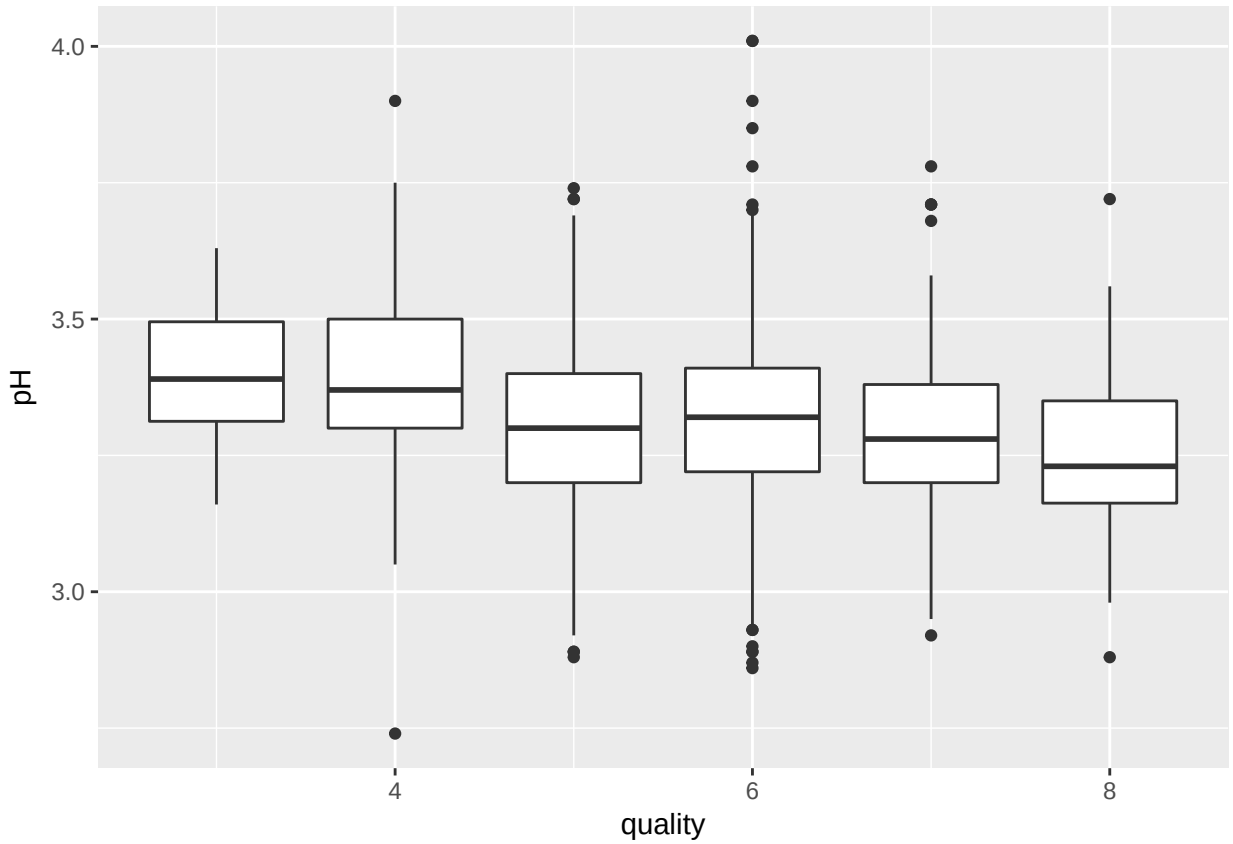

fixed.acidity
16
12
8
3.0
3.5
4.0
pH

```
## [1] 0.2349373
```

```
## [1] -0.6829782
```

pH correlates more to fixed acidity than to volatile acidity. According to the scatterplot for fixed acidity, there is a clear negative correlation - as pH goes down, the fixed acidity level also goes down. The correlation coefficient of fixed acidity is -0.683. I think it's interesting that fixed acidity shows little correlation with pH and if it does, it is that higher pH increases volatile acidity, as shown by the medium to high amounts of volatile acidity in samples with pH greater than 3.4.

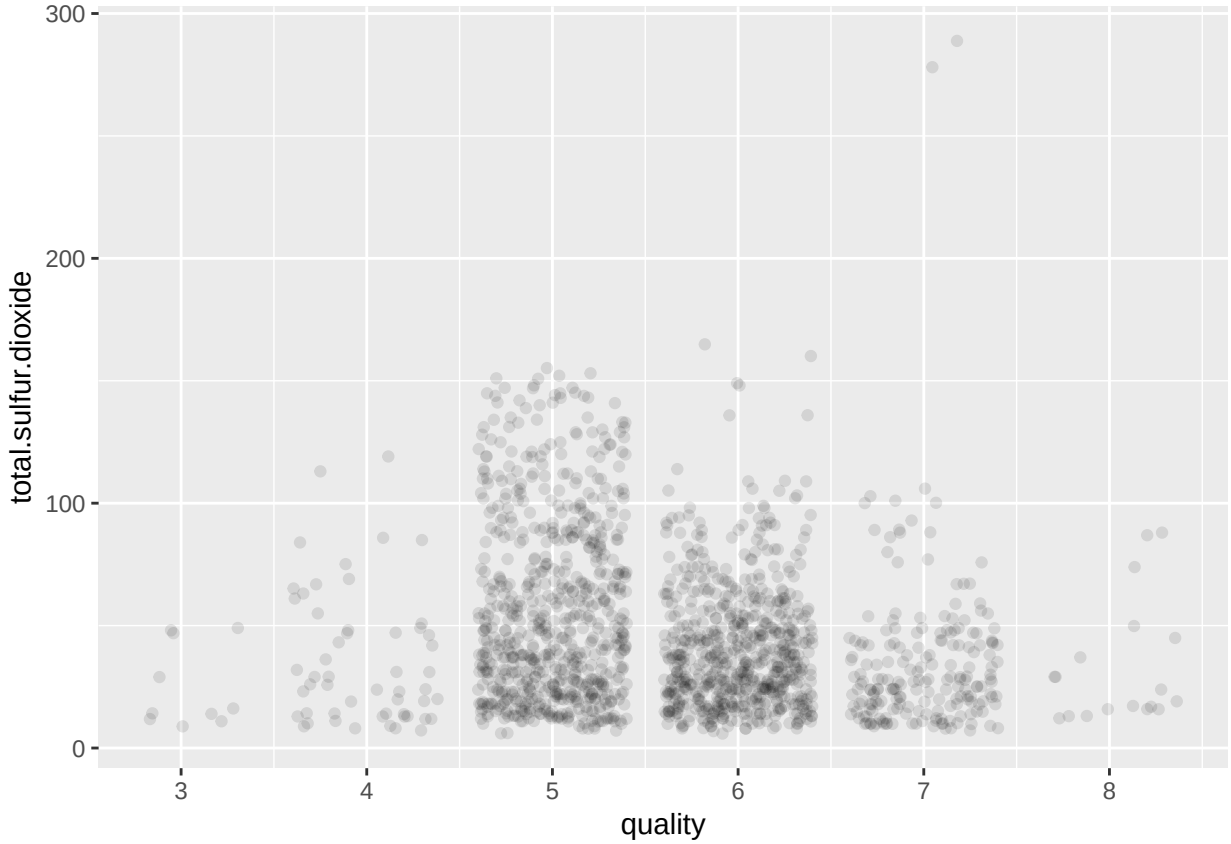


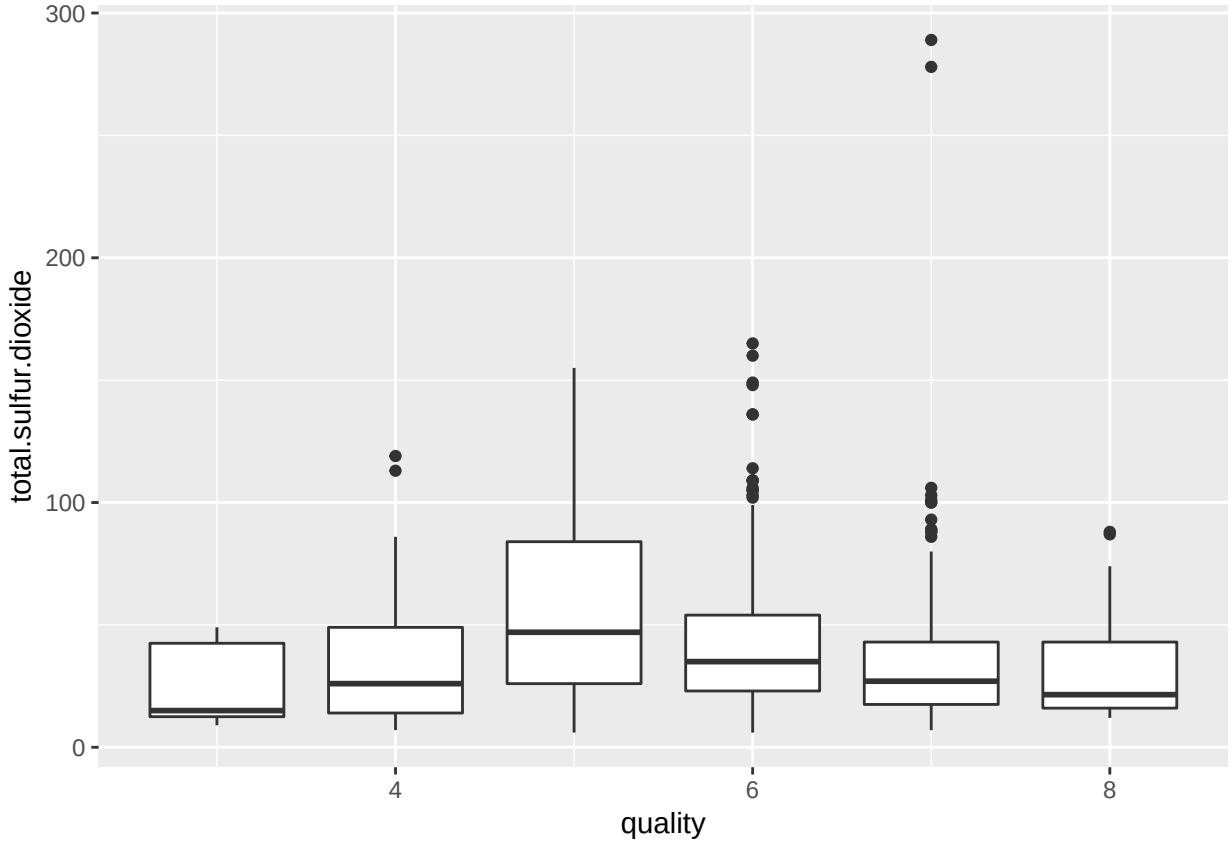



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010
```

```
## [1] -0.05773139
```

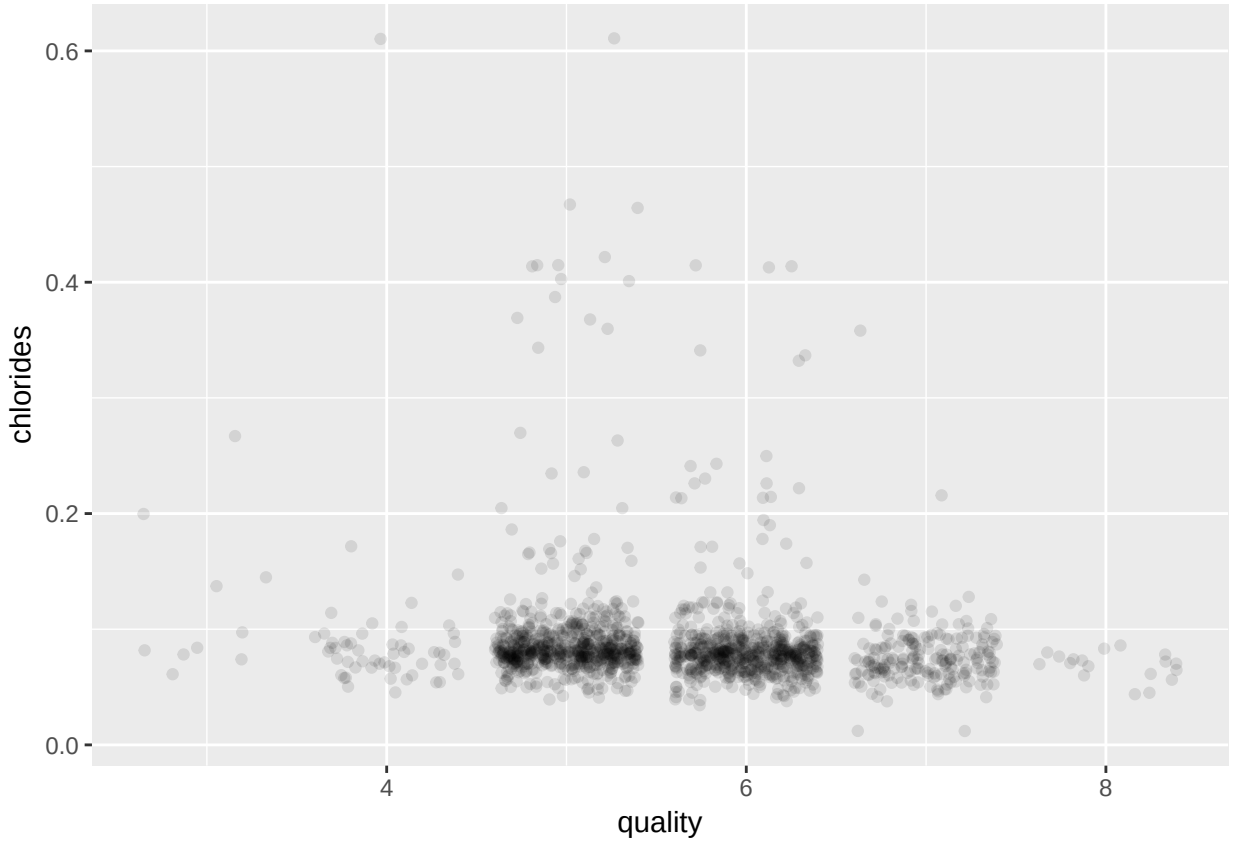
