# Exploration of Wine Quality Red by Jieying Yan

This report explores the red wine quality and investigates the relation between the observed variables and the red wine quality.

The dataset used for the analysis includes 1599 observations and 12 variables.

# Univariate Plots Section

## Summary of Data

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00
##  Median :0.07900   Median :14.00       Median : 38.00
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00
##     density             pH          sulphates         alcohol
##  Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40
##  1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50
##  Median :0.9968   Median :3.310   Median :0.6200   Median :10.20
##  Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42
##  3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10
##  Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.636
##  3rd Qu.:6.000
##  Max.   :8.000
```

```
## 'data.frame':    1599 obs. of  12 variables:
##  $ 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 ...
```

The dataset is a tidy dataset with 1599 observations and 12 variables.

# Introduction to Terminology

As the variables include terminology which might not be known to everyone, in the following a brief introduction to a few terminology of wine quality:

- $ fixed.acidity:

> This is given by the difference between the total acidity and the volatile acidity. (Source: chestofbooks)

- $ volatile.acidity:

> Most of the acids involved with wine are fixed acids with the notable exception of acetic acid, mostly found in vinegar, which is volatile and can contribute to the wine fault known as volatile acidity. (Source: Wikipedia)

- $ citric.acid

> While very common in citrus fruits, such as limes, citric acid is found only in very minute quantities in wine grapes. It often has a concentration about 1/20 that of tartaric acid. The citric acid most commonly found in wine is commercially produced acid supplements derived from fermenting sucrose solutions. (Source: Wikipedia)

- $ residual.sugar

> Among the components influencing how sweet a wine will taste is residual sugar. (Source: Wikipedia)

- $ chlorides

> These may be metal salts containing chloride ion such as sodium chloride, or more covalent chlorides of metals or nonmetals such as titanium(IV) chloride or carbon tetrachloride. (Source: Wikipedia)

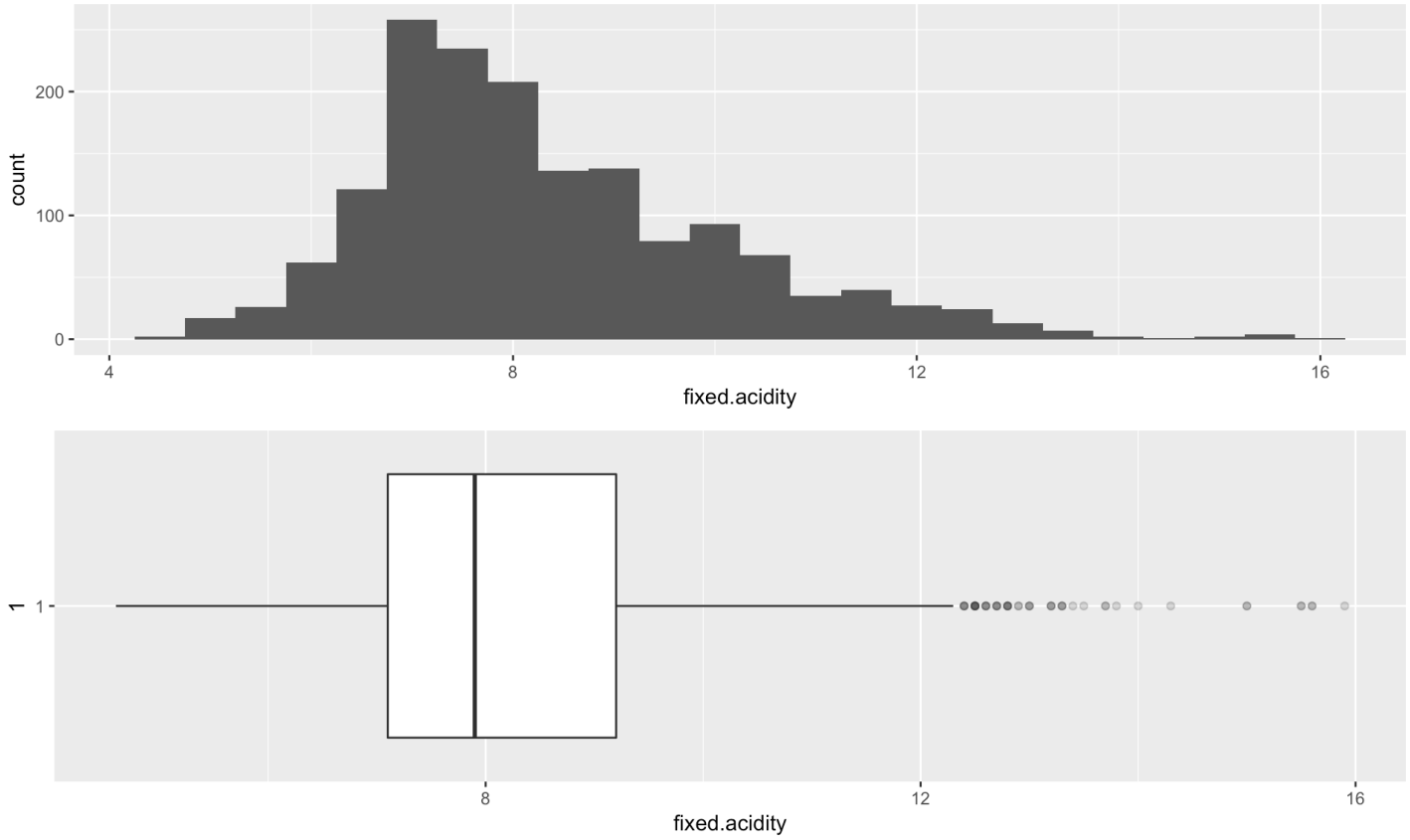
- $ free.sulfur.dioxide & $ total.sulfur.dioxide

> Sulfur dioxide was used by the Romans in winemaking when they discovered that burning sulfur candles inside empty wine vessels kept them fresh and free from vinegar smell. (Source: Wikipedia)

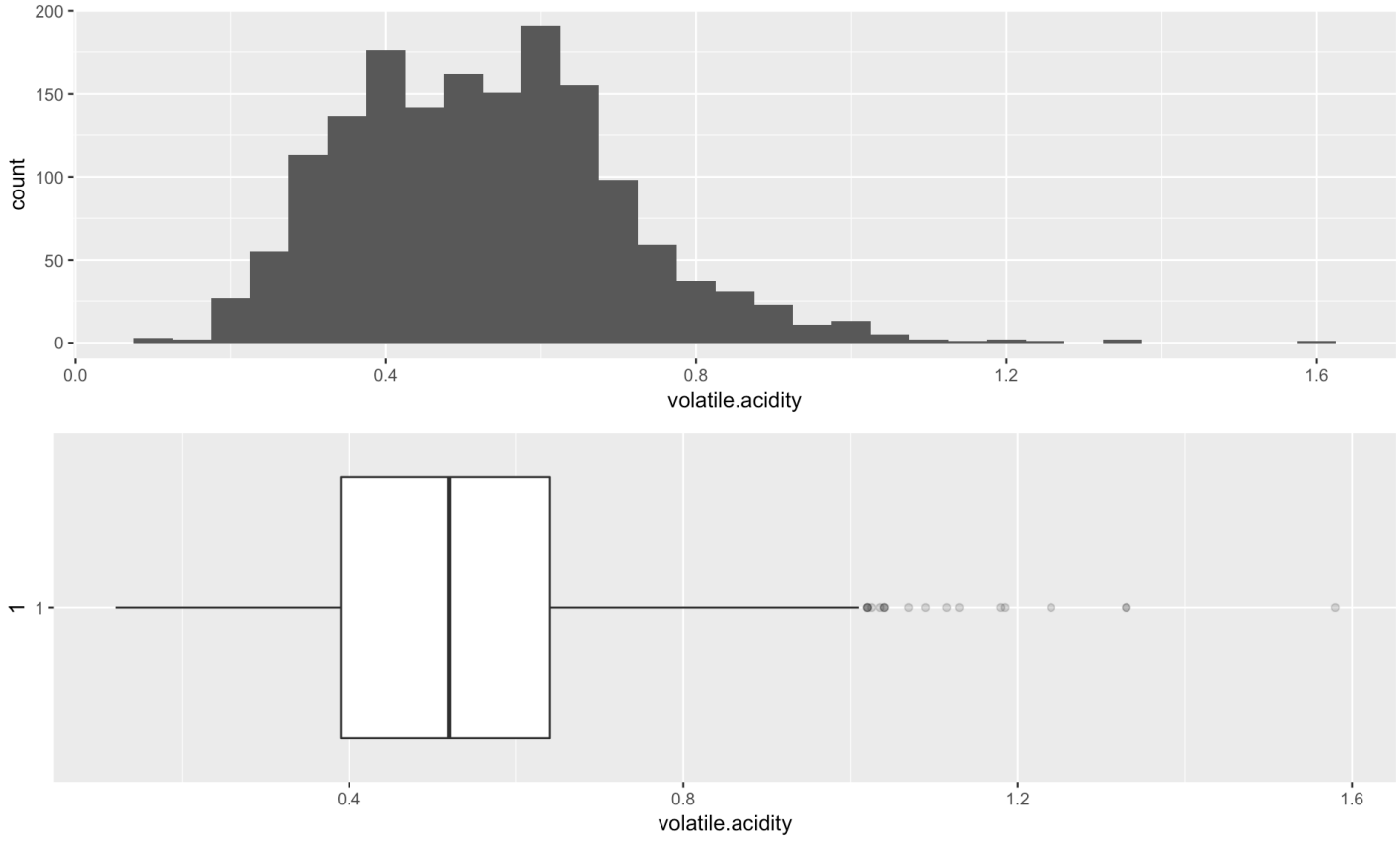
- $ sulphates

> The term ‘sulfites’ is an inclusive term for sulfur dioxide (SO2). SO2 is a preservative and widely used in winemaking (and most food industries), because of its antioxidant and antibacterial properties. SO2 plays a very important role in preventing oxidization and maintaining a wine’s freshness. (Source: thekichn)

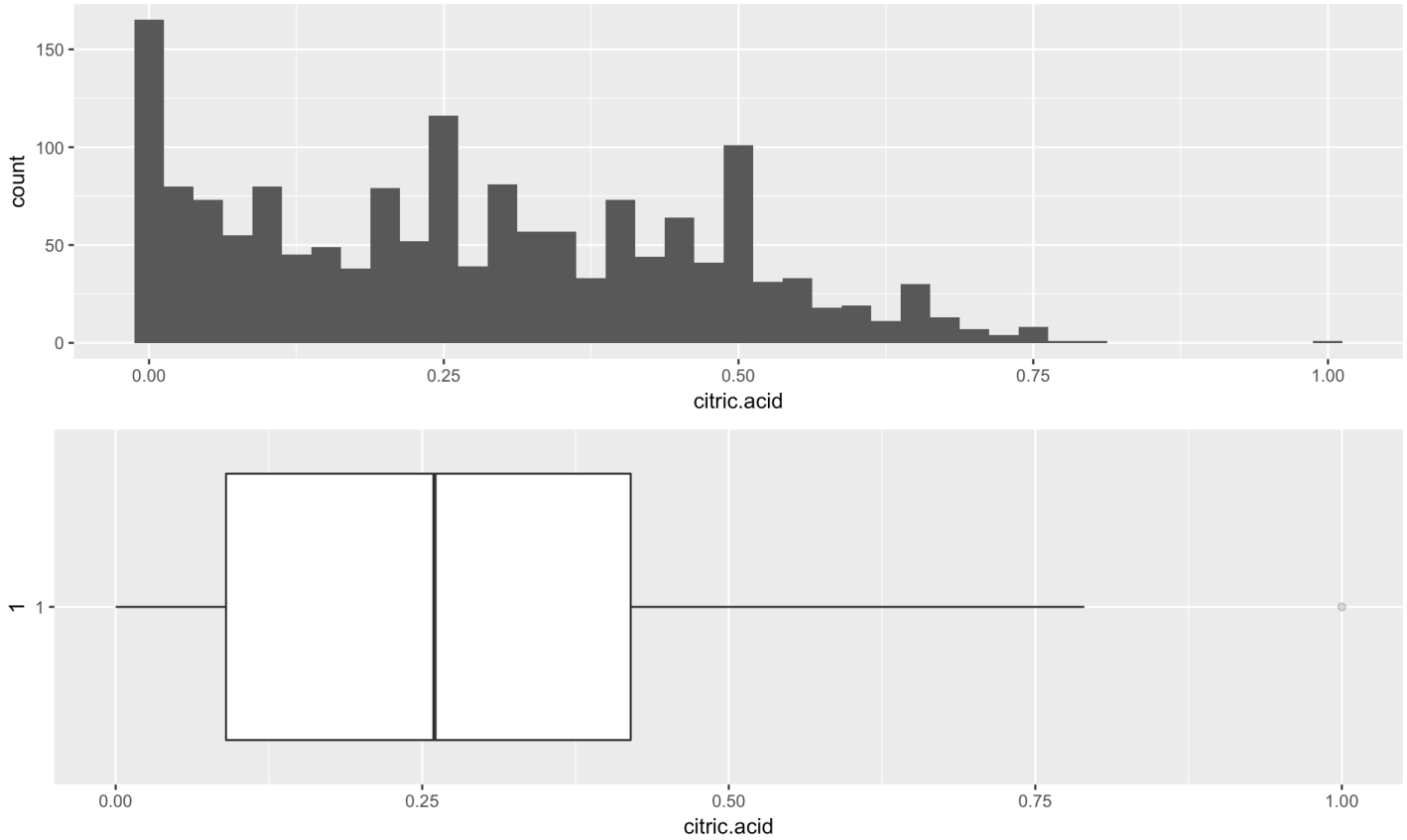
# Plotting of Variables



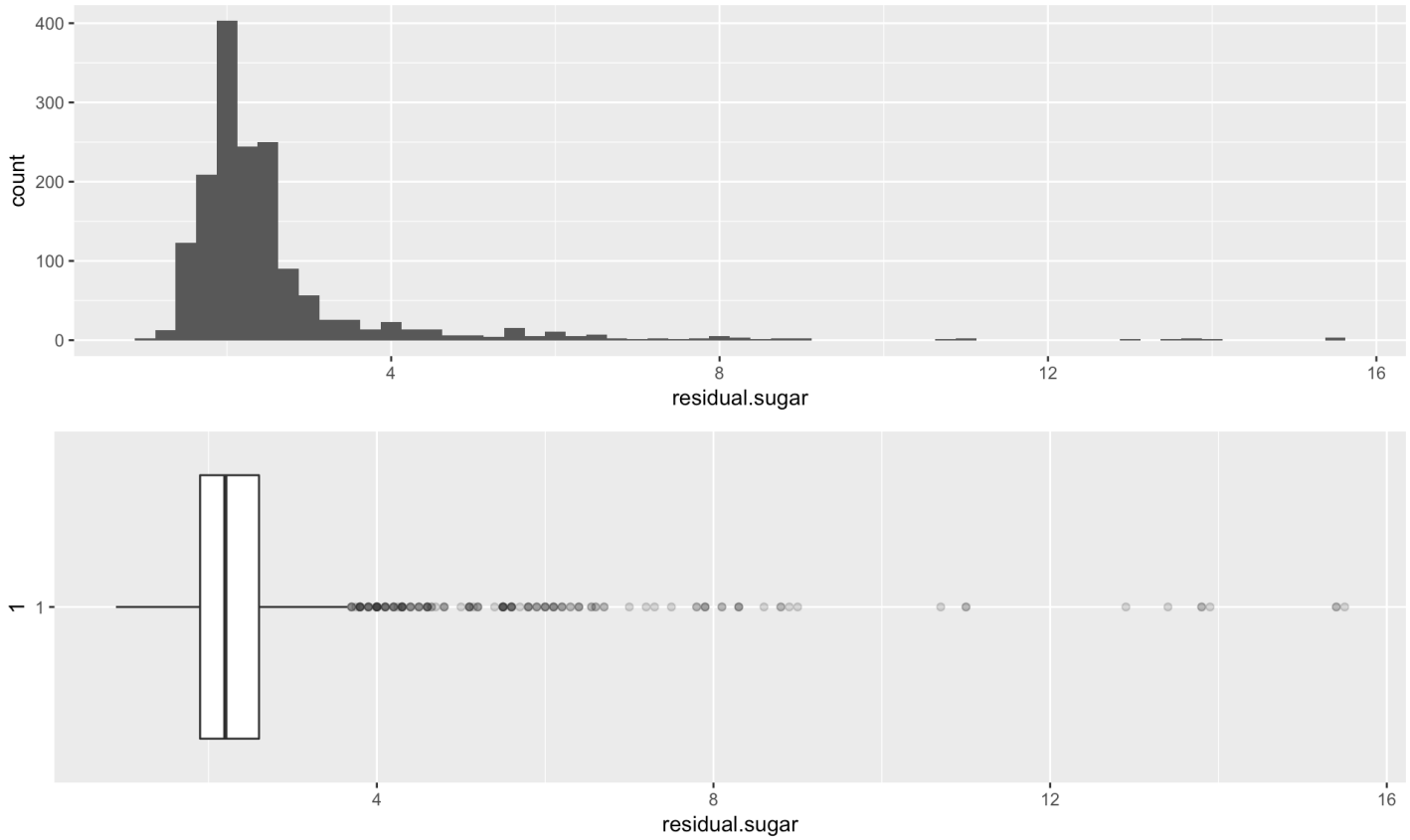

The distribution appears to be right skewed with outliers.



