Explore and summarize Data

Ihsan Alsaeedi

28-November-2018

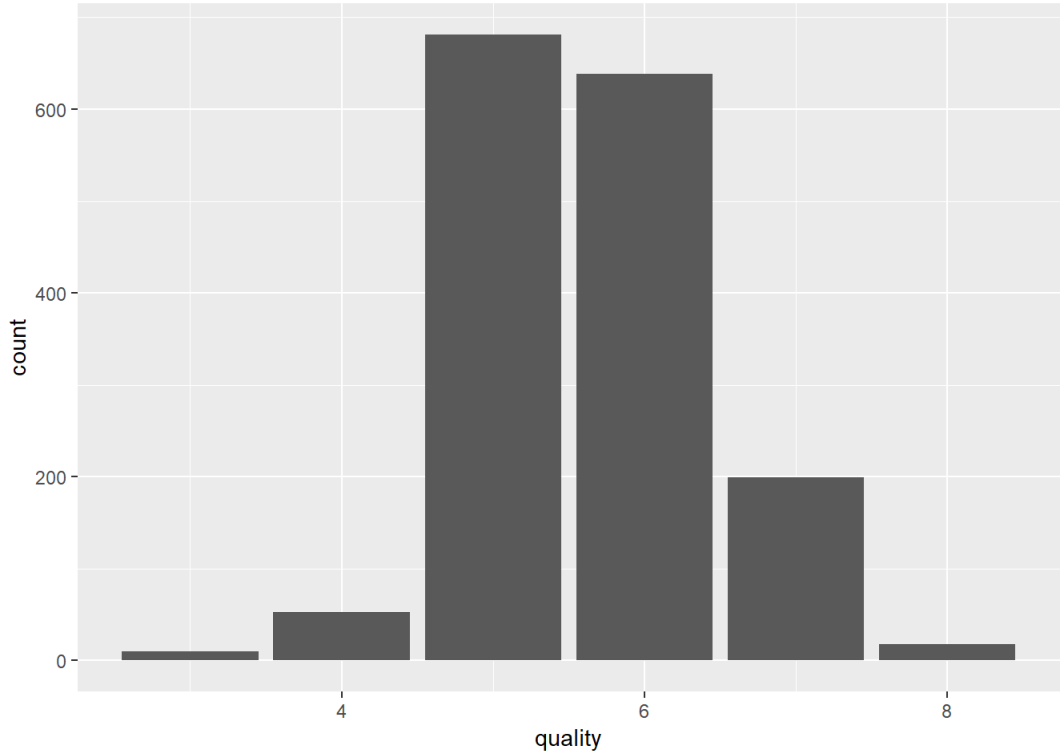
#Univariate Plots section

```
## '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 ...
```

- There are 1,599 observation with total 13 variables.

#Univariate Plots Section ## Quality

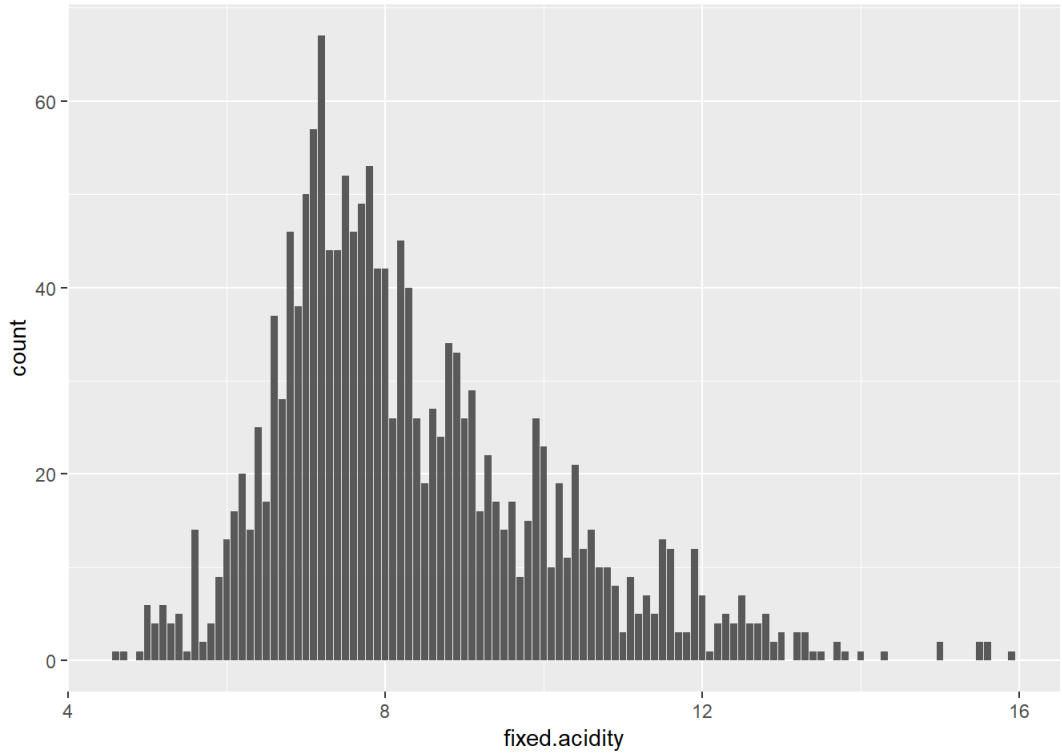
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.636   6.000   8.000
```



*The distribution of quality appears to be normal and concentrated around 5 and 6.

Fixed acidity

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

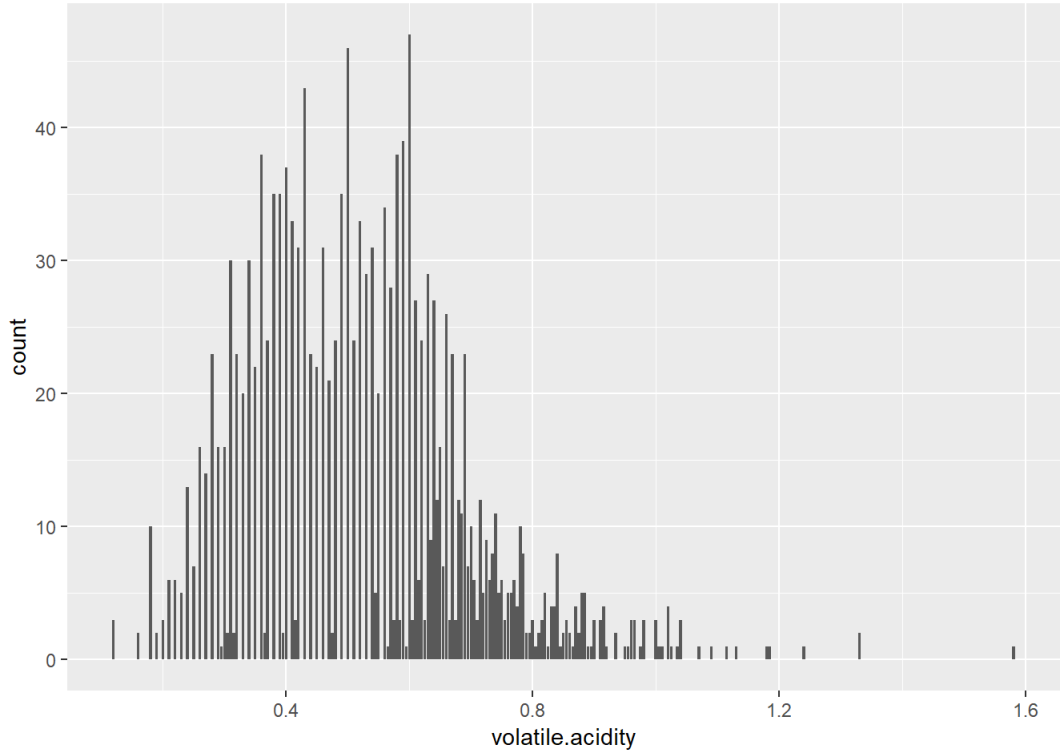


*The distribution of fixed

acidity is slightly right skewed and there are some outliers in the range(~ > 15).

Volatile acidity

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

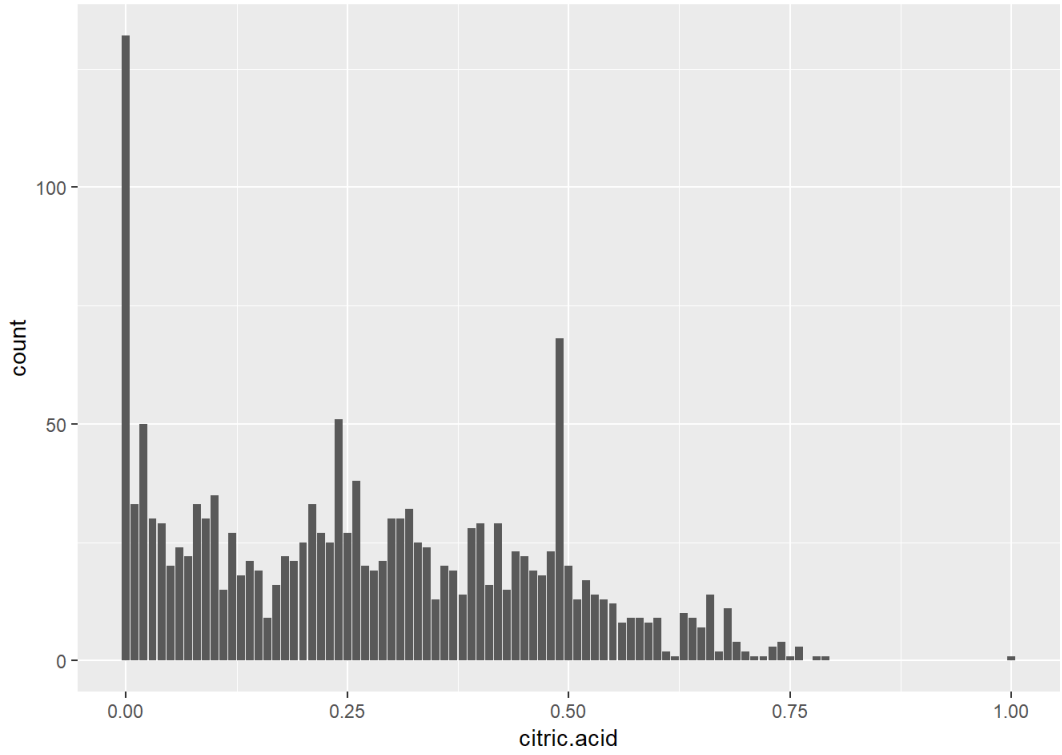


*The distribution of volatile

acidity is non-summetric and the outliers on the higher end of the scale are visible.

Citric acid

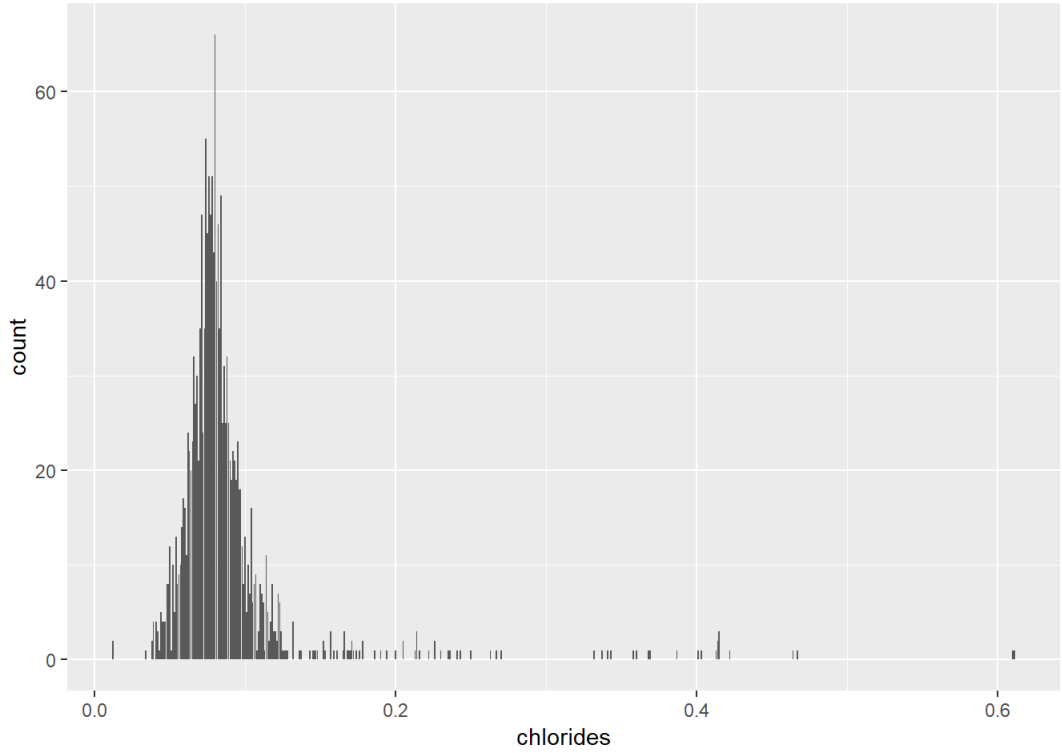
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```



*The distribution is right skewed.

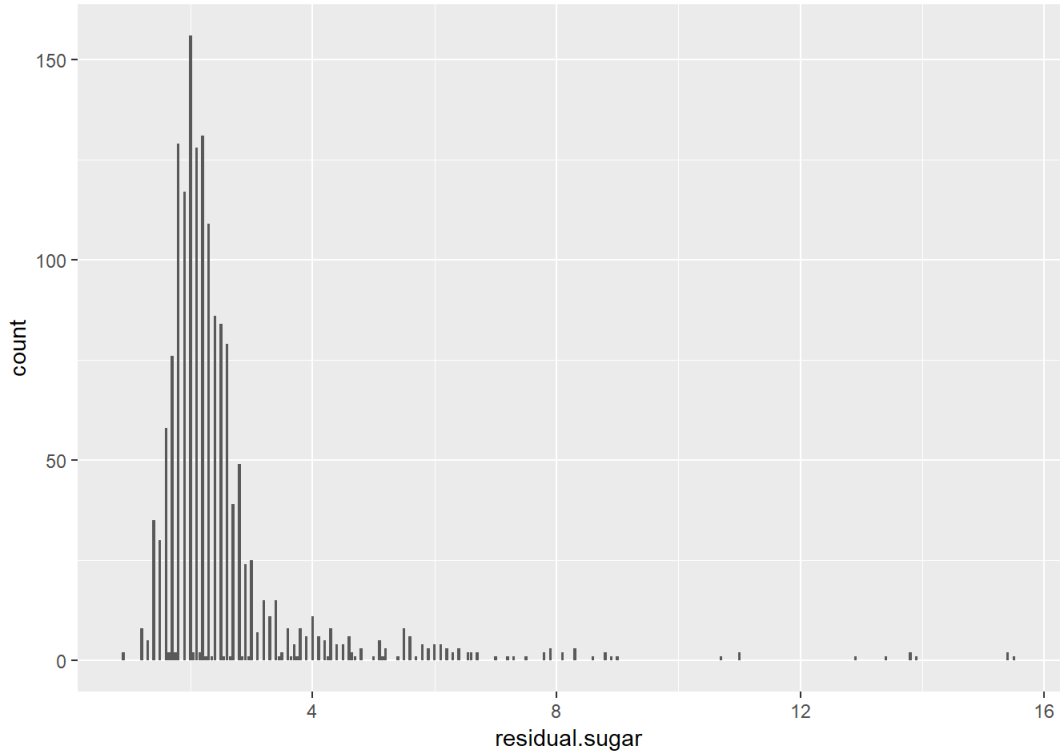
Chlorides

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



*The distribution with looks normal around its main peak but has a very long right tail.

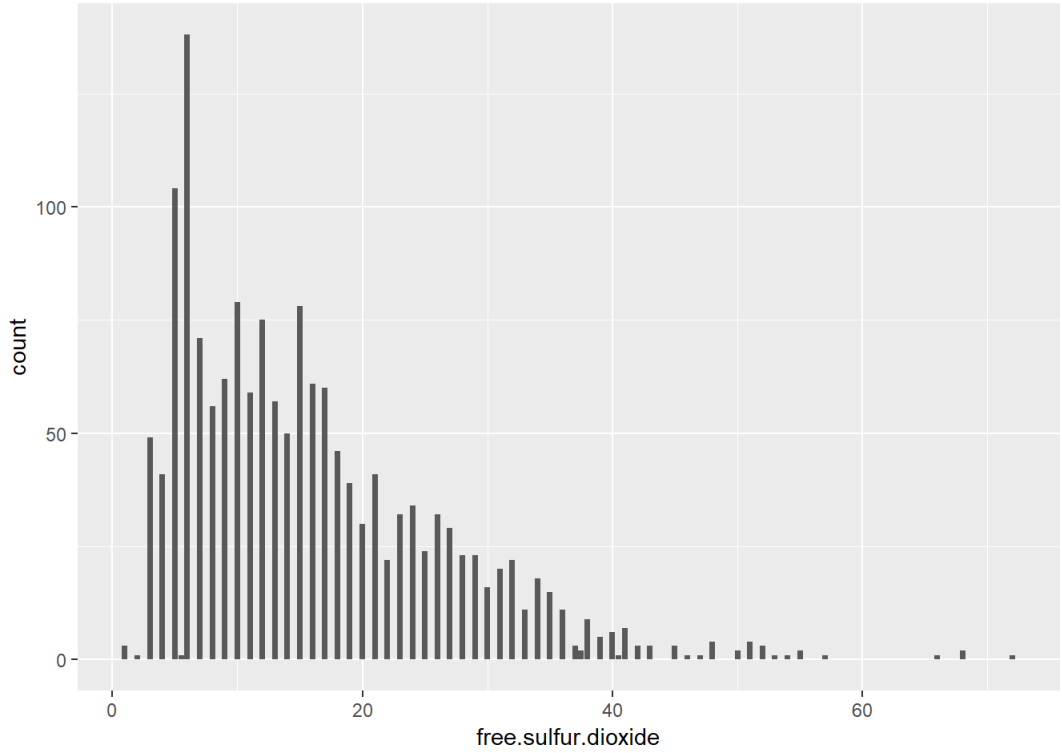
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```



*The distribution is right skewed with a long tail in the right side. There are many small bars on the right side of the main peak.

Free sulfur dioxide

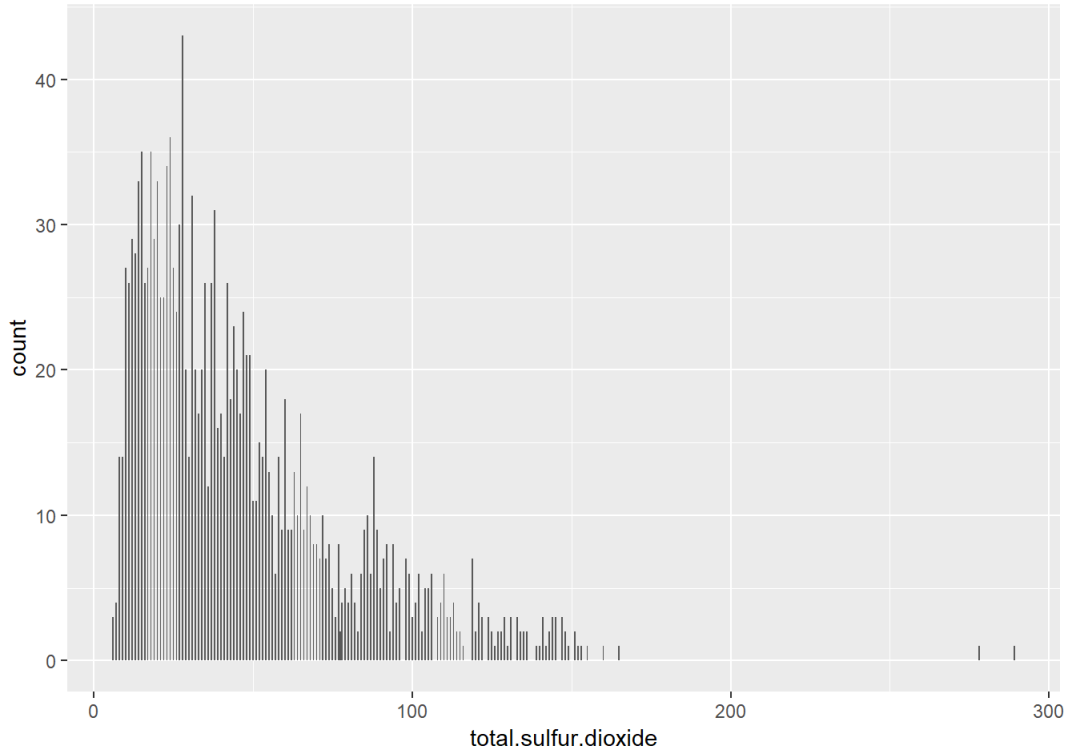
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
```



