Red Wine Quality Exploration by Sarah Alkhateeb

This report explores a dataset contains 1,599 red wines with 11 variables on the chemical properties (acidity, pH, density,. . .) of the wine.

Univariate Plots Section

```
## [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 ...
```

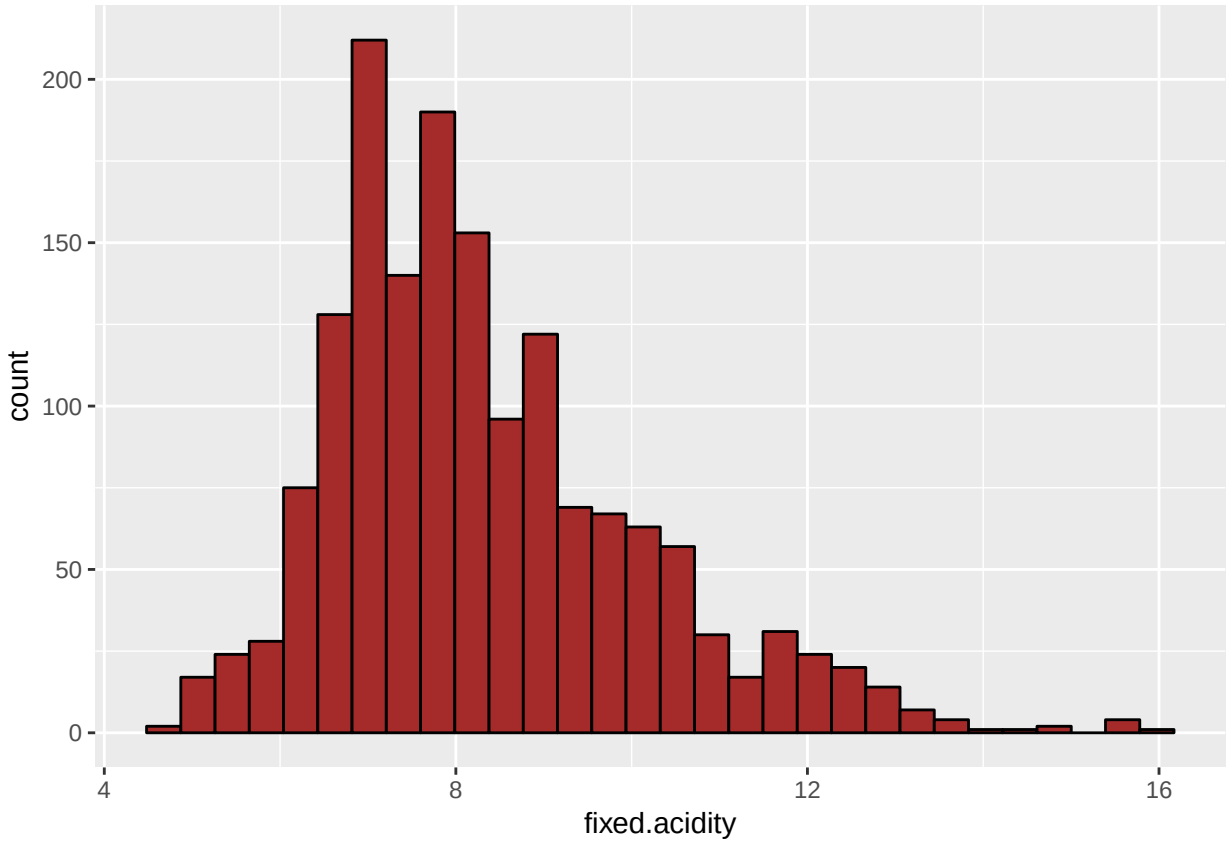
```
##        X          fixed.acidity   volatile.acidity  citric.acid
##  Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000
##  1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090
##  Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260
##  Mean   : 800.0   Mean   : 8.32   Mean   :0.5278   Mean   :0.271
##  3rd Qu.:1199.5   3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420
##  Max.   :1599.0   Max.   :15.90   Max.   :1.5800   Max.   :1.000
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.900   Min.   :0.01200   Min.   : 1.00
##  1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00
##  Median : 2.200   Median :0.07900   Median :14.00
##  Mean   : 2.539   Mean   :0.08747   Mean   :15.87
##  3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00
##  Max.   :15.500   Max.   :0.61100   Max.   :72.00
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  6.00       Min.   :0.9901   Min.   :2.740   Min.   :0.3300
##  1st Qu.: 22.00       1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500
##  Median : 38.00       Median :0.9968   Median :3.310   Median :0.6200
##  Mean   : 46.47       Mean   :0.9967   Mean   :3.311   Mean   :0.6581
##  3rd Qu.: 62.00       3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300
##  Max.   :289.00       Max.   :1.0037   Max.   :4.010   Max.   :2.0000
##     alcohol         quality
##  Min.   : 8.40   Min.   :3.000
##  1st Qu.: 9.50   1st Qu.:5.000
##  Median :10.20   Median :6.000
##  Mean   :10.42   Mean   :5.636
##  3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :14.90   Max.   :8.000
```

Obervations from the summary

Our dataset consist of 13 variables, with 1599 observations. 'X' is the unique identifier. '11 variables' are the chemical properties of the wine (fixed.acidity - alcohol) The last variable is the quality providing a rating between 0 (very bad) and 10 (very excellent)

Now, we will explore the variables individually

Fixed Acidity

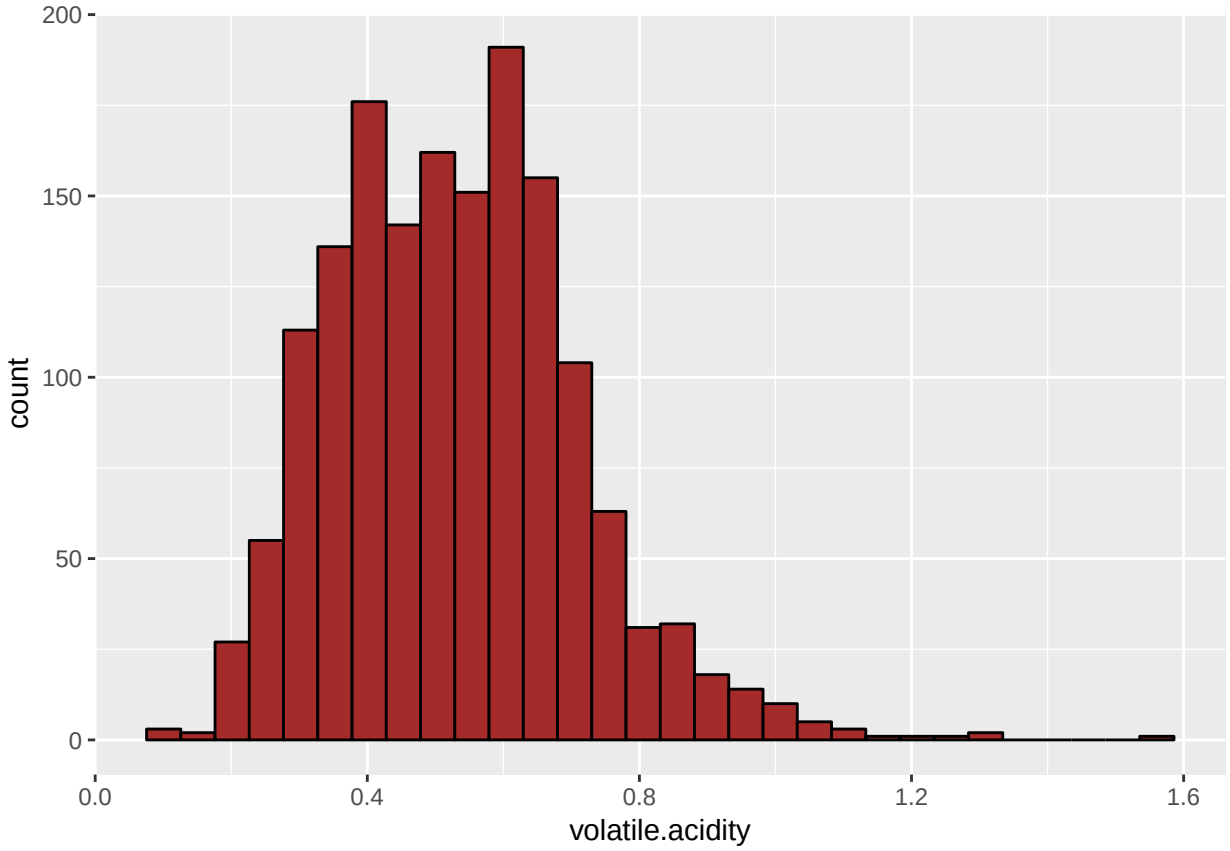



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
```

```
## [1] "7.2"
```

The distribution of fixed acidity is right skewed, and peaks at around 7

Volatile Acidity

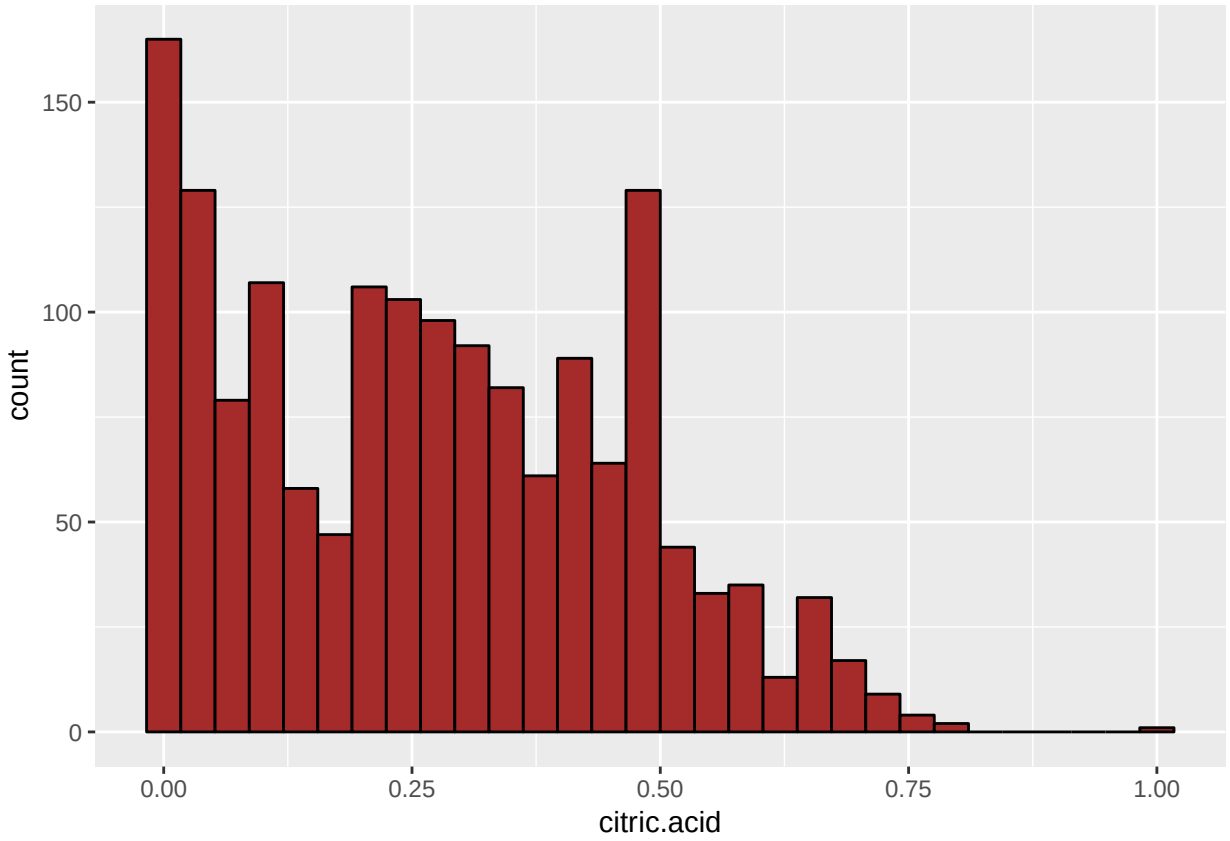



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```

