What Makes a Sound Wine?

```
## [1] "C:/Users/an-user/Desktop/UDA City/Term2/Part2/Final Project"
```

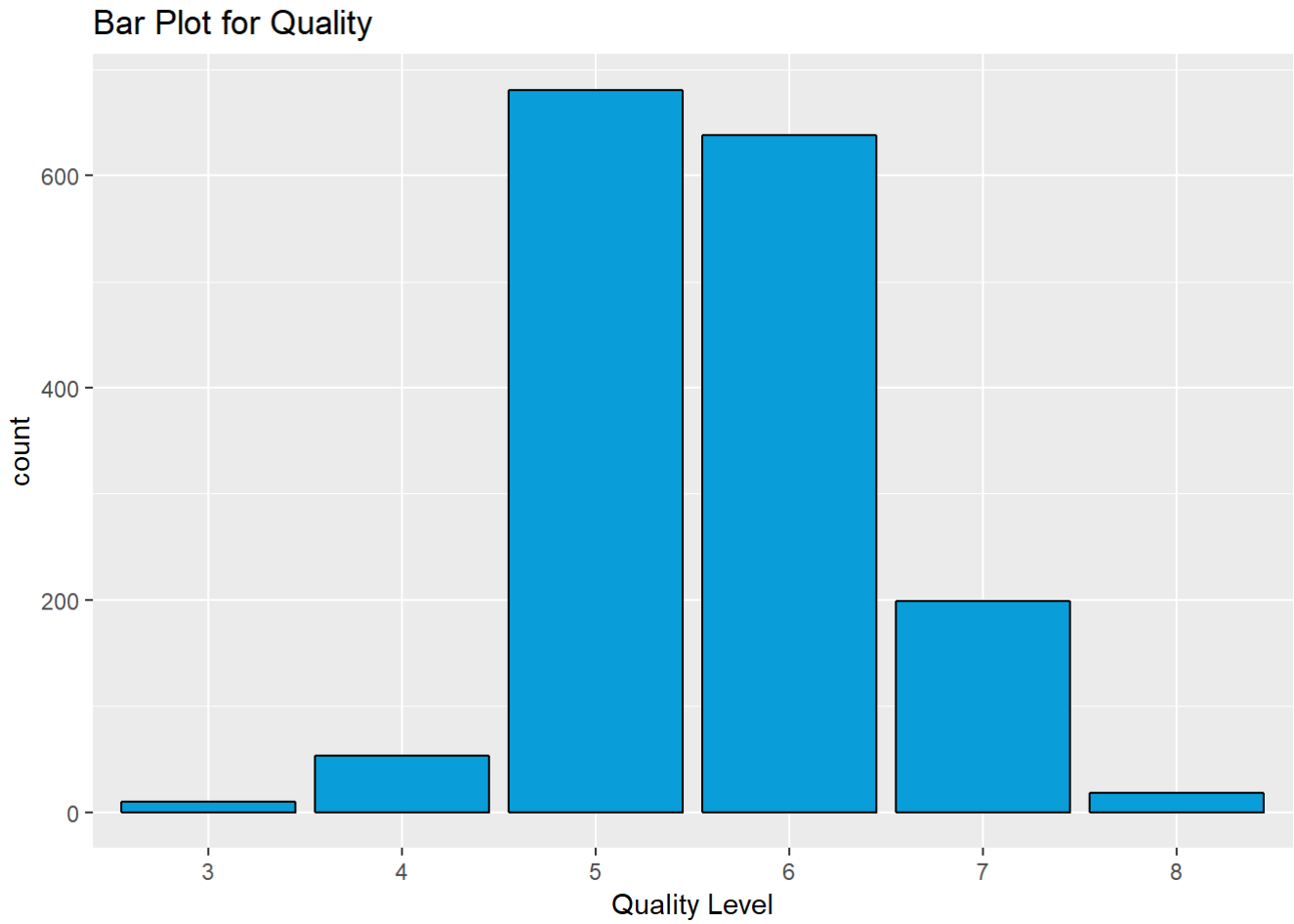
I have chosen red wine data among several options. When it comes to red wine, some people might think that quality of wine would matter the most. Of course, maximizing the quality of wine is important thing, but how efficiently people could brew wine is also significant. With limited amount of resources, if one can make more qualifed wine compared to competitors, that knowledge would be competitive advantage. So in this analysis, the dependent variable would be quality of wine, and other variables as independent variable.

Univariate Plots Section

From this code, I could get overall insight of all variables. Based on these results, I will going to look more deeply into each variable.

