Exploratory Data Analysis on Red Wine dataset by Trushit Vaishnav

In the following section, we will load the data and try to understand the structure of data.

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

Except quality, all our variables are continous variables. Quality is a discrete variable.

Let as create a new variable Quality2. If quality value is 3 or 4, we consider it as Bad. If quality variable is 5 and 6, we consider it as a Medium. If quality variable is 7 or 8, we consider it as Good.

```
## Warning: package 'bindrcpp' was built under R version 3.4.1
```

```
## [1] "Bad"    "Good"   "Medium"
```

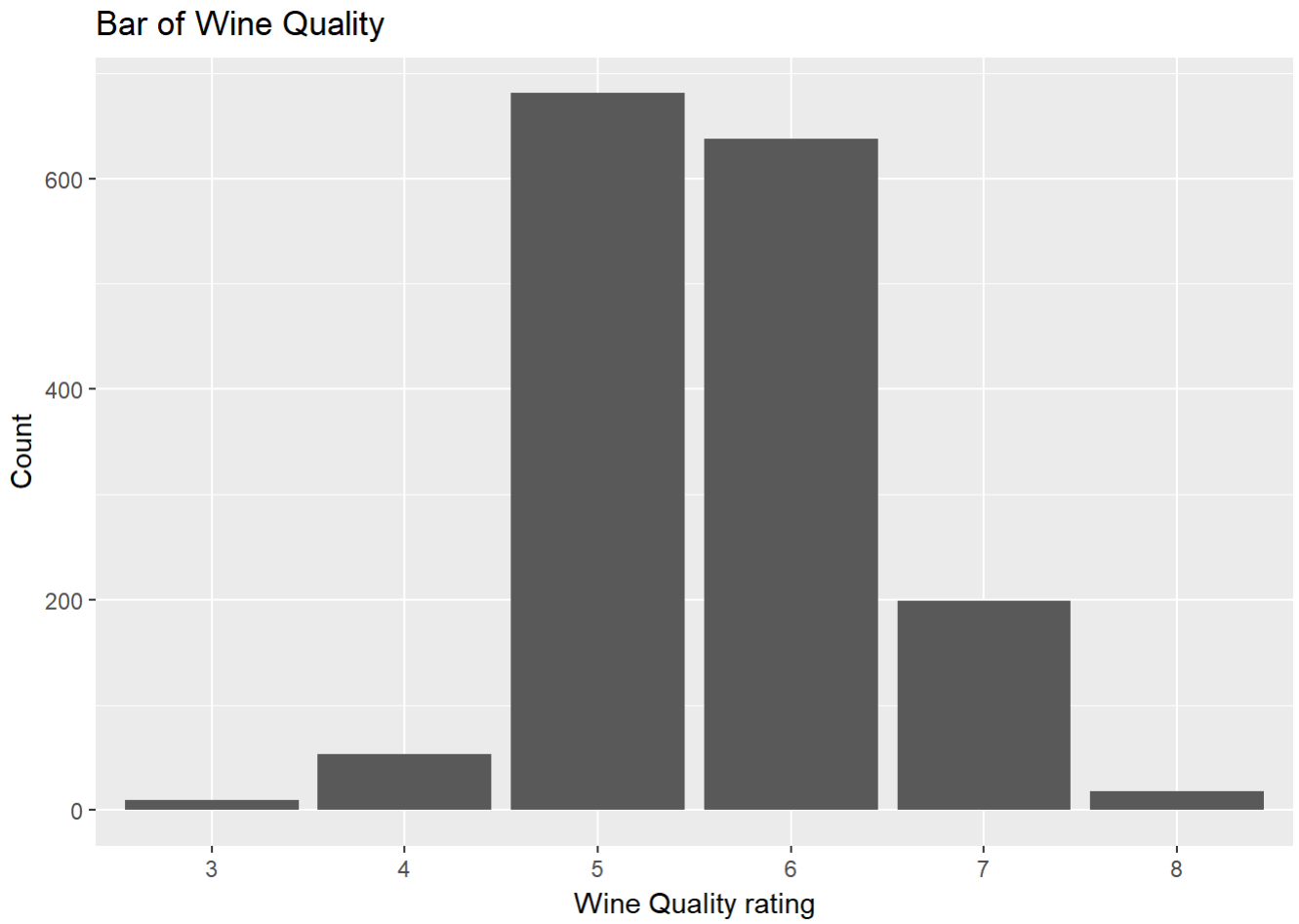
Now we shall perform visual univariate analysis. It is important to note that this is a exploratory analysis and not the explanatory analysis. The idea in this step is to get the idea about distribution of each variable.

Univariate Plots Section

We have used histograms to explore all the variables in the dataset except Wine quality which is an Ordinal variable. Bar plot has been used to visualize quality variable.

Quality

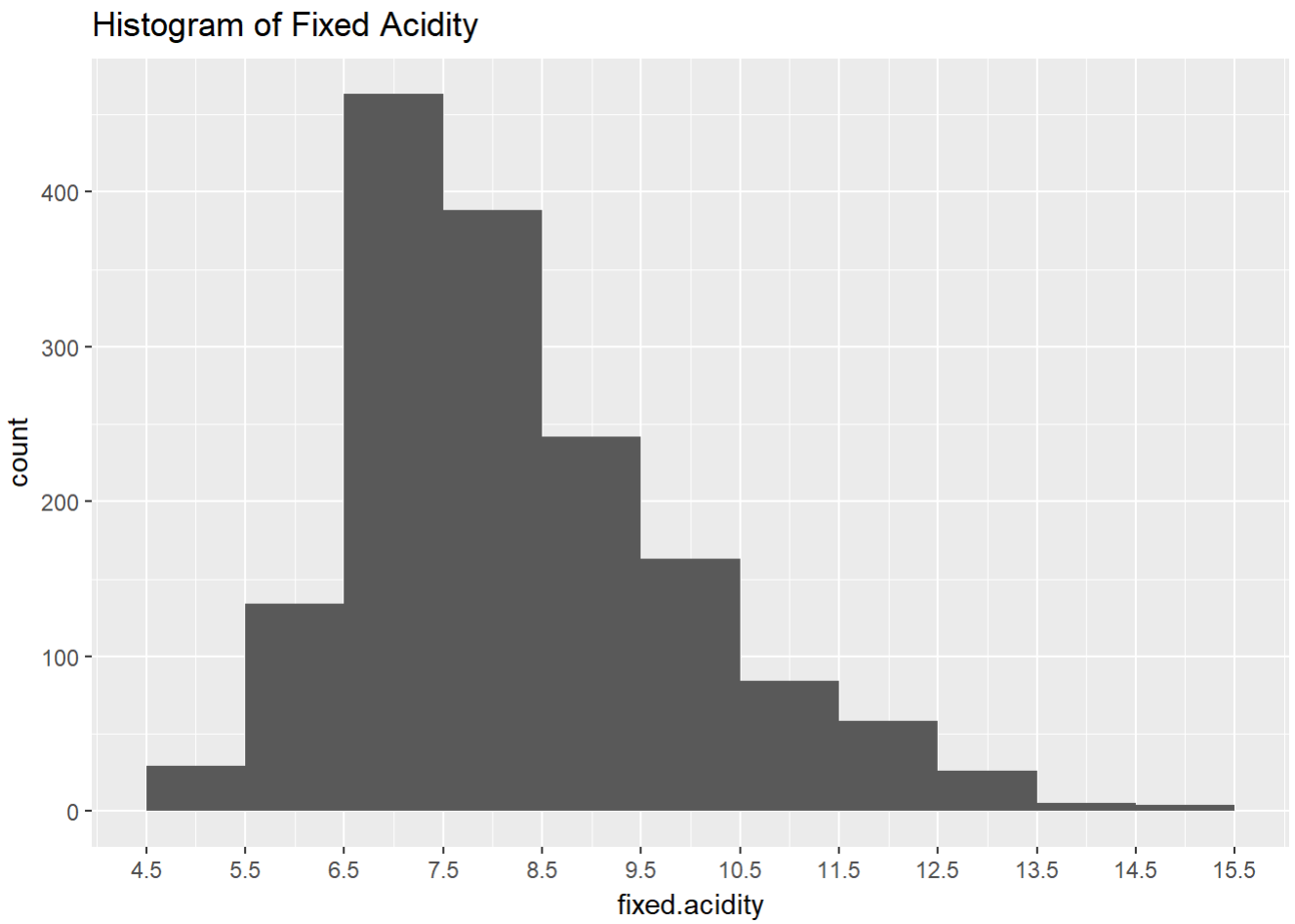
We can see that most of the wines fall into 5 or 6 followed by 7. Few wines are in quality 3,4 and 8.



Fixed Acidity

fixed.acidity seems to have a slight positve skew. Most of the wines have fixed.acidity value between 6.5 and 7.5.

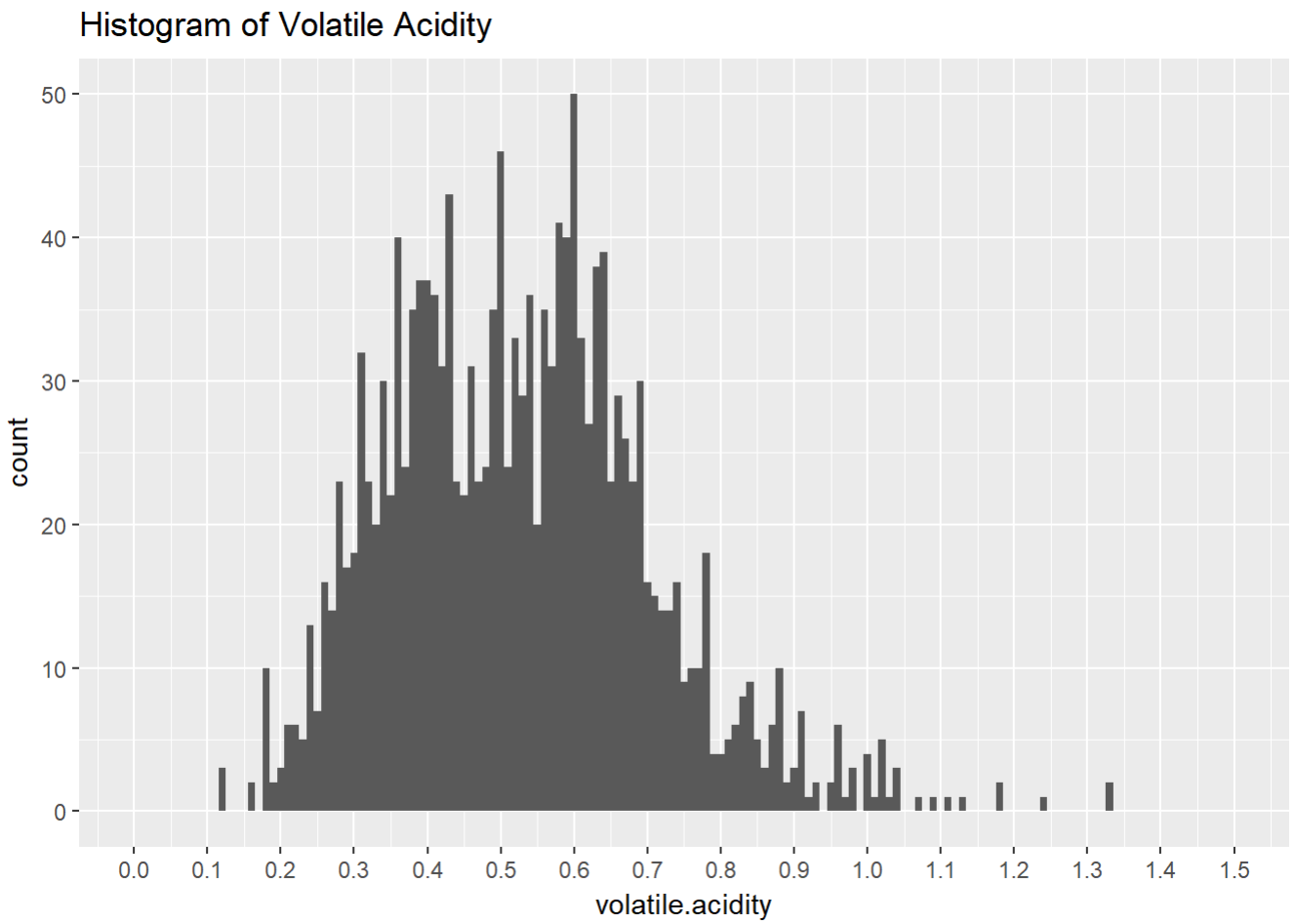
```
## Warning: Removed 3 rows containing non-finite values (stat_bin).
```



Volatile Acidity

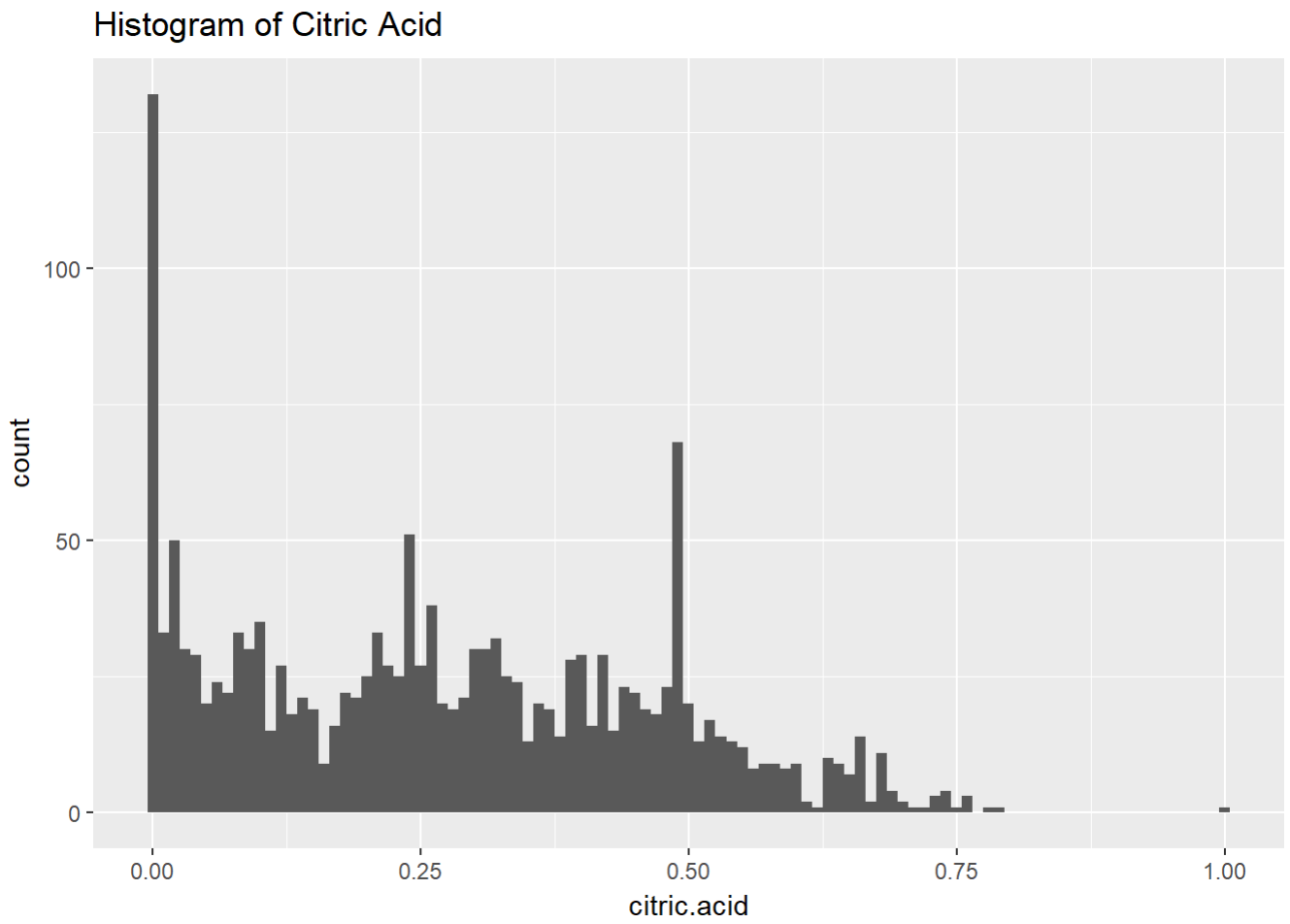
Volatile Acidity does not seem to have any particular kind of distribution. We can see that there are outliers present on the positive end of the distribution. This means there are few views with considerably higher amount of volatile acids.

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```



Citric Acid

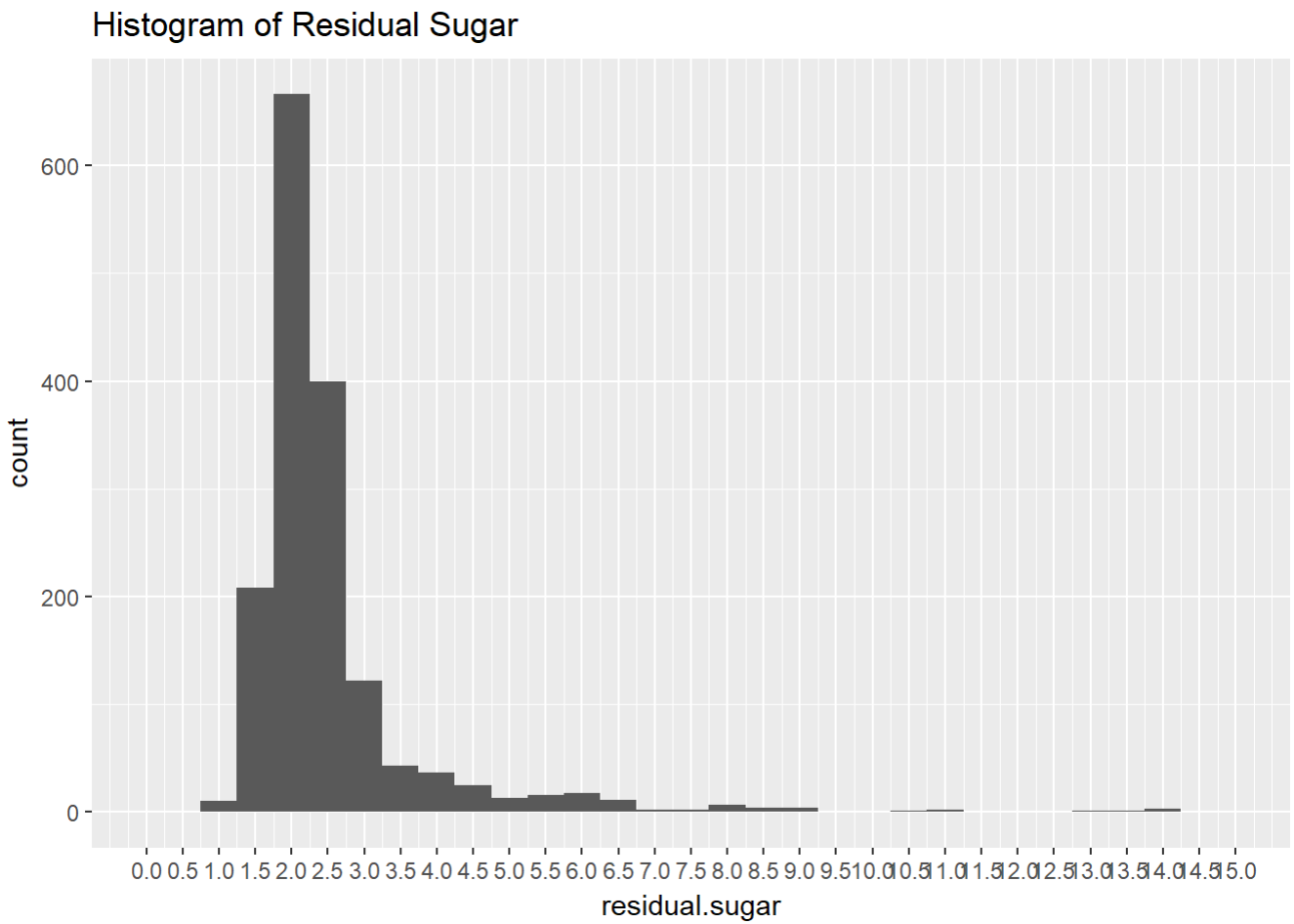
Histogram of citric acid has some interesting observations. First, though citric acid adds “freshness” to the wine, there are more than 100 wines that has no citric acid. We can also see a bump in the number of wines as amount of citric acid goes up by value of 0.25. This can be seen at 0.25 tick and 0.50 tick.



