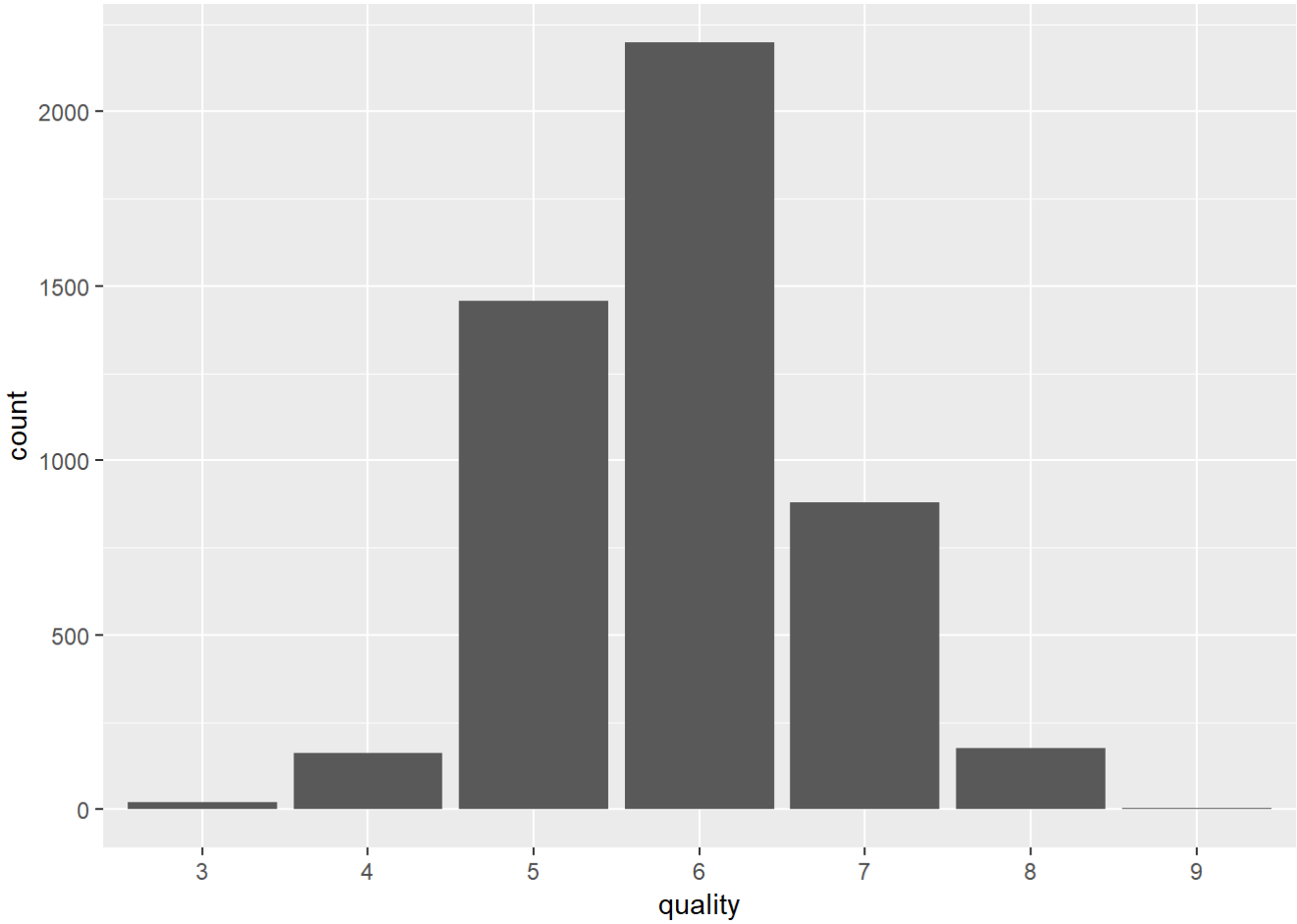
White Wine Exploration by Tianxing Zhai

Univariate Analysis

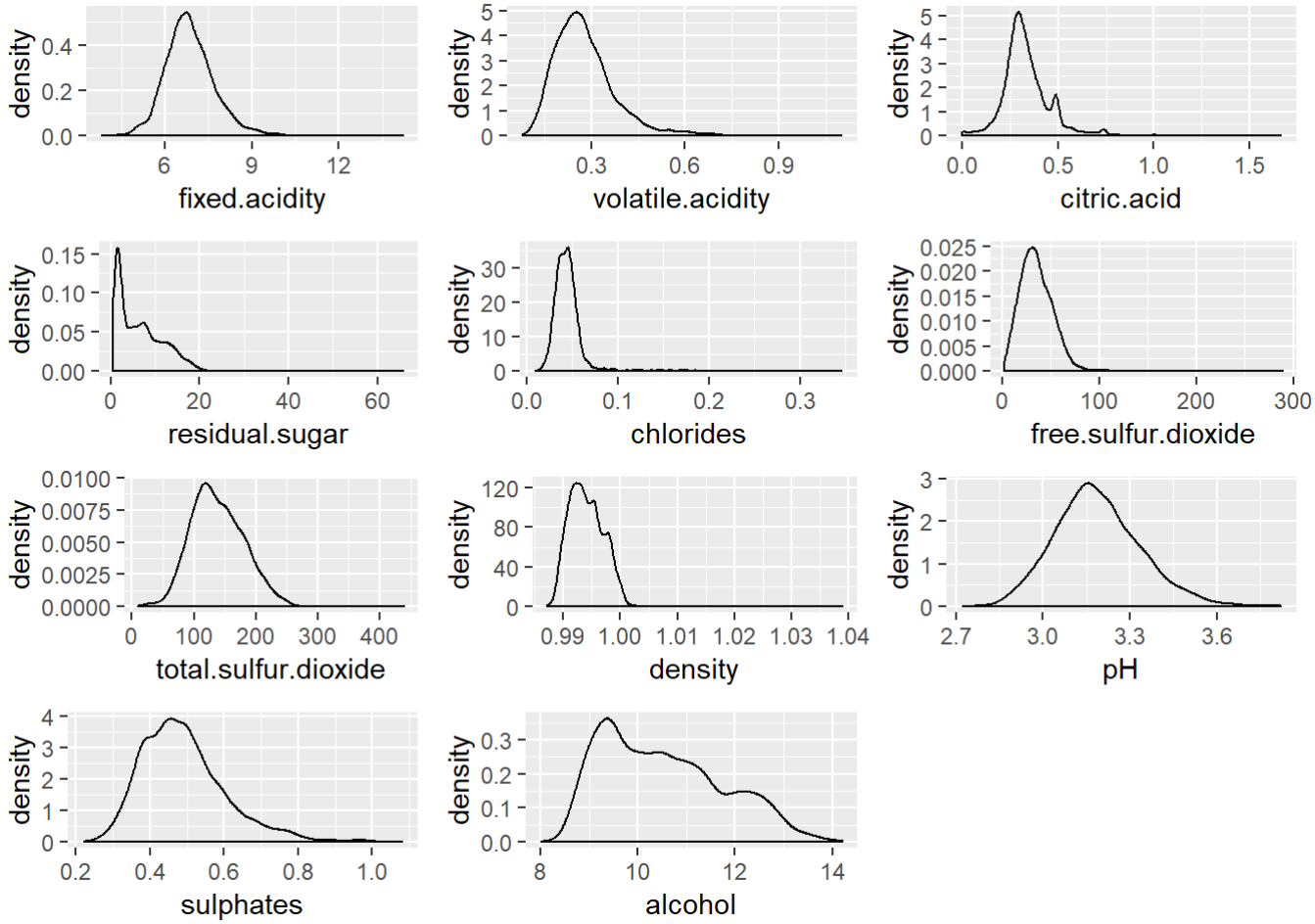
```
## 'data.frame':    4898 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : Ord.factor w/ 7 levels "3"<"4"<"5"<"6"<..: 4 4 4 4 4 4 4 4 4 4 ...
```

```
##        X        fixed.acidity    volatile.acidity  citric.acid
##  Min.   :   1   Min.   : 3.800   Min.   :0.0800   Min.   :0.0000
##  1st Qu.:1225   1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700
##  Median :2450   Median : 6.800   Median :0.2600   Median :0.3200
##  Mean   :2450   Mean   : 6.855   Mean   :0.2782   Mean   :0.3342
##  3rd Qu.:3674   3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900
##  Max.   :4898   Max.   :14.200   Max.   :1.1000   Max.   :1.6600
##
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.600   Min.   :0.00900   Min.   :  2.00
##  1st Qu.: 1.700   1st Qu.:0.03600   1st Qu.: 23.00
##  Median : 5.200   Median :0.04300   Median : 34.00
##  Mean   : 6.391   Mean   :0.04577   Mean   : 35.31
##  3rd Qu.: 9.900   3rd Qu.:0.05000   3rd Qu.: 46.00
##  Max.   :65.800   Max.   :0.34600   Max.   :289.00
##
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  9.0        Min.   :0.9871   Min.   :2.720   Min.   :0.2200
##  1st Qu.:108.0        1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100
##  Median :134.0        Median :0.9937   Median :3.180   Median :0.4700
##  Mean   :138.4        Mean   :0.9940   Mean   :3.188   Mean   :0.4898
##  3rd Qu.:167.0        3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500
##  Max.   :440.0        Max.   :1.0390   Max.   :3.820   Max.   :1.0800
##
##     alcohol       quality
##  Min.   : 8.00   3:  20
##  1st Qu.: 9.50   4: 163
##  Median :10.40   5:1457
##  Mean   :10.51   6:2198
##  3rd Qu.:11.40   7: 880
##  Max.   :14.20   8: 175
##                  9:   5
```

