Explore and Summarize Data by Satvik Sachdev

This project will apply exploratory data analysis techniques using R to analyze chemical properties of Red Wines in order to answer the following question: - “Which chemical properties influence the quality of red wines?”

The dataset contains information about 1,599 red wines with 11 variables on the chemical properties of the wine.

The quality of each wine is rated between 0(very bad) and 10(very excellent) by at least 3 wine experts.

Data Exploration

Summary of dataset

```
##  [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"
```

```
## [1] 1599   13
```

```
## 'data.frame':    1599 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

- The red wine dataset has 1599 observations.
- The total number of factors affecting each wines quality are 13.
- Input variables: all variables except quality.
- Output Variable: Quality is the output variable measured with the help of input variables. Each of the 12 variables contribute to the wine quality.

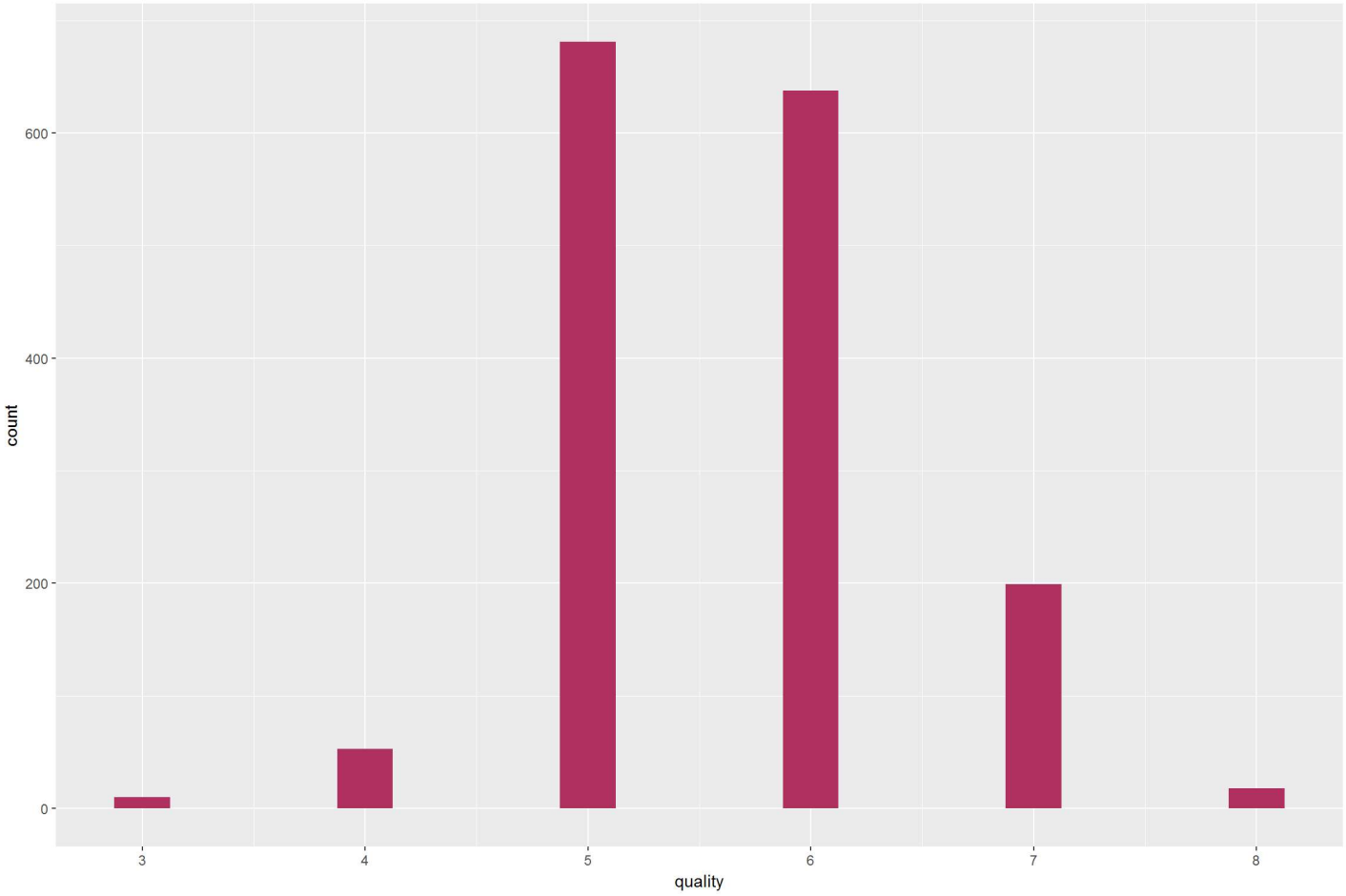
Univariate Plots Section

Generic function to create plots

Quality

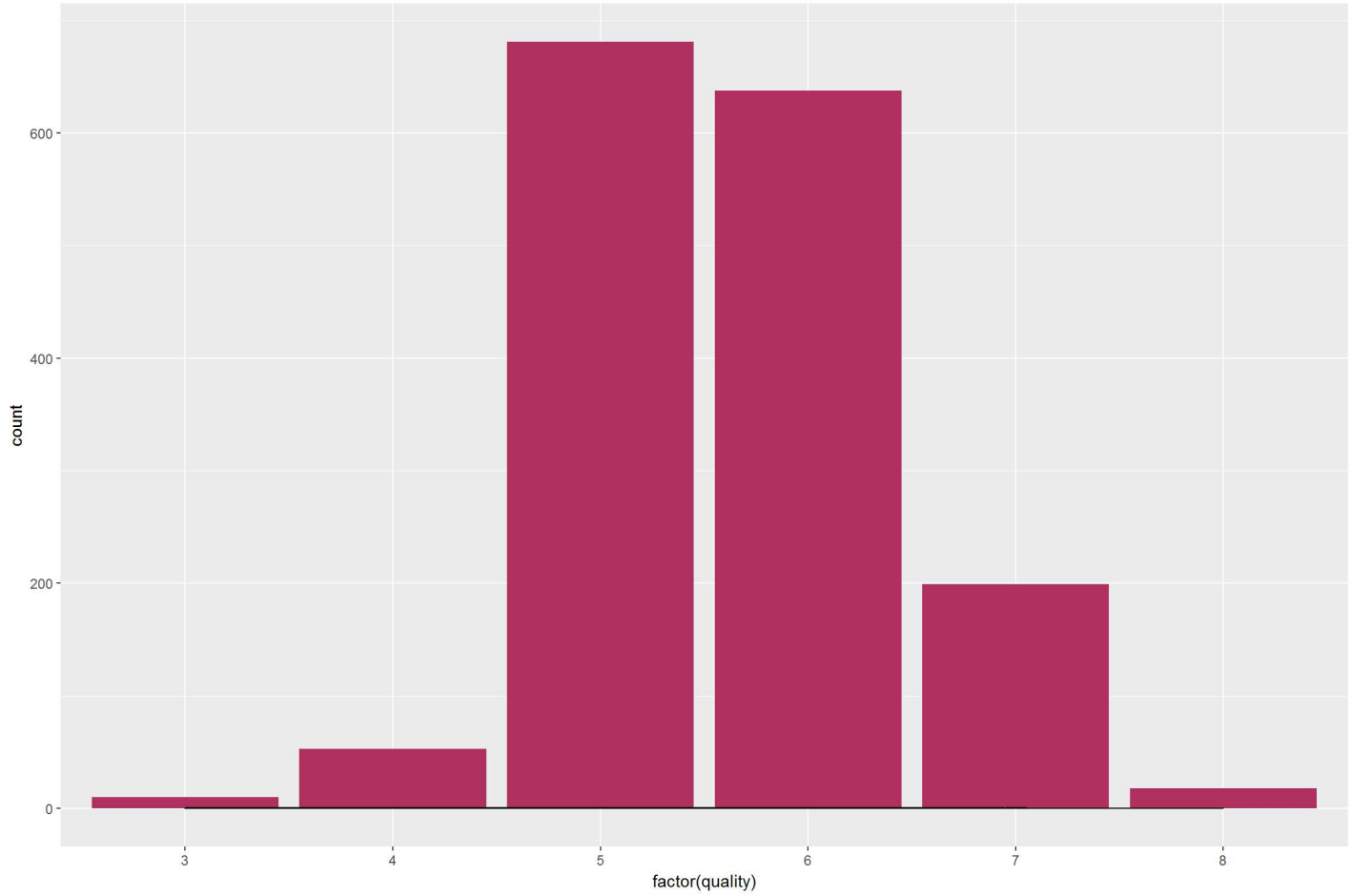
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.636   6.000   8.000
```

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



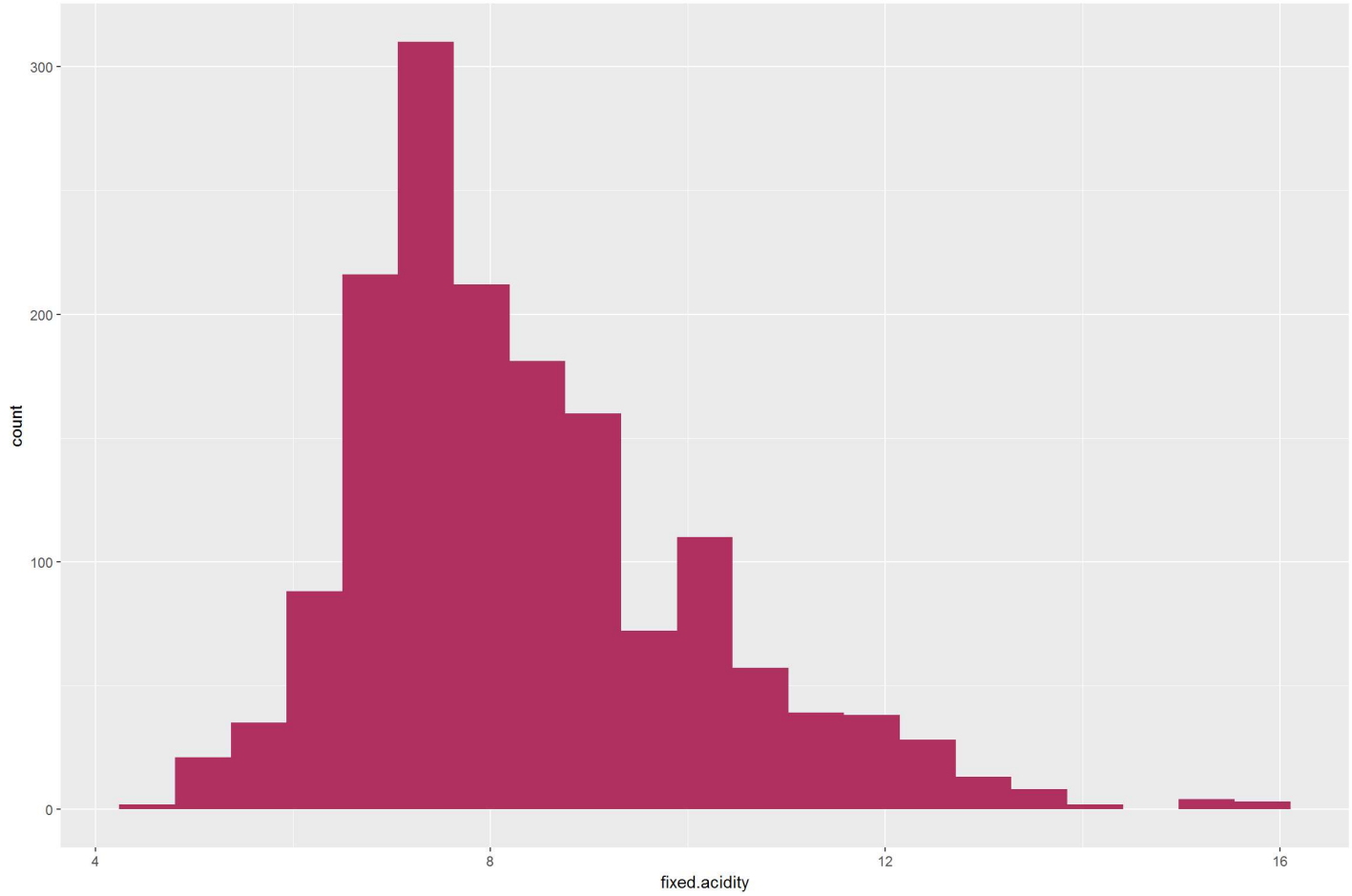
- From the graph and the summary statistics, we can infer that the distribution is almost normal with most wines having quality of 5 or 6.
- The lowest score is 3 and the highest score is 8.
- In our dataset, quality is a discrete variable while others are continuous.
- It would be easier to work with the dataset if quality was factored.

Quality - Factored



- Now that we have factored the quality variable, we can move on to other variables which affect the quality.

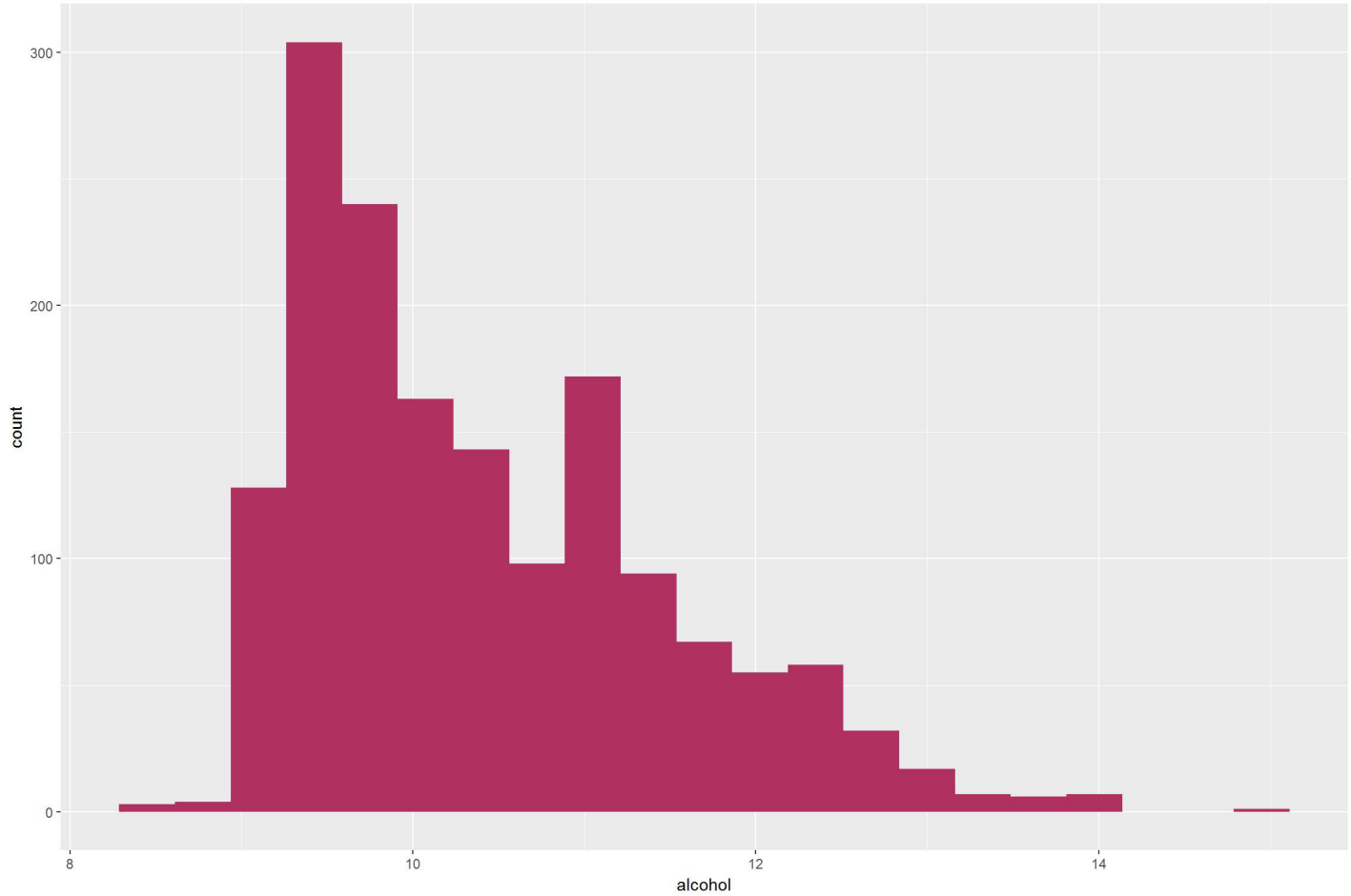
Fixed Acidity




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

- Fixed acidity peaks around 7. Some wines have values over 14 though.

