# Red Wine Analysis by Markus Fischer

This project is part of the Data Analysis Nanodegree (P4: EDA). This document includes an exploratory analysis of the red wine quality dataset.

## Choice of Data Set

I selected the wineQualityReds dataset for the purpose of this analysis.

## Overview / Guiding Questions / Loading the data

This tidy data set contains 1,599 red wines with 12 variables on the chemical properties of the wine (excluding X, which is an index). At least 3 wine experts rated the quality of each wine, providing a rating between 0 (very bad) and 10 (very excellent).

## [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"

One of the question we will attempt to answer is "Which chemical properties influence the quality of red wines?"

# Univariate Analysis

### What is the structure of the dataset?

The dataset includes 12 variables (excluding the X variable which is just an index of each observation) and 1,599 observations.

The variable names are included below:

## '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 : Factor w/ 6 levels "3","4","5","6",..: 3 3 3 4 3 3 3 5 5 3 ...

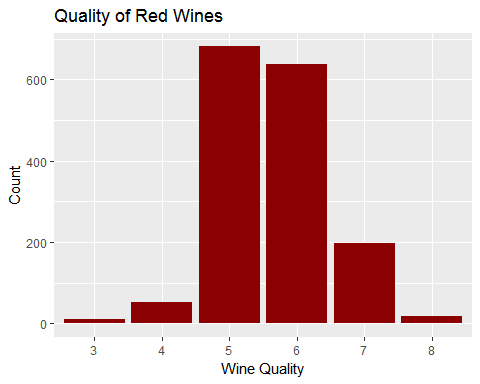
The structure of the data is as follows below. Quality of the wine is an integer from 0 to 10. In the process, we have transformed the quality variable into a factor (to help with plotting later on)

## 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 3: 10   
## 1st Qu.: 9.50 4: 53   
## Median :10.20 5:681   
## Mean :10.42 6:638   
## 3rd Qu.:11.10 7:199   
## Max. :14.90 8: 18

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

The main feature of interest in the dataset is the wine quality. We will analyze what factors impact wine quality and see whether any of those chemical factors have predictive power over the quality of the red wines.

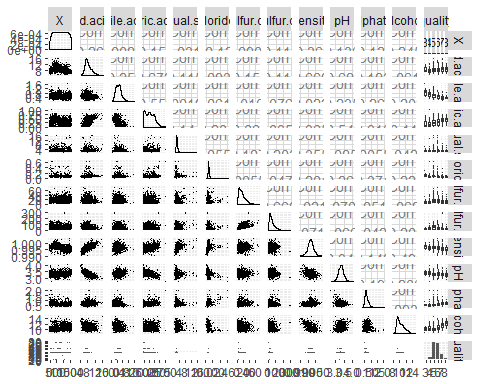
A couple of preliminary observations: 1. The sample consists of 1,599 red wines 2. The alcohol content varies from 8.40 to 14.90 3. The quality of the samples range from 3 to 8 with 6 being the median and 5.6 being the average 4. Fixed acidity, volatile acidity, residual sugar and dioxide variable all have large ranges 5. The range for density is relatively small ranging from 0.99 to 1.004 6. pH value varies from 2.740 to 4.010 with a median at 3.31

Let's plot the distribution of quality of wines in the dataset (which seems to broadly follow a normal distribution): 

The vast majority of wines have a medium quality of 5 or 6. The distribution seems to broadly follow a normal distribution.

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

In order to take a first look at all of the other features in the dataset, let's plot pairs of the various variables to examine whether any trends appear at first glance.

