White Wine Quality - Exploratory Data Analysis

Mohamad Zeini Jahromi

April 14, 2017

Contents

Introduction

In this project, we explore and analyze a dataset related to white variants of the Portuguese "Vinho Verde" wine. There are many factors that affect the taste and quality of wine such as: alcohol content, acidity and pH level, sugar content, chlorides and etc.

In this dataset, the inputs (physicochemical properties) include objective tests (e.g. pH values) and the output (quality of wine) is based on sensory data (median of at least 3 evaluations made by wine experts). Each expert graded the wine quality between 0 (very bad) and 10 (very excellent). We will investigate the relationships between different physicochemical properties of wine and quality of wine.

Dataset

The following shows different variables and their format in our data set

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

There are 4898 observations, 11 input variables which includes physicochemical properties of white wine and an output variable which is wine quality.

Input variables (based on physicochemical tests) and their description are as follows:

1 - fixed acidity: most acids involved with wine or fixed or nonvolatile (do not evaporate readily) (tartaric acid - g / dmˆ3)

2 - volatile acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste (acetic acid - g / dmˆ3)

3 - citric acid: found in small quantities, citric acid can add 'freshness' and flavor to wines (g / dmˆ3)

4 - residual sugar: the amount of sugar remaining after fermentation stops (g / dmˆ3)

5 - chlorides: the amount of salt in the wine (sodium chloride - g / dmˆ3

6 - free sulfur dioxide: he free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion (mg / dmˆ3)

7 - total sulfur dioxide: amount of free and bound forms of S02 (mg / dmˆ3)

8 - density: the density of water is close to that of water depending on the percent alcohol and sugar content (g / cmˆ3)

9 - pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic)

10 - sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels (potassium sulphate - g / dm3)

11 - alcohol: the percent alcohol content of the wine (% by volume)

Output variable (based on sensory data):

12 - quality (score between 0 and 10)

Summary of Dataset

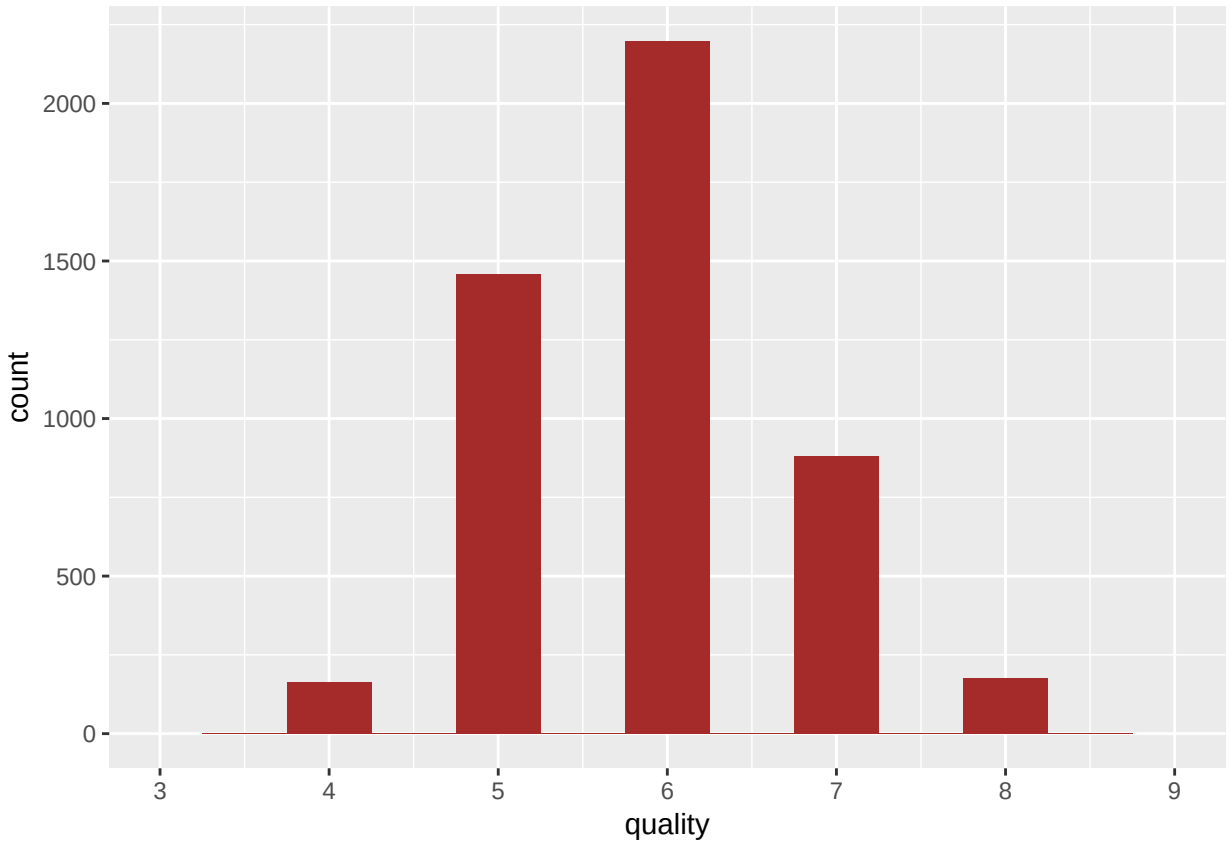
The following shows summary of our dataset.

```
##        X        fixed.acidity    volatile.acidity  citric.acid
##  Min.   :   1   Min.   : 3.800   Min.   :0.0800   Min.   :0.0000
##  1st Qu.:1225   1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700
##  Median :2450   Median : 6.800   Median :0.2600   Median :0.3200
##  Mean   :2450   Mean   : 6.855   Mean   :0.2782   Mean   :0.3342
##  3rd Qu.:3674   3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900
##  Max.   :4898   Max.   :14.200   Max.   :1.1000   Max.   :1.6600
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.600   Min.   :0.00900   Min.   :  2.00
##  1st Qu.: 1.700   1st Qu.:0.03600   1st Qu.: 23.00
##  Median : 5.200   Median :0.04300   Median : 34.00
##  Mean   : 6.391   Mean   :0.04577   Mean   : 35.31
##  3rd Qu.: 9.900   3rd Qu.:0.05000   3rd Qu.: 46.00
##  Max.   :65.800   Max.   :0.34600   Max.   :289.00
##  total.sulfur.dioxide    density             pH          sulphates
```

```
##  Min.   :  9.0        Min.   :0.9871   Min.   :2.720   Min.   :0.2200
##  1st Qu.:108.0        1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100
##  Median :134.0        Median :0.9937   Median :3.180   Median :0.4700
##  Mean   :138.4        Mean   :0.9940   Mean   :3.188   Mean   :0.4898
##  3rd Qu.:167.0        3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500
##  Max.   :440.0        Max.   :1.0390   Max.   :3.820   Max.   :1.0800
##     alcohol         quality
##  Min.   : 8.00   Min.   :3.000
##  1st Qu.: 9.50   1st Qu.:5.000
##  Median :10.40   Median :6.000
##  Mean   :10.51   Mean   :5.878
##  3rd Qu.:11.40   3rd Qu.:6.000
##  Max.   :14.20   Max.   :9.000
```

We can see, the average alcohol percentage in our dataset is about 10.51 with most of wines range between 9.5 and 11.4 percent. The fixed acidity ranges between 6.3 to 7.3 g/dmˆ3 The median of wine quality is 6 with most of wines fall between grade 5 and 6.

Distribution of Quality of Wine

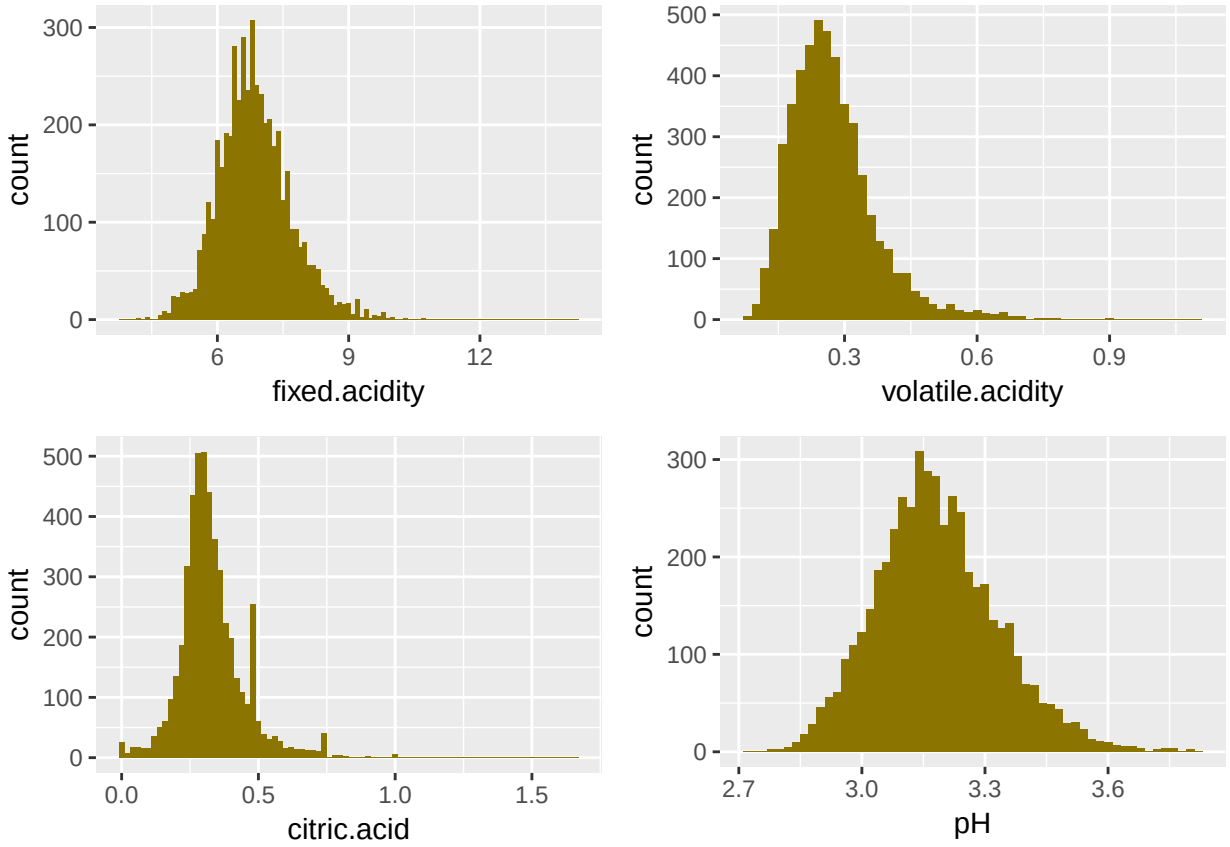


The quality of most of wines, fall between 5 and 7 and few of them have quality above 8 or below 4.

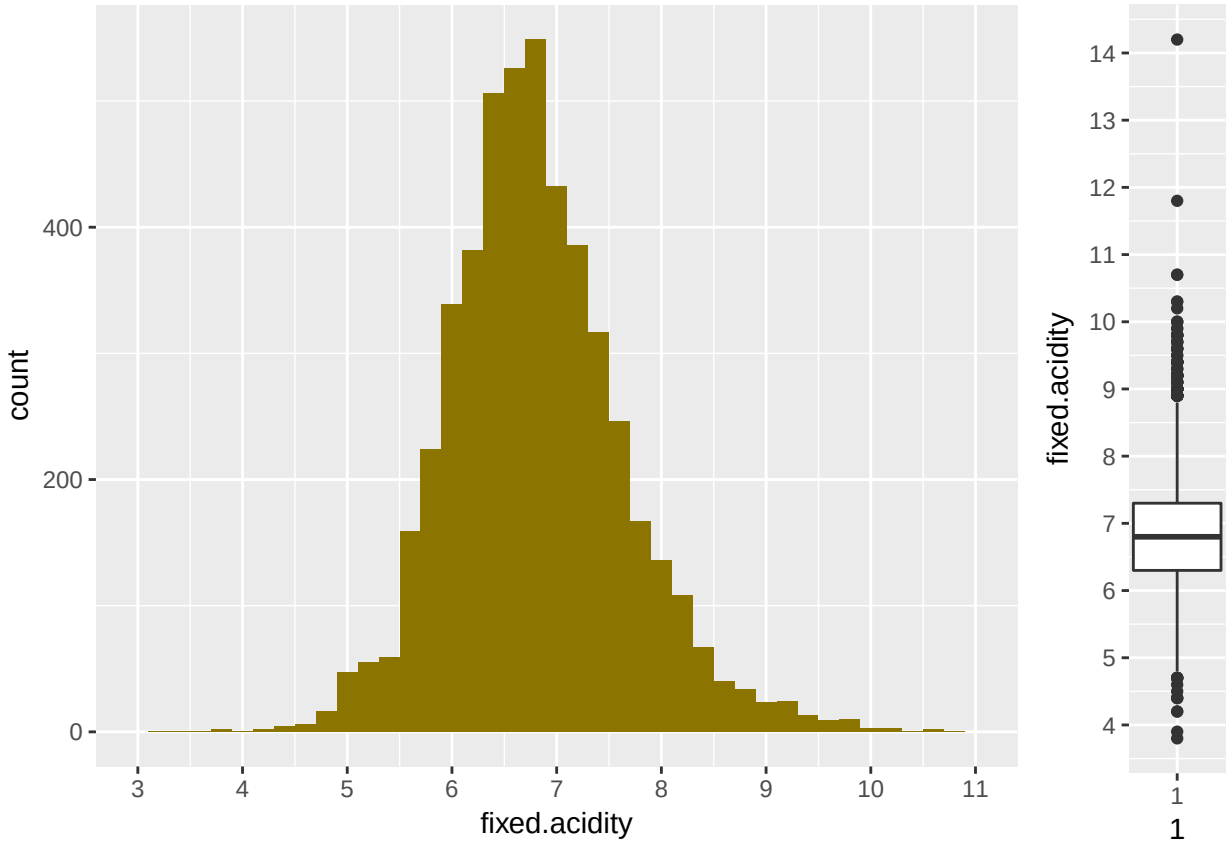
Univariate Plots - Distribution of Physicochemical Properties of Wine

We put variables in three groups based on the nature of their Physicochemical Properties. First group includes acidic components of wine and resulting pH (which is a measure of acidity). In the second group, we have all sulfur component of wine (as a solution or dissolved gas) and chlorides. Third group includes density, residual sugar and alcohol content.