The points seem to be evenly dispersed vertically in each of the quality levels for the scatterplot. This indicates a weak correlation between pH and quality. This is confirmed by the correlation coefficient which is just -0.058. Furthermore, the medians are all very similar to each other across quality.

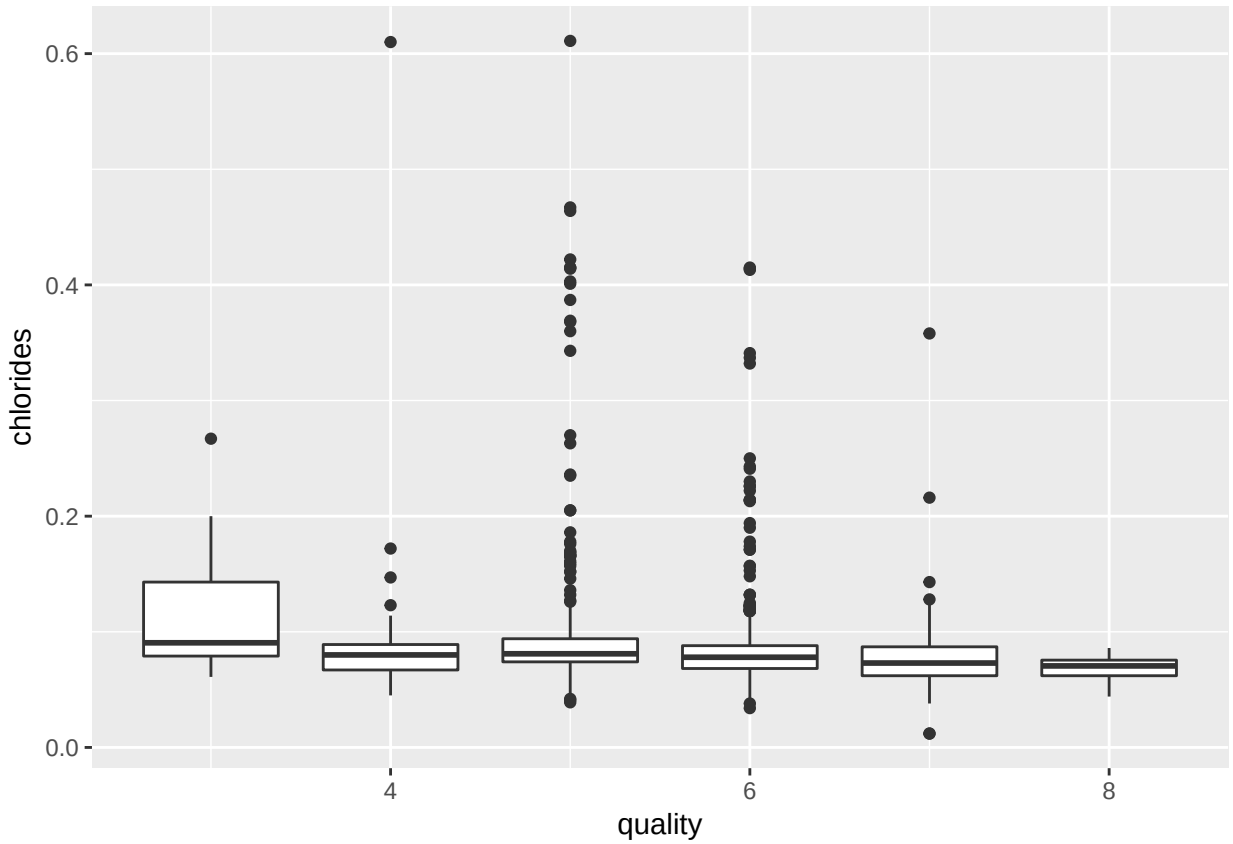

300
200
100
0
total.sulfur.dioxide
3
4
5
6
7
8
quality


300
200
100
0
total.sulfur.dioxide
4
6
8
quality

```
## [1] -0.1851003
```

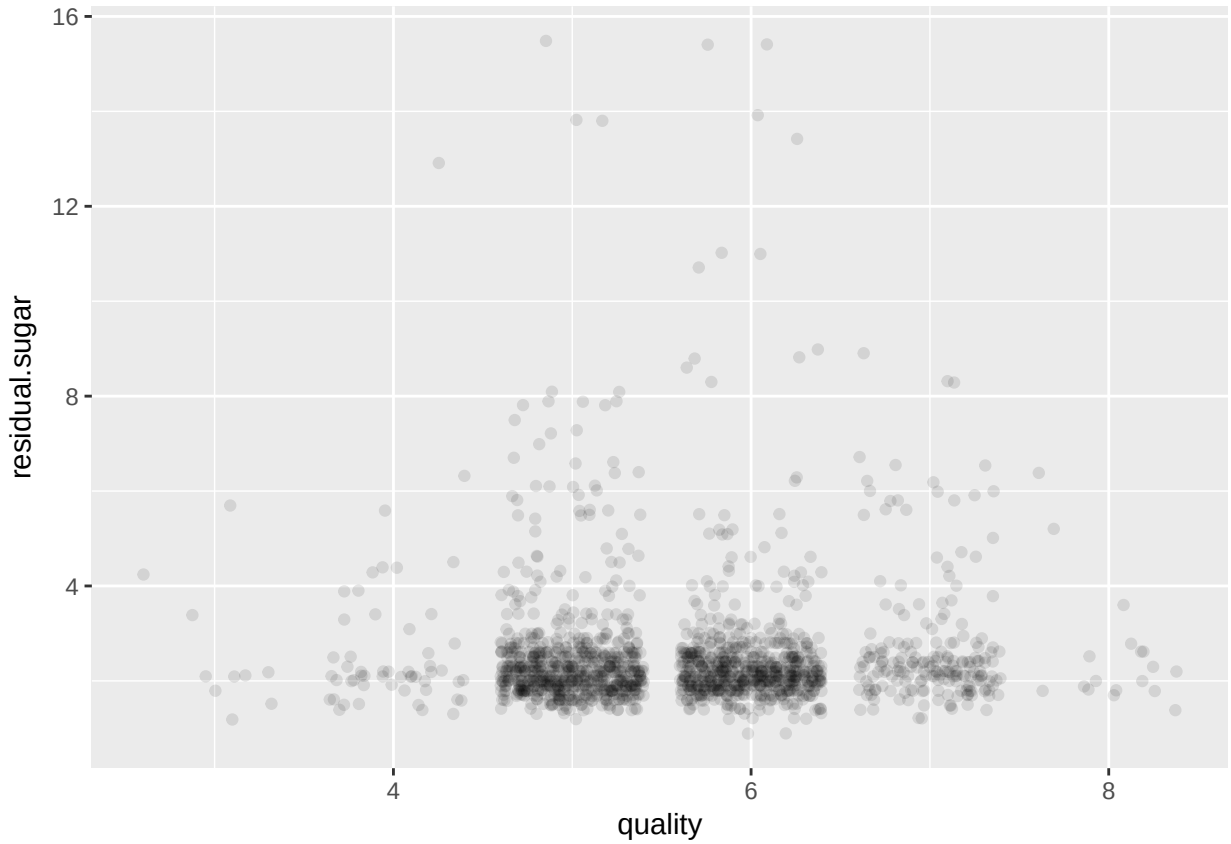
According to the scatter plot, the "bars" made by Qualities of 5 and 6 are the tallest. They start decreasing in height along the ends. Perhaps we can infer that to make an average (Quality 5 or Quality 6) wine, you can put more Sulfur Dioxide in it. The Quality 8 wine samples have very low total sulfur dioxide amounts, as confirmed by the negative correlation coefficient (-0.185).

