Ahmed Hisham El-sayed

exploratory data analysis 6th project of UDACITY nanodegree data analyst program

```
# Load all  packages i need
# chunk.
knitr::opts_chunk$set(echo = TRUE, fig.align='center', warning = FALSE,message = FALSE)

library(ggplot2)
library(GGally)
library(gridExtra)
library(grid)
library(foreign)
library(MASS)
library(Hmisc)
library(reshape2)
library(ggplot2)
```

```
# Load the Data
data <- read.csv('wineQualityWhites.csv')
str(data)
```

```
## 'data.frame':    4898 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
```

There are 4,898 observations and each observation has 13 variables of interest (X is count for each observation). There are 12 chemical properties the most important is quality , it is measure of each wine .

first Data Transformation

‘quality’ is now with type integer, so i thik it is better to transform it to be categorical ordeinal variable , and save this transformed data as quality.trans, but firstly i’m gonna drop X column (it is useless), and i’m gonna drop NA too

```
# since X variable is just a number represent each observation(which is useless) i'm gonna drop it and check
the latest view
data <- subset(data,select = -c(X))
head(data)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.0             0.27        0.36           20.7     0.045
## 2           6.3             0.30        0.34            1.6     0.049
## 3           8.1             0.28        0.40            6.9     0.050
## 4           7.2             0.23        0.32            8.5     0.058
## 5           7.2             0.23        0.32            8.5     0.058
## 6           8.1             0.28        0.40            6.9     0.050
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  45                  170  1.0010 3.00      0.45     8.8
## 2                  14                  132  0.9940 3.30      0.49     9.5
## 3                  30                   97  0.9951 3.26      0.44    10.1
## 4                  47                  186  0.9956 3.19      0.40     9.9
## 5                  47                  186  0.9956 3.19      0.40     9.9
## 6                  30                   97  0.9951 3.26      0.44    10.1
##   quality
## 1       6
## 2       6
## 3       6
## 4       6
## 5       6
## 6       6
```

```
# drop NA in the dataset
data <- na.omit(data)
str(data)
```

```
## 'data.frame':    4898 obs. of  12 variables:
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
```

```
# The resulting dataframe is same size, this means that there are no NA values
```

```
#  transform quality into a  factor categorical ordinal variable, save the column as quality.trasn instead o
f quality then check the lastest view
data$quality.trans <- ordered(data$quality)
str(data$quality.trans)
```

```
##  Ord.factor w/ 7 levels "3"<"4"<"5"<"6"<..: 4 4 4 4 4 4 4 4 4 4 ...
```

```
str(data)
```

```
## 'data.frame':    4898 obs. of  13 variables:
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
##  $ quality.trans       : Ord.factor w/ 7 levels "3"<"4"<"5"<"6"<..: 4 4 4 4 4 4 4 4 4 4 ...
```

A statistical summary of the data:

```
# a summary of the Data
summary(data)
```

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
##  Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
##  Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391
##  3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900
##  Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800
##
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.00900   Min.   :  2.00      Min.   :  9.0
##  1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0
##  Median :0.04300   Median : 34.00      Median :134.0
##  Mean   :0.04577   Mean   : 35.31      Mean   :138.4
##  3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0
##  Max.   :0.34600   Max.   :289.00      Max.   :440.0
##
##     density             pH          sulphates         alcohol
##  Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00
##  1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50
##  Median :0.9937   Median :3.180   Median :0.4700   Median :10.40
##  Mean   :0.9940   Mean   :3.188   Mean   :0.4898   Mean   :10.51
##  3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40
##  Max.   :1.0390   Max.   :3.820   Max.   :1.0800   Max.   :14.20
##
##     quality      quality.trans
##  Min.   :3.000   3:  20
##  1st Qu.:5.000   4: 163
##  Median :6.000   5:1457
##  Mean   :5.878   6:2198
##  3rd Qu.:6.000   7: 880
##  Max.   :9.000   8: 175
##                  9:   5
```

it is important to note at ‘quality’ and ‘quality.cat’ that there is no wine reached the lowest quality(3) nor the highest, quality (10) and it seems like there is only 20 wines got the minimum quality an only 5 got the maximum quality, however the most count of wines are in 4:8 quality, it difficult to draw any statistically significant conclusions about the extremes of the quality scale.

it appears there is fair spread in variables, so there is a meaningful differences between the min, median and max values, I will quantify one element of this spread by calculating the max:median ratio for each variable (excluding ‘quality.trans’):

```
# a function that calculates the maximum / median for any column
maxmedianratio = function(x)
                 {max(x)/median(x)}

# Apply the function to the dataset, without quality.trans
apply(subset(data,select = -c(quality.trans)),2,maxmedianratio)
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##             2.088235             4.230769             5.187500
##       residual.sugar            chlorides  free.sulfur.dioxide
##            12.653846             8.046512             8.500000
## total.sulfur.dioxide              density                   pH
##             3.283582             1.045525             1.201258
##            sulphates              alcohol              quality
##             2.297872             1.365385             1.500000
```

There is a fair amount of variance within variables, Density has the lowest ratio, with the maximum only 4.5% higher than the median. It remains to be seen whether this seemingly large spread amongst most of the variables is helpful for predicting wine quality.

by the way, I noticed that one variable (citric.acid) has a minimum of zero. Is this a missing data point or a true measurement? I will keep this in mind, but will not do anything with this observation at the moment.

Univariate Plots Section

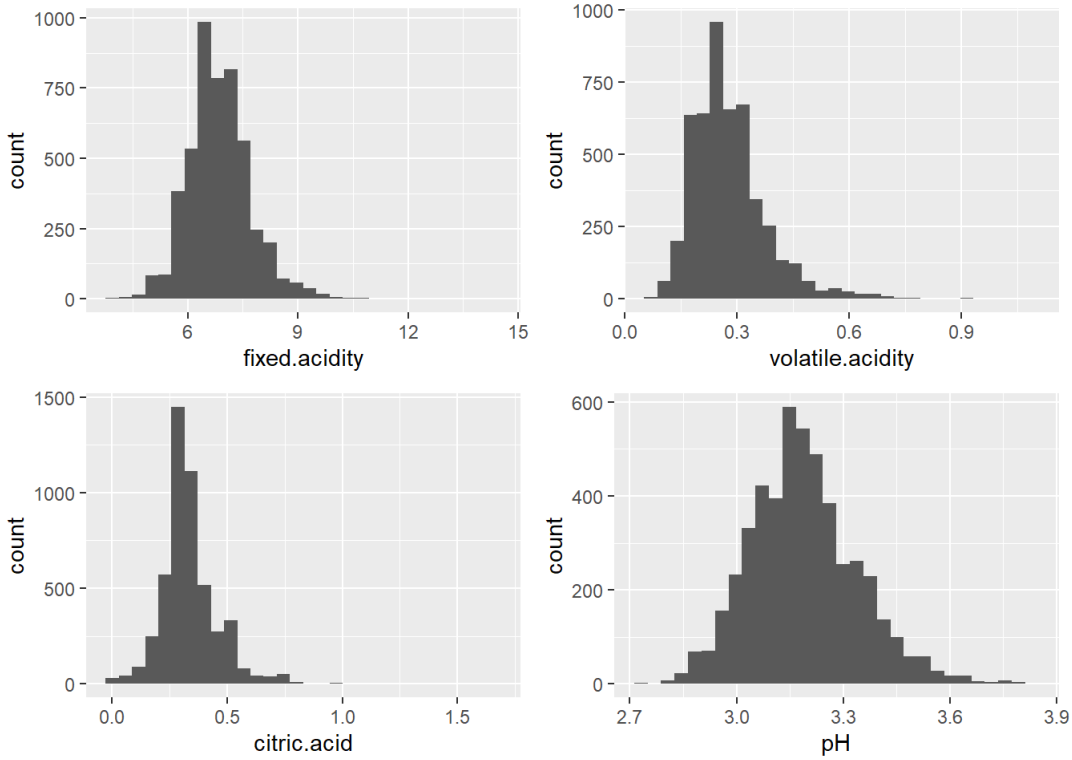
A good way to get an initial feel for the distribution of the data is via histograms. Rather than simply output 13 histograms, I will group the 13 properties into 3 different categories, and look at each category in turn. Since pH is a measure of acidity, I will group pH together with the

graphs showing the 3 acid levels (fixed.acidity, volatile.acidity, and citric.acid). Next, I will group together the 5 remaining concentration measurements (residual.sugar, chlorides,free.sulfur.dioxide, total.sulfur.dioxide, and sulphates). Finally, I will group together alcohol, density and quality.

first group

“Acidity” Related Histograms:

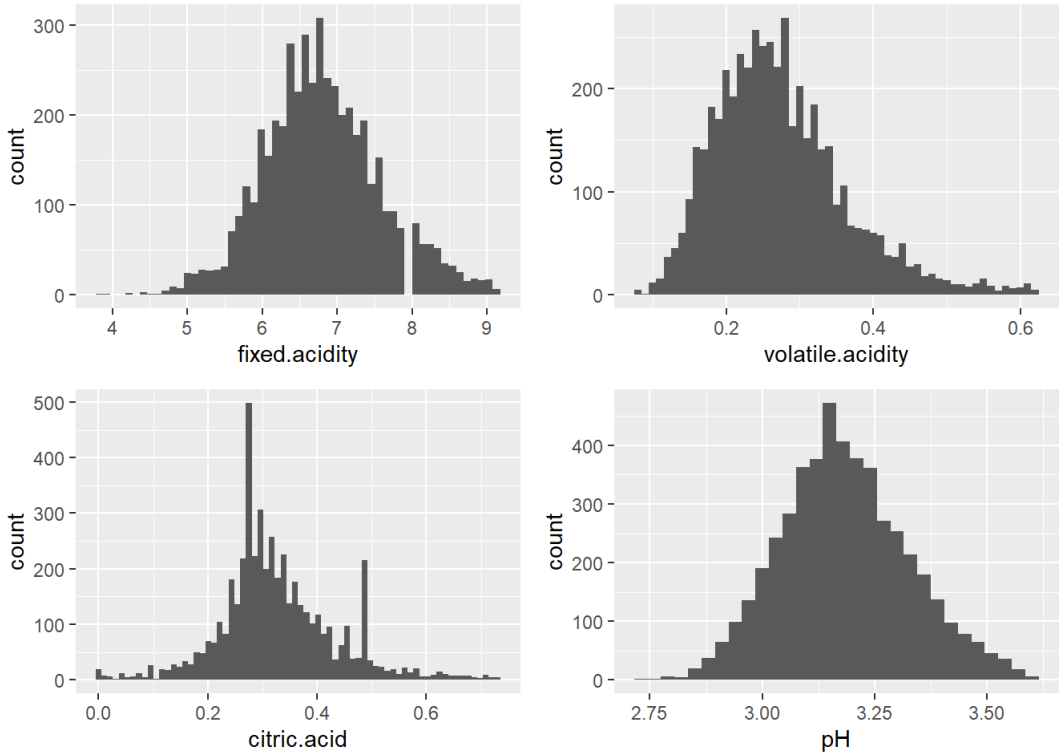
```
# Plot the 'acidity' related parameters as a group:
p1 <- ggplot(aes(fixed.acidity), data = data) + geom_histogram(bins = 30)
p2 <- ggplot(aes(volatile.acidity), data = data) + geom_histogram(bins = 30)
p3 <- ggplot(aes(citric.acid), data = data) + geom_histogram(bins = 30)
p4 <- ggplot(aes(pH), data = data) + geom_histogram(bins = 30)

grid.arrange(p1,p2,p3,p4,ncol=2)
```



These four parameters all look normally distributed. In all four cases, there is some positive skewing (right skewed and at this case it is better to consider median), with very low ‘count’ values for the higher x-axis values. it might make sense to exclude the upper most quantile (e.g. 1%) of each of these parameters which considered as outliers, to remove this skewing, which appears to impact only a small number of wines , lets take a look at what these graphs would look like if we exclude the top 1% quantile for each parameter and adjust the bin sizes a bit:

Plot the ‘acidity’ again, but with the 99+% quantile excluded:

```
p1 <- ggplot(aes(fixed.acidity), data = subset(
  data,data$fixed.acidity < quantile(data$fixed.acidity,0.99))) +
  geom_histogram(bins = 55) # we can use filter function instead of this

p2 <- ggplot(aes(volatile.acidity), data = subset(
  data,data$volatile.acidity < quantile(data$volatile.acidity,0.99))) +
  geom_histogram(bins = 55)

p3 <- ggplot(aes(citric.acid), data = subset(
  data,data$citric.acid < quantile(data$citric.acid,0.99))) +
  geom_histogram(bins = 70)

p4 <- ggplot(aes(pH), data = subset(
  data,data$pH < quantile(data$pH,0.99))) +
  geom_histogram(bins = 30)


