Red Wine Quality Analysis

Table of Contents

Introduction

Apply exploratory data analysis techniques to explore relationships in one variable to multiple variables.

Variables (based on physicochemical tests):

1. fixed acidity (tartaric acid - g / dm^3)
2. volatile acidity (acetic acid - g / dm^3)
3. citric acid (g / dm^3)
4. residual sugar (g / dm^3)
5. chlorides (sodium chloride - g / dm^3
6. free sulfur dioxide (mg / dm^3)
7. total sulfur dioxide (mg / dm^3)
8. density (g / cm^3)
9. pH
10. sulphates (potassium sulphate - g / dm3)
11. alcohol (% by volume)
12. quality (score between 0 and 10)

Question to answer

Which chemical properties and how are they influence the quality of red wines?

Univariate Plots Section

```
## [1] 1599   13
```

Red wine dataset has 1599 rows and 13 variables.

Check to see if there is any missing values.

```
## [1] 0
```

Displaying structure of red wine dataset to learn more about those 13 variables.

```
## 'data.frame':    1599 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ 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 ...
```

Displaying first 5 rows of red wine dataset.

```
##   X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 1           7.4             0.70        0.00            1.9     0.076
## 2 2           7.8             0.88        0.00            2.6     0.098
## 3 3           7.8             0.76        0.04            2.3     0.092
## 4 4          11.2             0.28        0.56            1.9     0.075
## 5 5           7.4             0.70        0.00            1.9     0.076
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                  34  0.9978 3.51      0.56     9.4
## 2                  25                  67  0.9968 3.20      0.68     9.8
## 3                  15                  54  0.9970 3.26      0.65     9.8
## 4                  17                  60  0.9980 3.16      0.58     9.8
## 5                  11                  34  0.9978 3.51      0.56     9.4
##   quality
## 1       5
## 2       5
## 3       5
## 4       6
## 5       5
```

Displaying summary of red wine dataset.

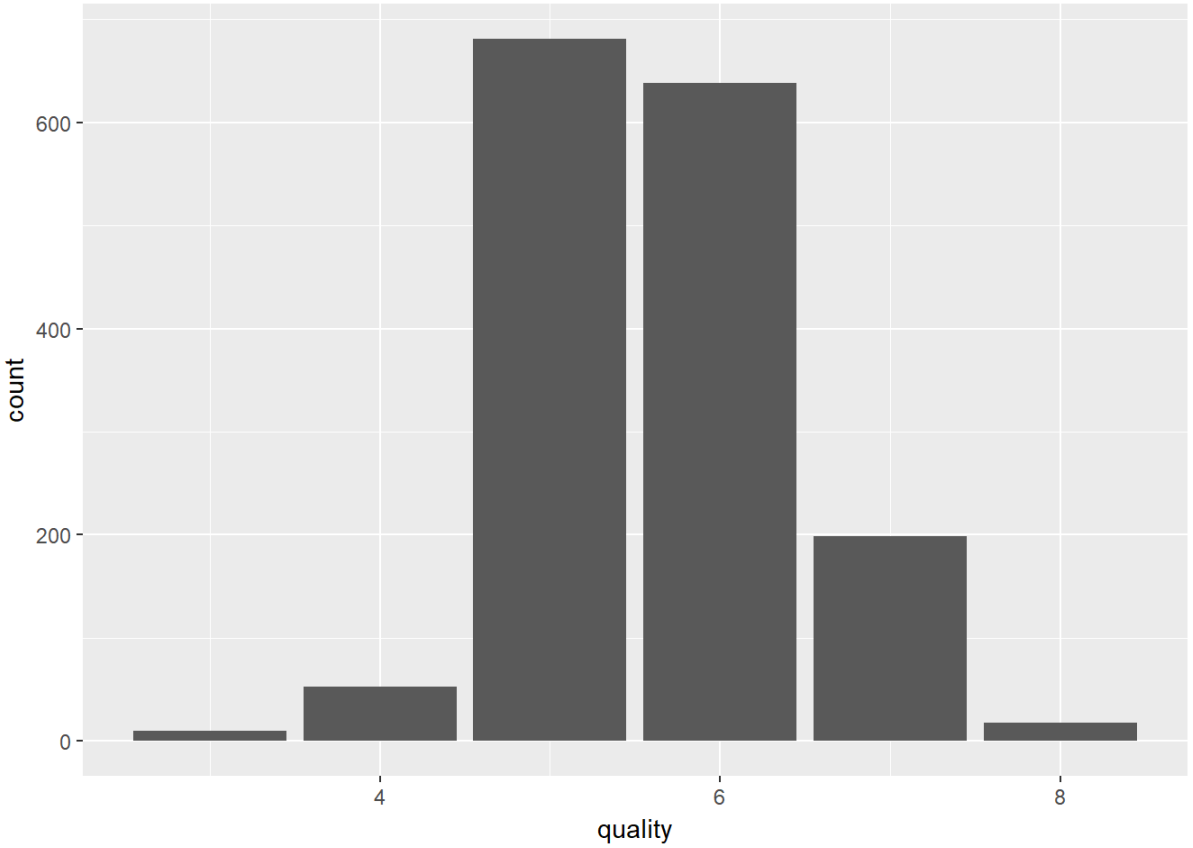
```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean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

Although quality can be represented in integer, I would like to create new column for categorical data type. Converting quality column to factor.

