Wine Dataset Exploration by Hajeong Noh

This project is my first Exploratory Data Analysis using R. The dataset includes several features of wines and its quality. The format includes Univariate, Bivariate and Multivariate analysis with a final summary and reflection. The original dataset is available here

Univariate Plots Section

Let's take a look at the dimensions, structure and summary of wine dataset.

```
## [1] 6497   13
```

```
## 'data.frame':    6497 obs. of  13 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 ...
##  $ type                : chr  "red" "red" "red" "red" ...
```

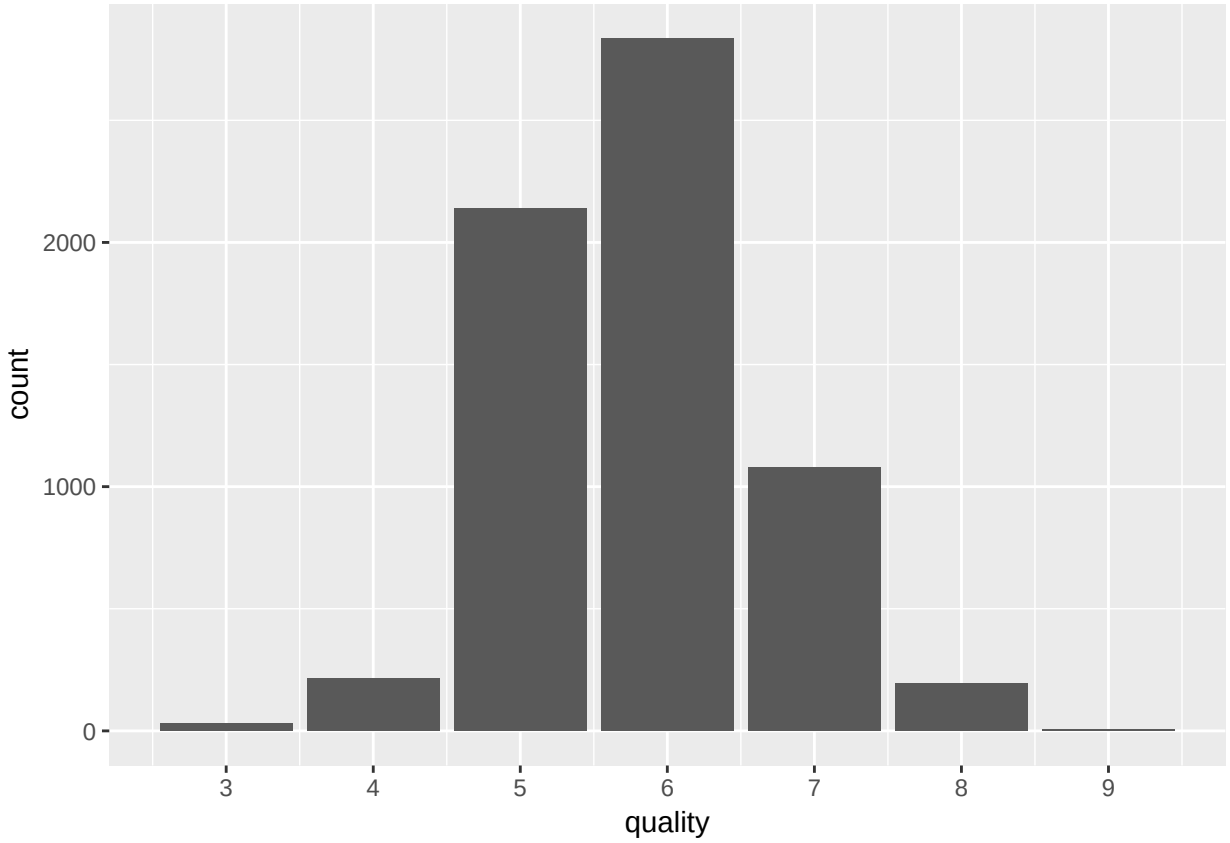
```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2500   1st Qu.: 1.800
##  Median : 7.000   Median :0.2900   Median :0.3100   Median : 3.000
##  Mean   : 7.215   Mean   :0.3397   Mean   :0.3186   Mean   : 5.443
##  3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900   3rd Qu.: 8.100
##  Max.   :15.900   Max.   :1.5800   Max.   :1.6600   Max.   :65.800
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.00900   Min.   :  1.00      Min.   :  6.0
##  1st Qu.:0.03800   1st Qu.: 17.00      1st Qu.: 77.0
##  Median :0.04700   Median : 29.00      Median :118.0
##  Mean   :0.05603   Mean   : 30.53      Mean   :115.7
##  3rd Qu.:0.06500   3rd Qu.: 41.00      3rd Qu.:156.0
##  Max.   :0.61100   Max.   :289.00      Max.   :440.0
##     density             pH          sulphates         alcohol
##  Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00
##  1st Qu.:0.9923   1st Qu.:3.110   1st Qu.:0.4300   1st Qu.: 9.50
##  Median :0.9949   Median :3.210   Median :0.5100   Median :10.30
##  Mean   :0.9947   Mean   :3.219   Mean   :0.5313   Mean   :10.49
##  3rd Qu.:0.9970   3rd Qu.:3.320   3rd Qu.:0.6000   3rd Qu.:11.30
##  Max.   :1.0390   Max.   :4.010   Max.   :2.0000   Max.   :14.90
##     quality          type
##  Min.   :3.000   Length:6497
##  1st Qu.:5.000   Class :character
##  Median :6.000   Mode  :character
##  Mean   :5.818
##  3rd Qu.:6.000
```

```
##  Max.   :9.000
```

The key stat quality has median value 6. minimum is 3 and maximum is 9. Even though quality is ranging from 0 to 10, only 3 through 9 are used.

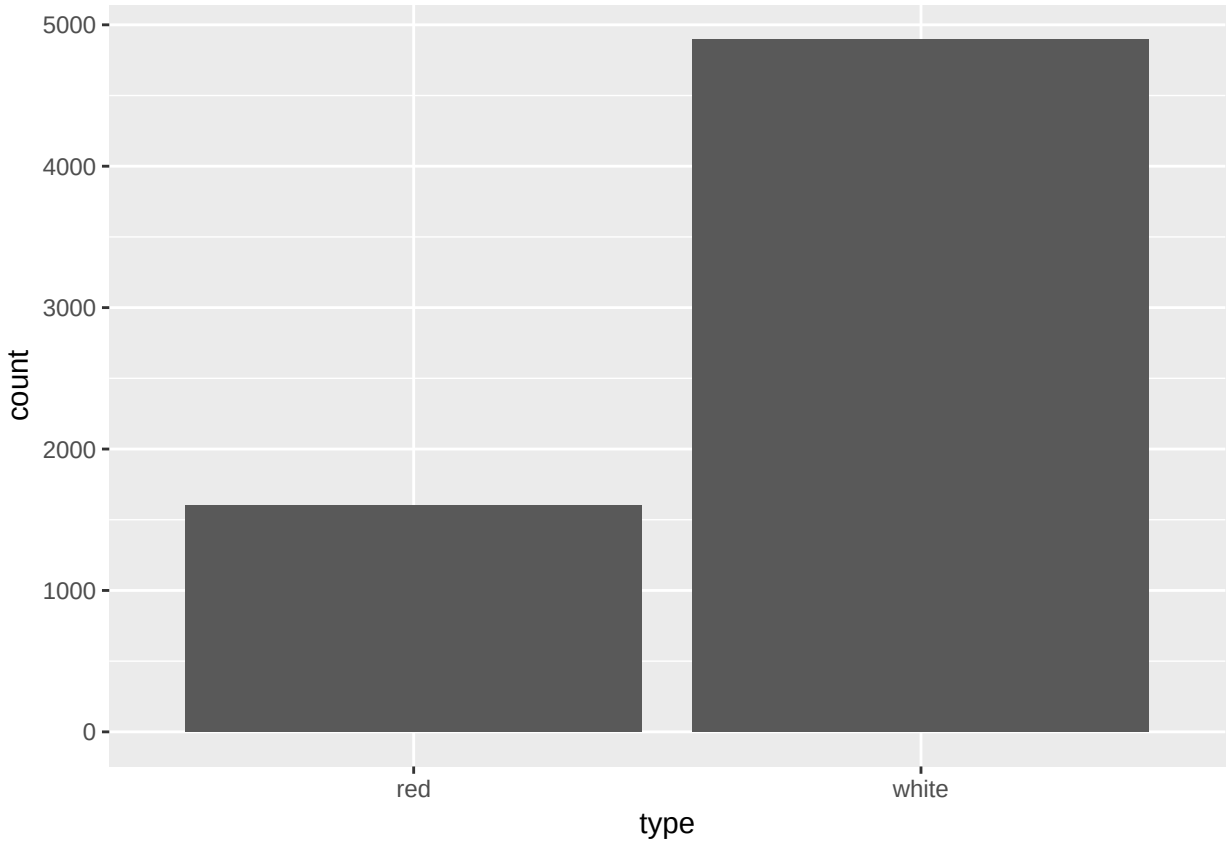
Most density values fall between 0.99 and 1.