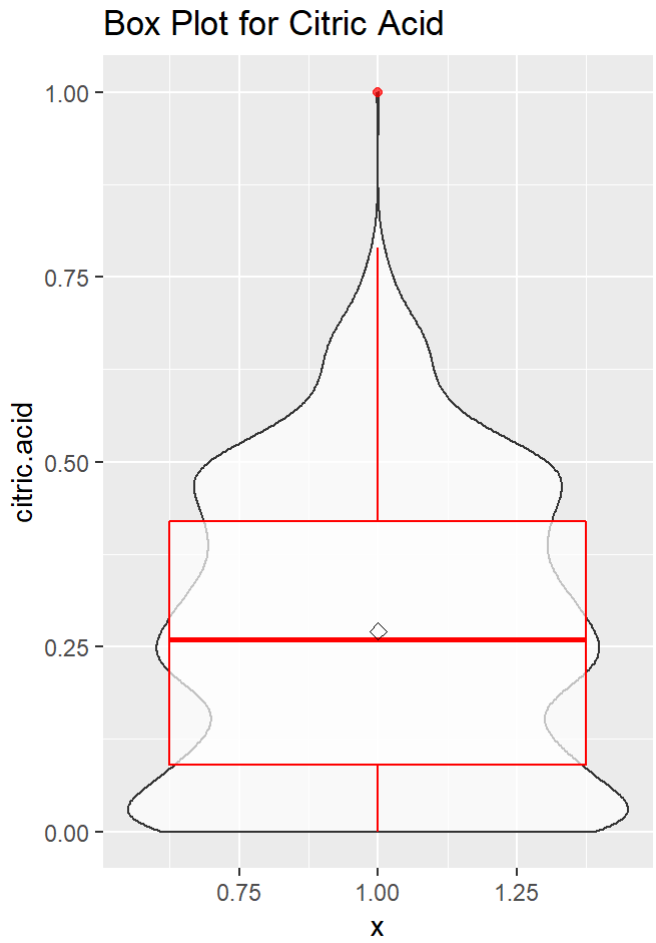
```
##        X          fixed.acidity   volatile.acidity  citric.acid
##  Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000
##  1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090
##  Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260
##  Mean   : 800.0   Mean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

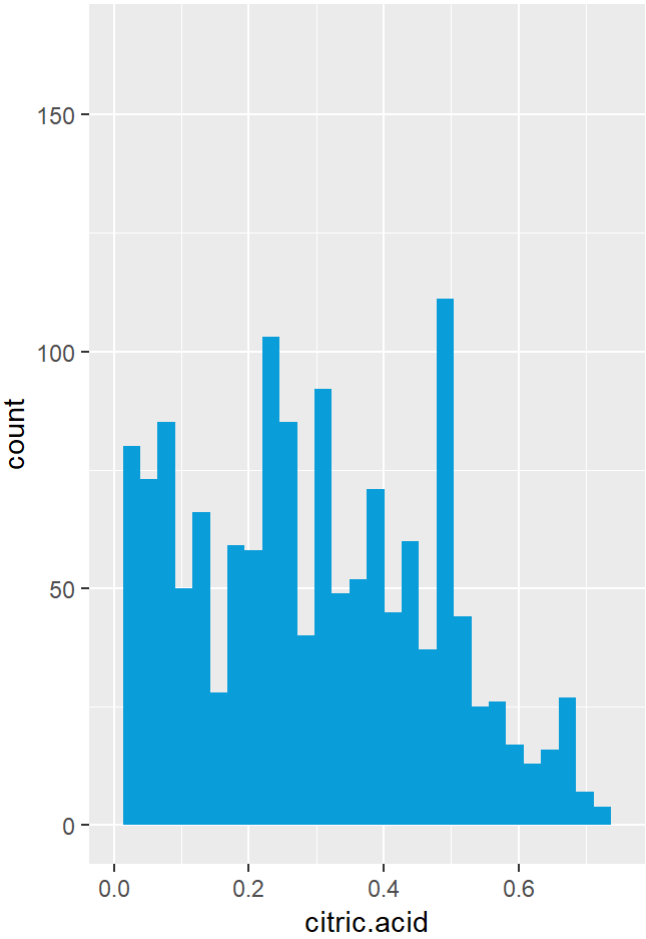
Part1 First of all, we should know the distribution of our dependent variable, which is quality of wines. The histogram showed us that most of qualities were concentrated between 5 and 7.



Part2 Second important varible is citric.acid, which means freshness of wine. This plot revealed that distribution of citric.acid is skewed to the right. This means that some of other variables might have caused citric.acid to follow such distribution.

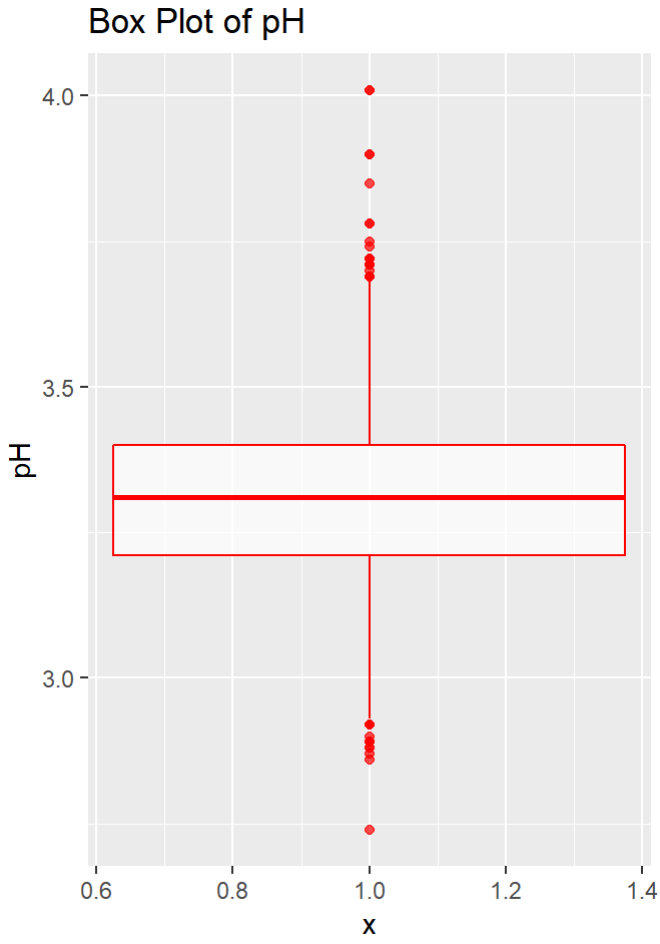
```
## $x
## [1] "citric.acid(g/dm^3)"
##
## $title
## [1] "Distribution of Citric Acid"
##
## attr(,"class")
## [1] "labels"
```

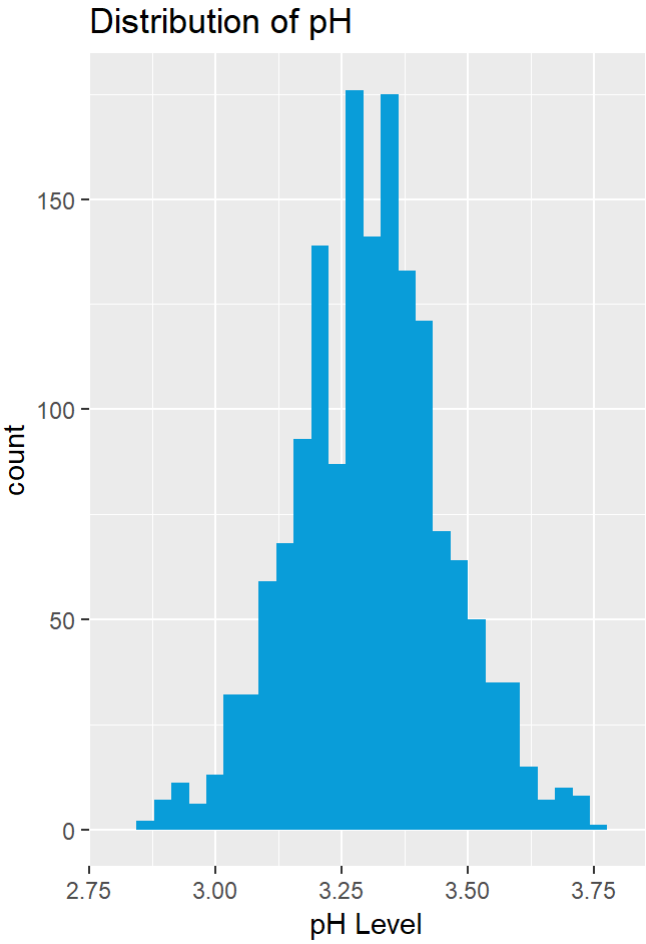




Citric acid showed the most unique distribution among all univariate variables. Omitting outlier, majority of data rows followed more like a bell-shape. This type of unique variable needs special attention.

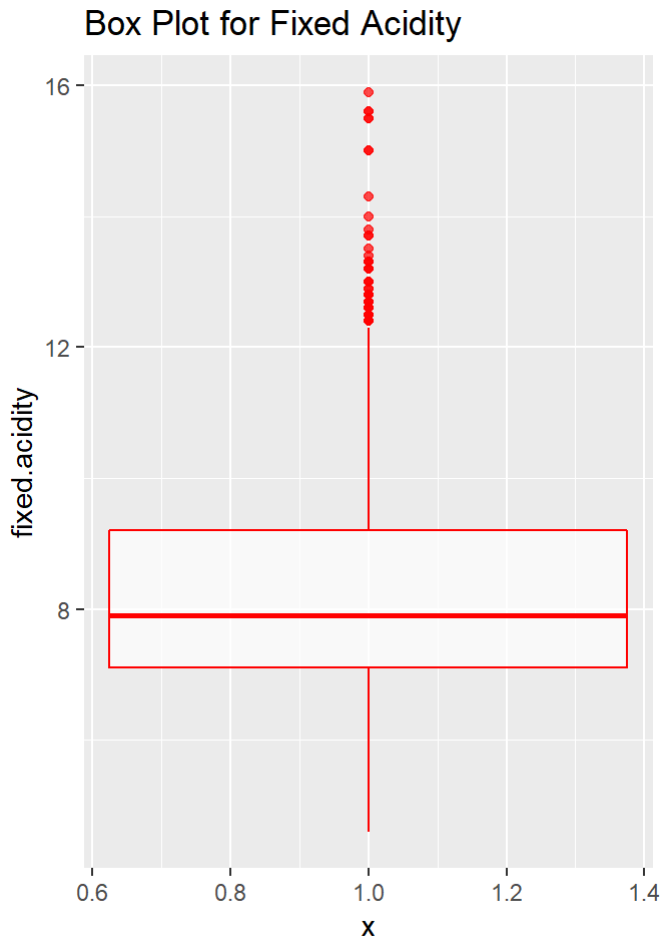
Part3 Third variable is pH level. Many people know that pH level represents the flavor of wine. I would like to see a distribution for this variable, and see if pH level has a impact on quality of wine in further research.

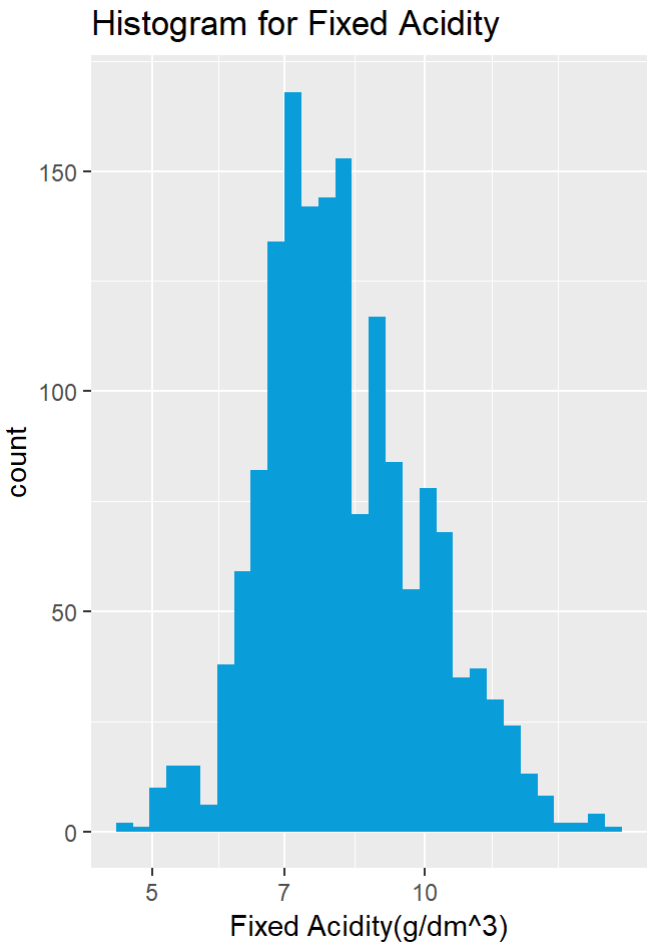




Part4 Other Original Variables In this section, I would like to investigate all the distribution information of each variables, so that I might get further insight in this research.

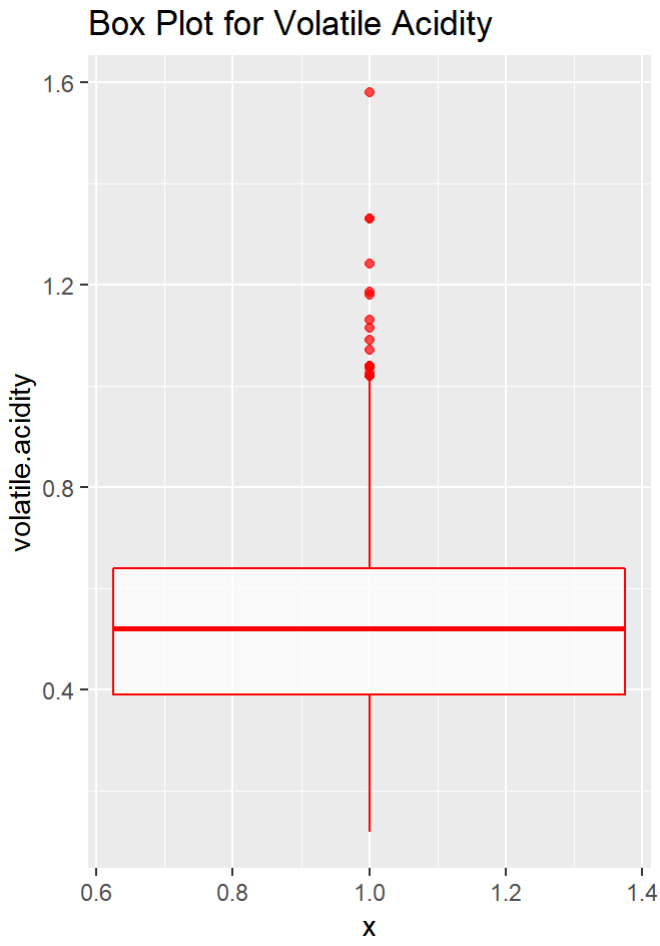
1. Fixed Acidity Distribution

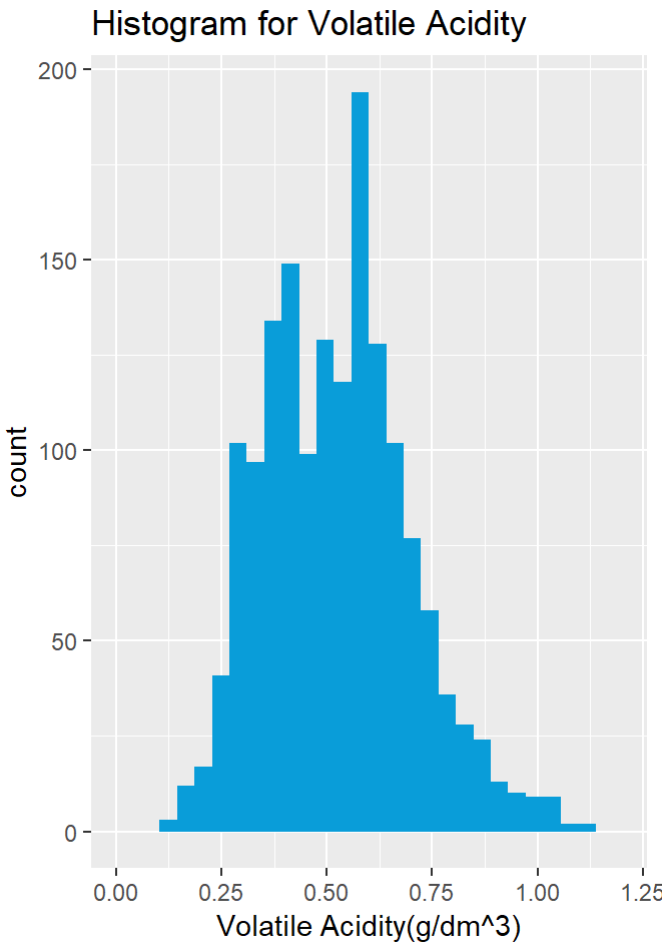




It is possible to see that there are a few outliers above 15. This part should be taken into account in later discussion.

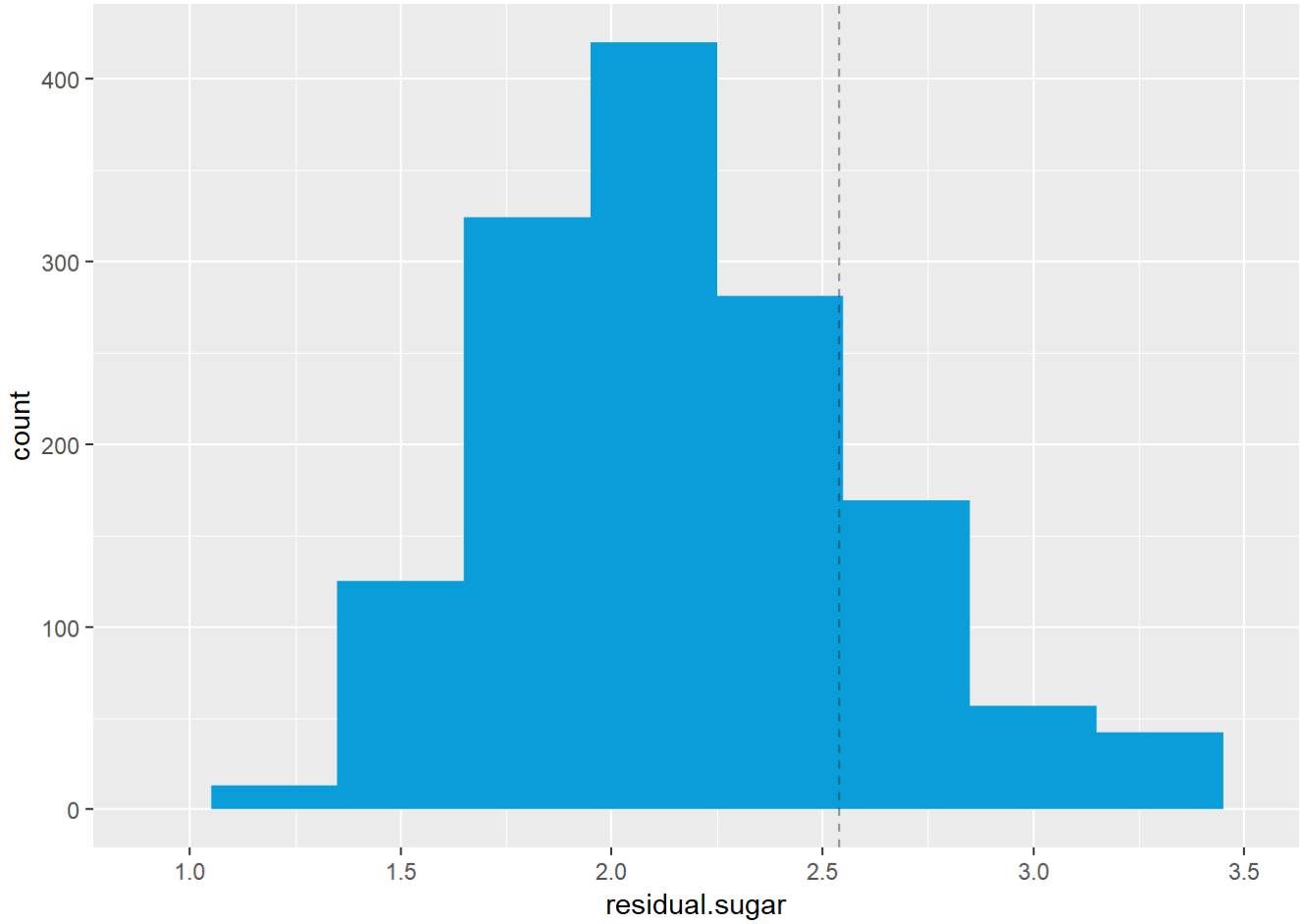
2. Distribution of Volatile acidity





The mean was about 0.5 and there were outliers above 1.3. Overall distribution was similar to normal distribution.

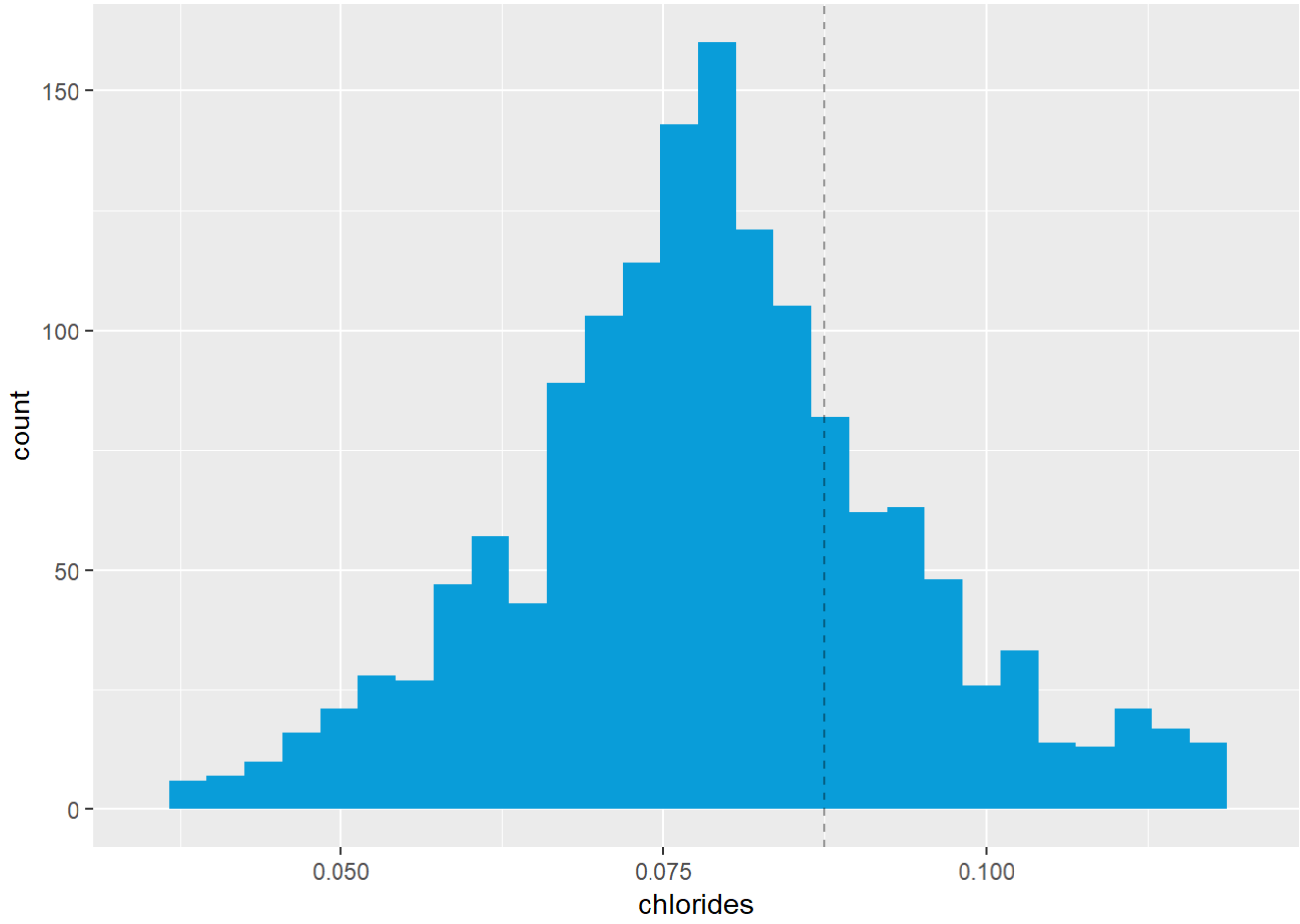
3. Residual Sugar




