RedWine

Anil S Bodepudi

6/25/2017

```
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
```

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

```
## 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 ...
```

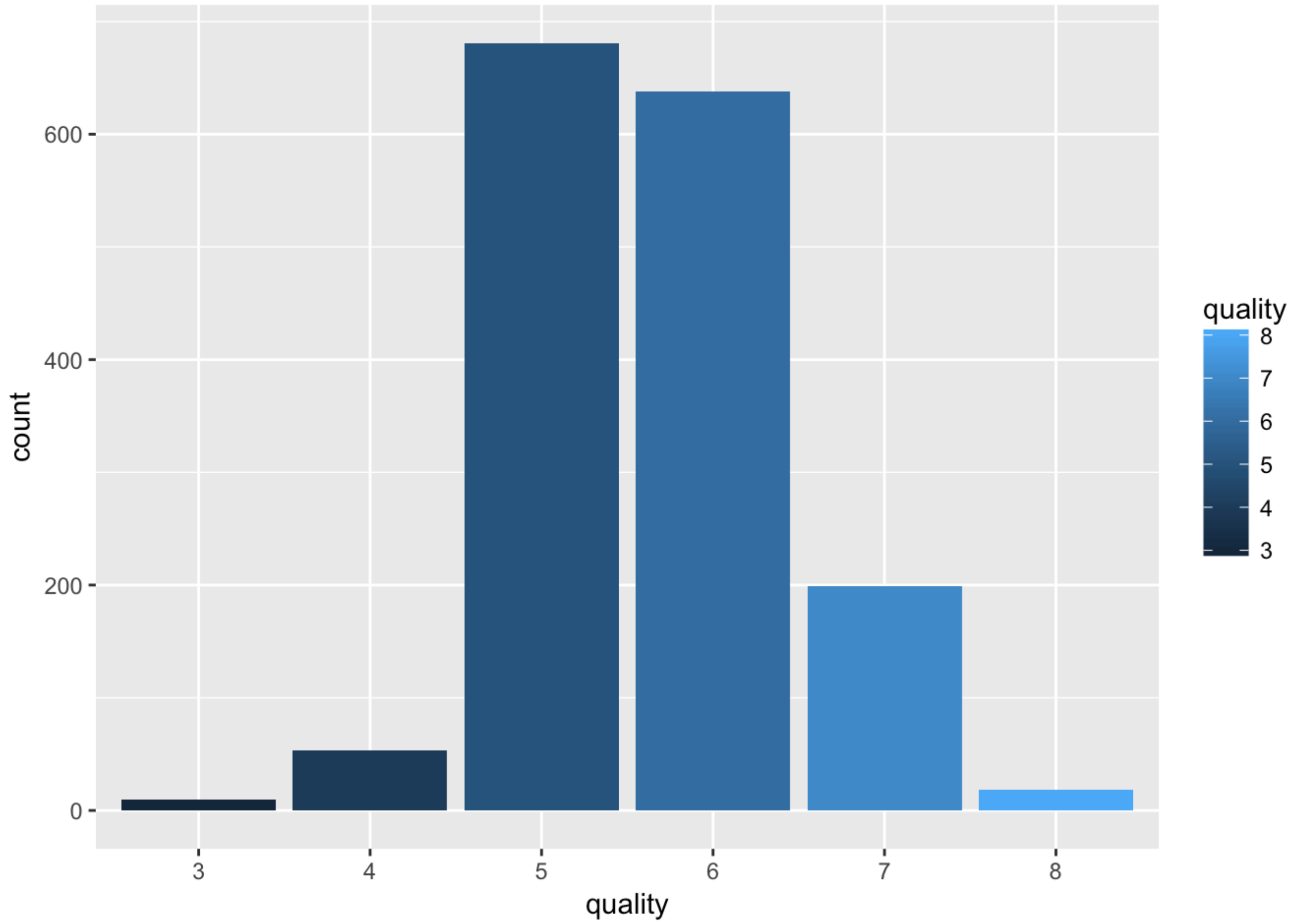
```
summary(wd)
```

```
##        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= quality), data = wd) + geom_bar() +theme_replace() +
 xlab("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

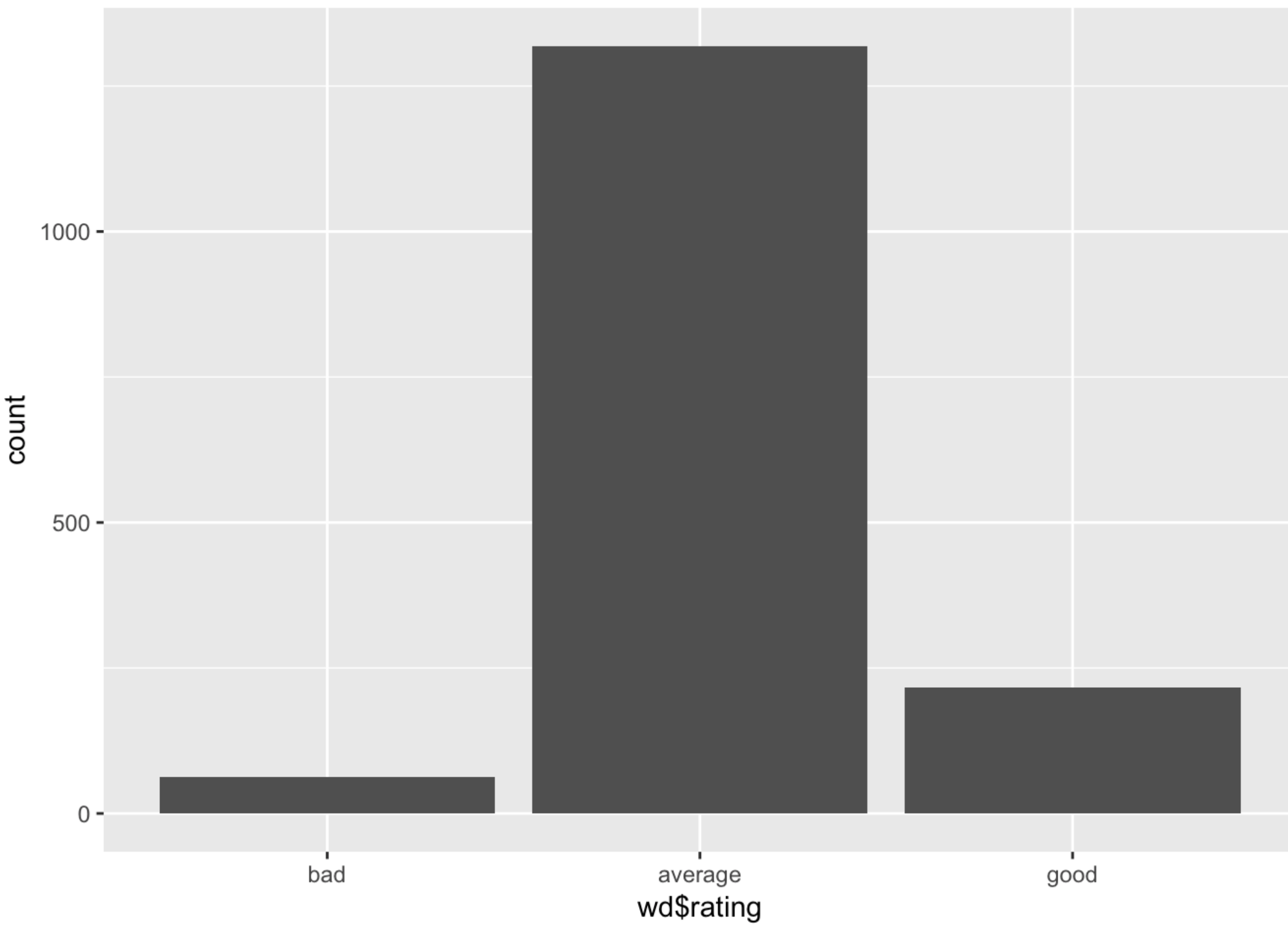
Looking at our first plot of wine quality, it roughly has a normal distribution with most rating being in 5 and 6. So lets create an another variable with variable ratings with following categories.

- 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)
```

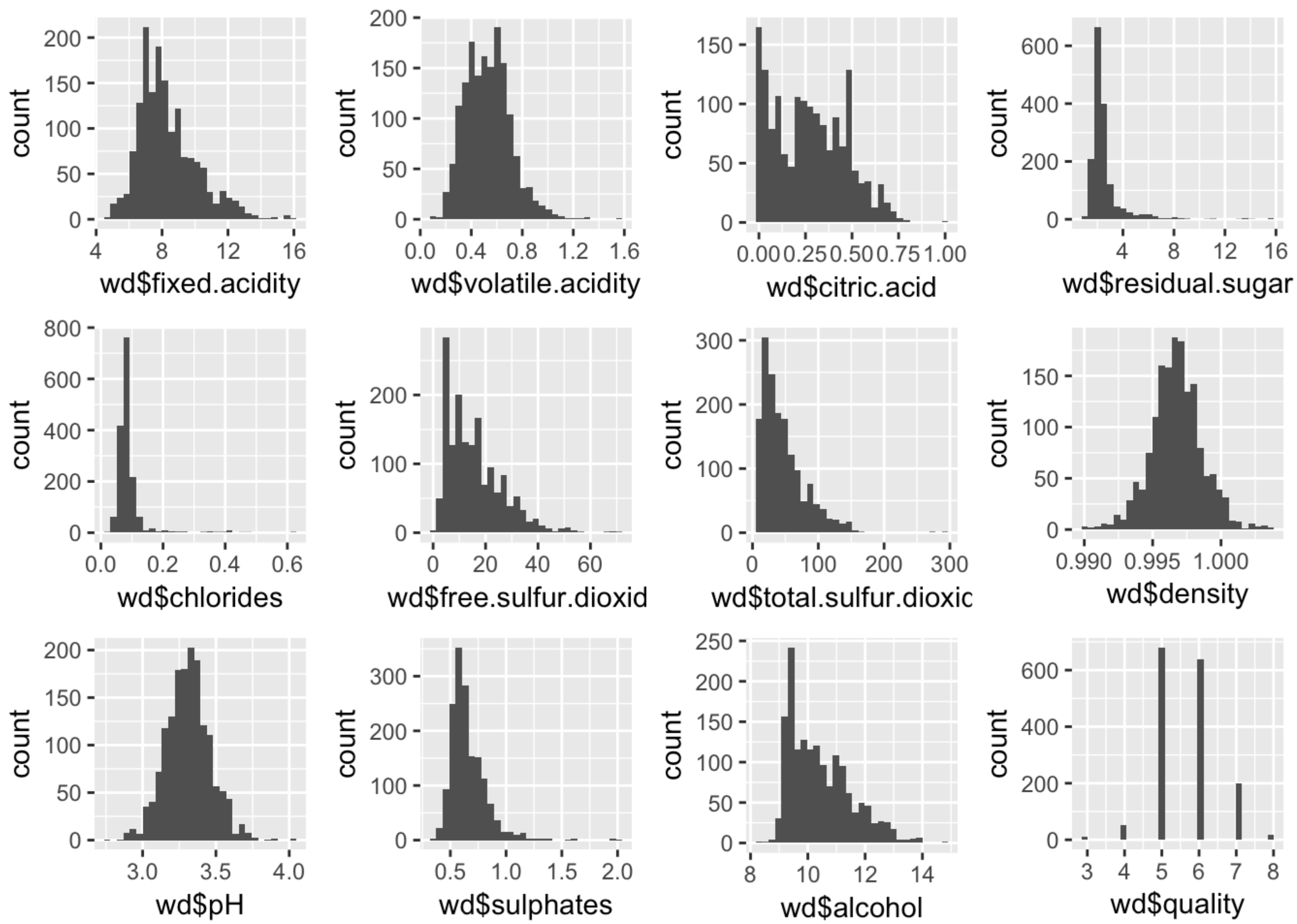
```
##     bad average    good
##      63    1319     217
```

