RedWine

12/16/2017

```r
knitr::opts_chunk$set(echo=TRUE, warning=FALSE, message=FALSE)
library(ggplot2)
library(GGally)
library(scales)
library(gridExtra)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:gridExtra':
##
##     combine

## The following object is masked from 'package:GGally':
##
##     nasa

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(knitr)
library(memisc)
```

```
## Warning: package 'memisc' was built under R version 3.4.3

## Loading required package: lattice

## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

##
## Attaching package: 'memisc'

## The following objects are masked from 'package:dplyr':
##
##     collect, recode, rename

## The following object is masked from 'package:scales':
##
##     percent

## The following objects are masked from 'package:stats':
##
##     contr.sum, contr.treatment, contrasts
```

```
## The following object is masked from 'package:base':
##
##     as.array
```

```
##Load Data
wd<- read.csv('wineQualityReds.csv')
```

Summary

Basic summary of the data is obtained with some basic commands in R.

```
str(wd)
```

```
## '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(wd)
```

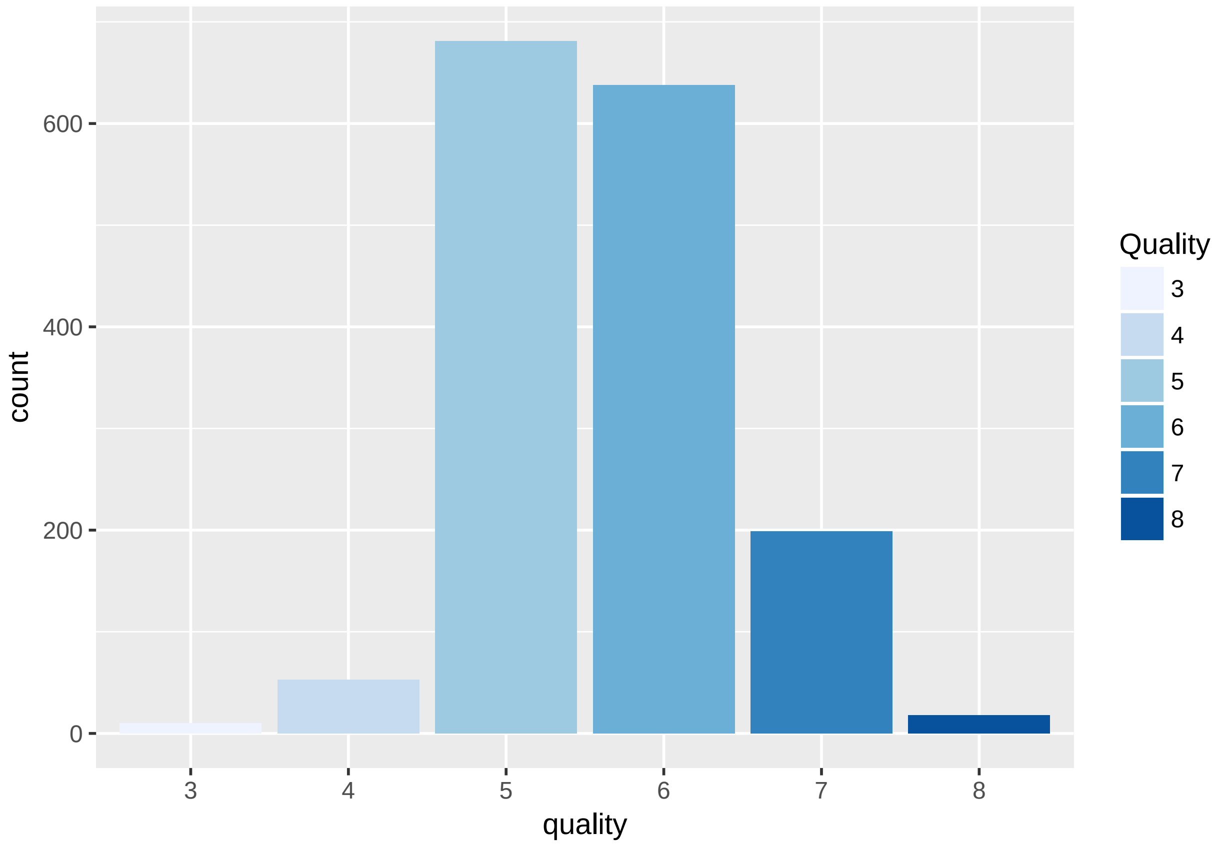
```
##        X          fixed.acidity   volatile.acidity  citric.acid
##  Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000
##  1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090
##  Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260
##  Mean   : 800.0   Mean   : 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
```

There are 1599 observations with 13 different variables. X is a unique identifier with a integer value. Quality is also an integer value. All other values are numeric but not necessary integers.

Here we are primary concerned with wine quality, so lets start with some basic plots.

```
ggplot(aes(as.factor(quality),fill= factor(quality)), data = wd) + geom_bar() +
  theme_replace() + xlab("quality")+
  scale_fill_brewer(type = 'seq',guide=guide_legend(title = 'Quality'))
```



From the data obtained until now some things can be inferred like,

- Quality lies between 3 and 8.
- Mean quality is 5.636.
- Median Quality being 6.

Univariate Analysis

Wine Quality

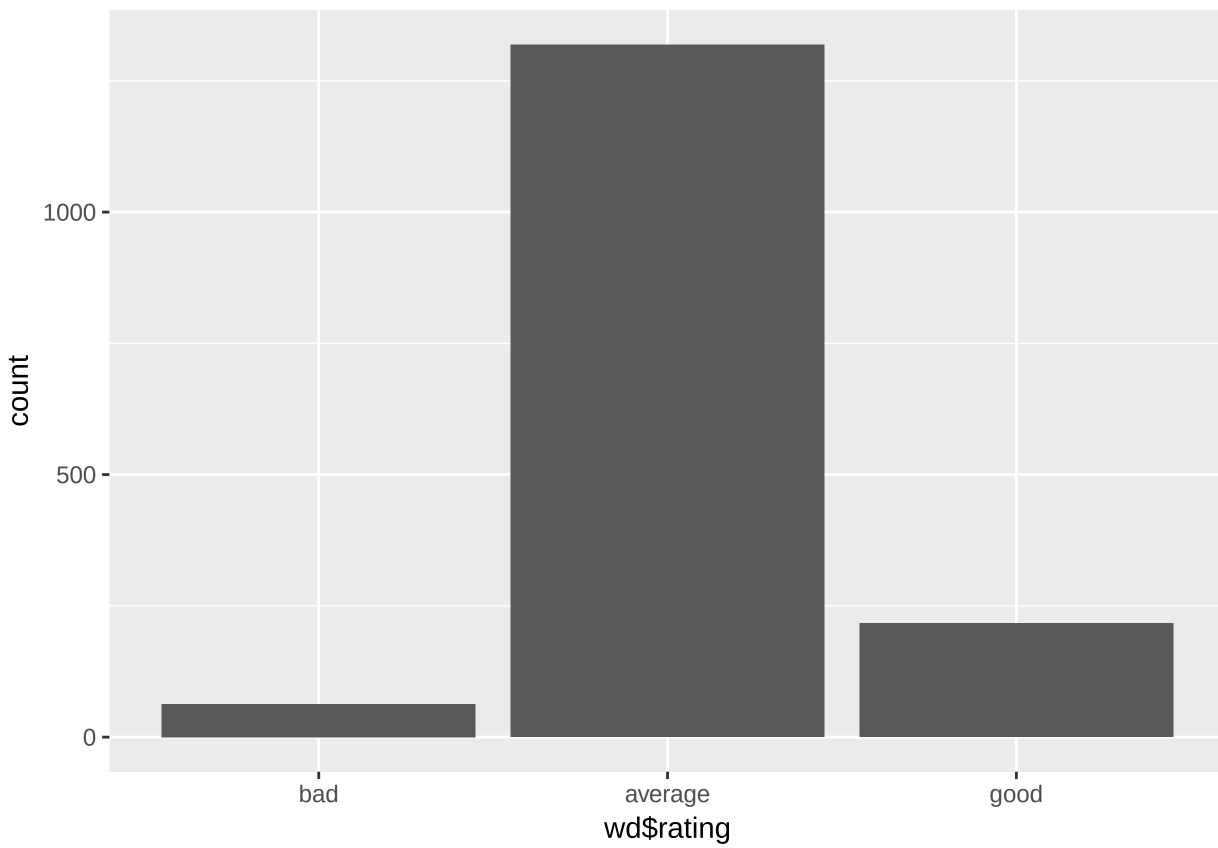
Looking at our first plot of wine quality, it roughly has a normal distribution with most rating being in 5 and 6. So lets create an another variable with variable ratings with following categories.

- 0-4 : poor
- 5-6: good
- 7-10 :ideal

```
wd$rating <- ifelse(wd$quality <5, 'bad', ifelse( wd$quality<7, 'average','good'))
wd$rating<- ordered(wd$rating, levels = c('bad','average','good'))
summary(wd$rating)
```

```
##     bad average    good
##      63    1319     217
```

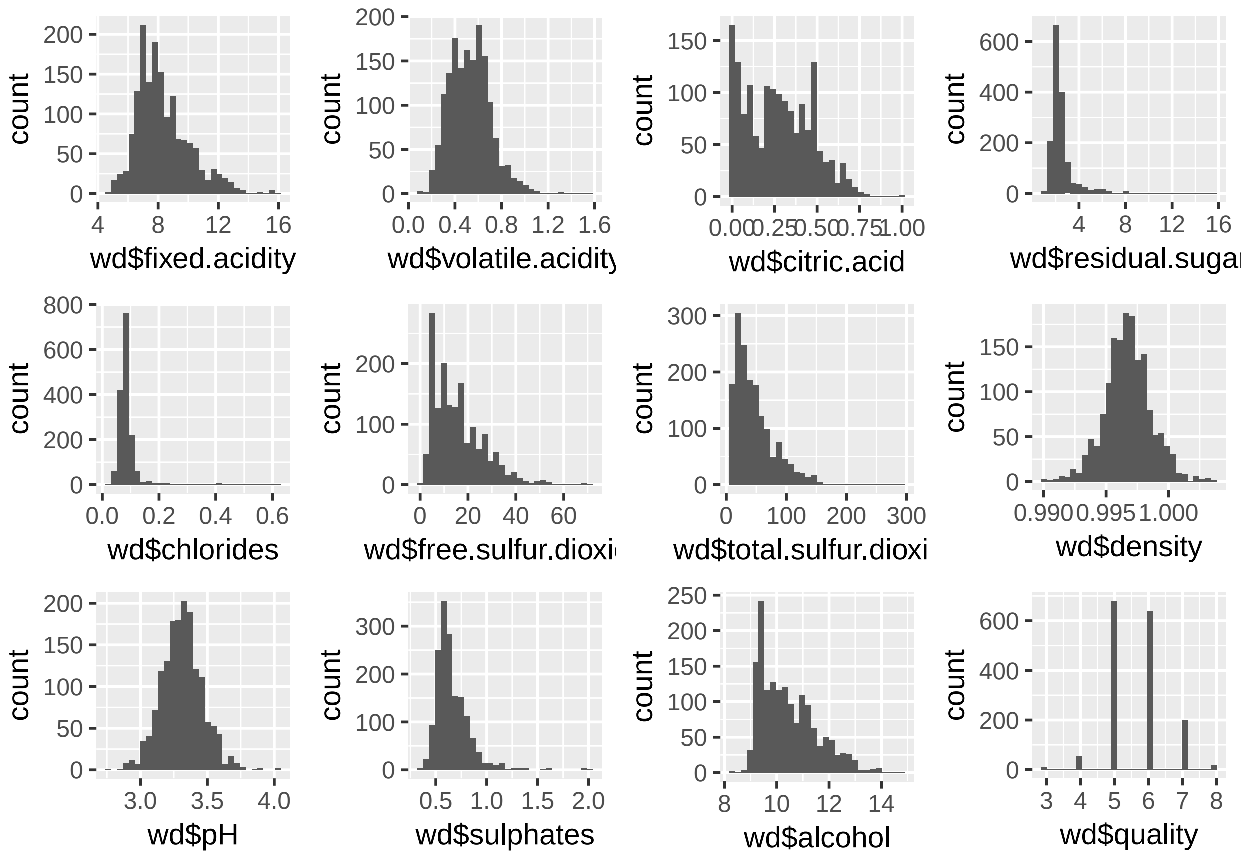
```
qplot(wd$rating)
```



Univaraiate plots section

```
grid.arrange(qplot(wd$fixed.acidity),
             qplot(wd$volatile.acidity),
             qplot(wd$citric.acid),
             qplot(wd$residual.sugar),
             qplot(wd$chlorides),
             qplot(wd$free.sulfur.dioxide),
             qplot(wd$total.sulfur.dioxide),
             qplot(wd$density),
             qplot(wd$pH),
             qplot(wd$sulphates),
             qplot(wd$alcohol),
```

```r
qplot(wd$quality),
ncol = 4)
```




```r
summary(wd$fixed.acidity)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
```

```r
summary(wd$volatile.acidity)
```

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

```r
summary(wd$citric.acid)
```

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

```r
summary(wd$residual.sugar)
```

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

```r
summary(wd$chlorides)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```

```r
summary(wd$free.sulfur.dioxide)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##    1.00    7.00   14.00   15.87   21.00   72.00
```

```r
summary(wd$total.sulfur.dioxide)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00
```

```r
summary(wd$density)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0037
```

```r
summary(wd$pH)
```

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

```r
summary(wd$sulphates)
```

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

```r
summary(wd$alcohol)
```

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