```
##  Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 5 5 5 6 5 5 5 7 7 5 ...
```

First, I will look at the distribution of quality.

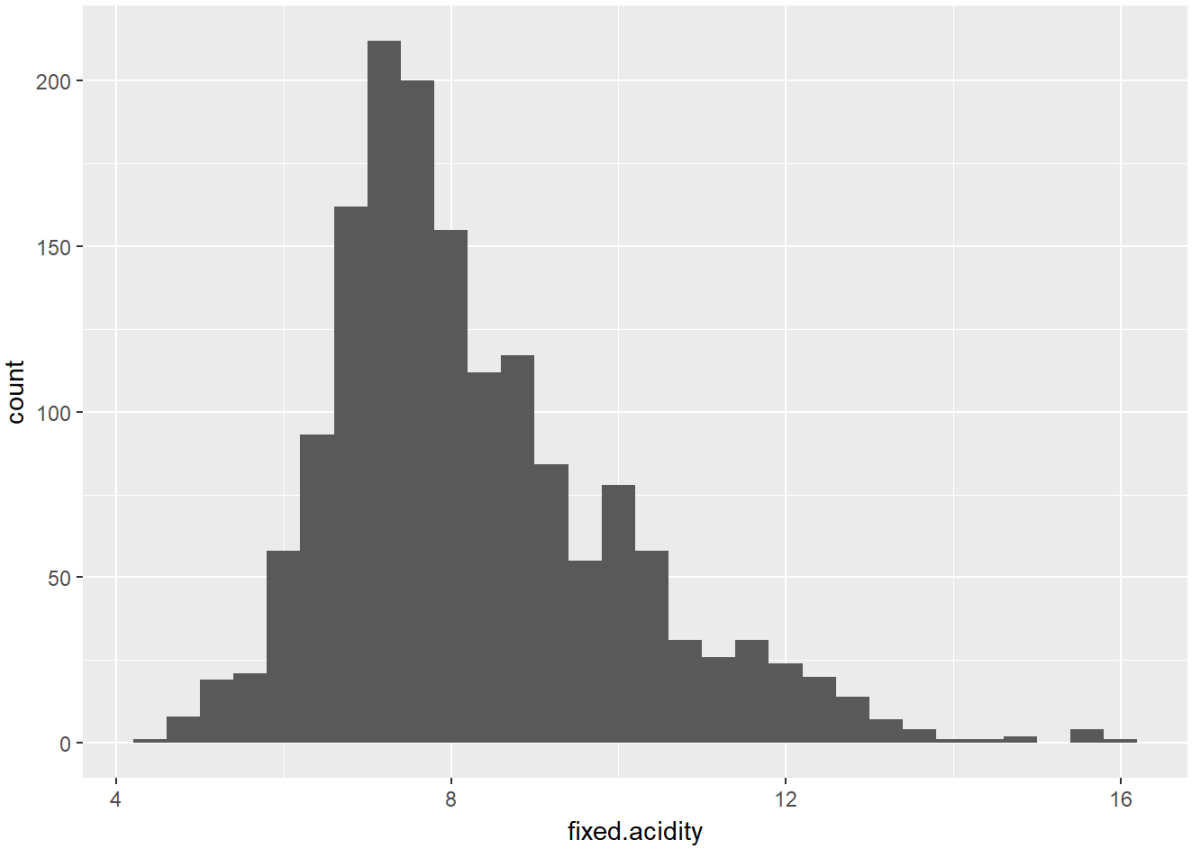
```
##
##   1   2   3   4   5   6   7   8   9  10
##   0   0  10  53 681 638 199  18   0   0
```



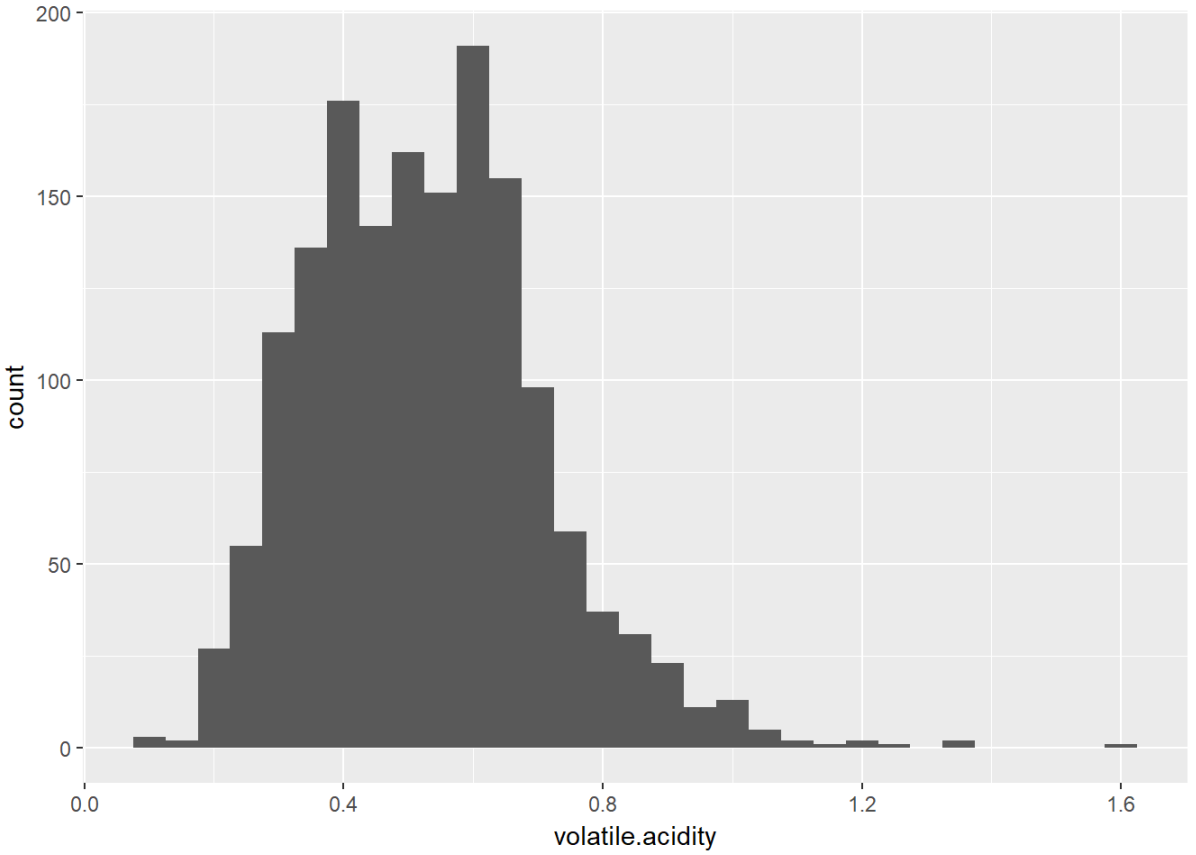
From the plot, majority of red wines in dataset fall into quality category 5 & 6. The lowest quality is 3, and the highest quality is 8. The mean is 5.636. The distribution is normal. Number of wines in quality category 3, 4, and 8 are very little compare to the ones in 5 and 6.

I want to explore which of wine features affect wine quality.

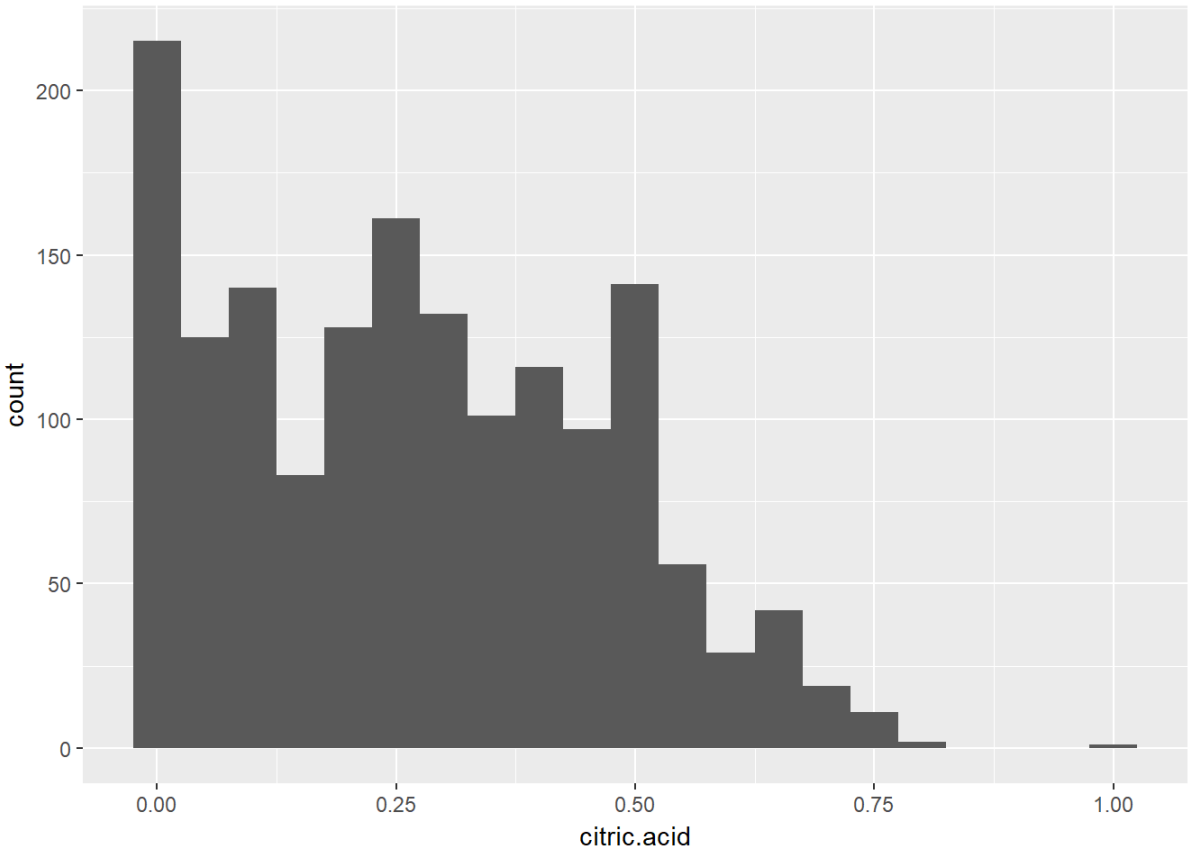
I will first look at features related to acids. - fixed acidity which impact the sourness or tartness in wine taste - volatile acidity which indicates wine spoilage - citric acid which adds freshness and flavors to wines



Majority of wine have fixed acidity between 6 and 11. The most popular value is somewhere around 7.5. I see some wines have fixed acidity of 15 - 16. However, they don't seem to be extreme, so I will consider them valid data. The distribution looks normal.

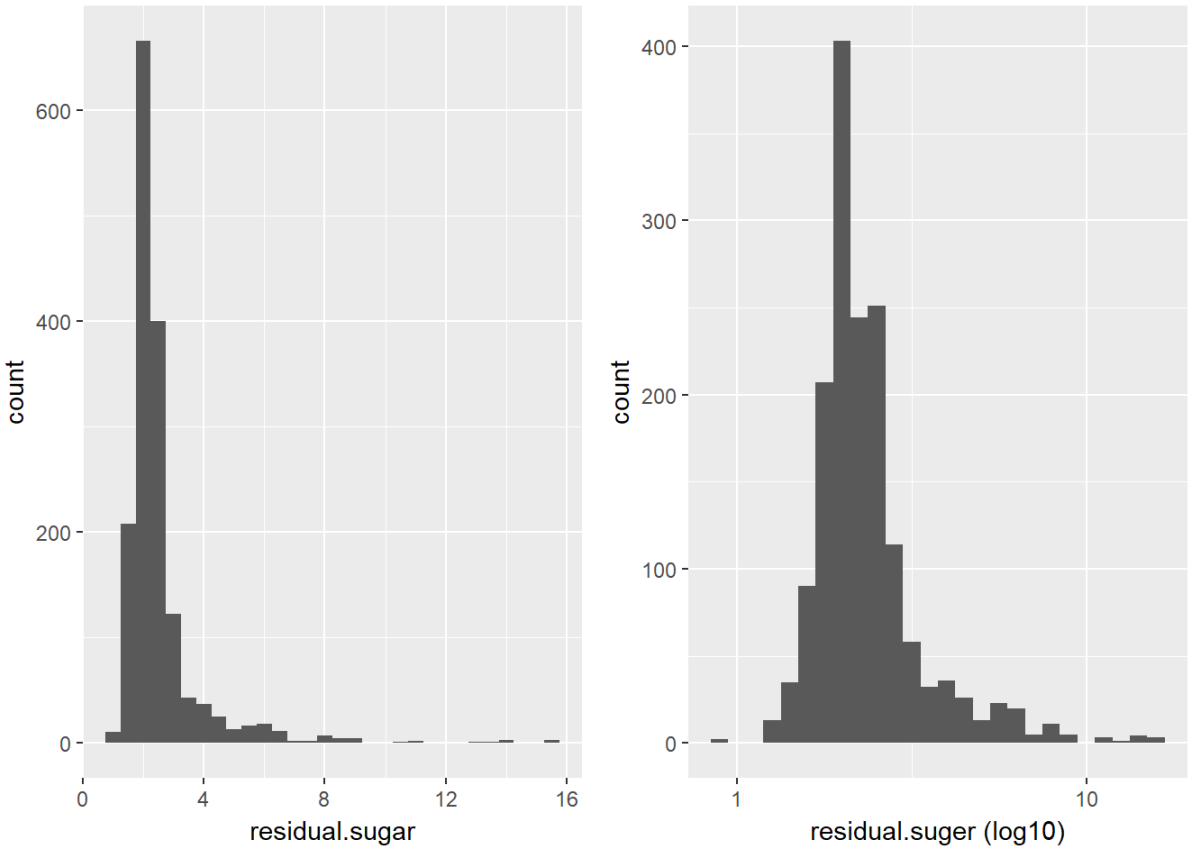


Most of wines have volatile acidity in 0.2 ~ 0.8. There are few wines have very low (<= 0.1) or very high (>= 1.1) volatile acidity. Even though volatile acidity is associated with wine spoilage, there is no wine with volatile acidity 0.0. I see some wines have volatile acidity of 1.6. However, they don't seem to be extreme, so I will consider them valid data. The distribution looks bimodal. There is a dip around 0.45 ~ 0.6.



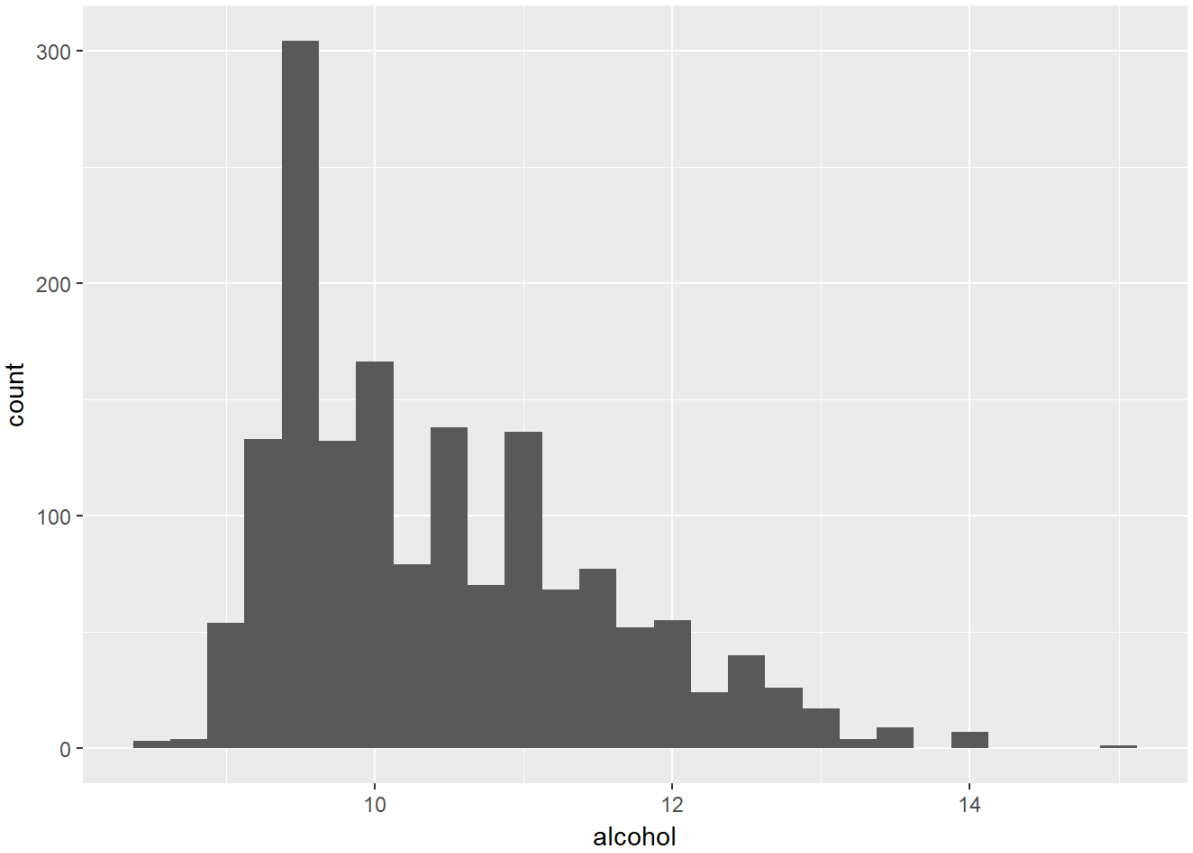
Majority of wines have citric acid between 0.00 to 0.50. There are very few wines have very high citric acid (> 0.75). I see some wines have citric acid of 1. However, they don't seem to be extreme, so I will consider them valid data.

Other features I would like to look at is residual suger and alcohol since they both play roles in sweet taste in wine.



Transformed the long tail data to better understand the distribution of residual sugar. After log-transformed, the distribution looks normal.

Most of wine has residual sugar between 2 and 5. Some wines have twice or closed to three times more residual sugar than majority of wines. The most popular residual sugar value is around 2.6 ~ 2.7.



Most of wines has alcohol percentage between 5 and 12. The most popular alcohol percentage is around 7.5. I see some wine has alcohol percentage 15, but I consider this data valid data. The distribution looks normal.

Univariate Analysis

What is the structure of your dataset?

There are 1599 red wines in dataset with 13 features.

1. Wine ID (integer)
2. Fixed acidity (numeric) - Acidity in wine
3. Volatile acidity (numeric) - Wine Fault, Defect, Spoilage
4. Citric acid (numeric) - Freshness and flavors of wine
5. Residual sugar (numeric) - Amount of sugar in wine
6. Chlorides (numeric) - Amount of salt in wine
7. Free sulfur dioxide (numeric) - Anti-microbial and anti-oxidant

8. Total sulfur dioxide (numeric) - Anti-microbial and anti-oxidant
9. Density (numeric) - Wine density
10. pH (numeric) - Measure of the acidity of wine, 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
11. Sulphates (numeric) - Preservative, Anti-microbial and anti-oxidant
12. Alcohol (numeric) - Percent alcohol content of wine
13. Quality (integer) - Wine Quality (score between 0 and 10)
14. Quality (factor) - Wine Quality (score between 0 and 10)

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

The main features in the dataset is quality and features impact sweetness and acidity of wine, i.e., fixed and volatile acidity, citric acid, residual sugar, and alcohol. I like to investigate which of those features have strong correlation to wine quality.

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

I think Chlorides which indicates salt in wine, pH level which measures the acidity, and density which is affected by sugar and alcohol content will support my investigation.

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

I have created new column for quality factor data type.

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 didn’t see anything unusual. There were few outliers in some of variables, but they didn’t seem extreme to indicate bad data. I also didn’t see any missing values in the dataset, either.

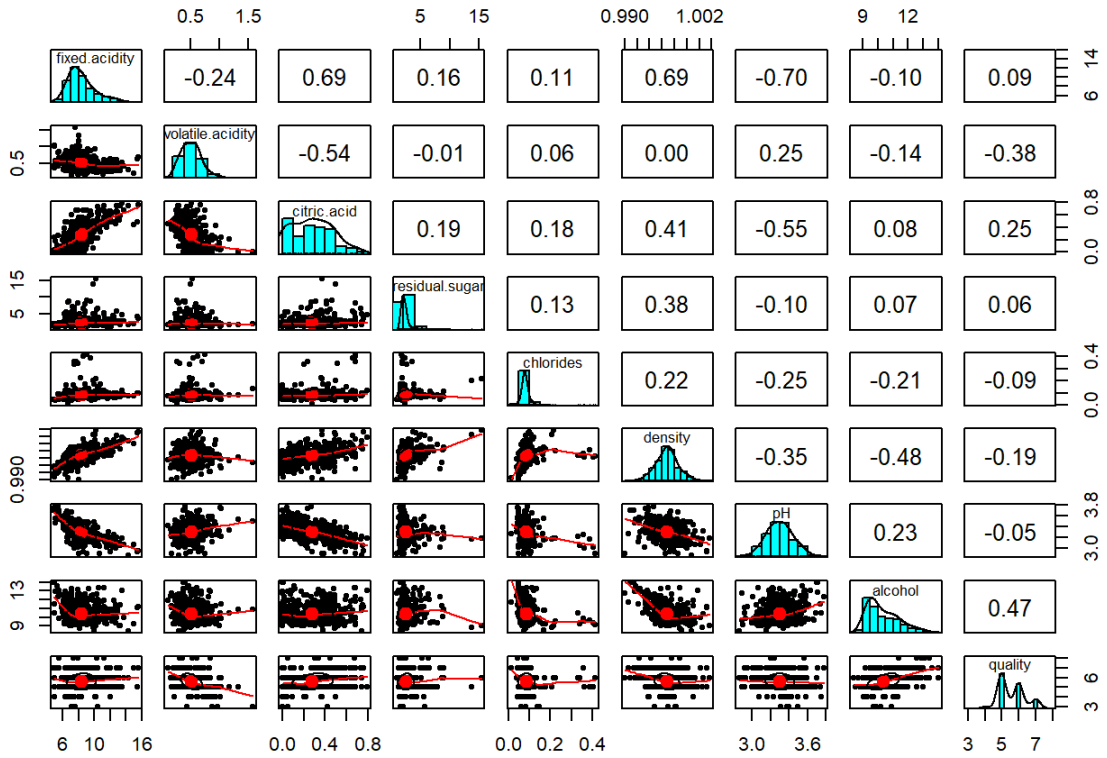
I log-transformed long tail data to better understand the distribution of residual sugar.

Bivariate Plots Section