Residual Sugars

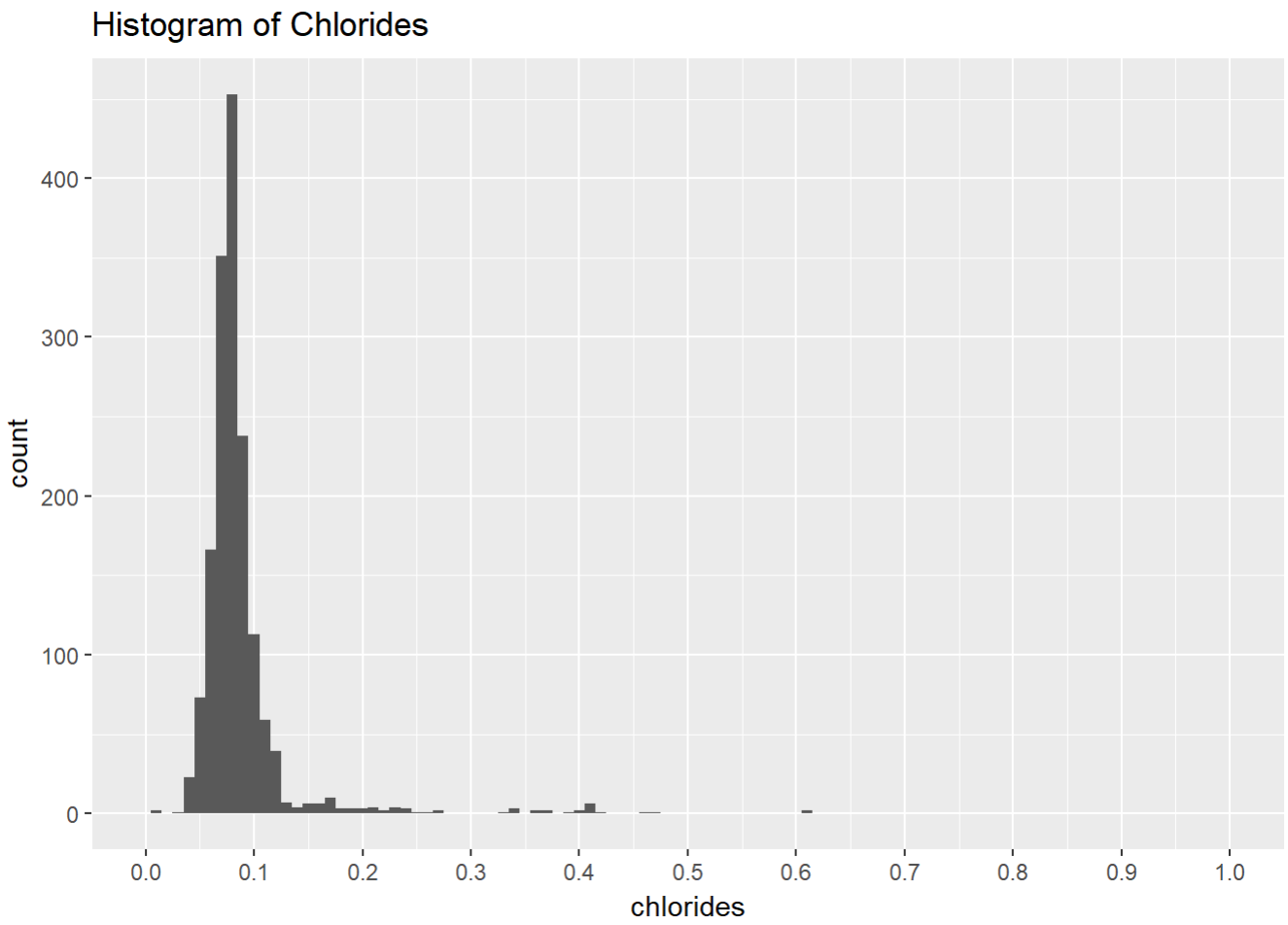
The distribution seems to have a positive skew. Most of the wines have little residual sugars while few wines have high quantity of residual sugar. We can also see few outliers on the positive end.

```
## Warning: Removed 3 rows containing non-finite values (stat_bin).
```



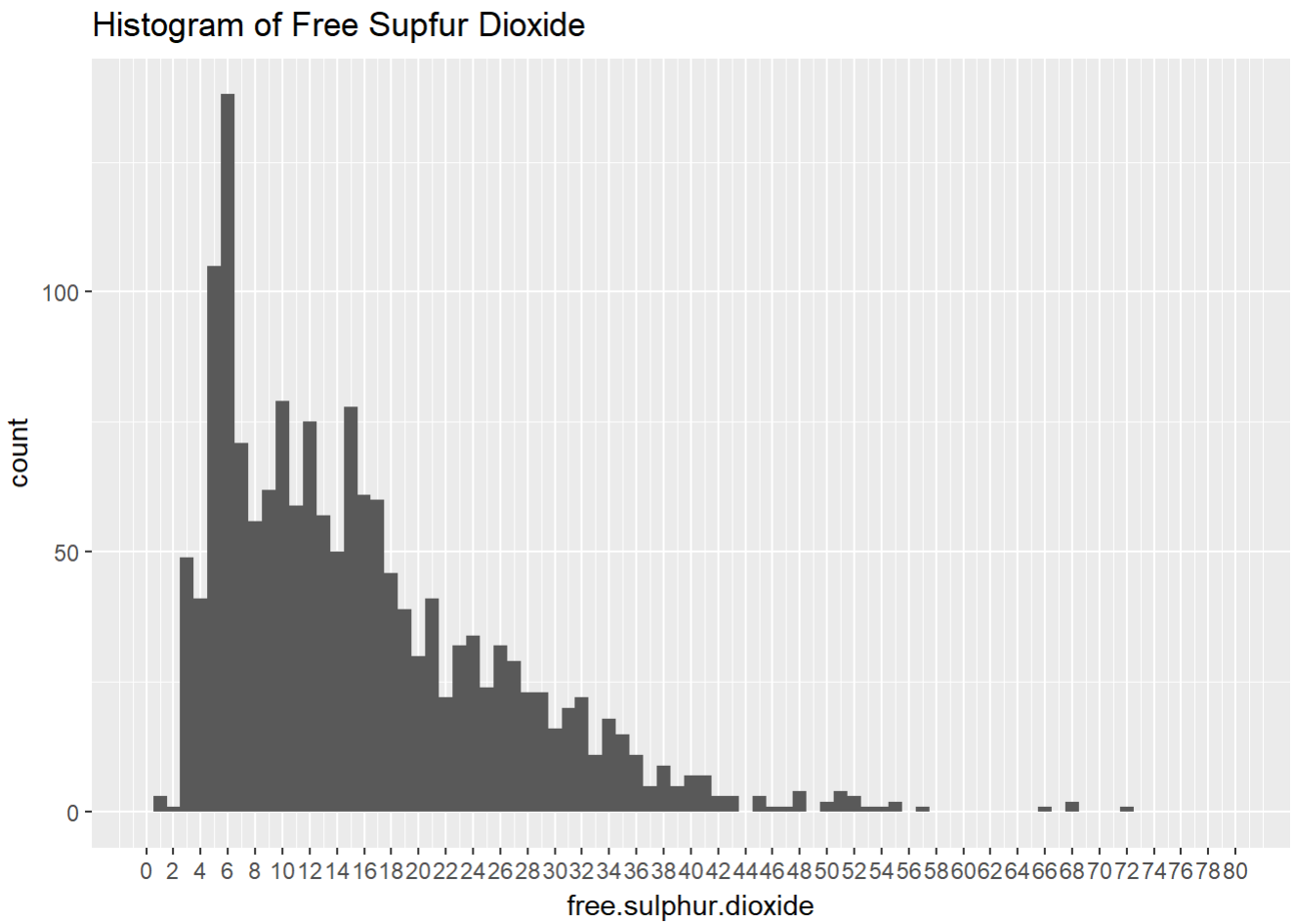
Chlorides

Chlorides seem to have a normal like distribution followed by a very thin tail and outliers on the positive end.



Free Sulphur Dioxide

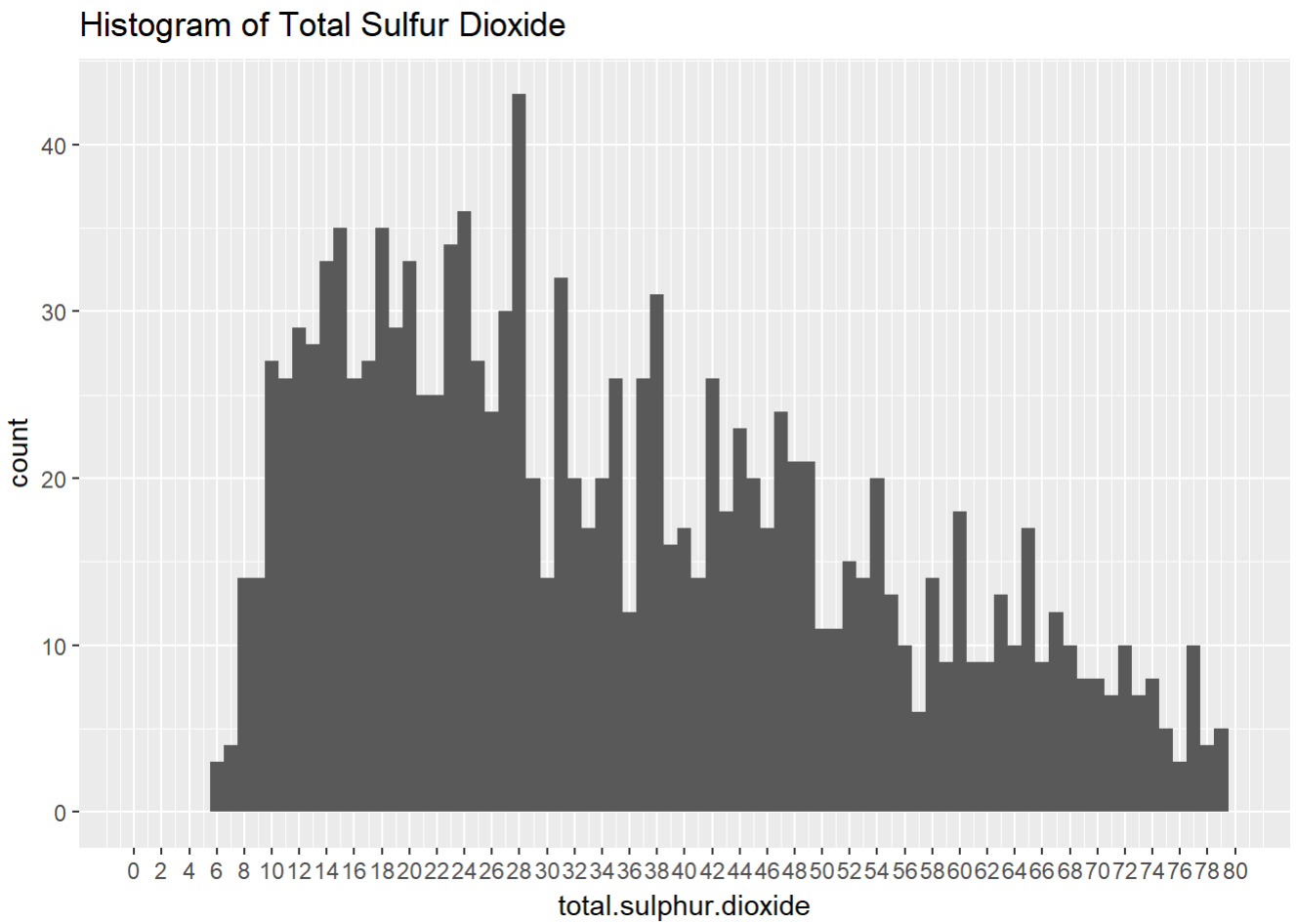
The distribution seems to have a positive skew to it. Outliers are also present at the positive end.



Total Sulphur Dioxide

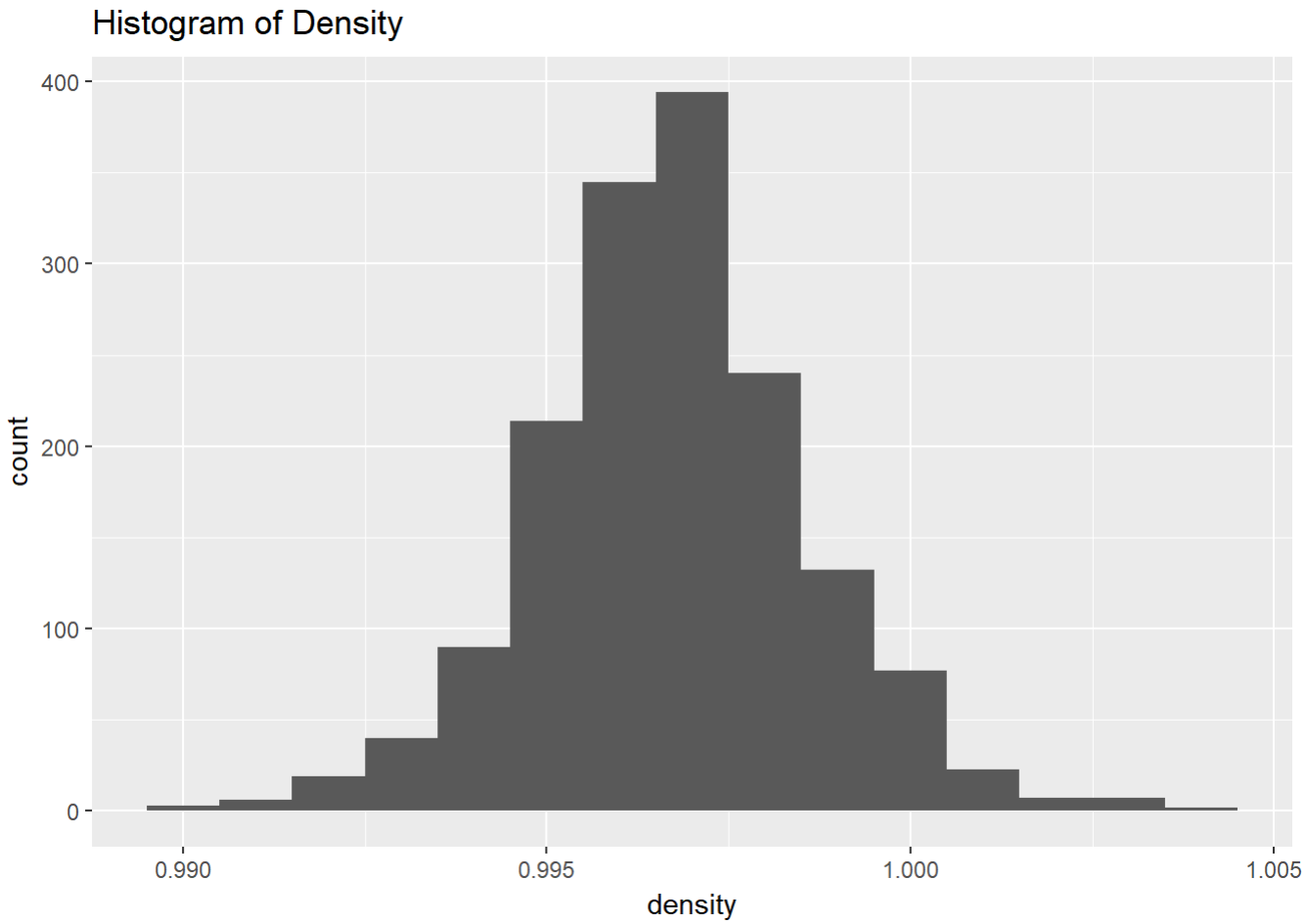
total.sulphur.dioxide seems to have a uniform like distribution with a little downwards gradient at the positive end.

```
## Warning: Removed 248 rows containing non-finite values (stat_bin).
```



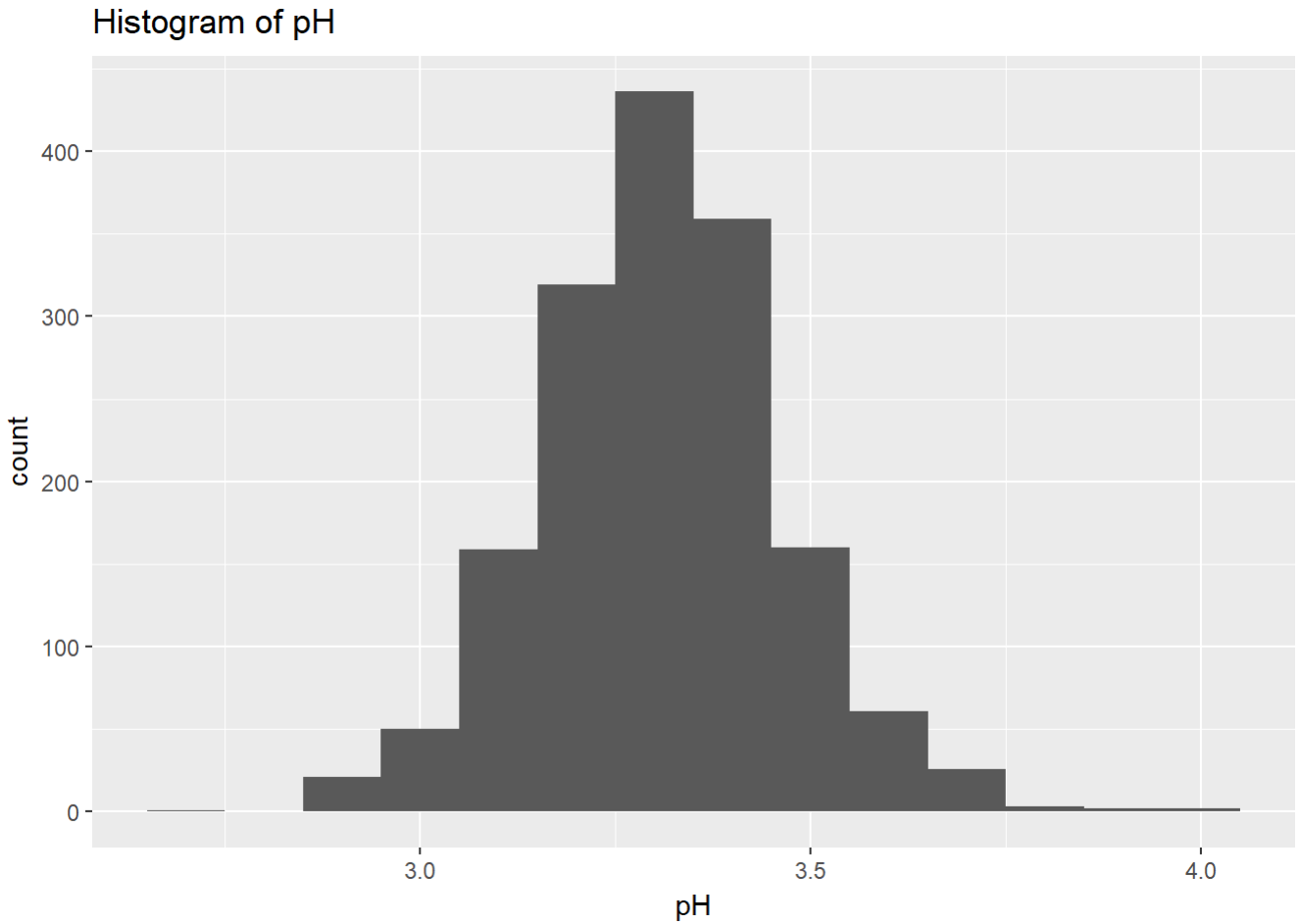
Density

Interestingly, density has a almost (perfectly) normal distribution.


Histogram of Density
400
300
200
100
0
count
0.990
0.995
1.000
1.005
density

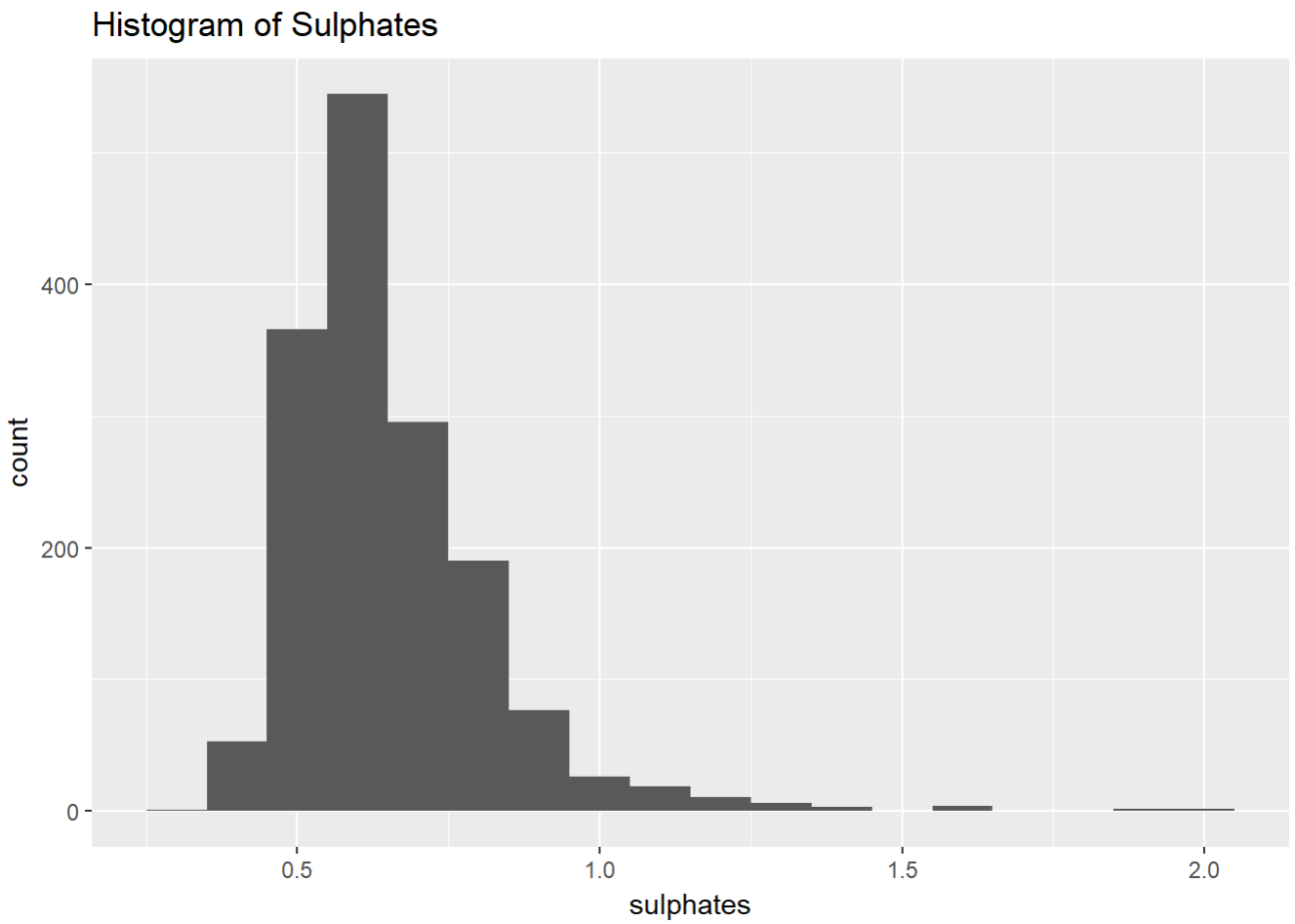
pH

pH value also seem to have a nearly normal distribution with outliers on both end.



