Chemical Properties Influencing Red Wine's Quality

Outline: Introduction | Summary of Data | Univariate Analysis | Bivariate Analysis | Multivariate Analysis | Regression Analysis | Summary

Introduction

The goal of this project is to determine chemical properties that influence the quality of red wines. Since this is a project for data uni-, bi-, and multivariate data visualization in R, the crux of the analyses focuses on such data visualization and miltivariate analyses. The visualizations provide depictions on variable distributions and relations between variables, which lend to model building in the regression analyses.

I don't drink, so, I can't provide much personal opinions on what chemical properties affecting red wine's quality. If this was my thesis, this section would be filled with literature reviews on wine preference and red wine. However, for this project, I'll limit my literature review based on what I gathered from the paper that included the scoring methodology and dataset.

Different researchers have different preference on how the approach data but the end goal should be very similar, i.e. to answer the research question(s). My approach is to combine my programming structure with the analysis story, which for the most, should coincide.

Summary of Data

Before reading and exploring data, I like to reserve the intro section of R programming to gather the needed packages and of course data reading process.

```
library(car); library(plyr); library(reshape2); library(data.table); library(ggplot2); library(GGally); library(memisc)
; library(grid); library(gridExtra); library(MASS); library(ordinal); library(tinytex)
# libz<-c("car","plyr","reshape2","data.table","ggplot2","GGally","memisc","grid","gridExtra","MASS","ordinal")
# lapply(x,install.packages(libz),character.only=T)
# lapply(libz,library,character.only=T)

############ Data Read ############
getwd() # Current Directory

## [1] "C:/Users/FA279J/Documents/Edu/DAND/rProject"

setwd("C:/Users/FA279J/Documents/Edu/DAND/rProject")
# list.files()
red1 <- read.csv("wineQualityReds.csv", header=TRUE, fill=TRUE)

############ Data Summary ############
cat("\n\nThe dataset's dimension is ",dim(red1),".\n")
```

```
##
##
## The dataset's dimension is  1599 13 .
```

```r
cat("\n\nUnivariate Statistics of Variables:-\n")
```

```
##
##
## Univariate Statistics of Variables:-
```

```r
summary(red1)
```

```
##       num         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
```

Obtaining data information and statistics allows researchers to not only get introduced to the data but before that to check if the data are read correctly. From experience, it's a good practice to observe the overall data outlook:

- names() lists all the variables: matched data description;
- dimension() provides the numbers of rows and columns: matched;
- summary() spits out each variable's distribution: seems to be believable and conforms to data description ranges;
- class() looks into variable classification: here, I look into data format for further analysis.
- head() would be used to double check data layout.

```
sapply(red1, class)

##                  num        fixed.acidity     volatile.acidity
##            "integer"            "numeric"            "numeric"
##          citric.acid       residual.sugar            chlorides
##            "numeric"            "numeric"            "numeric"
##  free.sulfur.dioxide total.sulfur.dioxide              density
##            "numeric"            "numeric"            "numeric"
##                   pH            sulphates              alcohol
##            "numeric"            "numeric"            "numeric"
##              quality
##            "integer"

head(red1)

##   num fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1   1           7.4             0.70        0.00            1.9     0.076
## 2   2           7.8             0.88        0.00            2.6     0.098
## 3   3           7.8             0.76        0.04            2.3     0.092
## 4   4          11.2             0.28        0.56            1.9     0.075
## 5   5           7.4             0.70        0.00            1.9     0.076
## 6   6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
## 4                  17                   60  0.9980 3.16      0.58     9.8
## 5                  11                   34  0.9978 3.51      0.56     9.4
## 6                  13                   40  0.9978 3.51      0.56     9.4
##   quality
## 1       5
## 2       5
## 3       5
```

```
## 4        6
## 5        5
## 6        5
```

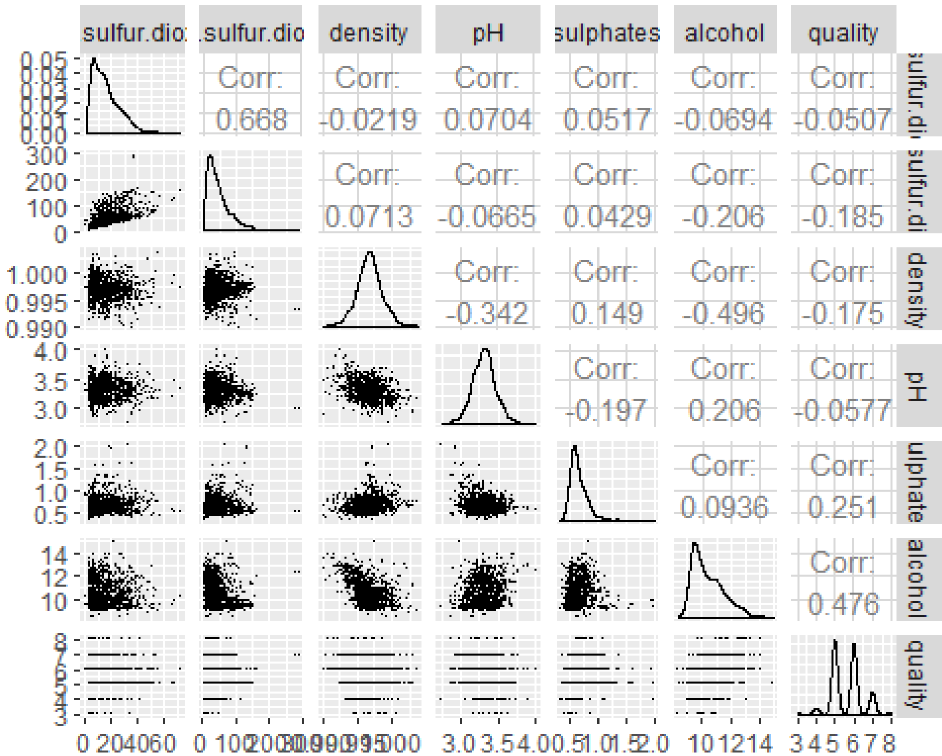
Univariate Analysis

Given that the dataset does not contain too many variables, I also start by looking into a scatter plot matrix. Initially, I look into the diagonally-placed distribution plots. Later on, I would revisit this plot to get a quick picture on relations between variables. From the scatterplot matrix below, I can quickly see the followings:

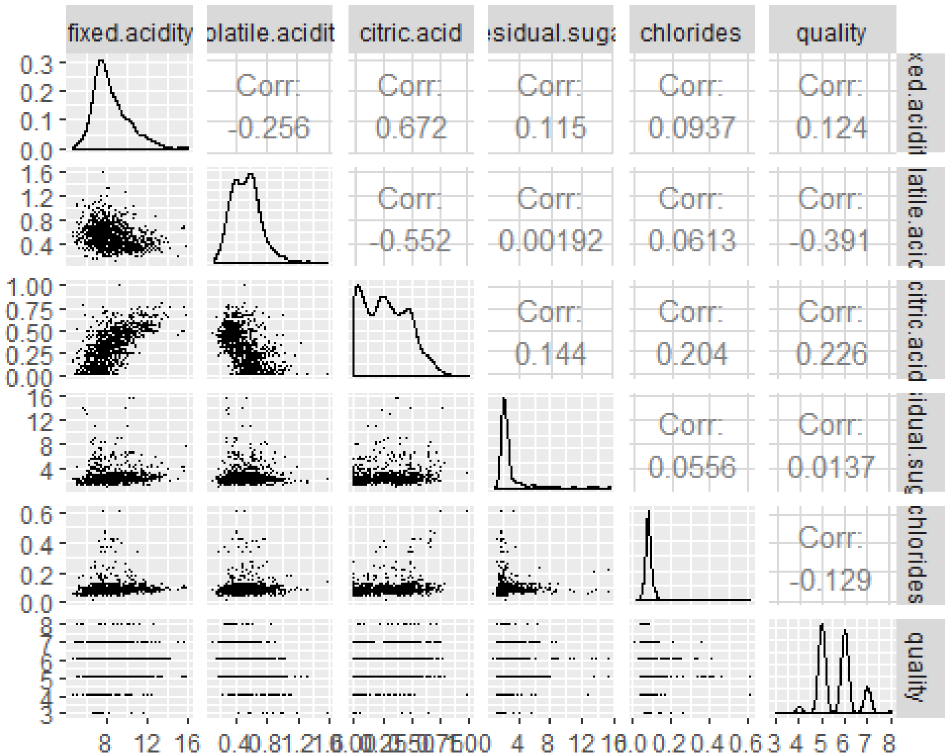
- residual.sugar, chlorides, and total.sulfur.dioxide variables need data transformation to obtain a more normal distribution
- quality has a weird up and down distribution. Based on the data description and bottom row scatter plots, this is probably to non-decimal round number quality score for red wine quality. Regarding quality, one can easily see that bulk of response for quality are in the two middle values of quality. This is especially disconcerting because lack of spread for a dependent variable may not produce a highly reliable linear regression estimation.
- the column for the dependent variable (quality) provides extra interesting insights.

