White Wine Quality Exploration

Citation: This dataset is public available for research. The details are described in [Cortez et al., 2009]. P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.

Question: With different variables in the white wine dataset, the analysis will identify chemical properties that influence the quality of white wines.

Univariate Plots Section

General exploration

```
## '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 dataset has 4,898 rows and 12 variables.

```
##  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
##  Min.   :3.000
##  1st Qu.:5.000
```

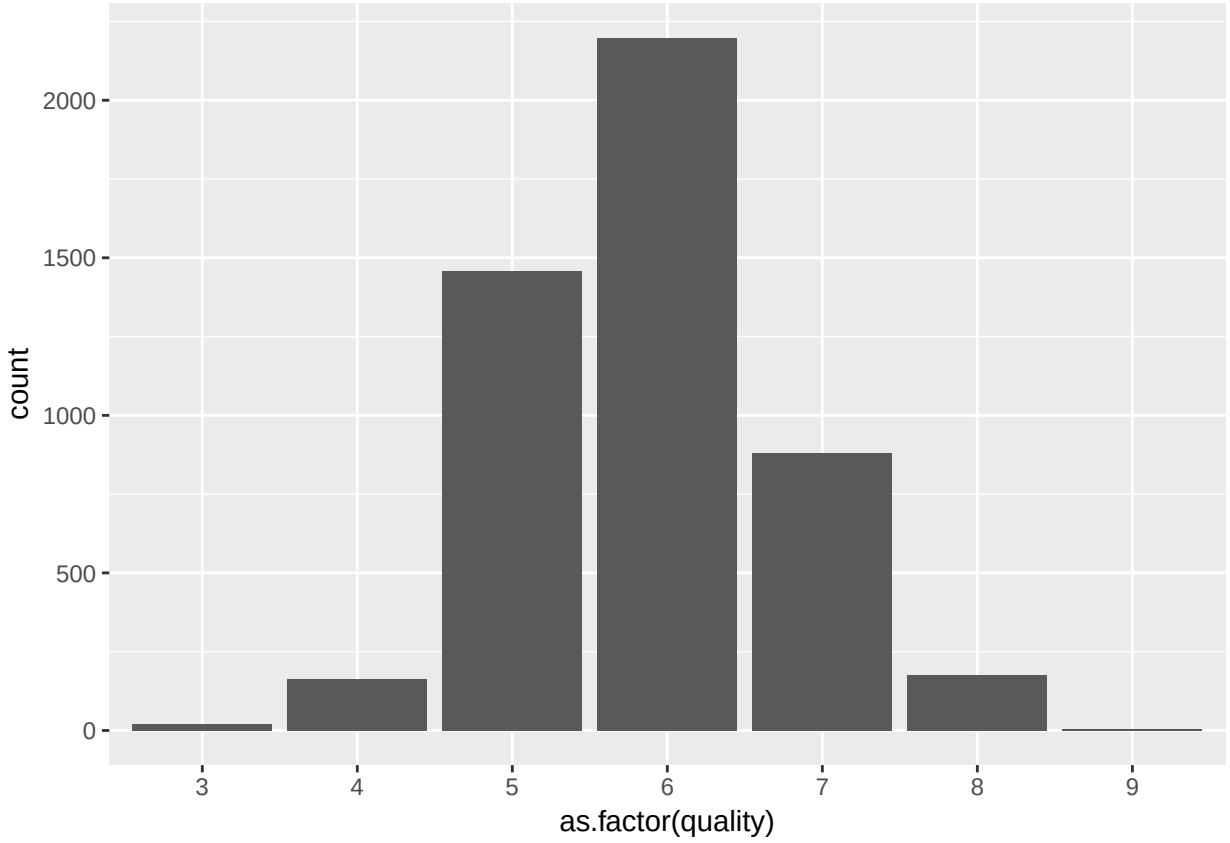
```
##  Median :6.000
##  Mean   :5.878
##  3rd Qu.:6.000
##  Max.   :9.000
```

Some variables (residual sugar, free sulfur dioxide, total sulfur dioxide) have max values that are far from median, which might be outliers.

Quality exploration

```
##
##    3    4    5    6    7    8    9
##   20  163 1457 2198  880  175    5
```

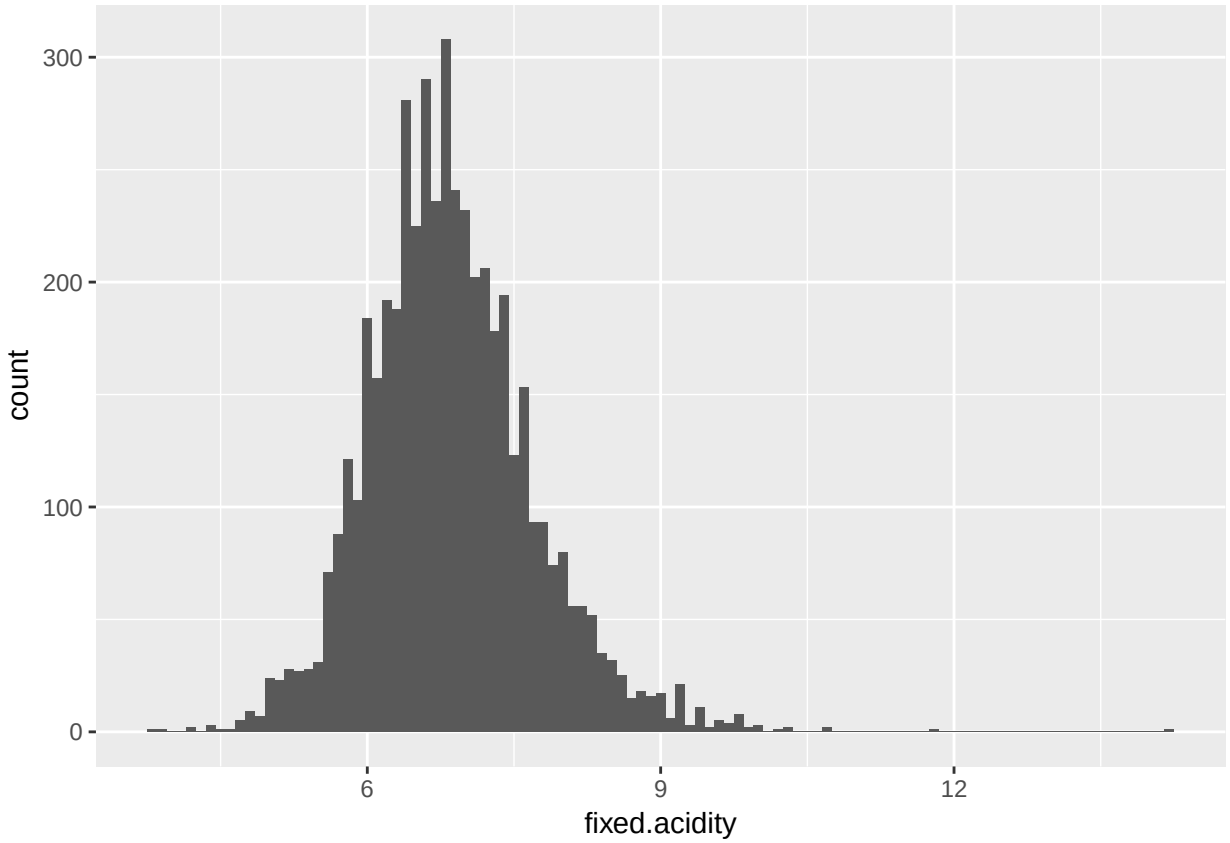
Most white wine are with quality of 6. There are few very good quality or bad quality white wine in the market.



The quality distribution looks very close to normal distribution.

Fixed acidity exploration

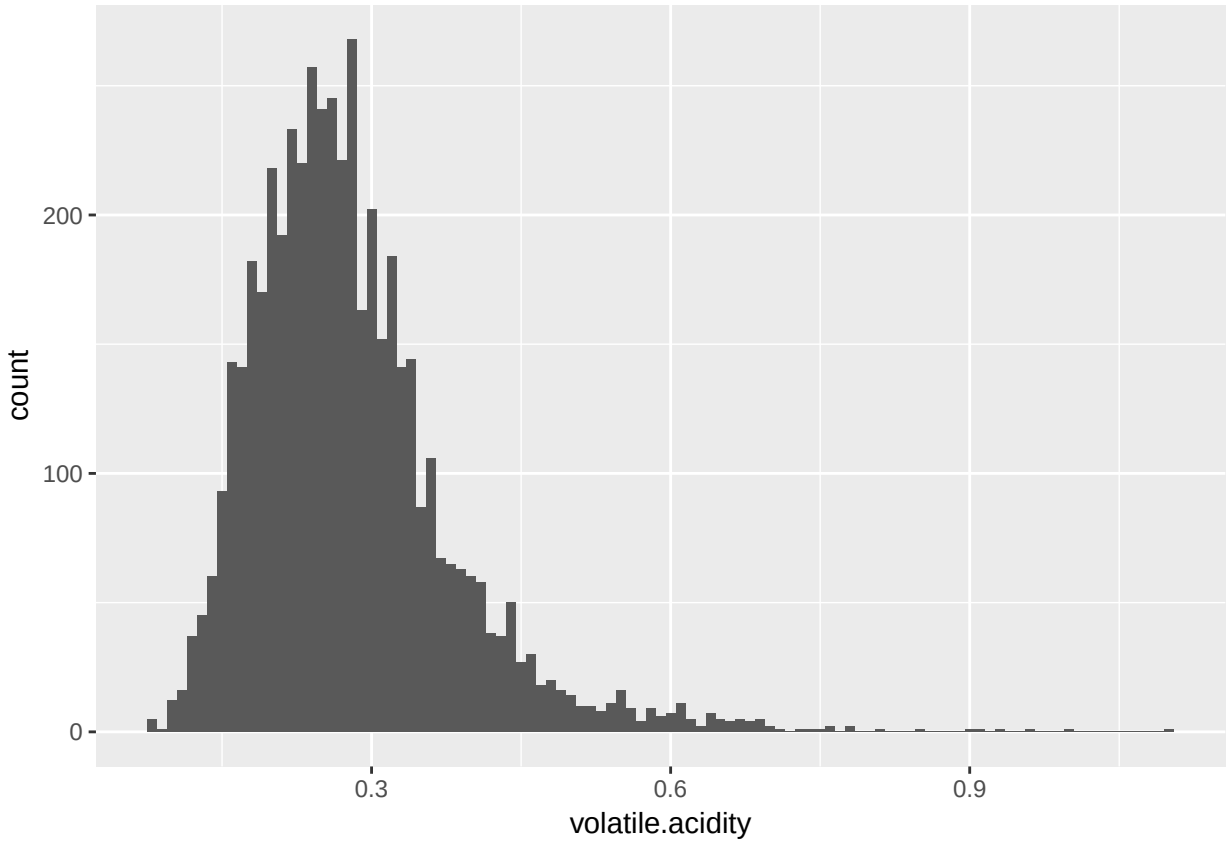
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.800   6.300   6.800   6.855   7.300  14.200
```



Most of the wine have fixed acidity between 6.3 - 7.3g/ dmˆ3: median 6.8g/ dmˆ3 and mean 6.855 g/ dmˆ3. The fixed acidity looks follow the normal distribution.

Volatile acidity exploration

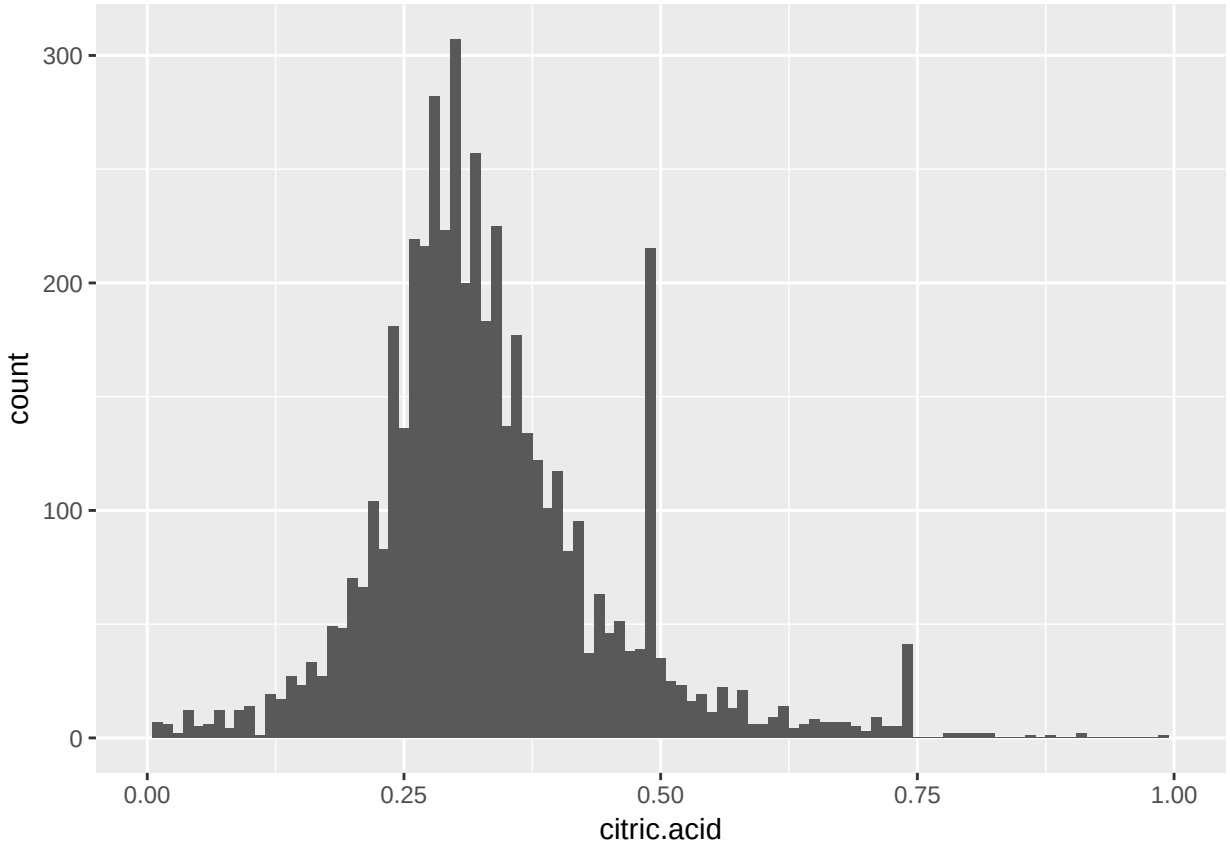
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.2100  0.2600  0.2782  0.3200  1.1000
```



Most of the wine have volatile acidity between 0.21 - 0.32g/ dmˆ3: median 0.26g/ dmˆ3 and mean 0.2782 g/ dmˆ3. The volatile acidity looks follow the normal distribution.

Citric acid exploration

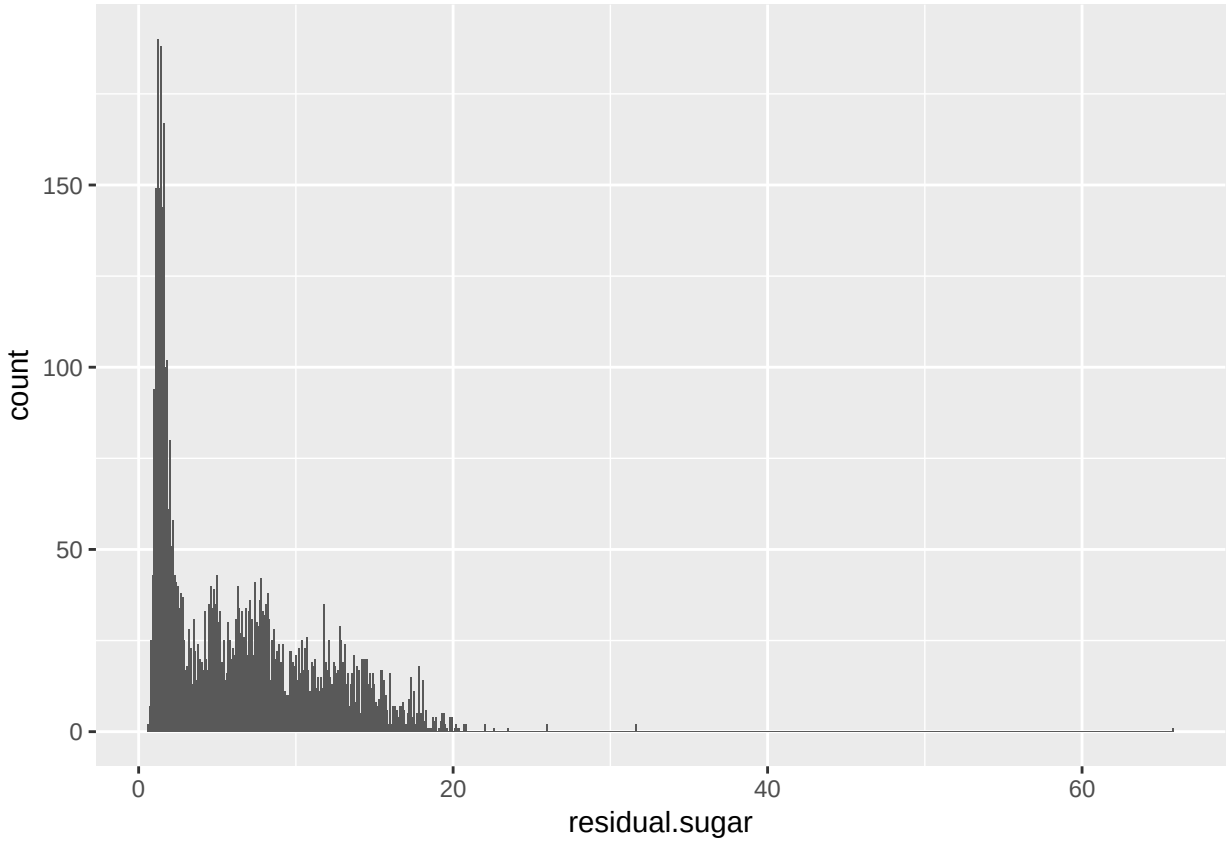
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2700  0.3200  0.3342  0.3900  1.6600
```