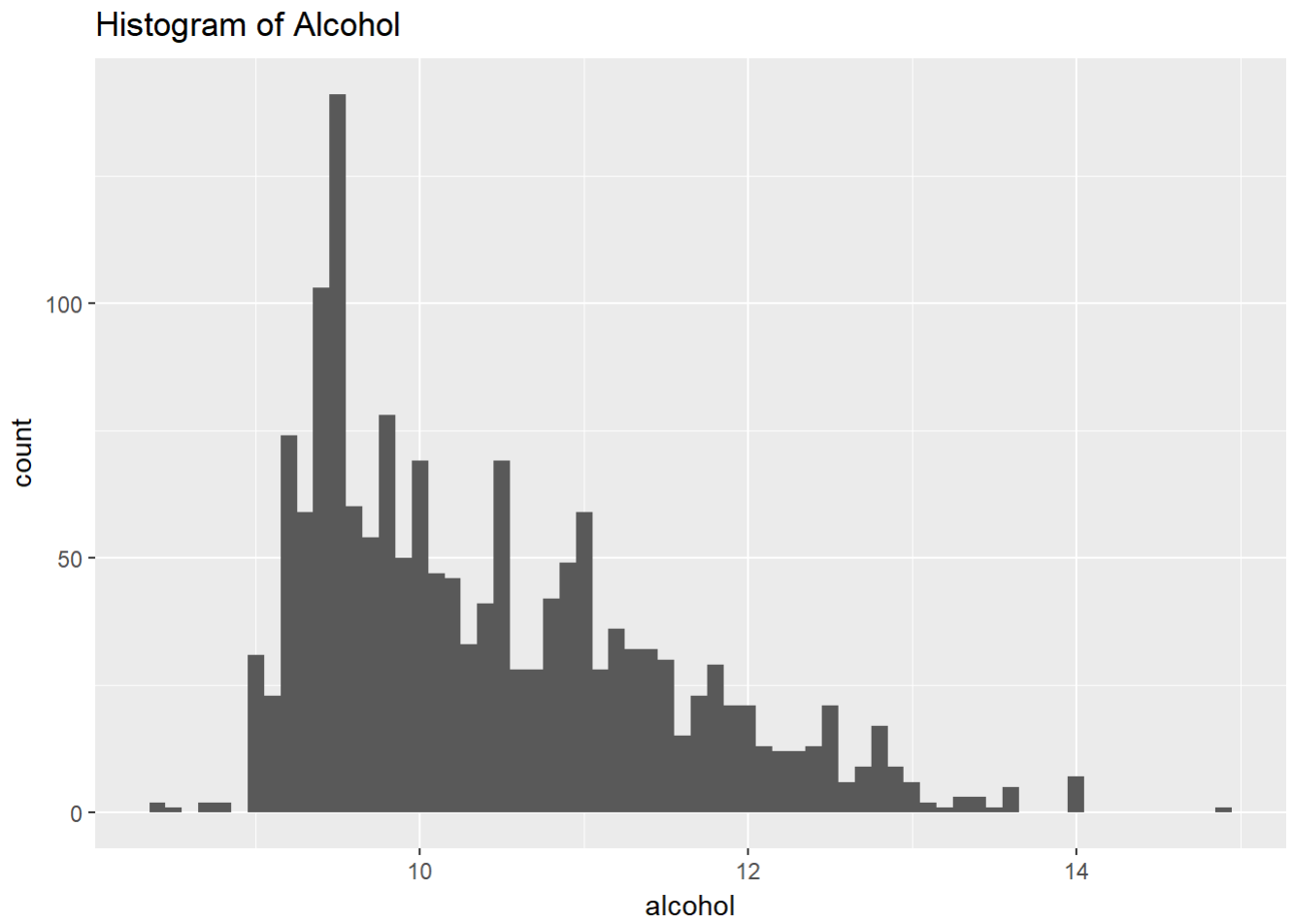
Sulphates

Sulphates clearly seem to have a positive skew with few outliers on positive side of the distribution.



Alcohol

Alcohol seems to have a positive skew. Majority of wines have alcohol value from 8 to 15.



Univariate Analysis

What is the structure of your dataset?

There are 1599 observations of wines with 12 features. All the variables are numerical variables except quality. Quality is integer (discrete) variable.

Most wines belong to the quality value of 5 or 6. Few belong to 2 or 3 and few belong to 7 or 8. Thus, wine quality has nearly normal distribution. I believe if we have more number of wine observations, the distribution can come more closer to normal distribution.

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

Quality of wine (quality) is the central feature of interest in the dataset. I am trying to model a relationship so as to predict the wine quality from other available features.

Since I did not have any intutive idea about wine tasting and wine parameters, I have performed data exploration to determine features that can predict wine quality.

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

I believe pH, alcohol,residual.sugar might contribute to the wine quality. However, this is a very vague assumption. Further, exploration is certainly needed.

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

A new variable called Quality2 has been created. It is a factor variable. Quality value of 3 and 4 is mapped as “Bad”, 5 and 6 as “Medium”" and 7 and 8 as “Good”. This helps converting the problem at hand into classification problem.

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?

I could make following observations about the plots:

While citric acid is related to freshness, there are more than 100 wines that has no cirtric acid. We can see that in the distribution of citric acid, a very high vertical line at 0.0.

Residual Sugar and chlorides have many outliers.

Also free.sulphur.dioxide has positive distribution. I have not performed any transformation to convert it to normal distribution since it is not needed at this point.

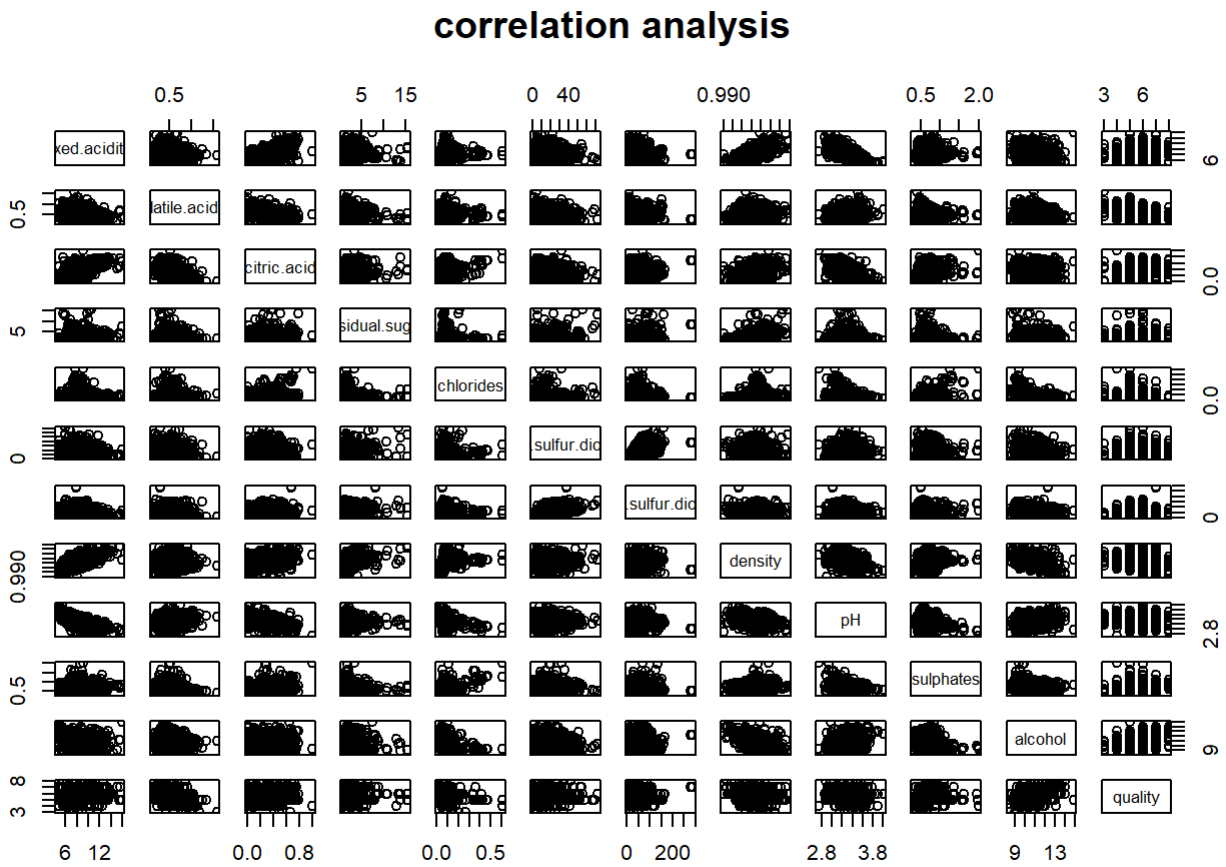
Bivariate Plots Section

We have performed correlation analysis in first step. Depending upon the strength of correlation, we have used box plots to study the relation between Quality2 and other parameters.

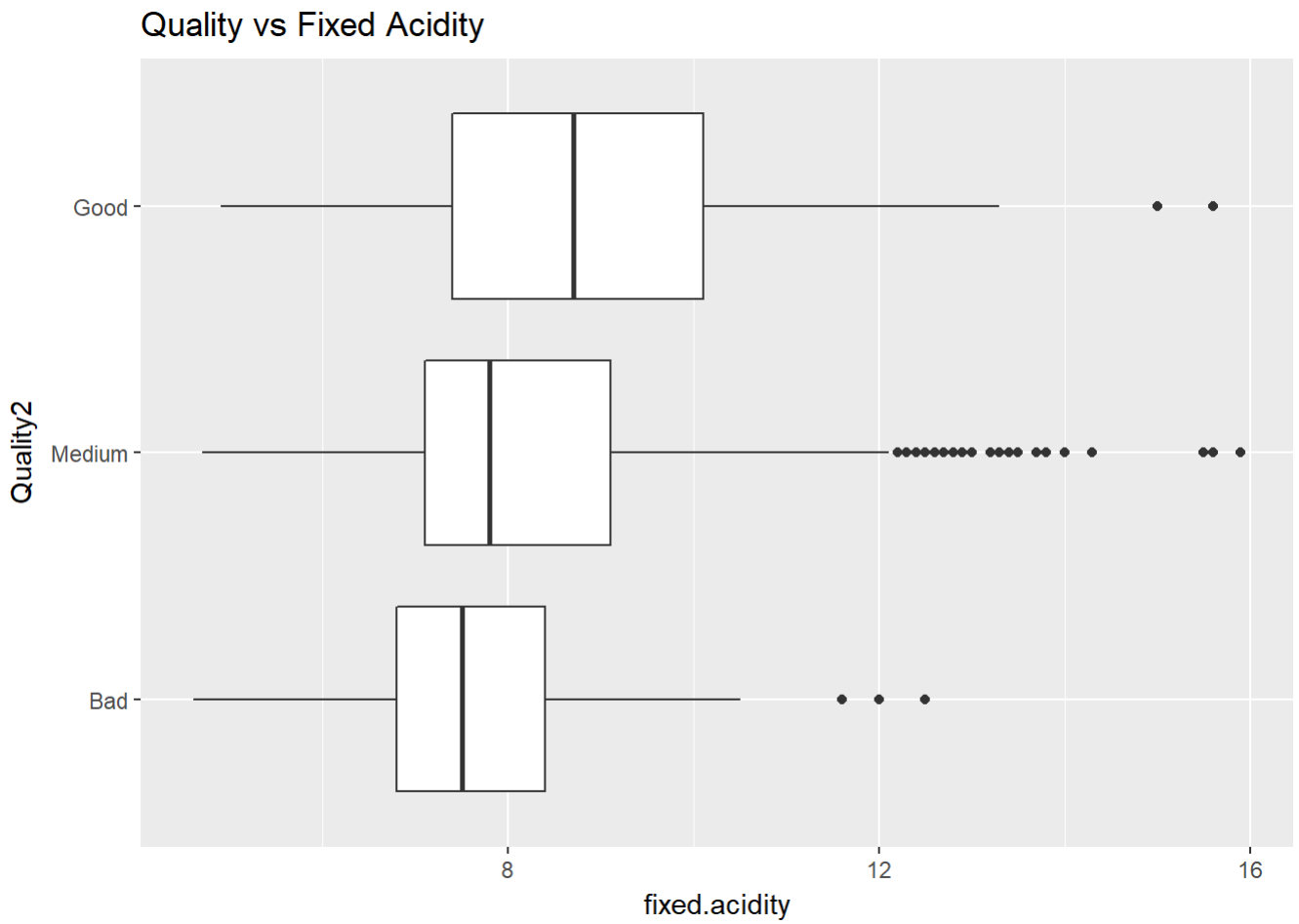
Correlation Analysis

```
##                       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
##                       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
##                       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
##                          sulphates     alcohol
## fixed.acidity          0.183005664 -0.06166827
## volatile.acidity      -0.260986685 -0.20228803
## citric.acid            0.312770044  0.10990325
```

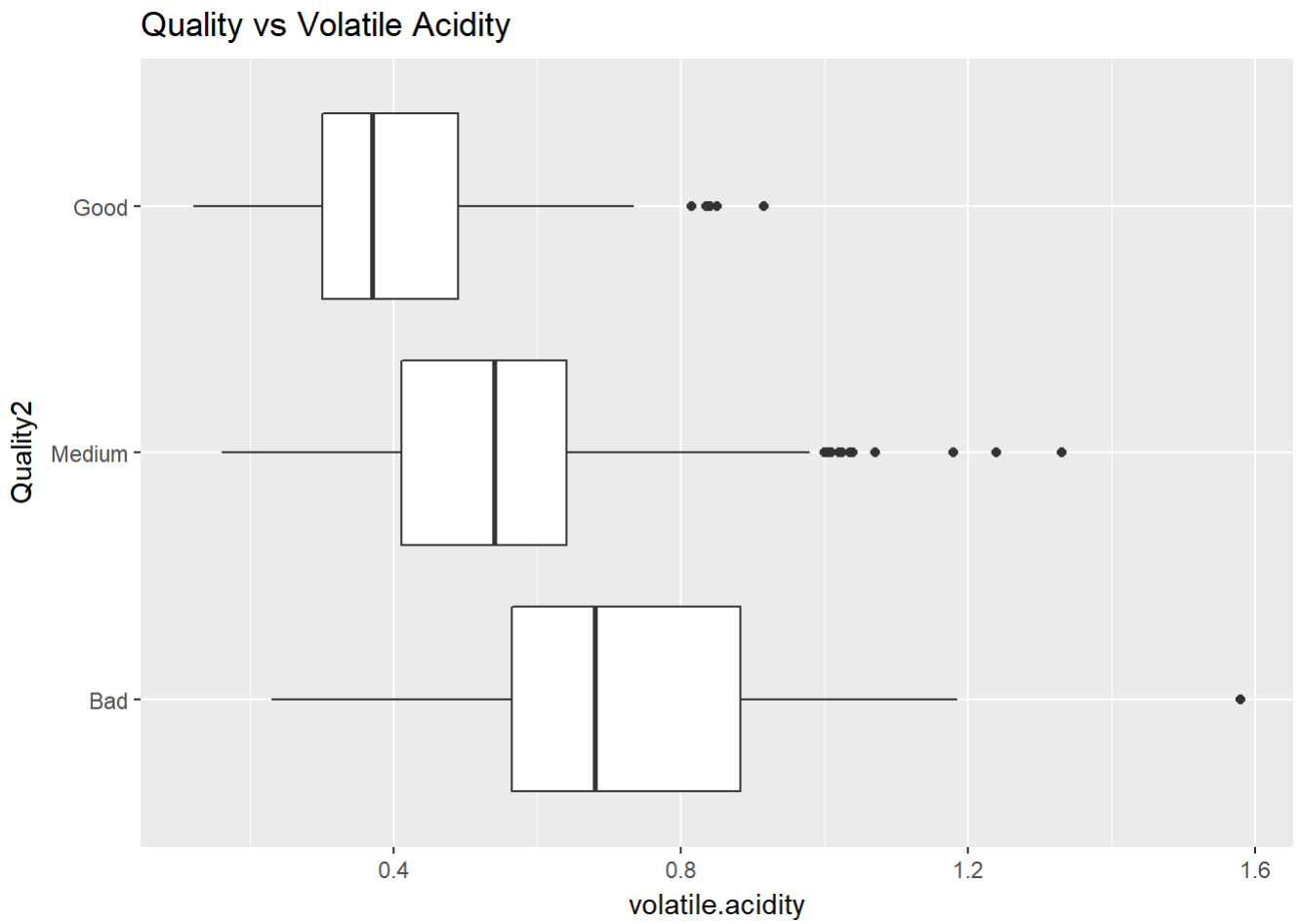
```
## residual.sugar         0.005527121  0.04207544
## chlorides              0.371260481 -0.22114054
## free.sulfur.dioxide    0.051657572 -0.06940835
## total.sulfur.dioxide   0.042946836 -0.20565394
## density                0.148506412 -0.49617977
## pH                    -0.196647602  0.20563251
## sulphates              1.000000000  0.09359475
## alcohol                0.093594750  1.00000000
```



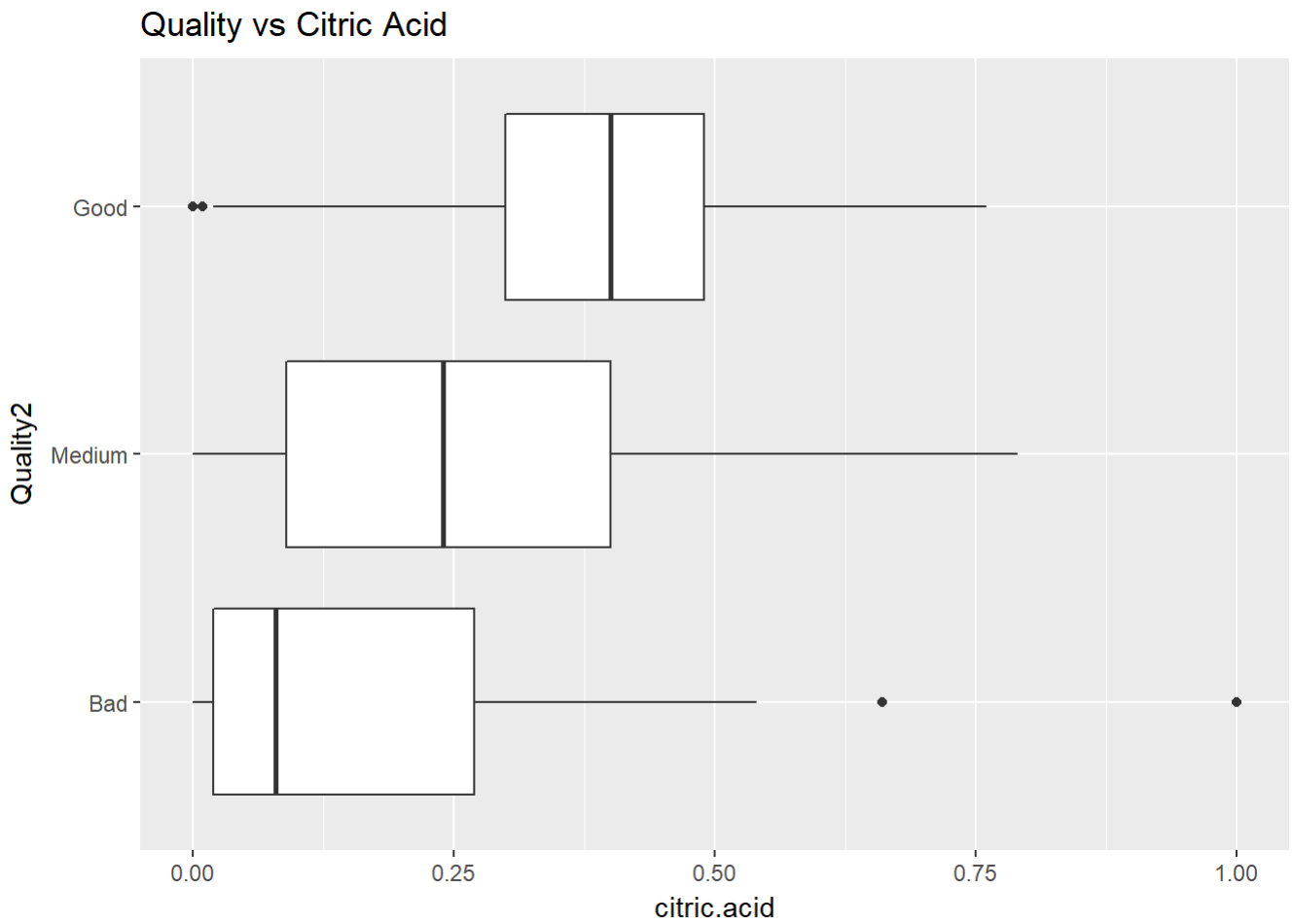
Quality of wine improves as fixed.acidity increases



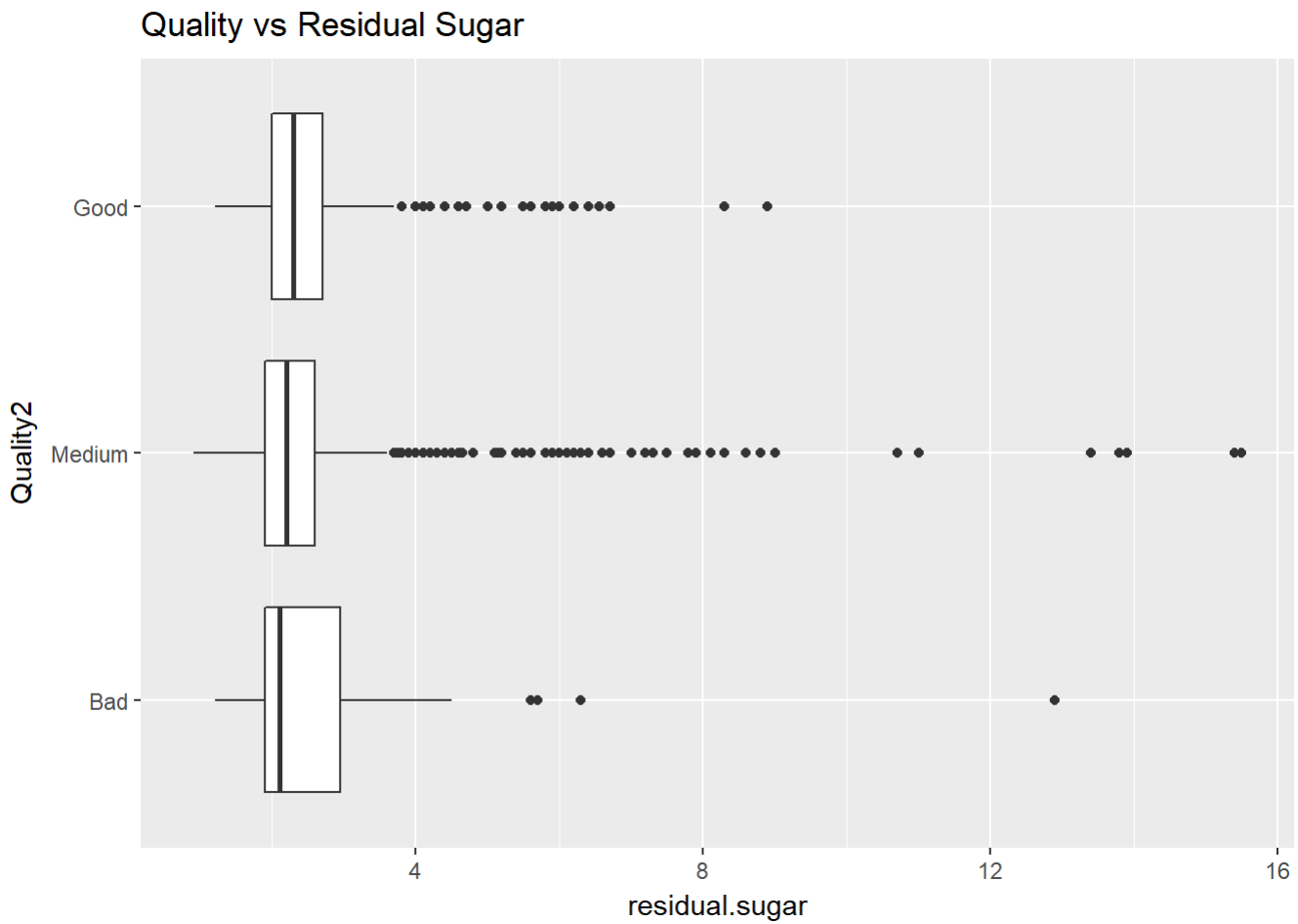
Quality of wine deteriorates as volatile.acidity increases



