Red Wine Quality Analysis

Bhavya Garg

Thursday, June 23, 2016

Analysis

Wine industry is a lucrative industry which is growing as social drining is on rise. There are many factors that make the taste and quality of wine unique.

In this project, I try to understand this dataset better and also try to find out if there is a relationship between quality of wine and different properties of it.

The scope of this analysis is to understand relationship of various parameters which impact the quality of Red Wine

Structure of the dataset

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

Description of attributes:

1) fixed acidity: most acids involved with wine or fixed or nonvolatile.

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

3)citric acid: found in small quantities, citric acid can add 'freshness' and flavor to wines

4)residual sugar: the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet

5)chlorides: the amount of salt in the wine

6)free sulfur dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine

7)total sulfur dioxide: amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine

8)density: the density of water is close to that of water depending on the percent alcohol and sugar content

9)pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale

10)sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant

11)alcohol: the percent alcohol content of the wine

12)quality (score between 0 and 10)

Summary of the Data Set

```
##        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
```

Observations from the Summary :

There are 1599 observations of 13 numeric variables.

X appears to be the unique identifier.

The quality of the samples range from 3 to 8, with a mean of 5.6 and median of 6.

All other variables seem to be continuous quantities.

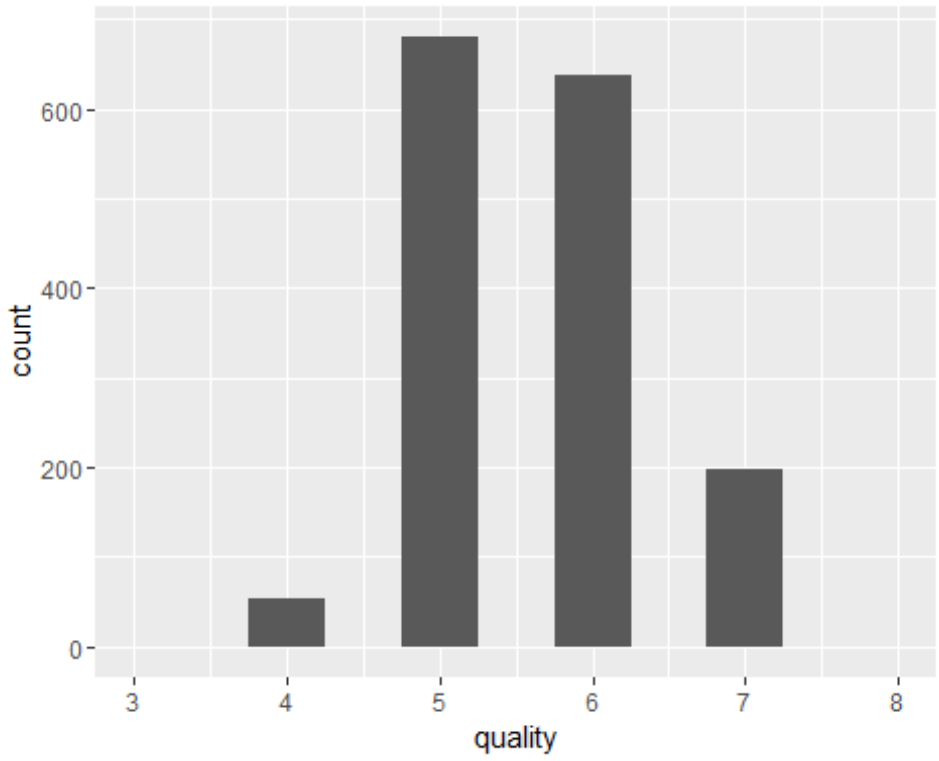
The alcohol content varies from 8.00 to 14.90 for the samples in dataset.

pH value varies from 2.720 to 4.010 with a median being 3.210.

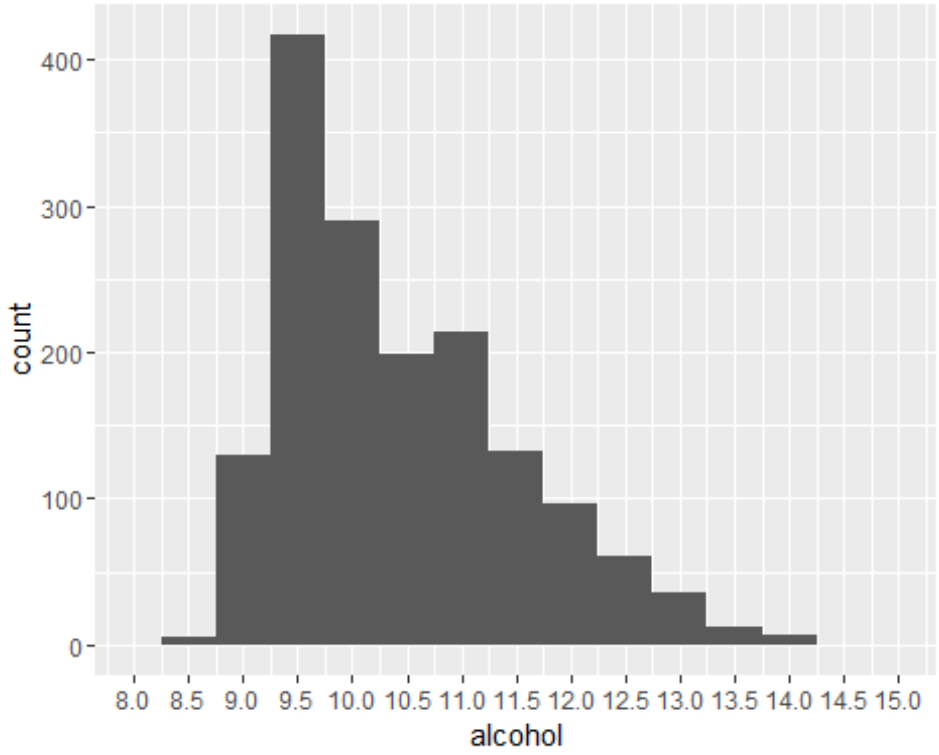
There is a big range for sulfur.dioxide (both Free and Total) across the samples.

Univariate Plots

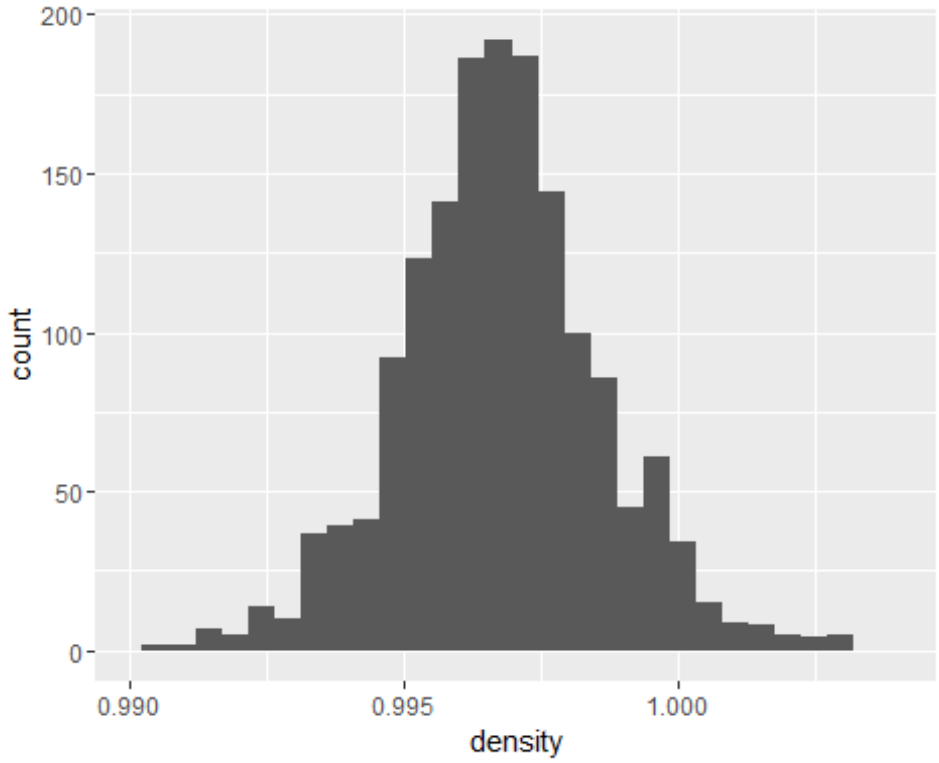
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.636   6.000   8.000

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



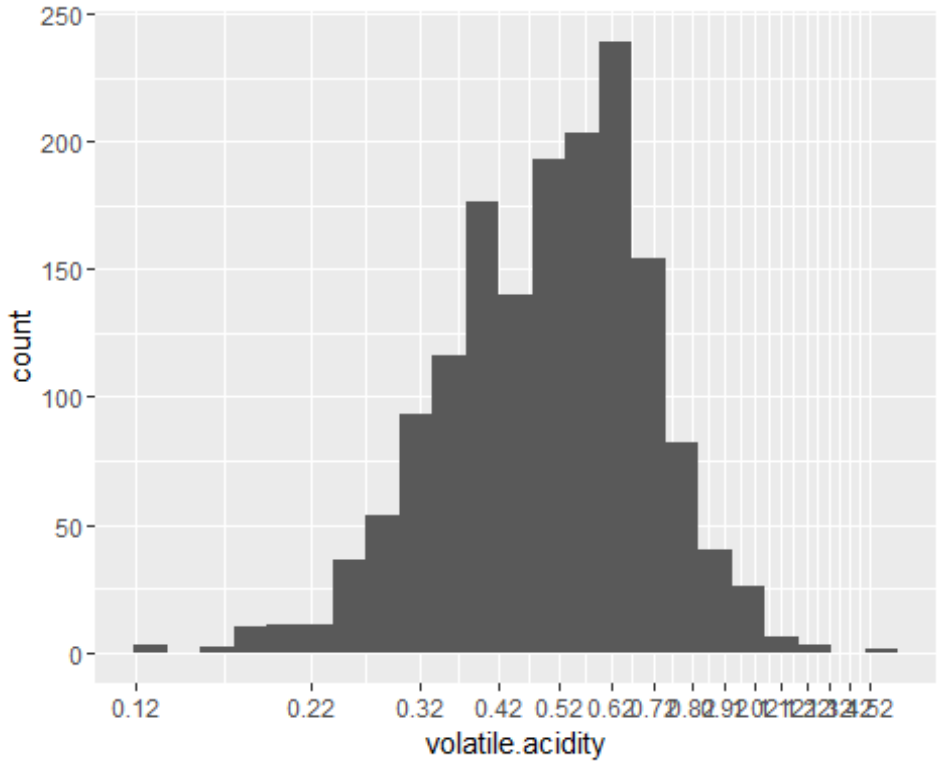

```
##
## average     bad    good
##    1319      63     217

##
##     bad average    good
##      63    1319     217

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



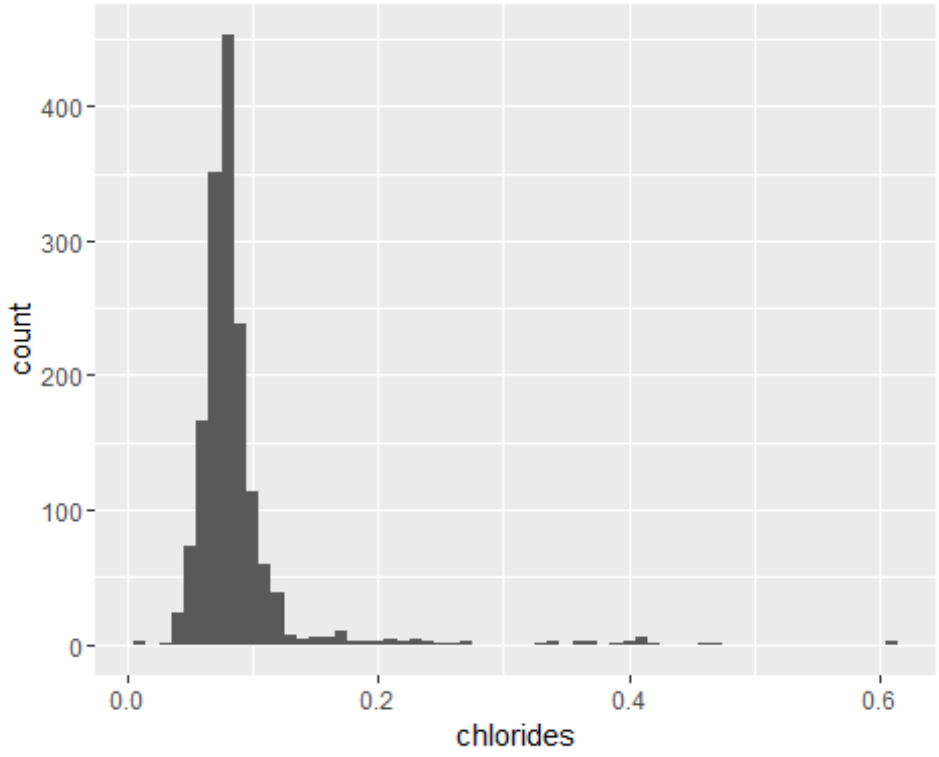

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0040
```