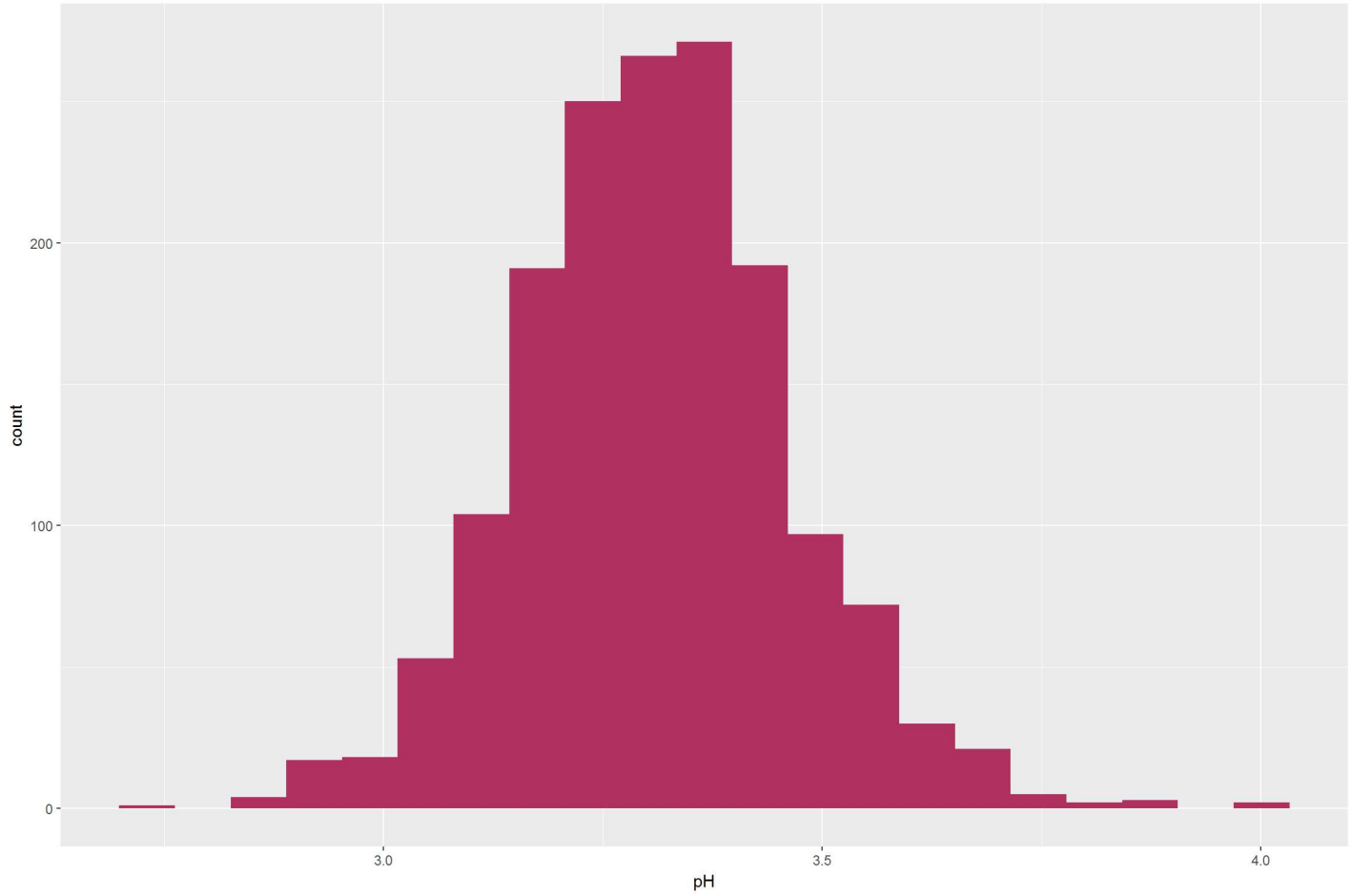
Alcohol




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

- Outlier - There seems to be an outlier with the value 14.9.
- The distribution is concentrated between 9 and 10.5

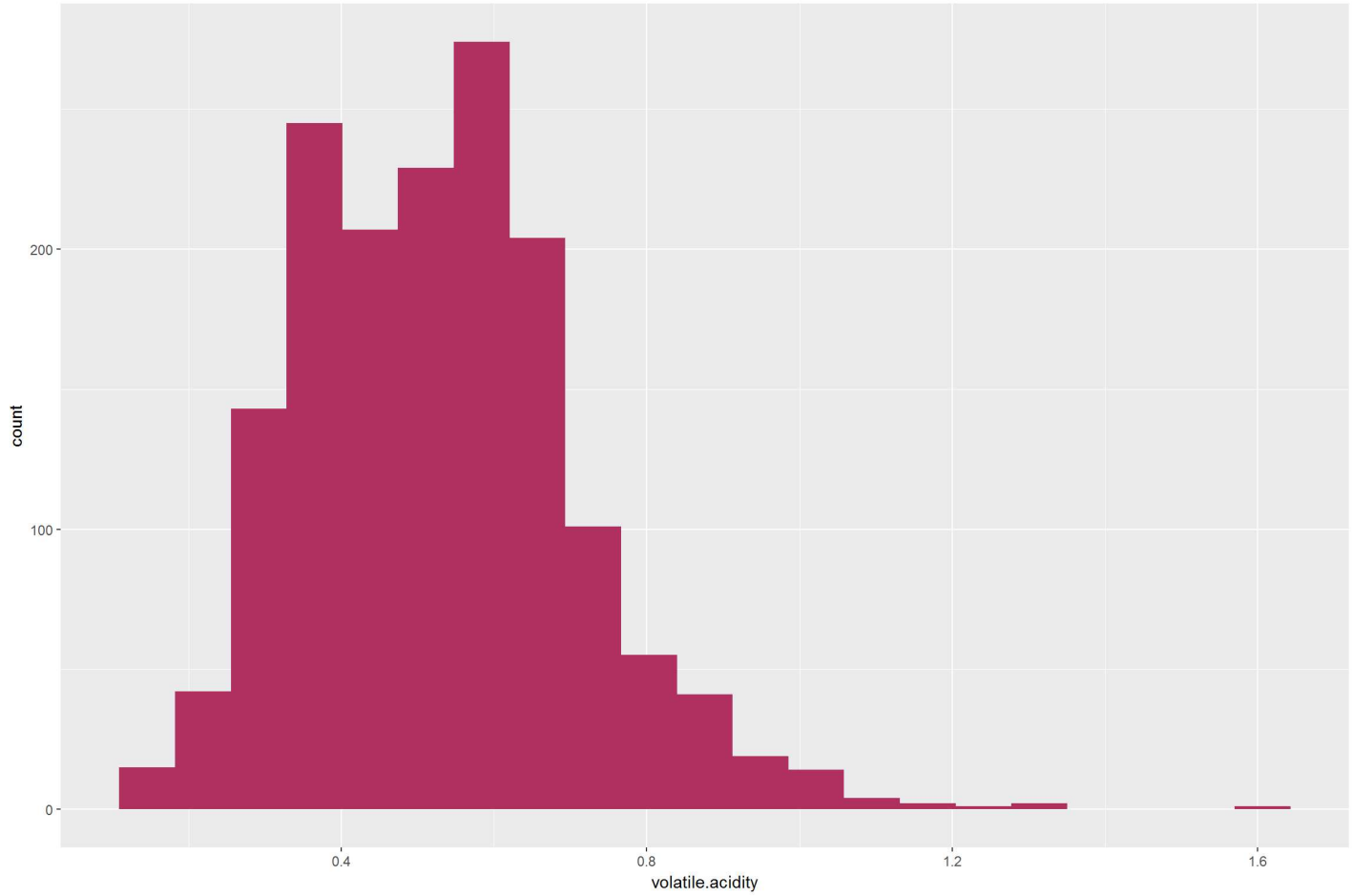
pH




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

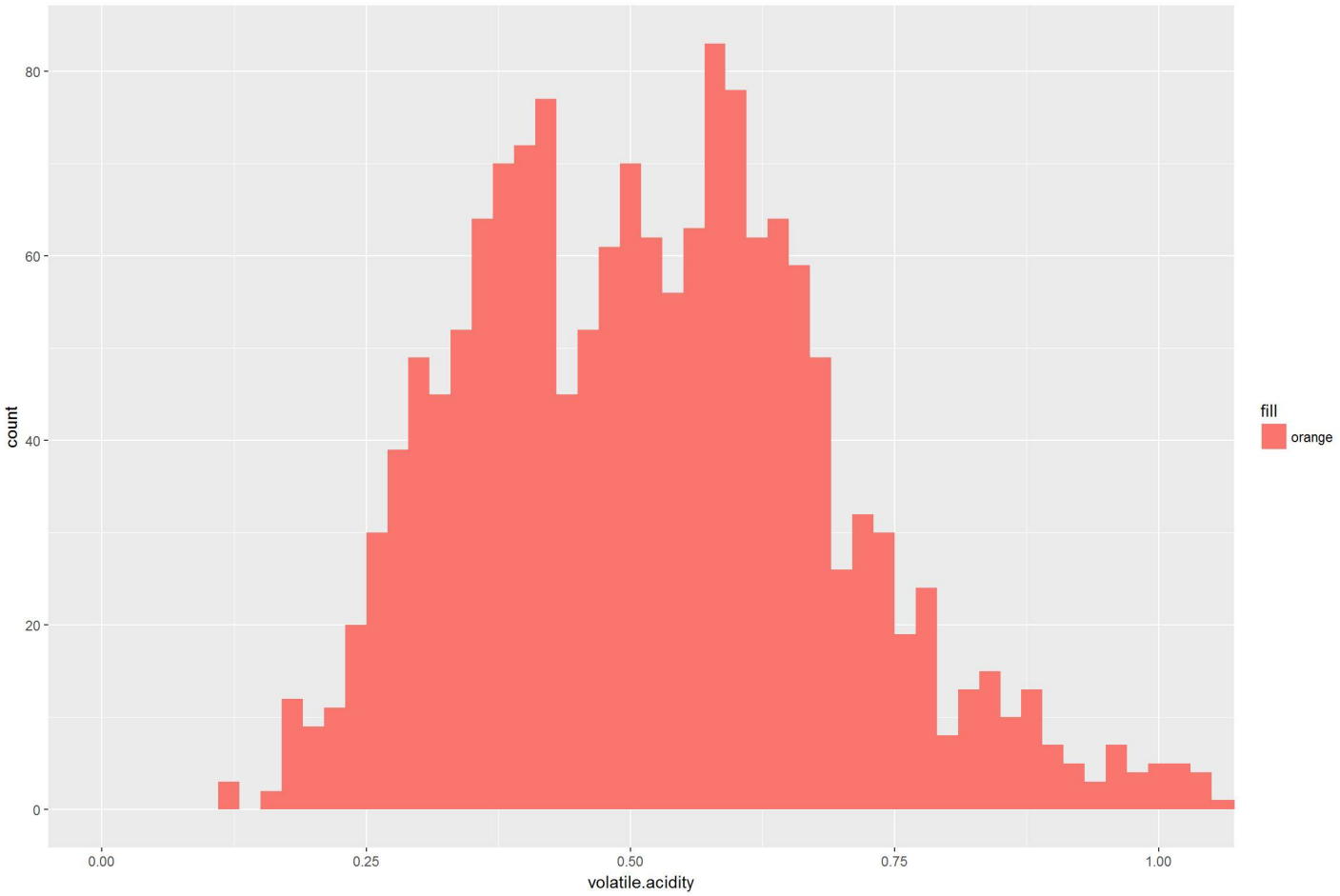
- A normal distribution is observed.
- There seem to be 2 outliers on each end of distribution.

Volatile Acidity

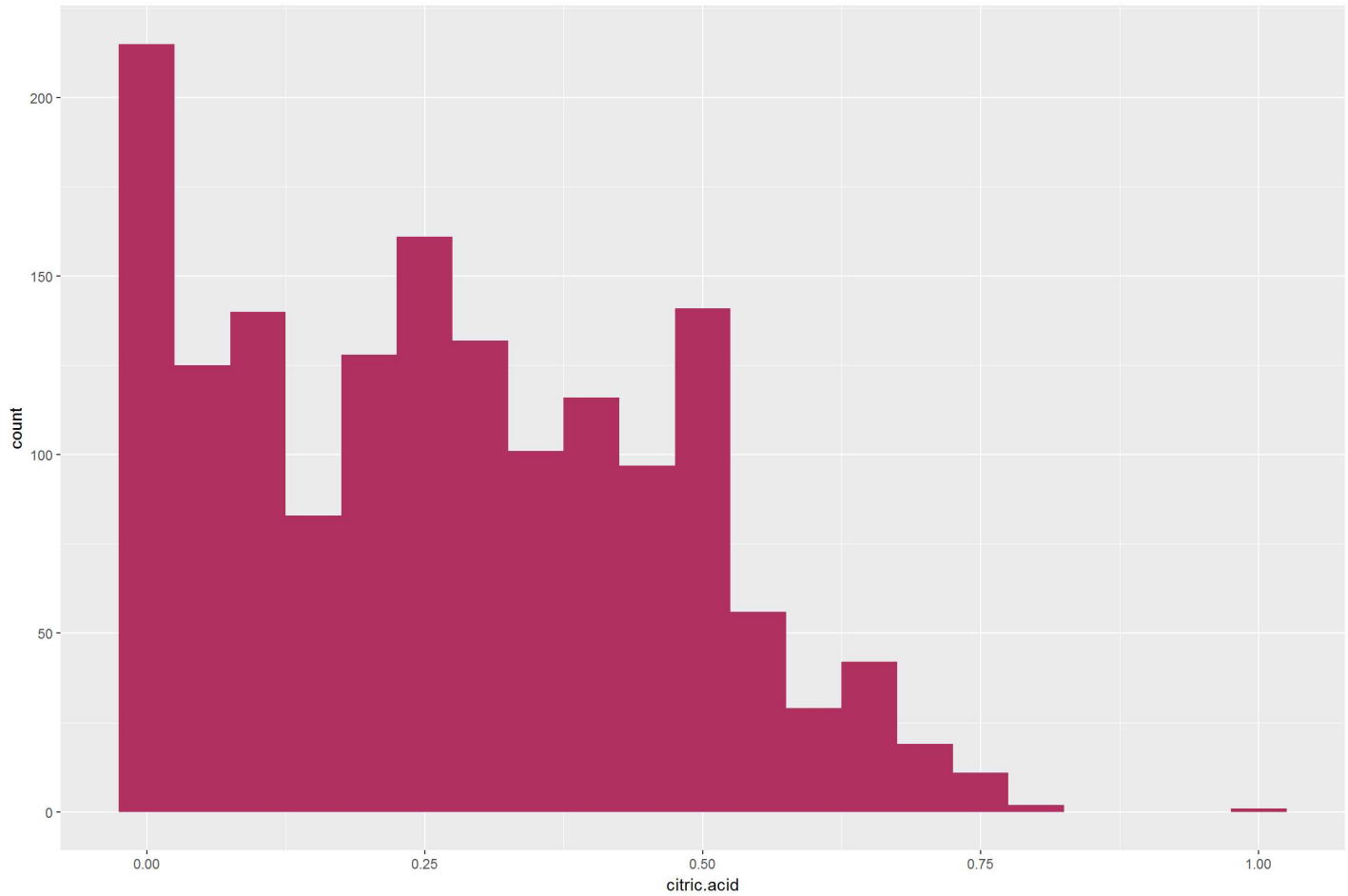



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```

- There is a big difference between the 3rd quartile and the maximum value. This may be due to outliers.
- Let's further analyze this by limiting the x-axis values.


