Red Wine Data Exploration

This report explores quality and other attributes of red wine for 1599 red wine samples and 14 variables

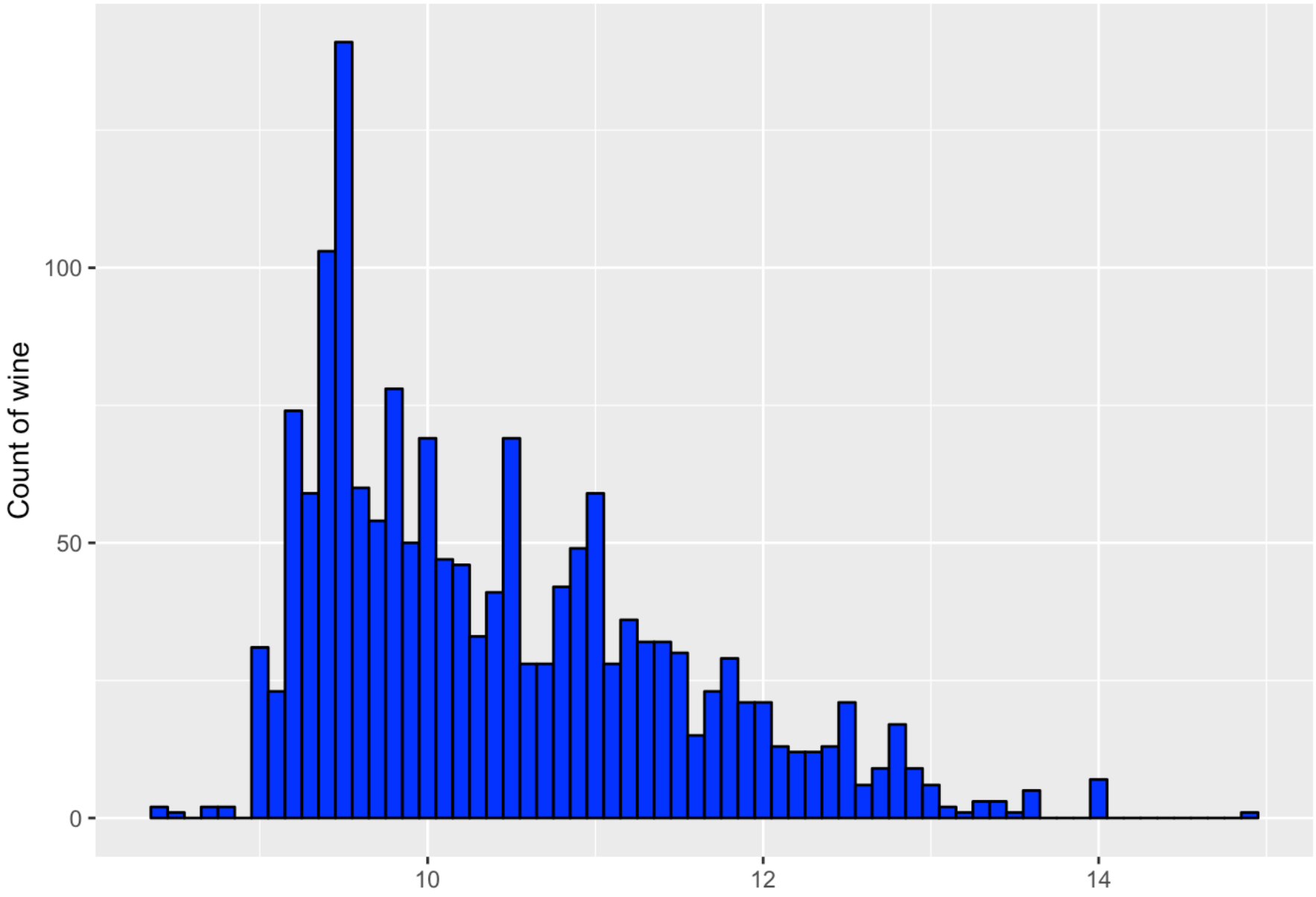
```
##        X           fixed.acidity   volatile.acidity  citric.acid
##  Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000
##  1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090
##  Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260
##  Mean   : 800.0   Mean   : 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      qualityFactor
##  Min.   : 8.40   Min.   :3.000   3: 10
##  1st Qu.: 9.50   1st Qu.:5.000   4: 53
##  Median :10.20   Median :6.000   5:681
##  Mean   :10.42   Mean   :5.636   6:638
##  3rd Qu.:11.10   3rd Qu.:6.000   7:199
##  Max.   :14.90   Max.   :8.000   8: 18
```

Univariate Plot Section

Lets explore some variables

Alcohol

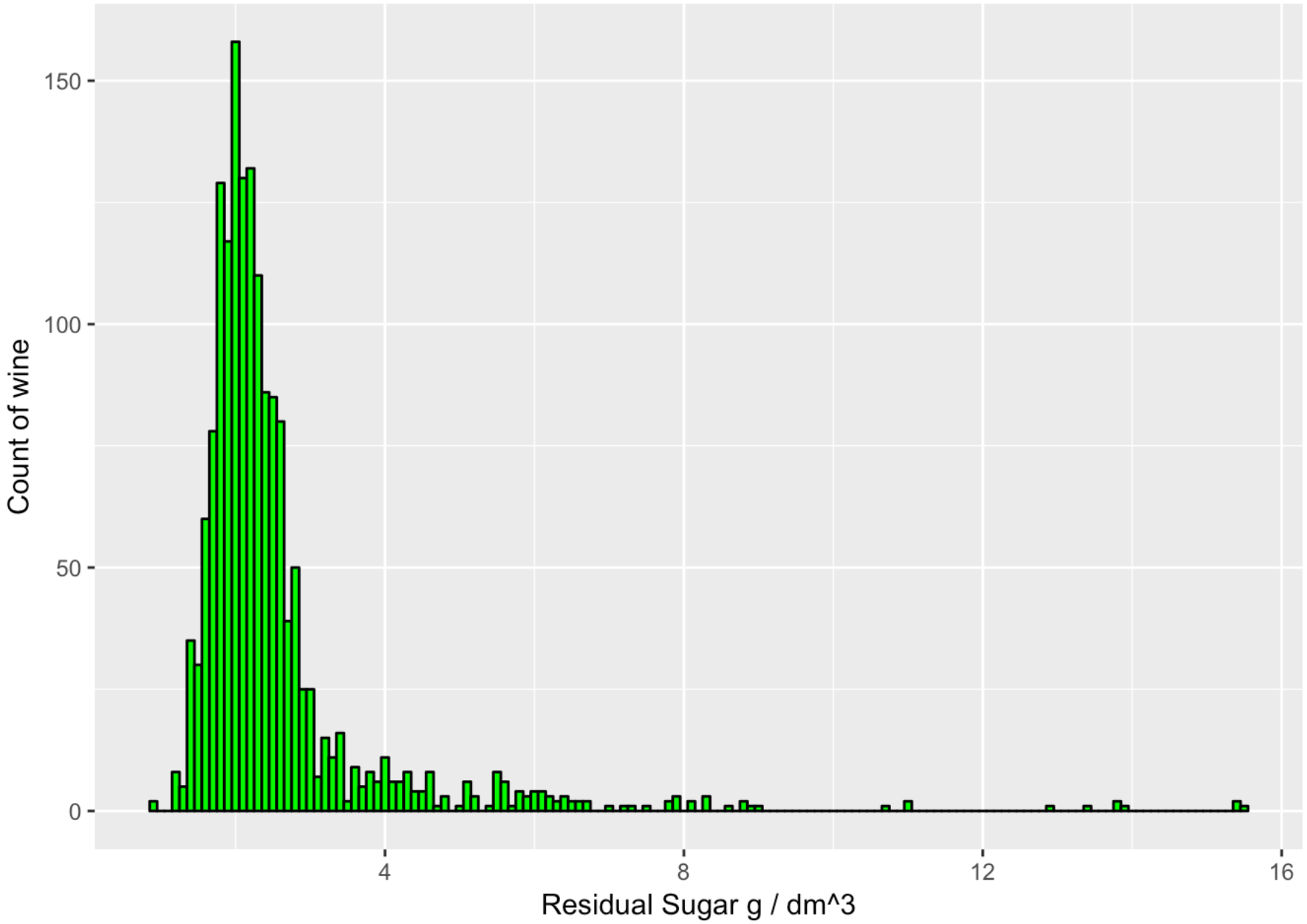
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```



The distribution appears left skewed with alcohol percent peaking approximately at 9.5 and 75% of the wines having alcohol less than 11.1

Residual Sugar

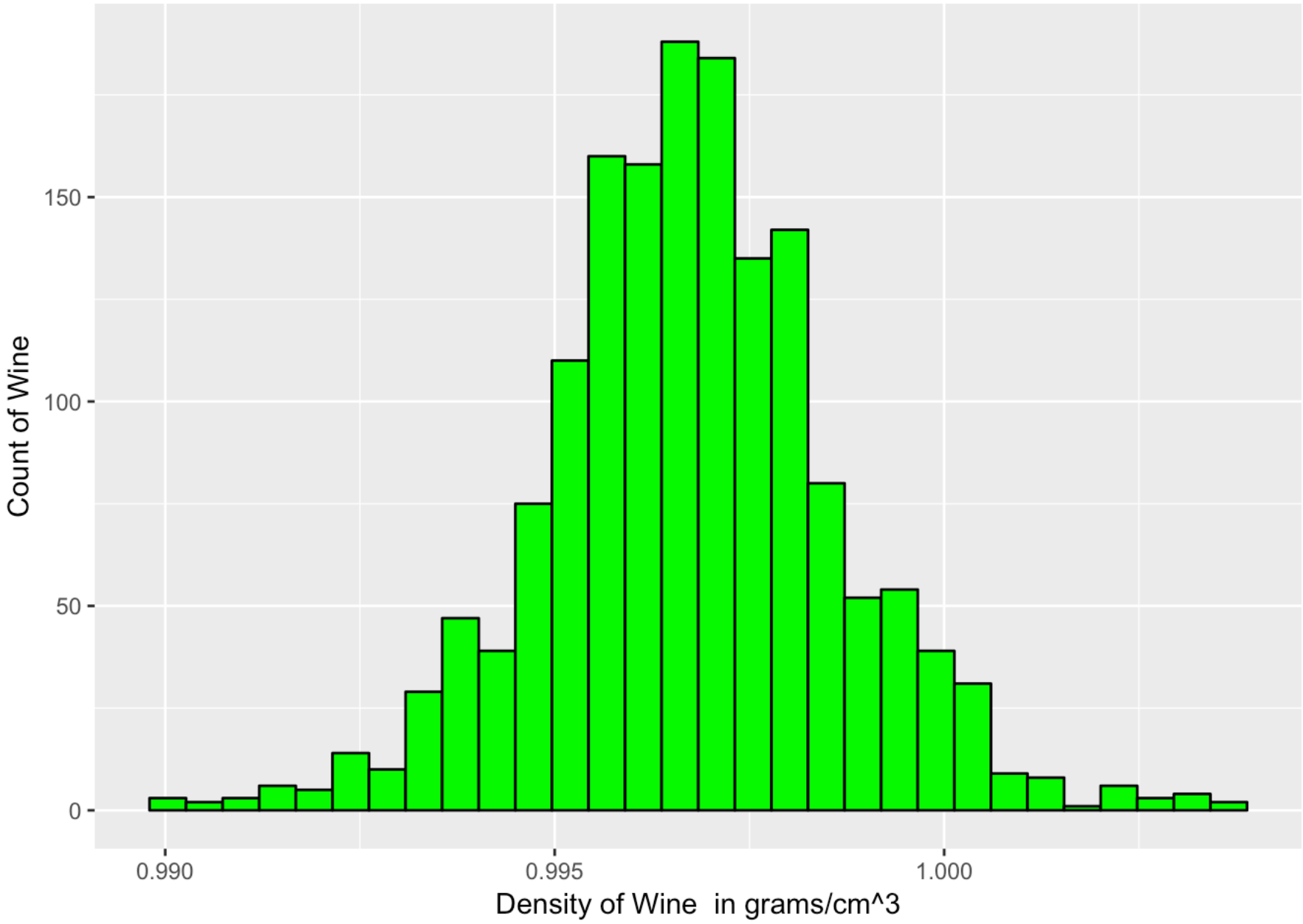
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```



Most of the wines have Residual Sugar value less than 4 grams/liter. We do have some wines where the residual sugar is high. I would like to see how these wines are rated. Are sweeter wines rated higher? Is sweetness a criteria in rating a wine?

DENSITY

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

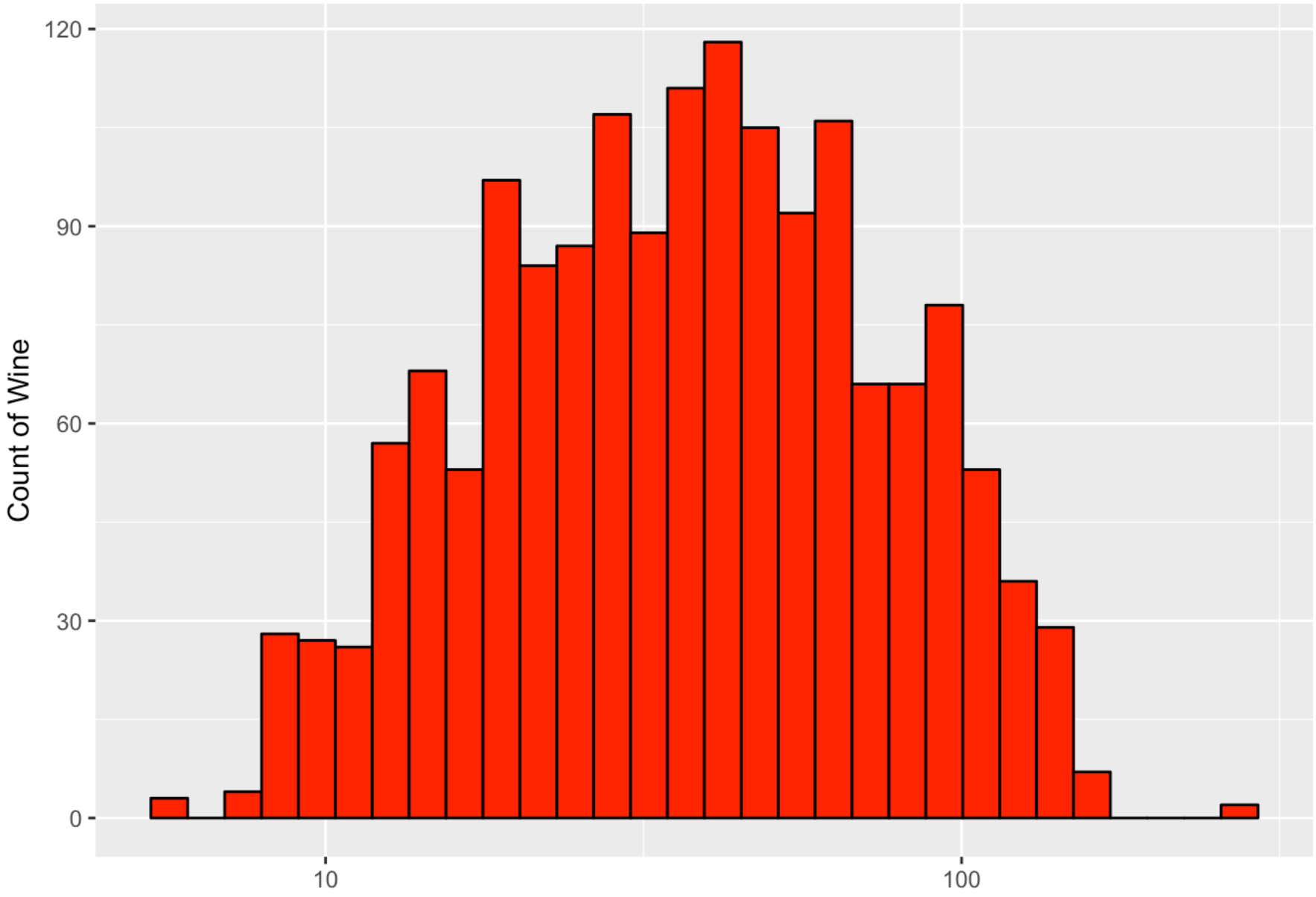


Pretty Normal distribution of Density. Mean .9967, Median .9968

Total Sulfur Dioxide

Represents the amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 mg/L, SO2 becomes evident in the nose and taste of wine

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

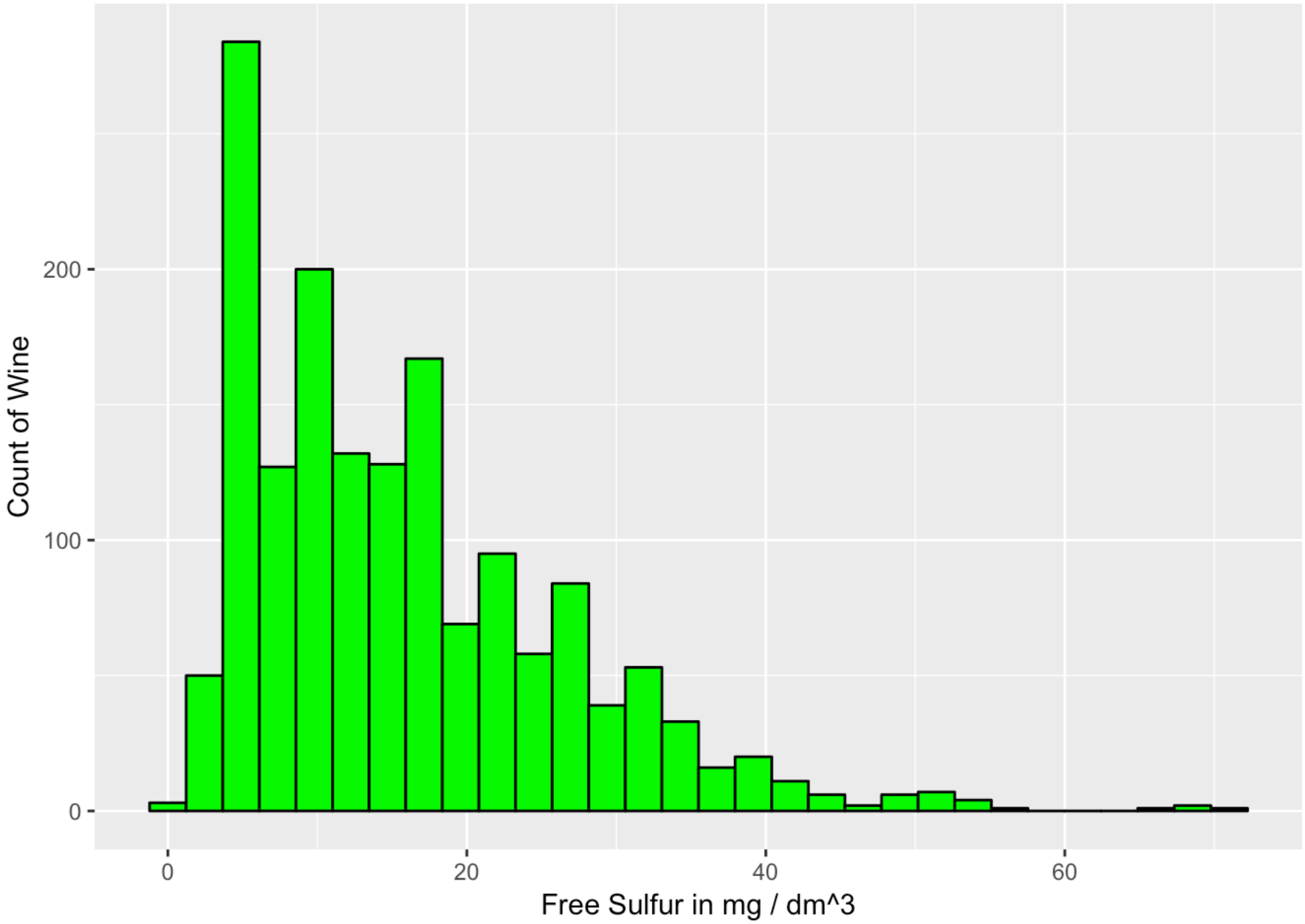


As can be seen above, majority of the wines are low in concentration of sulfur since So2 in high concentration can be detected in the taste of wine

