# Exploratory Analysis of Red Wine Quality by Jegan Kumar Ramanathan

I wouldn't call myself a wine connoisseur. But, I do like wines, especially, red wines. This dataset consists of measurements of several chemical properties of over 1500 red wines. In addition to these input variables, we also have the output variable named "quality" available for analysis. This output variable is based on sensory data and it's a median of at least 3 evaluations made by wine experts. Objective here is to explore this data through effective descriptive statistics and visualizations to understand the relationships that exist between these variables, which inturn will help us in developing a statistical model to predict the quality of red wines.

More information on this dataset can be found [here](https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityInfo.txt).

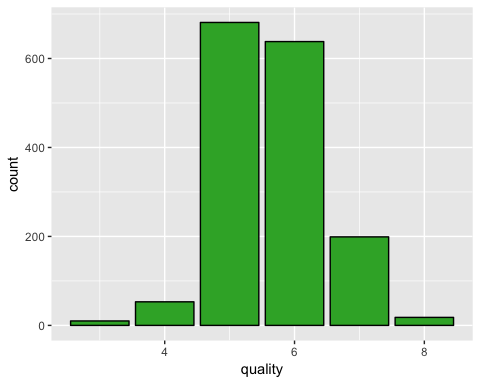
# Structure of the dataset

## 'data.frame': 1599 obs. of 12 variables:  
## $ 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 ...

Note: the row number column has been removed as it is not useful for this analysis

# Univariate analysis

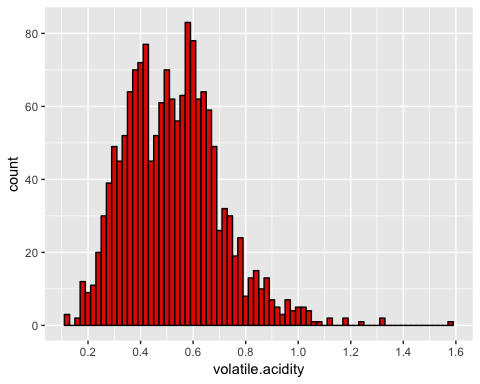
## Understanding the distribution of red wine quality



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

The quality variable seems to be more or less normally distributed. But, the values are concentrated around the centre i.e. lot of the red wines have been rated as a 5 or 6 on a scale of 10. So, there are a lot of decent or average red wines in this dataset according to wine experts.

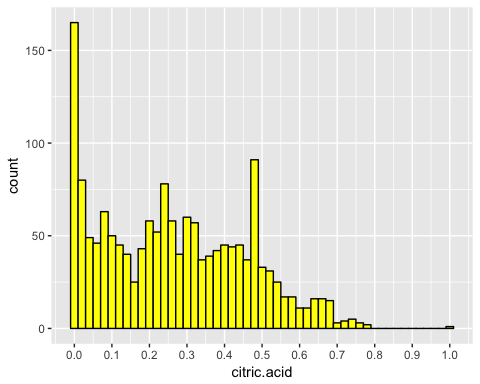
## Distribution of volatile acidity



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

Volatile acidity seems to be somewhat bi-modally distributed with a few outliers that have relatively high volatile acidity. High volatility in red wines results in an unpleasant, vinegar taste.

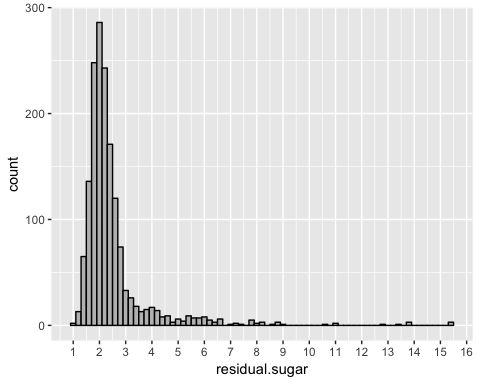
## Distribution of citric acid



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

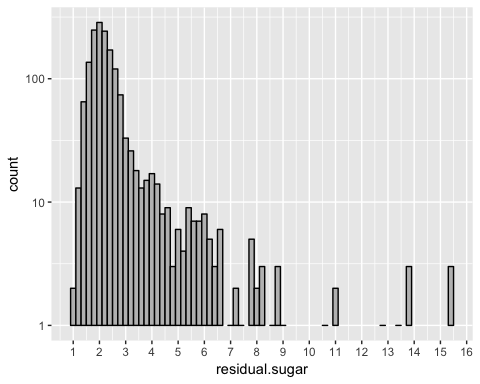
Citric acid distribution is relatively uniform with two peaks at 0 and then at around 0.5. It does taper off on or after 0.5. Citric acid adds 'freshness' and flavor to wines.

