Exploratory Data Analysis—Exploring-White-Wine-Data

Shawn Gong

May 16th, 2018

White Wine Quality Exploratory Data Analysis by Shawn Gong

Description of Dataset

The dataset contains 4898 rows of data regarding the quality of white wine. It contains 12 variables, while one of them is output variable, quality. The database contains only chemical attributes about white wines, and there's no selling price, grape type and branding available. Hence there will be no comparisons of white wine from different wineries or selling price.

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.

Available at: [@Elsevier] http://dx.doi.org/10.1016/j.dss.2009.05.016 [Pre-press (pdf)] http://www3.dsi.uminho.pt/pcortez/winequality09.pdf [bib] http://www3.dsi.uminho.pt/pcortez/dss09.bib

Variable information

- fixed acidity: most acids involved with wine or fixed or nonvolatile (do not evaporate readily)
- volatile acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
- citric acid: found in small quantities, citric acid can add 'freshness' and flavor to wines
- residual sugar: the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet g/L
- chlorides: the amount of salt in the wine
- free sulfur dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine mg/L
- total sulfur dioxide: amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine mg/L
- density: the density of water is close to that of water depending on the percent alcohol and sugar content
- pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
- sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant
- alcohol: the percent alcohol content of the wine %

- quality: The output variable, score between 0 and 10

Question about the dataset

1. Which attributes of white wine often associated with higher quality?
2. What are the factors that decrease the quality of white wine?

Potential biases

1. The white wine is rated by wine experts, which can result in possible biases as different wine experts may have different taste for white wine. As described in the documentation of this dataset, the quality rating is the median of at least 3 evaluations made by wine experts. The sensory quality rating may cause some level of bias, which cannot be mitigated as the dataset lacks data to point out this bias.
2. There's no year variable available, and there's no information in the documentation, either. There might be quality rating difference if a white wine is stored for years comparing against newly produced white wine.

First, I want to take a look at the data type of each variable.

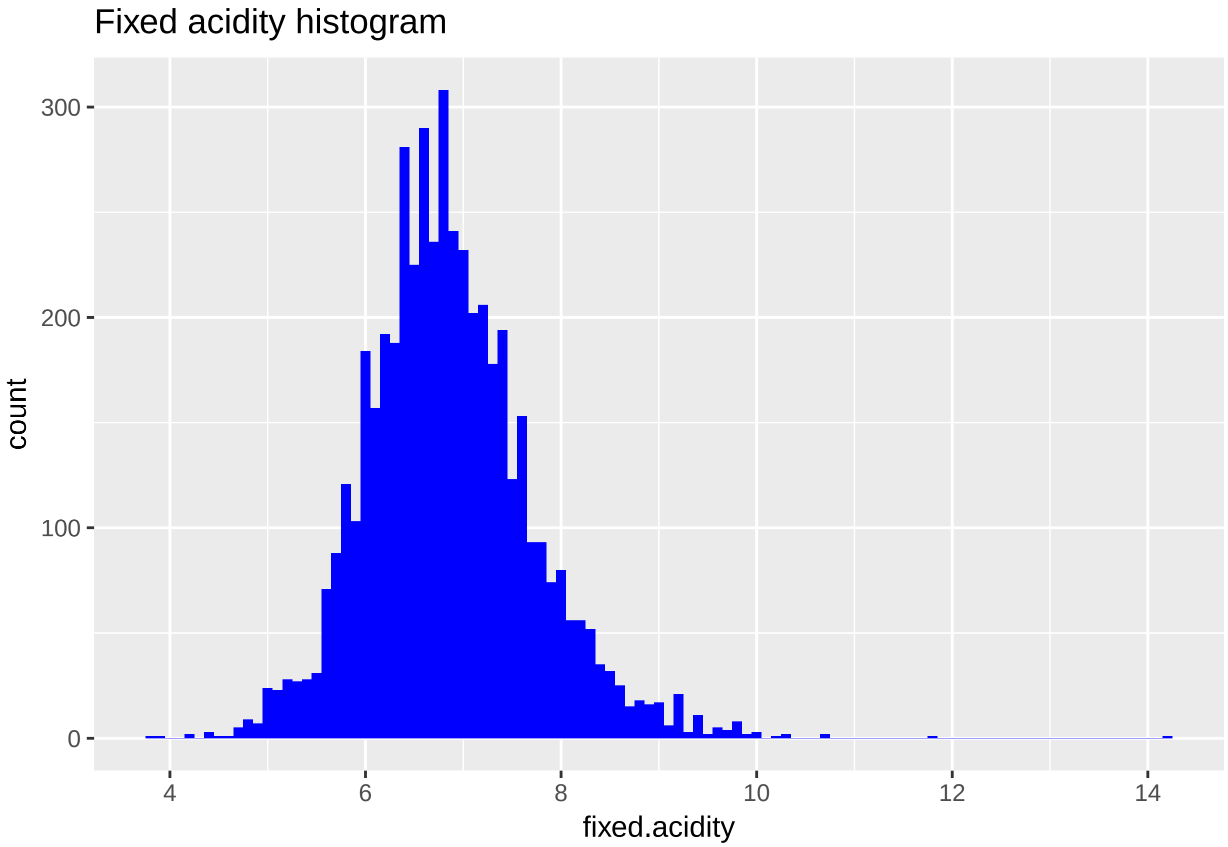
```
## 'data.frame':    4898 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num 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

The X variable is only representing the number of each observation, we are analyzing the entire dataset so that individual number is not relevant, it makes sense to remove the X variable.

```
white_wine = subset(white_wine, select = -c(X) )
```

Univariate Plots Section

Plot 1. Fixed Acidity



```
summary(white_wine$fixed.acidity)
```

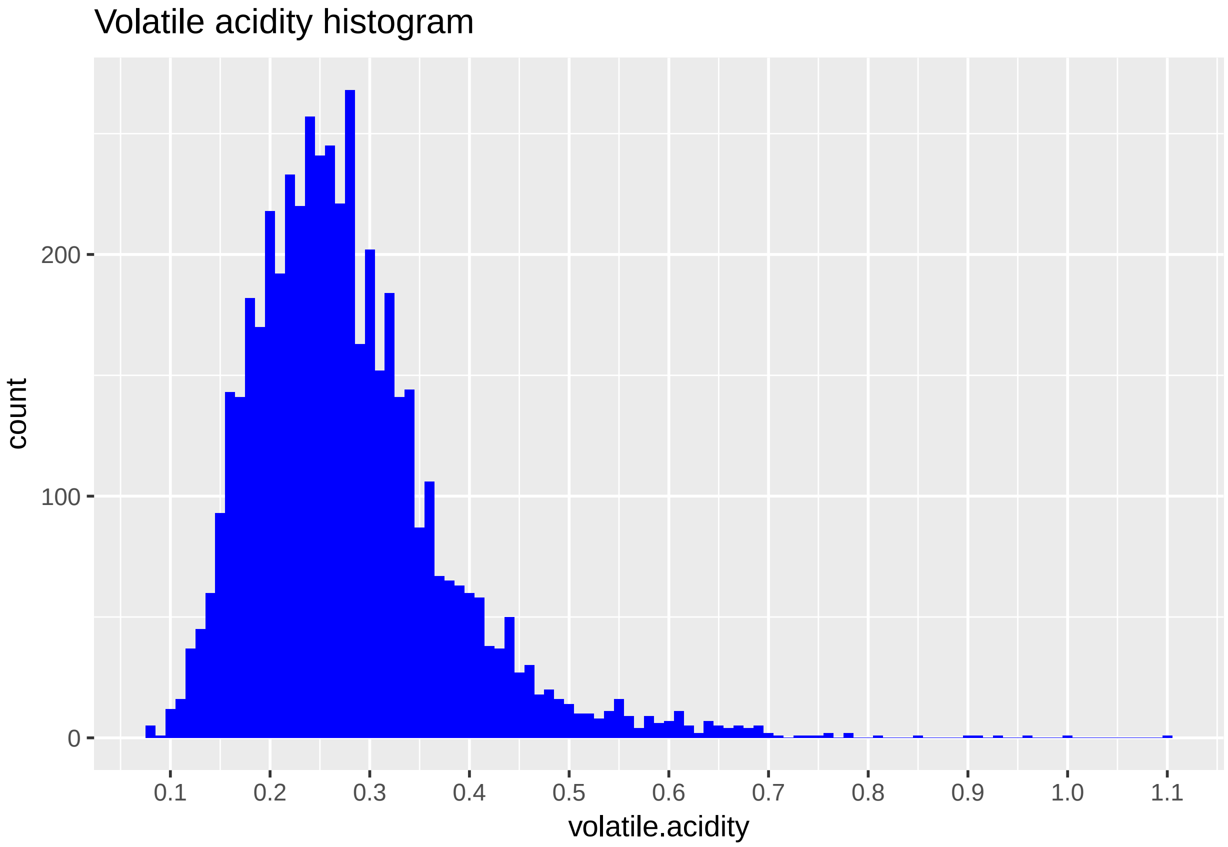
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.800   6.300   6.800   6.855   7.300  14.200
```

```
sd(white_wine$fixed.acidity)
```

```
## [1] 0.8438682
```

According to the histogram, the fixed acidity roughly follows a normal distribution. Most of the ibservations land in the range from 4 to 10. There are some outliers, as well.

Plot 2. Volatile Acidity



```
summary(white_wine$volatile.acidity)
```

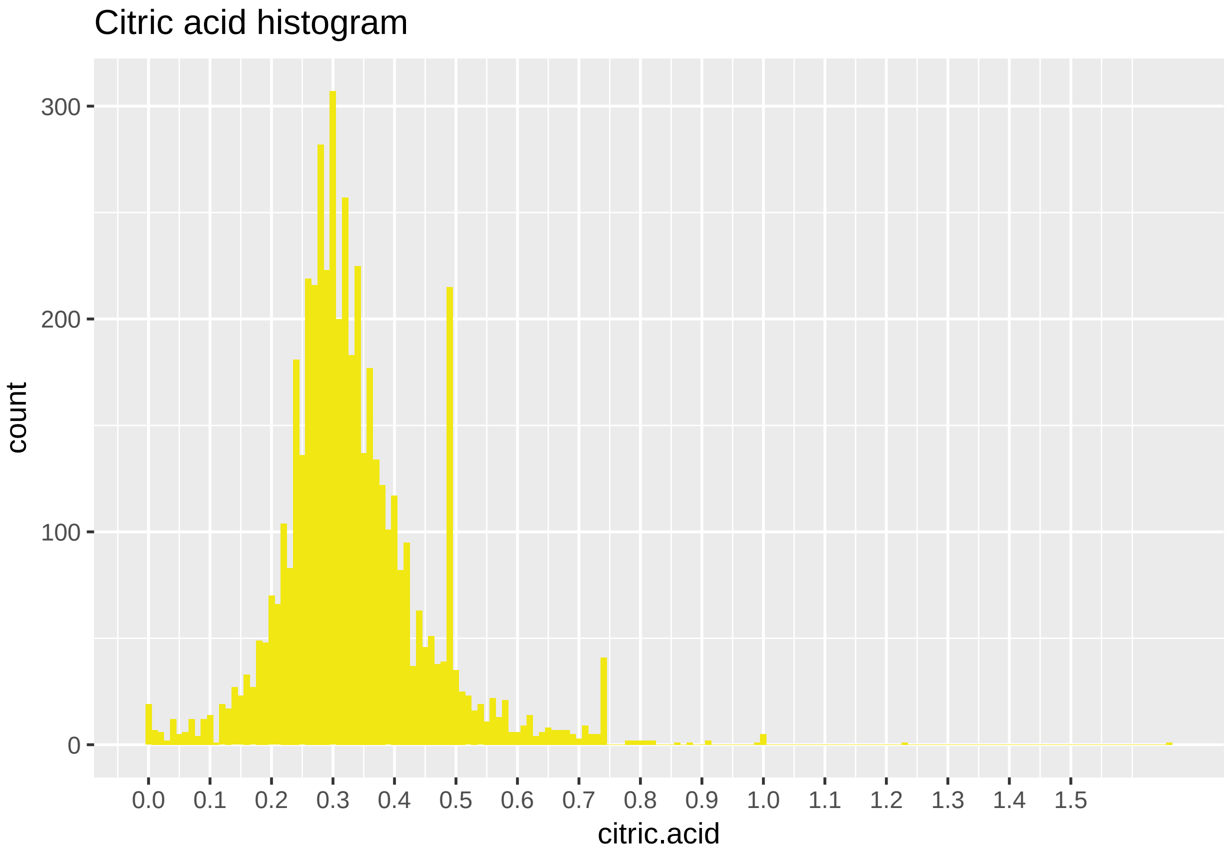
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.2100  0.2600  0.2782  0.3200  1.1000
```

```
sd(white_wine$volatile.acidity)
```

```
## [1] 0.1007945
```

Contrary to fixed acidity, the volatile acidity demonstrates a right skewed distribution.The data is ranging from 0.08 to 1.1, with a median of 0.26 and some extreme case where volatile acidity more than 1.1.

Plot 3. Citric Acid



```
summary(white_wine$citric.acid)
```

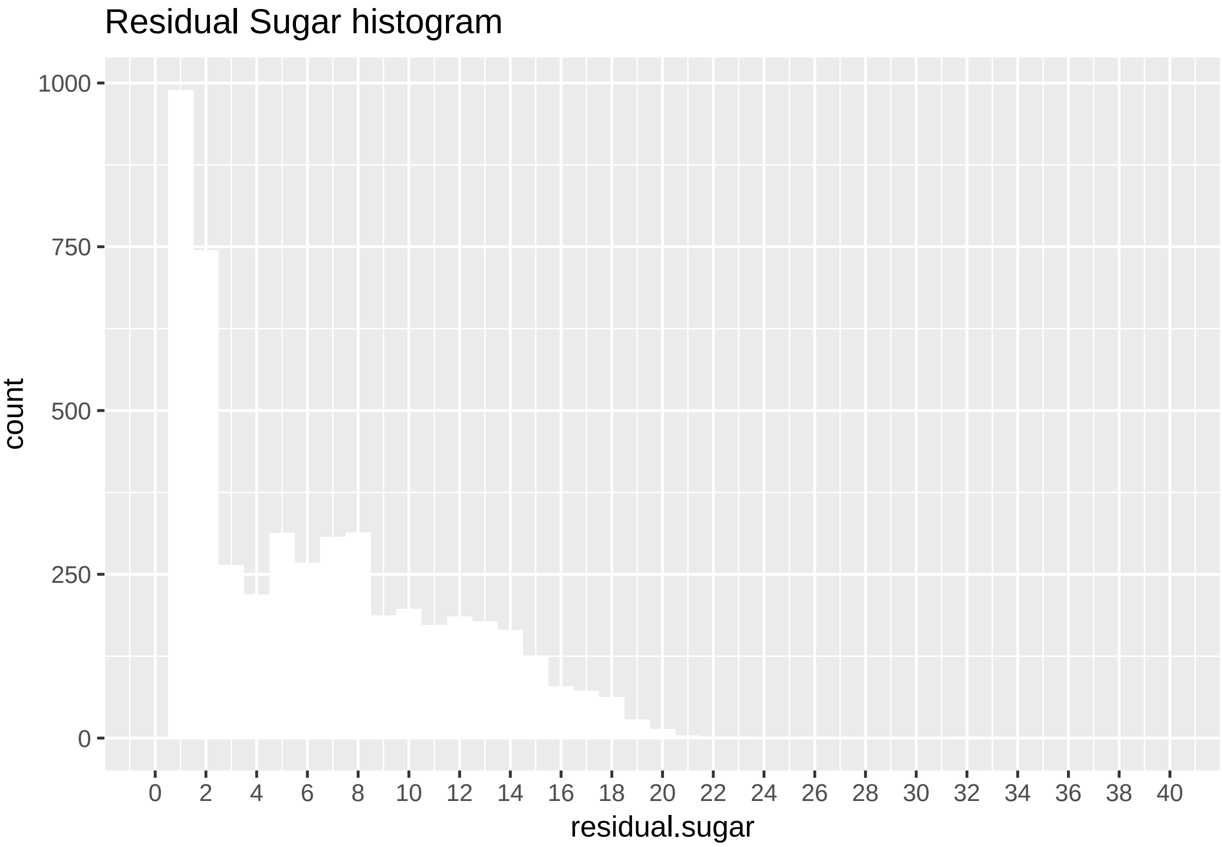
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.0000  0.2700  0.3200  0.3342  0.3900  1.6600
```

```
sd(white_wine$citric.acid)
```

```
## [1] 0.1210198
```

As the histogram shown, the citric acid variable demonstrates roughly a normal distribution but with a slight right skew. Additionally, there are unexpected high quantity of observations at the level of 0.49, which probably a result of the way of recording the data or a particular production process.