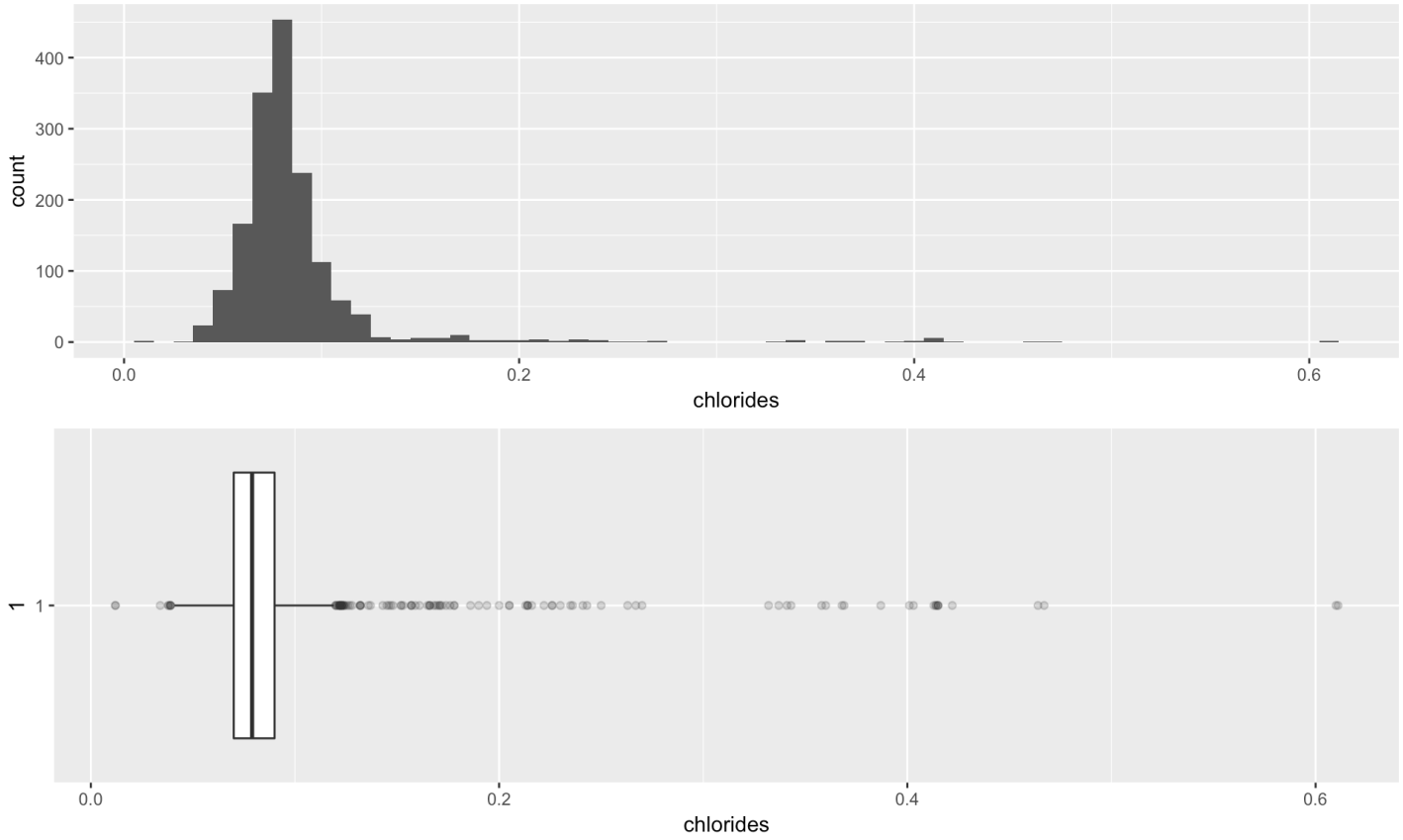

The distribution appears to be bimodal with ourliers.



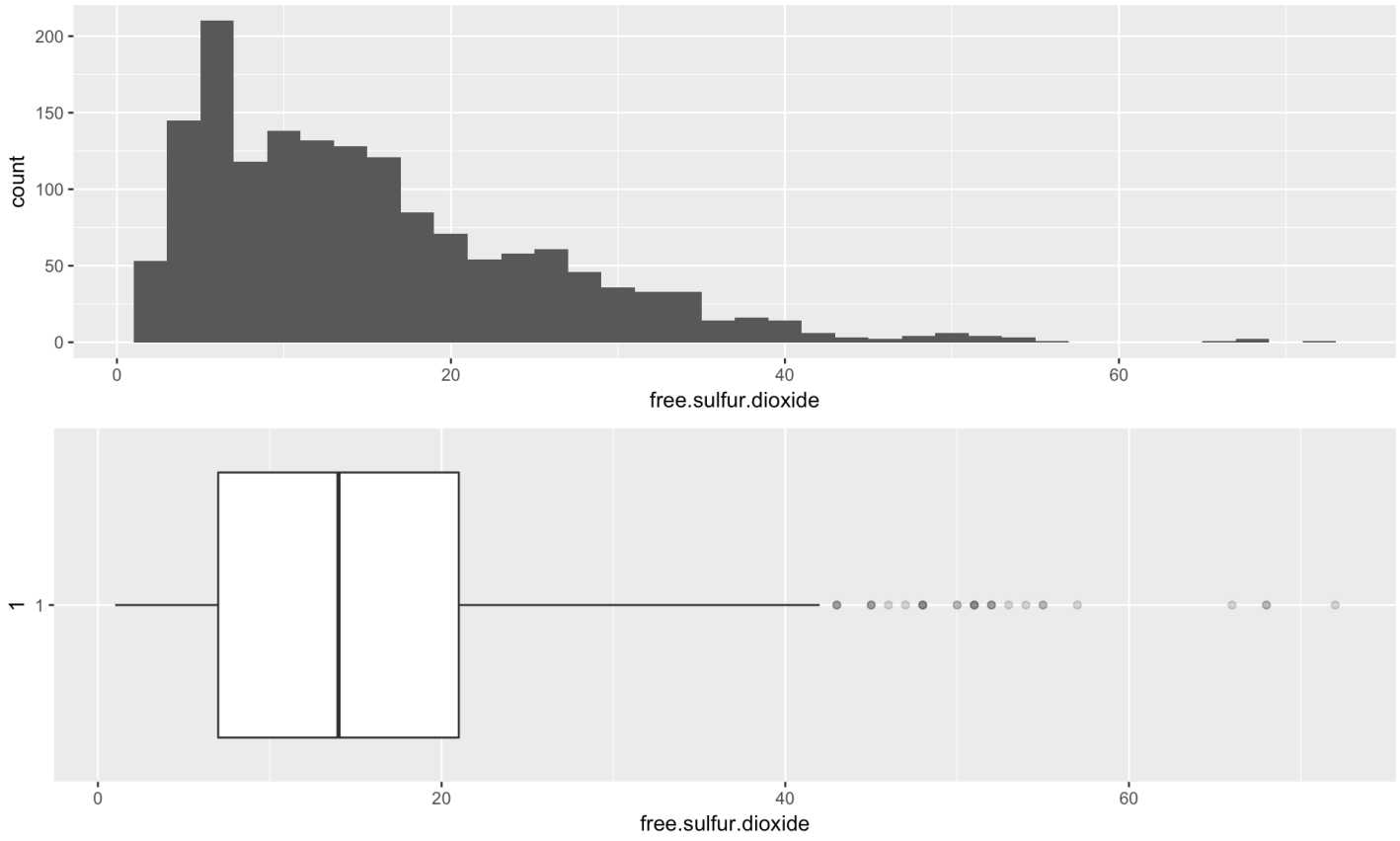

The distribution seems to be quite right skewed.



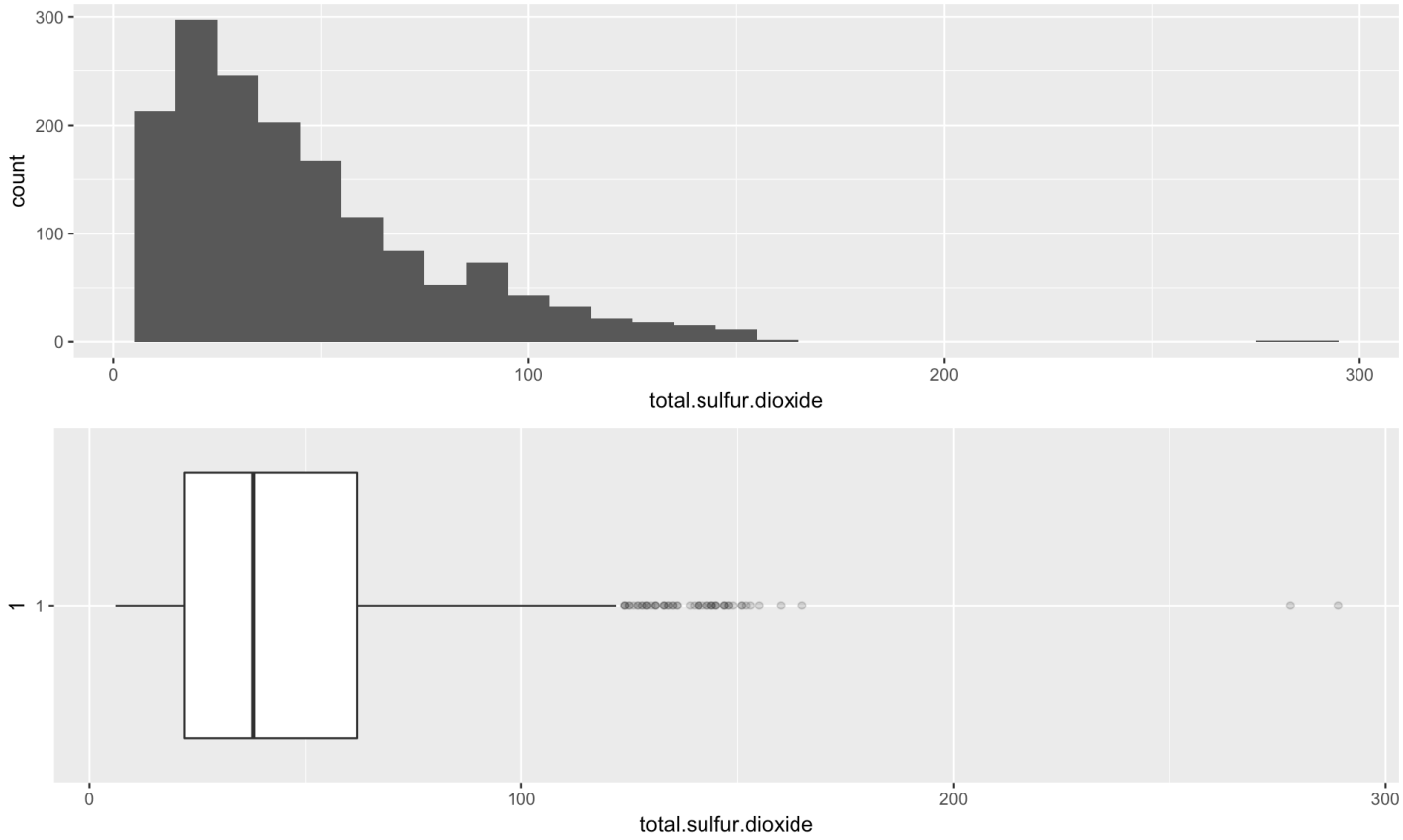

The distribution of residual.sugar is long tail with outliers.



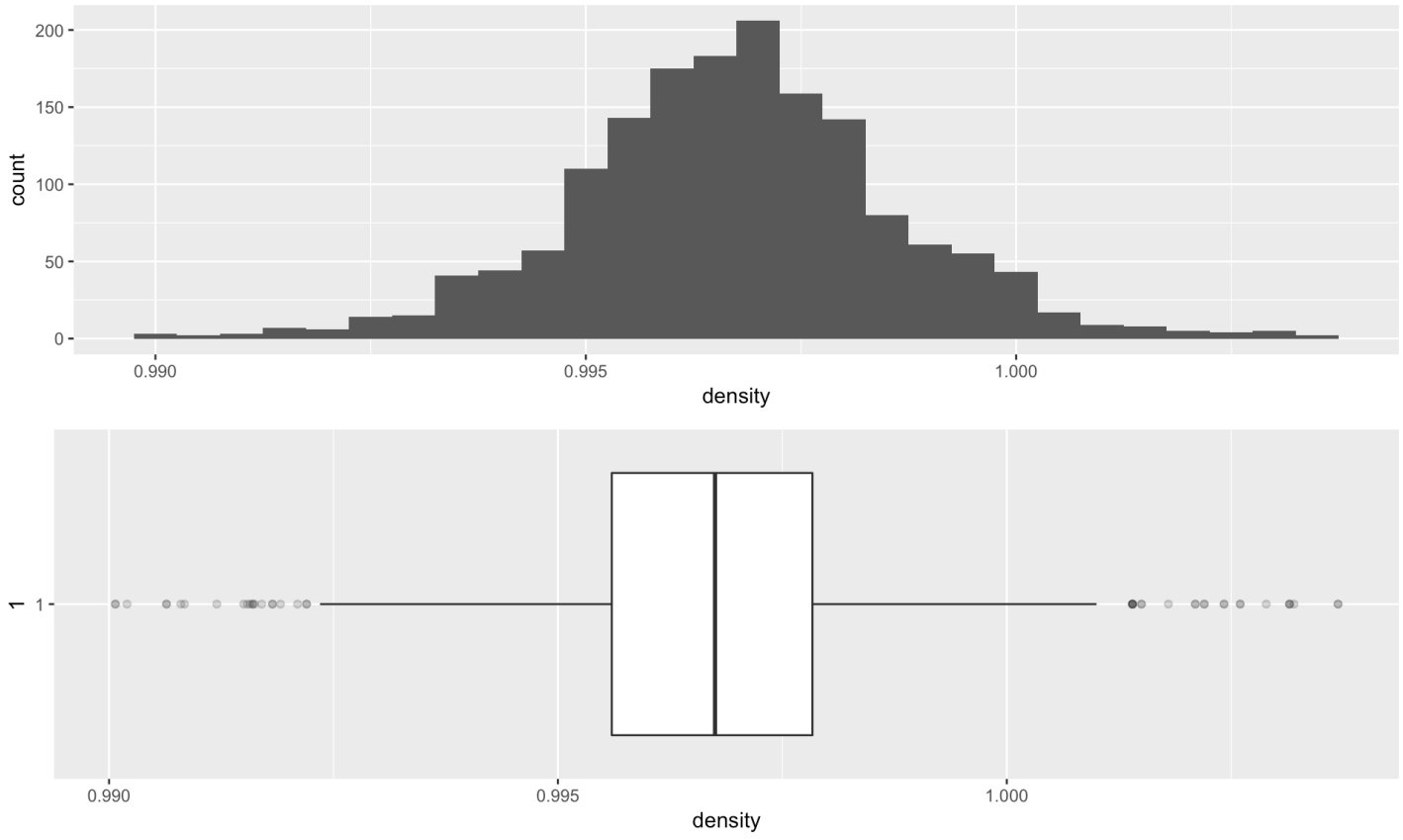

The distribution of chlorides is quite similar with residual.sugar and appears to be long tail with outliers.



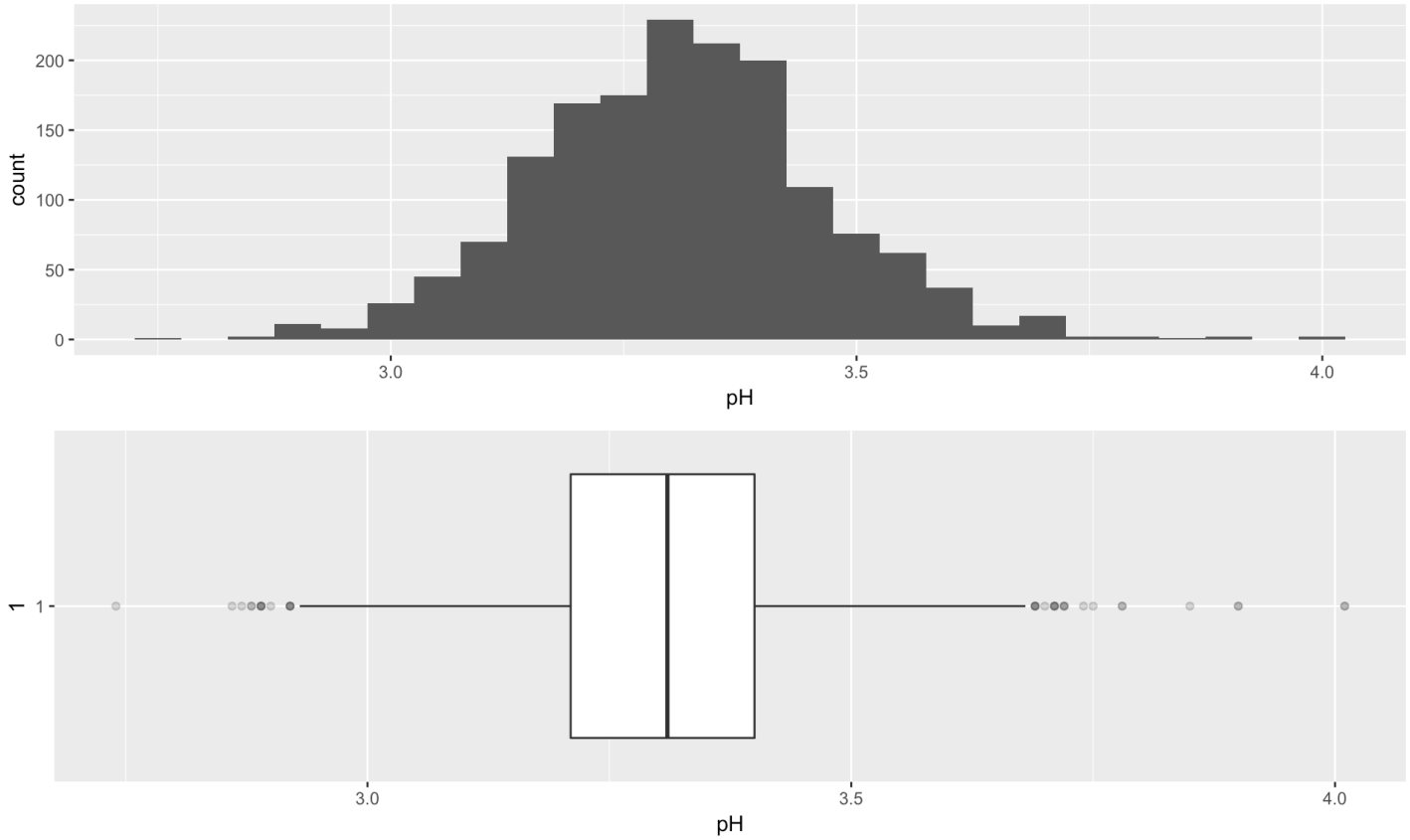

The distribution of free.sulfur.dioxide is right skewed with outliers.



