# Exploratory Data Analysis - Red Wine Quality

by Tim Roberts

# Table of Contents

**Dataset**: a data set from 2009 testing the chemical properties of the Portuguese “Vinho Verde” red wine variant. At least 3 wine experts rated the quality of each wine between 0 (very bad) and 10 (very excellent). More information can be found here. (https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityInfo.txt).

**Question**: What variables contribute to making the best quality Red Wine?

---

# 1. Initial Analysis: structure of the dataset.

In this section, I will first look at the structure of the data set. Then I will examine the distribution of each attribute individually by plotting its distribution.

**List of Variables**

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

## Field Descriptions:

1. **Fixed acidity**: most wine acids involved are fixed or nonvolatile. (do not evaporate readily)
2. **Volatile acidity**: amount of acetic acid in wine - can be unpleasant, vinegary taste if too high?
3. **Citric acid**: found in small quantities, can add ‘freshness’ and flavor to wines.
4. **Residual sugar**: sugar remaining after fermentation stops, rare < 1 gram/liter, > 45 grams/liter are considered sweet.
5. **Chlorides**: amount of salt in the wine.
6. **Free sulfur dioxide**: the free form of SO2 - prevents microbial growth and the oxidation of wine.
7. **Total sulfur dioxide**: free + bound forms of S02; in low concentrations, mostly undetectable in wine, free SO2 over 50 ppm, evident in the nose and taste of wine.
8. **Density**: the density of water is close to that of water (approx 1) depending on the percent alcohol and sugar content.
9. **pH**: acidic on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale.
10. **Sulphates**: anadditive which can contribute to S02 levels, acts as an antimicrobial and antioxidant.
11. **Alcohol**: the percent alcohol content of the wine.
12. **Quality (Output Variable)** - sensory score between 0 and 10.

**Data Structure**

```
## 'data.frame':    1599 obs. of  12 variables:
##  $ 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 ...
```

The red wine dataset contains 12 variables. There are 11 input numerical variables based on physicochemical tests and 1 categorical output variable (quality) based on sensory data, with 1599 observations.

**All Descriptive Stats**

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

**Quality Table**

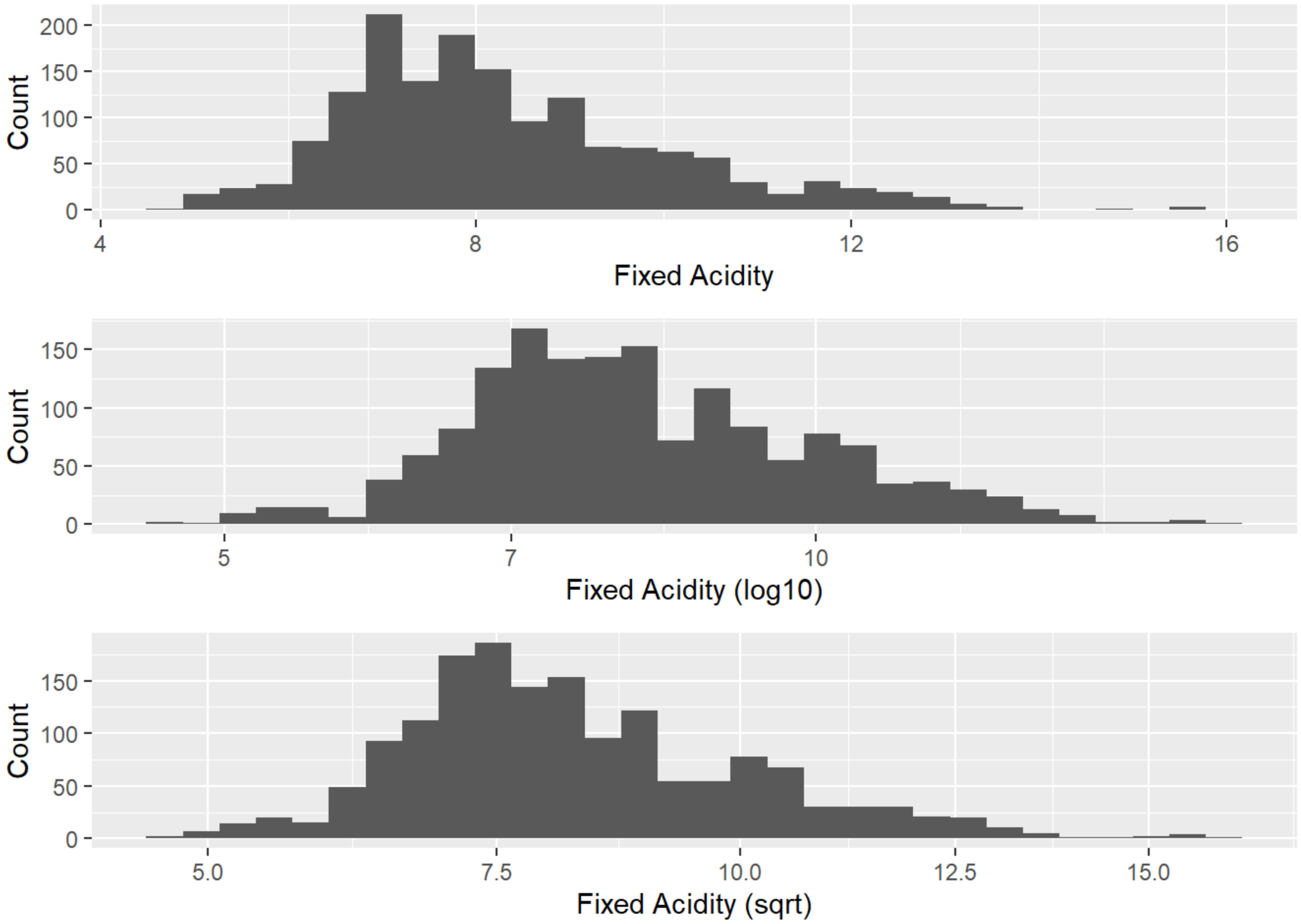
```
## 
##   3   4   5   6   7   8 
##  10  53 681 638 199  18
```

Fixed acidity values range between 0.1 and 1.6, with most values range between 0.3 and 0.7. The distribution is slightly positively skewed. Mean/Median seem to be relatively close on all variables except total.sulfur.dioxide and chlorides - long tailed? Thoughts at this stage - Quality range is between 3 and 8 - does this correlate with any of the other variables?

---

# 2. Univariate Exploration

**Fixed acidity**

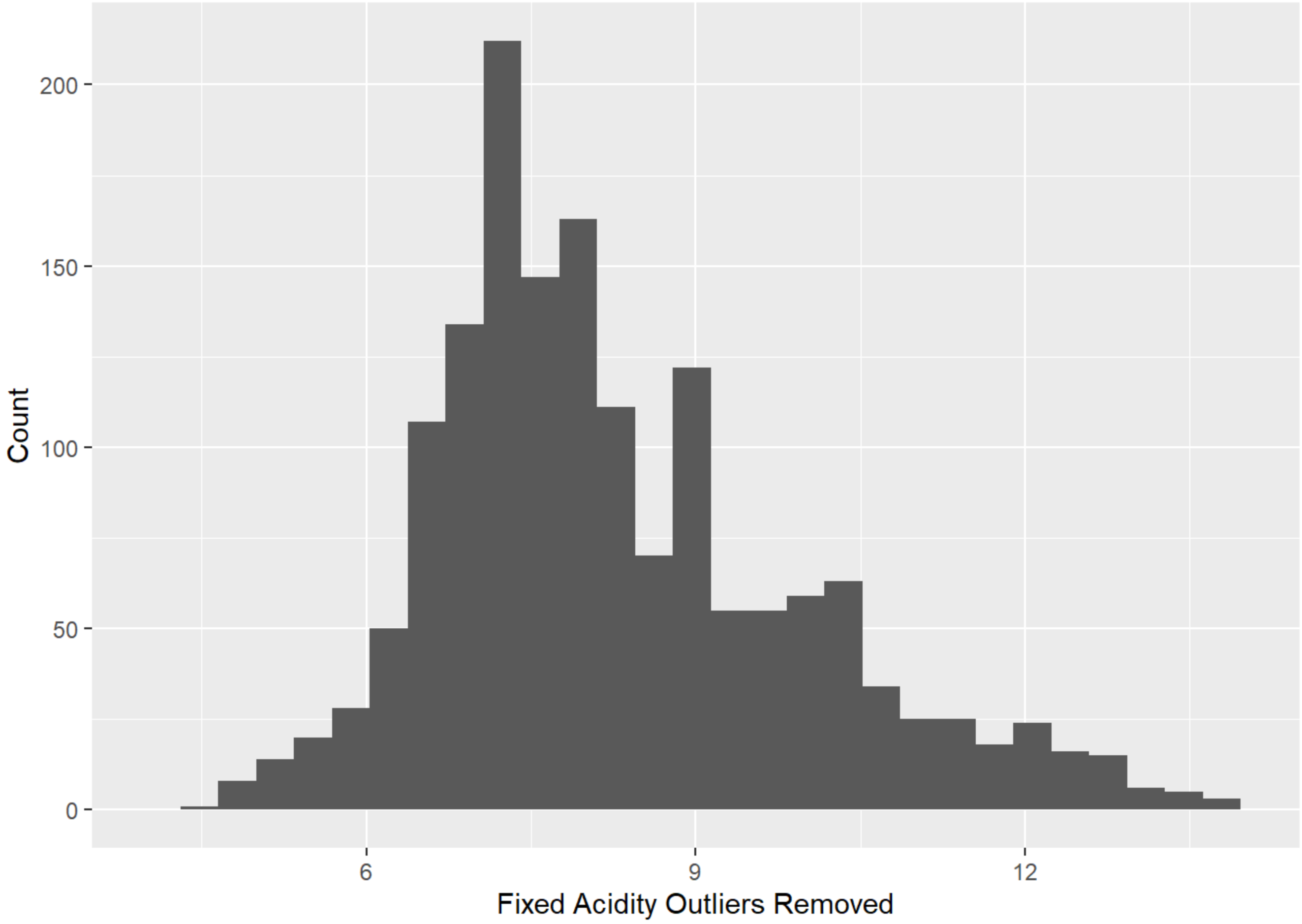


Distribution looks better using log10 transformation but it is long-tailed.

Quality Table: Fixed Acidity > 14

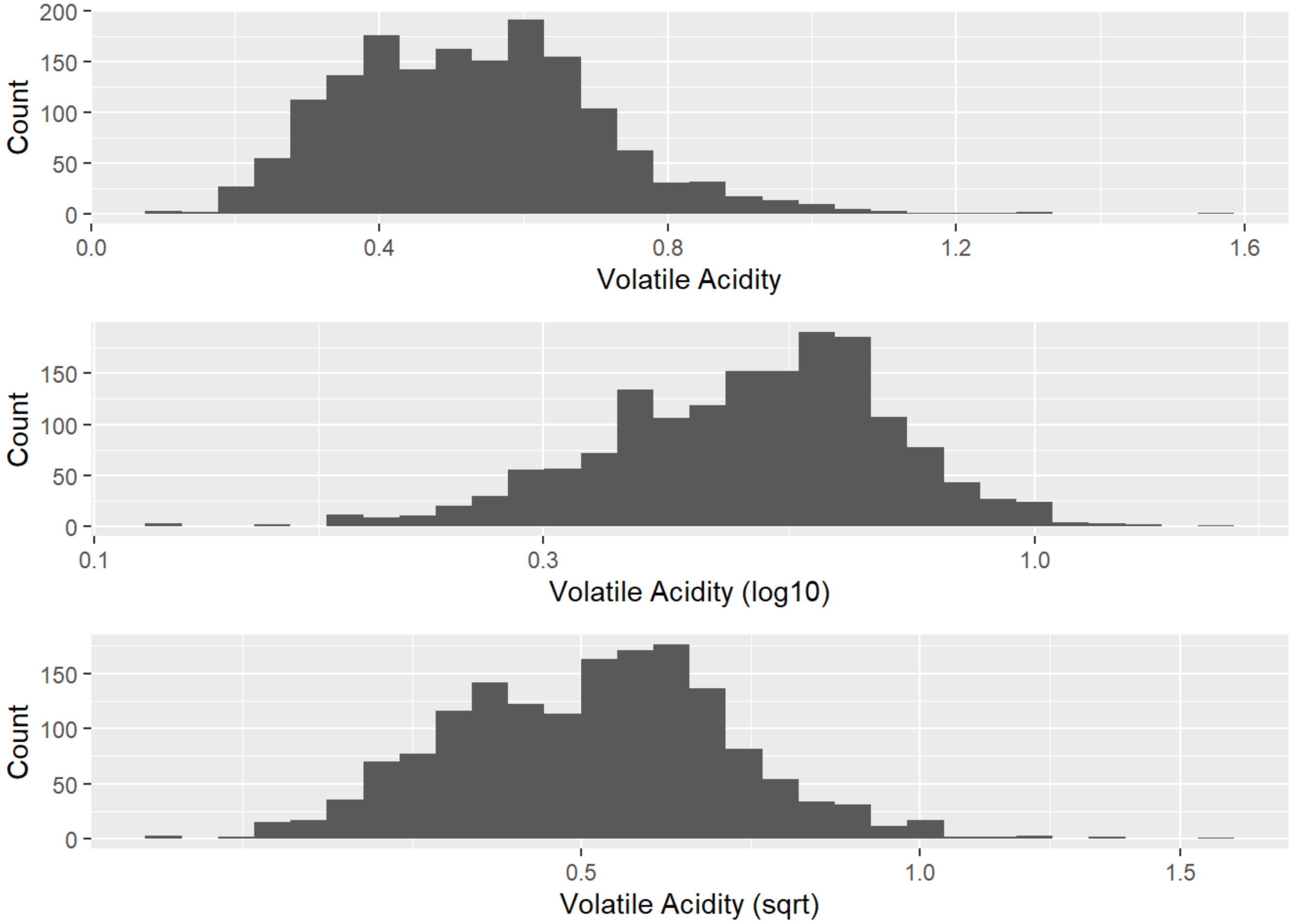
```
##
##     FALSE TRUE
##   3    10    0
##   4    53    0
##   5   677    4
##   6   637    1
##   7   196    3
##   8    18    0
```

A few outliers can be seen in the table above for Fixed acidity > 14.



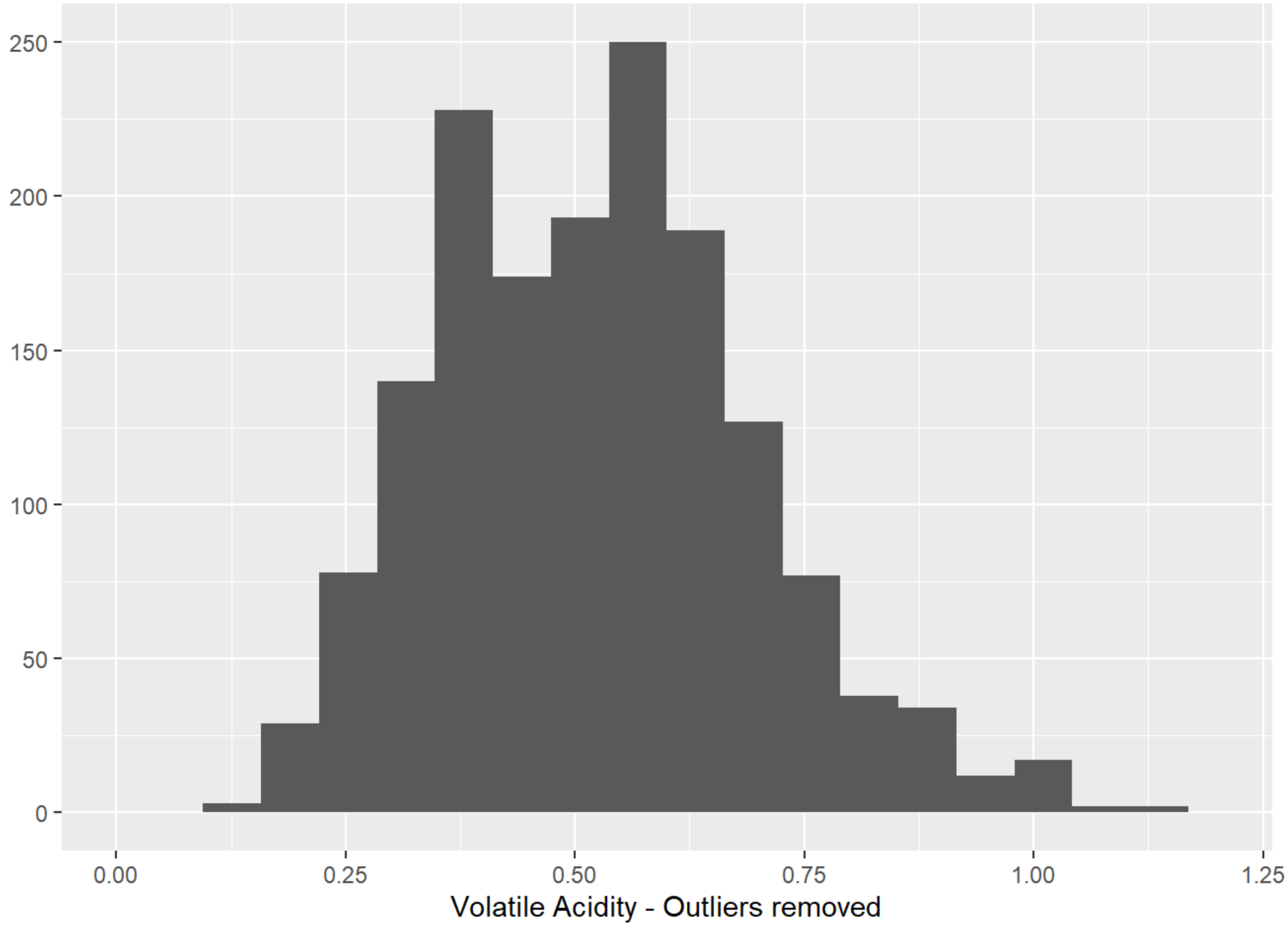
Distribution looks best using log10 transformation with outliers (> 14) removed.

**Volatile Acidity**



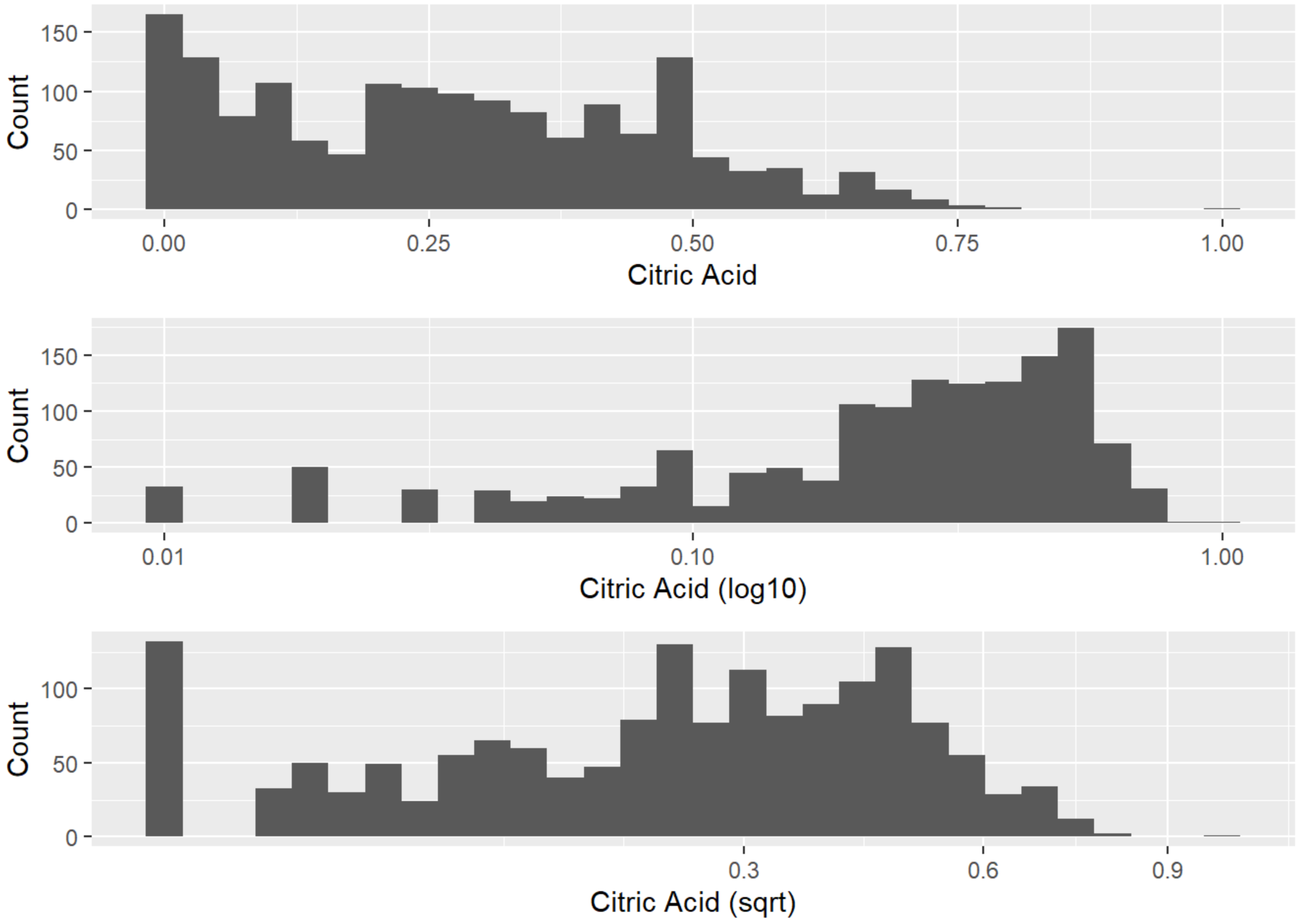
Quality Table: Volatile Acidity > 1.2

```
## 
##     FALSE TRUE
##   3     9    1
##   4    53    0
##   5   678    3
##   6   638    0
##   7   199    0
##   8    18    0
```



The best option seemed to be to used a sqrt transformation and remove 4 outliers (>1.2)

**Citric Acid**

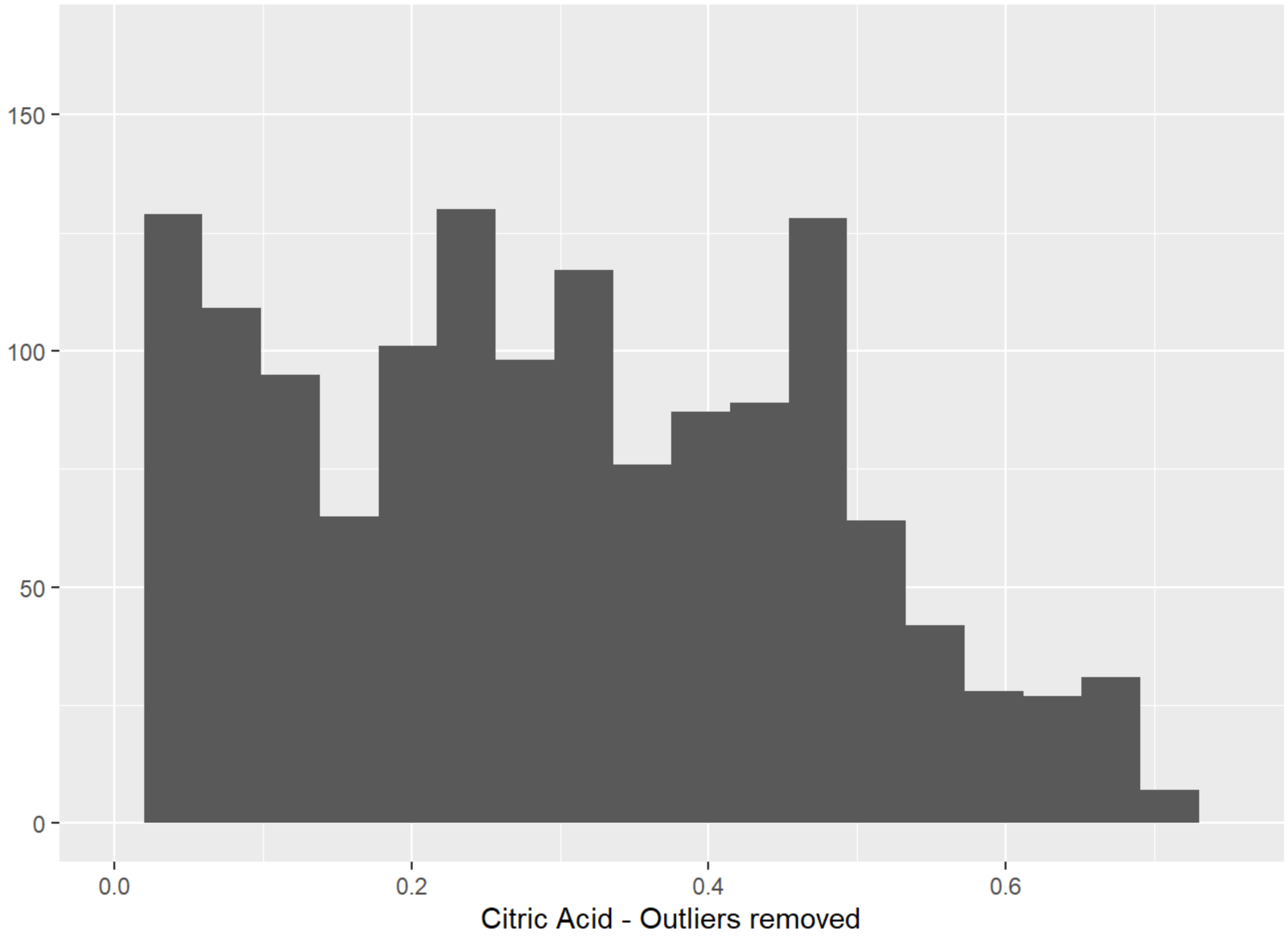


Quality Table: Citric Acid == 0

```
##
##     FALSE TRUE
##   3     7    3
##   4    43   10
##   5   624   57
##   6   584   54
##   7   191    8
##   8    18    0
```