```
## $x
## [1] "Residual Sugar(g/dm^3)"
##
## $title
## [1] "Histogram of Sugar"
##
## attr(,"class")
## [1] "labels"
```

In residual sugar, it is possible to know that it followed normal distribution, yet there are a few outliner above level of 7.

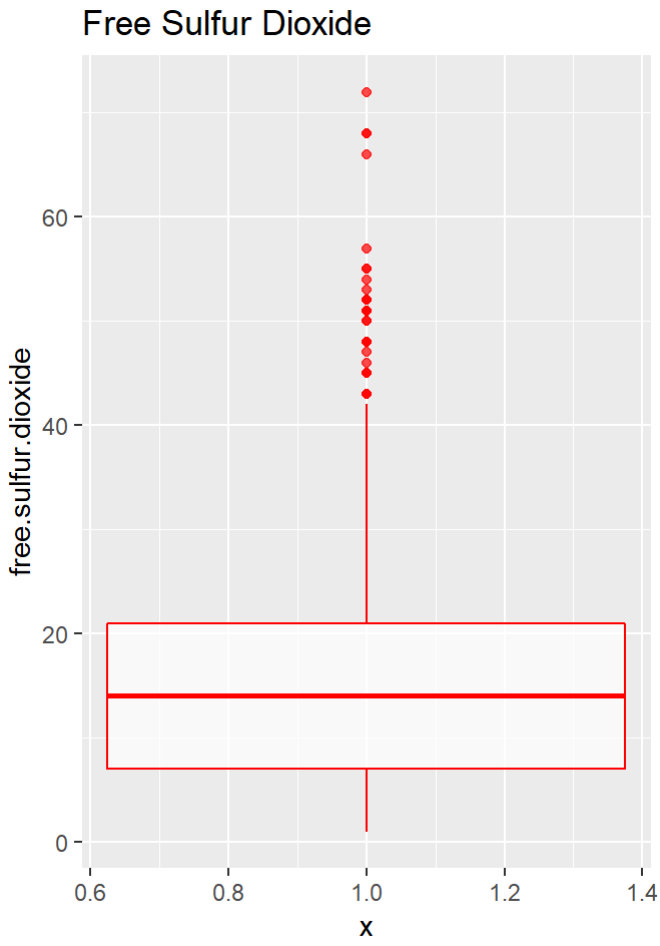
4. Chlorides(Salt)

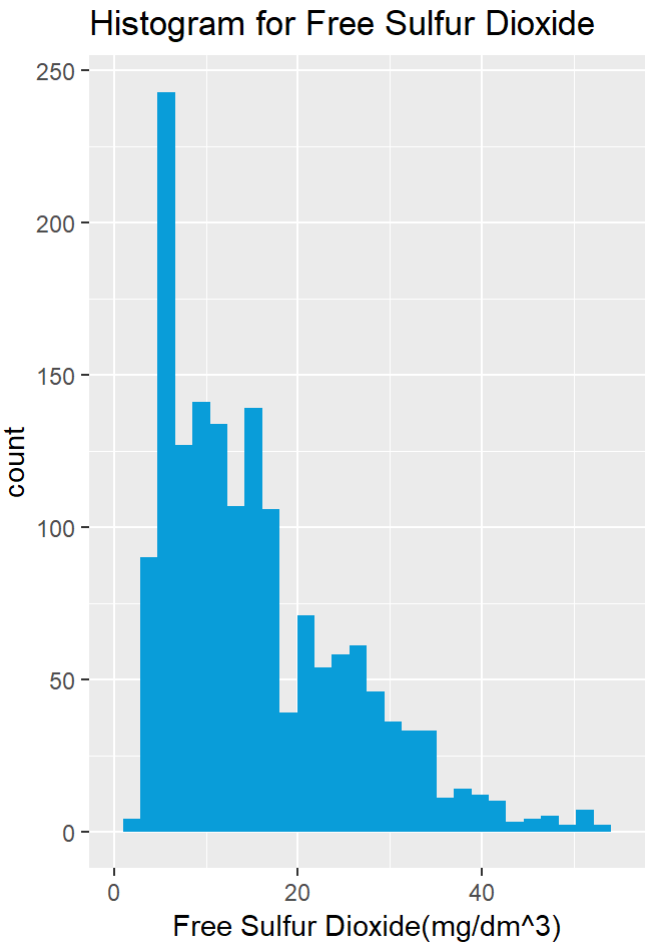



