WineQuality by Rui Maranhao

Summary Statistics

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

```
##  [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"
```

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

```
##        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
```

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

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

```
##  Ord.factor w/ 6 levels "3"<"4"<"5"<"6"<..: 3 3 3 4 3 3 3 5 5 3 ...
```

```
## [1] "3" "4" "5" "6" "7" "8"
```

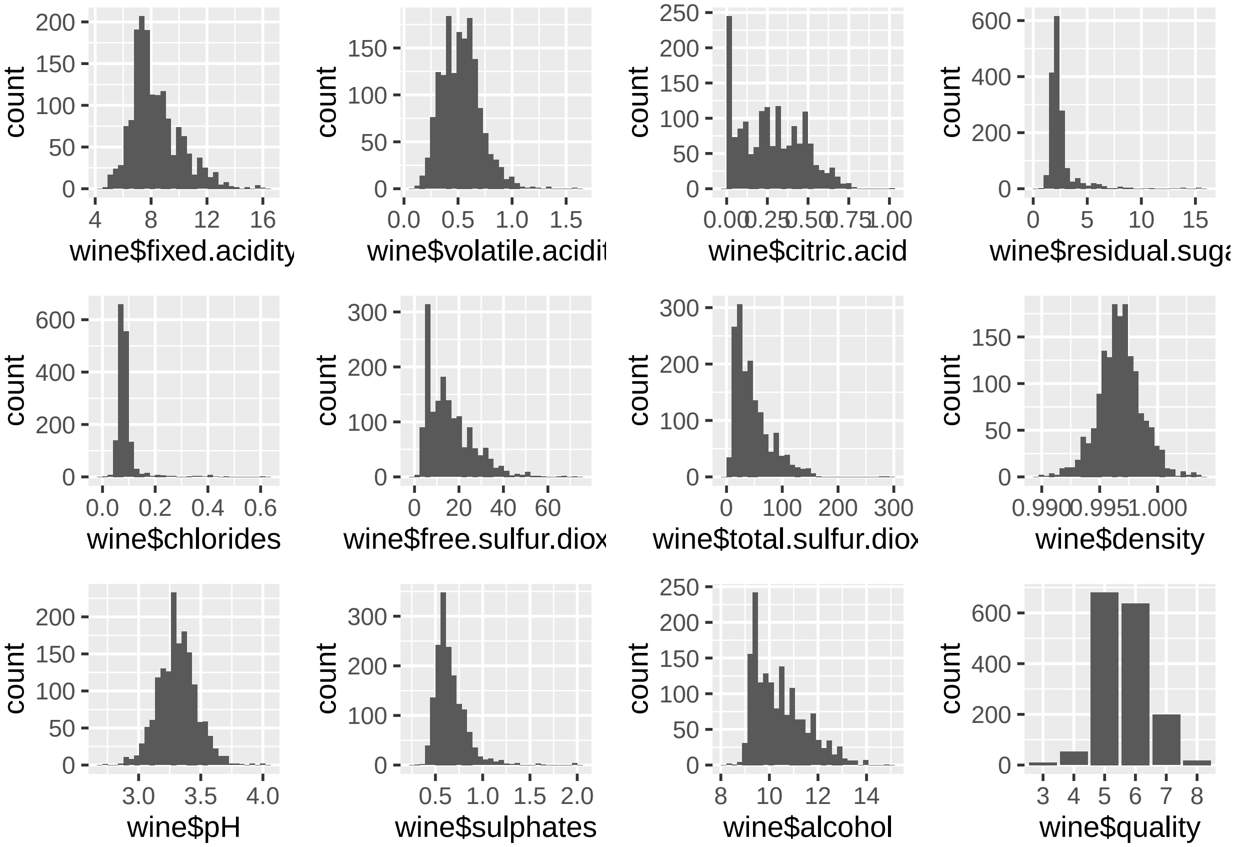
More information regarding wines [Cortez et al., 2009]. Input variables (based on physicochemical tests):

- fixed acidity (tartaric acid - g / dm^3)
- volatile acidity (acetic acid - g / dm^3)
- citric acid (g / dm^3)
- residual sugar (g / dm^3)
- chlorides (sodium chloride - g / dm^3
- free sulfur dioxide (mg / dm^3)
- total sulfur dioxide (mg / dm^3)
- density (g / cm^3)
- pH
- sulphates (potassium sulphate - g / dm^3)
- alcohol (% by volume)
- quality (score between 0 and 10)

Univariate Plots Section

Ploting all variables but X (which appears to be an unique identifier) to get the feeling about the distribution of the values.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