```
## [1] "0.6"
```

The distribution of volatile acidity appears to be bimodal, and is concentrated around 0.4 and 0.6.

Citric Acid

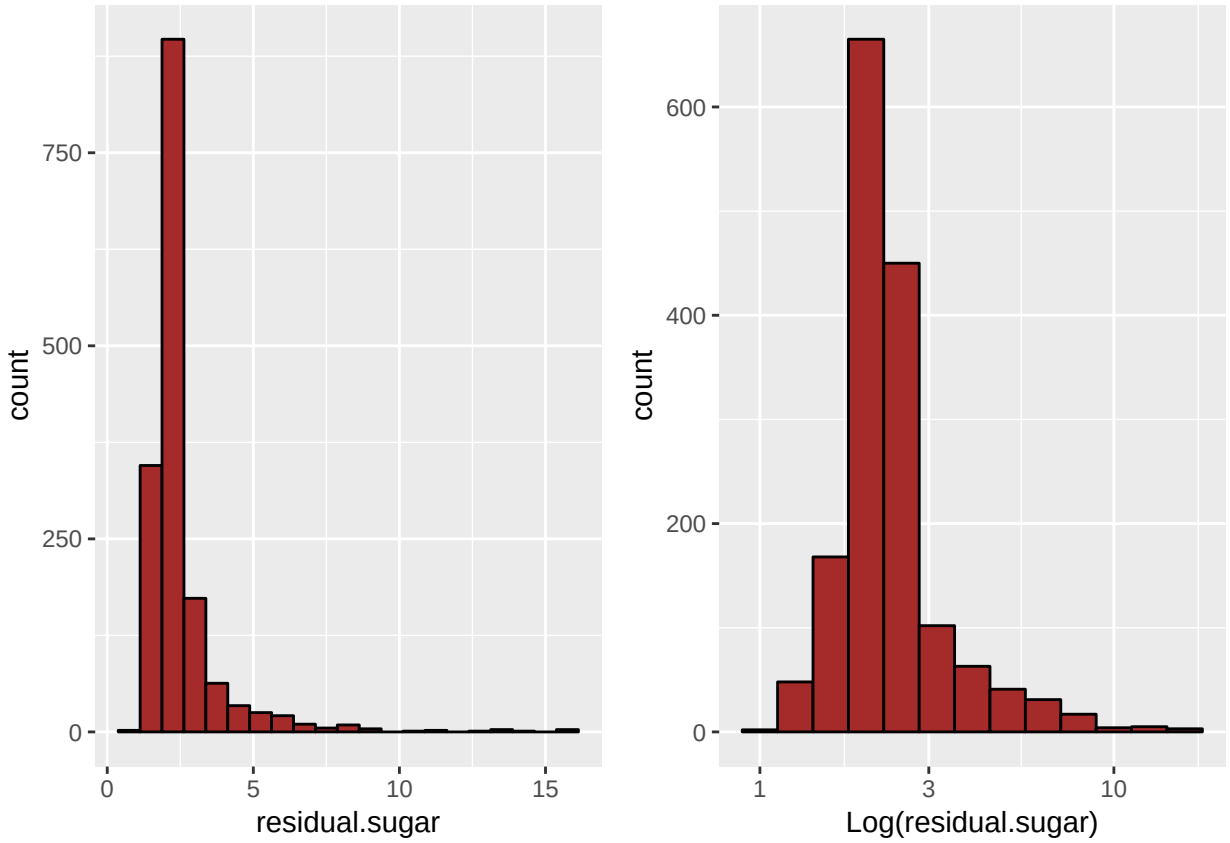



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```

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

It seems that most of the wines don't even have citric acid, the mode = 0. According to the description: citric acid found in small quantities, it can add 'freshness' and flavor to wines'

Residual Sugar

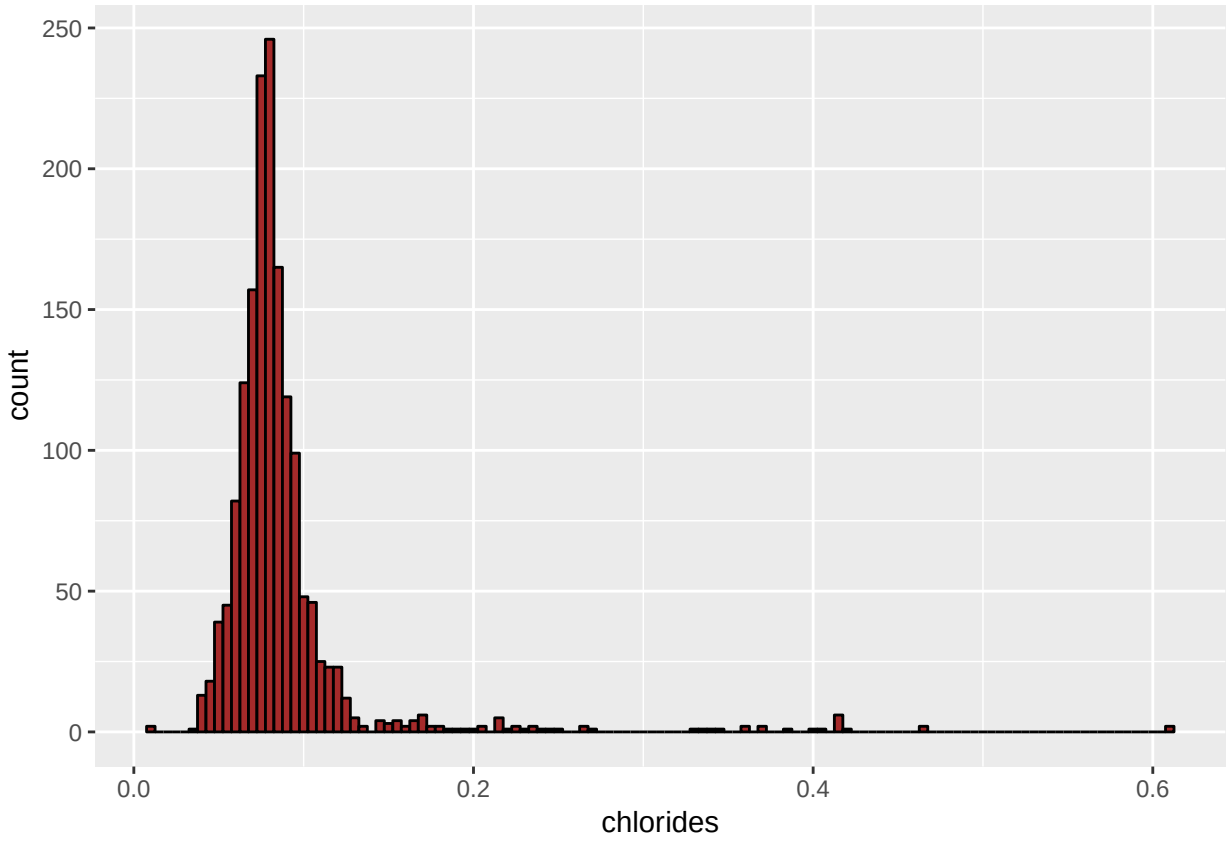



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```

```
## [1] "2"
```

The distribution is long tailed to the right (right skewed), The value with the highest bar/count (mode) = 2, Plotting histogram with a logarithmic scale (Log(residual.sugar) to better understand the distribution. The the log of the data restore the symmetry of the distribution.

Chlorides

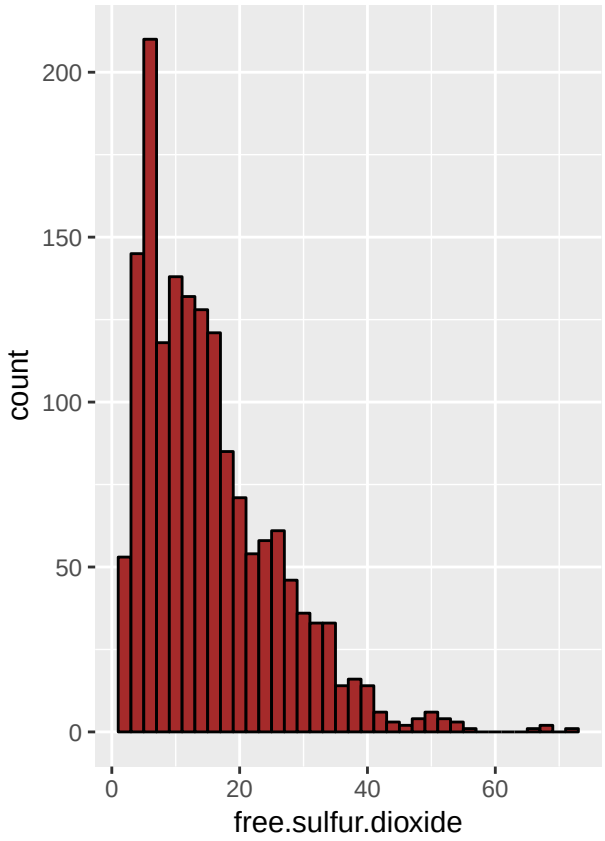


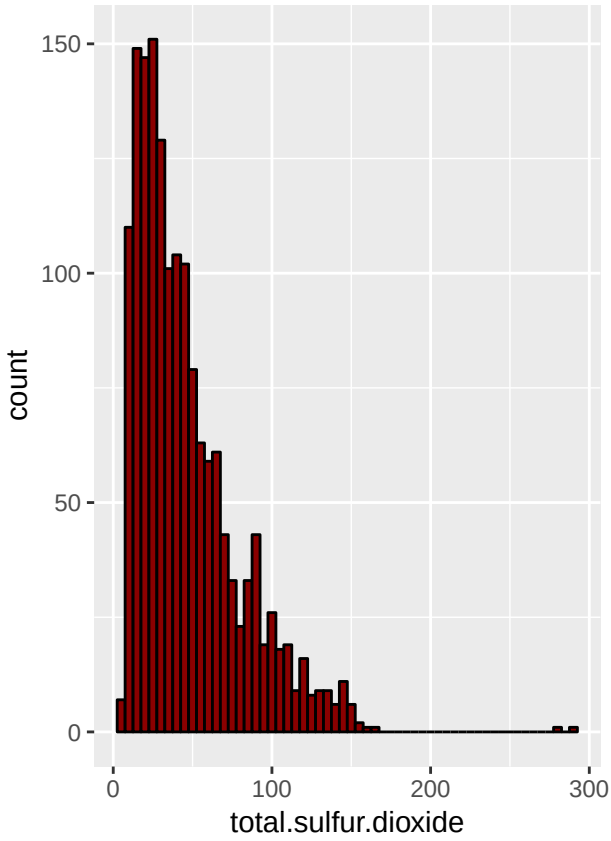

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```

```
## [1] "0.08"
```

The distribution is right skewed with an extreme outliers to the right. The chlorides count is higher around 0.08

Free Sulfur Dioxide & Total Sulfur Dioxide





Free Sulfur Dioxide

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