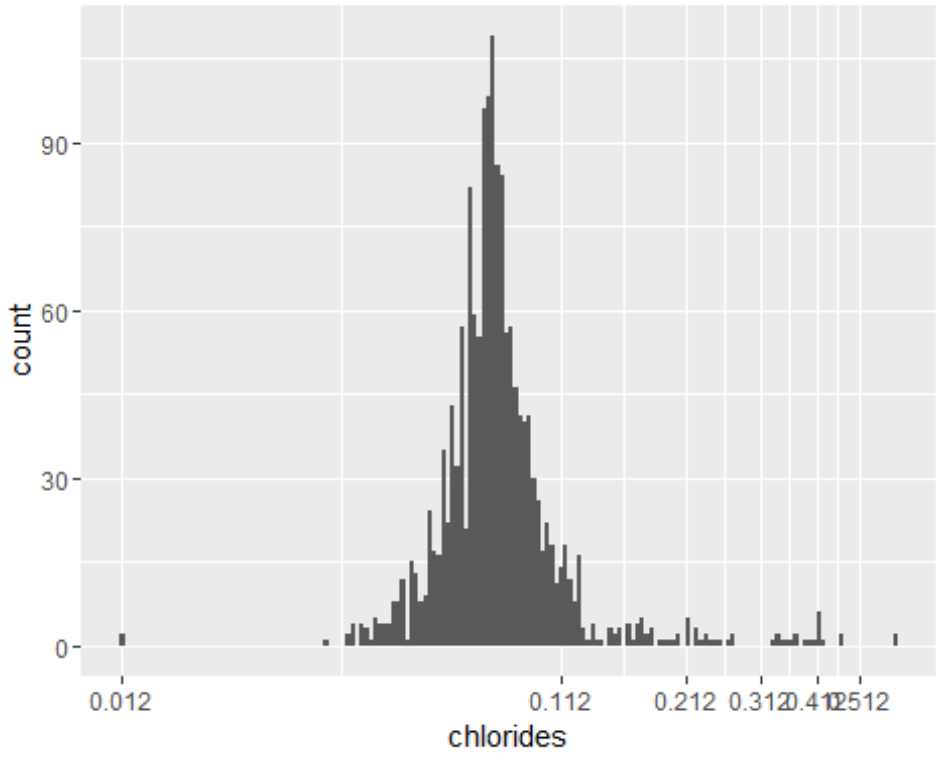


```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```

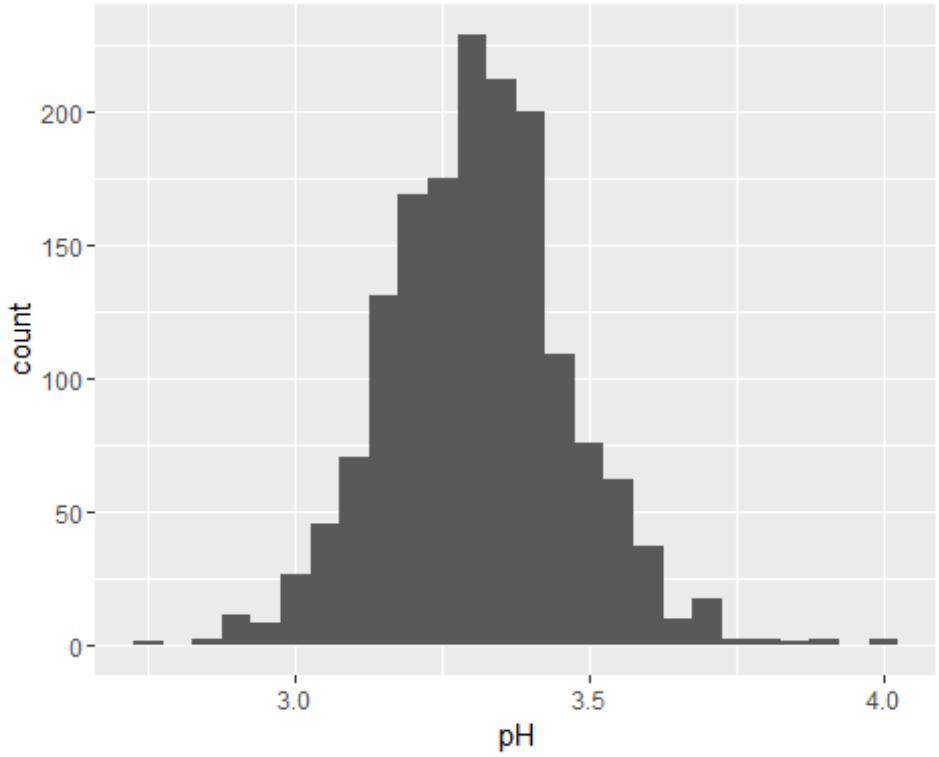



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```

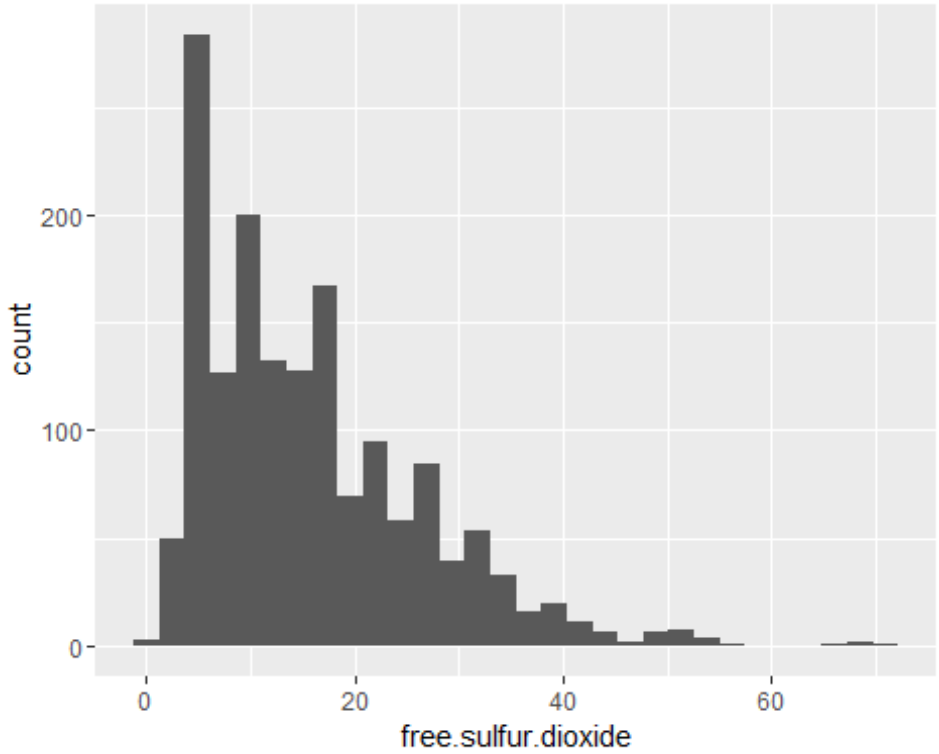




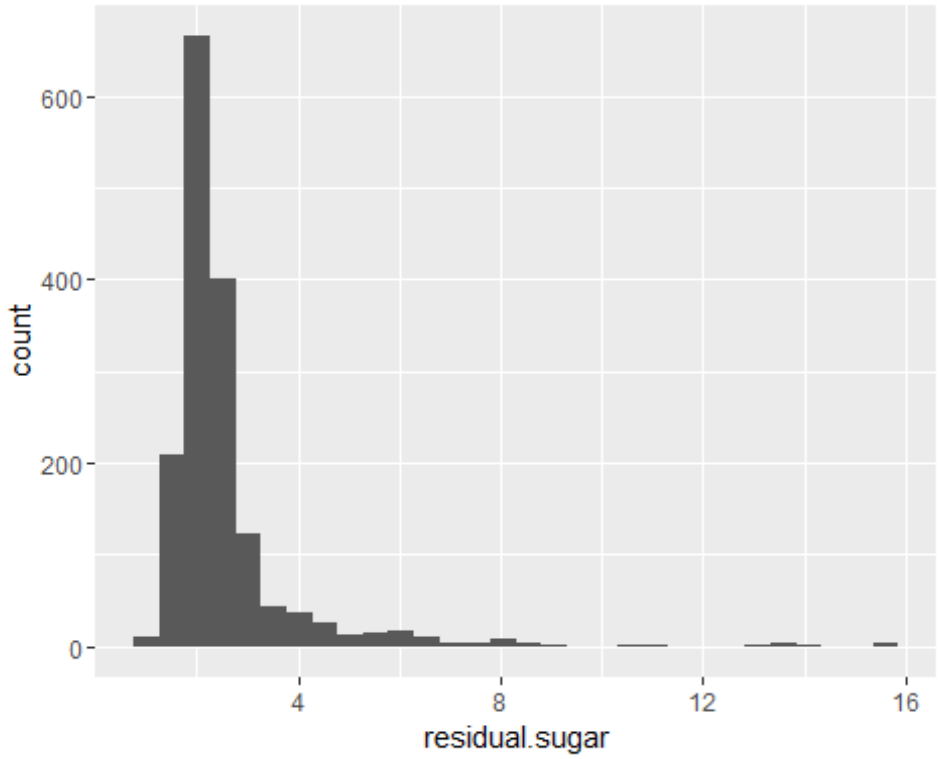

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010
```

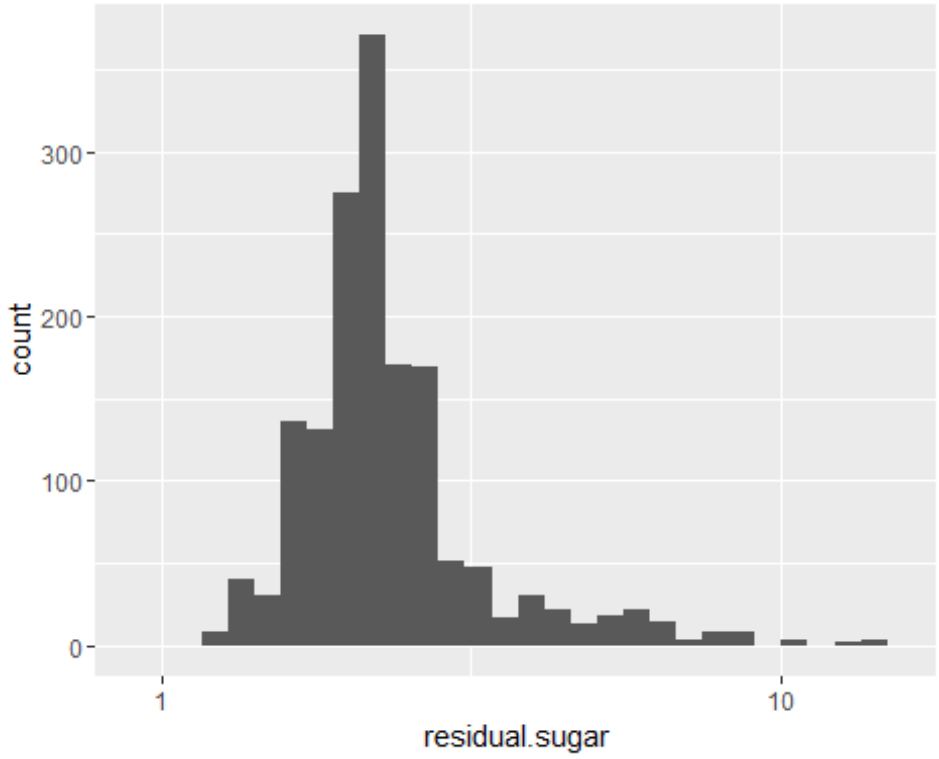



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
```



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





Univariate Analysis:

Some observations from the plot are as below:

The spread for the quality of Red Wine seems to exhibit a normal distribution.Also a large majority of the wines examined received ratings of 5 or 6, and very few received 3, 4, or 8

Alcohol level distribution looks skewed. Most frequently wines have around 10% of alcohol.

The density distribution of Red wine seems is normal.

Volatile acidity has normal distribution. I also suppose that more acetic wines have worse marks because high acidity can lead to unpleasant taste

Chlorides distribution initially is skewed so I used log10 to see the distribution clearer.

The pH value seems to dispaly a normal distribution with major samples exhibiting values between 3.0 and 3.5

The free sulfur dioxide seems to be skewed distribition with the longer tail towards right.

The amount of sugar remained after fermentation is rarely more than 4 g/litre.The distribution is highly skewed towards right.

The main feature of interest in the data is quality. I'd like to determine which features determine the quality of wines.

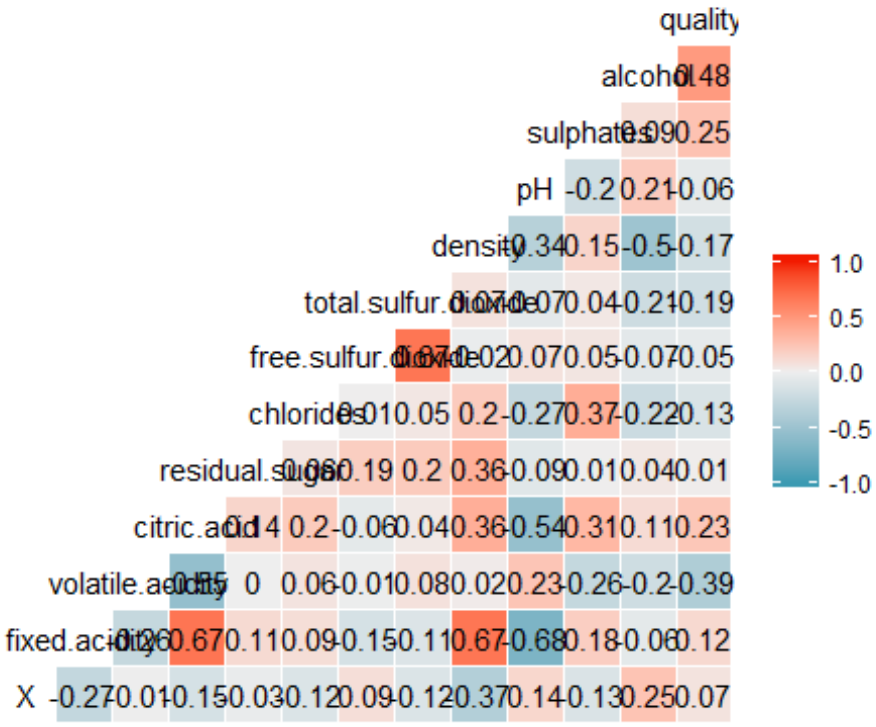
The variables related to acidity (fixed, volatile, citric.acid and pH) might explain some of the variance. I suspect the different acid concentrations might alter the taste of the wine. Also, residual.sugar dictates how sweet a wine is and might also have an influence in taste.

I created an ordered factor, Qualityrating, classifying Quality of the wine as 'bad', 'average', or 'good'.

Bivariate Plots and Analysis :

A correlation table for all variables will help understand the relationships between them