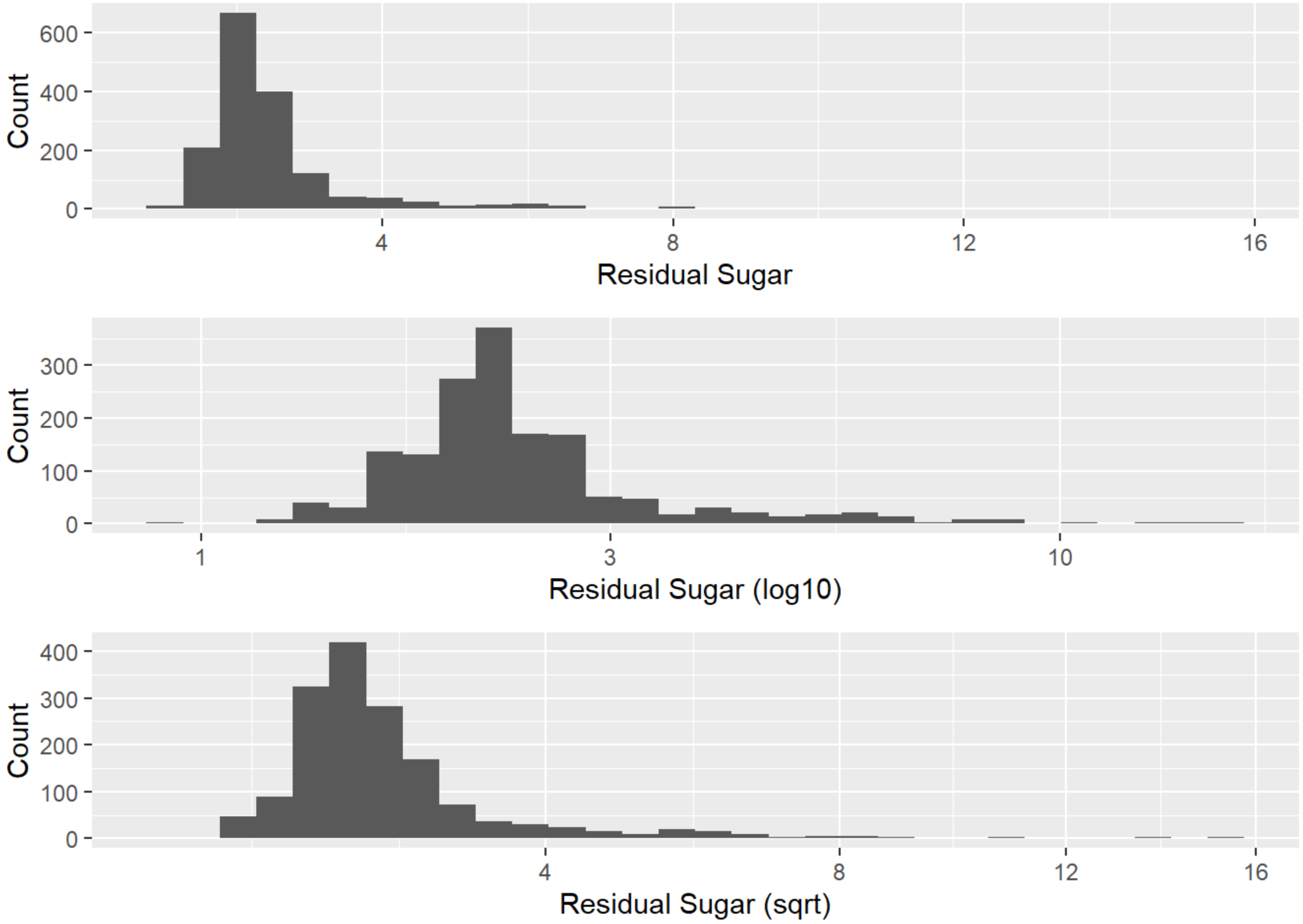
Quality Table: Citric Acid > 0.75

```
##
##     FALSE TRUE
##   3    10    0
##   4    52    1
##   5   679    2
##   6   637    1
##   7   197    2
##   8    18    0
```



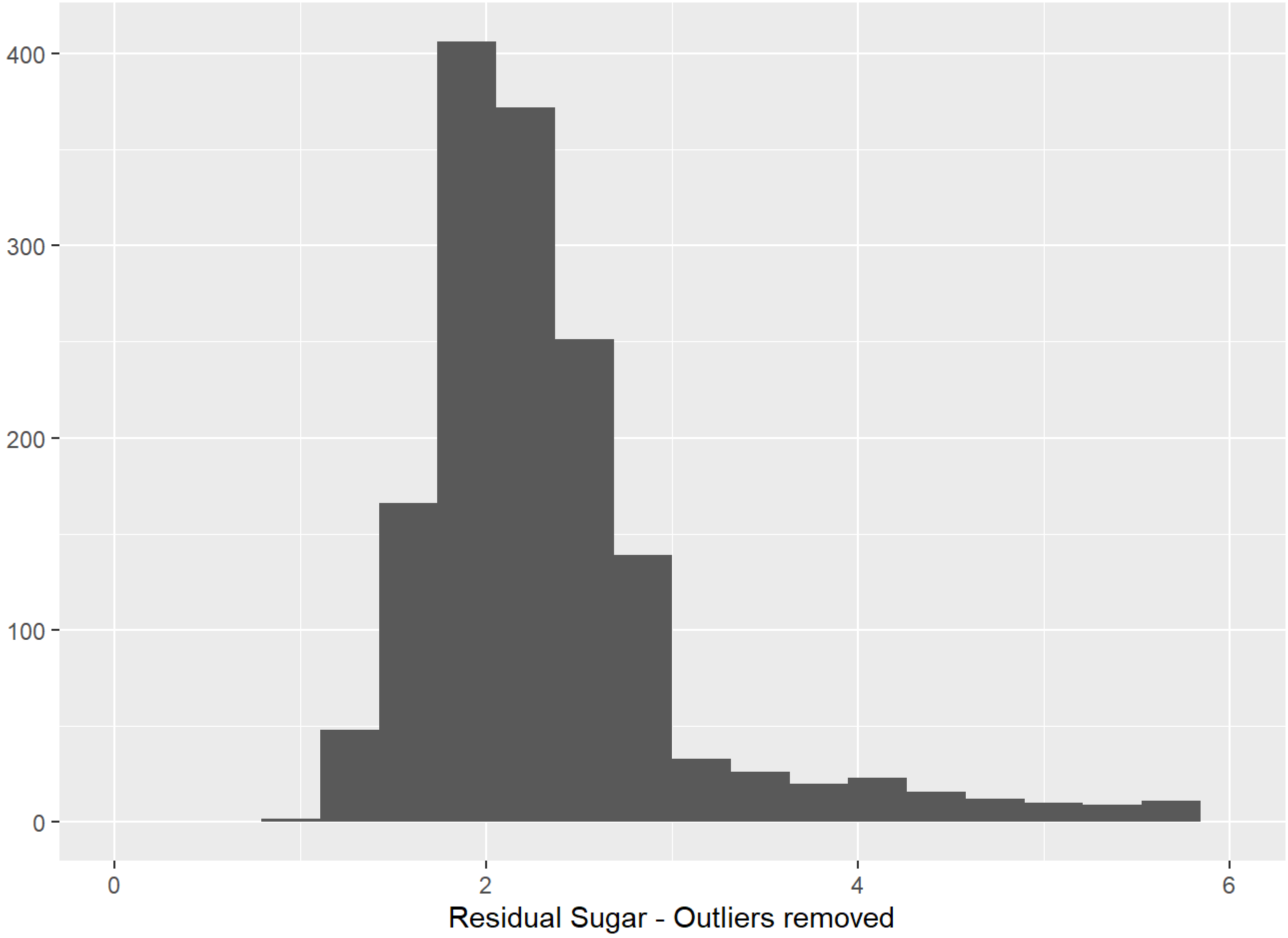
There are 132 0 values which seem to be pretty evenly distributed on quality. None of the transformations seem suitable. The best that can be done is to remove 6 outliers which spreads the distribution out.

**Residual Sugar**



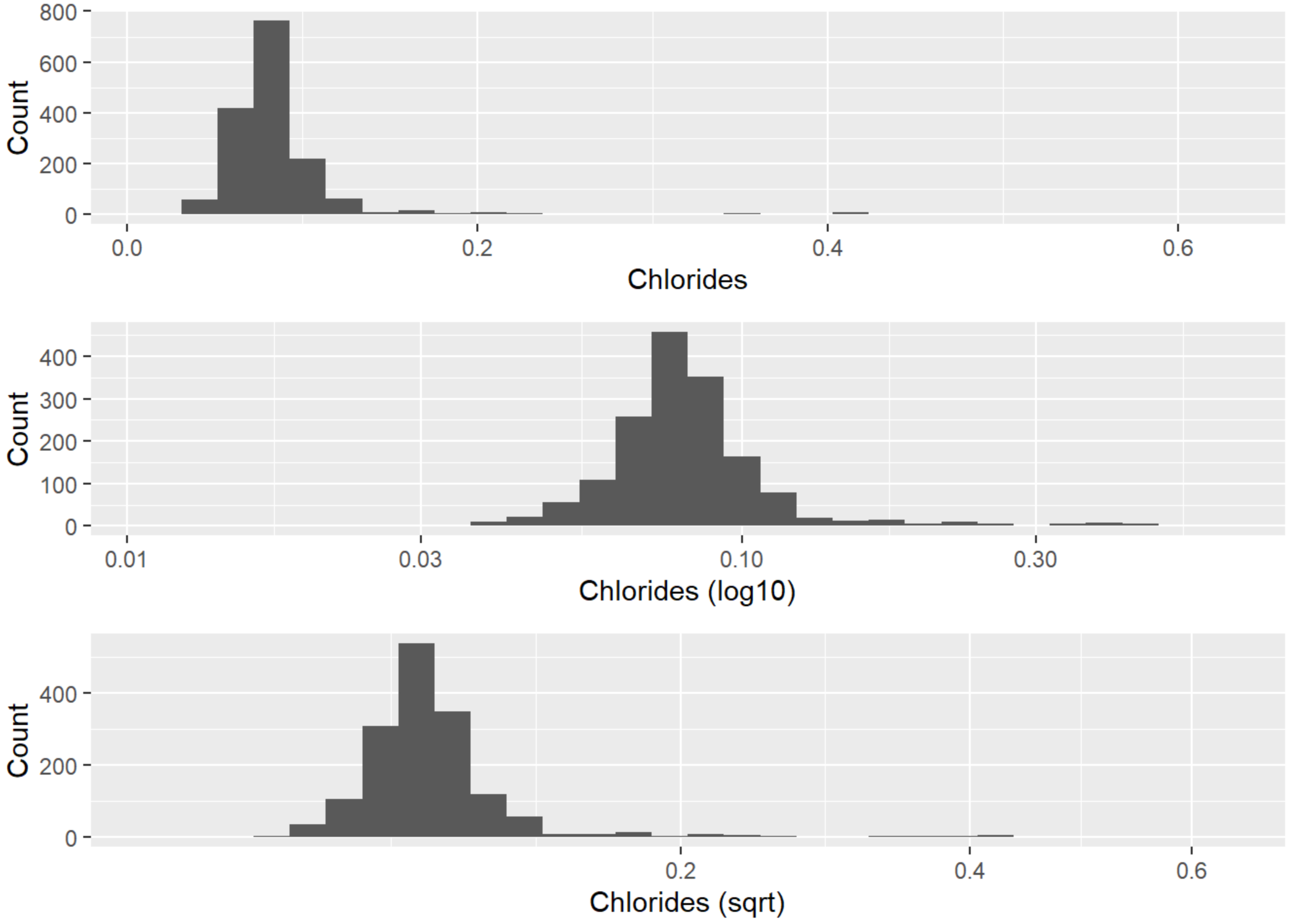
Quality Table: Residual Sugar > 6

```
## 
##     FALSE TRUE
##   3    10    0
##   4    51    2
##   5   658   23
##   6   624   14
##   7   191    8
##   8    17    1
```



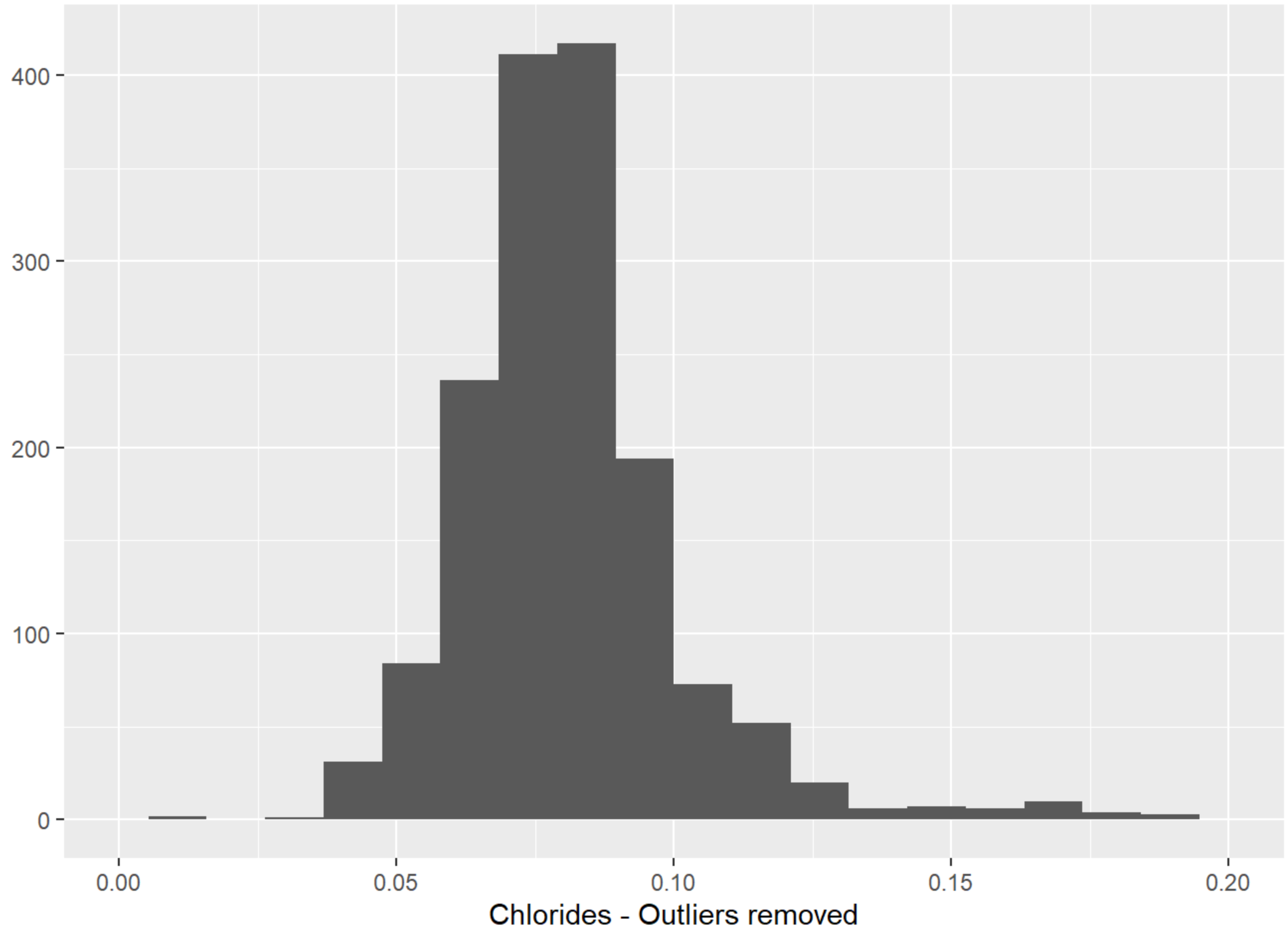
This time I went for a log10 transformation. Removing outliers > 6 (see graph) did had a significant impact on the distribution but there were 46 records which seemed to be too many.

### Chlorides



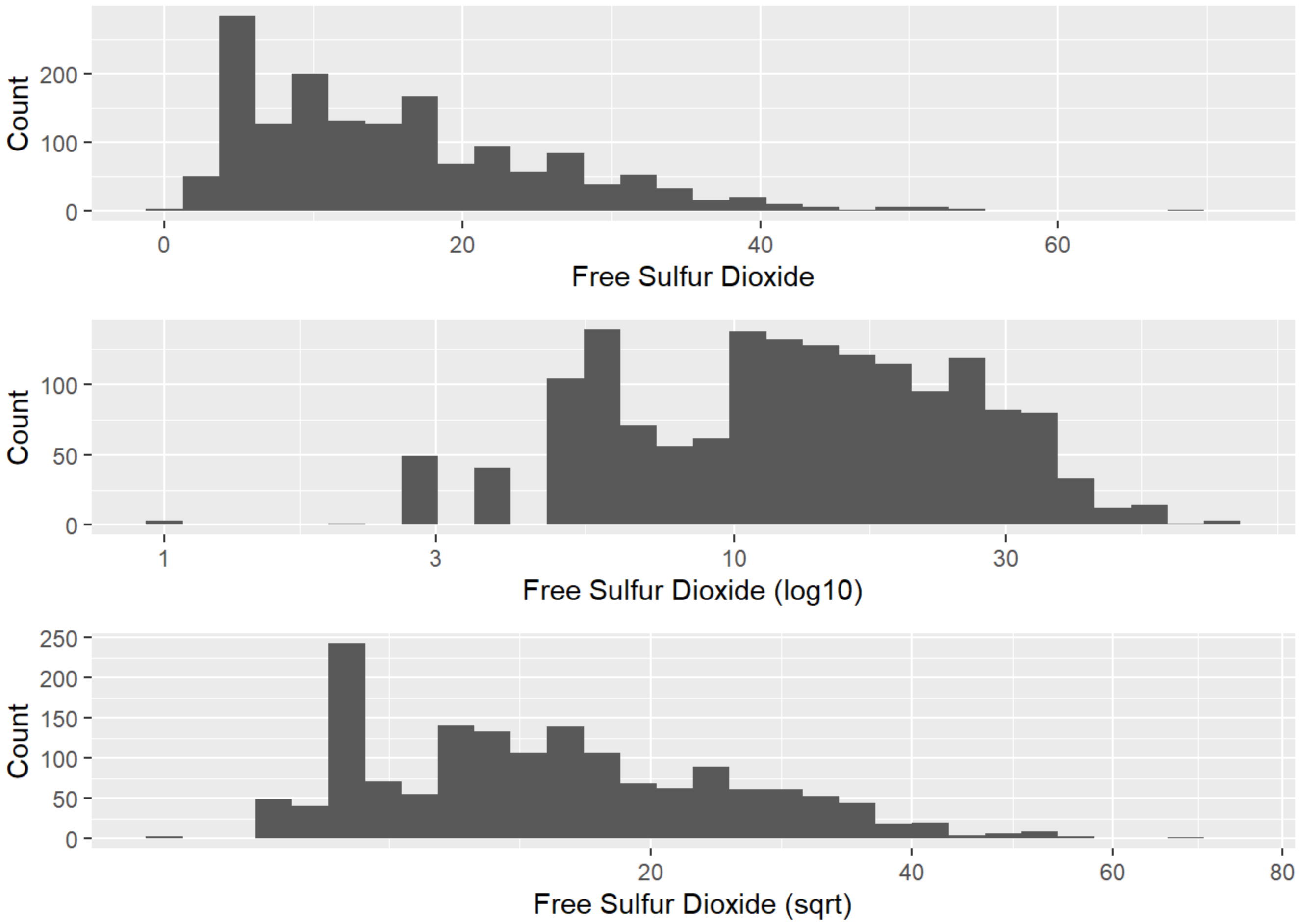
Quality Table: Chlorides > .2

```
##
##     TRUE
##   3   10
##   4   53
##   5  681
##   6  638
##   7  199
##   8   18
```



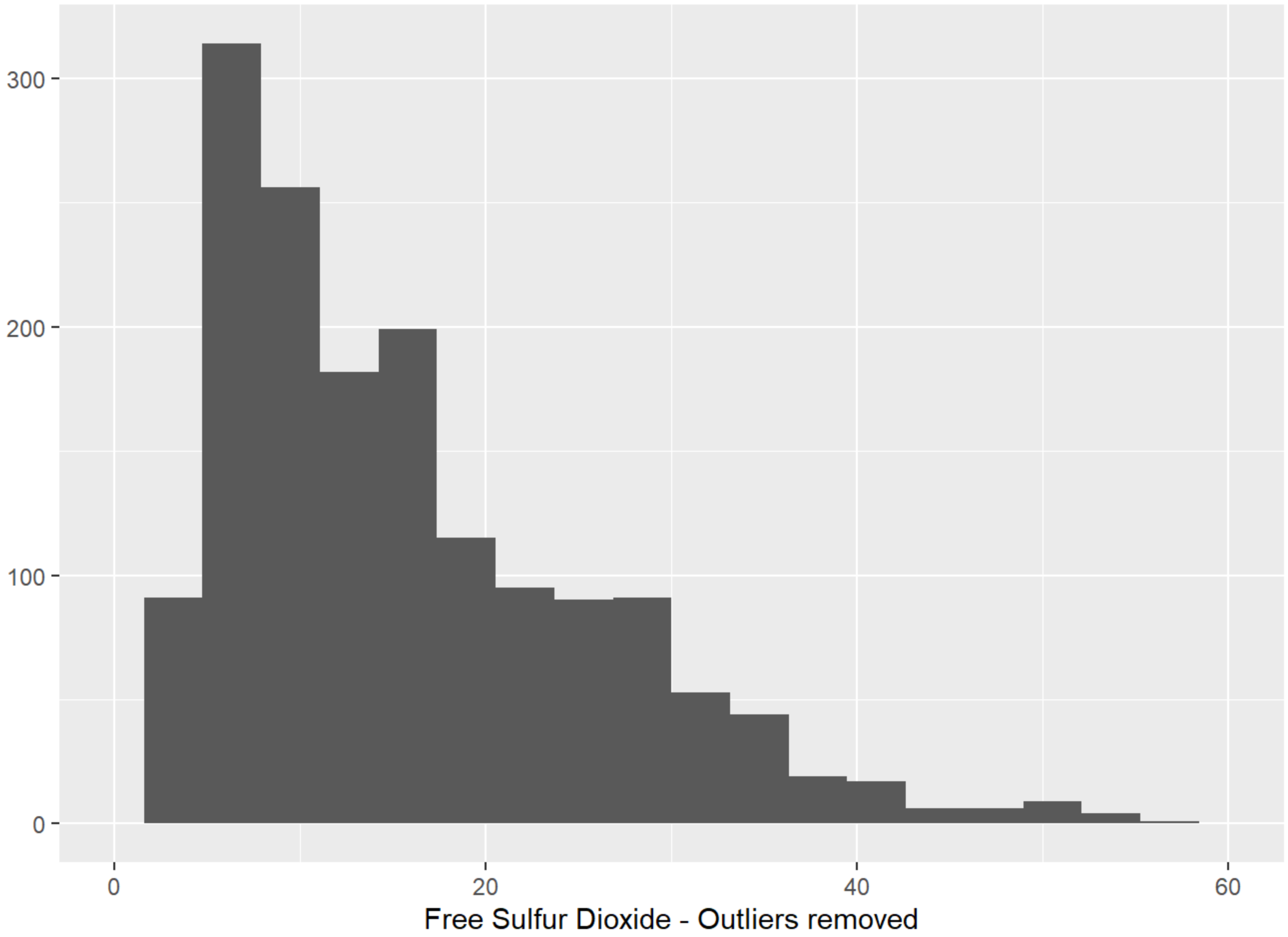
For chlorides the log10 transformation produced a central relatively evenly spread curve.

