

# Udacity DAND Project: ExploreAndSummarizeData

Jubin Soni

August 27, 2017

In this project I have performed Exploratory Data Analysis on the White wine dataset. In this dataset the wine quality is of particular interest to me. I have performed Univariate, Bivariate and Multivariate analysis and explored various factors that affect and correlate with the wine quality.

GitHub Link: EDA: Univariate, Bivariate and Multivariate Analysis on White Wine Dataset

([https://github.com/jubins/DAND-Nanodegree/blob/master/dandp5-](https://github.com/jubins/DAND-Nanodegree/blob/master/dandp5-exploreandsummarizedatainR/Udacity%20DAND%20Project_%20ExploreAndSummarizeData.pdf)

[exploreandsummarizedatainR/Udacity%20DAND%20Project\\_%20ExploreAndSummarizeData.pdf](https://github.com/jubins/DAND-Nanodegree/blob/master/dandp5-exploreandsummarizedatainR/Udacity%20DAND%20Project_%20ExploreAndSummarizeData.pdf))

## Loading all the packages

## Loading the Dataset

## Initial Data Exploration

```
## [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"
```

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

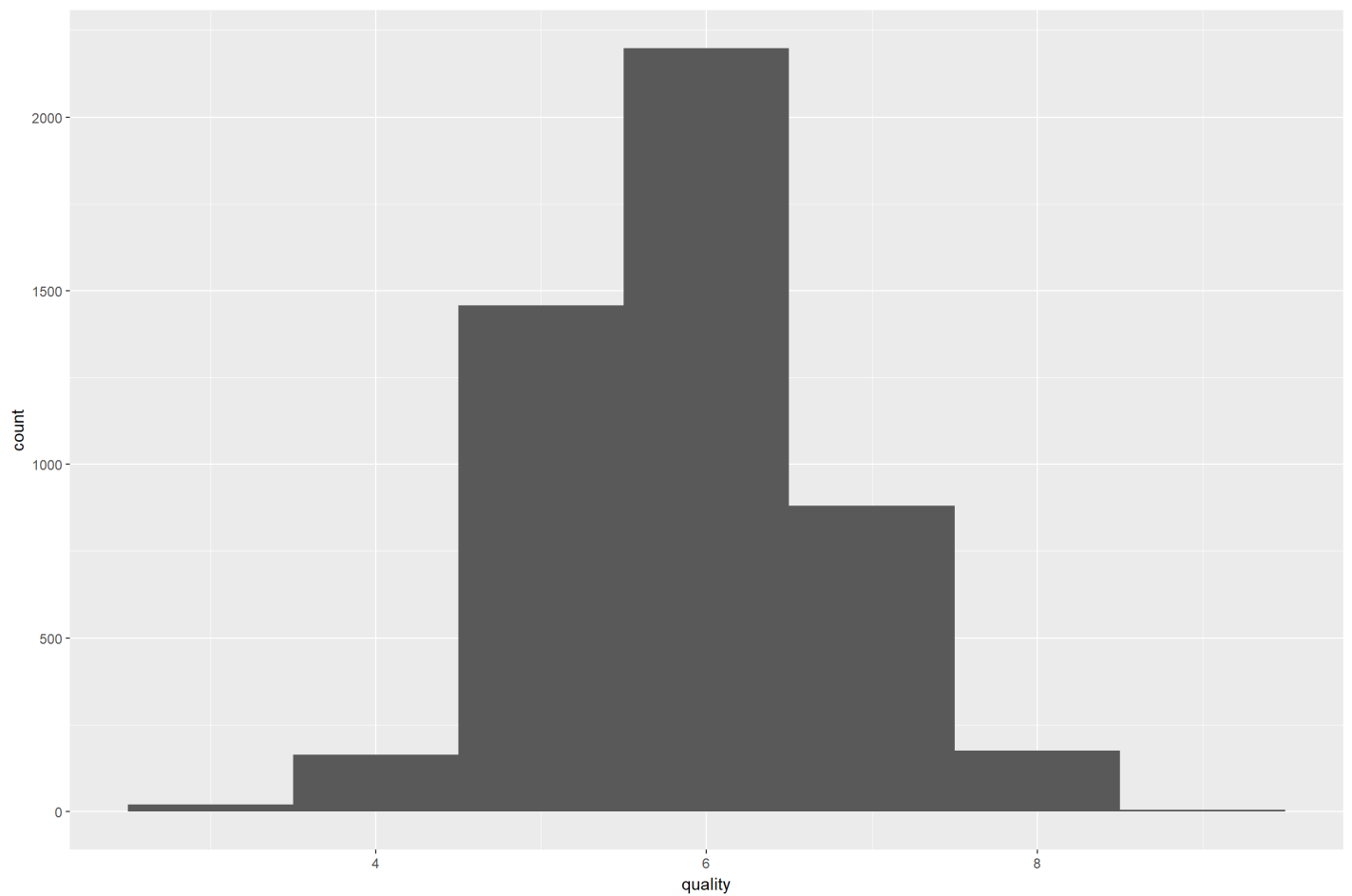
```
##      X      fixed.acidity  volatile.acidity  citric.acid
## Min.   : 1  Min.   : 3.800  Min.   :0.0800  Min.   :0.0000
## 1st Qu.:1225 1st Qu.: 6.300  1st Qu.:0.2100  1st Qu.:0.2700
## Median :2450 Median : 6.800  Median :0.2600  Median :0.3200
## Mean   :2450 Mean   : 6.855  Mean   :0.2782  Mean   :0.3342
## 3rd Qu.:3674 3rd Qu.: 7.300  3rd Qu.:0.3200  3rd Qu.:0.3900
## Max.   :4898 Max.   :14.200  Max.   :1.1000  Max.   :1.6600
## residual.sugar  chlorides      free.sulfur.dioxide
## Min.   : 0.600  Min.   :0.00900  Min.   : 2.00
## 1st Qu.: 1.700  1st Qu.:0.03600  1st Qu.: 23.00
## Median : 5.200  Median :0.04300  Median : 34.00
## Mean   : 6.391  Mean   :0.04577  Mean   : 35.31
## 3rd Qu.: 9.900  3rd Qu.:0.05000  3rd Qu.: 46.00
## Max.   :65.800  Max.   :0.34600  Max.   :289.00
## total.sulfur.dioxide  density      pH      sulphates
## Min.   : 9.0      Min.   :0.9871  Min.   :2.720  Min.   :0.2200
## 1st Qu.:108.0     1st Qu.:0.9917  1st Qu.:3.090  1st Qu.:0.4100
## Median :134.0     Median :0.9937  Median :3.180  Median :0.4700
## Mean   :138.4     Mean   :0.9940  Mean   :3.188  Mean   :0.4898
## 3rd Qu.:167.0     3rd Qu.:0.9961  3rd Qu.:3.280  3rd Qu.:0.5500
## Max.   :440.0     Max.   :1.0390  Max.   :3.820  Max.   :1.0800
## alcohol      quality
## Min.   : 8.00  Min.   :3.000
## 1st Qu.: 9.50  1st Qu.:5.000
## Median :10.40  Median :6.000
## Mean   :10.51  Mean   :5.878
## 3rd Qu.:11.40  3rd Qu.:6.000
## Max.   :14.20  Max.   :9.000
```

## Structure of Dataset

The dataset on white wines has 4898 observations of 13 variables. The main feature of interest is quality and alcohol content.

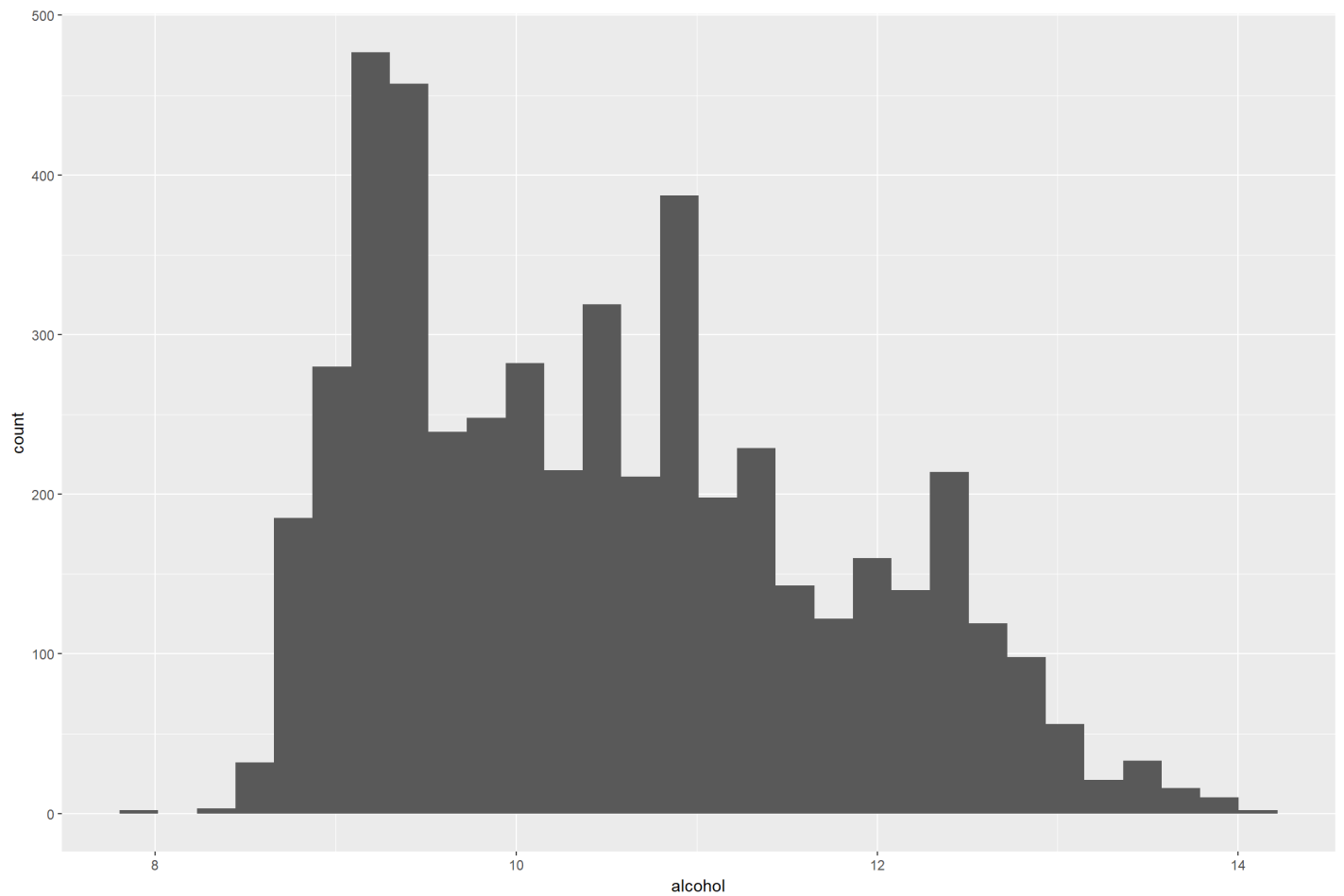
## Univariate Analysis

### Histogram of Quality of Wine



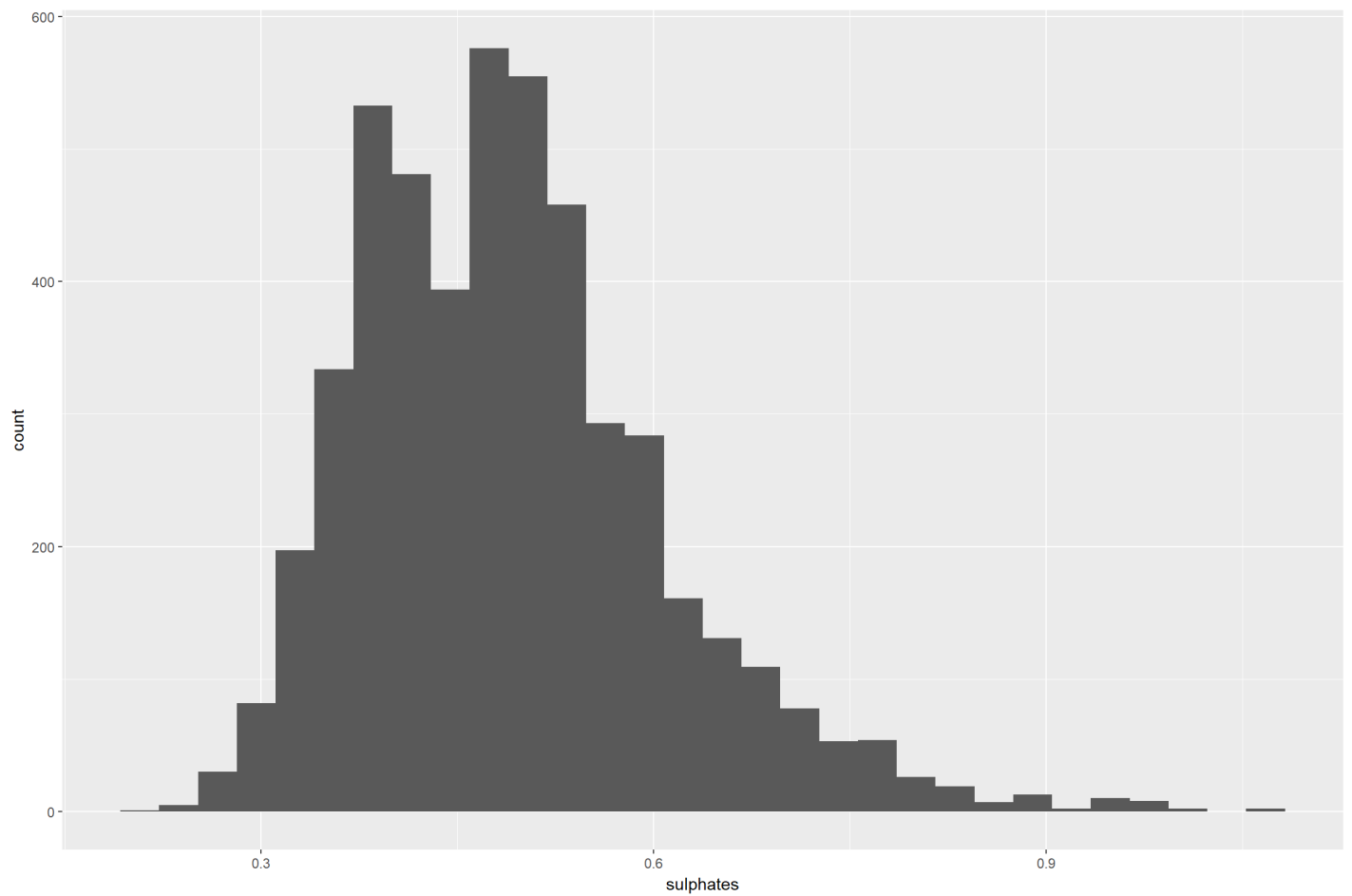
We see that the quality is close to a normal distribution, the graph is which is not fully symmetric because the quality 5 is higher than its corresponding value. However looking at the quality of 6 being highest we can say that the quality is above average.