*The distribution of free sulfur dioxide concentrations is right skewed.

##Total sulfur dioxide

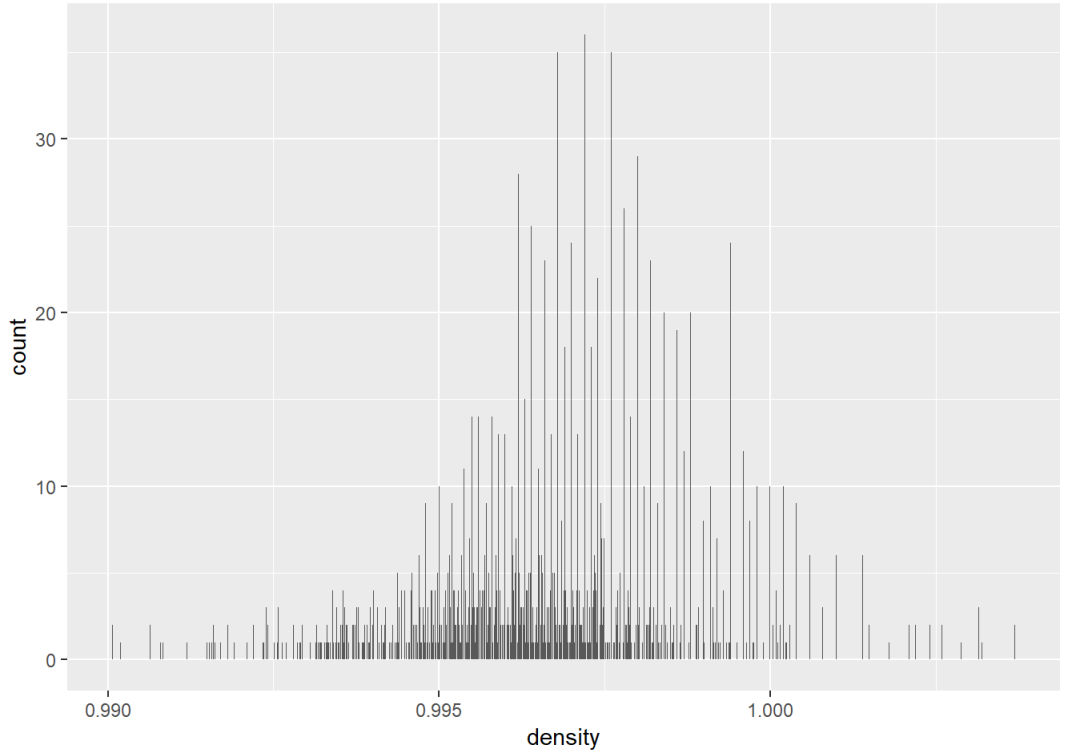
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00
```



*The distribution of total sulfur dioxide is right skewed.

Density

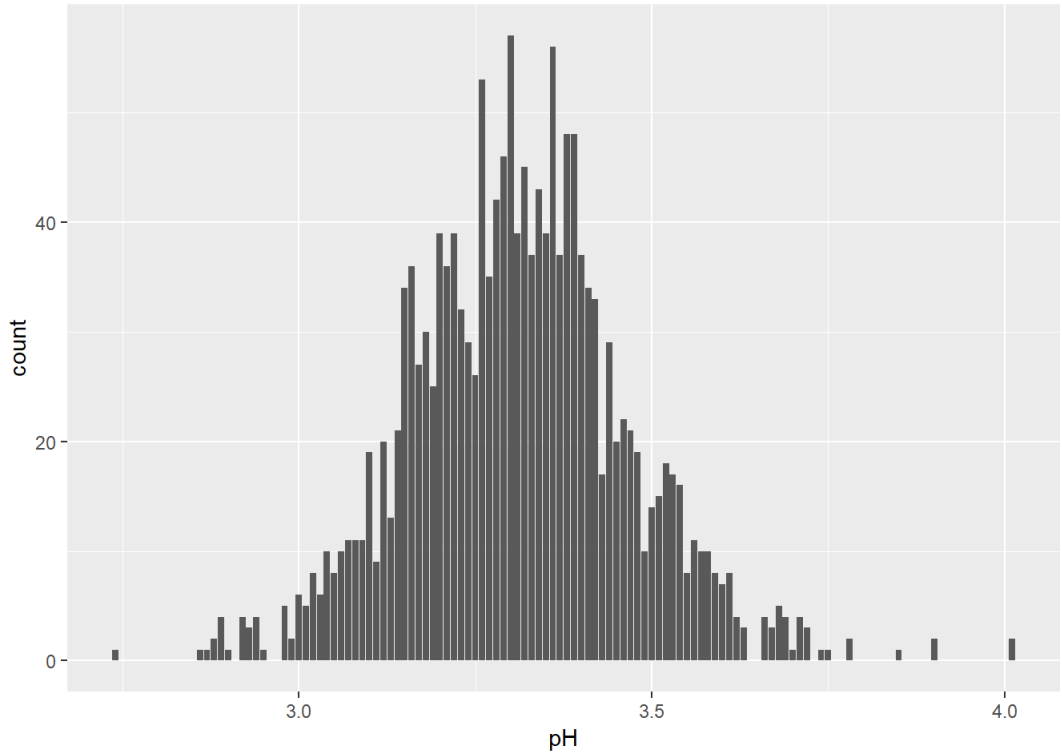
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0037
```



*The distribution is almost symmetric.