**Free Sulfur Dioxide**



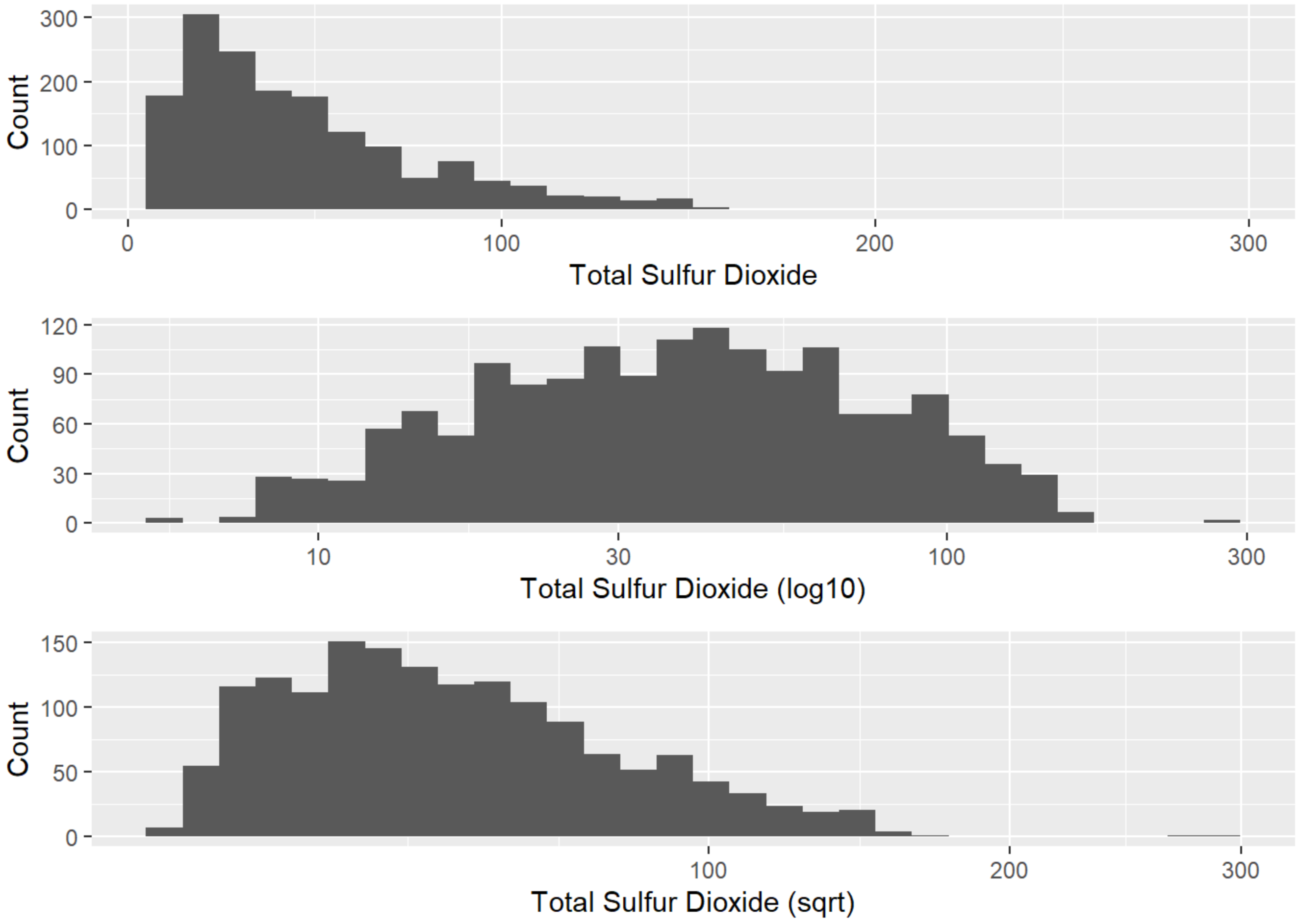
Quality Table: Free Sulfur Dioxide > 60

```
##
##      FALSE TRUE
##   3    10    0
##   4    53    0
##   5   678    3
##   6   637    1
##   7   199    0
##   8    18    0
```



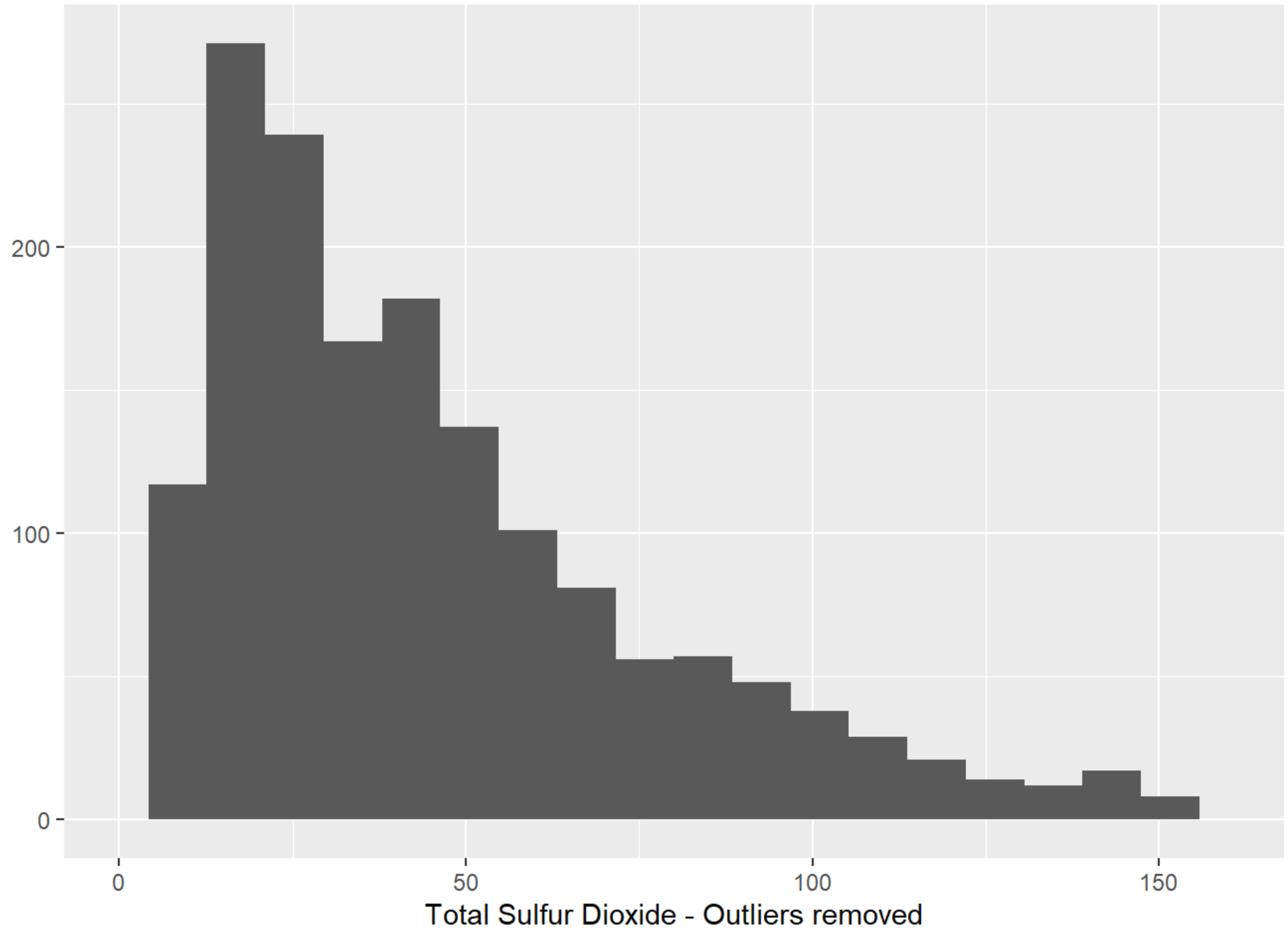
The sqrt transformation was used as this normalises better than log 10. Removing 4 outliers doesn’t alter the distribution significantly.

**Total Sulfur Dioxide**



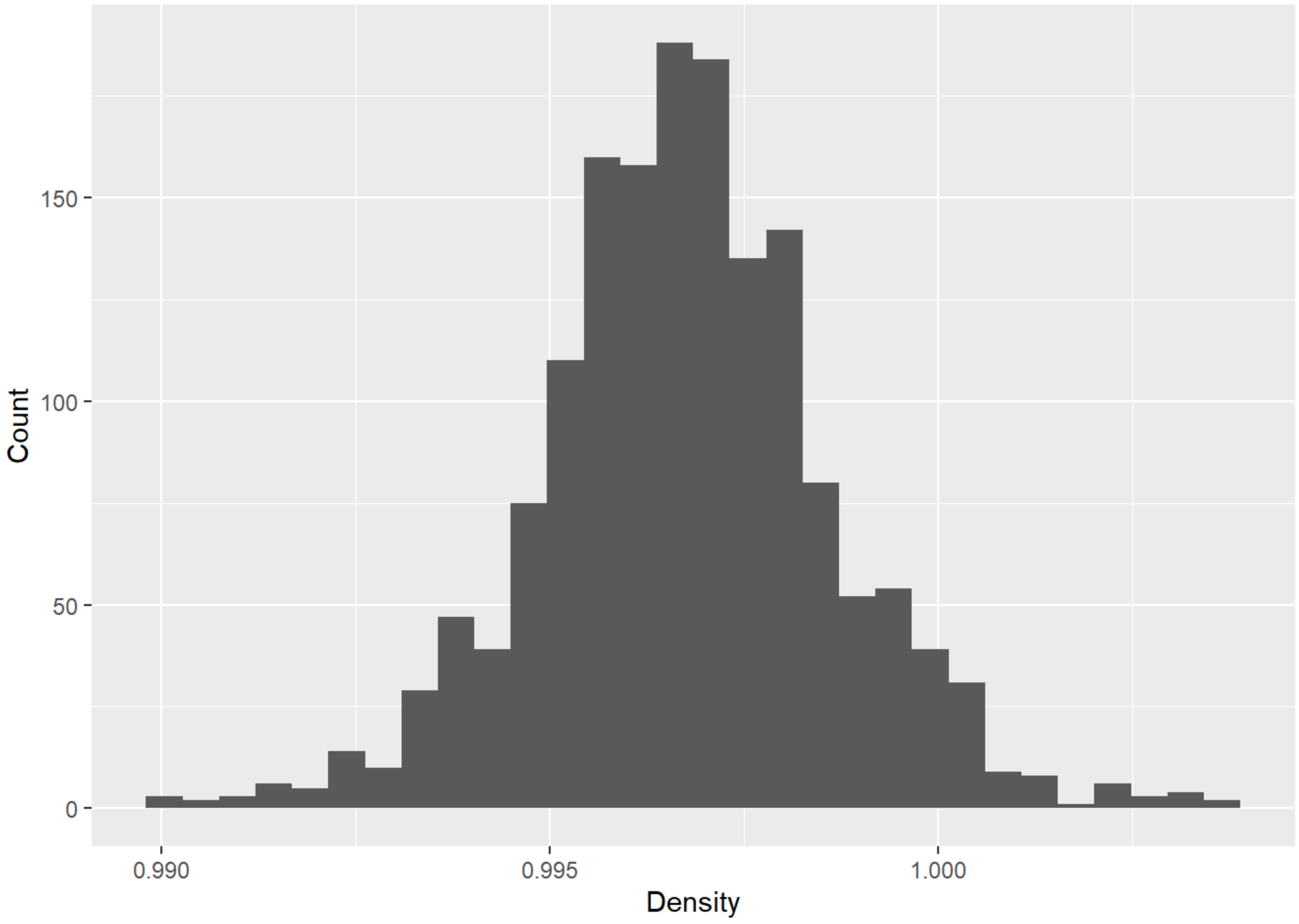
Quality Table: Total Sulfur Dioxide > 160

```
## 
##     FALSE TRUE
##   3    10    0
##   4    53    0
##   5   681    0
##   6   637    1
##   7   197    2
##   8    18    0
```



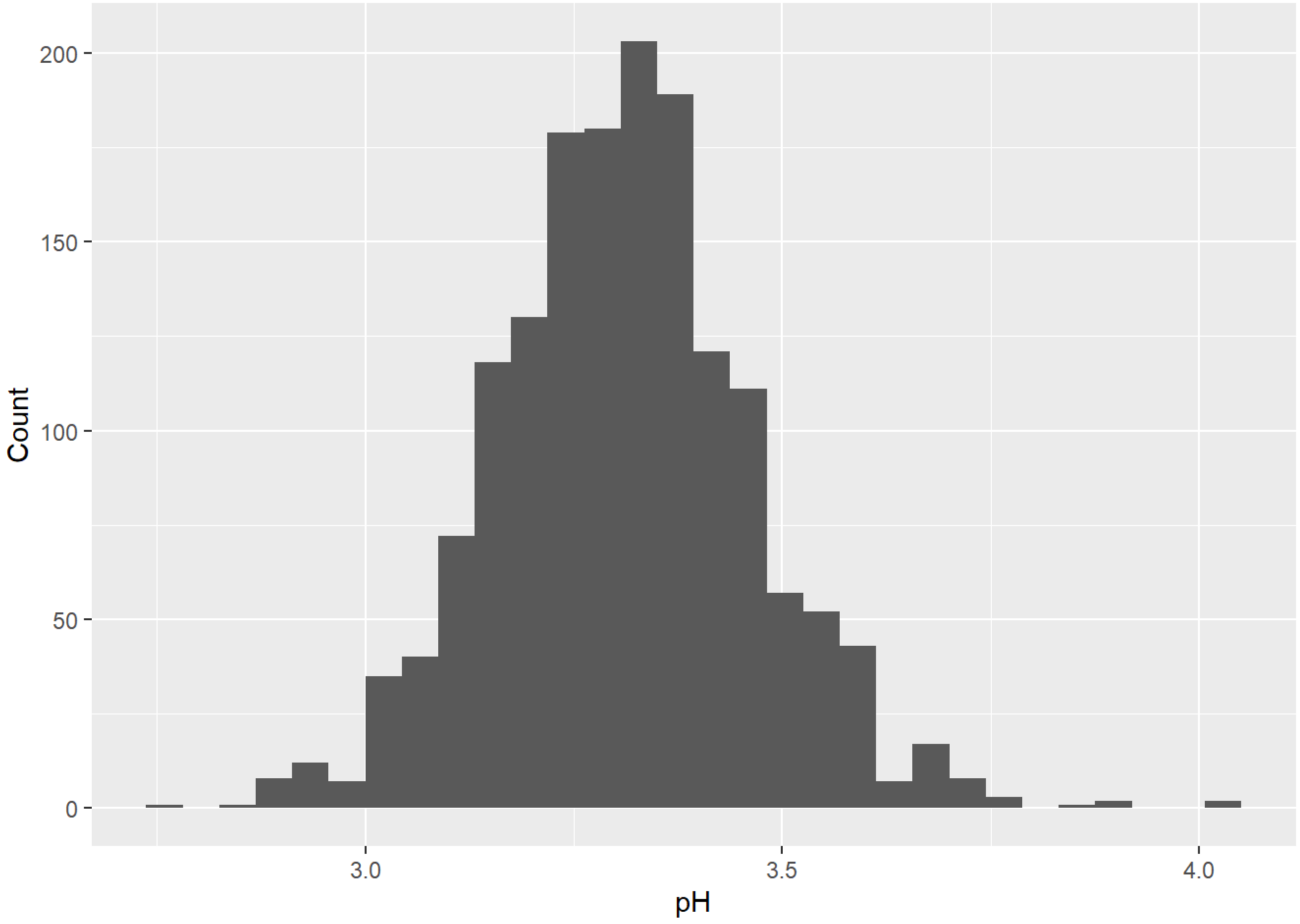
Removed 3 outliers for total sulfur dioxide (>160). The distribution is still heavily positively skewed but neither transformation option seems to have a significant effect.

**Density**



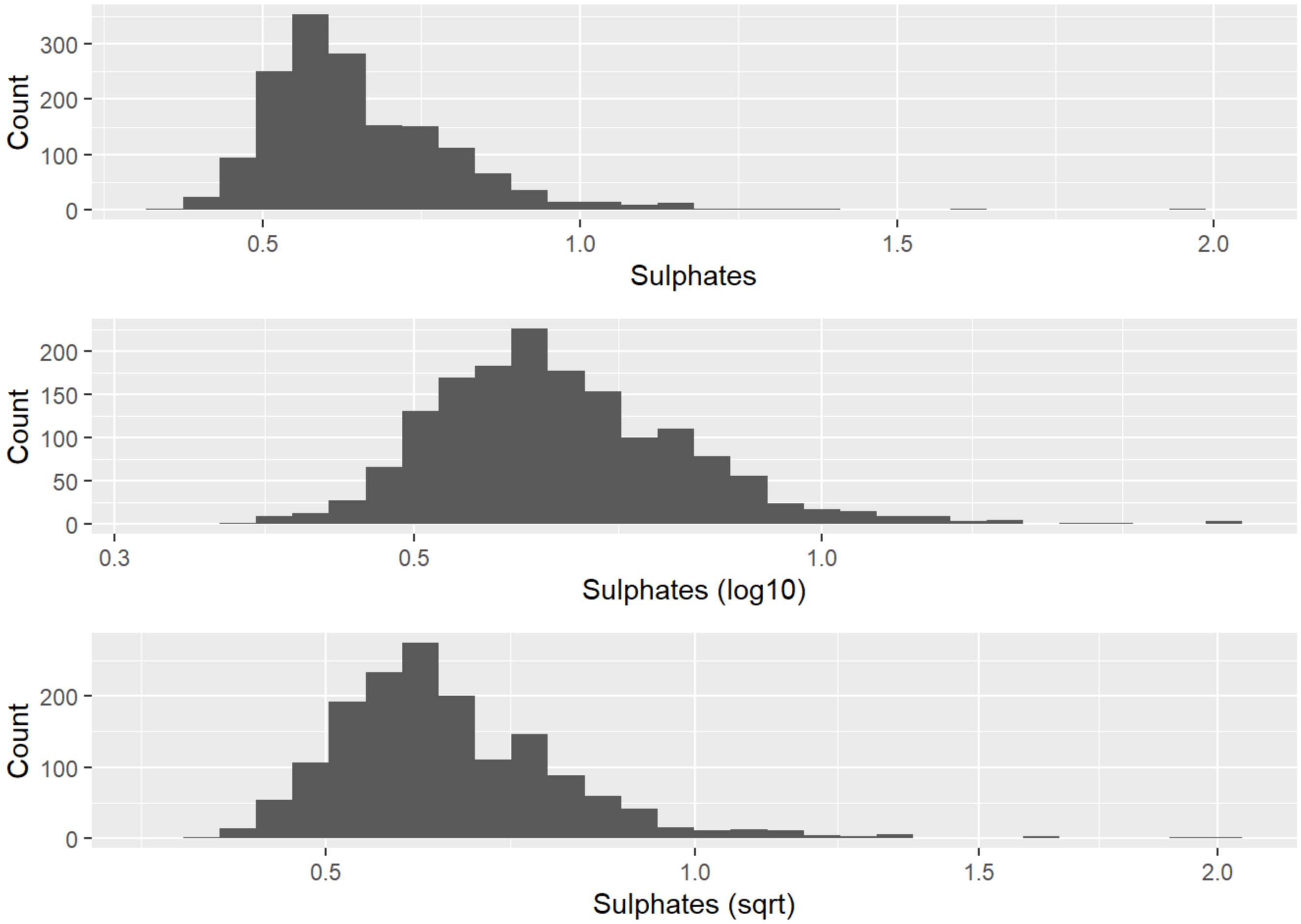
No change - normal already

**pH**



No change - normal already

### Sulphates

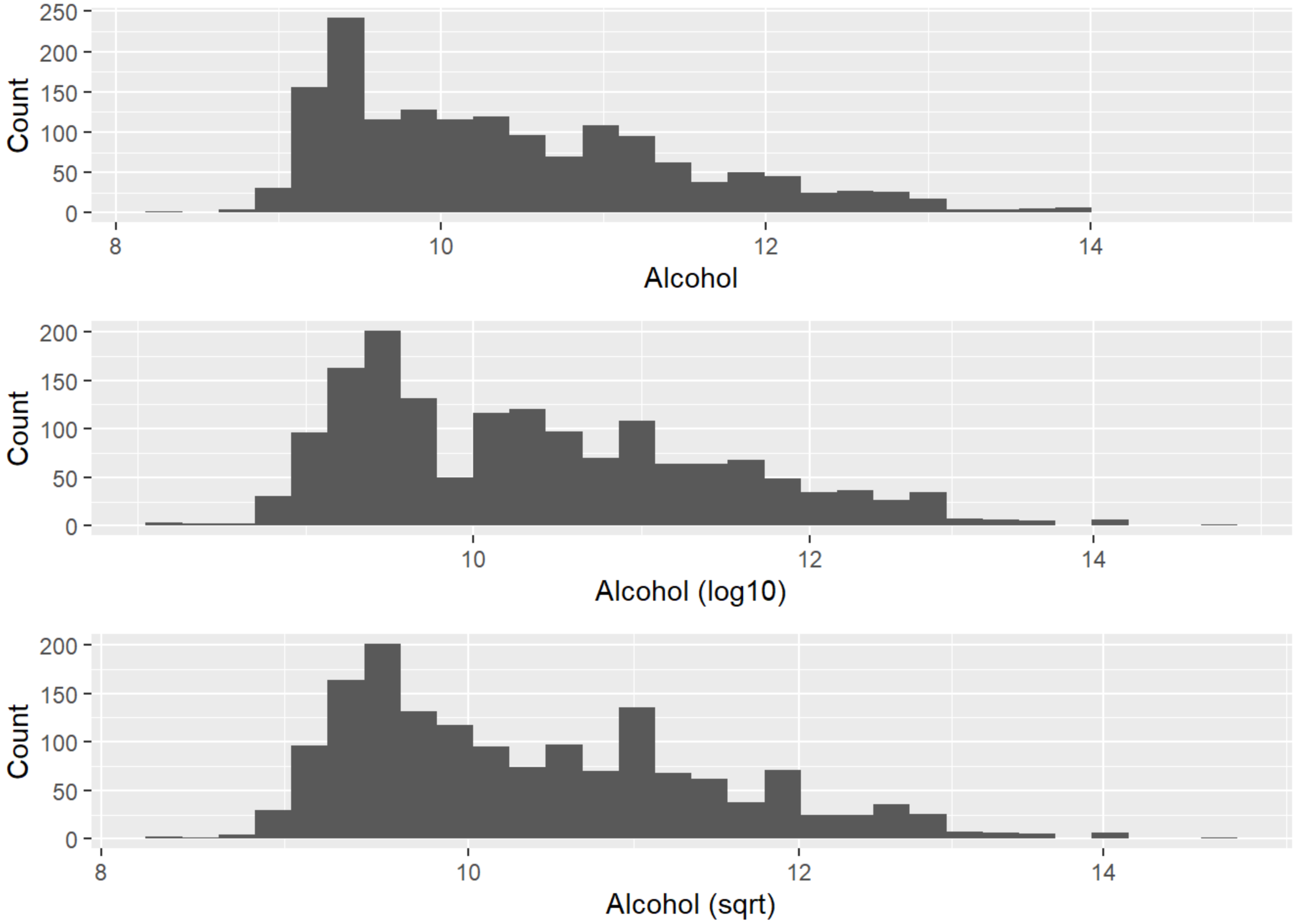


Quality Table: Sulphates > 1.5

```
## 
##     FALSE TRUE
##   3    10    0
##   4    52    1
##   5   677    4
##   6   635    3
##   7   199    0
##   8    18    0
```

Transformed the x-axis into log10 scale can make it more normally distributed. The outliers have low quality values (all under 6) which may be signficant.