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



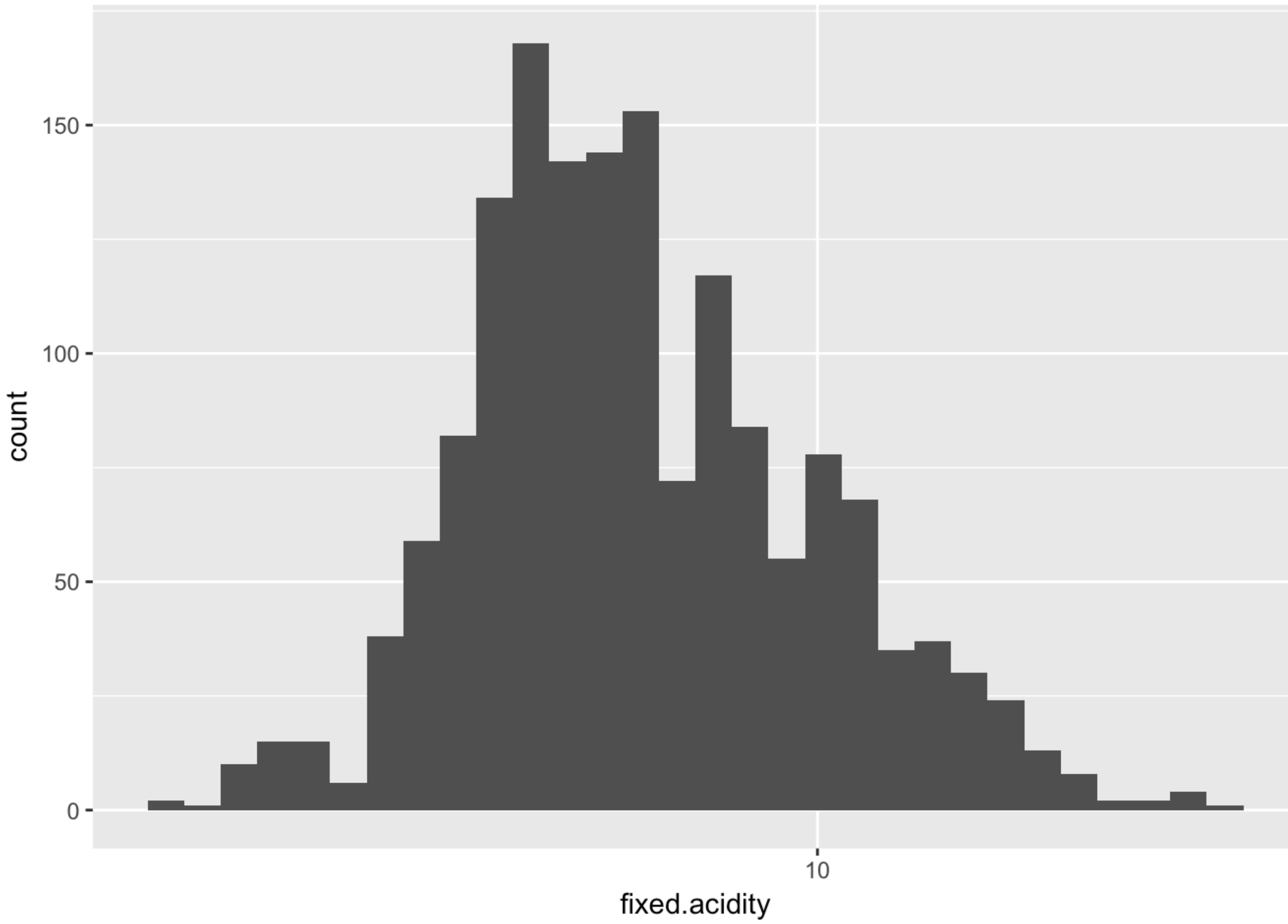
Distribution and Outliers

Looking at the plots above inferred details are as fallows,

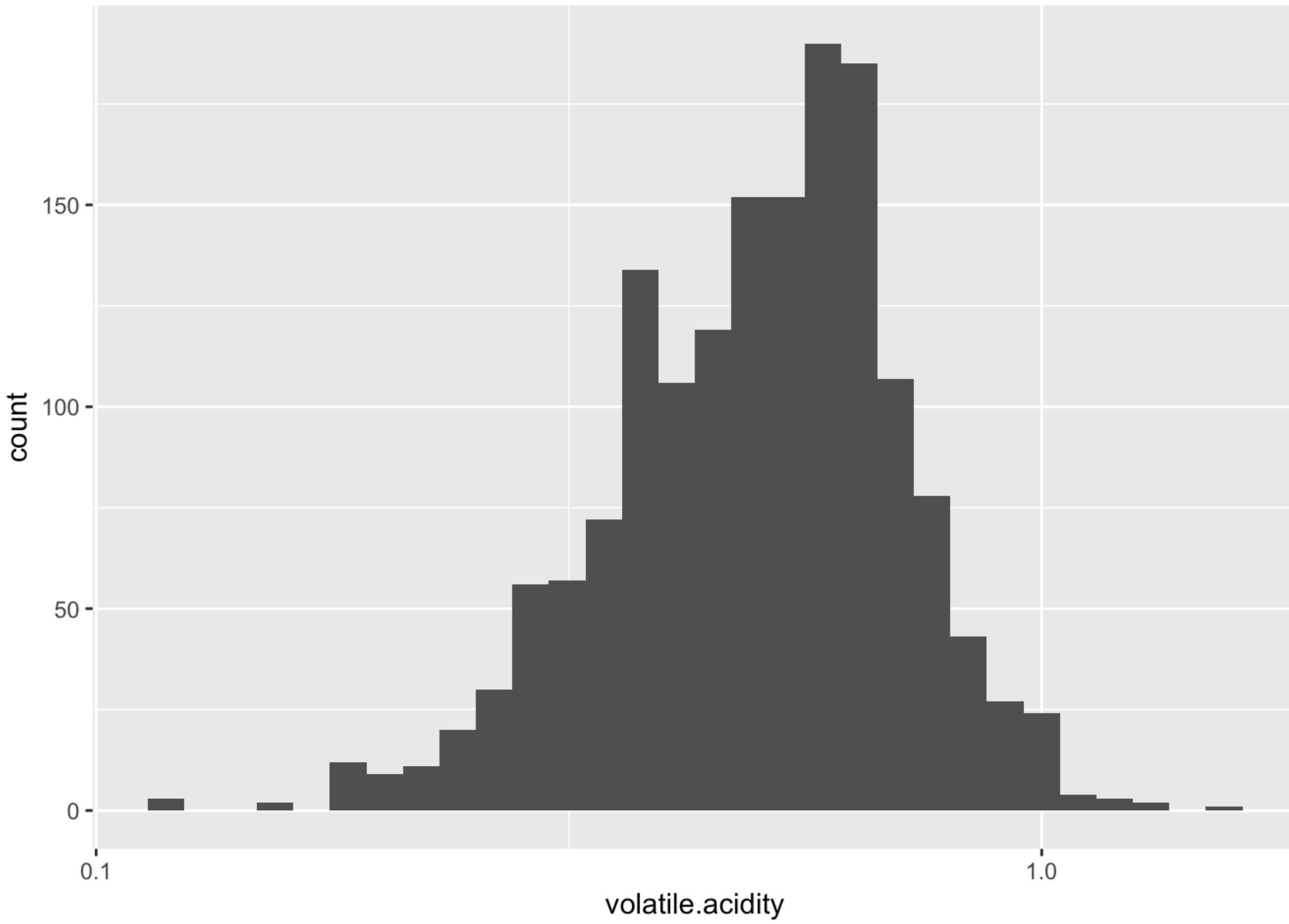
- Density and pH are normally distributed.
- Qualitatively, residual sugar and chlorides have extreme outlines.
- Fixed and volatile acidity, sulfur dioxides, sulphates, and alcohol seem to be long-tailed.
- Citric acid have many zero values,looks like there is some error in reporting but I am curious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

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

```
ggplot(data= wd,aes(x=fixed.acidity))+geom_histogram()+scale_x_log10()
```

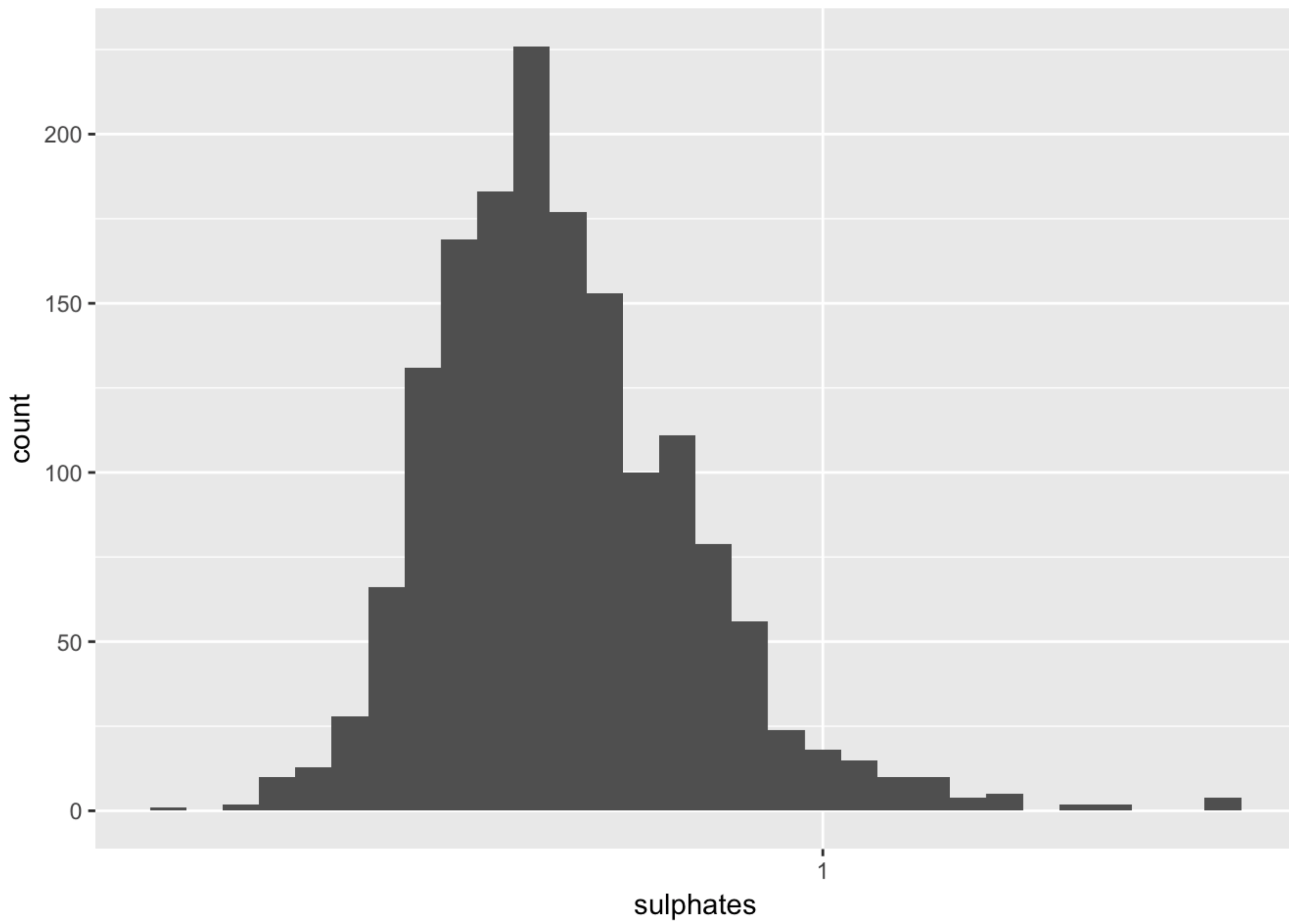



```
ggplot(data= wd,aes(x=volatile.acidity))+geom_histogram()+scale_x_log10()
```