grid.arrange(p1,p2,p3,p4,ncol=2
             )
```

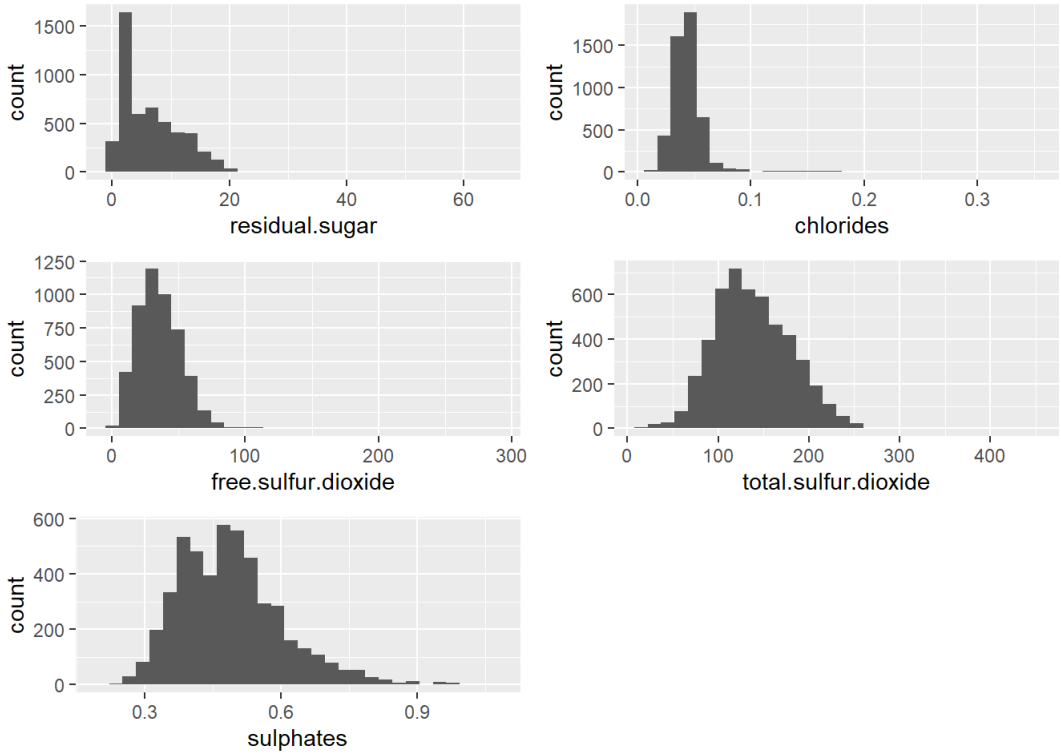


Once the top 1% of each parameter is excluded, it is easier to see the shape of the bulk of the data. There are two interesting ‘spikes’ in the citric acid profile, one near the median and a second smaller one near a value of 0.5, i tried to take log10 to ‘cetric.acid’ but it was not helpful so i suggests there might be something about the wine production process that generates an unusual citric acid profile.

second group

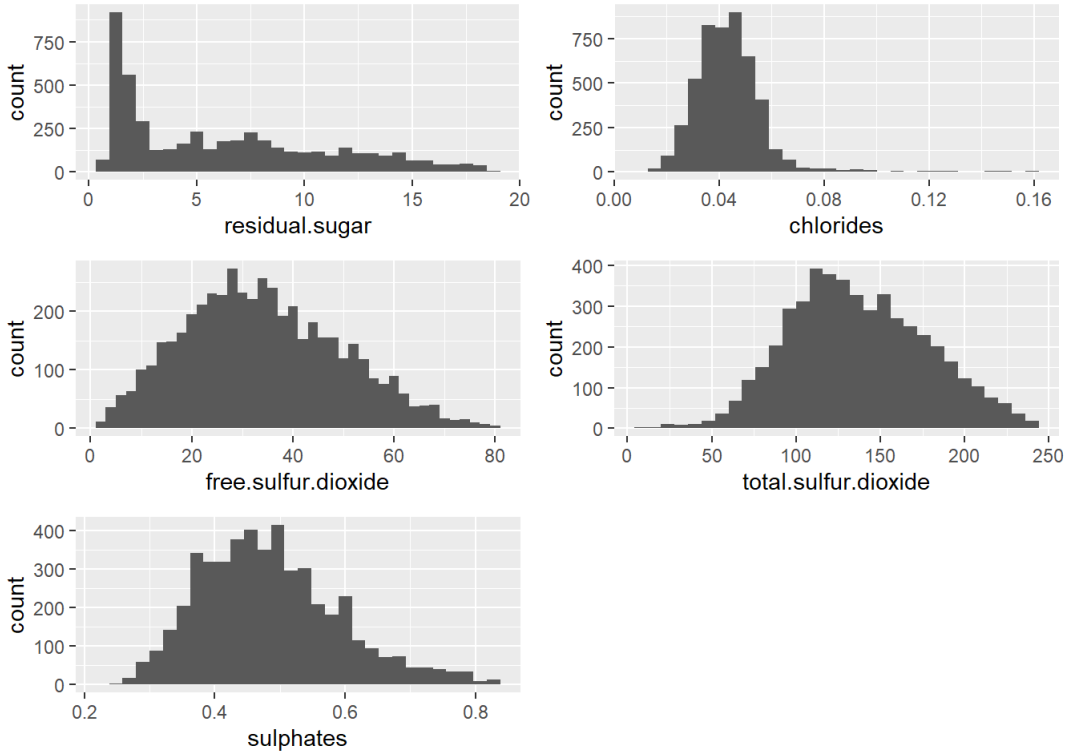
“Other Concentration” Related Histograms:

```
# Plot the 'other concentration' related parameters as a group:
p5 <- ggplot(aes(residual.sugar), data = data) + geom_histogram(bins = 30)

p6 <- ggplot(aes(chlorides), data = data) + geom_histogram(bins = 30)

p7 <- ggplot(aes(free.sulfur.dioxide), data = data) + geom_histogram(bins = 30)

p8 <- ggplot(aes(total.sulfur.dioxide), data = data) +
  geom_histogram(bins = 30)

p9 <- ggplot(aes(sulphates), data = data) + geom_histogram(bins = 30)


grid.arrange(p5,p6,p7,p8,p9,ncol=2)
```



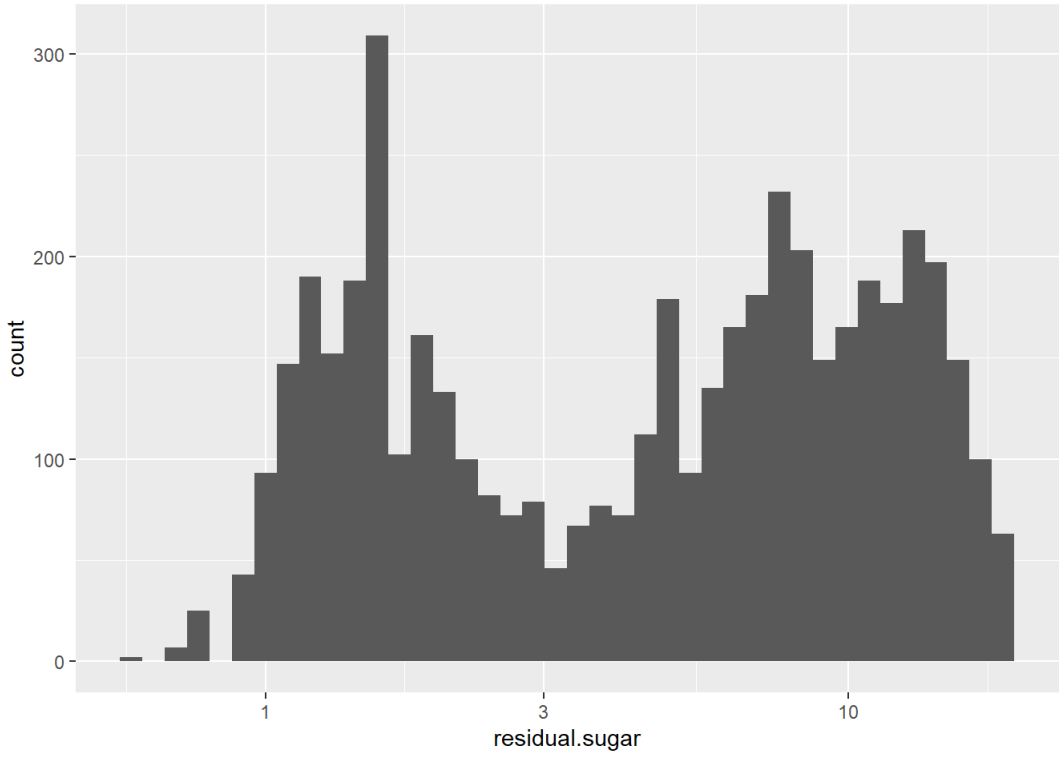
As was seen with the four “acid” related parameters, the five graphs above also exhibit positive skew. it appears that all of these parameters are normally distributed, with the exception of residual sugar, which is perhaps log normal. Lets again take a closer look, by excluding the top 1% values for each parameter:

```
# Plot the 'other concentration' again, but with the 99+% quantile excluded:

p5 <- ggplot( subset(
  data,data$residual.sugar < quantile(data$residual.sugar,0.99)),aes(residual.sugar) ) +
  geom_histogram(bins = 30)

p6 <- ggplot(aes(chlorides), data = subset(
  data,data$chlorides < quantile(data$chlorides,0.99))) +
  geom_histogram(bins = 30)

p7 <- ggplot(aes(free.sulfur.dioxide), data = subset(
  data,data$free.sulfur.dioxide < quantile(data$free.sulfur.dioxide,0.99))) +
  geom_histogram(bins = 40)

p8 <- ggplot(aes(total.sulfur.dioxide), data = subset(
  data,data$total.sulfur.dioxide < quantile(data$total.sulfur.dioxide,0.99))) +
  geom_histogram(bins = 30)

p9 <- ggplot(aes(sulphates), data = subset(
  data,data$sulphates < quantile(data$sulphates,0.99))) +
  geom_histogram(bins = 30)


grid.arrange(p5,p6,p7,p8,p9,ncol=2)
```



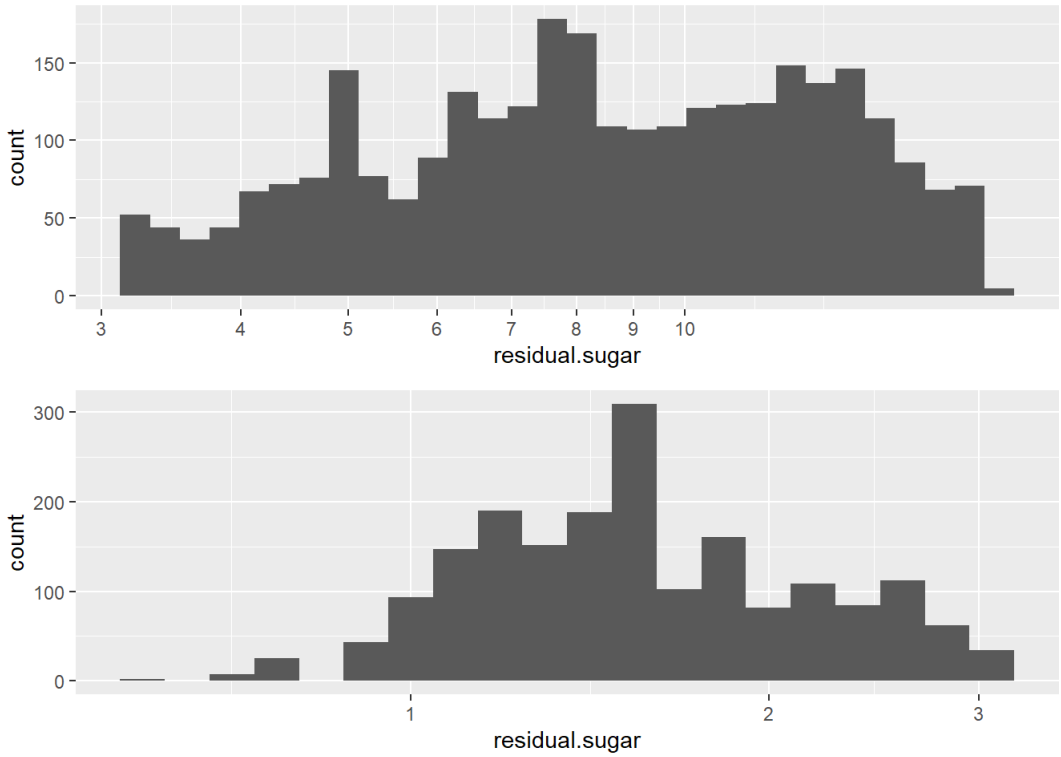
it is easier to see the shape of the bulk of the data after deleting 1 % quantile. Most parameters appear to be approximately normally distributed here, but residual.sugar it seems to be log normal. Lets make sure:

```
# Plot a log normal of residual.sugar distribution:

ggplot( data = subset(
  data,data$residual.sugar < quantile(data$residual.sugar,0.99)), aes(residual.sugar) ) +
  geom_histogram(bins = 40) +
  scale_x_log10()
```



it is a bimodal distribution might be valid and the center seems to be 3, Lets take a closer look at residual.sugar by blowing up the graph and graphing it as two separate parts, one for values above 3 and one for values below 3:

```
# Plot the residual.sugar graph twice to examine its behavior in detail:
# once for low values (<3) and once for higher values (3-20)

p10 <- ggplot(aes(residual.sugar), data = subset(
  data,data$residual.sugar < quantile(
    data$residual.sugar,0.99) & data$residual.sugar > 3.1)) +
  geom_histogram(bins = 30) +
  scale_x_log10(breaks=3:10)

p11 <- ggplot(aes(residual.sugar), data = subset(
  data,data$residual.sugar < quantile(
    data$residual.sugar,0.99) & data$residual.sugar <= 3.1)) +
  geom_histogram(bins = 20) +
  scale_x_log10(breaks=1:3)


grid.arrange(p10,p11,ncol=1)
```



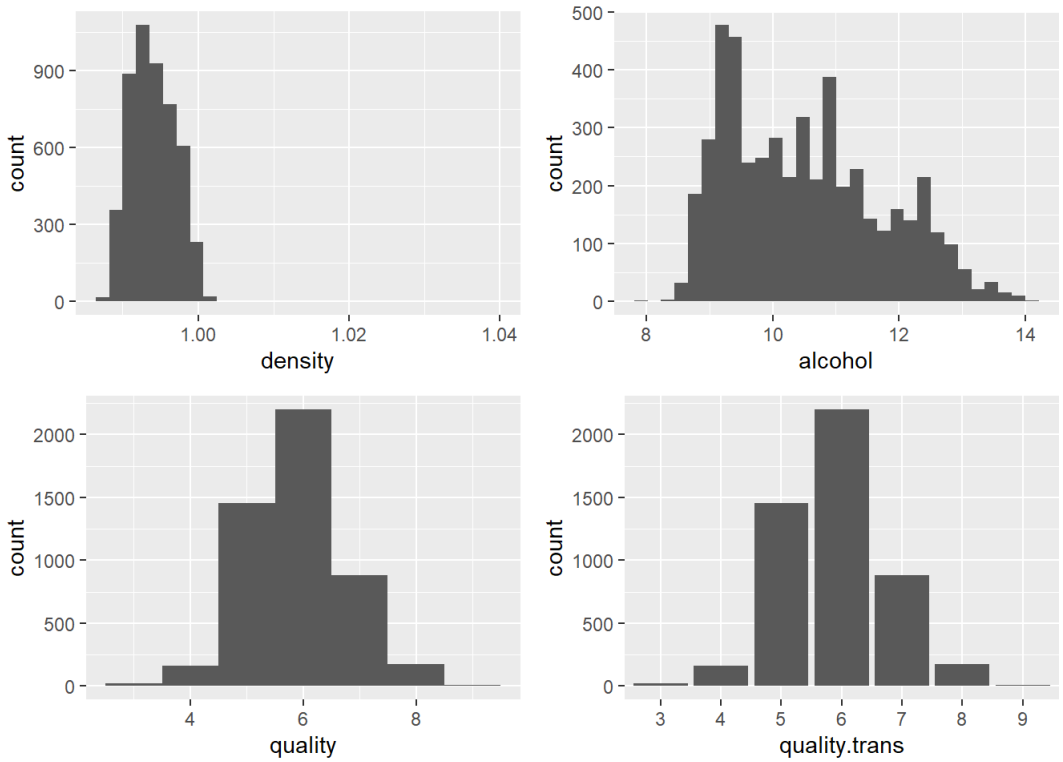
It is hard to draw any firm conclusions as to what is happening to explain the residual.sugar pattern, but the final graphs above suggest that a bimodal distribution might be a reasonable fit and that this could be a result of having two distinct methods for producing this wine type, one of which results in significantly lower residual sugar levels than the other.

group 3

“Other” Variables Histograms

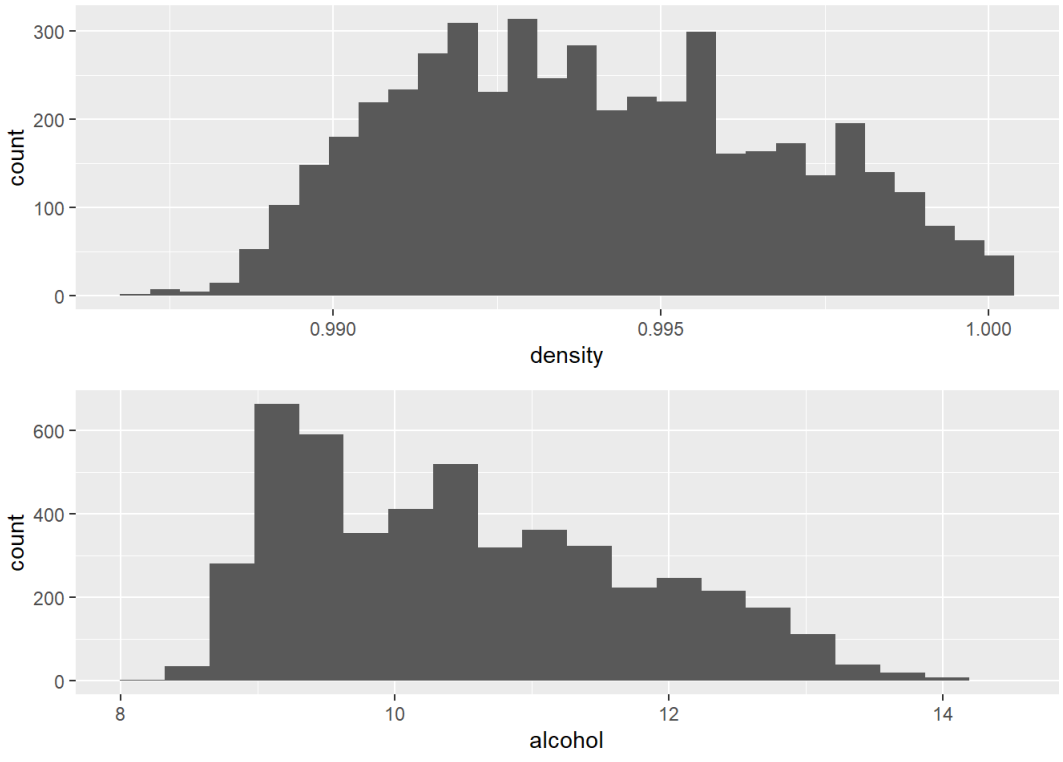
(Note: a bar chart is used in the case of ‘quality.cat’, since it is categorical):