## Distribution of residual sugar



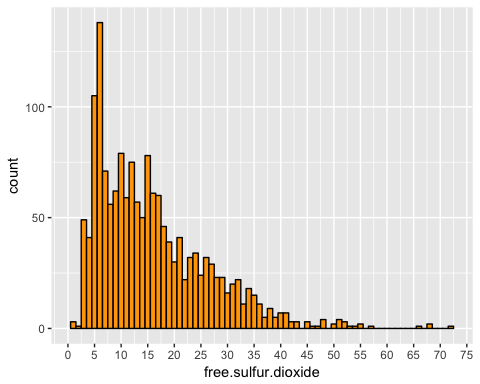
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.900 1.900 2.200 2.539 2.600 15.500

Residual sugar values are highly skewed and are right tailed. Residual sugar remaining after fermentation make the wines sweeter. Let's take a log tranform to get a different perspective on the distribution of this variable.



The distribution of the log10 transform of residual sugar further bolsters the observation that most of the red wines in this dataset have lower residual sugar content with a few exceptions.

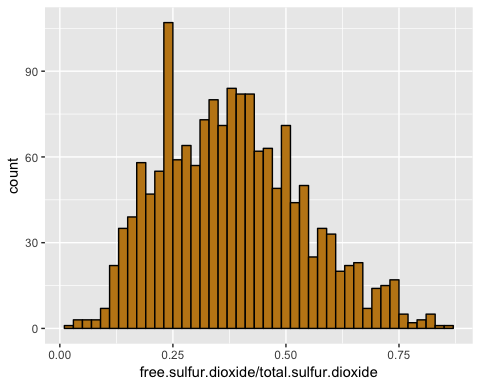
## Distribution of free sulfur dioxide



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

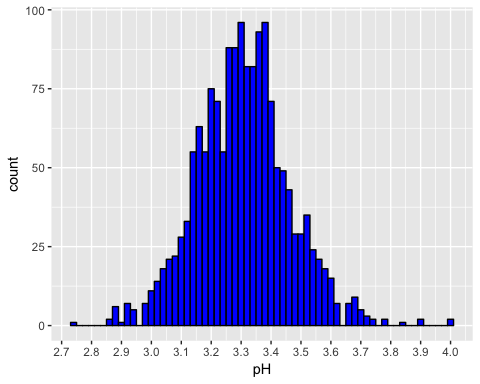
Free sulfur dioxide in high concentrations is increasingly noticeable through the smell and taste of wine. In this plot, we see a decreasing trend, as free sulfur dioxide increases, the count of red wines decreases. Also, there are a few outliers with relatively high amounts of free sulfur dioxide.

## Distribution of the ratio of free to total sulfur dioxide



Distribution of the ratio of free to total sulfur dioxide is normal but well distributed i.e. it has fatter tails.

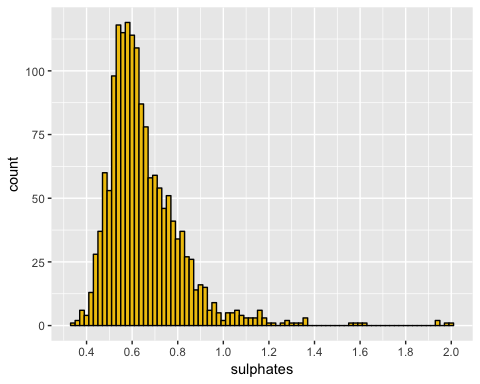
## Distribution of pH



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

pH values are normally distributed around the median/ mean of 3.31, indicating that most red wines are quite acidic.