```
## [1] "6"
```

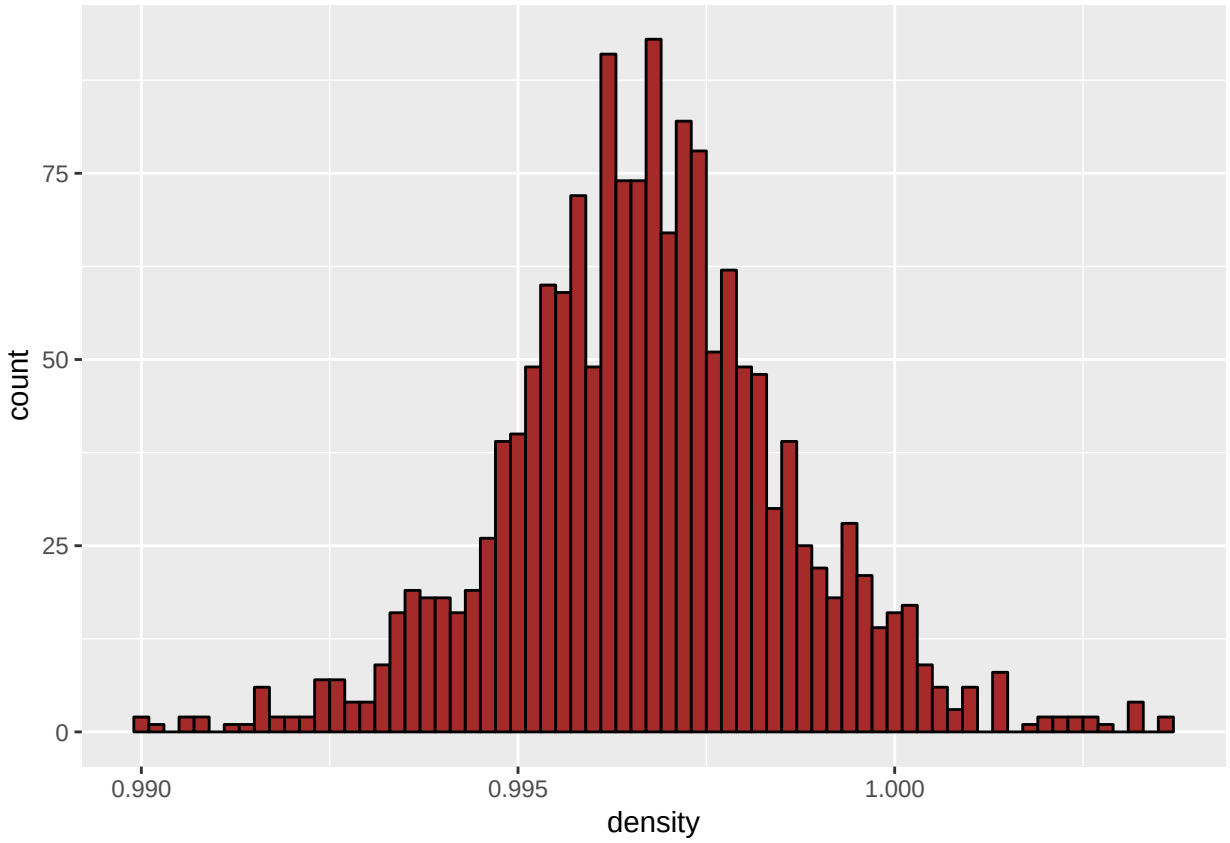
Total Sulfur Dioxide

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

```
## [1] "28"
```

Both distribution are right skewed. The Free Sulfur Dioxide distribution peaks at 6 while the Total Sulfur Dioxide distribution peaks at 28. Total Sulfur Dioxide has an extreme outliers (Total sulfurdioxide is the amount of free and bound forms of S02)

Density

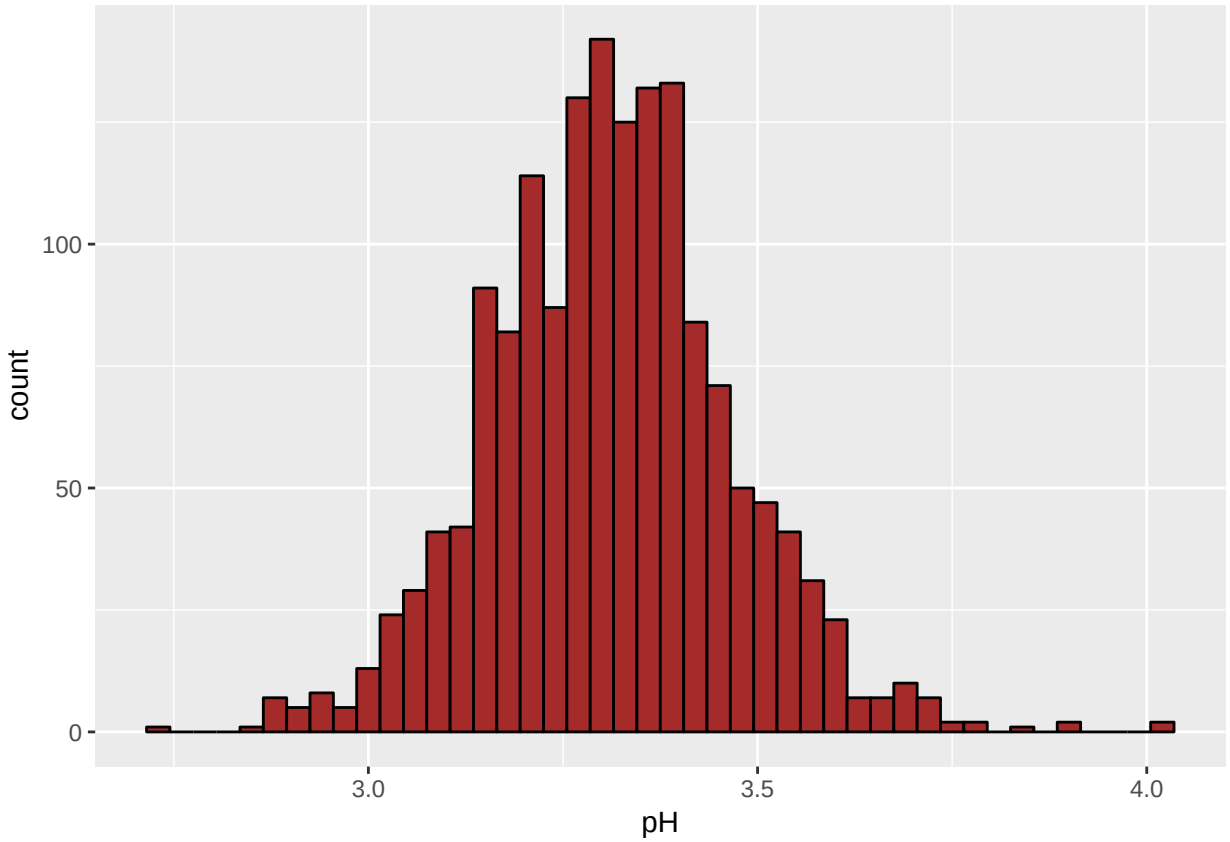



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0037
```

```
## [1] "0.9972"
```

the histogram is normally distributed with mean close to 1 (0.9967)

pH

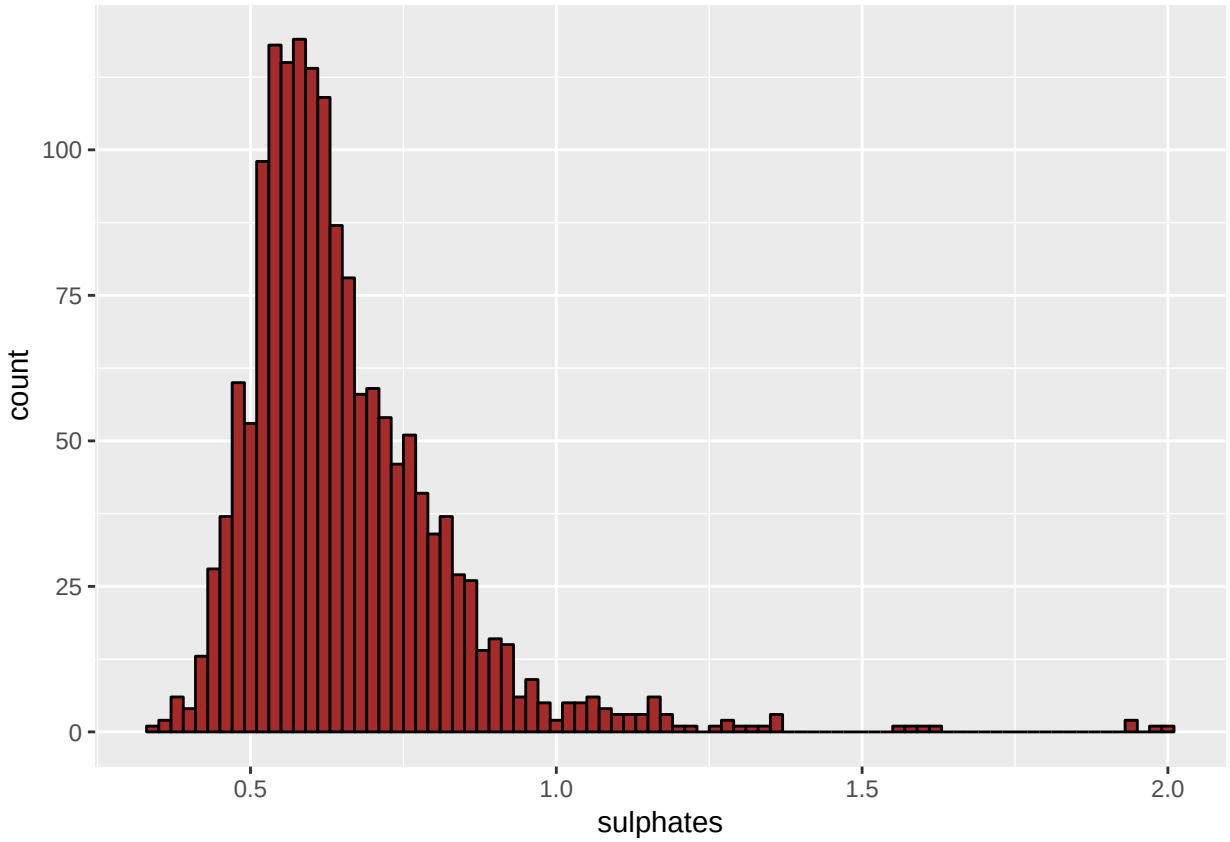



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010
```

```
## [1] "3.3"
```

We can see that the pH has a normaldistribution with a mean = 3.311 and mode = 3.3

Sulphates

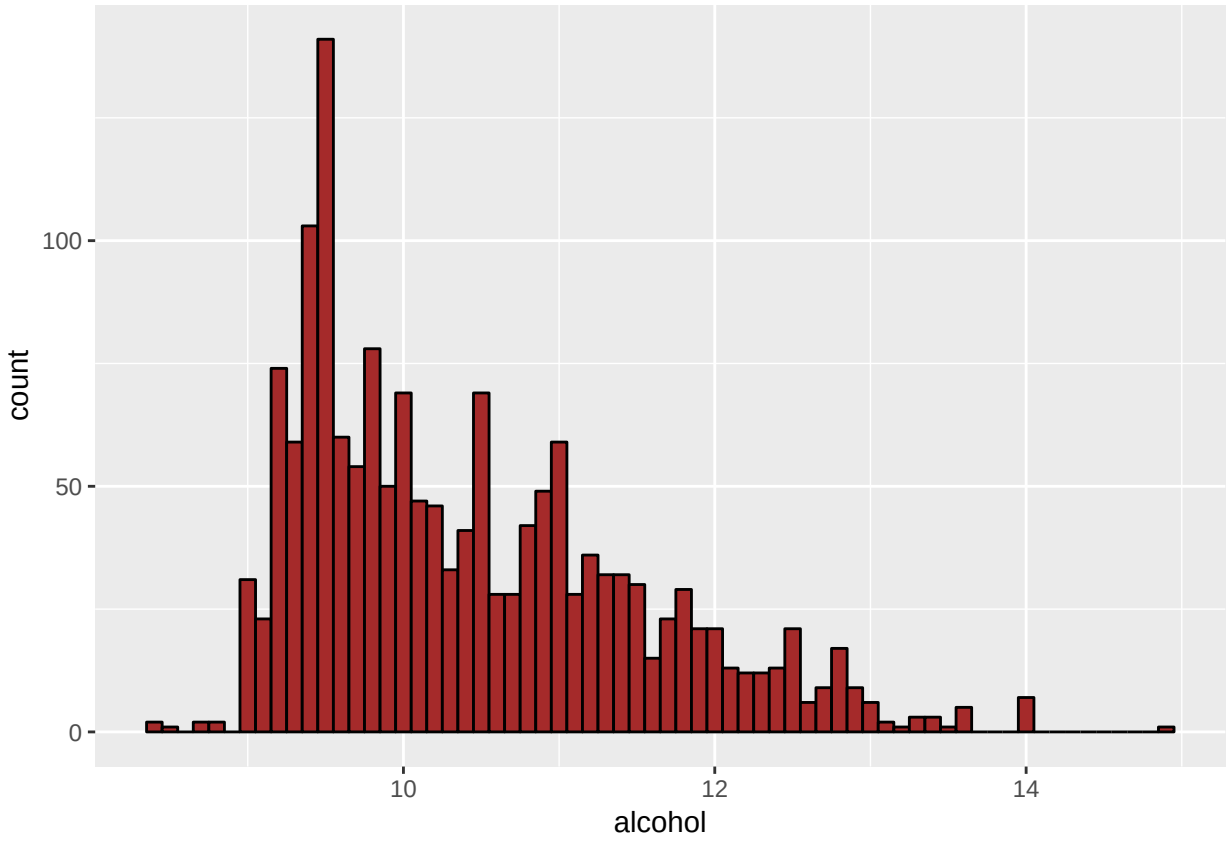



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```

