Explore and Summarize Red Wine quality Dataset by Deepa Sobhana Devi

Project Overview

This report explores a dataset containing 1599 attributes of the Portuguese “Vinho Verde” red wine.The dataset contains several physicochemical attributes and sensory classification made by wine experts.The variables are fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates, alcohol and quality.

Here we Use R and apply exploratory data analysis techniques to explore relationships in one variable to multiple variables and to explore a selected data set for distributions, outliers, and anomalies.

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

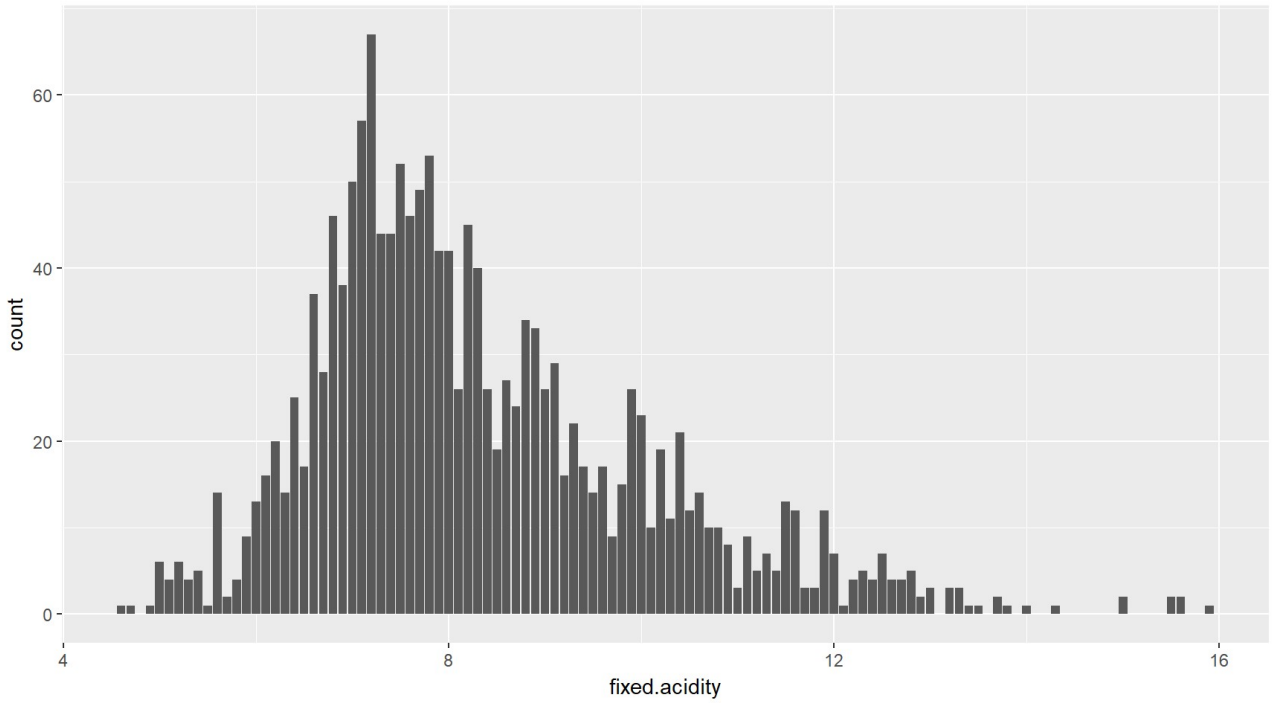
Our dataset consists of 13 variables, with 1599 observations.

Univariate Plots Section

In this section, we are performing some preliminary exploration of the dataset.

Fixed Acidity

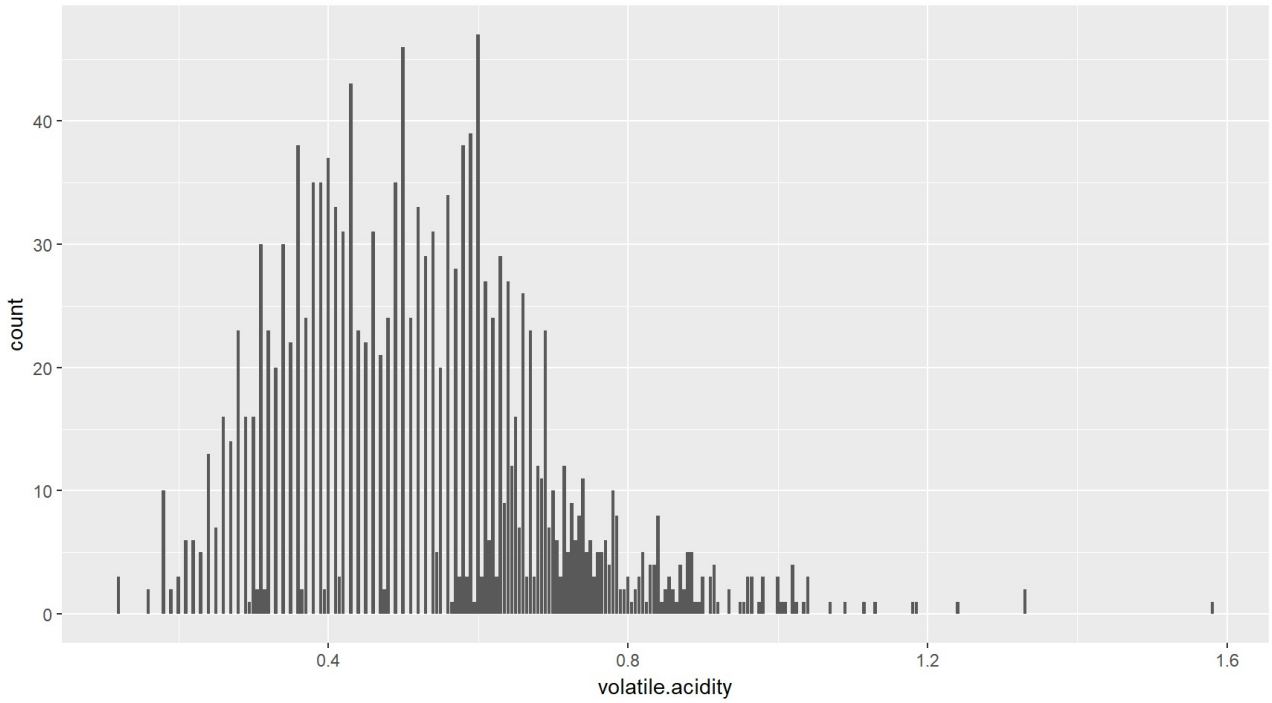
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
```



The distribution of fixed acidity is right skewed, and peaks at around 7. The median fixed acidity in the wines present in the dataset is 7.90 g/dm^3 . Most wines have an acidity between 7.10 and 9.20. The distribution of fixed acidity is slightly right skewed. There are some outliers in the higher range (~ >15).

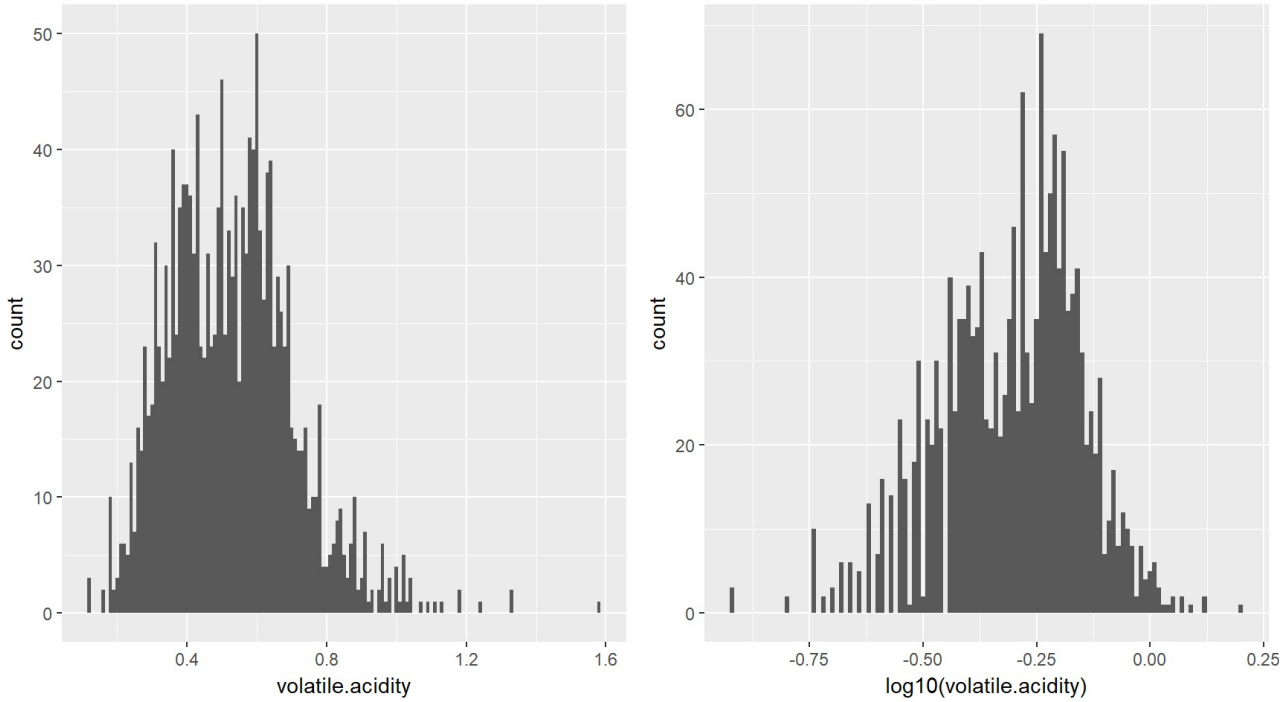
Volatile acidity

```
##    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 non-symmetric and bimodal with two peaks at 0.4 and 0.6. The median value is 0.52. Most observations fall in the range 0.39 - 0.64 and outliers on the higher end of the scale are visible.

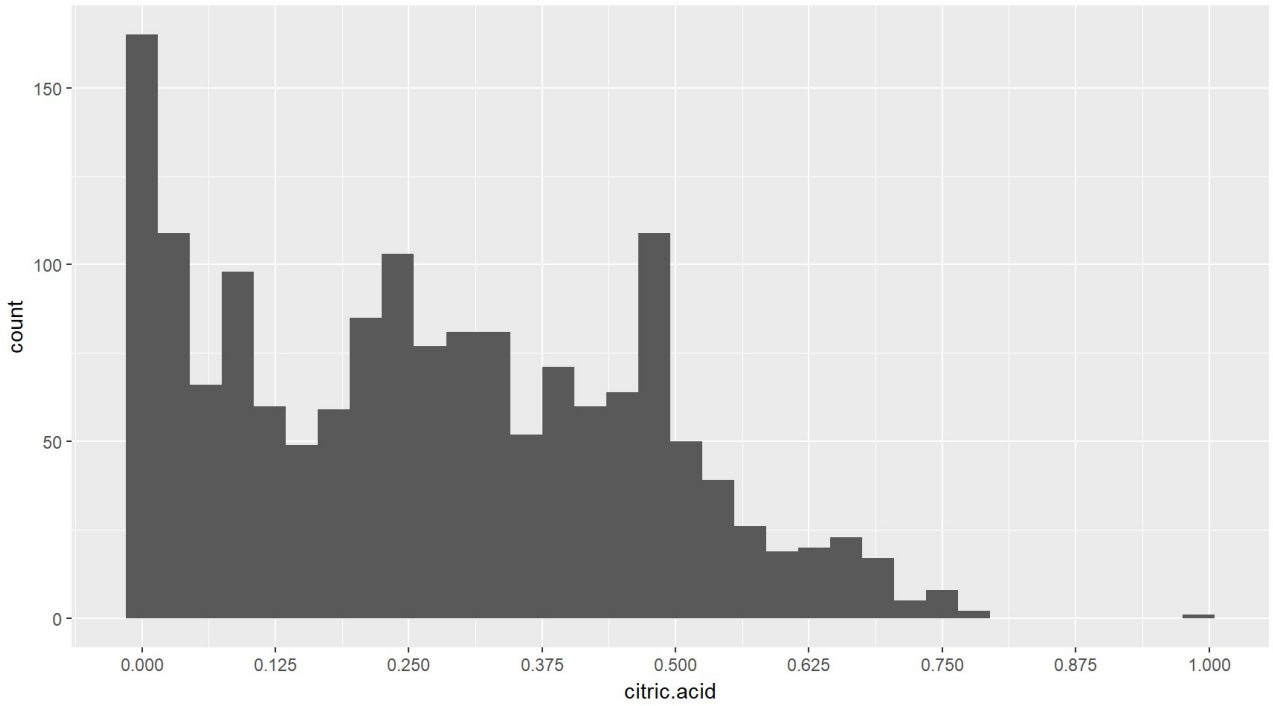
Compare log10 volatile acidity and volatile acidity to get a more normal distribution



By performing a log transformation on the volatile acidity distributions, allowed better visualizations of the data.

Citric acid

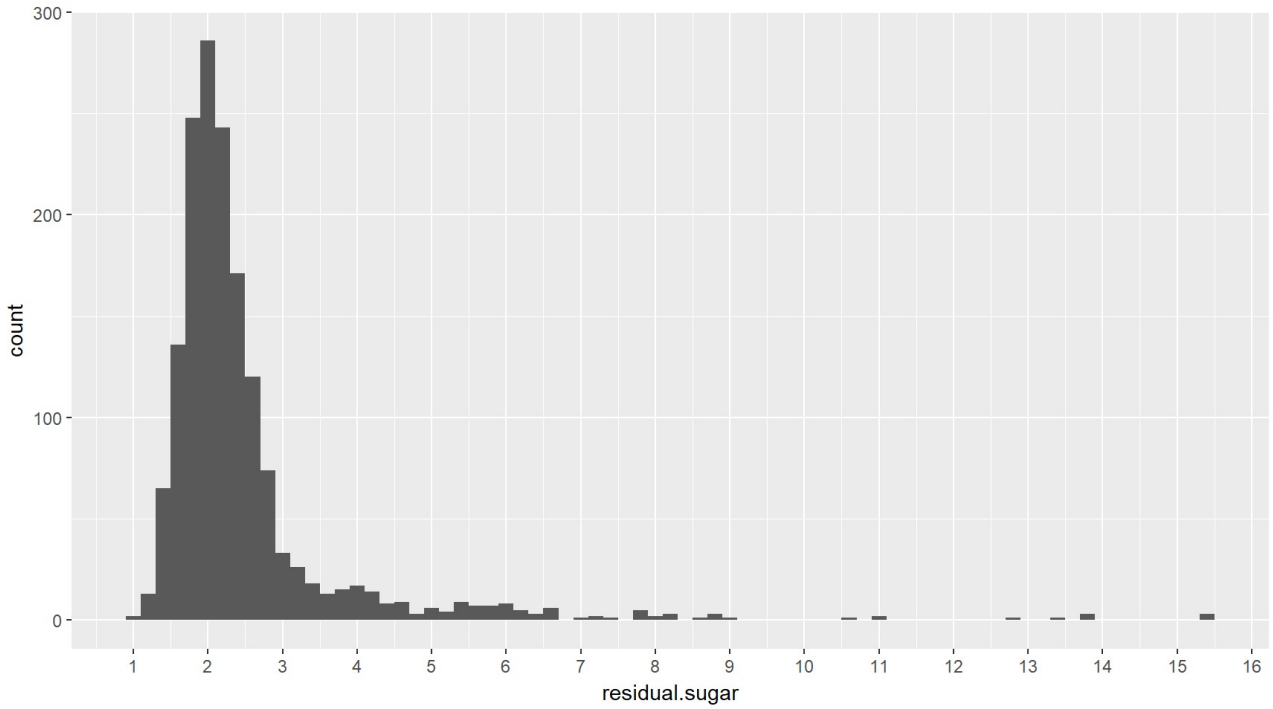
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```



The distribution of citric acid is not normal, most of the wines don’t even have citric acid.