pH

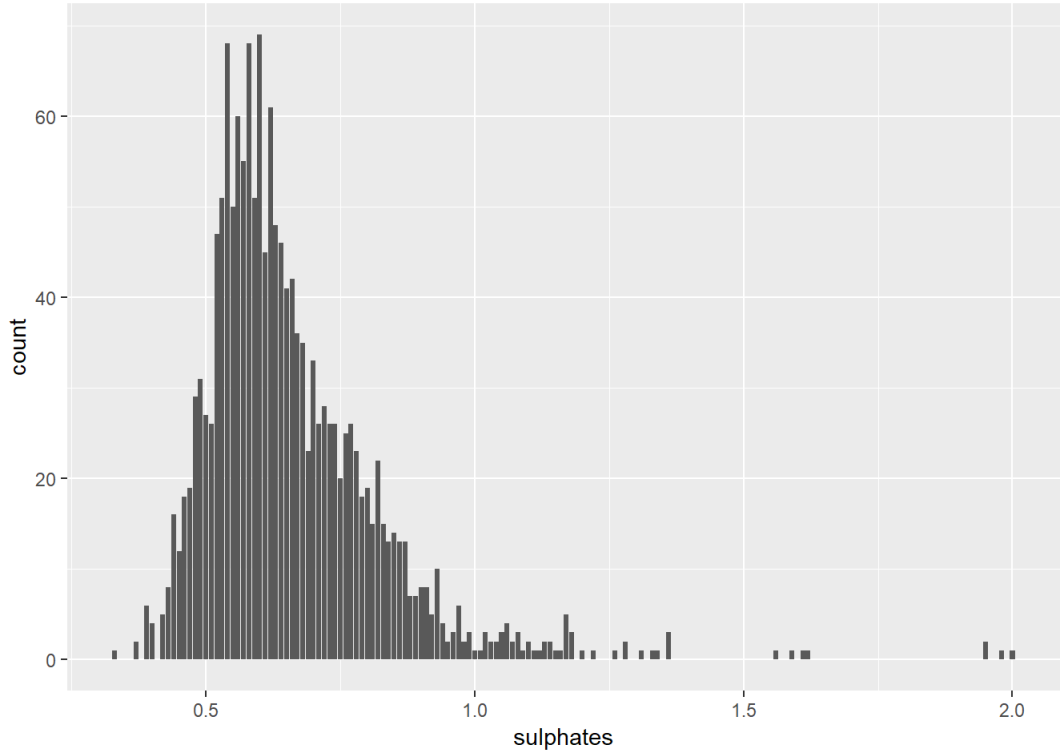
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010
```



*The distribution seems symmetrical or could be also considered bimodal with both peaks very close to each other.

Sulphates

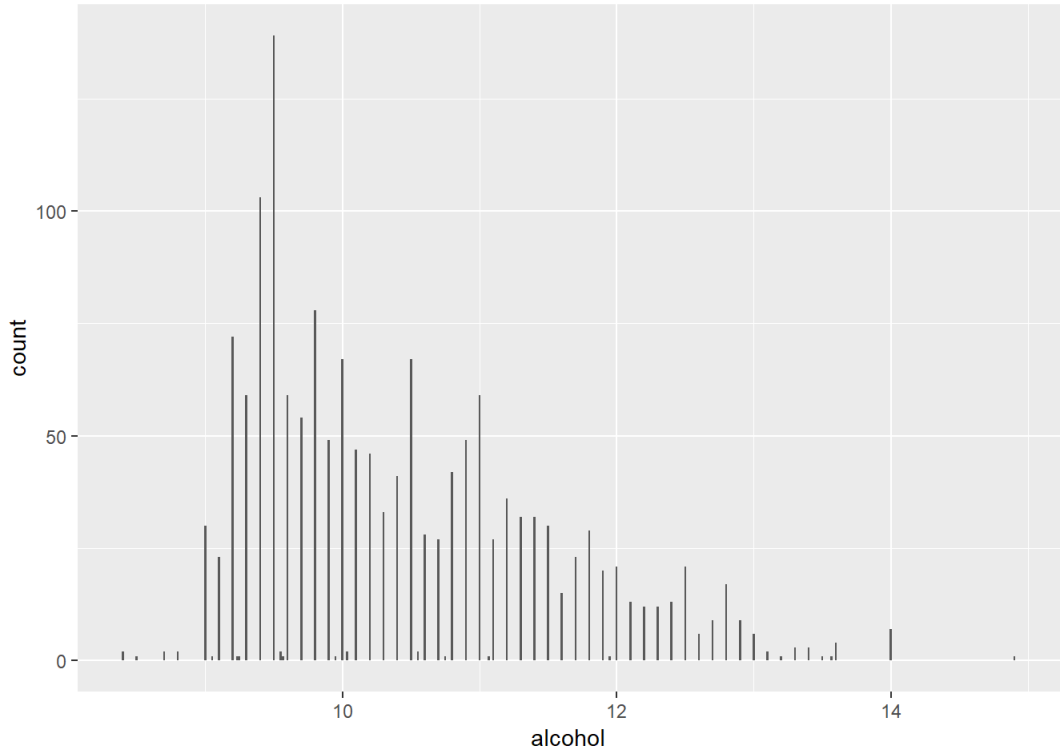
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```



*The distribution of sulphates is slightly right skewed. Some outliers on the right tail.

Alcohol

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



*The alcohol concentration distribution is right skewed.

#Univariate Analysis

What is the structure of your dataset?

There are 1599 red wines in the dataset with 11 features on the chemical properties of the wine.

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

The main features in the dataset are pH and quality.

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

I think all the physicochemical test results may help support the investigation. All of them are related to characteristics which may affect the flavor of the wine. They correspond to concentration of molecules which may have an impact on taste. Density is a physical property which will depend on the percentage of alcohol and sugar content, which will also affect taste.

Some variables may have strong correlation with each other. For instance, the pH will depend on the amount of acid molecules, while total sulfur dioxide may always follow a similar distribution of free sulfur dioxide.

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

No new variables were created in the dataset.

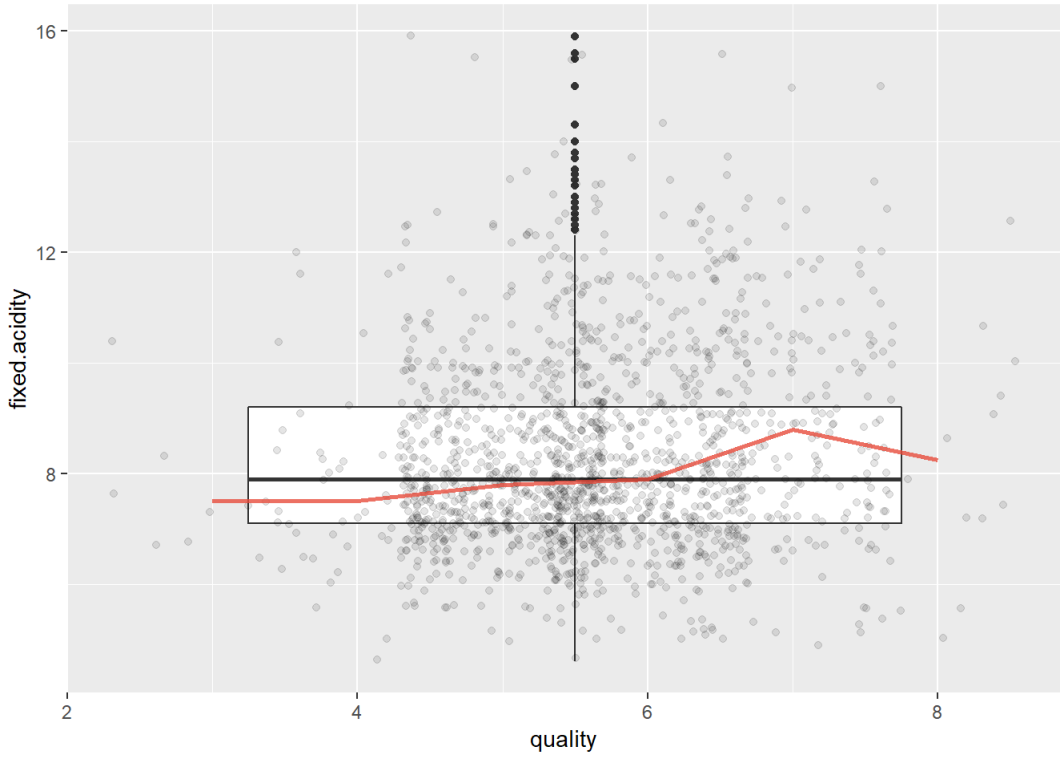
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

There were no unusual distributions, no missing values and no need to adjust the data. The dataset presented is already tidy which makes it an ideal dataset for a learning project as this one.

Bivariate Plots Section

Fixed Acidity vs. Quality

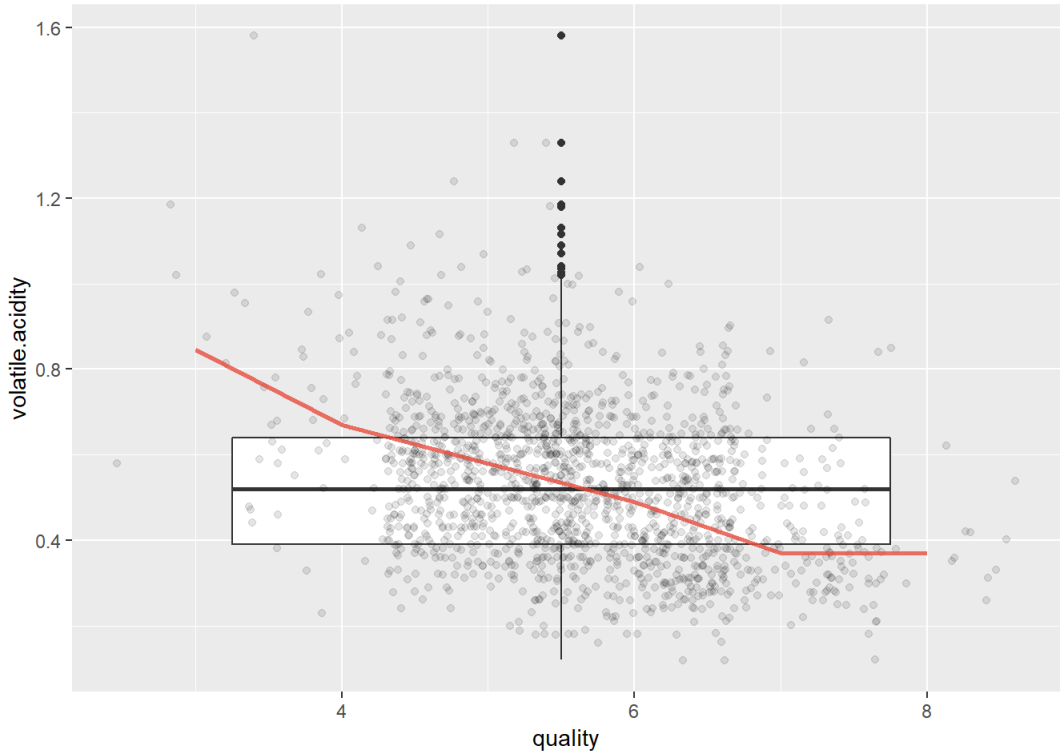
```
## [1] "Median of fixed.acidity by quality:"
## wine$quality: 3
## [1] 7.5
## ------------------------------------------------------------
## wine$quality: 4
## [1] 7.5
## ------------------------------------------------------------
## wine$quality: 5
## [1] 7.8
## ------------------------------------------------------------
## wine$quality: 6
## [1] 7.9
## ------------------------------------------------------------
## wine$quality: 7
## [1] 8.8
## ------------------------------------------------------------
## wine$quality: 8
## [1] 8.25
```



*We see a very slight upwards trend of higher quality with higher fixed acidity.

Volatile Acidity vs. Quality

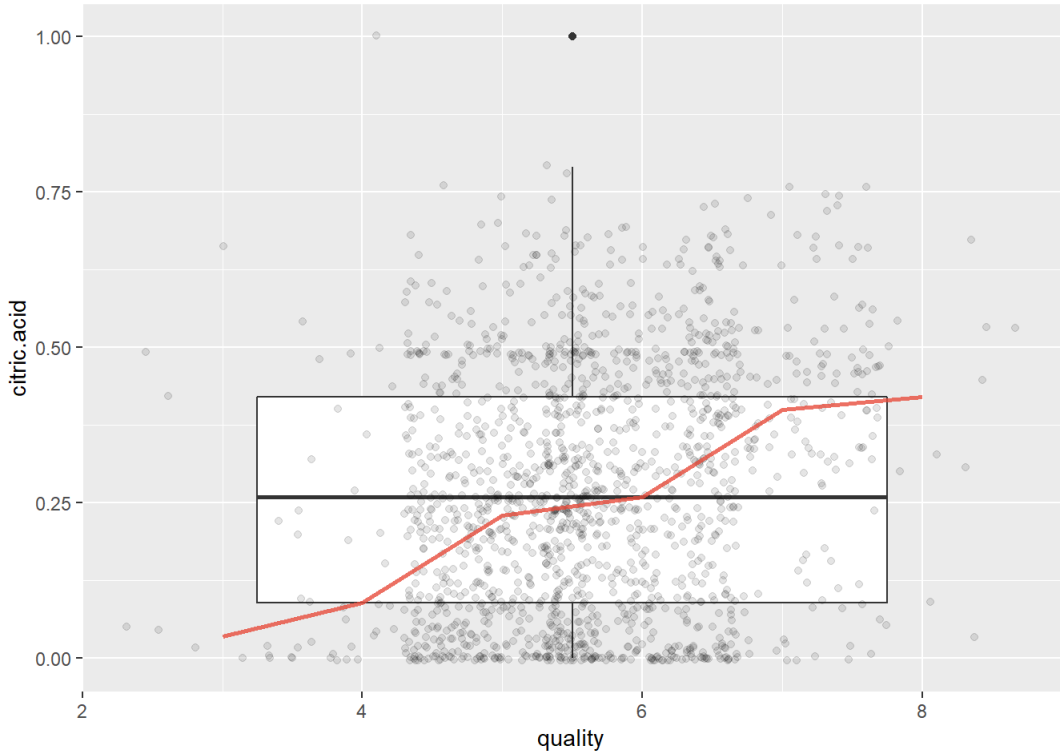
```
## [1] "Median of volatile.acidity by quality:"
## wine$quality: 3
## [1] 0.845
## ------------------------------------------------------------
## wine$quality: 4
## [1] 0.67
## ------------------------------------------------------------
## wine$quality: 5
## [1] 0.58
## ------------------------------------------------------------
## wine$quality: 6
## [1] 0.49
## ------------------------------------------------------------
## wine$quality: 7
## [1] 0.37
## ------------------------------------------------------------
## wine$quality: 8
## [1] 0.37
```



*Lower volatile acidity seems to mean higher wine quality.

Citric Acid vs. Quality

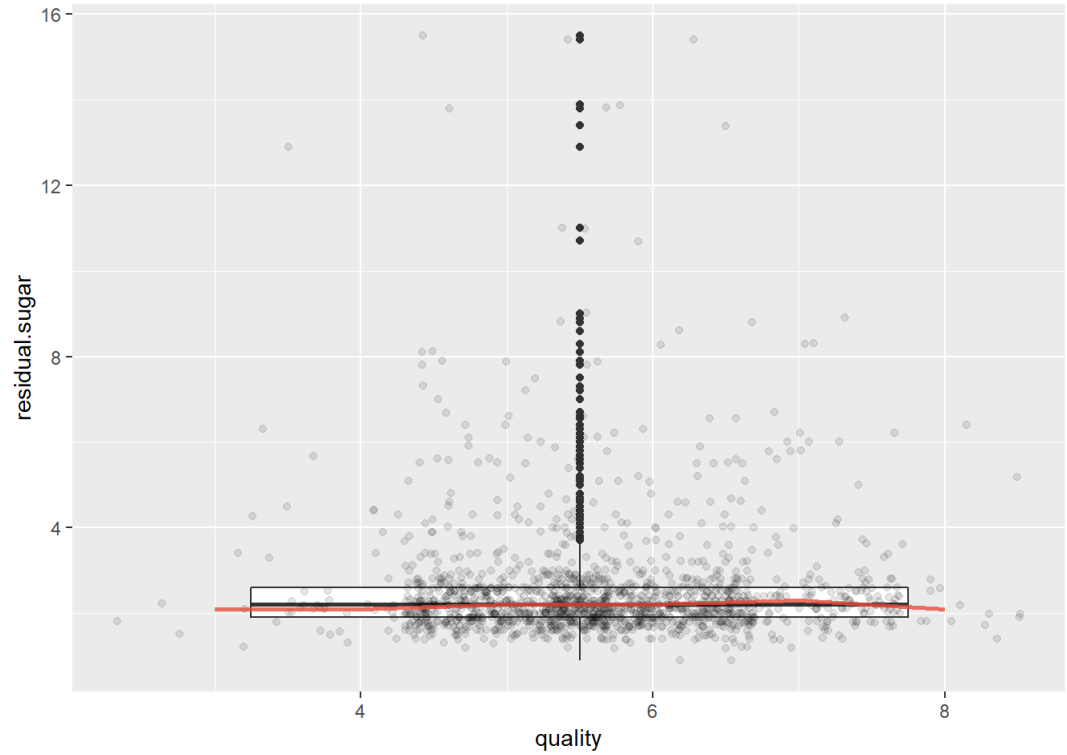
```
## [1] "Median of citric.acid by quality:"
## wine$quality: 3
## [1] 0.035
## ------------------------------------------------------------
## wine$quality: 4
## [1] 0.09
## ------------------------------------------------------------
## wine$quality: 5
## [1] 0.23
## ------------------------------------------------------------
## wine$quality: 6
## [1] 0.26
## ------------------------------------------------------------
## wine$quality: 7
## [1] 0.4
## ------------------------------------------------------------
## wine$quality: 8
## [1] 0.42
```



*Higher citric acid seems to mean a higher quality wine. The citric acid is always in low concentrations and in the univariate plots we saw that the distribution peaked at the zero value.

Residual Sugar vs. Quality

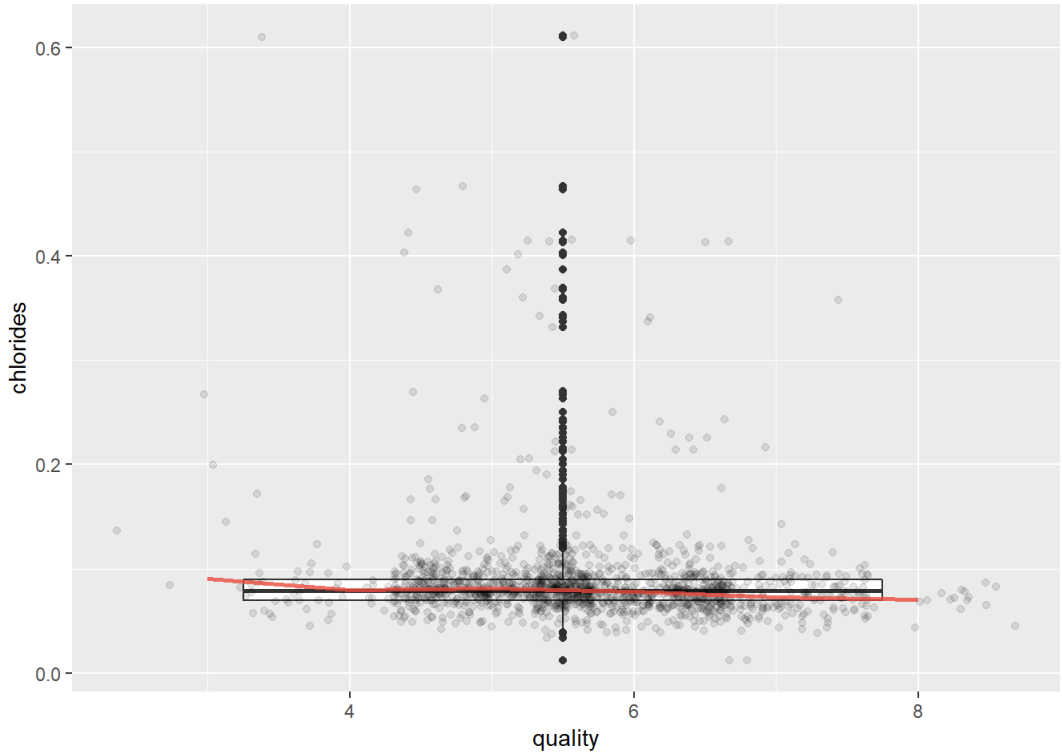
```
## [1] "Median of residual.sugar by quality:"
## wine$quality: 3
## [1] 2.1
## ------------------------------------------------------------
## wine$quality: 4
## [1] 2.1
## ------------------------------------------------------------
## wine$quality: 5
## [1] 2.2
## ------------------------------------------------------------
## wine$quality: 6
## [1] 2.2
## ------------------------------------------------------------
## wine$quality: 7
## [1] 2.3
## ------------------------------------------------------------
## wine$quality: 8
## [1] 2.1
```



*Residual sugar seems to have a low impact in the quality of the wine.

Chlorides vs. Quality