Free Sulfur Dioxide

Prevents microbial growth and oxidation of wine

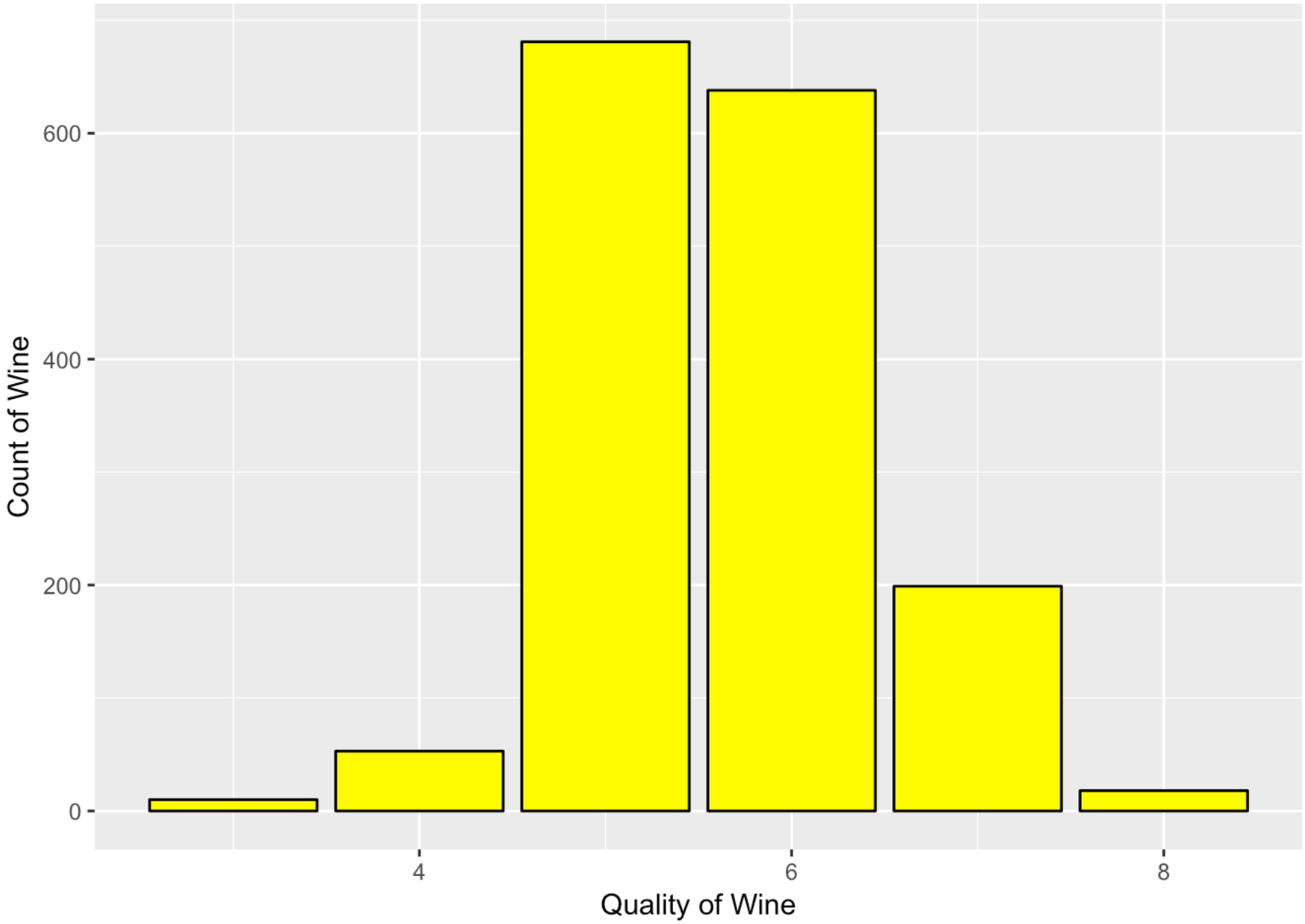
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
```



I would like to explore how amount of free sulfur impacts the quality of wine since sulfur helps with wine antioxidation and prevents microbial growth. The above

Quality

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

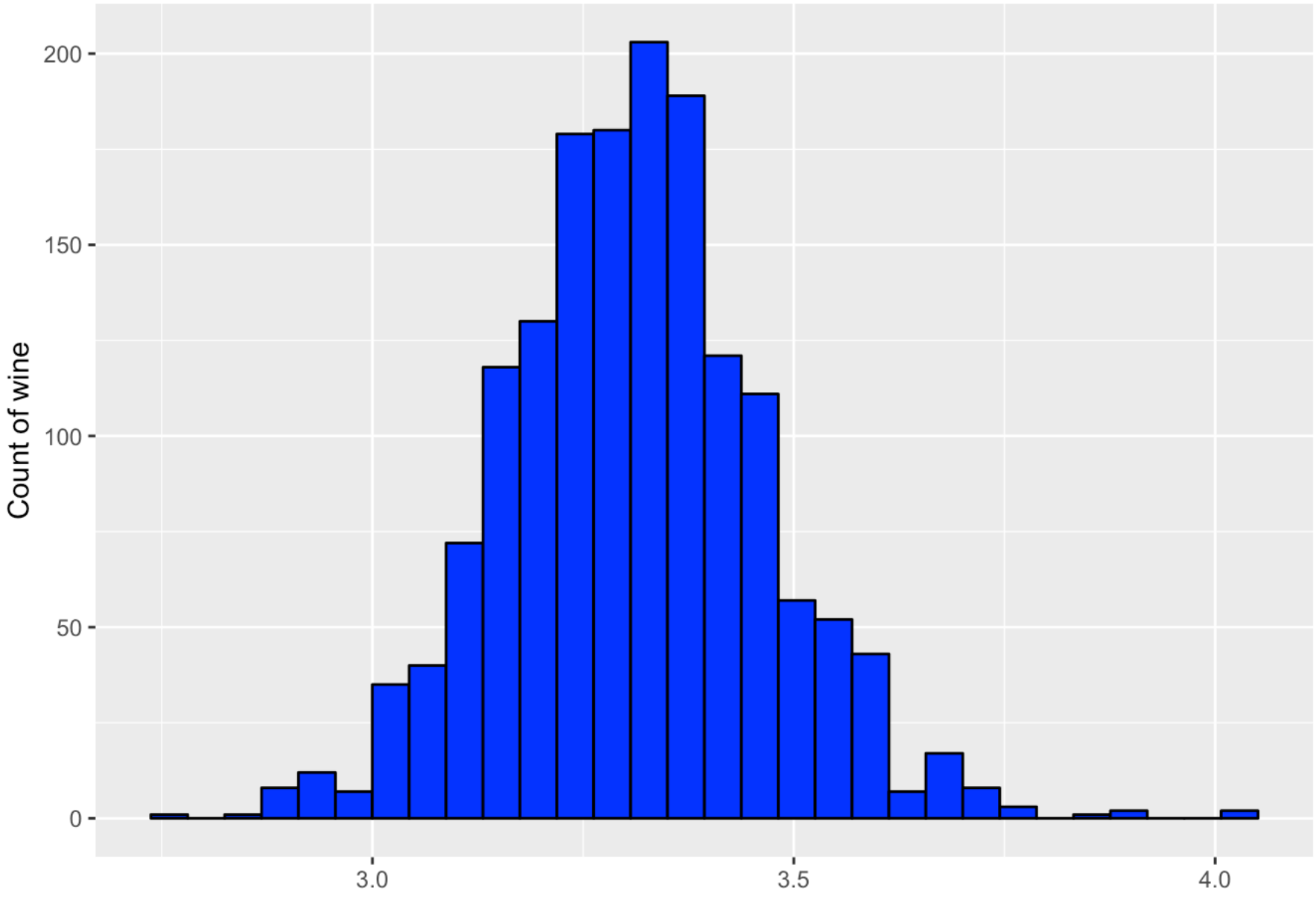


Most of the wines are of quality 5 and 6 and the range of wine is [3,8]

pH

pH tells how acidic wine is . Range is from 0 (very acidic) to 14(very basic). 7 is neutral

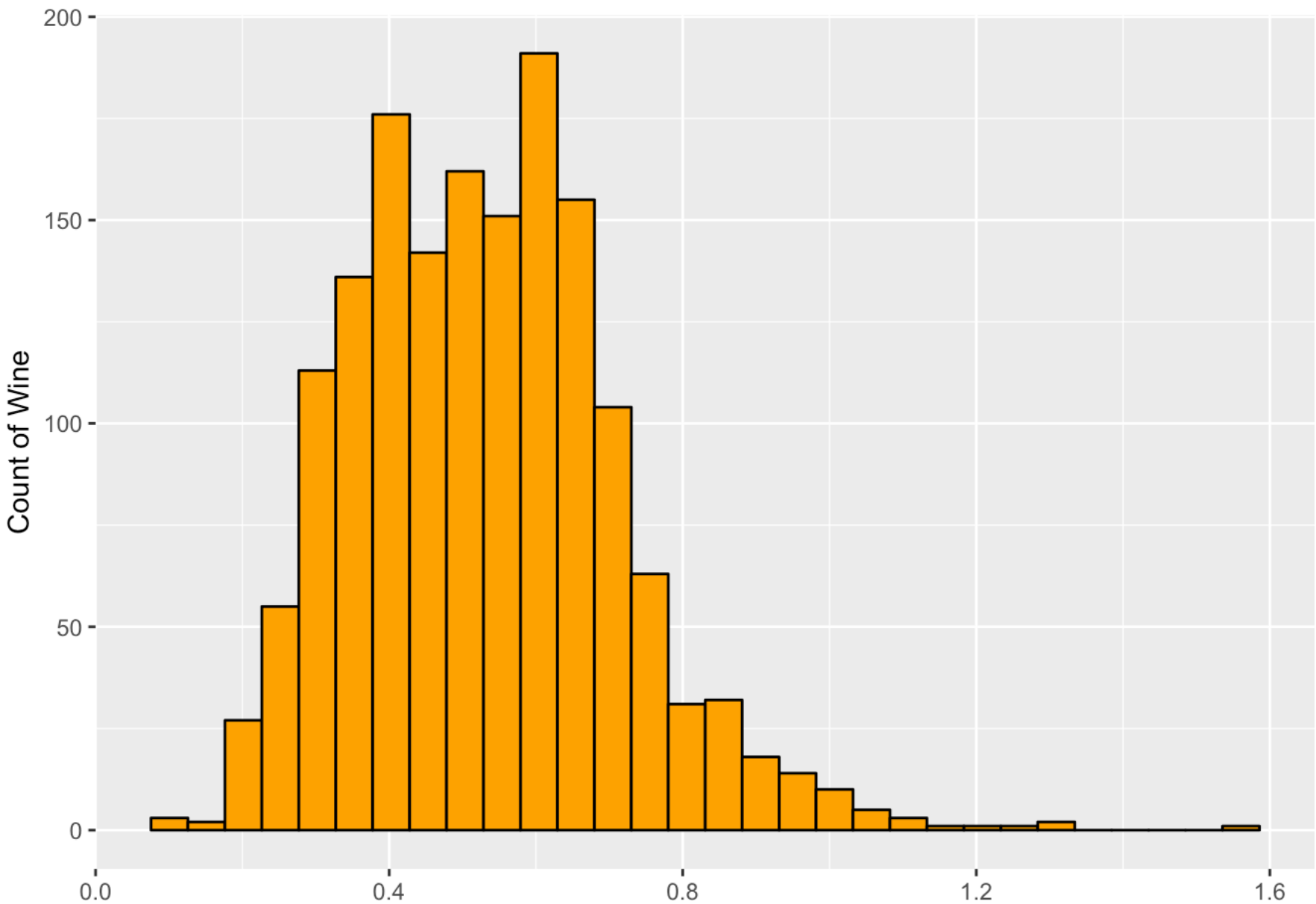
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   2.740   3.210   3.310   3.311   3.400   4.010
```



Looking at the above graph, All the wines are acidic in nature

Volatile Acidity

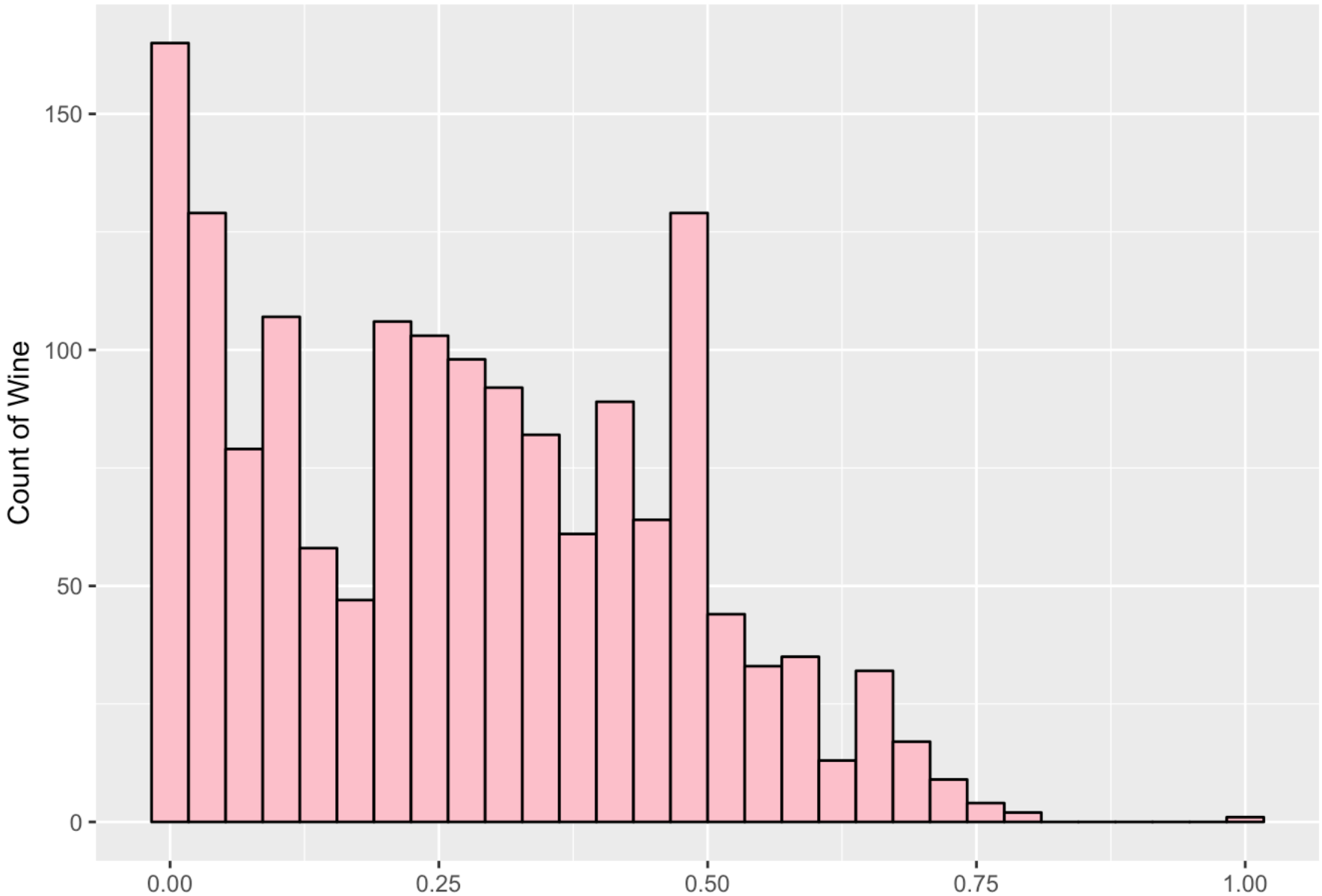
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```



Most of the wines have volatile acidity less than 1 gram/liter which is towards the lower side. Higher values of volatile acidity can give wine a vinegar kind of taste.I am curious to find out if volatile acidity impacts the quality rating of the wine.I believe it should. There also appears to be some outliers on the right side of the graph

