Red Wine Quality

Saeideh Shahrokh Esfahani

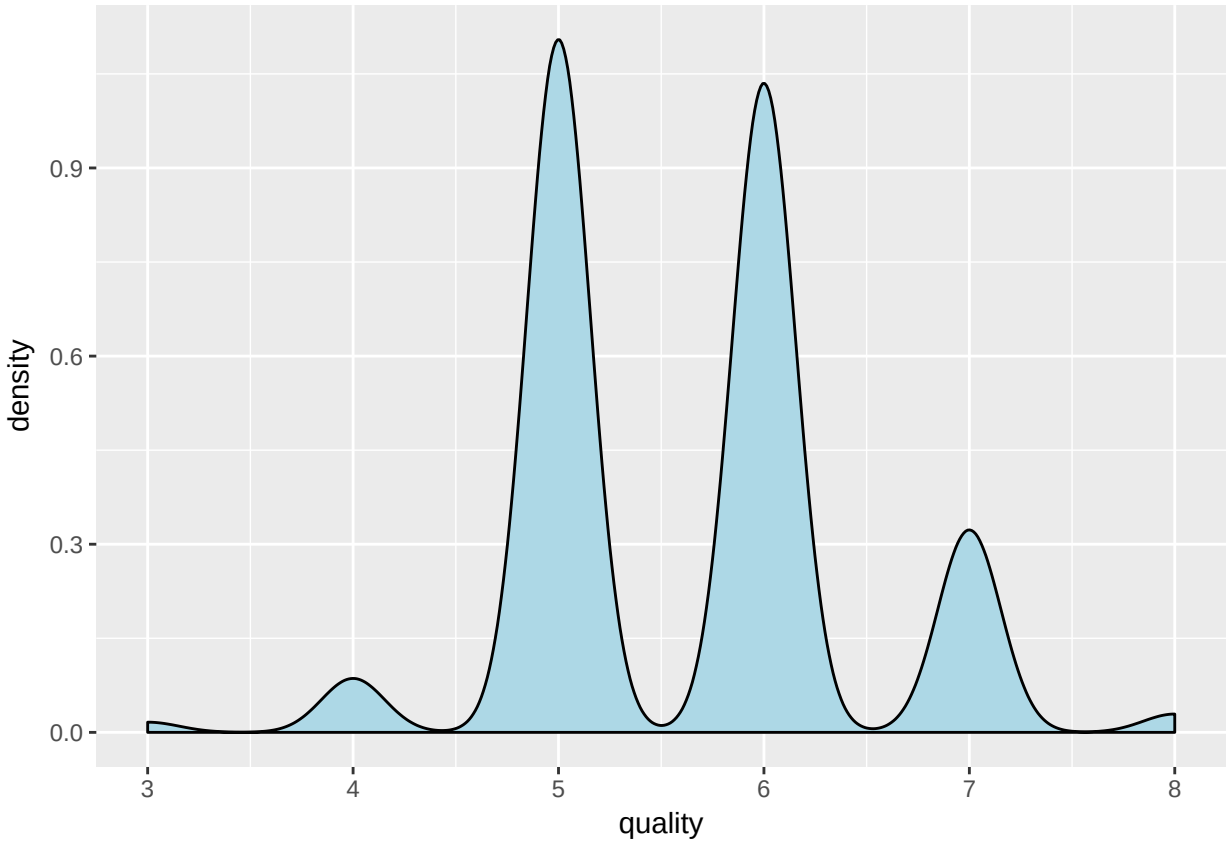
4/17/2017

```
knitr::opts_chunk$set(echo = TRUE)
```

This dataset containes red wine samples. The inputs include objective tests (e.g. alcohol value) and the output is based on data which came from a median of at least 3 evaluations made by wine experts. The quality of red wine was graded between 0 (very bad) to 10 (very excellent) by experts.

I will start my investigation of data with univariate plots where I will use them to get some sence about the variables. Then, I will move on through the bivariate plots, to have more investigation and find out the potential transformation needed for further investigations. Finally, I will implement multivariate plot. In this section I will work on predicting the relationship between diffrenet ingredient and quality of wine using Lasso feature selection.

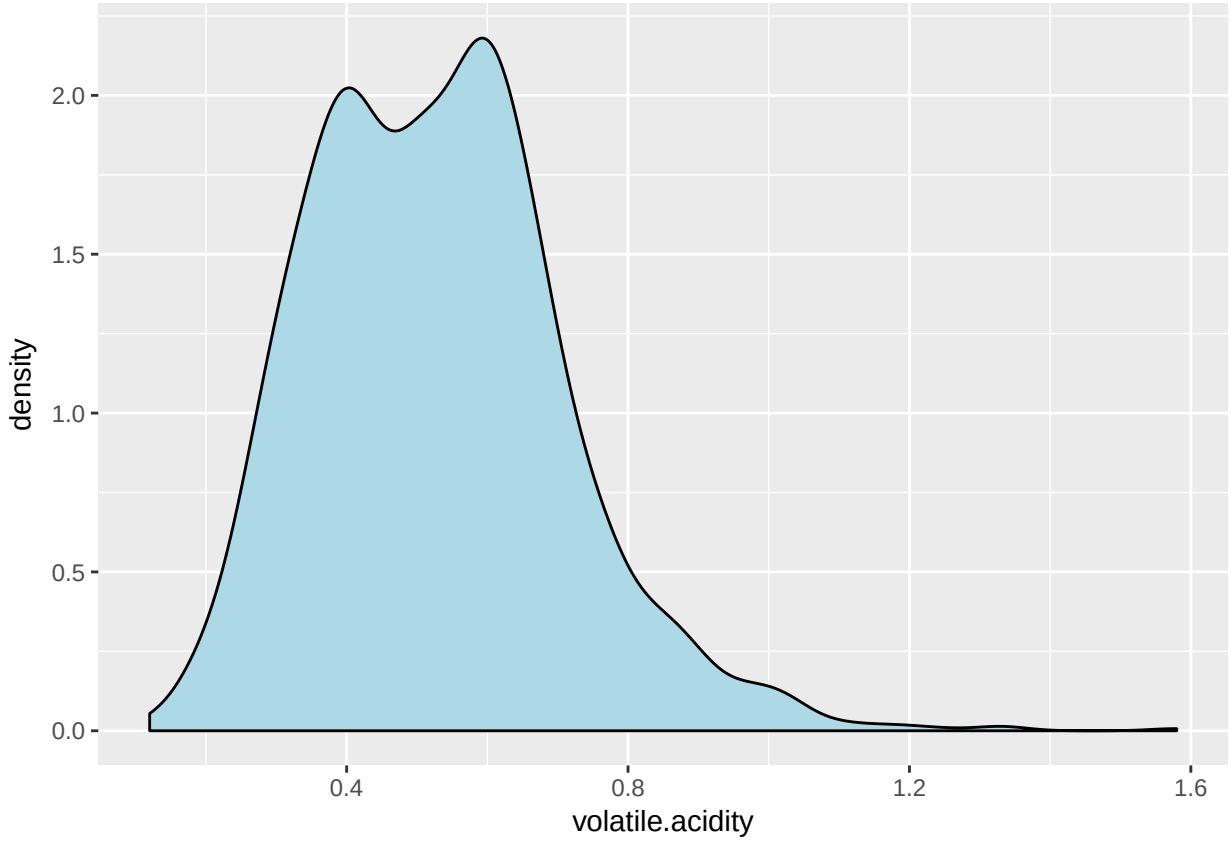
Univariate Plots Section



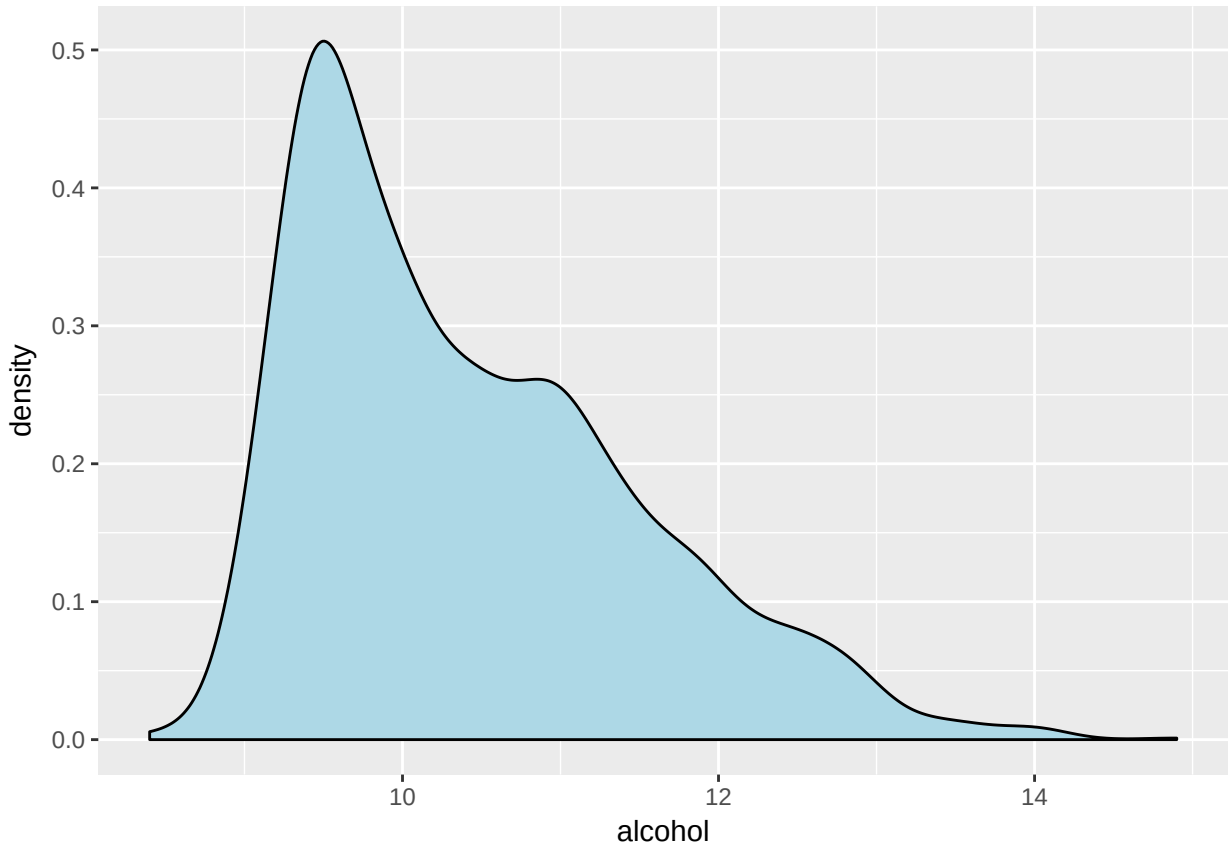
Based on the plot, the number of sample with quality of 5 and 6 are higher than athers. We can see also the the same result with using table function for the number of different qualities:

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

Here, I am going to exploring the distribution of different ingredients of red wine. The summary of each plot comes after the its plot.