# Univariate Analysis

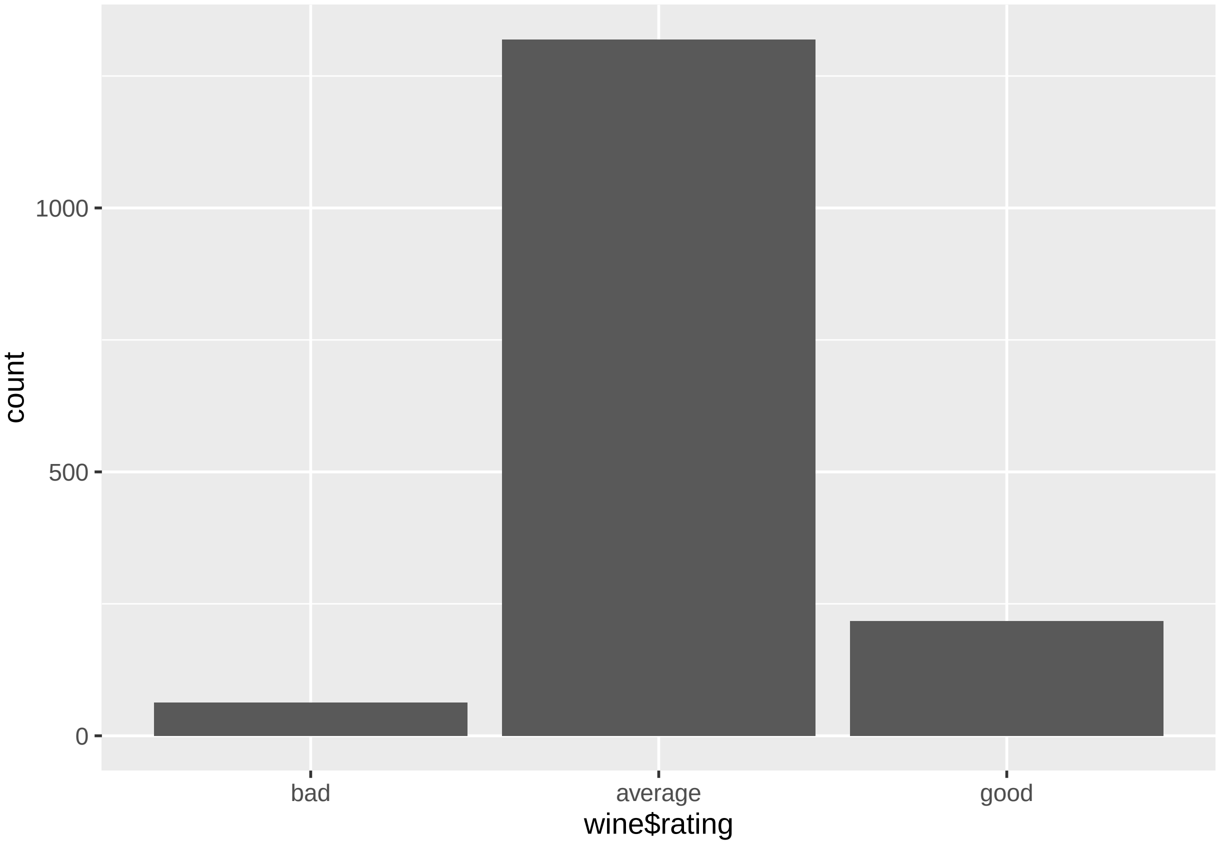
What is the structure of your dataset?

There are 1599 observations of 13 numeric variables.

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

The first feature of interest is the wine quality. The range of the wine quality is between 3 and 8, being 6.0 the median quality 6.0. A vast majority of the wines have a rating of either 5 or 6.

I have decided to instantiate a categorical variable rating wines qualitatively as bad, average, and good. One can see that most wines have been rated as average.



Second, I investigated the citric acid because I noticed that there are many observations that equal 0 in the dataset (132 observations to be precise). According to me this would require further investigation to find out whether this values were properly reported.

```
## [1] 132
```

Other variables of interest are pH and densitiy as they seem to be normally distributed. Alcohol and total/free sulphur dioxide look to have a long tail and being skewed towards 0.

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

I think further investigation wether pH could be classified into categorial (acid, base, neutral) could potentially be of interest. Also the relationship between fixed.acidity, volatile.acidity, and citric.acid (in particular this one due to the large number of 0s) could be interesting to further investigate in order to understand if the values are properly reported. This would require to understand the theoretical relationship between these variables.

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

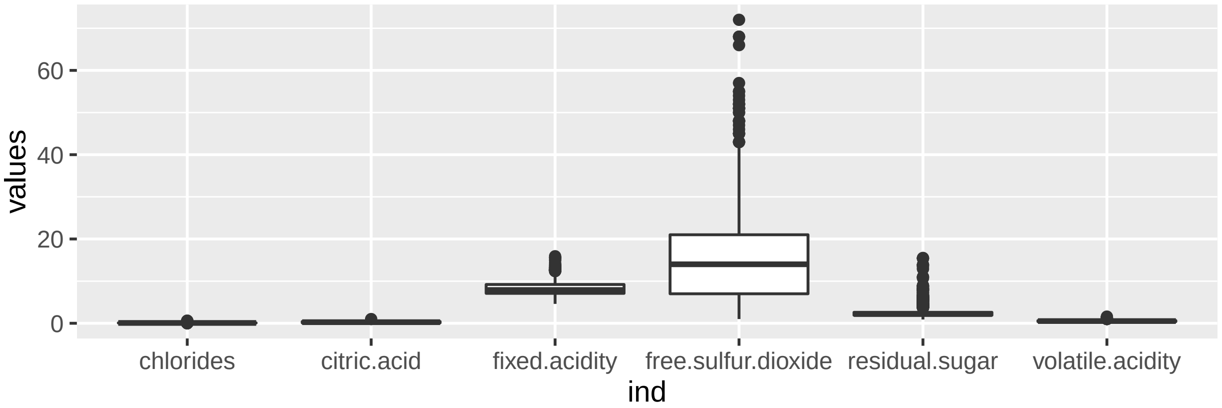
I only created the variable (ordered factor) rating to classify wine as good, bad or average. The information that the other variables store there does not seem to be any other variable that would fit well a sub-classification into categorical variables, perhaps with the exception of residual pH (neutral, base, acid).