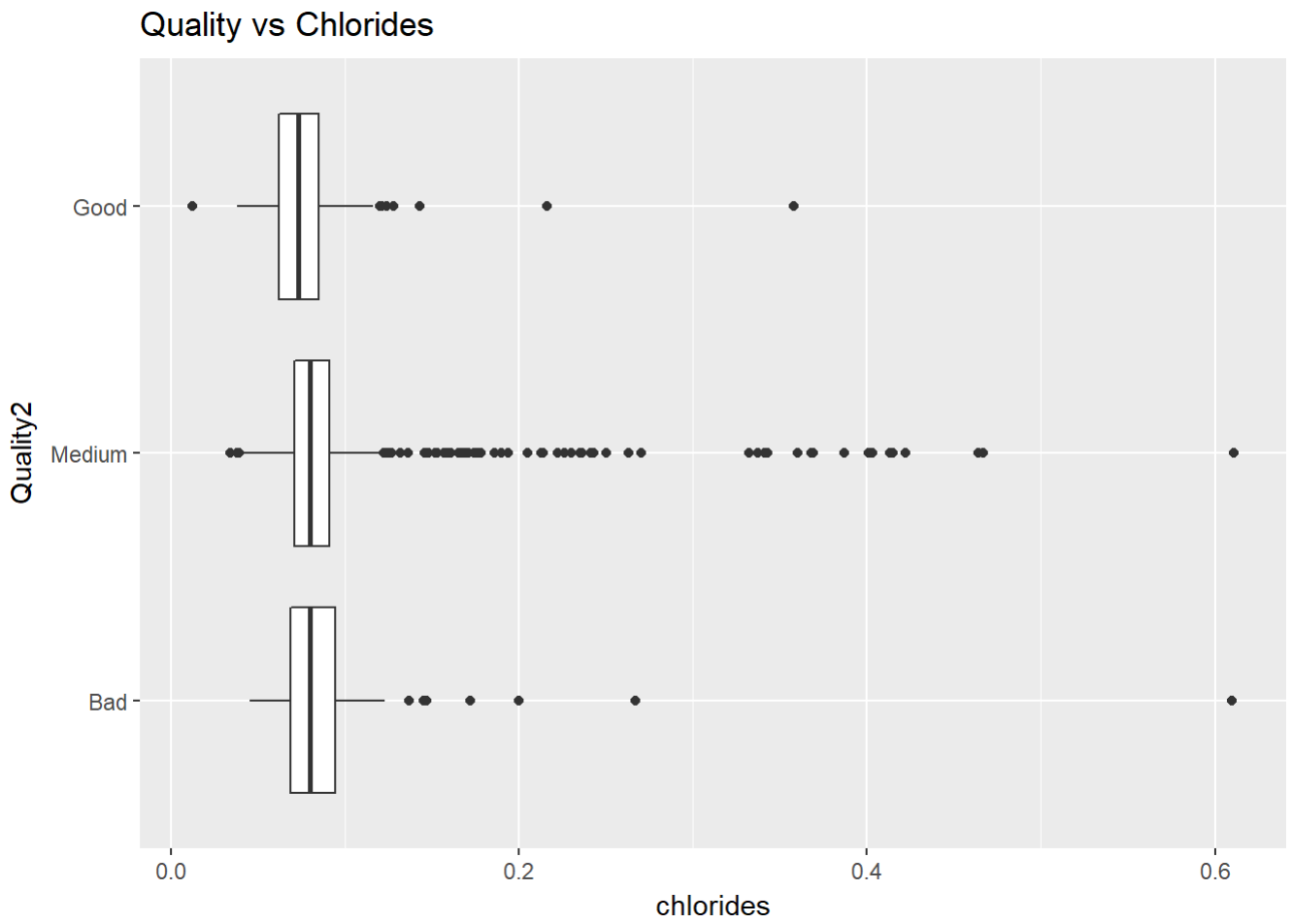
Quality of wine improves as citric.acid increases



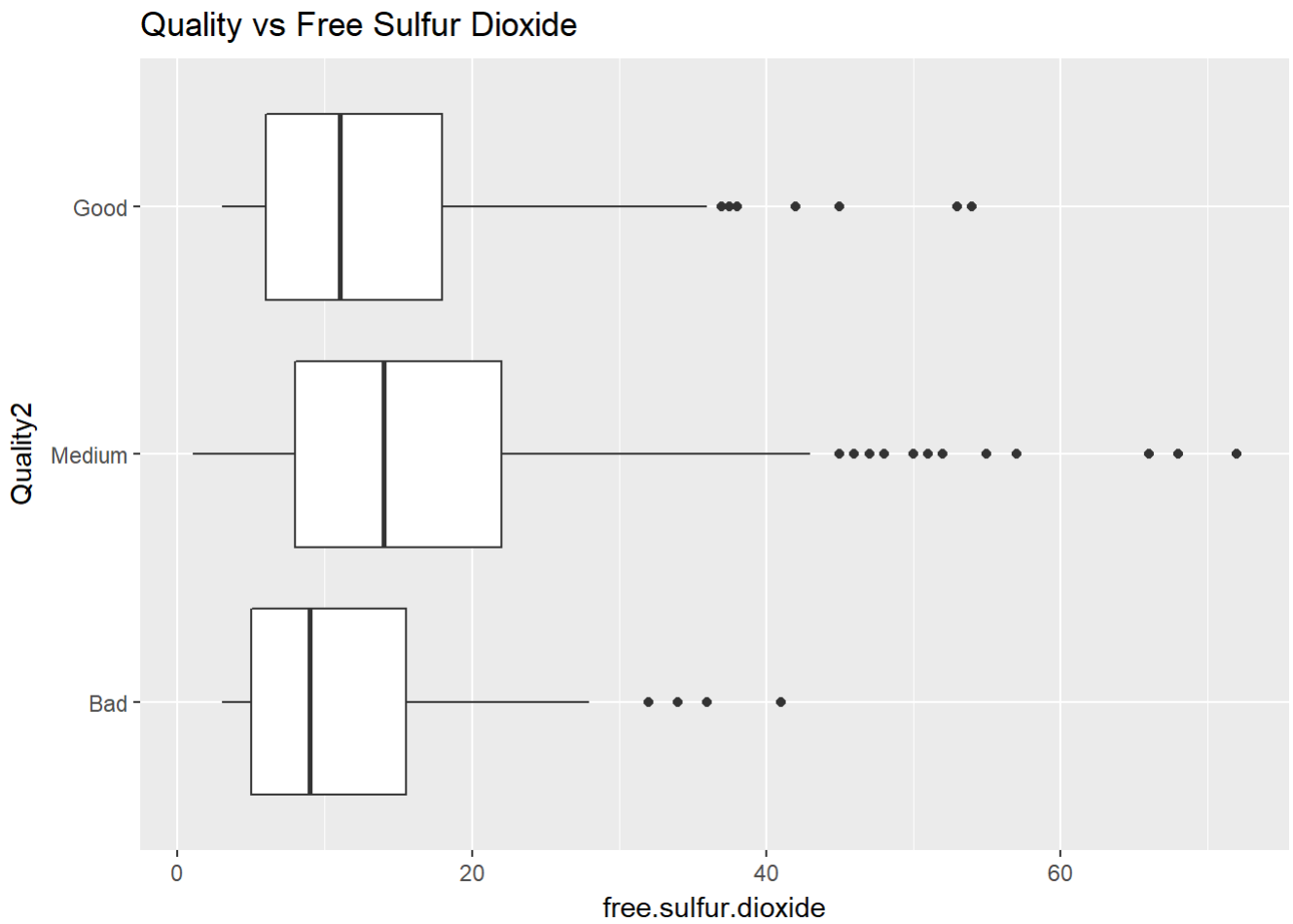
Relation between Quality of wine and residual.sugar cannot be established



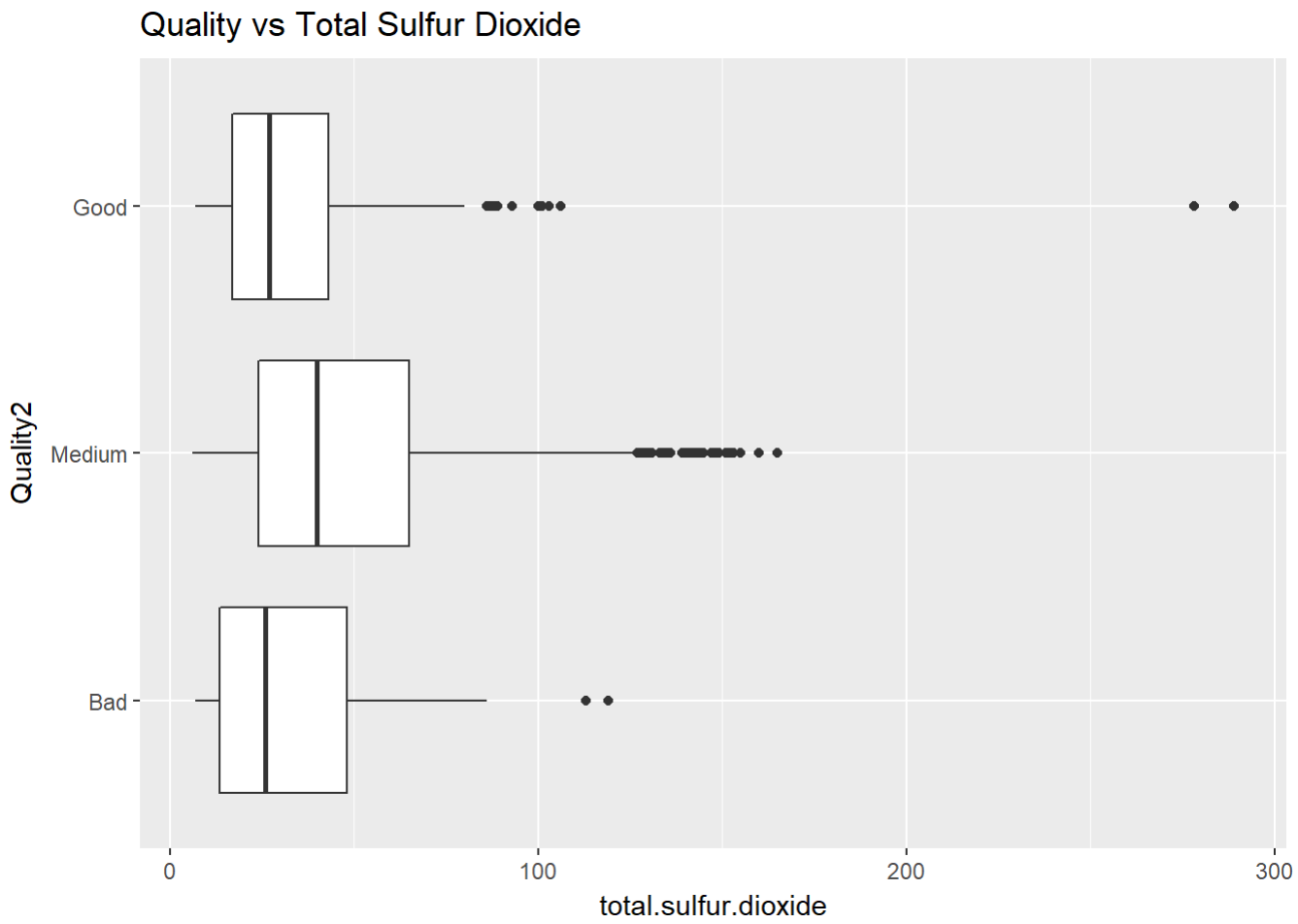
Relation between Quality of wine and chlorides cannot be established



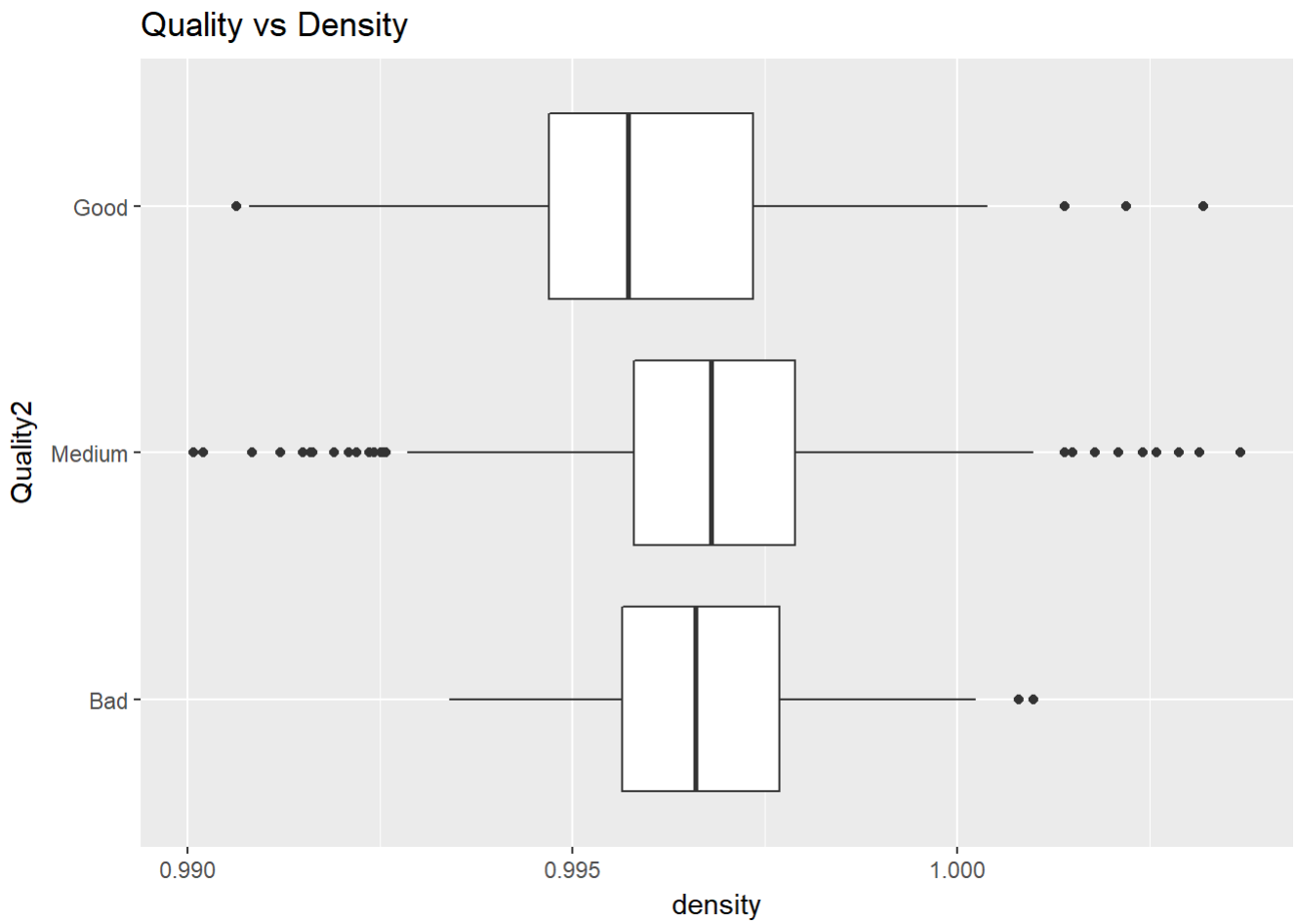
Relation between Quality of wine and free.sulfur.dioxide cannot be established



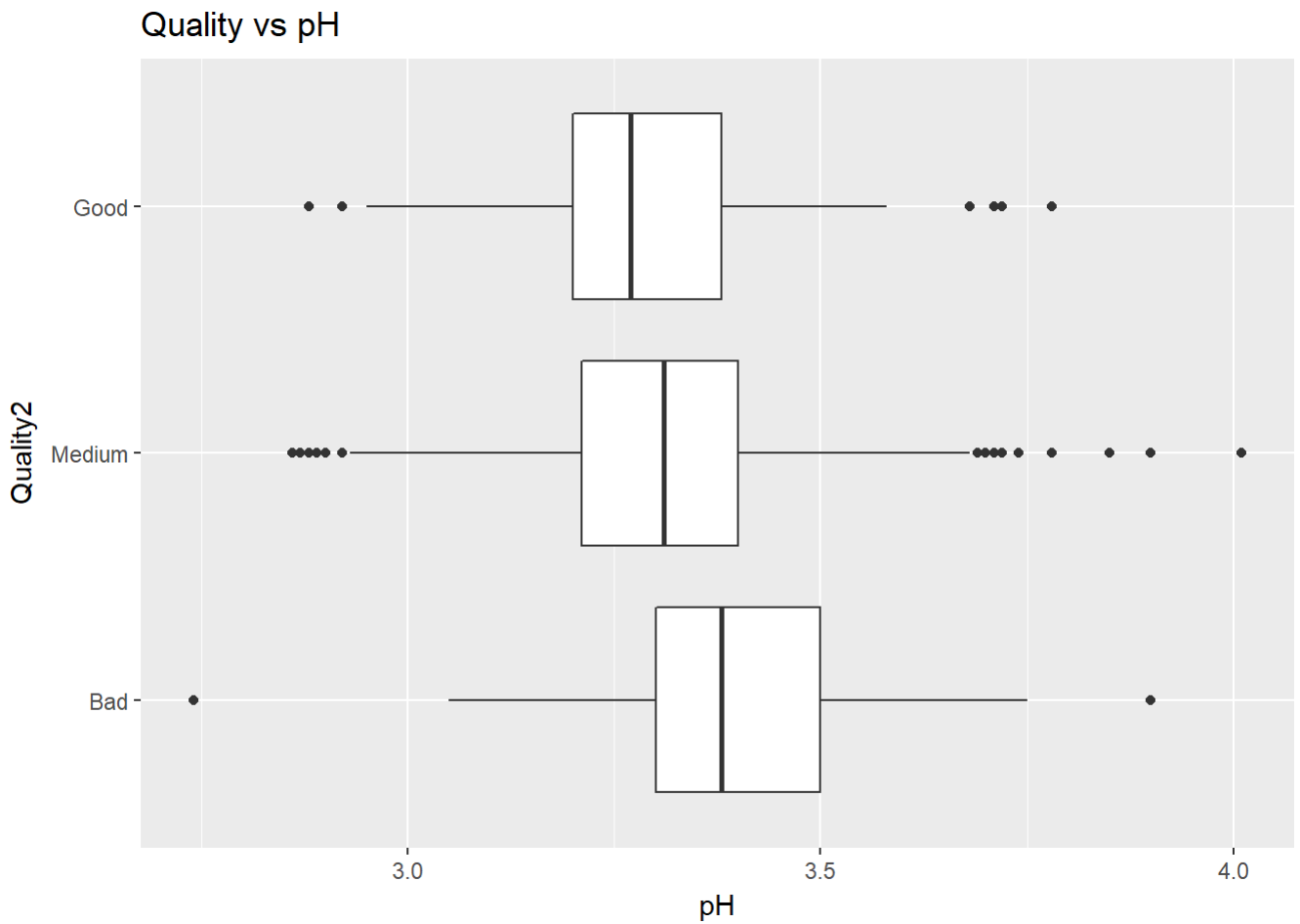
Relation between Quality of wine and total.sulfur.dioxide cannot be established



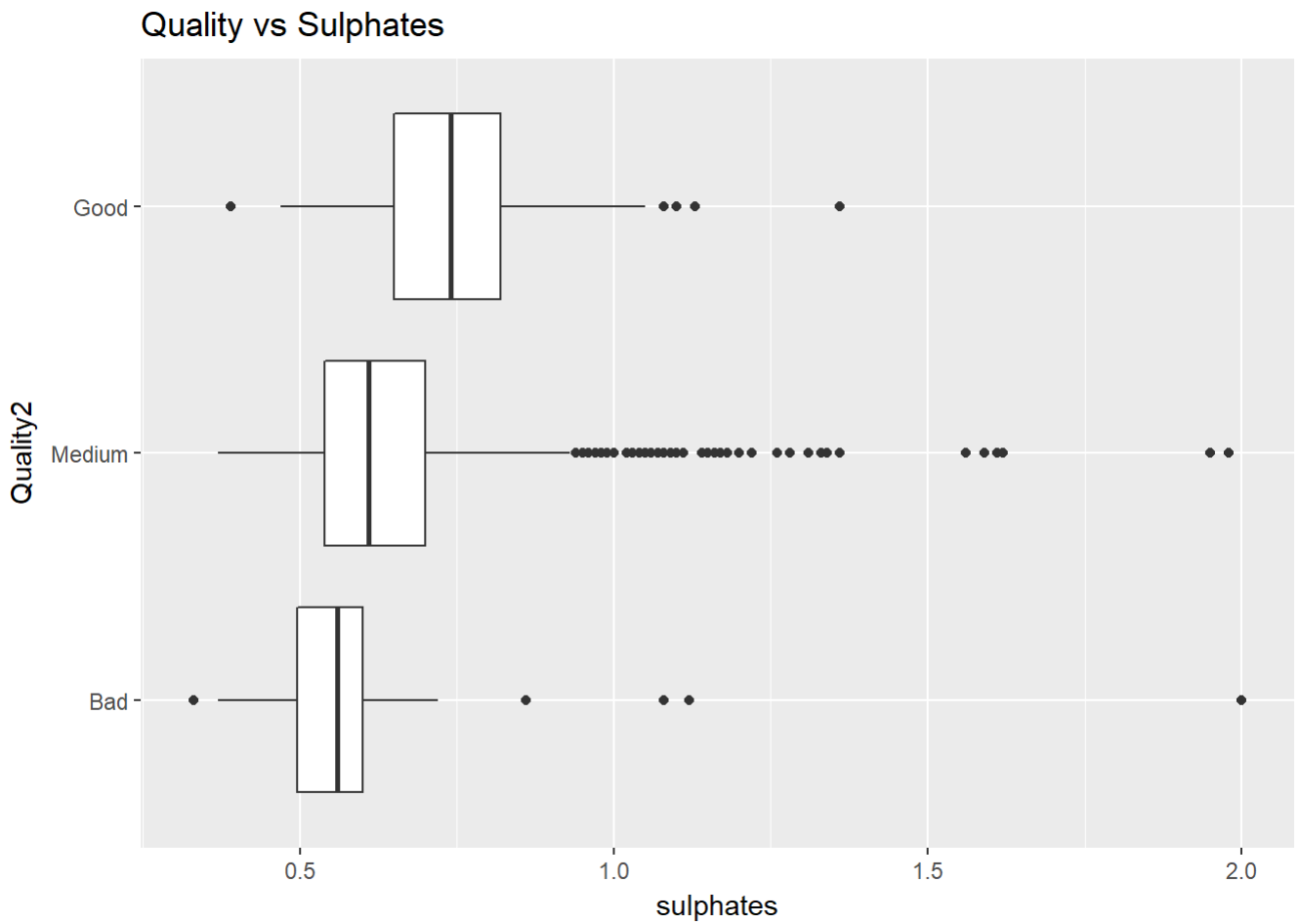
Relation between Quality of wine and density cannot be established