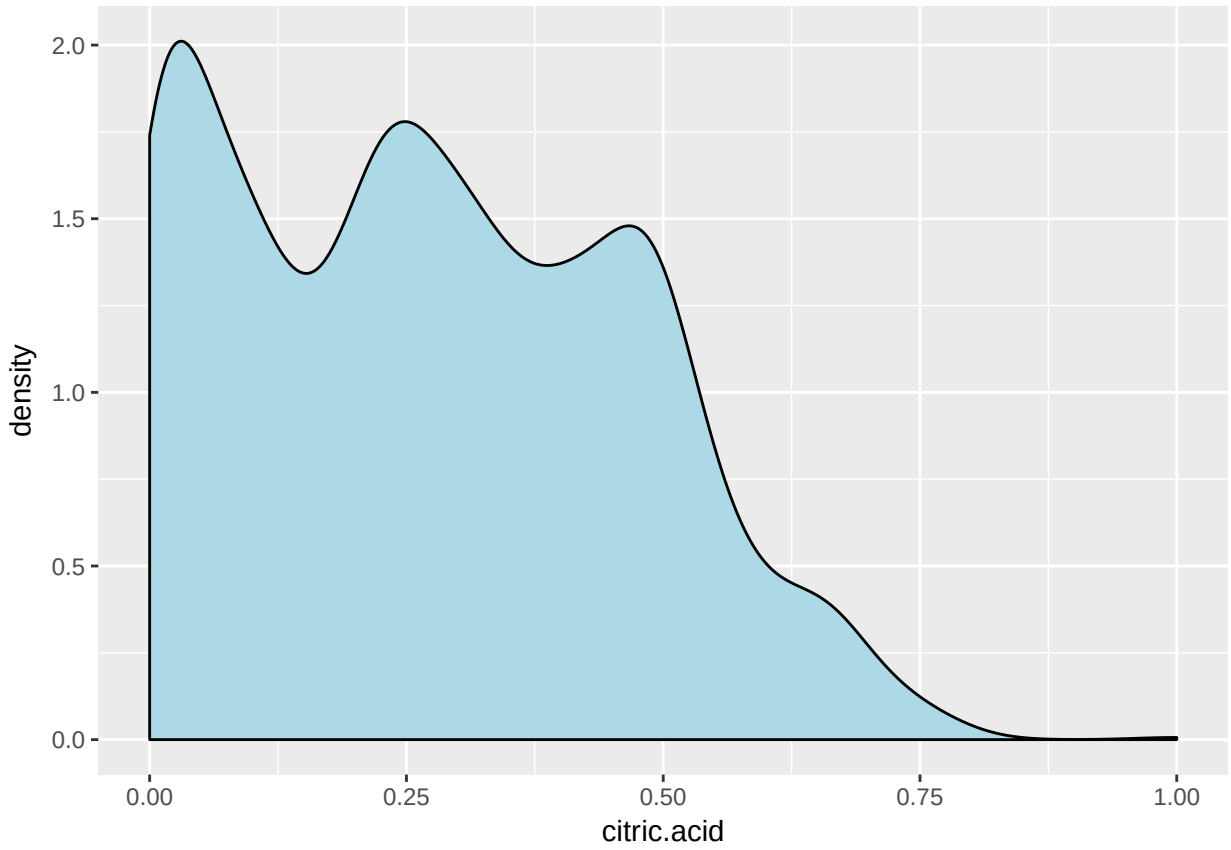



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```



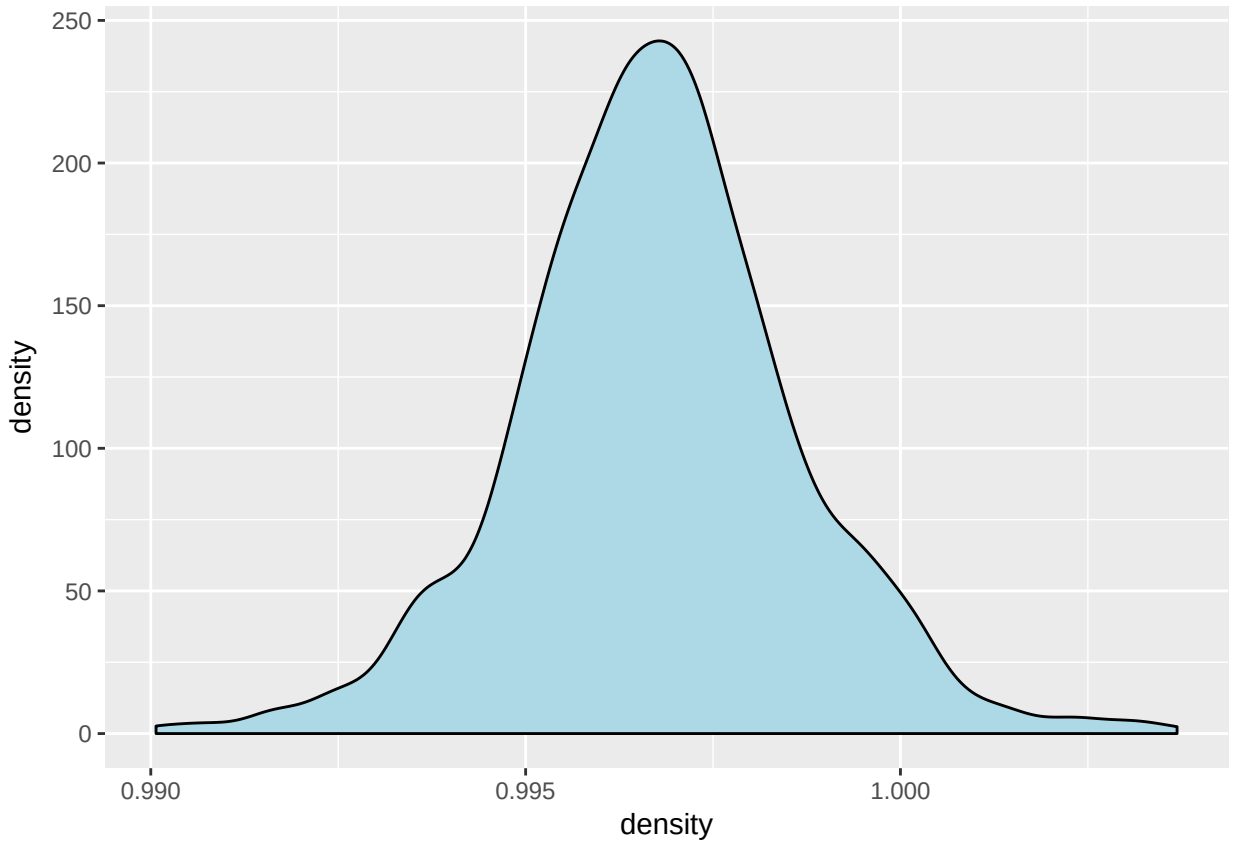

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```

It seems that all plots showed above are rigeht-skewed (positive skewness).

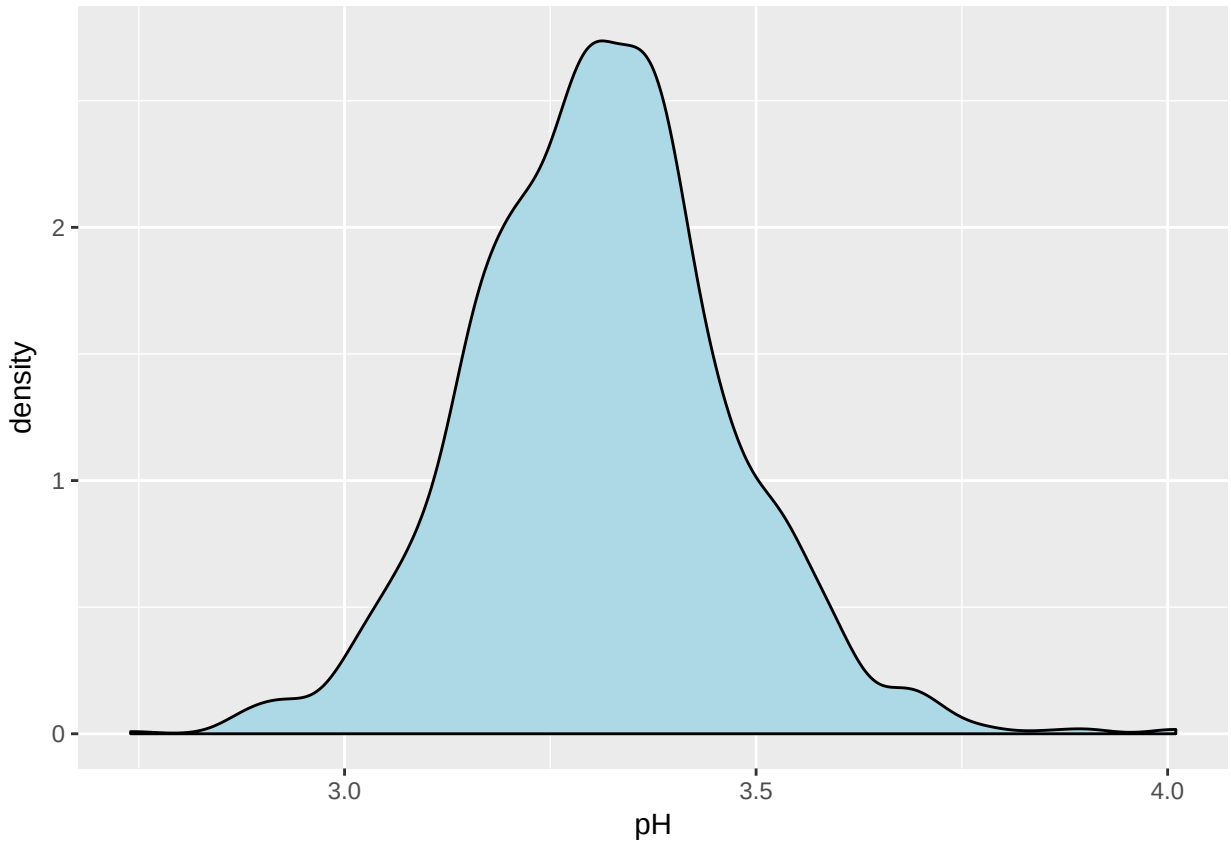



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```

The plot related to citric acid is not only right-skewed also has 3 peacks.




```
##    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.
##   2.740   3.210   3.310   3.311   3.400   4.010
```

As the plots depict, the pH and density have quite well normal distribution.

Univariate Analysis

The structure of the dataset is tidy data.

The feature that I am mostly interested in is how different wine ingredients could statistically affect quality of the red wine. Hence, a winer can in theory develop a line of wine production with some expected wine quality.

Further more, other features in the dataset that I think will help support my investigation into my feature(s) of interest could be included was the variance among wine experts' opinions. This dataset only provides a consensus of the qualities given by experts, however it would be important to see how this assessment varies across the experts.