```
# Plot the remaining parameters as a group:
p12 <- ggplot(aes(density), data = data) + geom_histogram(bins = 30)

p13 <- ggplot(aes(alcohol), data = data) + geom_histogram(bins = 30)

p14 <- ggplot(aes(quality), data = data) + geom_histogram(bins = 7)

p15 <- ggplot(aes(quality.trans), data = data) + geom_bar(stat="count")

grid.arrange(p12,p13,p14,p15,ncol=2)
```



The quality rating appears to be normally distributed, Density appears normal too, but with some positive skew. The alcohol content is an interesting one, it seems to be trimodal . Lets take a closer look at density and alcohol content, by replotting without the top 1% quantile:

```
# Plot density and alcohol again, excluding the top 99+% quantile for density:

p12 <- ggplot(aes(density), data = subset(
  data,data$density < quantile(data$density,0.99))) + geom_histogram(bins = 30)

p13 <- ggplot(aes(alcohol), data = data) + geom_histogram(bins = 20)


grid.arrange(p12,p13,ncol=1)
```



Density looks fairly normally distributed, whereas alcohol content might be bimodal or even trimodal.

Create New Variables

the chlorides to sulphates ratio might be a far more important measure of quality than the individual levels of either ion. Perhaps this ratio is

important for wine too, so I will create a chlorides-to-sulphate ratio variable.

In addition, I decided that the free-to-total sulfur dioxide ratio might be interesting .

I think the ratio of volatile acidity to fixed acidity might be important, since there might be a chemical interplay between the two forms of acidity. Finally, I decided to also create a sugar-to-alcohol ratio, since both variables exhibited strange, bimodal like behavior and intuitively it seemed there might be some interplay here, with a sugary taste potentially masking the sometimes unpalatable taste of a higher alcohol content. The new ratios were created and their descriptive statistics and histograms (excluding top 1% quantile) are presented below:

```
# Review the data
head(data)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.0             0.27        0.36           20.7     0.045
## 2           6.3             0.30        0.34            1.6     0.049
## 3           8.1             0.28        0.40            6.9     0.050
## 4           7.2             0.23        0.32            8.5     0.058
## 5           7.2             0.23        0.32            8.5     0.058
## 6           8.1             0.28        0.40            6.9     0.050
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  45                  170  1.0010 3.00      0.45     8.8
## 2                  14                  132  0.9940 3.30      0.49     9.5
## 3                  30                   97  0.9951 3.26      0.44    10.1
## 4                  47                  186  0.9956 3.19      0.40     9.9
## 5                  47                  186  0.9956 3.19      0.40     9.9
## 6                  30                   97  0.9951 3.26      0.44    10.1
##   quality quality.trans
## 1       6             6
## 2       6             6
## 3       6             6
## 4       6             6
## 5       6             6
## 6       6             6
```

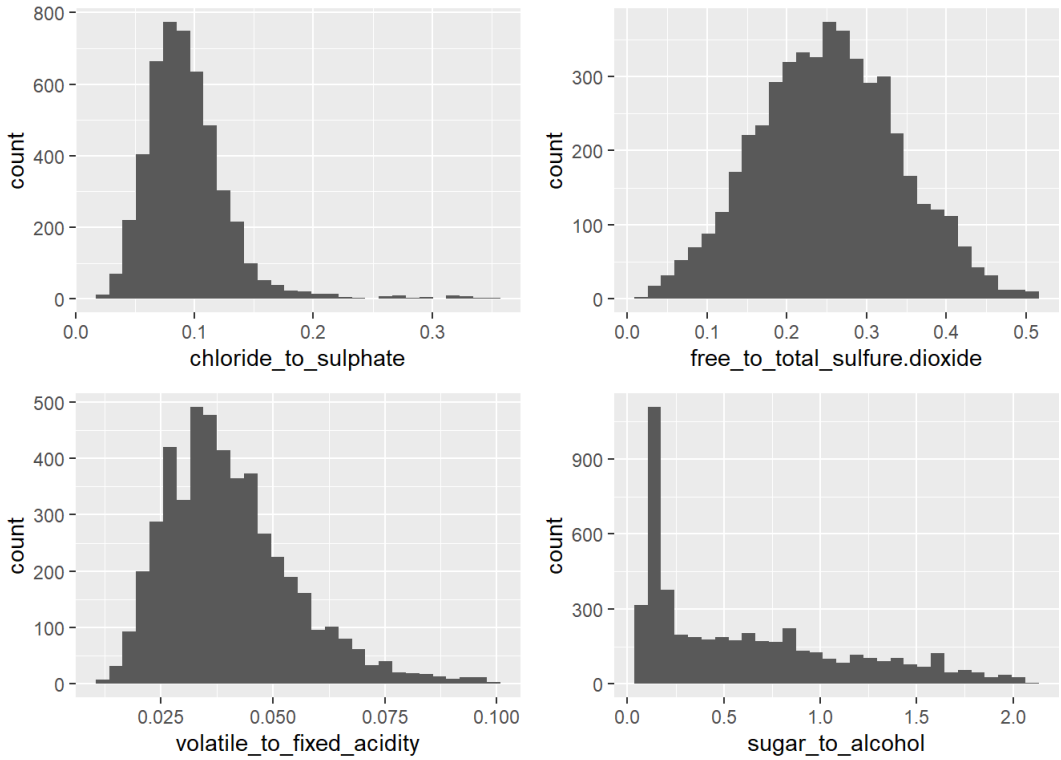
```
# Create and add four new variables to the dataframe:
data$chloride_to_sulphate <-with(data,chlorides / sulphates)

data$free_to_total_sulfure.dioxide <-with(
  data,free.sulfur.dioxide / total.sulfur.dioxide)

data$volatile_to_fixed_acidity <-with(data,volatile.acidity / fixed.acidity)

data$sugar_to_alcohol <-with(data,residual.sugar / alcohol)

# Output summary data on the new variables:
str(subset(data,select = c(chloride_to_sulphate,free_to_total_sulfure.dioxide,
                           volatile_to_fixed_acidity,sugar_to_alcohol)))
```

```
## 'data.frame':    4898 obs. of  4 variables:
##  $ chloride_to_sulphate         : num  0.1 0.1 0.114 0.145 0.145 ...
##  $ free_to_total_sulfure.dioxide: num  0.265 0.106 0.309 0.253 0.253 ...
##  $ volatile_to_fixed_acidity    : num  0.0386 0.0476 0.0346 0.0319 0.0319 ...
##  $ sugar_to_alcohol             : num  2.352 0.168 0.683 0.859 0.859 ...
```

```
summary(subset(data,
               select = c(chloride_to_sulphate,free_to_total_sulfure.dioxide,
                           volatile_to_fixed_acidity,sugar_to_alcohol)))
```

```
##  chloride_to_sulphate free_to_total_sulfure.dioxide
##  Min.   :0.02121      Min.   :0.02362
##  1st Qu.:0.07143      1st Qu.:0.19093
##  Median :0.08980      Median :0.25368
##  Mean   :0.09774      Mean   :0.25558
##  3rd Qu.:0.11053      3rd Qu.:0.31579
##  Max.   :0.62708      Max.   :0.71053
##  volatile_to_fixed_acidity sugar_to_alcohol
##  Min.   :0.01111           Min.   :0.0566
##  1st Qu.:0.03030           1st Qu.:0.1575
##  Median :0.03836           Median :0.4906
##  Mean   :0.04126           Mean   :0.6423
##  3rd Qu.:0.04848           3rd Qu.:0.9773
##  Max.   :0.18033           Max.   :5.6239
```

```
# Plot the new parameters as a group:
p16 <- ggplot(aes(chloride_to_sulphate), data = subset(
  data,data$chloride_to_sulphate < quantile(data$chloride_to_sulphate,0.99))) +
  geom_histogram(bins = 30)

p17 <- ggplot(aes(free_to_total_sulfure.dioxide), data = subset(
  data,data$free_to_total_sulfure.dioxide <
    quantile(data$free_to_total_sulfure.dioxide,0.99))) +
  geom_histogram(bins = 30)

p18 <- ggplot(aes(volatile_to_fixed_acidity), data = subset(
  data,data$volatile_to_fixed_acidity <
    quantile(data$volatile_to_fixed_acidity,0.99))) +
  geom_histogram(bins = 30)

p19 <- ggplot(aes(sugar_to_alcohol), data = subset(
  data,data$sugar_to_alcohol < quantile(data$sugar_to_alcohol,0.99))) +
  geom_histogram(bins = 30)


grid.arrange(p16,p17,p18,p19,ncol=2)
```



The free:total sulfur dioxide graph looks normally distributed. The chloride:sulphate, volatile:fixed acidity and sugar:alcohol graphs look positively skewed. In addition, the sugar:alcohol graph exhibts the same potentially bimodal behavior exhibited by the sugar and the alcohol graphs.

I am now ready to look at the relationship between the various parameters.

Univariate Analysis

What is the structure of your dataset?

normal distribution is the most spread here in most variables

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

acidity

What other features in the dataset do you think will help support your

###investigation into your feature(s) of interest? yes, sulfure

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

no

Of the features you investigated, were there any unusual distributions?

yes , at alchole

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? No

Bivariate Plots Section

I would like to start the bivariate analysis by looking at the correlation coefficients between the variables, as given below:

```
# Determine correlation coefficients among the variables
cor(subset(data,select = -c(quality.trans)))
```

```
##                              fixed.acidity volatile.acidity  citric.acid
## fixed.acidity                   1.00000000      -0.02269729  0.289180698
## volatile.acidity               -0.02269729       1.00000000 -0.149471811
## citric.acid                     0.28918070      -0.14947181  1.000000000
## residual.sugar                  0.08902070       0.06428606  0.094211624
## chlorides                       0.02308564       0.07051157  0.114364448
## free.sulfur.dioxide            -0.04939586      -0.09701194  0.094077221
## total.sulfur.dioxide            0.09106976       0.08926050  0.121130798
## density                         0.26533101       0.02711385  0.149502571
## pH                             -0.42585829      -0.03191537 -0.163748211
## sulphates                      -0.01714299      -0.03572815  0.062330940
## alcohol                        -0.12088112       0.06771794 -0.075728730
## quality                        -0.11366283      -0.19472297 -0.009209091
## chloride_to_sulphate            0.02109576       0.05938449  0.088940695
## free_to_total_sulfure.dioxide  -0.13945918      -0.19616085  0.016241396
## volatile_to_fixed_acidity      -0.33775891       0.93662196 -0.245794409
## sugar_to_alcohol                0.09363299       0.04575732  0.102730408
##                              residual.sugar   chlorides
## fixed.acidity                    0.08902070  0.02308564
## volatile.acidity                 0.06428606  0.07051157
## citric.acid                      0.09421162  0.11436445
## residual.sugar                   1.00000000  0.08868454
## chlorides                        0.08868454  1.00000000
## free.sulfur.dioxide              0.29909835  0.10139235
## total.sulfur.dioxide             0.40143931  0.19891030
## density                          0.83896645  0.25721132
## pH                              -0.19413345 -0.09043946
## sulphates                       -0.02666437  0.01676288
## alcohol                         -0.45063122 -0.36018871
## quality                         -0.09757683 -0.20993441
## chloride_to_sulphate             0.07800801  0.90145468
## free_to_total_sulfure.dioxide    0.05142979 -0.03321768
## volatile_to_fixed_acidity        0.01515041  0.04457791
## sugar_to_alcohol                 0.99001187  0.11932114
##                              free.sulfur.dioxide total.sulfur.dioxide
## fixed.acidity                      -0.0493958591          0.091069756
## volatile.acidity                   -0.0970119393          0.089260504
```

```
## citric.acid                          0.0940772210          0.121130798
## residual.sugar                       0.2990983537          0.401439311
## chlorides                            0.1013923521          0.198910300
## free.sulfur.dioxide                  1.0000000000          0.615500965
## total.sulfur.dioxide                 0.6155009650          1.000000000
## density                              0.2942104109          0.529881324
## pH                                  -0.0006177961          0.002320972
## sulphates                            0.0592172458          0.134562367
## alcohol                             -0.2501039415         -0.448892102
## quality                              0.0081580671         -0.174737218
## chloride_to_sulphate                 0.0793673567          0.109439559
## free_to_total_sulfure.dioxide        0.7386321024         -0.013447850
## volatile_to_fixed_acidity           -0.0848067079          0.039437265
## sugar_to_alcohol                     0.3143238443          0.429487399
##                                    density            pH   sulphates
## fixed.acidity                   0.26533101 -0.4258582910 -0.01714299
## volatile.acidity                0.02711385 -0.0319153683 -0.03572815
## citric.acid                     0.14950257 -0.1637482114  0.06233094
## residual.sugar                  0.83896645 -0.1941334540 -0.02666437
## chlorides                       0.25721132 -0.0904394560  0.01676288
## free.sulfur.dioxide             0.29421041 -0.0006177961  0.05921725
## total.sulfur.dioxide            0.52988132  0.0023209718  0.13456237
## density                         1.00000000 -0.0935914935  0.07449315
## pH                             -0.09359149  1.0000000000  0.15595150
## sulphates                       0.07449315  0.1559514973  1.00000000
## alcohol                        -0.78013762  0.1214320987 -0.01743277
## quality                        -0.30712331  0.0994272457  0.05367788
## chloride_to_sulphate            0.18691463 -0.1454060343 -0.36185381
## free_to_total_sulfure.dioxide -0.06552475  0.0008012900 -0.02236186
## volatile_to_fixed_acidity     -0.07540469  0.1136748292 -0.02891024
## sugar_to_alcohol                0.87168339 -0.2013195265 -0.01803066
##                                    alcohol      quality
## fixed.acidity                  -0.12088112 -0.113662831
## volatile.acidity                0.06771794 -0.194722969
## citric.acid                    -0.07572873 -0.009209091
## residual.sugar                 -0.45063122 -0.097576829
## chlorides                      -0.36018871 -0.209934411
## free.sulfur.dioxide            -0.25010394  0.008158067
## total.sulfur.dioxide           -0.44889210 -0.174737218
## density                        -0.78013762 -0.307123313
## pH                              0.12143210  0.099427246
## sulphates                      -0.01743277  0.053677877
## alcohol                         1.00000000  0.435574715
## quality                         0.43557472  1.000000000
## chloride_to_sulphate           -0.30635643 -0.192803276
## free_to_total_sulfure.dioxide   0.06446642  0.197214077
## volatile_to_fixed_acidity       0.11281181 -0.141314426
## sugar_to_alcohol               -0.53683146 -0.134750485
##                                chloride_to_sulphate
## fixed.acidity                           0.02109576
## volatile.acidity                        0.05938449
## citric.acid                             0.08894070
## residual.sugar                          0.07800801
## chlorides                               0.90145468
## free.sulfur.dioxide                     0.07936736
## total.sulfur.dioxide                    0.10943956
## density                                 0.18691463
## pH                                     -0.14540603
## sulphates                              -0.36185381
## alcohol                                -0.30635643
## quality                                -0.19280328
## chloride_to_sulphate                    1.00000000
## free_to_total_sulfure.dioxide           0.00332709
## volatile_to_fixed_acidity               0.03753703
## sugar_to_alcohol                        0.10129093
##                                free_to_total_sulfure.dioxide
## fixed.acidity                                     -0.13945918
## volatile.acidity                                  -0.19616085
## citric.acid                                        0.01624140
## residual.sugar                                     0.05142979
## chlorides                                         -0.03321768
## free.sulfur.dioxide                                0.73863210
## total.sulfur.dioxide                              -0.01344785
```