```
## [1] "0.6"
```

Sulphates is a wine additive which can contribute to sulfur dioxide gas (S02) levels. The distribution is right skewed and the majority of sulphate values lies between 0.5 and 0.7, and peaks at 0.6

Alcohol

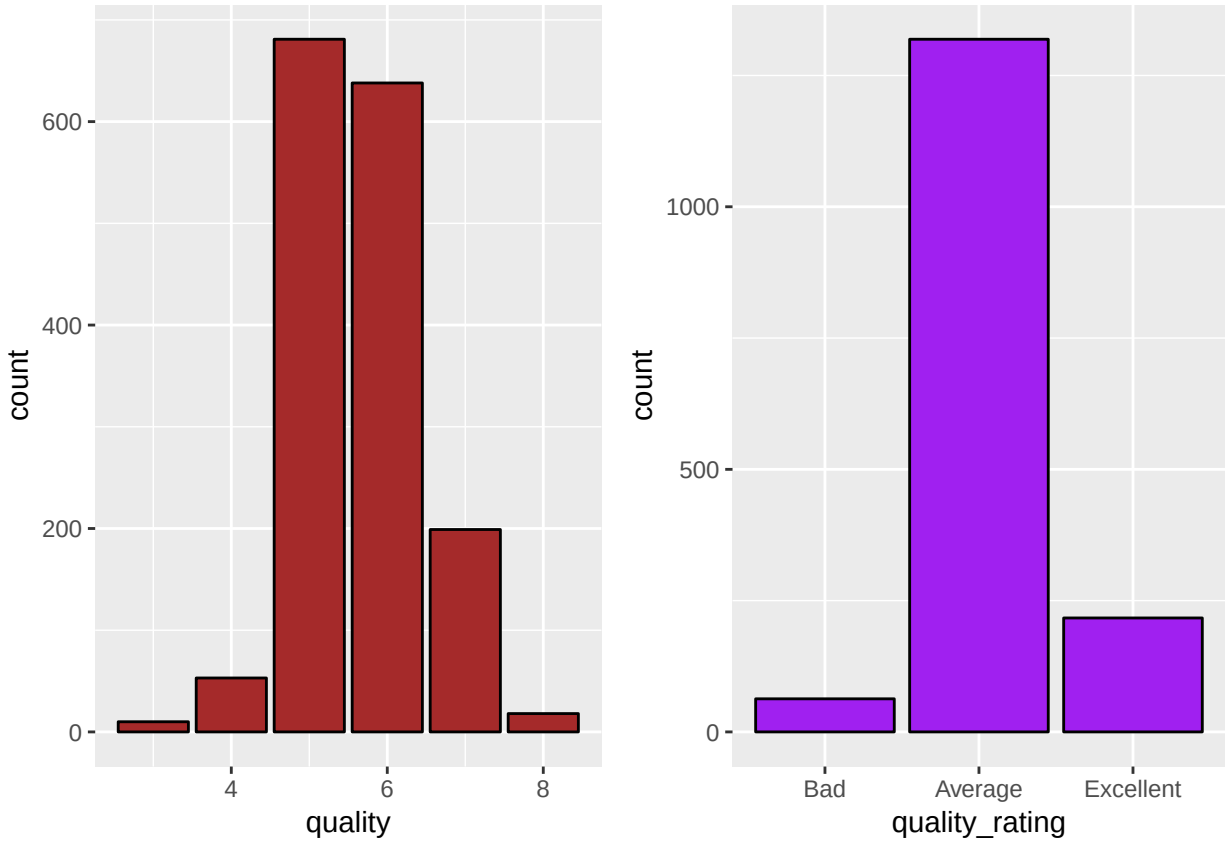



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```

```
## [1] "9.5"
```

The distribution is right skewed with mean = 10.2 and a maximum value = 14.9

Quality & Quality rating



We can see that there are much more Average wines than Excellent or Bad ones. Out of 10, very few wines scored high as 8 or low as 3 and 4. There is no score of 0, 1, 2, 9 or 10.

Univariate Analysis

What is the structure of your dataset?

The red_wine dataset consists of 1,599 observations and 13 variables (11 of which are chemical properties of the red wine (fixed.acidity, volatile.acidity, citric.acidity, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates, and alcohol). I have created an extra variable called quality_rating to classify the red wine quality to 3 categories: Bad, Average and Excellent.

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

The main feature that interests me is the 'quality'. I would like to determine the relationship between quality and other variables and discover which variables influence the quality of the wine and how can we determine if a certain wine is bad or good.

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

From the initial investigation, I think that Acidity, Alcohol, sugar, and density may have an effect on the quality of a wine.

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

Yes, I created the quality_rating variable which contains 3 levels based on the quality variable: 0 - 4: Bad, 5 -6: Average, 7 - 10: Excellent This classification will be easier for the comparison

**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?**

Most of the distributions were right skewed and contained extreme outliers. I performed a transformation on the residual. sugar using log, the log help in restoring the normal distribution of the data.

Bivariate Plots Section

First, We'll see the correlation between the quality and the other variables

```
## Call:corr.test(x = red_wine[2:12], y = red_wine[13], method = "pearson")
## Correlation matrix
##                       quality
## fixed.acidity            0.12
## volatile.acidity        -0.39
## citric.acid              0.23
## residual.sugar           0.01
## chlorides               -0.13
## free.sulfur.dioxide     -0.05
## total.sulfur.dioxide    -0.19
## density                 -0.17
## pH                      -0.06
## sulphates                0.25
## alcohol                  0.48
## Sample Size
## [1] 1599
## Probability values  adjusted for multiple tests.
##                       quality
## fixed.acidity            0.00
## volatile.acidity         0.00
## citric.acid              0.00
## residual.sugar           0.58
## chlorides                0.00
## free.sulfur.dioxide      0.09
## total.sulfur.dioxide     0.00
## density                  0.00
## pH                       0.06
## sulphates                0.00
## alcohol                  0.00
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

The variables which have the greater correlation with the quality are: alcohol, volatile.acidity, citric.acid and sulphates.