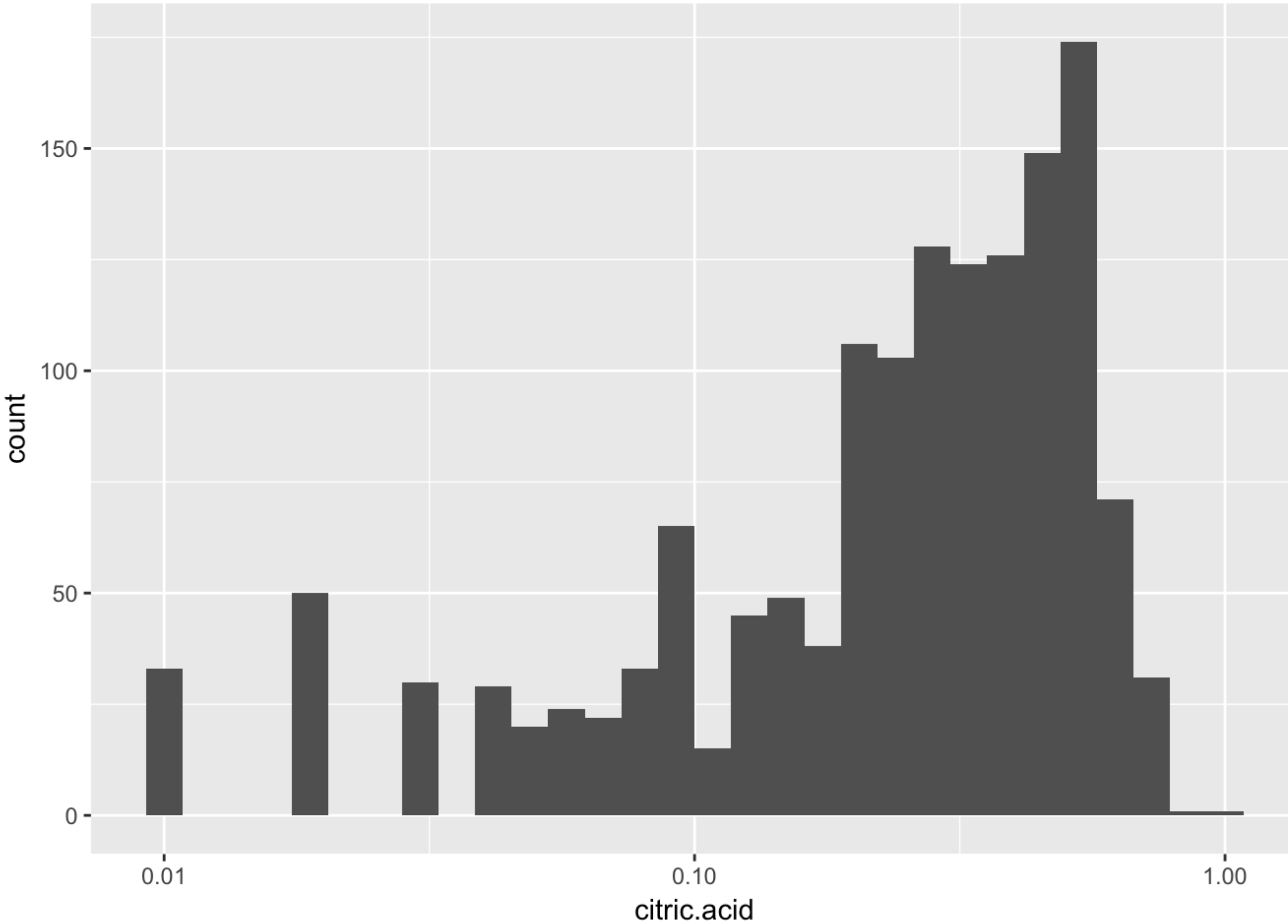


Similarly I plotted citric acid and sulphates to find out if they are normally distributed but found out only sulphates are normally distributed.

```
ggplot(data= wd,aes(x=sulphates))+geom_histogram()+scale_x_log10()
```




```
ggplot(data= wd,aes(x=citric.acid))+geom_histogram()+scale_x_log10()
```



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

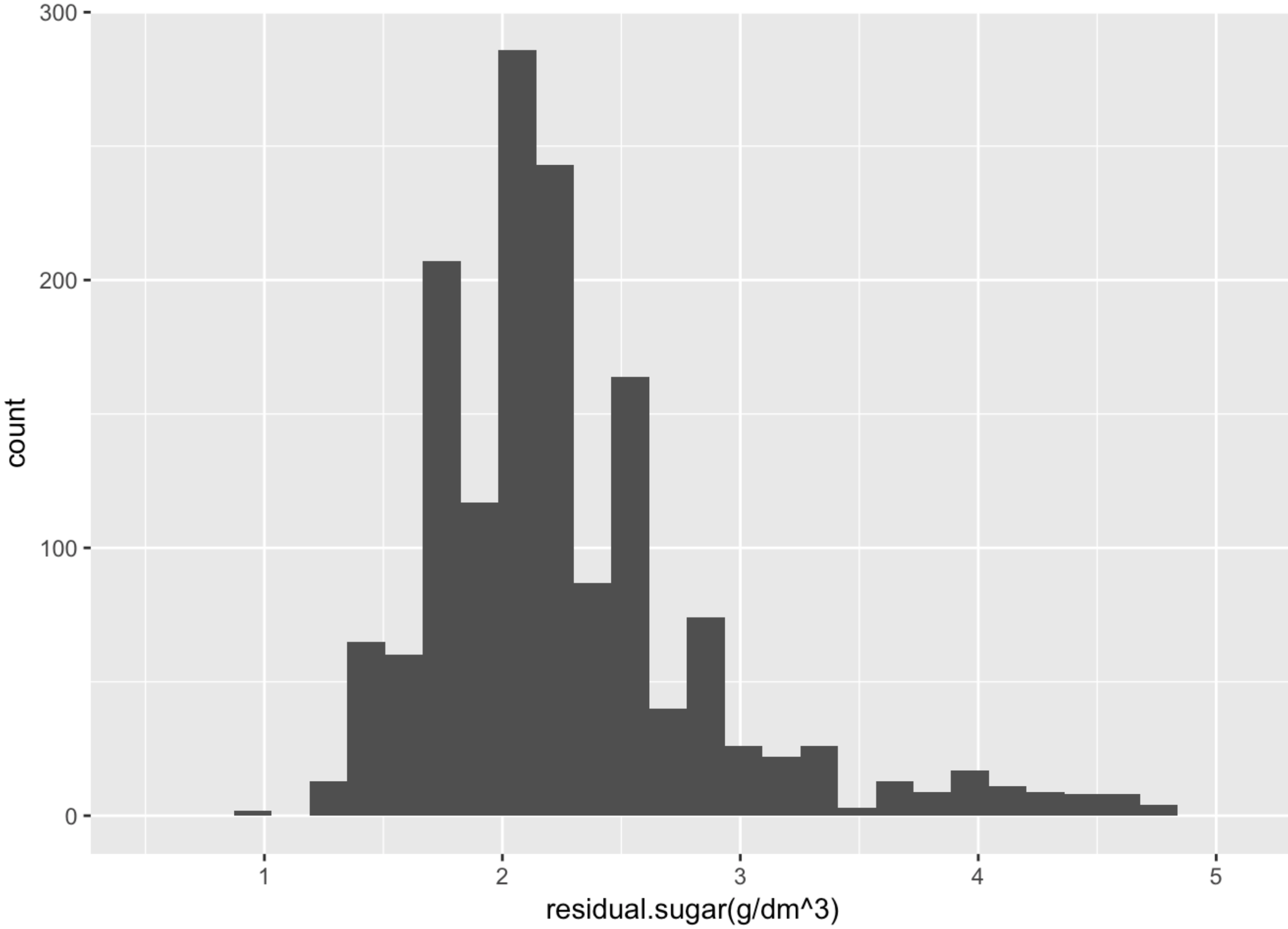
```
length(subset(wd, citric.acid==0)$citric.acid)
```

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

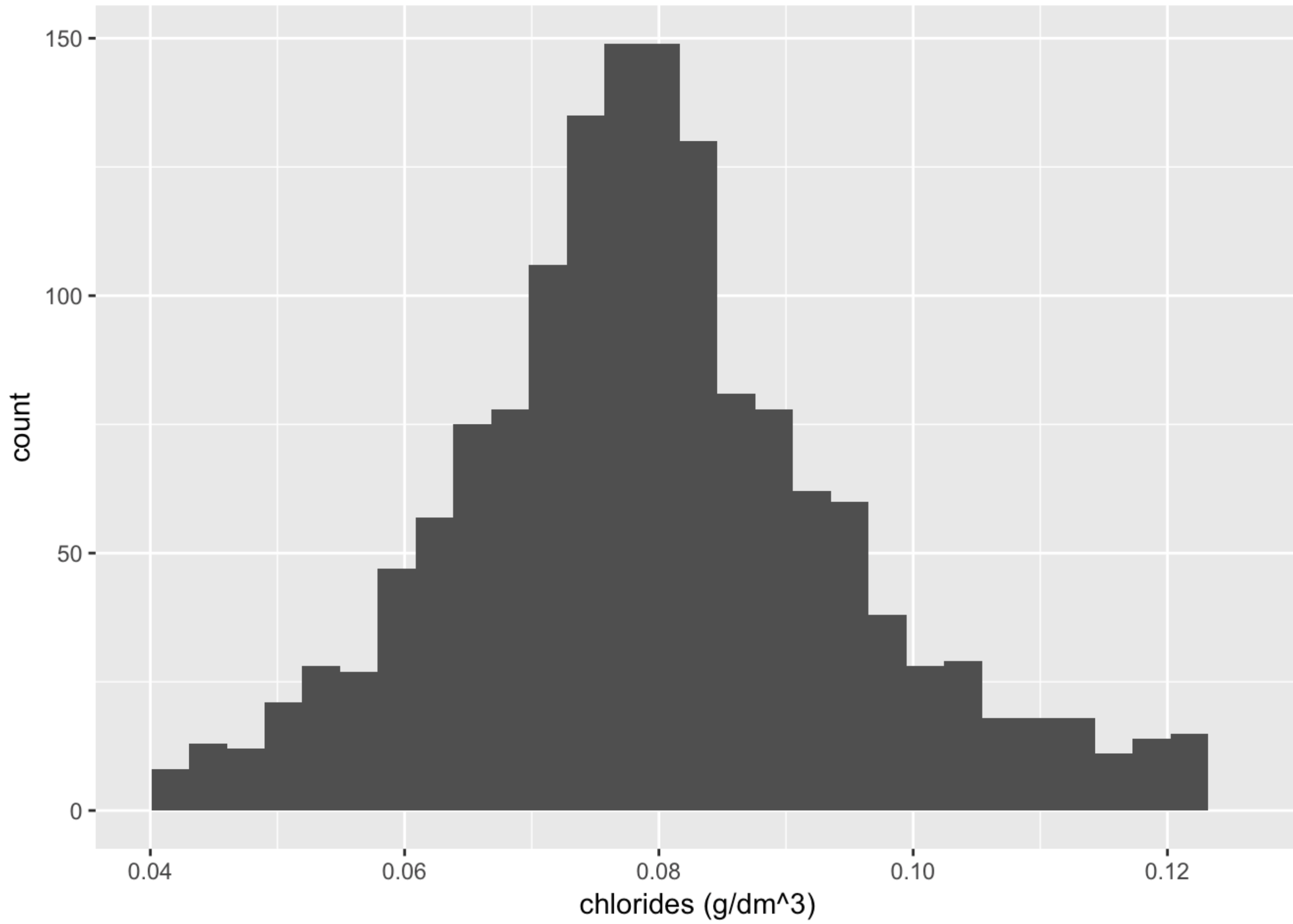
Plots in residual.sugar and chlorides

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.su
gar(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’.

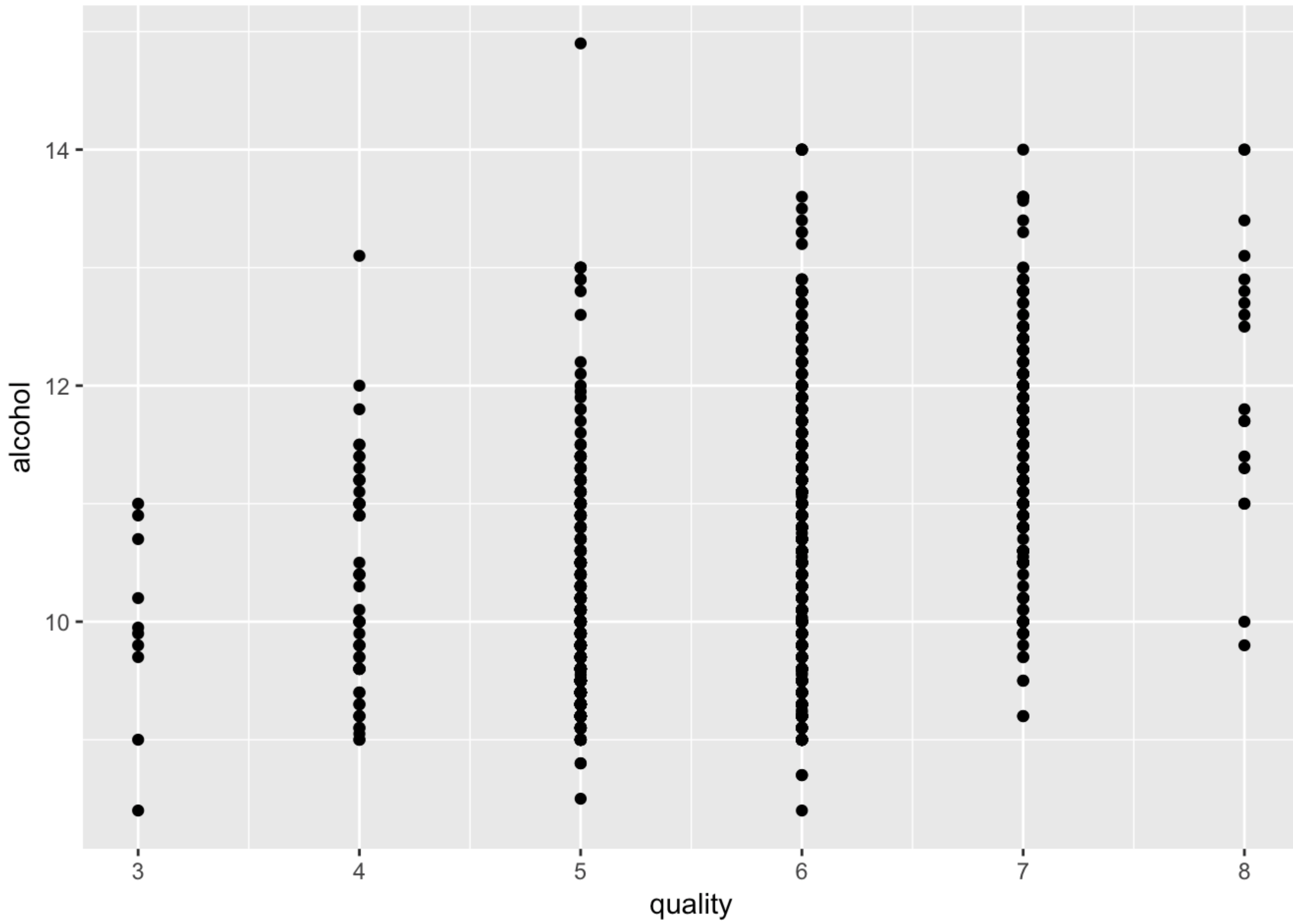
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