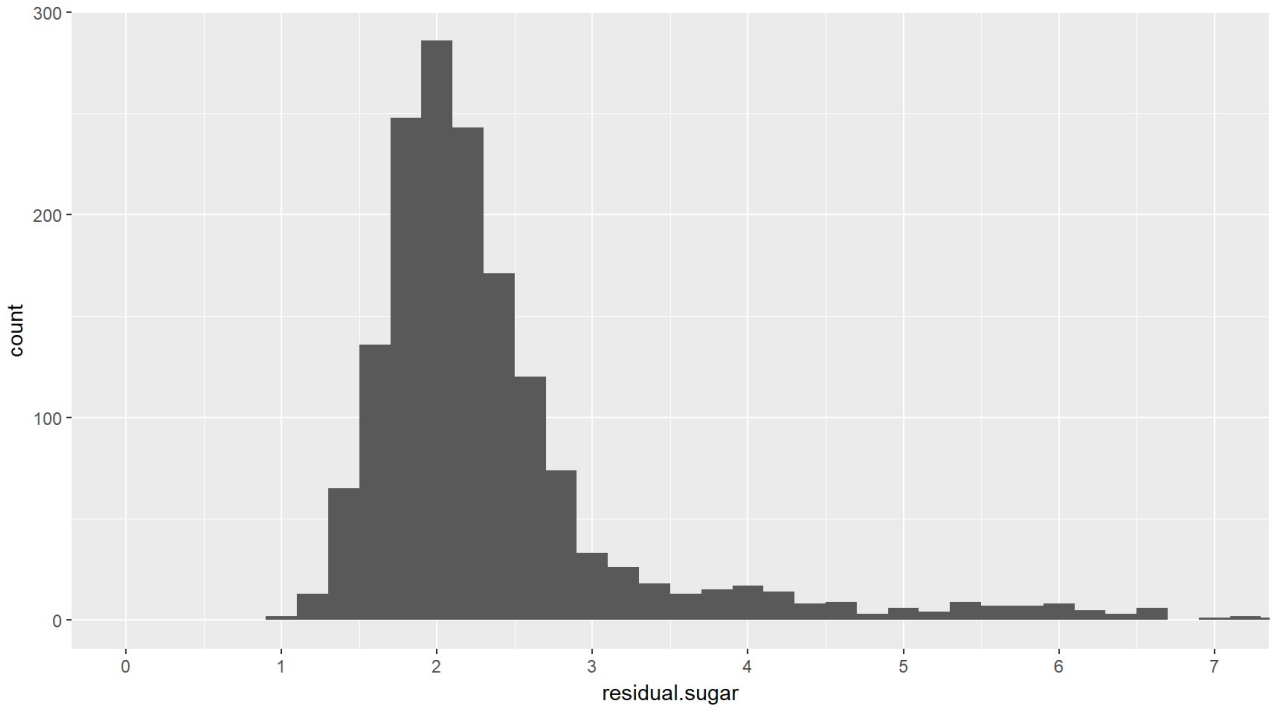
residual.sugar

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



In order to get a clear visualization of distribution, we are limiting range between 0 and 7 in x-axis.

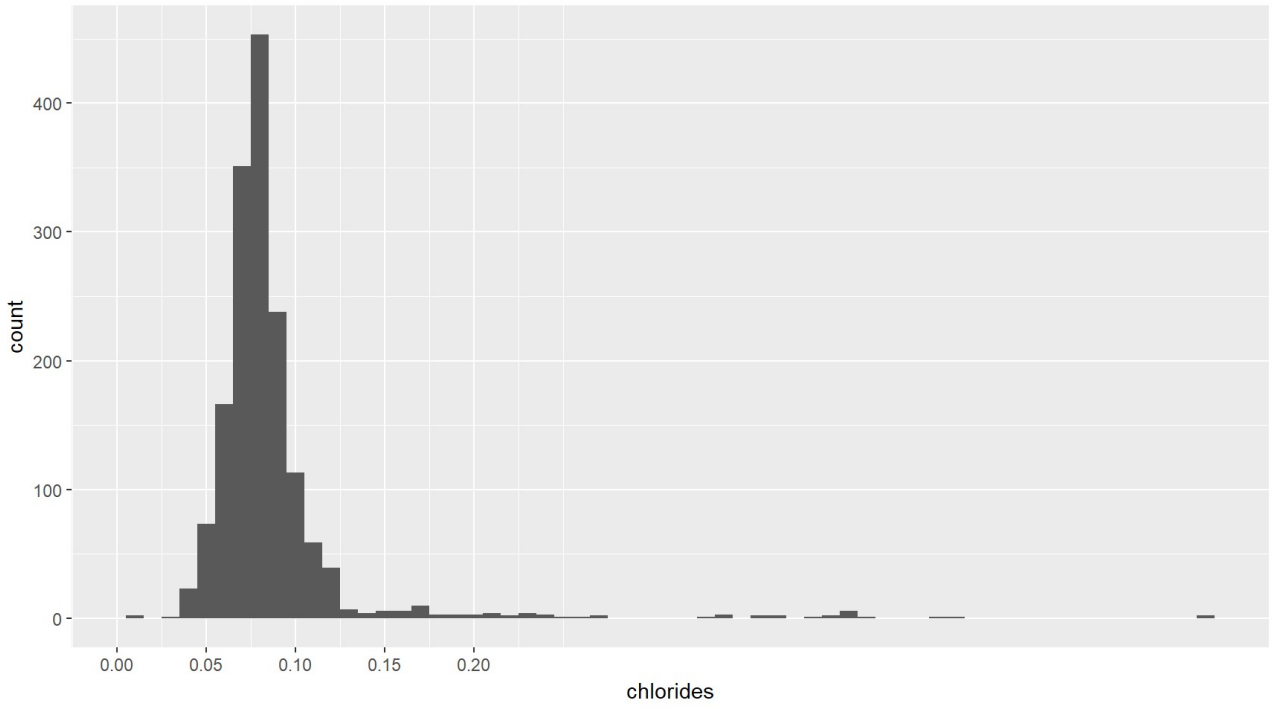
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```



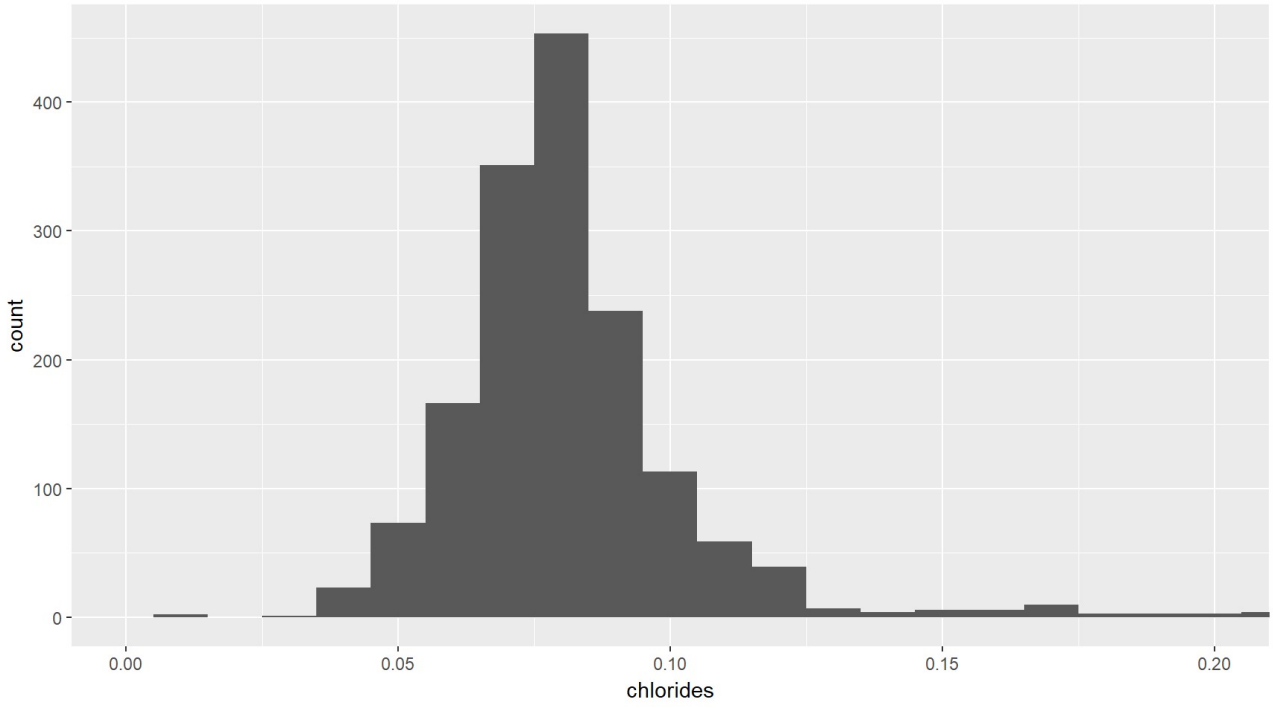
The distribution of residual sugar has a median value of 2.2 g/dm^3 . The distribution is right skewed with a long tail in the right side. The distribution peaks around $2g/dm^3$. There are many small bars on the right side of the main peak.

chlorides

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



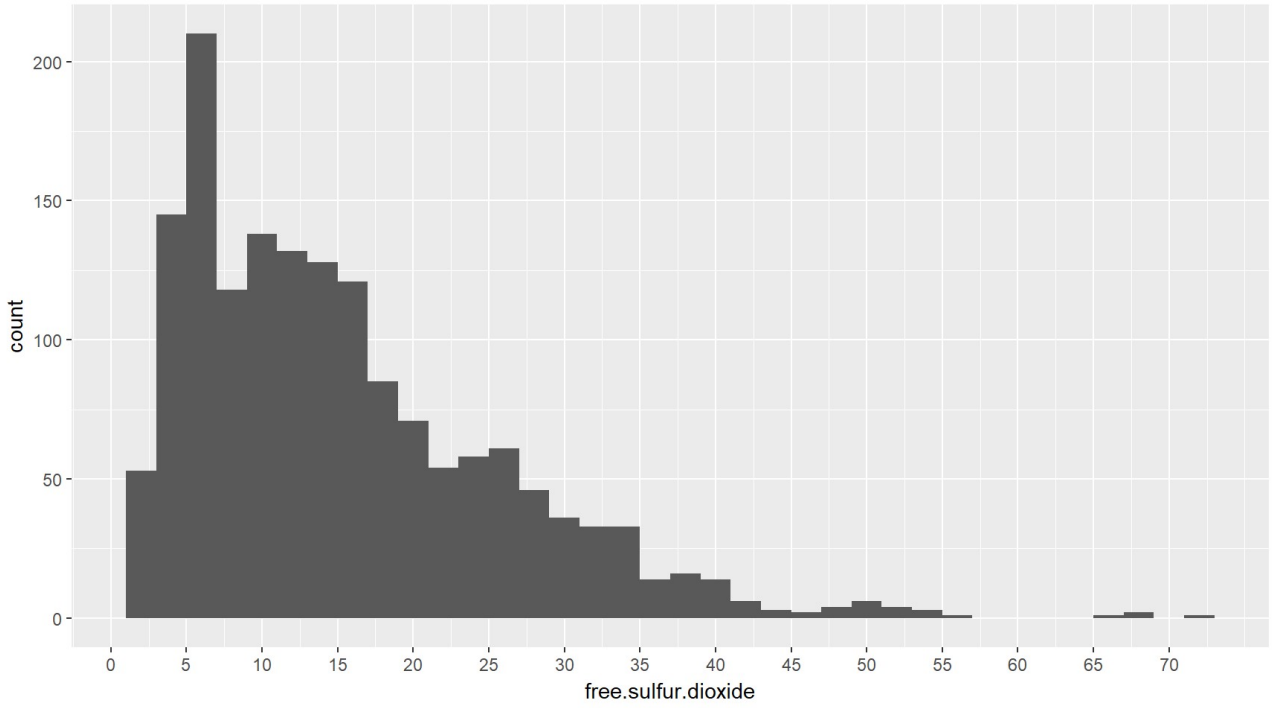
Since the distribution of chlorides is right skewed, let us zoom into the range between 0 and 0.2 values of chlorides,for better visualization and understanding .



The amount of chlorides in the wines has a median value of 0.079 $g/dm3$. The distribution of chlorides is right skewed and concentrated at around 0.09, with small counts of wines with values until 0.611 g/dm^3 .

free.sulfur.dioxide

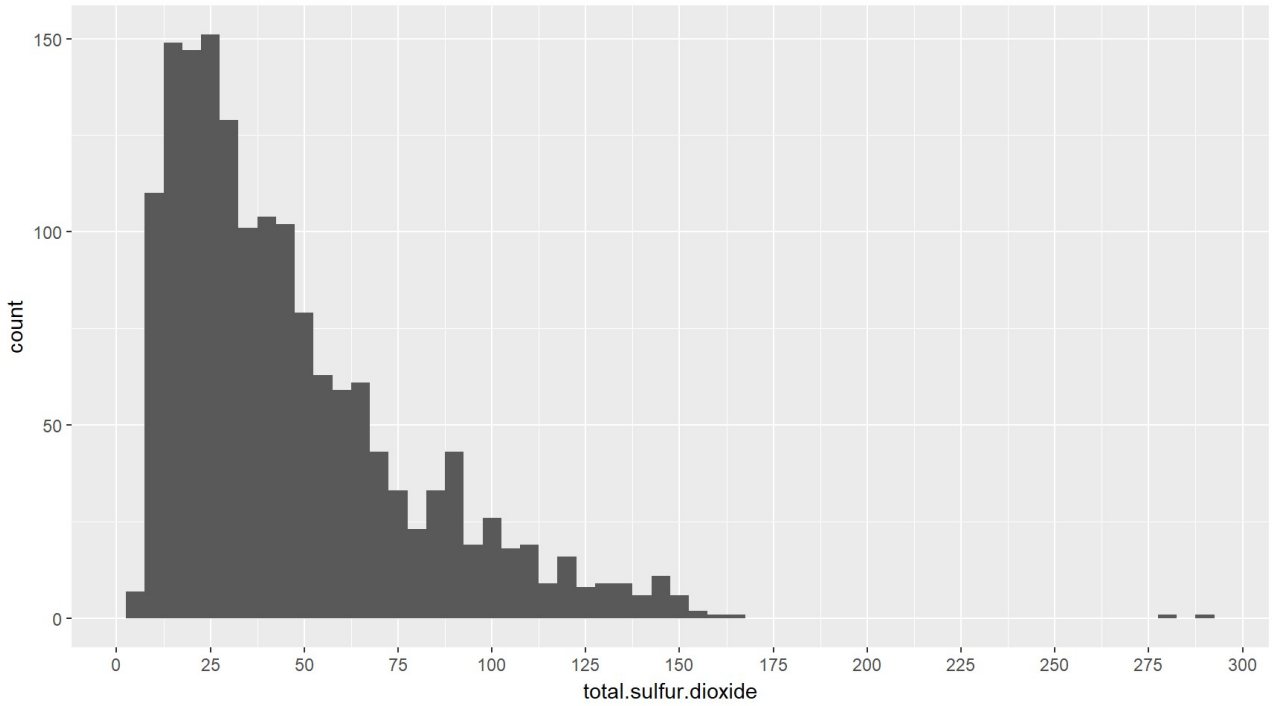
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
```



The Distribution of free sulfur dioxide is right skewed and the median value is 14 mg/dm^3 .The right tail extends until a maximum of 72 mg/dm^3 .

total.sulfur.dioxide

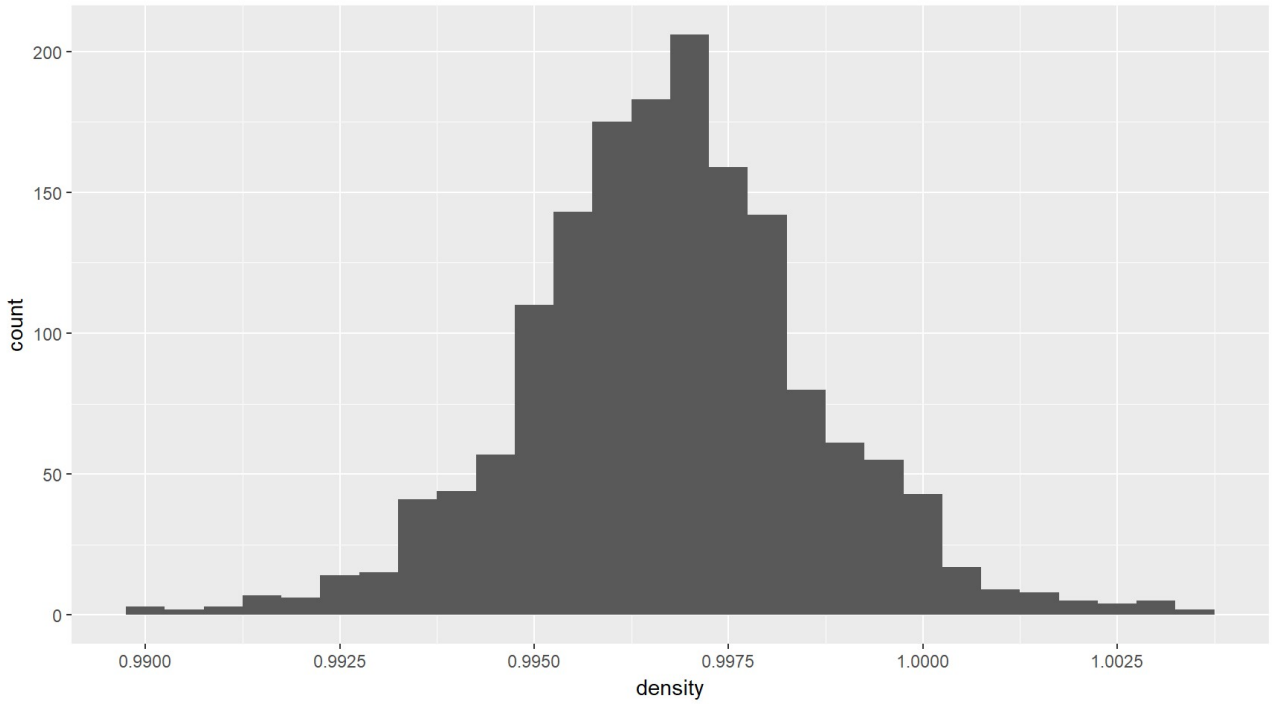
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00
```