Bivariate Plots Section

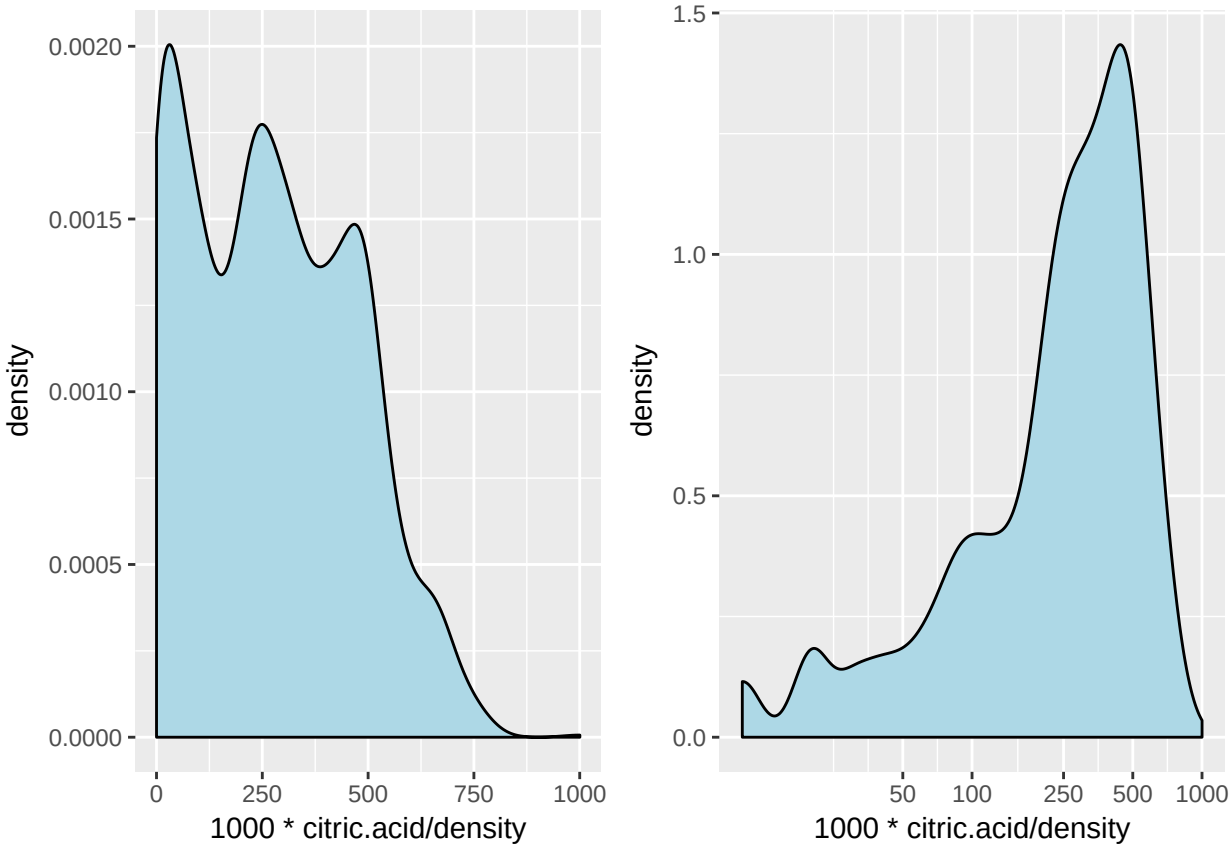
As plots in the Univariate section depicted,I am interested to investigate 3 relationship wich I briefly explaine them as follow:

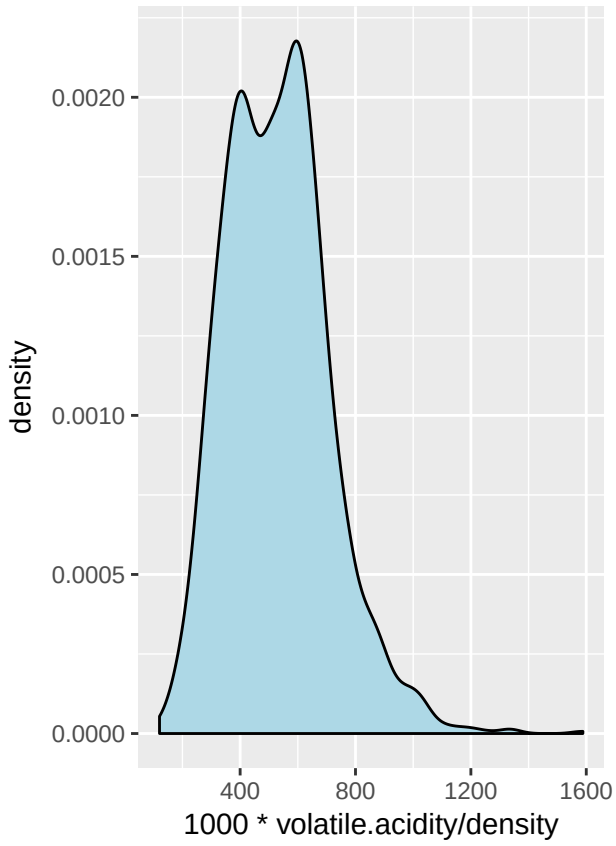
1) Regardless of pH and density distribution which had quite normal distributions, others ingridiend had right-skewed and citric acid had also 3 peacks. Hence, I am going to plot the logarithm of ratio of each

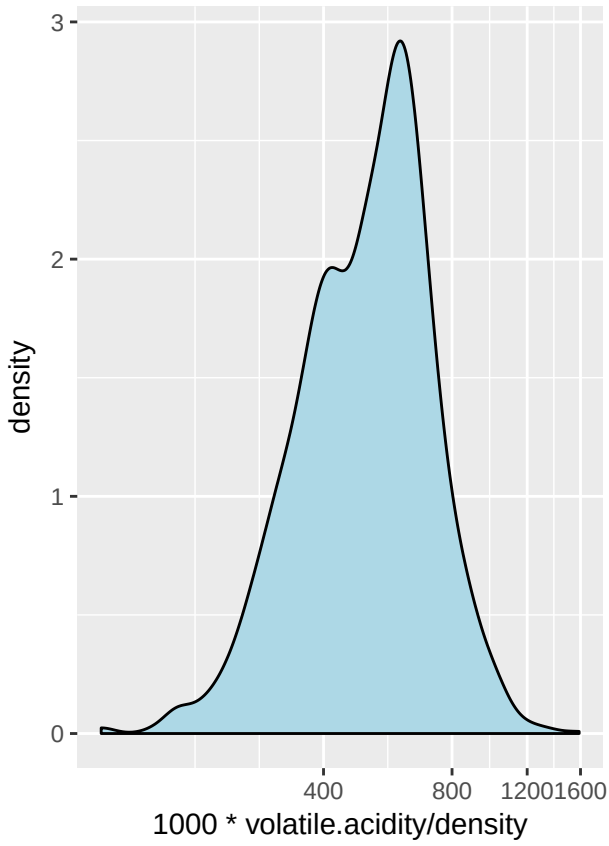
skewed ingredients over density. In this transformation of data I consider also the dimention being unique to "g/cmˆ3".

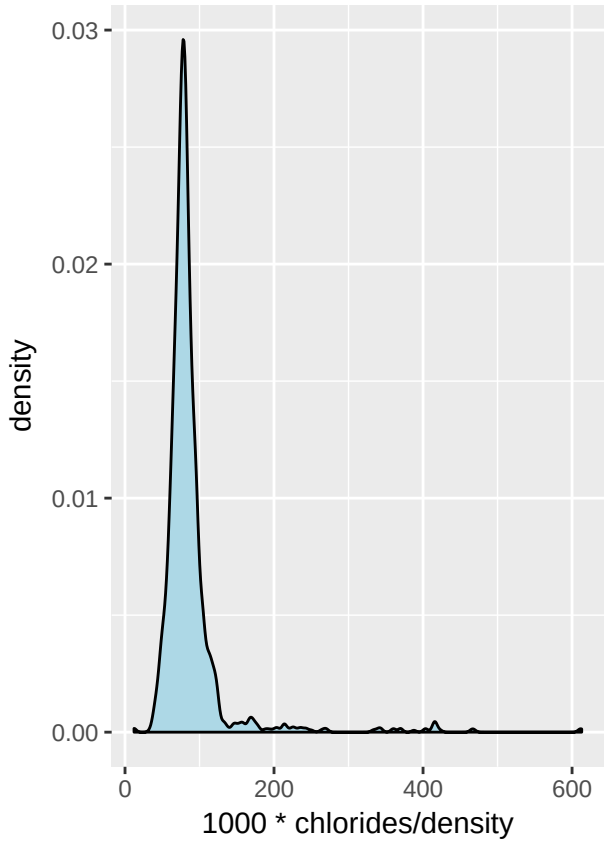
2) Also, Based on investigation in the univariate plots, I am curiuos to figure out different relationships between pH and different acids that inovolves in the ingredients of the red wine. Based on chemistry, I would expect higher acidity would lead to a lower pH.
3) Finally I would like to explore the relationship between the ingredients and quality.

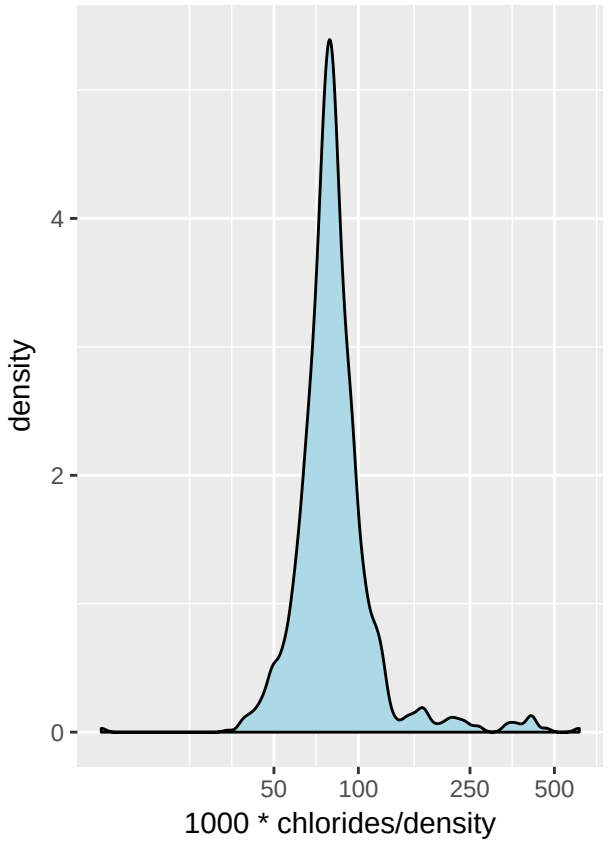
1) Data Transformation

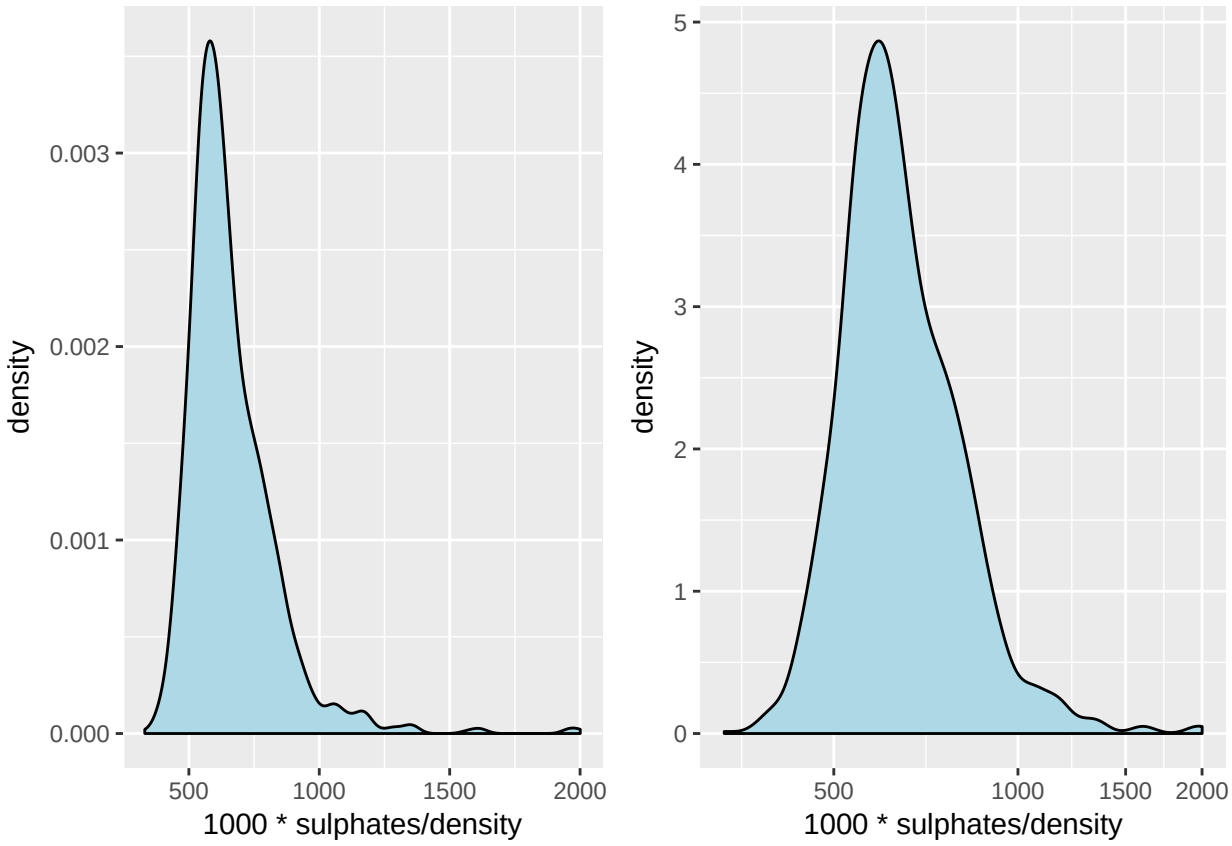



density
0.0020
0.0015
0.0010
0.0005
0.0000
400
800
1200
1600
1000 * volatile.acidity/density


