Red Wine Quality

Arpit Kanodia

August 14, 2015

Importing the CSV file

# Load the Data  
dataframe <- read.csv('wineQualityReds.csv')

## About the Project

In this project, I will explore the quality of wine wrt different properties and chemicals. I will produce the different plots, tables and numeric stats value for finally defining the results.

The data wineQualityReds.csv is available on this link [wineQualityReds.csv](https://www.google.com/url?q=https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityReds.csv&sa=D&usg=AFQjCNFDGQM6iS1viDgkVnXfjlJUymYerA)

And the details about the data is available here [Text file about Details of Data](https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityInfo.txt)

## Basic Details about the Dataset

Lets examine the structure and schema of dataset

str(dataframe)

## '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 ...

summary(dataframe)

## 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

As our main goal of this project is quality, it will be intresting to see statistics of only quality.

str(dataframe$quality)

## int [1:1599] 5 5 5 6 5 5 5 7 7 5 ...

summary(dataframe$quality)

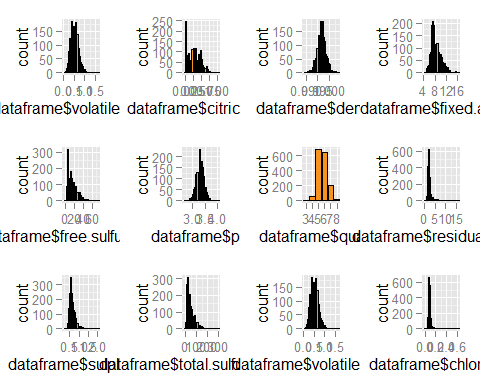
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.000 5.000 6.000 5.636 6.000 8.000

Some initial basic details of the dataset are observed as following:

* There are 1599 obs. with 13 variables.
* From the literature, the value of quality is in range of 0-10, and was rated by atleast 3 whine experts. In the dataset value range from 3-8, with median of 6 and mean at 5.636.
* X appears to be a identifier for each data.

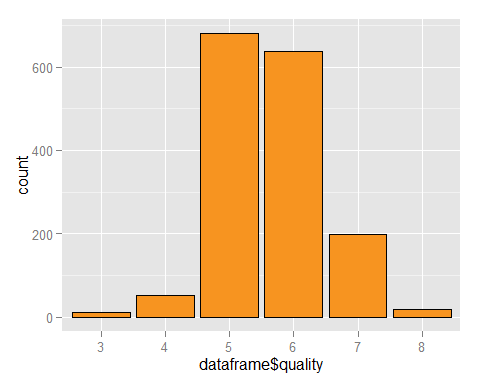
## Plots of Single Variables

Just to explore the data visully, I'll draw all the graphs for 12 variables quickly. The motive behind to see the distribution. Histograms are required to analyze initially about the distribution of the different variables.



## Analysis of Single Variable

## Quality of the Wine

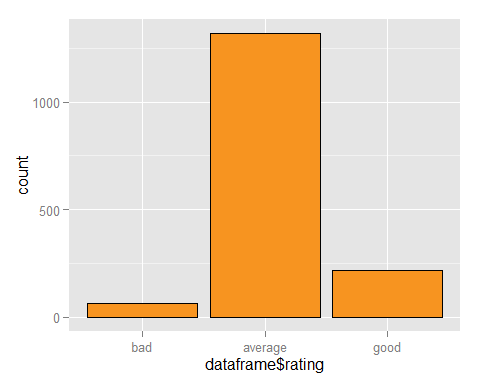
The quality of wine is in discrete value range from 0-10 as mentioned in description text of dataset. but with graph of quality one can definetely see the quality is in between 3-8, where as most of the data lie between 5-7, with very small data at 3,4 and at 8. 

For rating of wine, I thought to divide the quality into 'bad','average', 'good' and 'very good'. As the data range only from 3-8, dividing the data into 4 parameters, whereas 'very good' is non existent as there are no data for 9 and 10, is nonsensical. I finally decided to divide the data into only 3 parameters 'bad', 'average' and 'good'.