```
## $x
## [1] "Chlorides(g/dm^3)"
##
## $title
## [1] "Histogram of Salt"
##
## attr(,"class")
## [1] "labels"
```

In chlorides(simply meaning salt), it is possible to know that it followed normal distribution, yet there are a few outliner above level of 0.3.

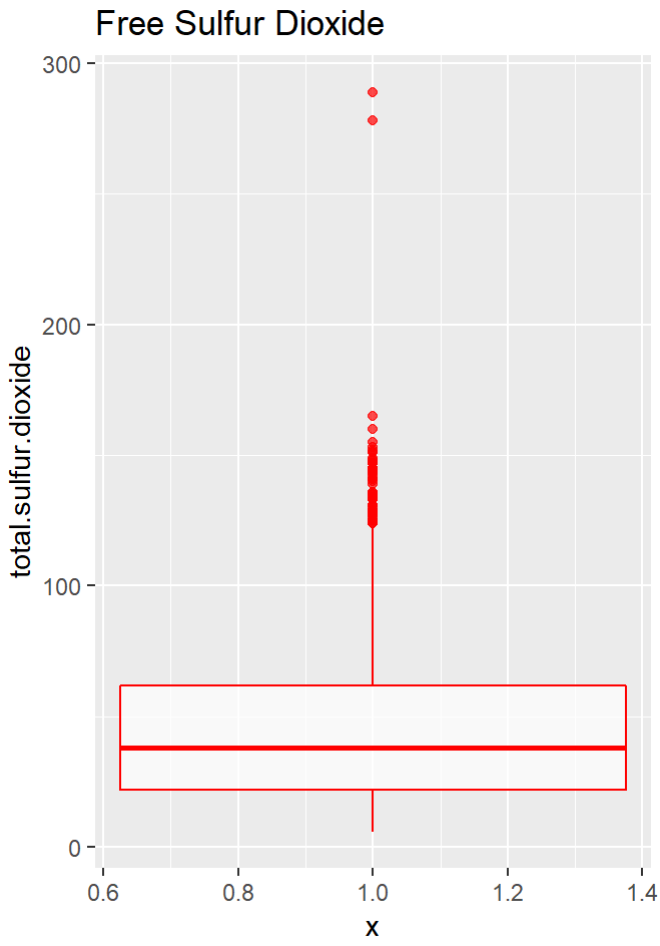
5. Free Sulfur Dioxide

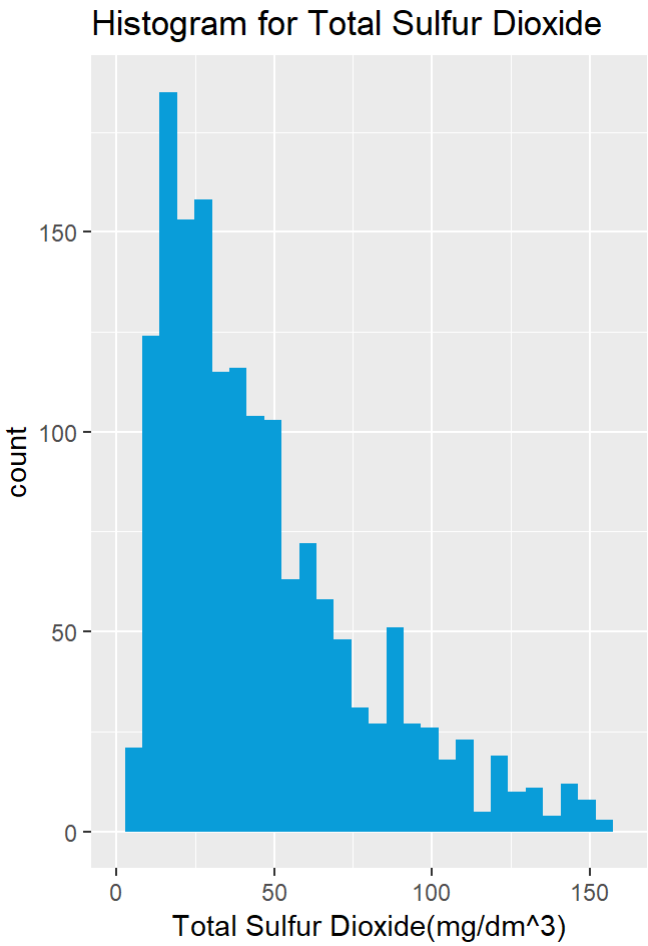




In Free Sulfur Dioxide distribution, the form was similar to right skewed distribution and had outliers above level of 60.

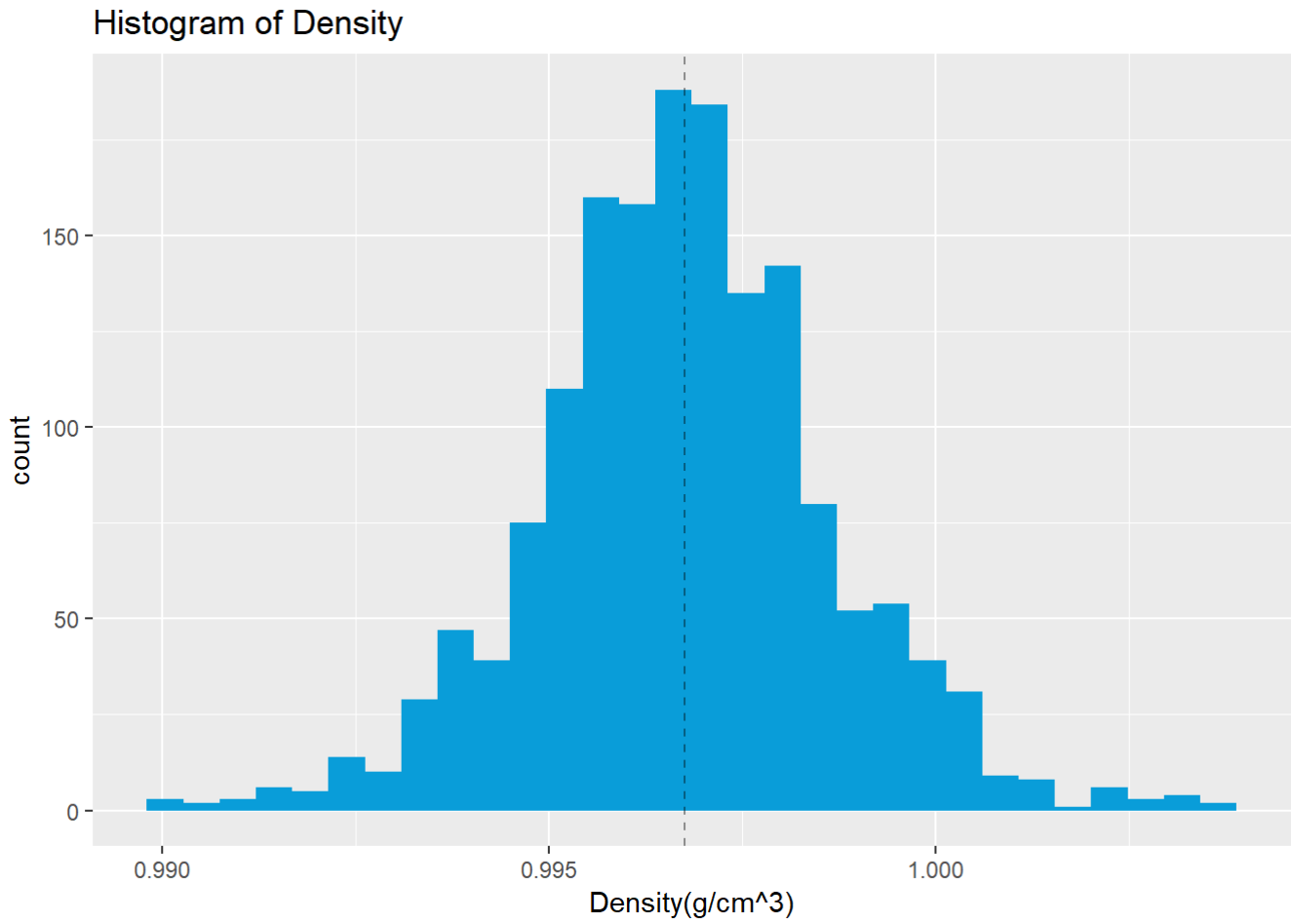
(6)Total Sulfur Dioxide





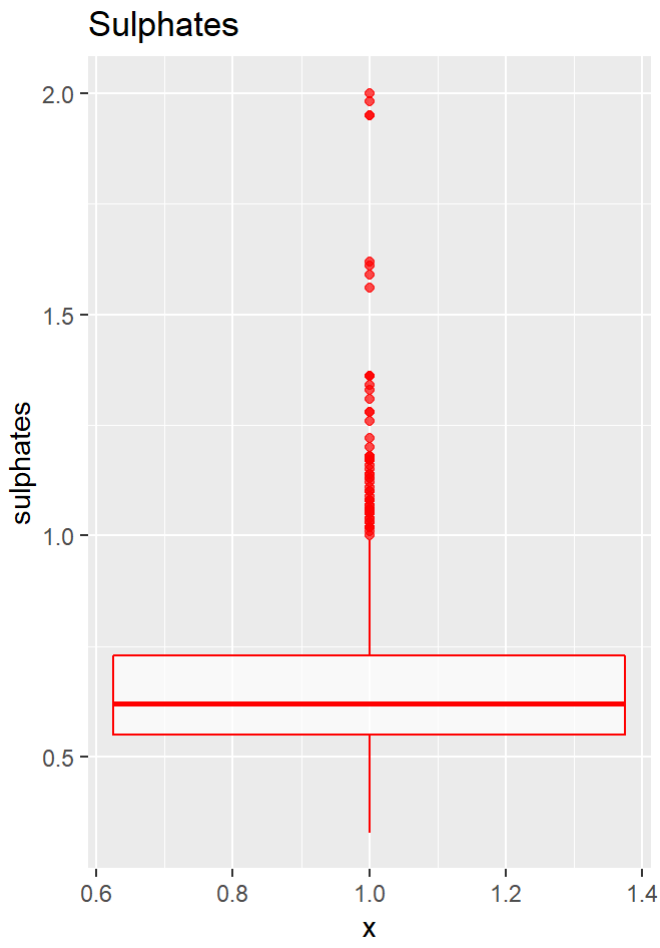
In Total Sulfur Dioxide distribution, the form was similar to right skewed distribution and had outliers above level of 150.

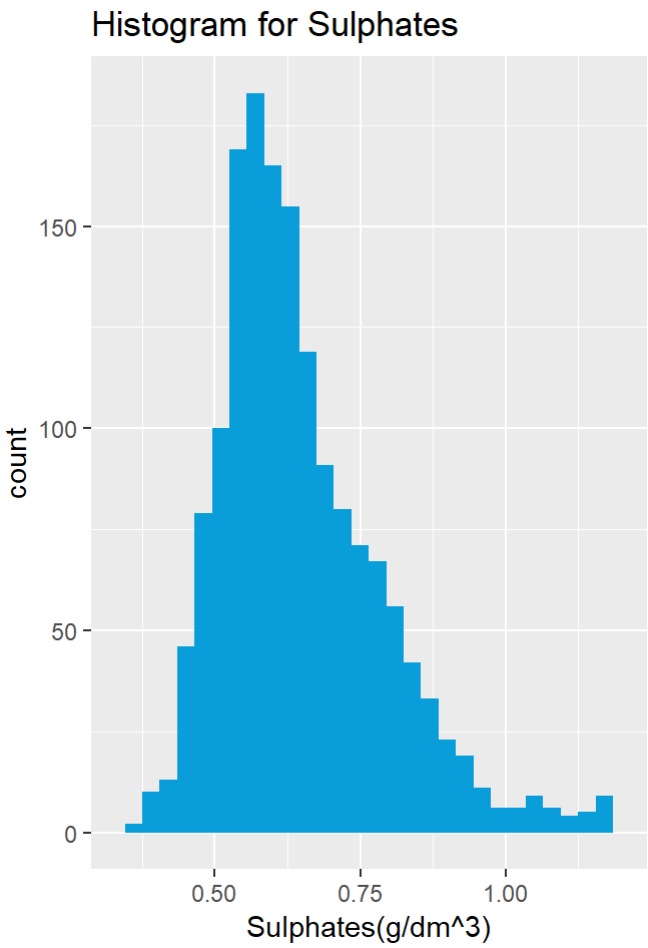
7. Density



Ditribution of density showed us almost perfet fit for normal distirbution with mean of 0.996.

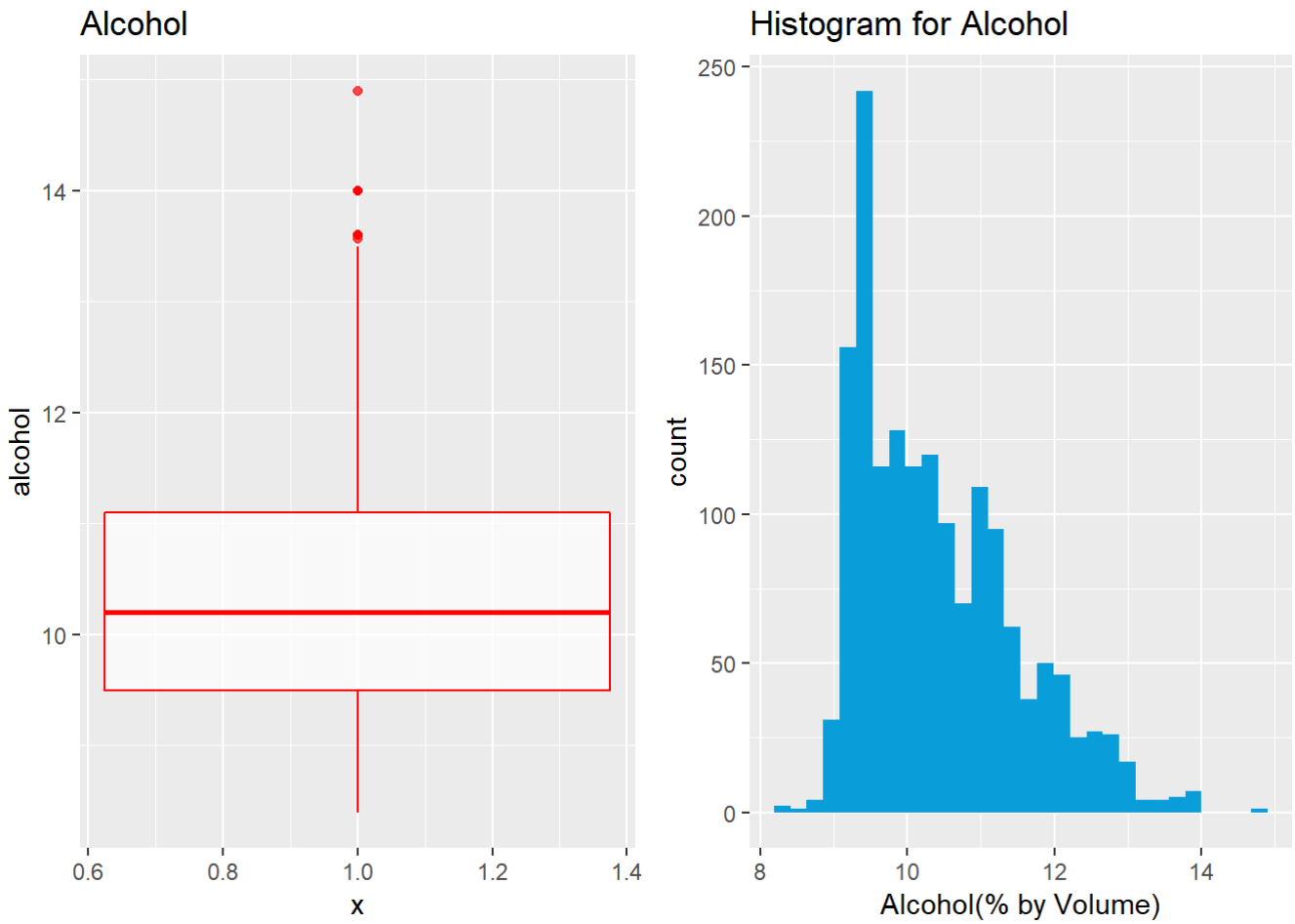
8. Sulphates





In this section, it is better to take causion on units(legends), because other sulfur related variables were measured with mg yet only sulphates were measured with g. This will be dealt with in later section. The means was about 0.55 with normal distribution. There are outliers above level of 1.5.

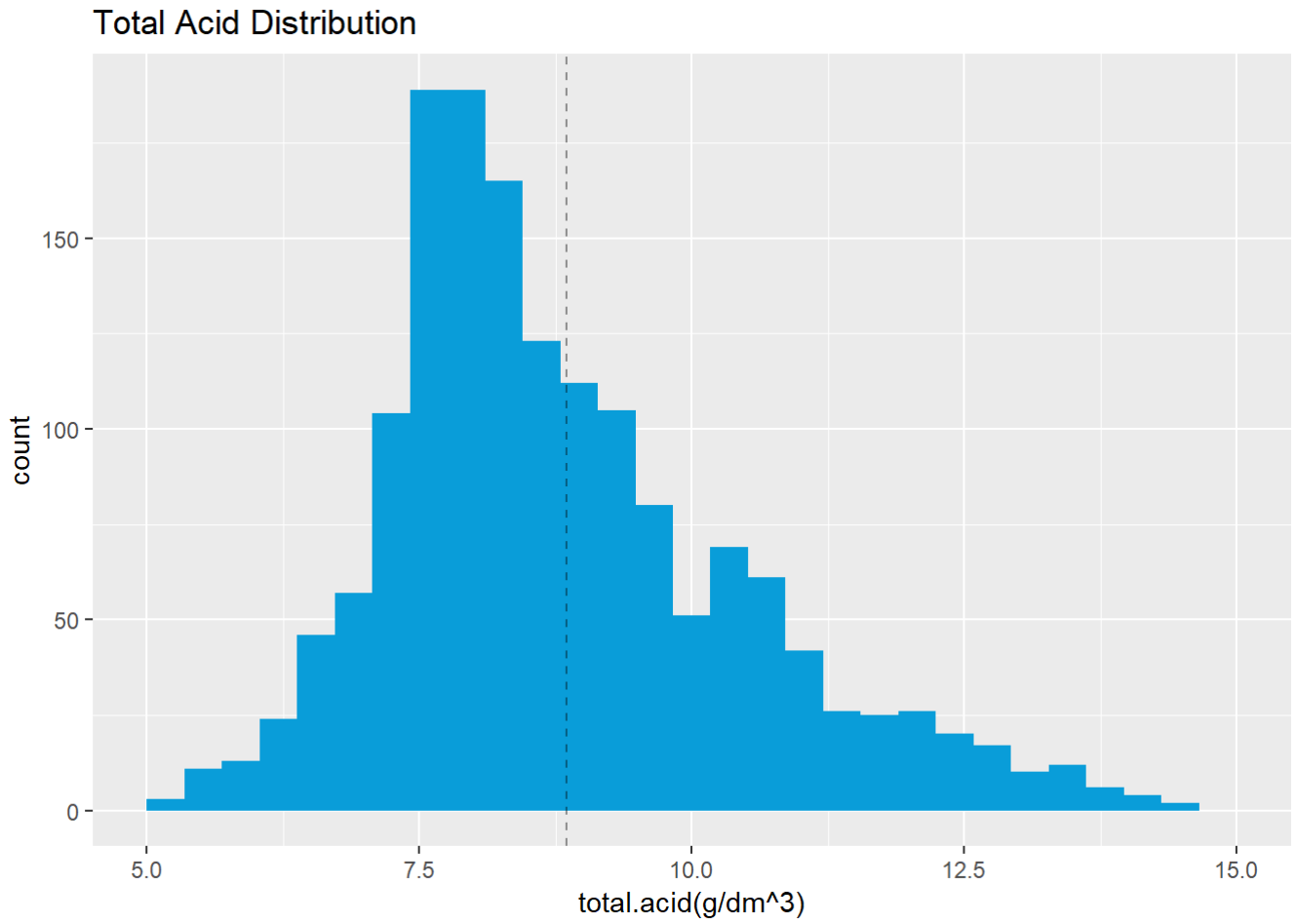
9. Alcohol



The distribution showed form of right skewed with almost no outliers.

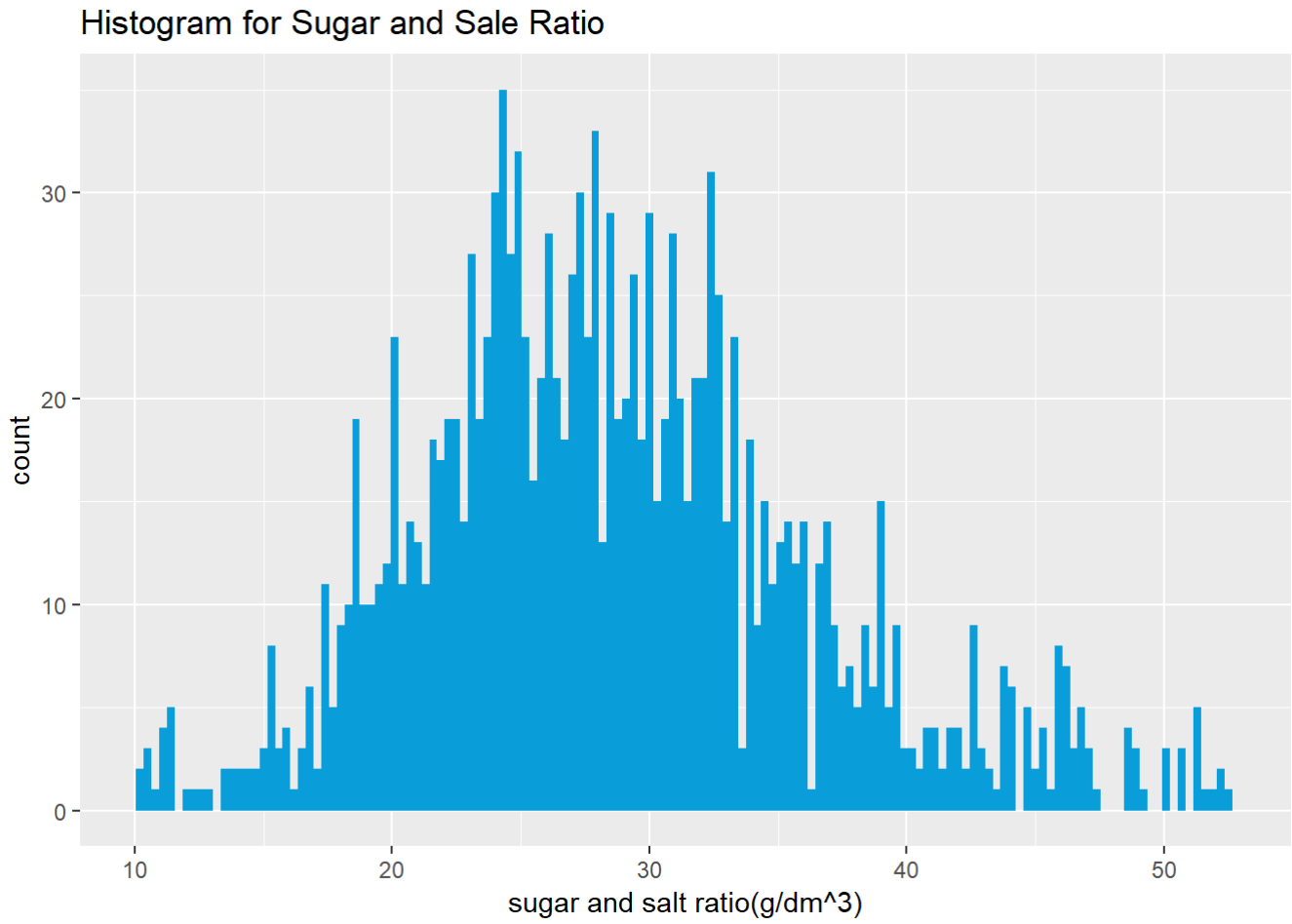
Part5 Manipulated Variables For other variables, there should be a few manipulations such as fixed.acidity and volatile.acidity. These two variables could be summed up and represented as total acitidy, and in turn this total index could affect pH level. For another example, chlorides, which meas amount of salt, would have to be considered with residual sugar. In common sense, it is the combination of sugar and salt that decides overall flavor of certain drinks or foods, not just total amount of sugar and salt. So these types of manipulations is going to be done in following sections.

Part5-1 This code is for making total acidity variable by adding fixed and volatile acidity. These two variables could be added directly since those two share same unit(dm^3). It seems like total acidity shows somewhat normal distribution.



Part5-2 This is a code for representing the ratio of sugar and salt. In personal opinion, the objective amount of salt or sugar does not soley make people feel certain food delicious. It is the proper combination of salt and sugar that produces best flavor.

residual sugar went to numerator and salt(chlorides) went to denominator. This result means that how much sugar is in the wine per one unit of salt. These two varibles could be multiplied or divided because they share same unit(dm^3)



Summary of Univariate Analysis

1. What is the structure of your dataset? It might seem that the data has only one dependent variable(DV), and simply other variables directly affect the DV. However, other variables should be combined into new variable such as fixed aggregating fixed acidity and volatile acidity into total acidity. This total acidity might be decisive variable in terms of pH level.
2. What is/are the main feature(s) of interest in your dataset? variables are closely related to each other rather than independent. Correlated variables are not that much useful when it comes to establishing a model. So several manipulations are neede in order to make new independent variables.
3. What other features in the dataset do you think will help support your investigation into your feature(s) of interest? Another feature is that most of variables show normal distributions . This feature is very important because distribution type could affect which models to use such as regression, MANOVA and etc.
4. Did you create any new variables from existing variables in the dataset? Yes, total acidity and sugar.salt.ratio. Specific reasons are articulated in the above section.

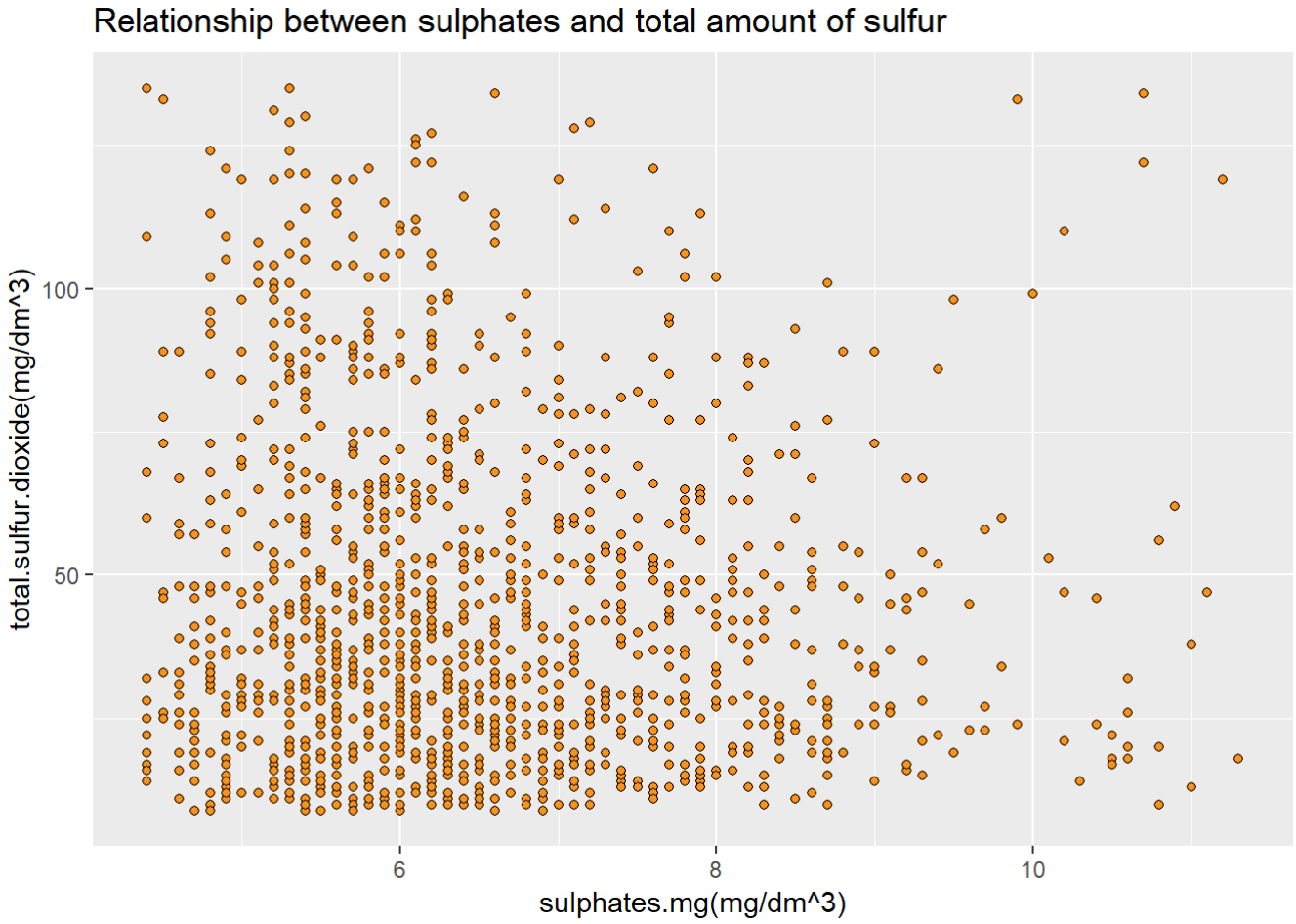
Bivariate Plots Section

Based on what you saw in the univariate plots, what relationships between variables might be interesting to look at in this section? Don’t limit yourself to relationships between a main output feature and one of the supporting variables. Try to look at relationships between supporting variables as well. This section is composed of four parts, each section containing plots, insights, statisical models and interpretations.

Part1 - Sulphates and total amount of sulfur dioxide First bivariate analysis that I have conducted is between sulphates and total.sulfur.dioxide. According to the information document, suliphates could increase level of SO2, which acts as antimicrobial and antioxidant. This concept directly affects total sulfur dioxide, which is about amount of free and bound forms of S02. SO2 could be a significant to quality of wine within certain level, but it might also be possible that too much SO2 could negatively affect quality. So it is essential to know what affects SO2 level, and the two variables are the ones that have high possibility.

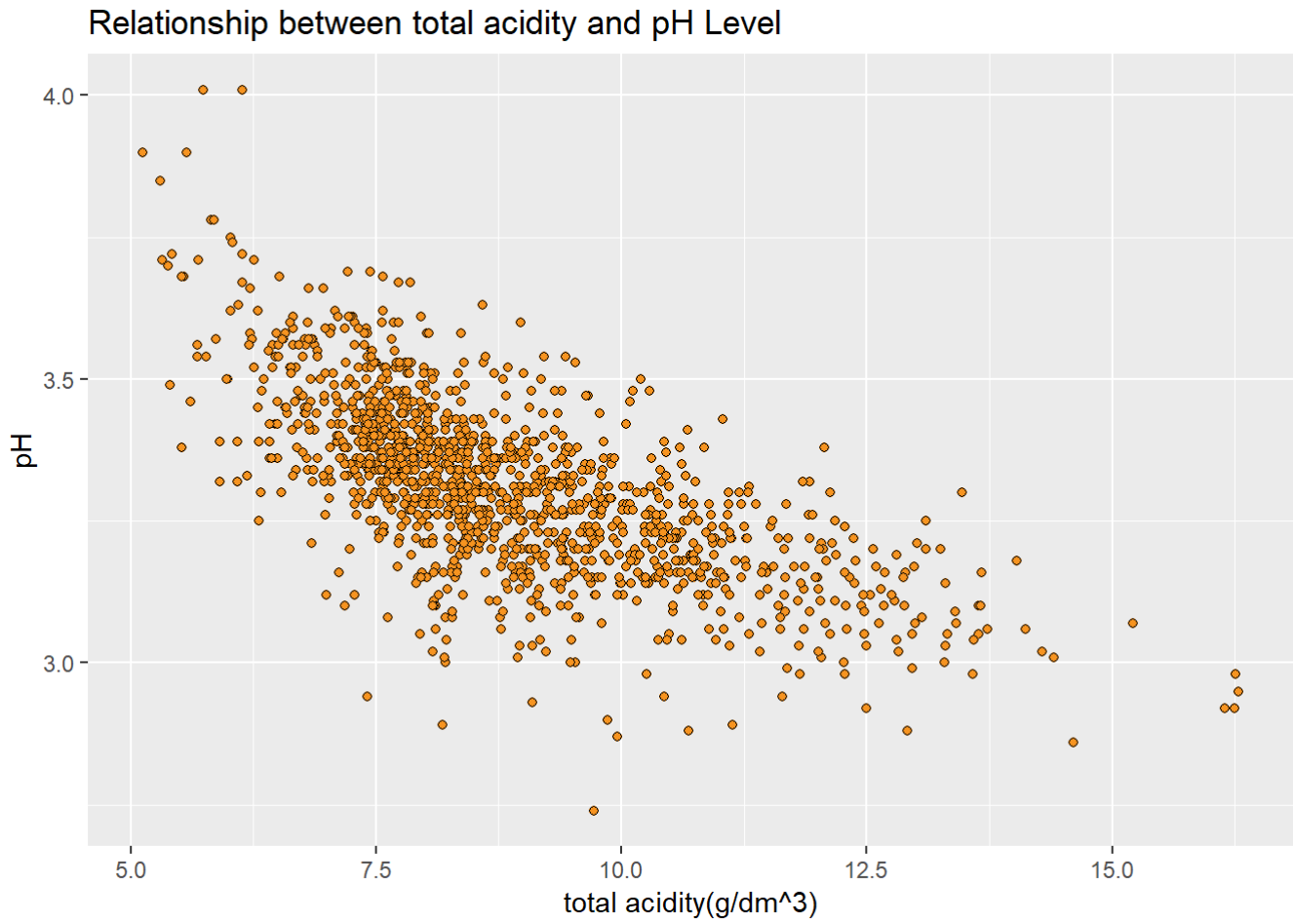
This is code equalizing the unit of sulphates and total sulfur dioxide