Now, we'll see the correlation between the supporting variables

```
## Call:corr.test(x = red_wine[2:12], y = red_wine[2:12], method = "pearson")
## Correlation matrix
##                       fixed.acidity volatile.acidity citric.acid
## fixed.acidity                  1.00            -0.26        0.67
## volatile.acidity              -0.26             1.00       -0.55
## citric.acid                    0.67            -0.55        1.00
## residual.sugar                 0.11             0.00        0.14
## chlorides                      0.09             0.06        0.20
## free.sulfur.dioxide           -0.15            -0.01       -0.06
## total.sulfur.dioxide          -0.11             0.08        0.04
## density                        0.67             0.02        0.36
## pH                            -0.68             0.23       -0.54
## sulphates                      0.18            -0.26        0.31
## alcohol                       -0.06            -0.20        0.11
##                       residual.sugar chlorides free.sulfur.dioxide
## fixed.acidity                   0.11      0.09               -0.15
## volatile.acidity                0.00      0.06               -0.01
## citric.acid                     0.14      0.20               -0.06
## residual.sugar                  1.00      0.06                0.19
## chlorides                       0.06      1.00                0.01
## free.sulfur.dioxide             0.19      0.01                1.00
## total.sulfur.dioxide            0.20      0.05                0.67
## density                         0.36      0.20               -0.02
## pH                             -0.09     -0.27                0.07
## sulphates                       0.01      0.37                0.05
## alcohol                         0.04     -0.22               -0.07
##                       total.sulfur.dioxide density    pH sulphates alcohol
## fixed.acidity                      -0.11    0.67 -0.68      0.18   -0.06
## volatile.acidity                    0.08    0.02  0.23     -0.26   -0.20
## citric.acid                         0.04    0.36 -0.54      0.31    0.11
## residual.sugar                      0.20    0.36 -0.09      0.01    0.04
## chlorides                           0.05    0.20 -0.27      0.37   -0.22
## free.sulfur.dioxide                 0.67   -0.02  0.07      0.05   -0.07
## total.sulfur.dioxide                1.00    0.07 -0.07      0.04   -0.21
## density                             0.07    1.00 -0.34      0.15   -0.50
## pH                                 -0.07   -0.34  1.00     -0.20    0.21
## sulphates                           0.04    0.15 -0.20      1.00    0.09
## alcohol                            -0.21   -0.50  0.21      0.09    1.00
## Sample Size
## [1] 1599
## Probability values  adjusted for multiple tests.
##                       fixed.acidity volatile.acidity citric.acid
## fixed.acidity                  0.00             0.00        0.00
## volatile.acidity               0.00             0.00        0.00
## citric.acid                    0.00             0.00        0.00
## residual.sugar                 0.00             1.00        0.00
## chlorides                      0.01             0.41        0.00
## free.sulfur.dioxide            0.00             1.00        0.41
## total.sulfur.dioxide           0.00             0.09        1.00
## density                        0.00             1.00        0.00
## pH                             0.00             0.00        0.00
## sulphates                      0.00             0.00        0.00
## alcohol                        0.41             0.00        0.00
##                       residual.sugar chlorides free.sulfur.dioxide
```

```
## fixed.acidity                 0.00      0.01               0.00
## volatile.acidity              1.00      0.41               1.00
## citric.acid                   0.00      0.00               0.41
## residual.sugar                0.00      0.63               0.00
## chlorides                     0.63      0.00               1.00
## free.sulfur.dioxide           0.00      1.00               0.00
## total.sulfur.dioxide          0.00      1.00               0.00
## density                       0.00      0.00               1.00
## pH                            0.03      0.00               0.18
## sulphates                     1.00      0.00               0.86
## alcohol                       1.00      0.00               0.19
##                      total.sulfur.dioxide density   pH sulphates alcohol
## fixed.acidity                        0.00    0.00 0.00      0.00    0.41
## volatile.acidity                     0.09    1.00 0.00      0.00    0.00
## citric.acid                          1.00    0.00 0.00      0.00    0.00
## residual.sugar                       0.00    0.00 0.03      1.00    1.00
## chlorides                            1.00    0.00 0.00      0.00    0.00
## free.sulfur.dioxide                  0.00    1.00 0.18      0.86    0.19
## total.sulfur.dioxide                 0.00    0.17 0.25      1.00    0.00
## density                              0.17    0.00 0.00      0.00    0.00
## pH                                   0.25    0.00 0.00      0.00    0.00
## sulphates                            1.00    0.00 0.00      0.00    0.01
## alcohol                              0.00    0.00 0.00      0.01    0.00
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

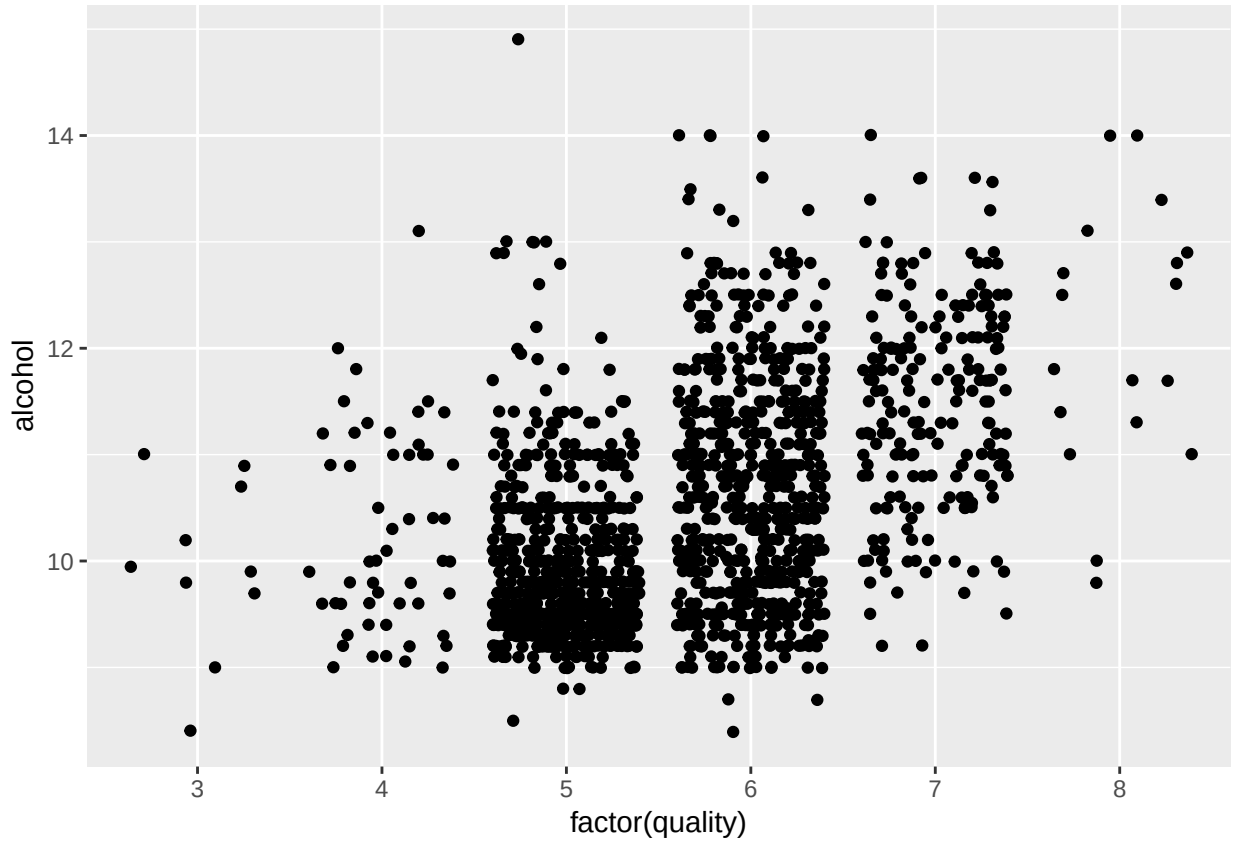
Observations from the above correlations:

*We can see some related variables such as: fixed.acidity with citric.acid & density/ citric.acid & volatile.acidity/ free.sulfur.dioxide & total.sulfur.dioxide/ density & alcohol/ pH with fixed.acidity & citric.acid

*Density is strong, positive correlated with Fixed acidity and strong, negative correlated with Alcohol.

Let's start plotting bivariate relations:

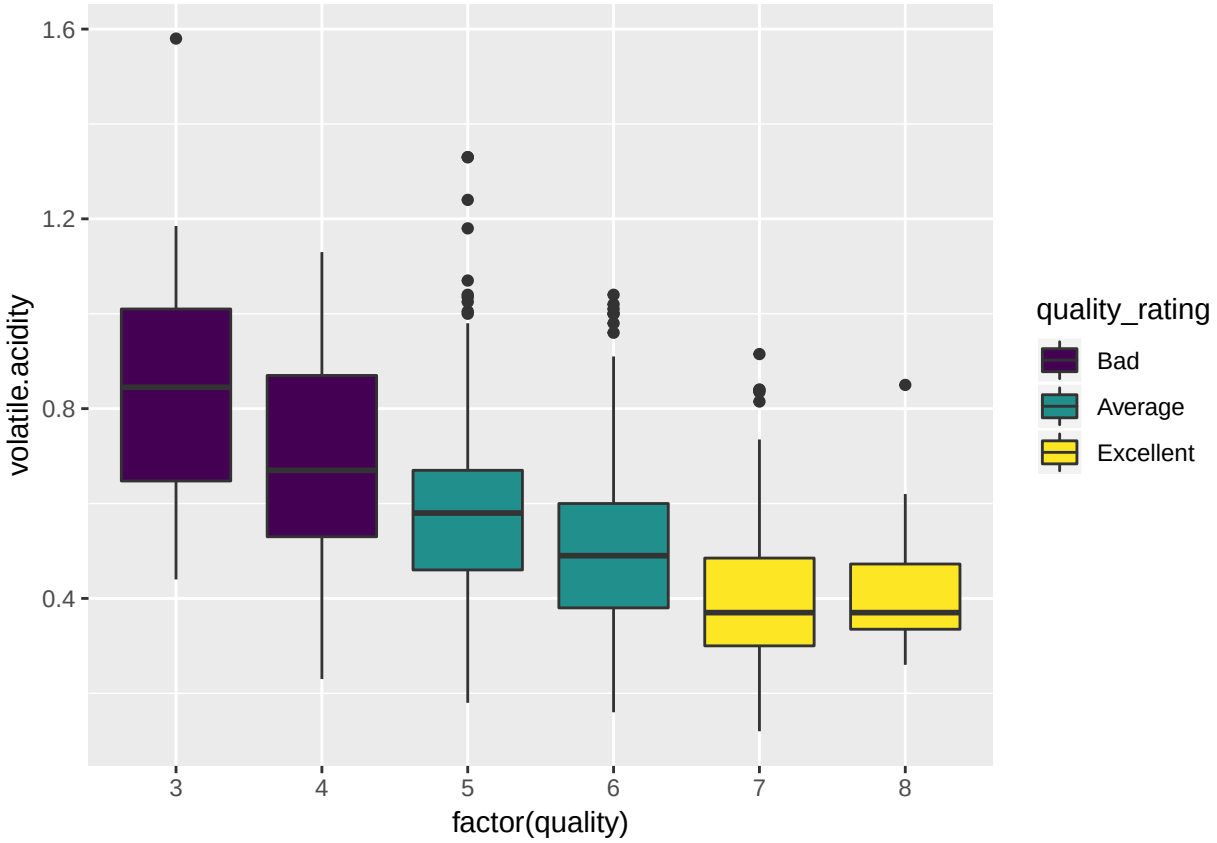
Quality & Alcohol




