# Red Wine Analysis by Deepashree Jayaramu

Loading the already curated CSV file that was downloaded from the link: <https://docs.Google.com/document/d/e/2PACX-1vRmVtjQrgEPfE3VoiOrdeZ7vLPO_p3KRdb_o-z6E_YJ65tDOiXkwsDpLFKI3lUxbD6UlYtQHXvwiZKx/pub?embedded=true>

This tidy data set contains 1,599 red wines with 11 variables on the chemical properties of the wine. At least 3 wine experts rated the quality of each wine, providing a rating between 0 (very bad) and 10 (very excellent).

We will explore and analyse the data through various statistical visualizations using the R language. The goal of this analysis will be to discover which chemical properties influence the quality of red wines.

Let us first see the structure and summary of the data frame using some basic functions.

## '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

We see that each column represents each variable in the dataset. All the variables except for ‘X’ and ‘quality’ have numeric values. The variable ‘X’ is an unique identifier for the different wine samples where as the ‘quality’ variable has an integer value indicating the quality of the wine defined by 3 experts. These values are from ‘0’ as very bad to ‘10’ as very good. However, we have the values of quality ranging from 3 to 8 only.

## 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

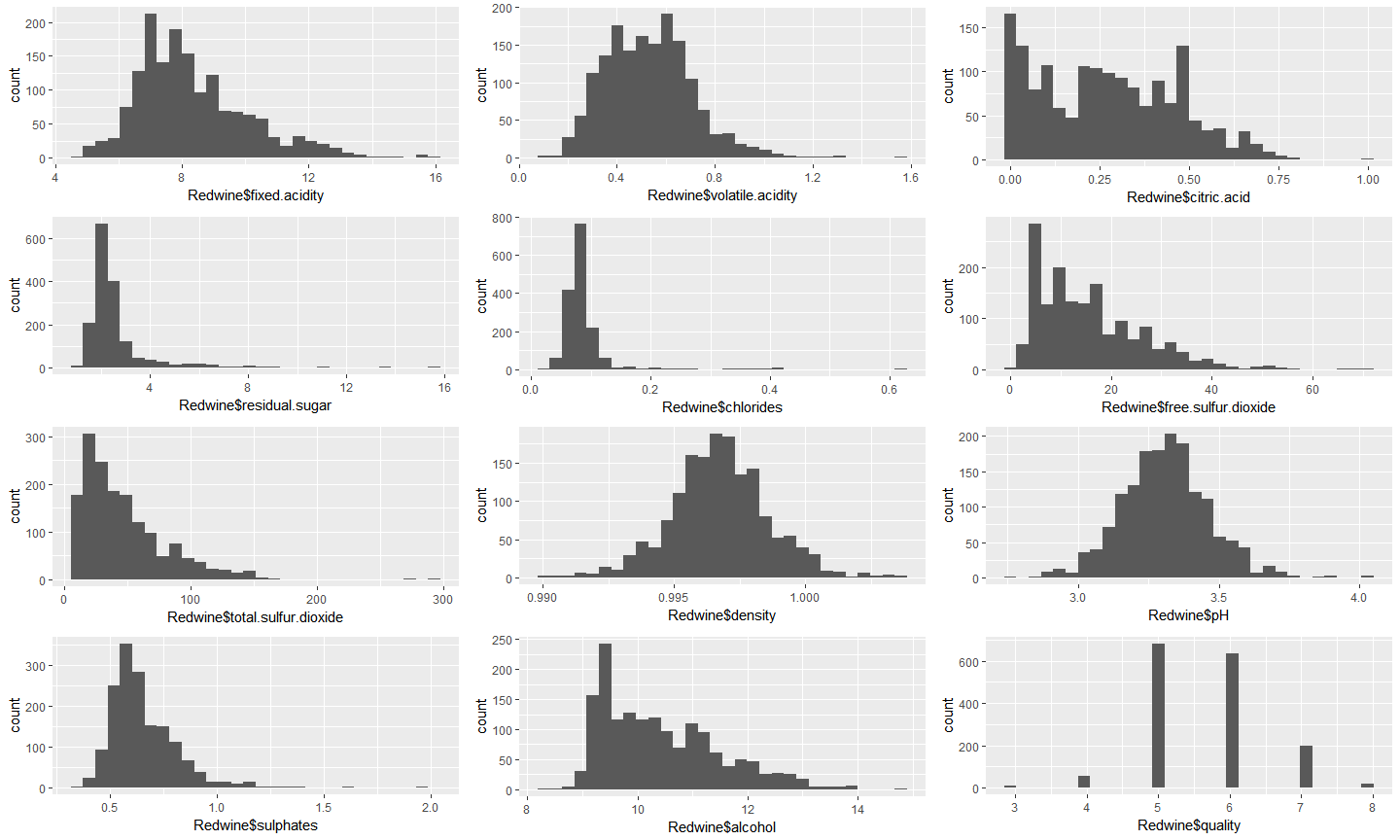
When we look at the summary of the quality in ‘Redwine’ data samples, we see that the quality is an ordered, categorical, discrete variable. The values ranged only from 3 to 8, with a mean of 5.6 and median of 6.

