Explore White Wine Quality

By Ruchita Maheshwary

In this report, I analyzed the data set for around 5000 white wines.

This data set is used to model wine preferences by data mining from physicochemical properties. In the dataset, the inputs include objective tests (e.g. PH values) and the output is based on sensory data (median of at least 3 evaluations made by wine experts). Each expert graded the wine quality between 0 (very bad) and 10 (very excellent).

Brief description about the dataset variables:

- fixed acidity: most acids involved with wine or fixed or nonvolatile (do not evaporate readily)
- volatile acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
- citric acid: found in small quantities, citric acid can add ‘freshness’ and flavor to wines
- 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
- chlorides: the amount of salt in the wine
- 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
- 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
- density: the density of water is close to that of water depending on the percent alcohol and sugar content
- 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
- sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant
- alcohol: the percent alcohol content of the wine
- quality (score between 0 and 10): Output variable (based on sensory data)

Univariate Plots Section

Below is a brief summary of the data:

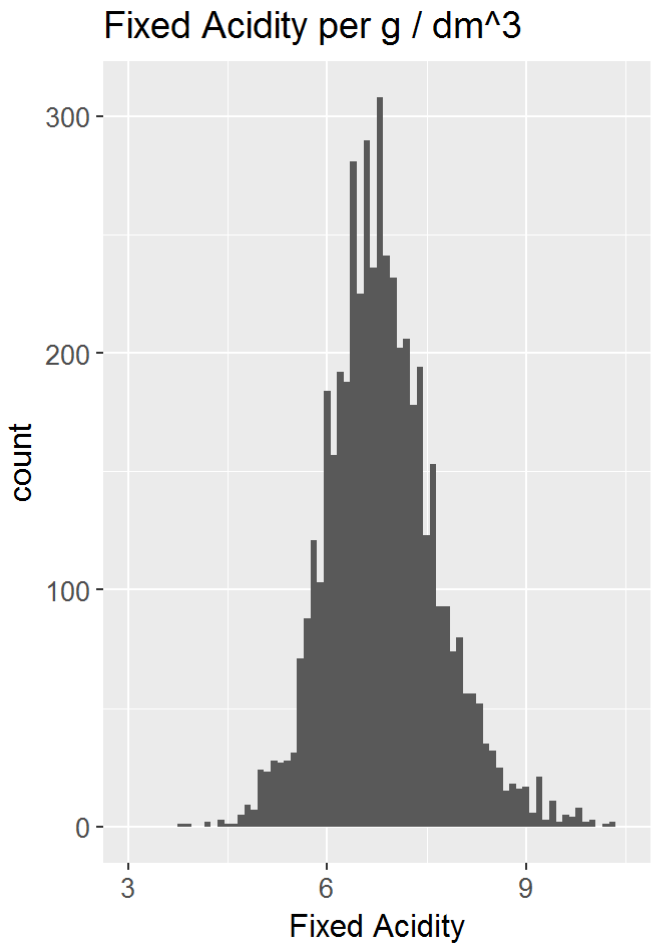
```
## [1] 4898   12
```

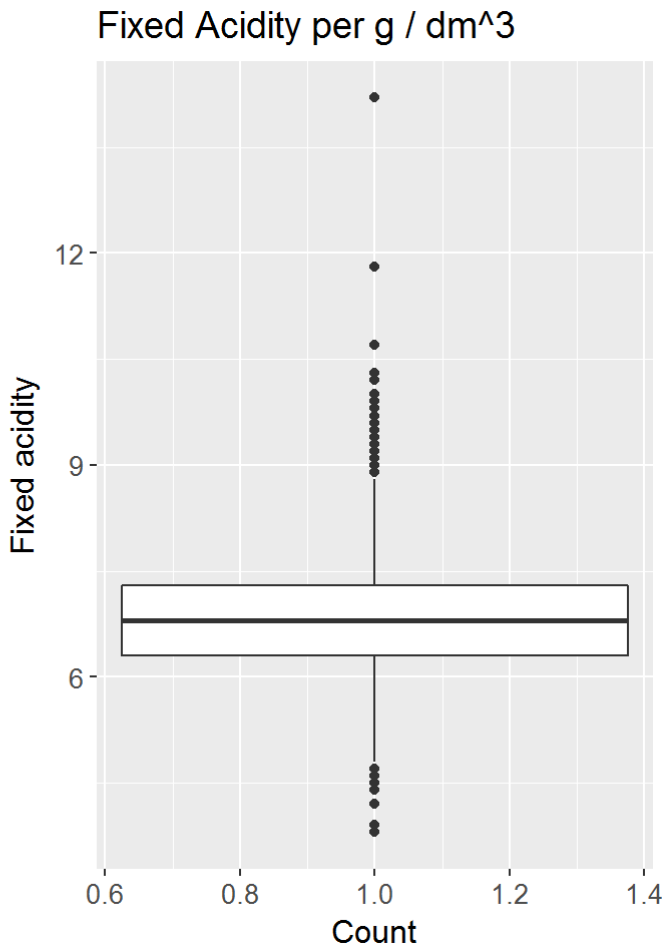
```
## 'data.frame':    4898 obs. of  12 variables:
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ 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 ...
```

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

Here in this section I would like to explore the significance of couple of variables in the data set.

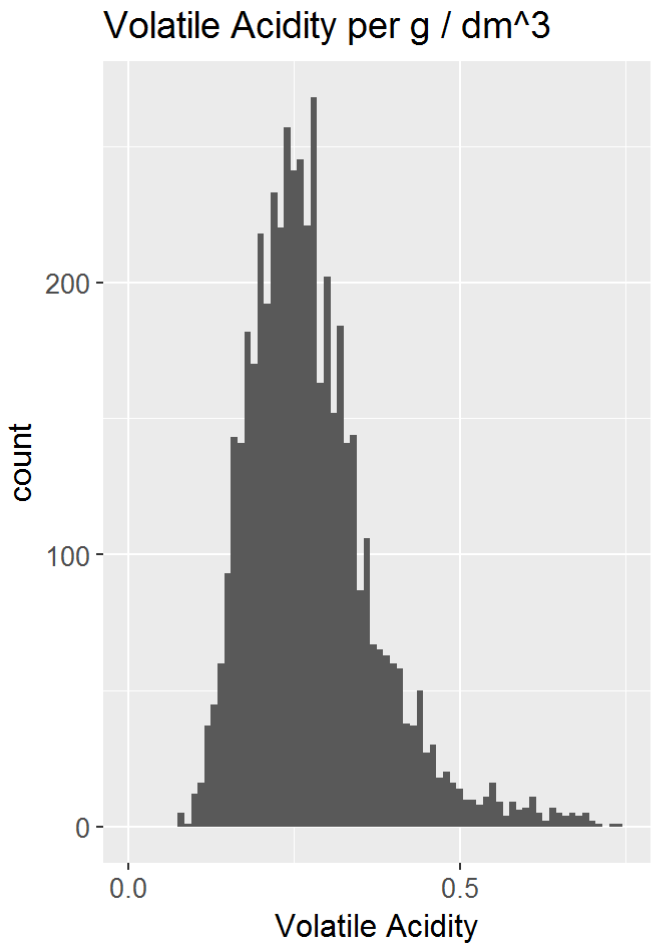
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.800   6.300   6.800   6.855   7.300  14.200
```

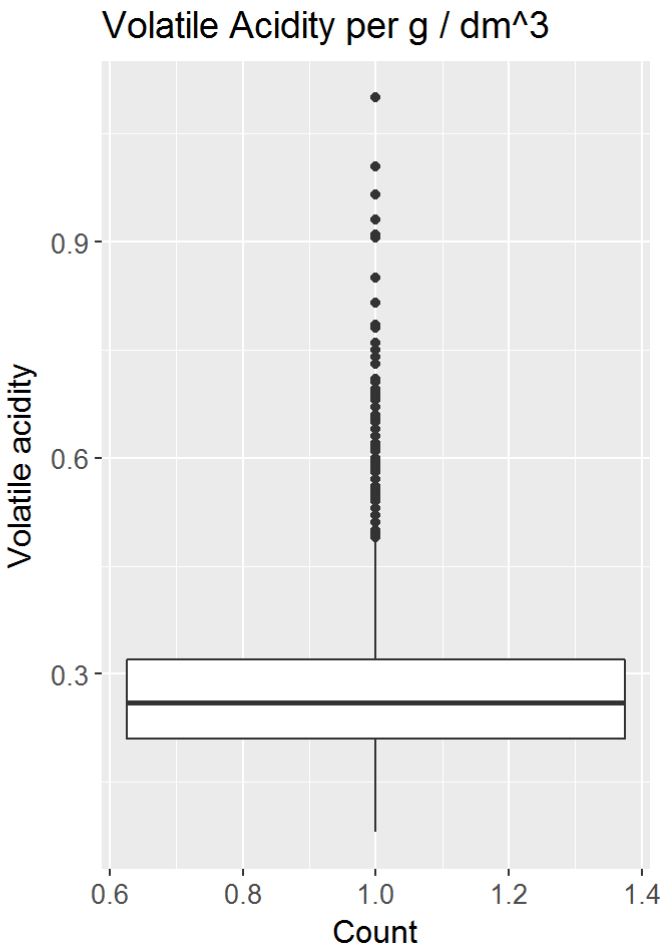




White wine's fixed acidity plot is a normal distribution with highest concentration of fixed acidity at 6.8. Around 50% of wines have fixed acidity more than 6.8 and around top 25% lie in the range 6.8-7.3 with an outlier at 14.2.

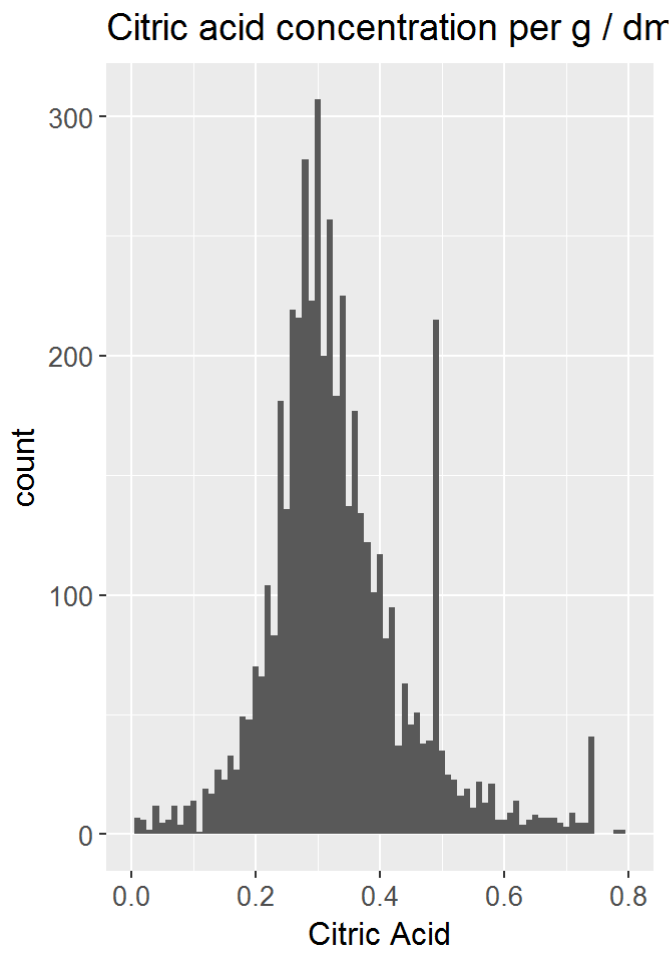
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.2100  0.2600  0.2782  0.3200  1.1000
```

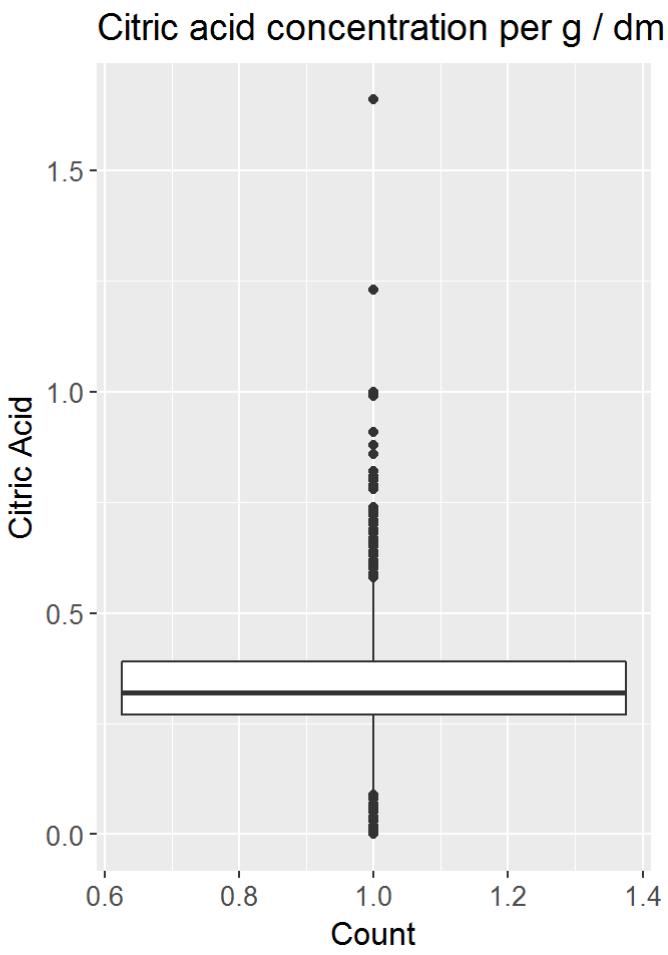




The summary statistics show that maximum volatile acidity is at 0.26. Around 50% of wines have volatile acidity more than 0.26 and around top 25% lie in the range 0.26-0.32 with an outlier at 1.1.

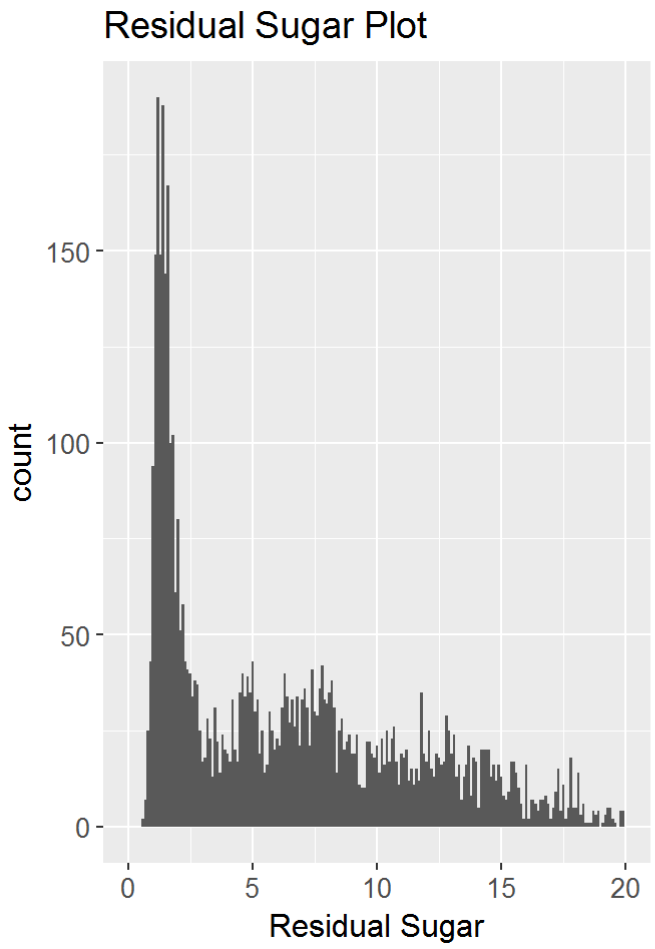
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2700  0.3200  0.3342  0.3900  1.6600
```

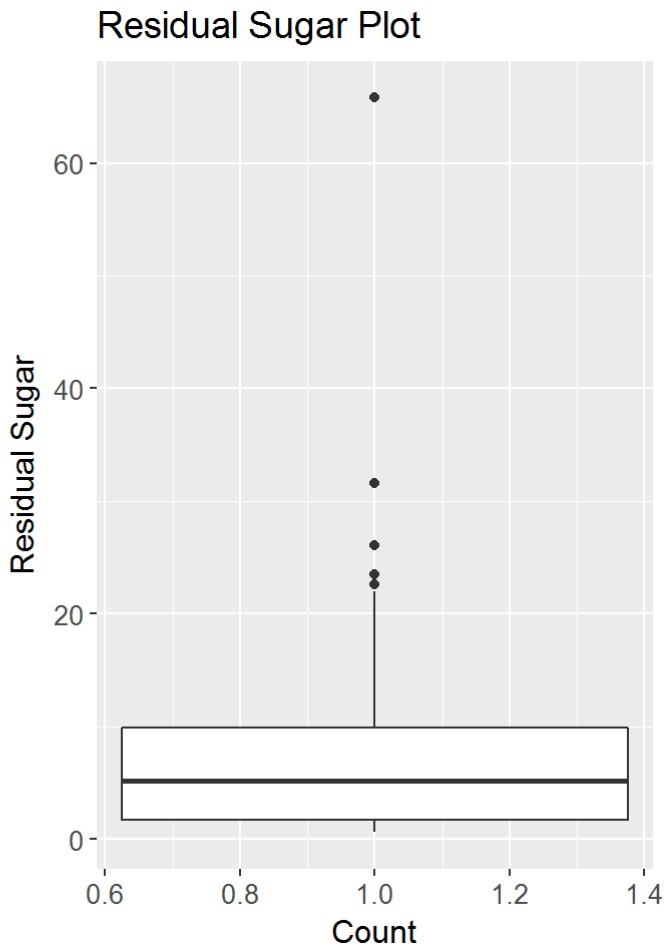




The plot has long tailed which after re - adjusting the axis could be normally distributed. Here the median of citric acid concentration is 0.32 g/dm^3 with an outlier at around 1.7.

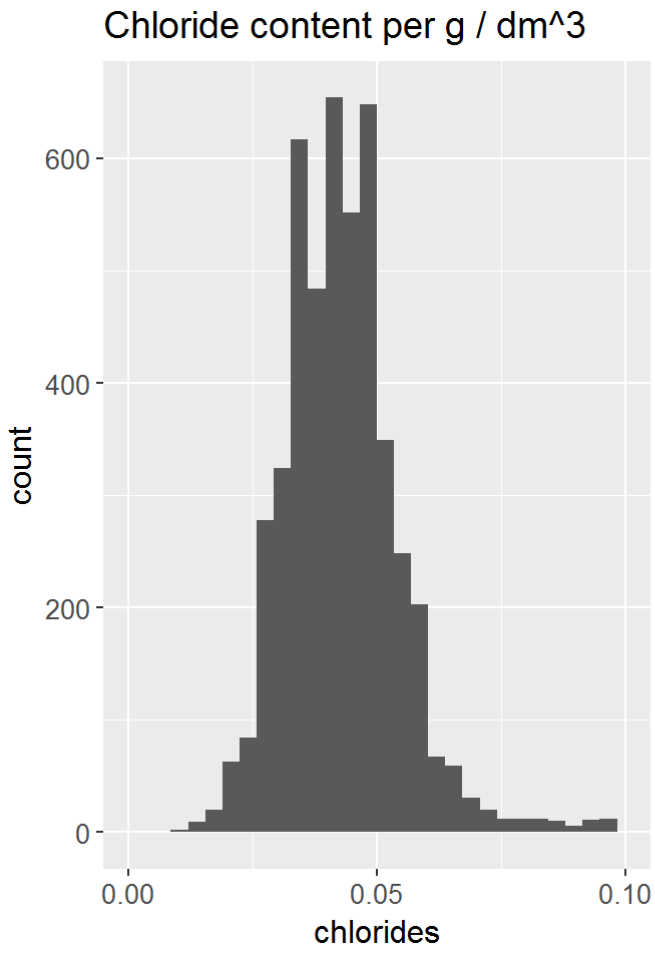
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.600   1.700   5.200   6.391   9.900  65.800
```

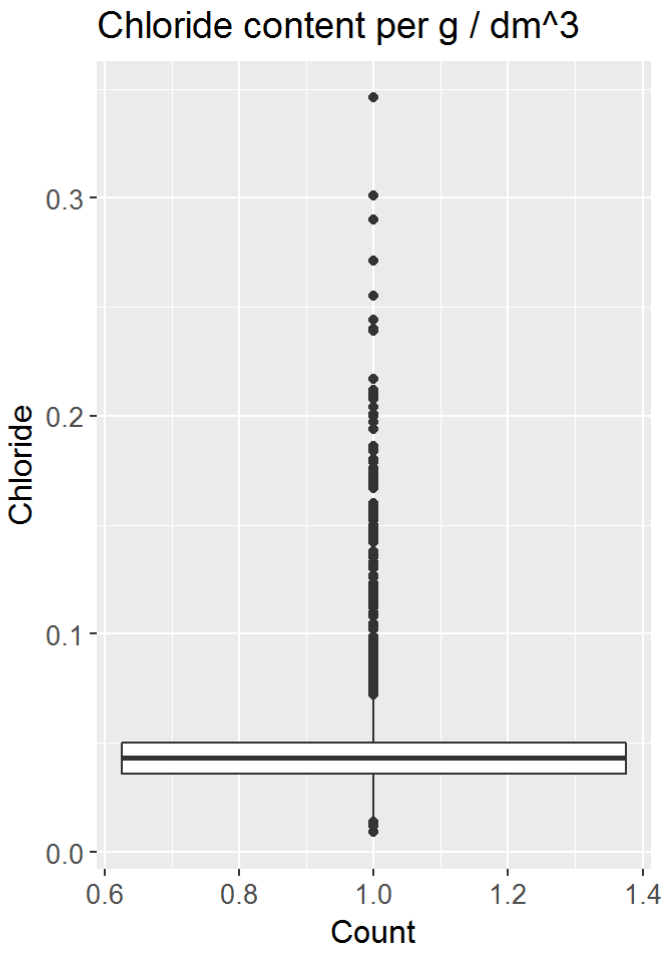




Here the median residual sugar is 5.2 g/dm^3 with an outlier at 65.8. We can perform readjust the plot to get a better understanding of it. This visualisation of residual sugar appears skewed to the right with a high peak at around 3.