```
df$sulphates.mg = df$sulphates*10
```



Part2 - total acidity and pH level

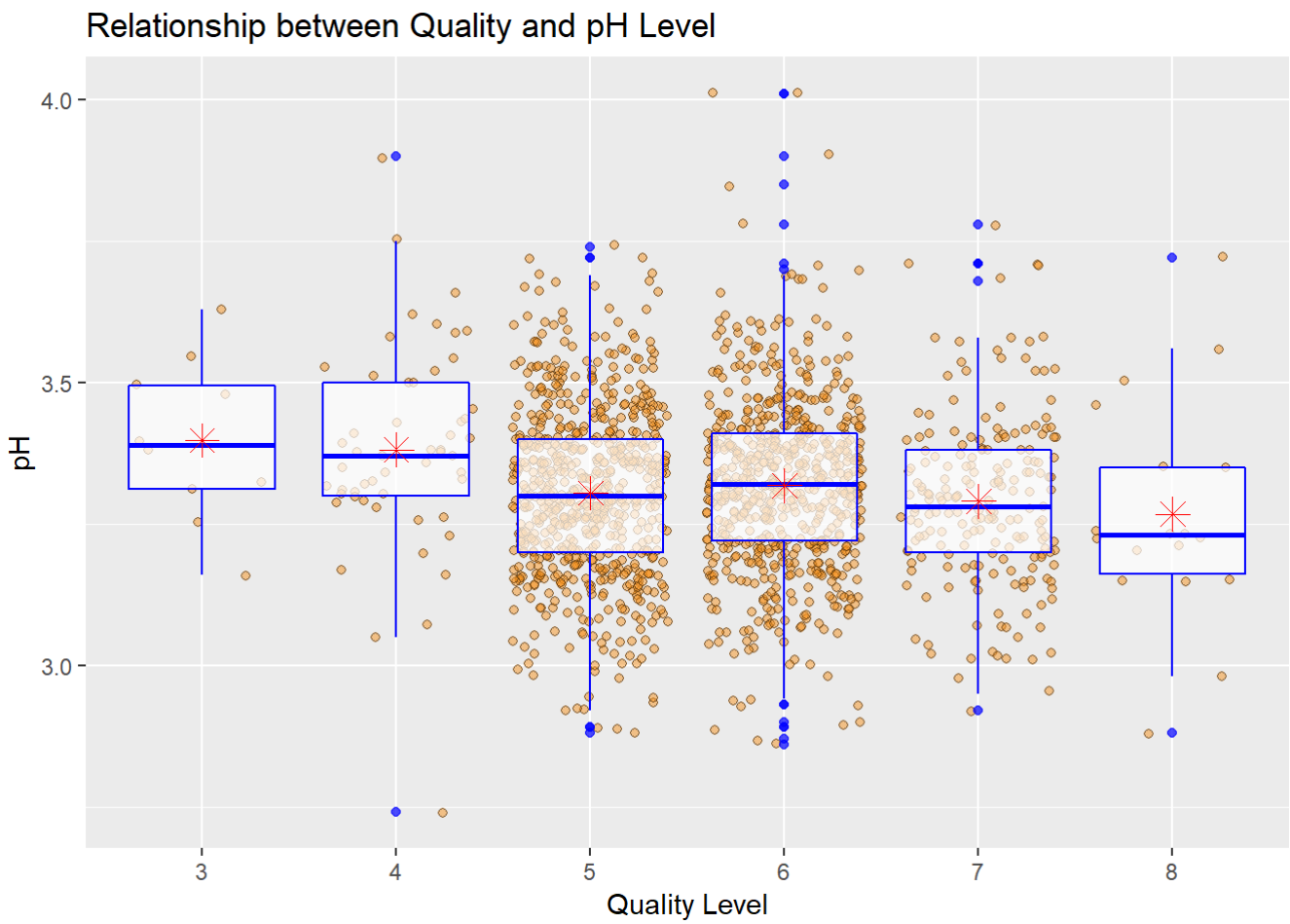
I wanted verity whether total acidity affects pH level. Based on the below graph, it is apparent that total acidity affects pH level. For next step, let's run linear regression to make sure this relationship is true or not.



Since significant the p-value level is insured, it is natural to conclude that total acidity had negative impact with -.06 coefficient.

