Assesing Wine Qualities

Krushab Gandhi

7/18/2018

Red Wine Exploration by Krushab Gandhi

In this project, we will explore a data set on wine quality. The dataset was obatined from the UCI Machine Learning Repository. The objective is to explore which chemical properties influence the quality of red wines.

We will start by exploring the data using the statistical program, R. As interesting relationships in the data are discovered, we will produce and refine plots to illustrate them.

Firstly, we will **load the Wine dataset** and **analyze** its structure

```
Wine <- read.csv('~/Desktop/Resume/Code-and-Dataset/wineQualityReds.csv')
str(Wine)
```

```
## '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 ...
```

Let's see the **summary** of the Wine dataset

```
summary(Wine)
```

```
##        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** of **13 numeric** variables.
- **X** appears to be the unique identifier.
- **Quality** is an ordered, categorical, discrete variable. From the literature, this was on a 0-10 scale, and was rated by at least 3 wine experts. The values ranged only from 3 to 8, with a **mean** of 5.64 and **median** of 6.
- All **other variables** seem to be continuous quantities with the **exception** of the .sulfur.dioxide suffixes
- From the variable descriptions, it **appears** that **fixed.acidity ~ volatile.acidity** and **free.sulfur.dioxide ~ total.sulfur.dioxide** may possibly be dependent subsets of each other.

Converting the **X** identifier into **factors** and **quality** attribute into **numeric**

```
Wine$quality <- as.numeric(Wine$quality)
Wine$X <- as.factor(Wine$X)
```

We use the **table()** function to calculate how many wines we have for each quality

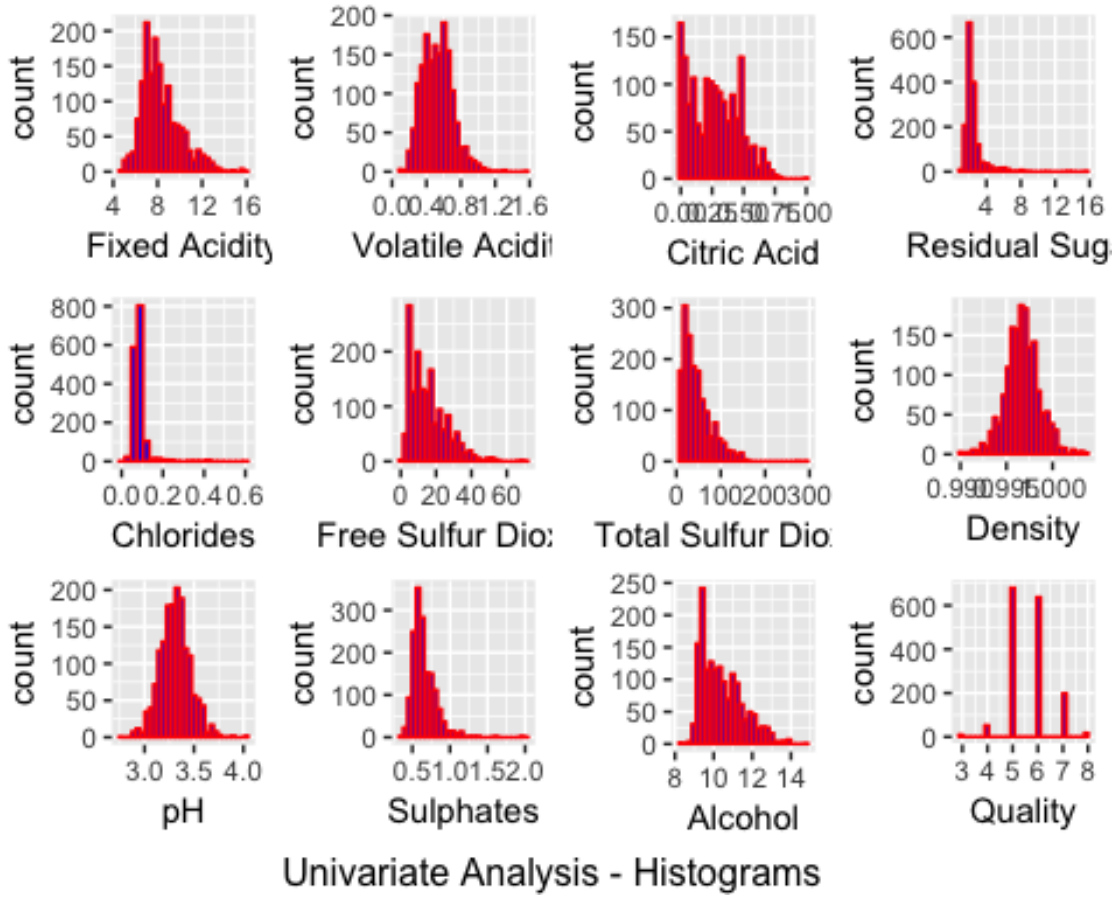
```
table(Wine$quality)
```

```
##
##   3   4   5   6   7   8
##  10  53 681 638 199  18
```

UNIVARIATE PLOTS SECTION
Analyzing each and every attribute in the Wine dataset.

```
library(ggplot2)
library(gridExtra)
grid.arrange(qplot(Wine$fixed.acidity, geom = "histogram",  xlab = "Fixed
Acidity", fill=I("blue"),
                   col=I("red")),
             qplot(Wine$volatile.acidity, geom = "histogram", xlab =
"Volatile Acidity", fill=I("blue"),
                   col=I("red")),
             qplot(Wine$citric.acid, geom = "histogram", xlab = "Citric
Acid", fill=I("blue"),
                   col=I("red")),
             qplot(Wine$residual.sugar, geom = "histogram", xlab = "Residual
Sugar", fill=I("blue"),
                   col=I("red")),
             qplot(Wine$chlorides, geom = "histogram", binwidth = 0.03, xlab
= "Chlorides", fill=I("blue"),
                   col=I("red")),
             qplot(Wine$free.sulfur.dioxide, geom = "histogram", xlab = "Free
Sulfur Dioxide",
                   fill=I("blue"), col=I("red")),
             qplot(Wine$total.sulfur.dioxide, geom = "histogram", xlab =
"Total Sulfur Dioxide",
                   fill=I("blue"), col=I("red")),
             qplot(Wine$density, geom = "histogram", xlab = "Density",
fill=I("blue"),
                   col=I("red") ),
             qplot(Wine$pH, geom = "histogram", xlab = "pH", fill=I("blue"),
                   col=I("red")),
             qplot(Wine$sulphates, geom = "histogram", xlab = "Sulphates",
fill=I("blue"),
                   col=I("red")),
             qplot(Wine$alcohol, geom = "histogram", xlab = "Alcohol",
fill=I("blue"),
```

```
                  col=I("red")),
            qplot(Wine$quality, geom = "histogram", xlab = "Quality",
fill=I("blue"),
                  col=I("red")),
            ncol = 4, bottom = "Univariate Analysis - Histograms")
```



UNIVARIATE ANALYSIS

From the plots, we come to know the following:

- Looking at **Wine Quality**, we find that it forms a **normal distribution**, with most of the data structured around wine quality of 5 or 6.
- It appears that **Density** and **pH** are **normally distributed**, with a **few outliers**.
- **Fixed and volatile acidity, sulfur dioxides, sulphates,** and **alcohol** seem to be **long-tailed.**
- Qualitatively, **residual sugar** and **chlorides** have **extreme outliers.**
- **Citric acid** appeared to have a large number of zero values. This might be a case of **non reporting.**

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

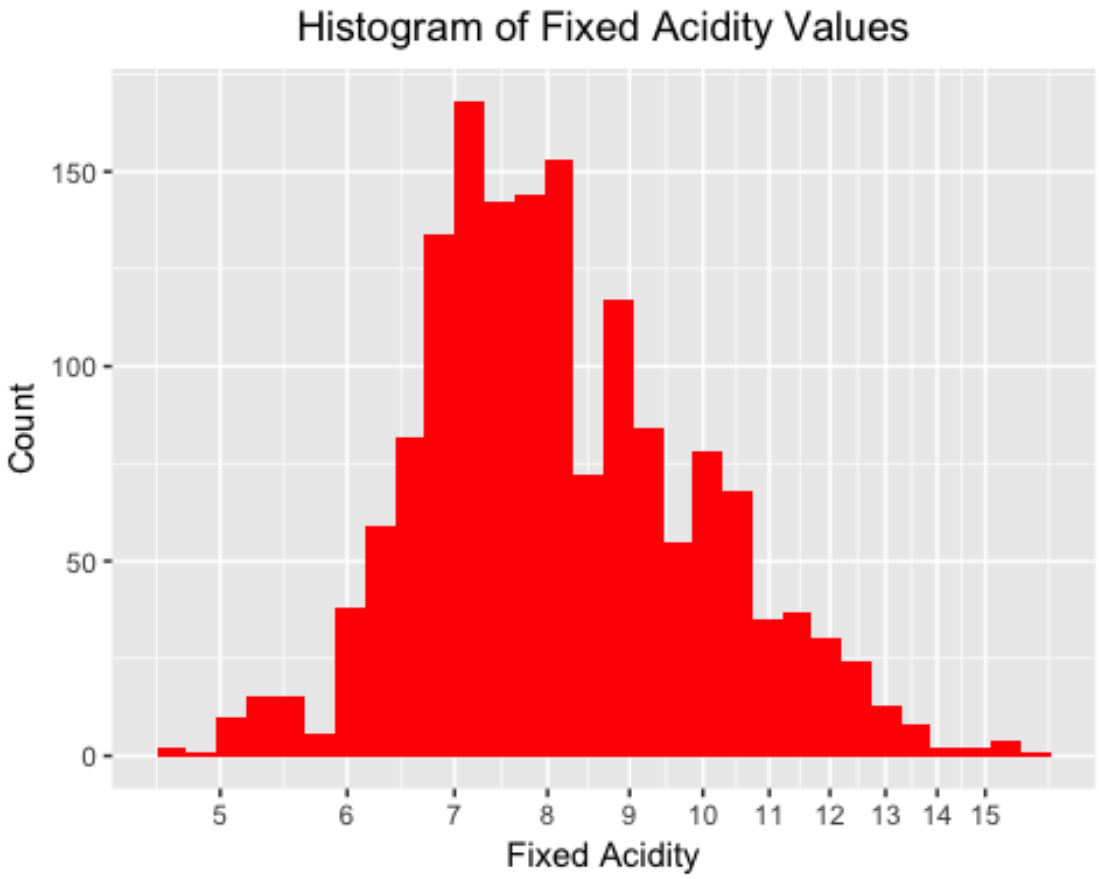
```
##     bad average    good
##      63    1319     217
```

```
table(Wine$rating)
```

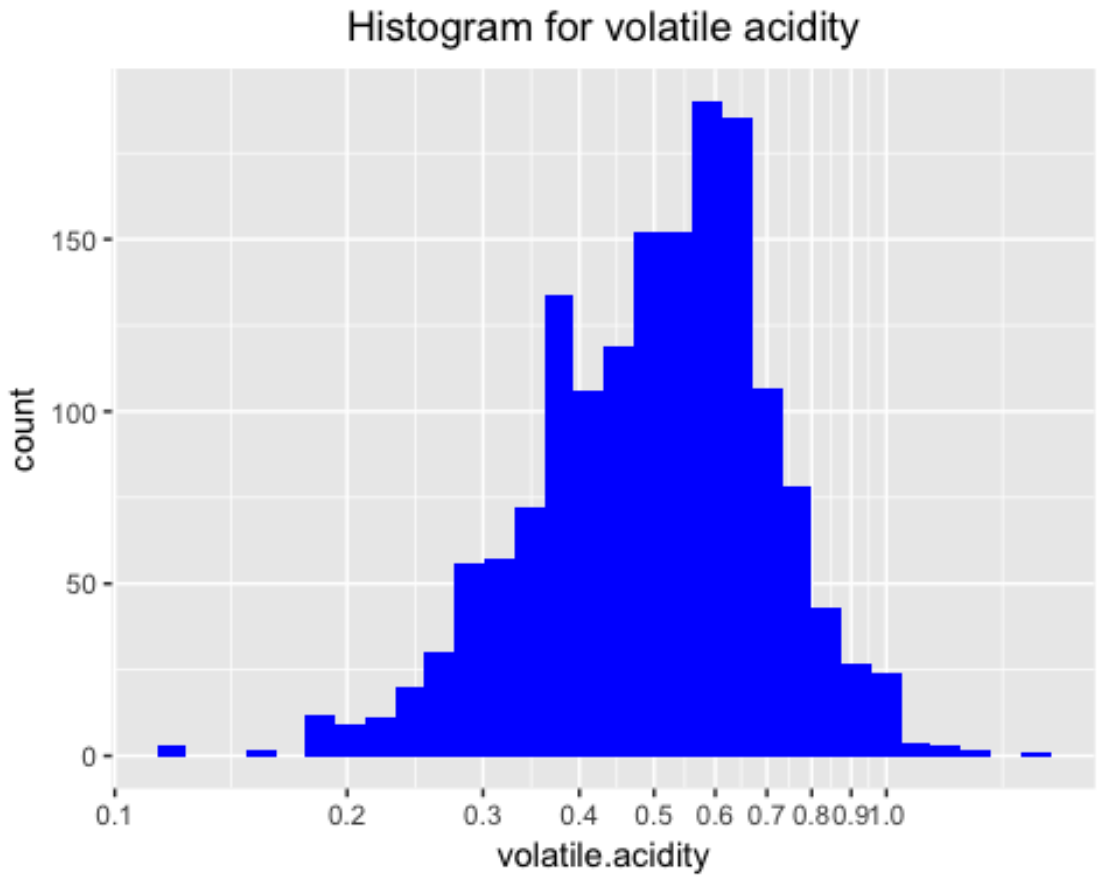
```
##
##     bad average    good
##      63    1319     217
```

To explore better, we created a new variable **Rating**, which classifies Wines into **Good, Bad, and Average** depending on their quality scores.
- **BAD** is Quality **less than 5.**
- **AVERAGE** is Quality **less than 7** but **greater than or equal to 5.**
- **GOOD** is any number **above and including 7.**

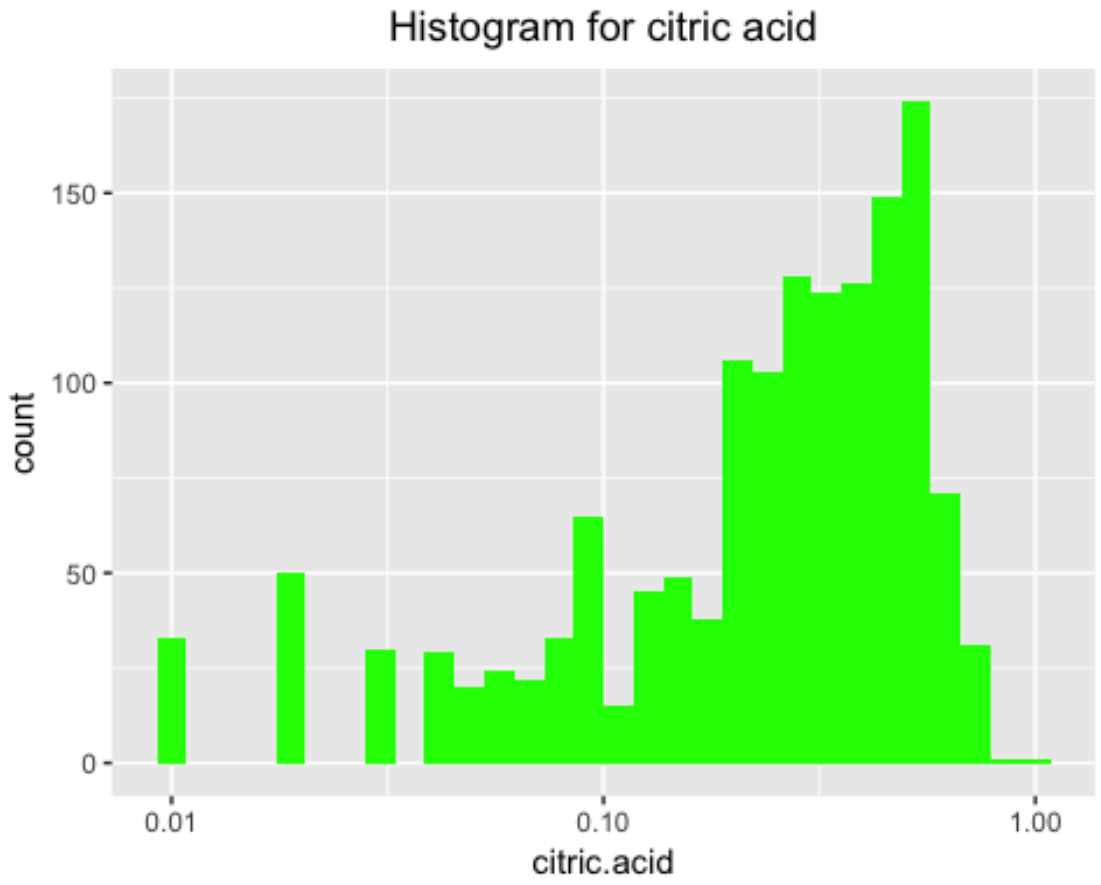
```
ggplot(Wine,aes(x=fixed.acidity)) + geom_histogram(fill='red') +
scale_x_log10(breaks=1:15) +
       xlab('Fixed Acidity') + ylab('Count') + ggtitle('Histogram of Fixed
Acidity Values') +
       theme(plot.title = element_text(hjust=0.5))
```