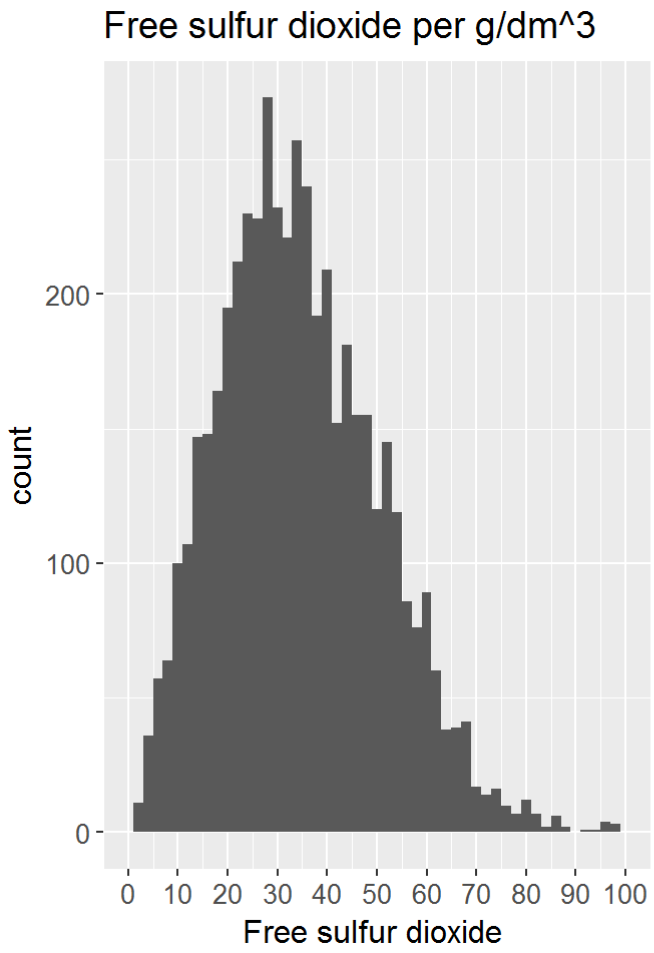
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00900 0.03600 0.04300 0.04577 0.05000 0.34600
```

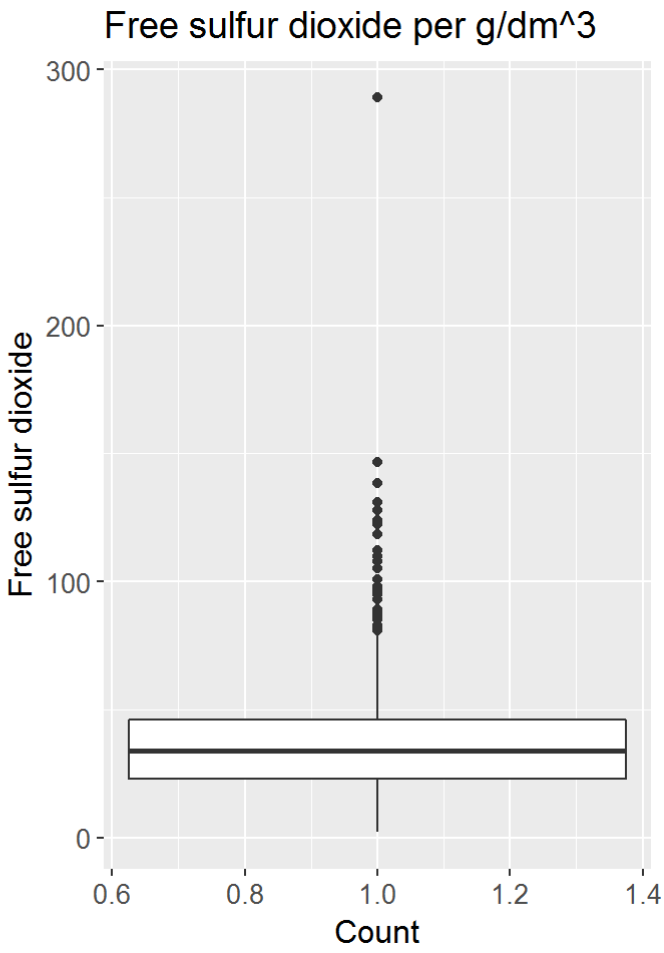




This visualization has a long tail with an outlier at around 0.34.This visualization appears slightly bimodal with rough peaks at around 0.03,0.05

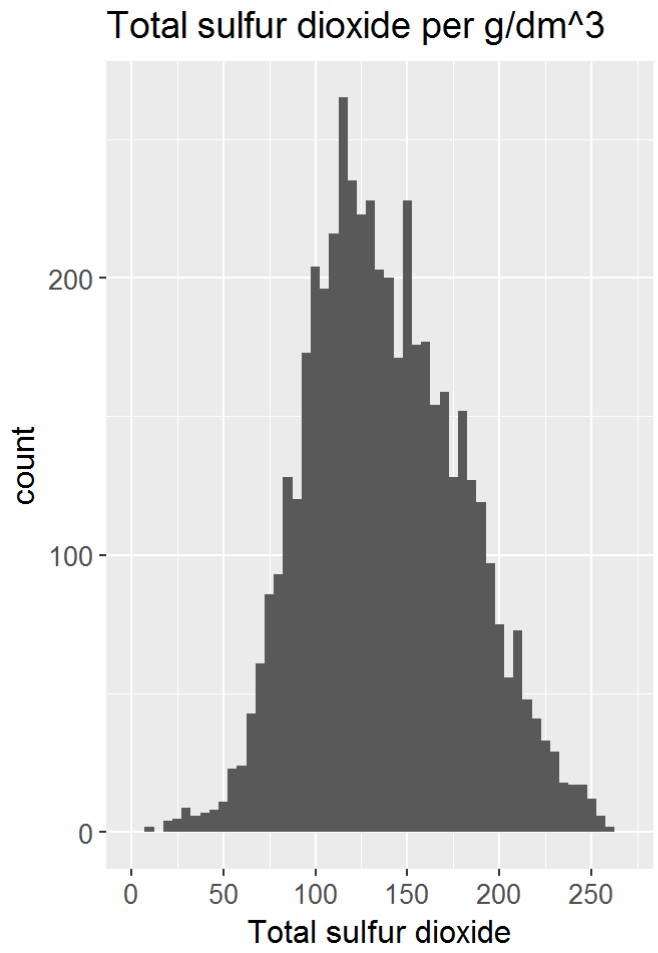
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   23.00   34.00   35.31   46.00  289.00
```

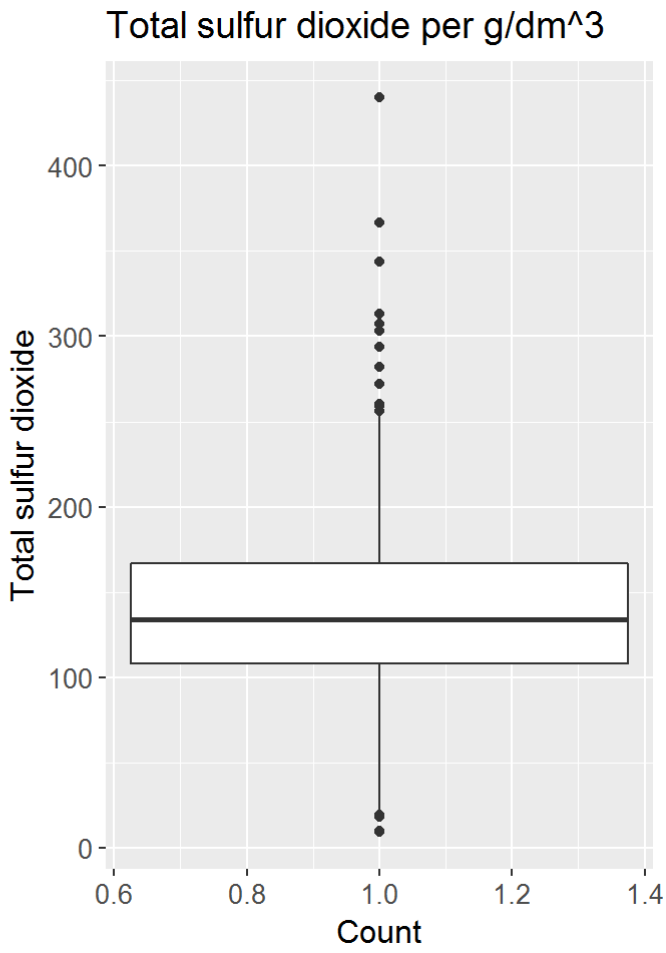




This graph has a median of free sulfur dioxide of 34 g/dm^3 and have an outlier at 289. On re-adjusting the axis to obtain an understanding of the shape of the plot.After readjusting the axis to remove the outlier and after limit the range of x-axis to 0-100 g/dm^3, the plot is clearly a normal distribution.

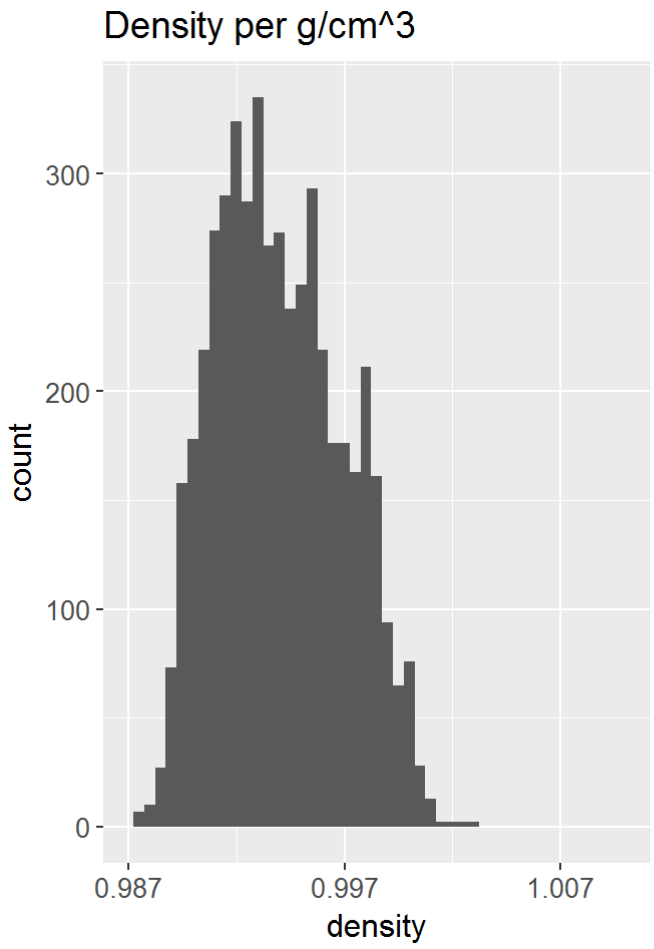
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.0   108.0   134.0   138.4   167.0   440.0
```

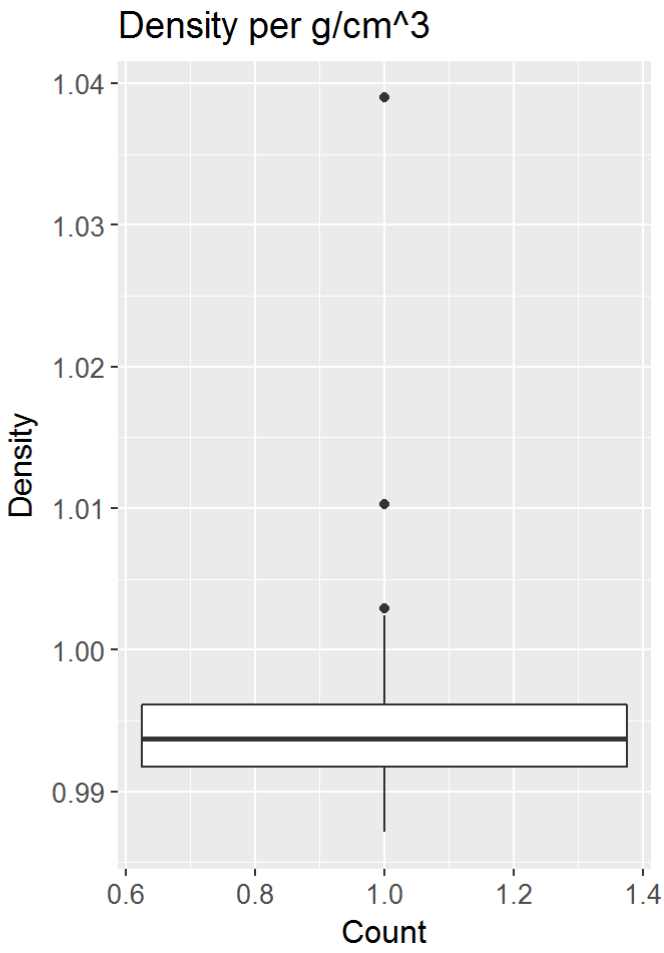




This plot is normally distributed with an outlier at 440 g/dm^3 with maximum concentration of sulfur dioxide at around 134 g/dm^3.

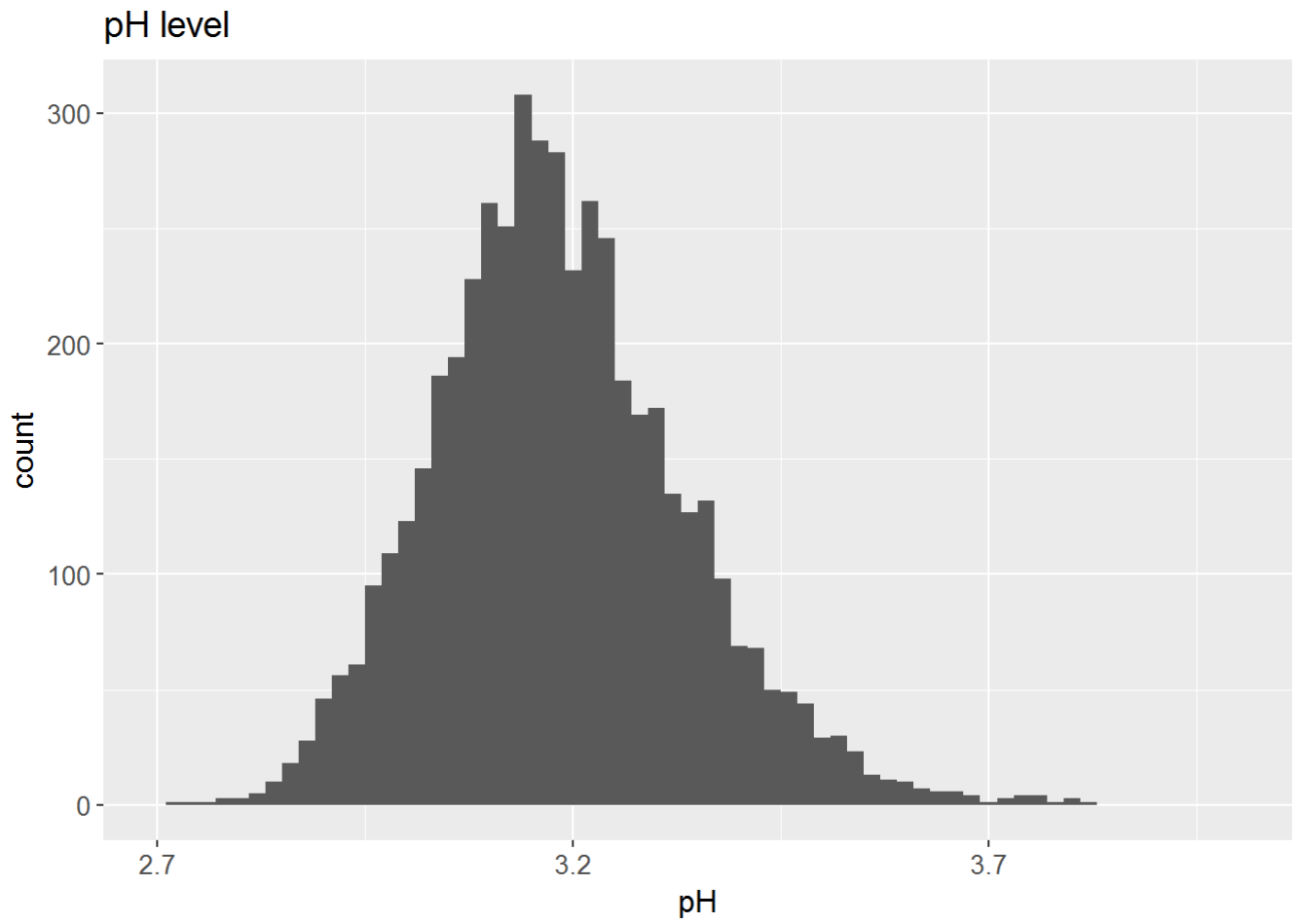
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9917  0.9937  0.9940  0.9961  1.0390
```





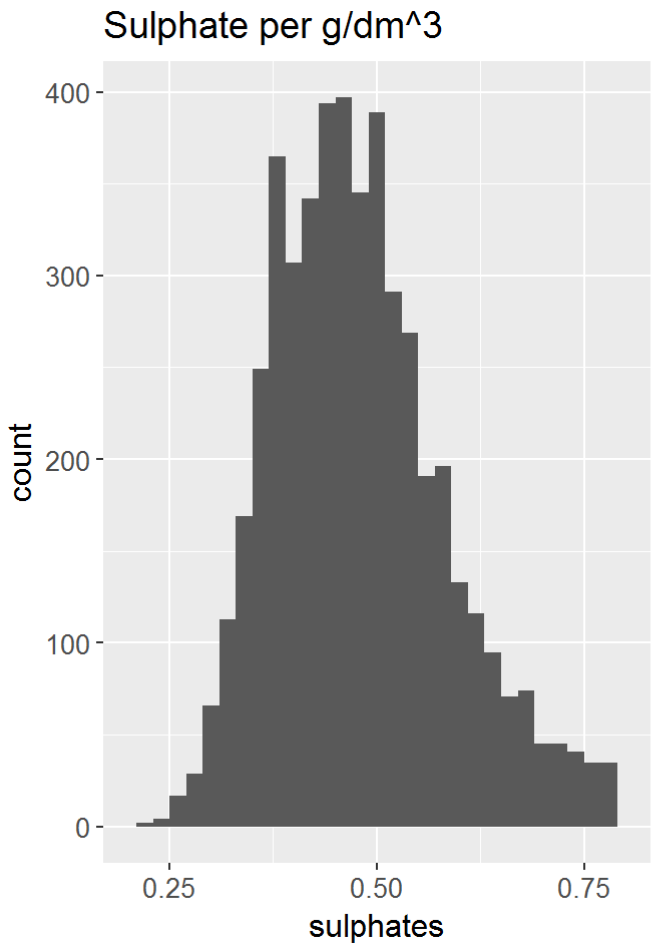
This plot had large number of outliers after 1.01 g/cm^3 with max being 1.0390 which had been removed after re-adjusting the x-axis. Now the plot appears to be normally distributed.The median of density had been at 0.9937 g/cm^3.

