Exploring Red Wine Quality

Shane Kao

Wednesday, February 18, 2015

Goal

Which chemical properties influence the quality of red wines?

Data Overview

This tidy data set contains 1,599 red wines with 11 variables on the chemical properties of the wine. At least 3 wine experts rated the quality of each wine, providing a rating between 0 (very bad) and 10 (very excellent).

```
setwd("C:/Users/asus/Downloads")
data=read.csv("wineQualityReds.csv",stringsAsFactors=FALSE)
str(data)
```

```
## '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(data)
```

```
##        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
```

- The feature `X` is row index of data, it may provide no further information.
- The feature `quality` is an ordered, categorical, discrete variable.

```
data$quality<-as.ordered(data$quality)
```

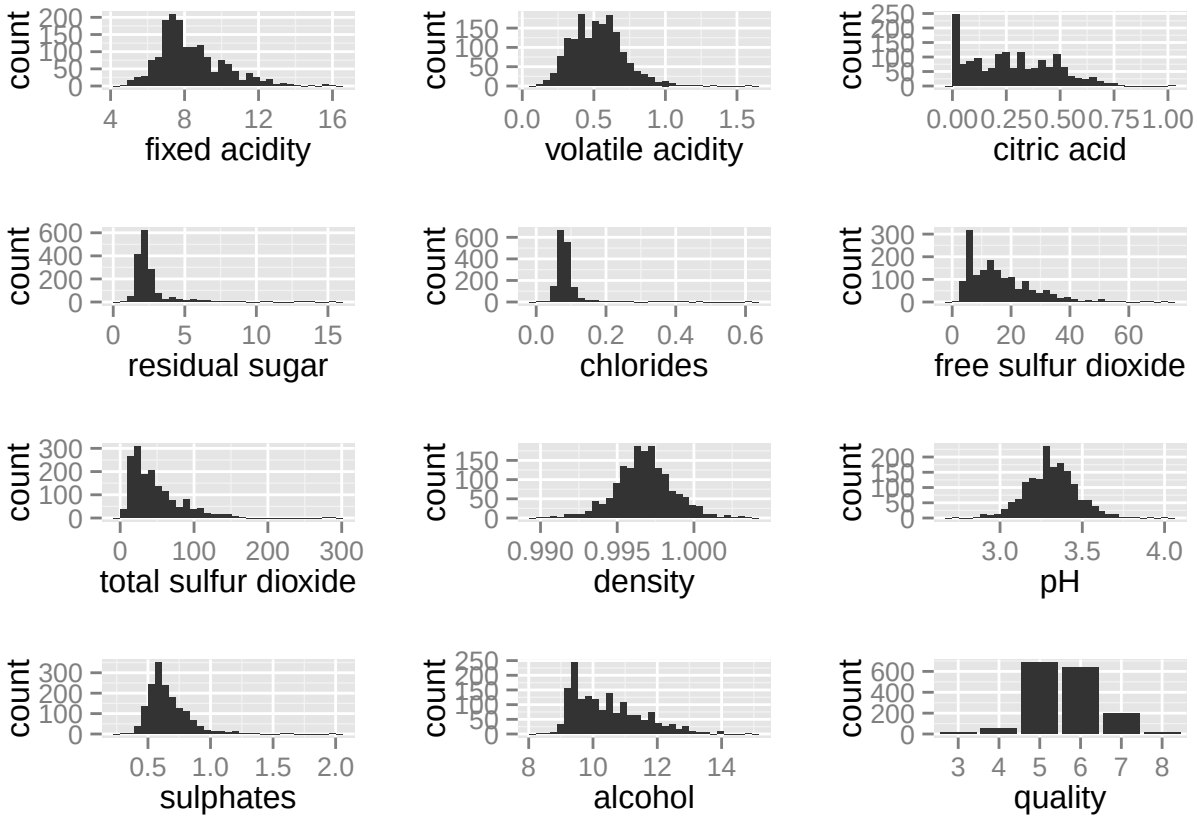
- From the variable descriptions, {`fixed.acidity`,`volatile.acidity`,`citric.acid`} and {`free.sulfur.dioxide`,`total.` may strongly correlated.