#Dividing data according to Quality  
dataframe$rating <- ifelse(dataframe$quality < 5, 'bad',   
 ifelse(dataframe$quality < 7, 'average', 'good'))  
#Adding levels  
dataframe$rating <- ordered(dataframe$rating,  
 levels = c('bad', 'average', 'good'))  
summary(dataframe$rating)

## bad average good   
## 63 1319 217

qplot(dataframe$rating, color = I('black'),fill = I('#F79420'))

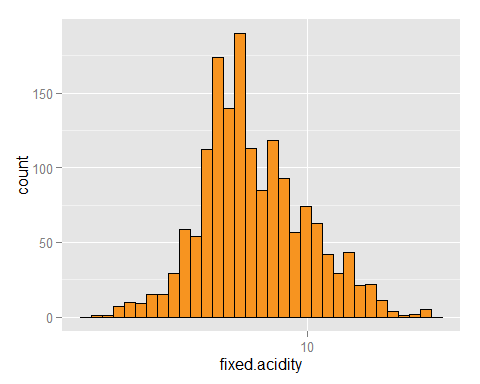
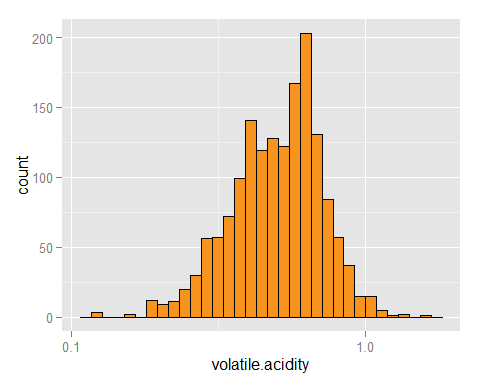
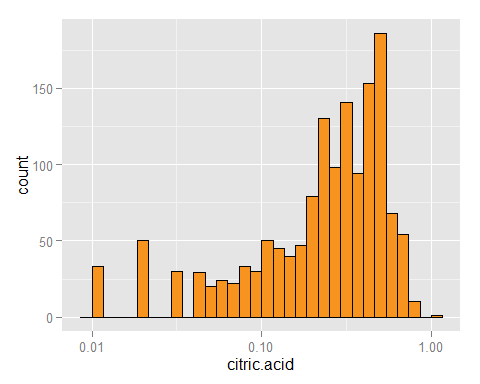


## Analysis of Distribution

* It is obviously appearing in graphs that maybe pH and density are normally distributed. -Sulphur dioxide,sulphates,alcohol, fixed and volatile acidity looks like to be long tailed.
* Residual sugar and chlorides have a extreme outliers.
* Citric acid seems to have large number of zero values.

#### Log based Plots

To furthur analyze the distribution, log10 based transformation is required.

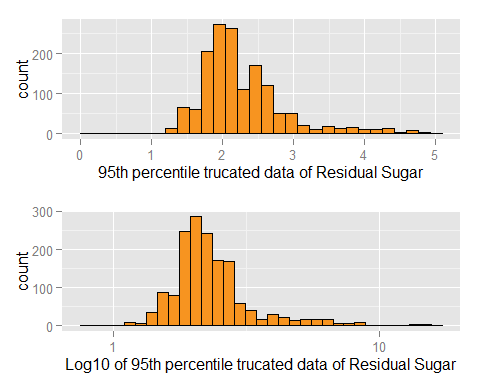
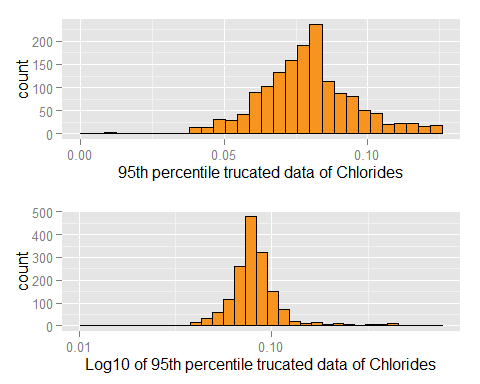
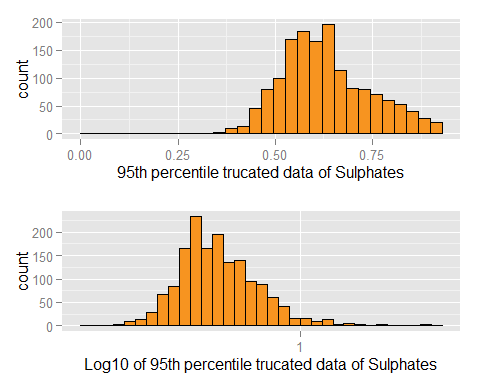
  