The distribution of total sulfur dioxide is right skewed with a median value of 38 mg/dm^3 .

density

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



Density is normally distributed. Most of the density values varies between 0.9956 and 0.9967and has

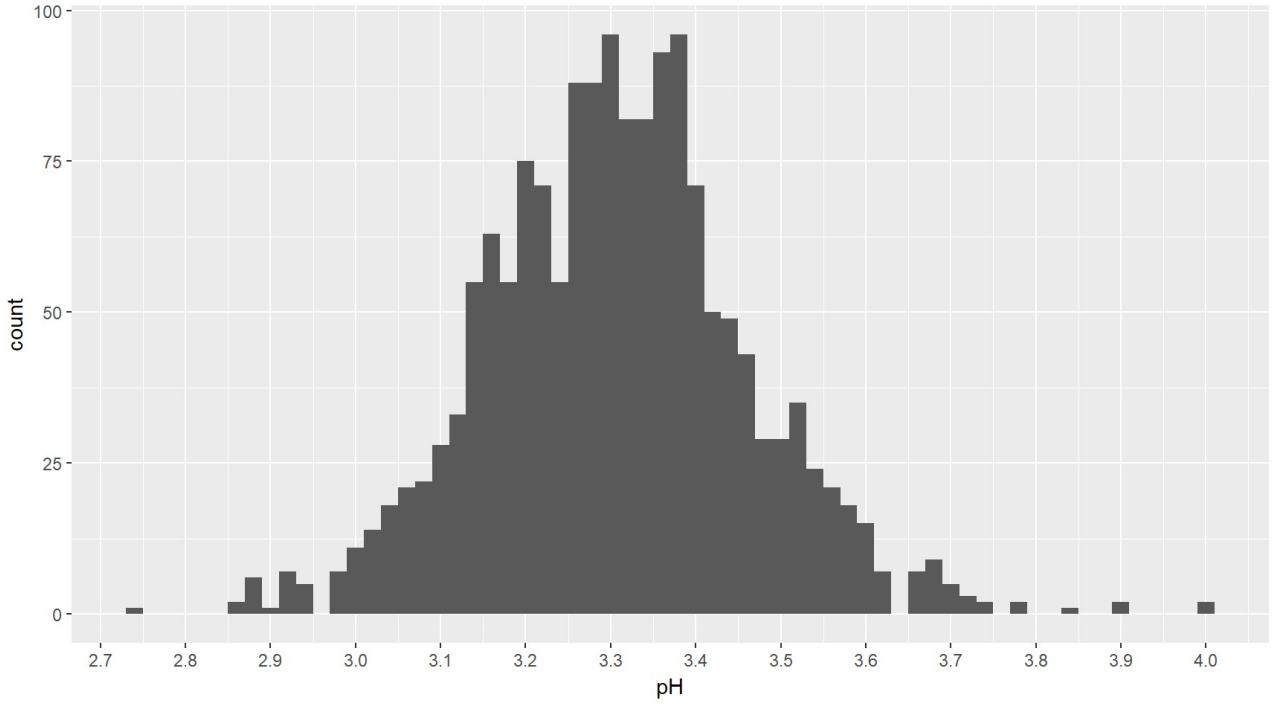
median value of 0.9968 .

g/c

pH

$m3$

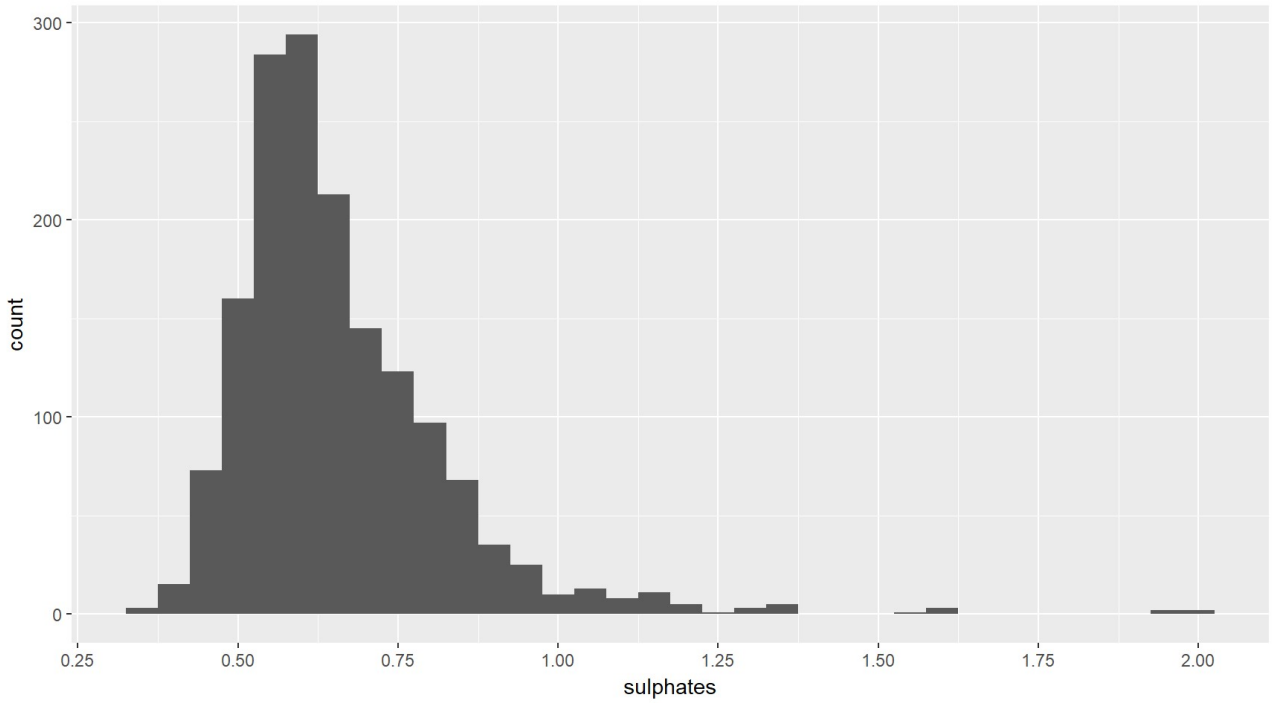
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010
```



pH is normally distributed or could be also considered bimodal with both peaks very close to each other. The median value is 3.31, and most wines have a pH between 3.21 and 3.4.

sulphates

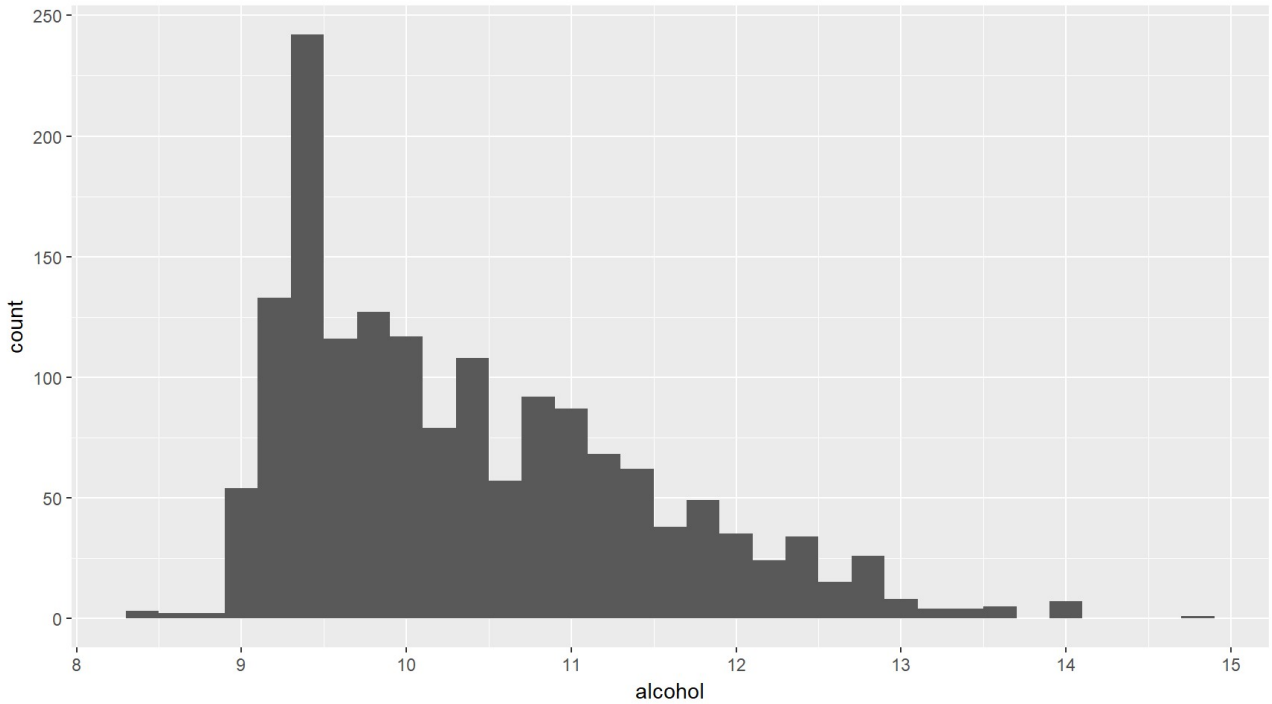
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```



The distribution of sulphates is right skewed with outliers on the right tail around 2 g/dm^3 of sulphates. The median value of sulphates is 0.62 and most wines have a concentration between 0.55 and 0.73.

alcohol

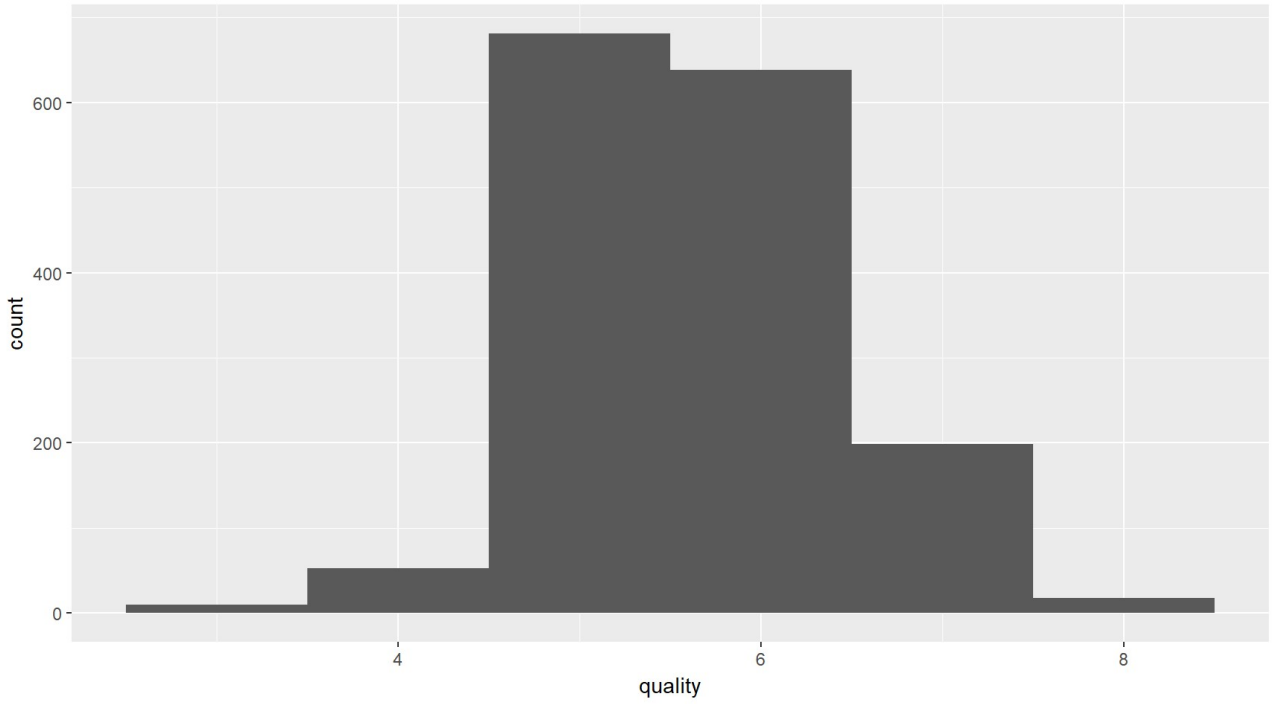
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```



The alcohol distribution seems to be right skewed.The minimum amount of alcohol is 8.4 % alcohol,which maybe the minimum alcohol needed for the wine. The highest peak of the distribution is at 9.5 % alcohol and the median value is 10.20%. The maximum amount of alcohol present in the wine is 14.90%.

Quality

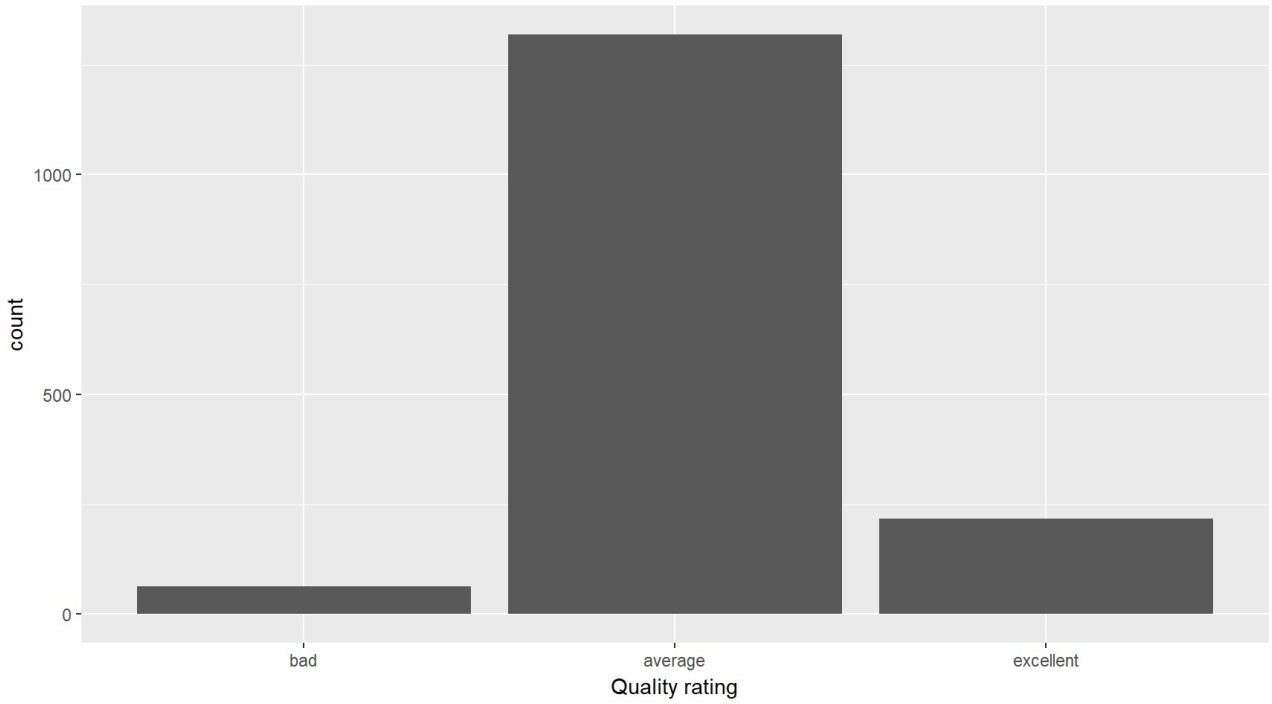
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.636   6.000   8.000
```



We can say the distribution of quality appears to be normal with many wines at average quality (4-5) and fewer wines at low quality and high quality. There are no wines with a quality worse than 3 and no wines with quality higher than 8. The vast majority of red wines have a quality ranking of 5 and 6.

It appears that we can actually group wine quality into three distinct categories: bad, average, and excellent. Most of the red wines were average, with a few having excellent tastes and then bad. Let’s explore what makes a wine excellent or bad .

```
##       bad   average excellent
##        63      1319       217
```



Univariate Analysis

What is the structure of your dataset?

The redwine dataset consists of 12 variables and 1599 observations.Among the 12 variables, 11 variable correspond to the result of a physicochemical test and one variable (`quality`) corresponds to evaluations of quality made by wine experts, varying from 0 (very bad) to 10 (very excellent).Each observation corresponds to a red wine sample.

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

The main feature of interest is the quality.I'd like to determine which features influence the wine quality and then building a predictive model of quality using these variables.

What other features in the dataset do you think will help support your

investigation into your feature(s) of interest?

Most features have an aproximately normal distribution, just like the quality variable. I think all the physicochemical test results may help support the investigation.

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

Yes, I created a new variable named ‘rating’. This variable divides the quality into 3 different categories, ‘bad’, ‘average’ and ‘excellent’. Even though the red wine is given a rating of 1 to 10, the lowest grade wine has a quality of 3, and the highest grade wine has a quality of 8.

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?

There were no missing values and no need to adjust the data. The dataset presented is already tidy which makes it an ideal dataset.