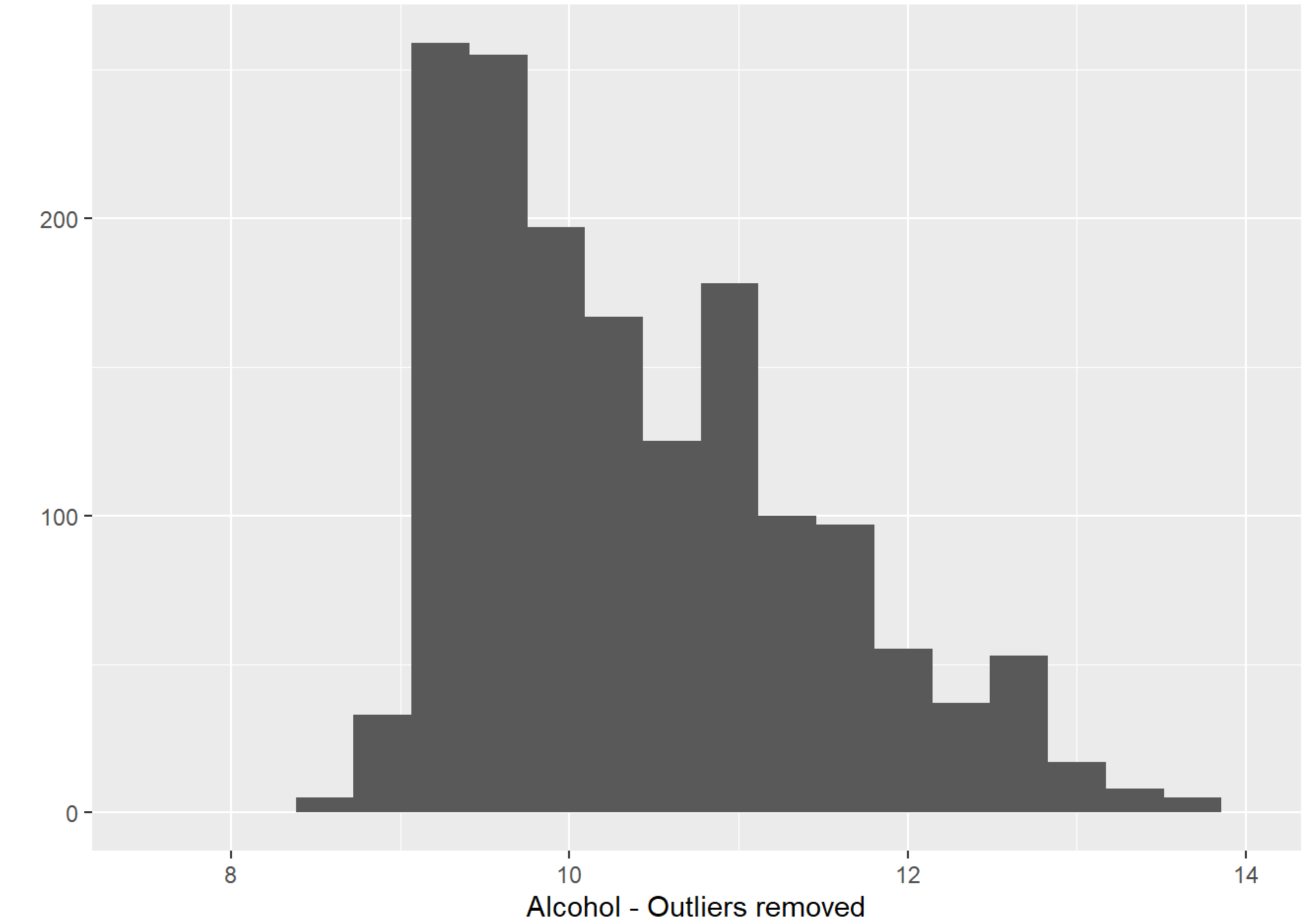
### Alcohol



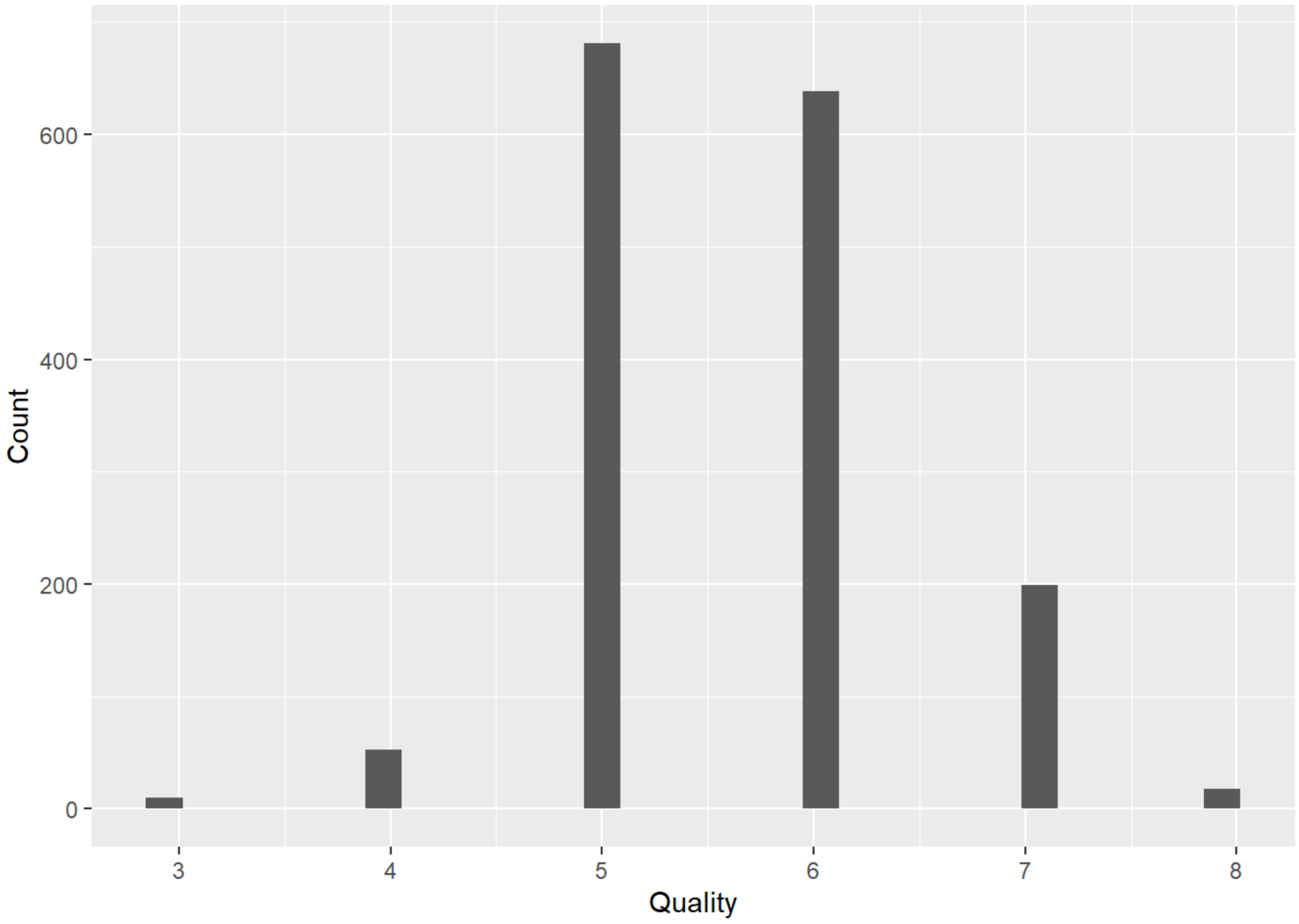
Quality Table:

Alcohol > 14

```
##
##     FALSE TRUE
##   3    10    0
##   4    53    0
##   5   680    1
##   6   638    0
##   7   199    0
##   8    18    0
```



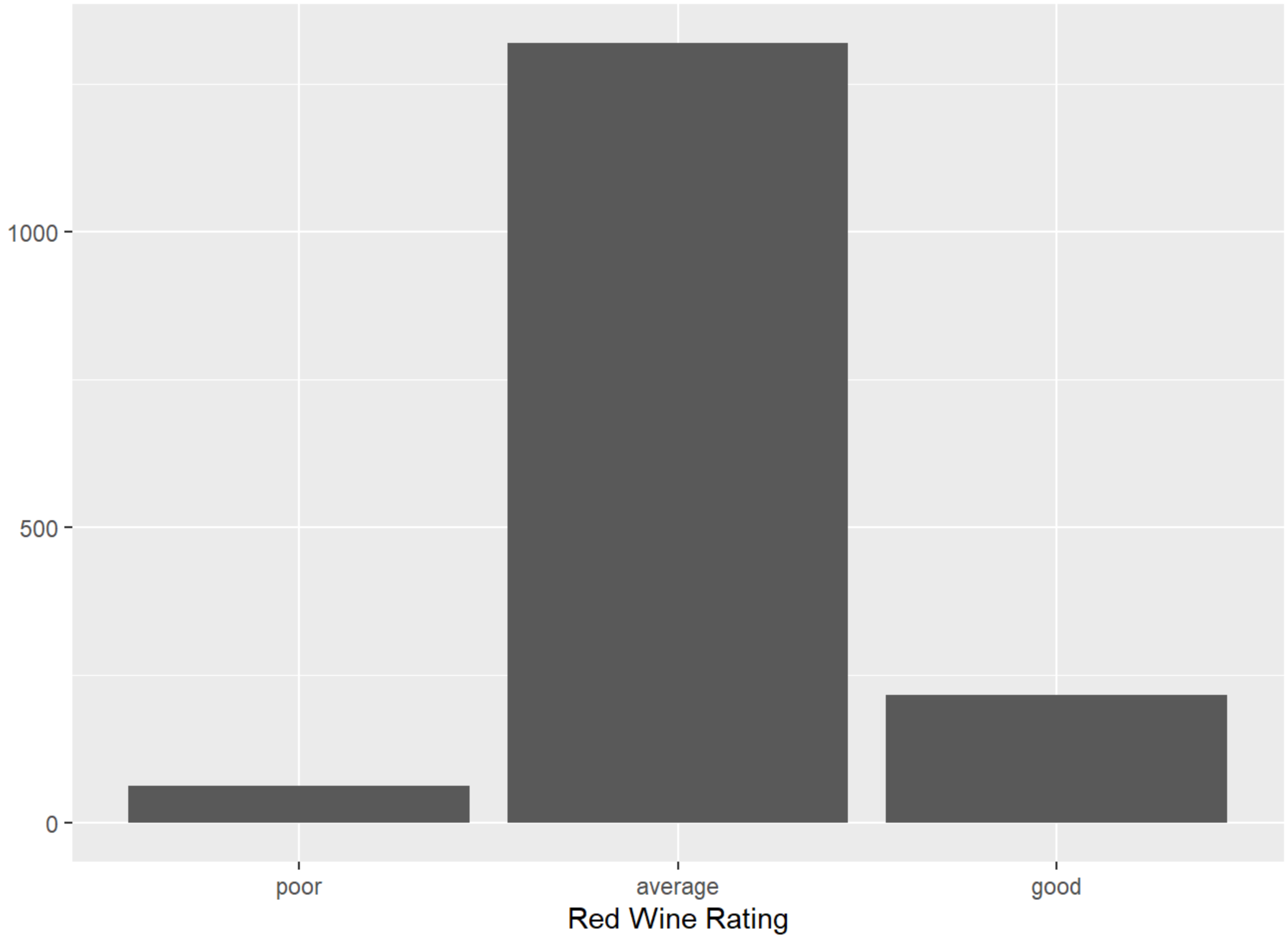
**Wine Quality**



Looking at the first plot of wine quality, it roughly has a normal distribution with most rating being in 5 and 6. Will create another variable called rating with following categories.

- 0-4 : poor
- 5-6: average
- 7-10 : good

```
##    poor average    good 
##      63    1319     217 
```



# Univariate Analysis

**What is the structure of your dataset?**

There are 12 attributes in the dataset. 11 of them (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol) are input attributes based on physicochemical tests. The other attribute (quality) is the output attribute based on sensory data. Each row corresponds to one particular wine with total 1599 different red wines in the data set.

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

The main feature of interest is the output attribute quality. I want to find out which of the 11 input attributes contribute to a high quality value.

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

The 11 input attributes are equally likely to contribute to the quality value at this point. The bivariate exploration will look more closely at how each of the attributes is distributed with a given quality value.

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

Yes I created a rating variable to rate wines as poor, average or good.

**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?**

Yes there are some distributions that are unusual. I adjusted these plots by taking log10 or sqrt values for the plots where appropriate because more accurate trends can be inferred from bivarite plots.

---

# 3. Bivaiate Exploration

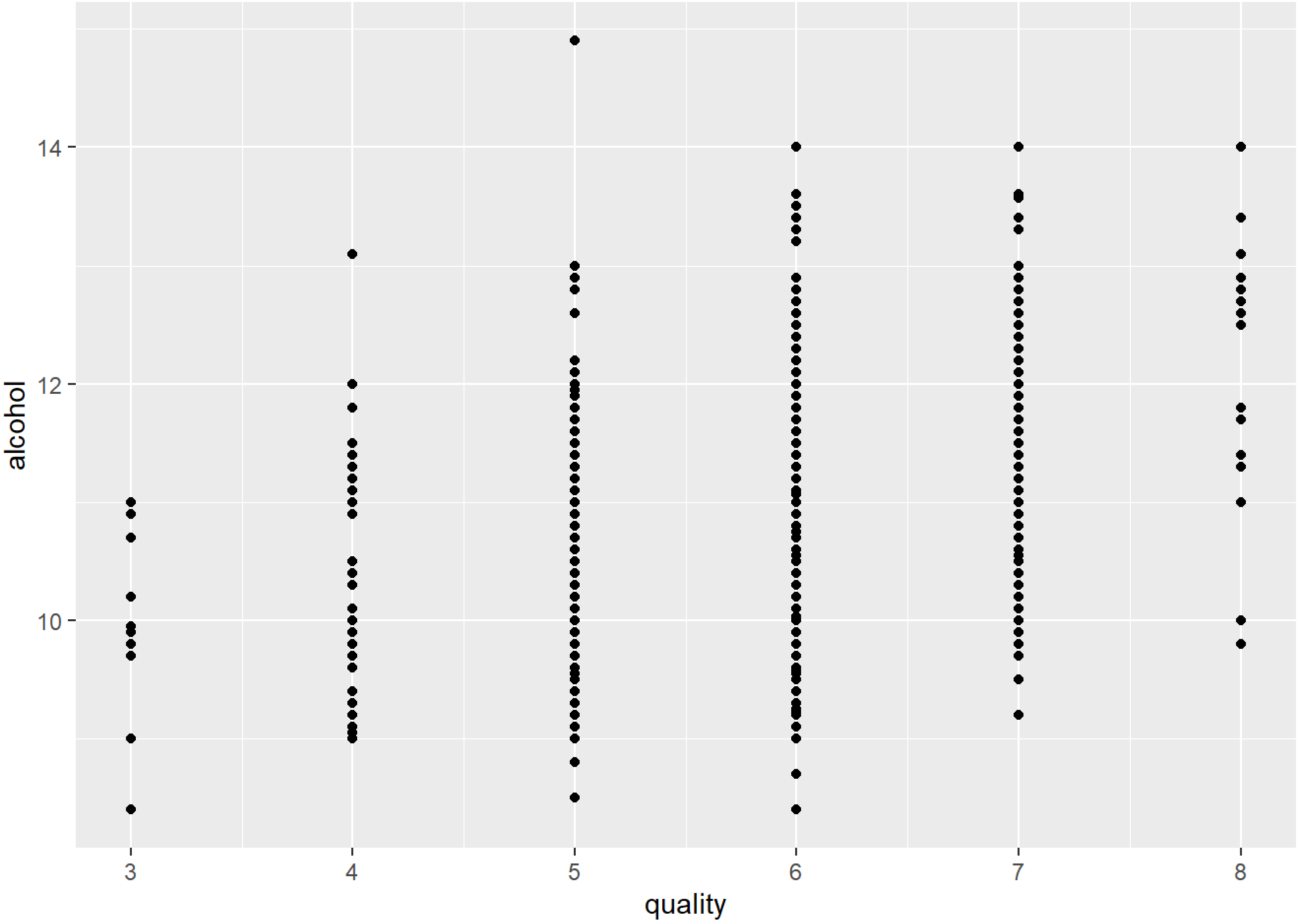
**Correlation of variables**

Correlation of variables against quality is calculated to further explore.

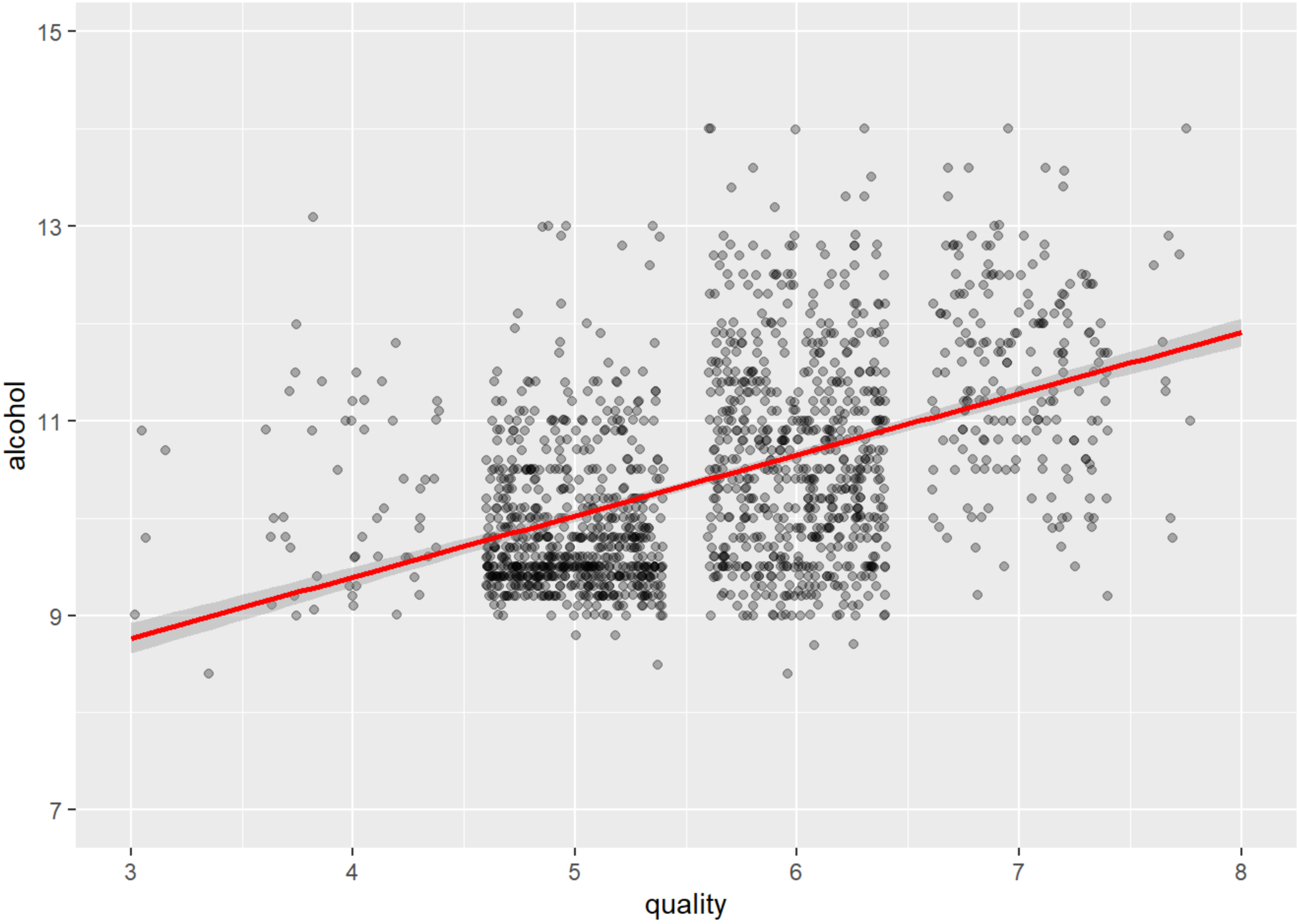
```
##   log10.fixed.acidity sqrt.volatile.acidity      sqrt.citric.acid
##            0.11423756           -0.39394603            0.20668220
##  log10.residual.sugar       log10.chlordies         sqrt.free.S02
##            0.02353331           -0.17613996           -0.05147378
##        sqrt.total.S02               density                    pH
##           -0.18419180           -0.17491923           -0.05773139
##       log10.sulphates          sqrt.alcohol
##            0.30864193            0.47682047
```

**Positive Correlations**

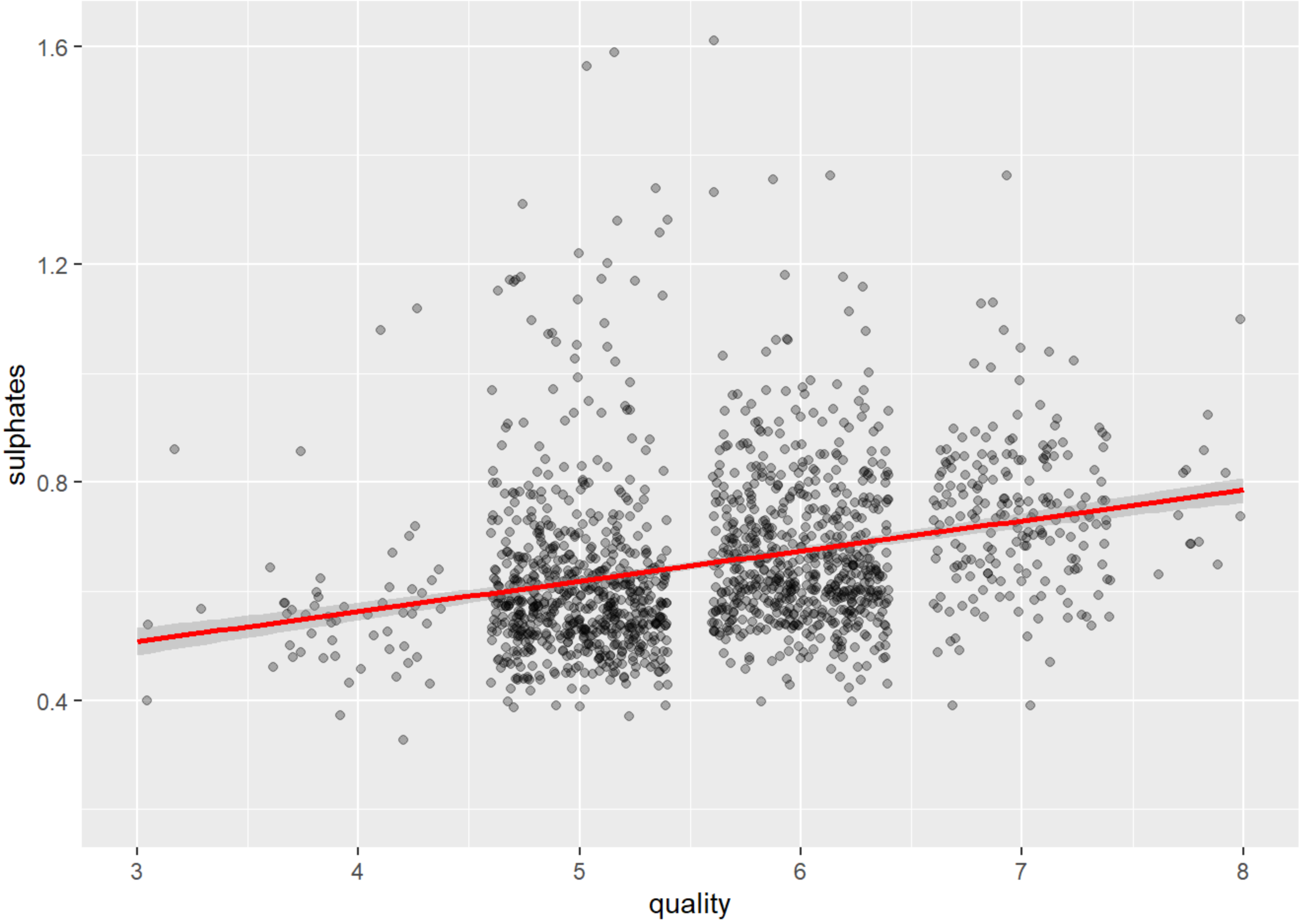
Alcohol Content has biggest correlation value to wine quality, so lets start with a basic scatter plot of the both.