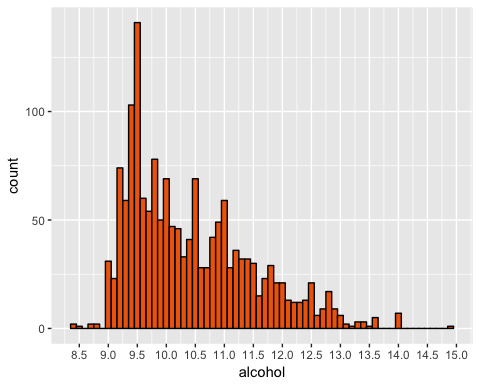
## Distribution of sulphates



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

Sulphates are additives used in red wines to increase sulfur dioxide gas levels which acts as an antimicrobial and antioxidant. This sulphate distribution is sort of skewed and normally distributed, with quite a few outliers.

## Distribution of alcohol

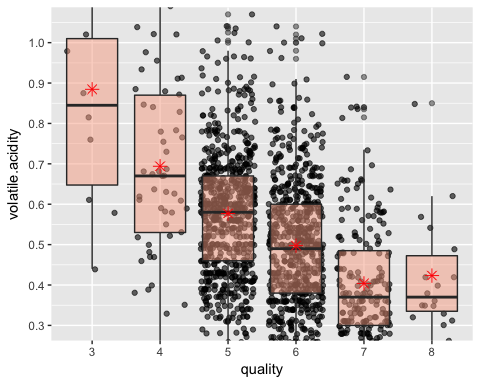


## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 8.40 9.50 10.20 10.42 11.10 14.90

Alcohol percentage by volume seems to mostly range between 9 - 13. But, as alcohol % increases the count of red wines reduces i.e. they have a negative relationship. Like other variables, there are some outliers here too.

# Bivariate analysis

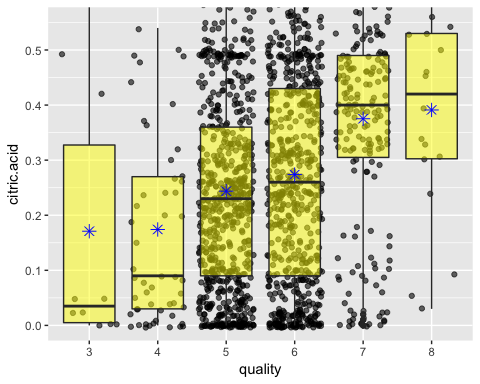
## Volatile acidity vs. Quality



As you can see clearly from the above box plots, as the quality increases the volatile acidity reduces. This makes sense as high volatile wines are considered unpleasant.

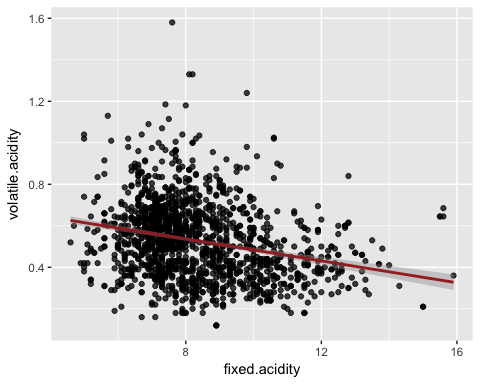
Note: the quality variable was changed from integer to factor to facilitate the generation of certain types of plots

## Citric acid vs. Quality



Unlike volatile acidity, as citric acid values increase so does the quality i.e. there seems to be a positive relationship between the amount of citric acid in red wines and its quality. This also is was expected as citric acids give 'freshness' and flavor to wines which I would think is appreciated by the experts.

## Volatile vs. fixed acidity

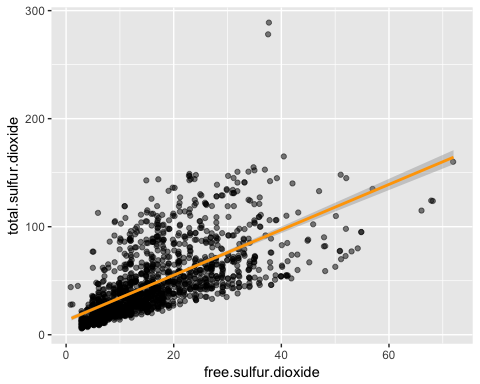


## [1] 0.2263725

As you can see from the scatter plot and correlation coefficient there is a negative relationship between volatile and fixed acidity, though the relationship is weak.

Note: the data was reloaded to change the quality variable data type to integer to facilitate generation of certain types of plots.

