RED WINE QUALITY DATA ANALYSIS

Author : Gorkem Berk SAHAN

Date : 1-Sept-2017

==========

Intro from data source :

This introduction help us to understand data structure ;

Attribute information:

Input variables (based on physicochemical tests):

1 - fixed acidity (tartaric acid - g / dm^3)

2 - volatile acidity (acetic acid - g / dm^3)

3 - citric acid (g / dm^3)

4 - residual sugar (g / dm^3)

5 - chlorides (sodium chloride - g / dm^3

6 - free sulfur dioxide (mg / dm^3)

7 - total sulfur dioxide (mg / dm^3)

8 - density (g / cm^3)

9 - pH

10 - sulphates (potassium sulphate - g / dm3)

11 - alcohol (% by volume)

Output variable (based on sensory data):

12 - quality (score between 0 and 10)

Description of attributes:

1 - fixed acidity: most acids involved with wine or fixed or nonvolatile (do not evaporate readily)

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

3 - citric acid: found in small quantities, citric acid can add ‘freshness’ and flavor to wines

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

5 - chlorides: the amount of salt in the wine

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

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

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

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

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

11 - alcohol: the percent alcohol content of the wine

Output variable (based on sensory data): 12 - quality (score between 0 and 10)

Red wine data has 1599 rows and 13 variables like below;

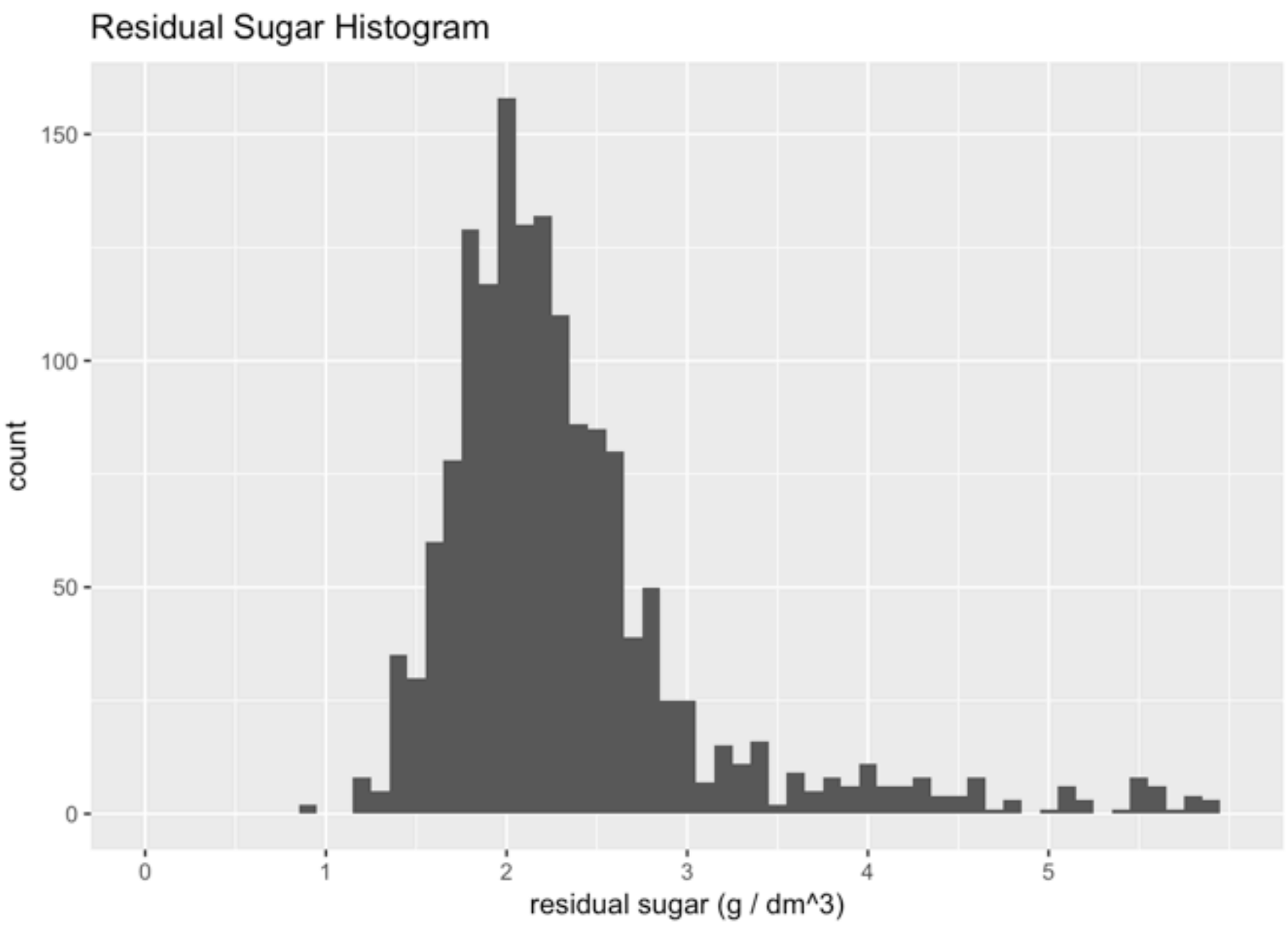
```
##  [1] "X"                   "fixed.acidity"        "volatile.acidity"
##  [4] "citric.acid"         "residual.sugar"       "chlorides"
##  [7] "free.sulfur.dioxide" "total.sulfur.dioxide" "density"
## [10] "pH"                  "sulphates"            "alcohol"
## [13] "quality"
```

```
##        X          fixed.acidity   volatile.acidity  citric.acid
##  Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000
##  1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090
##  Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260
##  Mean   : 800.0   Mean   : 8.32   Mean   :0.5278   Mean   :0.271
##  3rd Qu.:1199.5   3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420
##  Max.   :1599.0   Max.   :15.90   Max.   :1.5800   Max.   :1.000
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.900   Min.   :0.01200   Min.   : 1.00
##  1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00
##  Median : 2.200   Median :0.07900   Median :14.00
##  Mean   : 2.539   Mean   :0.08747   Mean   :15.87
##  3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00
##  Max.   :15.500   Max.   :0.61100   Max.   :72.00
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  6.00       Min.   :0.9901   Min.   :2.740   Min.   :0.3300
##  1st Qu.: 22.00       1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500
##  Median : 38.00       Median :0.9968   Median :3.310   Median :0.6200
##  Mean   : 46.47       Mean   :0.9967   Mean   :3.311   Mean   :0.6581
##  3rd Qu.: 62.00       3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300
##  Max.   :289.00       Max.   :1.0037   Max.   :4.010   Max.   :2.0000
##     alcohol         quality
##  Min.   : 8.40   Min.   :3.000
##  1st Qu.: 9.50   1st Qu.:5.000
##  Median :10.20   Median :6.000
##  Mean   :10.42   Mean   :5.636
##  3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :14.90   Max.   :8.000
```

Univariate Plots Section

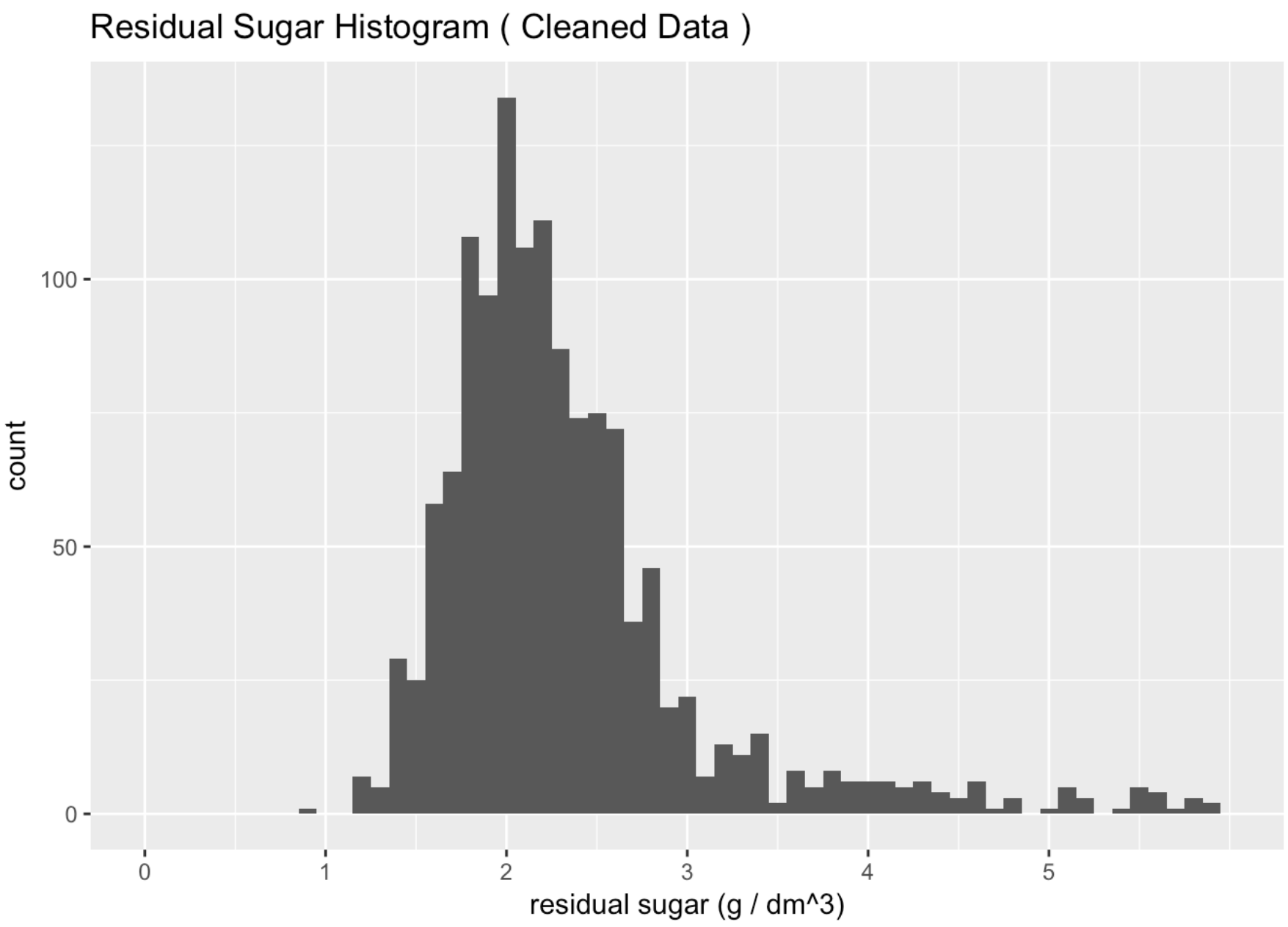
General distrubitions of each variable

First of all, lets see the all variables distrubition and try to understand what data say to us.

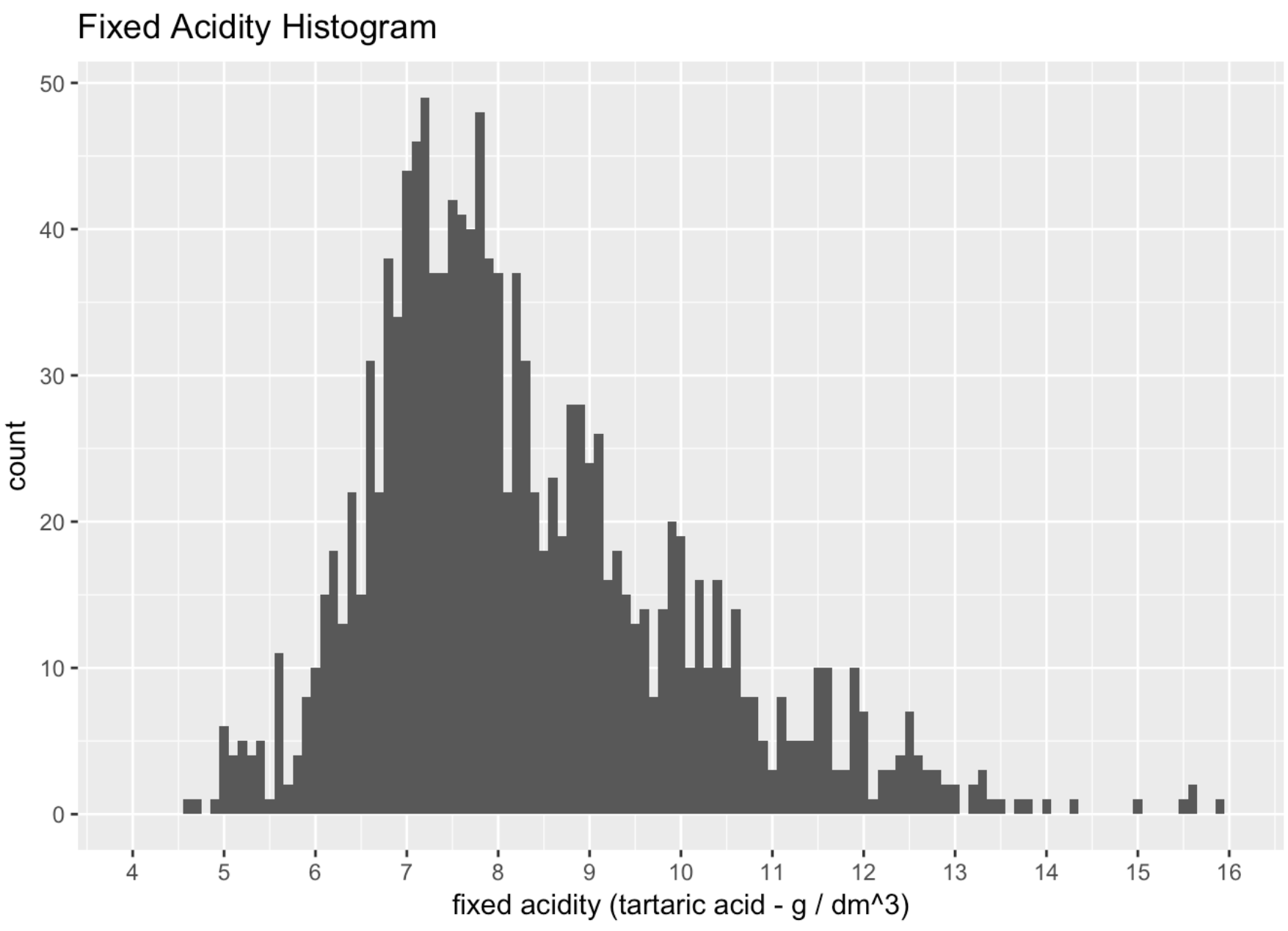


- As mentioned in data desc. it’s rare to find wines with less than 1 gram/liter sugar, in our dataset only 2 and these two wine which its datas has same values. so we have to clear data dublicates.

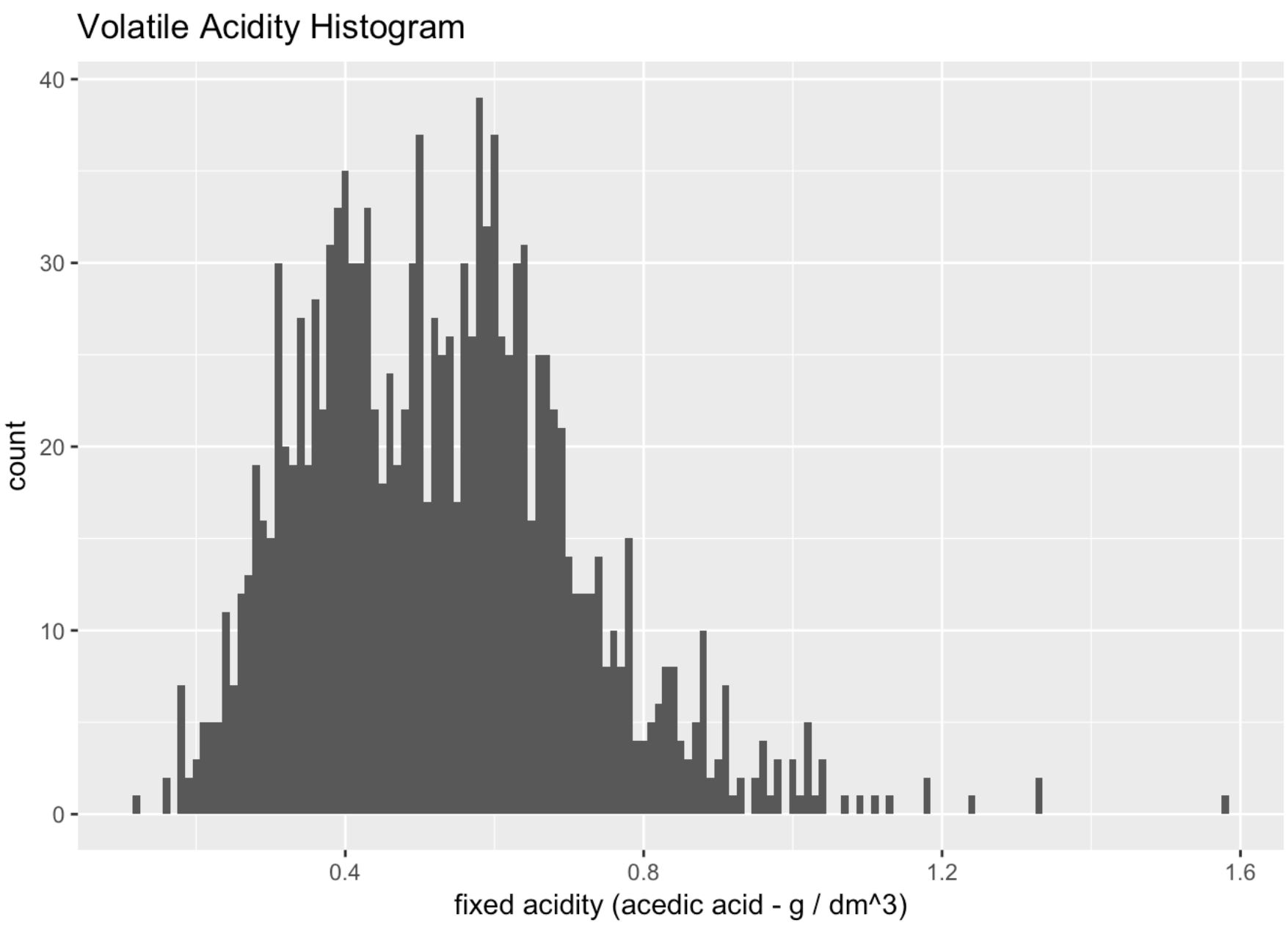
First column of dataset is rownum so we can ignore this column and expect this column , and after cleaning dublicates we have 1359 unique datas. so lets see the barchart above again



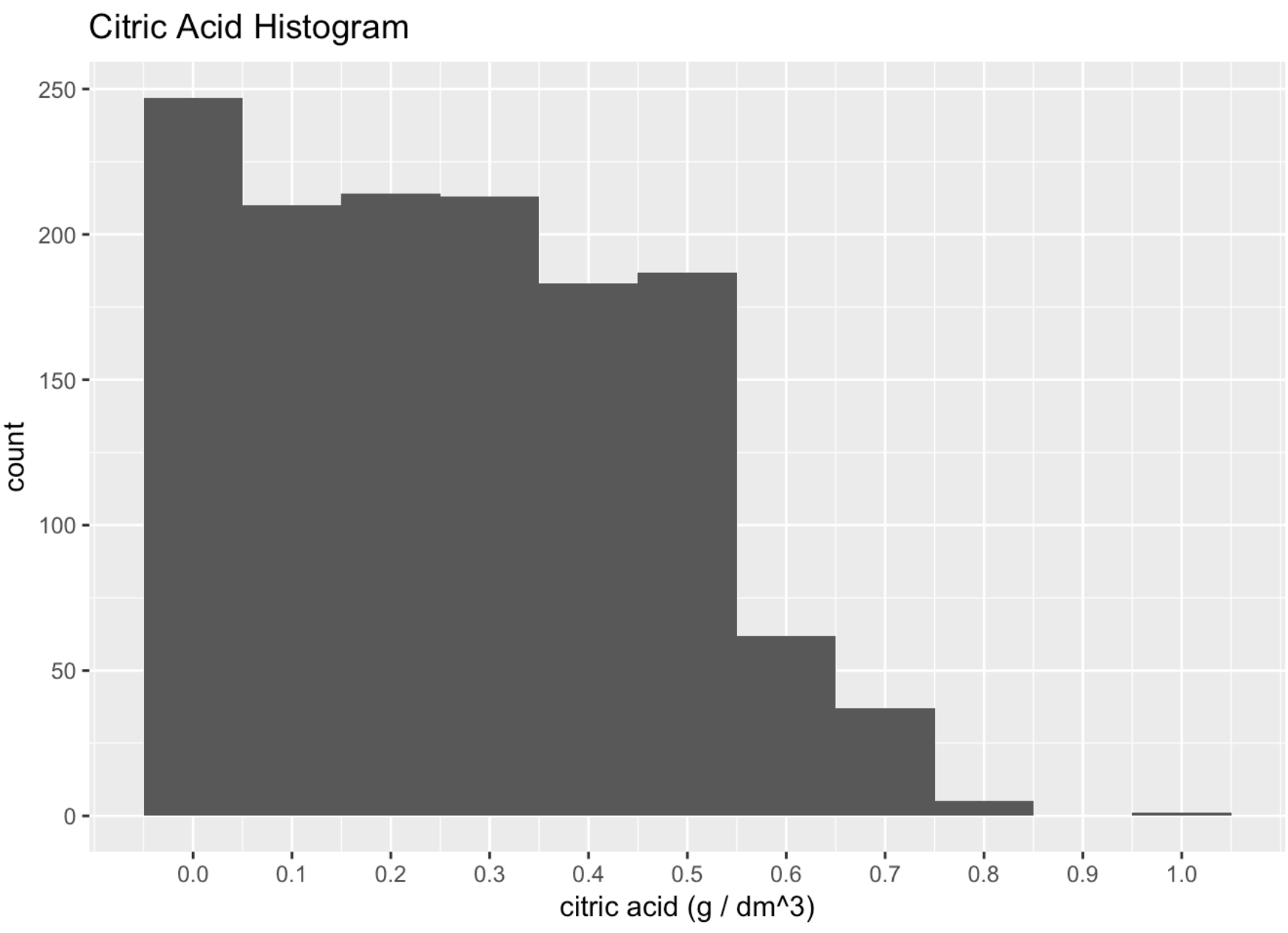
There is not any massive changing after removed dublicates…