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

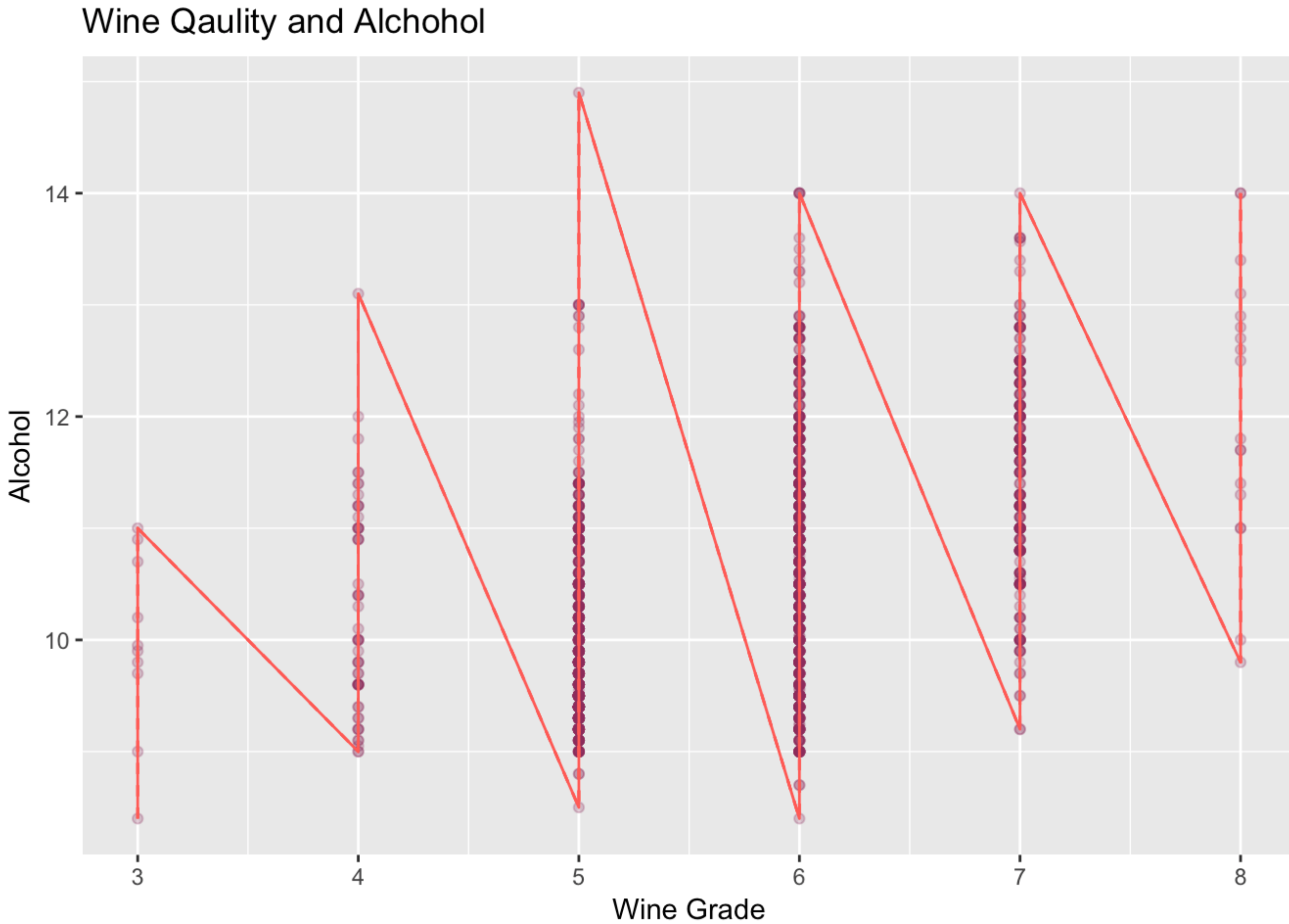
Wine quality has biggest correlation value to wine quality, so lets start with a basic scatter plot of the both.

```
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(x=quality, y=alcohol), data = wd) +
  geom_point(color='#993366', alpha = 1/4) +
  geom_line(stat = 'summary', fun.y=quantile, probs= 0.5, color='#FF6660') + geom_line(s
tat='summary',fun.y = quantile,probs = .9, linetype =2, color='#FF6660') +
  geom_line(stat = 'summary', fun.y=quantile, probs= 0.1, linetype =2, color='#FF6660')+
            xlab("Wine Grade") + ylab("Alcohol") +
    ggtitle("Wine Qaulity and Alchohol")
```



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

Wine Quality in categories

Here box plots are used to represent categorical values.

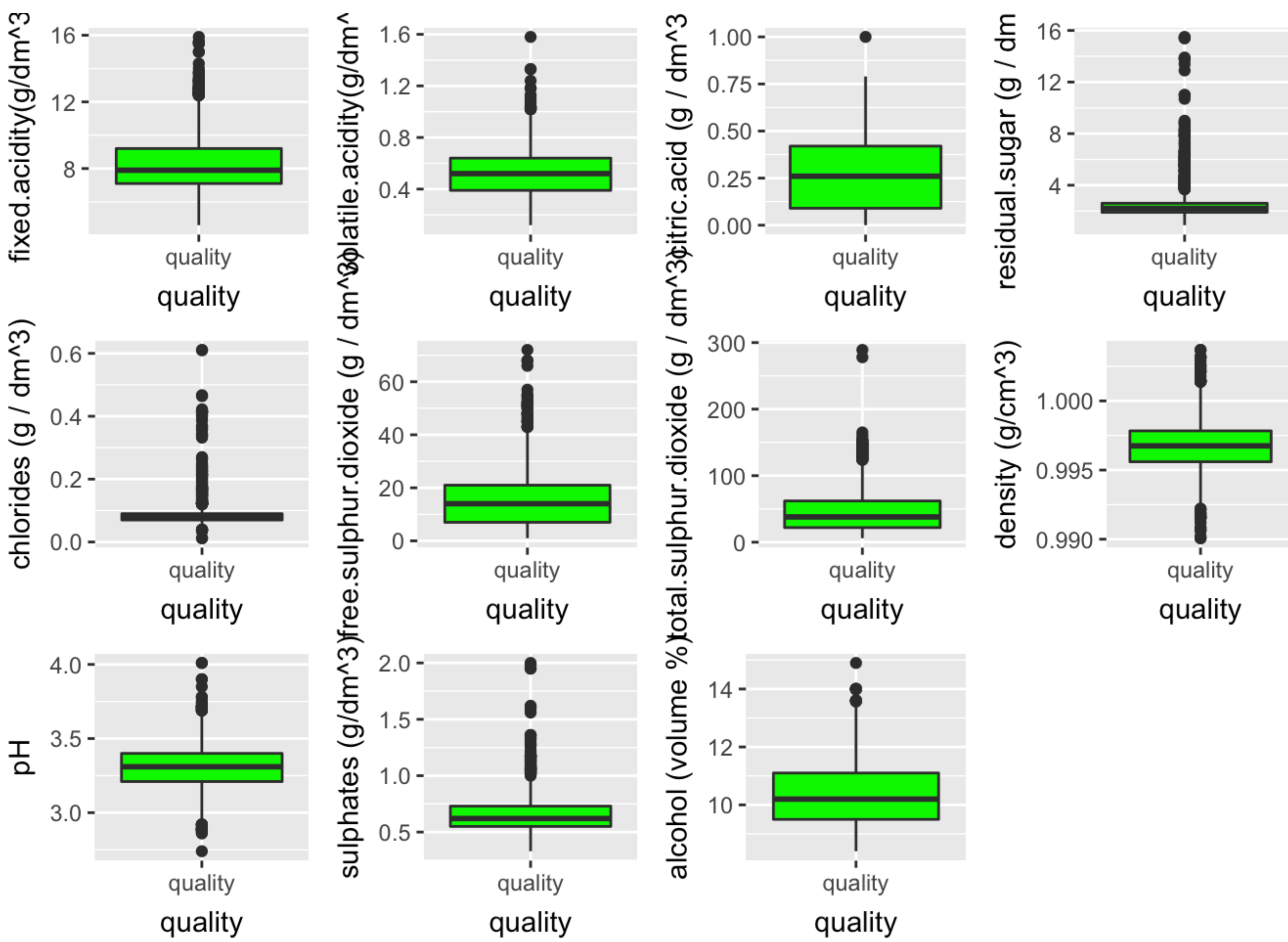
BoxPlot of quality

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

}

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

ncol= 4)
```

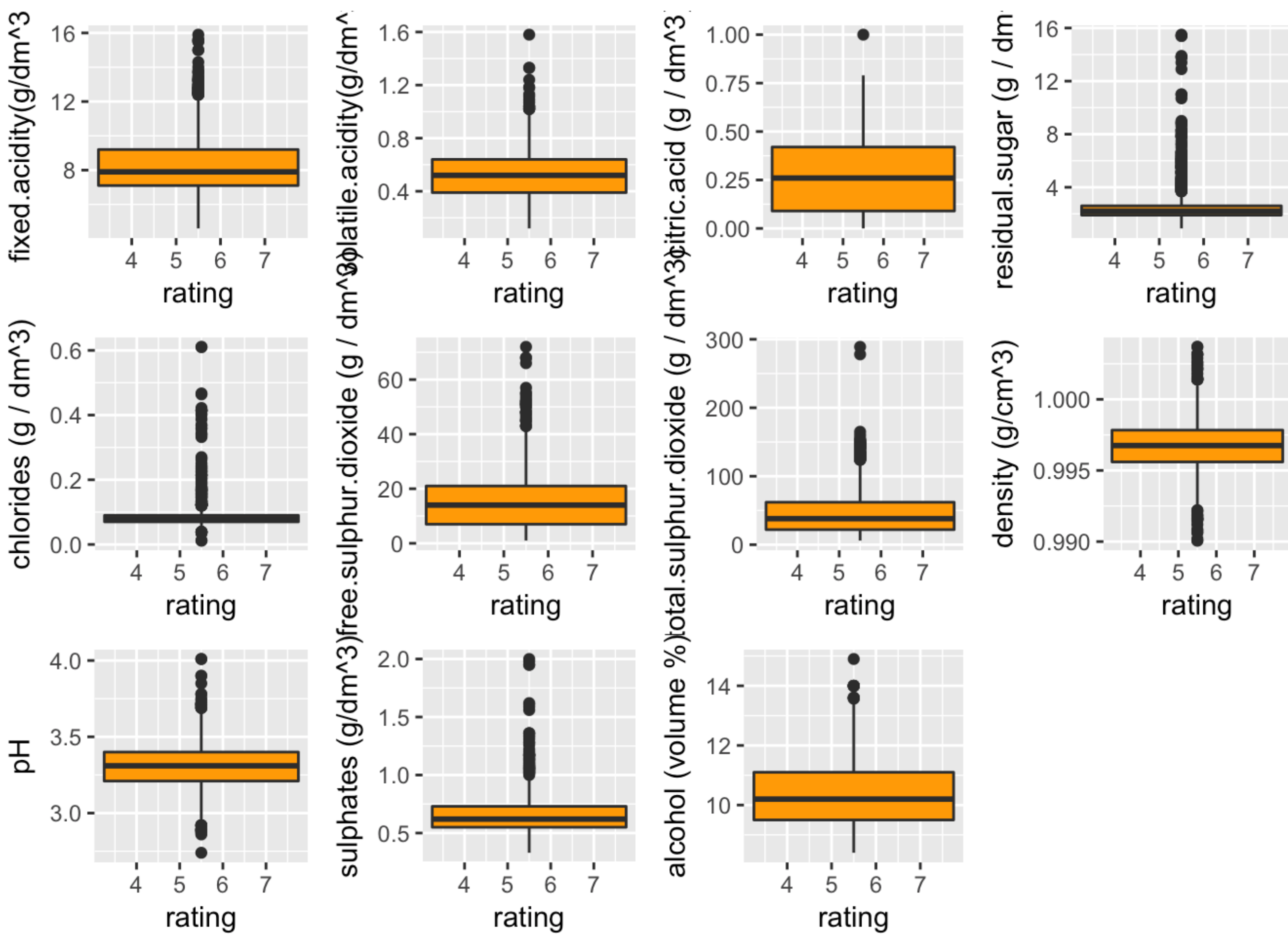


BoxPlot of rating

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

grid.arrange( rating_plot( 'quality', 'fixed.acidity', 'fixed.acidity(g/dm^3)'),
rating_plot('quality', 'volatile.acidity', 'volatile.acidity(g/dm^3)'),
rating_plot('quality', 'citric.acid', 'citric.acid (g / dm^3)'),
rating_plot('quality', 'residual.sugar', 'residual.sugar (g / dm^3)'),
rating_plot('quality', 'chlorides', 'chlorides (g / dm^3)'),
rating_plot('quality', 'free.sulfur.dioxide', 'free.sulphur.dioxide (g / dm^3)'),
rating_plot('quality', 'total.sulfur.dioxide', 'total.sulphur.dioxide (g / dm^3)'),
rating_plot('quality', 'density', 'density (g/cm^3)'),
rating_plot('quality', 'pH', 'pH'),
rating_plot('quality', 'sulphates', 'sulphates (g/dm^3)'),
rating_plot('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)
names(correlations) <- c('fixed.acidity','volatile.acidity',                          'c
itric.acid',                                                'log10.residual.sugar',
                              'log10.chlordies',
  'free.sulfur.dioxide',                                        'total.sulfur.dioxide', 'd
ensity',                          'pH', 'log10.sulphates', 'alcohol')
correlations
```