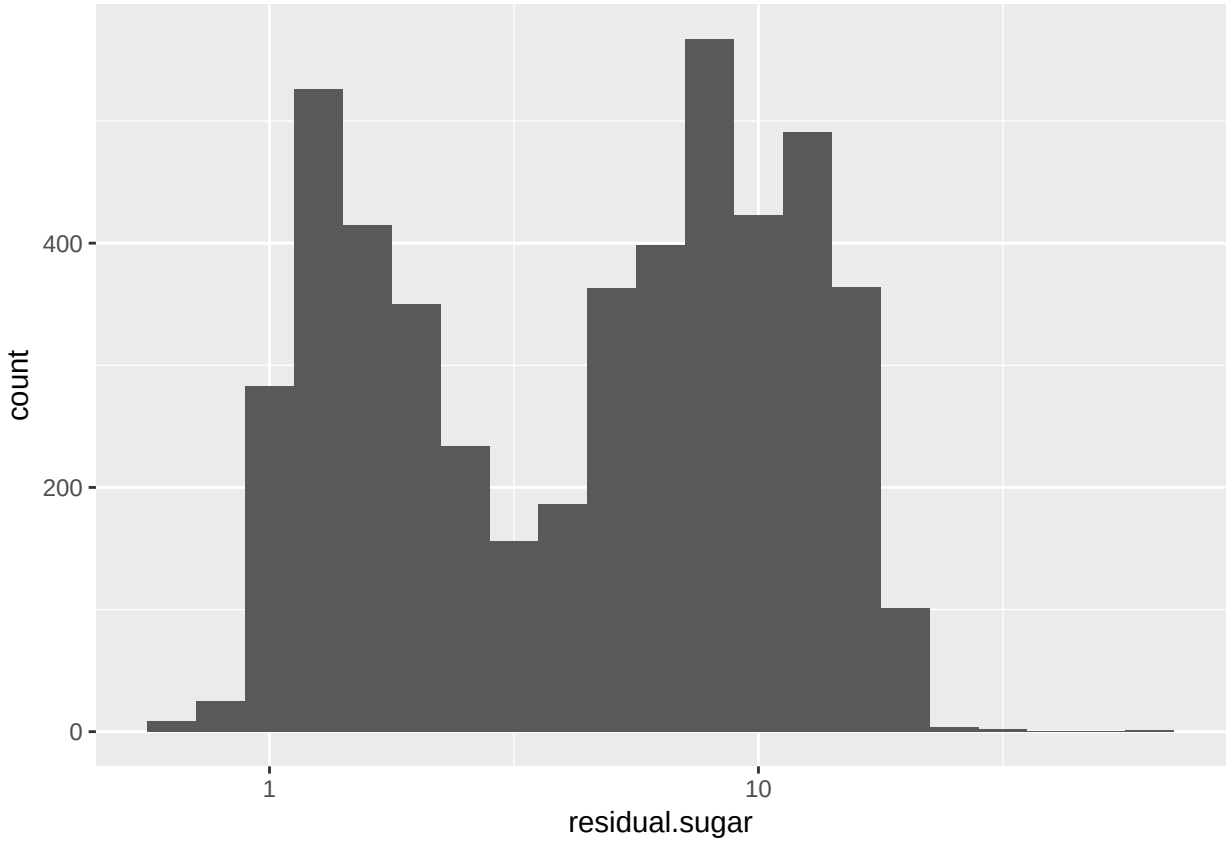
Most of the wine have citric acid between 0.27 - 0.39g/ dmˆ3: median 0.32g/ dmˆ3 and mean 0.3342 g/ dmˆ3. The citric acid looks follow the normal distribution but with an unsual spike at 0.49g/ dmˆ3 and 0.74g/ dmˆ3.

Residual sugar exploration

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.600   1.700   5.200   6.391   9.900  65.800
```



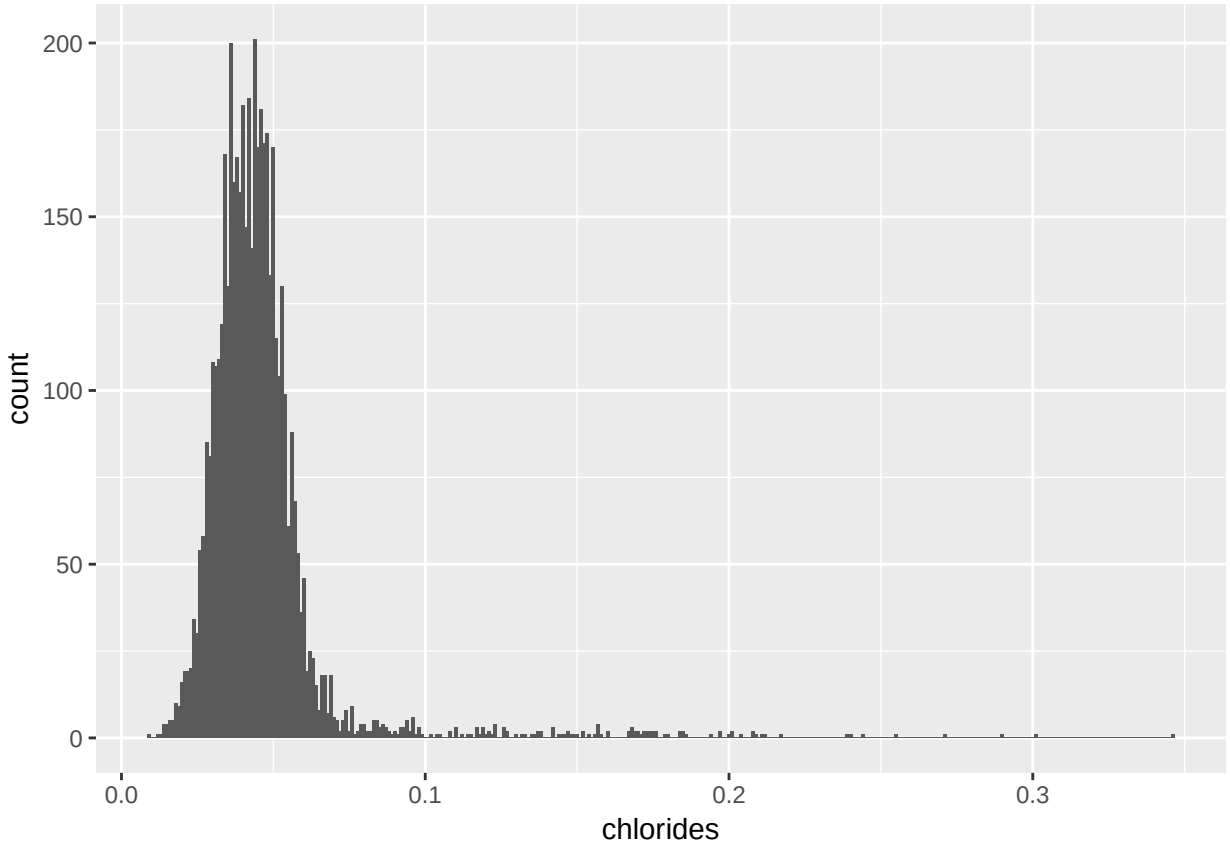
Most of the wine have residual sugar between 1.7 - 9.9g/ dmˆ3: median 5.2g/ dmˆ3 and mean 6.39 g/ dmˆ3. The residual sugar has a right tailed chart and I'd like to see the distribution of transformed variable.



The transformed residual sugar distribution appears bimodal. There might be two group of wine, one group are more sweet than the other one.

Chlorides exploration

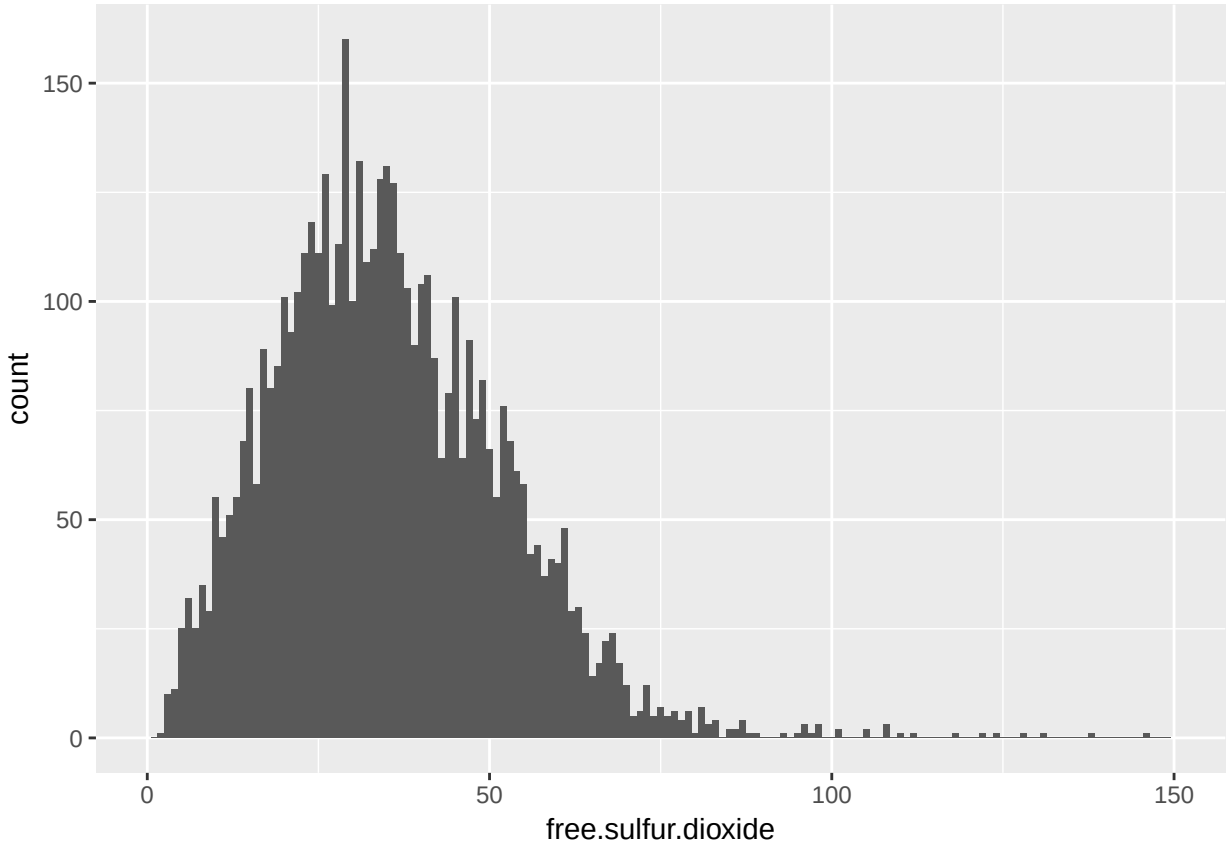
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00900 0.03600 0.04300 0.04577 0.05000 0.34600
```



Most of the wine have chlorides between 0.036 - 0.05g/ dmˆ3: median 0.043g/ dmˆ3 and mean 0.057 g/ dmˆ3. The chlorides looks follow the normal distribution but with outliers above 0.1g/ dmˆ3.