```
ggplot(Wine) + geom_histogram(aes(x=volatile.acidity),fill='blue') +
scale_x_log10(breaks=seq(0.1,1,0.1)) +
  ggtitle("Histogram for volatile acidity") + theme(plot.title =
element_text(hjust=0.5))
```

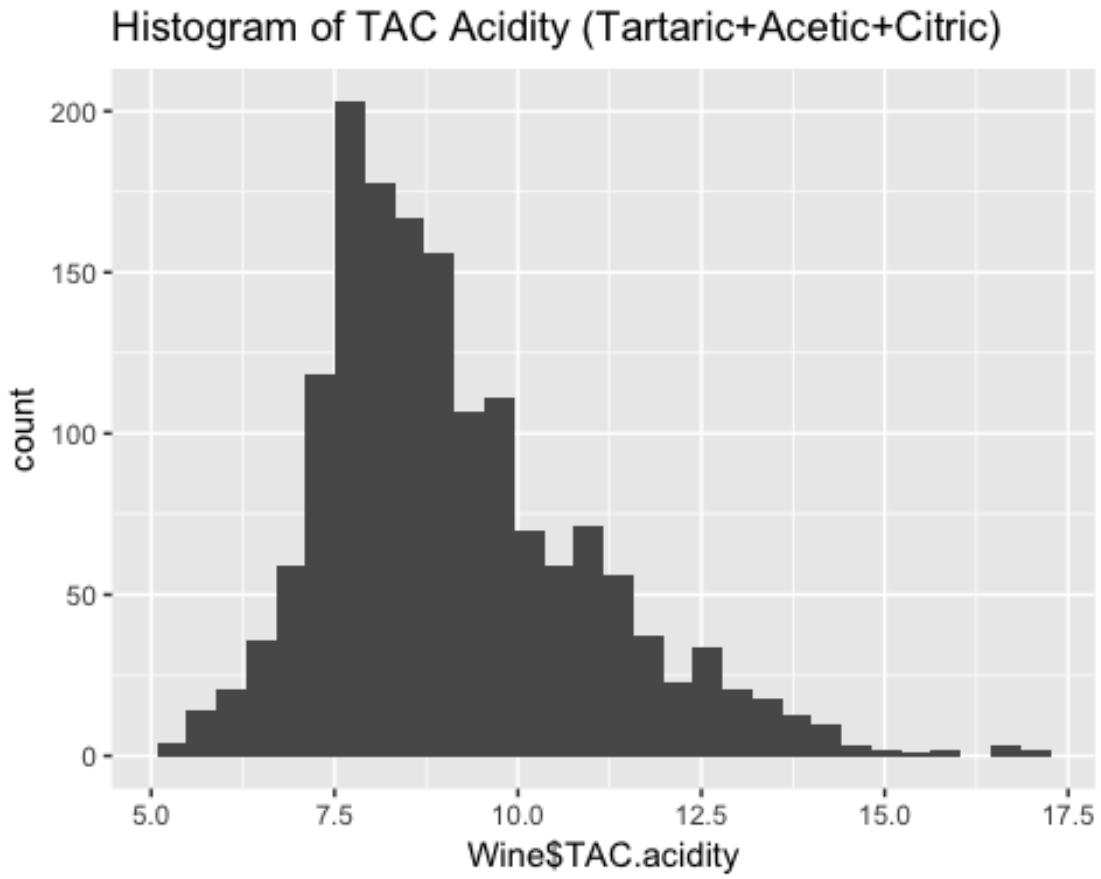



```
ggplot(Wine) + geom_histogram(aes(x=citric.acid), fill='green') +
scale_x_log10() +
  ggtitle("Histogram for citric acid") + theme(plot.title =
element_text(hjust=0.5))
```



- As we could clearly see, **citric acid** was one feature that was found to be **not normally distributed** on a logarithmic scale. The **transformation caused 132 data points to be in the infinite range**, telling us that 132 values are 0, since we know **Log(0) is infinity.**
- While exploring the **univariate histogram distributions**, there did not appear to be any **bimodal or multimodal distributions** that would warrant sub-classification into categorical variables. I considered potentially splitting **residual.sugar** into 'sweet wine' and 'dry wine', but a residual sugar of **greater than 45 g/L or g/m^3** is needed to classify as a **sweet wine.**
- We instantiated an ordered factor **rating**, classifying each wine sample as **'bad', 'average', or 'good'.**

```
library(plotly)
Wine$TAC.acidity <- Wine$fixed.acidity + Wine$volatile.acidity +
Wine$citric.acid
qplot(Wine$TAC.acidity,main = 'Histogram of TAC Acidity
(Tartaric+Acetic+Citric)')
```

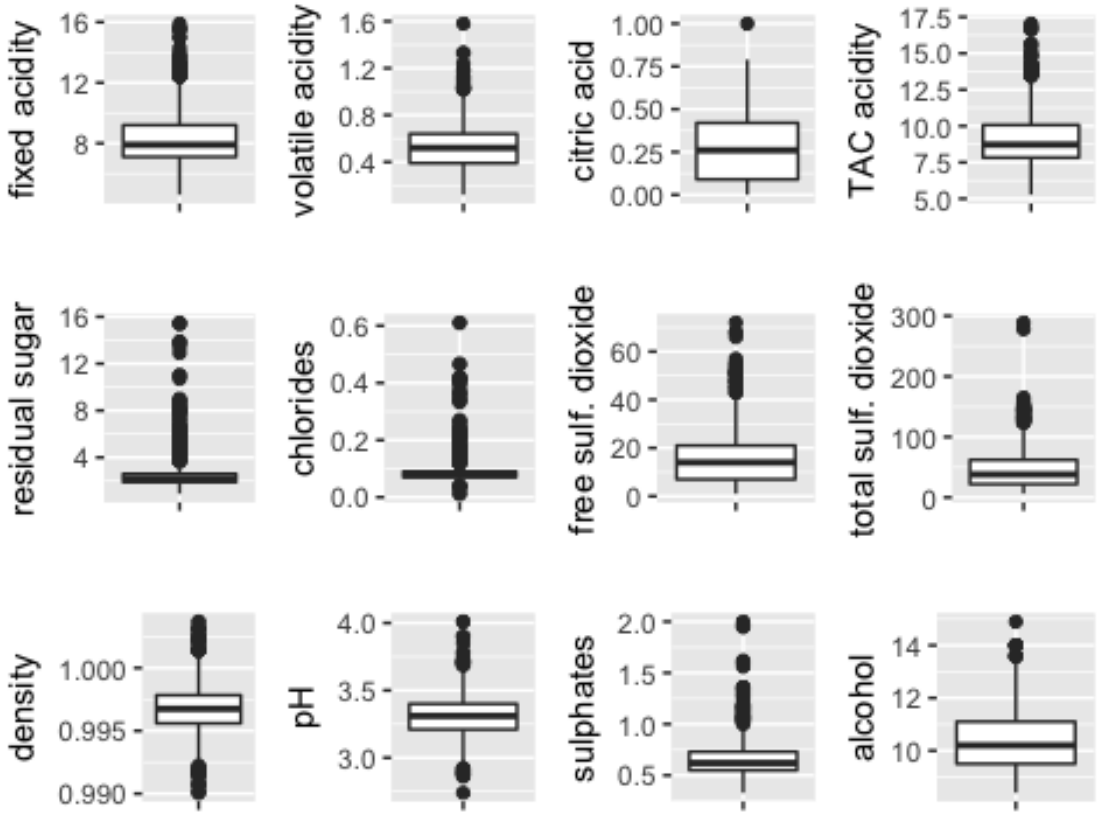



```
# p <- plot_ly(Wine,x=~TAC.acidity) %>% layout(xaxis = list(title =
"Histogram of TAC Acidity using plolty"))
# p ---- Using plotly to make the visualization more interactive
```

- Upon further examination of the **data set documentation**, it appears that **fixed.acidity and volatile.acidity** are different types of acids; tartaric acid and acetic acid. We decided to create a combined variable, **TAC.acidity**, containing the sum of **tartaric, acetic, and citric acid.**
- I have also used the **plotly** library to make the plot more interactive.

```
get_simple_boxplot <- function(column, ylab) {
return(qplot(data = Wine, x = '',
y = column, geom = 'boxplot',
xlab = '',
ylab = ylab))
}

grid.arrange(get_simple_boxplot(Wine$fixed.acidity, 'fixed acidity'),
get_simple_boxplot(Wine$volatile.acidity, 'volatile acidity'),
get_simple_boxplot(Wine$citric.acid, 'citric acid'),
```

```
get_simple_boxplot(Wine$TAC.acidity, 'TAC acidity'),
get_simple_boxplot(Wine$residual.sugar, 'residual sugar'),
get_simple_boxplot(Wine$chlorides, 'chlorides'),
get_simple_boxplot(Wine$free.sulfur.dioxide, 'free sulf. dioxide'),
get_simple_boxplot(Wine$total.sulfur.dioxide, 'total sulf. dioxide'),
get_simple_boxplot(Wine$density, 'density'),
get_simple_boxplot(Wine$pH, 'pH'),
get_simple_boxplot(Wine$sulphates, 'sulphates'),
get_simple_boxplot(Wine$alcohol, 'alcohol'),
ncol = 4)
```



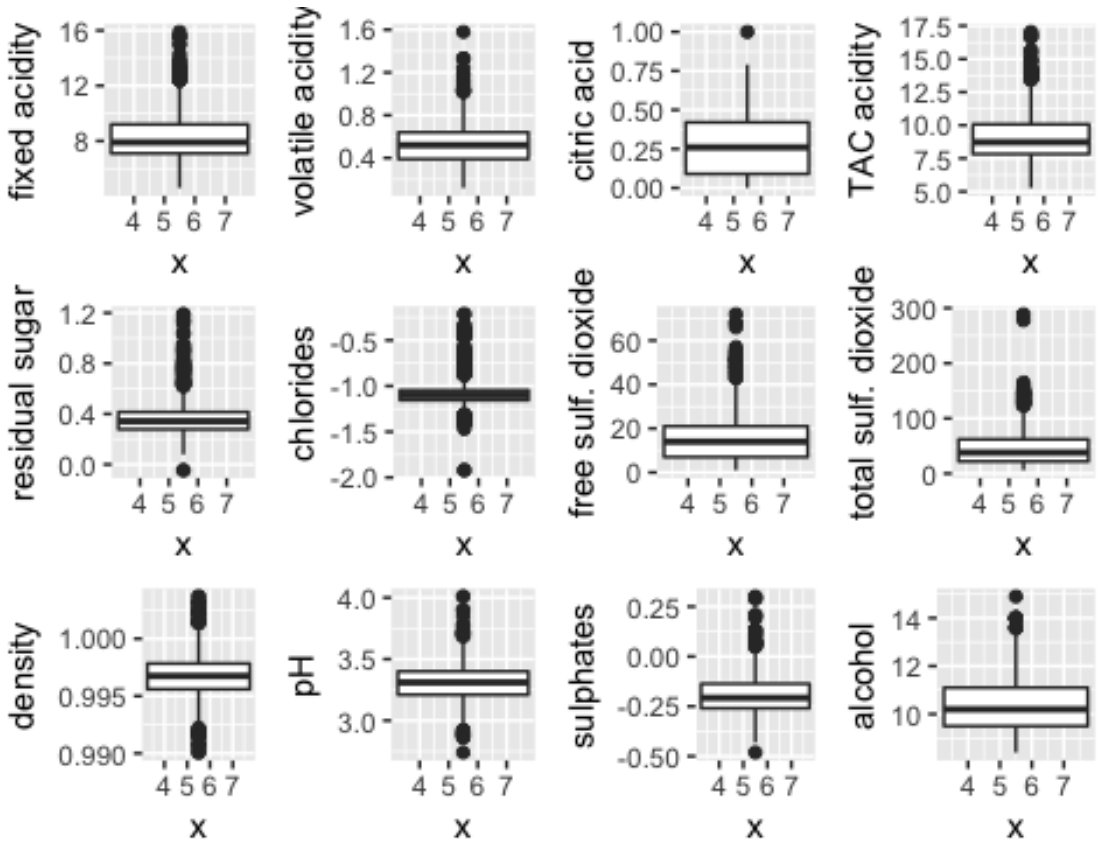
- In **univariate analysis**, we chose not to **tidy or adjust any data**, short of plotting a select few on logarithmic scales. Again making use of the **plolty** library for interactive plots.

- **Bivariate boxplots**, with **X as rating or quality**, will be more interesting in showing trends with wine quality.