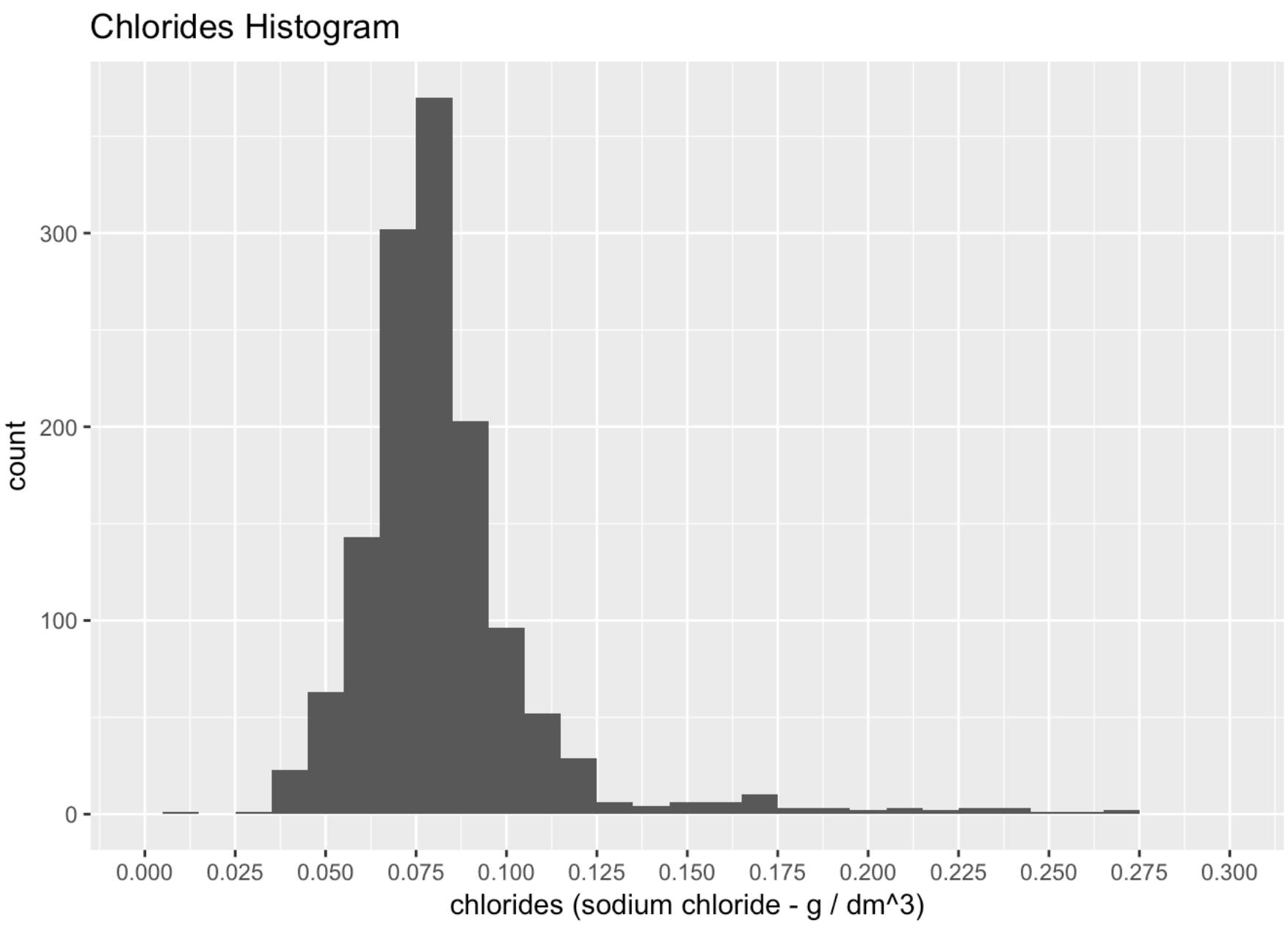
Fixed acidity is normal distrubuted, most of wines are between 7-9 g / L(dm^3)



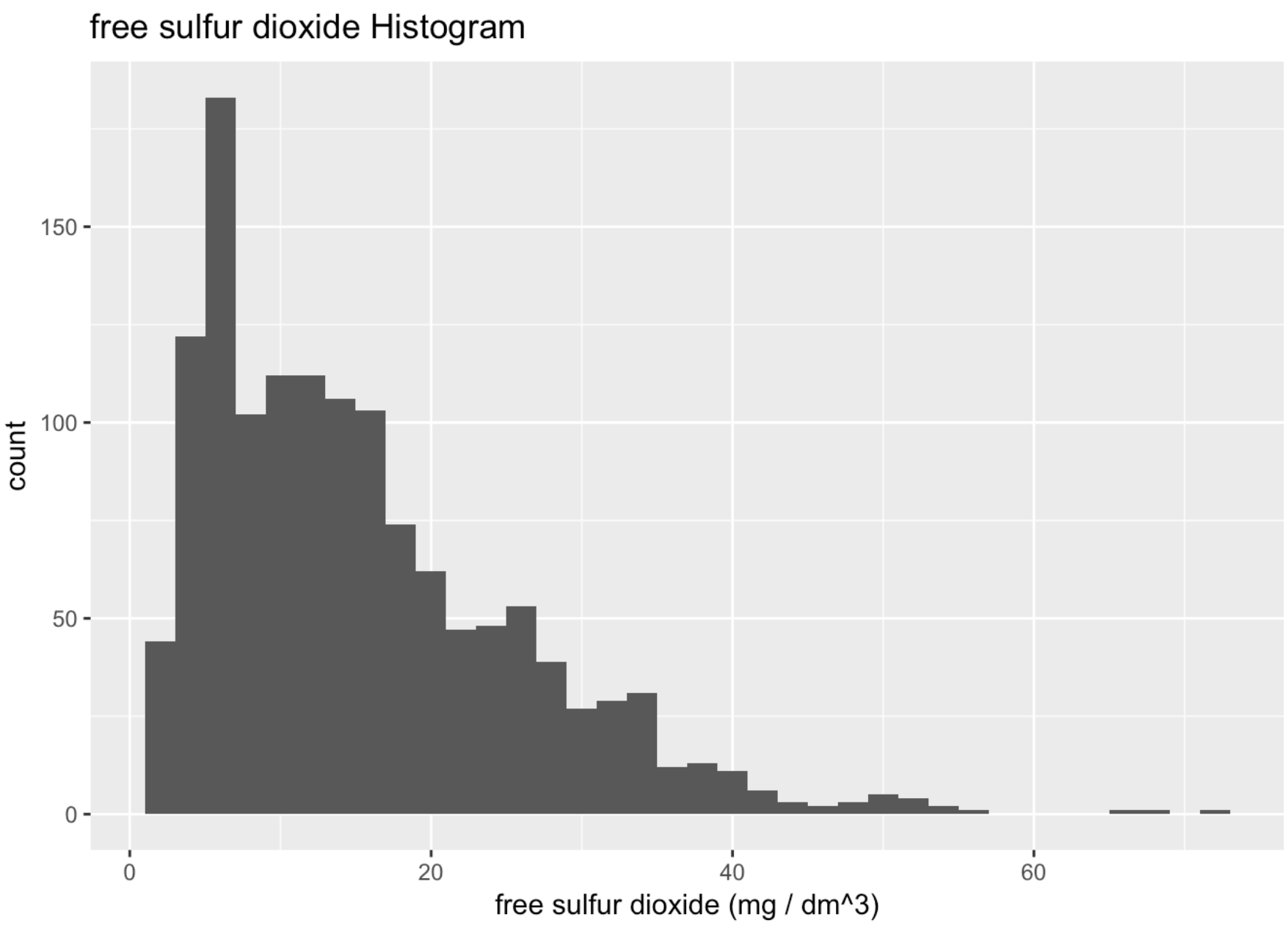
Volatile acidity is normal distrubuted. and most of wines between 0.4-0.6 g/L



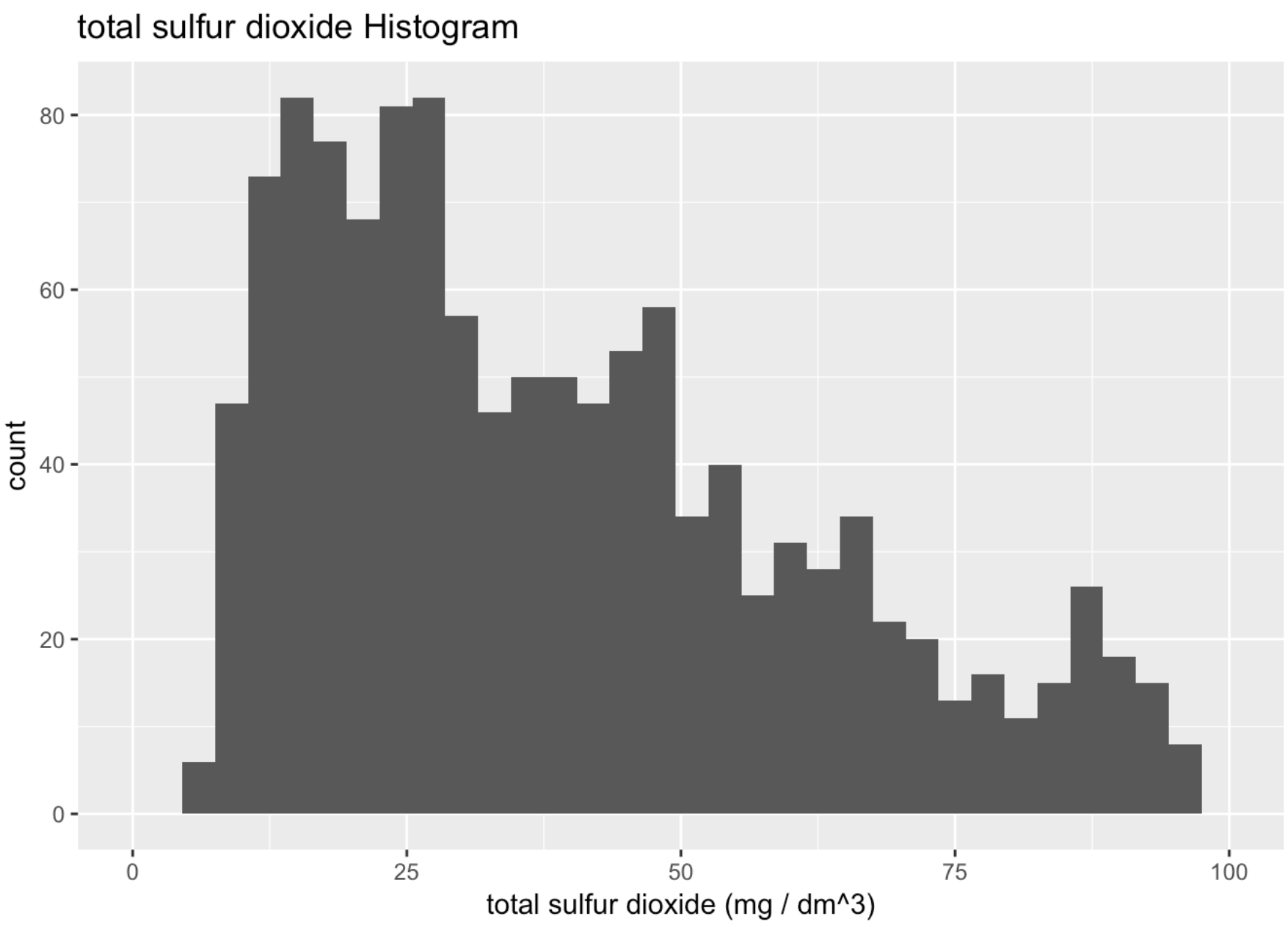
some wines hasn’t got citric acid, “citric acid can add ‘freshness’ and flavor to wines”, so we can consider relation of citric acid and quality



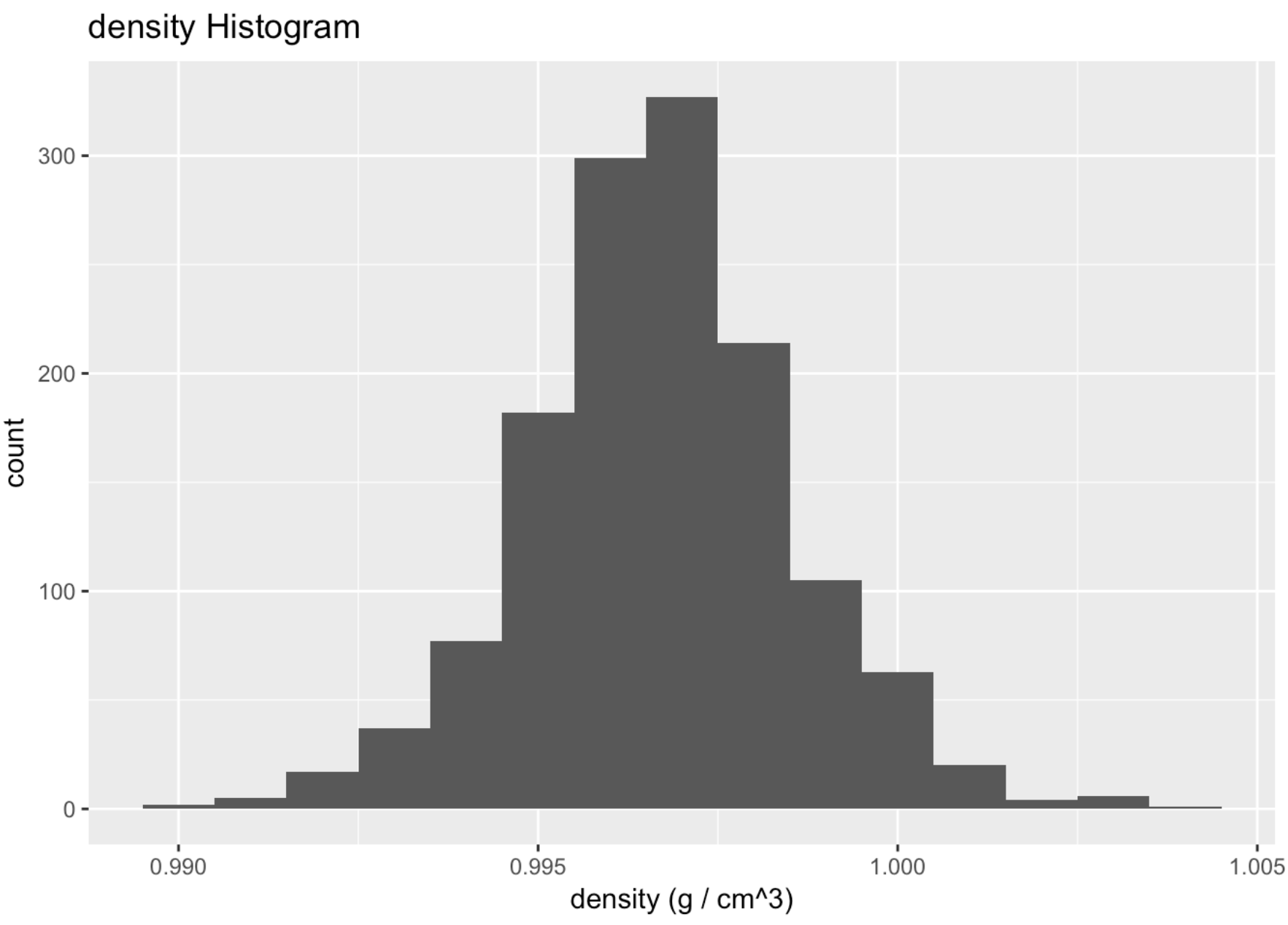
chlorides: the amount of salt in the wine, and most of it has 0.1 g/L



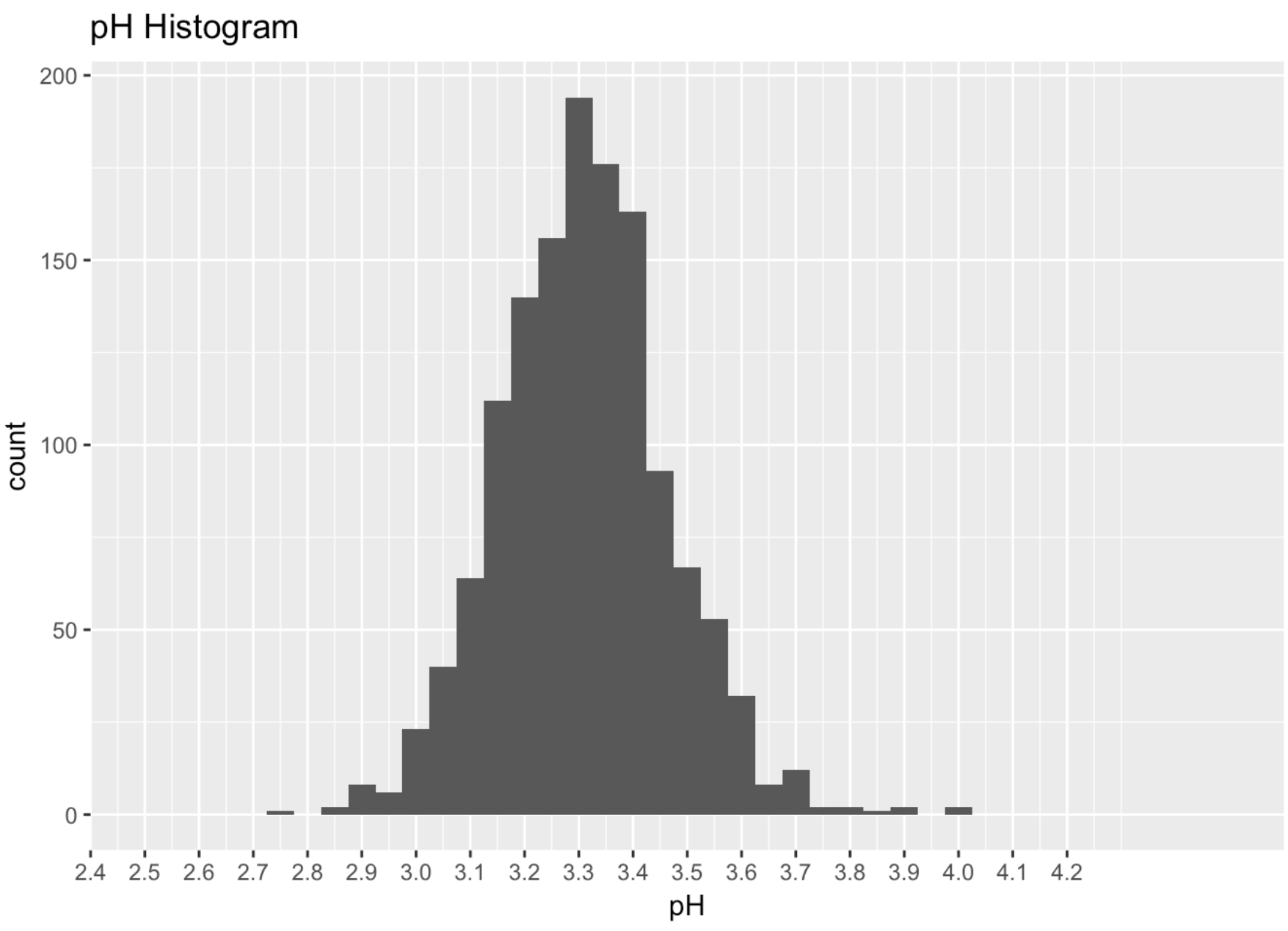
free sulfur dioxide : it prevents microbial growth and the oxidation of wine it is positively skewed ,