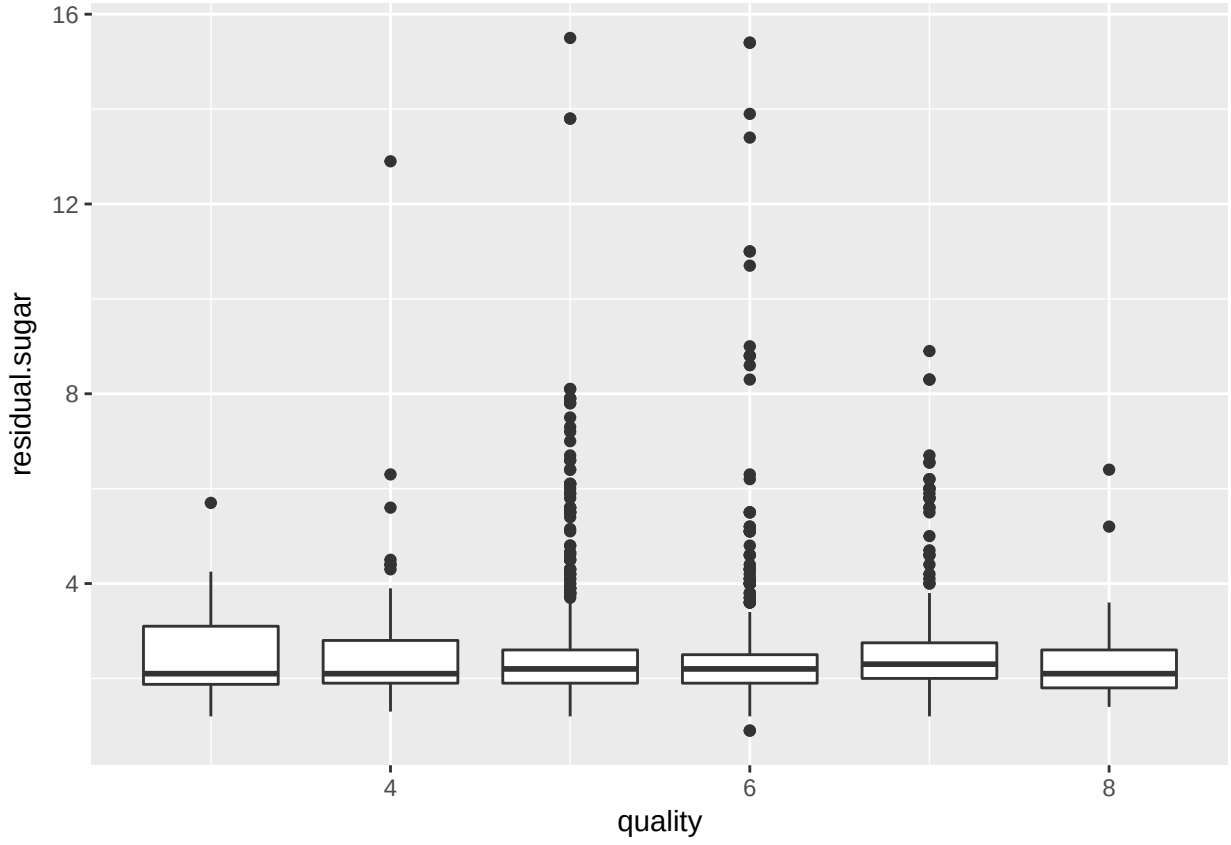





```
## [1] -0.1289066
```

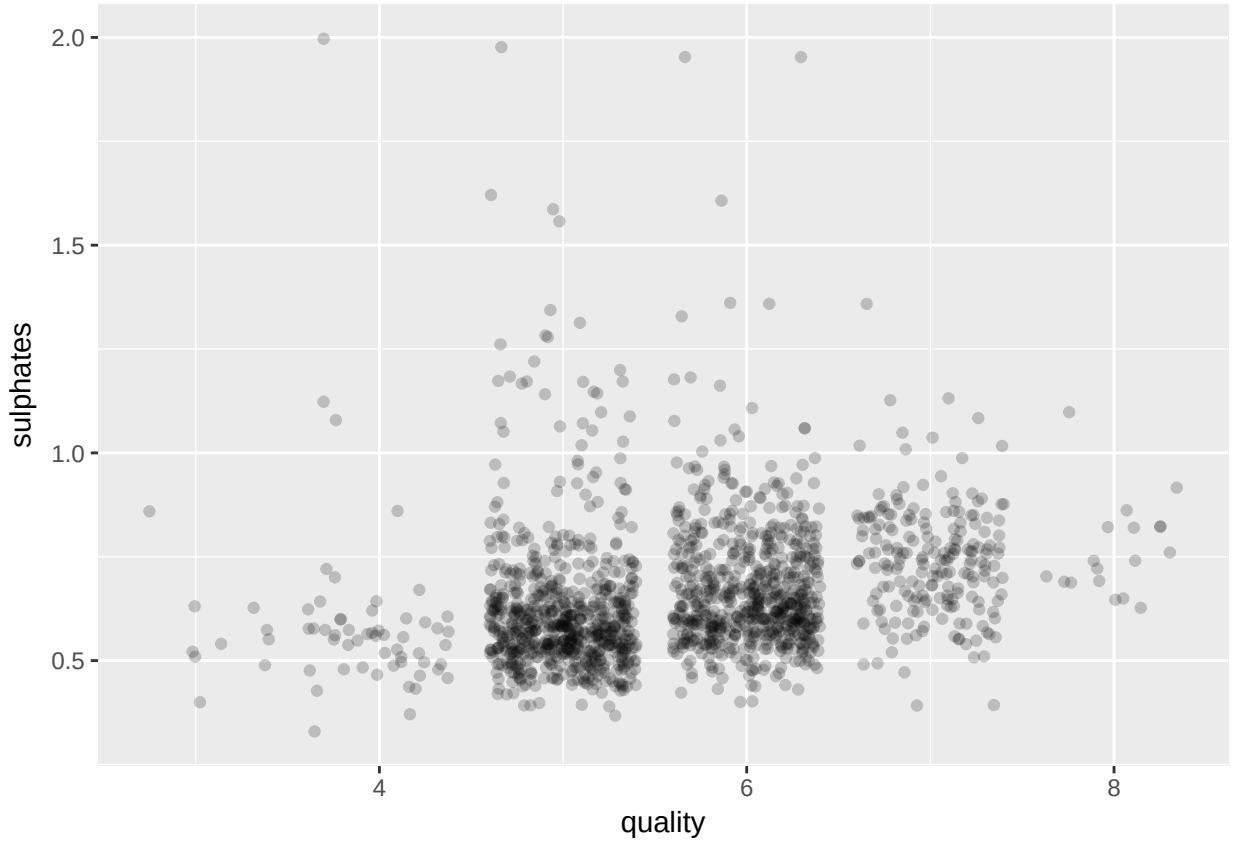
According to the scatter plot, as Quality improves, there is very little difference in the general amount of chloride. The boxplot reinforces this because the medians are all very close to each other with the amount of chloride.