```
# plot_ly(Wine,y=~alcohol,type='box') ------- Using plotly to make the
visualization more interactive
```

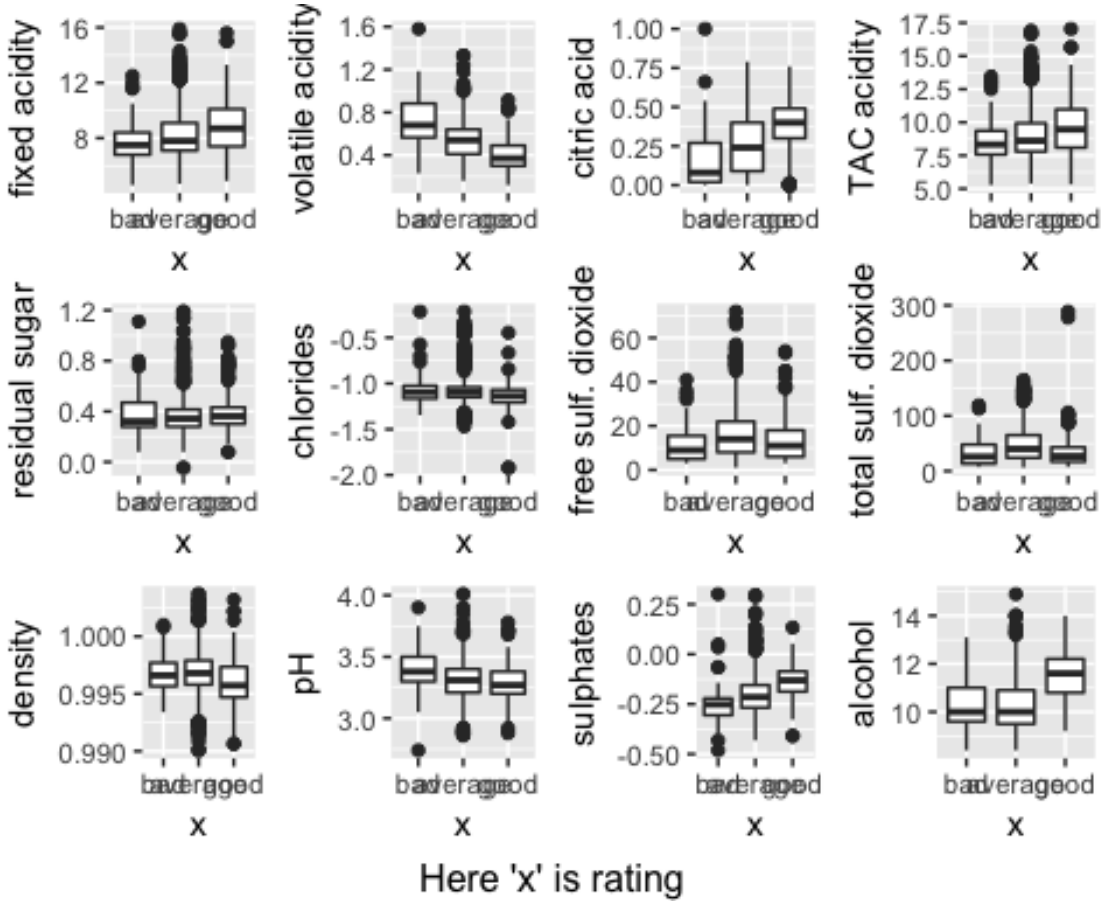
BIVARIATE PLOTS SECTION

```
set.seed(1)
Wine_sample <- Wine[,-which(names(Wine) %in% c('X',
'rating'))][sample(1:length(Wine$quality), 40), ]

get_bivariate_boxplot <- function(x, y, ylab) {
return(qplot(data = Wine, x = x, y = y, geom = 'boxplot', ylab = ylab))
}

grid.arrange(get_bivariate_boxplot(Wine$quality, Wine$fixed.acidity,
'fixed acidity'),
get_bivariate_boxplot(Wine$quality, Wine$volatile.acidity,
'volatile acidity'),
get_bivariate_boxplot(Wine$quality, Wine$citric.acid,
'citric acid'),
get_bivariate_boxplot(Wine$quality, Wine$TAC.acidity,
'TAC acidity'),
get_bivariate_boxplot(Wine$quality, log10(Wine$residual.sugar),
'residual sugar'),
get_bivariate_boxplot(Wine$quality, log10(Wine$chlorides),
'chlorides'),
get_bivariate_boxplot(Wine$quality, Wine$free.sulfur.dioxide,
'free sulf. dioxide'),
get_bivariate_boxplot(Wine$quality, Wine$total.sulfur.dioxide,
'total sulf. dioxide'),
get_bivariate_boxplot(Wine$quality, Wine$density,
'density'),
get_bivariate_boxplot(Wine$quality, Wine$pH,
'pH'),
get_bivariate_boxplot(Wine$quality, log10(Wine$sulphates),
'sulphates'),
get_bivariate_boxplot(Wine$quality, Wine$alcohol,
'alcohol'),
ncol = 4, bottom = "Here 'x' is quality")
```




```
grid.arrange(get_bivariate_boxplot(Wine$rating, Wine$fixed.acidity,
'fixed acidity'),
get_bivariate_boxplot(Wine$rating, Wine$volatile.acidity,
'volatile acidity'),
get_bivariate_boxplot(Wine$rating, Wine$citric.acid,
'citric acid'),
get_bivariate_boxplot(Wine$rating, Wine$TAC.acidity,
'TAC acidity'),
get_bivariate_boxplot(Wine$rating, log10(Wine$residual.sugar),
'residual sugar'),
get_bivariate_boxplot(Wine$rating, log10(Wine$chlorides),
'chlorides'),
get_bivariate_boxplot(Wine$rating, Wine$free.sulfur.dioxide,
'free sulf. dioxide'),
get_bivariate_boxplot(Wine$rating, Wine$total.sulfur.dioxide,
'total sulf. dioxide'),
get_bivariate_boxplot(Wine$rating, Wine$density,
'density'),
get_bivariate_boxplot(Wine$rating, Wine$pH,
'pH'),
get_bivariate_boxplot(Wine$rating, log10(Wine$sulphates),
'sulphates'),
get_bivariate_boxplot(Wine$rating, Wine$alcohol,
```

```
'alcohol'),
ncol = 4, bottom = "Here 'x' is rating")
```



Example using Plotly

```
# plot_ly(Wine,x=~quality,y=~alcohol) -------- Using plotly to make the
visualization more interactive
```

Results from Bivariate Analysis are as follows:

From exploring these plots, it seems that a 'good' wine generally has these trends: - **Higher** fixed acidity (tartaric acid) and citric acid
- **Lower** volatile acidity (acetic acid)
- **Lower** pH (i.e. more acidic)
- **Higher** sulphates
- **Higher** alcohol
- To a **lesser** extent, **lower chlorides and lower density**
- **Residual sugar and sulfur dioxides** did not seem to have a dramatic impact on the

quality or rating of the wines.
- Interestingly, it appears that different types of acid affect wine quality different; as such, **TAC.acidity** saw an attenuated trend, as the presence of volatile (acetic) acid accompanied decreased **quality.**

```
simple_cor_test <- function(x, y) {
return(cor.test(x, as.numeric(y))$estimate)
}

correlations <- c(
simple_cor_test(Wine$fixed.acidity, Wine$quality),
simple_cor_test(Wine$volatile.acidity, Wine$quality),
simple_cor_test(Wine$citric.acid, Wine$quality),
simple_cor_test(Wine$TAC.acidity, Wine$quality),
simple_cor_test(log10(Wine$residual.sugar), Wine$quality),
simple_cor_test(log10(Wine$chlorides), Wine$quality),
simple_cor_test(Wine$free.sulfur.dioxide, Wine$quality),
simple_cor_test(Wine$total.sulfur.dioxide, Wine$quality),
simple_cor_test(Wine$density, Wine$quality),
simple_cor_test(Wine$pH, Wine$quality),
simple_cor_test(log10(Wine$sulphates), Wine$quality),
simple_cor_test(Wine$alcohol, Wine$quality))

correlations
```

```
##         cor         cor         cor         cor         cor         cor
##  0.12405165 -0.39055778  0.22637251  0.10375373  0.02353331 -0.17613996
##         cor         cor         cor         cor         cor         cor
## -0.05065606 -0.18510029 -0.17491923 -0.05773139  0.30864193  0.47616632
```

```
names(correlations) <- c('fixed.acidity', 'volatile.acidity', 'citric.acid',
'TAC.acidity', 'log10.residual.sugar',
'log10.chlordies', 'free.sulfur.dioxide',
'total.sulfur.dioxide', 'density', 'pH',
'log10.sulphates', 'alcohol')
correlations
```

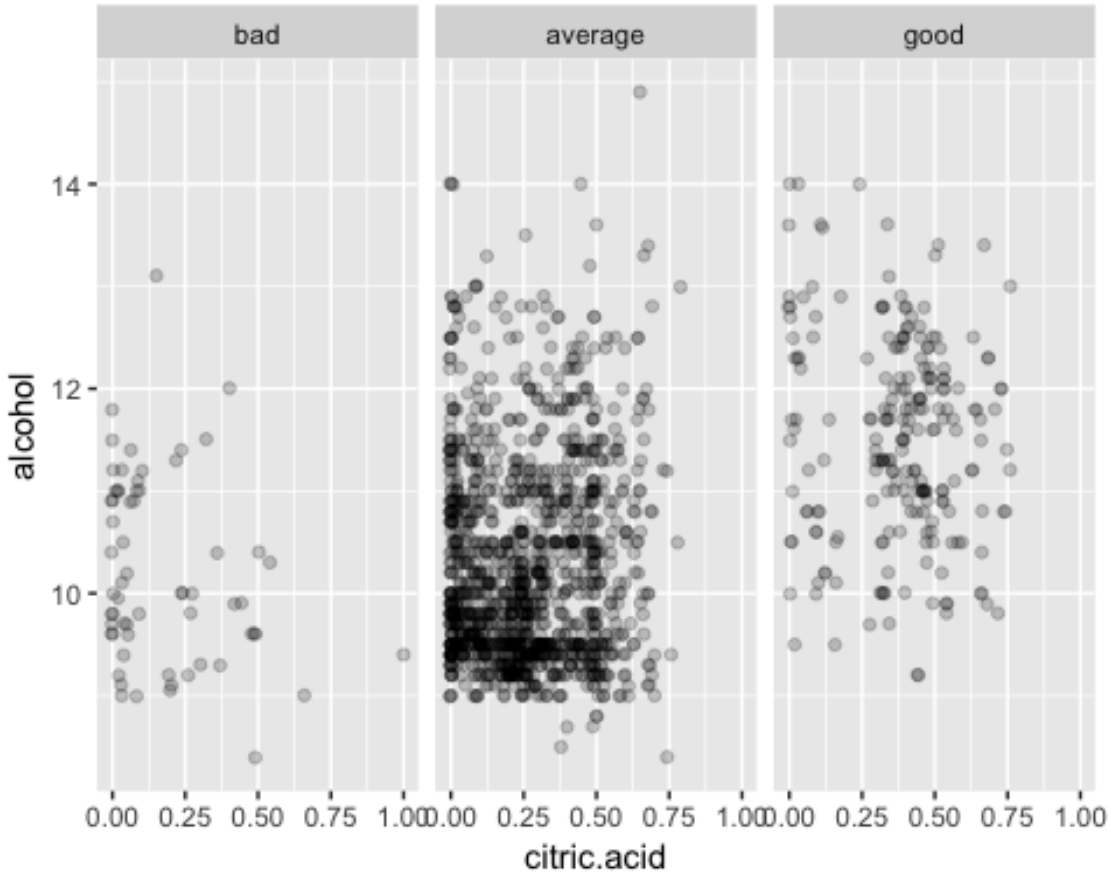
```
##        fixed.acidity     volatile.acidity          citric.acid
##           0.12405165          -0.39055778           0.22637251
##          TAC.acidity log10.residual.sugar      log10.chlordies
##           0.10375373           0.02353331          -0.17613996
##  free.sulfur.dioxide total.sulfur.dioxide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

By utilizing cor.test, I calculated the correlation for each of these variables against quality.