## Histogram of Alcohol



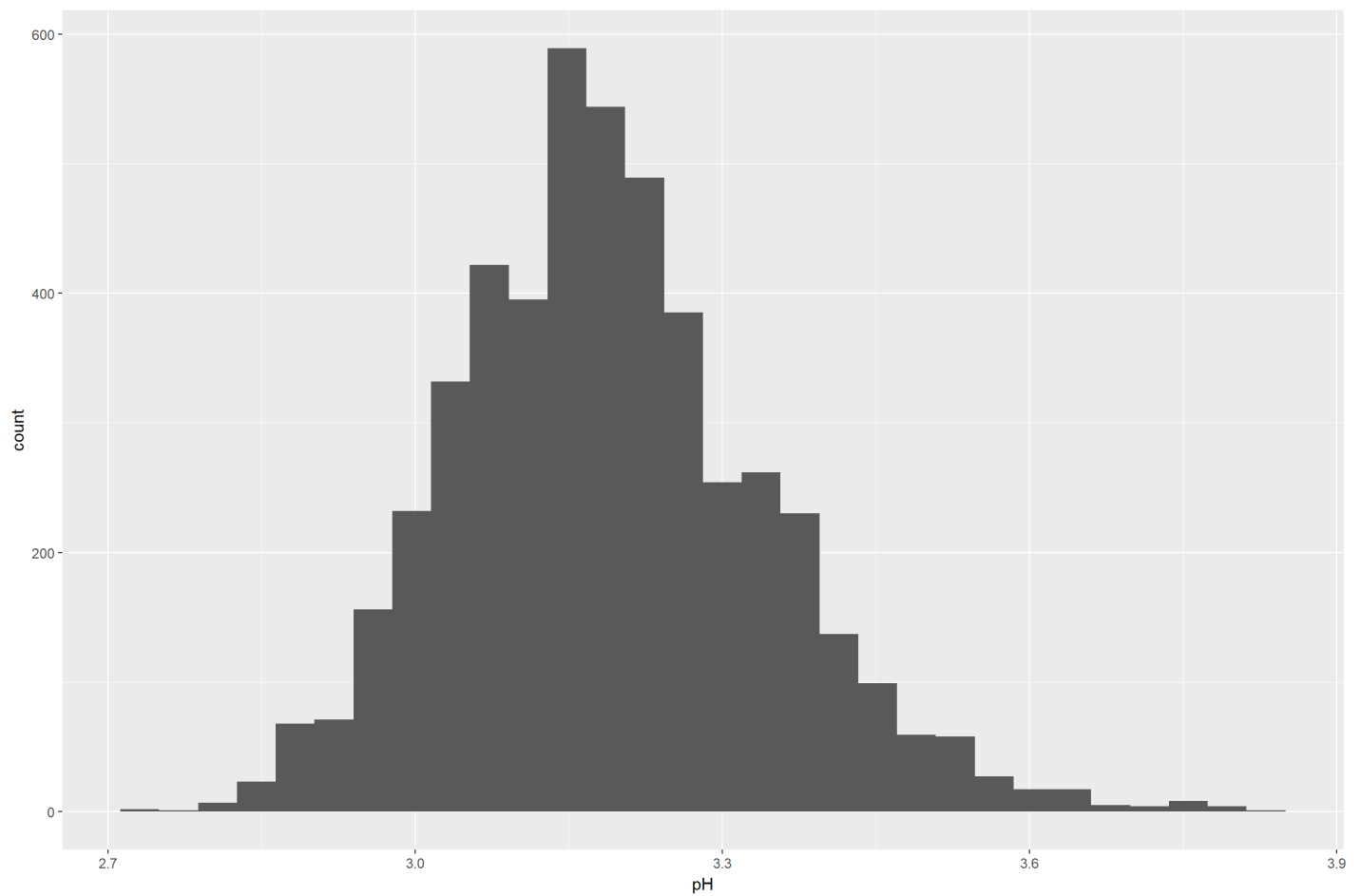
This histogram shows us that the majority of the wines in dataset have alcohol content between 9 to 12.

## Histogram of Sulphates



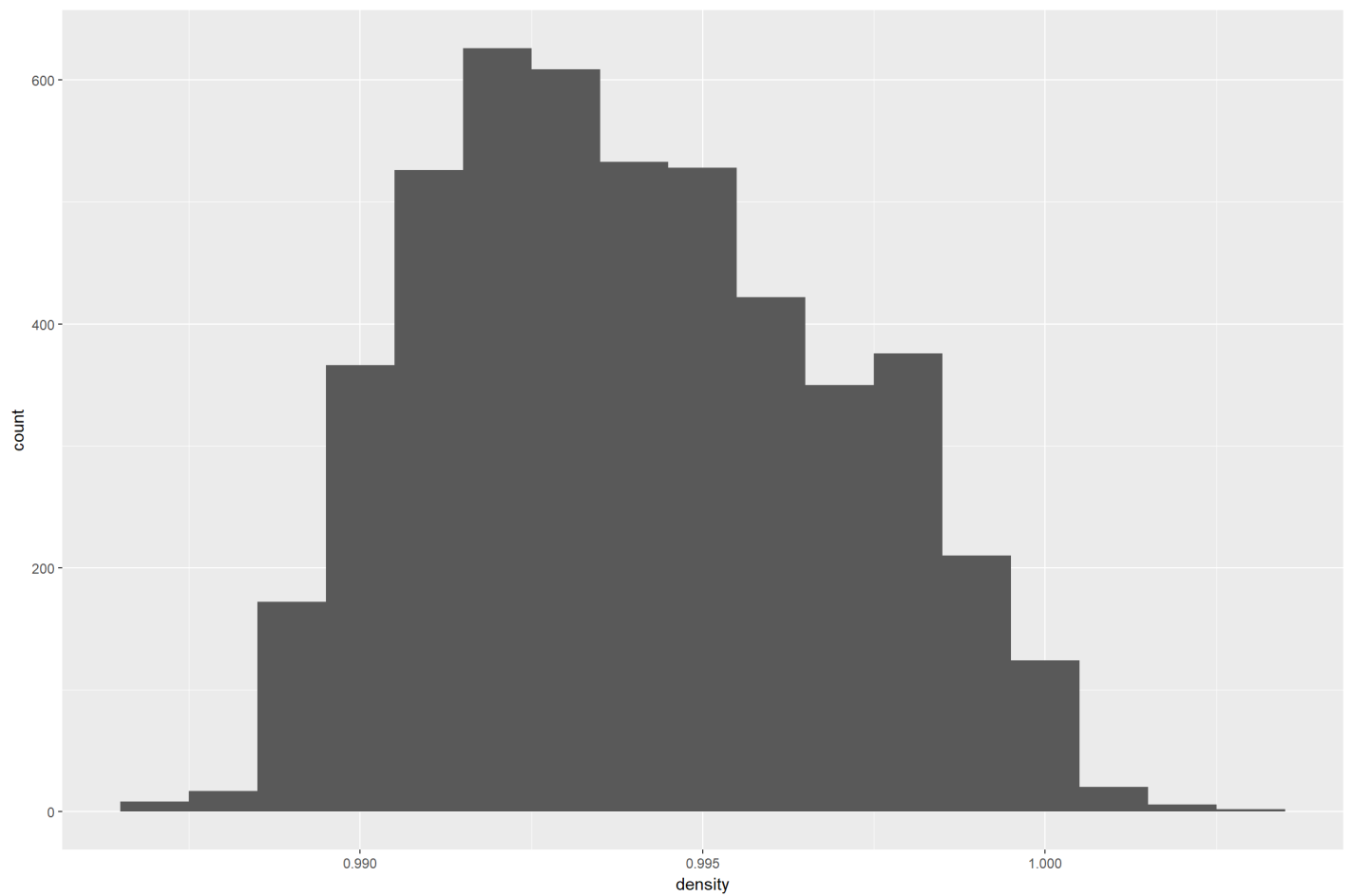
The histogram shows that majority of sulphates in wine are between 0.3 to 0.6, other are less than 200.

## Histogram of pH



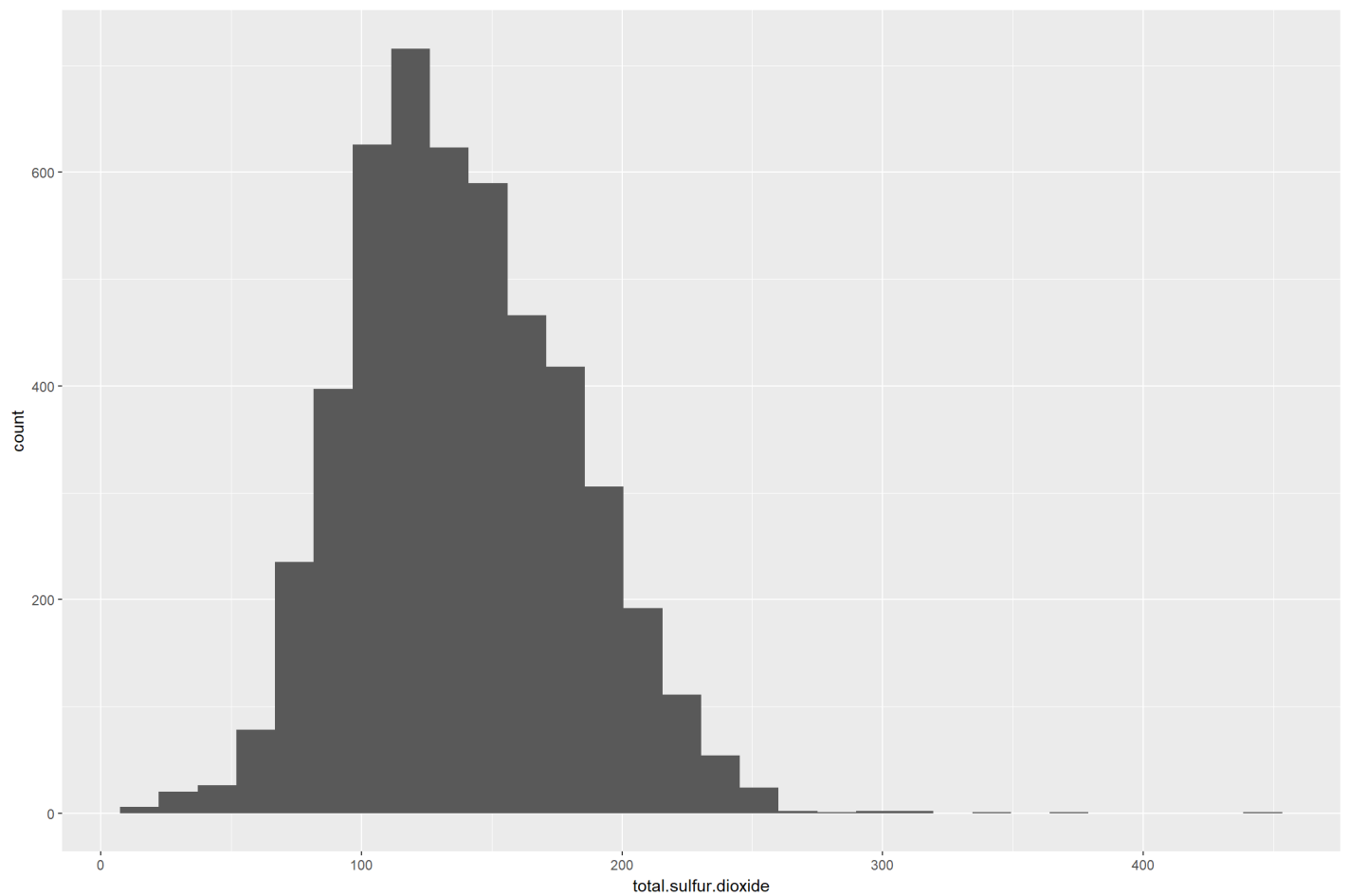
The pH is very narrow for White Wine between 2.7 to 3.8, so it seems like these values are within the acidity range and no alkaline (which is above 7)

## Histogram of Wine Density



The distribution below shows that density is within 0.9 to 1.0 range.

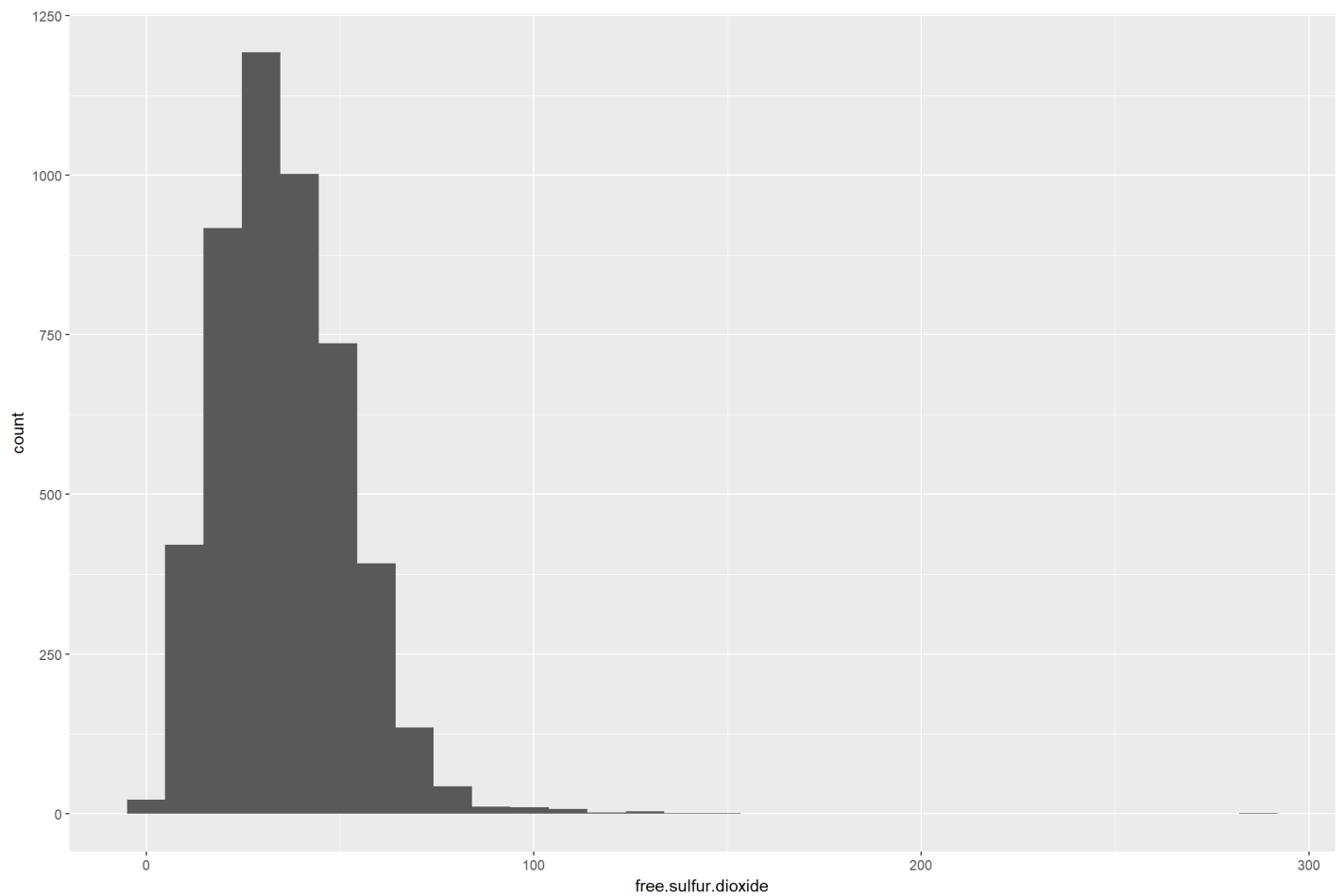
## Histogram of Total.Sulfur.Dioxide



The histogram of Sulfur Dioxide shows normal distribution in the White wine.