80
60
40
20
0
count
0.00
0.25
0.50
0.75
1.00
volatile.acidity
fill
orange

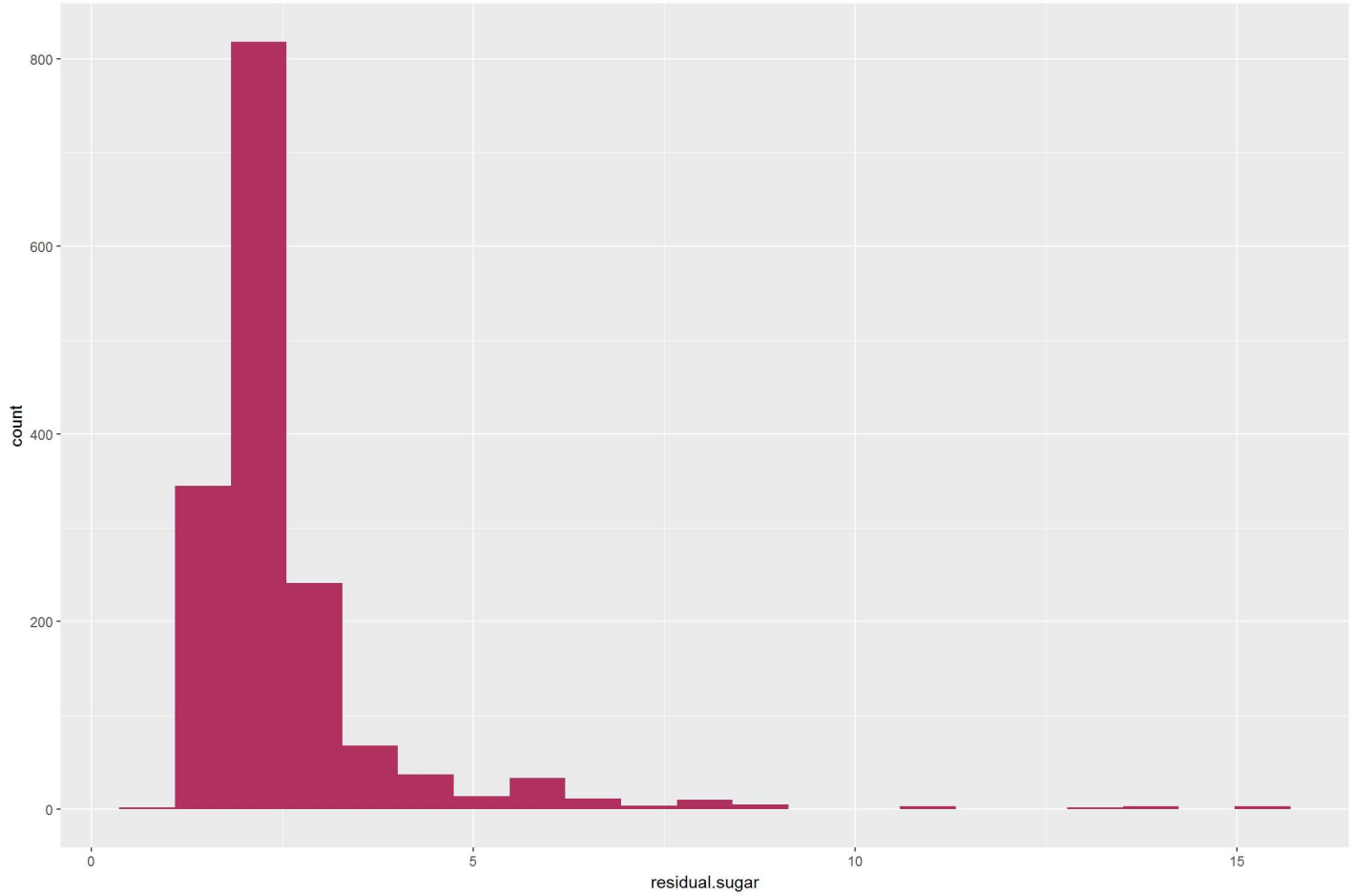
Citric Acid




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

- A left skewed distribution is observed.
- 1 outlier with the value of 1 is present.

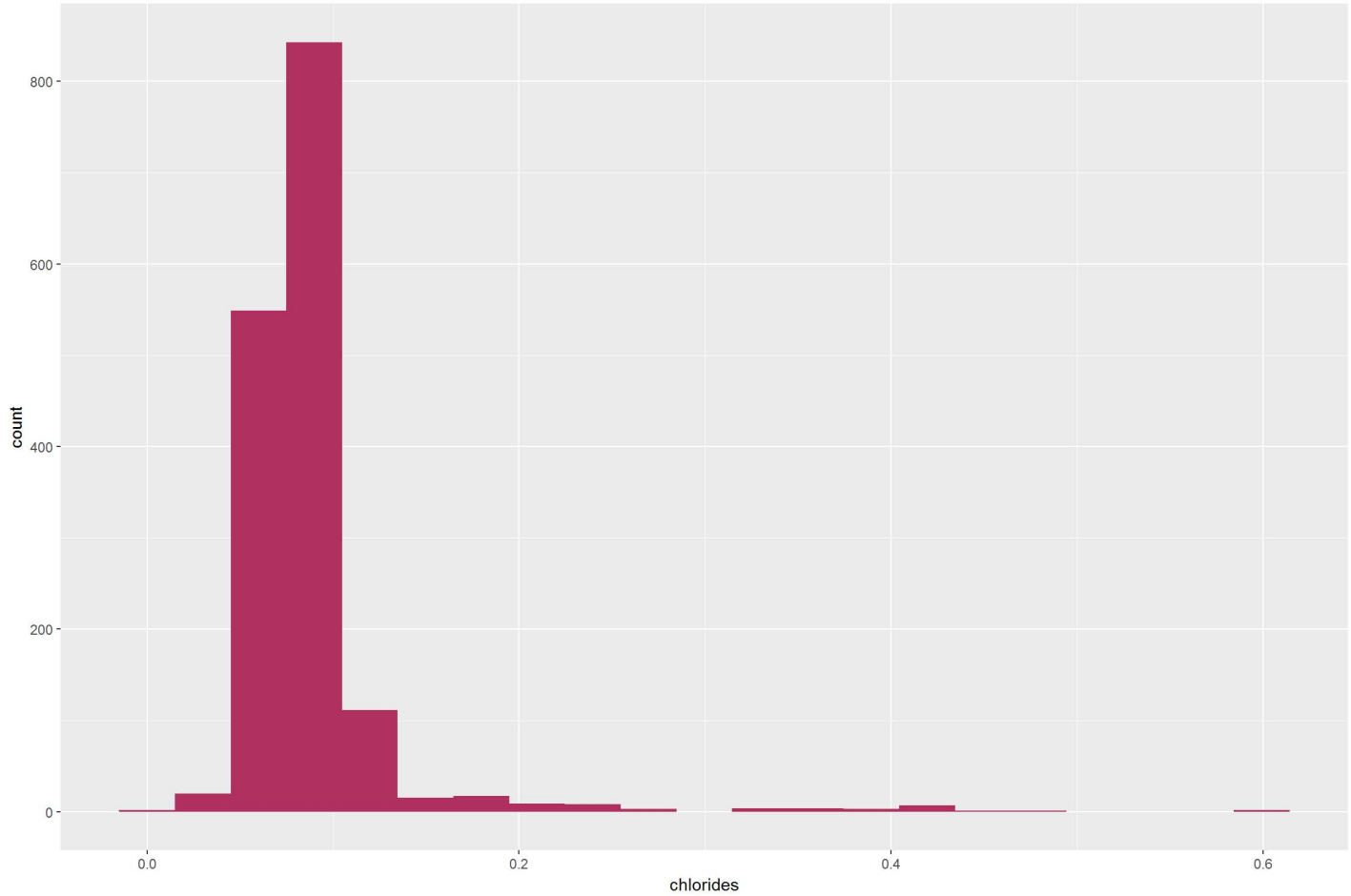
Residual Sugar




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

- Highly right skewed distribution is observed.
- Values of over 1 seem to be outliers.