```r
summary(wd$quality)
```

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

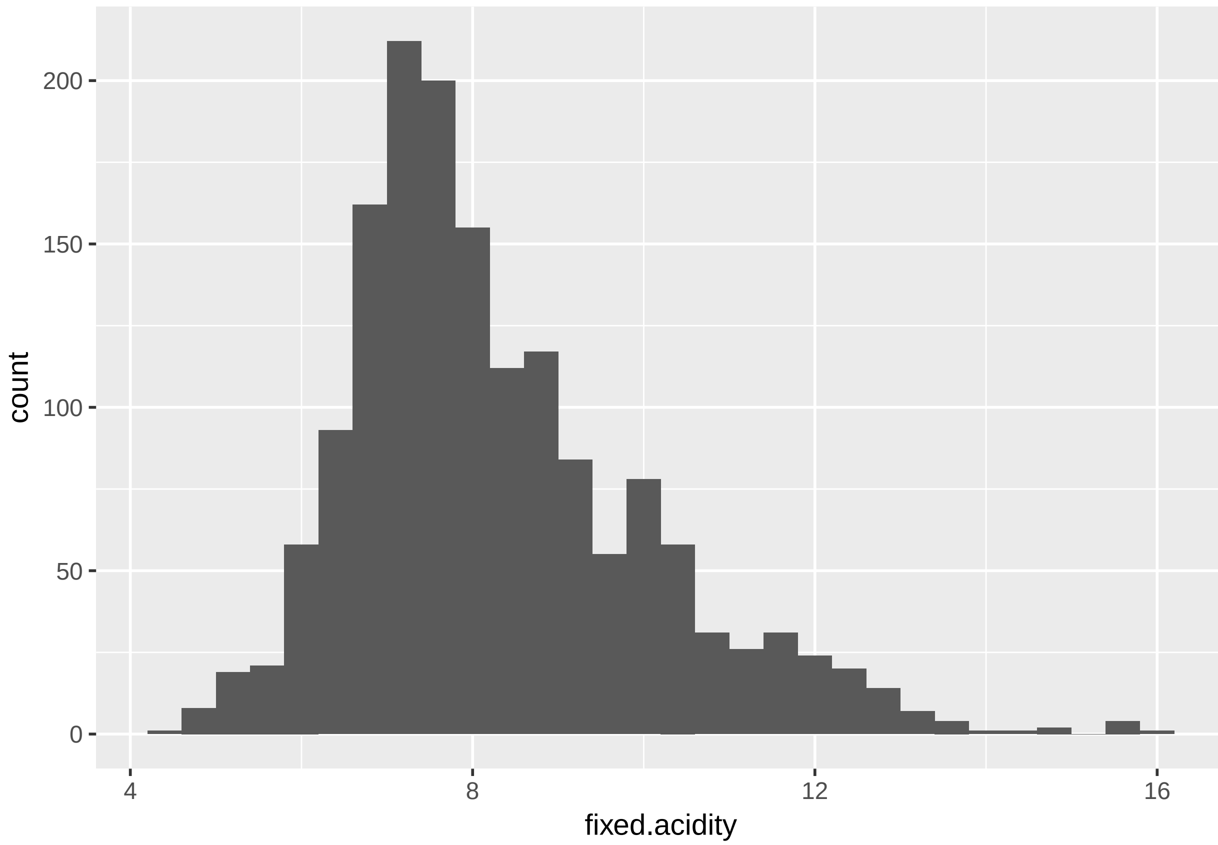
Distribution and Outliers

Looking at the plots above inferred details are as fallows,

- Density and pH are normally distributed.
- Qualitatively, residual sugar and chlorides have extreme outlines.
- Fixed and volatile acidity, sulfur dioxides, sulphates, and alcohol seem to be long-tailed.
- Citric acid have many zero values,looks like there is some error in reporting but I am curious to know.

Since fixed and volatile acidity are long tailed I plotted them in log10 scale and found them to be normally distributed.

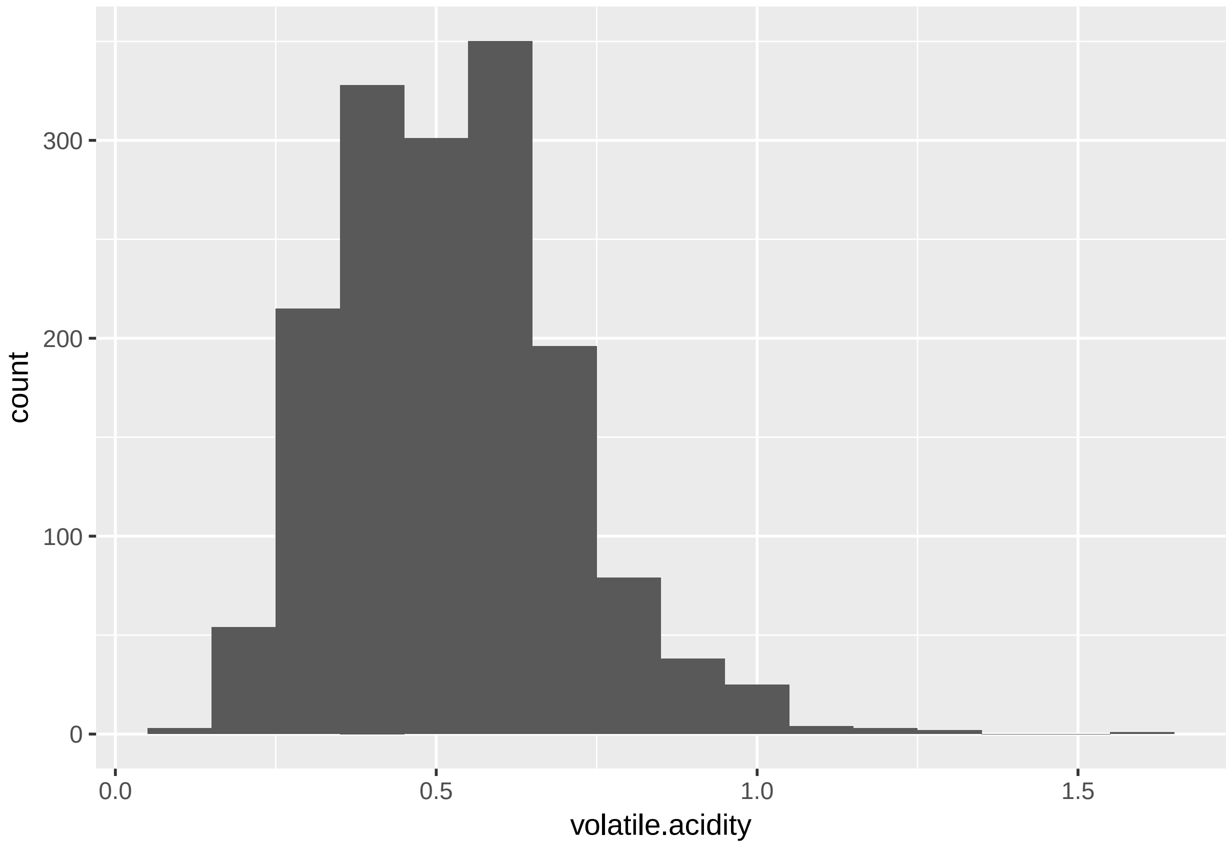
```r
ggplot(data= wd,aes(x=fixed.acidity))+geom_histogram(binwidth = 0.4)
```




```
summary(wd$fixed.acidity)
```

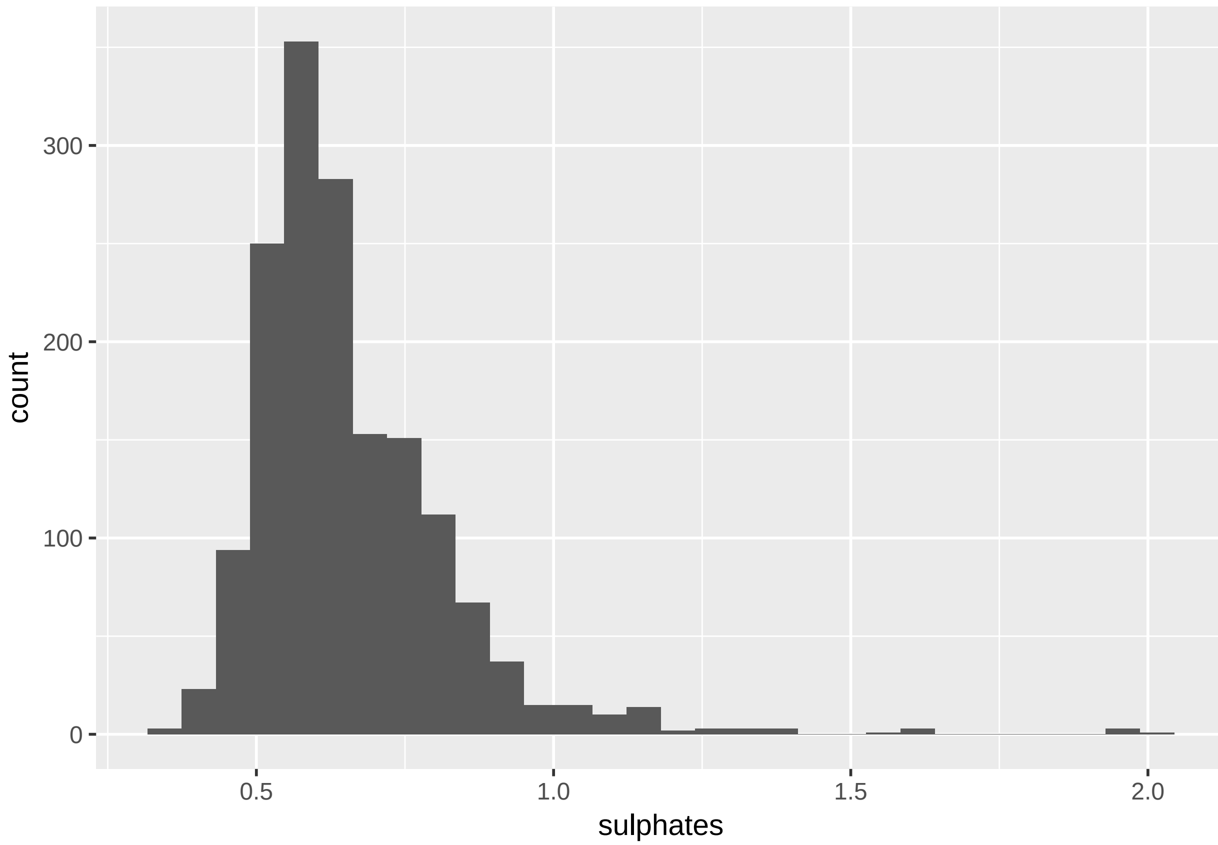
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
```

```
ggplot(data= wd,aes(x=volatile.acidity))+geom_histogram(binwidth = 0.1)
```

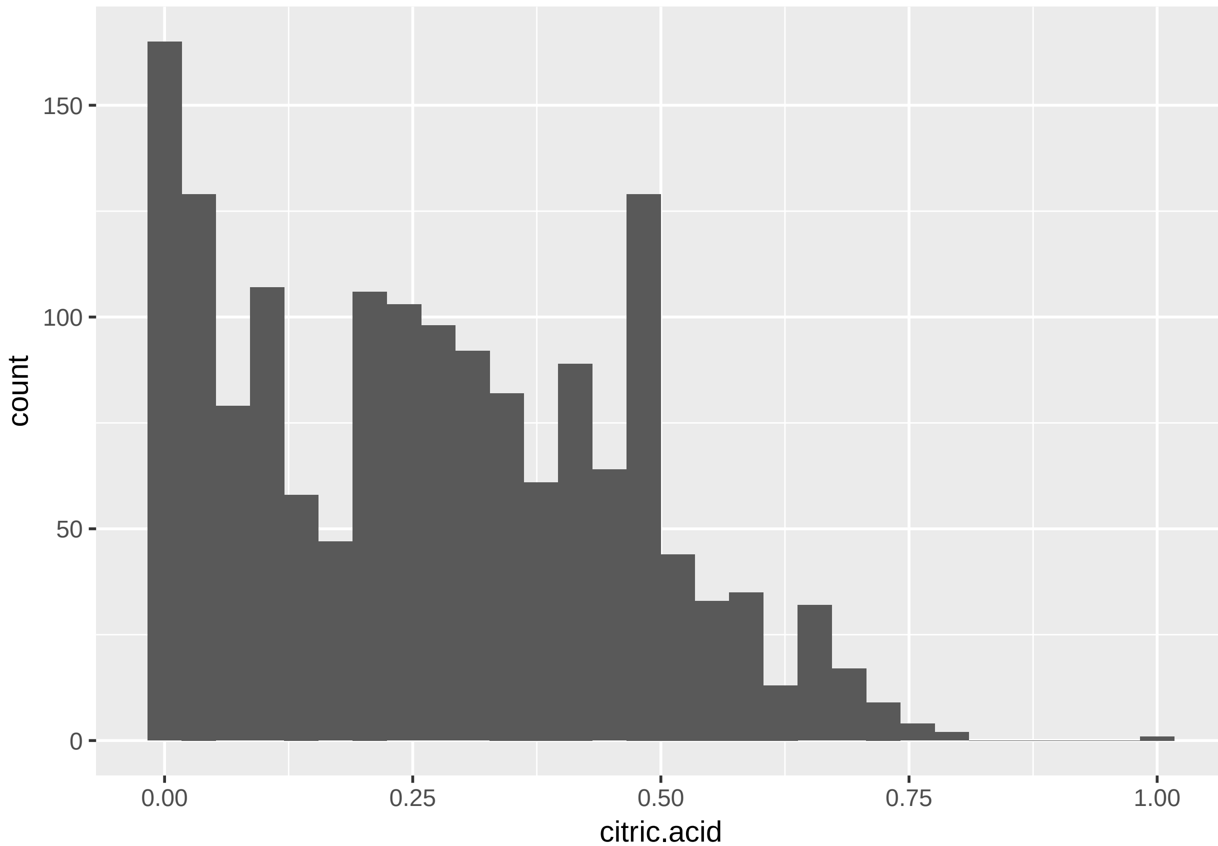


Similarly I plotted citric acid and sulphates to find out if they are normally distributed but found out only sulphates are normally distributed.

```
ggplot(data= wd,aes(x=sulphates))+geom_histogram()
```




```r
ggplot(data= wd,aes(x=citric.acid))+geom_histogram()
```



Further investigating the data on total number of zero entries I found that there are 132 in total.

