Exploratory Data Analysis of the Wine Quality DataSet

Nasser Aloqayli

May 10, 2017

Introduction

The aim of this report is to explore a dataset containing information and attributes about the quality of wine samples. The dataset tagged "WineQuality" contains 12 attributes for about 1599 wines. This report contains 3 main sections:

1. Exploratory Analysis
2. Final Plots and Summary
3. Reflection

The exploratory analysis section contains the main exploratory analysis carried out accompanied by some contextual description of the thought process involved. The final plots and summary section contains more polished version of some plots from the exploratory analysis section. The reflection section will contain a summary of final thoughts about the analysis and ideas for future analysis.

Exploratory Analysis

Load Packages

We'll start this section by loading important packages that will be used during the analysis process. We'll load `ggplot2` package for data visualization and `dplyr` package for handling and wrangling our dataset

Data Inspection and Handling

Next we'll load the `Wine Quality Reds` data using the popular `read.csv()` function in R after which we'll inspect the data and compute some descriptives.

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

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

From the output above, we see that our data contains 13 variables on 1599 observation (wine samples). Furthermore, all of the 13 variables are numeric variables with the `quality` and `X` variables being integers.

Careful inspection shows that the `quality` variable contains wine quality ratings with values ranging from 3 to 8. This is an indication that the `quality` variable is actually a ordered factor variable, consequently, we'll create another factor variable called `quality.factor` that contains the `quality` variable converted to a factor.

```
## 'data.frame':    1599 obs. of  14 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 ...
##  $ quality.factor      : Factor w/ 6 levels "3","4","5","6",..: 3 3 3 4 3 3 3 5 5 3 ...
```

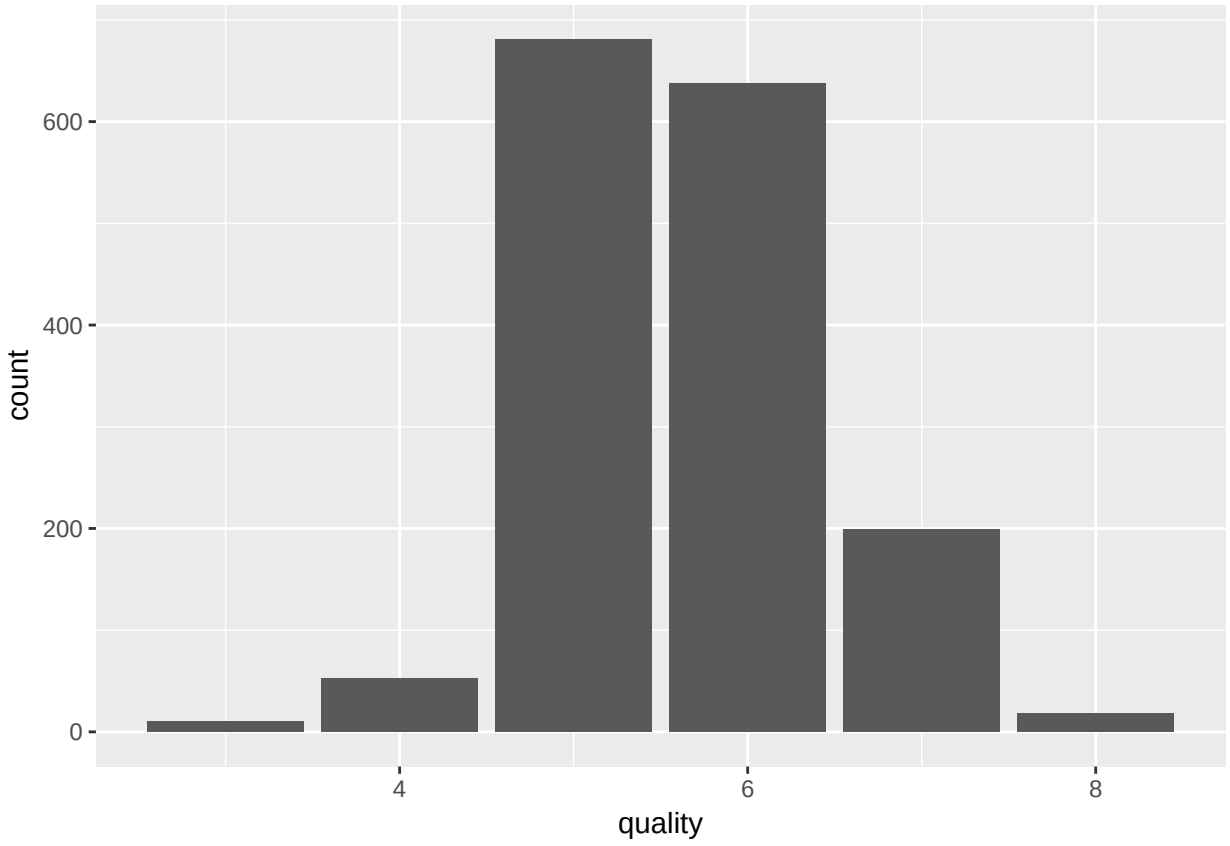
The `quality.factor` variable is now added to the data. Before we proceed to exploring the data with univariate plots, we'll check for missing values in our data

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

The winequality data does not contain any missing value which is a very good thing for our analysis.

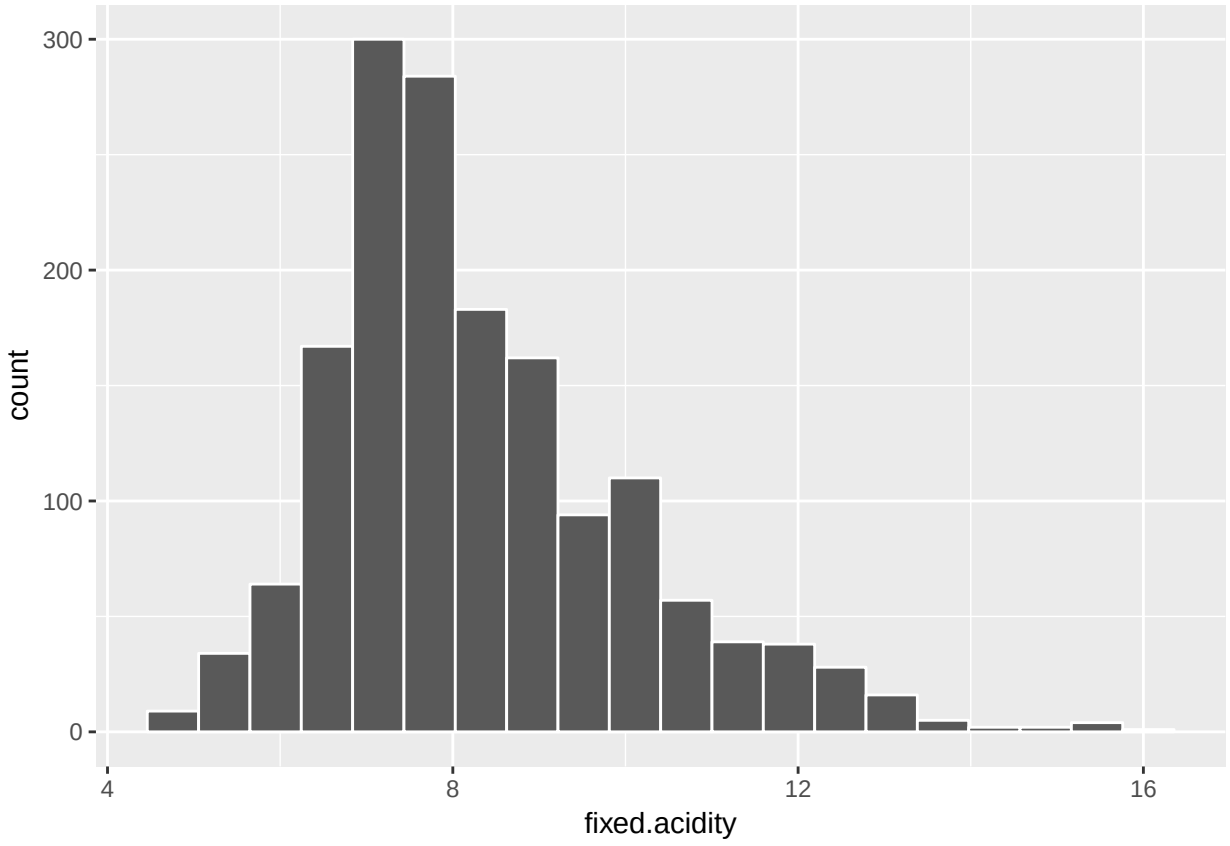
Univariate Plots

It's of interest to know the distribution of the wine quality



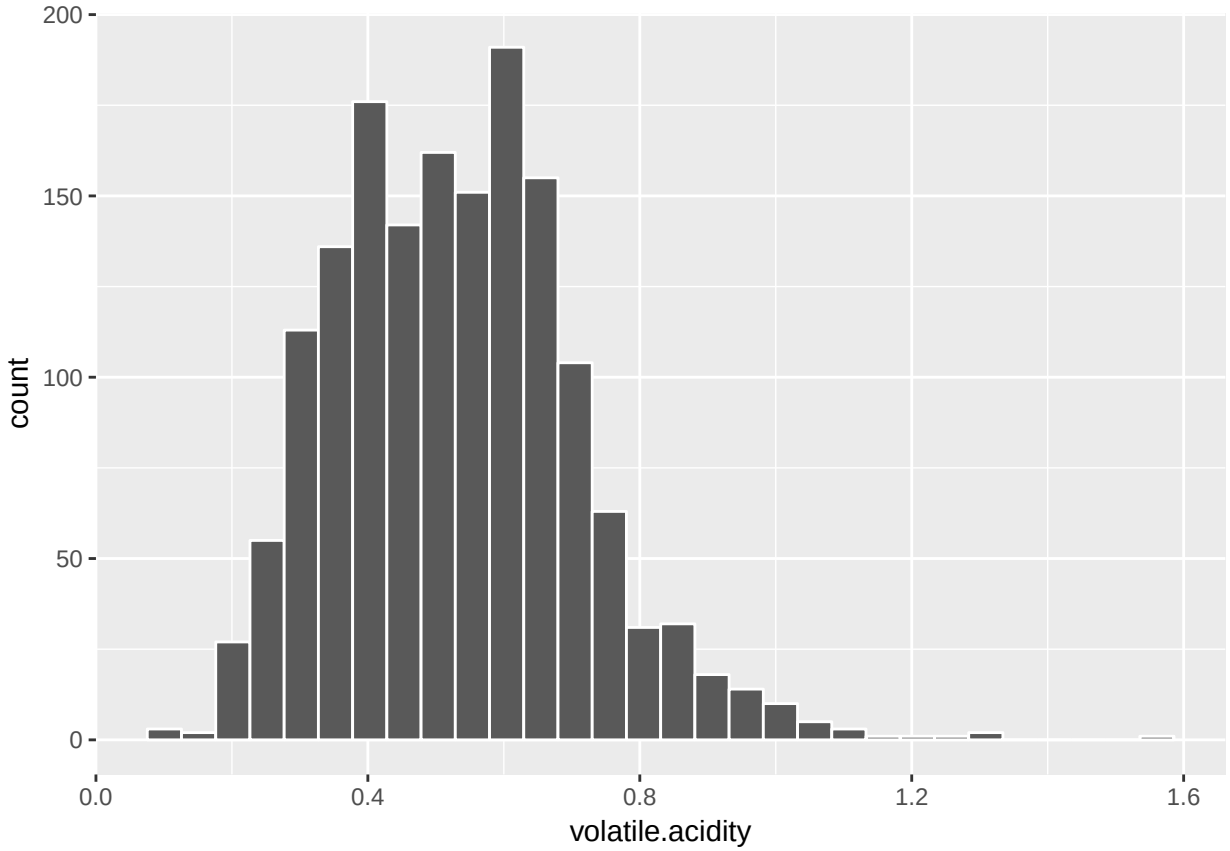

```
## quality
##   3   4   5   6   7   8
##  10  53 681 638 199  18
```

Most of the wines have a quality rating between 5 and 6 while few of the wines have quality rating of 8 or 3. The distribution of wine quality seems to follow normal distribution. Next we'll check the distribution of the other variables of the dataset