There is a wine that contains residual sugar way more than other wines. According to the guideline, this wine is sweet (residual sugar is more than 45).



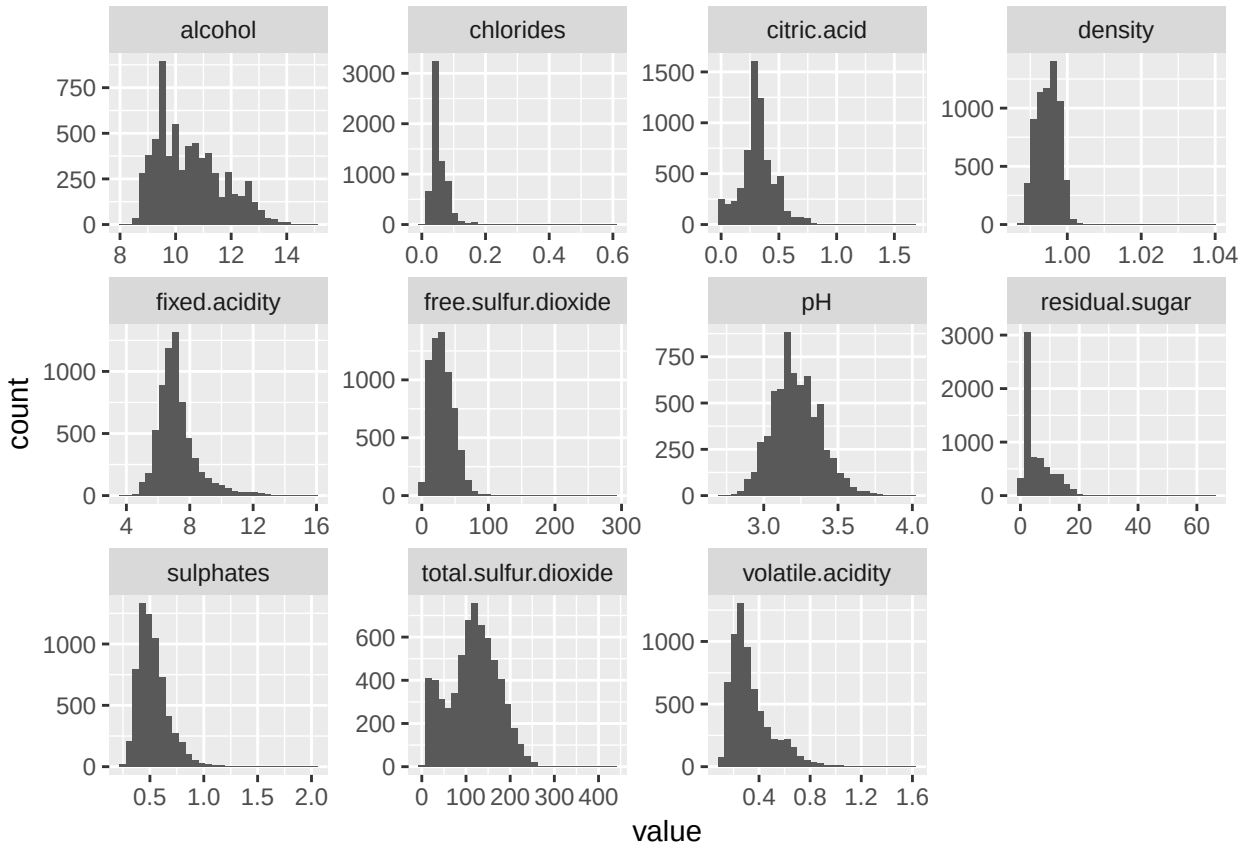

```
##
##    3    4    5    6    7    8    9
##   30  216 2138 2836 1079  193    5
```

The most common wine quality rating is 6 followed by 5 and 7. Few observations are at 3 and 9, which can be considered as very poor and very good, respectively.

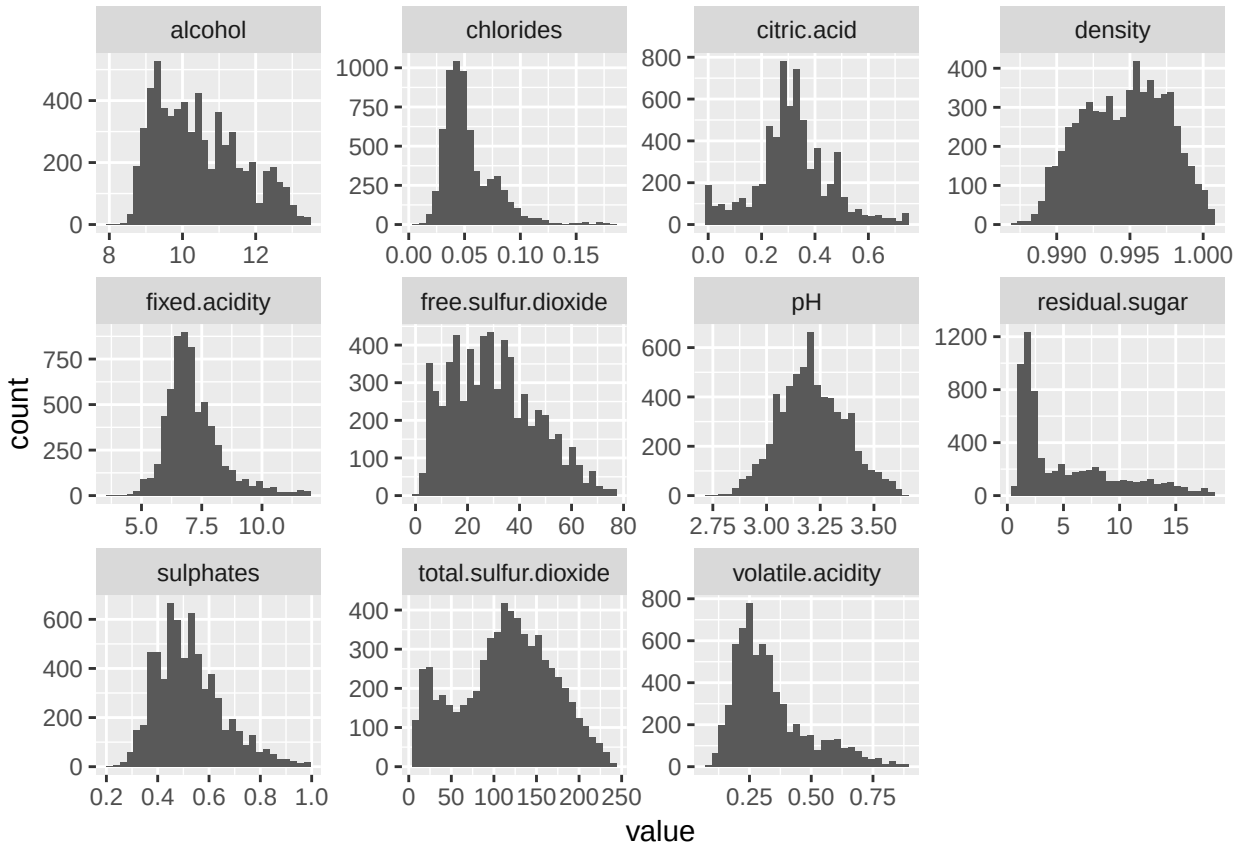



```
##
##   red white
##  1599  4898
```