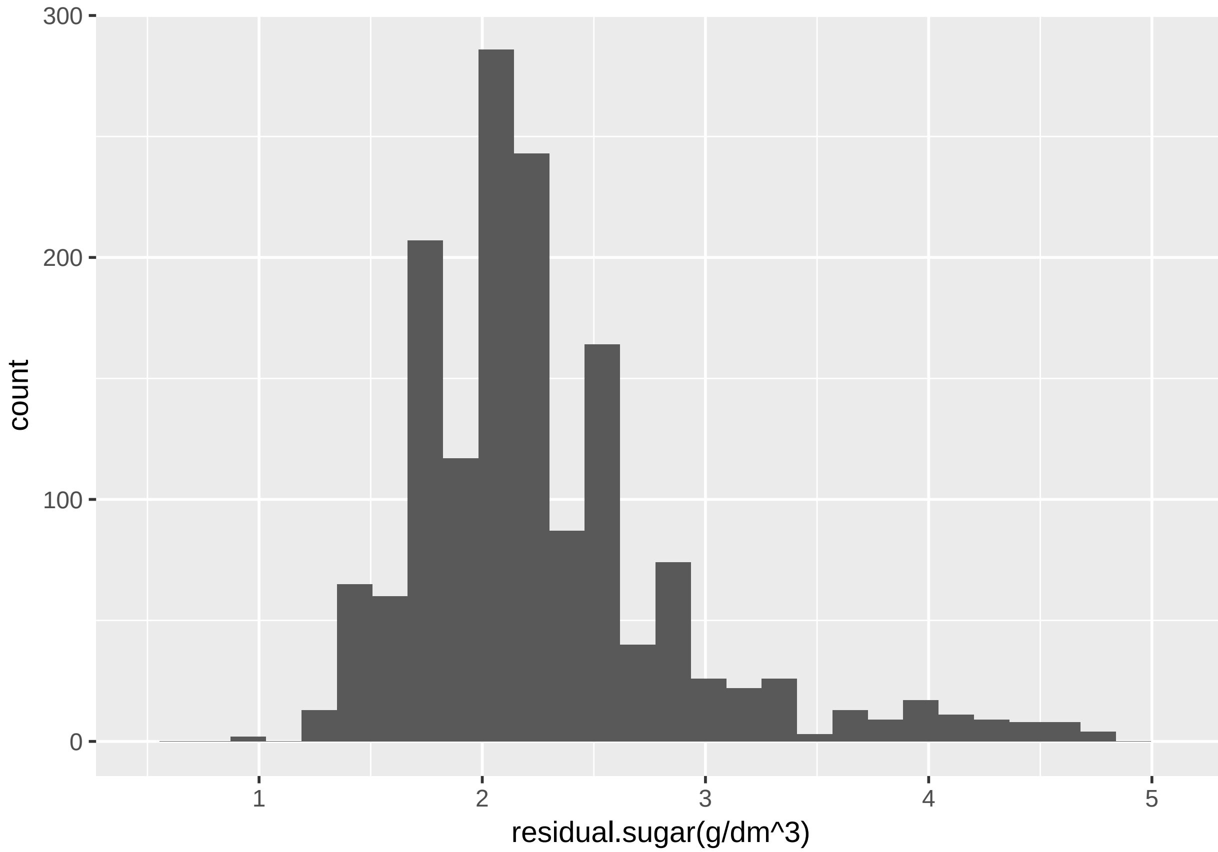
```
length(subset(wd, citric.acid==0)$citric.acid)
```

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

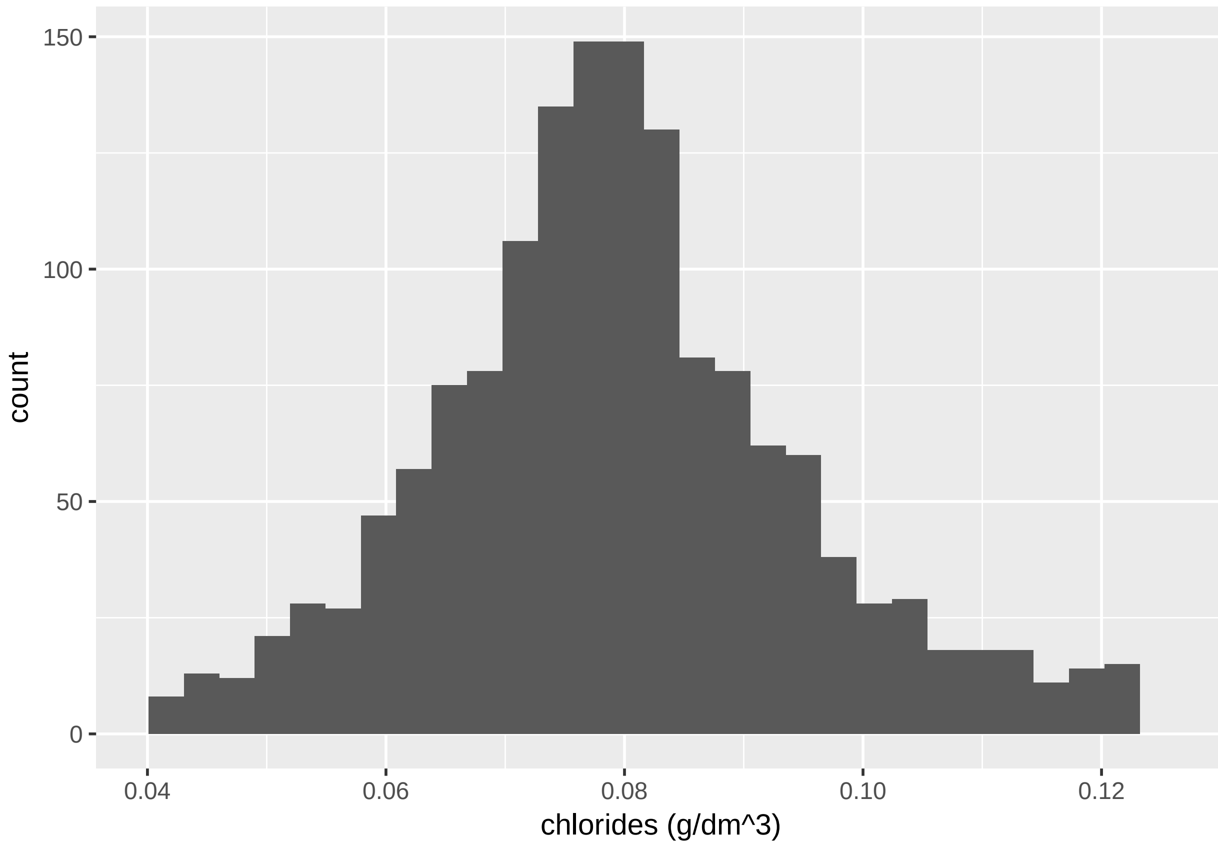
Plots in residual.sugar and chlorides

After removing some extreme outliers in the data, the following plots are obtained.

```
ggplot(data=wd,aes(x=residual.sugar)) + geom_histogram() +
  scale_x_continuous(lim= c(0.5, quantile(wd$residual.sugar, 0.95))) +
  xlab('residual.sugar(g/dm^3)')
```




```
ggplot(data=wd,aes(x=chlorides)) + geom_histogram() +
  scale_x_continuous(lim= c(0.04, quantile(wd$chlorides, 0.95))) +
  xlab('chlorides (g/dm^3)')
```



Observing the obtained plots, chlorides seems to follow normal distribution now. Residual sugars is nearly normal with some ouliers between 1-4(generally ideal).

Questions

What is the structure of your dataset?

```
str(wd)
```

```
## 'data.frame':    1599 obs. of  14 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 ...
##  $ rating              : Ord.factor w/ 3 levels "bad"<"average"<..: 2 2 2 2 2 2 2 3 3 2 ...
```

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

Yes, I created an ordered factor for rating level and names as 'good', 'poor', 'ideal'.

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

The main feature in the data is quality. I'd like to determine which features determine the quality of wines.

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

The variables related to acidity (fixed, volatile, citric.acid and pH) might explain some of the variance. I suspect the different acid concentrations might alter the taste of the wine. Also, residual.sugar dictates how sweet a wine is and might also have an influence in taste.

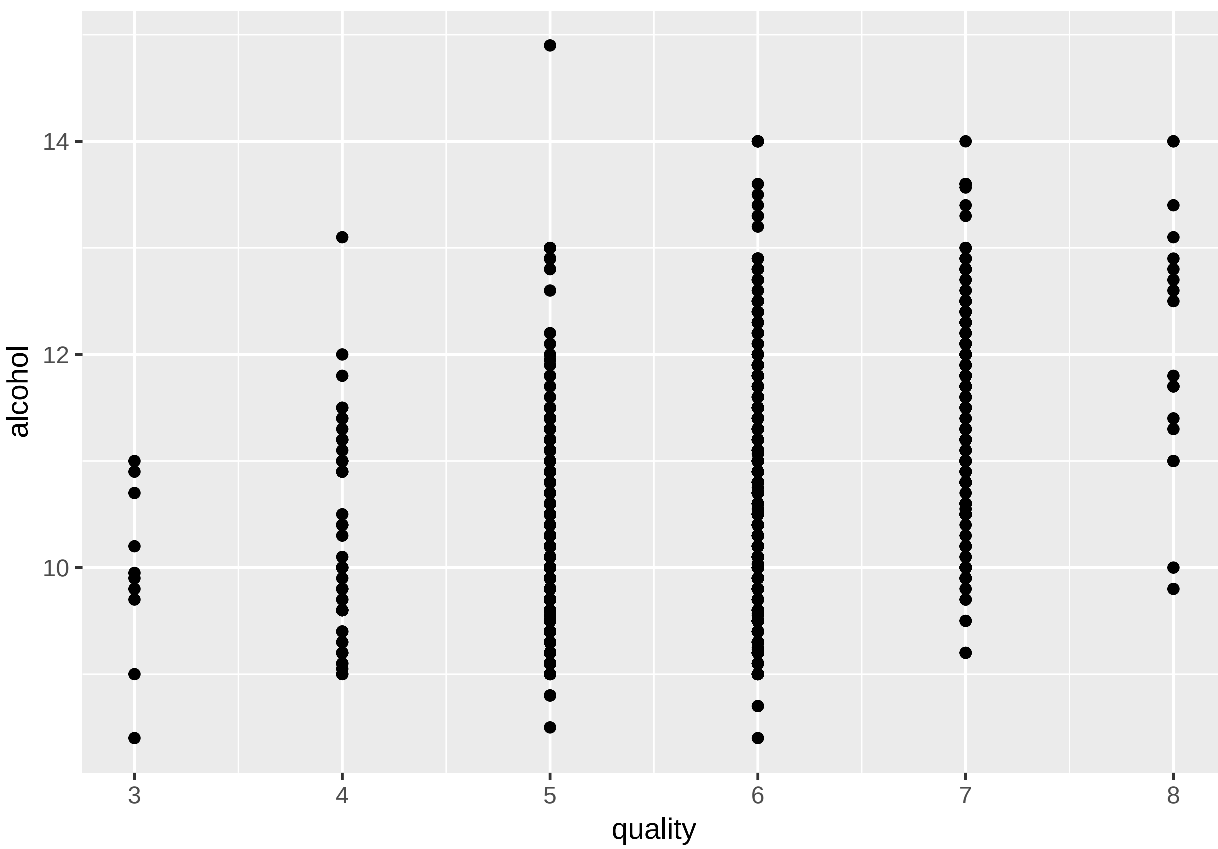
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

Yes there are some distributions that are unusual. I adjusted these plots by taking log10 values for the plots because more accurate trends can be inferred from bivarite plots.