```
## red_wine$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.400   9.725   9.925   9.955  10.575  11.000
## ------------------------------------------------------------
## red_wine$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00    9.60   10.00   10.27   11.00   13.10
## ------------------------------------------------------------
## red_wine$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     8.5     9.4     9.7     9.9    10.2    14.9
## ------------------------------------------------------------
## red_wine$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.80   10.50   10.63   11.30   14.00
## ------------------------------------------------------------
## red_wine$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.20   10.80   11.50   11.47   12.10   14.00
## ------------------------------------------------------------
## red_wine$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.80   11.32   12.15   12.09   12.88   14.00
```

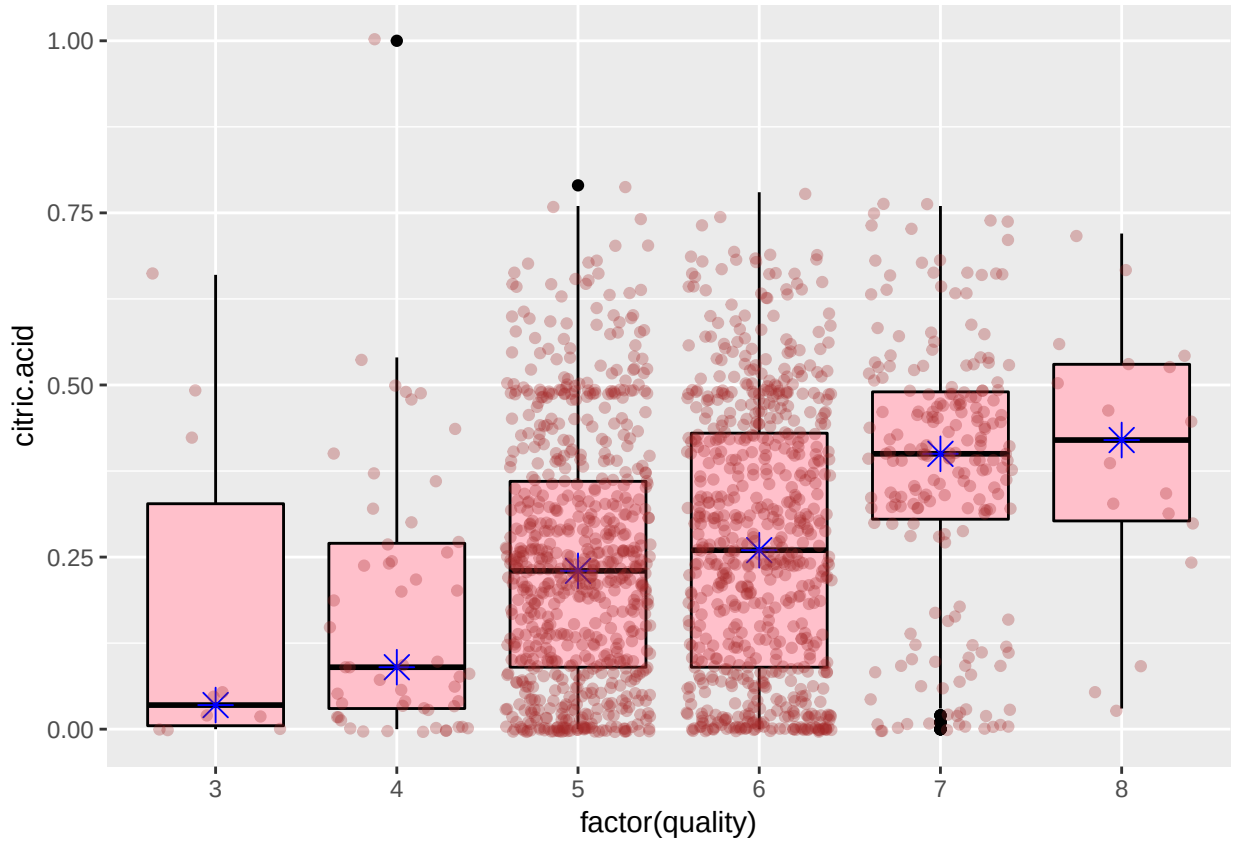
We can notice that better wines have higher alcohol, but we can see number of outliers here. The correlation is not too strong to say that alcohol alone influence the quality of the wine.

Quality & Volatile Acidity



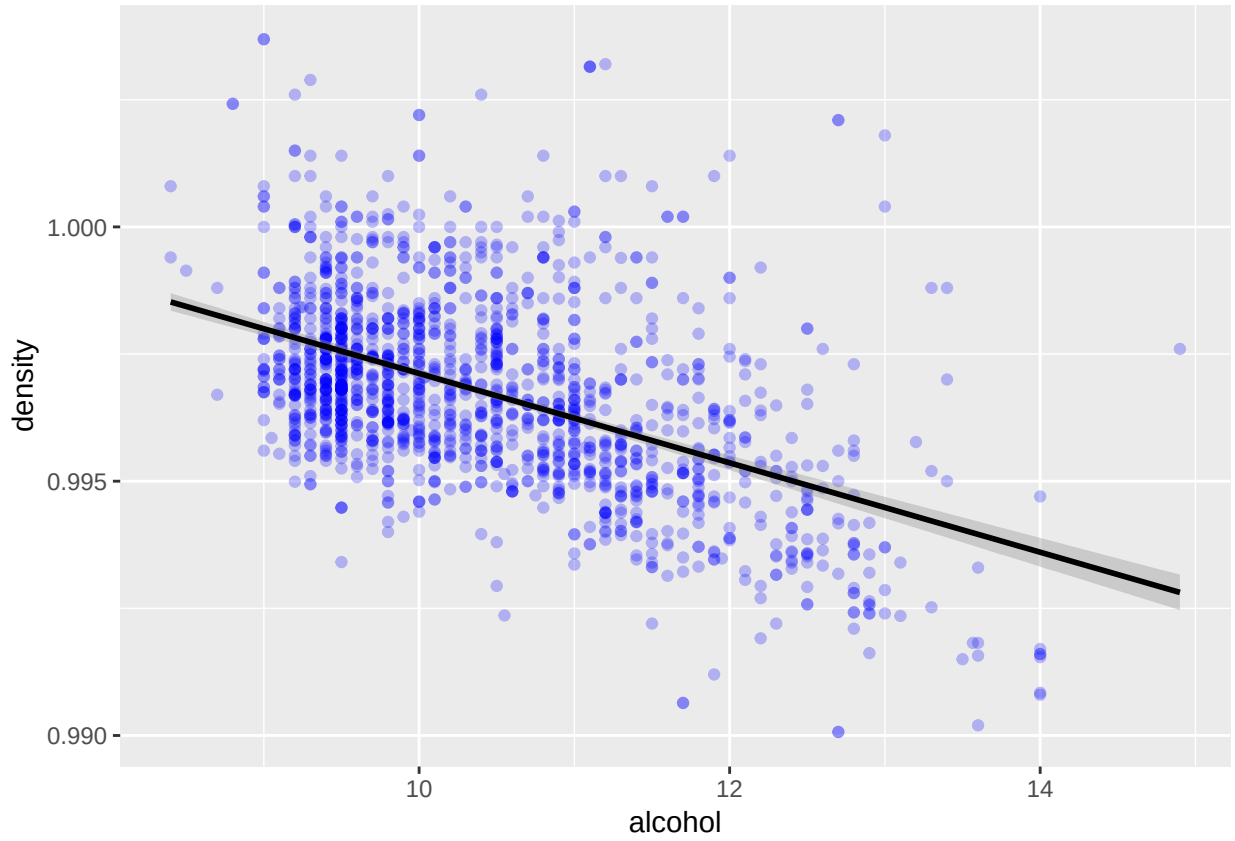
On the other hand, Volatile Acidity seems to have a negative effect (correlation) on quality. We can see that the better the wine is, the lower the volatile acidity.

Quality & Citric Acid

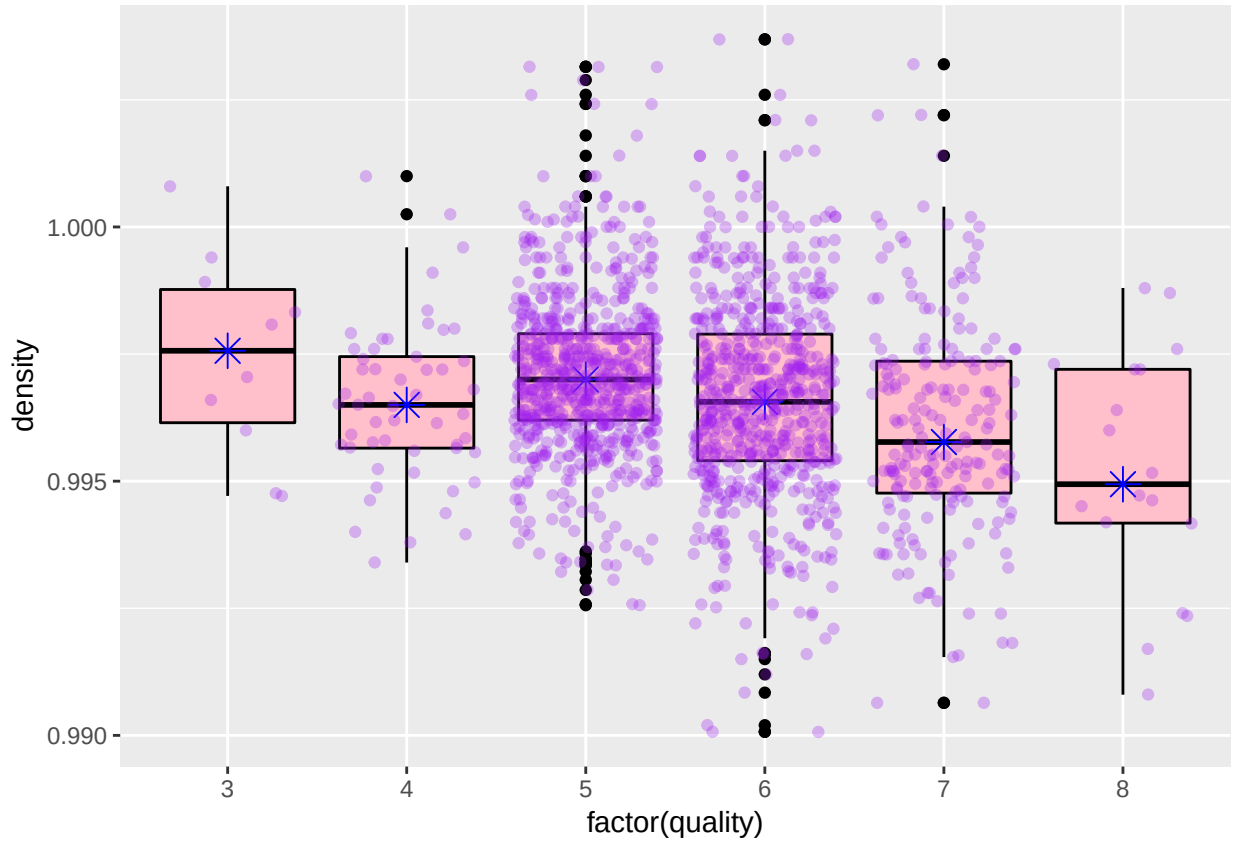


There is a positive correlation between Citric Acid and wine quality. We can see that better wines have higher median values of citric acid.

Alcohol & Density



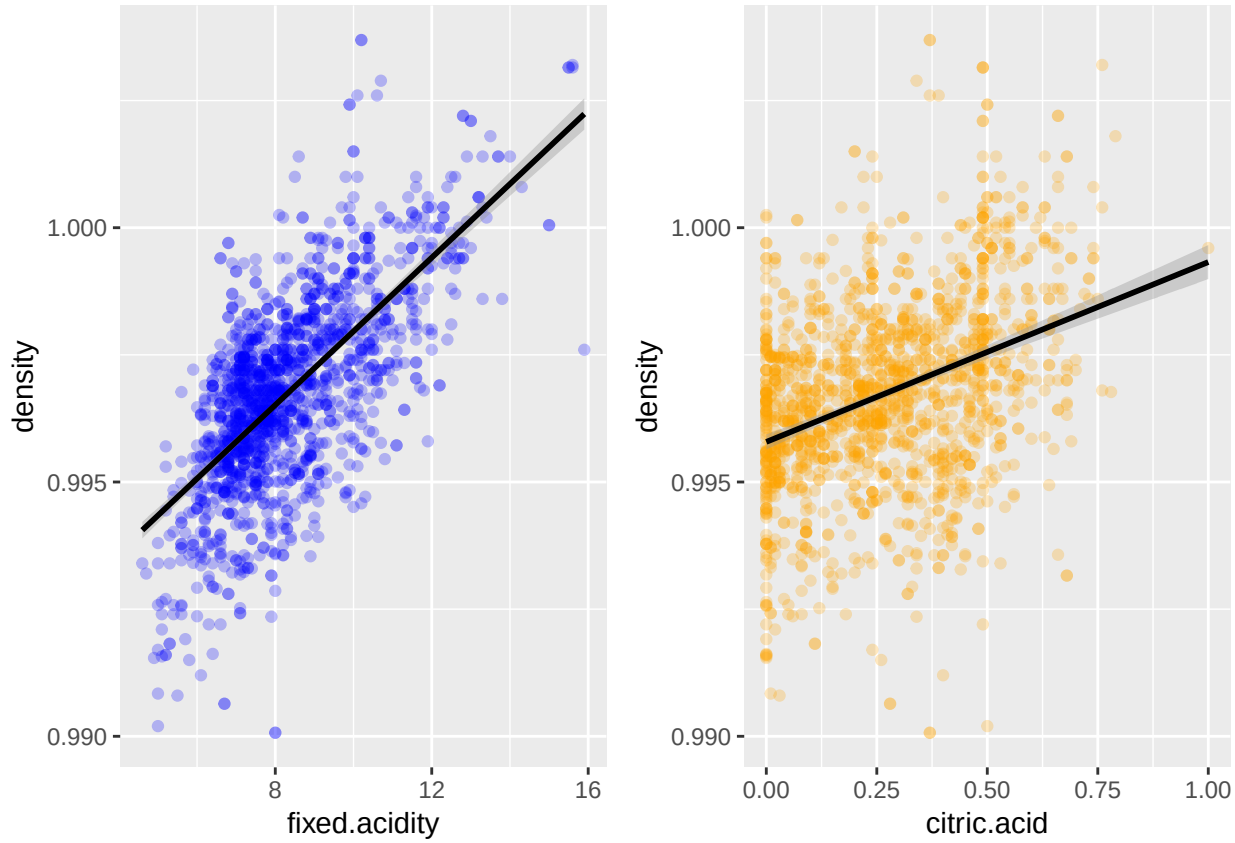
Quality & Density




```
## red_wine$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9947  0.9961  0.9976  0.9975  0.9988  1.0008
## ------------------------------------------------------------
## red_wine$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9934  0.9957  0.9965  0.9965  0.9974  1.0010
## ------------------------------------------------------------
## red_wine$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9926  0.9962  0.9970  0.9971  0.9979  1.0031
## ------------------------------------------------------------
## red_wine$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9954  0.9966  0.9966  0.9979  1.0037
## ------------------------------------------------------------
## red_wine$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9906  0.9948  0.9958  0.9961  0.9974  1.0032
## ------------------------------------------------------------
## red_wine$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9908  0.9942  0.9949  0.9952  0.9972  0.9988
```

At first sight, it looks like better quality wines have lower Density. However, we can see that density is correlated with alcohol, which seems like an interesting observation.

Fixed Acidity & Density/ Citric Acid & Density



We can see that density has a strong correlation with Fixed Acidity and Citric Acidity.