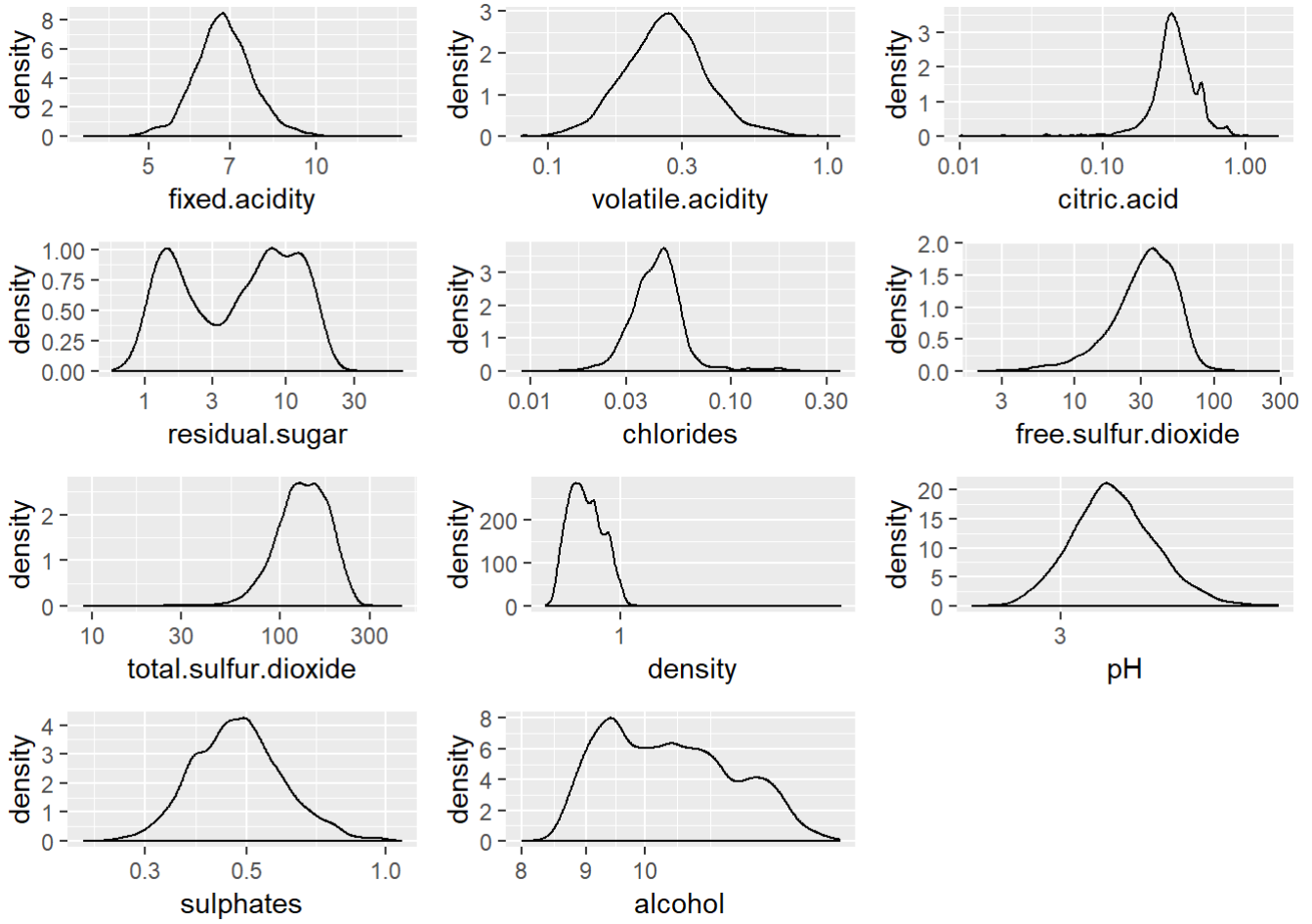
There are 4898 observations (whitewines) in the dataset with 13 variables as listed above. The ‘x’ variable is id of whitewines. The ‘quality’ variable is the quality of whitewines, which scores between 0 (worst) and 10 (best). All other variables are continuous variables which stand for chemical concentration or chemical features of whitewines (I will call these variables ‘chemical variables’).



The main feature in the data set is quality. All other variables (except ‘x’, the id) are measured for finding relation with quality. As we can see, most whitewines score 6 and fewest whitewines have scores of 3-4 (worst) and 8-9 (best).



The distributions of all chemical variables, except alcohol, are bell-shape with positive skew. So it it better to analyse their medians rather than means. Then I use log tranformation to normalize:



That looks more normal.

What is the structure of your dataset?

There are 4898 observations (whitewines) in the dataset with 13 variables as listed above. The ‘x’ variable is id of whitewines. The ‘quality’ variable is the quality of whitewines, which scores between 0 (worst) and 10 (best). All other variables are continuous variables which stand for chemical concentration or chemical features of whitewines (I will call these variables ‘chemical variables’).

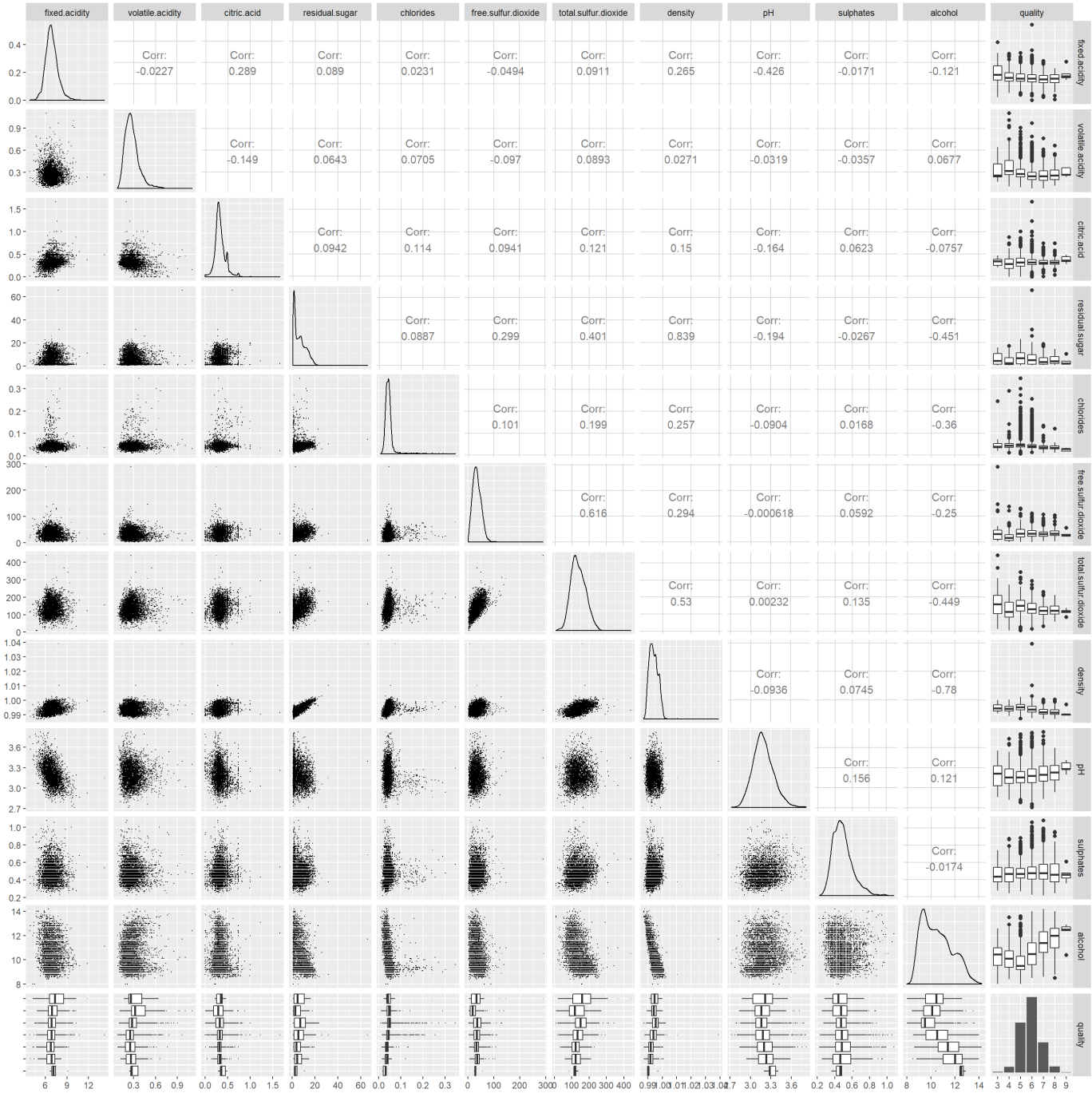
What is/are the main feature(s) of interest in your dataset?

The main feature in the data set is quality. All other variables (except ‘x’, the id) are measured for finding relation with quality. As we can see, most whitewines score 6 and fewest whitewines have scores of 3-4 (worst) and 8-9 (best).

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

All chemical variables are potential predict factors of quality. I will put most efforts on Bivariate Analysis.