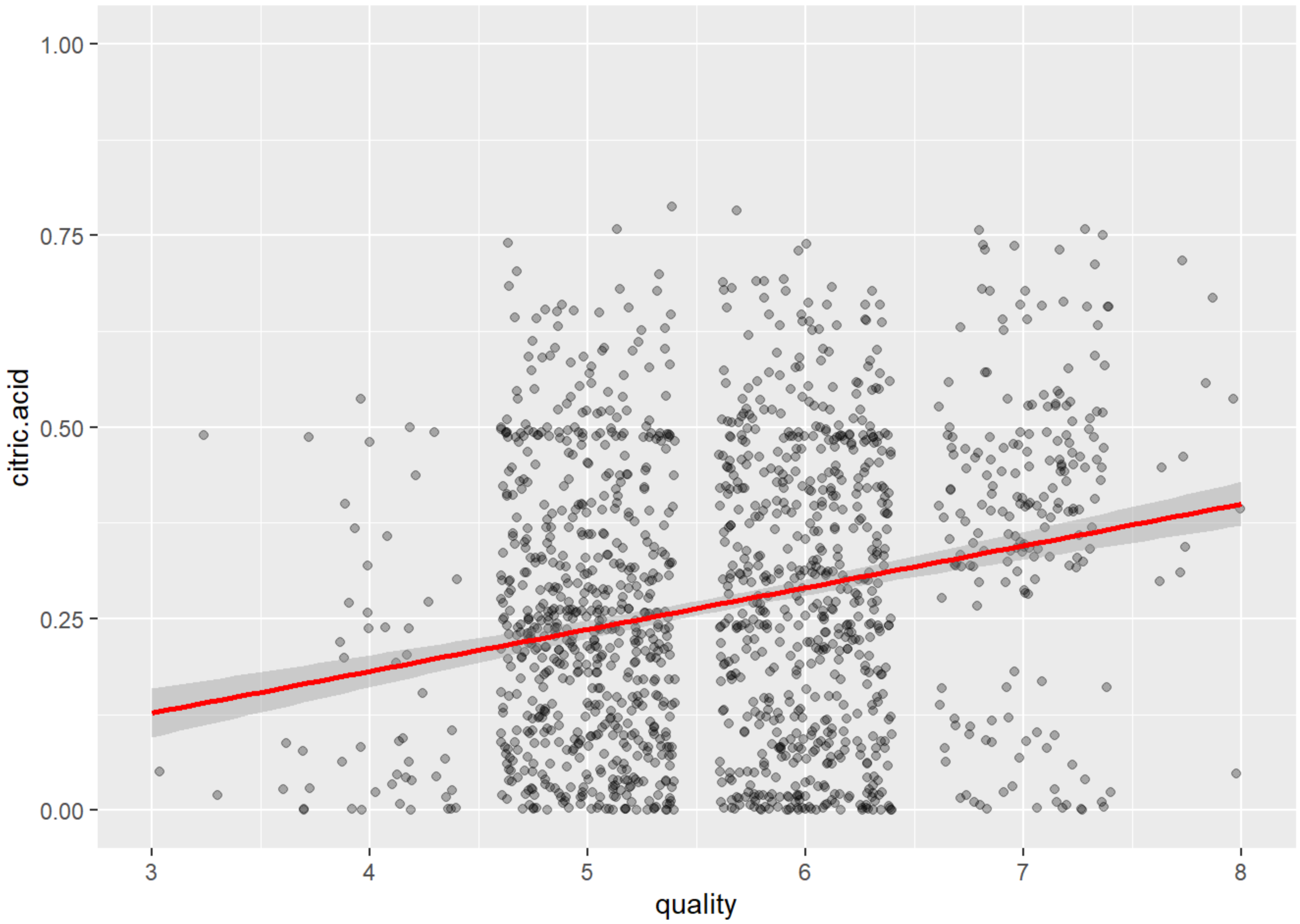
Original plot is over-crowded will add alpha values and jitter with trend line to observe the general trend.



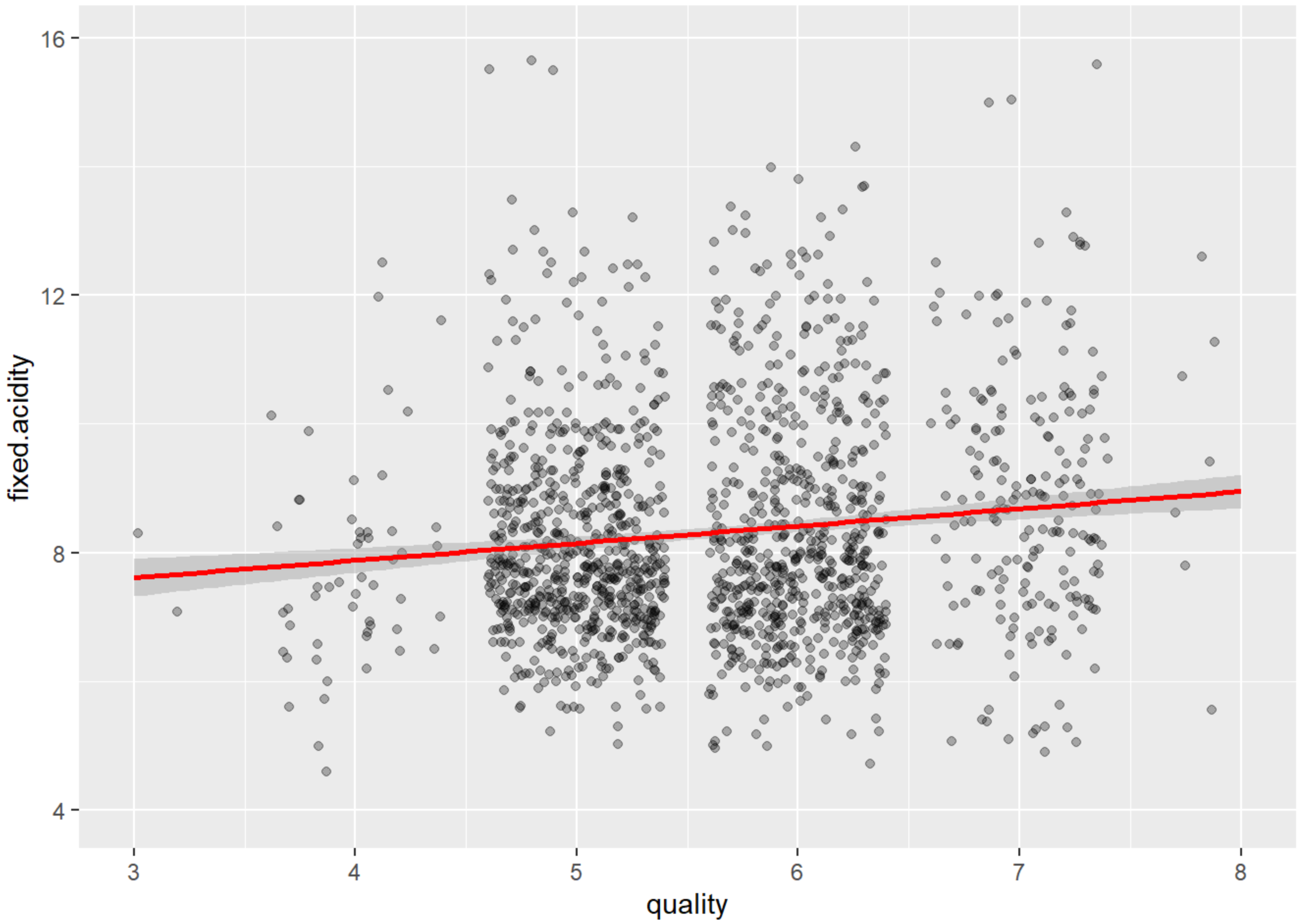
Next will look at sulfates.



Citric Acid.

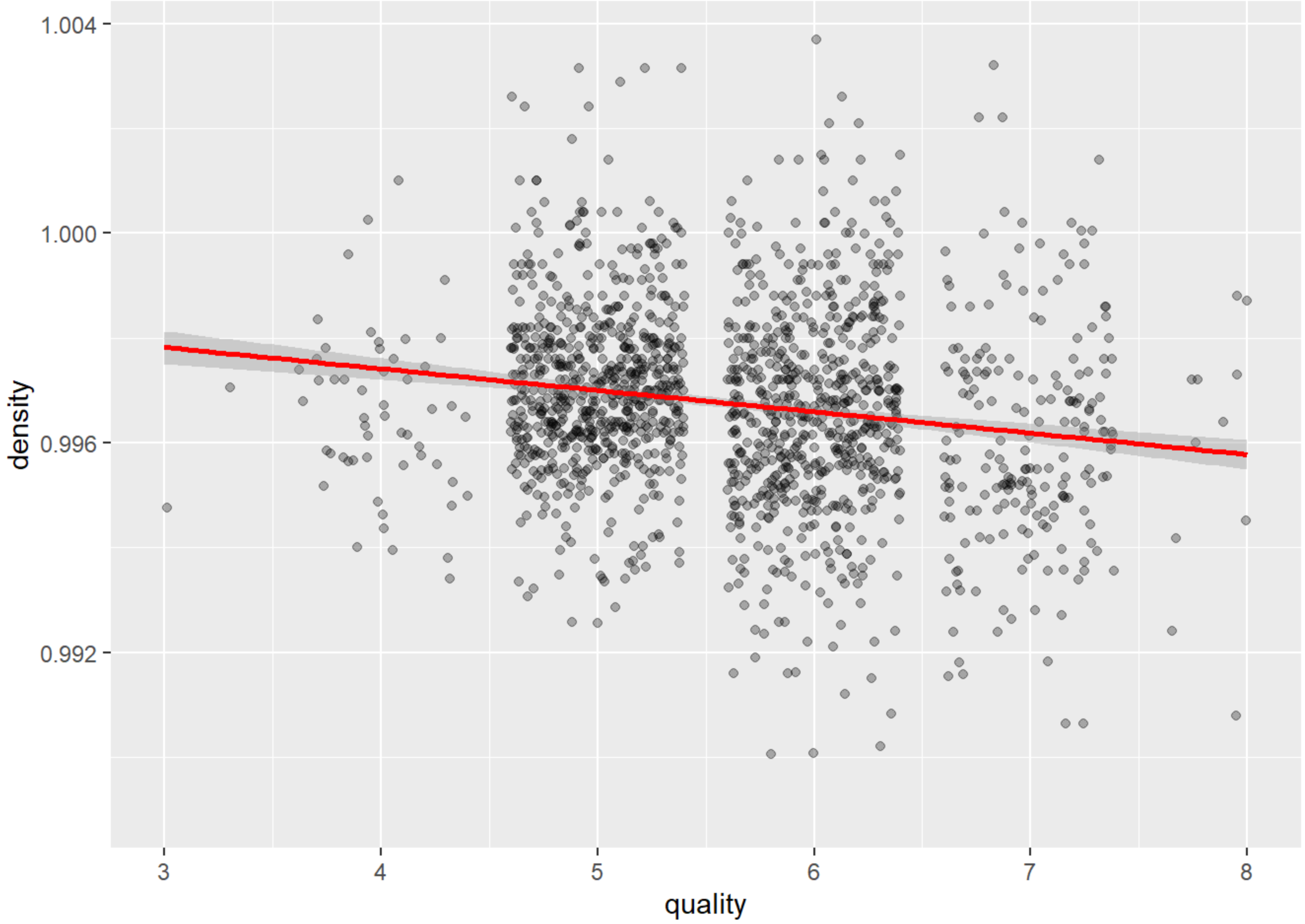


Fixed Acidity.

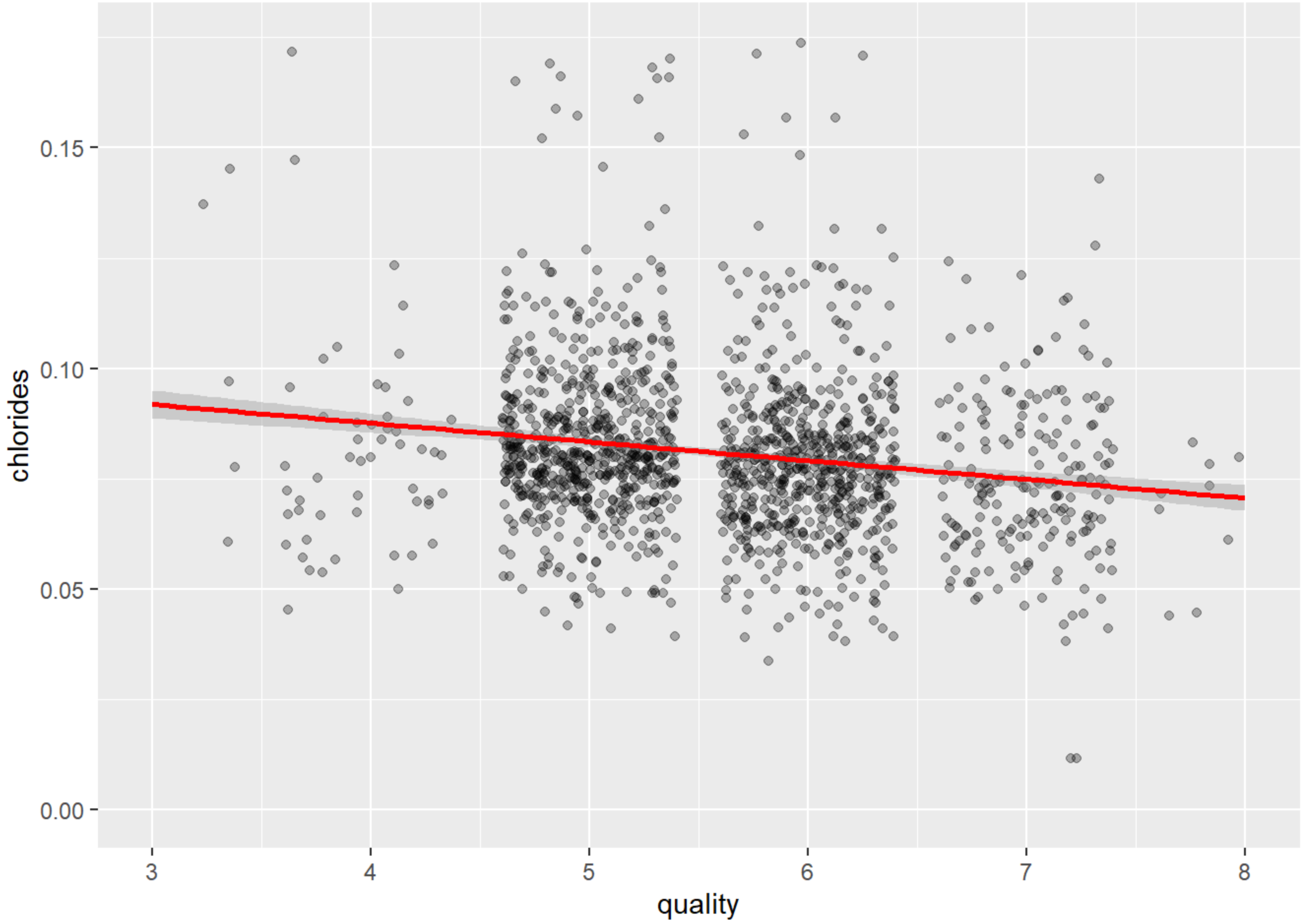


**Negative Correlations**

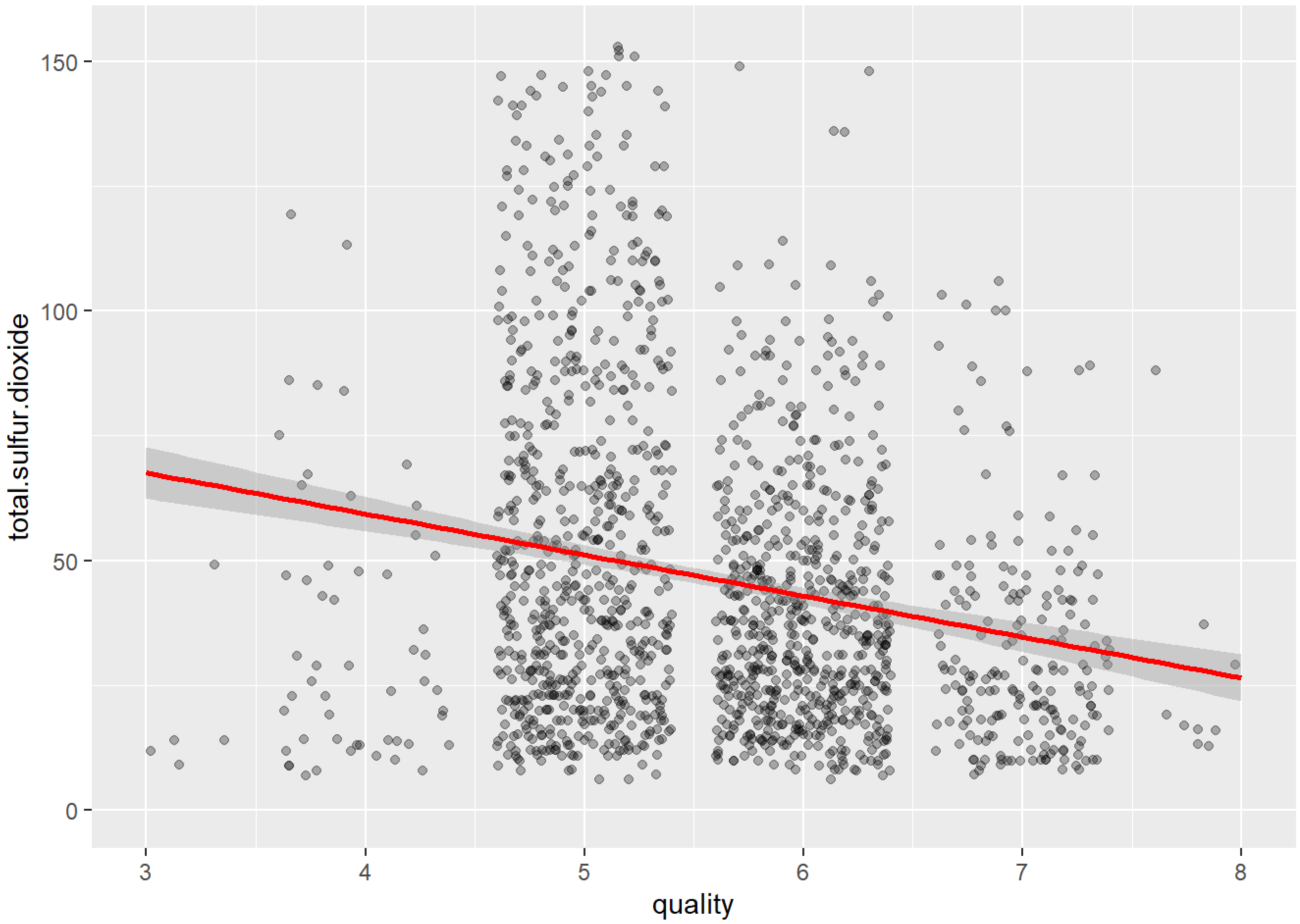
Density.



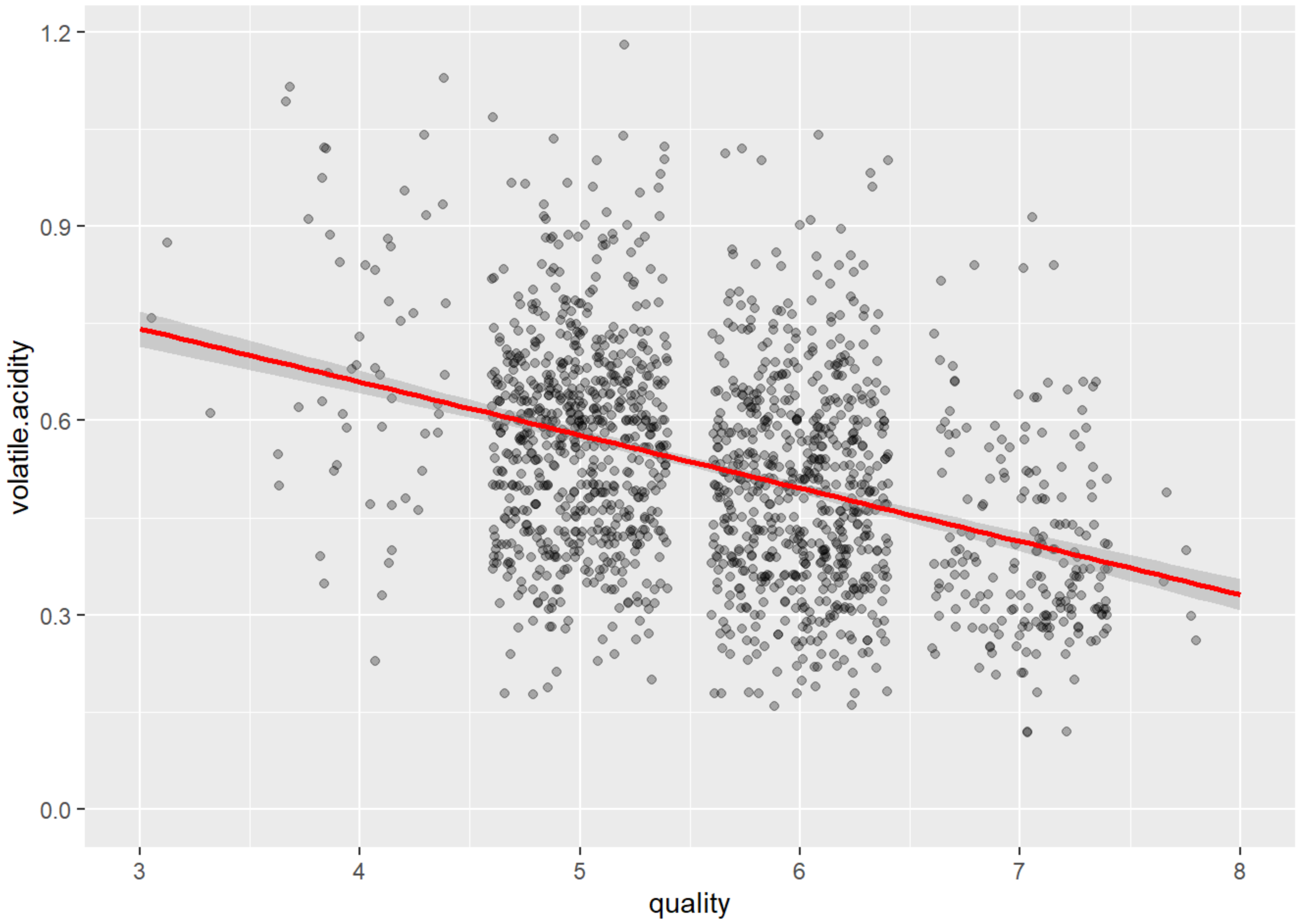
Chlorides.



Total Sulfur Dioxide.

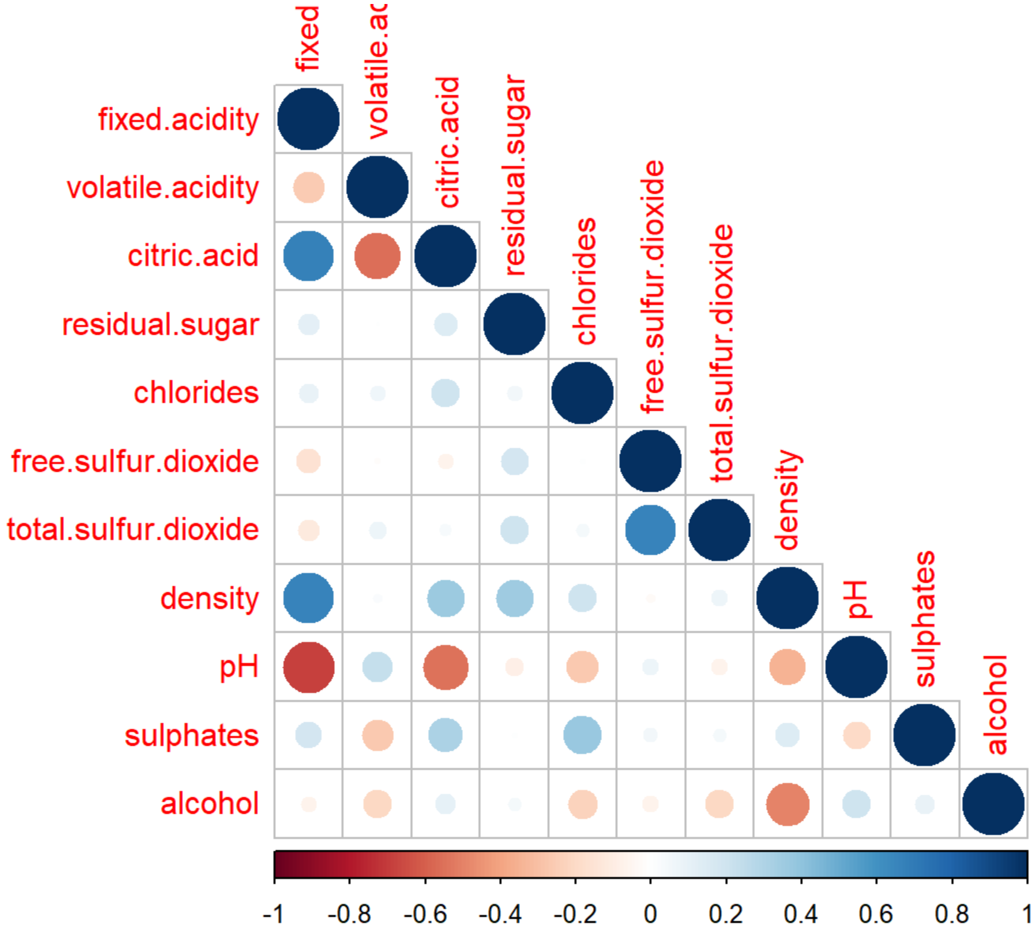


Volatile Acidity.