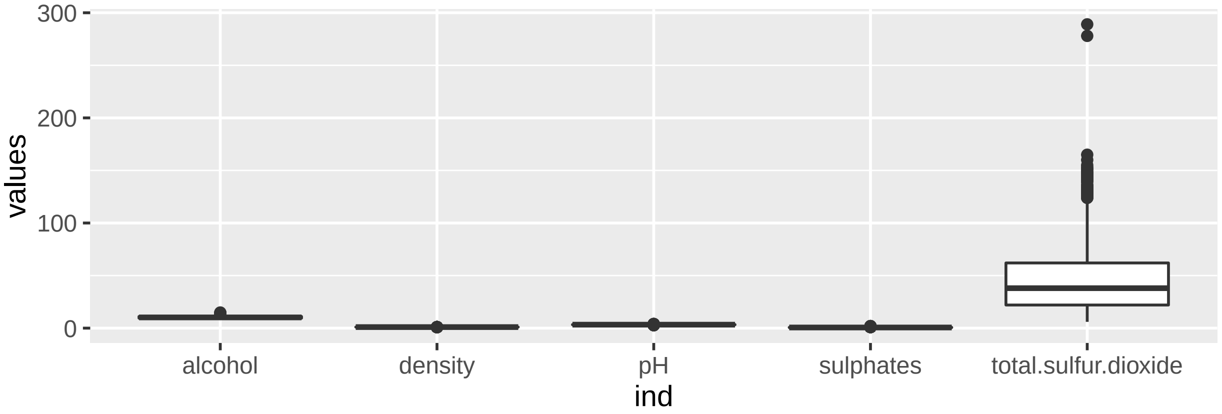
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

I've not found the need to tidy, adjust, or change the form of the data.

I've discussed the distributions in the section about the features of interest.

Below I use boxplots to gain a better understanding wrt. outliers (except for X for being an index, and quality for being categorical). I've also plotted the data using a log10 scale (I plotted for all variables, although only those with long tails are interesting for the log10 scale). These plots have shown that fixed acidity and to some extent pH, chlorides, densitiy, sulphates, volatile acidity to follow a normal distribution. As for the acidity variables, this is aligned with the fact that pH seems to be normally distributed, apart from the citric acid. The reason for the latter might be the number of 0 (potentially non-responses) discussed earlier. pH is normally distribution which suggests that the data is good, since by definition it is a measure of acidity and is on a logarithmic scale.