Univariate Plots Section

Univariate Analysis

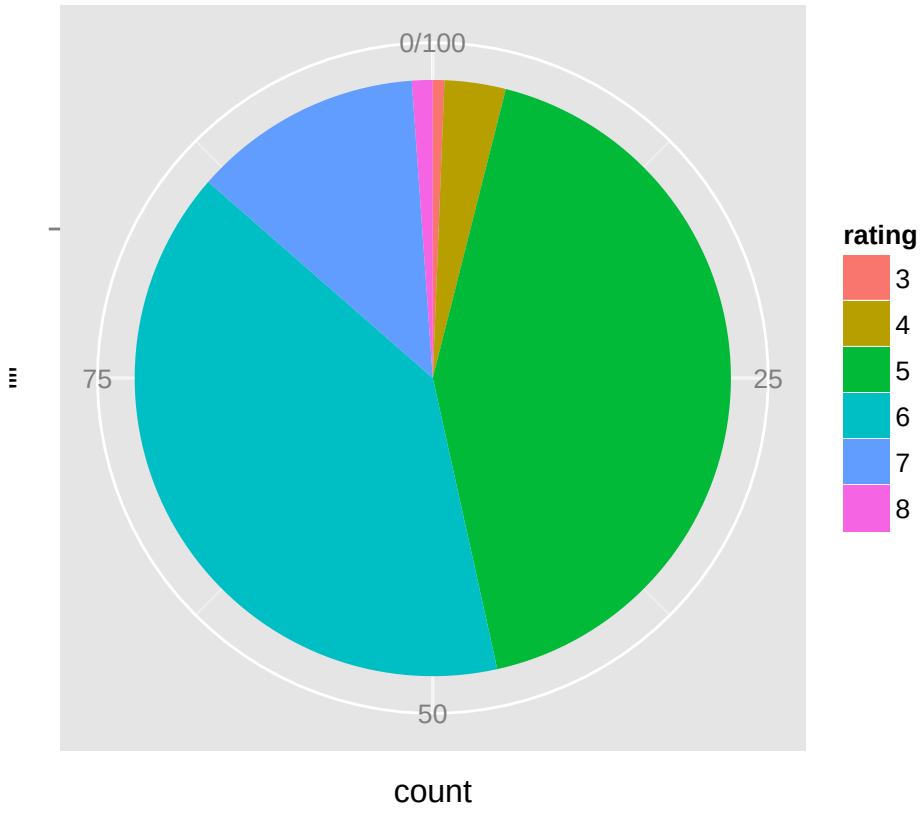
The following figure shows distributions for all features:

```
grid.arrange(qplot(data$fixed.acidity,xlab="fixed acidity"),
             qplot(data$volatile.acidity,xlab="volatile acidity"),
             qplot(data$citric.acid,xlab="citric acid"),
             qplot(data$residual.sugar,xlab="residual sugar"),
             qplot(data$chlorides,xlab="chlorides"),
             qplot(data$free.sulfur.dioxide,xlab="free sulfur dioxide"),
             qplot(data$total.sulfur.dioxide,xlab="total sulfur dioxide"),
             qplot(data$density,xlab="density"),
             qplot(data$pH,xlab="pH"),
             qplot(data$sulphates,xlab="sulphates"),
             qplot(data$alcohol,xlab="alcohol"),
             qplot(data$quality,xlab="quality"),
             ncol = 3)
```



- `density` and `pH` appear to be normally-distributed.
- As we discussed before, the feature `quality` is an categorical variable, pie chart may be better choice

```
data_quality=data.frame("count"=100*as.numeric(lapply(split(data$quality,data$quality),length))/dim(data
ggplot(data_quality, aes(x="", y=count, fill=rating))+
        geom_bar(width = 1, stat = "identity")+ coord_polar("y", start=0)
```



Short questions

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

- It is convenient to interpret the result by creating variable `rating` , classifying each wine as low, medium and high which is equally sized, randomly assign `quality` $= 5$ to low or medium level and `quality` $= 6$ to medium or high level.

```
data$rating<-rep("",dim(data)[1])
data[data$quality%in%c(3,4),"rating"]<-"low"
index_low=sample(data[data$quality==5,"X"],dim(data)[1]/3-length(data$rating[data$rating=="low"]))
data[index_low,"rating"]<-"low"
data[data$quality%in%c(7,8),"rating"]<-"high"
index_high=sample(data[data$quality==6,"X"],dim(data)[1]/3-length(data$rating[data$rating=="high"]))
data[index_high,"rating"]<-"high"
data$rating[data$rating==""]<-"medium"
data$rating<-ordered(data$rating, levels = c("low", "medium", "high"))
table(data$rating)
```

```
##
##    low medium   high
##    533    533    533
```

```
table(data$rating,data$quality)
```

```
##
##             3   4   5   6   7   8
##    low     10  53 470   0   0   0
##    medium   0   0 211 322   0   0
##    high     0   0   0 316 199  18
```

- I create a combined variable, `acid`, taking average of `fixed.acidity`, `volatile.acidity` and `citric.acid` after standardization.