density
3
2
1
0
400
800
1200
1600
1000 * volatile.acidity/density


density
0.03
0.02
0.01
0.00
0
200
400
600
1000 * chlorides/density

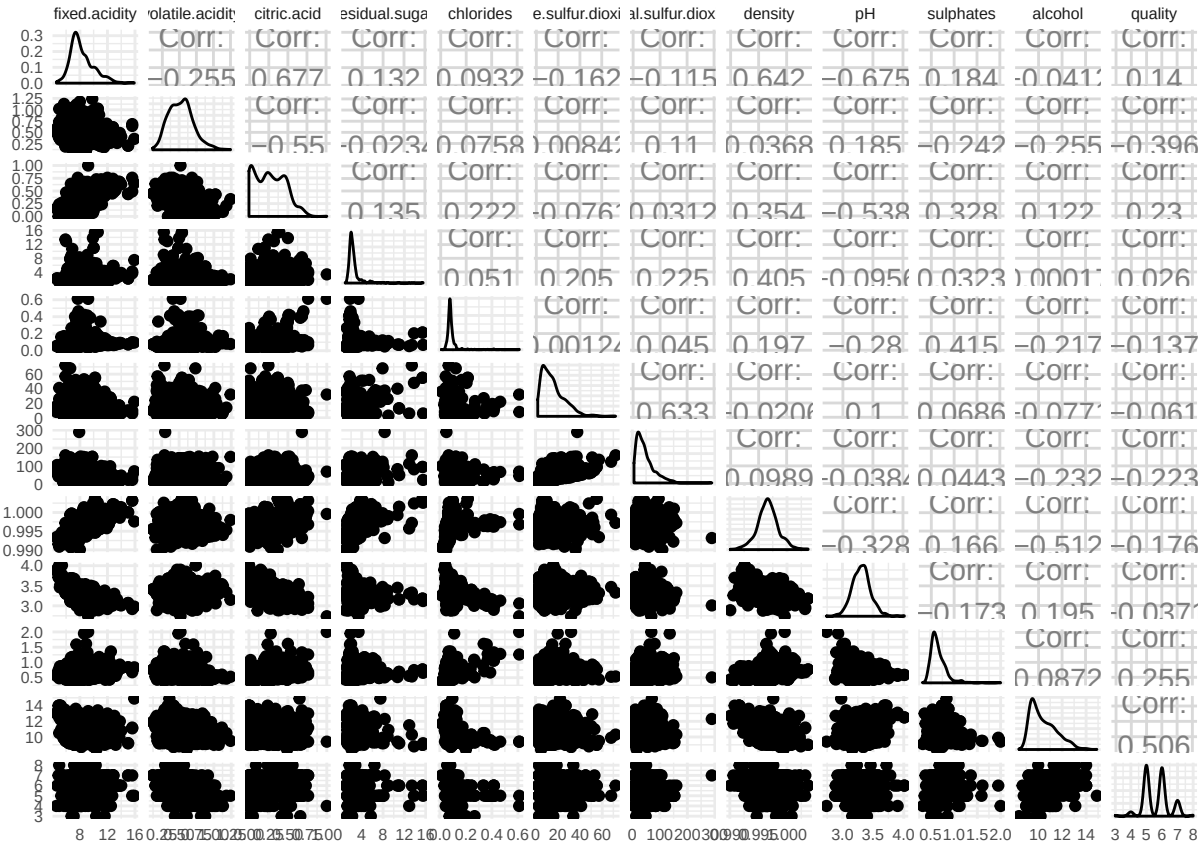

density
4
2
0
50
100
250
500
1000 * chlorides/density



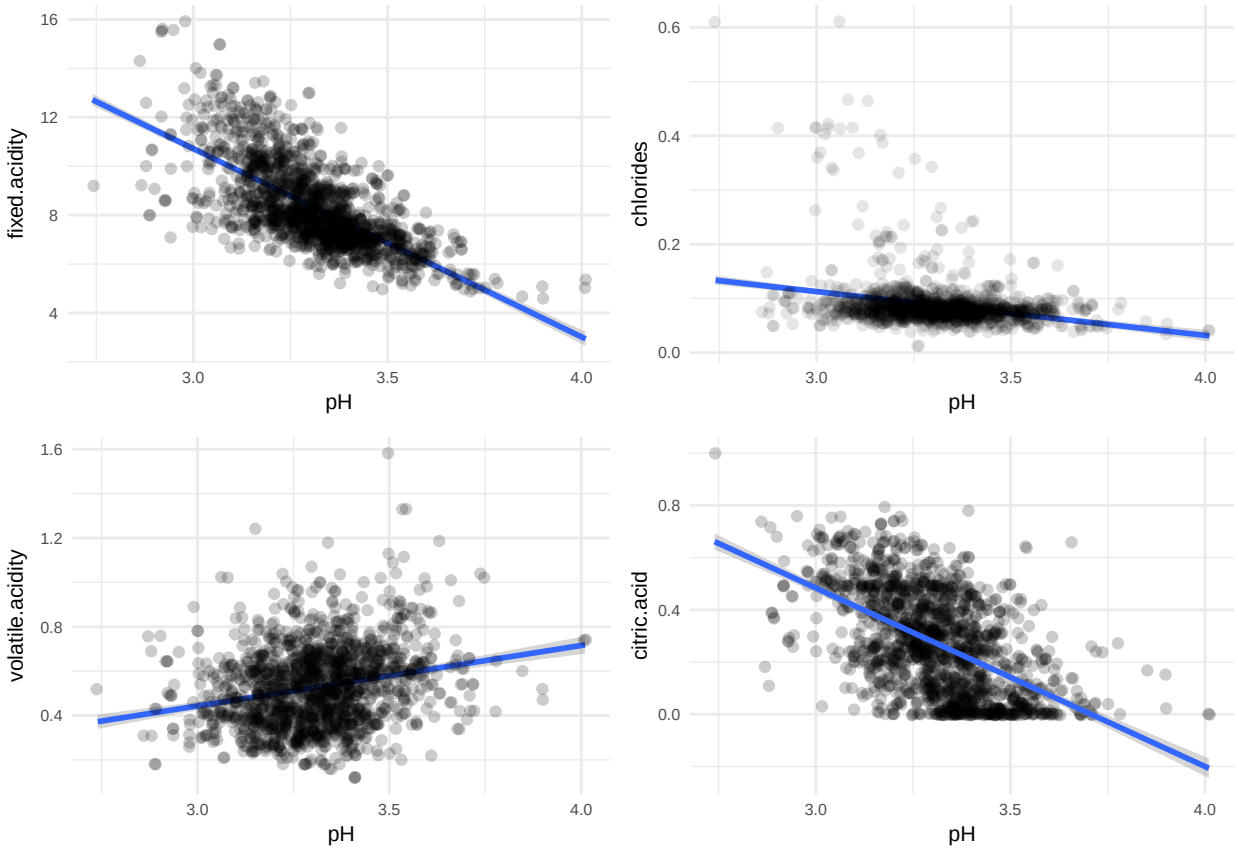
As the plots showed, the distribution of ingredient would be improved by this tranformation. For all of them, without considering of outliers, one may notice fairly normal distribution in logarithmic distribution.

2) Correlation exploration:

```
##  [1] "fixed.acidity"        "volatile.acidity"     "citric.acid"
##  [4] "residual.sugar"       "chlorides"            "free.sulfur.dioxide"
##  [7] "total.sulfur.dioxide" "density"              "pH"
## [10] "sulphates"            "alcohol"              "quality"
```


fixed.acidity
olatile.acidity
citric.acid
esidual.suga
chlorides
e.sulfur.dioxi
l.sulfur.diox
density
pH
sulphates
alcohol
quality

Investigation of relationship between Acids and pH



As I expected the correlation of pH and acidity was negative one, however one might notice that for volatile acidity (acetic acid), this correlation is positive. Also it seems that the amount of choloride acid in red wine is less than oter acids. In order to more exploration I find the correlation of pH and acids as follow:

pH and fixed acidity:

```
## [1] -0.6829782
```

pH and chlorides:

```
## [1] -0.2650261
```

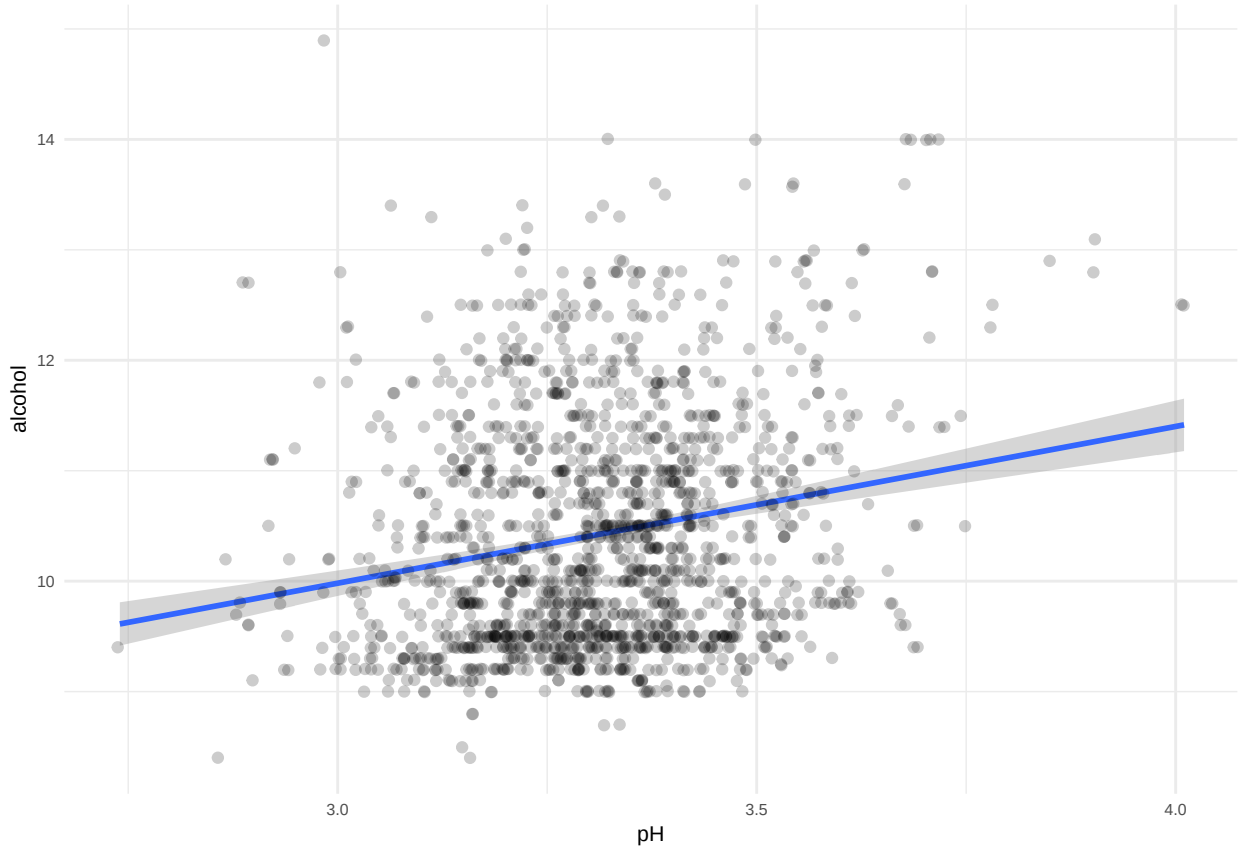
pH and volatile acidity:

```
## [1] 0.2349373
```