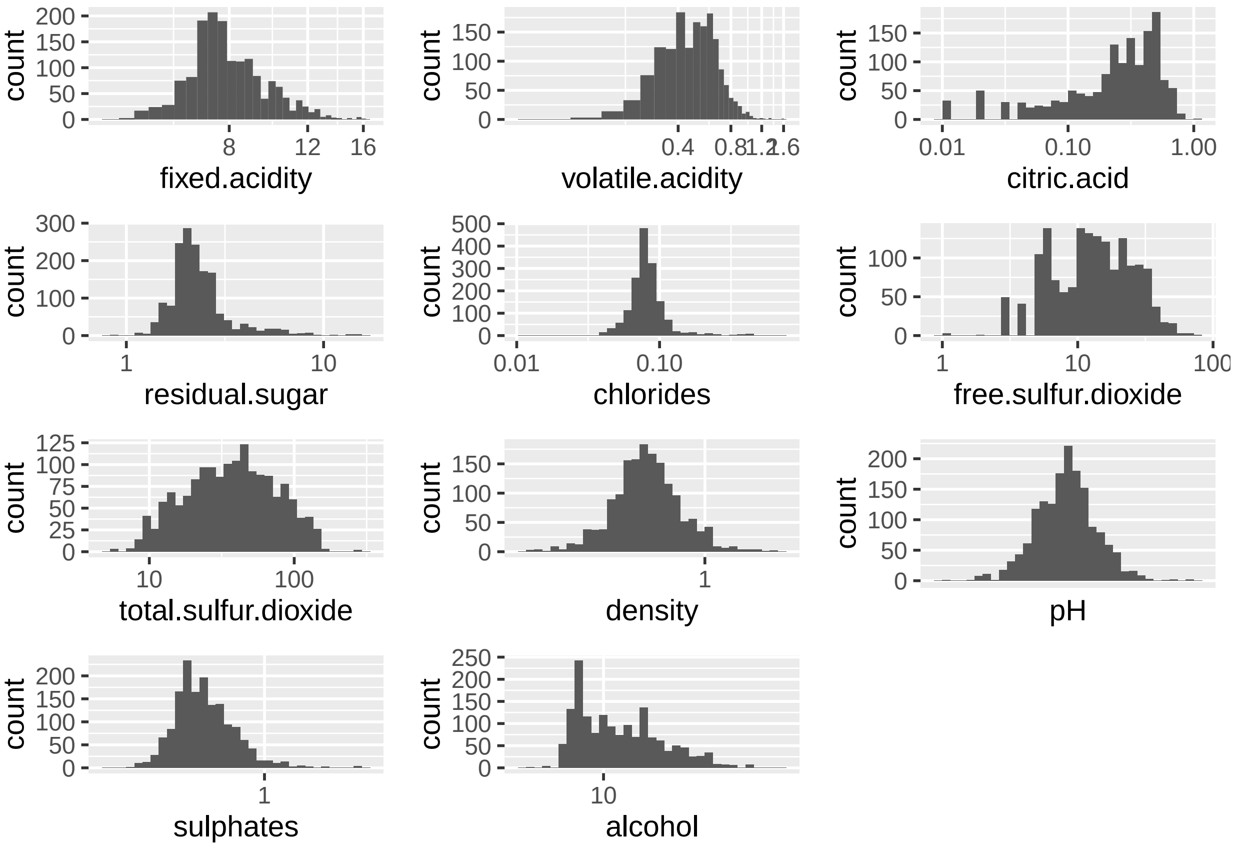


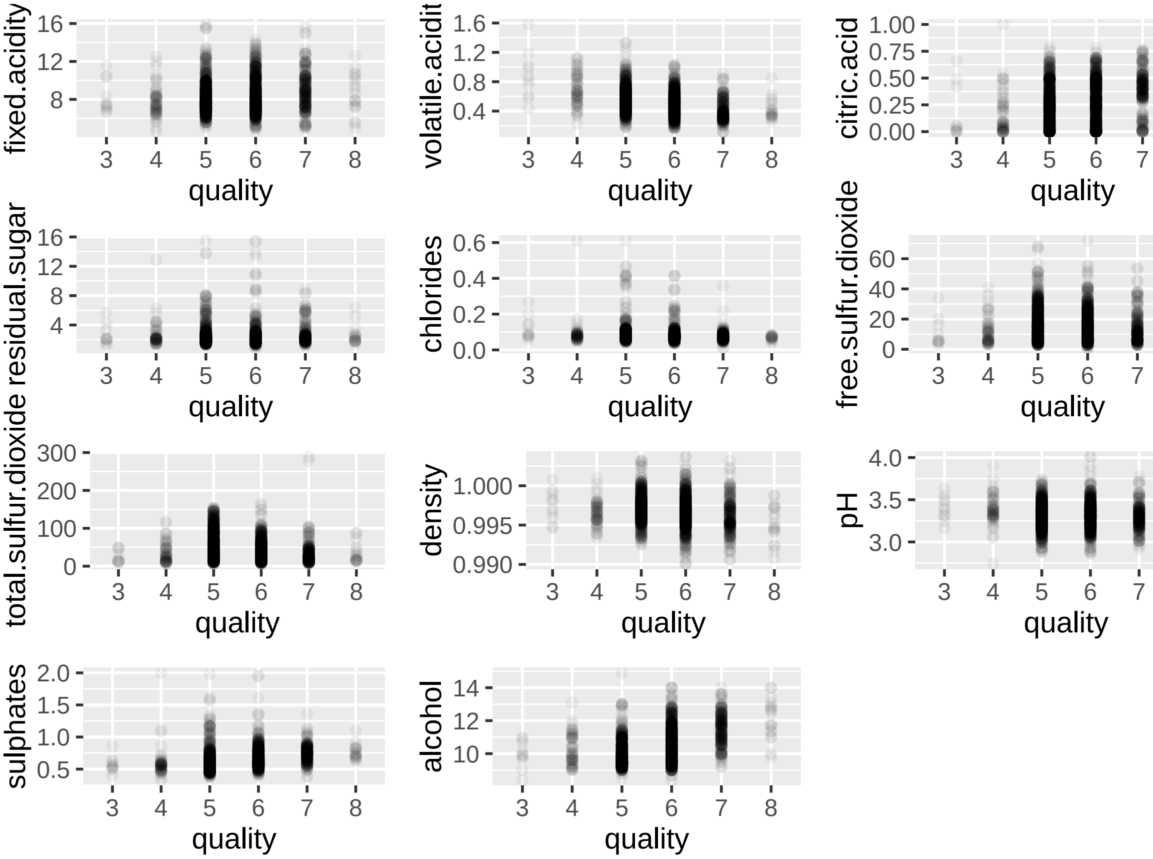


```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 132 rows containing non-finite values (stat_bin).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```





# Bivariate Plots Section
# 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?

It looks like the following characteristics yields 'good' wines:

- Acitiy: higher fixed acitity and citric acid; lower volative acitity;
- Lower pH (~3.5)
- Higher sulphtes
- Higher alcohol

It looks like the following yields 'average'/'good' wines:

- Sulfur dioxide: higher free and total sulfur dioxide
- Lower density

Residual sugar and chlorides did not seem to have an impact on the quality or rating of the wines.

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

I studied the relationship between the variables that correlate the most with quality: citric acid, sulphates, alcohol, and volatile acidity (see next question).

Appart from the higher the alchol the better the wine, the Pearson's test revealed that volatile acidity and citric acid have a strong negative correlation.

```
##
##  Pearson's product-moment correlation
##
## data:  citric.acid and alcohol
## t = 4.4188, df = 1597, p-value = 1.059e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.06121189 0.15807276
## sample estimates:
##       cor
## 0.1099032
```

```
##
##  Pearson's product-moment correlation
##
## data:  sulphates and alcohol
## t = 3.7568, df = 1597, p-value = 0.0001783
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.04477906 0.14196454
## sample estimates:
##        cor
## 0.09359475
```

```
##
##  Pearson's product-moment correlation
##
## data:  volatile.acidity and alcohol
## t = -8.2546, df = 1597, p-value = 3.155e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2488416 -0.1548020
## sample estimates:
##       cor
## -0.202288
```

```
##
##  Pearson's product-moment correlation
##
## data:  sulphates and citric.acid
## t = 13.159, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2678558 0.3563278
## sample estimates:
##     cor
## 0.31277
```