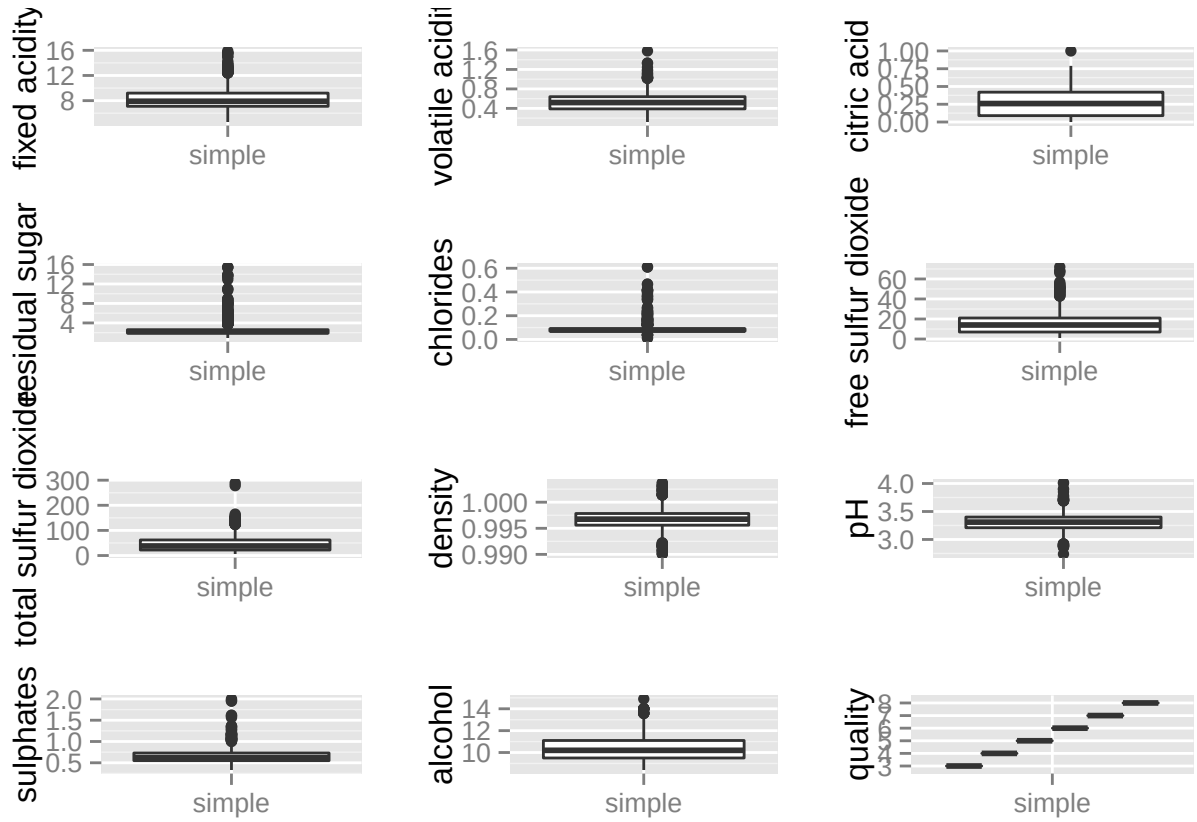
```
data$acid<-((data$fixed.acidity-mean(data$fixed.acidity))/sd(data$fixed.acidity)+
(data$volatile.acidity-mean(data$volatile.acidity))/sd(data$volatile.acidity)+
(data$citric.acid-mean(data$citric.acid))/sd(data$citric.acid))/3
```

- I create a combined variable, `dioxide`, taking average of `free.sulfur.dioxide` and `total.sulfur.dioxide` after standardization.

```
data$dioxide<-((data$free.sulfur.dioxide-mean(data$free.sulfur.dioxide))/sd(data$free.sulfur.dioxide)+
(data$total.sulfur.dioxide-mean(data$total.sulfur.dioxide))/sd(data$total.sulfur.dioxide))/2
```