Citric Acid

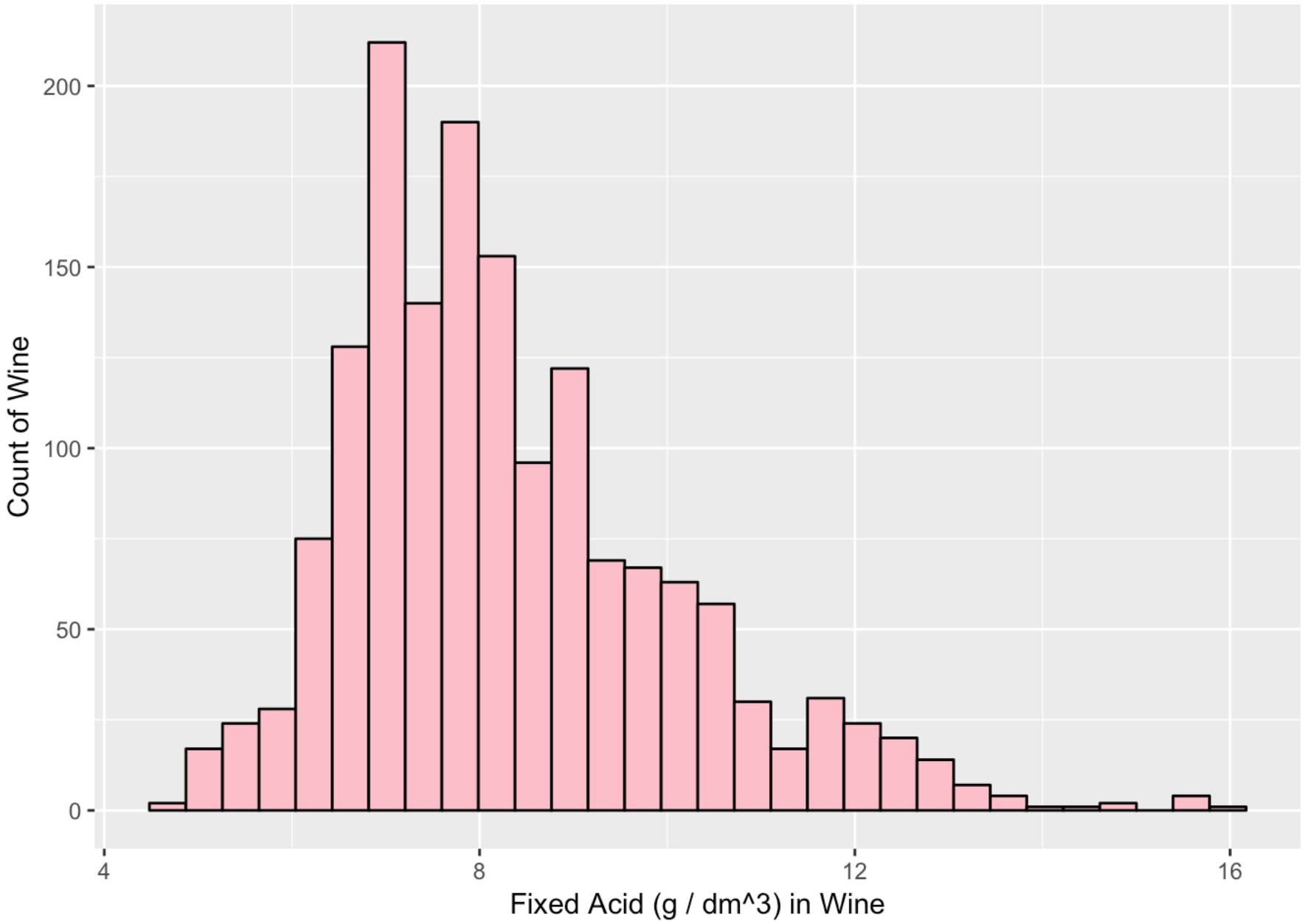
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   0.090   0.260   0.271   0.420   1.000
```



Citric Acid adds freshness to the wine. Most of the wines have citric acid less than 0.75. I am curious to see if citric acid impacts the quality rating since it adds freshness to the wine.

Fixed Acidity

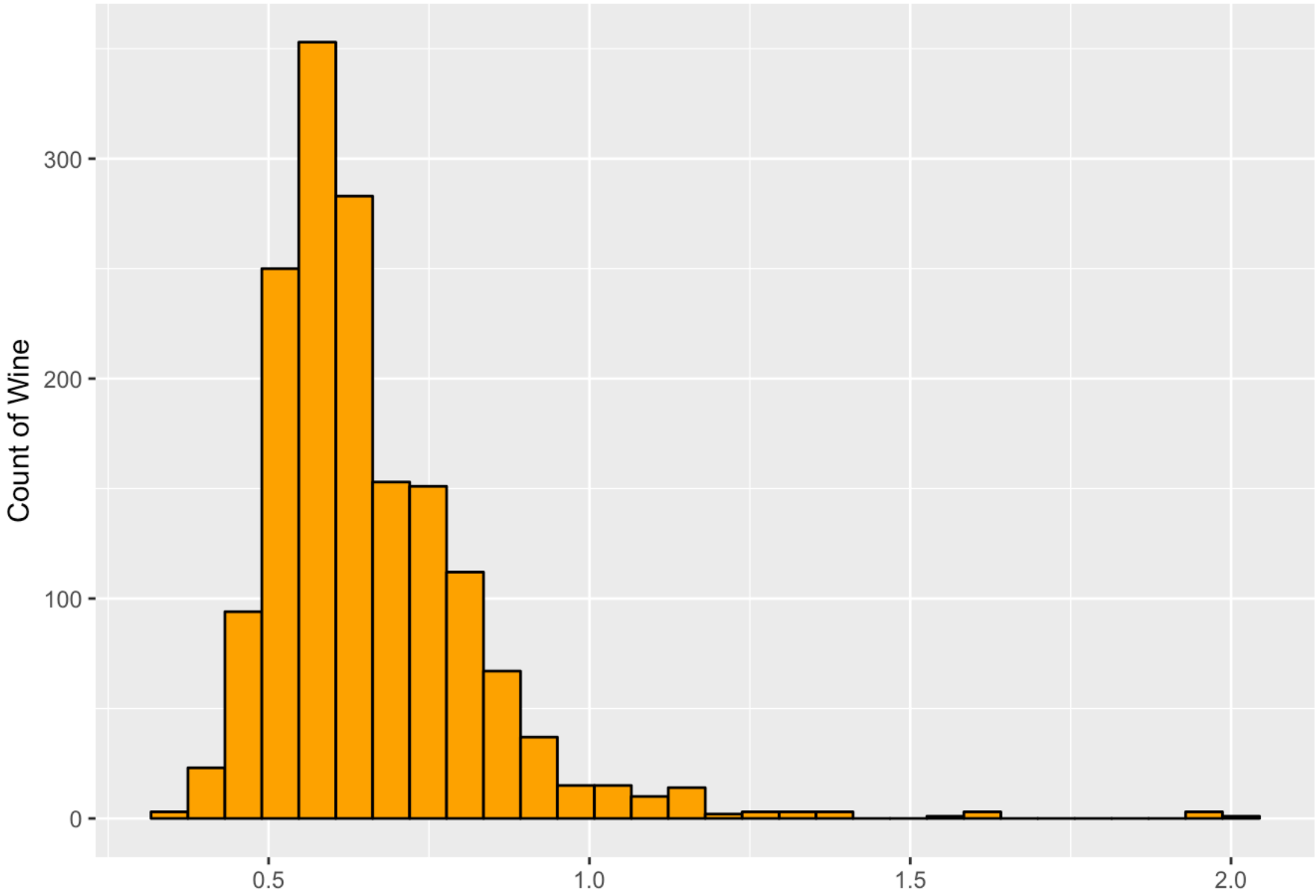
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
```



Bell shaped graph with median value 7.9 and mean 8.32

Sulphate

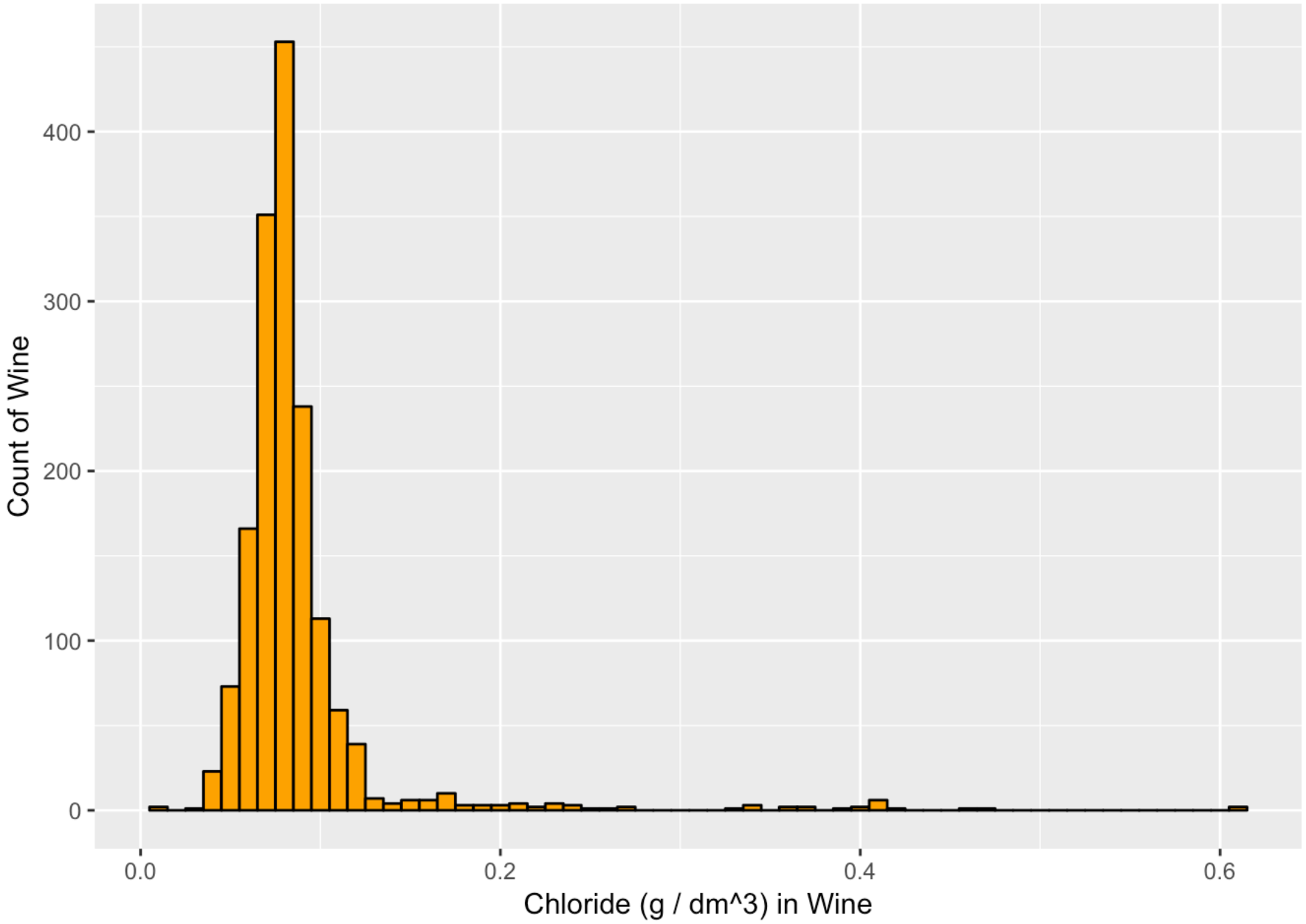
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```



Most wines have sulphates less than 1. Since sulphate contributes to SO2 levels that helps with antioxidation and antimicrobial, I am curious to see how sulphate concentration impacts the quality rating of wine

Chloridres

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



Most of the wine chloride is under .4. Few outliers

What is the structure of the dataset?

The dataset has 14 variables and 1599 observation. Quality is numerical and all other variable are non numerical

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

I would like to explore how quality is dependent on pH, density, alcohol

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

investigation into your feature(s) of interest?

I think citric acid and volatile acidity may also impact quality. So I would like to explore these as well.

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

Yes , I created a quality factor variable

Of the features you investigated, were there any unusual distributions?

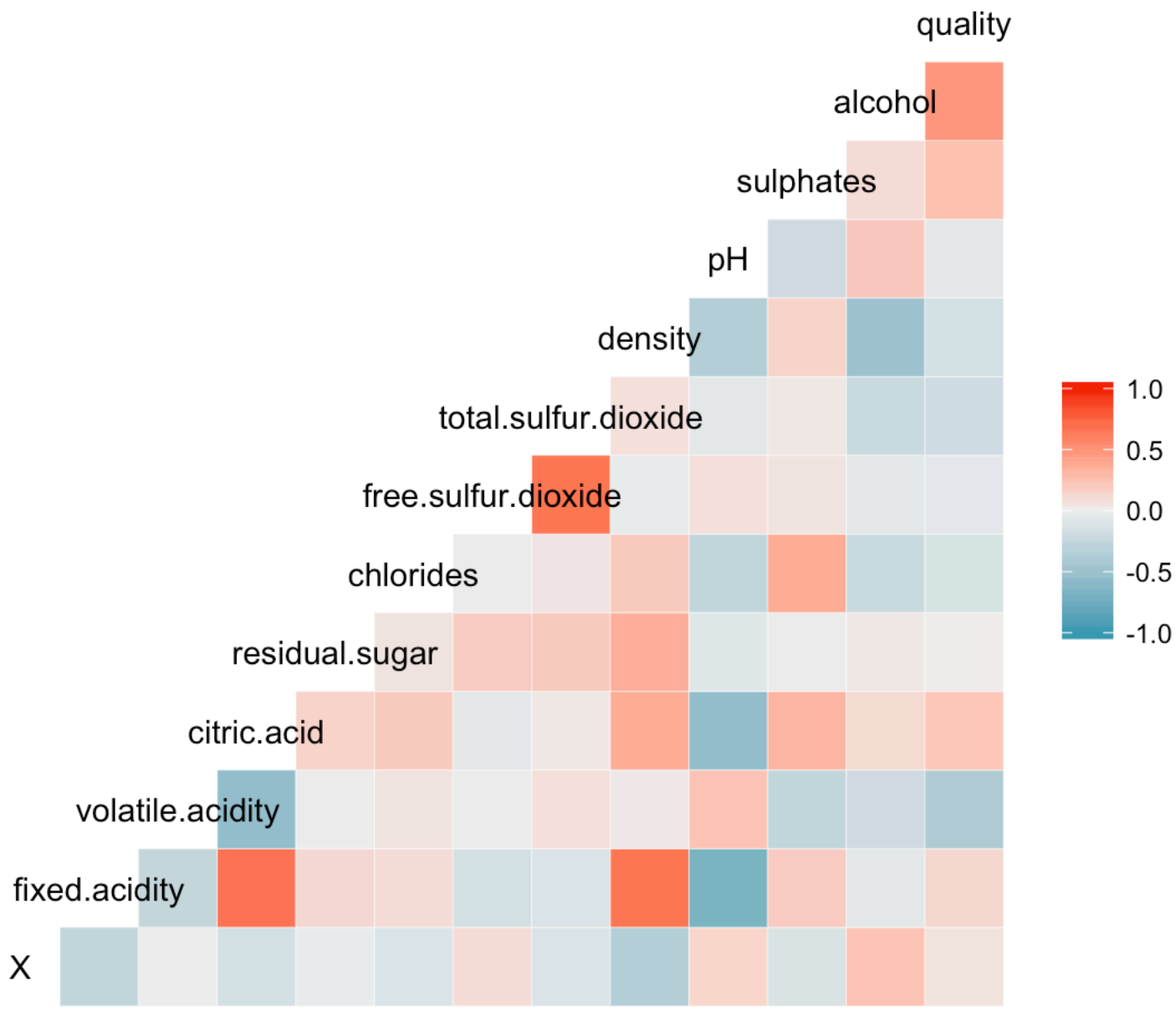
Did you perform any operations on the data to tidy, adjust, or change the form

of the data? If so, why did you do this?

CItric acid contains an outlier at 1 gm/liter. Sugar also has few outliers exceeding 12 g. I wonder if it is the same wine that has these outliers.

BIVARIATE PLOT SECTION

Preparing plot between all the variables to see all correlated variable and decide which individual plots to dig into deeper

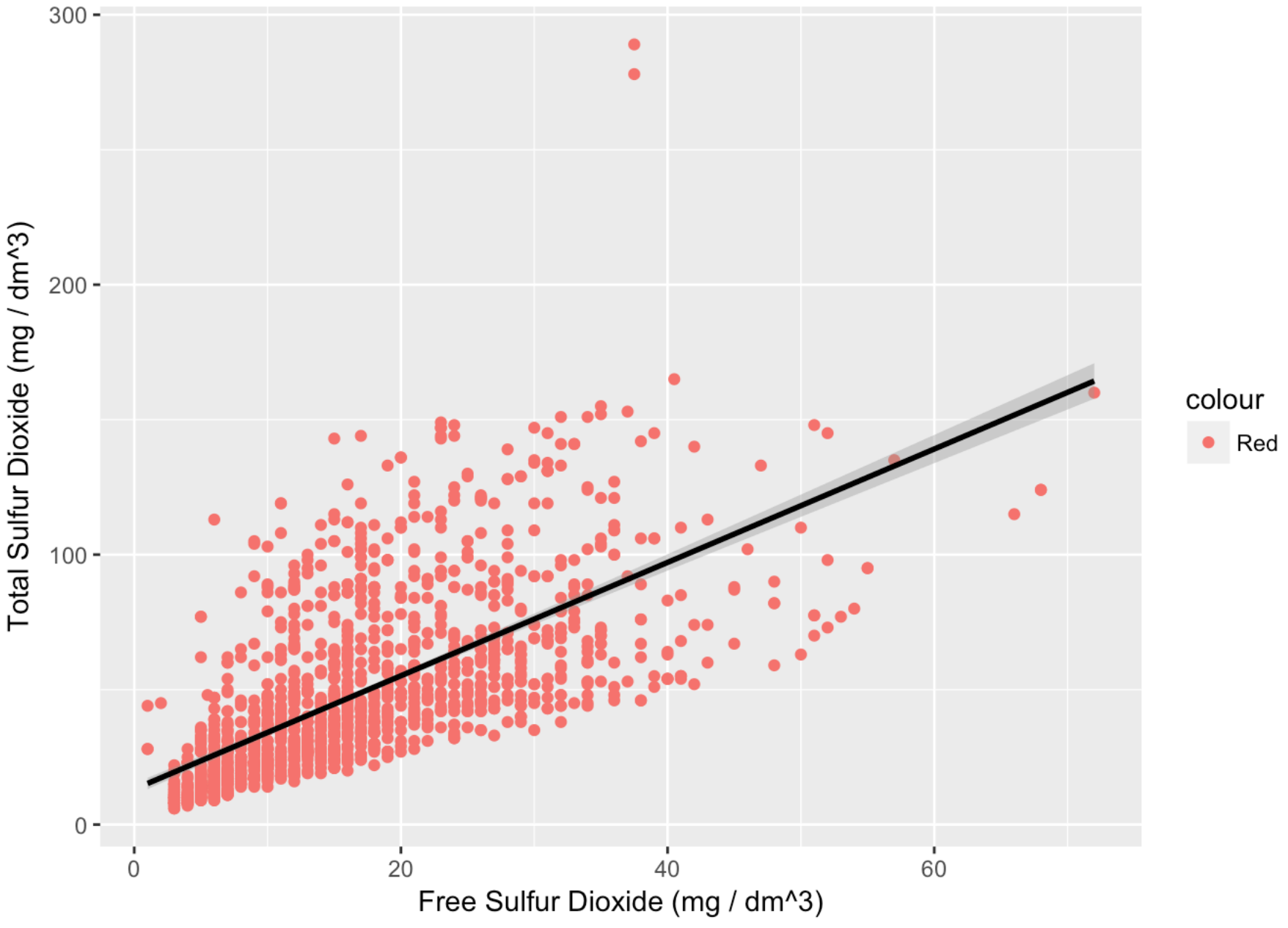


1. The strongest correlation of quality is with alcohol and volatile acidity. It also does have some correlation with sulphates and citric acid.