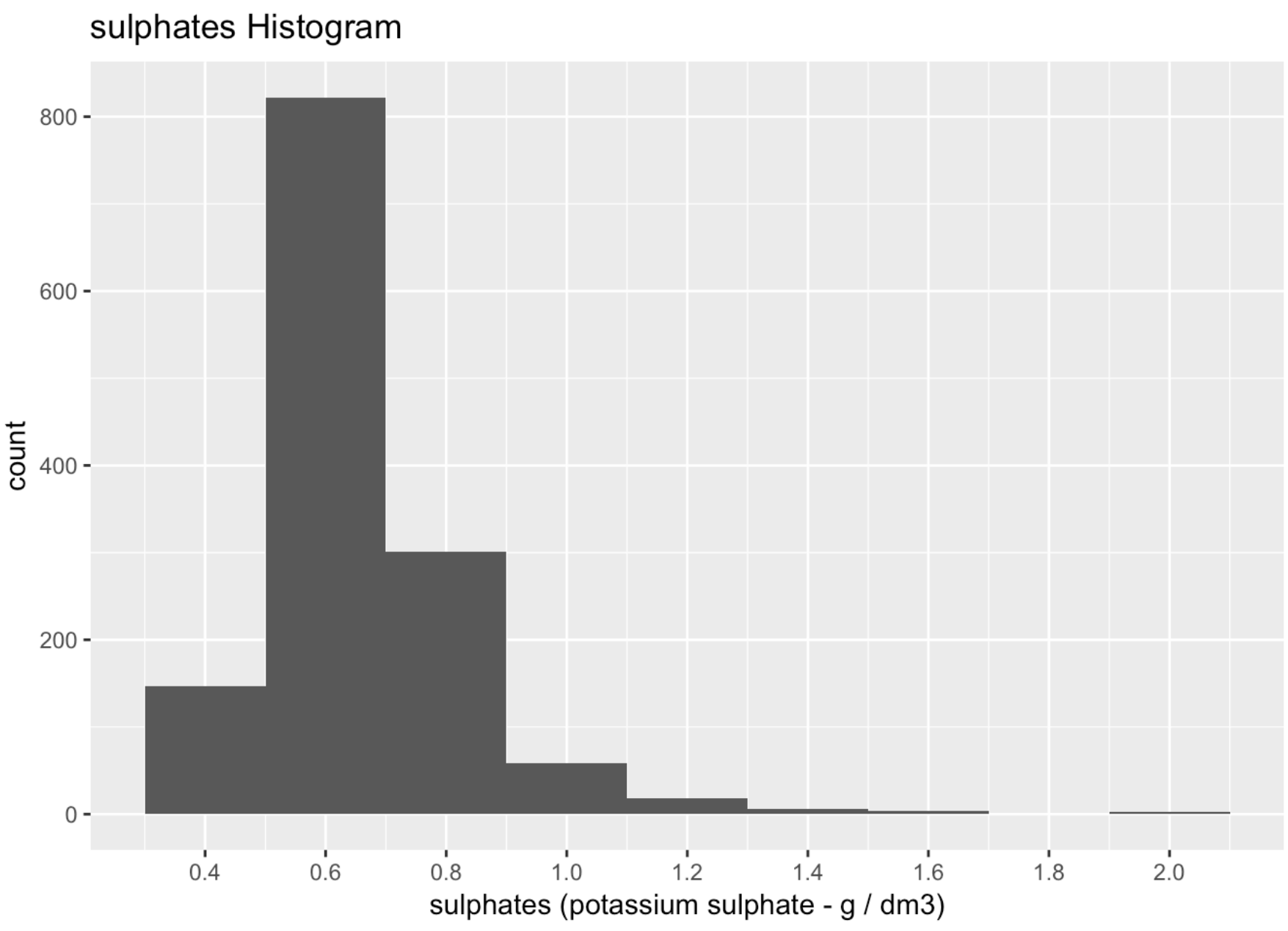
from descriptions of data “over 50 ppm, SO2 becomes evident in the nose and taste of wine” and searching on Google for ppm, 50 ppm means % 0.005



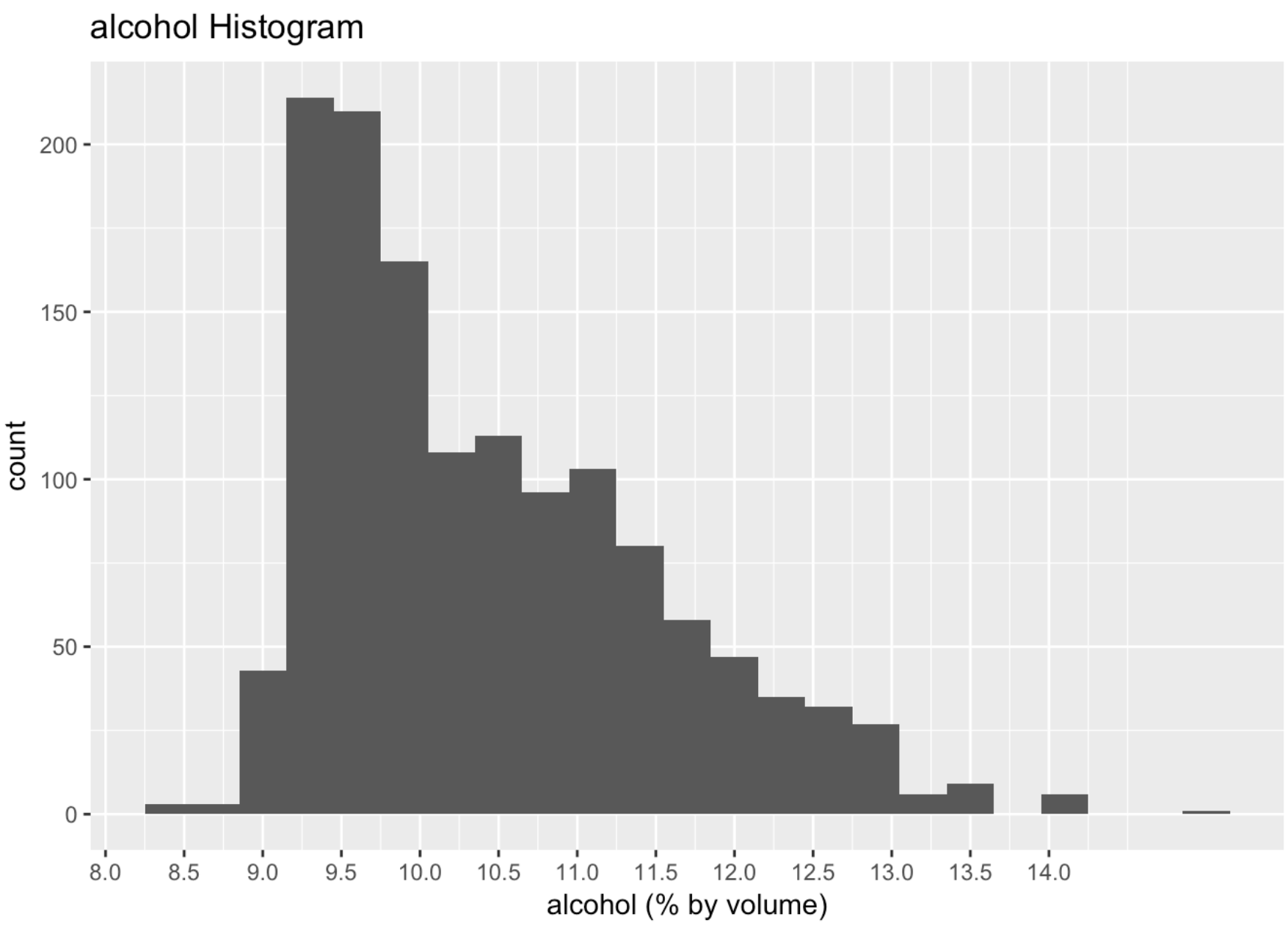
density is normal distributed. Density depend on solvent and solver, what is the relation of density , alcohol and sugar ? we are going to see realtion on bivariate section.



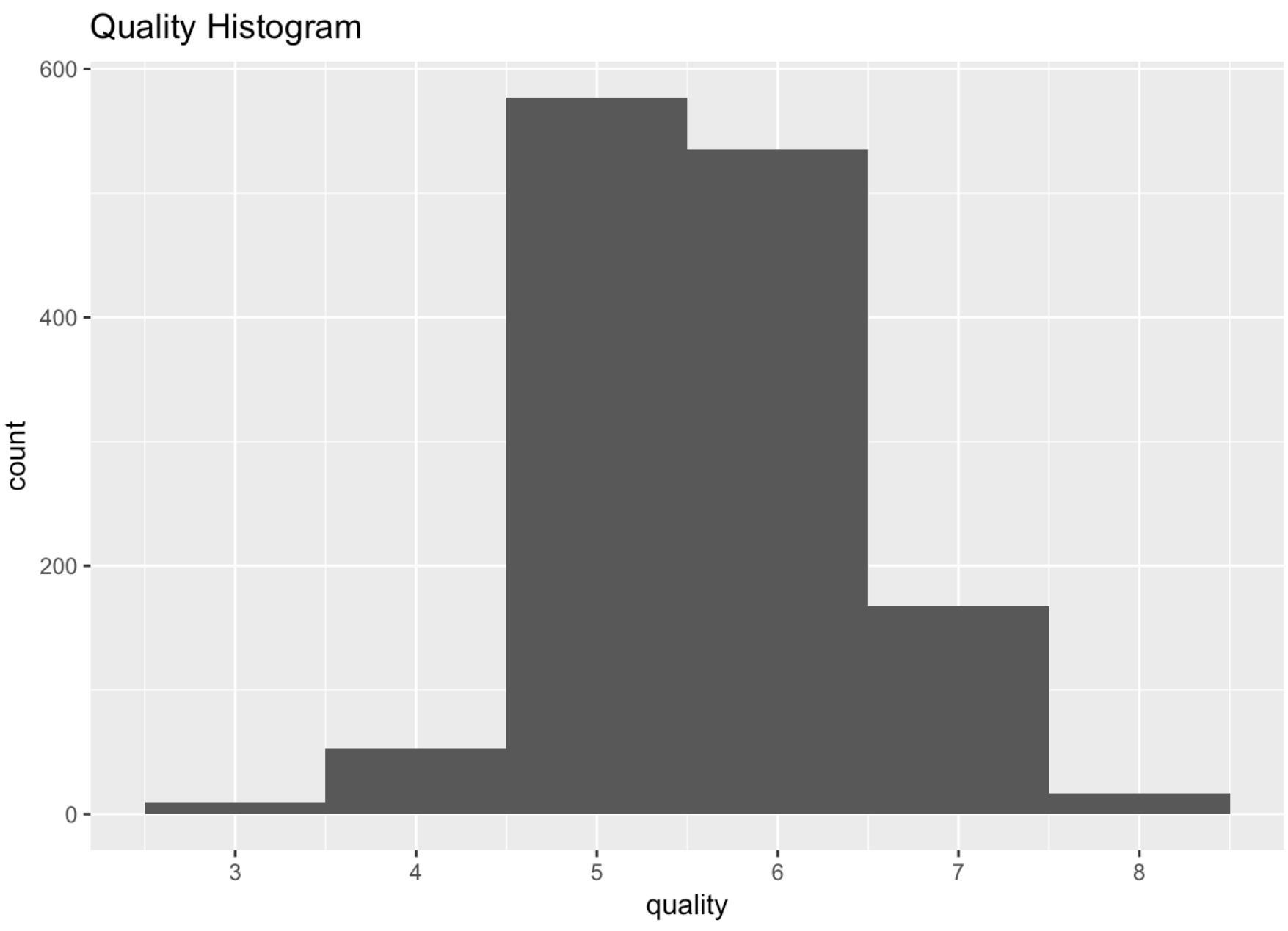
pH level is normal distributed and all wines' pH levels are between 2.7 - 4.0 that means acidic



Sulphates level is normal distributed, After googling sulphades, we can see that sulphate is a salt. so it effects on wine’s taste so we can consider what is relation of quality and sulphade ?



Alcohol level is positively skewed.



Tested wines quality is normal distributed and mean is 5.623

Univariate Analysis

We have 1359 unique data and 13 variables all input variables are number and we don’t have any ordered factor variables, but quality is a output variable and may use ordered factor variable

low Quality —> high Quality

```
0 ----------> 10
```

other observations :

- Most wines contains 1.5g - 2.5g sugar per liter
- Most wines contains 0.05 - 0.1 g salt per liter
- Most wines contains 7g - 9g tartaric acid per liter

What is the structure of your dataset?

Input variables (based on physicochemical tests):

1 - fixed acidity (tartaric acid - g / dm^3)

2 - volatile acidity (acetic acid - g / dm^3)

3 - citric acid (g / dm^3)

4 - residual sugar (g / dm^3)

5 - chlorides (sodium chloride - g / dm^3

6 - free sulfur dioxide (mg / dm^3)

7 - total sulfur dioxide (mg / dm^3)

8 - density (g / cm^3)

9 - pH

10 - sulphates (potassium sulphate - g / dm3)

11 - alcohol (% by volume)

Output variable (based on sensory data):

12 - quality (score between 0 and 10)

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

main features are density, pH, alcohol , chlorides. I'd like to see input variables relations each other and create a predictive model from input variables to output(quality) variable.

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

density is depend on sugar and alcohol content, we can consider its effect on quality, and suphates acts as an antimicrobial and antioxidant so it matters for quality

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

yes, we create an variable that keeps if volatile acid level is high (if outlier then 1 else 0). As we know high level volatile acid causes bad taste ,so we can control effects on quality.

Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the formof the data? If so, why did you do this?

No, all distributions is usual, I didn't need any trans

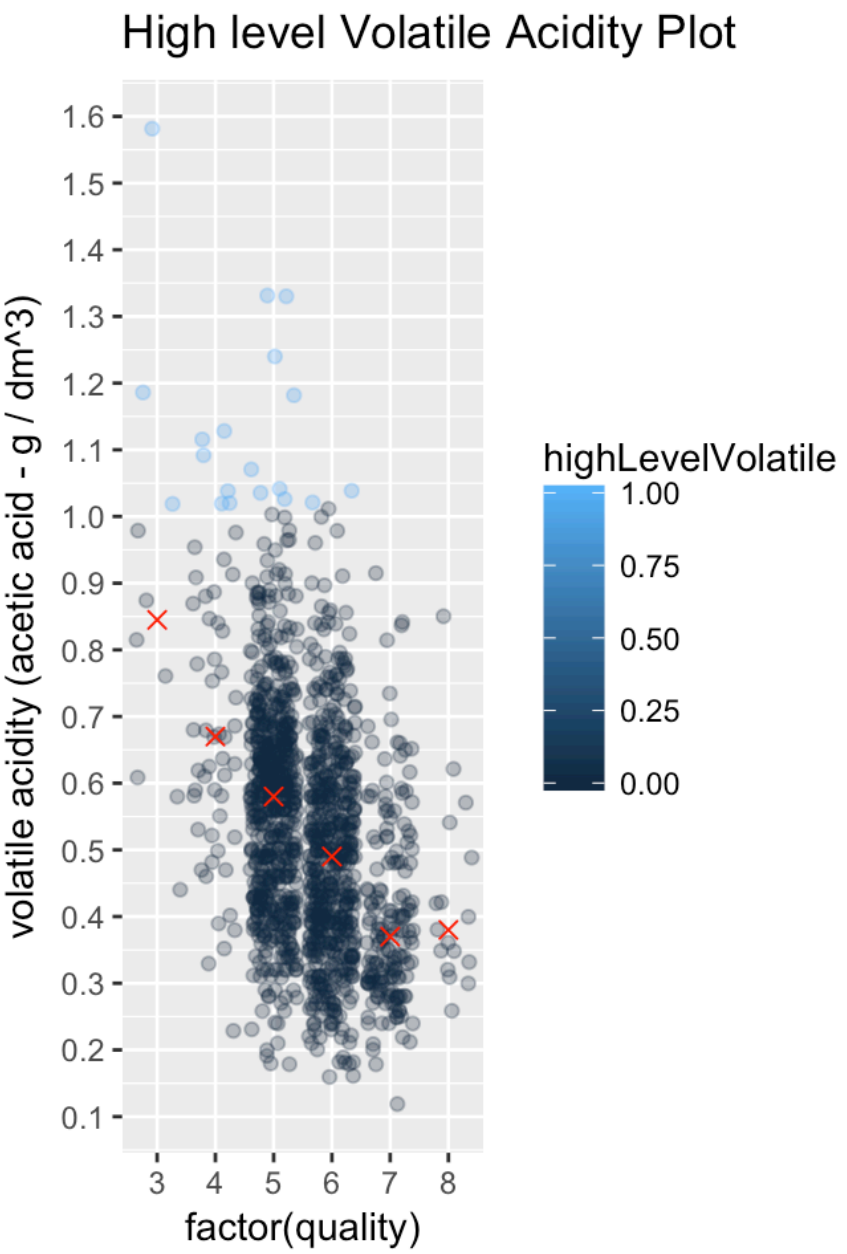
Bivariate Plots Section

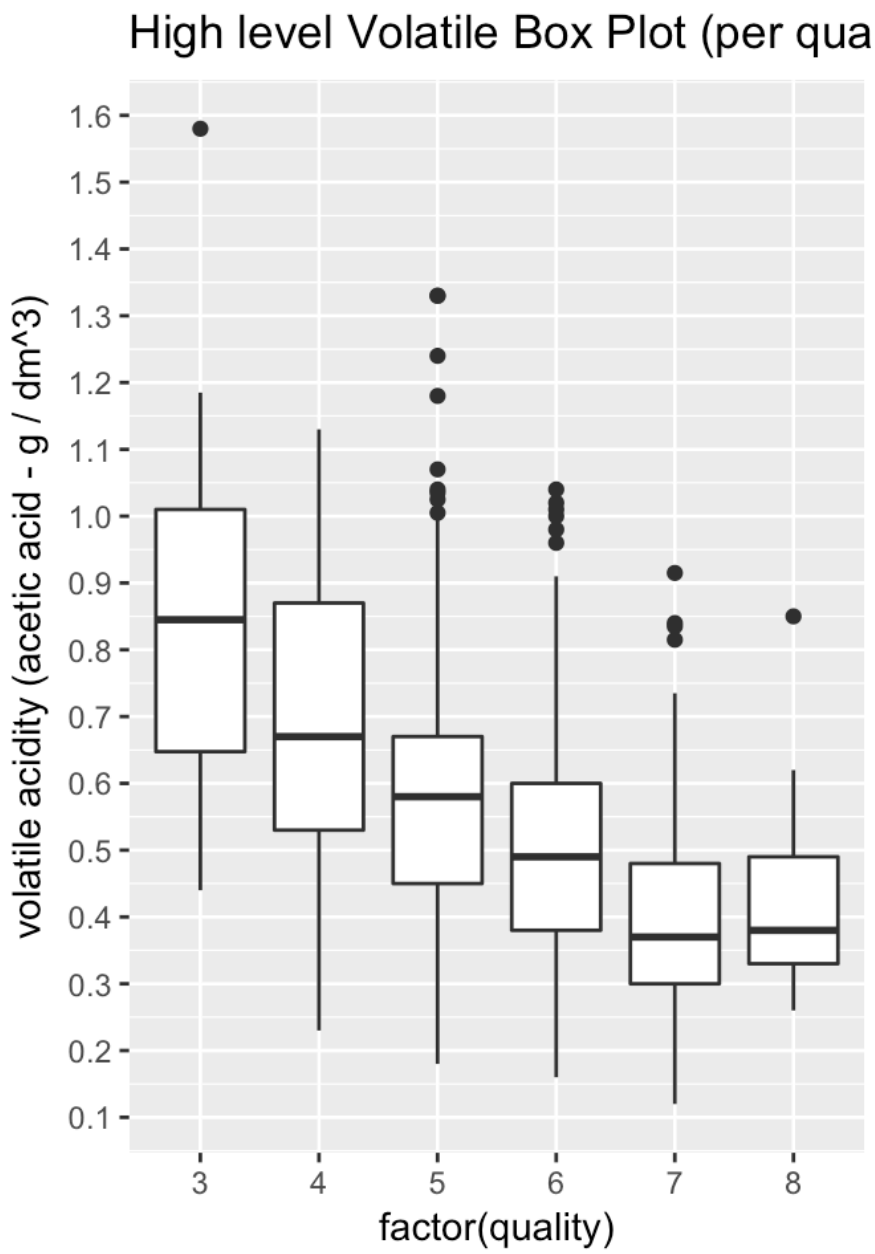
some variables' relationship each other ;

When look at the data structure and its descriptions , we can see that some quality descriptions like ;

- volatile acidity : which at too high of levels can lead to an unpleasant, vinegar taste so lets take a