One can see that alcohol is strongly correlated to quality with fixed.acidity, volatile.acidity and sulphates having weak correlations to quality. These correlations may in the end tell which chemical properties are important. Though, it's important note that these are untransformed variables and the strength of variables may be strengthen or dampen when other variables are considered in a model.

```r
ggpairs(red1[c(7:13)],
  lower = list(continuous = wrap("points", shape = I('.'))),
  upper = list(combo = wrap("box", outlier.shape = I('.'))))
```




```r
ggpairs(red1[c(2:6,13)],
  lower = list(continuous = wrap("points", shape = I('.'))),
  upper = list(combo = wrap("box", outlier.shape = I('.'))))
```



The correlation matrix is great at showing an overall picture. However, some relations or charts may need some detailed inspections, which may lead to data transformation. From the correlation matrix above, we can see that:

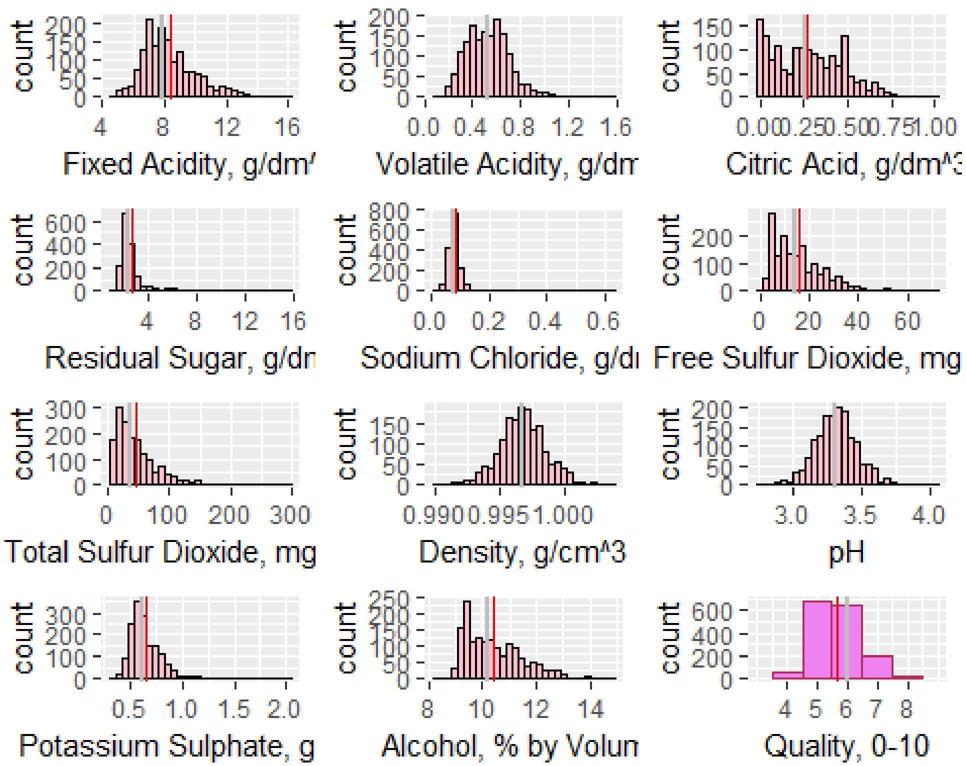
- quality has a weird distribution;
- residual.sugar and chlorides have serious deviation from normal distribution; and
- fixed.acidity, volatile.acidity, citric.acid, free.sulfur.dioxide, total.sulfur.dioxide, sulphates, and alcohol have considerable deviation from a normal distribution.

After getting an overall picture of variable set, I may step back to look into each variable's histogram. Visualization below is a compilation of each variable's distribution with the means (red vertical lines) and medians (grey vertical lines) included in the charts. For seasoned researchers, the histogram mat below does not add any significant information to the correlation matrix above. The correlation matrix should be able to indicate if a variable's mean is greater that its median - e.g. based on a distribution skewness. I considered having the mean and median values for each variable in the plot but the chart is getting too busy.

```
p01 <- ggplot(red1, aes(x=fixed.acidity)) +
  geom_histogram(color = 'black',fill = I('pink')) +
  xlab('Fixed Acidity, g/dm^3') +
  geom_vline(aes(xintercept = mean(fixed.acidity)),col='red',size=0.5) +
```

```
  geom_vline(aes(xintercept = median(fixed.acidity)), col = 'grey', size=1)

p02 <- ggplot(red1, aes(x=volatile.acidity)) +
  geom_histogram(color = 'black',fill = I('pink')) +
  xlab('Volatile Acidity, g/dm^3') +
  geom_vline(aes(xintercept = mean(volatile.acidity)),col='red',size=0.5) +
  geom_vline(aes(xintercept = median(volatile.acidity)), col = 'grey', size=1)

p03 <- ggplot(red1, aes(x=citric.acid)) +
  geom_histogram(color = 'black',fill = I('pink')) +
  xlab('Citric Acid, g/dm^3') +
  geom_vline(aes(xintercept = mean(citric.acid)),col='red',size=0.5) +
  geom_vline(aes(xintercept = median(citric.acid)), col = 'grey', size=1)

p04 <- ggplot(red1, aes(x=residual.sugar)) +
  geom_histogram(color = 'black',fill = I('pink')) +
  xlab('Residual Sugar, g/dm^3') +
  geom_vline(aes(xintercept = mean(residual.sugar)),col='red',size=0.5) +
  geom_vline(aes(xintercept = median(residual.sugar)), col = 'grey', size=1)

p05 <- ggplot(red1, aes(x=chlorides)) +
  geom_histogram(color = 'black',fill = I('pink')) +
  xlab('Sodium Chloride, g/dm^3') +
  geom_vline(aes(xintercept = mean(chlorides)),col='red',size=0.5) +
  geom_vline(aes(xintercept = median(chlorides)), col = 'grey', size=1)

p06 <- ggplot(red1, aes(x=free.sulfur.dioxide)) +
  geom_histogram(color = 'black',fill = I('pink')) +
  xlab('Free Sulfur Dioxide, mg/dm^3') +
  geom_vline(aes(xintercept = mean(free.sulfur.dioxide)),col='red',size=0.5) +
  geom_vline(aes(xintercept = median(free.sulfur.dioxide)), col = 'grey', size=1)

p07 <- ggplot(red1, aes(x=total.sulfur.dioxide)) +
  geom_histogram(color = 'black',fill = I('pink')) +
  xlab('Total Sulfur Dioxide, mg/dm^3') +
  geom_vline(aes(xintercept = mean(total.sulfur.dioxide)),col='red',size=0.5) +
  geom_vline(aes(xintercept = median(total.sulfur.dioxide)), col = 'grey', size=1)

p08 <- ggplot(red1, aes(x=density)) +
  geom_histogram(color = 'black',fill = I('pink')) +
  xlab('Density, g/cm^3') +
```

```r
  geom_vline(aes(xintercept = mean(density)),col='red',size=0.5) +
  geom_vline(aes(xintercept = median(density)), col = 'grey', size=1)

p09 <- ggplot(red1, aes(x=pH)) +
  geom_histogram(color = 'black',fill = I('pink')) +
  xlab('pH') +
  geom_vline(aes(xintercept = mean(pH)),col='red',size=0.5) +
  geom_vline(aes(xintercept = median(pH)), col = 'grey', size=1)

p10 <- ggplot(red1, aes(x=sulphates)) +
  geom_histogram(color = 'black',fill = I('pink')) +
  xlab('Potassium Sulphate, g/dm3') +
  geom_vline(aes(xintercept = mean(sulphates)),col='red',size=0.5) +
  geom_vline(aes(xintercept = median(sulphates)), col = 'grey', size=1)

p11 <- ggplot(red1, aes(x=alcohol)) +
  geom_histogram(color = 'black',fill = I('pink')) +
  xlab('Alcohol, % by Volume') +
  geom_vline(aes(xintercept = mean(alcohol)),col='red',size=0.5) +
  geom_vline(aes(xintercept = median(alcohol)), col = 'grey', size=1)

p12 <- ggplot(red1, aes(x=quality)) +
  geom_histogram(binwidth=1.0, color = 'maroon',fill = I('violet')) +
  xlab('Quality, 0-10') +
  scale_x_continuous(limits = c(3,9), breaks = c(4,5,6,7,8)) +
  geom_vline(aes(xintercept = mean(quality)),col='red',size=0.5) +
  geom_vline(aes(xintercept = median(quality)), col = 'grey', size=1)

grid.arrange(p01, p02, p03, p04, p05, p06, p07, p08, p09, p10, p11, p12, ncol=3)
```



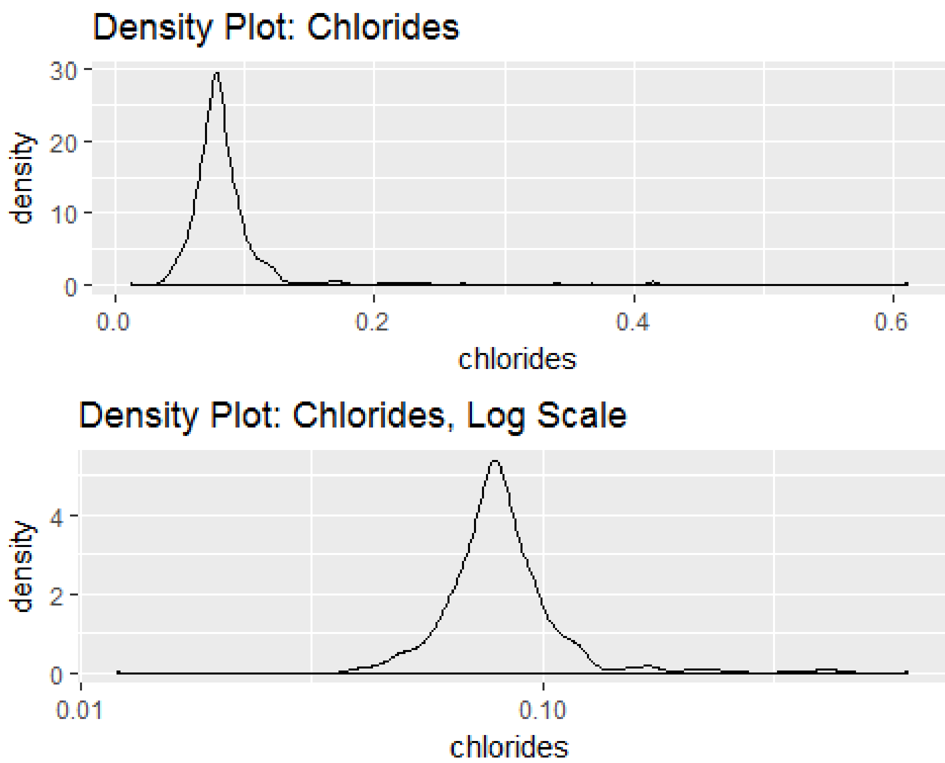
Non-normally distributed variables may pose problems in the regression analyses. If these variables seriously deviate from a normal distribution, researchers should transform these variables or categorize them accordingly. Plots below represent examples used to test for variable normality. I've used it for the three sets of non-normal distribution listed above. The plottings of all the questionable take up space. So, I am sharing some plot examples.