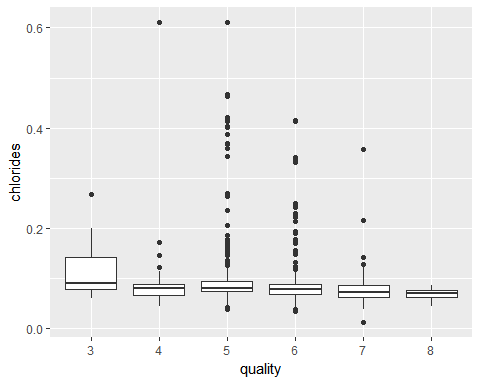


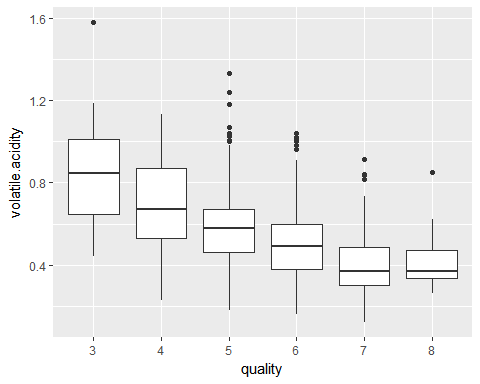
Looking at the boxplots on the right, a couple of variables seem to show patterns linked to wine quality. For instance, citric acidity seems to be present in higher quality wines and volatile acidity seems to have an inverse relationship with the quality of wines. Alcohol content seems also to play a role. Also, the scatter plots in the lower half seem to indicate that some variables have close to linear relationships. For instance fixed acidity and density seem to have a linear reltionship. Naturally, fixed acidity and pH also exhibit a relationship, which we would expect (as pH is an indicator of acidity).

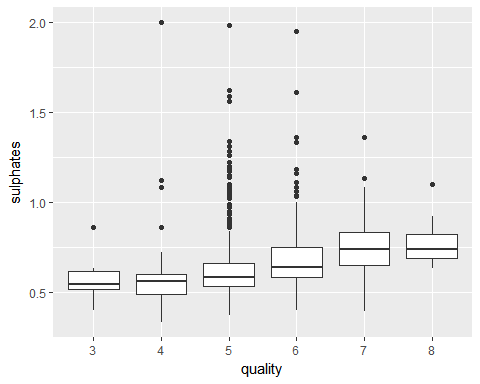
### Did you create any new variables from existing variables in the dataset?

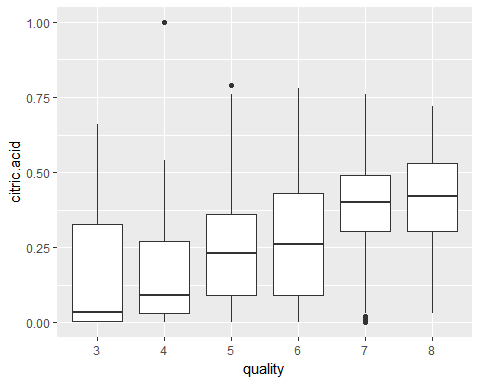
# Bivariate Plots Section

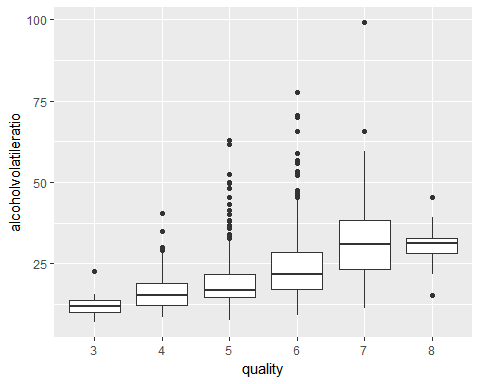
Lets start by plotting a couple of boxplots based on the analysis above Let's take one variable that doesn't seem to exhibit a strong relationship to quality:

 It is interesting to note that in general, chlorides don't seem to display a strong relationship - however very high quality wines (i.e. 8) all seem to have chloride levels below 0.2, which isn't the case for wines of quality 4 to 7.

A variable that does seem to have a strong relationship with quality is volatile acidity. Let's create a boxplot representing the relationship:  The higher quality wines seem broadly to have a lower volatile acidity than lower quality wines (although the dots do show some significant outliers to this rule). After some research, this makes a lot of sense: volatile acidity is typically present when the fermentation turns wine into vinegar - therefore high contents of volatile acidity might indicate a rotten / overly fermented wine.

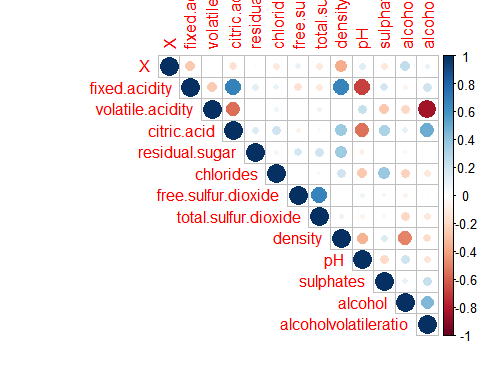
Let's also look at the sulphate content:  Sulfate content seems to be higher in better quality wines.