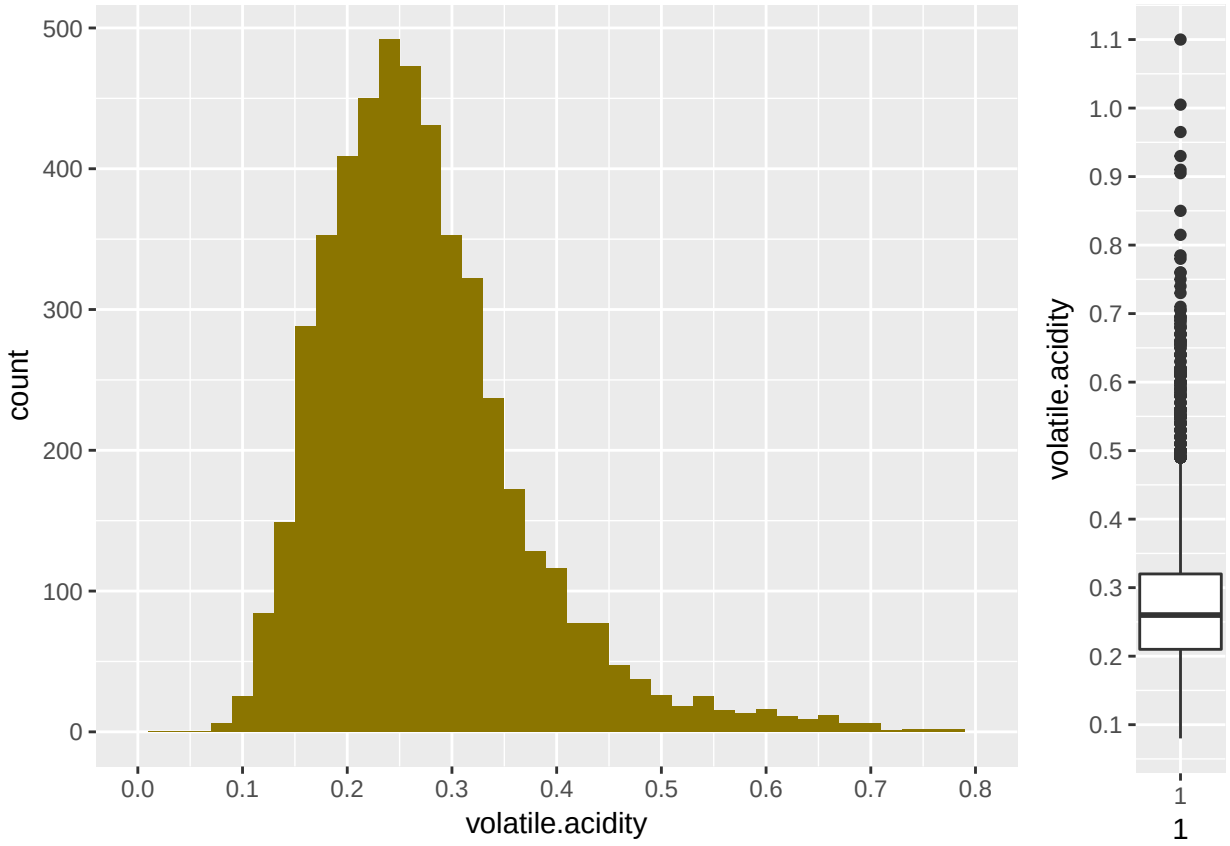
Acidity and pH



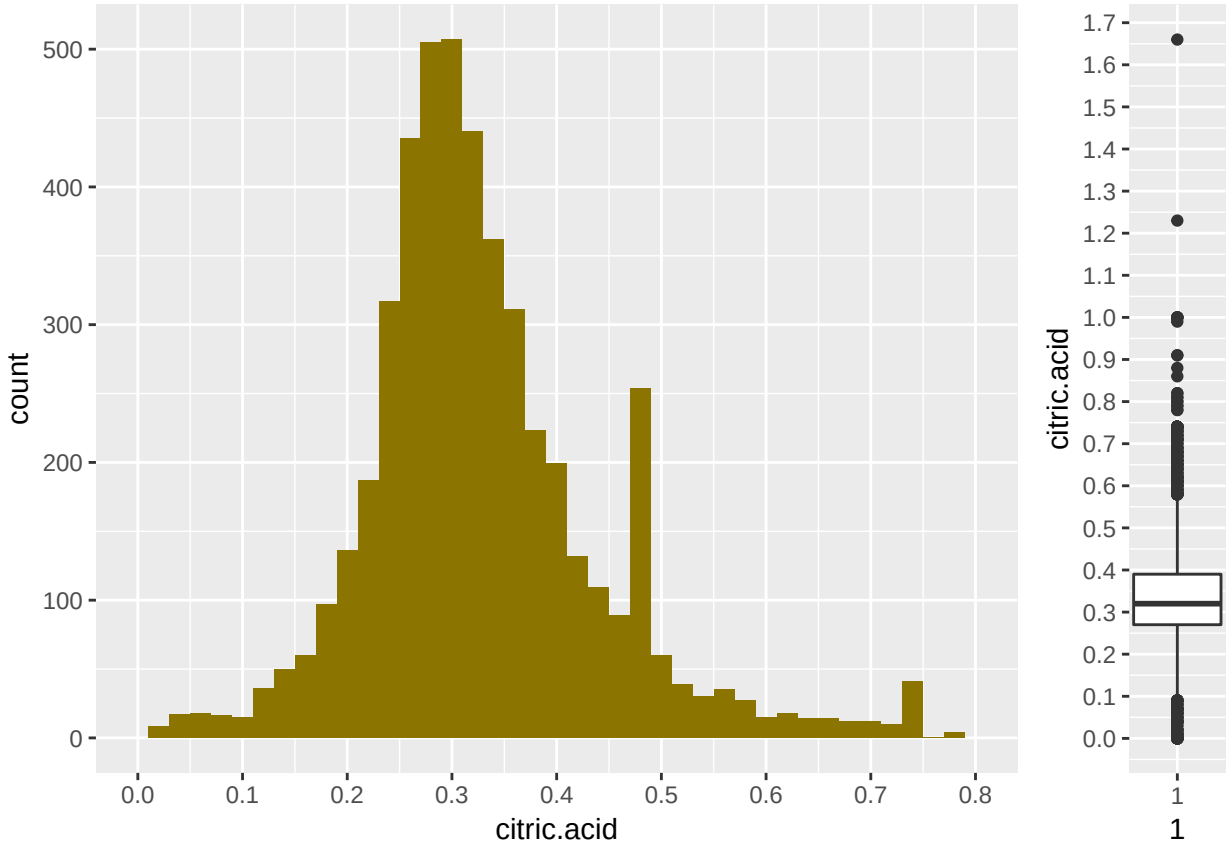
The fixed acidity, citric acid and pH plots show normal distributions but volatile acidity is slightly right skewed. Let's remove outliers and look at these plots one by one.



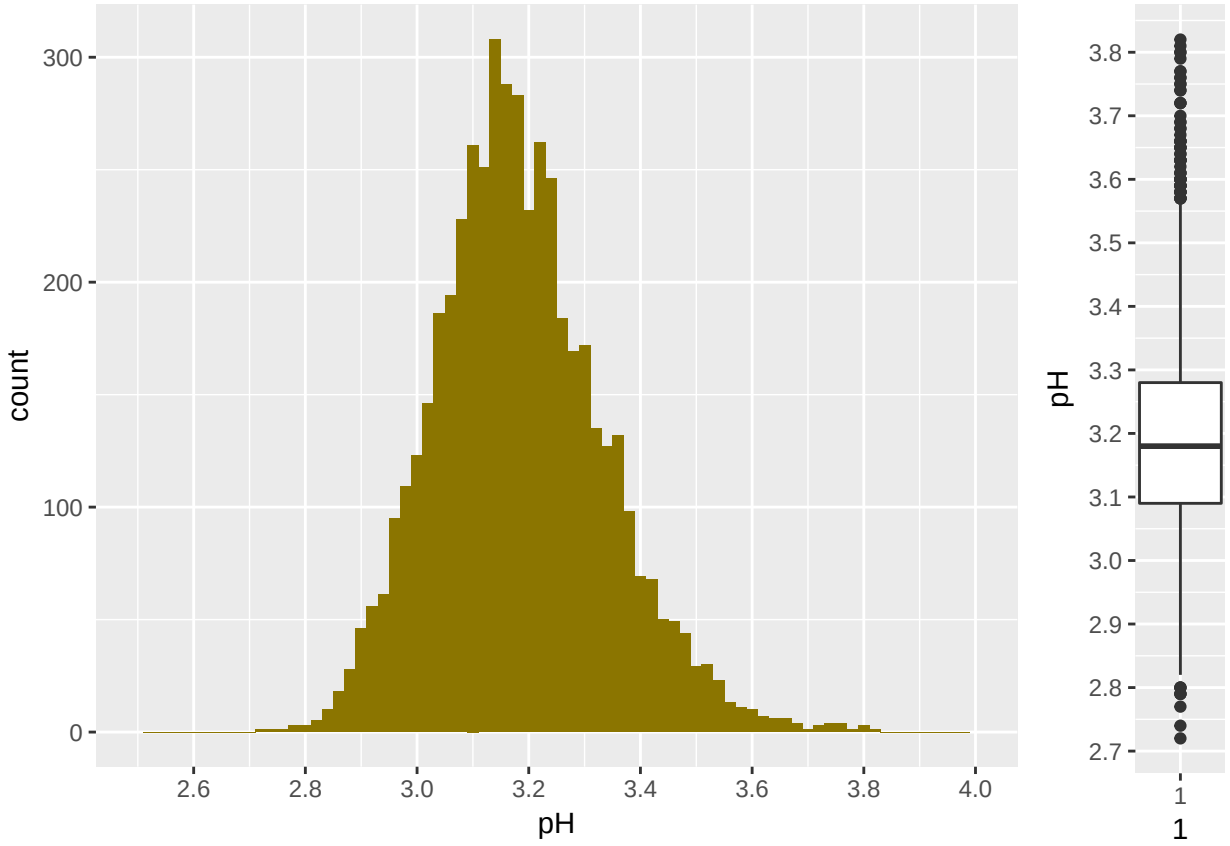
Majority of white wines fixed acidity (nonvolatile acids like tartaric acid) are between 6.3 and 7.3 g/dmˆ3. The mean and median are around 6.8 g/dmˆ3 which shows outliers has a little effect on central values. The right side boxplot shows the range of outliers.



The 1st and 3rd quartiles of volatile acidity (acetic acid) distribution are 0.21 and 0.23 g/dmˆ3 and the mean and median are around 0.28 and 0.26 g/dmˆ3. The distribution is slightly right skewed which shows outliers effect on the mean value increase. The right side boxplot shows the range of outliers.