pH and citric acid:

```
## [1] -0.5419041
```

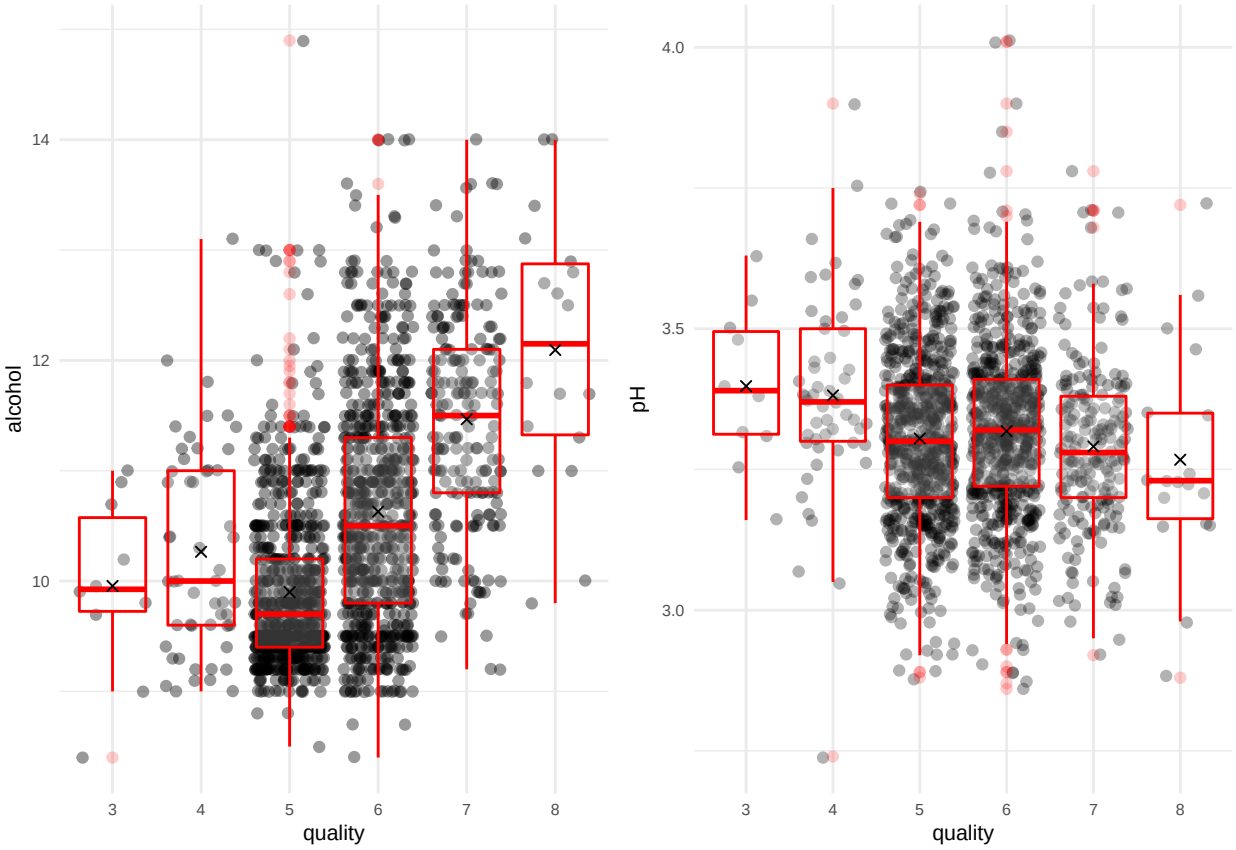
The following scatter plot shows the correlation between pH and alcohol:



The Pearson correlation is:

```
## [1] 0.2056325
```

3) Investigation of relationship between ingredients and quality

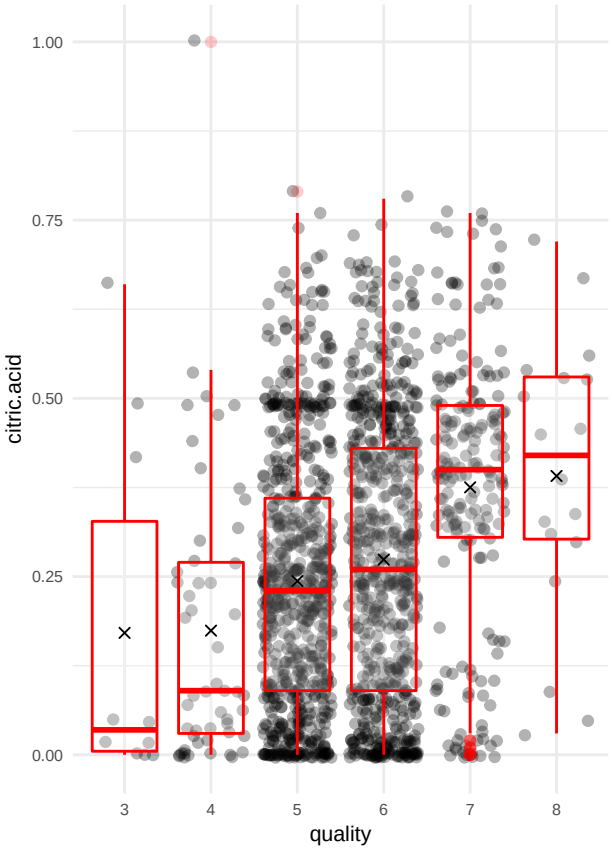


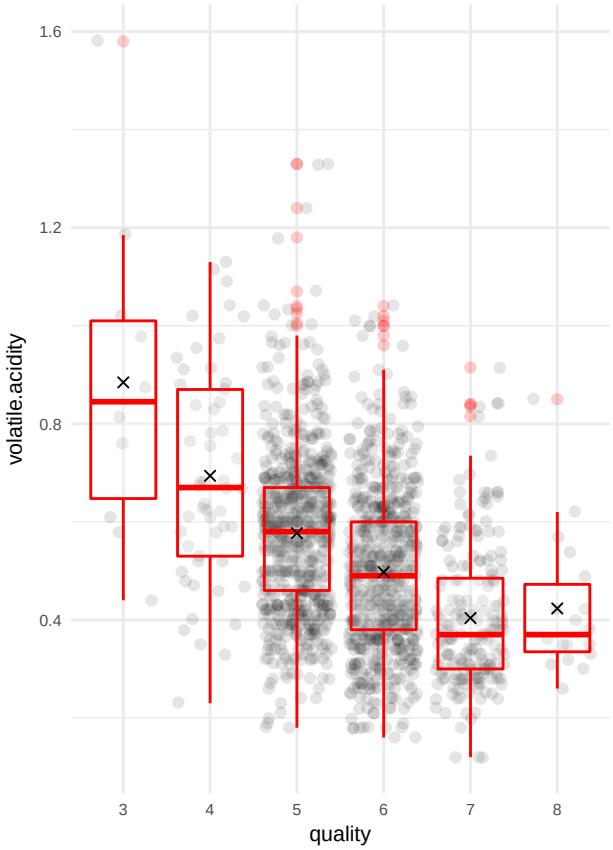
The correlation between quality and alcohol:

```
## [1] 0.4761663
```

The correlation between quality and pH:

```
## [1] -0.05773139
```



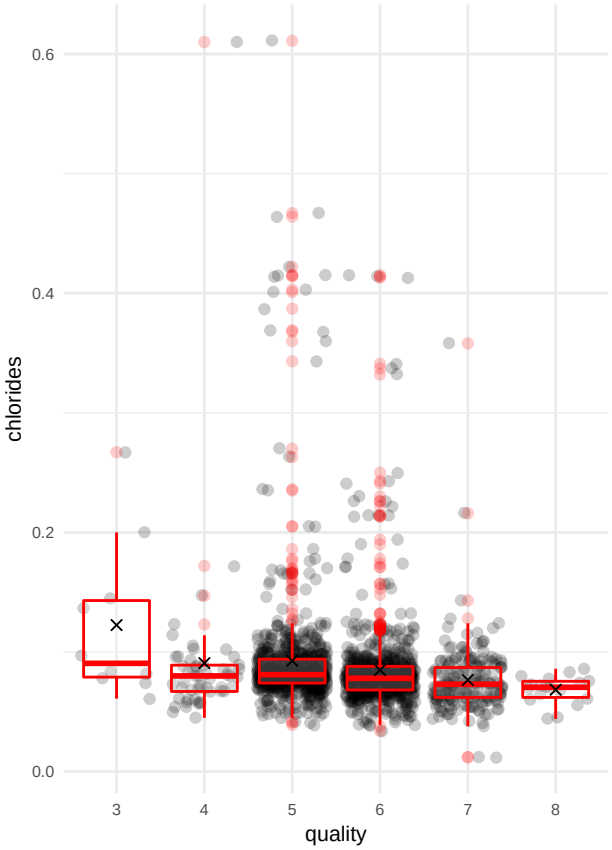


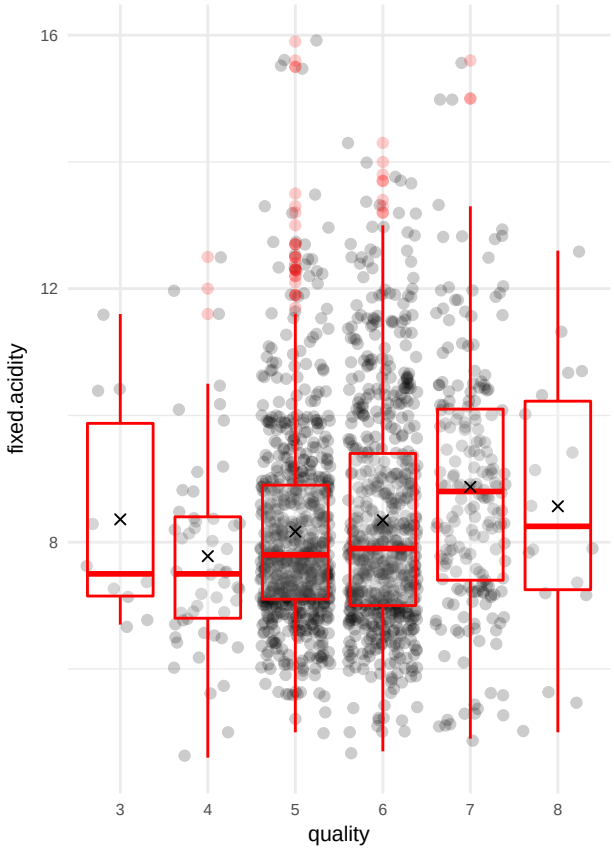
The correlation between quality and citric acid:

```
## [1] 0.2263725
```