The table of the Quality variable indicates that, 3 4 5 6 7 8 10 53 681 638 199 18

there are very few of the very bad and the very good quality wines. Most of the wine samples in this data are rated 5 or 6 making the majority an average quality :)

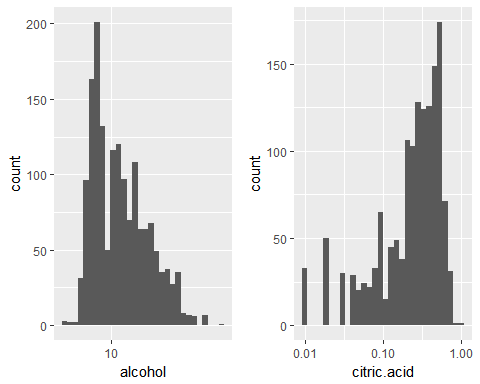
# Univariate Plots Section

In order to visualize the data and get a hang of the plots, let us draw the values of all the variables against the number of samples in this data frame into histogram plots arranging them all in one grid.



By observing the above plots, quality, density, pH, fixed & volatile acidity seems to be normally distributed. The total sulfur dioxide, and free sulfur dioxide are right skewed.

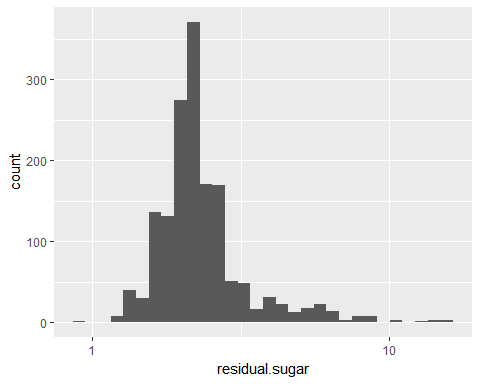
Let us plot the ‘alcohol’ and ‘citric acid’ values on log 10 scales to know the nature of their distribution.



From the above distribution we find that the ‘0’ values for alcohol that we see for 132 samples of wine were not reported.

One variable that I am curious is the residual sugar in the wine that can determine the wine to be a sweet wine. Let us check how many classify into sweet wines in our dataset.

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



Well, from the information on the Wikipedia, <https://en.Wikipedia.org/wiki/Sweetness_of_wine#Residual_sugar>, a wine’s sweetness is measured by the left over grams of sugar (not fermented sugar) per liter of wine. The link above indicates that it is rare to fine that the values less than 1 gm/ Lt in the driest of wines. On the contrary, a residual value of 45gm/Lt or more qualifies for the wine to be a sweet wine. So, clearly on running the summary function on the residual.sugar variable, we do not have any sweet wines in our dataset. The maximum value of residual sugar in the dataset is of 15.5

# Univariate Analysis

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

Quality

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

1. Residual sugar values can classify the wine into sweet or dry wines. Also, we will have to see if residual sugar has any influence on the quality of the wine.
2. To check which variables contribute to determine the quality of the wine.

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

There are not any new variables created at this point. I will have to create a variable ‘quality.factor’ by using the factor of it in order to make it a categorical variable.

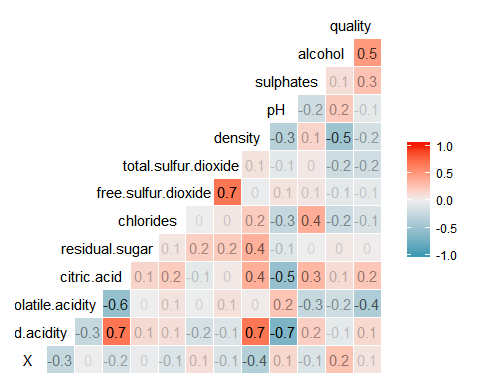
### Of the features you investigated, were there any unusual distributions? you perform any operations on the data to tidy, adjust, or change the form the data? If so, why did you do this?

In order to get the better picture of the distribution, the x-axes of the residual sugar was log transformed.

# Bivariate Plots Section

Let us begin with finding the correlation coefficient factors of the main feature of our interest, ‘quality’ with other features one by one.

