# Wine quality by Prasad Pagade (1/25/2017)

==========

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

```
##        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
```

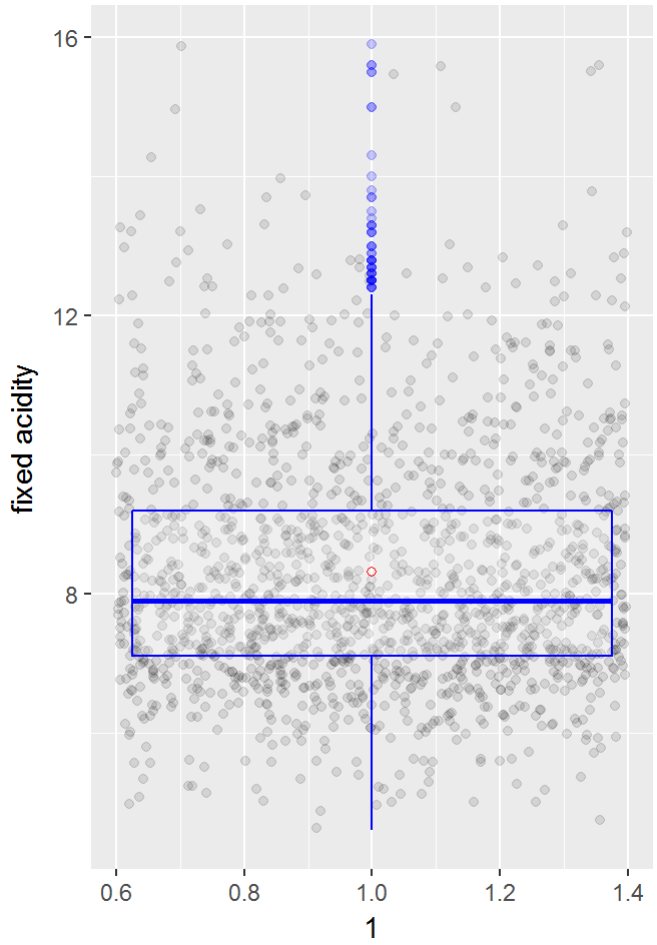
```
##   X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 1           7.4             0.70        0.00            1.9     0.076
## 2 2           7.8             0.88        0.00            2.6     0.098
## 3 3           7.8             0.76        0.04            2.3     0.092
## 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
```

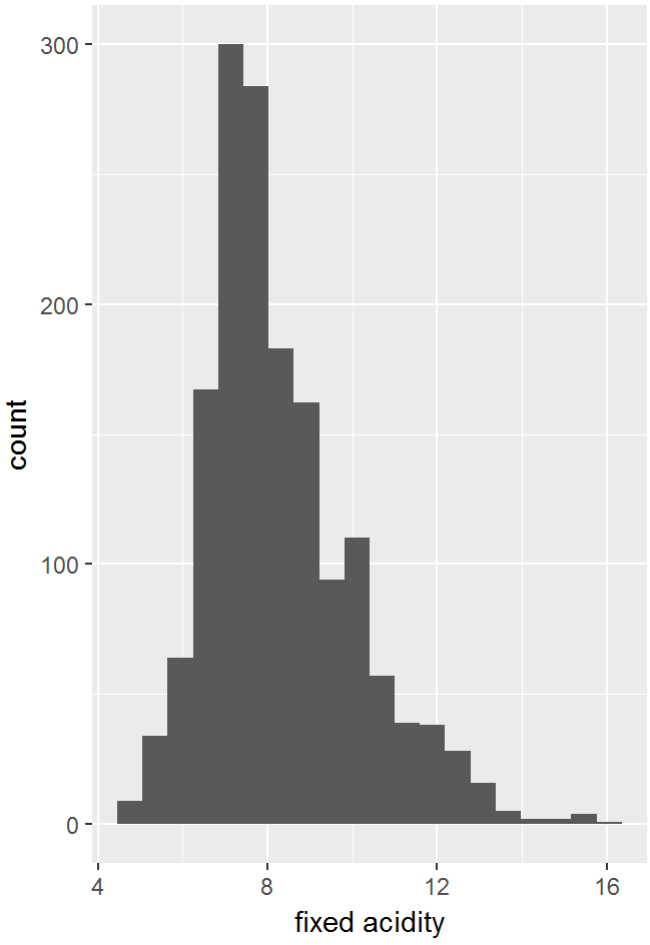
There are 1599 observations with 13 variables. Most wine quality are in the median range of 6. Observed large difference between mean and max values for variables like free.sulphur.dioxide, total.sulphur.dioxide and sugar.

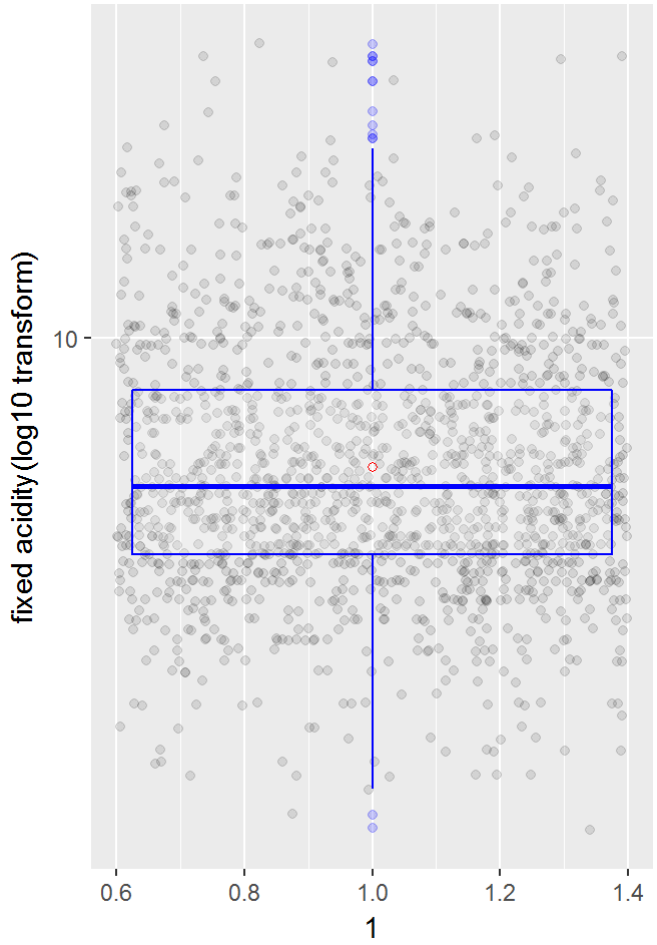
# Univariate Plots Section

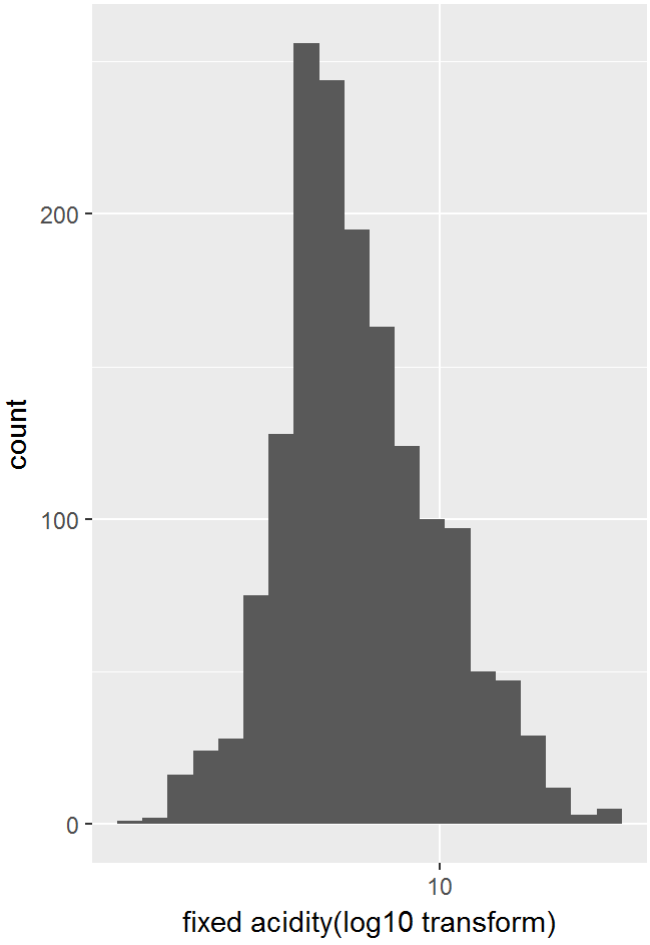
Let us take a first look at some variables by plotting them below

## 1) Fixed acidity


16
12
8
fixed acidity
0.6
0.8
1.0
1.2
1.4
1


300
200
100
0
count
4
8
12
16
fixed acidity

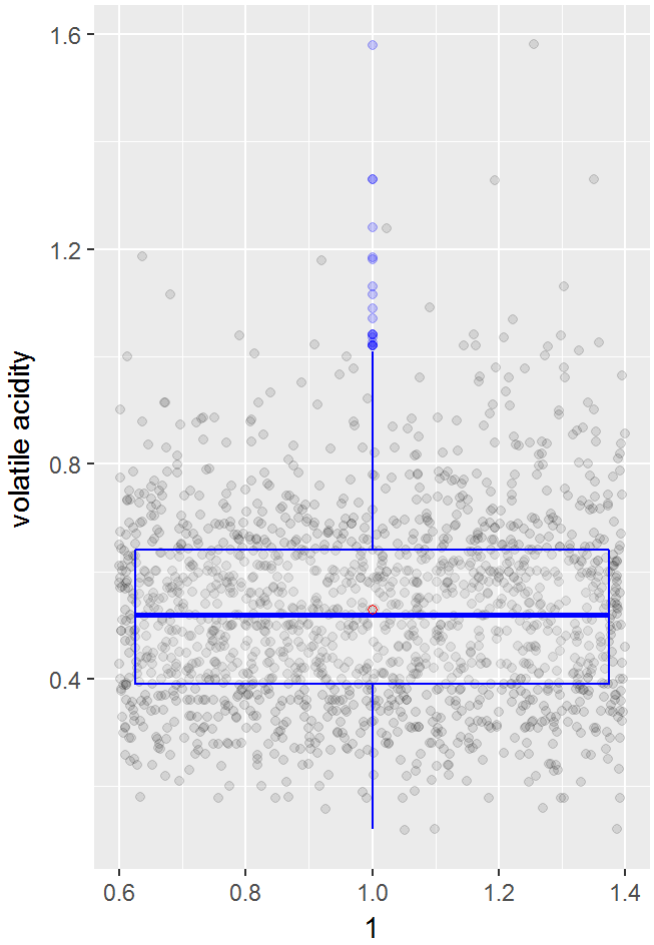

10
fixed acidity(log10 transform)
0.6
0.8
1.0
1.2
1.4
1

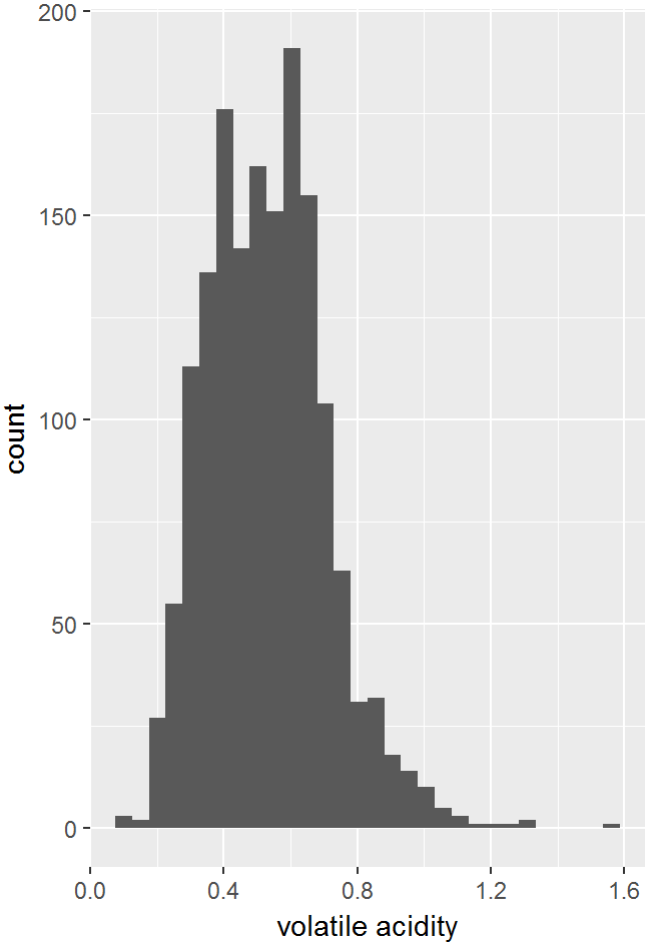

200
100
0
count
10
fixed acidity(log10 transform)