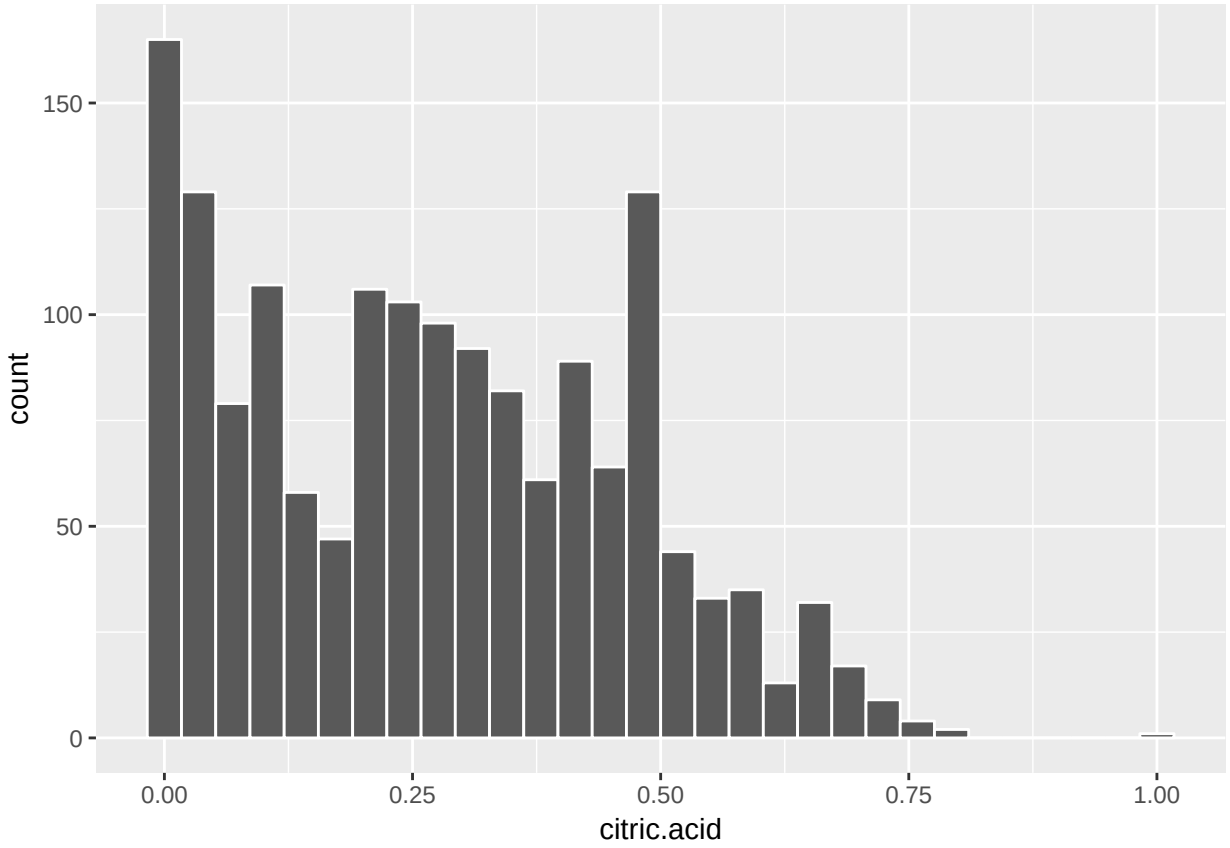

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
```

The distribution of fixed acidity is skewed to the right, consequently, the mean fixed acidity is greater than the median. The distribution of volatile acidity is also shown below



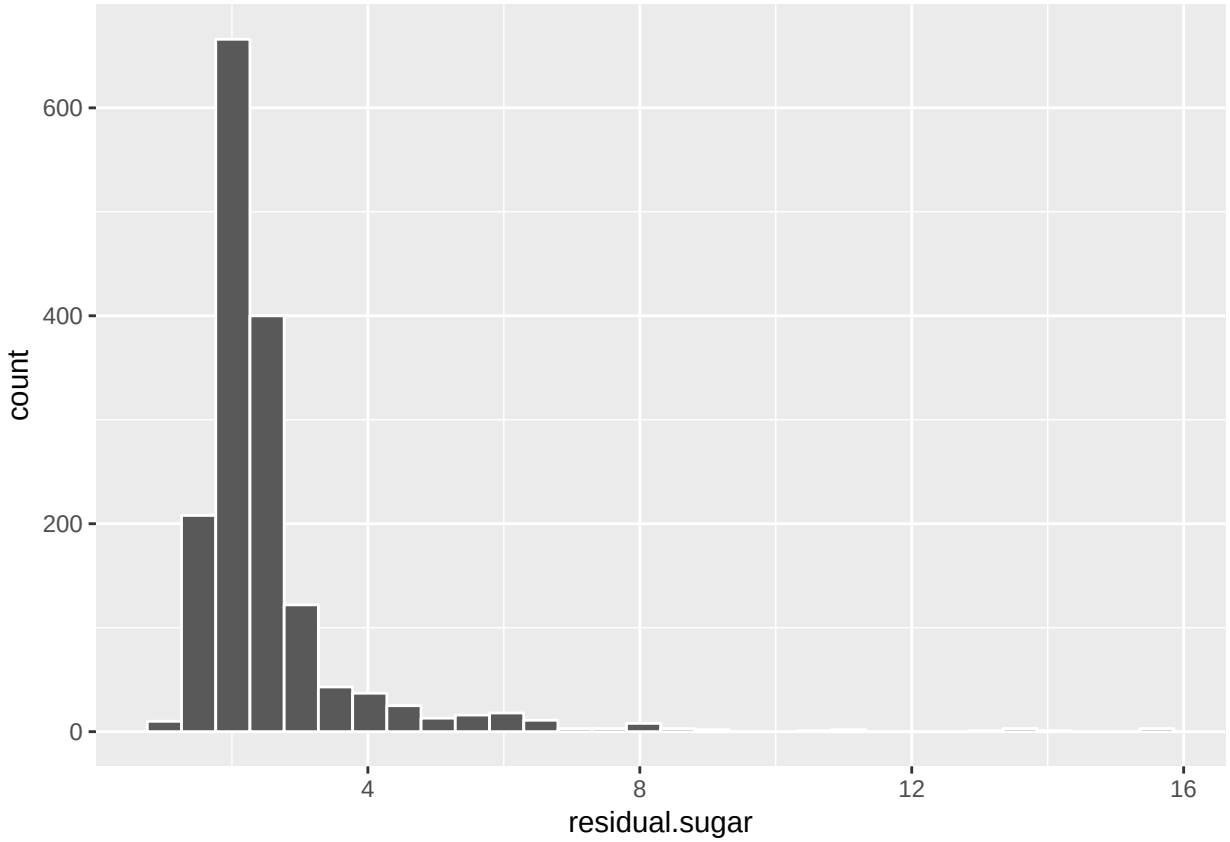

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```

The distribution of volatile acidity is slightly skewed to the right.



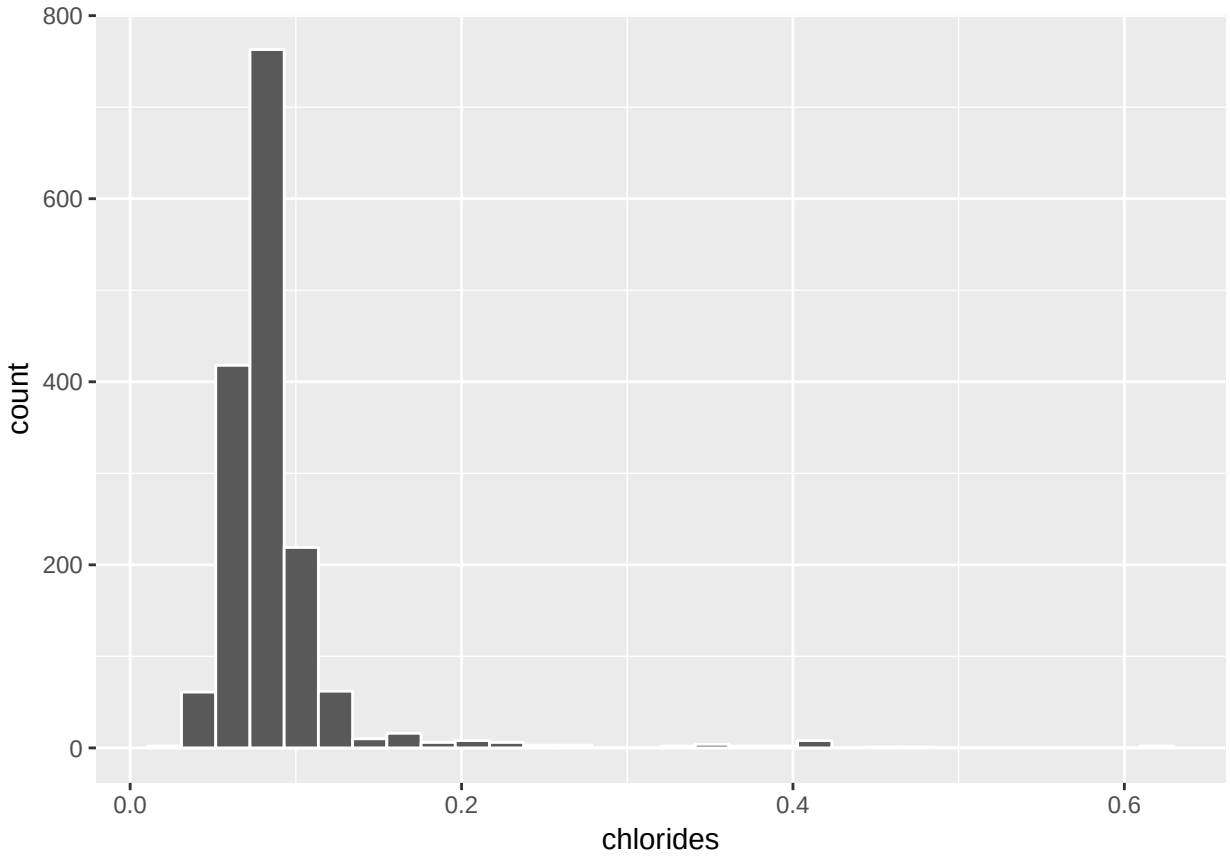

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```

The distribution of citric acid concentration has a long positive tail and it seems to be a bimodal distribution. Also worthy of note is the unusual spike at 0.00 concentration and 0.5 concentration. This indicates that lot of wine contained in the dataset have 0.00 and 0.5 citric acid concentration. Due to the right skewed distribution, the mean of citric acid concentration is greater than the median.



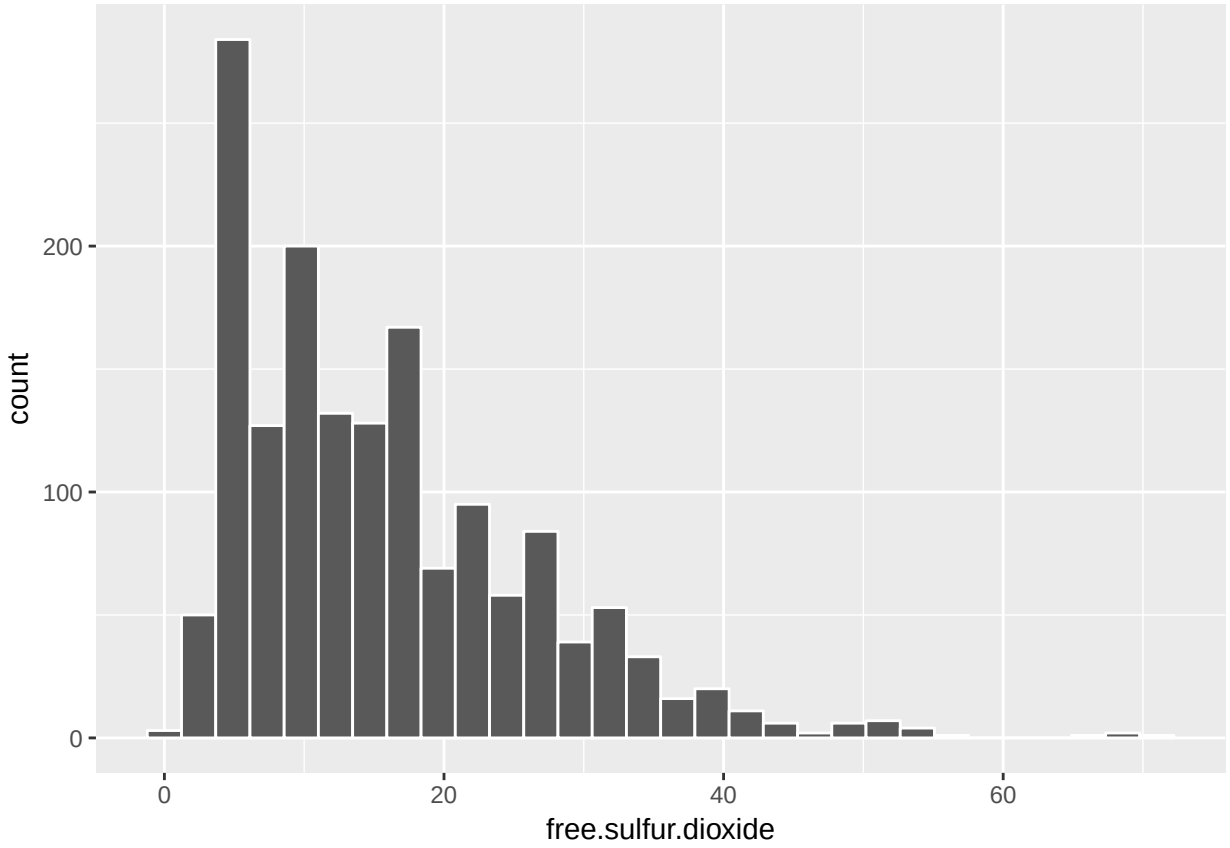

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```

The distribution of residual sugar content of the wines is unimodal with a long right tail (right skew). Most of the wines have a residual sugar content of about 2. As expected of a right skewed distribution, the mean is greater than the median.

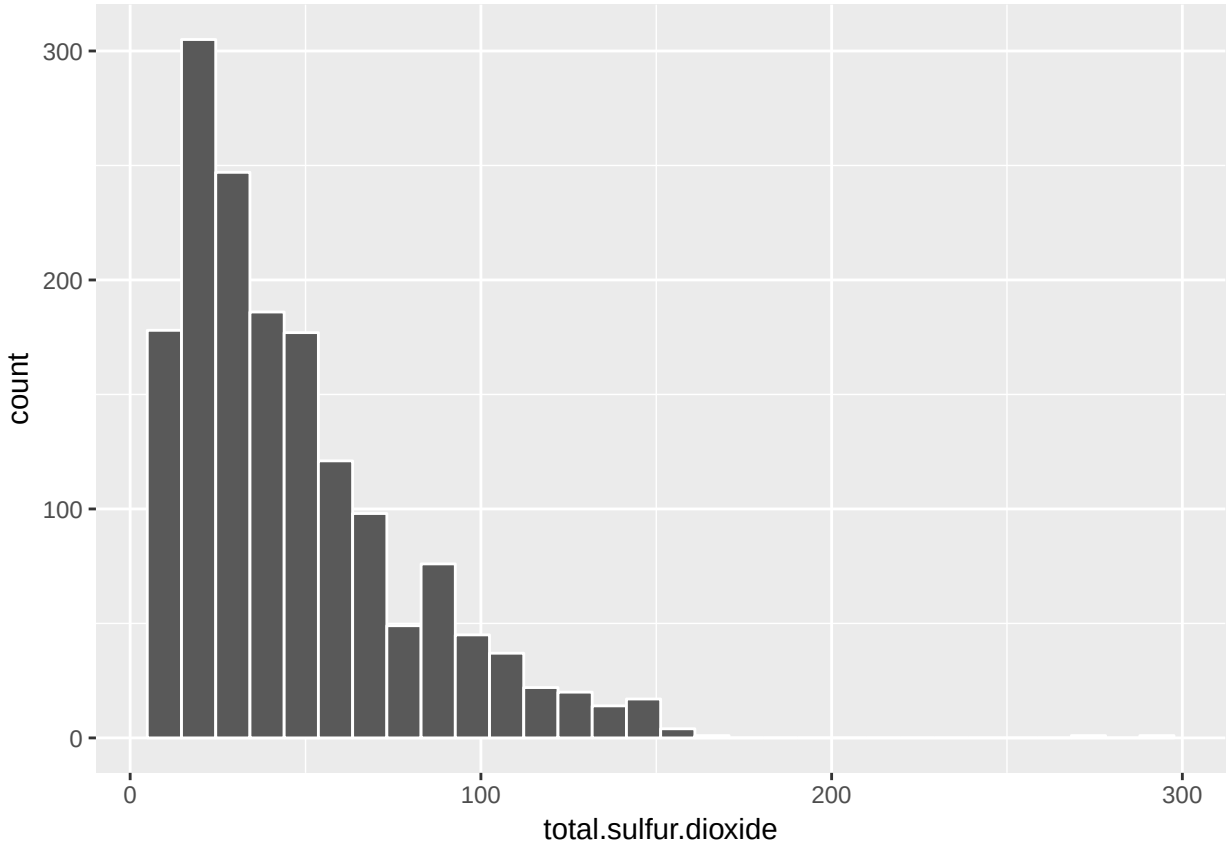



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```

The distribution of chlorides is quite similar to that of the residual sugar content with a unimodal right skewed distribution. Most of the wines have chloride contents of about $0.1 g/dm^3$. The wine with the lowest chloride content has $0.012 g/dm^3$ of chloride while the wine with maximum chloride content has $0.611 g/dm^3$ of chloride.

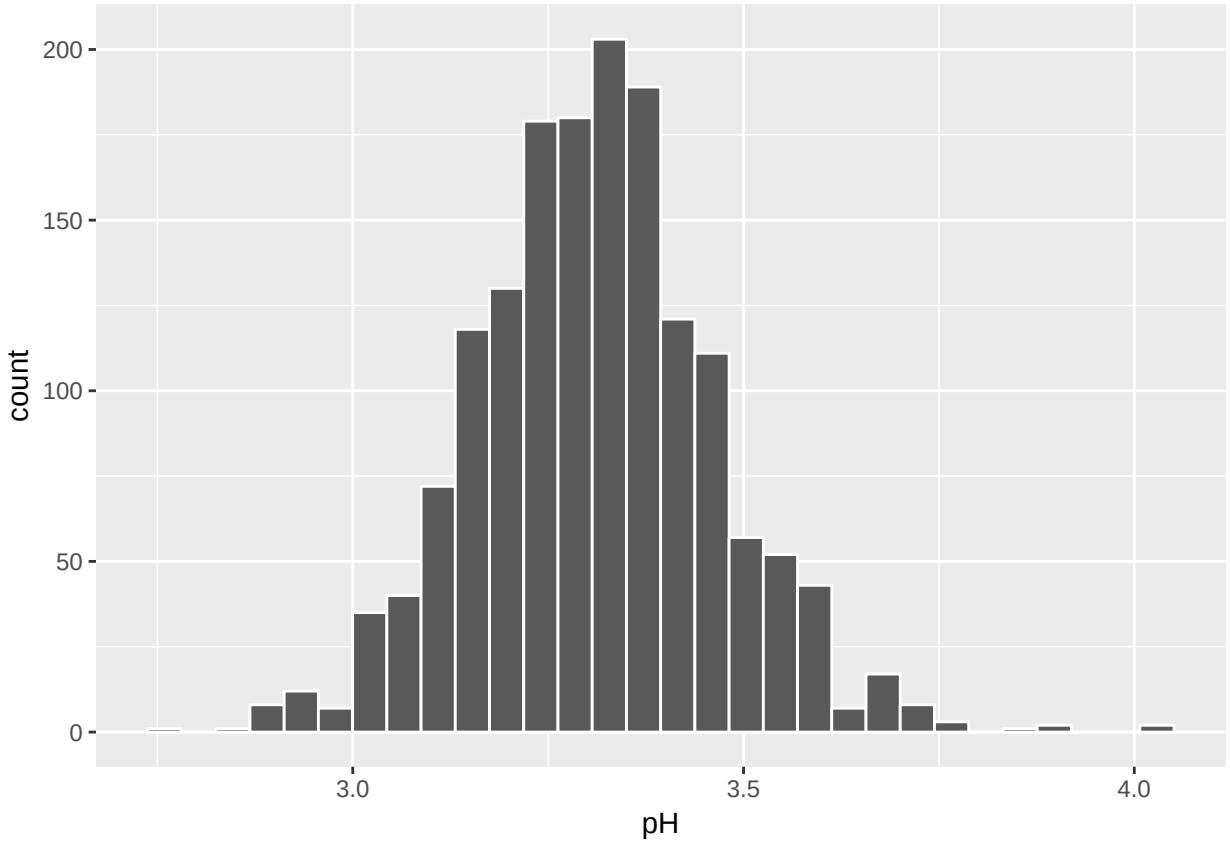



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
```