Bivariate Plots Section

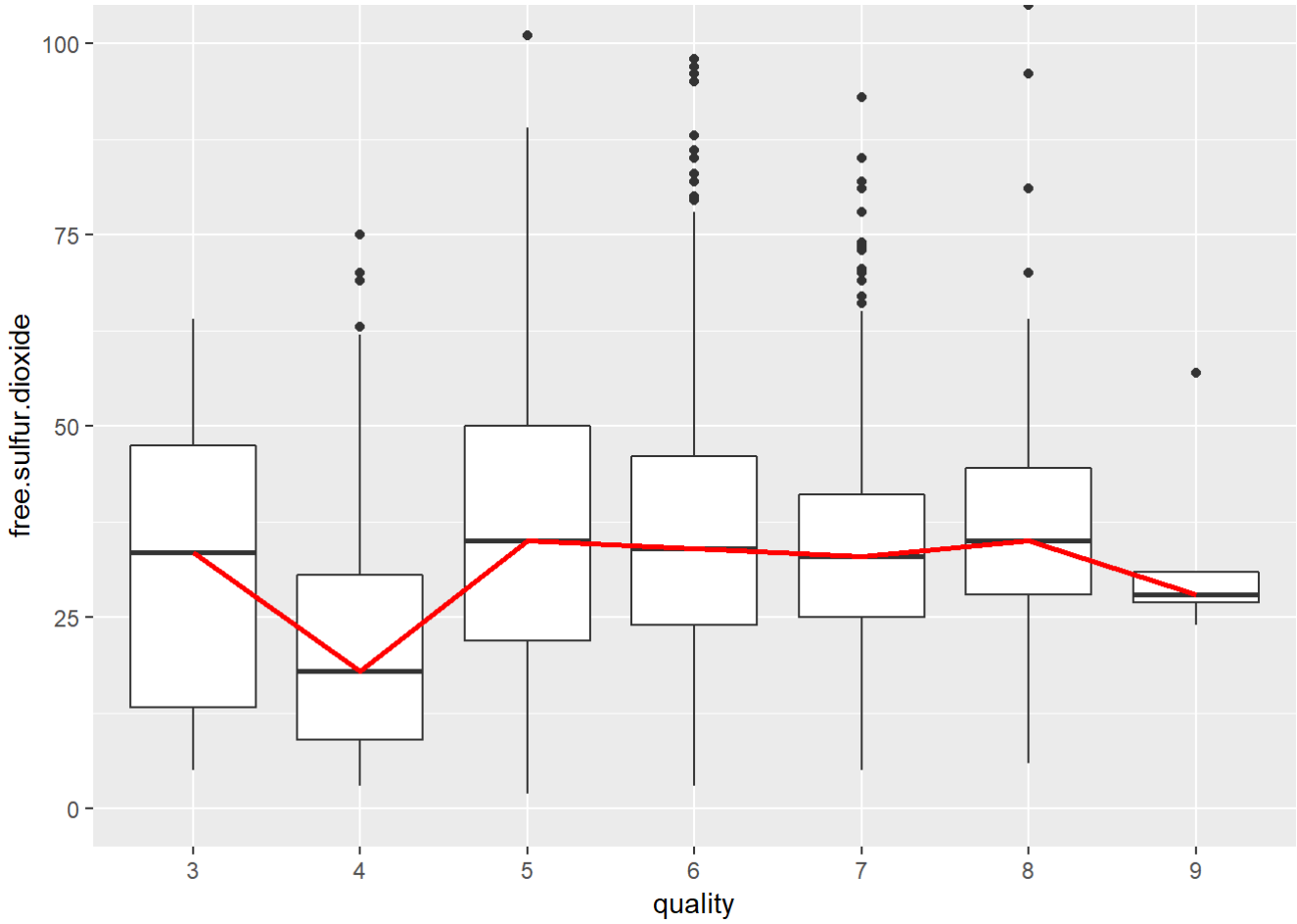


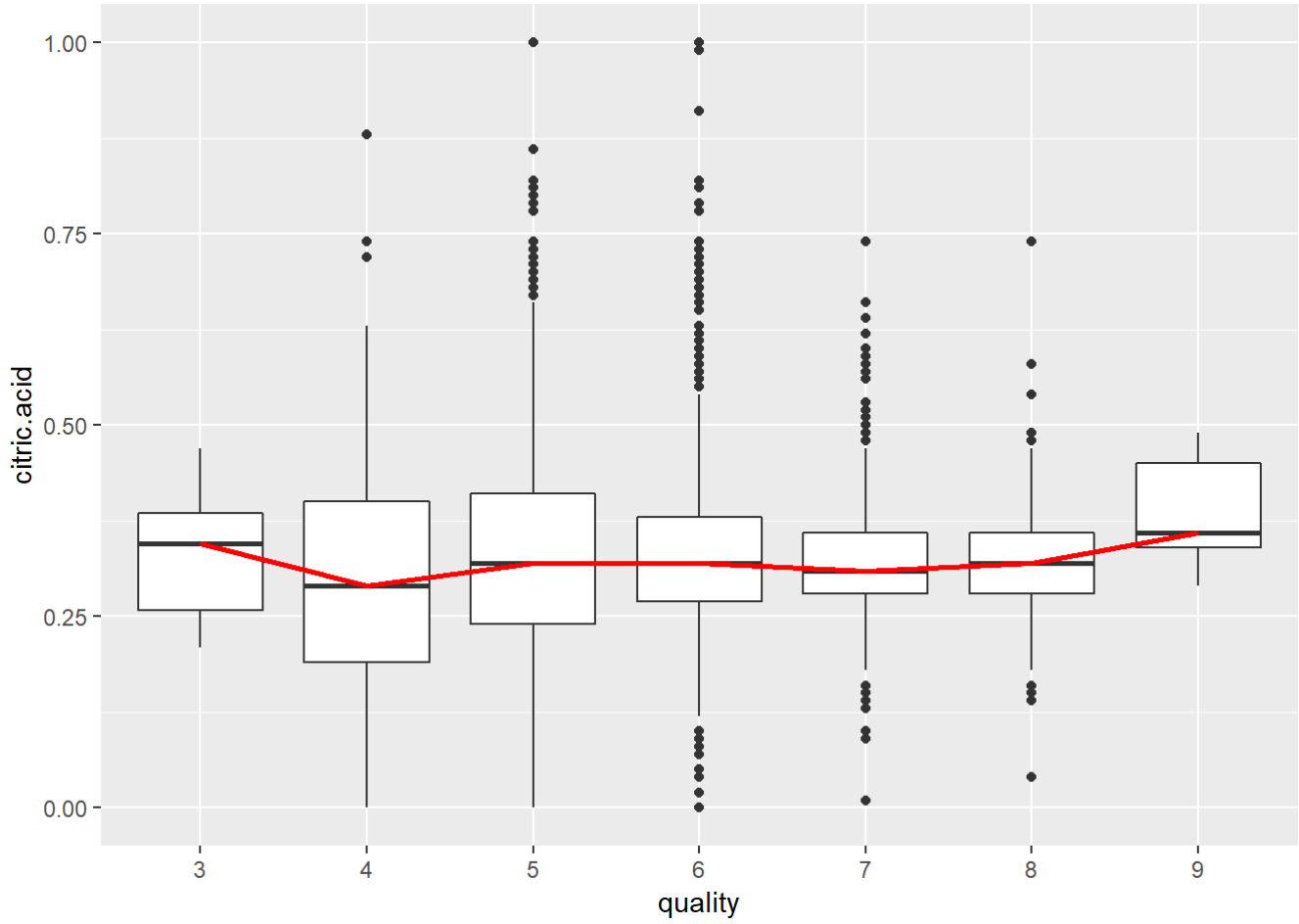
Above is the pair plots of all variables except X. It is hard to find correlation in the plot. We need to run correlation tests and find which variables have corelation with quality:

```
##              chemicals             P           R
## 1              alcohol 0.000000e+00  0.435574716
## 2            chlorides 0.000000e+00 -0.209934411
## 3          citric.acid 5.193459e-01 -0.009209091
## 4              density 0.000000e+00 -0.307123312
## 5        fixed.acidity 1.332268e-15 -0.113662831
## 6  free.sulfur.dioxide 5.681271e-01  0.008158067
## 7                   pH 3.080647e-12  0.099427246
## 8       residual.sugar 7.724044e-12 -0.097576829
## 9            sulphates 1.709792e-04  0.053677877
## 10 total.sulfur.dioxide 0.000000e+00 -0.174737218
## 11    volatile.acidity 0.000000e+00 -0.194722969
```

```
##              chemicals         P            R
## 3          citric.acid 0.5193459 -0.009209091
## 6 free.sulfur.dioxide 0.5681271  0.008158067
```

'free.sulfur.dioxide' and 'citric.acid' do not have significant correlation (p > 0.05) with quality. Let's make boxplots to see detailly: (the red line in boxplots represent the change of median, similarly hereinafter)



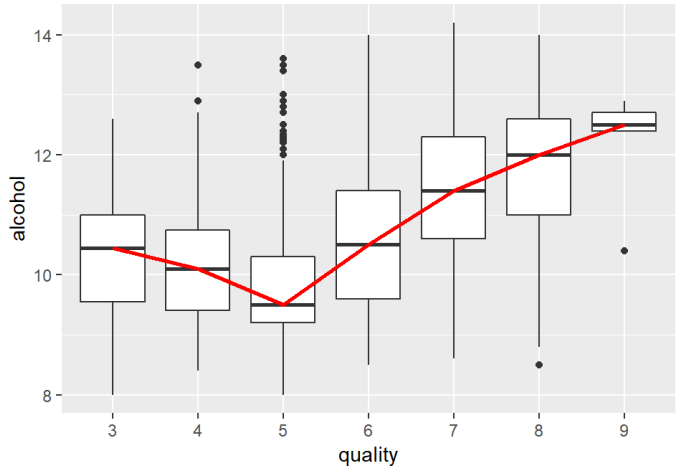


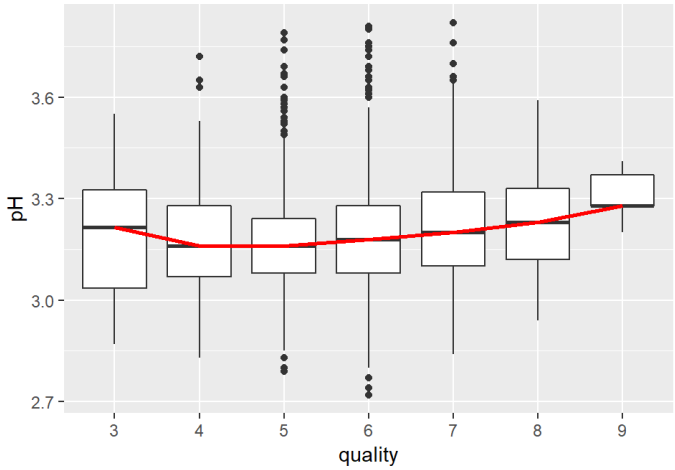
There no obvious patterns above. So I will exclude them from potential predict factors.

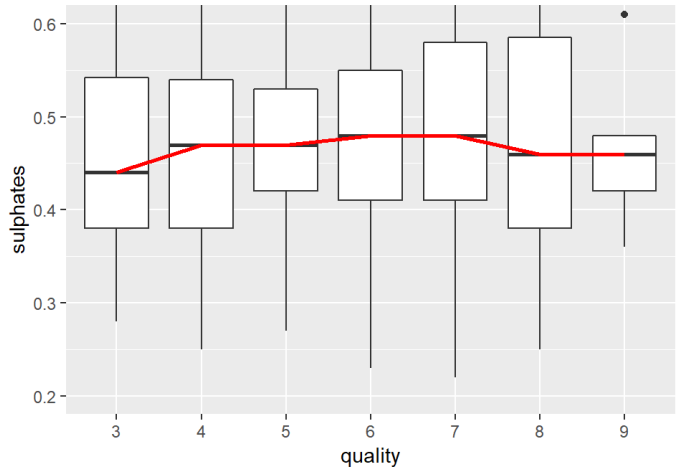
Among remaining 9 variables, 3 of them positively correlated with quality and 6 negatively correlated with quality.

```
##   chemicals            P          R
## 1   alcohol 0.000000e+00 0.43557472
## 7        pH 3.080647e-12 0.09942725
## 9 sulphates 1.709792e-04 0.05367788
```

'Alcohol', 'pH' and 'sulphates' positively correlated with quality. Similarly, I made boxplots to see detailly.