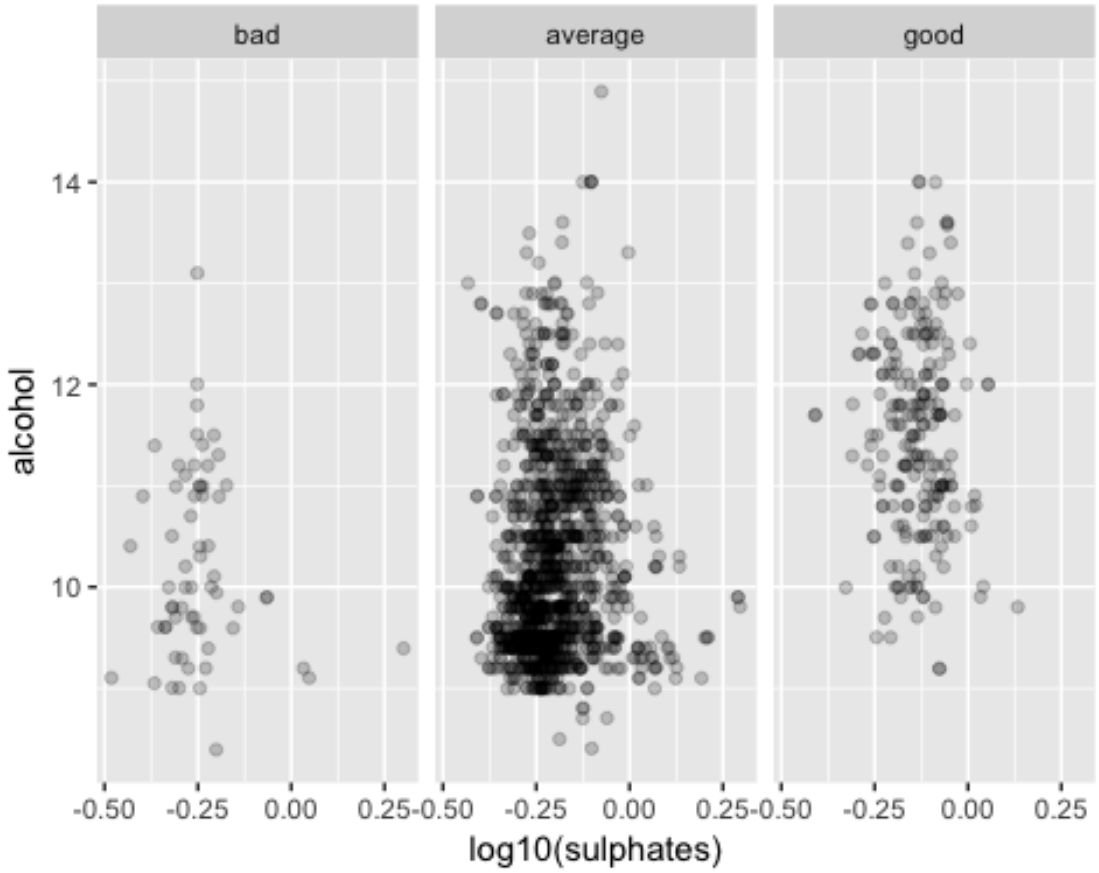
As we see, the **top 4** come to be: - **Alcohol**
- **Sulphates (log10)**
- **Volatile acidity**
- **Citric acid**

Let's plot them against each other using RATING as a facet.

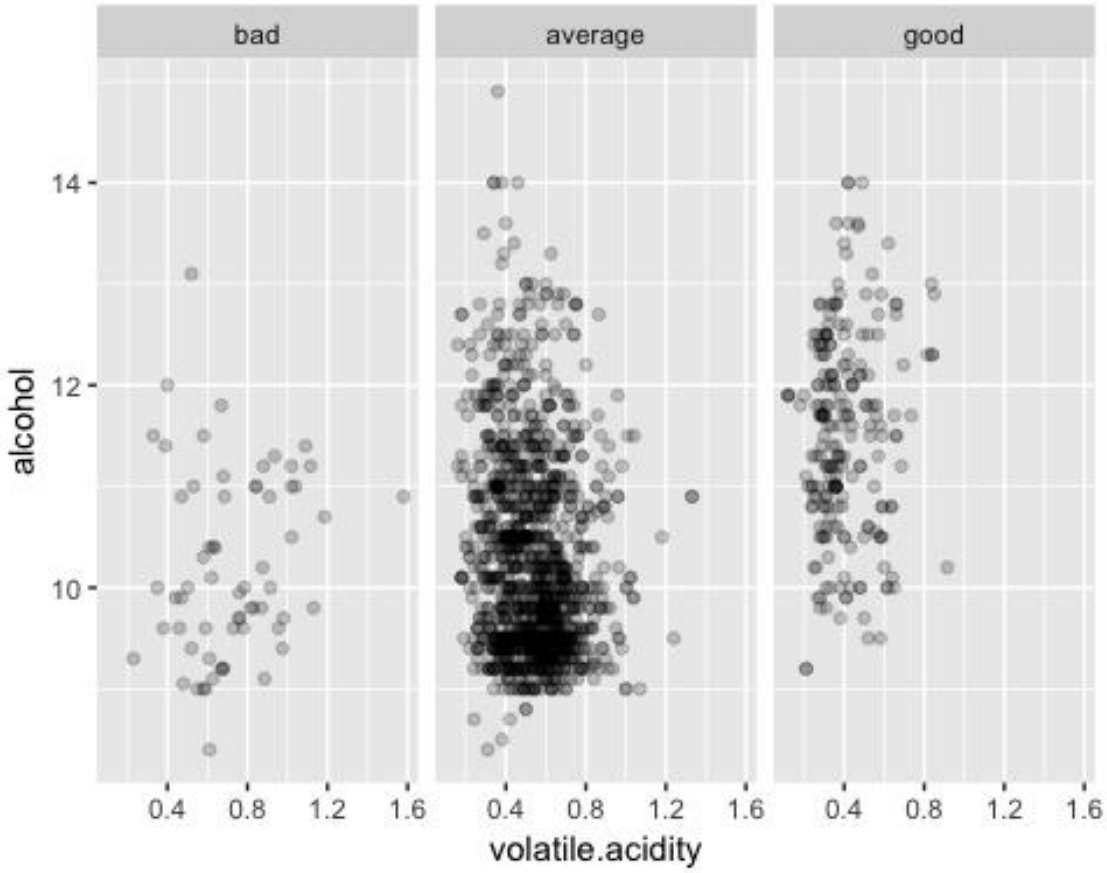
```
ggplot(data=Wine,aes(x=citric.acid,y=alcohol))+facet_wrap(~rating)+geom_jitte
r(alpha=0.2)
```



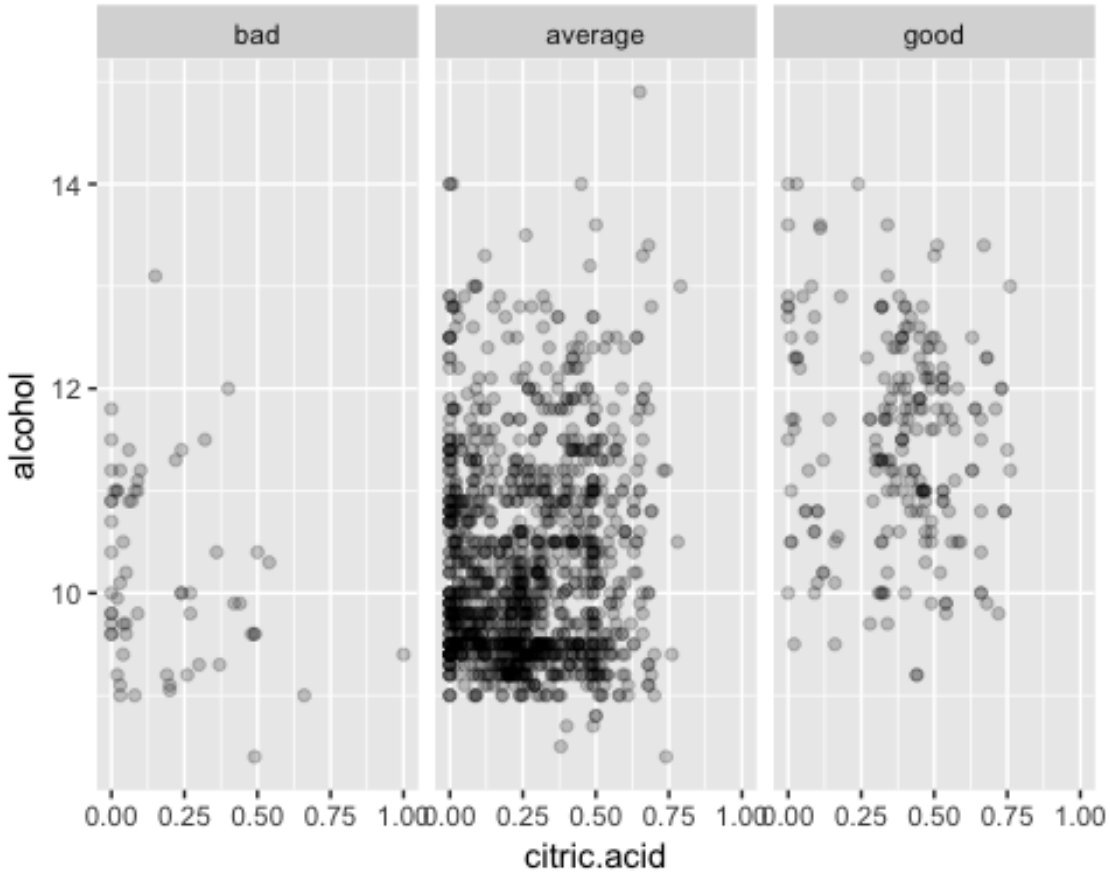

```
ggplot(data = Wine, aes(x = log10(sulphates), y = alcohol)) +
  facet_wrap(~rating) +
  geom_jitter(alpha=0.2)
```



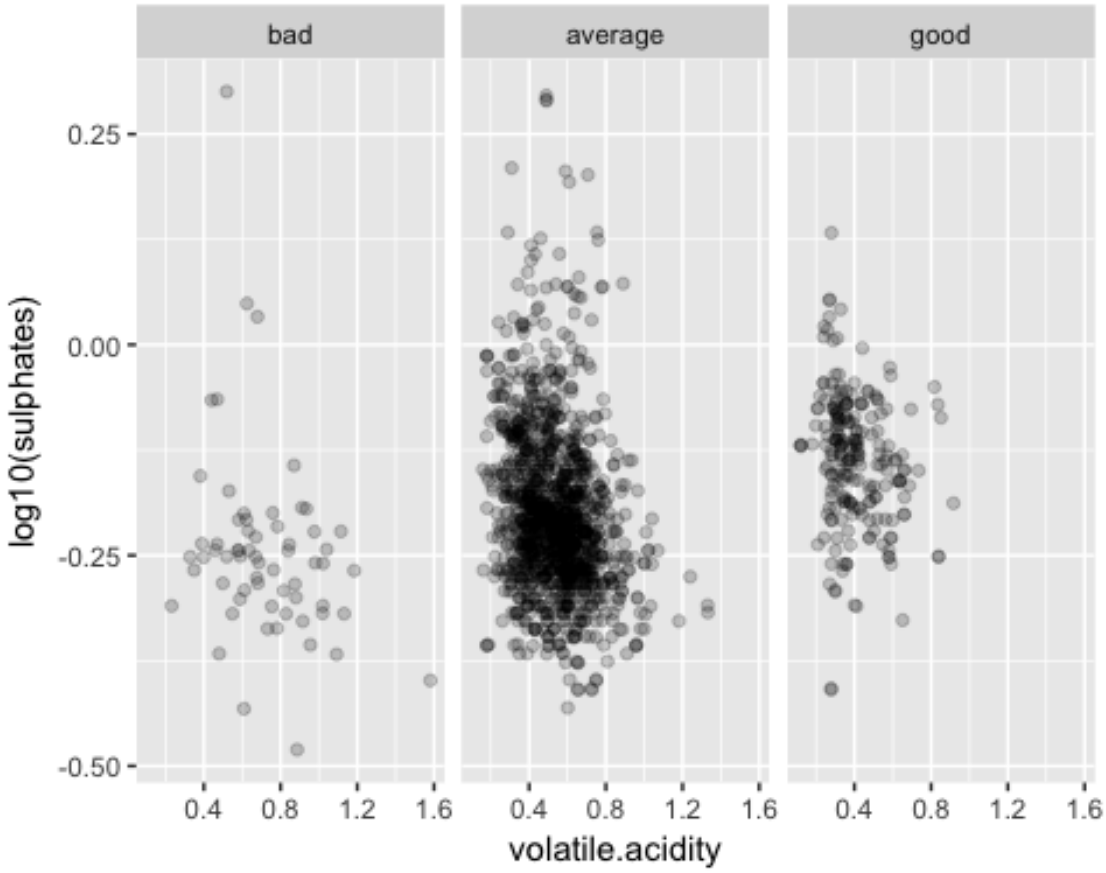

```
ggplot(data = Wine, aes(x = volatile.acidity, y = alcohol)) +
  facet_wrap(~rating) +
  geom_point(alpha=0.2)
```



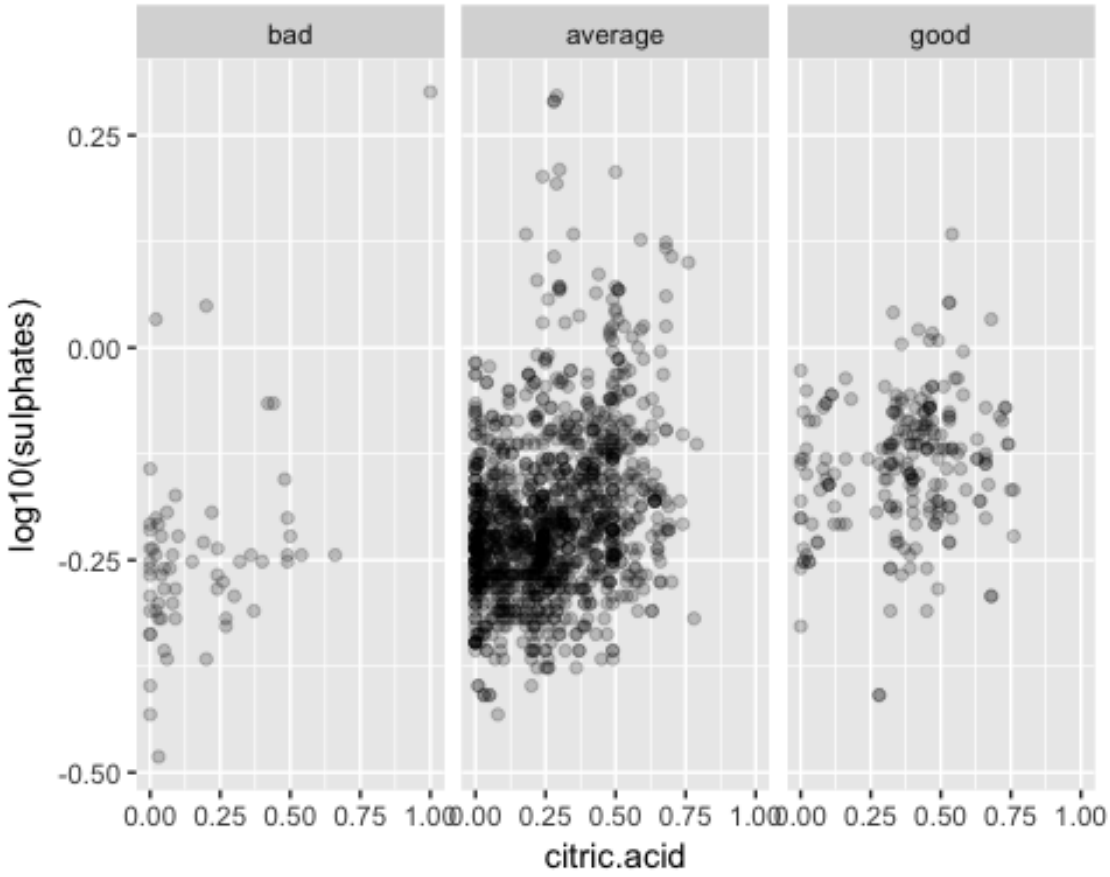

```
ggplot(data = Wine, aes(x = citric.acid, y = alcohol)) +
  facet_wrap(~rating) +
  geom_point(alpha=0.2)
```