```
## density                                          -0.06552475
## pH                                                0.00080129
## sulphates                                        -0.02236186
## alcohol                                           0.06446642
## quality                                           0.19721408
## chloride_to_sulphate                              0.00332709
## free_to_total_sulfure.dioxide                     1.00000000
## volatile_to_fixed_acidity                        -0.13913499
## sugar_to_alcohol                                  0.04818126
##                              volatile_to_fixed_acidity sugar_to_alcohol
## fixed.acidity                             -0.337758911      0.093632991
## volatile.acidity                           0.936621961      0.045757325
## citric.acid                               -0.245794409      0.102730408
## residual.sugar                             0.015150413      0.990011869
## chlorides                                  0.044577906      0.119321139
## free.sulfur.dioxide                       -0.084806708      0.314323844
## total.sulfur.dioxide                       0.039437265      0.429487399
## density                                   -0.075404688      0.871683392
## pH                                         0.113674829     -0.201319527
## sulphates                                 -0.028910236     -0.018030664
## alcohol                                    0.112811806     -0.536831461
## quality                                   -0.141314426     -0.134750485
## chloride_to_sulphate                       0.037537028      0.101290928
## free_to_total_sulfure.dioxide             -0.139134993      0.048181261
## volatile_to_fixed_acidity                  1.000000000     -0.002877822
## sugar_to_alcohol                          -0.002877822      1.000000000
```

Based on the correlations, it appears several chemicals negatively impact quality (correlations are shown in parentheses below): * fixed.acidity (-0.11)
* volatile.acidity (-0.19)
* citric.acid (-0.01)
* residual.sugar (-0.10)
* chlorides (-0.21)

Let's create a new variable, 'bad_solids', that adds them together . The new variable has the following statistics and correlation coefficient with quality:

```
# Create and add a new variable, 'bad_solids', to the dataframe:
data$bad_solids <-with(data,fixed.acidity + volatile.acidity +
                         citric.acid + residual.sugar + chlorides)
# Output summary data on the new variable:
str(data$bad_solids)
```

```
##  num [1:4898] 28.38 8.59 15.73 16.31 16.31 ...
```

```
summary(data$bad_solids)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.883   9.470  12.669  13.904  17.472  75.239
```

Correlation with quality:

```
# Determine the correlation of bad_solids with quality
cor(data$bad_solids,data$quality)
```

```
## [1] -0.1175407
```

This new variable has negatively correlation with quality. Its correlation coefficient (-0.117) is weaker than or basically equal to the individual correlations of many of its components. So this avenue looks like a dead end and I will not utilize this particular variable going forward.

I would like to narrow down the analysis to those variables that have a modest correlation with quality (say a coefficient with an absolute value on the order of 0.15). The list is as follows, with the correlations versus quality shown in parentheses: * volatile.acidity (-0.19)
* chlorides (-0.21)
* total.sulfur.dioxide (-0.17)
* density (-0.31)
* alcohol (0.44)
* chloride_to_sulphate (-0.19)
* free_to_total_sulfure.dioxide (0.20)

* sugar_to_alcohol (-0.13)
* volatile_to_fixed_acidity (-0.14)

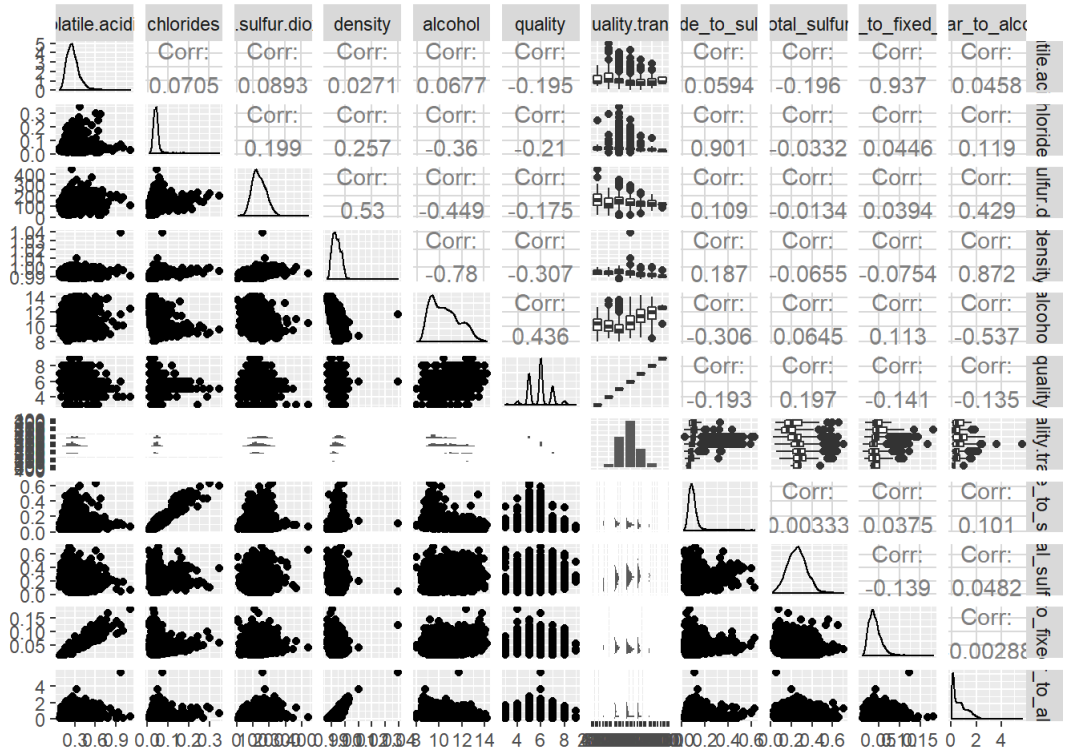
The list of variables being dropped (since their correlations with quality aren't high enough) are as follows: * fixed.acidity
* citric.acid
* residual.sugar
* free.sulfur.dioxide
* pH
* sulphates
* bad_solids

A scatterplot matrix can be a helpful early step in EDA. This matrix will allow us to get a sense as to whether there are trends between various variables in the dataset. First, I'll generate a scatterplot matrix using all the selected variables:

```
# Drop the variables identified previously:
data_subset <- subset(data,select= -c(fixed.acidity,citric.acid,residual.sugar,
 free.sulfur.dioxide,pH,sulphates,bad_solids))

# view data_subset
head(data_subset)
```

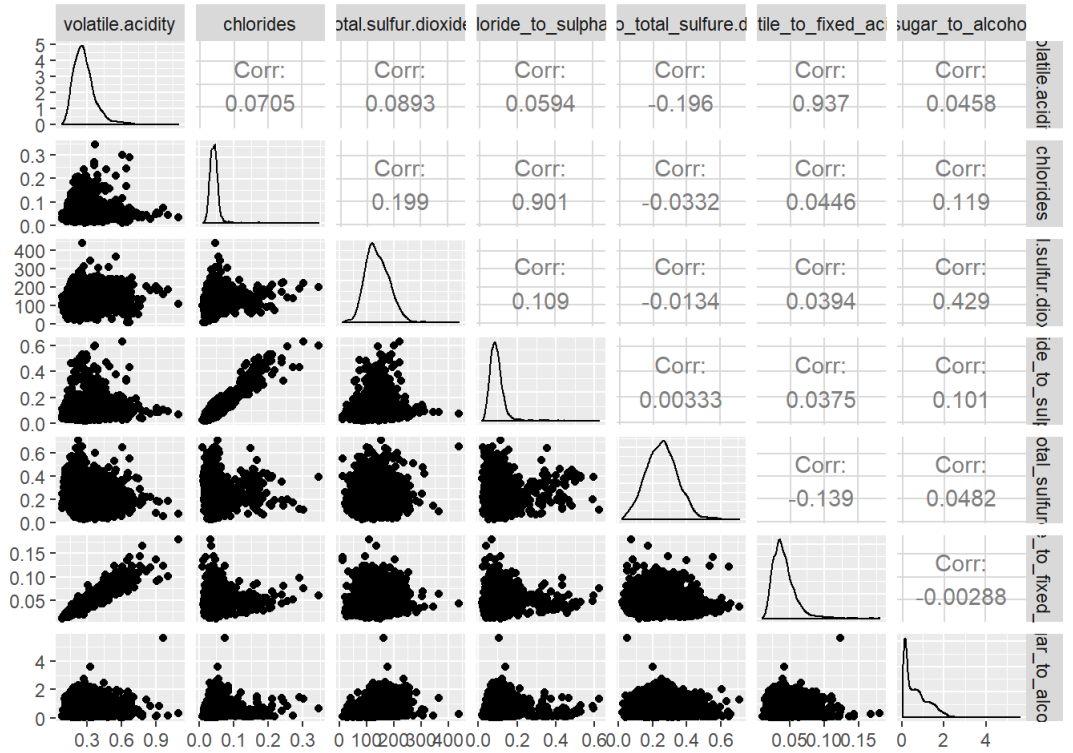
```
##   volatile.acidity chlorides total.sulfur.dioxide density alcohol quality
## 1             0.27     0.045                 170  1.0010     8.8       6
## 2             0.30     0.049                 132  0.9940     9.5       6
## 3             0.28     0.050                  97  0.9951    10.1       6
## 4             0.23     0.058                 186  0.9956     9.9       6
## 5             0.23     0.058                 186  0.9956     9.9       6
## 6             0.28     0.050                  97  0.9951    10.1       6
##   quality.trans chloride_to_sulphate free_to_total_sulfure.dioxide
## 1             6            0.1000000                     0.2647059
## 2             6            0.1000000                     0.1060606
## 3             6            0.1136364                     0.3092784
## 4             6            0.1450000                     0.2526882
## 5             6            0.1450000                     0.2526882
## 6             6            0.1136364                     0.3092784
##   volatile_to_fixed_acidity sugar_to_alcohol
## 1                0.03857143        2.3522727
## 2                0.04761905        0.1684211
## 3                0.03456790        0.6831683
## 4                0.03194444        0.8585859
## 5                0.03194444        0.8585859
## 6                0.03456790        0.6831683
```

```
# Generate the scatterplot to figure out the correlation between variables to each others:
ggpairs(data_subset)
```

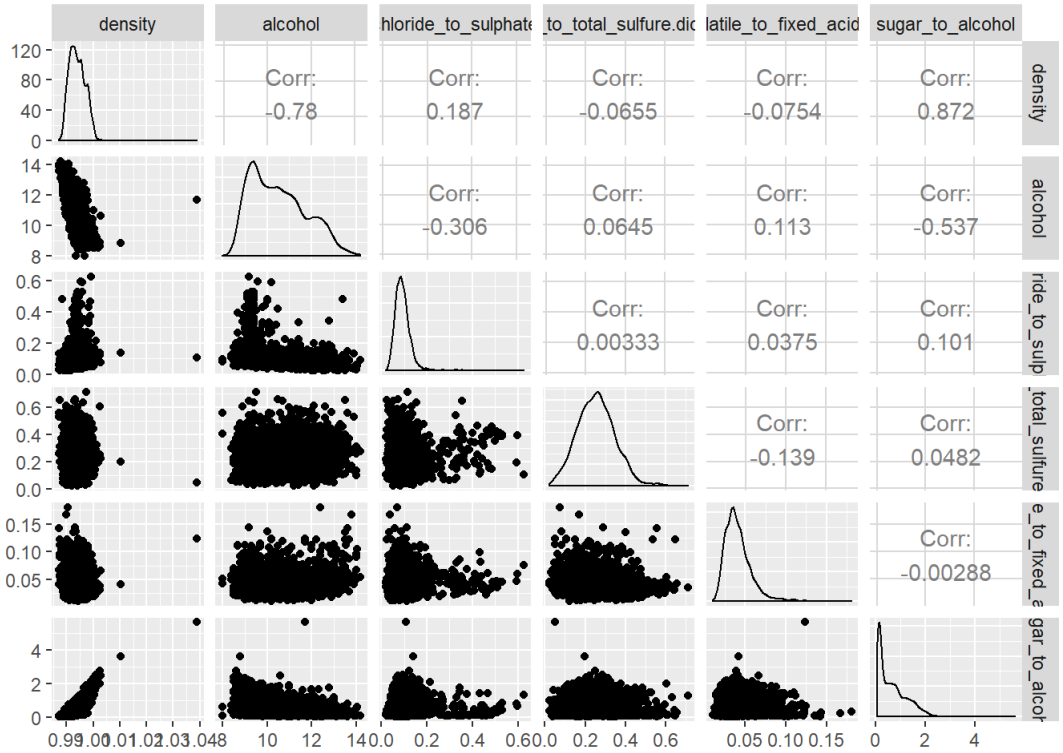


Although there appear to be some trends, the plot is too dense for any meaningful analysis, so I will split it up a bit. First, I will generate two scatterplot matrices that involve the primary feature of interest (quality):

```
# Plot subsets of the data, for ease of viewing:
data_subset3 <- subset(data_subset,
                       select=c(volatile.acidity,chlorides,total.sulfur.dioxide,
                                chloride_to_sulphate,free_to_total_sulfure.dioxide,
                                volatile_to_fixed_acidity,sugar_to_alcohol))
ggpairs(data_subset3)
```