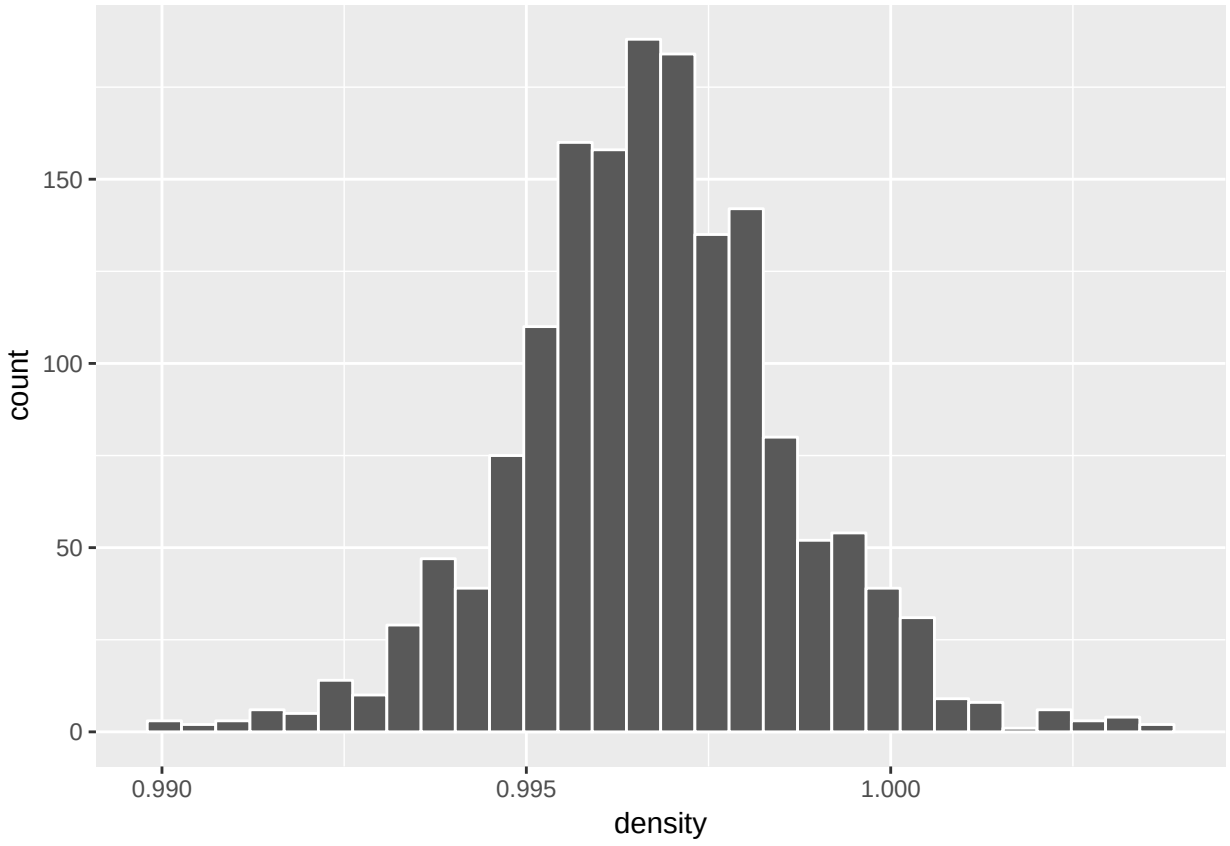



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00
```

The distribution of free and total sulfur dioxide of the wines are both skewed to the right. Next we'll have a look at the distributions of pH and density of the wines

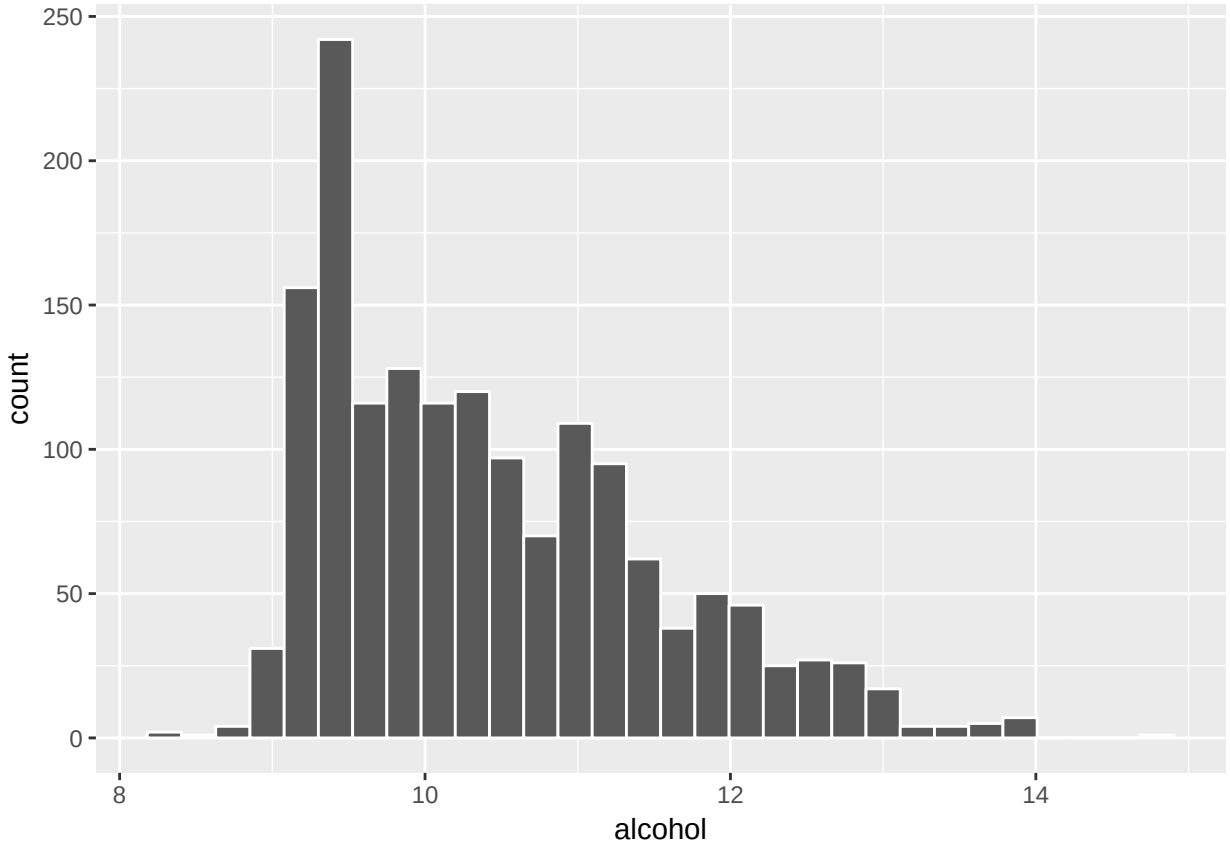



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010
```

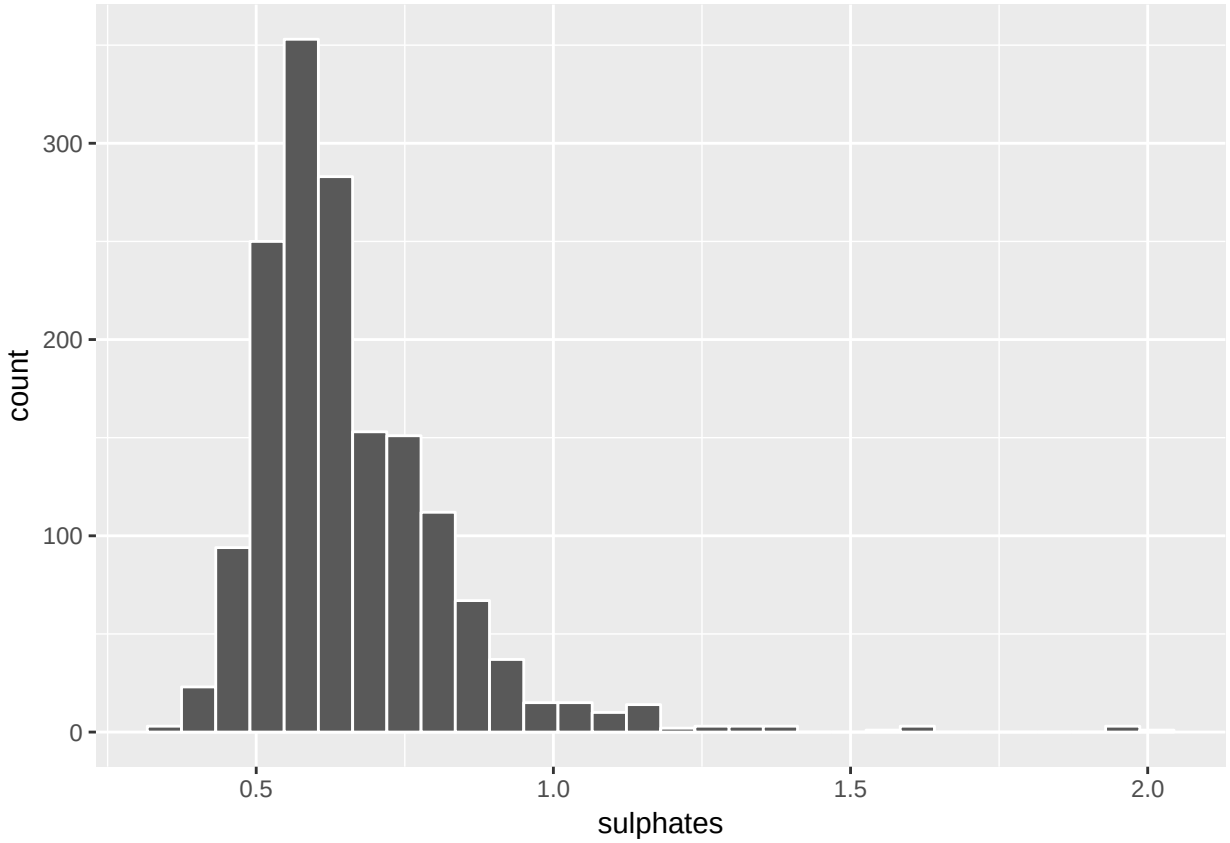



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0037
```

Surprisingly, the density and pH values of the wines follow a normal distribution with mean values 3.311 and 0.9967 for pH and density respectively. Finally, we'll have a look at the distribution of alcohol content and sulphates of the wines




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```

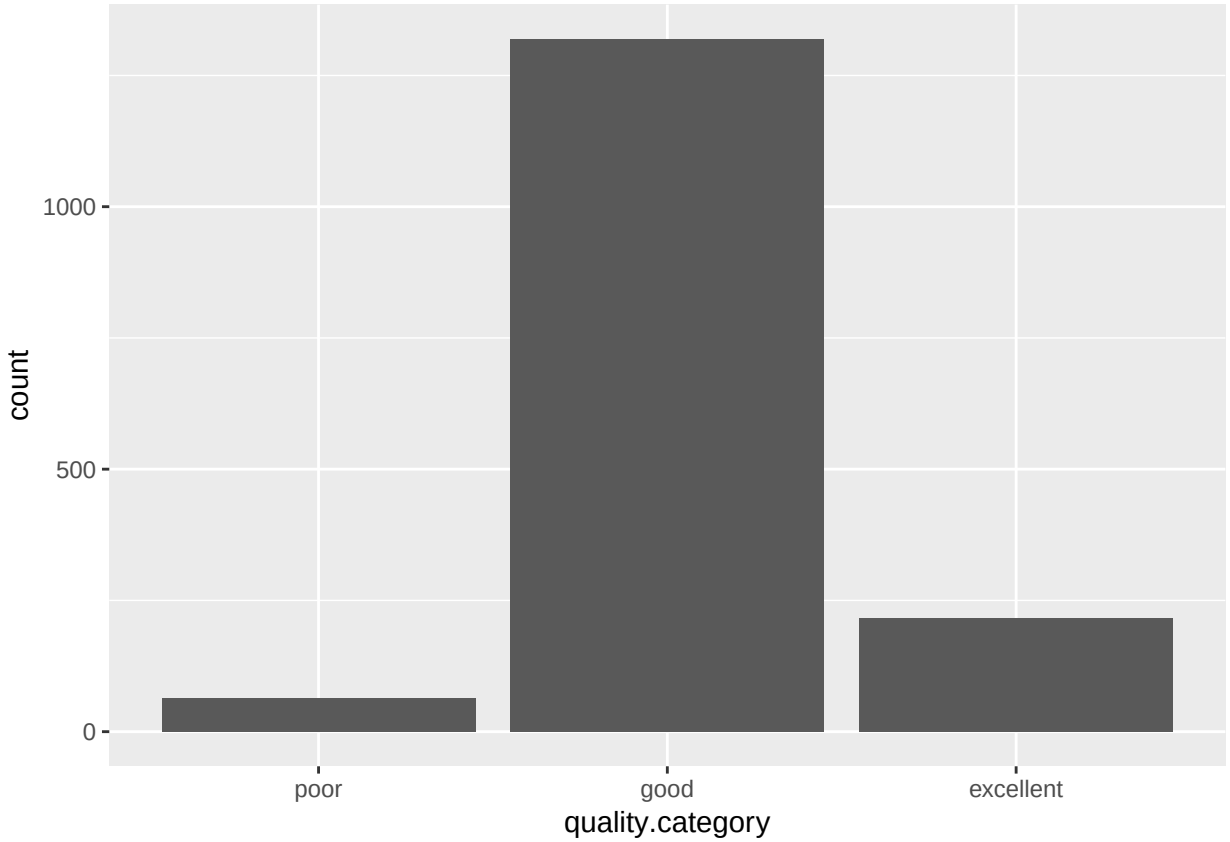



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```

Bivariate Plots

Here we'll try to explore relationships between attributes of the wines contained in the dataset with particular interest in wine quality. For the purpose of the this section, we'll create a new variable `quality.category` which indicates whether the quality of a wine is `poor` (quality rating between 3 and 4), `good` (quality rating between 5 and 6) and `excellent` (quality rating between 7 and 8).

The distribution of `quality.category` is shown below

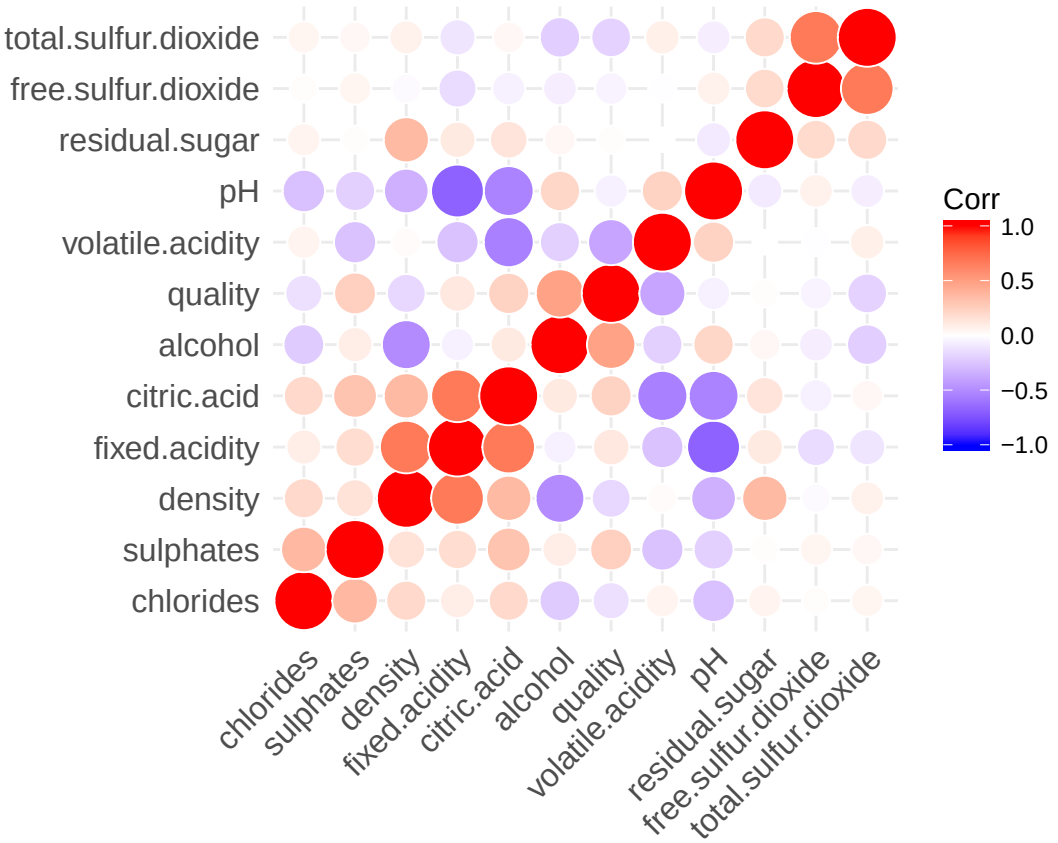


To proceed with bivariate plots, we'll have a look at the correlation matrix of the the attributes of the data.