```
##                  fixed.acidity volatile.acidity citric.acid residual.sugar
## fixed.acidity            1.000           -0.256       0.672          0.115
## volatile.acidity        -0.256            1.000      -0.552          0.002
## citric.acid              0.672           -0.552       1.000          0.144
## residual.sugar           0.115            0.002       0.144          1.000
## chlorides                0.094            0.061       0.204          0.056
## density                  0.668            0.022       0.365          0.355
## pH                      -0.683            0.235      -0.542         -0.086
## alcohol                 -0.062           -0.202       0.110          0.042
## quality                  0.124           -0.391       0.226          0.014
##                  chlorides density     pH alcohol quality
## fixed.acidity        0.094   0.668 -0.683  -0.062   0.124
## volatile.acidity     0.061   0.022  0.235  -0.202  -0.391
## citric.acid          0.204   0.365 -0.542   0.110   0.226
## residual.sugar       0.056   0.355 -0.086   0.042   0.014
## chlorides            1.000   0.201 -0.265  -0.221  -0.129
## density              0.201   1.000 -0.342  -0.496  -0.175
## pH                  -0.265  -0.342  1.000   0.206  -0.058
## alcohol             -0.221  -0.496  0.206   1.000   0.476
## quality             -0.129  -0.175 -0.058   0.476   1.000
```

Looking at the correlation coefficient between quality and other variables I am interested in from the univariate analysis, volatile acidity (-0.391) and alcohol (0.476) are moderately correlated with quality and citric acid is somewhat correlated with quality.

All other features I assumed to have effect on quality wine, i.e., Fixed acidity, residual sugar, chlorides, density, and pH level, have weak correlation with quality.



Other things I observed in the matrix above is I see some correlation between features related to each other.

Alcohol and residual sugar are moderately correlate with density. This make sense as density is determined by alcohol and sugar contents of wine. pH moderately correlates with fixed acidity and citric acid as it measures acidity of wine. Lastly, fixed acidity and volatile acidity correlates with citric acid.

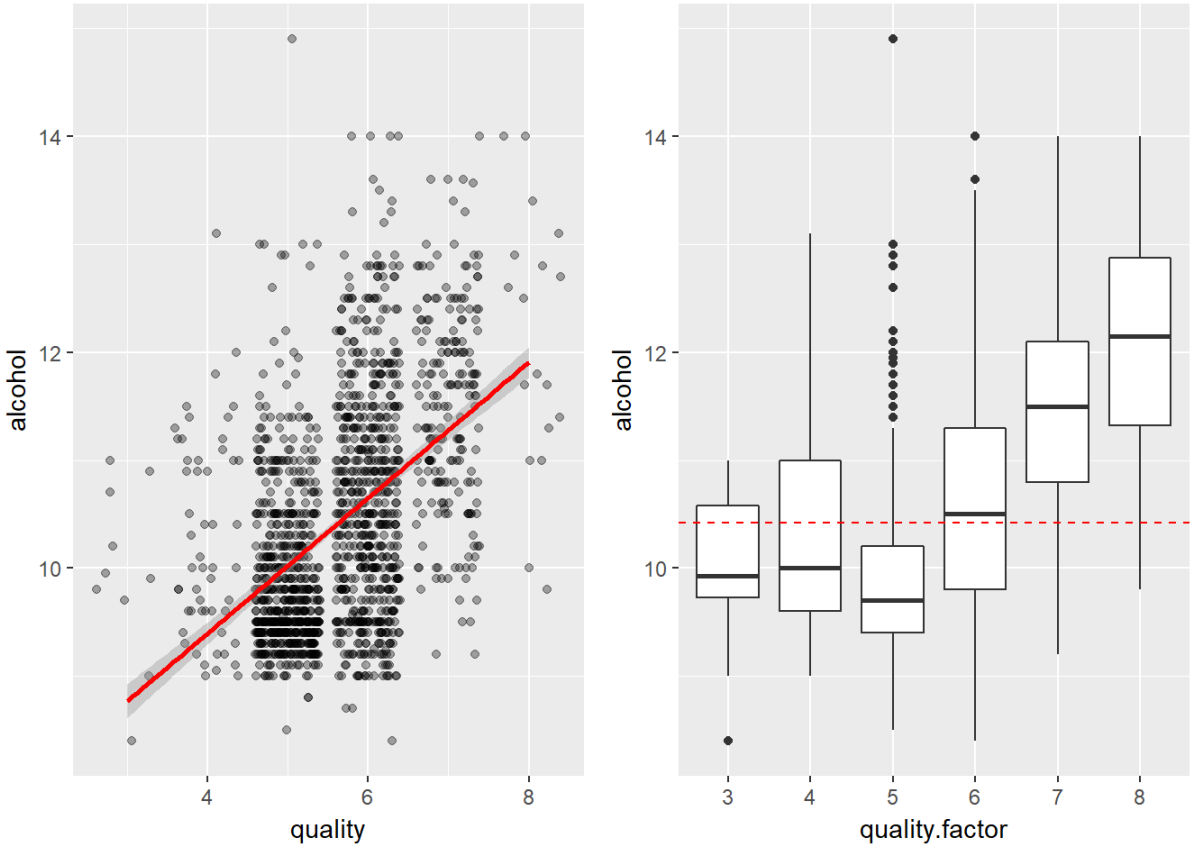
I will explore 3 most correlated variables, volatile acidity, alcohol, and citric acid, with quality further and investigate how they affect wine quality.

Before I start exploring 3 most correlated variables, I am going to look at summary of quality once again.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.636   6.000   8.000
```