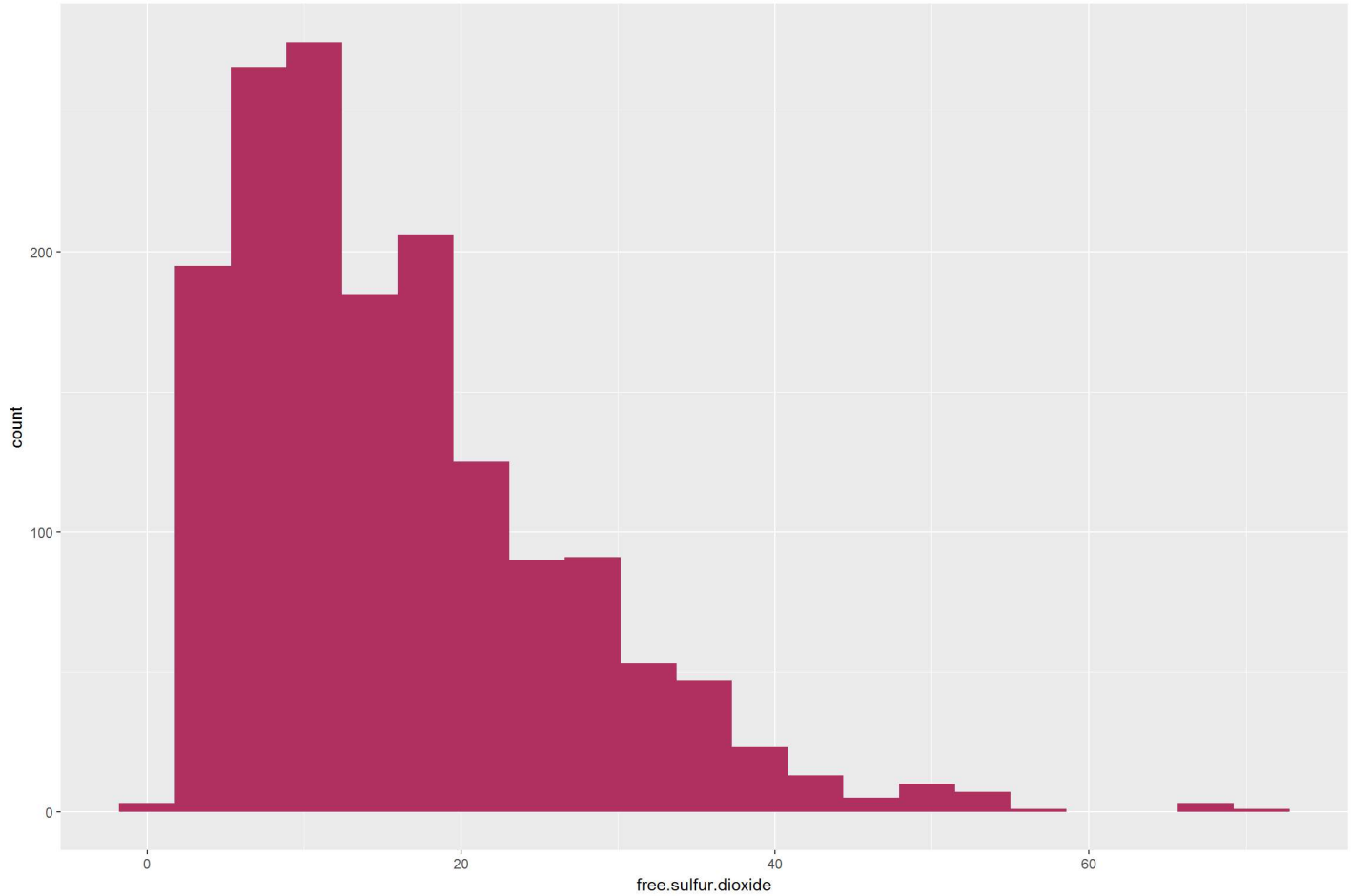
Chloride




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

- SOme outliers are present

Sulfur Dioxide

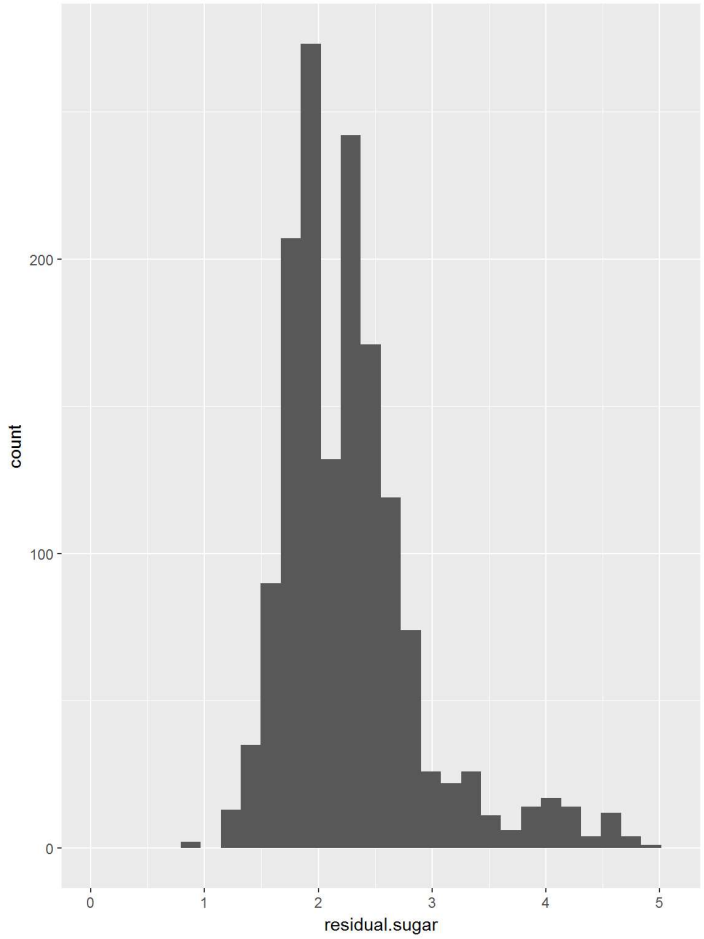


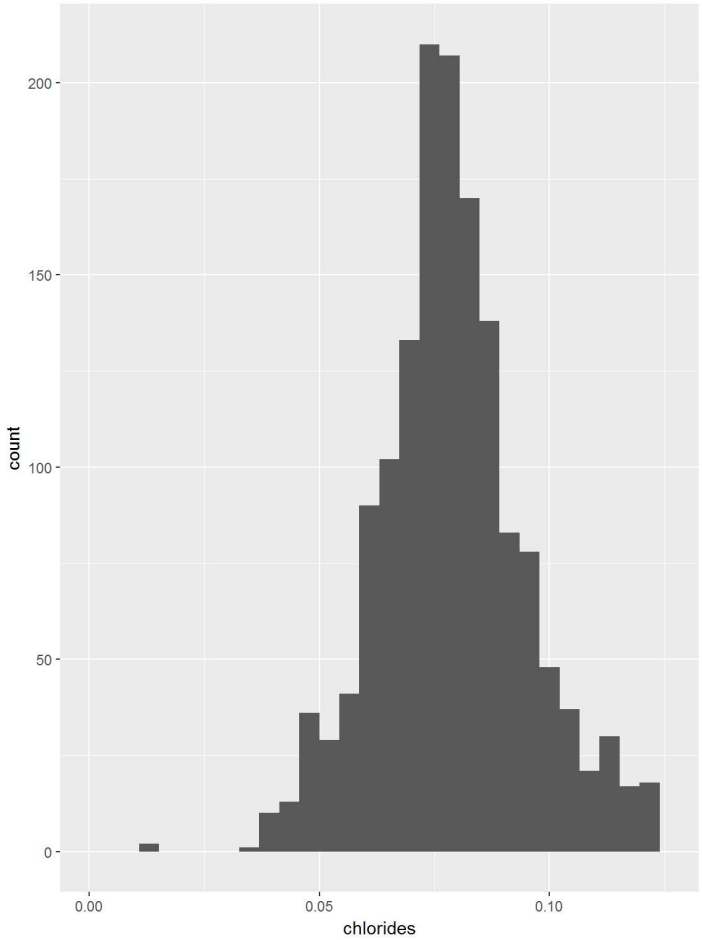

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
```

- Left skewed distribution with some outliers.

Exclude Outliers

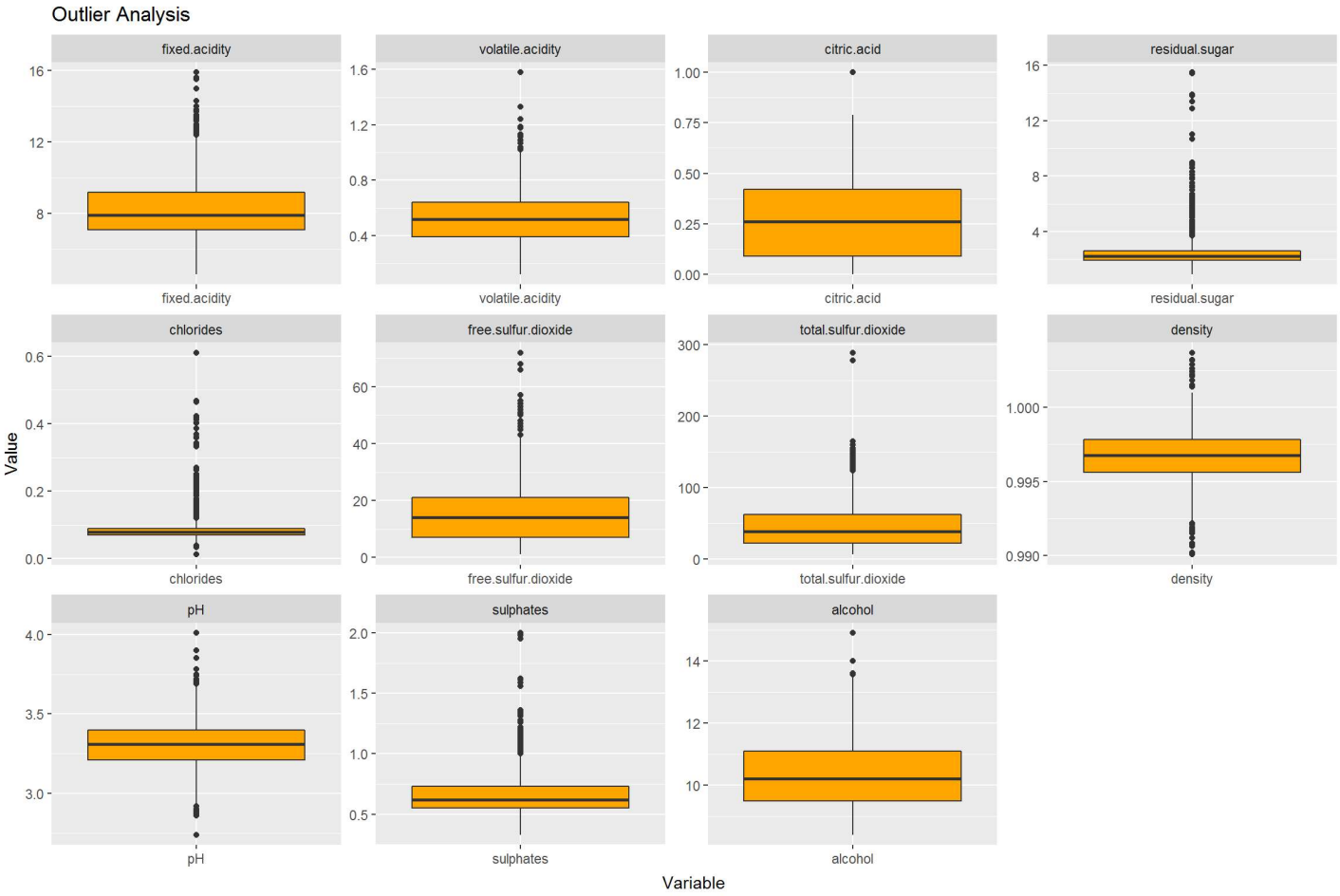
- The outliers for regular sugar and chlorides make it hard to see the shape of the distribution.
- Let's remove the 95th percentile for residual sugar and chlorides.



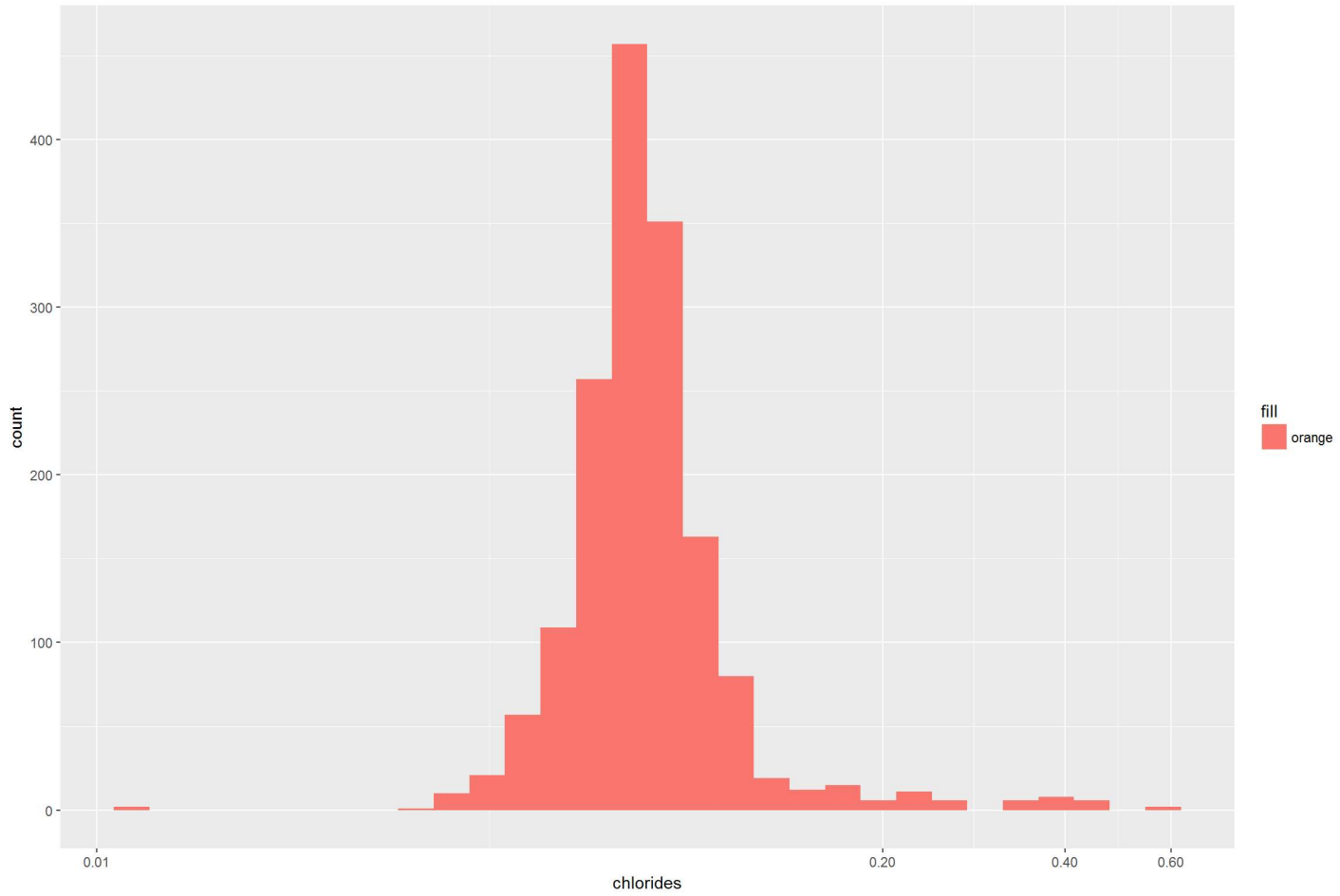


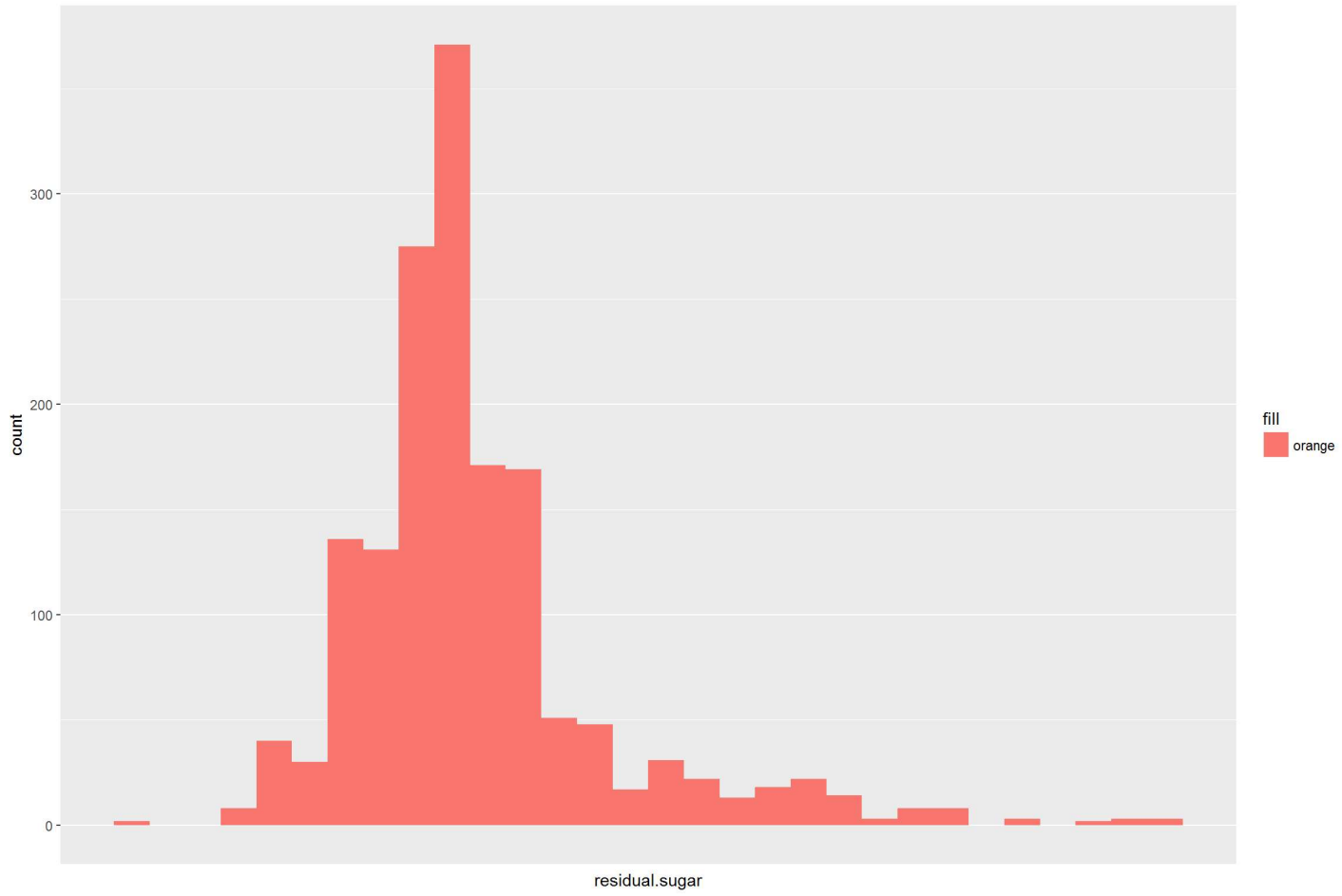
- As we can see, by removing the outliers we observe a normal distribution.

Outlier Analysis



- Outliers are present for many vairbales.
- It appears that “Residual Sugar” and “Chlorides” have the most outliers.
- I will be focussing on these 2 for outlier removal.
- Log transformation can be applied to scale the distribution from skewed to normal.


count
400
300
200
100
0
0.01
0.20
0.40
0.60
chlorides
fill
orange


count
300
200
100
0
residual.sugar
fill
orange

Univariate Analysis

What is the structure of your dataset?

The dataset contains 1599 obsevations with 12 continuous and 1 discrete variable. Input Variables: 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) 12 - X (index variable)

Output Variable: 1 - Quality (Score between 0 and 8)

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

- The main feature of interest for this dataset is how to determine quality given the input variables.
- I will also try to detect relationships between input variables.
- I will try to determine which variables influence quality.

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

These are the secondary features which may support by investigation: 1 - Alcohol 2 - Sulphates 3 - Citric Acid 4 - Volatile Acid

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

- I did not create any new variable.
- I did factor the variable quality into 3 levels for better categorization of wines based 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 form
of the data? If so, why did you do this?