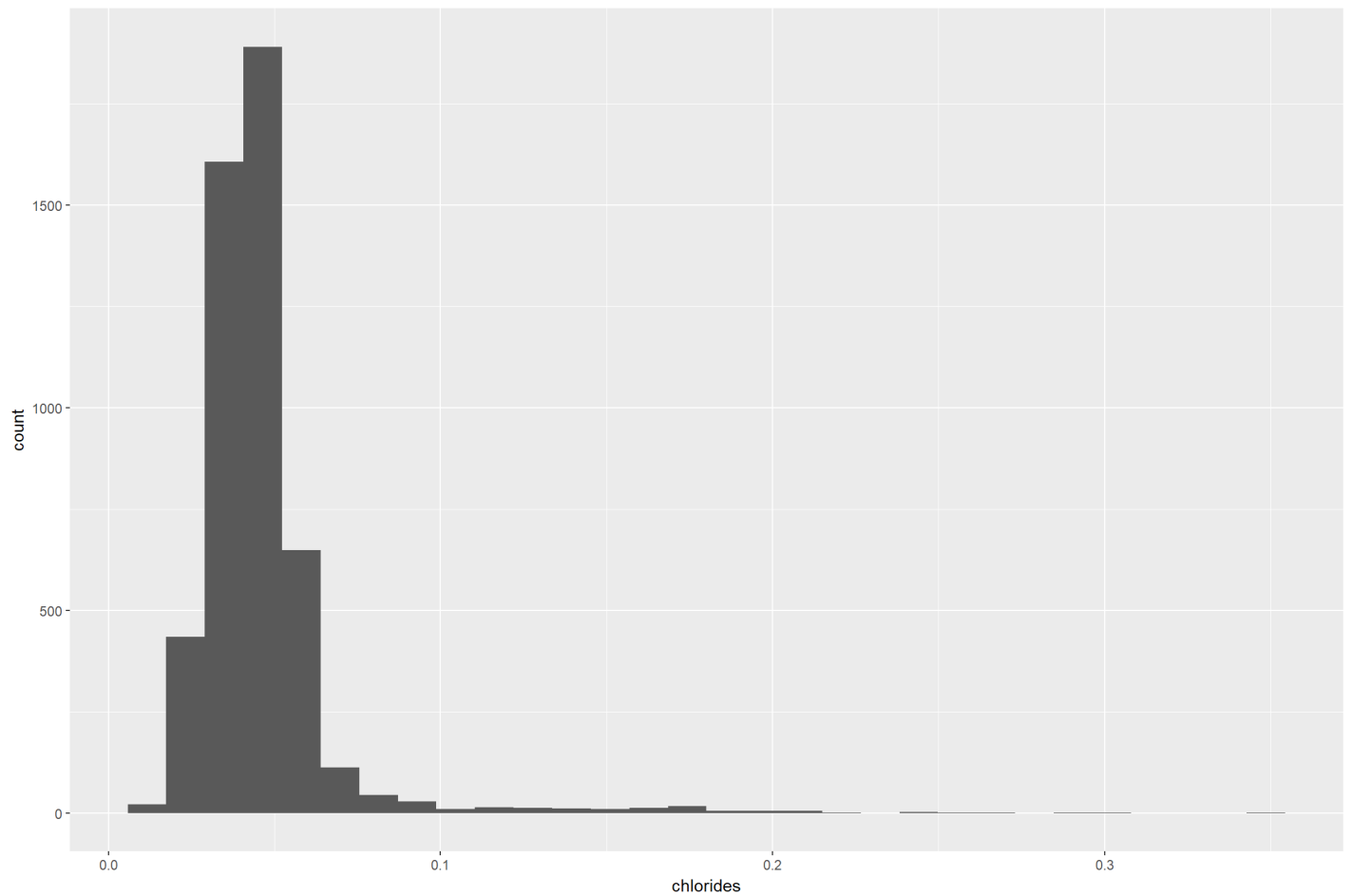
## Histogram of Free.Sulfur.Dioxide





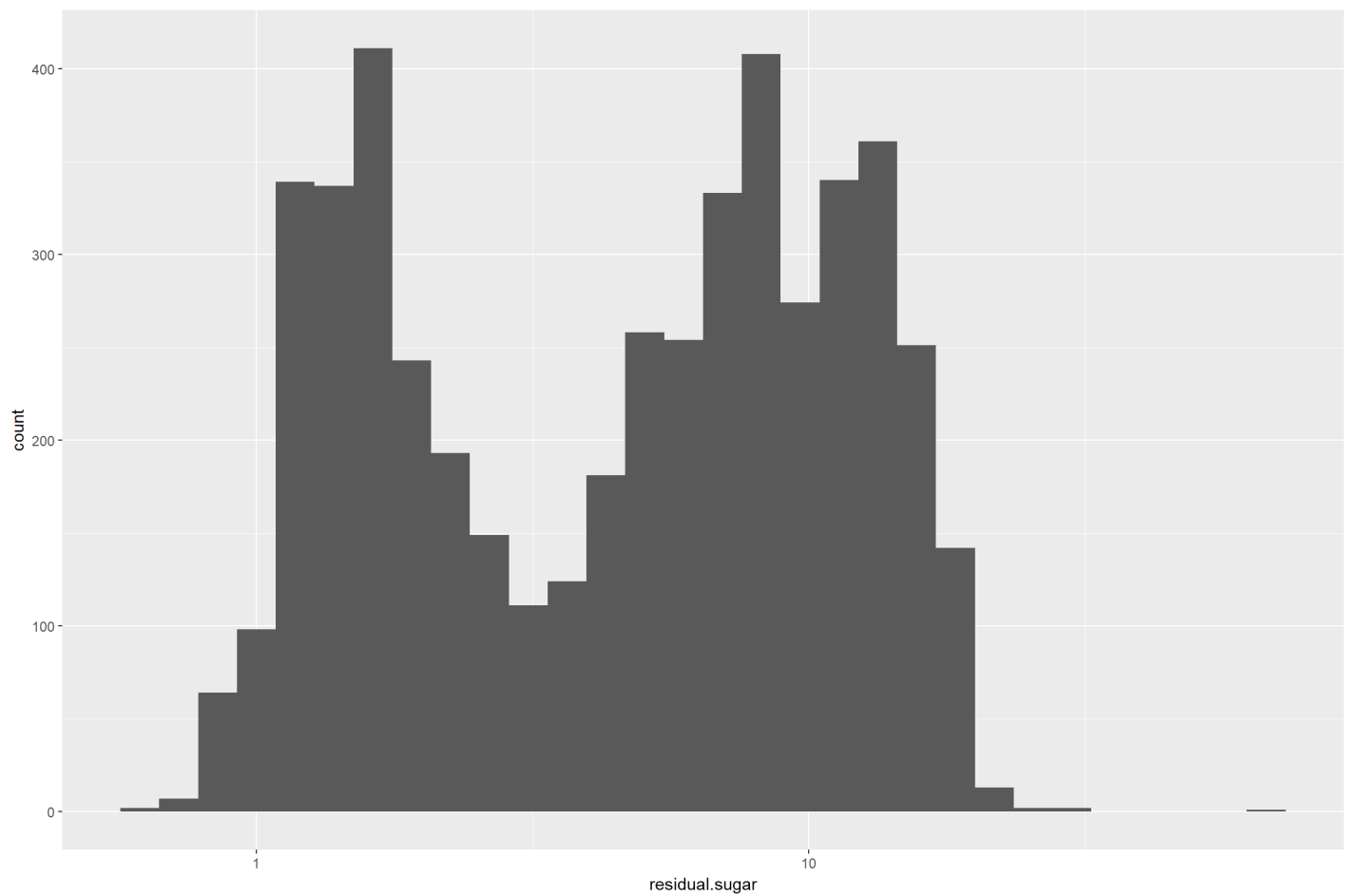
This histogram shows us that free.sulfur.dioxide in wine range from 0 to 100, and show normal distribution.

## Histogram of Chlorides



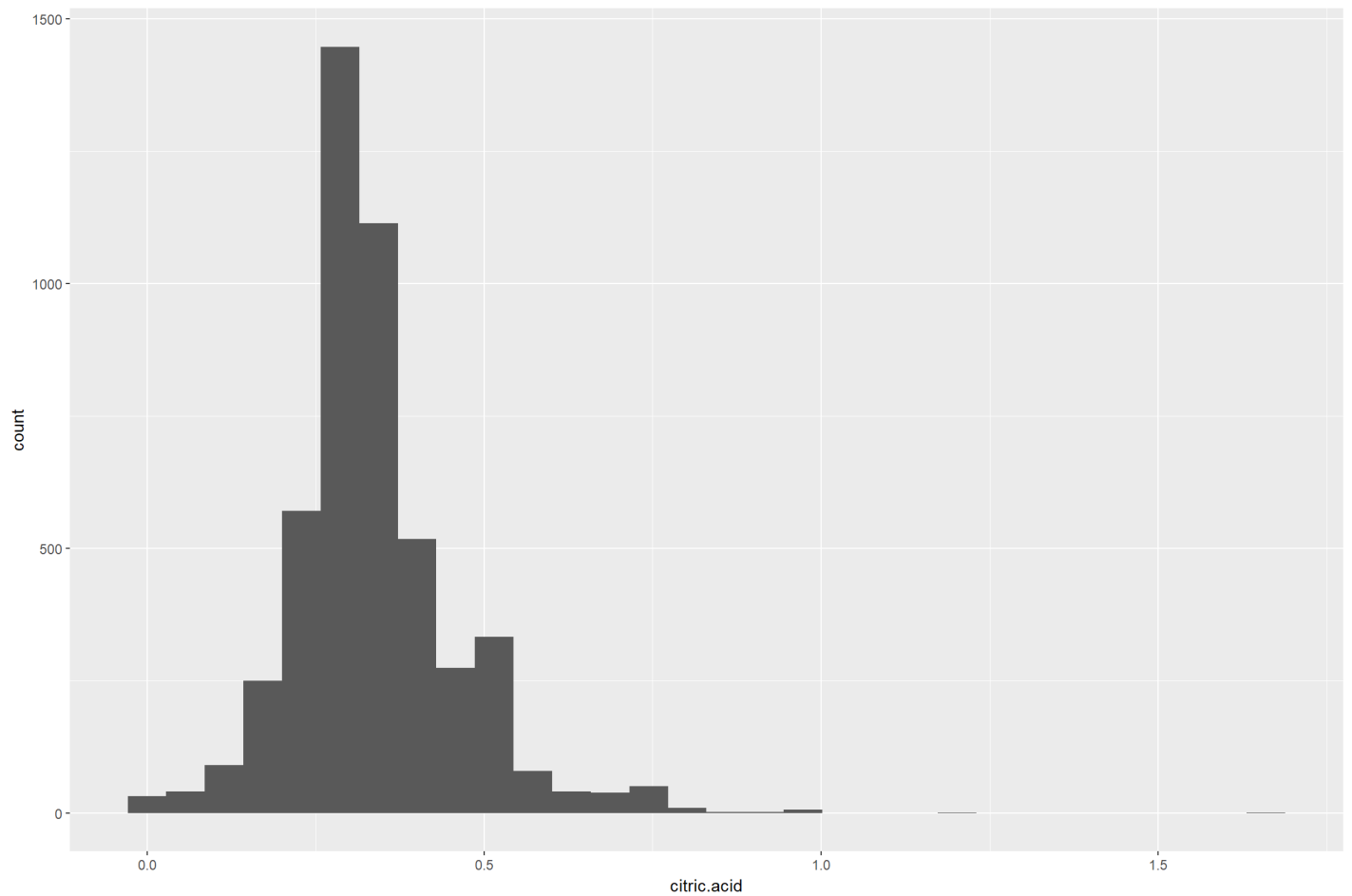
Majority of chlorides in wine are above 1500 and range between 0 to 0.1

## Histogram of Residual.sugar



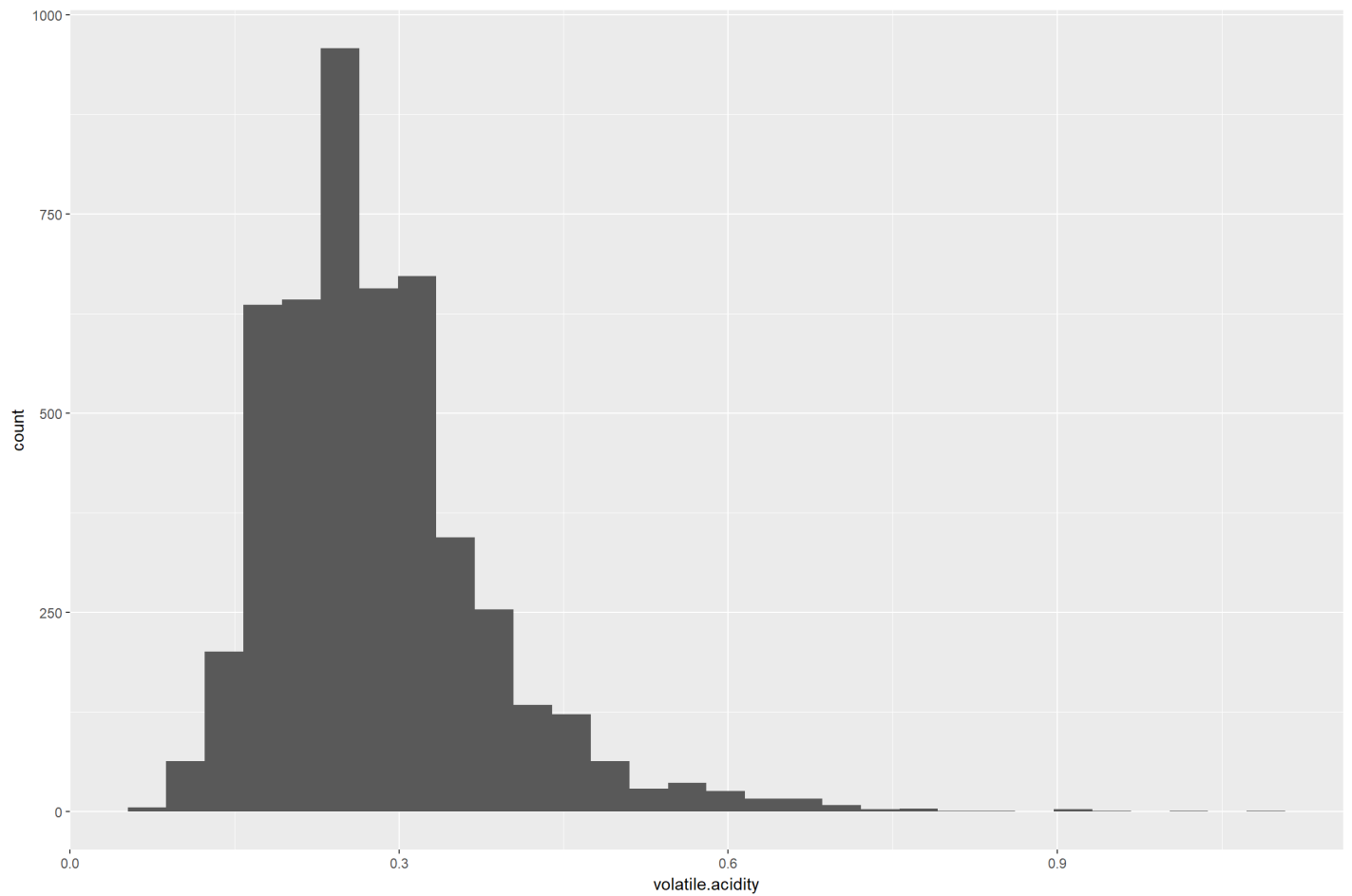
The histogram of residual sugar shows that as the count of sugar in whitewine does not follow any standard distribution. It has low peaks at left and right and low peak in the middle of distribution.