Free sulfur dioxide exploration

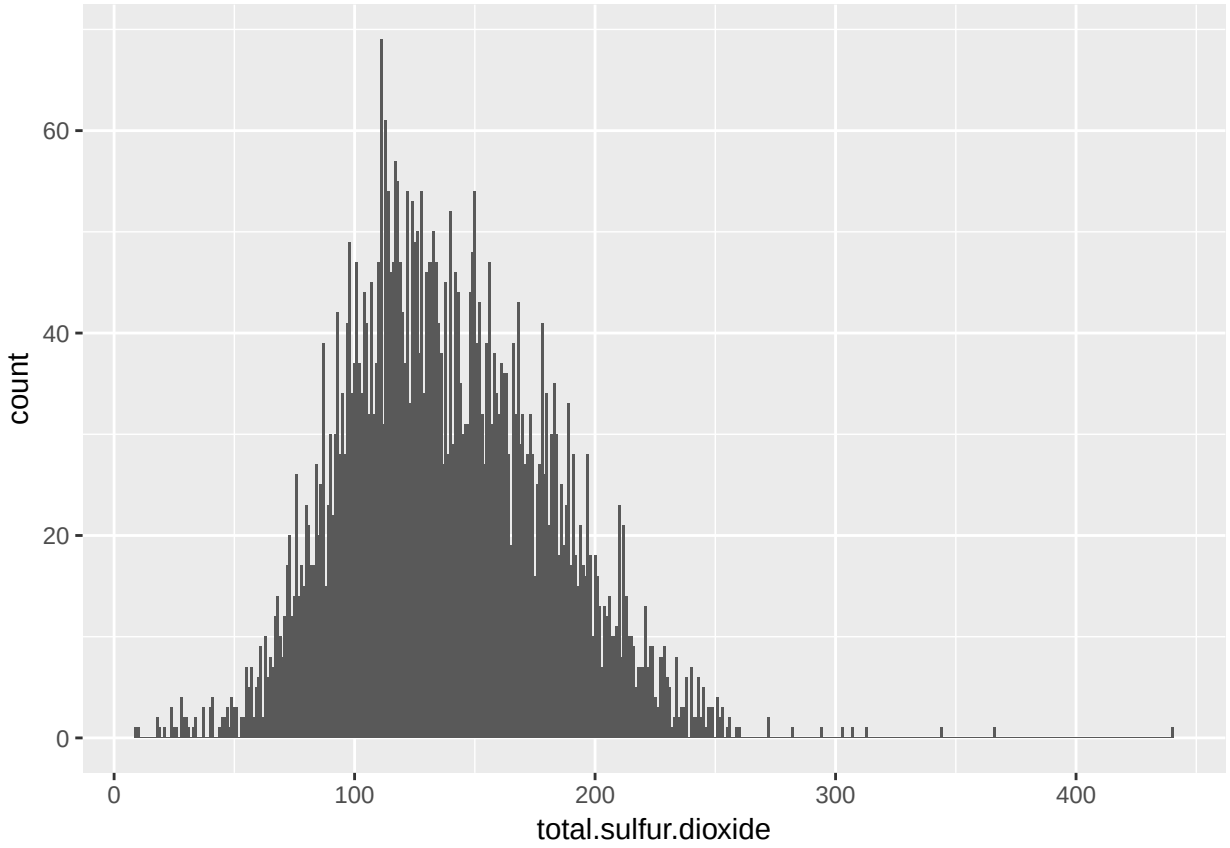
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   23.00   34.00   35.31   46.00  289.00
```



Most of the wine have free.sulfur.dioxide between 23 - 46g/ dmˆ3: median 34g/ dmˆ3 and mean 35.31 g/ dmˆ3. The free.sulfur.dioxide looks follow the normal distribution.

Total sufur dioxide exploration

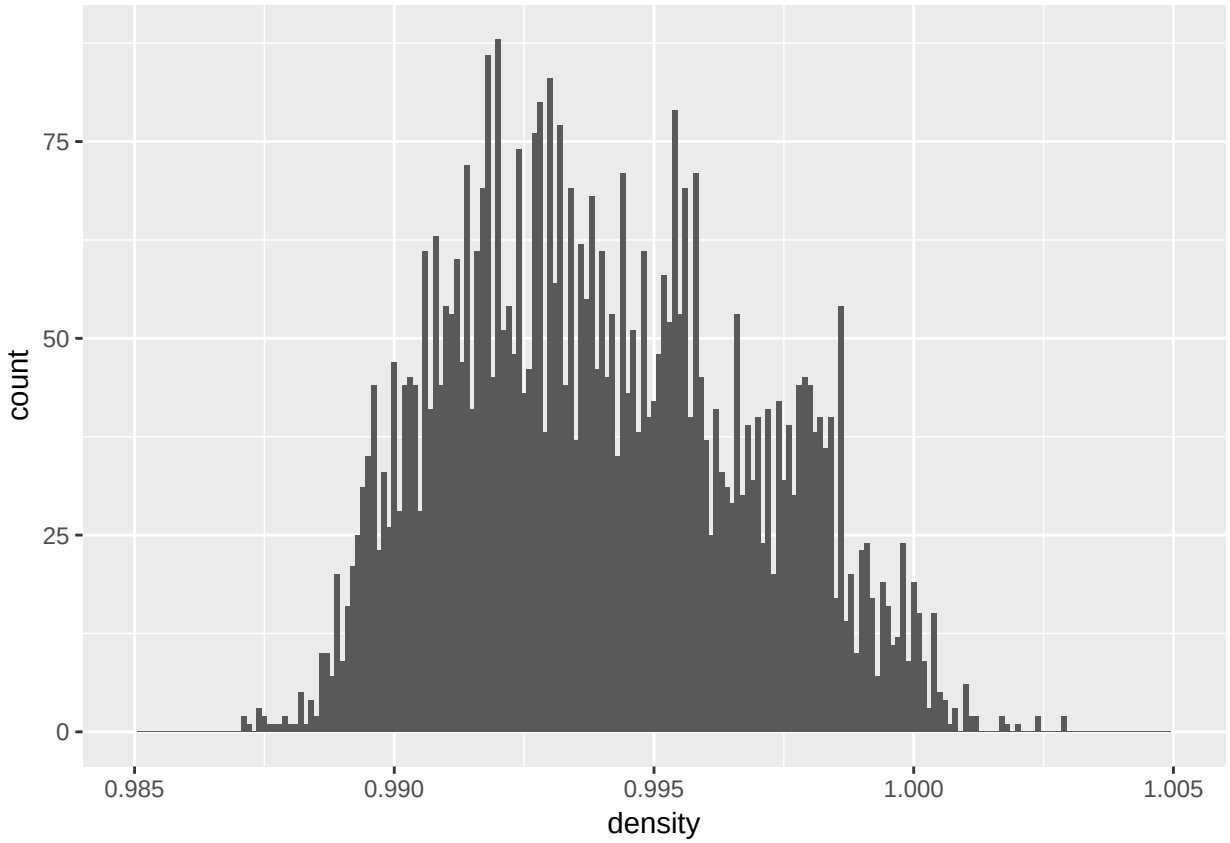
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.0   108.0   134.0   138.4   167.0   440.0
```



Most of the wine have total sulfur dioxide between 108 - 167g/ dmˆ3: median 134g/ dmˆ3 and mean 138.4 g/ dmˆ3. The total sulfur dioxide looks follow the normal distribution.

Density exploration

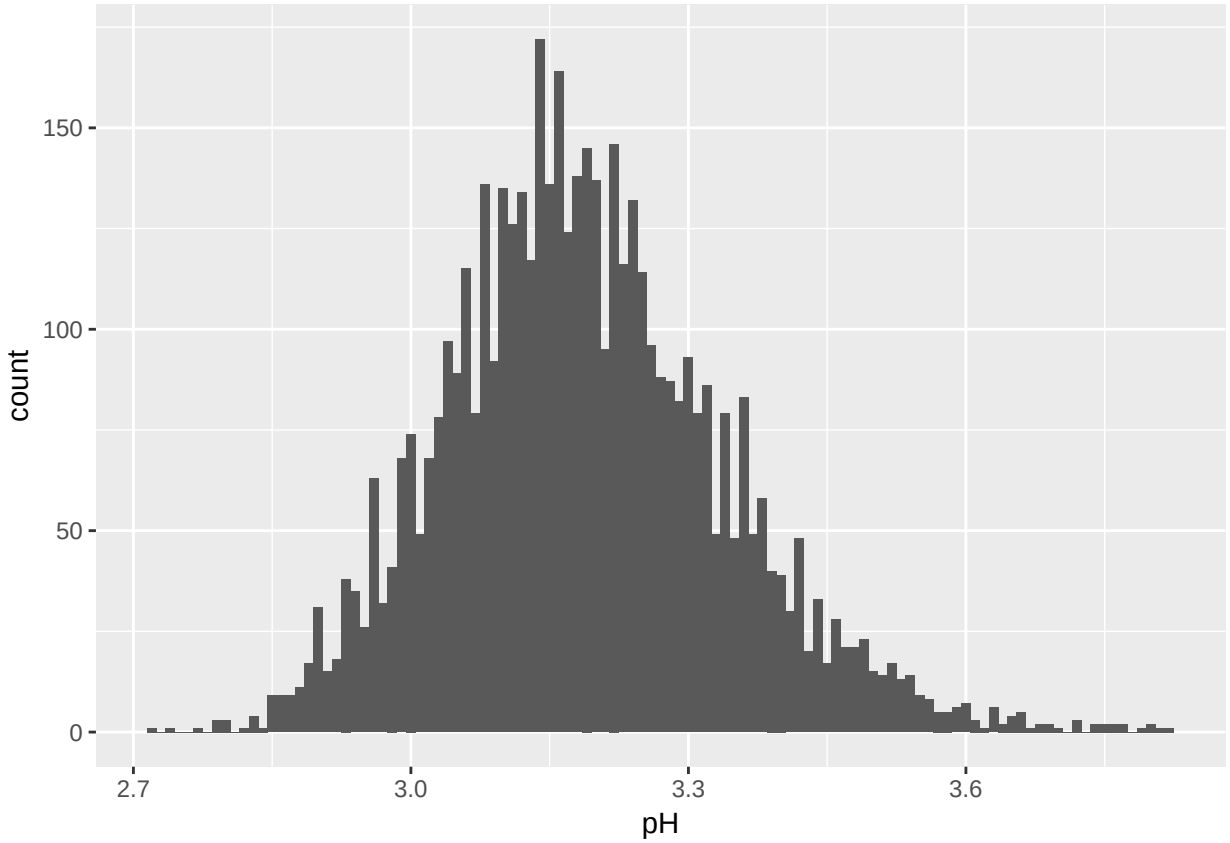
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9917  0.9937  0.9940  0.9961  1.0390
```



Almost all the density of the wine are close to 1.

PH exploration

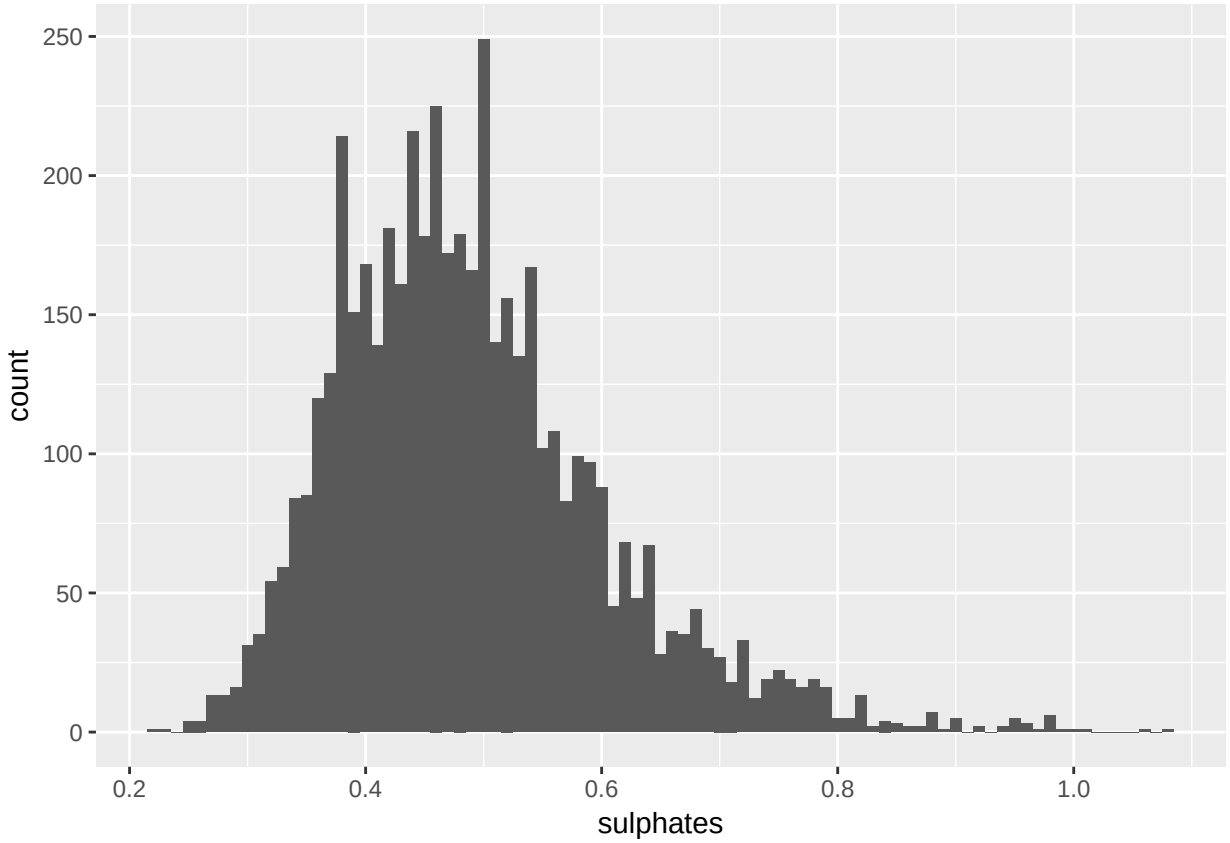
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.720   3.090   3.180   3.188   3.280   3.820
```



Most of the wine have pH value between 3 - 3.3: median 3.18 and mean 3.188. The pH looks follow the normal distribution.

Sulphates exploration

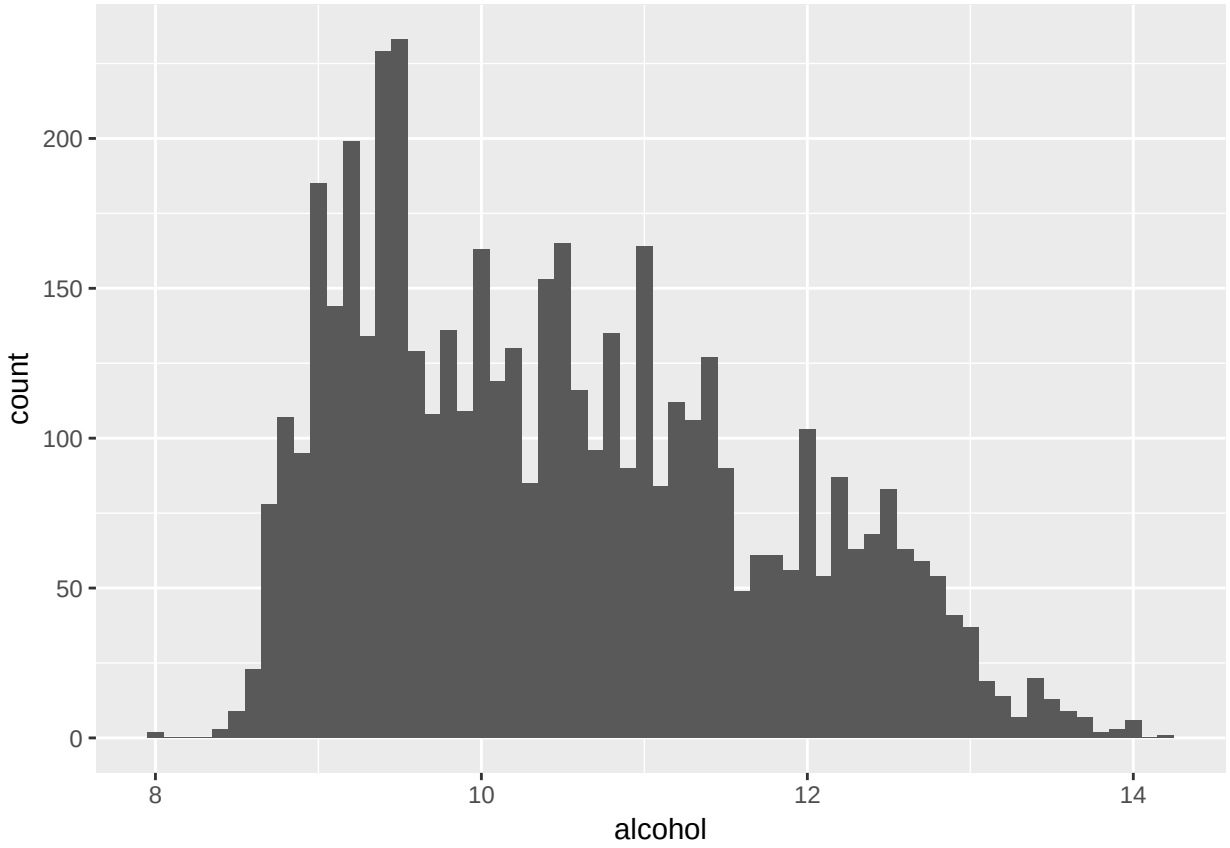
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2200  0.4100  0.4700  0.4898  0.5500  1.0800
```



Most of the wine have sulphates value between 0.41 - 0.55g/ dmˆ3: median 0.47g/ dmˆ3 and mean 0.489g/ dmˆ3 The sulphates looks follow the normal distribution.

Alcohol exploration

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.50   10.40   10.51   11.40   14.20
```



Most of the wine have alcohol value between 9.5 - 11.4g/ dmˆ3: median 10.4g/ dmˆ3 and mean 10.51g/ dmˆ3 The alcohol looks follow the normal distribution.