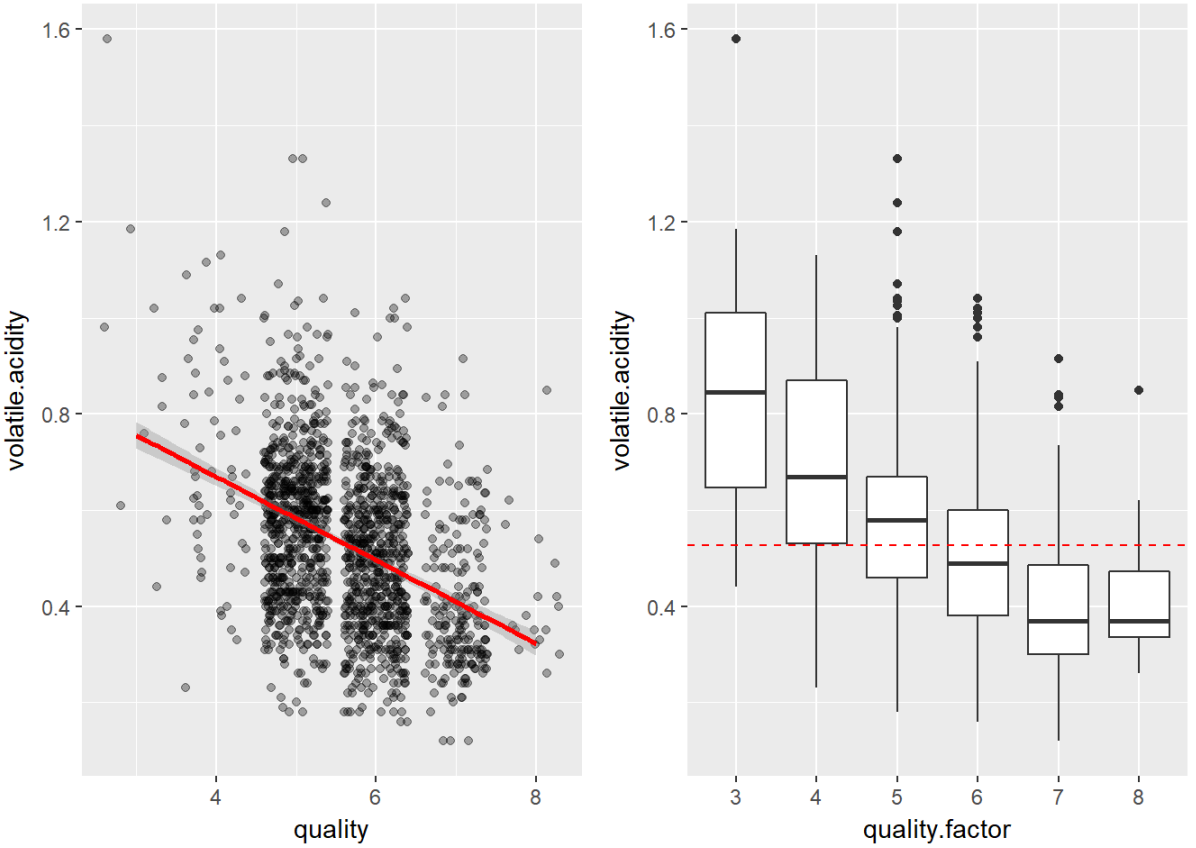
Mean is 5.636. Third quartile value is 6, and maximum value 8. I will consider good quality wine as any wines categorized as quality 6.

I will first look at alcohol as it has strongest correlation with quality among given features.



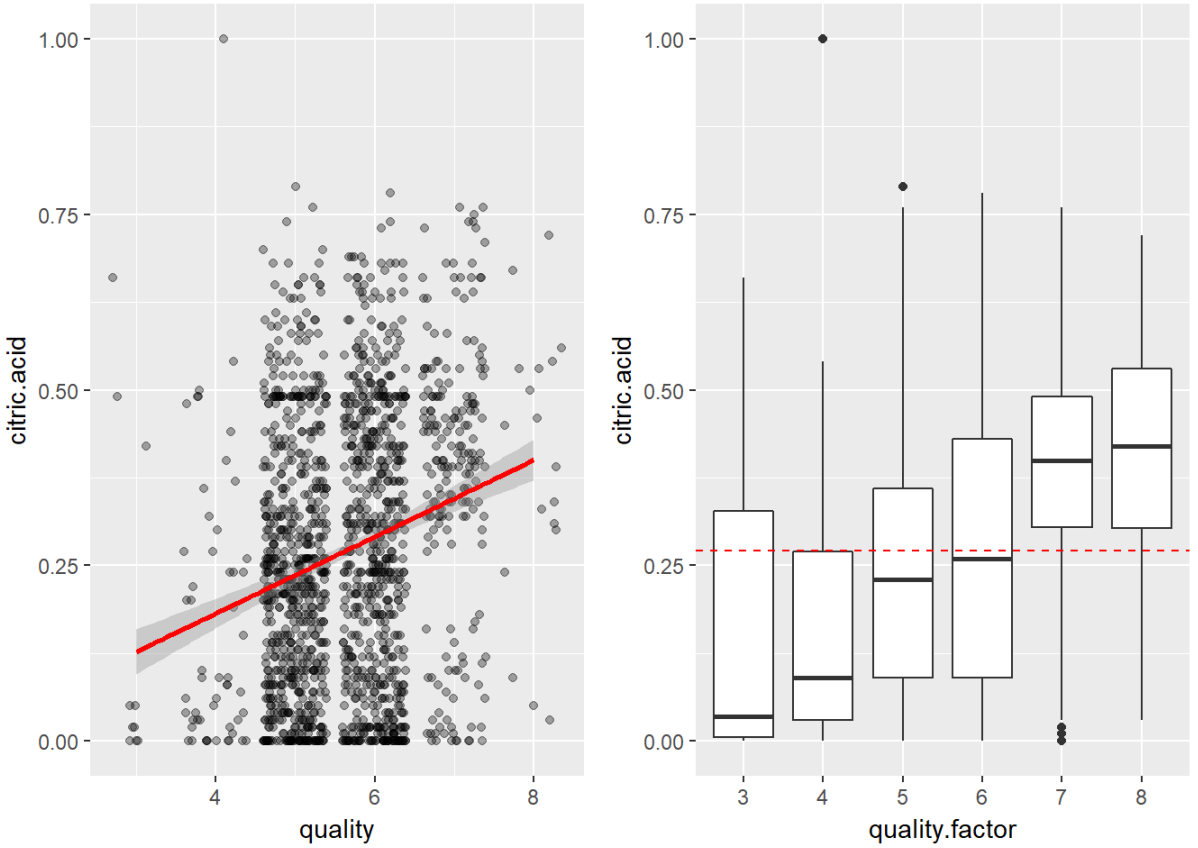
Scatter plot of alcohol confirms that there is moderate positive relationship to quality.