2. Some of the strongest correlations are citric and fixed acidity citric and volatile acidity(negative) fixed acidity and density ph and fixed acidity (negative) alcohol and density (negative)
3. Since the strongest relation of quality is with alcohol and volatile acidity, and since volatile has relation with citric and alcohol has relation with density, we may be able to explain quality through some of these indirect variables like density and citric acid. Plotting these variables will help figure out what all quality is dependent upon
4. Free sulfur dioxide and total sulfur dioxide has strong correlation. That may be because total sulfurdioxide is a combination of free and bound form of sulfurdioxide
5. Sulphates can contribute to SO2 but it doesn't have any relation with free or total sulfur dioxide which is strange. May be its because we have very limited data
6. Residual sugar has extremely weak correlation with quality. It has some correlation with total dulfur dioxide and free sulfur dioxide and has the strongest correlation with density

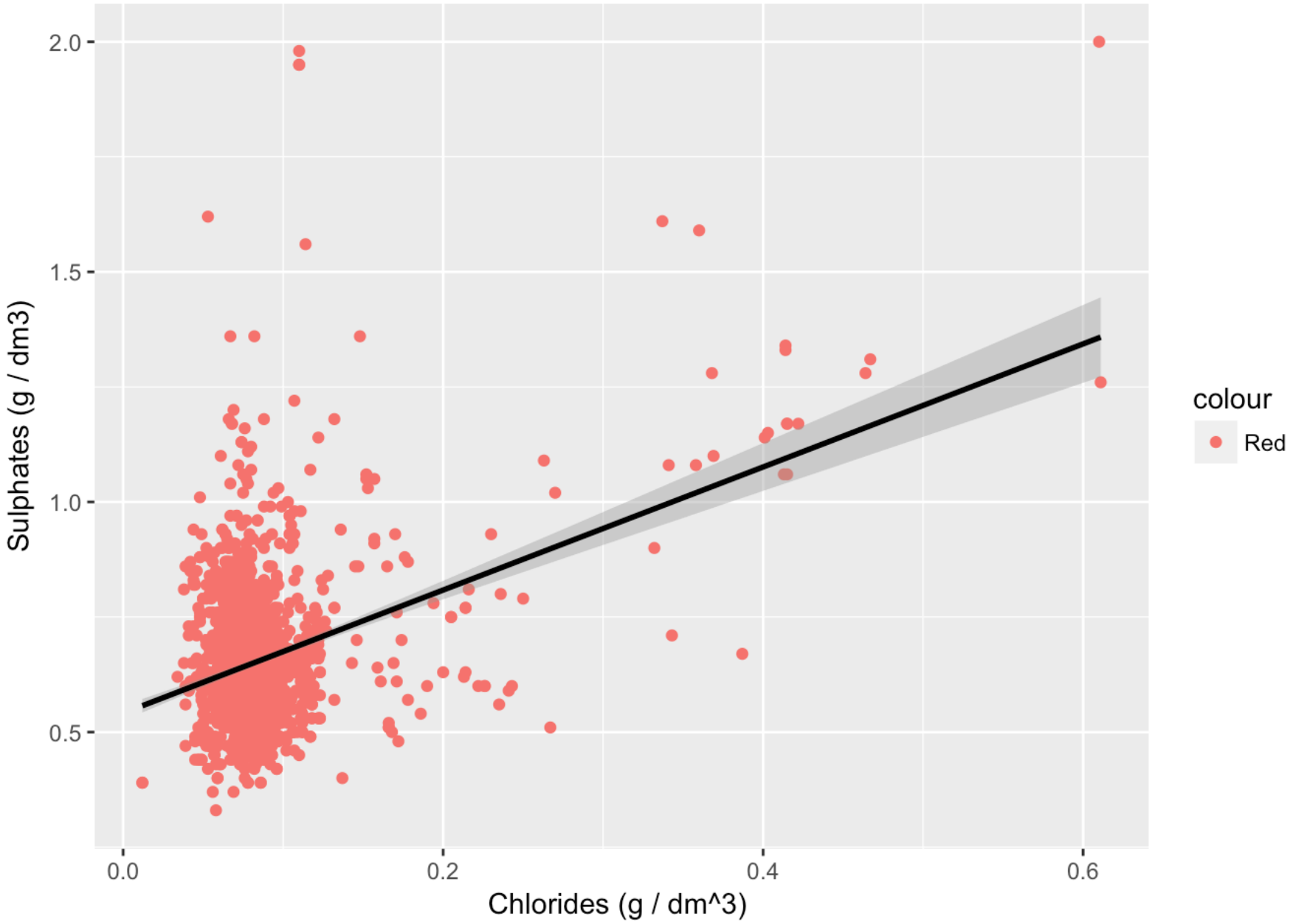
Lets create some box plots to figure out the relation between some of the above variables and how do they impact quality

Free Sulfur Dioxide and Total SUlfur Dioxide



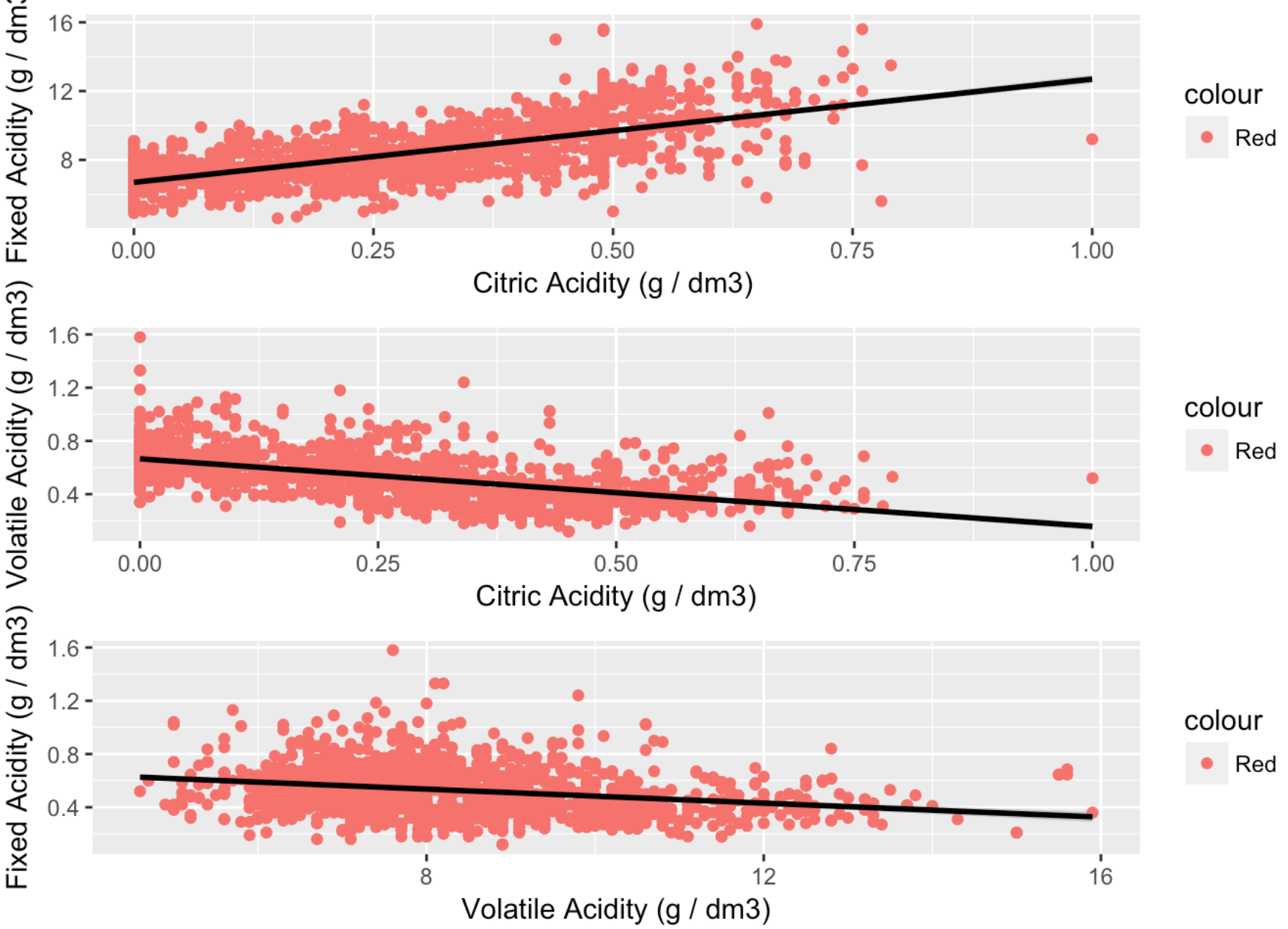
Above graph makes sense, since free sulfur dioxide is a part of total sulfur dioxide

Chlorides and Sulphates



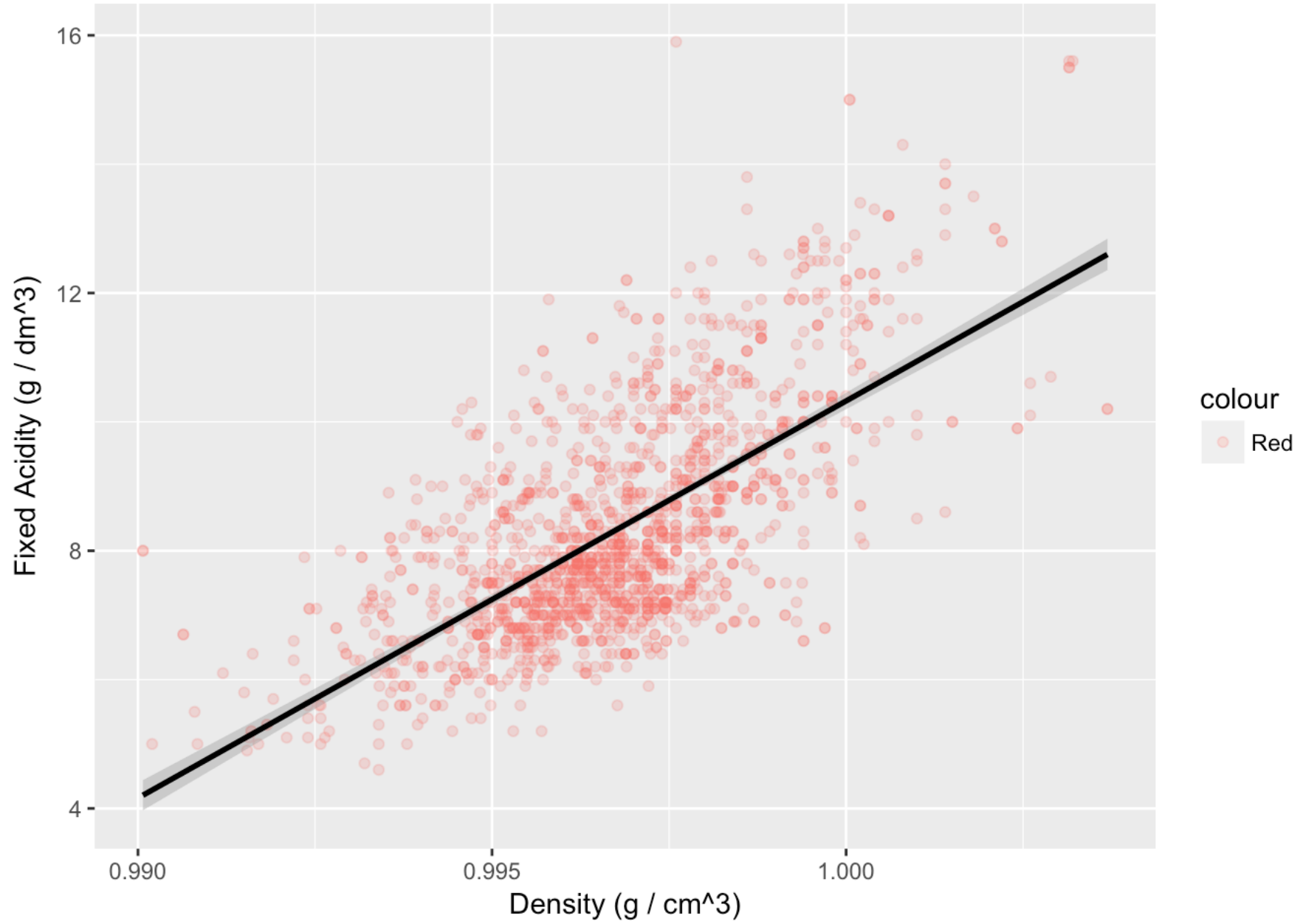
It appears that as chlorides concentration increaases , sulphate concentration also increases

Scatter plot between fixed acidity ,citric acidity and volatile acidity



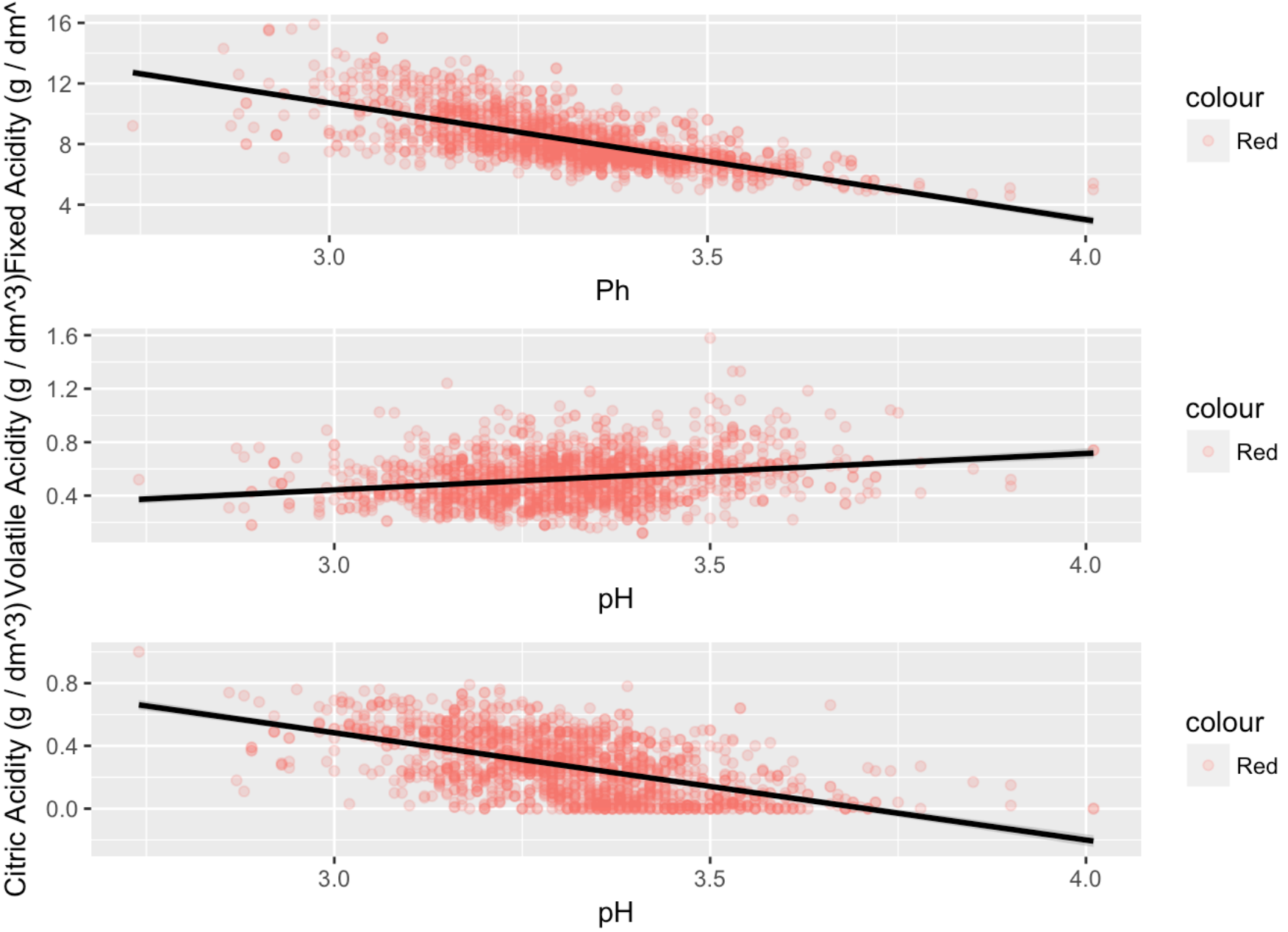
As we can see fixed acidity increase also increases citric acidity but an increase in citric acidity decreases volatile acidity. Thus fixed acidity and citric acidity has a positive relation with each other and a negative correlation with volatile acidity

Relation between density and fixed acidity



As fixed acidity increases, density also increases. They have a strong correlation

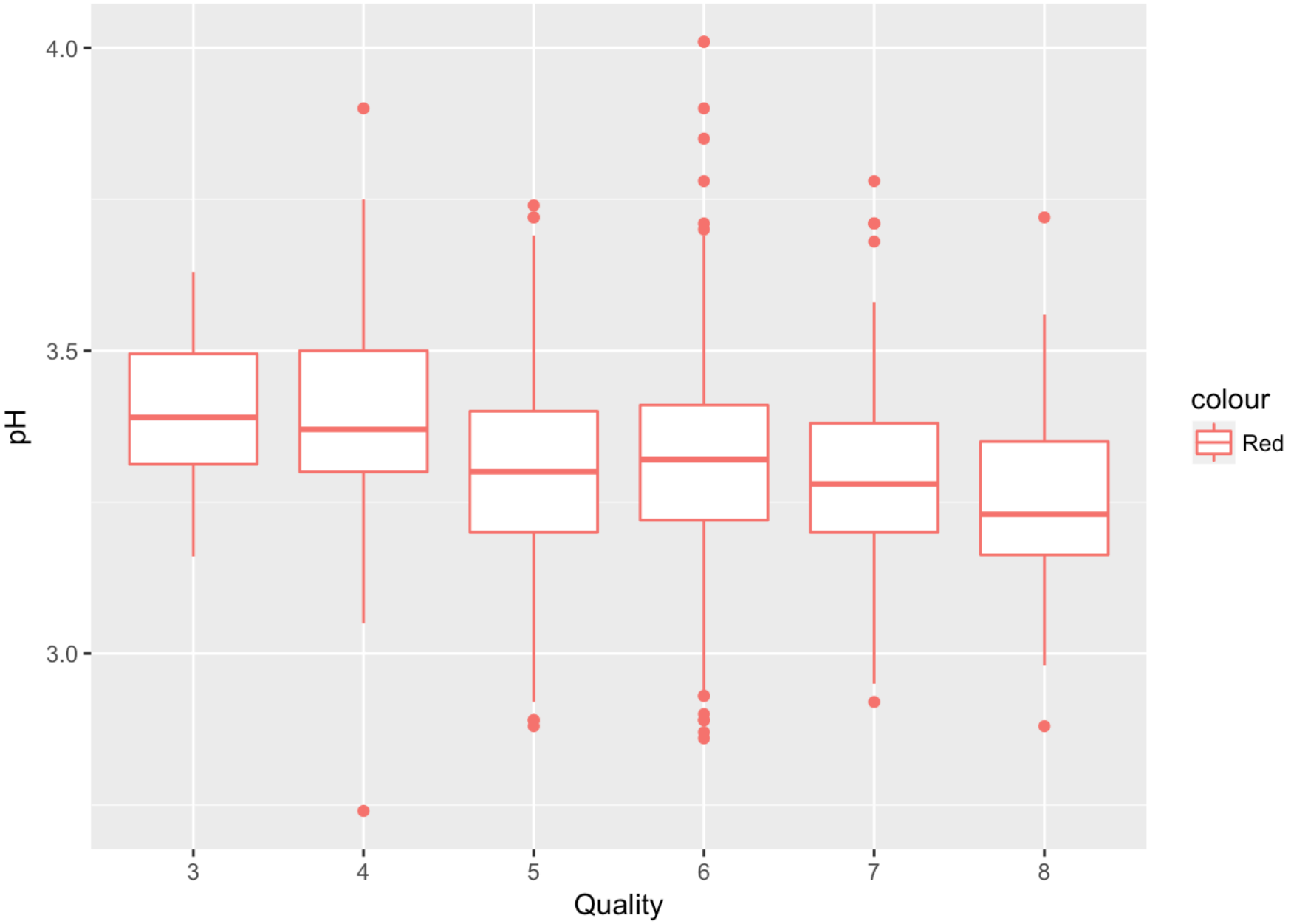
Relation between pH value and different kinds of acids in wine



As can be seen , increase in pH decreases the fixed acidity since higher pH has low acidity.Well, its interesting to see how increase in pH is increasing the volatile acidity but decreasing the fixed acidity. Higher pH should have low acidity but that true for relation between pH and fixed acidity and not between pH and volatile acidity. I wonder why.