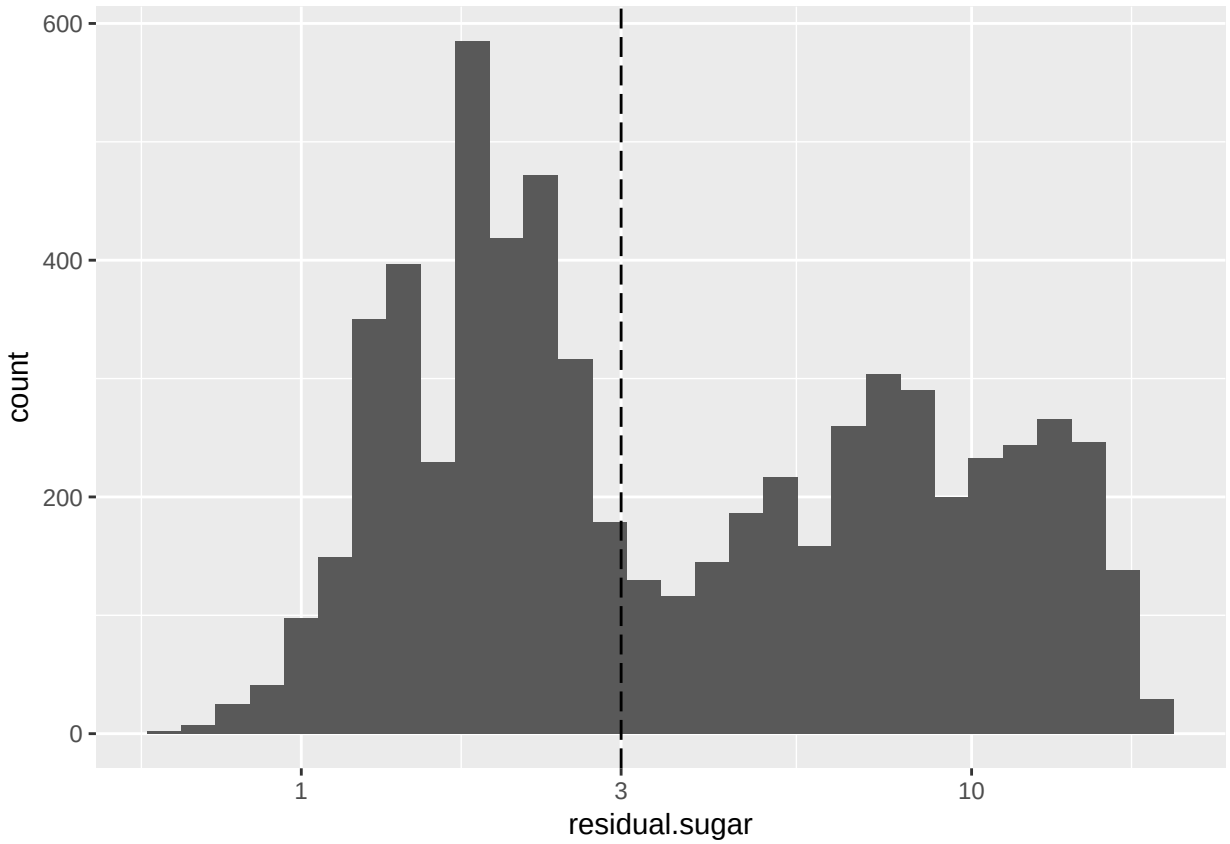
White wine is approximately 3 times more observed in the dataset.



The histograms are relatively located on the left side of each plot, since there are outliers with high values. Let's exclude outliers to observe distributions better.



Most of the variables follow normal distribution except residual sugar. Residual sugar distribution follow long-tail distribution. Let's transform the x-axis to observe the distribution better.



After x-axis transformation, the distribution follows bimodal. I can divide the wines into 2 groups of which has residual sugar less than 3 and which has more than 3 for further analysis.

Univariate Analysis

What is the structure of your dataset?

There are 6,497 observations with 11 quantitative features (alcohol, chlorides, citric acid, density, fixed acidity, free sulfur dioxide, pH, residual sugar, sulphates, total sulfur dioxide and volatile acidity), quality (discrete variable) and type (categorical).

Other observations: - The most observed value of quality is 6 followed by 5 and 7. - The average alcohol percentage is 10.51%. - About 75% of white wines have residual sugar less than 9.9. - The median citric acid is 0.32. - The bimodality of residual histogram shows that we can group the wines with more sugar and less sugar.

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

The main features in the dataset is quality and type. I'd like to explore which features show the most difference as the type differs. I also want to determine which feature contributes the most for predicting the quality of a wine.

What other features in the dataset do you think affect the quality of wine?

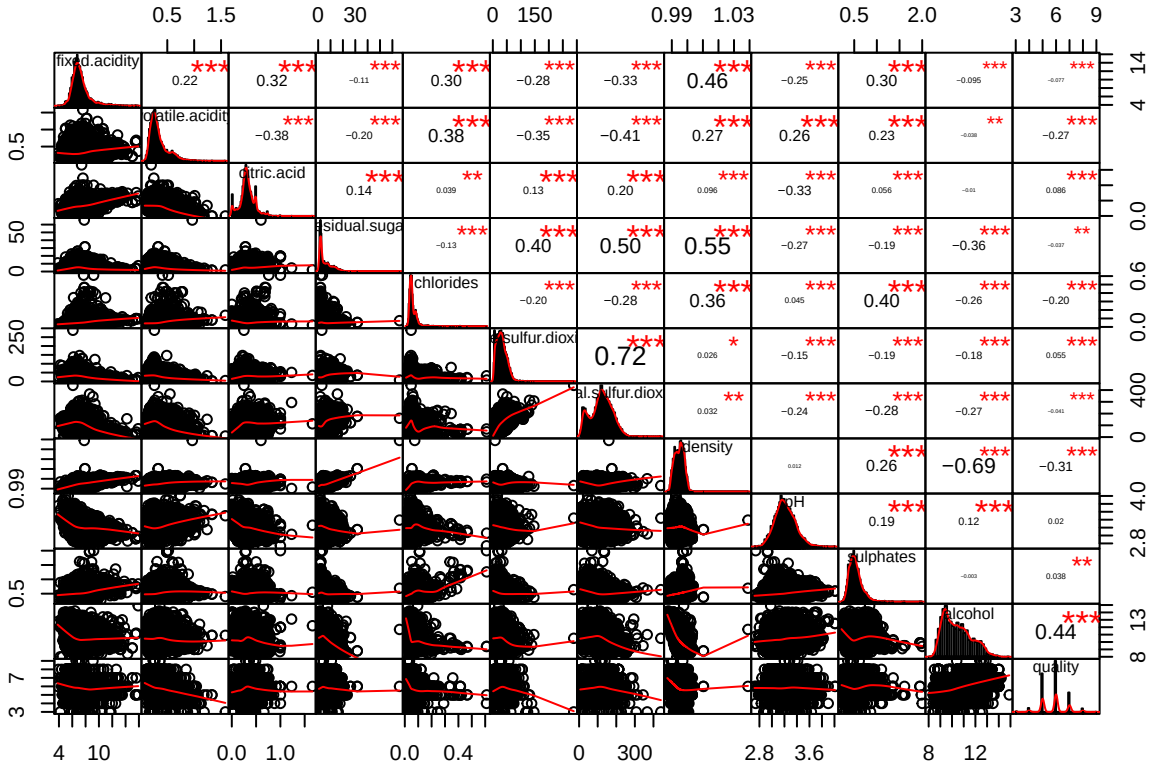
I expect that residual sugar and citric acid contribute to the quality of white wines. I think cholorides may also play a significant role.

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 made an adjustment when I tried to see each histogram of a variable. I found out many histograms are located on the left side of plot, so I removed outliers to check the distribution clearly.

Another adjustment I did was transform the x-axis of residual sugar histogram, since the original histogram show long-tailed distribution. After transformation I could observe that the distribution follows bimodal, which shows that I can divide less and more sugar group for further analysis.

Bivariate Plots Section



The matrix above is plotted using PerformanceAnalytics package. Here are some of the correlations involving quality.

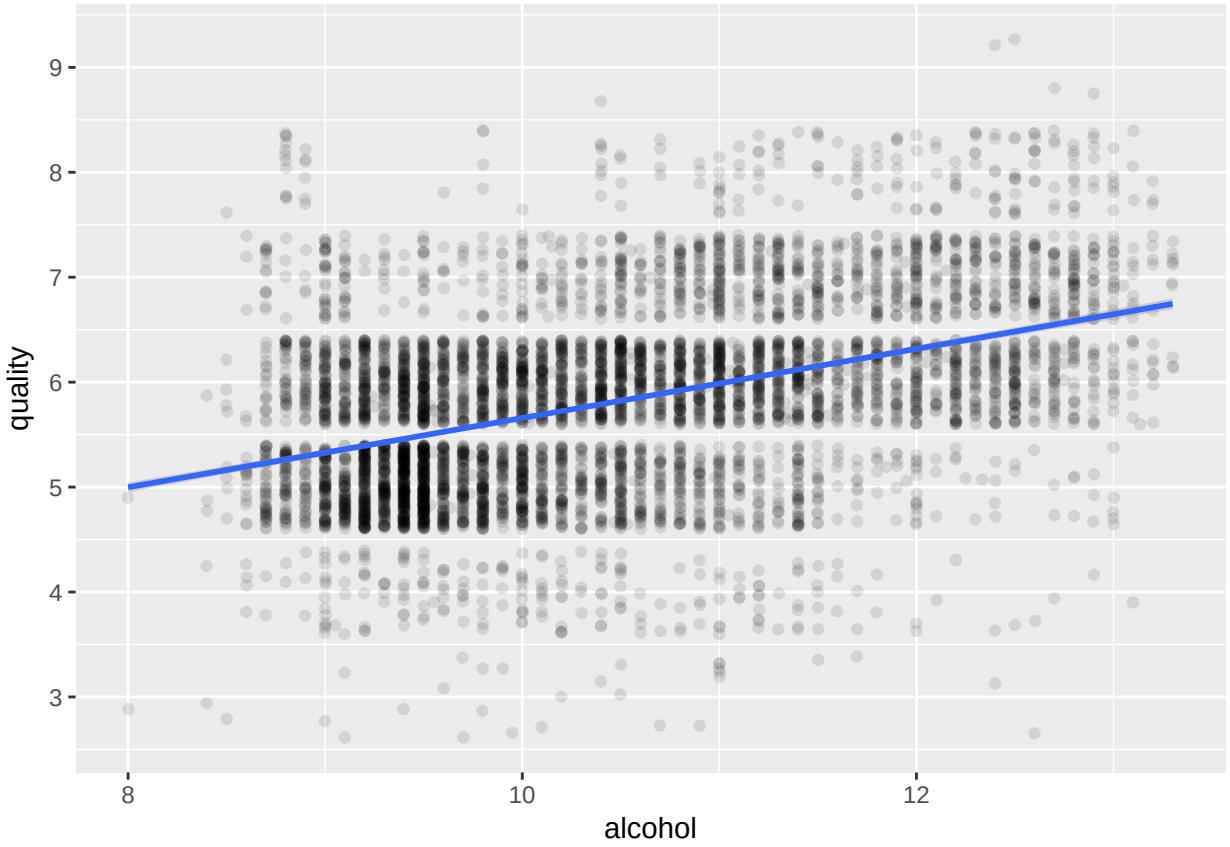
- Quality and Alcohol: 0.44
- Quality and Density: -0.31
- Quality and Volatile acidity: -0.27