The following plot gives us a visualization of the correlation table between each feature of the wine dataset.



We see that the following features are positively and highly correlated with Quality.

1)Alcohol  
2)Sulphates  
3)Citric Acid

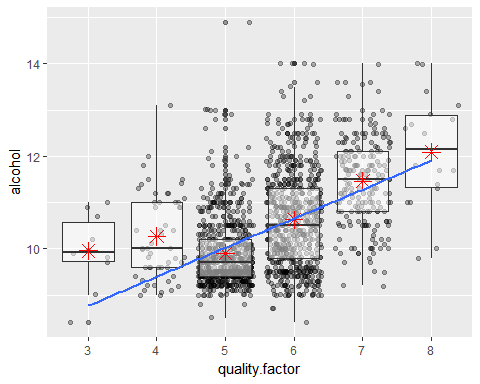
We also find out that the correlation between the ‘residual sugar’ and ‘quality’ is neutral. So, the residual sugar actually does not influence the quality of the wine as we thought earlier in the Univariate Analysis.

The following features are highly but negatively correlated with the Quality.

1)Density  
2)Volatile Acidity

Now let us visualize these correlations.

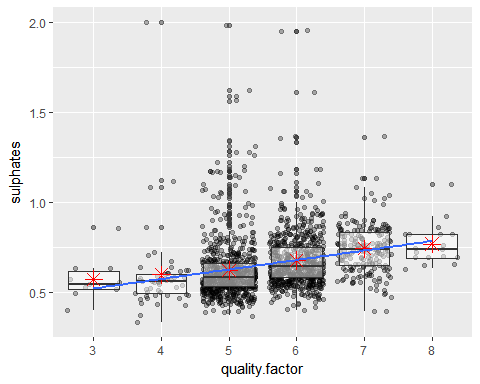
### Alcohol Vs Quality



The Alcohol content is seen more in the higher quality wines. And Alcohol content correlates with the quality of wine by a value 0.476. But with many outliers, we cannot decide the quality of wine with just the alcohol contents, leading us to seek more factors that might constitute in making better quality wine along with its alcohol content.

[1] 0.4761663

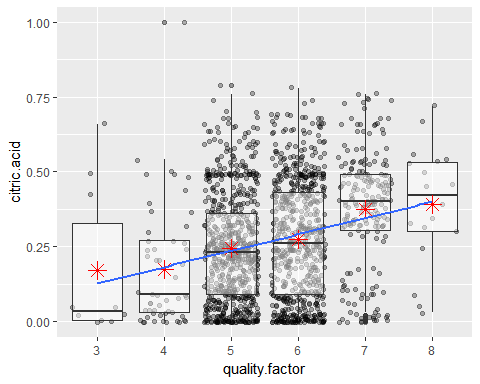
### Sulphates Vs quality



Sulphates and quality are positively correlated with a value, 0.25 only. There are many outliers in this distribution.

[1] 0.2513971

### Citric Acid Vs Quality

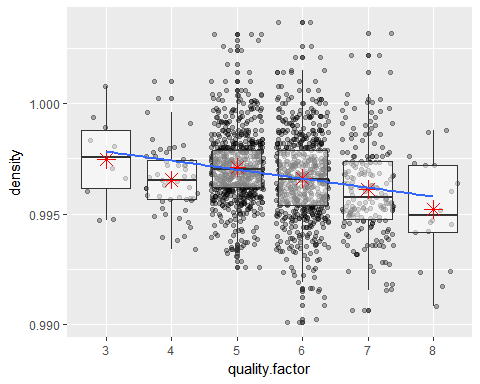


Citric acid and quality factors are positively correlated too with a value, 0.22

[1] 0.2263725

Now let us focus on to the negatively but highly correlated visualizations.

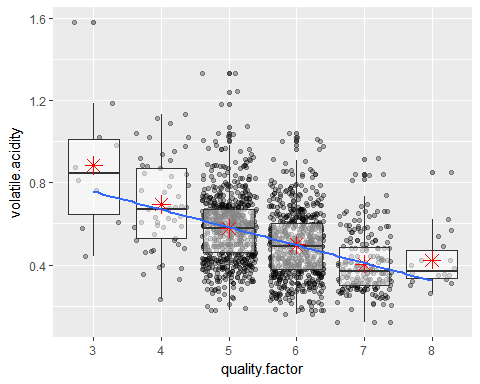
### Density Vs Quality



Well, the blue line is indicating that it is a negative correlation and the function below gives a value, -0.174. The density is pretty much similar for all samples, although we see many outliers. But if we observe the Y-Axis, the interval is very less on which densities are plotted, making all of them similar.