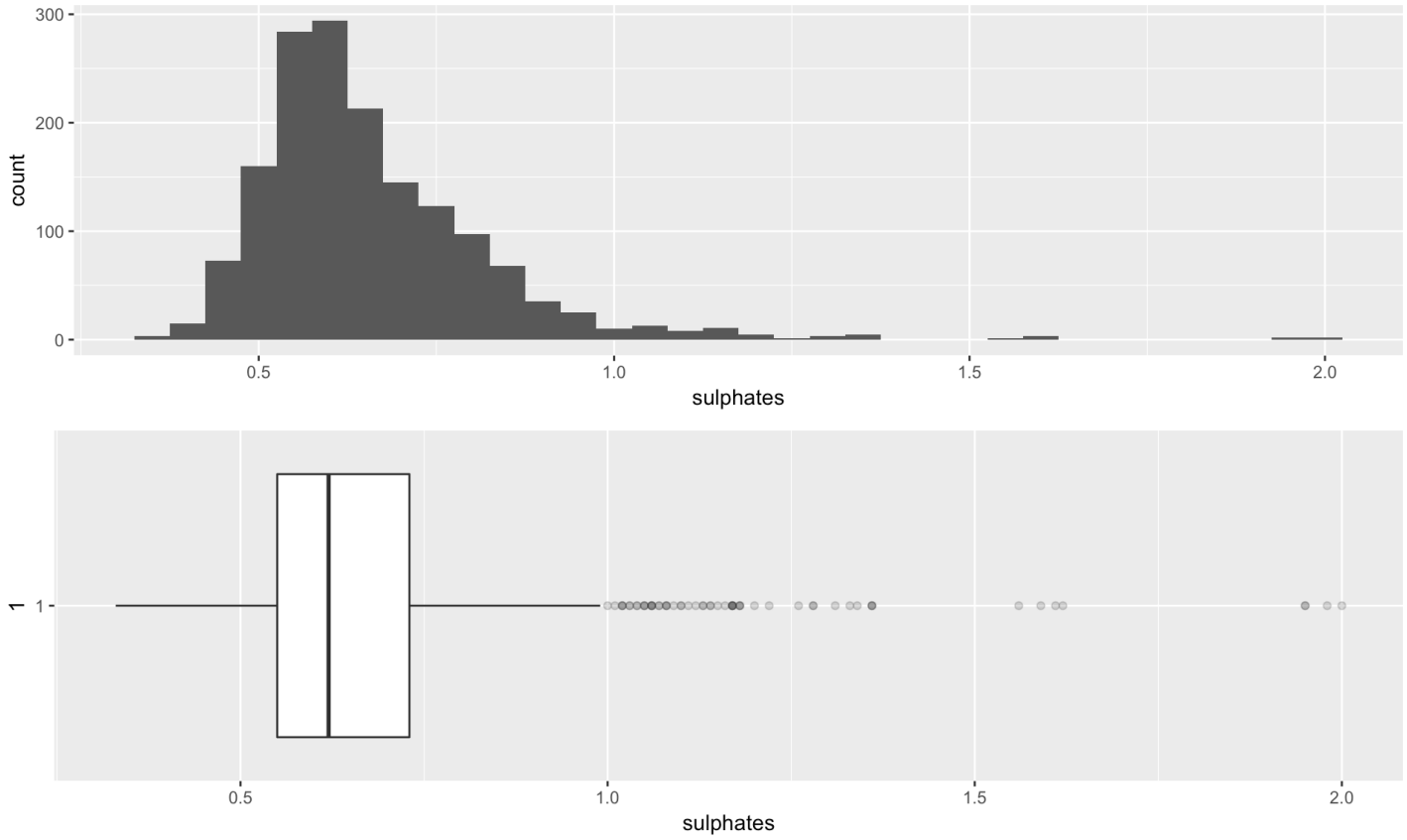

The distribution of total.sulfur.dioxide appears to be quite similar with free.sulfur.dioxide. It is right skewed with ourliers



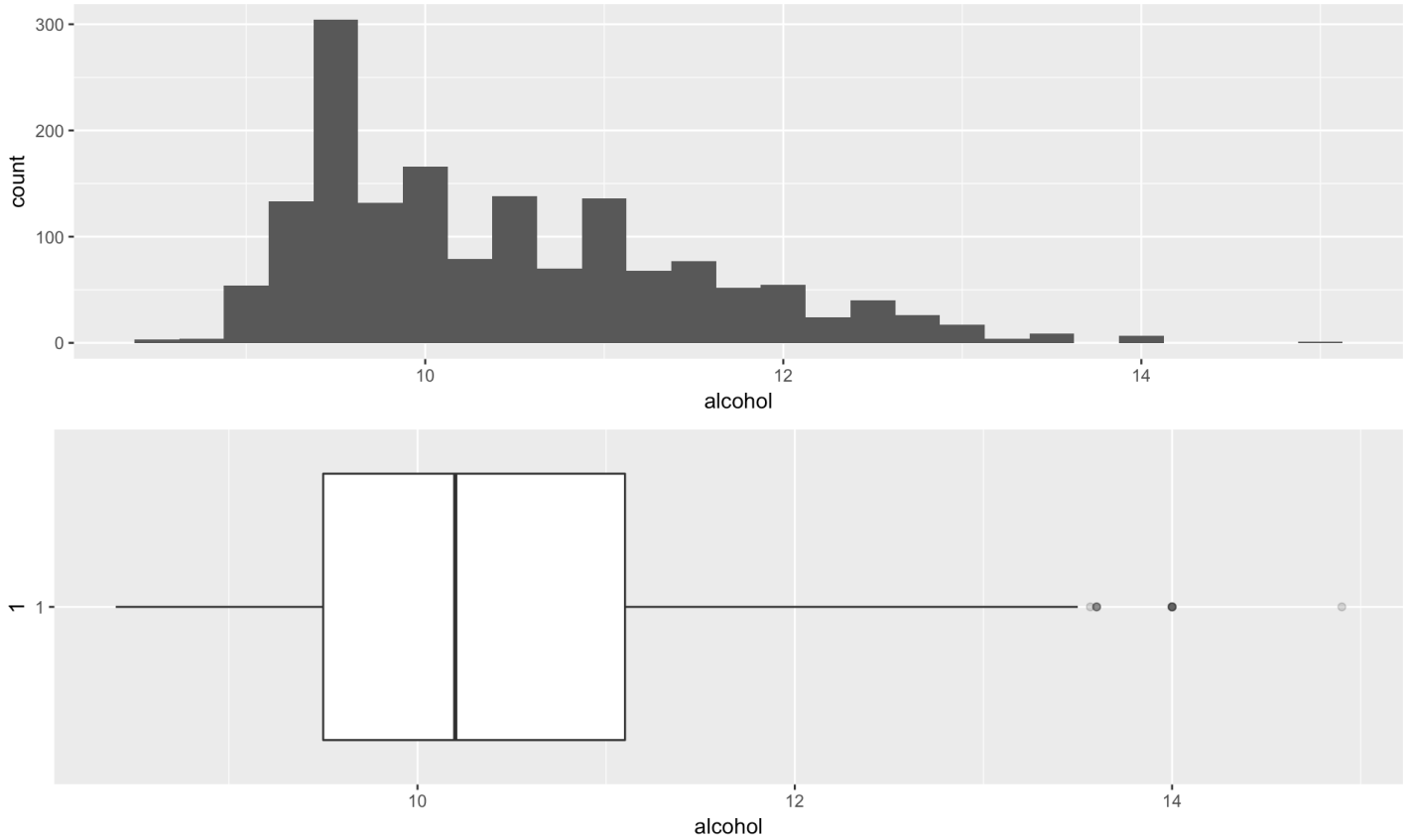

The distribution of density appears to be normal distribution with ourliers.



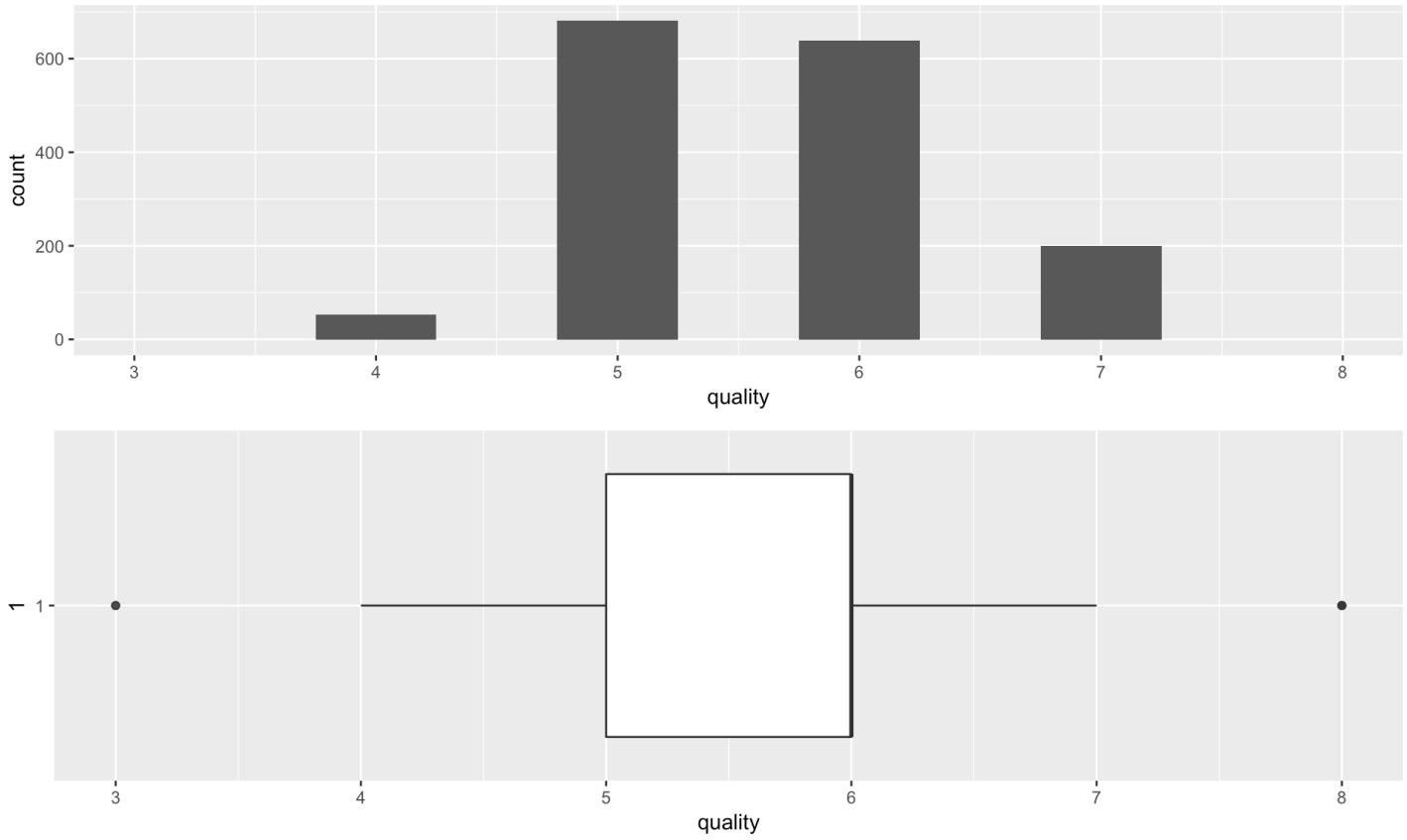

Similar with density, pH turns to be normally distributed with outliers.




The distribution of sulphates is right skewed with outliers.




The distribution of alcohol appears to be right skewed.



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

The distribution of quality ranges from 3-8 and appears to be normal distribution.

# Univariate Analysis

## What is the structure of your dataset?

- The dataset includes 1599 observations and 12 variables.
- Distribution of variables:

    1. Normal: volatile.acidity, density, PH, quality
    2. Right skewed: fixed.acidity, citric.acid, free.sulfur.dioxide, total.sulfur.dioxide, sulphates, alcohol,
    3. Long tail: residual.sugar, chlorides

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

The main feature of interest is $ quality.

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

> Quality is assessed best by tasting. (Source: BBR-wine-knowledge)

It would be interesting to look at variables which might influence the taste of wine:

- Acidity: $ fixed.acidity, $ volatile.acidity, $ citric.acid, $pH
- Sweetness: $ residual.sugar
- Freshness & smell: $ free.sulfur.dioxide, $ total.sulfur.dioxide, $ sulphates