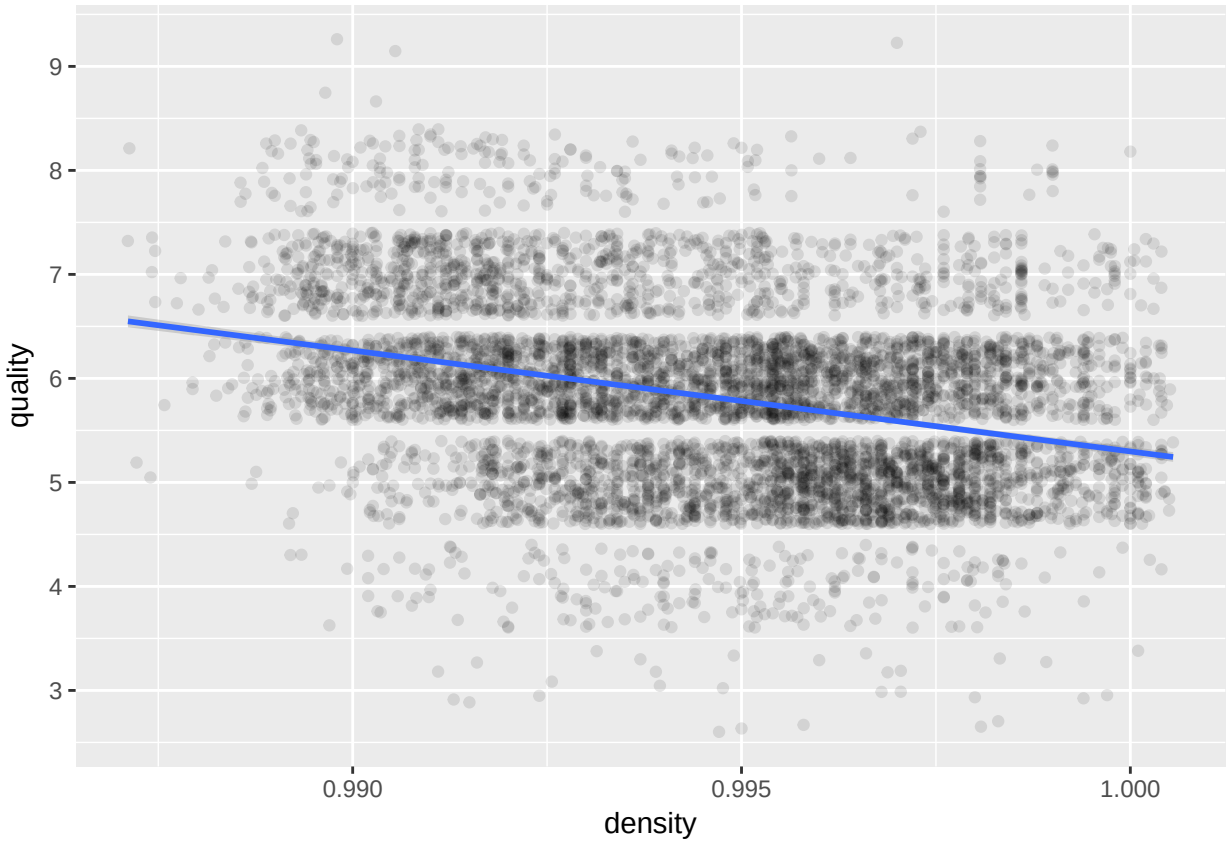
None of those show a strong correlation.

Other interesting correlations are observed between other features.

- Free sulfur dioxide and Total sulfur dioxide: 0.72
- Density and Alcohol: -0.69
- Density and Residual sugar: 0.55

Let's take a look ate the relationship between features and quality in more detail.