[1] -0.1749192

### Volatile Acidity Vs Quality

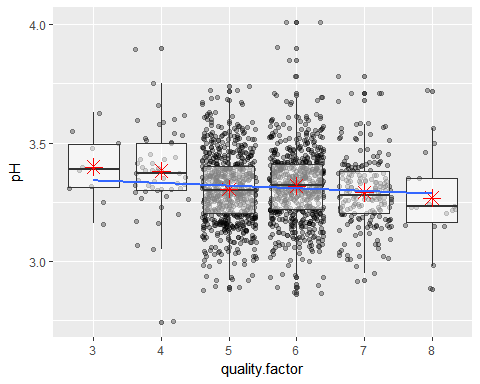


With many outliers yet a correlation value of -0.39, we see that the volatile acidity is decreasing as the quality of the wine increases.

[1] -0.3905578

### pH and Quality

Knowing that pH is the acidity versus the alkaline quotient and a factor that makes a wine good or bad, I was interested in analyzing this chemical property of the wine and it’s influence on the quality.

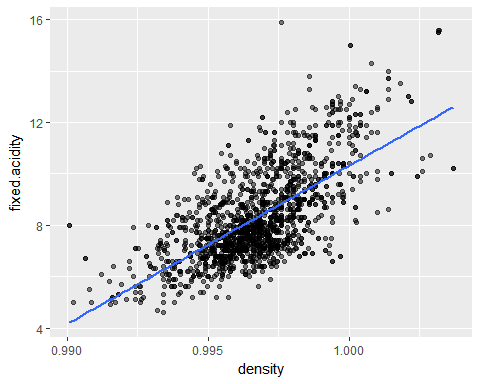


We know that the higher pH in the Red wines have less acidity and thus have a better texture. But the lower pH in red wines also allows them to retain their color and attain greater stability. However, by the plot above, we cannot conclude that the higher pH wines can be rated as better quality wines, as the correlation factor is negative (-0.057) with quality.

[1] -0.05773139

Let us look at the correlation Matrices again to analyse few other features of wine samples with higher correlation factors.

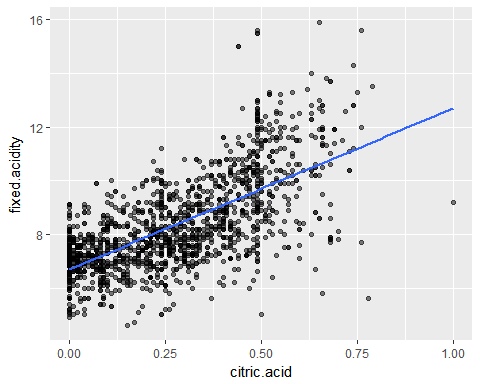
### Fixed Acidity and Density



Density of the wine is increasing as the fixed acidity of the wine increases with a correlation coefficient of 0.668.

[1] 0.6680473

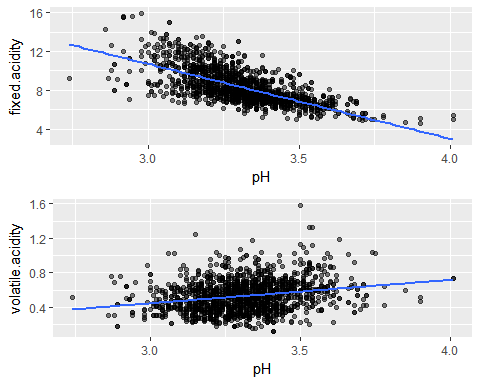
### Fixed Acidity and Citric ACid



Although we see outliers, the function below gives a strong correlation coefficient of 0.67 with increase in the fixed acidity of the wine samples as the citric acid increases in them.