Plot 4. Residual Sugar



```
summary(white_wine$residual.sugar)
```

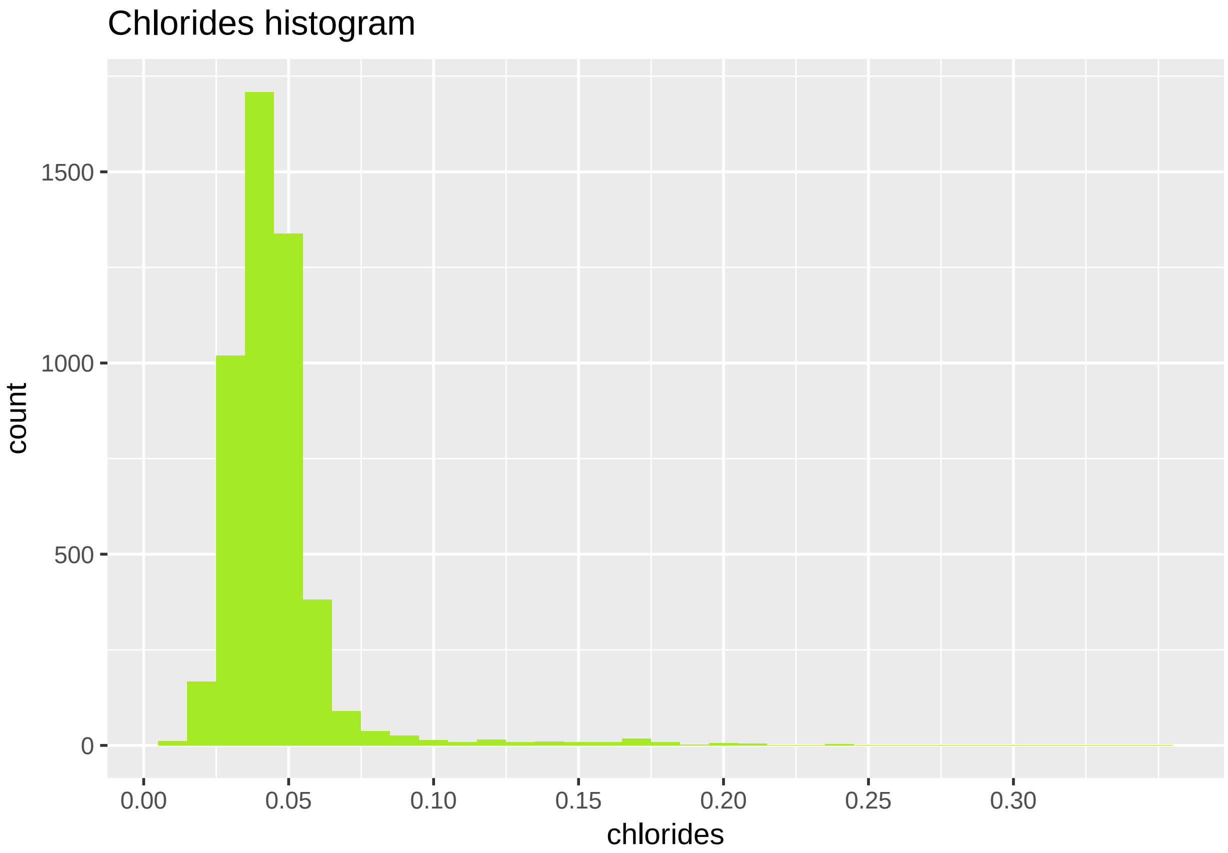
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.600   1.700   5.200   6.391   9.900  65.800
```

```
sd(white_wine$residual.sugar)
```

```
## [1] 5.072058
```

The residual sugar variable shows a right skewed distribution. The residual sugar level lies between 0.6 gram per liter to gram per litrewhile the median lies on 5.2 gram per litre. This outlier possibly is a result from the nature of the white wine. Some ice wine or noble rotten wine can be very sweet, as the residual sugar level can be as high as 340 gram per litre. Judging by the histogram, this dataset mainly contains the testing data of "normal" white wine instead of sweet white wine. Also the number of observations of sweet wine is very low, we can exclude the outlier.

Plot 5. Chlorides



```
summary(white_wine$chlorides)
```

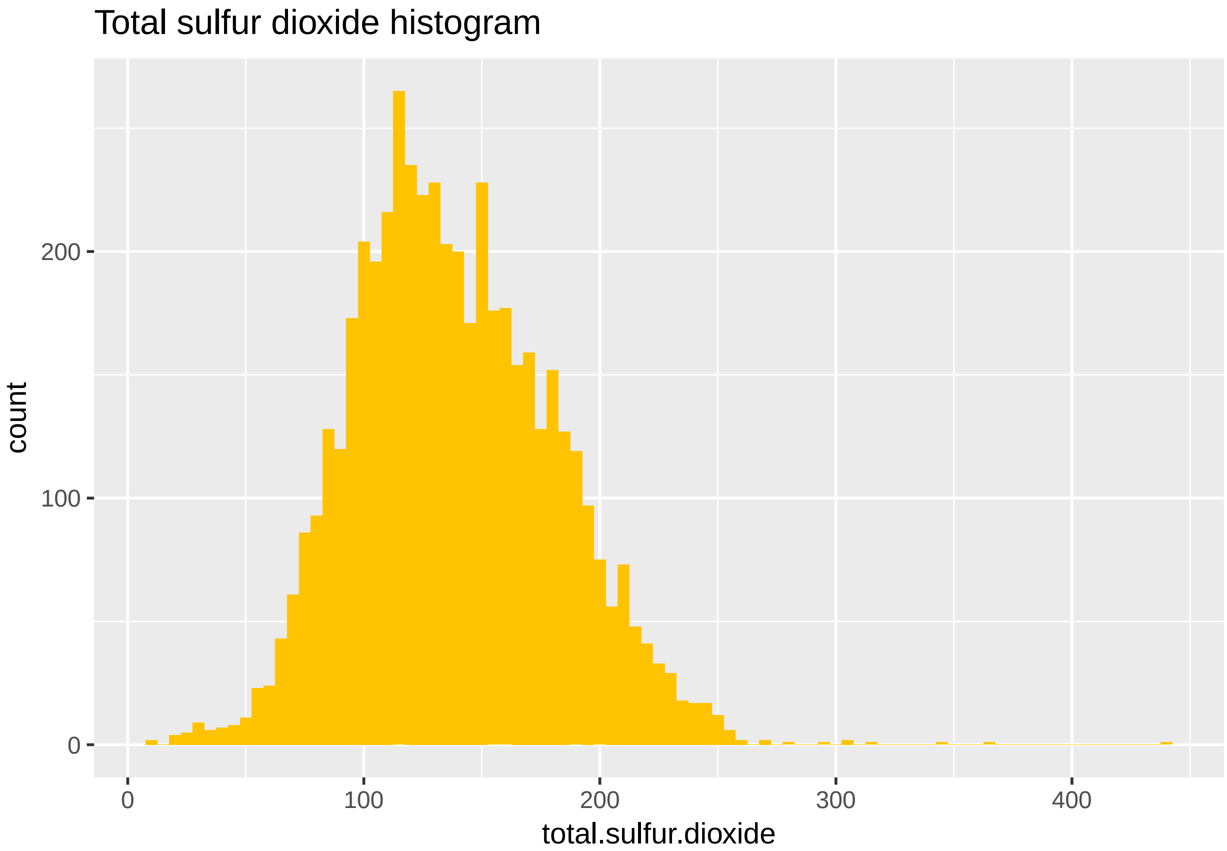
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00900 0.03600 0.04300 0.04577 0.05000 0.34600
```

```
sd(white_wine$chlorides)
```

```
## [1] 0.02184797
```

We can see that the level of chlorides in the white wine roughly follows a right skewed distribution, as well. Most of the observations in the dataset have chlorides level ranging from 0.25 to 0.10.

Plot 6. Total Sulfur Dioxide



```r
summary(white_wine$total.sulfur.dioxide)
```

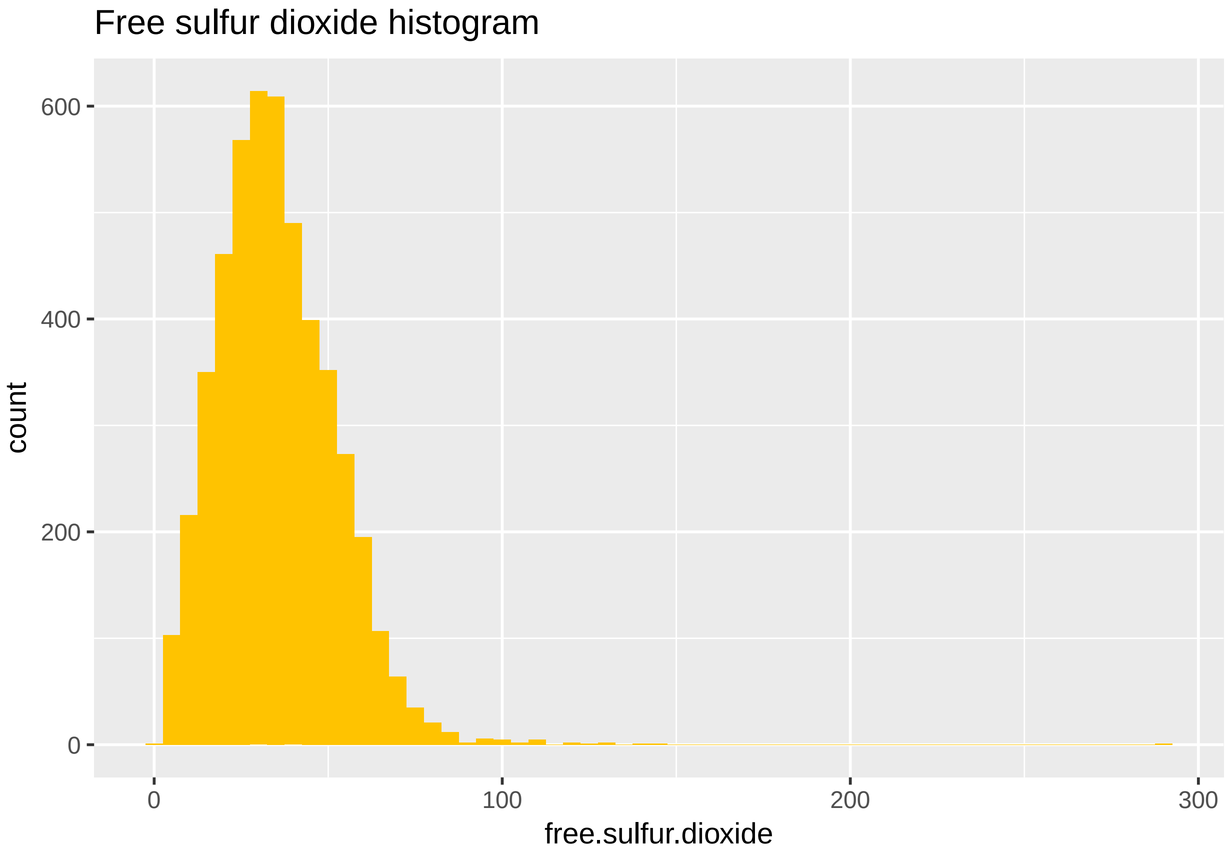
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.0   108.0   134.0   138.4   167.0   440.0
```

```r
sd(white_wine$total.sulfur.dioxide)
```

```
## [1] 42.49806
```

The total sulfur dioxide level in white wine follows a right skewed distribution. I am very curious about how does sulfur dioxide potentially influence the taste of white wine, given that the sulfur dioxide seems not likely to be presented in any production process according to common sense of most people.

Plot 7. Free Sulfur Dioxide



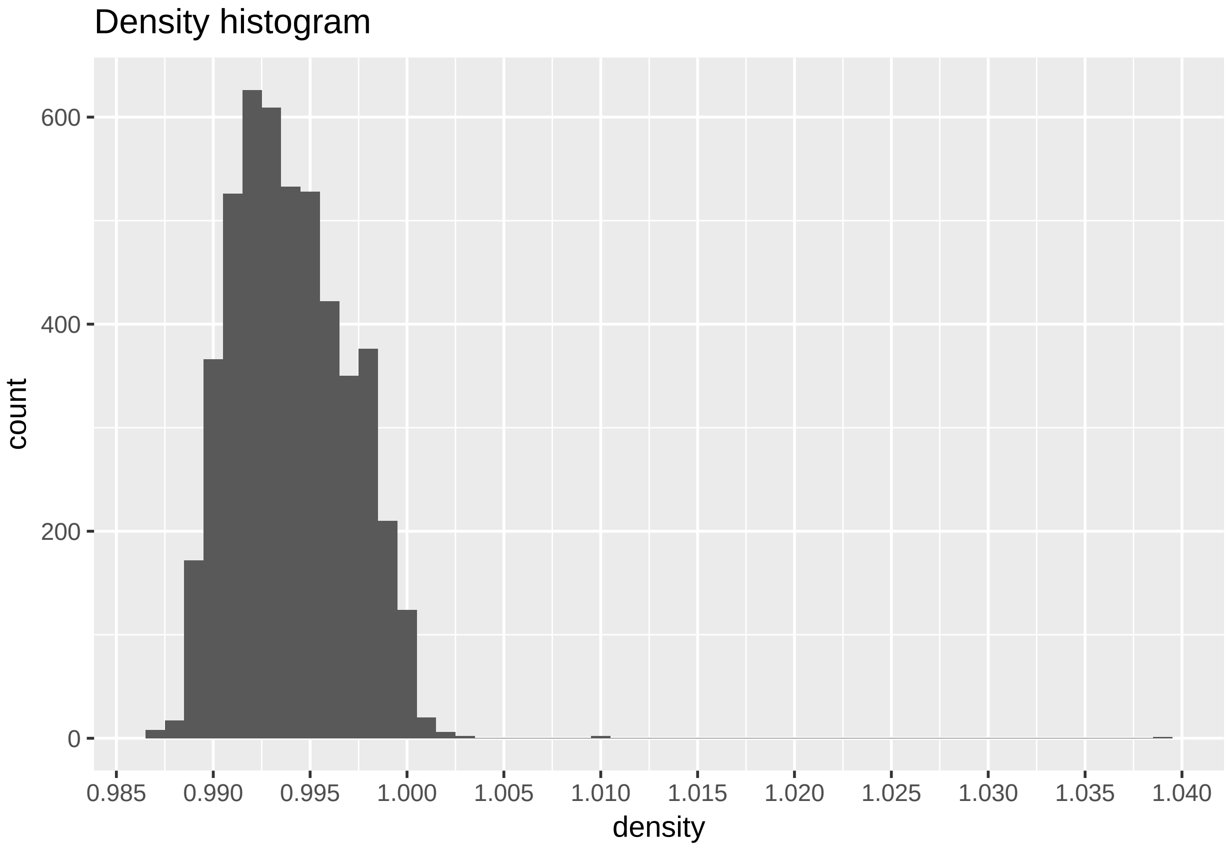
```
summary(white_wine$free.sulfur.dioxide)
```

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