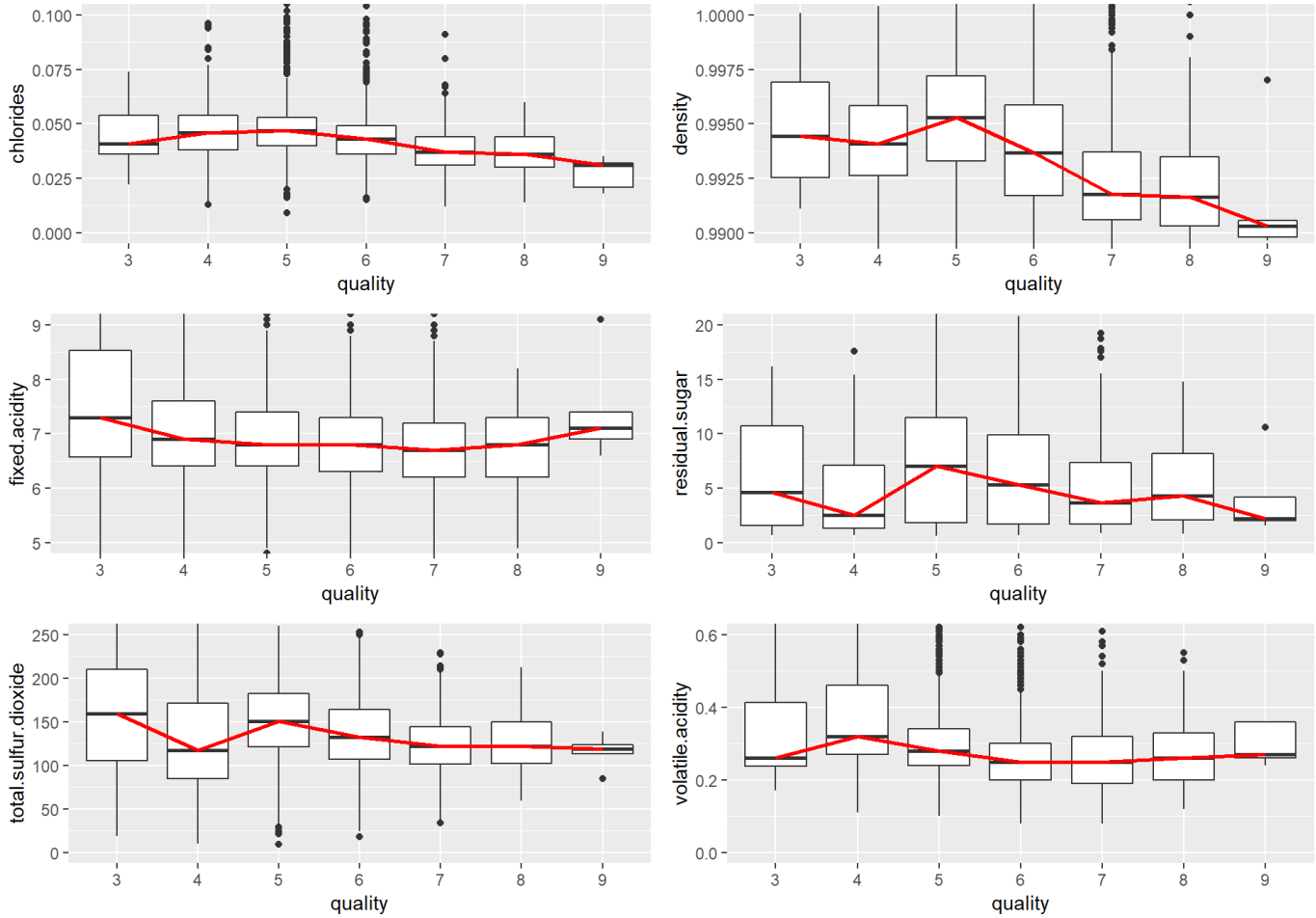




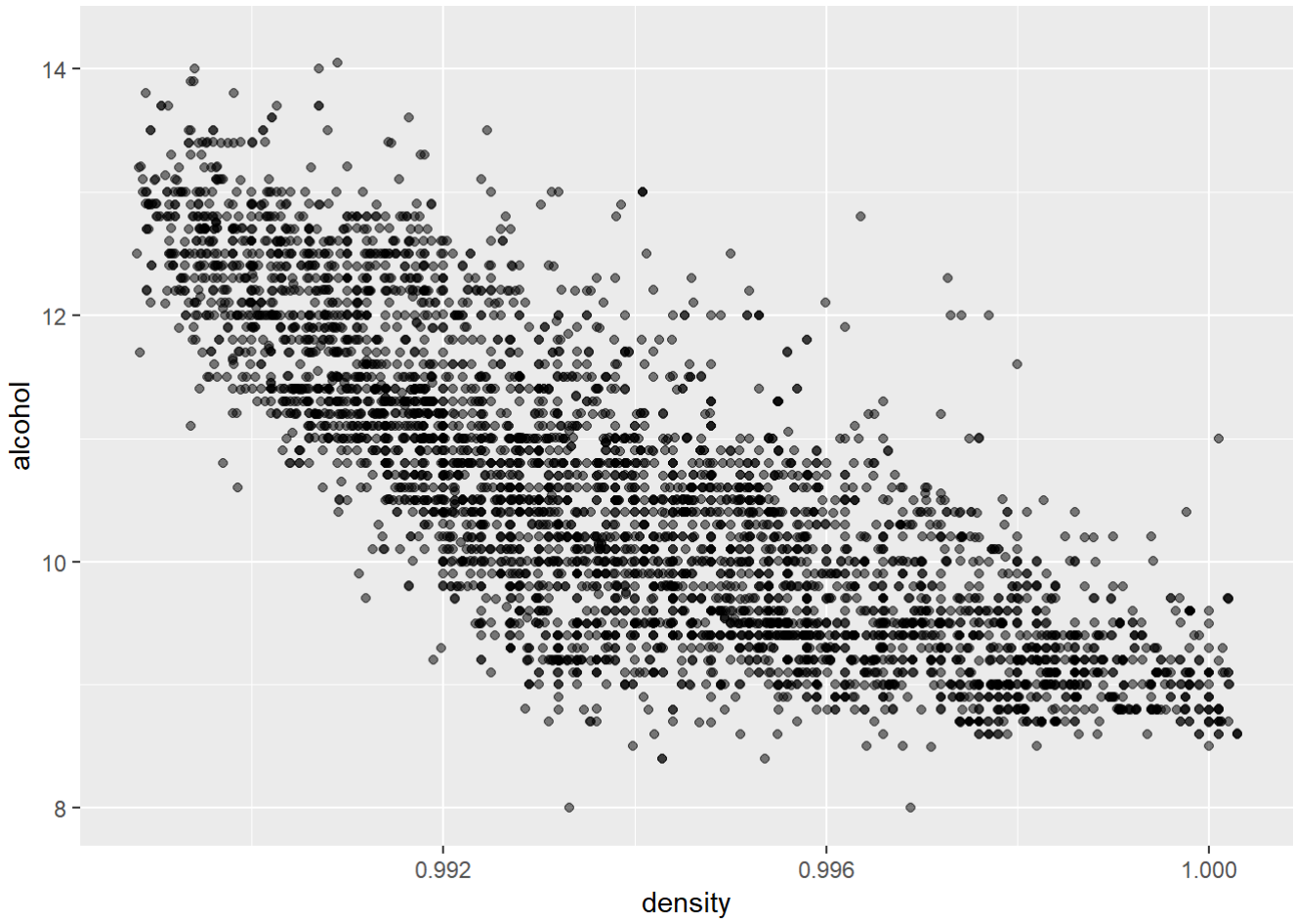
We can see apparent increase of median alcohol content from that of quality 5 to quality 9. So alcohol may be a strong predictor of quality. The increase of median pH from quality 4 to quality 9 is mild, but at least it is monotonous. As to sulphates concentration, there are neither obvious difference between quality groups, nor monotonous increase or decrease trends. Considering the R(0.05367788) is too small, I will exclude sulphates concentration from predict factors.

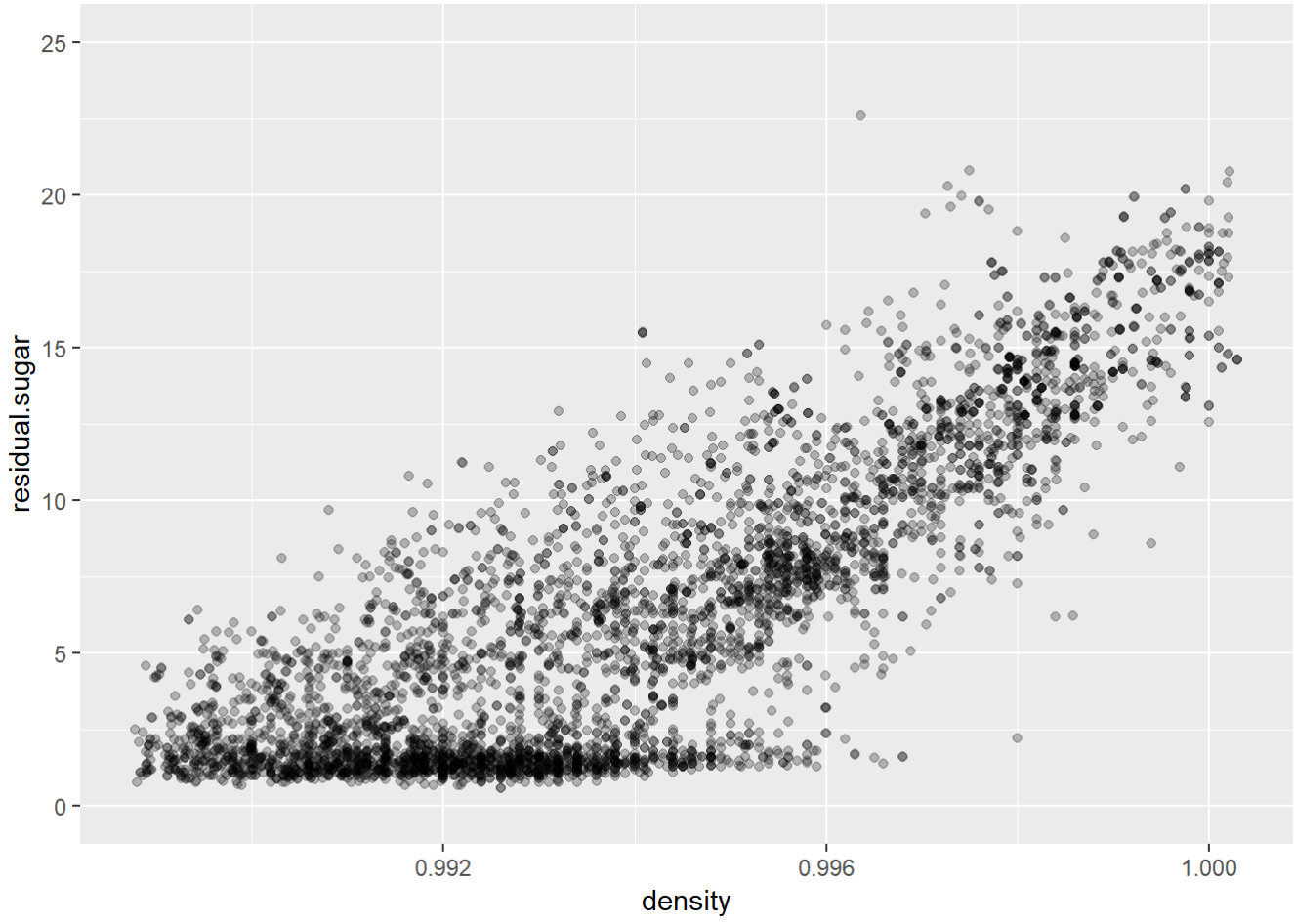
Here are potential negative predictors and their boxplots:

```
##               chemicals            P          R
## 2             chlorides 0.000000e+00 -0.20993441
## 4               density 0.000000e+00 -0.30712331
## 5         fixed.acidity 1.332268e-15 -0.11366283
## 8      residual.sugar 7.724044e-12 -0.09757683
## 10 total.sulfur.dioxide 0.000000e+00 -0.17473722
## 11    volatile.acidity 0.000000e+00 -0.19472297
```