[1] 0.6717034

### Fixed Acidity and pH Versus Volatile Acidity and pH

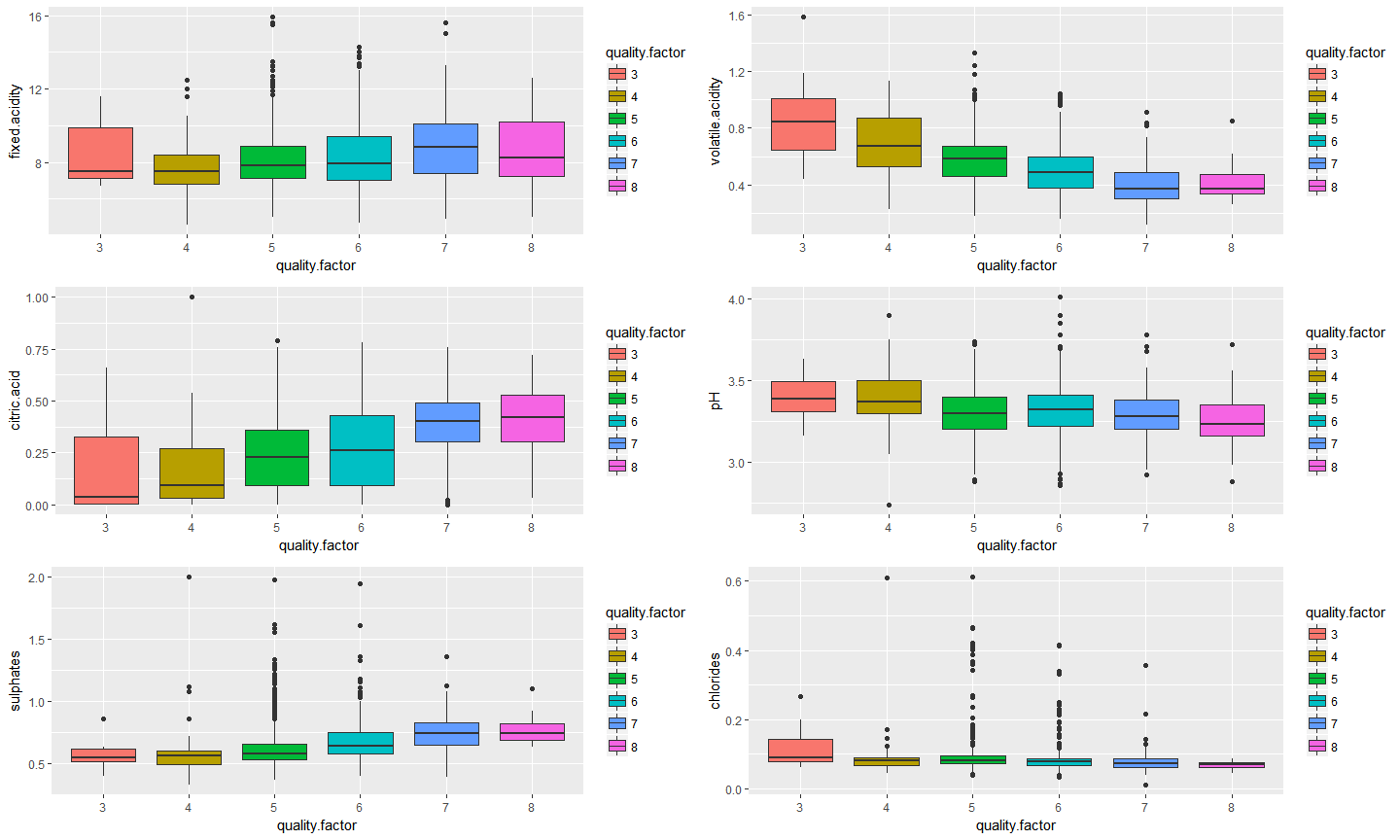


Strongly but negatively correlated factors are Fixed Acidity and the pH. As intended, the pH increases as the fixed acidity decreases. However, surprisingly, the pH slightly increases as the Volatile acidity increases with a positive correlation of 0.23!!

[1] -0.6829782

[1] 0.2349373

### Other components Vs quality



These separate plots show that the higher fixed acidity, higher citric acid, lower volatile acidity and hence a lower pH are found in a better quality wine, although as seen earlier, the pH differences among all wines are negligible!! Higher chlorides worsens the wine quality whereas higher sulphates are found in better quality wines.

# Bivariate Analysis

The analysis was based on the correlation coefficient factor table that we created in the beginning of the Bivariate Analysis section.

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

As expected, alcohol is one chemical component that influences the quality of the wine. Higher its content better the quality of wine. There is negative correlation between volatile acidity and quality. Citric acid and Fixed Acidity strongly correlate negatively with pH as expected.

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

By observing the correlation matrix, we can infer that chlorides and sulphates together correlate positively, but sulphates correlate with quality positively where as the chlorides correlate with quality negatively. This is little confusing. Contrary to my belief that higher pH of Red wines means the better wine quality, the plot here shows that the pH factor surprisingly has very less to do with quality of wine, with a negative correlation coefficient of -0.57.

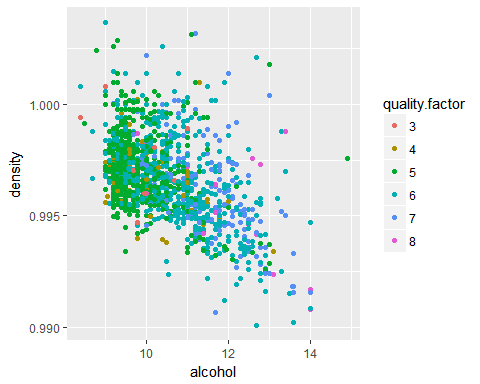
### What was the strongest relationship you found?