```
## [1] "Median of chlorides by quality:"
## wine$quality: 3
## [1] 0.0905
## ------------------------------------------------------------
## wine$quality: 4
## [1] 0.08
## ------------------------------------------------------------
## wine$quality: 5
## [1] 0.081
## ------------------------------------------------------------
## wine$quality: 6
## [1] 0.078
## ------------------------------------------------------------
## wine$quality: 7
## [1] 0.073
## ------------------------------------------------------------
## wine$quality: 8
## [1] 0.0705
```

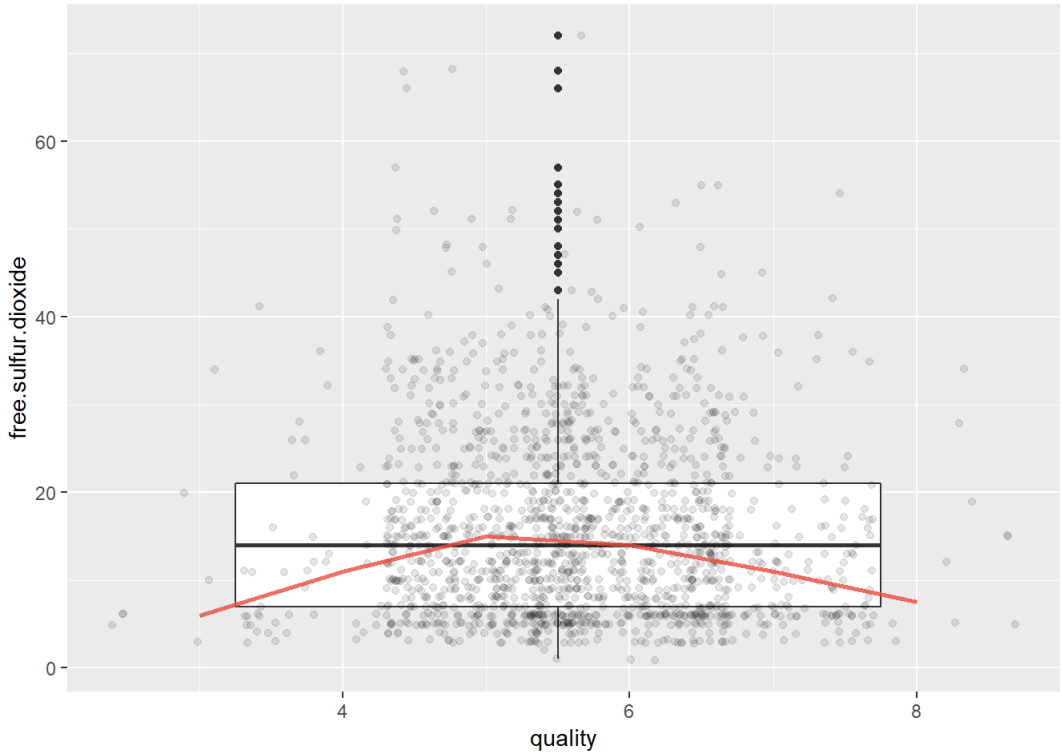


*A slight relation. Less

chlorides means higher quality.

Free sulfur dioxide vs. Quality

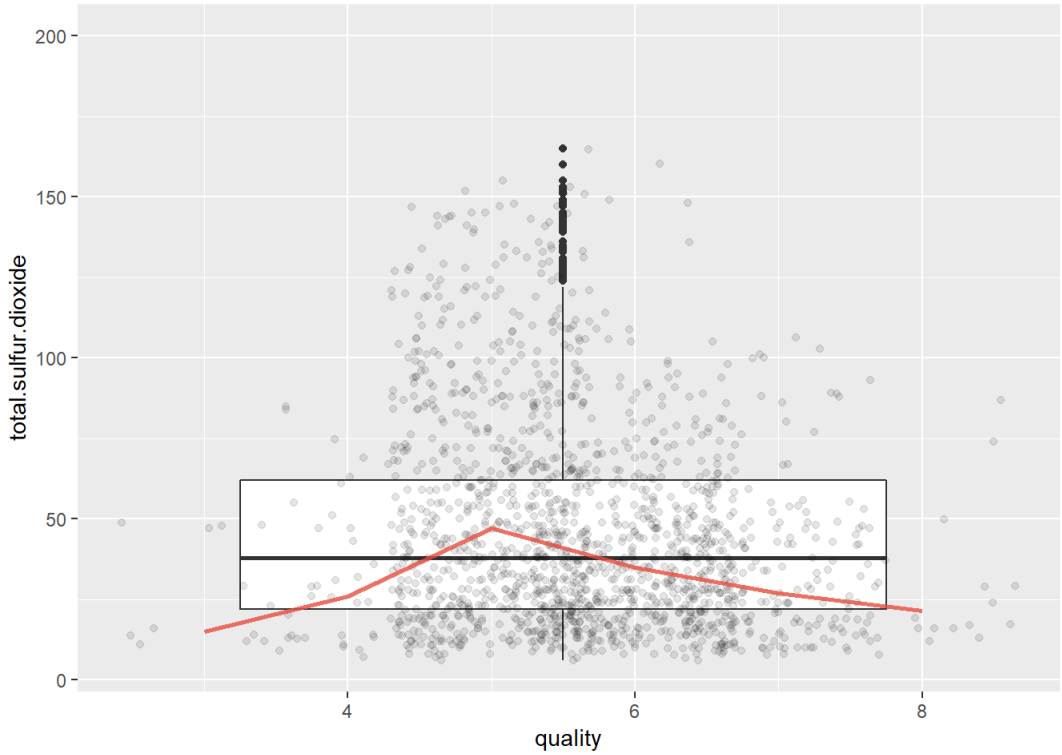
```
## [1] "Median of free.sulfur.dioxide by quality:"
## wine$quality: 3
## [1] 6
## ------------------------------------------------------------
## wine$quality: 4
## [1] 11
## ------------------------------------------------------------
## wine$quality: 5
## [1] 15
## ------------------------------------------------------------
## wine$quality: 6
## [1] 14
## ------------------------------------------------------------
## wine$quality: 7
## [1] 11
## ------------------------------------------------------------
## wine$quality: 8
## [1] 7.5
```



*The middle quality classes seem to have higher free sulfur dioxide than both the low and high quality.

Total sulfur dioxide vs. Quality

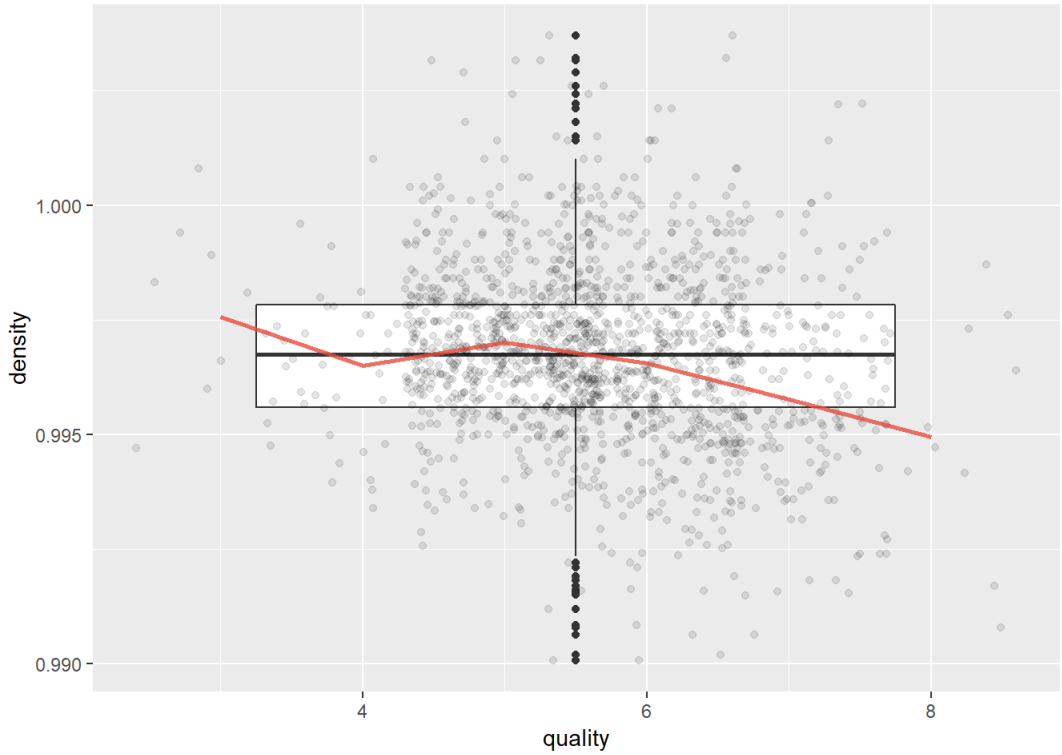
```
## [1] "Median of total.sulfur.dioxide by quality:"
## wine$quality: 3
## [1] 15
## ------------------------------------------------------------
## wine$quality: 4
## [1] 26
## ------------------------------------------------------------
## wine$quality: 5
## [1] 47
## ------------------------------------------------------------
## wine$quality: 6
## [1] 35
## ------------------------------------------------------------
## wine$quality: 7
## [1] 27
## ------------------------------------------------------------
## wine$quality: 8
## [1] 21.5
```



*Similar relation as with free sulfur dioxide. The middle classes have higher concentration than both the low and high.

Density vs. Quality