Univariate Analysis

What is the structure of your dataset?

There are 4,898 white wines with 11 variables(fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol).

Other observations:

- Most of the variable follow normal distribution except the residual sugar.
- Most of the wine are quality 6.
- Most of the wine have PH between 3 - 3.3.

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

Quality is the dependent variable and will be the main feature. I'd like to find the features that best predict the quality of the wine. Currently, I guess the

1. volatile acidity, which will lead to unpleasant scent, and
2. free sulfur dioxide, which can prevent microbial growth and the oxidation of wine,

may be the key factors to determine wine quality.

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

- Citric acid
- Residual Sugar
- Alcohol

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

No.

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?

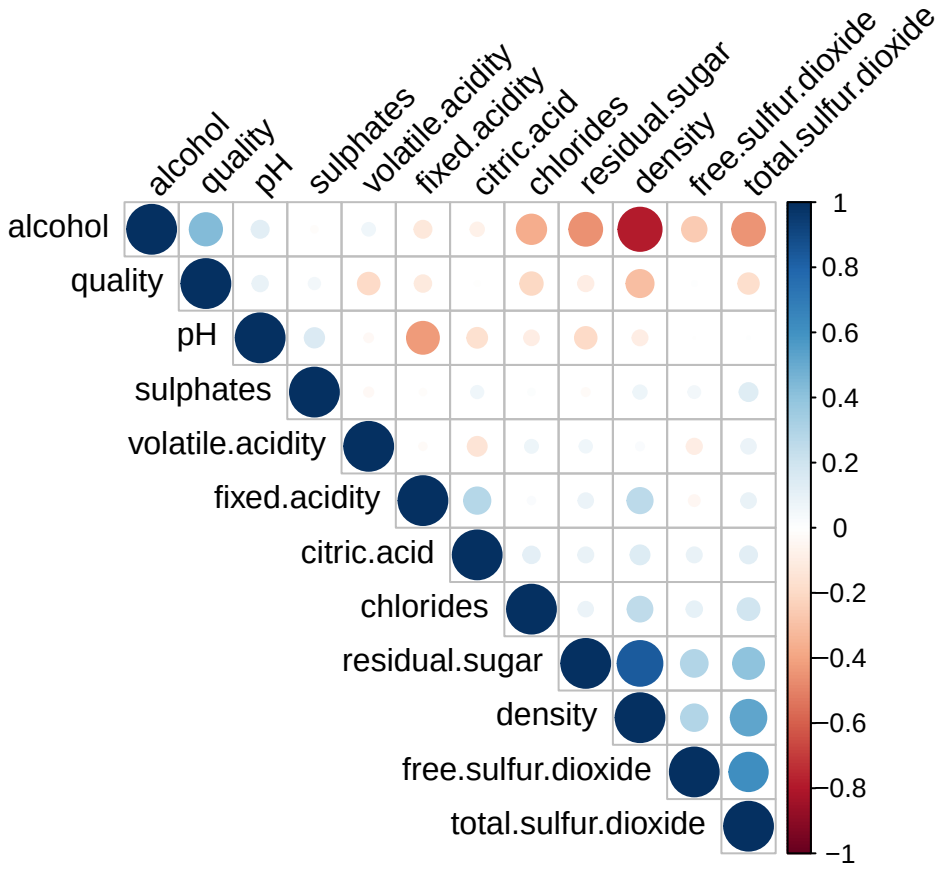
Residual sugar is right tailed and I made log transformation on it to find a normal distribution.

Bivariate Plots Section

Correlation

```
##                       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
##                       residual.sugar   chlorides free.sulfur.dioxide
## fixed.acidity            0.08902070  0.02308564       -0.0493958591
## volatile.acidity         0.06428606  0.07051157       -0.0970119393
## citric.acid              0.09421162  0.11436445        0.0940772210
## residual.sugar           1.00000000  0.08868454        0.2990983537
## chlorides                0.08868454  1.00000000        0.1013923521
## free.sulfur.dioxide      0.29909835  0.10139235        1.0000000000
## total.sulfur.dioxide     0.40143931  0.19891030        0.6155009650
## density                  0.83896645  0.25721132        0.2942104109
## pH                      -0.19413345 -0.09043946       -0.0006177961
## sulphates               -0.02666437  0.01676288        0.0592172458
## alcohol                 -0.45063122 -0.36018871       -0.2501039415
## quality                 -0.09757683 -0.20993441        0.0081580671
##                       total.sulfur.dioxide     density           pH
## fixed.acidity                  0.091069756  0.26533101 -0.4258582910
## volatile.acidity               0.089260504  0.02711385 -0.0319153683
## citric.acid                    0.121130798  0.14950257 -0.1637482114
```

```
## residual.sugar               0.401439311  0.83896645 -0.1941334540
## chlorides                    0.198910300  0.25721132 -0.0904394560
## free.sulfur.dioxide          0.615500965  0.29421041 -0.0006177961
## total.sulfur.dioxide         1.000000000  0.52988132  0.0023209718
## density                      0.529881324  1.00000000 -0.0935914935
## pH                           0.002320972 -0.09359149  1.0000000000
## sulphates                    0.134562367  0.07449315  0.1559514973
## alcohol                     -0.448892102 -0.78013762  0.1214320987
## quality                     -0.174737218 -0.30712331  0.0994272457
##                           sulphates      alcohol      quality
## fixed.acidity        -0.01714299 -0.12088112 -0.113662831
## volatile.acidity     -0.03572815  0.06771794 -0.194722969
## citric.acid           0.06233094 -0.07572873 -0.009209091
## residual.sugar       -0.02666437 -0.45063122 -0.097576829
## chlorides             0.01676288 -0.36018871 -0.209934411
## free.sulfur.dioxide   0.05921725 -0.25010394  0.008158067
## total.sulfur.dioxide  0.13456237 -0.44889210 -0.174737218
## density               0.07449315 -0.78013762 -0.307123313
## pH                    0.15595150  0.12143210  0.099427246
## sulphates             1.00000000 -0.01743277  0.053677877
## alcohol              -0.01743277  1.00000000  0.435574715
## quality               0.05367788  0.43557472  1.000000000
```



There are some relationships between variables from the correlation chart.

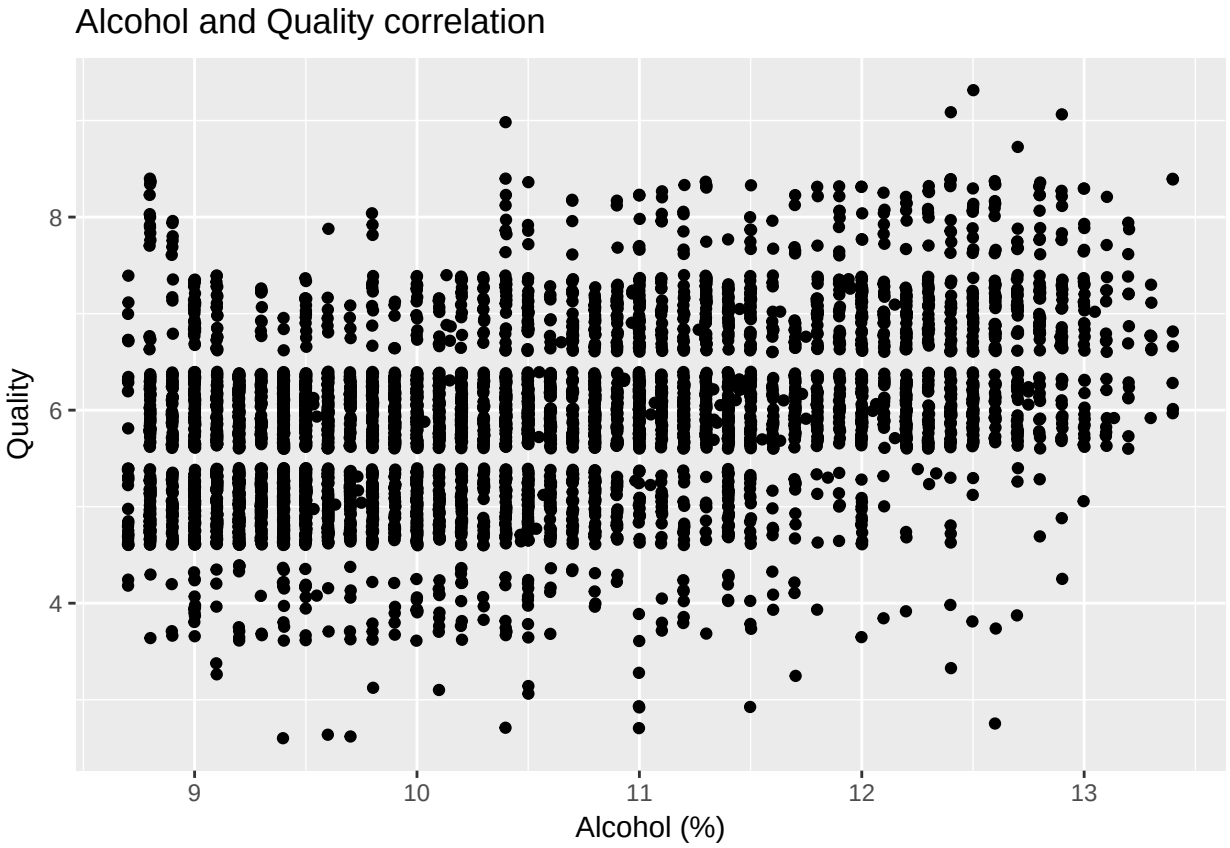
- fixed.acidity is moderately correlated to pH.
- Residual sugar is highly correlated to density and moderately correlated to total.sulfur.dioxide and

alcohol.
- chlorides is moderately correlated to alcohol.
- free.sulfur.dioxide is highly correlated to total.sulfur.dioxide and density.
- total.sulfur.dioxide is correlated to density, alcohol.
- density is highly correlated to alcohol.

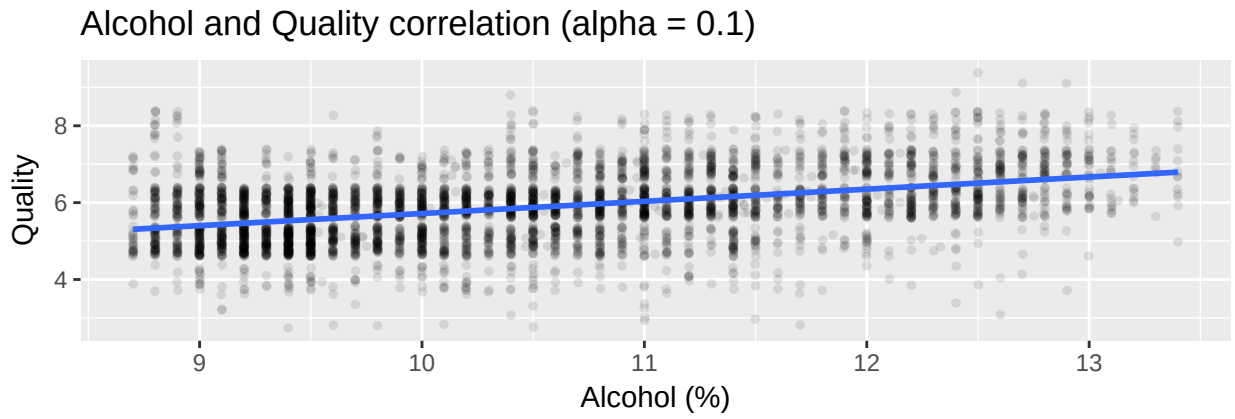
As I mentioned earlier, it looks that there are some outliers in the dataset and I'll remove top and bottom 1% of the data for the following analysis.

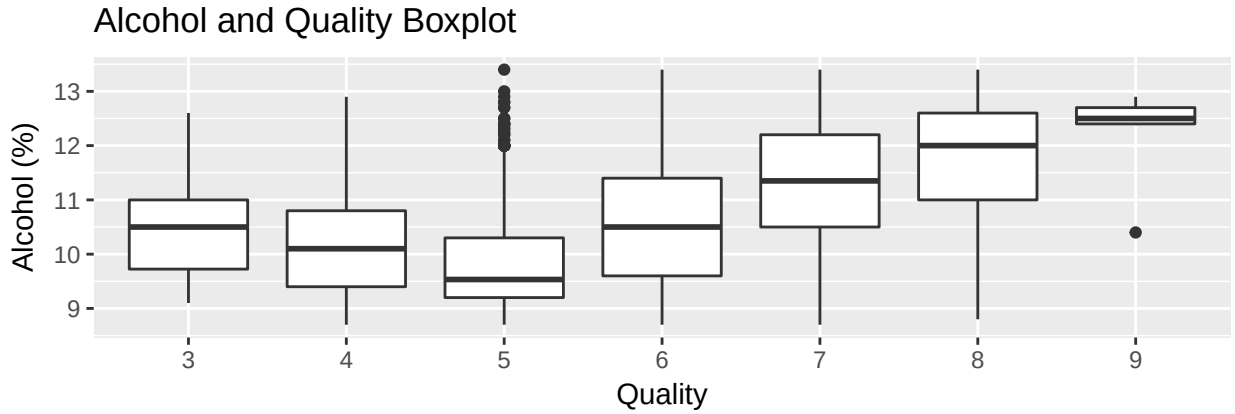
Since quality is the target variable and moderately correlated to alcohol and density, I want to explore on these two variables first.

Alcohol



The scatter plot has many overlapped points. I will adjust the alpha to have a clear chart.