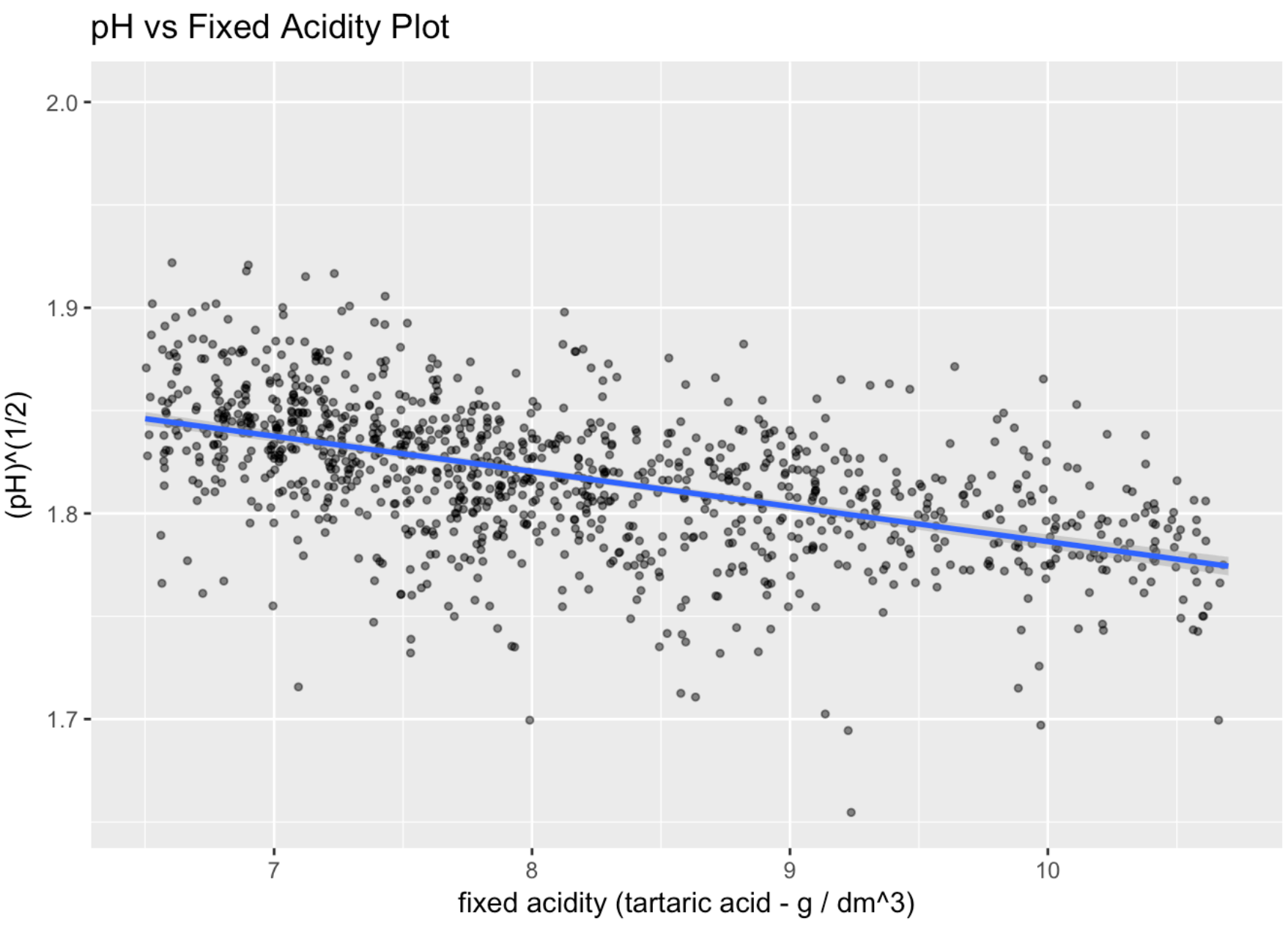
look to quality vs volatile acidity





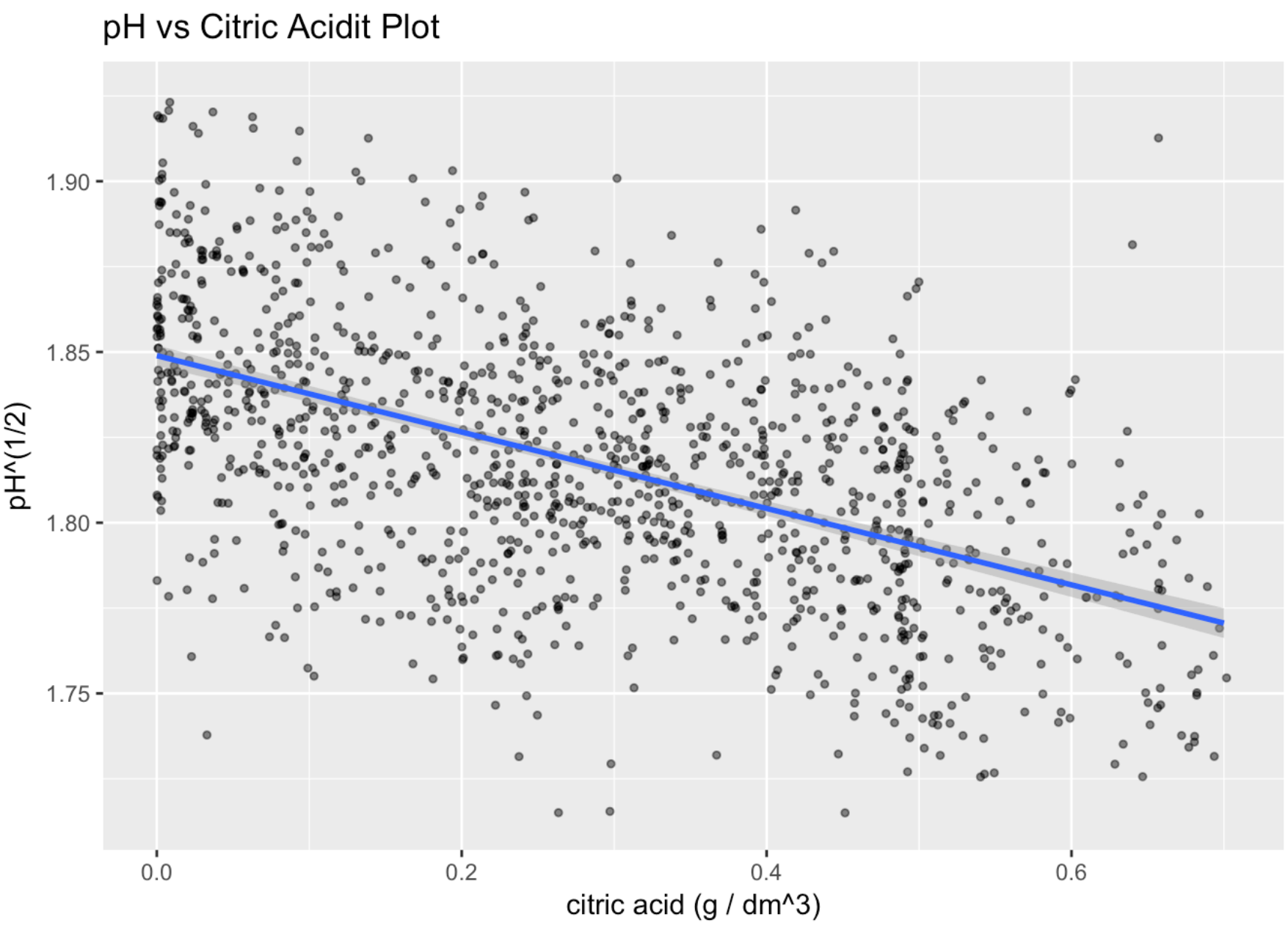
colored points are outlines, red cross are median of v.acidity per quality level, it seems like an relation between quality and volatile acidity but other variables can effect quality

```
##
##  Pearson's product-moment correlation
##
## data:  pH and fixed.acidity
## t = -34.797, df = 1357, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7137963 -0.6575190
## sample estimates:
##       cor
## -0.6866851
```

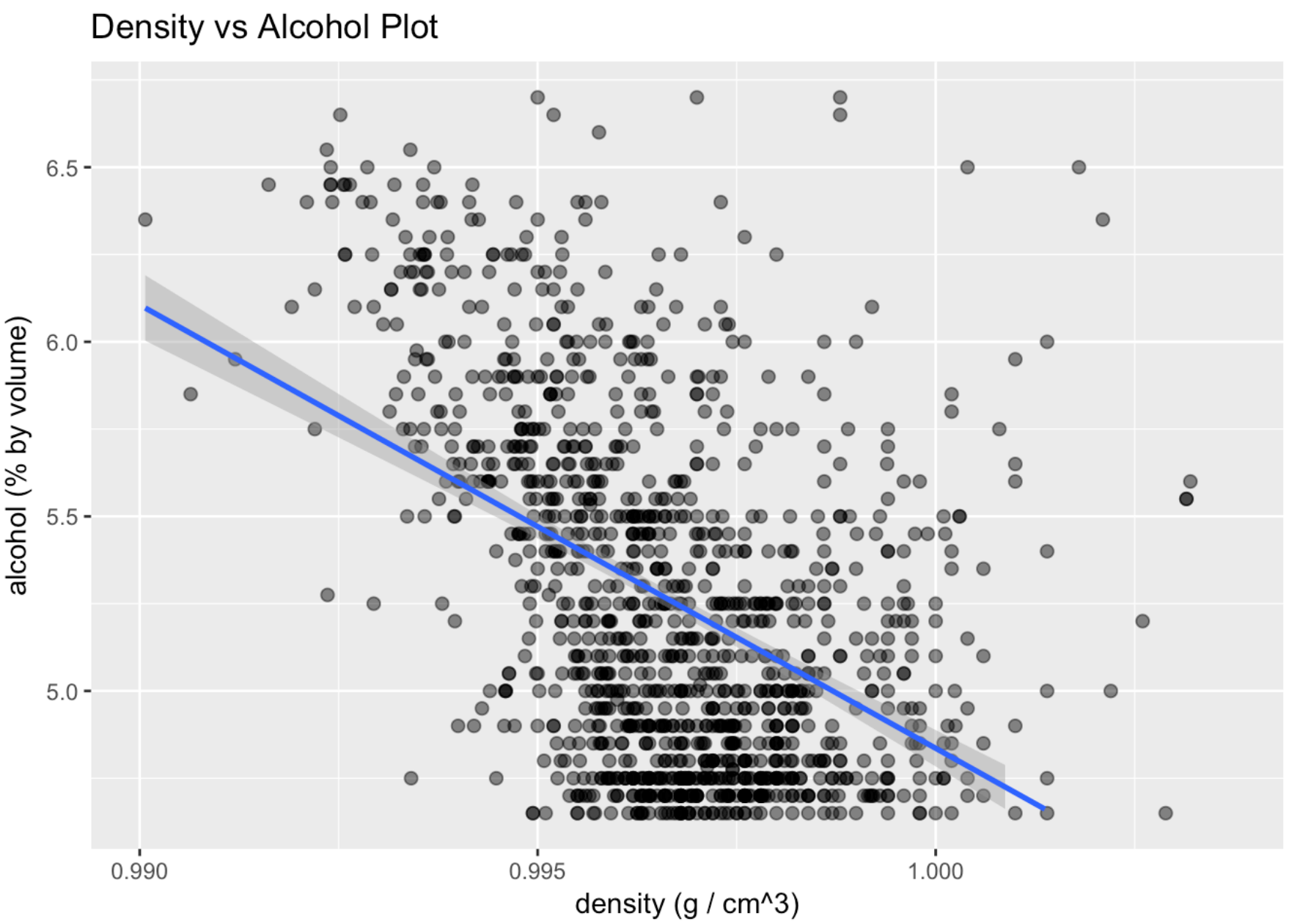


It seems fixed Acidity and pH level is related, correlation test is -0.6866851

```
##
##  Pearson's product-moment correlation
##
## data:  pH and citric.acid
## t = -24.279, df = 1357, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5863273 -0.5121207
## sample estimates:
##       cor
## -0.5503098
```



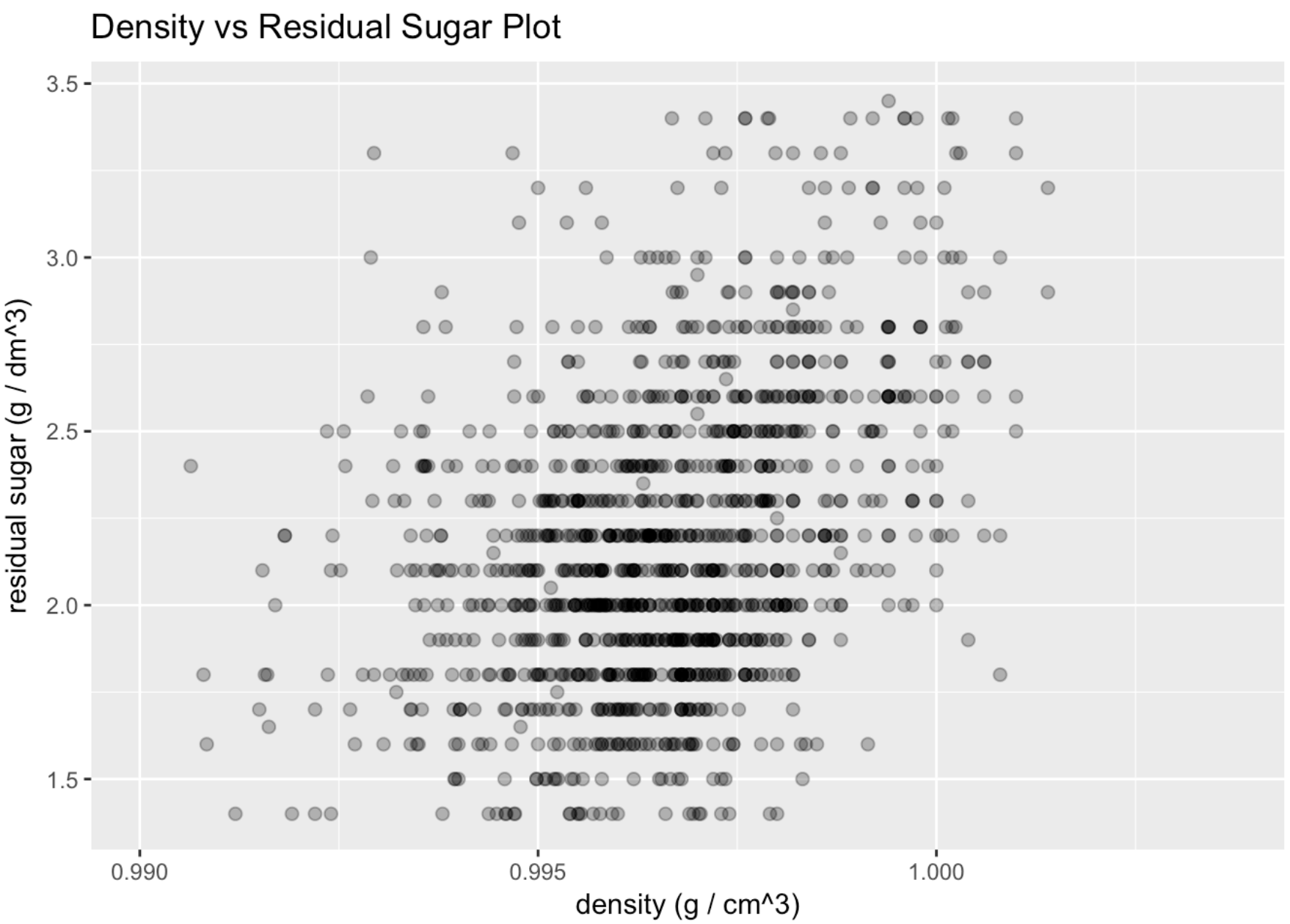
citric acid and pH is related, correlation is -0.5503098




```
##
##  Pearson's product-moment correlation
##
## data:  density and alcohol
## t = -21.553, df = 1357, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5435734 -0.4642874
## sample estimates:
##       cor
## -0.5049949
```

density and alcohol has an relation, correlation is -0.504

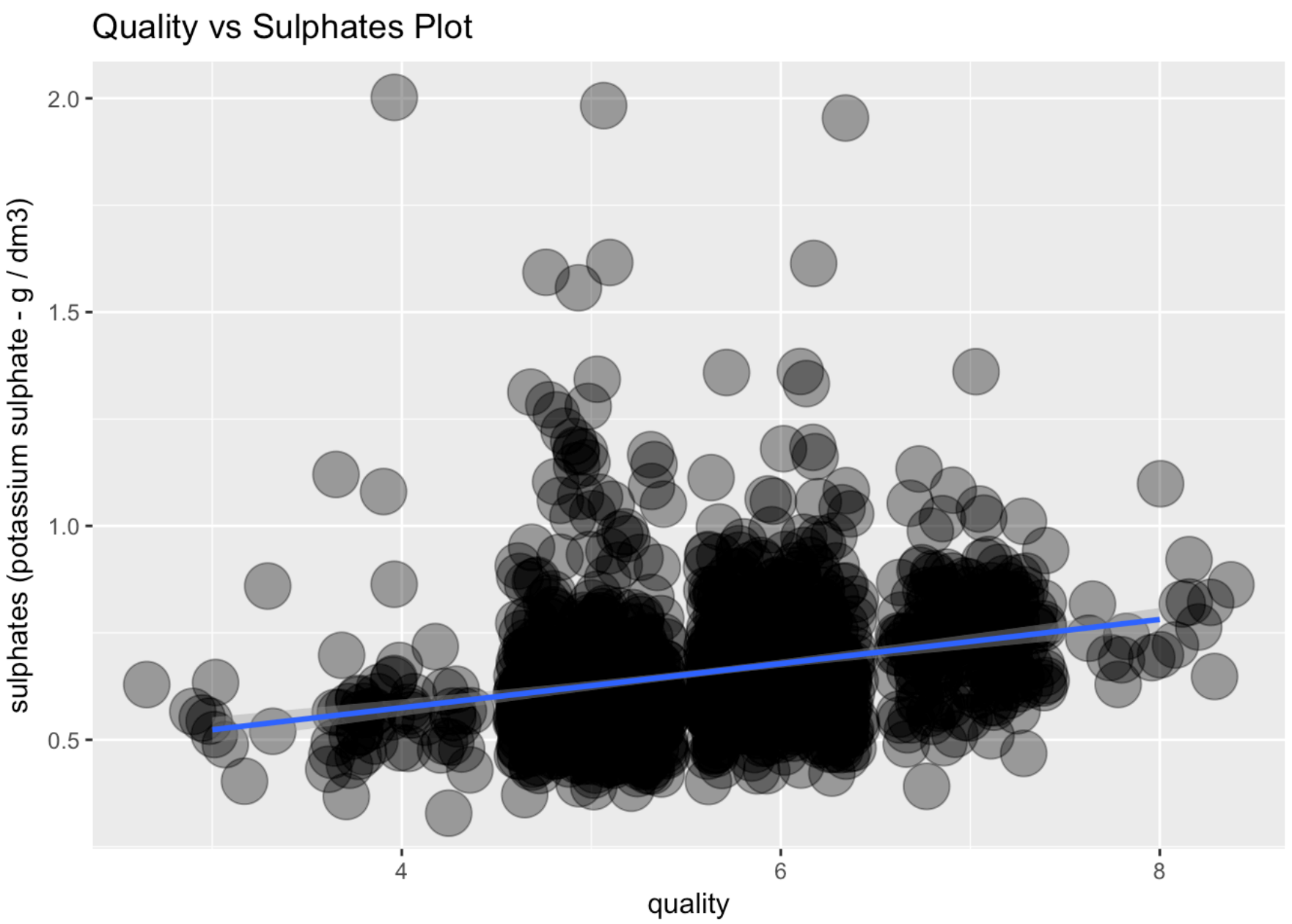
at the same time I hope that sugar has same relation with denstiy and ;




```
##
##  Pearson's product-moment correlation
##
## data:  density and residual.sugar
## t = 12.639, df = 1357, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2761121 0.3712904
## sample estimates:
##       cor
## 0.3245225
```

yes they have a relation but not much as alcohol and density , 0.3245

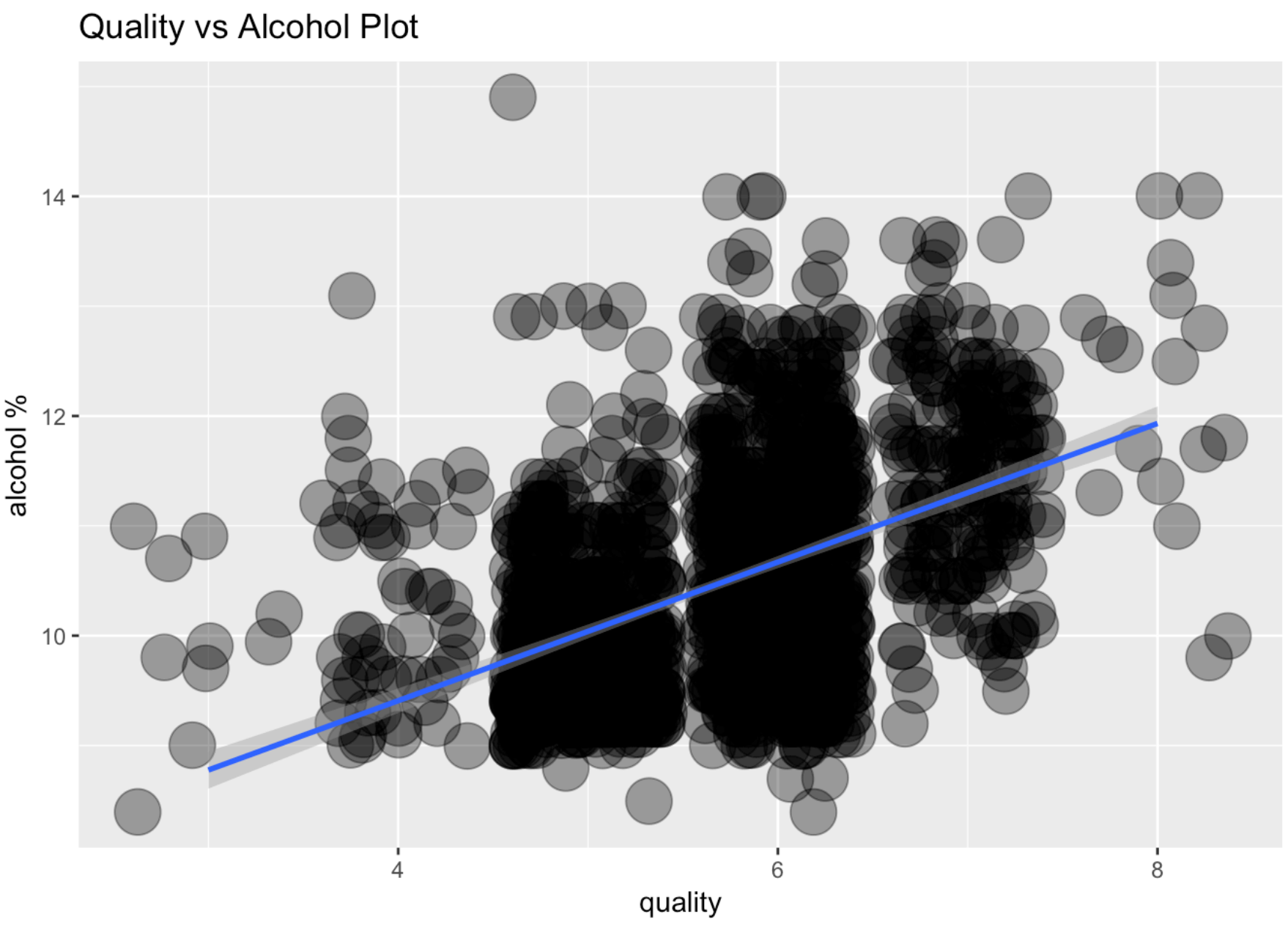
Relation between one input variable and output variable




```
##
##  Pearson's product-moment correlation
##
## data:  quality and sulphates
## t = 9.4641, df = 1357, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1982837 0.2980663
## sample estimates:
##       cor
## 0.2488351
```