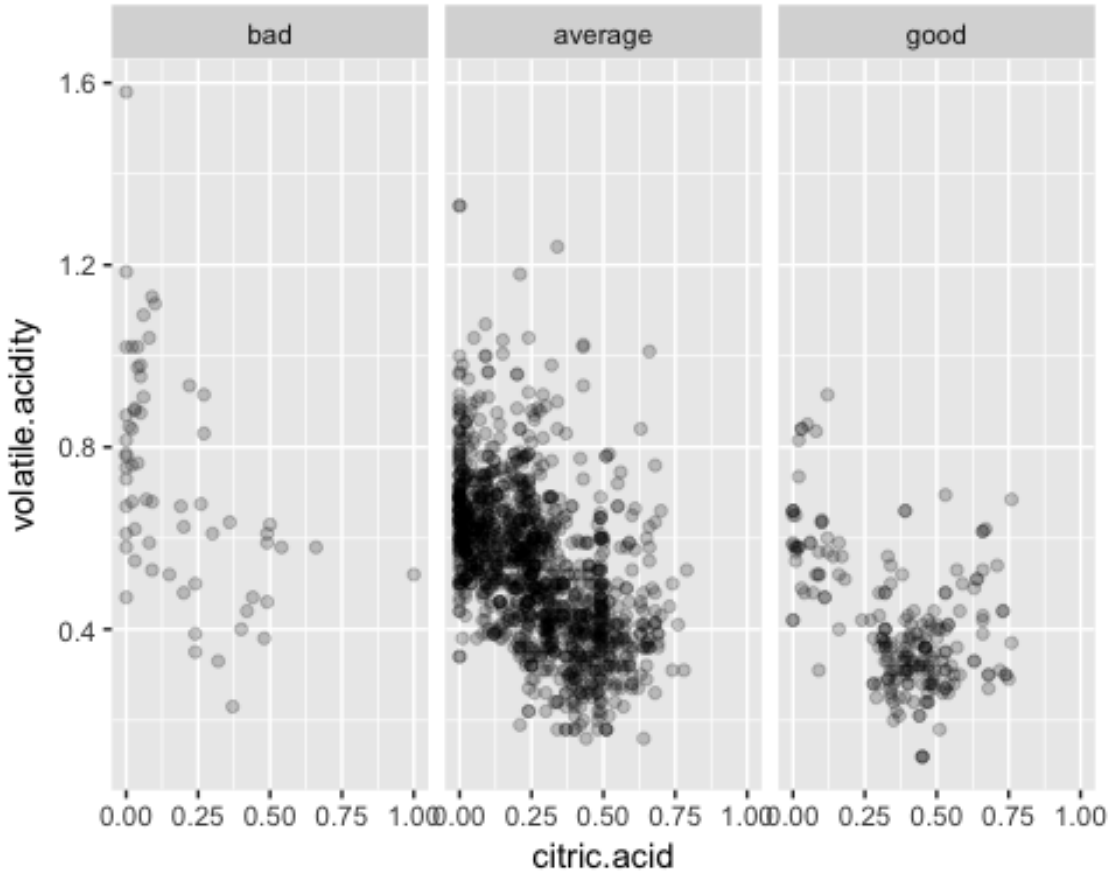

```
ggplot(data = Wine, aes(x = volatile.acidity, y = log10(sulphates))) +
  facet_wrap(~rating) +
  geom_jitter(alpha=0.2)
```




```
ggplot(data = Wine, aes(x = citric.acid, y = log10(sulphates))) +
  facet_wrap(~rating) +
  geom_point(alpha=0.2)
```



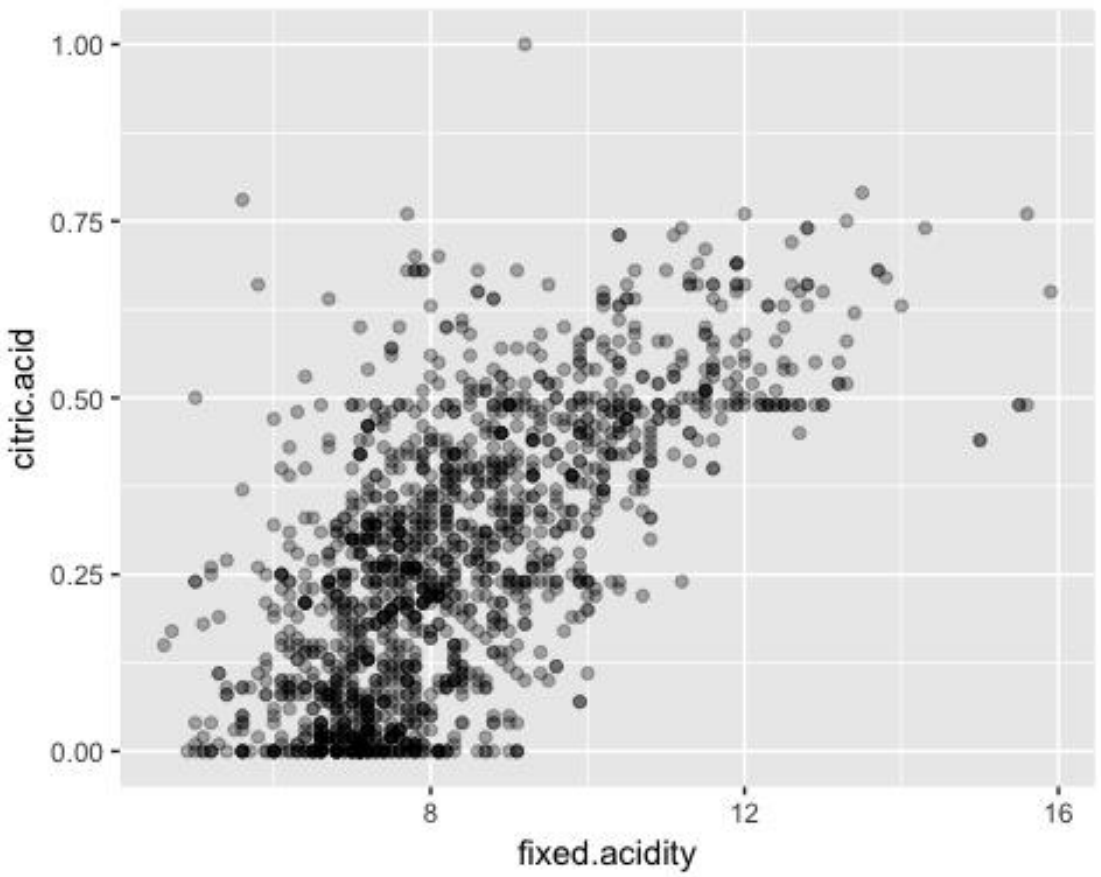

```
ggplot(data = Wine, aes(x = citric.acid, y = volatile.acidity)) +
  facet_wrap(~rating) +
  geom_point(alpha=0.2)
```



- The **relative value** of these scatterplots are suspect; if anything, it illustrates how **heavily alcohol** content affects **rating.**
- The **weakest bivariate relationship** appeared to be **alcohol vs. citric acid.**
- The plots were nearly **uniformly-distributed.**
- The **strongest relationship** appeared to be **volatile acididty vs. citric acid**, which had a **negative correlation.**

Lets examining the **acidity variables:**

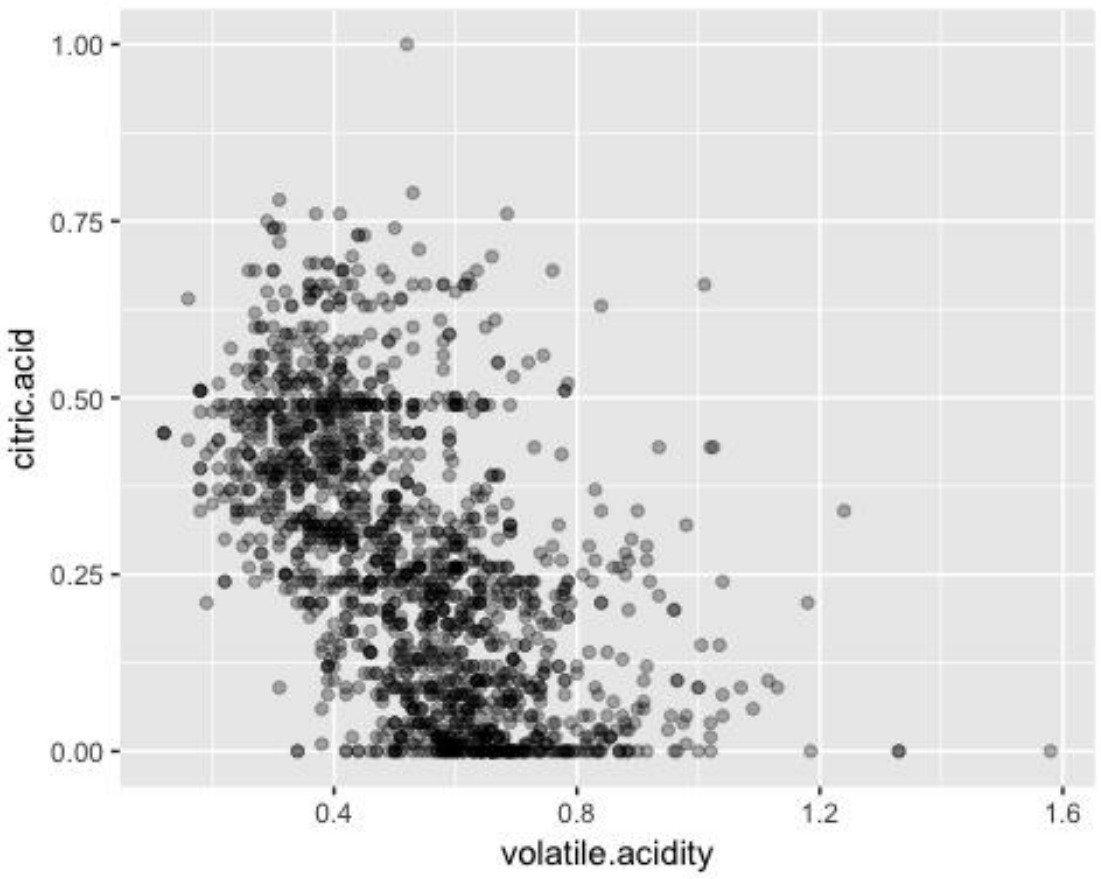
```
ggplot(data = Wine, aes(x = fixed.acidity, y = citric.acid)) +
  geom_point(alpha=0.3)
```




```
  cor.test(Wine$fixed.acidity, Wine$citric.acid)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Wine$fixed.acidity and Wine$citric.acid
## t = 36.234, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6438839 0.6977493
## sample estimates:
##       cor
## 0.6717034
```

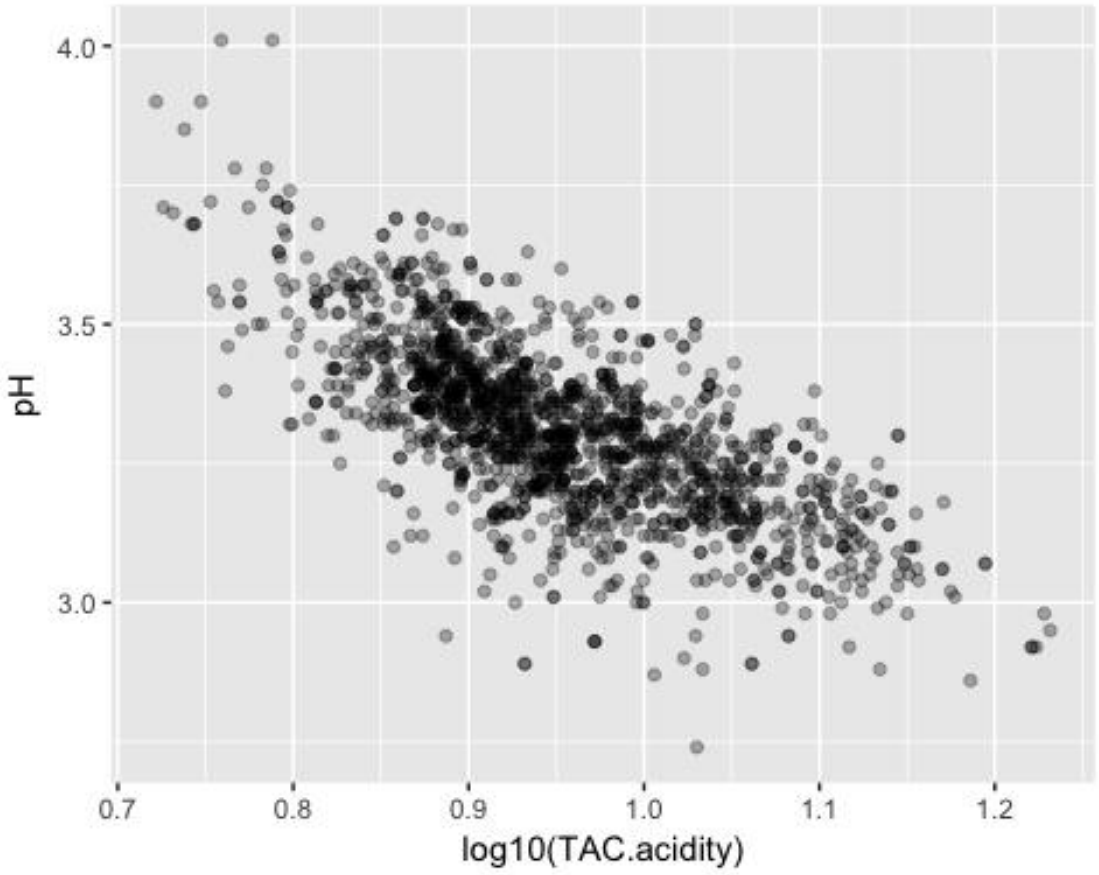
```
# plot_ly(Wine,x=~fixed.acidity,y=~citric.acid) ------ Using plotly to make
the visualization more interactive

ggplot(data = Wine, aes(x = volatile.acidity, y = citric.acid)) +
  geom_point(alpha=0.3)
```




```
cor.test(Wine$volatile.acidity, Wine$citric.acid)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Wine$volatile.acidity and Wine$citric.acid
## t = -26.489, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5856550 -0.5174902
## sample estimates:
##        cor
## -0.5524957
```

```
ggplot(data = Wine, aes(x = log10(TAC.acidity), y = pH)) +
  geom_point(alpha=0.3)
```




```r
cor.test(log10(Wine$TAC.acidity), Wine$pH)
```

```
##
##  Pearson's product-moment correlation
##
## data:  log10(Wine$TAC.acidity) and Wine$pH
## t = -39.663, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7283140 -0.6788653
## sample estimates:
##       cor
## -0.7044435
```