```
##
## Call:
## lm(formula = quality ~ alcohol, data = red_wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8442 -0.4112 -0.1690  0.5166  2.5888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.87497    0.17471   10.73   <2e-16 ***
## alcohol      0.36084    0.01668   21.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7104 on 1597 degrees of freedom
## Multiple R-squared:  0.2267, Adjusted R-squared:  0.2263
## F-statistic: 468.3 on 1 and 1597 DF,  p-value: < 2.2e-16
```

Based on the R2 score, alcohol alone explains about 22.7% of the variance in 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?

The Wine quality is correlated with Alcohol, Volatile Acidity, Sulphates and Citric Acid. None of these variables are strongly correlated with quality. The (Correlation Coefficient) r values are 0.48, -0.39, 0.25 and 0.23, respectively. From the R2 value, we saw that Alcohol alone explains about 22.7% of the variance in quality.

Fixed acidity, density, Residual sugar, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide and pH showed no or poor correlation with wine quality.

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

Fixed acidity is strongly correlated with density, fixed acidity increases the density. Alcohol is strongly correlated with density. While alcohol decreases the density, it also increases the quality.

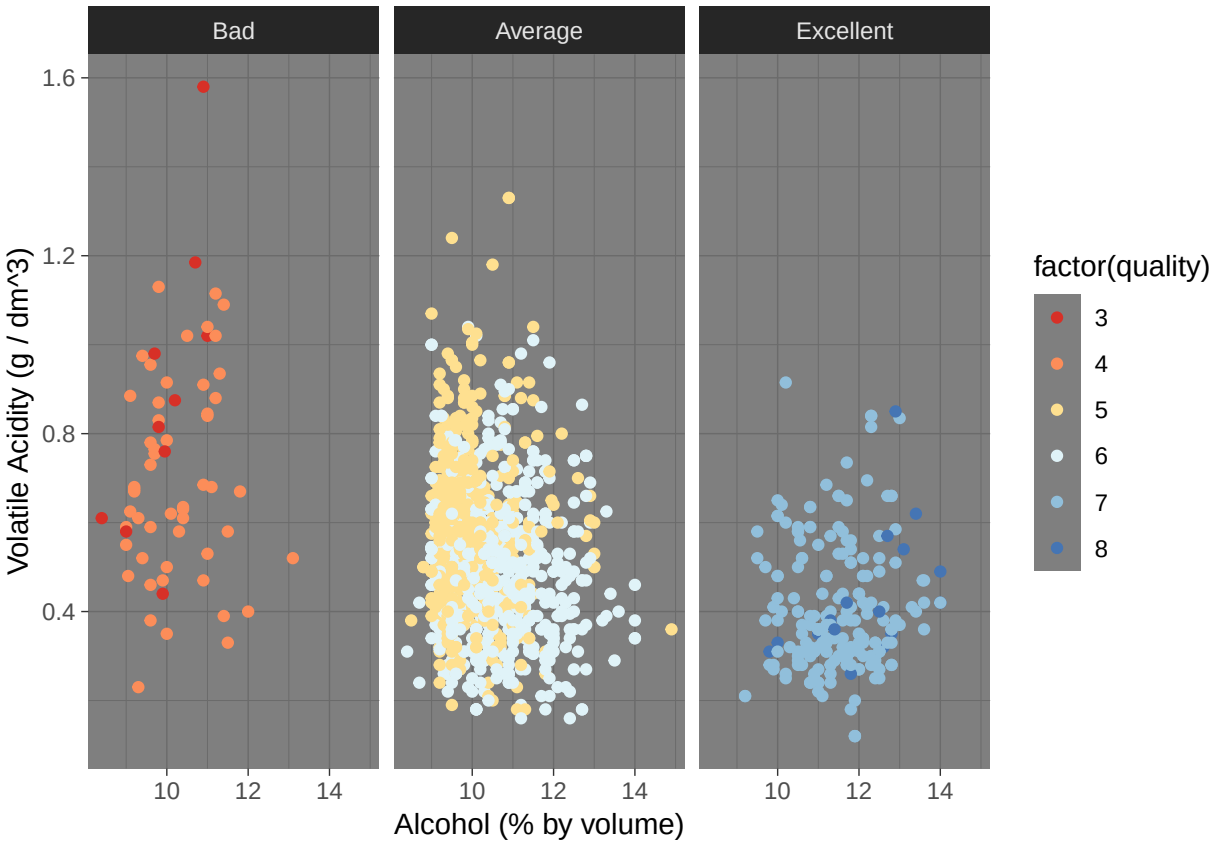
What was the strongest relationship you found?

The strongest relationship is between Alcohol & Quality. Concerning other variables, we'll find that the strongest relationships are between Fixed Acidity and Density (0.67),Fixed Acidity and Citric Acid (0.67) and Fixed Acidity & pH (-0.68).

Multivariate Plots Section

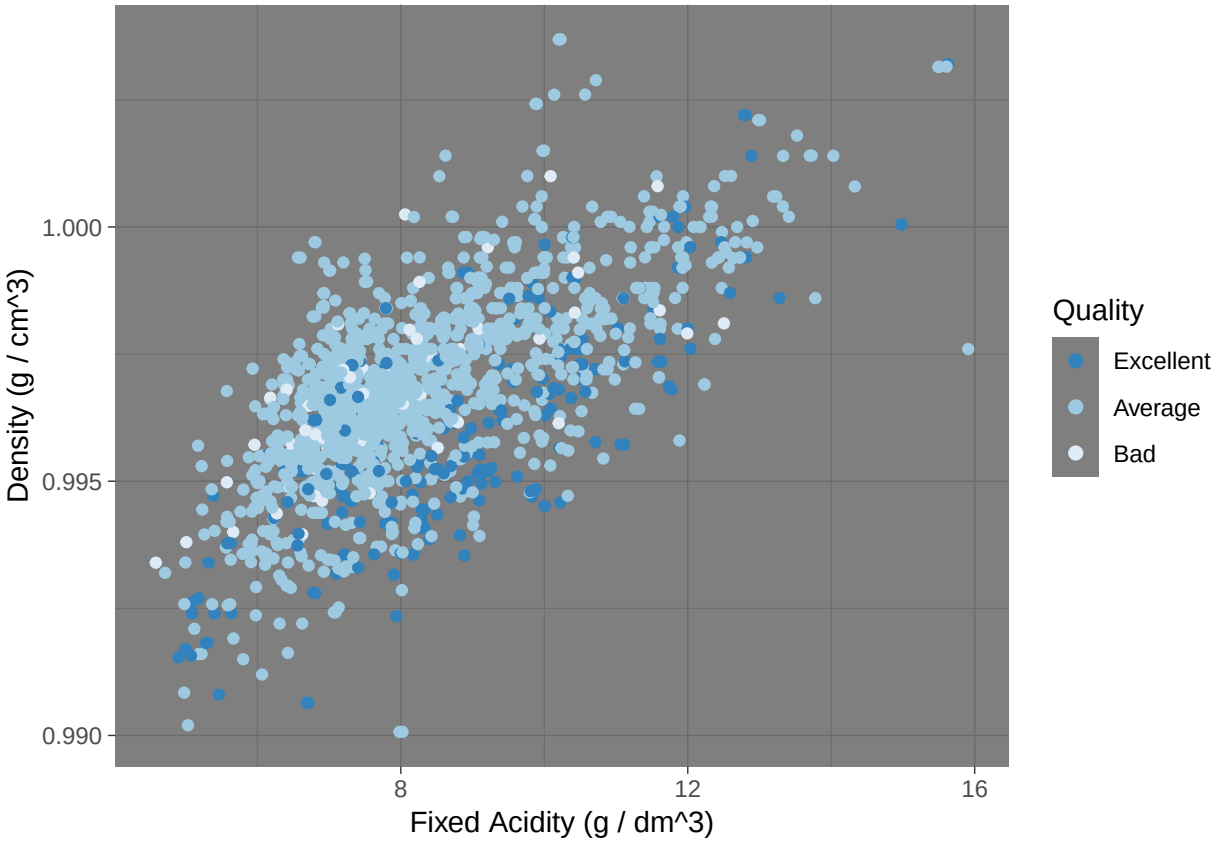
We're going to further explore relationships between Alcohol, Density, Acidity and the quality of the wine. The color of the dots will represent wine quality.

Quality & Alcohol & Volatile Acidity



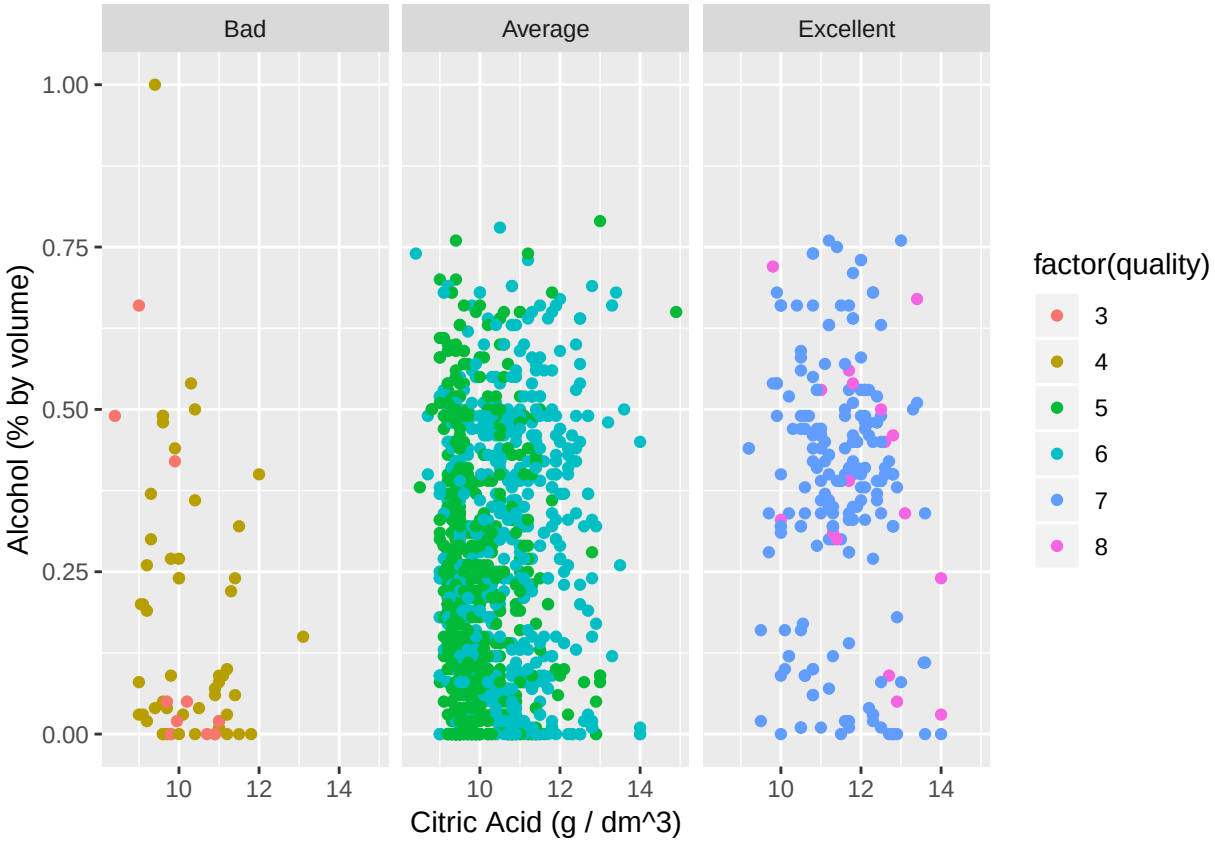
We can see that higher quality wines are concentrated in the bottom right of the plot, with higher values of Alcohol and Lower values of Volatile Acidity.

Fixed Acidity & Density & Quality



It appears that fixed acidity causes density levels to increase. It also seems that better quality has more density.

Citric Acid & Alcohol & Quality



No strong correlation here, either. But 'Average wines tend to be higher on the plot, meaning a greater concentration of citric acid.

Linear modelling

Let's build a linear model taking into account the variables that are more correlated with quality: Alcohol, Density, Volatile Acidity, Fixed Acidity and Citric Acid.

```
##
## Calls:
## m1: lm(formula = as.numeric(quality) ~ alcohol, data = red_wine)
## m2: lm(formula = as.numeric(quality) ~ alcohol + volatile.acidity,
##     data = red_wine)
## m3: lm(formula = as.numeric(quality) ~ alcohol + volatile.acidity +
##     fixed.acidity, data = red_wine)
## m4: lm(formula = as.numeric(quality) ~ alcohol + volatile.acidity +
##     fixed.acidity + citric.acid, data = red_wine)
## m5: lm(formula = as.numeric(quality) ~ alcohol + volatile.acidity +
##     fixed.acidity + density, data = red_wine)
##
## ================================================================================================
##                            m1            m2            m3            m4            m5
## ------------------------------------------------------------------------------------------------
##   (Intercept)           1.875***      3.095***      2.674***      2.622***      15.573
##                        (0.175)       (0.184)       (0.218)       (0.219)       (15.187)
##   alcohol               0.361***      0.314***      0.321***      0.325***      0.311***
```