The distribution of volatile acidity presented two unusual peaks which standed out of an otherwise normal distribution.

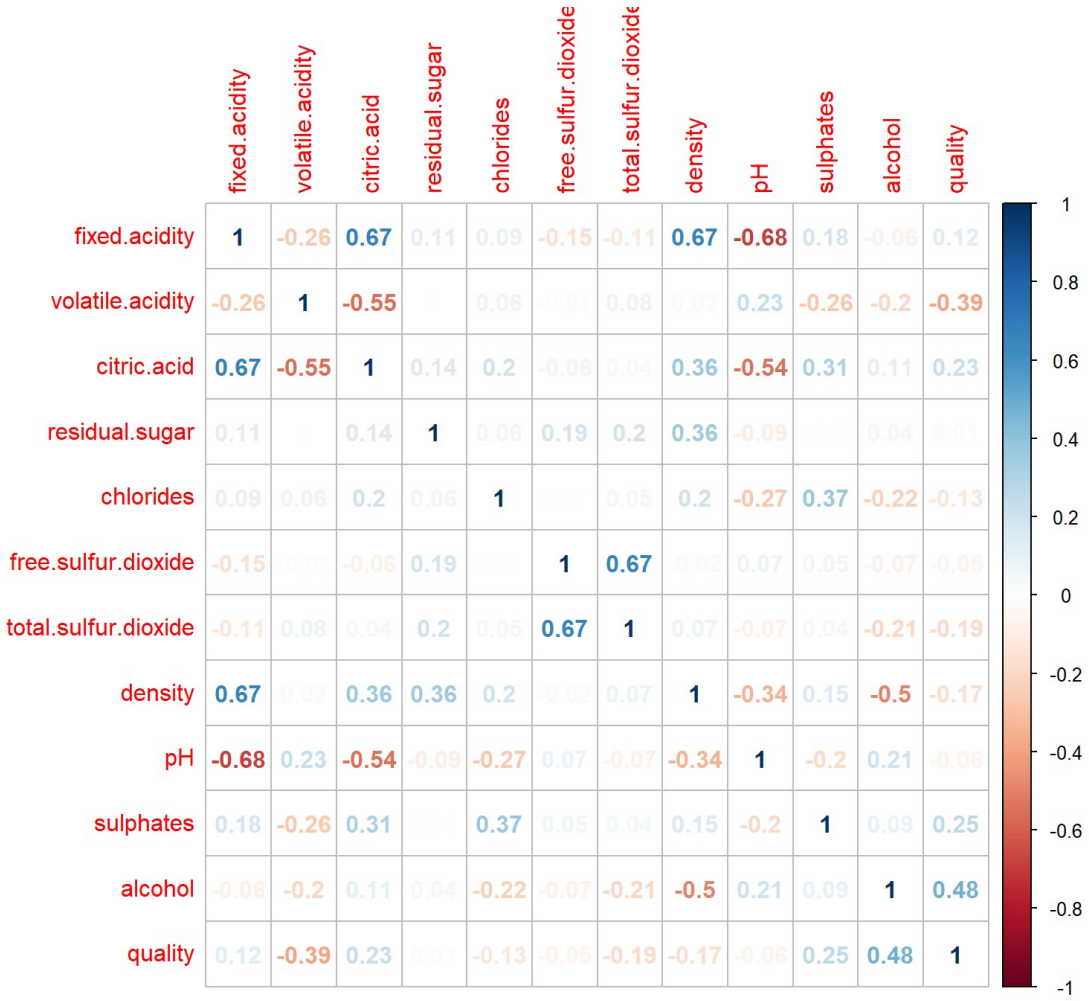
I preformed a log transformation on the volatile acidity distributions, because they were very skewed, and the transformations allowed better visualizations of the data.

Bivariate Plots Section

I used a correlation chart to help me find any relationships between the variables.

```
##                       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 top 4 correlation coefficients with quality are:

- alchol:quality = 0.48
- sulphates:quality = 0.25
- citric.acid:quality = 0.23
- fixed.acidity:quality = 0.12

So as we saw earlier, alcohol content has a high correlation with red wine quality. Other important attributes correlated with red wine quality include sulphates, citric acid and fixed acidity.

The biggest negative corrlation coefficients with quality are:

- volatile.acidity:quality = -0.39
- total.sulfur.dioxide:quality = -0.19
- density:quality = -0.17
- chlorides:quality = -0.13

So we see that volatile acids are negatively correlated with red wine quality. Tota sulfur dioxide, density and chlorides are also negatively correlated with quality.

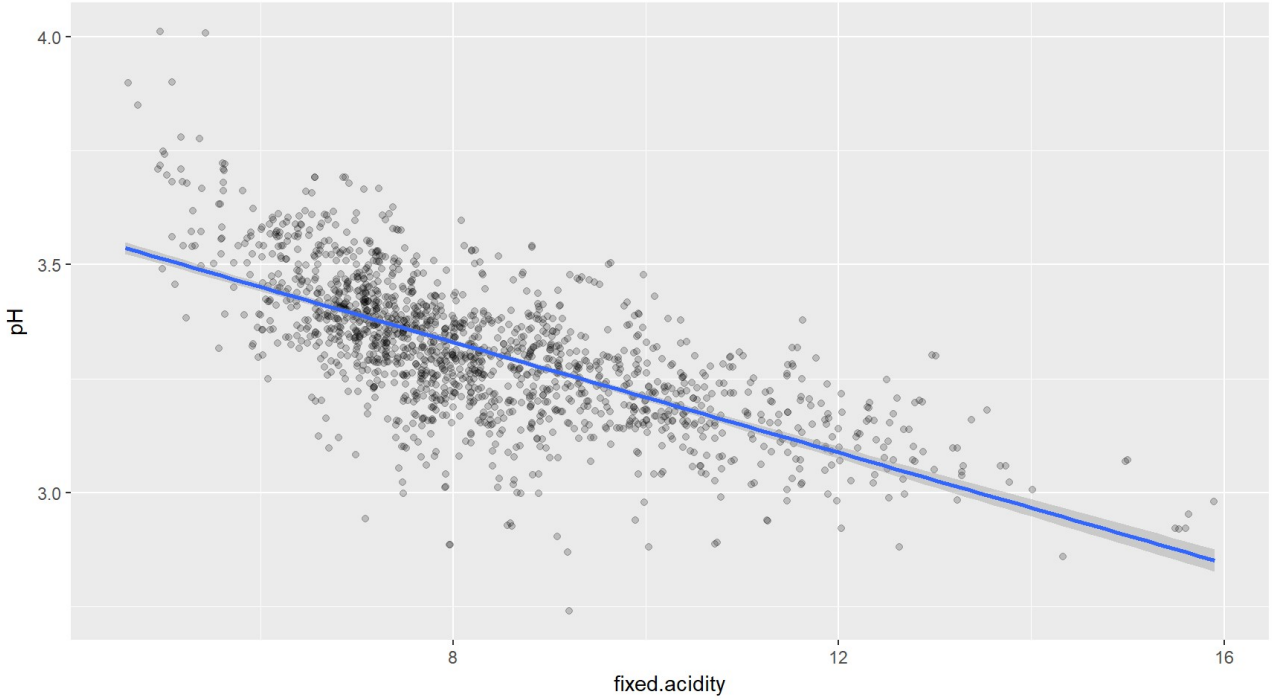
Variables with the highest (positive or negative) correlation include:

- fixed.acidity:citirc.acid = 0.67
- fixed.acidity:density = 0.67
- free.sulfur.dioxide:total.sulfur.dioxide = 0.67
- alcohol:quality = 0.48
- density:alcohol = -0.50
- citric.acid:pH = -0.54
- volatile.acidity:citirc.acid = -0.55
- fixed.acidity:pH = -0.68

Exploring relationships in a bit more detail.

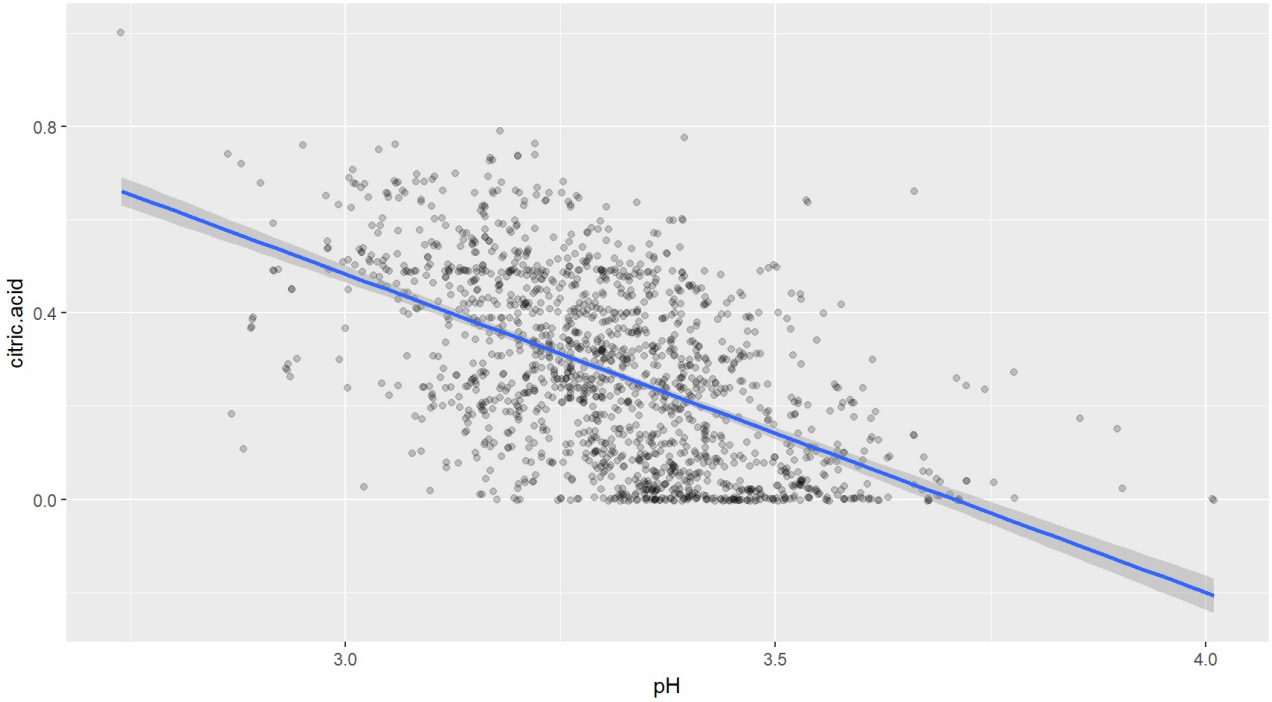
Acidity and pH:

Fixed acidity and pH:



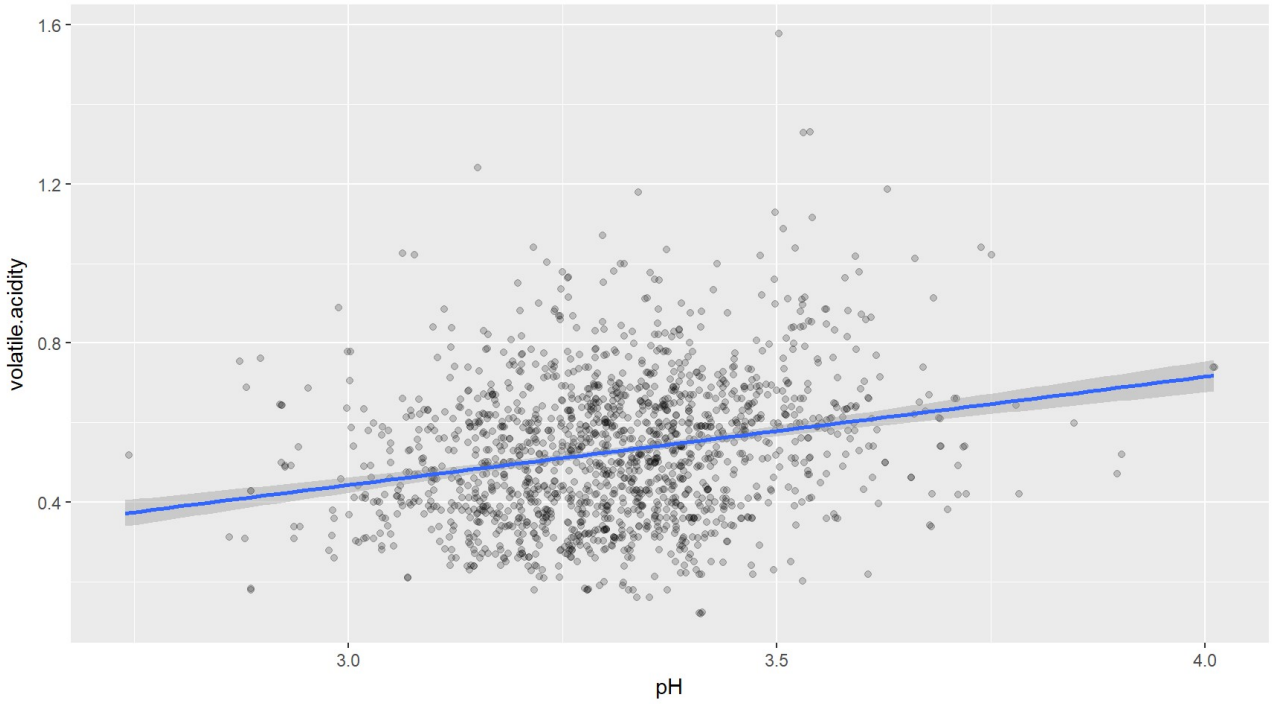
As expected the pH increases with the lower amount of acids. Fixed acidity accounts for most acids present in the wine.

Citric acid and pH



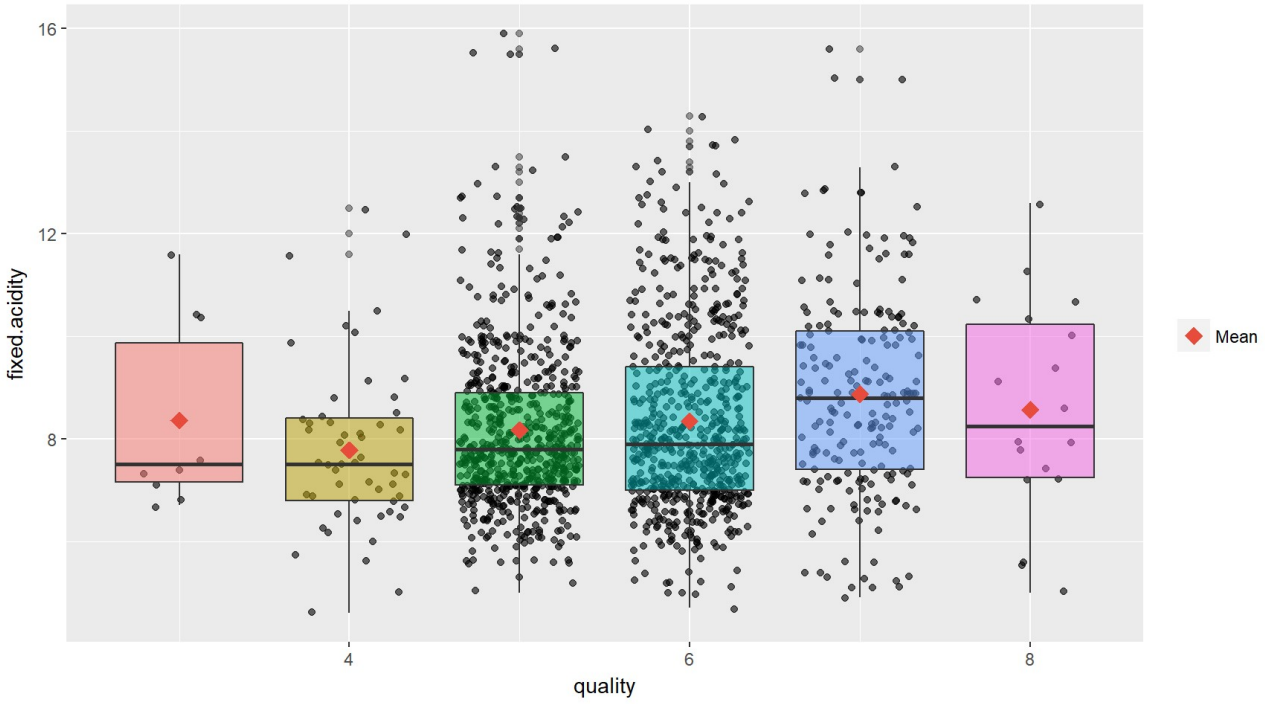
A similar relation is seen with the citric acid variable,pH increases with the lower amount of acids.

volatile acidity and pH



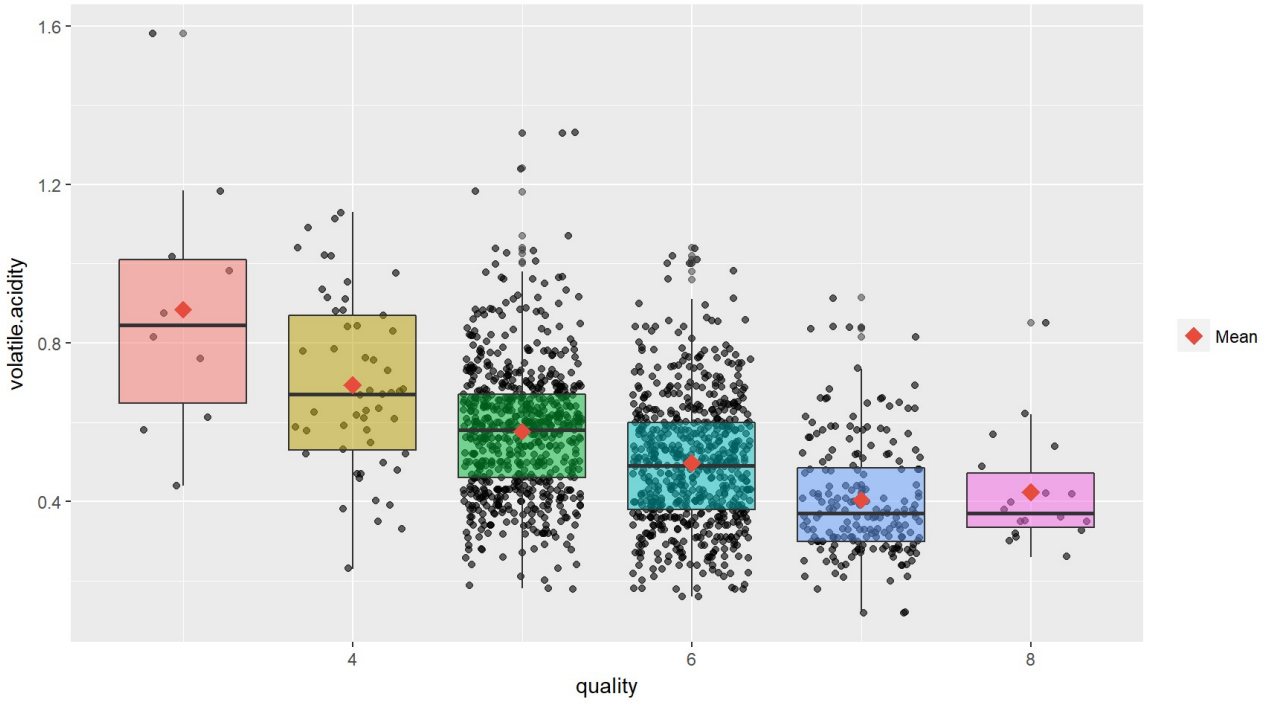
The volatile acidity seems to have either no relation with the pH or a slight positive correlation. The only acid concentration that shows some considerable correlation with pH is the fixed acidity.

Fixed Acidity vs. Quality



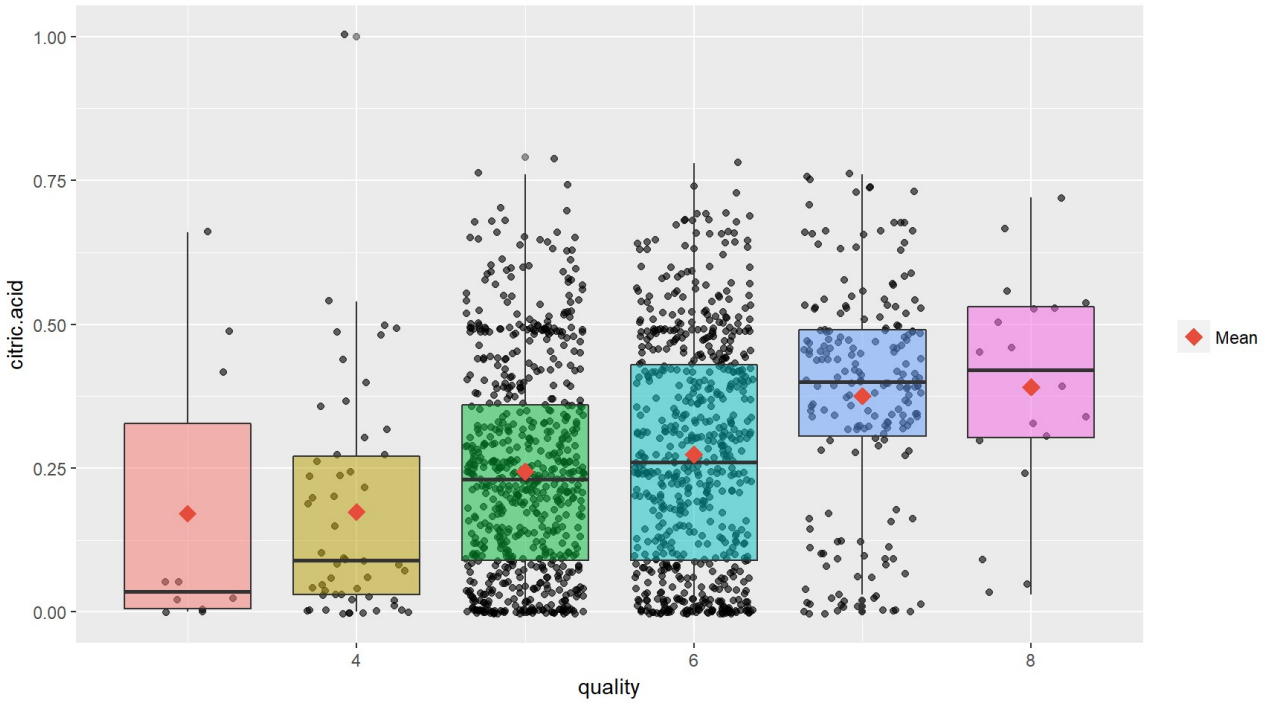
There is a slight upwards trend of higher quality with higher fixed acidity. However, the extreme quality classes (3 and 8) have less observations than the middle ones, which may make the mean value not so accurate. And we see a drop of acidity from 7 to the 8 quality class. Additionally, we see a big dispersion of acidity values across each quality scale. This may be a indicator that the quality cannot be predicted based only on the value of acidity and is the result of a combination of more variables.

Volatile Acidity vs. Quality



Here we observe Lower volatile acidity seems to mean higher wine quality.

Citric Acid vs. Quality



Here we observe Higher citric acid means higher quality wine. The citric acid is always in low concentrations and in the univariate plots we saw that the distribution peaked at the zero value.

Let’s see which proportion of wines has zero citric acid. For all the wines that proportion is:

```
## [1] 0.08255159
```

For each quality class the proportions are:

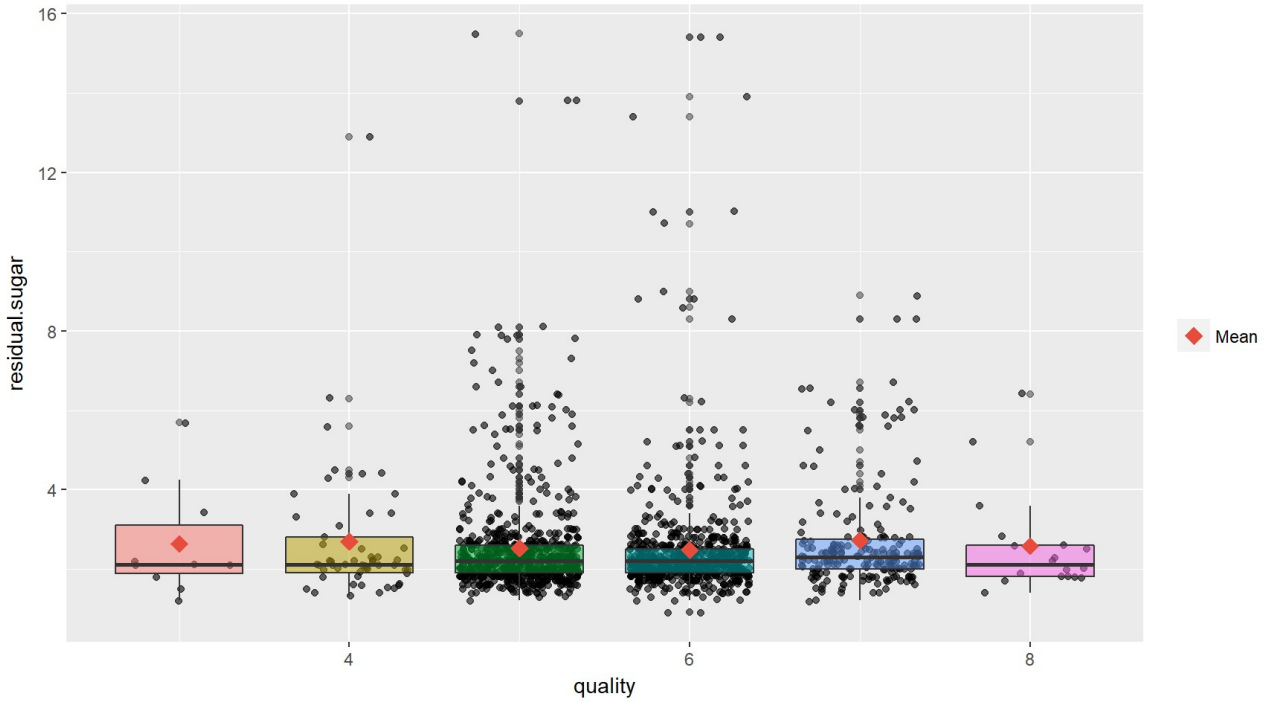
```
## # A tibble: 6 x 2
##   quality zero_citric
##     <int>       <dbl>
## 1       3  0.30000000
## 2       4  0.18867925
## 3       5  0.08370044
## 4       6  0.08463950
## 5       7  0.04020101
## 6       8  0.00000000
```

We see a decreasing proportion of wines with zero citric acid on the higher quality classes.

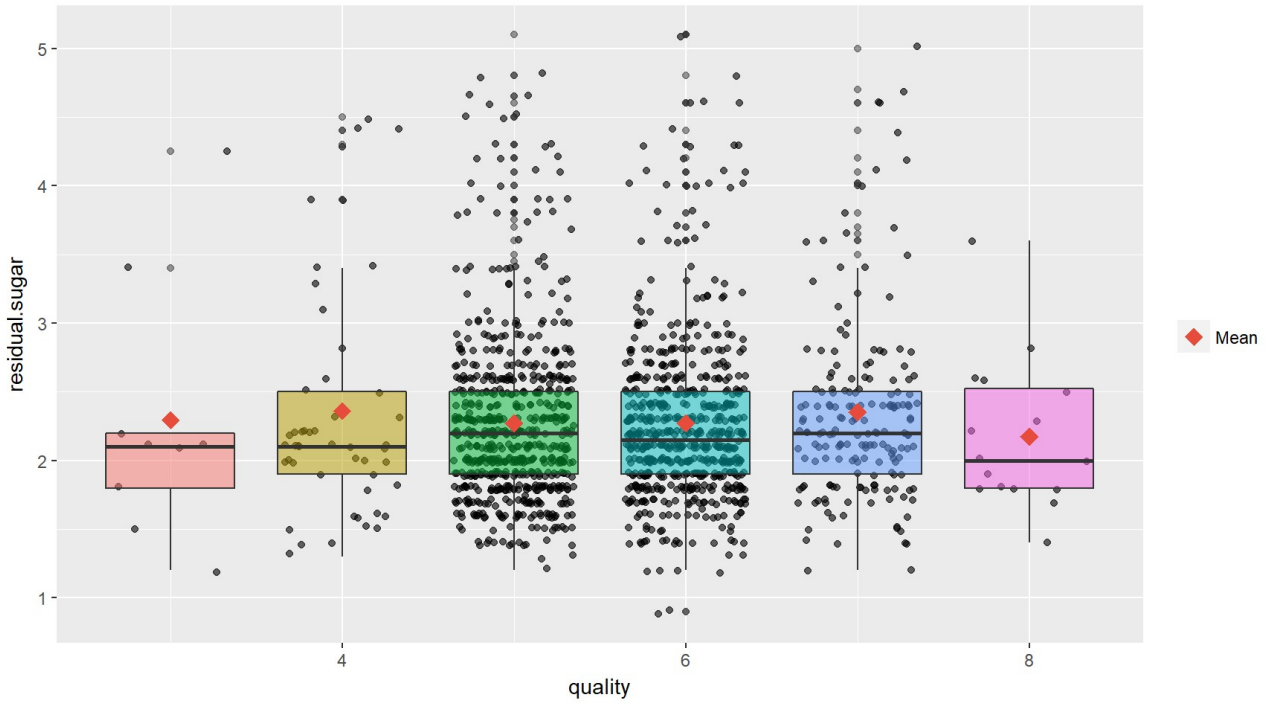
So, this supports our initial finding, the higher citric acid concentration relates to higher quality wines.

Here are the summary statistics for residual sugar:

Residual Sugar vs. Quality

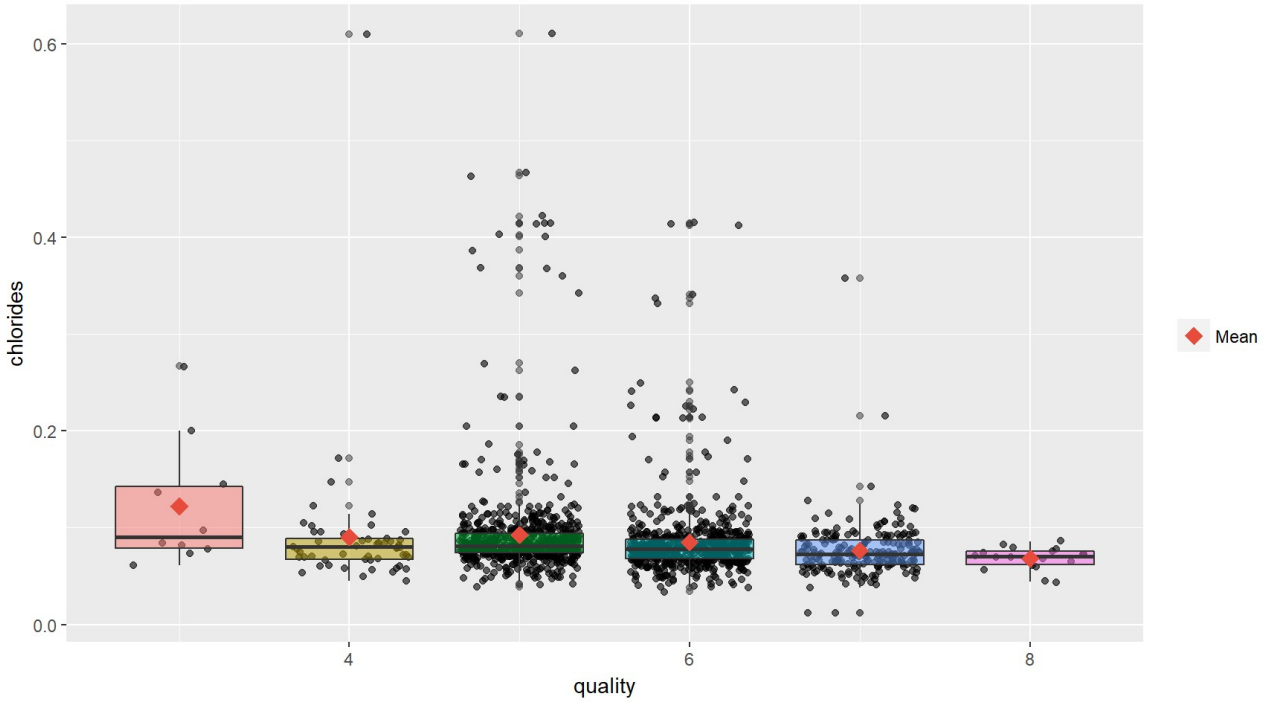


removing the top 5% to be able to have a better look



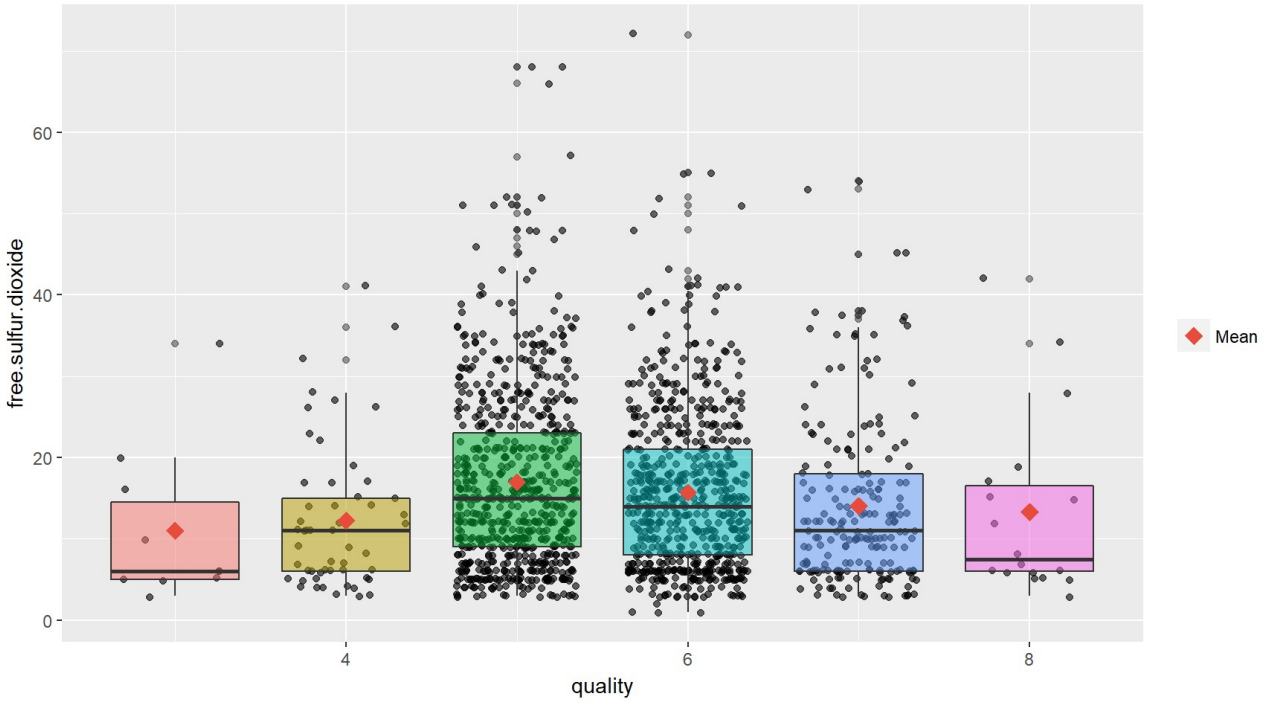
Residual sugar seems to have a low impact in the quality of the wine.

Chlorides vs. Quality



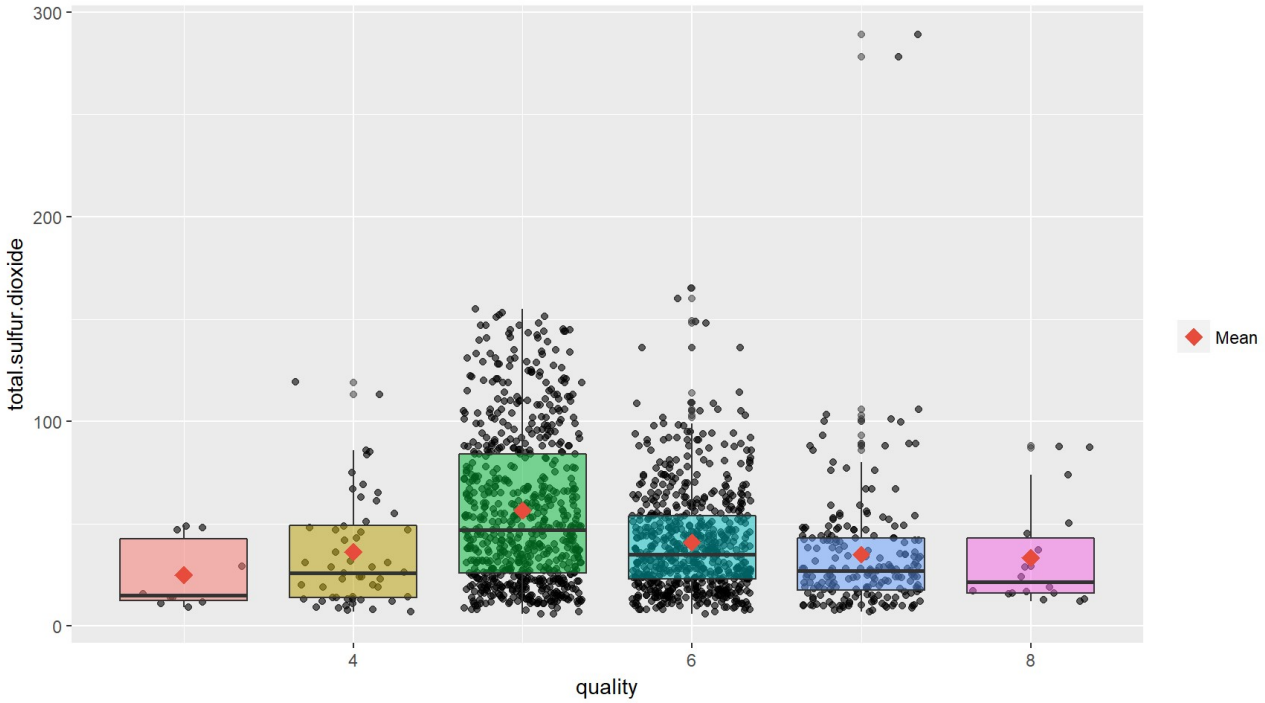
Here in the plot it shows slight variation. Less chlorides means higher quality.

Free sulfur dioxide vs. Quality



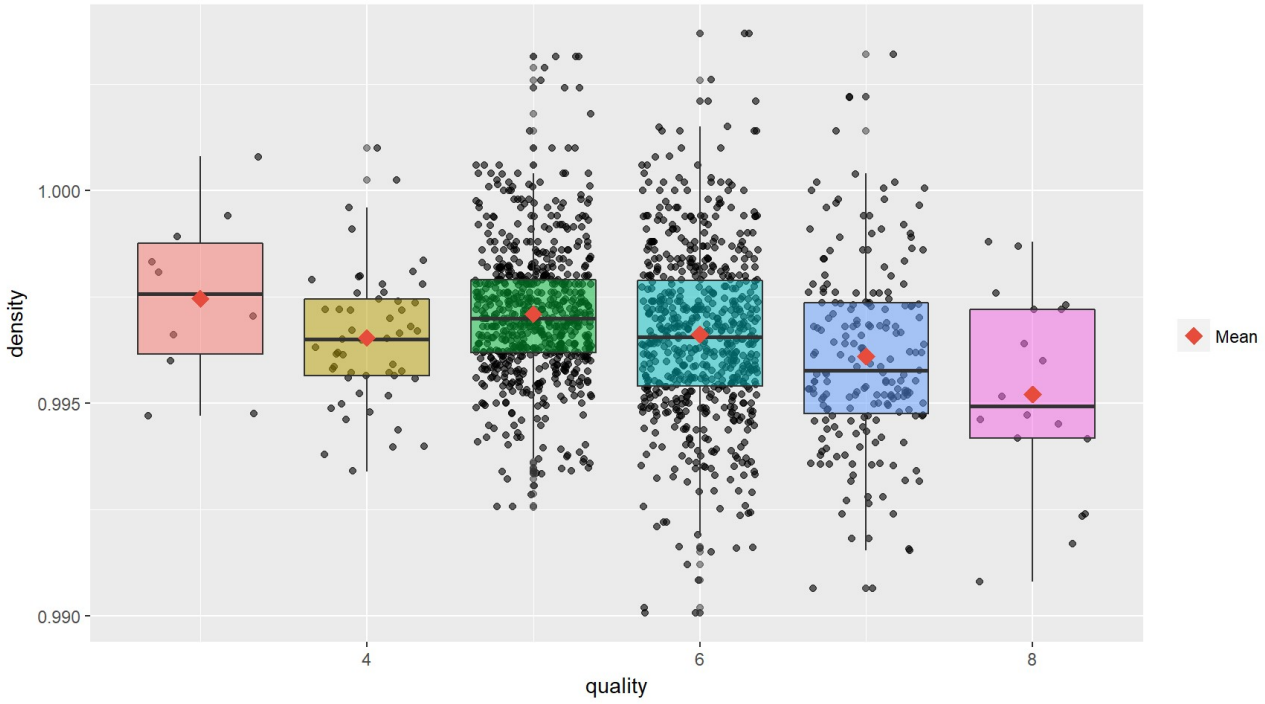
The middle quality classes seem to have higher free sulfur dioxide than both the low and high quality.

Total sulfur dioxide vs. Quality



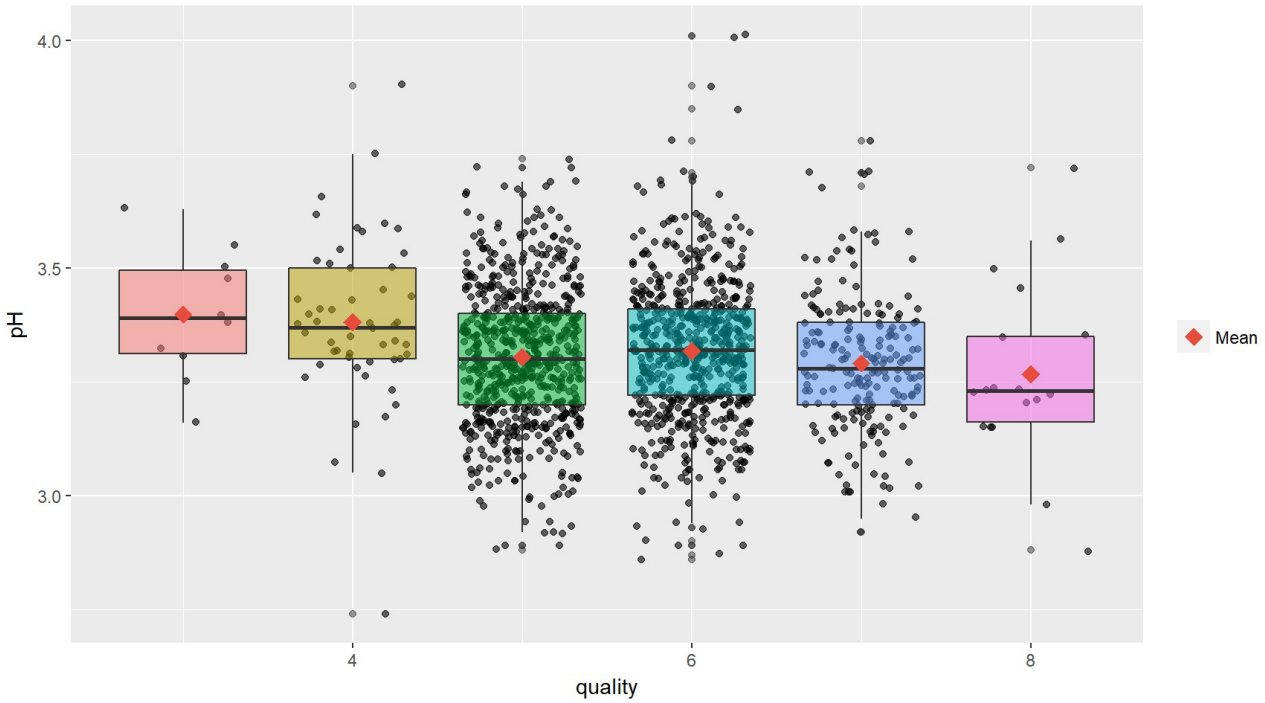
Above plot shows Similar relation as with free sulfur dioxide. The middle classes have higher concentration than both the low and high quality wine.

Density vs. Quality



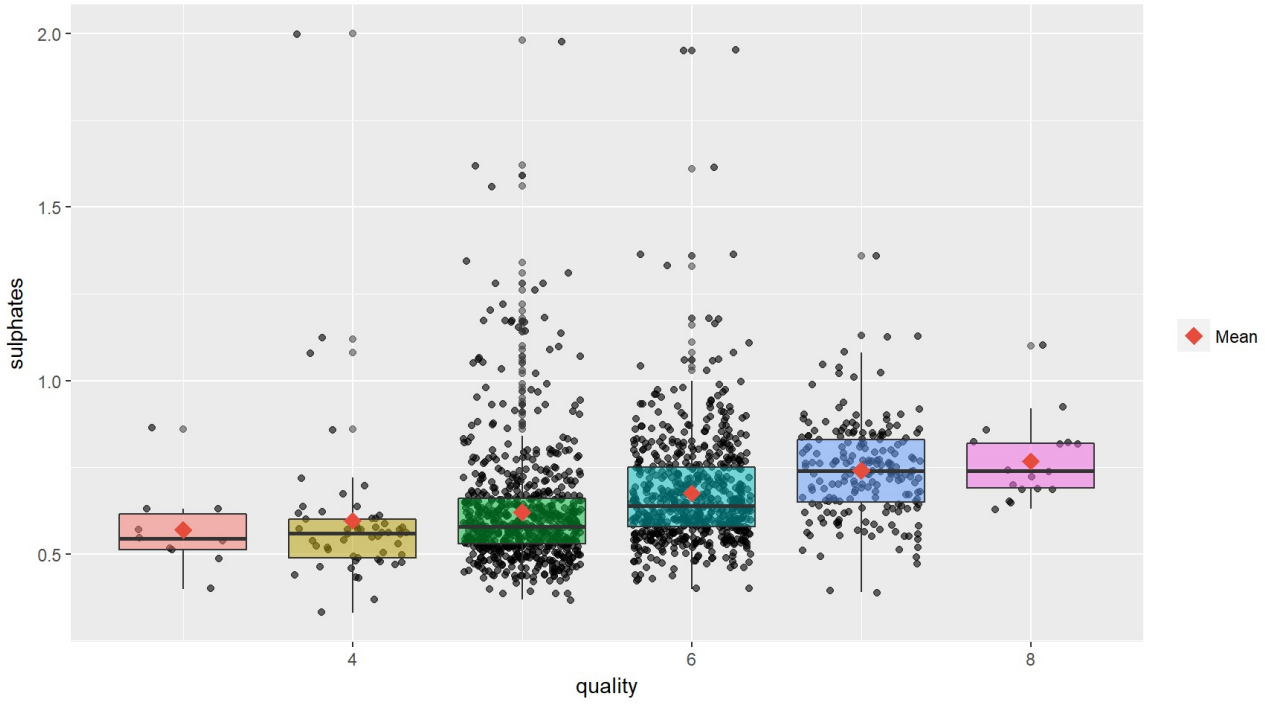
From the plot it shows Lower density means higher quality.

pH vs. Quality



There seems to be a trend of higher quality with lower pH.

Sulphates vs. Quality

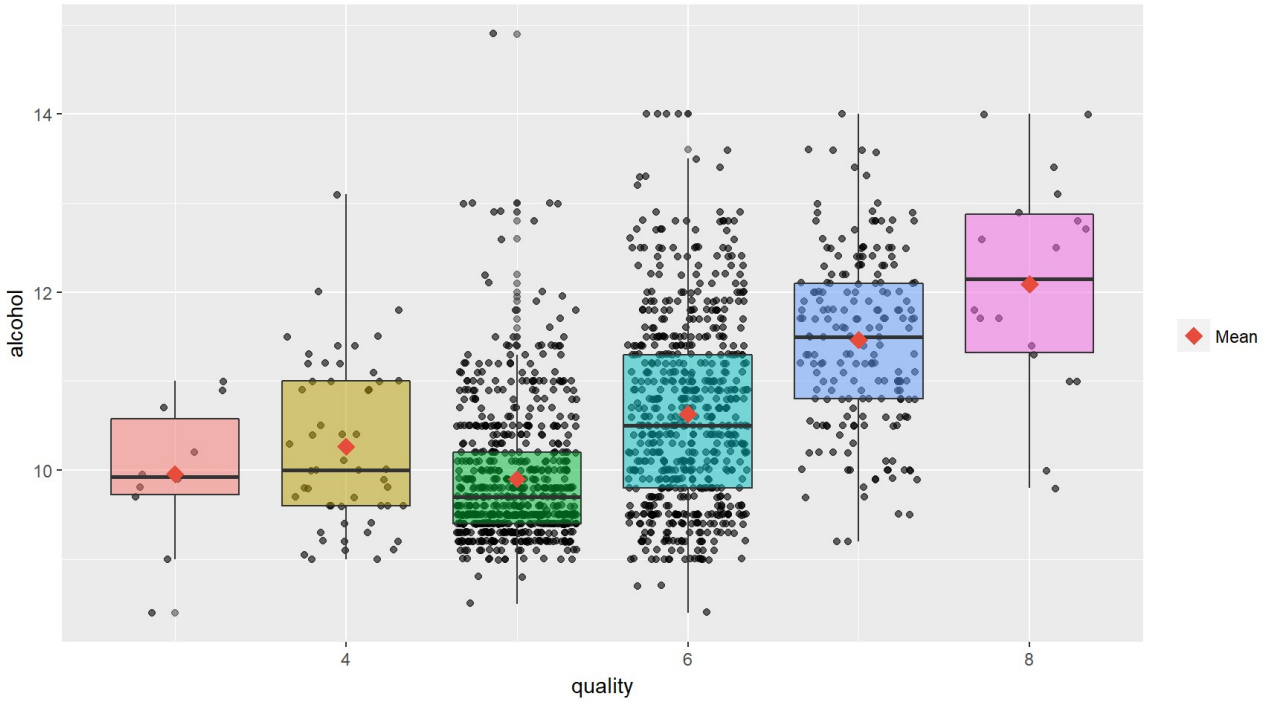


It seems that Higher sulphates concentration means higher quality.

And here are the summary statistics for sulphates at each quality level:

```
## $`3`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4000  0.5125  0.5450  0.5700  0.6150  0.8600
##
## $`4`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.4900  0.5600  0.5964  0.6000  2.0000
##
## $`5`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.370   0.530   0.580   0.621   0.660   1.980
##
## $`6`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4000  0.5800  0.6400  0.6753  0.7500  1.9500
##
## $`7`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3900  0.6500  0.7400  0.7413  0.8300  1.3600
##
## $`8`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.6300  0.6900  0.7400  0.7678  0.8200  1.1000
```

Alcohol vs. Quality



It looks like the red wines with a higher alcohol content tend to have a higher quality rating. The main

anomoly to this trend appears to be red wines having a quality ranking of 5.

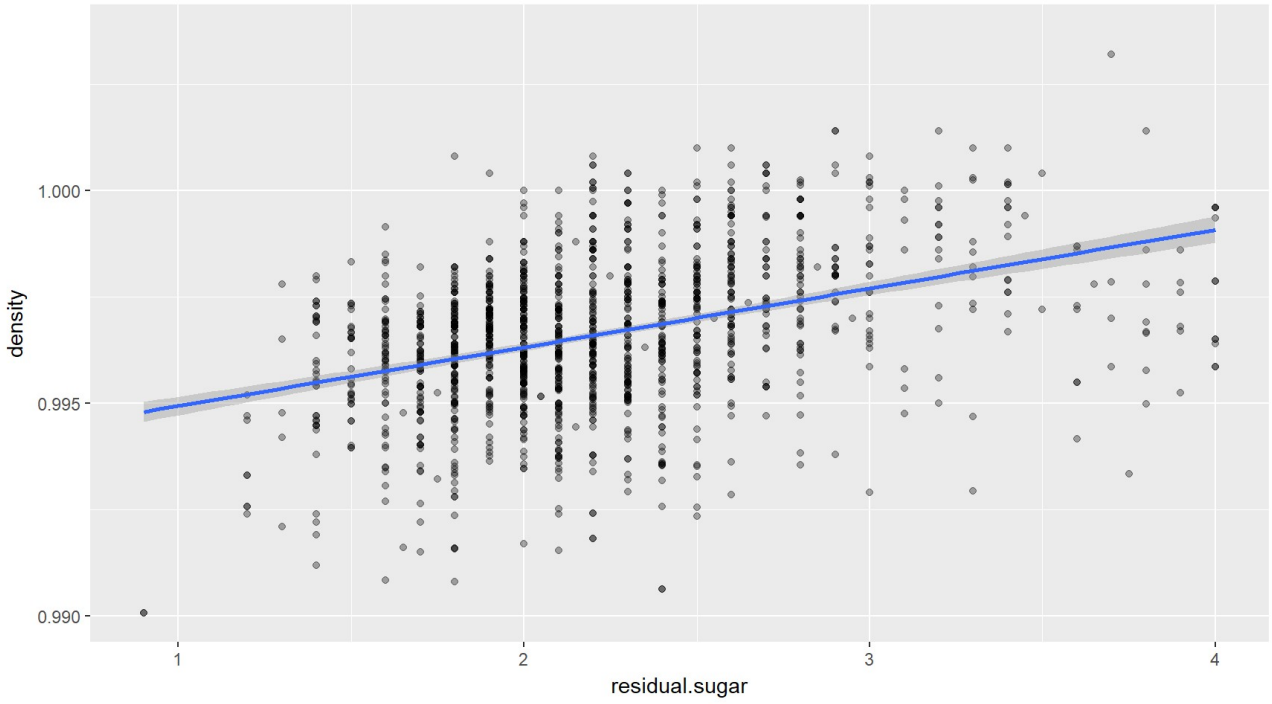
Here are the summary statistics for alcohol content at each quality level:

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

Density, Sugar and Alcohol Content

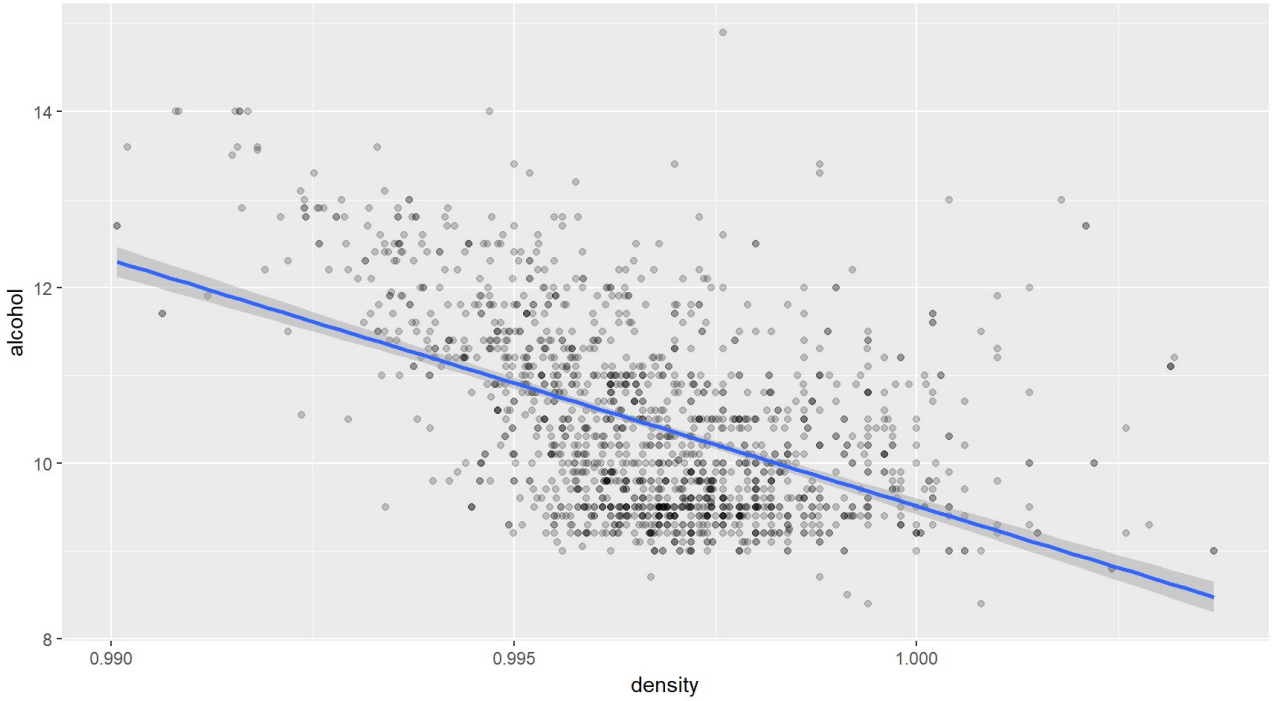
The density of wine should be close to the water density, and will change depending on the percent of alcohol and sugar content.

Density and residual sugar:



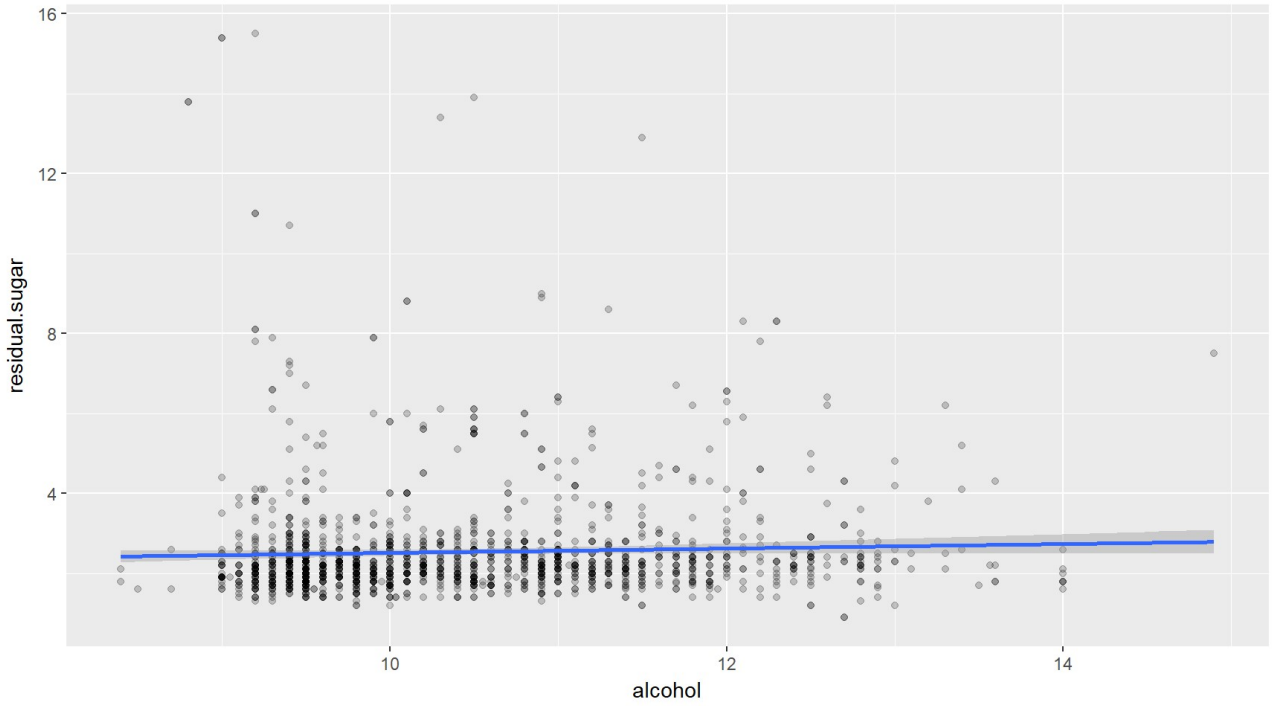
There is an increase of density with increase of residual sugar.

Density and alcohol:



There is a strong correlation between density and alcohol as we can see from this scatter graph.And we see a decrease of density with increase of alcohol content.

Residual sugar and alcohol:




```
##
##  Pearson's product-moment correlation
##
## data:  residual.sugar and alcohol
## t = 1.6829, df = 1597, p-value = 0.09258
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.006960058  0.090909069
## sample estimates:
##       cor
## 0.04207544
```

Here I expected a stronger correlation between the alcohol content and the residual sugar, since the alcohol comes from the fermentation of the sugars.

Maybe some of the alcohol in wines and yeast in the sugar have different metabolic behaviors which do not allow to establish a strong linear relationship between sugar fermentation and alcohol production.

Bivariate Analysis

The higher quality wine has stronger relationship with the fixed acidity, citric acid, sulphates and alcohol content. For the free and total sulfur dioxide we have seen in the plots that the medium quality levels (5 and 6) have both higher content than the low and higher quality levels.

Alcohol content has a high correlation with red wine quality. Other important attributes correlated with

red wine quality include sulphates, citric acid and fixed acidity.

We see that volatile acids are negatively correlated with red wine quality. Total sulfur dioxide, density and chlorides are also negatively correlated with quality.

As expected the pH increases with the lower amount of acids. Fixed acidity accounts for most acids present in the wine. The only acid concentration that shows some considerable correlation with pH is the fixed acidity.

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 Fixed acidity tend to have a strong correlation with pH and density. Fixed acidity decreases the pH and increases the density. Alcohol is also strongly correlated with density and quality. While alcohol decreases the density, it also increases the quality.

Did you observe any interesting relationships between the other features

(not the main feature(s) of interest)?

An interesting relationshipis observed between the density and the alcohol and sugar content.

I observed the relation between the pH and acidity level, which is the expected one.

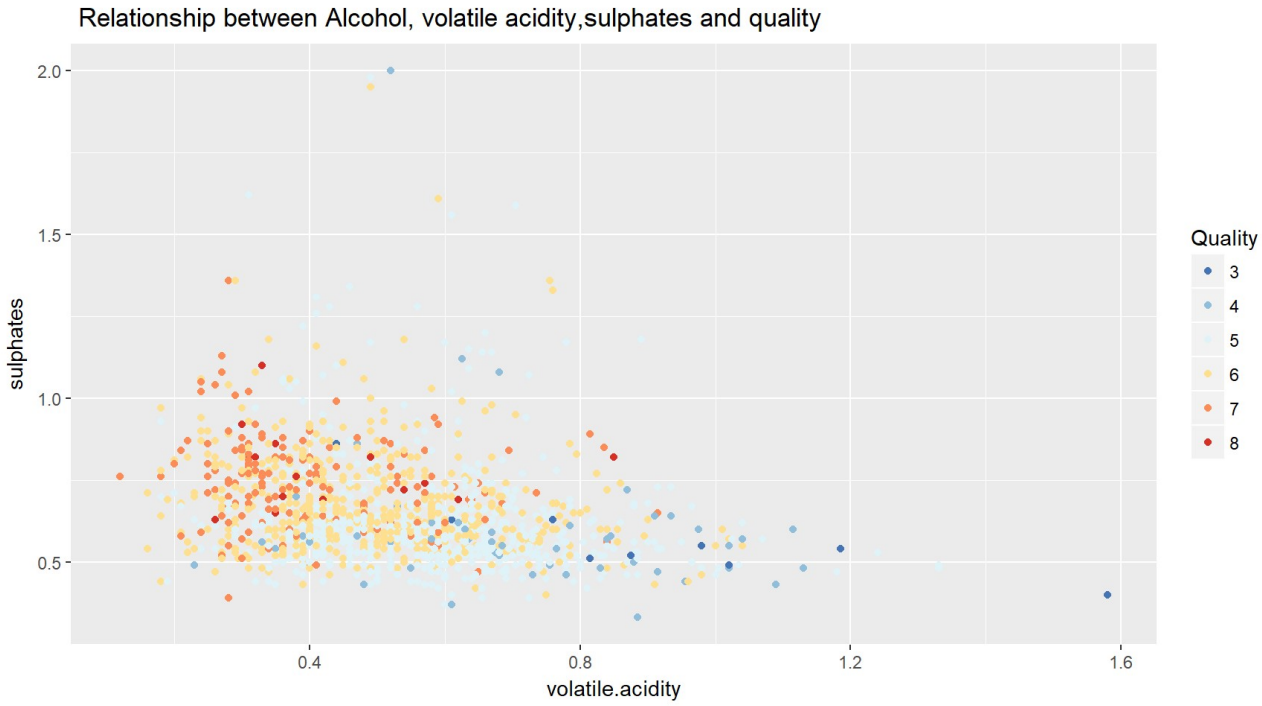
I was surprised by not finding a stronger relation between the residual sugar and alcohol level, since the alcohol comes from the fermentation of sugars.

What was the strongest relationship you found?

The correlation coefficients show that the variable with the strongest relationship with quality is the alcohol content.

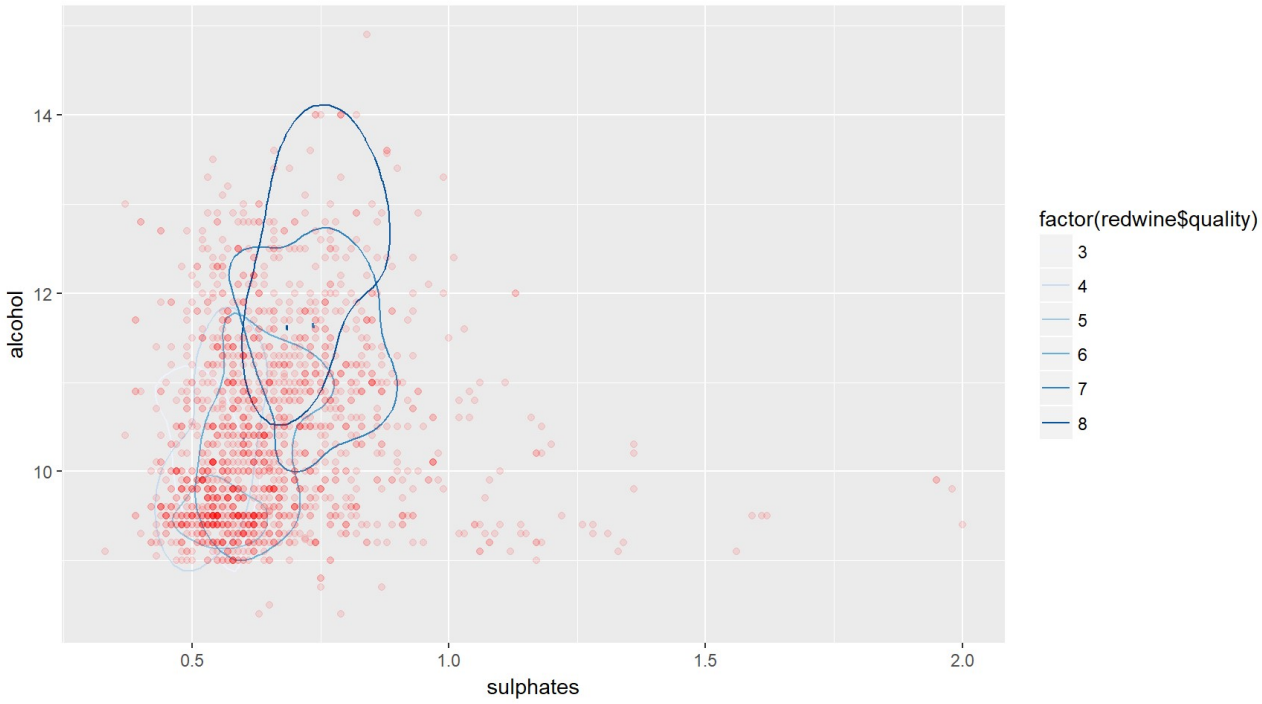
Multivariate Plots Section

Alcohol, volatile acidity,sulphates and quality



It looks like the higher quality red wines tend to be concentrated in the top left of the plot. This tends to be where the higher alcohol content (larger dots) are concentrated as well.

Let's try summarizing quality using a contour plot of alcohol and sulphate content:

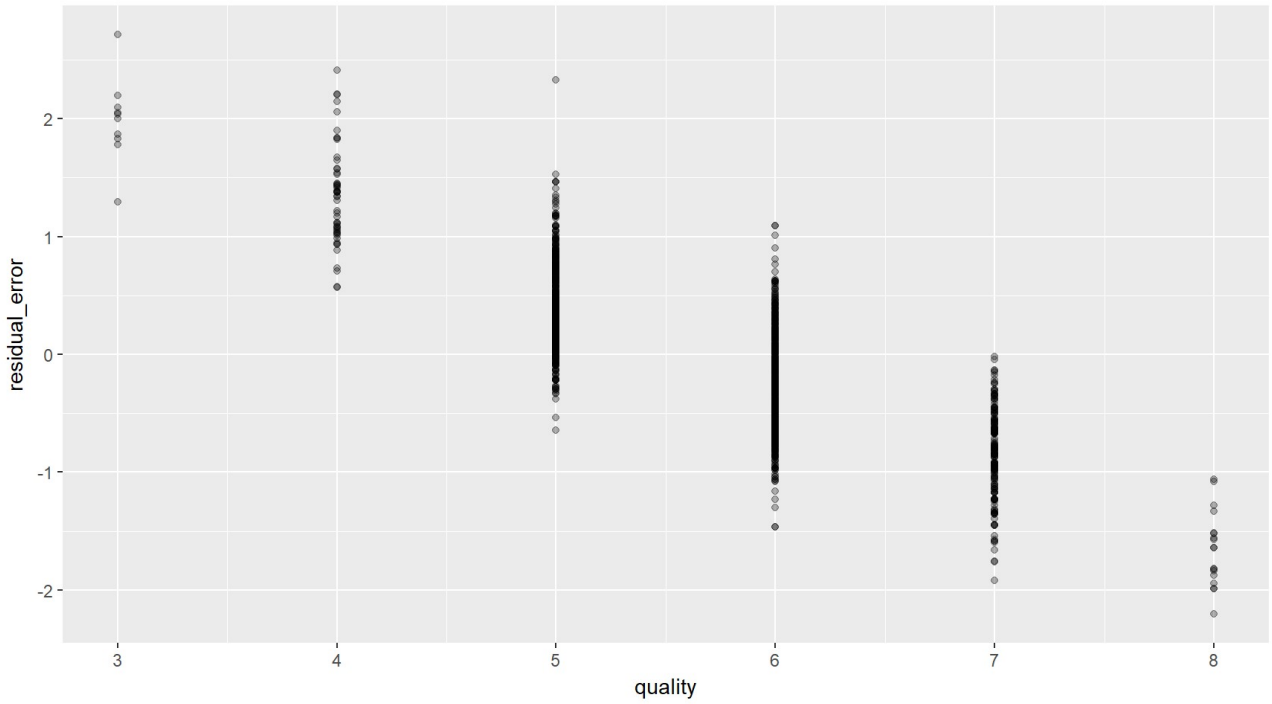


This shows that higher quality red wines are generally located near the upper right of the scatter plot (darker contour lines) wheras lower quality red wines are generally located in the bottom right.

Linear Model

Below are the data values related to a linear model created from four major variables: alcohol, sulphates, citric acid, and volatile acidity. These were all compared to quality and the below graph displays the average residual, or error, of the predictions for each quality.