```
sd(white_wine$free.sulfur.dioxide)
```

```
## [1] 17.00714
```

Plot 8. Density



```
summary(white_wine$density)
```

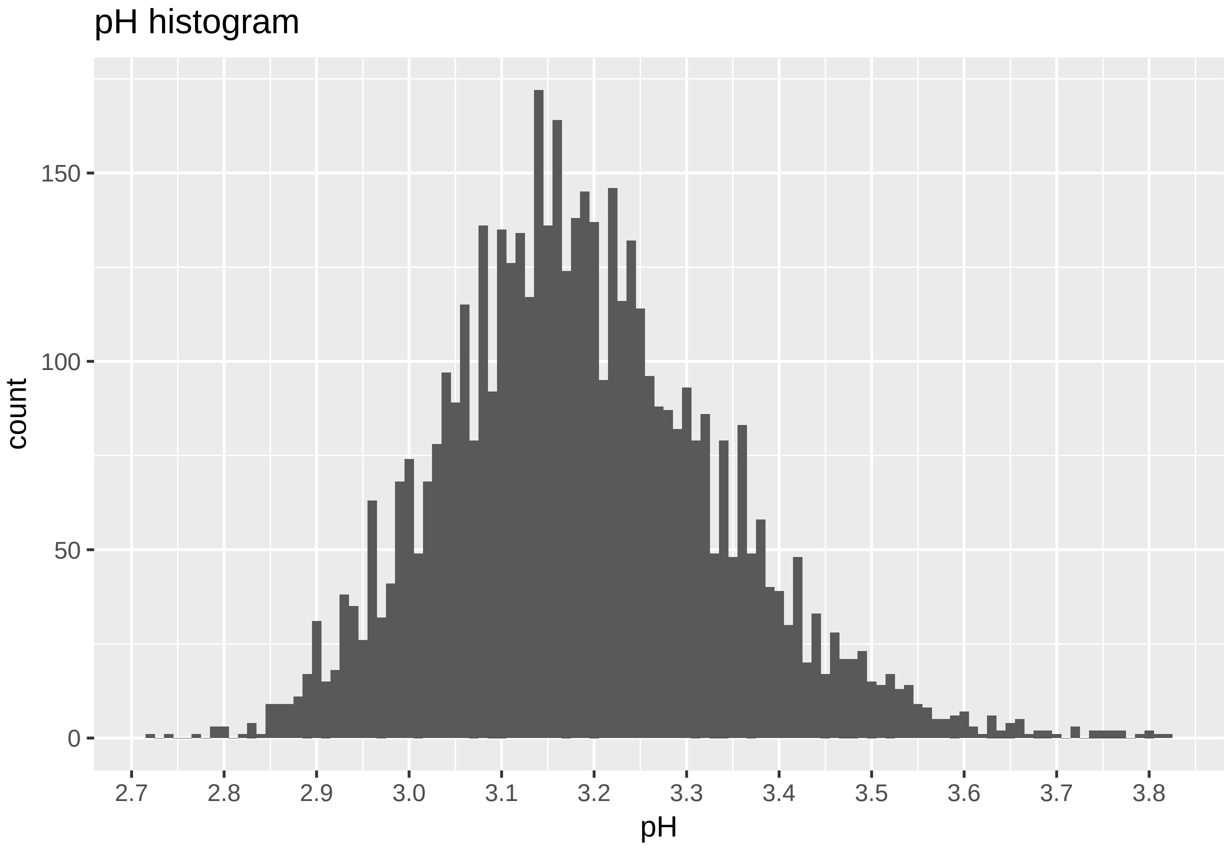
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9917  0.9937  0.9940  0.9961  1.0390
```

```
sd(white_wine$density)
```

```
## [1] 0.002990907
```

As we can see in the histogram, most of the observations have less than 1.0 density. It makes sense as the density of alcohol is lower than water, which makes this variable highly correlated to the alcohol variable. There are some observations with density more than 1.0 however, probably a result from other chemical component of white wine.

Plot 9. pH



```r
summary(white_wine$pH)
```

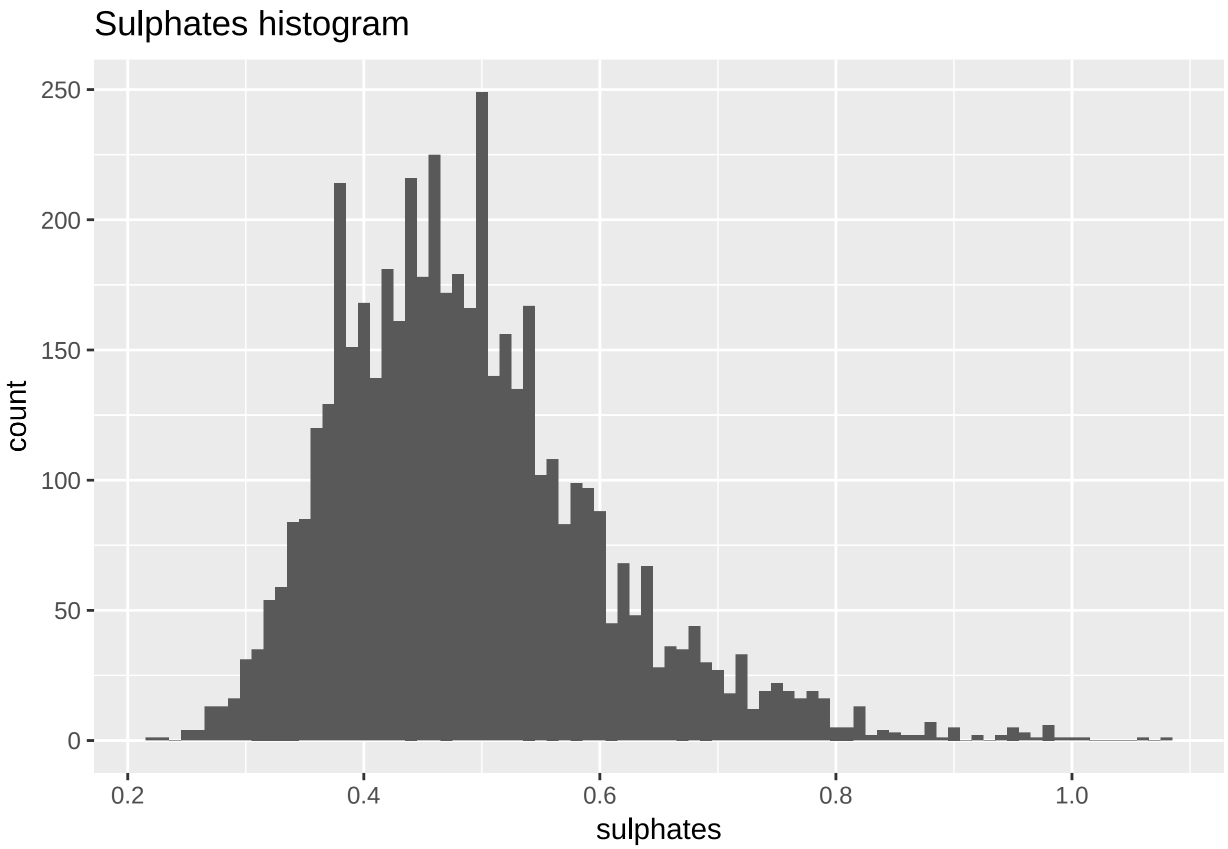
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   2.720   3.090   3.180   3.188   3.280   3.820
```

```r
sd(white_wine$pH)
```

```
## [1] 0.1510006
```

We can see that most of the white wine has a pH ranging from 2.7 to 3.8. Compare to other food, white wine has similar pH to soda, orange juice and lemon juice.

Plot 10. Sulphates



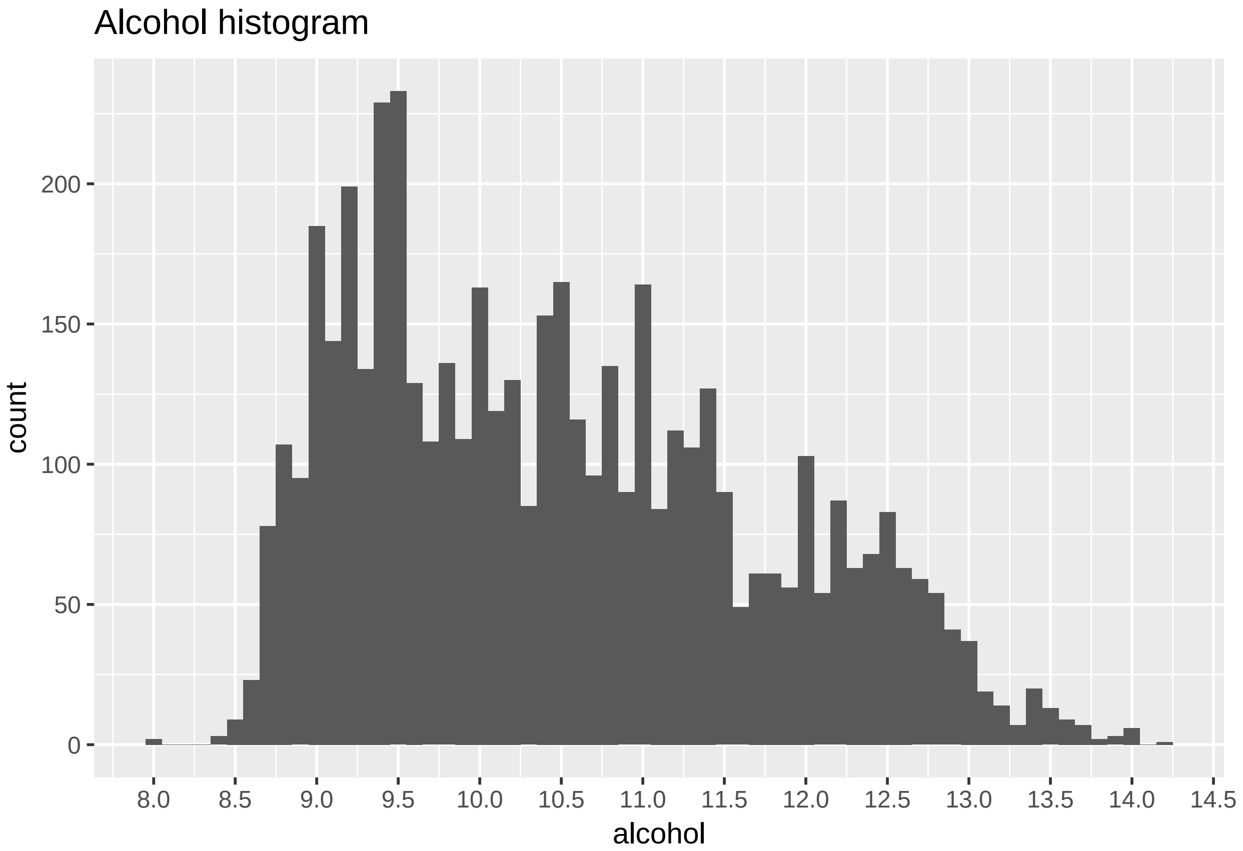
```
summary(white_wine$sulphates)
```

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

```
sd(white_wine$sulphates)
```

```
## [1] 0.1141258
```

Plot 11. Alcohol



```
summary(white_wine$alcohol)
```

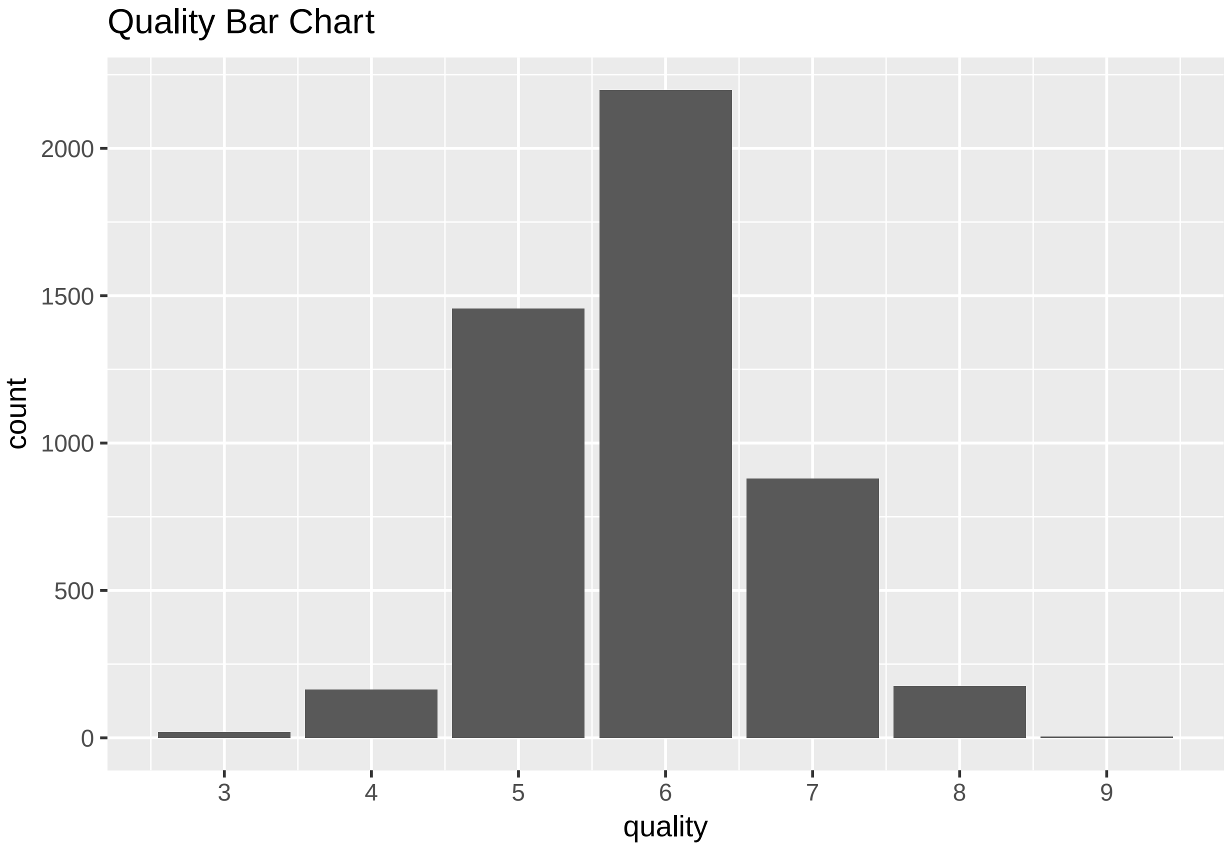
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.50   10.40   10.51   11.40   14.20
```

```
sd(white_wine$alcohol)
```

```
## [1] 1.230621
```

The alcohol component does not follow a common distribution. The alcohol content becomes variant after 9.5%. The alcohol component is ranging from 8.0% to 14.2% as the plot shown.