The correlation between quality and volatile acidity:

```
## [1] -0.3905578
```



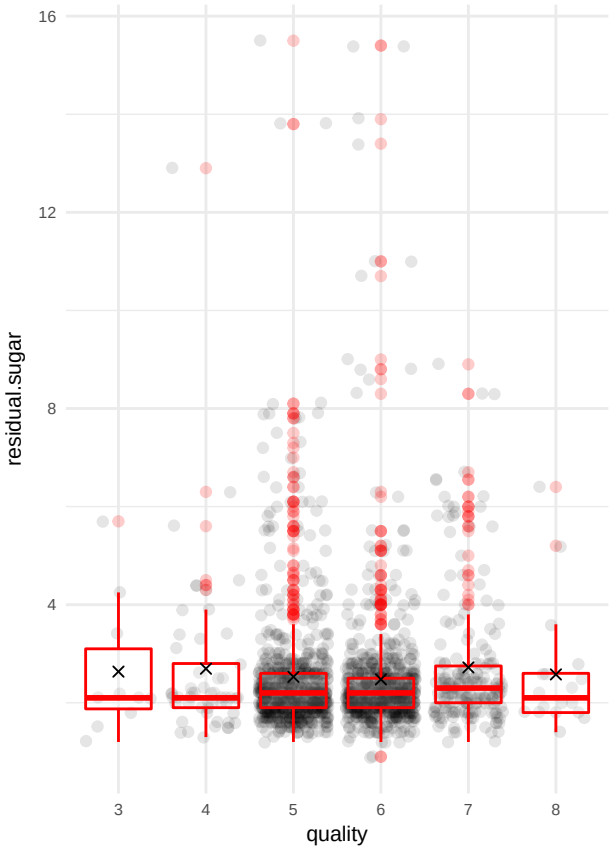


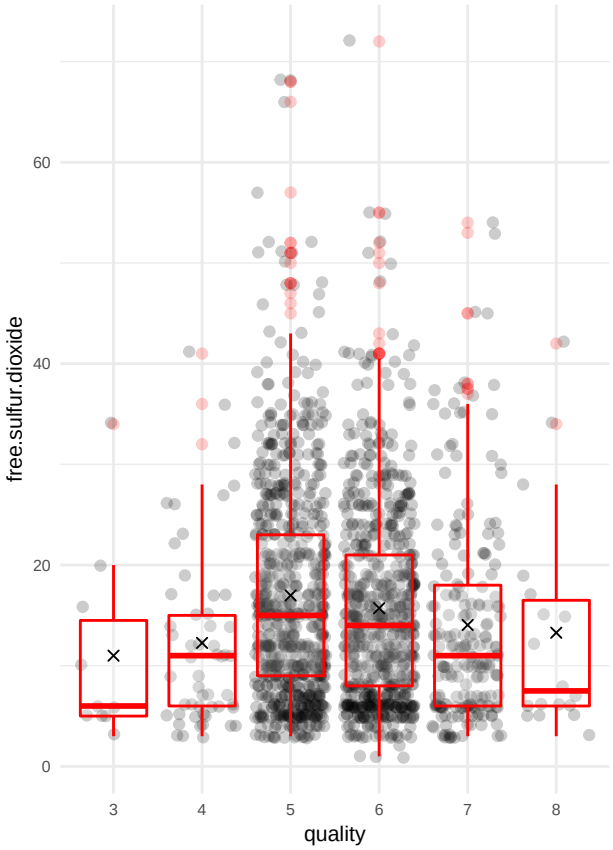
The correlation between quality and chlorides:

```
## [1] -0.1289066
```

The correlation between quality and fixed acidity:

```
## [1] 0.1240516
```





The correlation between quality and residual sugar:

```
## [1] 0.01373164
```

The correlation between quality and free sulfur dioxide:

```
## [1] -0.05065606
```

The correlation between quality and total sulfur dioxide:

```
## [1] -0.1851003
```

Bivariate Analysis

Based on the plot shown above there is not a strong correlation between quallity and pH (almost -0.06). In contrast it seems there is a correlation between alcohol and quality, but this correlation may not be a linear one.

As we expected that higher acidity, less pH, there are negative correlations between pH and citric acid, tartaric acid (related to fixed acidity parameter) and acid chorolide. However the acetic acid surprisingly has a positive correlation with pH.

I observed the interesting relationships between the feature that I will explaine brieflly bellow.

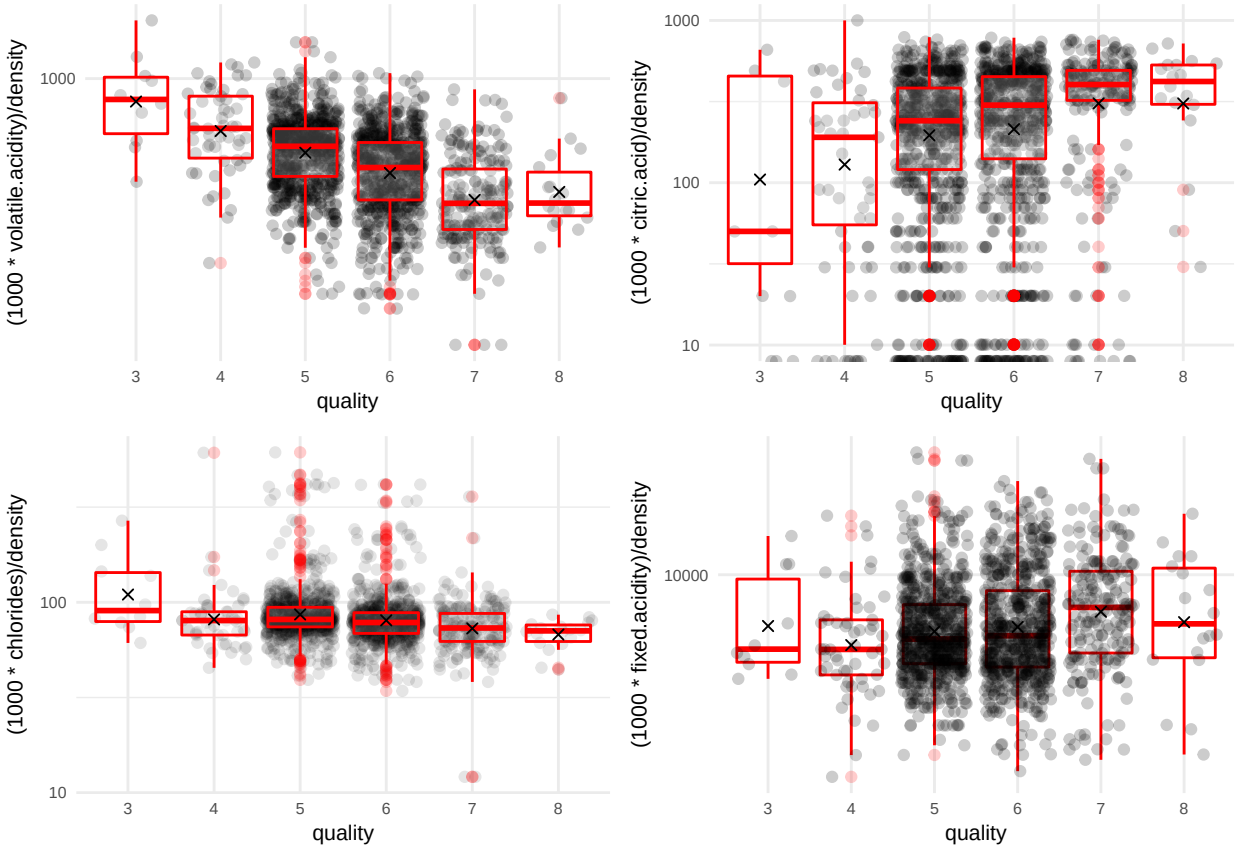
Apparantly acetic acid and pH manifest positive correlation, while I would have expected a negative one, because higher acidity would mean less pH. It might happen becuase of the fact that acetic acid is not a strong acid like citric acid and two others, or it could be due to other interacting features, i.e. acetic acid is not the sole factor, and that the correlation could be confounded by the variabiliy in other features.

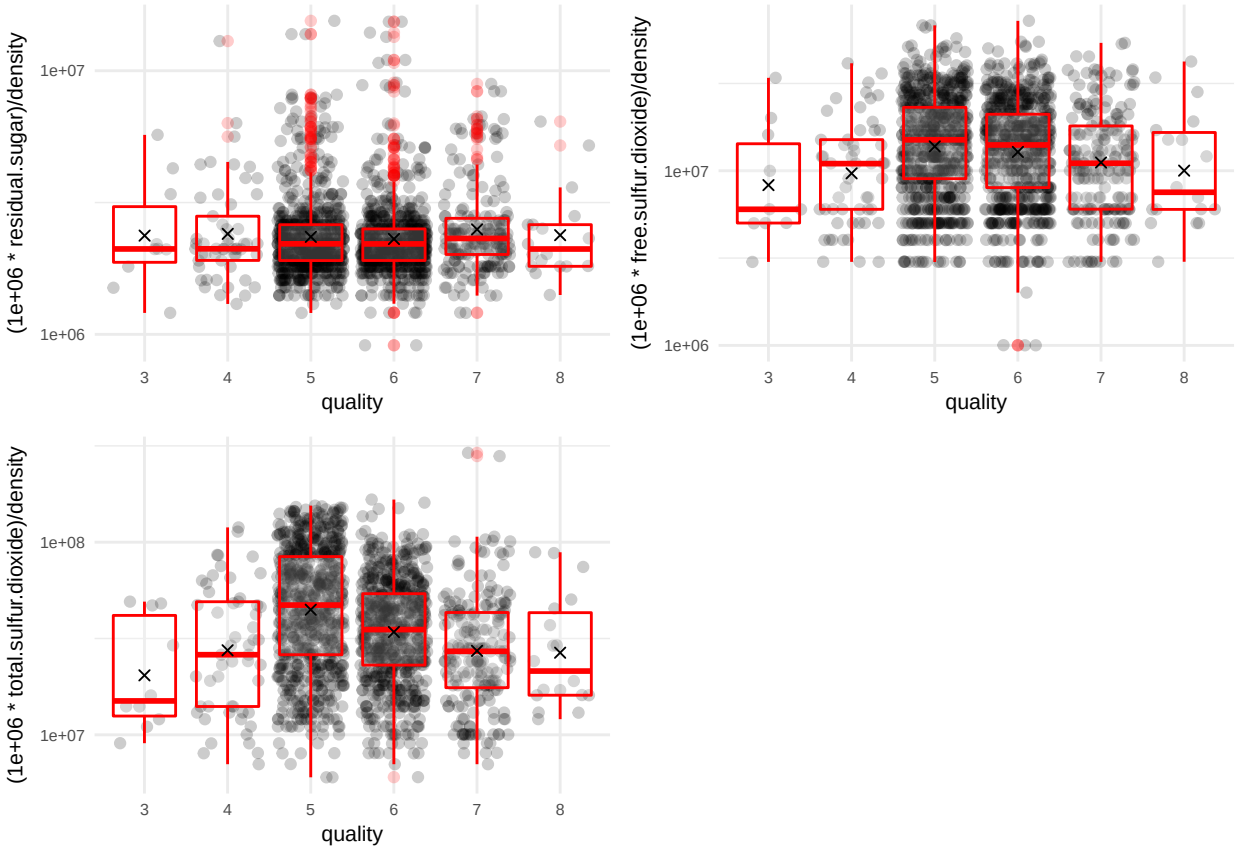
Another interesting fact that I realized was the relationship between quality and different acids. One would see both positive and negative correlations between diffrent acids and quality.

I think there are in fact strong relationships between quality and the “logarithm of ratio of acids to densities”.