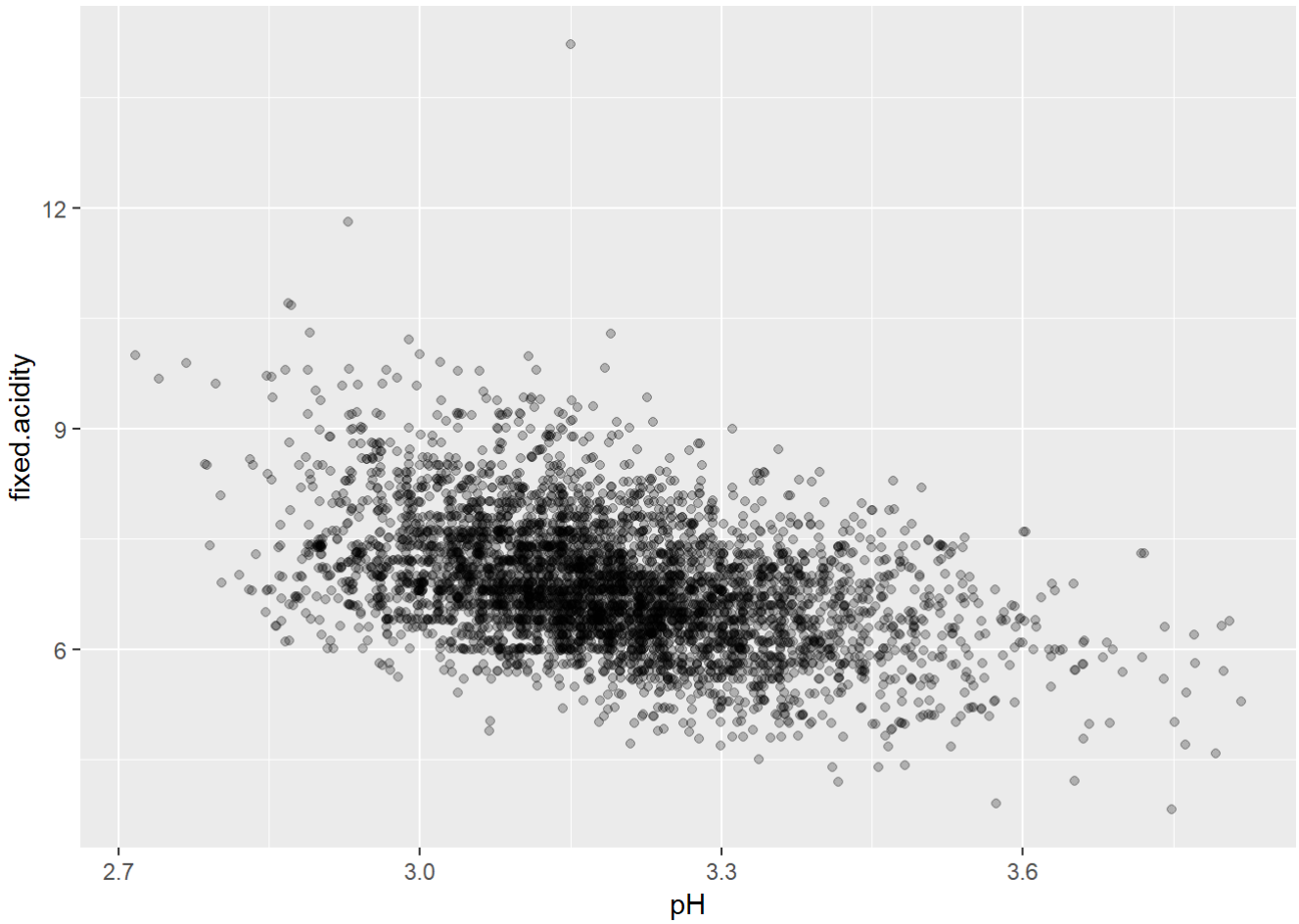


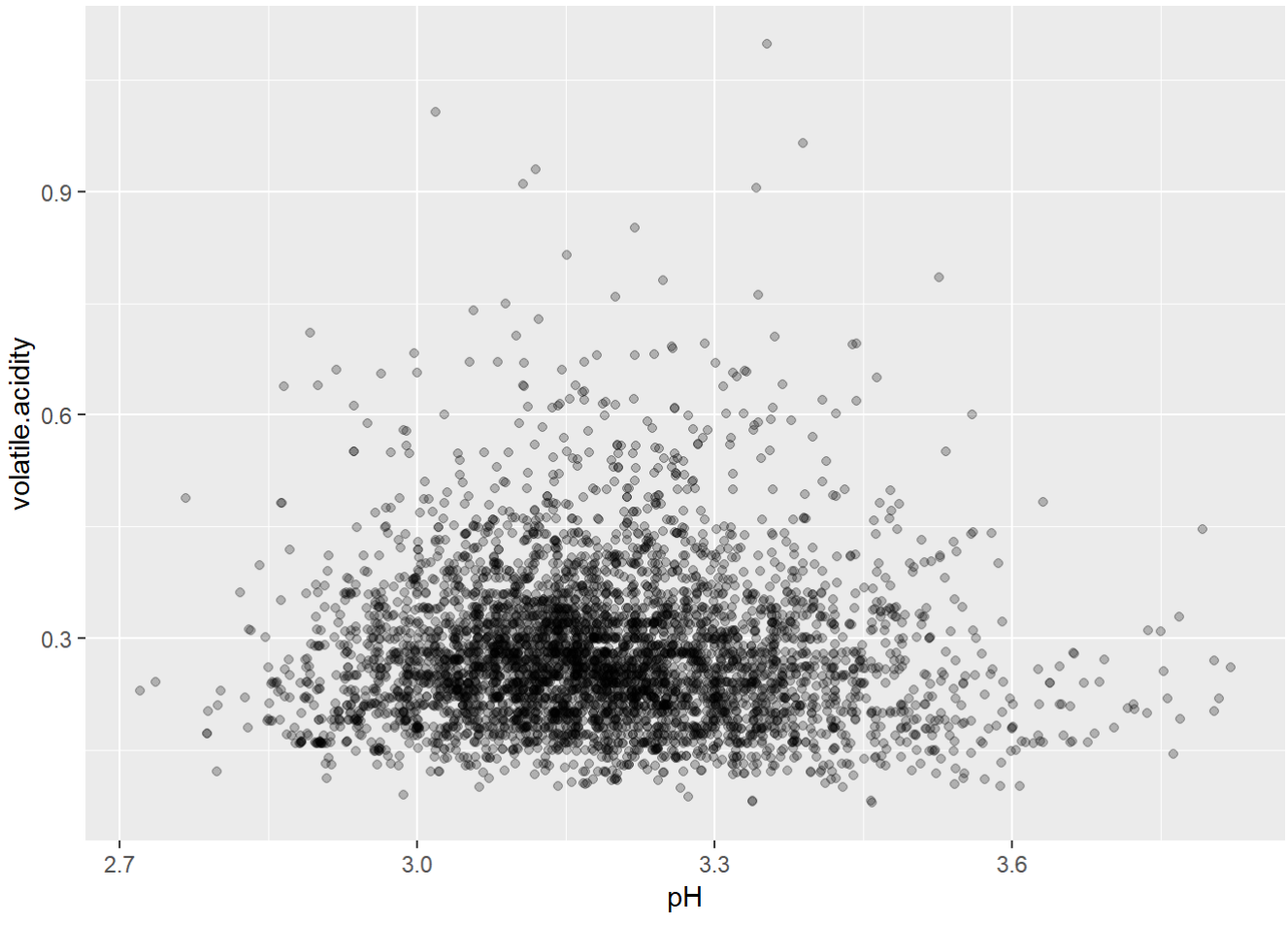
Using the same criteria as above, we can conclude that: Density may be strong predict factors. Chlorides and residual sugar may be weak predict factors. Fixed acidity, total sulfur dioxide and volatile acidity may not be predict factors.