In the examples below, we see that both chlorides and total.sulfur.dioxide seem more normal after a natural log transformation. With a ggplot, I may adjust the binwidth to obtain a better distribution depiction. I have a personal preference for a natural log transformation due to its meaningful insights compared to other transformation - although later on, I tested for non-linear relationship with squared variables. In a linear regression analysis, for instance, a natural-log transformed independent variable's parameter estimate may be interpreted as "a one-percent change in the independent variable is related to a increase in the dependent variable".

Below are two examples on how I checked for variable's normality distribution. I'd change the variables accordingly.

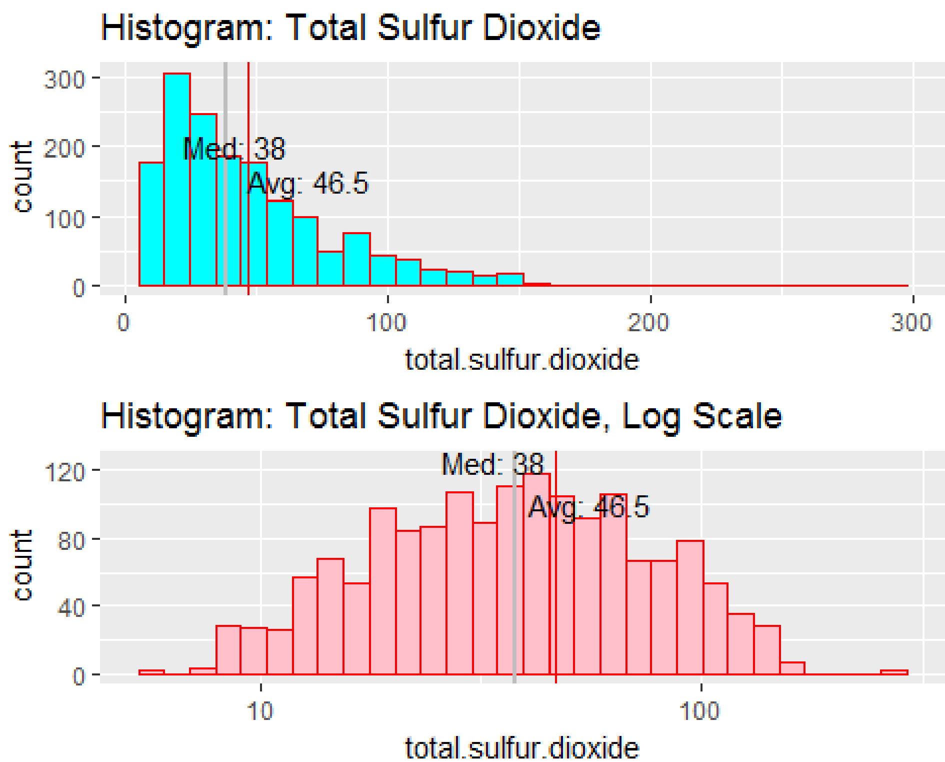
```
### Examples
# 1. Density Plot

d01 <- ggplot(red1,aes(x=chlorides)) + geom_density() +
  ggtitle('Density Plot: Chlorides')
```

```
d02 <- ggplot(red1,aes(x=chlorides)) + geom_density() + scale_x_log10() +
    ggtitle('Density Plot: Chlorides, Log Scale')

grid.arrange(d01,d02,ncol=1)
```




```
### Examples
# 2. Histogram
h01 <- ggplot(red1, aes(x=total.sulfur.dioxide)) +
  geom_histogram(color='red', fill='cyan') +
  ggtitle('Histogram: Total Sulfur Dioxide') +
  geom_vline(aes(xintercept = mean(total.sulfur.dioxide)),col='red',size=0.5) +
  geom_vline(aes(xintercept = median(total.sulfur.dioxide)), col = 'grey', size=1) +
  annotate("text", x = mean(red1$total.sulfur.dioxide) * 1.5, y = 150,
  label = paste0("Avg: ", round(mean(red1$total.sulfur.dioxide),1))) +
  annotate("text", x = median(red1$total.sulfur.dioxide) * 1.1, y = 200,
  label = paste0("Med: ", round(median(red1$total.sulfur.dioxide),1)))

h02 <- ggplot(red1, aes(x=total.sulfur.dioxide)) +
```

```
  geom_histogram(color='red', fill='pink') +
  scale_x_log10()+
  ggtitle('Histogram: Total Sulfur Dioxide, Log Scale') +
  geom_vline(aes(xintercept = mean(total.sulfur.dioxide)),col='red',size=0.5) +
  geom_vline(aes(xintercept = median(total.sulfur.dioxide)), col = 'grey', size=1) +
  annotate("text", x = mean(red1$total.sulfur.dioxide) * 1.2, y = 100,
  label = paste0("Avg: ", round(mean(red1$total.sulfur.dioxide),1))) +
  annotate("text", x = median(red1$total.sulfur.dioxide) * 0.9, y = 125,
  label = paste0("Med: ", round(median(red1$total.sulfur.dioxide),1)))

grid.arrange(h01, h02, ncol=1)
```



Variable Transformation

Besides a natural log transformation, I considered categorizing continuous variables and squared categorization. If my knowledge on these chemical properties were solid, I'd group them according to these meaningful categorization. I considered a binary, three-group, and four-group categorizations based on variable distribution and regression analysis results. The resulting categorization should not be too small percentage-wise that it may cause regression estimation issues.

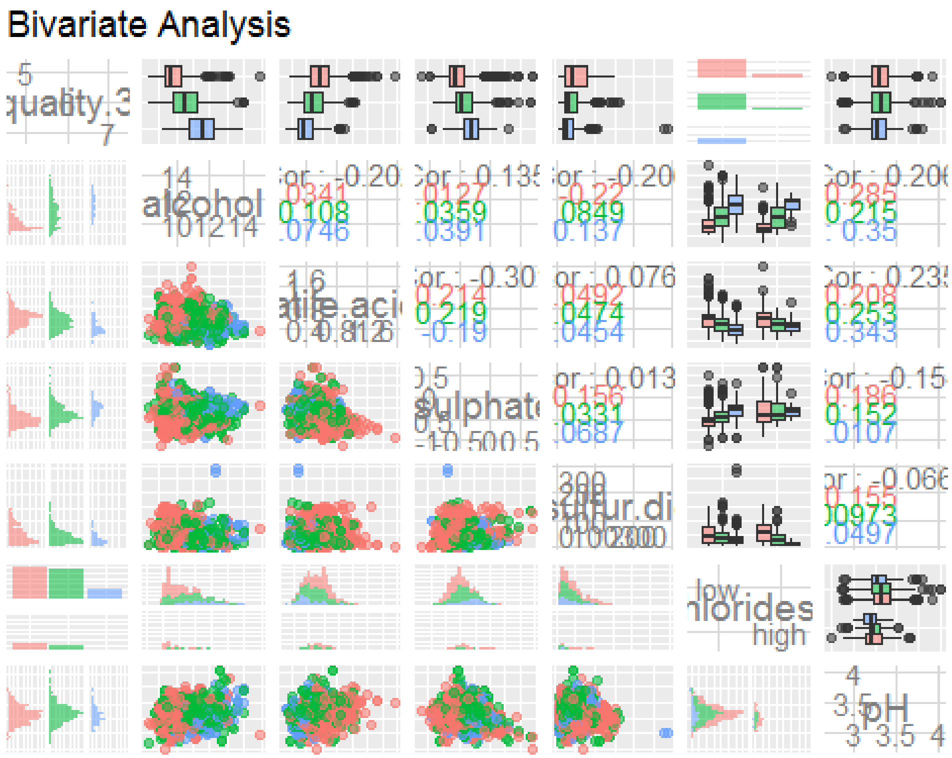
```r
### Categorization
# Quality, fixed.acidity, volatile.acidity, chlorides
red1$quality.f <- cut(red1$quality, breaks = c(0,4,5,6,10), labels=c("4","5","6","7"))
red1$quality.3 <- cut(red1$quality, breaks = c(0,5,6,10), labels=c("5","6","7"))

# Residual Sugar and Chlorides Transformations
red1$residual.sugar3 <- cut(red1$residual.sugar, breaks=c(-Inf, 2.2, 3, Inf), labels=c("low","middle","high"))
red1$ln.residual.sugar <- log(red1$residual.sugar)

red1$ln.chlorides <- log(red1$chlorides)
red1$chlorides4 <- cut(red1$chlorides, breaks=c(-Inf, 0.07,0.08, 0.1, Inf), labels=c("Q1","Q2","Q3","Q4"))
red1$chlorides2 <- cut(red1$chlorides, breaks=c(-Inf, 0.1, Inf), labels=c("low","high"))

### Other Transformations
red1$ln.fixed.acidity <- log(red1$fixed.acidity)
red1$ln.volatile.acidity <- log(red1$volatile.acidity)
red1$ln.sulphates <- log(red1$sulphates)
red1$ln.total.sulfur.dioxide <- log(red1$total.sulfur.dioxide)

# non-linear
red1$residual.sugar.sq <- (red1$residual.sugar)**2
red1$alcohol.sq <- (red1$alcohol)**2
```

Bivariate Analysis

The initial correlation matrix in this report actually serves an important bivariate analysis tool. For instance, I can immediate see chemical properties that influence the quality of red wine. In my multivariate analysis, I will immediately include alcohol (r=0.48) and volatile.acidity (-0.39) in the quality model from the get-go. Then, I need to consider total.sulfur.dioxide, and citric acid.