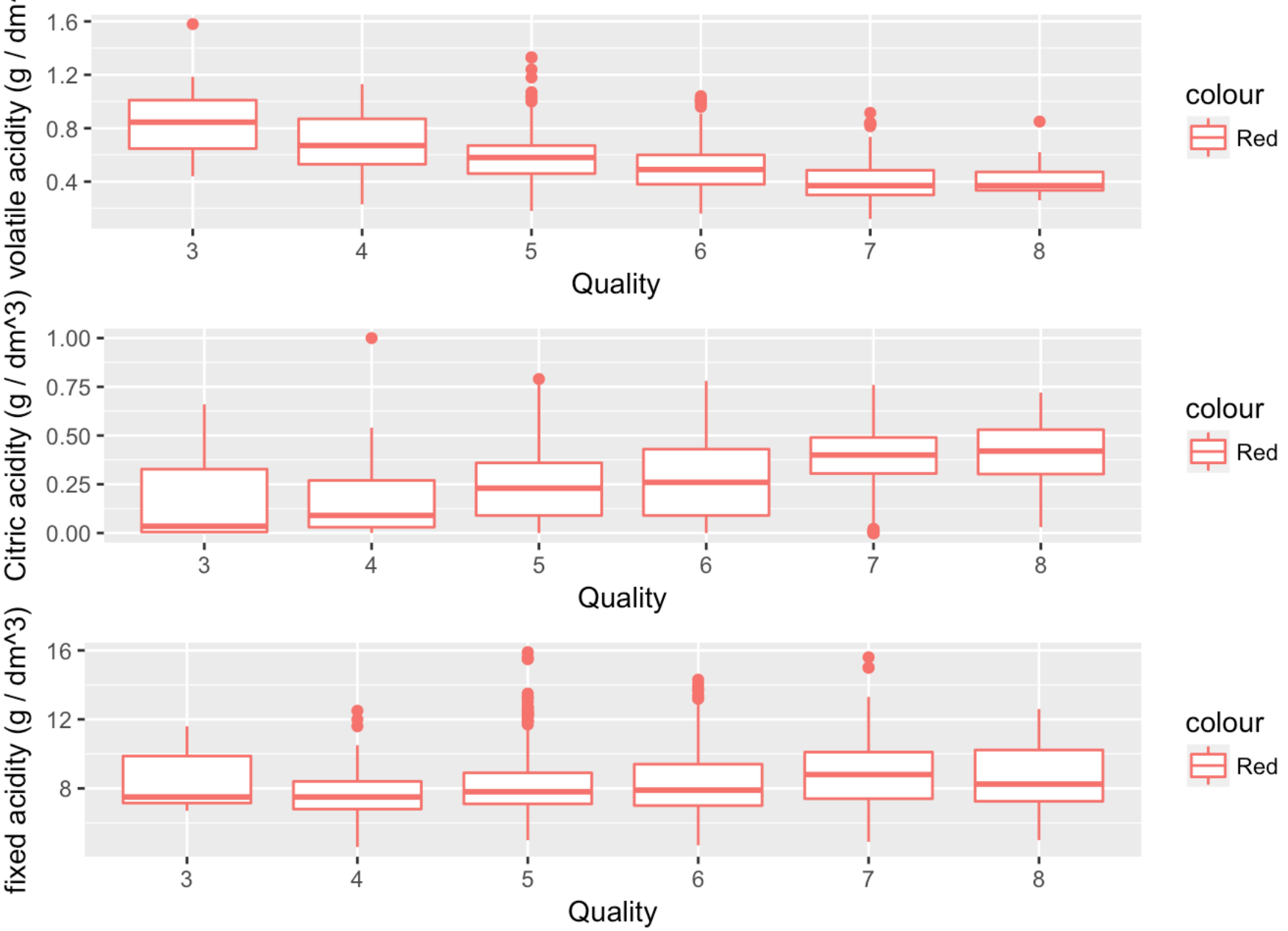
Lets see how quality is related to various acids in the wine

Relation between pH and quality of wines



Lower quality wines tend to have higher pH levels which means lower quality wines should have less acidity

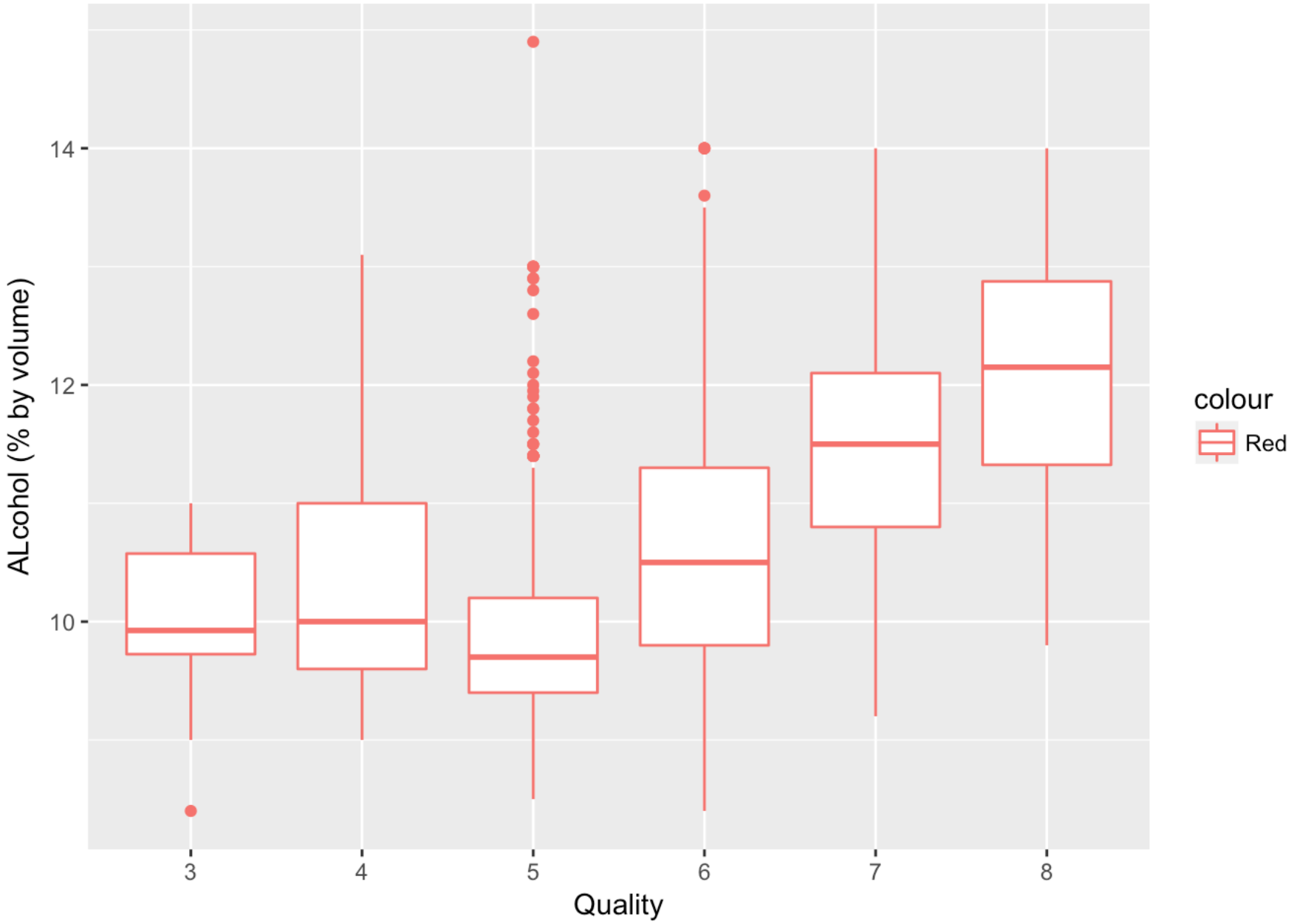
Relation between various acidities and quality of a wine



As expected in this dataset, Higher quality wines have lower ph and lower volatile acidity (since ph and volatile acidity are positively related suprisingly) and higher quality wines have higher median value of citric acid and fixed acidity.

Lets see relation between quality and alcohol

Relation between quality and alcohol

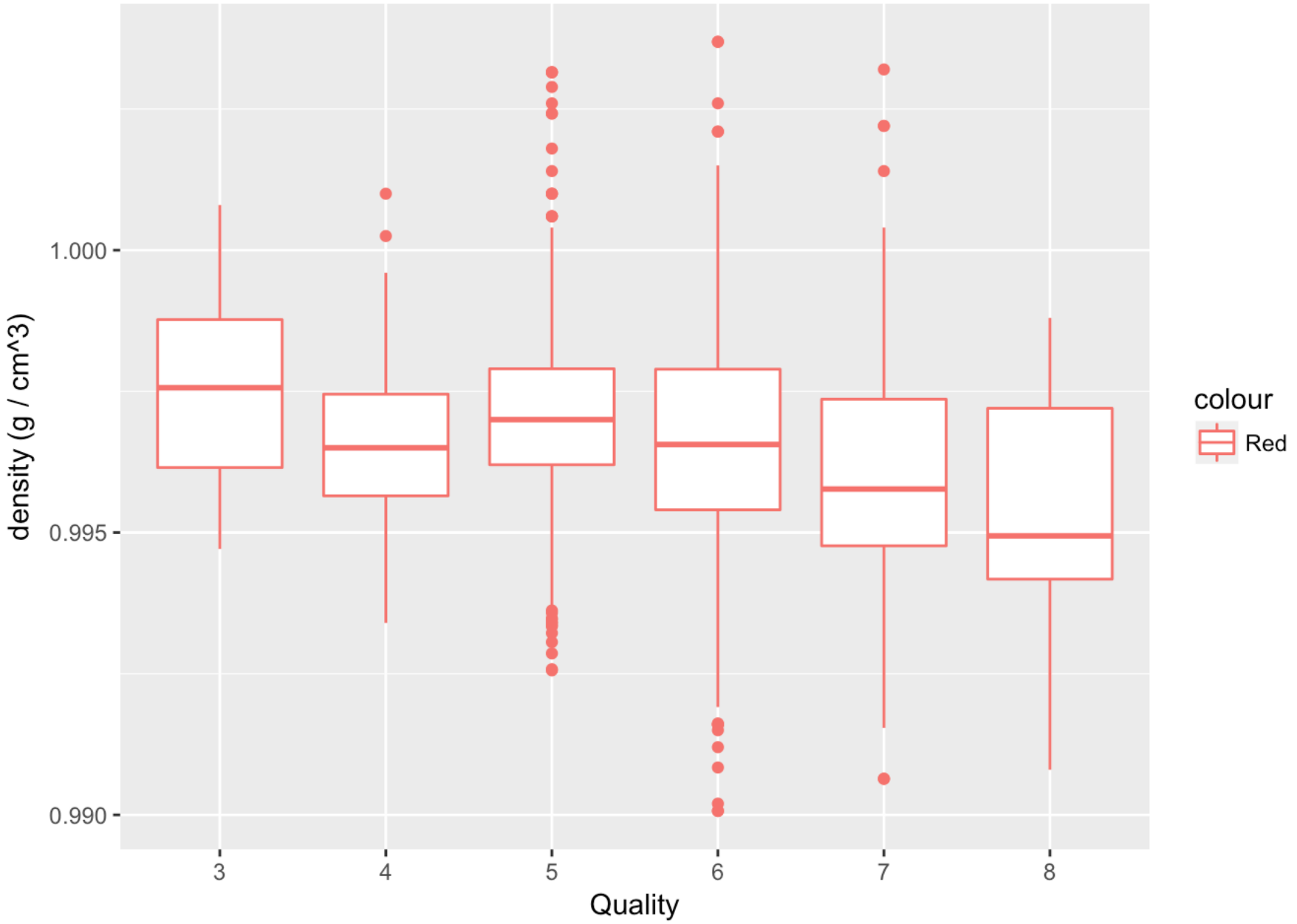


With the exception of qines with quality 5, alcohol content appears to be increasing in higher quality wines. The middle quality also have a lot of outliers that may be the reason of a low median in that quality.

Relation between alcohol and density

As the alcohol content increases , density decreases. This makes sense since the density of water is higher than alochol. Since higher quality wines have high alcohol content, the density should reduce as the quality increases. Let plot density and quality to confirm the relationship

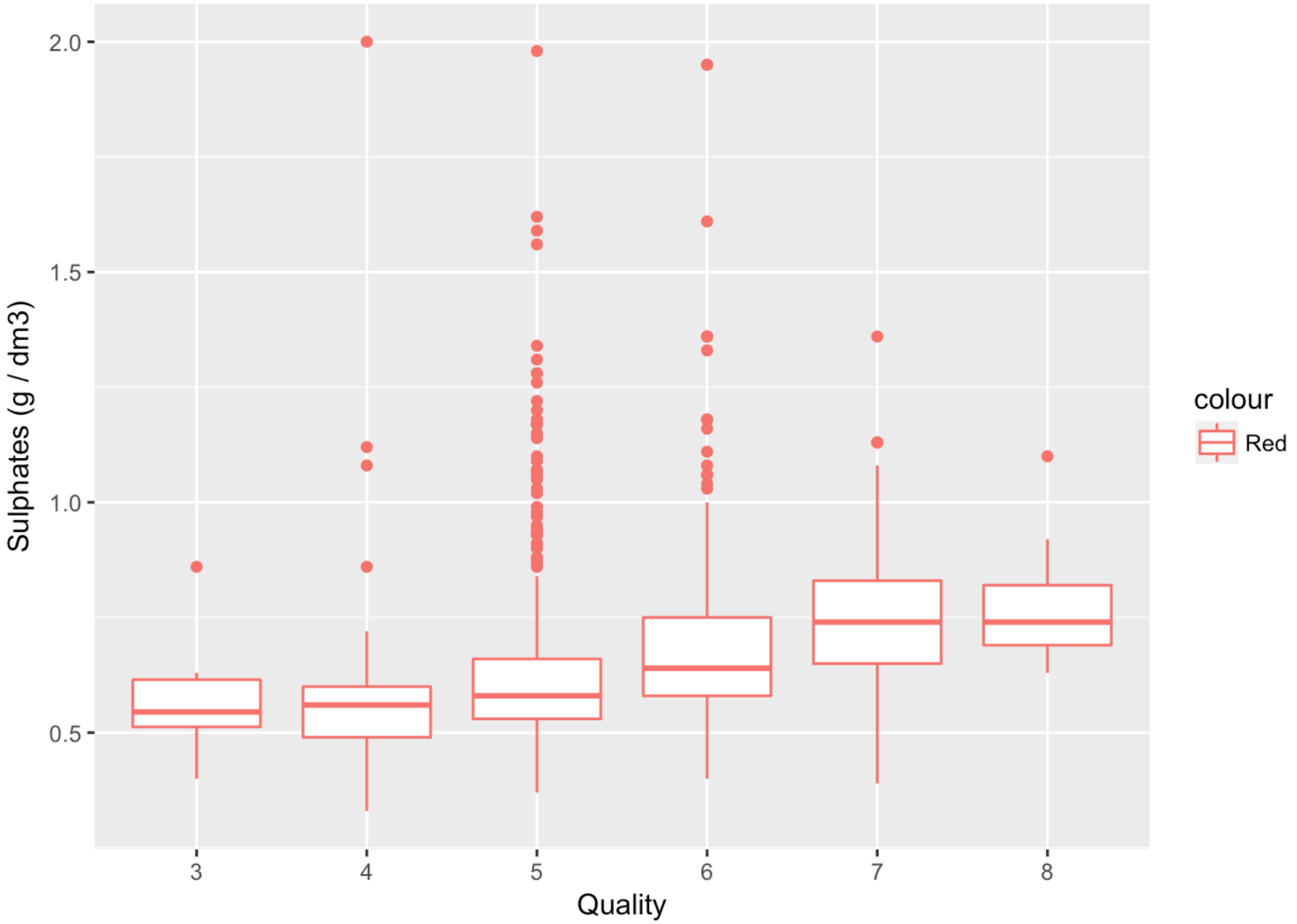
Relation between quality and density



With the exception of quality 5 wines, the density appears to be reducing as the quality increases. The quality 5 may be behaving this way because of many outliers as seen above. But in general , the density reduces as the quality increases

Lets figure out the relation between sulphates and quality. As per the correlation matrix, sulphates do have some correlation with quality

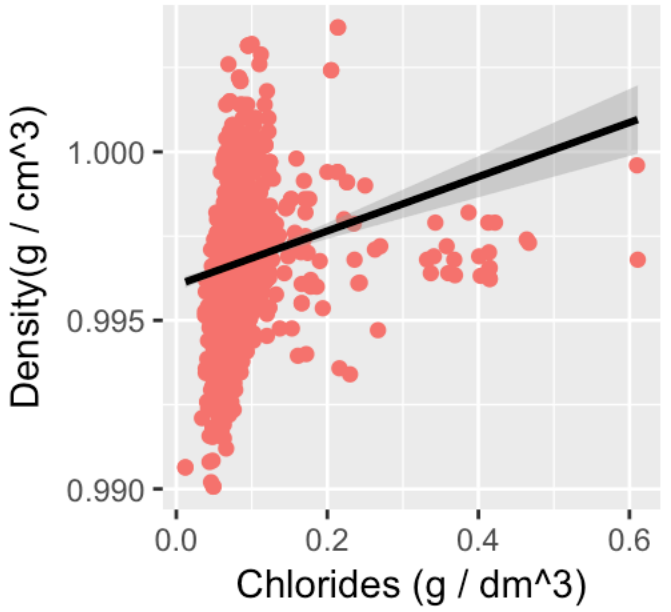
Relation between quality and sulphates

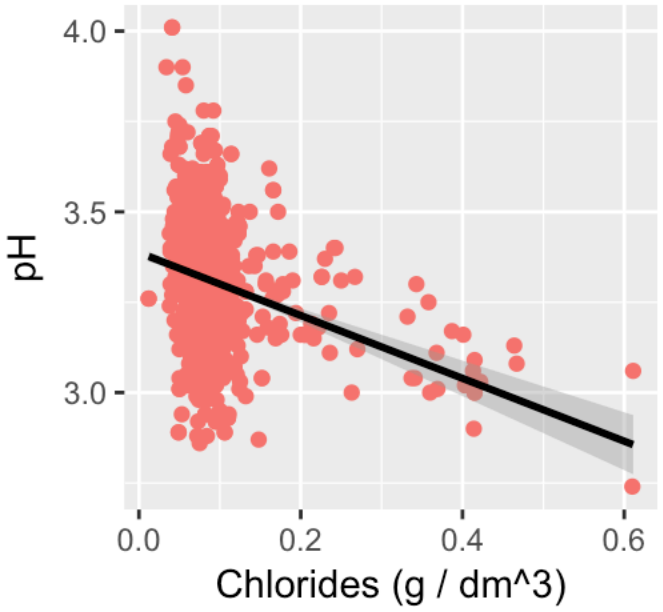


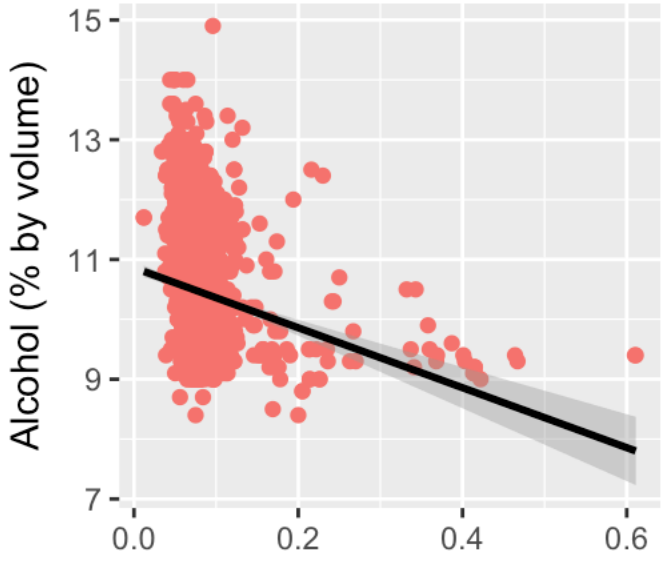
There are quiet a few outliers here so we may need to remove outliers to figure out the exact relationship. Looking at the above plot, it appears that higher quality wines have high sulphate content.