```
## '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 ...
##  $ Qualityrating       : Ord.factor w/ 3 levels "bad"<"average"<..: 2 2 2
2 2 2 2 3 3 2 ...
```



We can see some correlation in pairs like:

alcohol vs. density-----negative correlation(-0.5)

fixed acidity vs. density ------positive correlation(0.67)

residual sugar vs. density------positive correlation(0.36)

chlorides vs. sulphates------positive correlation(0.37)

quality vs. alcohol------positive correlation(0.48)

sulphate vs. citric acid----------positive correlation(0.31)

density vs. pH-----negative correlation(-0.34)

density vs. citric acid----------positive correlation(0.37)

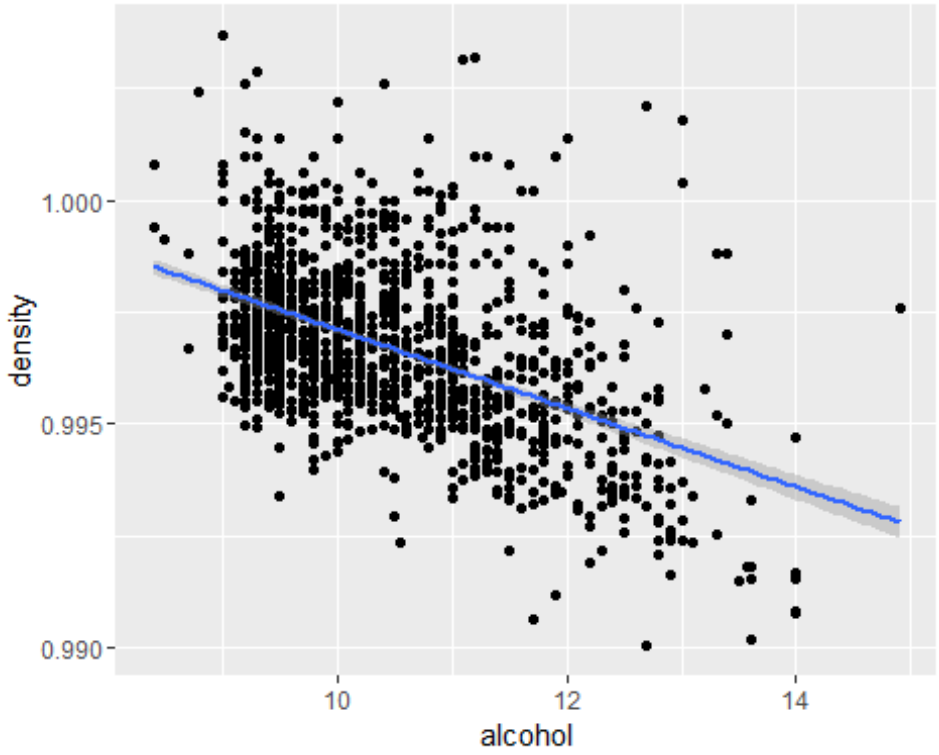
Total sulphur dioxide vs. free sulphur dioxide----positive correlation(0.67)

citric acid vs. fixed acidity ----------positive correlation(0.67)

citric acid vs. volatile acidity--------negative correlation(-0.55)

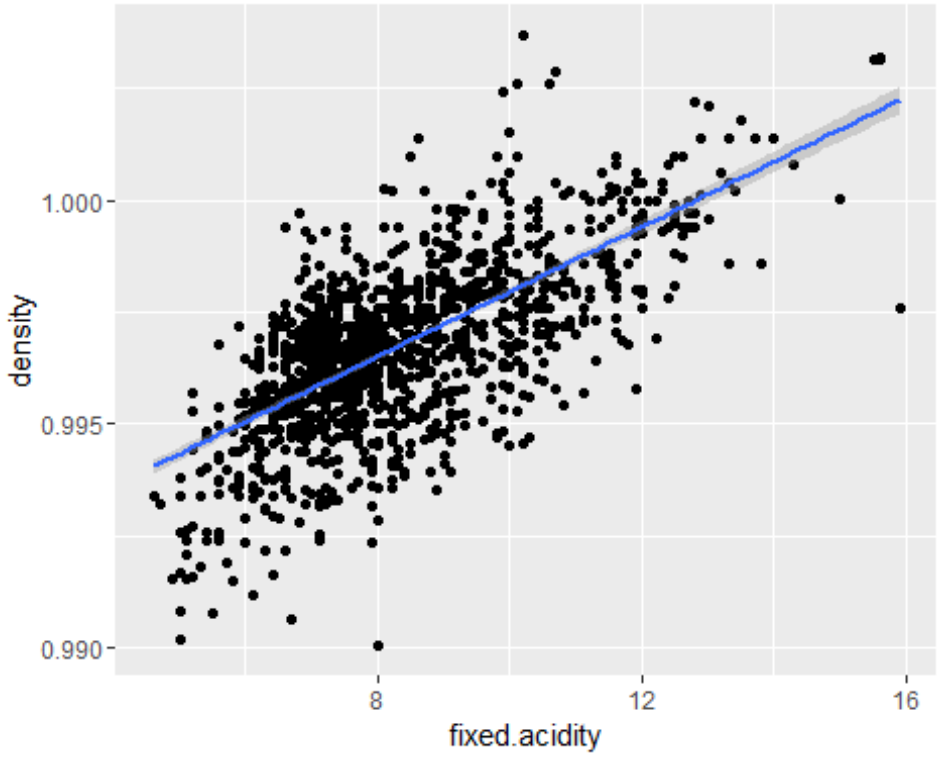
BIVARIATE PLOTS TO COMPARE THE CORRELATIONS BETWEEN THE INPUT VARIABLES

```
##
##  Pearson's product-moment correlation
##
## data:  wine$alcohol and wine$density
## t = -22.8382, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5322547 -0.4583061
## sample estimates:
##       cor
## -0.4961798
```



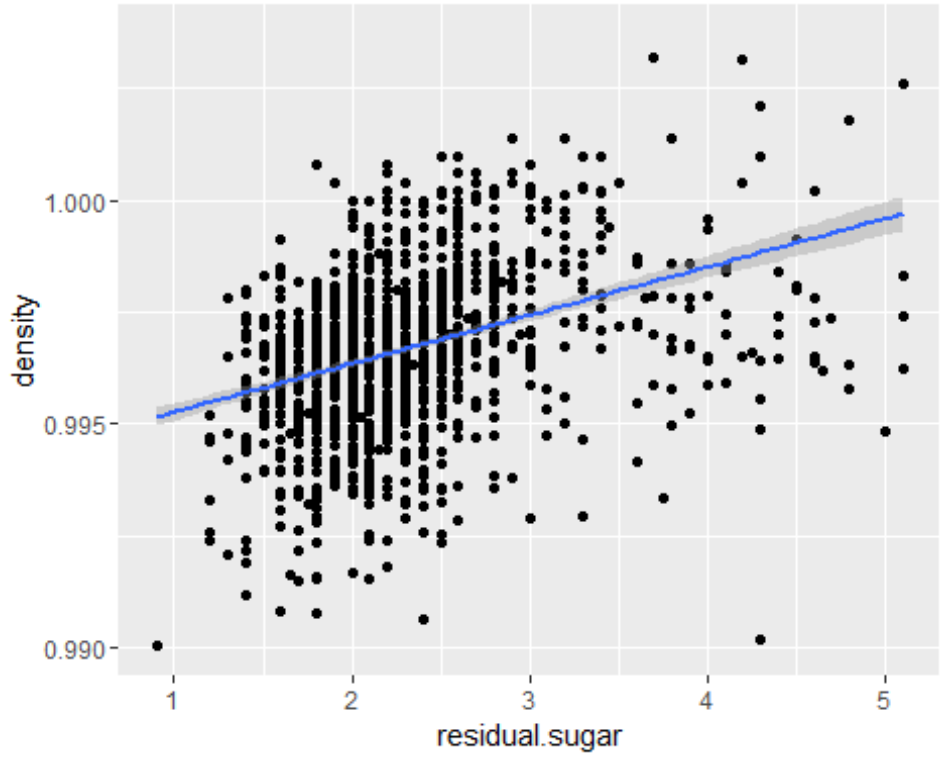
Alcohol has negative correlation with density. This is expected as alcohol is less dense than water

```
##
##  Pearson's product-moment correlation
##
## data:  wine$fixed.acidity and wine$density
## t = 35.8771, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6399847 0.6943302
## sample estimates:
##       cor
## 0.6680473
```



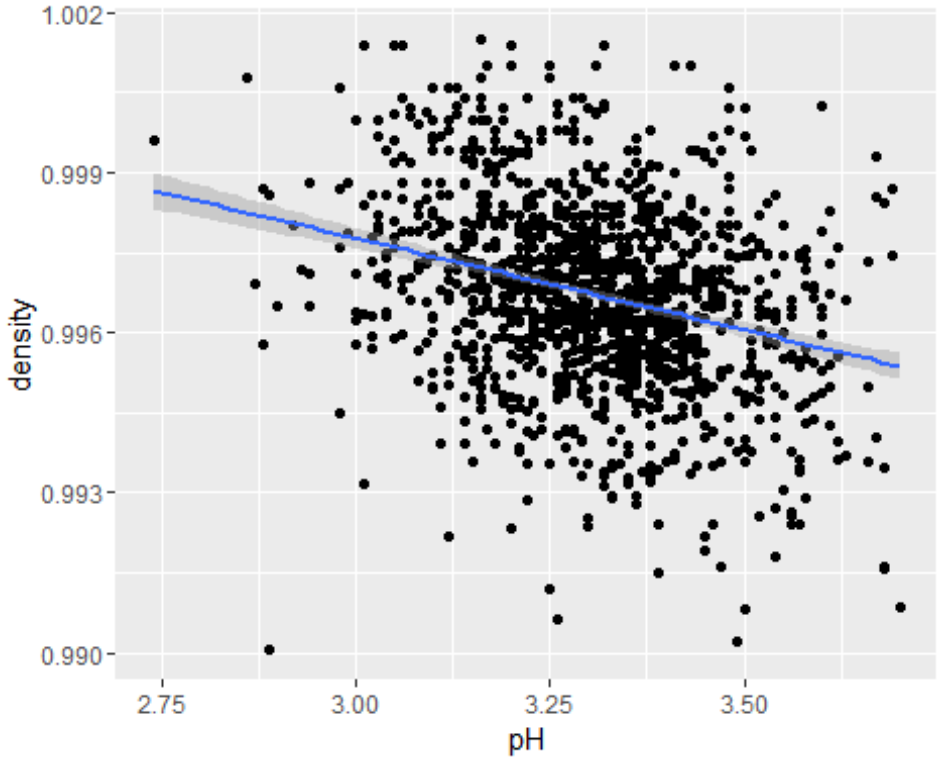
Density has a very strong correlation with fixed.acidity.