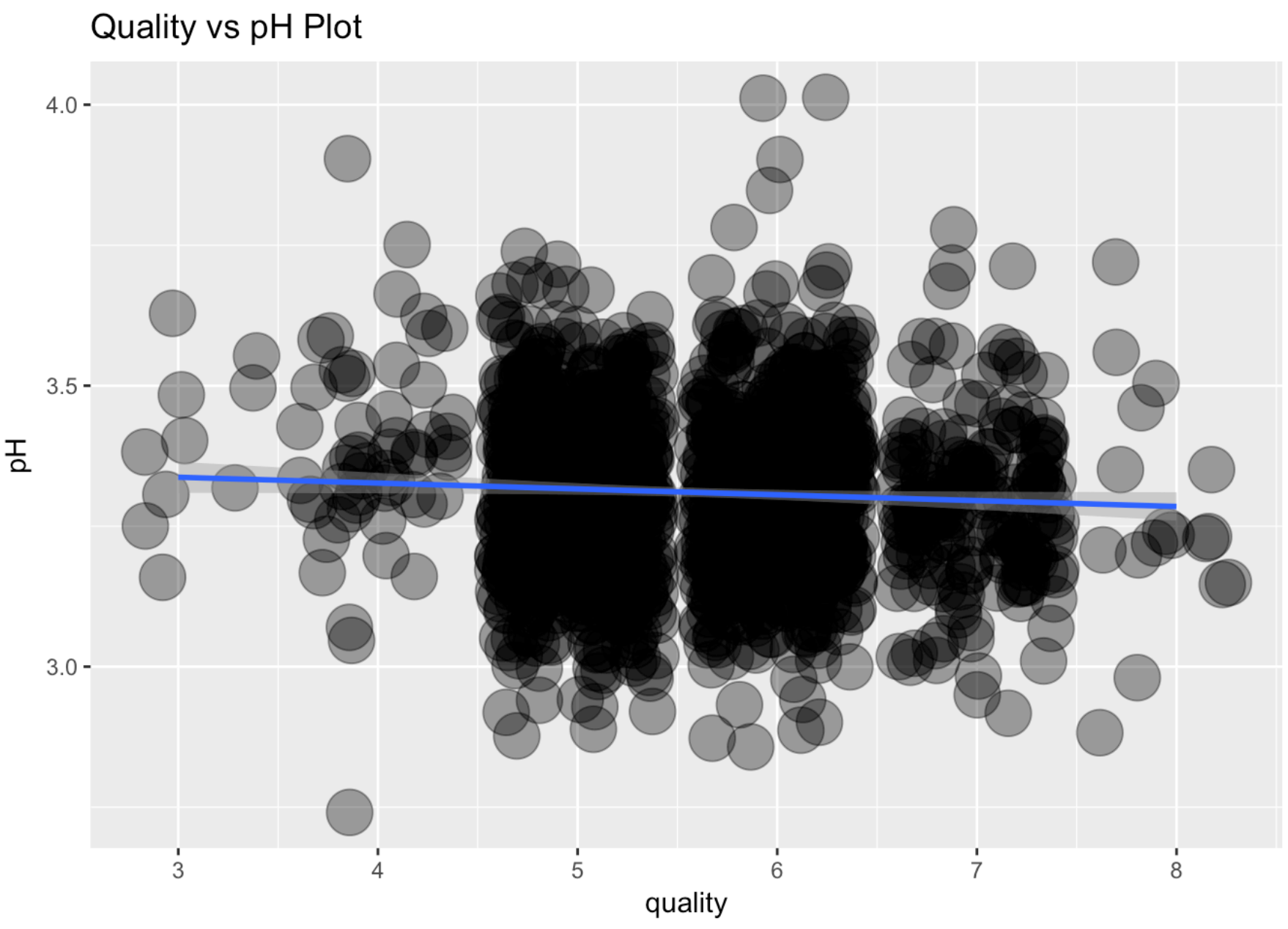
quality and sulphates relation, correlation is 0.2488351

```
##
##  Pearson's product-moment correlation
##
## data:  quality and alcohol
## t = 20.174, df = 1357, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4383646 0.5202301
## sample estimates:
##       cor
## 0.4803429
```

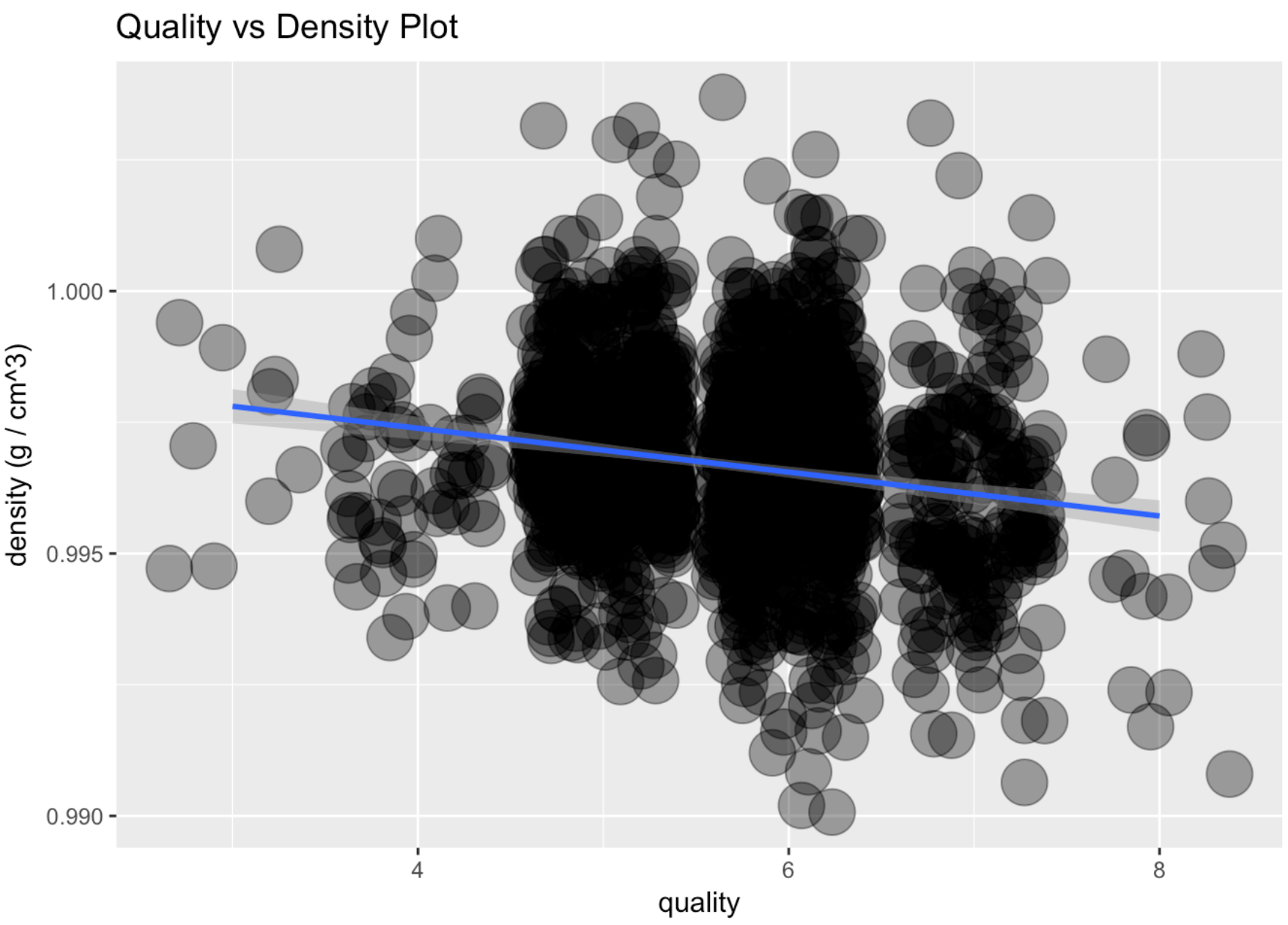


alcohol and quality relation , correlation is 0.4803429

```
##
##  Pearson's product-moment correlation
##
## data:  quality and pH
## t = -2.0382, df = 1357, p-value = 0.04172
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.108102657 -0.002076103
## sample estimates:
##         cor
## -0.05524511
```

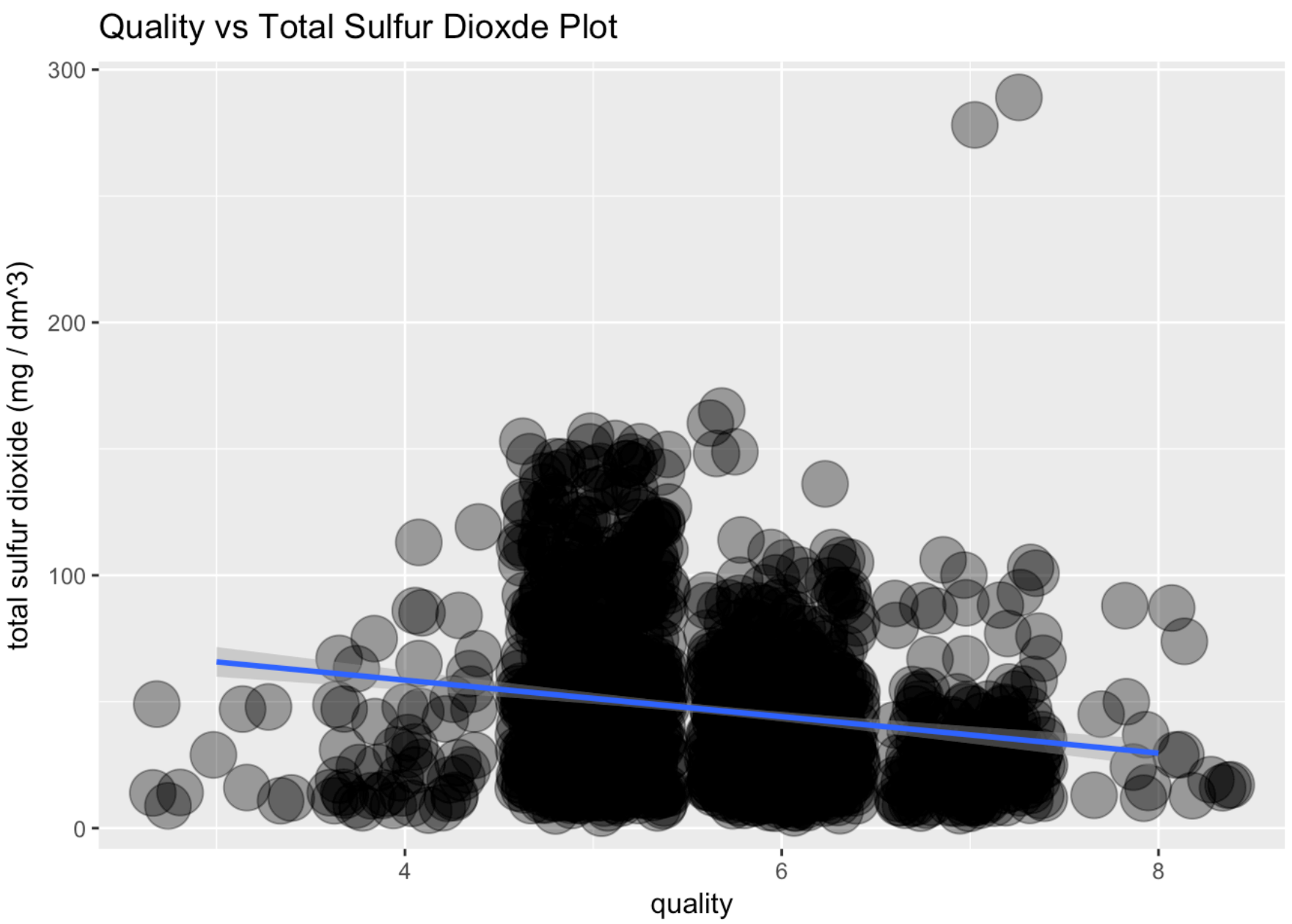


quality and pH relation is low so much, correlation is -0.05524511



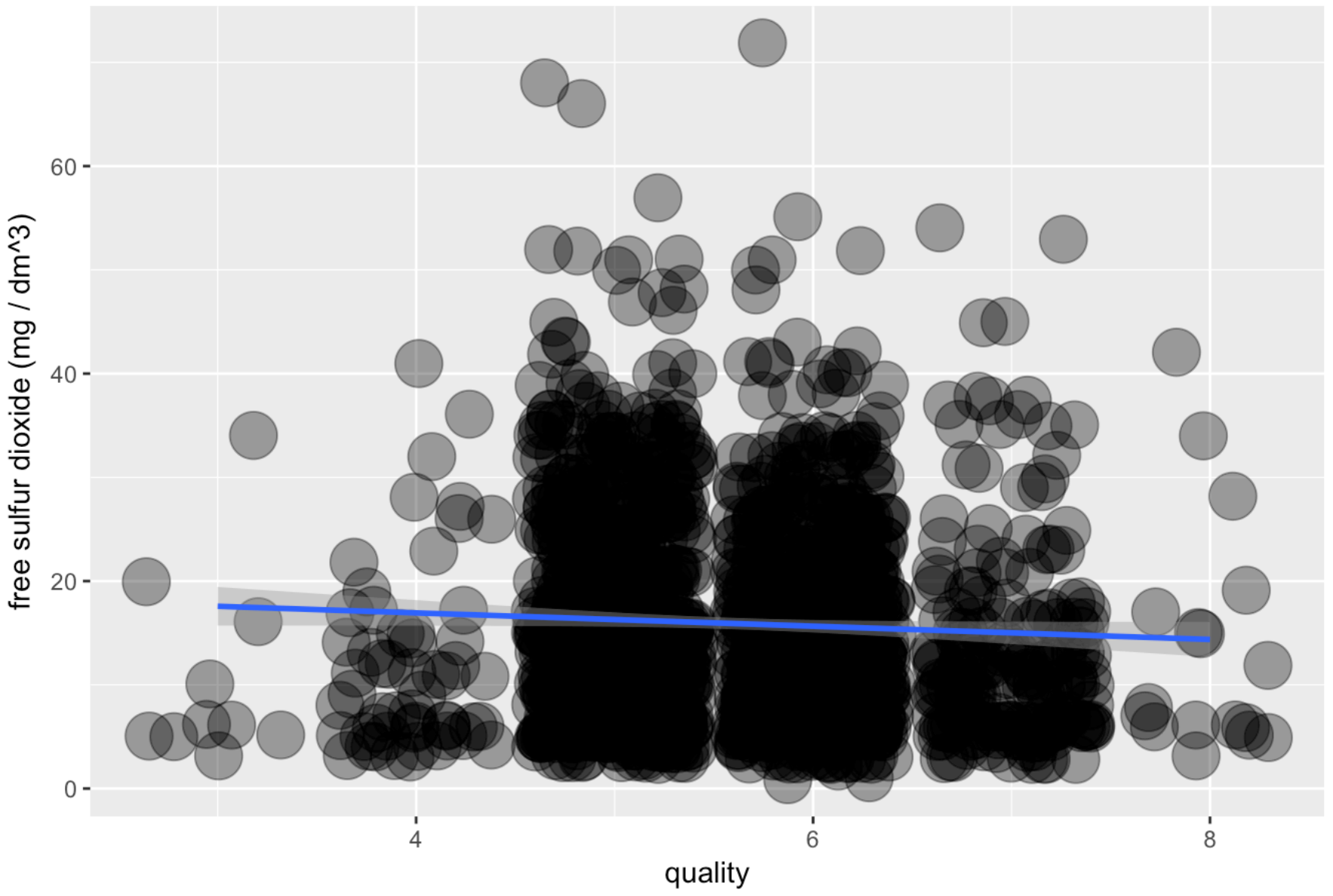

```
##
##  Pearson's product-moment correlation
##
## data:  quality and density
## t = -6.9056, df = 1357, p-value = 7.658e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2351231 -0.1323735
## sample estimates:
##       cor
## -0.1842517
```

density and quality relation is not much, correlation is -0.1842517



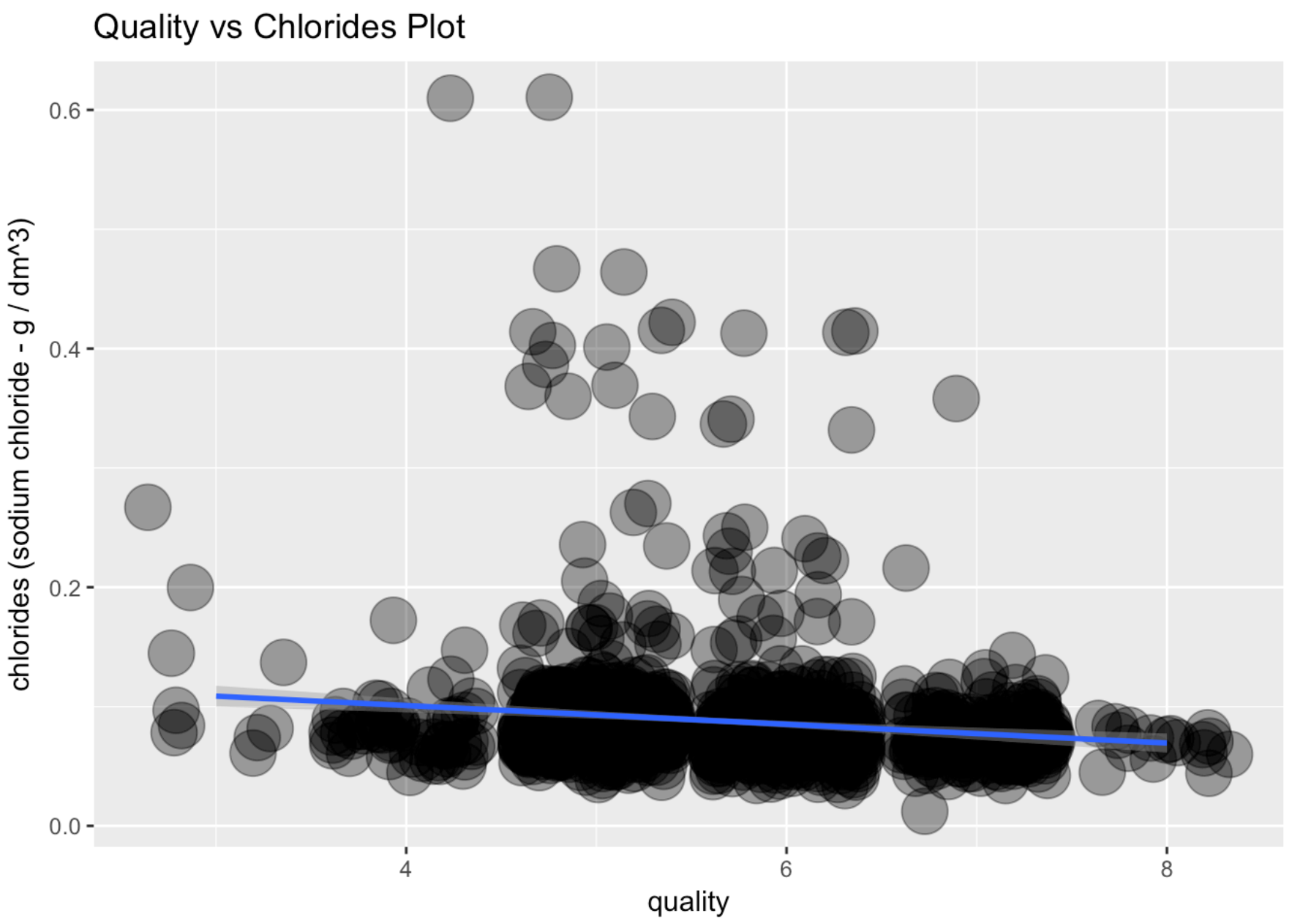

```
##
##  Pearson's product-moment correlation
##
## data:  quality and total.sulfur.dioxide
## t = -6.6579, df = 1357, p-value = 4.022e-11
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2288660 -0.1258707
## sample estimates:
##       cor
## -0.1778554
```