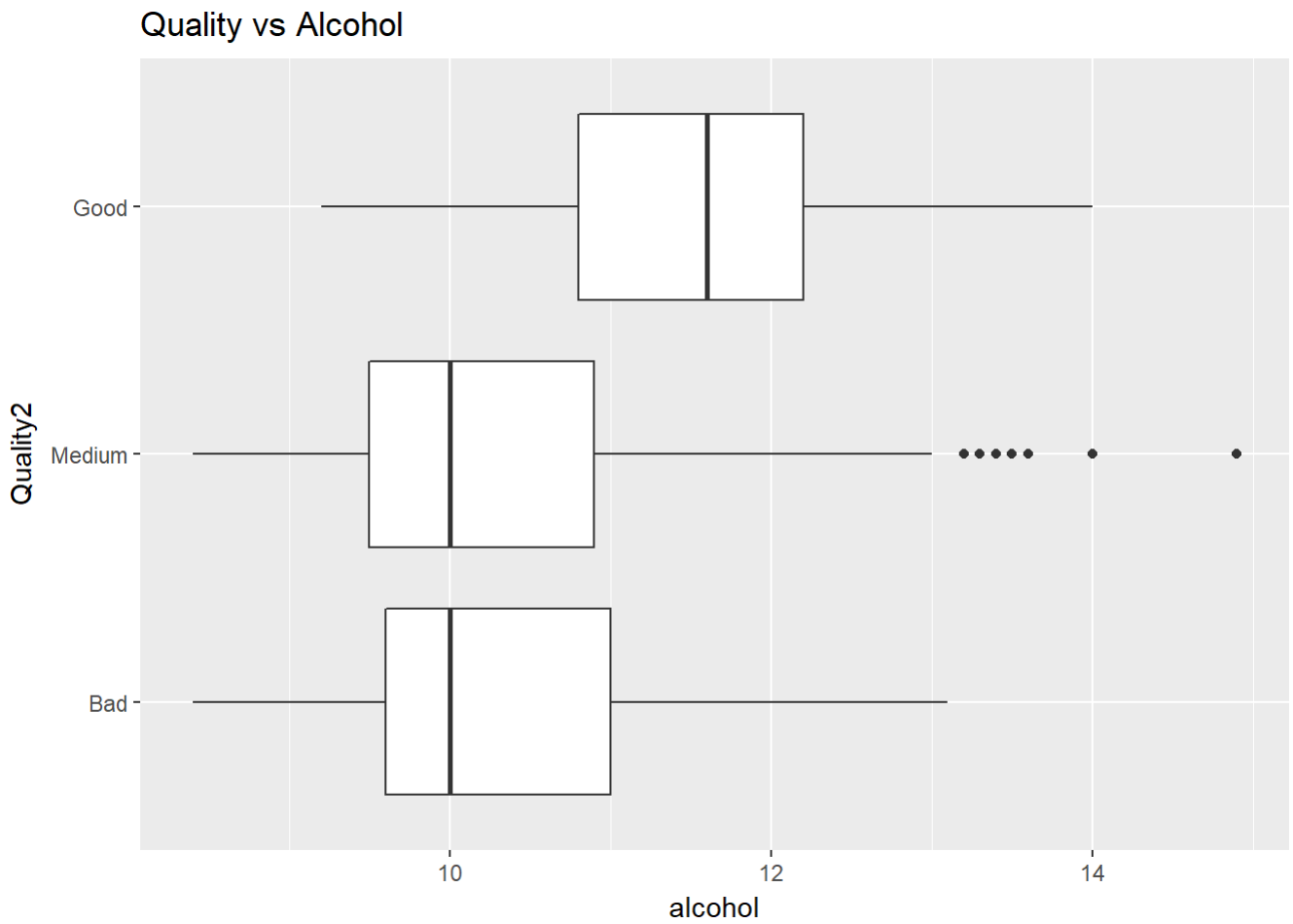
Quality deteriorates as the pH increases



Quality of wine improves as the sulphates increases.

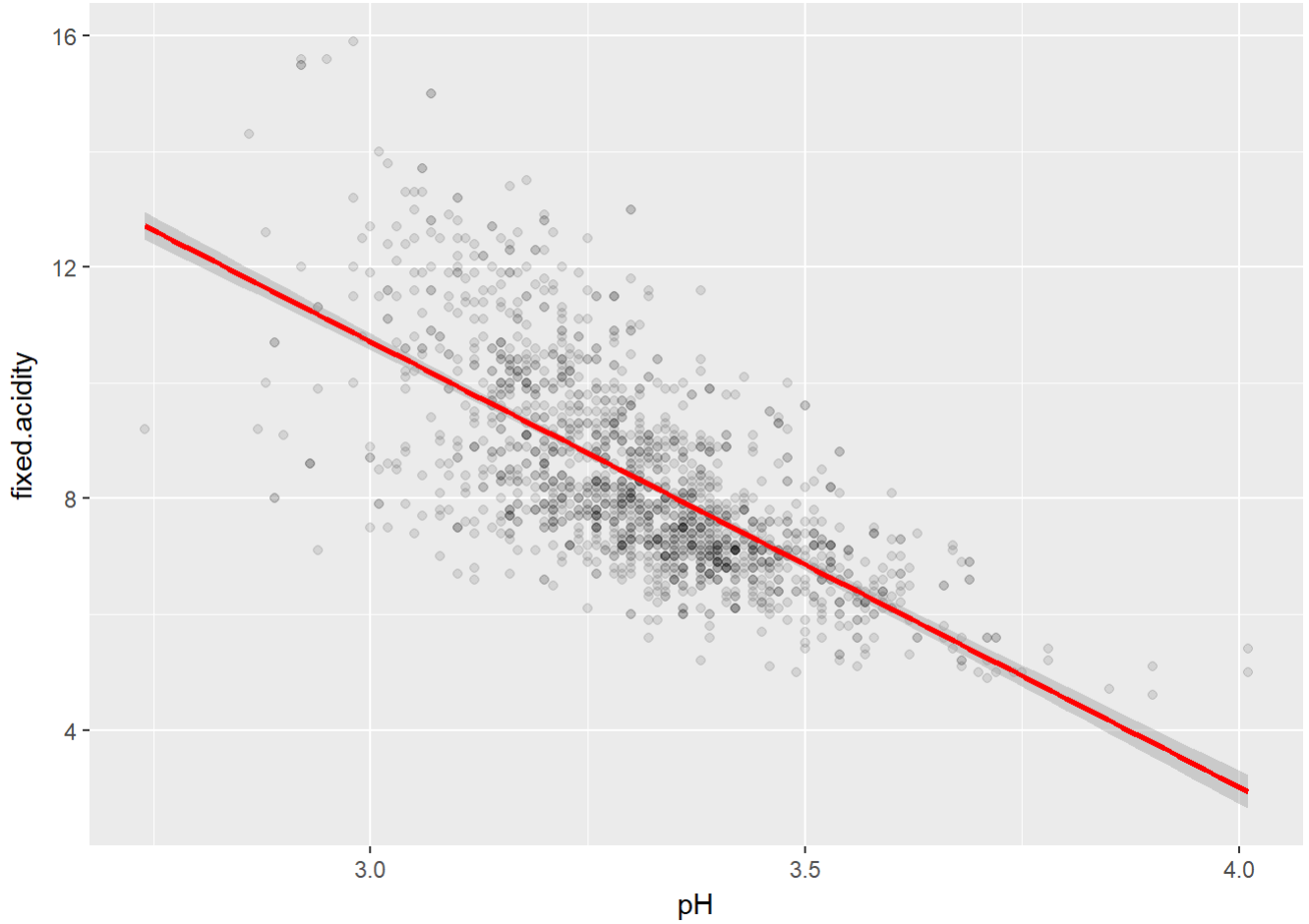


Quality of wine improves as the alcohol content increases

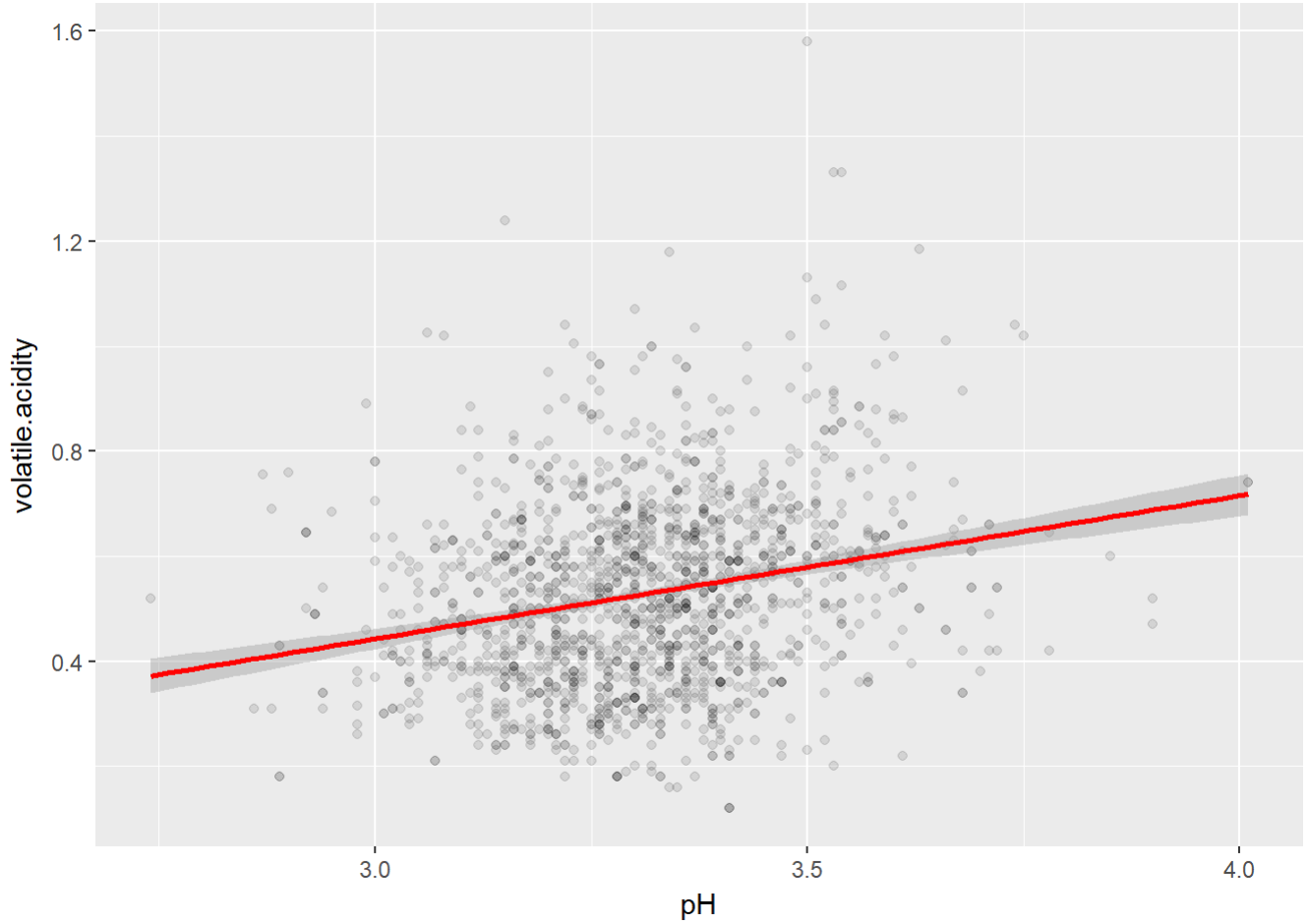


Exploration of relationship other than the one with feature of interest

When I explored the relationship between the pH and fixed.acidity and that between pH and volatile.acidity, I found that pH and fixed.acidity are negatively correlated. This was expected. However, volatile.acidity and pH are somewhat positively correlated. This was completely unexpected on my part.

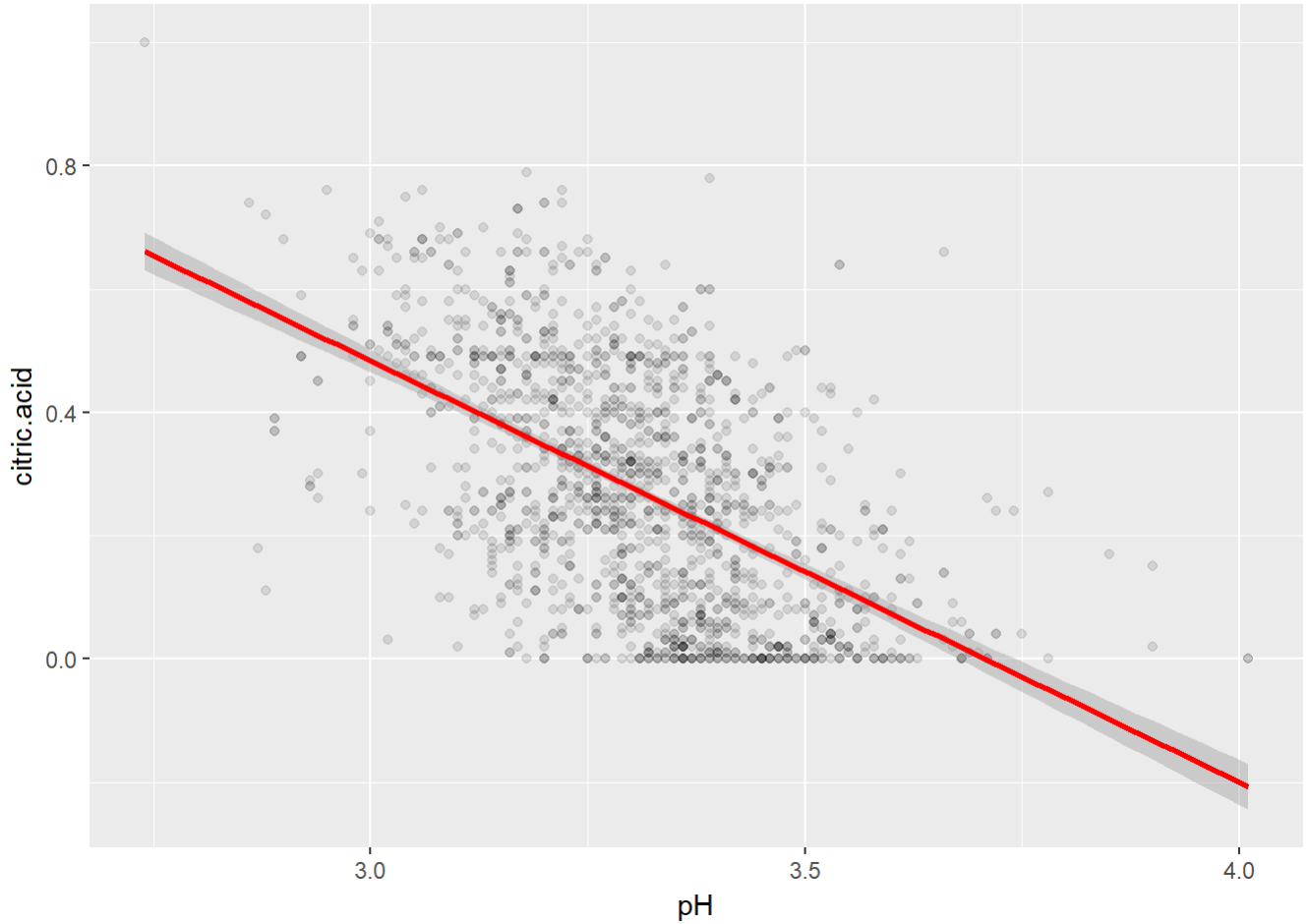

16
12
8
4
fixed.acidity
3.0
3.5
4.0
pH

```
##
## Call:
## lm(formula = pH ~ fixed.acidity, data = rw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51780 -0.06547  0.00164  0.06488  0.52207
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.814959   0.013776  276.93   <2e-16 ***
## fixed.acidity -0.060561   0.001621  -37.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1128 on 1597 degrees of freedom
## Multiple R-squared:  0.4665, Adjusted R-squared:  0.4661
## F-statistic:  1396 on 1 and 1597 DF,  p-value: < 2.2e-16
```

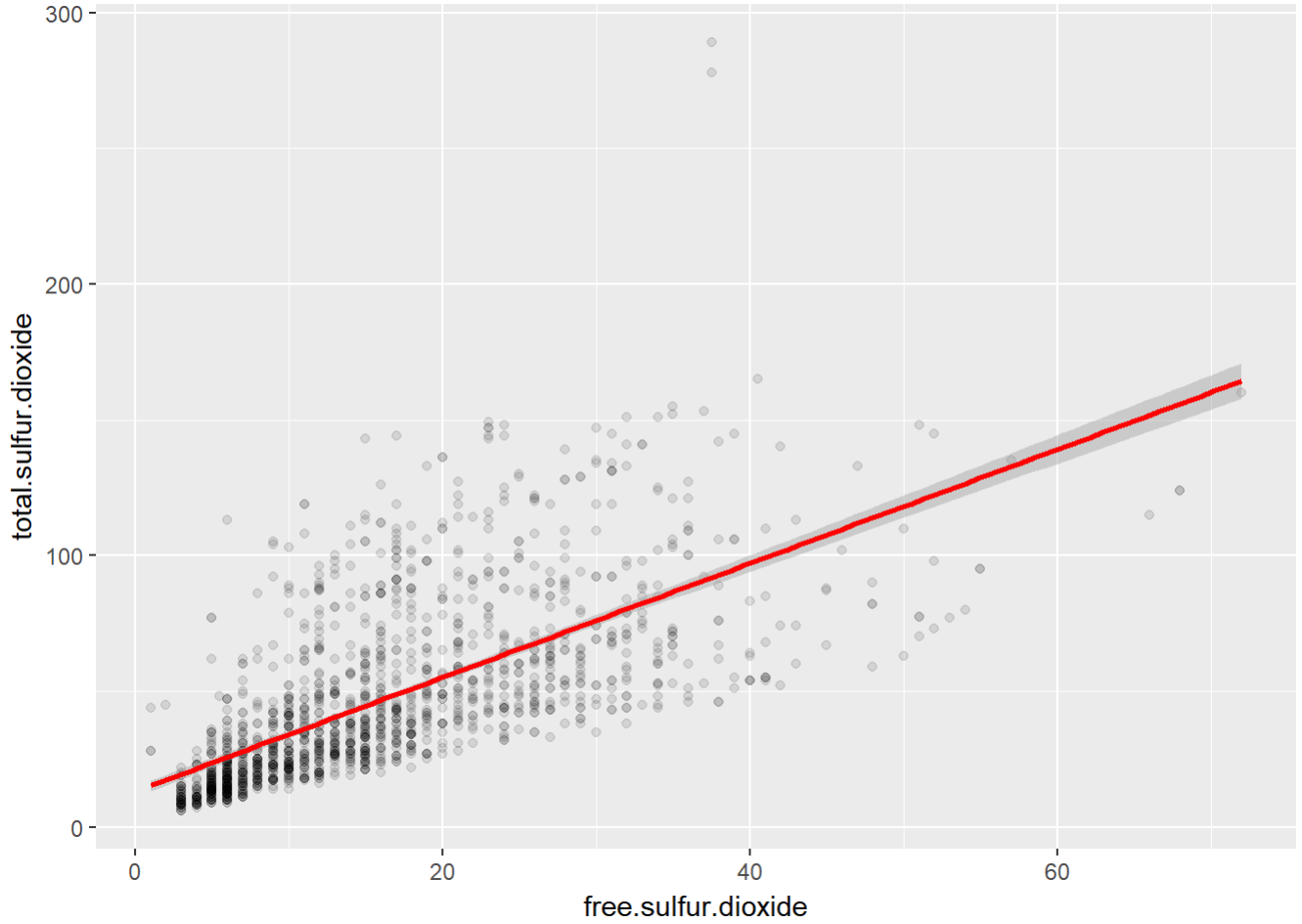

1.6
1.2
0.8
0.4
volatile.acidity
3.0
3.5
4.0
pH

```
##
## Call:
## lm(formula = pH ~ volatile.acidity, data = rw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.56953 -0.09895  0.00098  0.09301  0.65591
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.20420    0.01169 274.134   <2e-16 ***
## volatile.acidity  0.20256    0.02097   9.659   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1501 on 1597 degrees of freedom
## Multiple R-squared:  0.0552, Adjusted R-squared:  0.0546
## F-statistic:  93.3 on 1 and 1597 DF,  p-value: < 2.2e-16
```

Citric acid and pH are negatively correlated as well.



total.sulfur.dioxide and free.sulfur.dioxide are also positively correlated as expected.



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?

Quality2 variable is independent variable. All the variables that have correlation of more than 0.5 with Quality variables were plotted on box plot.

OBSERVATIONS:

From the box plots, we can observe that in general,

Quality improves as the fixed.acidity increases. Quality deteriorates as the volatile.acidity increases. Quality improves as the citric.acid increases. Quality deteriorates as the pH increases. Quality improves as the sulphates increases. Quality improves as the alcohol content increases.

For all the other variables, there is no specific relation. Also, there are too many outliers and it is difficult to establish anything visually.

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

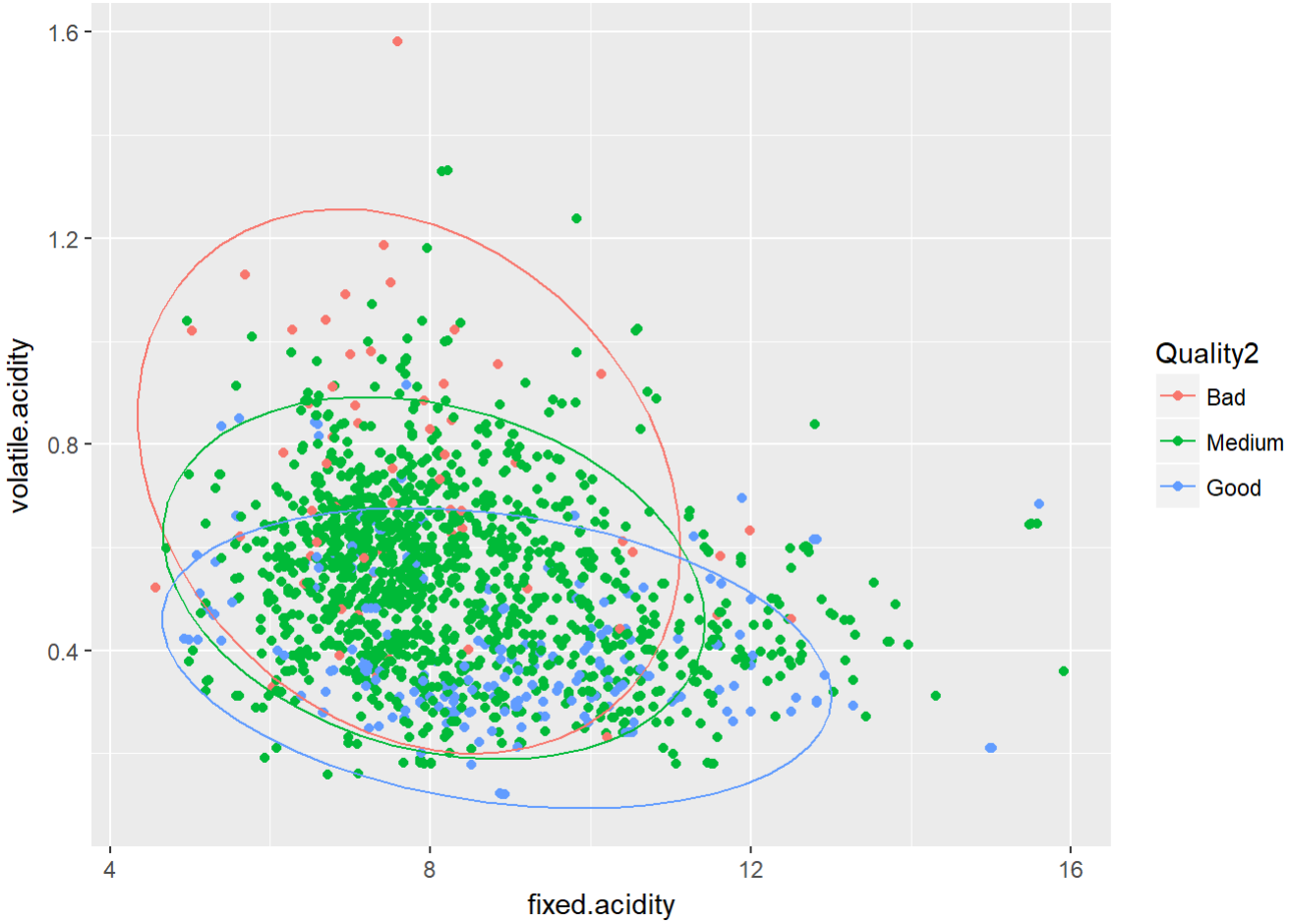
As expected, pH and fixed acidity are negatively correlated (-0.68). Also linear model explains about 40% is explain by the model.