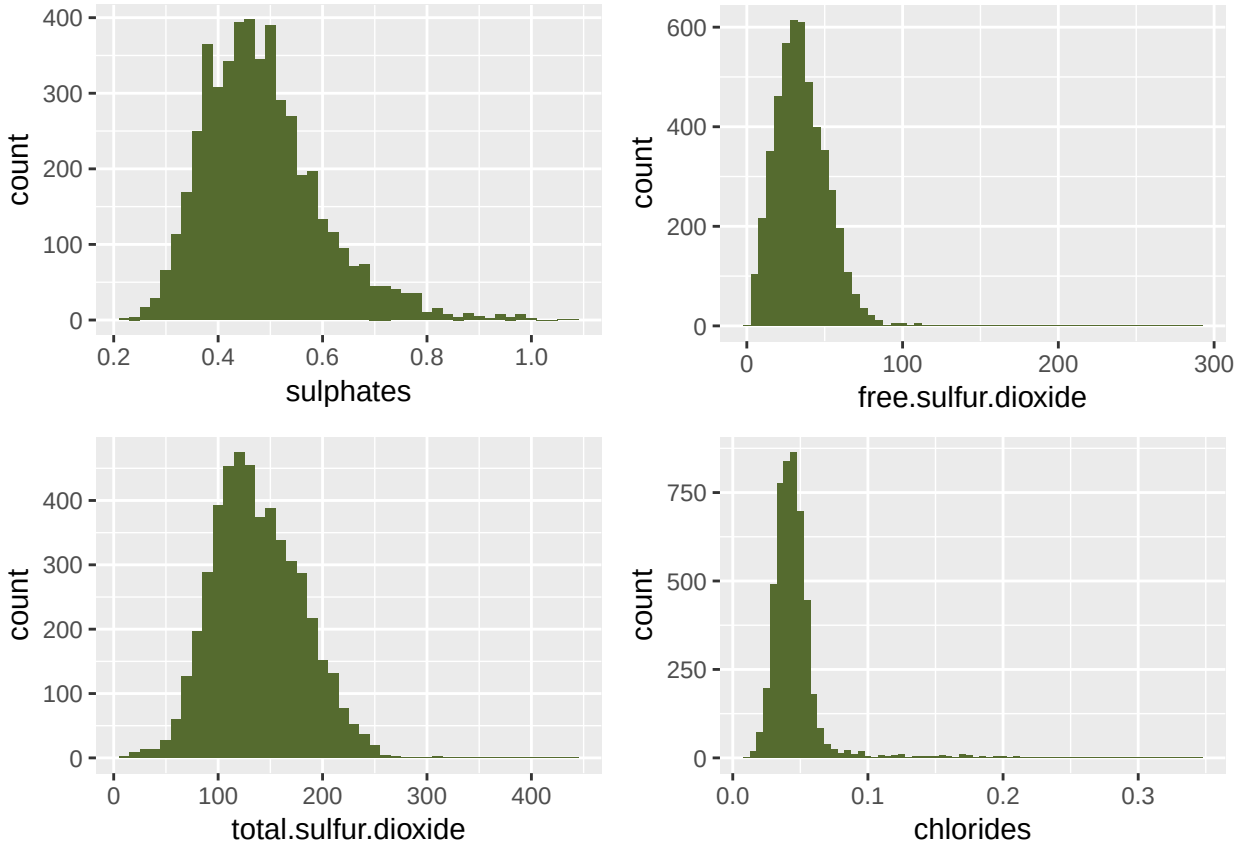


Citric acid add 'freshness' and flavor to wines and its quantity in most of white wines are between 0.27 and 0.39 g/dmˆ3. The mean and median are around 0.33 g/dmˆ3 which shows outliers has a little effect on central values. There is an unexpected spike around 4.9 g/dmˆ3. Since the corresponding pH plot shows no sign of sharp increase in acidity level, the spike might represent inaccurate data. The right side boxplot shows the range of outliers.

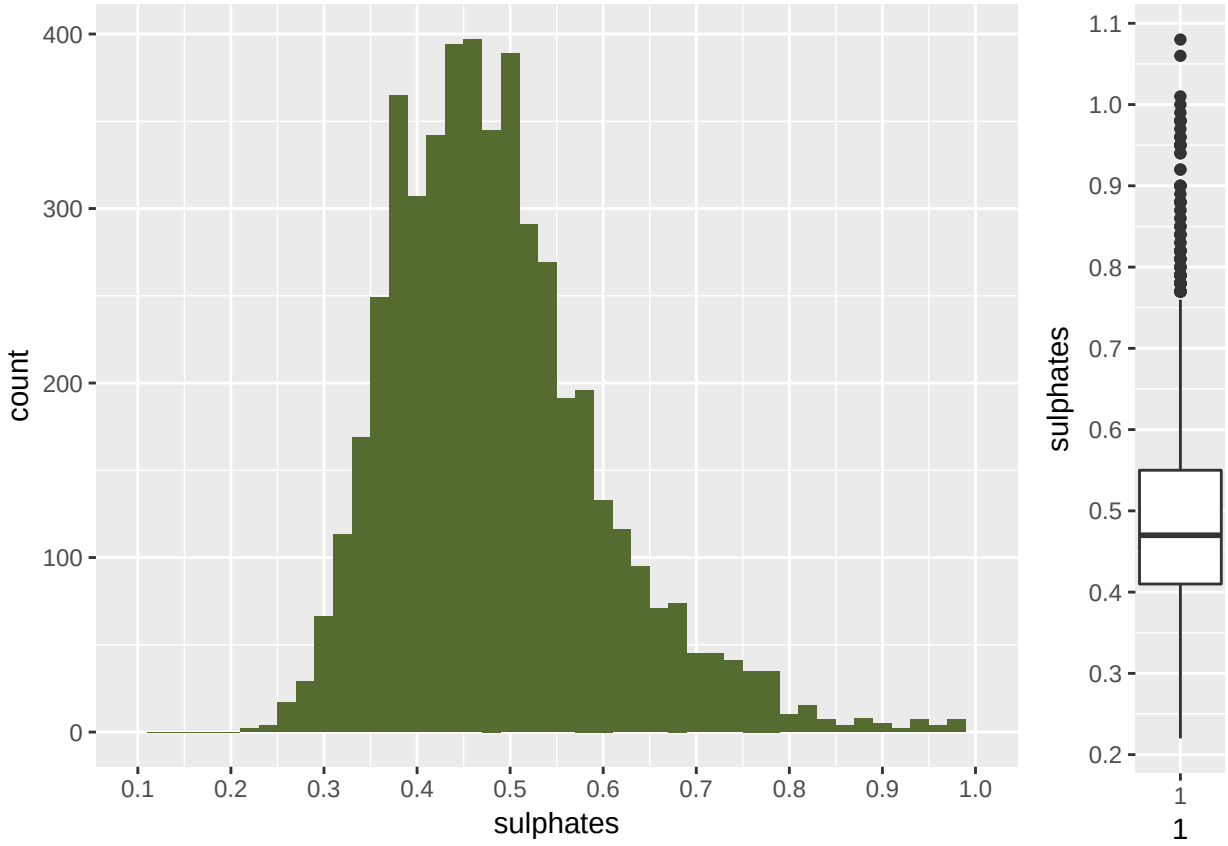


Above figure shows normal distribution of wine pH. Since all pH values are below 3.8, we can say all the wines are acidic in nature. Majority of pH values are between 3.1 to 3.3 with an average and median of 3.2. The wine fixed acidity and citric acidity also show the same behavior as pH but the volatile acidity distribution is slightly right skewed. The right side boxplot shows the range of outliers.

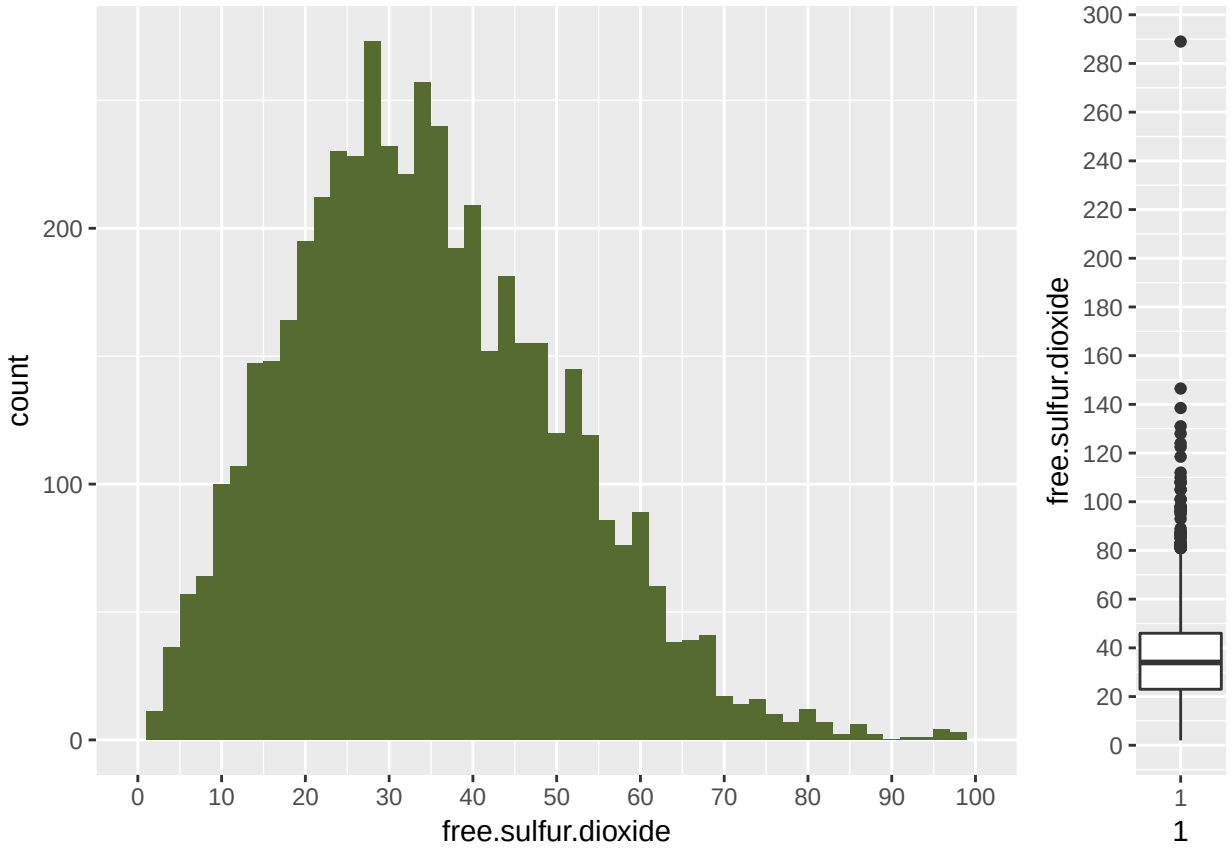
Sulfur and Chlorides



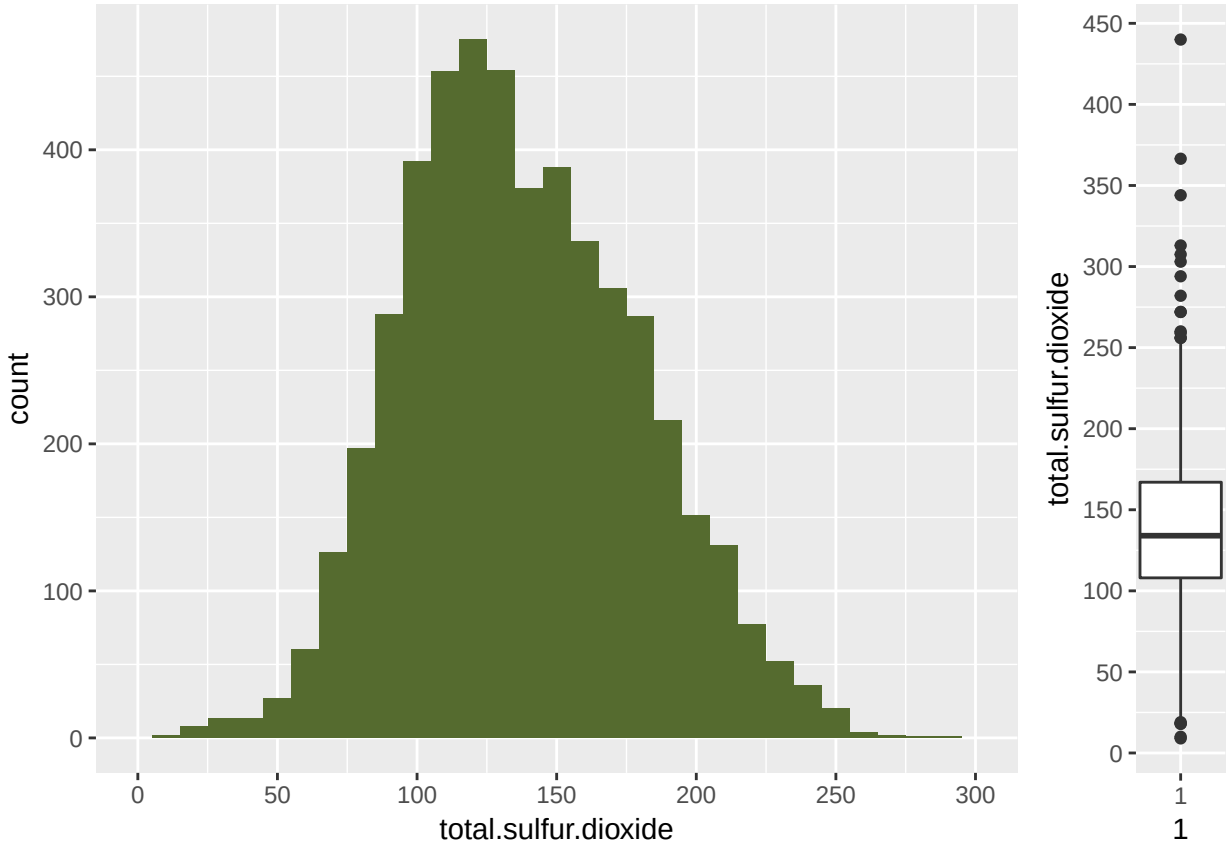
It looks like that sulfates, free sulfur dioxide and total sulfur dioxide have right skewed distributions and chlorides have a normal distribution with a wide range of outliers. Let's explore each plot more carefully.