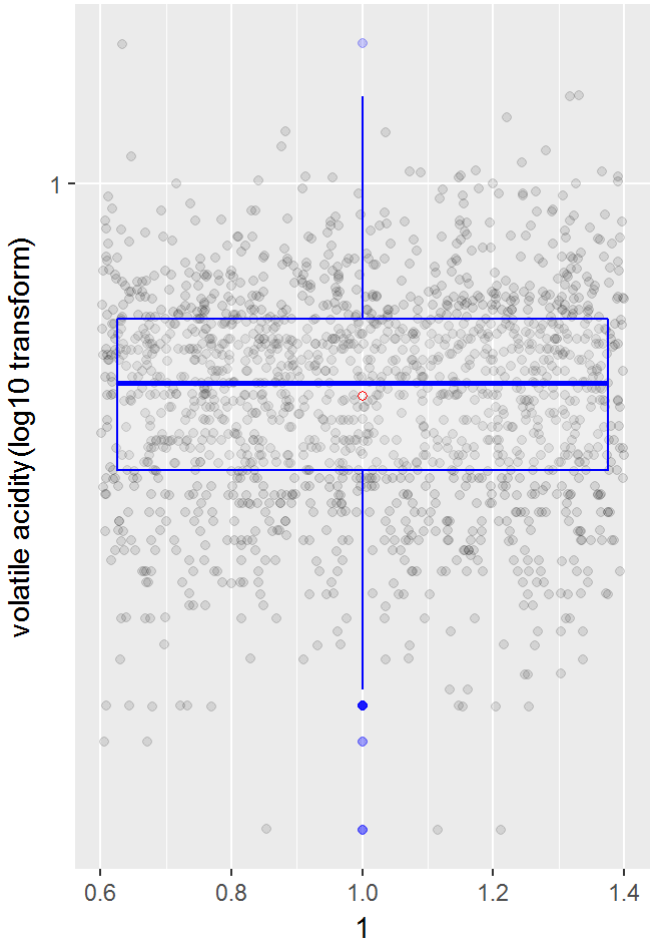
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
```

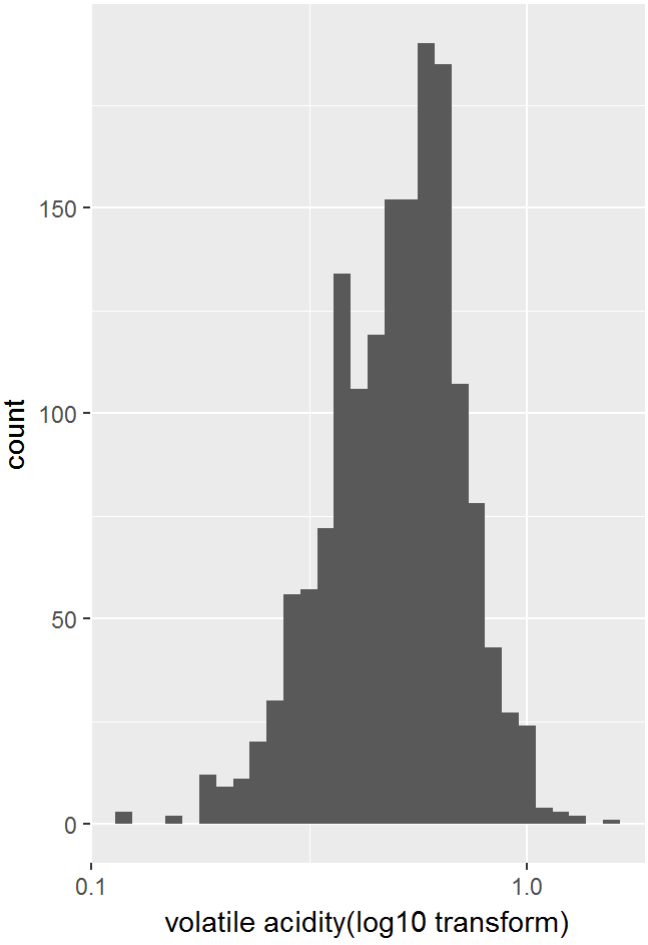
Fixed acidity is long-tailed distribution. The log transform does not reveal anything new but it normalizes the distribution.

# 2) Volatile Acidity


volatile acidity
1.6
1.2
0.8
0.4
0.6
0.8
1.0
1.2
1.4
1


count
200
150
100
50
0
0.0
0.4
0.8
1.2
1.6
volatile acidity

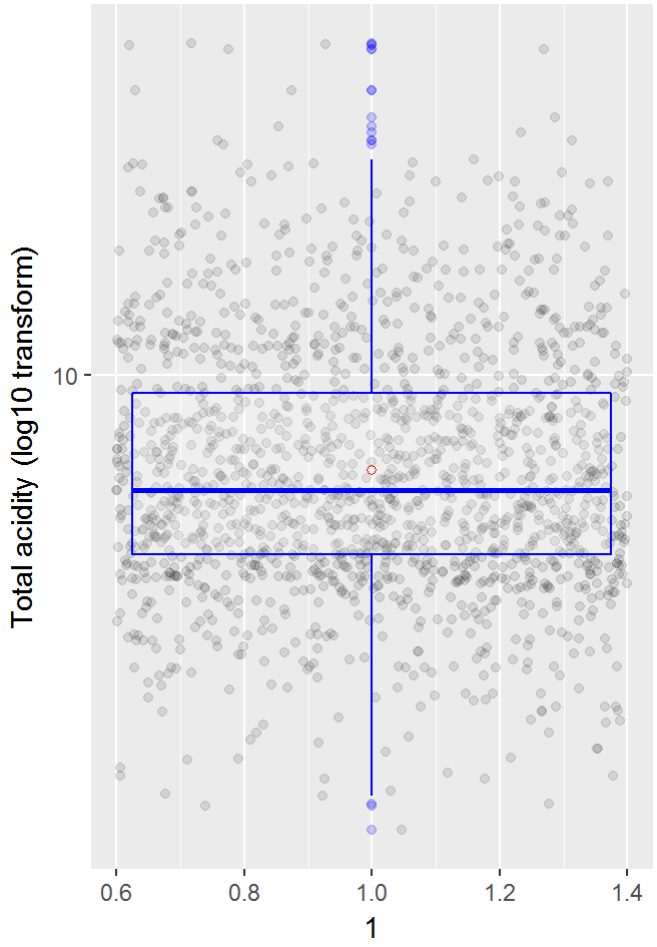

volatile acidity(log10 transform)
1
0.6
0.8
1.0
1.2
1.4
1

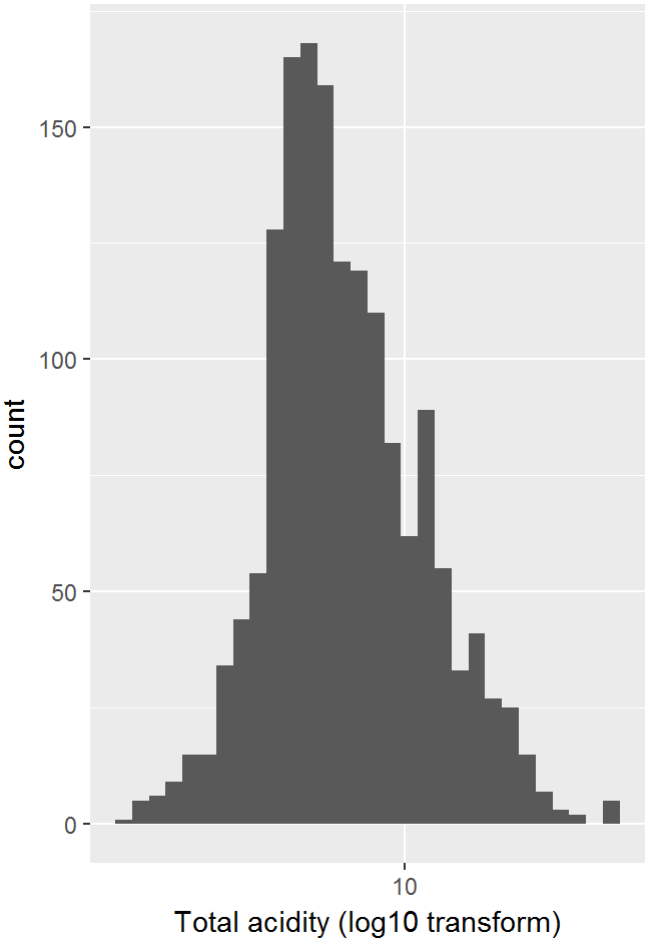

count
150
100
50
0
0.1
1.0
volatile acidity(log10 transform)

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```

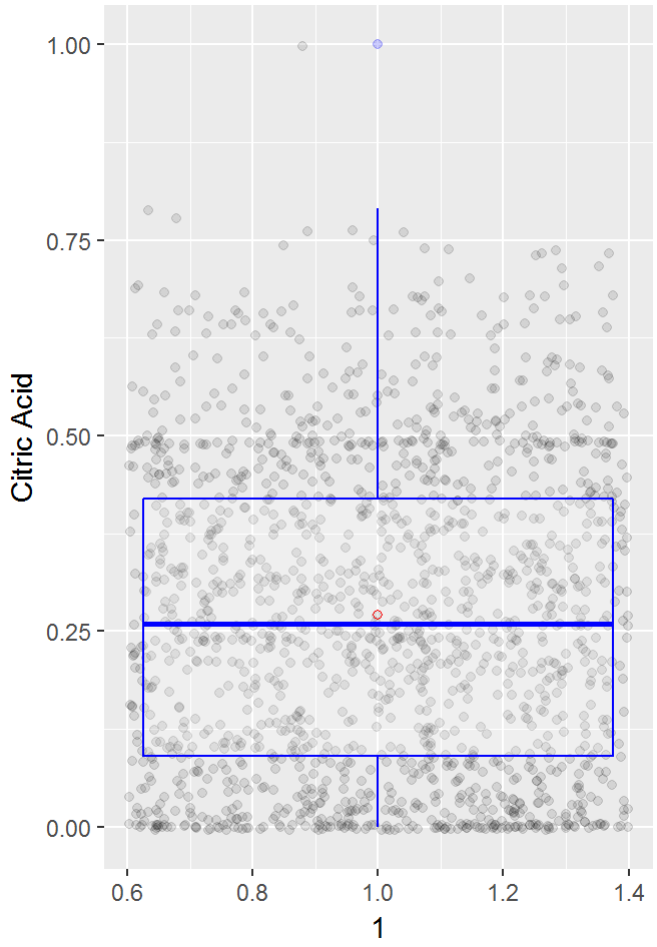
Similar to fixed acidity, volatile acidity also has a long tail distribution. However, when we look at the log transforms, we can see that the distribution looks a little binominal.

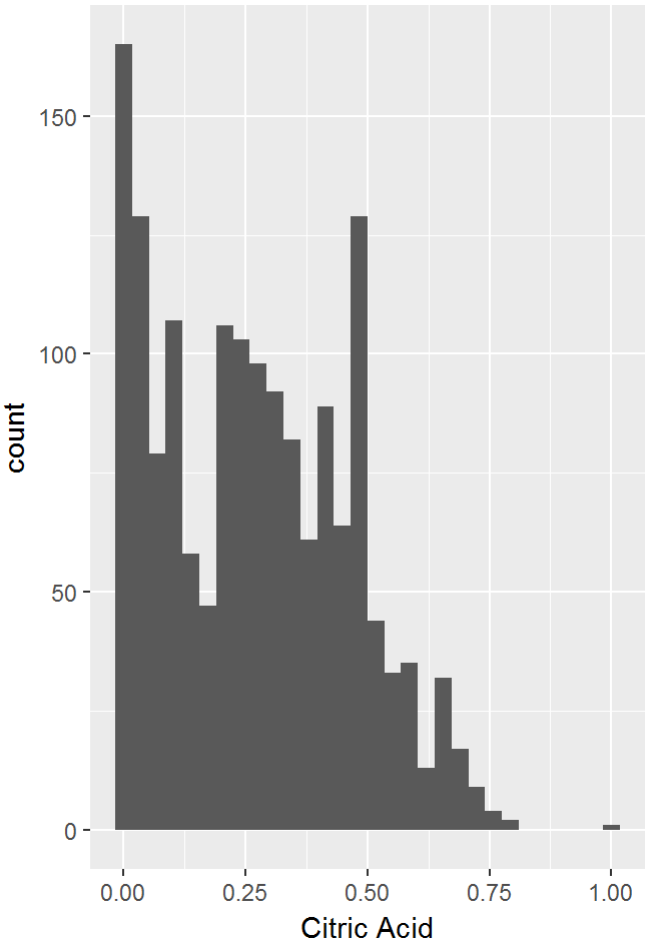
# 2.1) Total acidity (fixed acidity + volatile acidity)

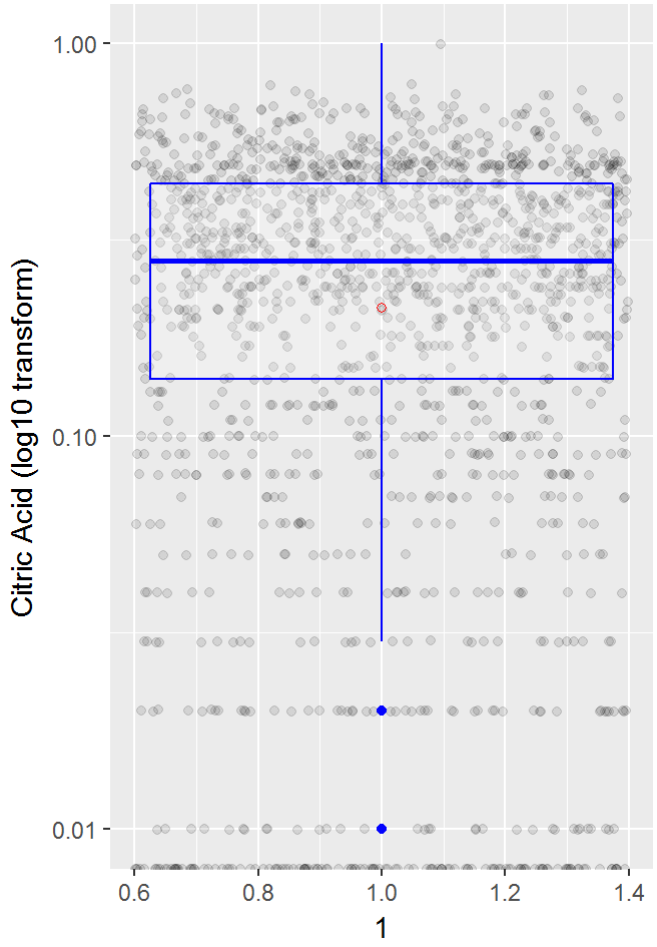


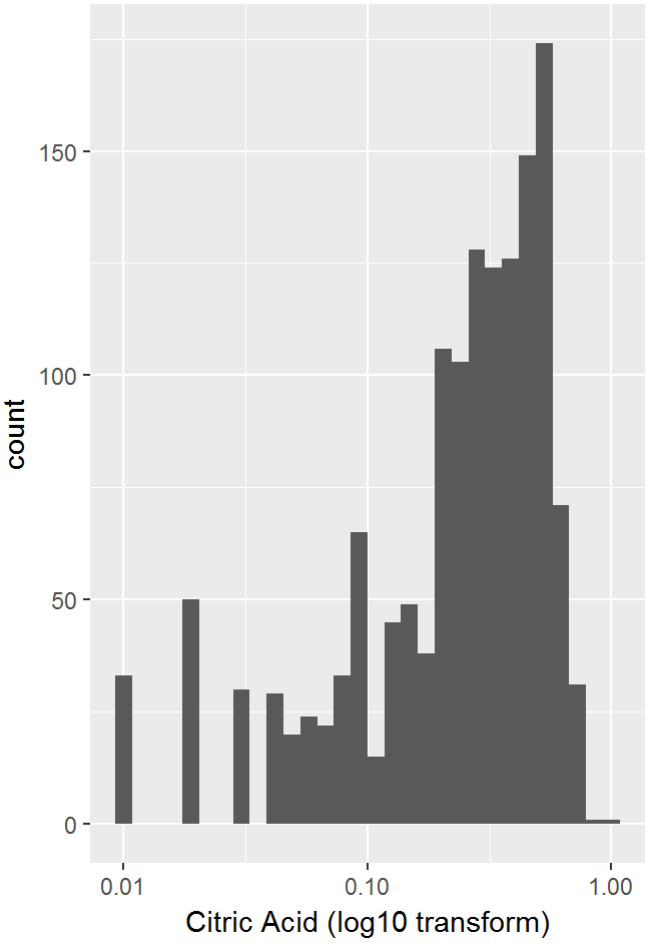




# 3) Citric Acid


1.00
0.75
0.50
0.25
0.00
Citric Acid
0.6
0.8
1.0
1.2
1.4
1

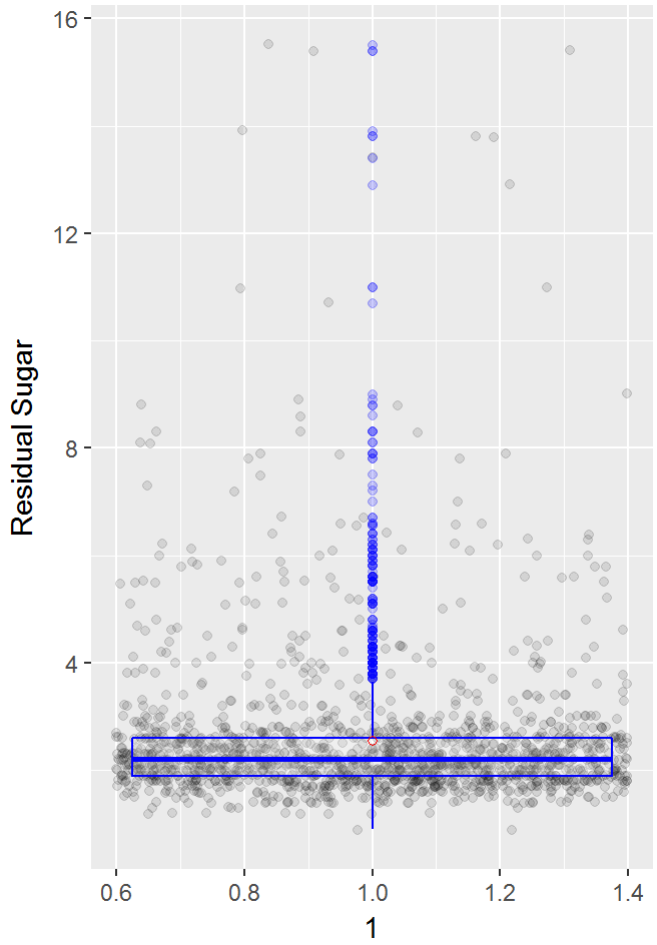

150
100
50
0
count
0.00
0.25
0.50
0.75
1.00
Citric Acid

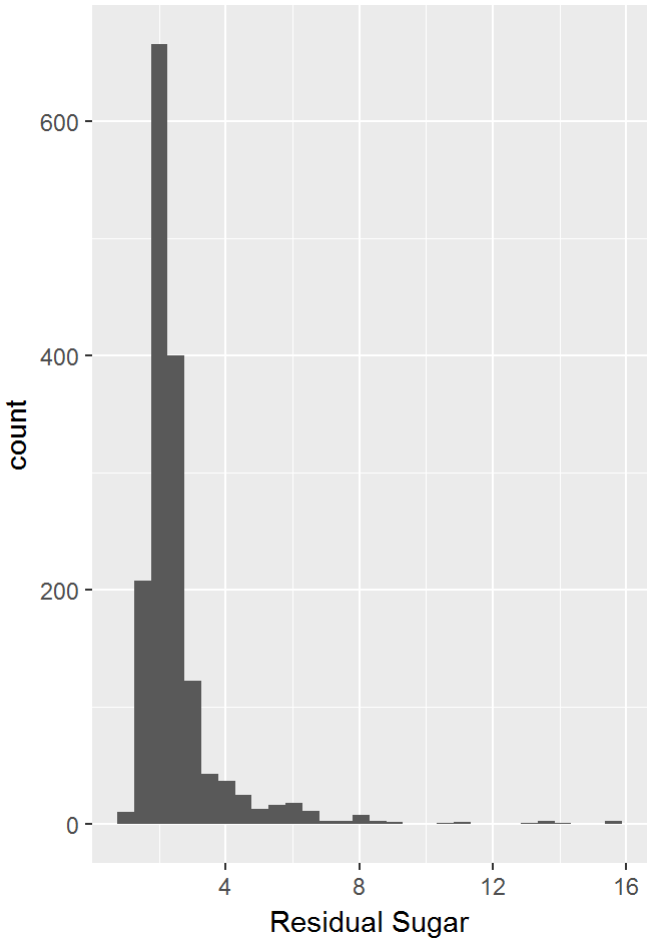

1.00
0.10
0.01
Citric Acid (log10 transform)
0.6
0.8
1.0
1.2
1.4
1

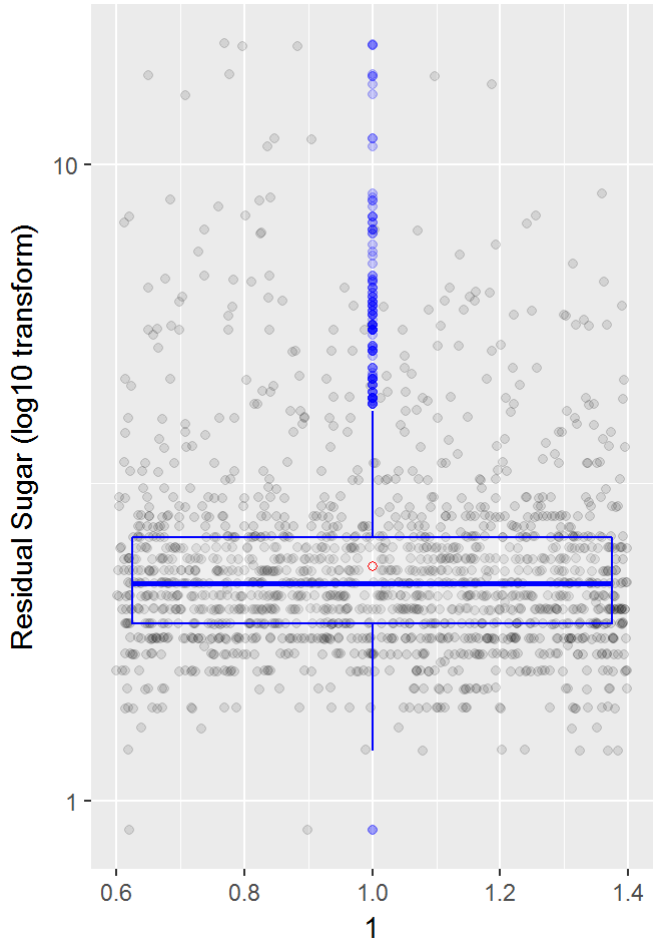

150
100
50
0
count
0.01
0.10
1.00
Citric Acid (log10 transform)

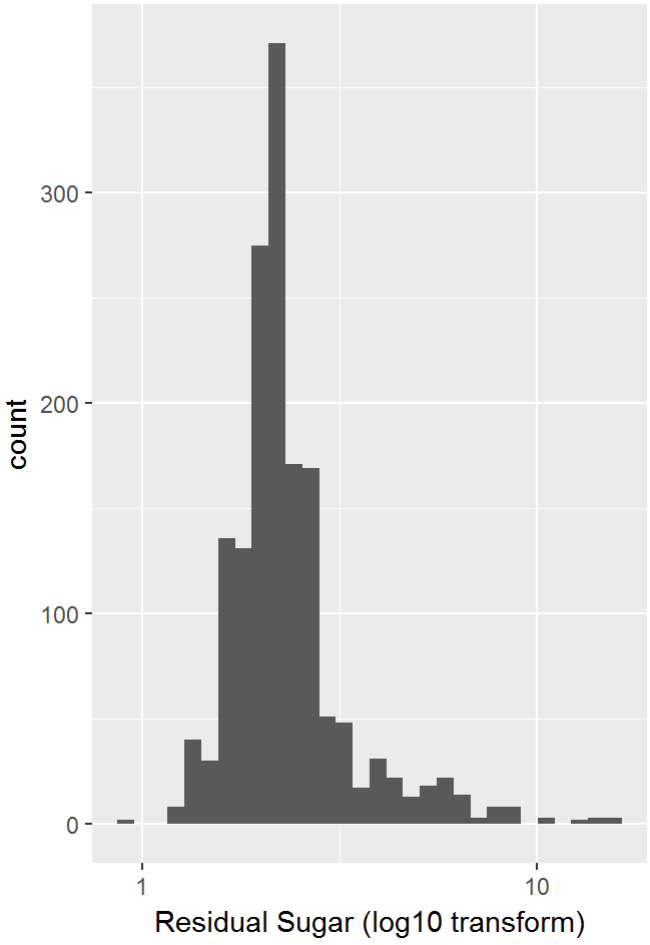
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```

# 4) Residual Sugar


Residual Sugar
16
12
8
4
0.6
0.8
1.0
1.2
1.4
1

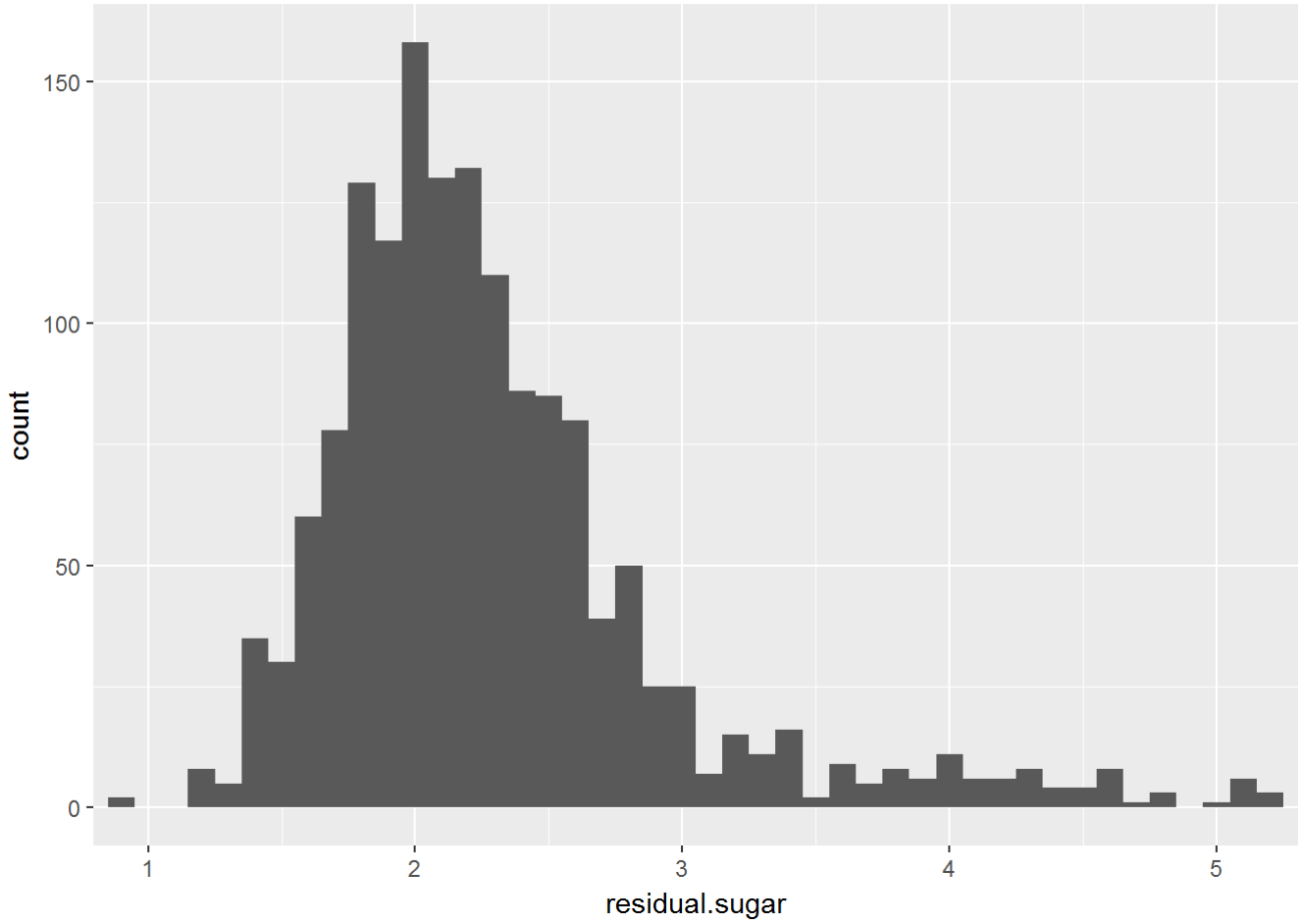

count
600
400
200
0
4
8
12
16
Residual Sugar


Residual Sugar (log10 transform)
10
1
0.6
0.8
1.0
1.2
1.4
1


count
300
200
100
0
1
10
Residual Sugar (log10 transform)

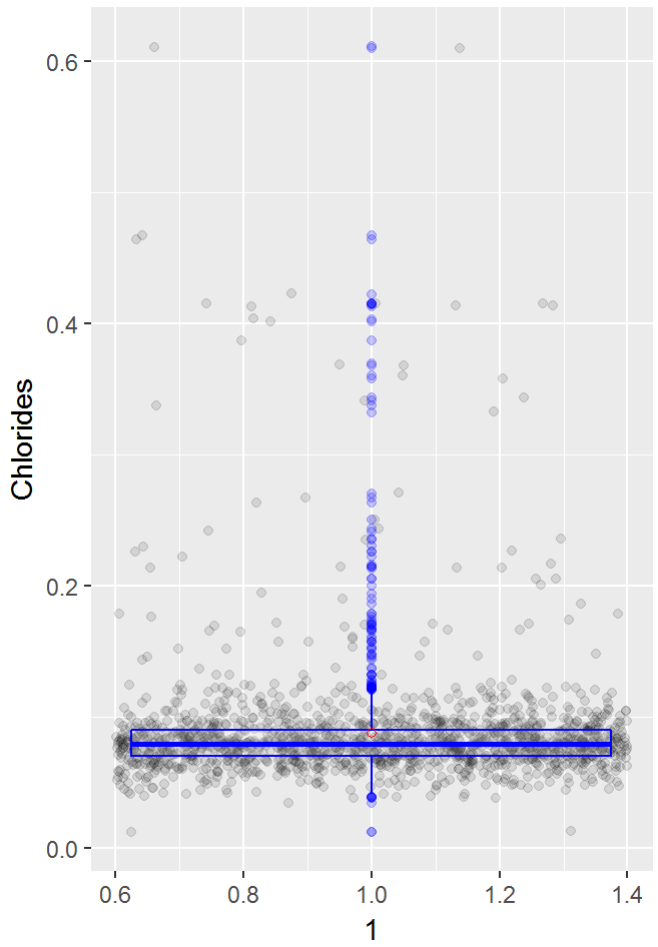
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```

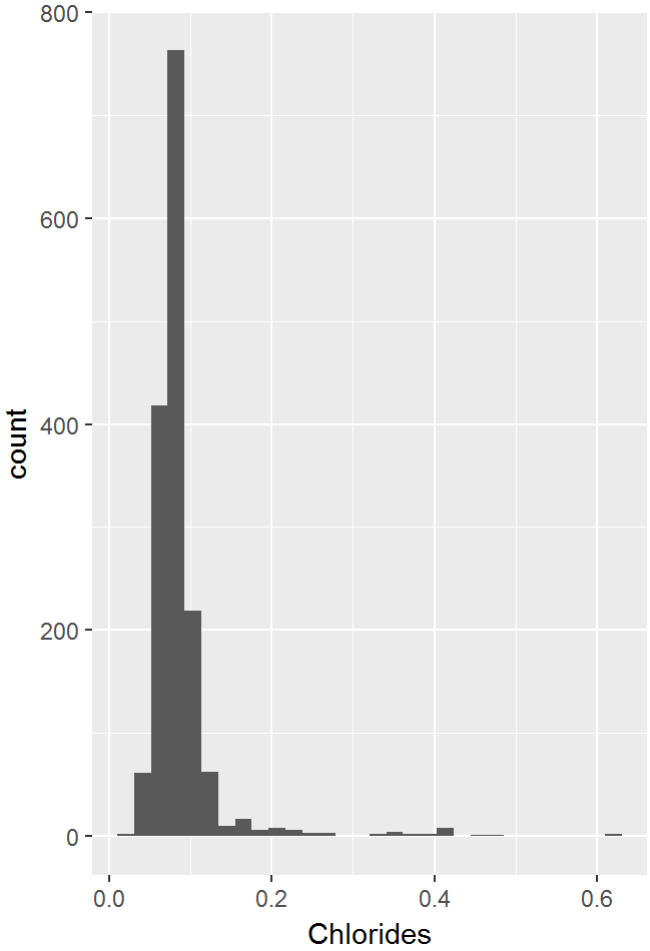
Residual sugar has a very long-tail distribution with many outliers. We can also note that there are a lot of outliers in this distribution. It will be interesting to see how these outliers affect the quality of wine. In the log transform plots, the values are still very skewed, but it looks more like a normal distribution.

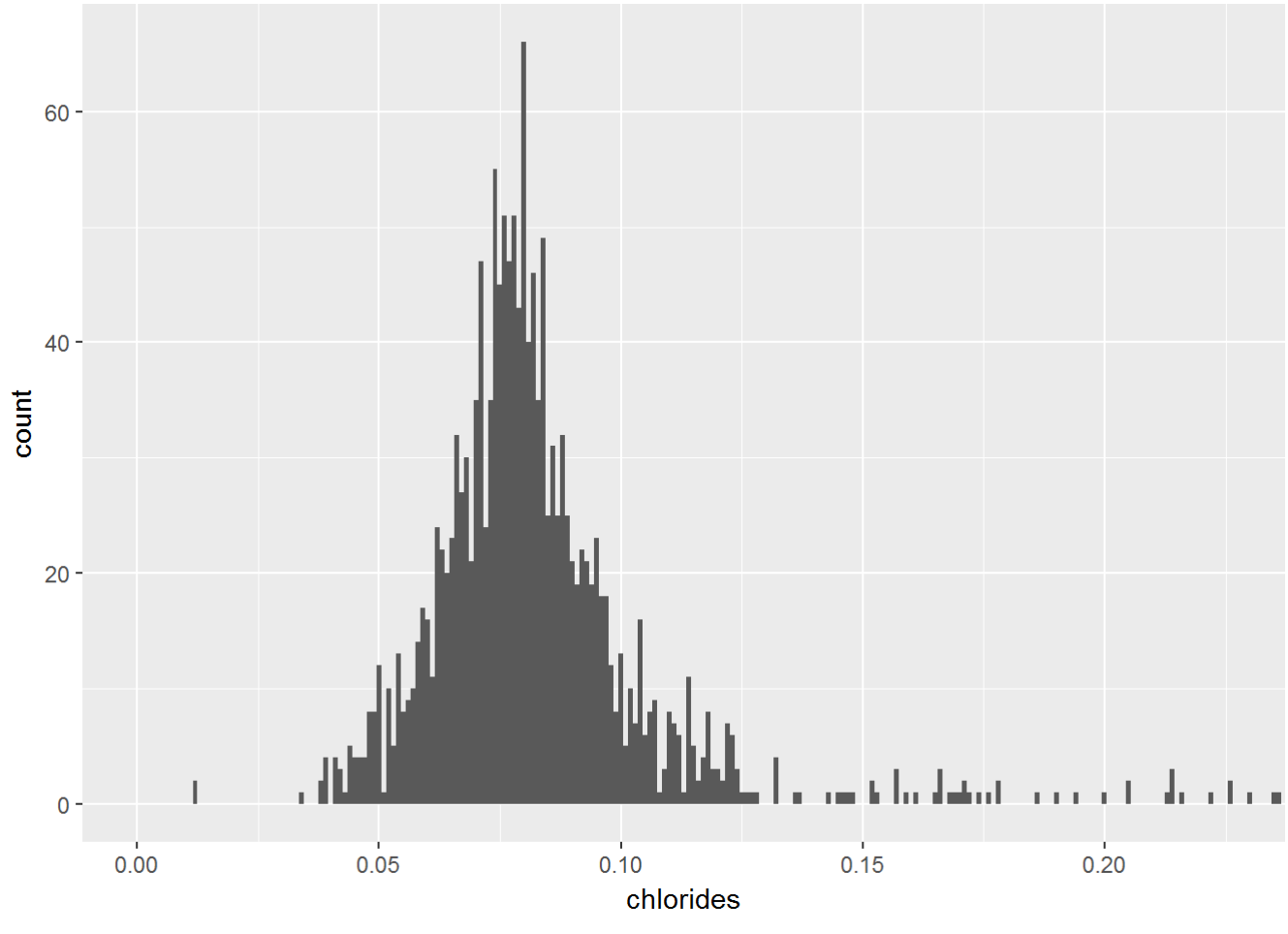


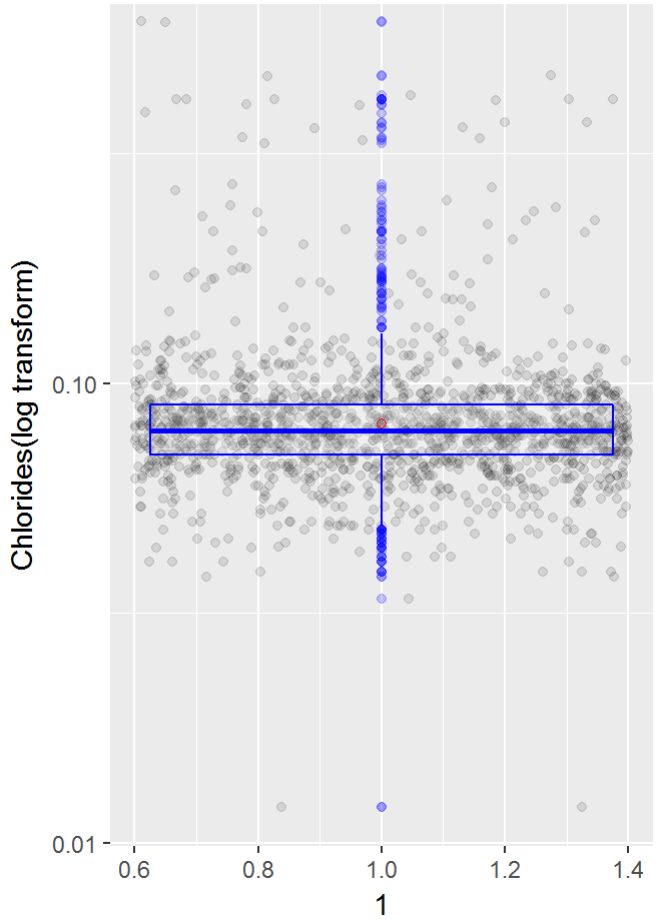

In the third plot, I removed the top five percent of data points to have a better understanding of the distribution. We see that most wines have residual sugar at around 2

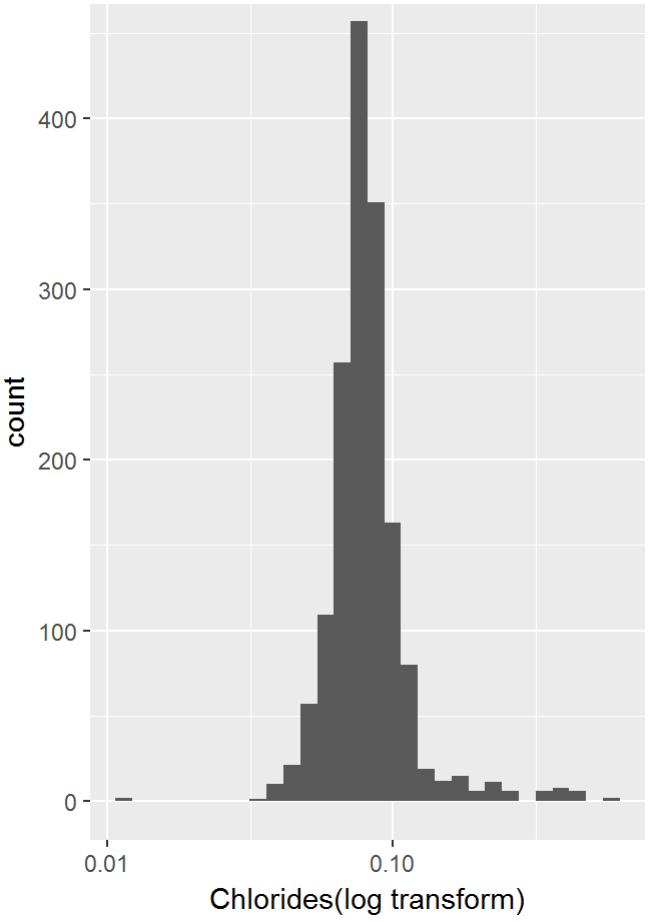
# 5)Cholorides


Chlorides
0.6
0.4
0.2
0.0
0.6
0.8
1.0
1.2
1.4
1


count
800
600
400
200
0
0.0
0.2
0.4
0.6
Chlorides


count
60
40
20
0
0.00
0.05
0.10
0.15
0.20
chlorides

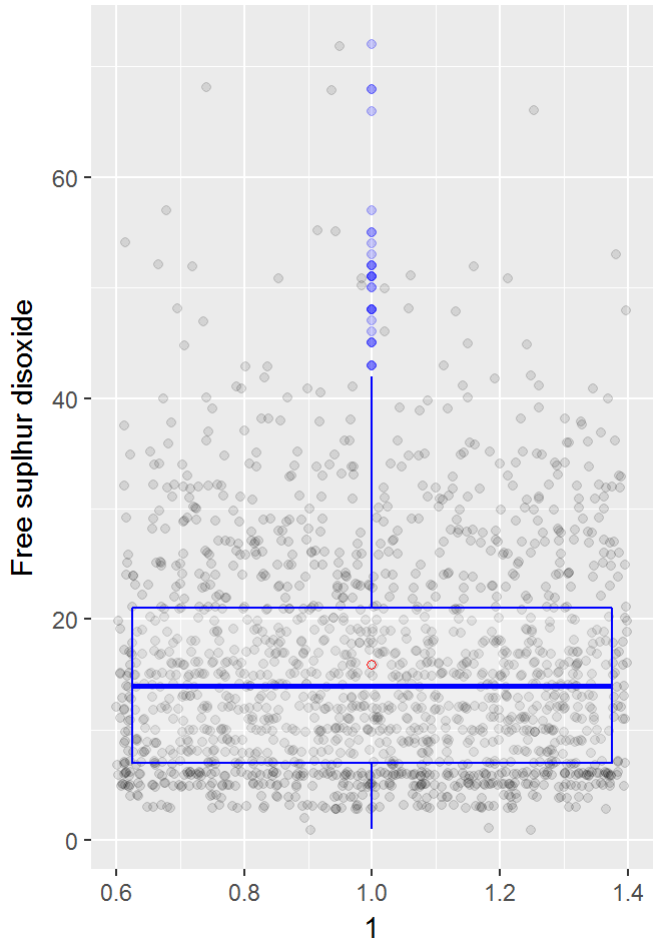


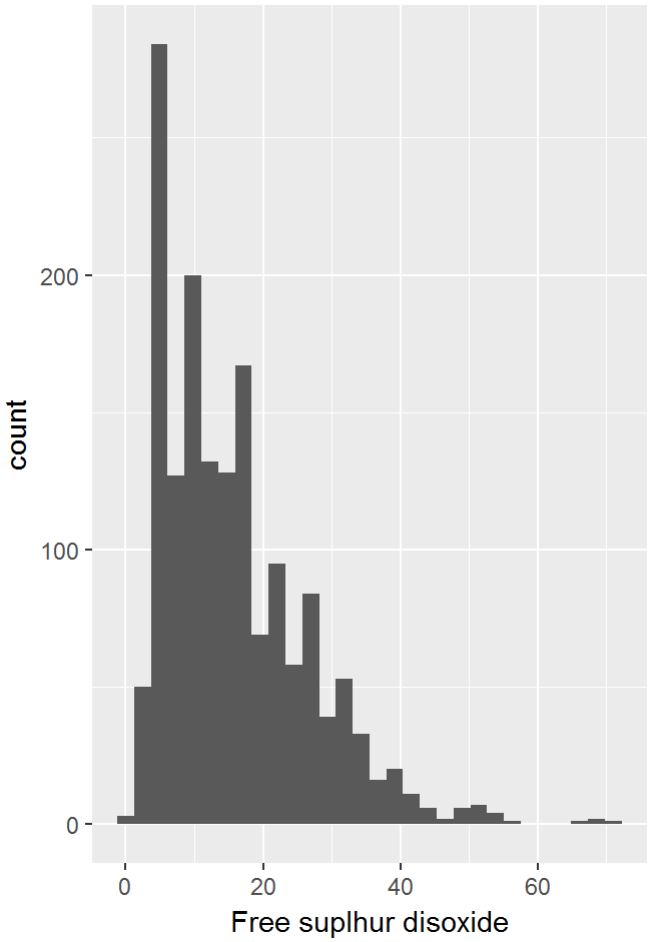



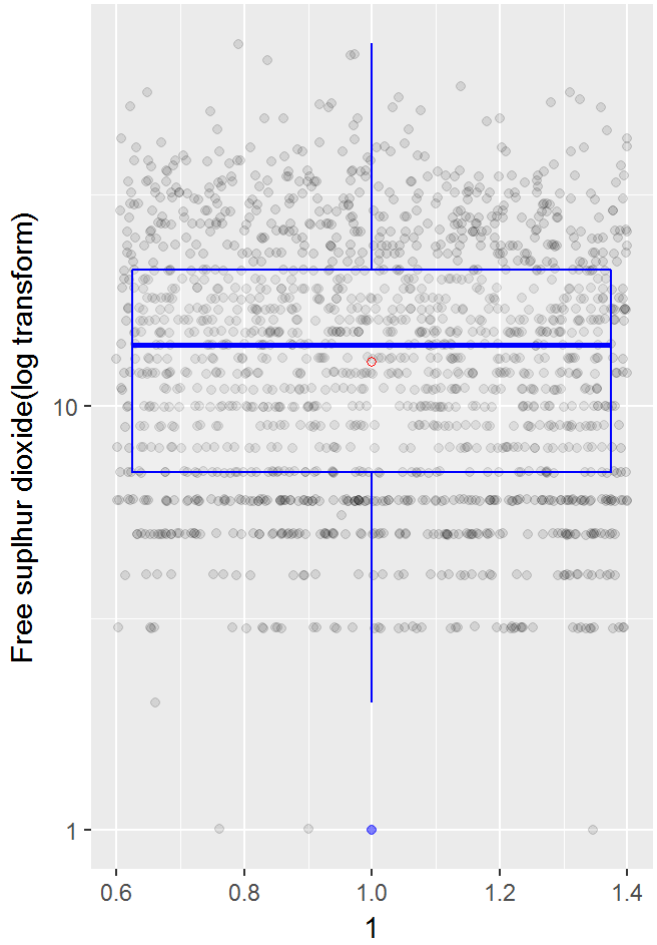

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```

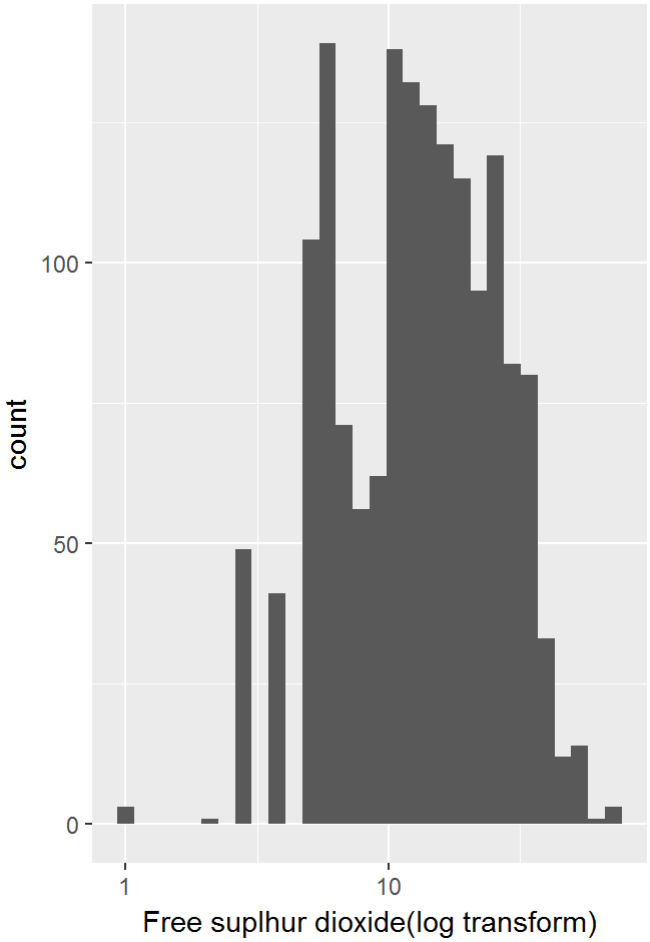
Chlorides have distribution similar to residual sugar and have a strong concentration around the median. We also note a lot of outliers from the box plot. In the second plot, the top two percent of data points were removed to help understand the distribution of points around the median.

# 6) Free sulphur dioxide


Free suplhur disoxide
60
40
20
0
0.6
0.8
1.0
1.2
1.4
1


count
200
100
0
0
20
40
60
Free suplhur disoxide

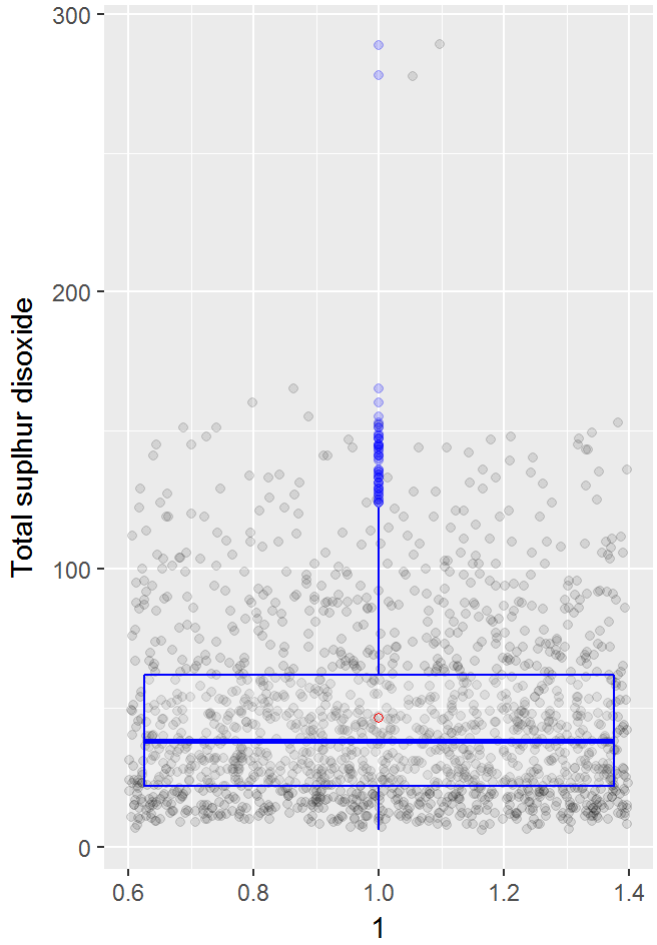

Free suplhur dioxide(log transform)
10
1
0.6
0.8
1.0
1.2
1.4
1

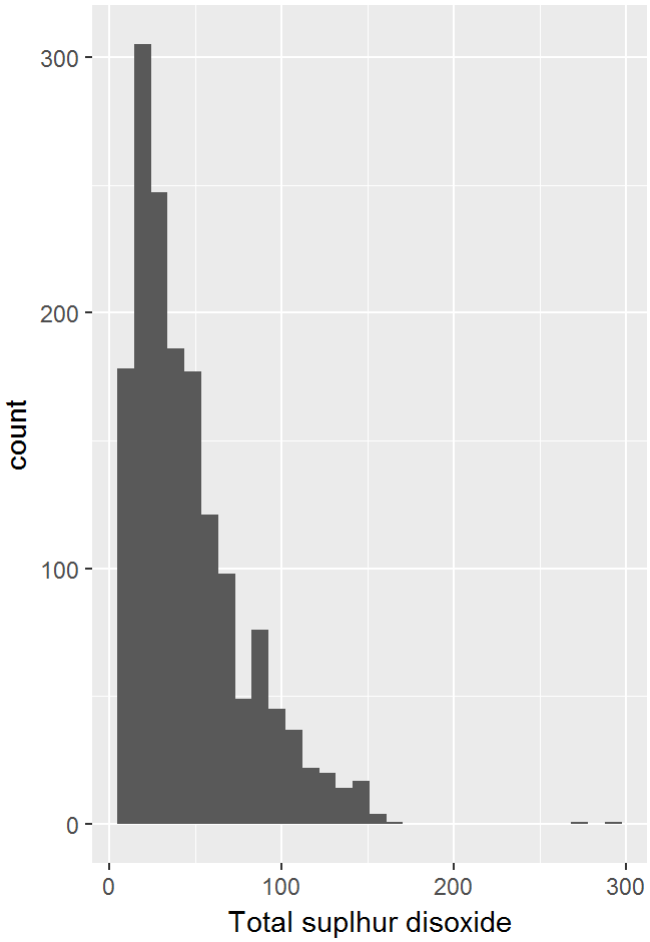

count
100
50
0
1
10
Free suplhur dioxide(log transform)

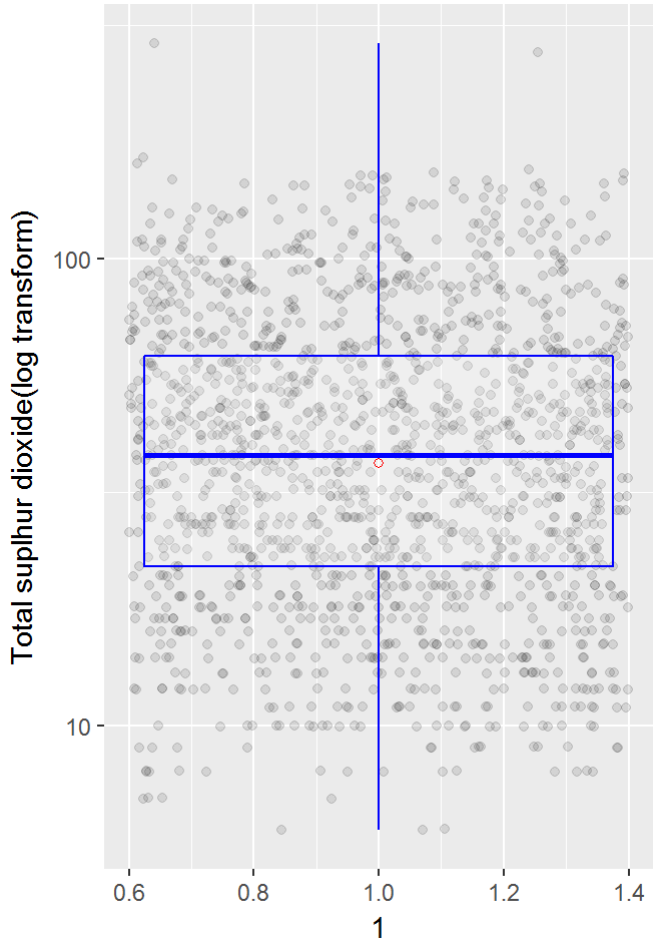
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
```