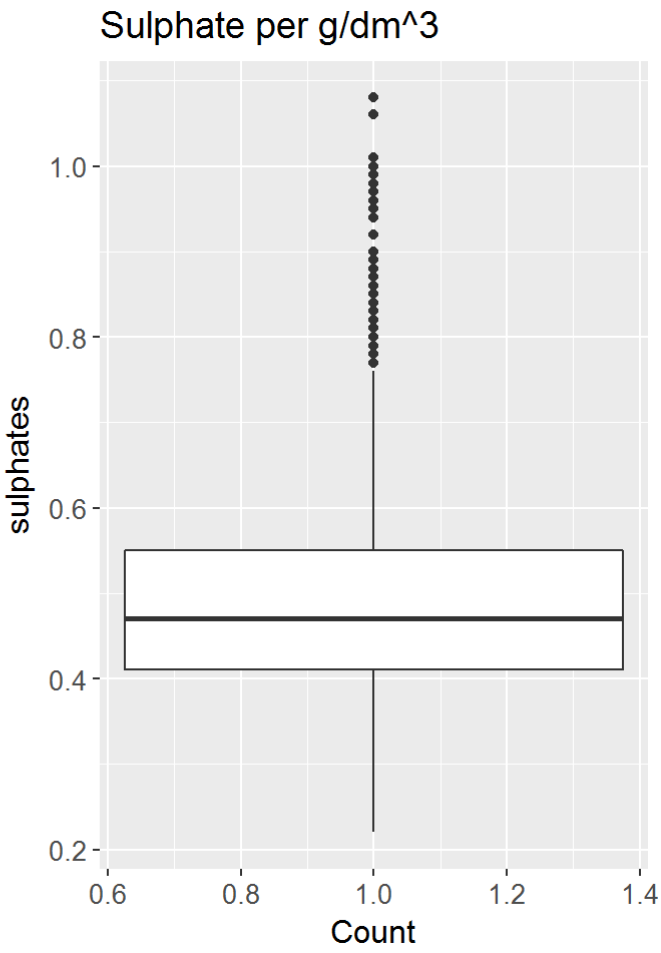
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.720   3.090   3.180   3.188   3.280   3.820
```



This plot of pH level in the white wines is normally distributed with the median at 3.188. Around 50% of white wines had pH level of more than 3.188 with the maximum being at 3.820.

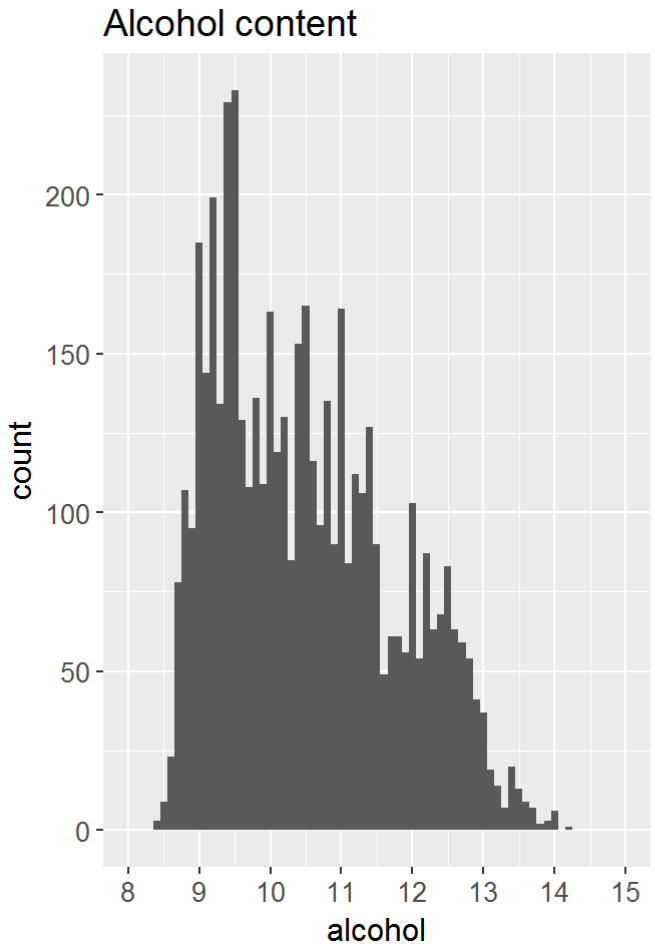
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2200  0.4100  0.4700  0.4898  0.5500  1.0800
```

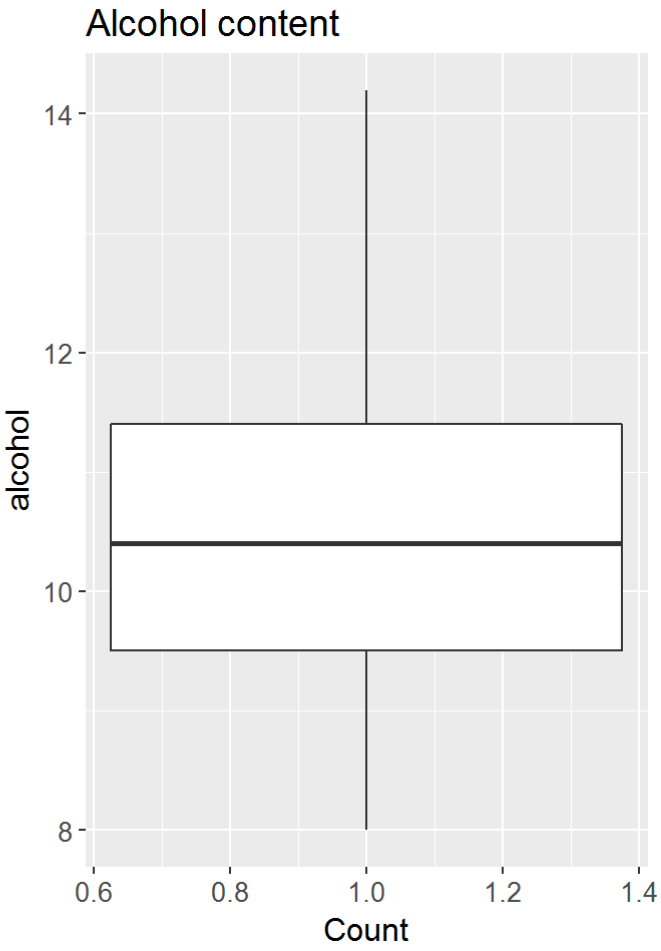




This plot is skewed to the right with a long tail and having an outlier at 1.08. The median of sulphate concentration is 0.47 g/dm^3.

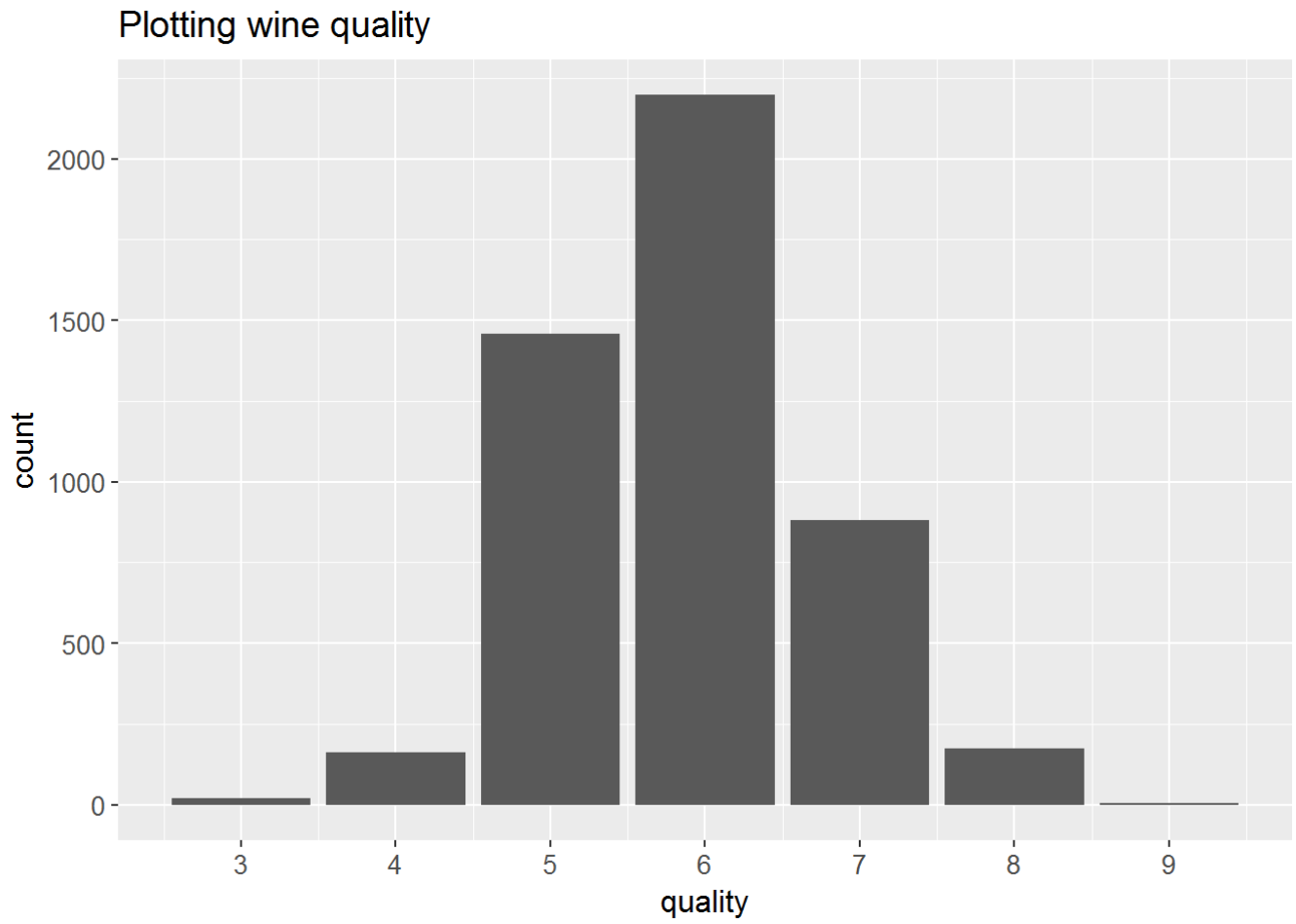
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.50   10.40   10.51   11.40   14.20
```





This plot appears to be skewed towards the right having mean of 10.51 and and an outlier at 14.20.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.878   6.000   9.000
```



This bar plot shows that the median of wine quality has been 6 forming around 44.8% of the dataset with the best quality of wine having quality as 9 which forms around 1.02% of the dataset.

Univariate Analysis

What is the structure of your dataset?

There are 4898 observations of white wine in this dataset with 13 variables. The first variable 'X' is a unique identifier for each record so it is ignored in my analysis. Other variables like fixed.acidity, volatile.acidity, citric.acid,residual.sugar,chlorides,free.sulfur.dioxide,total.sulfur.dioxide, density,pH,sulphates,alcohol are all numeric variables with quality being an integer variable.

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

My main point of interest is determing the factors which determine the quality of white wine.

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

I would need to compute the correlation of each of the variables like fixed.acidity, volatile.acidity,citric.acid,residual.sugar,chlorides, free.sulfur.dioxide,total.sulfur.dioxide, density,pH,sulphates,alcohol with the quality to determine which variables influence the quality and hence the

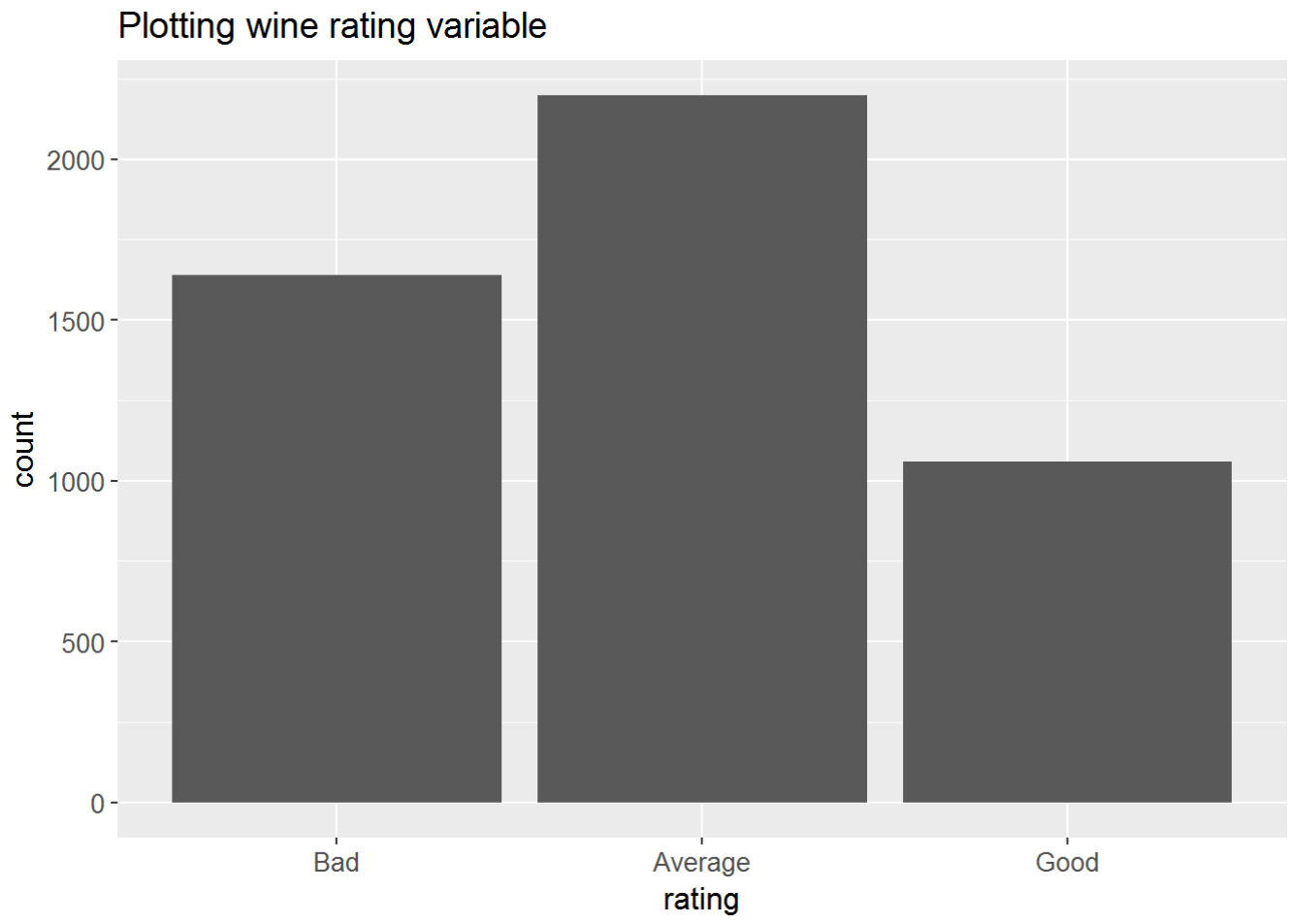
taste of white wine. I will explore these relationships in the later sections of my report.

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

I have added the following variable in the dataset :
* **rating** : which is calculated on the basis of the quality of the white wine. With 3-5 for “Bad”, 6 for “Average” and 7-9 as “Good” rating. I have chosen these scales to somewhat equally distribute the wine data set into various groups. This factored variable will help me in later investigating the relationship of variables with the wine quality
* **total.acidity** : Calculated as sum of fixed and volatile acidity
* **wine.type** : Classifies wine on the basis of sugar content, with wines less than 10 sugar as “Dry”, more than 30 as “sweet” and the others as “Medium” * **citric_acid_bucket** : Groups citric acid into buckets of [0-0.3],(0.3-0.6], (0.6,1.7]
* **total_sulfur_dioxide_bucket** : Groups total sulfur dioxide into buckets of [0-120],(120-210],(210,440]
* **alchol_content** : Classifies wine on the basis of alcohol content, with wines content less than 12.5 as “Low”, between 12.5 to 13.5 as “Moderately Low” and the others as “High”
* **chloride_sulfate_ratio** : Classifies the proportion of chloride with sulfates with proportion less than 0.5 as “Very Bitter” and others as “Bitter”

Since the wine rating variable will be used in my further analysis, I want to plot a visualization depicting the distribution of wines in the different “rated” buckets

```
##     Bad Average    Good
##    1640    2198    1060
```



From this visualization it is clear that maximum wines have “average” rating(quality=6), followed by “bad”(quality between 3 to 5) and then “good”(quality between 7 to 9)

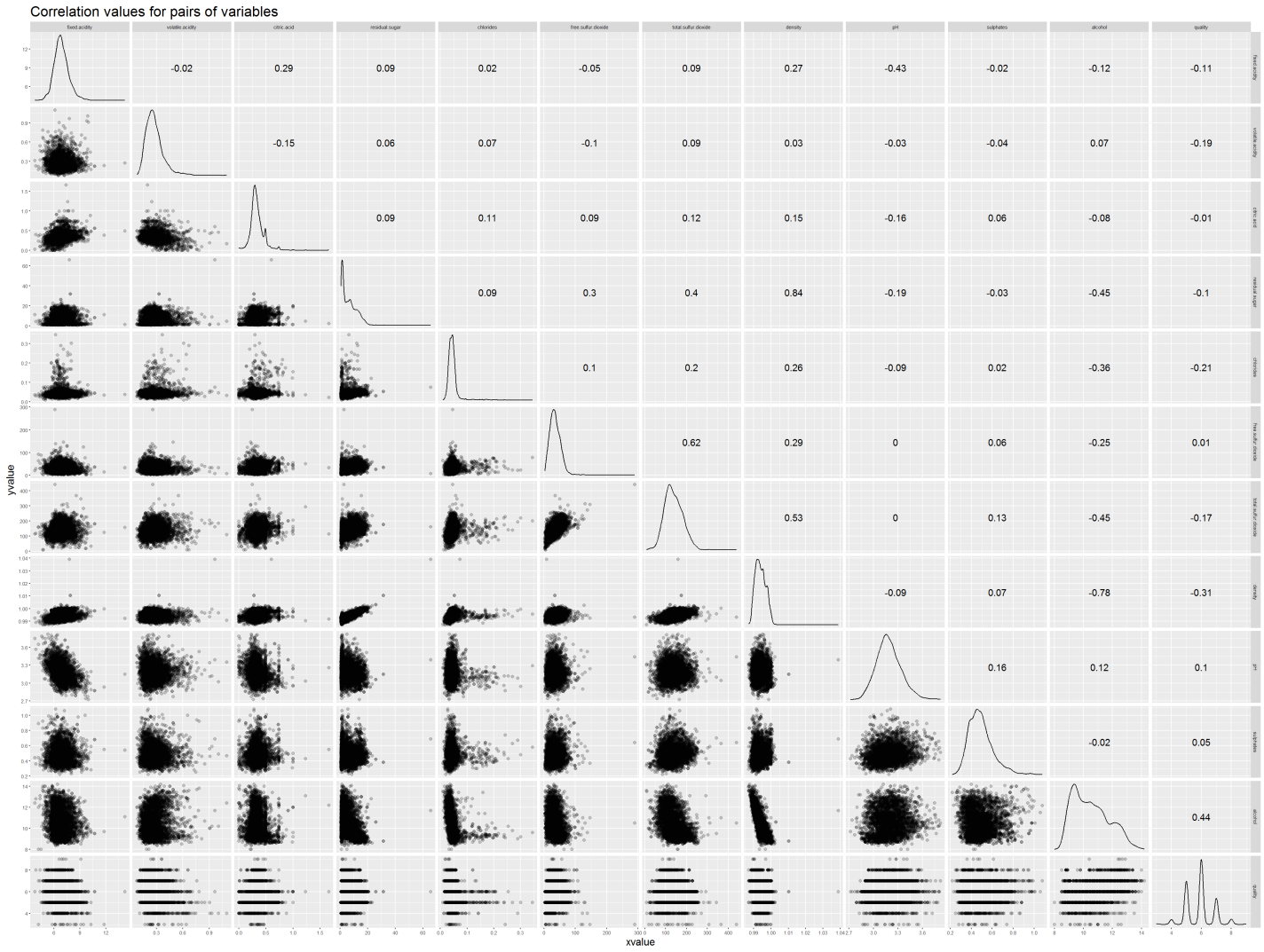
Of the features you investigated, were there any unusual distributions?
Did you perform any operations on the data to tidy, adjust, or change the form
of the data? If so, why did you do this?

The fixed acidity,chloride and alcohol variables had to be log transformed to make the plot more normally distributed. Other than that citric acid, sulfur dioxide and residual sugar plots x- axis had to be re adjusted to have a better understanding of the visualization.

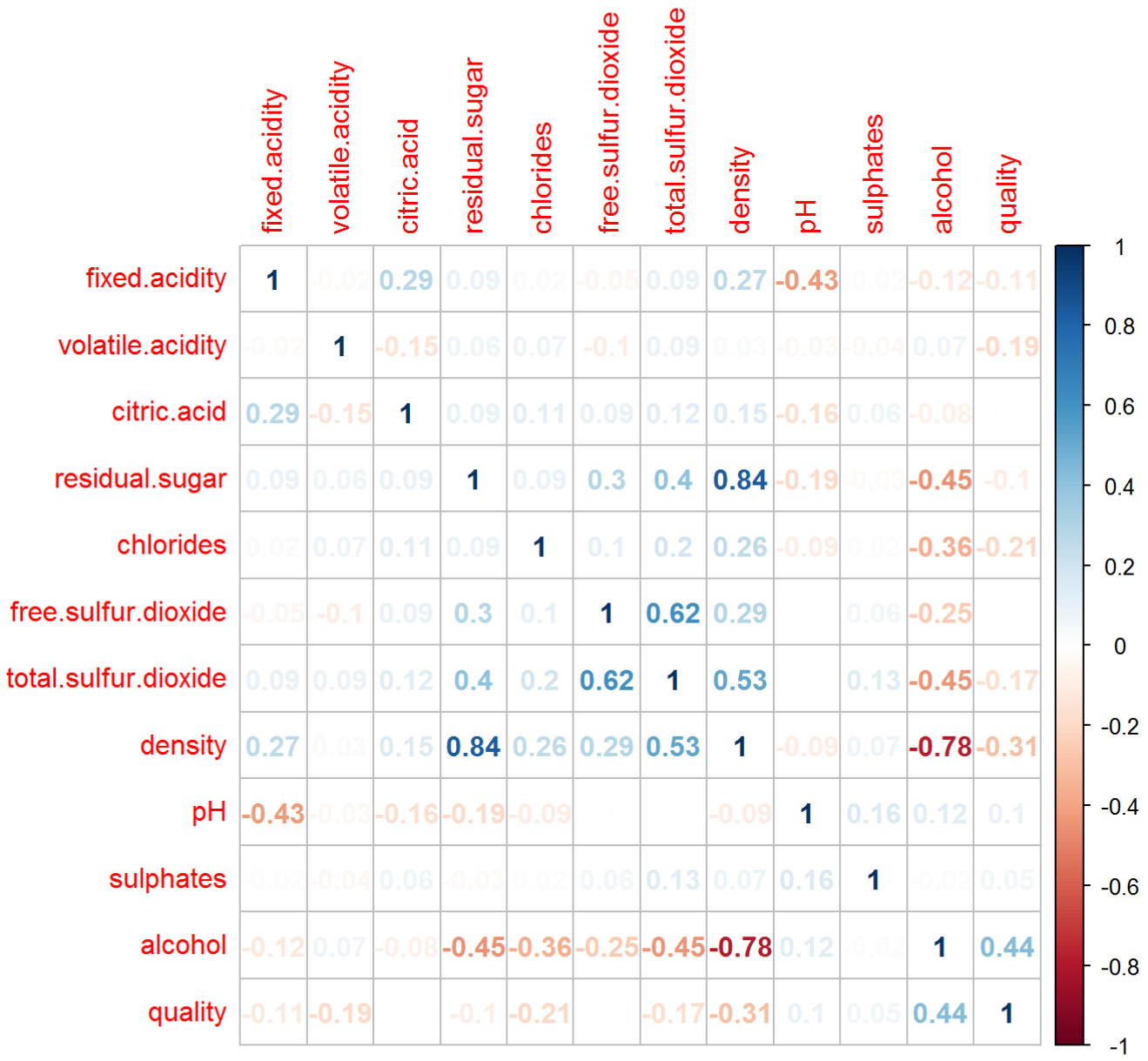
Bivariate Plots Section

Here in this section I would like to explore the relationship among couple of variable pairs which have a strong/weak relationship among them.