## Histogram of Citric.Acid



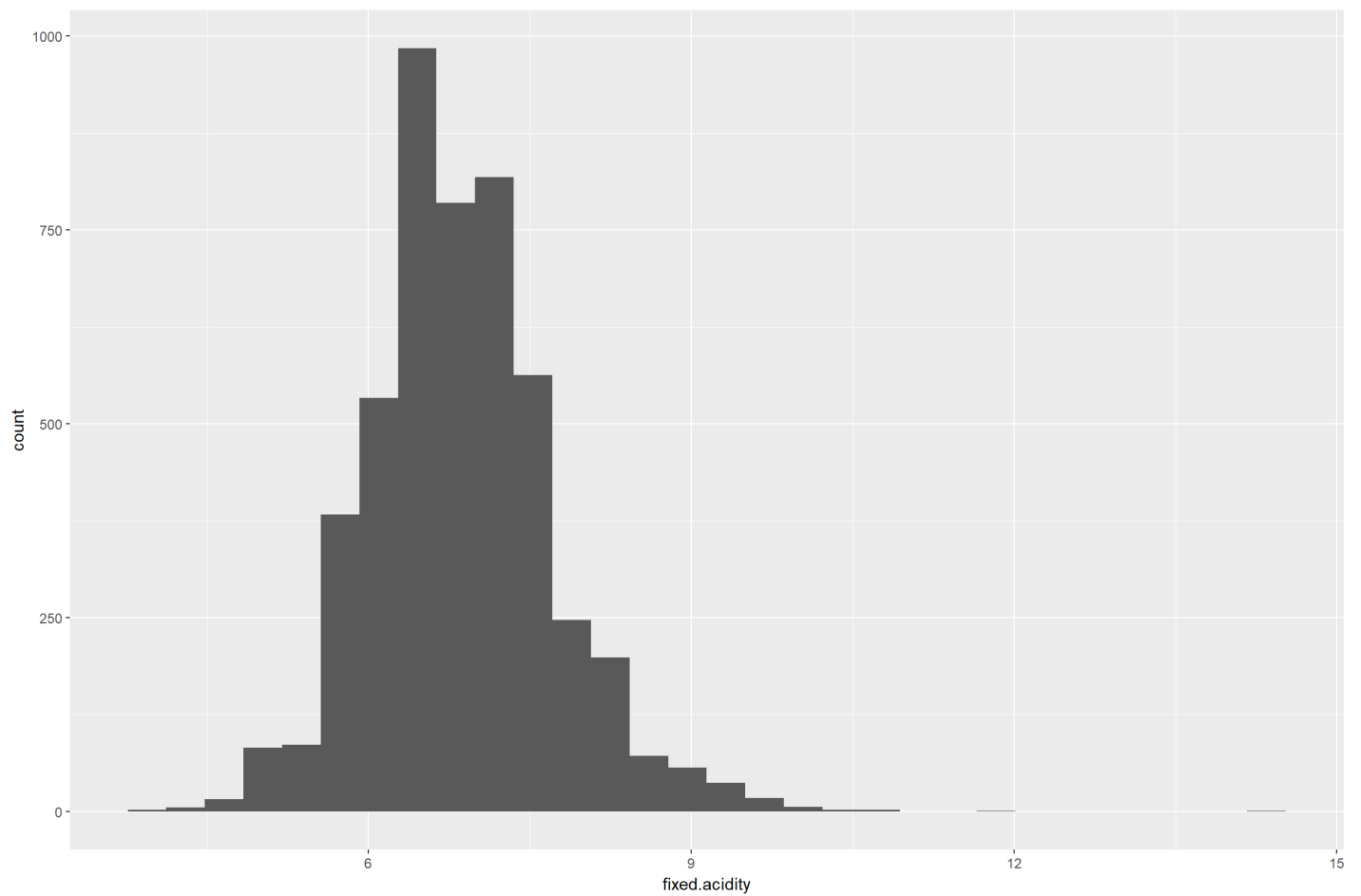
The Citric.Acid value ranges from 0 to 0.5 and the distribution is normal in white wine.

## Histogram of Volatile.Acidity



There are lot of 0.29 valued entries in volatile.acidity.

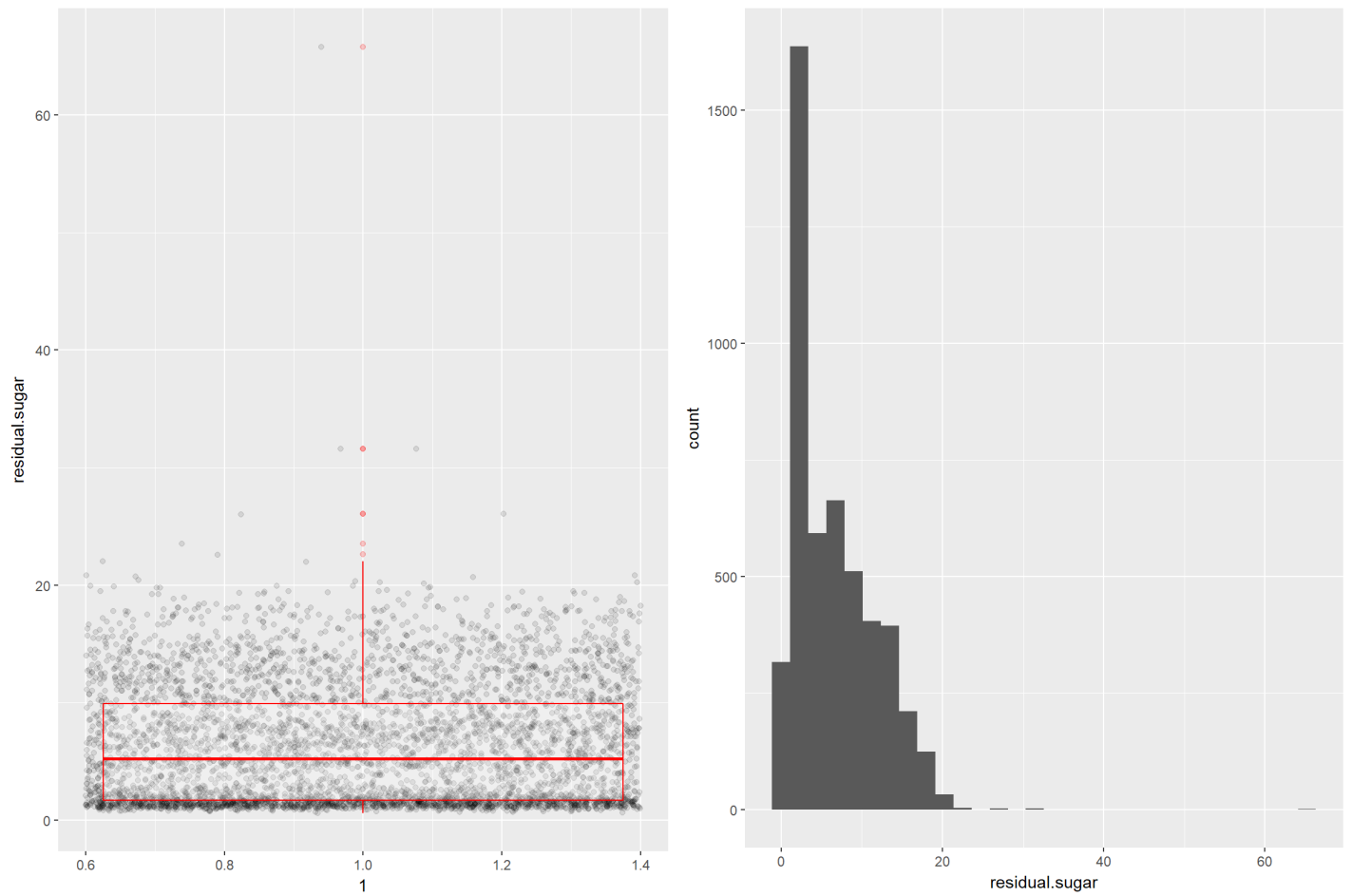
## Histogram of fixed.acidity value of wine



The histogram of acidity of potential shows normal distribution.

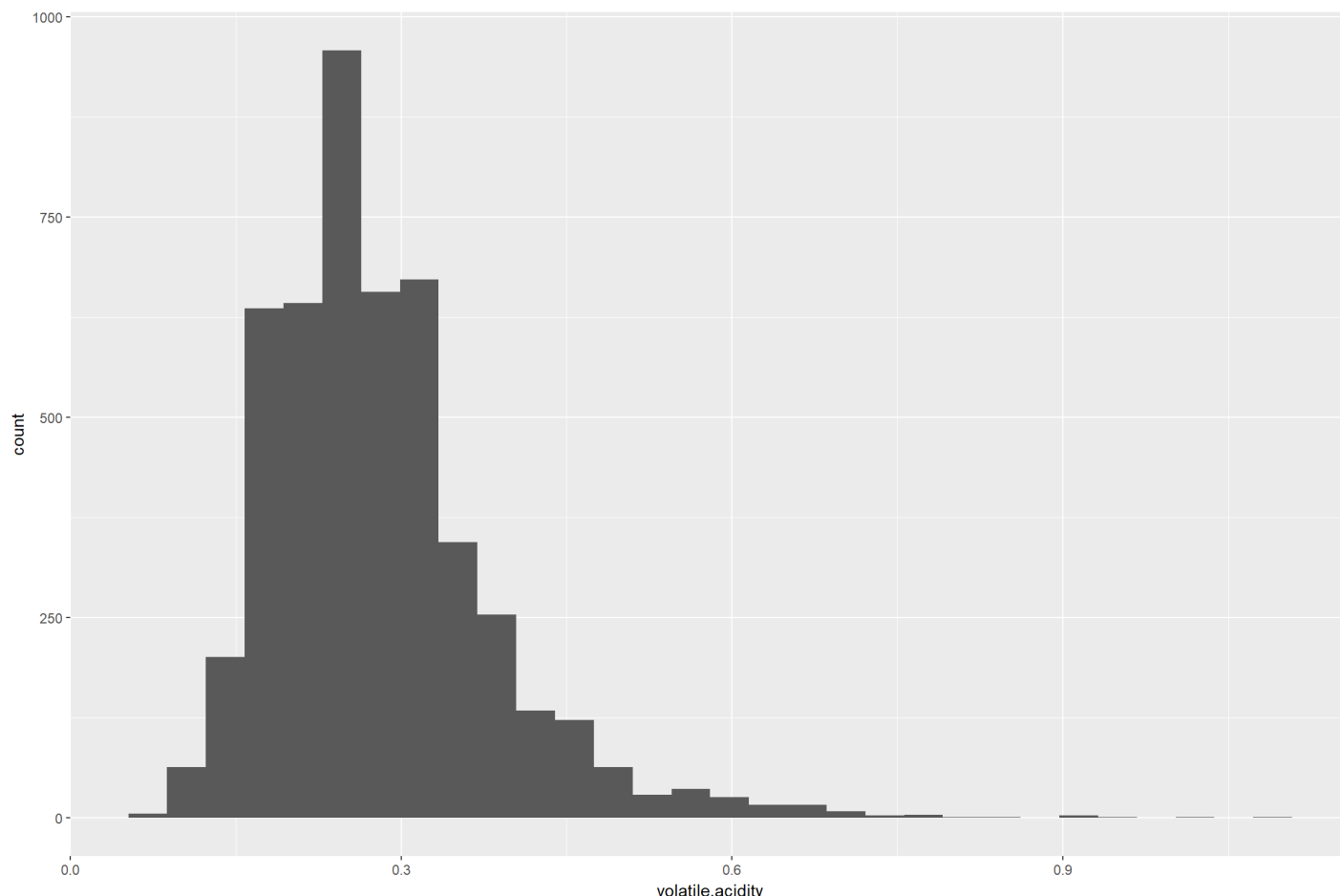
## Outliers

Residual.sugar



There is unusual distribution of 1s of the residual.sugar in the dataset.

Volatile.Acidity



There is a lot distributions valued 0.29 in the dataset. However this does not seem to be an issue because there are other values in Volatile.Acidity as well.

## Creating new variable: Rating

### Univariate Analysis Questions and Findings

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

I think further investigation whether pH could be classified into categorical (acid, base, neutral) could potentially be of interest. Also the relationship between fixed.acidity, volatile.acidity, and citric.acid (in particular this one due to the large number of 0s) could be interesting to further investigate in order to understand if the values are properly reported. This would require to understand the theoretical relationship between these variables.

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

I created new rating variable, to visualize the quality of wine by setting threshold on quality values, if  $\text{wine.quality} < 5$ , rating is 'bad', if  $5 < \text{wine.quality} < 7$  rating is 'average' else the rating is 'good'.

I created the variable (ordered factor) rating to classify wine as good, bad or average. The information that the other variables store there does not seem to be any other variable that would fit well a sub-classification into categorical variables, perhaps with the exception of residual pH (neutral, base, acid).