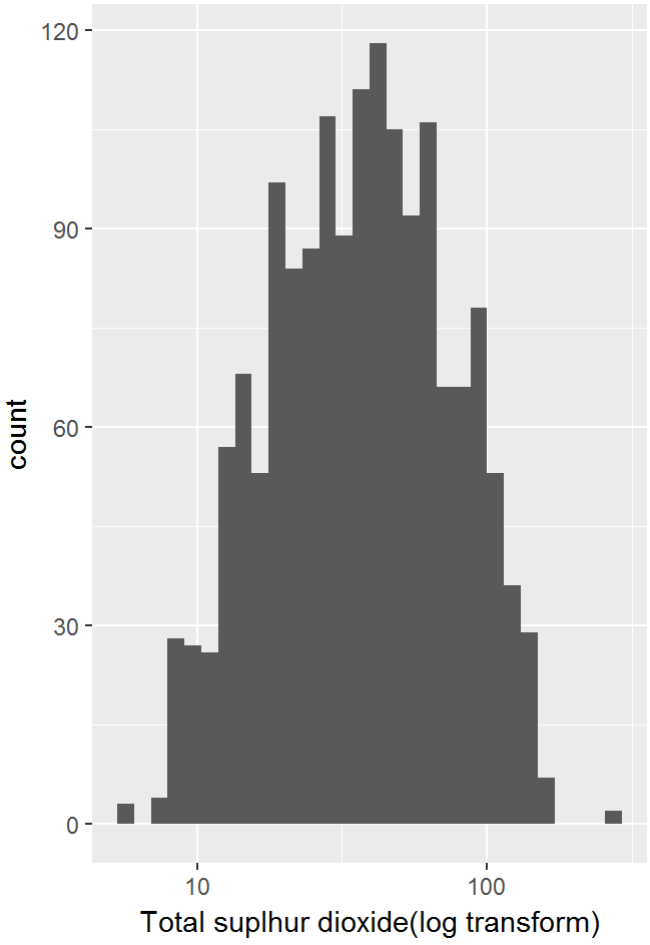
Interesting to note that the free sulphur dioxide has a bi-modal distribution when we take the log transform. We also note that data is well spread out compared to the other features we have seen yet.