Analysis of relationship between the variables



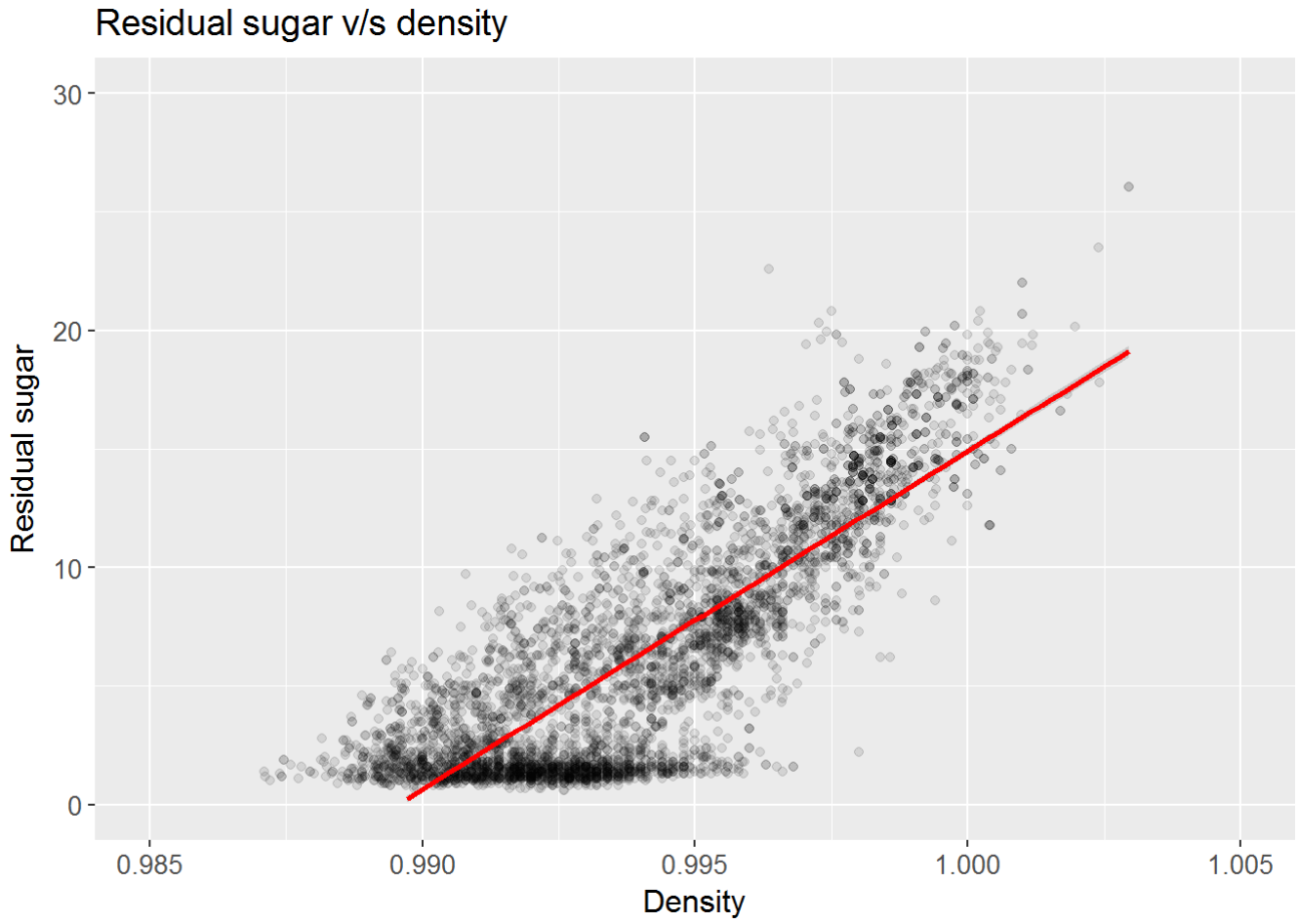
Numeric correlations among the variables


fixed.acidity
volatile.acidity
citric.acid
residual.sugar
chlorides
free.sulfur.dioxide
total.sulfur.dioxide
density
pH
sulphates
alcohol
quality
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

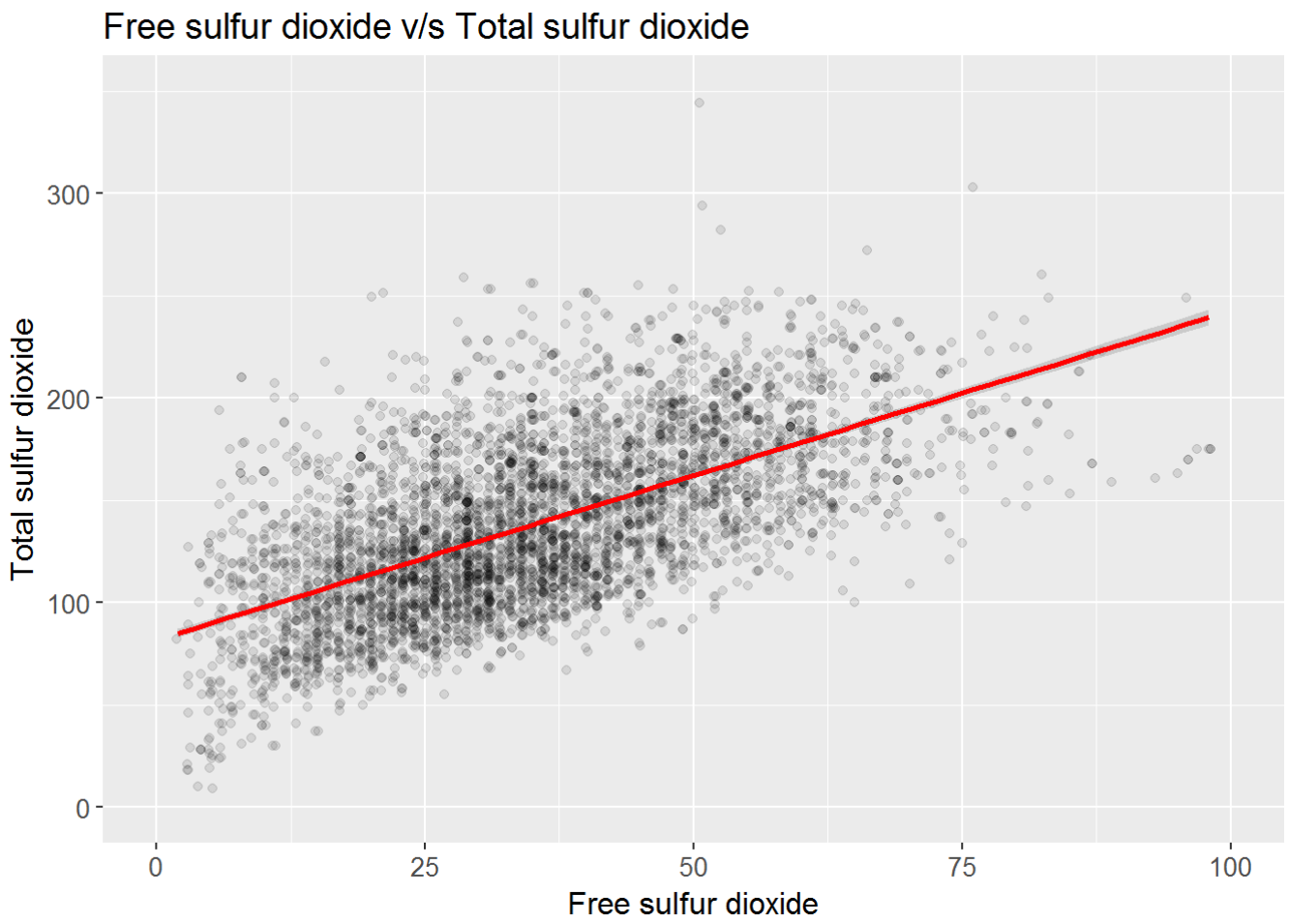
From the above plot the top most positive and negative correlation coefficients, r based on strength of association[1] are

- Residual sugar: density = 0.8389665
- Free sulfur dioxide: Total sulfur dioxide=0.615501
- Alcohol: quality = 0.4355747
- Total sulfur dioxide: density = 0.5298813
- Fixed acidity:Citric acid=0.2891807
- Chlorides:Density=0.2572113
- Density:quality= -0.3071233
- Alcohol:Residual sugar=-0.4506312
- Alcohol:Density= -0.7801376
- Fixed.acidity:pH= -0.4258583
- Alchohol: Total sulfur dioxide = -0.4488921
- Alcohol: Chloride = -0.3601887

I will exploring some of these relationships in this section of the report.



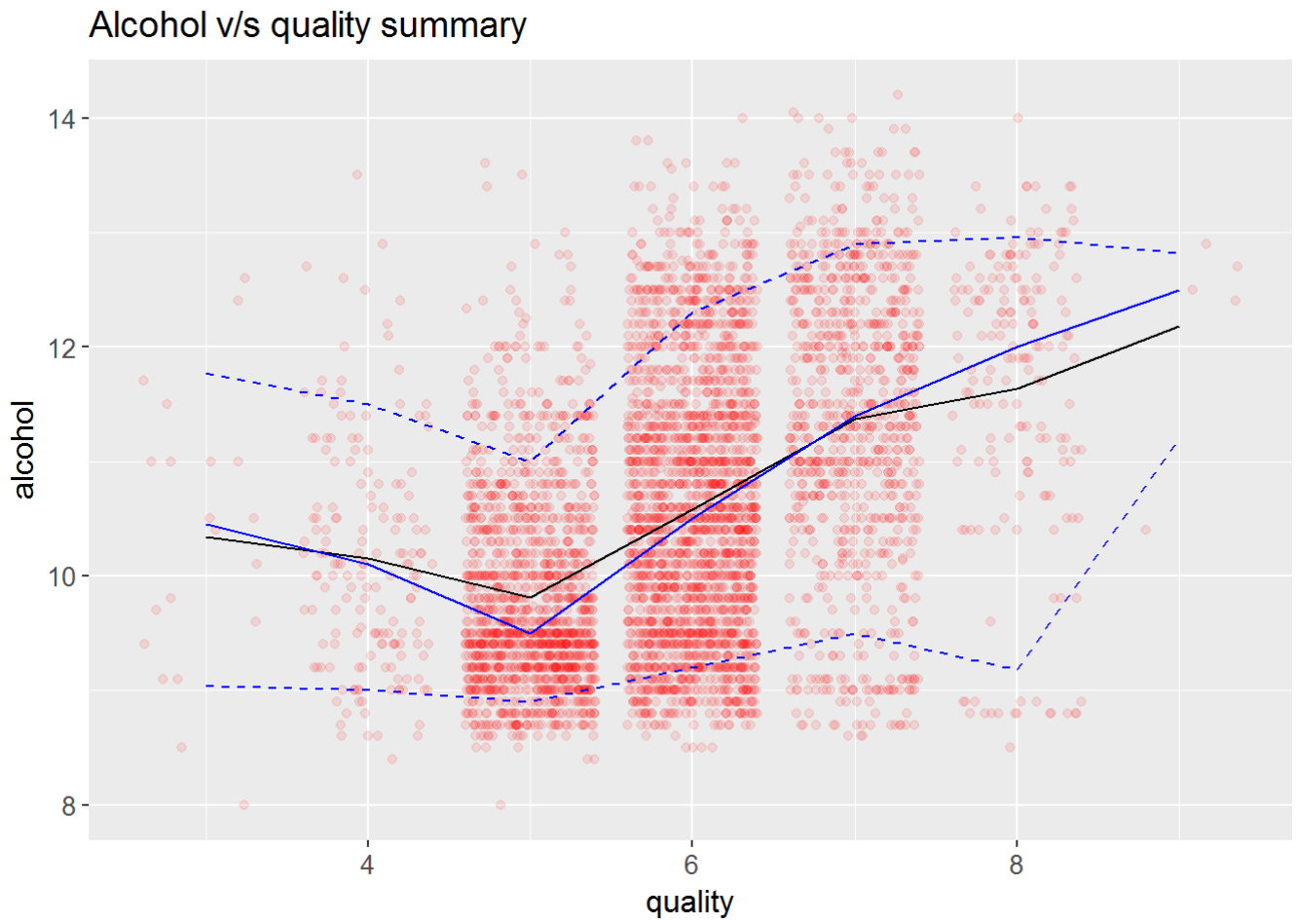
This plot shows that as residual sugar in white wine increases, the density increases.



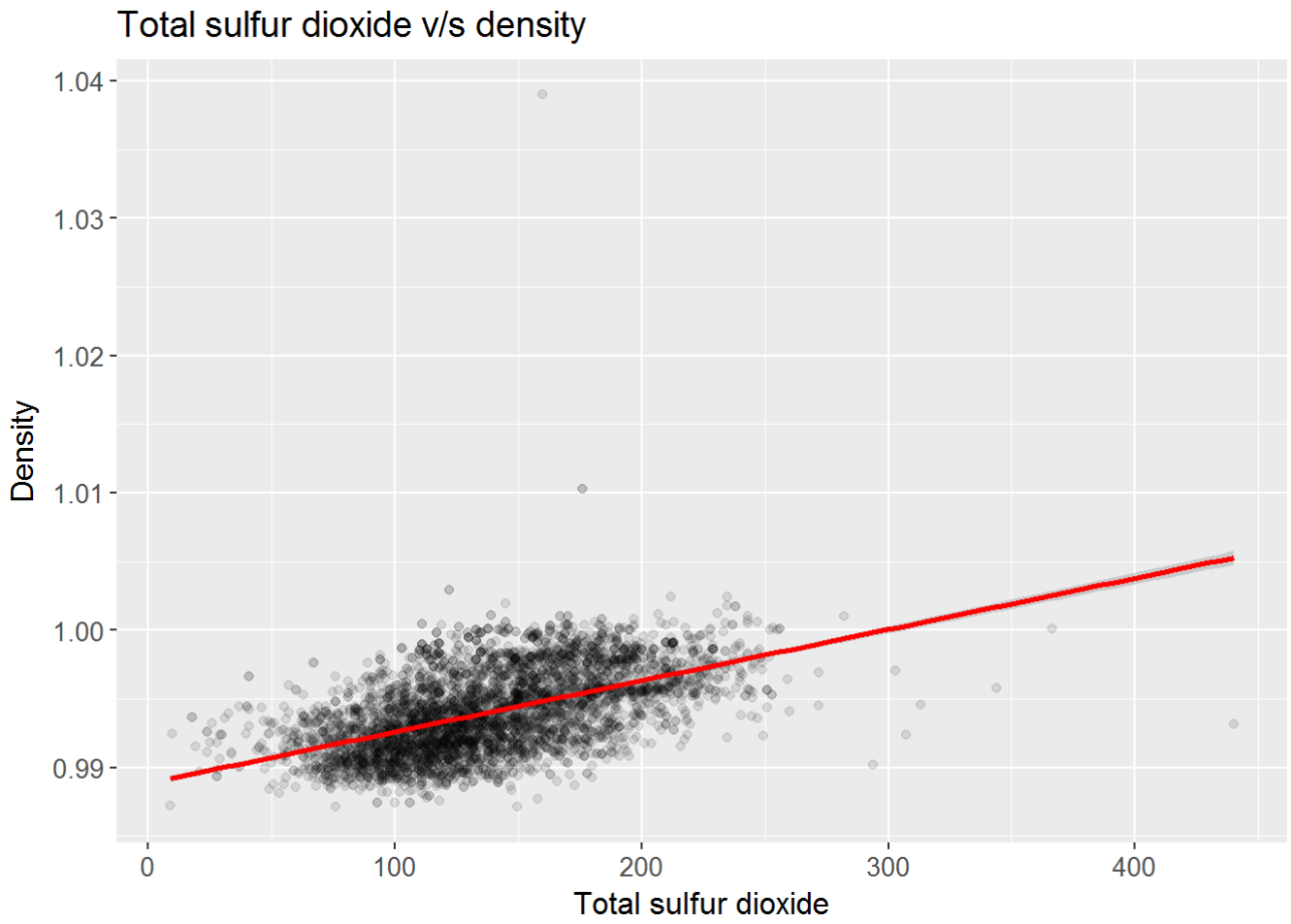
This plot shows that with increase in free sulfur dioxide the proportion of total sulfur dioxide increases.

```
## white_wine_df$rating: Bad
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.20    9.60    9.85   10.40   13.60
## ------------------------------------------------------------
## white_wine_df$rating: Average
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50    9.60   10.50   10.58   11.40   14.00
## ------------------------------------------------------------
## white_wine_df$rating: Good
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50   10.70   11.50   11.42   12.40   14.20
```

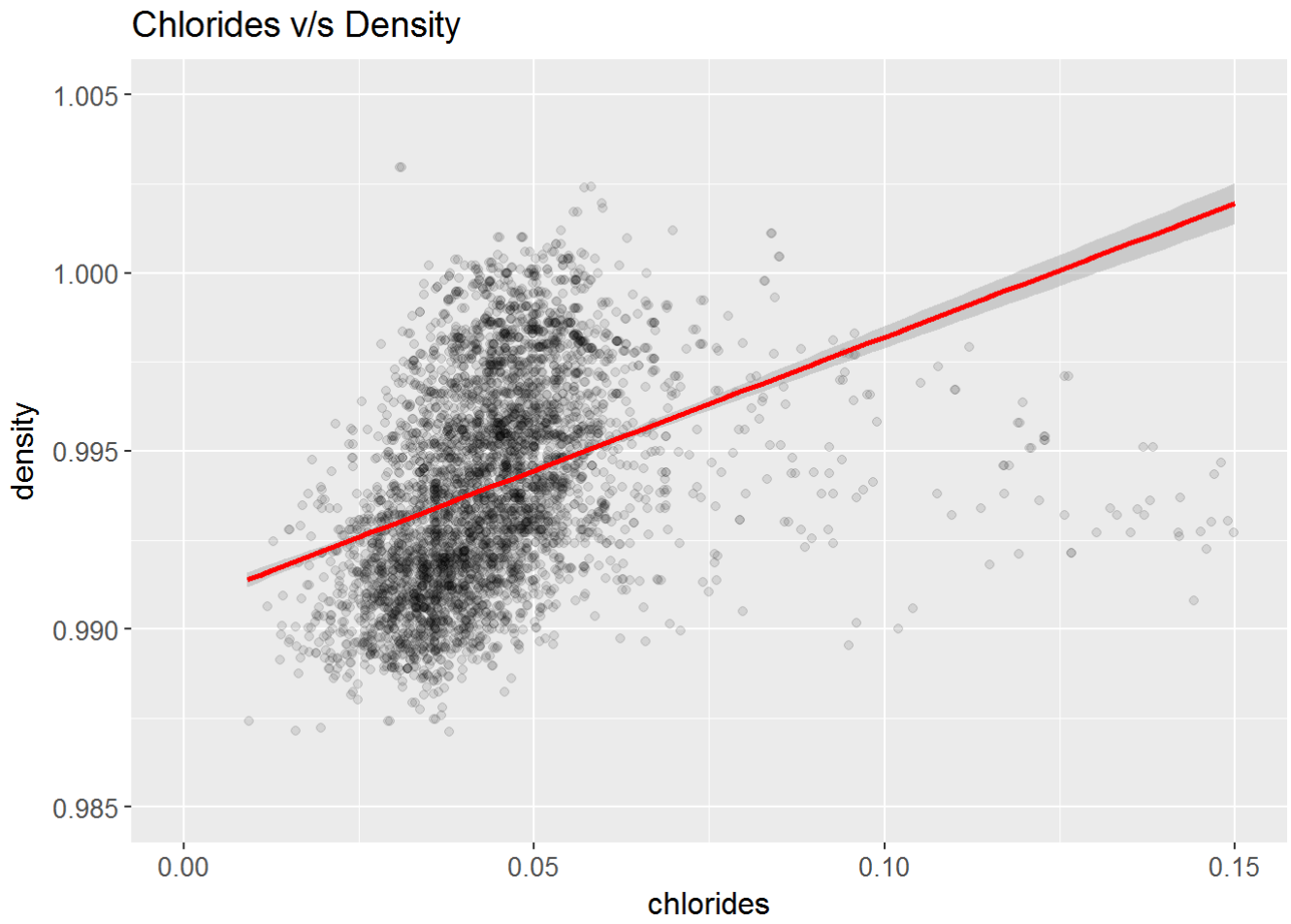
This plot shows that good quality of wine had high median of 12.0 alcohol. This visualisation shows that as alcohol content increase the quality of wine increases.



This plot the summary statistic between alcohol and quality with black line representing the mean, and the 0.9,0.5,0.1 quantile relationship between the two.