Bivariate Plots

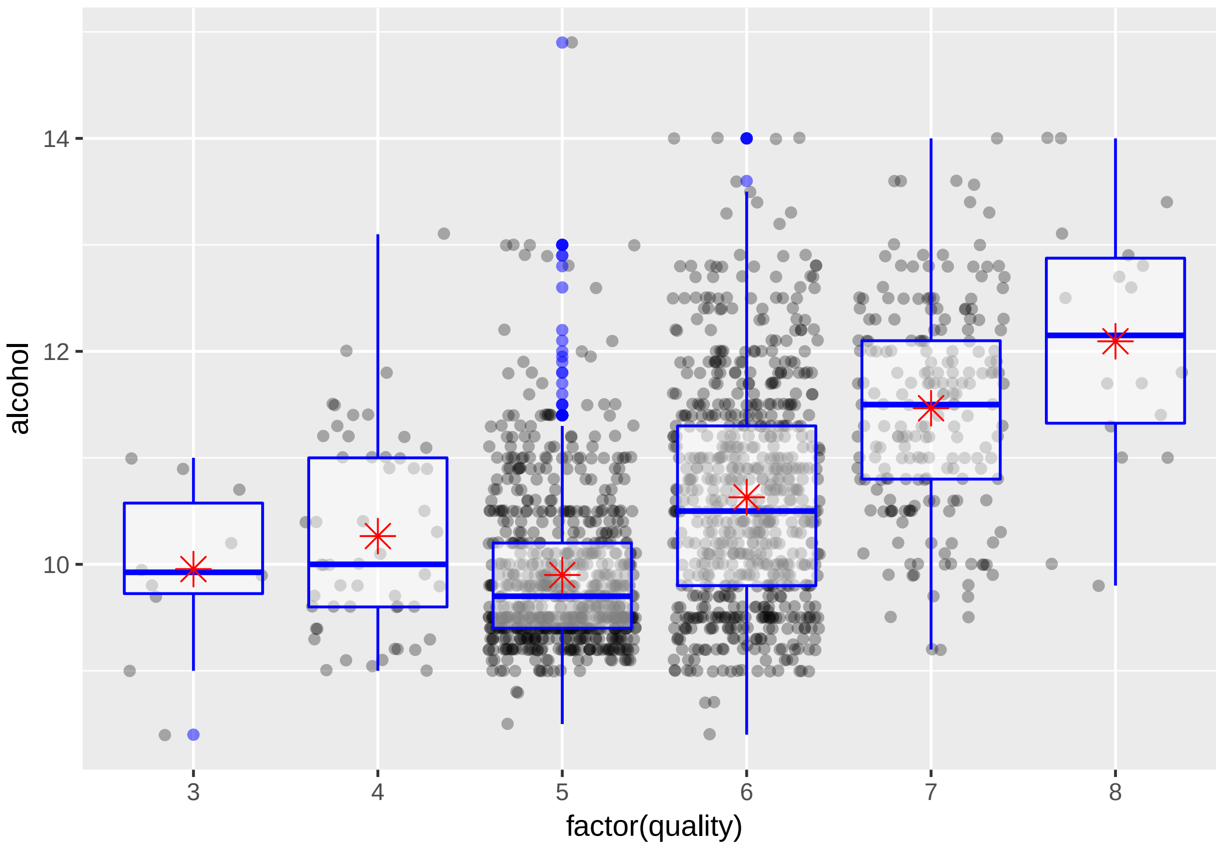
Wine quality has biggest correlation value to wine quality, so lets start with a basic scatter plot of the both.

```
ggplot(aes(x=quality, y=alcohol), data = wd) +
  geom_point()
```



Since the original plot is over crowded with too many points lets add alpha values and 0.1, 0.5 and .09 percentile line to observe the general trends.

```
ggplot(aes(factor(quality),
            alcohol),
        data = wd) +
  geom_jitter( alpha = .3)  +
  geom_boxplot( alpha = .5,color = 'blue')+
  stat_summary(fun.y = "mean",
               geom = "point",
               color = "red",
               shape = 8,
               size = 4)
```



Plot clearly shows trends in increasing wine quality with alcohol content.

Wine Quality in categories

Here box plots are used to represent categorical values.

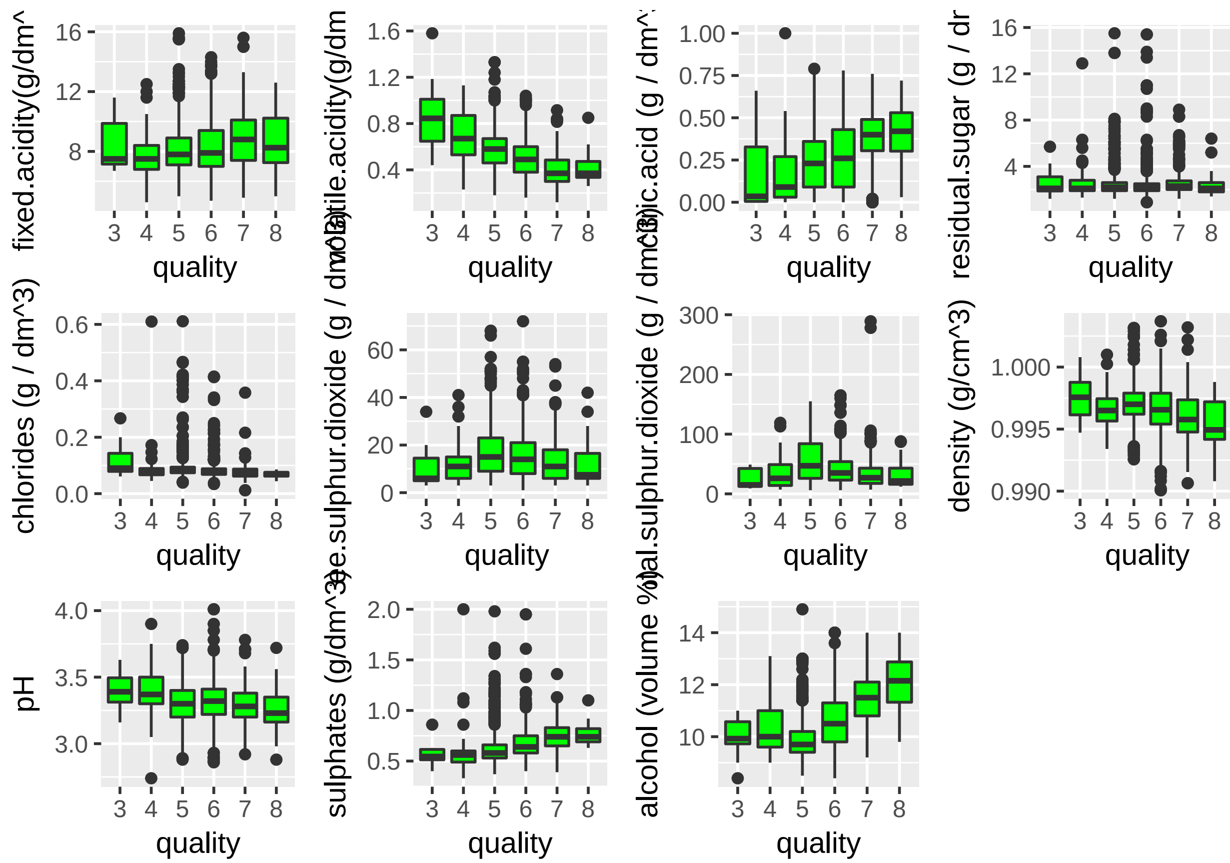
BoxPlot of quality

```
quality_plot <- function (x, y, ylab) {
return (ggplot(data = wd, aes_string(x,y)) +
geom_boxplot(fill = 'green') +
xlab ('quality') + ylab(ylab))
```

```
}

grid.arrange( quality_plot( 'factor(quality)', 'fixed.acidity',
                            'fixed.acidity(g/dm^3)'),
quality_plot('factor(quality)', 'volatile.acidity', 'volatile.acidity(g/dm^3)'),
quality_plot('factor(quality)', 'citric.acid', 'citric.acid (g / dm^3)'),
quality_plot('factor(quality)', 'residual.sugar', 'residual.sugar (g / dm^3)'),
quality_plot('factor(quality)', 'chlorides', 'chlorides (g / dm^3)'),
quality_plot('factor(quality)', 'free.sulfur.dioxide',
             'free.sulphur.dioxide (g / dm^3)'),
quality_plot('factor(quality)', 'total.sulfur.dioxide',
             'total.sulphur.dioxide (g / dm^3)'),
quality_plot('factor(quality)', 'density', 'density (g/cm^3)'),
quality_plot('factor(quality)', 'pH', 'pH'),
quality_plot('factor(quality)', 'sulphates', 'sulphates (g/dm^3)'),
quality_plot('factor(quality)', 'alcohol', 'alcohol (volume %)'),

ncol= 4)
```



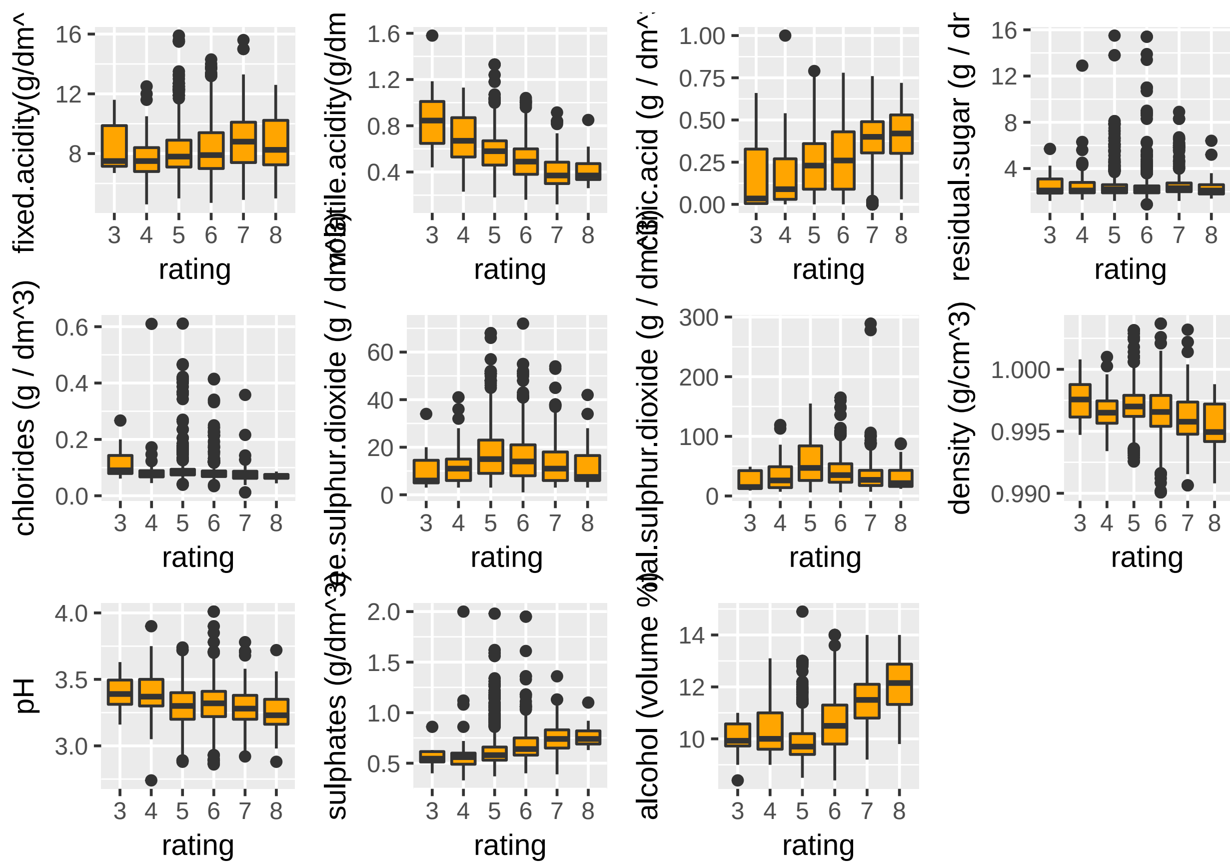
BoxPlot of rating

```
rating_plot <- function(x, y, ylab) {
  return (ggplot(data = wd, aes_string(x, y)) +
   geom_boxplot(fill = 'orange') +
```

```r
  xlab('rating') + ylab(ylab))
}

grid.arrange( rating_plot( 'factor(quality)', 'fixed.acidity', 'fixed.acidity(g/dm^3)'),
rating_plot('factor(quality)', 'volatile.acidity', 'volatile.acidity(g/dm^3)'),
rating_plot('factor(quality)', 'citric.acid', 'citric.acid (g / dm^3)'),
rating_plot('factor(quality)', 'residual.sugar', 'residual.sugar (g / dm^3)'),
rating_plot('factor(quality)', 'chlorides', 'chlorides (g / dm^3)'),
rating_plot('factor(quality)', 'free.sulfur.dioxide', 'free.sulphur.dioxide (g / dm^3)'),
rating_plot('factor(quality)', 'total.sulfur.dioxide', 'total.sulphur.dioxide (g / dm^3)'),
rating_plot('factor(quality)', 'density', 'density (g/cm^3)'),
rating_plot('factor(quality)', 'pH', 'pH'),
rating_plot('factor(quality)', 'sulphates', 'sulphates (g/dm^3)'),
rating_plot('factor(quality)', 'alcohol', 'alcohol (volume %)'),

ncol= 4)
```



Observing the above plots some things can be inferred for a good wine,

- Higher sulphur.dioxide and volatile.acidity,
- Lower pH,
- Higher density,
- lower fixed.acidity and citric.acid.