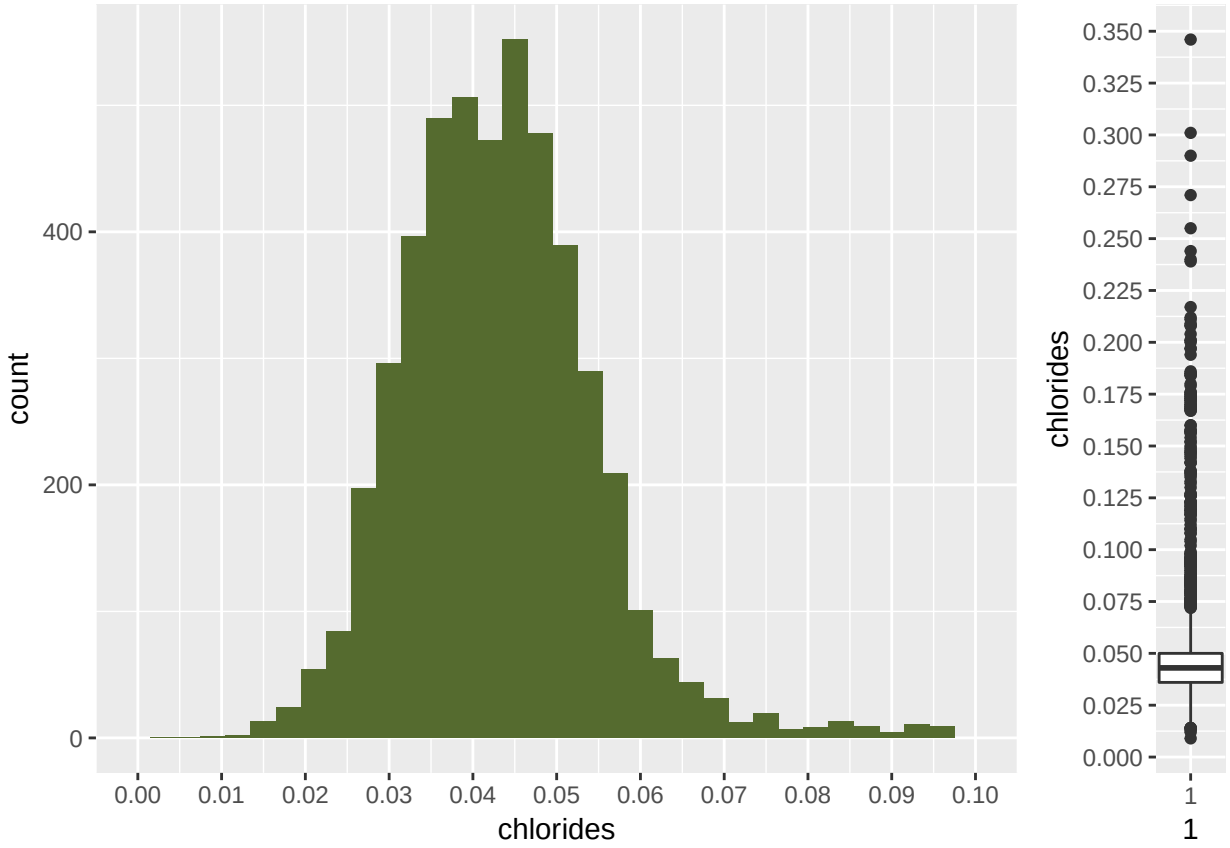
The sulfates (potassium sulphate) distributions is slightly right skewed with majority of values between 0.41 and 0.55 g/dmˆ3. The mean and median are 0.49 and 0.47 g/dmˆ3 that shows outliers effect on the mean value increase. The right side boxplot shows the range of outliers.



The free sulfur dioxide distribution is similar to the sulfates distributions and right skewed with majority of values between 23 and 46 mg/dmˆ3. The mean and median are 35 and 34 mg/dmˆ3 that shows outliers has a little effect on the mean value. The right side boxplot shows the range of outliers.

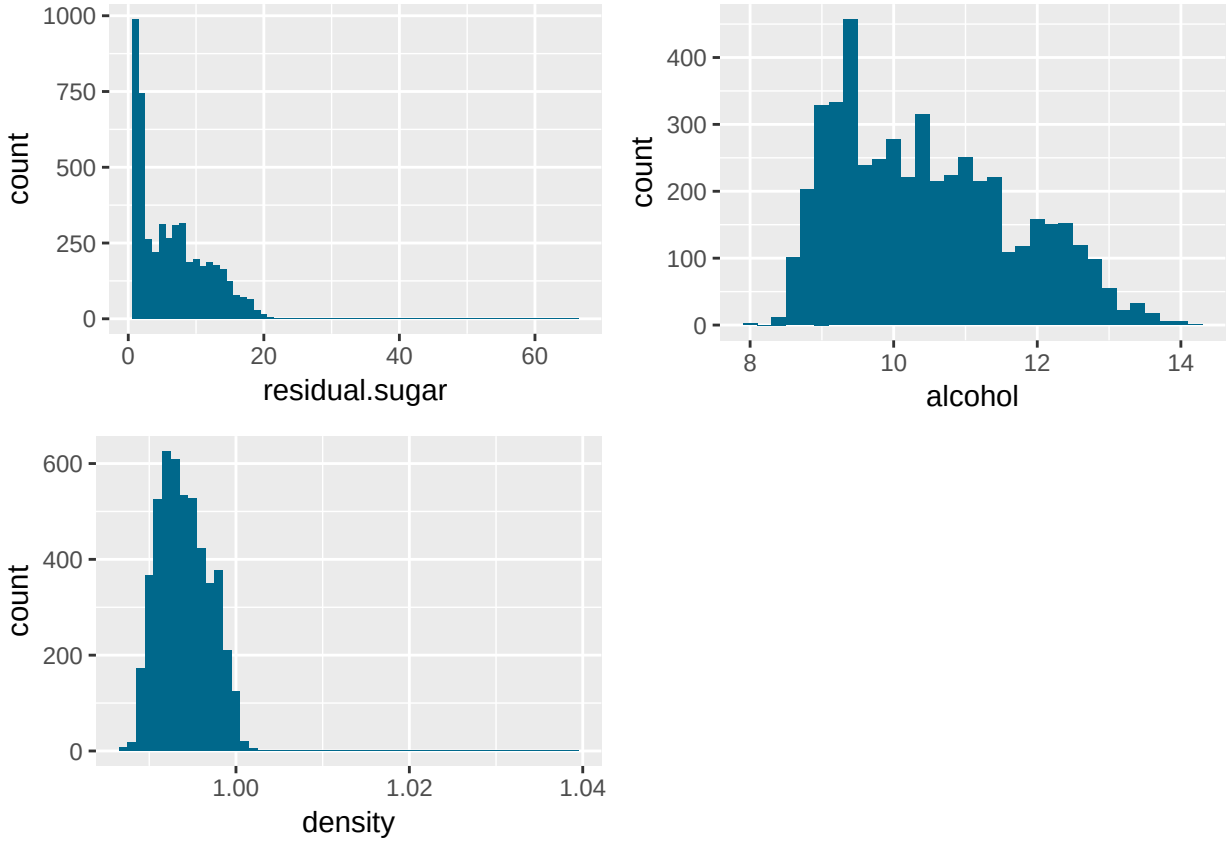


Comparing to the sulfates and free sulfur dioxide distributions, the total sulfur dioxide is less skewed. Most of wines have a total sulfur dioxide between 108 and 167 mg/dmˆ3. The mean and median are around 135 mg/dmˆ3 that shows outliers effect are insignificant. The right side boxplot shows the range of outliers.

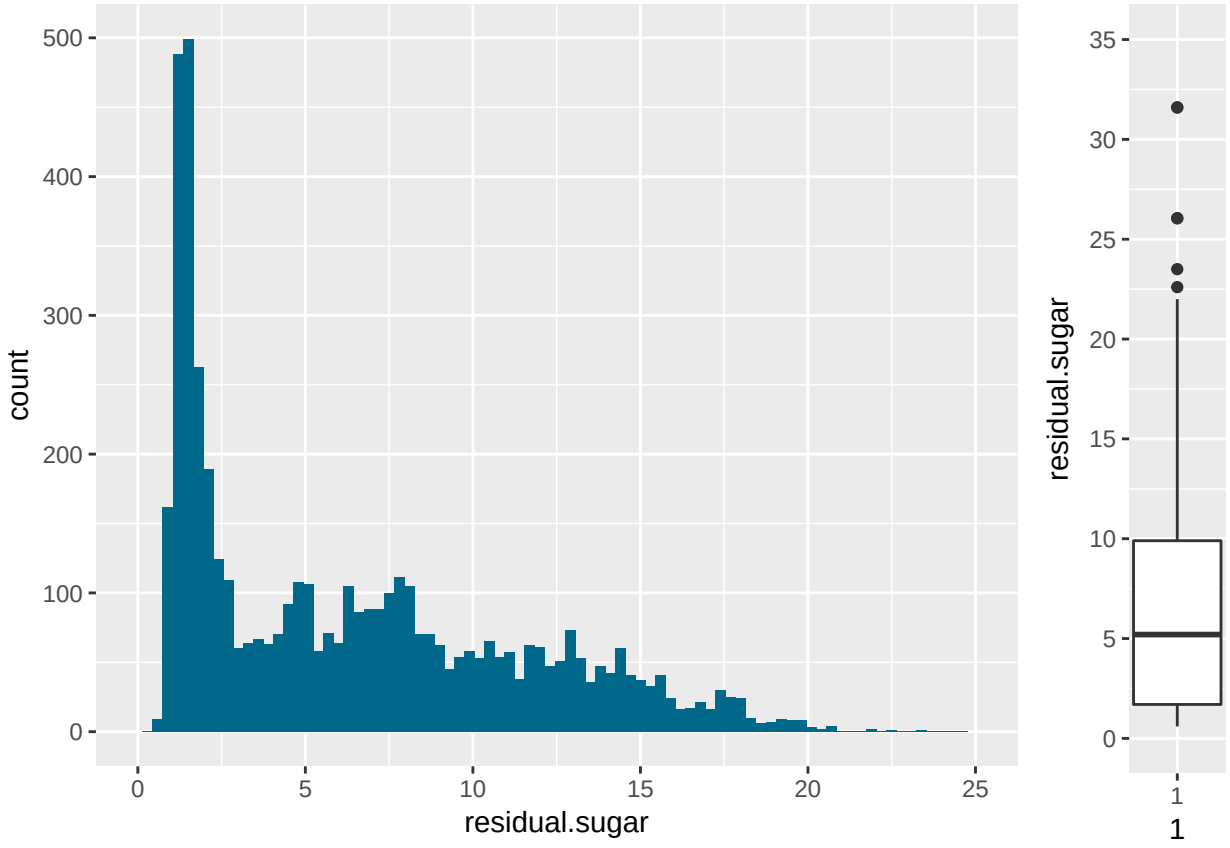


Finally, above plot shows that the chlorides have a normal distribution with a wide range of outliers. Let's explore each plot more carefully. The 1st and 3rd quartiles of distribution are 0.036 and 0.050 g/dmˆ3 and the mean and median values are 0.043 and 0.045 g/dmˆ3. The outliers pushed the mean value to the right side of median. The right side boxplot shows the range of outliers.