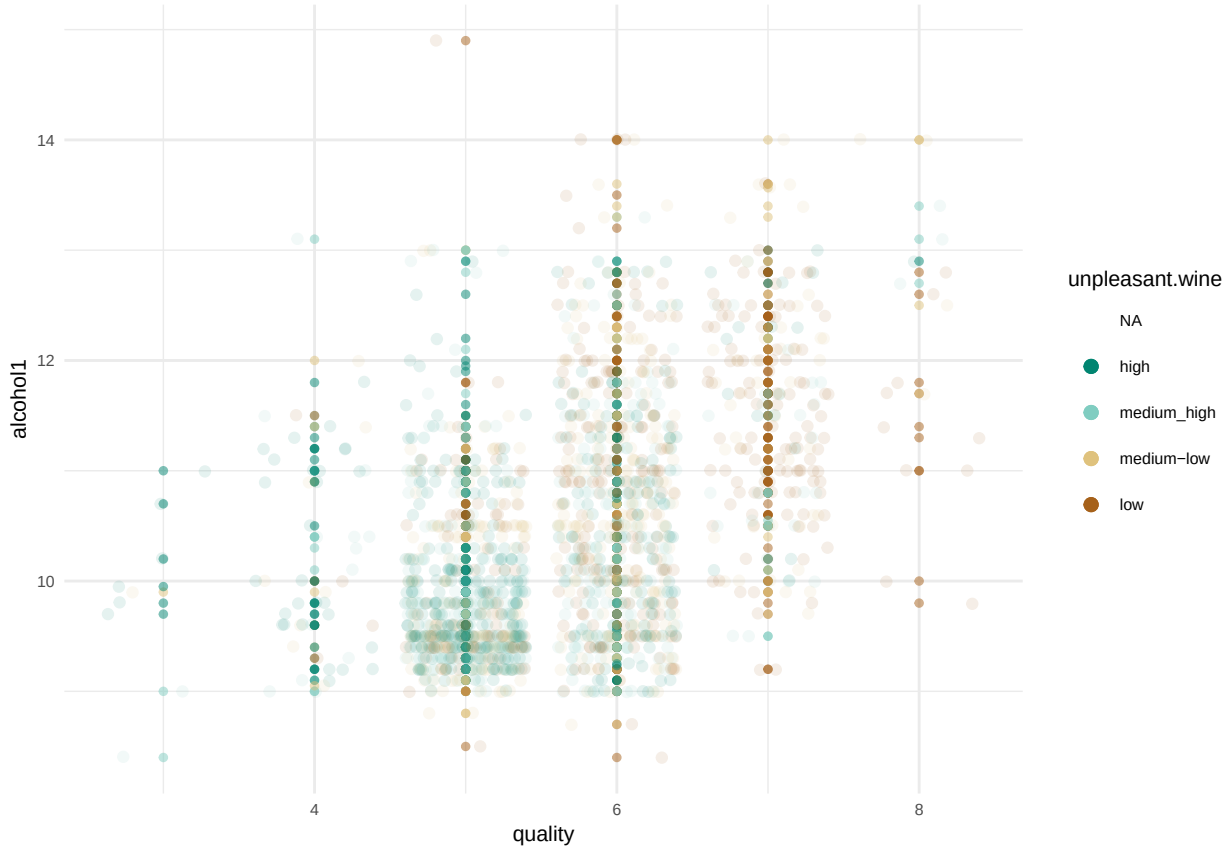
Also, based on the polts the data seemed somewhat skewed with some outliers, and hence I decided to consider the logarithmic distribution of the ratio of ingredient to density. So I add the logarithm of the ratio of ingerdiens and density. I also consider the dimansion to be unique to g/cm^3.

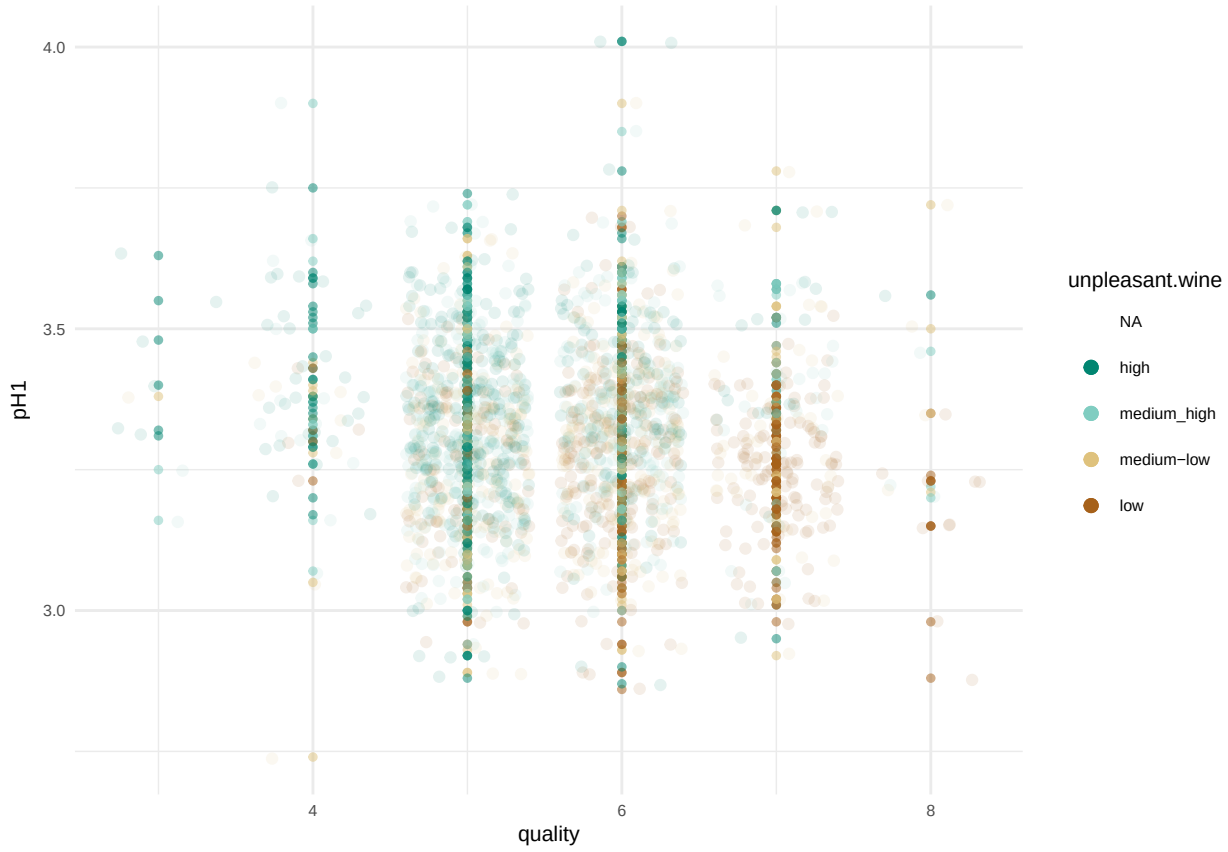
Multivariate Plots Section





The high levels of acetic acid (or the volatile acidity) in wine can lead to an unpleasant wine, vinegar taste. In order to investigate this fact, first I defined a 4-level categorical variable based on the volatile acidity as follows: [Min.,1st Qu.),[1st Qu., Median), [Median,3rd Qu), [3rd Qu, Max.]. Then I mapped this ranges to "low", "medium-low", "medium_high", "high" respectively to create the following plot.


alcohol1
14
12
10
4
6
8
quality
unpleasant.wine
NA
high
medium_high
medium−low
low


pH1
4.0
3.5
3.0
4
6
8
quality
unpleasant.wine
NA
high
medium_high
medium−low
low

Multivariate Analysis

pH and acidity

In order to consider the role of 4 different acids in pH, I use lineare regresion in order to investigate the effect of each one in predection of red wine pH.

```
##
## Calls:
## m1: lm(formula = pH ~ rfixed.acidity, data = wnew)
## m2: lm(formula = pH ~ rfixed.acidity + rcitric.acid, data = wnew)
## m3: lm(formula = pH ~ rfixed.acidity + rcitric.acid + rchlorides,
##     data = wnew)
## m4: lm(formula = pH ~ rfixed.acidity + rcitric.acid + rchlorides +
##     rvolatile.acidity, data = wnew)
##
## ================================================================
##                          m1         m2         m3         m4
## ----------------------------------------------------------------
##   (Intercept)         8.279***   7.790***   7.903***   7.881***
##                      (0.124)    (0.135)    (0.133)    (0.146)
##   rfixed.acidity     -0.551***  -0.489***  -0.467***  -0.467***
##                      (0.014)    (0.015)    (0.015)    (0.015)
##   rcitric.acid                  -0.014***  -0.014***  -0.014***
##                                 (0.002)    (0.002)    (0.002)
##   rchlorides                               -0.071***  -0.071***
##                                            (0.008)    (0.008)
##   rvolatile.acidity                                    0.003
##                                                       (0.009)
## ----------------------------------------------------------------
##   R-squared              0.501      0.521      0.543      0.543
##   adj. R-squared         0.500      0.520      0.542      0.542
##   sigma                  0.109      0.107      0.105      0.105
##   F                   1601.748    868.200    630.781    472.862
##   p                      0.000      0.000      0.000      0.000
##   Log-likelihood      1274.382   1307.609   1344.445   1344.512
##   Deviance              19.016     18.242     17.421     17.419
##   AIC                -2542.764  -2607.218  -2678.890  -2677.024
##   BIC                -2526.633  -2585.710  -2652.005  -2644.761
##   N                   1599       1599       1599       1599
## ================================================================
```

As one may notice from the result, the fitted model did not perform well due to the low amount of R-squered. Therefor, in order to investigate about the effect of factors for having a high quality red wine, I use the Lasso for feauture slections.

To do so, I create matrix x for my training set and matrix y for the out put. I use the logarithmic dataset that I made in the Bivariate section.
Moreover, in order to have all the values in same range, I normalized all features with their means and standard deviations, so I can have trustable evaluation from the lasso results for the coefficients.

```
##        rfixed.acidity     rvolatile.acidity          rcitric.acid
##           0.012530257          -0.160006211           0.000000000
##       rresidual.sugar            rchlorides  rfree.sulfur.dioxide
##           0.005781414          -0.083096176           0.066986211
## rtotal.sulfur.dioxide            rsulphates              alcohol1
```

```
##          -0.106473315           0.168845134           0.290937023
##                   pH1
##          -0.063366647
```