quality and total sulfur dioxide relation is not much, correlation is -0.1778554



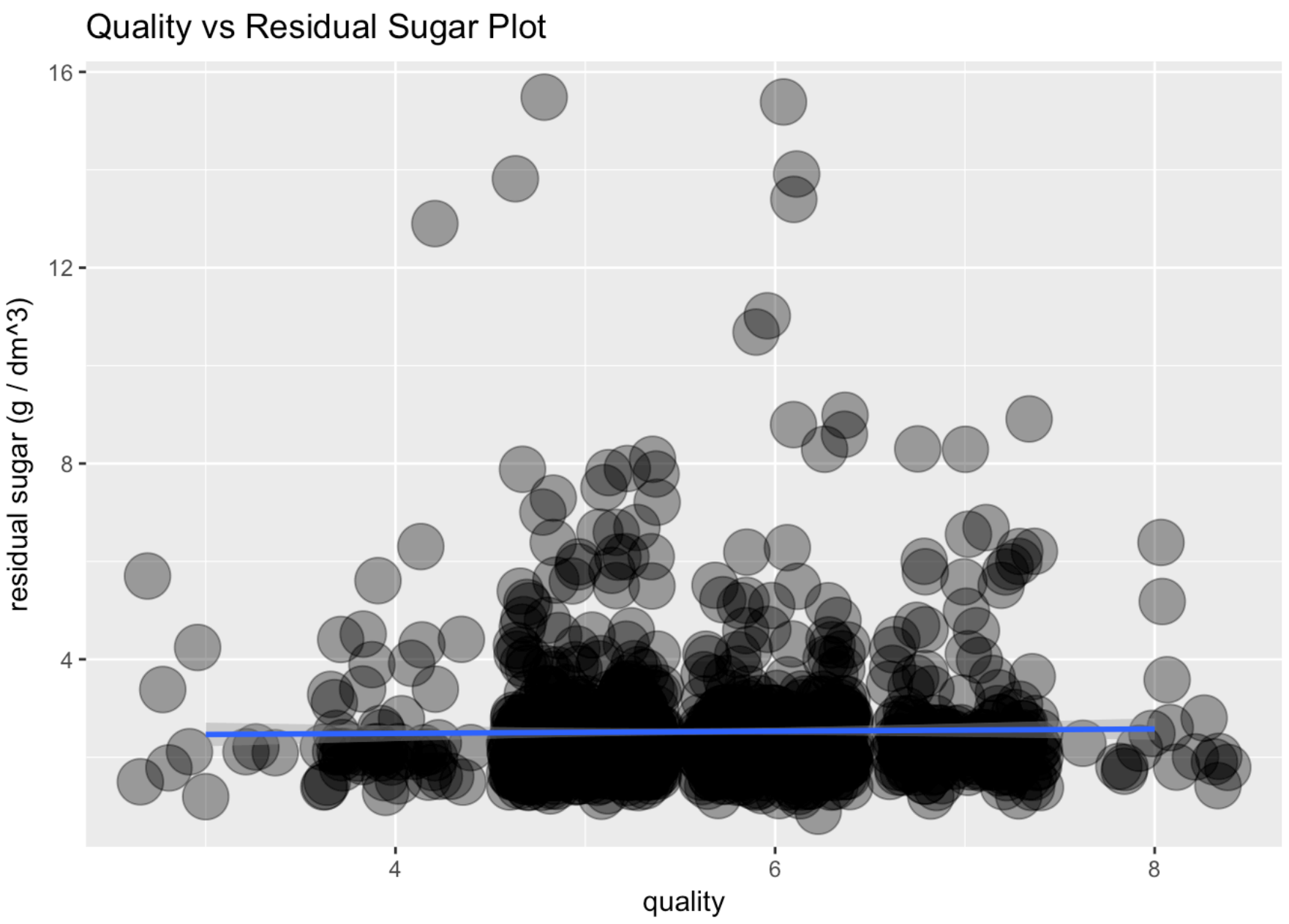

```
##
##  Pearson's product-moment correlation
##
## data:  quality and free.sulfur.dioxide
## t = -1.8613, df = 1357, p-value = 0.06292
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.103360524  0.002719642
## sample estimates:
##         cor
## -0.05046277
```

quality and free sulfur dioxide relation is not much, correctional is -0.05046277



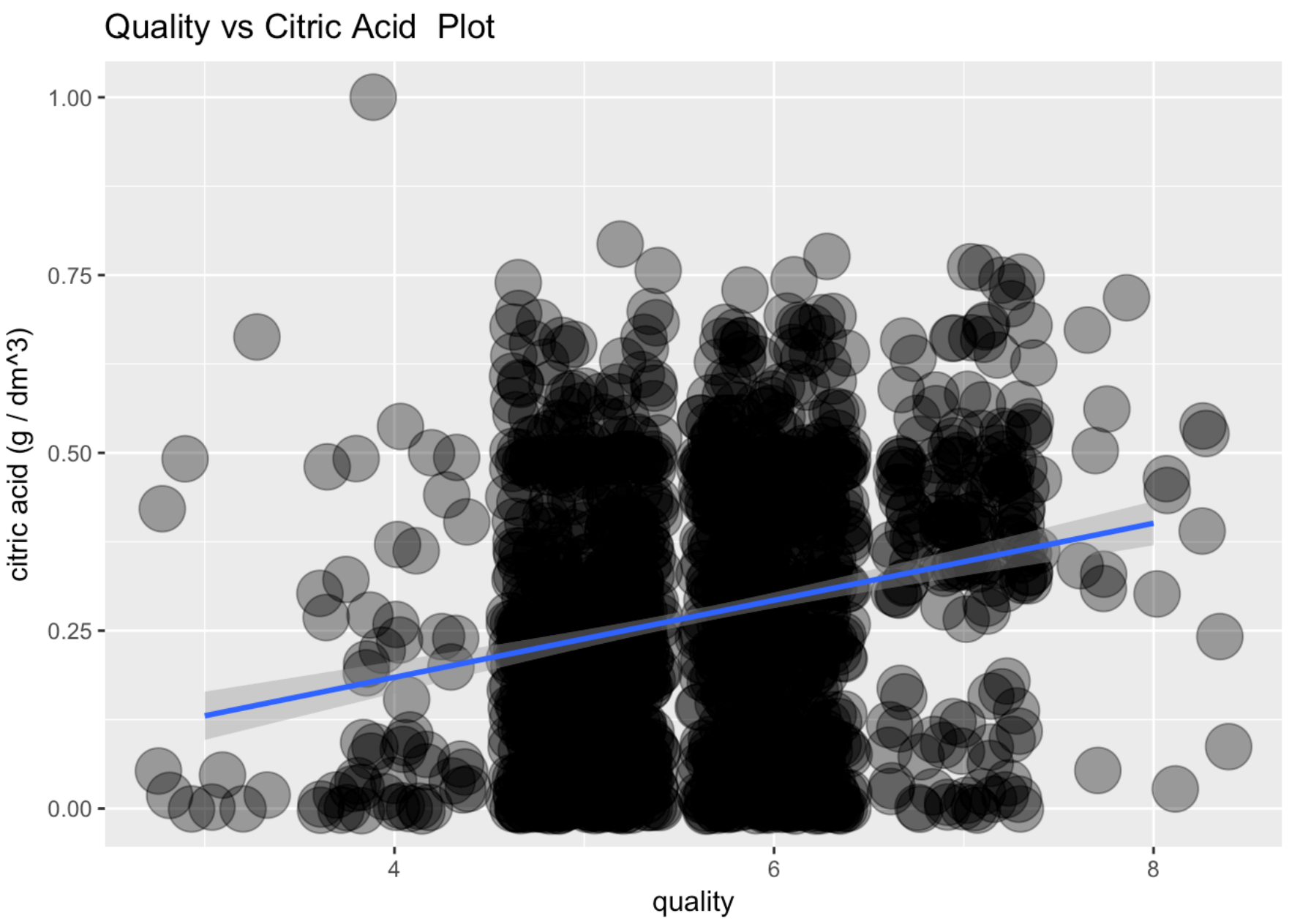

```
##
##  Pearson's product-moment correlation
##
## data:  quality and chlorides
## t = -4.8672, df = 1357, p-value = 1.264e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1828896 -0.0783591
## sample estimates:
##       cor
## -0.1309884
```

chlorides and quality relation is not much, correlation is -0.1309884



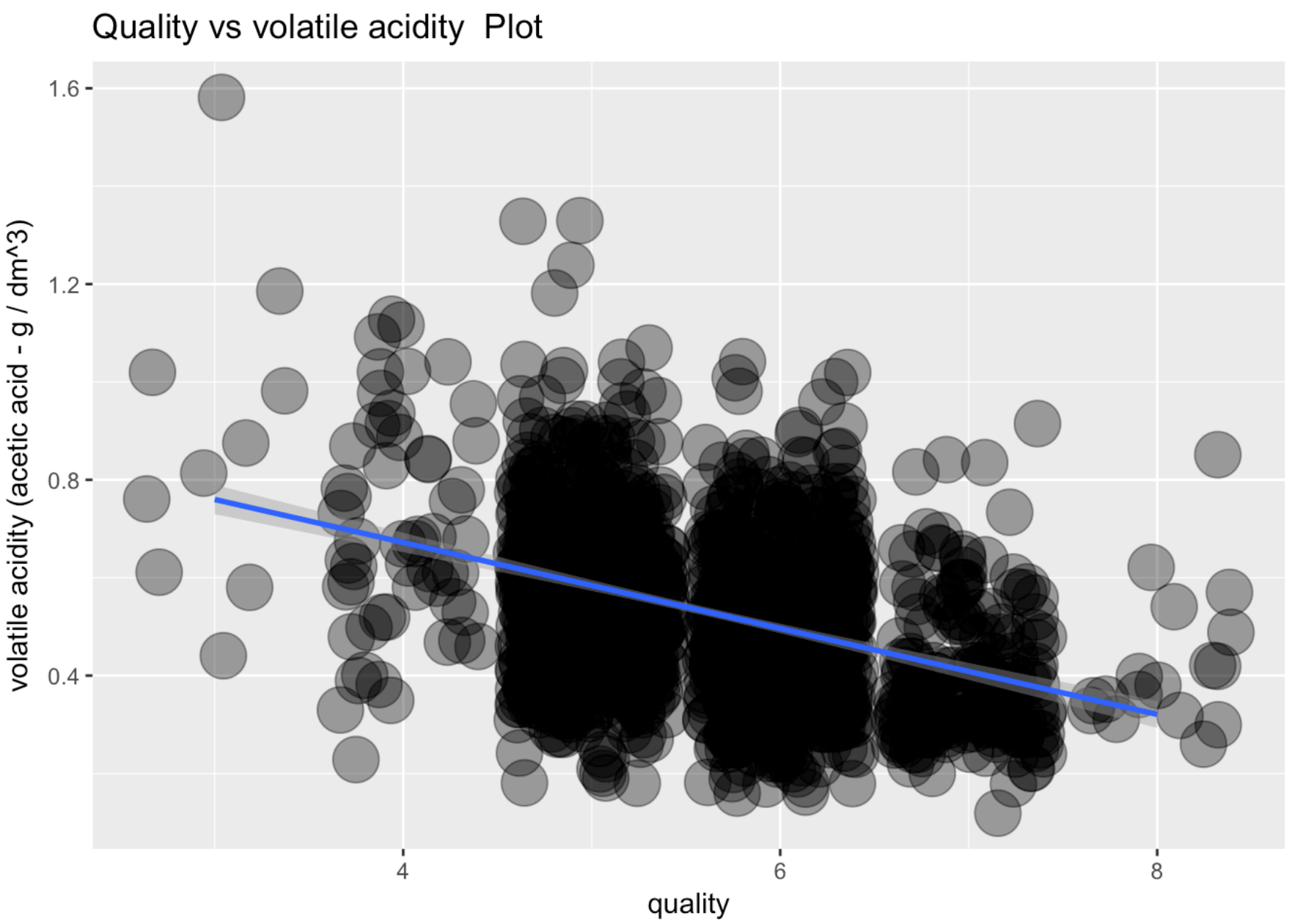

```
##
##  Pearson's product-moment correlation
##
## data:  quality and residual.sugar
## t = 0.50253, df = 1357, p-value = 0.6154
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.03956334  0.06676715
## sample estimates:
##       cor
## 0.01364047
```

quality and residual sugar relation is not much, corelation is 0.01364047



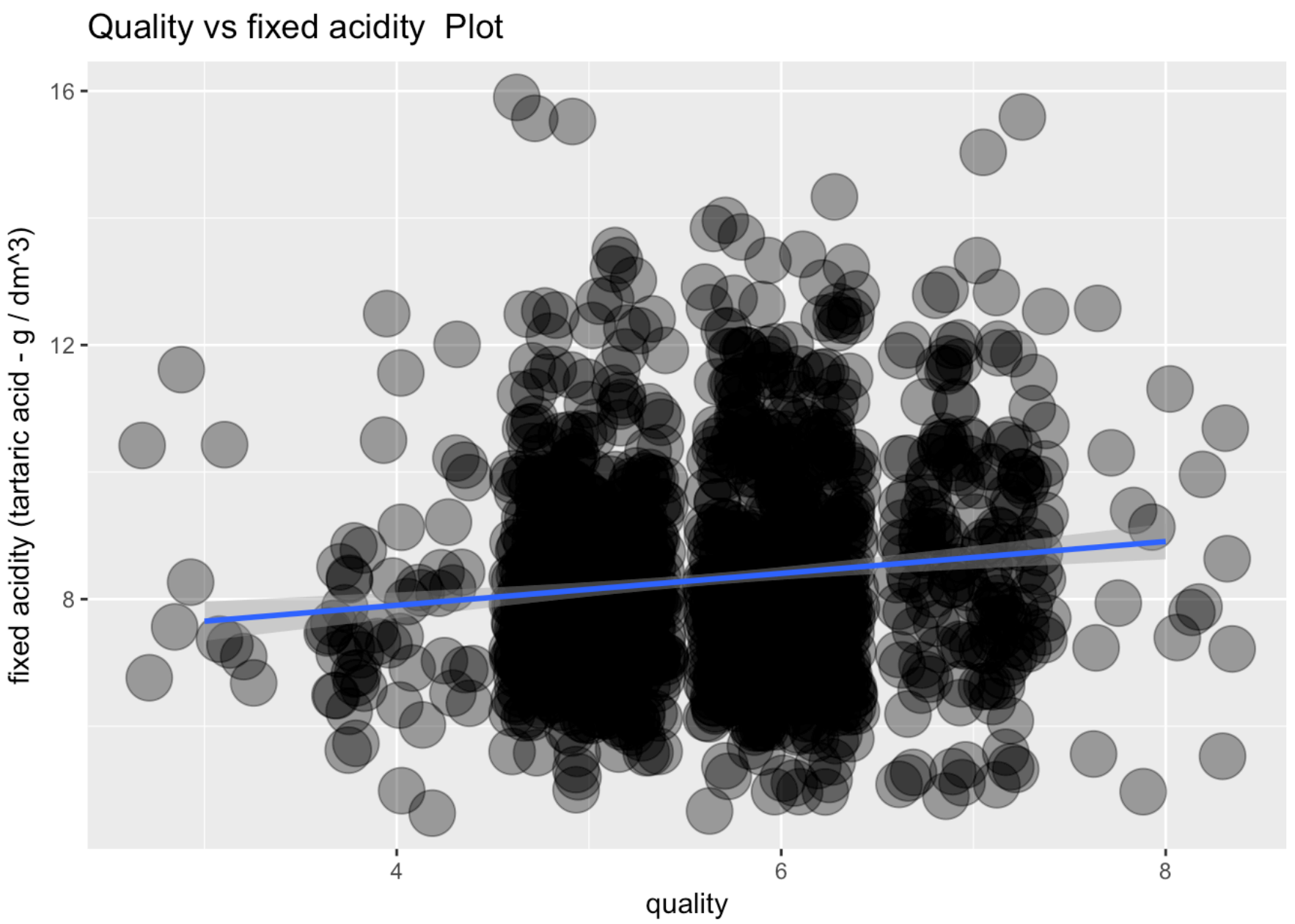

```
##
##  Pearson's product-moment correlation
##
## data:  quality and citric.acid
## t = 8.6284, df = 1357, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1770292 0.2778629
## sample estimates:
##       cor
## 0.2280575
```

quality and citric acid relation is not much , correlation is 0.2280575 . We know that citric acid effects on wines taste positively , there is an positive relation but not trend.



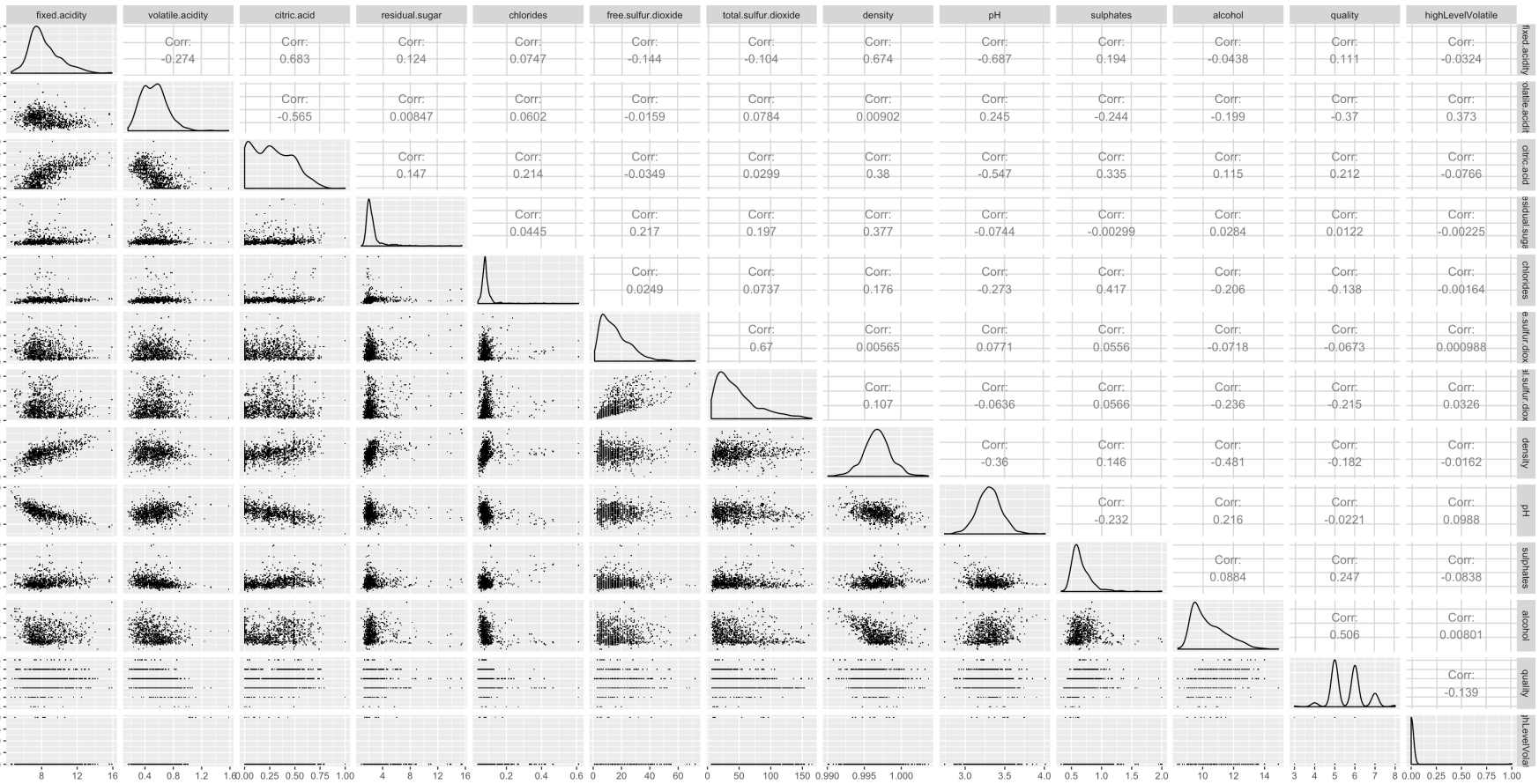

```
##
##  Pearson's product-moment correlation
##
## data:  quality and volatile.acidity
## t = -15.849, df = 1357, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4391596 -0.3493810
## sample estimates:
##       cor
## -0.3952137
```