More than 50% of wines in quality category 6 and more than 75% wines in quality category 7 & 8 have alcohol percentage greater than its mean. What's interesting to see is that quality category 5 has the most number of wines (681), yet it has lowest median and narrowest IQR (9.4 - 10.2). It also has the most number of outliers.



Scatter plot of volatile acidity confirms that there is moderate negative relationship to quality.

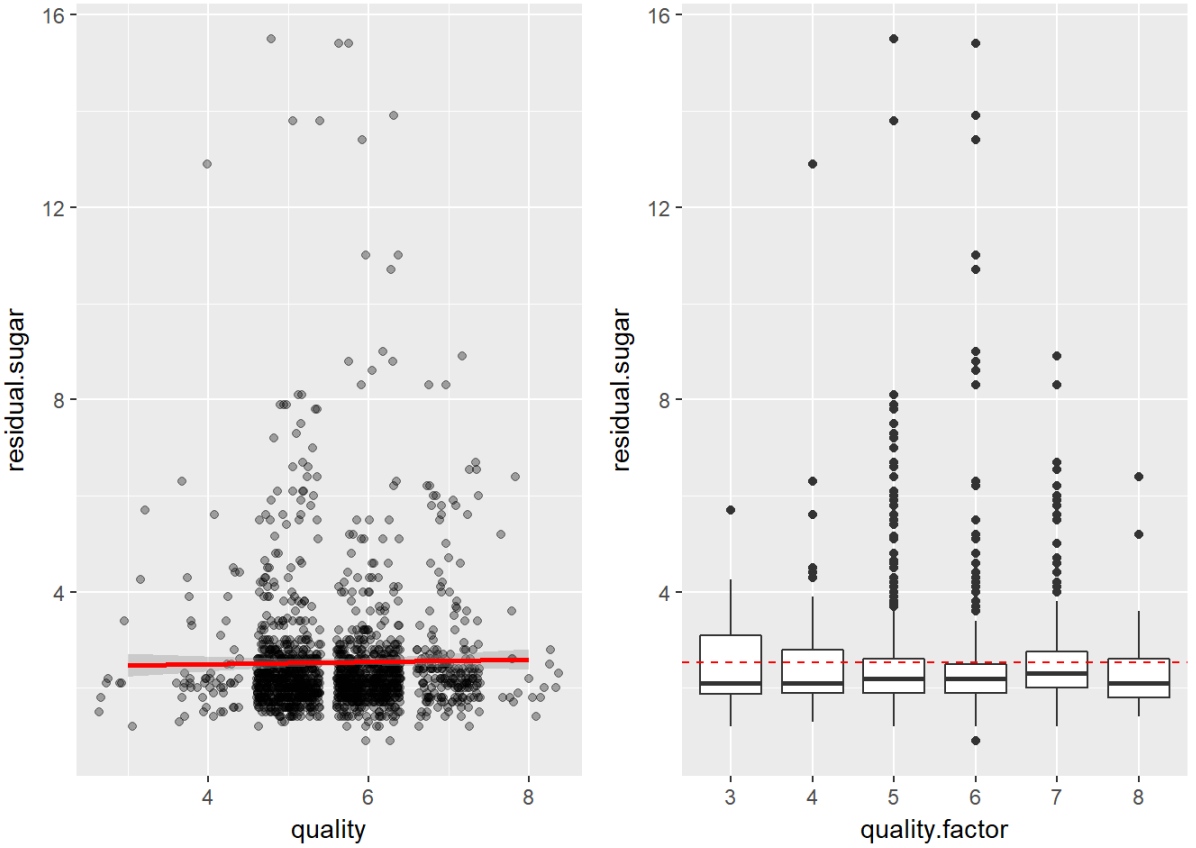
The trend between volatile acidity and quality is clearer, with worst quality category 3 having the highest median and the widest IQR. More than 50% of wines in quality category 6 and more than 75% wines in quality category 7 & 8 have volatile acidity lower than its mean. The median of quality category 7 & 8 are about the same, although the variation in volatile acidity is narrower in quality category 8.



Scatter plot of citric acid confirms that there is moderate positive relationship to quality.

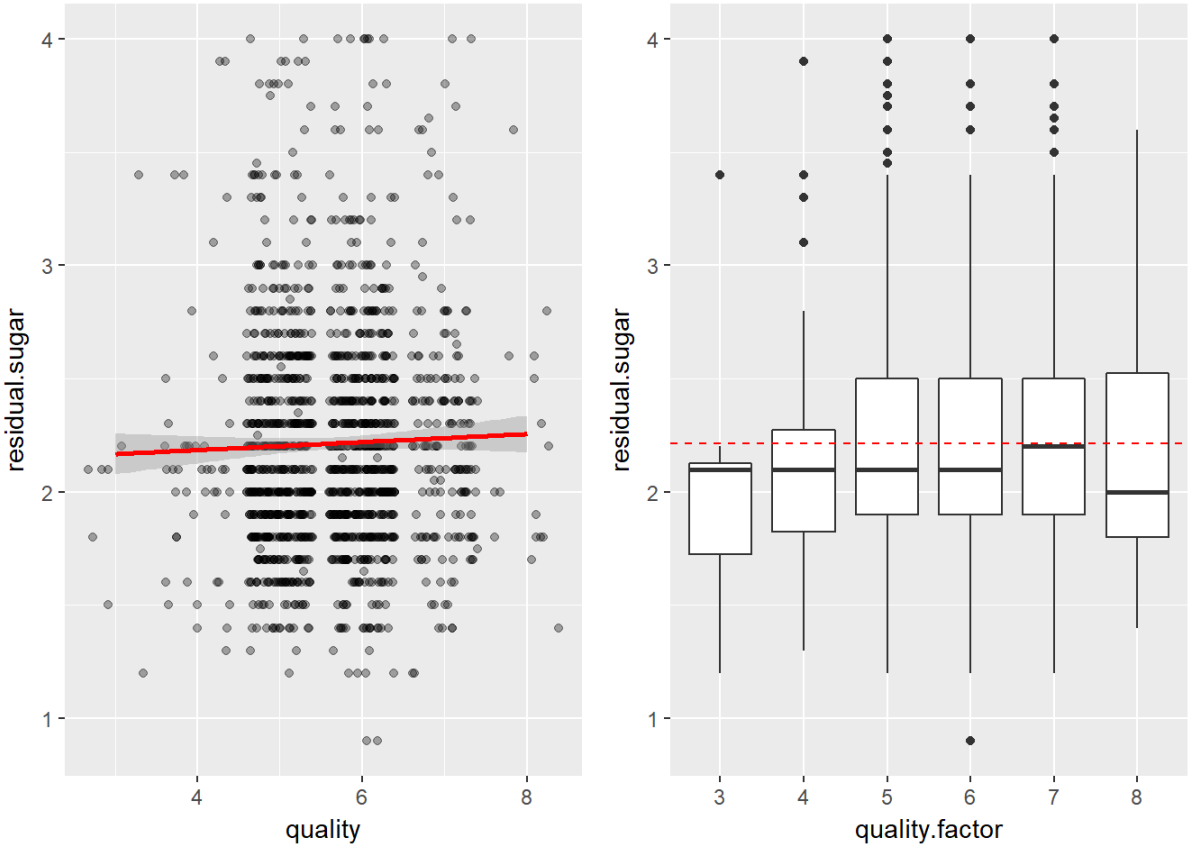
The trend between citric acid and quality is very clear, with worst quality category 3 having the lowest median. More than 50% of wines in quality category 6 and more than 75% wines in quality category 7 & 8 have greater than its mean. IQR of quality category under 7 are wider than those in 7 & 8.

Lastly, I will look at residual sugar.



To my surprise, it has very weak correlation to quality. Scatter plot of residual sugar shows there is almost no or very weak correlation with quality.

The boxplot of residual sugar shows that about 75% of wines in most quality categories are below its mean. Most of wines which have residual sugar higher than 4 are marked as outliers. I wonder a result would be different if I subset the residual sugar data to less than equal to its value 4 to remove outliers.



There is very weak correlation between residual sugar and quality even after I remove outliers.

Interesting thing to see in the boxplot is quality category 8 has the lowest median and the widest IQR. But there are no indication higher residual sugar leads to the better or poor 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?

I observed alcohol, volatile acidity, and citric acid are 3 strongest relationship to quality. When Alcohol or citric acid increases, quality of wine increases. When volatile acidity decreases, quality of wine decreases.

For alcohol percentage, quality category 5 have the most number of wines, yet it has the lowest median and the narrowest IQR. It also has the most number of outliers. Quality category 6 has number of wines close to quality category 5, yet it has one of the widest IQR of alcohol percentage.

The trend between volatile acidity and quality is clear. Worst quality category 3 having the highest median and the widest IQR, and best quality category 8 having the lowest median and the narrowest IQR.

The trend between citric acid and quality is very clear. Worst quality category 3 having the lowest median and the narrowest IQR, and best quality category 8 having the highest median and the widest IQR.

To my surprise, residual sugar has very weak correlation with quality.

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

I observed moderate correlation between features which are related to each other.

Alcohol and residual sugar are moderately correlate with density. This make sense as density is determined by alcohol and suger contents of wine. pH moderately correlates with fixed acidity and citric acid as it measures acidity of wine. Lastly, fixed acidity and volatile acidity correlates with citric acid.

What was the strongest relationship you found?

The strongest correlation I found with quality was alcohol percentage. It's a positive relationship, so higher alcohol percentage means better quality.