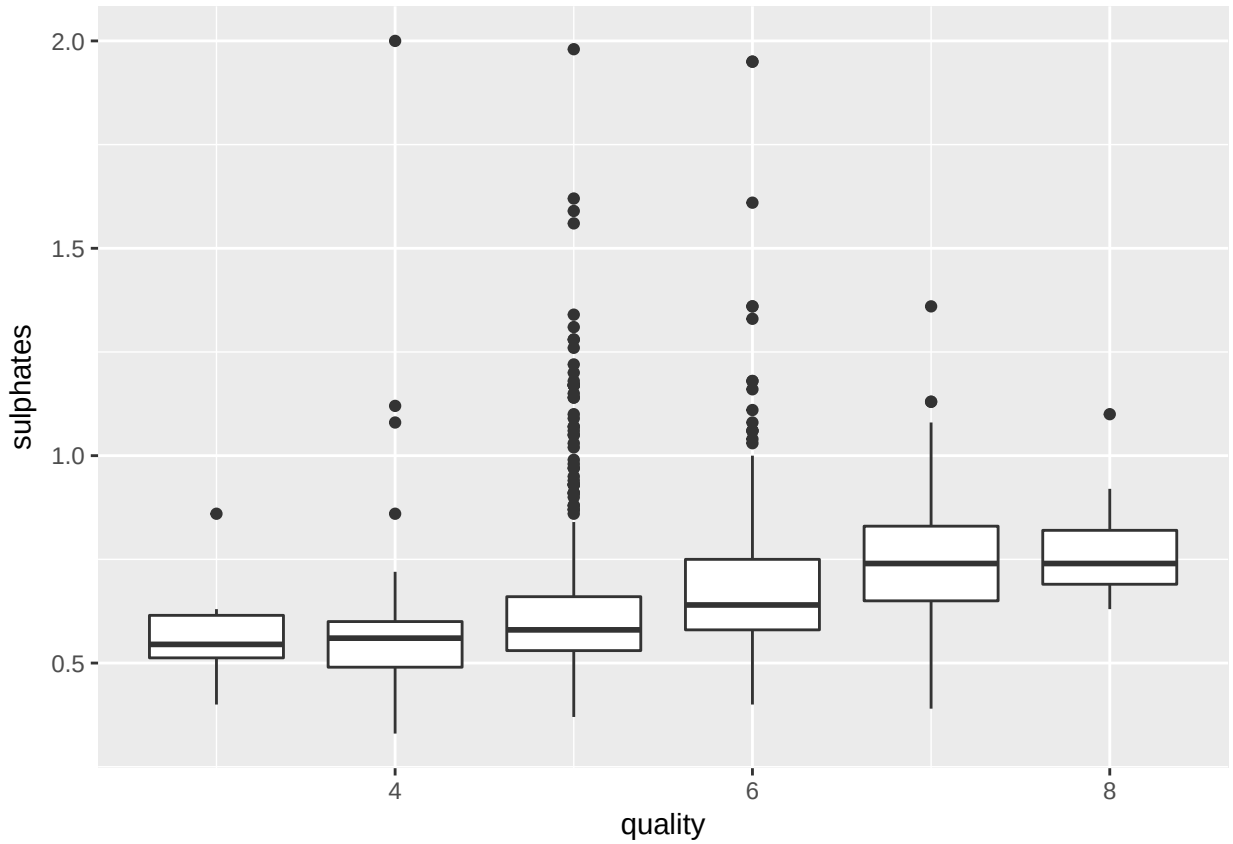

16
12
8
4
residual.sugar
4
6
8
quality


16
12
8
4
residual.sugar
4
6
8
quality

```
## [1] 0.01373164
```

Similar to chloride, there seems to be little correlation of residual sugar and quality. The points are all dispersed similarly in each quality level, showing no cause and effect relationship between residual sugar and quality (Since there are more samples with 5 and 6, it is only heavier in the bottom for those qualities.) In the box plots, the medians are all similar and so are the IQRs. Furthermore, the correlation coefficient is very close to 0.

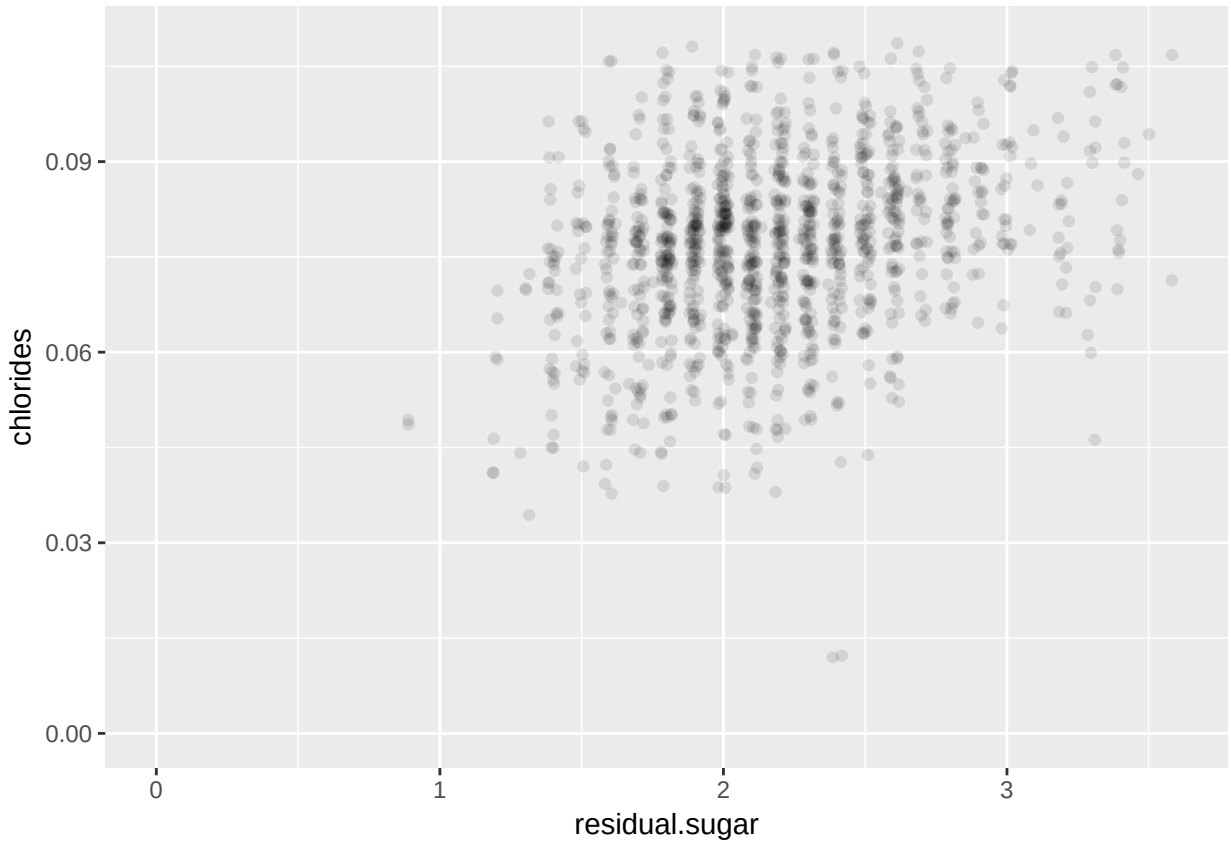





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```

```
## [1] 0.2513971
```

There's a slight positive correlation with sulphates and quality. As the quality is increasing, the points are going upward. Furthermore, the boxes in the boxplot are also moving upward as quality increases. And a correlation coefficient of 0.25 is solid enough to indicate some positive relationship.




```
##
##  Pearson's product-moment correlation
##
## data:  redwine$residual.sugar and redwine$chlorides
## t = 2.2257, df = 1597, p-value = 0.02617
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.006606405 0.104346223
## sample estimates:
##        cor
## 0.05560954
```

Lastly, let's look at the relationship between chlorides and residual sugar as we mentioned in the first section that their distributions look similar. In the scatterlot, clearly, there is little correlation between the amount of residual sugar and chlorides since the points are all scattered. Furthermore, the correlation coefficient is just 0.05.

Bivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

I think the relationship between alcohol and quality, sulphates and quality, and citric acid and quality were the most interesting. These were the top 3 strongest predictors of quality, with alcohol being on top by a clear margin. I think if we were to make a model, alcohol would be the strongest variable in that model

followed by sulphates then citric acid. I was expecting pH to have a stronger correlation with the quality of wine, but it had very little correlation.