The same can be said of citric acidity:  Going back to volatile acidity, let's reflect upon the fermetation process of wines. The aim of the fermentation process is to ensure the wine develops the necessary alcohol levels. Nonetheless, the more the wine ferments, the higher the risk of volatile acidity and risk that the wine rots. Therefore, I would expect a high quality wine to have high alcohol content relative to the volatile acidity content.

Let's create a new variable representing the ratio of alcohol to volatile acidity and plot it against the quality of the wine:  The alcohol to volatile acidity ratio is higher in wines of higher quality - thus confirming our expectations.

# Bivariate Analysis

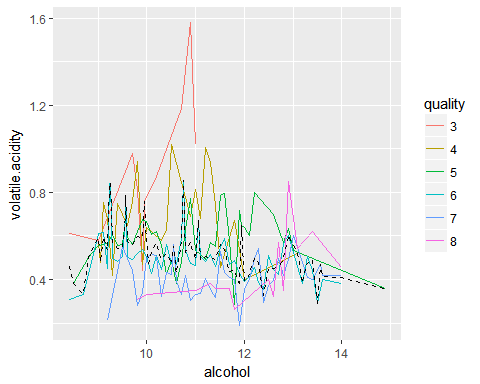
### Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

Let's examine the correlation between all of the variables: 

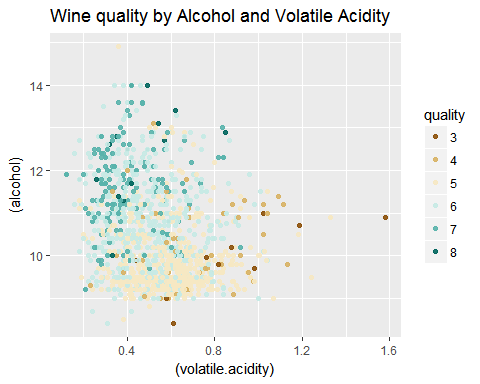
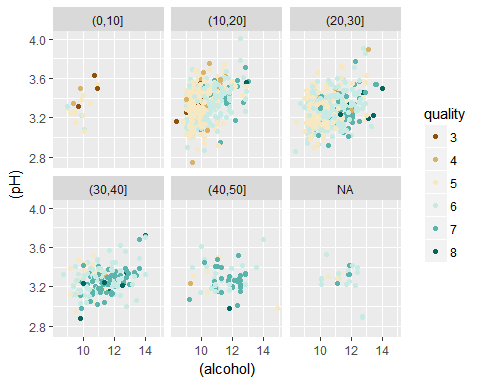
### Did you observe any interesting relationships between the other features (not the main feature(s) of interest)? What was the strongest relationship you found?

The strongest relationship we found was that between pH and fixed acidity - an almost perfect correlation. Or fixed acidity and sulfates. This makes perfect sense as the sulfates are part of the cause of the fixed acidity and pH is a measure of acidity. Citric acidity and fixed acidity seem to have somewhat of a relationship.

# Multivariate Plots Section

We identified a relationship between volatile acidity, alcohol and quality of the wine. As such, let's use a line plot to visualize this otherwise:  It is interesting to note tht at high volatile acidity and low alcohol levels, the lower quality wines appear. On the other end of the graph, where alcohol is high and volatile aciditiy is low, we see the 6 - 8 quality wines appear.

We can see this otherwise, using a scatterplot:

 Let's facet the scatterplot above by grouping the datapoints by several levels of alcohol / volatile.acidity levels:  Again, at levels of alcohol/volatile acidity ratio above ~25, we only see the higher quality wines, whereas at levels below, we seem ore of the lower quality wines.

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

We created a linear model to potentially predict the wine quality. The predictive power of the model isn't particularly strong, with R-squared levels of 0.3 or less.