Correlation of varaiables

Correlation of variables against quality is calculated to further explore,

```
correlations <- c(

  cor.test(wd$fixed.acidity, wd$quality)$estimate,
  cor.test(wd$volatile.acidity, wd$quality)$estimate,
  cor.test(wd$citric.acid, wd$quality)$estimate,
  cor.test(log10(wd$residual.sugar), wd$quality)$estimate,
  cor.test(log10(wd$chlorides), wd$quality)$estimate,
  cor.test(wd$free.sulfur.dioxide, wd$quality)$estimate,
  cor.test(wd$total.sulfur.dioxide, wd$quality)$estimate,
  cor.test(wd$density, wd$quality)$estimate,
  cor.test(wd$pH, wd$quality)$estimate,
  cor.test(log10(wd$sulphates), wd$quality)$estimate,
  cor.test(wd$alcohol, wd$quality)$estimate,
  cor.test(wd$alcohol,wd$pH)$estimate)
names(correlations) <- c('fixed.acidity','volatile.acidity','citric.acid','log10.residual.sugar', 'log1
correlations
```

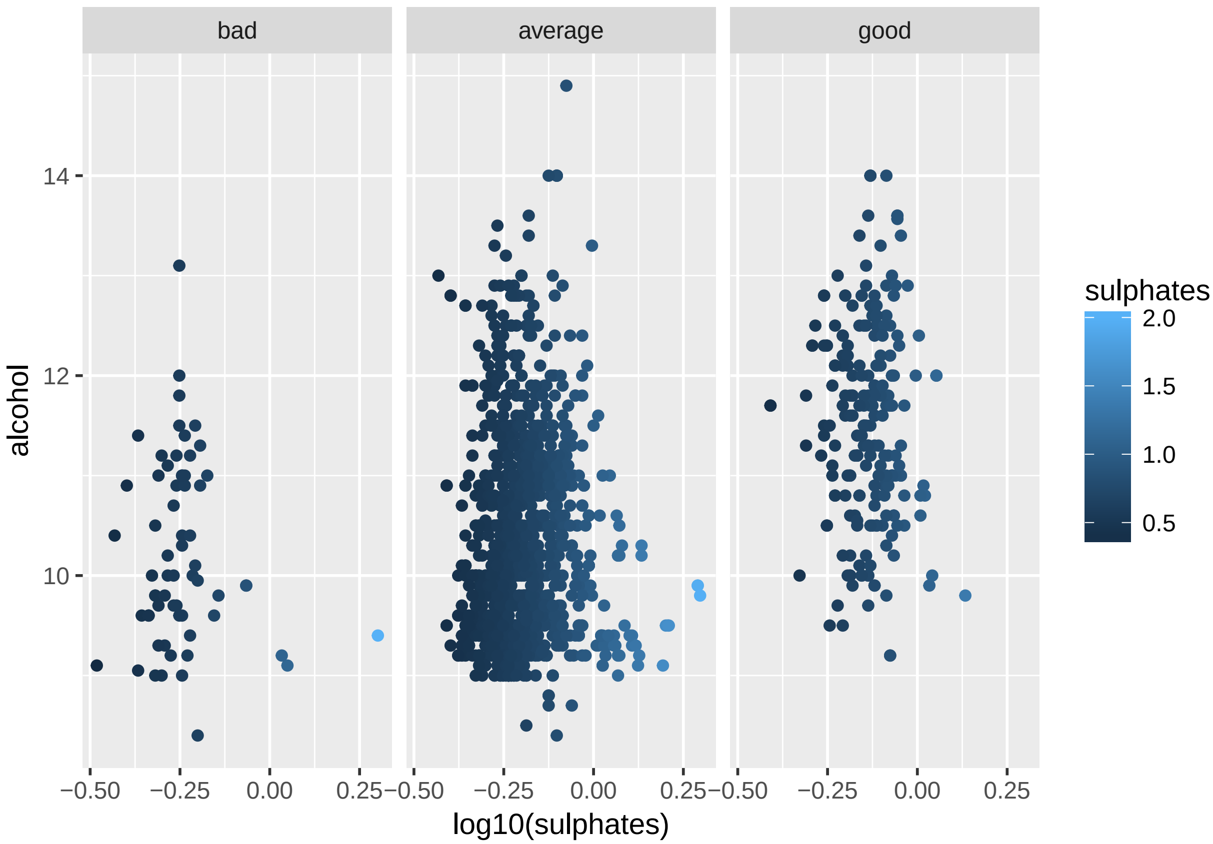
```
##        fixed.acidity     volatile.acidity          citric.acid
##           0.12405165          -0.39055778           0.22637251
## log10.residual.sugar      log10.chlordies  free.sulfur.dioxide
##           0.02353331          -0.17613996          -0.05065606
## total.sulfur.dioxide              density                   pH
##          -0.18510029          -0.17491923          -0.05773139
##      log10.sulphates              alcohol        alochol vs pH
##           0.30864193           0.47616632           0.20563251
```

Observing the above results following show a strong correaltion with quality,

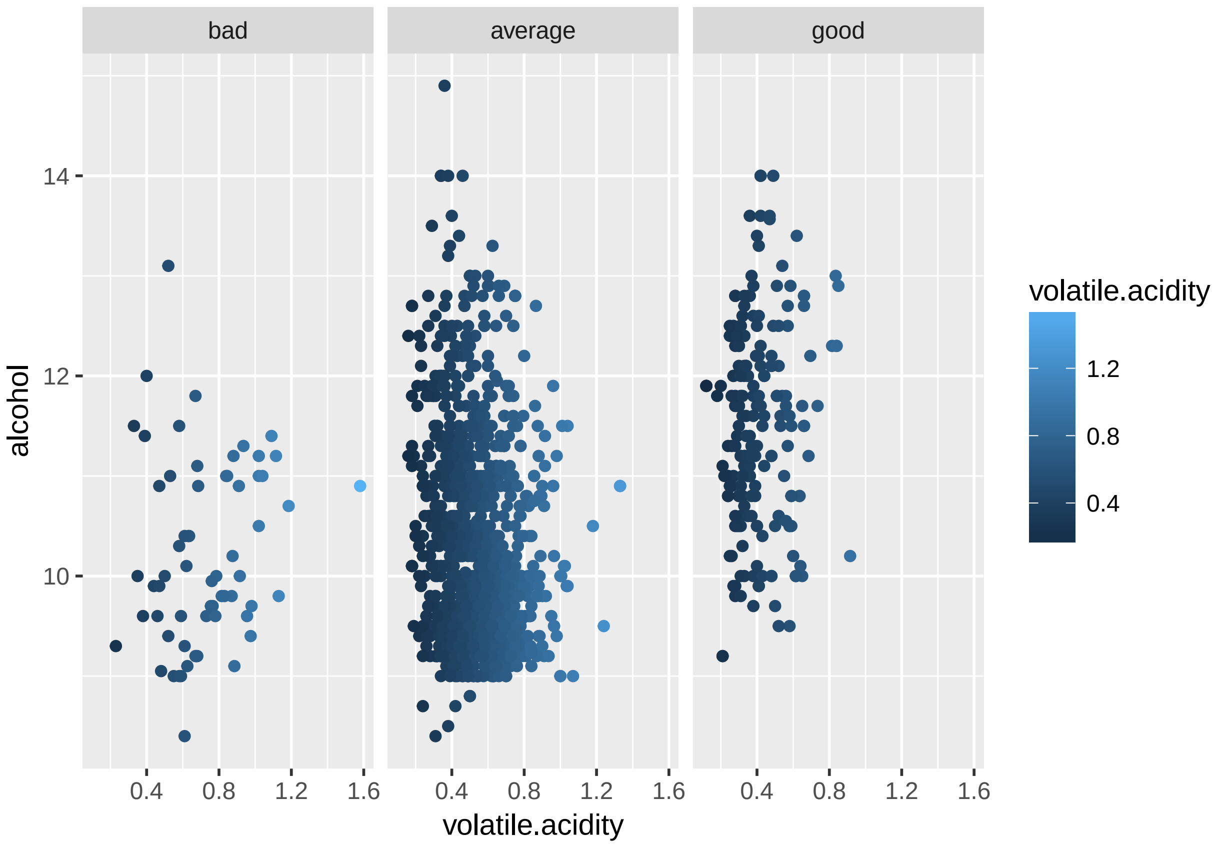
- alcohal
- sulphates
- citric.acid
- fixed.acidity

To further explore lets plot these highly correlated variables with rating:

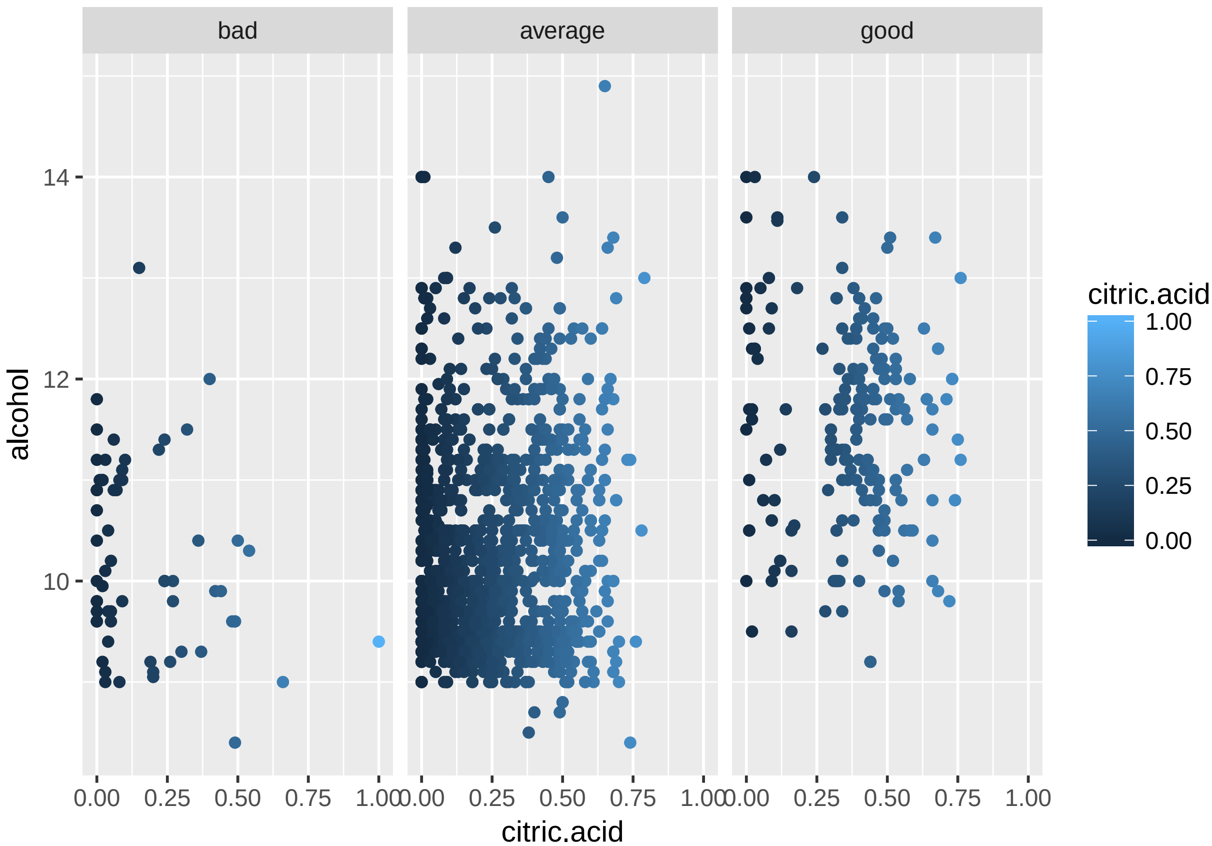
```
ggplot( data = wd, aes(x= log10(sulphates), y= alcohol, color =sulphates )) +
  facet_wrap(~rating) +
  geom_point()
```



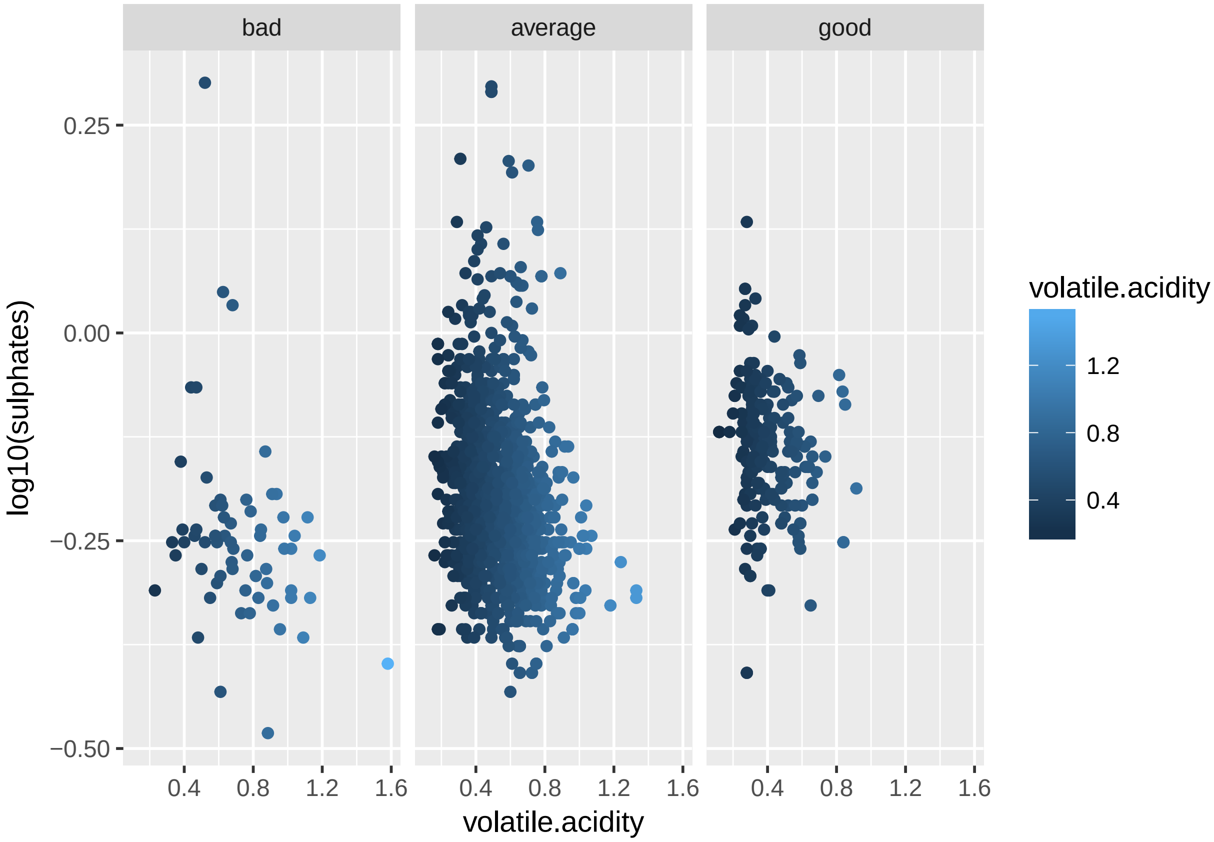

```r
ggplot(data = wd, aes(x = volatile.acidity, y = alcohol, color =volatile.acidity )) +
  facet_wrap(~rating) +
  geom_point()
```