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



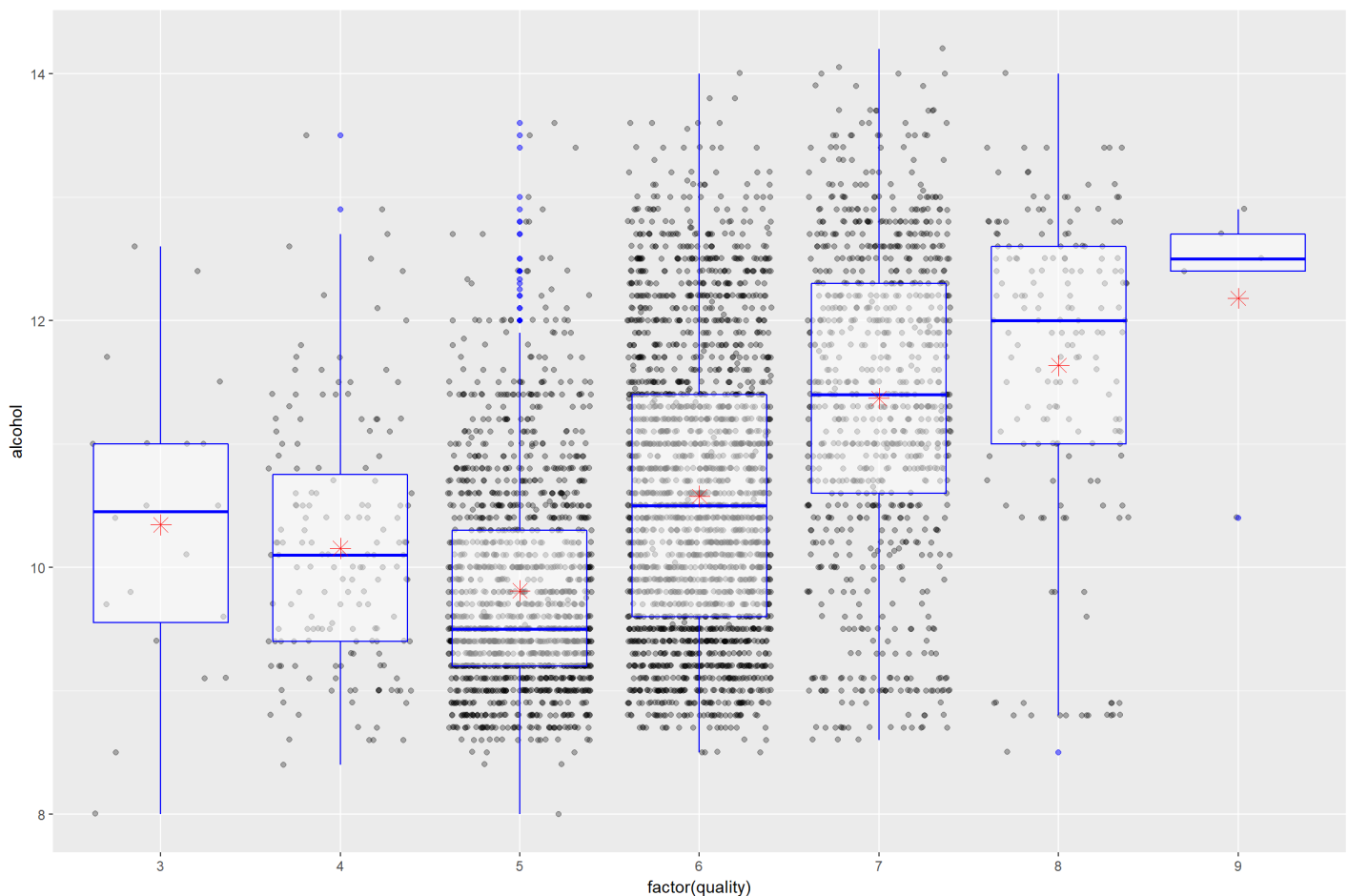
I created Boxplots in Bivariate analysis to gain a better understanding wrt. outliers (except for X for being an index, and quality for being categorical). I've also plotted the data using a log10 scale in Univariate analysis (I plotted for all variables, although only those with long tails are interesting for the log10 scale – plots below the box plots). These plots have shown that fixed acidity and to some extent pH, chlorides, density, sulphates, volatile acidity to follow a normal distribution. As for the acidity variables, this is aligned with the fact that pH seems to be normally distributed, apart from the citric acid. The reason for the latter might be the number of 0 (potentially non-responses) discussed earlier. pH is normally distribution which suggests that the data is good, since by definition it is a measure of acidity and is on a logarithmic scale.

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've not found the need to tidy, adjust, or change the form of the data.

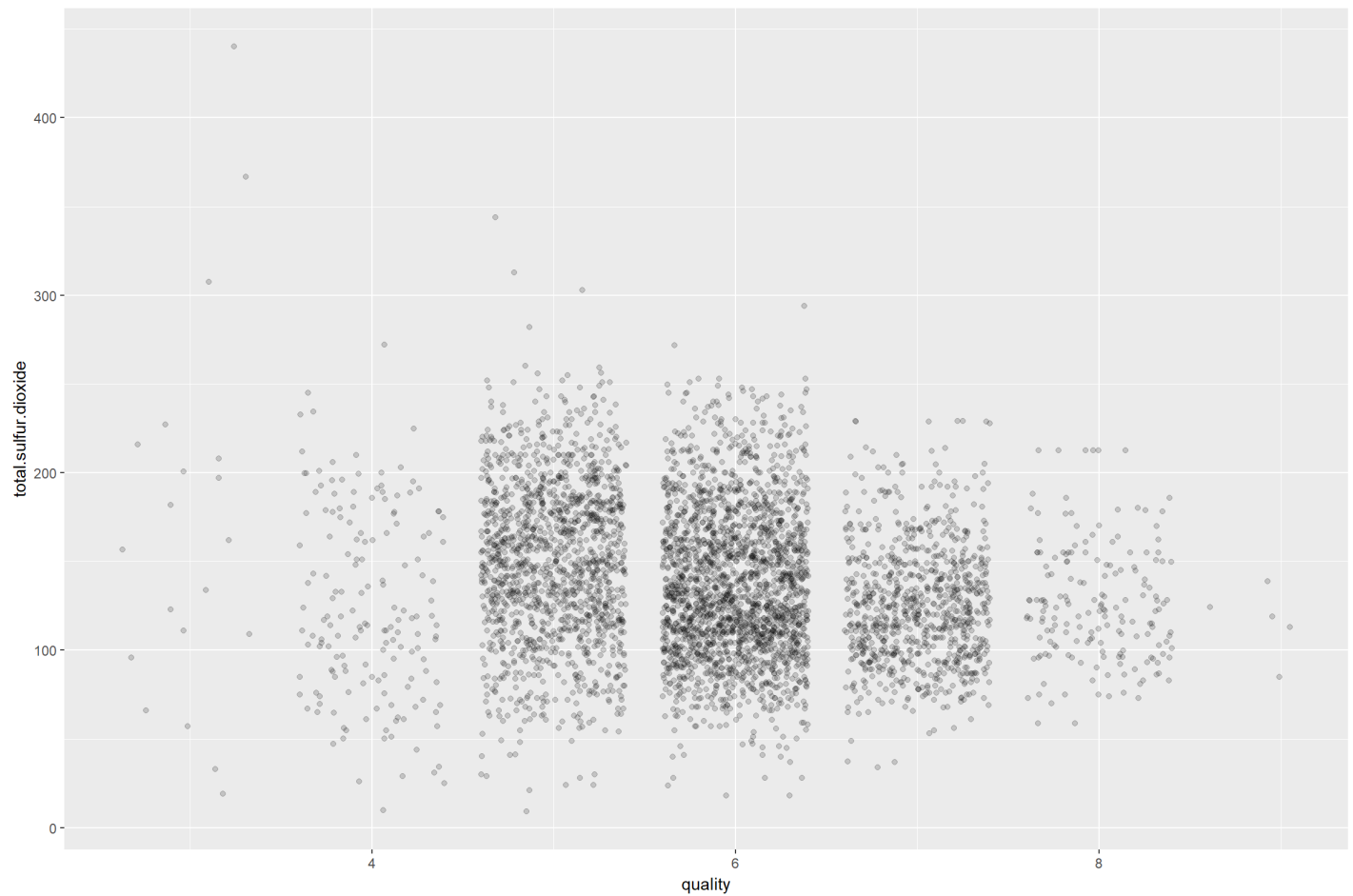
## Bivariate Analysis

Let us check how is the distribution of quality and alcohol content in wine. To do this creating a new variable alcohol.bucket.



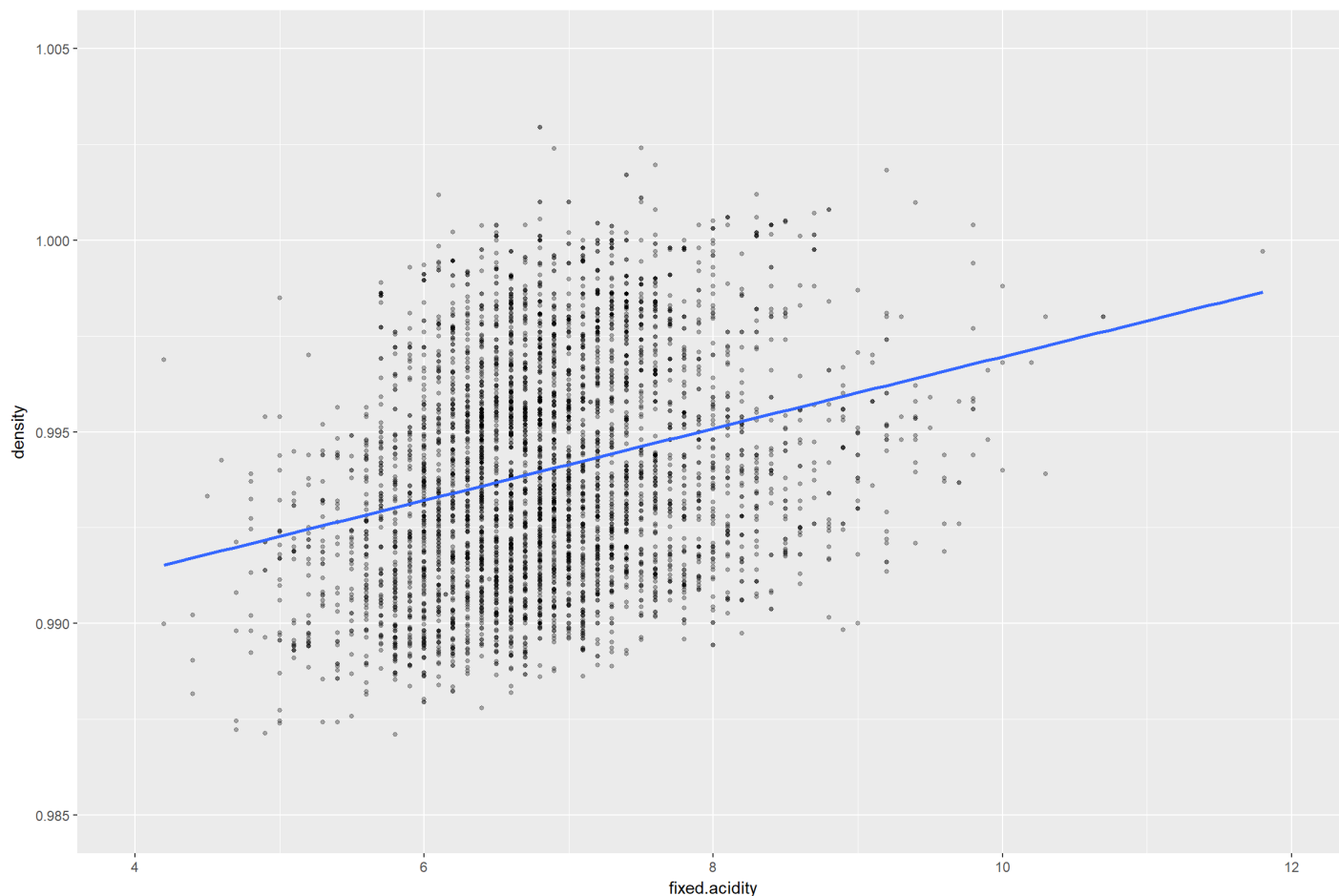
The graphs shows us that alcohol and wine quality have clear correlation.

Scatterplot of quality and Sulfur.Dioxide



This plot shows that in white wine quality of 5,6,7 has majority of Sulfur.Dioxide in it.

## Scatter plot of Fixed.Acidity and Density



We see that there is linear relationship between fixed.acidity and density of wine. As the fixed.acidity increases the density increases however between fixed.acidity between 5 and 8 the density is between 0.9 to 1.0.

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?

I noticed that the following characteristics yield 'good' wines:

- Acidity: higher fixed acidity and citric acid; lower volatile acidity
- Lower pH (~3.5)
- Higher Sulphates
- Higher Alcohol
- Lower density
- Regular chlorides and sugar do not seem to impact much on the quality.

Did you observe any interesting relationships between other features and not the main features of interest?

Below I have calculated the correlation between variables that correlate the most with quality: citric acid, sulphates, alcohol and volatile acidity.