```
##                       fixed.acidity volatile.acidity citric.acid
## fixed.acidity           1.00000000     -0.256130895  0.67170343
## volatile.acidity       -0.25613089      1.000000000 -0.55249568
## citric.acid             0.67170343     -0.552495685  1.00000000
## residual.sugar          0.11477672      0.001917882  0.14357716
## chlorides               0.09370519      0.061297772  0.20382291
## free.sulfur.dioxide    -0.15379419     -0.010503827 -0.06097813
## total.sulfur.dioxide   -0.11318144      0.076470005  0.03553302
## density                 0.66804729      0.022026232  0.36494718
## pH                     -0.68297819      0.234937294 -0.54190414
## sulphates               0.18300566     -0.260986685  0.31277004
## alcohol                -0.06166827     -0.202288027  0.10990325
## quality                 0.12405165     -0.390557780  0.22637251
##                       residual.sugar    chlorides free.sulfur.dioxide
## fixed.acidity           0.114776724  0.093705186       -0.153794193
## volatile.acidity        0.001917882  0.061297772       -0.010503827
## citric.acid             0.143577162  0.203822914       -0.060978129
## residual.sugar          1.000000000  0.055609535        0.187048995
## chlorides               0.055609535  1.000000000        0.005562147
## free.sulfur.dioxide     0.187048995  0.005562147        1.000000000
## total.sulfur.dioxide    0.203027882  0.047400468        0.667666450
## density                 0.355283371  0.200632327       -0.021945831
## pH                     -0.085652422 -0.265026131        0.070377499
## sulphates               0.005527121  0.371260481        0.051657572
## alcohol                 0.042075437 -0.221140545       -0.069408354
```