# 7) Total suplhur dioxide


Total suplhur disoxide
300
200
100
0
0.6
0.8
1.0
1.2
1.4
1


count
300
200
100
0
0
100
200
300
Total suplhur disoxide


Total suplhur dioxide(log transform)
100
10
0.6
0.8
1.0
1.2
1.4
1

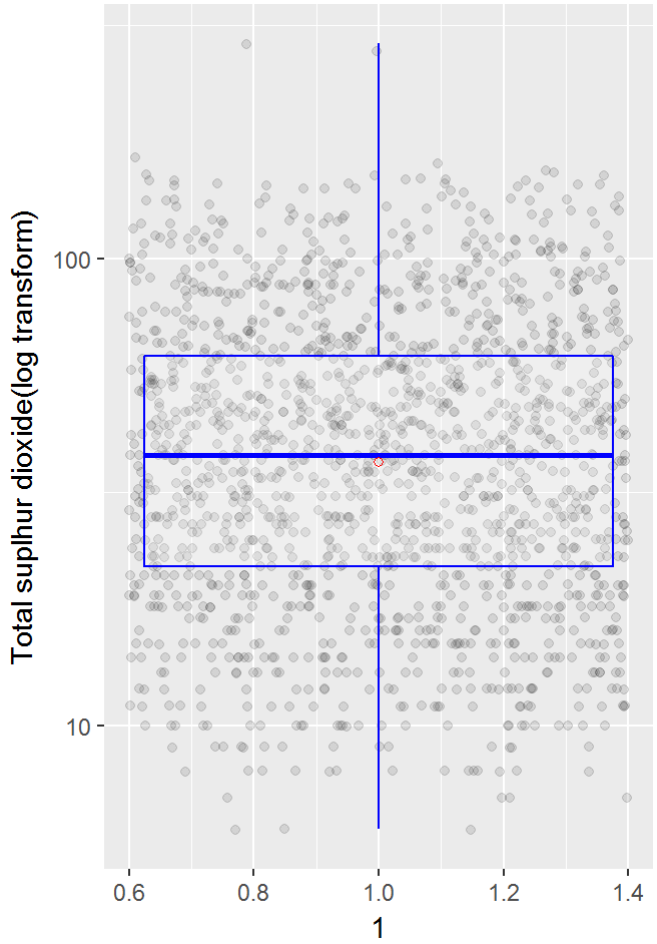

count
120
90
60
30
0
10
100
Total suplhur dioxide(log transform)

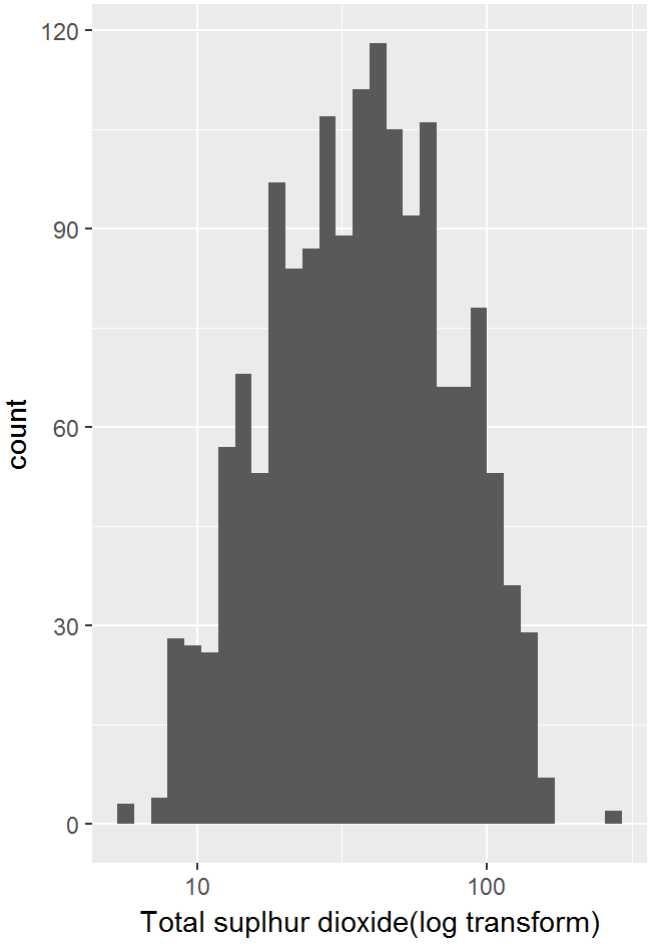
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00
```

Total sulfur dioxide is similar in ways to free sulfur dioxide. I would argue that its points are not quite a dispersed, as there are fewer outliers and its interquartile range does not look quite as large. It also has a long-tail distribution, but when we look at its log10 plot, the points are rather normally distributed.

# 7.1) Bound sulphur dioxide

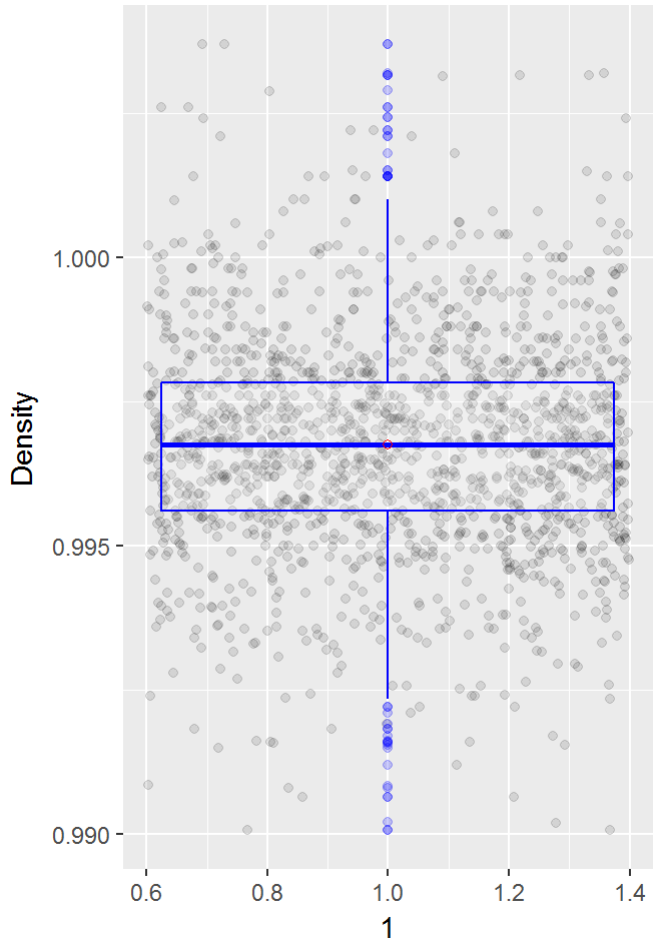
I created a new variable from the bound sulfur dioxide (total sulfur dioxide - free sulfur dioxide) to see if it has any intereseting pattern. Let's see the comparison of all three features.

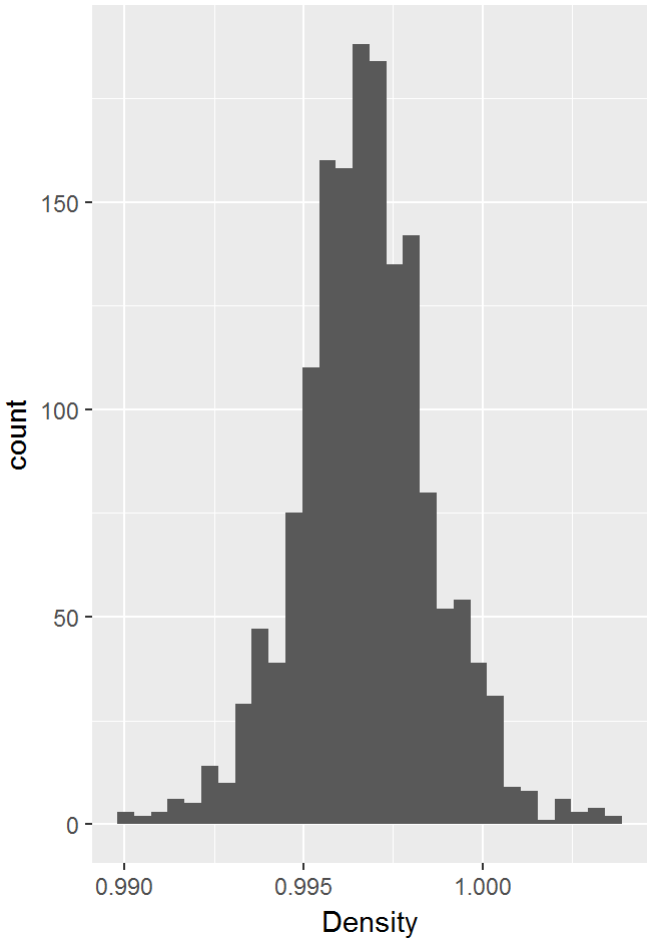


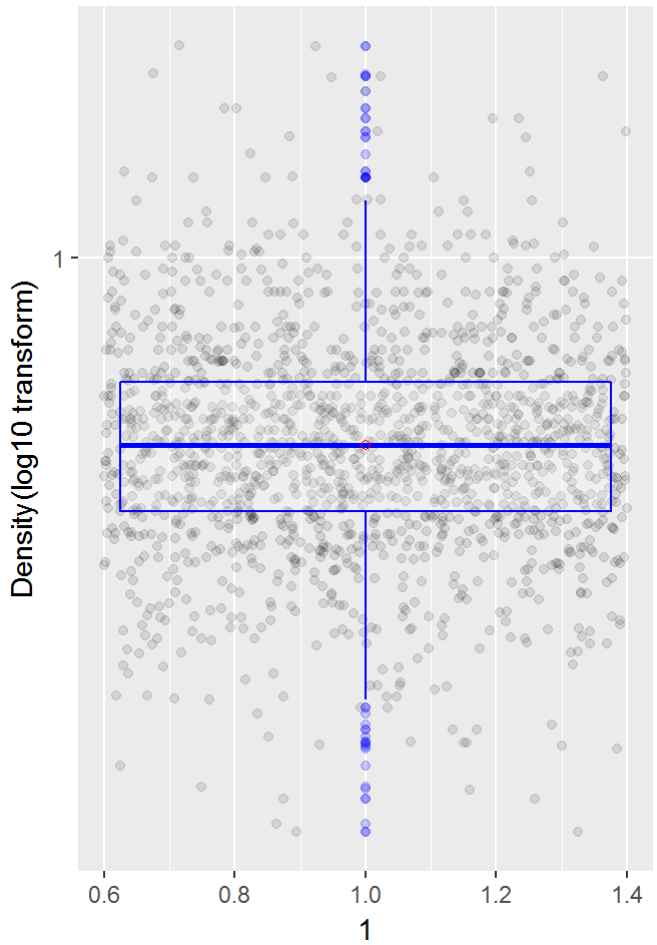



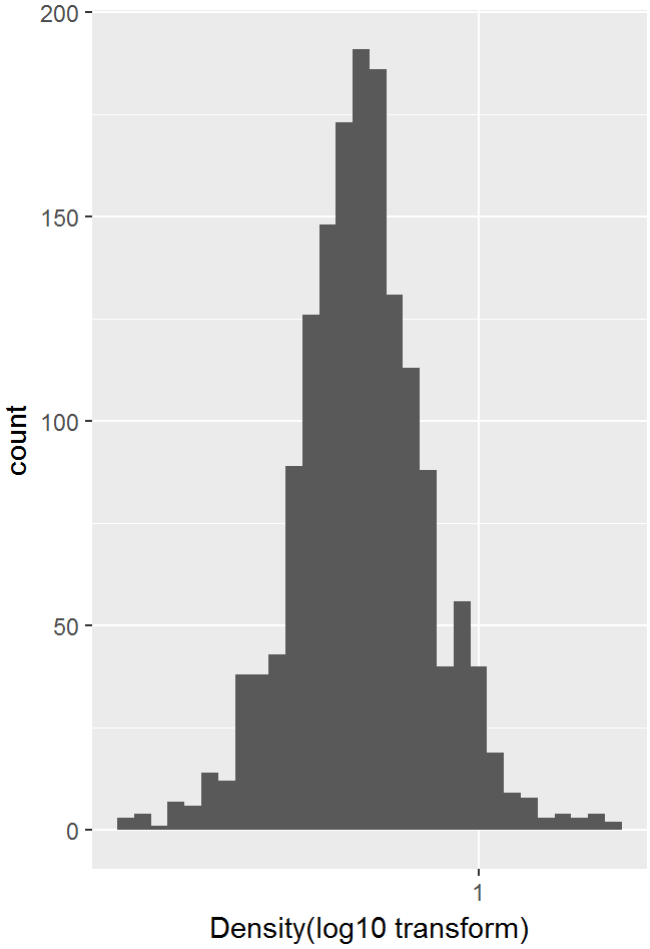

We note that there is no significant difference in the distribution of the new variable.

# 8) Density


Density
1.000
0.995
0.990
0.6
0.8
1.0
1.2
1.4
1


count
150
100
50
0
0.990
0.995
1.000
Density

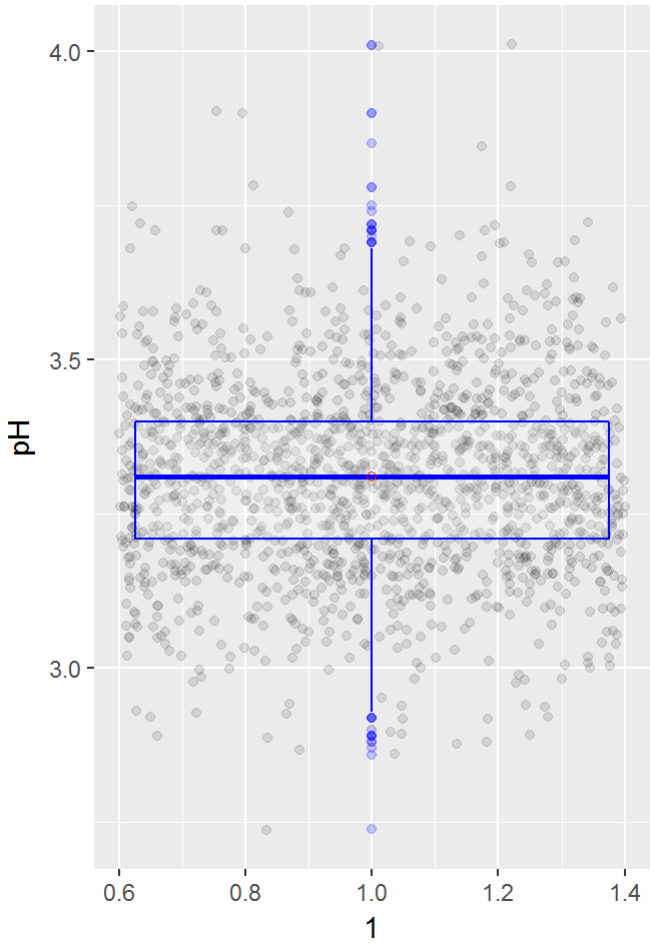

Density(log10 transform)
1
0.6
0.8
1.0
1.2
1.4
1

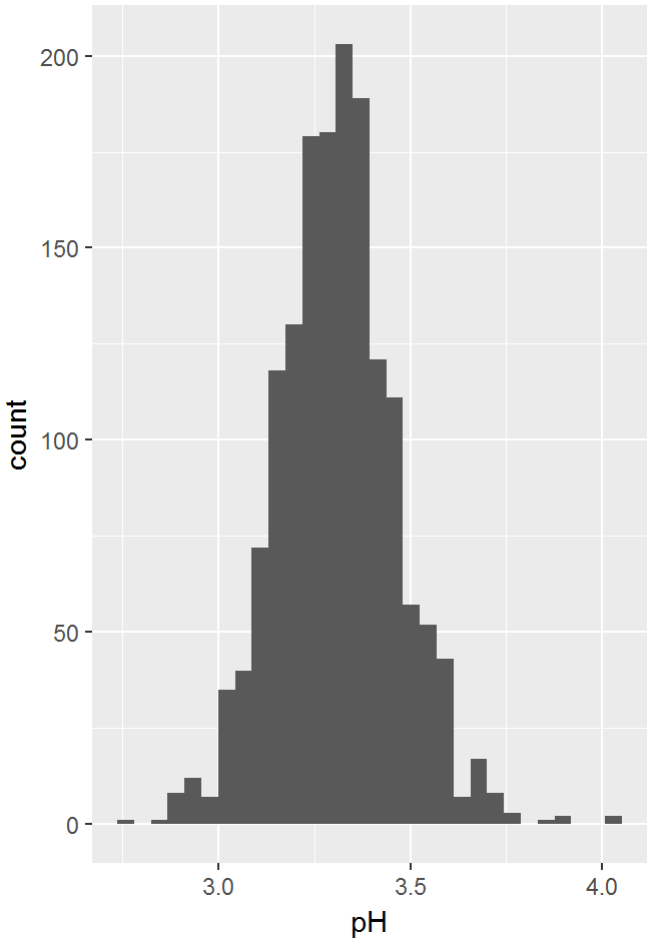

count
200
150
100
50
0
1
Density(log10 transform)

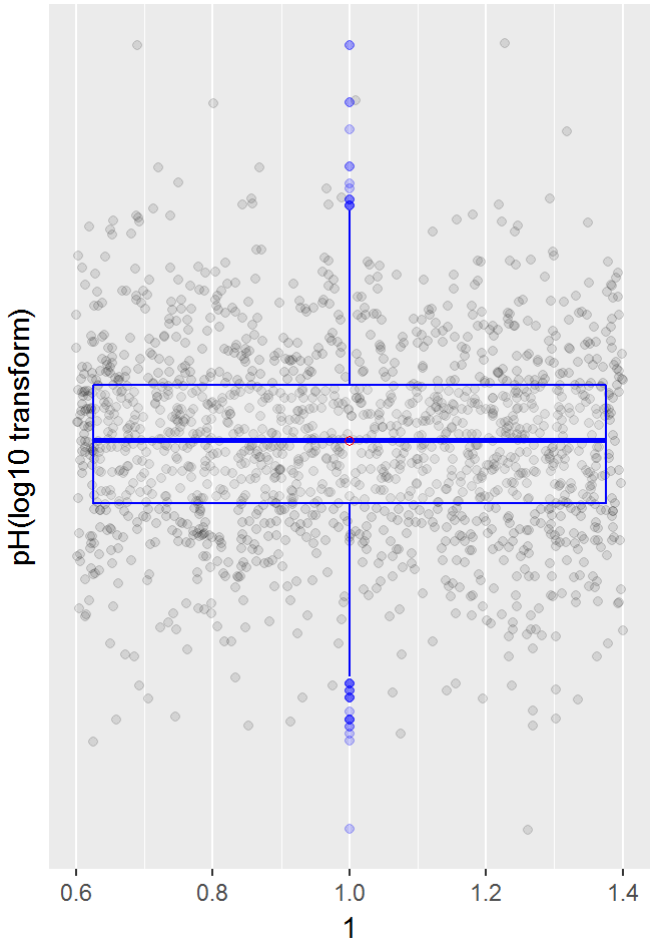
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0040
```

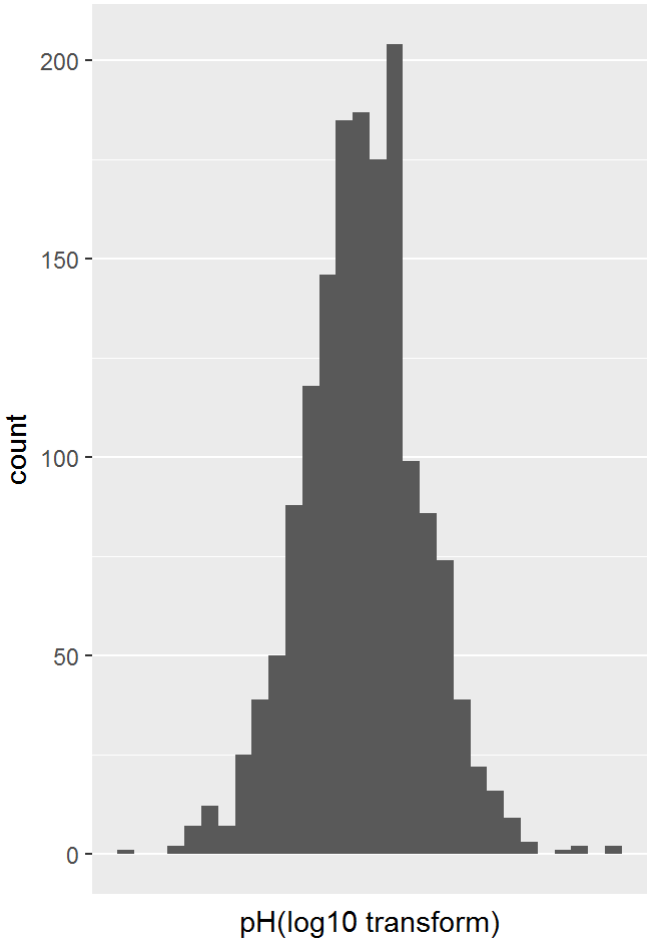
Density has a very normal looking distribution with most of the values falling between 0.995 and 1. For comparison, water has a density of 1, so most of our wine is less dense than water. There are very few outliers.

# 9) pH


4.0
3.5
3.0
pH
0.6
0.8
1.0
1.2
1.4
1


200
150
100
50
0
count
3.0
3.5
4.0
pH

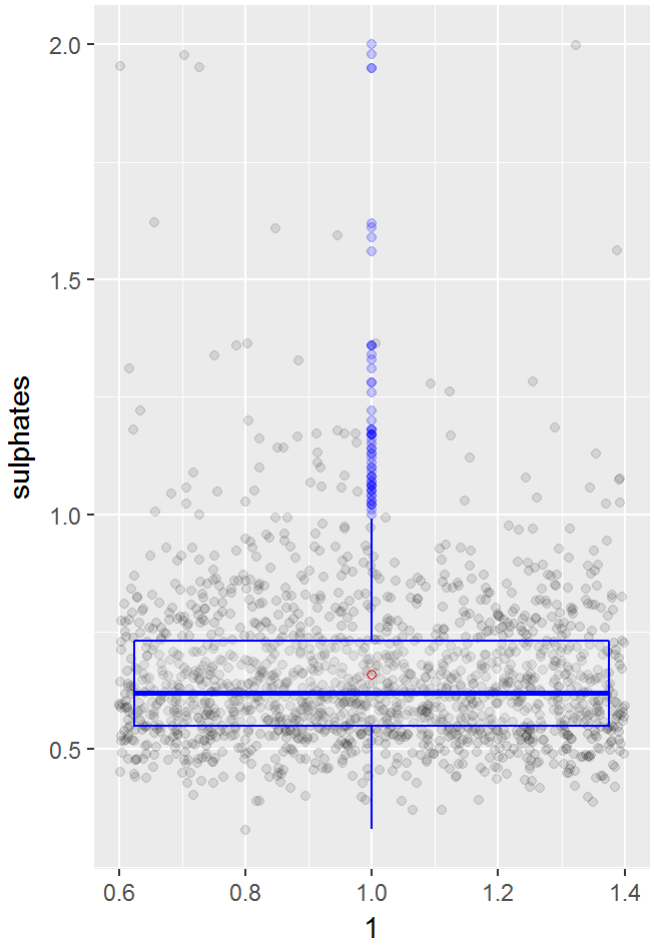

pH(log10 transform)
0.6
0.8
1.0
1.2
1.4
1

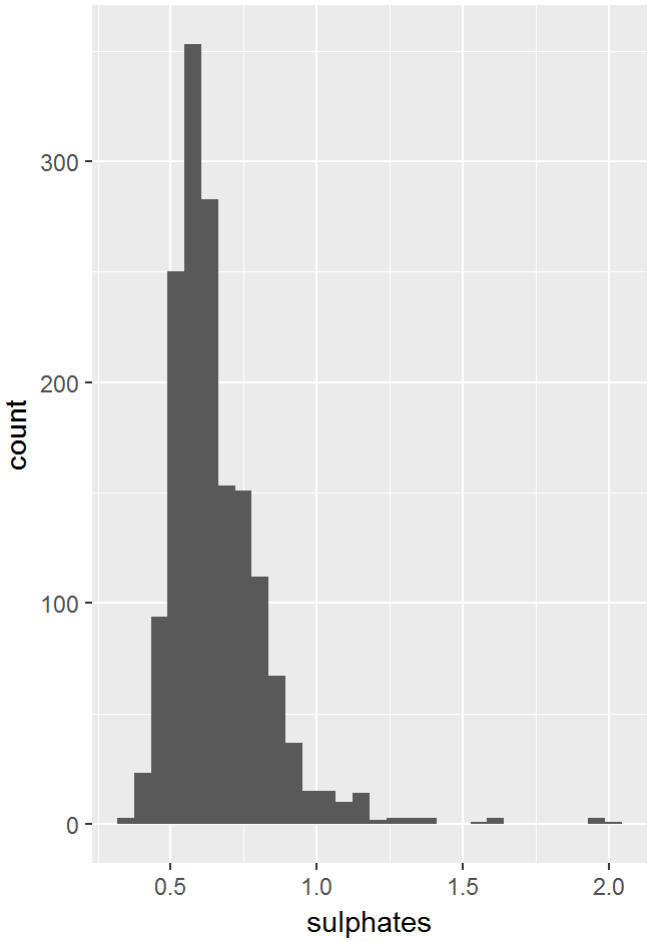

200
150
100
50
0
count
pH(log10 transform)