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?

```
grid.arrange(qplot(x = 'simple',y=data$fixed.acidity,geom='boxplot',ylab="fixed acidity",xlab=""),
             qplot(x = 'simple',y=data$volatile.acidity,geom='boxplot',ylab="volatile acidity",xlab="")
             qplot(x = 'simple',y=data$citric.acid,geom='boxplot',ylab="citric acid",xlab=""),
             qplot(x = 'simple',y=data$residual.sugar,geom='boxplot',ylab="residual sugar",xlab=""),
             qplot(x = 'simple',y=data$chlorides,geom='boxplot',ylab="chlorides",xlab=""),
             qplot(x = 'simple',y=data$free.sulfur.dioxide,geom='boxplot',ylab="free sulfur dioxide",xla
             qplot(x = 'simple',y=data$total.sulfur.dioxide,geom='boxplot',ylab="total sulfur dioxide",x
             qplot(x = 'simple',y=data$density,geom='boxplot',ylab="density",xlab=""),
             qplot(x = 'simple',y=data$pH,geom='boxplot',ylab="pH",xlab=""),
             qplot(x = 'simple',y=data$sulphates,geom='boxplot',ylab="sulphates",xlab=""),
             qplot(x = 'simple',y=data$alcohol,geom='boxplot',ylab="alcohol",xlab=""),
             qplot(x = 'simple',y=data$quality,geom='boxplot',ylab="quality",xlab=""),
             ncol = 3)
```



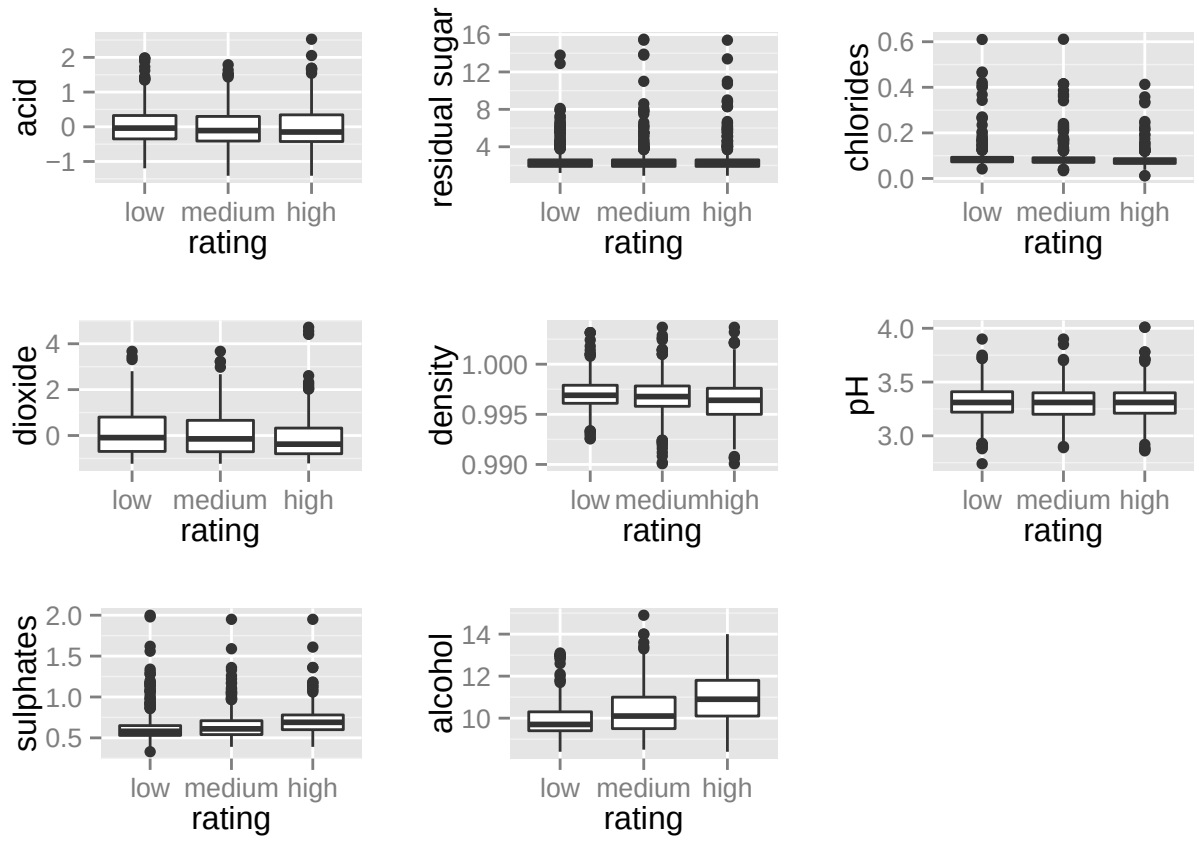
- `residual.sugar` and `chlorides` have extreme outliers.
- I don't tidy or adjust any data.