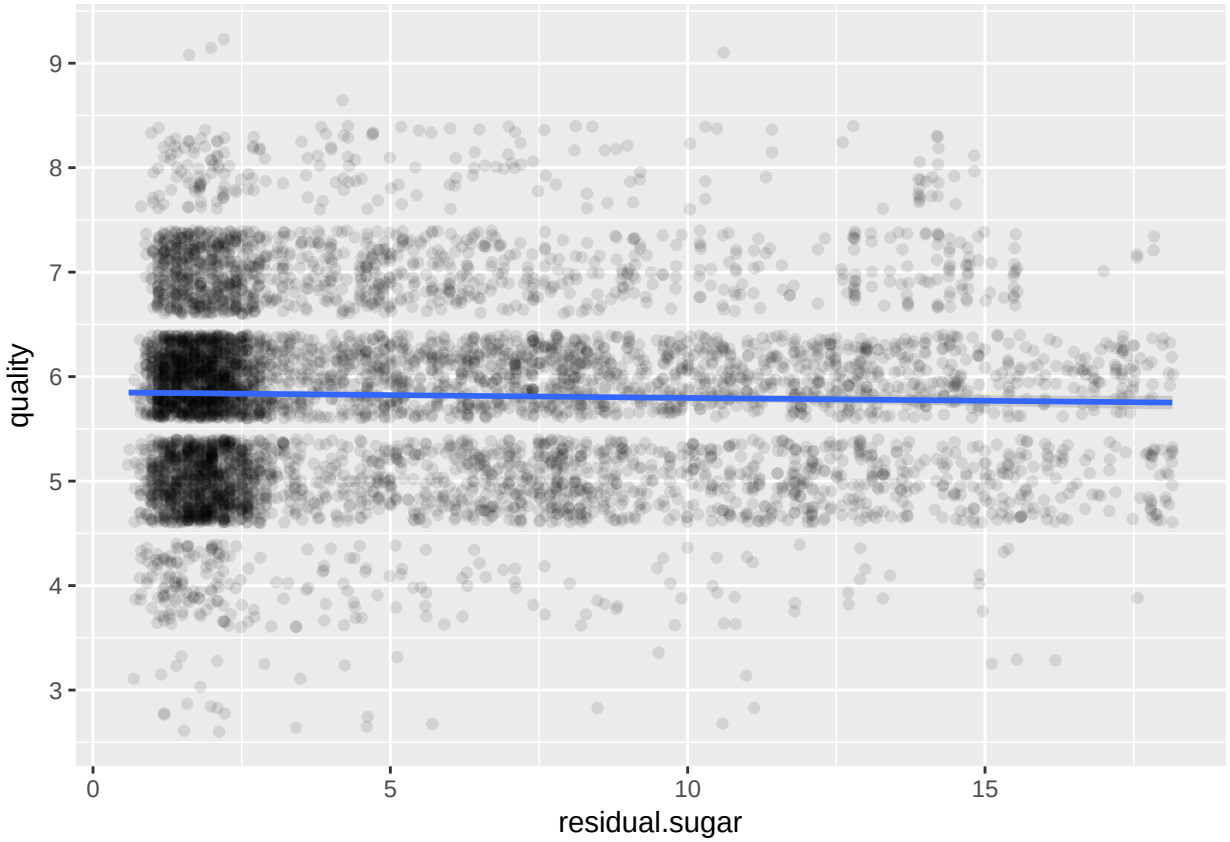


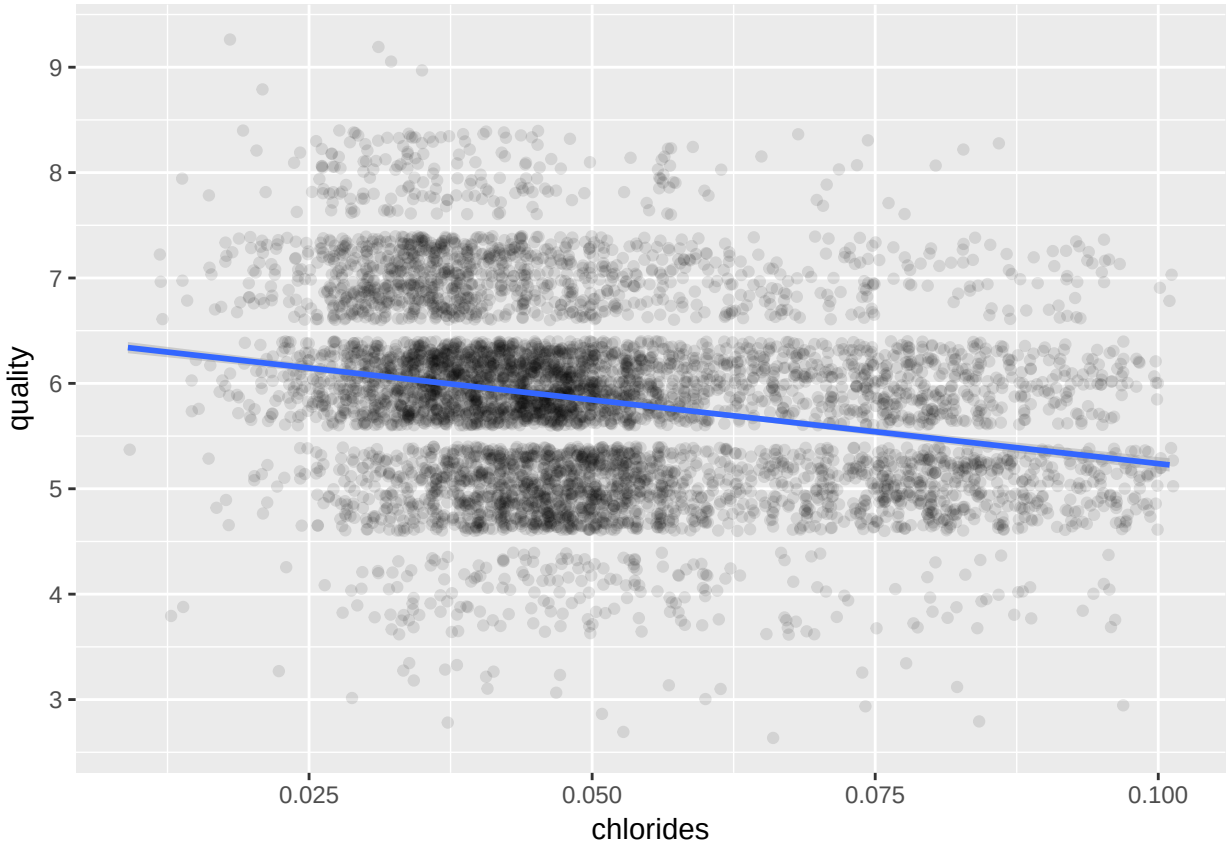
The scatter plots above show the relationship of features which showed the moderate correlation with quality ($0.7 > \text{cor} >= 0.3$) in the correlation matrix. I jittered the plots to visualize trends better and exclude top 1% outliers to check the general trend.

The first scatter plot shows there is a positive relationship between alcohol and quality, even though the correlation is not so strong.

The second scatter plot shows on the other hand, there is a negative relationship between density and quality. Heavy wine can harm the taste.

What about the other variables that I expected to affect the quality? Let's find out!

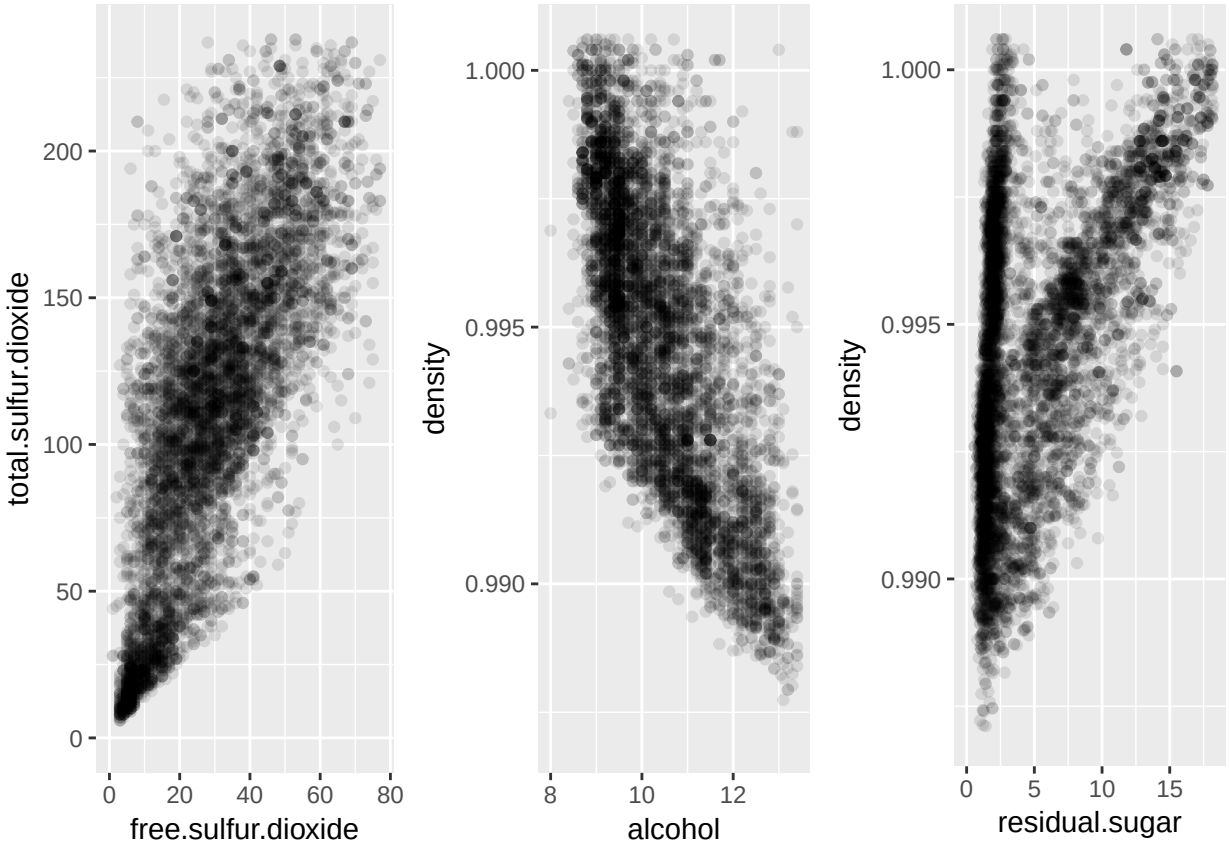

quality
9
8
7
6
5
4
3
0
5
10
15
residual.sugar



The correlation of residual sugar and quality is 0.07, which is very weak. Moreover, the distribution looks very even.

The second plot shows that chlorides and quality has a negative relationship. However, the correlation of chlorides and quality is -0.20, which shows a weak relationship between features.

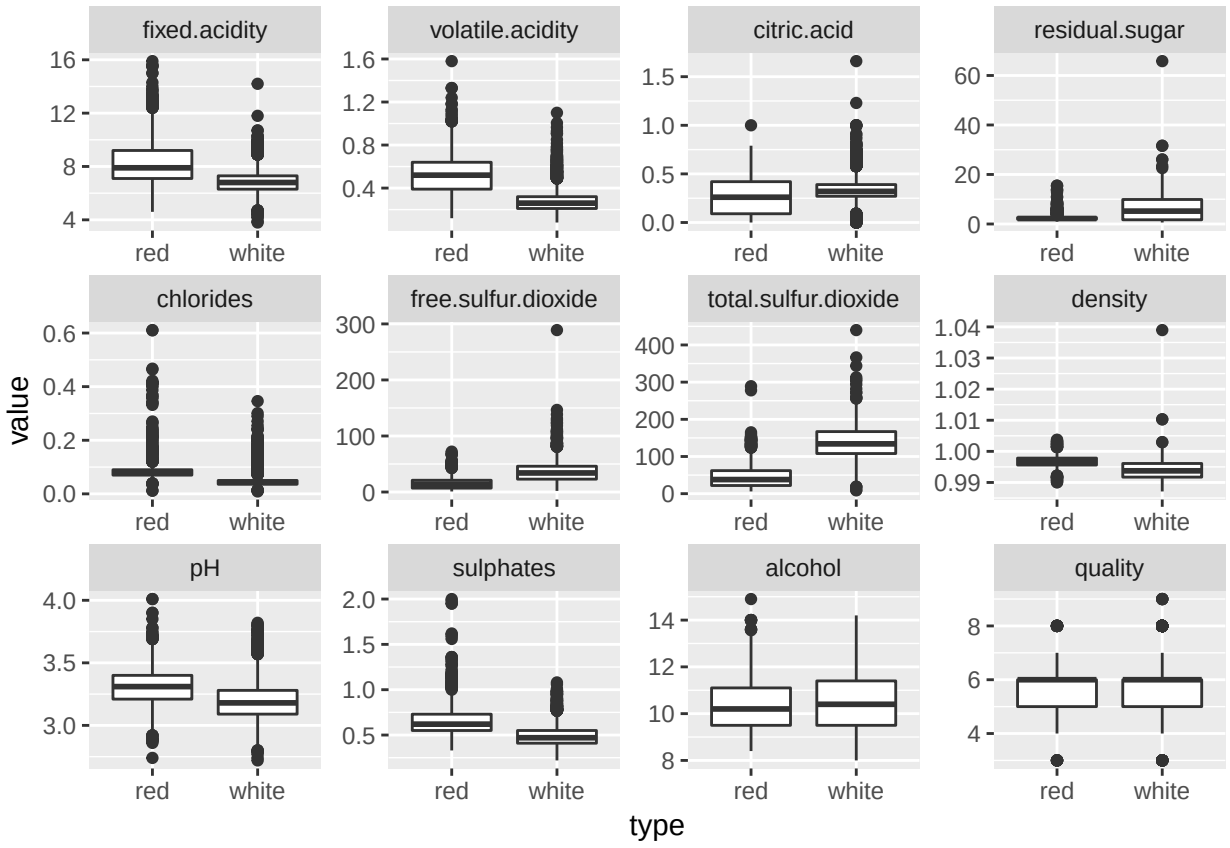
Now let's search for an interesting scatter plot between features except quality. Drawing plots which showed high correlations in the correlation matrix will be interesting.