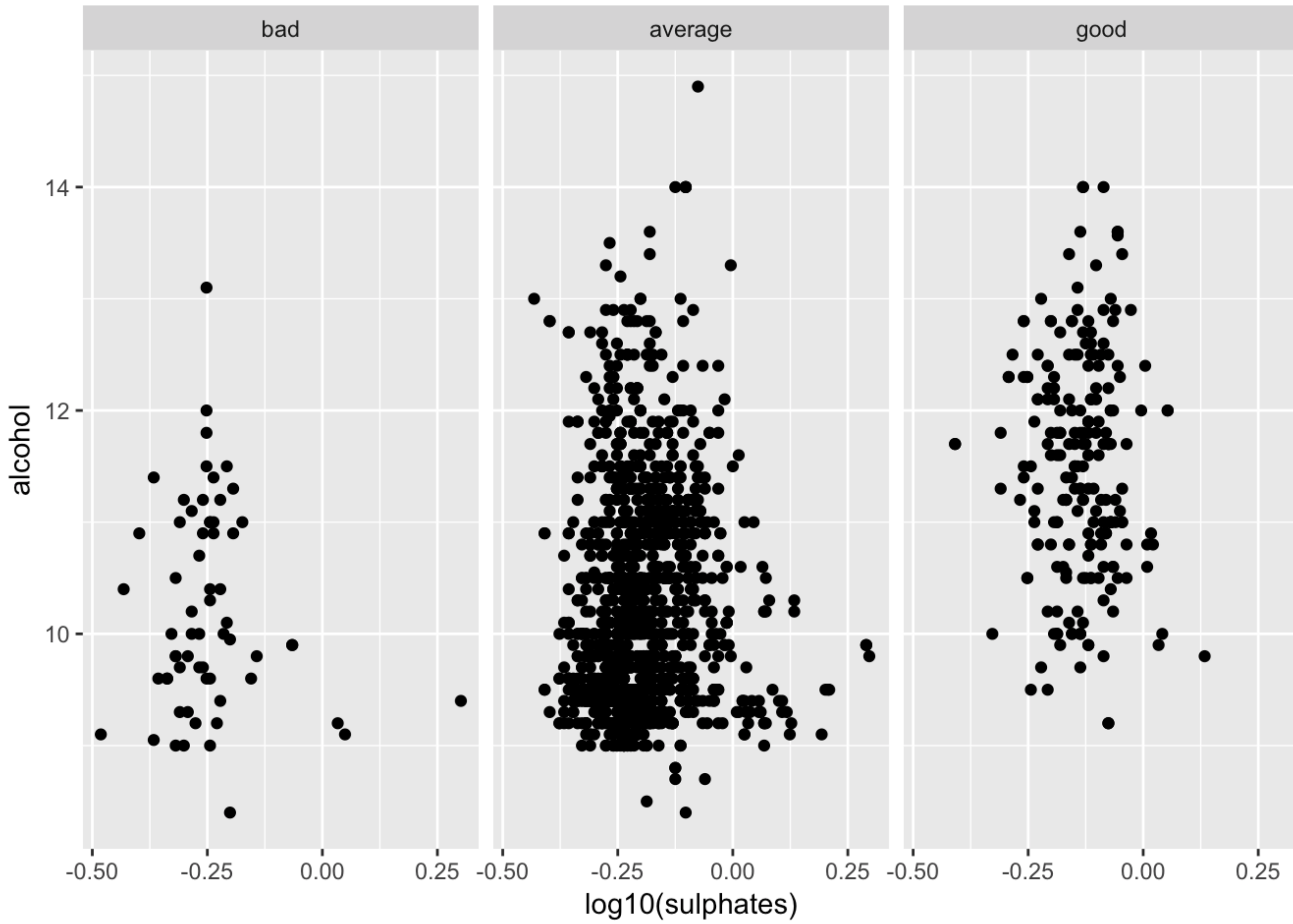
```
##        fixed.acidity     volatile.acidity          citric.acid
##           0.12405165          -0.39055778           0.22637251
## log10.residual.sugar      log10.chlordies  free.sulfur.dioxide
##           0.02353331          -0.17613996          -0.05065606
## total.sulfur.dioxide         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

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

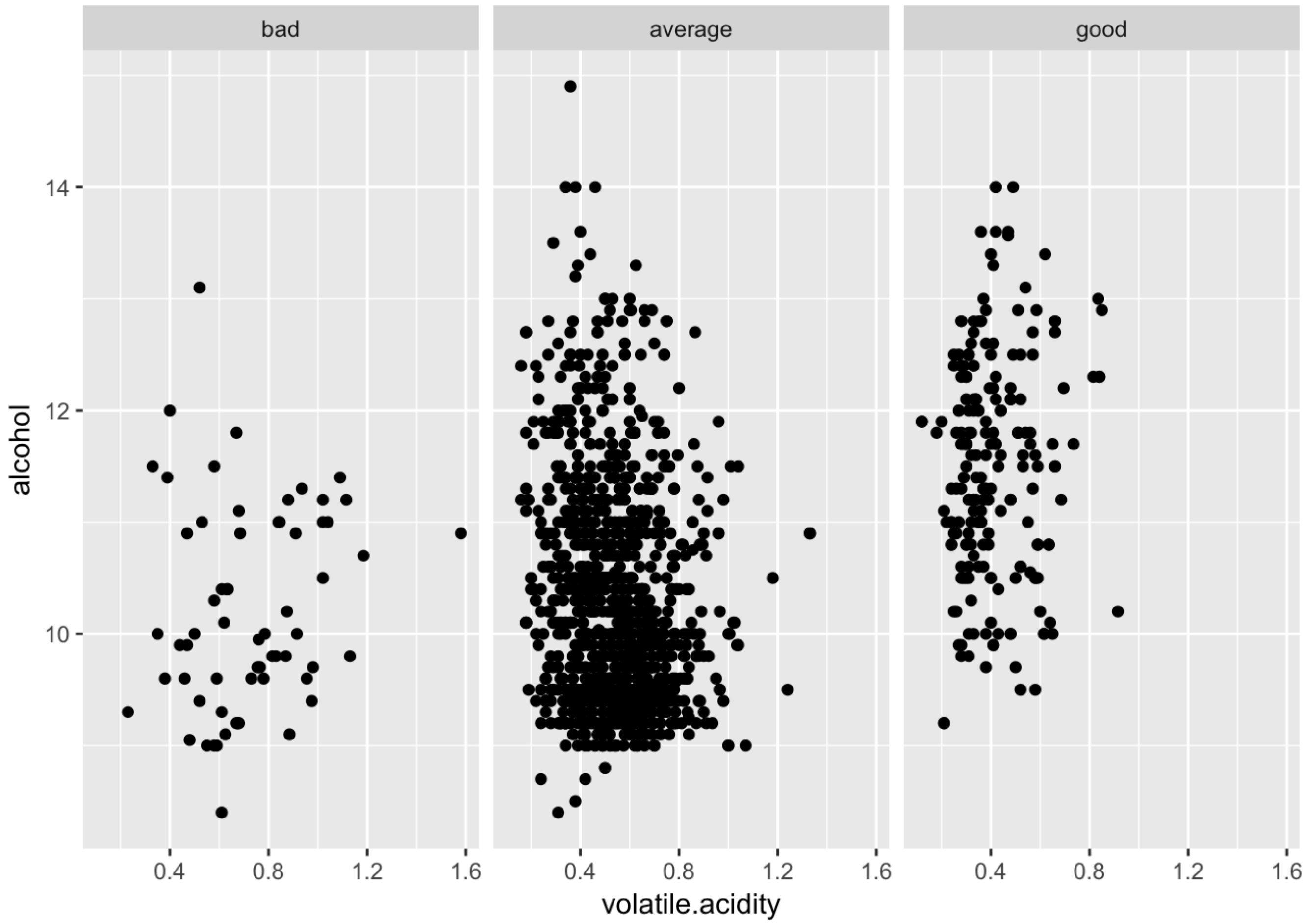
- alcohal
- sulphates
- citric.acid
- fixed.acidity

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

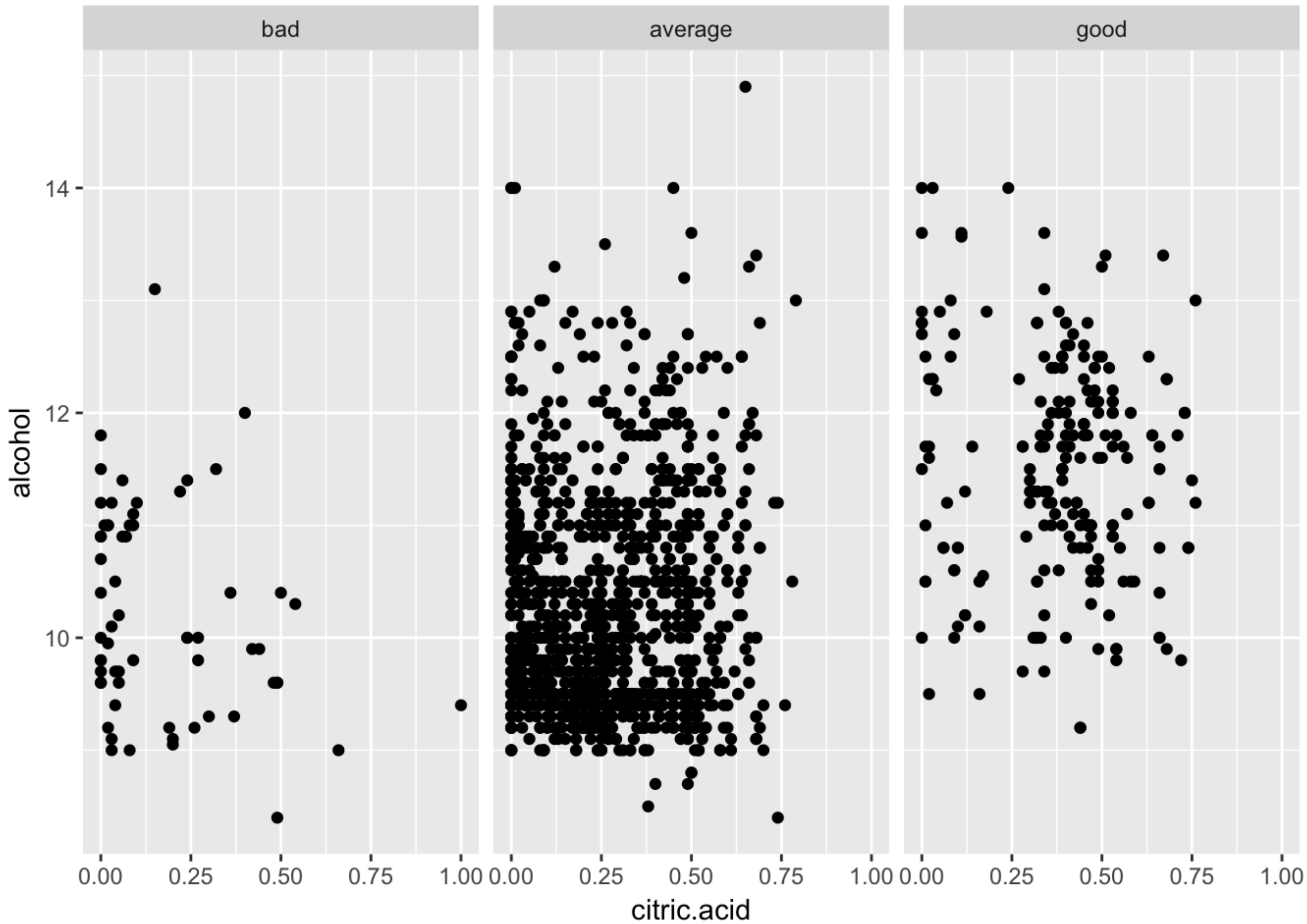
```
ggplot( data = wd, aes(x= log10(sulphates), y= alcohol)) +
  facet_wrap(~rating) +
  geom_point()
```



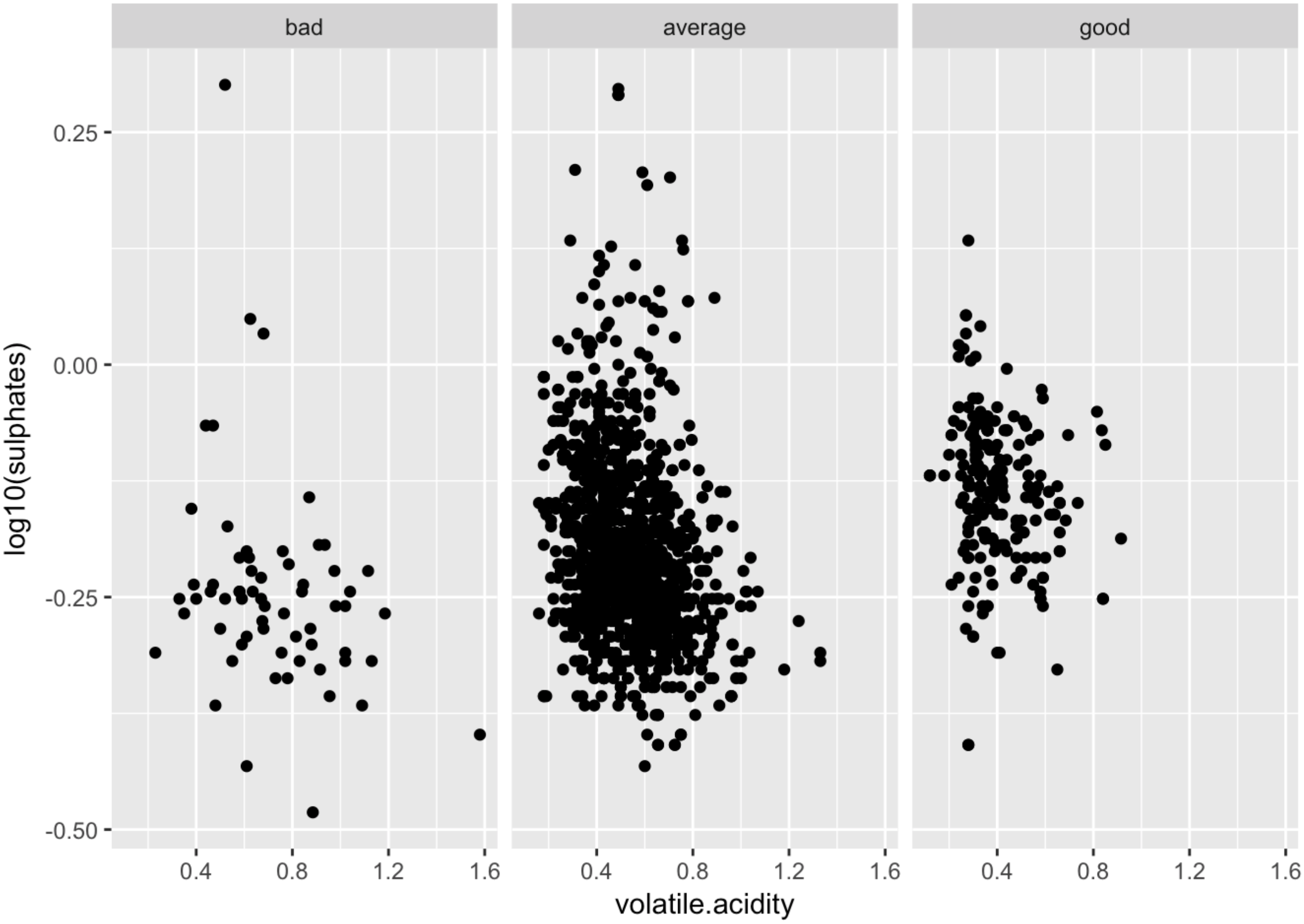