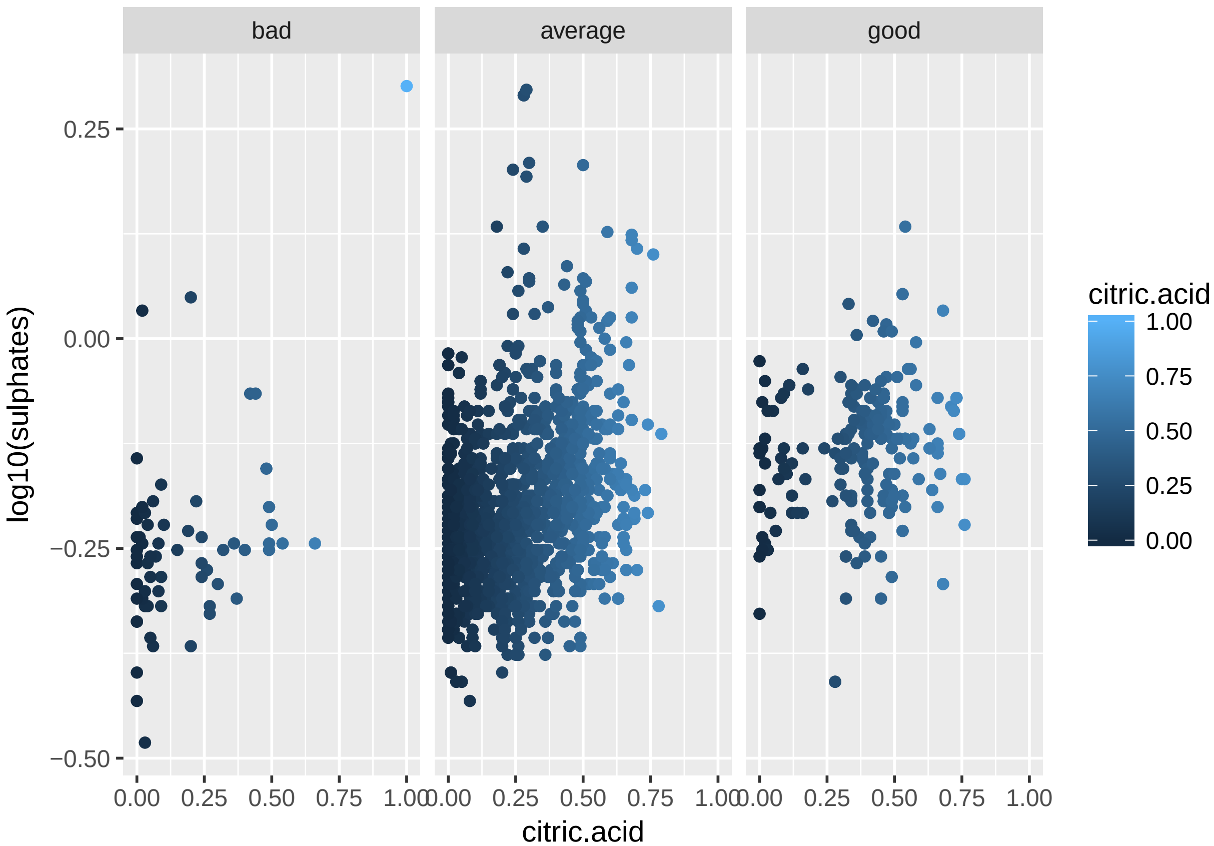

```
ggplot(data = wd, aes(x = citric.acid, y = alcohol, color = citric.acid)) +
  facet_wrap(~rating) +
  geom_point()
```



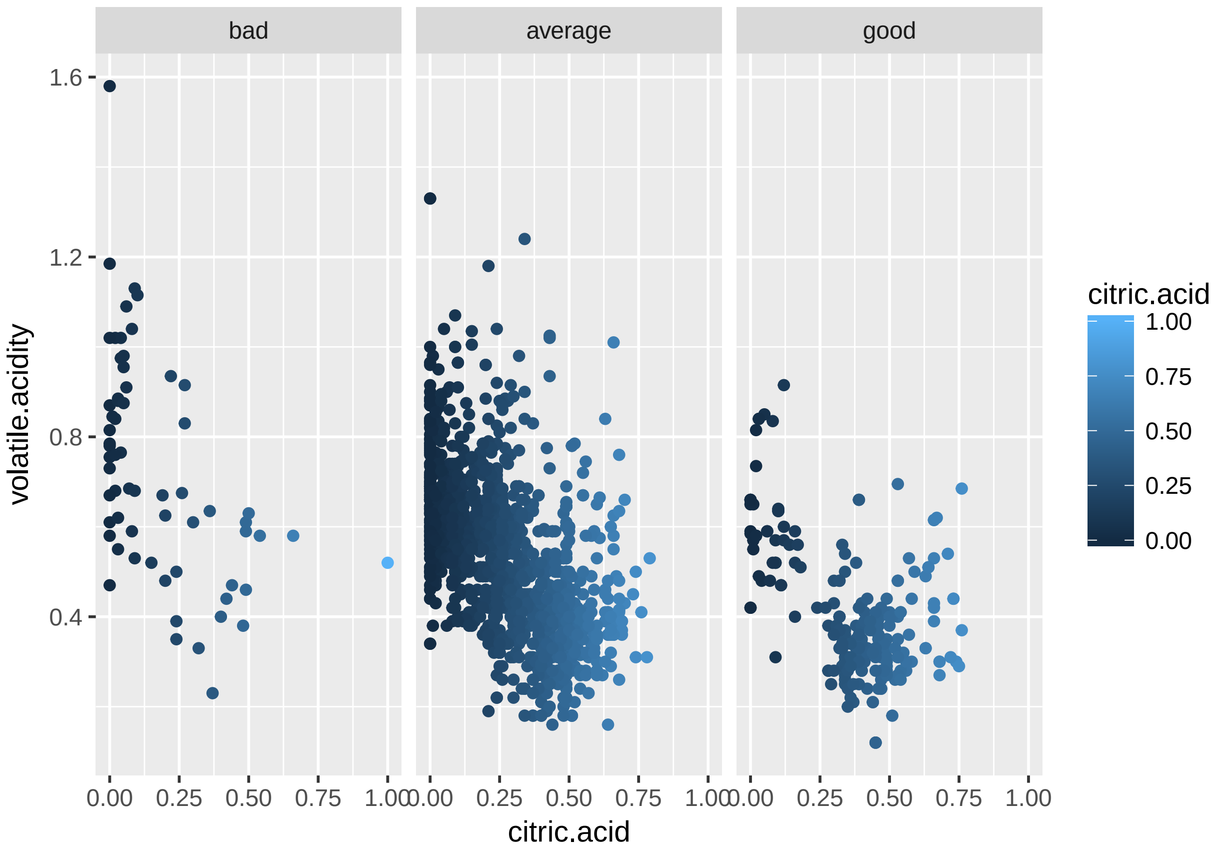

```
ggplot(data = wd, aes(x = volatile.acidity, y = log10(sulphates), color =volatile.acidity )) +
  facet_wrap(~rating) +
  geom_point()
```




```r
ggplot(data = wd, aes(x = citric.acid, y = log10(sulphates), color = citric.acid)) +
  facet_wrap(~rating) +
  geom_point()
```




```r
ggplot(data = wd, aes(x = citric.acid, y = volatile.acidity, color = citric.acid)) +
  facet_wrap(~rating) +
  geom_point()
```



From the above plots only one thing is clear: alcohol content heavely effects rating.

** 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? **

- Fixed.acidity seems to have little to no effect on quality.
- Quality seems to go up when volatile.acidity goes down. The higher ranges seem to produce more average and poor wines.
- Better wines tend to have higher concentration of citric acid.
- Contrary to what I initially expected residual.sugar apparently seems to have little to no effect on perceived quality. -Altough weakly correlated, a lower concentration of chlorides seem to produce better wines. -Better wines tend to have lower densities. -In terms of pH it seems better wines are more acid but there were many outliers. Better wines also seem to have a higher concentration of sulphates. -Alcohol graduation has a strong correlation with quality, but like the linear model showed us it cannot explain all the variance alone. We're going to need to look at the other variables to generate a better model.

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

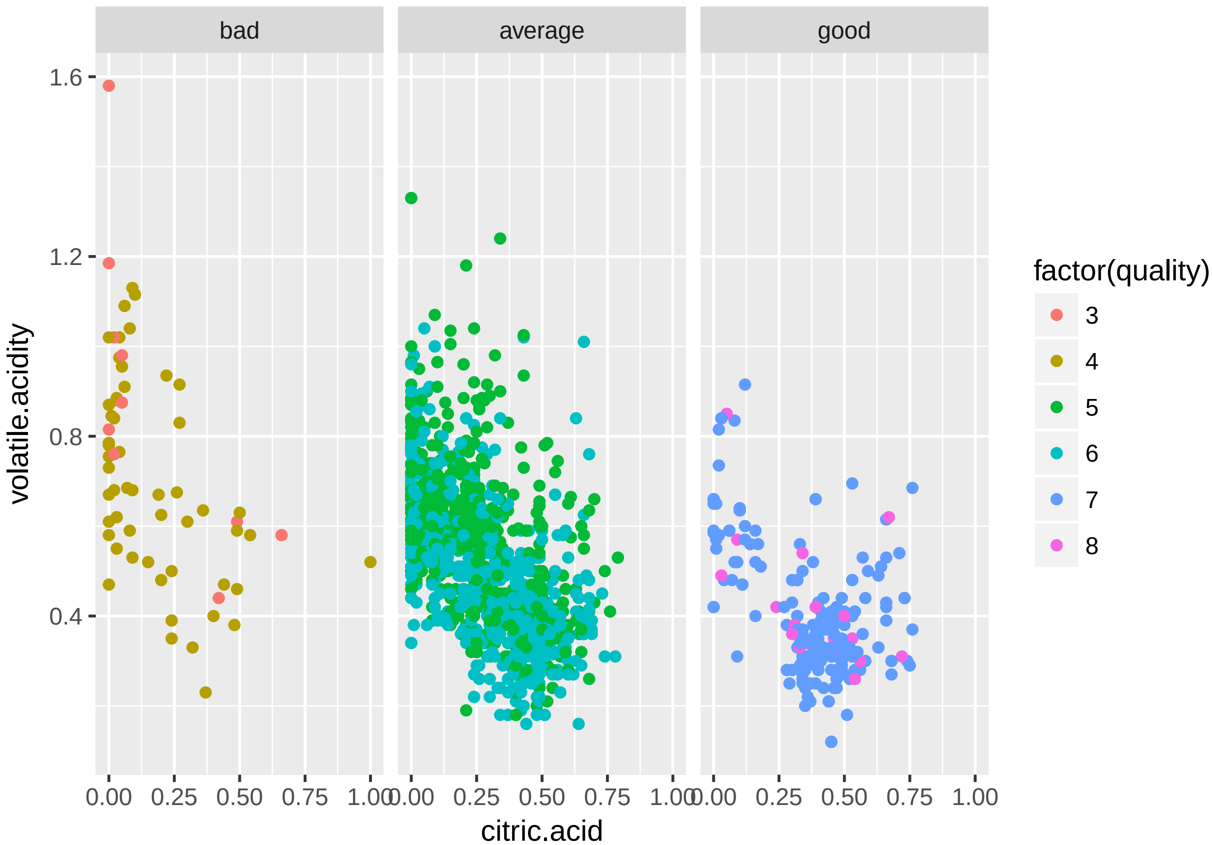
Volatile.acidity surprised me with a positive coefficient for the linear model.

** What was the strongest relationship you found? **

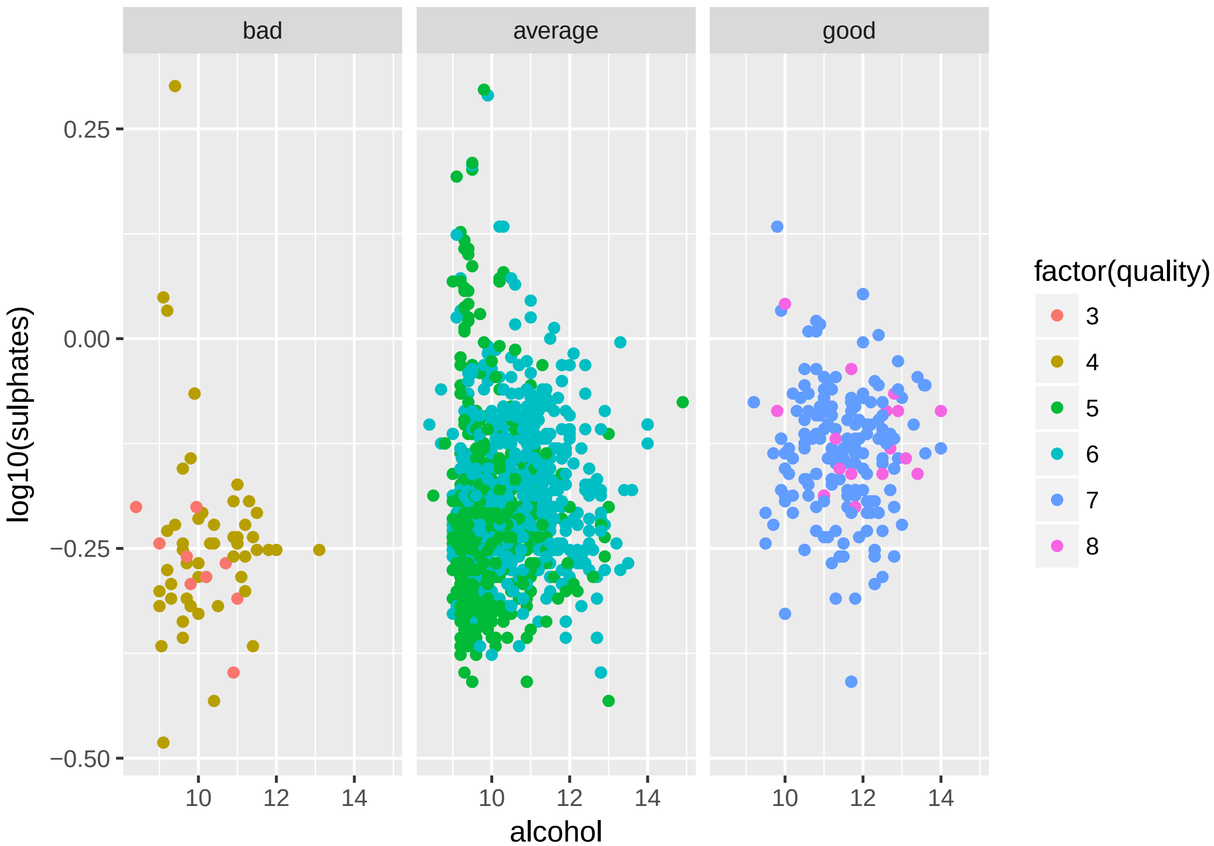
The relationship between the variables total.sulfur.dioxide and free.sulfur.dioxide.