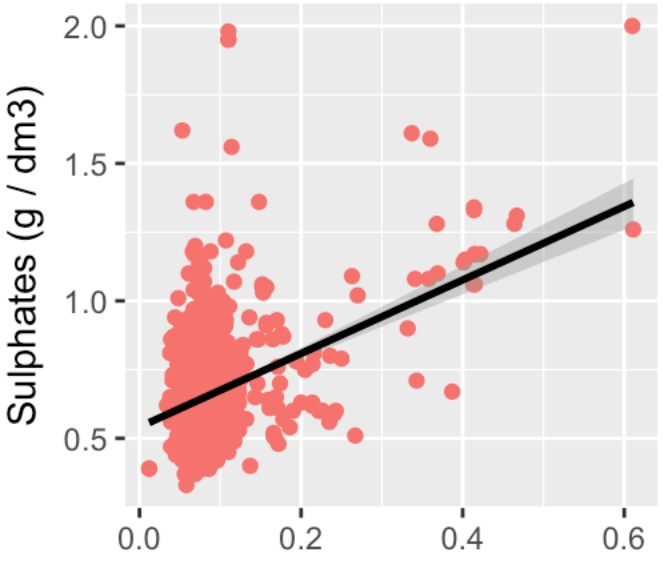
Chlorides appears to be having some correlation with density, pH, sulphates and alcohol. Since some of these attributes have relation with quality, I would like to see what relation chloride have with these variables and if chloride is anywhere determining quality. Lets plot chloride with these attributes

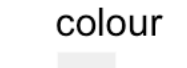
Relation of Chlorides with Density, pH, Alcohol, Sulphates



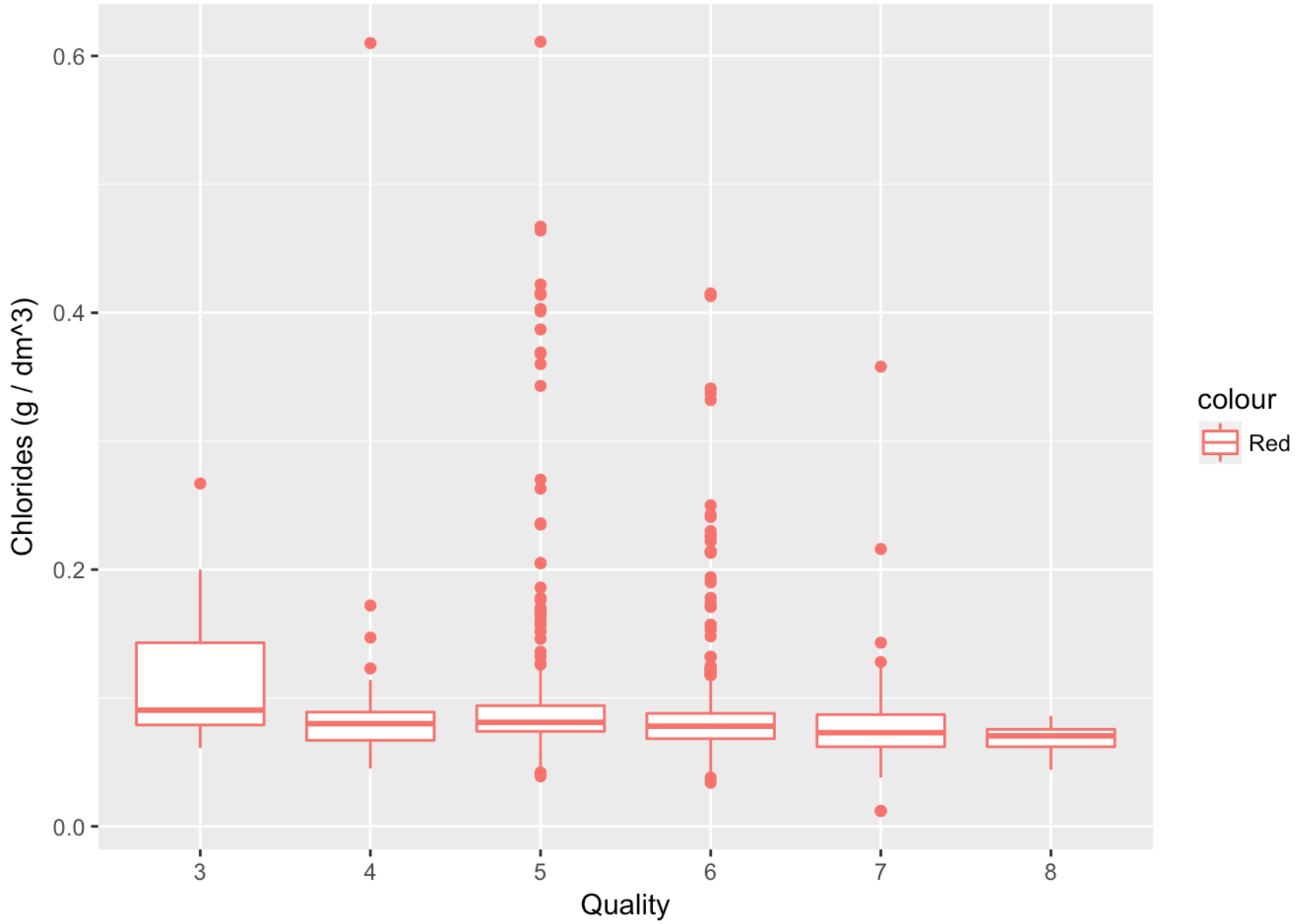






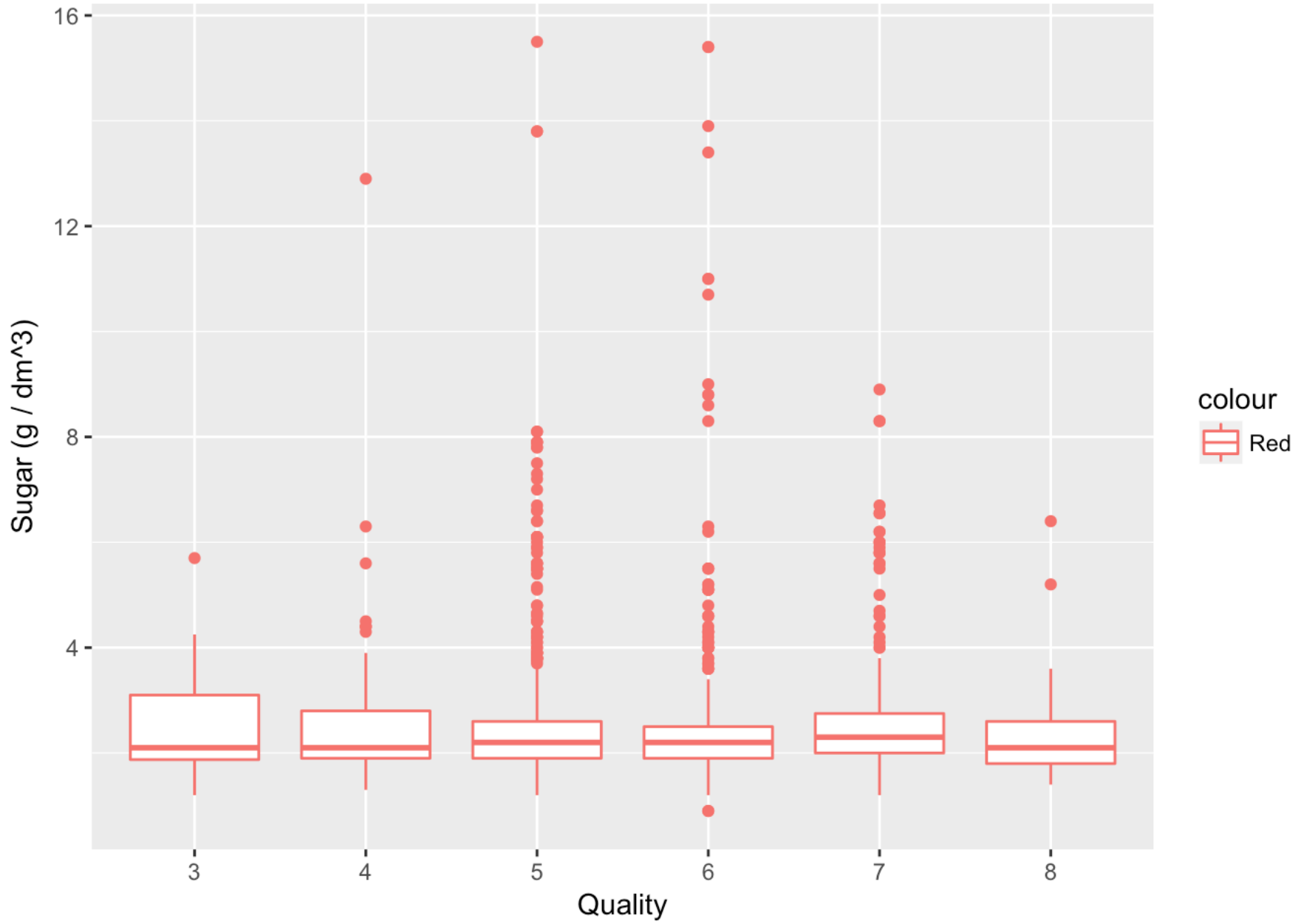


Relation between chlorides and quality



As per the above graph, these doesnt appear to be any significant relation between quality and chlorides. Chlorides have a positive relation with sulphates and sulphates have a positive relation with quality but chlorides doesnt appear to be impacting quality

Relation between residual Sugar and quality



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?

1. As quality increases, alcohol content increases
2. As quality increases , citric acid increases, volatile acidity decreases and fixed acidity increases
3. As quality increases, density decreases
4. As quality increases, ph value decreases
5. As quality increases, sulphates content increases
6. Chlorides doesnt have an impact on quality
7. Residual Sugar doesnt have an impact on quality

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

1. Higher alcohol content has lower density

2. Fixed acidity increases as density increases
3. Fixed acidity decreases as ph Increases
4. Citric acid decreases as pH increases
5. Volatile acidity increases as pH increases
6. Free sulfur dioxide is positively related to total sulfur dioxide
7. Chlorides have a slight positive relation with sulphates

What was the strongest relationship you found?

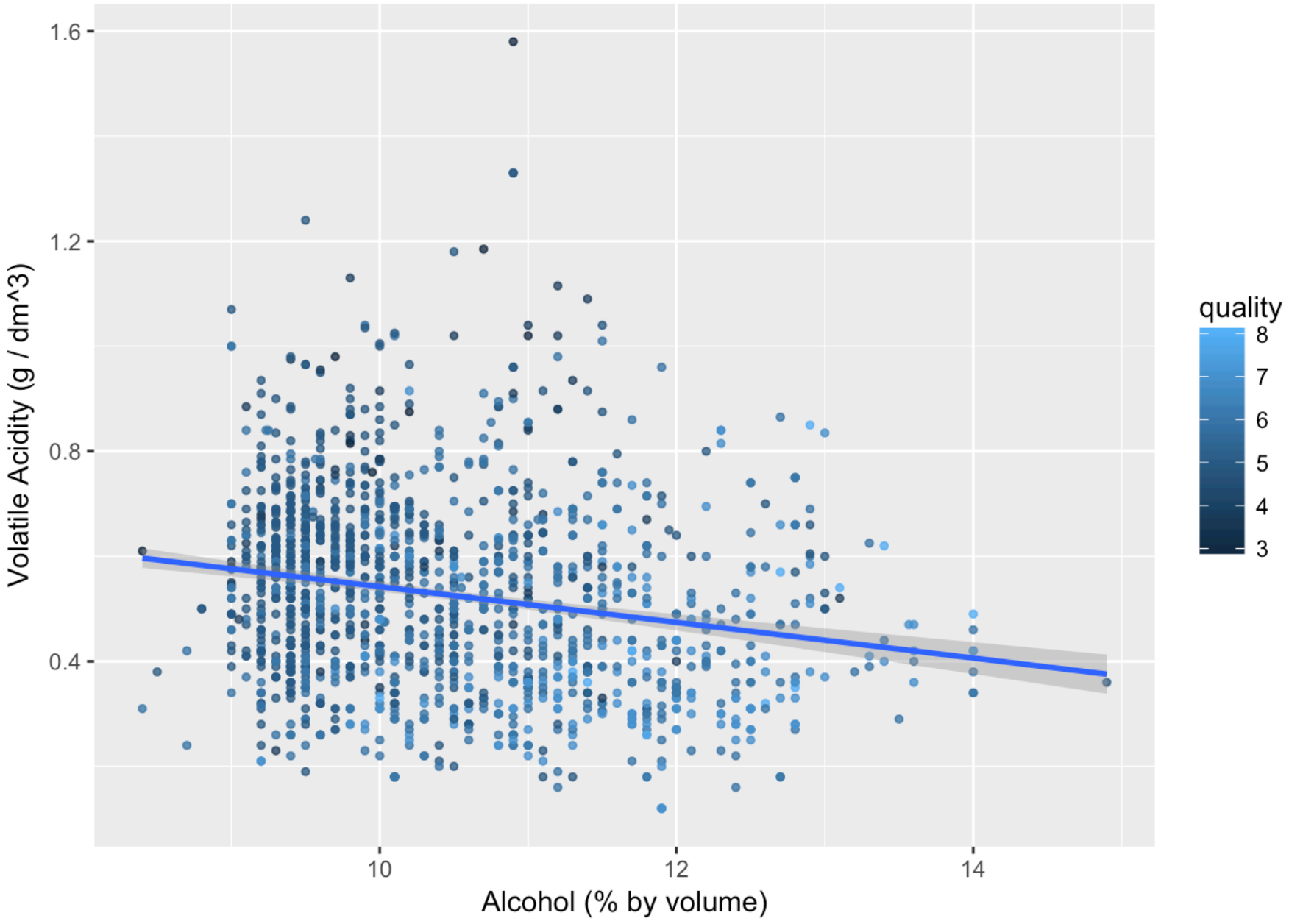
1. Quality is positively related to alcohol. Correlation .476
2. Free Sulfur Dioxide is positively related to total sulfur dioxide .668

MULTIVARIATE PLOT

Relationship between citric acid, alcohol and quality of wine

Looking at the above graph, it can be said that wines with higher alcohol content and higher citric acid content seem to be of higher quality

Relationship between volatile acidity, alcohol and quality of wine

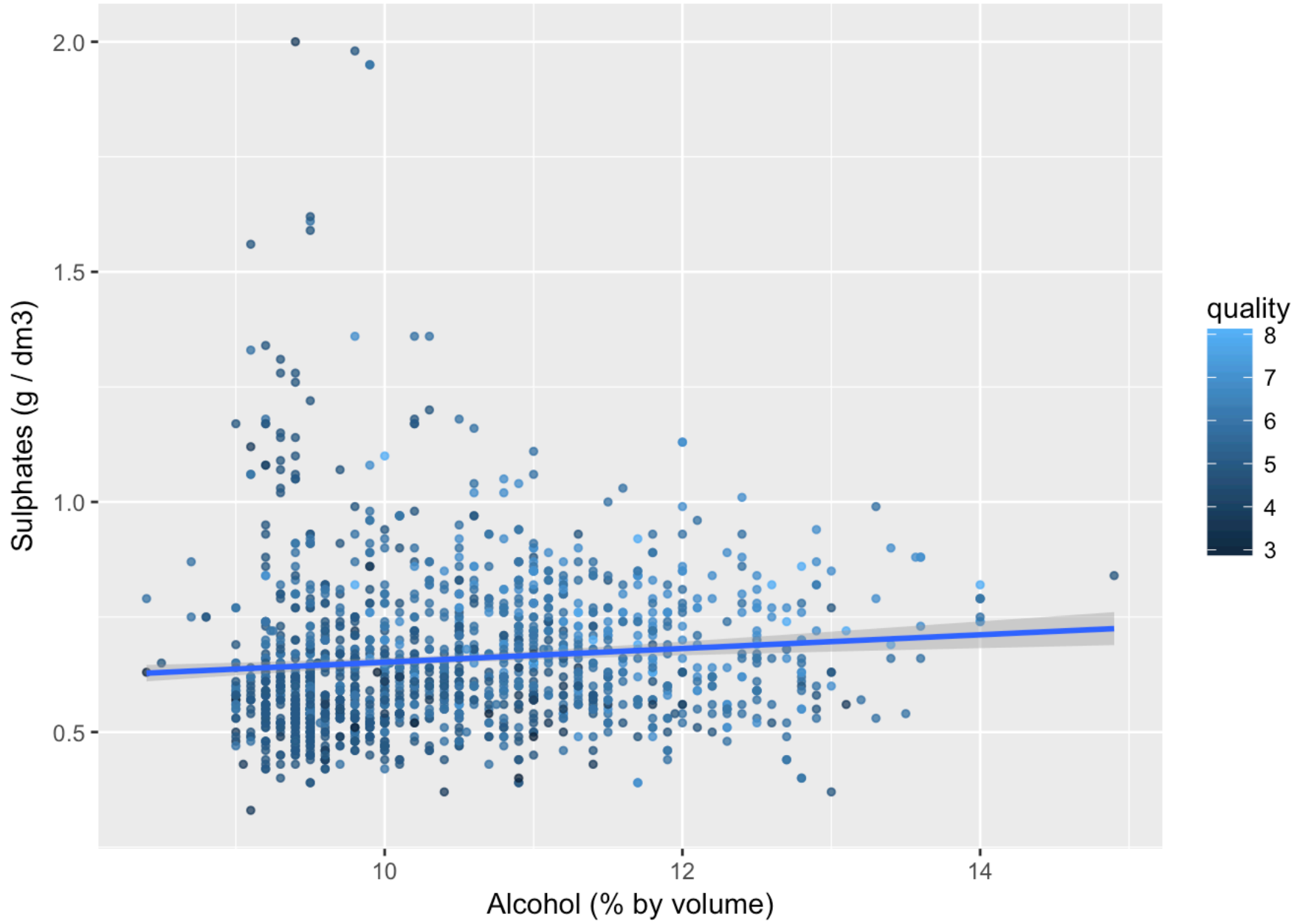


It appears that higher alcohol and lower volatile acidity tends to produce better quality wines