Also the Pearson test revealed that volatile acidity and citric acid have strong negative correlation.

```
##
## Pearson's product-moment correlation
##
## data: citric.acid and alcohol
## t = -5.3141, df = 4896, p-value = 1.119e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.1035156 -0.0478237
## sample estimates:
## cor
## -0.07572873
```

```
##
## Pearson's product-moment correlation
##
## data: sulphates and alcohol
## t = -1.22, df = 4896, p-value = 0.2225
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.04541705 0.01057885
## sample estimates:
## cor
## -0.01743277
```

```
##
## Pearson's product-moment correlation
##
## data: volatile.acidity and alcohol
## t = 4.7492, df = 4896, p-value = 2.1e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## 0.03978694 0.09554320
## sample estimates:
## cor
## 0.06771794
```

```
##
## Pearson's product-moment correlation
##
## data: sulphates and citric.acid
## t = 4.3699, df = 4896, p-value = 1.269e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## 0.03438451 0.09017997
## sample estimates:
## cor
## 0.06233094
```

```
##
## Pearson's product-moment correlation
##
## data: sulphates and volatile.acidity
## t = -2.5015, df = 4896, p-value = 0.0124
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.063670893 -0.007729425
## sample estimates:
##          cor
## -0.03572815
```

```
##
## Pearson's product-moment correlation
##
## data: volatile.acidity and citric.acid
## t = -10.578, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.1767384 -0.1219760
## sample estimates:
##          cor
## -0.1494718
```

## What was the strongest relationship you found?

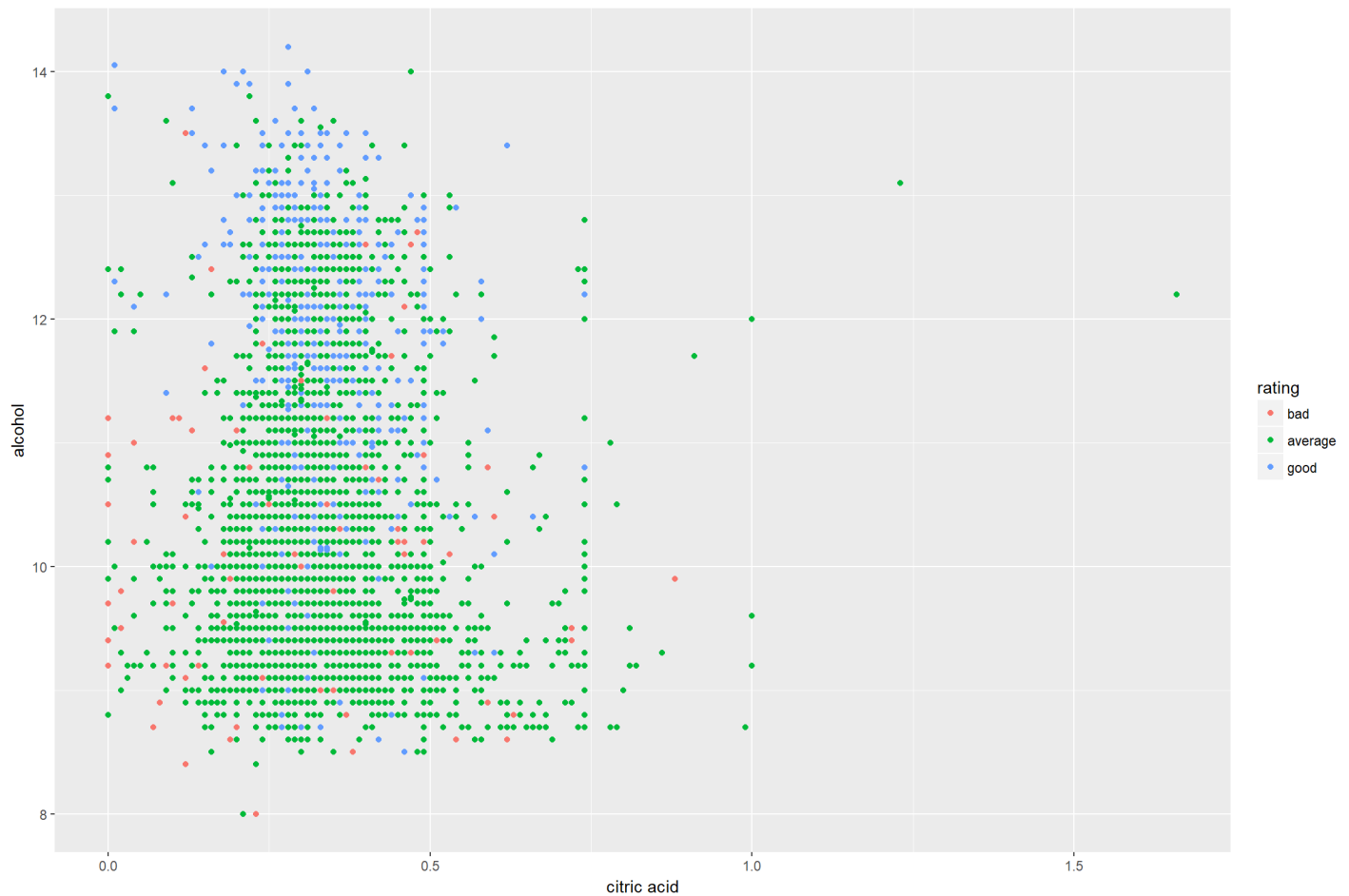
The quality shows the strongest correlation with alcohol (0.4761663). I also observed that volatile.acidity and alcohol, and sulphates and alcohol show weakest bi-variate relationship.

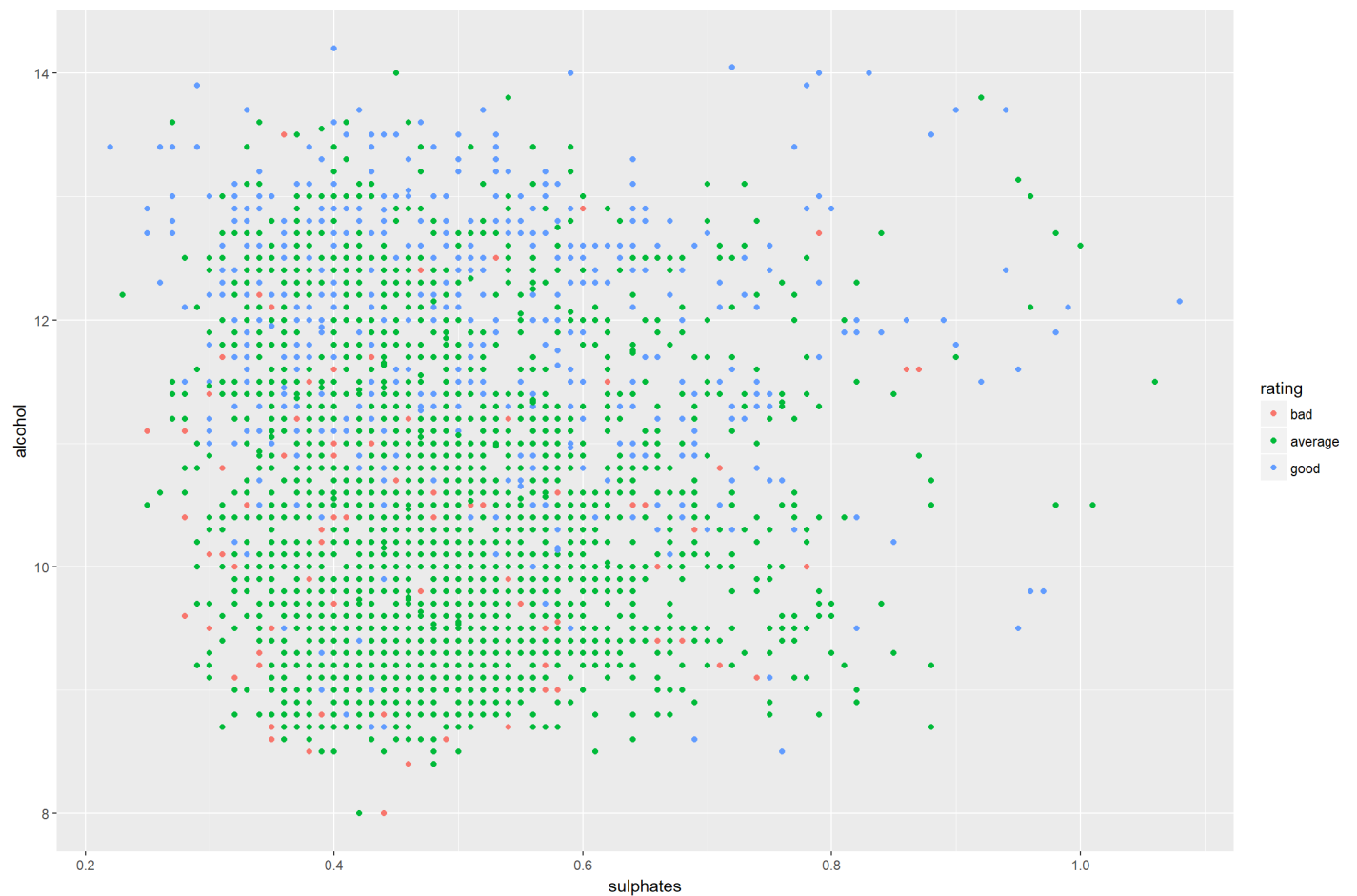
```
##
## Pearson's product-moment correlation
##
## data: quality and alcohol
## t = 33.858, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## 0.4126015 0.4579941
## sample estimates:
##          cor
## 0.4355747
```

## Multivariate Analysis

Talk about some of the relationships you observed in this investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

I focused on 4 features that showed strong correlation with quality and pH, I found that the key factors that help in achieving good wine quality are higher sulphates, higher alcohol, lower volatile acid and higher citric acid.

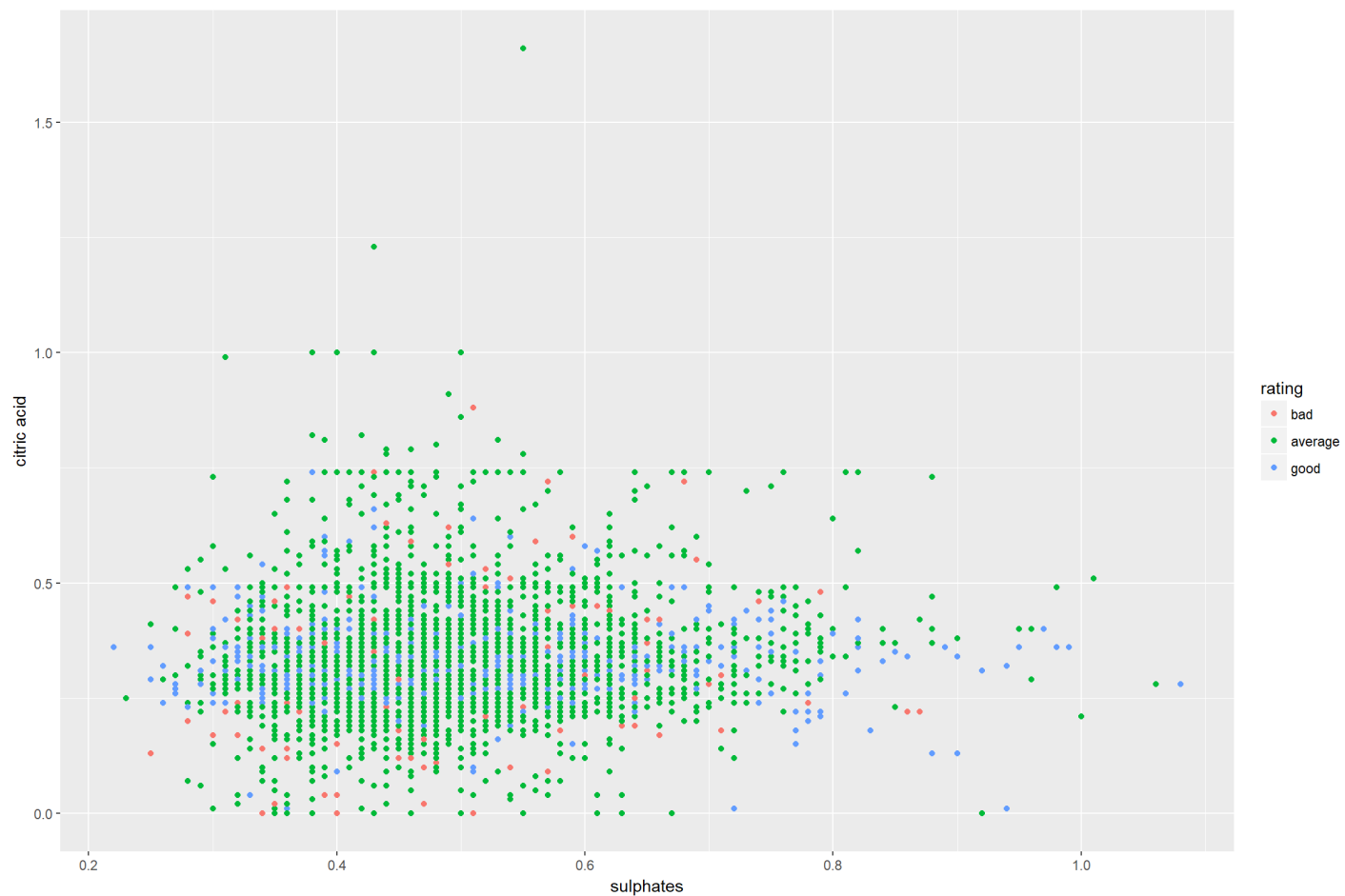
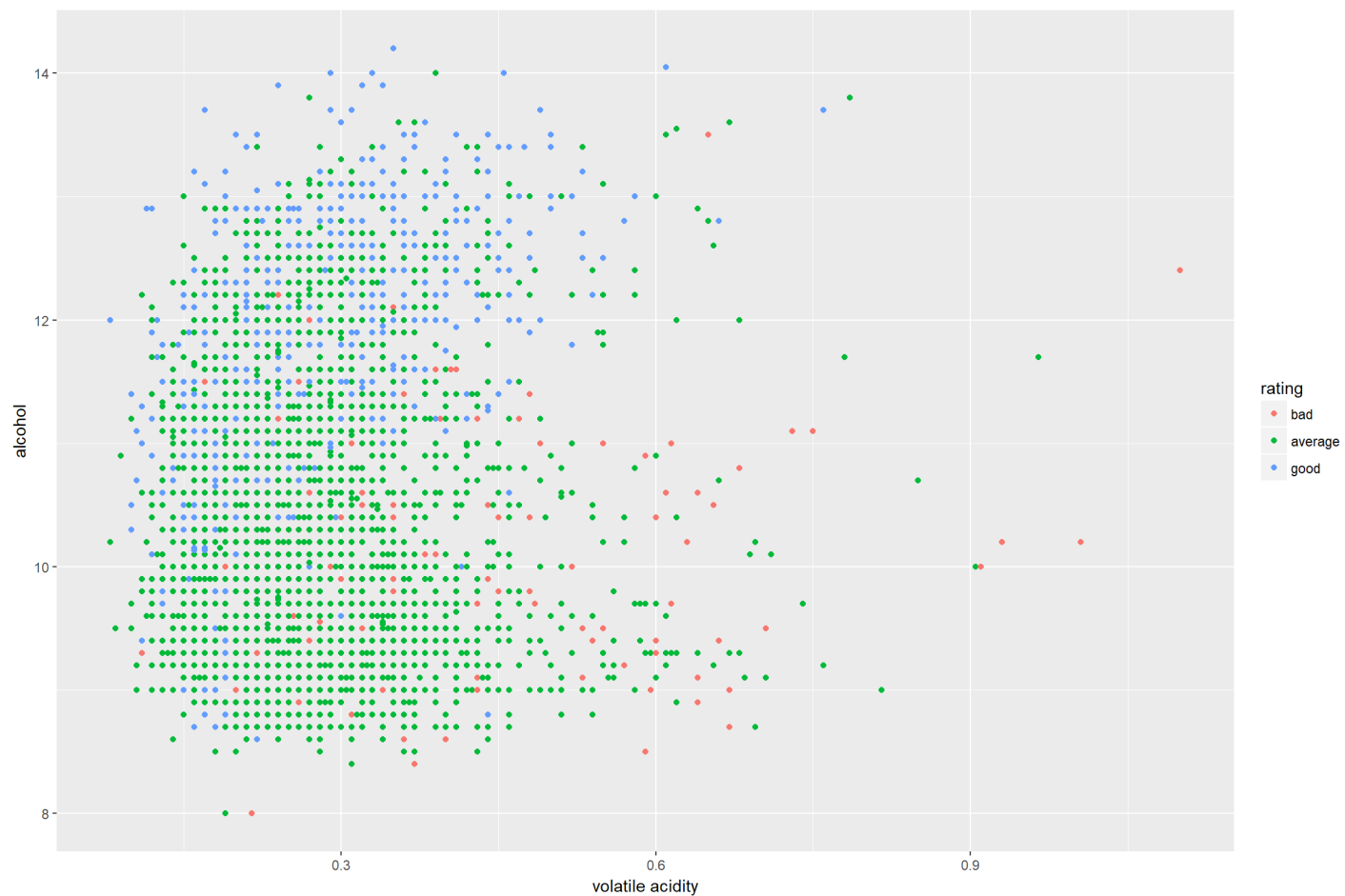