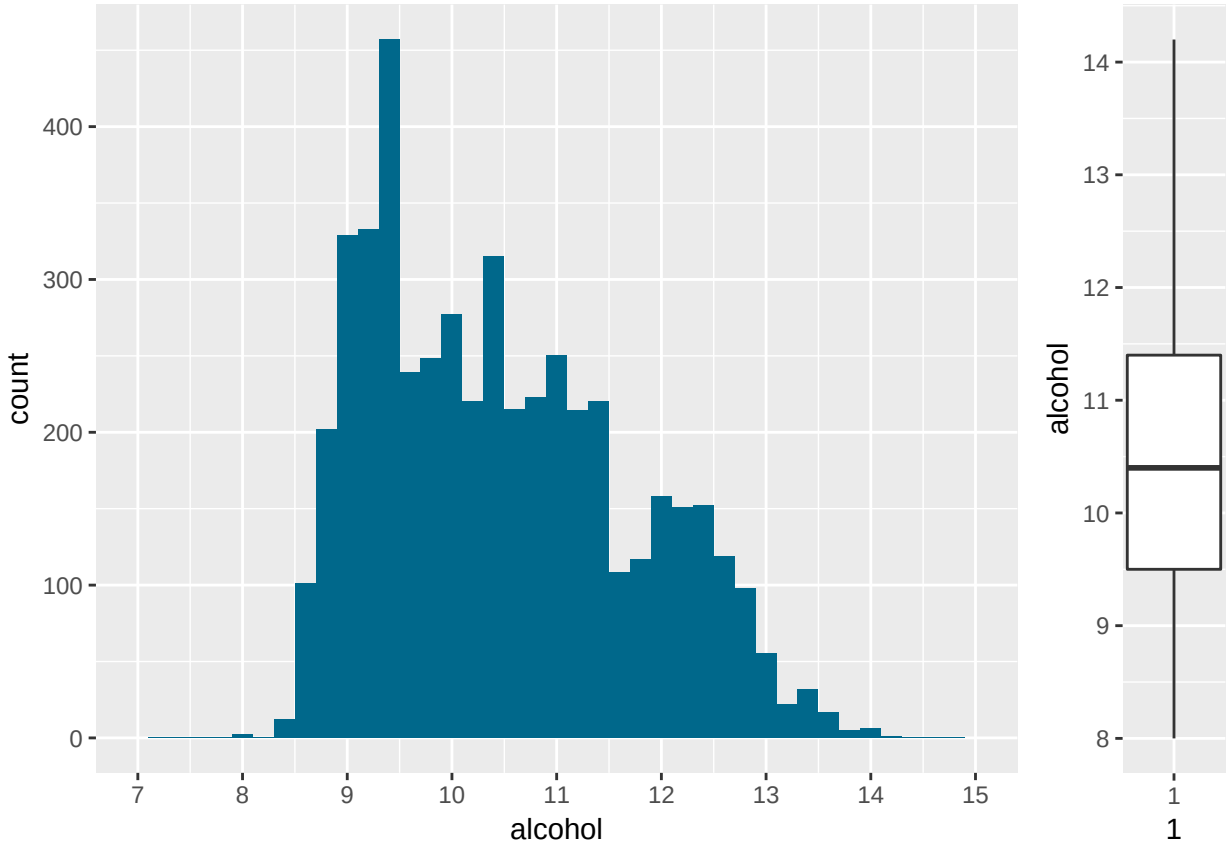
Sugar and Alcohol Contents and Density



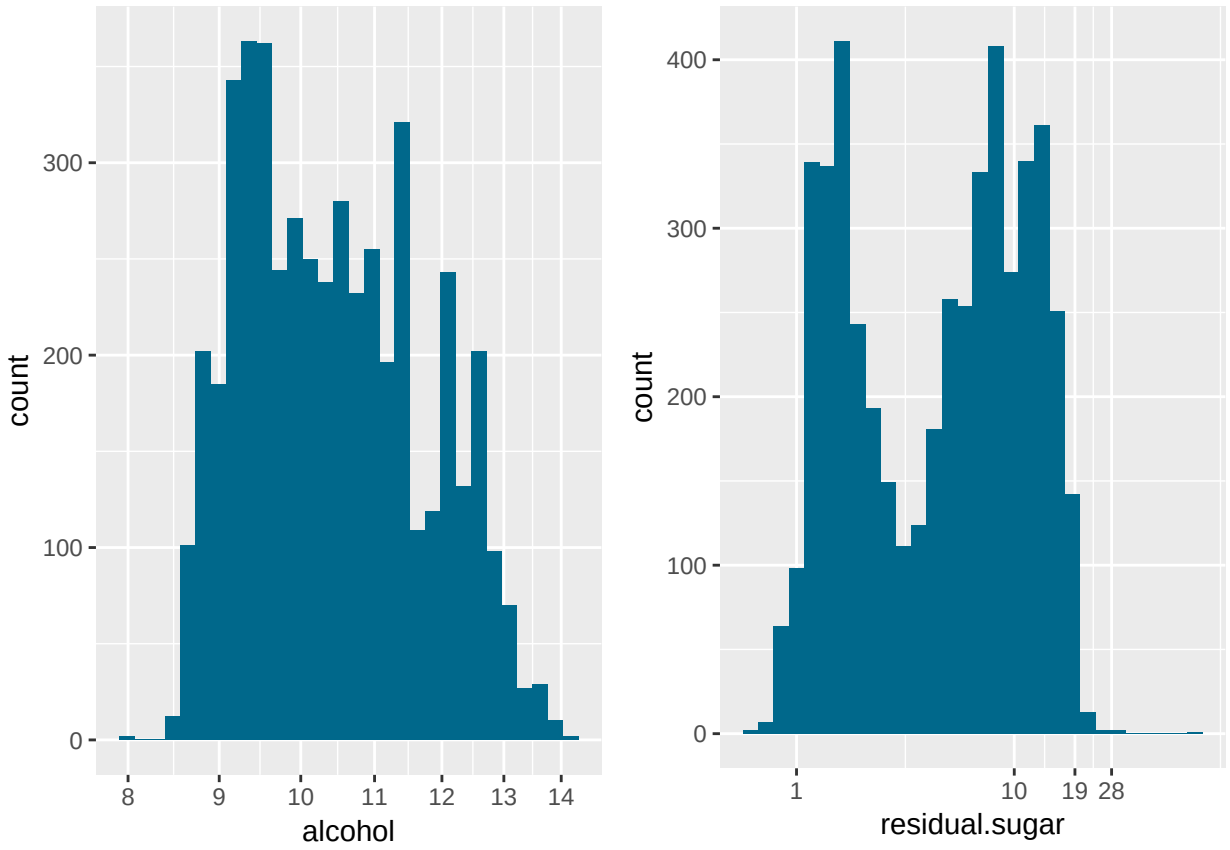
The third group of plots shows that alcohol and residual sugar distributions are not normal but density plot shows a normal distribution. Let's remove outliers and look at these plots one by one.



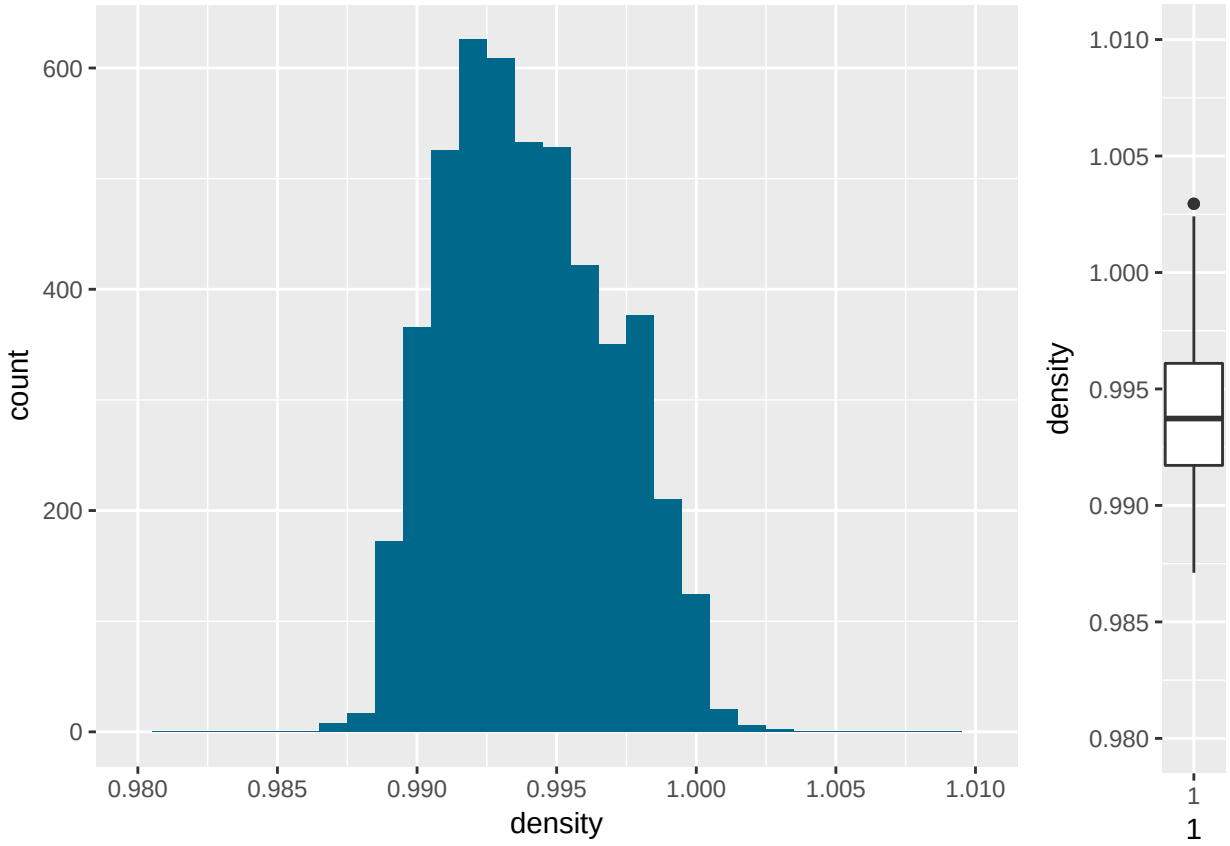
The residual sugar (amount of sugar remained after fermentation) distribution is highly right skewed with multiple peaks. Majority of white wines have residual sugar amount between 1.7 and 9.9 g/dmˆ3 with the mean and median around 6.4 and 5.2 g/dmˆ3. The right side boxplot shows the range of outliers. There is one data point at 70 g/dmˆ3 which we removed it from the boxplot. Residual sugar and alcohol content are closely associated. Next, we look at the alcohol distribution.



The alcohol content distribution is highly right skewed too with multiple peaks. Majority of white wines have alcohol percentage between 9.5 and 11.4 with the mean and median around %10.5. The right side boxplot shows almost no outlier. The residual sugar and alcohol content seems to have similar distributions. Let's look at their log distributions for further explorations of peak points.

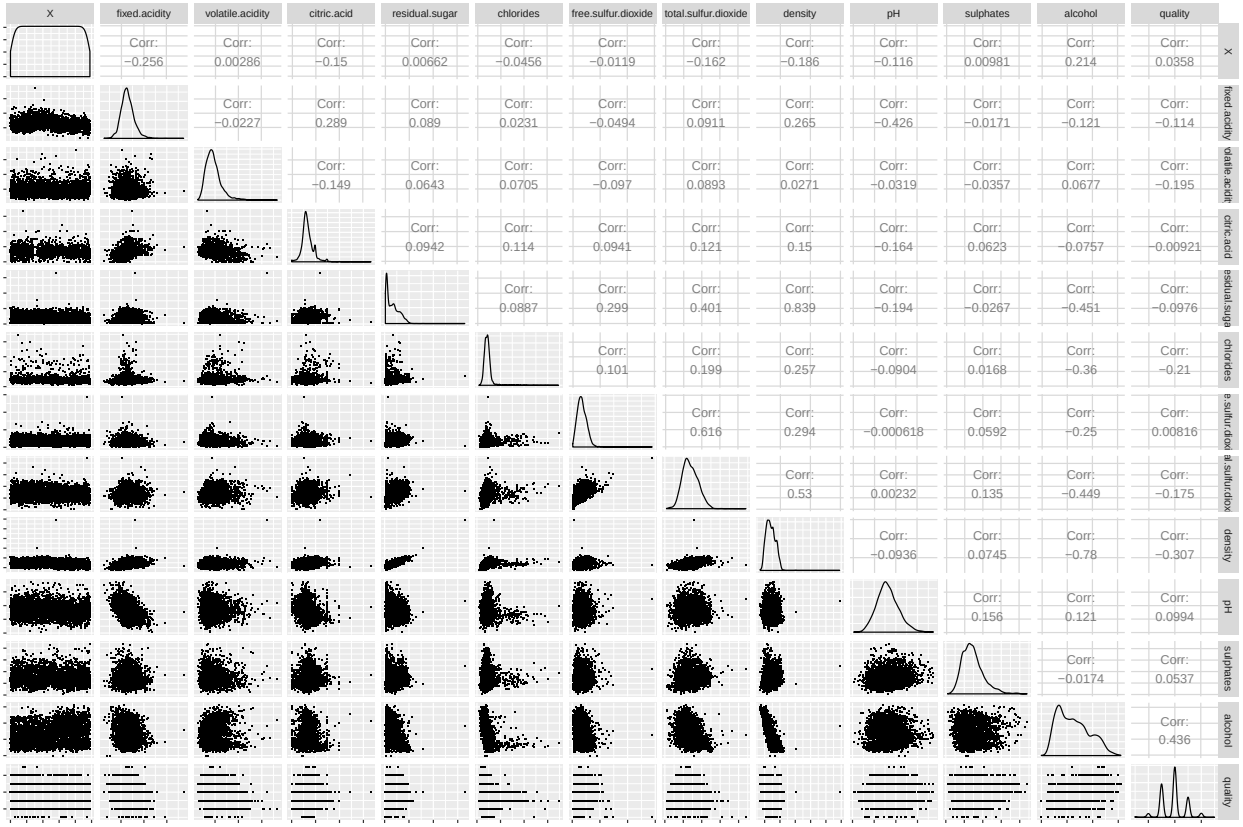


The log distribution of alcohol content is close to a normal distribution but the residual sugar distribution has two distinct peaks. However, the trend of increase and decrease of alcohol percentage is similar to the residual sugar trend.



The density of wine is directly related to alcohol and residual sugar contents. The density distribution normal with majority of white wines have density between 0.9917 and 0.9961 g/dmˆ3 and the mean and median around 0.9940 g/dmˆ3. The right side boxplot shows the range of outliers. There is one data point at 1.04 g/dmˆ3, the same as the extreme outlier in the residual sugar distribution, which we removed it from the boxplot.

Bivariate Plots and Correlations



The correlation matrix shows all distributions and correlations among input and output variables in our dataset. For example, we can observe a positive relationship between density and residual sugar and a negative relationship between density and alcohol. This totally makes sense since sugar solutions and alcohol have a higher and lower densities than water (or wine) respectively and dissolving each in water follows the same rule.

Moreover, there is a positive relationship between total SO2 and both free SO2 and chlorides.