When plots are made on the base of log 10, fixed.acidity and volatile.acidity appeared to be normally distributed, as the pH is normally distributed which is measure of acidity in log scale. But, citric.acid not appeared to be normally distributed. On furthur observations.

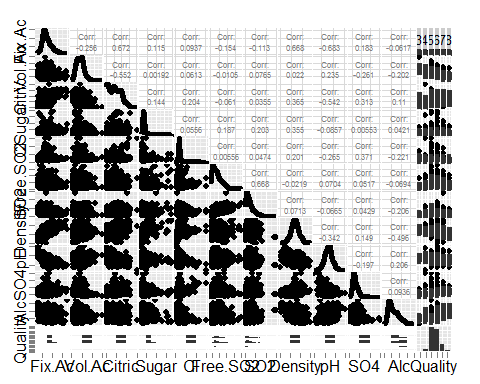
length(subset(dataframe, citric.acid == 0)$citric.acid)

## [1] 132

This is clear that 132 observations are actually zero.

## Checking the Overall picture of Dataset

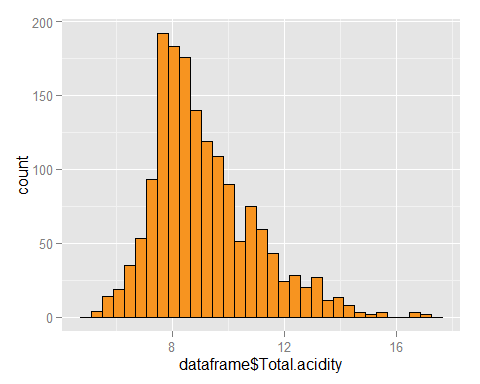


## Some Short Questions

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

Yes, rating based on quality is created that is bad,average and good wine. In last plot I also created AlcohalRating according to % of Alcohol in Wine.

Furthur I created a new variable for acidity, it seems like fixed.acidity and volatile.acidity caused because of different acids in the Wine. I created a new Total.acidity containing the sum of fixed.acidity,volatile.acidity and citric.acidity.



I searched on internet and I found fixed acidity caused because of tartaric, malic, citric, and succinic acids in wine. And volatile acidity causes because of acetic acid.

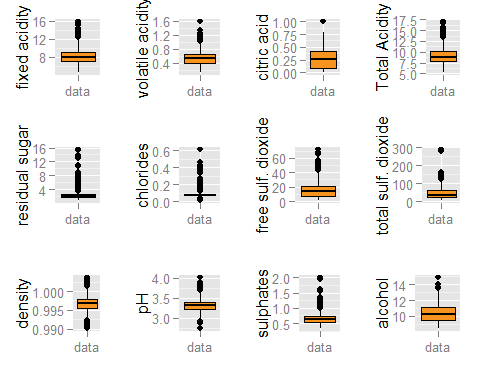
Source:

[Fixed Acidity](http://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity) [Volatile Acidity](http://waterhouse.ucdavis.edu/whats-in-wine/volatile-acidity)