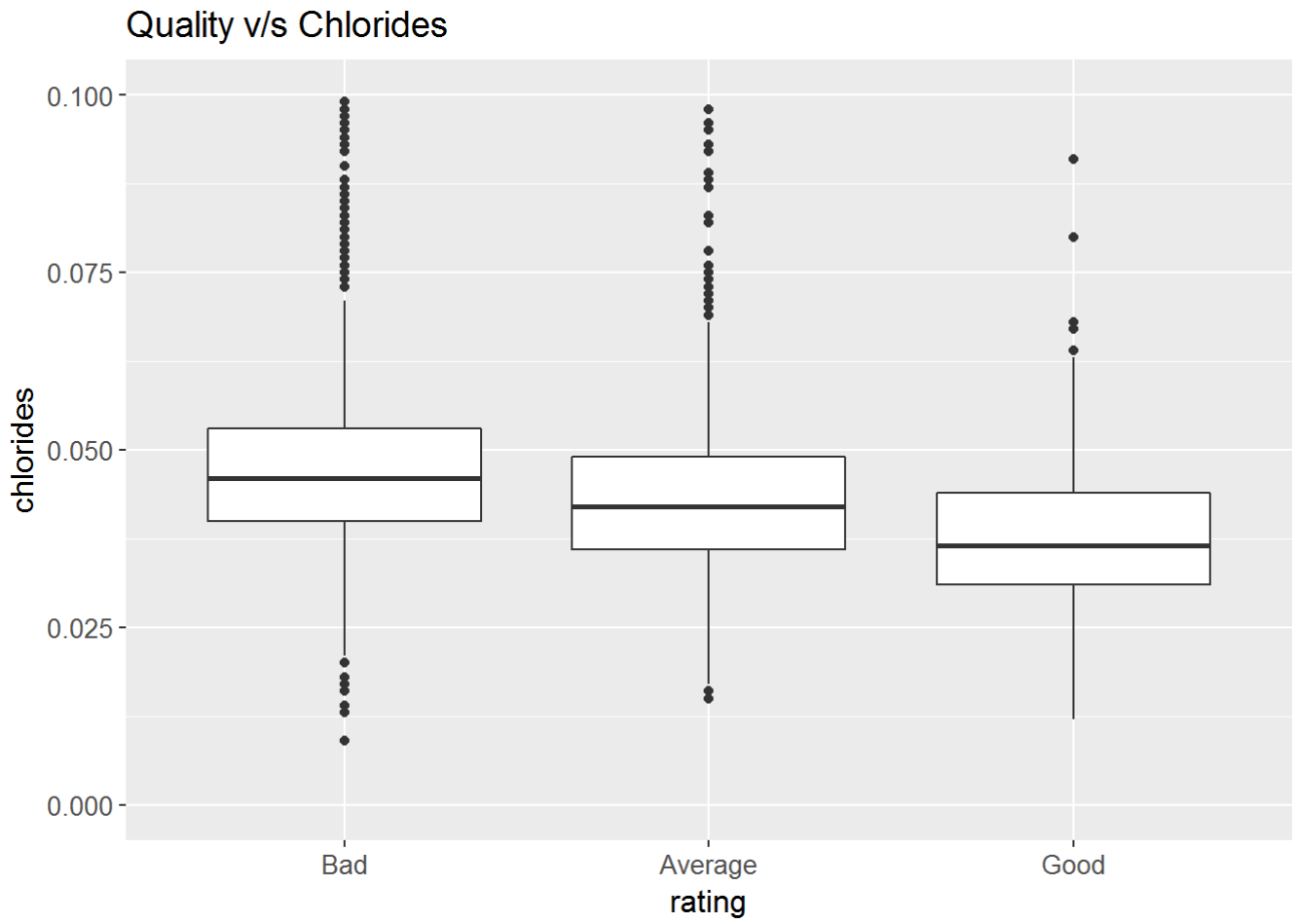


This plot shows that with increase in total sulfur dioxide the density of white wine increases.



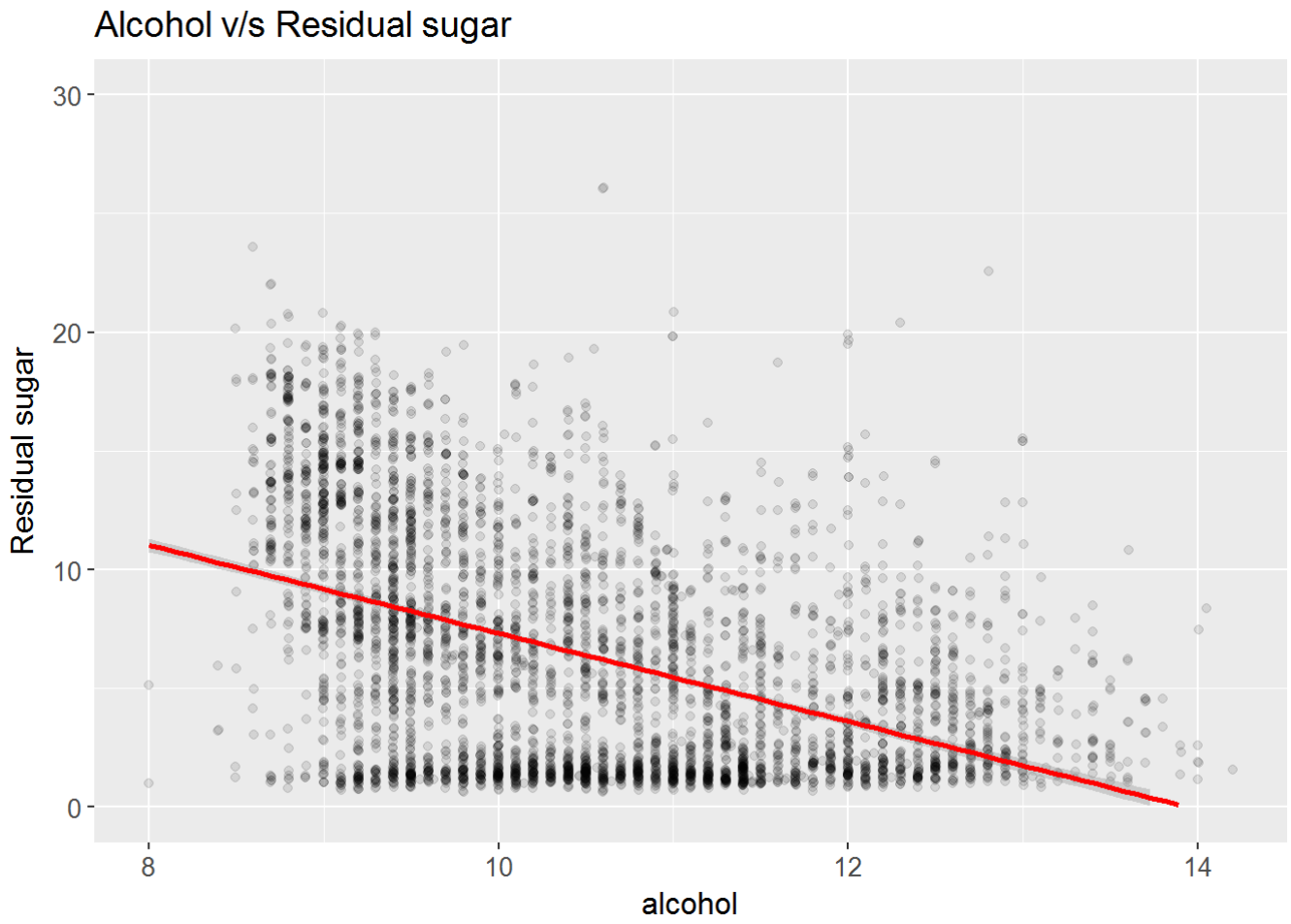
This plot shows that with increase in chlorides the density of white alcohol as well increases.

```
## white_wine_df$rating: Bad
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00900 0.04000 0.04700 0.05144 0.05300 0.34600
## ------------------------------------------------------------
## white_wine_df$rating: Average
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01500 0.03600 0.04300 0.04522 0.04900 0.25500
## ------------------------------------------------------------
## white_wine_df$rating: Good
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.03100 0.03700 0.03816 0.04400 0.13500
```

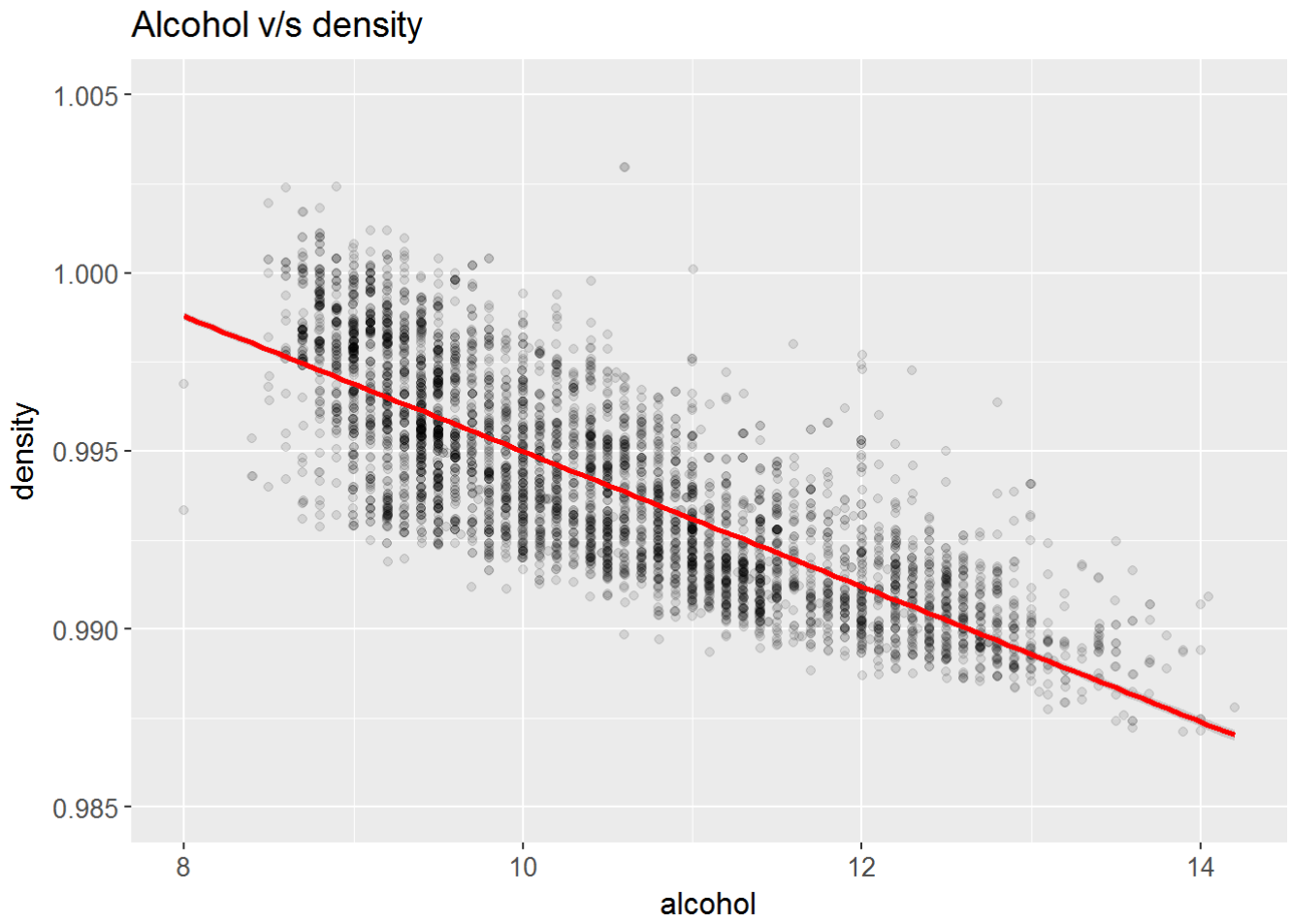


This mean that with good quality of wine the proportion of chlorides is less and for average quality of wine the proportion of chlorides in more. The chloride proportion increases with better quality of wine and then decreases.