## Free vs. total sulfur dioxide



## [1] 0.6676665

It is obvious that there should be a relatively high correlation between free and total sulfur dioxide as they are dependent variables, total SO2 is a sum of free and bound SO2. The scatter plot and correlation coefficients just restate this obvious relationship.

## Computing correlations of all 11 input variables vs quality

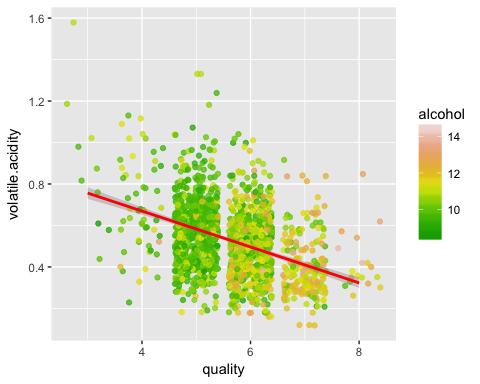
## Pearson's correlation coefficient of fixed.acidity and quality: 0.1240516   
## Pearson's correlation coefficient of volatile.acidity and quality: -0.3905578   
## Pearson's correlation coefficient of citric.acid and quality: 0.2263725   
## Pearson's correlation coefficient of residual.sugar and quality: 0.01373164   
## Pearson's correlation coefficient of chlorides and quality: -0.1289066   
## Pearson's correlation coefficient of free.sulfur.dioxide and quality: -0.05065606   
## Pearson's correlation coefficient of total.sulfur.dioxide and quality: -0.1851003   
## Pearson's correlation coefficient of density and quality: -0.1749192   
## Pearson's correlation coefficient of pH and quality: -0.05773139   
## Pearson's correlation coefficient of sulphates and quality: 0.2513971   
## Pearson's correlation coefficient of alcohol and quality: 0.4761663

There is a medium strong negative correlation between volatile acidity & quality, and there is a weak positive correlation between citric acid and quality. This observation is in agreement with what we noticed in the scatter/ box plots before.

There is also a medium strong positive correlation between alcohol % and quality. So, it is worth investigating this relationship a bit further. Sulphate content also seems to have a weak positive correlation with quality which can be analyzed further as well.

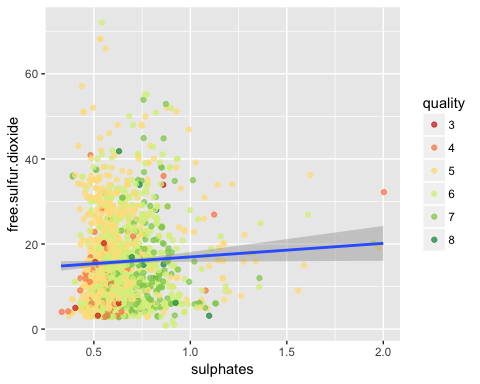
# Multivariate analysis

## Volatile acidity vs. Quality by Alcohol %



As you can see in the above plot, the alcohol content increases as the quality increases. This plot summarizes both the medium strong positive relationship of alcohol and quality, and the medium strong negative relationship of volatile acidity and quality.

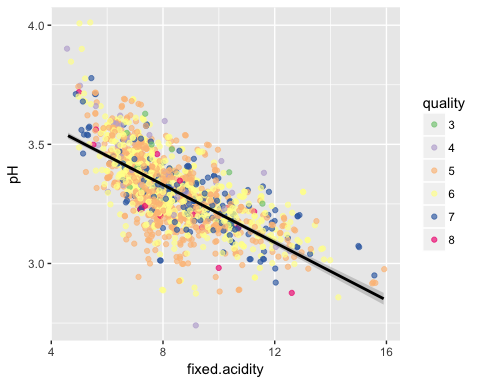
## Free sulfur dioxide vs sulphates colored by quality



As discussed before, sulphates are wine additives which increase SO2 gas levels. So, we would expect a correlation between free sulphur oxide (dissolved gas) and sulphates i.e. a positive relationship. But, in the above plot, there is a very weak positive correlation between these two variables. In addition to testing this relationship, this plot also shows how the quality of red wines vary with both sulphates and free sulfur dioxide. You could argue that as sulphates increase the quality increases (correlation coefficients calculated before suggest the same), but this is not the case with free sulfur dioxide. I guess what's happening here is that if the amount of free SO2 is low in certain wines, then just the right amount of sulphates are added to improve the wine.