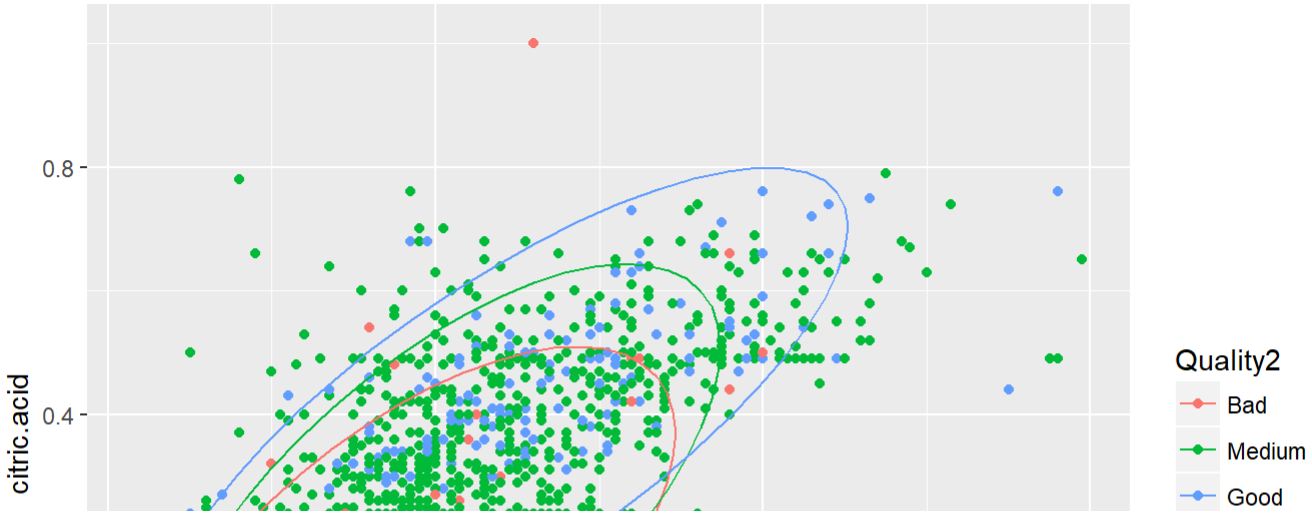
What was the strongest relationship you found?

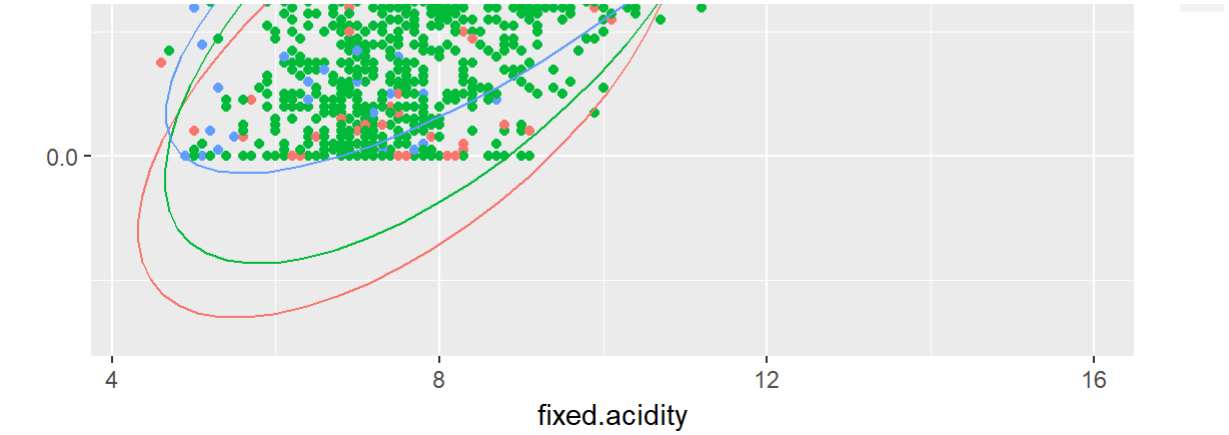
I think relation between free.shlfur.dioxide and total.sulfur.dioxide is strongest bivariate relation based on the correlation value.

Multivariate Plots Section

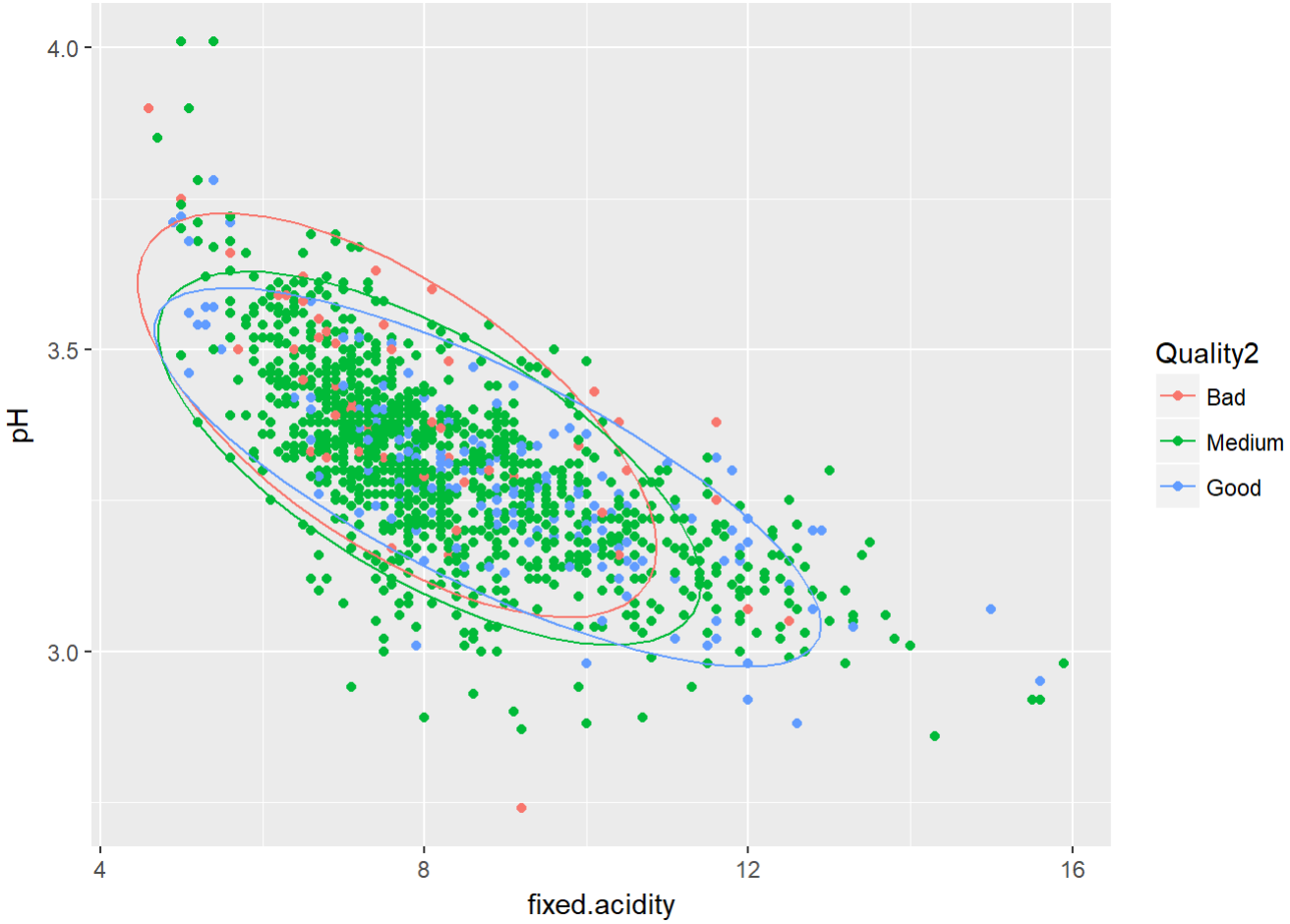
In this section we have created scatter plots between a pair of independent variable and mapped quality2 as a color variable. Further, we have tried to identify classes with the help of ellipses.

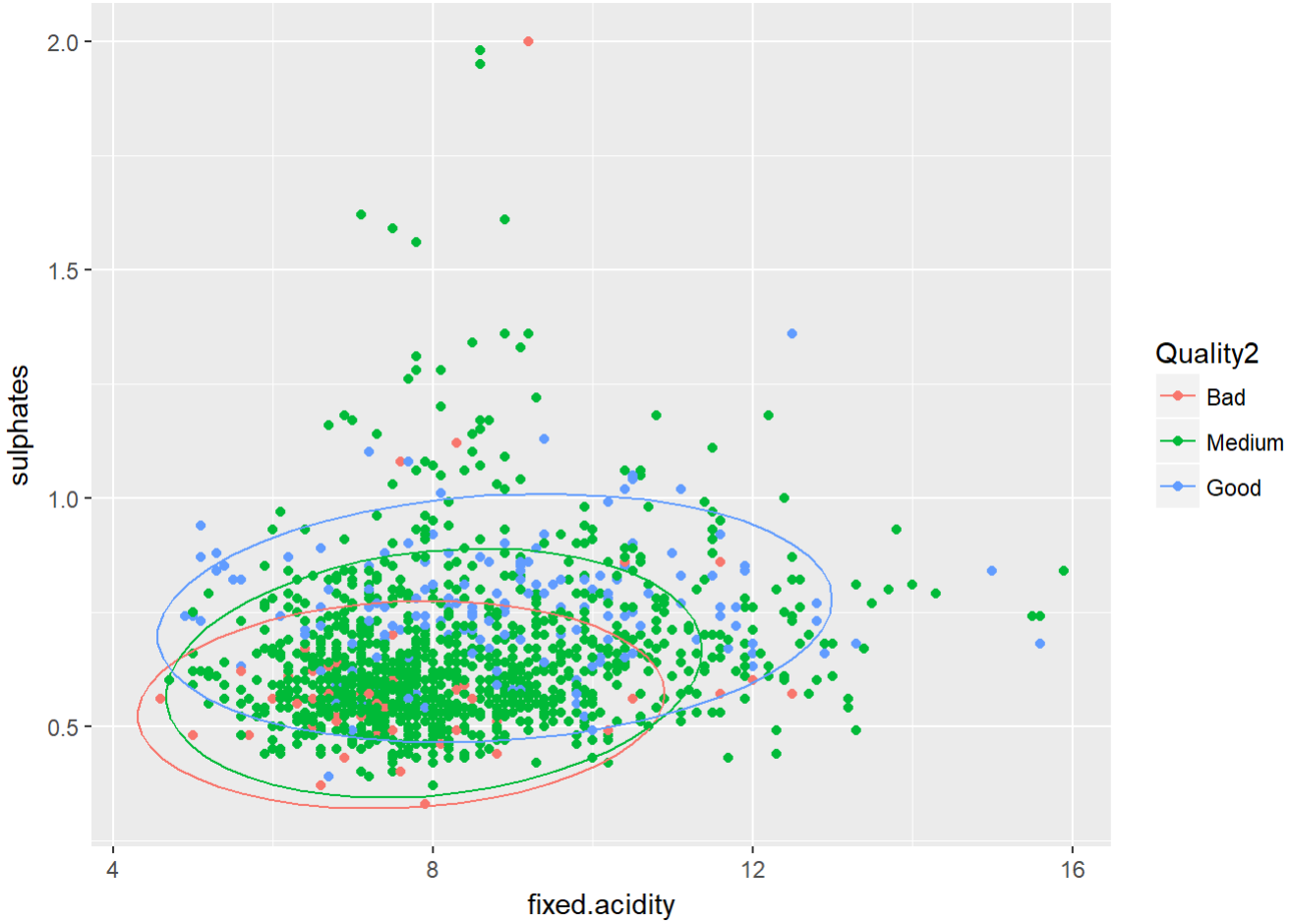

1.6
1.2
0.8
0.4
volatile.acidity
4
8
12
16
fixed.acidity
Quality2
Bad
Medium
Good