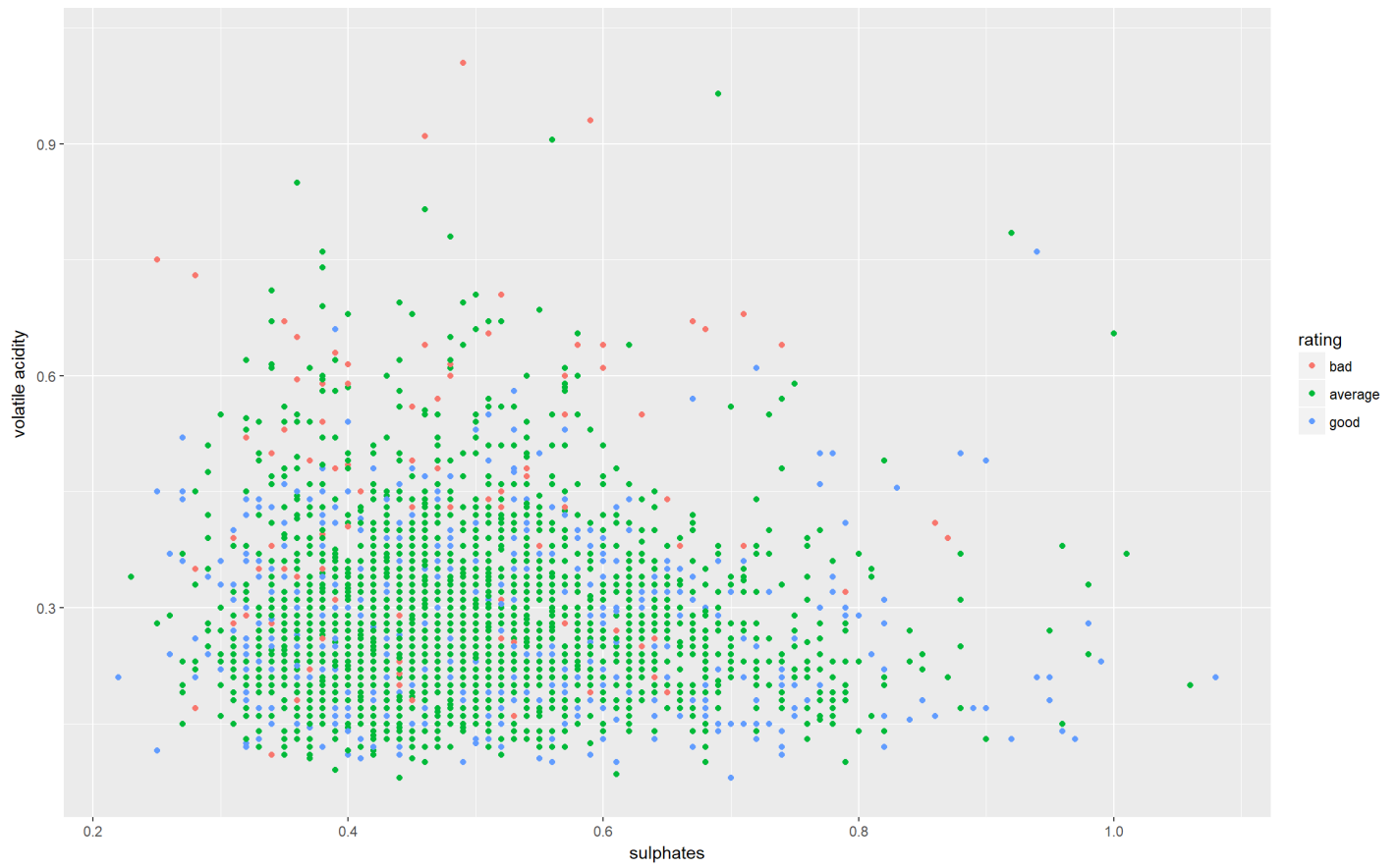


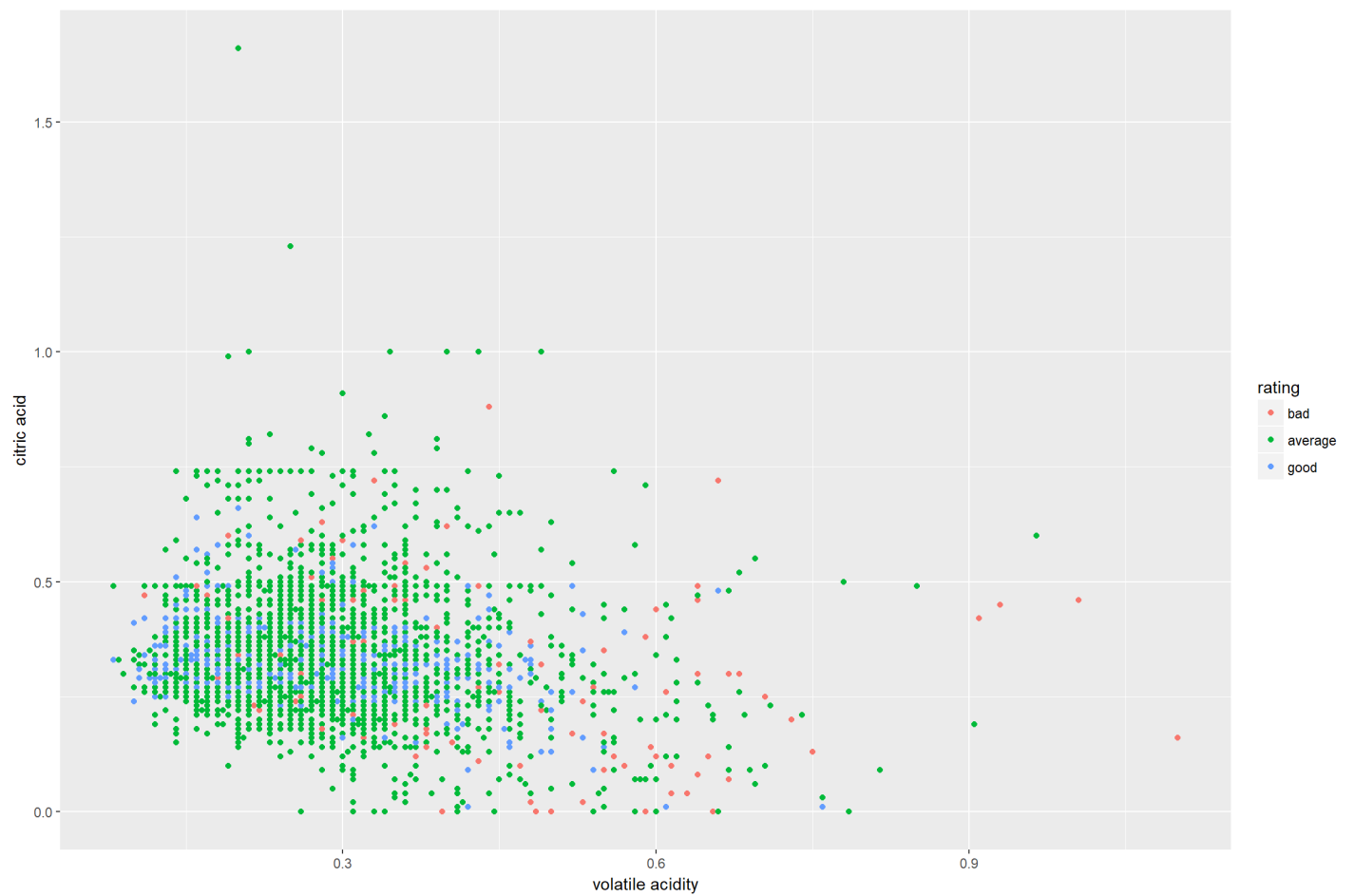
This scatter plot shows us that the majority of the ratings given to alcohol are average and it depends on the key factors above. Also when the alcohol content is low there are majority of bad rated wines and where alcohol content is high, wine quality is rated as good.



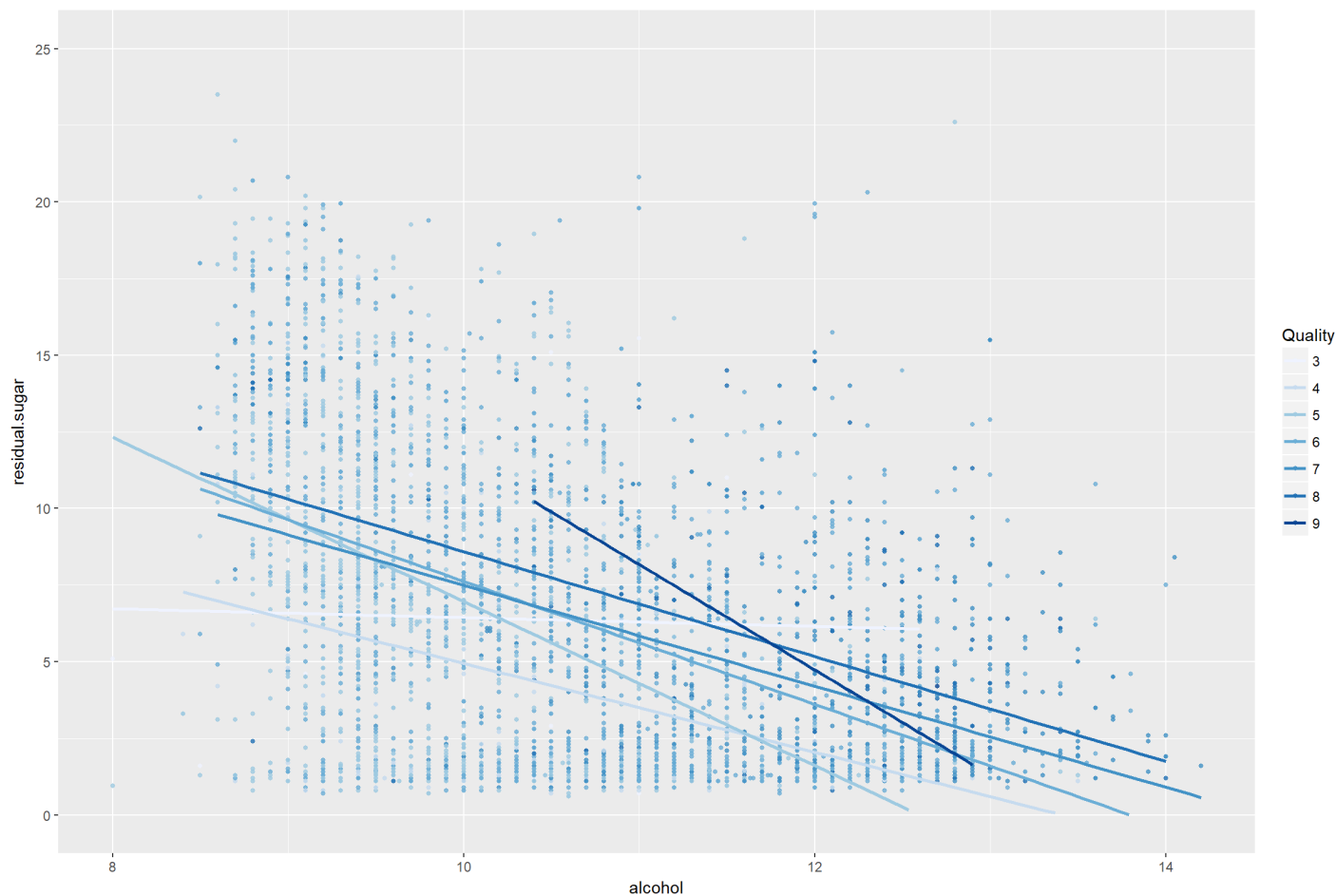








In the plots above, Sulphates and Citric.Acid, Sulphates and Volatile.Acidity we see a good number are rated as average and there is mix of good and bad ratings so we cannot infer anything. The pH and Alcohol relationship shows that similar to Sulphates and Alcohol, majority of the wines have good rating where alcohol content is higher, and bad rating where alcohol content is lower, even though pH is fairly constant in both cases. And Higher Volatile.Acidity causes the wine quality to decrease as with it the Citric.Acid increases.

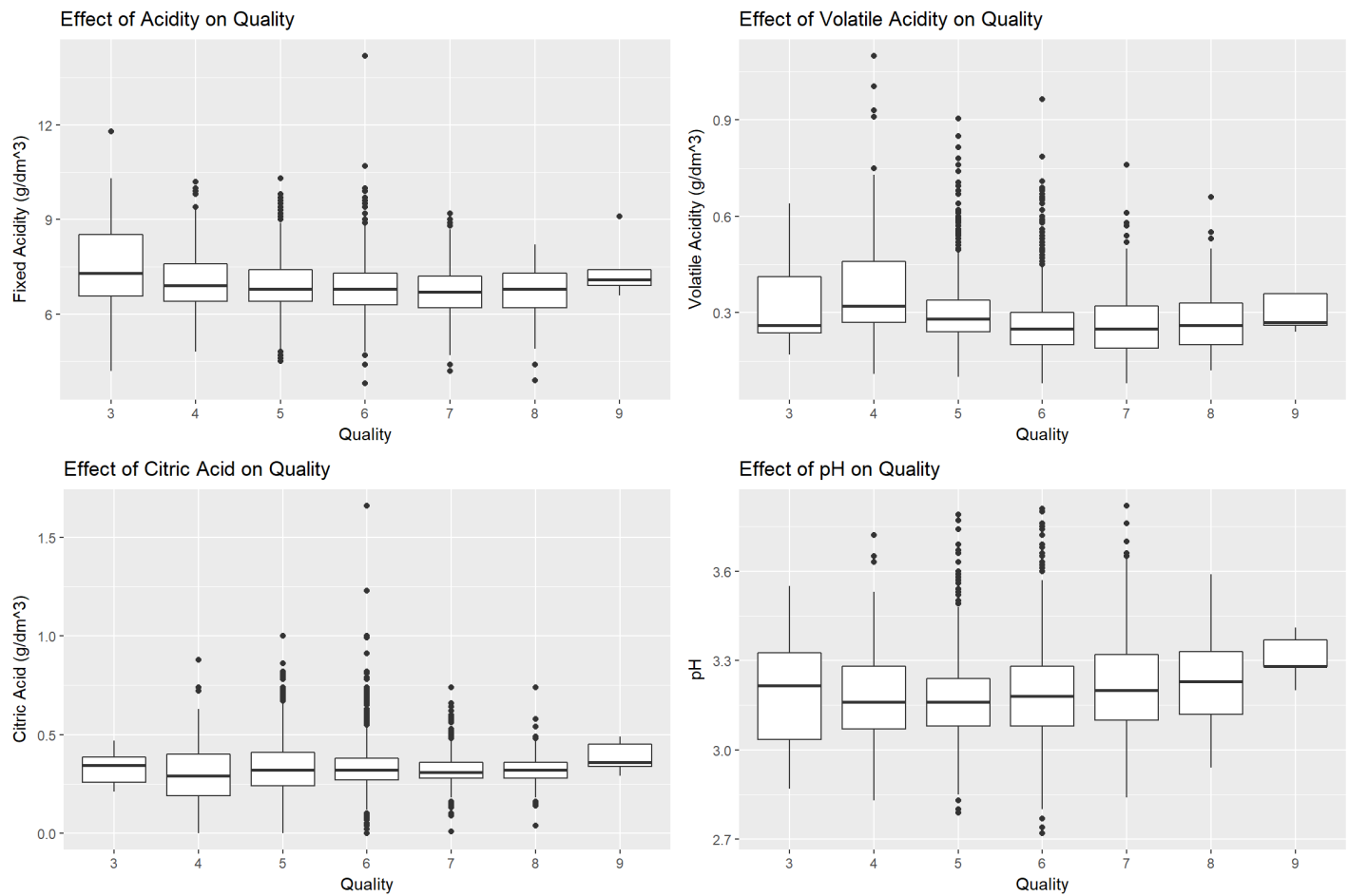


From the above plot we can see that lower residual sugar and higher alcohol content leads to better wine quality.