Since quality can be analyzed as a categorical variable, I also compared the quality and other variables treating quality as a categorical variable. Due to smaller proportions for qualities of '4 or less' and '8 or more', I grouped them together with '5' and '7' respectively - what I coined as 'quality.3'. Hopefully, a bivariate analysis between quality.3 and other variables will provide more hints of additional important variables.

```r
ggpairs(data=red1,
        columns=c("quality.3","alcohol","volatile.acidity","ln.sulphates","total.sulfur.dioxide","chlorides2","pH"),
        title="Bivariate Analysis",
        aes(color = quality.3, alpha = .9),
        axisLabels = 'internal')
```



After a few iterations with the stratified boxplot and distribution -as depicted in the bivariate analysis above-, it seems that ln.sulphates definitely needs to be considered based on the boxplot. Other variables that I will look into include total.sulfur.dioxide, chlorides2, and pH. I may consider more variables but I have to keep in mind that some of these variables are highly correlated and maybe related to other variables. For instance, we expect free.sulfur.dioxide to be a component of total.sulfur.dioxide - they actually have a correlation of 66%. So, I would be wary of including both variables in the model due to multicollinearity, which should be reflected in the regression statistics when comparing models with one of the two variables vs. both variables.

Other relationships may not be obvious such as alcohol and pH. Alcohol tends to be neutral, thus, having a pH value around 7. The data depicts a low 21% correlation between these two variable. Two important correlations here are:

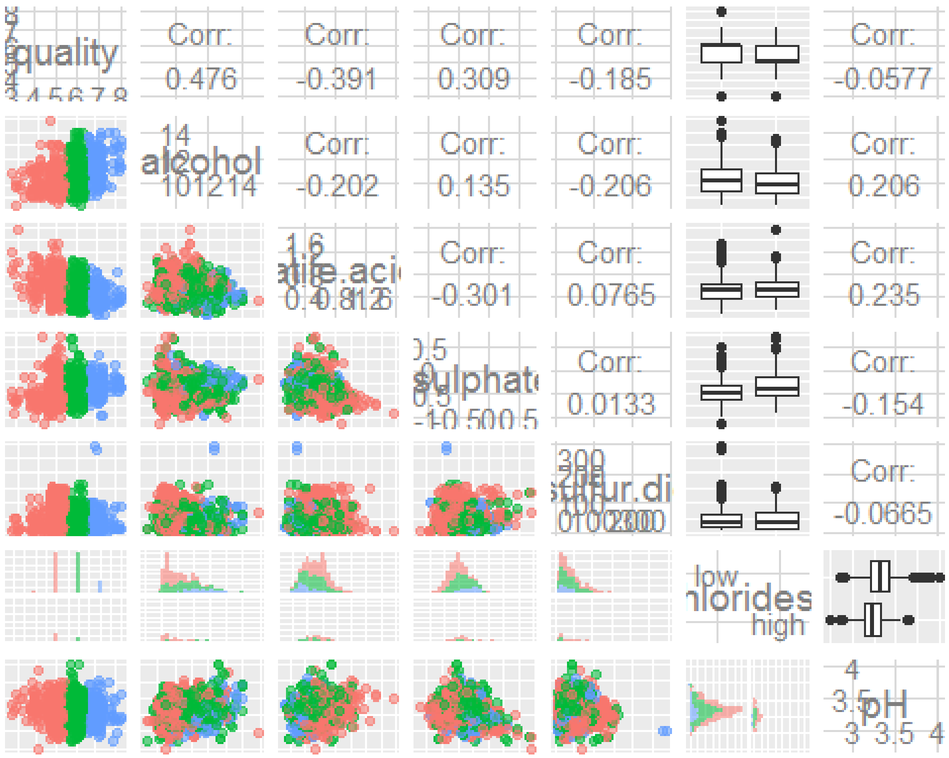
- r(pH, fixed.acidity)= - 68% but
- r(pH, volatile.acidity)= 23%.

Since acid have lower pH value, I expected a strong negative correlation of acidic substance with the pH value, as shown by that of fixed.acidity. However, volatile.acidity and pH shows a positive but weak correlation.

Multivariate Statistics

I am a big fan of using the matrix to look into overall trends and variable association. So, for the multivariate statistics, I am doing the same. I employed the same matrix plot but this time, I added the three-level categorical variable as the color in the correlation matrix.

```
ggpairs(red1, columns=c("quality","alcohol","volatile.acidity","ln.sulphates","total.sulfur.dioxide","chlorides2","pH")
,
        lower = list(mapping = aes(color = quality.3, alpha = .6),continuous=wrap("points", position="jitter")),
        axisLabels = 'internal')
```

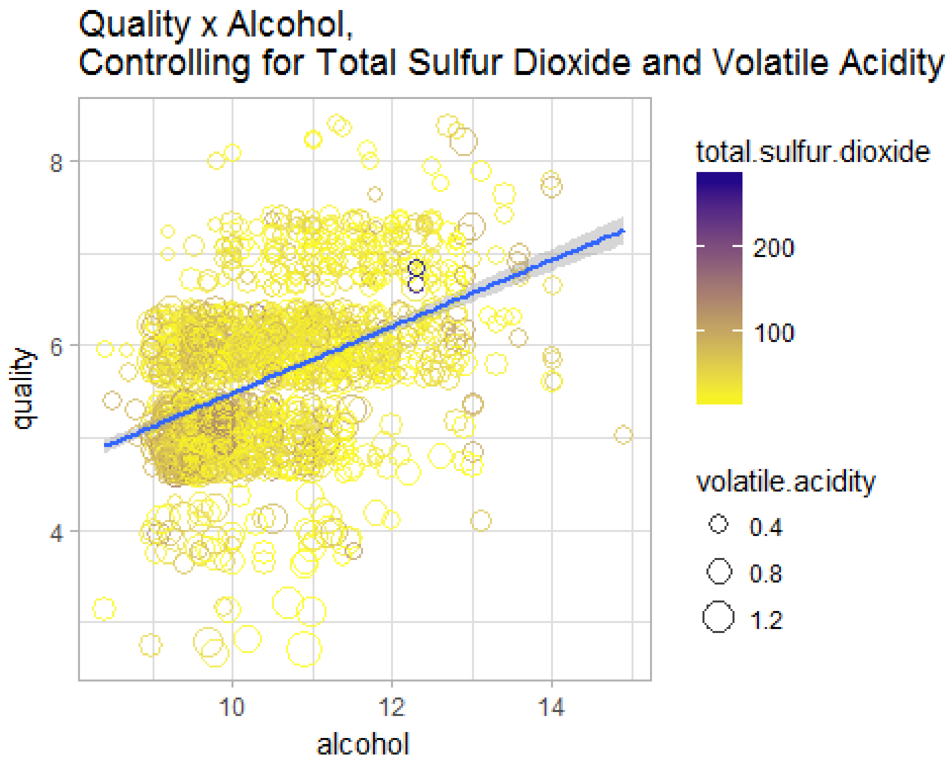


My takeaways on what variables to test for the regression analyses were very similar to the ones obtained in the univariate and bivariate sections. Based on the chart below:

- alcohol and quality seems to be positively related based on the trend line;
- volatile.acidity and quality seems to be negatively related based on lower quality tend to have larger circle points; and
- total.sulfur.dioxide and quality may not be related because the the darker circles are scattered along different quality values.