```
## [1] "Median of density by quality:"
## wine$quality: 3
## [1] 0.997565
## ------------------------------------------------------------
## wine$quality: 4
## [1] 0.9965
## ------------------------------------------------------------
## wine$quality: 5
## [1] 0.997
## ------------------------------------------------------------
## wine$quality: 6
## [1] 0.99656
## ------------------------------------------------------------
## wine$quality: 7
## [1] 0.99577
## ------------------------------------------------------------
## wine$quality: 8
## [1] 0.99494
```

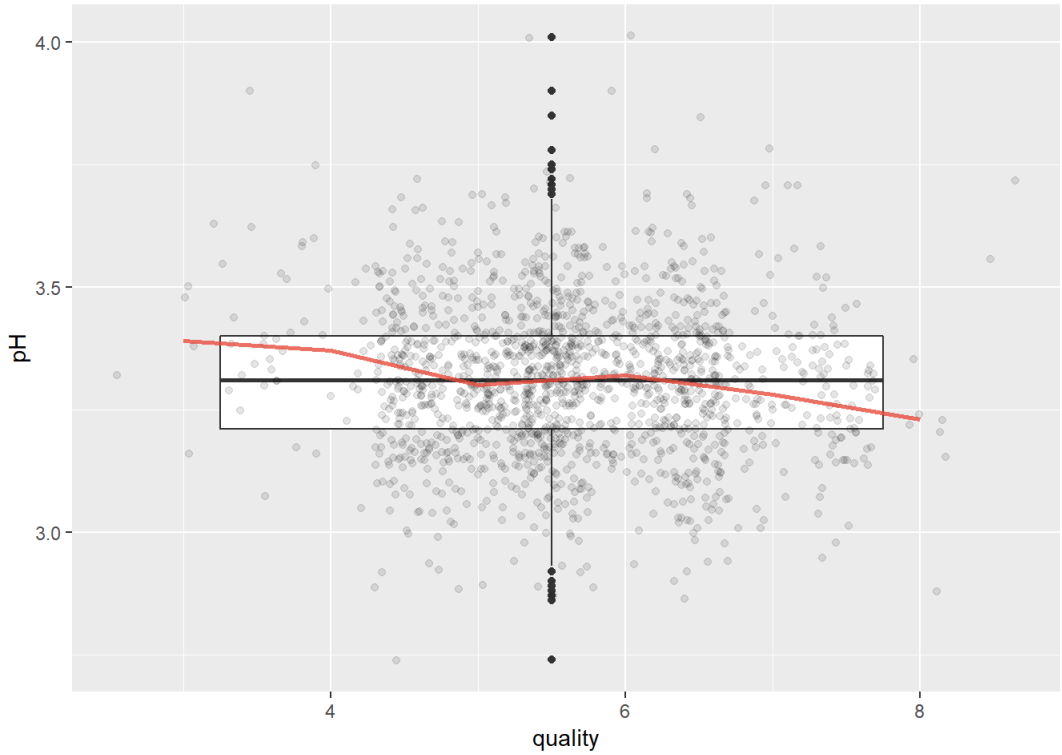


*Lower density means higher

quality. From the dataset descriptions we know that the density will depend on the percentage of alcohol and sugar content. We should check those relationships later.

pH vs. Quality

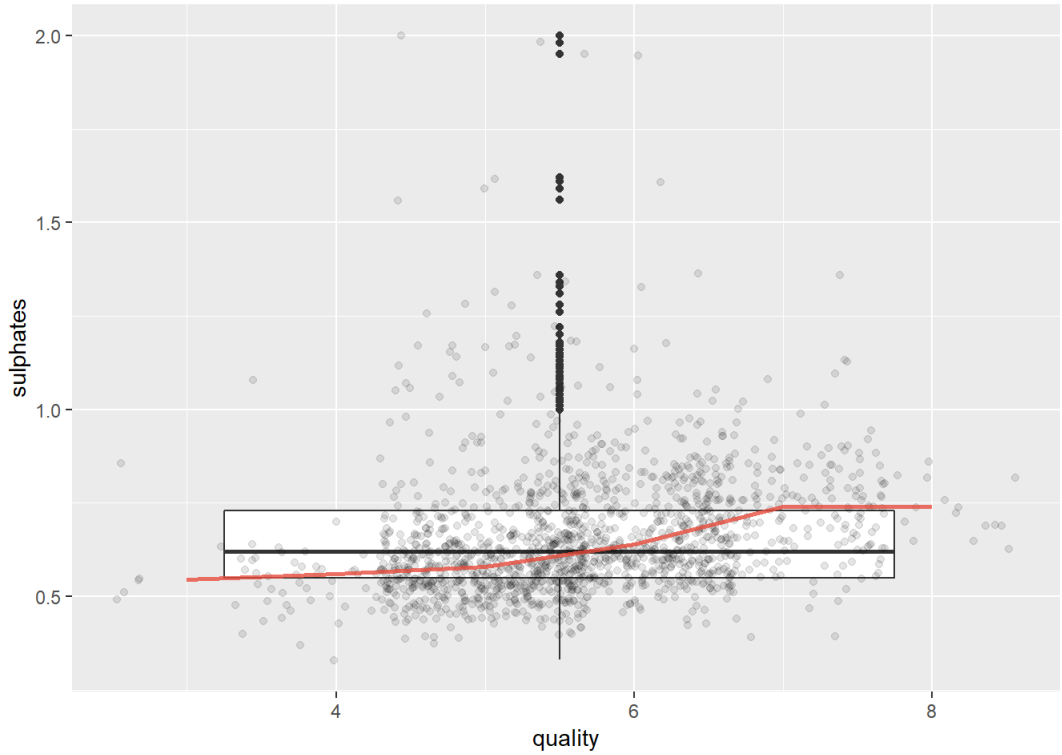
```
## [1] "Median of pH by quality:"
## wine$quality: 3
## [1] 3.39
## ------------------------------------------------------------
## wine$quality: 4
## [1] 3.37
## ------------------------------------------------------------
## wine$quality: 5
## [1] 3.3
## ------------------------------------------------------------
## wine$quality: 6
## [1] 3.32
## ------------------------------------------------------------
## wine$quality: 7
## [1] 3.28
## ------------------------------------------------------------
## wine$quality: 8
## [1] 3.23
```



*There seems to be a trend of higher quality with lower pH. Higher quality with more acid content? We should check correlations between pH and the acidity levels.

Sulphates vs. Quality

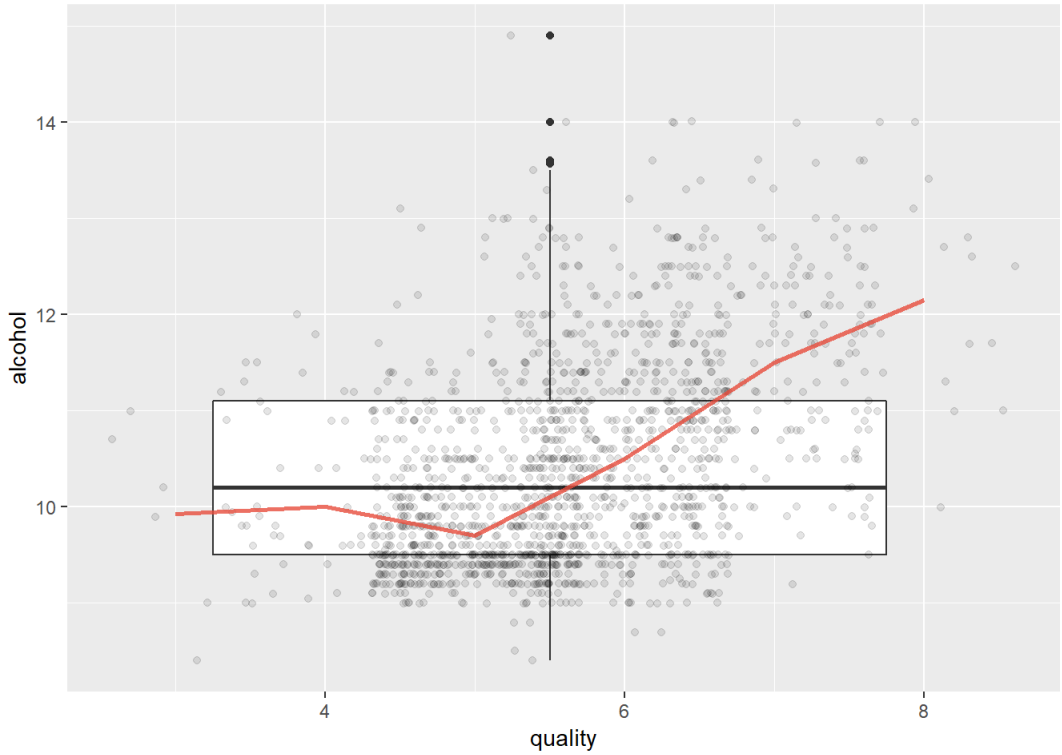
```
## [1] "Median of sulphates by quality:"
## wine$quality: 3
## [1] 0.545
## ------------------------------------------------------------
## wine$quality: 4
## [1] 0.56
## ------------------------------------------------------------
## wine$quality: 5
## [1] 0.58
## ------------------------------------------------------------
## wine$quality: 6
## [1] 0.64
## ------------------------------------------------------------
## wine$quality: 7
## [1] 0.74
## ------------------------------------------------------------
## wine$quality: 8
## [1] 0.74
```



*Higher sulphates concentration means higher quality.

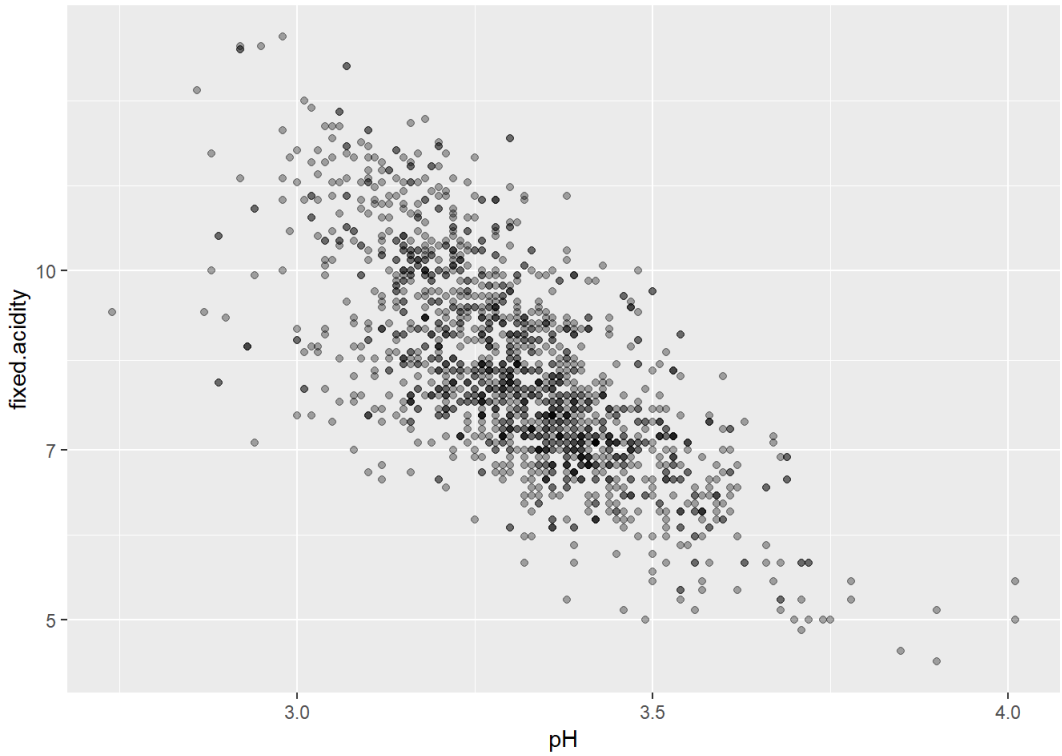
Alcohol vs. Quality