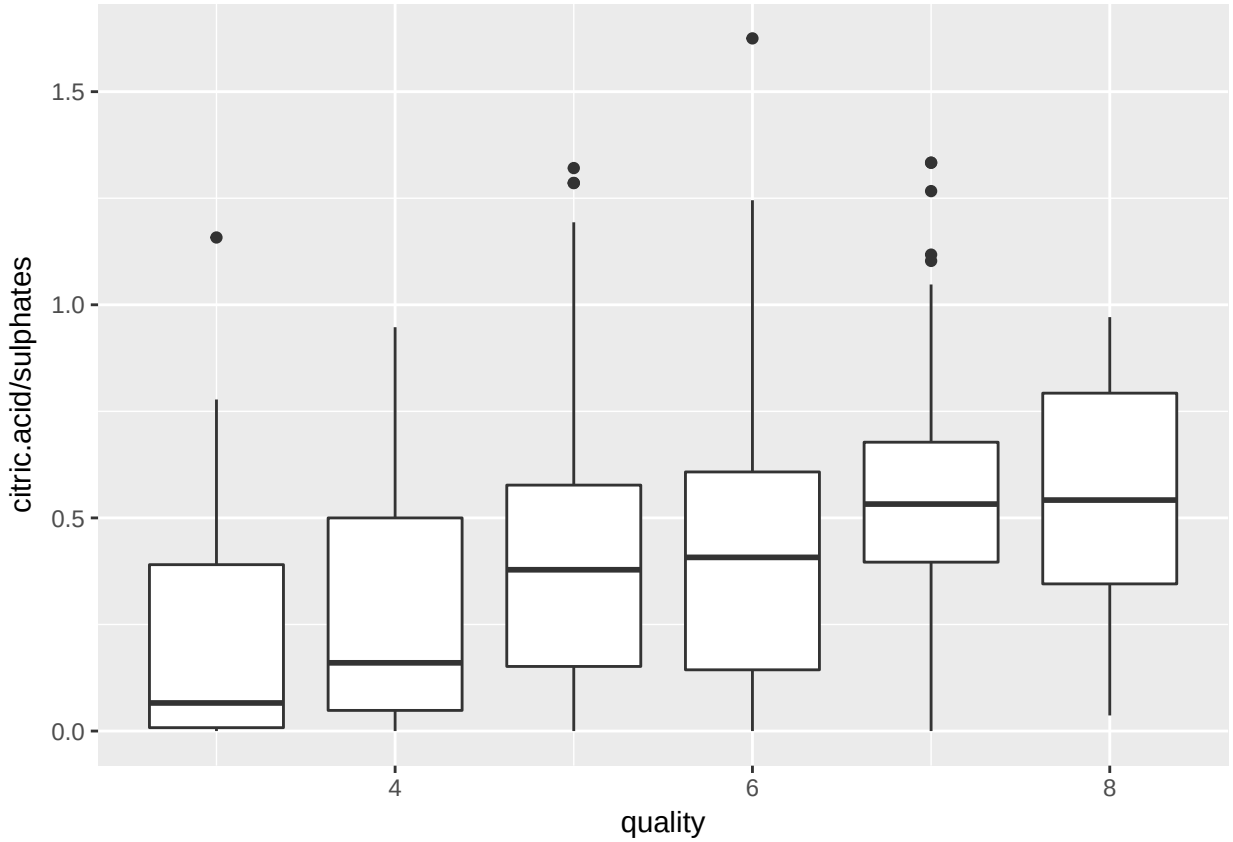
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

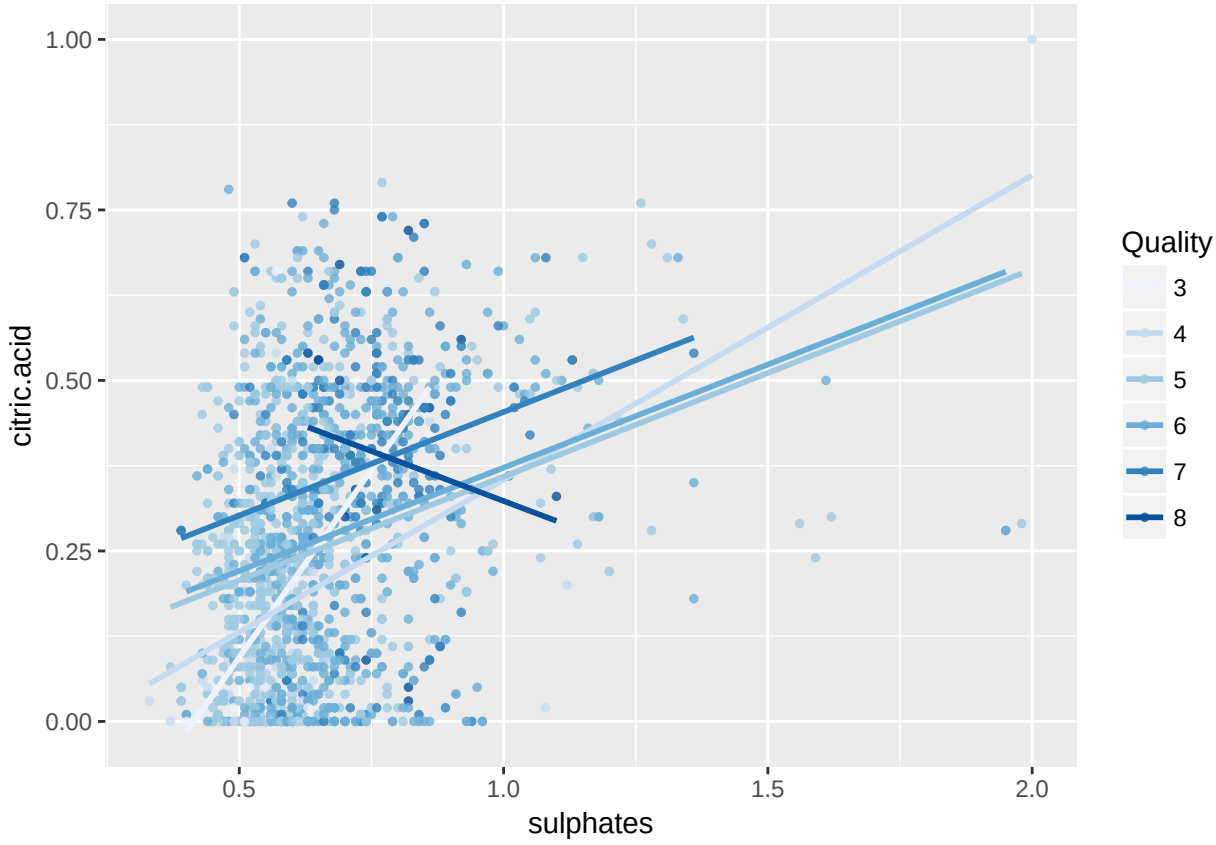
The relationhip between pH and fixed acidity was very strong, which was expected. However, I was expecting the relationship between pH and volatile acidity to be stronger and negative, similar to fixed acidity and pH.

What was the strongest relationship you found?

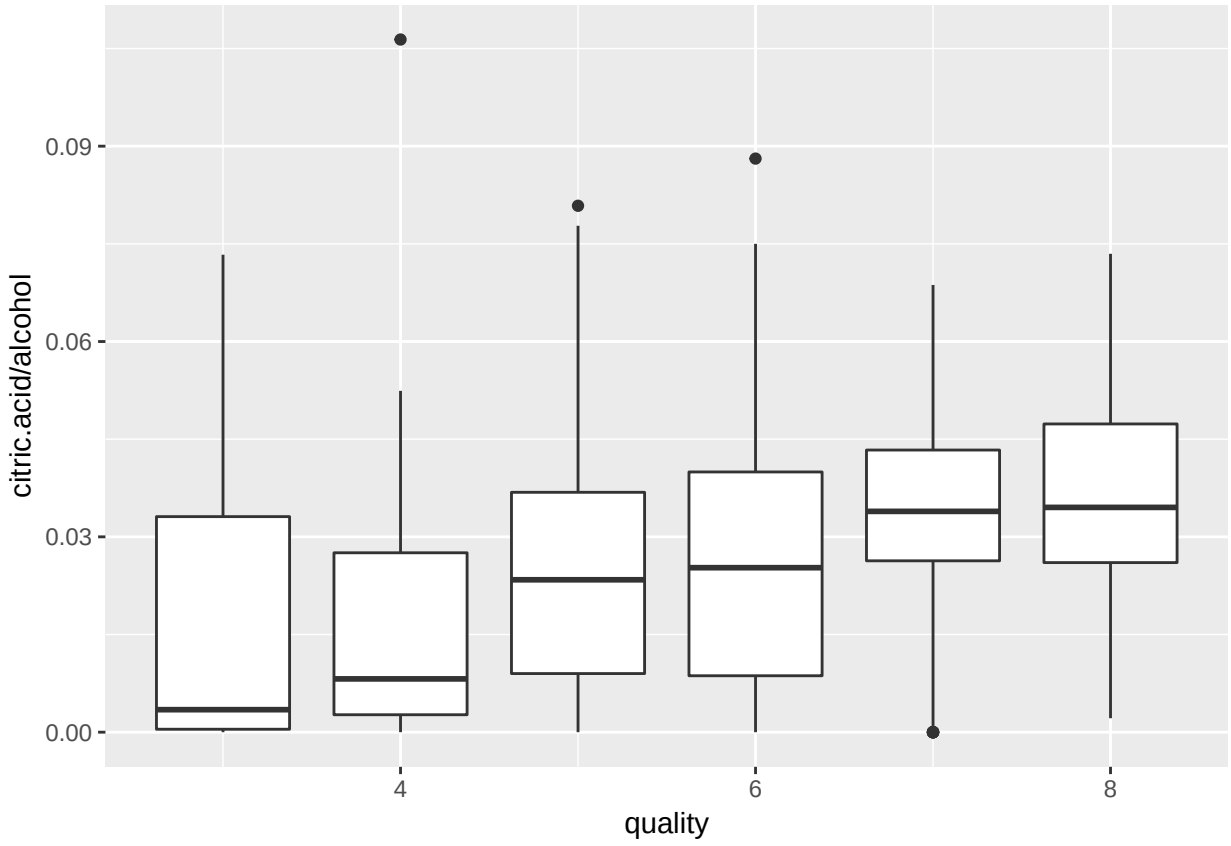
The relationship of pH and fixed acidity was the strongest. However, if we're talking about the Feature of interest, Quality, then Quality and Alcohol was the strongest.