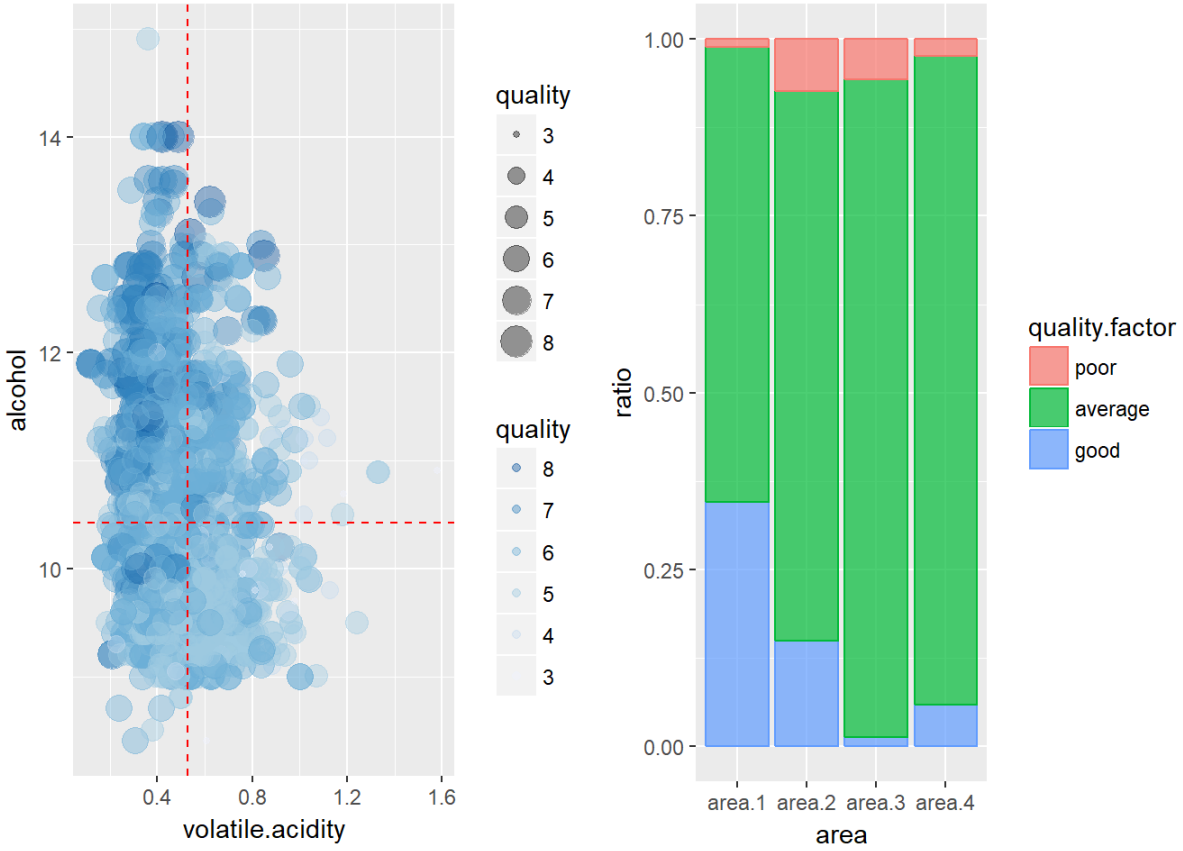
Multivariate Plots Section

By looking at individual features, I observed that higher percentage of alcohol, lower volatile acidity, and higher citric acid each leads to quality wines. I would like to investigate if this statement is still true when combination of these features is present in wines.

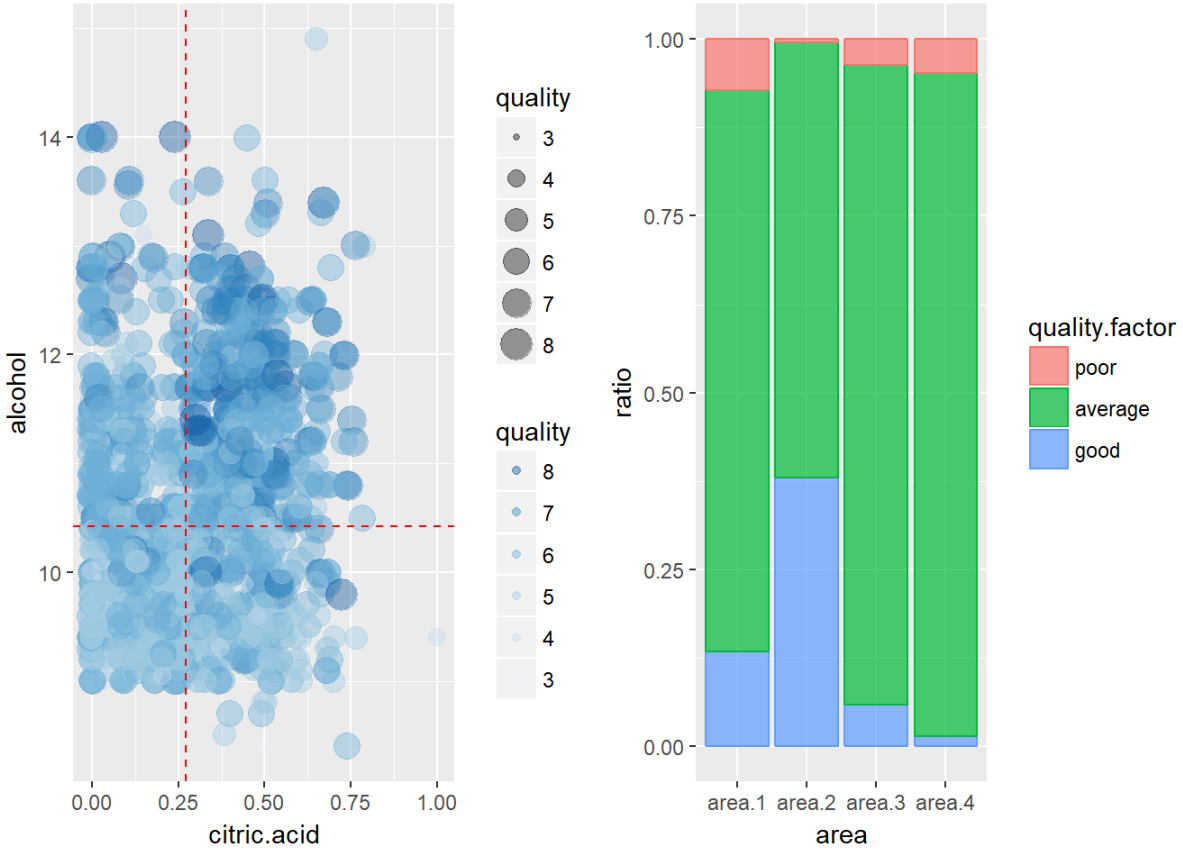
I am dividing areas of scatter plot to area 1, 2, 3, and 4.
Area 1 has values greater than its mean for y axis and less than its mean for x axis.
Area 2 has values greater than its mean for both x and y axis.
Area 3 has values less than its mean for y axis and greater than its mean for x axis.
Area 4 has values less than its mean for both x and y axis.