```
##
##  Pearson's product-moment correlation
##
## data:  sulphates and volatile.acidity
## t = -10.804, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3060917 -0.2147125
## sample estimates:
##        cor
## -0.2609867
```

```
##
##  Pearson's product-moment correlation
##
## data:  volatile.acidity and citric.acid
## t = -26.489, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5856550 -0.5174902
## sample estimates:
##        cor
## -0.5524957
```

What was the strongest relationship you found?

The quality shows the strongest correlation with alcohol (0.4761663). Note that there is also a negative strong correlation with volatile acidity (-0.3905578), and sulphates and alcohol show the weakest bi-variate relationship.

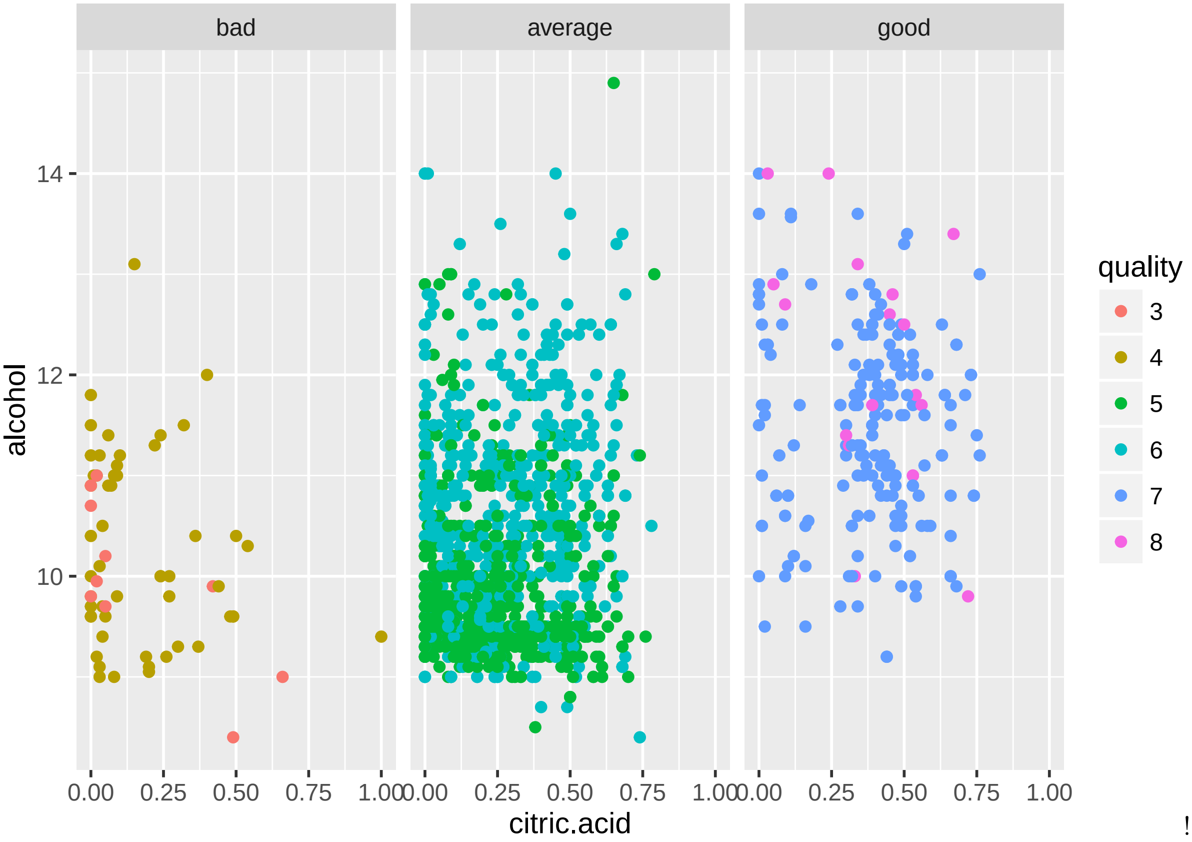
```
##
##  Pearson's product-moment correlation
##
## data:  fixed.acidity and as.numeric(quality)
## t = 4.996, df = 1597, p-value = 6.496e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.07548957 0.17202667
## sample estimates:
##       cor
## 0.1240516
```

```
##
##  Pearson's product-moment correlation
##
## data:  volatile.acidity and as.numeric(quality)
## t = -16.954, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4313210 -0.3482032
## sample estimates:
##        cor
## -0.3905578
```

```
##
##  Pearson's product-moment correlation
##
```

```
## data:  citric.acid and as.numeric(quality)
## t = 9.2875, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1793415 0.2723711
## sample estimates:
##       cor
## 0.2263725

##
##  Pearson's product-moment correlation
##
## data:  residual.sugar and as.numeric(quality)
## t = 0.5488, df = 1597, p-value = 0.5832
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.03531327  0.06271056
## sample estimates:
##        cor
## 0.01373164

##
##  Pearson's product-moment correlation
##
## data:  chlorides and as.numeric(quality)
## t = -5.1948, df = 1597, p-value = 2.313e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.17681041 -0.08039344
## sample estimates:
##        cor
## -0.1289066

##
##  Pearson's product-moment correlation
##
## data:  free.sulfur.dioxide and as.numeric(quality)
## t = -2.0269, df = 1597, p-value = 0.04283
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.099430290 -0.001638987
## sample estimates:
##         cor
## -0.05065606

##
##  Pearson's product-moment correlation
##
## data:  total.sulfur.dioxide and as.numeric(quality)
## t = -7.5271, df = 1597, p-value = 8.622e-14
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2320162 -0.1373252
## sample estimates:
##        cor
## -0.1851003