```
ggplot(data = wd, aes(x = volatile.acidity, y = alcohol)) +
  facet_wrap(~rating) +
  geom_point()
```



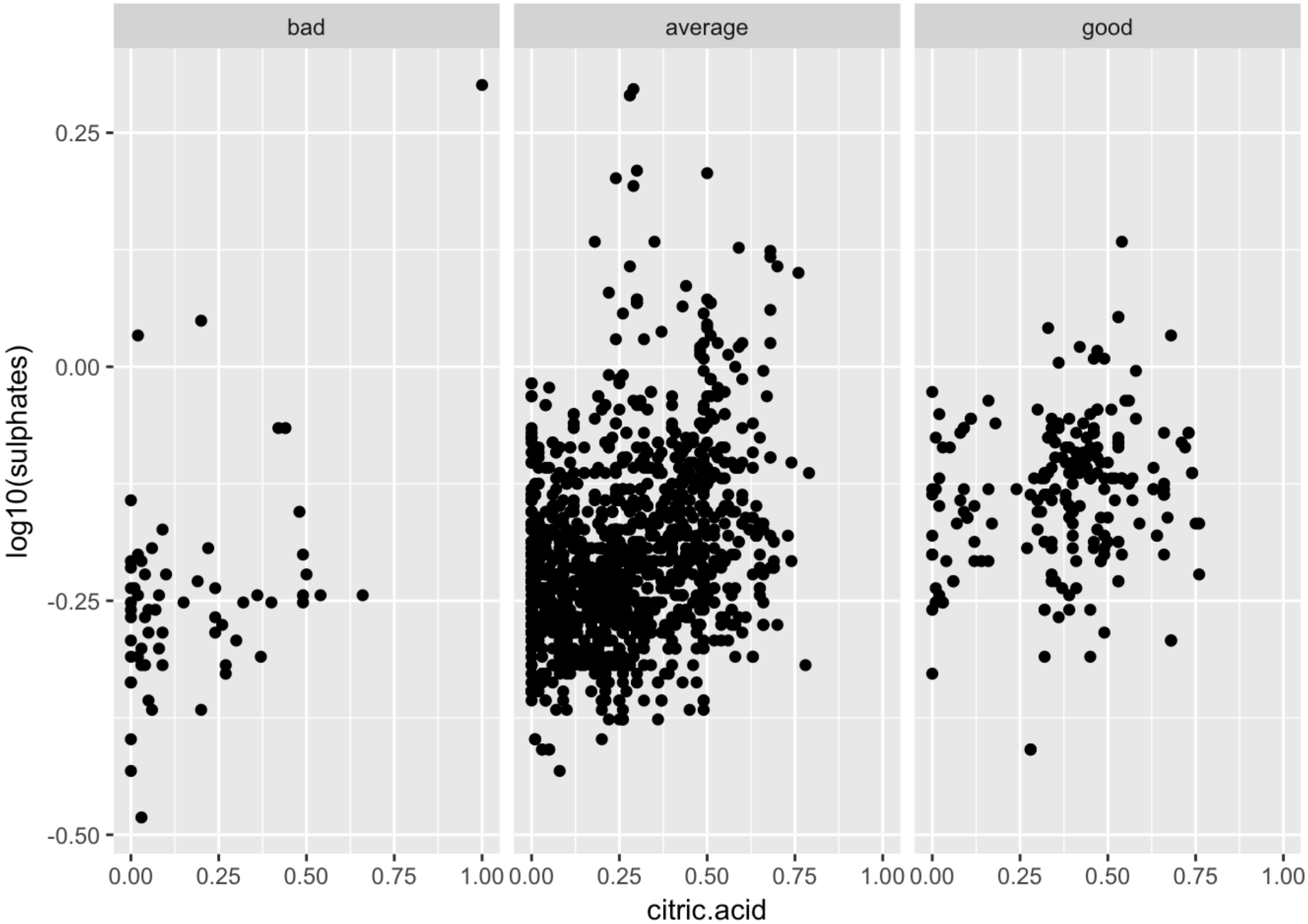

```
ggplot(data = wd, aes(x = citric.acid, y = alcohol)) +
  facet_wrap(~rating) +
  geom_point()
```



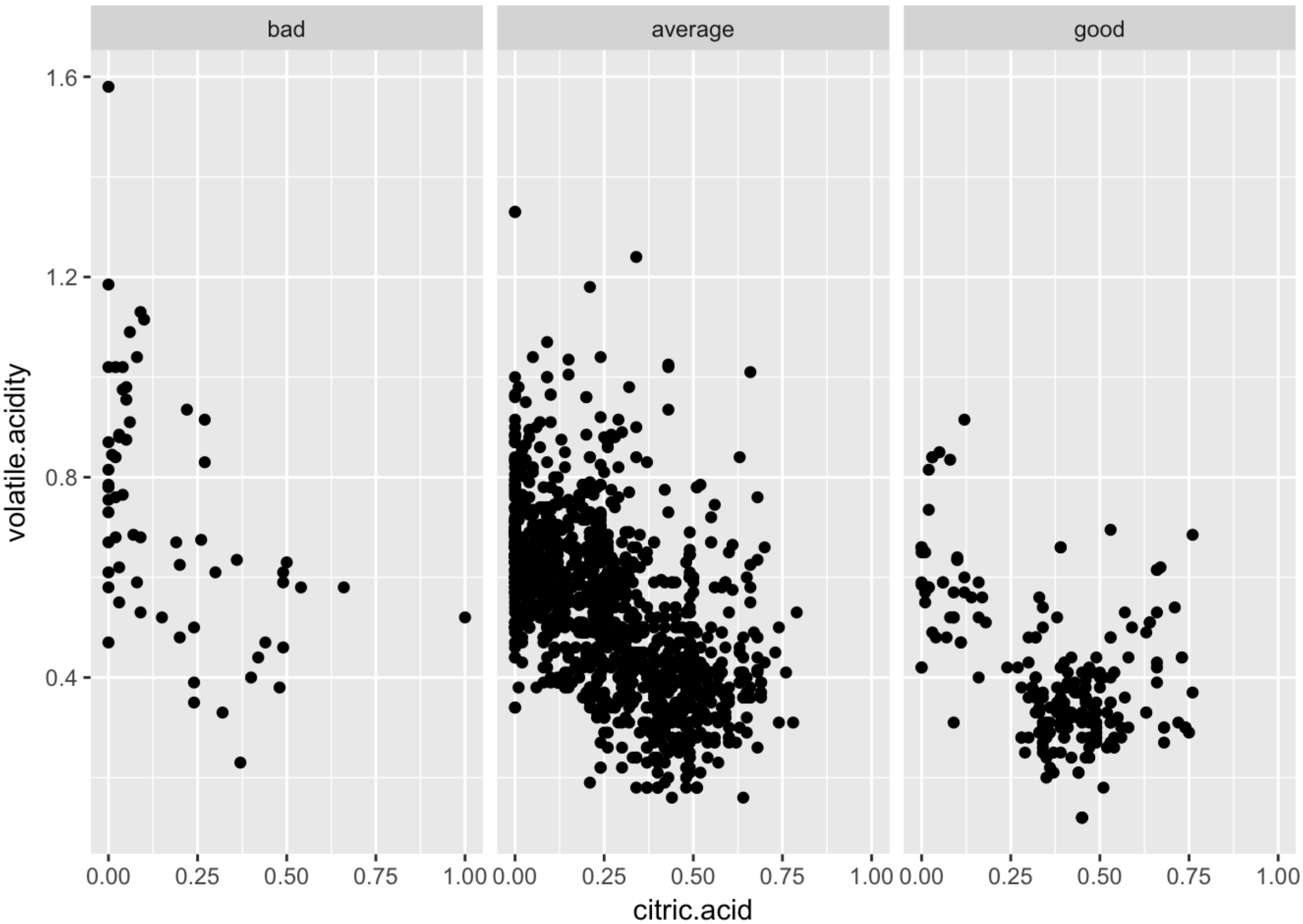