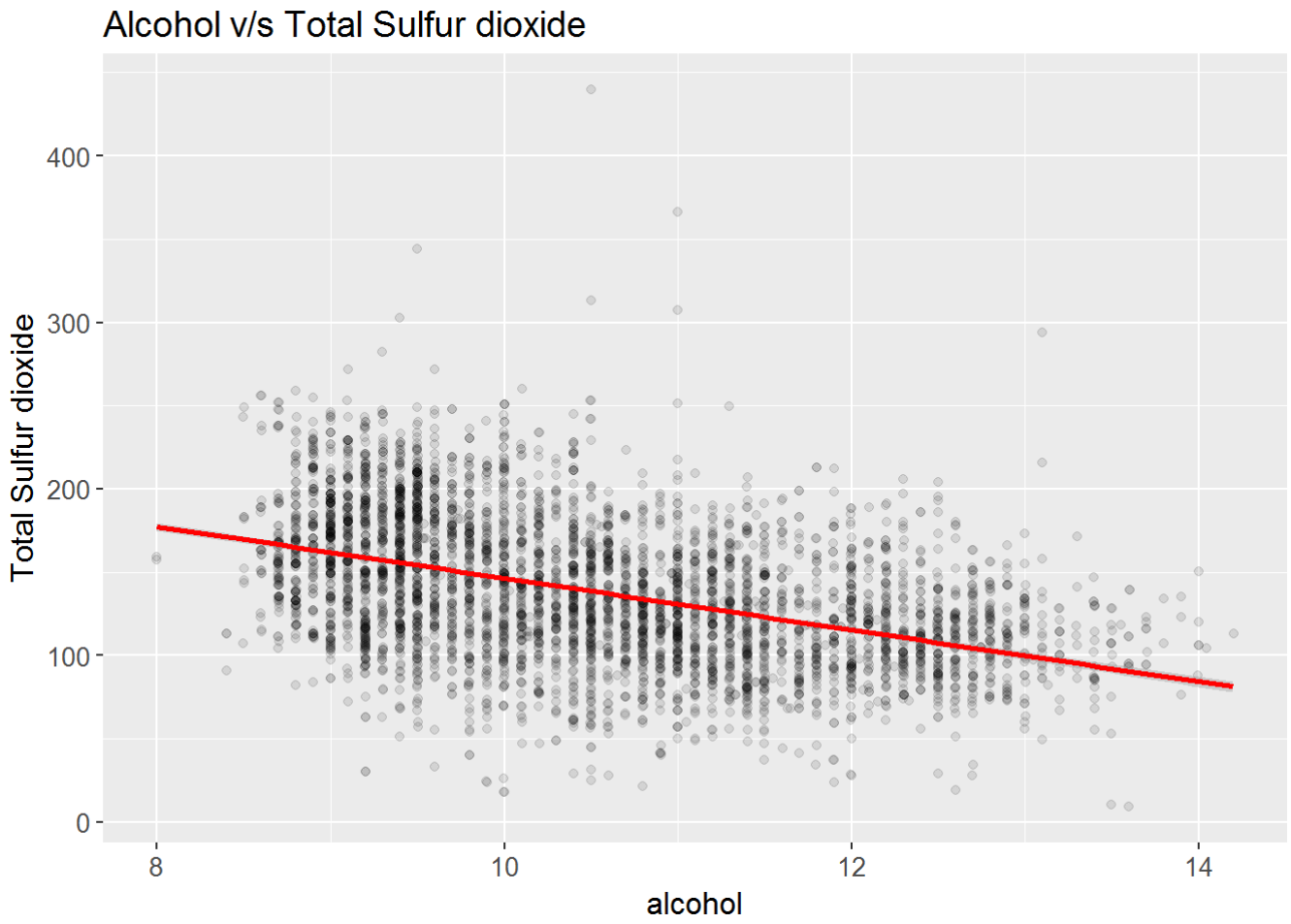
```
##               Df Sum Sq Mean Sq F value Pr(>F)
## rating         2 0.1147 0.05735   126.3 <2e-16 ***
## Residuals   4895 2.2228 0.00045
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



This plot shows that large number of wines have residual sugar content decrease as alcohol proportion increases.



This plot shows that large number of wines have density reduced as the alcohol proportion increases



These plot show that as alcohol in white wine increase the total sulfur dioxide content decreases.

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?

Here in this investigation I observed strong relation between alcohol and quality,residual sugar and density,free and total sulfur dioxide,fixed acidity and citric acid, chlorides and density.

The correlation constant values of other variables in the dataset wrt to quality are :

- citric.acid:quality = -0.009209091
- free.sulfur.dioxide:quality = 0.008158067
- fixed.acidity: quality= -0.1136628
- volatile.acidity:quality = -0.194723
- residual.sugar:quality = -0.09757683
- chlorides:quality = -0.2099344
- total.sulfur.dioxide:quality = -0.1747372
- density:quality = -0.3071233

- pH:quality = 0.09942725
- sulphates:quality=0.05367788
- alcohol: quality=0.4355747

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

I observed positive correlation between free and total sulfur dioxide(0.615501), alcohol and quality(0.4355747),Fixed acidity and Citric acid(0.2891807), Chlorides and Density(0.2572113) and negative correlation between density and quality(-0.3071233), Alcohol and Residual sugar(-0.4506312), Alcohol and Density(-0.7801376) and Fixed acidity and pH(-0.4258583)

What was the strongest relationship you found?

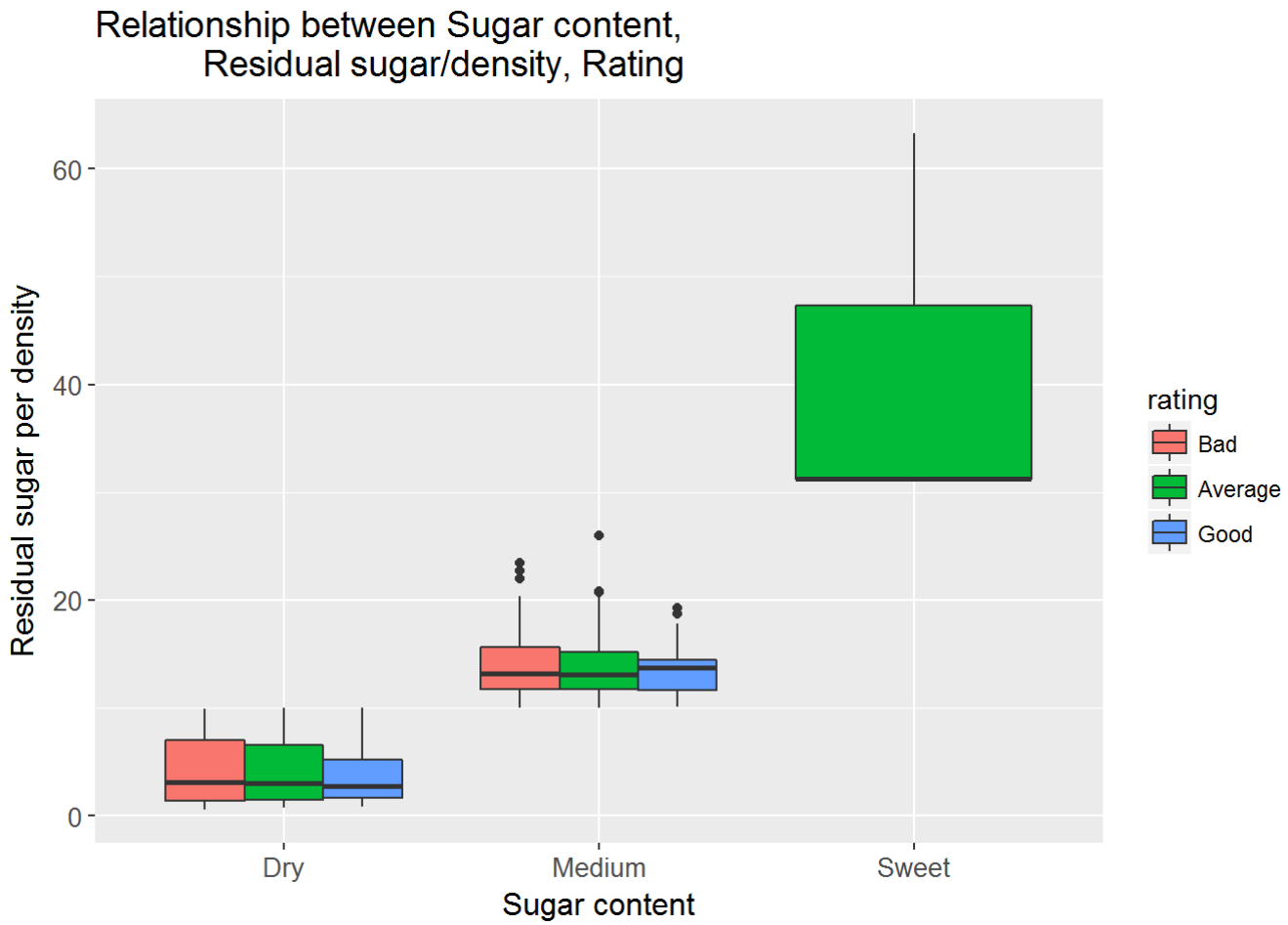
I found the strongest positive correlation between residual sugar and density with correlation constant value of 0.8389665.

Multivariate Plots Section

Here in this section I will like to explore the relationship of 2 or more variables with respect to wine rating

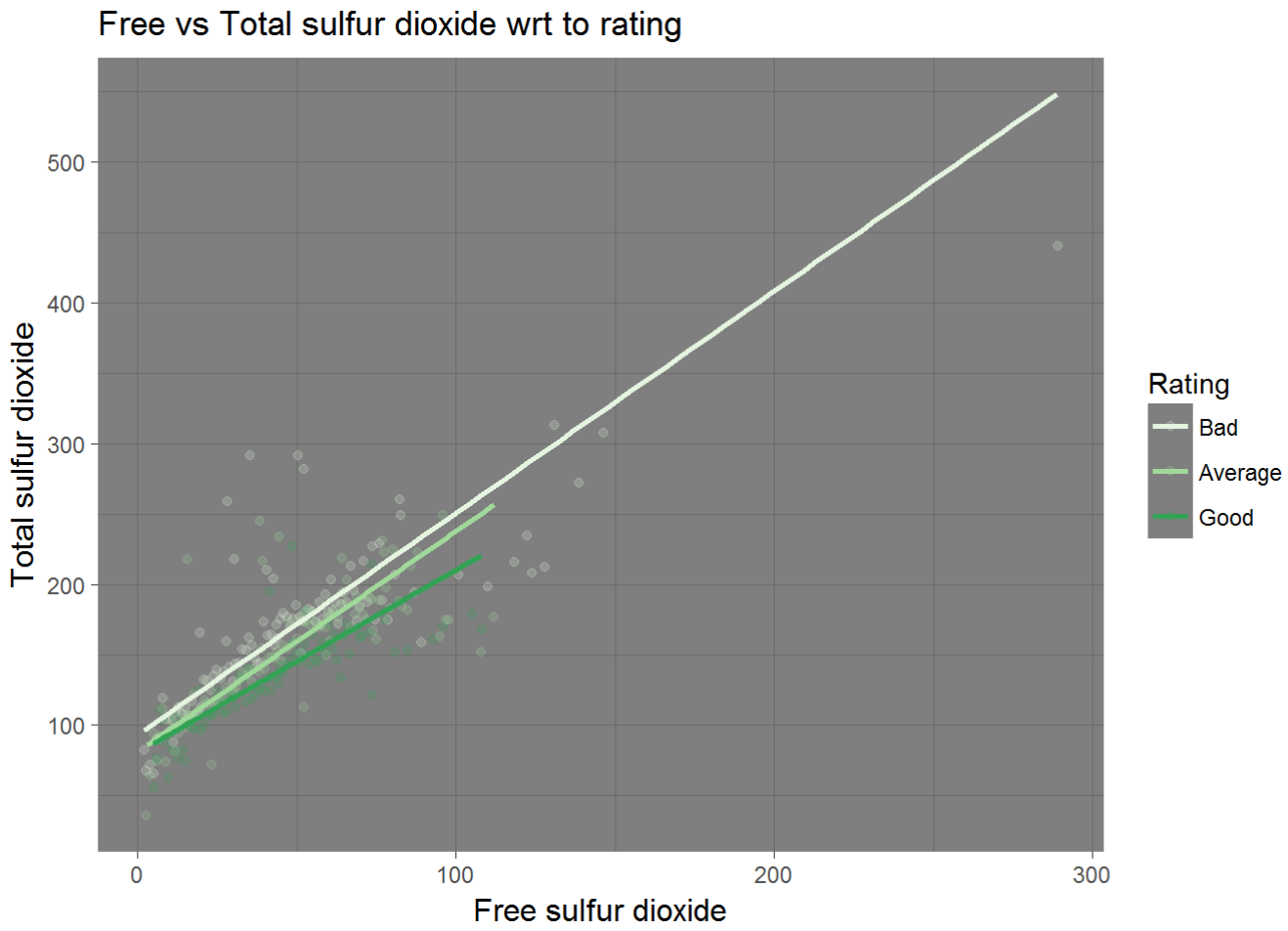
In wine tasting, humans are least sensitive to the taste of sweetness (in contrast to sensitivity to bitterness or sourness) with the majority of the population being able to detect sugar or “sweetness” in wines between 1% and 2.5% residual sugar. Additionally, other components of wine such as acidity and tannins can mask the perception of sugar in the wine.

First I will like to explore the relationship of residual sugar with density and does it impact the wine rating



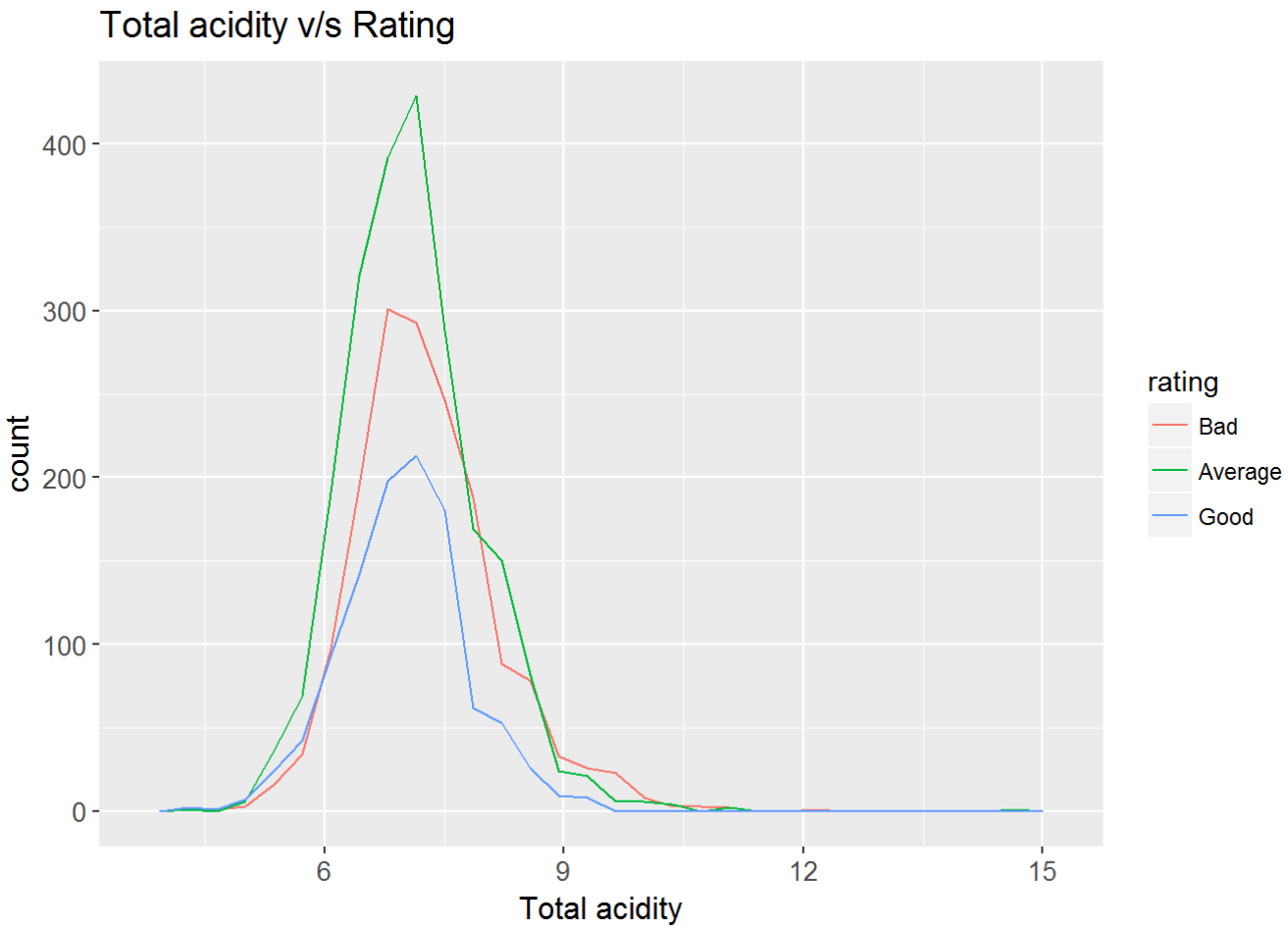
This plot explores the relationship between wine sweetness to residual sugar per density and how it impacts the wine rating. It shows that as sweetness level of wines increase, the density as well increases. And it shows that sweet wines have higher sugar/density and are just averagely rated.

Lets go upon exploring the relationship between free and total sulfur dioxide with respect to the wine rating



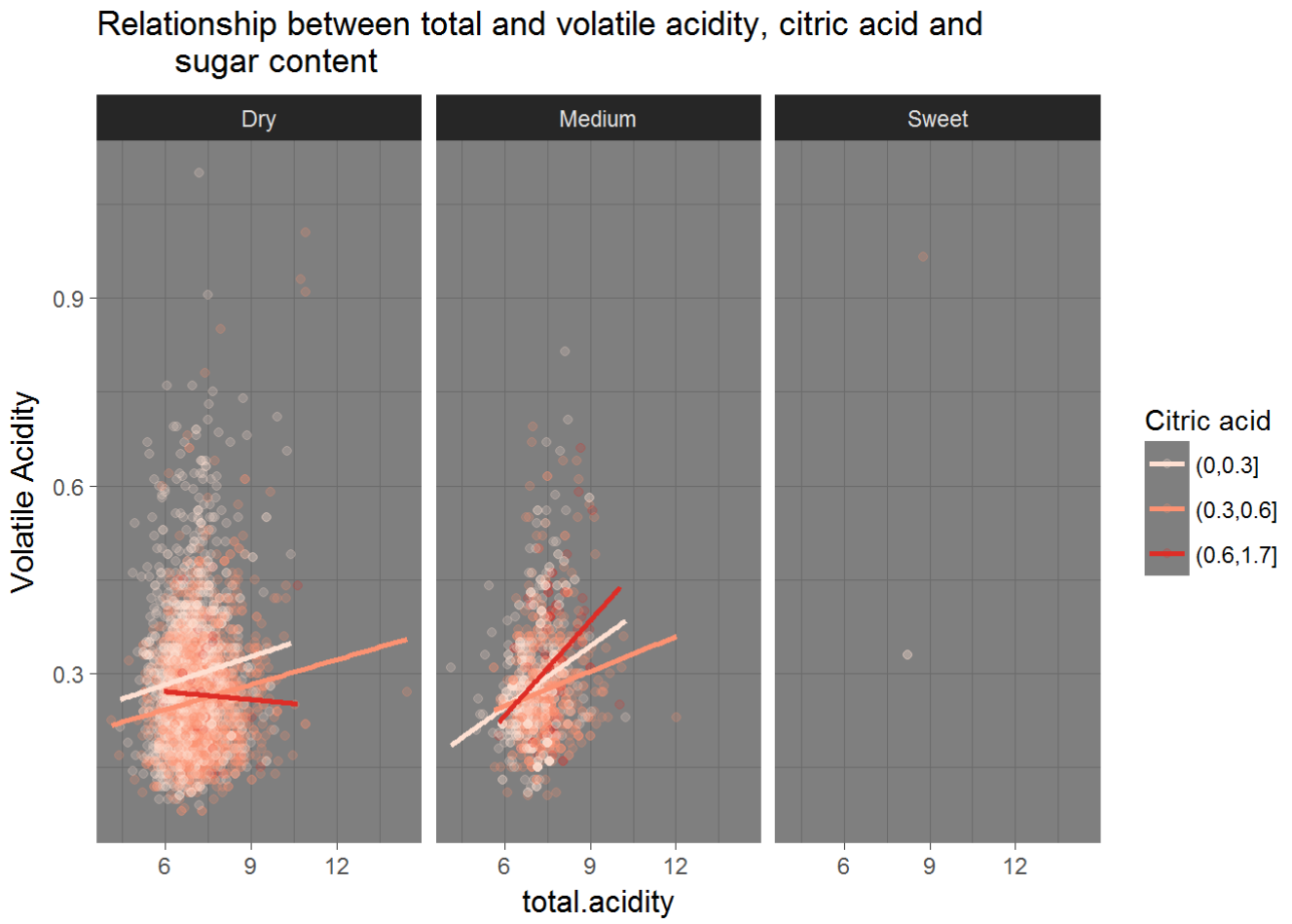
This plot shows that “Good” rated wines have lower proportion of free v/s total sulfur dioxide proportion as compared to “Bad” wines.This plot shows that for “Good” rated wines, the proportion of free sulfur dioxide vs total sulfur dioxide has to be less. As the wine quality increases the proportion of free vs total sulfur dioxide decreases

Acids are one of 4 fundamental traits in wine (the others are tannin, alcohol and sweetness). Acidity gives wine its tart and sour taste. SO let us go on by exploring the total acidity(fixed+volatile) with respect to wine rating.



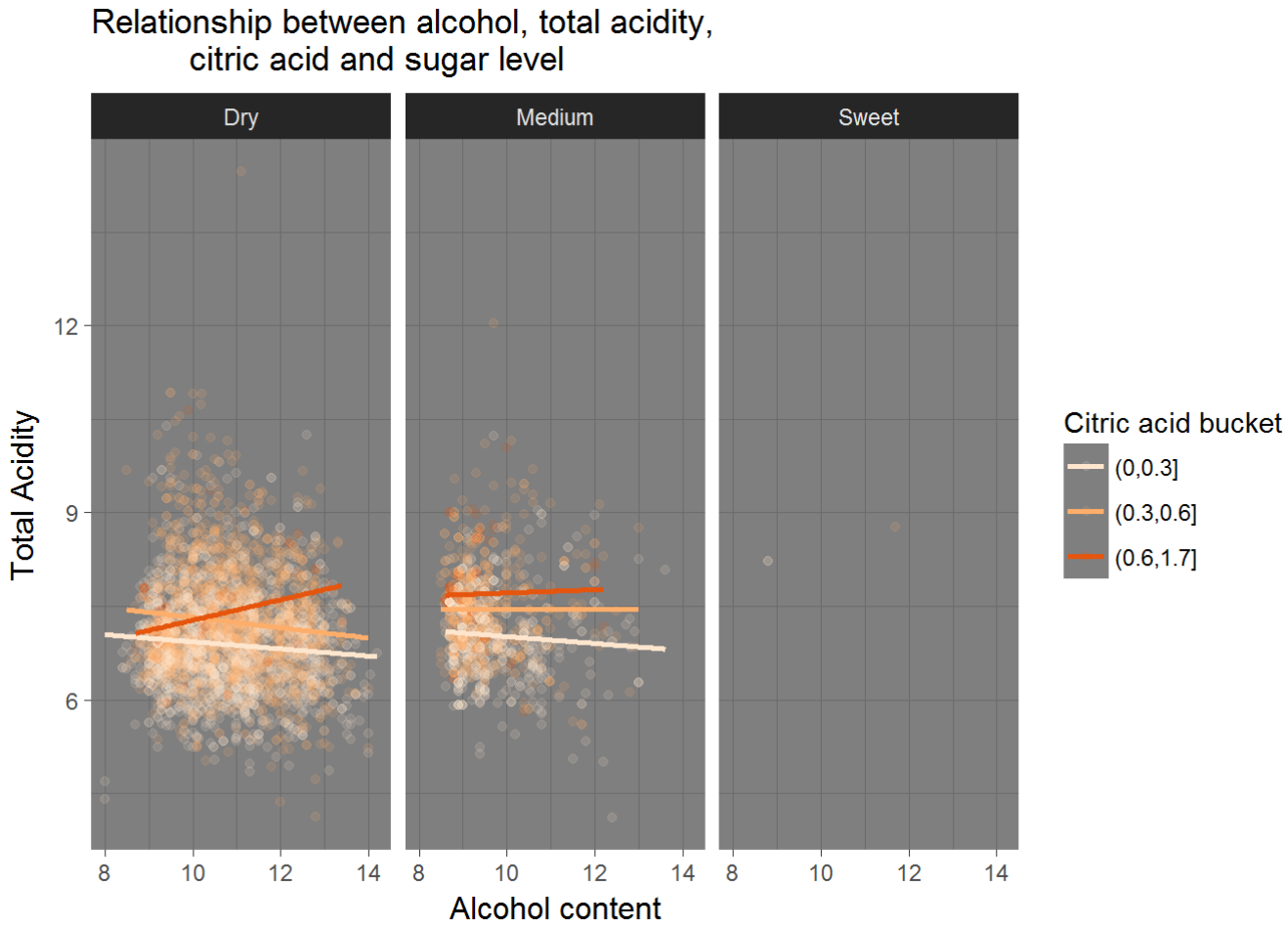
This plot shows the density curve for differently rated wines - Good, Bad and average with respect to their total acidity. Here the density of total acidity of average rated wines is the maximum, followed by bad wines and then by good wines.

Citric acid is a weak organic acid, which is often used as a natural preservative or additive to food or drink to add a sour taste to food.In terms of wines it is often added to wines to increase acidity,or to give a “fresh” flavor. Let us explore the relationship of citric acid with total and volatile acidity and corresponding sugar level in the wines



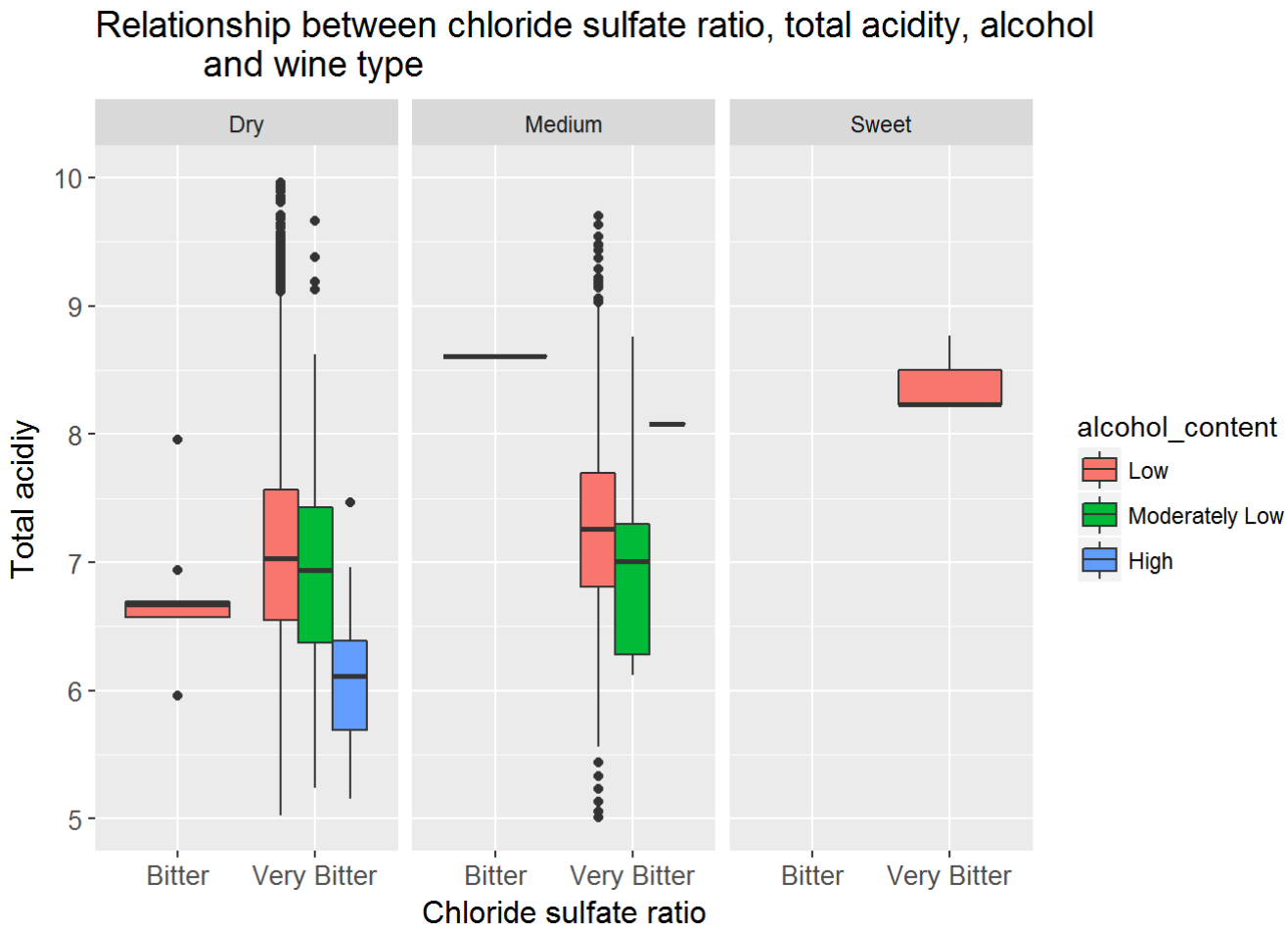
This multi variate scatter plot groups wines based on sugar content and citric acid content. For dry wines, most of citric acid content falls within bucket of (0,0.3] with low level of total acidity and medium volatile acidity. Dry wines also contain prominent wines with citric acid in bucket of (0.3,0.6] which are scattered for volatile acidity and fixed acidity. Medium wines have higher proportion of wines in (0.6,1.7] citric acid bucket as compared to dry wines. And for wines with content between 0-0.6 citric acid content is medium with total and volatile acidity proportion.

Let us explore the relationship of citric acid with alcohol and total acidity and sweetness.

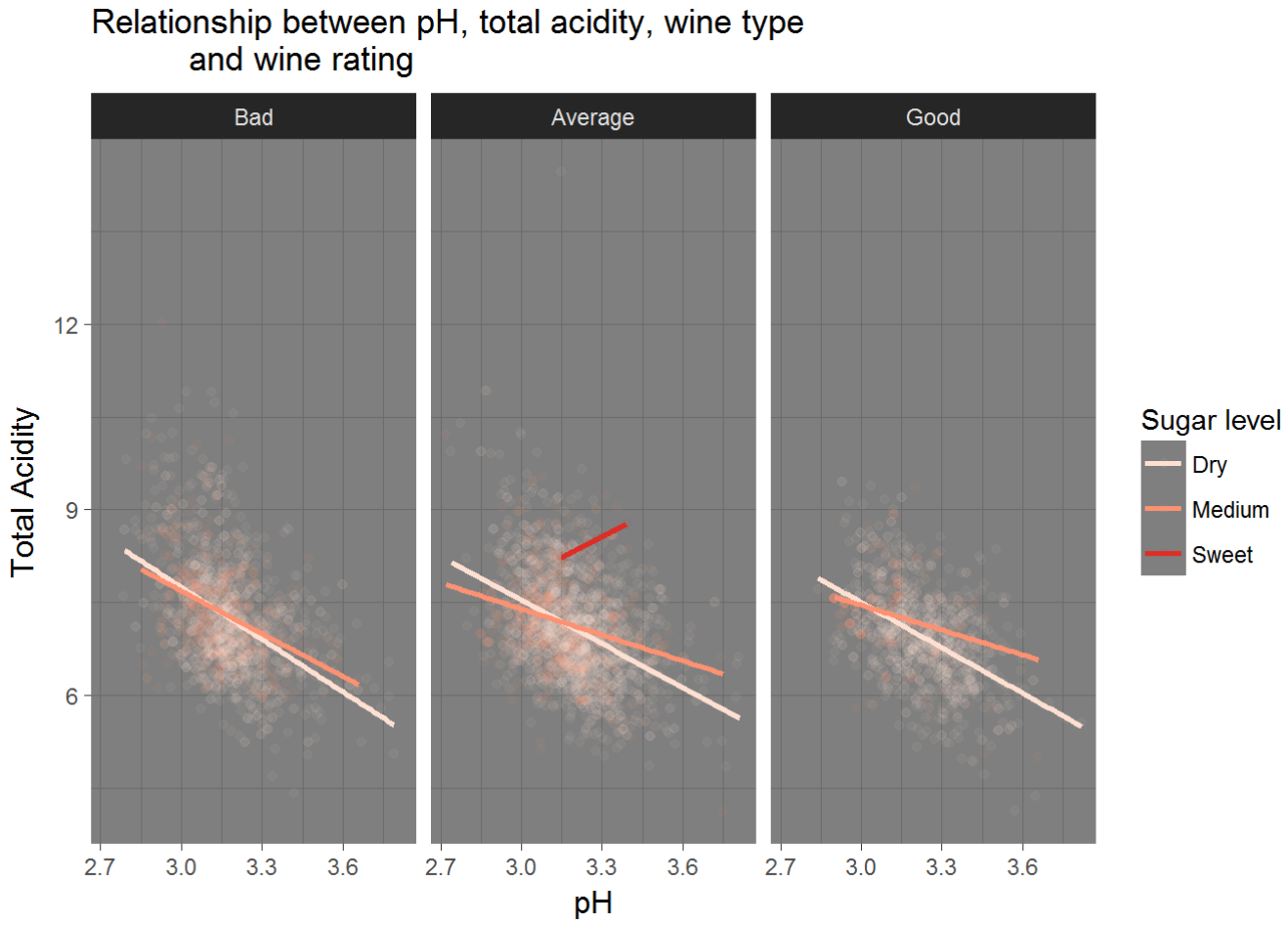


This plot shows relationship among the variables i.e. alcohol content, total acidity wrt to citric acid and sweetness in wine. This shows that dry wines have normally moderately low level of alcohol and low/medium level of
citric acid. Medium sweeted wines also have moderately low level of alchol with large number of wines containing medium level of citric acid.

Chloride concentrations in excess tend to give a full malty taste while Sulfate ions, in contrast, tend to accentuate hop flavors and bitterness. So let us analyze this relationship wrt to alcohol, the total acidity and sugar level to determine the balance in flavours

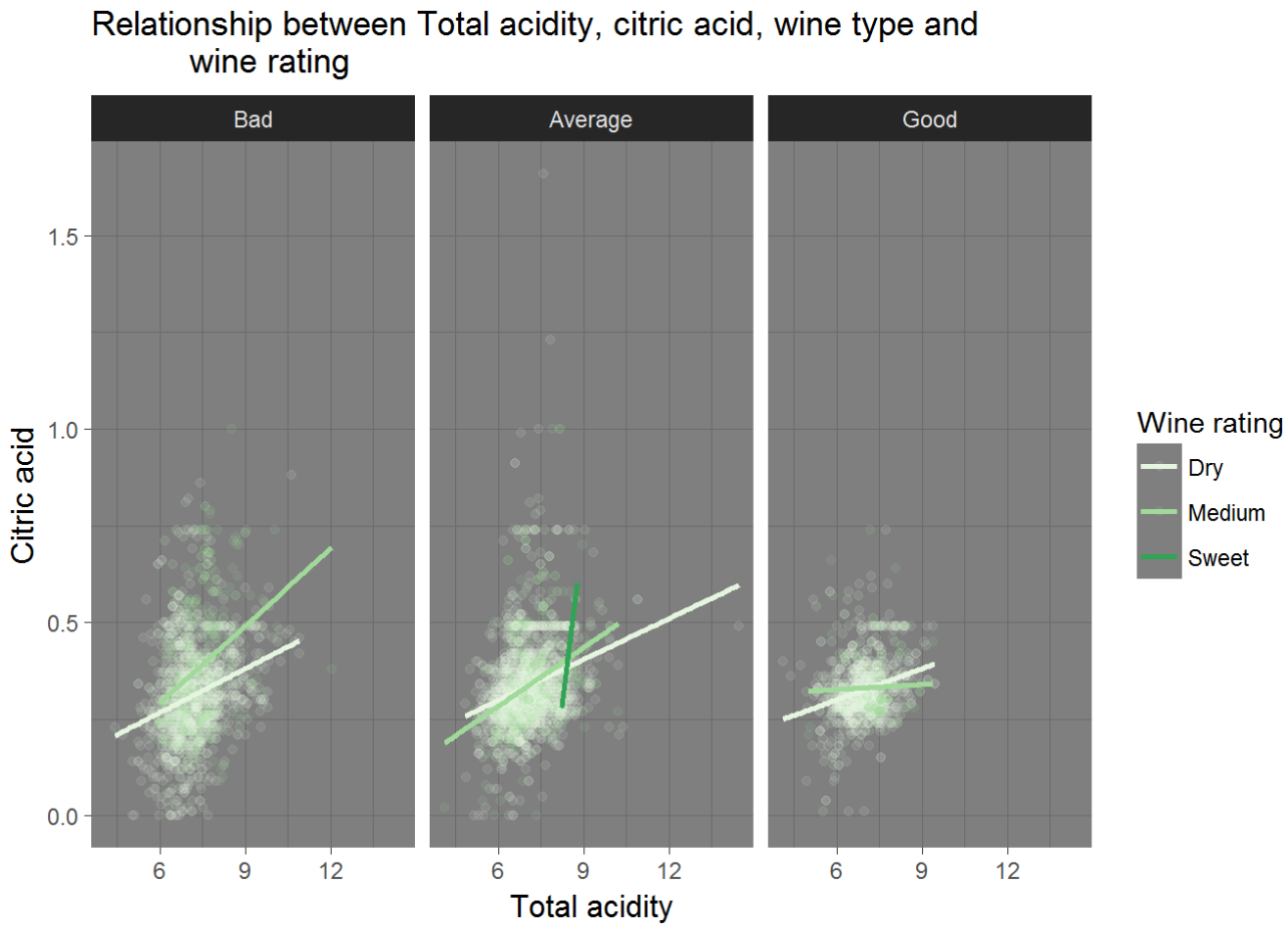


This plot explores a weak relationship between chloride sulfate ratio v/s total acidity wrt to alcohol content and sweetness. Here dry wines are normally very bitter with low level of alcohol and medium level of acidity. Medium sweeted wines are also very bitter with low level of alcohol with high level of acidity. Even sweet wines have high proprortion of chloride/sulfate ratio to balance out the sweetness

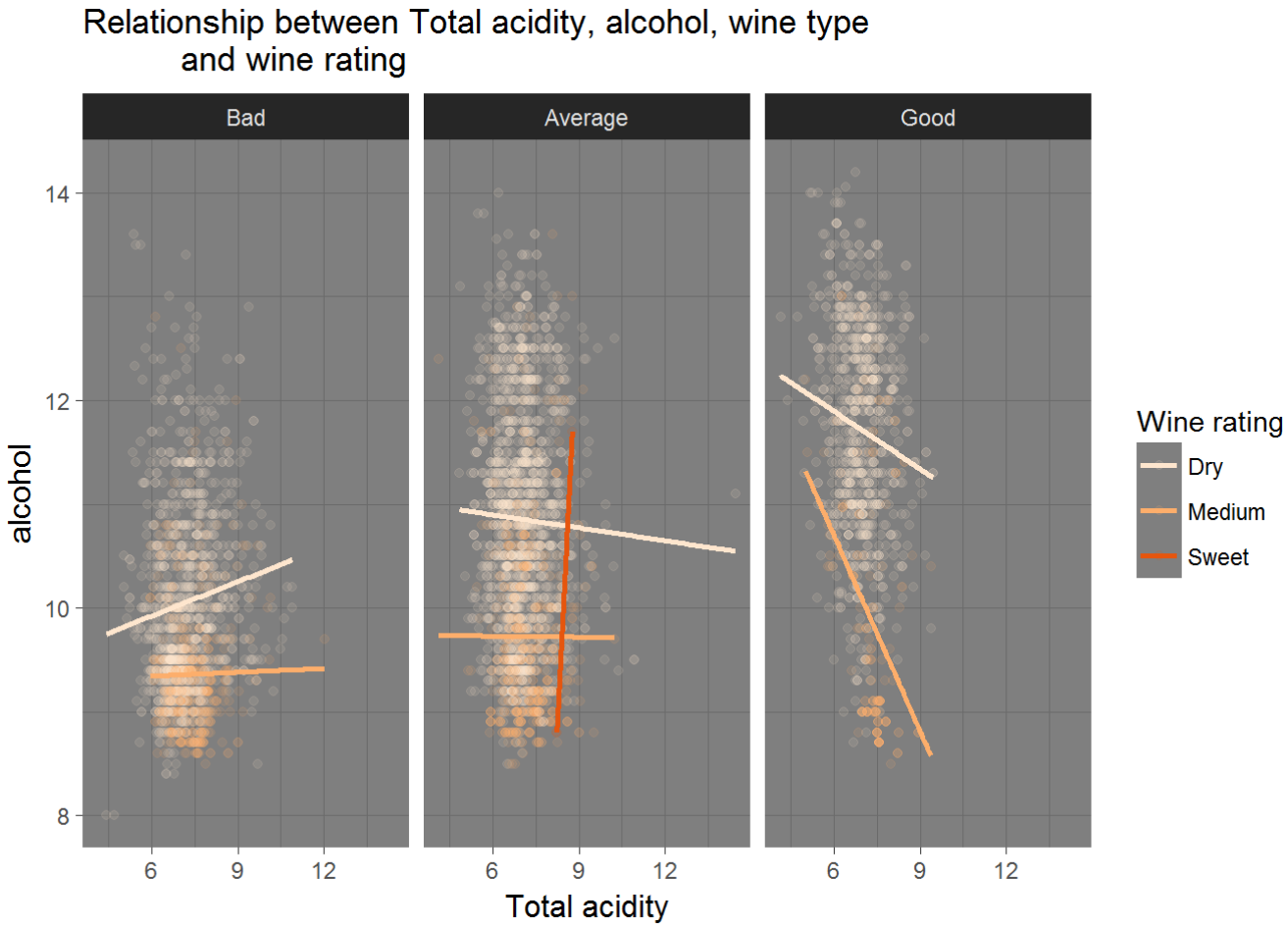


This plot explores the relationship of pH wrt to fixed acidity on basis of sweetness and wine rating. This shows that good wines haves low level of total acidity and are normally dry/medium level of sweetness.And bad wines more dry/medium level of sweetness as compared to good wines.

Let us determine the relationship if any present between total acidity, citric acid, sweetness and wine rating

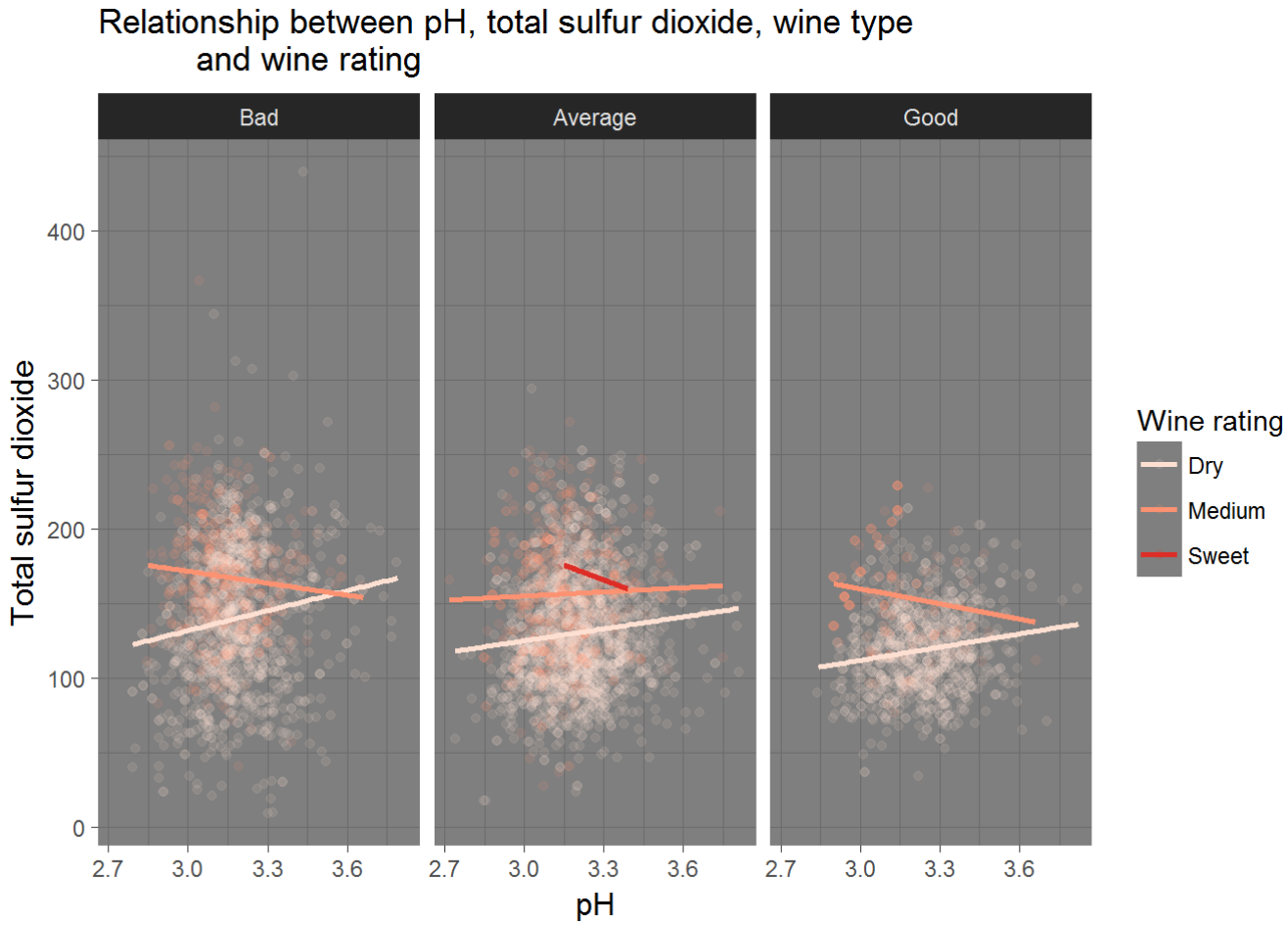


This plot explores the relationship between citric and total acidity wrt to sweetness and wine rating. Here for good wines, a very small proportion is medium sweet and else are normally dry. The citric acid proportion also is less. For bad wines, high percentage of wines are dry followed by medium wines. The citric acid level is high as compared to good wines.



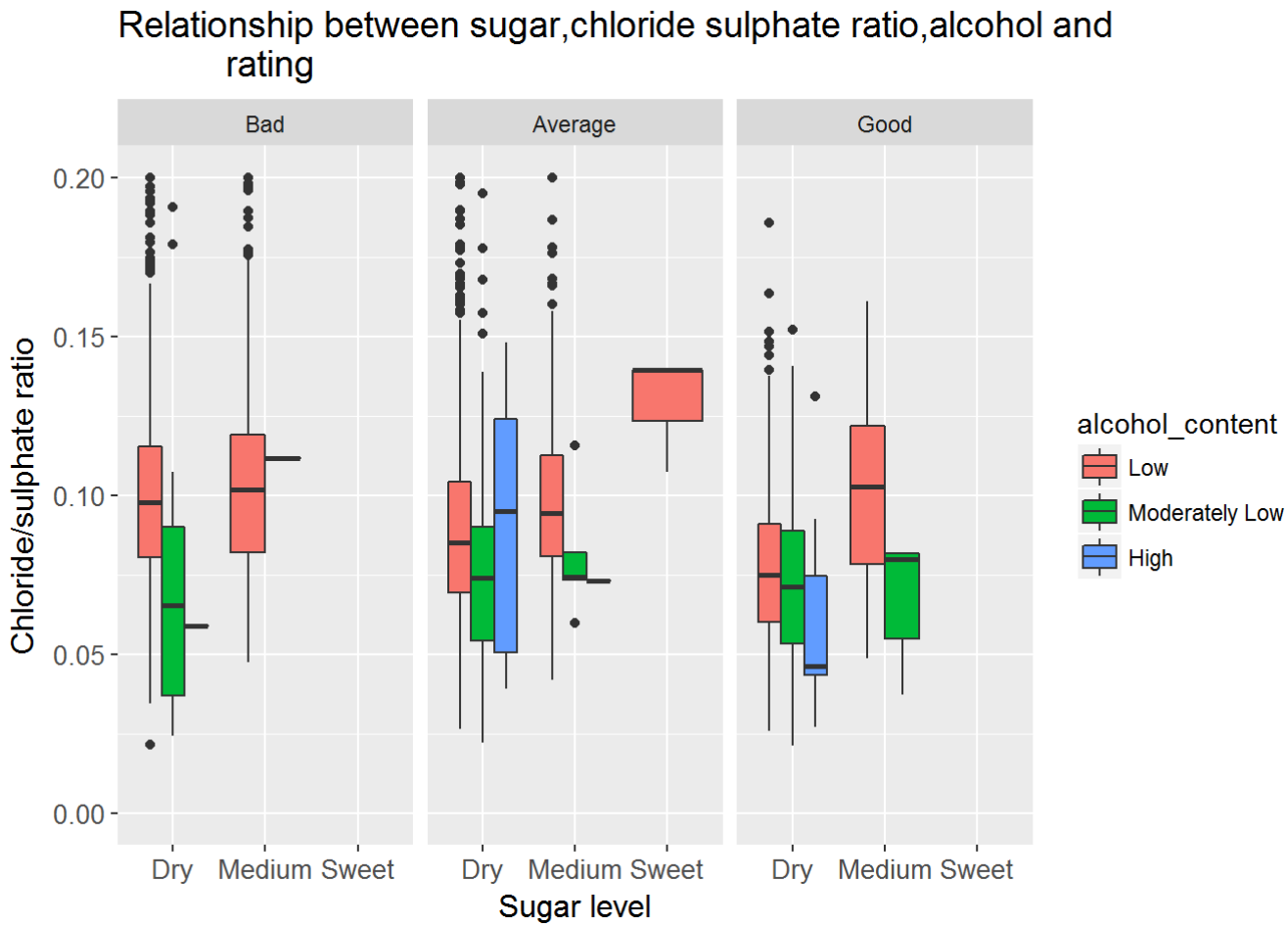
This plot explores the relationship between total acidity and alcohol wrt to wine sweetness and rating. A very low proportion falls into “good” wine category followed by bad rating. Good and bad wines have medium/dry level of sweetness with alcohol in the range 8-14.

Today, the use of sulfur dioxide (SO2) is widely accepted as a useful winemaking aide. It is used as a preservative because of its anti-oxidative and anti-microbial properties in wine. Let us explore this relationship with wrt to ph, sugar and wine rating.



This plot explores the relationship between pH and total sulfur dioxide wrt to sweetness and wine rating. Here the good wines have average count of total sulfur dioixe with dry/medium level of sweetness. The bad wines have large number of dry wines with scattered pH and high total sulfur dioxide.

Let us explore a relationship between sugar level/ chloride sulphate ratio, alcohol level and how it impacts the wine rating



From this plot we can see that good rated wines have a balance of sugar and saltiness with proportionate level of sugar and chloride/sulphate ratio and have normally lower level level of alcohols. Bad wines have high level of sugar and chloride/sulphates and low level of alcohols

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?

In this part of investigation, I explored relationship among multiple variables int the data set. I found that as sugar content increases in wine, the density increases linearly and the wine quality decreases.Also high rated wines have lower proportion of free v/s total sulfur dioxide proportion as compared to poor rated wines.As total acid density increases in wine the quality decreases, average rated wines are more dry as compared to bad wines which are more mediumly sweeted.Good rated wines have more balance of total acidity with citric acid ratio, dry wines have normally moderately low level of alcohol with high total acidity and medium level of citric acid. Dry wines have normally moderately low level of alcohol and low/medium level of citric acid. Medium sweeted wines also have moderately low level of alchol with large number of wines

containing medium level of citric acid. From my analysis I cannot pin point a single variable which could directly impact the quality, as they say its more about balance! It has to be a balance of acidity, sweetness, alcohol and sugar.

Were there any interesting or surprising interactions between features?

An interesting observation made is that no single variable can by itself be responsible for the quality of the wine. Wines need to be a balanced in terms of acidity, alcohol, sugar, saltiness etc. I found that a combination of medium total acidity and dry/medium sweetness, low level of citric acid, less of total sulfur dioxide, low/moderately low level of alcohol wrt to low level of residual sugar and chloride/sulphate ratio would be atleast required to increase the chance of that wine sample to be among the higher quality wine set.

To conclude I would like to add some inferential statistical tests to determine statistical significance of the various variables wrt wine quality. **Analysis of variance (ANOVA)** is a collection of statistical models used to analyze the differences among group means and their associated procedures (such as “variation” among and between groups). It is a quick, easy way to rule out un-needed variables that contribute little to the explanation of a dependent variable. So let us determine the relationship of quality with newly added categorical variables like alcohol_content, citric_acid_bucket,wine.type, total_sulfur_dioxide_bucket and total.acidity