```
##
##  Pearson's product-moment correlation
##
## data:  wine$residual.sugar and wine$density
## t = 15.189, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3116908 0.3973835
## sample estimates:
##       cor
## 0.3552834
```



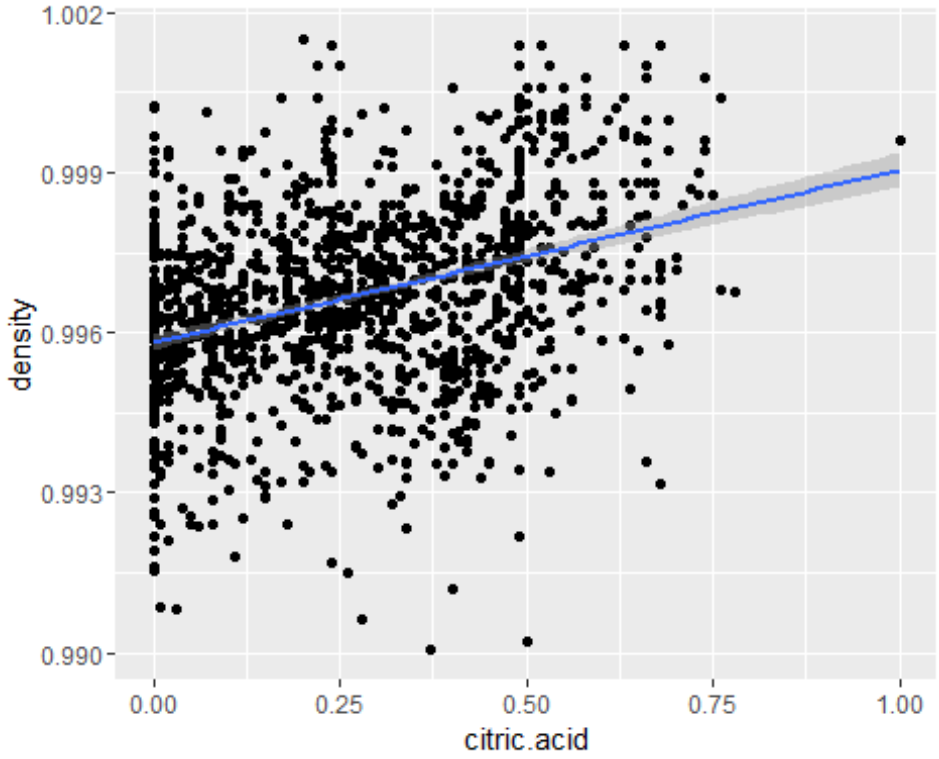
There exists a positive correlation between residual sugar and density

```
##
##  Pearson's product-moment correlation
##
## data:  wine$pH and wine$density
## t = -14.5297, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3842835 -0.2976642
## sample estimates:
##       cor
## -0.3416993
```

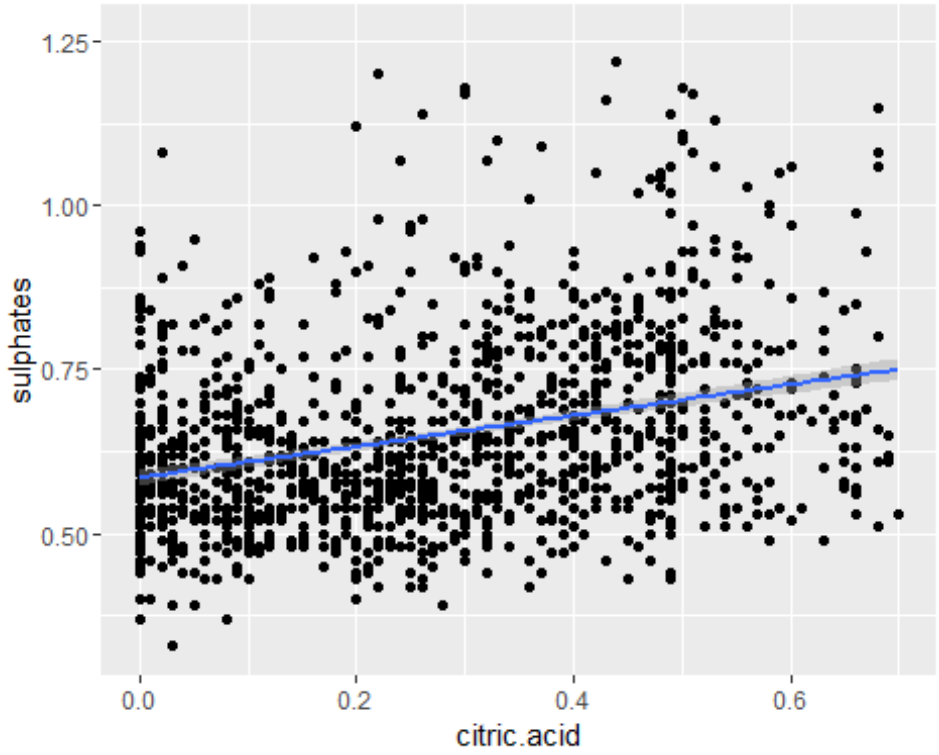


Negative correlation exists between density and pH.

```
##
##  Pearson's product-moment correlation
##
## data:  wine$citric.acid and wine$density
## t = 15.6646, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3216809 0.4066925
## sample estimates:
##       cor
## 0.3649472
```

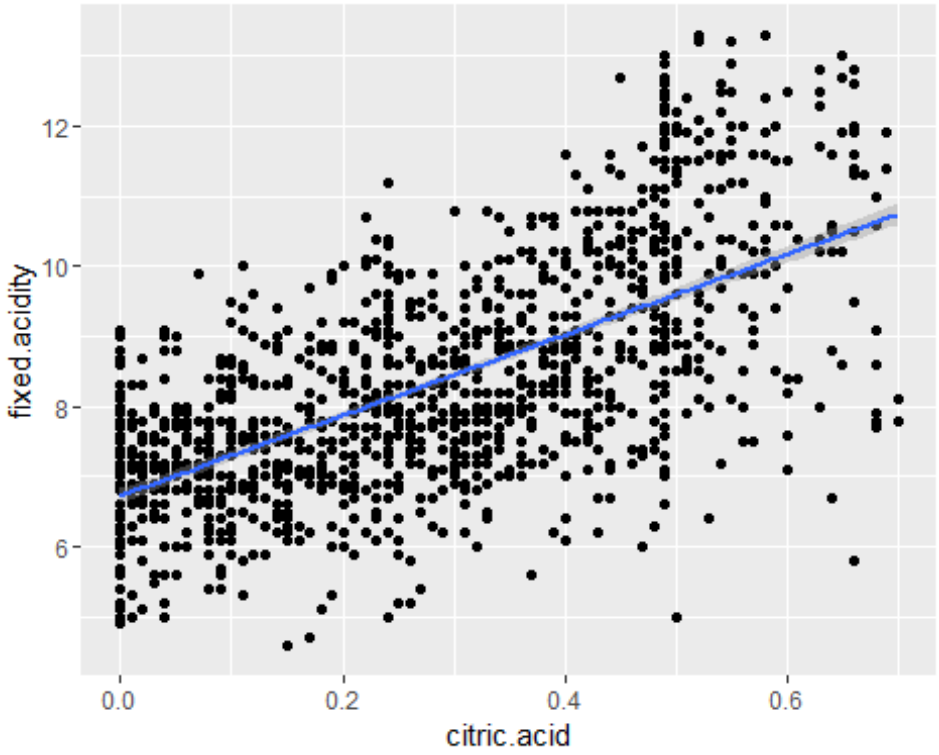



```
##
##  Pearson's product-moment correlation
##
## data:  wine$citric.acid and wine$sulphates
## t = 13.1593, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2678558 0.3563278
## sample estimates:
##     cor
## 0.31277
```



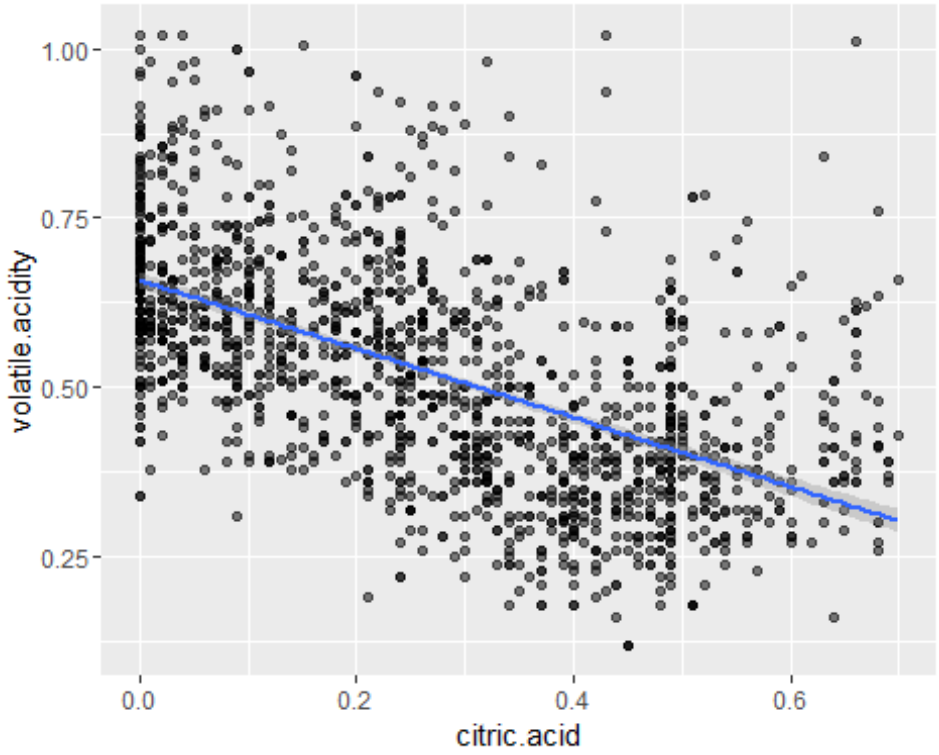
Positive correlation between sulpahte and citric acid.

```
##
##  Pearson's product-moment correlation
##
## data:  wine$citric.acid and wine$fixed.acidity
## t = 36.2341, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6438839 0.6977493
## sample estimates:
##       cor
## 0.6717034
```



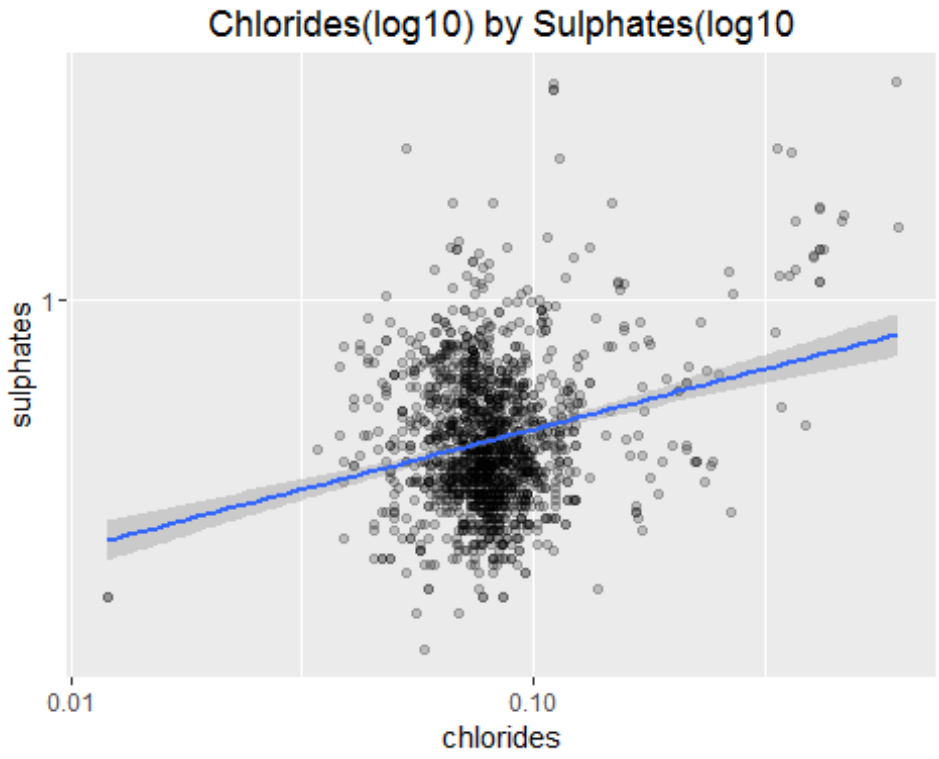
citric acid and fixed acidity are strongly correlated

```
##
##  Pearson's product-moment correlation
##
## data:  wine$citric.acid and wine$volatile.acidity
## t = -26.4891, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5856550 -0.5174902
## sample estimates:
##       cor
## -0.5524957
```

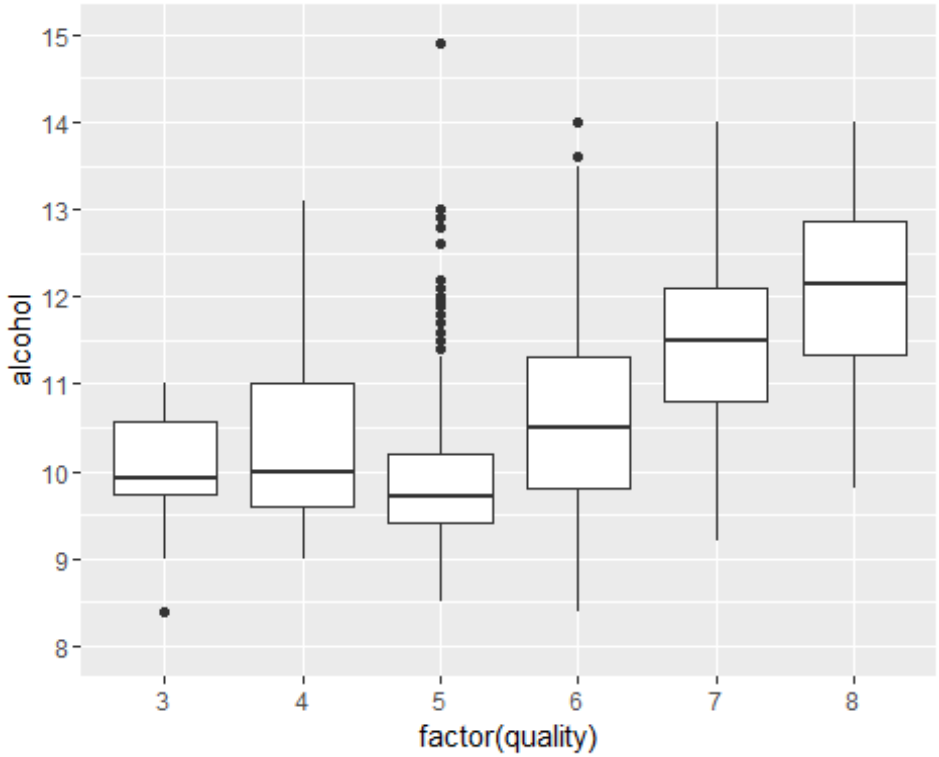


citric acid and volatile acidity are negatively correlated