```
## [1] "Median of alcohol by quality:"
## wine$quality: 3
## [1] 9.925
## ------------------------------------------------------------
## wine$quality: 4
## [1] 10
## ------------------------------------------------------------
## wine$quality: 5
## [1] 9.7
## ------------------------------------------------------------
## wine$quality: 6
## [1] 10.5
## ------------------------------------------------------------
## wine$quality: 7
## [1] 11.5
## ------------------------------------------------------------
## wine$quality: 8
## [1] 12.15
```



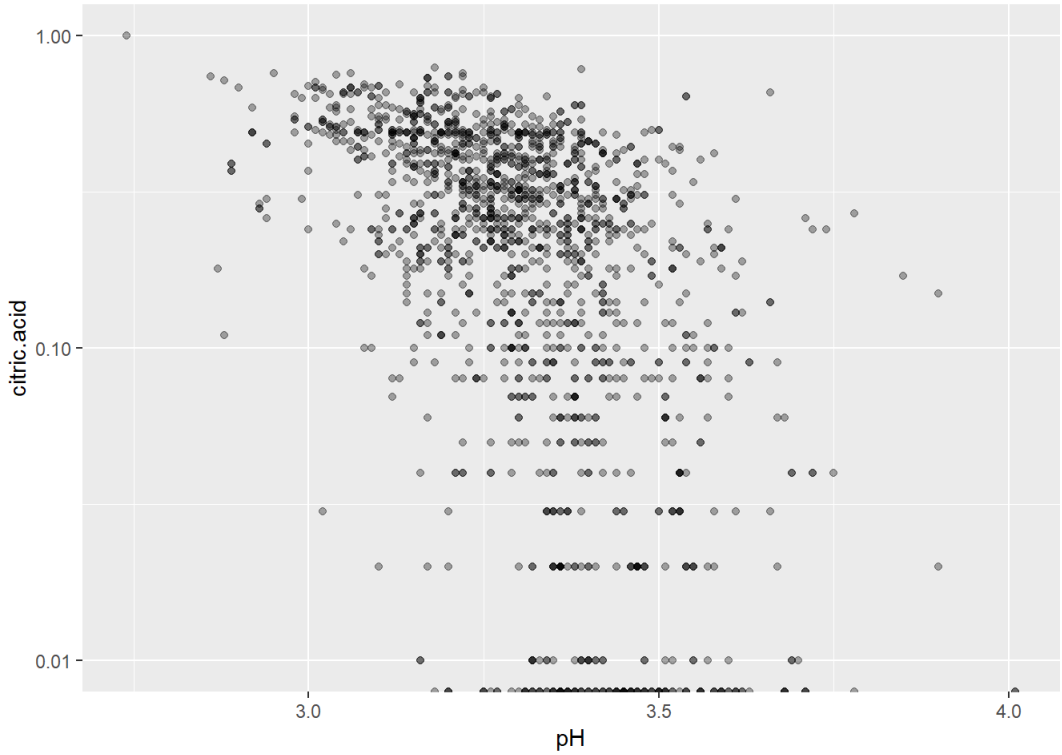
*Besides the small downward bump in the quality class 5, the higher the alcohol content, the higher rated the wines get.

Acidity and pH



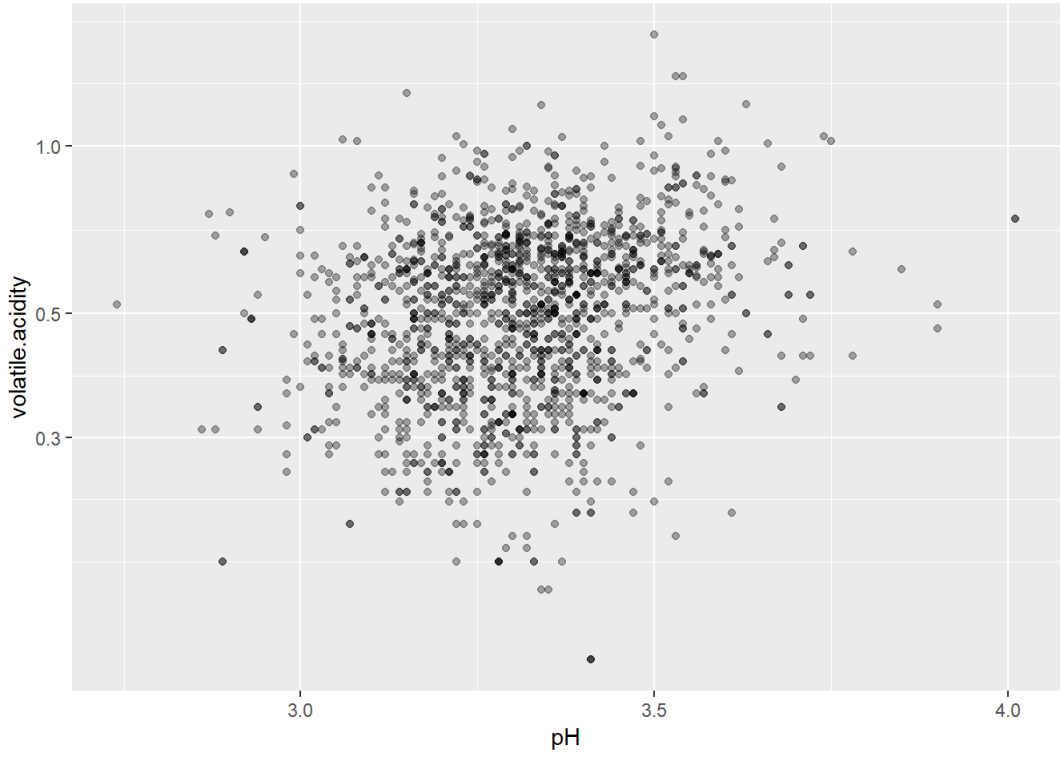
*As expected the pH increases with the lower amount of acids. Fixed acidity accounts for most acids present in the wine.

Citric acid and pH



*A similar relation is seen with the citric acid variable. But since the citric acid is at lower concentrations, the relation is not so strong. pH will be dominated by the other acids.

volatile acidity

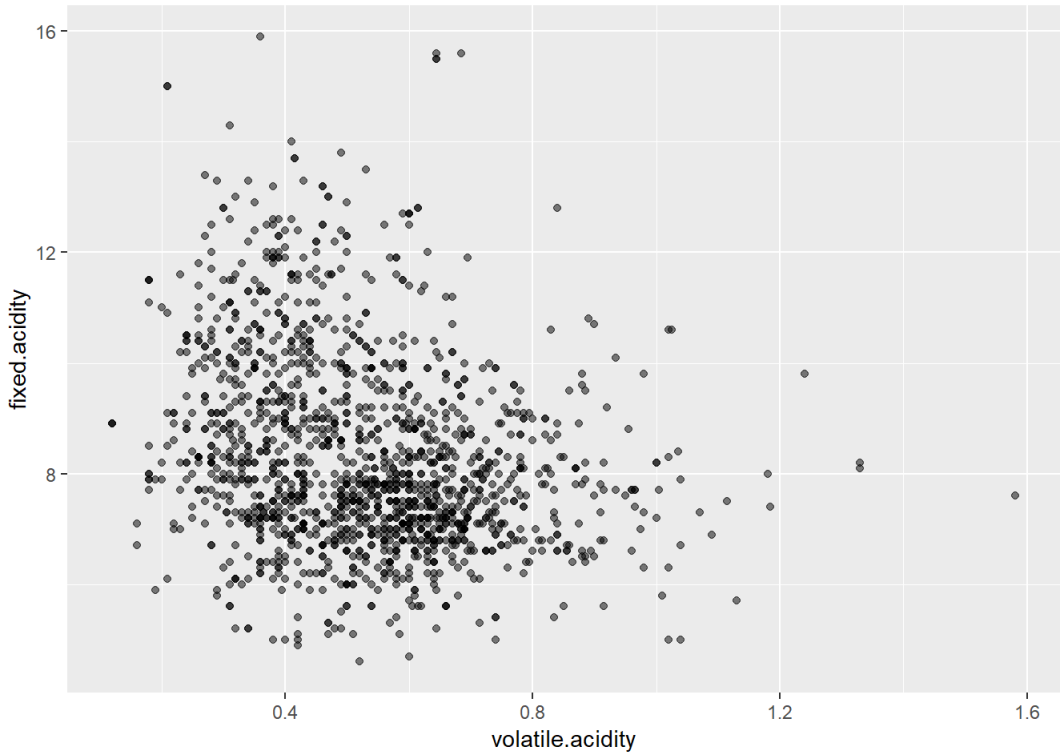


*The volatile acidity seems to have either no relation with the pH or a slight positive correlation.

Correlation coefficient

```
##
##  Pearson's product-moment correlation
##
## data:  pH and log10(volatile.acidity)
## t = 9.1468, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1760195 0.2691923
## sample estimates:
##       cor
## 0.2231154
```

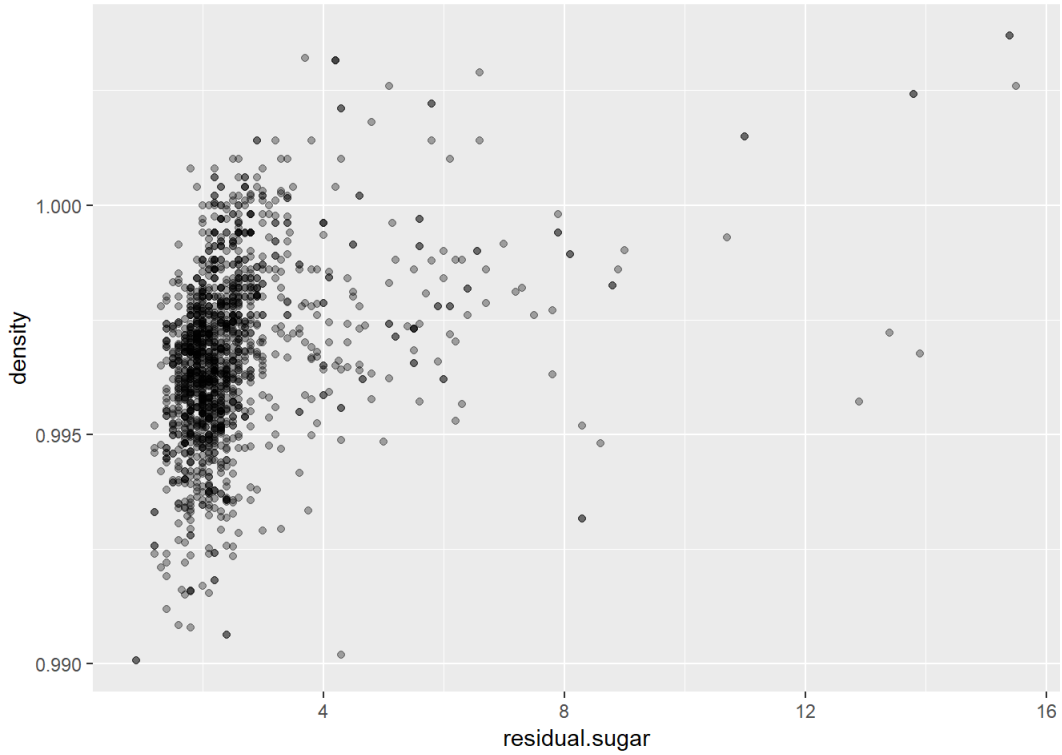
*The correlation coefficient shows a weak positive correlation of volatile.acidity with the pH. Maybe when the volatile acids are present in higher concentration, the concentration of the remaining acids is lower and that contributes to the increase of pH.

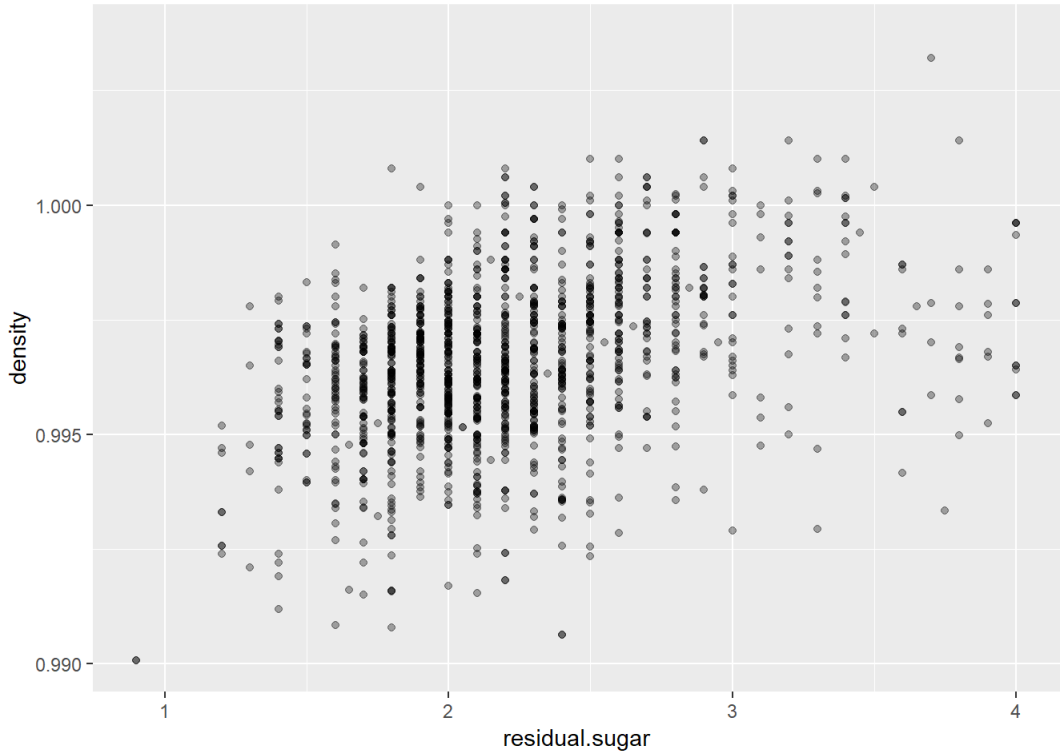