```
##
## Call:
## lm(formula = pH ~ total.acidity, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51790 -0.06765  0.00121  0.06752  0.53377
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.850688   0.015106  254.91   <2e-16 ***
## total.acidity -0.060986   0.001677  -36.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1142 on 1597 degrees of freedom
## Multiple R-squared:  0.4531, Adjusted R-squared:  0.4528
## F-statistic:  1323 on 1 and 1597 DF,  p-value: < 2.2e-16
```

Part3 - pH and quality relationship Many people know that pH level affects a taste of wine. In this research, I wanted to make sure whether this was actualy true or not. So in an intuitive sense, I draw a plot that might explain a relationship between the two variables. In later analysis, I want to run regression model to find out there is significant difference.

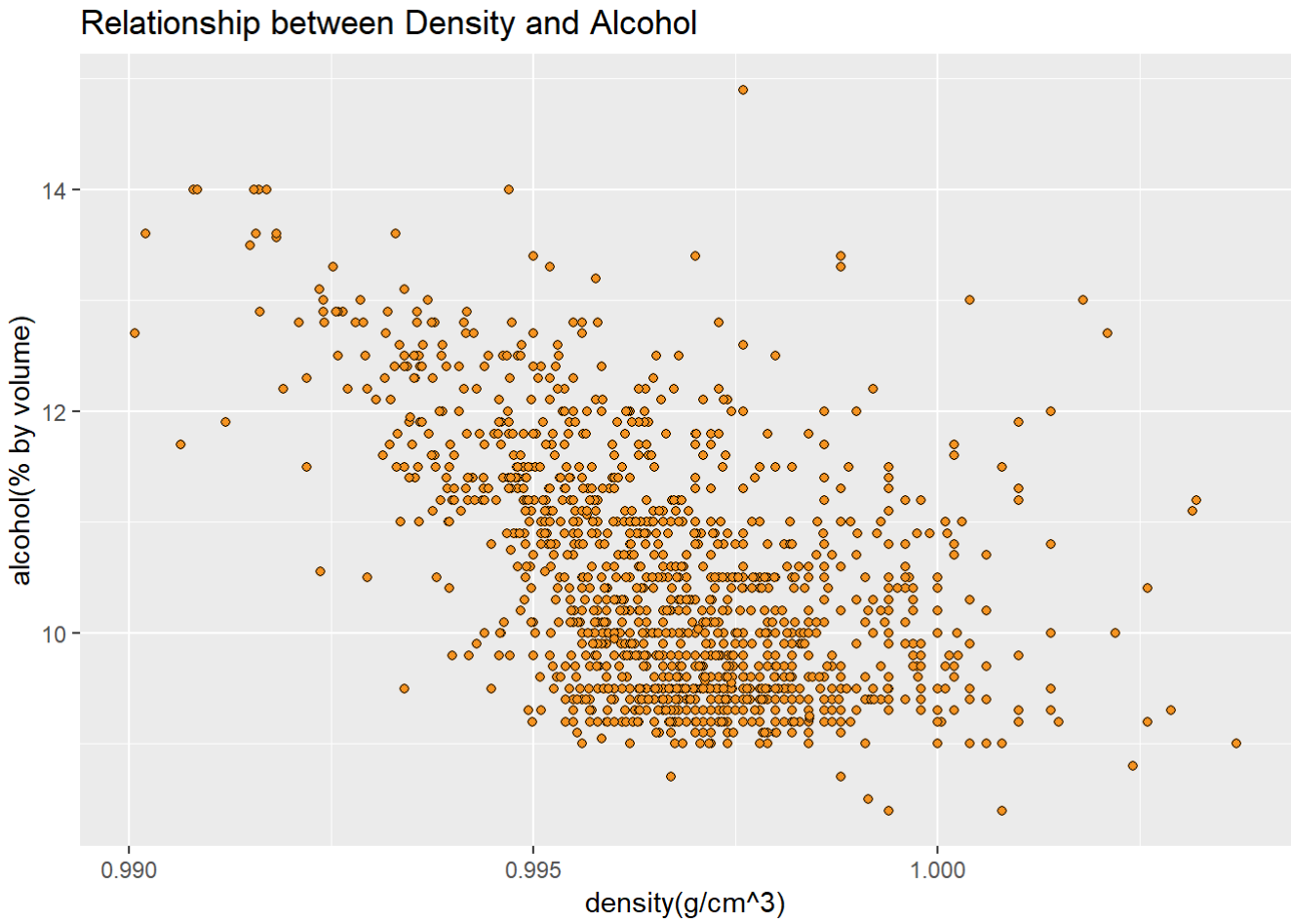


Out of curiosity, I ran linear regression analysis with quality and pH variable. The result revealed us that there are significant differences between groups, show meaningful p-value, 0.021.

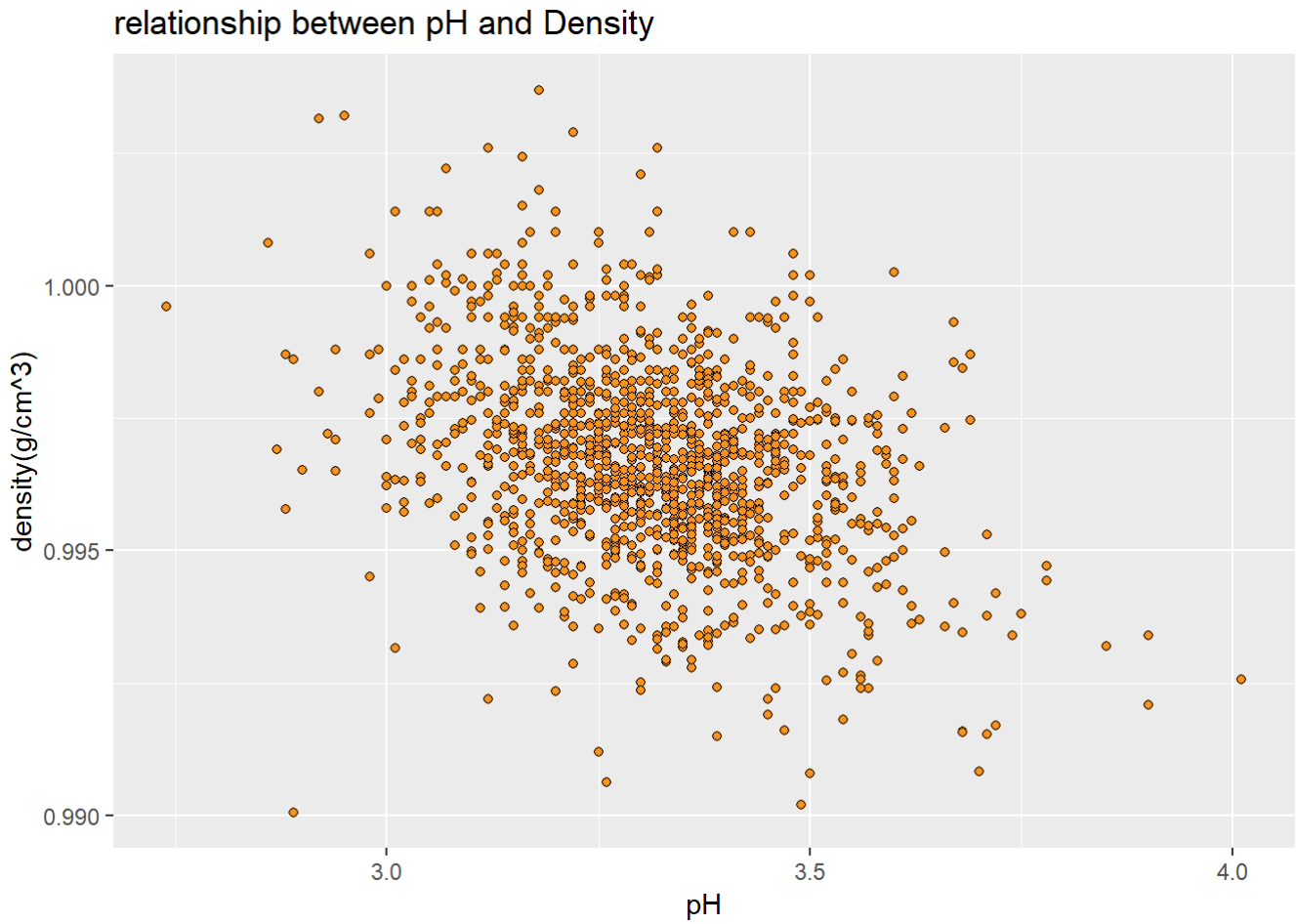
```
fit <- lm(quality ~ pH, data = df)
summary(fit)
```