The plots above show the distribution between features that has a relatively strong correlation. Free sulfur dioxide and total sulfur dioxide evidently has a strong positive correlation 0.72.

Another interesting realationships are density with alcohol and residual sugar. Density tends to go down as alcohol percentage gets higher and as residual sugar is less contained in the wine. I assume that alcohol is less dense than other ingredients and residual sugar is more dense.

Now, let's make plots that shows the difference of red wine and white wine!



The box plots above provide very interesting intuitions about the difference of red wine and white wine. While there is no big difference between quality and alcohol in both wines, white wines includes has less acidity and but more sugar and sulfur dioxide in average.

Bivariate Analysis

Talk about some of the relationships you observed in this part of the investigation.

Quality correlates moderately with alcohol and density.

As alcohol percentage increases, the quality tends to increase as well. However, the slope is very gentle. The correlation efficient of quality and alcohol is 0.436.

On the other hand, as density increases, the quality tends to decrease. The correlation efficient of quality and density is -0.307.

I also found there are a difference in ingredients as the type of wine differs. While there is no big difference between quality and alcohol in red and white wines, white wines includes has less acidity and but more sugar and sulfur dioxide in average.

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

I observed that density correlates strongly with alcohol and residual sugar.

As residual sugar increases, the density increase as well.The relationship appears to be linear. The correlation efficient of density and alcohol is 0.839.

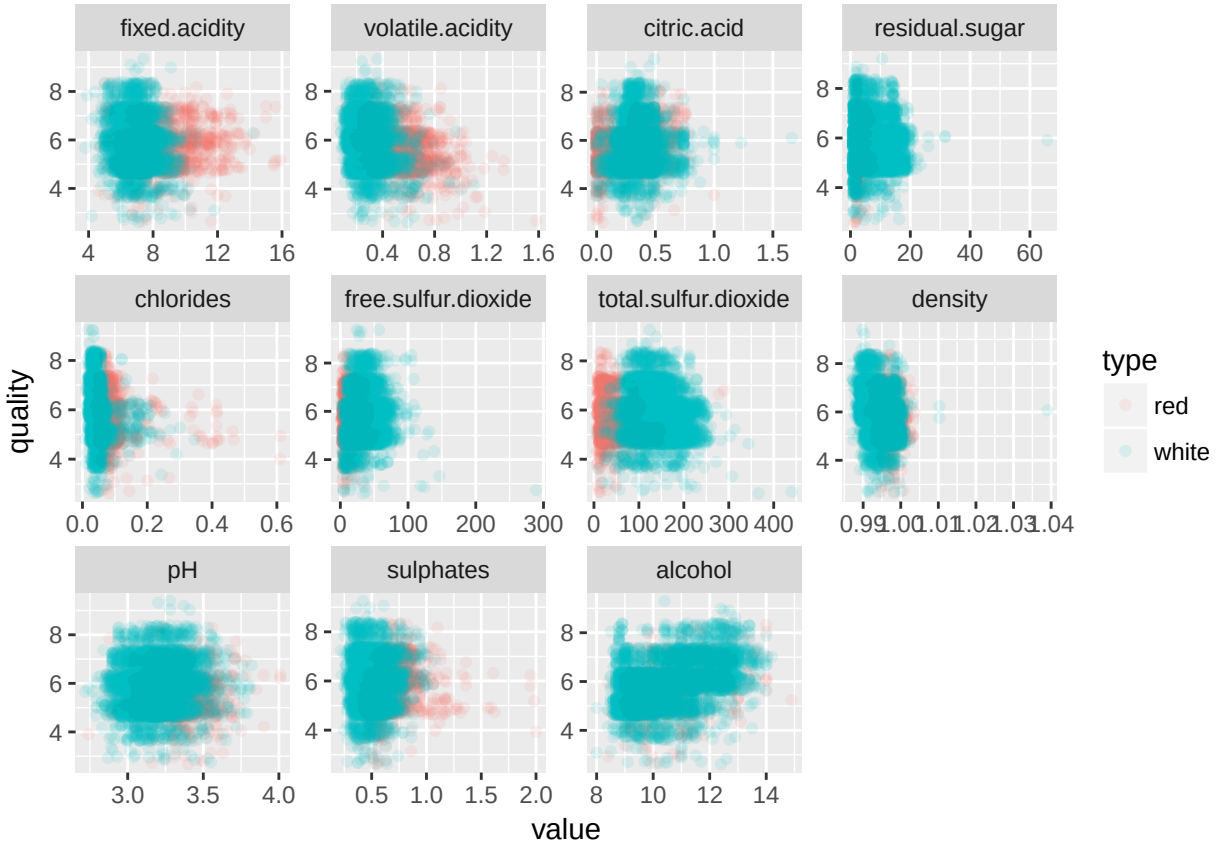
On the other hand, as alcohol increases, density tends to decrease. The relationship between alcohol and density appears to be linear. The correlation efficient of quality and density is -0.780.

What was the strongest relationship you found?

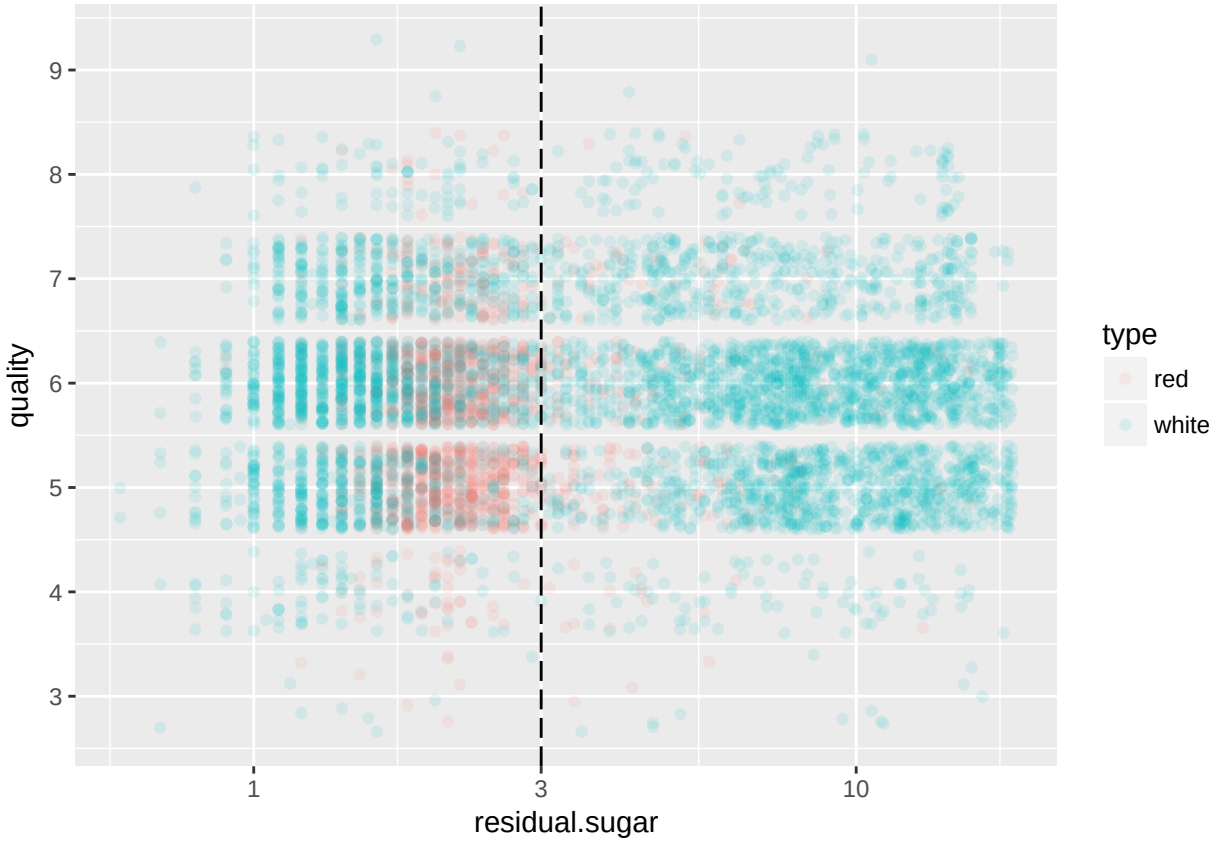
The strongest relationship that was between free sulfur dioxide and total sulfur dioxide with correlation 0.72. This seems to be a logical result.