```
##
##  Pearson's product-moment correlation
##
## data:  wine$chlorides and wine$sulphates
## t = 15.9785, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3282127 0.4127694
## sample estimates:
##       cor
## 0.3712605
```

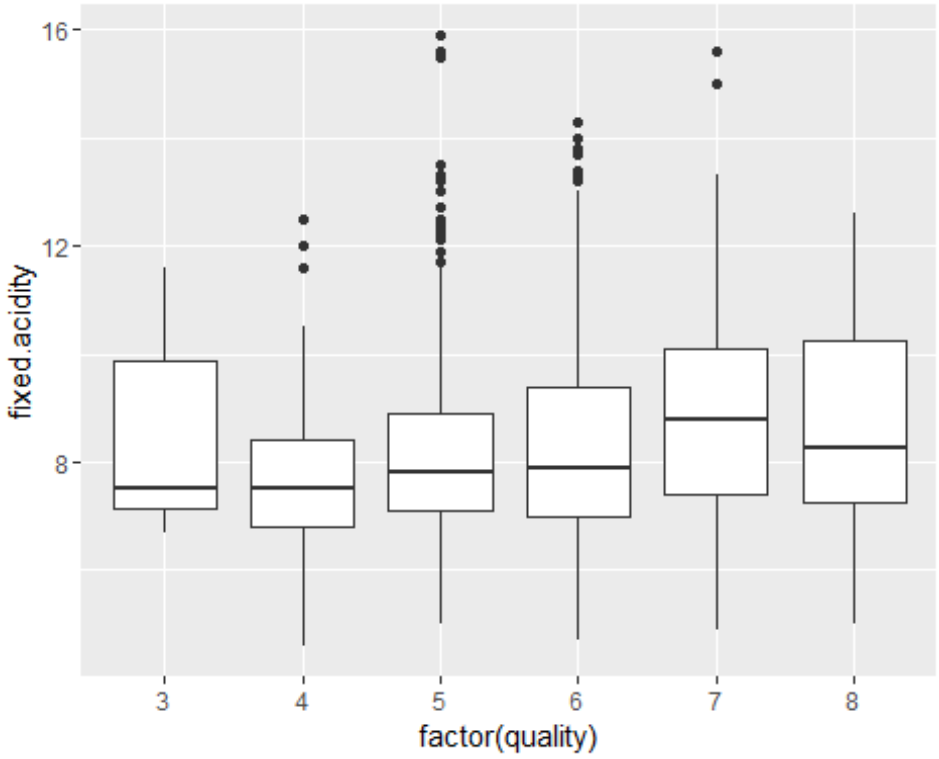


Let's use boxplots to further examine the relationship between some varibles and quality.

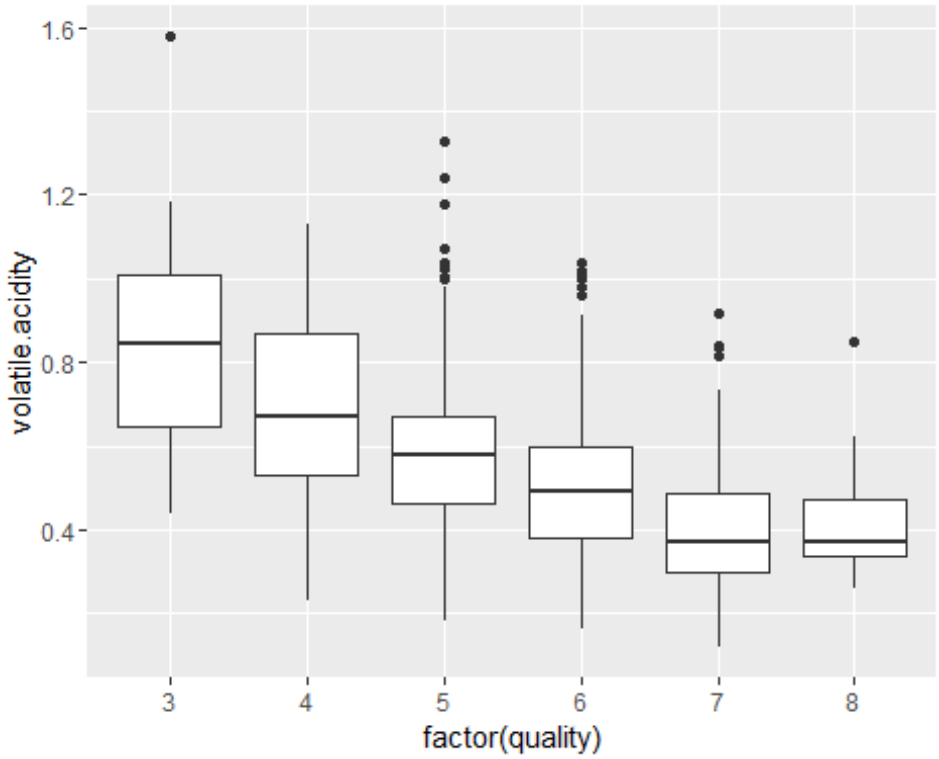