```
data_subset4 <- subset(data_subset,
                       select=c(density,alcohol, chloride_to_sulphate,
                                free_to_total_sulfure.dioxide,
                                volatile_to_fixed_acidity,sugar_to_alcohol))
ggpairs(data_subset4)
```



The most interesting observations I gleam from these additional scatterplots are as follows: * In these plots, I observed some very strong correlations (0.9+) between the ratio variables I created and their components (e.g. the chloride:sulfate ratio has a 0.90 correlation with the chloride level). While some correlation is obviously expected, since the derived variable contains the component variable, a correlation at this level is suggestive that there is a link of some sort between the components themselves, and that the ratios therefore might have statistical significance.

- The strongest correlation observed amongst variables that do not inovolve derivatives of themselves is the 0.87 observed between the sugar:alcohol ratio and density.

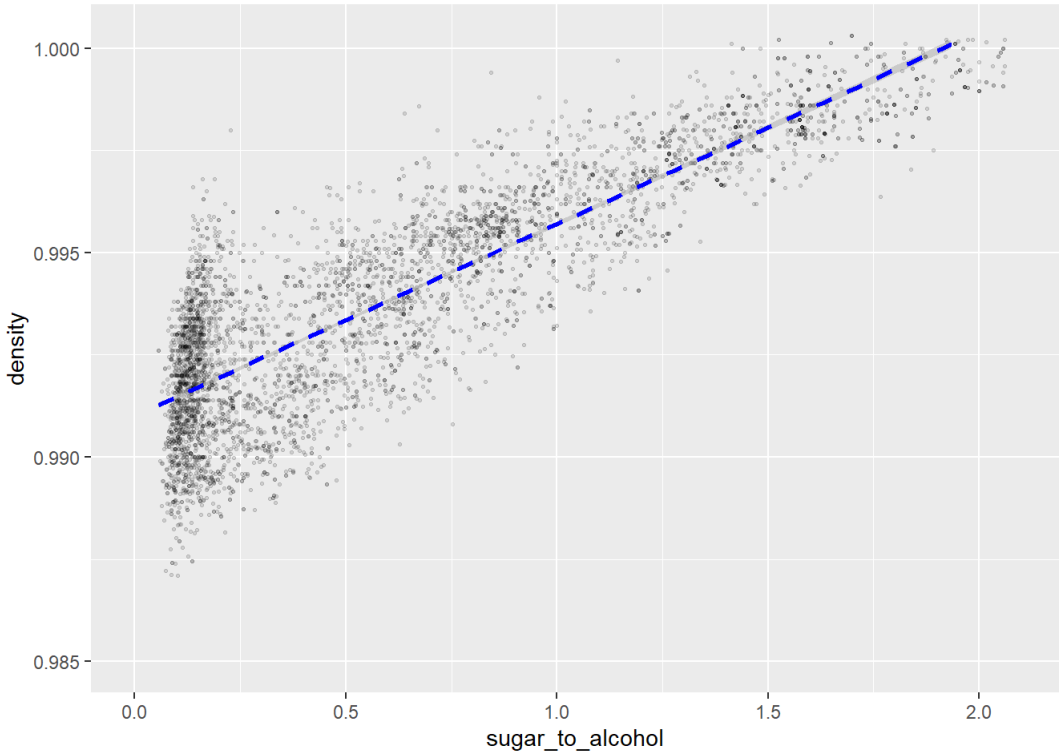
Let's now take a closer look at some of the interesting bivariate pairs:

Bivariate Analysis

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?

####1. Density and the Sugar:Alcohol Ratio

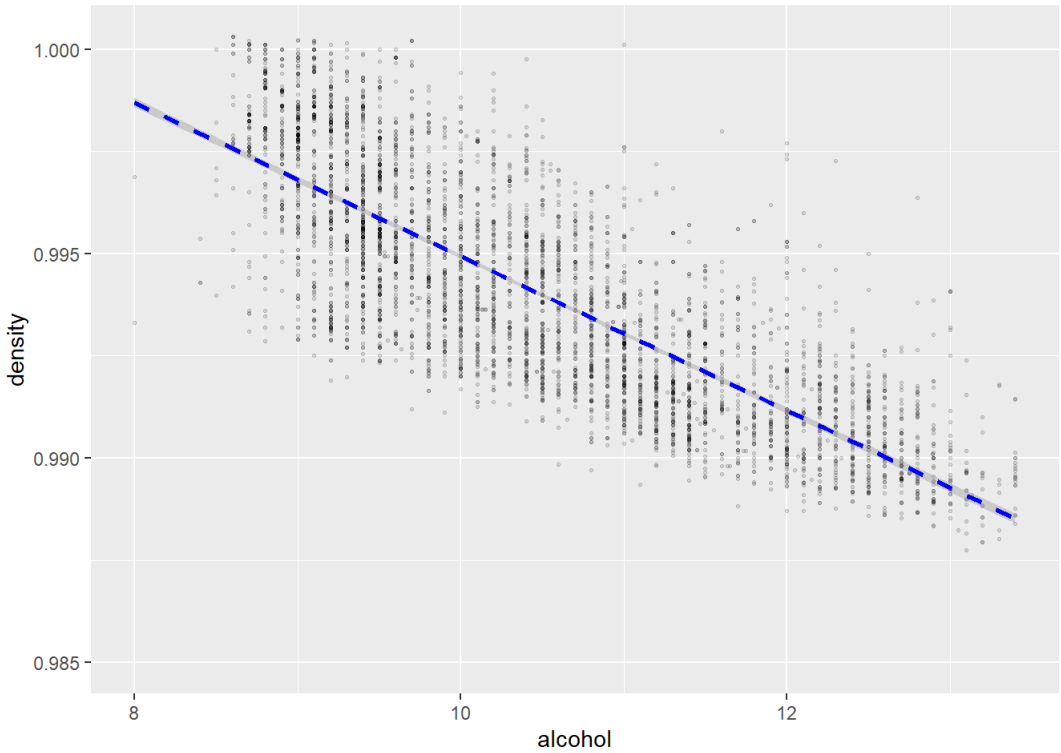
```
ggplot(aes(x=sugar_to_alcohol, y=density),
       data = data_subset) +
  geom_point(alpha = 1/8,size=0.5) +
  geom_smooth(method='lm',color = 'blue',linetype=2) +
  xlim(0,quantile(data_subset$sugar_to_alcohol,0.99)) +
  ylim(0.985,quantile(data_subset$density,0.99))
```



the relationship between density and the sugar:alcohol ratio is strong positive linear approximation appears .

####2. Density and Alcohol Content: Density was also observed to have a strong inverse correlation with the alcohol content (-0.78). Let's consider a graph of these two variables:

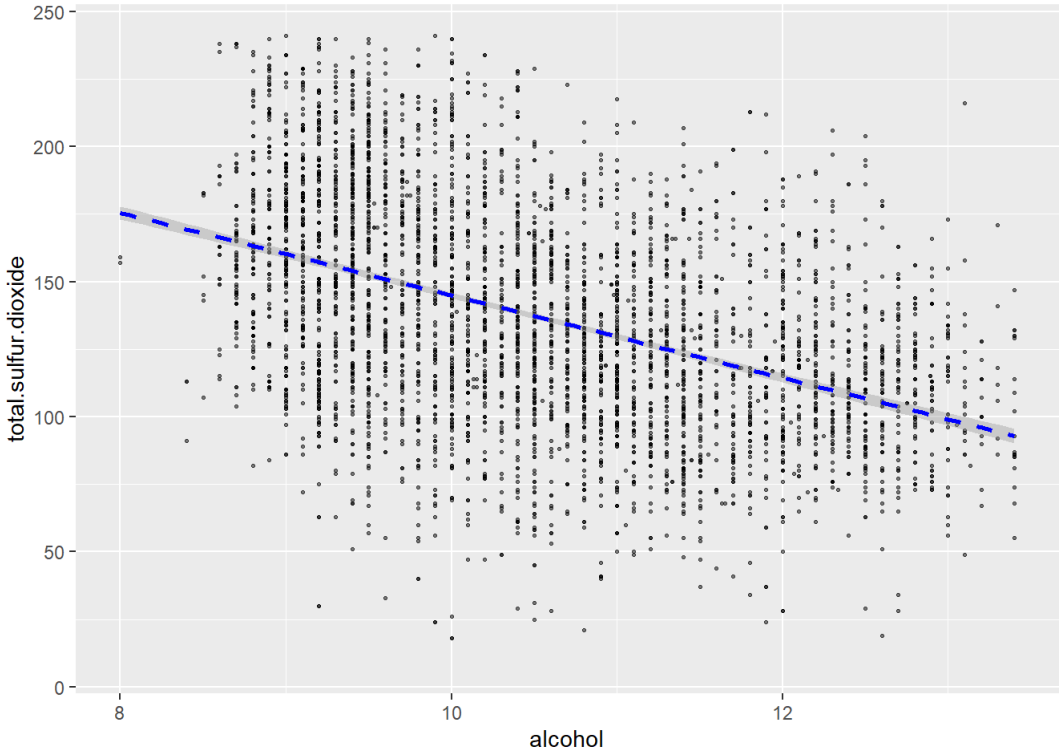
```
ggplot(aes(x=alcohol, y=density),
       data = data_subset) +
  geom_point(alpha = 1/8,size=0.5) +
  geom_smooth(method='lm',color = 'blue',linetype=2) +
  xlim(8,quantile(data_subset$alcohol,0.99)) +
  ylim(0.985,quantile(data_subset$density,0.99))
```



This inverse relationship strong negative linear relationship

####3. Total Sulfur Dioxide and Alcohol Level:

```
ggplot(aes(x=alcohol, y=total.sulfur.dioxide),
       data = data_subset) +
  geom_point(alpha = 1/2,size=0.5) +
  geom_smooth(method='lm',color = 'blue',linetype=2) +
  xlim(min(data_subset$alcohol),quantile(data_subset$alcohol,0.99)) +
  ylim(min(data_subset$total.sulfur.dioxide),
       quantile(data_subset$total.sulfur.dioxide,0.99))
```



it is a moderate negative linear relationship

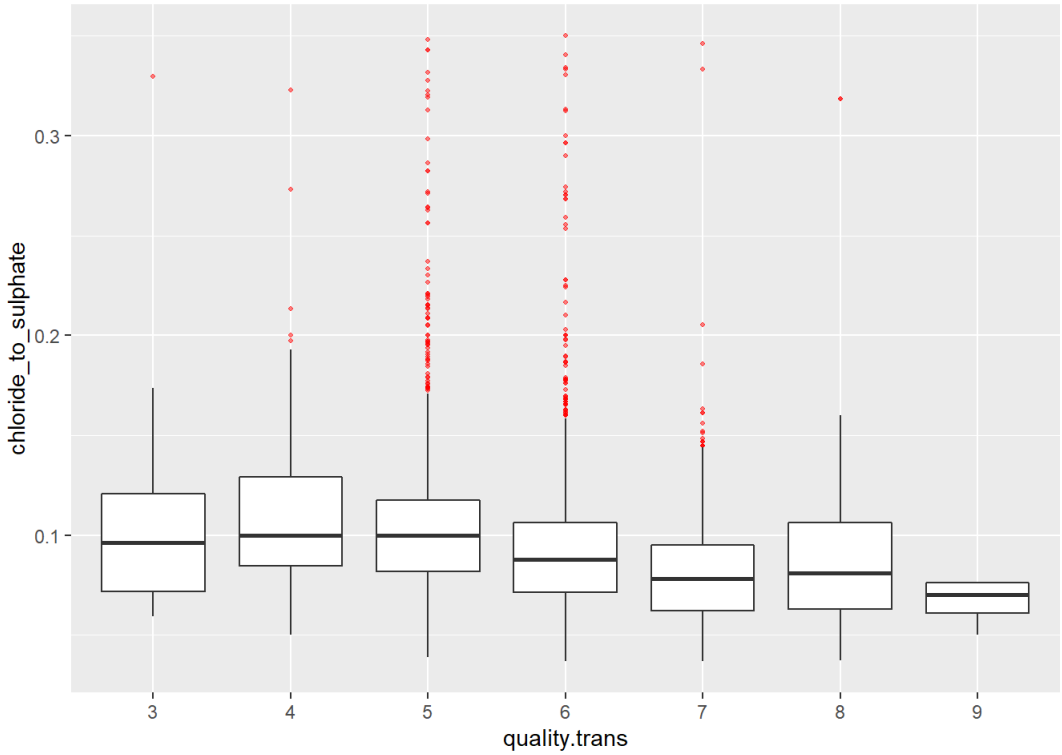
Next, lets consider the relationship between the quality measurement and various parameters.

Did you observe any interesting relationships between the other features

(not the main feature(s) of interest)? yes, as shown below…

4a. Quality and the Chloride:Sulphate Ratio

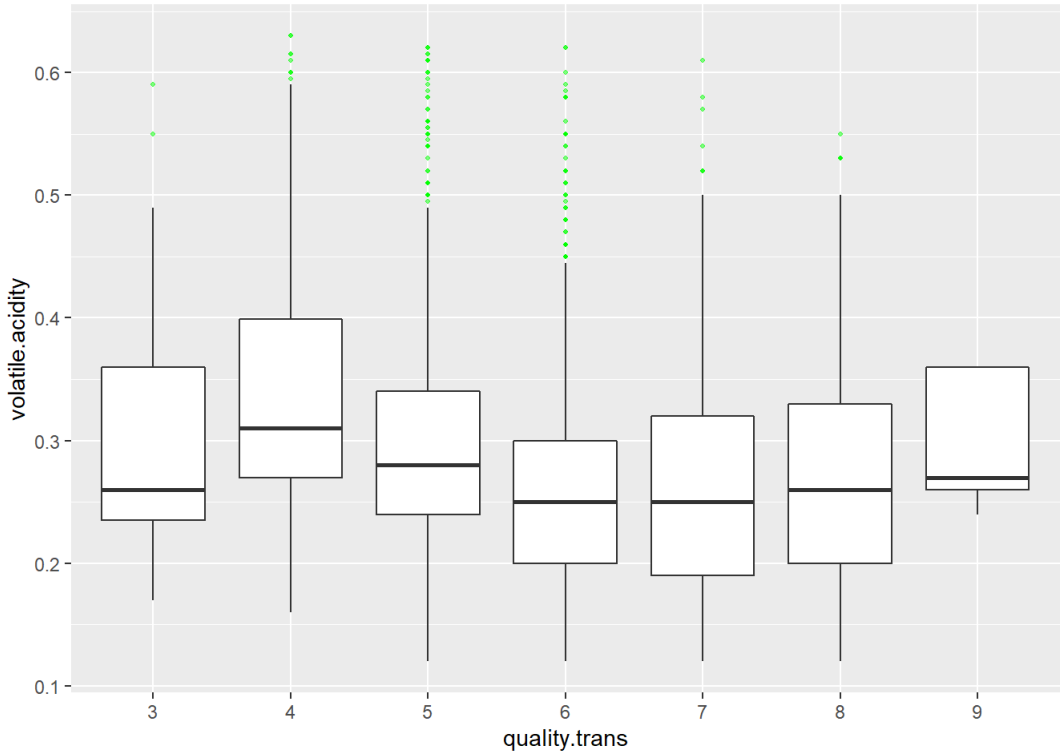
```
ggplot(aes(x=quality.trans, y=chloride_to_sulphate),
       data = data_subset) +
  geom_boxplot(outlier.alpha = 0.5,outlier.color = 'red',outlier.size = 0.75) +
  ylim(quantile(data_subset$chloride_to_sulphate,0.01),
       quantile(data_subset$chloride_to_sulphate,0.99))
```



It appears that in general, higher quality wines have lower chloride:sulphate ratios

4b. Quality and Volatile Acidity

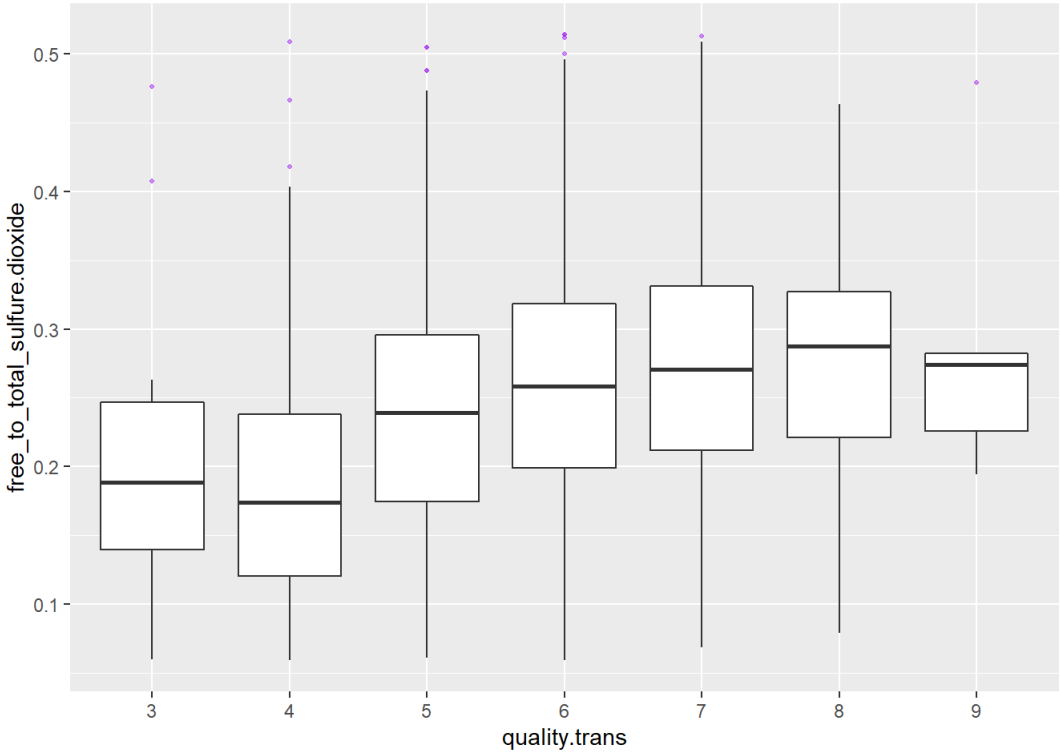
```
ggplot(aes(x=quality.trans, y=volatile.acidity),
       data = data_subset) +
  geom_boxplot(outlier.alpha= 0.5,outlier.color= 'green',outlier.size = 0.75) +
  ylim(quantile(data_subset$volatile.acidity,0.01),
       quantile(data_subset$volatile.acidity,0.99))
```



There does not appear to be any particularly promising trend between the volatile acidity level and quality, since the median volatile acidity rises and falls with changes in the quality category, with no apparent trend.

4c. Quality and the Free:Total Sulfur Dioxide Ratio

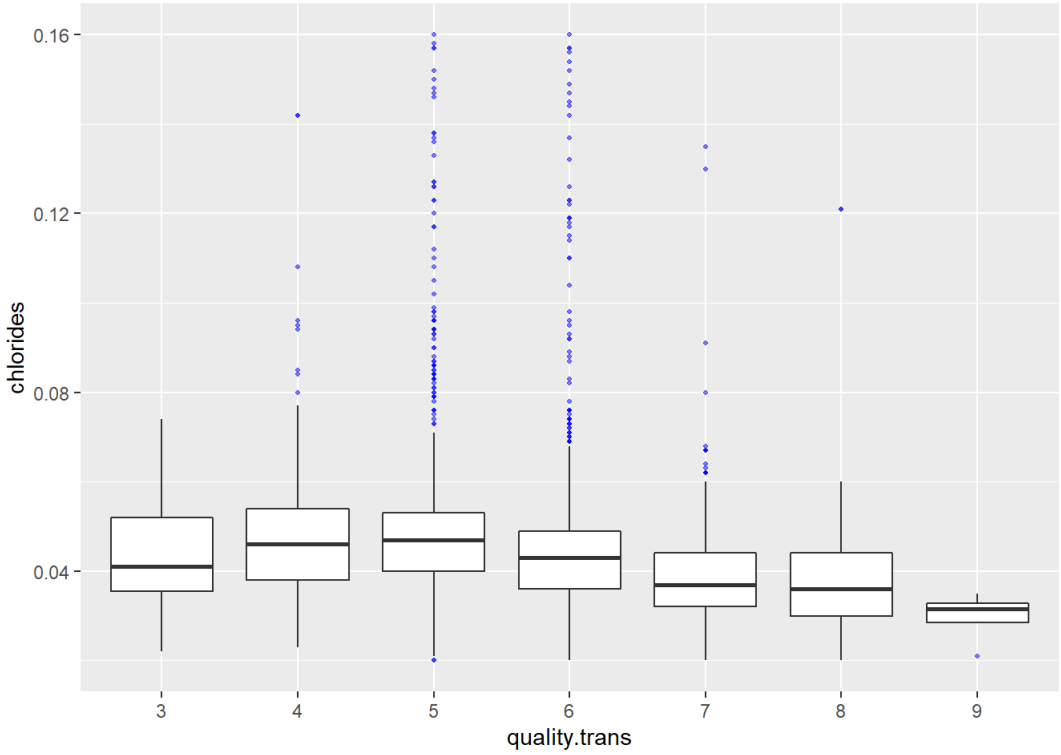
```
ggplot(aes(x=quality.trans, y=free_to_total_sulfure.dioxide),
       data = data_subset) +
  geom_boxplot(outlier.alpha=0.5,outlier.color='purple',outlier.size = 0.75)  +
  ylim(quantile(data_subset$free_to_total_sulfure.dioxide,0.01),
       quantile(data_subset$free_to_total_sulfure.dioxide,0.99))
```



It appears that in general, higher quality wines have higher free:total sulfur dioxide ratios, since the median values appear to consistently increase as quality increases.

4d. Quality and Chloride Level

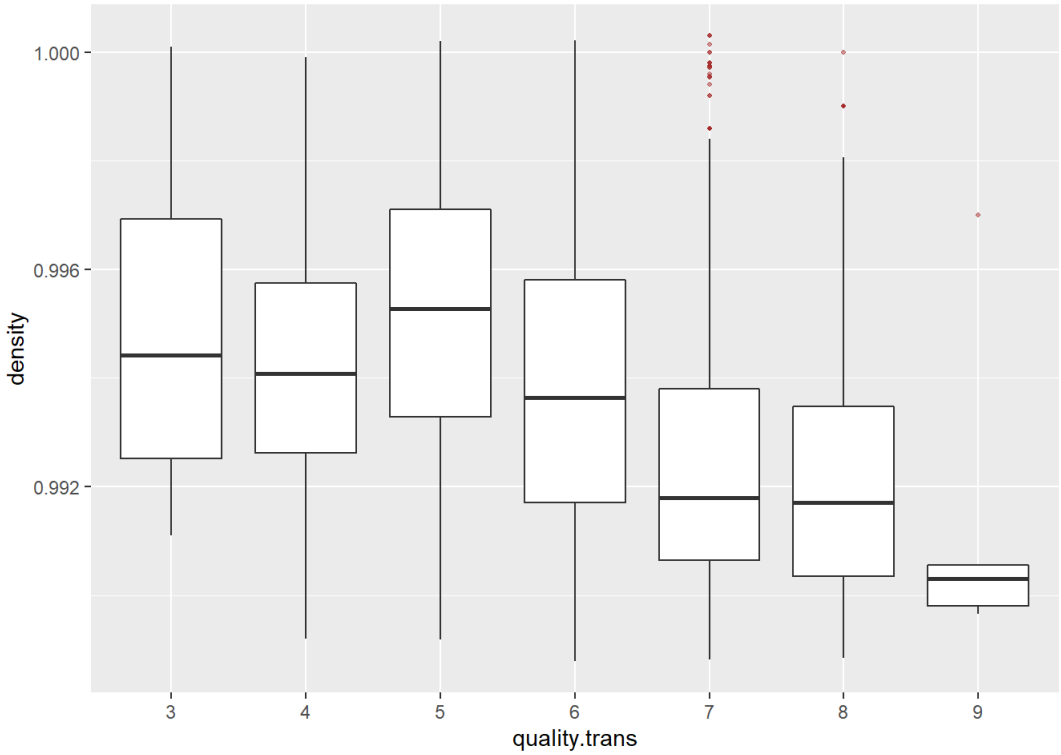
```
ggplot(aes(x=quality.trans, y=chlorides),
       data = data_subset) +
  geom_boxplot(outlier.alpha= 0.5,outlier.color= 'blue',outlier.size = 0.75)  +
  ylim(quantile(data_subset$chlorides,0.01),
       quantile(data_subset$chlorides,0.99))
```



It appears that in general, higher quality wines have lower chloride levels, since the median value of chlorides drops with increasing quality.

4e. Quality and Density

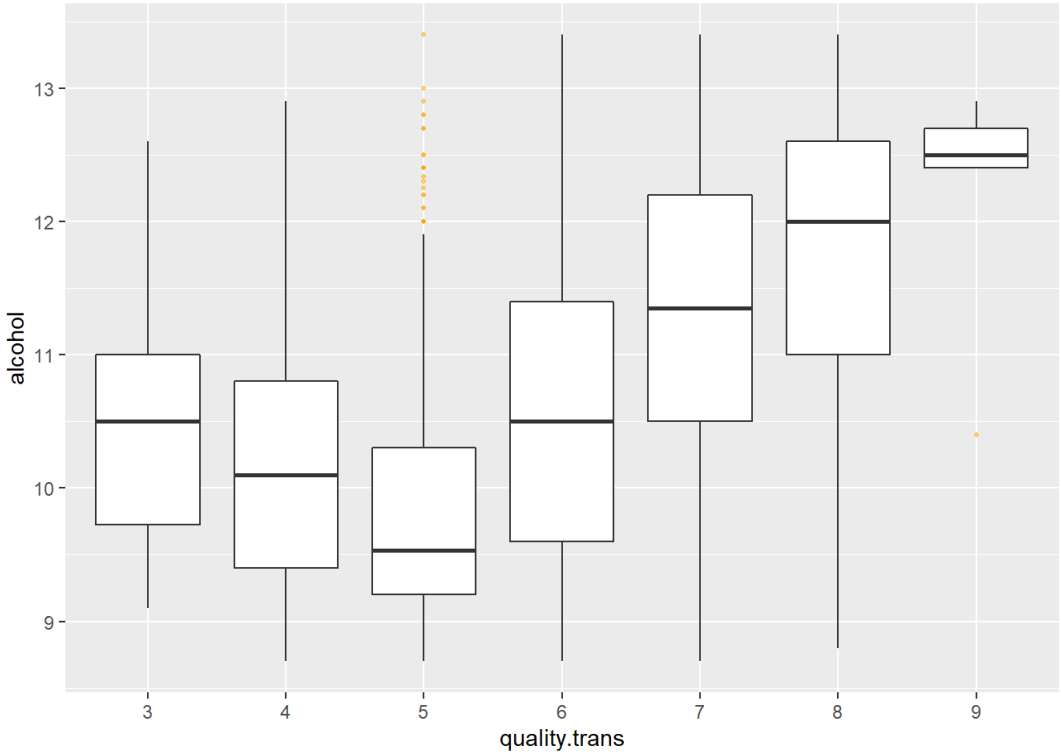
```
ggplot(aes(x=quality.trans, y=density),
       data = data_subset) +
  geom_boxplot(outlier.alpha=0.5,outlier.color='brown',outlier.size = 0.75) +
  ylim(quantile(data_subset$density,0.01),quantile(data_subset$density,0.99))
```



The relationship between density and quality appears to be quite strong: higher quality wines (quality rating of 7 or higher) appear to be lower density compared to lower quality wines (quality rating of 5 or lower), based on the large differences in the median density observed between the quality extremes.

4f. Quality and Alcohol Content

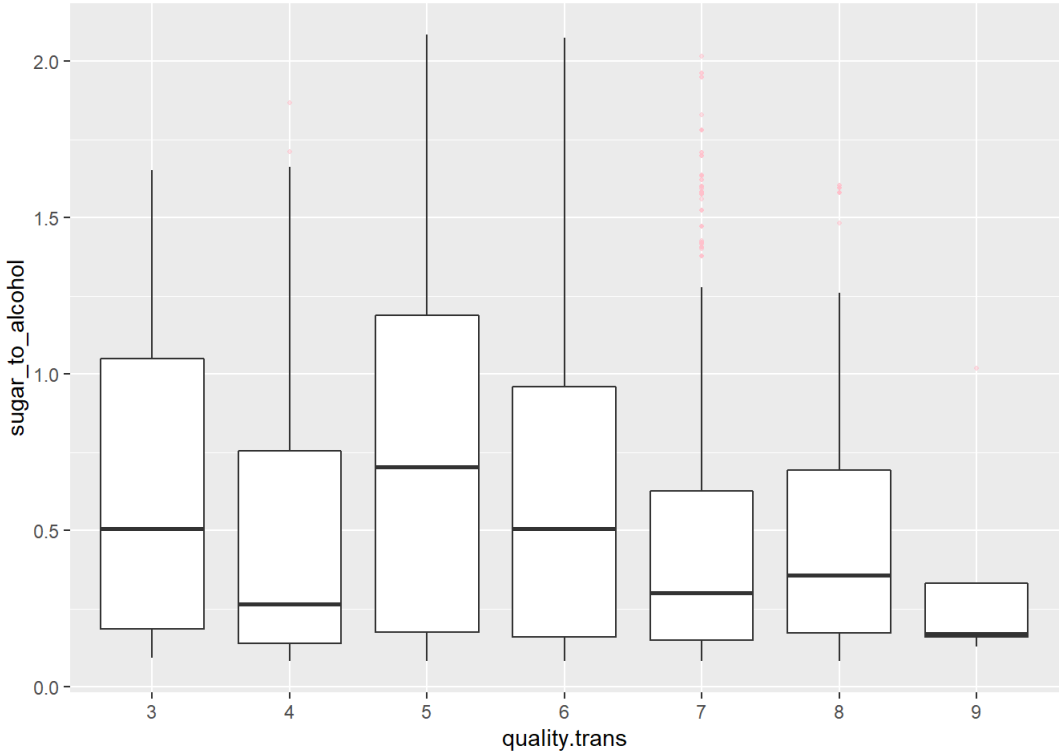
```
ggplot(aes(x=quality.trans, y=alcohol),
       data = data_subset) +
  geom_boxplot(outlier.alpha=0.5,outlier.color= 'orange',outlier.size = 0.75) +
  ylim(quantile(data_subset$alcohol,0.01),quantile(data_subset$alcohol,0.99))
```



The relationship between alcohol content and quality appears potentially promising, particularly at the higher end of the quality scale, where there is a clear upwards trend in quality (from levels 5 through 9) as the median alcohol content increases.

4g. Quality and Sugar:Alcohol Ratio