The strongest positive correlation is between citric.acid and fixed.acidity (0.67), where as the strongest negative correlation is between pH and fixed.acidity (-0.68) However, alcohol has the strongest positive correlation with quality (0.47) and volatile.acidity has the strongest negative correlation with quality (-0.39)

# Multivariate Plots Section

We already know that higher alcohol content in the wine increases the quality of the wine. We shall now look into other chemical components along with alcohol and checking for the quality of wines.

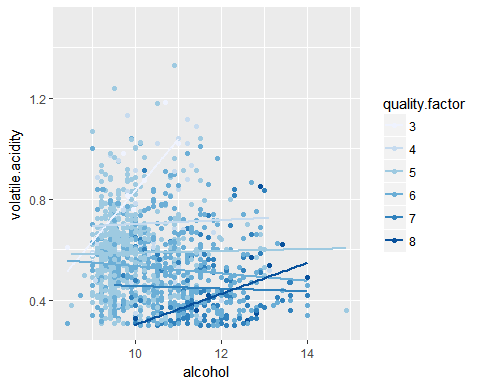
### Density and Alcohol faceted by Quality



We already depicted that density does not influence the quality in bivariate analysis. The above scatter plot confirms that density does not affect the quality even with the increase in alcohol content.There are good wines (color code 8) across the density scale.

### Volatile Acidity and Alcohol

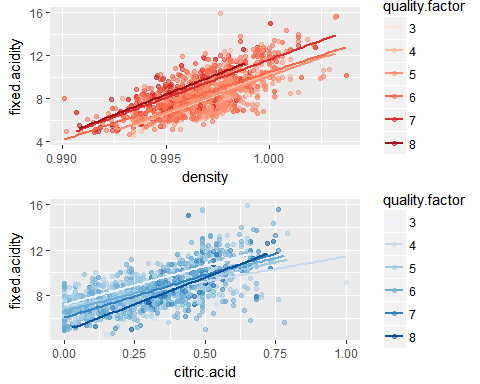
Now let us check how volatile acidity and alcohol affect the quality.



Clearly, the quality lines seen on the above scatter plot indicates as depicted earlier that lower volatile acidity yields in better quality wines, more so with increased amount of alcohol in this case.

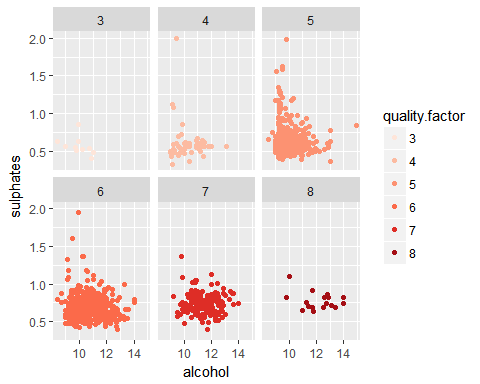
### Fixed Acidity and Density Vs Fixed Acidity and Citric Acid

Let us now compare how fixed acidity, density and Citric acid affect the quality of wine.



Again, considering the lines of quality factors 7 and 8, better wines are prepared with higher fixed acidity across all densities. Higher citric acid and lower fixed acidity yields in better quality wines.

### Sulphates and alcohols facetted by quality factor



Apart from a few outliers in quality factor 4, it is clear from the plots that the higher quality wines have higher sulphate components along with higher alcohol content.