The strength of the model is its simplicity, running a regression based on the variables. Unfortunately, the model has severl limitations: 1. the variable we are predicting is assumed to be a numerical continuous variable, when in fact it is a discrete value (i.e. takes an integer value from 3 to 8) 2. THe model assumes all variables are independent - which is clearly not the case (i.e. pH and fixed acidity are heavily correlated) 3. The model does not cater for more complex relationships between the variables

##   
## Calls:  
## m1: lm(formula = quality ~ alcoholvolatileratio, data = wines)  
## m2: lm(formula = quality ~ alcoholvolatileratio + pH, data = wines)  
## m3: lm(formula = quality ~ alcoholvolatileratio + pH + fixed.acidity,   
## data = wines)  
## m4: lm(formula = quality ~ alcoholvolatileratio + pH + fixed.acidity +   
## citric.acid, data = wines)  
## m5: lm(formula = quality ~ alcoholvolatileratio + pH + fixed.acidity +   
## citric.acid + residual.sugar, data = wines)  
## m6: lm(formula = quality ~ alcoholvolatileratio + pH + fixed.acidity +   
## citric.acid + residual.sugar + chlorides, data = wines)  
## m7: lm(formula = quality ~ alcoholvolatileratio + pH + fixed.acidity +   
## citric.acid + residual.sugar + chlorides + total.sulfur.dioxide,   
## data = wines)  
## m8: lm(formula = quality ~ alcoholvolatileratio + pH + fixed.acidity +   
## citric.acid + residual.sugar + chlorides + total.sulfur.dioxide +   
## density, data = wines)  
##   
## ==================================================================================================================  
## m1 m2 m3 m4 m5 m6 m7 m8   
## ------------------------------------------------------------------------------------------------------------------  
## (Intercept) 4.836\*\*\* 4.838\*\*\* 3.853\*\*\* 3.881\*\*\* 3.864\*\*\* 4.531\*\*\* 5.551\*\*\* 142.289\*\*\*   
## (0.045) (0.400) (0.620) (0.622) (0.622) (0.652) (0.666) (16.011)   
## alcoholvolatileratio 0.035\*\*\* 0.035\*\*\* 0.035\*\*\* 0.035\*\*\* 0.036\*\*\* 0.034\*\*\* 0.031\*\*\* 0.023\*\*\*   
## (0.002) (0.002) (0.002) (0.002) (0.002) (0.002) (0.002) (0.002)   
## pH -0.000 0.226 0.208 0.207 0.075 -0.110 0.410\*   
## (0.119) (0.161) (0.164) (0.164) (0.168) (0.169) (0.176)   
## fixed.acidity 0.030\* 0.035\* 0.034\* 0.022 -0.009 0.129\*\*\*   
## (0.014) (0.016) (0.016) (0.017) (0.017) (0.023)   
## citric.acid -0.085 -0.098 0.037 0.225 0.231   
## (0.147) (0.148) (0.153) (0.155) (0.151)   
## residual.sugar 0.009 0.010 0.026 0.072\*\*\*   
## (0.013) (0.013) (0.013) (0.014)   
## chlorides -1.396\*\*\* -1.587\*\*\* -0.798   
## (0.421) (0.418) (0.419)   
## total.sulfur.dioxide -0.004\*\*\* -0.003\*\*\*   
## (0.001) (0.001)   
## density -140.100\*\*\*   
## (16.392)   
## ------------------------------------------------------------------------------------------------------------------  
## R-squared 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.3   
## adj. R-squared 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.3   
## sigma 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7   
## F 385.3 192.5 130.1 97.6 78.1 67.4 64.3 68.0   
## p 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0   
## Log-likelihood -1753.9 -1753.9 -1751.7 -1751.5 -1751.3 -1745.8 -1727.4 -1691.5   
## Deviance 839.6 839.6 837.3 837.2 836.9 831.2 812.3 776.6   
## AIC 3513.7 3515.7 3513.4 3515.0 3516.5 3507.5 3472.8 3403.0   
## BIC 3529.8 3537.2 3540.3 3547.3 3554.2 3550.5 3521.2 3456.8   
## N 1599 1599 1599 1599 1599 1599 1599 1599   
## ==================================================================================================================

# Reflection

The red wine data set contains information on 1,599 winesacross 10+ variables. I started by exploring the individual variables in the data set. As I built an understanding of the variables and data, I explored more intricate relationships between variables

There was a clear relationship between alcohol level, volatile acidity and the quality of the wine. While I was able to identify broad relationships between those variables, it was very quickly clear that more goes into evaluating the quality of the wine - the individual variables we had were not sufficient in creating a powerful predictive model. Nonetheless, we identified interesting relationships that reflect the quality of the fermentation process and ultimately the quality of the wine.

Sources: <https://winemakermag.com/676-the-perils-of-volatile-acidity> <http://extension.psu.edu/food/enology/wine-production/wine-made-easy-fact-sheets/volatile-acidity-in-wine> <https://www.winebusiness.com/wbm/?go=getArticleSignIn&dataId=123076>