```
ggplot(aes(x=quality.trans, y=sugar_to_alcohol),
       data = data_subset) +
  geom_boxplot(outlier.alpha= 0.5,outlier.color= 'pink',outlier.size = 0.75)  +
  ylim(quantile(data_subset$sugar_to_alcohol,0.01),
       quantile(data_subset$sugar_to_alcohol,0.99))
```



It is hard to get clear trend between the sugar:alcohol ratio and a wine's quality, and the median values move up and down as the quality improves.

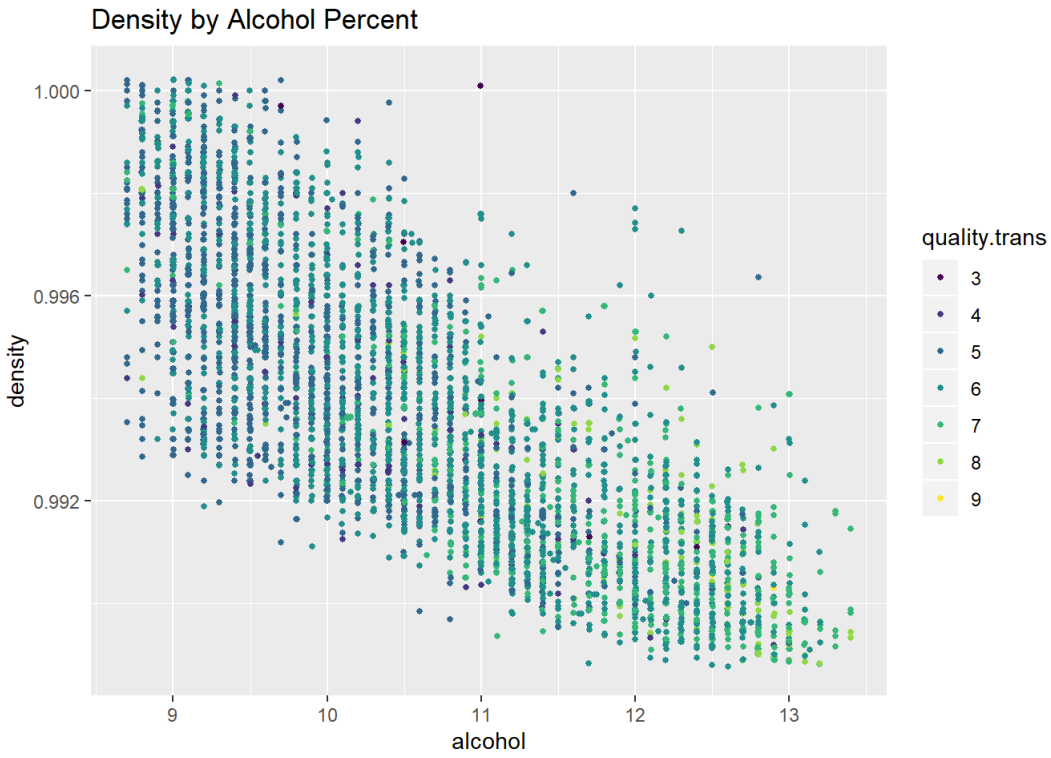
What was the strongest relationship you found?

The relationship between density and quality

Multivariate Plots Section

I will now consider the interaction of multiple variables. First, it was observed in the bivariate analysis that there is a relatively strong inverse relationship between density and the alcohol content (correlation coefficient of -0.78). The quality levels can be layered onto that graph as well:

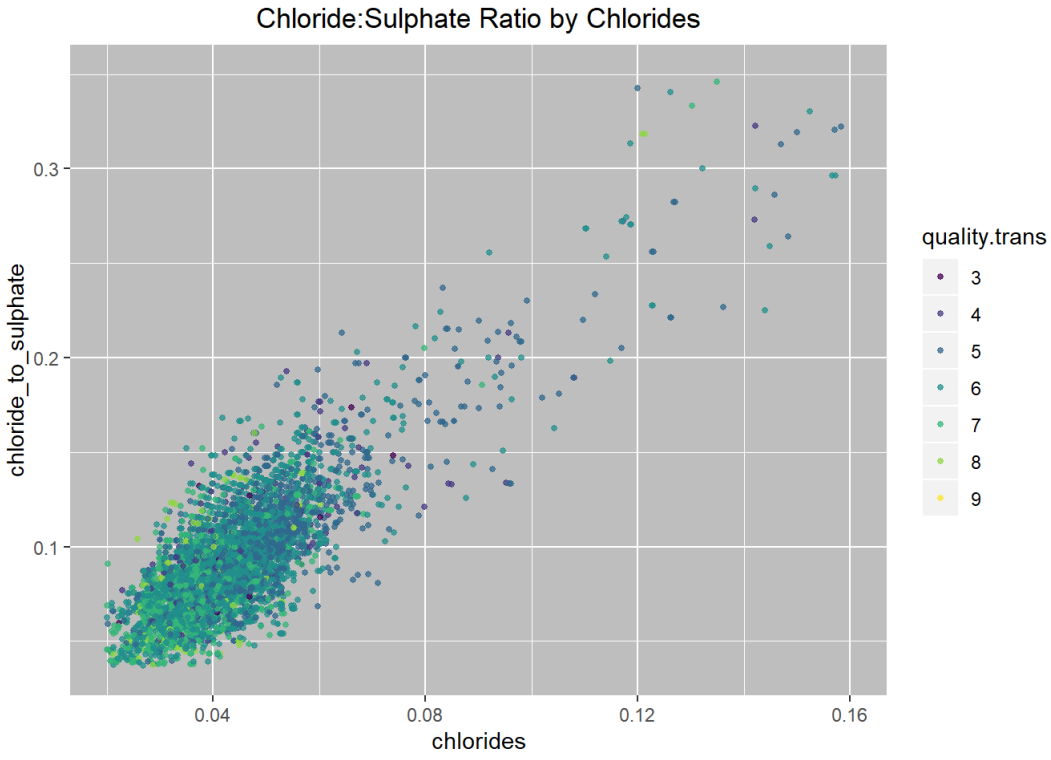
```
ggplot(aes(x = alcohol, y = density,color=quality.trans), data = data_subset) +
  geom_point(alpha = 1, size = 1, position = 'jitter') +
  xlim(quantile(data_subset$alcohol,0.01),quantile(data_subset$alcohol,0.99)) +
  ylim(quantile(data_subset$density,0.01),quantile(data_subset$density,0.99)) +
  ggtitle('Density by Alcohol Percent')
```



the higher quality wines tend to have high alcohol content and also low density.

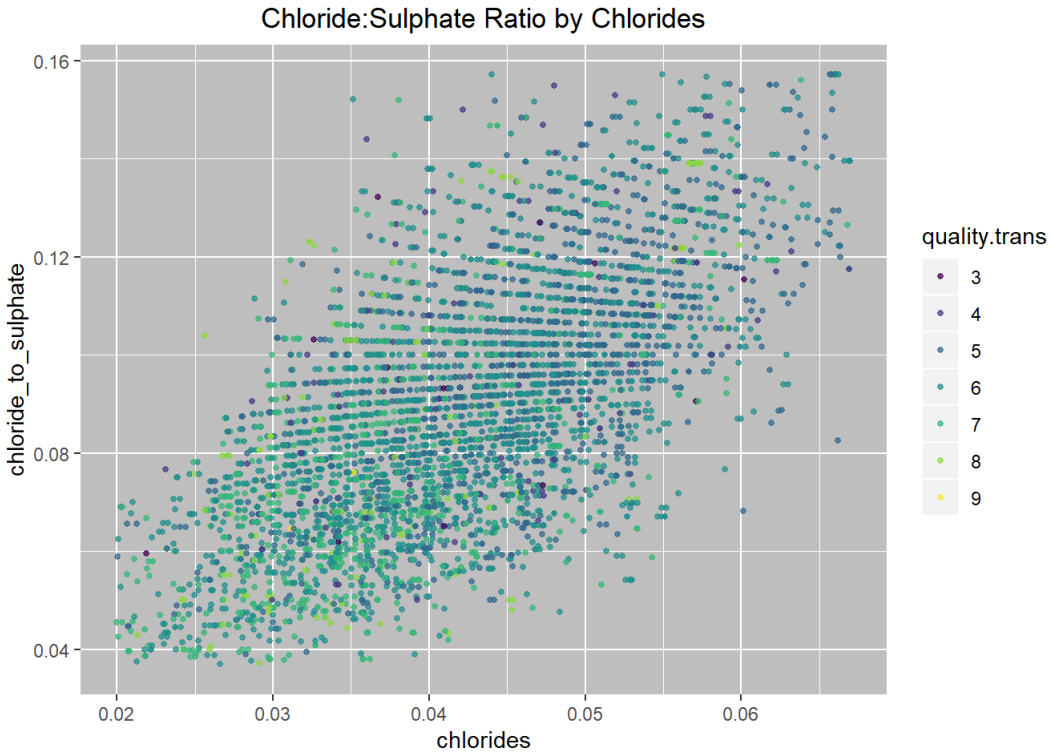
It was observed during the bivariate analysis that there was a strong correlation between the chloride level and the chloride:sulphate ratio. The quality levels can be layered onto that graph as well:

```
ggplot(aes(x = chlorides, y = chloride_to_sulphate,color=quality.trans),
       data = data_subset) +
  geom_point(alpha = 0.75, size = 1, position = 'jitter') +
  xlim(quantile(data_subset$chlorides,0.01),
       quantile(data_subset$chlorides,0.99)) +
  ylim(quantile(data_subset$chloride_to_sulphate,0.01),
       quantile(data_subset$chloride_to_sulphate,0.99)) +
  ggtitle('Chloride:Sulphate Ratio by Chlorides')+
     theme(plot.title = element_text(hjust = 0.5),panel.background = element_rect(fill = "gray"))
```



It appears there might be a tendency for high quality wines to be low chloride and low chloride:sulphate ratio. Let's zoom in on the lower left portion of the graph, which contains most of the data points, by truncating out the top 5% quantile for each variable:

```
ggplot(aes(x = chlorides, y = chloride_to_sulphate,color=quality.trans),
       data = data_subset) +
  geom_point(alpha = 0.75, size = 1, position = 'jitter') +
  xlim(quantile(data_subset$chlorides,0.01),
       quantile(data_subset$chlorides,0.95)) +
  ylim(quantile(data_subset$chloride_to_sulphate,0.01),
       quantile(data_subset$chloride_to_sulphate,0.95)) +
  ggtitle('Chloride:Sulphate Ratio by Chlorides') +
     theme(plot.title = element_text(hjust = 0.5),panel.background = element_rect(fill = "gray"))
```



There does indeed appear to be a tendency for the higher quality wines to be lower in chlorides and chloride:sulphate ratio.

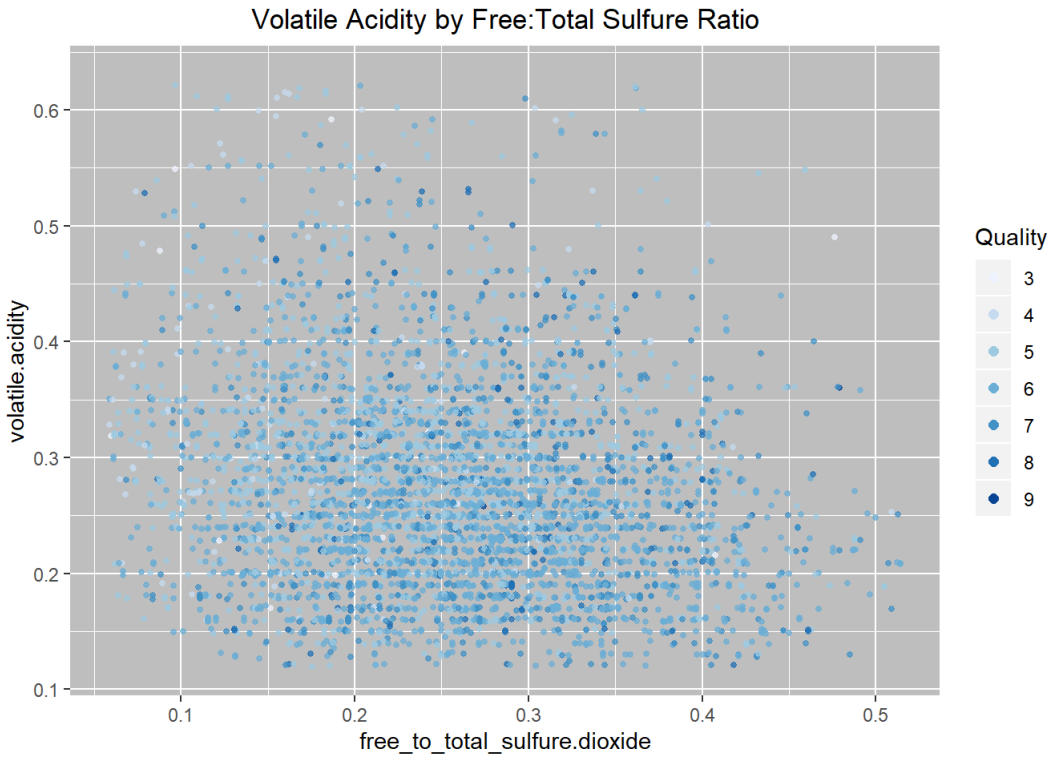
Multivariate Analysis

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?

Next, lets consider volatile acidity and the free:total sulfure dioxide ratio. During the bivariate analysis, this variable pair was observed to have one of the strongest correlations with quality (-0.19), so it seems worth considering in a multivariate format too, where quality is layered on the graph:

```
ggplot(aes(x = free_to_total_sulfure.dioxide,
           y = volatile.acidity,color=quality.trans), data = data_subset) +
  geom_point(alpha = 0.75, size = 1, position = 'jitter') +
  scale_color_brewer(type = 'seq',
    guide = guide_legend(title = 'Quality', reverse = F,
    override.aes = list(alpha = 1, size = 2))) +
  xlim(quantile(data_subset$free_to_total_sulfure.dioxide,0.01),
       quantile(data_subset$free_to_total_sulfure.dioxide,0.99)) +
  ylim(quantile(data_subset$volatile.acidity,0.01),
       quantile(data_subset$volatile.acidity,0.99)) +
  ggtitle('Volatile Acidity by Free:Total Sulfure Ratio') +
     theme(plot.title = element_text(hjust = 0.5),panel.background = element_rect(fill = "gray"))
```

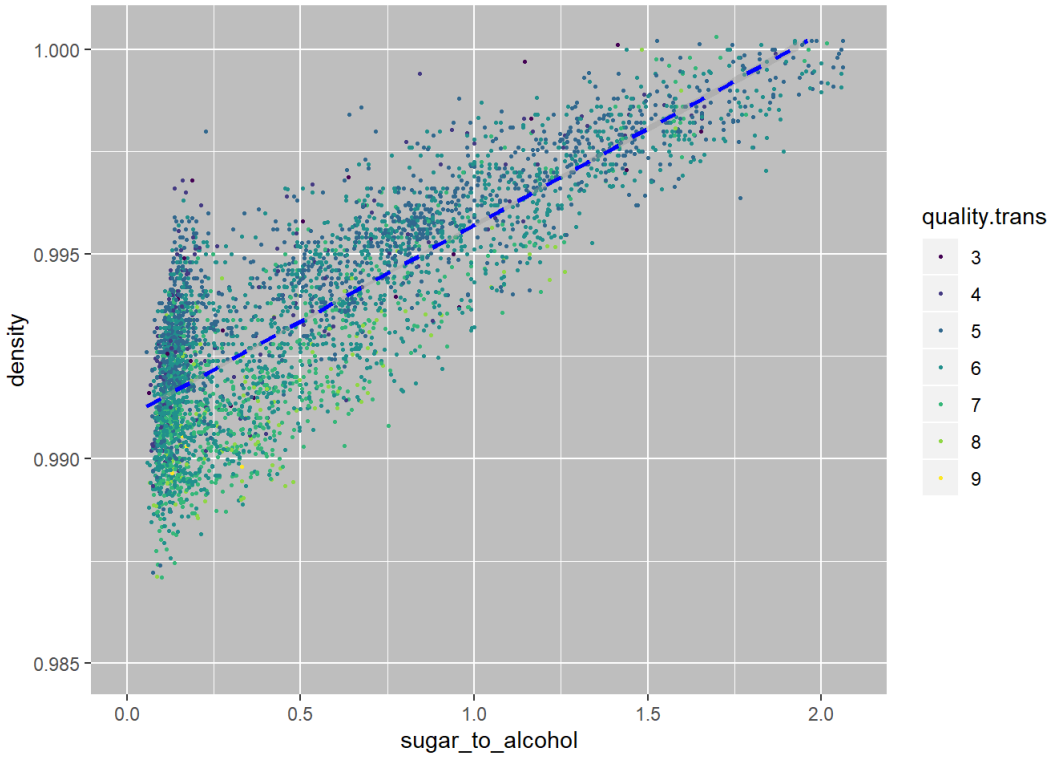


there is no strong pattern regarding where the higher versus lower quality wines fall on the graph. The quality points are dispersed throughout, even though there might be some weak relationships in terms of where they tend to fall.

Were there any interesting or surprising interactions between features?

Now lets look at the bivariate pair that exhibited the highest correlation coefficient, namely density and the sugar:alcohol ratio, which had a correlation coefficient of 0.87. To deepen the insight into these two variables and how they might impact wine quality, lets layer quality onto the graph:

```
ggplot(aes(x=sugar_to_alcohol, y=density,color = quality.trans),
       data = data_subset) +
  geom_point(alpha = 1,size=0.5) +
  geom_smooth(method='lm',color = 'blue',linetype=2) +
  xlim(0,quantile(data_subset$sugar_to_alcohol,0.99)) +
  ylim(0.985,quantile(data_subset$density,0.99)) +
     theme(panel.background = element_rect(fill = "gray"))
```



A very interesting graph results, where there appears to be a strong tendency for the higher quality wines to cluster below the trendline

whereas the lower quality wines tend to cluster above the trendline. In other words, for a given sugar:alcohol ratio, higher quality wines tend to be less dense, and above a certain sugar:alcohol ratio (approximately 1.5), there appear to be very few good quality wines.

final Data Transformation

It was observed at the very beginning of the analysis that one drawback of this data set is the relatively small number of samples for wines at the extreme ends of the quality spectrum. For example, of the nearly 5,000 wines in the dataset, there were zero wines of qualities 0,1,2 or 10. There were only 20 wines of quality 3 and only 5 wines of quality 9. Given the tiny number of samples on the extremes of the quality spectrum, it is possible that the dataset is being partitioned too finely. This seems particularly possible given that 'quality' is ultimately an expert's judgement call rather than an easy-to-measure number, so one might expect a legitimate quality level 7 wine to be tagged as a 6 or an 8, depending on which expert makes the judgement.

To address this, I would like to consider how things might look if the quality categories are more 'coarse' and hence each category has many more samples. To do so, lets consider any wine with a 3-5 rating as 'bad', a wine with a 6 rating as 'ok' and a wine with a 7-9 rating as 'good'. When the data is split along these lines, one obtains the following sample count per category:

```
# create a function for bucketing the data into three new quality categories:
mytrans <- function(x) {
  if(x < 6){
    'bad'
  }
  else if(x < 7){
    'ok'
  }
  else 'good'
}
# apply the function to create a new categorical variable, 'good_bad':
data_subset$good_bad <- ordered(
  sapply(data_subset$quality,mytrans),levels=c('bad','ok','good'))
summary(data_subset$good_bad)
```

```
##  bad   ok good
## 1640 2198 1060
```

so now we have categorical quality data

OPTIONAL: Did you create any models with your dataset? Discuss the

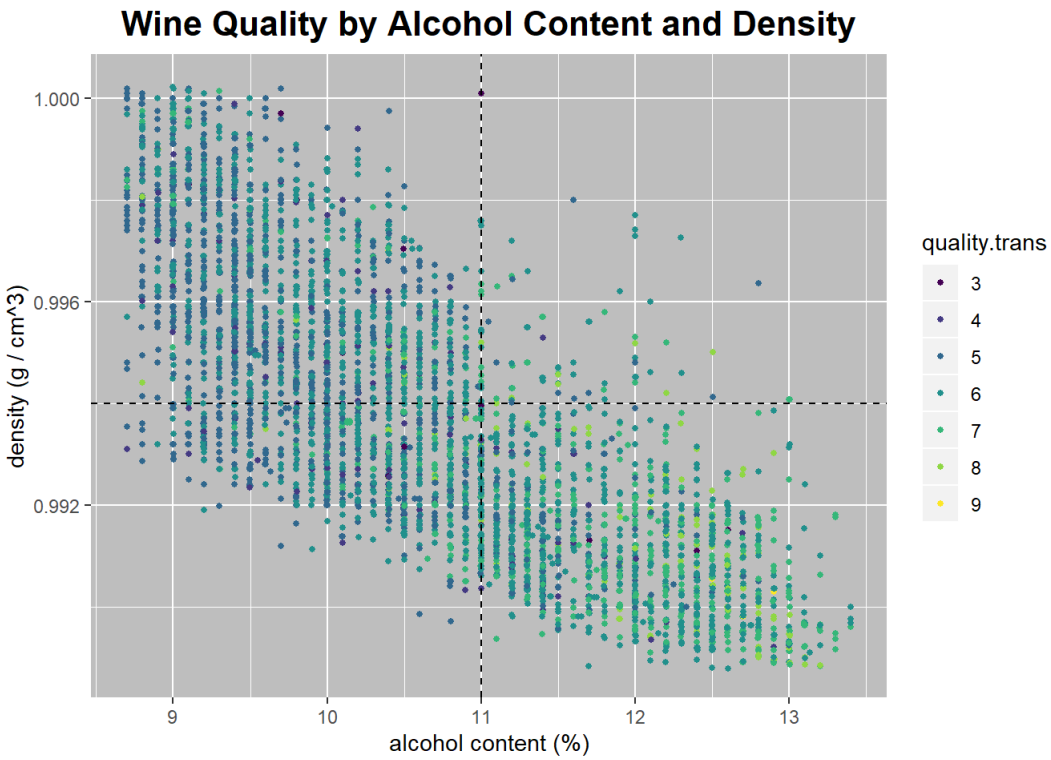
###strengths and limitations of your model. NO ——

Final Plots and Summary

In this section, three particularly interesting graphs that help summarize the key findings from the EDA are presented.

Plot One

```
ggplot(aes(x = alcohol, y = density,color=quality.trans), data = data_subset) +
  geom_point(alpha = 1, size = 1, position = 'jitter') +
  xlim(quantile(data_subset$alcohol,0.01),quantile(data_subset$alcohol,0.99)) +
  ylim(quantile(data_subset$density,0.01),quantile(data_subset$density,0.99)) +
  geom_vline(xintercept = 11,linetype=2) +
  geom_hline(yintercept = 0.994,linetype=2) +
  ggtitle('Wine Quality by Alcohol Content and Density') +
     theme(plot.title = element_text(hjust = 0.5,size = 16, face = "bold"),
           panel.background = element_rect(fill = "gray")) +
  xlab("alcohol content (%)") +
  ylab("density (g / cm^3)")
```

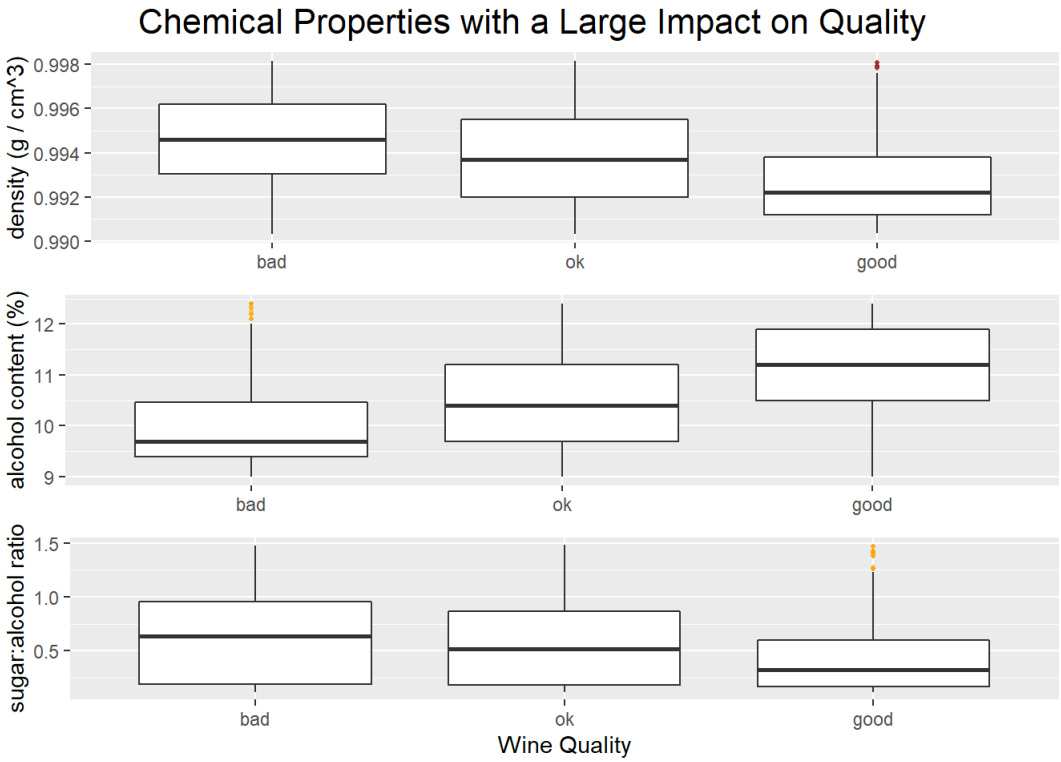


Description One

This plot demonstrates that in general, the high quality wines (quality 7-9) tend to have high alcohol content and low density, Conversely, the poor quality wines (quality 3-5) tend to have low alcohol content and high density, dominating the two left side quadrants.

Plot Two

```
# graph three variables of interest vs 'bad','ok','good' quality categories:
p24 <- ggplot(aes(x=good_bad, y=density),
       data = data_subset) +
  geom_boxplot(outlier.alpha = 0.5,outlier.color ='brown',outlier.size= 0.75) +
  ylim(quantile(data_subset$density,0.1),quantile(data_subset$density,0.9)) +
  theme(axis.title.x=element_blank()) +
  ylab("density (g / cm^3)")

p25 <- ggplot(aes(x=good_bad, y=alcohol),
       data = data_subset) +
  geom_boxplot(outlier.alpha= 0.5,outlier.color= 'orange',outlier.size= 0.75) +
  ylim(quantile(data_subset$alcohol,0.1),quantile(data_subset$alcohol,0.9)) +
  theme(axis.title.x=element_blank()) +
  ylab("alcohol content (%)")

p26 <- ggplot(aes(x=good_bad, y=sugar_to_alcohol),
       data = data_subset) +
  geom_boxplot(outlier.alpha= 0.5,outlier.color= 'orange',outlier.size= 0.75) +
  ylim(quantile(data_subset$sugar_to_alcohol,0.1),
       quantile(data_subset$sugar_to_alcohol,0.9))  +
  xlab("Wine Quality") +
  ylab("sugar:alcohol ratio")

grid.arrange(p24,p25,p26, ncol=1, top=textGrob("Chemical Properties with a Large Impact on Quality",gp=gpar(
fontsize=16,face='bold')))
```

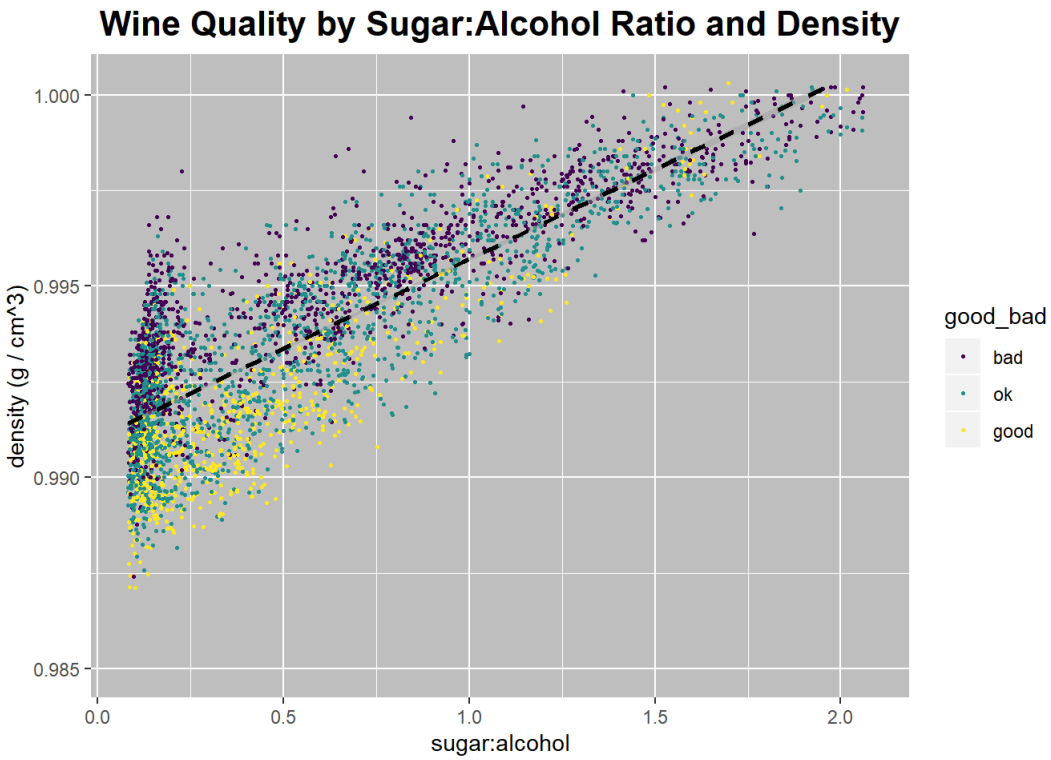


Description Two

This plot demonstrates that once wine quality is transformed into more coarse bins (i.e. ‘bad’,‘ok’ and ‘good’ instead of integers 3-9) then consistent trends emerge in the impact of various chemical properties on wine quality. Specifically, as the density and the sugar:alcohol ratio decrease, the wine quality increases and as the percent alcohol increases the wine quality increases.

Plot Three

```
ggplot(aes(x=sugar_to_alcohol, y=density,color = good_bad),
       data = data_subset) +
  geom_point(alpha = 1,size=0.5) +
  geom_smooth(method='lm',color = 'black',linetype=2) +
  xlim(quantile(data_subset$sugar_to_alcohol,0.01),
       quantile(data_subset$sugar_to_alcohol,0.99)) +
  ylim(0.985,quantile(data_subset$density,0.99)) +
  ggtitle('Wine Quality by Sugar:Alcohol Ratio and Density') +
     theme(plot.title = element_text(hjust = 0.5,size = 16, face = "bold"),
           panel.background = element_rect(fill = "gray")) +
  xlab("sugar:alcohol") +
  ylab("density (g / cm^3)")
```



Description Three

This plot summarizes the key findings from the EDA exercise: at a given sugar:alcohol level, high quality wines tend to have lower densities than low quality wines. Further, beyond a certain sugar:alcohol ratio (approximately 1.0 - 1.5) there is a preponderance of bad quality wines compared to good quality wines.

Reflection

One major struggle is to figure out the relation between each variable to another

Despite this struggle , bar plots and bar charts and scatterplots helped me to understand the relations between variables

I think the most interesting thing was the chemical analysis which was the first time i work on chemical analysis like this

I think the most interesting area for future exploration with this dataset would be to utilize machine learning techniques