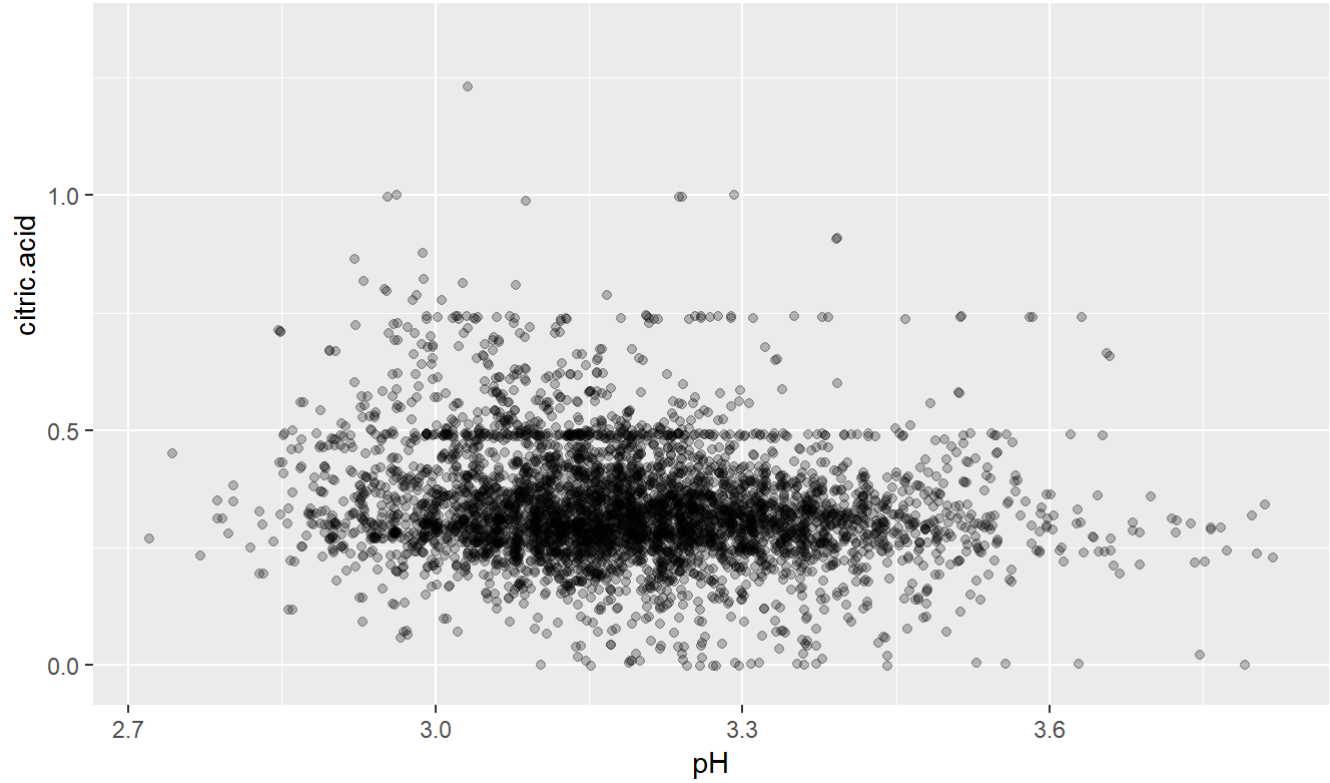


Density has strong corelation with alcohol and residual sugar. The effect of density on quality may be confounded by alcohol or residual sugar.


fixed.acidity
12
9
6
2.7
3.0
3.3
3.6
pH


volatile.acidity
0.9
0.6
0.3
2.7
3.0
3.3
3.6
pH


1.5



Among all three acids: fixed acidity(tartaric acid), volatile acidity(acetic acid) and citric acid, tartaric acid has strongest effect on pH. # Bivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

The alcohol content and density may be strong predict factors for quality. Whitewines with higher alcohol content and lower density may have better quality. Chlorides, residual sugar and pH may be weak predict factors for quality.

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

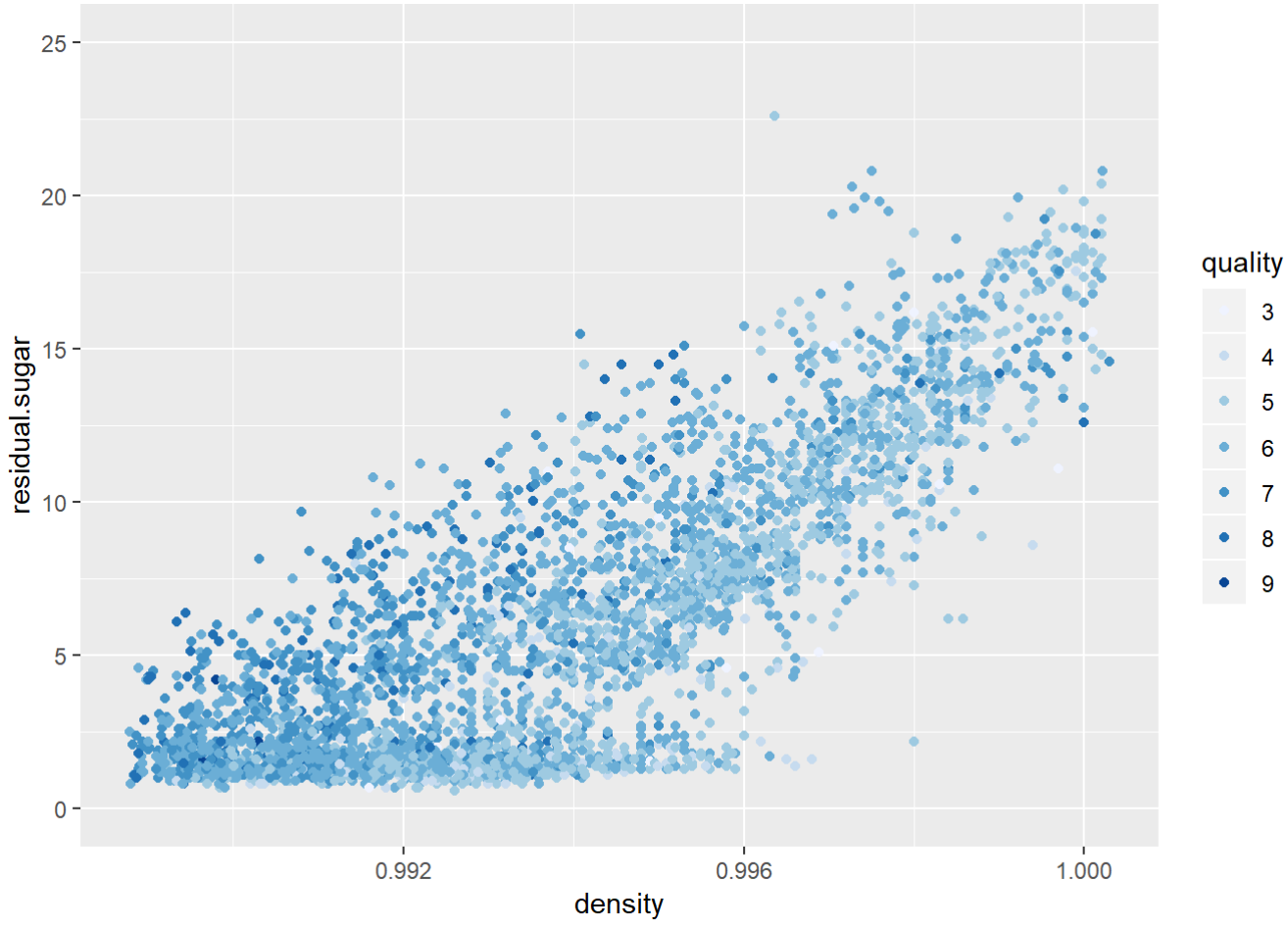
Density has strong corelation with alcohol and residual sugar. The effect of density on quality may be confounded by alcohol or residual sugar. Among all three acids: fixed acidity(tartaric acid), volatile acidity(acetic acid) and citric acid, tartaric acid has strongest effect on pH.

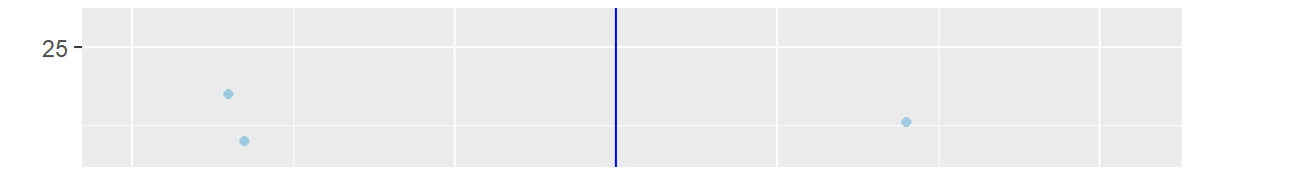
What was the strongest relationship you found?

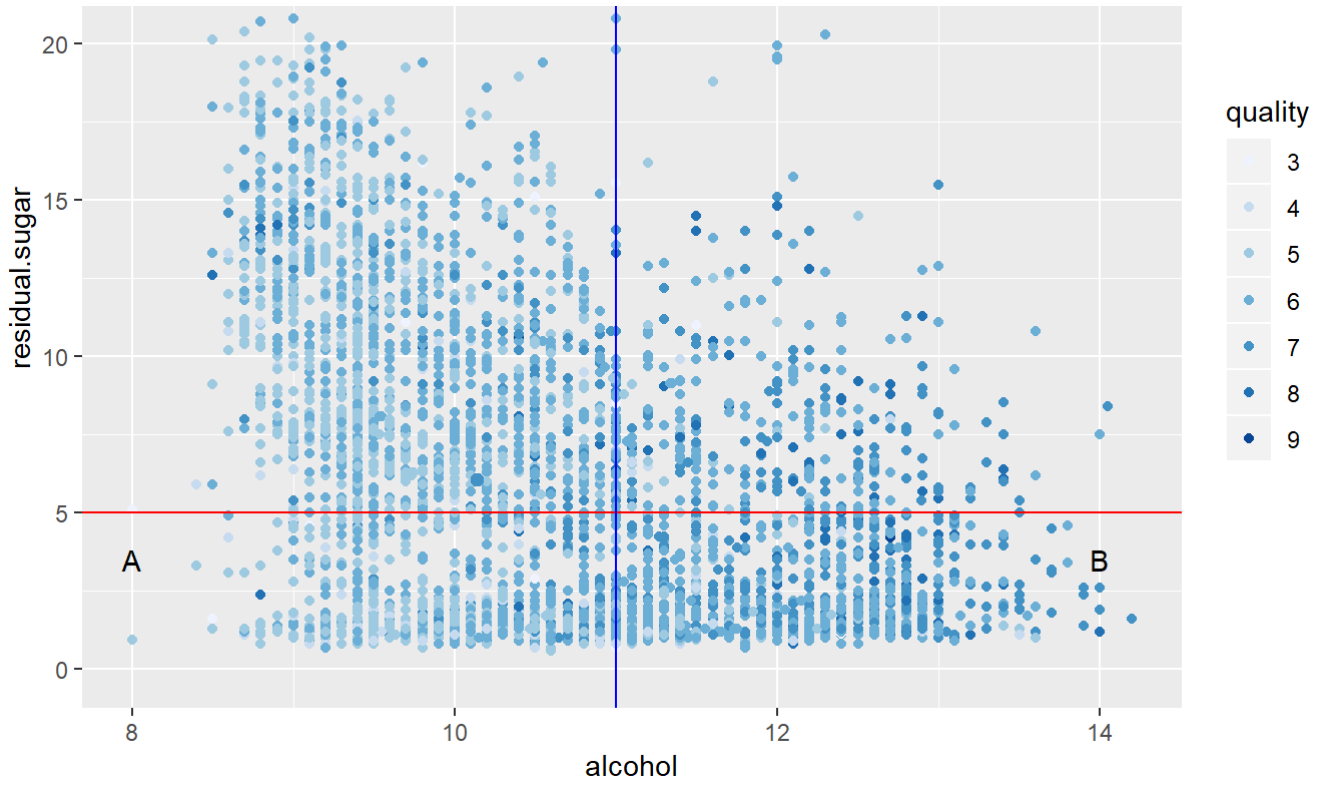
The relationship between density and residual sugar/ alcohol.