```
##
##  Pearson's product-moment correlation
##
## data:  fixed.acidity and volatile.acidity
## t = -10.589, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3013681 -0.2097433
## sample estimates:
##        cor
## -0.2561309
```

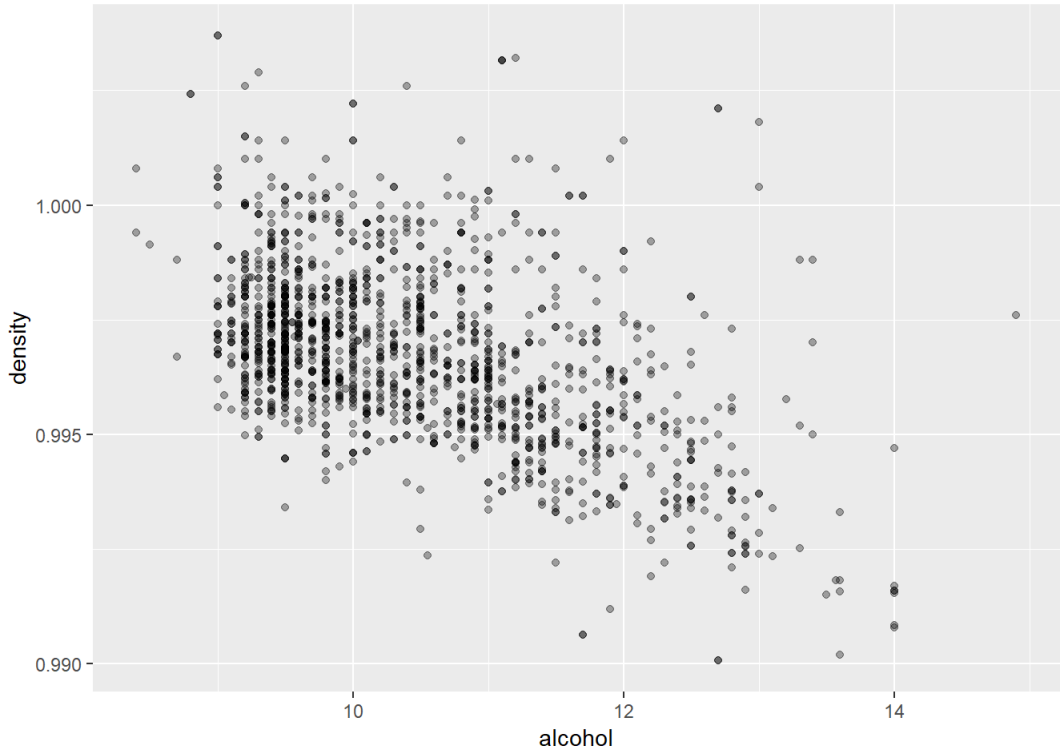
*We can see a weak negative correlation. On the plot, both variables seems to be have a natural limit on the lower sides. We have seen on the univariate plots that both are right skewed.

Density, Sugar and Alcohol Content

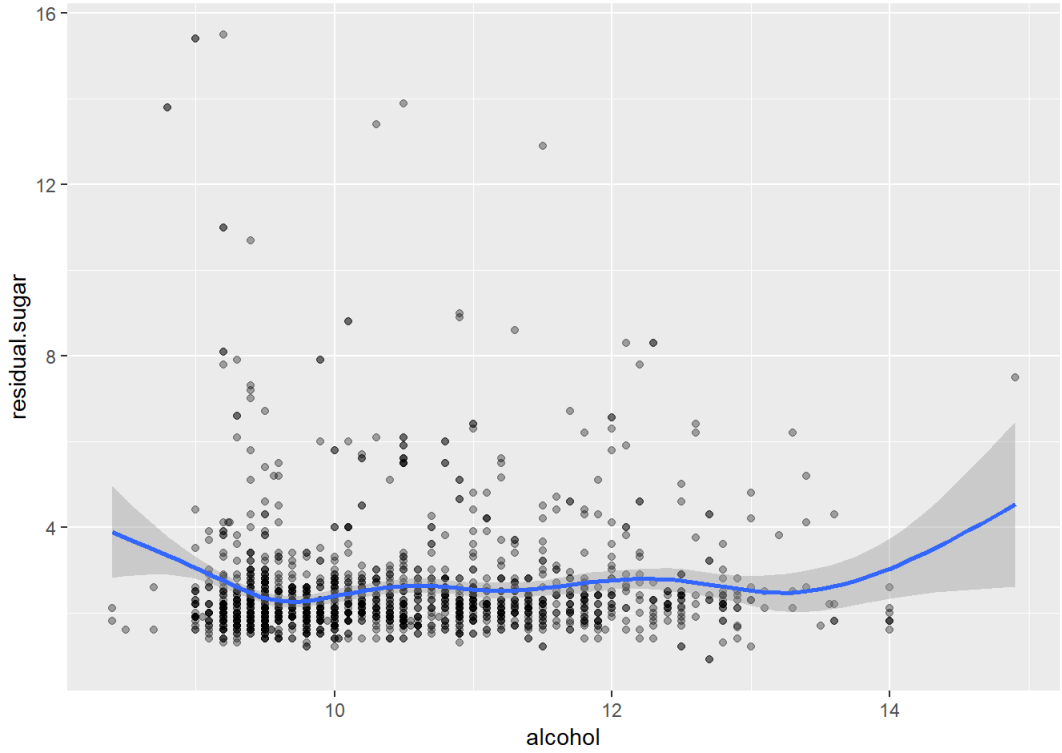


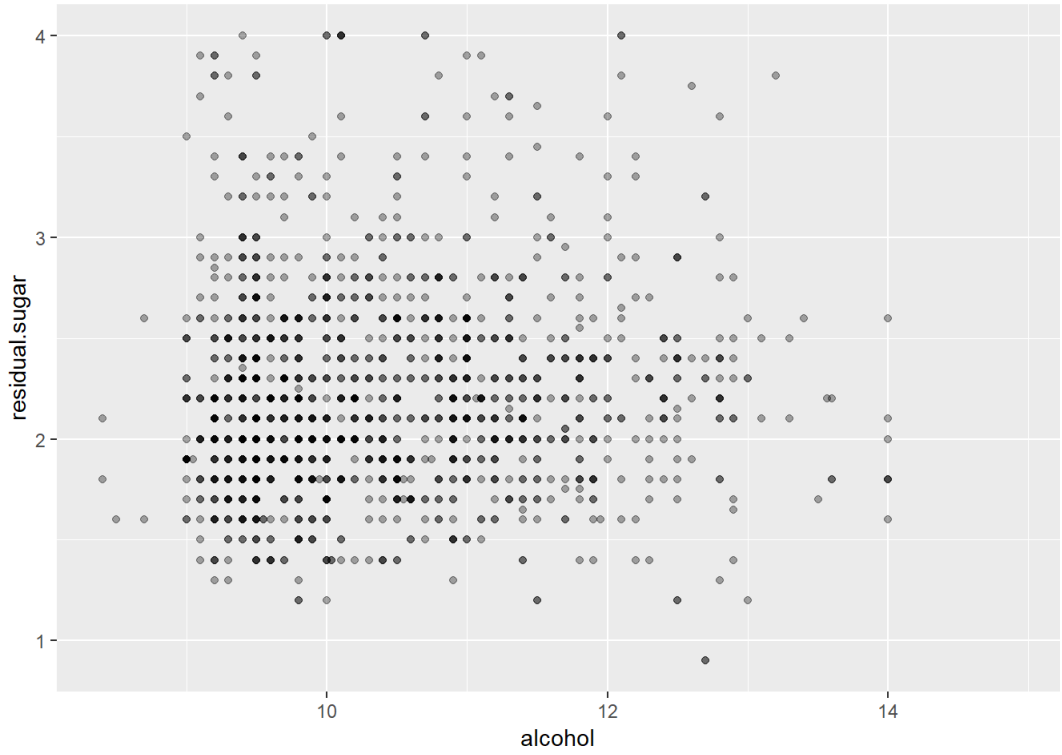


*We see a increase of density with increase of residual sugar.



*And we see a decrease of density with increase of alcohol content.



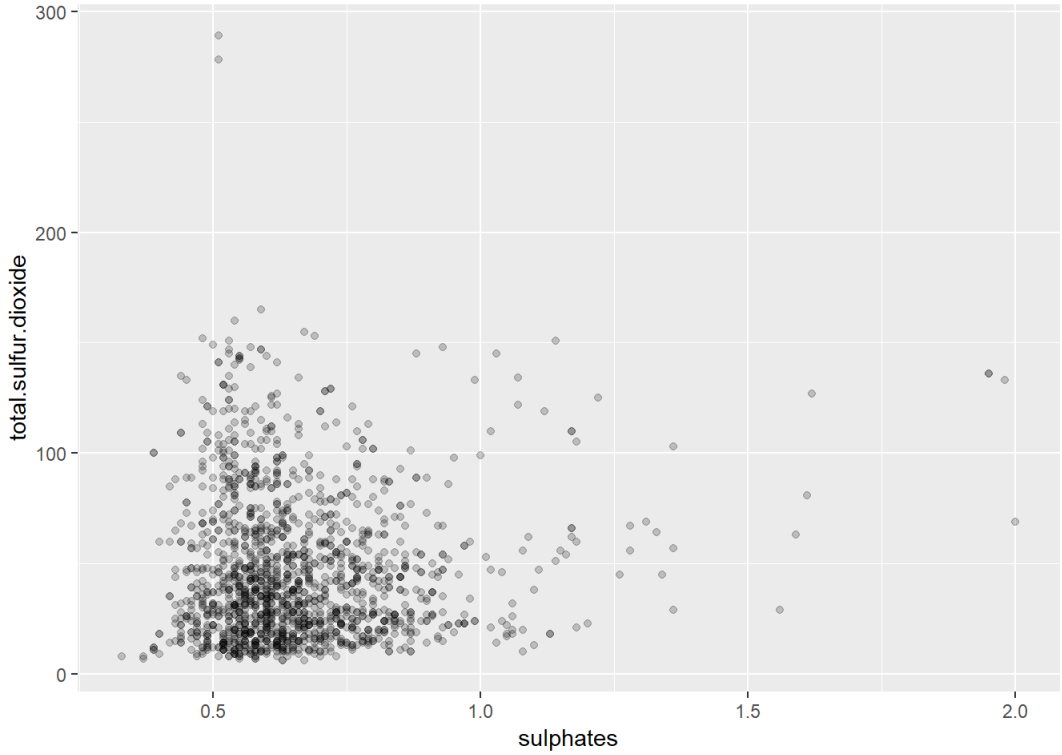


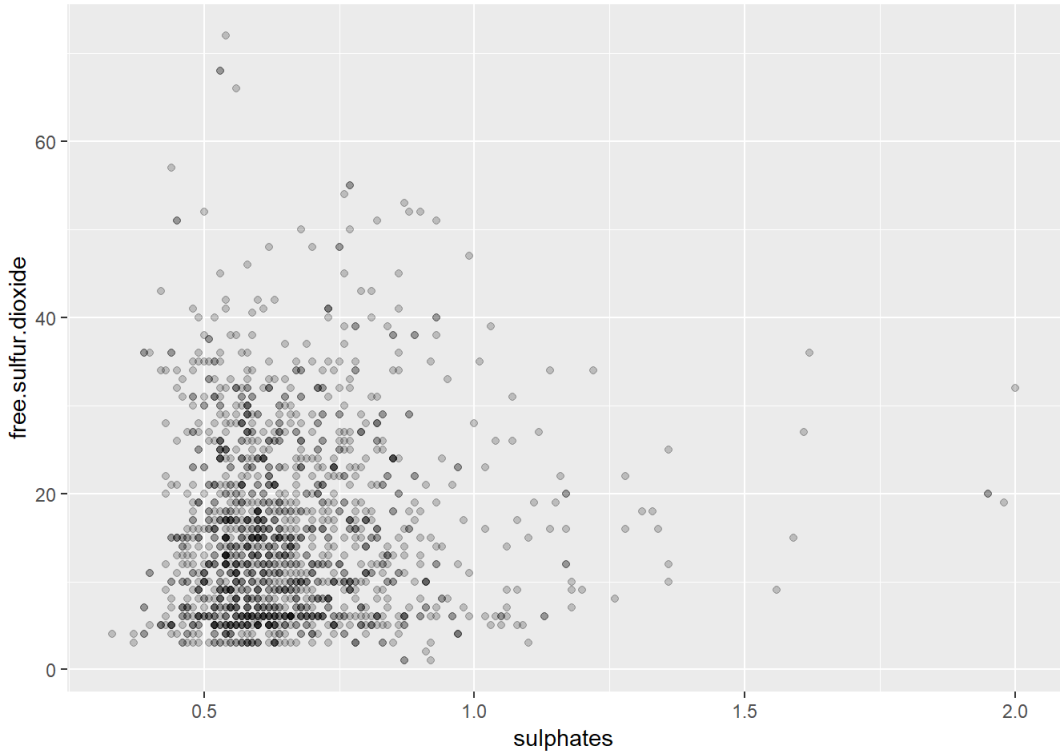

```
##
##  Pearson's product-moment correlation
##
## data:  residual.sugar and alcohol
## t = 1.6829, df = 1597, p-value = 0.09258
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.006960058  0.090909069
## sample estimates:
##       cor
## 0.04207544
```

*I was expecting a stronger correlation between the alcohol content and the residual sugar, since the alcohol comes from the fermentation of the sugars.

Maybe some of the wines are fortified with extra alcohol added that does not come from the fermentation of the sugar, or the yeast strains have different metabolic behaviors which do not allow to establish a linear relationship between sugar fermentation and alcohol production. Also, we don’t know which grape types were used, which may have different sugar contents.

Sulphates and sulfur oxide






```
##
##  Pearson's product-moment correlation
##
## data:  total.sulfur.dioxide and sulphates
## t = 1.7178, df = 1597, p-value = 0.08602
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.006087119  0.091774762
## sample estimates:
##       cor
## 0.04294684
```

```
##
##  Pearson's product-moment correlation
##
## data:  free.sulfur.dioxide and sulphates
## t = 2.0671, df = 1597, p-value = 0.03888
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.002643125 0.100424406
## sample estimates:
##       cor
## 0.05165757
```

*The relationship between sulphate levels and sulfur dioxide is very weak.

Correlations with quality

```
##                          [,1]
## X                    0.08392803
## fixed.acidity         0.11408367
## volatile.acidity     -0.38064651
## citric.acid           0.21348091
## residual.sugar        0.03204817
## chlorides            -0.18992234
## free.sulfur.dioxide  -0.05690065
## total.sulfur.dioxide -0.19673508
## density              -0.17707407
## pH                   -0.04367193
## sulphates             0.37706020
```

##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?

The wine quality is higher has stronger relationship with the volatile acidity, citric acid, sulphates and alcohol content. The correlation coefficients show us the strength of the relationship with the remaining variables.