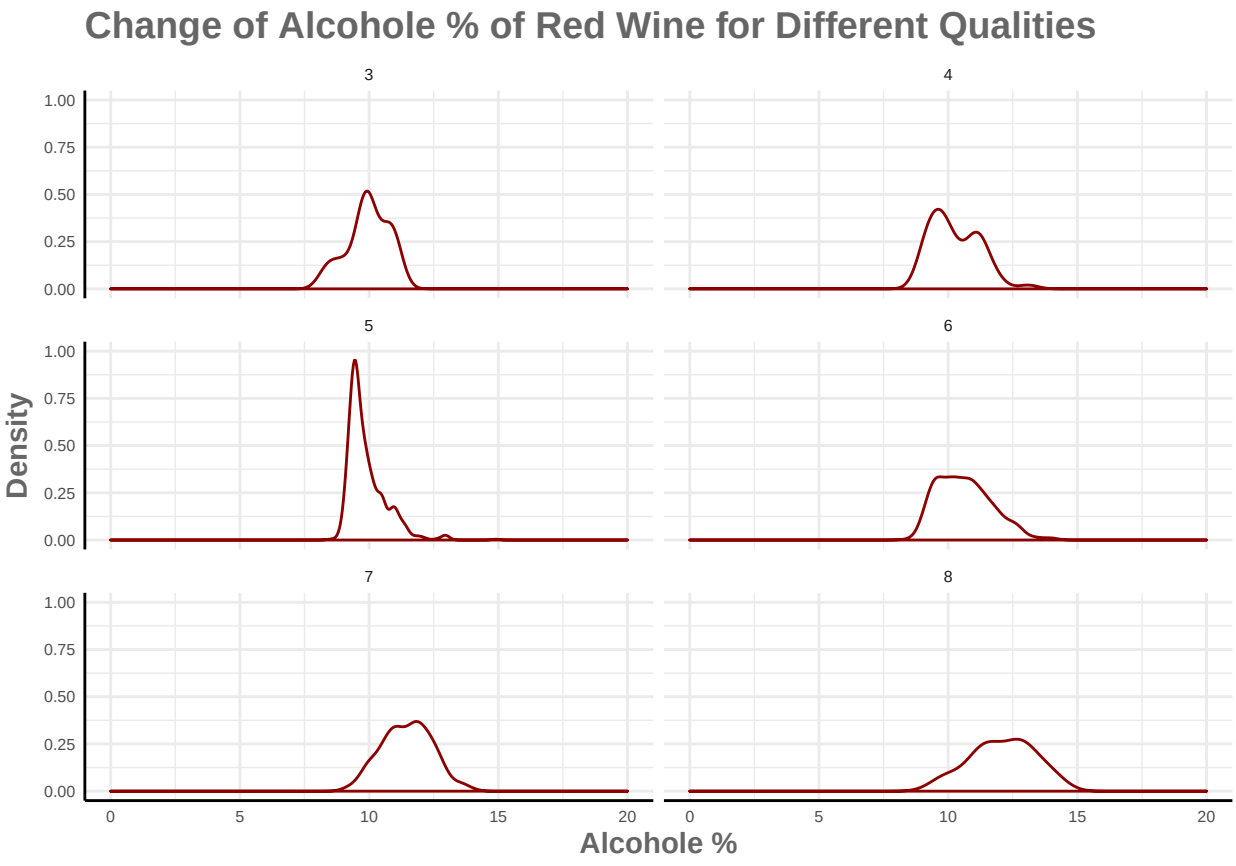
The given results from lasso show that the citric acid may not have enough power to affect on red wine quality. On the other hand, alcohol has high impact between other factors.

until now, I have used linear regression for determining how the acidity and pH are related to each other in red wine ingredient. However, the R-square results showed that the model have had poor performance.

Therefor, I assume that the different features may have their impact for having good red wine quality. So I used Lasso regression to do feature selection systematically. The given results from Lasso, make sence, since the volatile acidity feature (acetic acid) which is an unpleasent factor (at too high of levels) has the most negative coefficients and alcohol which is the high impact in alcoholic beverage, has the most positive coefficients.

Final Plots and Summary

Plot One

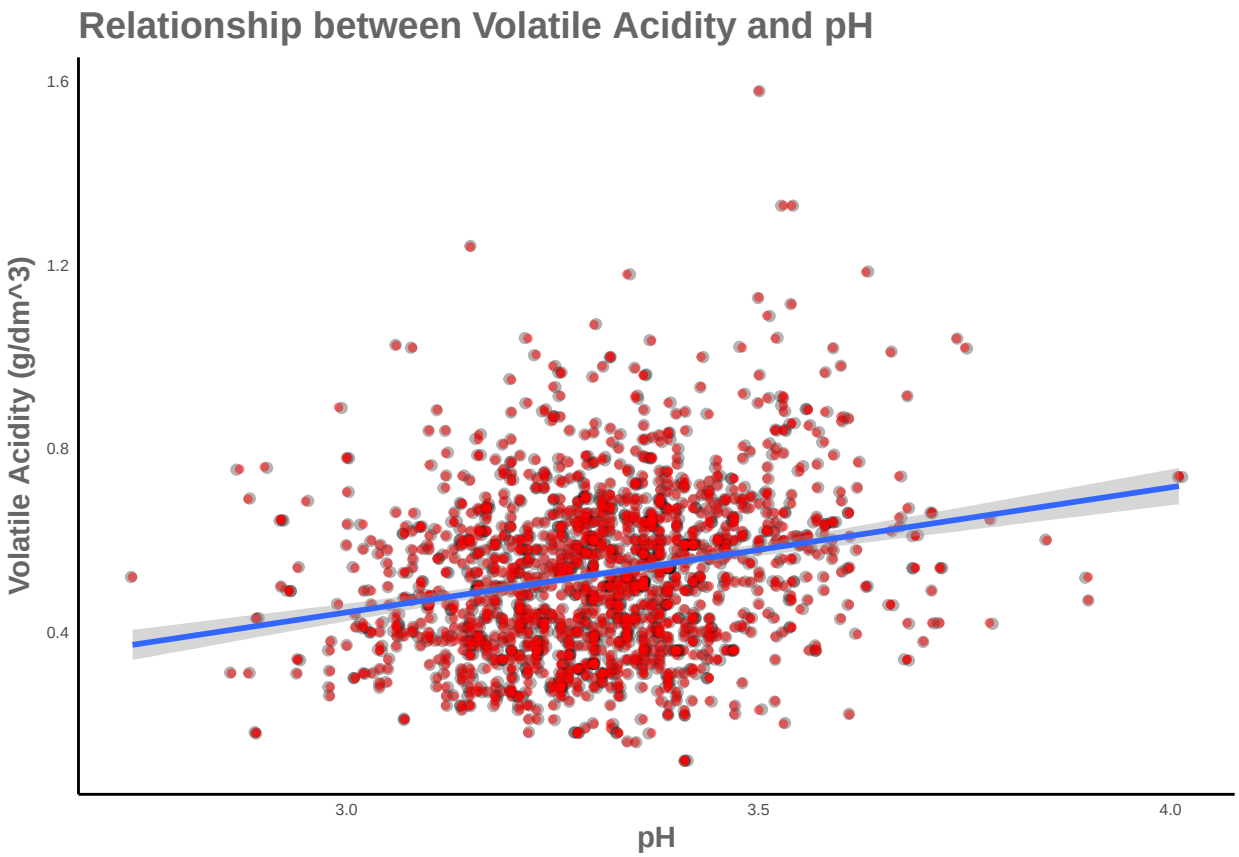


Description One

The plots very well depict how alcohol changes in different red wine quality. The peak of plot moves from almost positive skewed distribution in low quality to fairly normal distributin with smoth peak in high quality

red wine.

Plot Two



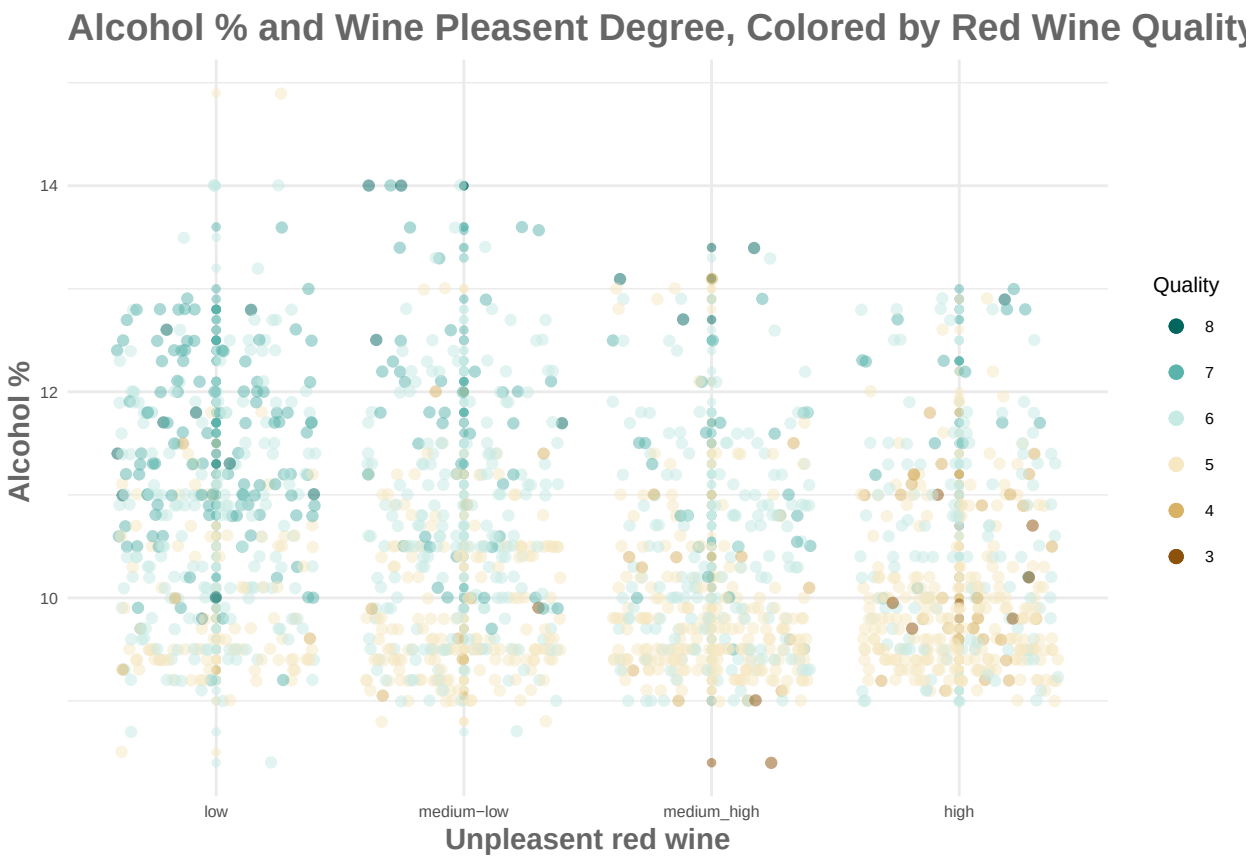
The relationship of pH and volatile acidity:

```
## [1] 0.2349373
```

Description Two

The interesting thing about this plot is that although we know that higher acidity makes less pH, we notice that by increasing acetic acid in red wine, the pH also increases.

Plot Three



Description Three

This plot indicates that in low amount of acetic acid which leads to pleasent wine, we have better quality (mostly, 8,7,6).

While, with increasing of acetic acid where we have more unpleasentness, the quality decreases, and in high level the plot does not show any high quality wine. This plot greatly shows another evidence for what we expected based on the feature selection with our Lasso model. ——

Reflection

The red wine dataset contains information for 1599 observation of twelve red wine ingredients and its quality. I started by understanding of each variables in the dataset and exploring the summary of each variable. Then I developed some questions and started to answer them by obseving various plots. Based on the plots, I decided to consider the logarithem of the ratio of ingredients over the density for the ingredients. Finally, I use linear regresion to investigate the effect of differen acids on pH of the red wine. However, due to the finding of low R-squared for the linear model, I went through another method. The second method was lasse. I implemented this model for feature selection. This time the results completely match with the plots as I explained in the following.

Feature selection model (Lasso) explained how all input variables are relevant. Also the plots clearly showed these relevancy such as the trends between volatile acidity (citric acid) and alcohol with the quality of red wine. The amount of citric acid in the red wine has an important factor which can lead to either an unpleasant

or pleasant taste of red wine. On the other hand, I was suprised that the correlation between pH and citric acid was positive which baced on chemistry rule it has to be the negative one.

On the other hand the dataset has some limitations including the size of the dataste. I believe that 1599 sample data is not enough for having good predction and also overcomming with high variance problem for the model prediction. Furthermore, the data is for 2009, and due to improvement of technology in labratory meaurment devices, it would be much better having the data after 2009 which may be more accurate than this dataset and leads to have better feature selection and precise prediction as well. Moreover, I would interested to use support vector machine in order to predict the quality of the red wine.