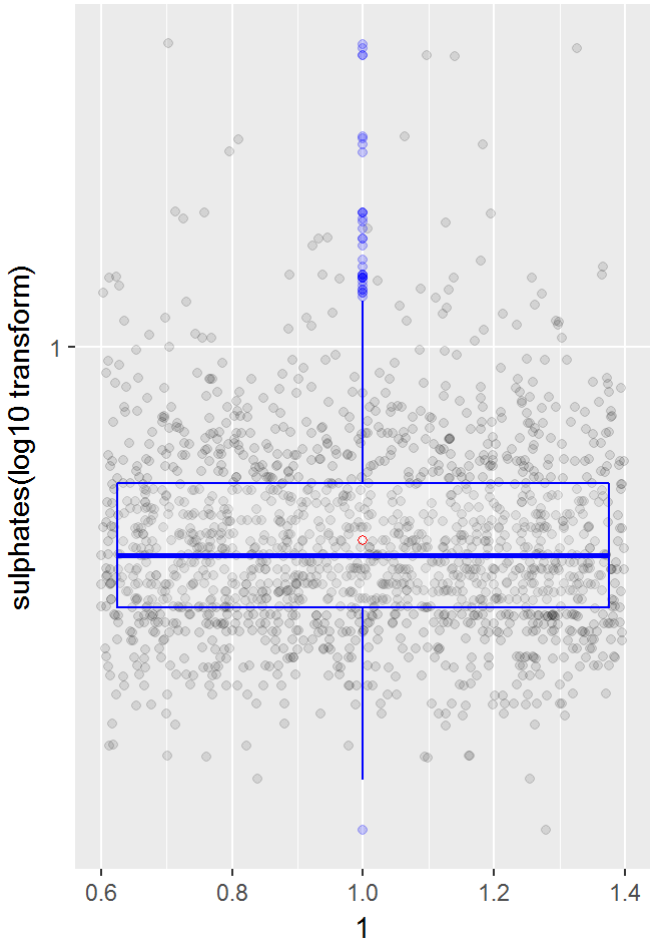
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010
```

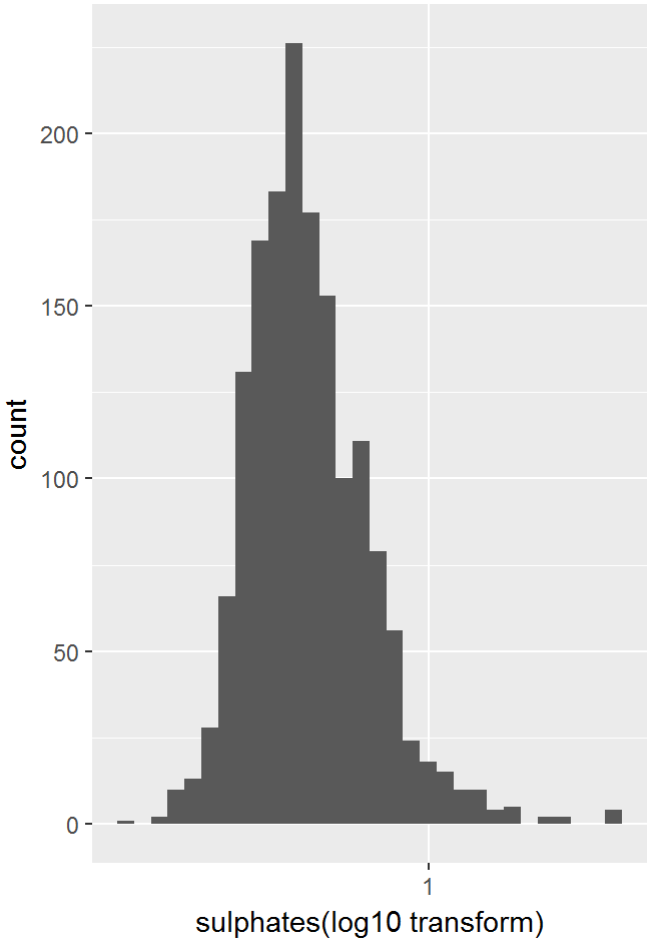
Another normal looking distribution, with most of the pH values falling between 3.1 and 3.5. There are a few outliers.

# 10) Sulphates


sulphates
2.0
1.5
1.0
0.5
0.6
0.8
1.0
1.2
1.4
1


count
300
200
100
0
0.5
1.0
1.5
2.0
sulphates

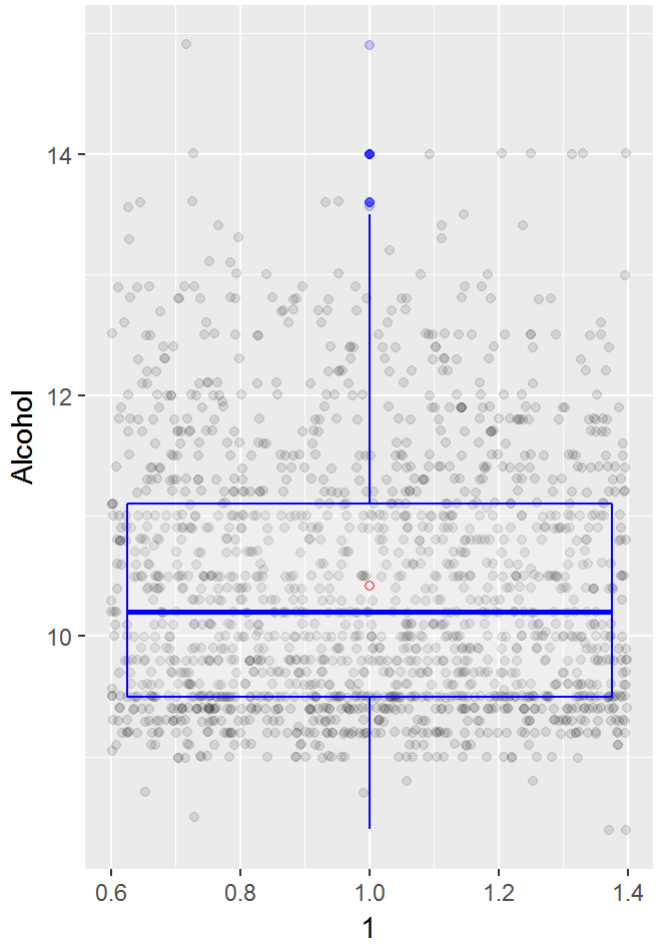

sulphates(log10 transform)
1
0.6
0.8
1.0
1.2
1.4
1

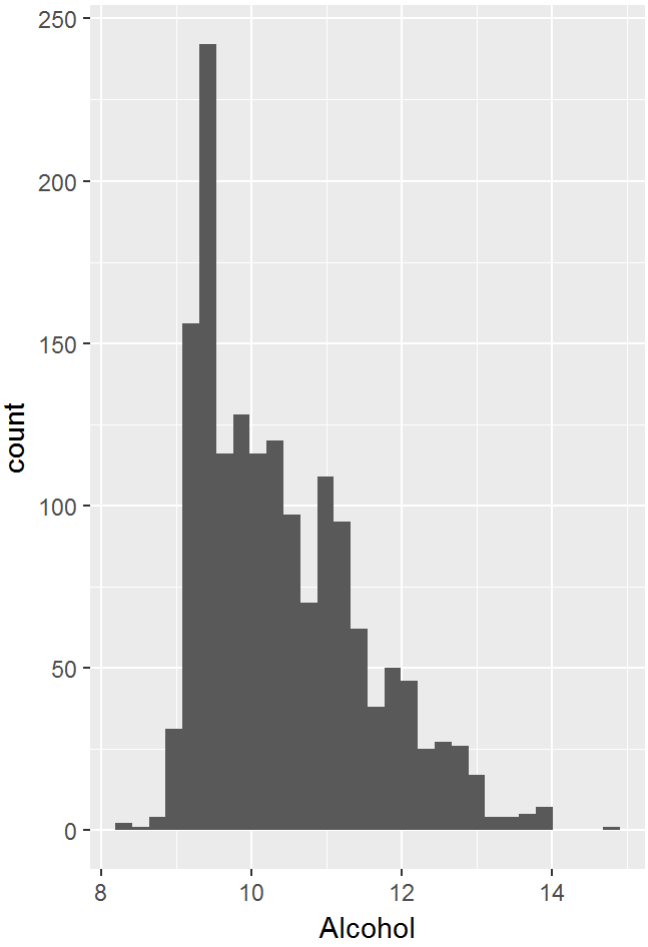

count
200
150
100
50
0
1
sulphates(log10 transform)

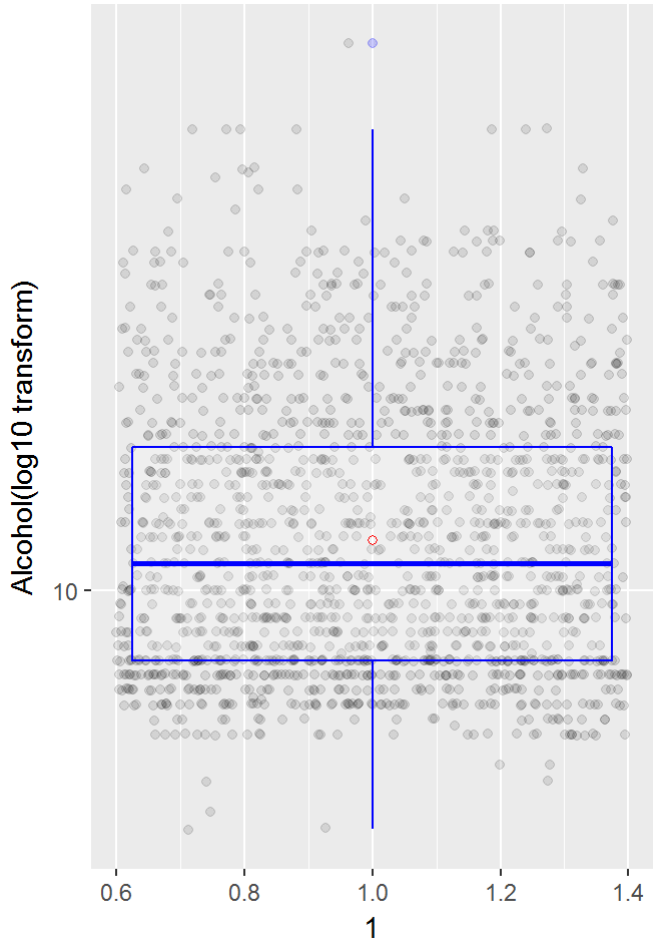
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```

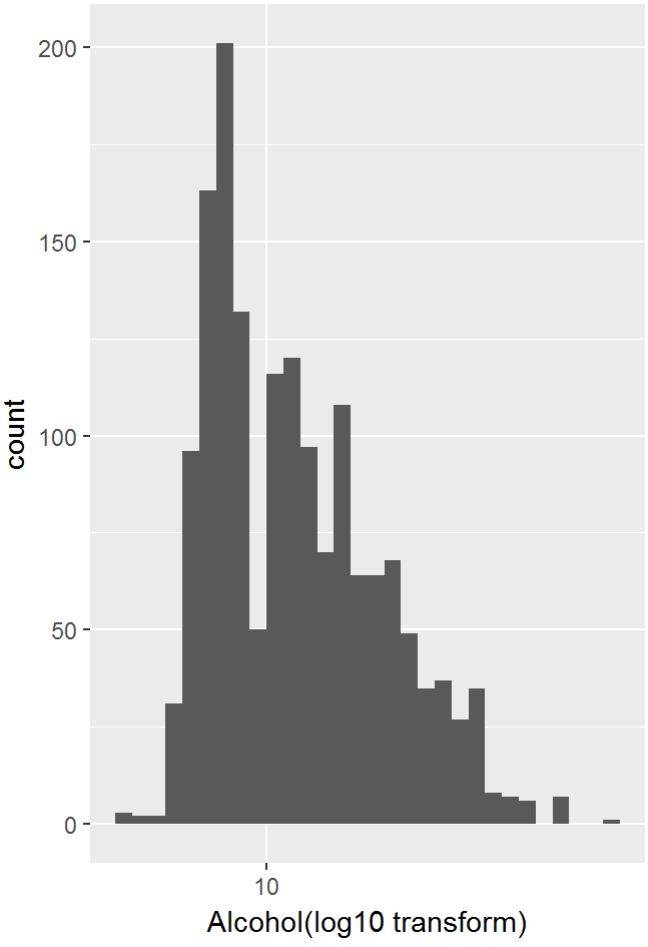
Sulphates is more long-tail than density and pH, it still looks rather normally distributed, as most of the values are clustered around 0.6. An interesting point about sulphates, is that some of its outliers are very far away from median. It will be interesting to see how that affects the quality of wine. Looking at its log transforms, sulphates are much more normally distributed, and there are still some outliers, despite the transformation.

# 11) Alcohol


Alcohol
14
12
10
0.6
0.8
1.0
1.2
1.4
1


count
250
200
150
100
50
0
8
10
12
14
Alcohol

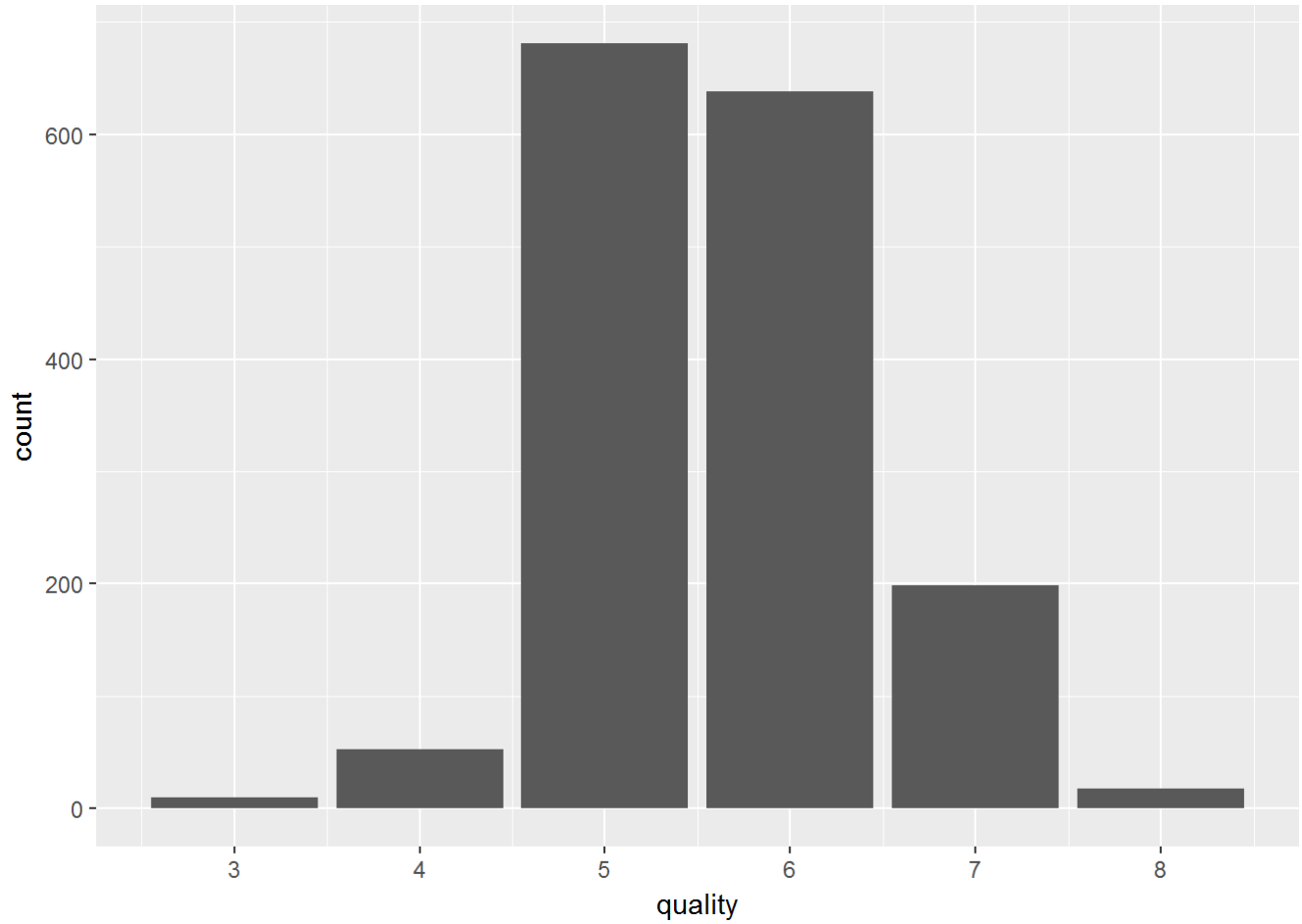

Alcohol(log10 transform)
10
0.6
0.8
1.0
1.2
1.4
1


count
200
150
100
50
0
10
Alcohol(log10 transform)

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```

Alcohol has a long-tail distribution, with there only being a few outliers. Looking at the log transforms does not reveal anything new, except that it still has a long-tail distribution which is similar to the original plots. Most wines have less than 11% alcohol which is true to my limited knowledge as I rarely have picked up a wine personally that is more than 11% in alcohol content.

## 12) Quality




Quality is on a 1-10 scale, which means that most of the wines we will look at in the analysis are average wines. It will be interesting to try to find what can make a wine very good or very bad, and to see if there is much correlation between the variables.

# Univariate Analysis

## What is the structure of your dataset?

The dataset is a tidy set and it has 1599 observations with 13 variables for each one. All of the observations are numerical. The first variable is an index. The “quality” variable (score between 0 and 10) has only 6 discrete values: 3, 4, 5, 6, 7, 8. Most of the data is concentrated for wines having quality score of 5 and 6.

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

Quality is main interest in the dataset. It would be interesting to see which features contribute most to the quality of the wine. My end goal is to make a model that takes in the different input features of the wine and is able to predict the quality score for that composition.

## What other features in the dataset do you think will help support your investigation into your feature(s) of interest?

I expect alcohol, pH, residual sugar, and total acidity will contribute most to the quality of the wine. After a little research on red wine through google, I found that wine experts seem to enjoy a red wine that is neither tart, nor sweet, nor dry, but smooth and wet. It would be interesting to see the composition of different features for the good quality(7 or 8) wines in our dataset.

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

I created three new variables:

- bound.sulfur.dioxide: the result of subtract free.sulfur.dioxide from total.sulfur.dioxide
- total.acidity: the result from addition of acidity and volatile.acidity.
- class: to group wines in three classes -> bad (qualities 3 and 4), regular (qualities 5 and 6) and good (qualities 7 and 8).

## 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?

I noted that most of the wines have quality score of either 5 or 6. This could make it more difficult to determine what makes a good wine, as there is less data about them. Same is true for wines that have a quality score of 3 and 4. If we had more even distribution of wine qualities then it would be easier to contrast the differences between a good quality and bad quality wine.

Most of the data have an alcohol value between 9 and 12, with a median of 10. The wines having quality scores of 5 and 6 have less alcohol content compared to higher quality wines.

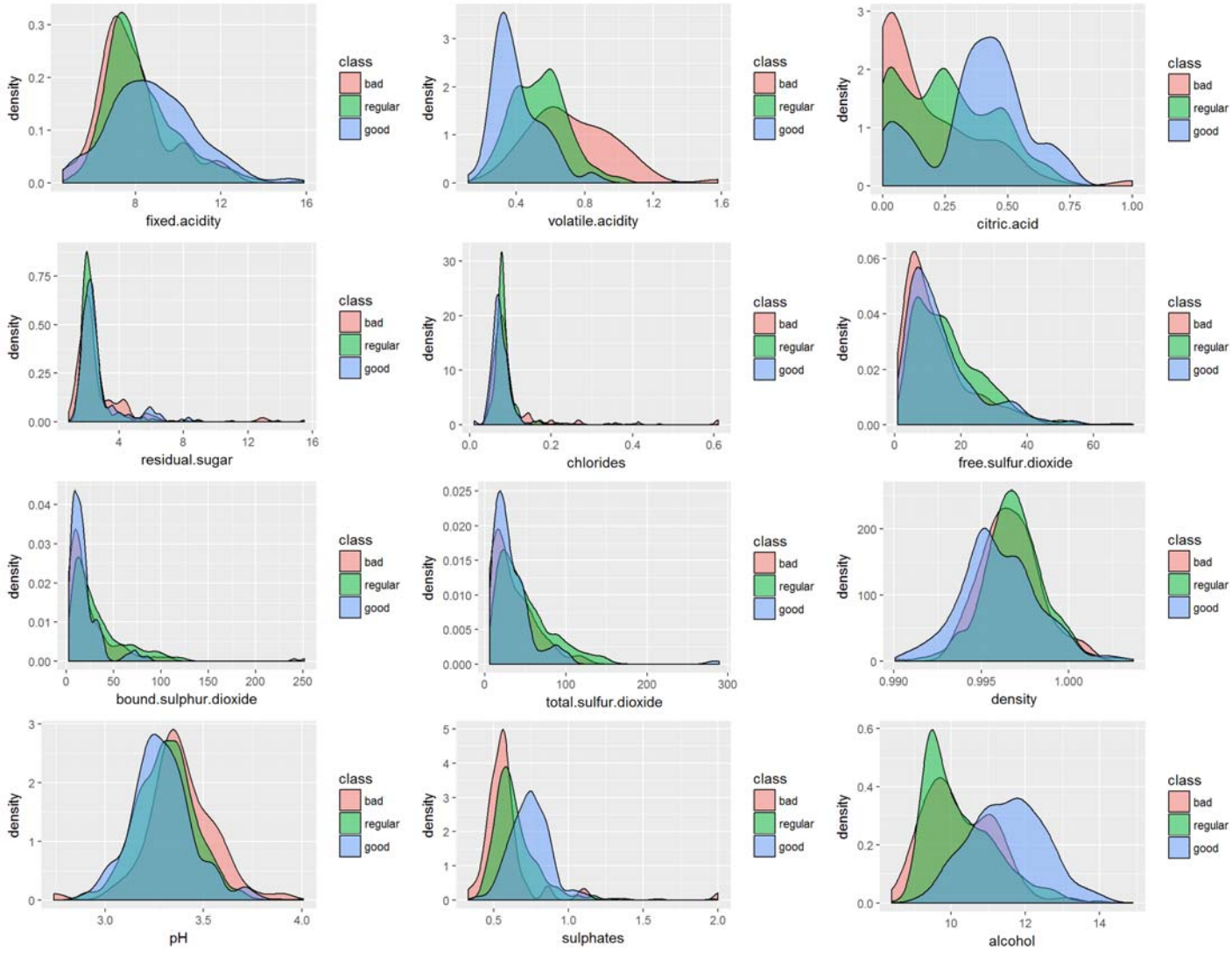
Regarding the new variables, bound sulfur dioxide (“nonfree.sulfur.dioxide”) tends to have bigger values than free sulfur dioxide. The percentage of free sulfur dioxide (“pfree.sulfur.dioxide”) has a distribution almost normal, with mean around 0.4.

For some of the features, there were noticible amount of outliers. I removed the top few percent of data points when looking at an additional plot. This was to have a better view of the core of the data, i.e. the interquartile range and how it is distributed.

As mentioned above I categorized the “quality” feature into bad, regular and good for better visualization plots for analysis later.

# Bivariate Plots Section

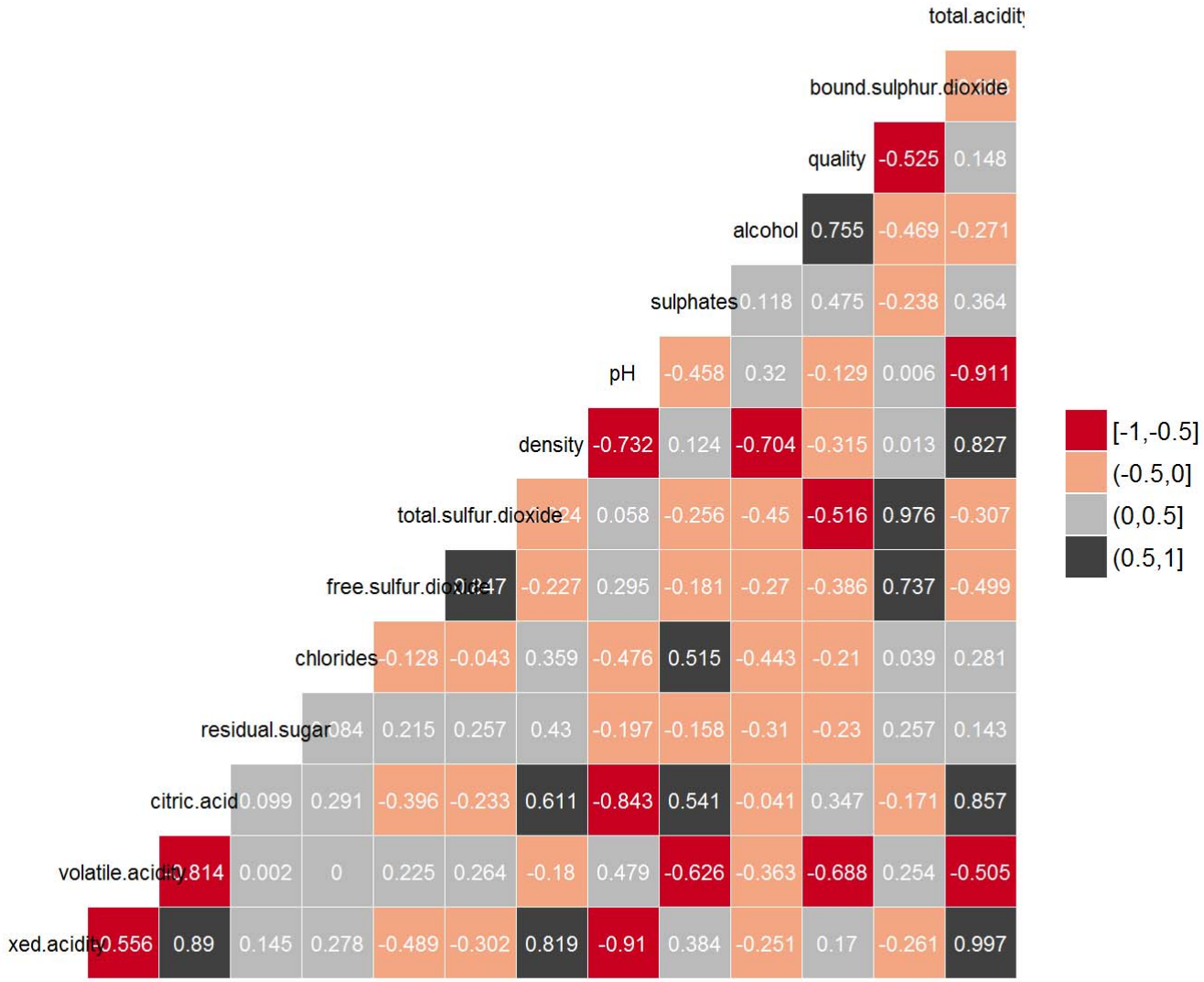
# 1) Let's plot and compare the distribution of different features with quality. We will use the new variable class to plot the density plots below




This was an interesting plot. I observed three points regarding wine quality:

- Bad wines have a bigger volatile acidity distribution.
- Bad wines have the least citric acid concentration.
- Good wines have more percentage of alcohol.

# 2) Now let's build a plot to see the corelation of the features with one another



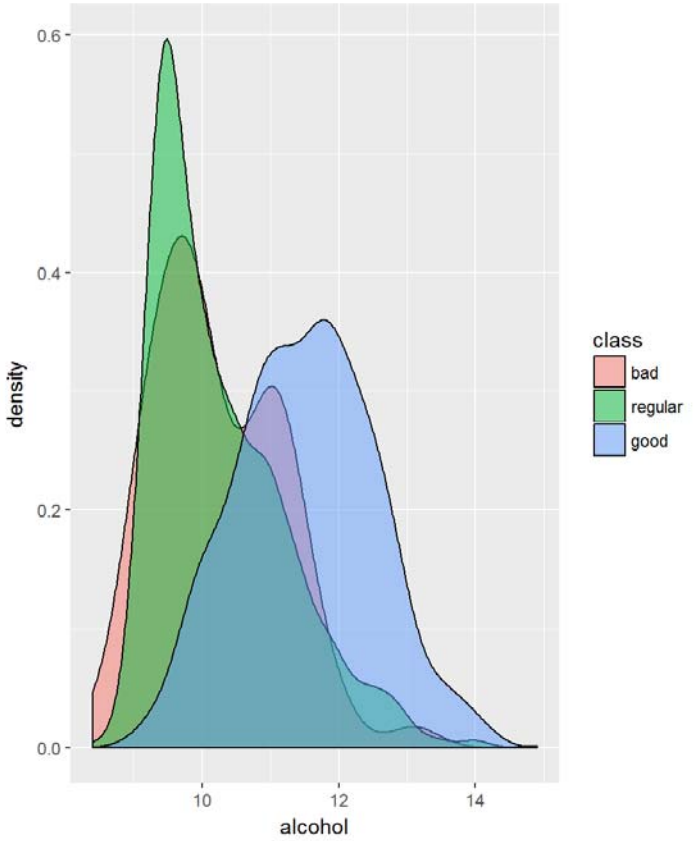

Intresting! We find that our intial guesses were true about which factors would be related to determine the quality of wine. Looking that the quality bar we note that: - “citric.acid”, “sulphates” and “alcohol” shows the bigger postive correlation values with quality - “volatile acidity” has a high negative correlation with quality

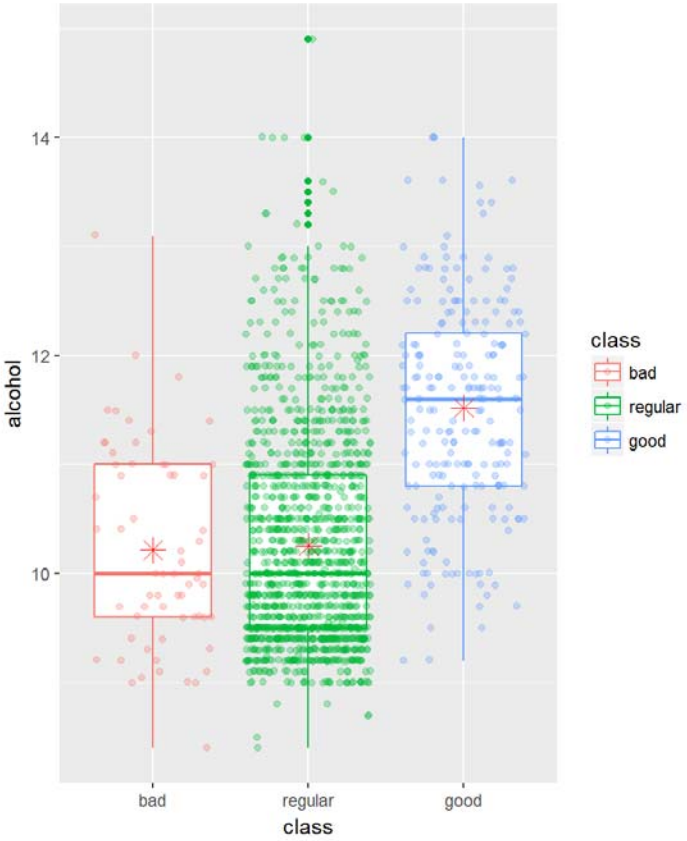
My assumption that sugar will be an important factor for wine quality seems to be incorrect based on what the plot shows.

# 3) Let’s explore these few found relations in depth

Let’s plot our feature with respect to class.

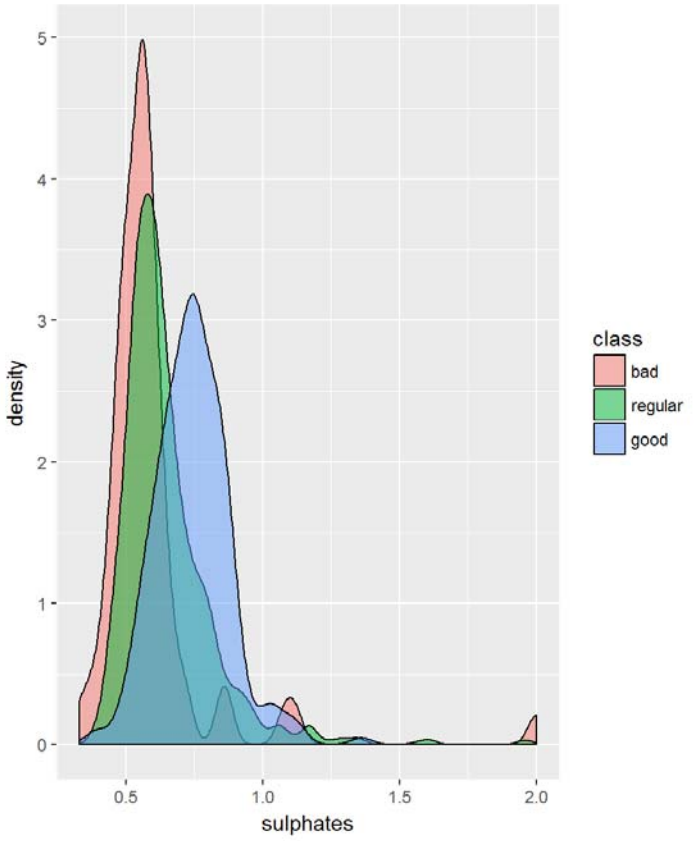