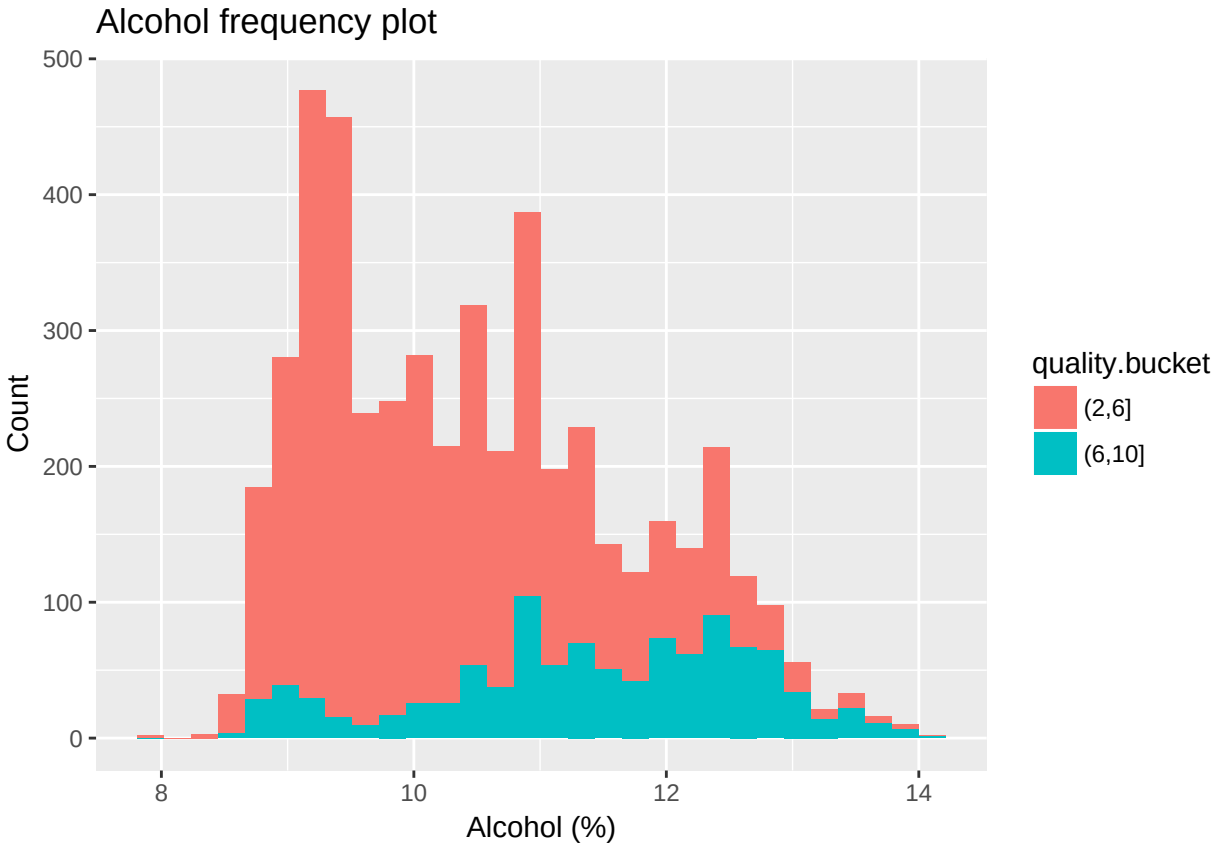


From both scatter plot with alpha adjustment and boxplot, it looks that higher quality wine usually have slightly higher alcohol. The alcohol might be one of the key factors for the Quality of wine.

```
## wine$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.55   10.45   10.35   11.00   12.60
## ------------------------------------------------------------
## wine$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.40   10.10   10.15   10.75   13.50
## ------------------------------------------------------------
## wine$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.000   9.200   9.500   9.809  10.300  13.600
## ------------------------------------------------------------
## wine$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50    9.60   10.50   10.58   11.40   14.00
## ------------------------------------------------------------
## wine$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.60   10.60   11.40   11.37   12.30   14.20
## ------------------------------------------------------------
## wine$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50   11.00   12.00   11.64   12.60   14.00
## ------------------------------------------------------------
```

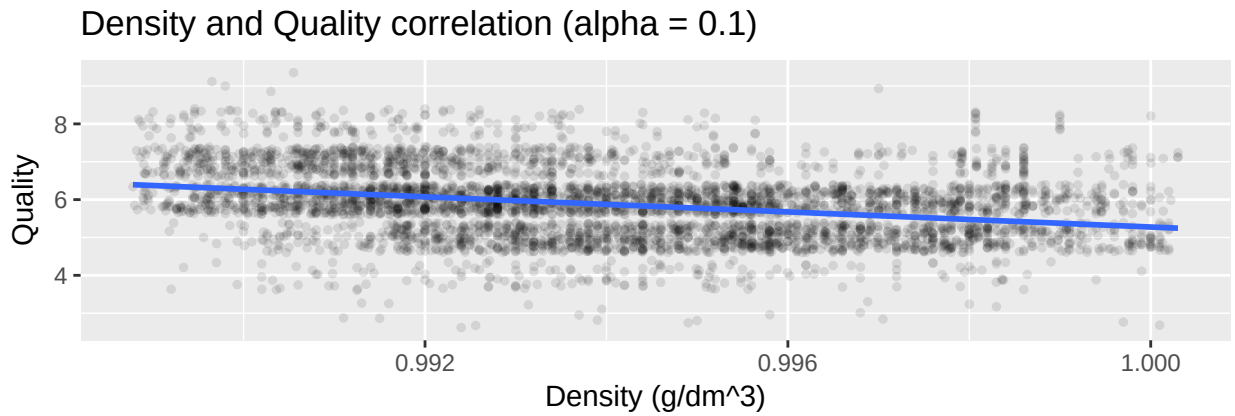
```
## wine$quality: 9
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   12.40   12.50   12.18   12.70   12.90
```

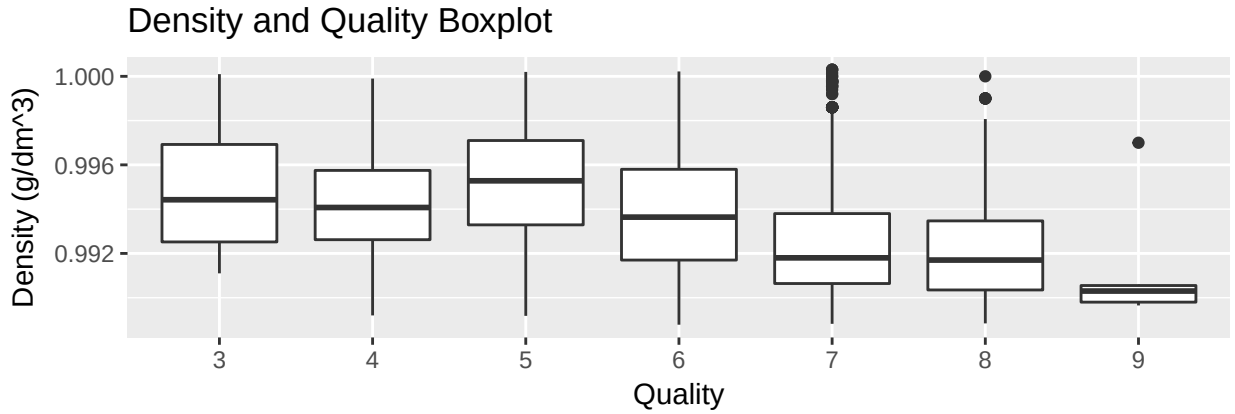
Another takeout is that for quality lower than 6, most alcohol of wines are less than 10.5%, while wines with quality above 6 typically have alcohol more than 11.4%.



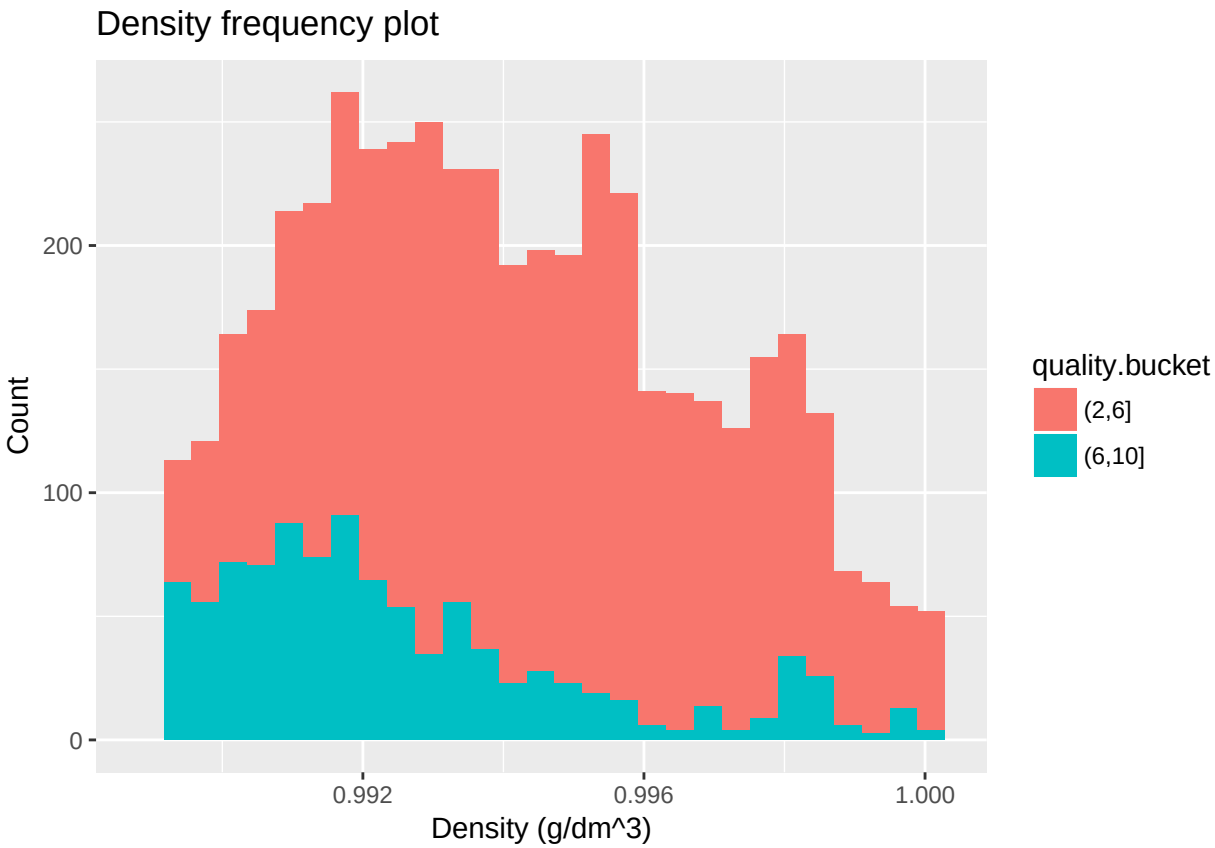
From the frequency plot, high quality wine skew to high alcohol value while low quality wine skew to low alcohol value.

Density



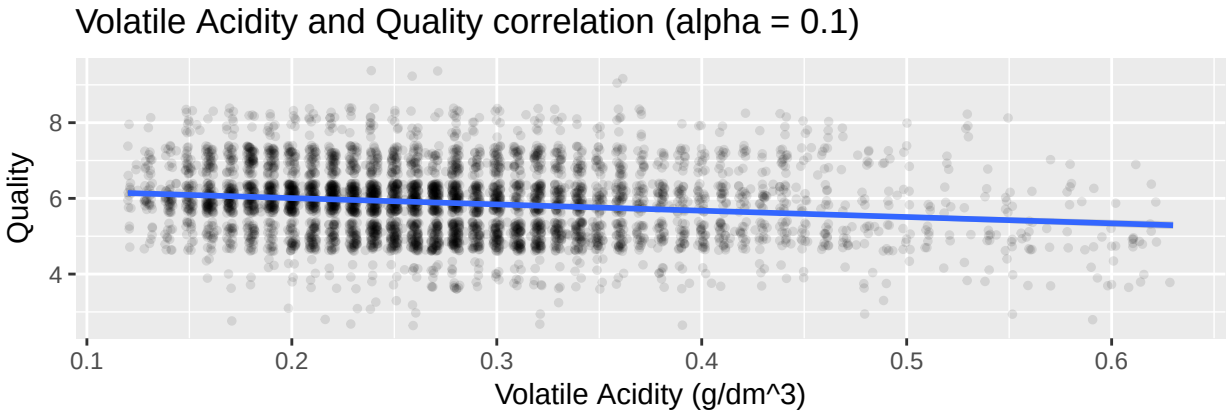


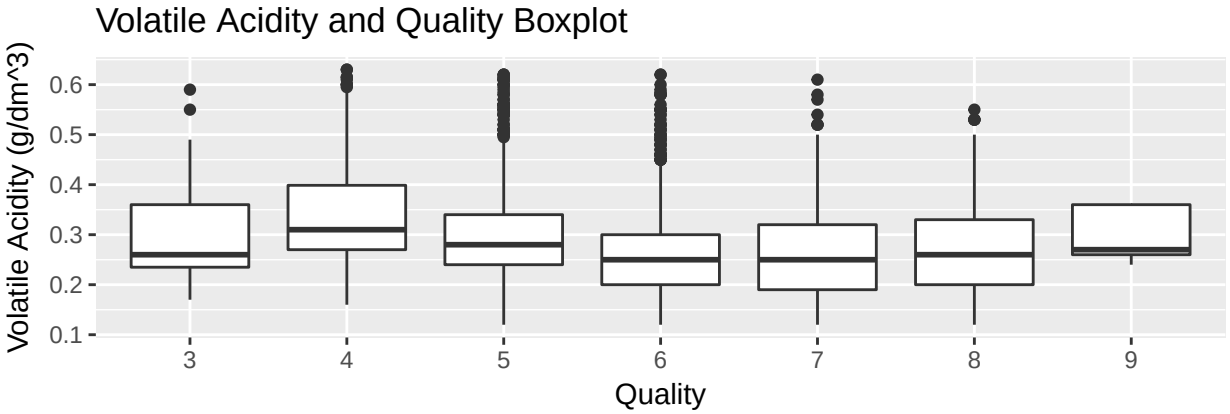
From both scatter plot with alpha adjustment and boxplot, it looks that higher quality wine usually have slightly lower density.



Besides alcohol and dnesity, I'd like to take a look at other interest variables (volatile acidity, free sulfur dioxide, citric acid, and residual sugar).

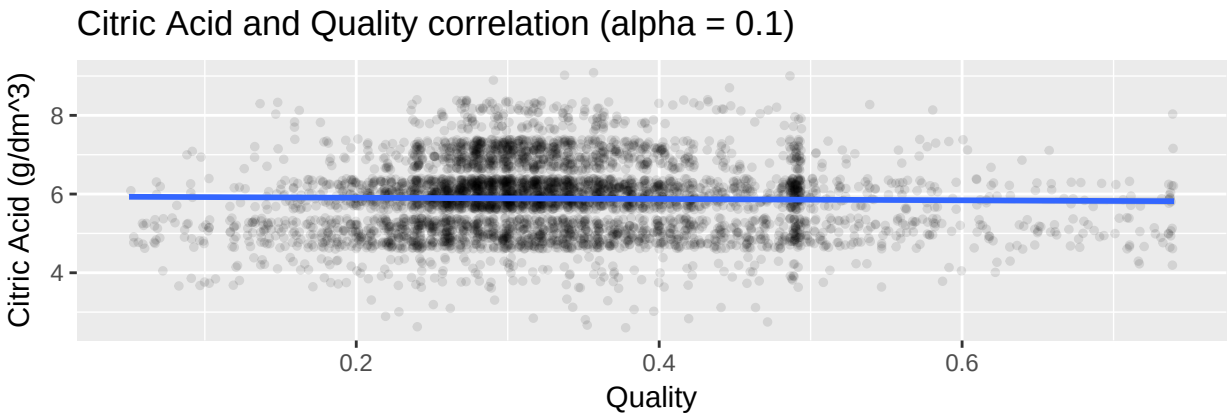
Volatile acidity

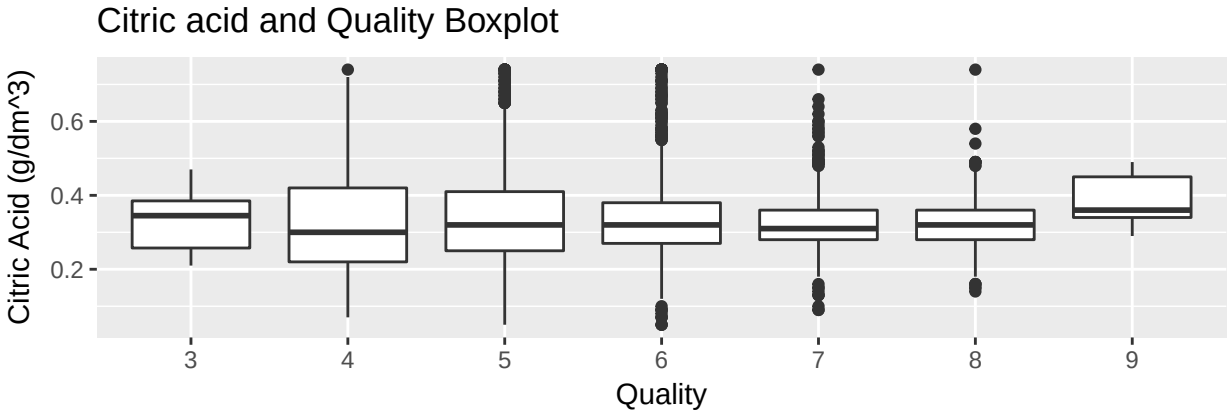




It seems that higher quality wines usually have lower volatile.acidity.