Multivariate Plots Section


alcohol
14
12
10
8
0.992
0.996
1.000
density
quality
3
4
5
6
7
8
9

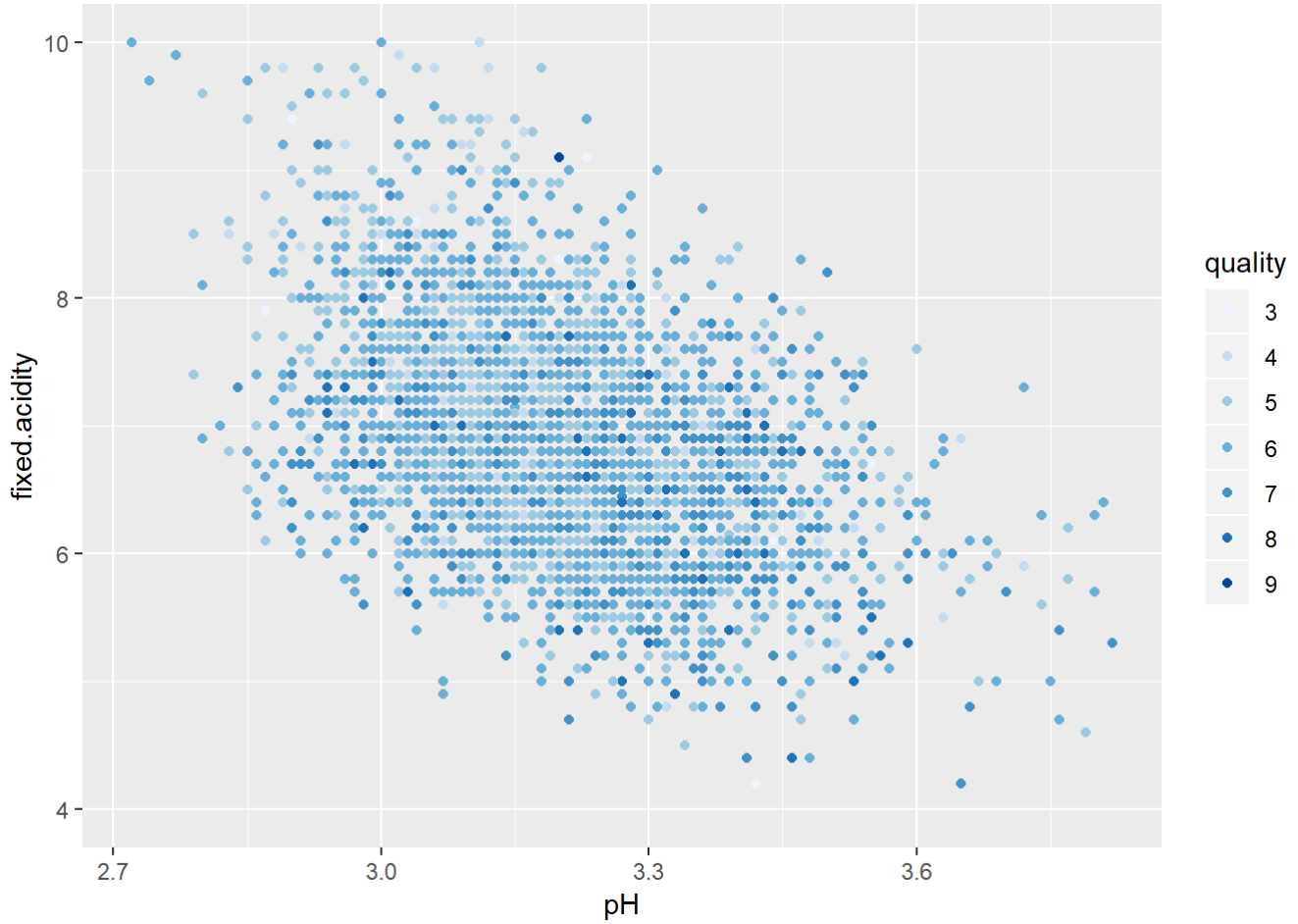

residual.sugar
25
20
15
10
5
0
0.992
0.996
1.000
density
quality
3
4
5
6
7
8
9


25



In the first figure, dots with dark colors(quality 7 and 8) gather in the top left corner. In the second figure, similar pattern is observed as well: dots with dark colors(quality 7 and 8) gather in the bottom left corner. These mean that whitwines with more alcohol, less sugar and lower density tend to have better quality.

I also notice that density mianly depends on the alcohol and sugar content. More alchol and less sugar, lower the density will be. This implies that the effect of density on the prediction of quality is confounded by alcohol and sugar. The effect strength of alcohol is greater than that of sugar because: In figure 3, keep the sugar content in a low interval (< 5). When alcohol level is low, there are few dark dots in that area (A area). However, when alcohol level is high, there are many dark dots in that area (B area).



The pattern here is much more unconspicuous. But we can still recognize that dark dots tend to center in the right. Besides, we can notice that fixed acid do not have effect on the quality, because the dark dots center in the right, not the right top or bottom corners.

```
##
## Calls:
## m1: lm(formula = quality ~ alcohol, data = df_quality.as.numeric)
## m2: lm(formula = quality ~ alcohol + density, data = df_quality.as.numeric)
## m3: lm(formula = quality ~ alcohol + density + pH, data = df_quality.as.numeric)
## m4: lm(formula = quality ~ alcohol + density + pH + chlorides, data = df_quality.as.numeric)
## m5: lm(formula = quality ~ alcohol + density + pH + chlorides + residual.sugar,
##     data = df_quality.as.numeric)
##
## ============================================================================================
##                         m1            m2            m3            m4            m5
## --------------------------------------------------------------------------------------------
##   (Intercept)          0.582***    -24.492***    -25.288***    -23.956***    111.428***
##                       (0.098)       (6.165)       (6.161)       (6.160)       (12.828)
##   alcohol              0.313***      0.360***      0.356***      0.340***      0.203***
##                       (0.009)       (0.015)       (0.015)       (0.015)       (0.019)
##   density                           24.728***     24.688***     23.675***    -112.542***
##                                     (6.079)       (6.072)       (6.067)       (12.846)
##   pH                                              0.276***      0.261***      0.577***
##                                                   (0.076)       (0.076)       (0.079)
##   chlorides                                                     -2.286***     -1.408*
##                                                                 (0.558)       (0.555)
##   residual.sugar                                                              0.065***
##                                                                               (0.005)
## --------------------------------------------------------------------------------------------
##   R-squared             0.190         0.192         0.195         0.197         0.220
##   adj. R-squared        0.190         0.192         0.194         0.197         0.219
##   sigma                 0.797         0.796         0.795         0.794         0.782
##   F                  1146.395       583.290       394.271       300.857       276.407
##   p                     0.000         0.000         0.000         0.000         0.000
##   Log-likelihood    -5839.391     -5831.127     -5824.480     -5816.089     -5745.260
##   Deviance           3112.257      3101.773      3093.365      3082.784      2994.902
##   AIC               11684.782     11670.255     11658.961     11644.177     11504.521
##   BIC               11704.272     11696.241     11691.444     11683.157     11549.997
##   N                  4898          4898          4898          4898          4898
## ============================================================================================
```

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?

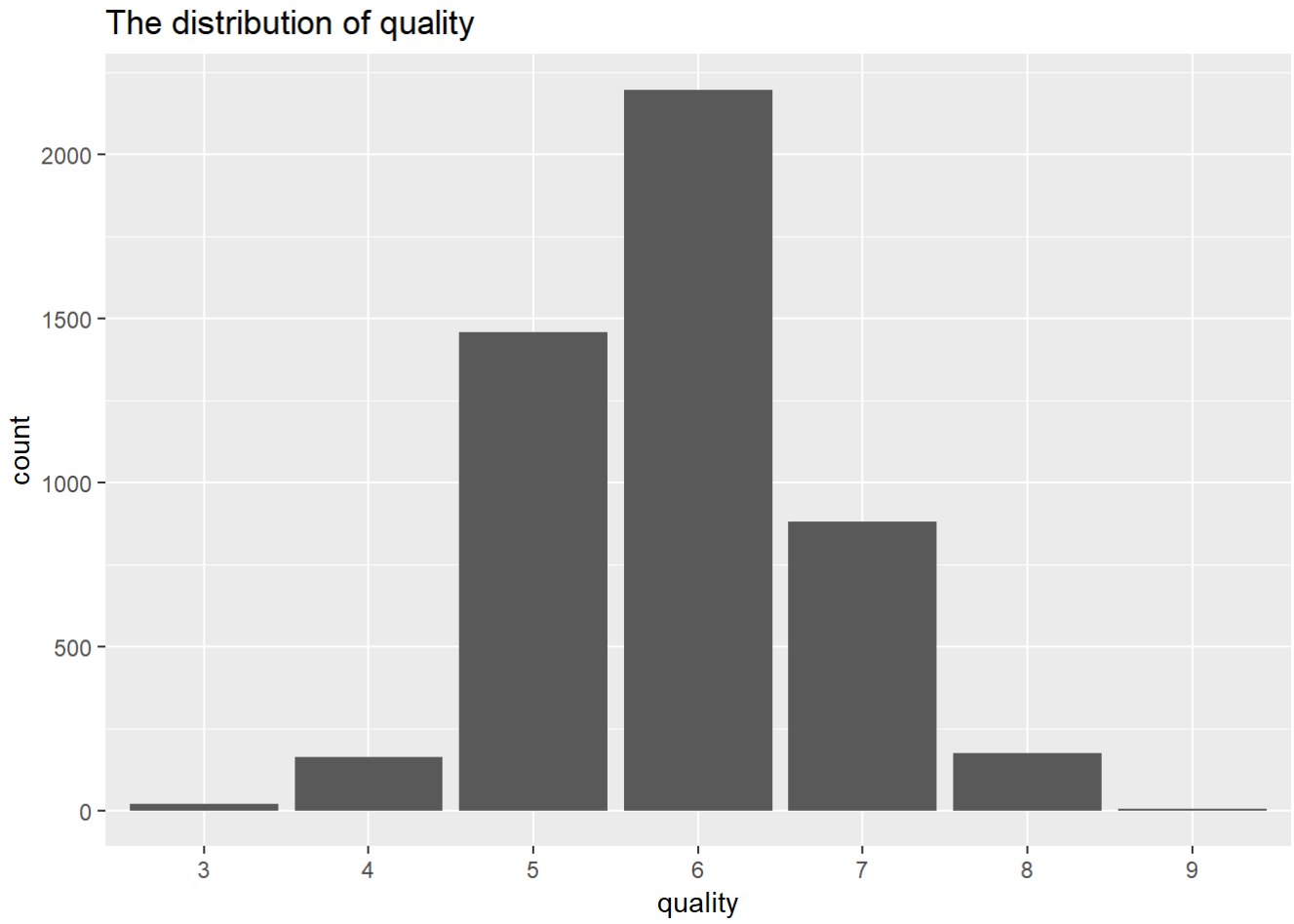
The quality of whitewines is mainly affected by alchol, sugar content and density. But the effect of density on quality is confounded by alcohol and sugar. The confounding strength of alcohol is greater than that of sugar.