- I found some unusual distributions. I fixed them by performing outlier removal.

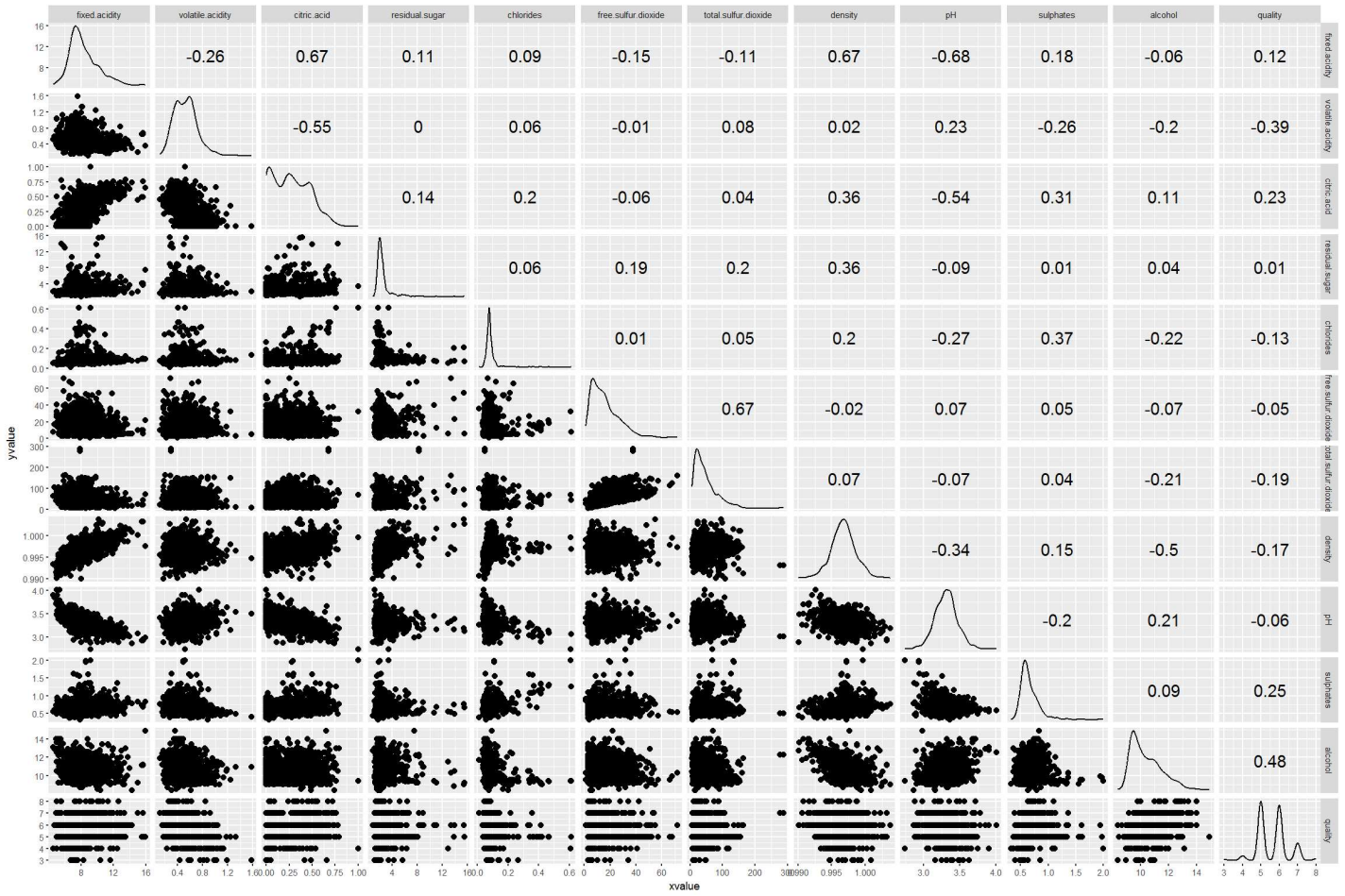
Bivariate Plots Section

Correlation Plot

- I will be using product-moment correlation to determine and visualize the relationship between 2 variables.
- The objective is to find out how variables affect quality and how one variable may affect another variable.

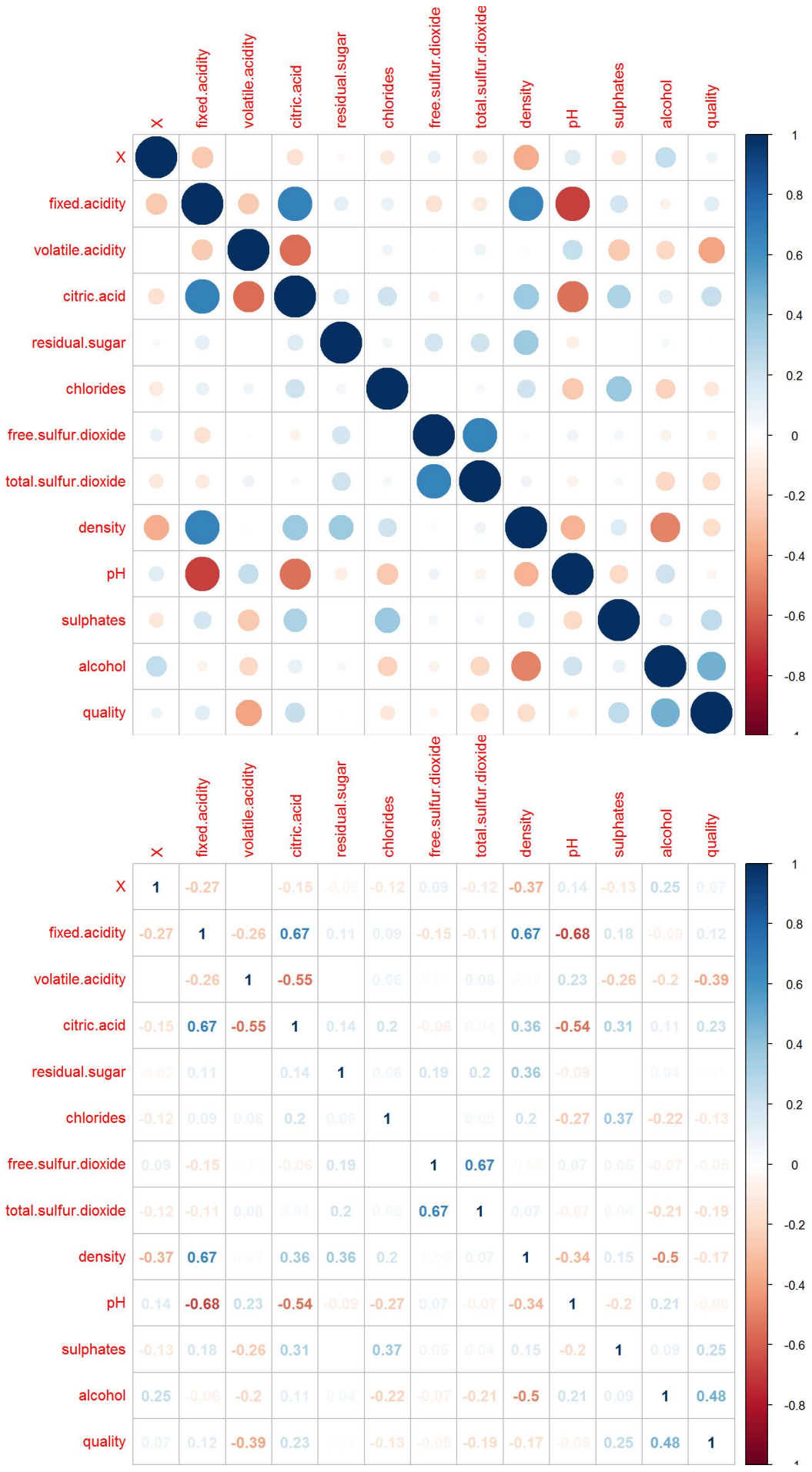
```
##  [1] "fixed.acidity"        "volatile.acidity"     "citric.acid"
##  [4] "residual.sugar"       "chlorides"            "free.sulfur.dioxide"
##  [7] "total.sulfur.dioxide" "density"              "pH"
## [10] "sulphates"            "alcohol"              "quality"
```

Correlation Table:



Create correlation table with the help of corrplot.

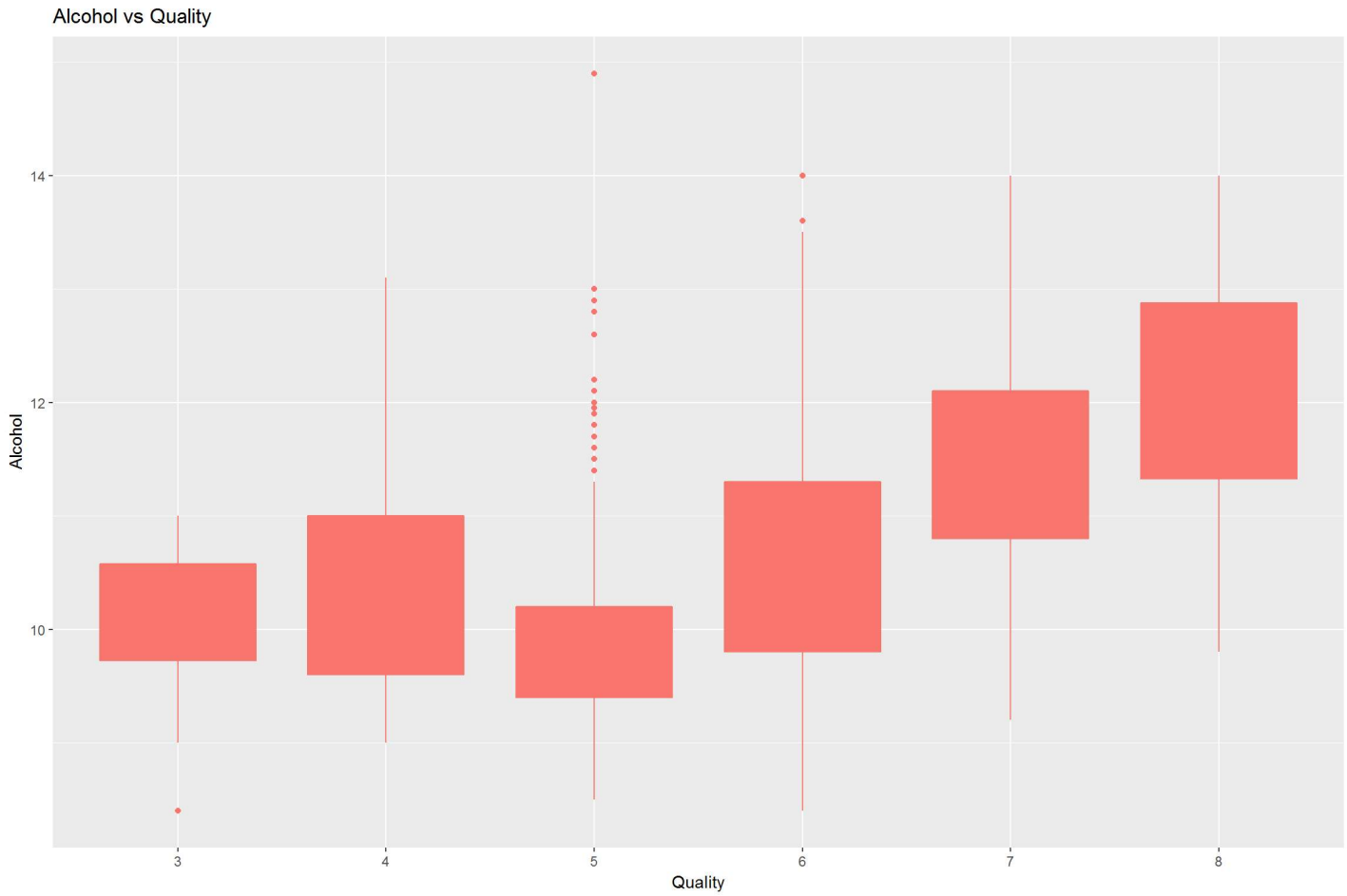
- convert digital matrix to a graph
- display data vividly
- may help to find clusters in data



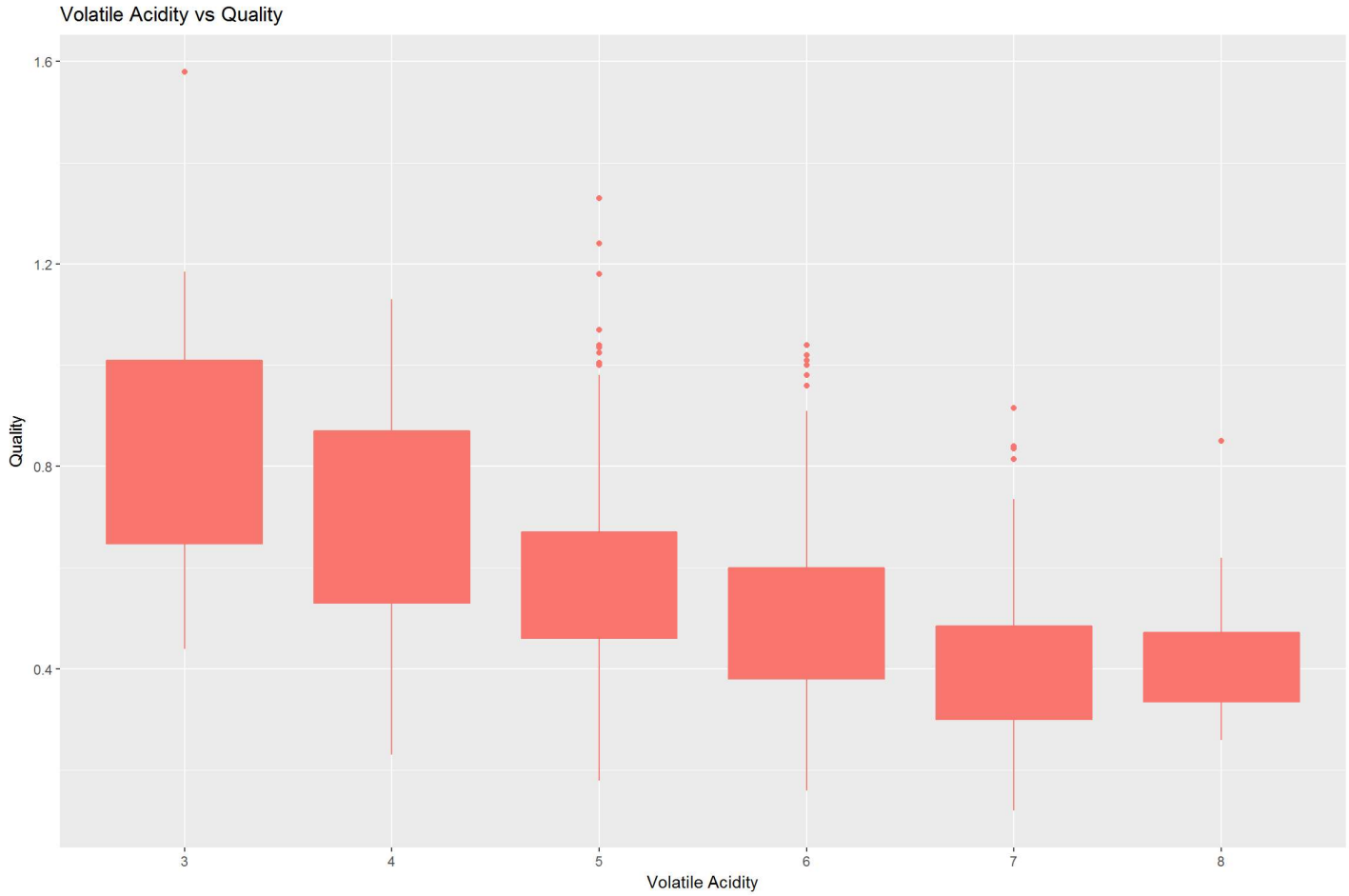
- The corrplot helps better visualize strong correlations.

- The variables with positive correlation to quality are:
  - alcohol:quality = 0.48
  - sulphates:quality = 0.25
  - citric.acid:quality = 0.23
  - fixed.acidity:quality = 0.12
  - residual sugar:quality = 0.01
- We see that alohol content and quality are closely related.
- The variables with negative correlation to quality are:
  - volatile.acidity:quality = -0.39
  - total.sulfur.dioxide:quality = -0.19
  - density:quality = -0.17
  - chlorides:quality = -0.13
- We see that volatile acidity of red wine and quality are inversely related.
- Variables with highest(+ve or -ve) correlation are:
  - fixed.acidity:citirc.acid = 0.67
  - fixed.acidity:density = 0.67
  - free.sulfur.dioxide:total.sulfur.dioxide = 0.67
  - alcohol:quality = 0.48
  - density:alcohol = -0.50
  - citric.acid:pH = -0.54
  - volatile.acidity:citirc.acid = -0.55
  - fixed.acidity:pH = -0.68

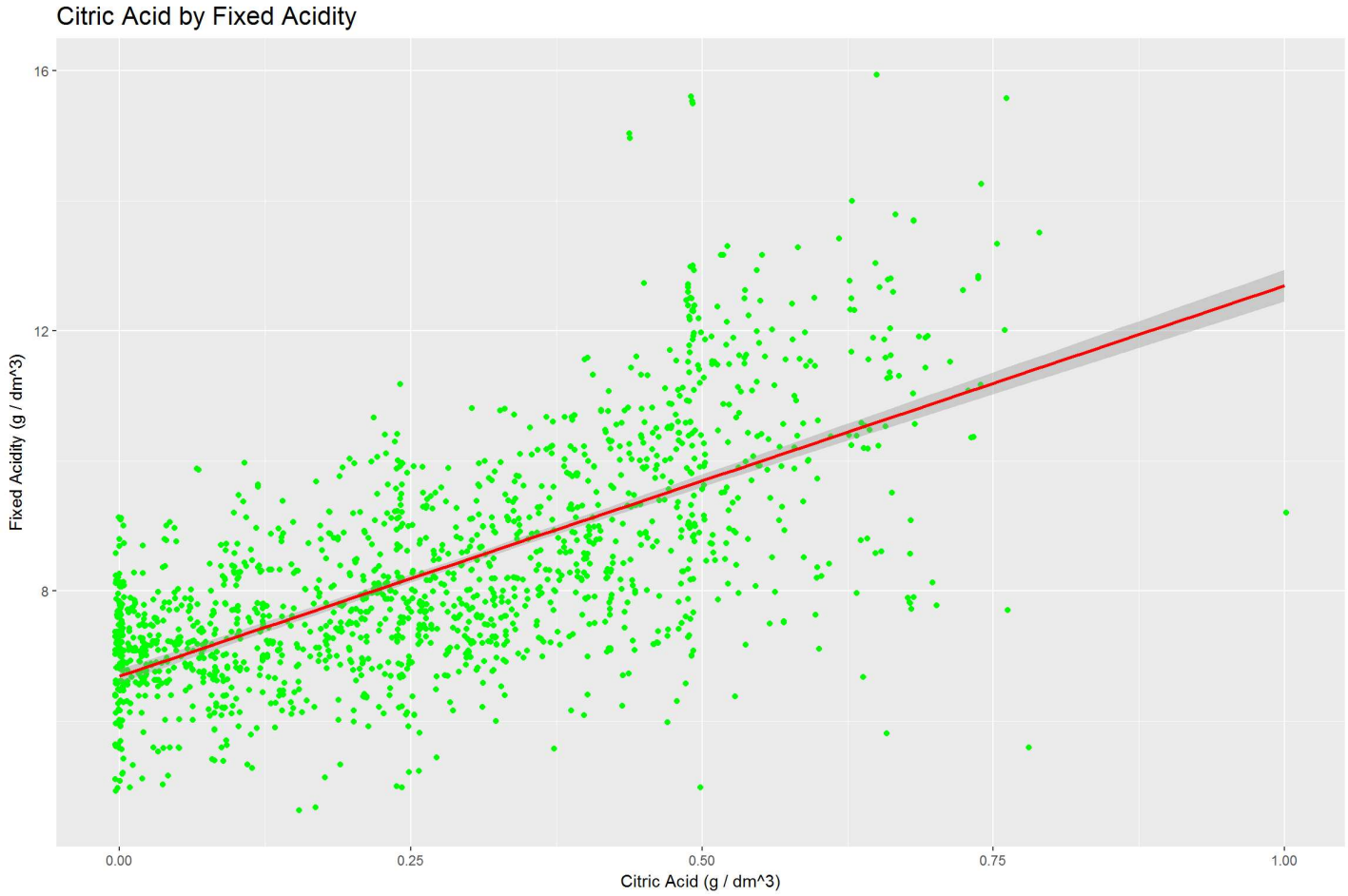
Plot Aclohol and Wine Quality



Plot Volatile acidity vs Quality



Plot Citric Acid vs Fixed Acidity

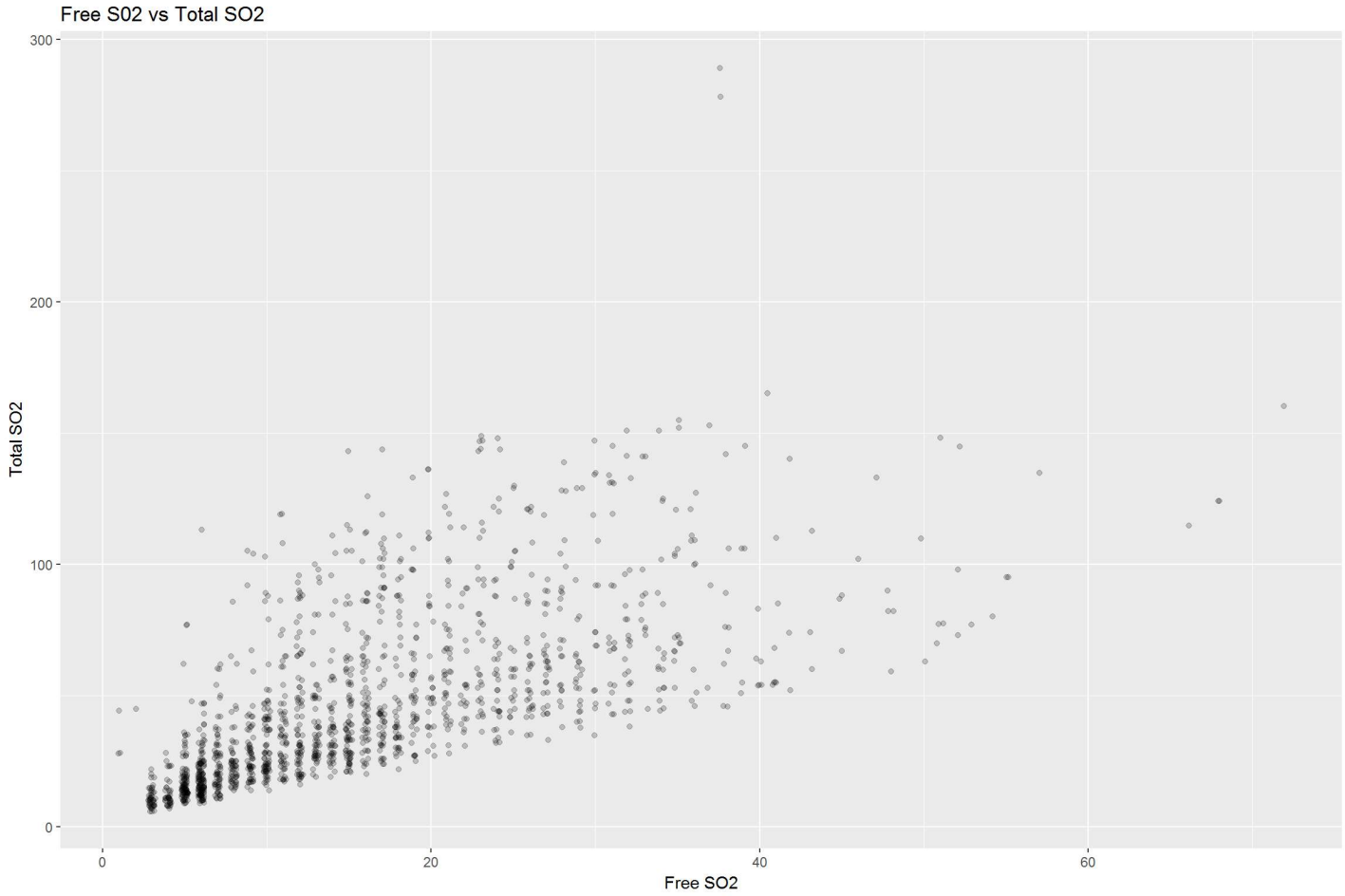


Closer look at relationships

Lets summarise wine by volatile acidity:

```
## # A tibble: 6 x 6
##   quality mean_volatile_acidity median_volatile_acidity
##     <int>                 <dbl>                   <dbl>
## 1       3             0.8845000                   0.845
## 2       4             0.6939623                   0.670
## 3       5             0.5770411                   0.580
## 4       6             0.4974843                   0.490
## 5       7             0.4039196                   0.370
## 6       8             0.4233333                   0.370
## # ... with 3 more variables: var_volatile_acidity <dbl>,
## #   std_volatile_acidity <dbl>, n <int>
```

Plot Free SO2 vs Total SO2



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?

-> Fixed Acidity - Shows positive correlation with citric acid since citirc acid is one of the fixed acid. It also shows high correlation with density. - Shows high negative correlation with pH and volatile acidity.

-> Volatile Acidity - Shows negative correlation with quality and citric acid.

-> Density - Shows high negative correlation with alcohol, acidity and pH.

-> Free sulphur dioxide - High positive correlation with total sulphur dioxide and very little correlation with sulphates.

-> Quality - Positive correlation with alcohol and negative correlation with volatile acidity.

```
    -> Quality shows negative correlation with volatile acidity.

    -> It looks like wine with higher volatile acidity has better quality as well.

    -> We can infer from the QUality vs Volatile QUality graph that as the volatile acidity decr
eases, the wine quality increases.