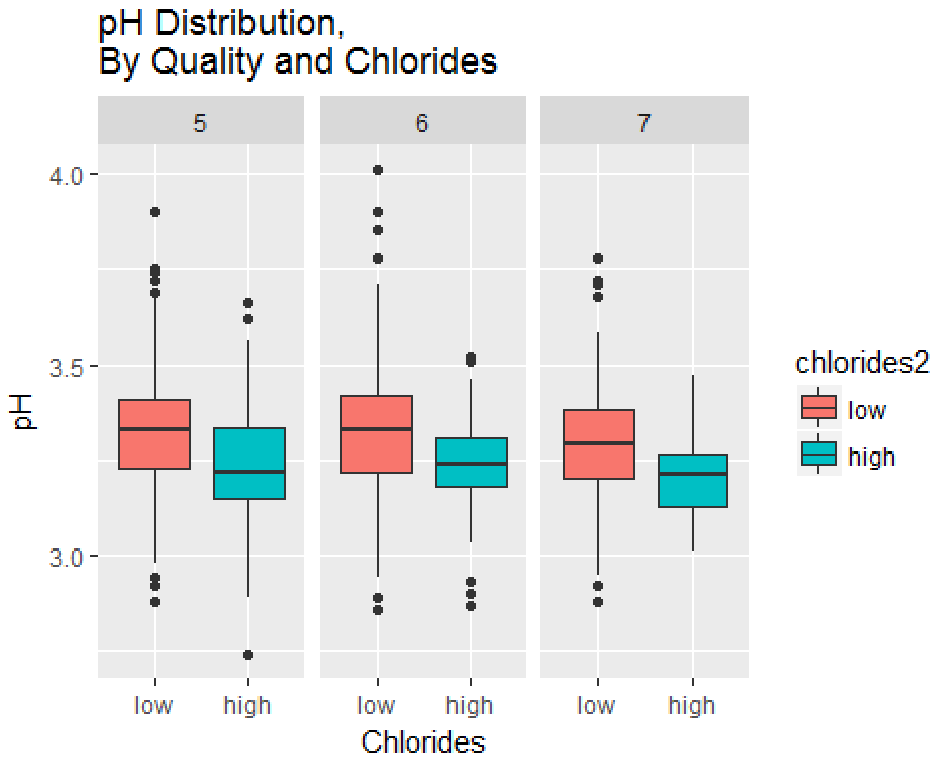
```
ggplot(aes(y = quality, x = alcohol, color = total.sulfur.dioxide), data = red1) +
  scale_color_gradient(low="yellow", high="darkblue") +
```

```
  geom_point(aes(size=volatile.acidity), alpha = 0.75, position = 'jitter', shape=1) +
  geom_smooth(method = lm) +
  ggtitle('Quality x Alcohol,\nControlling for Total Sulfur Dioxide and Volatile Acidity') +
  theme_light()
```



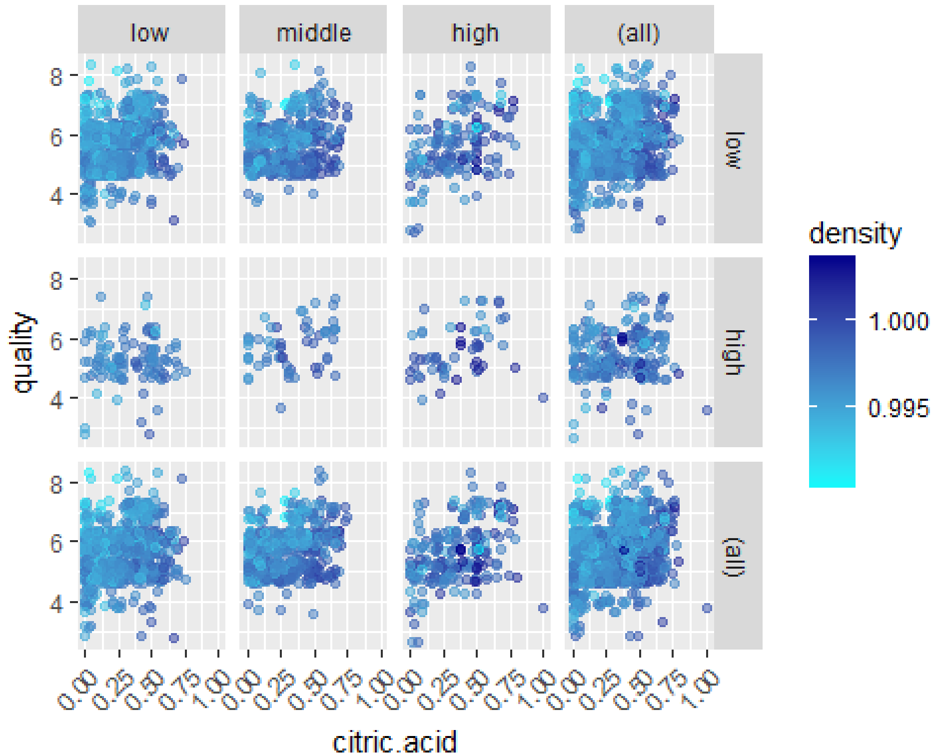
The section on quality vs. chlorides from the multivariate matrix plot above was too small. So, I created a stratified box plot below, hoping to gain more insights on the relationship of pH and chlorides to red wine quality. It's still hard to gauge if there's such a relationship. High chloride wines tend to have slightly lower quality.

```
ggplot(red1, aes(x=chlorides2, y=pH, fill=chlorides2)) +
  geom_boxplot() +
  labs(title="pH Distribution, \nBy Quality and Chlorides", x='Chlorides', y='pH') +
  facet_wrap(~quality.3)
```

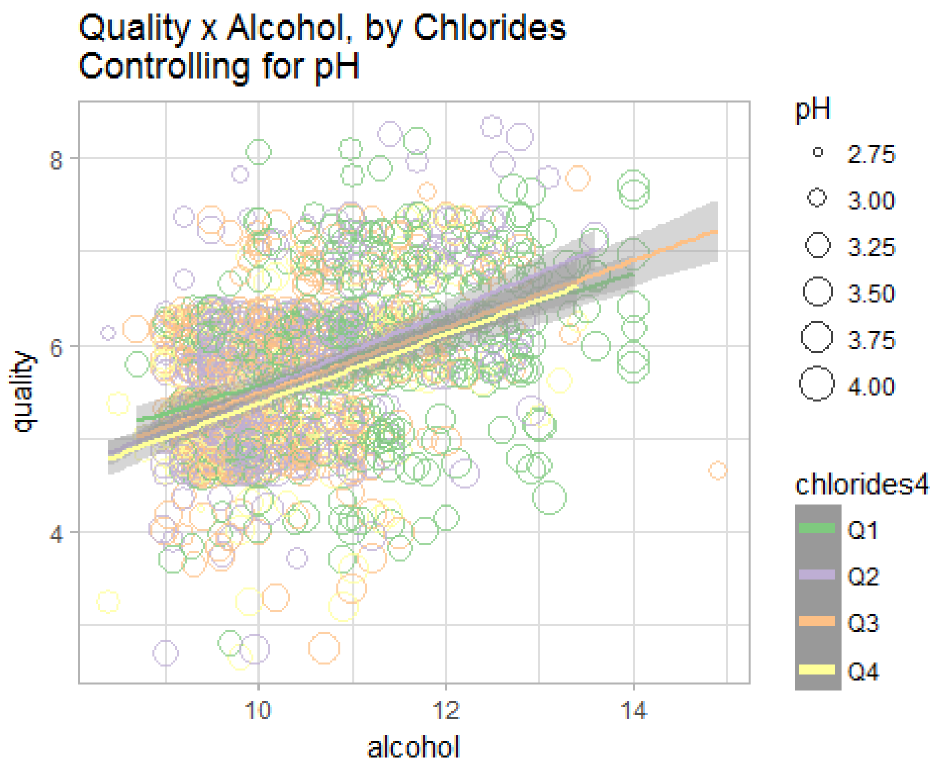


I also looked into how citric.acid and density may be related to quality. Based on the chart below, the relationships to quality are spurious. There are no noticeable trends even when the analyses were stratified by citric.acid and color-coded by density. The only thing that popped up was a positive quality-citric.acid relationship for high citric.acid group.

```
ggplot(red1, aes(citric.acid, quality, colour = density)) +
  scale_color_gradient(low="cyan", high="darkblue") +
  geom_jitter(alpha = .5) +
  facet_grid(chlorides2 ~ residual.sugar3, margins = TRUE) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1))
```

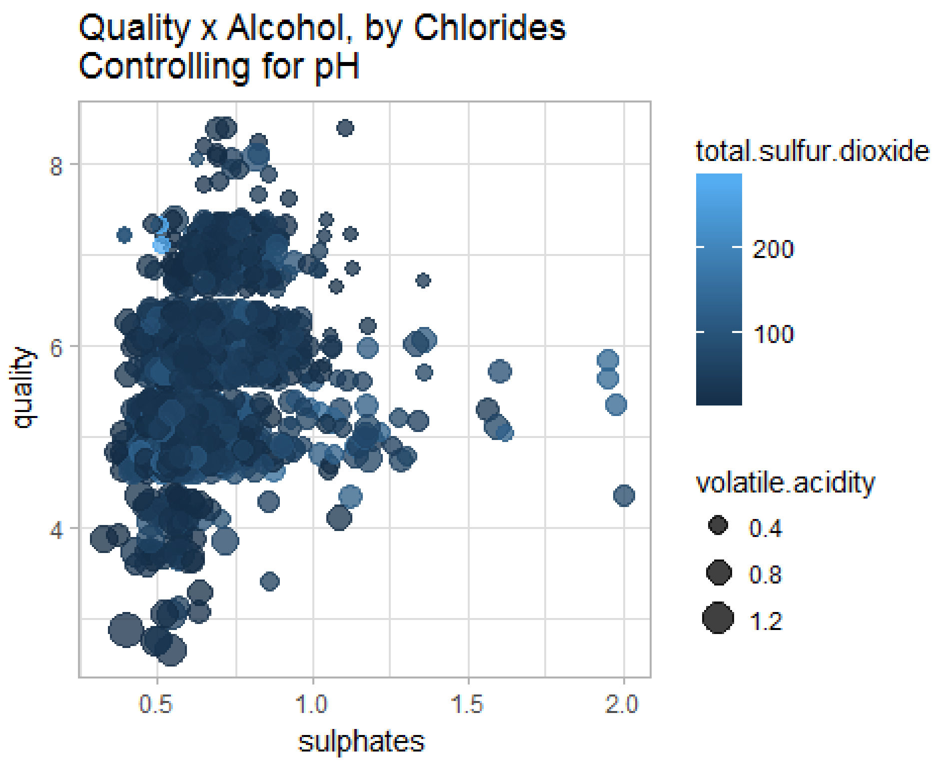



```
ggplot(aes(y = quality, x = alcohol, color = chlorides4), data = red1) +
  geom_point(aes(size=pH), alpha = 0.75, position = 'jitter', shape=1) +
  scale_color_brewer(type = 'qual',
                     guide = guide_legend(title = 'chlorides4',  reverse = F,
                                          override.aes = list(alpha = 1, size = 2))) +
  geom_smooth(aes(group=chlorides4), method = lm) +
  ggtitle('Quality x Alcohol, by Chlorides\nControlling for pH') +
  theme_light()
```



Using the template below, I look into other variables but none of them seem to be strong.

```
ggplot(aes(y = quality, x = sulphates, color = total.sulfur.dioxide), data = red1) +
  geom_point(aes(size=volatile.acidity), alpha = 0.75, position = 'jitter') +
  ggtitle('Quality x Alcohol, by Chlorides\nControlling for pH') +
  theme_light()
```



Regression Analysis

In the bivariate and multivariate analyses, I've identified variables that are likely to be influencing red wine quality. As previously mentioned, based on correlation results, I expect alcohol (r=0.48) and volatile.acidity (-0.39) to important variables.