```
##                          [,1]
## X                    0.08392803
## fixed.acidity         0.11408367
## volatile.acidity     -0.38064651
## citric.acid           0.21348091
## residual.sugar        0.03204817
## chlorides            -0.18992234
## free.sulfur.dioxide  -0.05690065
## total.sulfur.dioxide -0.19673508
## density              -0.17707407
## pH                   -0.04367193
## sulphates             0.37706020
```

*For the free and total sulfur dioxide we have seen in the plots that the medium quality levels (5 and 6) have both higher content than the low and higher quality levels. This may hint at some interaction with the other variables.

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

I observed the expected relation between the pH and acidity level.

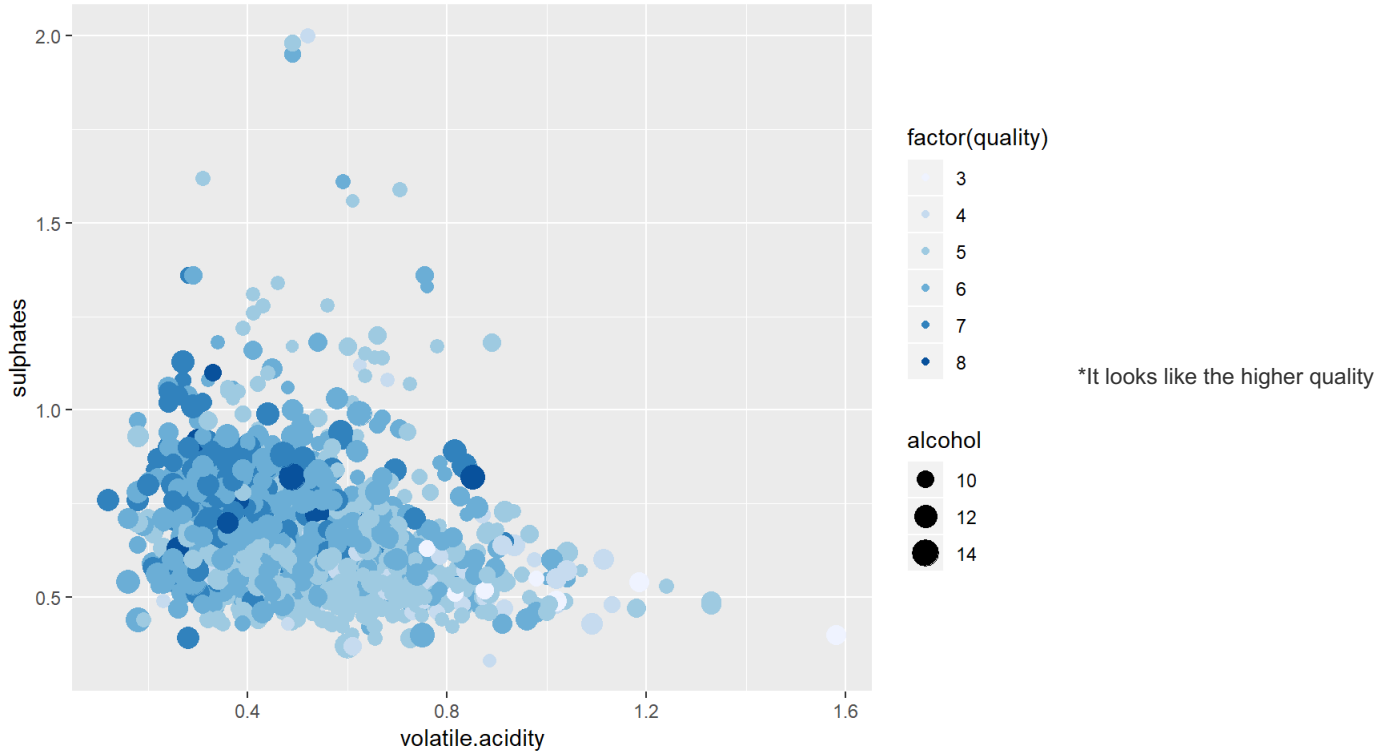
It was interesting to observe the relation between the density and the alcohol and sugar content.

I was surprised by not finding a stronger relation between the residual sugar and alcohol level, since the alcohol comes from the fermentation of sugars.

What was the strongest relationship you found?

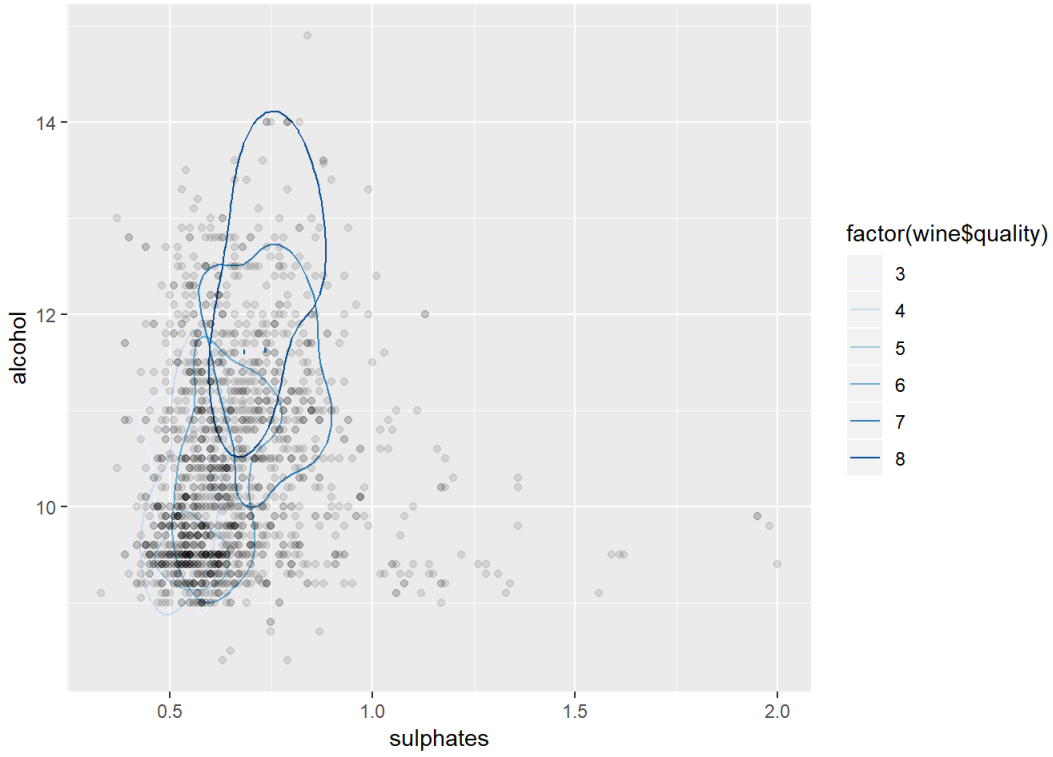
The correlation coefficients show that the variable with the strongest relationship with quality is the alcohol content.

Multivariate Plots Section



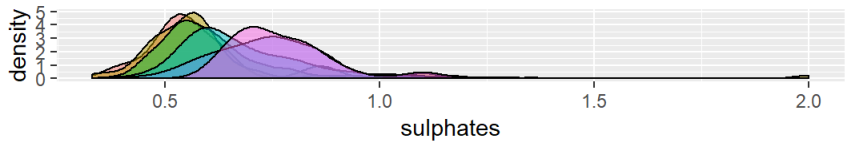
*It looks like the higher quality red wines tend to be concentrated in the top left of the plot. This tends to be where the higher alcohol content (larger dots) are concentrated as well.

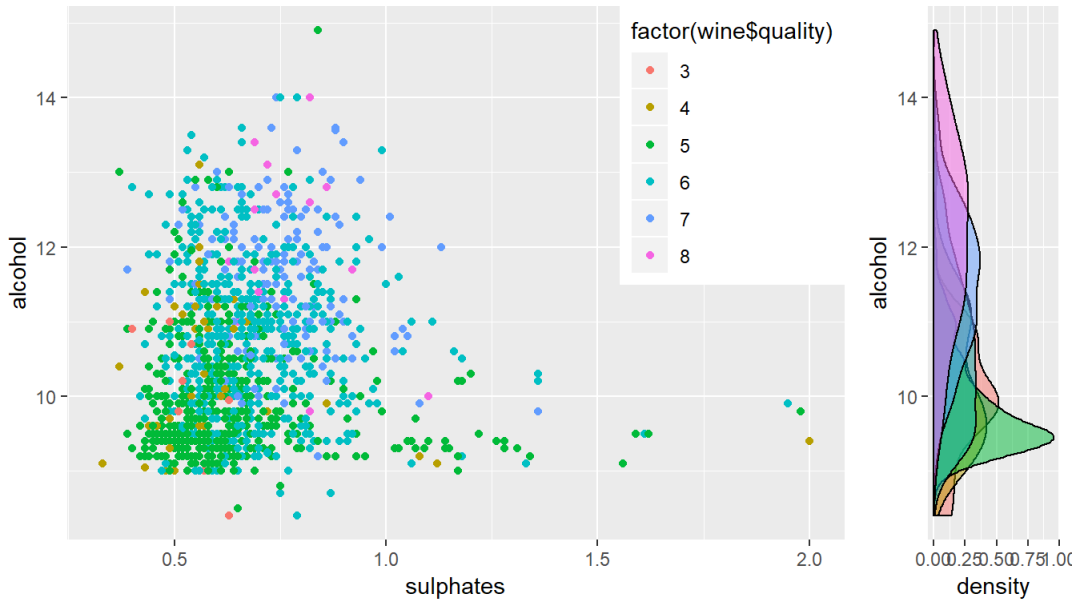
Let's try summarizing quality using a contour plot of alcohol and sulphate content:



*This shows that higher quality red wines are generally located near the upper right of the scatter plot (darker contour lines) wheras lower quality red wines are generally located in the bottom right.

Let's make a similar plot but this time quality will be visualized using density plots along the x and y axis and color :

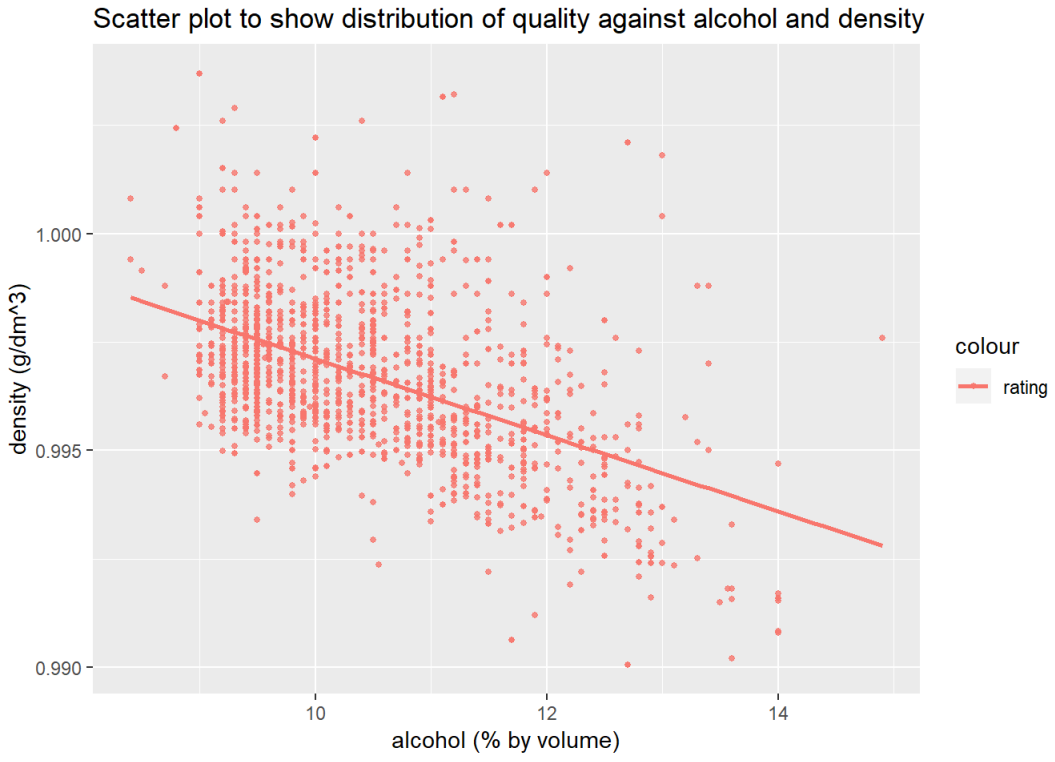




*Again, this clearly illustrates that higher quality wines are found near the top right of the plot.

Final Plots and Summary

Plot One



Description One

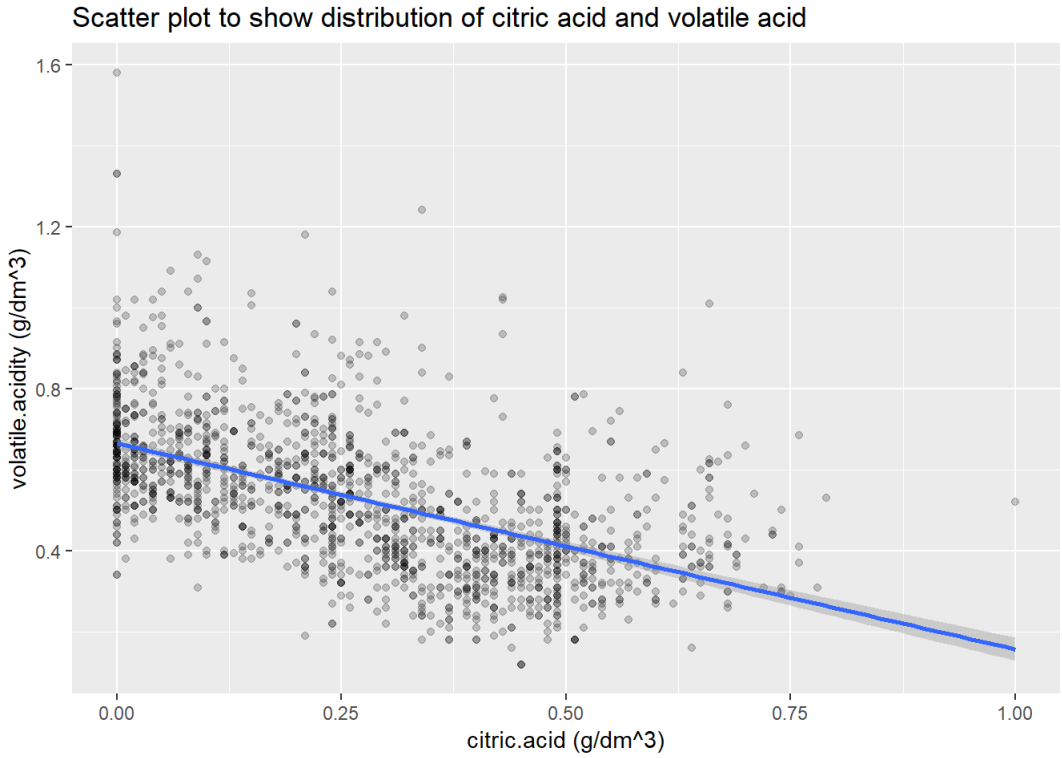
The most promising plot supporting the argument that we have started from the very beginning - "Good wines have high alcohol content and lesser density with medium pH.

###Plot Two

Description Two

We can see that the clustering of good wines is near a place where the density is low but the acidity is high. This supports our previous argument that good wines have high acidic content.

Plot Three



Description Three

The argument that winemakers uses citric acid to pull up the acidic content of the wine in cases where volatile acidity is low. So, if the natural acidity is in good proportions, artificial citric ones are not included.

Reflection

This project was an interesting opportunity to put all the knowledge of the R plotting functionality to explore a real dataset. The dataset was

put together for the purpose of applying machine learning techniques and was therefore already very well organized without any missing data. The only downside was the unbalancing of the classes: much more wines at the middle levels than at the low and high ends.

When working with such a dataset the first challenge is to choose in which direction to steer our exploration. Luckily, the dataset description file already hints at some variables of interest. For example, it tells us that citric acid can add freshness to wines, while acetic acid can add an unpleasant vinegar taste. This shows how important it is to have specific domain knowledge while performing a data analysis. Without it we are left adrift and will spend much time exploring in the wrong directions. When we do not have that domain knowledge, consulting with an expert in the field will be incredibly valuable to save us some time.

Another challenge I faced was interpreting the multivariate plots. When adding a third dimension - in this project, a color was mostly used - it becomes harder to grasp trends. There is no longer a nice line to guide our eyes, but instead the change of color should tell in which directions are our variables evolving. The use of a correlation matrix to find which variables have the biggest correlations helped to trim down the combinations to explore and made it easier to find interesting patterns.

As a follow up exercise, we could think of bringing the white wine dataset into this analysis and explore if the same trends that we found here apply on the different sort of wines.

Finally, having identified the main trends in the data, prediction models could be build to see how good this trends can be used to predict the wine quality based on the physicochemical attributes.