    -> The strongest correlation between Alcohol and Quality can be observed. It appears that, w
ine
    with high alcohol content tend to have good quality. Similarly, wines with low and medium al
ochol
    content have low and medium wine quality respectively.
```

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

- Yes, I observed a positive correlation for the following:
  - fixed.acidity:citirc.acid = 0.67
  - fixed.acidity:density = 0.67
  - free.sulfur.dioxide:total.sulfur.dioxide = 0.67
- Free SO2 vs Total SO2 graph validates the positive correlation between the 2 variables.

What was the strongest relationship you found?

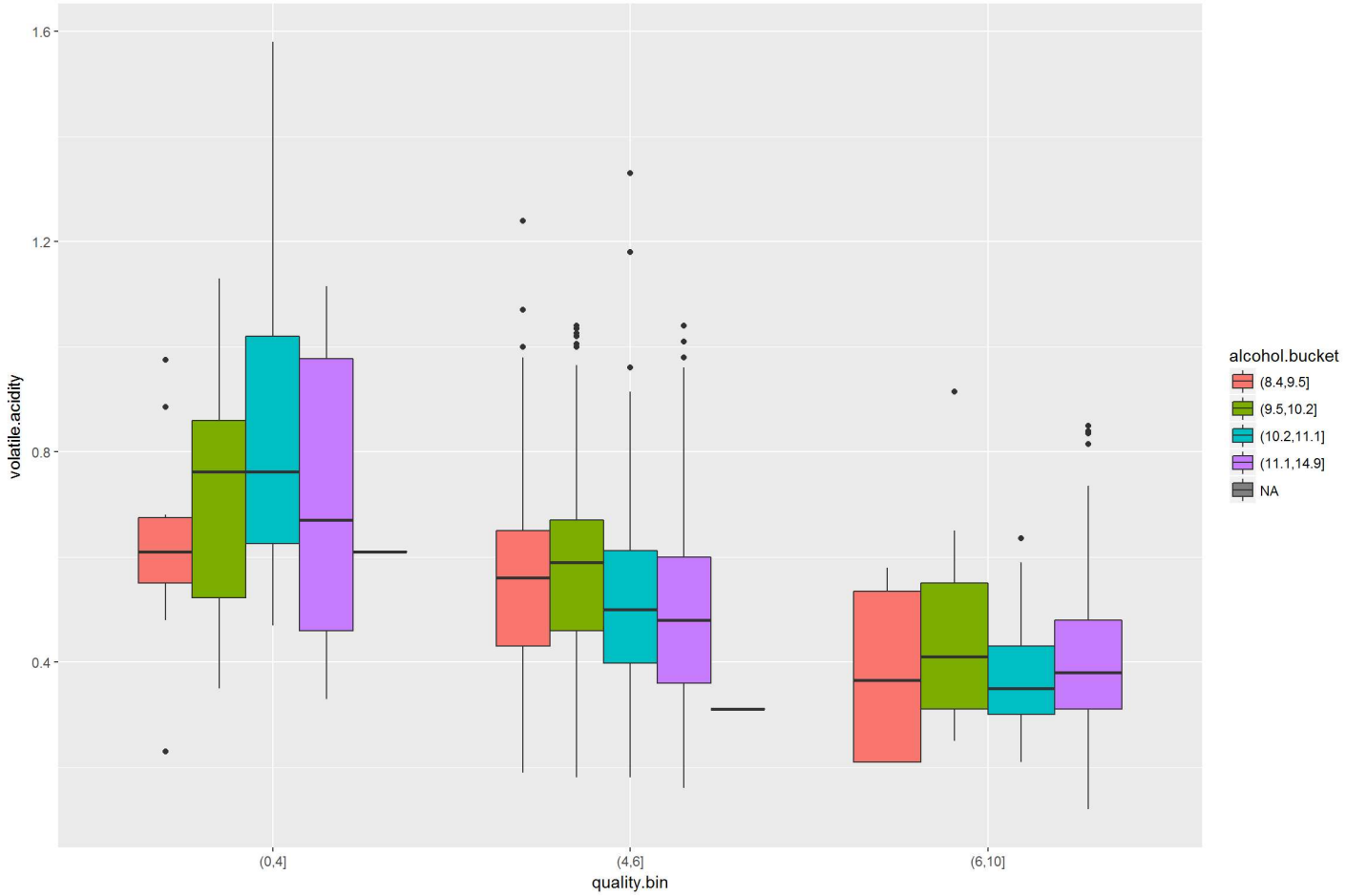
- Strongest relationship was found between: POsitive: - fixed.acidity : citirc.acid = 0.67 - fixed.acidity : density = 0.67 - free.sulfur.dioxide : total.sulfur.dioxide = 0.67 Negataive: - fixed.acidity : pH - volatile.acidity : citric.acid

Multivariate Plots Section

Quality takes in numerical values which range between 3-8. TO make our analysis better, it

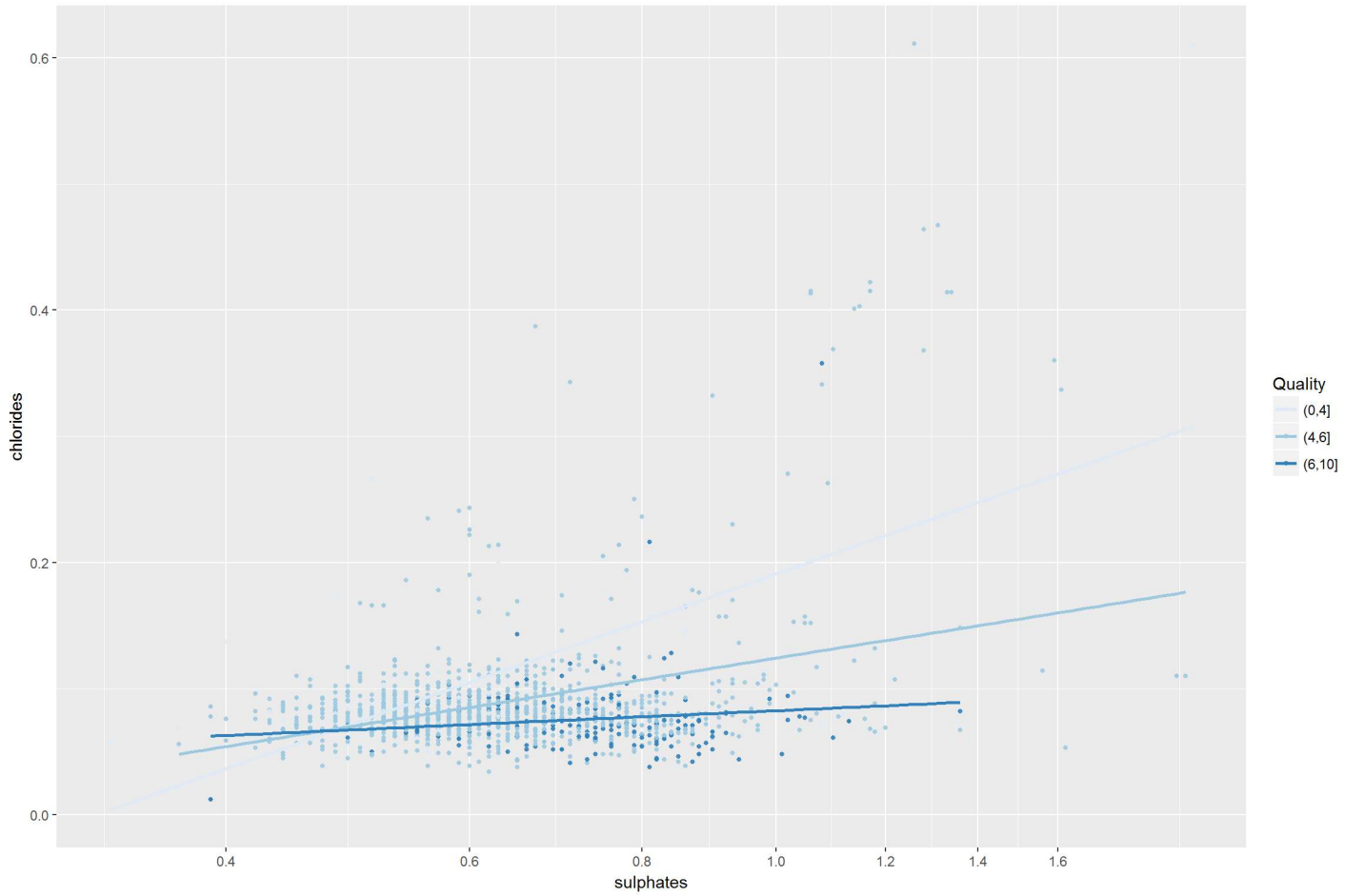
is useful to factor quality. ### Let’s convert quality into discrete variable:

Plot the effect of volatile.acidity and alcohol on quality



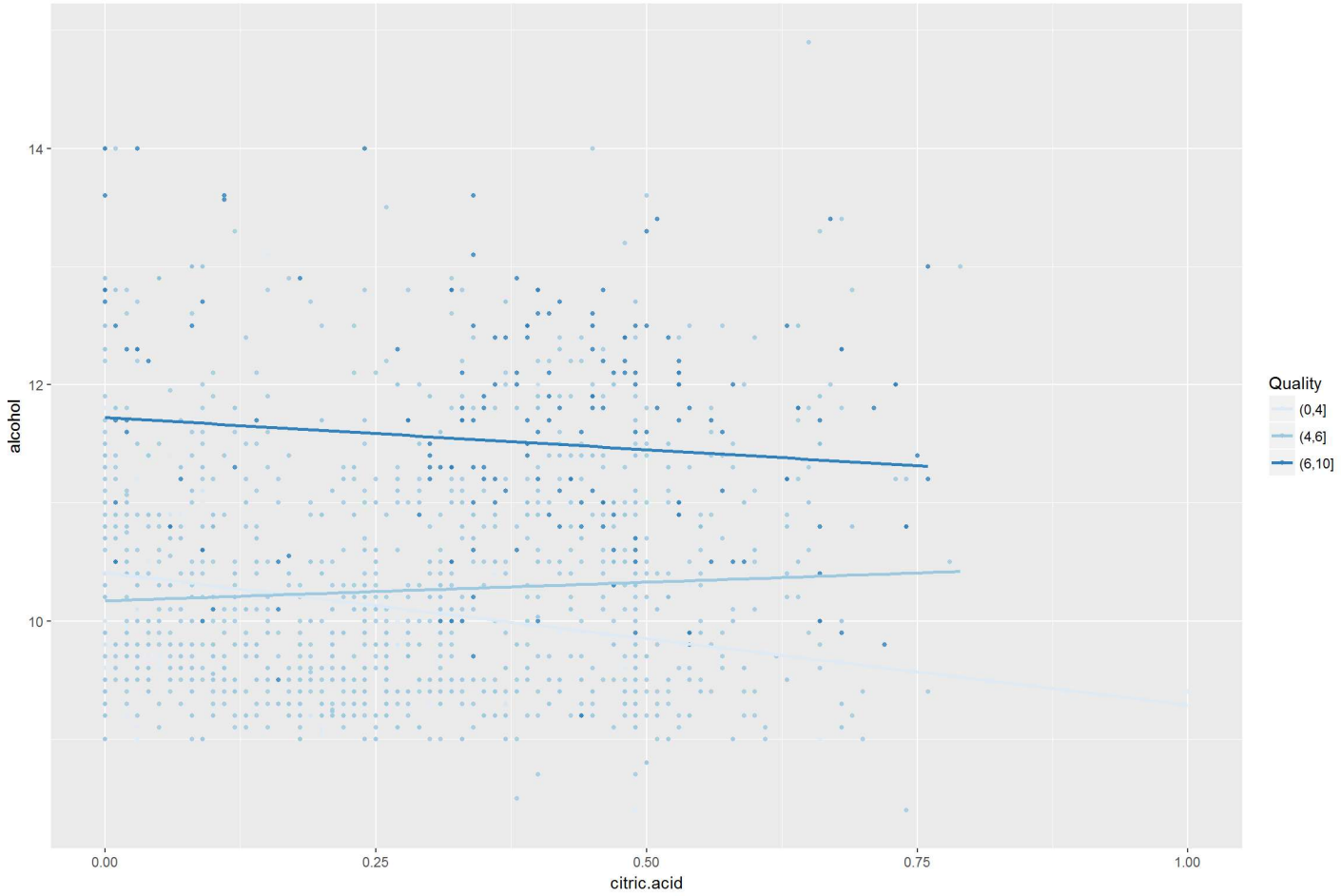
- It looks like wines with low to medium volatile acidity and medium to high alcohol content have better quality.
- Wines with high volatile acidity definitely have poor quality.

Plot the effect of chlorides and sulphates on quality



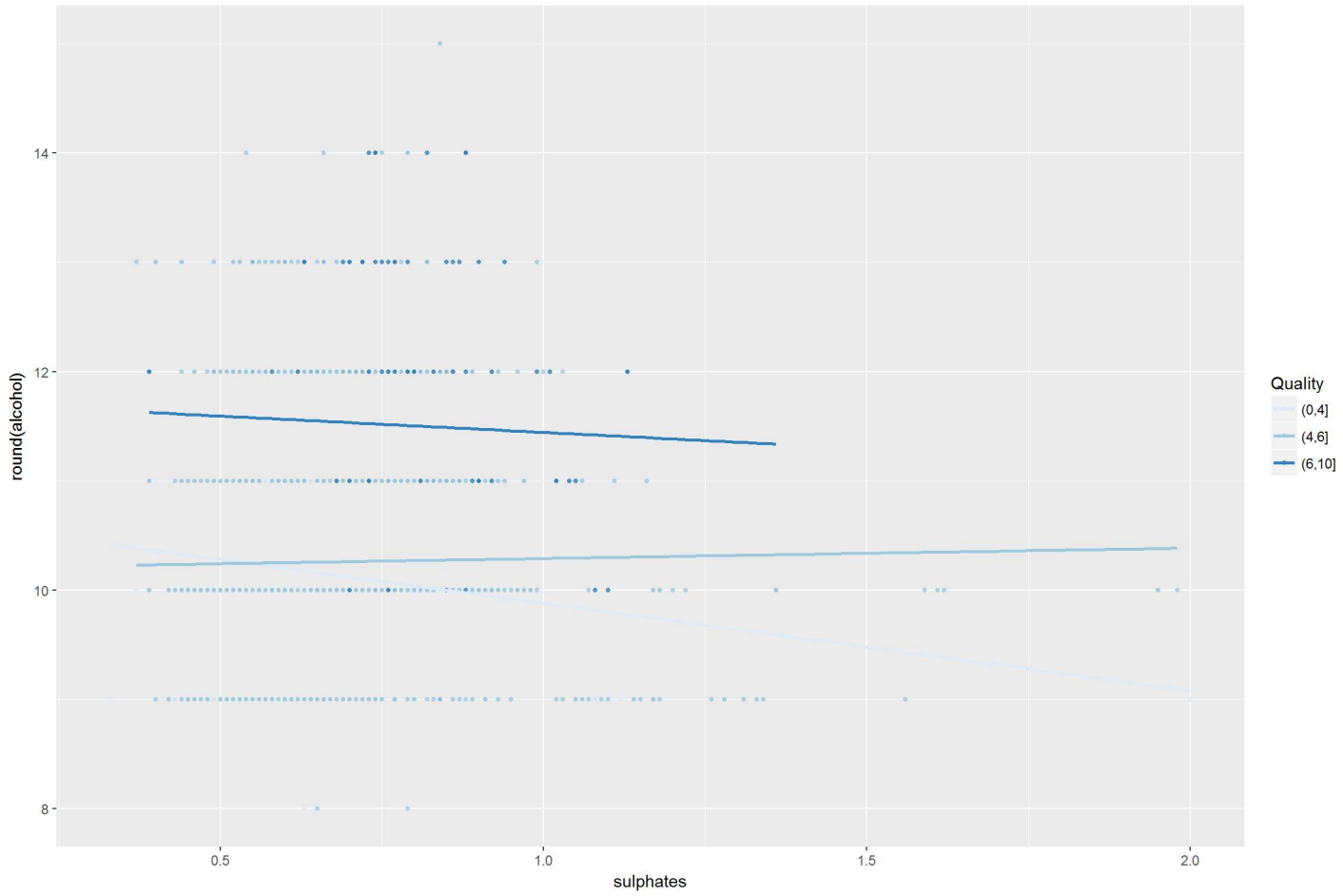
- The data points for medium quality wine is all over the graph. It is hard to determine a relationship here.
- It can be said thoug, if a wine has very lowe chloride and sulphate content then it would have poor quality as well.
- High quality wines sees to have medium suplhate and low choloride content.

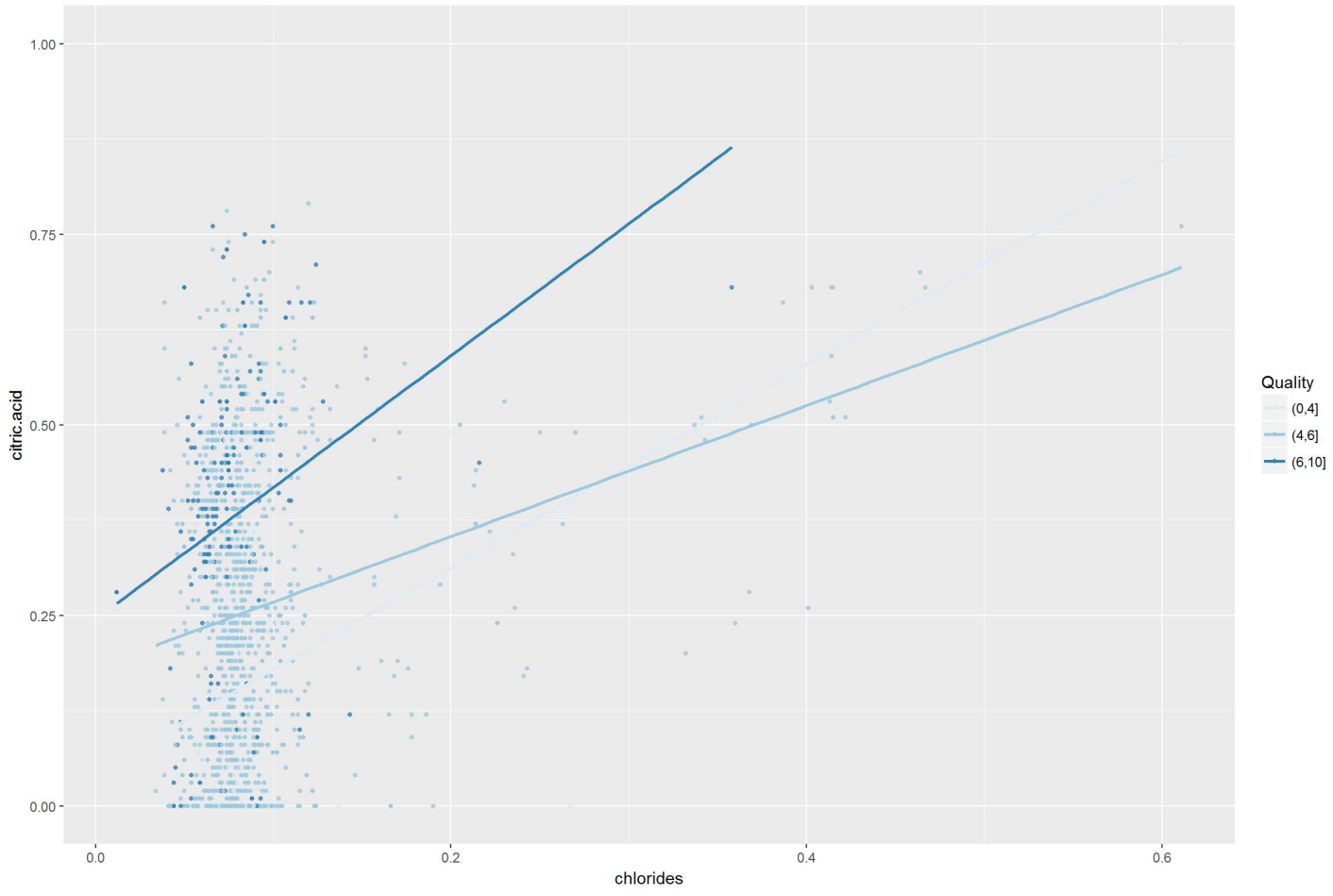
Plot affect of citric acid and alcohol on quality



- High quality wines have high amount of alochol and low amount of citric acid.
- Medium quality wine distribution is all over the place. This may suggest that some other variable plays a role.
- Low quality wine points are also spread out.

Plot the effect of sulphates and alcohol


round(alcohol)
14
12
10
8
0.5
1.0
1.5
2.0
sulphates
Quality
(0,4]
(4,6]
(6,10]


citric.acid
1.00
0.75
0.50
0.25
0.00
0.0
0.2
0.4
0.6
chlorides
Quality
(0,4]
(4,6]
(6,10]

- High quality wines have high alcohol and medium sulphate value.
- Medium quality wine have low alochol and low sulphate value.
- Low quality wines typically have low alochol and a very low suplhate value.

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?

- For high quality wine, volatile acidity is low and alcohol percentage is high.
- High quality wines have low sulphates and chloride values.
- High quality wines have higher alcohol percentage and lower citric acid values.
- High quality wines tend to have higher alcohol content than sulphate(medium-high).
- High chloride and citric acid values reduce the quality of the wine.

Were there any interesting or surprising interactions between features?

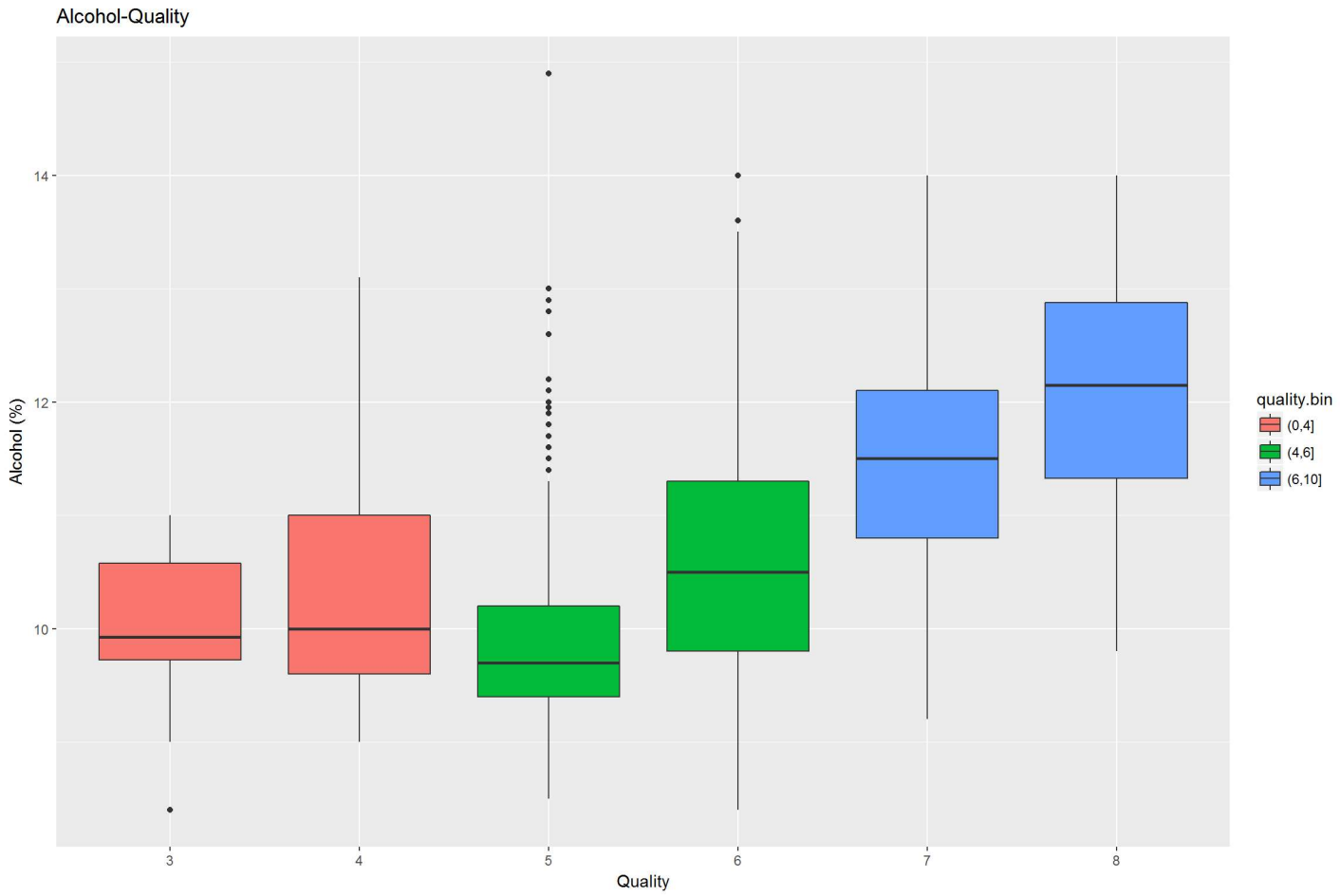
- No one variable affects the quality of the wine.
- Usually, higher alcohol content has high quality.
- Wines with medium-high sulphate content, medium-high citric quality and low-medium volatile acidity have good wine quality.

Final Plots and Summary

Plot 1

Alcohol and Quality have strong relationship. Let’s demonstrate this:

- By individual values:



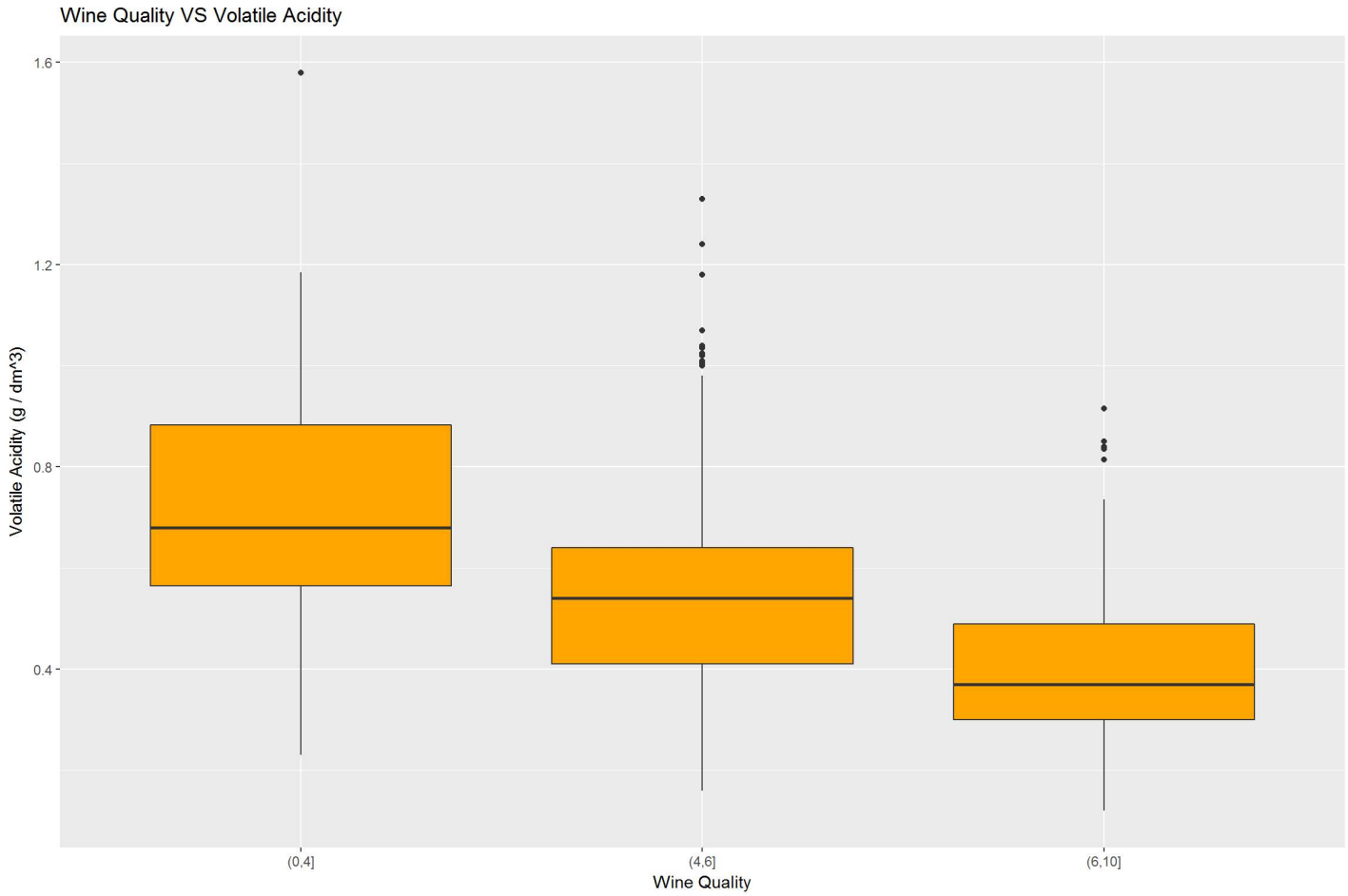
- Summary Statistics:

```
## wine$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.400   9.725   9.925   9.955  10.580  11.000
## ------------------------------------------------------------
## wine$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00    9.60   10.00   10.27   11.00   13.10
## ------------------------------------------------------------
## wine$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     8.5     9.4     9.7     9.9    10.2    14.9
## ------------------------------------------------------------
## wine$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.80   10.50   10.63   11.30   14.00
## ------------------------------------------------------------
## wine$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.20   10.80   11.50   11.47   12.10   14.00
## ------------------------------------------------------------
## wine$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.80   11.32   12.15   12.09   12.88   14.00
```

Description 1

- I talked above how alcohol affects wine quality. It can be seen here that quality range 6-10 has higher median quality value versus other.
- There is one exception to the alcohol %. Wines with quality 5 tend to have lower alcohol percentage than wines with quality 3 and 4. This can be further explored.

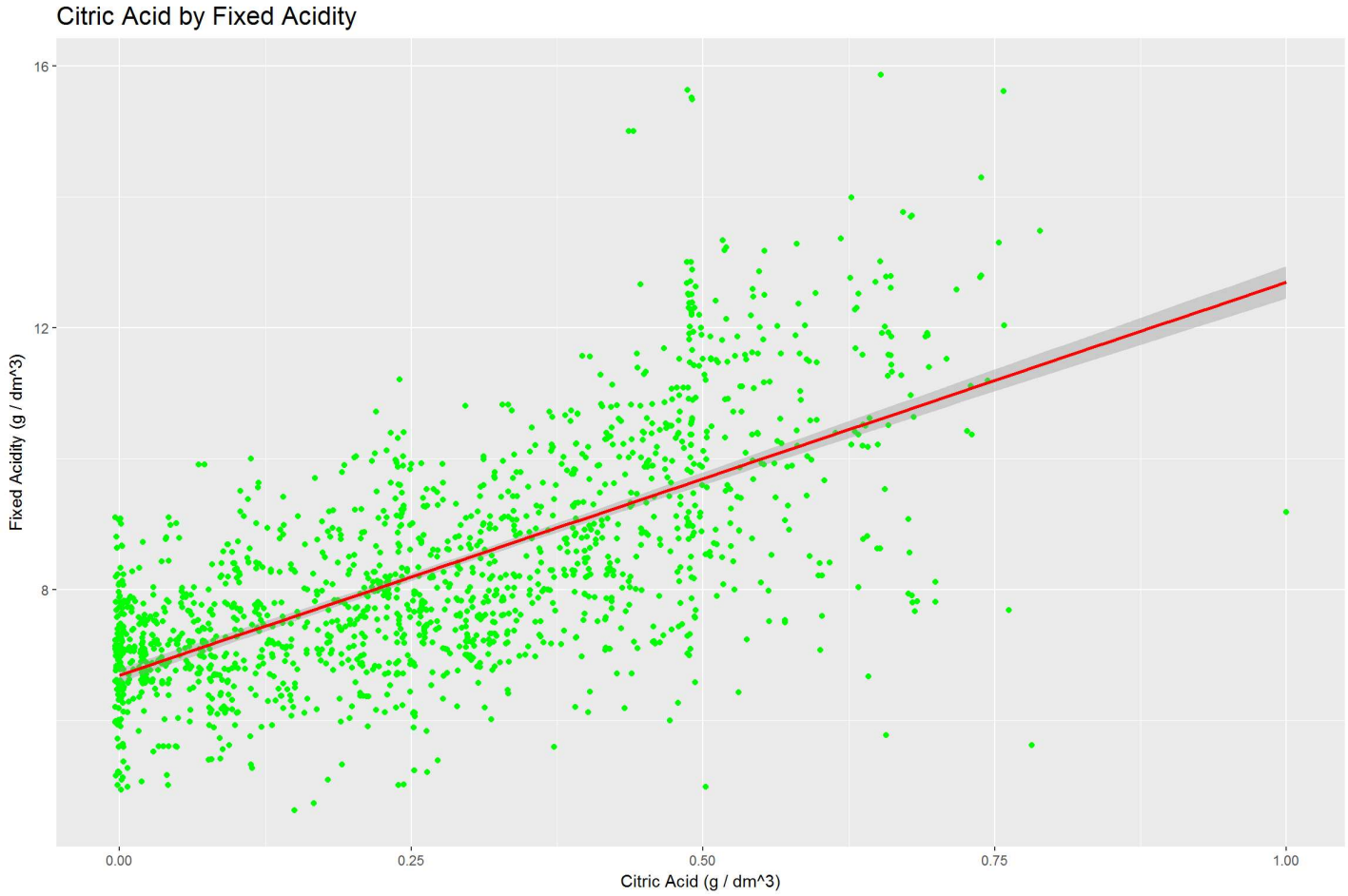
Plot 2



Description 2

- High quality wines usualyy have low values of volatile acidity.
- Medium quality wines typically have medium value of volatile acidity. But many medium quality wines have high value of volatile acidity as well.
- Low quality wines have medium to high value for volatile acidity.

Plot 3



Description 3

- These two variables demonstrate high positive correlation.
- High quality wines have high amount of alochol and low amount of citric acid.
- The above mentioned point may suggest that high quality wines may have low critric acid and fixed acidity value.

Reflection

This was truly a very good case study to explore and understand univariate, bivariate and multivariate analysis. I started my exploration by first understanding the nature of the data and what values variables hold. Once I started exploring variables one at a time, I became more comfortable and intrigued with the dataset. Exploring one variable helped me develop a better understanding of the data. It was interesting to find out if one single variable can truely affect the quality of wine.

To better understand quality of wines, I factored quality into discrete variable by creating quality.bin. This helped in better categorization of wines. Then I moved on to bivariate analysis where I explored if any 2 variables have an effect on the quality of wine together.

I used the help of a correlation matrix to determine which variables I want to explore. I focussed my attention on the strong correlationship variables.

After performing bivariate analysis, I determined that wines with higher alcohol content in conjunction with medium-high sulphate content usually have higher quality.

TO further explore the relationships between variables, I performed multivariate analysis, where I assessed the main feature quality against 2 variables at a time. This helped determine if 2 variables in conjunction can have an effect on quality. I observed that volatile acidity and alcohol can help determine quality of the wine to an extent.

Multivariate analysis helped explore interesting relationships among variables.

The main limitations of the dataset is its age - the dataset is quite old. Also, the dataset is not global - it is limited to a region.

Future scope of this analysis can have: - Predictive modellinng of data where a model is fed 2 variables and quality is determined - Determining the set of variables which would help make a good predictive model. - The data can be extended to include more acedic values.