# Other correlations to Consider

**Correlation Matrix**

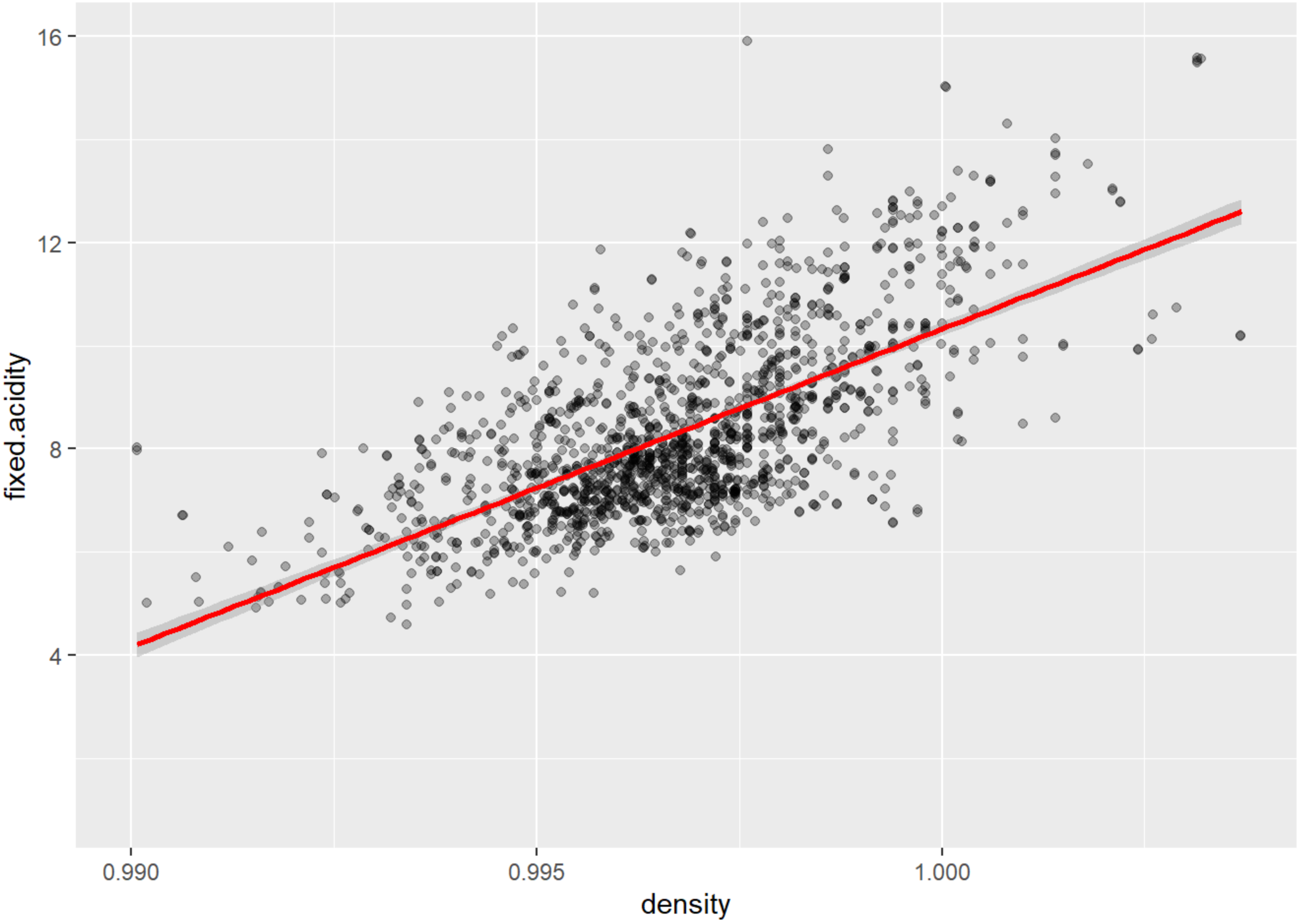


**Correlation of variables calculated to further explore.**

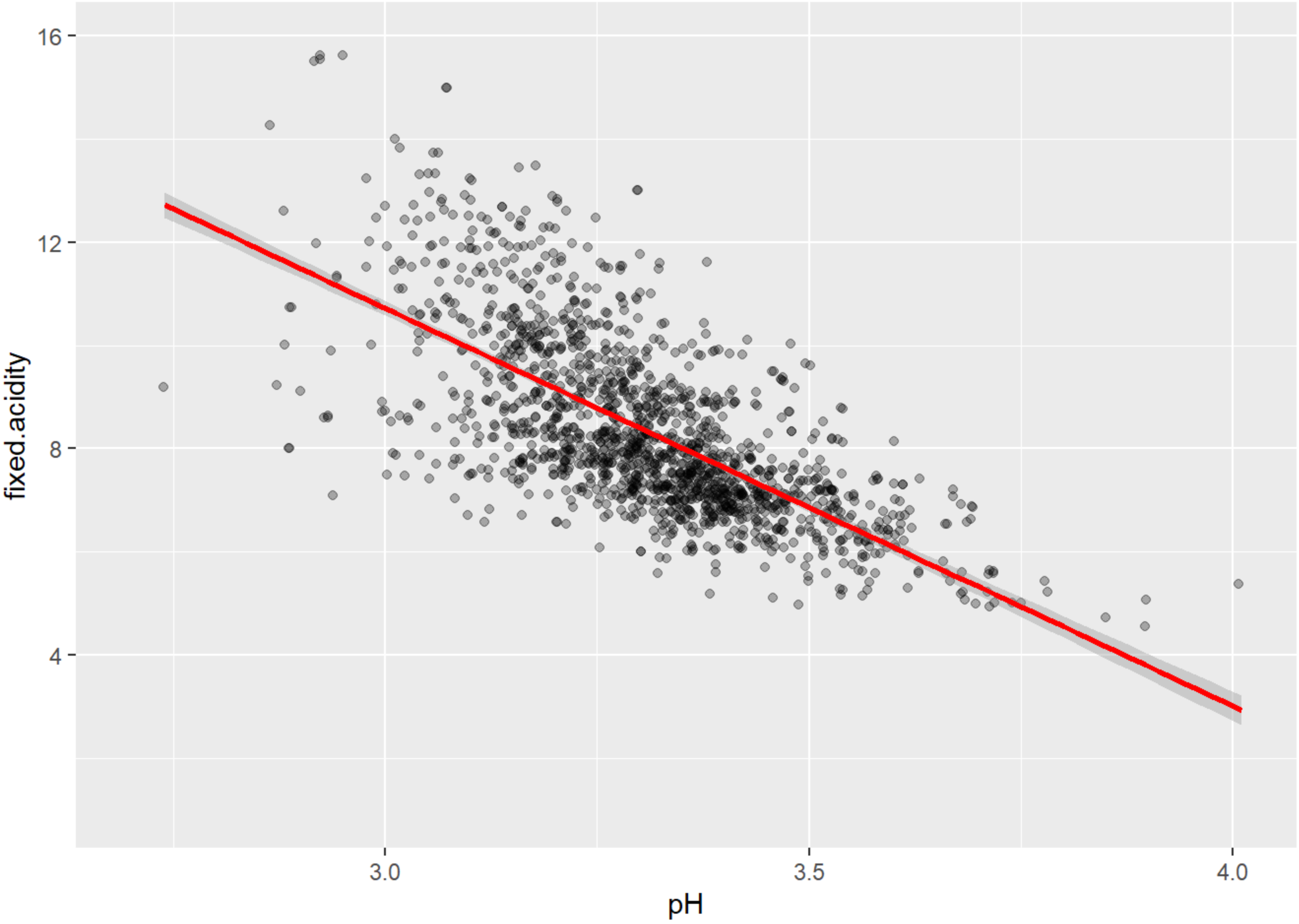
```
## fixed.acidity vs. density      fixed.acidity vs. pH        citric.acid vs. pH
##                  0.6747701                -0.7063602                -0.5387327
##   acids citric vs. fixed  acids citric vs. volatile       Alcohol vs. density
##                  0.6210402                -0.5668516                -0.4937366
##         SO2 free vs. total
##                  0.7390765
```

**Correlation plots to explore.**

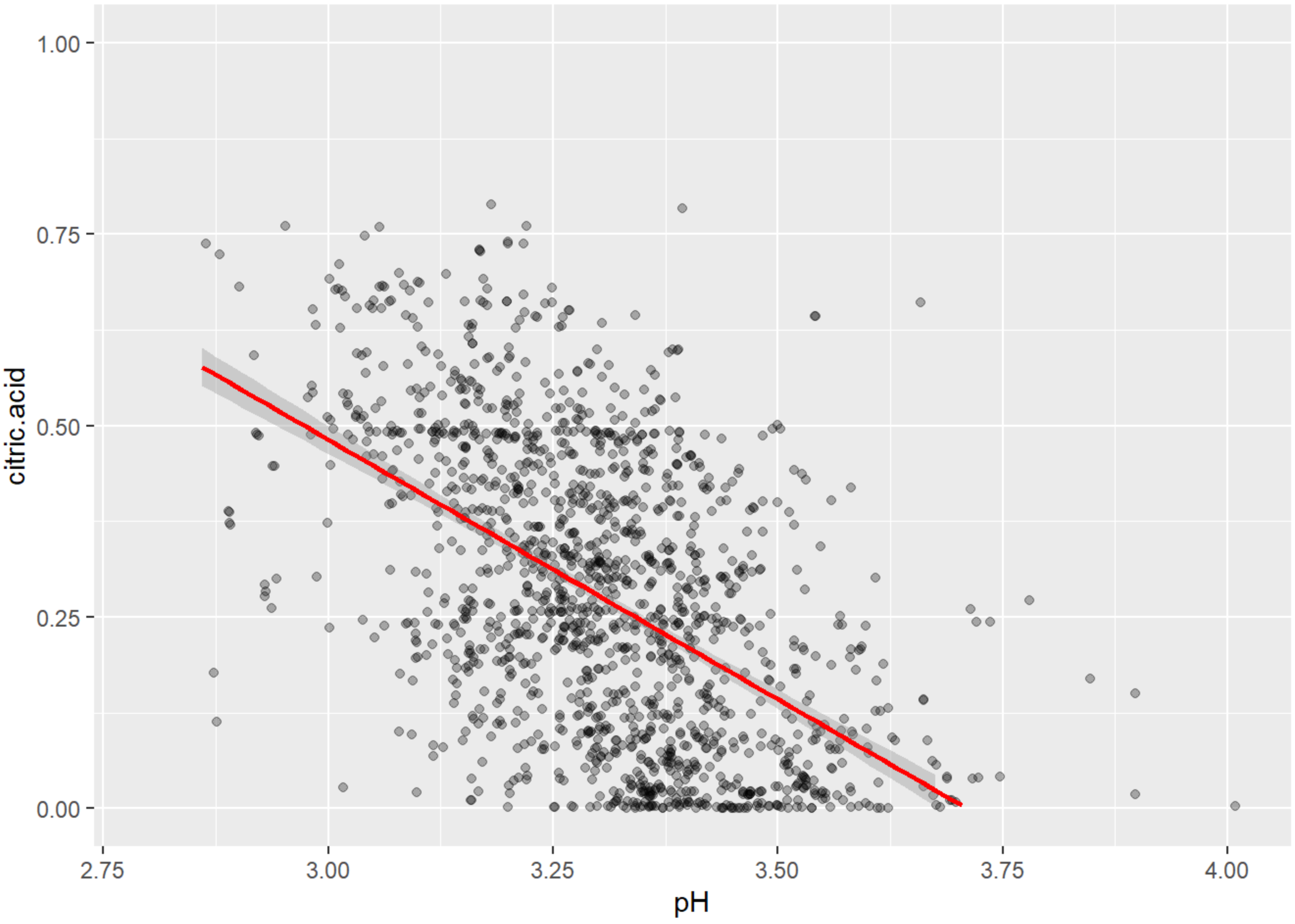
Fixed Acidity vs Density.



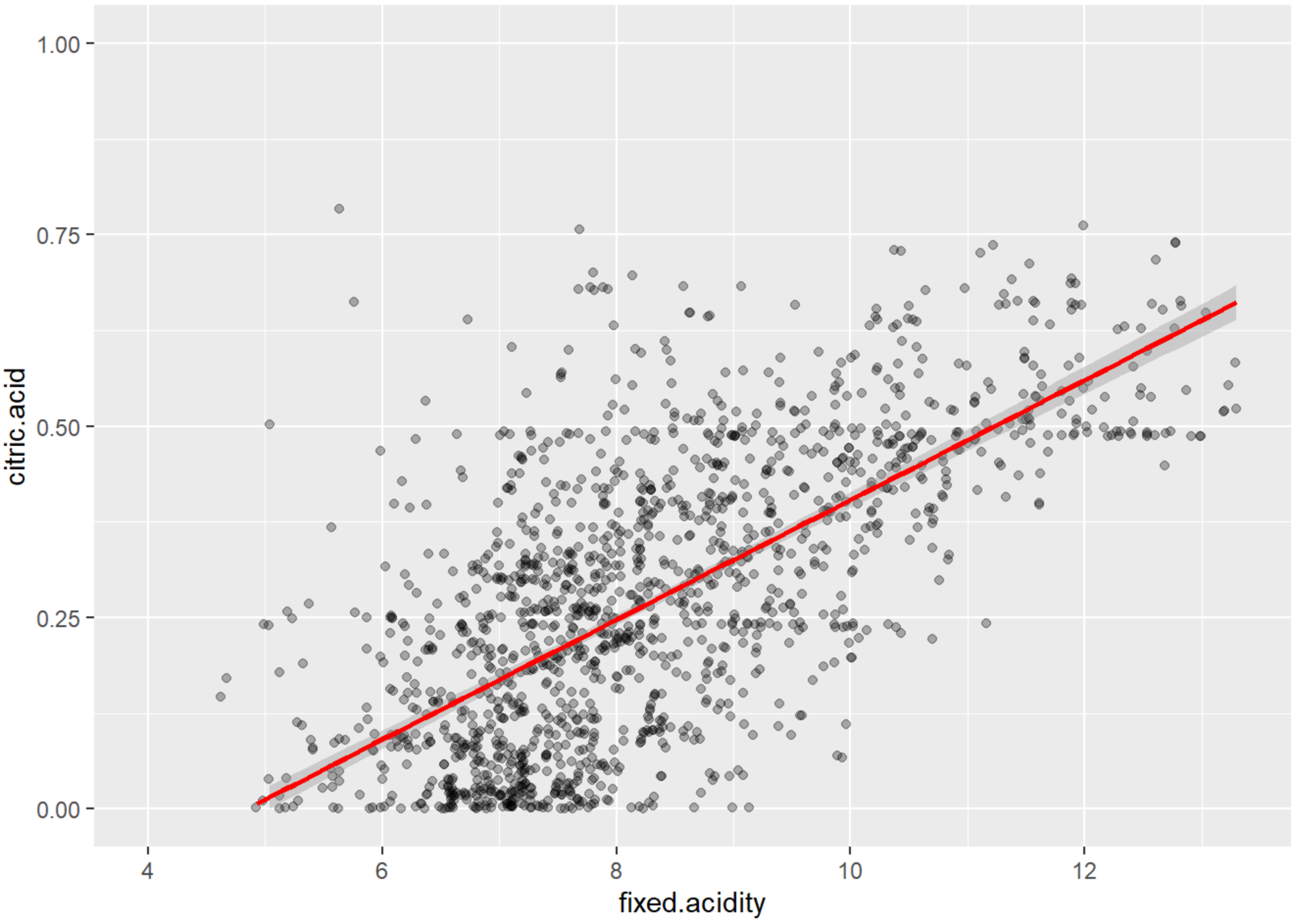
Fixed Acidity vs pH.



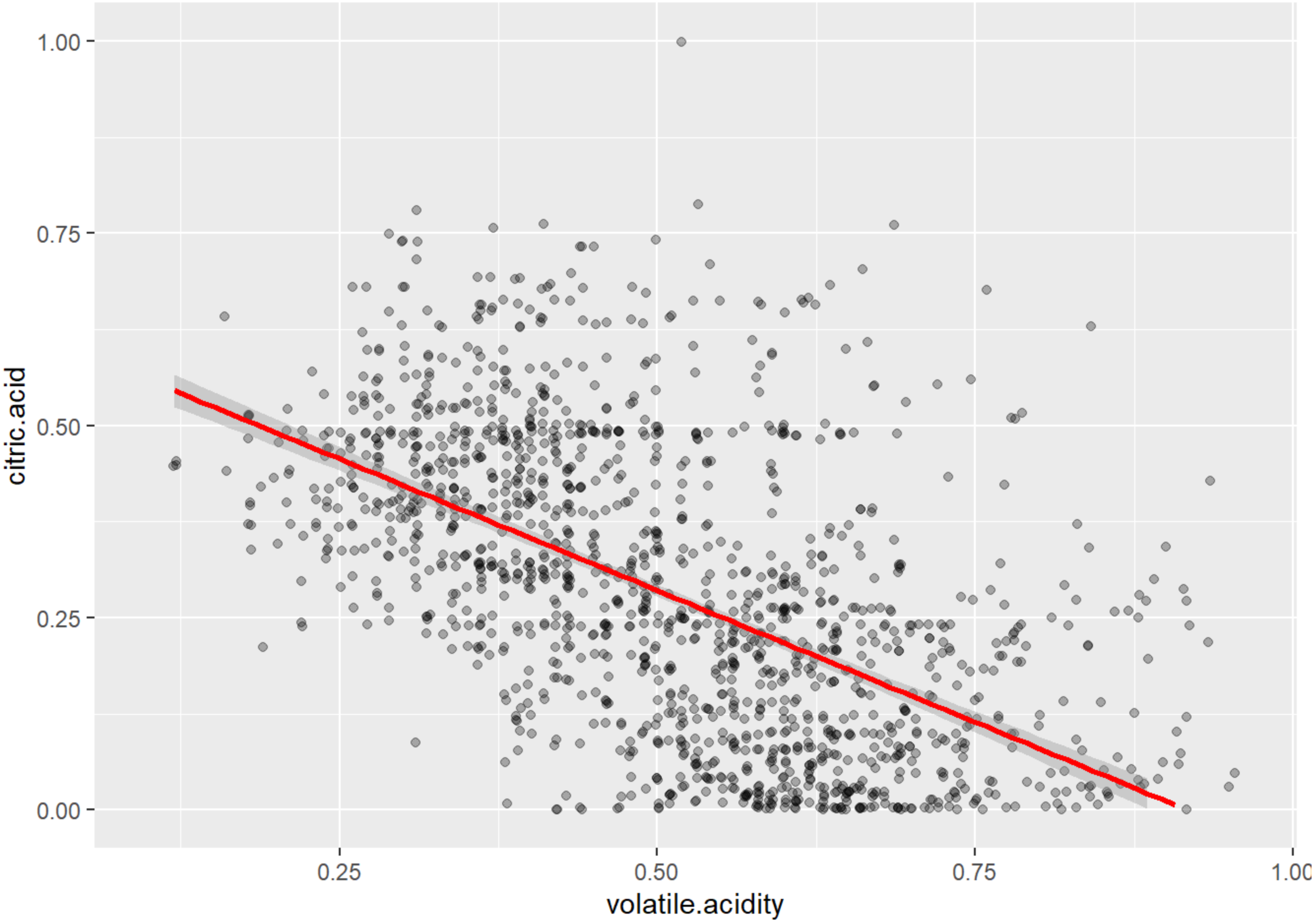
Citric acid vs pH.



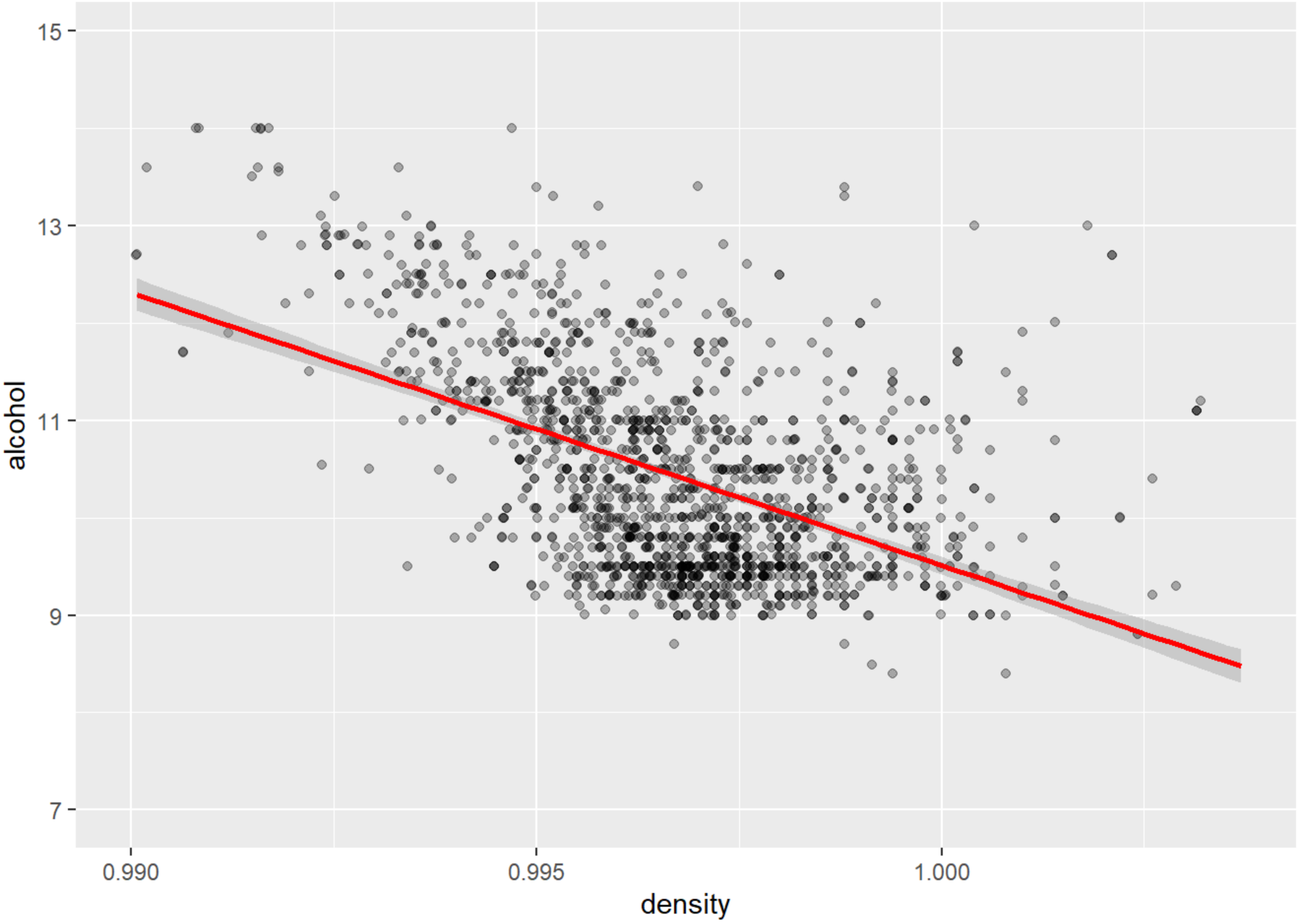
Citric Acid vs Fixed Acidity.