Multivariate Plots Section

Let's take a look at the scatter plots between quality and other features with colored type.

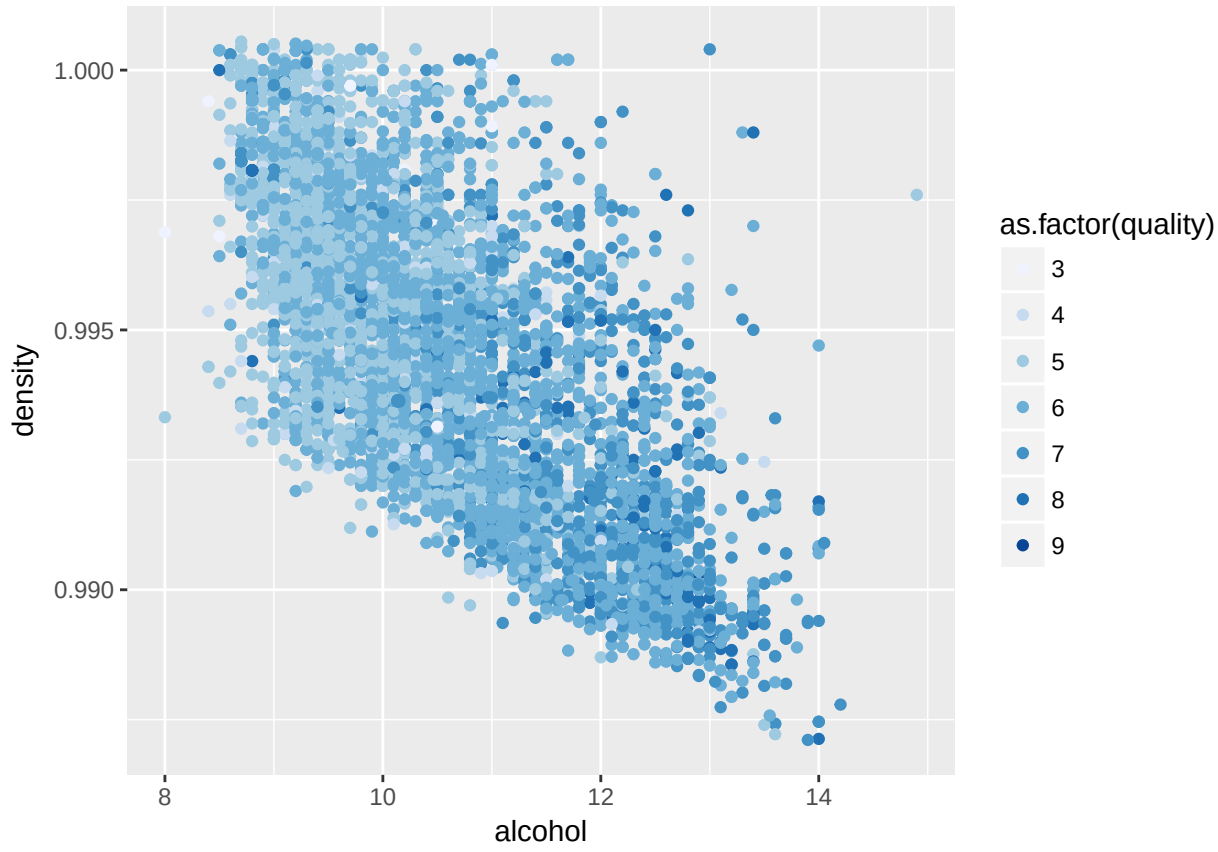


As we have seen in the boxplots, we can see some separation of red wines and white wines in each box plots. red wines are observed in more acidity but less sulfur dioxide area. However, in the residual sugar-quality plot, it is not so clear how the data points are distributed. Let's draw a seprate plot.

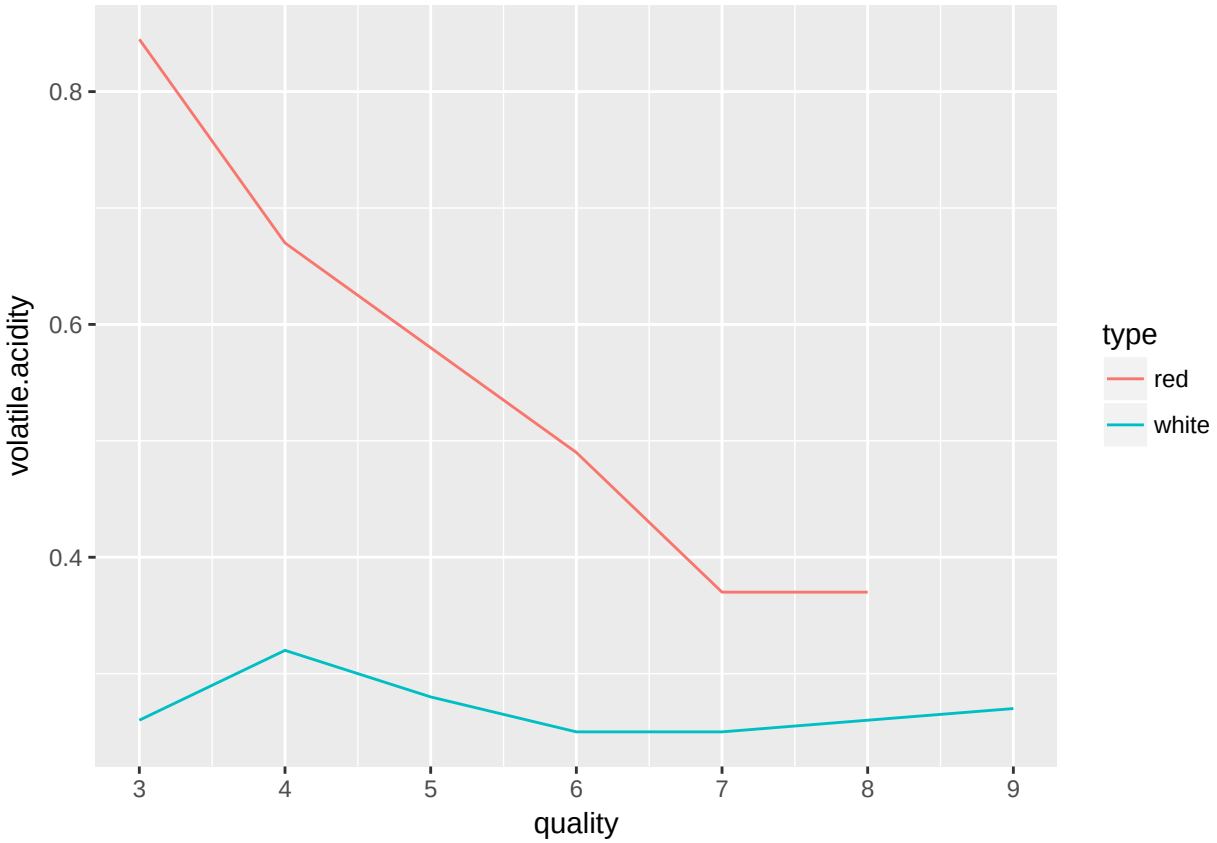


X-axis is scaled to logarithm, as we did for residual sugar it in the univariate section. The dashed line is where residual sugar is 3. We can observe that the most red wines are located on the left side of the dashed line. I think red wines containes relatively less sugar and white wines has diverse sorts of wines in terms of residual sugar amount.

Let's plot the most correlated features with quality, alcohol and density with quality.



As we observed in the bivariate section, alcohol and density scatter plot shows a negative relationship. lighter blue points are observed more on the left upper part of the plot, which represents less alcohol and high density wines get low ratings by judges.



Based on the correlation matrix, volatile acidity and quality has a negative relationship with correlation efficient -0.27. However, we can see that the quality change line appears totally different by the wine type.

The median volatile acidity of red wines decreases as the quality of red wine increases. On the other hand, the median volatile acidity of white wines does not show big differences as the quality changes.

I can assume that the volatile acidity is a good criteria to measure a quality of a red wine, but not for a white wine.

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?