## Final Plots and Summary

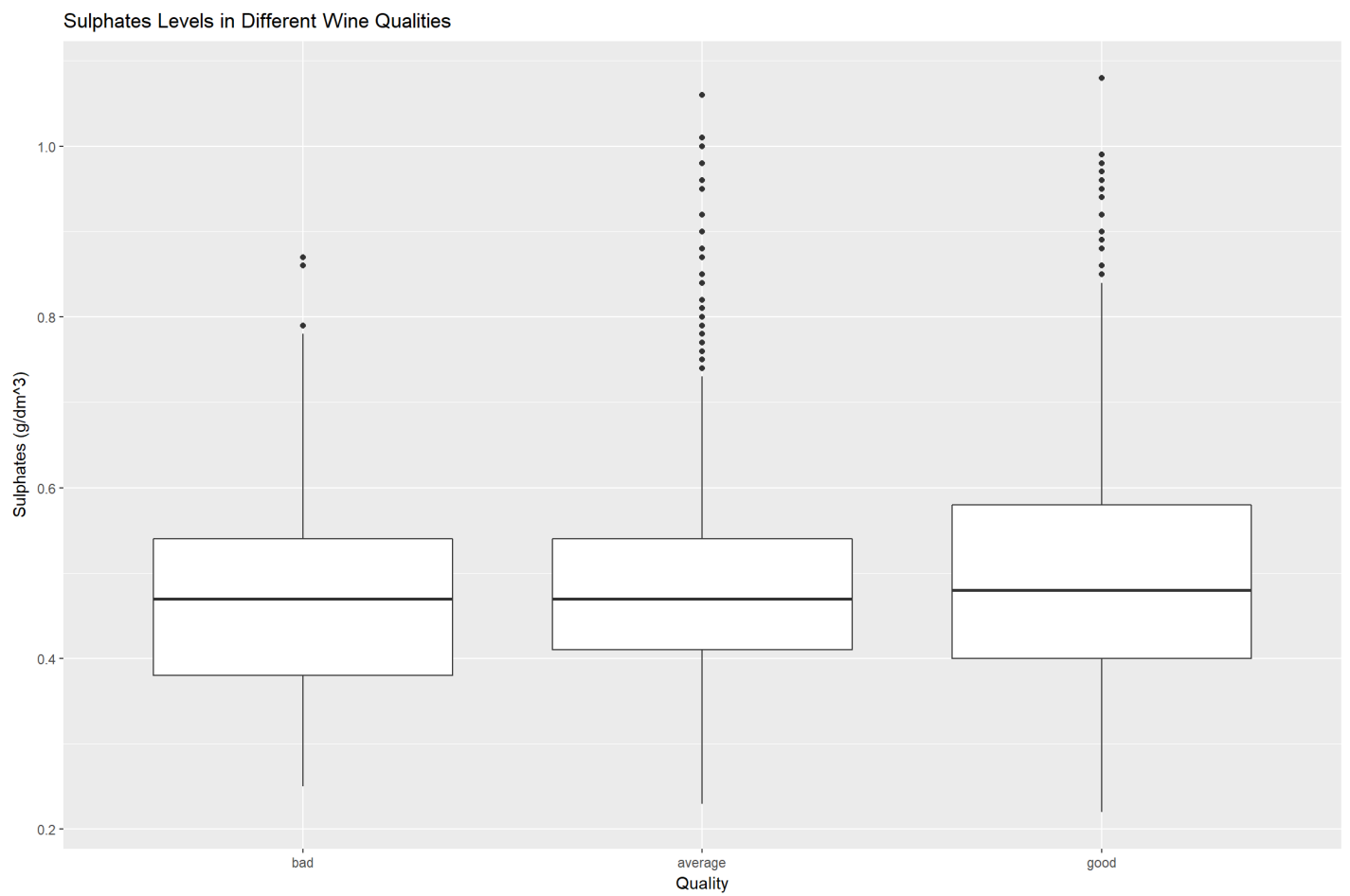
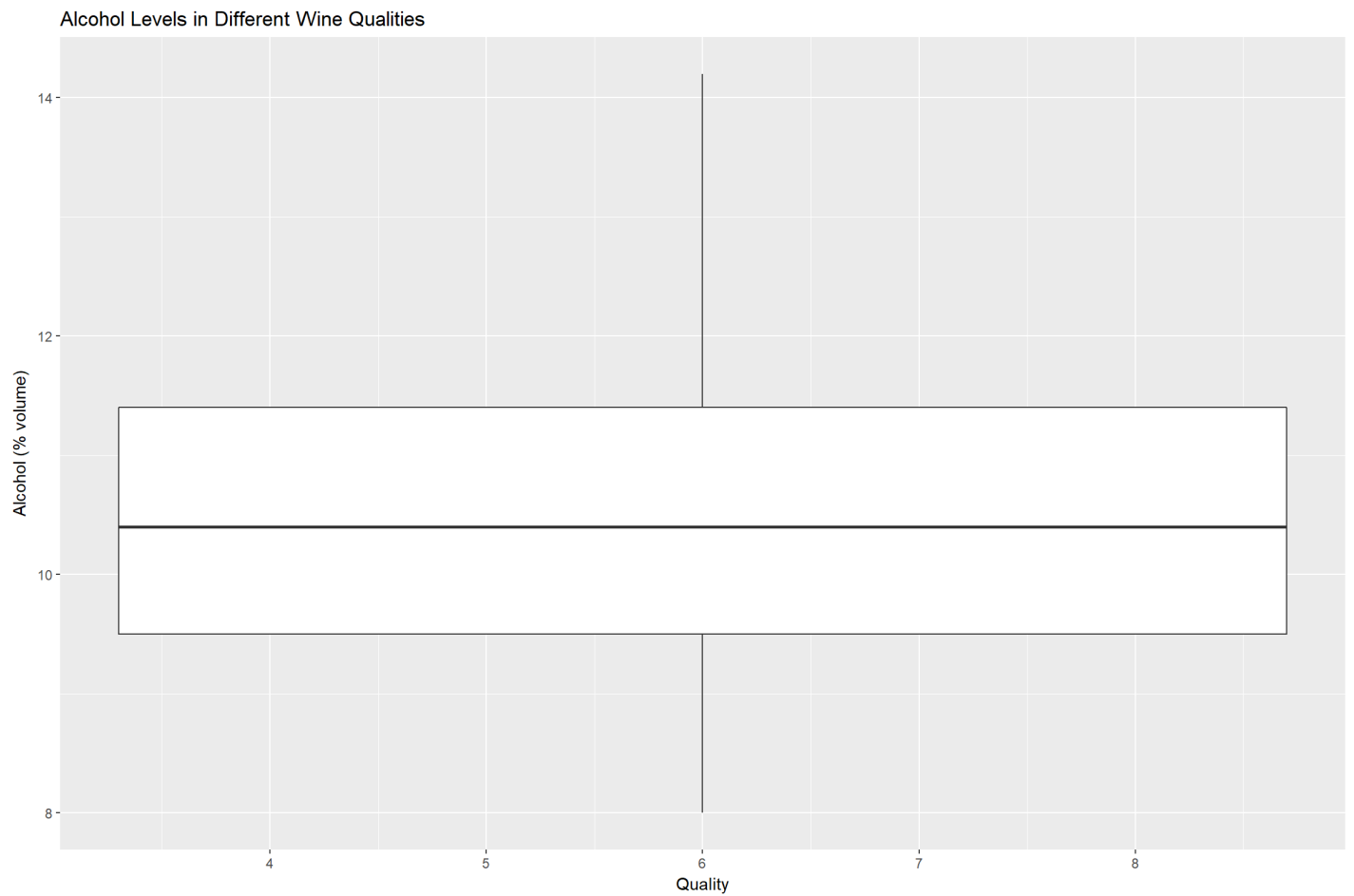
### Plot One

These plots show the effect of acidity and pH on wine quality. I observe that higher acidity (lower pH), apart from the volatile acid, is shown to yield better wines. Higher acidity and lower pH also makes sense because these variables are correlated and this is why acidity is inline with pH (making a balance for good wine).



## Plot Two

These plots show the effect of alcohol on wine quality. On average, we observe that the wine quality increases with alcohol, except for wine quality 5. To further understand the difference between quality 4 and 5, I computed the stats considering only wines with that rating.



```
## [1] "Alcohol summaries for quality = 4"
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      8.40   9.40   10.10   10.15   10.75   13.50
```

```
## [1] "Alcohol summaries for quality = 5"
```

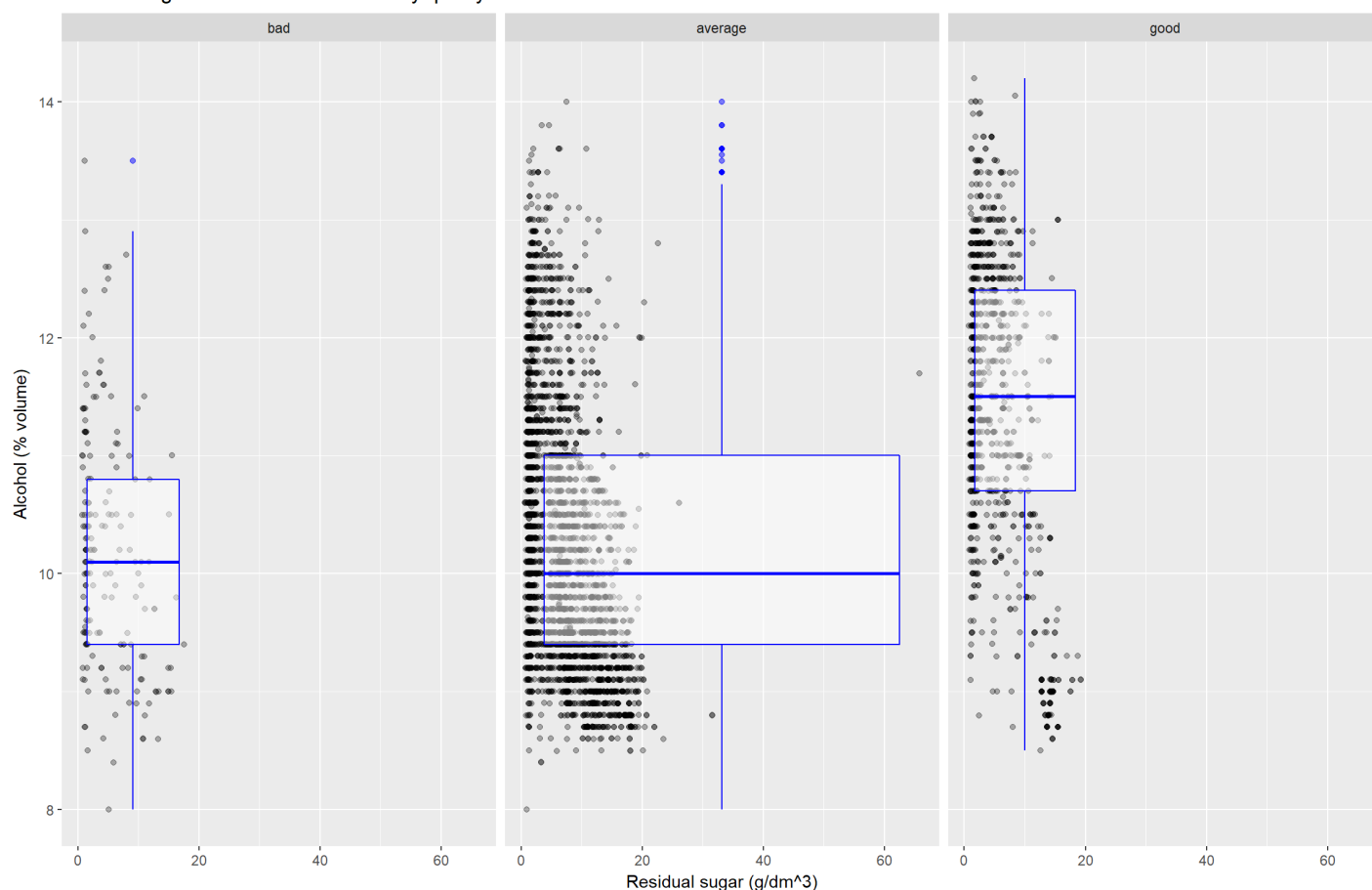
```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      8.000   9.200   9.500   9.809   10.300   13.600
```

I noticed tha the wine quality increases with sulphates. However, having more of it will not help much sulphates are only good for quality upto a certain level.

## Plot Three

This boxplot shows the correlation between quality and residual sugar vs. alcohol. I observe that lower residual sugars and higher alcohol leads to better wine quality.

residual sugar vs. alcohol correlation by quality



## Reflections

This study about white wine reveals following:

- Most wines in the dataset are rated as 'average' just small number is rated as bad.
- The data about citric acid is not accurate because many entries are set to 0.

- Higher alcohol level impacts the wine rating. That is, higher alcohol means better wine.
- More alcohol and low residual sugars yield to good red wines.
- Residual sugar and chlorides do not seem to have much impact on wine quality.
- The lower volatile acidity the better wine (as there is negative correlation).

Given that only alcohol correlated with quality of wine, it suggests that alcohol has impact on wine quality. However there are other factors like harvest year, vineyard location, temperature of harvest missing from the dataset. It will be interesting to do further exploration on these factors and see how it affects the quality of wine.

My major struggle in this project was to find more information about the wine that would have been important in this study. I spent sometime online researching about information on wines and this project increased my knowledge about wines. I think next time while drinking wine I will recall this project and check if my rating matches with what I have discovered.

## Thank you