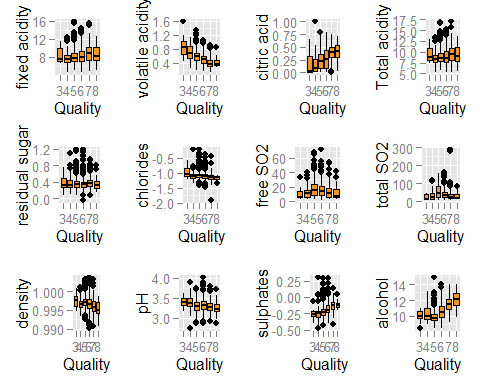
#### Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

I investigated the distribution in above sections. For outliers I decided to plot the boxplots. In Single Variable plot, I didnt tidy or adjust the data, but plotted some on log scale.

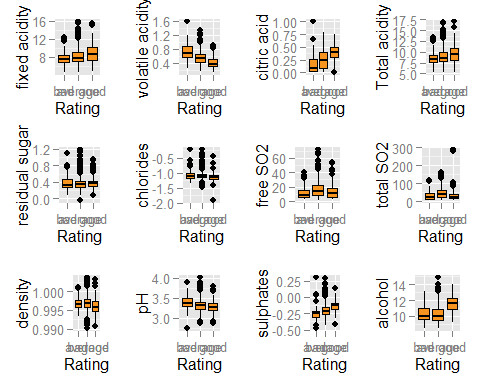
#### Boxplots of Single Variable

For defining the median across quality and rating it is required to plot the Boxplot. 

#### Boxplots for Two variable (Quality V/S Different Variables)

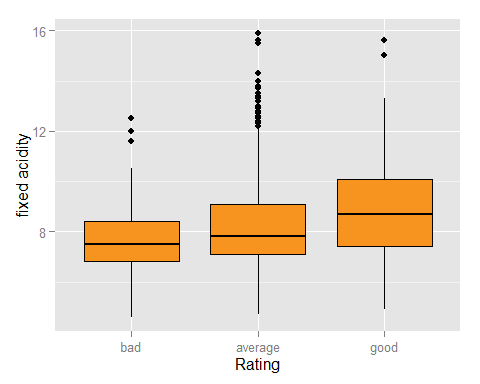


#### Boxplots for Two variable (Rating V/S Different Variables)

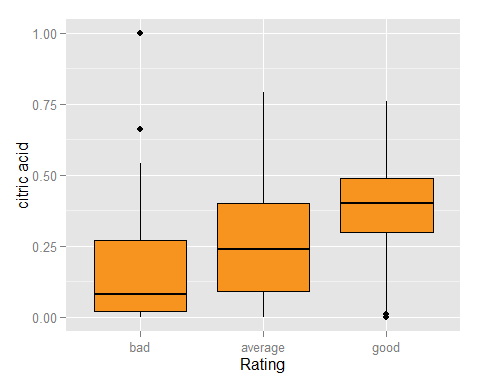


#### Analysis of Boxplots

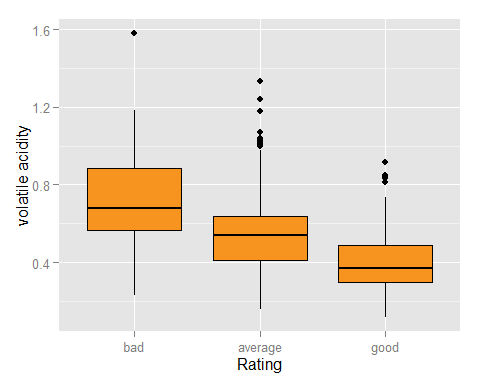
By examining the plots, some conclusions can be drawn:



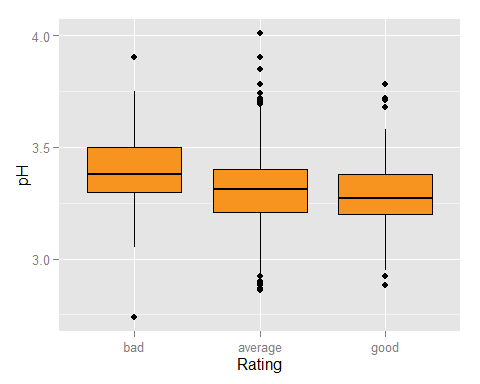
* A good wine causes because of high fixed.acidity (tartaric, malic, citric, and succinic acids).