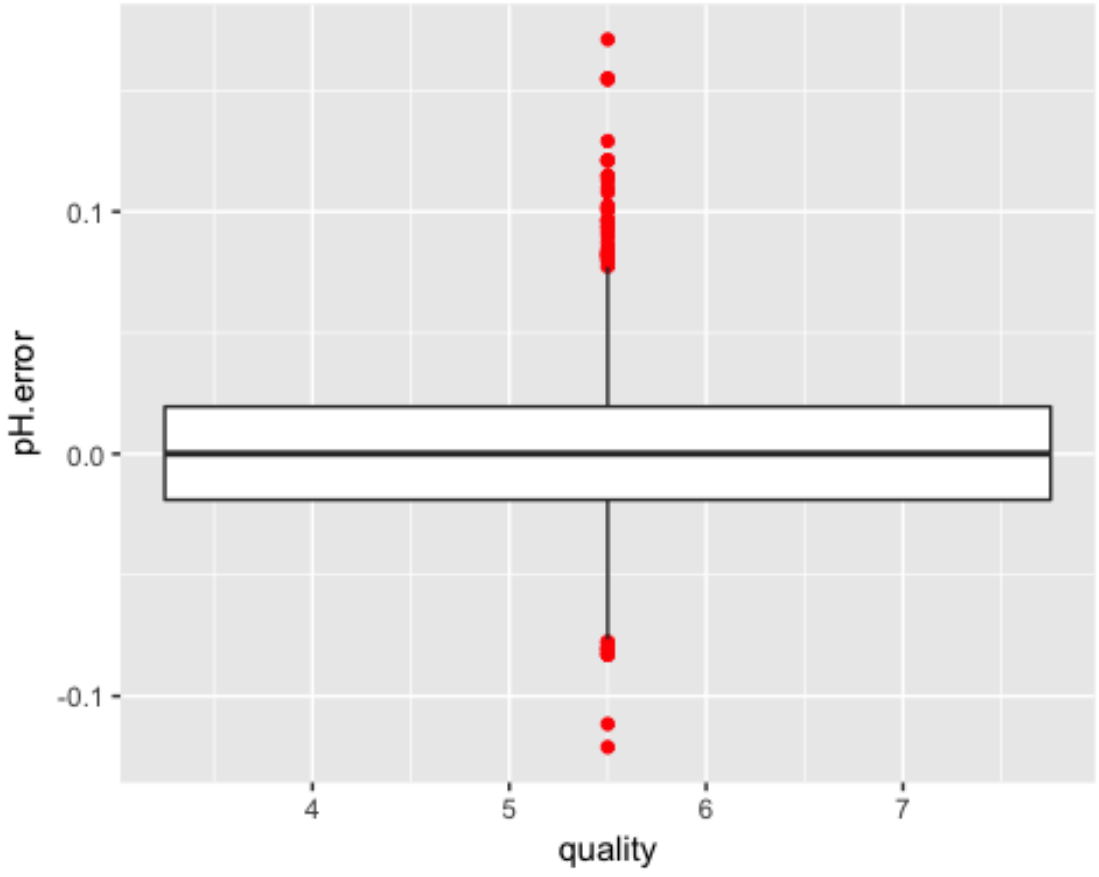
Upon examining, I observed **strong correlations** between the acidity variables.

- Most notably, base 10 logarithm **TAC.acidity** correlated very well with **pH**. This is certainly expected, as pH is essentially a measure of acidity.

An **interesting question** to pose, using basic chemistry knowledge, is to ask **what other components other than the measured acids are affecting pH**. We can quantify this

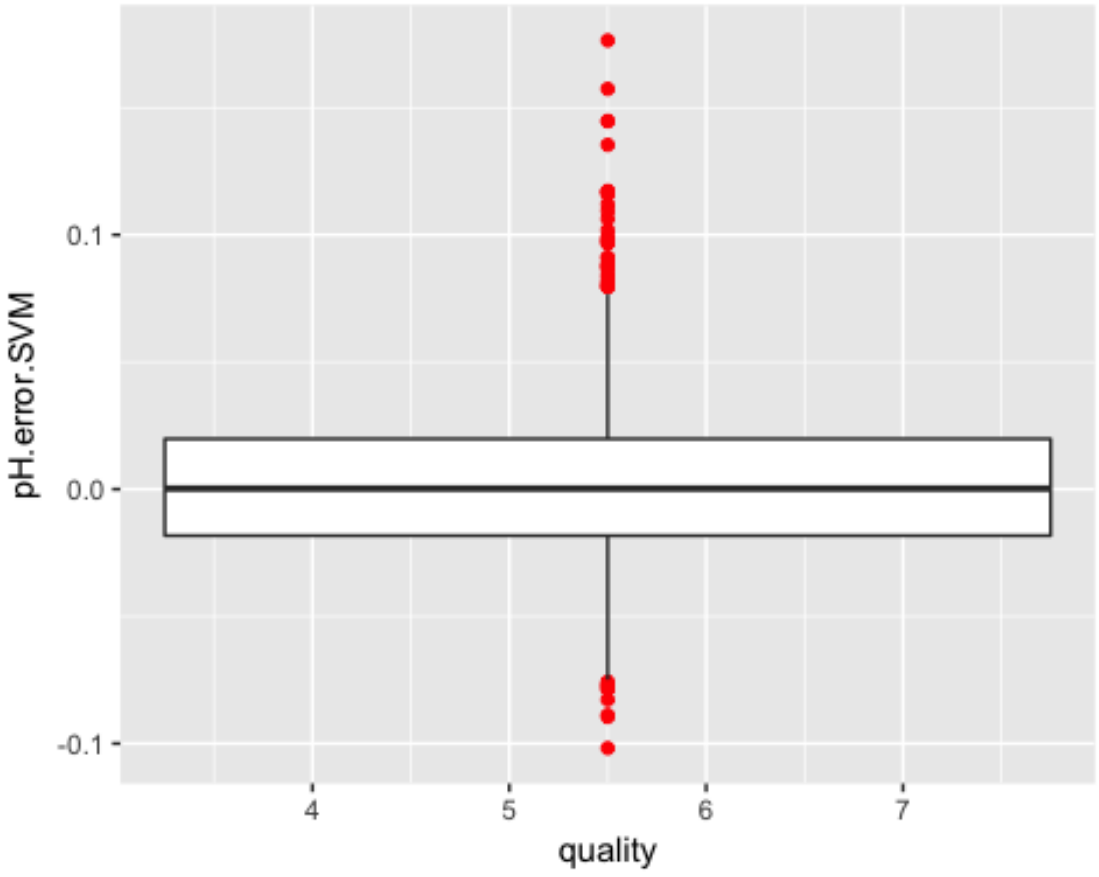
difference by **building a predictive linear model**, to **predict** pH based on TAC.acidity and capture the % difference as a new variable.

```
# Linear model
m <- lm(I(pH) ~ I(log10(TAC.acidity)), data = Wine)
Wine$pH.predictions <- predict(m, Wine)
# predict(m,Wine) --- to see linear prediction values

# (observed - expected) / expected
Wine$pH.error <- (Wine$pH.predictions - Wine$pH)/Wine$pH
# Wine$pH.error ----- to see linear prediction errors

ggplot(data = Wine, aes(x = quality, y = pH.error)) +
  geom_boxplot(outlier.colour = 'red')
```




```
# We can also add something interesting to our model, to check its accuracy.
# The RMS Error.

rmse <- function(error)
{
  sqrt(mean(error^2))
```

```
}

rmse(m$residuals)

## [1] 0.1095431

#Now, we train a Support Vector Machine.

library(e1071)

SVM <- svm(I(pH) ~ I(log10(TAC.acidity)), data = Wine)
Wine$pH.Predict.SVM <- predict(SVM,Wine)
# predict(SVM,Wine)  --- to see the SVM prediction values

Wine$pH.error.SVM <- (Wine$pH.Predict.SVM - Wine$pH)/Wine$pH

ggplot(data = Wine, aes(x = quality, y = pH.error.SVM)) +
  geom_boxplot(outlier.colour = 'red')
```




```
rmse(SVM$residuals)

## [1] 0.106785
```

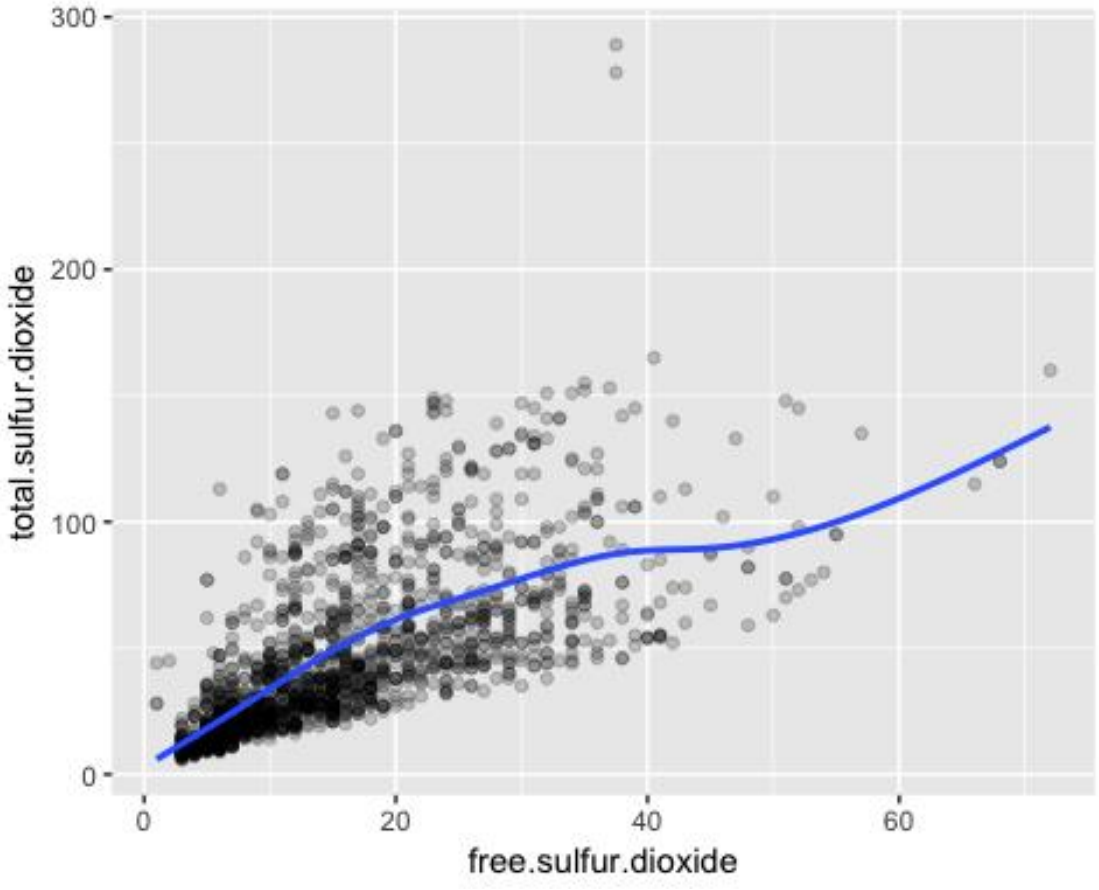
Linear Model and Support Vector Machine

We see that a **SVM** functions **slightly better** than a **LM**.
- The **first RMS** is of the **LM**, the **second** of the **SVM**.
- The **median % error** hovered at or **near zero** for most wine qualities.
- Notably, **wines** rated with a **quality of 3** had **large negative error.**
- We can interpret this finding by saying that for many of the 'bad' wines, total acidity from tartaric, acetic, and citric acids were a worse predictor of pH. **Simply** put, it is likely that there were other components–possibly impurities–that changed and affected the pH.