Univariate Plots Section

Univariate Analysis

The following boxplot shows how the features affect `rating`:

```
grid.arrange(qplot(x = data$rating,y=data$acid,geom='boxplot',ylab="acid",xlab="rating"),
             qplot(x = data$rating,y=data$residual.sugar,geom='boxplot',ylab="residual sugar",xlab="rati
             qplot(x = data$rating,y=data$chlorides,geom='boxplot',ylab="chlorides",xlab="rating"),
             qplot(x = data$rating,y=data$dioxide,geom='boxplot',ylab="dioxide",xlab="rating"),
             qplot(x = data$rating,y=data$density,geom='boxplot',ylab="density",xlab="rating"),
             qplot(x = data$rating,y=data$pH,geom='boxplot',ylab="pH",xlab="rating"),
             qplot(x = data$rating,y=data$sulphates,geom='boxplot',ylab="sulphates",xlab="rating"),
             qplot(x = data$rating,y=data$alcohol,geom='boxplot',ylab="alcohol",xlab="rating"),
             ncol = 3)
```

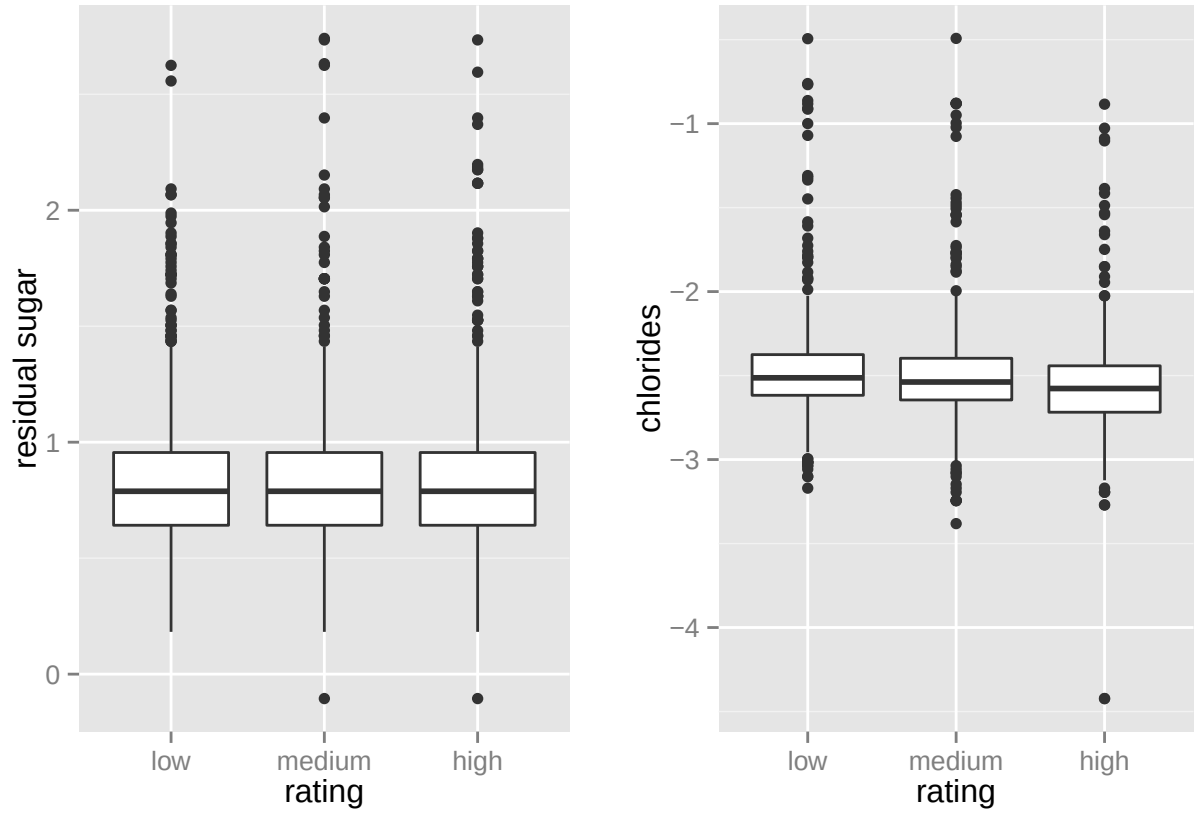


Short questions

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

As we discussed before, `residual.sugar` and `chlorides` have extreme outliers, it is difficut to verify the relationship with `rating`. We plotted on a base 10 logarithmic scale as follows:

```
grid.arrange(
             qplot(x = data$rating,y=log(data$residual.sugar),geom='boxplot',ylab="residual sugar",xlab=
             qplot(x = data$rating,y=log(data$chlorides),geom='boxplot',ylab="chlorides",xlab="rating")
             ncol = 2)
```



What was the strongest relationship you found?