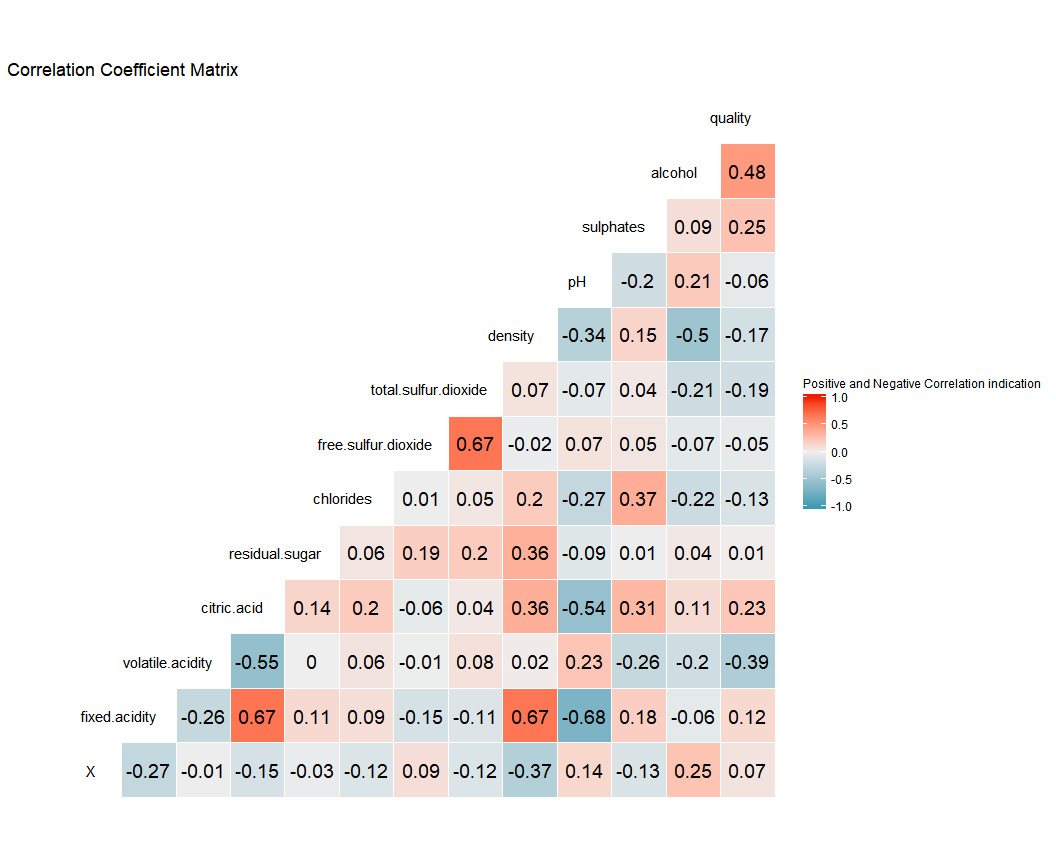
# Multivariate Analysis

### Talk about some of the relationships you observed in this part of the . Were there features that strengthened each other in terms of at your feature(s) of interest?

Yes, higher alcohol along with higher sulphates were found in better quality wines. It is clear that better wines have lower volatile acidity, higher citric acid and lower fixed acidity.