# 3.1) Alcohol vs Class

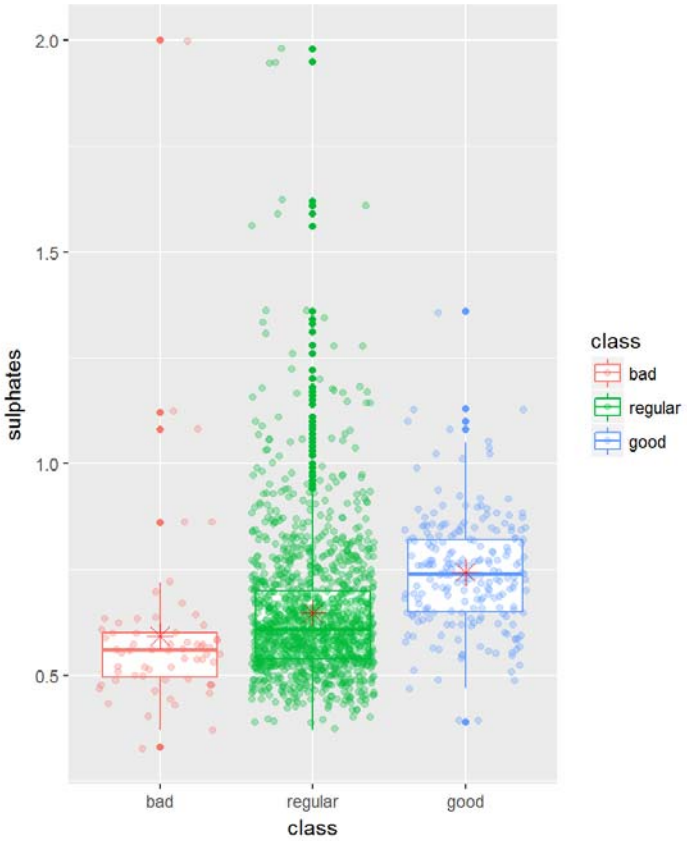



As per our observation, good quality wines have **higher** levels of alcohol.

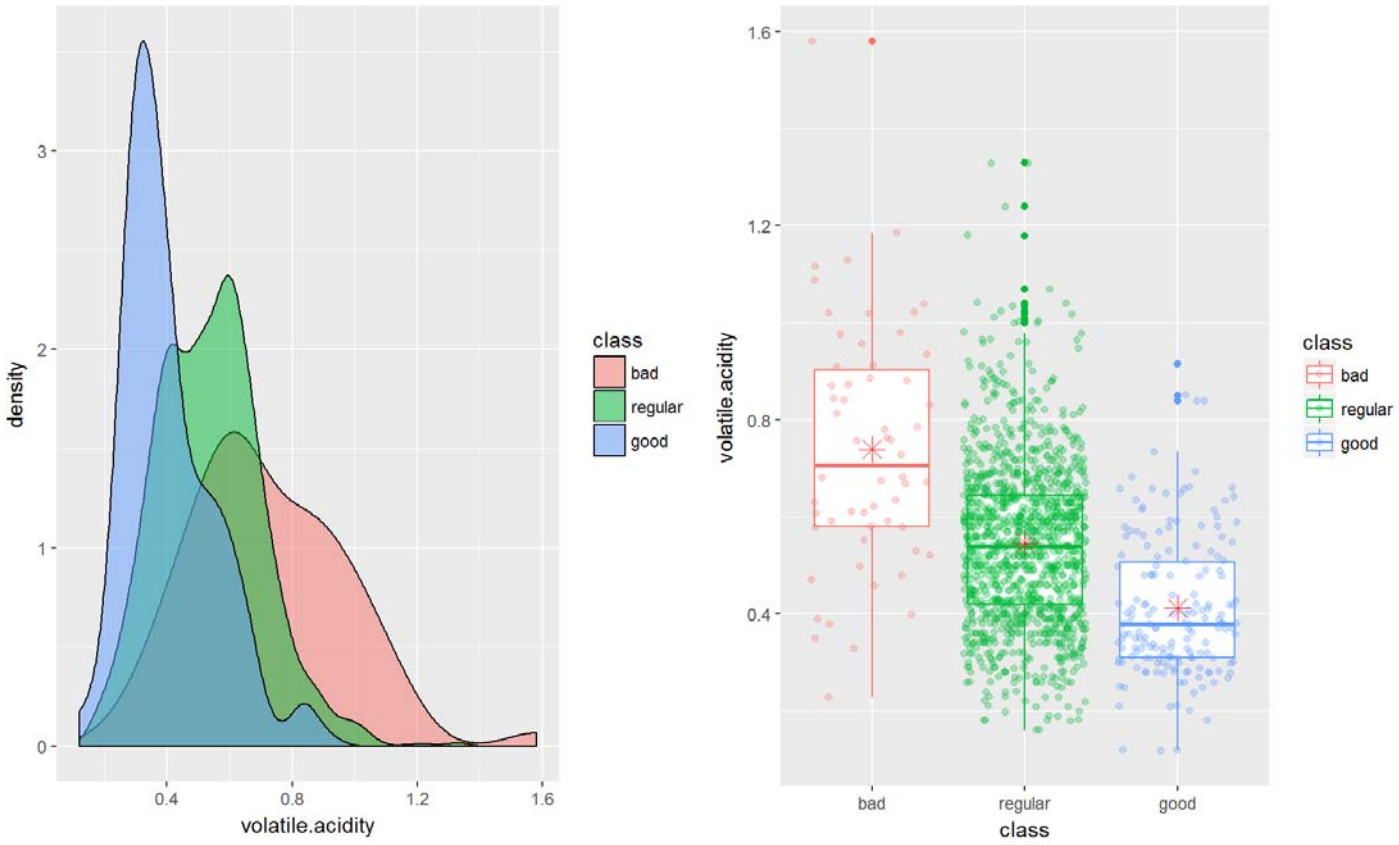
# 3.2) Sulphates vs Class



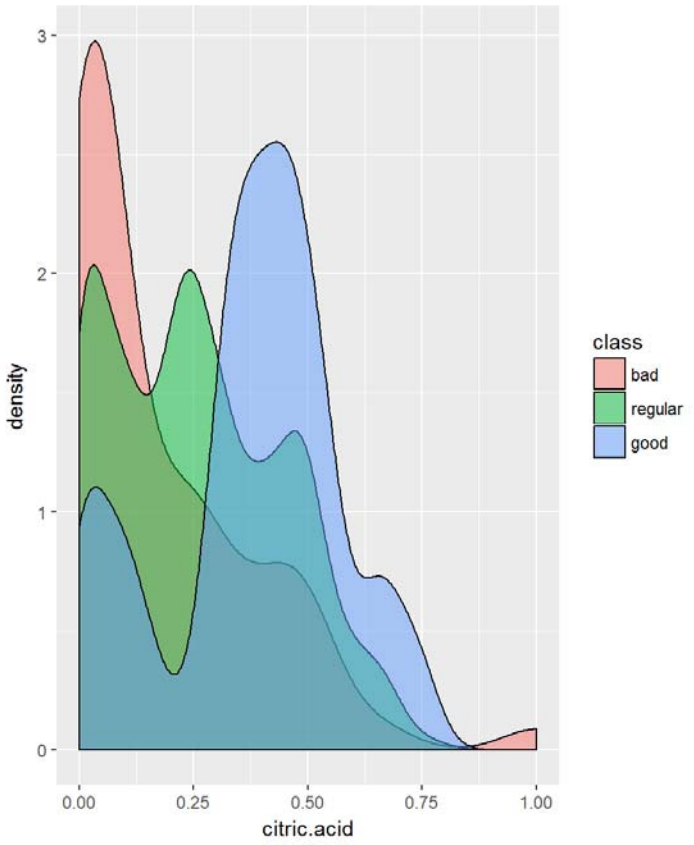
Again, per our observation we note that higher quality wines have **higher** levels of sulphates

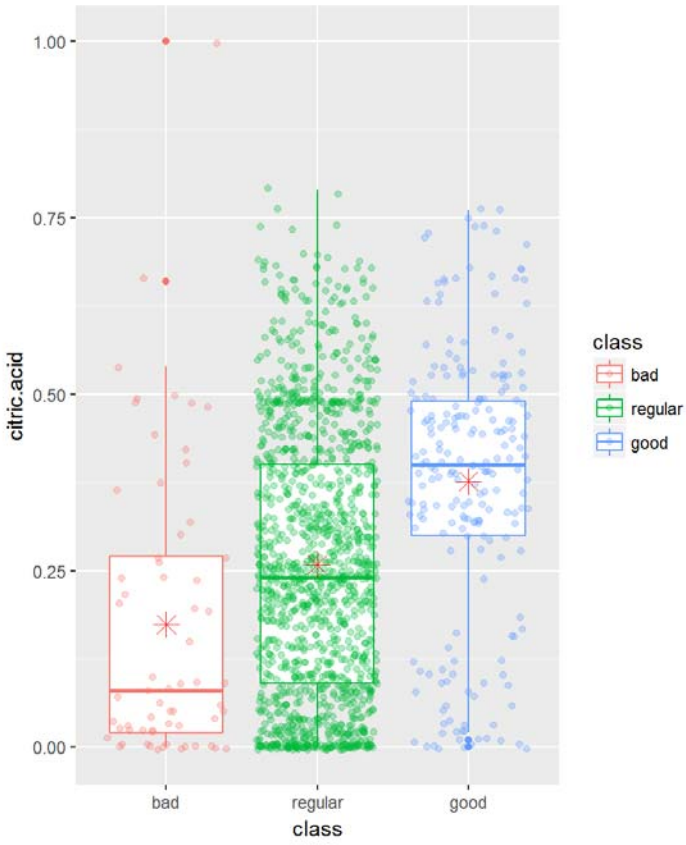
# 3.3) volatile acidity vs Class




This plot helps us note that high quality wines have **less** amount of volatile acidity.

# 3.4) citric acid vs class

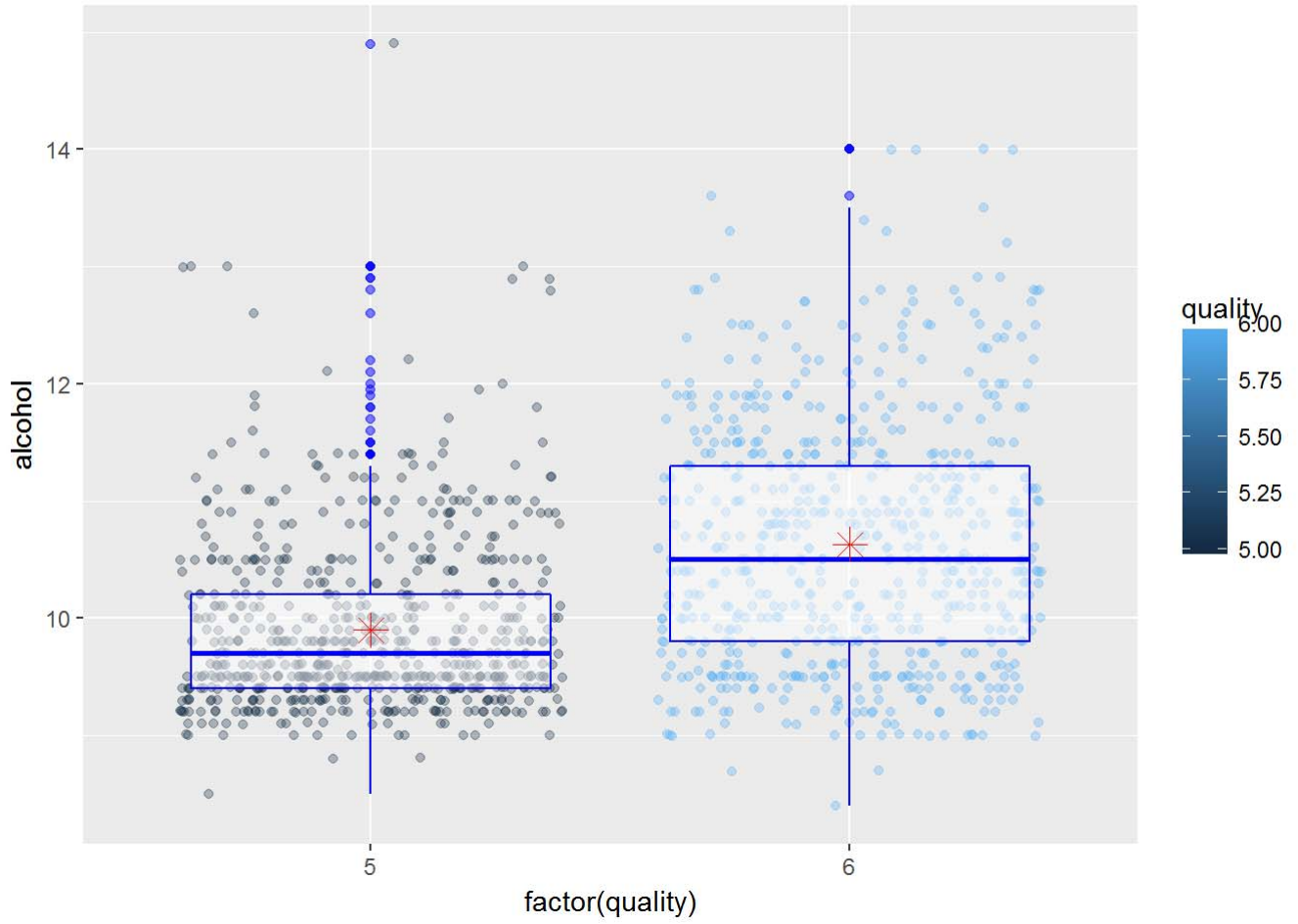



Again, per our observation we note that higher quality wines have **higher** levels of citric acid

# 3.5) Further dive into Alcohol vs quality

Let's take a look at the regular wines (having quality of 5 and 6) to figure out if there is any alcohol level difference that sets them apart.




There is a jump for alcohol variable between qualities 5 and 6. Maybe this is a separation between potentially bad wines and potentially good wines.

# 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?

Based on our previous analysis, we have been checking some correlations. We were able to explore some tendancy across the quality for: “volatile.acidity”, “citric.acid”, “sulphates” and “alcohol”. All the cases except “volatile.acidity” are positive correlations. This is normal, because “volatile.acidity” is the concentration of acetic acid in wine, which present in too much concetration can lead to a sharp vinegar taste. For values of 5 in the “quality” variable the values for “alcohol” are very spread, although the tendency is that good wines (quality 7 or 8) have the highest median level of alcohol.

Furthermore, correlation matrices have given us a global overview of all pairwise relations in a numerical and graphical ways.

## Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

I was surprised to see that pH and volatile acidity are positively correlated, since a higher pH value means less acidity.

As expected, citric acid, acidity, and pH are all rather correlated, given that they all measure acidity.

Lastly, I was wrong in assuming that residual sugar may have a significant impact in the quality of the wine. Infact, it hardly contributes towards quality.

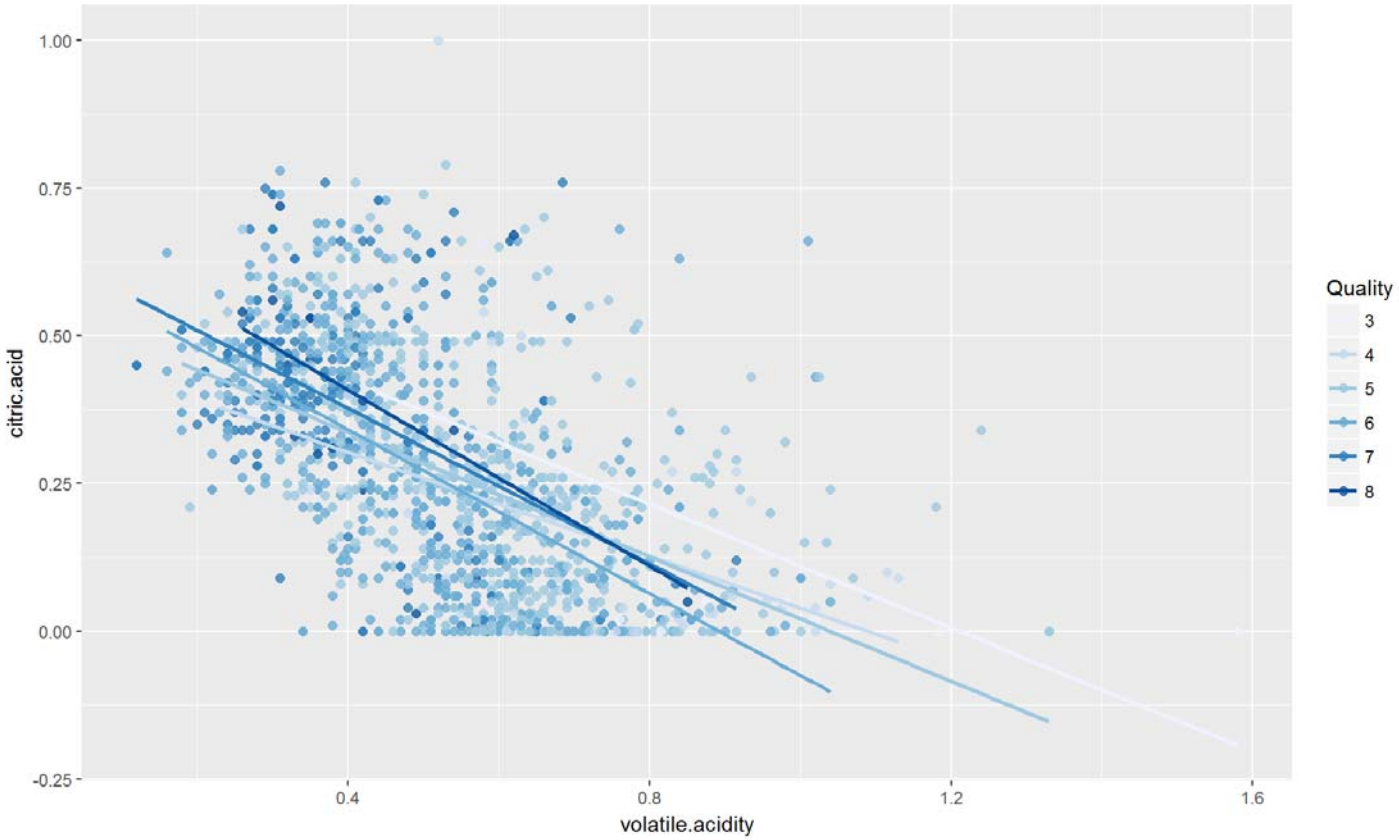
## What was the strongest relationship you found?

The strongest relationship, ignoring that between “total.sulfur.dioxide” and “bound.sulfur.dioxide”, is the negative correlation (-0.68) between “fixed.acidity” and “pH”. The best positive correlation (0.76) was founf between “alcohol” and “quality”.

# Multivariate Plots Section

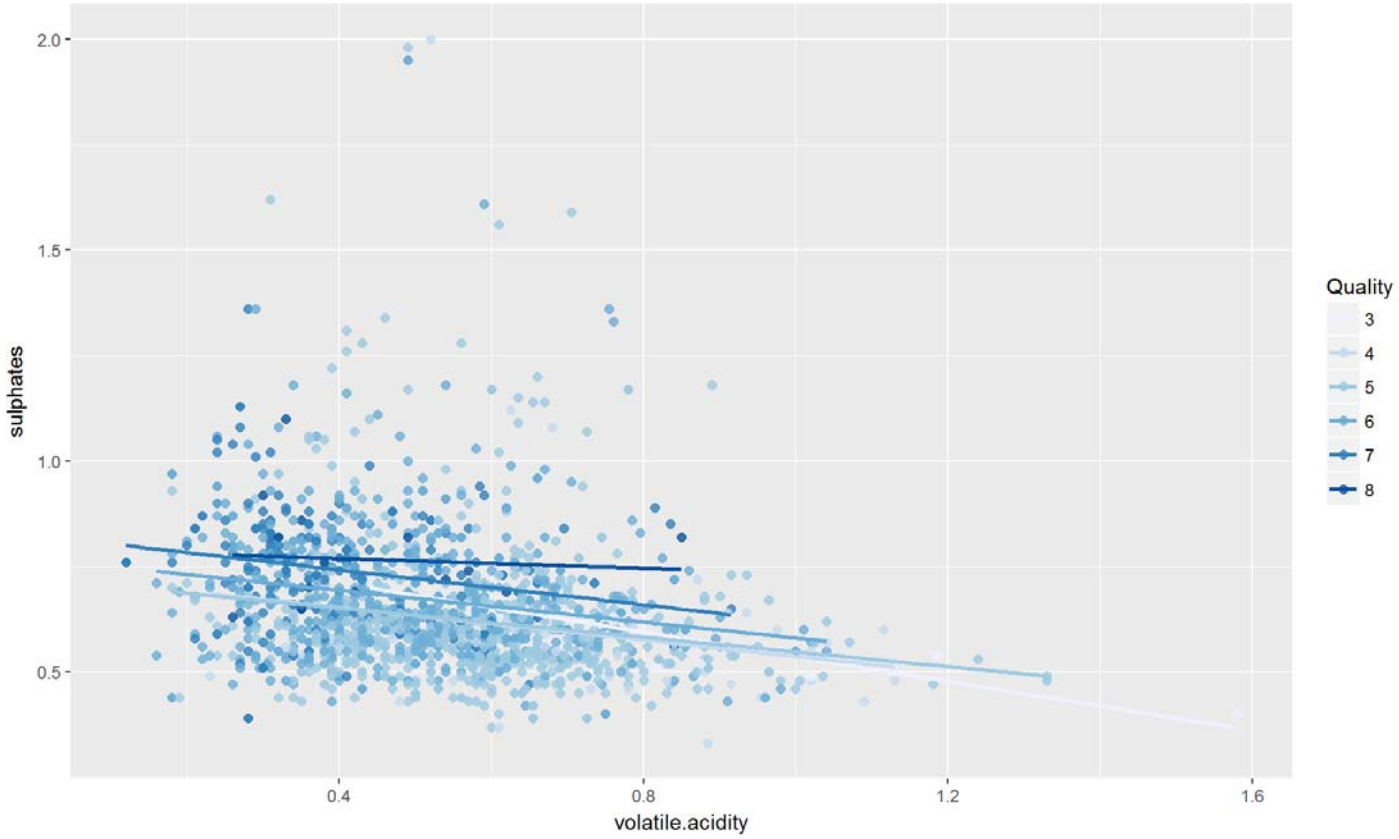
## 1) For this part, we will focus on our 4 main features we explored in the earlier plots and come up with a predictor model for quality.

### a) Citric acid vs volatile acidity



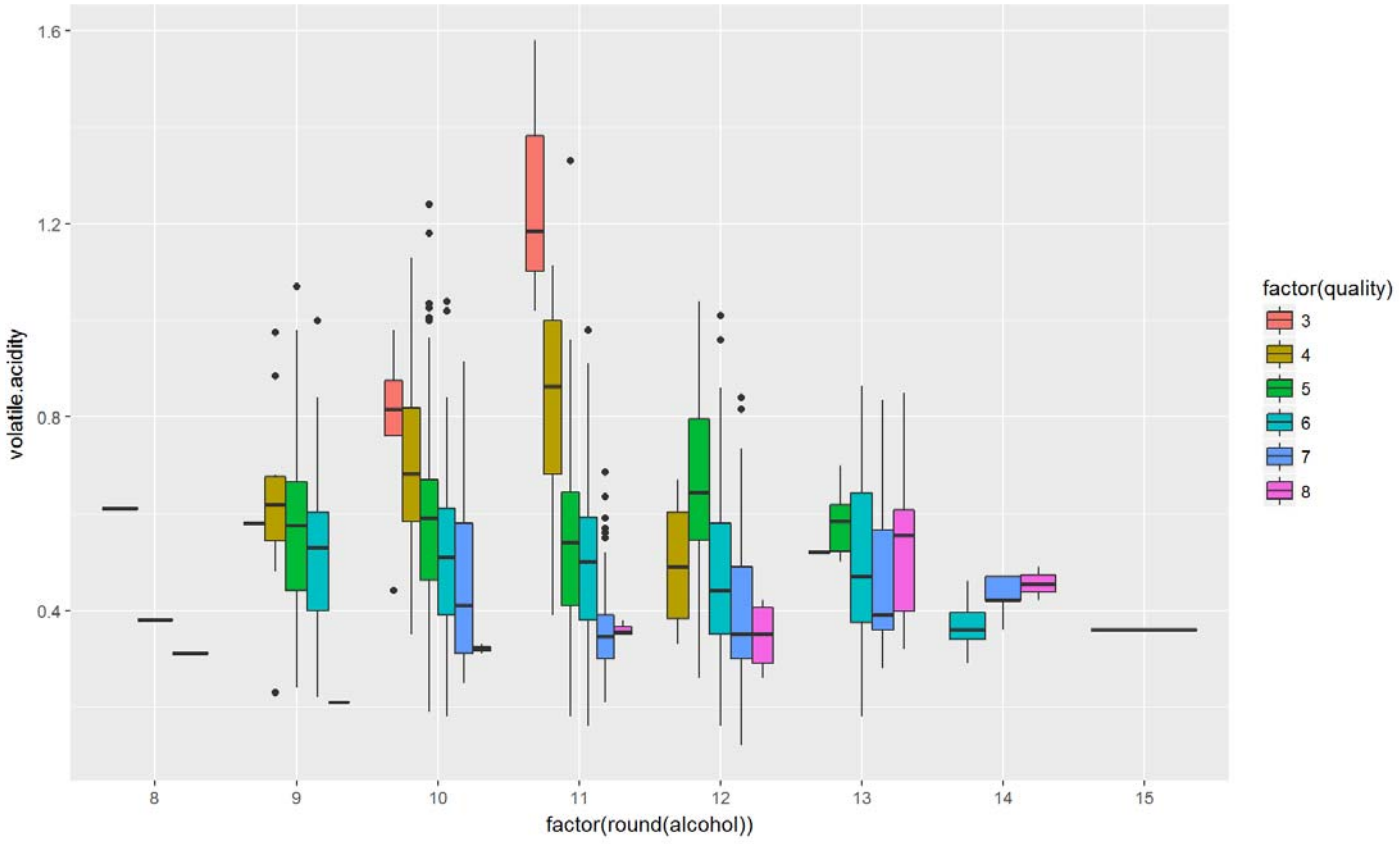

In the citric acid and volatile acidity we can see that the relationship is inversely proportional. The amount of volatile acidity decreases as the level of citric acid increases in the wine.

### b) volatile acidity vs sulphates



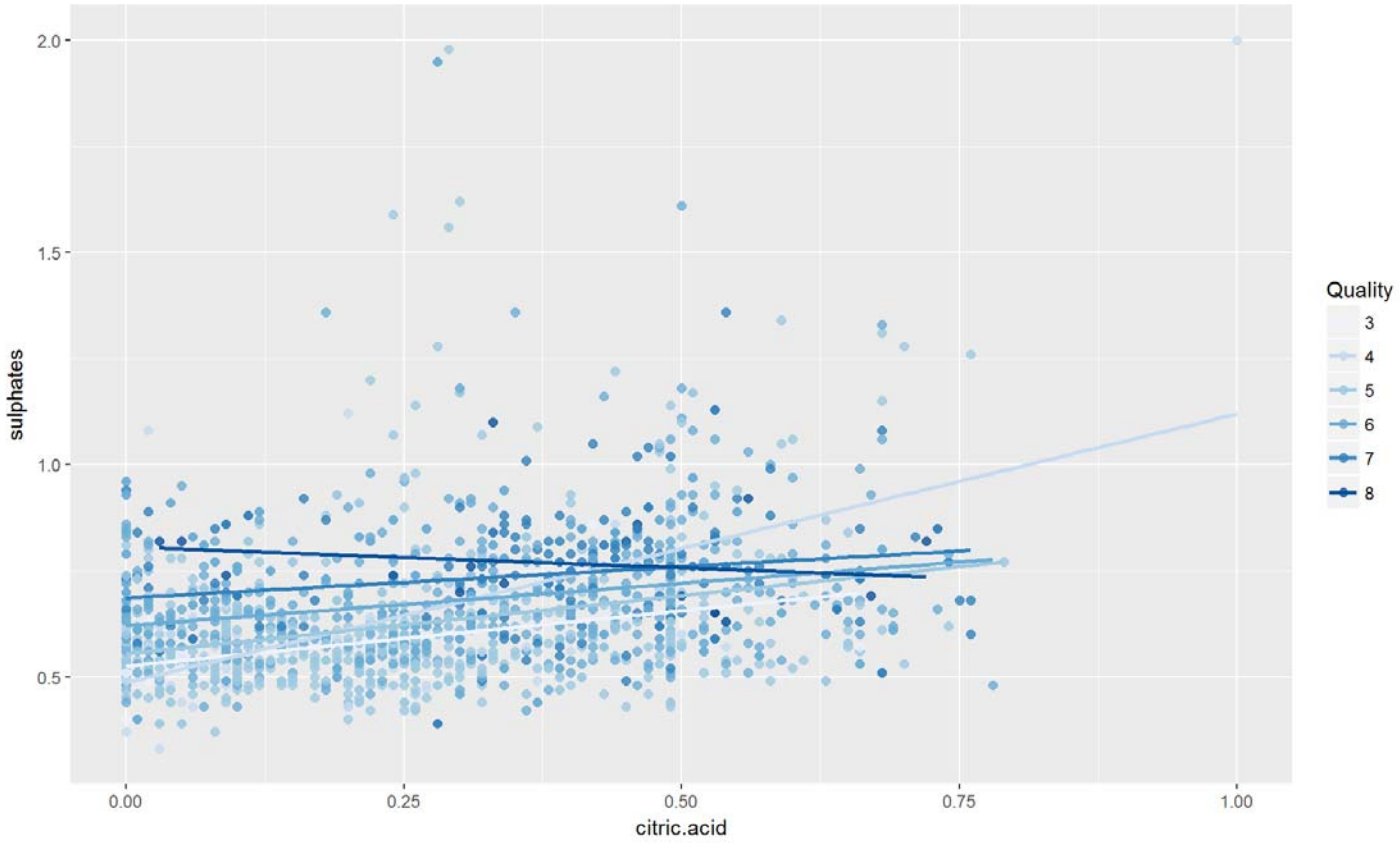

Here we note that the volatile acidity decreases as the level of sulphates increases. However, for good quality wines the levels remain more or less the same.

## c) volatile acidity vs alcohol



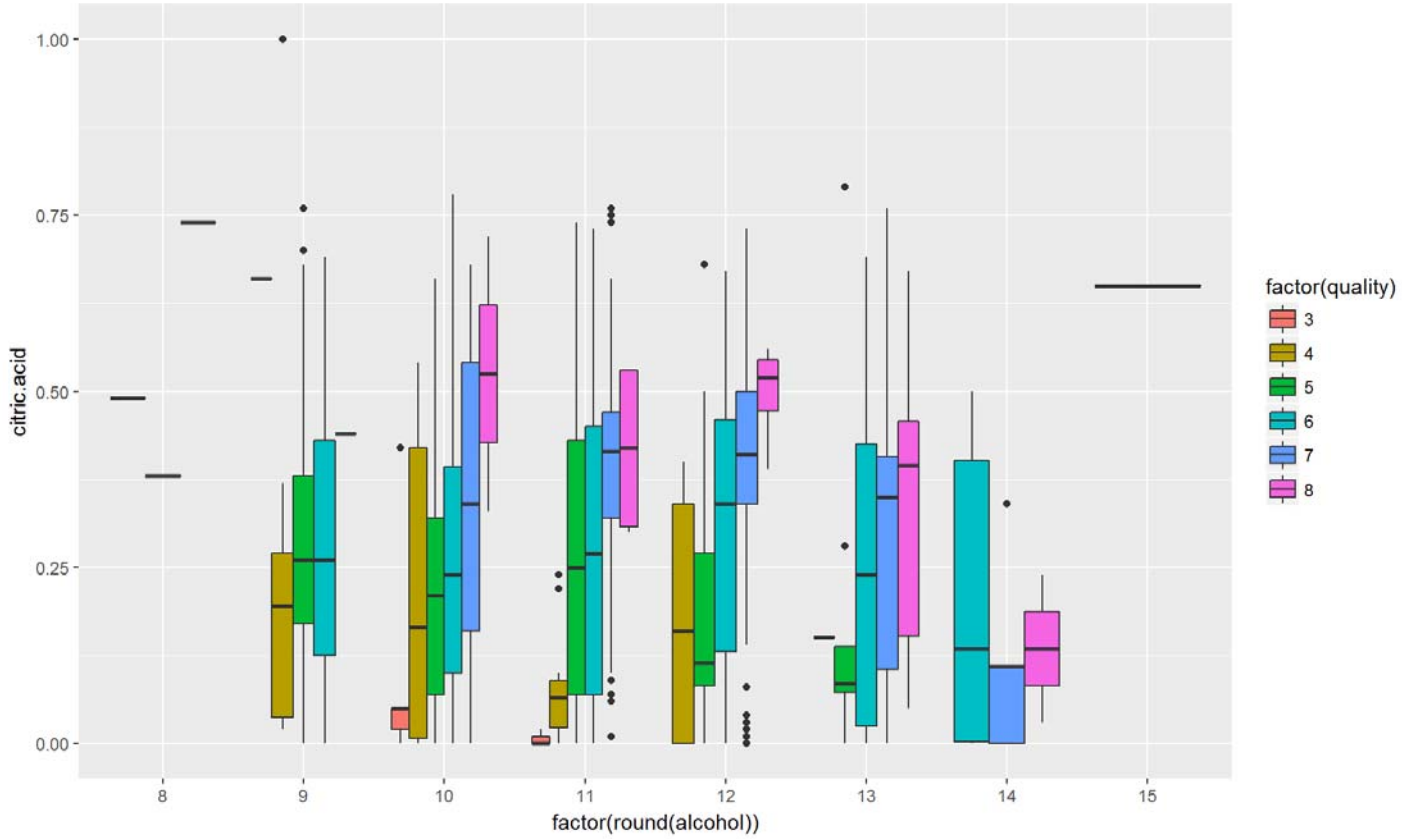

For alcohol between 9% to 12% we note that the level of volatile acidity decreases as the level of alcohol in the wines increase. For higher alcohol content (>12% ) ranges from 0.4g/dm^3 to 0.6g/dm^3.

## d) citric.acid vs sulphates



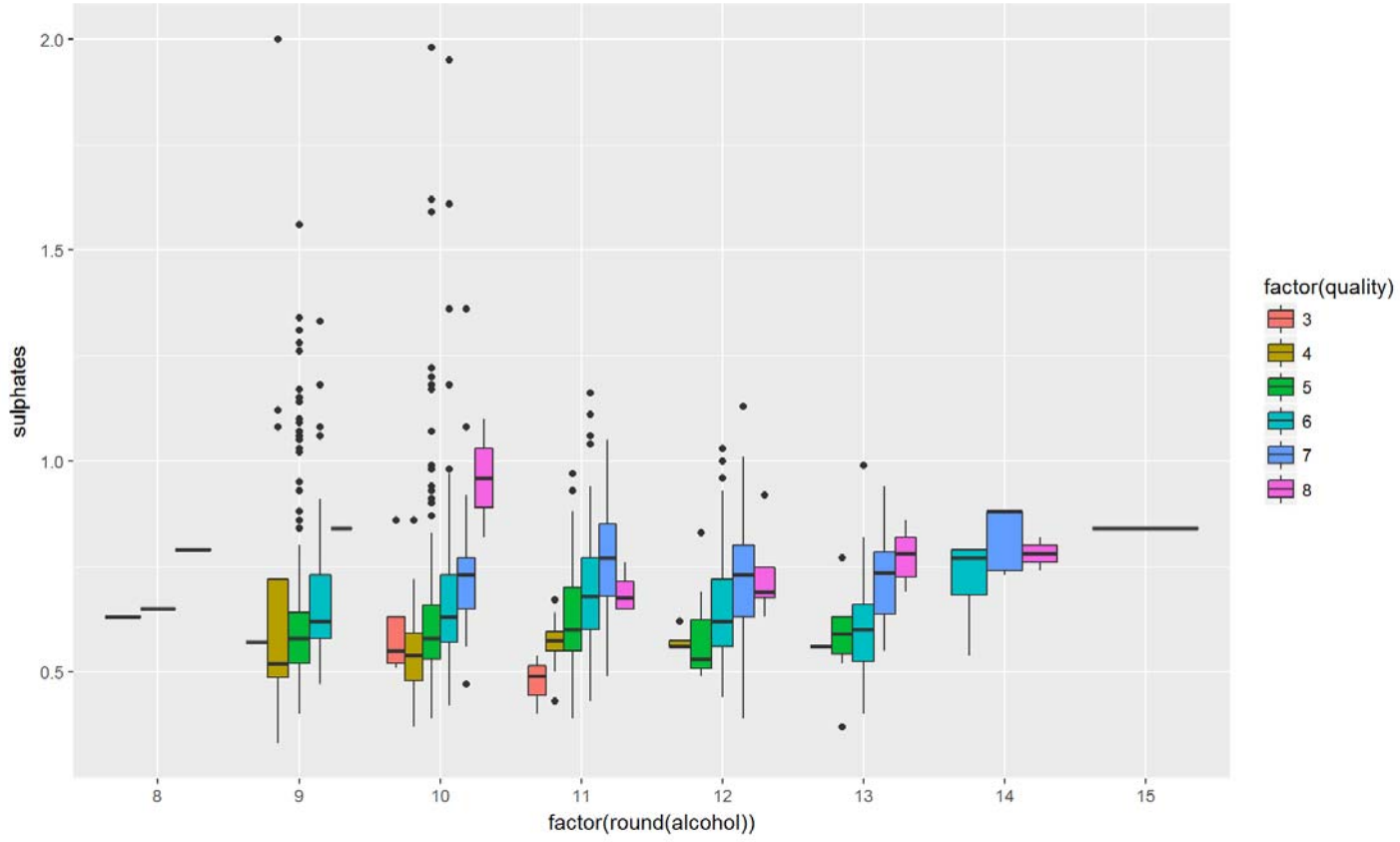

We can see that for lower quality wines the relationship is linear between sulphates and citric acid. However, high quality wines(8) more or less tend to have a flat relationship.

## e) citric.acid vs alcohol




Here we note that for each aclohol percentage level, as the quality of the wine increases there is a slight rise in the level of citric acid. The only exception was found for alcohol 14% where the citric acid level drop when the quality of the wine increases.

## f) sulphates vs alcohol



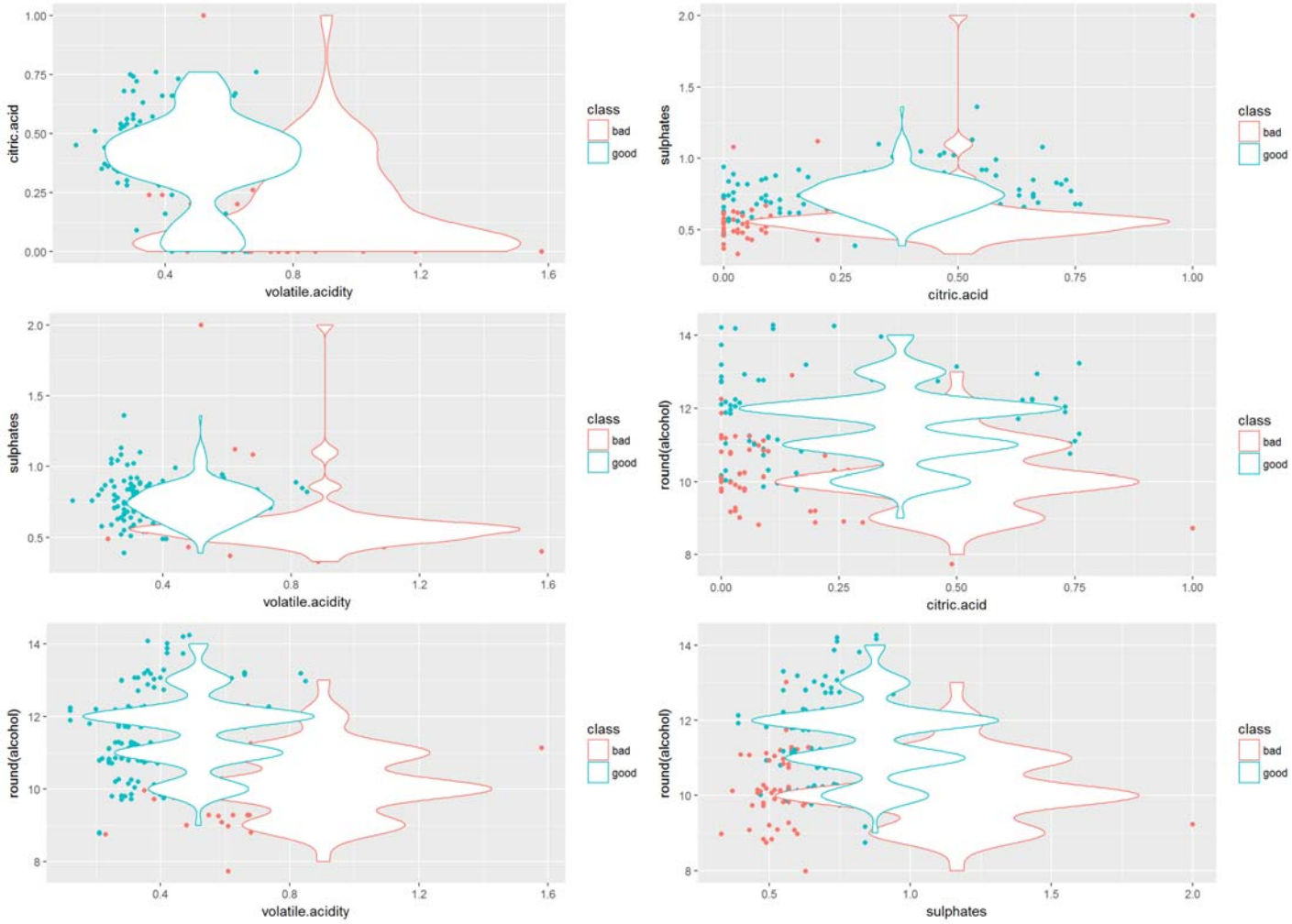
Here we note that for each percentage level of alcohol in the wine, the quantity of sulphates increases as the quality of the wine increases.

# 2) Data Modelling

For prediction purposes, we have two main problems:

- Unbalanced spread of quality feature (too many regular wines)
- The regular wines are very spread across feature values, so they are mixed with bad and good classes.

Maybe what we should try is to predict good (or bad) wines, not to try to classify into the three classes. Lets check only bad wines against good wines. In this case, we also add some density 2D maps in order to see where are located the clusters or groups for each combination of features:



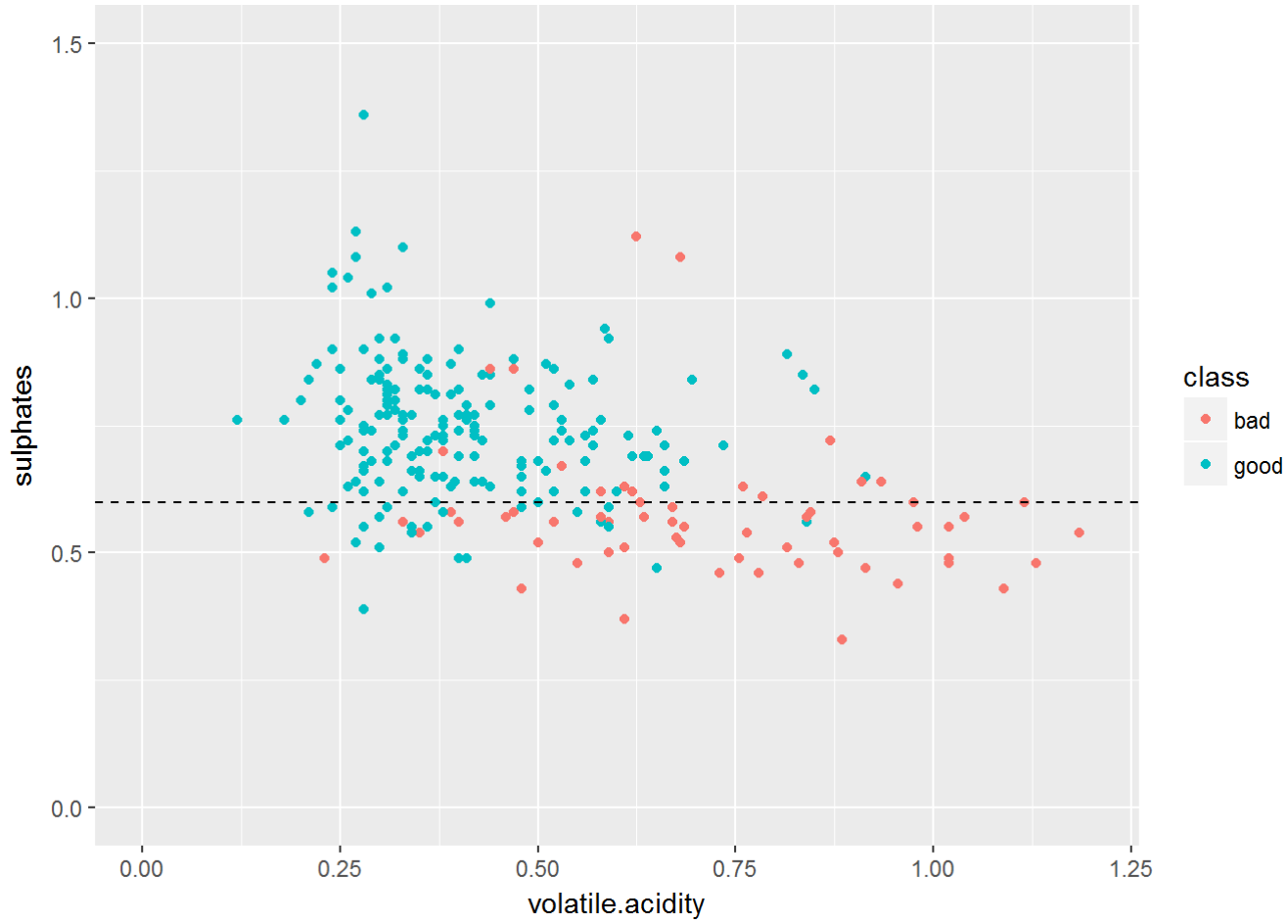

We can see more clear distinct differences after removing the “regular” quality wines in our data and helps focus on the trends specifically.

Below are our observations from the plot:

- volatile acidity vs citric acid
    - Good wines: low volatile acidity and medium citric acid
    - Bad wines: Both medium
- volatile acidity vs sulphates
    - Good wines: low volatile acidity and low sulphates

    - Bad wines: medium to high volatile acidity and low sulphates
- volatile acidity vs alcohol
    - Good wines: low volatile acidity and high alcohol
    - Bad wines: medium volatile acidity and low alcohol
- citric acid vs sulphates
    - Good wines: low to medium citric acid and low sulphates
    - Bad wines: low to medium citric acid and low sulphates
- citric acid vs alcohol
    - Good wines: medium citric acid and high alcohol
    - Bad wines: low to medium citric acid and low to medium alcohol
- sulphates vs alcohol
    - Good wines: low to medium sulphates and high alcohol
    - Bad wines: low sulphates and low to medium alcohol

For the combination “citric.acid” vs “sulphates”, we see more or less an horizontal line separating good and bad wines:




Let’s build a simple linear model using our four main features(alcohol,volatile.acidity,sulphates and citric.acid).