```
## quality                 0.013731637 -0.128906560        -0.050656057
##                      total.sulfur.dioxide     density          pH
## fixed.acidity                -0.11318144  0.66804729 -0.68297819
## volatile.acidity              0.07647000  0.02202623  0.23493729
## citric.acid                   0.03553302  0.36494718 -0.54190414
## residual.sugar                0.20302788  0.35528337 -0.08565242
## chlorides                     0.04740047  0.20063233 -0.26502613
## free.sulfur.dioxide           0.66766645 -0.02194583  0.07037750
## total.sulfur.dioxide          1.00000000  0.07126948 -0.06649456
## density                       0.07126948  1.00000000 -0.34169933
## pH                           -0.06649456 -0.34169933  1.00000000
## sulphates                     0.04294684  0.14850641 -0.19664760
## alcohol                      -0.20565394 -0.49617977  0.20563251
## quality                      -0.18510029 -0.17491923 -0.05773139
##                         sulphates     alcohol     quality
## fixed.acidity         0.183005664 -0.06166827  0.12405165
## volatile.acidity     -0.260986685 -0.20228803 -0.39055778
## citric.acid           0.312770044  0.10990325  0.22637251
## residual.sugar        0.005527121  0.04207544  0.01373164
## chlorides             0.371260481 -0.22114054 -0.12890656
## free.sulfur.dioxide   0.051657572 -0.06940835 -0.05065606
## total.sulfur.dioxide  0.042946836 -0.20565394 -0.18510029
## density               0.148506412 -0.49617977 -0.17491923
## pH                   -0.196647602  0.20563251 -0.05773139
## sulphates             1.000000000  0.09359475  0.25139708
## alcohol               0.093594750  1.00000000  0.47616632
## quality               0.251397079  0.47616632  1.00000000
```

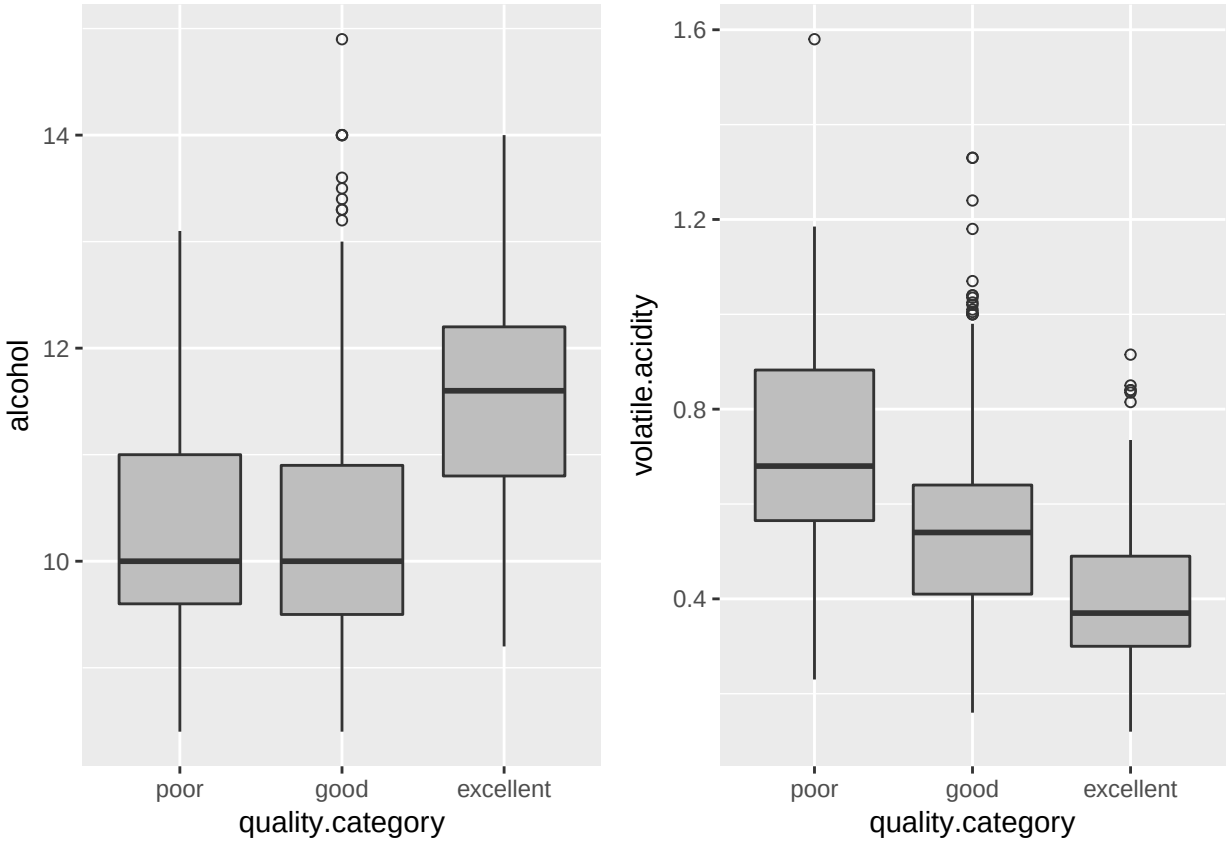
The above matrix can be better visualised with the correlation heat map below



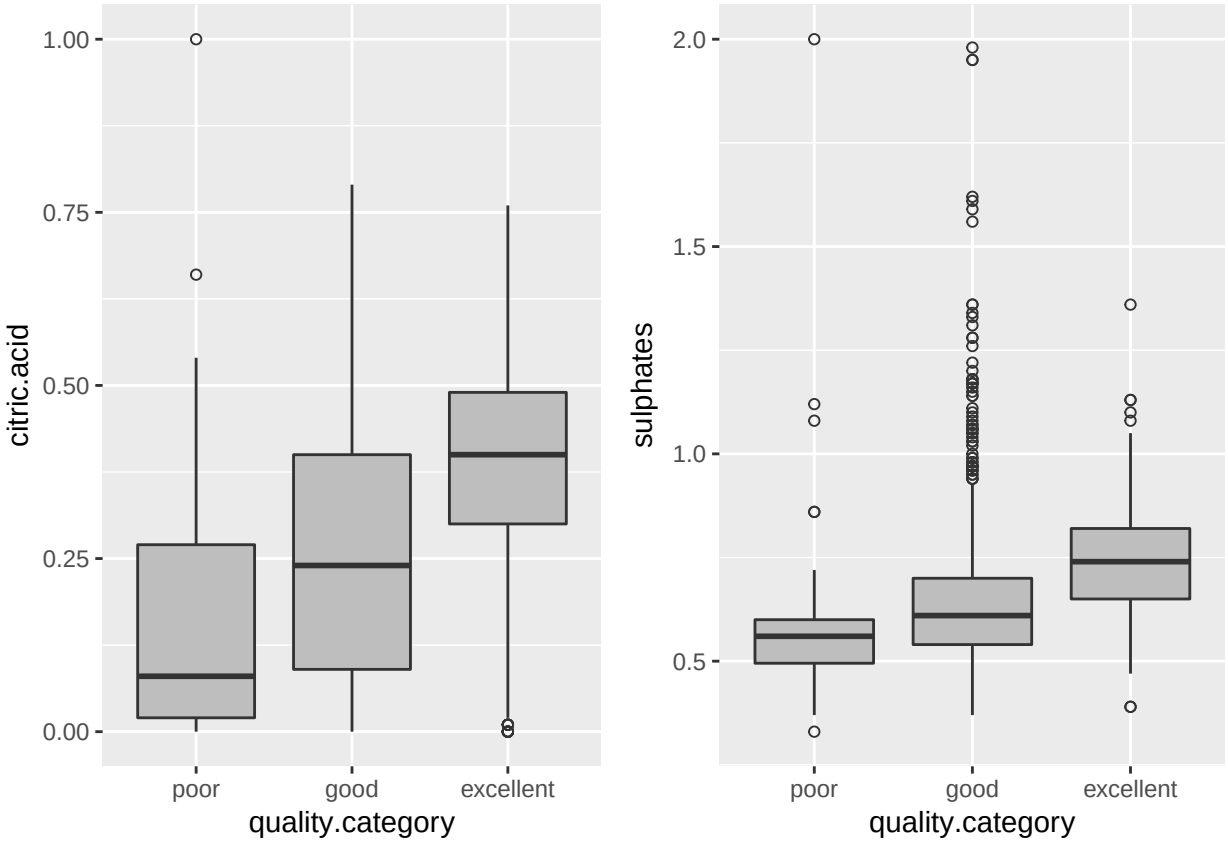