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

Yes. I make a simple linear regression model to predict qulity. But strictly speaking, ‘quality’ is not a continous variable. So it is not so good to use linear regression model. The R-square is 0.220, too low for a eligible linear regression model.

Final Plots and Summary

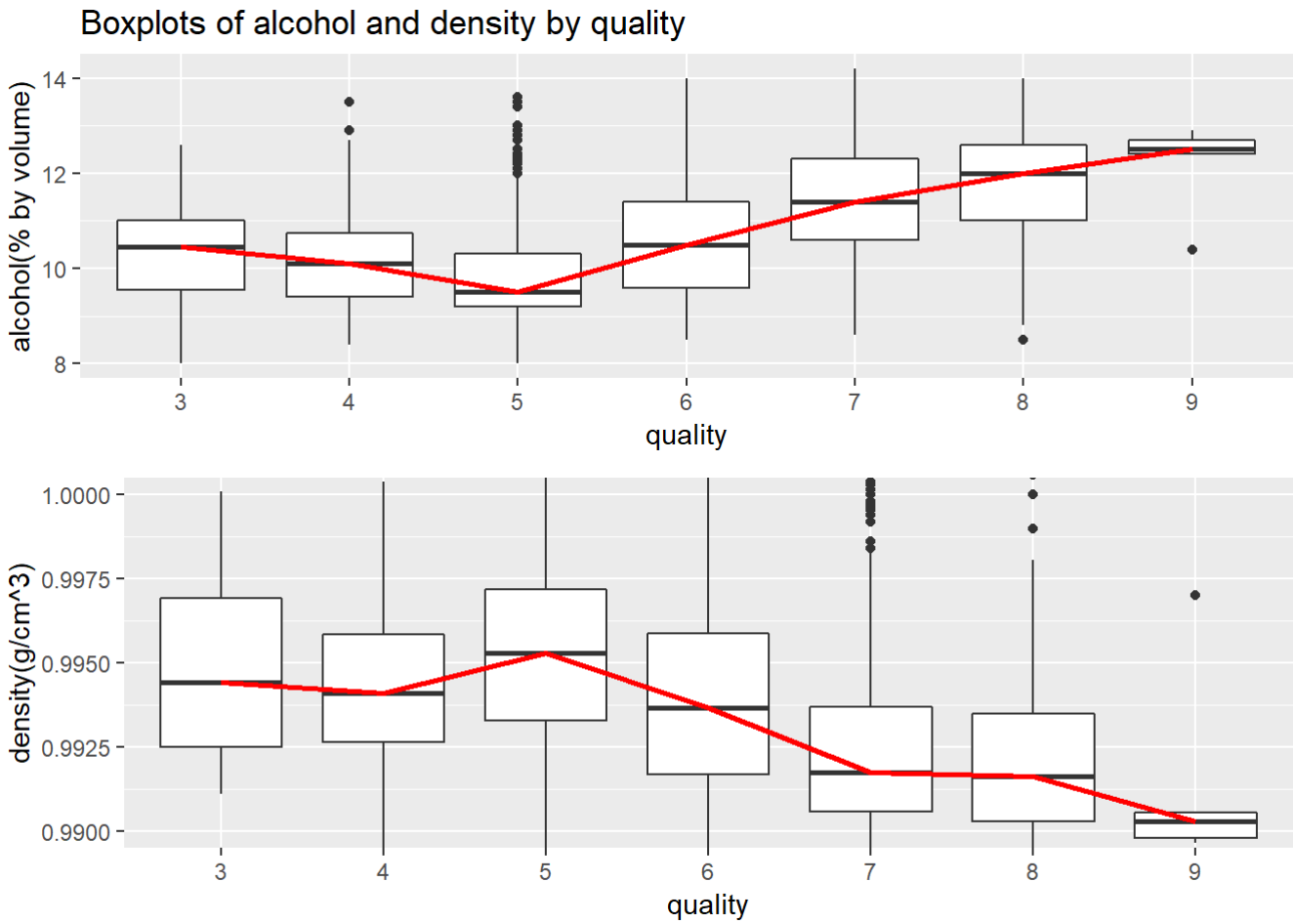
Plot One



Description One

Most whitewines have middle level quality(6). Fewer of them have quality 5 and 7, which are worse and better than quality 6, respectively. Fewest wines have quality 3,4 and 8,9, which mean the worst and best, respectively.

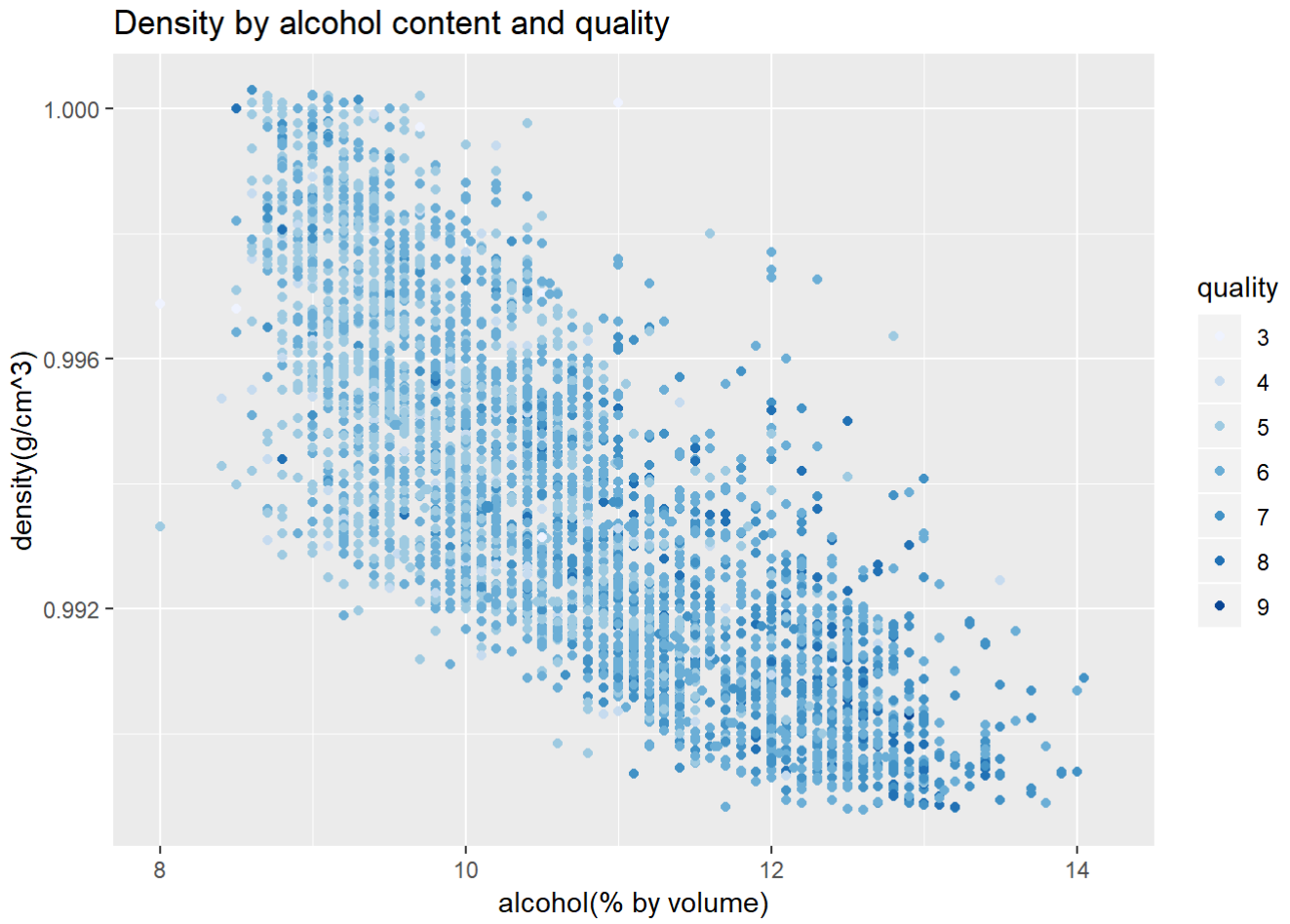
Plot Two



Description Two

Among all 11 chemical variables, alcohol and density are the strongest influence factors of quality. Better wines have higher alcohol content and lower density. Although there is a slight decline of median alcohol content from quality 3 to quality 5, the increase from quality 5 to 9 is much more obvious. The trend is similar for median density: The median starts decreasing from quality 5 to 9, and the trend is also apparent.

Plot Three



Description Three

Alcohol and density are the strongest influence factors of quality. But the only one which actually affects quality is alcohol, not density. Density has a strong linear relation between alcohol content: more alcohol, lower density. The logic chain is: Whitewines with high alcohol content tend to have better quality and lower density. So seemingly, density 'affects' the quality.

Reflection

This data set contains information on 4898 whitewines across 13 variables. I started by understanding the chemical variable and quality in the data set, and then I explored interesting questions and leads as I continued to make observations on plots. Eventually, I explored the quality of whitewines across many chemical variables and created a linear model to predict quality.

There was a clear trend between the alcohol and density of a wine and its quality. I was very surprised that other chemical factors like acidity and sulfur dioxide content did not have a strong correlation with quality. I also found that alcohol and density have strong linear relation. This remiands there may be confounding: Whitewines with high alcohol content tend to have better quality and lower density. So seemingly, density 'affects' the quality. Although the true effector is not density, I still suggest that measuring density can be a rule of thumb for predicting wine quality. Because density have an almost perfect linear relation with the true effector, alcohol. And it is much easier to measure density than alcohol content.

The limitations of my analysis are: 1. I did not quantify the effect of confounding. 2. I treat quality as categorical data, so it is not appropriate to do linear regression on it. And we can see that the final R-square of my model is very low (0.220). Maybe in the future work, after I learn more on statistics, I will try to sovle these two problems.