The following table shows correlations among input variables and wine quality.

```
##                              [,1]
## fixed.acidity        -0.113662831
## volatile.acidity     -0.194722969
## citric.acid          -0.009209091
## residual.sugar       -0.097576829
## chlorides            -0.209934411
## free.sulfur.dioxide   0.008158067
## total.sulfur.dioxide -0.174737218
## density              -0.307123313
## pH                    0.099427246
## sulphates             0.053677877
## alcohol               0.435574715
```

```
##                              [,1]
## fixed.acidity        -0.113662831
## volatile.acidity     -0.194722969
## citric.acid          -0.009209091
```

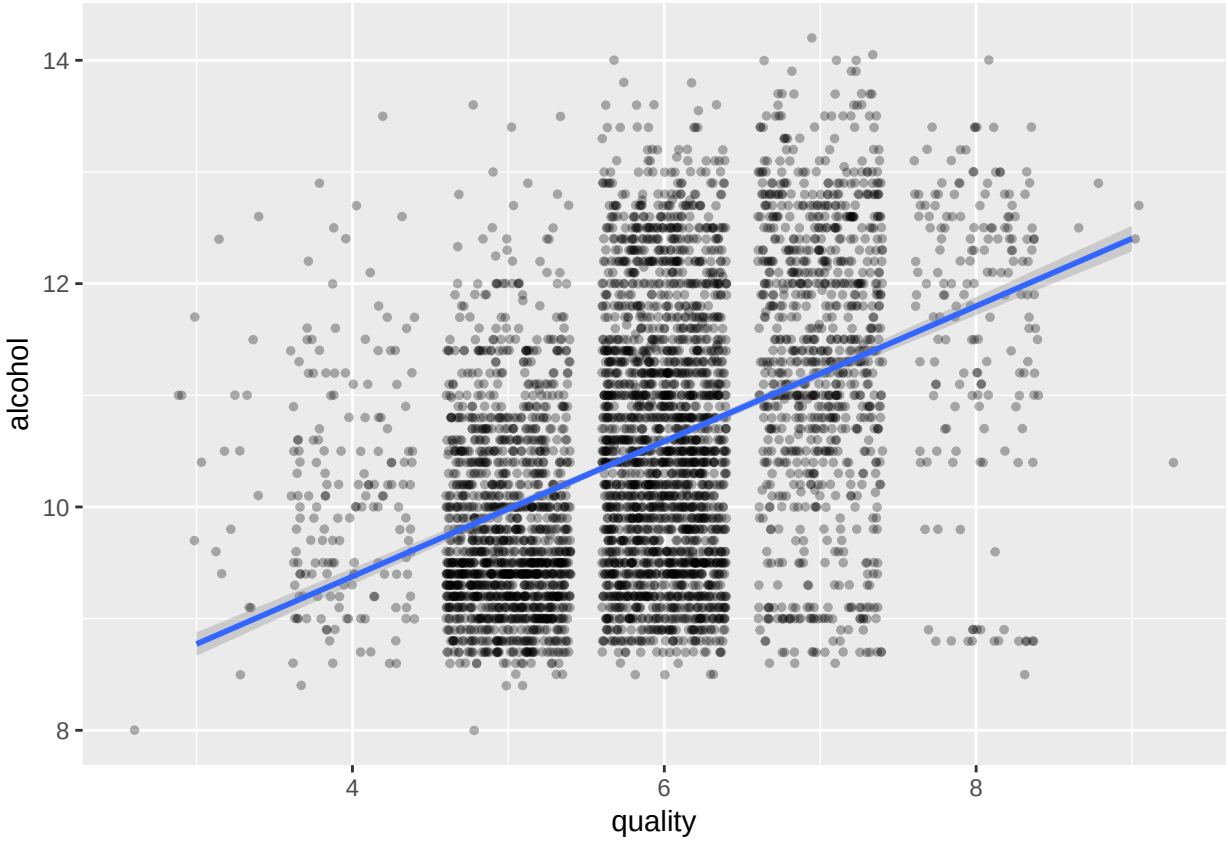
```
## residual.sugar       -0.097576829
## chlorides            -0.209934411
## free.sulfur.dioxide   0.008158067
## total.sulfur.dioxide -0.174737218
## density              -0.307123313
## pH                    0.099427246
## sulphates             0.053677877
## alcohol               0.435574715
## quality               1.000000000
```

Analyzing correlation matrix and correlation values, we can see a relatively strong correlation between wine quality and alcohol. In addition, this also shows that wine quality has negative correlations with chlorides and density.

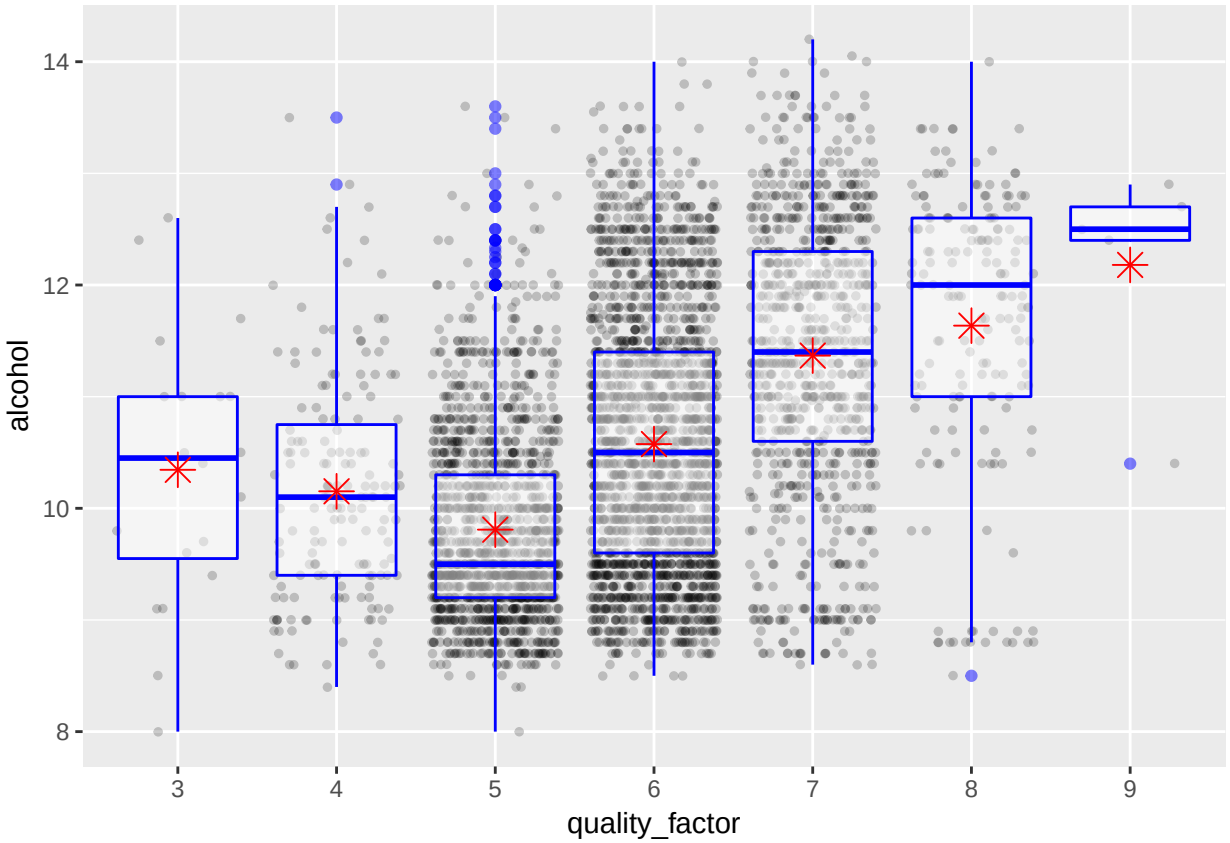
Next, we will plot different relationship between wine quality and its properties and try to find out which one can predict the quality of wine more accurately.

Wine Quality vs. Alcohol, Density and Chlorides

Since alcohol has the highest correlation of any of the features, let's start by looking at the scatterplot along with the linear model for the alcohol content and wine quality

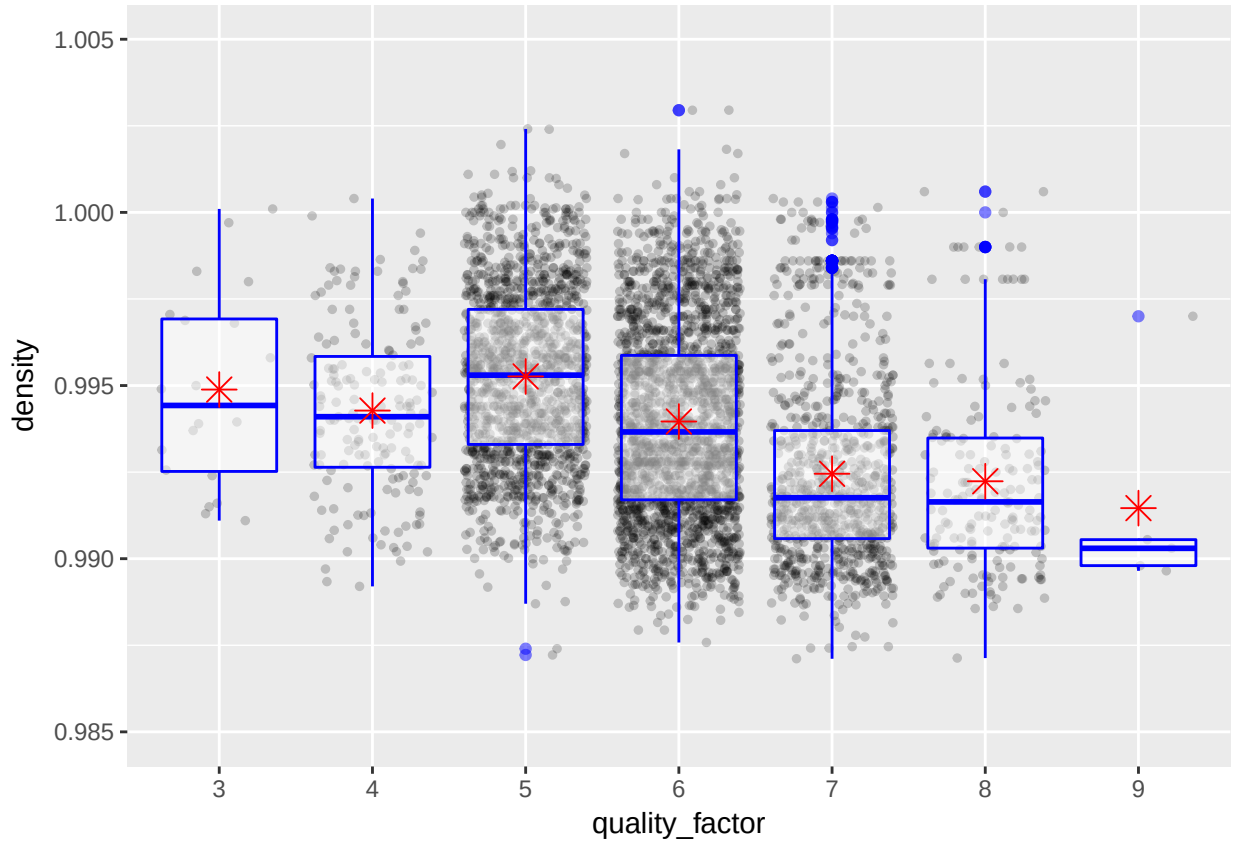


Although data points are somewhat scattered but the linear model shows a relationship between alcohol and quality (with a correlation value of 0.44). Let's explore further and check the boxplot of alcohol percent versus wine quality too.

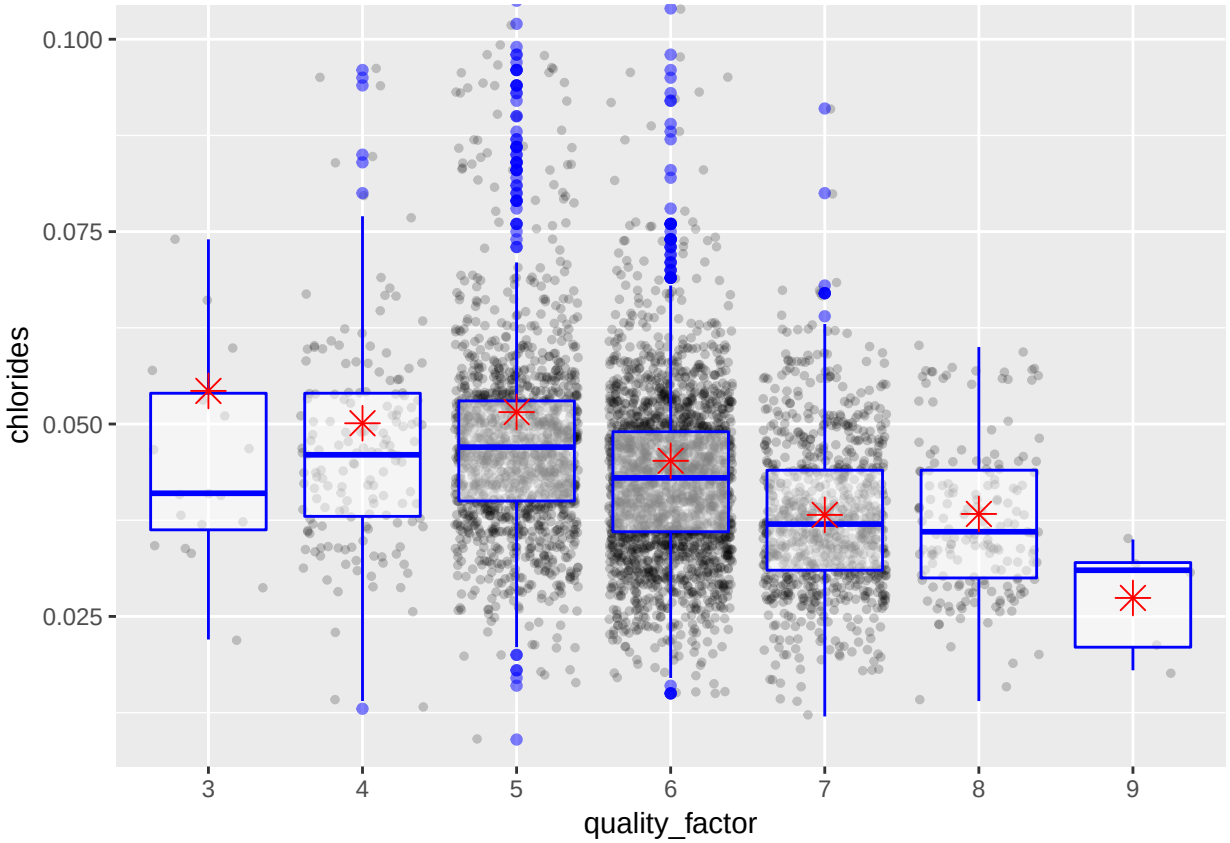


Based on the boxplot of alcohol percent versus wine quality and considering the fact that majority of wines have quality of 5 to 8, we can say, as the average alcohol content increases the wine quality increases as well.