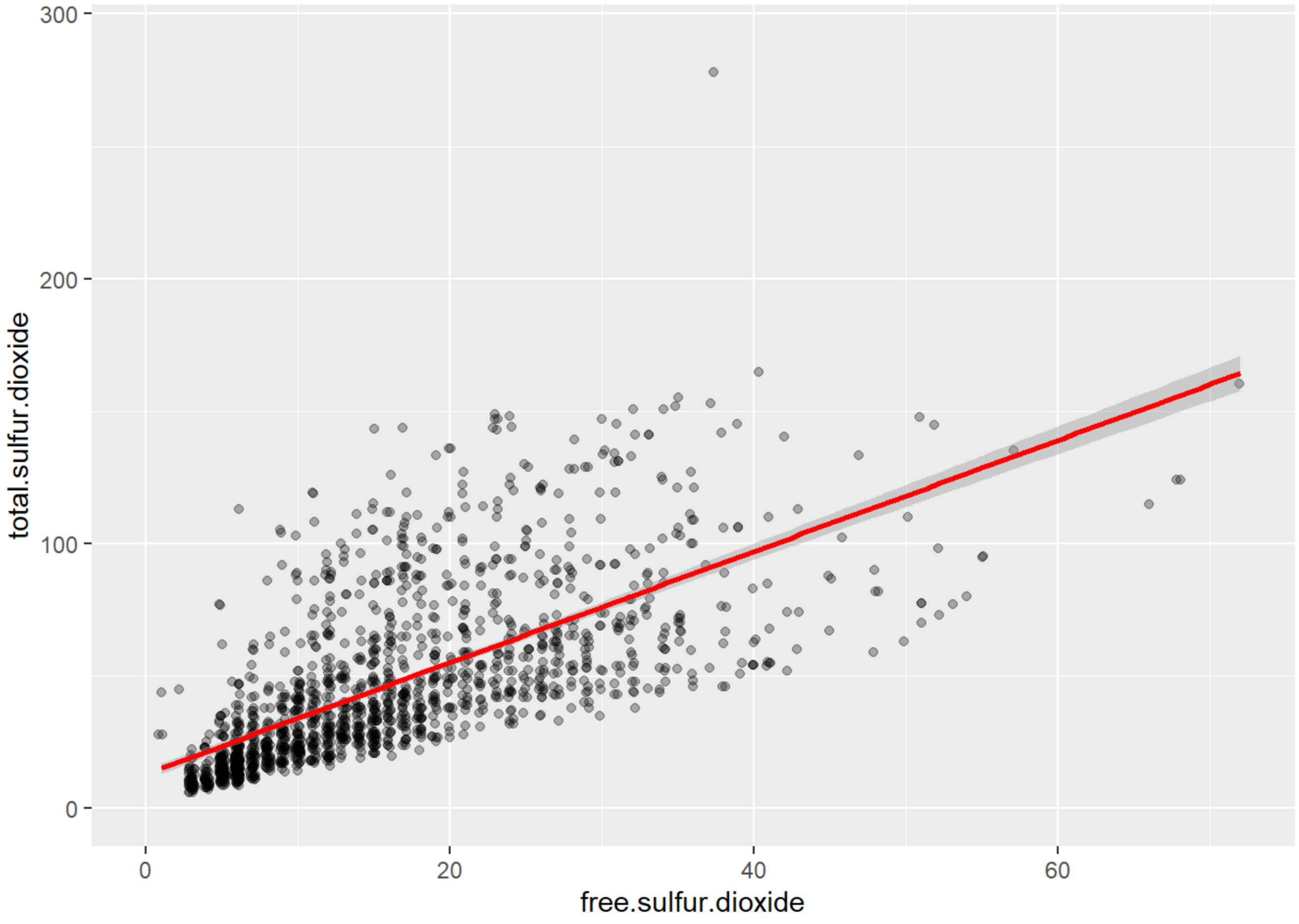
Citric Acid vs Volatile Acidity.



Alcohol vs. Density.



Sulfur Dioxide: Free vs. Total.



## 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?**

There are a few attributes exhibit some trends that look promising to be used to predict quality.
* Quality is positively correlated with alcohol, citric acid, sulphates, and fixed acidity.
* Quality is negatively correlated with volatile acidity, chlorides, density, and total sulfur dioxide.

We can summarize these relations in the following table:

| Attribute Name | Relation with Quality |
|---|---|
| fixed acidity | ~ |
| volatile acidity | - |
| citric acid | + |
| residual sugar | ~ |
| chlorides | - |
| free sulfur dioxide | ~ |
| total sulfur dioxide | ~ |
| density | - |
| pH | - |
| sulphates | + |

| Attribute Name | Relation with Quality |
|---|---|
| alcohol | + |

- ‘~’ means the attribute exhibits no clear trend with quality
- ‘-’ means the attribute is negatively correlated with quality
- ‘+’ means the attribute is positively correlated with quality

**Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?** There are a few attributes that are correlated based on physical and chemical principles:

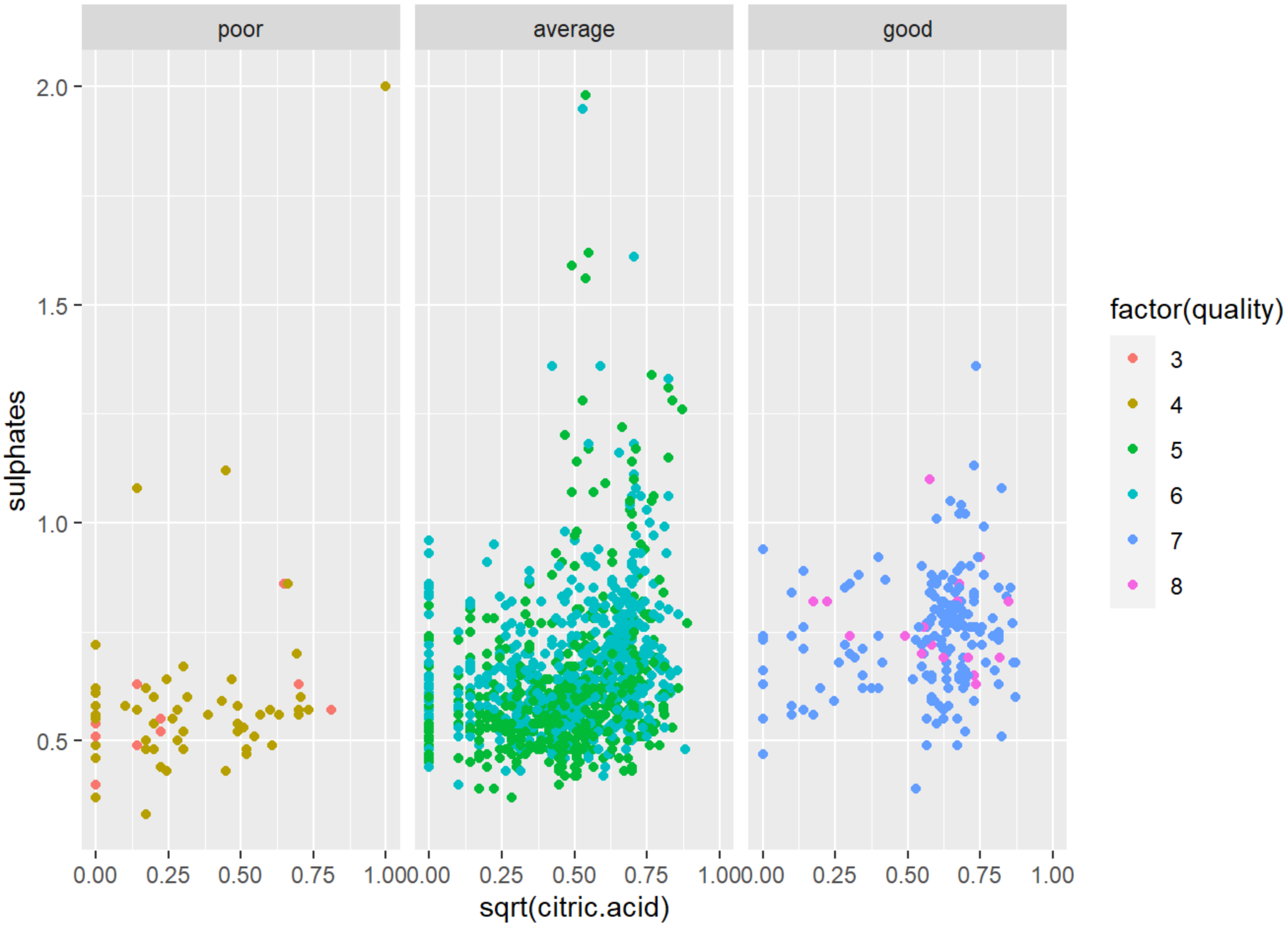
- Fixed acidity and density are positively correlated because the main fixed acids in wine, tartaric acid, has a higher density than water, therefore wines that contain more tartaric acid have a higher density.
- Fixed acidity and pH are negatively correlated because higher concentration of fixed acidity makes the wine more acidic, therefore the wine has a lower pH.
- Citric acid and pH are negatively correlated because higher concentration of citric acid, which is non-volatile, makes the wine more acidic, therefore the wine has a lower pH.
- Fixed acidity and citric acid are positively correlated because the fixed acidity includes citric acid.
- Citric acid and volatile acidity are negatively correlated because during the fermentation process the oxygen in wine is kept to a minimum which contributes to volatile acidity which ruins the wine. Citric acid is used to supplement the fermentation process in wine.
- Density and alcohol are negatively correlated because alcohol has a lower density than water, therefore wines that contain more alcohol have a lower density.
- Total sulfur dioxide and free sulfur dioxide are positively correlated because total sulfur dioxide includes free sulfur dioxide.

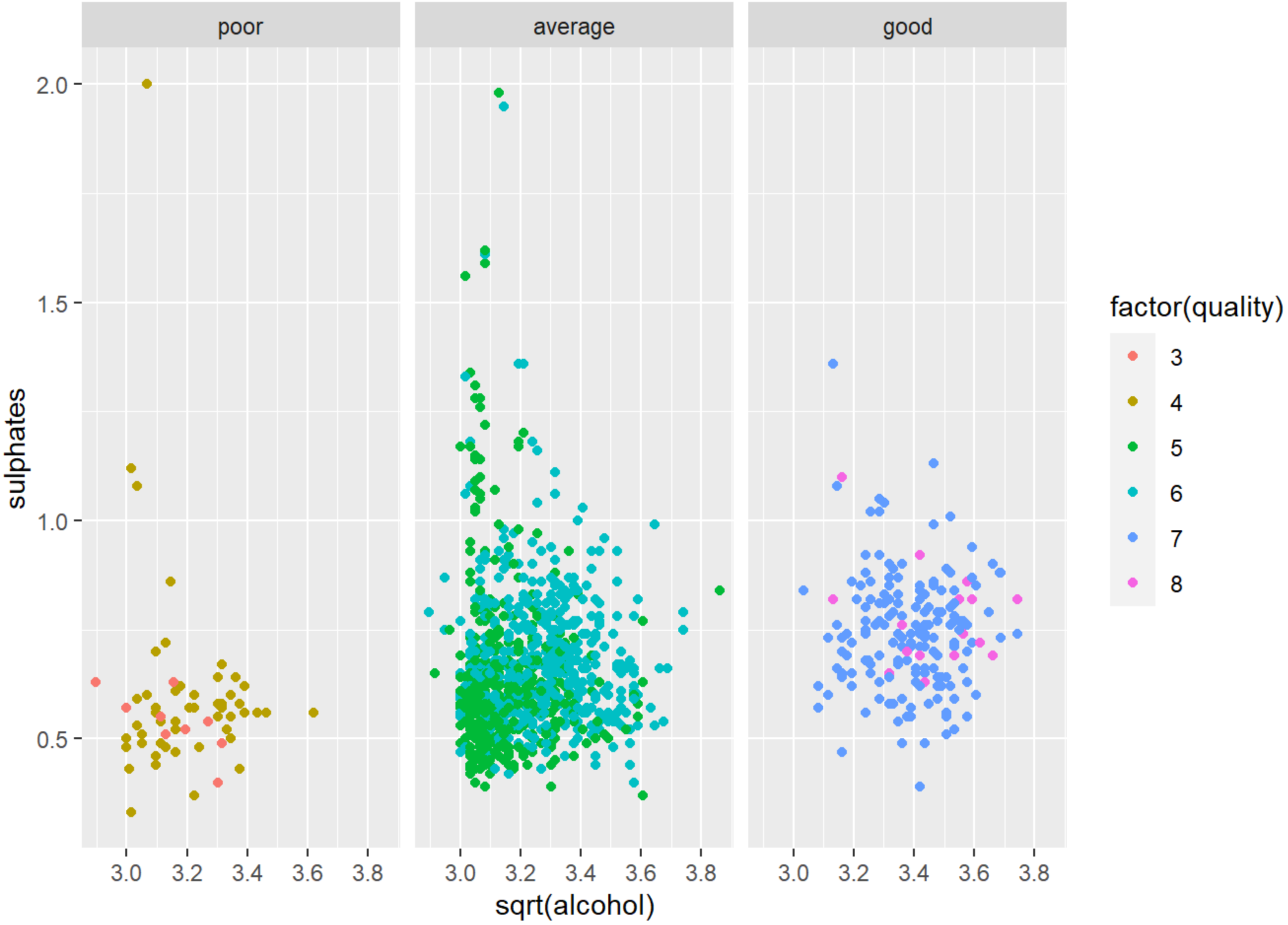
**What was the strongest relationship you found?** Observing from the plot, alcohol has the strongest relationship with quality.

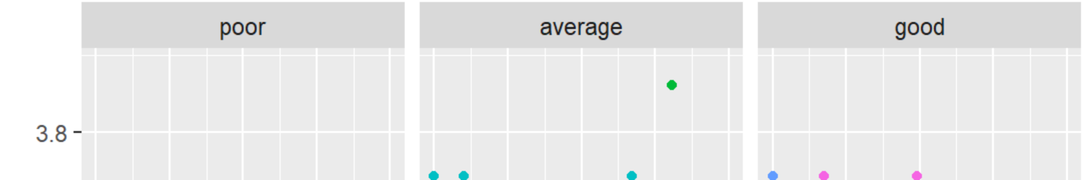
# 4. MultiVariate Exploration

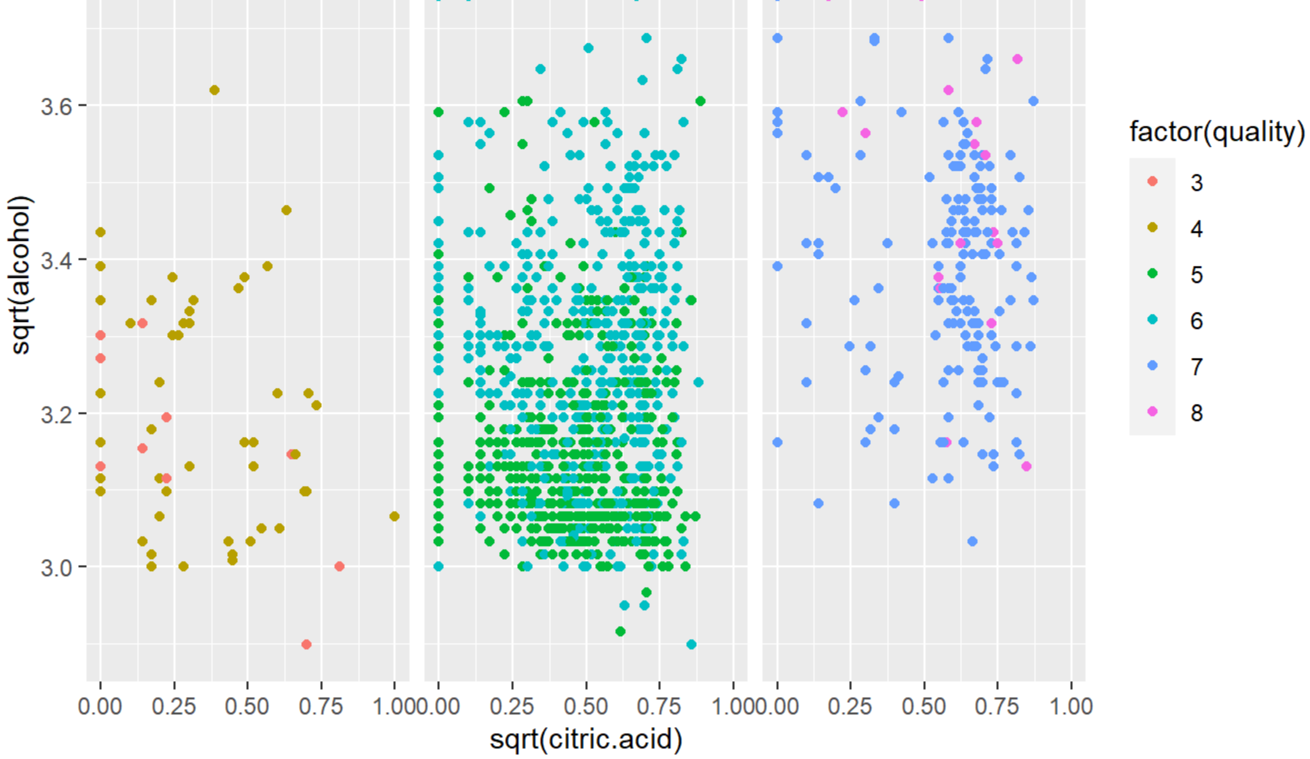
## Multivariate Plots

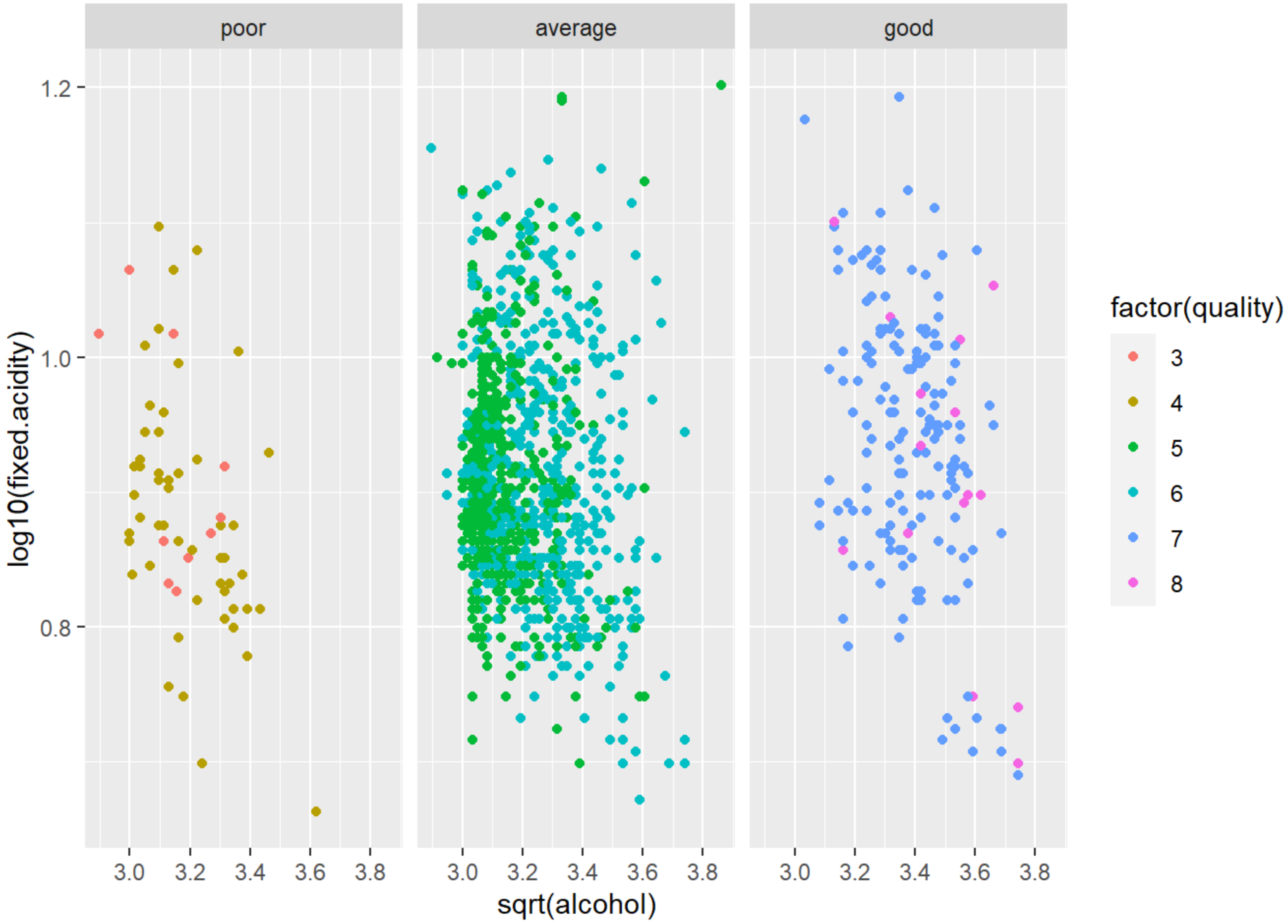
The scatter plots are facet wraped by rating. Graphs for four variables citric.acid, fixed.acidity, sulphates, and alcohol which show high correlations with quality. Other variables that affect quality are fixed acidity, pH, and density.

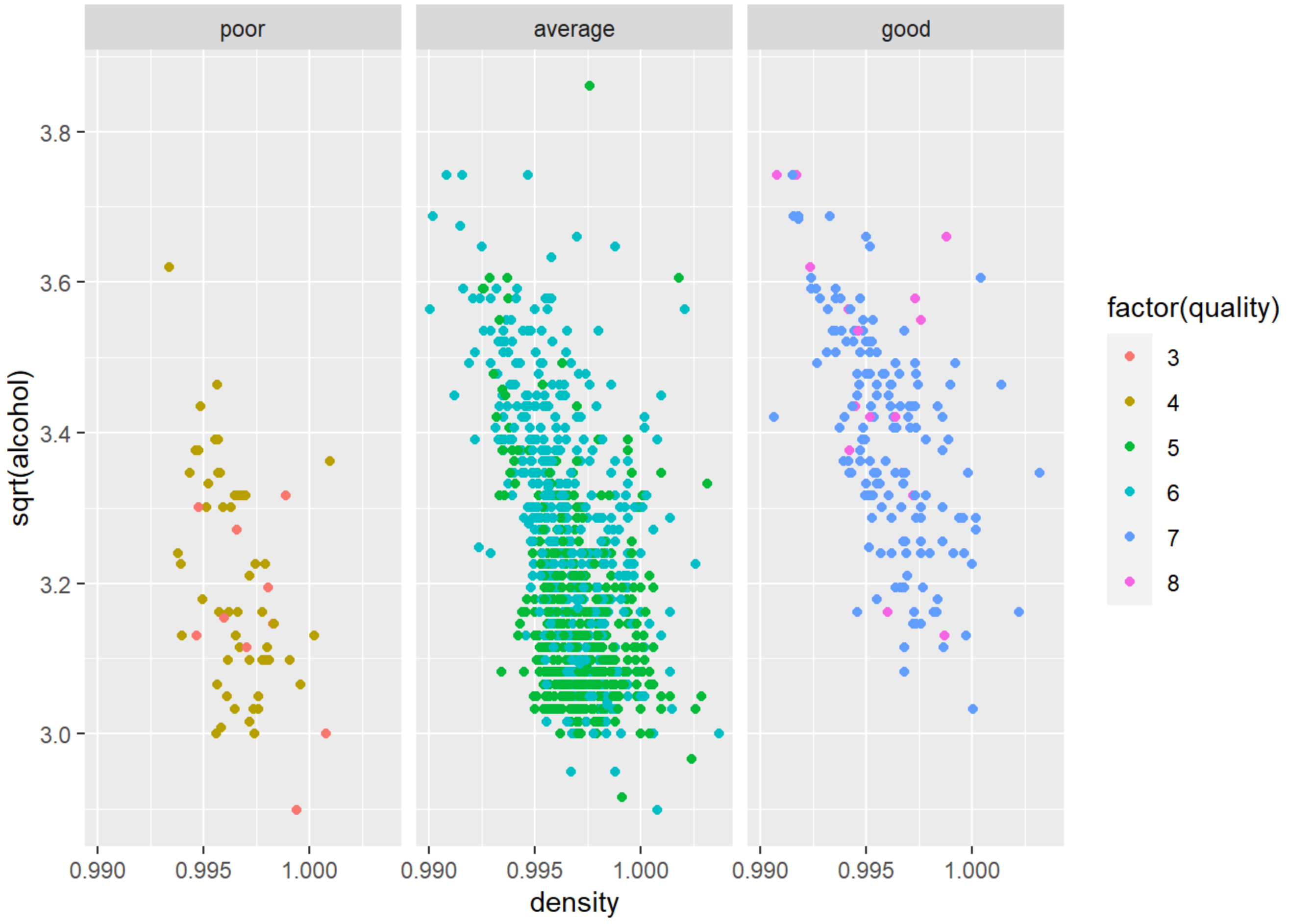

poor
average
good
sulphates
2.0
1.5
1.0
0.5
0.00
0.25
0.50
0.75
1.00
sqrt(citric.acid)
factor(quality)
3
4
5
6
7
8