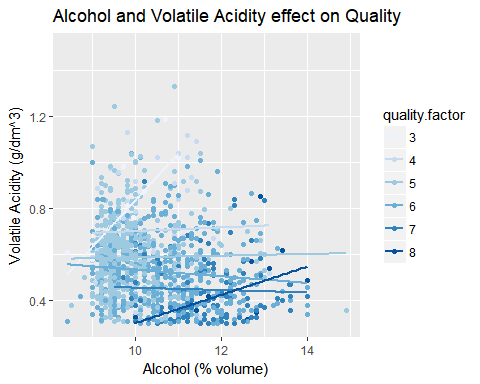
# Final Plots and Summary

### Plot One: Correlation Table



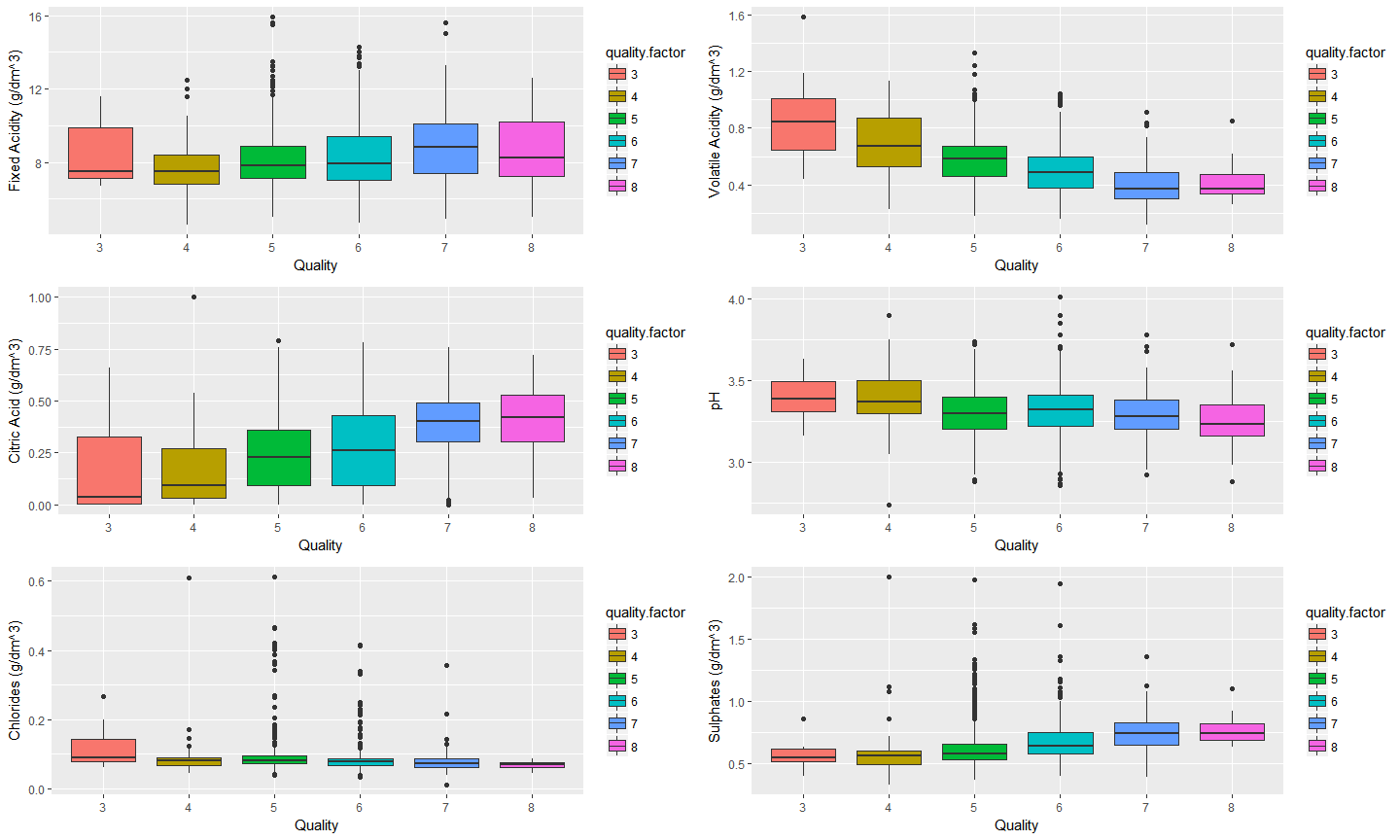
##### This Correlation Coeffiecient Matrix is the basis of all our explorations and visualizations throughout the analysis! This gives us a clear indication of the positively and negatively correlated factors.

### Plot two: Volatile Acidity and Alcohol



##### This is perhaps the most clearily and easily understandable plots of all. We see that the quality factor lines, 6,7 and 8 are seen at the bottom (indicating lower volatile acidity levels) and only at the higher level of alcohol levels. This visualization shows that the wine quality is strongly affected by its alcohol and volatile acidity contents. Higher alcohol and lower volatile acidity levels are seen in better quality wines.

### Plot Three: Acidity, pH, Chlorides and sulphates versus Quality



##### A simple yet important visualisation to understand how various components of a wine can effect our main feature of interest, the quality. It is clear that relatively lower volatile acidity, higher citric acid and hence the lower pH but slightly higher fixed acidity (negligible relative to volatile acidity) and higher sulphates - all constitute to make a better quality wine. Many outliers in the plots also mean that not one component alone is responsible for the quality of wines.

# Reflection

Well, all through the analysis, we had one goal to achieve: Finding various chemical components of the wine to match up with our focus of interest, the 'Quality' of wine. It is easier to taste and decide on a fine wine, where as the process of preparing the best quality wines is rigorous!! Atleast we could say so now, as we go through various aspects that make the quality better. Who knew there are so many controlling factors like pH, Acidity, Sulphates, Chlorides to name a few in making the quality better?!  
  
  
Finally we know that higher alcohol, Sulphates, Citric acid with lower Volatile Acidity and pH constitutes to a better quality wine.   
  
It was fun to explore data of Red wines as I also learnt about the process of making wine through internet on discovering the knowledge on many variables of the data set.  
  
  
As few improvements to this analysis, I could have also derived few linear models in order to show the significant margins in the correlation between variables.   
  
Although I tried to figure out the relation between many factors, the free.sulfur.dioxide and total.sulfur.dioxide are the two variables that did not interest me much as they seemed negligible during the analysis.   
There were not many samples in quality factor 3 and 8 and no data below 3 and above 8 quality-ratings. More wine samples in these categories might have lead to clearer osbervation during visualizations.  
There were no sweet wines in the data. I wonder if there are better quality sweet wines or any wine that is sweet is given a low rating!!

# References:

<https://rawgit.com/mabelvj/data-analyst-nanodegree/master/P4-explore-and-summarize-data/P4-Analysis_of_a_dataset.html>