```
##                                                                      Df
## alcohol_content                                                       2
## citric_acid_bucket                                                    2
## wine.type                                                             2
## total_sulfur_dioxide_bucket                                           2
## total.acidity                                                         1
## alcohol_content:citric_acid_bucket                                    3
## alcohol_content:wine.type                                             2
## citric_acid_bucket:wine.type                                          3
## alcohol_content:total_sulfur_dioxide_bucket                           3
## citric_acid_bucket:total_sulfur_dioxide_bucket                        4
## wine.type:total_sulfur_dioxide_bucket                                 2
## alcohol_content:total.acidity                                         2
## citric_acid_bucket:total.acidity                                      2
## wine.type:total.acidity                                               1
## total_sulfur_dioxide_bucket:total.acidity                             2
## alcohol_content:citric_acid_bucket:wine.type                          1
## alcohol_content:citric_acid_bucket:total_sulfur_dioxide_bucket        3
## alcohol_content:wine.type:total_sulfur_dioxide_bucket                 1
## citric_acid_bucket:wine.type:total_sulfur_dioxide_bucket              4
## alcohol_content:citric_acid_bucket:total.acidity                      3
## alcohol_content:wine.type:total.acidity                               1
## citric_acid_bucket:wine.type:total.acidity                            2
## alcohol_content:total_sulfur_dioxide_bucket:total.acidity             2
## citric_acid_bucket:total_sulfur_dioxide_bucket:total.acidity          4
## wine.type:total_sulfur_dioxide_bucket:total.acidity                   2
## alcohol_content:citric_acid_bucket:wine.type:total_sulfur_dioxide_bucket   1
## alcohol_content:citric_acid_bucket:wine.type:total.acidity            1
## alcohol_content:citric_acid_bucket:total_sulfur_dioxide_bucket:total.acidity   1
## alcohol_content:wine.type:total_sulfur_dioxide_bucket:total.acidity    1
## citric_acid_bucket:wine.type:total_sulfur_dioxide_bucket:total.acidity    3
## Residuals                                                          4815
##                                                                  Sum Sq
## alcohol_content                                                     298
## citric_acid_bucket                                                   26
## wine.type                                                             7
## total_sulfur_dioxide_bucket                                          27
## total.acidity                                                        48
## alcohol_content:citric_acid_bucket                                    2
## alcohol_content:wine.type                                             1
## citric_acid_bucket:wine.type                                         17
## alcohol_content:total_sulfur_dioxide_bucket                           7
## citric_acid_bucket:total_sulfur_dioxide_bucket                        7
## wine.type:total_sulfur_dioxide_bucket                                 7
## alcohol_content:total.acidity                                         0
## citric_acid_bucket:total.acidity                                      7
## wine.type:total.acidity                                               5
## total_sulfur_dioxide_bucket:total.acidity                            10
## alcohol_content:citric_acid_bucket:wine.type                          1
## alcohol_content:citric_acid_bucket:total_sulfur_dioxide_bucket        2
## alcohol_content:wine.type:total_sulfur_dioxide_bucket                 1
## citric_acid_bucket:wine.type:total_sulfur_dioxide_bucket             12
## alcohol_content:citric_acid_bucket:total.acidity                      1
```