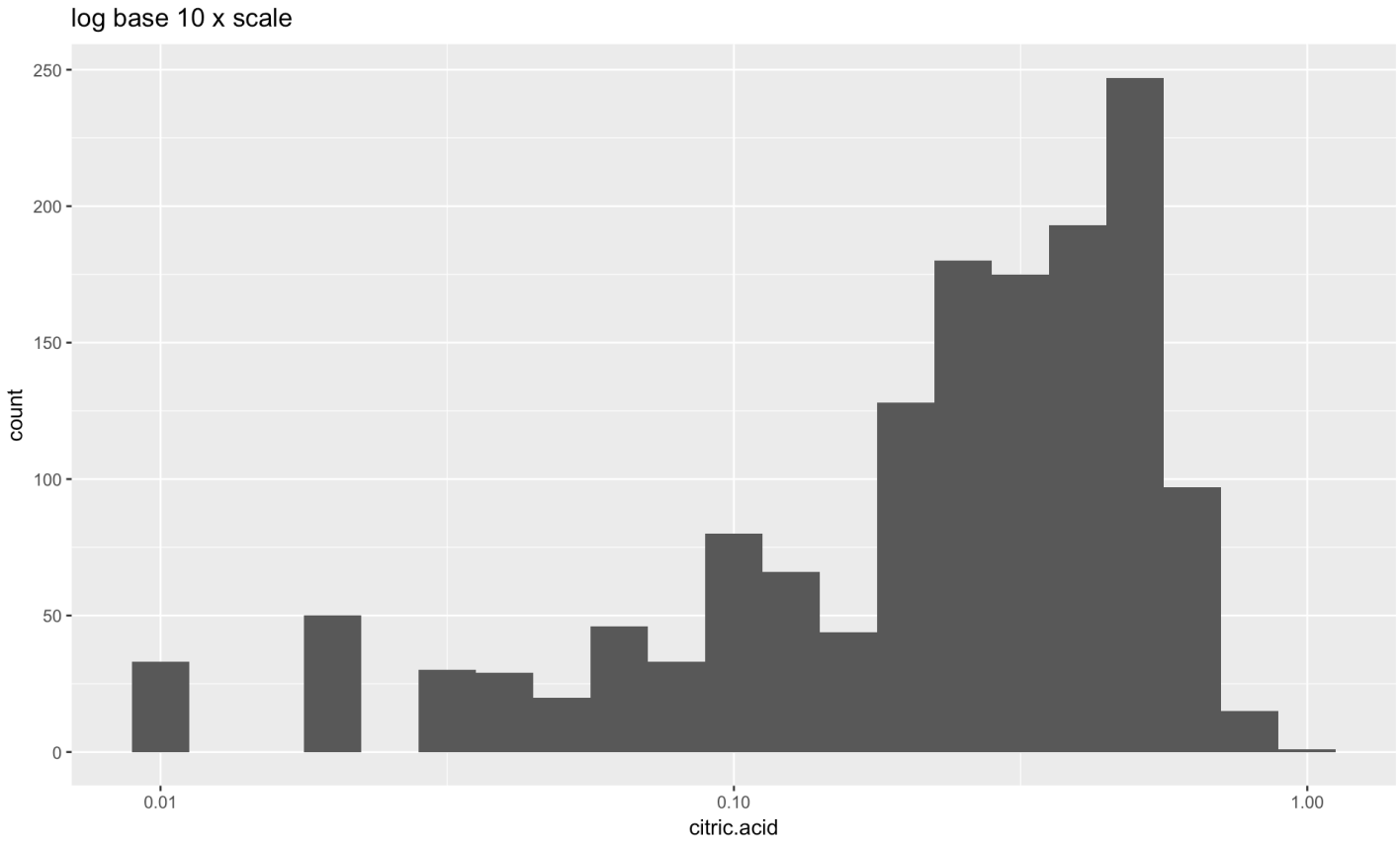
## Did you create any new variables from existing variables in the dataset?

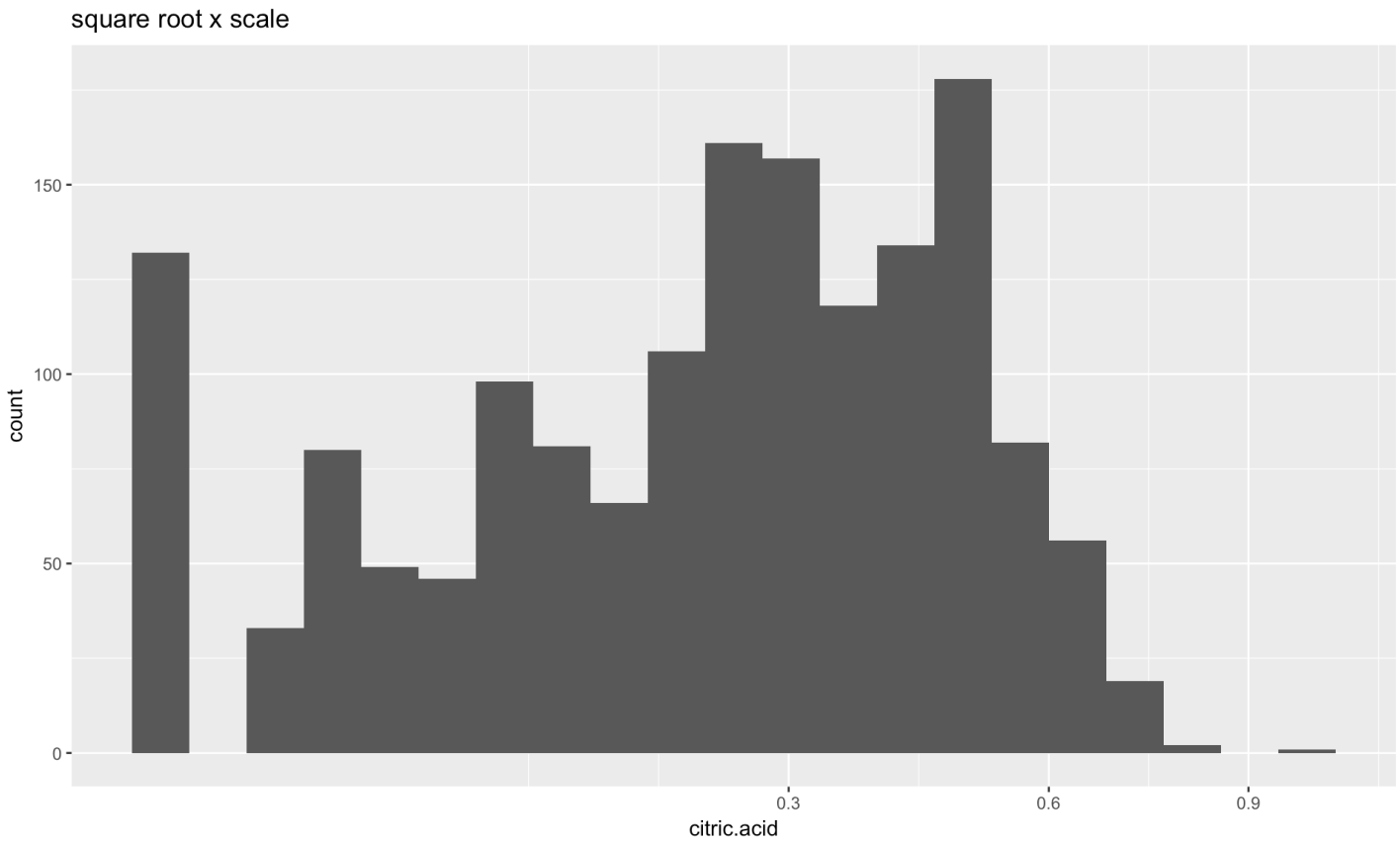
Yes. In order to increase the number in each quality category, a new varible quality.bucket is created. The value in the varible includes 3 levels: (2,4], (4,6], (6,8]

Moreover, in order to compute correlation in the below analysis, the variable quality will be changed to numeric.

# 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?

For the variables whoes distribution are right skewed or long tail, we can apply log base 10 or square root to scale x. Taking the citric.acid as an example.



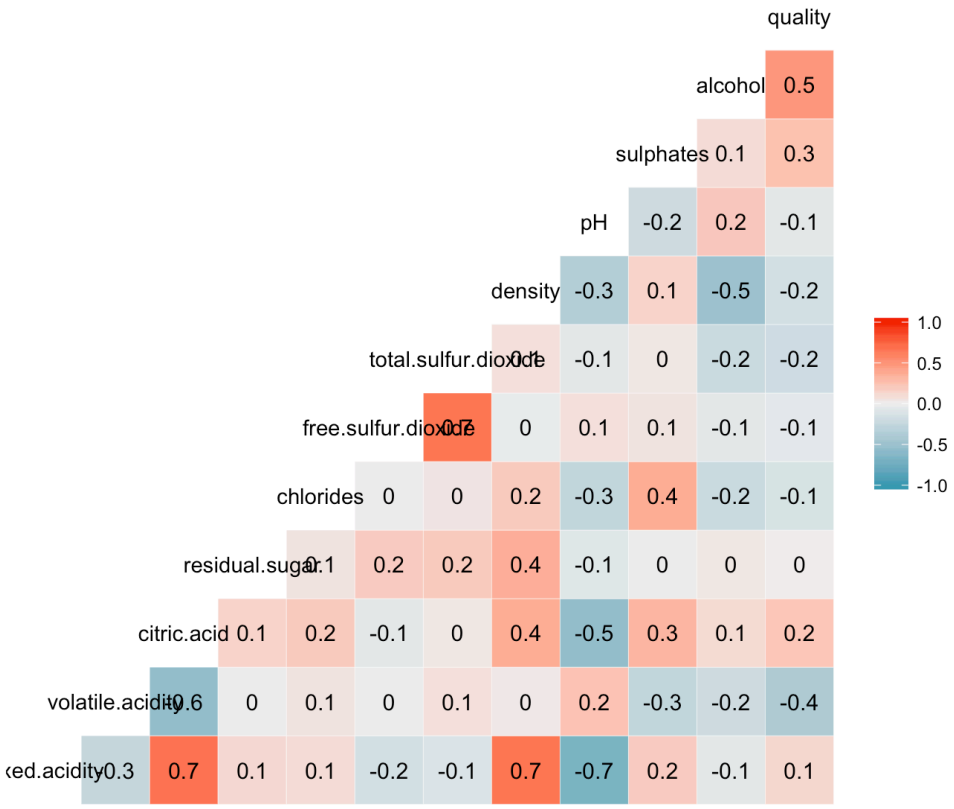



# Bivariate Plots Section

## Summary of correlation between variables

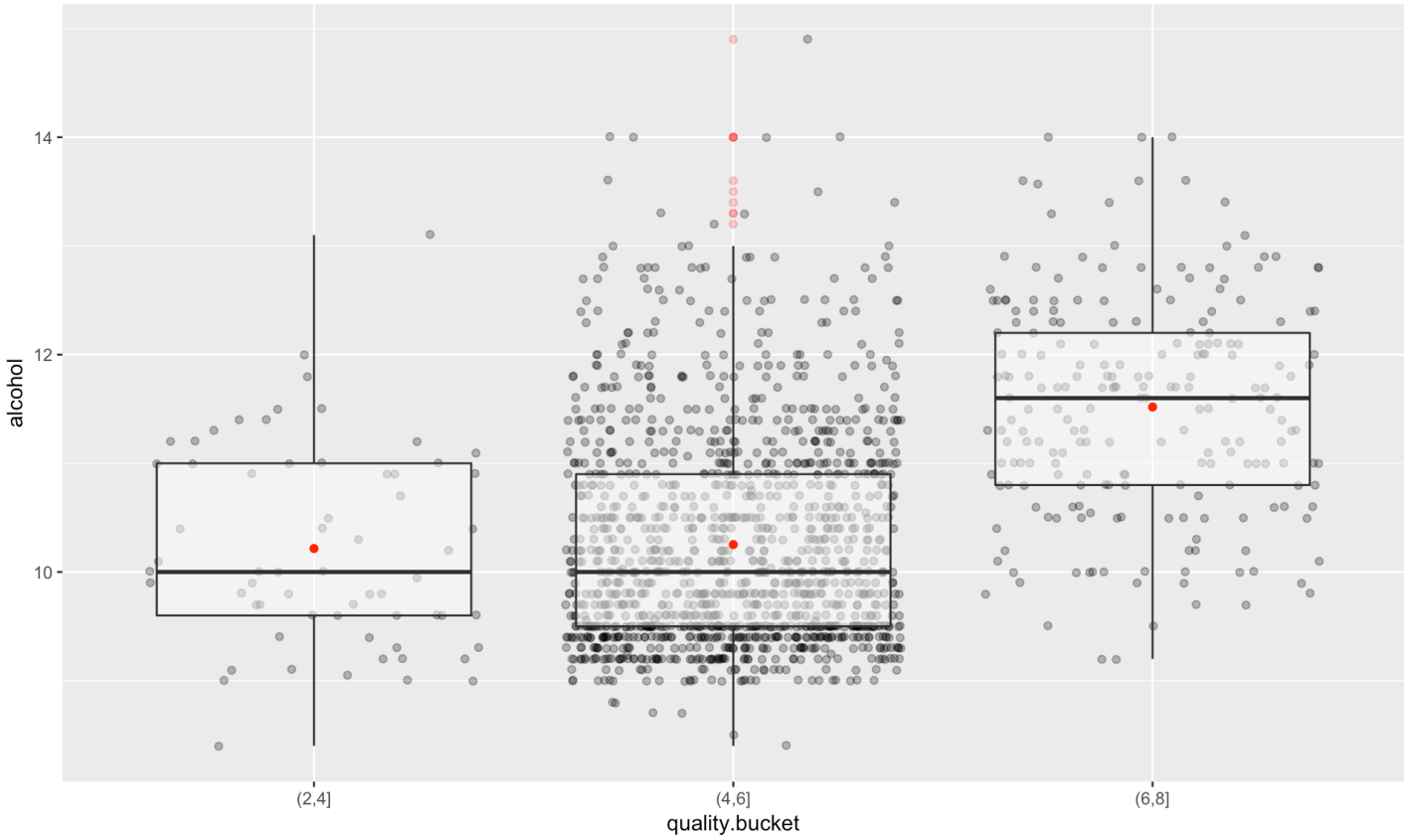
```
##                       fixed.acidity volatile.acidity citric.acid
## fixed.acidity            1.00000000     -0.256130895  0.67170343
## volatile.acidity        -0.25613089      1.000000000 -0.55249568
## citric.acid              0.67170343     -0.552495685  1.00000000
## residual.sugar           0.11477672      0.001917882  0.14357716
## chlorides                0.09370519      0.061297772  0.20382291
## free.sulfur.dioxide     -0.15379419     -0.010503827 -0.06097813
## total.sulfur.dioxide    -0.11318144      0.076470005  0.03553302
## density                  0.66804729      0.022026232  0.36494718
## pH                      -0.68297819      0.234937294 -0.54190414
## sulphates                0.18300566     -0.260986685  0.31277004
## alcohol                 -0.06166827     -0.202288027  0.10990325
## quality                  0.12405165     -0.390557780  0.22637251
##                       residual.sugar    chlorides free.sulfur.dioxide
## fixed.acidity            0.114776724  0.093705186        -0.153794193
## volatile.acidity         0.001917882  0.061297772        -0.010503827
## citric.acid              0.143577162  0.203822914        -0.060978129
## residual.sugar           1.000000000  0.055609535         0.187048995
## chlorides                0.055609535  1.000000000         0.005562147
## free.sulfur.dioxide      0.187048995  0.005562147         1.000000000
## total.sulfur.dioxide     0.203027882  0.047400468         0.667666450
```

```
## density                 0.355283371  0.200632327        -0.021945831
## pH                     -0.085652422 -0.265026131         0.070377499
## sulphates               0.005527121  0.371260481         0.051657572
## alcohol                 0.042075437 -0.221140545        -0.069408354
## quality                 0.013731637 -0.128906560        -0.050656057
##                      total.sulfur.dioxide     density          pH
## fixed.acidity                -0.11318144  0.66804729 -0.68297819
## volatile.acidity              0.07647000  0.02202623  0.23493729
## citric.acid                   0.03553302  0.36494718 -0.54190414
## residual.sugar                0.20302788  0.35528337 -0.08565242
## chlorides                     0.04740047  0.20063233 -0.26502613
## free.sulfur.dioxide           0.66766645 -0.02194583  0.07037750
## total.sulfur.dioxide          1.00000000  0.07126948 -0.06649456
## density                       0.07126948  1.00000000 -0.34169933
## pH                           -0.06649456 -0.34169933  1.00000000
## sulphates                     0.04294684  0.14850641 -0.19664760
## alcohol                      -0.20565394 -0.49617977  0.20563251
## quality                      -0.18510029 -0.17491923 -0.05773139
##                        sulphates     alcohol     quality
## fixed.acidity         0.183005664 -0.06166827  0.12405165
## volatile.acidity     -0.260986685 -0.20228803 -0.39055778
## citric.acid           0.312770044  0.10990325  0.22637251
## residual.sugar        0.005527121  0.04207544  0.01373164
## chlorides             0.371260481 -0.22114054 -0.12890656
## free.sulfur.dioxide   0.051657572 -0.06940835 -0.05065606
## total.sulfur.dioxide  0.042946836 -0.20565394 -0.18510029
## density               0.148506412 -0.49617977 -0.17491923
## pH                   -0.196647602  0.20563251 -0.05773139
## sulphates             1.000000000  0.09359475  0.25139708
## alcohol               0.093594750  1.00000000  0.47616632
## quality               0.251397079  0.47616632  1.00000000
```



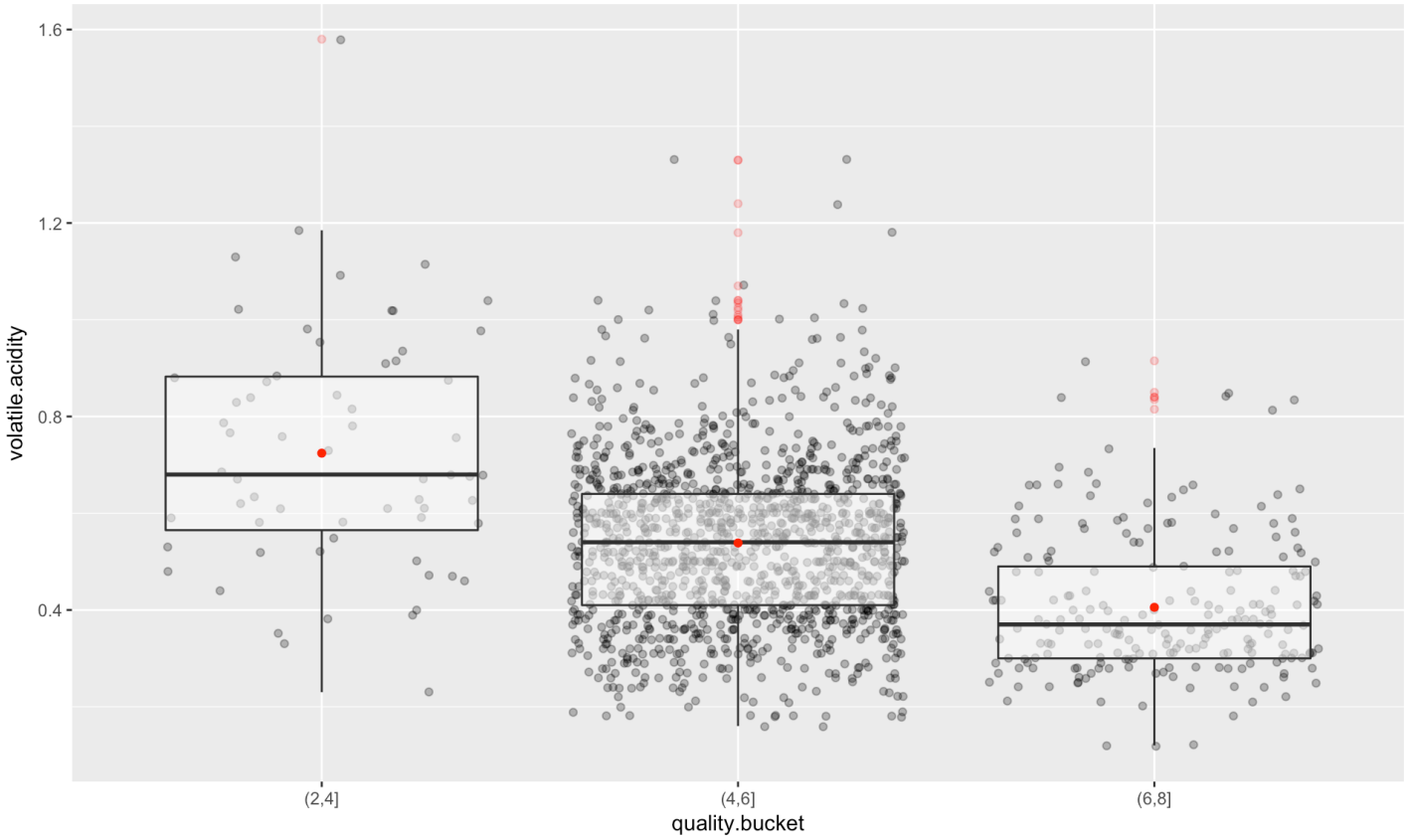

- For the variable quality, we can see a relatively strong correlation in the pairs below:
    1. Correlation between quality and alcohol: 0.476
    2. Correlation between quality and volatile acidity: -0.391
    3. Correlation between quality and sulphates: 0.251
    4. Correlation between quality and citric acid: 0.226
- For other variales:
    1. What is interesting is that the variable fixed.acidity seems to be correlative to a few other variables: citric.acid, density, pH.