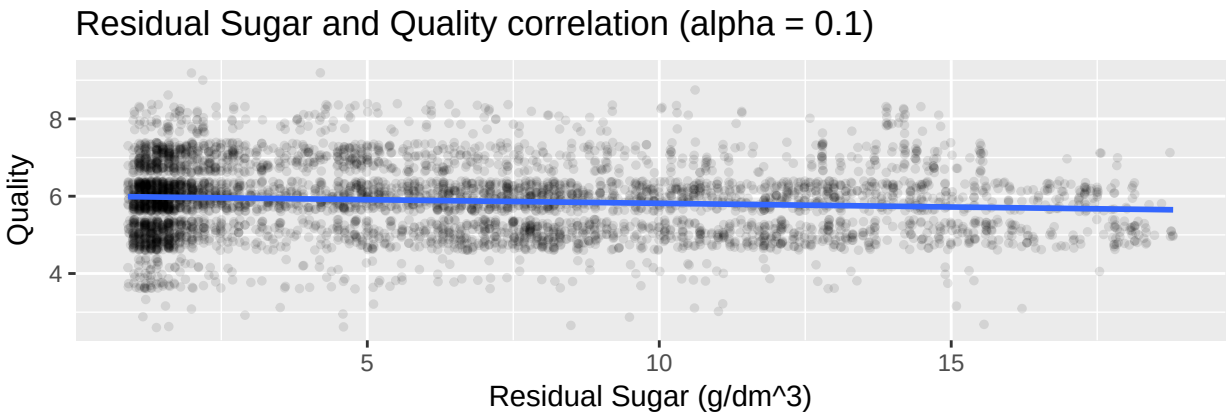
Citric acid

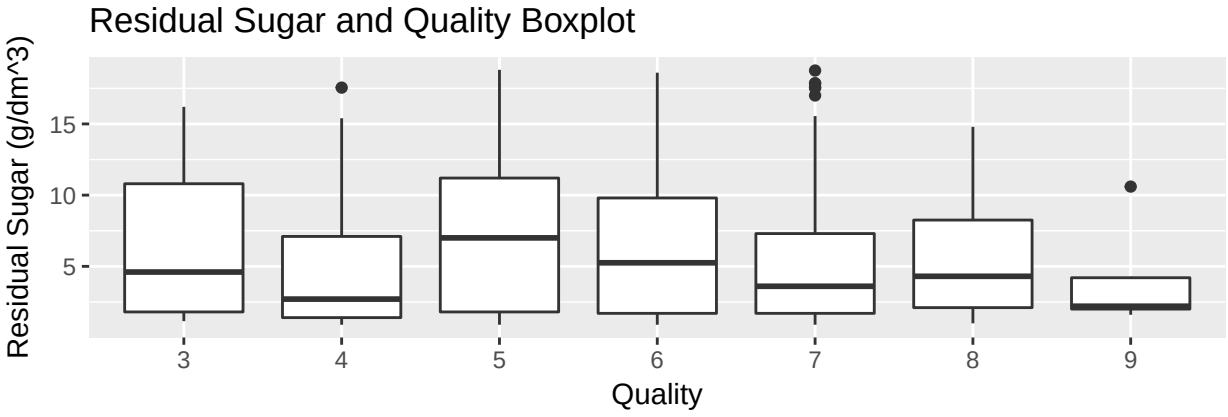




Citric.acid don't have much impact on quality.

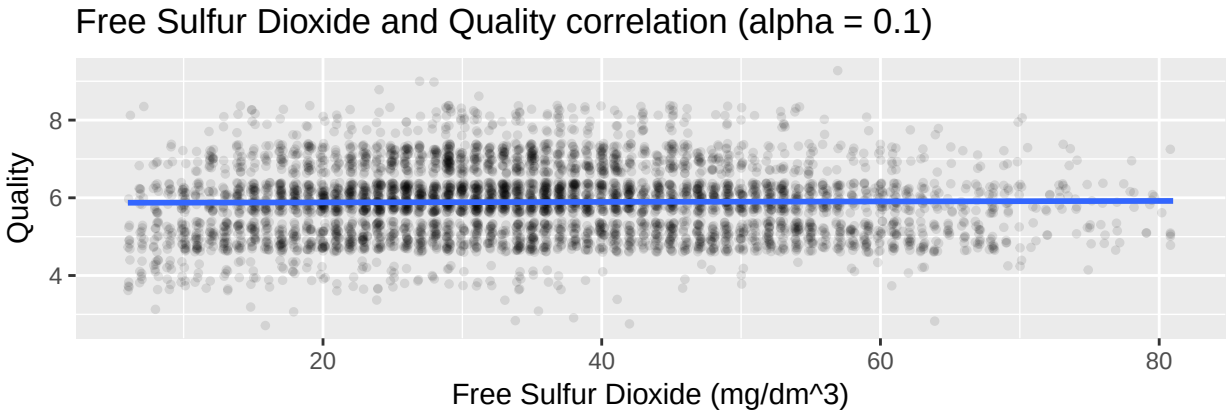
Residual.sugar

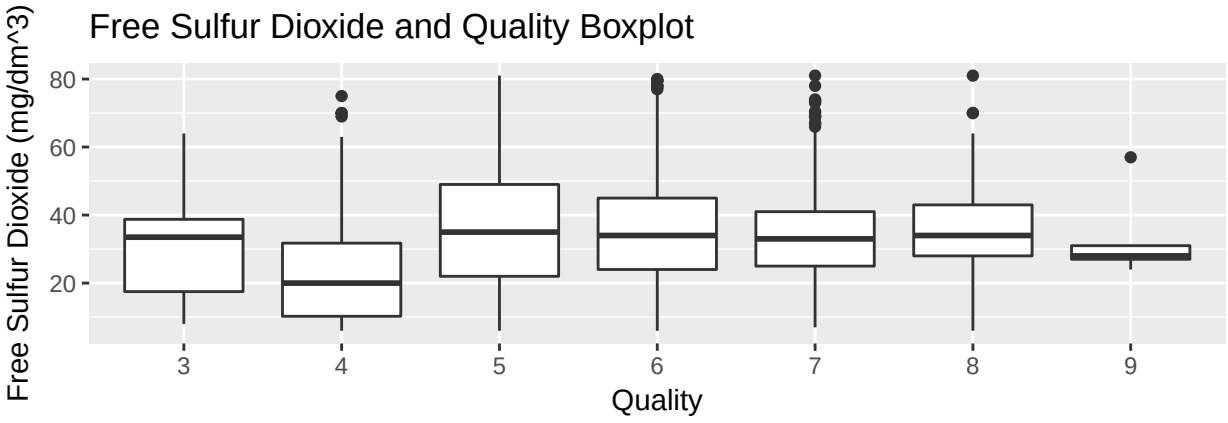




Residual.sugar don't have much impact on quality.

Free.sulfur.dioxide





Free.sulfur.dioxide don't have much impact on quality.

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?

- When quality increases, the alcohol slightly increases as well.
- When quality increases, the density value will decrease accordingly.
- Since free.sulfur.dioxide is included in total.sulfur.dioxide, they are highly correlated to each other.

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

- alcohol is highly correlated to density and moderately correlated to residual.sugar, chlorides, total.sulfur.dioxide, and quality.
- fixed.acidity is moderately correlated to pH.
- Residual sugar is highly correlated to density and moderately correlated to total.sulfur.dioxide and alcohol.

What was the strongest relationship you found?

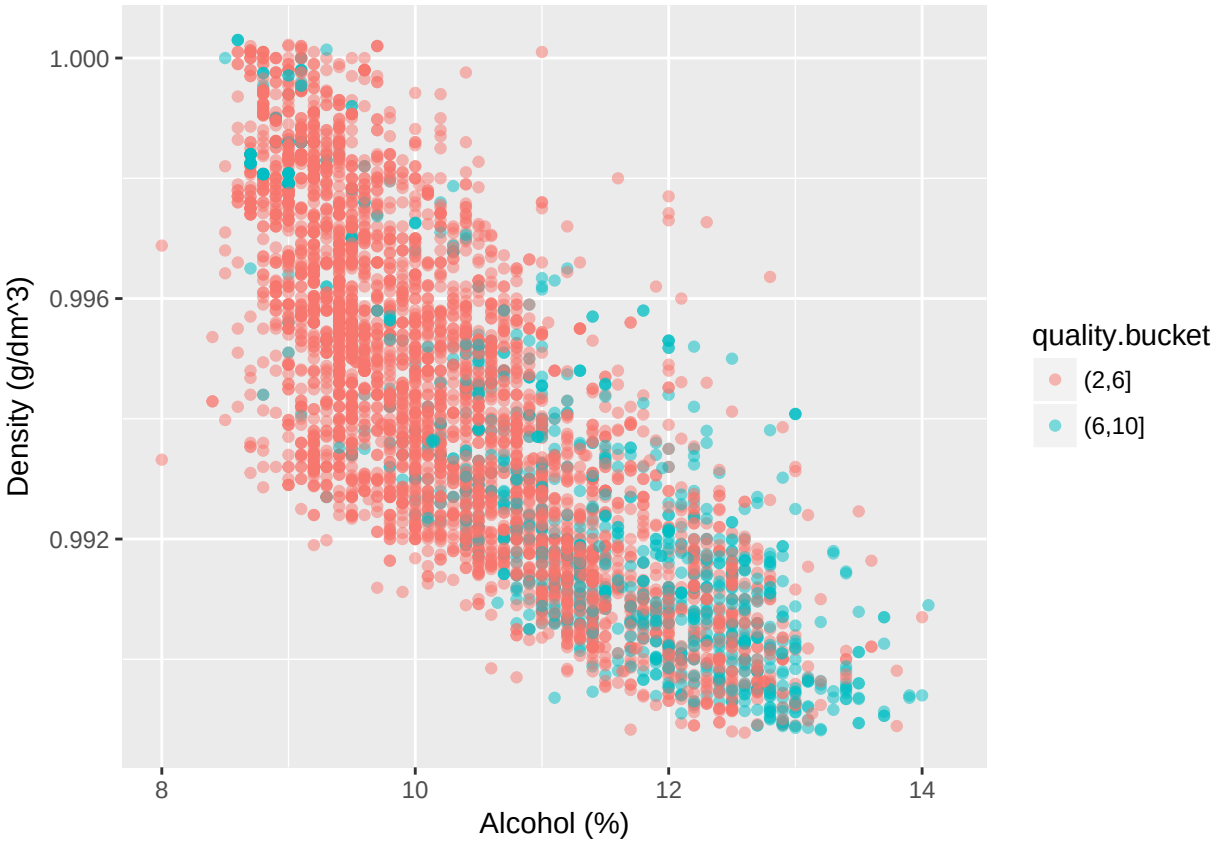
Density and residual.sugar have the strongest correlation of 0.839.

Multivariate Plots Section

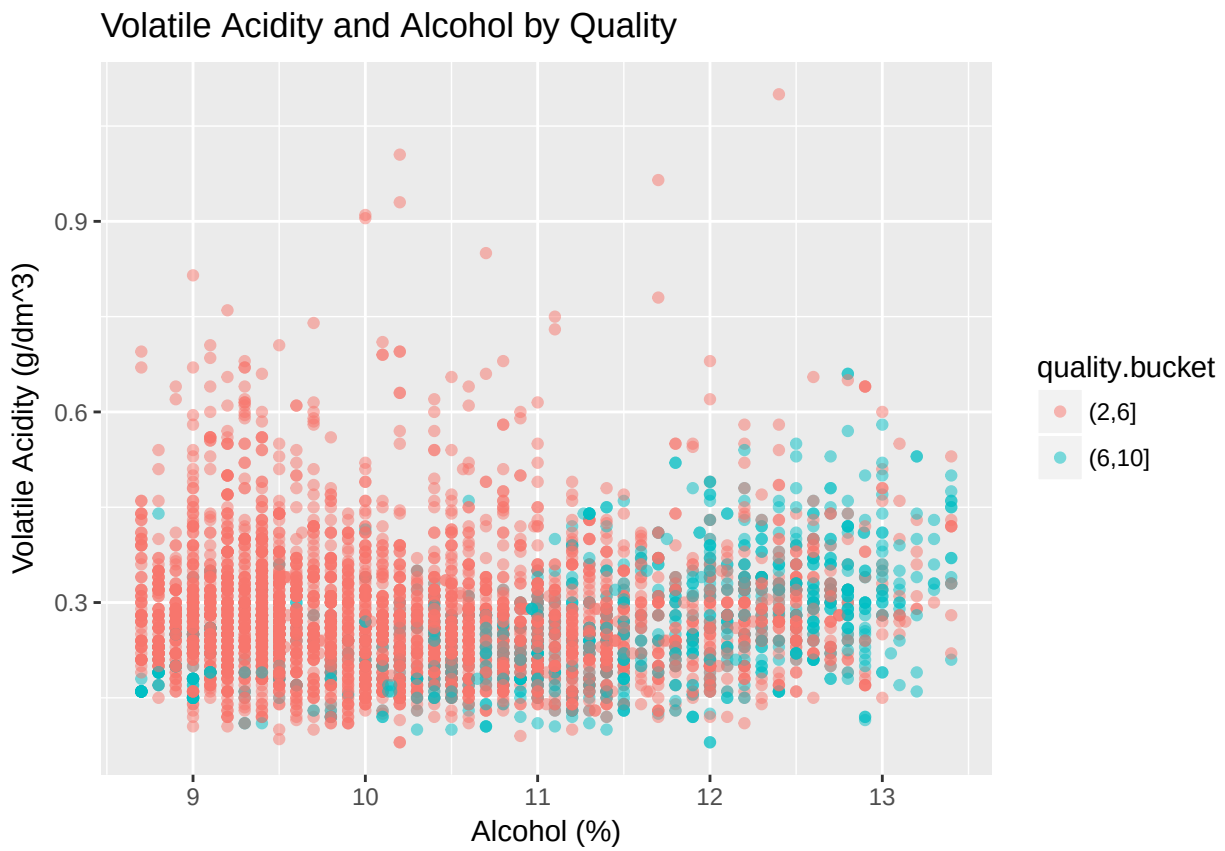
Alcohol

Since alcohol along can not predict the quality very well, I'd like to add other variables to interact with alcohol to find out more insights. Before any further action, I created two buckets for different quality groups, one group with lower quality (2 - 6) and another one with higher quality (7 - 10).

First, I'll include density to see how the plot changes.