```
##
## Call:
## lm(formula = quality ~ pH, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6817 -0.6394  0.3032  0.3878  2.4874
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.6359     0.4332  15.320   <2e-16 ***
## pH           -0.3020     0.1307  -2.311    0.021 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8065 on 1597 degrees of freedom
## Multiple R-squared:  0.003333,   Adjusted R-squared:  0.002709
## F-statistic:  5.34 on 1 and 1597 DF,  p-value: 0.02096
```

Part4 - Density and Alcohol By reading information document, density is related to pH level and alcohol. So I was curious if there is an certain distribution among these variables. It seems like there is negative correlation between density and alcohol, and between desnity and pH level.




```
##
##  Pearson's product-moment correlation
##
## data:  df$density and df$alcohol
## t = -22.838, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5322547 -0.4583061
## sample estimates:
##       cor
## -0.4961798
```



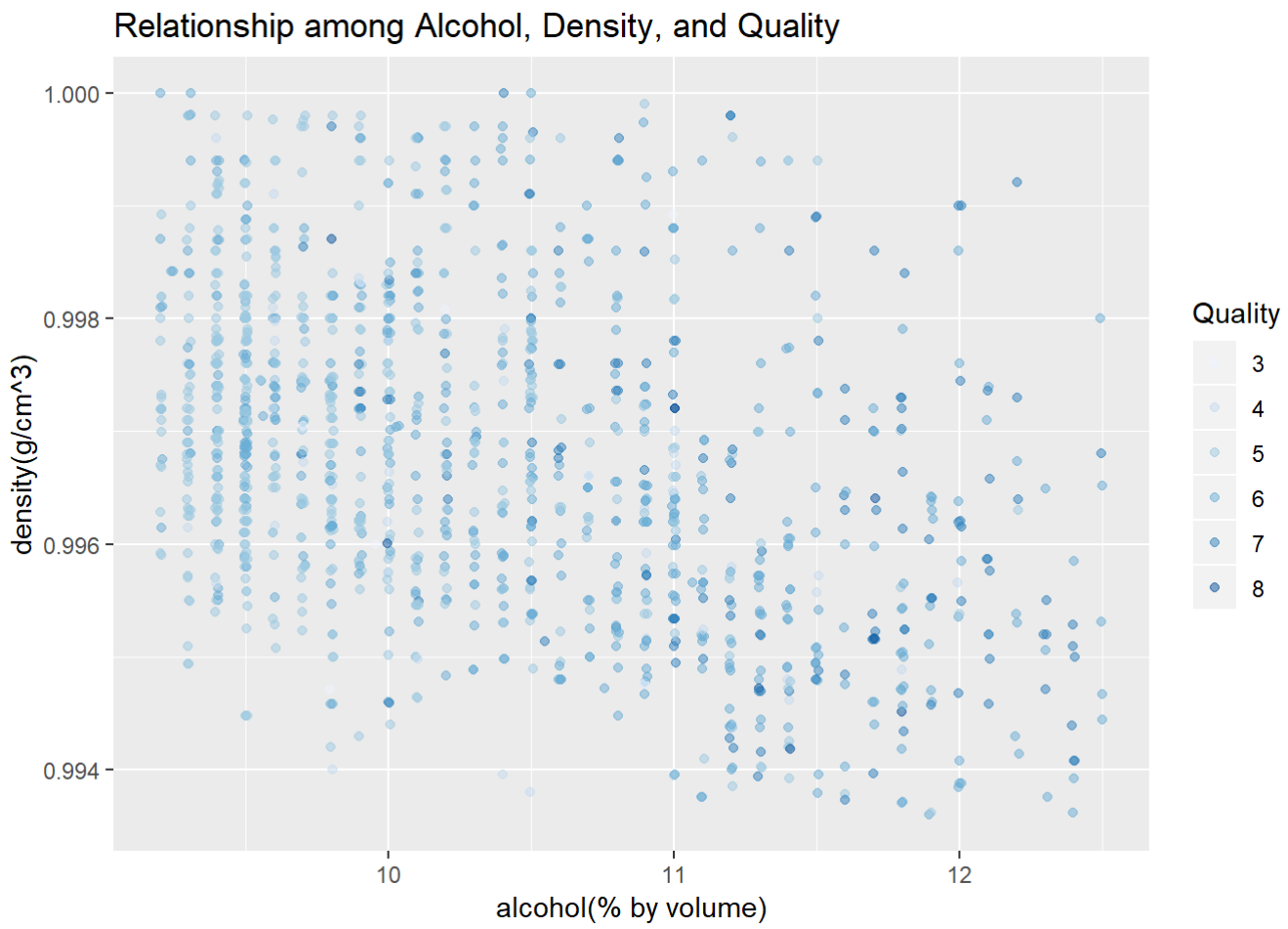