##
##  Pearson's product-moment correlation
##
## data:  density and as.numeric(quality)
## t = -7.0997, df = 1597, p-value = 1.875e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2220365 -0.1269870
## sample estimates:
##        cor
## -0.1749192

##
##  Pearson's product-moment correlation
##
## data:  pH and as.numeric(quality)
## t = -2.3109, df = 1597, p-value = 0.02096
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.106451268 -0.008734972
## sample estimates:
##         cor
## -0.05773139

##
##  Pearson's product-moment correlation
##
## data:  sulphates and as.numeric(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:  alcohol and as.numeric(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
```

Multivariate Plots Section



!!!!!!

# 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?

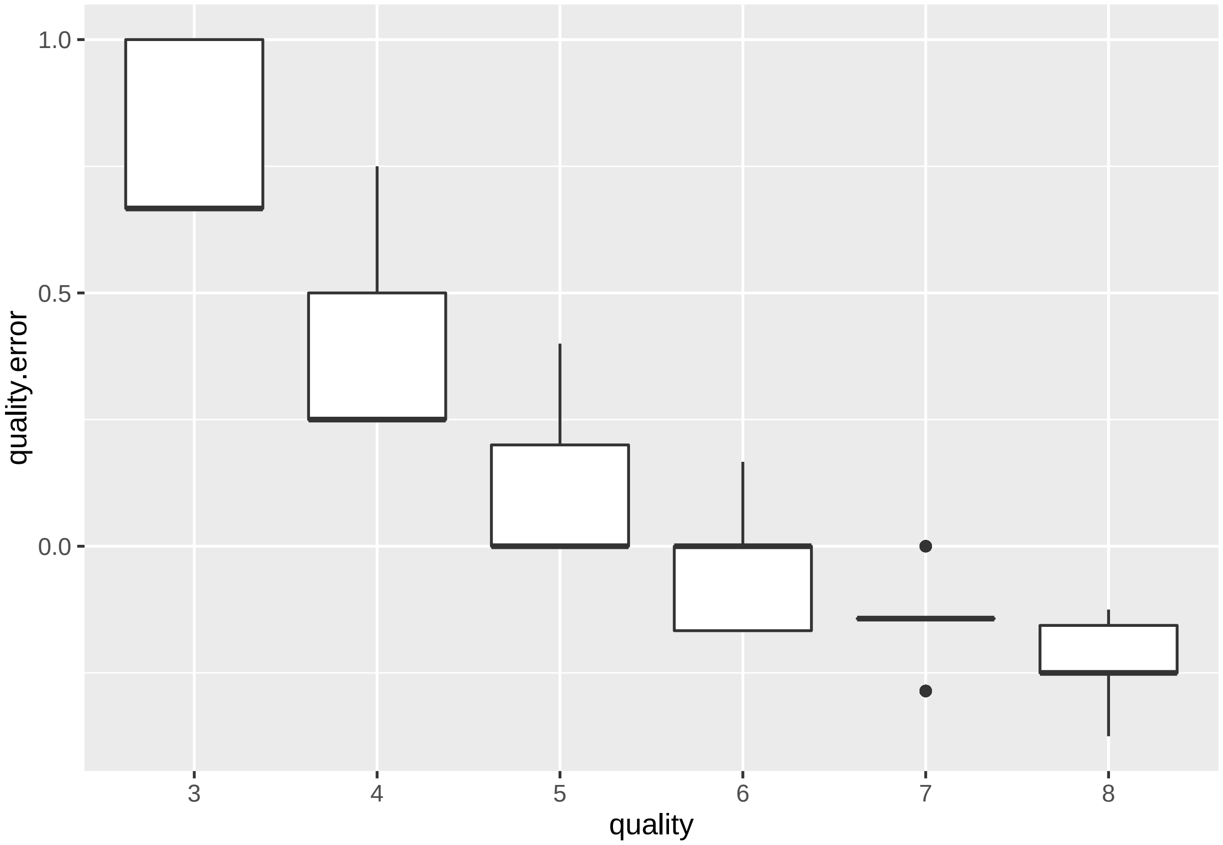
I focused on the 4 features that showed strong correlation with quality, plus pH. The plots show that a higher citric acid, higher sulphates, higher alchohol, and lower volatile acid are key factors to achieve a high quality wine.

Were there any interesting or surprising interactions between features?

pH does not seem to have a high impact on wine quality, despite the fact that the acids do play a role.