    2. The variable citric.acid appears to be correlative to a few other variables: fixed.acidity, density, pH
    3. The variable free.sulfur.dioxide and total.sulfur.dioxide is correlative with each other.

# Plotting the variables correlative with quality.bucket



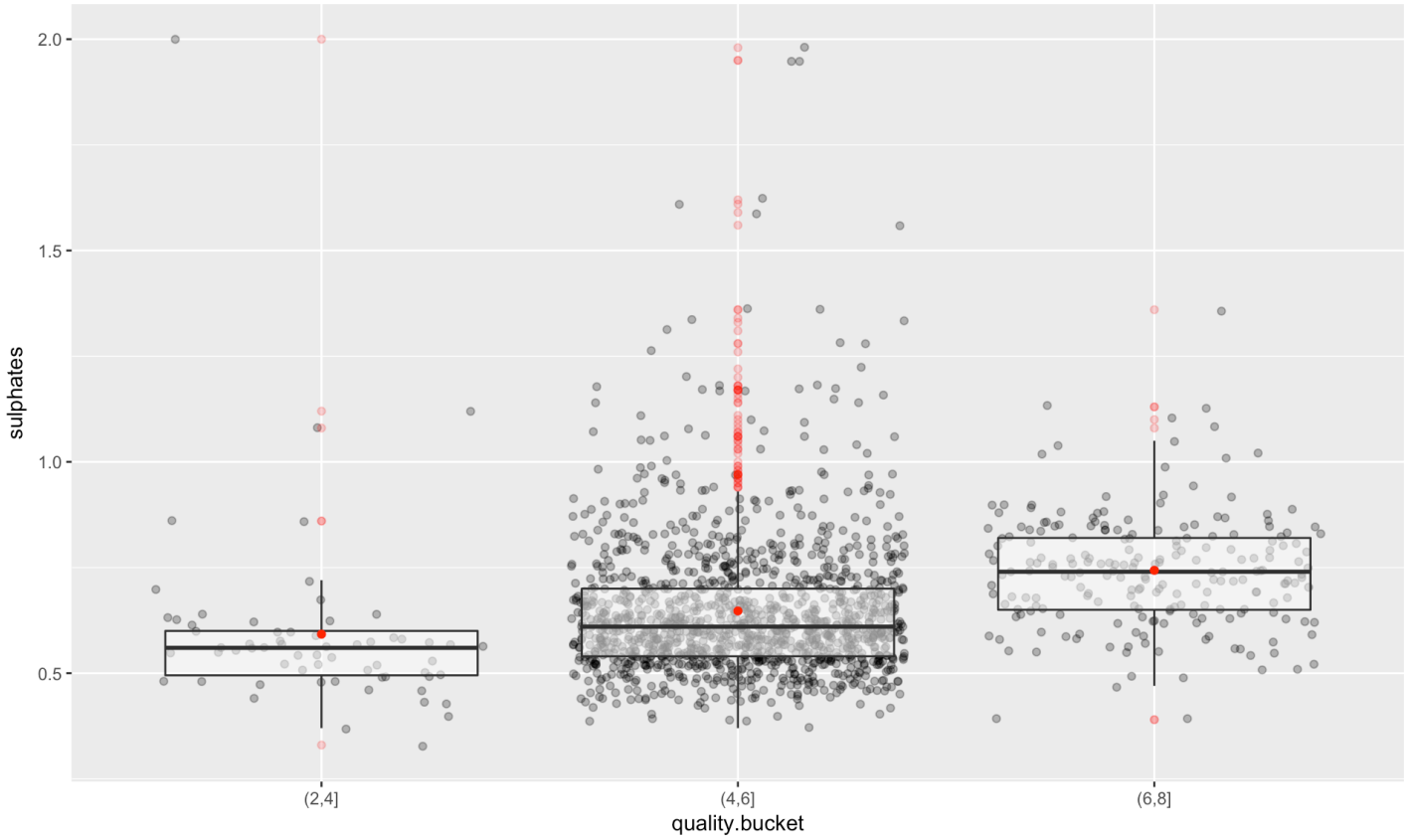

```
## $title
## [1] "Wine Quality group by Alcohol"
##
## $subtitle
## NULL
##
## attr(,"class")
## [1] "labels"
```

Better wine seems to have higher procent alcohol. But the difference is not obvious between wine in medium quality bucket and low quality bucket.



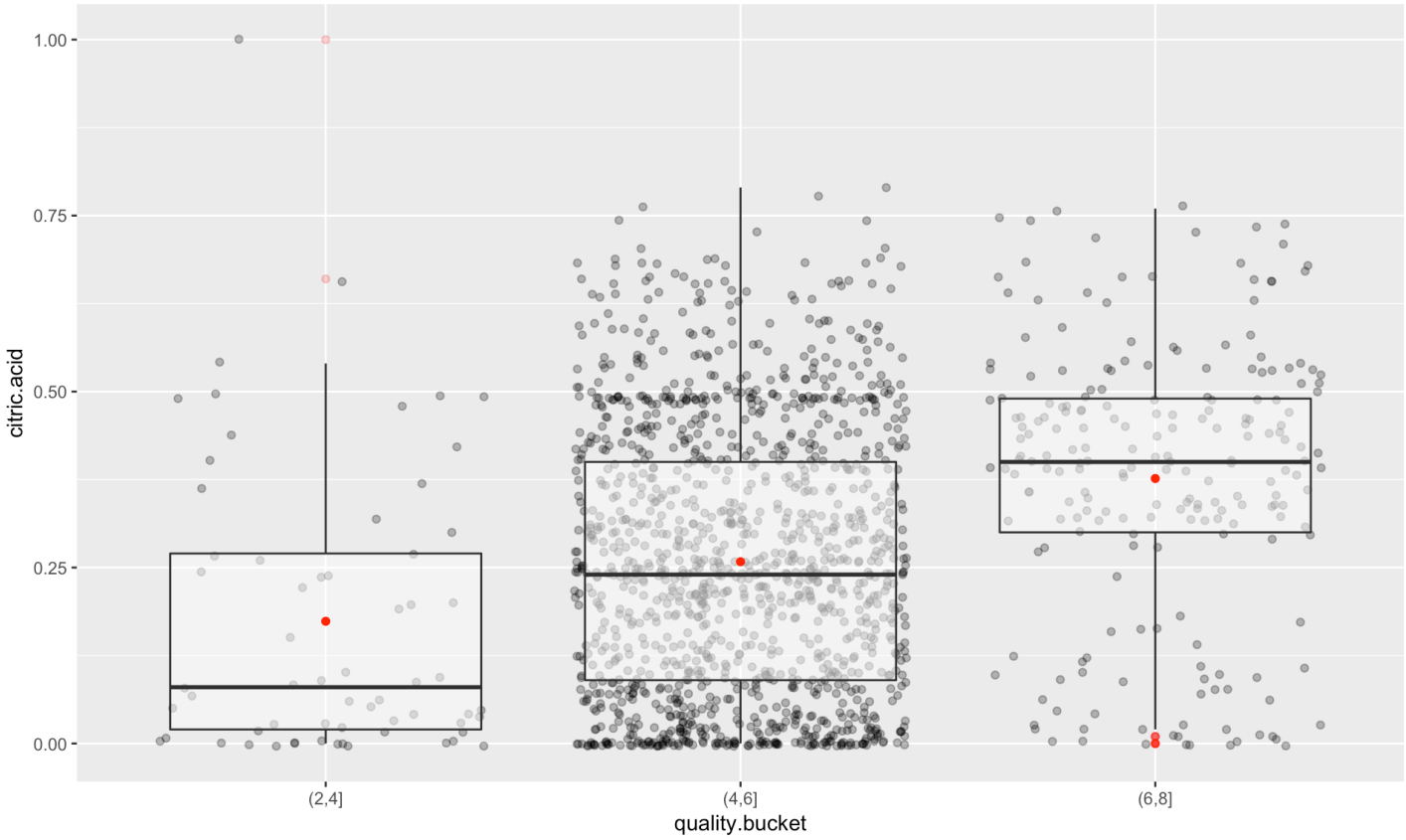

```
## $title
## [1] "Wine Quality group by Volatile Acidity"
##
## $subtitle
## NULL
##
## attr(,"class")
## [1] "labels"
```

Better wine contains to some certain sense less volatile acidity.




```
## $title
## [1] "Wine Quality group by Sulphates"
##
## $subtitle
## NULL
##
## attr(,"class")
## [1] "labels"
```

Sulphates, which is used to keep the wine fresh, is postively correlative with wine quality.

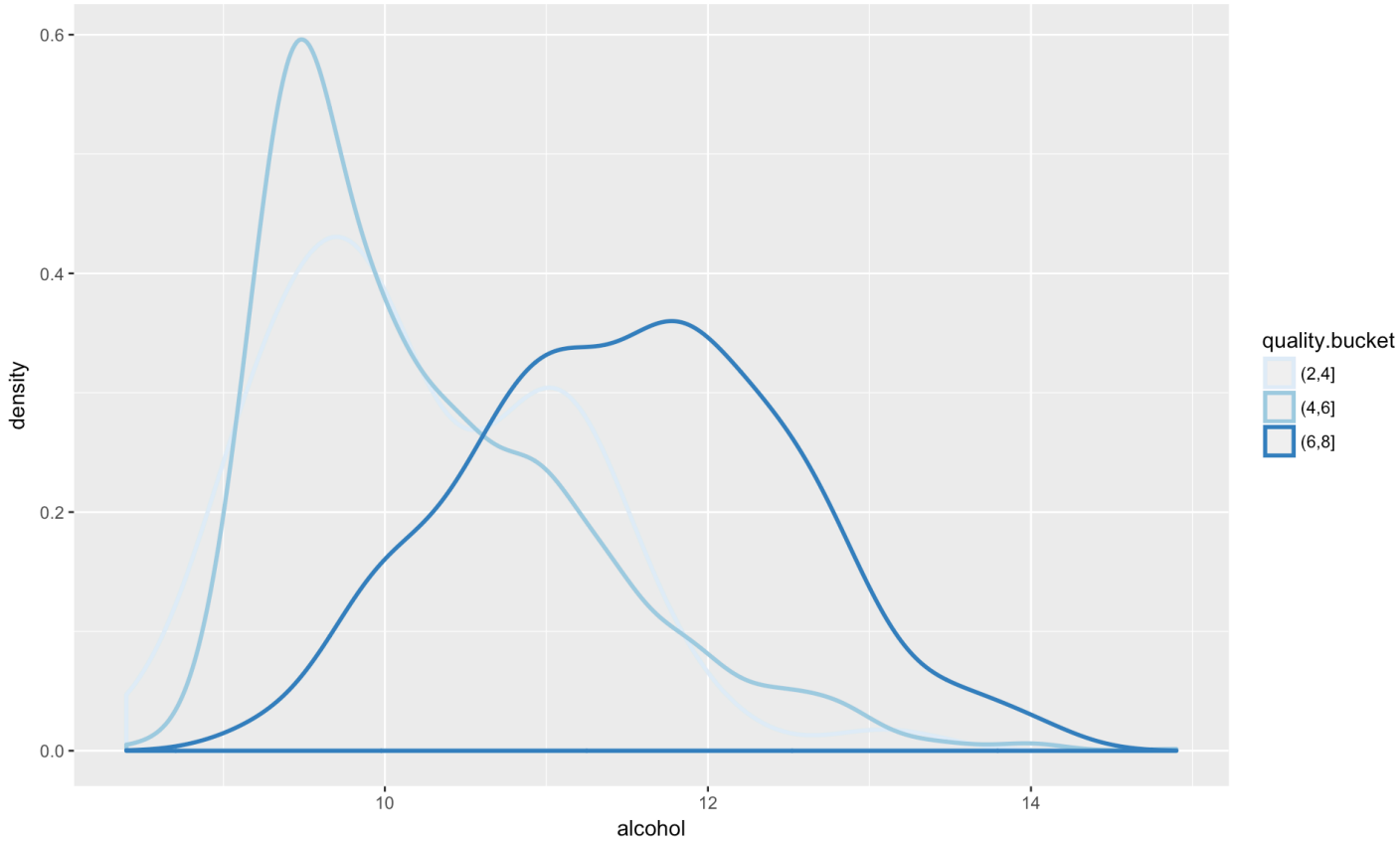



```
## $title
## [1] "Wine Quality group by Citric Acid"
##
## $subtitle
## NULL
##
## attr(,"class")
## [1] "labels"
```

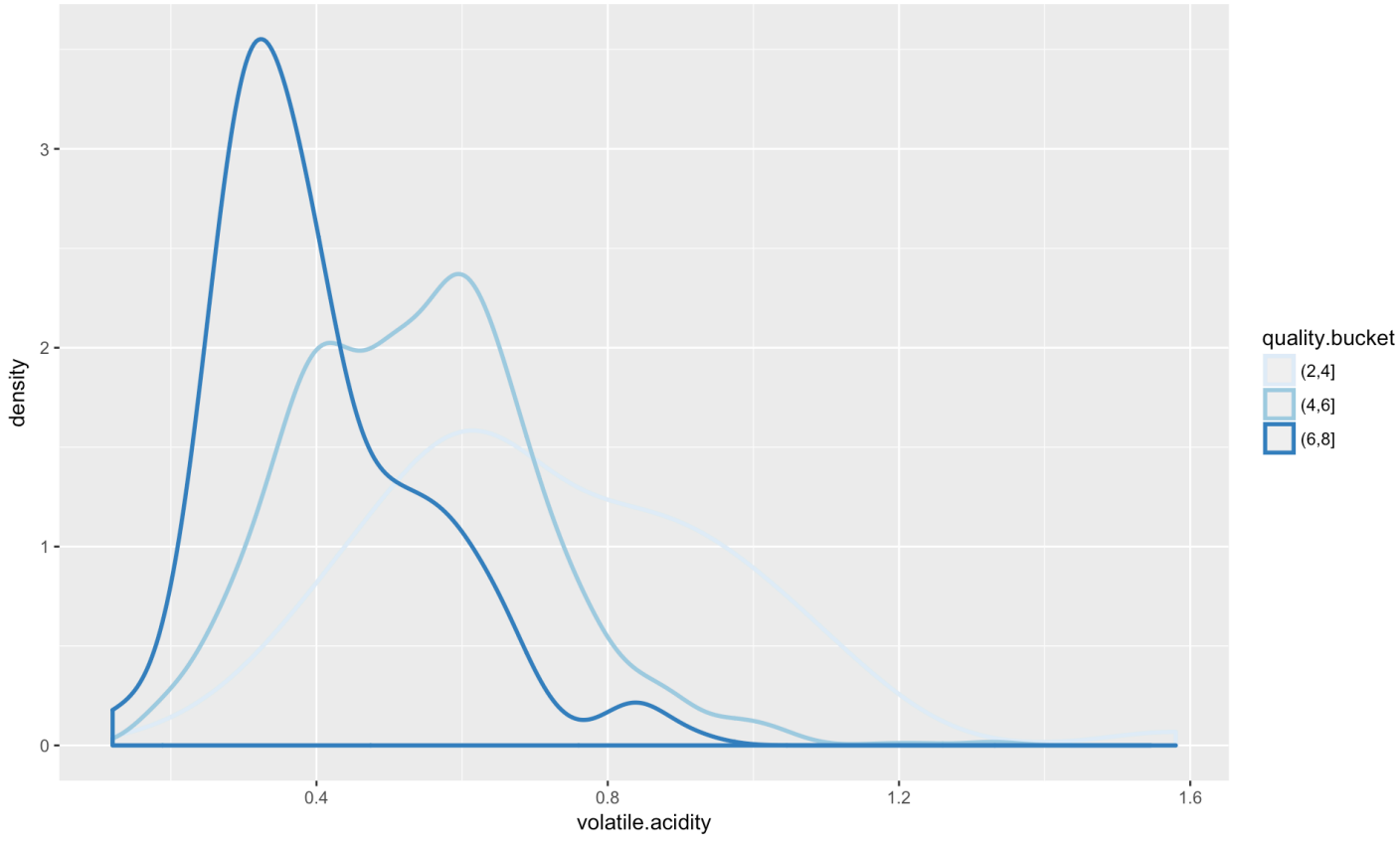
Better quality wine contains also to some certain sense more percentage of citric acid.