```
##
## Calls:
## m1: lm(formula = I(quality ~ alcohol), data = red_wine)
## m2: lm(formula = quality ~ alcohol + volatile.acidity, data = red_wine)
## m3: lm(formula = quality ~ alcohol + volatile.acidity + sulphates,
##     data = red_wine)
## m4: lm(formula = quality ~ alcohol + volatile.acidity + sulphates +
##     citric.acid, data = red_wine)
##
## ================================================================
##                         m1         m2         m3         m4
## ----------------------------------------------------------------
##   (Intercept)        1.875***   3.095***   2.611***   2.646***
##                     (0.175)    (0.184)    (0.196)    (0.201)
##   alcohol            0.361***   0.314***   0.309***   0.309***
##                     (0.017)    (0.016)    (0.016)    (0.016)
##   volatile.acidity             -1.384***  -1.221***  -1.265***
##                                (0.095)    (0.097)    (0.113)
##   sulphates                                0.679***   0.696***
##                                           (0.101)    (0.103)
##   citric.acid                                        -0.079
##                                                      (0.104)
## ----------------------------------------------------------------
##   R-squared              0.2        0.3        0.3        0.3
##   adj. R-squared         0.2        0.3        0.3        0.3
##   sigma                  0.7        0.7        0.7        0.7
##   F                    468.3      370.4      268.9      201.8
##   p                      0.0        0.0        0.0        0.0
##   Log-likelihood     -1721.1    -1621.8    -1599.4    -1599.1
##   Deviance             805.9      711.8      692.1      691.9
##   AIC                 3448.1     3251.6     3208.8     3210.2
##   BIC                 3464.2     3273.1     3235.7     3242.4
##   N                   1599       1599       1599       1599
## ================================================================
```

We can see that adding the “sulphates” adds small improvement but “citric.acid” does not improve the model(we saw this from our plots). The model is not such a good one as the R_square value is low(0.3 for model 3). Let’s check the accuracy of the model:

Successful prediction by quality (0-10):