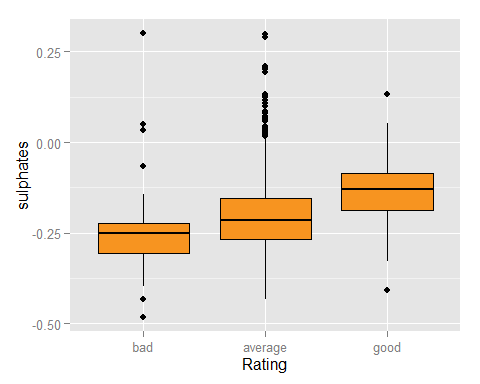
* A good wine causes because of high citric.acid.



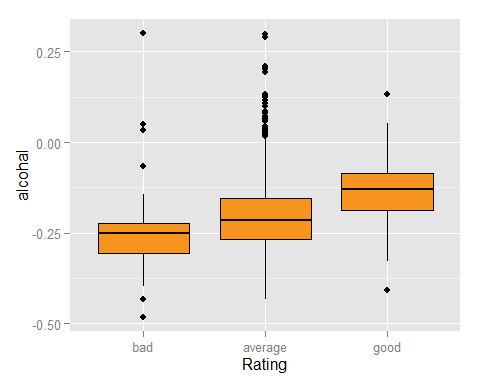
* A good wine causes because of low volatile.acidity(acetic acid).



* A good wine causes because of lower pH value, means higher acidity.



* A good wine causes because of higher sulphates.



* A good wine causes because of higher alcohal.
* Little bit, lesser chlorides and density also affect the wine quality.
* SO2 and Residual Sugar donot seems to have any affect on wine quality.

#### Correlation in between Quality and Different Variables

## fixed.acidity volatile.acidity citric.acid   
## 0.12405165 -0.39055778 0.22637251   
## Total.acidity log10.residual.sugar log10.chlordies   
## 0.10375373 0.02353331 -0.17613996   
## free.SO2 total.SO2 density   
## -0.05065606 -0.18510029 -0.17491923   
## pH log10.sulphates alcohol   
## -0.05773139 0.30864193 0.47616632