# Distribution of chemical properties in wine of different quality bucket

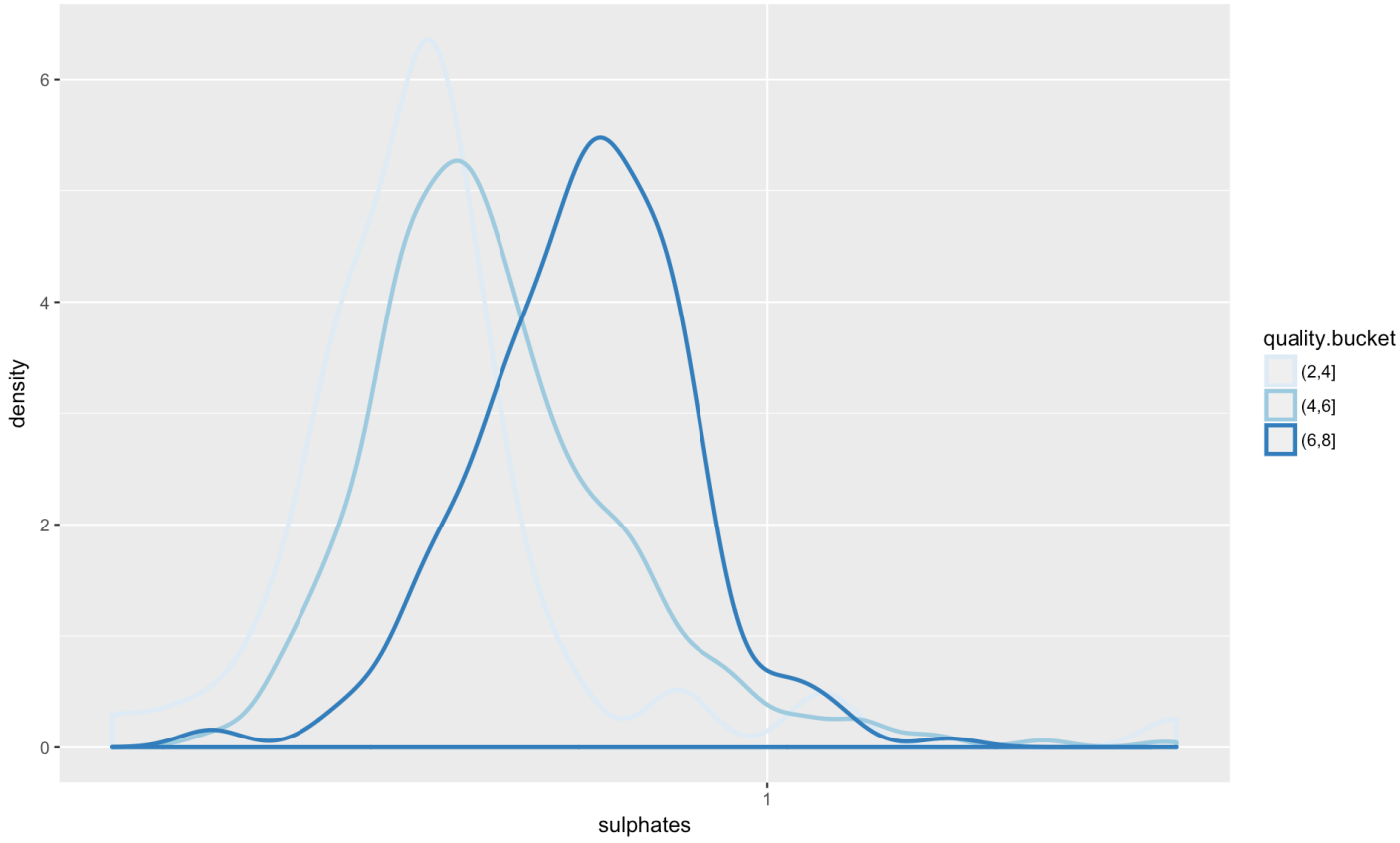
The plot will used the quality.bucket instead of quality to investigate the distribution in order to have more observations in each group to generalize the trend.



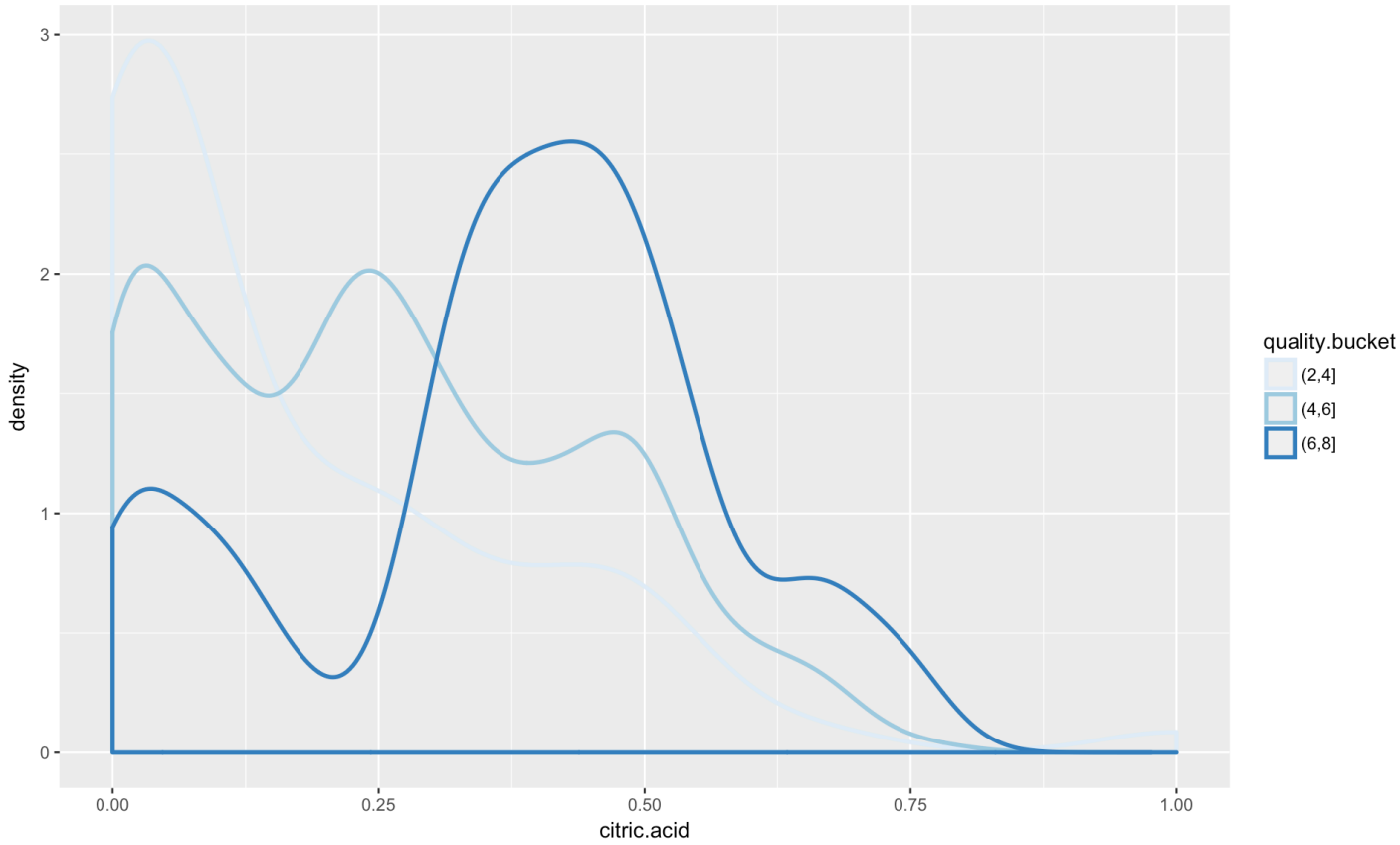

From the plot we see a clear difference of alcohol between different wine quality bucket: Better quality wine in bucket (6,8] have in average about 12% alcohol, whoes median is greater than the other two quality buckets.




The distrbution of volatile.acidity for the best quality bucket is relatively slim with a median at about 0.35.




The plot applies a log base 10 x scale and shows the median of diffrent quality.bucket better.




The distribution of citric.acid for different quality.bucket seems to be quite complicated. The distribution for the bucket (4,6] and (6,8] turns to be multimodal. However, the plot still shows a tendency that better quality wine contains to certain sense more citric acid.

# 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?

From the plotting above the following 4 variables are strong / relatively strong correlative with the red wine quality:

- Alcohol: The variable alcohol is stronly positively correlative with the quality: The best quality wine on average contains about 12% alcohol.
- Volatile acidity: The volatile acidity is also quite strongly correlative with the quality. Better quality wine contains less percentage of volatile acidity.
- Sulphates: With a log base 10 x scale we can see a difference of median of percentage of sulphates for different quality bucket. However, as the median is still quite close to each other and the distribution is quite slim and tall, we can only say that the sulphates is to some certain sense postively correlative to the wine quality.
- Citric acid: From the distribution we see some overlapping between different quality bucket. And the distribution is also quite complicated with multimodal. However, we can still see that better quality wine contains higher percentage of citric acid.

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

Some relationships for other variales: 1. What is interesting is that the variable fixed.acidity seems to be correlative to a few other variables: citric.acid, density, pH. 2. The variable citric.acid appears to be correlative to a few other variables: fixed.acidity, density, pH 3. The variable free.sulfur.dioxide and total.sulfur.dioxide is correlative with each other.