Multivariate Plots

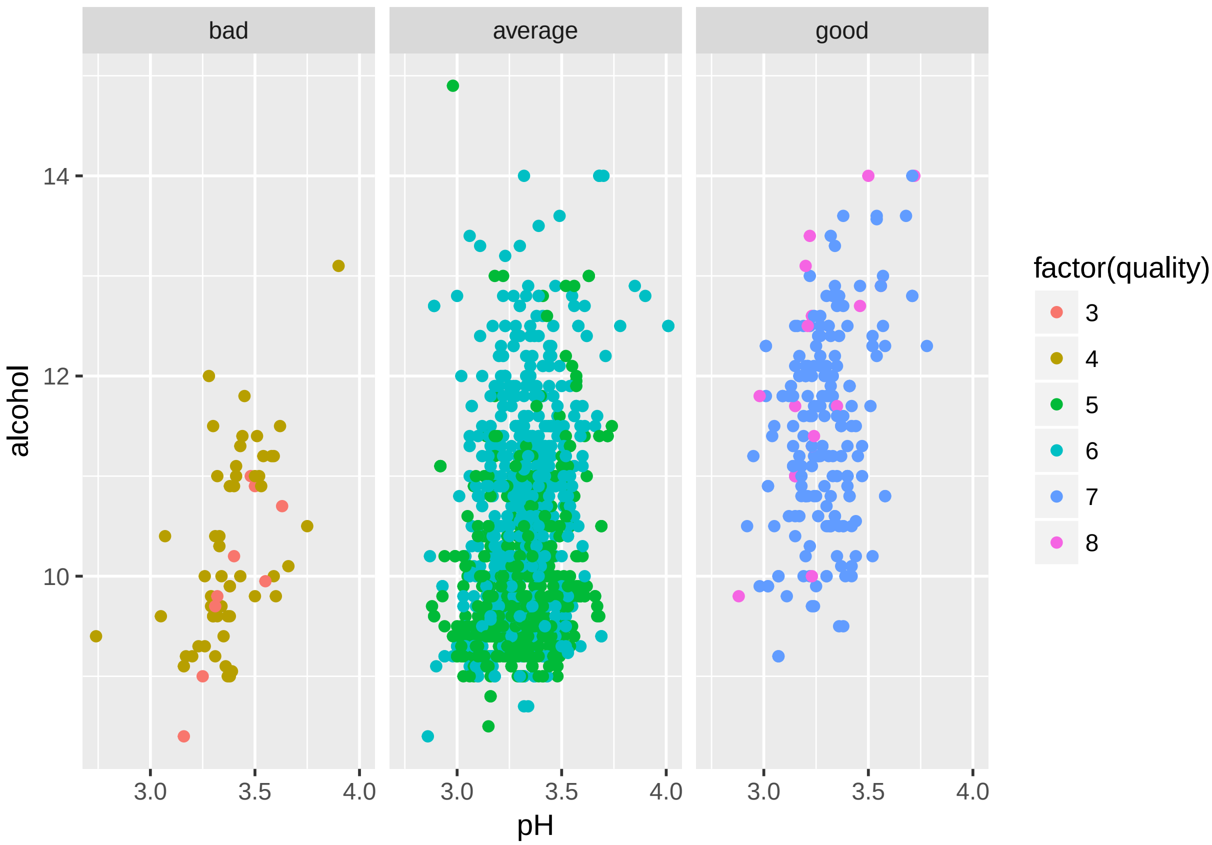
```
ggplot(data = wd,
       aes(x = citric.acid, y = volatile.acidity,
           color = factor(quality))) +
  geom_point() +
  facet_wrap(~rating)
```




```
ggplot(data = wd,
       aes(x = alcohol, y = log10(sulphates),
           color = factor(quality))) +
  geom_point() +
  facet_wrap(~rating)
```




```
ggplot(data = wd,
       aes(x = pH, y = alcohol, color = factor(quality))) +
  geom_point() +
  facet_wrap(~rating)
```



** 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? ** High alcohol contents and high sulphate concentrations seems to produce better wine.

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

Density and alcohol had a stronger negative correlation than others. Adding features to the model that have similar effects probably just overcomplicates the model.

** What was the strongest relationship you found?** The strongest relationship definetly is corelation between pH and fixed acidity.

Analysis

These scatter plots are too crowded so I tried to facet by rating. Graphs between four variables citric.acid, fixed.acidity, sulphates and alcohol which shown high correlations with quality and faceted them with rating. I conclude that higher citric.acid and lower fixed.acidity yields better wines. Better wines also have higher alcohol and sulphates and lower pH.

Linear Multivariable Model

Linear multivariable model was created to predict the wine quality based on chemical properties.

```
# regression
  m1<-lm(quality ~ volatile.acidity,data=wd)
  m2<-update(m1,~. + alcohol)
```

```r
m3<-update(m2,~. + sulphates)
m4<-update(m3,~. + citric.acid)
m5<-update(m4,~. + chlorides)
m6<-update(m5,~. + total.sulfur.dioxide)
m7<-update(m6,~. + density)
mtable(m1,m2,m3,m4,m5,m6,m7)
```

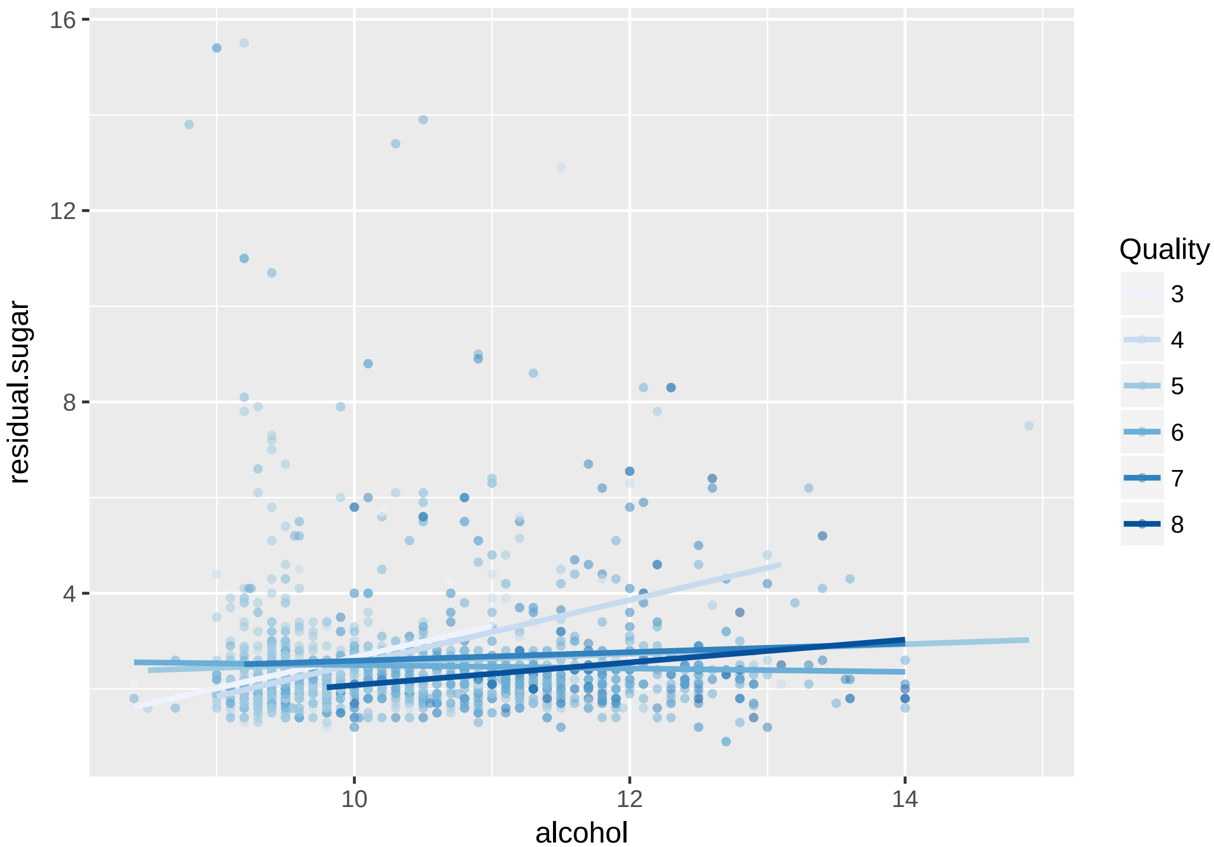
```
##
## Calls:
## m1: lm(formula = quality ~ volatile.acidity, data = wd)
## m2: lm(formula = quality ~ volatile.acidity + alcohol, data = wd)
## m3: lm(formula = quality ~ volatile.acidity + alcohol + sulphates,
##     data = wd)
## m4: lm(formula = quality ~ volatile.acidity + alcohol + sulphates +
##     citric.acid, data = wd)
## m5: lm(formula = quality ~ volatile.acidity + alcohol + sulphates +
##     citric.acid + chlorides, data = wd)
## m6: lm(formula = quality ~ volatile.acidity + alcohol + sulphates +
##     citric.acid + chlorides + total.sulfur.dioxide, data = wd)
## m7: lm(formula = quality ~ volatile.acidity + alcohol + sulphates +
##     citric.acid + chlorides + total.sulfur.dioxide + density,
##     data = wd)
##
## ==========================================================================================
##                               m1           m2           m3           m4           m5           m(
## ------------------------------------------------------------------------------------------
##   (Intercept)                6.566***     3.095***     2.611***     2.646***     2.769***     2.
##                             (0.058)      (0.184)      (0.196)      (0.201)      (0.202)      (0.
##   volatile.acidity          -1.761***    -1.384***    -1.221***    -1.265***    -1.155***    -1.
##                             (0.104)      (0.095)      (0.097)      (0.113)      (0.115)      (0.
##   alcohol                                 0.314***     0.309***     0.309***     0.292***     0.
##                                          (0.016)      (0.016)      (0.016)      (0.016)      (0.
##   sulphates                                            0.679***     0.696***     0.871***     0.
##                                                       (0.101)      (0.103)      (0.111)      (0.
##   citric.acid                                                      -0.079        0.021        0.
##                                                                    (0.104)      (0.106)      (0.
##   chlorides                                                                     -1.663***    -1.
##                                                                                 (0.405)      (0.
##   total.sulfur.dioxide                                                                       -0.
##                                                                                              (0.
##   density
##
## ------------------------------------------------------------------------------------------
##   R-squared                  0.153        0.317        0.336        0.336        0.343        0.
##   adj. R-squared             0.152        0.316        0.335        0.334        0.341        0.
##   sigma                      0.744        0.668        0.659        0.659        0.656        0.
##   F                        287.444      370.379      268.912      201.777      166.407      143.
##   p                          0.000        0.000        0.000        0.000        0.000        0.
##   Log-likelihood         -1794.312    -1621.814    -1599.384    -1599.093    -1590.662    -1580.
##   Deviance                 883.198      711.796      692.105      691.852      684.595      675.
##   AIC                     3594.624     3251.628     3208.768     3210.186     3195.324     3176.
##   BIC                     3610.756     3273.136     3235.654     3242.448     3232.964     3219.
##   N                       1599         1599         1599         1599         1599         1599
## ==========================================================================================
```

The model of 6 features has the lowest AIC (Akaike information criterion) number. As the number of features increase the AIC becomes higher. The parameter of the predictor also changed dramatically which shows a sign of overfitting.