Plot 12. Quality Rating



```
summary(white_wine$quality)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   3.000   5.000   6.000   5.878   6.000   9.000
```

```
quantile(white_wine$quality, 0.05)
```

```
## 5% 
##  5
```

```
quantile(white_wine$quality, 0.95)
```

```
## 95% 
##   7
```

```
sd(white_wine$quality)
```

```
## [1] 0.8856386
```

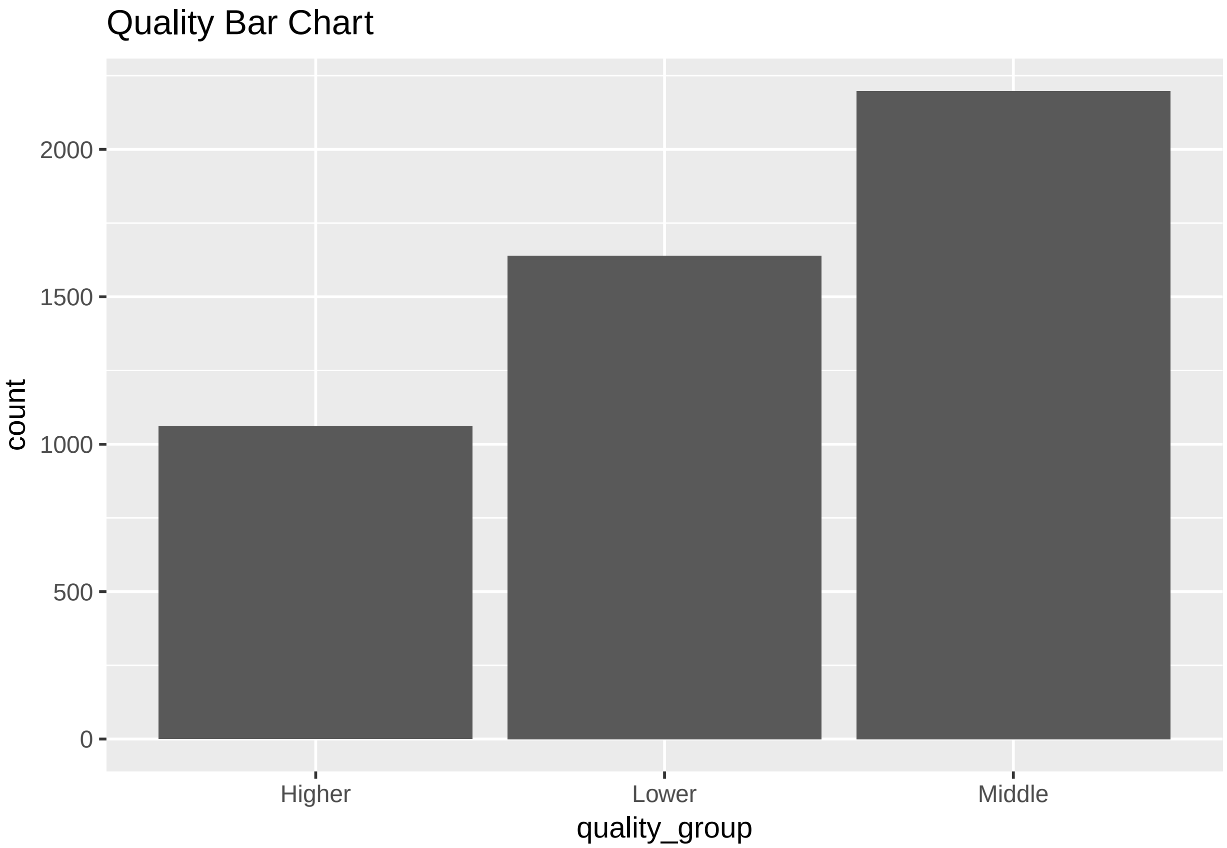
It seems that the most common quality rating for white wine is 6, while white wine with quality rating higher than 6 is less than those which received a lower score. According to the quantile, the middle 90% of observations lies in between 5 to 7. There was no observation with quality score as low as 0, 1 nad 2, nor there was white wine received 10 as quality score. The observation shows that the wine experts are more likely to give out conservative quality scores that is not too high or too low, and it is hard to be perfect in their eyes.

Quick wrangling to make the data more evenly distributed

```
white_wine$quality_group <- ifelse(white_wine$quality == 6, "Middle",
                                   ifelse(white_wine$quality <=5, "Lower", "Higher"))
white_wine$quality_group <- factor(white_wine$quality_group)
```

As we can see in the quality bar chart, the quantity of white wine which received 6 as quality score is much higher than any other quality rating, so that we can aggregate the other quality ratings as summarise them as a group to reduce the difference of each quality group.

Plot 13. Quality group



This time it is much better looking. We will be using this quality group variable in some of the analysis coming in later.

Univariate Analysis

What is the structure of your dataset?

The dataset contains 4898 observations and 13 variables. Among them, there are 11 variables describe the chemical component of white wine, and there's one categorical variable called quality, which is the quality rating given by wine experts for each white wine in the dataset. There is no null data either. Overall the data is quite clean, although there are some outliers in some of the variables. There's no clear evidence that suggest how do these outliers appear.

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

The dataset describes white wine quality using chemical component breakdown. There's no branding, production year available, so we will not be discussing the effect of these factors to white wine. Additionally, the quality rating is based on sensory taste of wine expert, which may cause bias because wine experts could misjudge the quality of white wine as they have different preference for white wine and tend to give conservative quality scores. Especially for medium score, critics can give different scores for white wine that has "average" taste. Additionally, due to the way it records the data (Median of quality from at least 3 wine experts), we tend to get average scores because of the nature of median, that also explains the fact that there is no quality score of 0, 1, 2 and 10. But we have to use sensory observation to measuring the quality of white wine, and the results from wine expert are probably the best available.

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

There are variables follow the similar distributions to each other. It might be useful as sometimes a similar distribution can translate to a higher correlation.

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

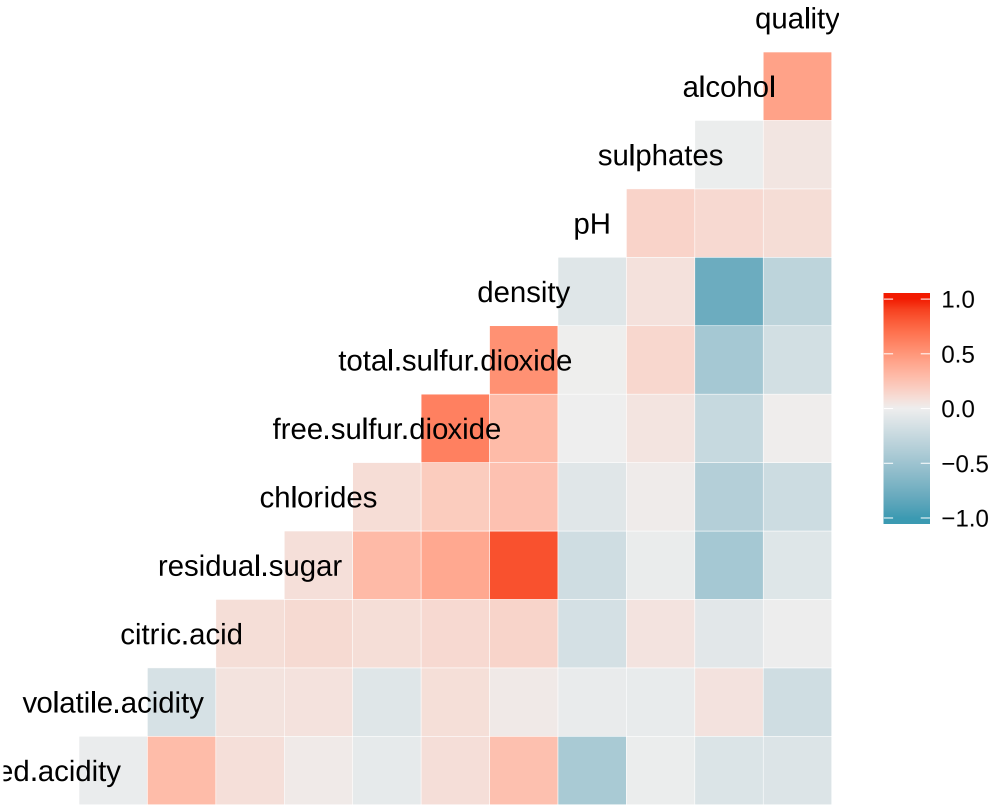
Yes, although each variable in this dataset describes one particular chemical attribute of white wine, which makes it not possible to create variable based on existing variables, I created a quality group variable for better analysis. Most of the existing variales are continuous variables, and we cannot divide them by group based sheer on the range of observations as the quantity of chemical component might not be significant enough to change the taste. But I found that the quality rating variable can be modified to better version. I divided the existing quality score into three different groups: Lower, Middle and Higher. The Lower quality group contains quality rating from 3 to 5, the Middle group only contains observations which received 6 as their result, and the Higher group contains observations received 7 to 9 as the final results. That way we can balance the quantity in each quality group thus benefit further analysis.

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?

The alcohol content demonstrates an unusual distribution where the data becomes variant on the right side of the peak of distribution. This variation might due to that there's no "standard" alcohol content for white wine, the production process of white wine may result in minor variance in alcohol content. I removed the X variable as it only reflects the number of each observation, which is not relevant to our analysis.

Bivariate Plots Section

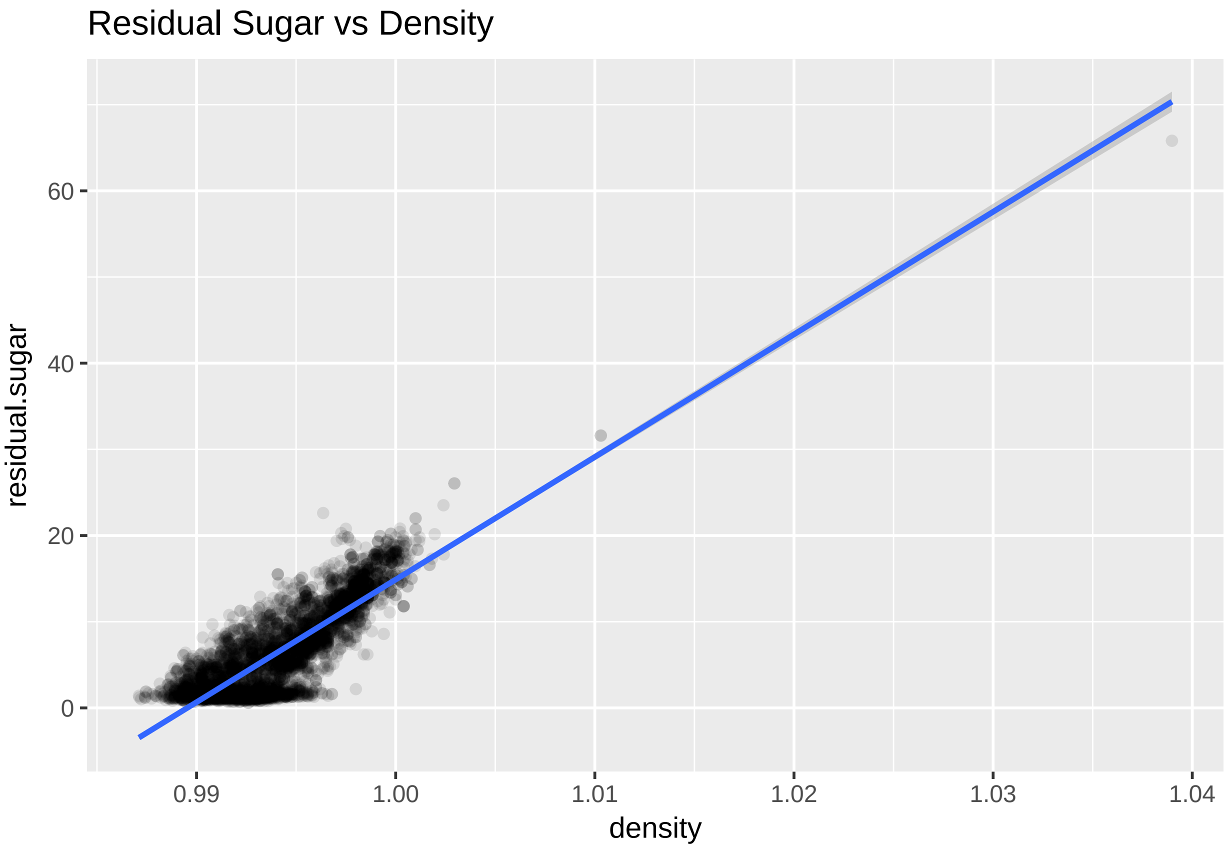
Correlation matrix



Creating a correlation matrix will give us a better idea of how does each variable correlate to each other. As we can see in the plot, Density has strong positive correlation with residual sugar variable, it also has a strong negative correlation with the alcohol variable. The density also have a strong positive correlation with Total Sulfur Dioxide variable, but not as strong as the other correlationship. Additionally, the Total Sulfur Dioxide has a strong correlation with Free Sulfur Dioxide. The above correlations are making sense as they reflect our common sense. We will be focusing on these variables in the following section.

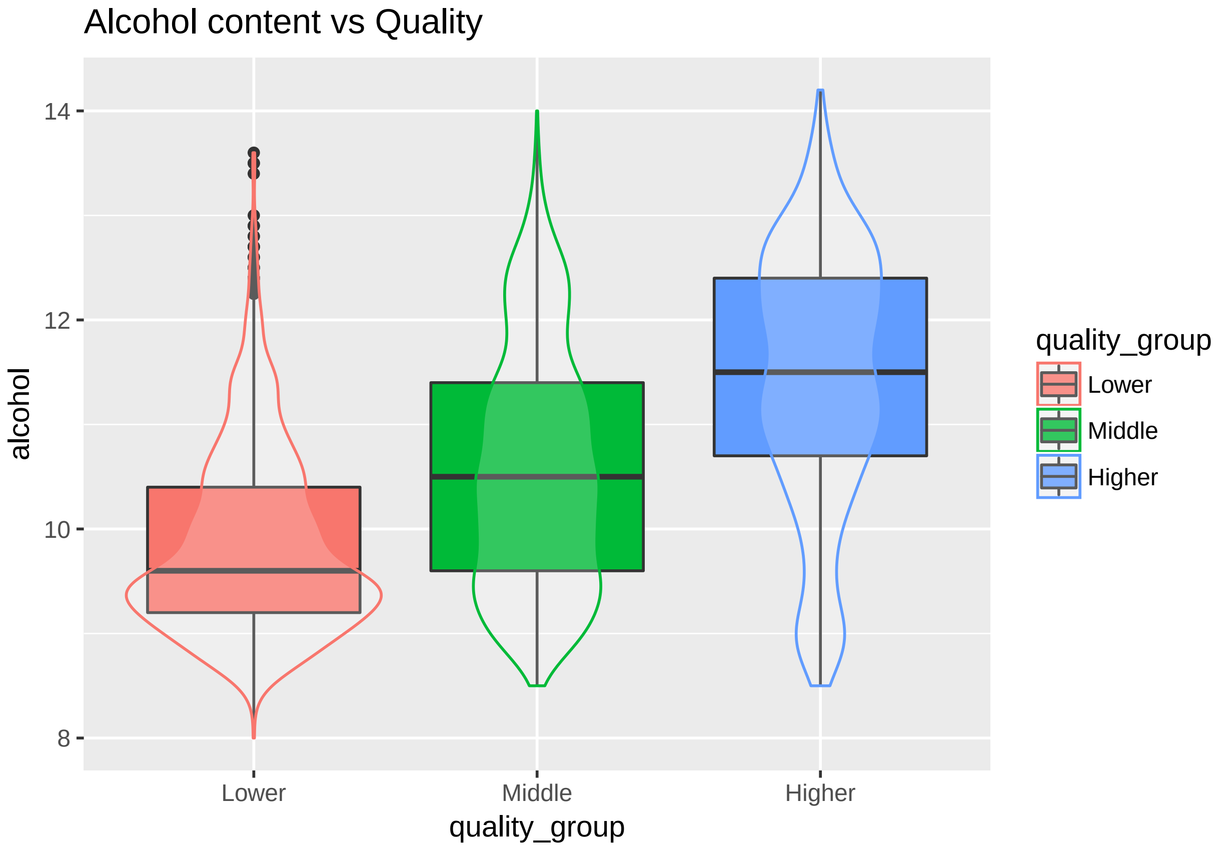
Residual Sugar vs Density

```
ggplot(aes(x = density, y= residual.sugar), data = white_wine) +
  geom_point(position = "jitter", alpha = 1/10) +
  geom_smooth(method = lm) +
  ggtitle("Residual Sugar vs Density")
```



We can see a clear correlation between density and residual sugar, which makes sense as sugar has a higher density than water and alcohol.