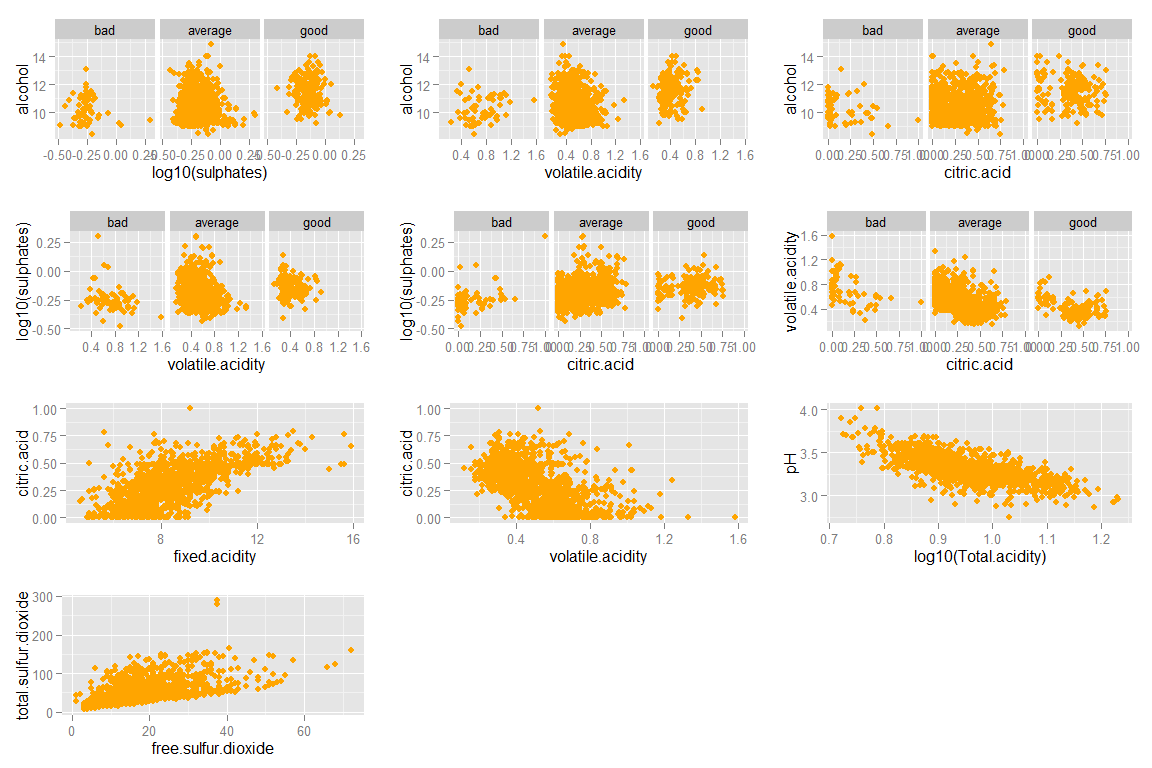
Strong Correlation found in quality and

* Alcohal
* Volatile Acidity
* Log10 of Sulphates
* Citric Acid

Definetely Alcohal, Volatile Acidity, Log10 of Sulphates and Citric acid are the main parameters that affecting the quality of wine.

#### Scatter Plot

To graphically check the correlation and the represent the relatationship graphically, it required to plot the Scatter plot.



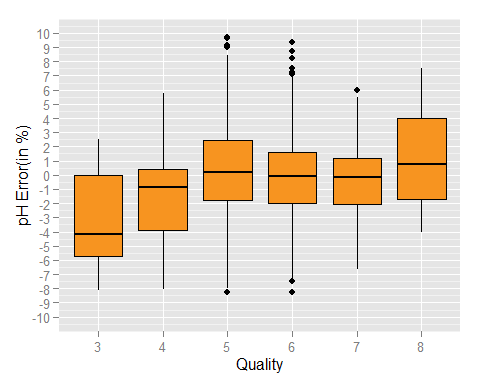
## log10(Sulphates) v/s Alcohol   
## 0.1351562   
## volatile.acidity v/s Alcohol   
## -0.2022880   
## volatile.acidity v/s log10(sulphates)   
## -0.3005487   
## citric.acid v/s log10 (sulphates)   
## 0.3315162   
## citric.acid vs volatile.acidity   
## -0.5524957   
## fixed.acidity v/s citric.acid   
## 0.6717034   
## volatile.acidity v/s citric.acid   
## -0.5524957   
## Total.acidity v/s pH   
## -0.7044435   
## free SO2 v/s Total SO2   
## 0.6724531

I found very strong correlation between Total.acidity and pH, and the strongest correlation after that is between free.SO2 and total.SO2.

This is obvious that the strong correlation found in Acidity and pH because of pH is inversely proportional to acidity. One question need to ask what other components affecting pH other than acids. We can see this difference by building a linear model.