The model can be described as:

wine_quality = 2.985 + 0.276xalcohol - 2.985xvolatile.acidity + 0.908xsulphates + 0.065xcitric.acid - - 1.763*chlorides - 0.002xtotal.sulfur.dioxide

```
ggplot(aes(x = alcohol,
           y = residual.sugar  , color = factor(quality)),
       data = wd) +
      geom_point(alpha = 0.5, size = 1) +
      geom_smooth(method = "lm", se = FALSE,size=1)  +
  scale_color_brewer(type='seq',
                   guide=guide_legend(title='Quality'))
```

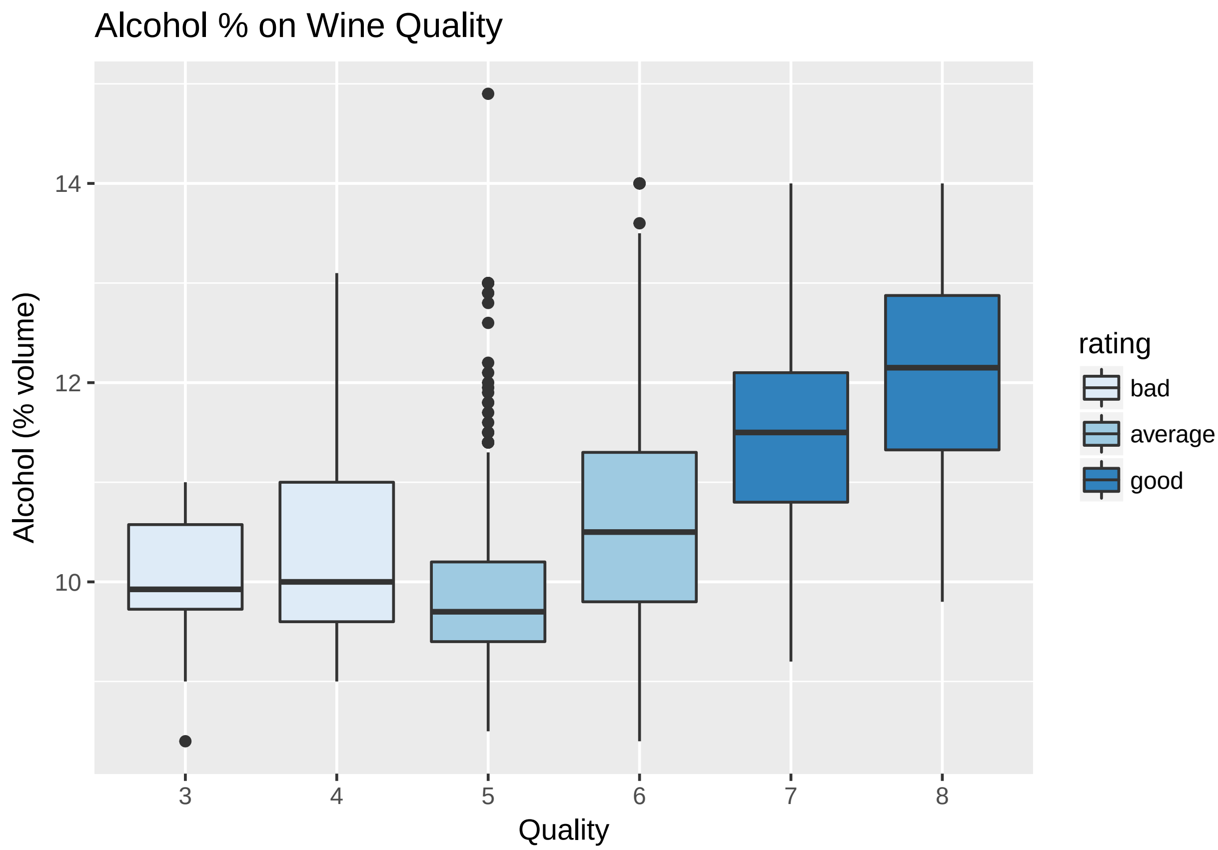


Final Plots and Summary

Alcohol and Wine quality

```
ggplot(data = wd, aes(as.factor(quality), alcohol, fill = rating)) +
  geom_boxplot() +
  ggtitle('Alcohol % on Wine Quality') +
  xlab('Quality') +
```

```
  ylab('Alcohol (% volume)') +
  scale_fill_brewer(type = 'seq', palette = 1)
```

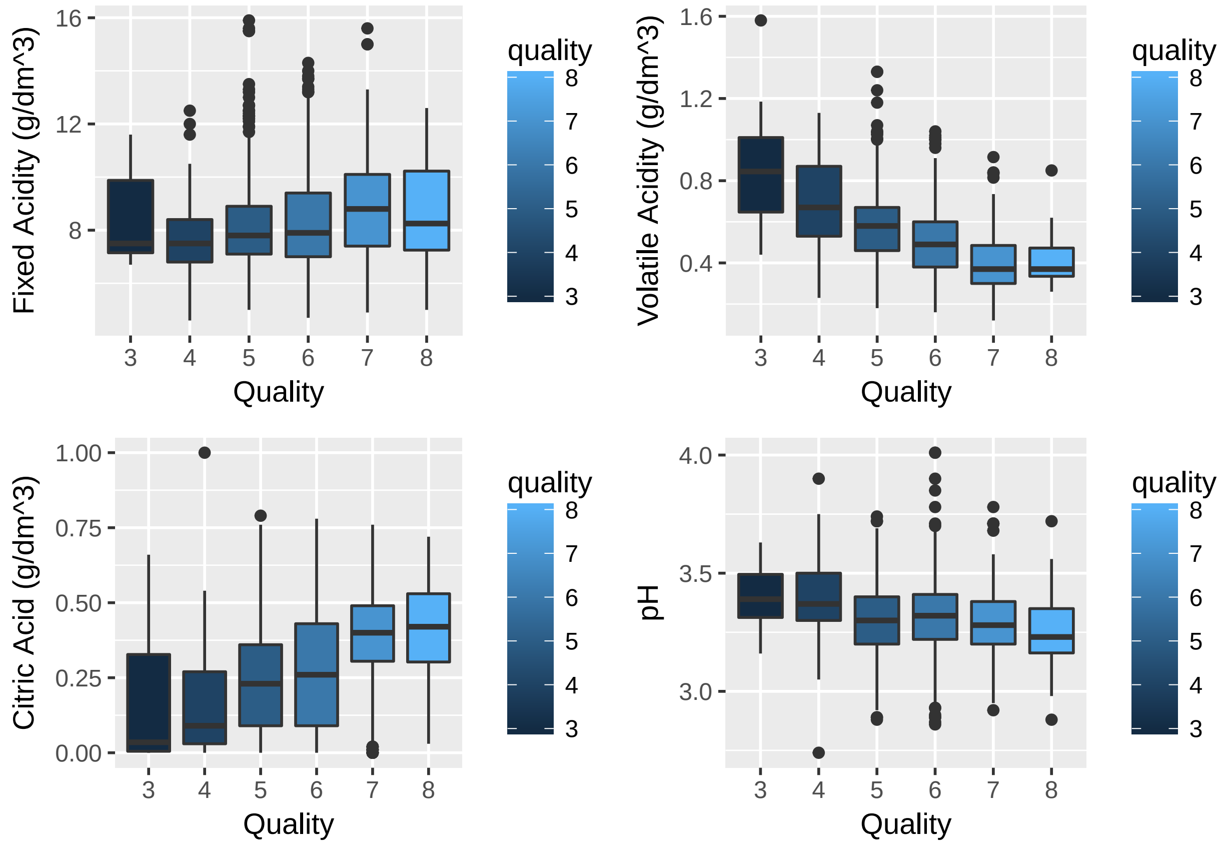


From the above plot it is clear that wine quality increases with % of alcohol in it. Intrestingly the alcohol percentage of higher quality wines(quality> 6) incresed with quality but some lower quality wines doest have the lowest alcohol percentage.

Acids and Wine quality

```
grid.arrange(ggplot(data = wd, aes(x = factor(quality),y =fixed.acidity,
                                   fill = quality)) +
               ylab('Fixed Acidity (g/dm^3)') +
               xlab('Quality') +
               geom_boxplot(),
             ggplot(data = wd, aes(x = factor(quality),y = volatile.acidity,
                                   fill = quality)) +
               ylab('Volatile Acidity (g/dm^3)') +
               xlab('Quality') +
               geom_boxplot(),
             ggplot(data = wd, aes(x = factor(quality), y = citric.acid,
                                   fill = quality)) +
               ylab('Citric Acid (g/dm^3)') +
               xlab('Quality') +
               geom_boxplot(),
             ggplot(data = wd, aes(x = factor(quality), y = pH,
                                   fill = quality)) +
```

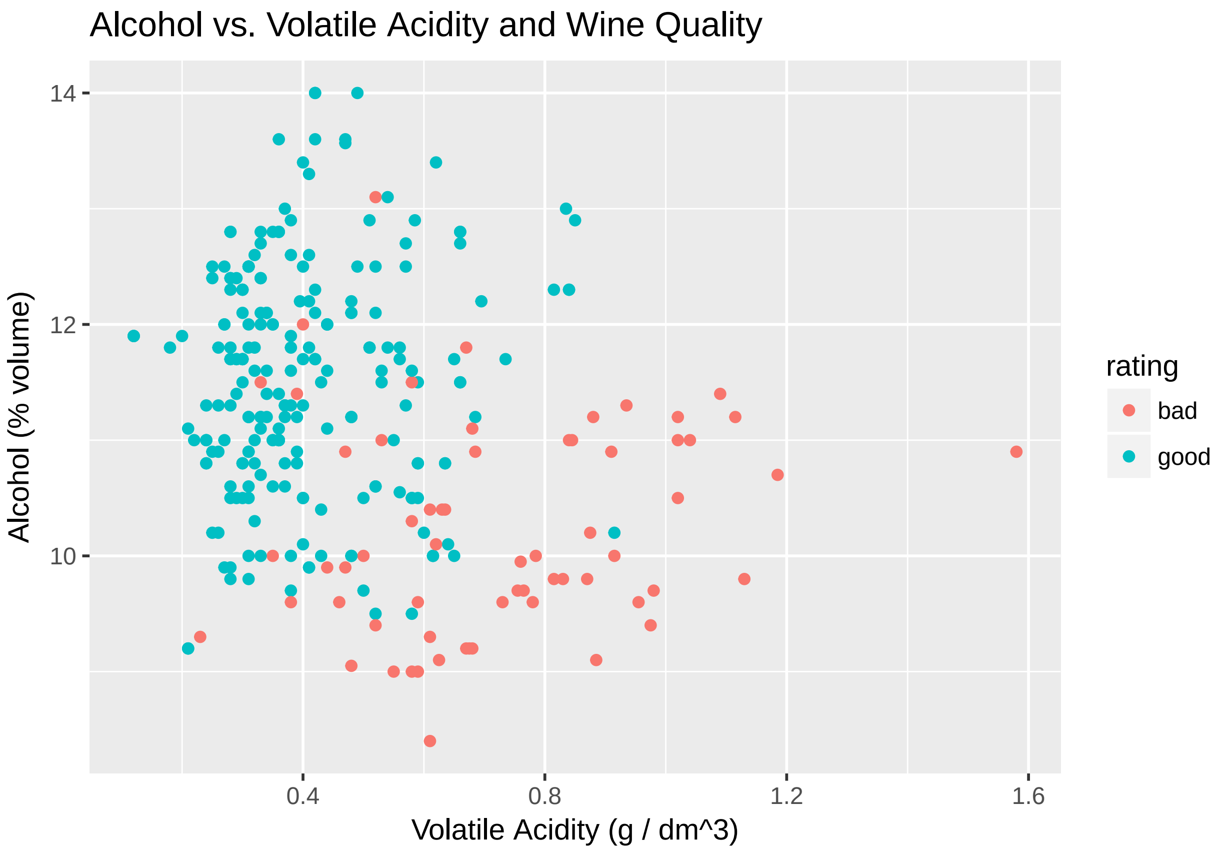
```r
               ylab('pH') +
               xlab('Quality') +
               geom_boxplot())
```



From the above plots it is clear that higher acidic(lower pH) content is seen in highly rated wines and on the contrary low volotalie acidic wines are good quality wines.

Good and Bad wines

```r
ggplot(data = subset(wd, rating != 'average'),
       aes(x = volatile.acidity, y = alcohol,
                      color = rating)) +
  geom_point() +
  ggtitle('Alcohol vs. Volatile Acidity and Wine Quality') +
  xlab('Volatile Acidity (g / dm^3)') +
  ylab('Alcohol (% volume)')
```



Above plots includes only good and bad wines, some things that can be inferred from the plot are:

- High volatile acidity–with few exceptions–kept wine quality down.
- A combination of high alcohol content and low volatile acidity produced better wines.

Reflection

Wine quality depends on many features, through this exploratory data analysis I was able to relate some of the key factors like alcohol content, sulphates, and acidity. The correlations for these variables are within reasonable bounds. The graphs adequately illustrate the factors that make good wines 'good' and bad wines 'bad'. This dataset has 11 physiochemical properties of 1599 red wines. I read up on information about each property so I understood overall implications as I looked at the dataset further. After looking at the distributions of some variables, I looked at the relationship between two- and, eventually, three-variable combinations.

In this data, my main struggle was to get a higher confidence level when predicting factors that are responsible for the production of different quality of wines especially the 'Good' and the 'Bad' ones. As the data was very centralized towards the 'Average' quality, my training set did not have enough data on the extreme edges to accurately build a model which can predict the quality of a wine given the other variables with lesser margin of error. So maybe in future, I can get a dataset about Red Wines with more complete information so that I can build my models more effectively.

For future studies, it would be interesting to mesure more acid types in the analysis. Wikipedia for example, suggests that malic and lactic acid are important in wine taste and these were not included in this sample.

Also, I think it would be interesting to include each wine critic judgement as separate entry in the dataset. After all, each individual has a different taste and is subject to prejudice and other distorting factors. I believe that having this extra information would add more value to the analysis.

References

- http://www.winegeeks.com/articles/85/high_alcohol_is_a_wine_fault_not_a_badge_of_honor/
- http://www.winegeeks.com/articles/85/high_alcohol_is_a_wine_fault_not_a_badge_of_honor/
- https://onlinecourses.science.psu.edu/stat857/node/223
- https://github.com/Dalaska/Udacity-Red-Wine-Quality/blob/master/redwine_final.rmd