```
##
## Call:
## lm(formula = quality ~ alcohol, data = wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8442 -0.4112 -0.1690  0.5166  2.5888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.87497    0.17471   10.73   <2e-16 ***
## alcohol      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
```

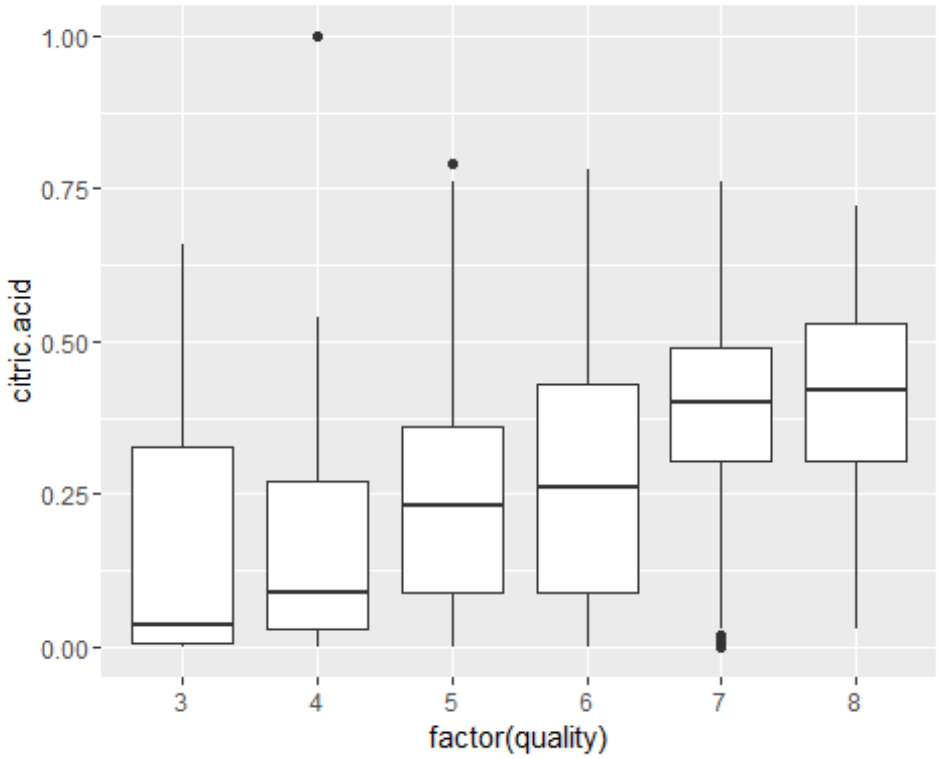
The correlation is clear here. With an increase in alcohol content, we see an increase in the concentration of better graded wines.Based on the R-squared value it seems alcohol alone only explains about 22% of the variance in quality. We're going to need to look at the other variables to generate a better model.



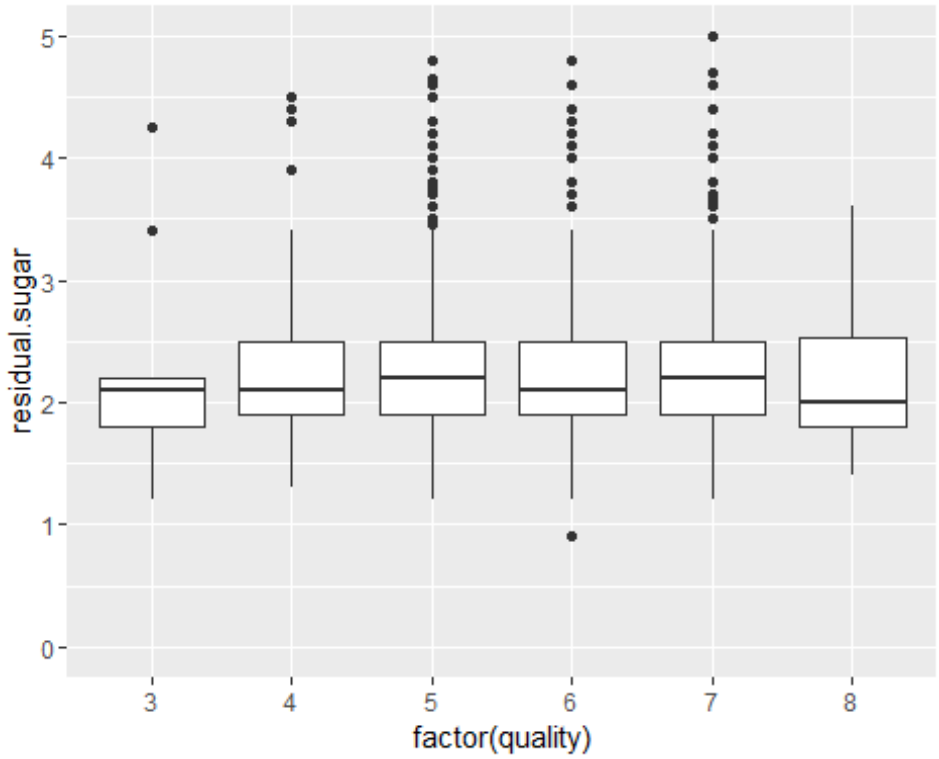
As the boxplot showed, fixed.acidity seems to have little to no effect on quality.



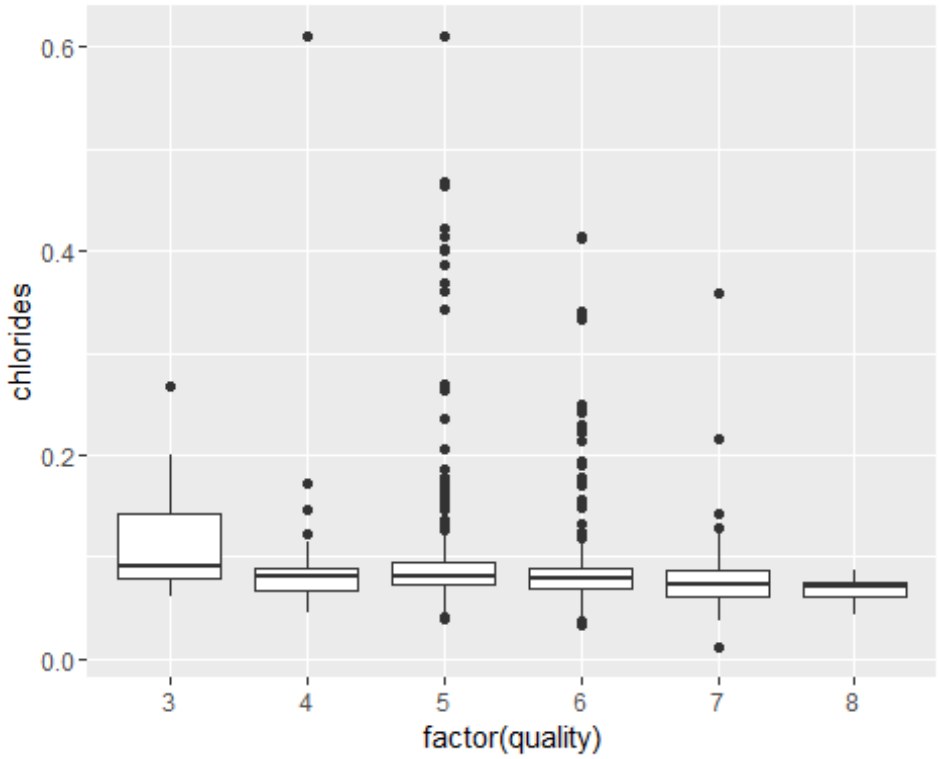
Quality seems to go up when volatile.acidity is low.The higher amount of acetic acid in wine seem to produce more average and poor wines.



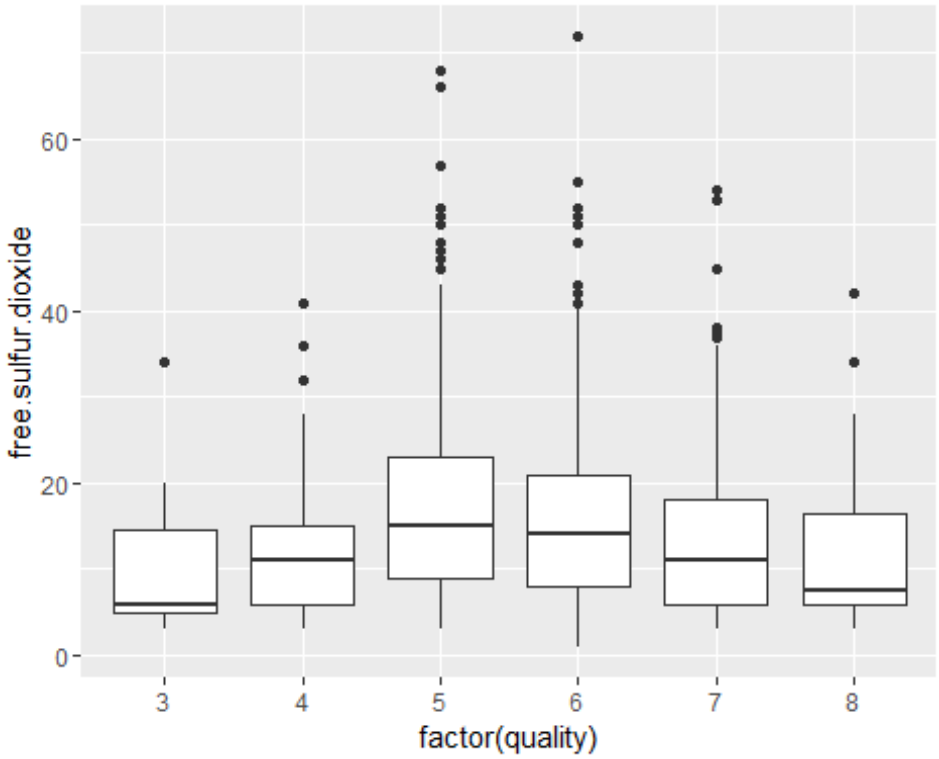
We can see the soft correlation between these two variables. Better wines tend to have higher concentration of citric acid.



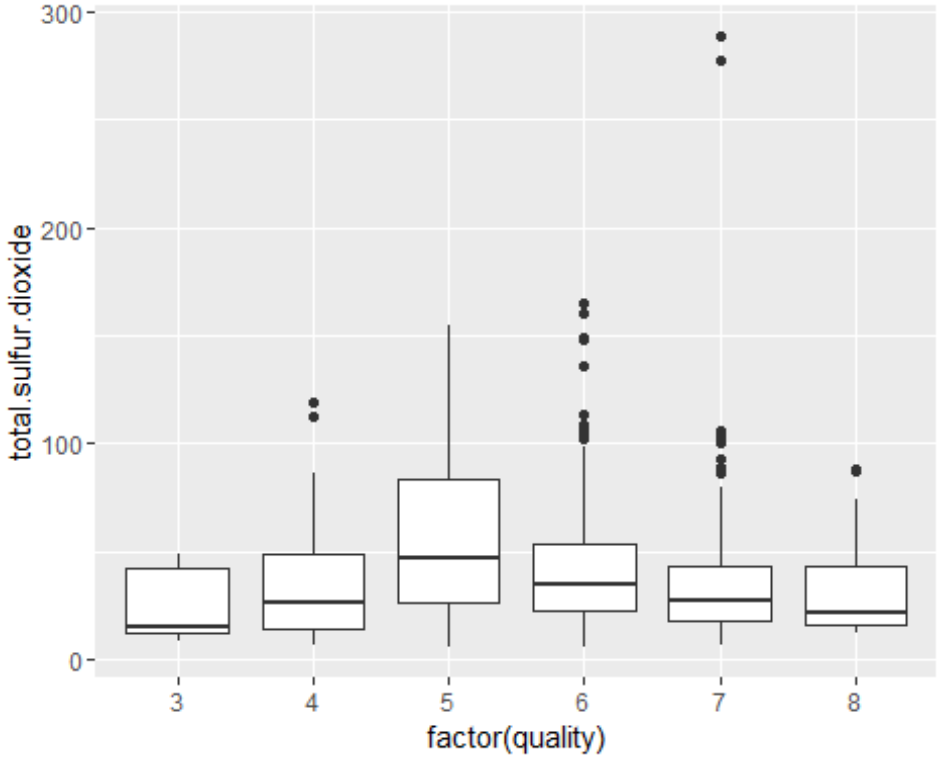
residual.sugar apparently seems to have little to no effect on perceived quality.Also there seems to have so many outliers.



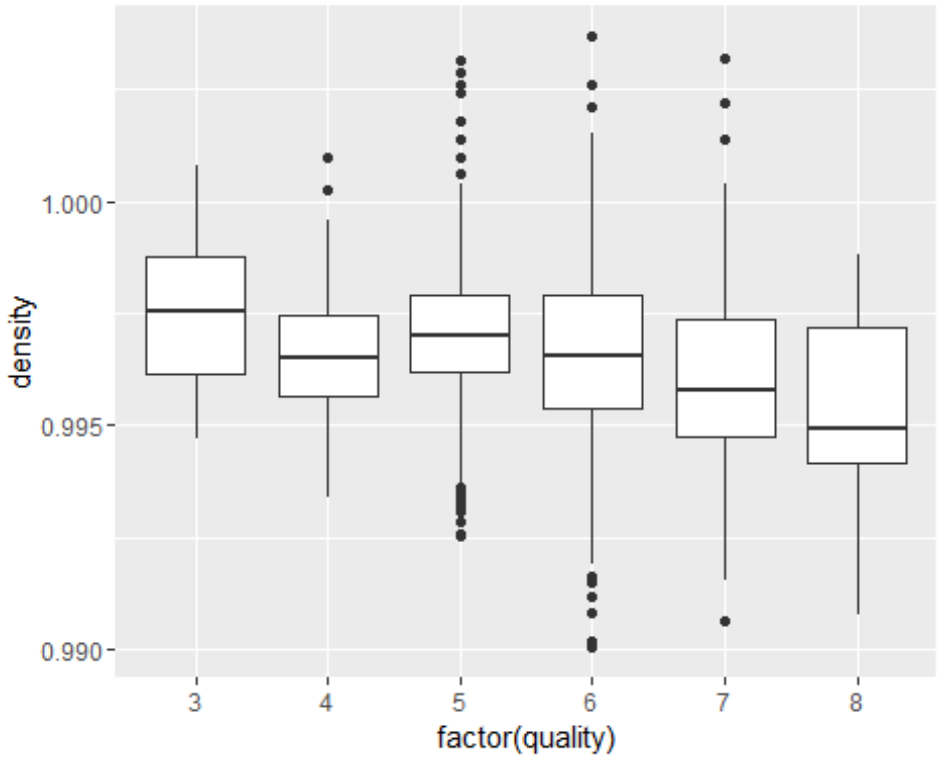
Altough weakly correlated, a lower concentration of chlorides seem to produce better wines.



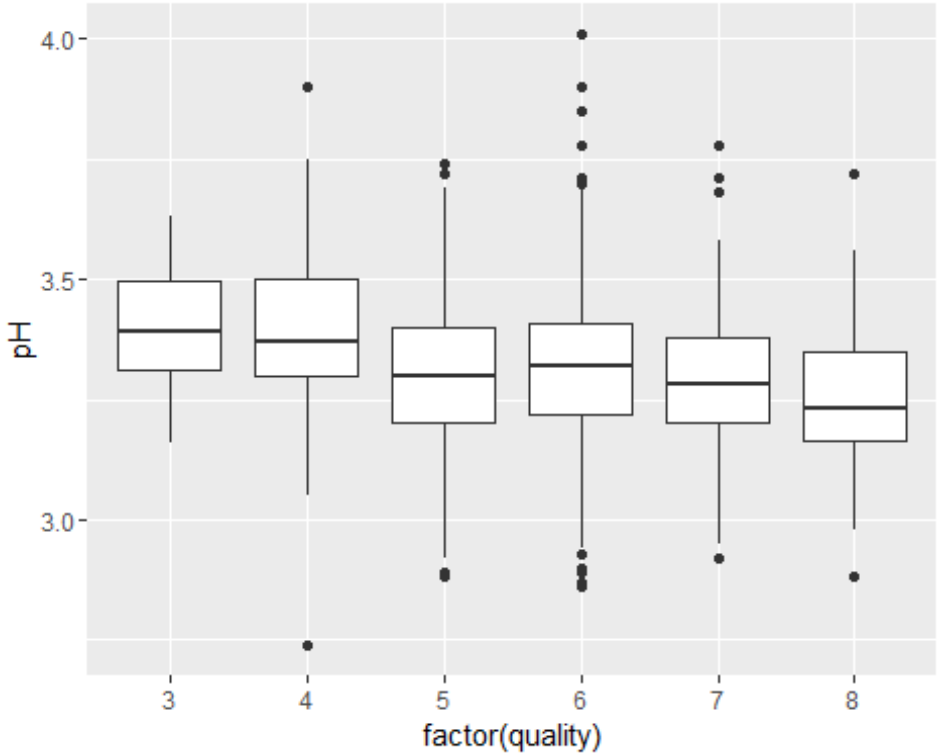
Free sulphur dioxide seems to be an unwanted feature of wine.



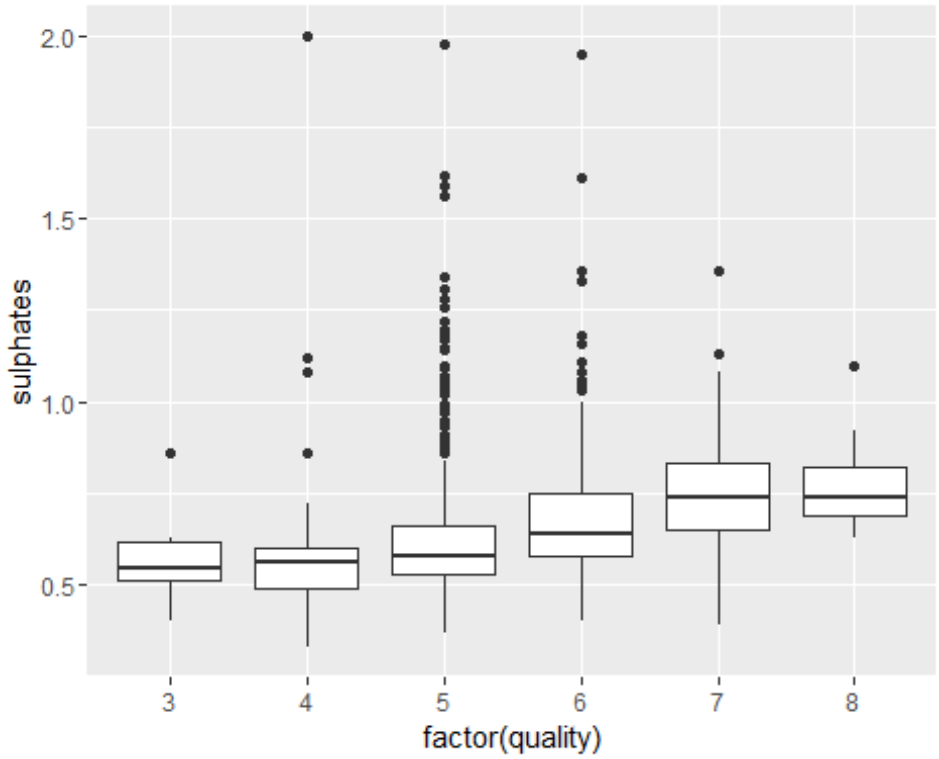
As a superset of free.sulfur.dioxide there is no surprise to find a very similar distribution here.



Better wines tend to have lower densities, but this is probably due to the alcohol concentration. I wonder if density still has an effect if we hold alcohol constant.



Altough there is definitely a trend (better wines being more acid) there are some outliers.



Interesting. Altough there are many outliers in the medium wines, better wines seem to have a higher concentration of sulphates.

Bivariate Analysis:

Fixed.acidity seems to have little to no effect on quality

Quality seems to go up when volatile.acidity goes down. The higher ranges seem to produce more average and poor wines.Better wines tend to have higher concentration of citric acid.

Contrary to what I initially expected residual.sugar apparently seems to have little to no effect on perceived quality.Altough weakly correlated, a lower concentration of chlorides seem to produce better wines.Better wines tend to have lower densities.

In terms of pH it seems better wines are more acid but there were many outliers. Better wines also seem to have a higher concentration of sulphates.

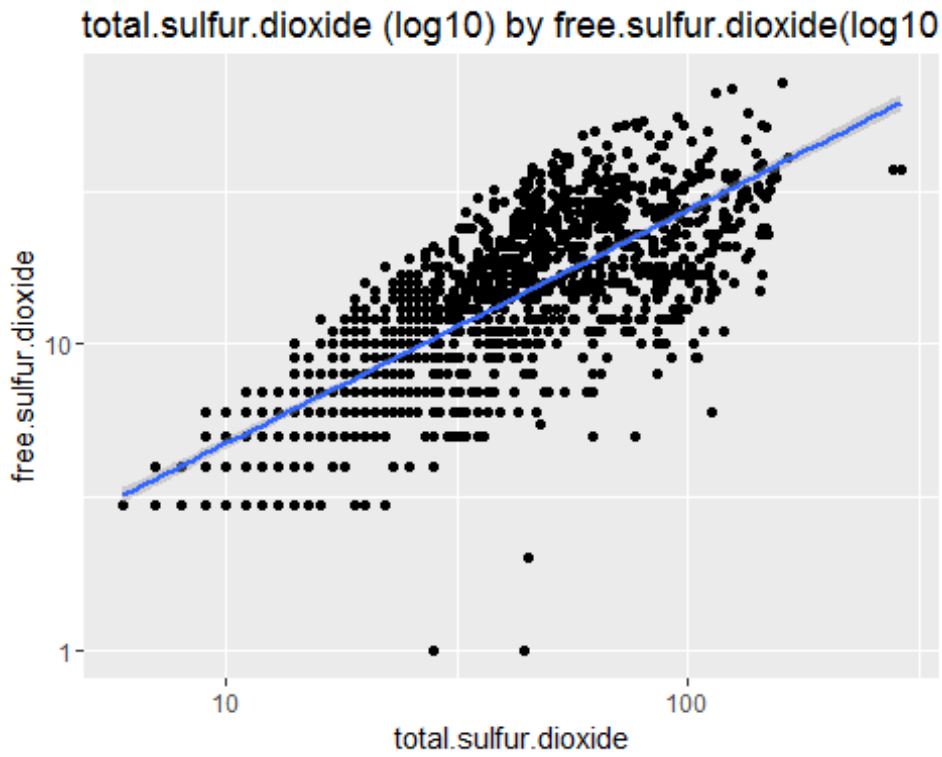
Alcohol graduation has a strong correlation with quality, but like the linear model showed us it cannot explain all the variance alone. We're going to need to look at the other variables to generate a better model.

Other than the main feature of interest i observe interesting relationships between the other below features:

I verified the strong relation between free and total sulfur.dioxide.

Strong relation between fixed acidity and citric acid.

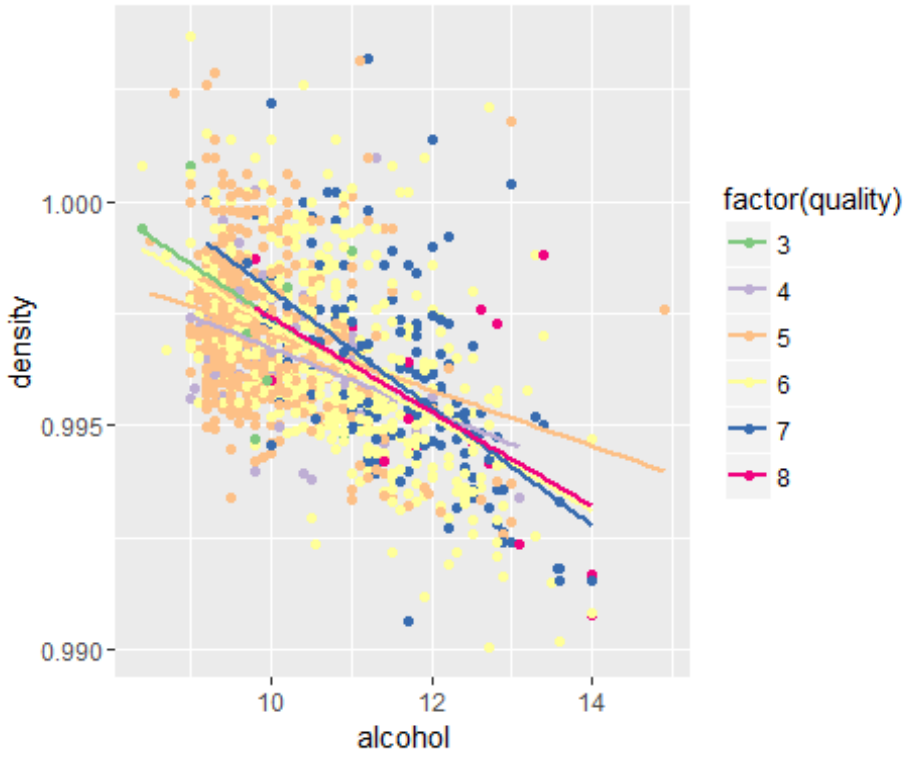
Strong relation between fixed acidity and volatile acidity.



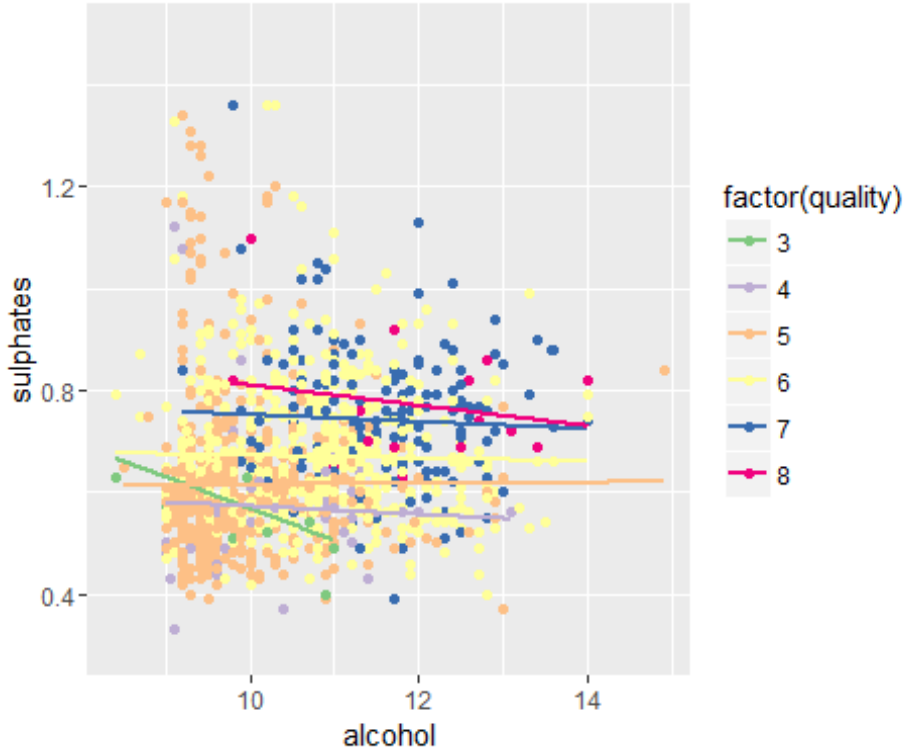