volatile acidity and quality relation is more , correlation is -0.3952137 this is conjecturable because of volatile acidity’s effect on taste. There would be and negative relation and it is.




```
##
##  Pearson's product-moment correlation
##
## data:  quality and fixed.acidity
## t = 4.4159, df = 1357, p-value = 1.086e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.06626797 0.17111577
## sample estimates:
##       cor
## 0.1190237
```

quality and fixed acidity relation is less, correlation is 0.1190237



Bivariate Analysis

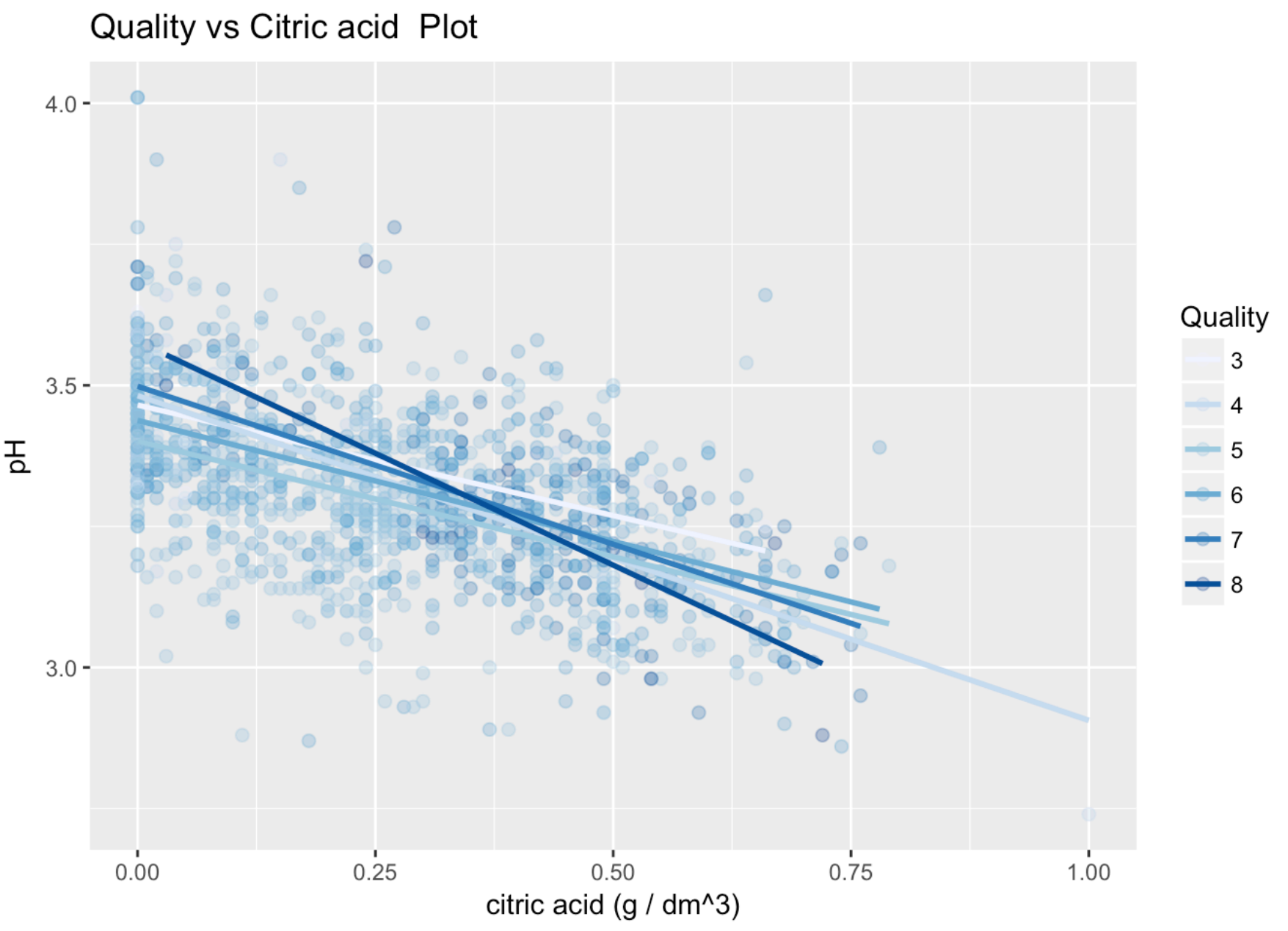
when I control the relationship of variables with each other, I saw that pH and fixed acidty is related with each other and pH and citric acid is related too. their correlation coefficent are -0.68 and -0.56

As we know density depends on solvent and solute, correlation of density vs alcohol is -0.504 and correlation of density and sugar is 0.3245

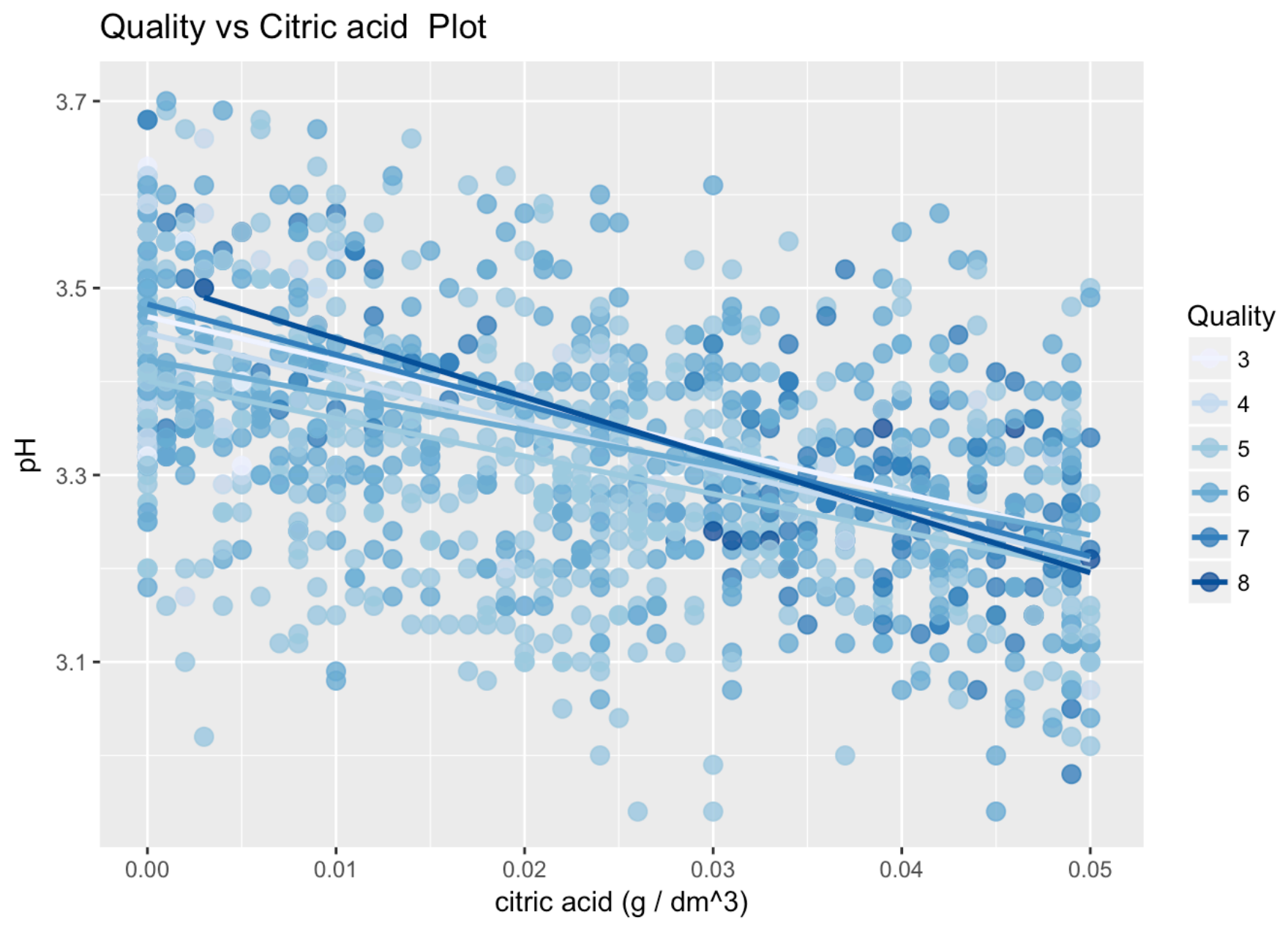
It is interesting that density and fixed acidity relation is strong, correlation coefficent is 0.678

strongest relationship is among pH and fixed acidty.

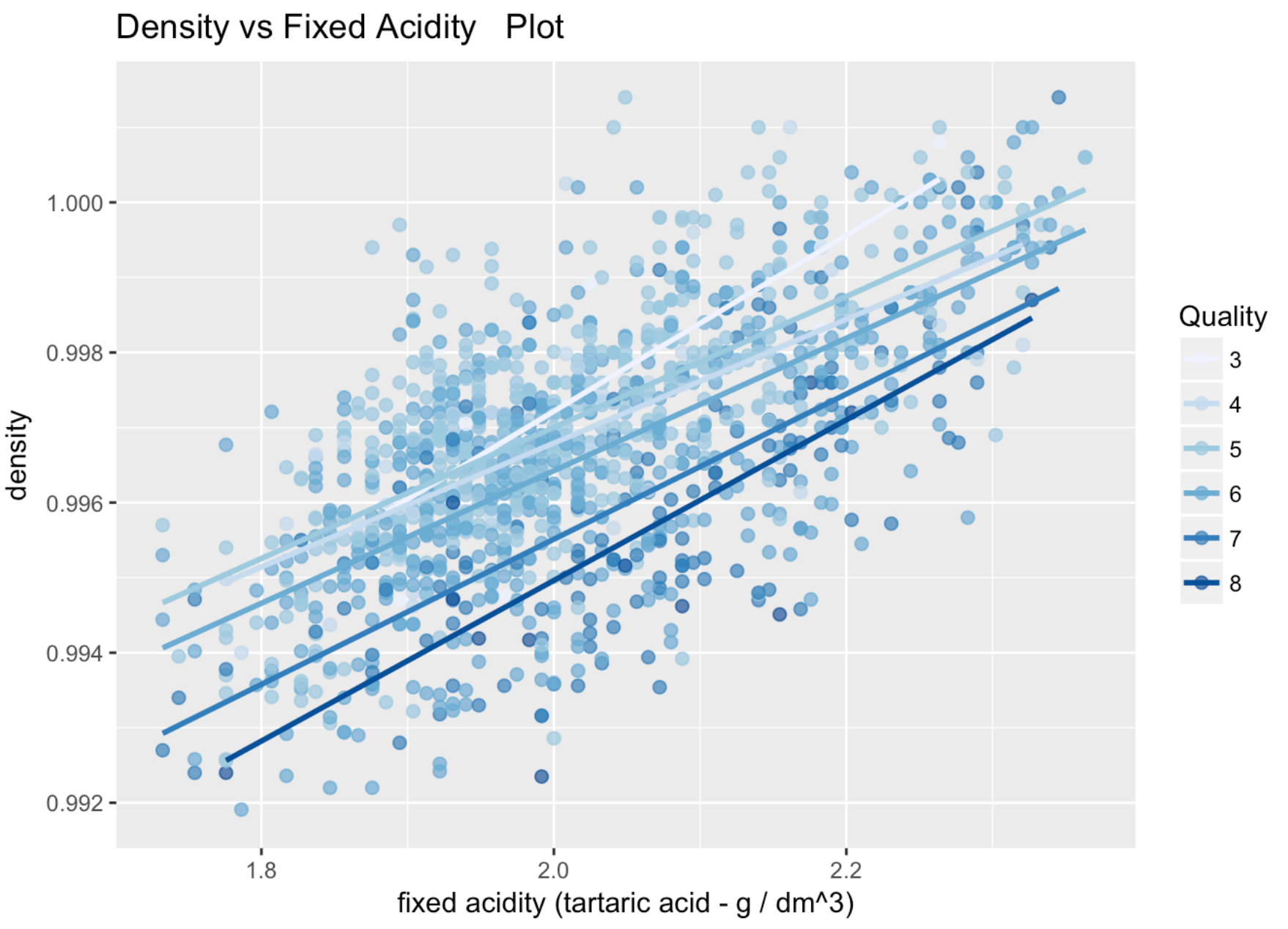
Multivariate Plots Section



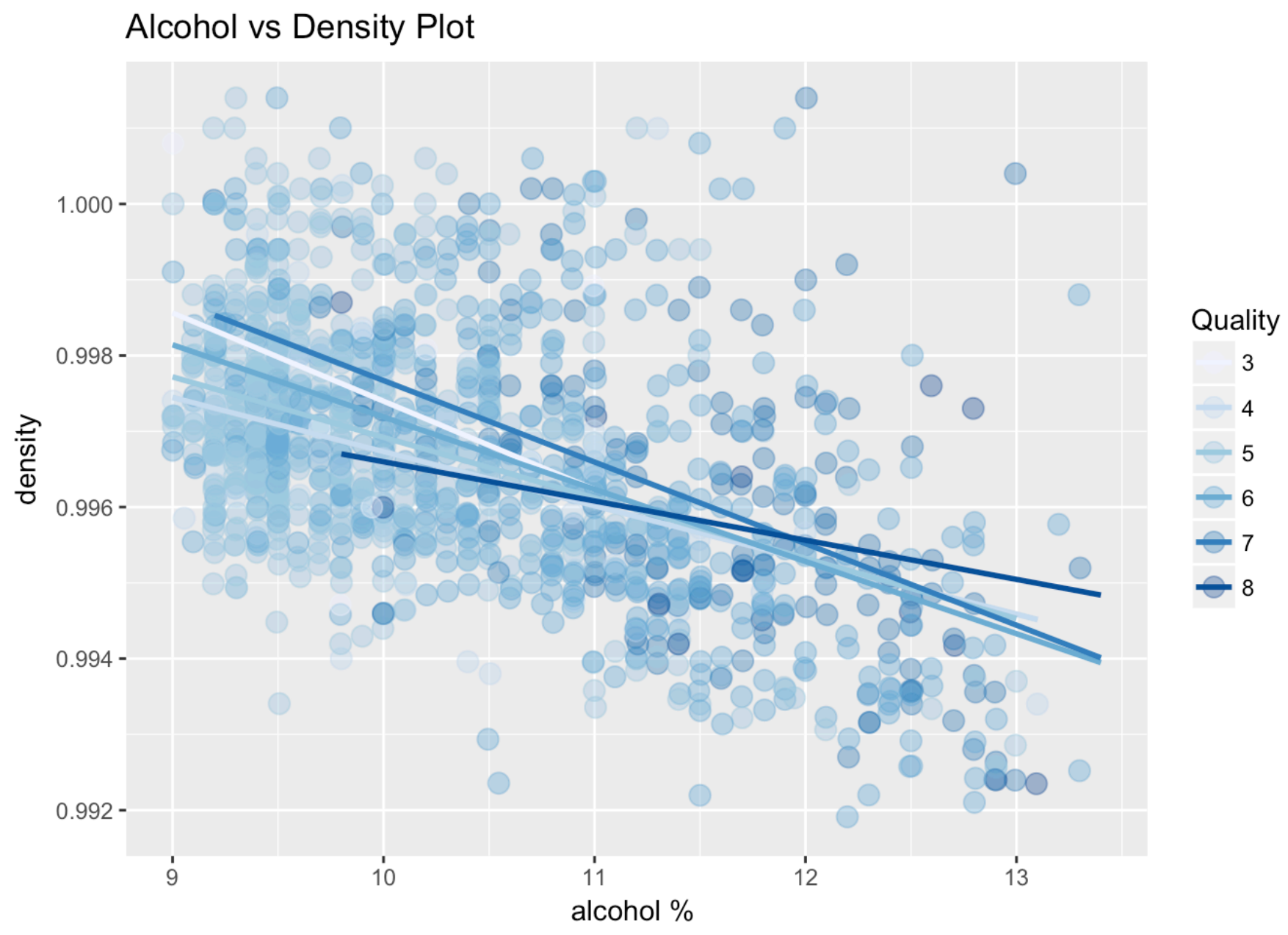
I want to see that quality distribution on this plot, so I used color for quality but It doesn’t show a clear relation;



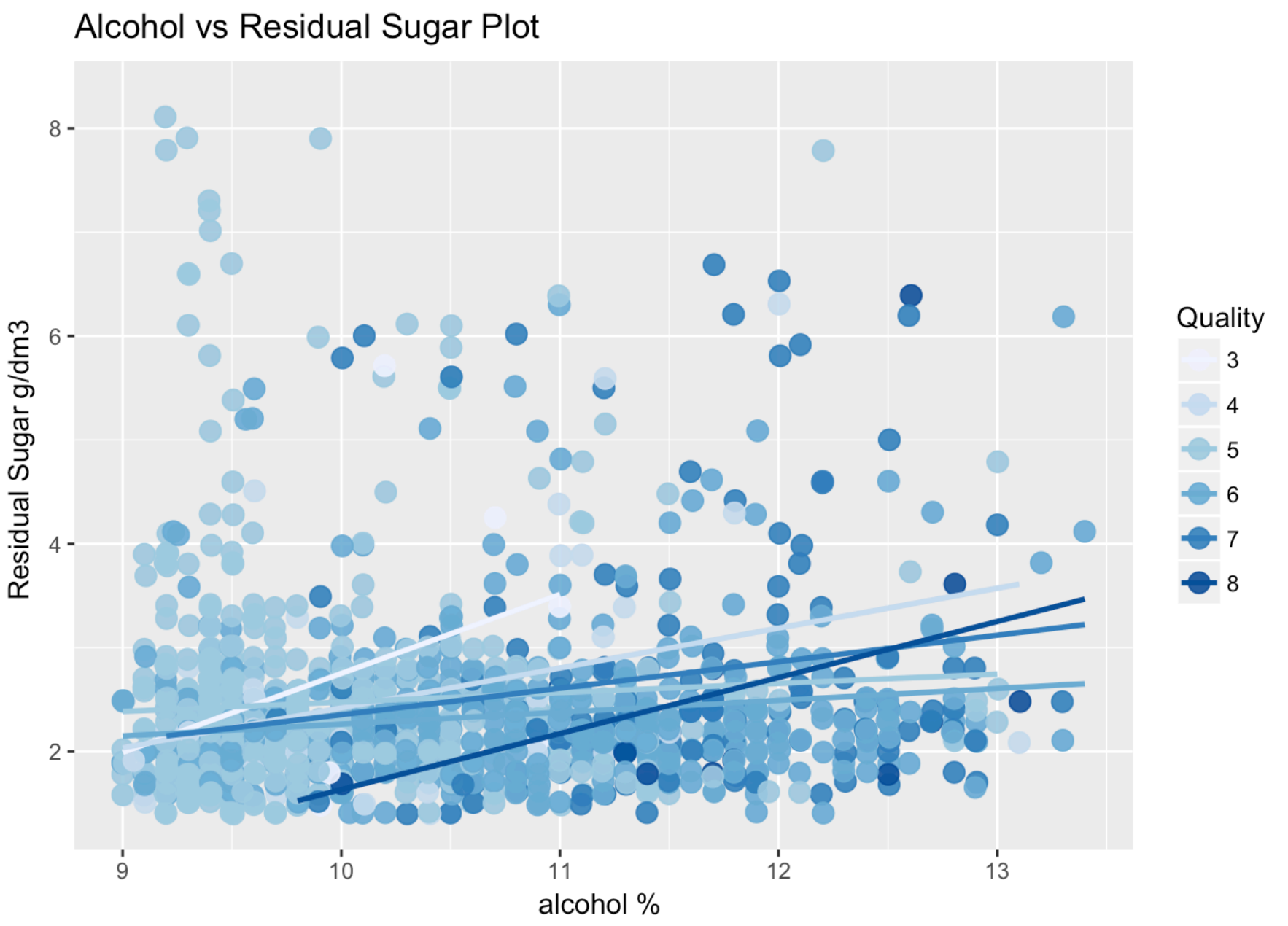
where citric acid level zero, it seems having low quality.