From the correlation heat map above, we notice that

- wine quality is fairly and positively correlated with alcohol content and negatively correlated with volatile acidity.
- total and free sulfur dioxide are highly correlated with each other positively.
- fixed acidity is positively correlated with density and citric acid and negatively correlated with pH.
- citric is negatively correlated with volatile acidity and pH.

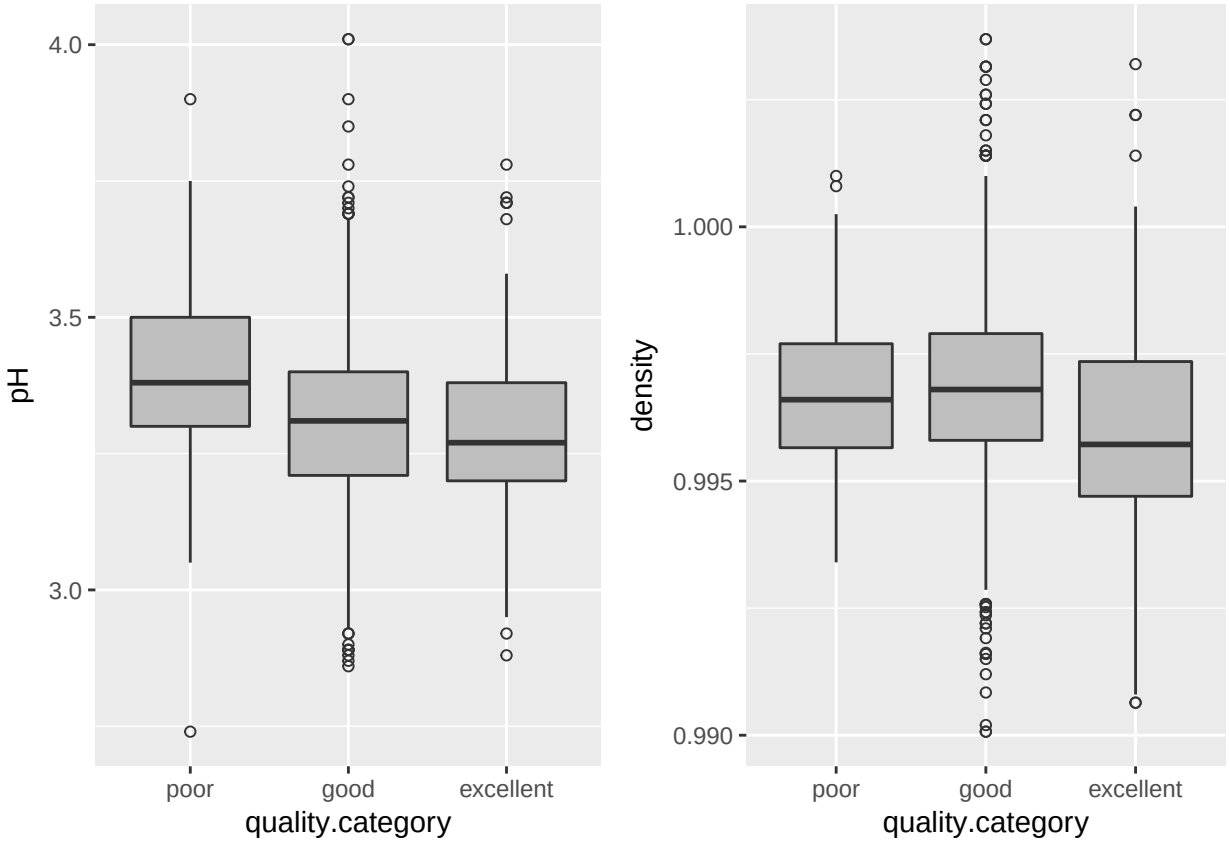
From the observation above about the correlation between alcohol content and wine quality with volatile acidity, one wonders if alcohol contents and volatile acidity differ across the three quality categories.



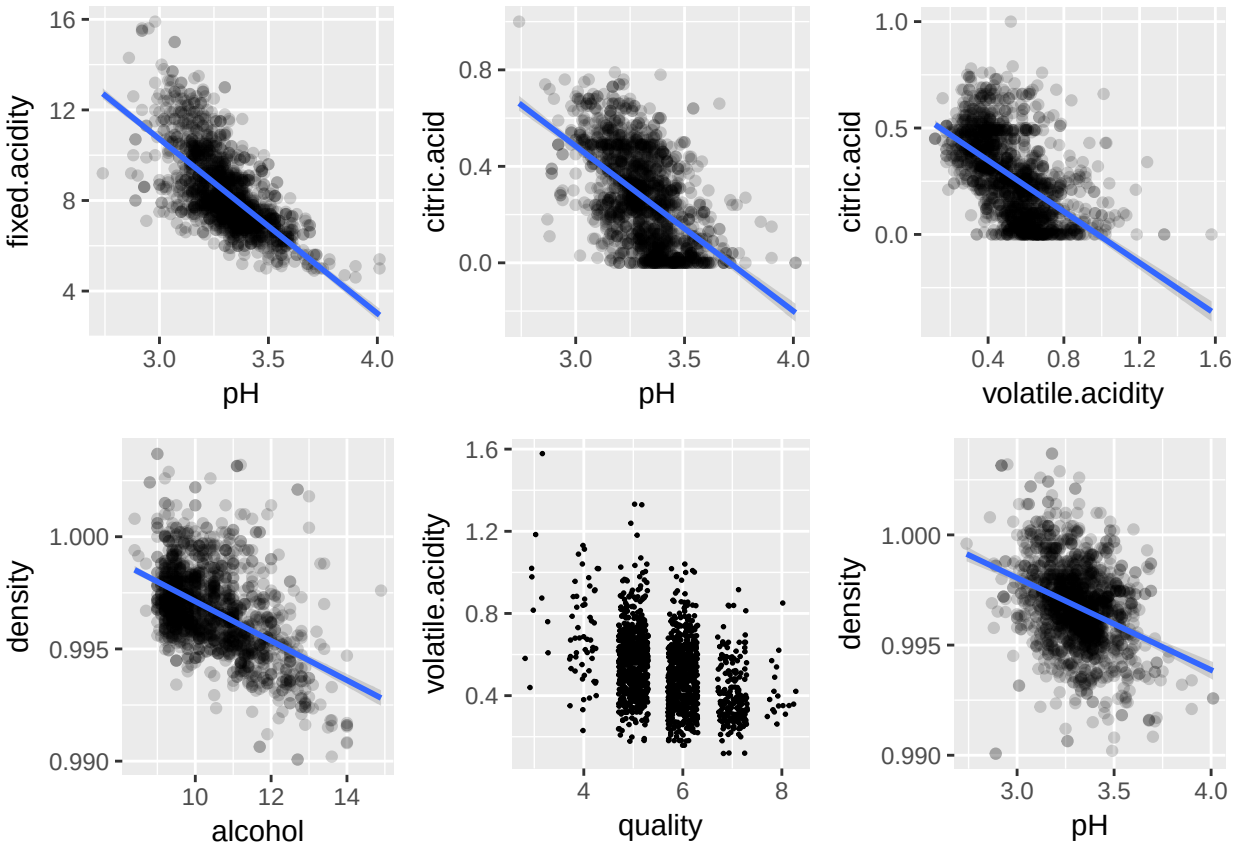
And the plot above suggests that alcohol content are higher in wines of excellent quality while volatile acidity are much lower in wines with excellent quality. Furthermore, citric acid and sulphates seem to be positively correlated (albeit a weak correlation) with wine quality. So we check if those also significantly differ among quality categories



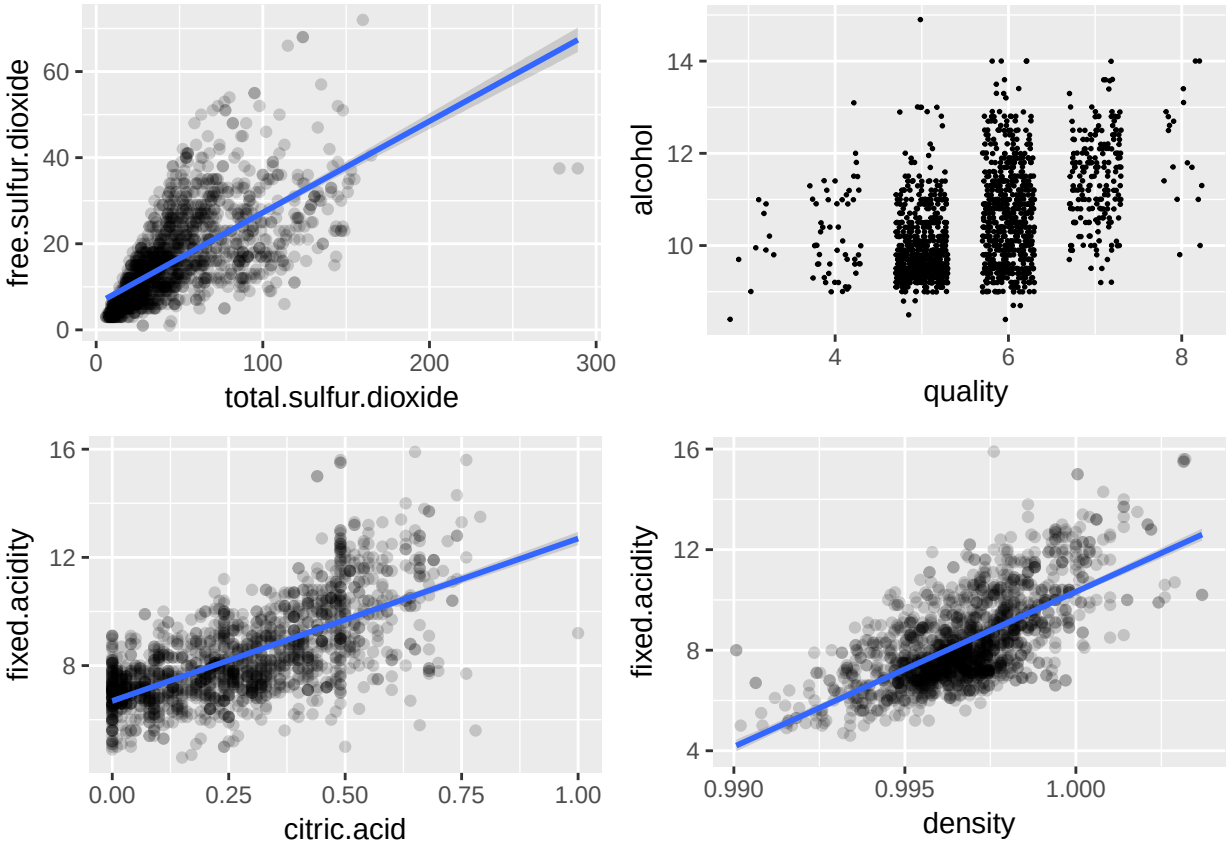
And we see from the boxplots that citric acid and sulphates seem to increase with wine quality with highest levels observed among wines of excellent quality. Unlike citric acid and sulphates; pH levels, and density seem to decrease with increase in wine quality



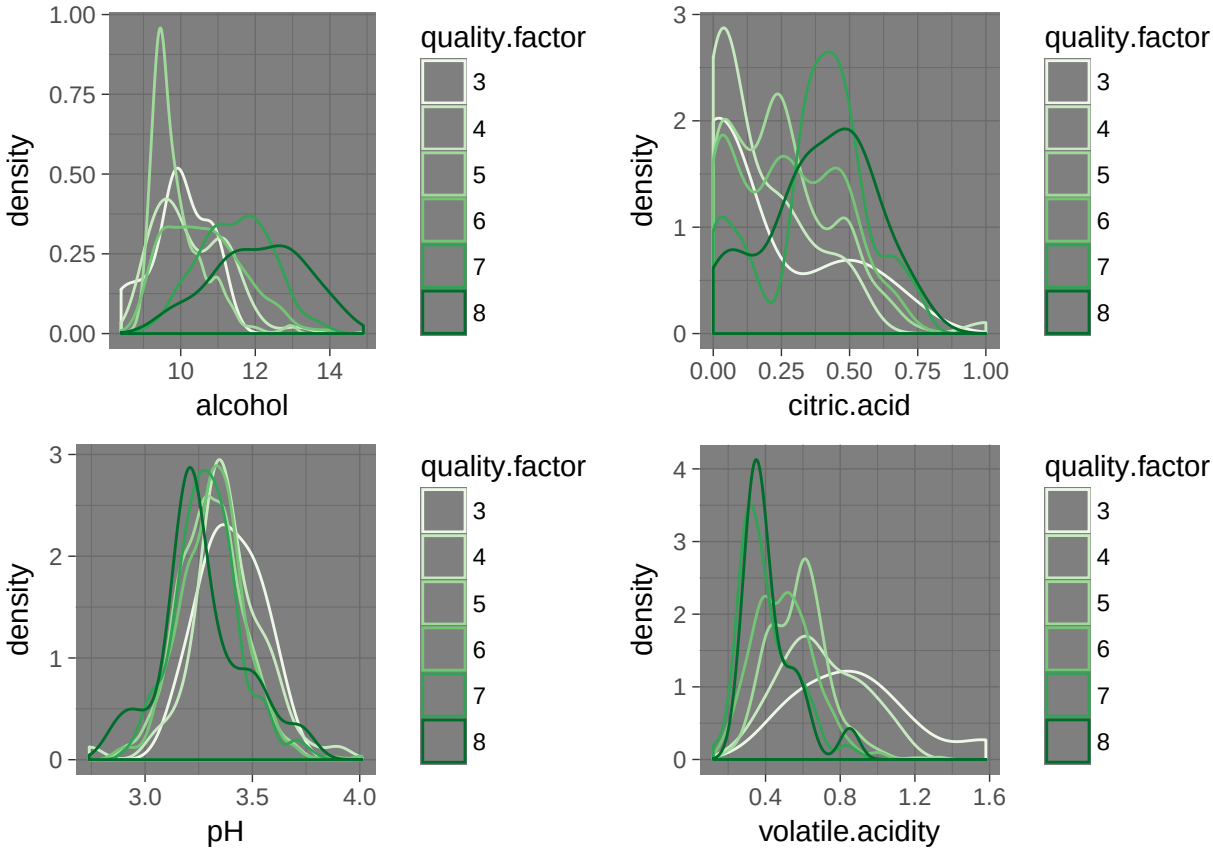
The following plot contains scatter plots of variables with the highest negative correlations as shown on the correlation plot



Furthermore, the following plot contains scatter plots of variables with the highest positive correlations as shown on the correlation plot



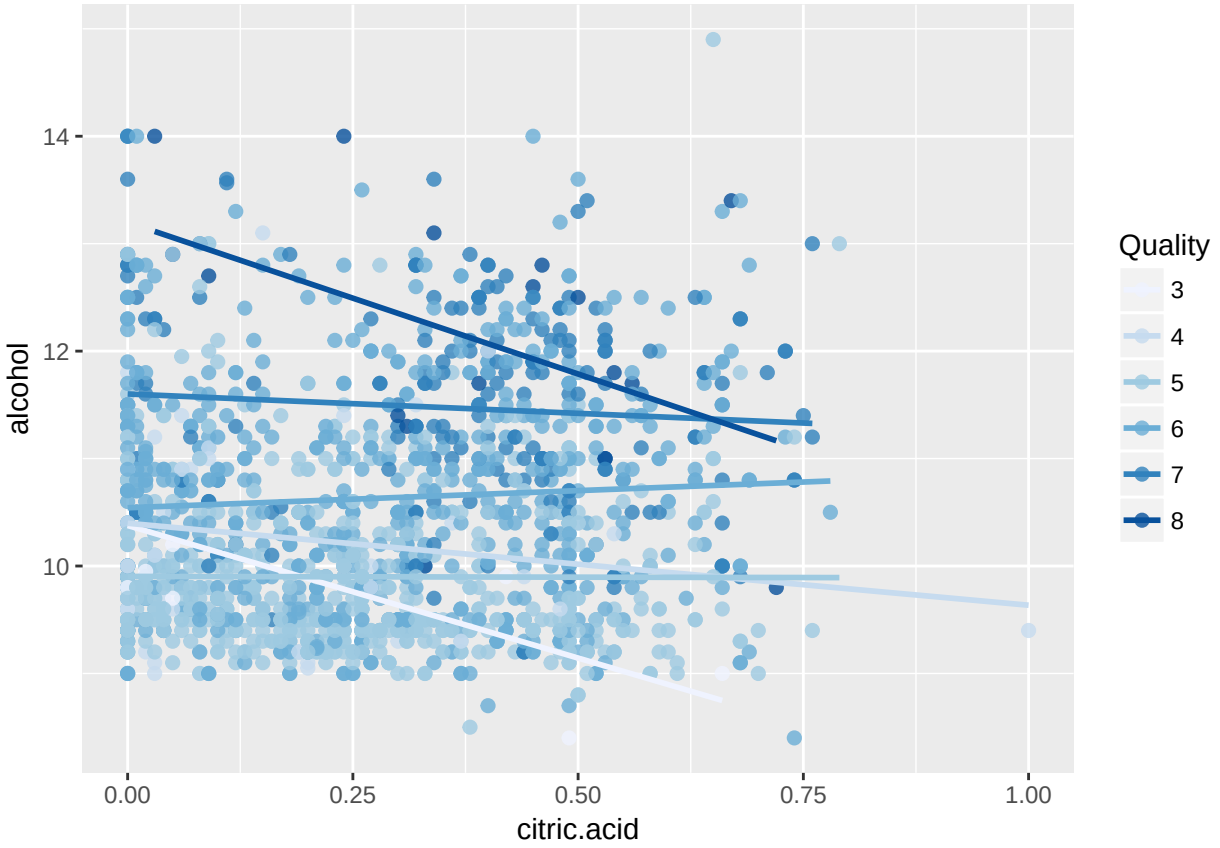
Since alcohol content, volatile acidity, pH levels and citric acid all seem to differ among different wine qualities, it's of interest to see their distribution grouped by different wine quality ratings