Also I'm defining -
Good Wine : Quality 7 & 8
Average Wine : Quality 5 & 6

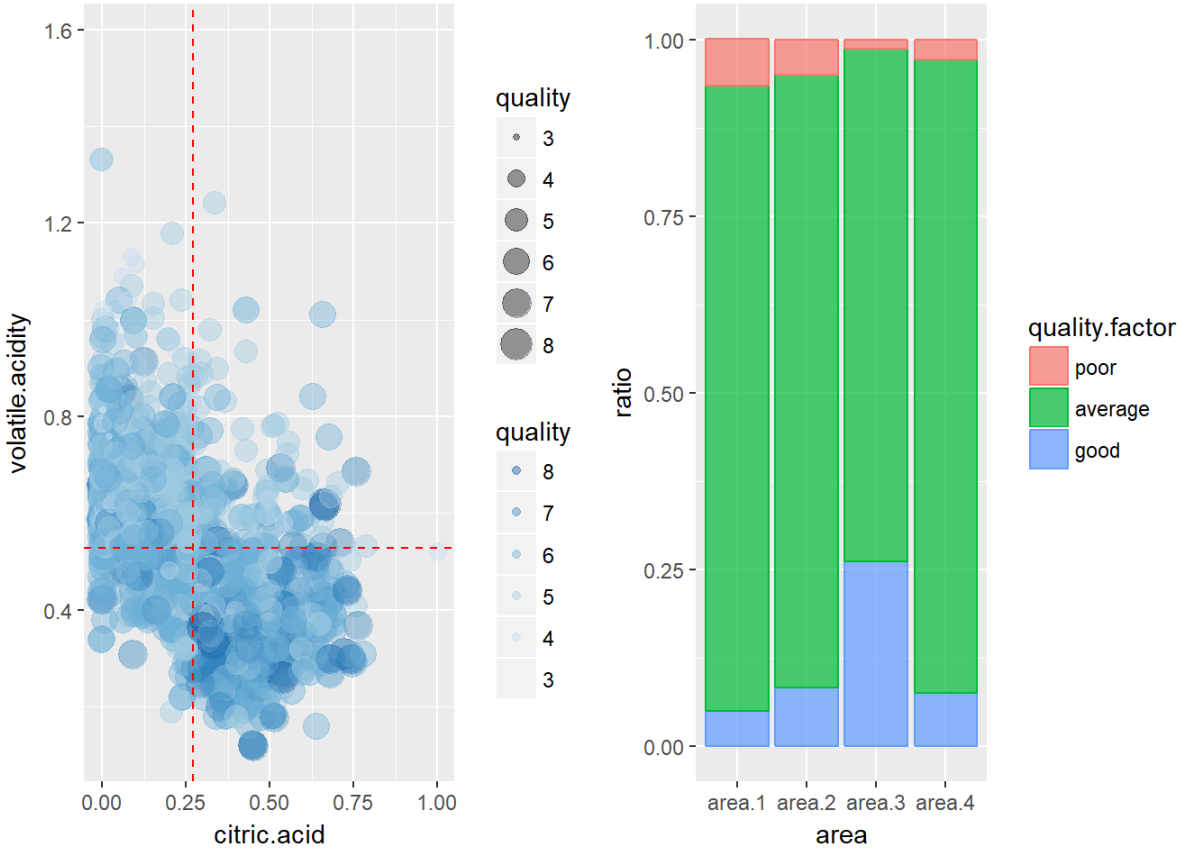
Poor Wine : Quality 3 & 4



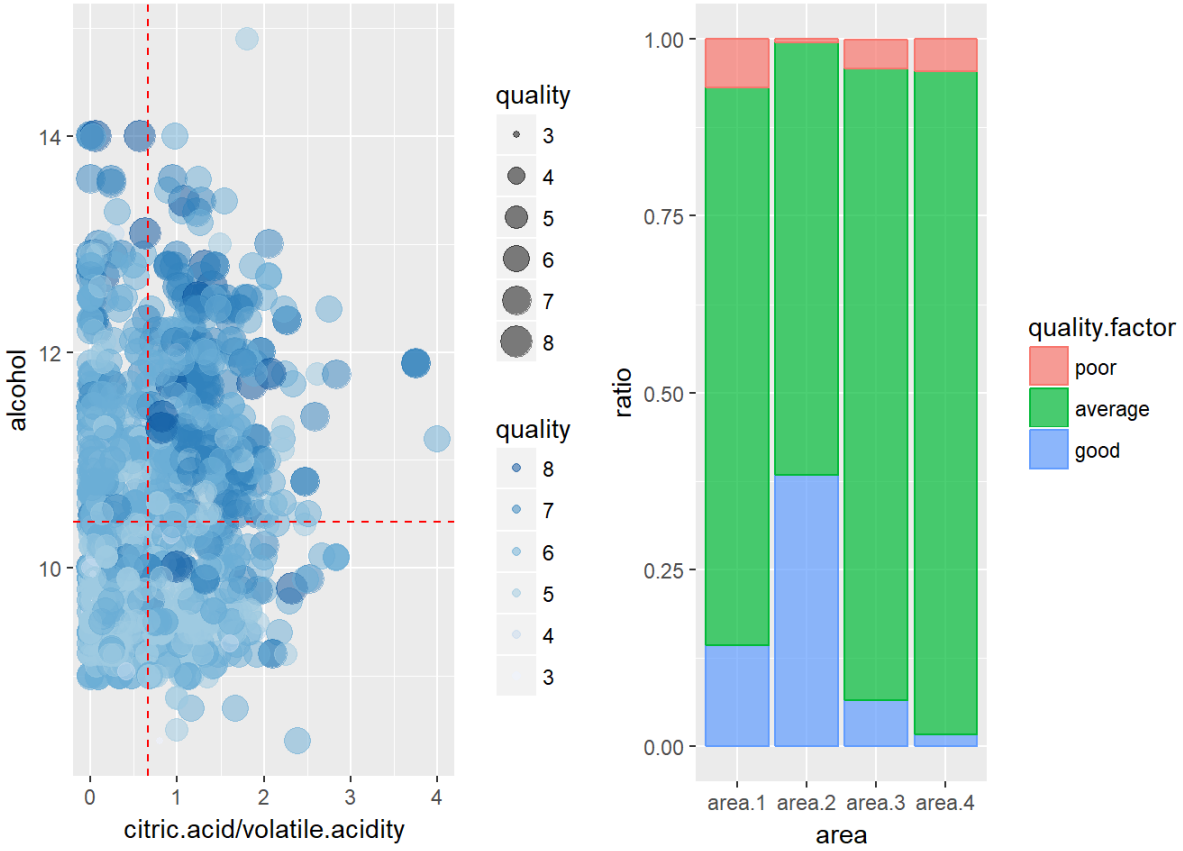
Wines with alcohol percentage greater than its mean and volatile acidity lower than its mean (Area 1) lean towards better quality. Wines with alcohol percentage less than its mean and volatile acidity greater than its mean (Area 3) lean towards better quality.



Wines with alcohol percentage greater than its mean and citric acid greater than its mean (Area 2) lean towards better quality.
Wines with alcohol percentage less than its mean and citric acid less than its mean (Area 4) lean towards poor quality.



Wines with citric acid greater than its mean and volatile acidity lower than its mean (Area 3) lean towards better quality.
Wines with citric acid lower than its mean and volatile acidity greater than its mean (Area 1) lean towards poor quality.



Note:
citric acid/volatile acidity > mean -> high citric acid & low volatile acidity
citric acid/volatile acidity < mean -> low citric acid & high volatile acidity

Wines with alcohol percentage greater than its mean and citric acid/volatile acidity greater than its mean (Area 2) lean towards better quality.
Wines with alcohol percentage less than its mean and citric acid/volatile acidity less than its mean (Area 4) lean towards poor 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?

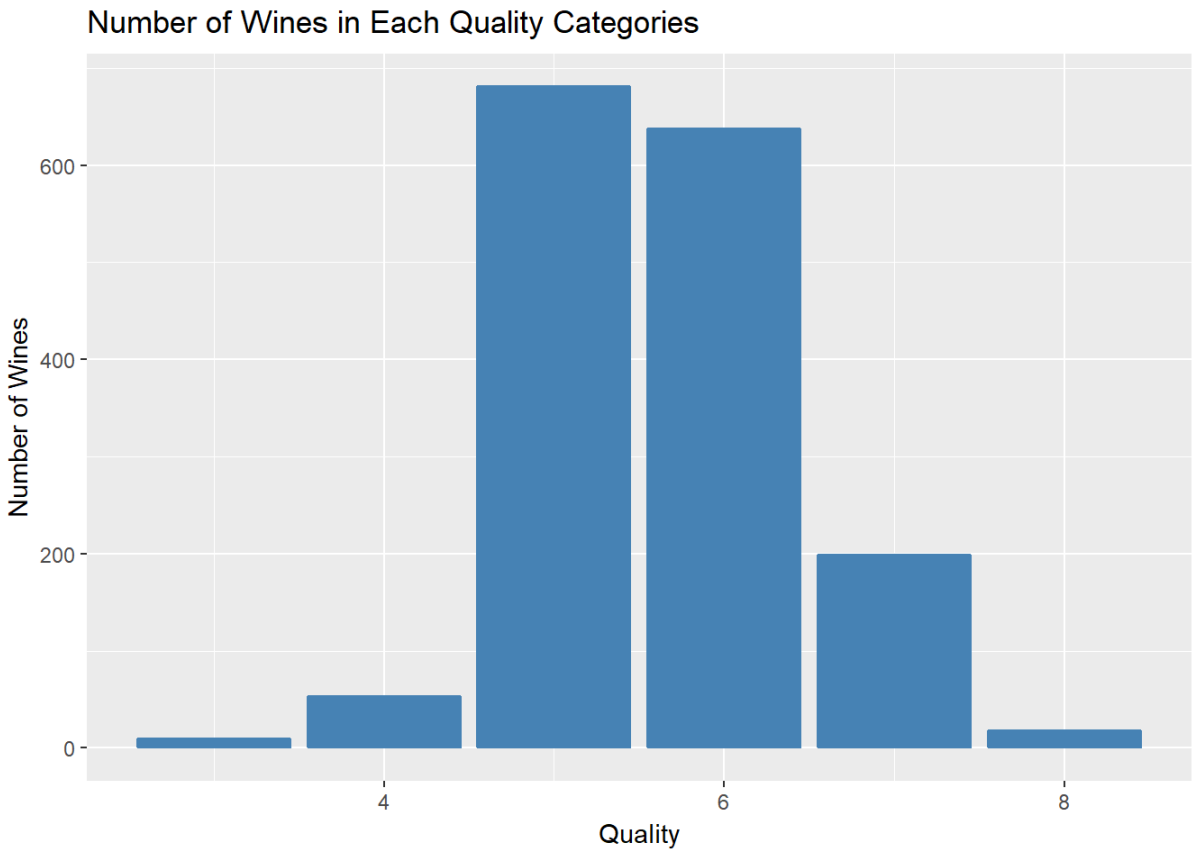
In Bivariate analysis, I observed that higher percentage of alcohol, lower volatile acidity, and higher citric acid each leads to better quality wines. In Multivariate analysis, I investigated combination of the features and see if wines have all 3 properties of features leads to quality wines. That seem to be true.

Were there any interesting or surprising interactions between features?

There was not anything unusual or surprising interactions between features.