```
ggplot(data = wd, aes(x = volatile.acidity, y = log10(sulphates))) +
  facet_wrap(~rating) +
  geom_point()
```




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



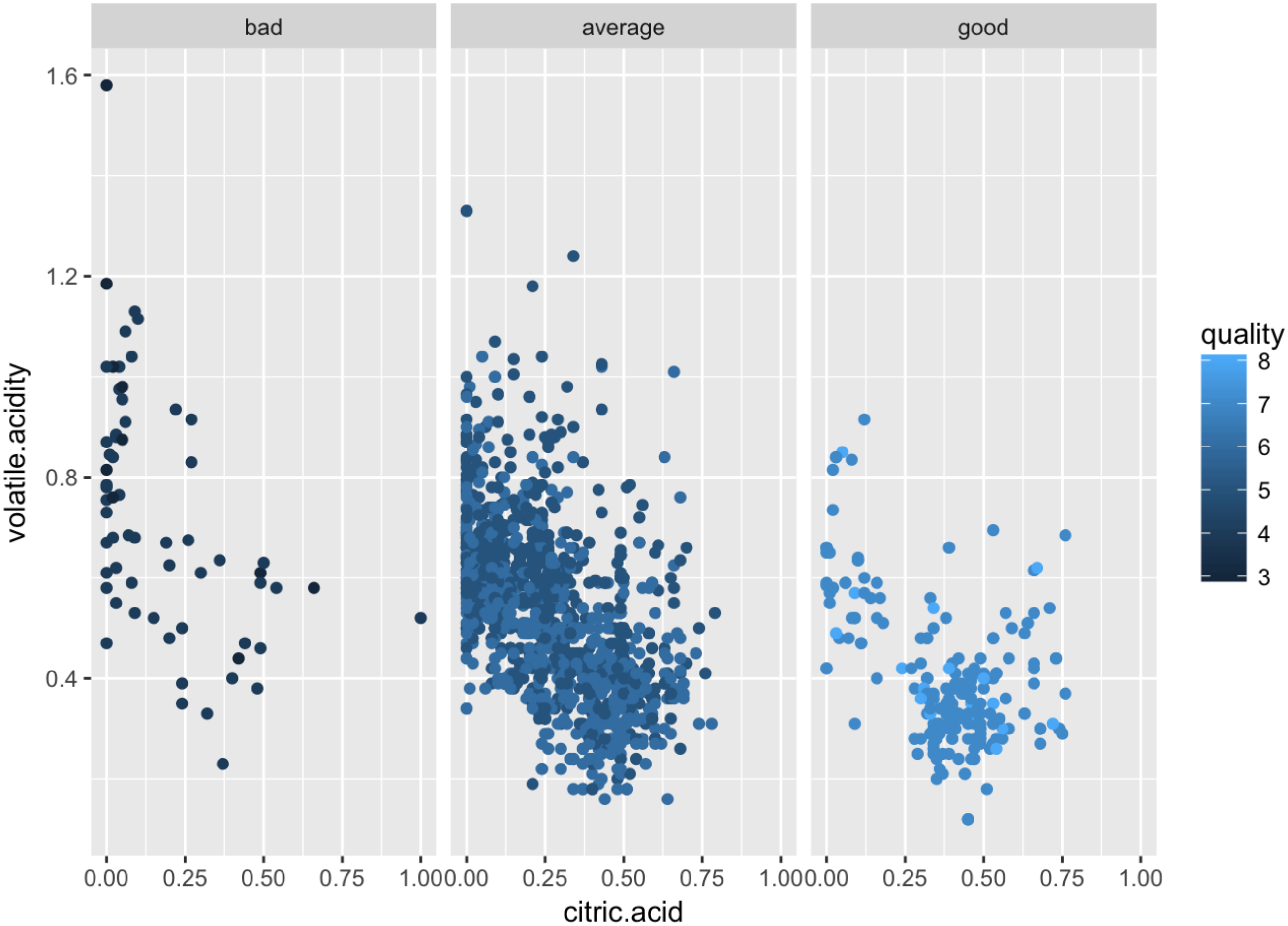