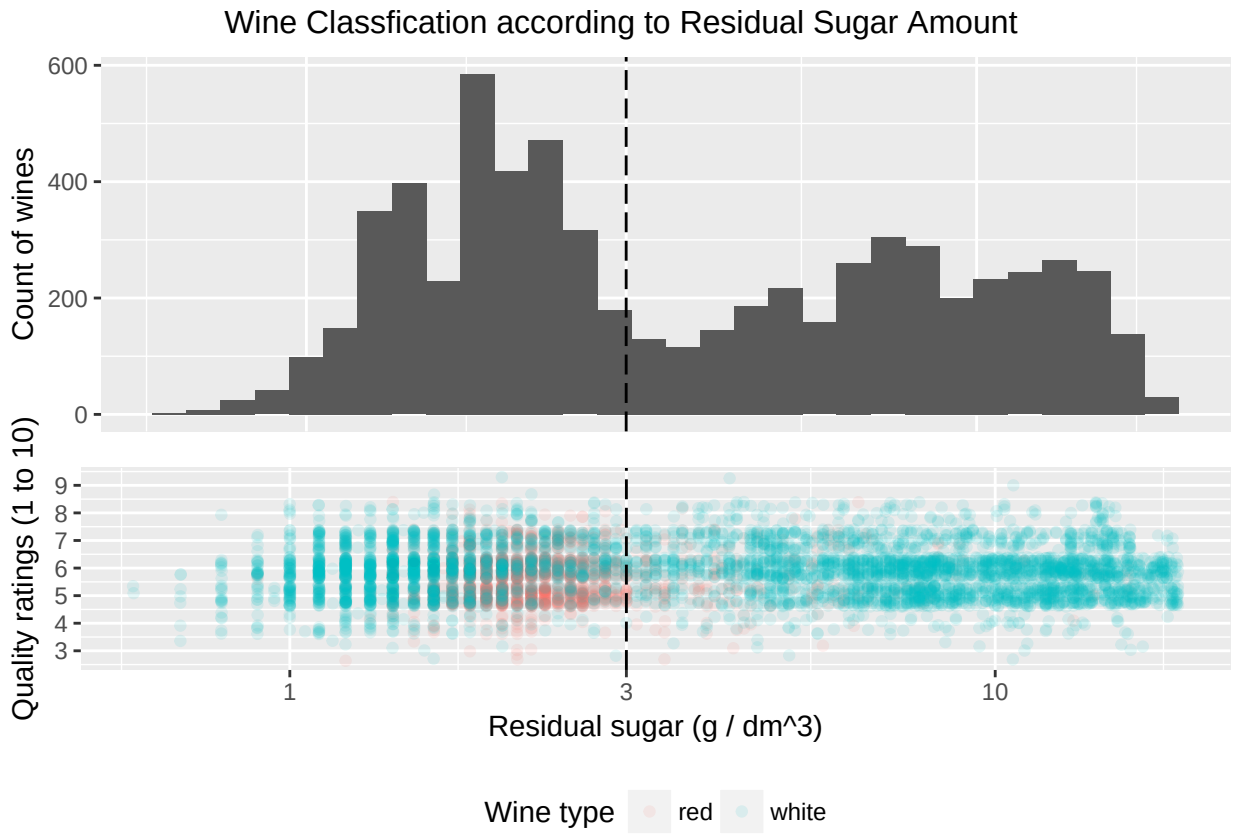
Using multivariate Analysis, I was able to visualize how the data points of red wines and white wines are located. I could find out that the residual sugar of red wines are mostly below 3. Also, the relationship between quality, alcohol and density is strengthened.

Were there any interesting or surprising interactions between features?

I was able to observe that the average volatile acidity trend as the quality increases totally differs by the type of wine. I found out the volatile acidity is a good criteria to measure a quality of a red wine, but not for a white wine.

Final Plots and Summary

Plot One

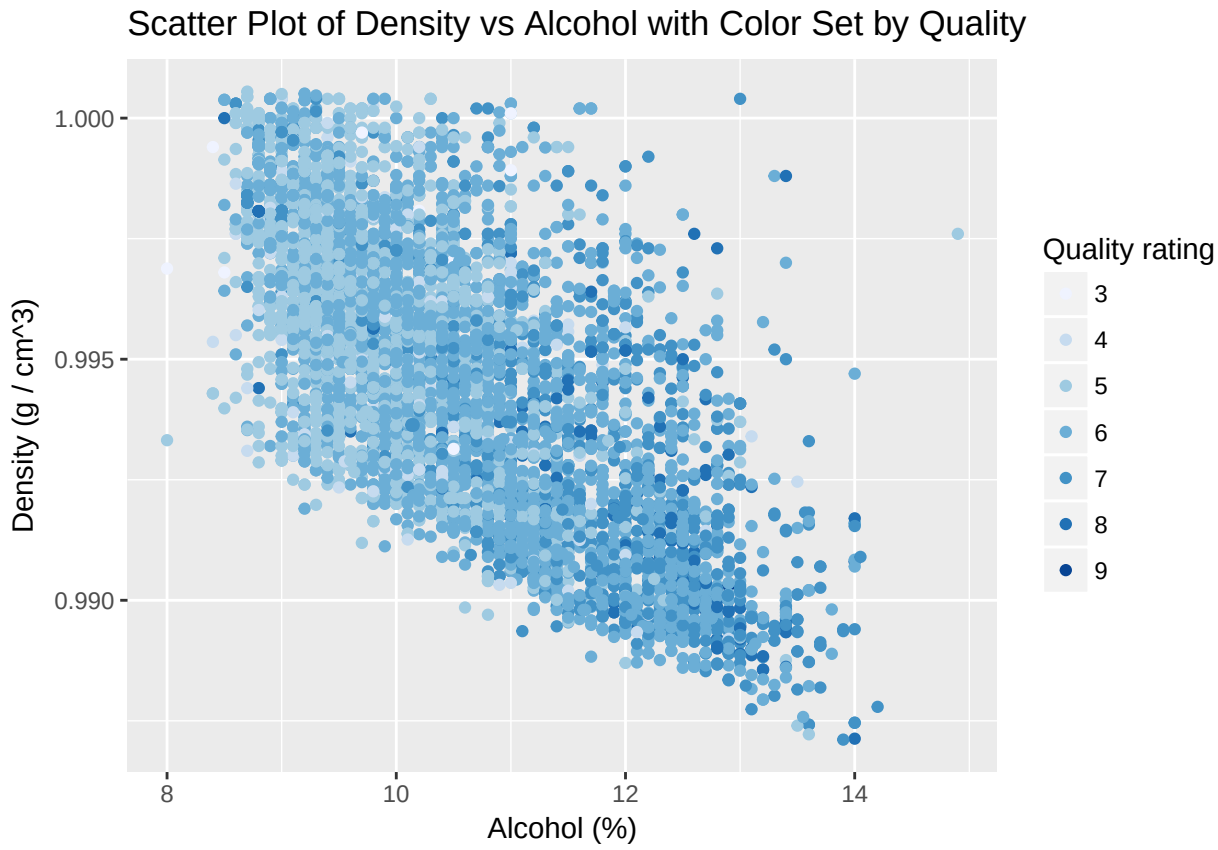


Description One

The distribution of residual sugar of wines appear to be bimodal on log scale. So I divided the group to the one that has residual sugar less than 3 and the other that has residual sugar more than 3.

By plotting it in a scatter plot with coloring the types, I discovered residual sugar and quality has no noticeable relationship. Also, most of the red wines are observed in the less residual sugar group. And white wines are observed quite evenly over all the residual sugar range.

Plot Two



Description Two

The plot reflects the relationship between alcohol, density and quality in a single plot.

In general, density tends to decrease as alcohol percentage increases, even though the features are not strongly correlated. I guess it is because alcohol is less dense than other ingredients. Also, we can observe that light blue data points are more observed on the left upper part of plot, which represents that less alcohol and high density wines are relatively poorly rated.

Plot Three

Description Three

The Median of volatile acidity for each quality rating differs evidently by the type of wine.

The median volatile acidity of red wines decreases as the quality of red wine increases. On the other hand, the median volatile acidity of white wines does not show big differences as the quality changes.

Perhaps, the volatile acidity is a good criteria to measure a quality of a red wine, but not for a white wine.

Reflection

The wine data set contains information on 6,497 wines across 13 variables. I started to explore the dataset by plotting variables and continued to search for interesting relationships. Eventually I discovered some interesting relationship between quality, type and other features.

In early phase of exploration, I thought there must be a main variable that explaines the quality rating of wines. I was surprised that it is possible that it is hard to explain the quality of variable even if I consider every variable in the dataset.

Also, I discovered the general trend of a variable can be a lot different when you seperate another variable and observe the trends of each seprated variable. For example, volatile acidity and quality shows a negative correlation in general. However, if you divide the dataset into red wine and white wine and observe the

volatile acidity by quality, you can see red wine shows a rapid decrease in volatile acidity as quality rating increases. On the other hand, the volatile acidity of white wine cannot explain much about the change of quality ratings.

To investigate the data further, I would like to get the information of how long the wines are aged. Normally well-aged wines are considered as quality wines, so I think this variable could predict the quality of wine better. Also, it would be interesting to trace the change of ingredients as the wine ages.