Relationship between density, alcohol and quality of wine

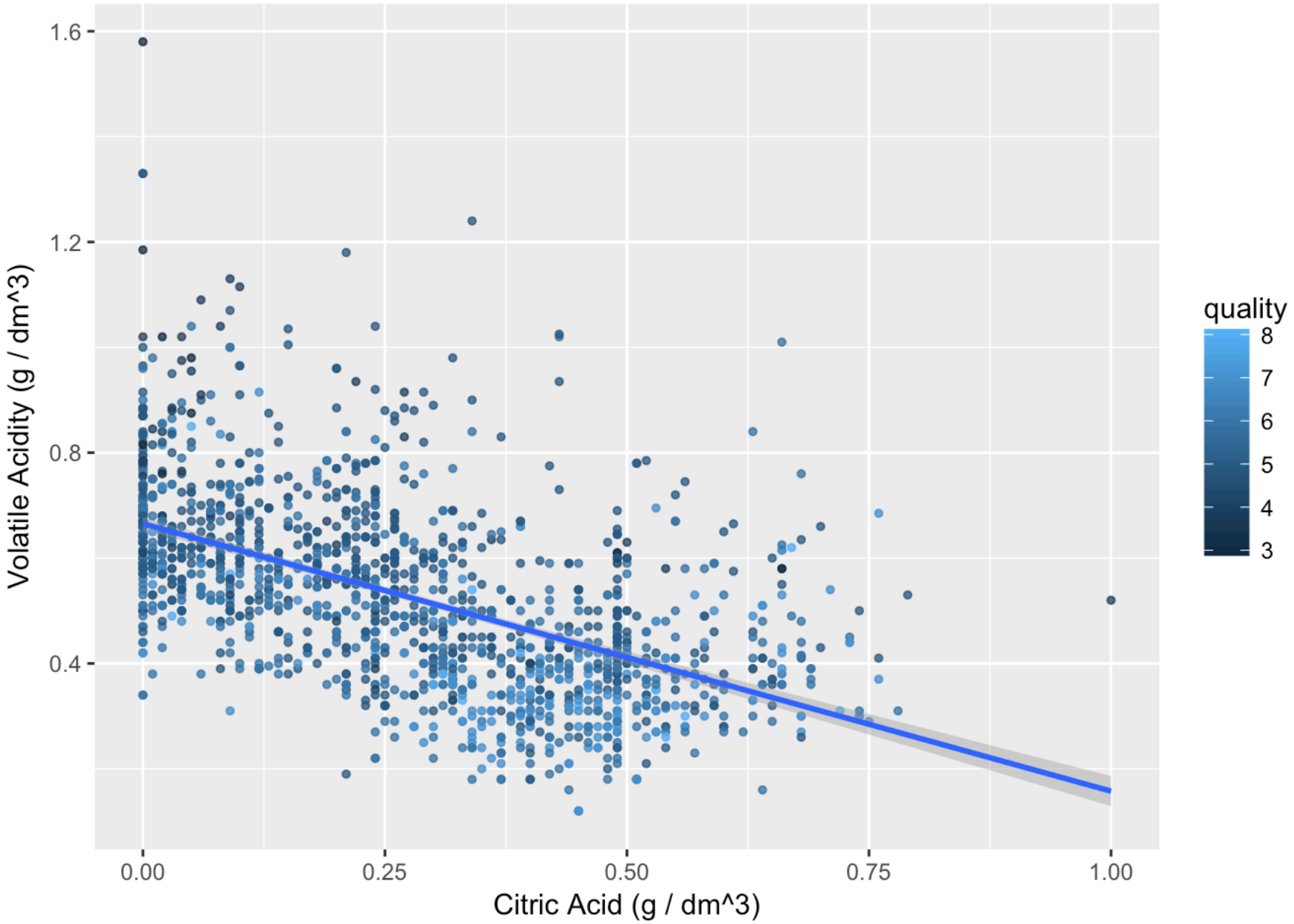
Lower density and higher alcohol content seem to be producing better quality wines

Relationship between sulphates, alcohol and quality of wine



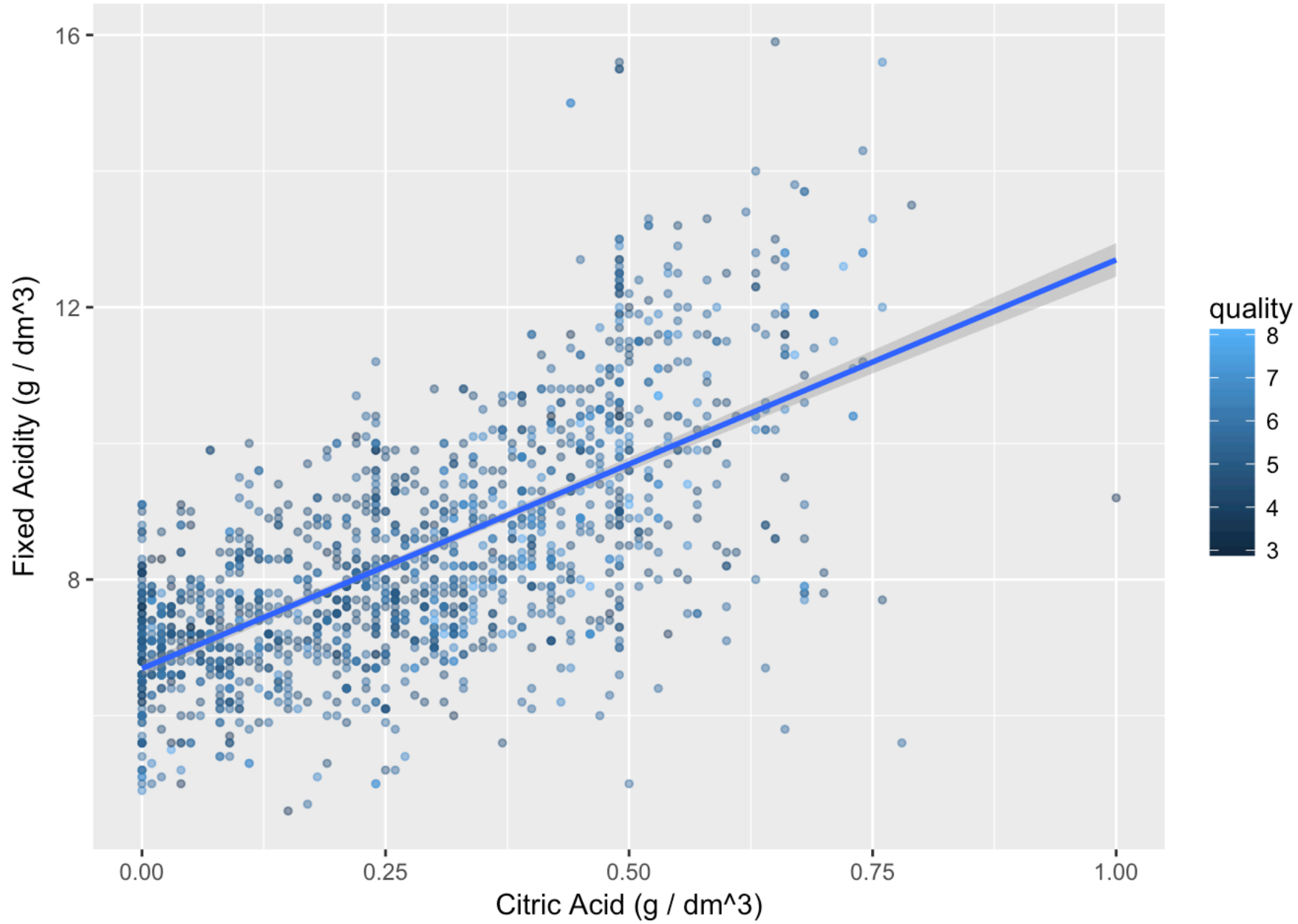
Higher sulphate and higher alcohol content produces better quality wines Try to figure out impact of different kinds of acids on quality

Relationship between volatile acidity, citric.acid and quality of wine



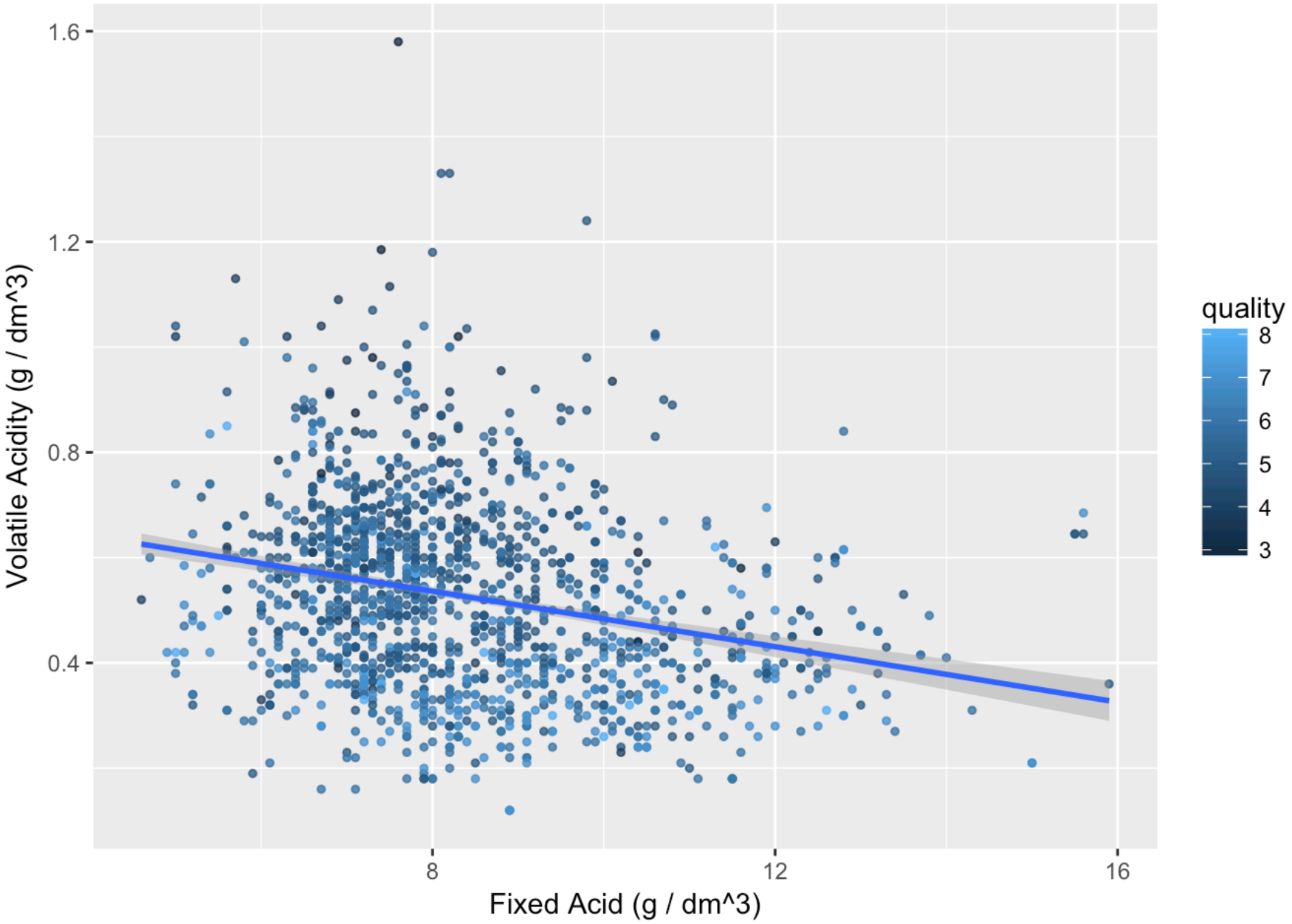
Lower volatile acidity and higher citric acid produces better wine