Alcohol vs Quality



```
cor(x = white_wine$alcohol, y = white_wine$quality)
```

```
## [1] 0.4355747
```

```
## Lower quality group
with(subset(white_wine, quality_group == "Lower"), summary(alcohol))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.20    9.60    9.85   10.40   13.60
```

```
## Middle quality group
with(subset(white_wine, quality_group == "Middle"), summary(alcohol))
```

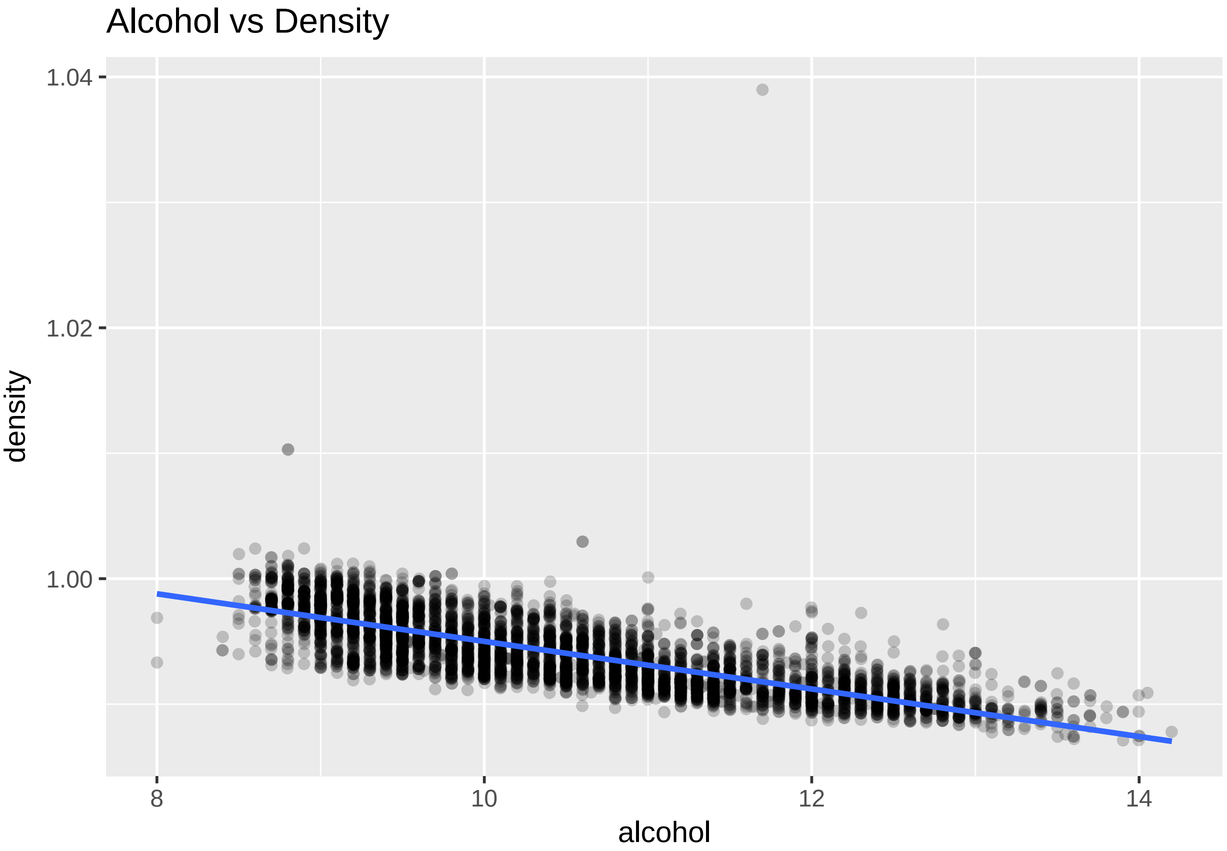
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50    9.60   10.50   10.58   11.40   14.00
```

```
## Higher quality group
with(subset(white_wine, quality_group == "Higher"), summary(alcohol))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50   10.70   11.50   11.42   12.40   14.20
```

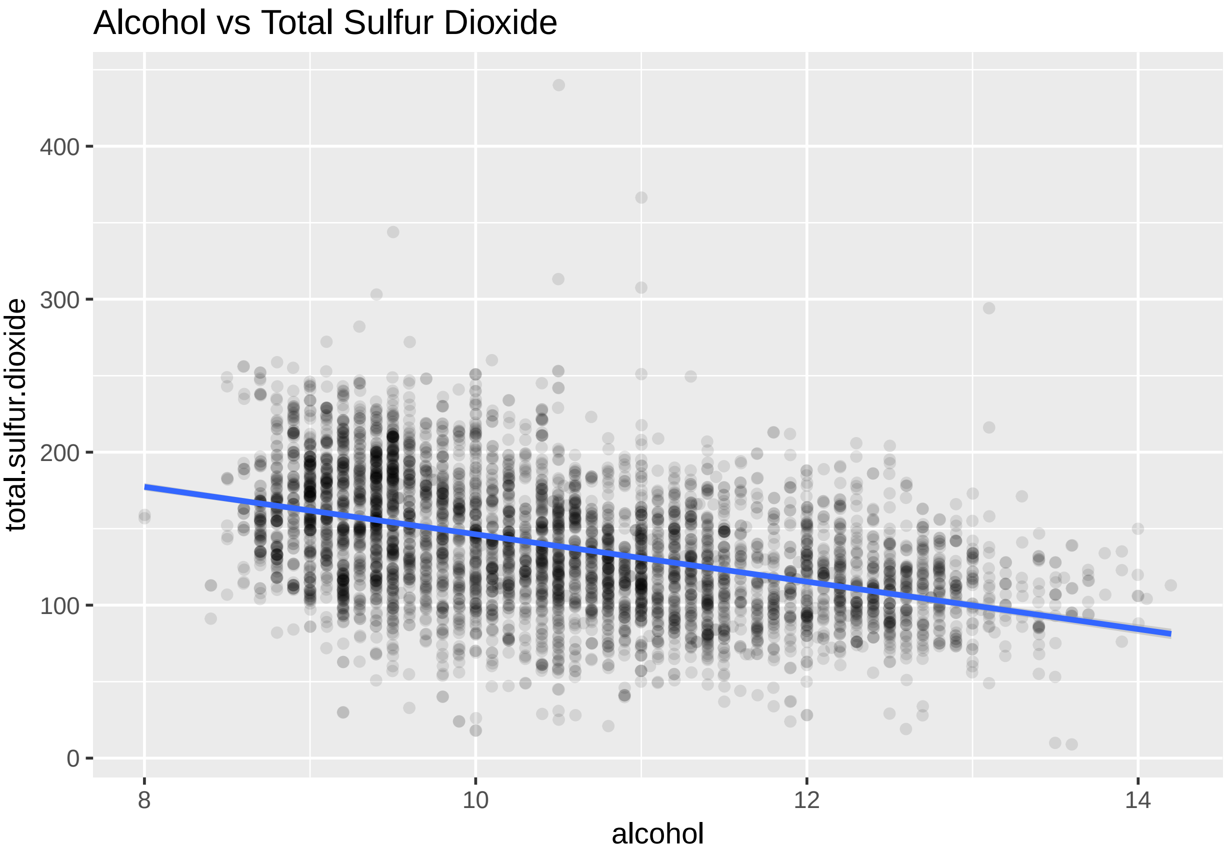
From the plot we can see a interesting phenomenon: white wine with higher alcohol content tend to received higher quality score. Judging from the boxplot, the median of the alcohol content of each quality group is significantly different from each other. Another interesting fact is that the median of alcohol content of Lower quality group locates near the 25% quantile of the Middle group, and the median of the alcohol content of Middle quality group locates near the 25% quantile of the Higher quality group. The correlation between quality and alcohol content is 0.43, which is a positive correlation relationship. But the quality score variable is not suitable for analysis here, as the quantity of observation under each quality score is variant, which may result in lower reliability. But after dividing them by quality group, we can tell that the alcohol content has a positive influence on the quality rating. Higher alcohol content readablily improved the quality rating of white wine.

Alcohol vs. Density



The scatter plot demonstrates a strong correlation between It is common sense that alcohol has lower density compare to water. So the proportion of alcohol in white wine will have a direct influence on the density of white wine, as well.

Alcohol vs. Total Sulfur Dioxide

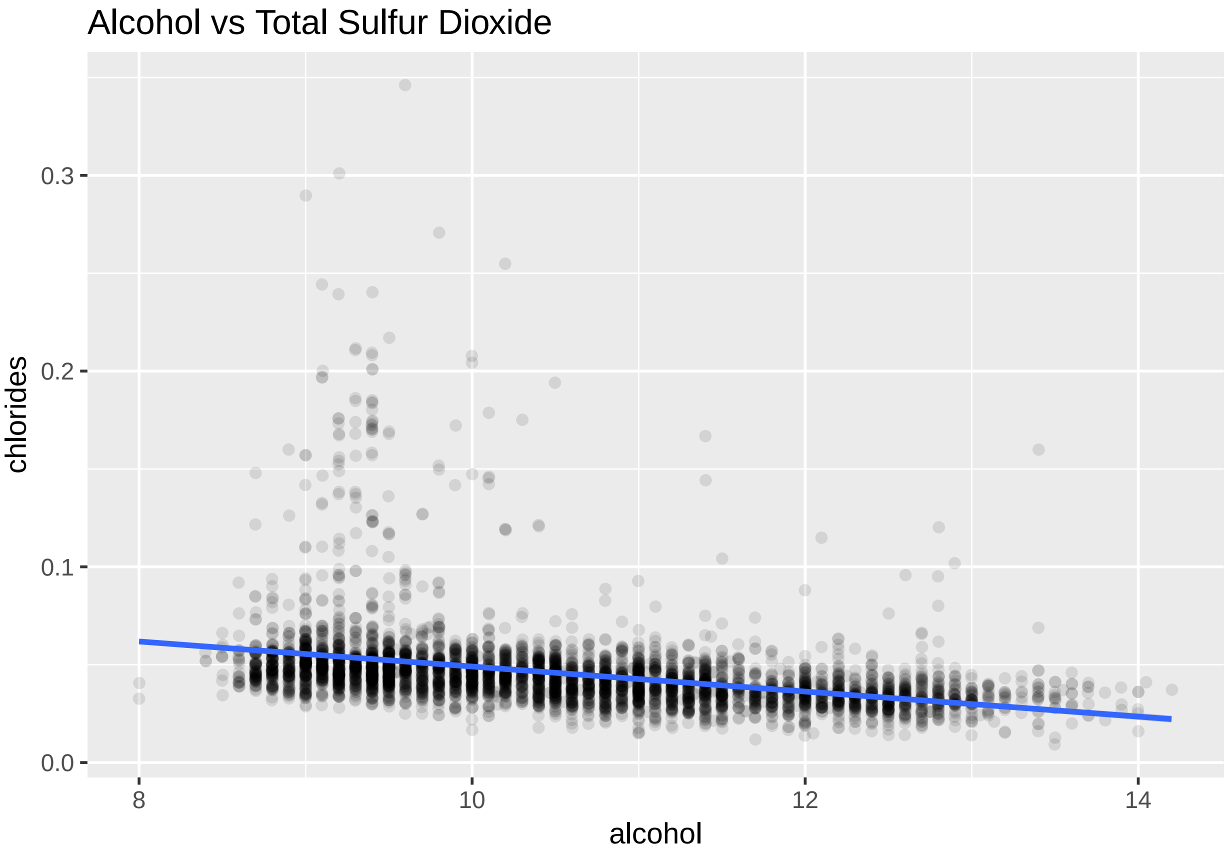


```
## Correlation between alcohol and total sulfur dioxide
cor(white_wine$alcohol, white_wine$total.sulfur.dioxide)
```

```
## [1] -0.4488921
```

We can see that the alcohol is negative correlated with the total sulfur dioxide by -0.45, which means the present of sulfur dioxide can somewhat reduce the alcohol content.

Alcohol vs. Chlorides

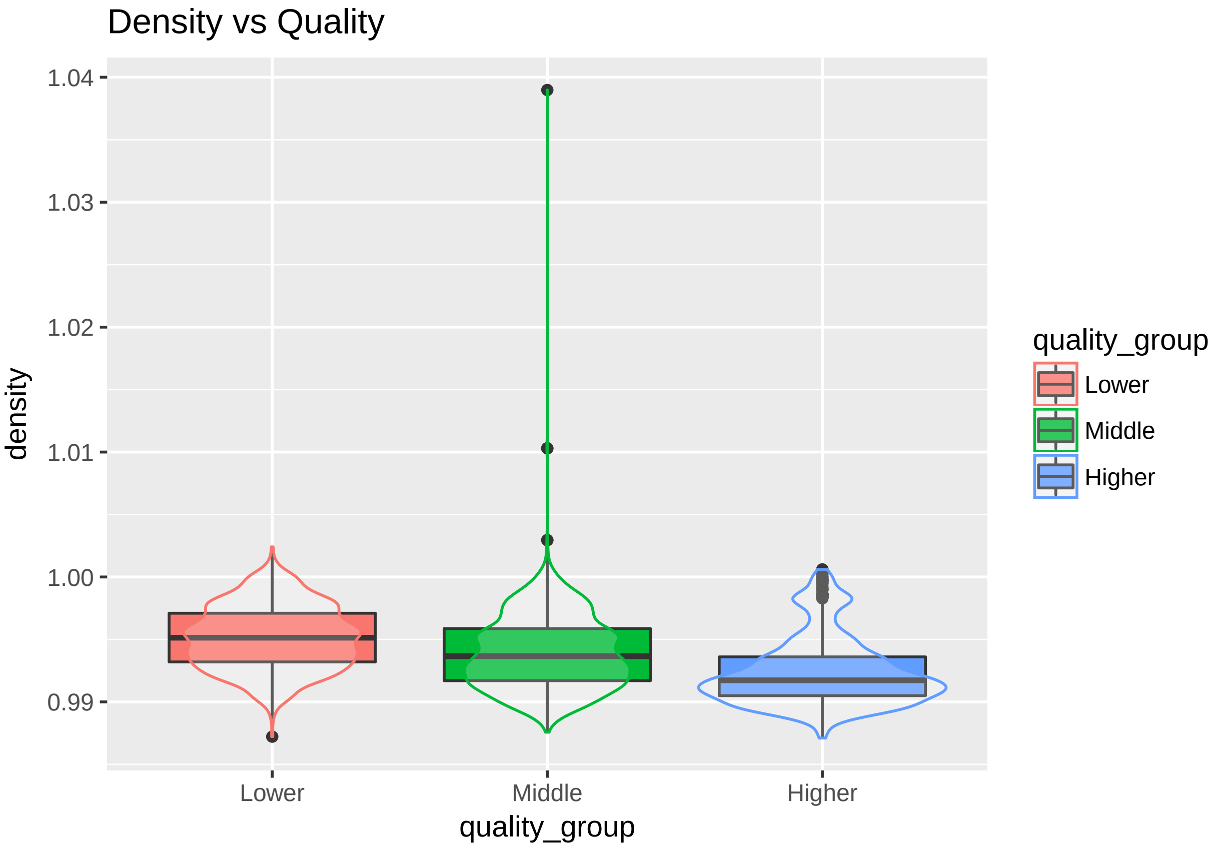