Next, we investigate the density and chlorides scatterplots and boxplots.



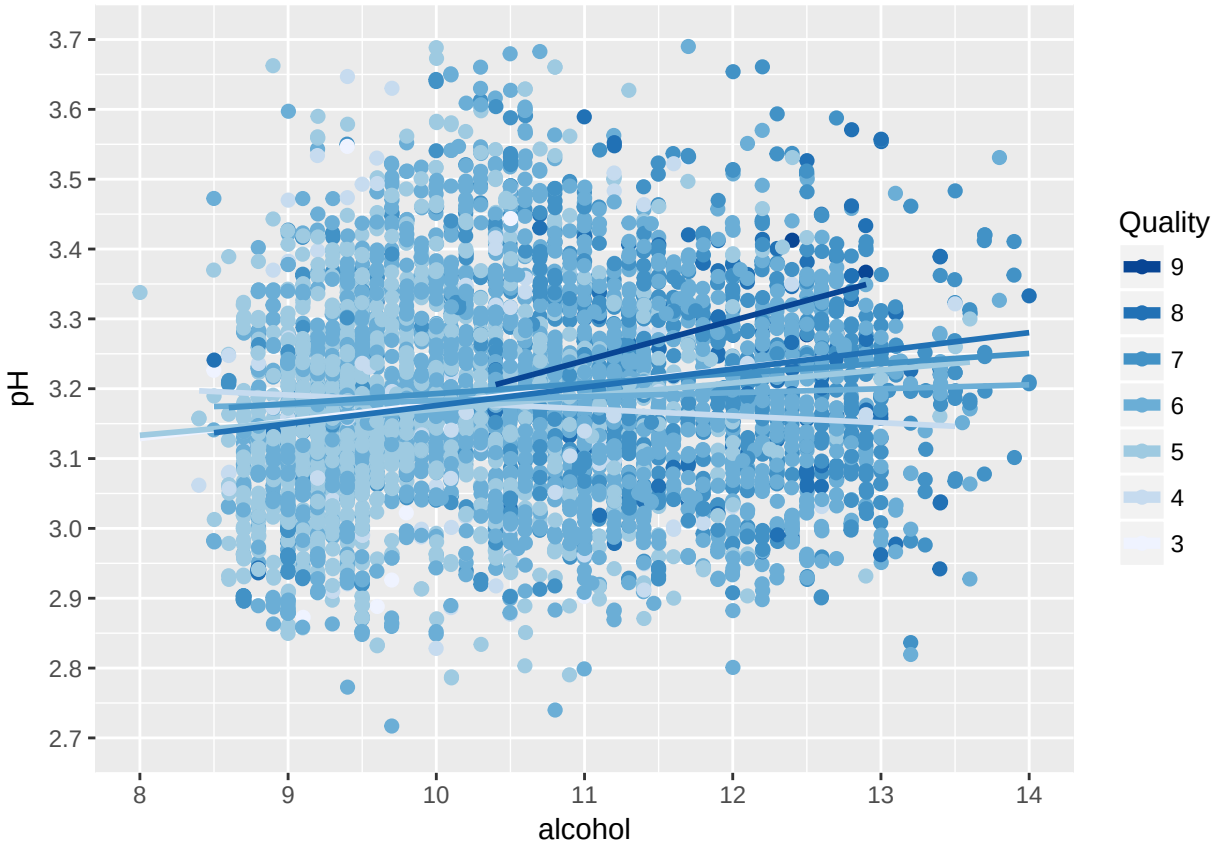
The boxplot of density versus wine quality shows a negative relationship where as the density increases the wine quality decreases. Again we are just looking at the majority of wine qualities (5 to 8). The negative correlation value of -0.3 shows the same relationship too.



Here, we can see the same trend between chlorides and wine quality. Although the correlation value (-0.2) is not significant, the boxplot shows some negative relationship between chlorides and the majority of wine qualities (5 to 8).

Multivariate Plots

We are investigating multivariate plots to find the effect of different variables on wine quality. First let's look at the relationship between pH and alcohol and quality of the wine.

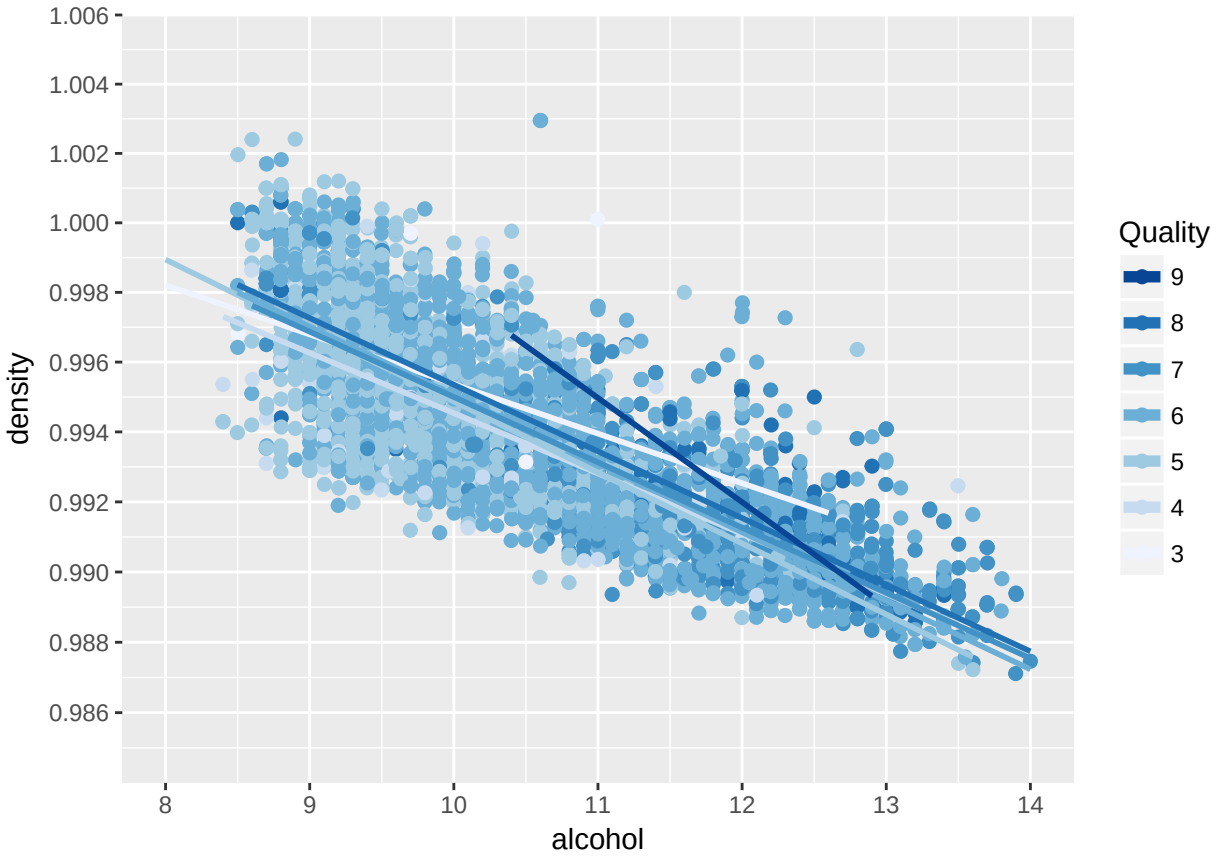


The above scatterplot shows the relationship between pH and alcohol percentage colored by quality of the wine. We also included regression lines for each quality rating to depict the separation more clearly.

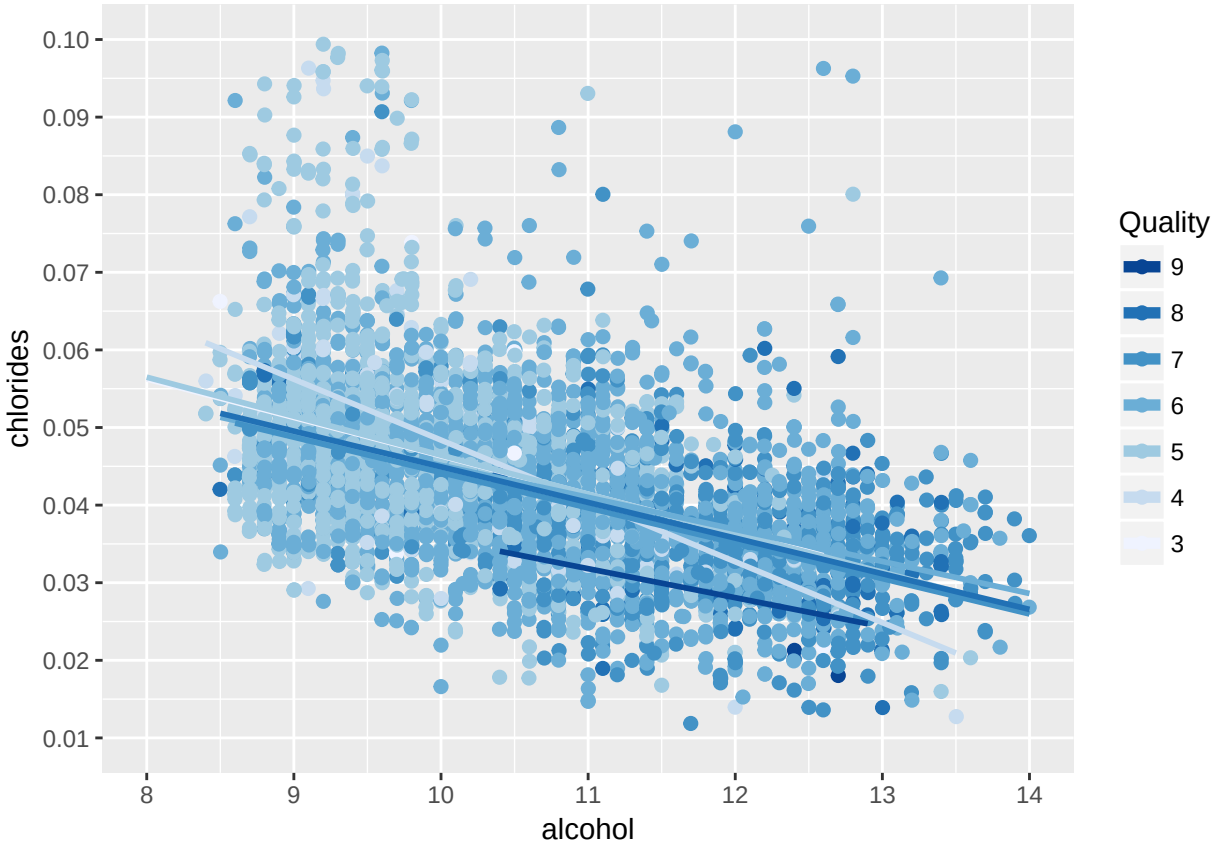
Most of great wines are in the right side of the plot. More specifically, if the alcohol percentage is above 11 percent, there seems to be a good chance that we will have a wine with quality of 7 or above.

On the other hand, as we expected, there is no relationship between pH and wine quality. The high quality wines have a wide range of pH as the low quality wines. Therefore, we can say pH impact on wine quality is insignificant.