Linear Regression Analysis

I ended up with a linear regression with six significant regressors: alcohol, volatile.acidity, ln.sulphates, total.sulfur.dioxide, and chlorides2 + pH. The overall model was highly significant based on the F test. The R-squared for this final model was 36.7%.

```
m1 <- lm(quality ~ alcohol, data = red1)
m2 <- update(m1, ~ . + volatile.acidity)
m3 <- update(m2, ~ . + ln.sulphates)
m4 <- update(m3, ~ . + total.sulfur.dioxide)
m5 <- update(m4, ~ . + chlorides2)
m6 <- update(m5, ~ . + pH)
mtable(m1, m2, m3, m4, m5, m6)
```

```
##
## Calls:
## m1: lm(formula = quality ~ alcohol, data = red1)
## m2: lm(formula = quality ~ alcohol + volatile.acidity, data = red1)
## m3: lm(formula = quality ~ alcohol + volatile.acidity + ln.sulphates,
##     data = red1)
## m4: lm(formula = quality ~ alcohol + volatile.acidity + ln.sulphates +
##     total.sulfur.dioxide, data = red1)
## m5: lm(formula = quality ~ alcohol + volatile.acidity + ln.sulphates +
##     total.sulfur.dioxide + chlorides2, data = red1)
## m6: lm(formula = quality ~ alcohol + volatile.acidity + ln.sulphates +
##     total.sulfur.dioxide + chlorides2 + pH, data = red1)
##
## ==============================================================================================================
##                                 m1            m2            m3            m4            m5            m6
## --------------------------------------------------------------------------------------------------------------
##   (Intercept)                  1.875***      3.095***      3.369***      3.612***      3.722***      4.962***
##                               (0.175)       (0.184)       (0.184)       (0.191)       (0.192)       (0.373)
##   alcohol                      0.361***      0.314***      0.303***      0.290***      0.283***      0.299***
##                               (0.017)       (0.016)       (0.016)       (0.016)       (0.016)       (0.016)
##   volatile.acidity                          -1.384***     -1.156***     -1.134***     -1.082***     -0.975***
##                                             (0.095)       (0.097)       (0.097)       (0.097)       (0.101)
##   ln.sulphates                                             0.641***      0.660***      0.747***      0.729***
##                                                           (0.077)       (0.077)       (0.079)       (0.079)
##   total.sulfur.dioxide                                                  -0.002***     -0.002***     -0.002***
##                                                                         (0.001)       (0.001)       (0.001)
##   chlorides2: high/low                                                               -0.208***     -0.245***
##                                                                                      (0.049)       (0.049)
##   pH                                                                                               -0.442***
##                                                                                                    (0.114)
## --------------------------------------------------------------------------------------------------------------
##   R-squared                    0.227         0.317         0.345         0.353         0.361         0.366
##   adj. R-squared               0.226         0.316         0.344         0.352         0.359         0.364
##   sigma                        0.710         0.668         0.654         0.650         0.647         0.644
##   F                          468.267       370.379       280.646       217.574       179.608       153.472
##   p                            0.000         0.000         0.000         0.000         0.000         0.000
##   Log-likelihood           -1721.057     -1621.814     -1587.752     -1578.324     -1569.192     -1561.728
##   Deviance                   805.870       711.796       682.108       674.111       666.456       660.263
##   AIC                       3448.114      3251.628      3185.503      3168.648      3152.385      3139.456
##   BIC                       3464.245      3273.136      3212.389      3200.911      3190.025      3182.473
```

```
## N 1599 1599 1599 1599 1599 1599
## ==================================================================================================
```

Logistic Regression Analysis

When the regressors were ran in an ordinal logistic regression modal, all the regressors were also significant. Highly significant in the linear regression model, pH is almost significant at 1 percent significant level.

```
ol <- clm(quality.3 ~ alcohol + volatile.acidity + ln.sulphates + total.sulfur.dioxide + chlorides2 + pH, data=red1)
summary(ol)

## formula:
## quality.3 ~ alcohol + volatile.acidity + ln.sulphates + total.sulfur.dioxide + chlorides2 + pH
## data:    red1
##
##  link  threshold nobs logLik   AIC     niter max.grad cond.H
##  logit flexible  1599 -1219.49 2454.98 6(0)  9.85e-12 3.0e+06
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## alcohol               0.946801   0.058960  16.058  < 2e-16 ***
## volatile.acidity     -2.722549   0.352178  -7.731 1.07e-14 ***
## ln.sulphates          2.521606   0.265606   9.494  < 2e-16 ***
## total.sulfur.dioxide -0.012909   0.001904  -6.781 1.19e-11 ***
## chlorides2high       -0.821801   0.172255  -4.771 1.83e-06 ***
## pH                   -0.972504   0.379180  -2.565   0.0103 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##     Estimate Std. Error z value
## 5|6    3.098      1.221   2.537
## 6|7    6.007      1.230   4.883

cat("\n\nOdds Ratio \n")

##
##
## Odds Ratio

exp(coef(ol))
```

```
##                   5|6                   6|7               alcohol
##           22.15944000          406.43443193            2.57745243
##      volatile.acidity          ln.sulphates total.sulfur.dioxide
##            0.06570705           12.44857820            0.98717358
##        chlorides2high                    pH
##            0.43963900            0.37813505
```

```
cat("\n\nProportional Odds Test \n")
```

```
##
##
## Proportional Odds Test
```

```
nominal_test(ol)
```

```
## Tests of nominal effects
##
## formula: quality.3 ~ alcohol + volatile.acidity + ln.sulphates + total.sulfur.dioxide + chlorides2 + pH
##                      Df  logLik    AIC     LRT Pr(>Chi)
## <none>                  -1219.5 2455.0
## alcohol               1 -1219.0 2456.1 0.90611  0.34115
## volatile.acidity      1 -1218.5 2454.9 2.04794  0.15241
## ln.sulphates          1 -1218.0 2454.0 3.01188  0.08266 .
## total.sulfur.dioxide  1 -1219.5 2456.9 0.04466  0.83263
## chlorides2            1 -1219.3 2456.5 0.43857  0.50781
## pH
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Proportional odds ratio test insignificant results indicate that we can assume that the odds ratio between the three levels of quality are proportionate. This means that the current ordinal logistic regression can be used that I don't have to resort to a multinomial logistic regression instead.

Final Plots and Summary

Data Outlook

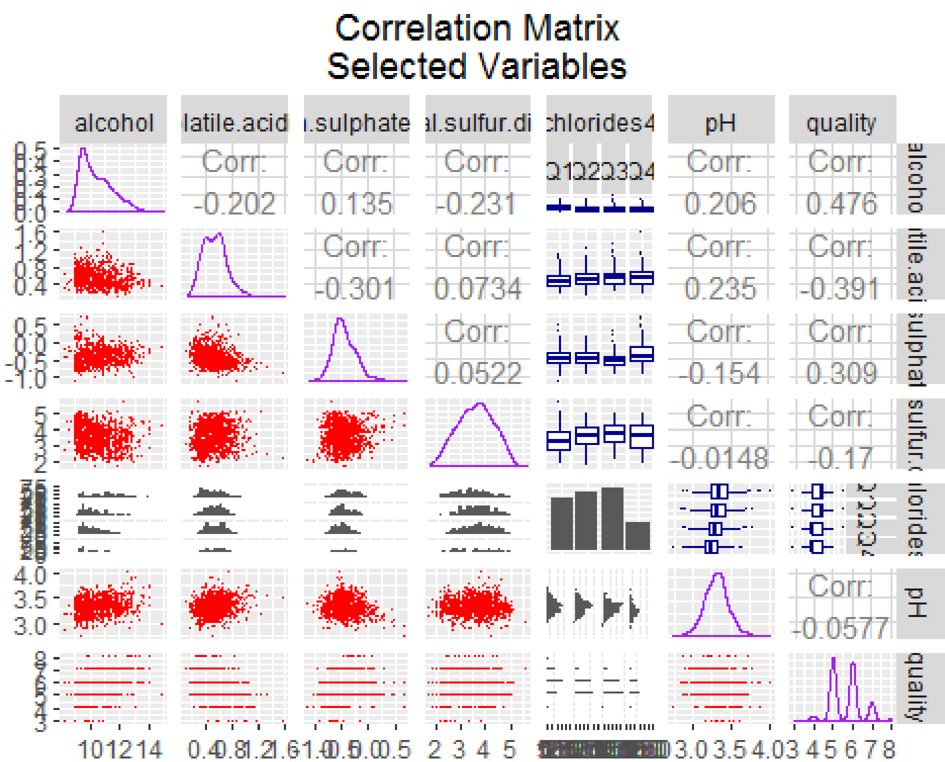
Given that the dataset does not contain too many variables, I started by looking into a scatter plot matrix. Initially, I look into the diagonally-placed distribution plots. Then, I revisited this plot to get a quick picture on relations between variables. From the scatterplot matrix below, I can quickly see the followings:

- residual.sugar, chlorides, and total.sulfur.dioxide variables need data transformation to obtain a more normal distribution

- quality has a weird up and down distribution. Based on the data description and bottom row scatter plots, this is probably to non-decimal round number quality score for red wine quality. Regarding quality, one can easily see that bulk of response for quality are in the two middle values of quality. This is especially disconcerting because lack of spread for a dependent variable may not produce a highly reliable linear regression estimation.
- the column for the dependent variable (quality) provides extra interesting insights.

One can see that alcohol is strongly correlated to quality with fixed.acidity, volatile.acidity and sulphates having weak correlations to quality. These correlations may in the end tell which chemical properties are important. Though, it's important note that these are untransformed variables and the strength of variables may be strengthen or dampen when other variables are considered in a model.