```
## Correlation between alcohol and chlorides
cor(white_wine$alcohol, white_wine$chlorides)
```

```
## [1] -0.3601887
```

Similar to total sulfur dioxide, the chlorides is negatively correlated with the alcohol content.

Density Vs. Quality



```
## Lower quality group
with(subset(white_wine, quality_group == "Lower"), summary(density))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.9872  0.9932  0.9951  0.9952  0.9971  1.0024
```

```
## Middle quality group
with(subset(white_wine, quality_group == "Middle"), summary(density))
```

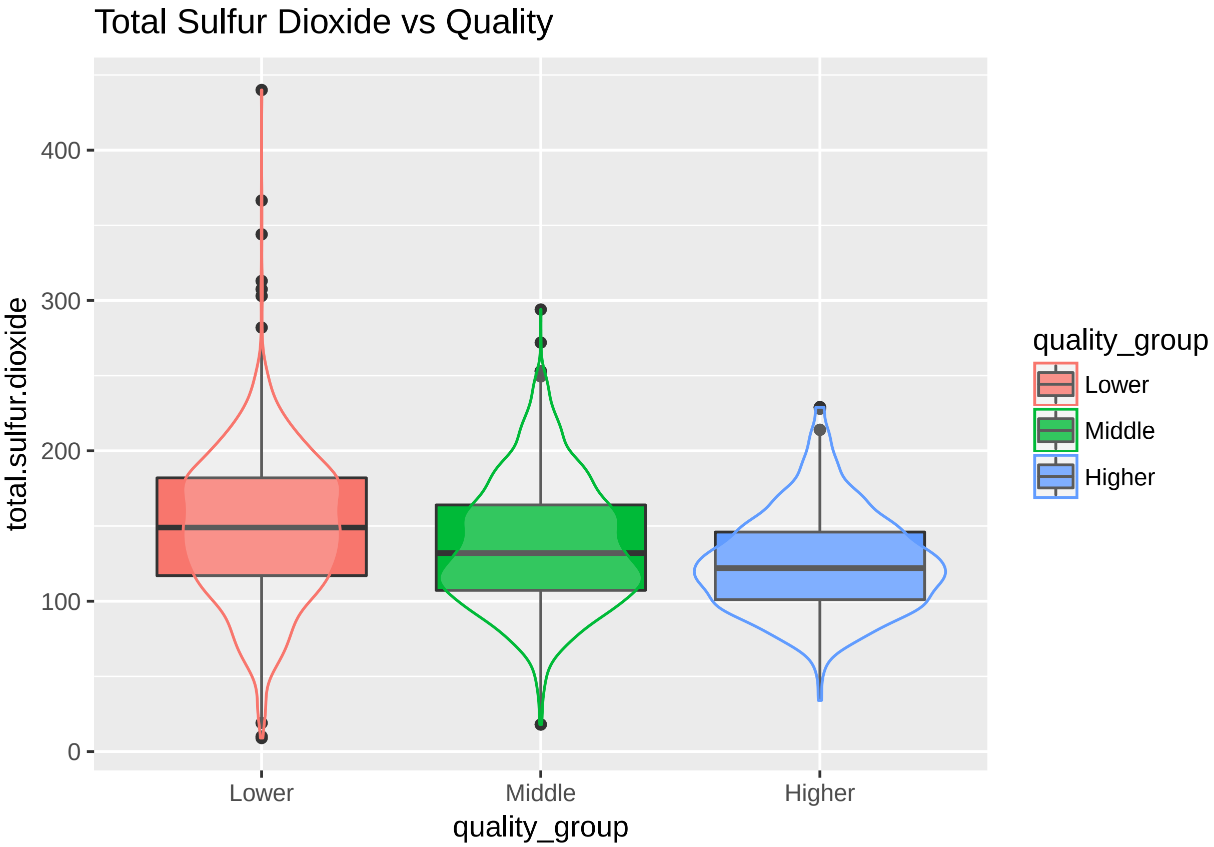
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.9876  0.9917  0.9937  0.9940  0.9959  1.0390
```

```
## Higher quality group
with(subset(white_wine, quality_group == "Higher"), summary(density))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.9871  0.9905  0.9917  0.9924  0.9936  1.0006
```

We can see that the density has a negative correlation with the quality of white wine. It makes sense as the density has a high correlation with alcohol.

Total Sulfur Dioxide vs. Quality



```
## Lower quality group
with(subset(white_wine, quality_group == "Lower"), summary(total.sulfur.dioxide))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.0   117.0   149.0   148.6   182.0   440.0
```

```
## Middle quality group
with(subset(white_wine, quality_group == "Middle"), summary(total.sulfur.dioxide))
```

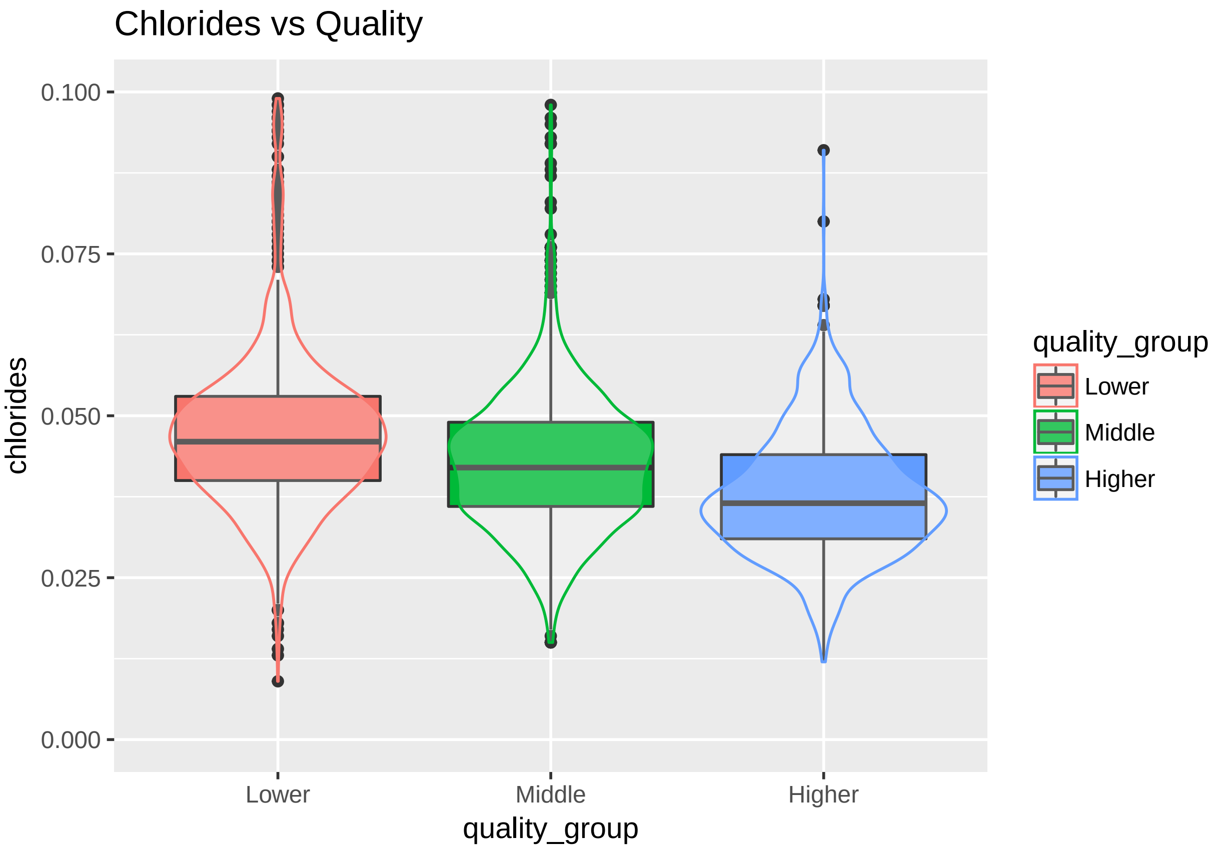
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    18.0   107.2   132.0   137.0   164.0   294.0
```

```
## Higher quality group
with(subset(white_wine, quality_group == "Higher"), summary(total.sulfur.dioxide))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    34.0   101.0   122.0   125.2   146.0   229.0
```

We can see that the total sulfur dioxide does have a effect on the quality score. The higher the total sulfur dioxide level, the white wine will less likely to get a high quality rating by wine experts. Although the correlation is not strong as the alcohol and density, we can still see the shift of total sulfur dioxide level in different quality group. I think sulfur dioxide might be a side product of one of the production process of white wine. According to the boxplot, there's no significant difference in total sulfur dioxide level across all the quality groups. This might be result from that the sulfur dioxide level is not high enough to affect the taste significantly.

Chlorides vs. Quality



```
## Lower quality group
with(subset(white_wine, quality_group == "Lower"), summary(chlorides))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00900 0.04000 0.04700 0.05144 0.05300 0.34600
```

```
## Middle quality group
with(subset(white_wine, quality_group == "Middle"), summary(chlorides))
```

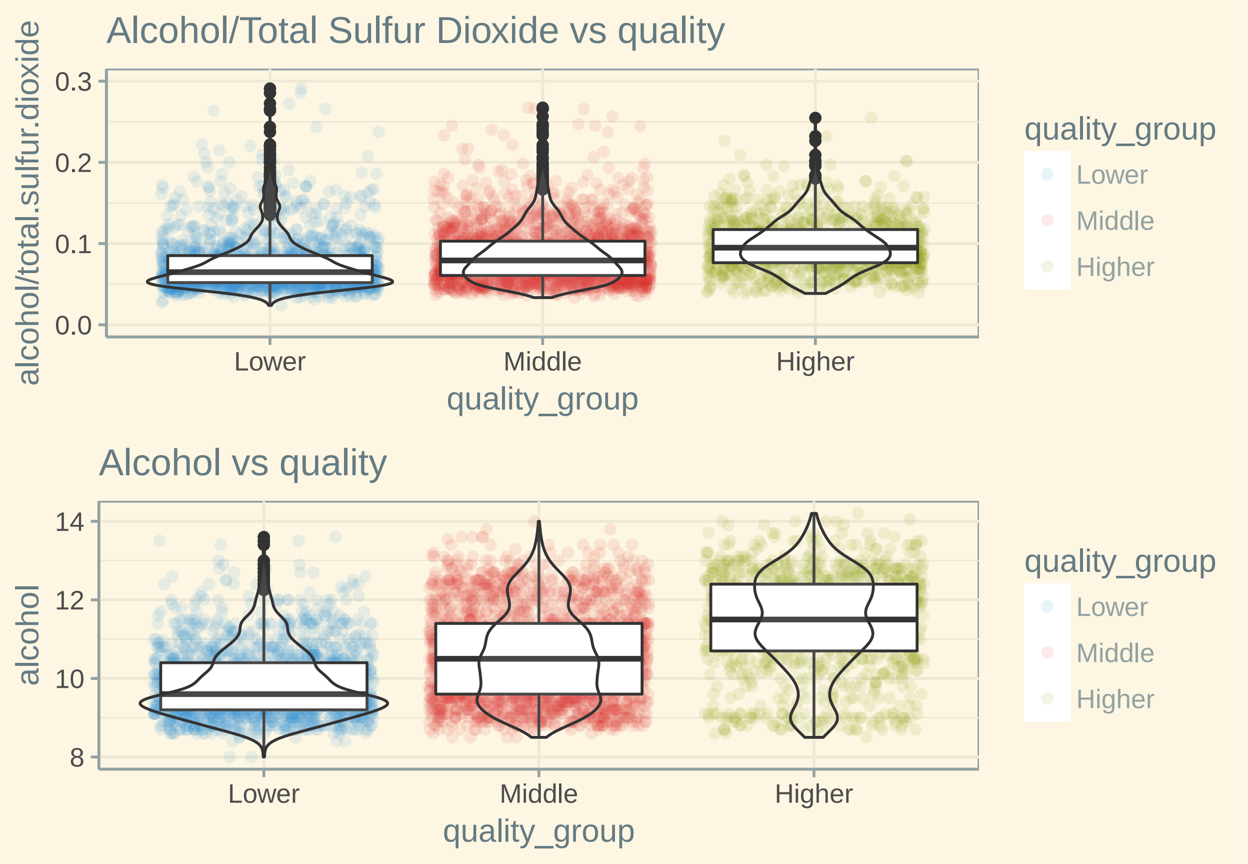
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01500 0.03600 0.04300 0.04522 0.04900 0.25500
```

```
## Higher quality group
with(subset(white_wine, quality_group == "Higher"), summary(chlorides))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.03100 0.03700 0.03816 0.04400 0.13500
```

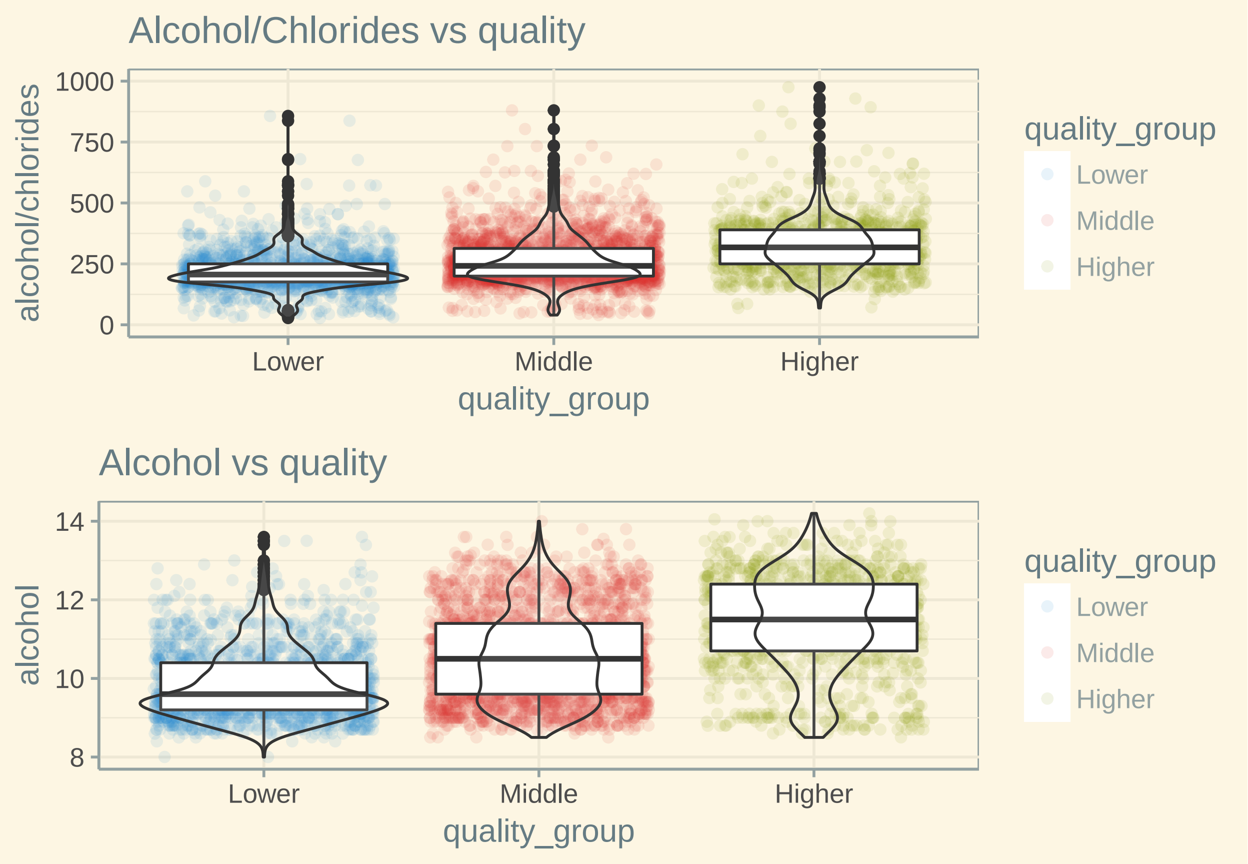
Although we can observe little difference in chlorides level across all the quality groups, we can see that the chlorides level is more concentrated in the lower range for higher quality group. Similar to sulfur dioxide, I think chlorides is a side product of one of the production process of white wine. The better white wine has better quality control, so that we can see there are few points out of the range of the boxplot.