As **annotated previously**, we **hypothesized that** free.sulfur.dioxide and total.sulfur.dioxide were dependent on each other.
Plotting this:

```
ggplot(data = Wine, aes(x = free.sulfur.dioxide, y = total.sulfur.dioxide)) +
  geom_point(alpha=0.2) +
  geom_smooth(se=F)
```

```
## `geom_smooth()` using method = 'gam'
```



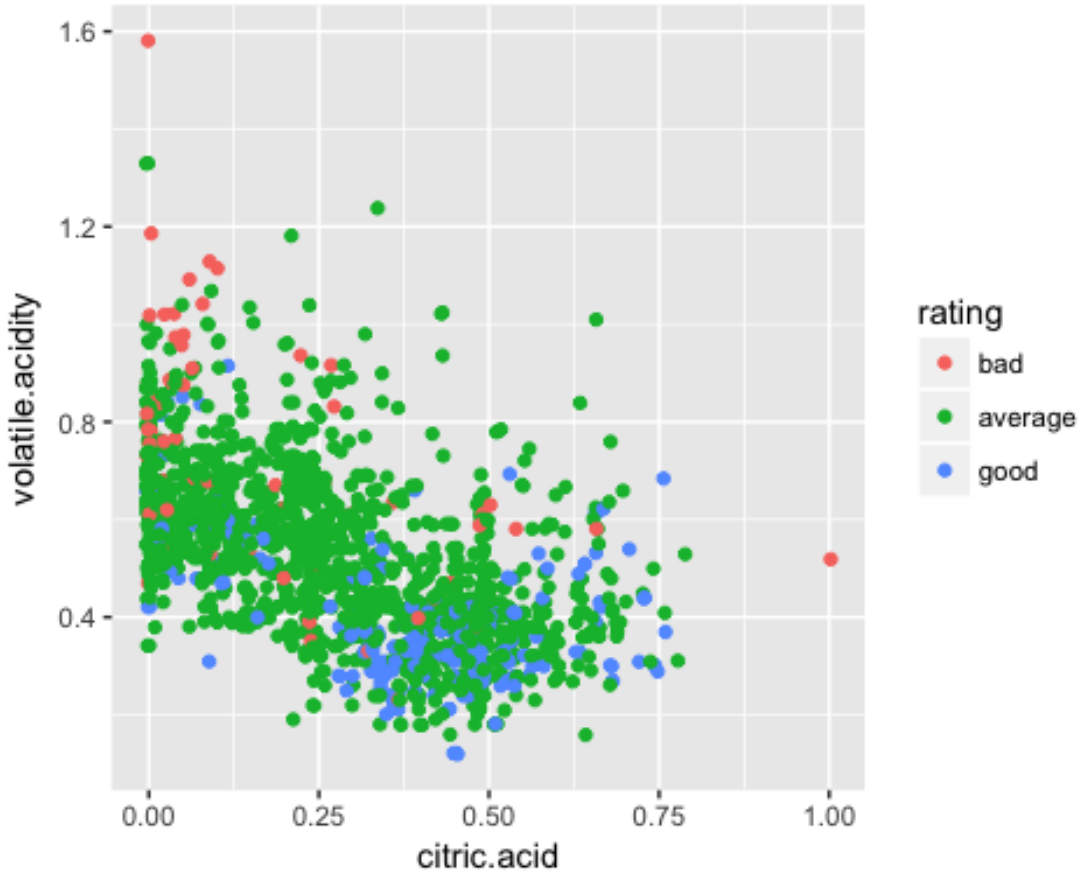

```
cor.test(Wine$free.sulfur.dioxide, Wine$total.sulfur.dioxide)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Wine$free.sulfur.dioxide and Wine$total.sulfur.dioxide
## t = 35.84, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6395786 0.6939740
## sample estimates:
##       cor
## 0.6676665
```

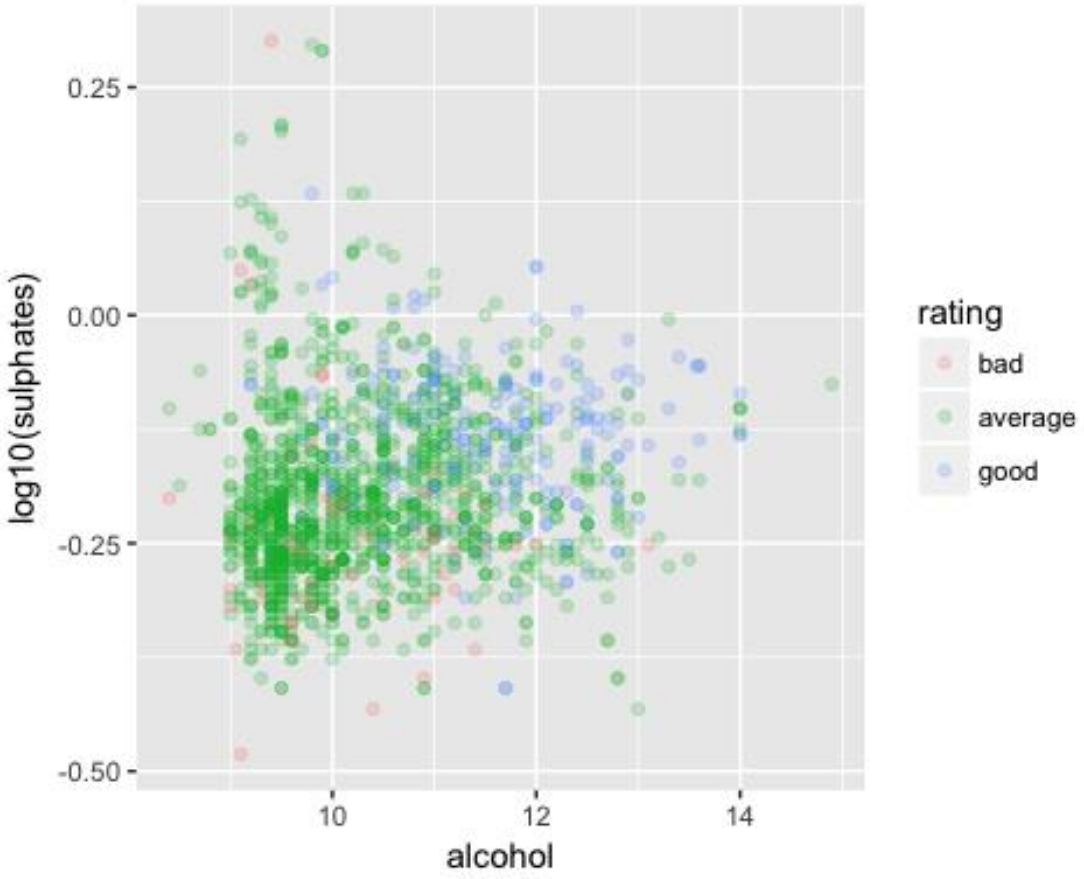
- It is clear that there is a very **strong relationship** between the two.
- Aside from TAC.acidity, this seemed to be the **strongest bivariate relationship.**
- **Additionally**, despite the telling name descriptions, the clear 'floor' on this graph hints that free.sulfur.dioxide is a subset of total.sulfur.dioxide.

MULTIVARIATE PLOTS SECTION

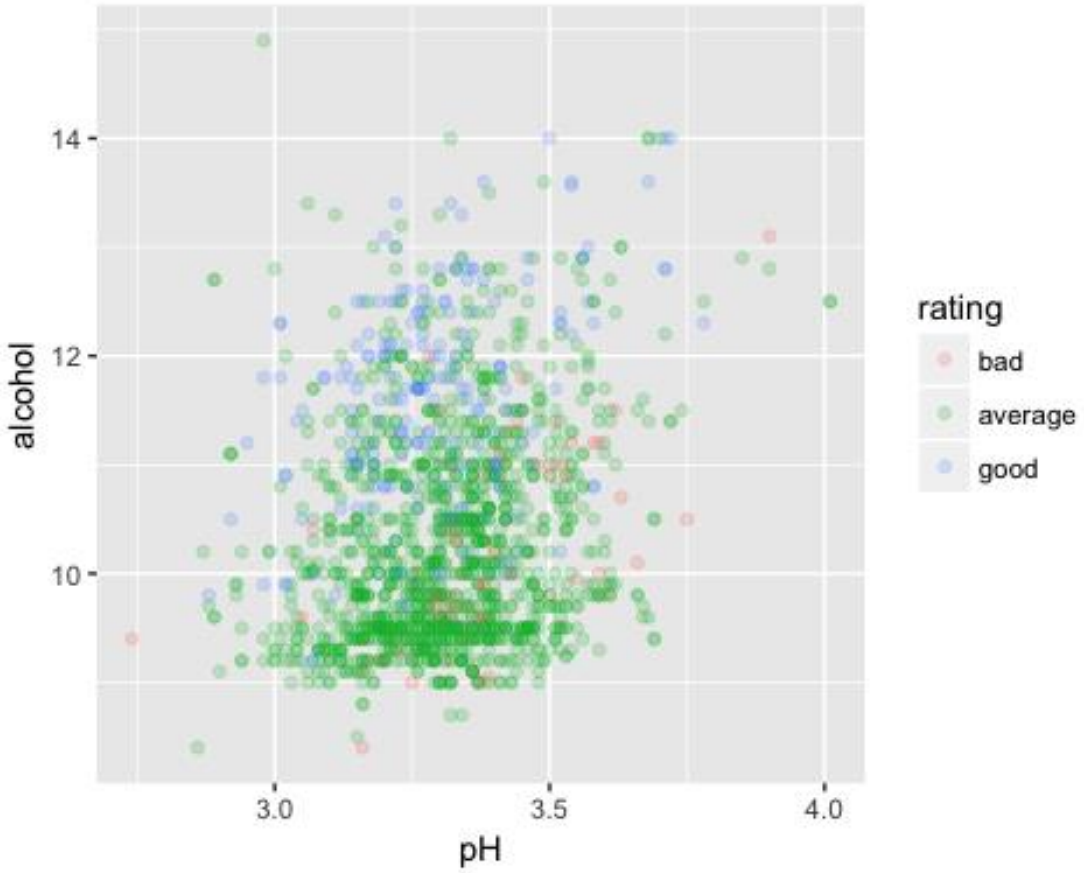
```
ggplot(data = Wine,
       aes(x = citric.acid, y = volatile.acidity,
           color = rating)) +
  geom_jitter(alpha=1)
```



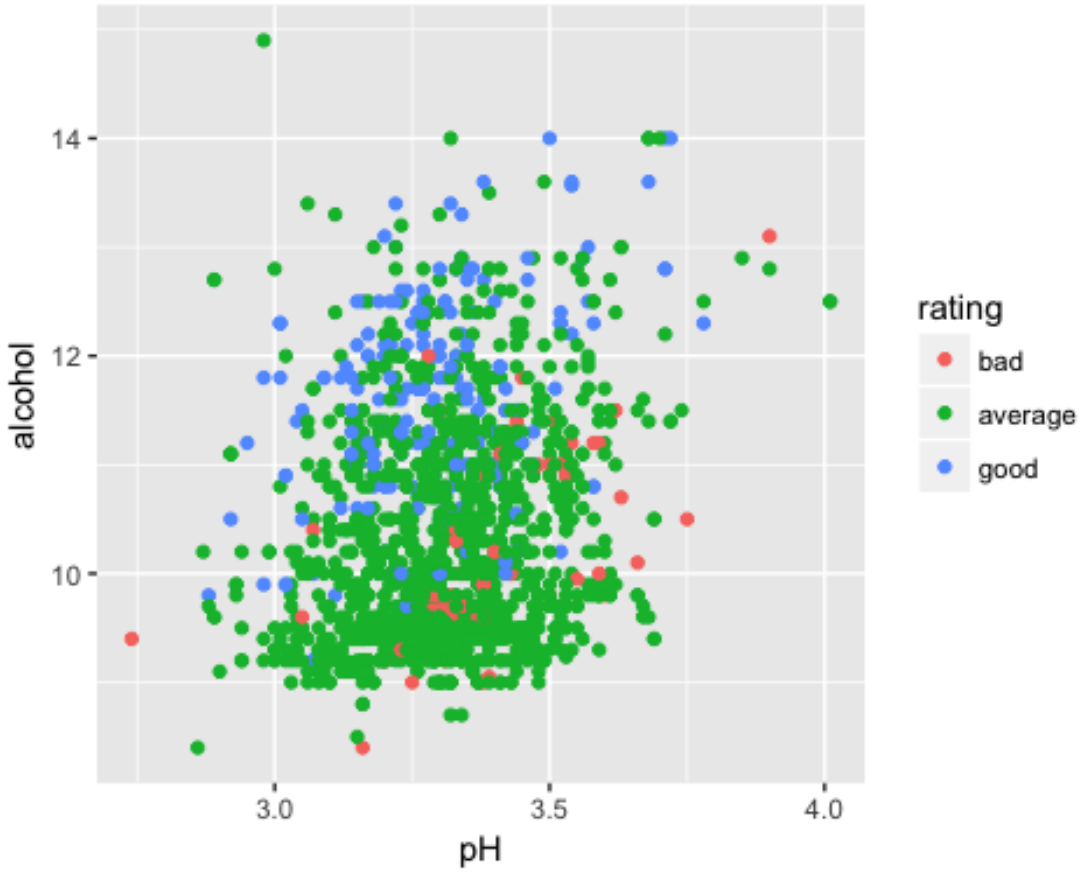

```
ggplot(data = Wine,
       aes(x = alcohol, y = log10(sulphates),
           color = rating)) +
  geom_point(alpha=0.2)
```




```
ggplot(data = Wine,
       aes(x = pH, y = alcohol, color = rating)) +
  geom_point(alpha=0.2)
```




```
ggplot(data = Wine,
       aes(x = pH, y = alcohol, color = rating)) +
  geom_point()
```



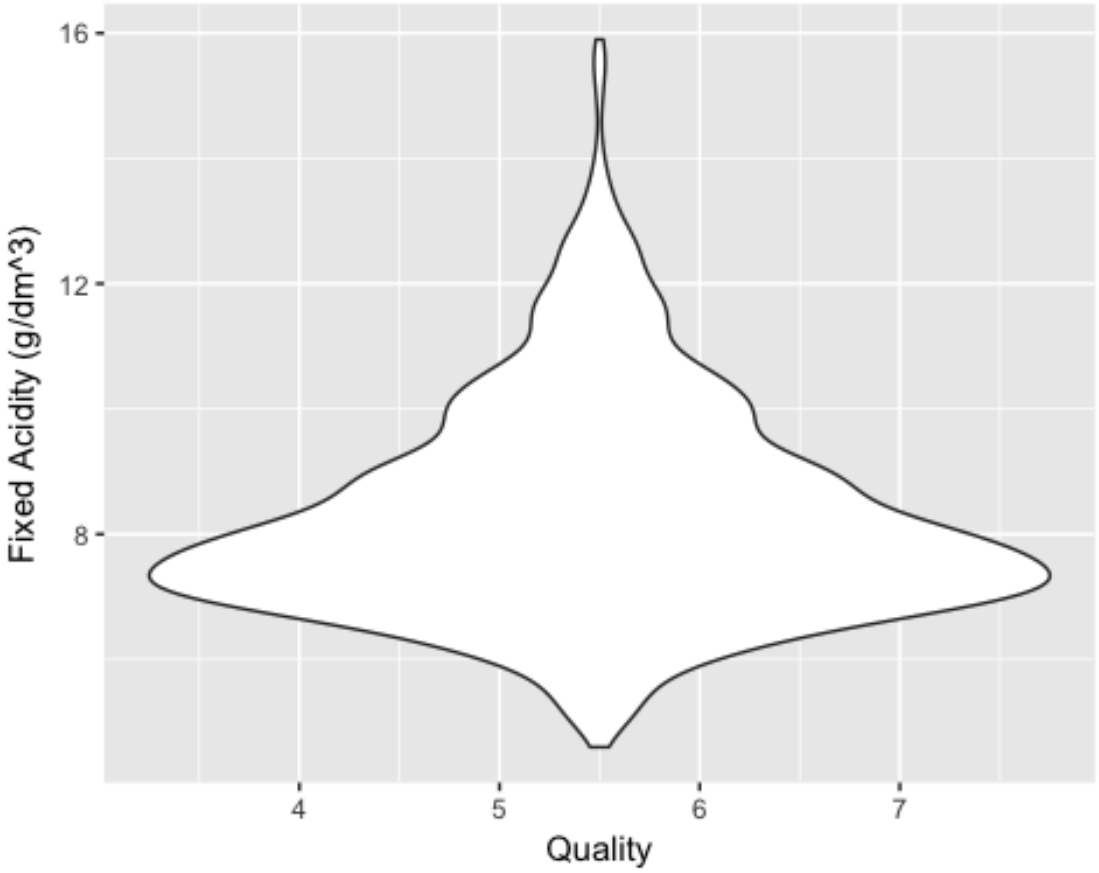
From the plots we come to know the following:

- We primarily examined the **4 features** which showed high correlation with **quality.**
- These scatterplots were a bit crowded, so we faceted by **rating AND by quality** in the final plot to illustrate clearly and a little more about the population differences between **good wines, average wines, and bad wines.**
- It's clear that a **higher citric acid and lower volatile (acetic) acid** contributes towards better wines. **Likewise**, better wines tended to have **higher sulphates and alcohol content.** Surprisingly, **pH** had very little visual impact on wine quality, and was shadowed by the **larger impact of alcohol.**
- Interestingly, this shows that what makes a **good wine** depends on the **type of acids** that are present.