```
## alcohol_content:wine.type:total.acidity                                          0
## citric_acid_bucket:wine.type:total.acidity                                       6
## alcohol_content:total_sulfur_dioxide_bucket:total.acidity                        3
## citric_acid_bucket:total_sulfur_dioxide_bucket:total.acidity                     3
## wine.type:total_sulfur_dioxide_bucket:total.acidity                              6
## alcohol_content:citric_acid_bucket:wine.type:total_sulfur_dioxide_bucket         0
## alcohol_content:citric_acid_bucket:wine.type:total.acidity                       1
## alcohol_content:citric_acid_bucket:total_sulfur_dioxide_bucket:total.acidity     1
## alcohol_content:wine.type:total_sulfur_dioxide_bucket:total.acidity              1
## citric_acid_bucket:wine.type:total_sulfur_dioxide_bucket:total.acidity           0
## Residuals                                                                     3309
##                                                                             Mean Sq
## alcohol_content                                                             148.78
## citric_acid_bucket                                                           12.89
## wine.type                                                                     3.32
## total_sulfur_dioxide_bucket                                                  13.60
## total.acidity                                                                48.02
## alcohol_content:citric_acid_bucket                                            0.68
## alcohol_content:wine.type                                                     0.29
## citric_acid_bucket:wine.type                                                  5.67
## alcohol_content:total_sulfur_dioxide_bucket                                   2.45
## citric_acid_bucket:total_sulfur_dioxide_bucket                                1.80
## wine.type:total_sulfur_dioxide_bucket                                         3.57
## alcohol_content:total.acidity                                                 0.06
## citric_acid_bucket:total.acidity                                              3.42
## wine.type:total.acidity                                                       5.16
## total_sulfur_dioxide_bucket:total.acidity                                     4.82
## alcohol_content:citric_acid_bucket:wine.type                                  0.84
## alcohol_content:citric_acid_bucket:total_sulfur_dioxide_bucket                0.75
## alcohol_content:wine.type:total_sulfur_dioxide_bucket                         0.57
## citric_acid_bucket:wine.type:total_sulfur_dioxide_bucket                      2.90
## alcohol_content:citric_acid_bucket:total.acidity                              0.32
## alcohol_content:wine.type:total.acidity                                       0.04
## citric_acid_bucket:wine.type:total.acidity                                    3.23
## alcohol_content:total_sulfur_dioxide_bucket:total.acidity                     1.69
## citric_acid_bucket:total_sulfur_dioxide_bucket:total.acidity                  0.82
## wine.type:total_sulfur_dioxide_bucket:total.acidity                           3.04
## alcohol_content:citric_acid_bucket:wine.type:total_sulfur_dioxide_bucket      0.01
## alcohol_content:citric_acid_bucket:wine.type:total.acidity                    1.21
## alcohol_content:citric_acid_bucket:total_sulfur_dioxide_bucket:total.acidity  1.14
## alcohol_content:wine.type:total_sulfur_dioxide_bucket:total.acidity           0.52
## citric_acid_bucket:wine.type:total_sulfur_dioxide_bucket:total.acidity        0.12
## Residuals                                                                     0.69
##                                                                             F value
## alcohol_content                                                             216.509
## citric_acid_bucket                                                           18.764
## wine.type                                                                     4.826
## total_sulfur_dioxide_bucket                                                  19.790
## total.acidity                                                                69.884
## alcohol_content:citric_acid_bucket                                            0.987
## alcohol_content:wine.type                                                     0.428
## citric_acid_bucket:wine.type                                                  8.249
## alcohol_content:total_sulfur_dioxide_bucket                                   3.564
## citric_acid_bucket:total_sulfur_dioxide_bucket                                2.621
```

```
## wine.type:total_sulfur_dioxide_bucket                                    5.201
## alcohol_content:total.acidity                                            0.080
## citric_acid_bucket:total.acidity                                         4.979
## wine.type:total.acidity                                                  7.514
## total_sulfur_dioxide_bucket:total.acidity                                7.018
## alcohol_content:citric_acid_bucket:wine.type                             1.220
## alcohol_content:citric_acid_bucket:total_sulfur_dioxide_bucket           1.090
## alcohol_content:wine.type:total_sulfur_dioxide_bucket                    0.835
## citric_acid_bucket:wine.type:total_sulfur_dioxide_bucket                 4.225
## alcohol_content:citric_acid_bucket:total.acidity                         0.462
## alcohol_content:wine.type:total.acidity                                  0.063
## citric_acid_bucket:wine.type:total.acidity                               4.703
## alcohol_content:total_sulfur_dioxide_bucket:total.acidity                2.462
## citric_acid_bucket:total_sulfur_dioxide_bucket:total.acidity             1.196
## wine.type:total_sulfur_dioxide_bucket:total.acidity                      4.430
## alcohol_content:citric_acid_bucket:wine.type:total_sulfur_dioxide_bucket  0.018
## alcohol_content:citric_acid_bucket:wine.type:total.acidity               1.767
## alcohol_content:citric_acid_bucket:total_sulfur_dioxide_bucket:total.acidity  1.664
## alcohol_content:wine.type:total_sulfur_dioxide_bucket:total.acidity      0.752
## citric_acid_bucket:wine.type:total_sulfur_dioxide_bucket:total.acidity   0.171
## Residuals
##                                                                             Pr(>F)
## alcohol_content                                                          < 2e-16
## citric_acid_bucket                                                      7.63e-09
## wine.type                                                               0.008055
## total_sulfur_dioxide_bucket                                             2.76e-09
## total.acidity                                                            < 2e-16
## alcohol_content:citric_acid_bucket                                      0.397825
## alcohol_content:wine.type                                               0.651780
## citric_acid_bucket:wine.type                                            1.80e-05
## alcohol_content:total_sulfur_dioxide_bucket                             0.013589
## citric_acid_bucket:total_sulfur_dioxide_bucket                          0.033166
## wine.type:total_sulfur_dioxide_bucket                                   0.005542
## alcohol_content:total.acidity                                           0.923037
## citric_acid_bucket:total.acidity                                        0.006914
## wine.type:total.acidity                                                 0.006144
## total_sulfur_dioxide_bucket:total.acidity                               0.000904
## alcohol_content:citric_acid_bucket:wine.type                            0.269505
## alcohol_content:citric_acid_bucket:total_sulfur_dioxide_bucket          0.351929
## alcohol_content:wine.type:total_sulfur_dioxide_bucket                   0.360997
## citric_acid_bucket:wine.type:total_sulfur_dioxide_bucket                0.002044
## alcohol_content:citric_acid_bucket:total.acidity                        0.708810
## alcohol_content:wine.type:total.acidity                                 0.801819
## citric_acid_bucket:wine.type:total.acidity                              0.009109
## alcohol_content:total_sulfur_dioxide_bucket:total.acidity               0.085367
## citric_acid_bucket:total_sulfur_dioxide_bucket:total.acidity            0.310299
## wine.type:total_sulfur_dioxide_bucket:total.acidity                     0.011967
## alcohol_content:citric_acid_bucket:wine.type:total_sulfur_dioxide_bucket 0.893645
## alcohol_content:citric_acid_bucket:wine.type:total.acidity              0.183785
## alcohol_content:citric_acid_bucket:total_sulfur_dioxide_bucket:total.acidity 0.197112
## alcohol_content:wine.type:total_sulfur_dioxide_bucket:total.acidity     0.386018
## citric_acid_bucket:wine.type:total_sulfur_dioxide_bucket:total.acidity  0.915911
## Residuals
##
```

```
## alcohol_content                                                    ***
## citric_acid_bucket                                                 ***
## wine.type                                                          **
## total_sulfur_dioxide_bucket                                        ***
## total.acidity                                                      ***
## alcohol_content:citric_acid_bucket
## alcohol_content:wine.type
## citric_acid_bucket:wine.type                                       ***
## alcohol_content:total_sulfur_dioxide_bucket                        *
## citric_acid_bucket:total_sulfur_dioxide_bucket                     *
## wine.type:total_sulfur_dioxide_bucket                              **
## alcohol_content:total.acidity
## citric_acid_bucket:total.acidity                                   **
## wine.type:total.acidity                                            **
## total_sulfur_dioxide_bucket:total.acidity                          ***
## alcohol_content:citric_acid_bucket:wine.type
## alcohol_content:citric_acid_bucket:total_sulfur_dioxide_bucket
## alcohol_content:wine.type:total_sulfur_dioxide_bucket
## citric_acid_bucket:wine.type:total_sulfur_dioxide_bucket           **
## alcohol_content:citric_acid_bucket:total.acidity
## alcohol_content:wine.type:total.acidity
## citric_acid_bucket:wine.type:total.acidity                         **
## alcohol_content:total_sulfur_dioxide_bucket:total.acidity          .
## citric_acid_bucket:total_sulfur_dioxide_bucket:total.acidity
## wine.type:total_sulfur_dioxide_bucket:total.acidity                *
## alcohol_content:citric_acid_bucket:wine.type:total_sulfur_dioxide_bucket
## alcohol_content:citric_acid_bucket:wine.type:total.acidity
## alcohol_content:citric_acid_bucket:total_sulfur_dioxide_bucket:total.acidity
## alcohol_content:wine.type:total_sulfur_dioxide_bucket:total.acidity
## citric_acid_bucket:wine.type:total_sulfur_dioxide_bucket:total.acidity
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 19 observations deleted due to missingness
```

In most cases we put significance at the alpha=.05 level, or we require the P value to be less then .05 to be considered statistically significant. Based upon the above results I would say to determine wine quality the following relationships are significant wrt to quality :

- alcohol_content
- citric_acid_bucket
- wine.type
- total.acidity
- citric_acid_bucket:wine.type
- wine.type:total_sulfur_dioxide_bucket
- citric_acid_bucket:total.acidity
- wine.type:total.acidity
- total_sulfur_dioxide_bucket:total.acidity

Final Plots and Summary

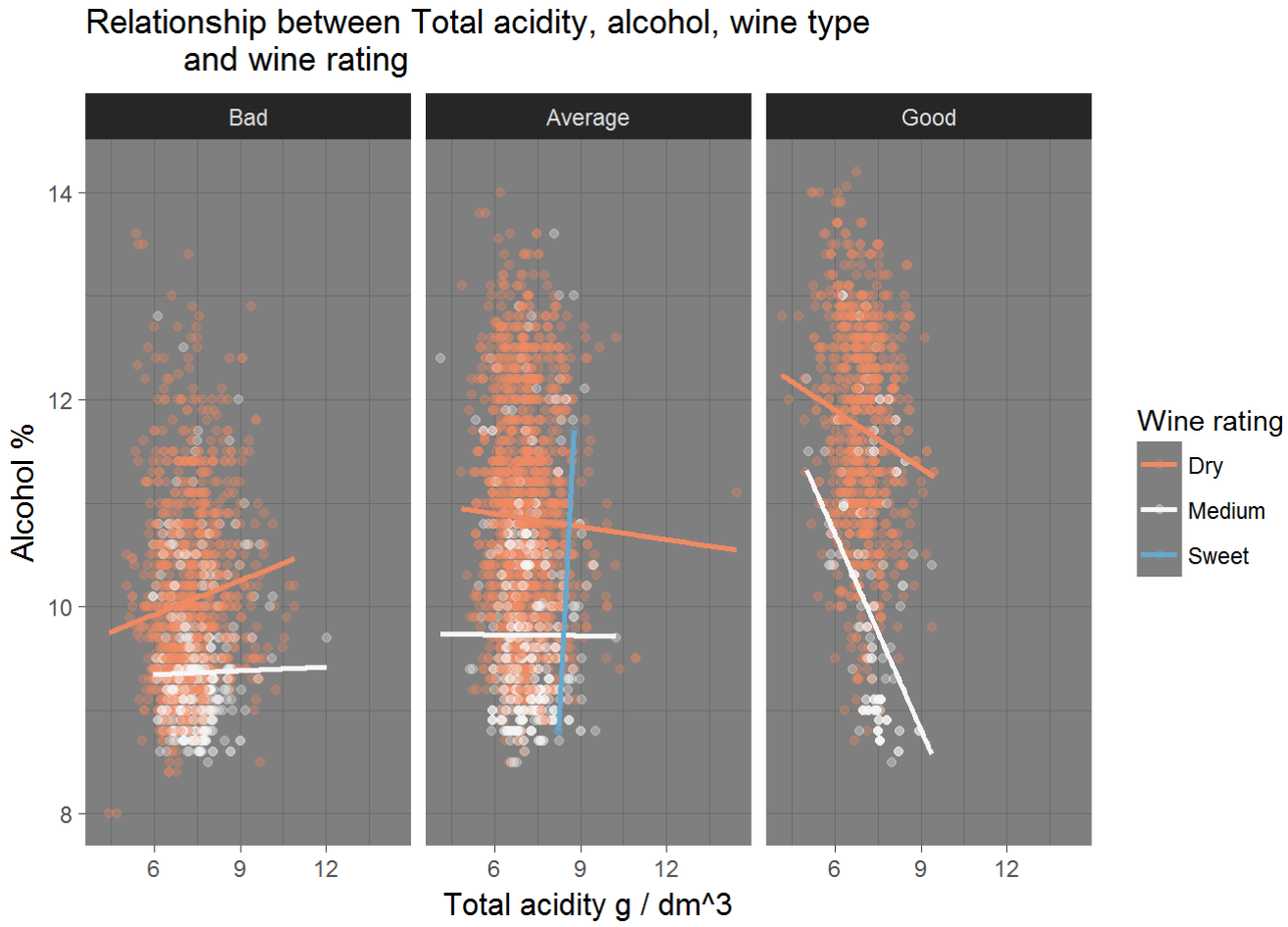
Plot One

```
## white_wine_df$rating: Bad
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.20    9.60    9.85   10.40   13.60
## ------------------------------------------------------------
## white_wine_df$rating: Average
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50    9.60   10.50   10.58   11.40   14.00
## ------------------------------------------------------------
## white_wine_df$rating: Good
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50   10.70   11.50   11.42   12.40   14.20
```

Description One

I have chosen this plot as per correlation matrix for bivariate analysis, alcohol and quality had the highest correlation. This plot shows that “Good”" quality of wine had high median of 12.0 followed by “average”" for 11.8 and “bad” for 10.4 % level of aldohol. This visualisation shows that as alcohol content increase the quality of wine increases.

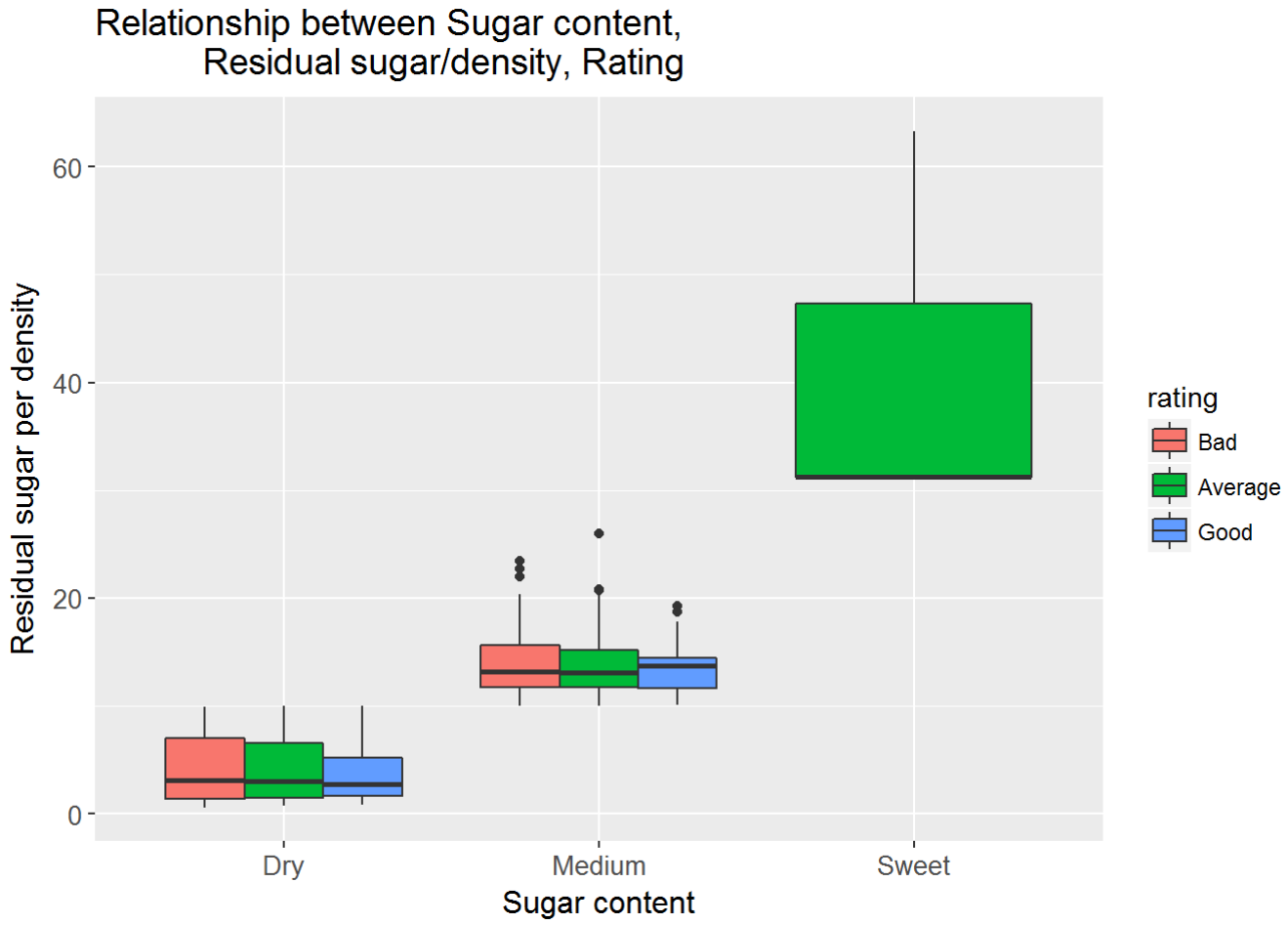
Plot Two



Description Two

This plot explores the relationship among various variables i.e. between total acidity and alcohol wrt to wine sweetness and rating. Since alcohol,sweetness, acidity are some of the major characteristics which influence wine quality I wanted to explore this relationship. A very low proportion of the wines in the dataset fell into “good” wine category followed by “bad” rate in thedataset. Maximum number of wines are averagely rated. It appears high level of alcohol is an important ingredient in wine preparation. Good wines have almost equal number of medium/dry sweeted wines with high level of alcohol. Bad wines have large number of wine samples with medium/dry level of sweetness with alcohol in the range 8-14.

Plot Three



Description Three

This plot explores the relationship between residual sugar wrt to residual sugar per density and how it impacts the wine rating. It shows that as sweetness level of wines increase, the density as well increases. And it shows that sweet wines have higher sugar/density and are just averagely rated while the dry wines have low sugar/density and are either/average/good rated.

Reflection

In this project I explored the white wine dataset which contained information about 4898 wines for 13 variables i.e X,fixed.acidity,volatile.acidity, citric.acid, residual.sugar,chlorides,free.sulfur.dioxide,total.sulfur.dioxide, density,pH,sulphates,alcohol,quality. I did exploratory analysis on the data set for single, two and multiple variables in the dataset using R plots/visualizations.

In the first section of univariate analysis, I plotted the distribution of each of the data set to better understand the sample data of white wines. I created some additional variables such as rating based upon the wine quality,total acidity which was the sum of fixed and variable acidity, wine type which determined the sugar level of the wines, grouped total sulfur dioxide and citric acid proportion into various buckets and chloride sulphate ratio.

Certain variables like fixed acidity,chloride and alcohol variables had to be log transformed to make the plot more normally distributed. Other than that citric acid, sulfur dioxide and residual sugar plots x- axis had to be re-adjusted to have a better understanding of the visualization.

Next in the bivariate section, I explored the correlation among the various variables where correlation among residual sugar and density appeared the highest with alcohol and density appeared to be the weakest.The variables having the highest correlation with wine quality were alcohol and the lowest being chlorides.

In the multivariate analysis, I plotted couple of visualizations which explored the relationship of wine quality with residual sugar, alcohol, total acidity, sulfur dioxide, chloride sulfate ratio,pH. An interesting observation made was that no single variable can by itself be responsible for the quality of the wine. Wines need to be a balanced in terms of acidity, alcohol, sugar, saltiness etc. I found that a combination of medium total acidity and dry/medium sweetness, low level of citric acid, less of total sulfur dioxide, low/moderately low level of alcohol wrt to low level of residual sugar and chloride/sulphate ratio would be atleast required to increase the chance of that wine sample to be among the higher quality wine set.

The road-block which I hit in this exercise was choosing the dataset. Initially I was keen on exploring the presidential election dataset and was half way through as well, but I hit a road block in between in the bivariate section where I could not establish correlation among the categorical variables in the section. So I thought of switching to the wine dataset. This was also not a smooth sailing for me as I could not establish a strong correlation among the variables in the multi variate section which directly influenced the wine quality. So I read couple of articles in the Internet to better understand wine preperation and the significance of various chemical constituents and then I was able to move forward in my exploration.

Coming back to the white wine data set, I feel additional research about the data set needs to be carried out considering other factors as well. Because there can be other non chemical factors like location,topography,soil,climate, temperature which might influence the wine preperation and hence impact the wine quality.Future scope includes creation of a linear predictive model for wine analysis.