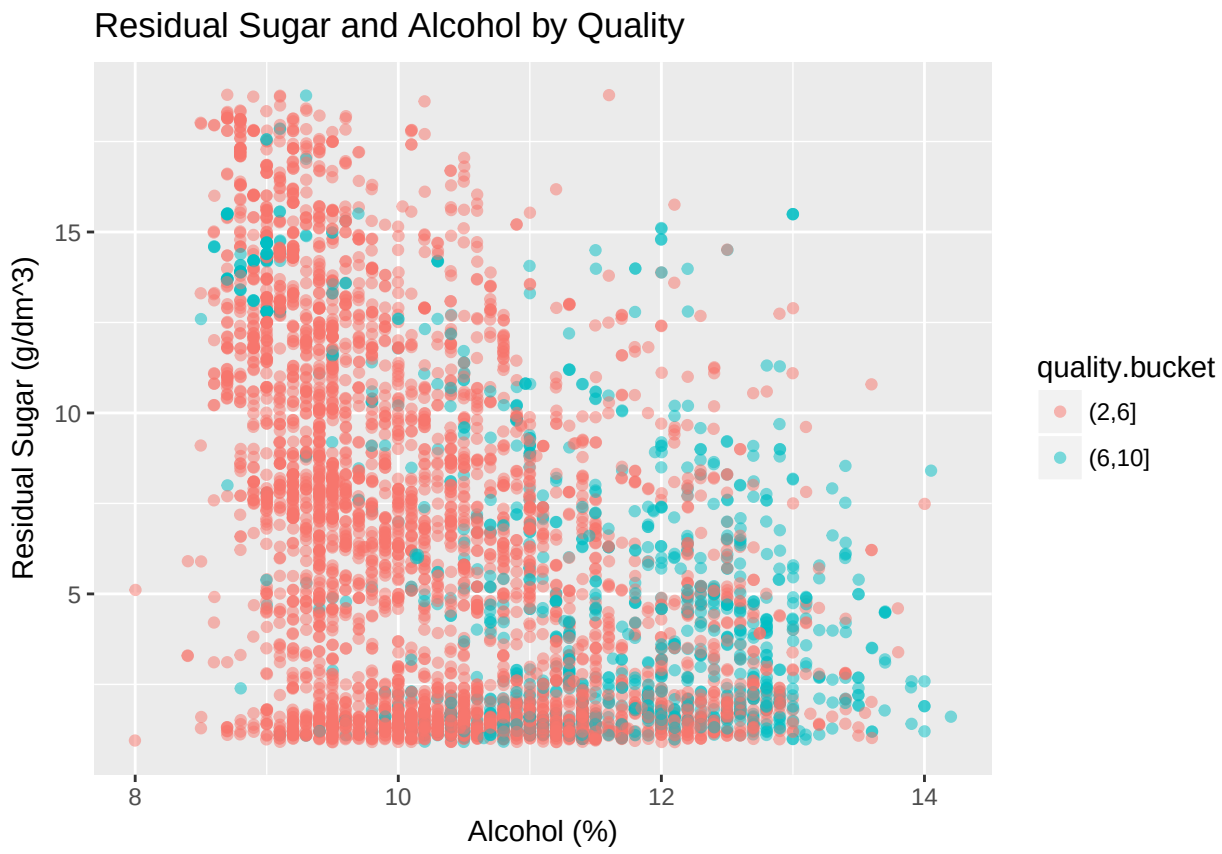


From the scatter plot, alcohol and density exist an negaitve relationship, which leads to high quality wine tends to have higher alcohol and lower density value.

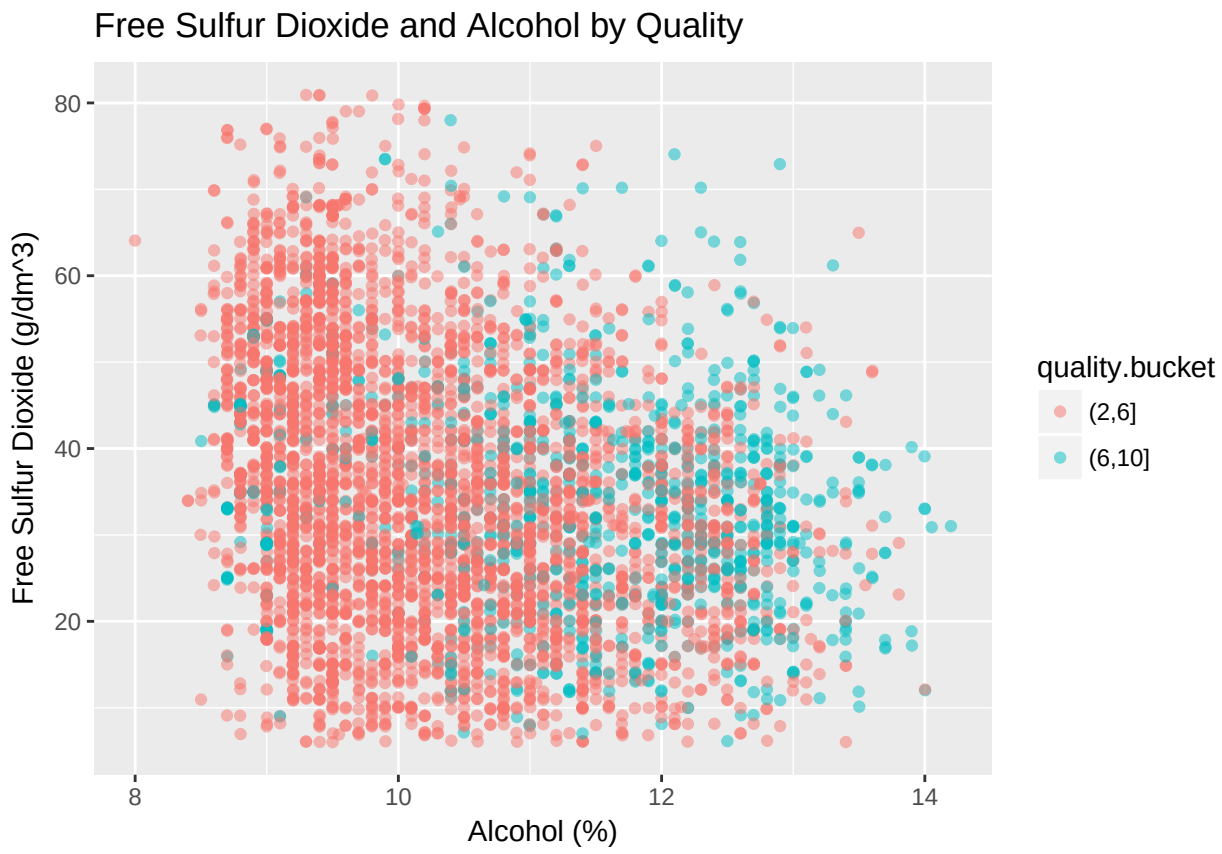


Then, I tried to put the volatile.acidity, which will cause unpleasant scent, in the chart to find out pattern if possible. From the scatter plot, it shows that higher quality wines usually exist in higher alcohol group. However, there is no specific trend for volatile.acidity in different quality groups.

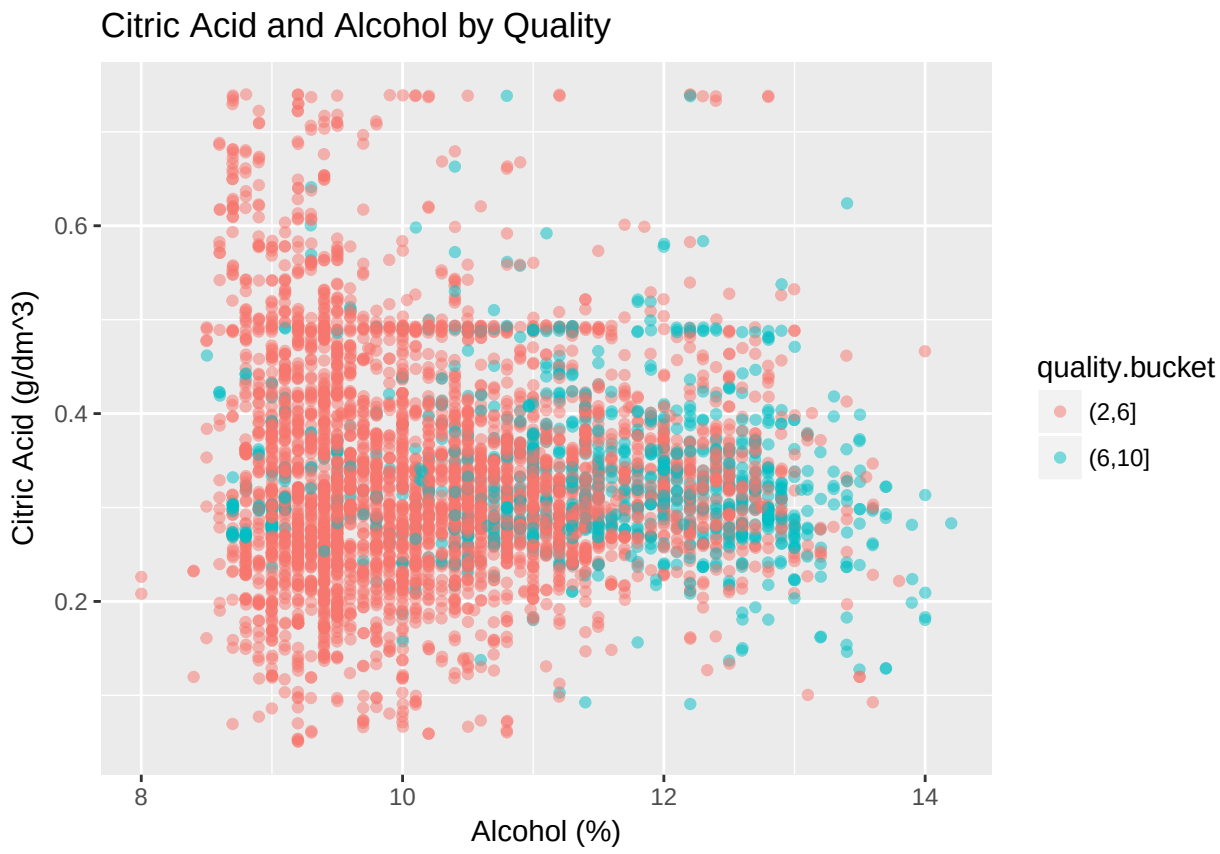
It didn't show much evidence that higher quality wines will have lower volatile.acidity as I expected.



It seems that sugar variance is smaller among high quality wine, sugar usually controlled under 10 g/dmˆ3.



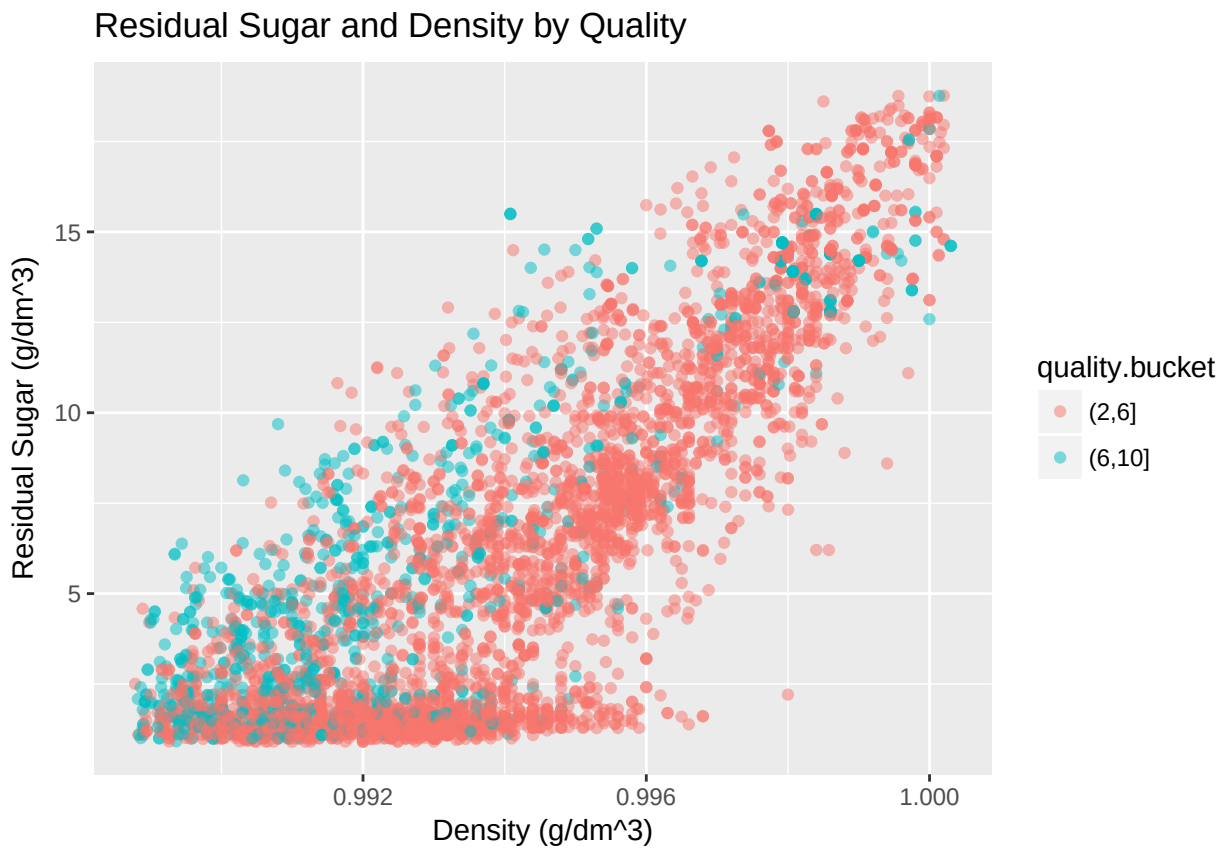
It didn't show much evidence that Free sulfur Dioxide has impact on wine quality.



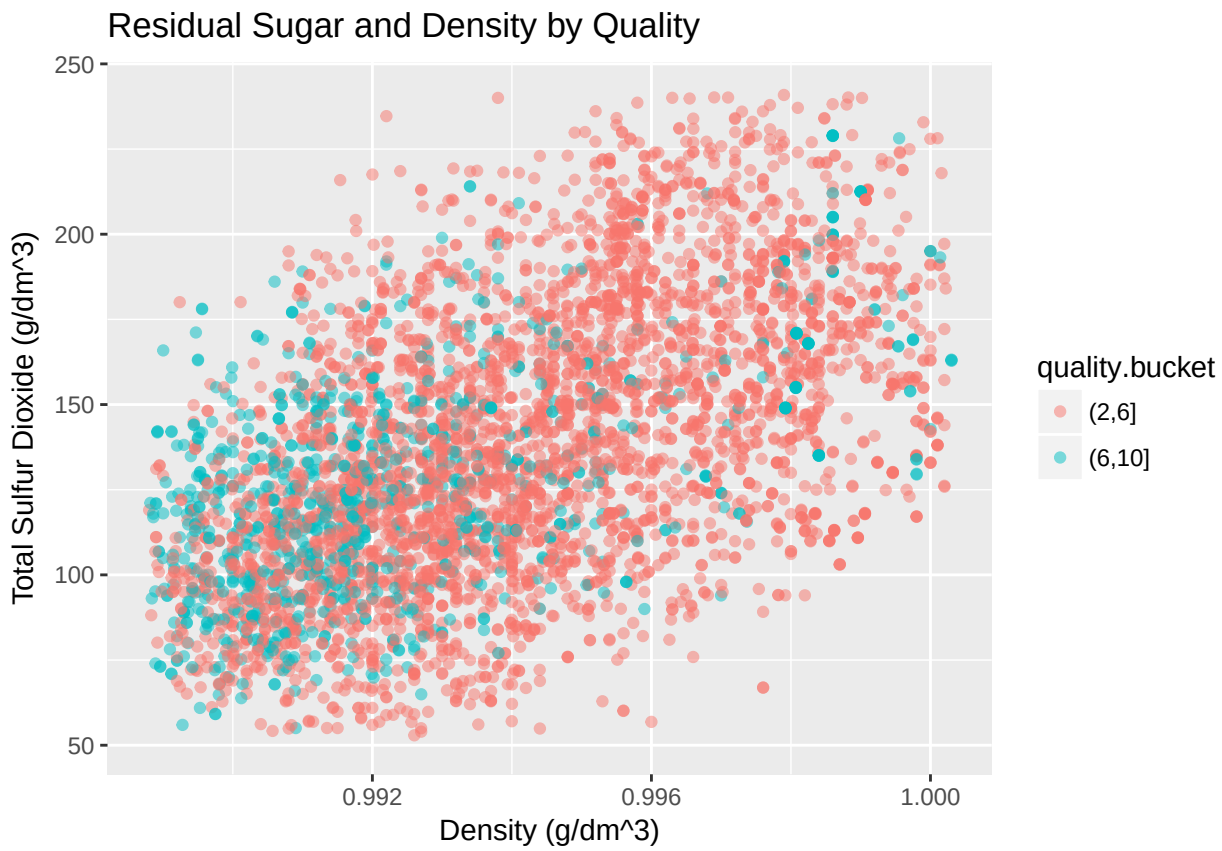
It didn't show much evidence that Citric has impact on wine quality.