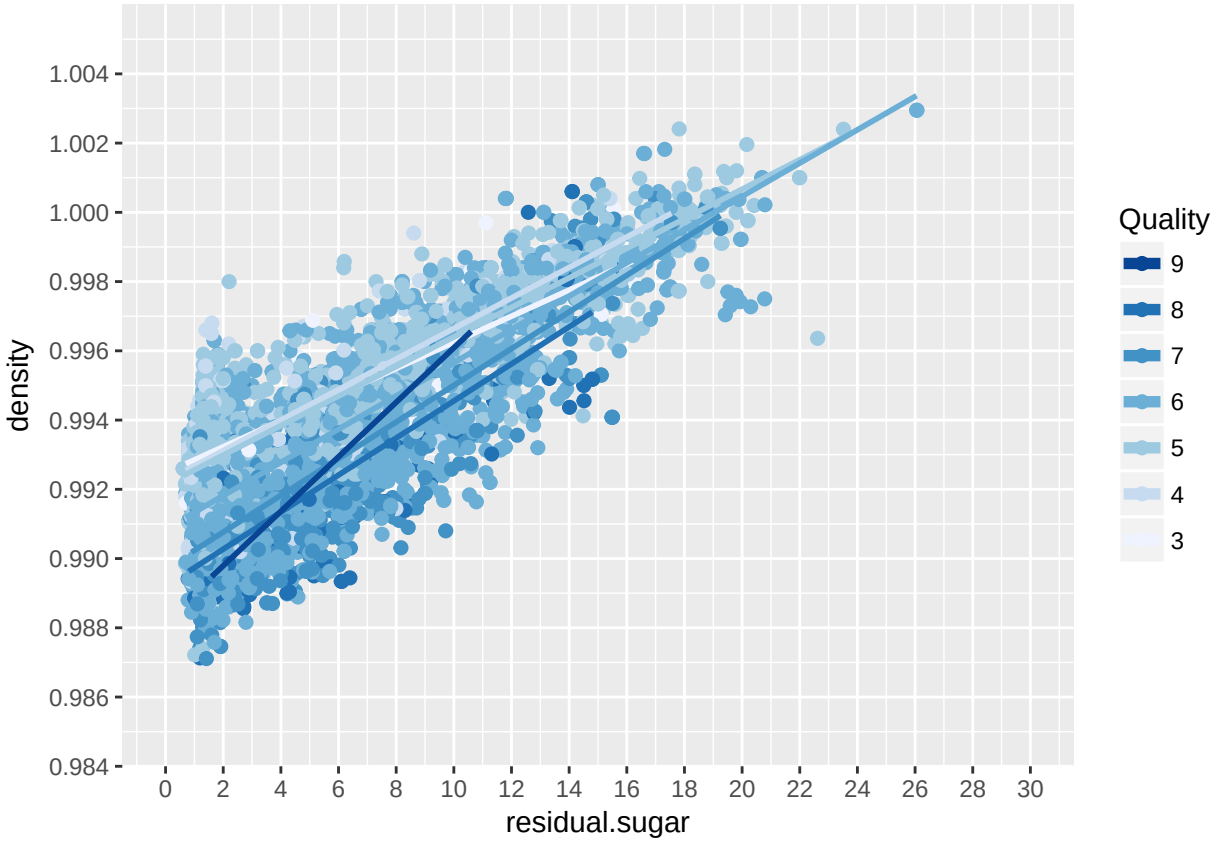
Next, we plot a similar scatterplot for density.



We can observe a strong correlation between density and alcohol percentage. As alcohol content increases, the density decreases. In addition, high quality wines tend to have less density and more alcohol.



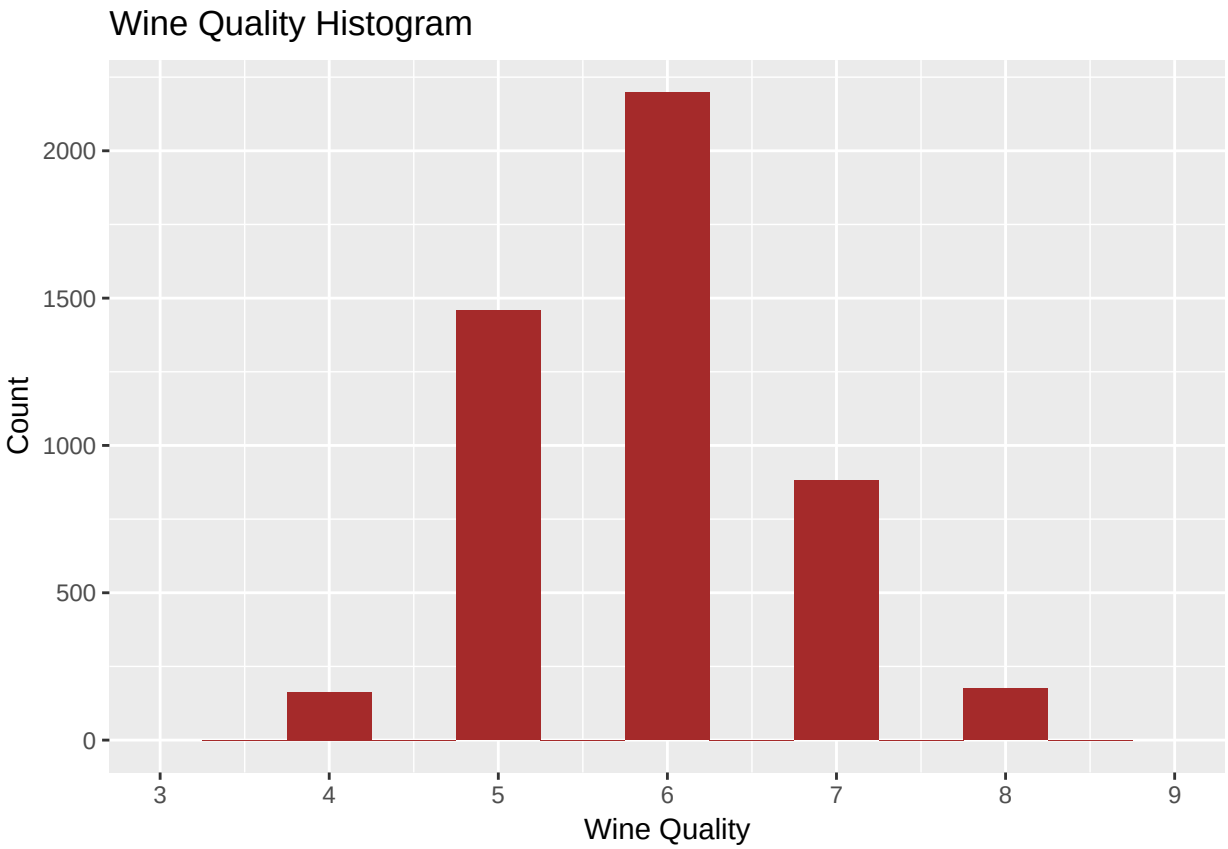
Considering scatterplot of chloride and alcohol for different wines qualities, we can see higher quality wines tend to have less chlorides. For chlorides levels higher than 0.05 g/dmˆ3, the quality of wine most likely will be 5 or less.



Finally, let's include residual sugar in the density scatterplot colored by wine quality. As we expected, increasing sugar content will increase density. Most of great wines are in the bottom part of the plot where density is lower and alcohol content is higher. More specifically, if the density is more than 0.995 g/cm^3 or residual sugar is more than 15 g/dm^3, there seems to be a good chance that we will have a wine with quality of 6 or below.

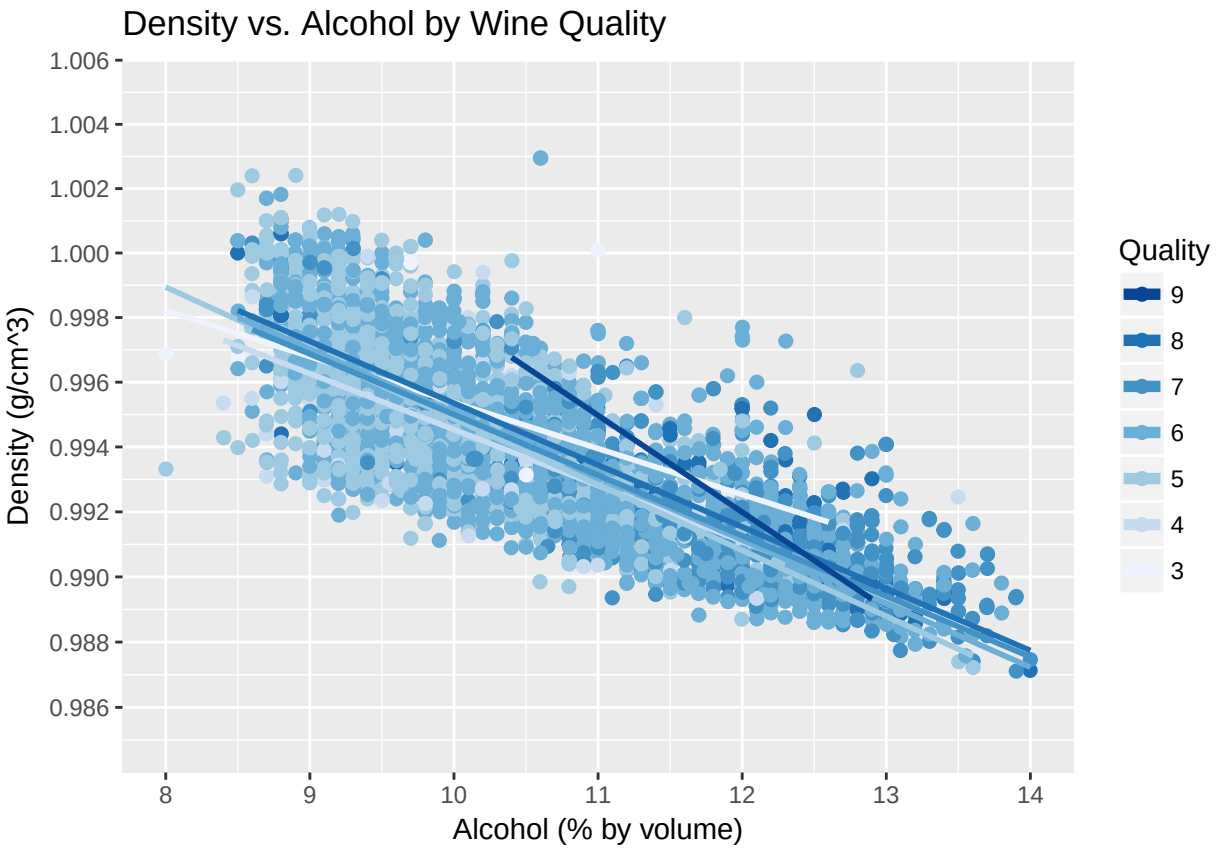
Final Plots and Summary

For the first plot , we present histogram of wine quality.



This histogram shows how wine quality is distributed in our dataset. Most of wines have a quality between 5 and 7 with the peak at the grade of 6. Therefore, in our next observations, as we investigate relationships, we can overlook wine qualities above 7 and below 5.

For the second plot, we choose scatterplot of density vs. alcohol colored by wine quality.



This multivariate plot shows a distinctive correlation between density and alcohol percentage. As alcohol content increases, the density decreases. While left side of the plot consists of poor wines (quality of 5 and below), right hand side of the plot mostly consist of good wines (quality of 7 and above). As a result, we can say, high quality wines tend to have less density or more alcohol percentage.

For the third plot, we present scatterplot of density vs. residual sugar colored by wine quality. The reason for this selection is that, residual sugar tends to have a strong relationship with density and also recognizing sugar taste in wines are easier than other chemicals of wine. So it could be a helpful hint for selecting good wines.

The scatterplot of residual sugar vs. density by wine quality shows that as sugar content increases, the density increases as well. This behavior is opposite of when alcohol content increases. Most of great wines are in the bottom part of the plot where density is lower and alcohol content is higher. More specifically, we can say, if the density is less than 0.995 g/cmˆ3 or residual sugar is less than 15 g/dmˆ3, there seems to be a good chance that we will have a wine with quality of 7 or above.

Comparing above plot with the previous plots which displayed the histogram of wine quality, and scatterplot of alcohol percentage vs. density by wine quality, one can conclude the following points:

- Most of white wines in this dataset are rated 5, 6 and 7. There are very few wines rated below 4 or above 7.
- Better wines (quality of 7 and above), tend to have higher percentage of alcohol. Majority of white wines with more than %12 of alcohol have a quality of 7 or above, and wines with less than %10.5 of alcohol are considered to be poor wines (quality of 5 and below).
- There is a strong correlation between density and residual sugar and alcohol content. Based on our observation, low density wines, tend to have less residual sugar and more alcohol. Most of great white wines (with quality of 7 or above) have density of less than 0.995 g/cmˆ3 or residual sugar of less than 15 g/dmˆ3.

Reflection and Future Analysis

In this project, we performed Exploratory Data Analysis (EDA) on the white wine dataset. We used an extensive set of libraries and new tools in R to explore the dataset. We also applied more advanced functions such as summarize, groupby, ggpairs, and quantile ranges and we used different types of plotting such as

scatterplots, histograms, boxplots, and line graphs throughout this study. We utilized useful tools such as transparency, jitter, smoothing, labeling, limiting axes, and facet wrapping to create more representative figures. Finally we used R markdown to generate a professional report.

The strongest tools in our analysis are multivariate analysis where we evaluating more than two variables at once, for different possibilities. We used powerful color packages which I found challenging in terms of finding the right packages and options for proper visualization and discovering relationships between different parameters and wine quality.

We found out that alcohol content is the most important factor to determine the quality of white wines. Also alcohol percentage is directly affected by the amount of residual sugar and both of them are changing the density linearly. In the fermentation process, the sugar converts to alcohol and the more sugar left after fermentation, the less the percentage of alcohol will be in the wine.

We have not find any significant correlation between white wine quality and sulphates or associated sulfur dioxide gas (SO2) levels. However, the amount of chlorides can affect the quality of a white wine. Higher quality wines tend to have less chlorides. For chlorides levels higher than 0.05 g/dmˆ3, the quality of wine most likely will be 5 or less.

Finally, all the white wines are acidic and have pH level between 3 and 3.5 and it cannot be a good factor to predict the quality of white wines.

There is definitely a great room to improve our analysis and come up with more accurate predictions. White wine dataset is such a rich dataset and many relationships and correlations can be extracted from it and here in this project, we investigated some distinctive relationships between wine qualities and their physicochemical properties.

For the next udacity projects, I'm going to build models for predicting the quality based on the different features. I'm also interested in extending this analysis by evaluating the red wine dataset and looking at the similarities and differences. I have learned many aspects of white wines and parameters affecting their quality and found it very interesting and will develop my EDA skills on new datasets in the future.