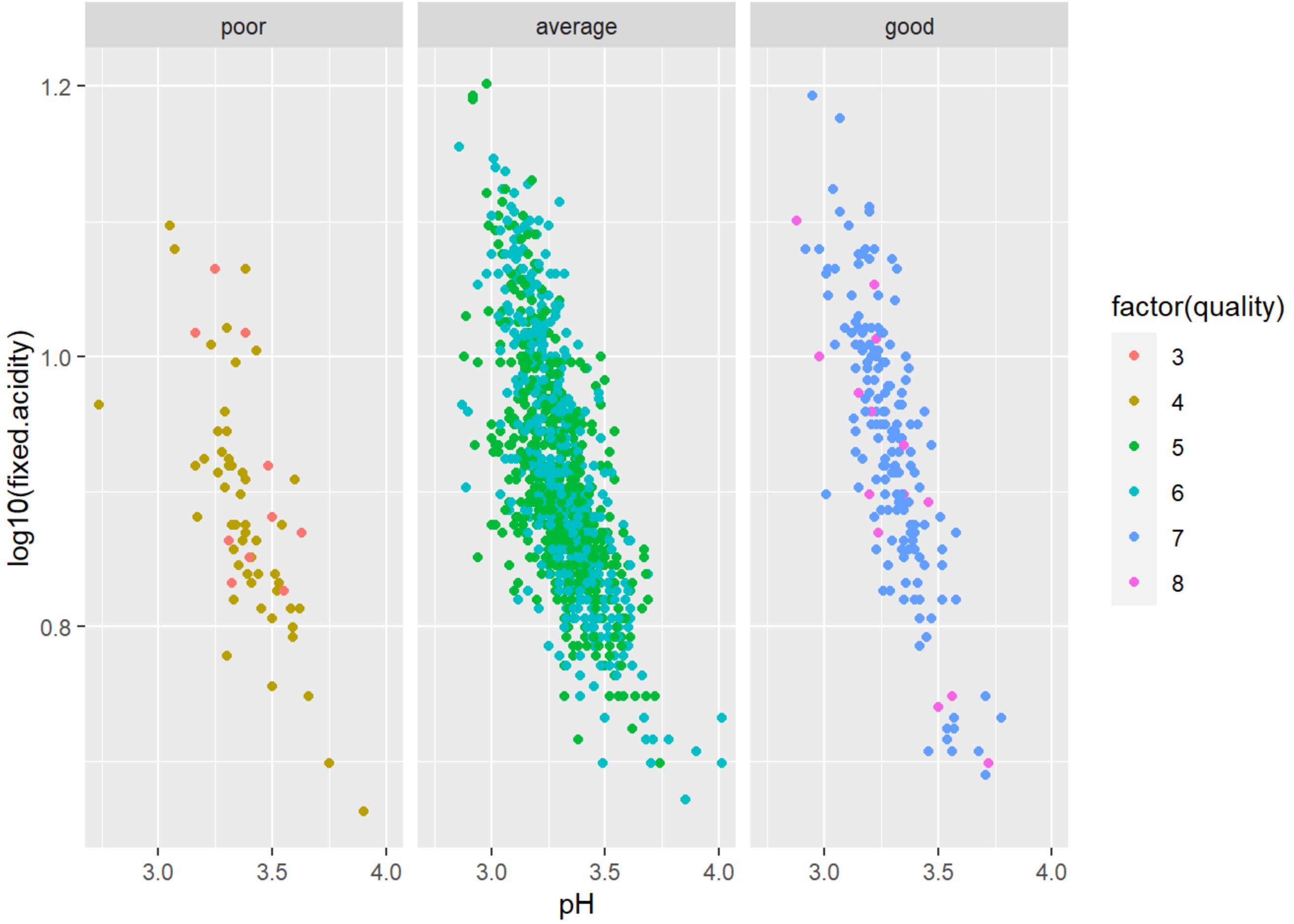

poor
average
good
sulphates
2.0
1.5
1.0
0.5
3.0
3.2
3.4
3.6
3.8
sqrt(alcohol)
factor(quality)
3
4
5
6
7
8

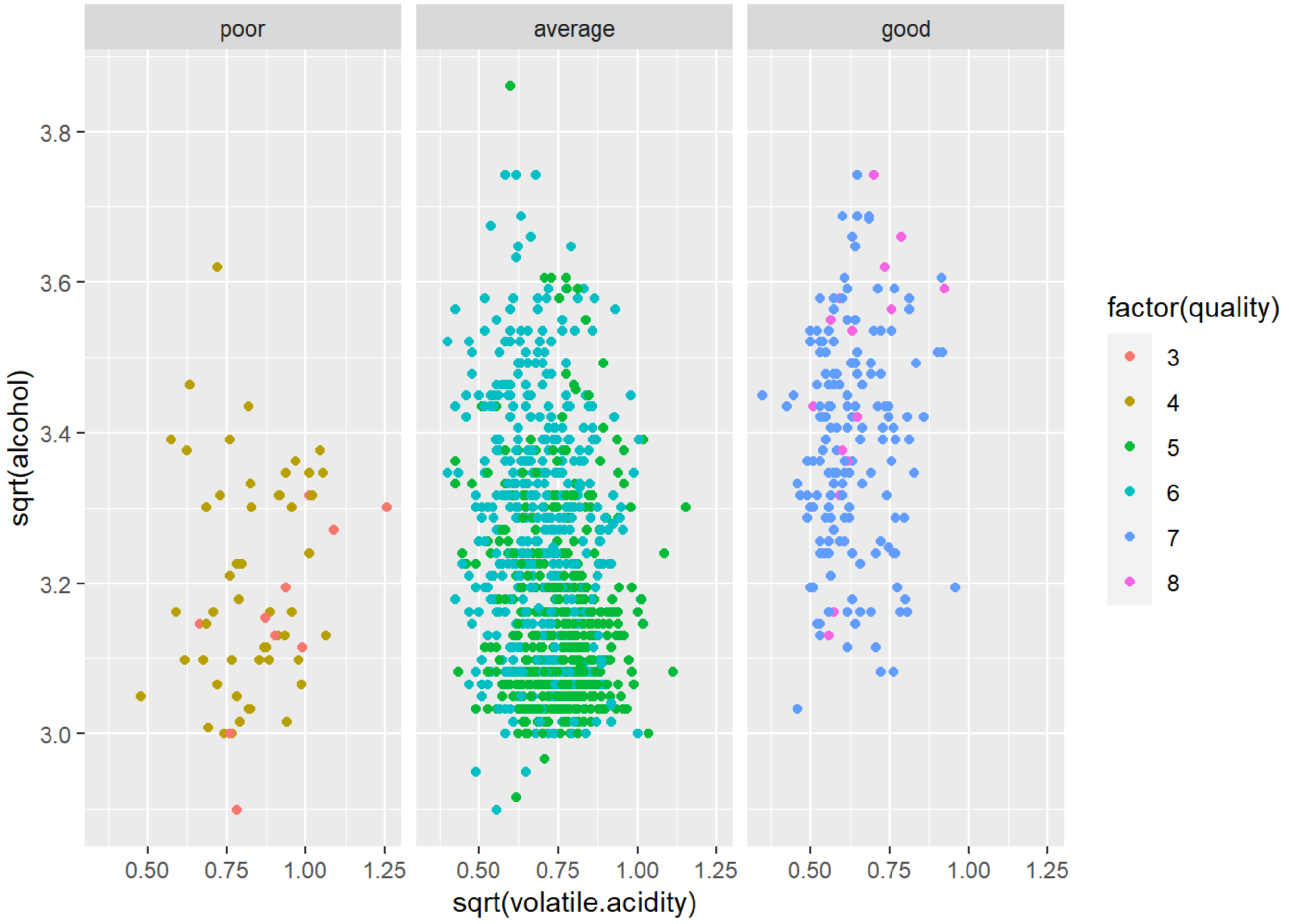

poor
average
good
3.8


sqrt(alcohol)
3.6
3.4
3.2
3.0
0.00
0.25
0.50
0.75
1.00
sqrt(citric.acid)
factor(quality)
3
4
5
6
7
8


poor
average
good
log10(fixed.acidity)
1.2
1.0
0.8
3.0
3.2
3.4
3.6
3.8
sqrt(alcohol)
factor(quality)
3
4
5
6
7
8


poor
average
good
sqrt(alcohol)
3.8
3.6
3.4
3.2
3.0
0.990
0.995
1.000
density
factor(quality)
3
4
5
6
7
8


poor
average
good
log10(fixed.acidity)
1.2
1.0
0.8
3.0
3.5
4.0
pH
factor(quality)
3
4
5
6
7
8



# Linear Model

Linear multivariable model was created to predict the wine quality based on chemical properties.

- First model I will only look at the most promising attribute alcohol from univariate exploration section.
- Second model I will add the attributes that exhibit a clear trend with quality from the bivariate exploration section besides alcohol.
- Third model I will add all the rest variables.

```
##
## Calls:
## m1: lm(formula = I(quality) ~ I(sqrt(alcohol)), data = red_wine)
## m2: lm(formula = I(quality) ~ I(sqrt(alcohol)) + sqrt(volatile.acidity) +
##     sulphates + sqrt(total.sulfur.dioxide) + sqrt(citric.acid) +
##     log10(fixed.acidity) + log10(chlorides) + density, data = red_wine)
## m3: lm(formula = I(quality) ~ I(sqrt(alcohol)) + sqrt(volatile.acidity) +
##     sulphates + sqrt(total.sulfur.dioxide) + sqrt(citric.acid) +
##     log10(fixed.acidity) + log10(chlorides) + density + log10(residual.sugar) +
##     sqrt(free.sulfur.dioxide) + pH, data = red_wine)
##
## =====================================================================
##                                    m1           m2           m3
## ---------------------------------------------------------------------
##   (Intercept)                   -2.024***    20.442       20.866
##                                 (0.354)     (14.999)     (22.744)
##   I(sqrt(alcohol))               2.376***     1.781***     1.750***
##                                 (0.110)      (0.131)      (0.178)
##   sqrt(volatile.acidity)                     -1.636***    -1.567***
##                                              (0.177)      (0.179)
##   sulphates                                   0.853***     0.853***
##                                              (0.108)      (0.112)
##   sqrt(total.sulfur.dioxide)                 -0.022**     -0.048***
##                                              (0.008)      (0.012)
##   sqrt(citric.acid)                          -0.188       -0.165
##                                              (0.113)      (0.116)
##   log10(fixed.acidity)                        1.194***     0.807
##                                              (0.334)      (0.526)
##   log10(chlorides)                           -0.471***    -0.531***
##                                              (0.133)      (0.136)
##   density                                   -21.377      -20.593
##                                             (15.043)     (23.209)
##   log10(residual.sugar)                                    0.177
##                                                           (0.147)
##   sqrt(free.sulfur.dioxide)                                0.050*
##                                                           (0.020)
##   pH                                                      -0.292
##                                                           (0.200)
## ---------------------------------------------------------------------
##   R-squared                      0.227        0.352        0.356
##   N                           1599         1599         1599
## =====================================================================
##   Significance: *** = p < 0.001; ** = p < 0.01; * = p < 0.05
```

The most promising attribute alcohol alone has R-squared value of 0.227. By adding the other 7 promsing attributes, R-squared value is a 1.55 times better becoming 0.352. But adding the remaining 3 attributes only increases the R-squared value by .004 to 0.356.

# Multivariate Analysis

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

High alcohol contents with sulphate concentrations above 0.5 and citric acid above 0.5 with a pH < 3.5 seems to produce a better wine.

**Were there any interesting or surprising interactions between features?**
Fixed acididty and pH had a stronger negative correlation than others.

**What was the strongest relationship you found?**
The strongest relationship is the correlation between alcohol and quality.

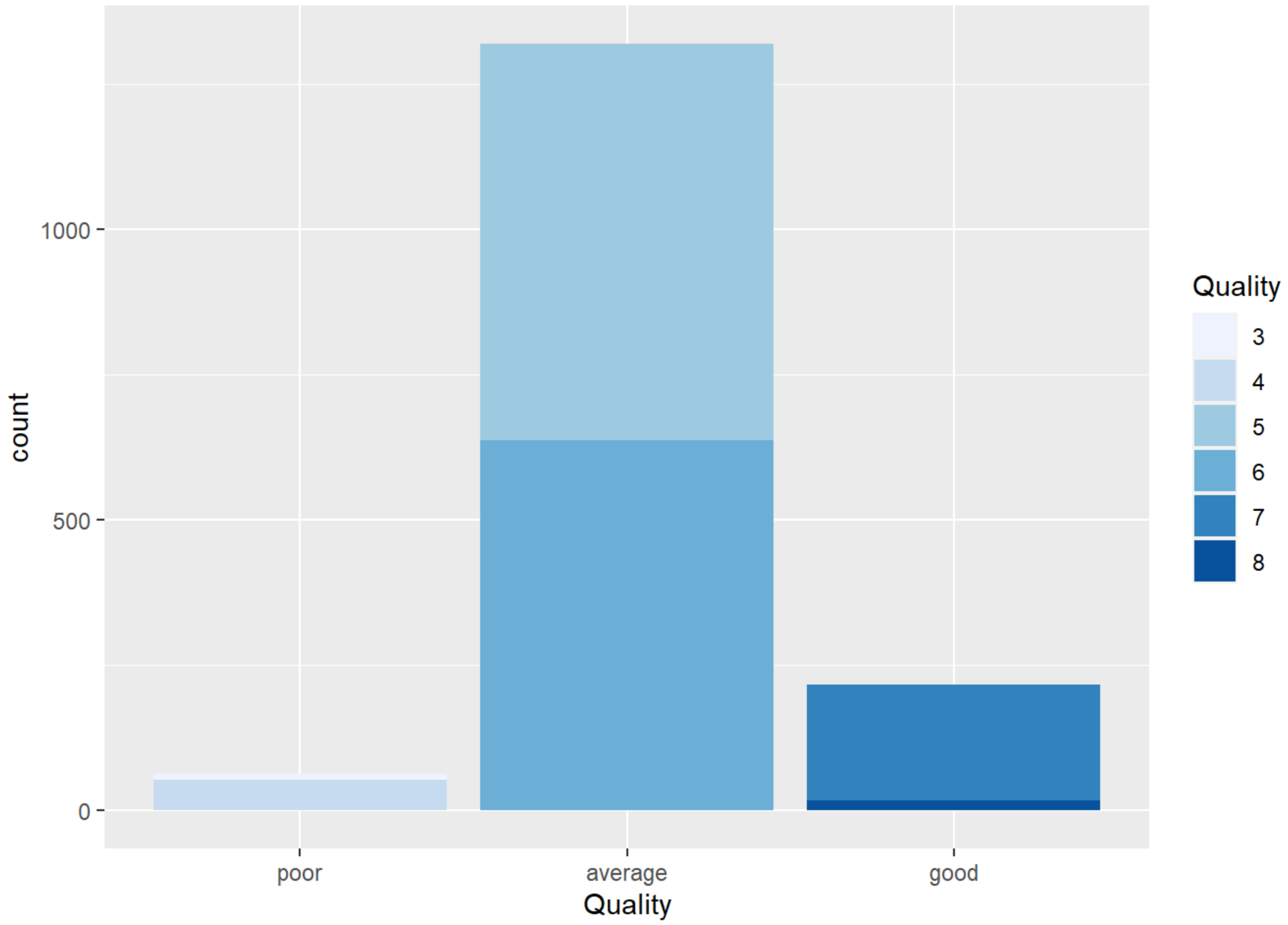
**OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model.**

I created three linear models to predict the output attribute quality. The strength of the model is that it is a simple linear model and it is easy to interpret. Due to the limitation of the dataset, only physical and chemical attributes are available, and other import attributes, such as price, color, smell, etc are missing. The other attributes may influence the quality values to a large extent.

---

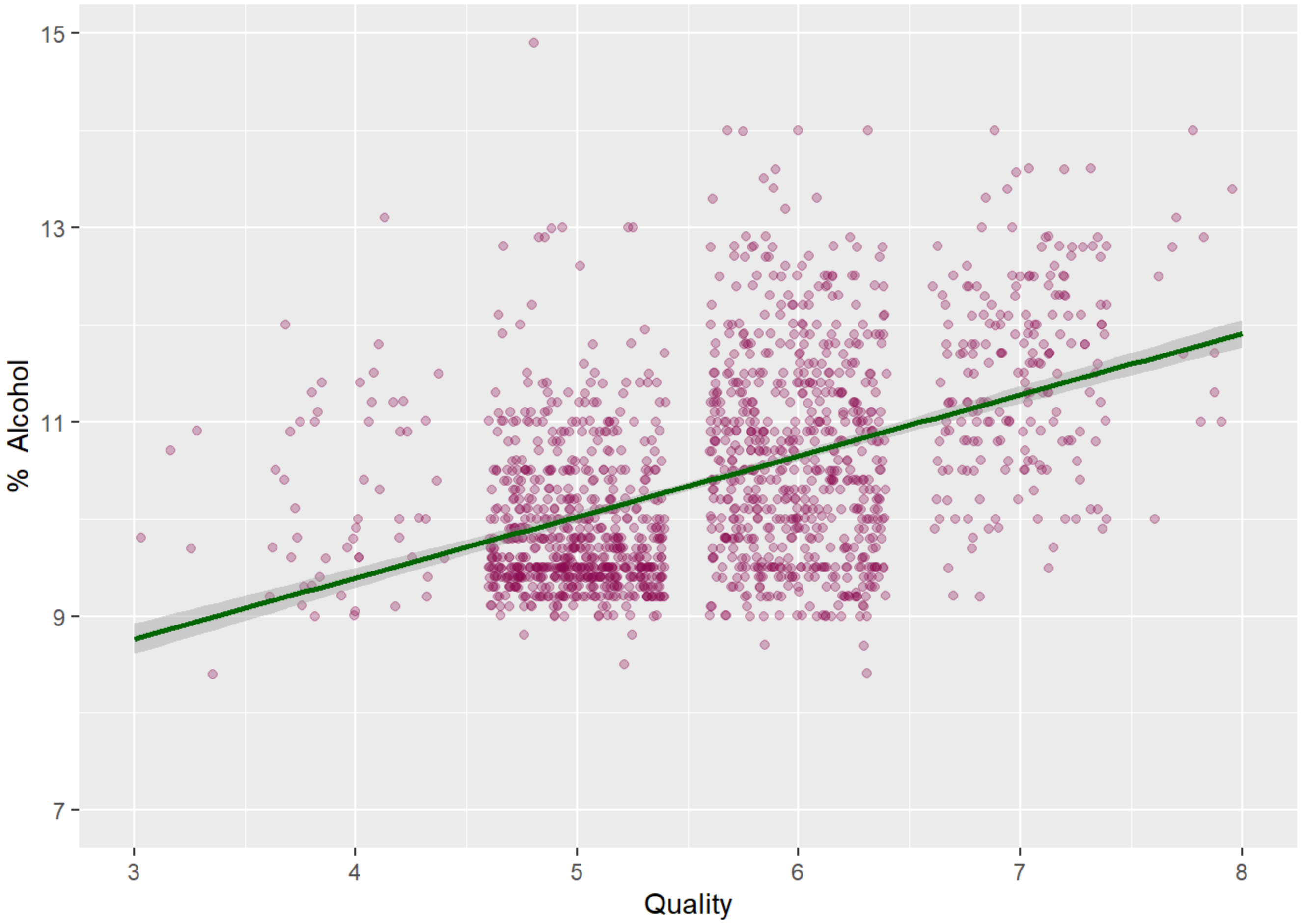
# 5. Final Plots

## Plot One



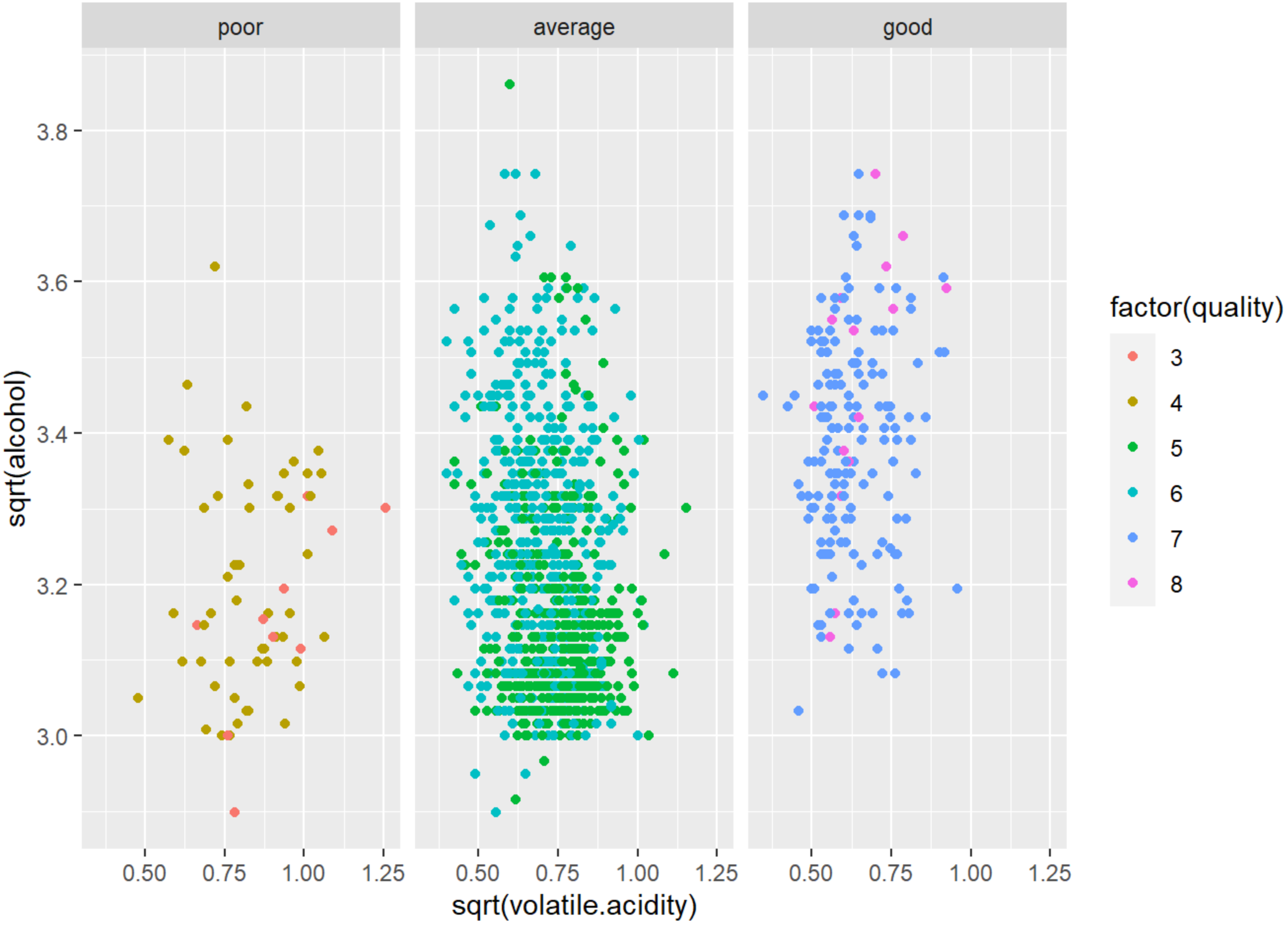
We can see the majority of the wines have a quality level of 5-6 which I consider average with very few wines in either extremes (3 or 8).

## Plot Two



From the above plot it is clear that wine quality increases with % of alcohol in it. Interestingly the alcohol percentage of higher quality wines( quality> 6) increased with quality but some lower quality wines do not have the lowest alcohol percentage.

# Plot Three



The above plots include all wines, some things that can be inferred from this plot are:

- High volatile acidity, with few exceptions, kept wine quality down.
- A combination of high alcohol content and low volatile acidity produced better wines.

# 6. Reflection

This dataset has 11 physicochemical properties of 1599 red wines. I read the information about each property so I understood the overall implications as I investigated the dataset further. After looking at the distributions of some variables, I examined the relationship between two and, eventually, three-variable combinations.

Wine quality depends on many features, through this exploratory data analysis I was able to relate some of the key factors like alcohol content, sulphates, and volatile acidity. The graphs adequately illustrate the factors that make good wines ‘good’ and poor wines ‘poor’.

In this data, my main struggle was to get a higher confidence level when predicting factors that are responsible for the production of good quality of wines. As the data was very centralized towards the ‘Average’ quality, this dataset did not have enough data on the extreme edges to accurately build a model that can predict the quality of a wine.

In the future, perhaps I can obtain a more robust dataset about Red Wines in order to build a more models. I believe by incorporating other types of attributes, such as price, color and smell, a better model can be built to predict the quality of wine.