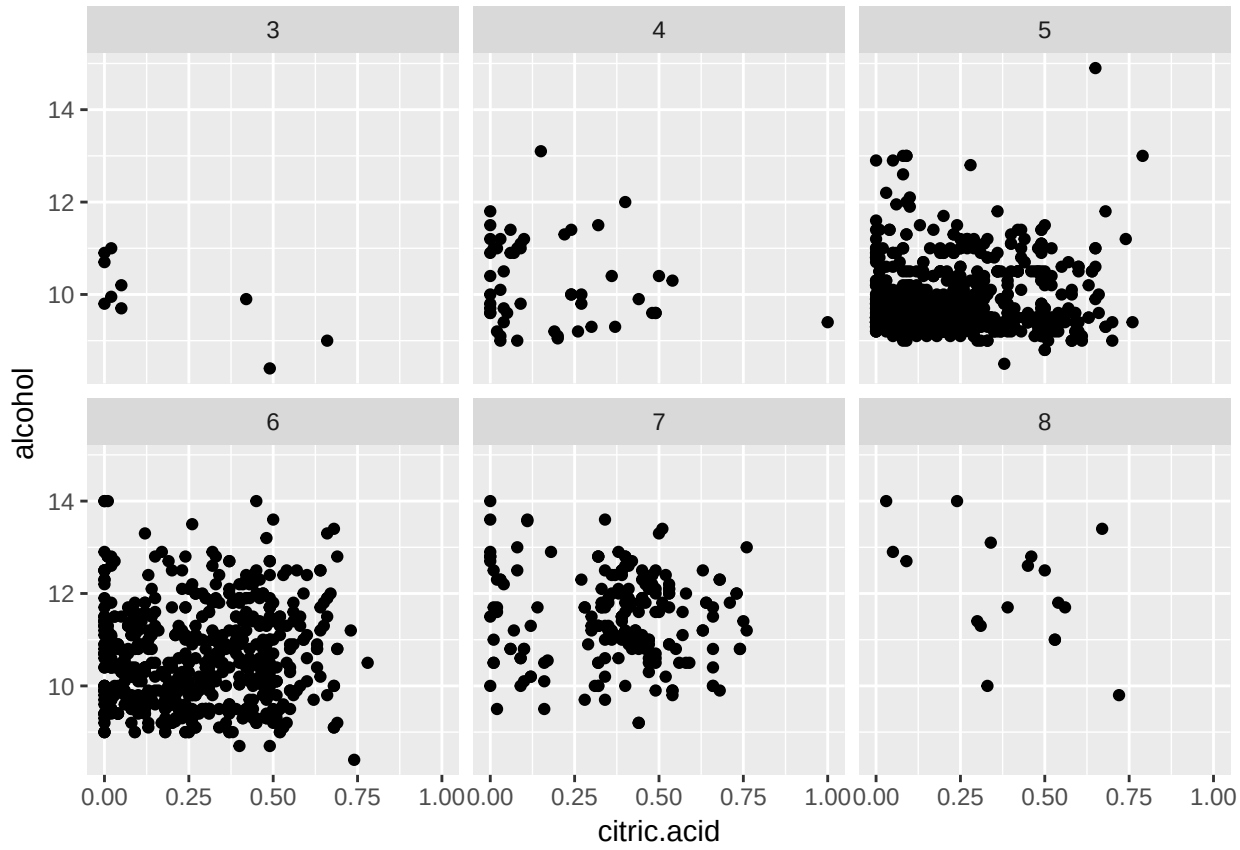
From the plots above, it is obvious that excellent quality wines have a higher alcohol and citric acid contents coupled lower volatile acidity and pH levels. Poor quality wines on the other hand have the least alcohol and citric acid contents coupled with the highest pH and volatile acidity levels.

Multivariate Plots

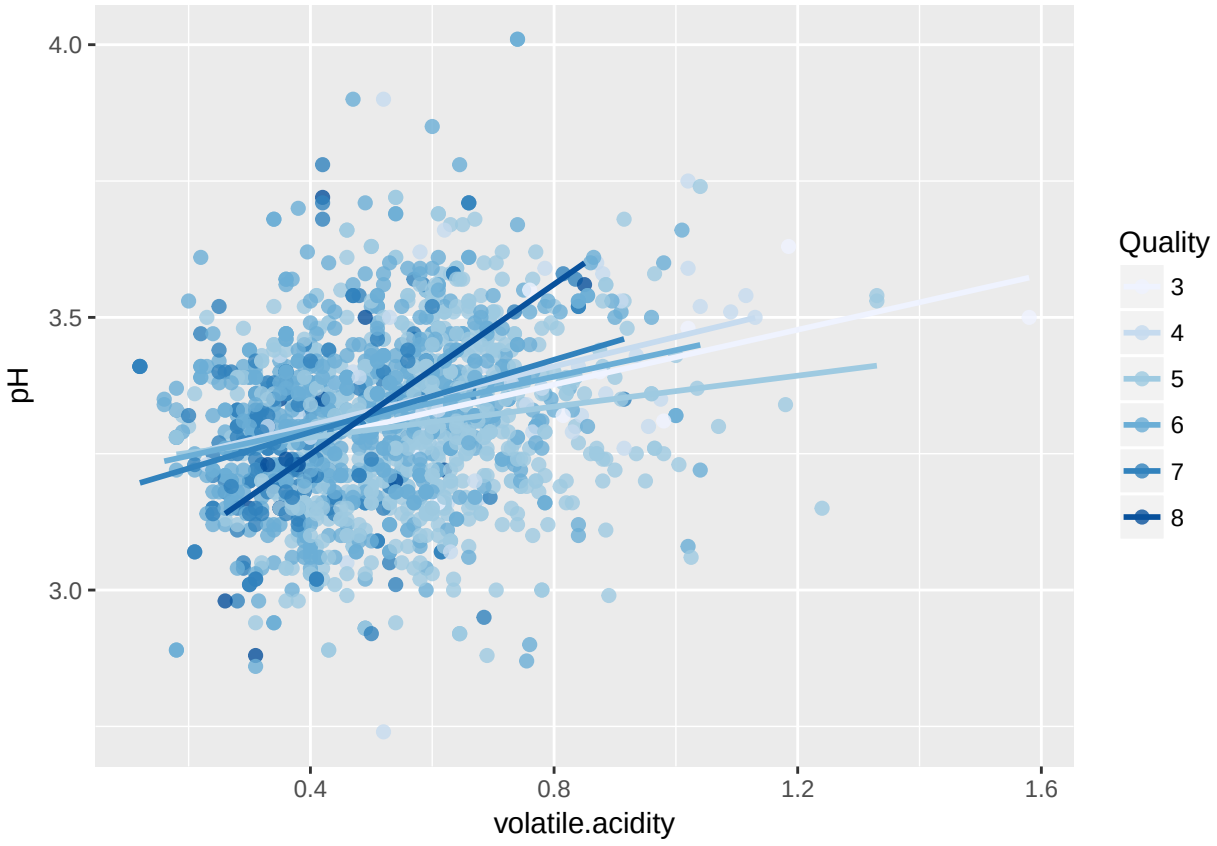
From the previous section, much higher levels of alcohol and citric acid were seen in wines with excellent quality than the others. In order to see how both factors affect wine quality, we make a scatter plot of alcohol and citric acid levels with points colored by wine qualities.



The plot reveals some important patterns in that most of the excellent wines (colored blue) are seen at high levels of both alcohol and citric acid. Furthermore, most of the good quality wines are spread out evenly along the medium levels of alcohol and citric acid. In addition, wines of excellent quality show a negative relationship between their alcohol and citric acid content. Another way to visualise this plot is using a faceted plot shown below



Likewise, the plot below shows that excellent quality wines are seen at lower levels of pH and volatile acidity with quite some few excellent wines seen at other regions of the plot. We also observe a strong positive relation between pH and volatile acidity of excellent quality wines.



Linear Model

From the plots in previous section, the variables alcohol, citric acid, pH, and volatile acidity all seems to be factors affecting the quality of a wine. It is of interest to know whether these variables will be a good predictor of wine quality. Thus we carry out a multiple linear regression model with quality as the dependent variable coupled with alcohol, citric acid, pH, and volatile acidity as predictors