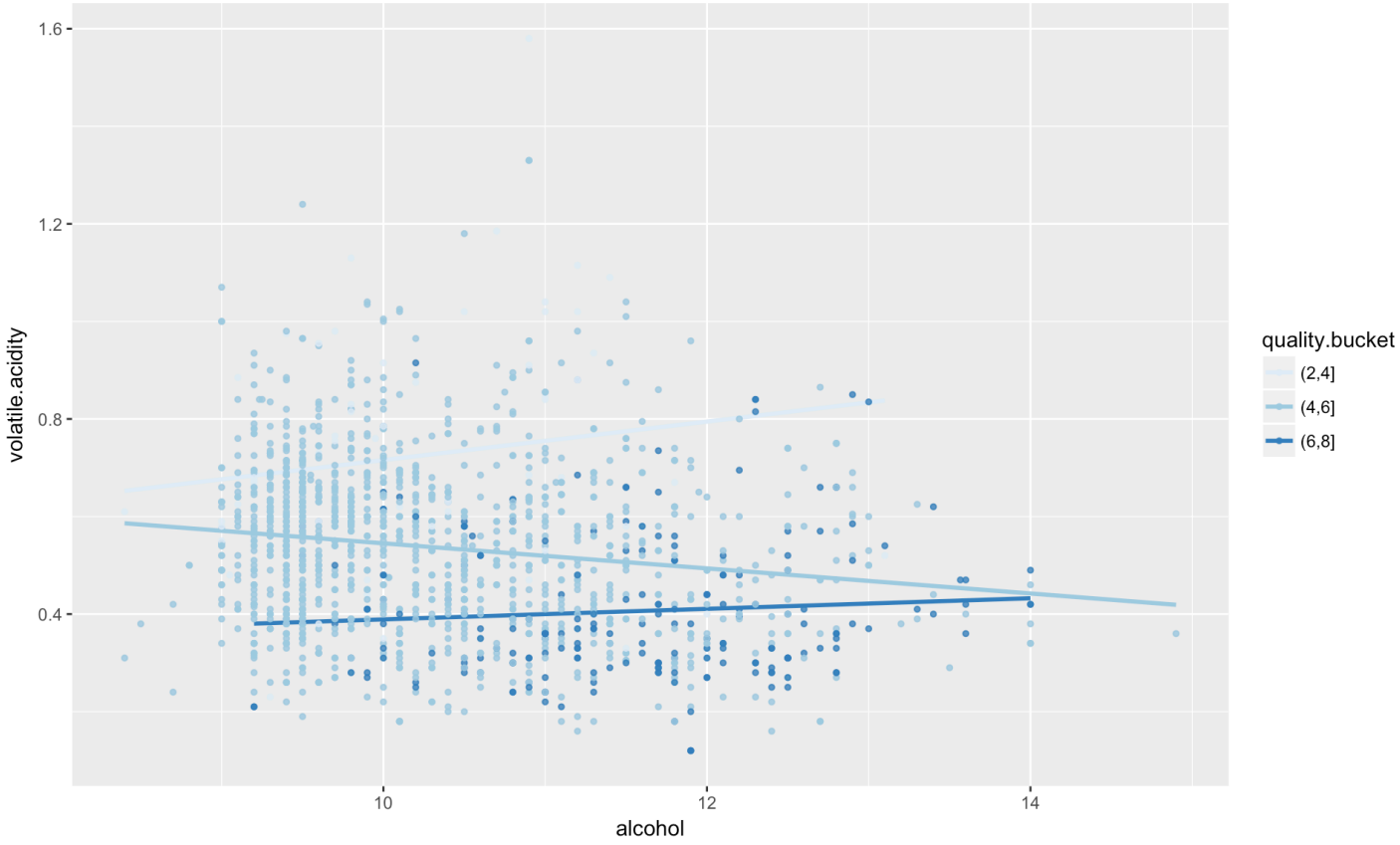
## What was the strongest relationship you found?

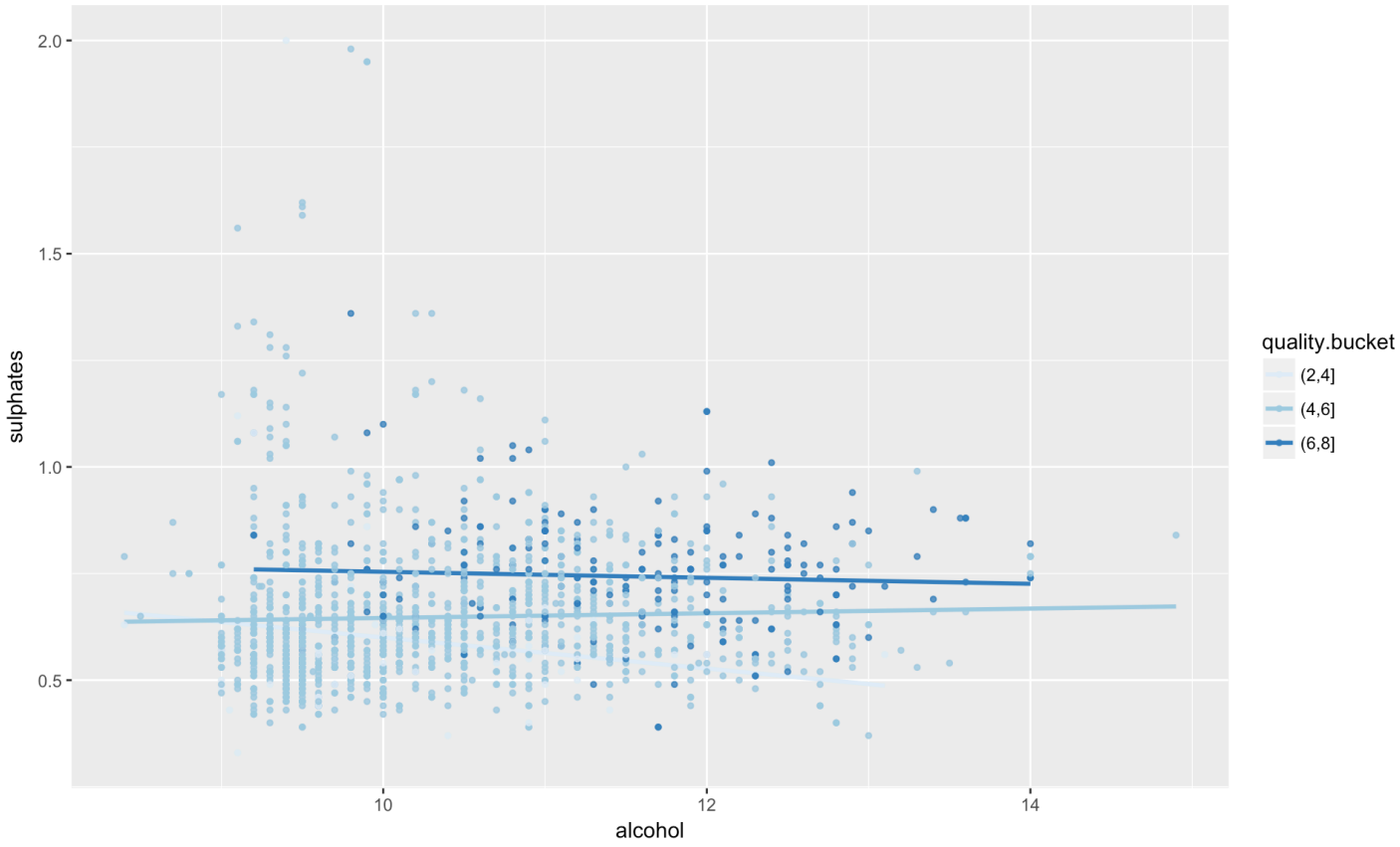
The fixed.acidity & pH shows a strong negative correlation: r = -0.68

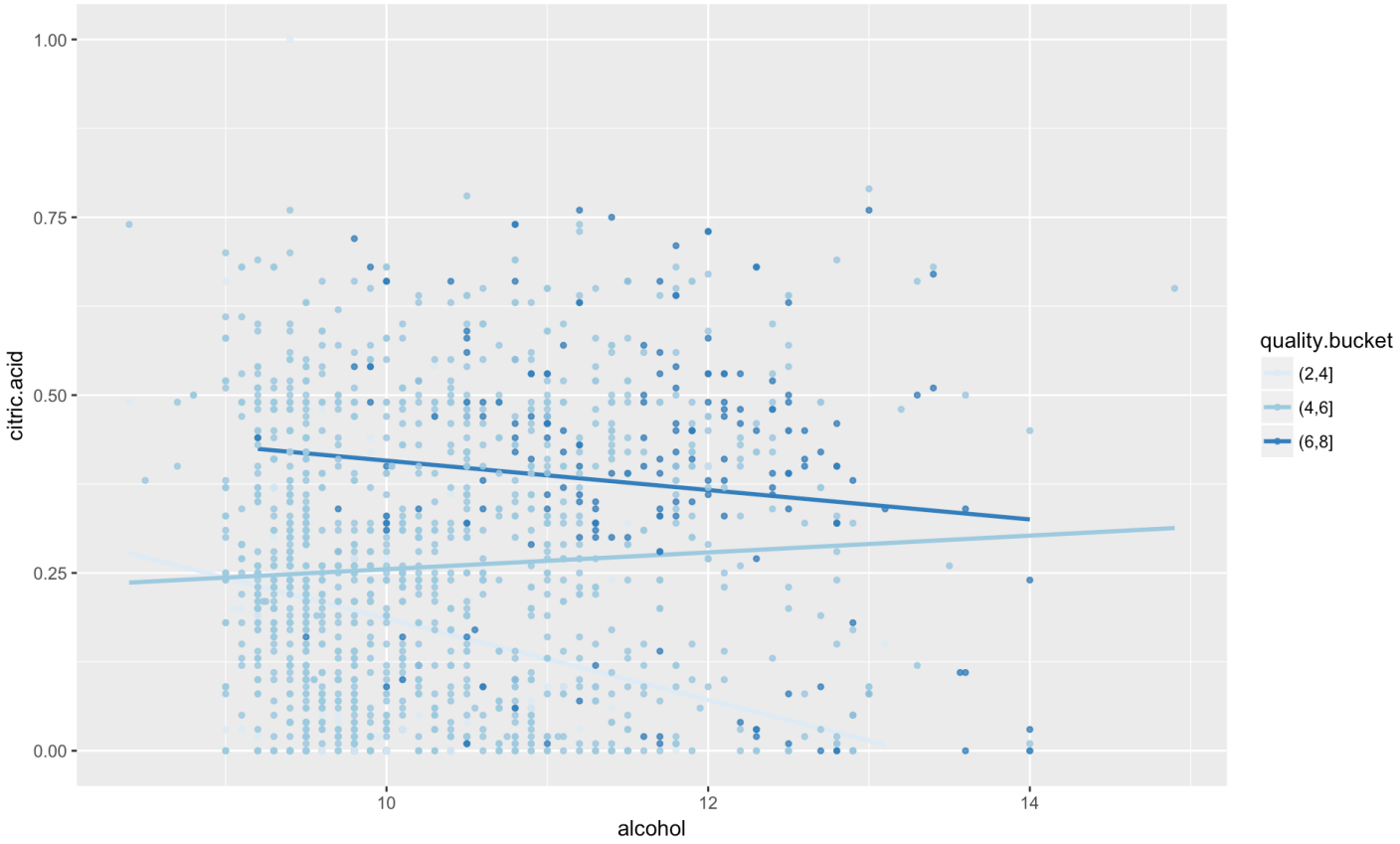
# Multivariate Plots Section

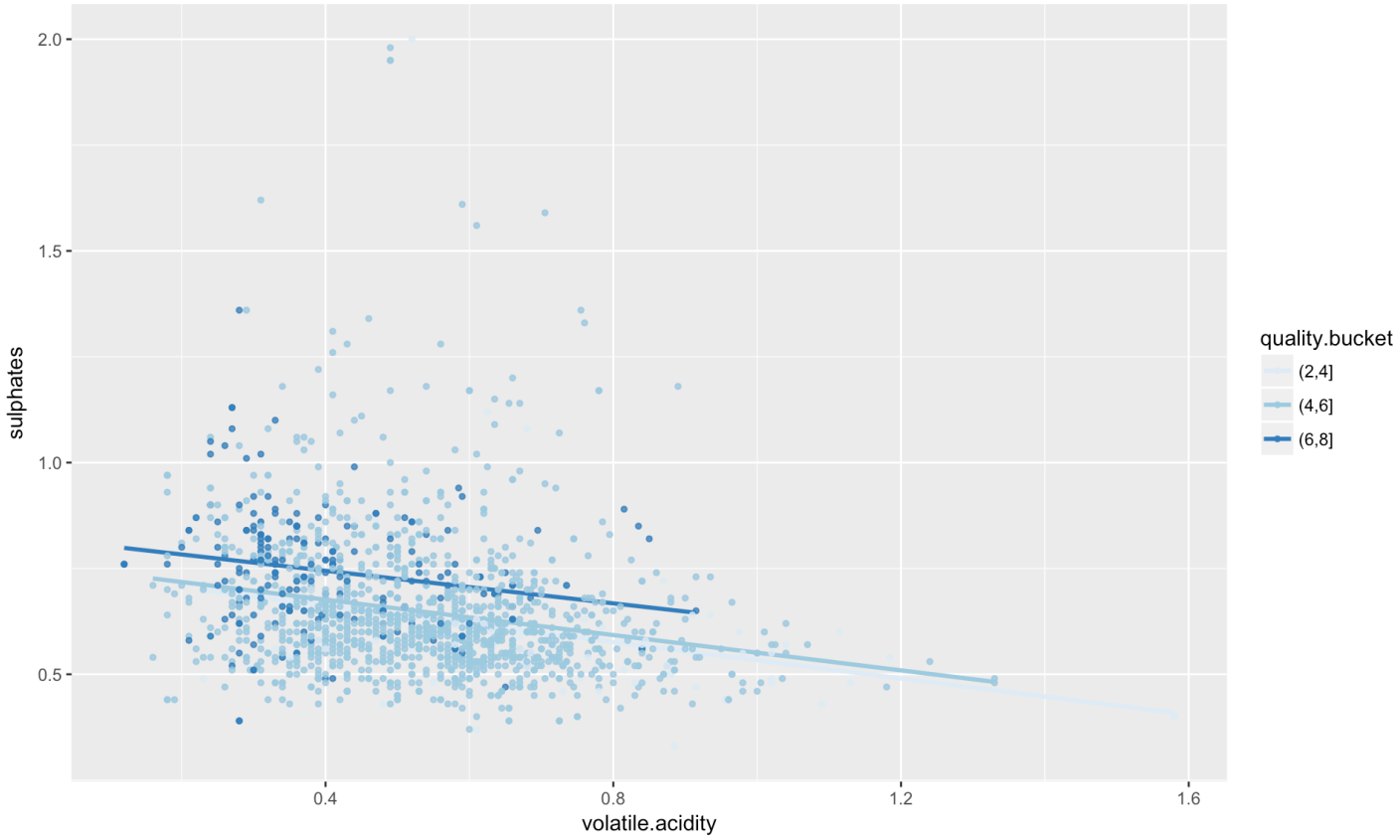
Since the variable alcohol and volatile acidity show a strong correlation with quality of wine, and the variable sulphates and citric acid show to some certain sense correlation with quality of wine, it would be interesting to combine to variables in the following ways:

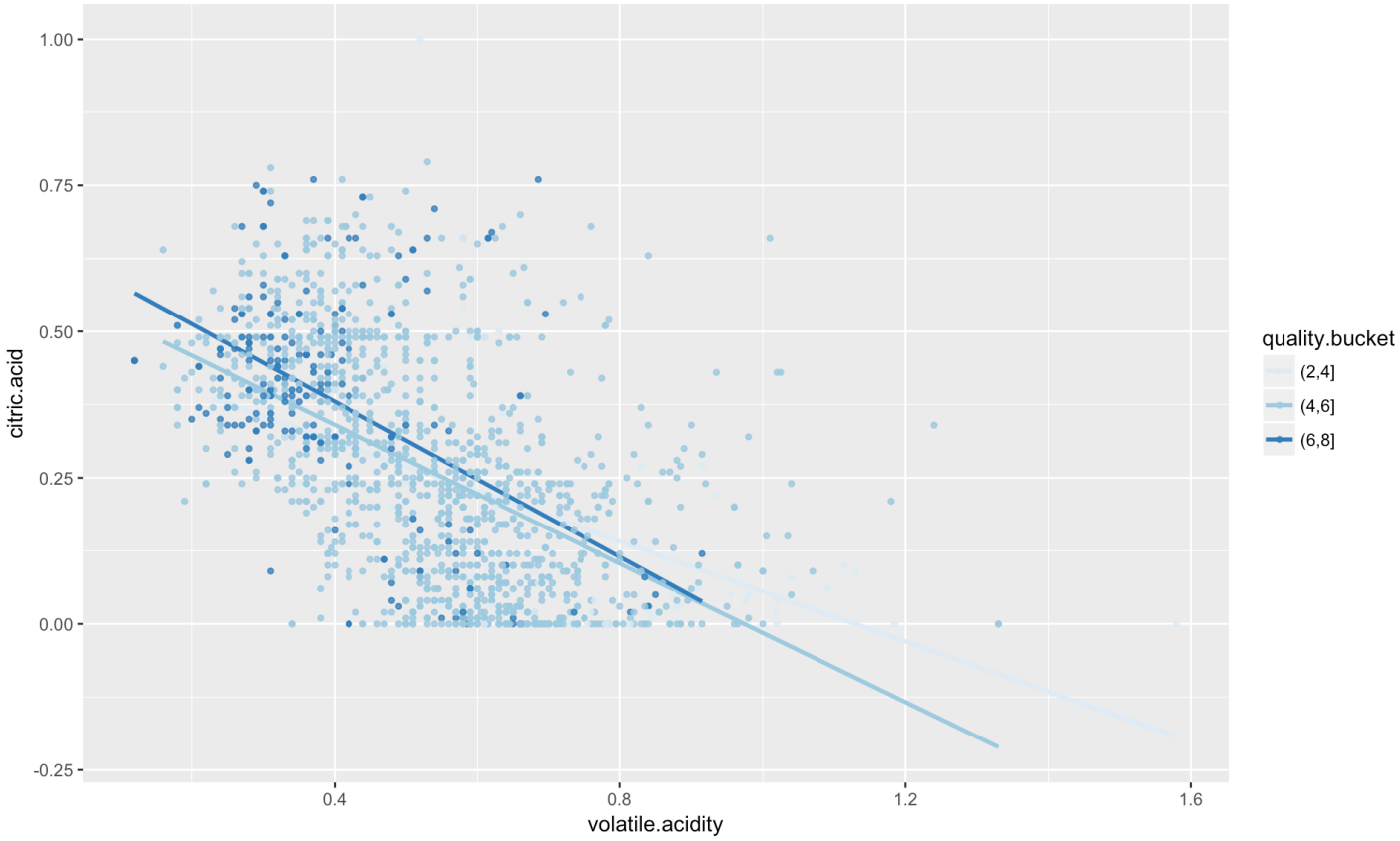
- Strong variable with strong variable
- Strong variable with relatively strong variable




1.00
0.75
0.50
0.25
0.00
citric.acid
10
12
14
alcohol
quality.bucket
(2,4]
(4,6]
(6,8]

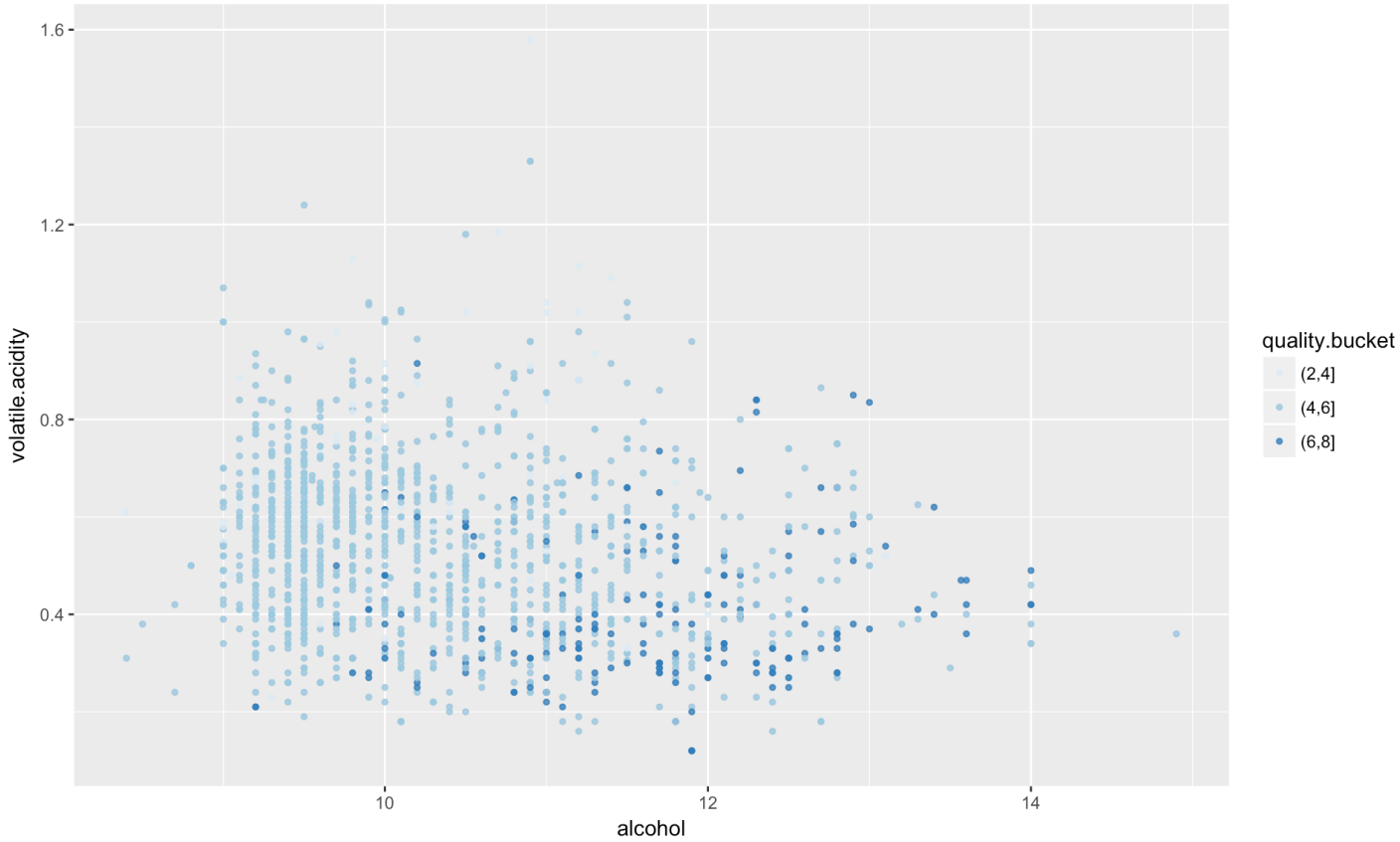

2.0
1.5
1.0
0.5
sulphates
0.4
0.8
1.2
1.6
volatile.acidity
quality.bucket
(2,4]
(4,6]
(6,8]




From the plots above, the following combination seems to have less overlapping between quality bucket:

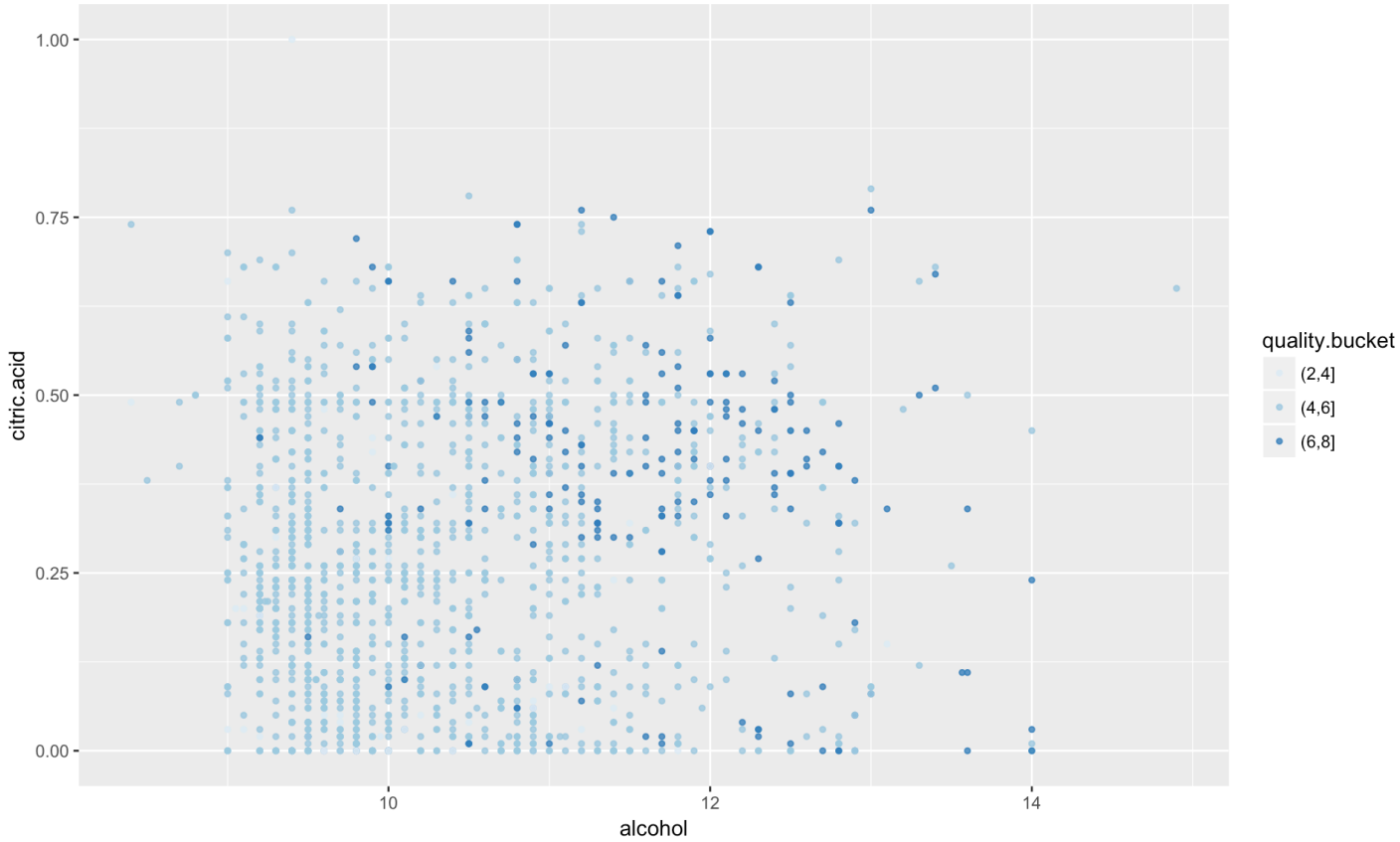
- Alcohol, volatile acidity and quality bucket
- Alcohol, citric acid and quality bucket

Let’s visualize the above combination with scatterplot.