```
##
## Calls:
## m1: lm(formula = as.numeric(quality) ~ alcohol, data = redwine)
## m2: lm(formula = as.numeric(quality) ~ alcohol + sulphates, data = redwine)
## m3: lm(formula = as.numeric(quality) ~ alcohol + sulphates + citric.acid,
##     data = redwine)
## m4: lm(formula = as.numeric(quality) ~ alcohol + sulphates + citric.acid +
##     volatile.acidity, data = redwine)
##
## ================================================================
##                         m1         m2         m3         m4
## ----------------------------------------------------------------
##   (Intercept)       1.875***   1.375***   1.434***   2.646***
##                    (0.175)    (0.177)    (0.176)    (0.201)
##   alcohol           0.361***   0.346***   0.338***   0.309***
##                    (0.017)    (0.016)    (0.016)    (0.016)
##   sulphates                    0.994***   0.814***   0.696***
##                               (0.102)    (0.107)    (0.103)
##   citric.acid                              0.513***  -0.079
##                                           (0.093)    (0.104)
##   volatile.acidity                                   -1.265***
##                                                      (0.113)
## ----------------------------------------------------------------
##   R-squared             0.2        0.3        0.3        0.3
##   adj. R-squared        0.2        0.3        0.3        0.3
##   sigma                 0.7        0.7        0.7        0.7
##   F                   468.3      295.0      210.5      201.8
##   p                     0.0        0.0        0.0        0.0
##   Log-likelihood    -1721.1    -1675.1    -1660.0    -1599.1
##   Deviance            805.9      760.9      746.6      691.9
##   AIC                3448.1     3358.3     3329.9     3210.2
##   BIC                3464.2     3379.8     3356.8     3242.4
##   N                  1599       1599       1599       1599
## ================================================================
```



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?

The main relationships explored were between the biggest correlators with quality.

We have seen how alcohol and volatile acidity relate with quality. Higher alcohol and lower acidity give in general better quality wines.

Also with sulphates we see the same trend of better quality when both the alcohol and sulphates become higher.

OPTIONAL: Did you create any models with your dataset? Discuss the strengths

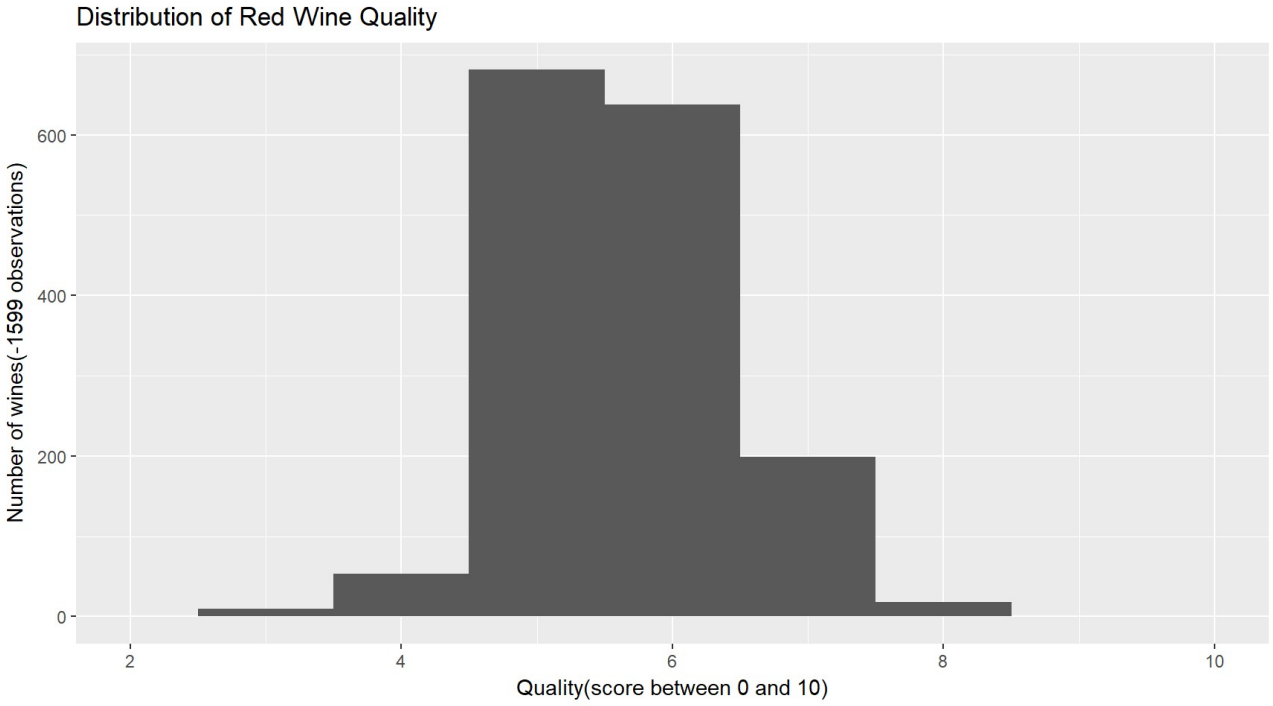
and limitations of your model.

I created a linear model for predicting quality. The R-squared value for the model was 0.2, which was a very low one. It indicates that a linear model probably is not the best fit for this dataset. Alcohol, volatile acidity and sulphates were the most important prediction variables. Since there is a large correlation between some of the variables, some sort of feature selection would improve the model.

Final Plots and Summary

Plot One

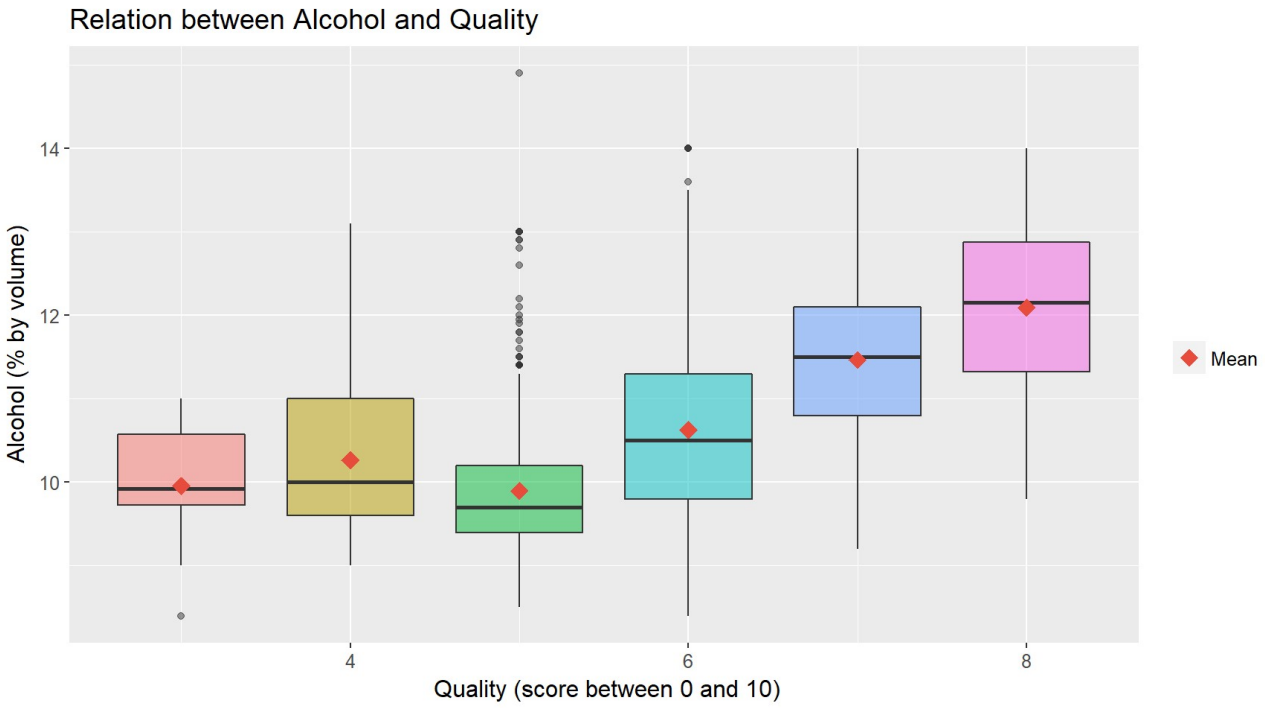
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.636   6.000   8.000
```



Description One

We can say the distribution of quality appears to be normal with many wines at average quality (4-5) and fewer wines at low quality and high quality. There are no wines with a quality worse than 3 and no wines with quality higher than 8.The vast majority of red wines have a quality ranking of 5 and 6.

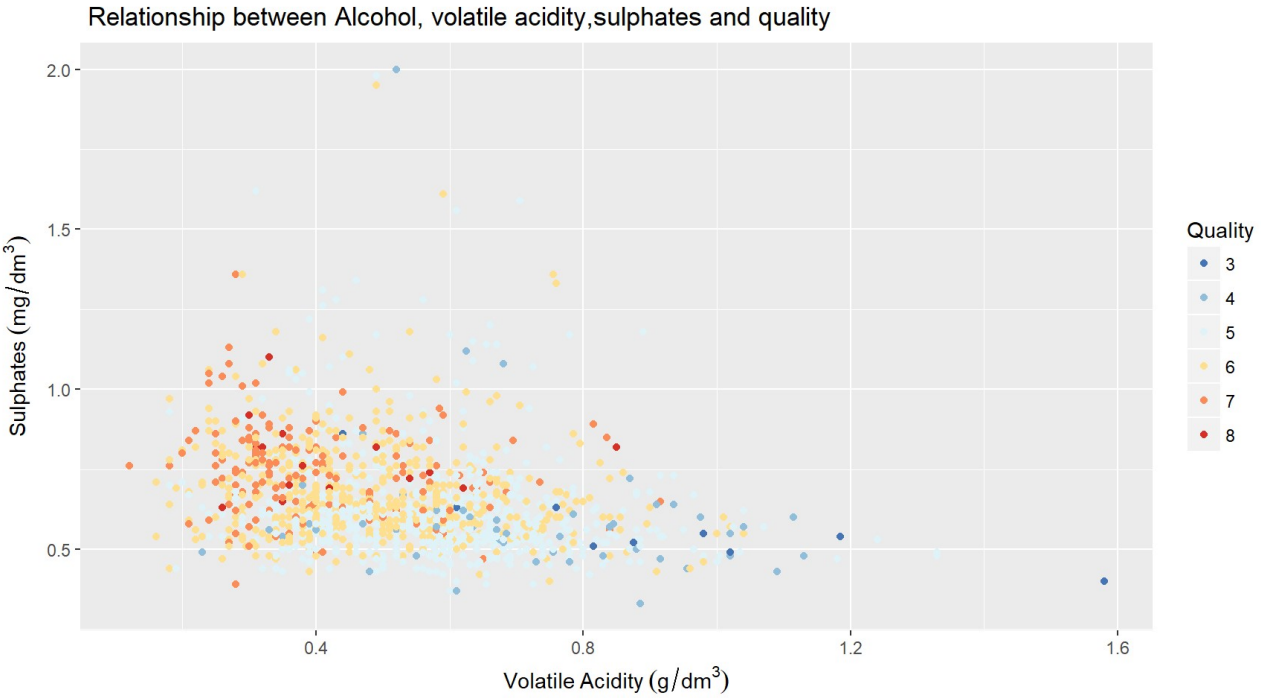
Plot Two



Description Two

A very interesting relation is shown in this chart. There is a positive trend of alcohol level on the quality of wines. For the quality classes 3 to 5, the effect is limited. The quality is probably being steered by another variable, but from the quality rating 5 to 8, we see a sharp increase in the alcohol content. The general trend is that Wines with higher alcohol content are rated higher in quality.

Plot Three



Description Three

This chart shows how quality relates with Alcohol,Volatile acidity and sulphates.It is noticeable that, for a higher alcohol content and lower acidity give in general better quality wines.Also with sulphates we see the same trend of better quality with higher alcohol content.It looks like the higher quality red wines tend to be concentrated in the top left of the plot. This tends to be where the higher alcohol content (larger dots) are concentrated as well.

Reflection

The redwine data set contains 1599 observations across 12 variables.I started by understanding the individual variables in the data set.At first a univariate, then bivariate and finally multivariate examinations are performed allow for a progressive understanding of the dataset and the relations between its features.Most of the univariate plots were right skewed, but density and pH were normally distributed.

When I started this project, I tried to understand the dataset,description and its variable,but the dataset description file already hints at some variables of interest. For example, it tells us that citric acid can add freshness to wines, while acetic acid can add an unpleasant vinegar taste. This shows how important it is to have specific domain knowledge while performing a data analysis.

The challenge I dealt with is in understanding meaningful relations in the multivariate plots.That is when adding a third element a color variation was mostly used ,it becomes harder to grasp trends.One thing that I found suprising was when analysing the acids, fixed acidity and citric acid decreases the pH but volatile acidity didn't, in fact it seemed to increase the pH. A lower pH is supposed to mean that the wine is more acidic.

The whole analysis process was a very valuable experience, in which I got an opportunity to practice plotting various types of charts, handling overplotting and choosing the best chart type to convey the intended message.

A linear model for predicting quality was built, but it performed poorly, indicating that the dataset did not behave very much linearly. In the future, a diferent set of quality prediction models could be applied, and an evaluation of the best fit could be performed.

References:

1.http://jamesmarquezportfolio.com/correlation_matrices_in_r.html

2.http://environmentalcomputing.net/plotting-with-ggplot-adding-titles-and-axis-names

3.https://s3.amazonaws.com/content.udacity-data.com/courses/ud651/diamondsExample_2016-05.html

4.https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityInfo.txt