```
ggpairs(red1[c(12,3,23,24,19,10,13)],
  diag = list(continuous = wrap("densityDiag", color = "purple", alpha = 0.5)),
  lower = list(continuous = wrap("points", color = "red", shape = I('.'))),
  upper = list(combo = wrap("box", color = "darkblue", outlier.shape = I('.')))) +
  ggtitle('Correlation Matrix\nSelected Variables') +
  theme(plot.title = element_text(hjust = 0.5))
```

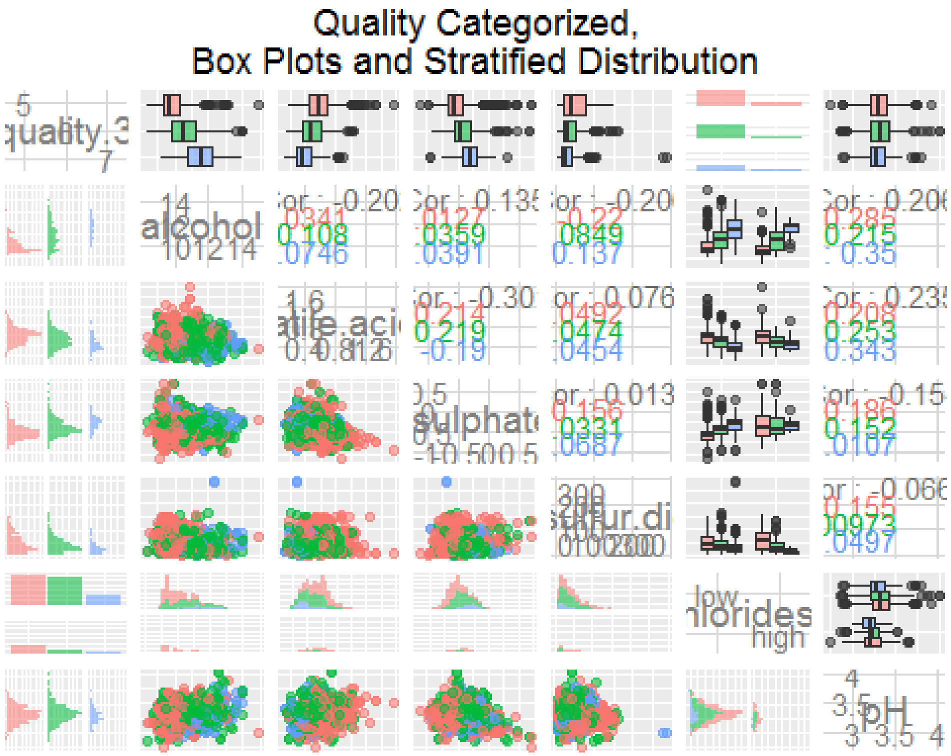


This correlation matrix in this report actually serves as an important bivariate analysis tool. For instance, I can immediate see chemical properties that influence the quality of red wine. In my multivariate analysis, I will immediately include alcohol (r=0.48) and volatile.acidity (-0.39) in the quality model from the get-go. Then, I need to consider total.sulfur.dioxide, and citric acid, including their transformed and categorized variables.

Quality as a Categorical Variable

Since quality can be analyzed as a categorical variable, I also compared the quality and other variables treating quality as a categorical variable. Due to smaller proportions for qualities of '4 or less' and '8 or more', I grouped them together with '5' and '7' respectively - what I coined as 'quality.3'. Hopefully, a bivariate analysis between quality.3 and other variables will provide more hints of additional important variables.

```
ggpairs(data=red1,
        columns=c("quality.3","alcohol","volatile.acidity","ln.sulphates","total.sulfur.dioxide","chlorides2","pH"),
        title="Bivariate Analysis",
        aes(color = quality.3, alpha = .9),
        axisLabels = 'internal') +
        ggtitle('Quality Categorized,\nBox Plots and Stratified Distribution') +
        theme(plot.title = element_text(hjust = 0.5))
```



After a few iterations with the stratified boxplot and distribution -as depicted in the bivariate analysis above-, it seems that ln.sulphates definitely needs to be considered based on the boxplot. Other variables that I will look into include total.sulfur.dioxide, chlorides2, and pH. I may consider more variables but I have to keep in mind that some of these variables are highly correlated and maybe related to other variables. For instance, we expect free.sulfur.dioxide to be a component of total.sulfur.dioxide - they actually have a correlation of 66%. So, I would be wary of including both variables in the model due to multicollinearity, which should be reflected in the regression statistics when comparing models with one of the two variables vs. both variables.

Other relationships may not be obvious such as alcohol and pH. Alcohol tends to be neutral, thus, having a pH value around 7. The data depicts a low 21% correlation between these two variable. Two important correlations here are:

- r(pH, fixed.acidity)= - 68% but
- r(pH, volatile.acidity)= 23%.

Since acid have lower pH value, I expected a strong negative correlation of acidic substance with the pH value, as shown by that of fixed.acidity. However, volatile.acidity and pH shows a positive but weak correlation. I would be concerned to introduce pH intovariable with fixed.acidity, but not one with volatile.acidity.

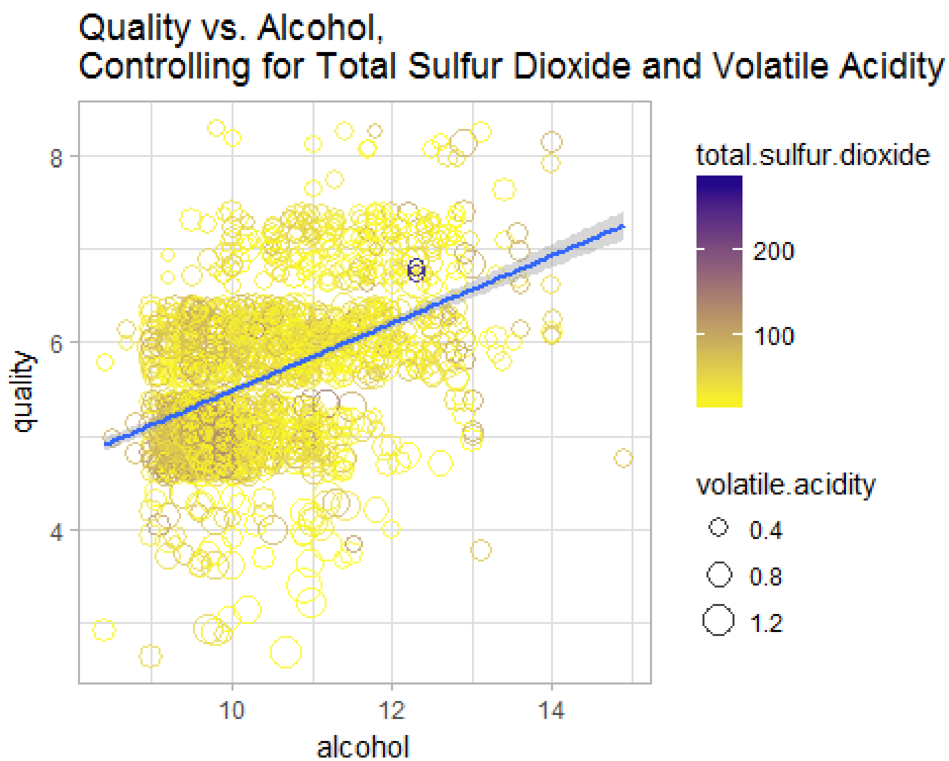
Multivariate Analysis Depiction

Two plot matrix above helped me to narrow down the variables that I should look into. I looked into variables with high correlation (ablsolute value) with quality. These variables are looked into as they are, transformed variables, and categorized variables. I am aware that sometimes a variable may not be highly correlated but may become significant when introduced into a model along with other variables.

In any case, a useful plot to depict various variables to quality can be represented in a chart like the one below. In this chart below: * alcohol and quality seems to be positively related based on the trend line; * volatile.acidity and quality seems to be negatively related based on lower quality tend to have larger circle points; and * total.sulfur.dioxide and quality may not be related because the the darker circles are scattered along different quality values.

I also tested several transformed variables in the model. One of my hypotheses is that properties that are very strong or even very week may not produce high quality red wine. Thus, I expected an inverse-U relationship for acid-like variables with quality.