```
##
##  Pearson's product-moment correlation
##
## data:  wine$free.sulfur.dioxide and wine$total.sulfur.dioxide
## t = 35.8402, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6395786 0.6939740
## sample estimates:
##       cor
## 0.6676665
```

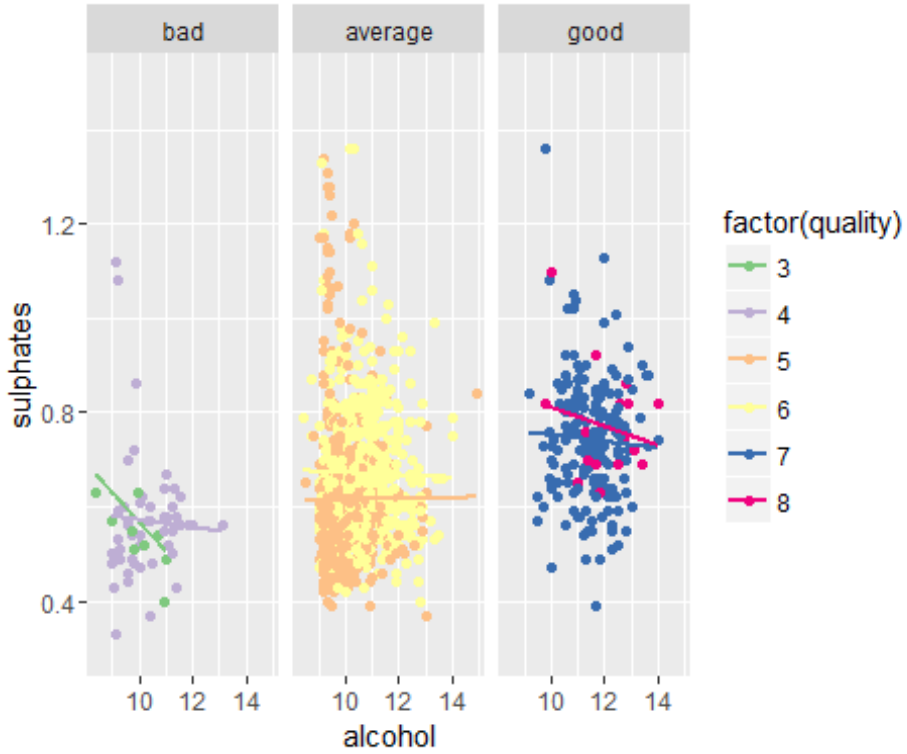
The strongest relationship I found was between the variables total.sulfur.dioxide and free.sulfur.dioxide.

Multivariate Plots:

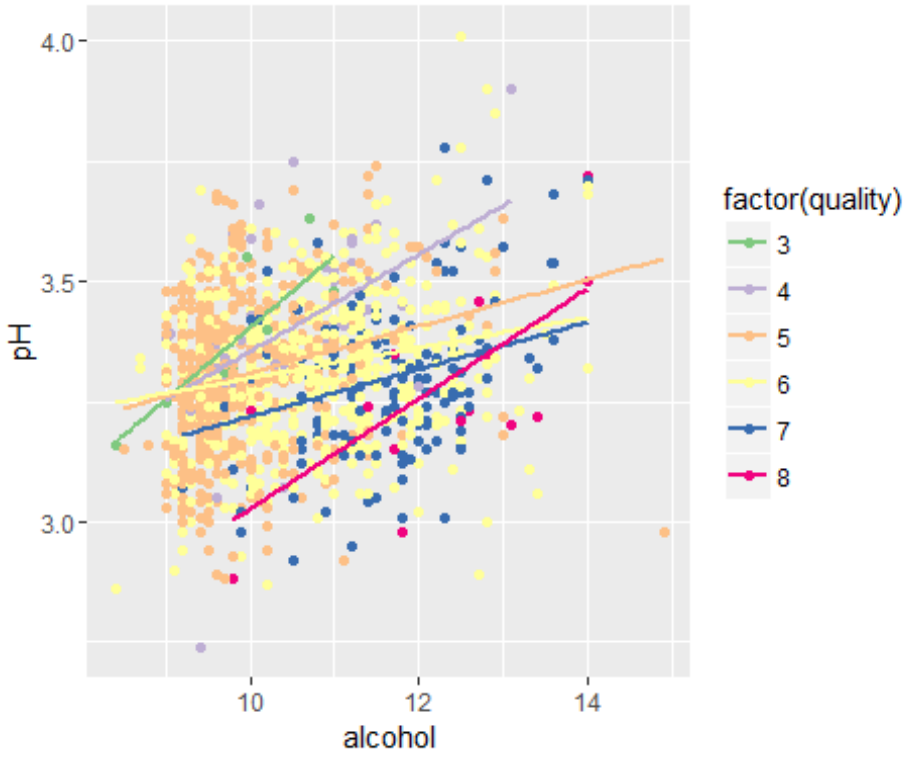


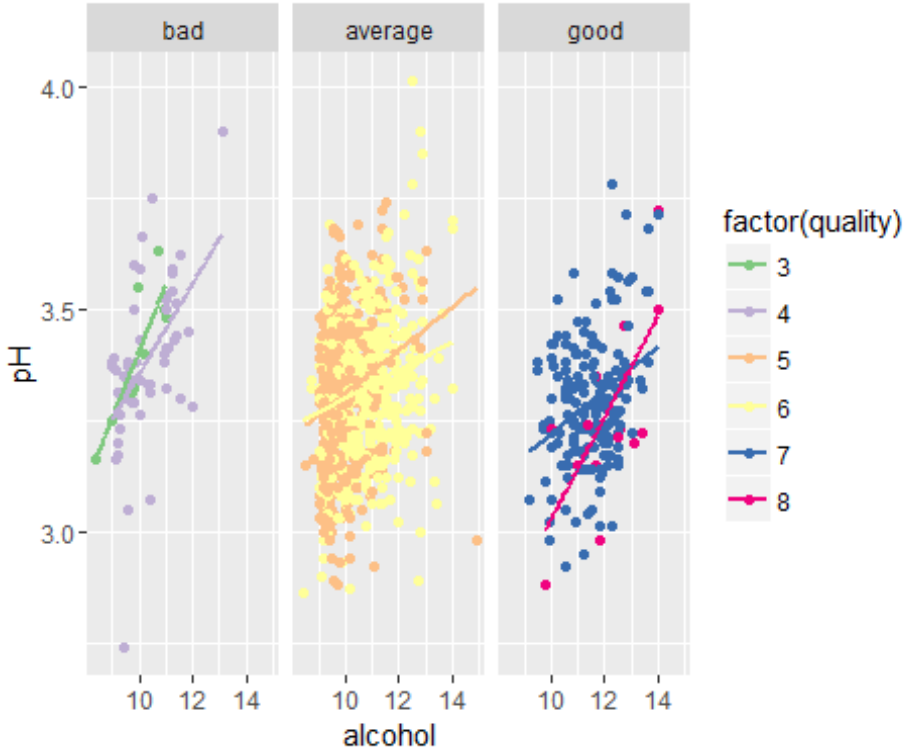
When we hold alcohol constant, there is no evidence that density affects quality whereas previous bivariate plot showed that density had an impact on the quality of wine.



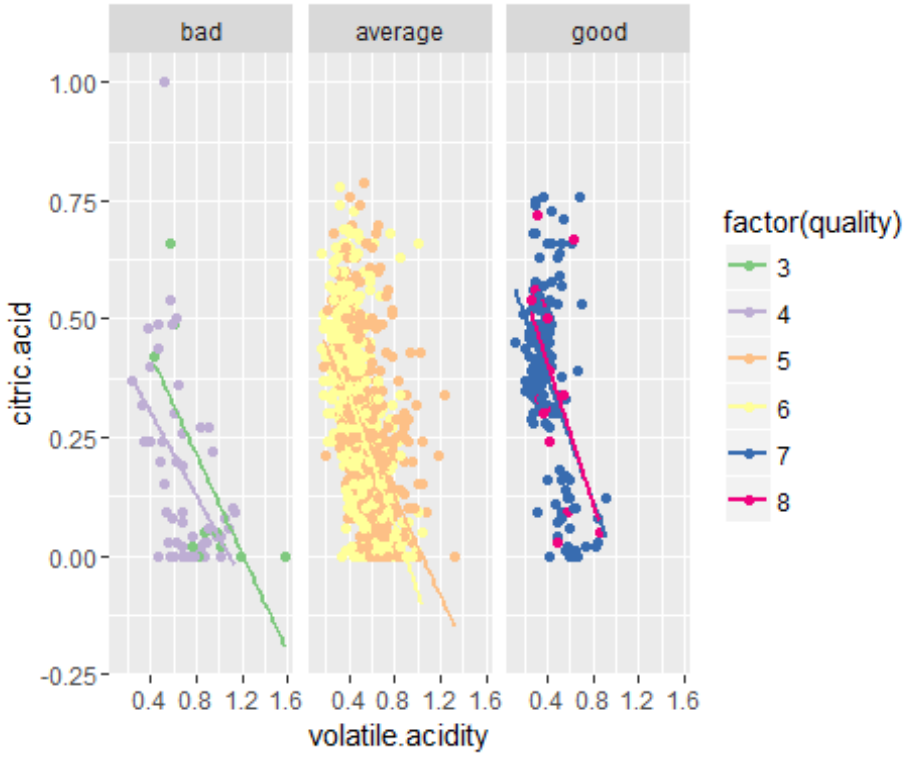


It seems that for wines with high alcohol content, having a higher concentration of sulphates produces better wines.





###Low pH and high alcohol concentration seem to be a good match.



###High citric acid and low acetic acid seems like a good combination.

Linear model