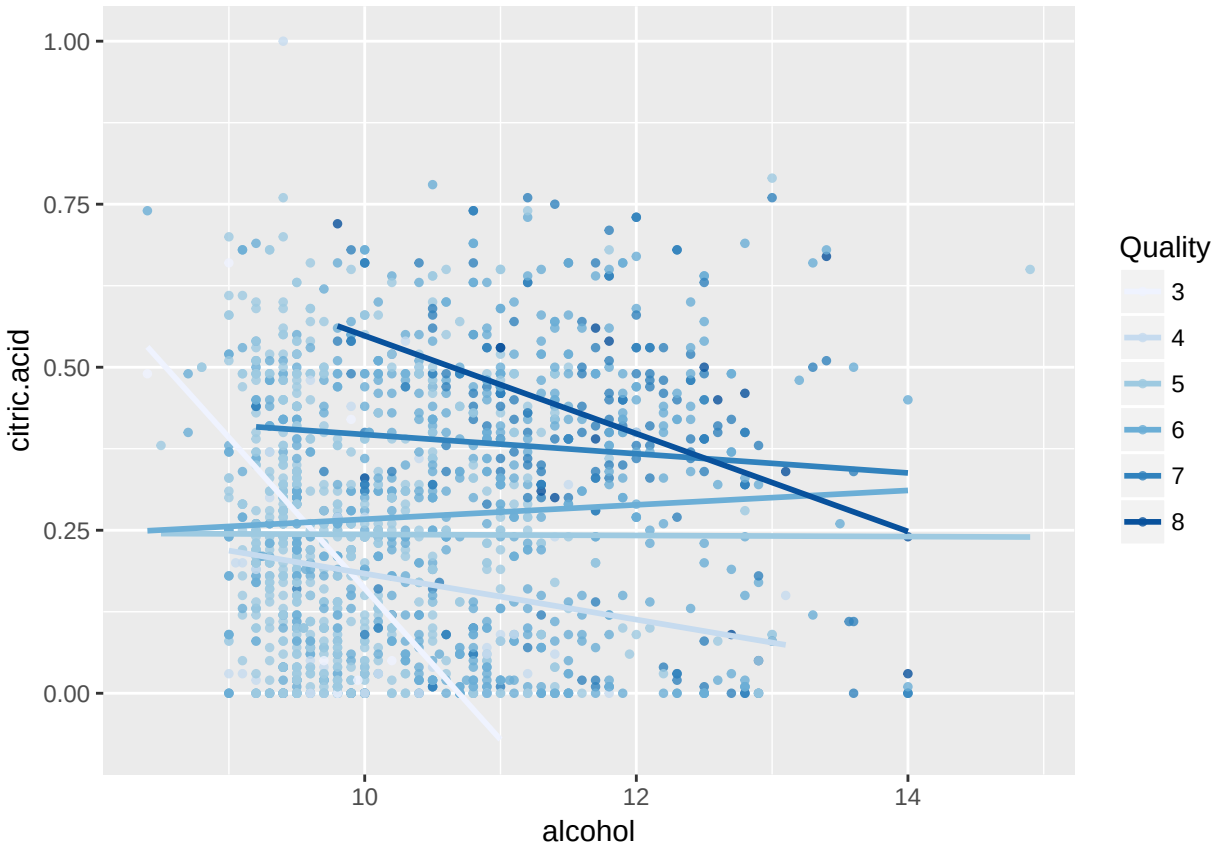
Multivariate Plots Section



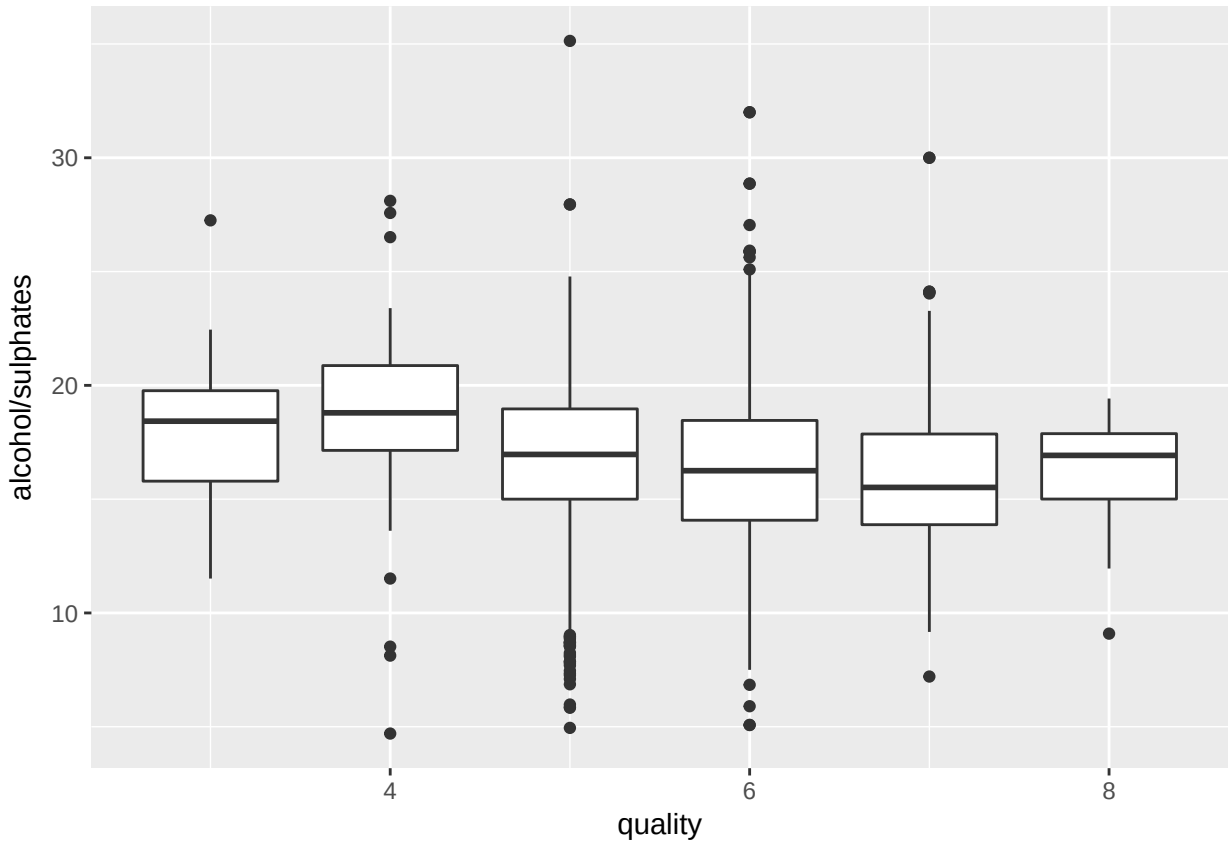


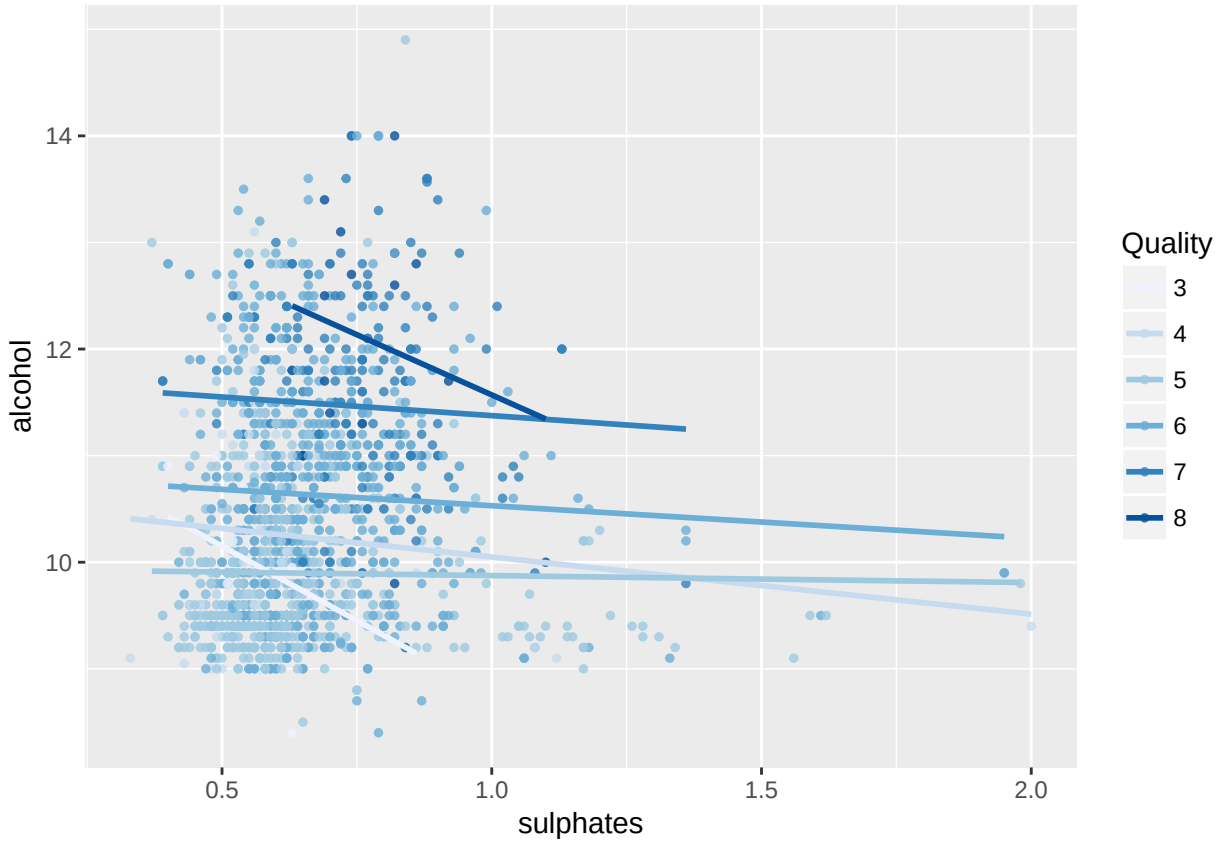
The scatterplot of sulphates and citric acid shows us that too much or too little sulphates give a bad effect to the quality. There's also a sweet spot for the amount of citric acid that should be there. I think the ideal amount of sulphates and citric acid should be somewhere near the light blue patch with sulphates being 0.75 and citric acid content being 0.50. I think if a wine sample has its sulphates and citric acid level in this area or even with citric acid values as low as 0.25 (but not lower than that), that would be ideal. According to the box plot, as the ratio of citric acid/sulphates gets higher, the quality increases. In other words, if the citric.acid to sulphates ratio is 0.5, there's a good chance of the wine sample being of 7 or 8 quality.