```
##                          (0.017)       (0.016)       (0.016)       (0.016)       (0.020)
##   volatile.acidity                   -1.384***     -1.286***     -1.420***     -1.272***
##                                      (0.095)       (0.099)       (0.115)       (0.100)
##   fixed.acidity                                     0.036***      0.056***      0.045**
##                                                     (0.010)       (0.013)       (0.015)
##   citric.acid                                                     -0.314*
##                                                                   (0.137)
##   density                                                                     -12.922
##                                                                               (15.214)
## ---------------------------------------------------------------------------------------
##   R-squared              0.227         0.317         0.322         0.325         0.323
##   adj. R-squared         0.226         0.316         0.321         0.323         0.321
##   sigma                  0.710         0.668         0.665         0.664         0.665
##   F                    468.267       370.379       253.055       191.617       189.939
##   p                      0.000         0.000         0.000         0.000         0.000
##   Log-likelihood     -1721.057     -1621.814     -1615.379     -1612.739     -1615.017
##   Deviance             805.870       711.796       706.090       703.763       705.771
##   AIC                 3448.114      3251.628      3240.758      3237.479      3242.034
##   BIC                 3464.245      3273.136      3267.643      3269.742      3274.297
##   N                   1599          1599          1599          1599          1599
## =======================================================================================
```

Based on the value of R2 (0.323), we can't predict or explain the quality of a wine.

Multivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

There were non-strong features that were highly corelated to each other. Fixed acidity increases the density levels, and It seems that good quality has more density. *I expected Density plays a big role in determining the quality of the wine ,but it doesn't have a strong realtion with quality.

Were there any interesting or surprising interactions between features?

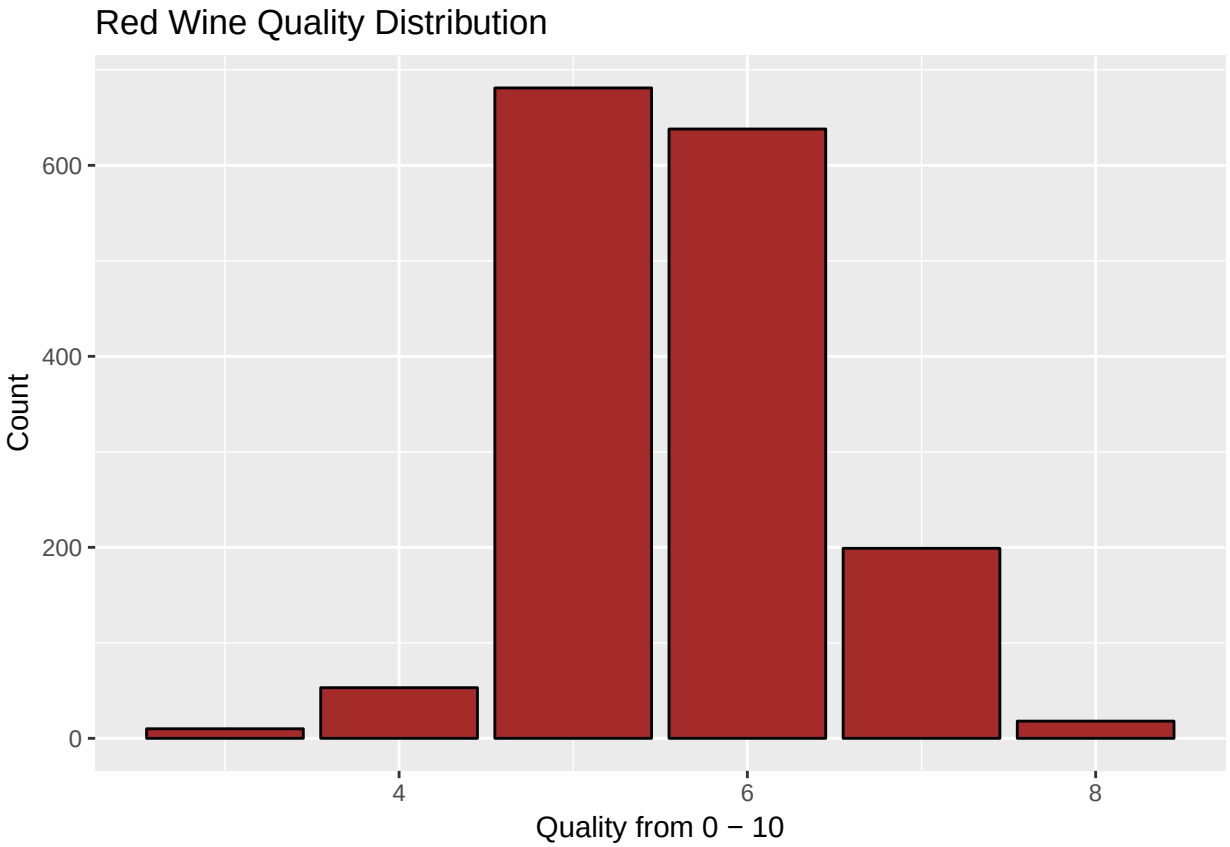
Yes, the correlation between Fixed Acidity and density appears to be strong, which something I didn't think of.

OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model?

I created a linear model but wasn't satisfied with the results. The R2 score was 0.323, which is insufficient to explain the variance in Wine Quality.

Final Plots and Summary

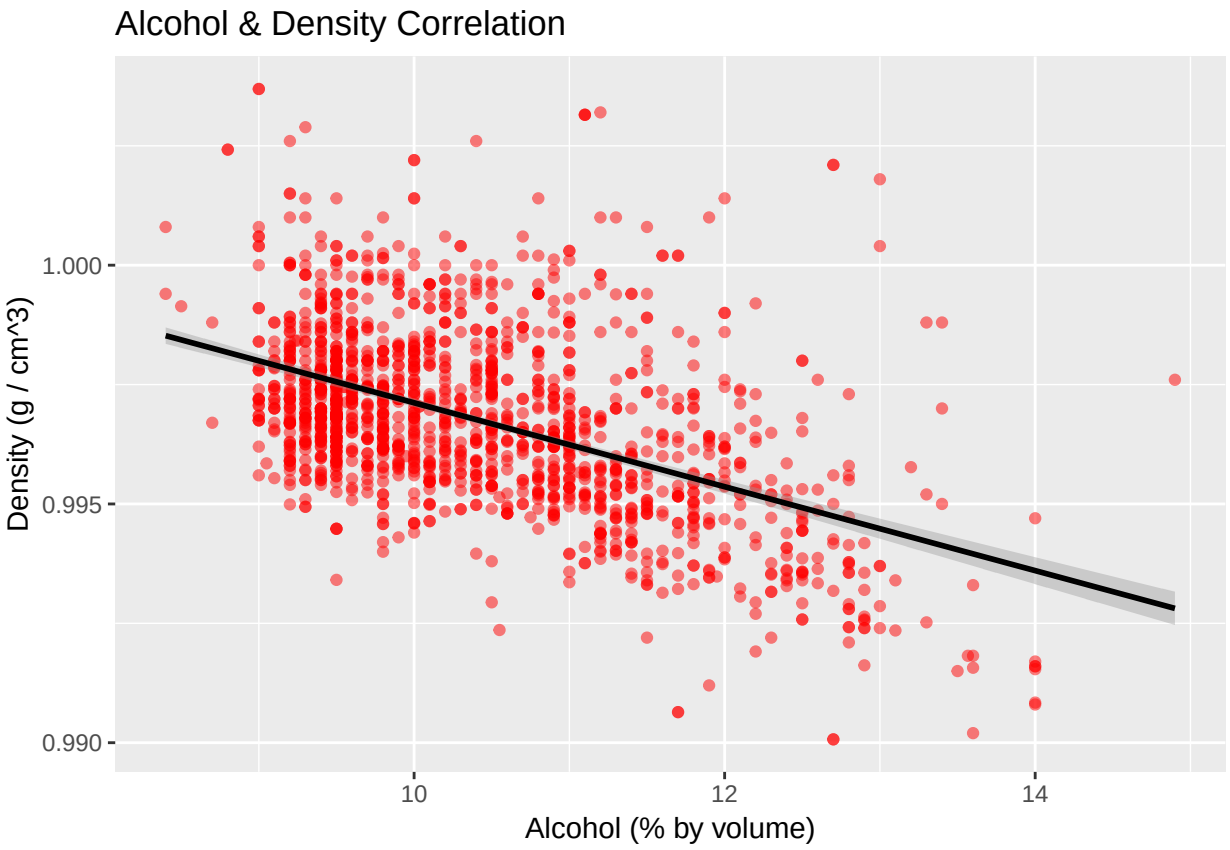
Plot One



Description One

Quality is the main feature I was interested in, it's an important property to pay special attention to it. We can see that the distribution is bell shaped (normal), and we have more Average wines than the Bad or Excellent ones.

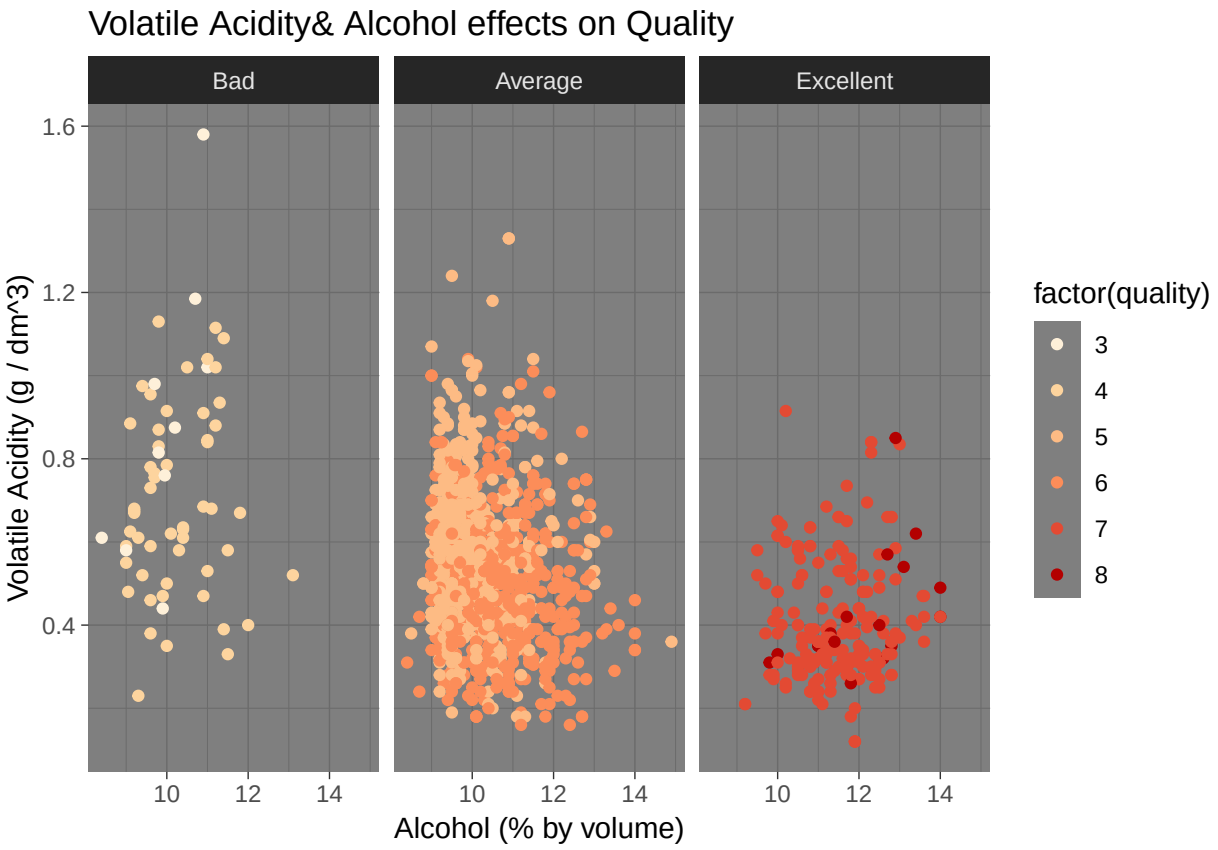
Plot Two



Description Two

*We can see that there is a strong negative correlation (r = -0.5) between Alcohol and Density. Alcohol increases the Quality (r = 0.48) while Density has a weak negative correlation with quality (-0.17), which seems interesting for further investigation.

Plot Three



Description Three

We can see the two variables that have a good correlation with the Wine Quality. The values in the "Average" facet look grouped between 9 & 12 in terms of Alcohol and higher rate of Volatile Acidity. Nevertheless, we can observe that the "Bad" quality values have no or very weak correlation with both Volatile Acidity or Alcohol.

Reflection

The Red Wine dataset contains 1599 observations with 13 variables represent various chemical properties of the wine. The first section (univariate plots) showed that most of the variables are right-skewed except the density and pH, which have normal distributions.

The first time I saw the data I thought that there would be strong correlations between the quality of the wine and other variables. However, after the plots, it turns out that very few variables have a strong correlation even though it is not very strong. Contrary to what I thought, Density has a weak negative correlation with the quality and have a Pearson correlation r = - 0.17. However, Alcohol has a negative correlation with density and I can say that alcohol decreases the overall density of the wine. Some limitations: The data set seems small and with more observations, we may discover stronger relations. Creating a linear model was

not very useful to predict the quality of a wine. We need more data to be able to perform a prediction and explain more relations among variables.

Are other variables such as grape types, wine brand, production year and wine selling price affect the quality of a wine and the rating?