```
ggplot(data = wd, aes(x = citric.acid, y = volatile.acidity)) +
  facet_wrap(~rating) +
  geom_point()
```



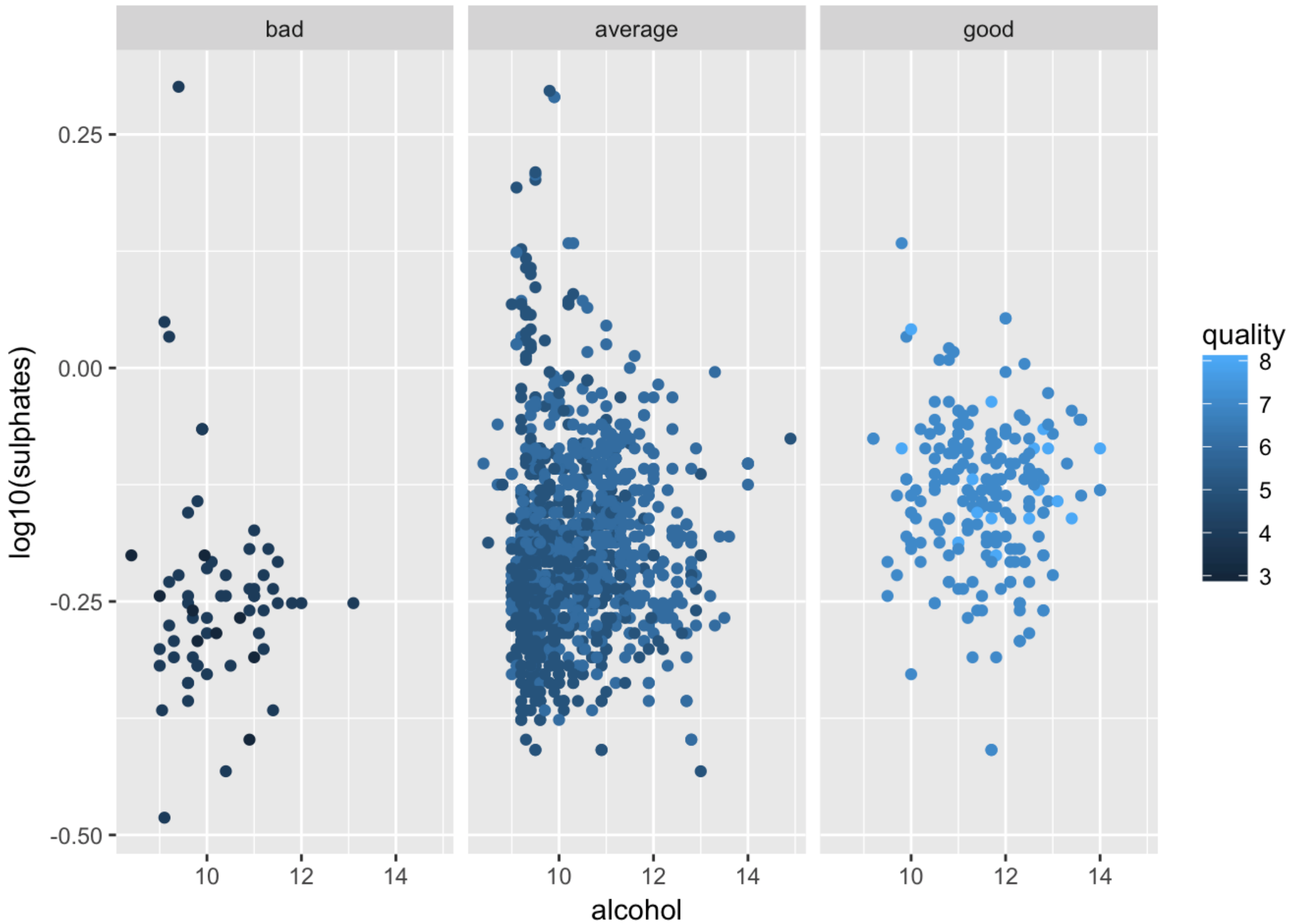
From the above plots only one thing is clear: alcohol content heavely effects rating.

Multivariate Plots

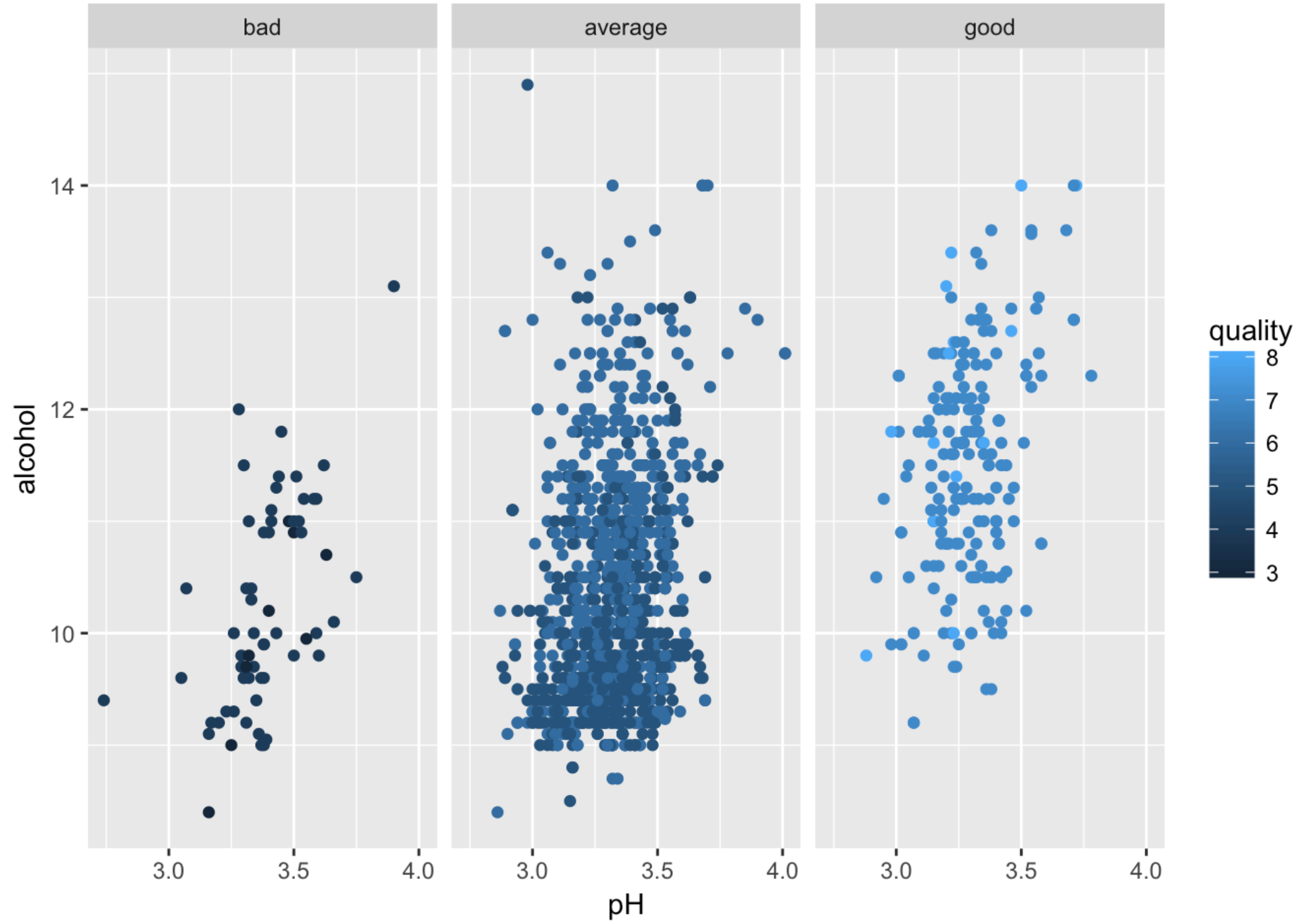
```
ggplot(data = wd,
       aes(x = citric.acid, y = volatile.acidity,
           color = quality)) +
  geom_point() +
  facet_wrap(~rating)
```




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




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



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)
  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         m6
         m7
## -------------------------------------------------------------------------------------
-----------------
##   (Intercept)             6.566***   3.095***   2.611***   2.646***   2.769***   2.985
***   -0.953
##                          (0.058)    (0.184)    (0.196)    (0.201)    (0.202)    (0.20
6)    (11.990)
##   volatile.acidity       -1.761***  -1.384***  -1.221***  -1.265***  -1.155***  -1.104
***   -1.114***
##                          (0.104)    (0.095)    (0.097)    (0.113)    (0.115)    (0.11
5)     (0.120)
##   alcohol                            0.314***   0.309***   0.309***   0.292***   0.276
***    0.280***
##                                     (0.016)    (0.016)    (0.016)    (0.016)    (0.01
7)     (0.020)
##   sulphates                                     0.679***   0.696***   0.871***   0.908
***    0.903***
##                                                (0.101)    (0.103)    (0.111)    (0.11
1)     (0.112)
##   citric.acid                                             -0.079      0.021      0.065
       0.044
##                                                           (0.104)    (0.106)    (0.10
6)     (0.124)
##   chlorides                                                          -1.663***  -1.763
***   -1.747***
##                                                                      (0.405)    (0.40
3)     (0.406)
##   total.sulfur.dioxide                                                          -0.002
***   -0.002***
##                                                                                 (0.00
1)     (0.001)
##   density
       3.923
##
```

```
    (11.944)
## ----------------------------------------------------------------------------------------------
-----------------
##   R-squared                     0.2        0.3        0.3        0.3        0.3
 0.4        0.4
##   adj. R-squared                0.2        0.3        0.3        0.3        0.3
 0.3        0.3
##   sigma                         0.7        0.7        0.7        0.7        0.7
 0.7        0.7
##   F                           287.4      370.4      268.9      201.8      166.4      14
3.9      123.3
##   p                             0.0        0.0        0.0        0.0        0.0
 0.0        0.0
##   Log-likelihood            -1794.3    -1621.8    -1599.4    -1599.1    -1590.7    -158
0.2    -1580.1
##   Deviance                    883.2      711.8      692.1      691.9      684.6      67
5.7      675.6
##   AIC                        3594.6     3251.6     3208.8     3210.2     3195.3     317
6.4     3178.3
##   BIC                        3610.8     3273.1     3235.7     3242.4     3233.0     321
9.4     3226.7
##   N                          1599       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.

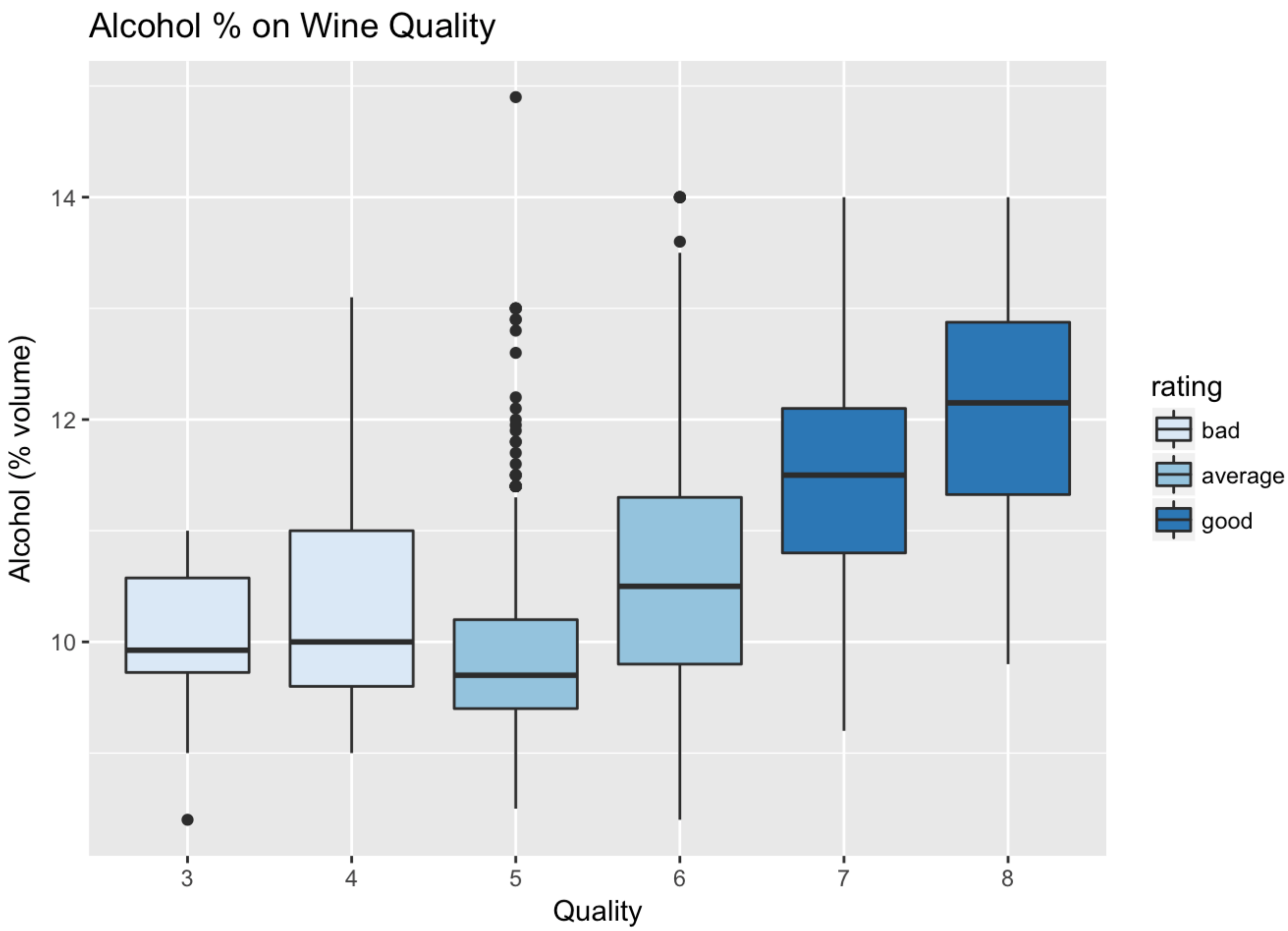
The model can be described as:

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

Final Plots and Summary

Alcohol and Wine quality

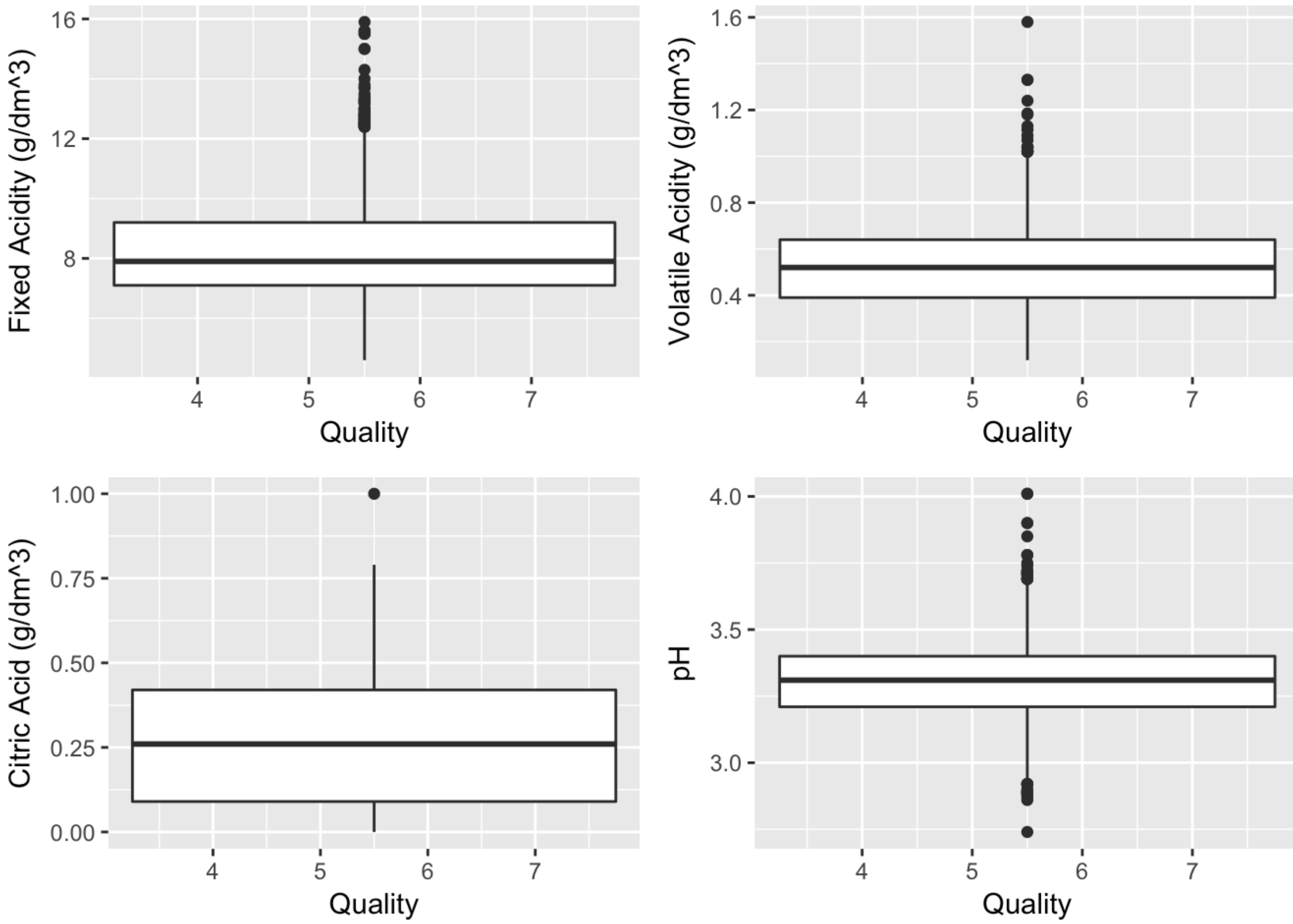
```
ggplot(data = wd, aes(as.factor(quality), alcohol, fill = rating)) +
  geom_boxplot() +
  ggtitle('Alcohol % on Wine Quality') +
  xlab('Quality') +
  ylab('Alcohol (% volume)') +
  scale_fill_brewer(type = 'seq', palette = 1)
```



From the above plot it is clear that wine quality increases with % of alcohol in it.

Acids and Wine quality

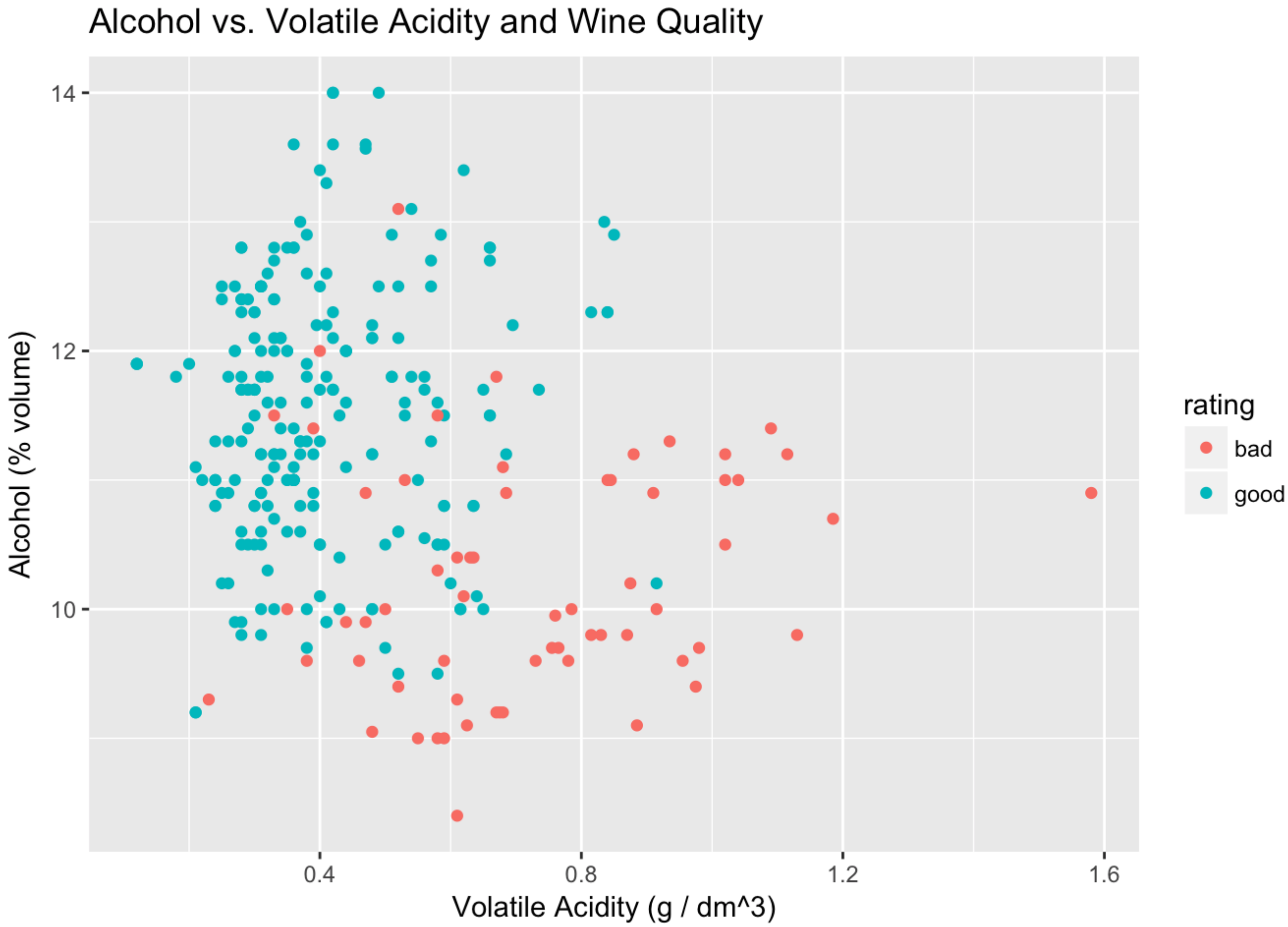
```
grid.arrange(ggplot(data = wd, aes(x = quality,y =fixed.acidity,
                                   fill = quality)) +
               ylab('Fixed Acidity (g/dm^3)') +
               xlab('Quality') +
               geom_boxplot(),
             ggplot(data = wd, aes(x = quality,y = volatile.acidity,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

Good and Bad wines

```
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'.

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/)
- 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://onlinecourses.science.psu.edu/stat857/node/223)

- https://github.com/Dalaska/Udacity-Red-Wine-Quality/blob/master/redwine_final.rmd (https://github.com/Dalaska/Udacity-Red-Wine-Quality/blob/master/redwine_final.rmd)