FINAL PLOTS AND SUMMARY

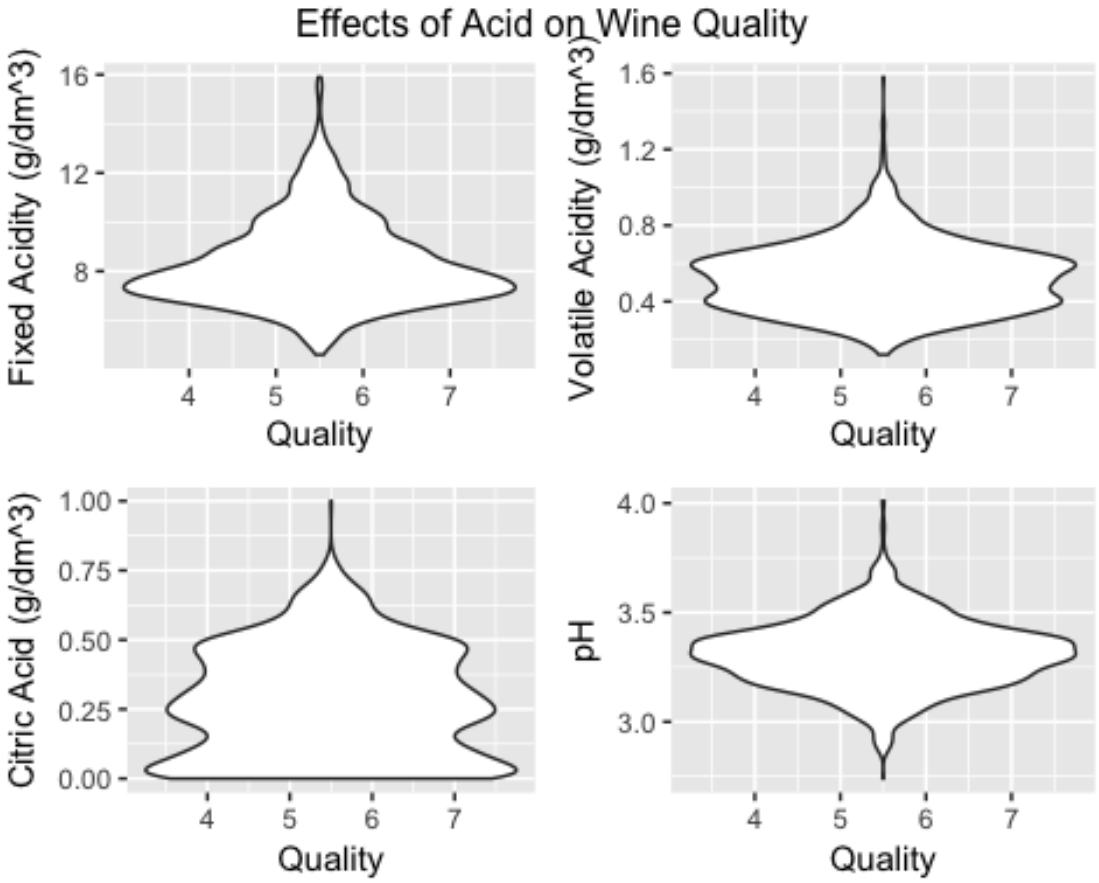
PLOT ONE: EFFECTS OF ACID ON WINE QUALITY

```
### Plot One: Effects of Acid on Wine Quality
ggplot(data = Wine, aes(x = quality, y = fixed.acidity,
fill = quality)) +
ylab('Fixed Acidity (g/dm^3)') +
xlab('Quality') +
geom_violin()+guides(fill=F)
```




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

```
xlab('Quality') +
geom_violin()+guides(fill=F),
ggplot(data = Wine, aes(x = quality, y = pH,
fill = quality)) +
ylab('pH') +
xlab('Quality') +
geom_violin()+guides(fill=F),top='Effects of Acid on Wine Quality')
```



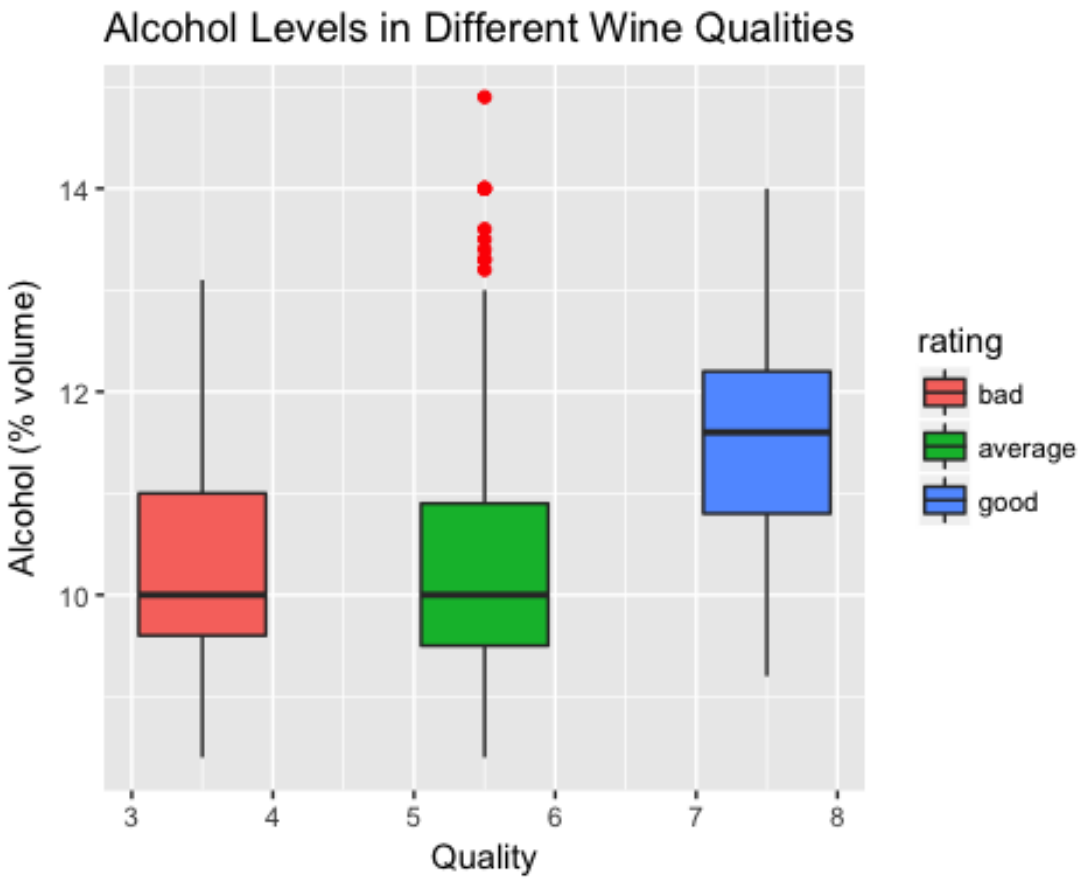
SUMMARY ONE

- These **subplots** were created to demonstrate the effect of acidity and pH on wine quality. Generally, **higher acidity** (or lower pH) is seen in **highly-rated wines.**

- To caveat this, a presence of **volatile (acetic) acid negatively affected wine quality. Citric acidity** had a **high correlation** with **wine quality**, while fixed (tartaric) acid had a smaller impact.

PLOT TWO: EFFECTS OF ALCOHOL ON WINE QUALITY

```
### Plot Two: Effect of Alcohol on Wine Quality
ggplot(data = Wine, aes(x = quality, y = alcohol,
fill = rating)) +
geom_boxplot(outlier.color = 'red') +
ggtitle('Alcohol Levels in Different Wine Qualities') +
xlab('Quality') +
ylab('Alcohol (% volume)')
```

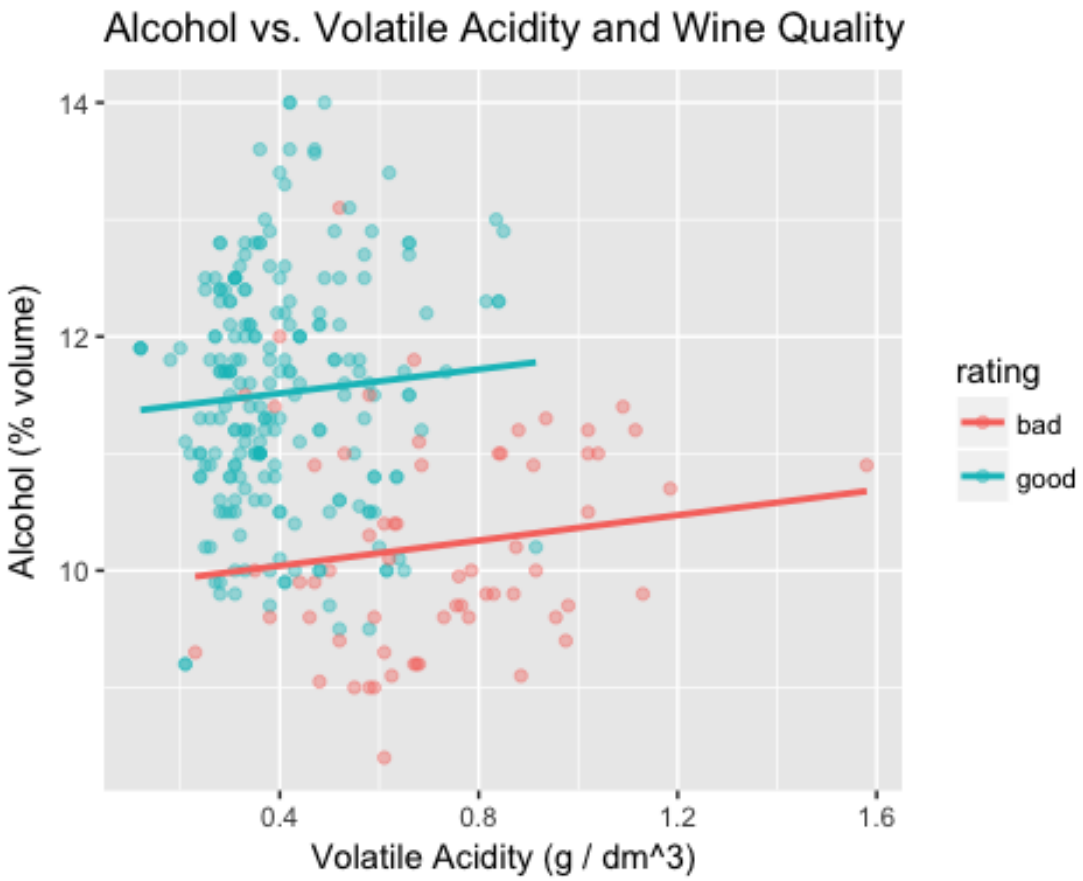


SUMMARY TWO

- These boxplots demonstrate the **effect of alcohol** content on wine **quality.**
- Generally, **higher alcohol content correlated with higher wine quality.**
- **However**, as the outliers and intervals show, **alchol** content alone did not produce a **higher quality.**

PLOT THREE: What makes good wines, good, and bad wines, bad?

```
ggplot(data = subset(Wine, rating != 'average'),
aes(x = volatile.acidity, y = alcohol,
color = rating)) +
geom_point(alpha=0.4) +
ggtitle('Alcohol vs. Volatile Acidity and Wine Quality') +
xlab('Volatile Acidity (g / dm^3)') +
ylab('Alcohol (% volume)') + geom_smooth(method = 'lm',se=F)
```




```
# plot_ly(Wine,x=~rating,y=~volatile.acidity,z=~alcohol,type='scatter3d') ---
-- Using plotly to make the visualization more interactive
```

SUMMARY THREE

- This is perhaps the **most telling graph.** We subsetted the data to remove the 'average' wines, or any wine with a rating of **5 or 6.** As the **correlation tests** show, **wine quality** was affected most **strongly by alcohol and volatile acidity.**

- **While the boundaries are not as clear cut or modal**, it's apparent that **high volatile acidity**–with few exceptions–kept **wine quality down.**

- A **combination** of high alcohol content and low volatile acidity **produced** better wines.

MAIN CONCLUSION

- Through this **exploratory data analysis**, we were able to identify the **key factors** that determine and drive wine quality, mainly: **alcohol content, sulphates, and acidity.**

- It is important to note, **however**, that **wine quality** is ultimately a **subjective measure**, albeit measured by wine experts. That said, the **correlations for these variables are within reasonable bounds.**
- The **graphs** adequately illustrate the factors that make good wines 'good' and bad wines 'bad'.
- **Further study with inferential statistics** could be done to quantitatively confirm these assertions.

- One of the **challenges that we faced** was the lack of a clear analysis about the **quality of our wine.** The ordered factor **"Quality"** was not very helpful and to overcome this, we created another variable **"Rating".**

- To make predictions of wine quality and any other if required, we trained two models. As we saw before, the **linear model and the Support Vector Machine.**
- The **SVM performed marginally better** and we decided to stick with it if we had to make any more predictions.