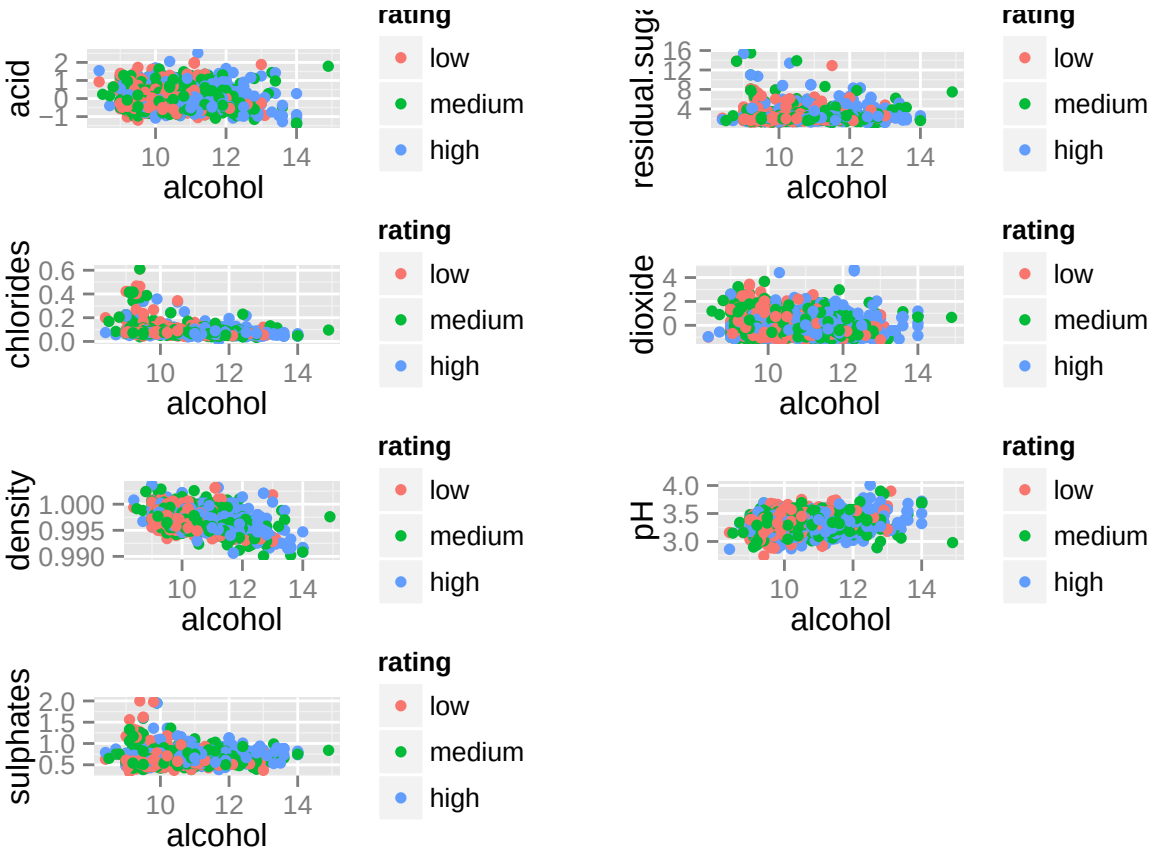
Higher `alcohol`, higher `rating`.

Multivariate Plots Section

Multivariate Analysis

I examined the scatter plot of all the pair of features containing `alcohol` between different `rating`.

```
grid.arrange(
ggplot(data = data,aes(x = alcohol, y = acid,color = rating))+geom_point() ,
ggplot(data = data,aes(x = alcohol, y = residual.sugar,color = rating)) +geom_point() ,
ggplot(data = data,aes(x = alcohol, y = chlorides,color = rating))+geom_point() ,
ggplot(data = data,aes(x = alcohol, y = dioxide,color = rating)) +geom_point() ,
ggplot(data = data,aes(x = alcohol, y = density,color = rating))+geom_point() ,
ggplot(data = data,aes(x = alcohol, y = pH,color = rating)) +geom_point() ,
ggplot(data = data,aes(x = alcohol, y = sulphates,color = rating))+geom_point() ,
ncol=2)
```