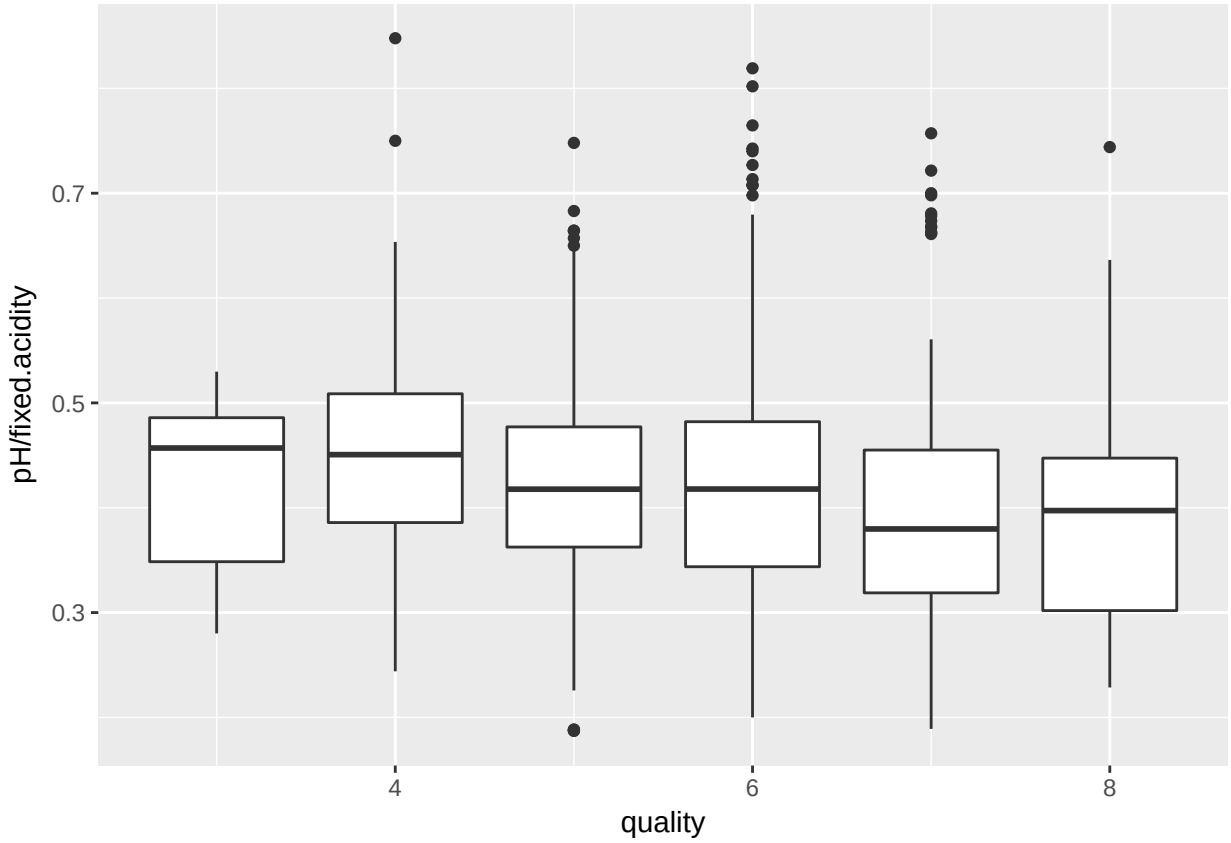
This scatterplot is interesting because it shows that extremes of either alcohol or citric acid without enough of the other will result in bad quality. This is because most points either too down or too to the left on the graph are darker colors, indicating a bad quality. The top right portion of points have a cluster of points that are light colored, indicating a good quality. The boxplot confirms this by indicating that when the ratio of citric acid to alcohol is higher (just below 0.04) it leads to greater quality then when the ratio is closer to 0.

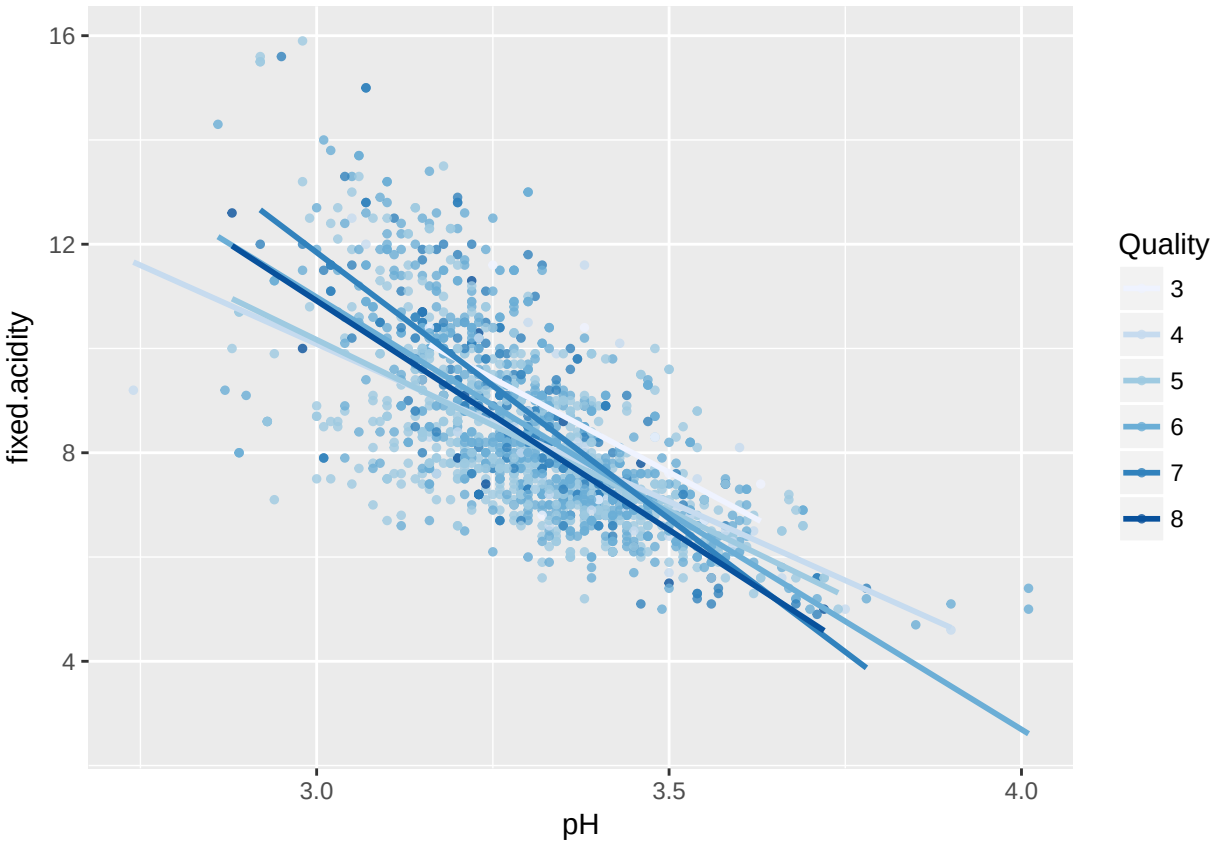




Again, this scatterplot is similar to the previous one in that when there is too much of either substance

(alcohol or sulphates) without the other, it results in lower quality. Most of the lighter colored dots are in the top right region again. It appears that the ideal sulphate level is around 0.75-1 and the alcohol level is around 12-13. The boxplot makes it seem as though the ratio does not make a big difference in quality. Maybe because dividing the alcohol content (a big number) by the sulphates content (very small number) makes the ratios end up all similar to each other, so it's hard to look for a trend/relationship.





According to this scatterplot, the lighter points are more concentrated in the middle than the ends, where pH is about 3.3 and fixed.acidity is near 8. This shows that a medium level of fixed acidity or pH is ideal. The box plot however shows that there is a slight decreasing of the ph to fixed acidity ratio as the quality increases.

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?

There seemed to be a relationship between citric acids, sulphates, and quality, as seen by the first plot. This relationship showed a sweetspot regarding the amount of citric acid and the amount of sulphates there should be in order to maximize the quality. For sulphates, too low or too much had a clear bad affect on quality, indicating darker dots. The ideal amount seemed to be 0.75 and for citric acid it seemed to be 0.5. I thought the patch of lighter colored blue dots in the top right area is really where quality was a little better. I noticed that extremes (low or high) of most chemicals caused an adverse effect on the quality. I thought the complentary relationship between citric acid and alcohol was interesting because it shows that too much of either of them without the other can have an adverse effect but if you have a good dose of both of them together then it leaves a good impact on quality.

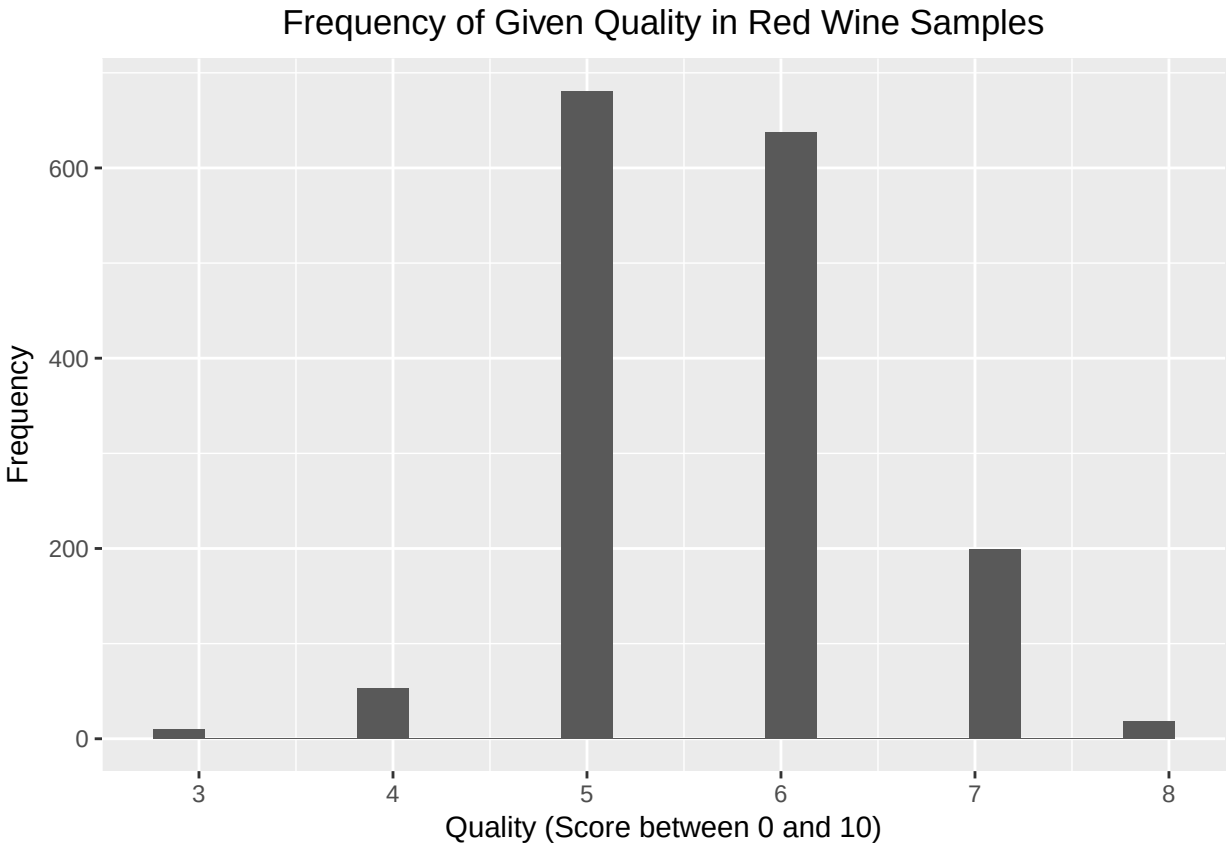
Were there any interesting or surprising interactions between features?