```
## [1] 0.5822389
```

Successful prediction by class:

```
## [1] 0.833646
```

Let’s see an example to predict the wine quality for the following features - alcohol= 11, volatile.acidity = 0.6 , sulphates= 0.7

The predicted wine quality is for a wine with composition - alcohol= 11, volatile.acidity = 0.6 , sulphates= 0.7 is :

```
## [1] 6
```

If we use rounded predicted quality values then our accuracy is 58% of the qualities. But if we use quality classes(bad, regular and good), then we increase the accuracy to 83%.

# 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?

In this section we compared our four features(volatile.acidity,citric.acid,sulphates and alcohol) in pair plots, taking into account different classes of wine. The plots revealed a large spread of points for regular quality(5 and 6) wines;also there is no clear cut off margin between bad and regular wines as well as regular and good quality wines. After taking a subset of data containing only bad and good wines, we were able to see more distinguishable characteristics from one another as we saw in the plot.

We noted that most of the good wines have medium values of citric acid and low values of volatile acidity. Bad wines usually have medium-high volatile acidity and low citric acid. This is similar for combinations of “volatile.acidity” with “sulphates” or “alcohol”: good wines are upper left and bad wines are lower right. This tendency is similar in “alcohol” vs “citric.acid” or “sulphates”, although in this case good wines are on the upper right and bad wines on the lower left. For the combination “citric.acid” vs “sulphates”, we see more or less a horizontal line separating good and bad wines.

## Were there any interesting or surprising interactions between features?

Yes, based on the bivariates plot, it seems that there is a positive correlation between “citric.acid” and “quality”. But if we observe the scatter plots by class of wine (only good and bad), we do not see a clear cutoff of “citric.acid” feature to distinguish good and bad wines.

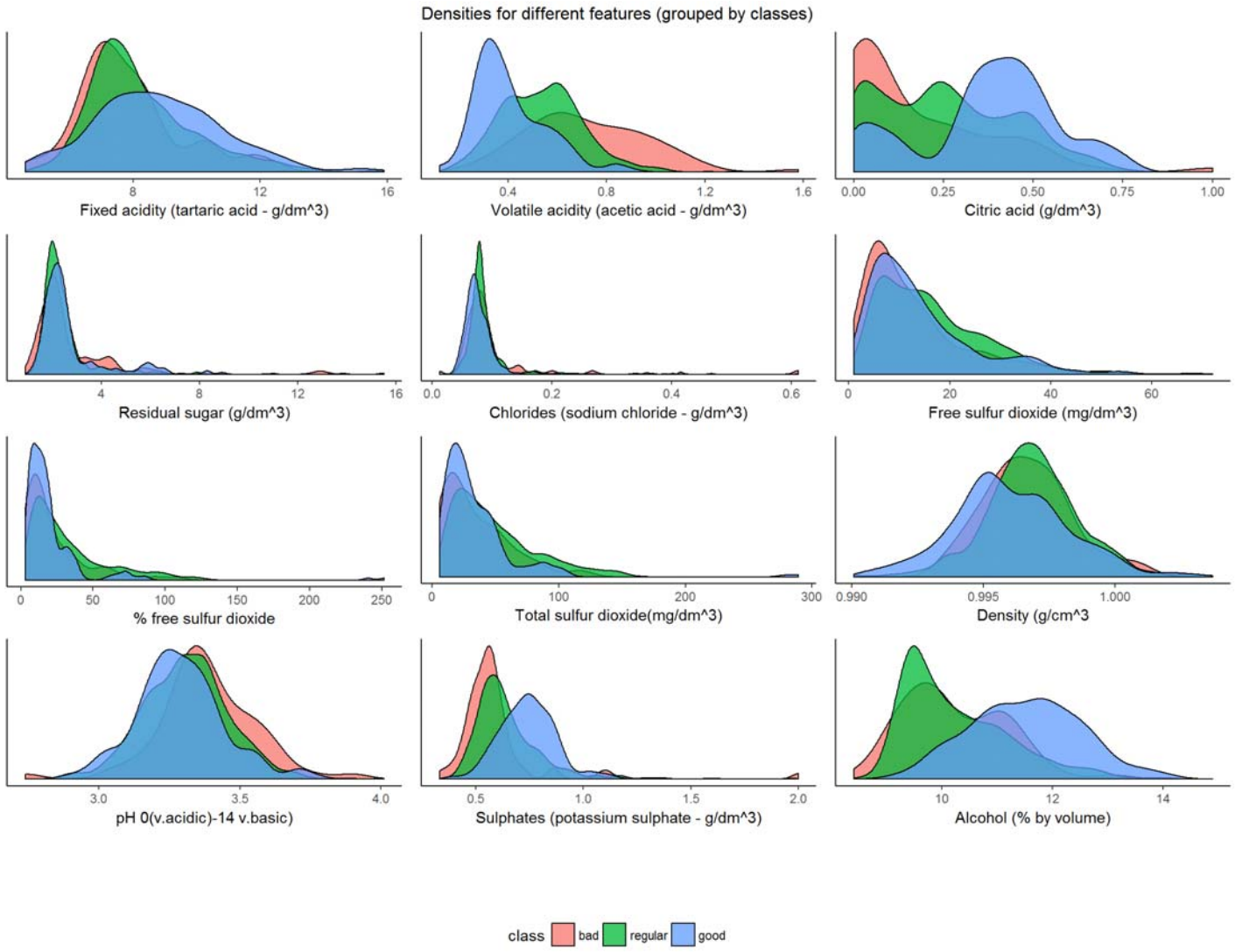
## OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model.

I created simple linear models using our four main features. The first model includes only “alcohol” as predictor. Then next model add “volatile.acidity”. Model 3 adds also “sulphates”, and the last model adds “citric.acid”. The R2 values of our models are not very good, the accuracy could also be a little misleading as we have a very unbalanced dataset (too many “regular” wines). Maybe the biggest problem for the model is to distinguish between bad and regular wines, and between good and regular wines.

---

# Final Plots and Summary

## Plot One

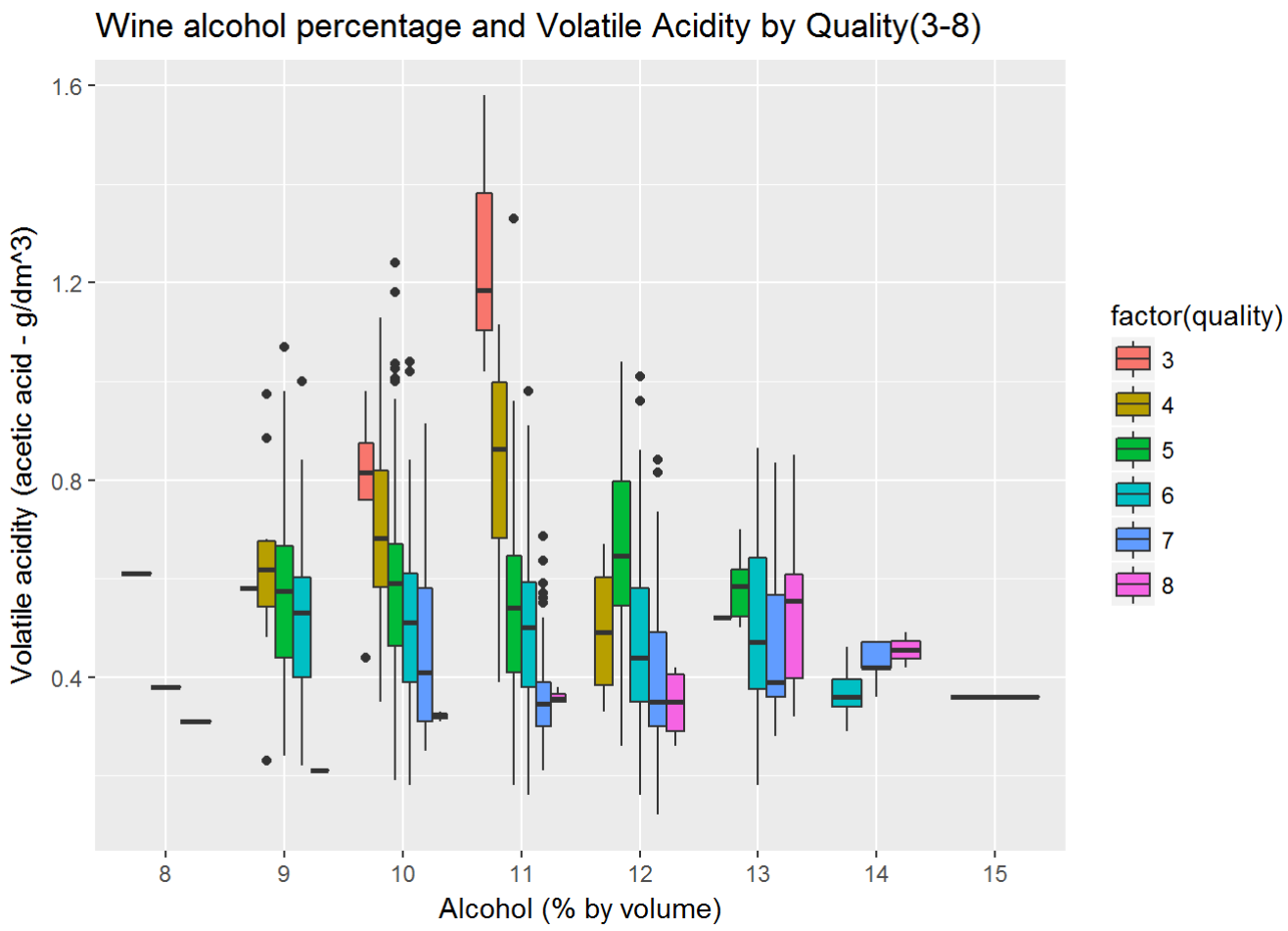



# Description One

This plot shows the densities for the distributions of all features in the dataset. We created three quality classes for wine: bad (4 or lower quality values) in red, regular (5 or 6) in green and good (7 or higher) in blue and grouped the features. Those variables with less overlapping in their density curves could help us to distinguish between quality classes. Four of the best features for this purpose are: volatile acidity, citric acid, sulphates and alcohol. Other variables also could help us to detect a specific class, like fixed acidity (good wines) and % free sulfur dioxide (regular wines).

Note: text, values and ticks of Y-axis were removed for clarity

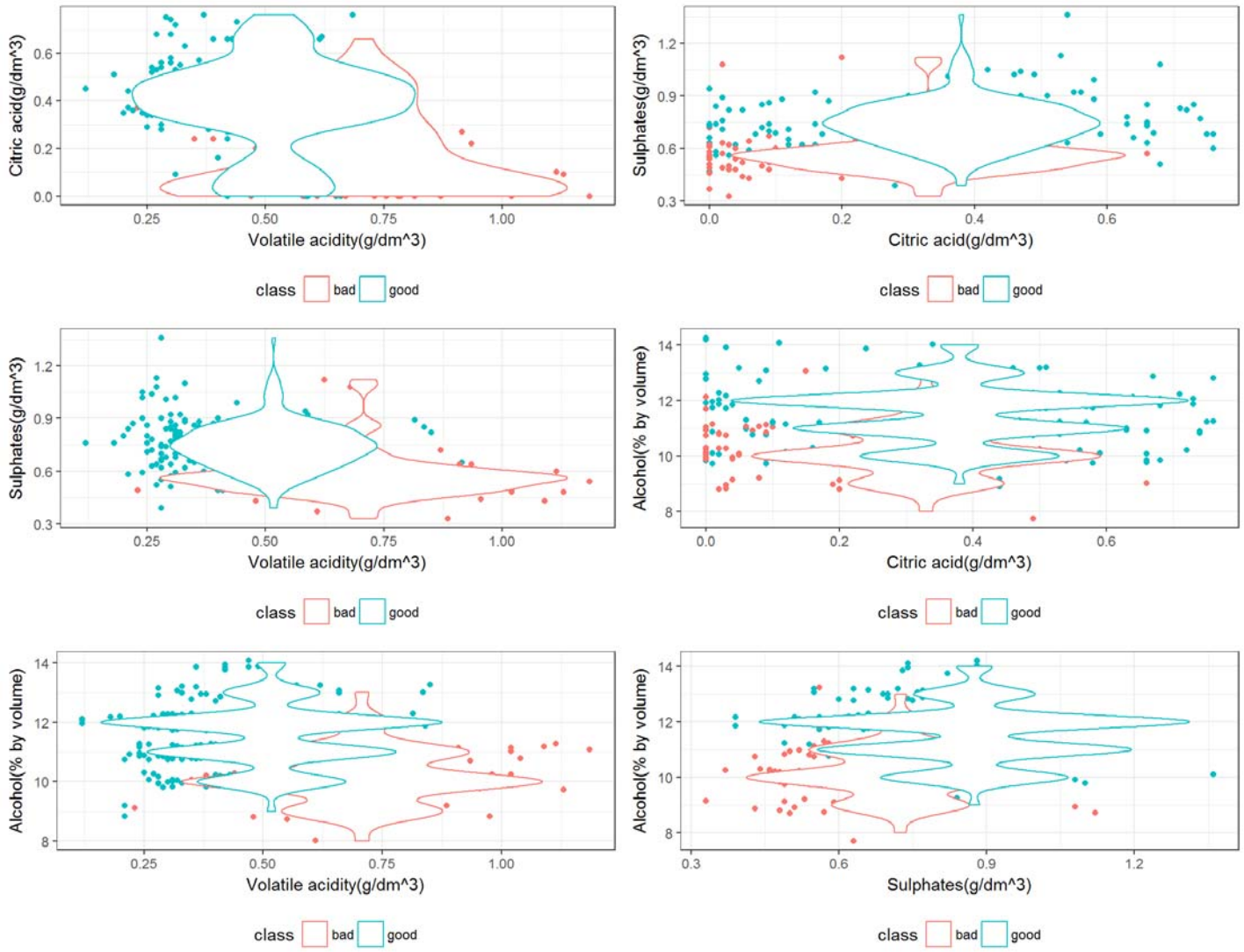
# Plot Two




# Description Two

Alcohol by volume and volatile acidity were the two most influential chemical properties related to quality in red wine based on our research. Alcohol had a positive relationship with quality, perhaps due to a higher concentration of flavor in wines with higher alcohol percentages. Volatile acidity had a negative relationship with quality rating, due to the fact that higher concentrations can lead to undesirable flavors. As shown by the the plot, the lowest quality wines tend to have lower alcohol percentages and higher volatile acidity concentrations, while the higher quality wines had higher alcohol percentages and lower volatile acidity concentrations.

# Plot Three




Pairwise comparison of the 4
influential features, with density maps (grouped by class)

## Description Three

This pair plot was influenced from the correlation we found from our correlation plot. We narrowed our focus to the 4 main identified features that contributes to the quality of the wine. We did a comparison of each feature to another. We filtered out the regular wines to make clear distinguishable patterns between good and bad wine. We also deleted some outliers for volatile acidity, citric acid and sulphates.

The idea is to show that these features could help to distinguish good wines from bad wines. This is beacuse when selecting a wine we are more conscious about picking up a good wine over a bad one.

# Reflection

We have been analysing a red wine dataset with almost 1599 observations with 13 variables. The target feature was quality of the wine which ranges from 1 to 10. The objective was to analyse the other features to know their influence in wine quality. After observing different distributions for the features, taking into account the qualities, we determined four of the features as the most influential: volatile acidity, citric acid, sulphates and alcohol. After grouping the qualities in three classes (bad, regular and good), we saw that there was a correlation with the main features. This correlation is positive in all cases, except for volatile acidity whose correlation is negative.

The main struggles came from the fact that our data was unbalanced. Majority values of quality of the wine data was for 5 and 6. This made it difficult to analyze scatter plots due to overplotting. I was able to get better results using boxplots which helped unlock relations between different features. The other problem due to large number regular quality wines was that it was difficult to find the differences in composition of bad, regular and good quality wines. I was able to get through this problem by removing the data for regular wines and observe plots for bad vs good wines. This helped a lot in indentifying our best features to predict wine quality.

According to our study, good wines seem to have lower volatile acidity, higher alcohol and medium-high sulphate values. Bad wines tend to have low values for citric acid.

For the predictive model, we used a simple linear model with only one main feature, and then adding one by one the other 3 main features. Although the R2 is small, the accuracy are more or less high. But this is mainly because we have a problem of unbalanced data: too many “regular” class observations.

In order to improve our predictive model we need a more balanced data. Also, using parameter selection algorithms like Kbest select would help us understand the importance of each feature in our model. I would also try using different machine learning algorithms to see if there is improvement in the accuracy.