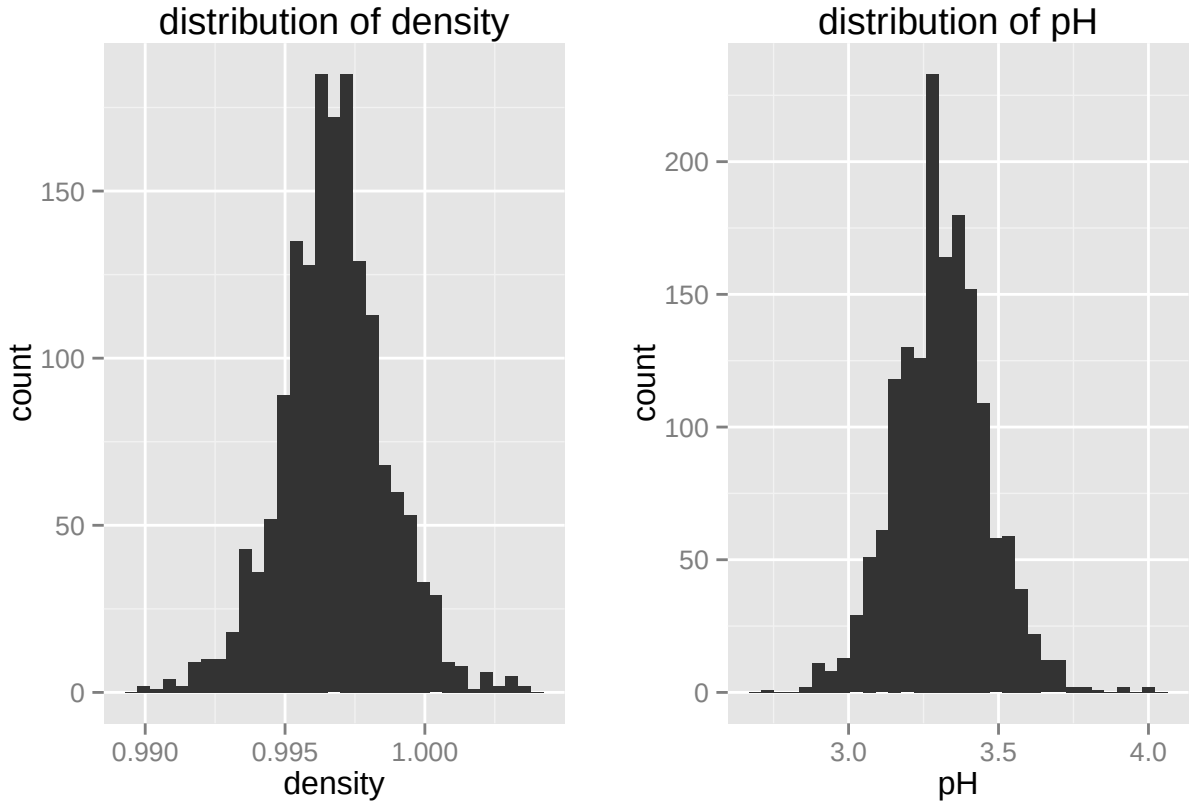
Short questions

Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

Lower `density` and higher `alcohol`, higher `rating`.