#### Errors in between Observerd pH and Predictive pH

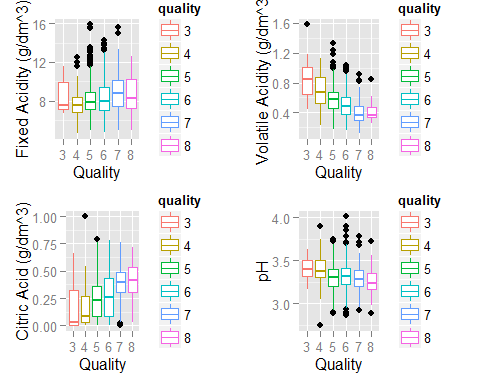
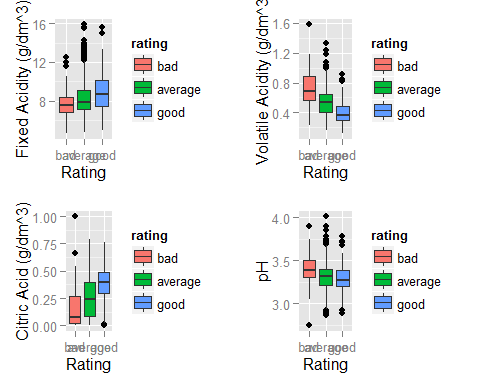
#Defining Linear Model  
m <- lm((pH) ~ log10(Total.acidity),data=dataframe)  
#Making predictions according to LM.  
dataframe$pH.predictions <- predict(m, dataframe)   
#Calculating % of Error  
dataframe$pH.error <- ((dataframe$pH.predictions - dataframe$pH)/dataframe$pH)\*100  
#Creating a Boxplot in between Quality V/S pH.error  
qplot(data = dataframe,   
 x = dataframe$quality, y = dataframe$pH.error,   
 geom = 'boxplot', ylab = 'pH Error(in %)',   
 xlab= 'Quality',color = I('black'), fill = I('#F79420')) +   
 scale\_y\_continuous(limits = c(-10,10), breaks = seq(-10,10,1))



The median % error in all cases remain somewhere close to 0 except in quality of 3. The median % of error of quality of 3 is at slightly lower than negative 4%. In this case we can conclude, the acids are not performing well in this case, or error is caused because of impurity in Acids.

# Summary and Plots

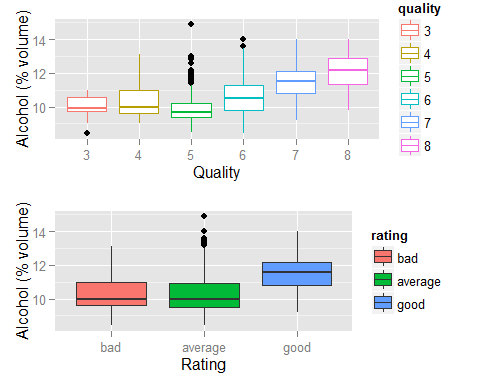
## Effect of acids on Wine

#### Conclusions on effect of acidity on Wine

* There is little effect of Fixed acidity on Wine. But lower than 8 g/dm^3 acetic acid can cause a wine to be a bad wine, and a slight larger value than 8 g/dm^3 acetic acid can convert a wine into a good wine.
* Volatile acidity which causes because of tartaric, malic, citric, and succinic acids should be slightly below 0.4 g/dm^3.
* If Citric acid is closer to 0 g/dm^3, then this can cause a wine to be bad. A value between 0.25-0.50 g/dm^3 is required for good wine.
* For a good wine pH value should be in between 3-3.5, but a closer value to 3.5 can cause a wine to be bad.

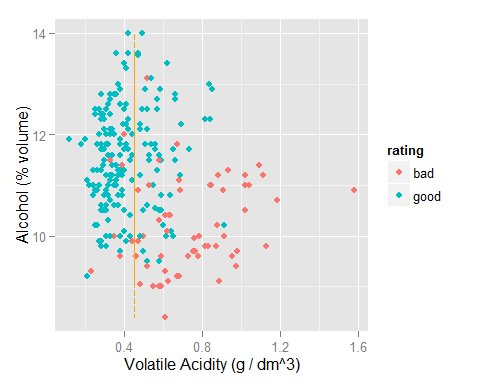
## Effects of Alcohol on Wine Quality



The boxplot show a higher quantity of alcohol may cause a good quality wine. But, as the outliners show, alchohol content alone didnot produce high quality wine.