Quality vs. Alcohol content and Total Sulfur Dioxide



By creating a facet plot, we can see the difference between the boxplot of alcohol/total sulfur dioxide and the boxplot measuring the actual alcohol content. In the original plot where we are comparing the alcohol content in each quality group, the data is more variant as we can see in the violin plot, the lower quality group has some observations that have high alcohol content but still received a low quality rating while the higher quality group also has a few observations that have low alcohol content. We can see that the plot looks better if we divide the alcohol content by total sulfur dioxide. The meaning of the plot then becomes the measurement of how much alcohol will there be if we have 1 unit of total sulfur dioxide. We can actually see the data points become more focused and less variant. In other words, the white wine will be purer. We will also get the same conclusion that the higher alcohol content is positively correlated with higher quality rating.

Quality vs. Alcohol content and Chlorides



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?

Alcohol positively correlated with the quality of white wine. The quality is also negatively affected by total sulfur dioxide, chlorides and density. Actually, except for alcohol content, other chemical attributes of white wine either have little positive correlation or negative correlation with the quality rating. There are two variables that have particular interest: total sulfur dioxide and chlorides. In my opinion, these two variables has a indirect indication on the quality control of white wine. We can see that the white wine with higher quality rating tend to have lower median and lower variance on these two variables from the boxplot. The presentation of these chemical residual is the result of some production processes. Taking a wild guess, according to the high variance of the level of these side products presented in the lower quality group, I think there could be other hidden variables outside of the dataset that affect the quality rating, because the side product reflects the quality control of the producer.

Additionally, in the exploration process I found that many high rating white wine can have low alcohol content and vice versa, so we need to include more factor to make the result more obvious. Thus, I involved the side product variables including total sulfur dioxide and chlorides to further explore the how the purity of white wine affects the quality rating. The result better demonstrated the effect of side product on the quality rating. It also shows that the higher alcohol content given the same total sulfur dioxide/chlorides level, the likelier the white wine receives higher rating.

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

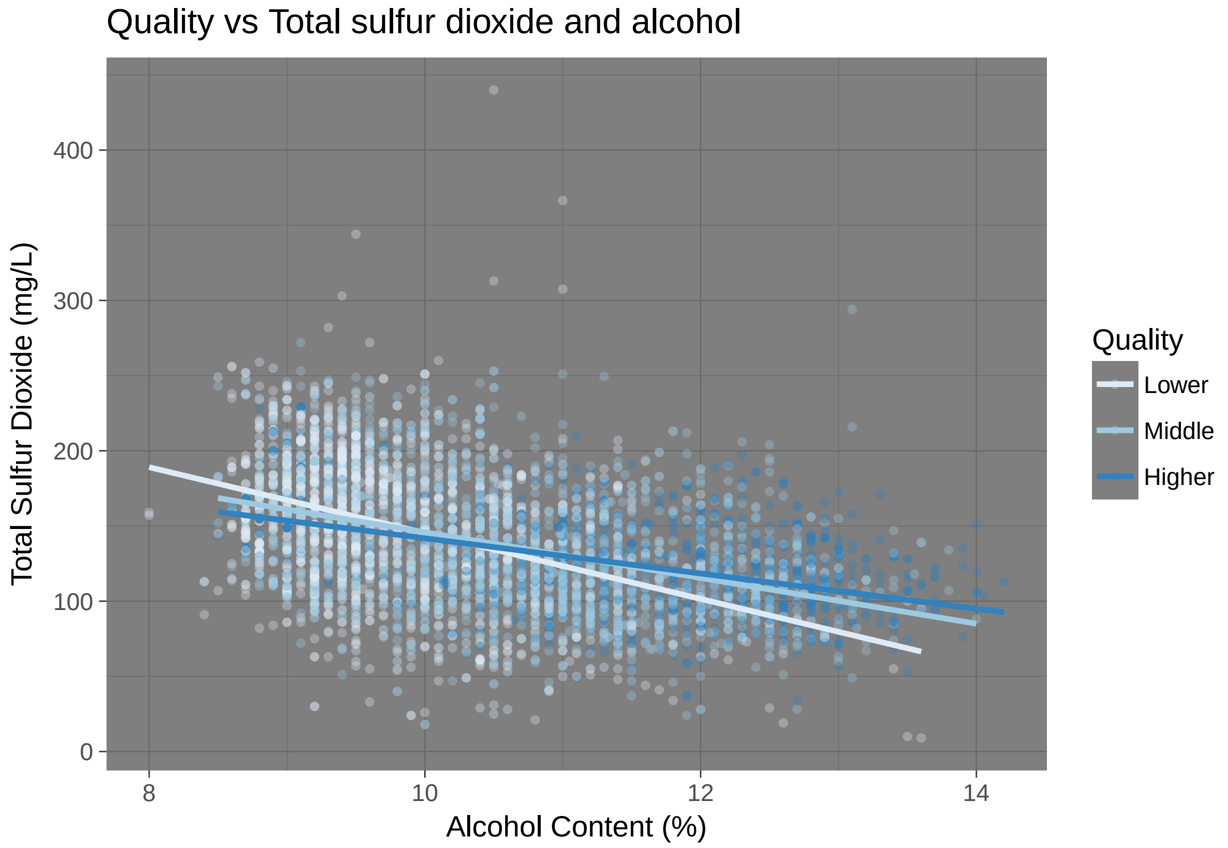
The total sulfur dioxide and chlorides not only negatively correlated to quality, but they also negatively correlated to alcohol content, as well. This interesting correlation has indirectly indicated that these factors reflects the effectiveness of production quality control. The alcohol content shows readabily increase with less of these chemical components presented in white wine.

What was the strongest relationship you found?

The strongeset relationship I found was density and residual sugar. The correlation between these two variables is 0.84, which is a strong correlation, which make prefect sense, as sugar has a density of 1.59 gram per cubic centimeter, while water has 1 gram per cubic centimeter. Sugar is more heavier than water given the same volume. So it is common sense that these two variables have the strongest correlation to each other. Similarly, alcohol content also has a strong negative correlation with the density of white wine, which is a result of the different in density of water and alcohol, too.

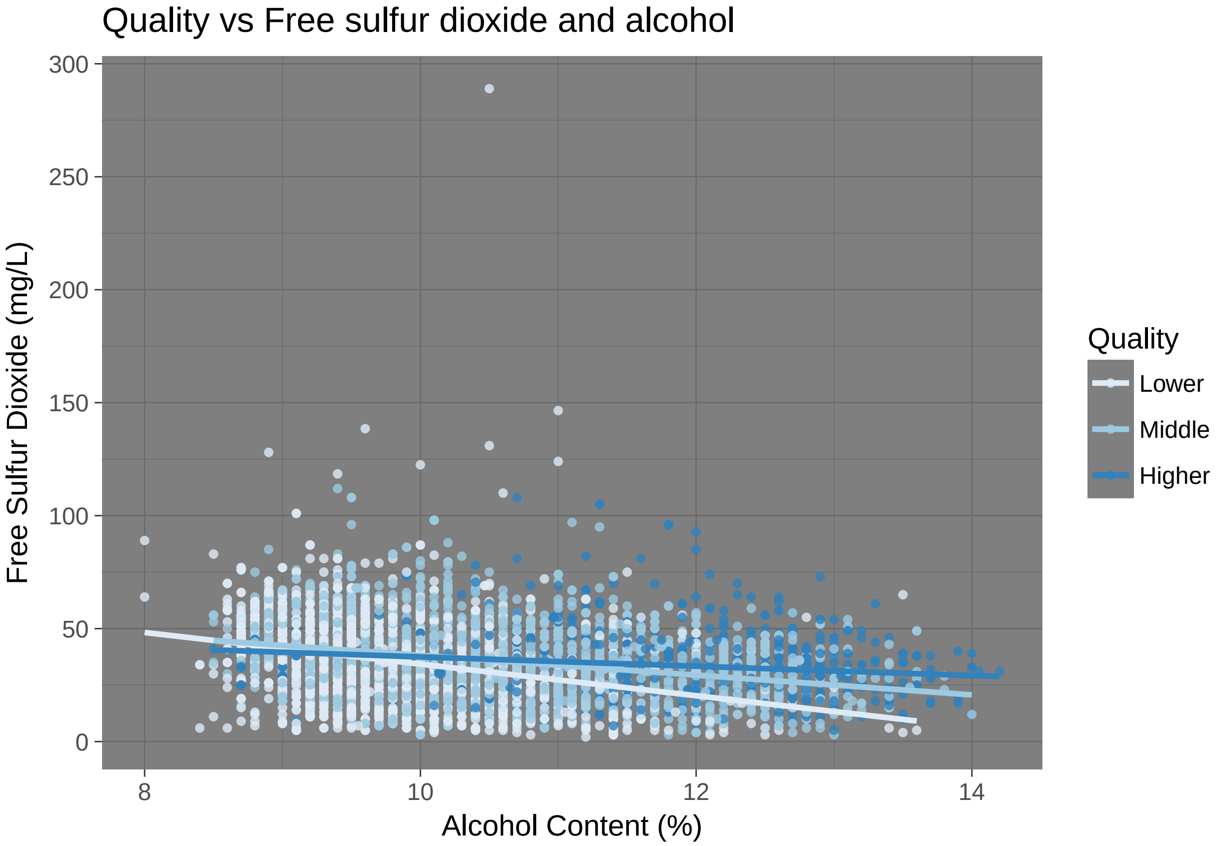
# Multivariate Plots Section

Quality vs Total Sulfur Dioxide and alcohol



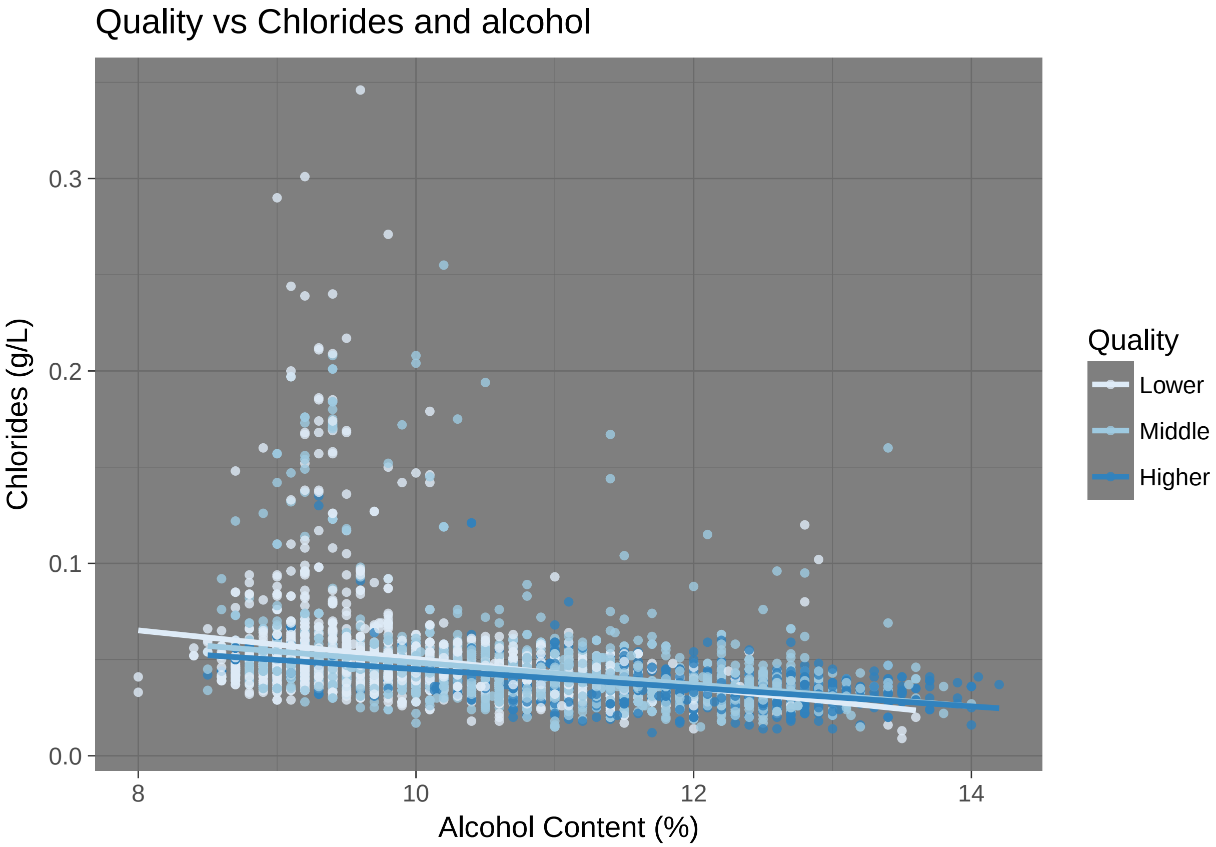
From the plot we can see that the higher quality group mostly has lower total sulfur dioxide level and higher alcohol content, which can be seen on both the linear model and the scatter plot. This interesting phenomenon reveals the fact that higher quality wine typically has better quality control thus it often associated with greater alcohol content and less total sulfur dioxide. I find it is even more interesting that such a minor difference can have a big influence on the taste of white wine. As we can see in the documentation, the sulfur dioxide would be evident to nose and taste when the free sulfur dioxide exceeds 50 mg/L. We will examine the relationship of free sulfur dioxide in the next plot.