Final Plots and Summary

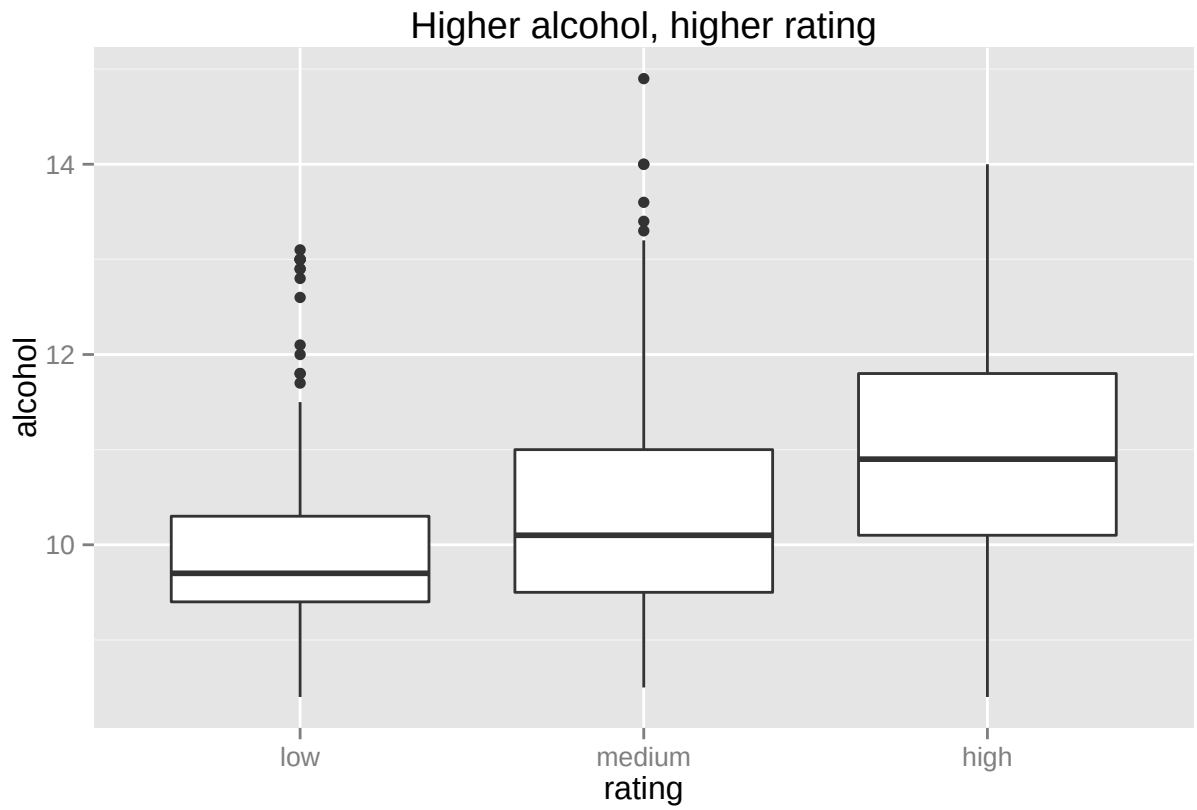
Plot One



Description One

`density` and `pH` appear to be normally-distributed.

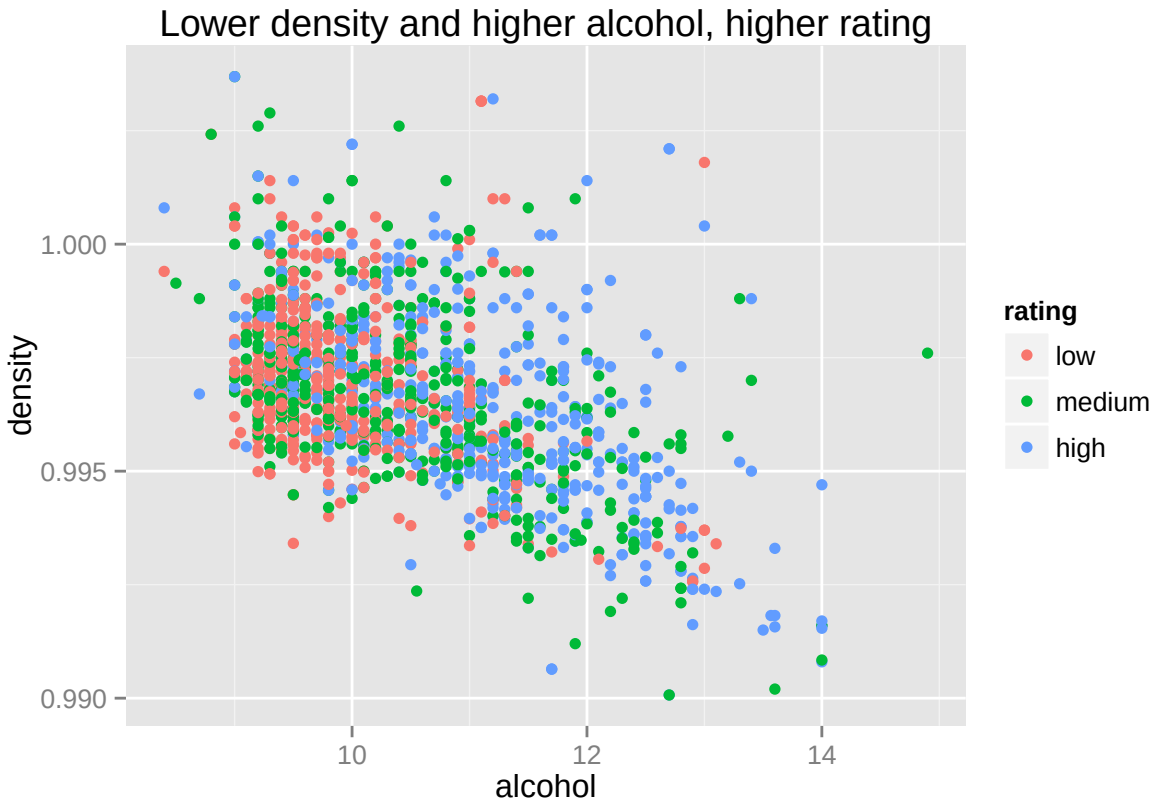
Plot Two



Description Two

Higher `alcohol`, higher `rating`.

Plot Three



Description Three

Lower `density` and higher `alcohol`, higher `rating`.

Reflection

Through this exploratory data analysis, I think feature `alcohol` influence the quality of red wines, however, wine experts give many 5 and 6 score of measure of wine quality, maybe just use the data of quality score {3,4} compare to {7,8} will show clearly patterns.