Final Plots and Summary

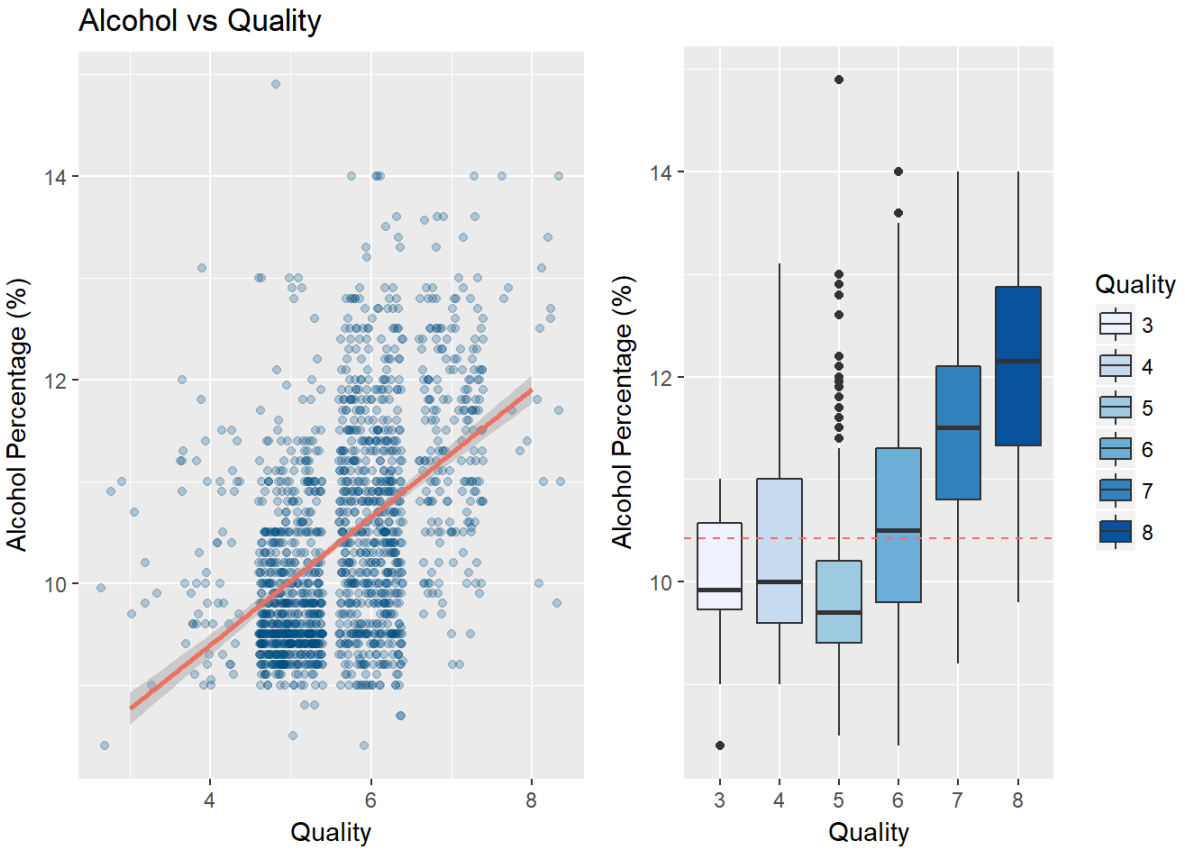
Plot One



Description One

The distribution of diamond prices appears to be normal. Most of average quality wine falls into wine quality category 5 and 6. We have very little data for wine category 3, 4, and 8.

Plot Two

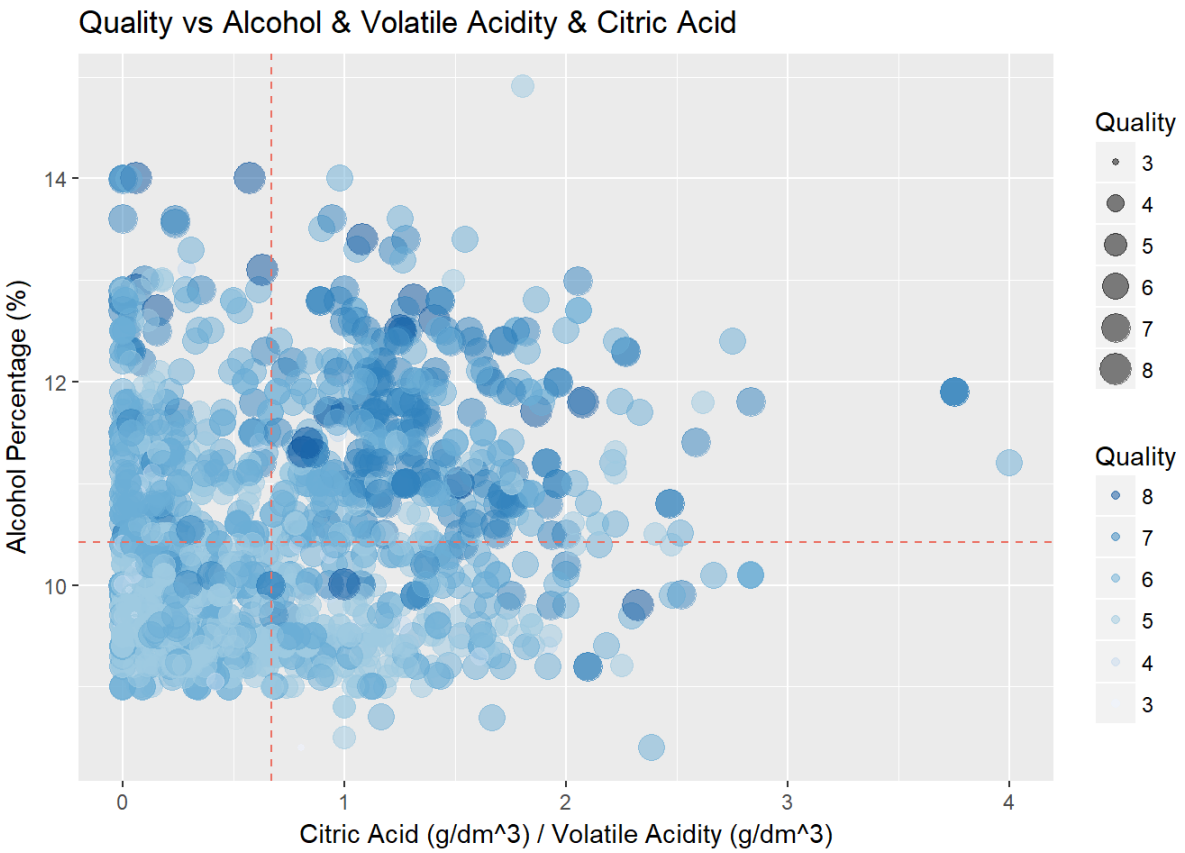


Description Two

Alcohol percentage is strongly correlated with wine quality among all variables. It has a positive relationship with quality.

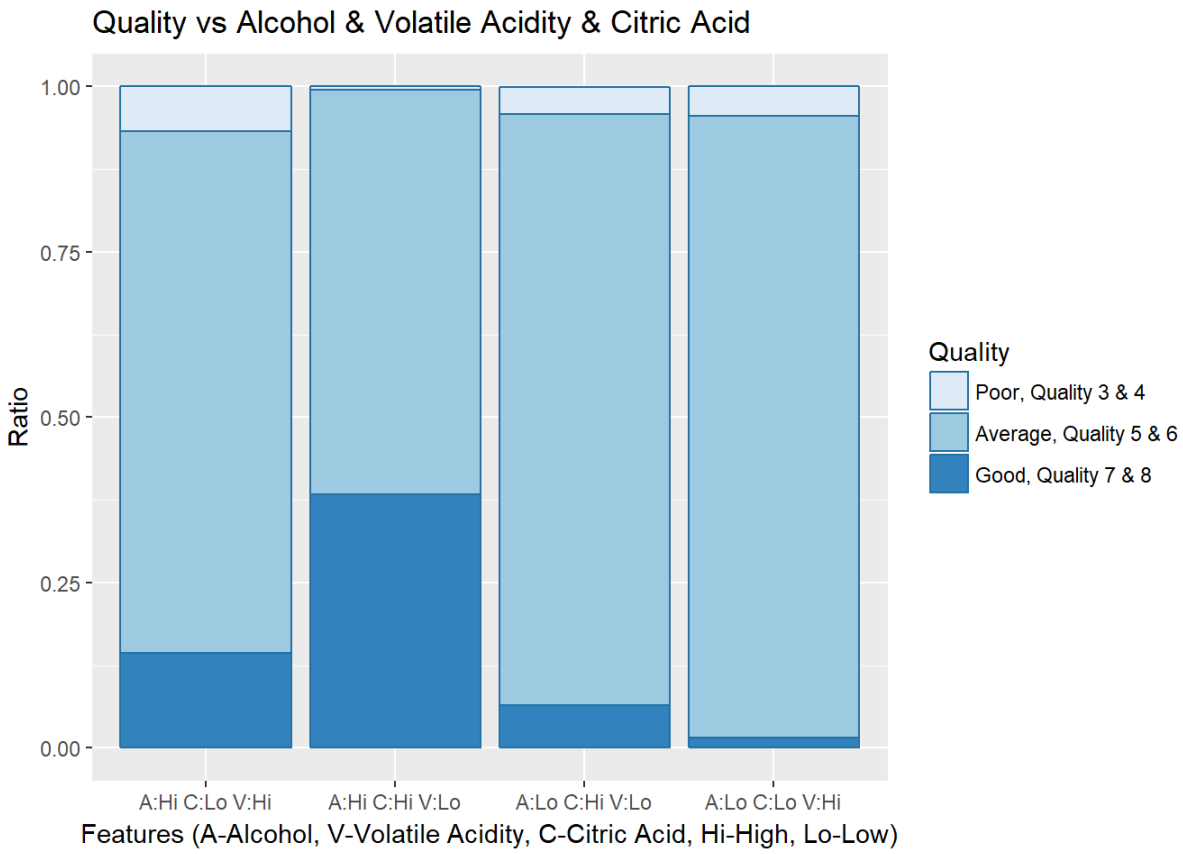
More than 50% of wines in quality category 6 and more than 75% wines in quality category 7 & 8 have alcohol percentage greater than its mean. The trend seems to be higher the alcohol percentage, better wine quality. However, quality category 5 has the lowest median alcohol percentage, narrowest interquartile range, and the most number of outliers. This might be because quality category 5 has the most number of wines out of all quality categories.

Plot Three



citric acid/volatile acidity > mean -> high citric acid & low volatile acidity

citric acid/volatile acidity < mean -> low citric acid & high volatile acidity



Description Three

Red wines which have higher alcohol percentage, higher citric acid, and lower volatile acidity lead to better wine quality. On the other hand, red wines with lower alcohol percentage, lower citric acid, and higher volatile acidity lead to poor wine quality.

Reflection

Red wine dataset has 1599 red wine data with 13 features. The question I wanted to answer was “Which and how features affect quality of wine?”

I first started by exploring each feature in dataset. Then I find out which features have relationship with quality of wine. Lastly, I investigated which combination of features and how they correlate with quatity of wine.

In the end, I was able to find, alcohol percentage, volatile acidity, and citric acid affect most to wine quality.

When I started invesigation, I had an assumption that sweetness (residual suger) and saliness (chlorides) of wine has something to do with wine quality, but I was wrong. To my surprise, they didn’t have strong correlation with wine quality.

I would have liked to have the dataset with more wine samples. Wine category 3 had 10 wine data, wine category 4 had 53, and wine category 8 and 18 data. These categories represent either poor or good quality, so I will be more confident with my findings if I were to do the same investigation with more samples.

Also, if the dataset included more information like year wine was produced, cost of wine, where it was produced, etc., I am sure I would find other features that might affect red wine quality.