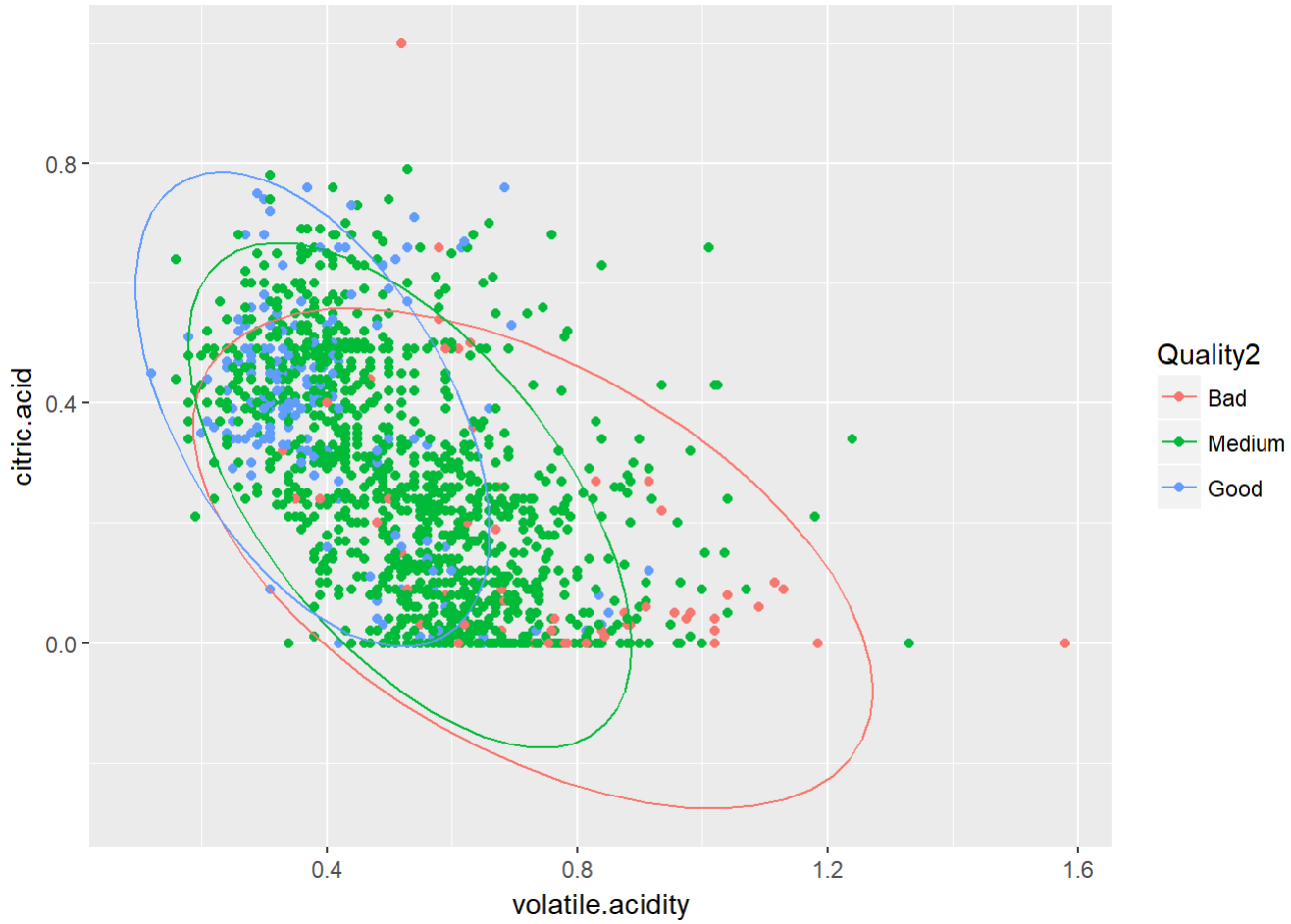

0.8
0.4
citric.acid
Quality2
Bad
Medium
Good

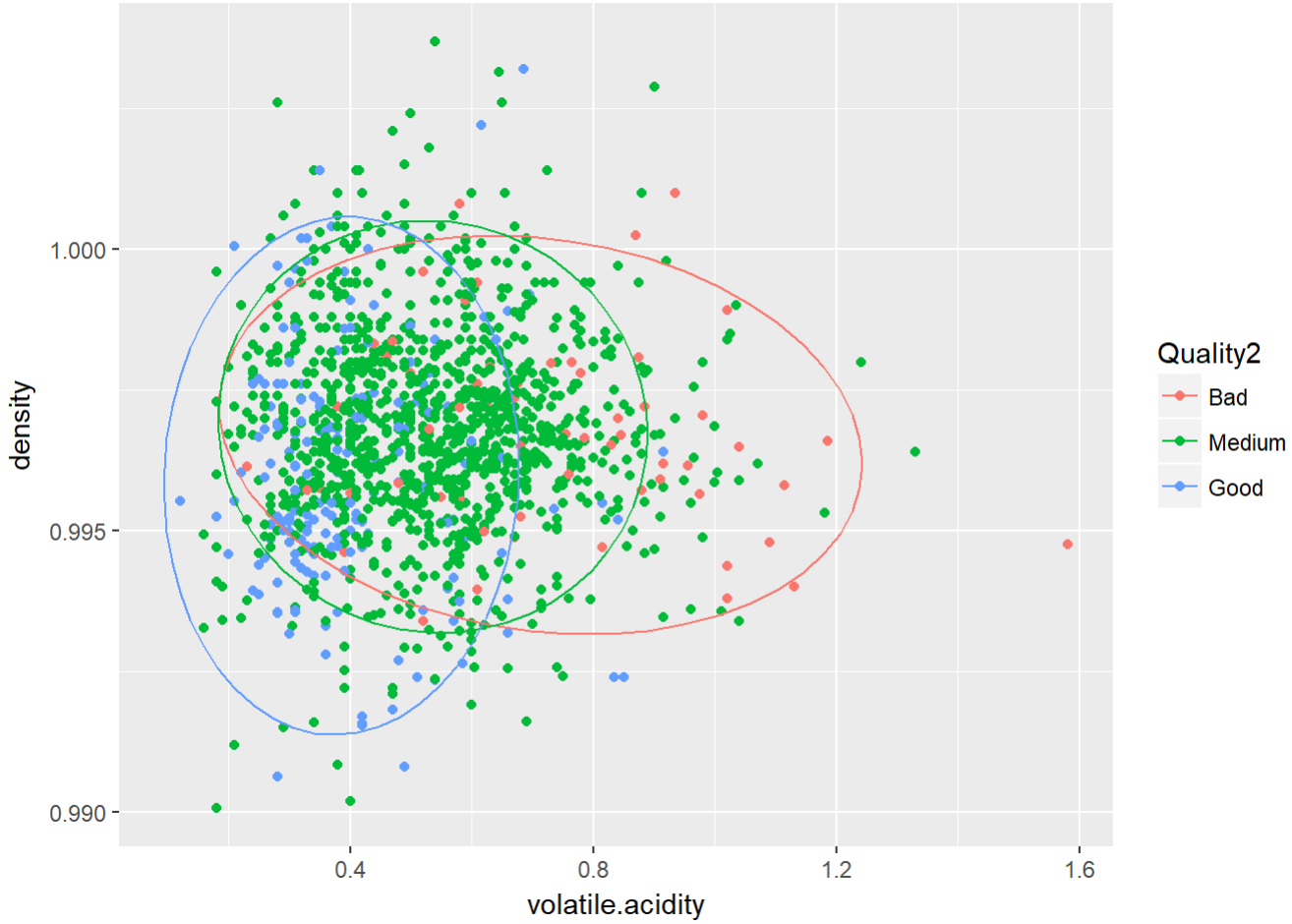

0.0
4
8
12
16
fixed.acidity

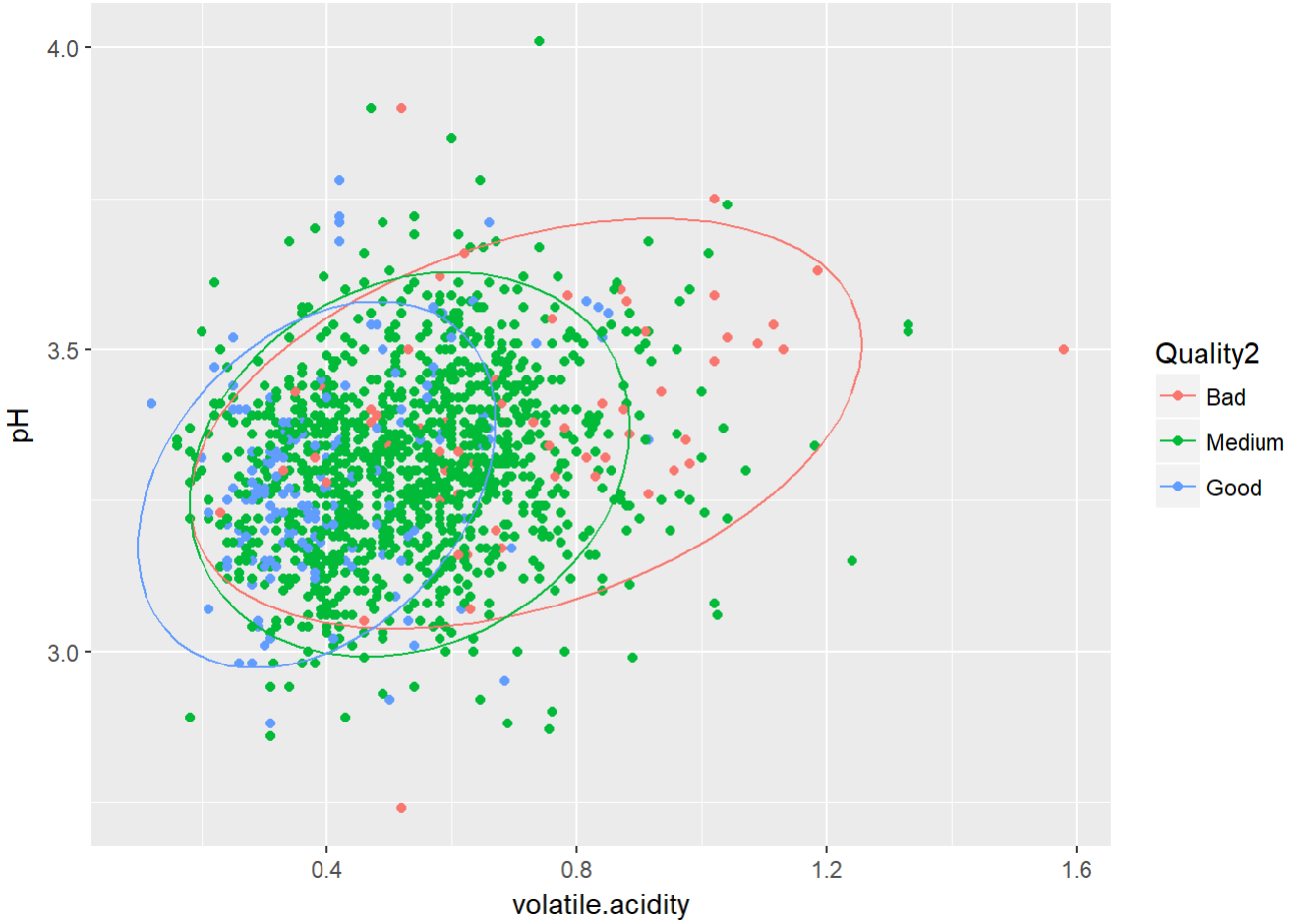

density
1.000
0.995
0.990
4
8
12
16
fixed.acidity
Quality2
Bad
Medium
Good

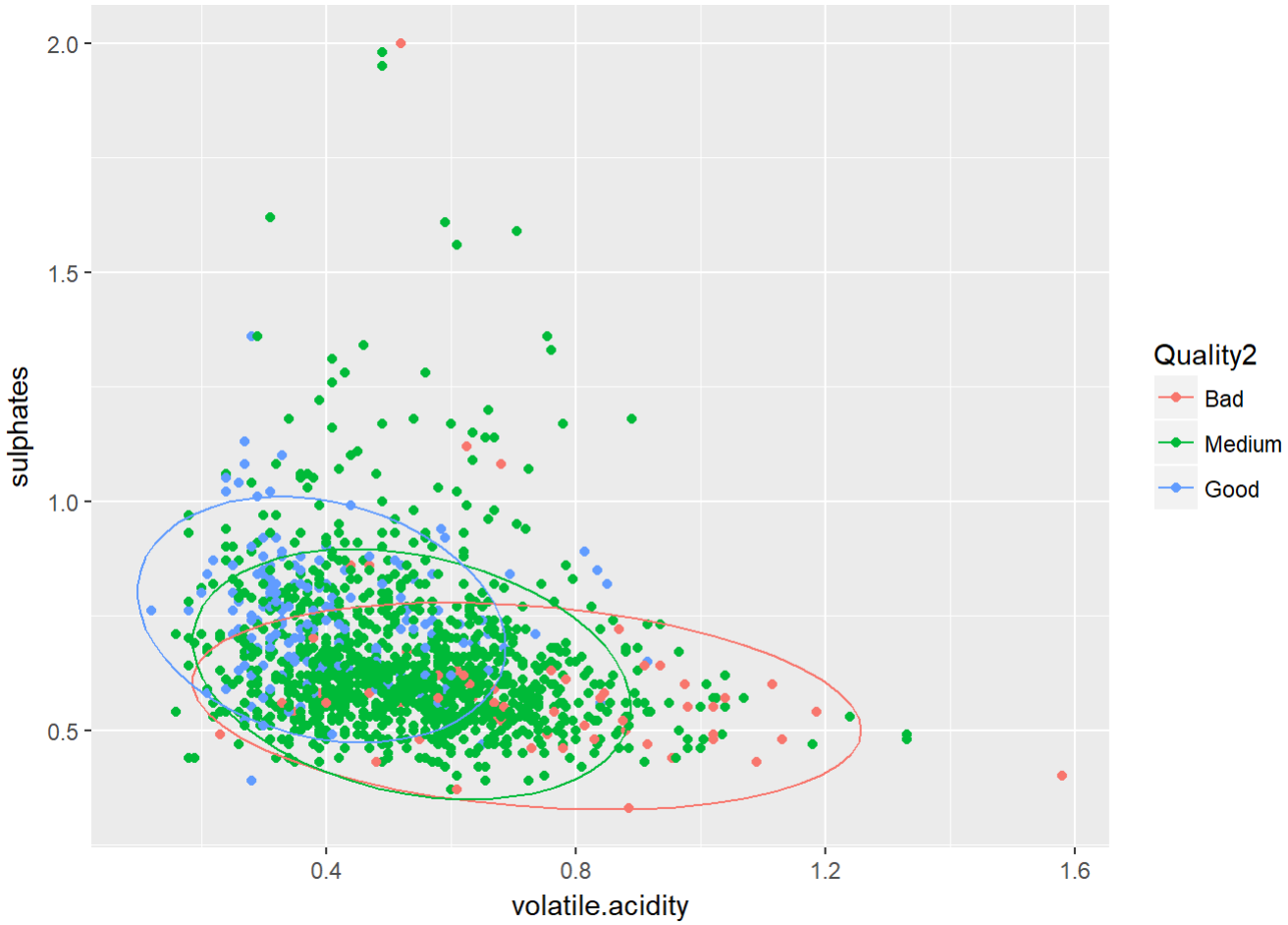

4.0
3.5
3.0
pH
4
8
12
16
fixed.acidity
Quality2
Bad
Medium
Good

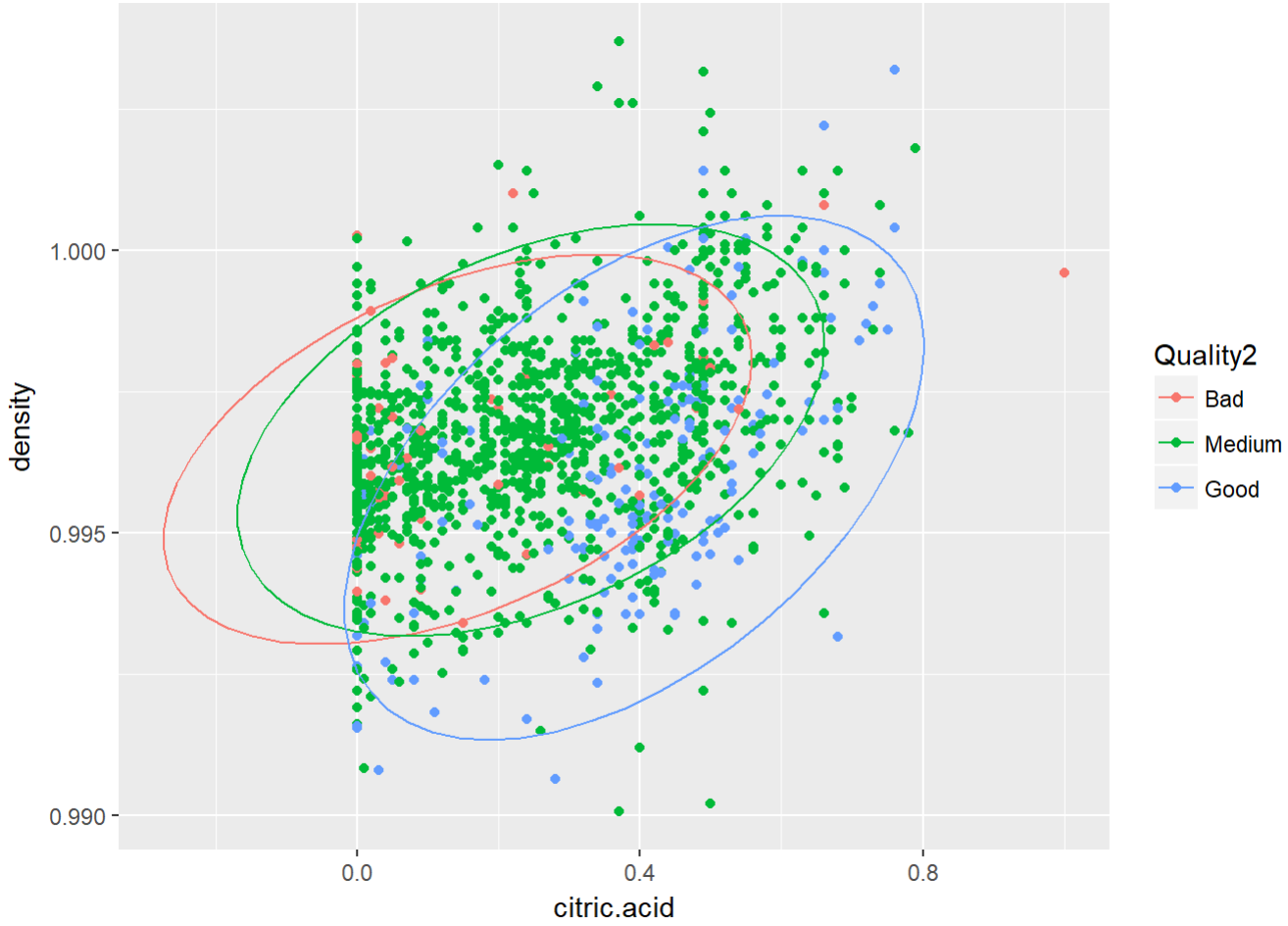

2.0
1.5
1.0
0.5
sulphates
4
8
12
16
fixed.acidity
Quality2
Bad
Medium
Good

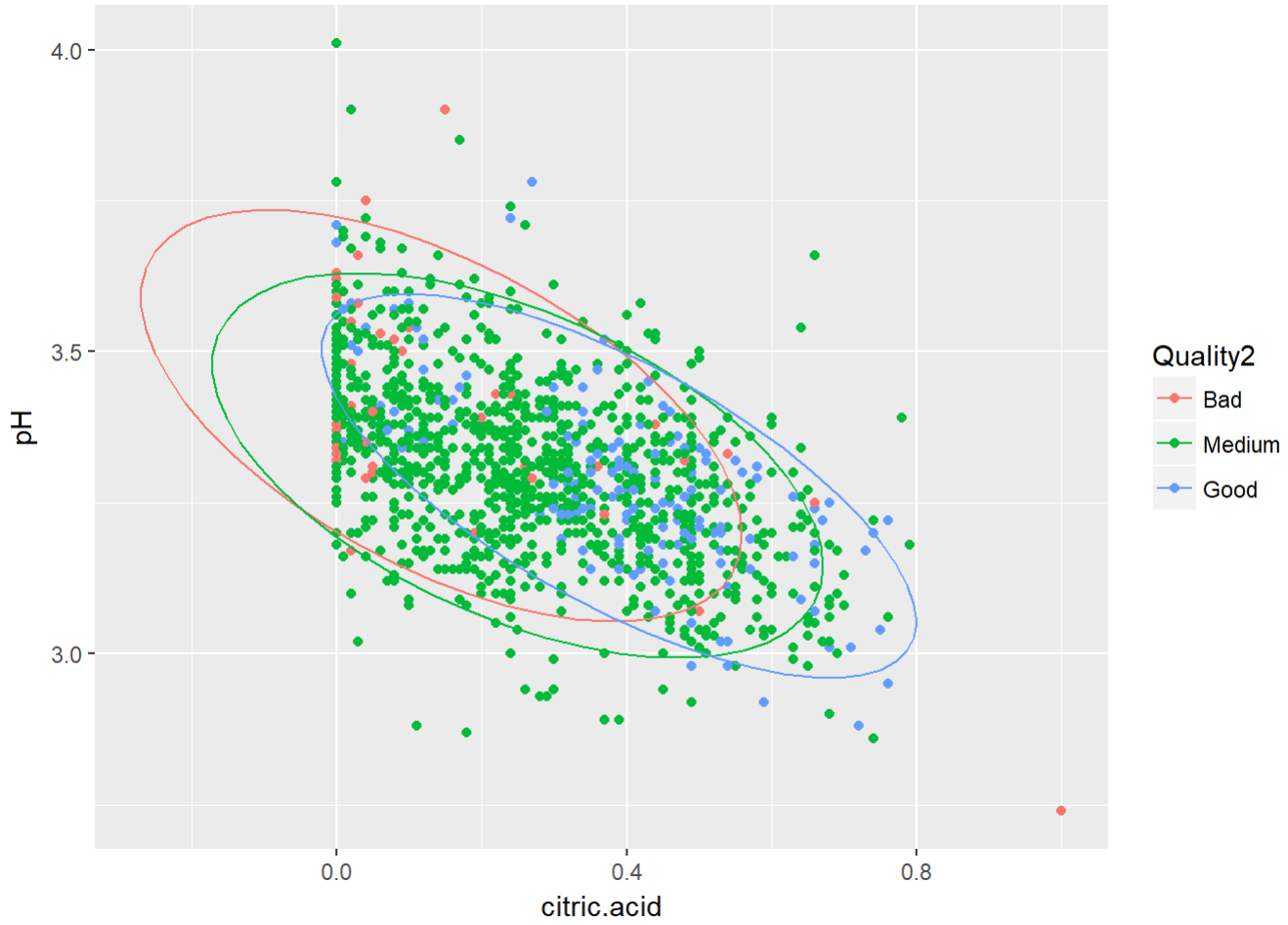

0.8
0.4
0.0
citric.acid
0.4
0.8
1.2
1.6
volatile.acidity
Quality2
Bad
Medium
Good

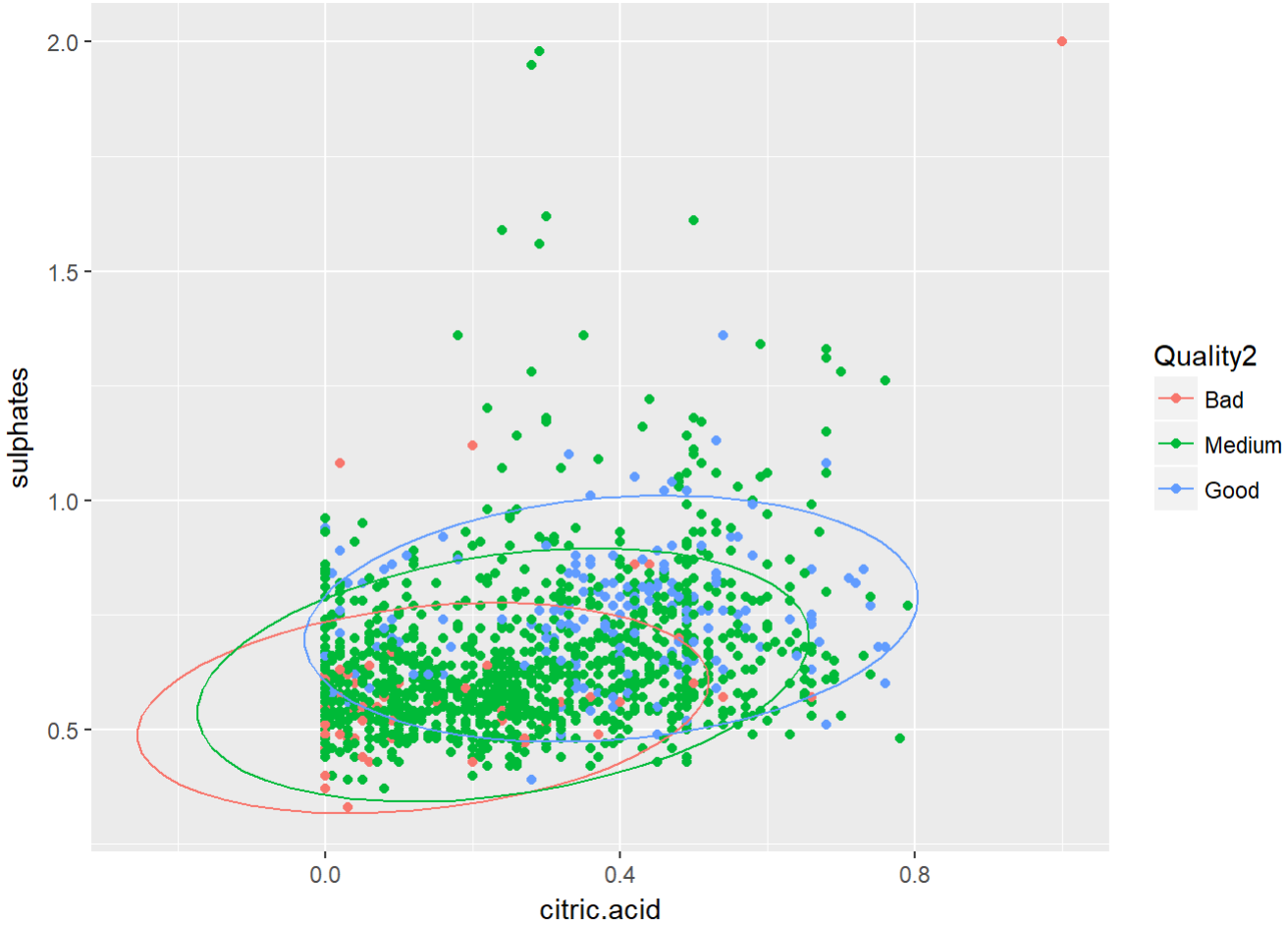

density
1.000
0.995
0.990
0.4
0.8
1.2
1.6
volatile.acidity
Quality2
Bad
Medium
Good


4.0
3.5
3.0
pH
0.4
0.8
1.2
1.6
volatile.acidity
Quality2
Bad
Medium
Good

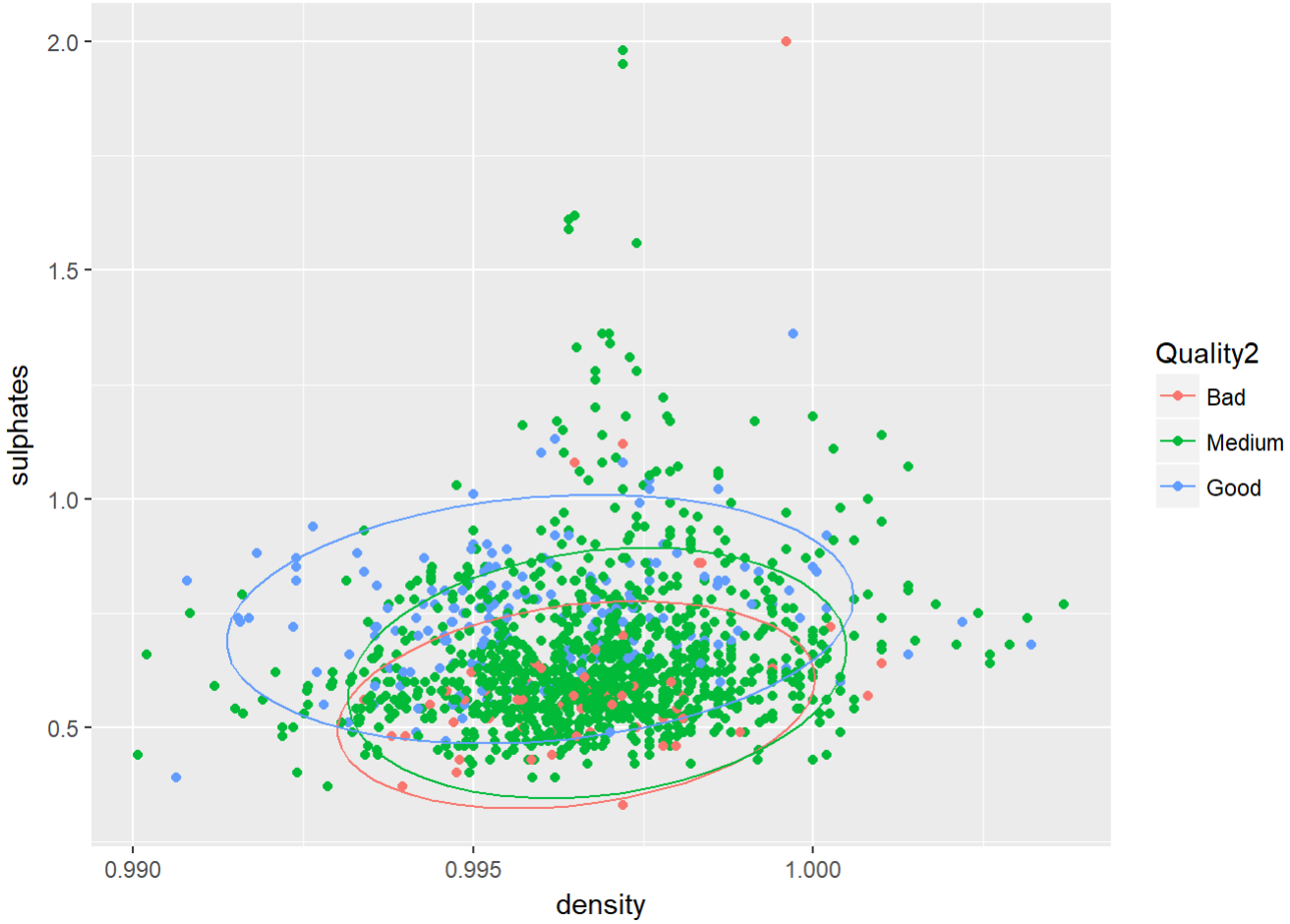

2.0
1.5
1.0
0.5
sulphates
0.4
0.8
1.2
1.6
volatile.acidity
Quality2
Bad
Medium
Good