Even though the plots between bucket (2,4] and (4,6] are sometimes mixed together, we can still see a relatively clear area where the best quality bucket (6,8] locates: This shows the following character on general with estimation:

- alcohol between 11% and 13%
- volatile.acidity between 0.3g / dm^3 and 0.6g / dm^3




The plots of bucket (2,4] and (4,6] are still missing together. However, we can still see that the plots of the best quality bucket (6,8] are relatively centralized in the area where:

- alcohol with estimation between 11% and 13%
- citric acid with estimation between 0.3g / dm^3 and 0.6g / dm^3

# 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?

Combining features together would strengthened each other in terms of detecting the best quality wine. The alcohol and the volatile acidity together can indicate the wine quality. A good quality wine is more likely to show the following character: * alcohol between 11% and 13% * volatile.acidity between 0.3g / dm^3 and 0.6g / dm^3

Moreover, the alcohol and the citric acid together can also indicate the wine quality. A good quality wine is likely to show the following character:

- alcohol with estimation between 11% and 13%
- citric acid with estimation between 0.3g / dm^3 and 0.6g / dm^3

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

Based on the above finding, it would be interesting to check the correlation of the following two pairs:

- Alcohol and volatile.acidity: r = -0.20
- Alcohol and citric.acid: r = 0.11

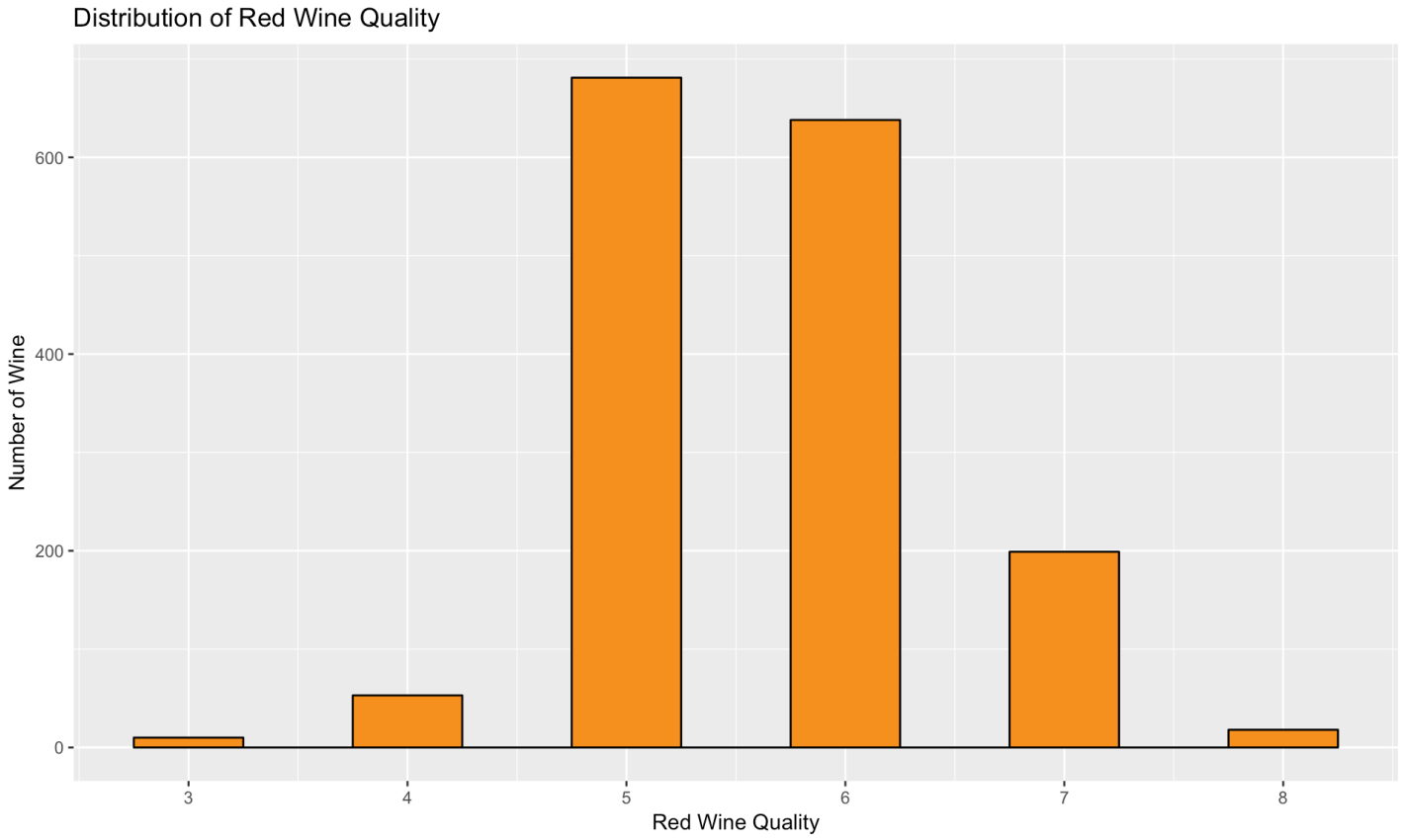
We see that alcohol and volatile acidity shows certain correlation: The higher degree the alcohol is, the percentage of volatile adicidy would be lower. Alcohol and citric acid seems to have also certain correlation.

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

---

# Final Plots and Summary

## Plot One

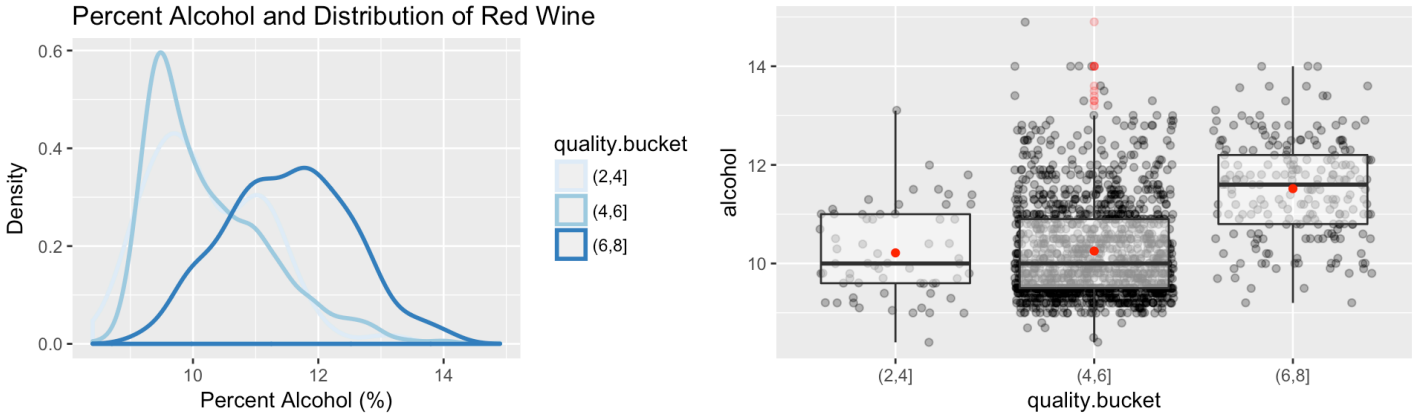



## Description One

The distribution of red wine quality ranges from quality 3 to quality 8, where a higher number indicates a better quality. The distribution is little bit left skewed. In the observation dataset the median wine quality is 6.

# Plot Two

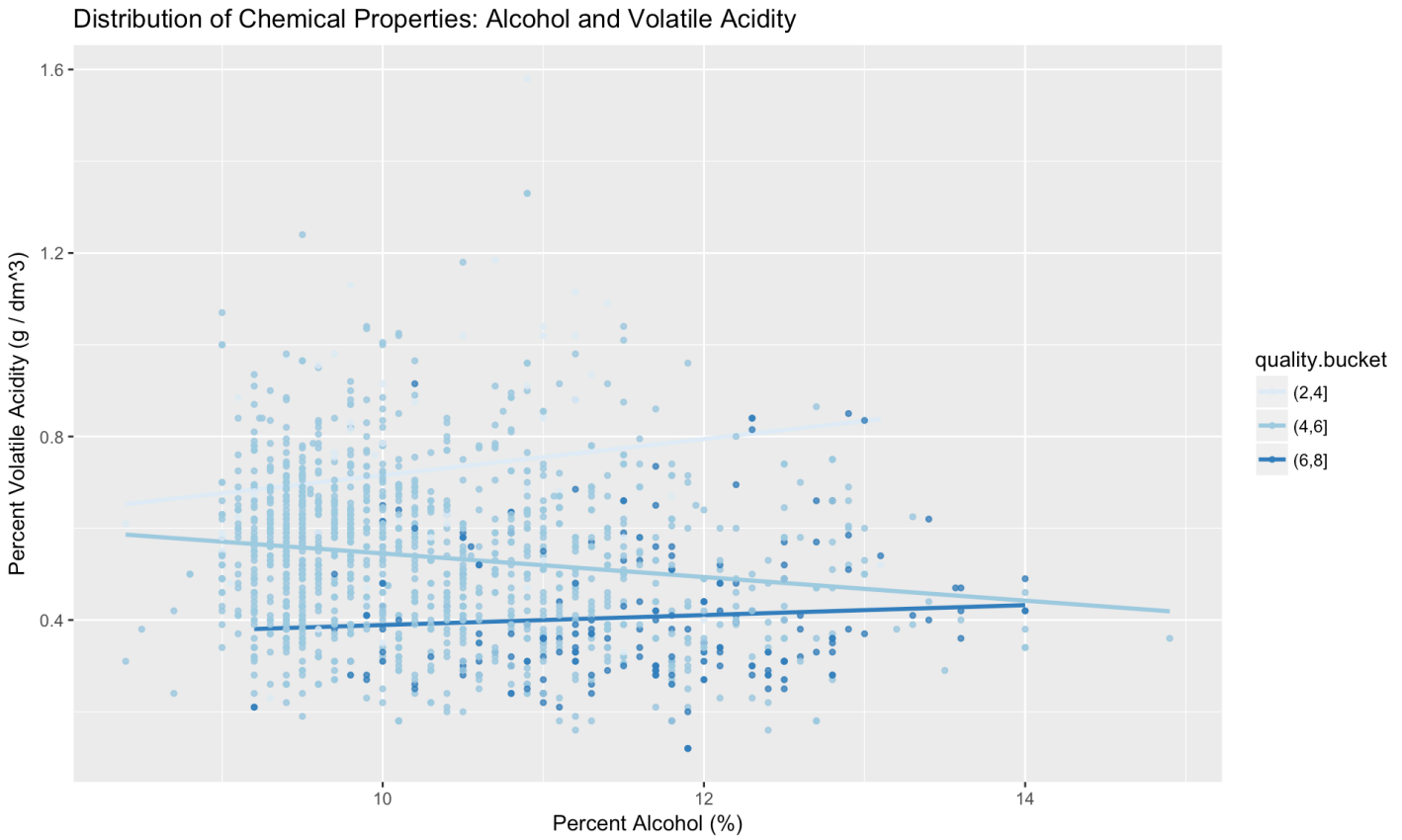
```
## $title
## [1] "Wine Quality group by Alcohol"
##
## $subtitle
## NULL
##
## attr(,"class")
## [1] "labels"
```




# Description Two

To make the visulization cleaner, quality.bucket instead of quality will be applied to build groups. We can see that the distribution between the quality bucket (2,4] and (4,6] in terms of alcohol have some overlapping together. This means that there would not be much difference in percent alcohol for wine in low quality and medium quality. However, a good quality wine within bucket (6,8] shows on average a higher percentage of alcohol.

# Plot Three




## Description Three

We can briefly see some center of plots for different quality bucket:

- Wine in good quality bucket (6,8] is more likely to have higher percentage of alcohol (between 11% and 13% with estimation) and lower percentage of volatile acidity (between 0.25g / dm^3 and 0.6g / dm^3 with estimation).
- Wine in low quality bucket (2,4] on average shows lower percentage of alcohol and higher percentage of volatile acidity.
- Wine in medium quality bucket (4,6] on average contains less percent of volatile acidity than wine in low quality bucket. At the same time it also contains less percentage of alcohol than wine in high quality bucket (6,8].

---

# Reflection

The red wine quality is a very interesting topic to look into as red wine comsumption is very common. From the analysis we find out four variables which are relatively strong indicators for the wine quality: Alcohol, volatile acidity, sulphates and citric acid.

One of the challenge of doing this analysis is that as a layman for wine chemical process and terminology, it is neccessary that I first spent some time google about the terminology in the dataset before I started to analyse it. This has help a lot in understanding the analysis.

It is surprising that the variable alcohol is strongly correlative with wine quality, as I thought that red wine is always with about 13.5% alcohol.

In the following a review of a few factors in the analysis which should be considered and improved in further study:

- Quanlity of Sample: The distrition of wine quality is quite unbalanced. With the 1599 observations in the dataset, most of the wine falls into the medium quality bucket (4,6]. Only 63 out of 1599 wine are within the low quality bucket, which could make the analysis for wine in low quality bucket not accuracy.
- Features which are interacting with each other: One of the challenge here is the chemical terminology and process. As a layman for the wine chemical process, it is difficult to indentify the interaction / relationhsip between different chemical properties. With more knowledge in the wine chemical process, the further study can group certain variables together / create new variables out of them to reduce the dimension.
- Last but not the least, the result of the study can be more comprehensive with interdisciplinary knowledge. For example, in the analysis we see that percentage of sulphates is positively correlative with wine quality. However, it would make sense that the analysis goes beyond merely statistical method and noting that there is also arguments on the benefit of sulfate-free products.

# Sources


- Chestofbooks: http://chestofbooks.com/food/beverages/Alcohol-Properties/Fixed-Acidity.html (http://chestofbooks.com/food/beverages/Alcohol-Properties/Fixed-Acidity.html)
- Wikipedia: https://en.wikipedia.org/wiki/Acids_in_wine#Citric_acid (https://en.wikipedia.org/wiki/Acids_in_wine#Citric_acid)
- Wikipedia: https://en.wikipedia.org/wiki/Sweetness_of_wine#Residual_sugar (https://en.wikipedia.org/wiki/Sweetness_of_wine#Residual_sugar)
- Wikipedia: https://en.wikipedia.org/wiki/Category:Chlorides (https://en.wikipedia.org/wiki/Category:Chlorides)
- Wikipedia: https://en.wikipedia.org/wiki/Sulfur_dioxide (https://en.wikipedia.org/wiki/Sulfur_dioxide)
- Thekichn: http://www.thekitchn.com/the-truth-about-sulfites-in-wine-myths-of-red-wine-headaches-100878 (http://www.thekitchn.com/the-truth-about-sulfites-in-wine-myths-of-red-wine-headaches-100878)
- BBR-wine-knowledge: https://www.bbr.com/wine-knowledge/faq-quality#betterquality (https://www.bbr.com/wine-knowledge/faq-quality#betterquality)
- Little-book-of-r: http://little-book-of-r-for-multivariate-analysis.readthedocs.io/en/latest/src/multivariateanalysis.html (http://little-book-of-r-for-multivariate-analysis.readthedocs.io/en/latest/src/multivariateanalysis.html)
- Beginning R: An Introduction to Statistical Programming. Author: Larry Pace, Joshua Wiley
- http://www.stat.wisc.edu/~larget/stat302/chap2.pdf (http://www.stat.wisc.edu/~larget/stat302/chap2.pdf)
- https://briatte.github.io/ggcorr/#controlling-the-coefficient-labels (https://briatte.github.io/ggcorr/#controlling-the-coefficient-labels)