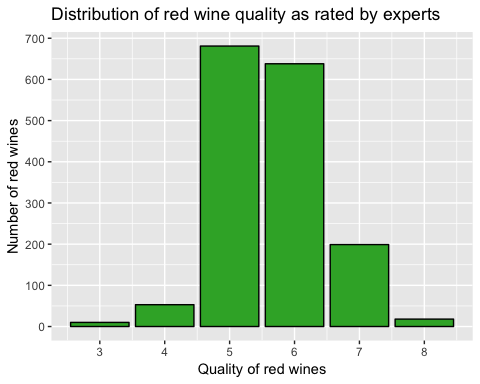
Note: the quality variable was changed from integer to factor to facilitate the generation of certain types of plots

## pH vs fixed acidity colored by quality



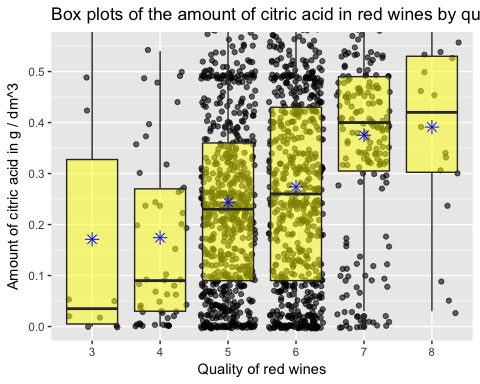
Low pH values indicate an acidic solution, and high pH values indicate a basic solution. So, there is a direct relationship between pH and fixed acidity. As shown in the plot above, when fixed acidity increases, pH decreases. I added in quality as a third variable represented by the color to see whether there are any noticeable relationships between quality and pH and/or fixed acidity. It doesn't look like there is.

# Final Plots and Summary



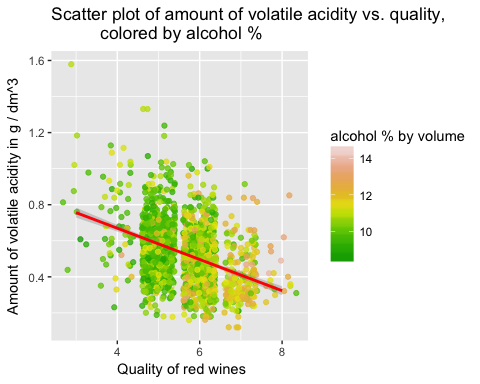
This plot is a histogram of the quality variable which is a median of 3 or more red wine expert ratings. Understanding the distribution of this variable is a critical first step for analyzing the dataset and then subsequently in developing a good model that can effectively predict the quality of red wines not part of this dataset.

Note: the data was reloaded to change the quality variable data type to integer to facilitate generation of certain types of plots.



These box plots clearly depict the positive relationship between citric acid and red wine quality. This could be used to develop a statistical model for red wine quality prediction.

Note: the quality variable was changed from integer to factor to facilitate the generation of certain types of plots



This colored scatter plot with straight line fitted through is an important visualization of how all three variables - volatile acidity, alcohol and quality are related to each other. I believe the understanding of these relationships will be crucial in developing a good model representative of red wine quality.

Note: the data was reloaded to change the quality variable data type to integer to facilitate generation of certain types of plots.

# Reflection

I explored all variables that I expected to have a reasonably strong relationship with red wine quality which would later help in developing a model for predicting the quality of other red wines not part of the dataset.

I did struggle to generate multivariate plots as there were no categorical variables and just one discrete variable. I also took some time to wrap my head around what would be a good multivariate relationship to explore.

Rest of the analysis was pretty smooth sailing. I was surprised by how alcohol was so strongly correlated with quality as I thought the amount of volatile acidity and citric acid were the key drivers of quality.

It would be good to develop a preliminary model for predicting wine quality with input variables as alcohol, volatile acidity, citric acid and sulphates, and test how the model performs.

It would also be useful to transform or create new variables from some of these input variables and analyze whether these transformations produce a better relationship.

Another area that can also be addressed are the outliers or anomalies found while analyzing these variables.