## What is the actual difference in Bad and Good Wine

To understand this, I removed average rating from plot.



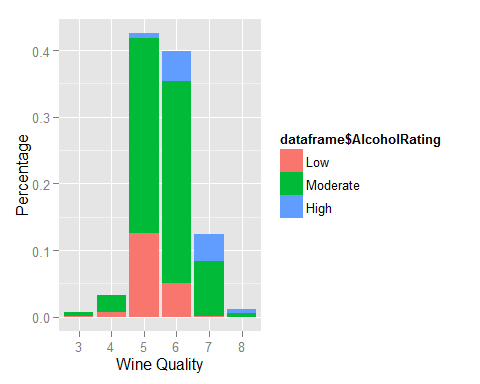
* From this plot, this can be concluded the good wine creating a cluster when quantity of volatile.acidity is low and quantity of alcohal is high compare to bad wine (except some cases).
* The best way to make a good wine can be keeping volatile.acidity low and keeping alcohol high.

## Histogram Based on % Volume of Alcohol

#Dividing data according to Alcohol %  
dataframe$AlcoholRating <- ifelse(dataframe$alcohol < 9.5, 'Low',   
 ifelse(dataframe$alcohol < 12, 'Moderate', 'High'))  
#Adding levels  
dataframe$AlcoholRating <- ordered(dataframe$AlcoholRating,  
 levels = c('Low', 'Moderate', 'High'))  
  
summary(dataframe$AlcoholRating)

## Low Moderate High   
## 297 1140 162

ggplot(data= dataframe, aes(x=as.factor(dataframe$quality) , fill = dataframe$AlcoholRating)) +   
 geom\_histogram(aes(y=..count../sum(..count..))) +  
 xlab("Wine Quality") + ylab("Percentage")



For plotting this, we divided the data into 3 section according to % of Alcohol according to this

* Low with less than 9.5%
* Moderate with inbetween 9.5 and 12%
* High above 12%.

I found as the quality increases, the % of alcohal also increases. The number of wine with Low % of alcohol vanished as Wine Quality increases.

# Summary

The findings for Good Wine as follows.

* A wine with greater 8 g/dm3 acetic acid (Fixed Acid) can cause a good wine.
* Volatile acid should be remain below 0.4 g/dm3.
* Citric Acid should be at 0.25-050 g/dm3.
* I dont able to find much difference of pH value in bad and good wine. But the value should be closer to 3.5
* Alcohol % should be above 9.5%

# Final Reflection

I was able to determine which factor affect the wine quality most i.e alcohol, sulphates and acidity. Important note is that quality or rating is subjective manner, and can change from person to person. The above graphs described what makes a wine good.

I never done the statistical analysis from boxplot before, it took me time to understand the boxplot.

Furthur in most of the links, like this [Link](http://www.realsimple.com/holidays-entertaining/entertaining/food-drink/alcohol-content-wine), I found a Strong/High Wine is called when Alcohol % is called when wine is above 14.5%. But I found 0 data which is above 14.5%. I searched pretty hard then I found some 10 years ago, above 12% was considered as Strong Wine.

This is possible the dataset is old, or dataset is made from old parameters of wine. I furthur faced many problems in many functions of R, as is grid.arrange, gridarrange and in creating functions. I pretty worked hard in understanding this in [stackoverflow](http://stackoverflow.com/).

On other note, the dataset is from which country is not defined. The various parameters defined differently in different country. Demographically and ethinically, a wine which is bad for one maybe good for other. For example, a French can rate a wine as good but a person from US maybe rate the same wine as bad.

## Future Analysis

* Future Analysis can be done by sub-classification is residual.sugar in sweet wine and dry wine. [Wikipedia](http://en.wikipedia.org/wiki/Sweetness_of_wine#Residual_sugar)
* Furthur examination of outliers.
* Rating and Quality based on region.