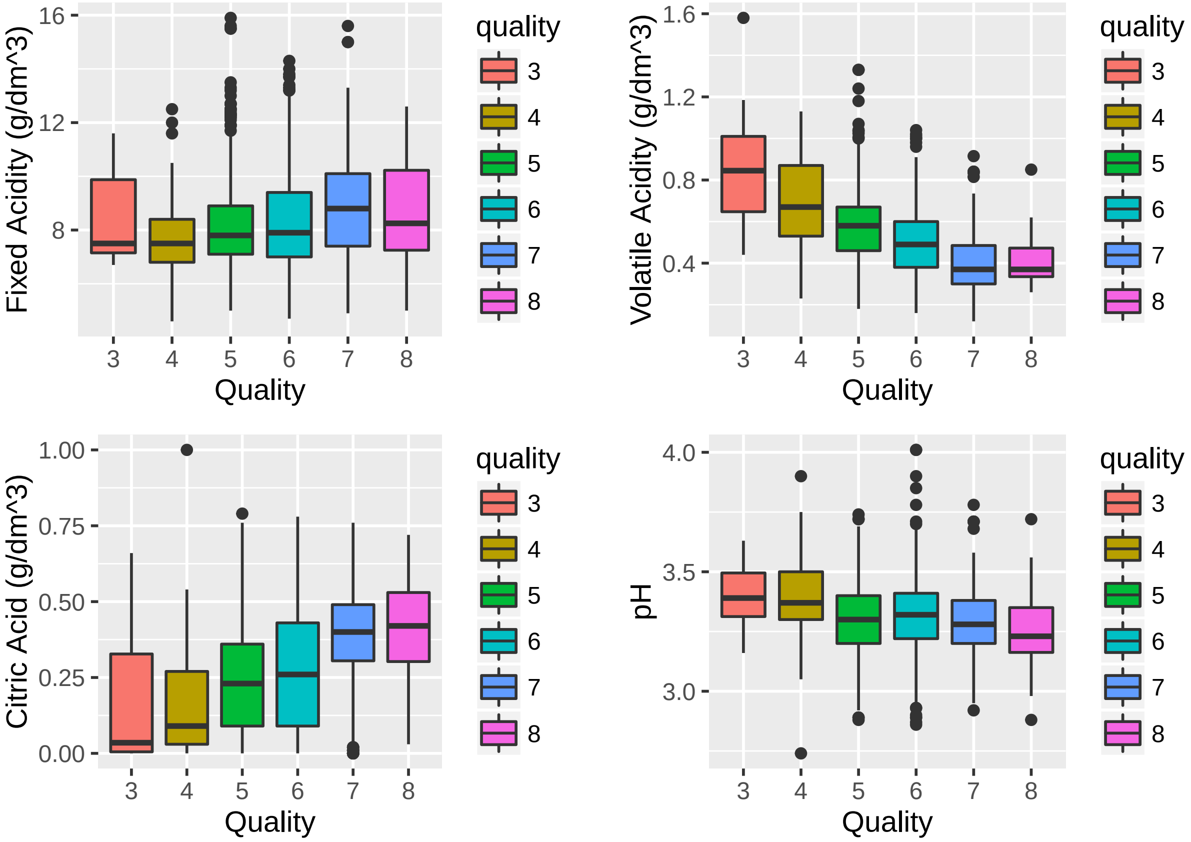
OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model.

I've created a model to predict the quality of wine given its alcohol. I did this because alcohol has revealed to have strong correlation with quality. The linear model didn't work well in practice. I think the main reason is because quality is a categorical value, hence not very suitable to linear models. It might be the case it is possible to predict pH from the acids.


1.0
0.5
0.0
quality.error
3
4
5
6
7
8
quality

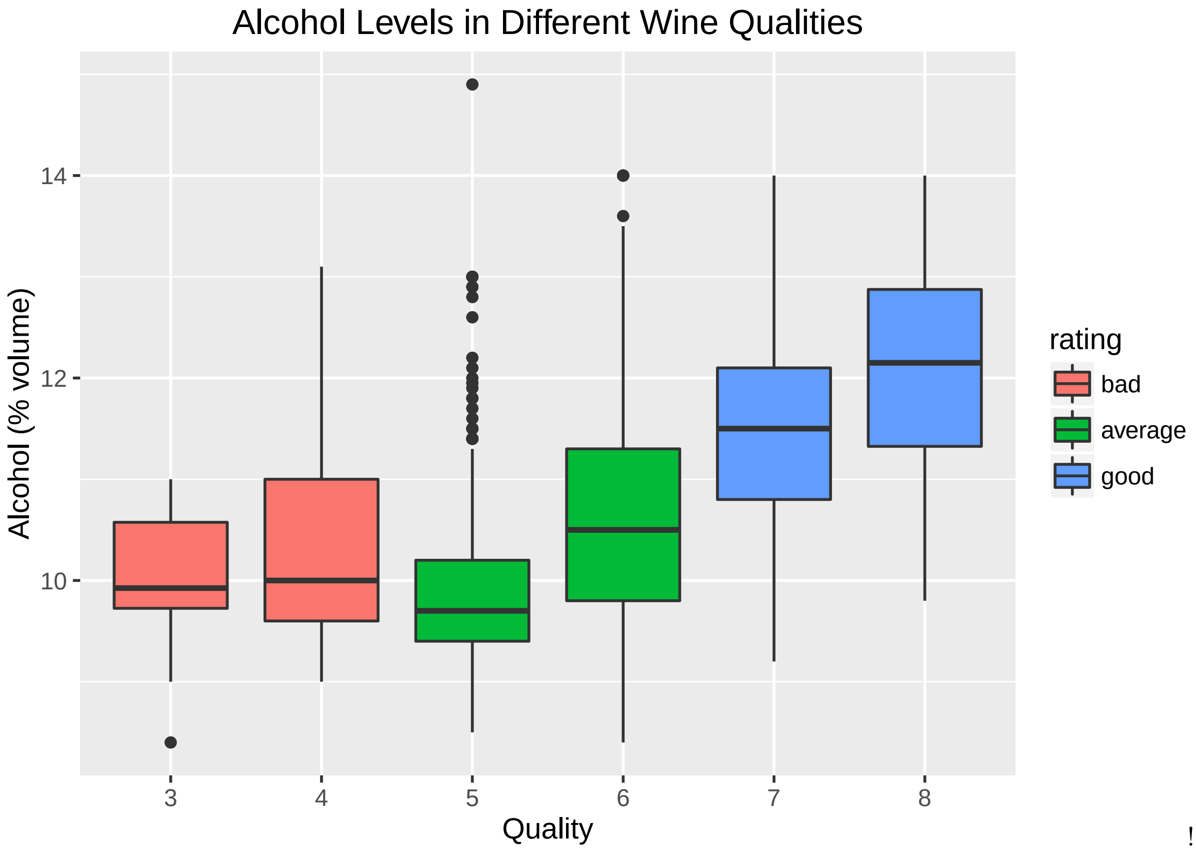
Final Plots and Summary

Plot One



### Description One These plots show the effect of acidity and pH on wine quality: higher acidity (hence, lower pH), apart from the volatile acid, is shown to yield better wines. The impact of fixed acid on quality is marginal.

Plot Two

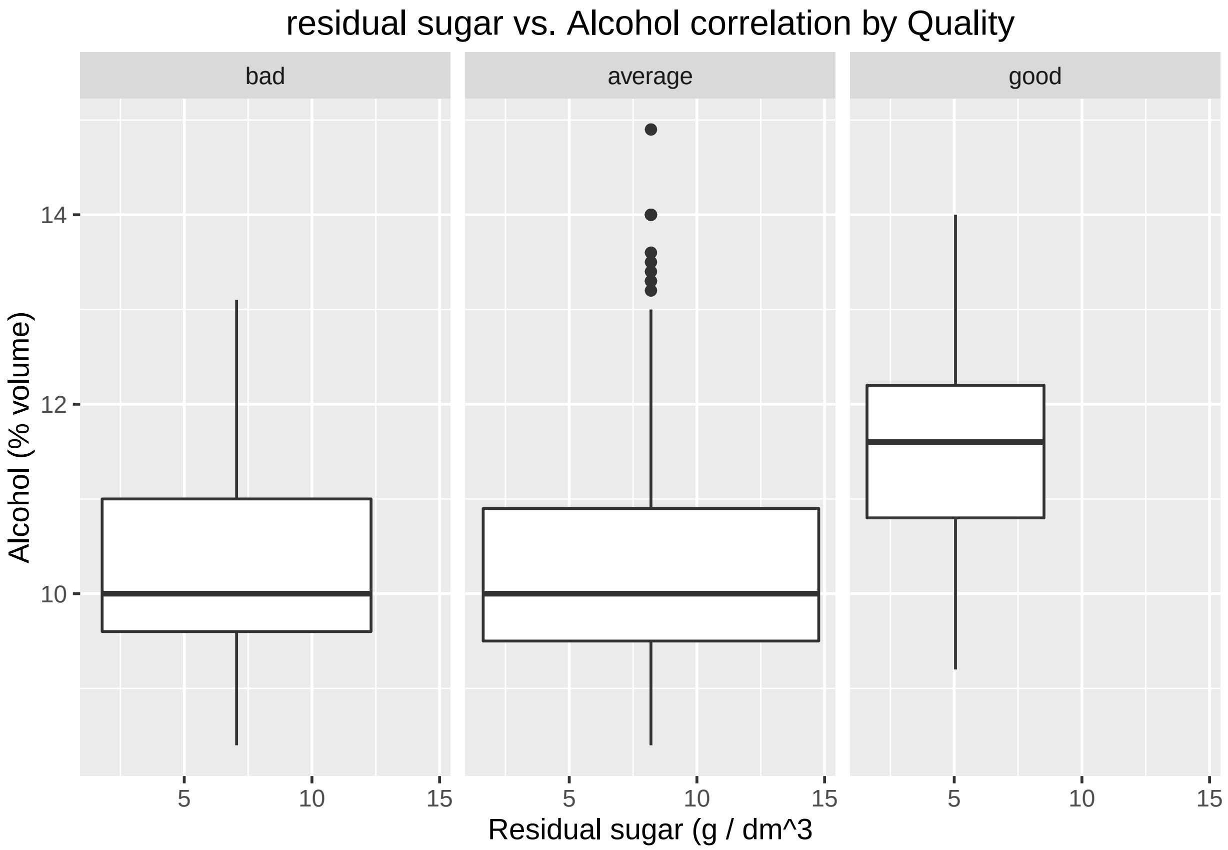


!

### Description Two These plots show the effect of alcohol on wine quality: higher alcohol, better wines. The same trend is observed of sulphates.

Plot Three

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```



### Description Three

This plot shows the the correlation betwwen quality and residual sugar vs. alcohol. The plot suggests that higher alcohol levels and lower residual sugars levels imply better red wines.

Reflection

The study about wine wines has revealed the following:

- Most wines in the dataset are rated as 'average', just a small number of wines are rated as 'bad'.
- There might be a problem with the citric acid, as there are many entries set to 0.
- Higher alchold positively impact the rating of the wines (i.e., better wines).
- More alcohol and low residual sugars yield to good red wines. [My intuition tells me that this might be different for white wines!]
- Residual sugar and chlorides did not seem to have an impact on the quality or rating of the wines.
- The lower volatile acidity the better the wine (string negative correlation)

Given that only alcohol correlated with quality of wine, this may suggest that rating a wine is not subject to the objectivity of the exports. There might also be the case that there are other factors which are not represented in the dataset (vintage, harvest year, location of the vineyards, temperatures before harvesting the wine, etc). Trying to obtain this information would be interesting for further exploration. Confirming this results using the white wine dataset woud also be of particular interest to me.

References

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.