```
##
## Calls:
## m1: lm(formula = quality ~ alcohol, data = wine)
## m2: lm(formula = quality ~ alcohol + sulphates, data = wine)
## m3: lm(formula = quality ~ alcohol + sulphates + volatile.acidity,
##     data = wine)
## m4: lm(formula = quality ~ alcohol + sulphates + volatile.acidity +
##     citric.acid, data = wine)
## m5: lm(formula = quality ~ alcohol + sulphates + volatile.acidity +
##     citric.acid + fixed.acidity, data = wine)
##
##
================================================================================
##                         m1         m2         m3         m4         m5
## --------------------------------------------------------------------------------
-
##   (Intercept)        1.875***   1.375***   2.611***   2.646***   2.202***
##                     (0.175)    (0.177)    (0.196)    (0.201)    (0.224)
##   alcohol            0.361***   0.346***   0.309***   0.309***   0.320***
##                     (0.017)    (0.016)    (0.016)    (0.016)    (0.016)
##   sulphates                     0.994***   0.679***   0.696***   0.701***
##                                (0.102)    (0.101)    (0.103)    (0.103)
##   volatile.acidity                         -1.221***  -1.265***  -1.343***
```

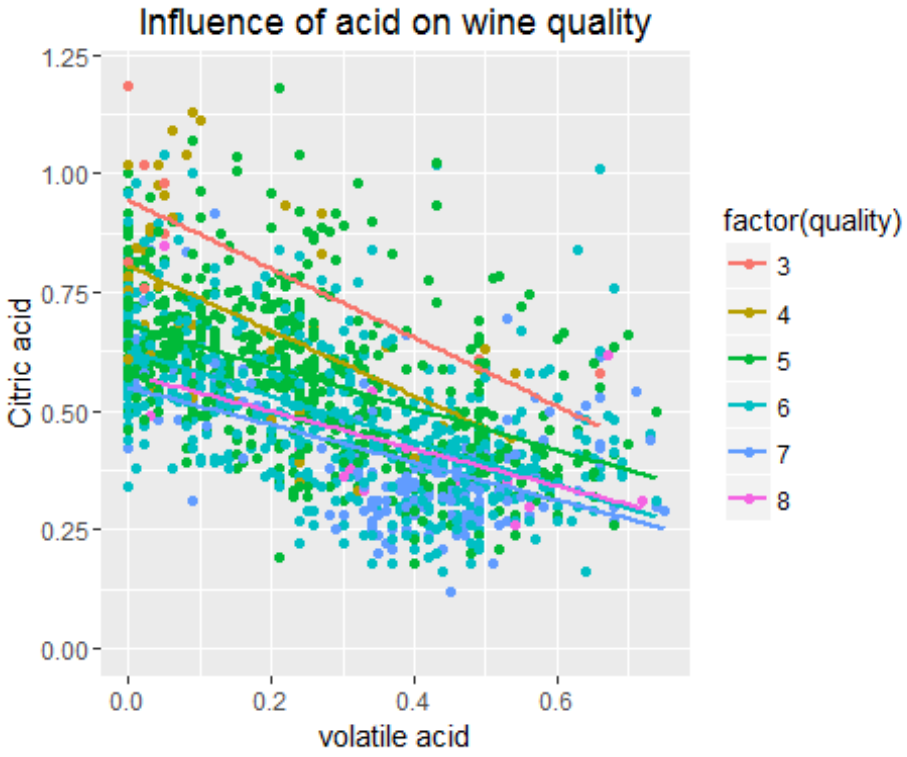
```
##                                          (0.097)    (0.113)    (0.113)
##   citric.acid                                       -0.079     -0.469***
##                                                     (0.104)    (0.137)
##   fixed.acidity                                                 0.057***
##                                                                (0.013)
## ------------------------------------------------------------------------
-
##   R-squared              0.2        0.3        0.3        0.3        0.3
##   adj. R-squared         0.2        0.3        0.3        0.3        0.3
##   sigma                  0.7        0.7        0.7        0.7        0.7
##   F                    468.3      295.0      268.9      201.8      167.0
##   p                      0.0        0.0        0.0        0.0        0.0
##   Log-likelihood     -1721.1    -1675.1    -1599.4    -1599.1    -1589.6
##   Deviance             805.9      760.9      692.1      691.9      683.7
##   AIC                 3448.1     3358.3     3208.8     3210.2     3193.3
##   BIC                 3464.2     3379.8     3235.7     3242.4     3230.9
##   N                   1599       1599       1599       1599       1599
##
========================================================================
```

High alcohol contents and high sulphate concentrations combined seem to produce better wines.

I created several models. The most prominent of them was composed of the variables alcohol, sulphates, and the acid variables. There are two problems with it. First the low R squared score suggest that there is missing information to propely predict quality.

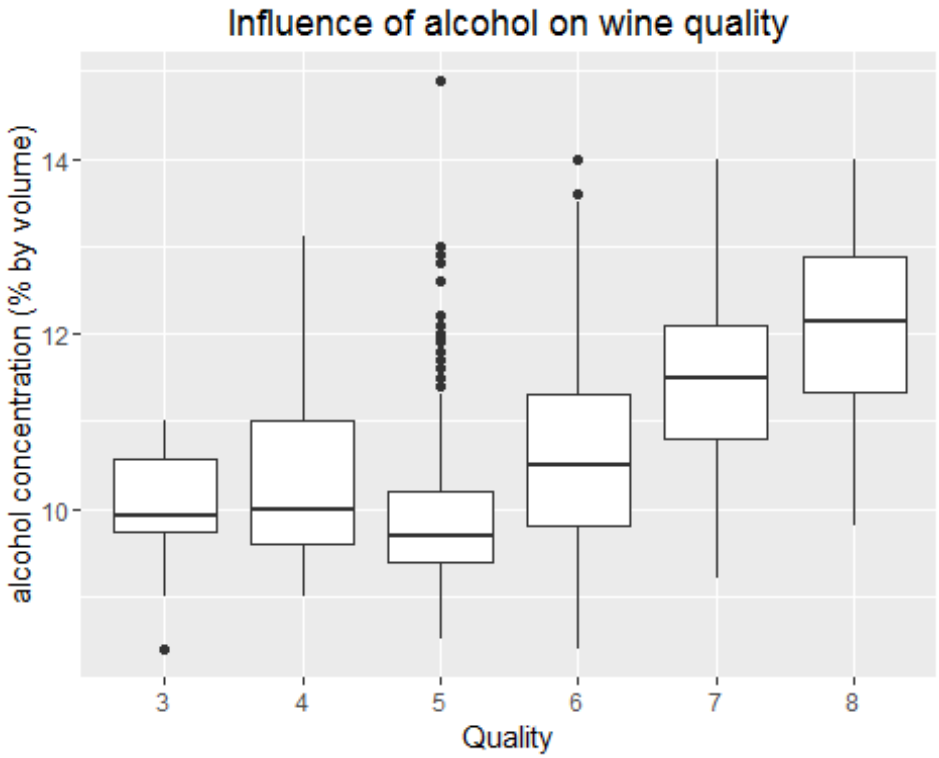
Final Plots and Summary

Plot 1: Effect of acid on wine quality



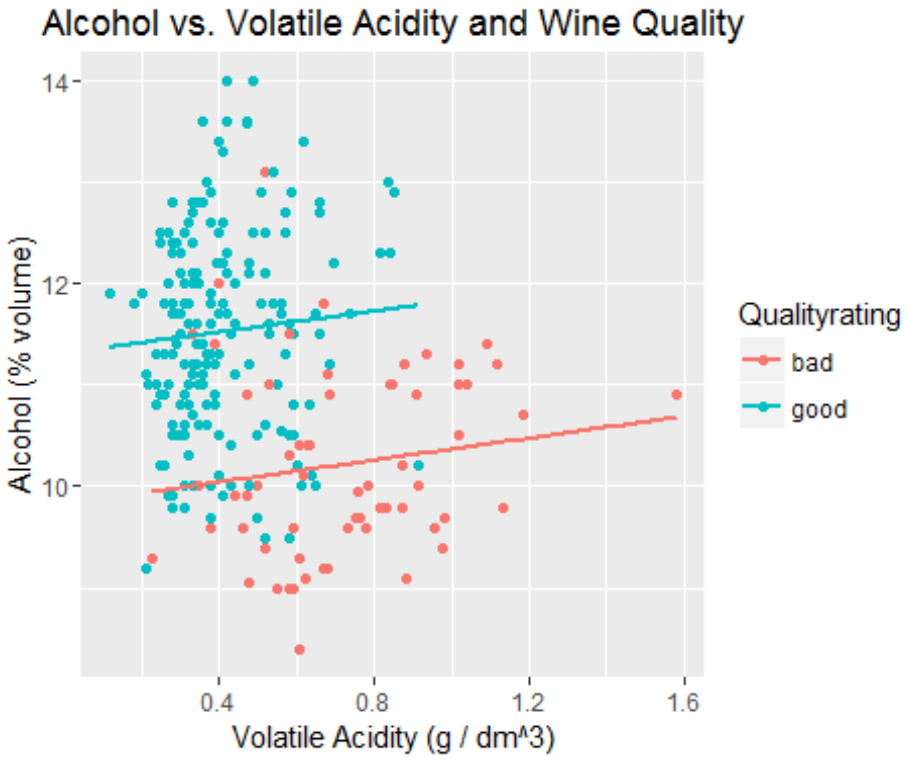
Plot shows that high citric acid and low volatile acid leads to somewhat better wine.

Plot 2: Effect of alcohol on wine quality



These boxplots demonstrate the effect of alcohol content on wine quality. Generally, higher alcohol content correlated with higher wine quality. However, as the outliers and intervals show, alchol content alone did not produce a higher quality.

Plot 3: What makes good wines, good, and bad wines, bad?



This is perhaps the most telling graph. I subsetted the data to remove the 'average' wines, or any wine with a rating of 5 or 6. As the correlation tests show, wine quality was affected most strongly by alcohol and volaticle acidity. While the boundaries are not as clear cut or modal, it's apparent that high volatile acidity--with few exceptions--kept wine quality down. A combination of high alcohol content and low volatile acidity produced better wines.

Reflection

The wine data set contains information on the chemical properties of a selection of wines collected in 2009. It also includes sensorial data (wine ranking).I started by looking at the individual distributions of the variables, trying to get a feel for each one.

The first thing I noticed was the high concentration of wines in the middle ranges of the ranking, that is, average tasting wines. This proved to be very problematic during the analysis as I kept questioning myself wether there was a true correlation between two variables or it was just a coincidence given the lack of "outlier" (poor and excellent) wines.

After exploring the individual variables, I proceded to investigate the relationships between each input variable and the outcome variable quality.The most promissing varibles were alcohol concentration, sulphates and the individual acid concentrations.

On the final part of the analysis I tried using multivariate plots to investigate if there were interesting combinations of variables that might affect quality. I also used a multivariate plot to confirm that density did not have an effect on quality when holding alcohol concentration constant.In the end, the produced model could not explain much of the variance in quality. This is further corroborated acidity analysis.

For future studies, it would be interesting to mesure more acid types in the analysis. Wikipedia for example, suggests that malic and lactic acid are important in wine taste and these were not included in this sample.