```
##
## Call:
## lm(formula = quality ~ alcohol + citric.acid + pH + volatile.acidity)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.51757 -0.41070 -0.07765  0.46079  2.20383
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        4.67244    0.45693  10.226  < 2e-16 ***
## alcohol            0.33396    0.01675  19.937  < 2e-16 ***
## citric.acid       -0.18040    0.12055  -1.496    0.135
## pH                -0.52858    0.13531  -3.906 9.76e-05 ***
## volatile.acidity -1.36074    0.11321 -12.019  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.665 on 1594 degrees of freedom
## Multiple R-squared:  0.3237, Adjusted R-squared:  0.322
## F-statistic: 190.7 on 4 and 1594 DF,  p-value: < 2.2e-16
```

From the output above, the coefficient of citric acid is not significantly different from zero indicating that its not contributing to the linear model. Consequently we remove it. Furthermore, sulphates and density also seem to affect the quality of wines from plots in previous section so we include them in our model

```
##
## Call:
## lm(formula = quality ~ alcohol + pH + volatile.acidity + sulphates +
##     density)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.67752 -0.37390 -0.06917  0.47129  2.07763
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.67993   10.77588   0.249   0.8036
## alcohol           0.32186    0.01856  17.337  < 2e-16 ***
## pH               -0.30362    0.11879  -2.556   0.0107 *
## volatile.acidity -1.15601    0.09998 -11.562  < 2e-16 ***
## sulphates         0.63391    0.10359   6.120 1.18e-09 ***
## density           0.80222   10.63096   0.075   0.9399
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6577 on 1593 degrees of freedom
## Multiple R-squared:  0.3388, Adjusted R-squared:  0.3367
## F-statistic: 163.3 on 5 and 1593 DF,  p-value: < 2.2e-16
```

Because the coefficient of density is not significant, we remove sulphates from the model

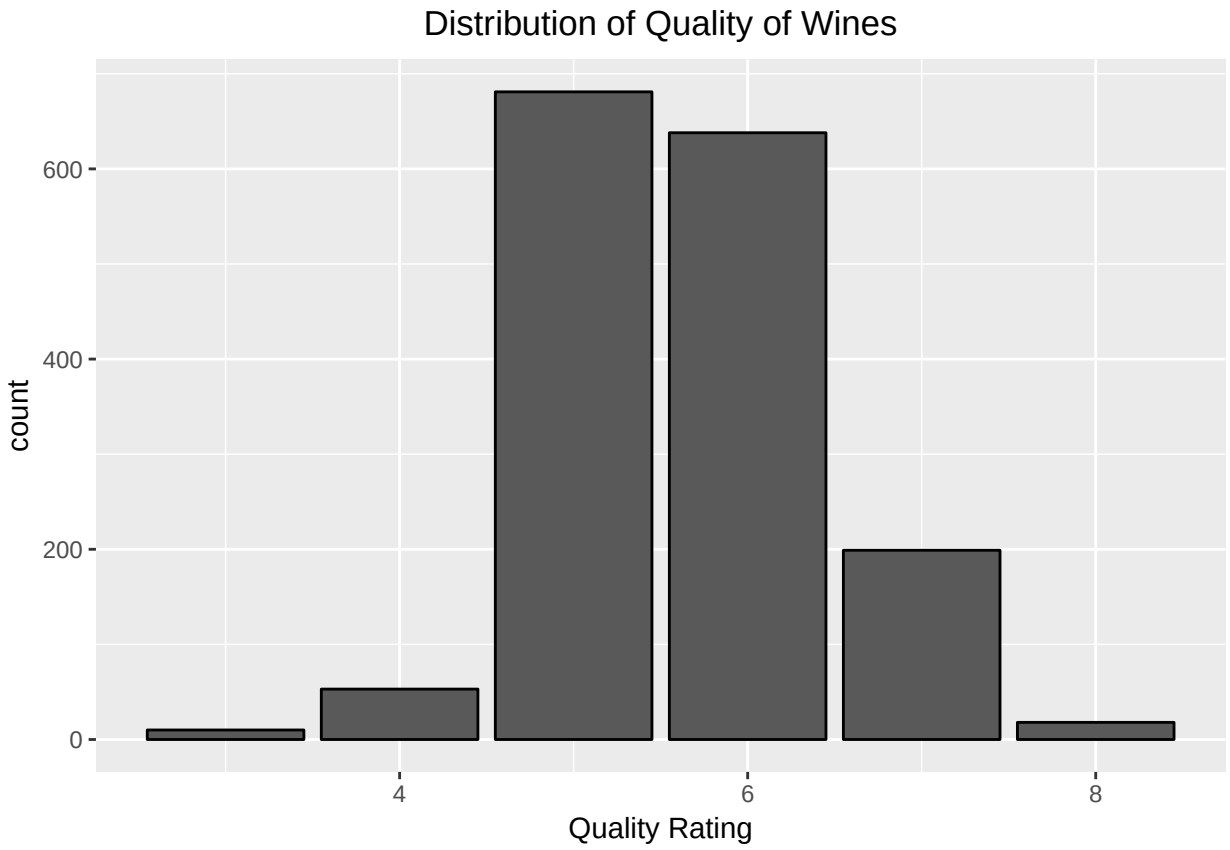
```
##
## Call:
## lm(formula = quality ~ alcohol + pH + volatile.acidity + sulphates)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.67671 -0.37422 -0.06988  0.47085  2.07767
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.49257    0.38537   9.063  < 2e-16 ***
## alcohol           0.32121    0.01641  19.576  < 2e-16 ***
## pH               -0.30580    0.11521  -2.654  0.00803 **
## volatile.acidity -1.15583    0.09993 -11.567  < 2e-16 ***
## sulphates         0.63528    0.10195   6.231 5.91e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6575 on 1594 degrees of freedom
## Multiple R-squared:  0.3388, Adjusted R-squared:  0.3372
## F-statistic: 204.2 on 4 and 1594 DF,  p-value: < 2.2e-16
```

From the model above we see that all coefficients are significantly different from zero indicating that all the predictors significantly contribute to the quality of wines. Furthermore, an adjusted R-squared value of 0.3372 indicate that the predictors alcohol, pH, volatile acidity and sulphates account for about 34 percent of the variation in wine quality.

This model is limited in that not all the predictors follow normal distribution. Furthermore, the dependent variable is ordinal discreet variable and more advance modelling techniques may be beneficial.

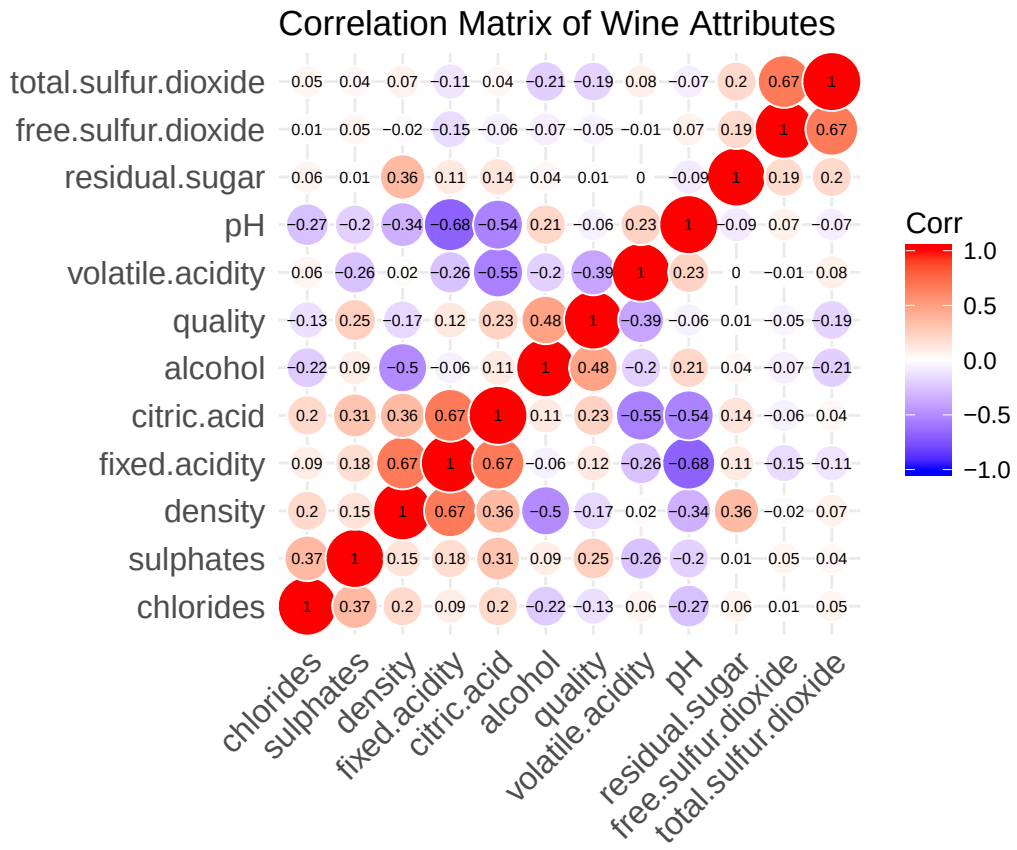
FInal Plots and Summary

Plot One



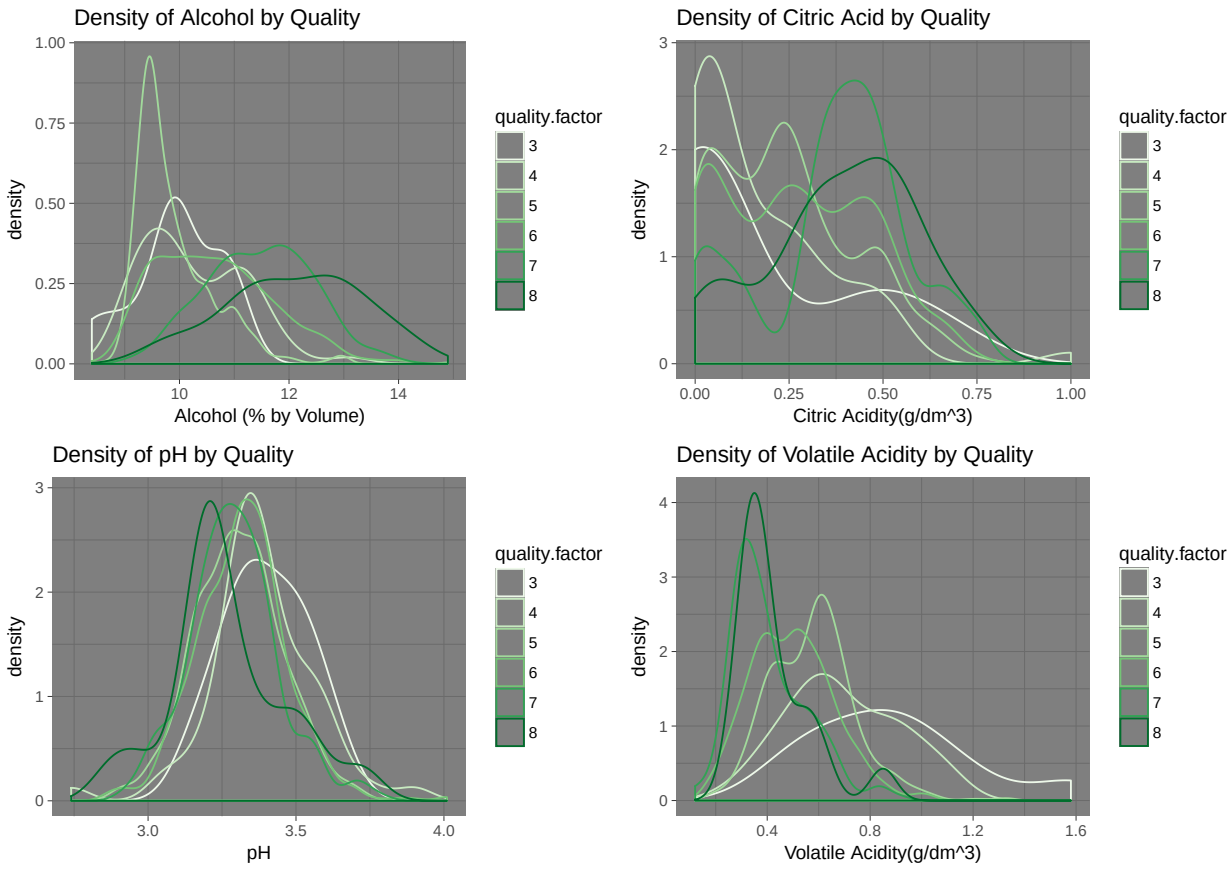
The plot above shows the distribution of quality of wines contained in the dataset. As seen from the plot, most wines have a quality rating of 5 and 6 with few observations seen at the tails of the distribution. This plot is of particular interest first because the main variable of interest in this data is wine quality and also because the distribution of the wine quality is normal.

Plot Two



The plot above shows a visualisation of the correlation matrix of the wine attributes. This plot is of interest because it revealed the nature of the relationships between the wine attributes and most importantly, the relationship between wine quality and the other attributes of the wines contained in the dataset. The plot shows that alcohol content has the highest positive correlation with wine quality and volatile acidity has the "highest" negative correlation with wine quality.

Plot Three



The plot above shows the densities of Alcohol, Citric Acid, pH and Volatile Acidity. This plot is of particular interest as it shows how the variables differ among the different levels of wine quality.

Reflection

The wine quality dataset contains information on 1599 wine samples across 12 variables. This report started with an exploratory analysis which involved individual variables followed by pairs of variables. The analysis involve lots of plots made to answer interesting questions. This was then followed by a multivariate analysis which involved fitting a linear model.

Analysis showed that there is a relationship between wine quality and alcohol content. Further investigation showed that excellent wines have the highest percent by volume of alcohol. Furthermore, there is a negative relationship between wine quality and volatile acidity with the excellent quality wines showing the lowest levels of volatile acidity. Likewise, pH and citric acid have a negative and positive relationship respectively with wine quality. We further investigated these variables by making them predictors in a linear model with wine quality as the dependent variable.

Model summary showed that citric acid did not significantly contribute to wine quality. Further iterations on the model revealed that alcohol, pH, volatile acidity and sulphates all significantly affect to wine quality. The low adjusted R squared value of about 34% was a bit of a let down as it indicated that these 4 predictors could only account for 34% of the variation in wine quality.

However, this is an opportunity for improvement in future analysis and investigation and more advanced modelling or classification techniques can be applied on the dataset for better and more accurate prediction

results. Even though recognising which of the variables are relevant in determining the quality of wines was quite difficult at the onset, the correlation matrix visialsation provided a lot of insight into variables worth exploring that could contribute towards the quality of wines.