I thought it was interesting how although very high levels of alcohol did have a bad effect on quality, the bad effect in general of higher levels of alcohol wasn't as noticeable as the bad effect of high levels of other

chemicals. I think that this may indicate that alcohol content is one of the more important indicators of quality of red wine.

Final Plots and Summary

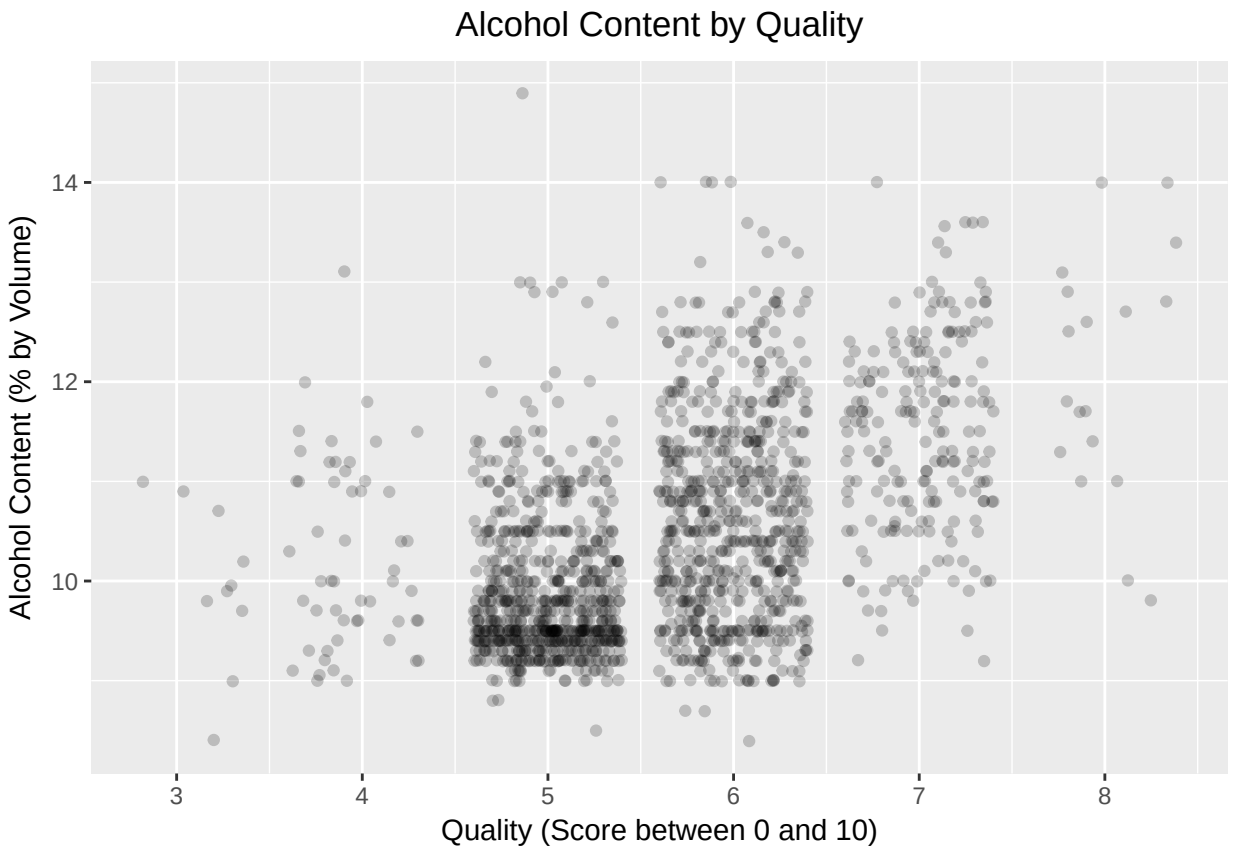
Plot One



Description One

The frequency of quality is unimodal because there is a single peak of frequency which occurs around the middle qualities (5 and 6).Both ends (3 and 8) have barely any scores, in other words, there's a big gap between the sum of 3 and 8 qualities and the sum of 5 and 6 qualities. If there were more samples with 3s and 8s, then perhaps we would be able to have a better idea of what makes good wine samples good and bad wine samples bad.

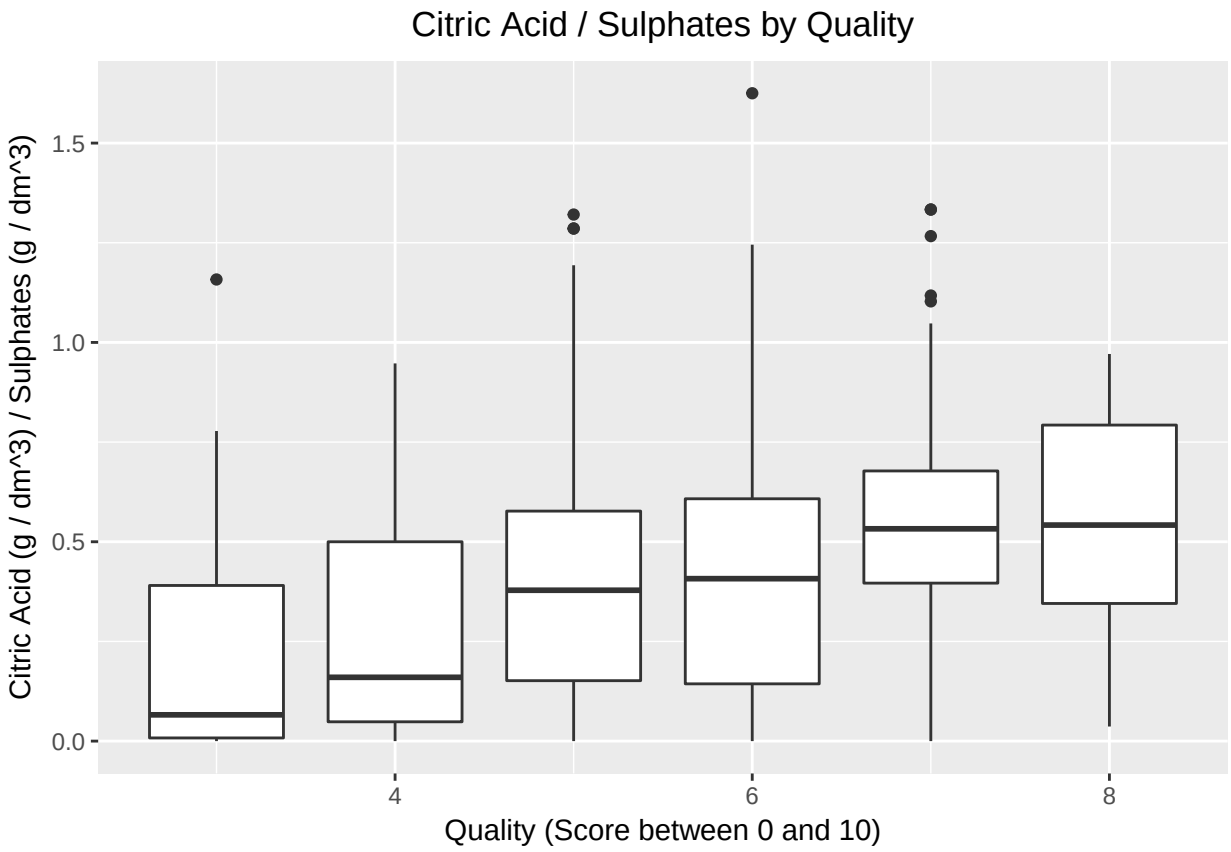
Plot Two



Description Two

As the Quality increases, more dots are shifting upwards in the alcohol content scale. There is a big density of dots for quality 5 volume 9% and this dense area is dispersed more evenly in quality 6. This indicates more alcohol content than 9% is likelier to give better quality.

Plot Three



Description Three

As quality increases, the ratio of citric acid by sulphates acid is also increasing, moving closer to 1. This indicates that when citric acids and sulphate acid quantities are both closer to each other, that's when there is better quality. Not only do the medians get higher but the top of the boxes (The 75th quartiles) also get bigger, so that means the trend is clear.

Reflection

I think it's interesting how important of a role alcohol played in the quality of red wine samples. More than any other factor, alcohol played the greatest role in determining the quality. Although it's a shame that there were not more samples of 8 in order to see what makes them so good, I still think that the findings suggest that alcohol is the strongest factor for quality. Besides alcohol, I think citric acid also plays a strong role in the quality of red wine. Both the bivariable and multivariable analysis proved this fact. I think this study could have been improved if there were more samples of 3s and 8s. It's unfortunate that the samples of quality were concentrated in the middle qualities of 5 and 6. This makes it more difficult to tell what makes good wine good and bad wine bad. This study should be done again with more bimodal sampling of wines, regarding their qualities. This way we can tell what makes good wine good and bad wine bad more clearly. I think what went successful with this study is the fact that I was able to narrow down alcohol's importance as a predictor of quality of red wine. Even citric acid was narrowed down as a factor, so that's another point of success. I think what was cause for struggle was that as mentioned before, there were not enough samples for

very bad qualities and very good qualities. Another struggle was that since Quality is the only categorical variable, I wasn't able to make a more variety of charts for the analysis. Instead, I was forced to always use Quality as the categorical variable in the analysis.