```
ggplot(aes(y = quality, x = alcohol, color = total.sulfur.dioxide), data = red1) +
  scale_color_gradient(low="yellow", high="darkblue") +
  geom_point(aes(size=volatile.acidity), alpha = 0.75, position = 'jitter', shape=1) +
  geom_smooth(method = lm) +
  ggtitle('Quality vs. Alcohol,\nControlling for Total Sulfur Dioxide and Volatile Acidity') +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme_light()
```



Conclusion from Regression Analyses

Both Linear and Logistic Regression analyses provided similar results. Variables that were positively related to quality were alcohol and ln.sulphates. Volatile.acidity, total.sulfur.dioxide, chlorides, and pH were negatively associated with quality.

Linear regression results indicates, given the other variables are held constant in the model:

- a one-percent increase in alcohol by volume is related to an **increase** in red wine quality by as score of 0.3;
- a one-percent increase in sulphates (potassium sulphate - g / dm3) is associated to an **increase** in red wine quality by as score of 0.73;
- a one-gram/dm3 increase in volatile acidity (acetic acid) is related to a **decrease** in red wine quality by a score of almost 1;
- a one-mg/dm3 increase in total.sulfur.dioxide is related to a **decrease** in red wine quality by a score of 0.002;
- red wines with high chloride tend to have a quarter **less** quality score than those with low chloride; and
- a one-unit increase in pH score is related to a **decrease** in red wine quality by almost a half score.

Logistic regression results indicates, given the other variables are held constant in the model:

- for a one percent **increase** in alcohol by volume, the odds of higher quality versus next highest quality categories are 2.6 times greater,

- for a 10% g/dm3 **increase** in sulphates (potassium sulphate), the odds of higher quality versus next highest quality categories are 3.3 times greater (Note: 1.1^(12.44)=3.3; See: "Interpretation of log transformed predictors in logistic regression)",
- for a 0.1 gram/dm3 **decrease** in volatile acidity (acetic acid), the odds of higher quality versus next highest quality categories are 1.5 times greater,
- for a one-mg/dm3 **decrease** in total.sulfur.dioxide, the odds of higher quality versus next highest quality categories are 0.01 times greater,
- red wines with high chloride tend to have a quarter quality score **less** than those with low chloride,
- for a one-unit **decrease** in pH score, the odds of higher quality versus next highest quality categories are 2.6 times greater.

The uni-, bi-, and multivariate analyses really helped to focus on the important variables, best transformation, and appropriate transformation. I'm aware that there are statistics for many of these assumption tests such as test for normality. However, visual representations really helped to see which transformations and categorization cuts points would be the better ones.

Reflections

Alcohol was the **best predictor** of quality, followed by acidity. The results truly make sense because alcohol is what makes wine an alcoholic beverage. As for acidity, I suppose people are attracted to the soda-like strength in a wine.

I was really hoping that the **squared variables** would be significant. To me, some chemical properties maybe associated to quality but after a point, the chemical properties may be too strong that the wine would be rated as low quality ones. Based on this results, one cannot recommend that we max out the chemical properties positively related to quality while minimize or eliminate those which are negatively related to quality. Since the red wines tested here tend to have average scores among expert, implications based on this study should **only be generalized to average red wine**.

Although squared variables were not significant, the hypothesis on **extreme chemical properties being less desired** is supported in chlorides findings. Chlorides actually had many small positive outliers. I found that extremely high chlorides tend to have lower quality compared to the first three quartiles. The variable was not significant when tested as a continuous and log-transformed variables. When included as quartiles, only the largest quartile was significant.

The validity of the study and findings may have been affected by the **data quality**. The methodology here plays a big part because I know that at least three testers were involved. If they tested several wines in a subsequent manner, their taste buds may be affected. And if they tested way too many wines at once, their judgments will certainly be impaired, resulting in invalid score.

We also do not have data on **raters** and if the **testing** were done in one session. Such data would allow researchers to control for the wine tester and testing session.

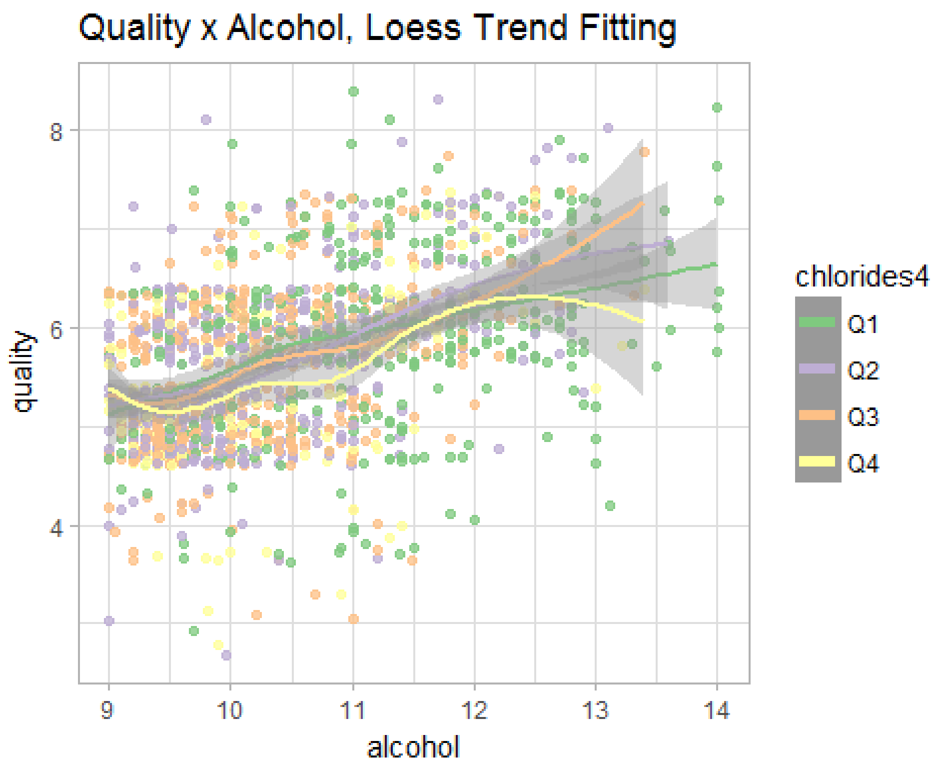
Most of the red wines were given average scores or **5, 6, and 7**. We could have benefited from testing **really bad** and **expensive wine** or wine with extreme chemical properties for more variance in the score. Perhaps, a dataset with more variance can better explain the variance in quality and also capture a U-inverse relationship. In addition, a better quality scoring could incorporate various taste elements such as the five **characteristics of wine** such as sweetness, acidicity, tannin, fruit, and body.

Also stated in the data note, the "**median** of at least 3 evaluations made by wine experts" were taken. Taking the median score may be problematic due to the tendency to arrive at scores closer to "5". If I had each evaluator's valuation for each wine, I may be able to remove scores with high variance based on evaluation validity. Since we are generalizing the results to the general population, I am not convinced that sampling by experts would necessarily represent the larger population taste and preference.

As laid out in the simple data note, "there is **no data** about grape types, wine brand, wine selling price". Grape types, wine brand, and wine selling price are arguably important factors in determining wine quality. One may argue that selling price may be a better reflection of wine quality although each wine's supply factor may serve as a counter-argument.

A nice segway to my last course of this Nanodegree - the multinomial analyses in this study focused on regression ones when there are other Machine Learning tools that may better capture chemical properties that influence wine quality.

```
ggplot(aes(y = quality, x = alcohol, color = chlorides4), data = subset(red1,alcohol>=9 & alcohol<=14)) +
  geom_point(alpha = 0.75, position = 'jitter') +
  scale_color_brewer(type = 'qual',
                     guide = guide_legend(title = 'chlorides4',  reverse = F,
                                          override.aes = list(alpha = 1, size = 2))) +
  geom_smooth(aes(group=chlorides4), method = loess) +
  ggtitle('Quality x Alcohol, Loess Trend Fitting') +
  theme_light()
```



From the chart above, one can see that quality may be affected by these variables in unique, non-systematic way which may require some algorithm. These imperfect association to quality were also depicted throughout this project. I look forward to learning various machine learning method in the Intro to Machine Learning course.

Resources

Changing Title in Multiplot ggplot2 Using grid.arrange: https://stackoverflow.com/questions/14726078/changing-title-in-multiplot-ggplot2-using-grid-arrange

Data Transformation with dplyr: CHEAT SHEET: https://github.com/rstudio/cheatsheets/raw/master/data-transformation.pdf

Draw a Trend Line Using ggplot: https://stackoverflow.com/questions/38412817/draw-a-trend-line-using-ggplot

ggplot2 Quick Reference: colour (and fill): http://sape.inf.usi.ch/quick-reference/ggplot2/colour

How Basic Wine Characteristics Help You Find Favorites: http://winefolly.com/review/wine-characteristics/

How to add mean, and mode to ggplot histogram?: https://stackoverflow.com/questions/47000494/how-to-add-mean-and-mode-to-ggplot-histogram

Interpretation of Log Transformed Predictors in Logistic Regression: https://stats.stackexchange.com/questions/8318/interpretation-of-log-transformed-predictors-in-logistic-regression

Ordinal Logistic Regression | R Data Analysis Examples: https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/

R Markdown Cheat Sheet: https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf

RDocumentation: nominal_test: https://www.rdocumentation.org/packages/ordinal/versions/2015.6-28/topics/nominal_test

Recode Data in R: http://rprogramming.net/recode-data-in-r/

Wine Quality Methodology: https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityInfo.txt