Relationship between volatile acidity, alcohol and quality of wine



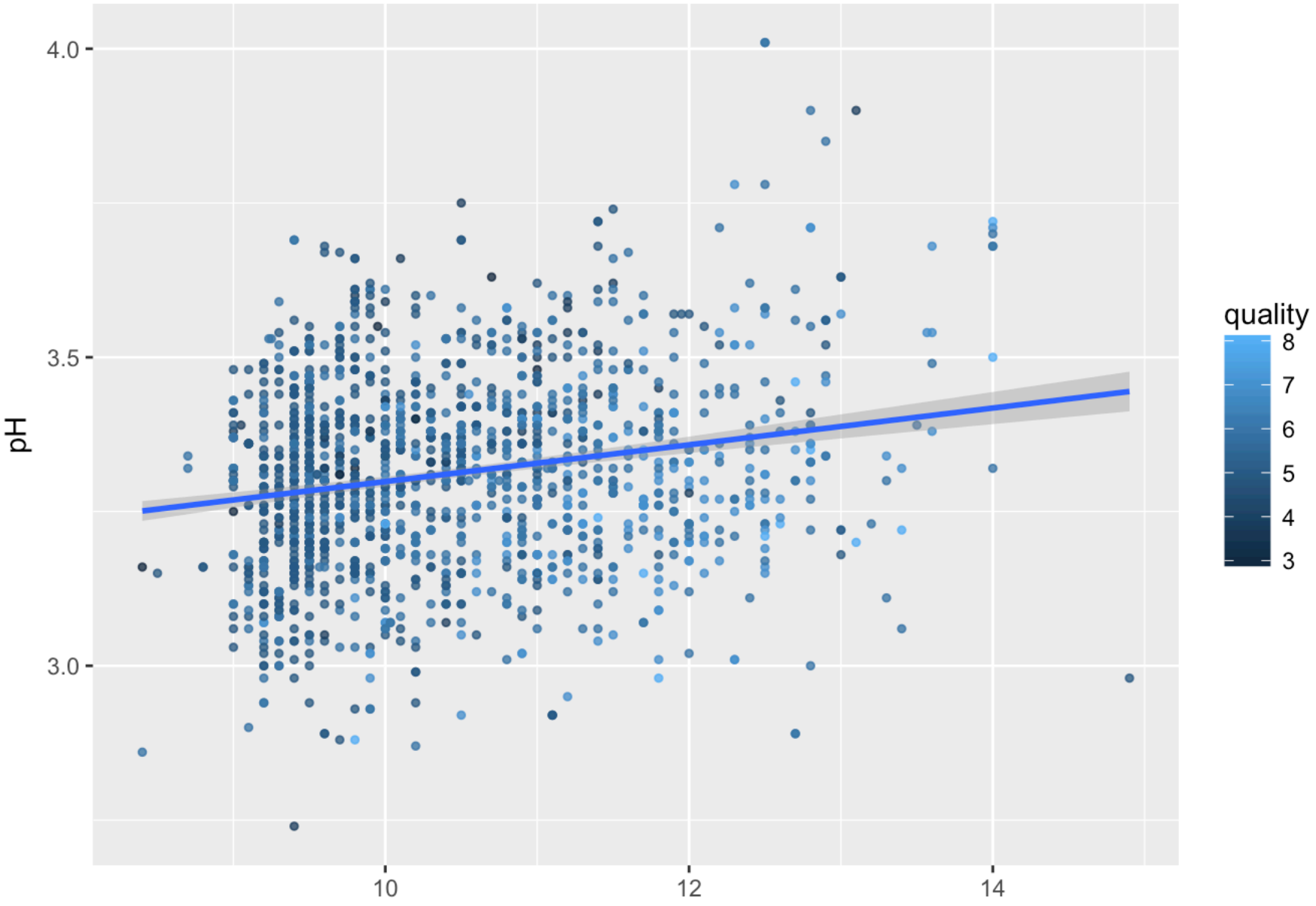
Not much correlation here

Relationship between volatile acidity, fixed acitidy and quality of wine



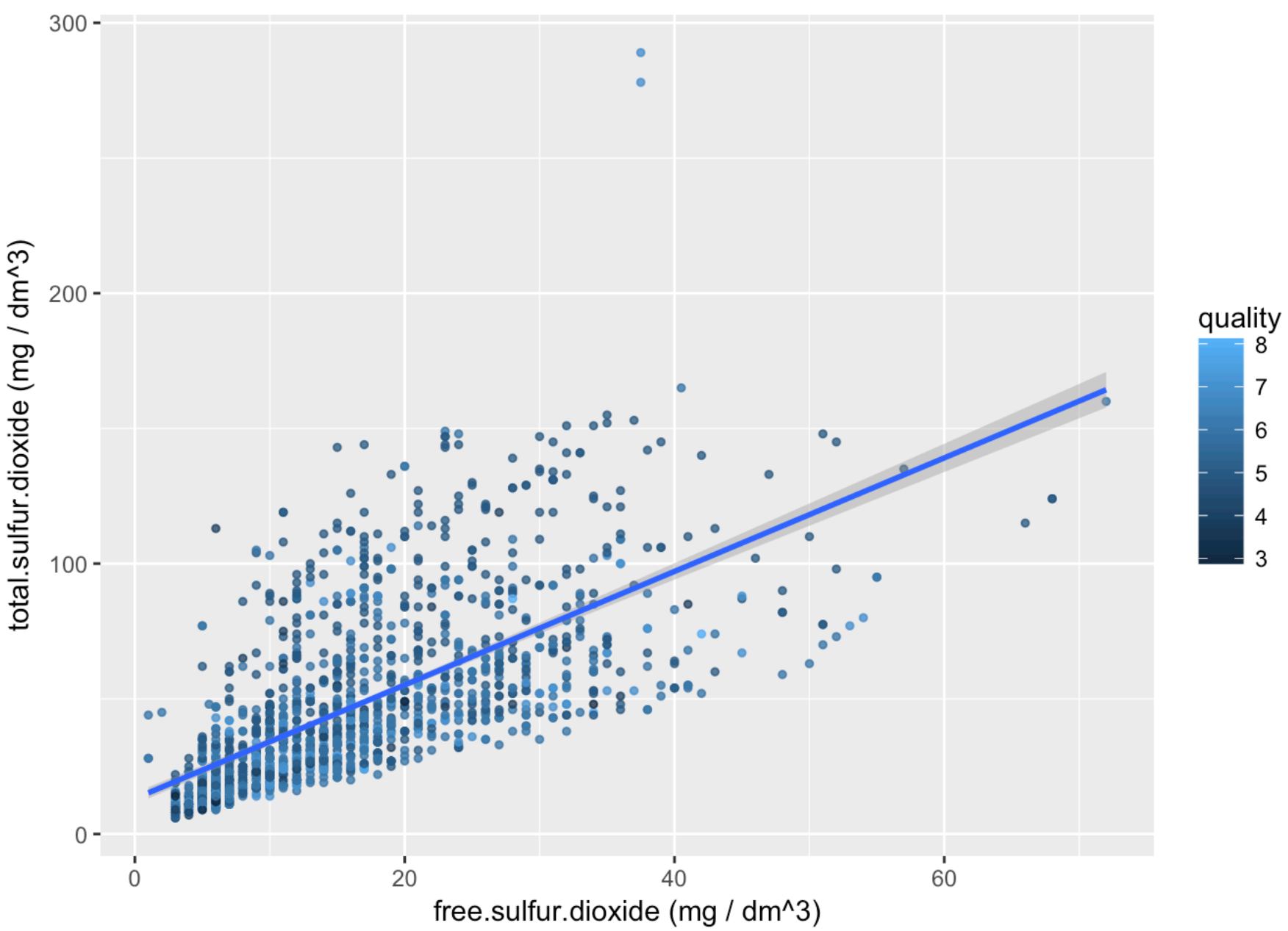
Not Much correlation here

Relationship between wine quality, alcohol and pH



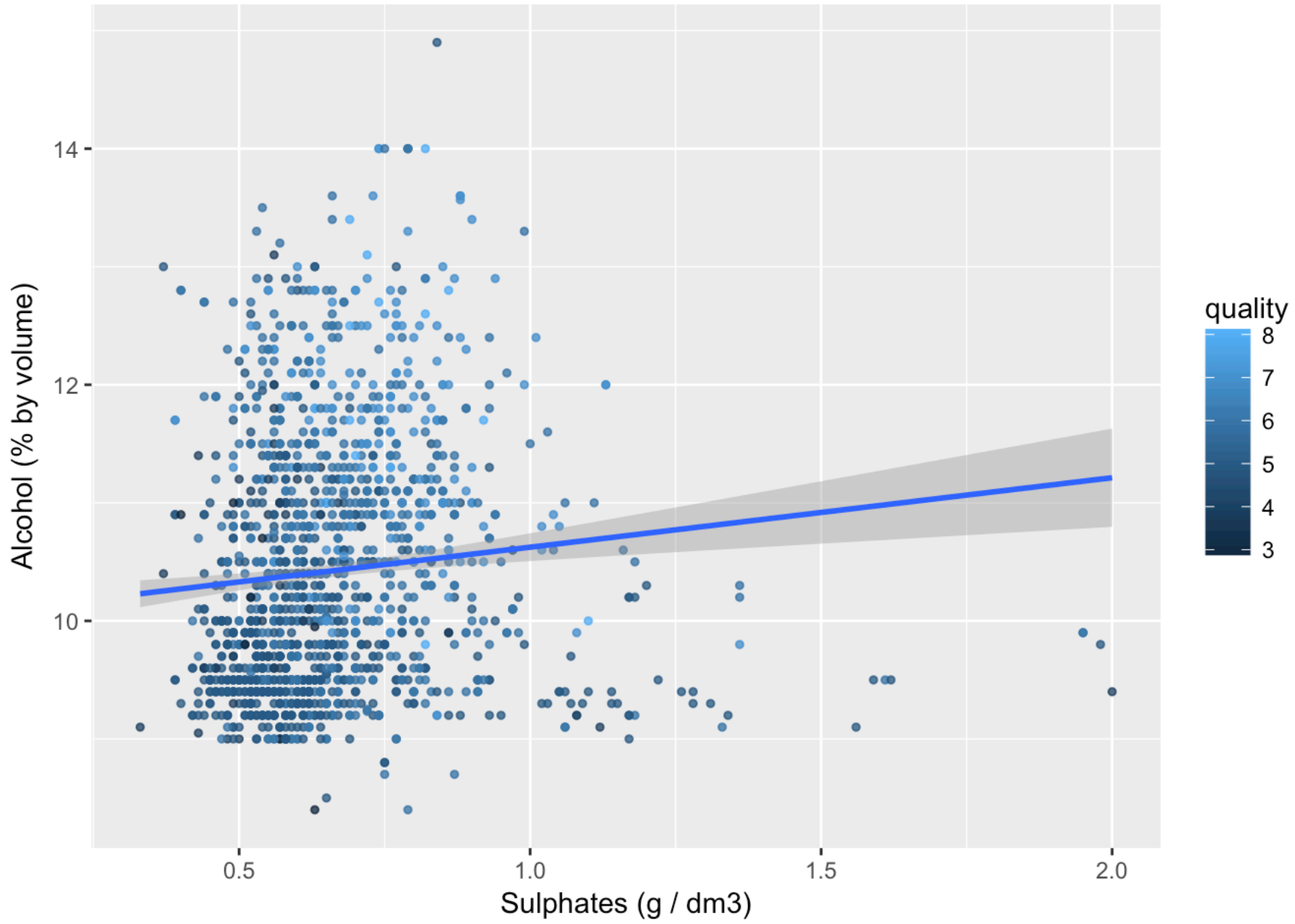
low pH and high alcohol content tend to produce better wines

Relationship between wine total.sulfur.dioxide, alcohol and free.sulfur.dioxide



Not much correlation here

Relationship between wine quality, alcohol and sulphates



It looks like wines with higher quality content are better if they have high sulphates in them.

Creating a model

```
##
## Calls:
## m1: lm(formula = as.numeric(quality) ~ alcohol, data = winequality)
## m2: lm(formula = as.numeric(quality) ~ alcohol + citric.acid, data = winequality)
## m3: lm(formula = as.numeric(quality) ~ alcohol + citric.acid + sulphates,
##     data = winequality)
## m4: lm(formula = as.numeric(quality) ~ alcohol + citric.acid + sulphates +
##     density, data = winequality)
## m5: lm(formula = as.numeric(quality) ~ alcohol + citric.acid + sulphates +
##     density + pH, data = winequality)
##
## ================================================================================
======
##                          m1             m2             m3             m4
m5
## --------------------------------------------------------------------------------
------
##   (Intercept)         1.875***       1.830***       1.434***      19.459
20.50
0
```

```
##                    (0.175)       (0.171)       (0.176)      (12.063)      (12.04
0)
##   alcohol           0.361***      0.346***      0.338***      0.321***      0.33
7***
##                    (0.017)       (0.016)       (0.016)       (0.020)       (0.02
1)
##   citric.acid                     0.730***      0.513***      0.583***      0.40
5***
##                                  (0.090)       (0.093)       (0.104)       (0.12
1)
##   sulphates                                     0.814***      0.829***      0.81
1***
##                                                (0.107)       (0.107)       (0.10
7)
##   density                                                    -17.931       -17.74
5
##                                                             (11.998)      (11.97
1)
##   pH                                                                        -0.40
2**
##                                                                           (0.13
9)
## ------------------------------------------------------------------------------
------
##   R-squared         0.227         0.257         0.284         0.285         0.28
8
##   adj. R-squared    0.226         0.256         0.282         0.283         0.28
6
##   sigma             0.710         0.696         0.684         0.684         0.68
2
##   F               468.267       276.595       210.501       158.556       129.11
0
##   p                 0.000         0.000         0.000         0.000         0.00
0
##   Log-likelihood  -1721.057     -1688.711     -1659.955     -1658.835     -1654.63
6
##   Deviance          805.870       773.917       746.576       745.532       741.62
6
##   AIC              3448.114      3385.421      3329.910      3329.671      3323.27
2
##   BIC              3464.245      3406.930      3356.795      3361.934      3360.91
2
##   N                1599          1599          1599          1599          1599
## ==============================================================================
======
```

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?

1. Higher alcohol content along with lower density produces better wines
2. Higher alcohol content along with higher sulphates produces better wines
3. Higher citric acid and lower volatile acid produces better wines
4. Higher alcohol content and lower volatile acidity produces better wines
5. Higher alcohol content and higher citric acid produces better wines

Were there any interesting or surprising interactions between features?

Relationship between volatile acidity, alcohol and quality of wine was surprising as they were not much related

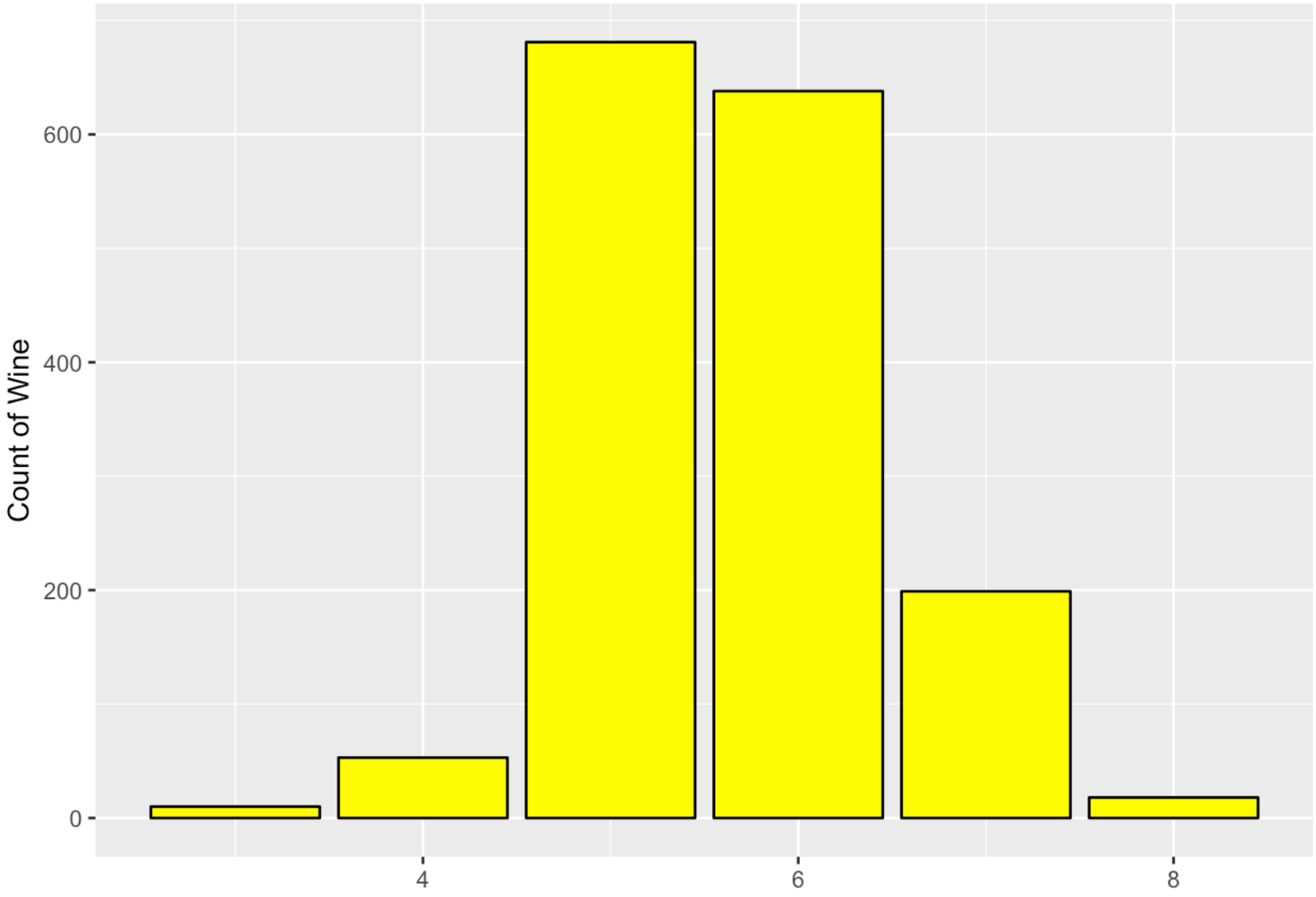
OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model.

Yes, I created a linear model and noticed that sulphates had the strongest impact on the quality of wine. pH, citric acid, and alcohol also are somewhat impacting the quality of wine but not as much as sulphates The dataset contains most of the wines with average quality. It would have been good if the data was varied and had wines from different qualities

FINAL PLOT AND SUMMARY

PLOT 1

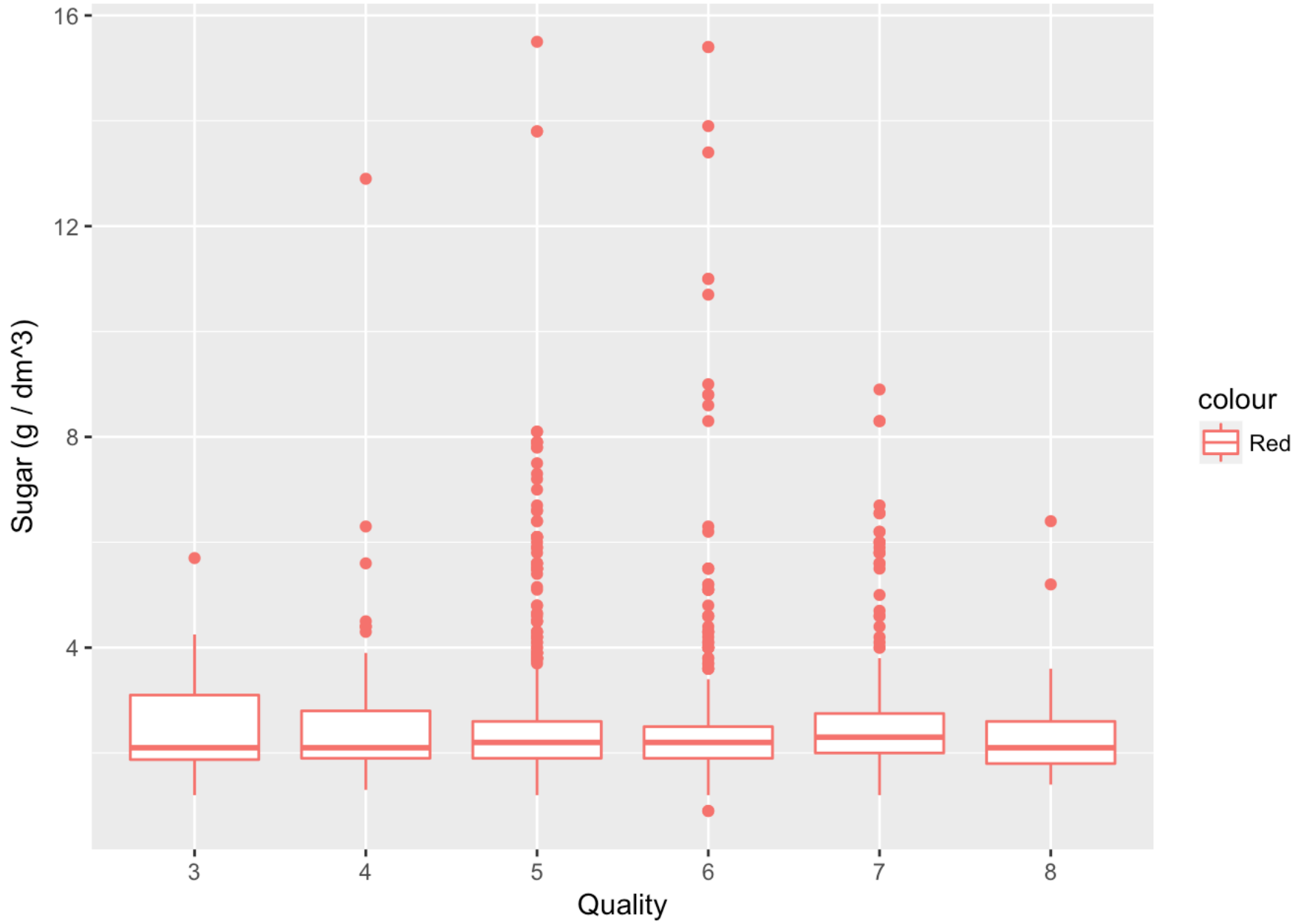
```
#Univariate Analysis of quality
ggplot(aes(x=quality , color=I('black'), fill=I('yellow')), data=winequality)+
  geom_bar()+
  xlab("Quality Rating of Wine")+
  ylab("Count of Wine")
```



The plot explains the quality in the dataset. Any conclusions made from this dataset will be applied to the “average” quality wines

PLOT 2

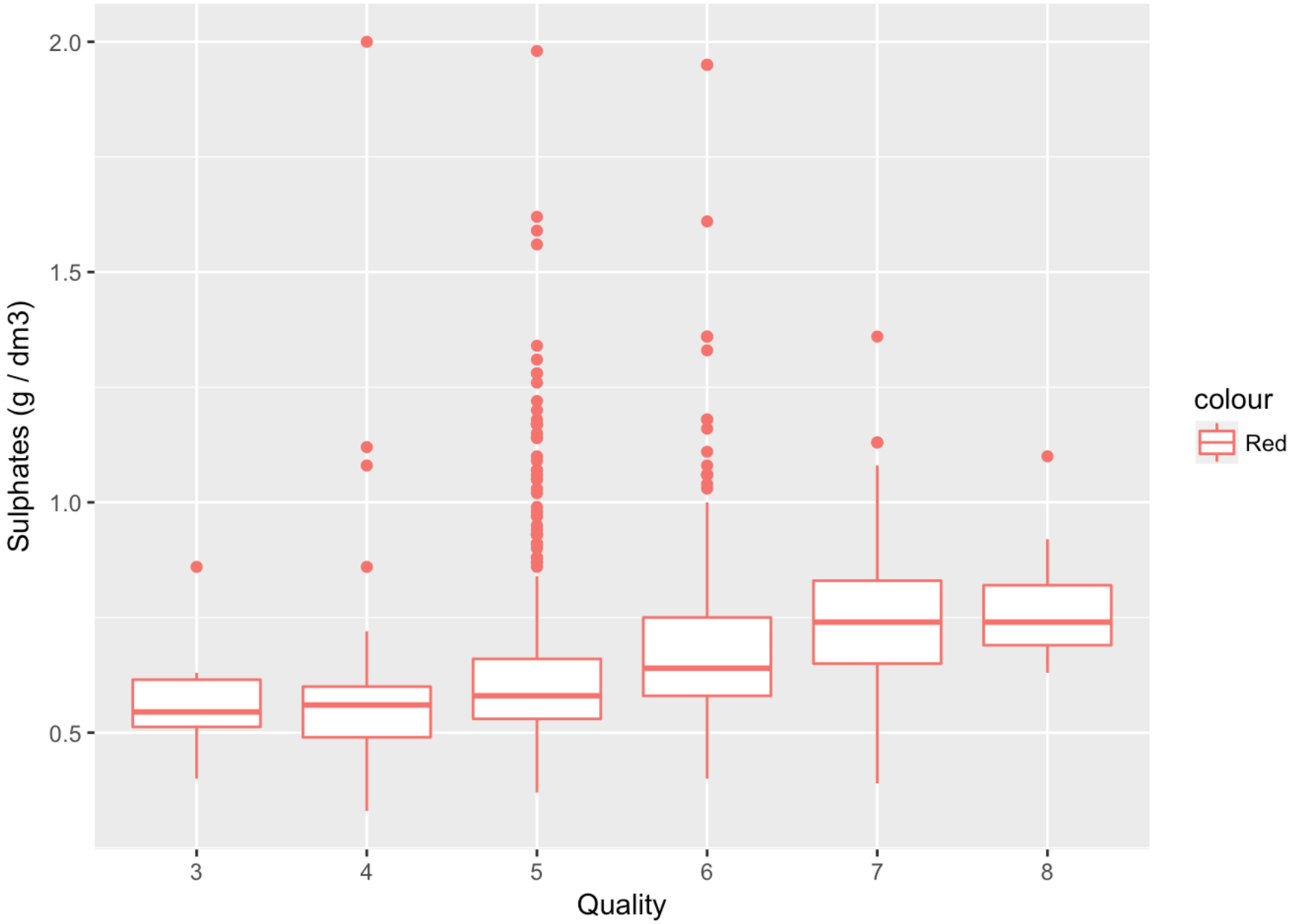
```
#Relation between quality and Sugar
ggplot(aes(y=residual.sugar, x=qualityFactor, color="Red"), data=winequality)+
  geom_boxplot()+
  xlab("Quality")+
  ylab("Sugar (g / dm^3)")
```



To my surprise, residual sugar has no impact on the quality of wine.

PLOT 3

```
#Relation between quality and sulphates
ggplot(aes(y=sulphates, x=qualityFactor, color="Red"), data=winequality)+
  geom_boxplot()+
  xlab("Quality")+
  ylab("Sulphates (g / dm3)")
```



Sulphates play a major role in determining the quality of the wine

REFLECTION

The dataset provided only had wines with average quality. Not many extreme quality wines were part of the dataset. As a result during exploration, it was very difficult to come to a satisfying conclusion provided that I knew most of the data is from wine quality 5 and 6.

I was also surprised to see how residual sugar and sulfur did not play as big a role in defining the quality as I assumed it will.

For future explorations, I would like to get more wines in quality other than 5 and 6. Also in the current dataset , there was just one categorical variable. Adding more categorical variable for future exploration will surely help