I also want to see how density interact with other variables for the following analysis.

Density



First, I include the residual.sugar variable and it looks that when density is the same, high quality wine has more sugar compared to low quality wine.



It didn't show much evidence that Total Sulfur Dioxide has impact on wine quality.

Linear regression models comparison

```
m1 <- lm(quality ~ alcohol, data = wine)
m2 <- update(m1, ~ . + density)
m3 <- update(m2, ~ . + residual.sugar)
m4 <- update(m3, ~ . + citric.acid)
m5 <- update(m4, ~ . + free.sulfur.dioxide)
m6 <- update(m5, ~ . + volatile.acidity)

mtable(m1, m2, m3, m4, m5, m6, sdigits = 3)
```

```
##
## Calls:
## m1: lm(formula = quality ~ alcohol, data = wine)
## m2: lm(formula = quality ~ alcohol + density, data = wine)
## m3: lm(formula = quality ~ alcohol + density + residual.sugar, data = wine)
## m4: lm(formula = quality ~ alcohol + density + residual.sugar + citric.acid,
##     data = wine)
## m5: lm(formula = quality ~ alcohol + density + residual.sugar + citric.acid +
##     free.sulfur.dioxide, data = wine)
## m6: lm(formula = quality ~ alcohol + density + residual.sugar + citric.acid +
##     free.sulfur.dioxide + volatile.acidity, data = wine)
##
```

```
## ================================================================================================
##                            m1           m2           m3           m4           m5           m6
## ------------------------------------------------------------------------------------------------
##   (Intercept)           2.582***   -22.492***    90.313***    95.983***    89.836***    68.95
##                        (0.098)      (6.165)     (12.374)     (12.558)     (12.519)     (12.1
##   alcohol               0.313***     0.360***     0.246***     0.240***     0.257***     0.3
##                        (0.009)      (0.015)      (0.018)      (0.018)      (0.018)      (0.0
##   density                           24.728***   -87.886***   -93.619***   -87.744***   -66.4
##                                     (6.079)     (12.317)     (12.506)     (12.465)     (12.1
##   residual.sugar                                  0.053***     0.055***     0.049***     0.0
##                                                  (0.005)      (0.005)      (0.005)      (0.0
##   citric.acid                                                   0.246**      0.194*      -0.0
##                                                                (0.095)      (0.095)      (0.0
##   free.sulfur.dioxide                                                       0.005***     0.0
##                                                                             (0.001)      (0.0
##   volatile.acidity                                                                       -2.0
##                                                                                          (0.1
## ------------------------------------------------------------------------------------------------
##   R-squared               0.190        0.192        0.210        0.211        0.220        0.2
##   adj. R-squared          0.190        0.192        0.210        0.211        0.219        0.2
##   sigma                   0.797        0.796        0.787        0.787        0.783        0.7
##   F                    1146.395      583.290      434.085      327.629      275.702      300.0
##   p                       0.000        0.000        0.000        0.000        0.000        0.0
##   Log-likelihood      -5839.391    -5831.127    -5776.812    -5773.443    -5746.636    -5587.0
##   Deviance             3112.257     3101.773     3033.737     3029.566     2996.585     2807.5
##   AIC                 11684.782    11670.255    11563.624    11558.886    11507.272    11190.0
##   BIC                 11704.272    11696.241    11596.107    11597.866    11552.748    11242.0
##   N                    4898         4898         4898         4898         4898         4898
## ================================================================================================
```

Although the R^2 increases as the variables increase in the model, all the variables together still can not explain very well about the quality.

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?

The higher quality wines usually have higher alcohol but volatile.acidity and free.sulfur.dioxide don't really impact on quality.

Were there any interesting or surprising interactions between features?

I thought that volatile.acidity would exist less in high quality of wines but it turns out that the volatile.acidity disperse similarly across differnt quality levels.

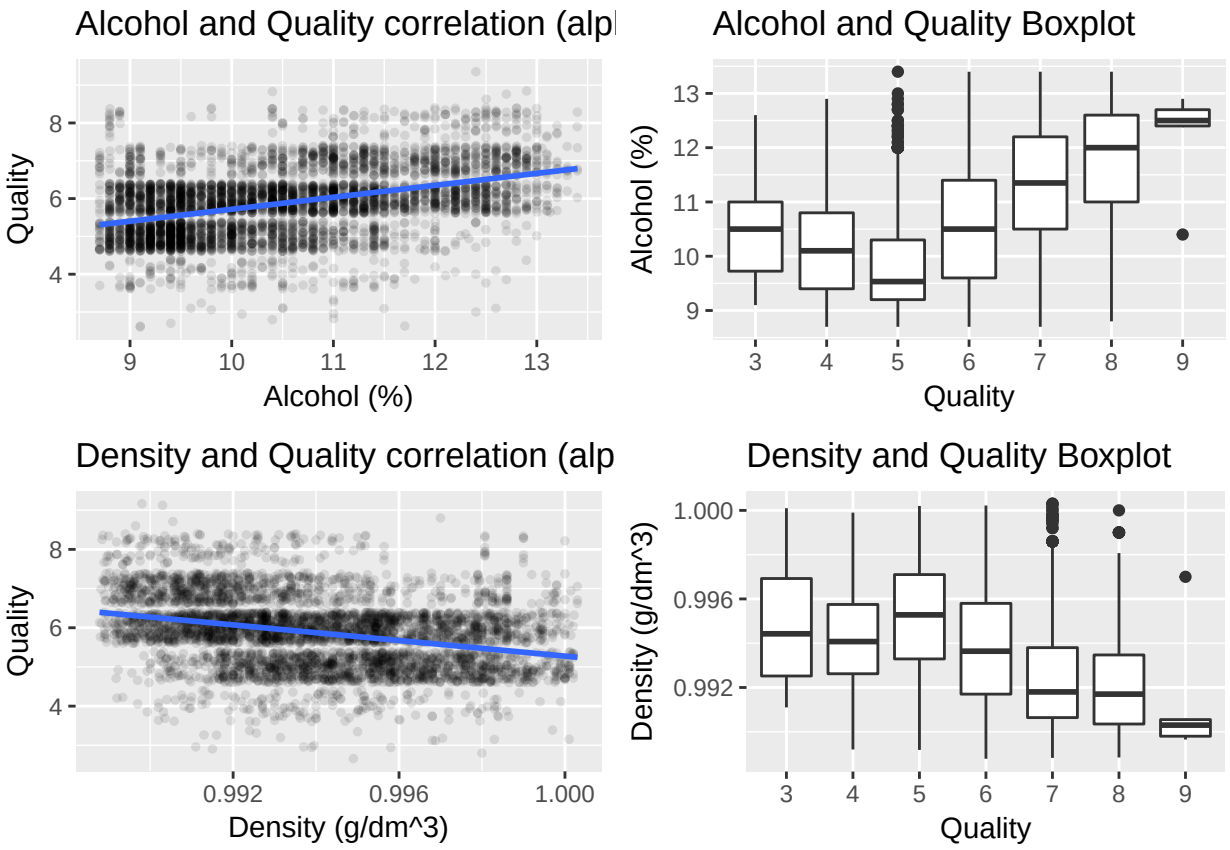
OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model.

Yes, I created models with different variables. The dataset is alreday in a tidy format which is convenient to build regression models. However, it might be lack of information about the place of production or the year

of production, which might affce the quality.

Final Plots and Summary

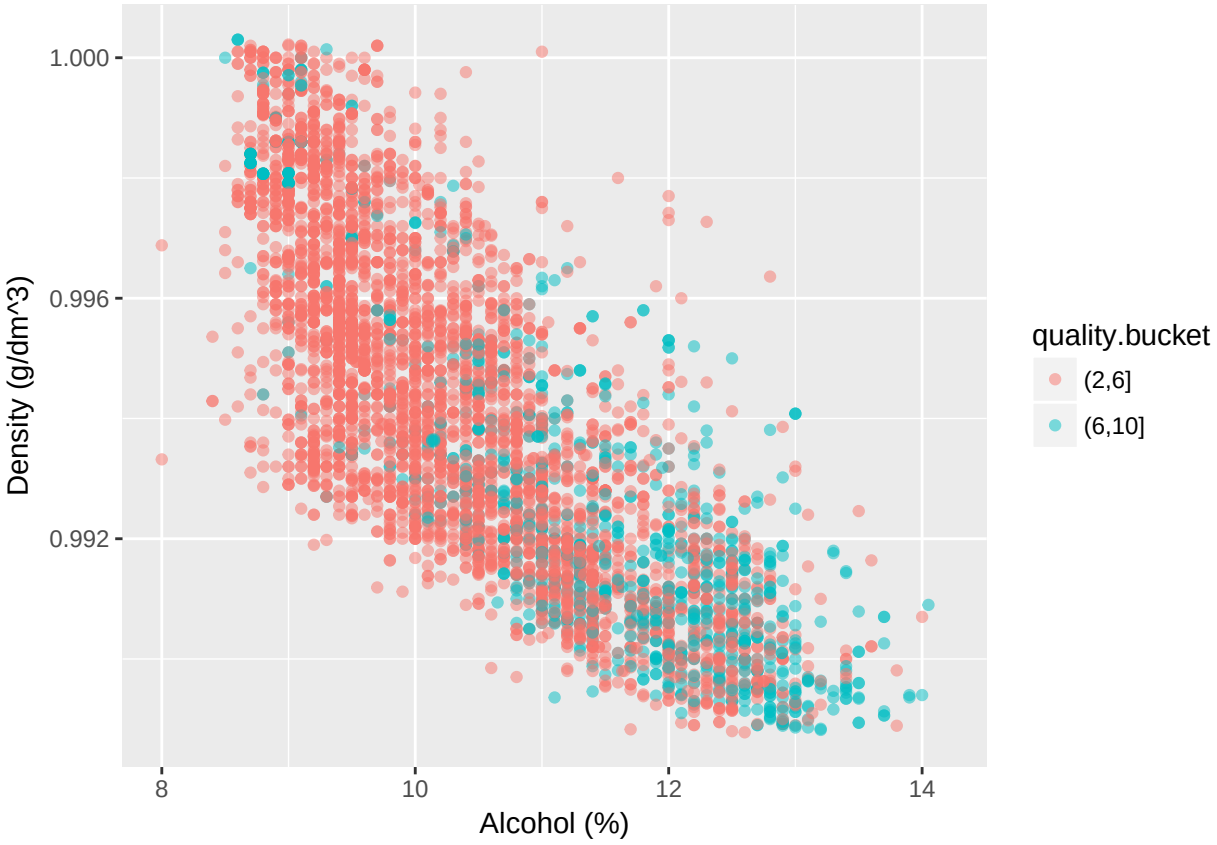
Plot One



Description One

My interest variable wine Quality has a relationship with alcohol(correlation of 0.436) and density(correlation of 0.3). From the Alcohol and Quality scatter plot, it looks that there is an upward trend, which shows that higher quality wine usually have higher alcohol. To take a closer look in related boxplot, it really shows that the median alcohol value increases from 9.5% at quality 5 to 12.5% at quality 9. I also noticed that high quality(7, 8, 9) wine mostly exist with alcohol greater than 11.5%. For another variable density, there is a downward trend in Density and Quality scatter plot, which indicates that lower quality wine usually have higher density. The related boxplot also shows that the median density value decreases from around 0.995 g/dmˆ3 at quality 5 to 0.99 g/dmˆ3 at quality 9

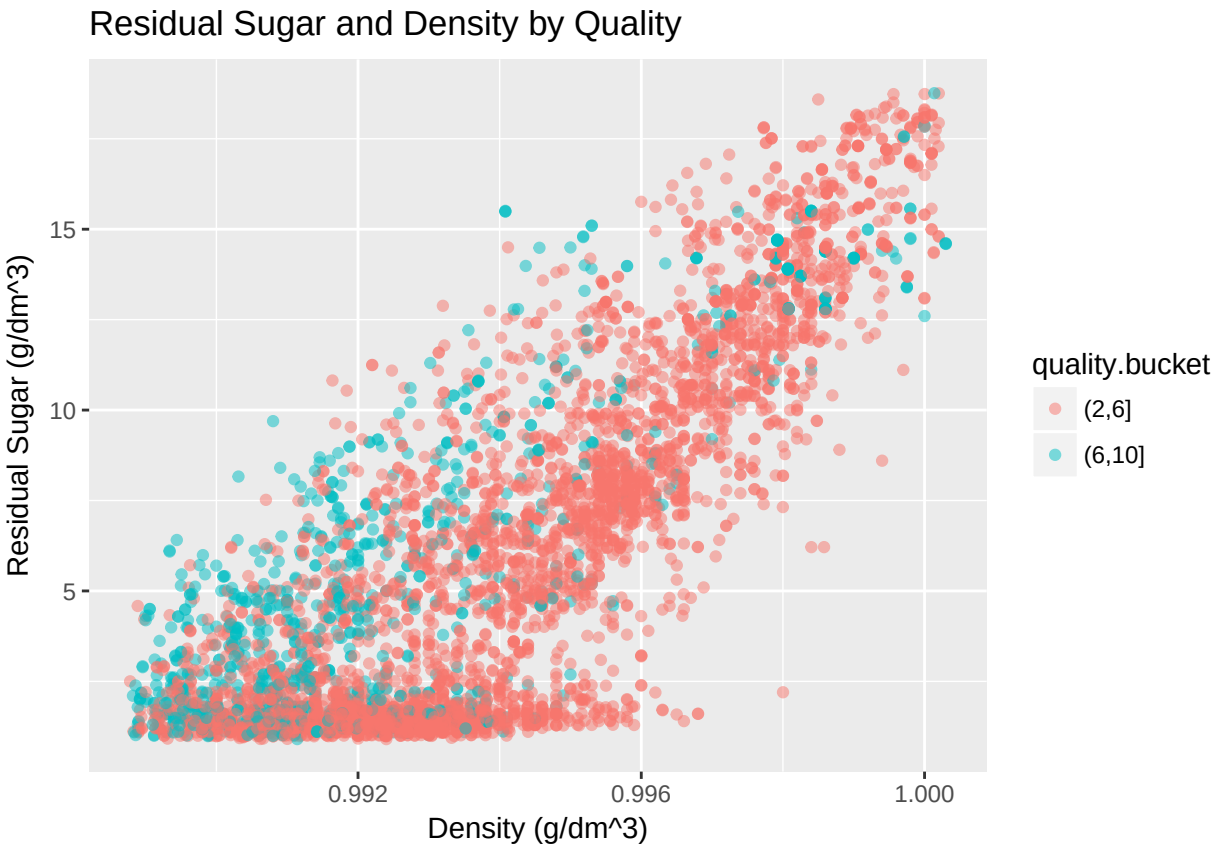
Plot Two



Description Two

Since Alcohol and Density are the two most related varialbels to quality, I put the variables together and find an interesting relationship. Alcohol and Density exist an negaitve relationship which leads to higher alcohol wine has lower density. The high quality wine usually exist with high alcohol and also tends to have lower density value.

Plot Three



Description Three

We know from previous plot that high quality wine tends to have lower density. From the plot three, we also notice that low density wine usually has low residual sugar. The high quality wine mostly show up under density of 0.995 g/dmˆ3 and under sugar of 10 g/dmˆ3.

Reflection

I feel hard to focus on specific variables only by univariate analysis, whcih might lead to wrong directions for the further analysis. The correlation plots help me to identify most relevant variables but meanwhile, I'm also not sure whether to ignore those irrelevant ones. As I tried different combinations of variables to generate bivariate or mulitivariate plots, I feel usefult to leverage the correlation information to identify trends for different quality groups, which I think I can apply to future works if I don't know where to start.