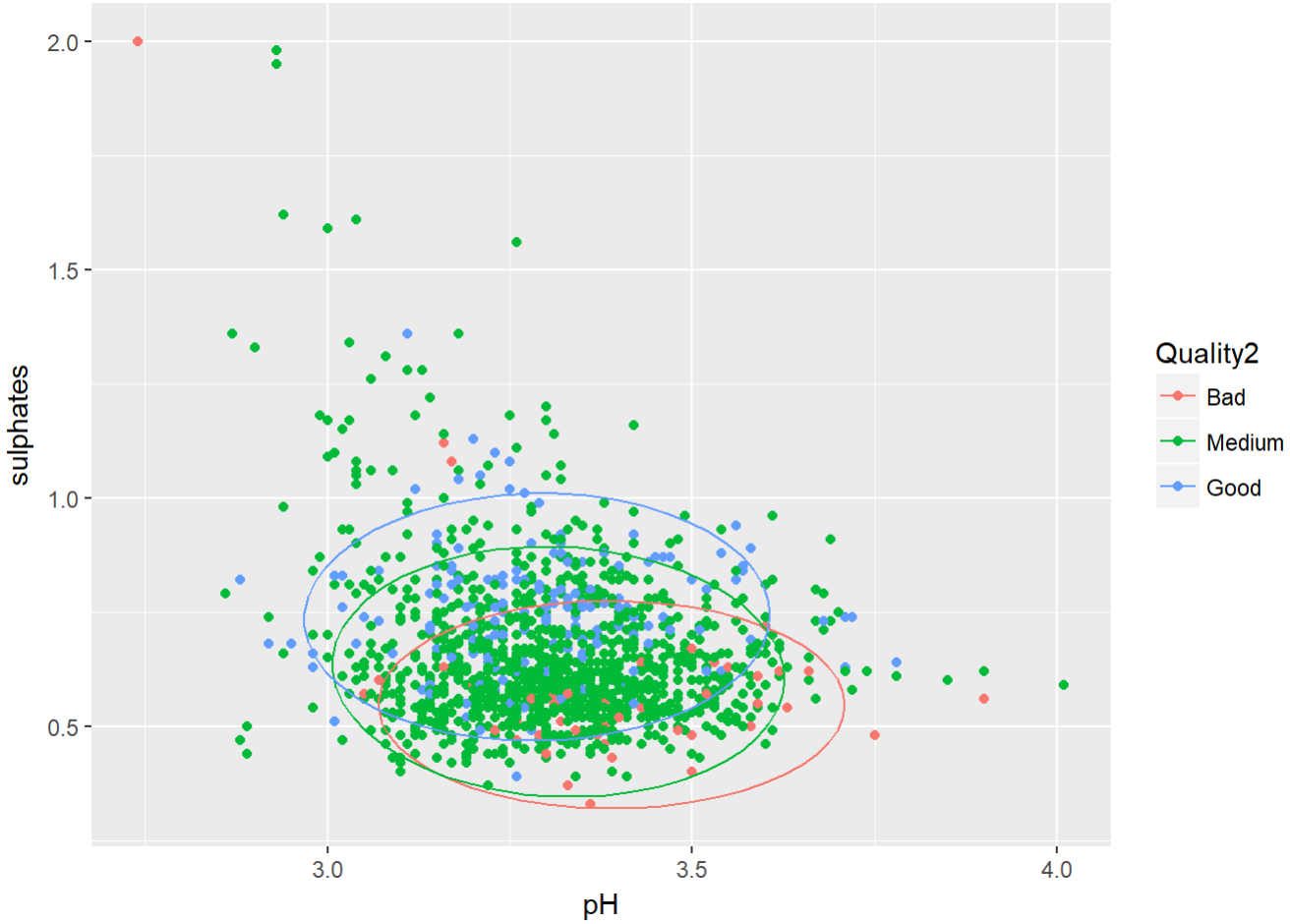

density
1.000
0.995
0.990
0.0
0.4
0.8
citric.acid
Quality2
Bad
Medium
Good


4.0
3.5
3.0
pH
0.0
0.4
0.8
citric.acid
Quality2
Bad
Medium
Good


2.0
1.5
1.0
0.5
sulphates
0.0
0.4
0.8
citric.acid
Quality2
Bad
Medium
Good


4.0
3.5
3.0
pH
0.990
0.995
1.000
density
Quality2
Bad
Medium
Good

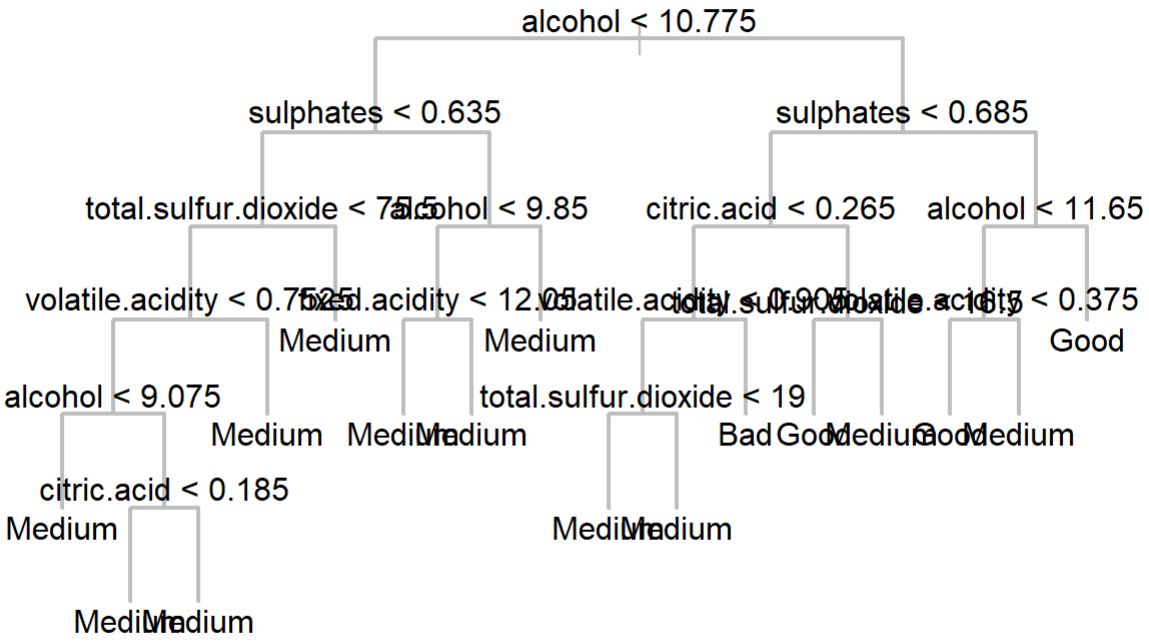

2.0
1.5
1.0
0.5
sulphates
0.990
0.995
1.000
density
Quality2
Bad
Medium
Good



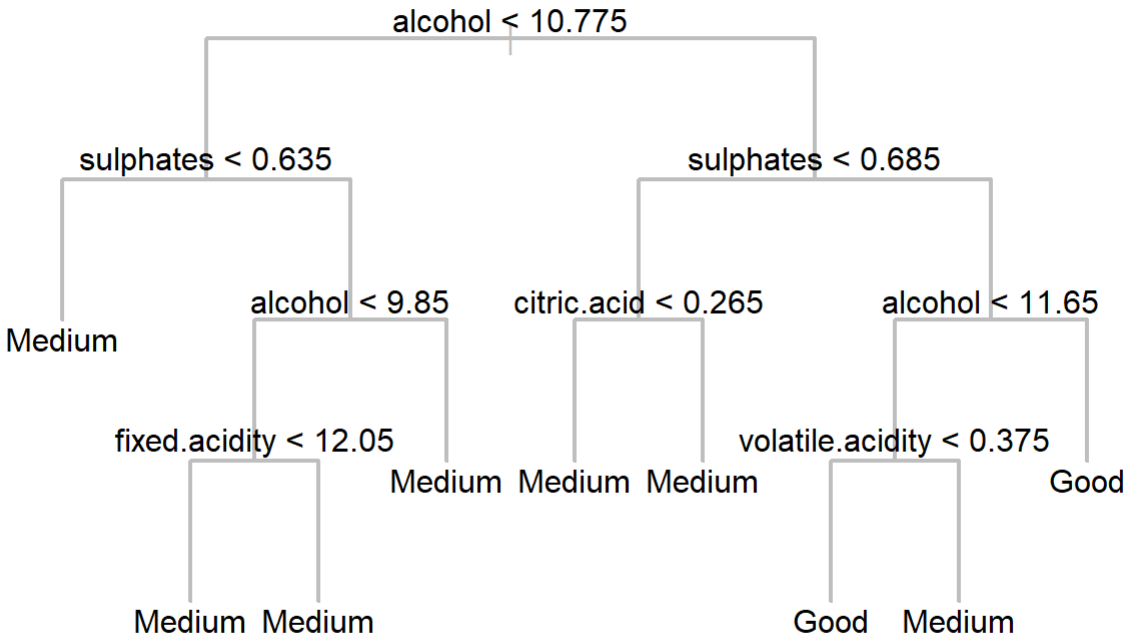
Decision Tree Model

Decision Tree, to classify wine in Quality classes has been created using all the features available in the dataset. Pruning is done to reduce the complexity of the tree.

Before pruning



After pruning



Accuracy before and after pruning respectively

```
## [1] 0.8275 0.8300
```

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?

After performing bivariate analysis, features that might influence the quality were identified.

For multivariate analysis, features that were idenfied in bivariate analysis were paired together and quality was mapped as a color on a scatter plot.

For instance, fixed.acidity and volatile.acidity were considered on X and Y axes and color was used to map the quality. While demarcation is not clear, cluster/class formation can been seen. Similar formation is seen for following pairs:

volatile.acidity vs fixed.acidity citric.acid vs volatile.acidity density vs volatile.acidity pH vs volatile.acidity density vs citric.acid density vs fixed.acidity citric.acid vs fixed.acidity pH vs citric.acid pH vs Density pH vs fixed.acidity sulphates vs fixed.acidity sulphates vs pH sulphates vs density sulphates vs citric.acid sulphates vs volatile.acidity

Were there any interesting or surprising interactions between features?

density and fixed.acidity are positively correlated. I did not expect this to happen.

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

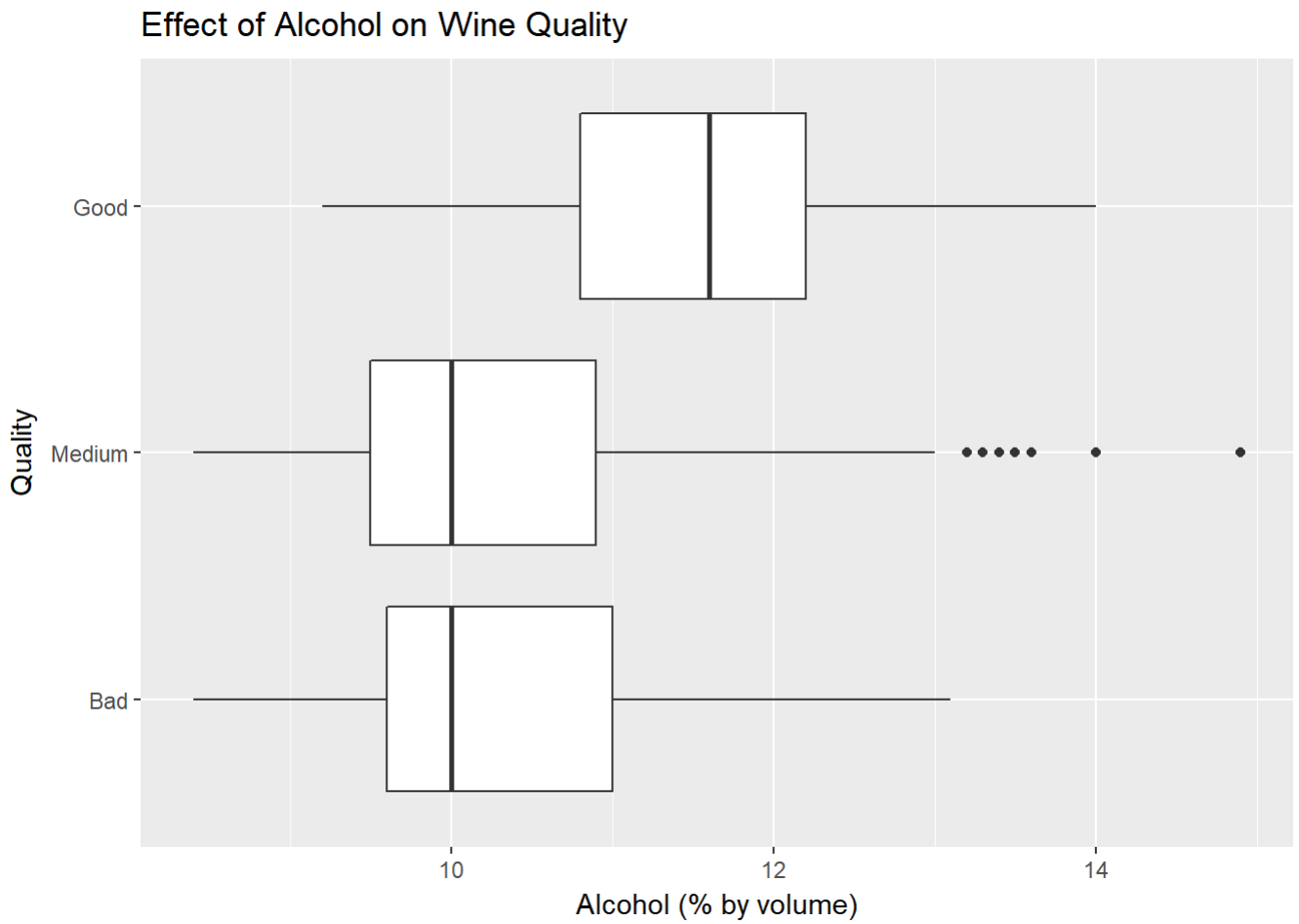
Yes. I created decision tree classifier to classify the wine with the given parameters into Good, Bad and Medium.

RandomForest, Bagging and boosting can be used to improve the performance however for this submission, I have used only basic decision tree without ensemble methods.

All the features in the dataset has been used since decision tree can perform automatic feature selection. Pruned Tree is also created so as to follow the principal of parsimony.

Final Plots and Summary

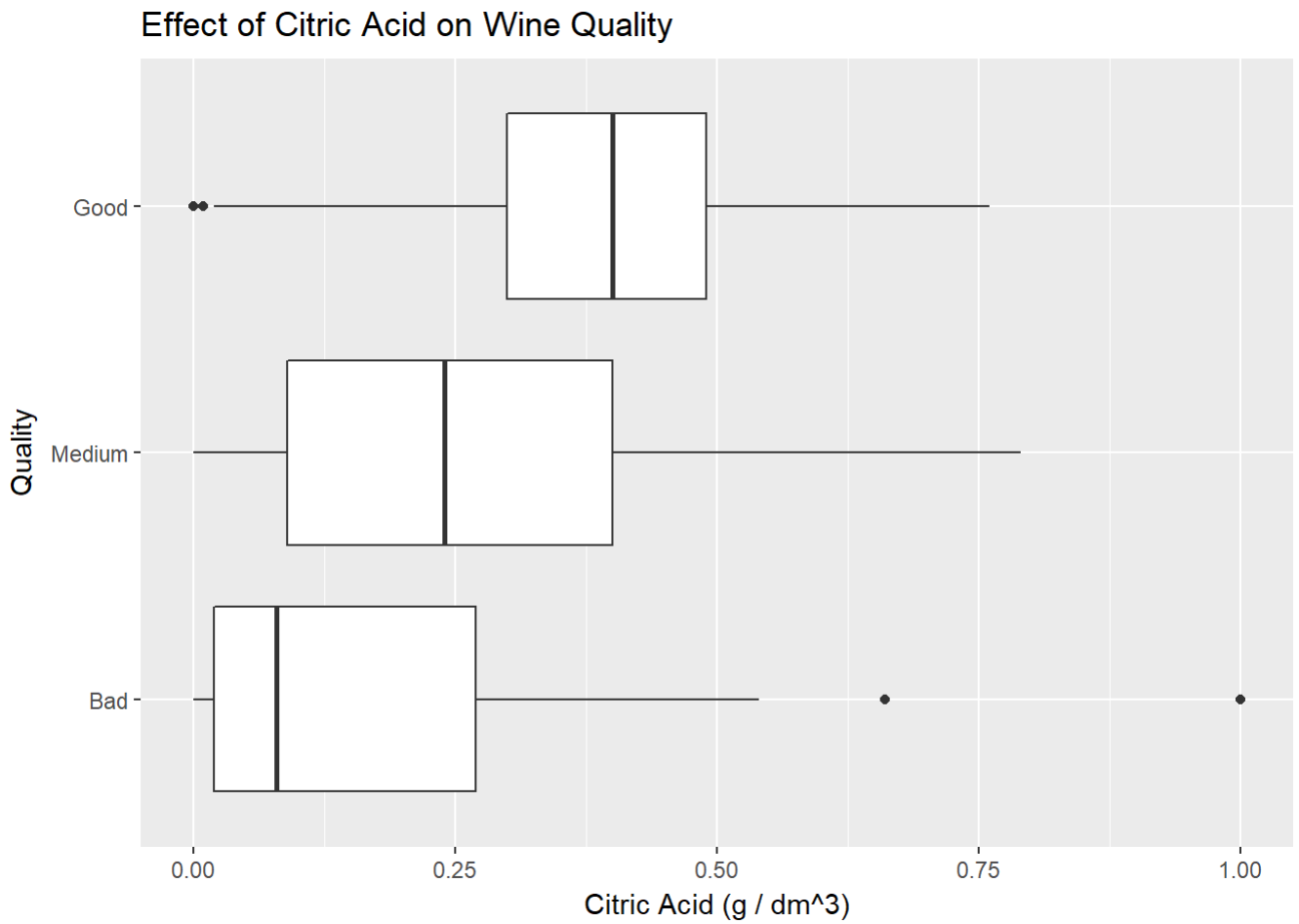
Plot One



Description One

We are exploring the information about wine. The fact that Alcohol is one of the most important content of wine makes the above plot essential. As seen from the above plot, quality of wine is better as Alcohol content is increased. There are few exceptions for medium category where even after having the high percentage of alcohol, their quality is labeled as medium. It will be interesting to study these outliers. However for this report, that investigation has not been performed.

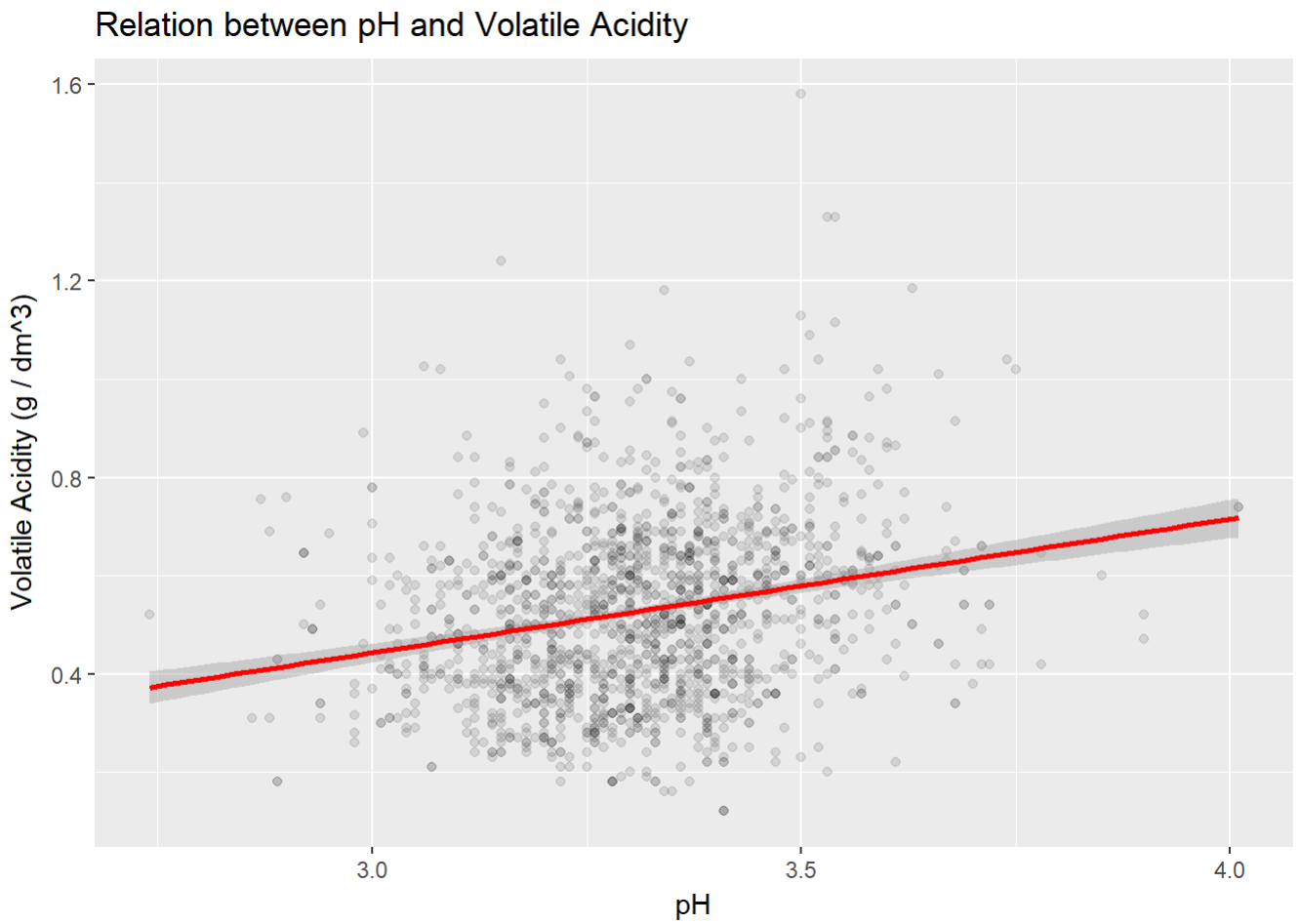
Plot Two



Description Two

Citric acid gives the sense of freshness. From the above plot we can see that wine quality improves as the amount of citric acid goes up. While this is the general trend, we have outliers on both sides of the coin. We have bad wines with very high citric acid and good wines with almost no citric acid.

Plot Three



Description Three

I am highlighting this plot because the relation is completely opposite to what I had expected. I though, any kind of acidity will be negatively related to pH. However, we can see that volatile.acidity and pH has a positive correlation. While relation is not very strong, general trend is positive.

Reflection

The red wine dataset has 1599 observations of wines with 12 features. All the variables are numerical variables except quality. Quality is a discrete variable.

Most wines belong to the quality value of 5 or 6. Few belong to 2 or 3 and few belong to 7 or 8. Thus, wine quality has nearly normal distribution. I believe if we have more number of wine observations, the distribution can come more closer to normal distribution.

The central idea of the analysis is to determine the variables that affect the wines. We did reveal interesting relations between wine quality and other variables. We also created a decision tree classifier from the available features in the dataset. After pruning the tree, the classfication of accuracy is almost 83 %. One of the limitation with this model is way quality variable was captured in the dataset. Quality variable is based on sensory data. I am sure multiple wine testers would have been consulted to create a dataset. However, this might not give good generalization to the model and have some bias.

One of the problems that I faced during the project was regarding formation of model. Decision Tree classifer, though have a decent accuracy, could not classify any Wine as Bad. This is because the dataset itself seems to have a bias. There are only 63 entires that could be labeled as bad. While there are 217 Good quality wines and 1319 Medium quality wines.

In the next iteration of the project, it would be interesting to see how a predictive models works if one does not classify wines into Bad, Medium and Good. Rather the original scale of wine quality from 1 to 10 is used in predict process.

Citation

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.