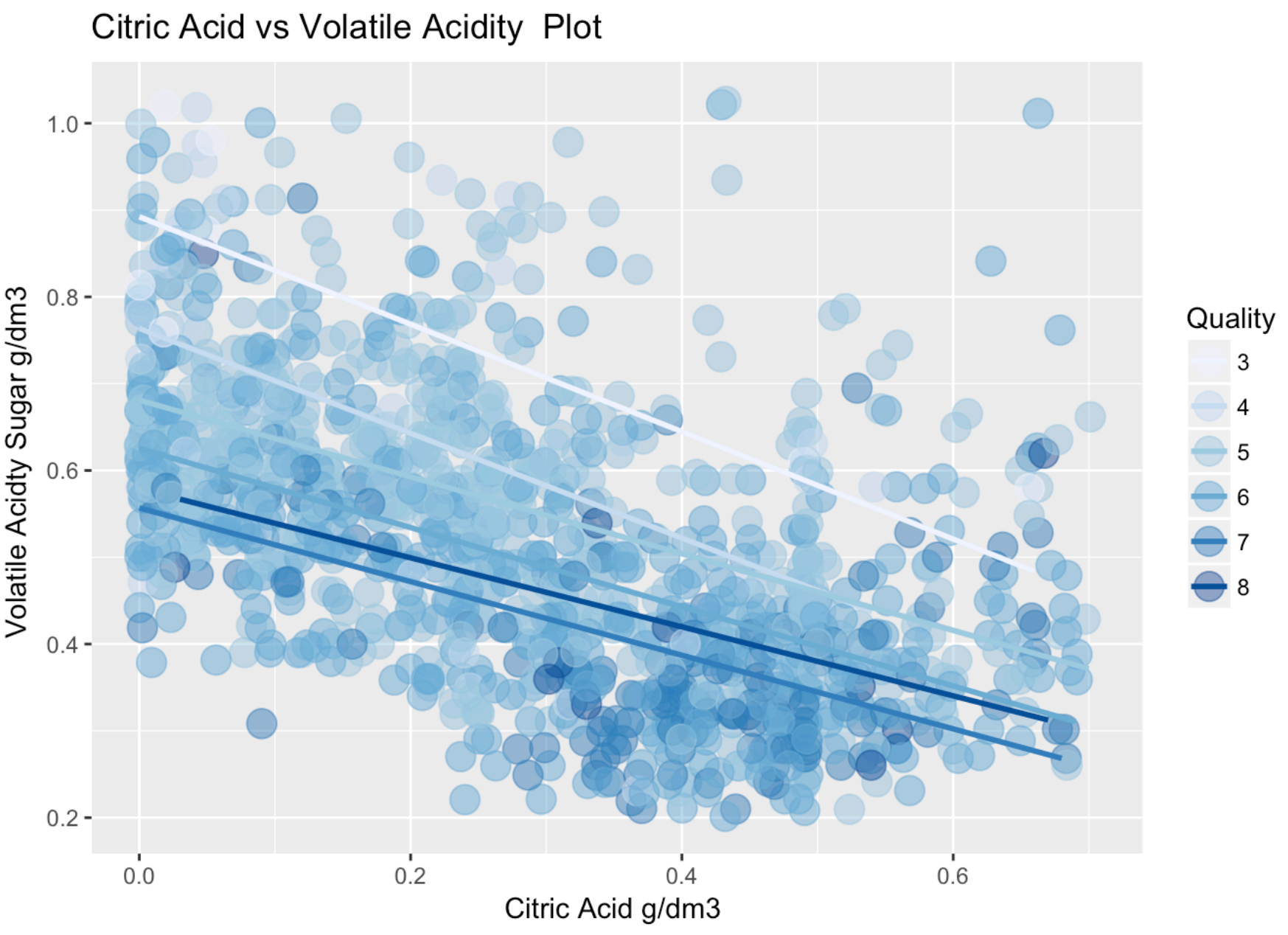
we can not see strong relation on this graph, we may say that low density and high fixed acidty wines have more quality.



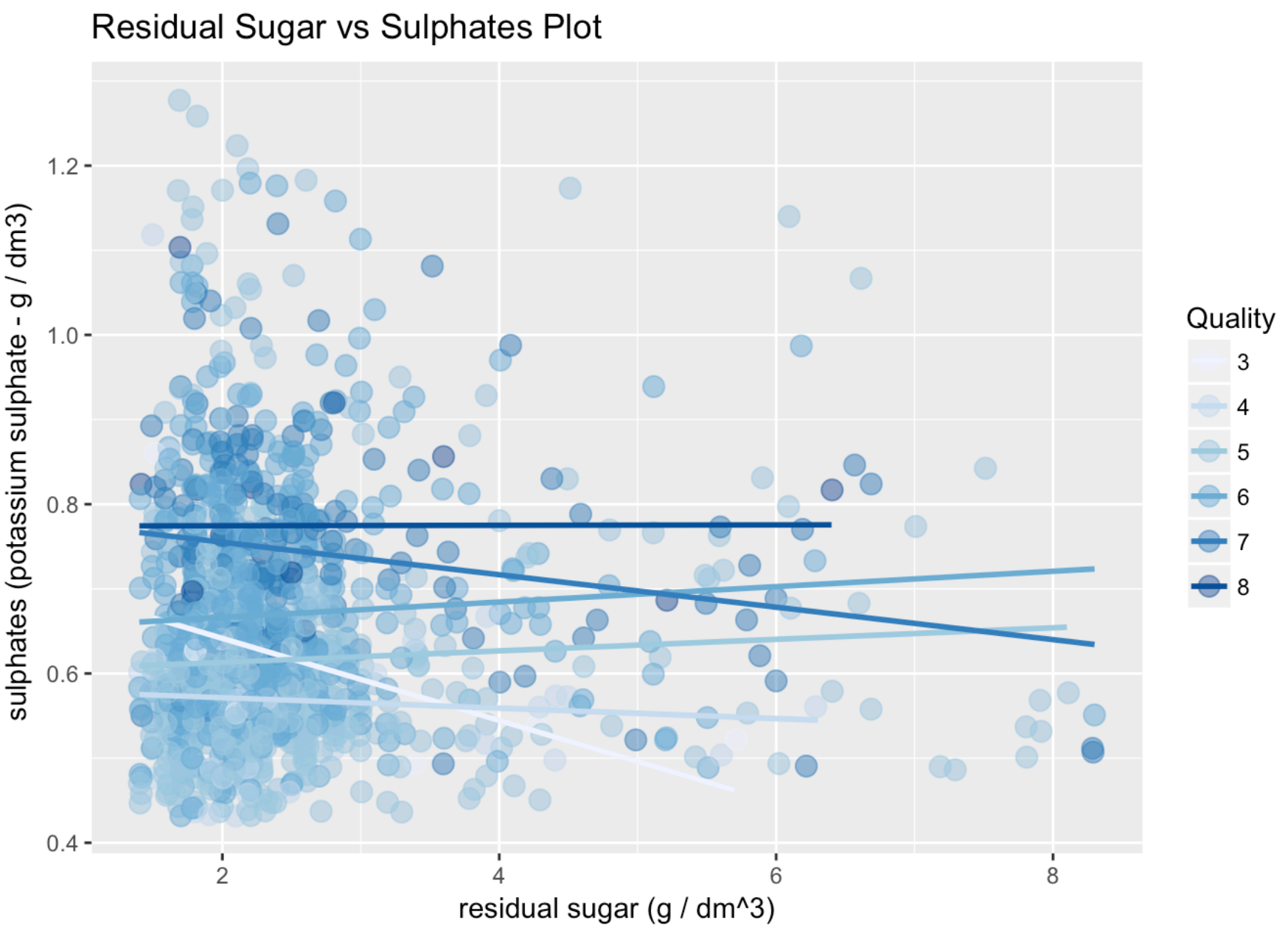
wines that has low density and high alcohol have higher quality



wines that has low sugar and high alcohol have higher quality



in this graph we can see th variables that effect on wines taste and as ew mention before, citric acid has positive efect but volatile acidity not.



Most of wines has between 0.45 - 1 sulphates and 1.5-2.5 sugar , It doesnt seem a strong relation with qualtiy.

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?

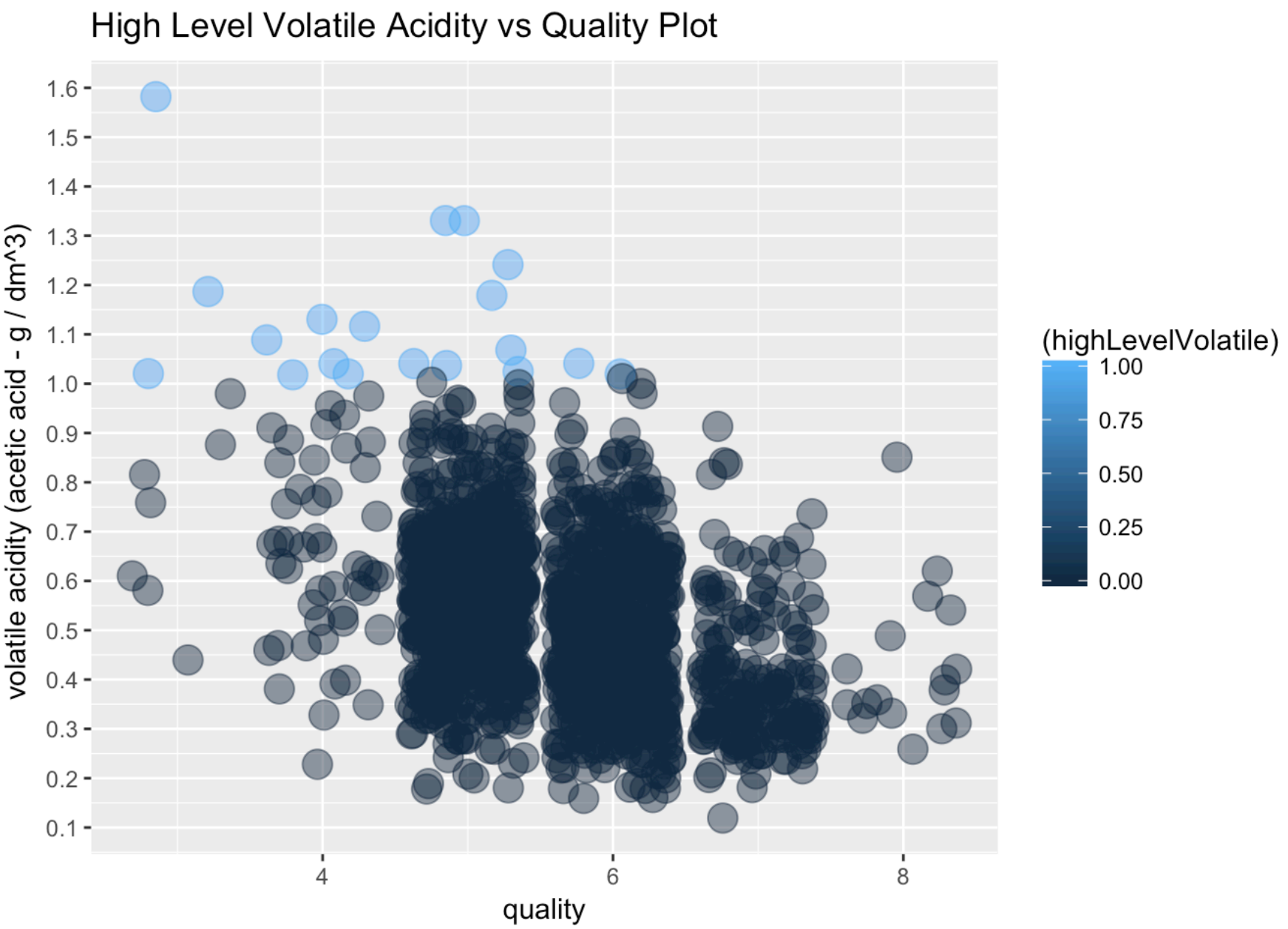
High quality wines has more alcohol and has less volatile acidity, and density

Citric acid and fixed acidty has strong relation, according to this relation we can say that wines has citric acid has less volacity acidity. pH and fixed acidiy relation is strong too, they have negative relation.

cirtic acid and pH plot which colored by quality, says that wines that contains about zero citric asid has less quality.

Final Plots and Summary

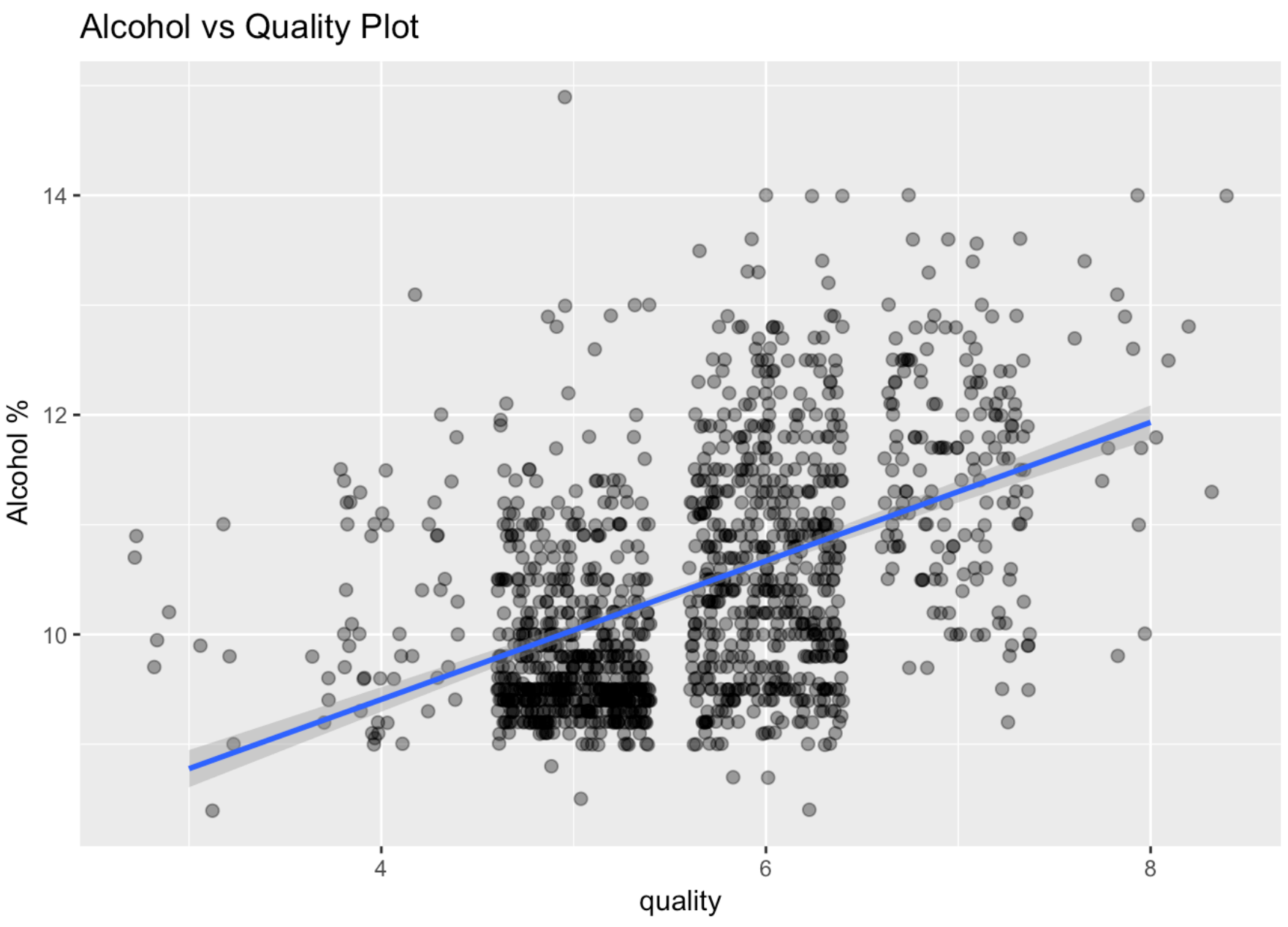
Plot One



Description One

colored point is outlier of volatile acidity and wines has high level volatilw acidity has bad taste , so we can say that taste effects on quality of wine.

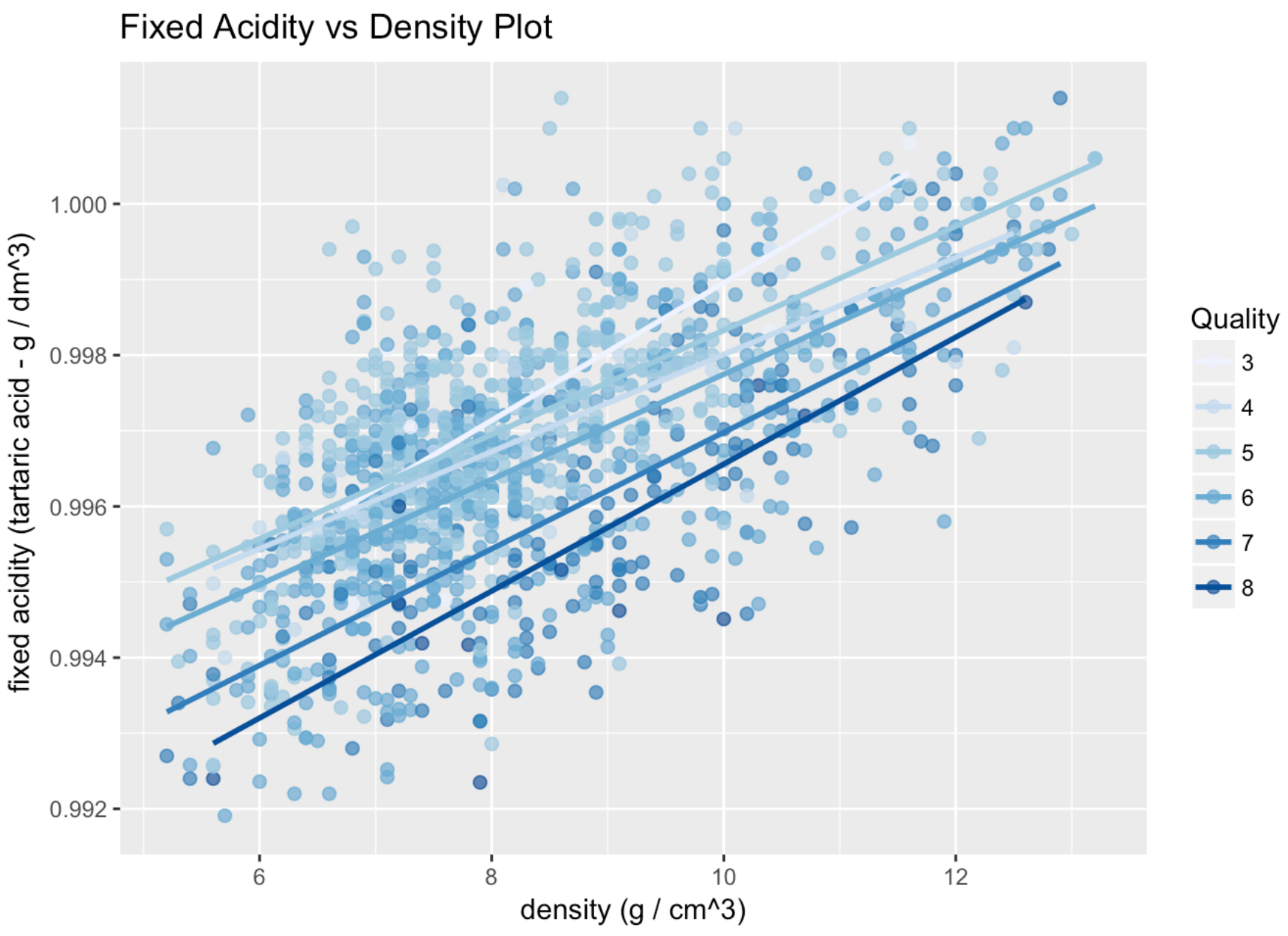
Plot Two



Description Two

wine quality and alcohol relation is the strongest relation in other variables, we can see easily from this plot

Plot Three



Description Three

Unexpected relation is between density and fixed acidity, they have strong relation and lower density and high fixed acidty wines have high quality.

Reflection

Our dataset has 1599 rows. But after some controls I noticed that first column only shows row number and other columns has some dublicate values. and I cleared it from dublicates and we have 1359 unique rows as result. All columns has numeric values and there are no factor variables.So we didnt need to group any variables.

Frist of all I started analysis by creating histograms per input and output variable, It doesn't seem any anormal distrubition, It was hard to guess what variables has releation with each other beacuse of I'm not chemist or someone like uses chemistry. So I have to study all variables' realtions with each other. Beacuse of the fact that I don't have time so much, I select some variables which may has realtion with each other by guessing according to my basic knowladge. And I focused input variables effects to output variables. I have created some plots and try to see the relations, but there wasn't a strong relation between input and output variable directly. According to datastructure and data description from data source, I try to found the variables that effect on wine's taste, I thought that these variables should have an relation with quality. And really the varibles that effects on taste has more strong relation with qualtiy.

It is interesting that density and fixed acidty relation, I didnt consider a relation, as I understand acid's density is higher then pure wine's, and so they have and positive relation. There is not a strong relation of quality with other variables as much as I expect. quality and alcohol realtion seems good but there is not a trend. Wine testers consider that wine's taste for qualty so main variables should be the variables that has effect on wines taste, so citric acid, alcohol and volatile acidity is our most related variables.

With this dataset we can see some relations but there are many variable that we can not get from data suppliers and I think that they effects on qualtiy like what type of grape is used for wine and how long time it take to fermantation. For future work if we have some variables that effects on taste , we can create an more reliable model with them. So far we know that alcohol is the most effective variable on wine's quality and volatile acidity effect on its taste. Some varibles that makes wine healther like sulphates has a relation with quality but I think testers dont realize about that they can only consider wine's taste.

Referances:

Data descriptions from udacity : https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityInfo.txt (https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityInfo.txt)