Quality vs Free Sulfur Dioxide and alcohol



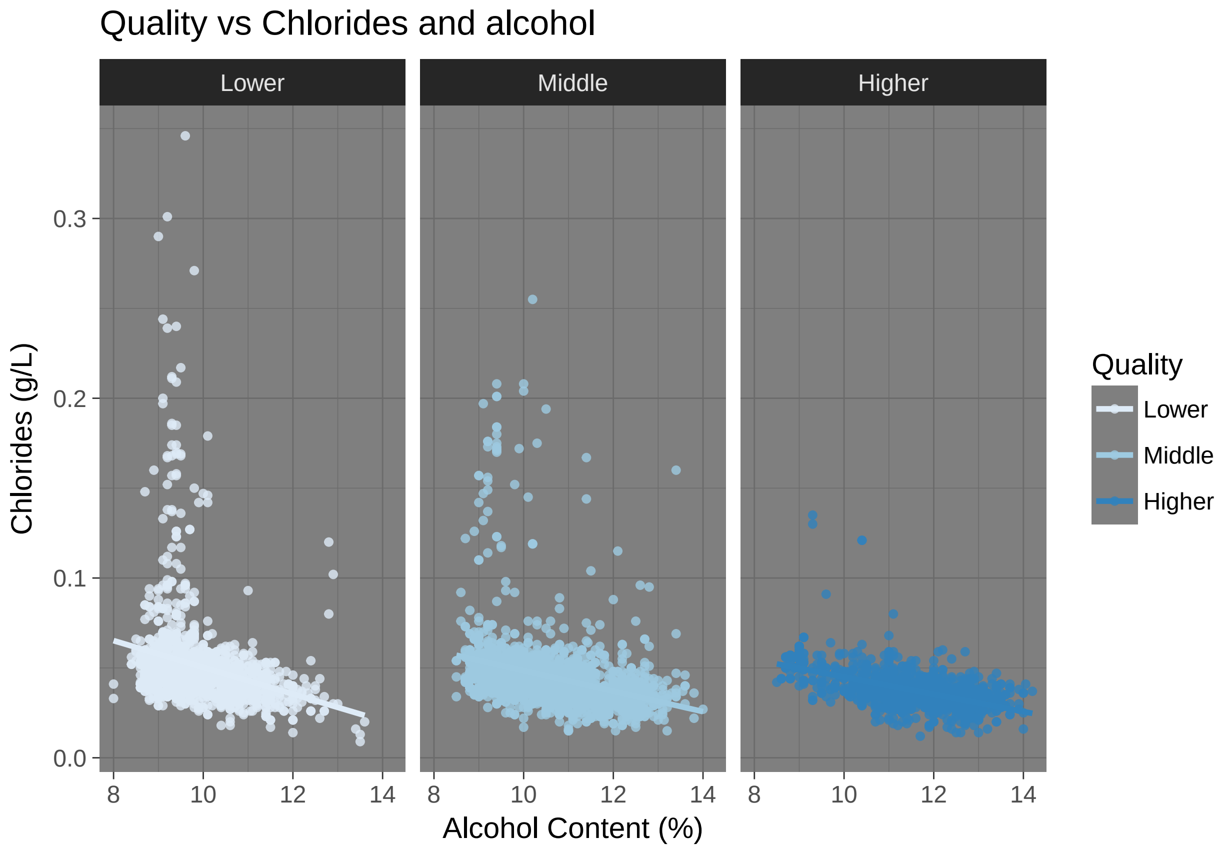
The above plot further reveals interesting phenomenon. As we can see in the plot, higher quality white wine does tend to have lower free sulfur dioxide overall, but there are lots of observations actually have free sulfur dioxide level that is higher that 50mg/L. Meanwhile there are also many observations from lower level quality group that have free sulfur dioxide lower than 50mg/L, as well. I think there might be human error or other factors not included in the dataset in effect. But overall, we can see the trend of quality increasing with lower free sulfur dioxide and higher alcohol content. The linear model suggests the same conclusion, too.

Quality vs Chlorides and alcohol



Furthermore, I want to take a look at the chlorides variable. As we can see in the scatter plot, the chlorides variable demonstrates similar pattern to sulfur dioxide: the higher quality group tend to have lower chlorides and higher alcohol content compare to lower quality group and middle quality group. But the plot also contains another interesting phenomenon: the lower quality and middle quality group tend to have higher variance in chlorides, while higher quality group stays in a lower level. My assumptions of seems to be strengthened by this analysis. There are many observations that have low level of residual chlorides, but they got a low rating regardlessly. The main difference in the lower quality group and higher quality group is that the lower quality group has much higher variance than the higher quality group. One of the possible explanation is that there is human error or ther are other hidden factors that are not included in the dataset reflect the quality control, which directly influence the taste of white wine.

To get a better sense of it, we can zoom in to see the difference:



As we can see in the facet plot, the variance in chlorides in the lower quality group is much higher that it in the higher quality group. Despite that some observations in both groups have the same alcohol content, the quality ratings are drastically different.

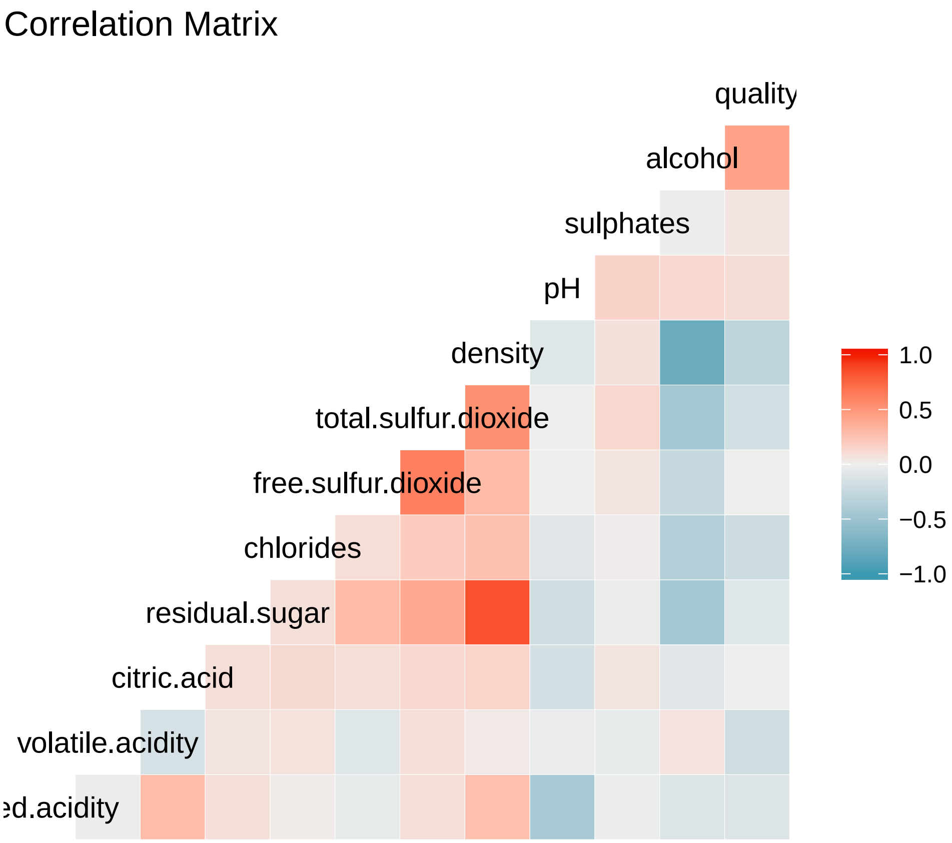
Multivariate Analysis

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 the multivariate plot section, I zoomed in to the "byproduct" variables including the total sulfur dioxide, free sulfur dioxide and chlorides. By further examining the correlation of variables, I found that my assumptions are strengthened. The byproduct is a reflection of quality control, and there are hidden variables influence the taste of wine not included in the dataset. It is fascinating to see the patterns in these plots. They all reveal interesting fact about the quality ratings of white wine. I think it will be more interesting if there's more data and more variable in this dataset.

Final Plots and Summary

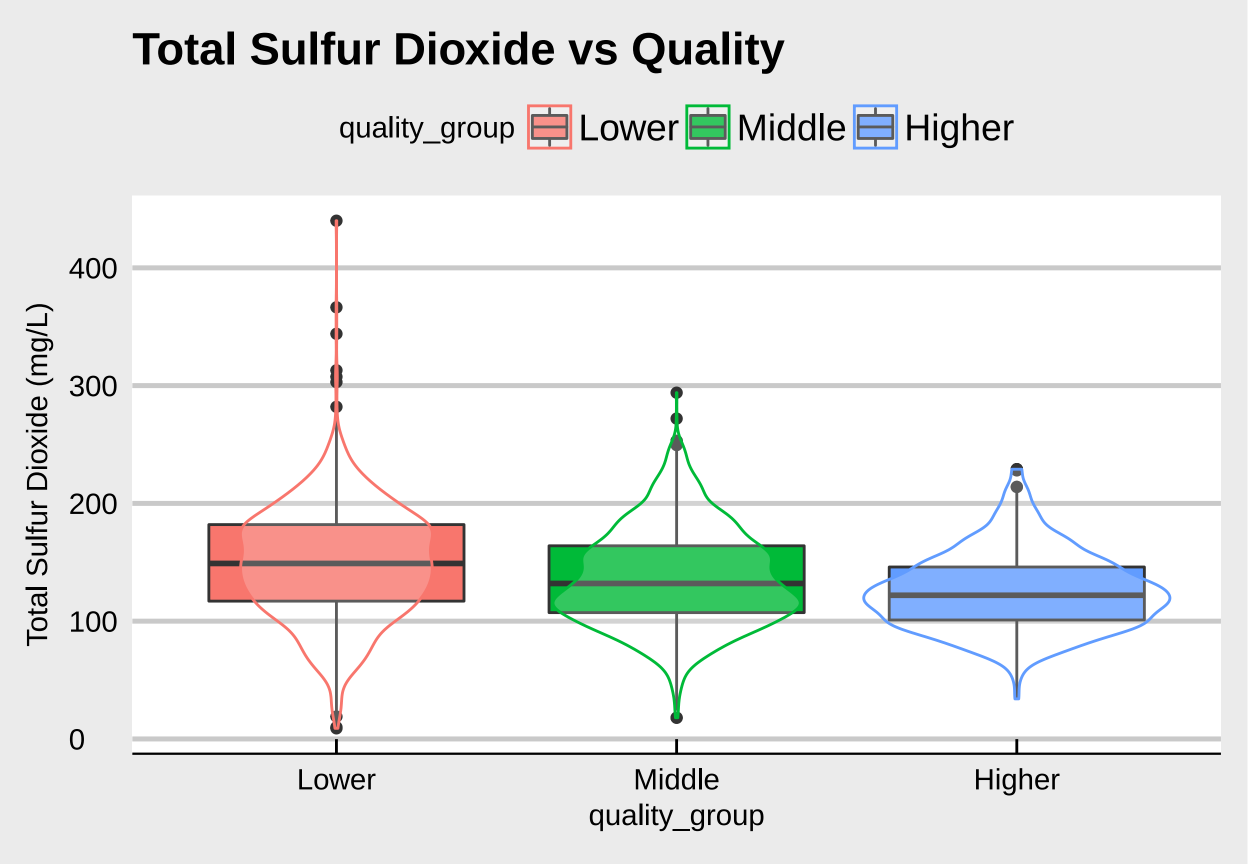
Plot One



Description One

This correlation matrix gives me the first glimpse on the relationship of every variable in this dataset. It uses color to demonstrate the correlation between each variable. This plot is very readable and it helped me determine which variable do I need to analyze. I can also see interesting facts from this correlation matrix, it turns out only few variables in this dataset have some level of influence on the quality rating, which include alcohol, density, total sulfur dioxide and chlorides. The dependence of each variable is clearly demonstrated on this plot, we can see that the density is highly correlated with alcohol and residual sugar, which makes perfect sense as it is common sense that the density of these chemical components is different from each other, and changing quantity of these components will affect the density of white wine accordingly.

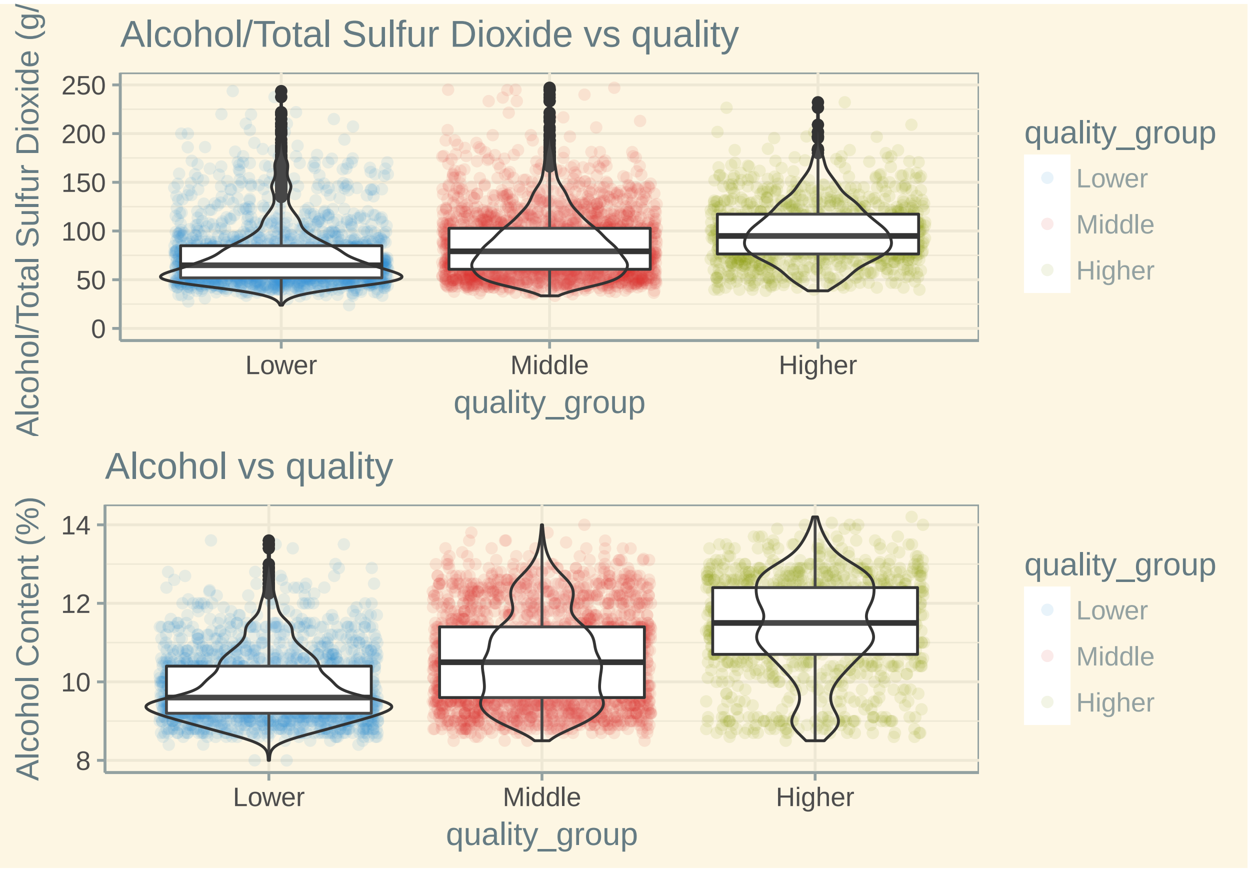
Plot Two



Description Two

I particularly like this plot because it reveals a interesting trend of variance. We can not only see in the plot that the higher quality group tend to have lower total sulfur dioxide in the white wine, but the variance also tend to be lower compare to the other two groups, and the variance shows a decreasing trend. This plot has inspired me to think more about the relationship of total sulfur dioxide and the quality control of white wine production process. Furthermore, the interesting pattern has also enabled me to think aboout how does the purity of alcohol affect the overall quality of white wine.

Plot Three



Description Three

This facet plot is also one of my favorite plot. Given that alcohol has positive correlation with quality rating, which total sulfur dioxide is negatively correlated with. I assume that the side products like sulfur dioxide are emerged from the production process, and higher alcohol content compare to the side products is a symbol of better quality control. I divided the alcohol content by total sulfur dioxide, which made the variance become lower and the data point and the violin plot were more focused. The plot reflects how the purity of alcohol content would affect the quality of white wine.

Reflection

By analyzing the white wine dataset, I found the following insights:

- The alcohol content has positive contribution to the quality of white wine.
- Side products like sulfur dioxide and chlorides has negative influence on the quality of white wine.
- The side product does not only reflect the actual by product in the production process, but they are also indicators for quality control in the production process, which may be demonstrated by other variables that are not included in the dataset.

- Other factors in the dataset such as residual sugar does not have strong affect on the quality of white wine.

I have also faced some challenges while completing this project. There's lack of categorical variables available in the dataset, and most variables are chemical attributes, which means I can hardly find a clue for suitable interaction between each variable. This confusion might be a result of lack of chemical knowledge or simply because I am not a wine expert. Despite that, I found the analysis went quite smoothly, and it is not always a requirement for data analysts/scientists to be an expert of the topic they research on. Asking the right question and communicating the right insight is much more important.

Another thing that I found challenging is the way of obtaining quality score. Due to the nature of median, there's a lot of 6 in the dataset, while the higher and lower scores are not comparable. To resolve this, I created new variable called quality group and divided them by three groups, which made the analysis much easier. After doing this, I found the trends are eaiser to be revealed by plots.