```
##
##  Pearson's product-moment correlation
##
## data:  df$pH and df$density
## t = -14.53, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3842835 -0.2976642
## sample estimates:
##       cor
## -0.3416993
```

All of the correlation tests were significant, which means there are certain relationships or might be even causal relationship that we could find more in this dataset. This point of analysis would be conducted in later multivariate section.

Multivariate Plots & Analysis Section

Multivariate plots and analysis section consists of four parts. Each part provide plots among variables, insights, statistical analysis and interpretations, so that I could give find critical variables that might determine wine quality.

Part1 - Alcohol, Density and Quality First analysis for multivariate was about density, alcohol and quality. In the above bivariate section, I have found negative correlationship. I wanted to go further regarding quality, because quality is our main focus.



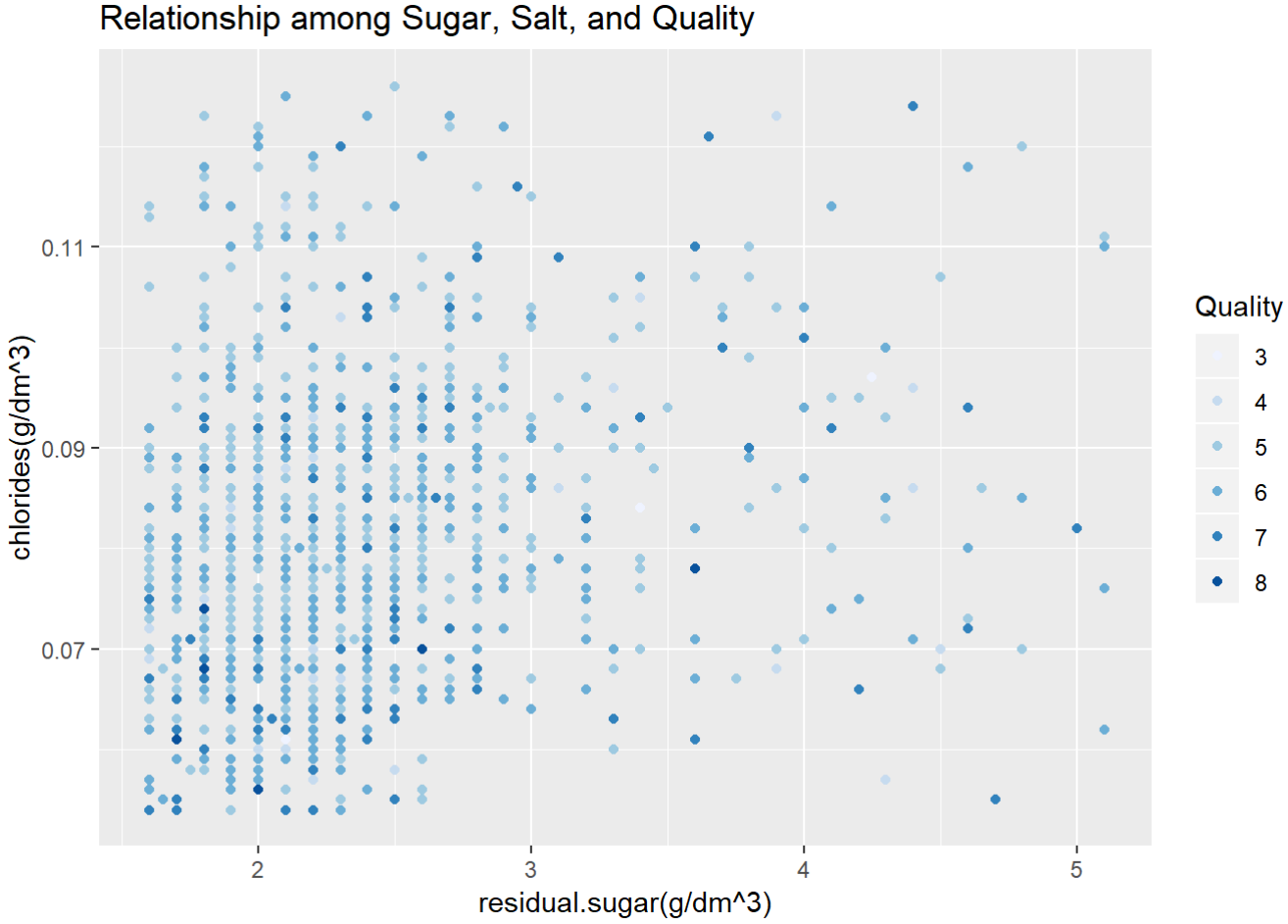
The result show us that there might be some causal relationship amoung these variables. If I could construct a model, then alcohol and density would be indepedent variable and quality would be dependent variable.

```
##
## Call:
## lm(formula = quality ~ alcohol + density, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9376 -0.3914 -0.1429  0.5263  2.5556
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -33.15238   10.87817  -3.048  0.00234 **
## alcohol       0.39144    0.01915  20.441  < 2e-16 ***
## density      34.82170   10.81292   3.220  0.00131 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7083 on 1596 degrees of freedom
## Multiple R-squared:  0.2317, Adjusted R-squared:  0.2308
## F-statistic: 240.7 on 2 and 1596 DF,  p-value: < 2.2e-16
```

With significant p-value on both IVs, we could conclude that density and alcohol have positive influence on quality. Especially, estimate for density is 34, which means that its impact on quality is very effective.

Part2 - Sugar, Salt and Quality For next multi-variate analysis, three variables were considered; sugar, salt and quality. First of all, I have plotted multivariate analysis. I have involved within the range between 5% and 95% on both sugar and salt axises. It is because too much or too less sugar and sale would not be proper product

for sale. The graph result showed that there might be positive relationship among three variables.



Based on the plot about sugar, salt and quality, it is time to verify whether they have real impact on quality. As all of the variables are metric(=numeric) scale, it is right to run analysis with multiple linear regression model.

```
fit <- lm(quality ~ residual.sugar + chlorides, data = df)
summary(fit)
```

```
##
## Call:
## lm(formula = quality ~ residual.sugar + chlorides, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6898 -0.6462  0.3047  0.3618  2.3617
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.80075    0.05431 106.813  < 2e-16 ***
## residual.sugar  0.01201    0.01424   0.843    0.399
## chlorides      -2.23185    0.42648  -5.233 1.89e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8012 on 1596 degrees of freedom
## Multiple R-squared:  0.01706,    Adjusted R-squared:  0.01582
## F-statistic: 13.85 on 2 and 1596 DF,  p-value: 1.092e-06
```

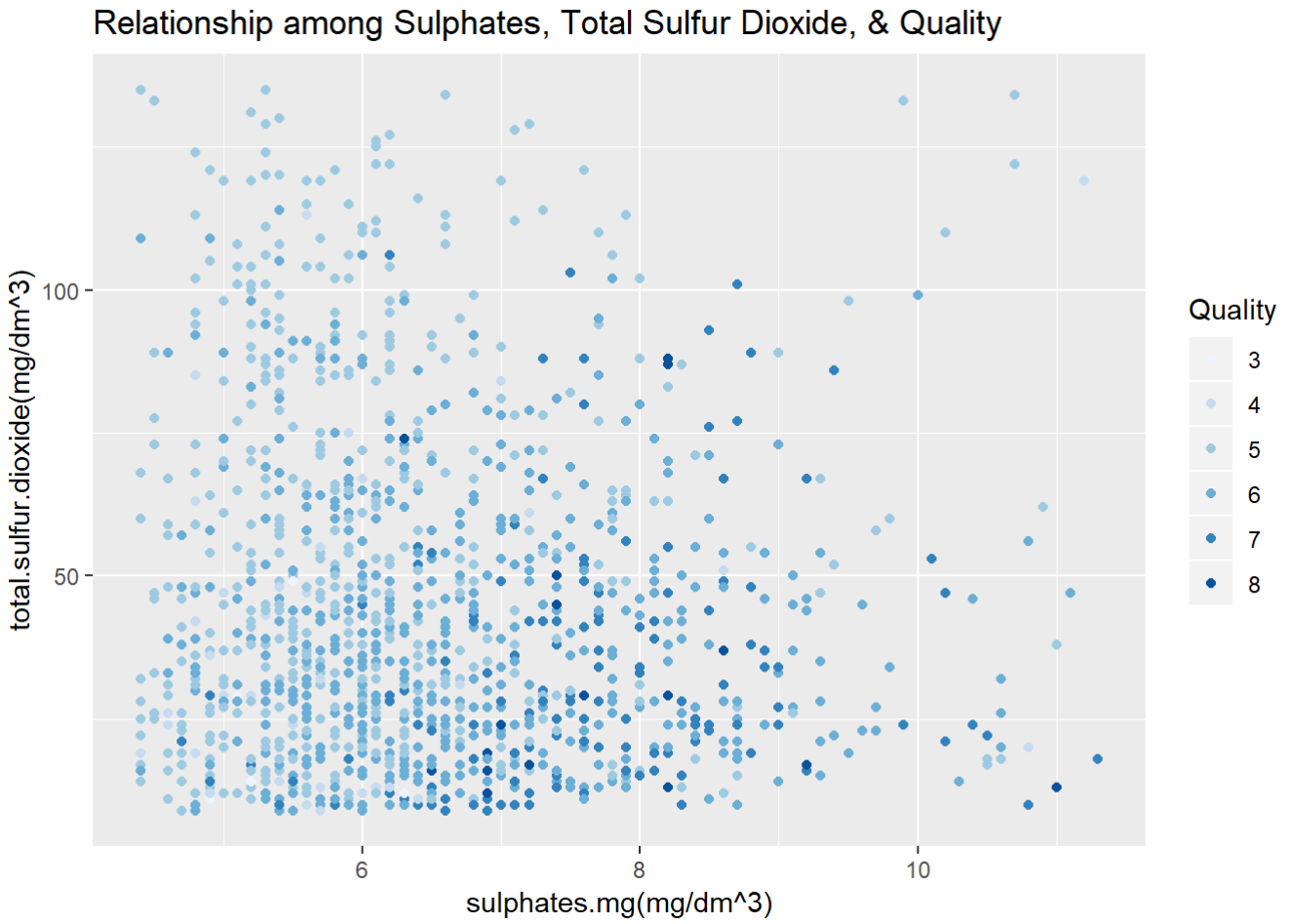
Based on the model summary, it seems that only salt(chlorides) has meaningful impact on quality with significant p-value and coefficient being -2.23. However, it is skeptical to completely skeptical to ignore residual sugar and focus on salt itself. In order to consider salt and sugar simultaneously, it is time for 'salt.sugar.ratio(new variable that I have made in the above section)' to take a role. For reminder, salt went into denominator and sugar went into numerator, which means that how much sugar were in the wine per salt. I have run linear regression so as to verify the casual relationship.

```
fit<- lm(quality ~ sugar.salt.ratio, data = df)
summary(fit)
```

```
##
## Call:
## lm(formula = quality ~ sugar.salt.ratio, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7933 -0.6161  0.3185  0.3957  2.4022
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.501729   0.038627 142.431  < 2e-16 ***
## sugar.salt.ratio 0.004195   0.001030   4.071 4.91e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8037 on 1597 degrees of freedom
## Multiple R-squared:  0.01027,    Adjusted R-squared:  0.009652
## F-statistic: 16.57 on 1 and 1597 DF,  p-value: 4.908e-05
```

The result clearly show us that sugar salt ratio has positive impact on quality of wine. P-value were completely significant with 0.004 coefficient. So from this part of multivariate analysis, I have found meaningful factor that might determine quality of wine.

Part3 - Sulphates and Total Amount of sulfur dioxide In the bivariate analysis, sulphates and total sulfur dioxide relationship were considered. It is time to go one step further with these variables by considering quality at the same time. In total, sulphates, total sulfur dioxide, and quality will be taken into account. First of all, lets plot in terms of these three variables



It seems that there might be certain causal relationship. So I have conducted multiple linear regression model.

```
fit <- lm(quality ~ sulphates.mg + total.sulfur.dioxide, data = df)
summary(fit)
```

```
##
## Call:
## lm(formula = quality ~ sulphates.mg + total.sulfur.dioxide, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1885 -0.5512  0.0395  0.4211  2.7159
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.045207   0.080447  62.714  < 2e-16 ***
## sulphates.mg          0.123787   0.011314  10.941  < 2e-16 ***
## total.sulfur.dioxide -0.004818   0.000583  -8.265 2.91e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7659 on 1596 degrees of freedom
## Multiple R-squared:  0.1016, Adjusted R-squared:  0.1005
## F-statistic: 90.29 on 2 and 1596 DF,  p-value: < 2.2e-16
```

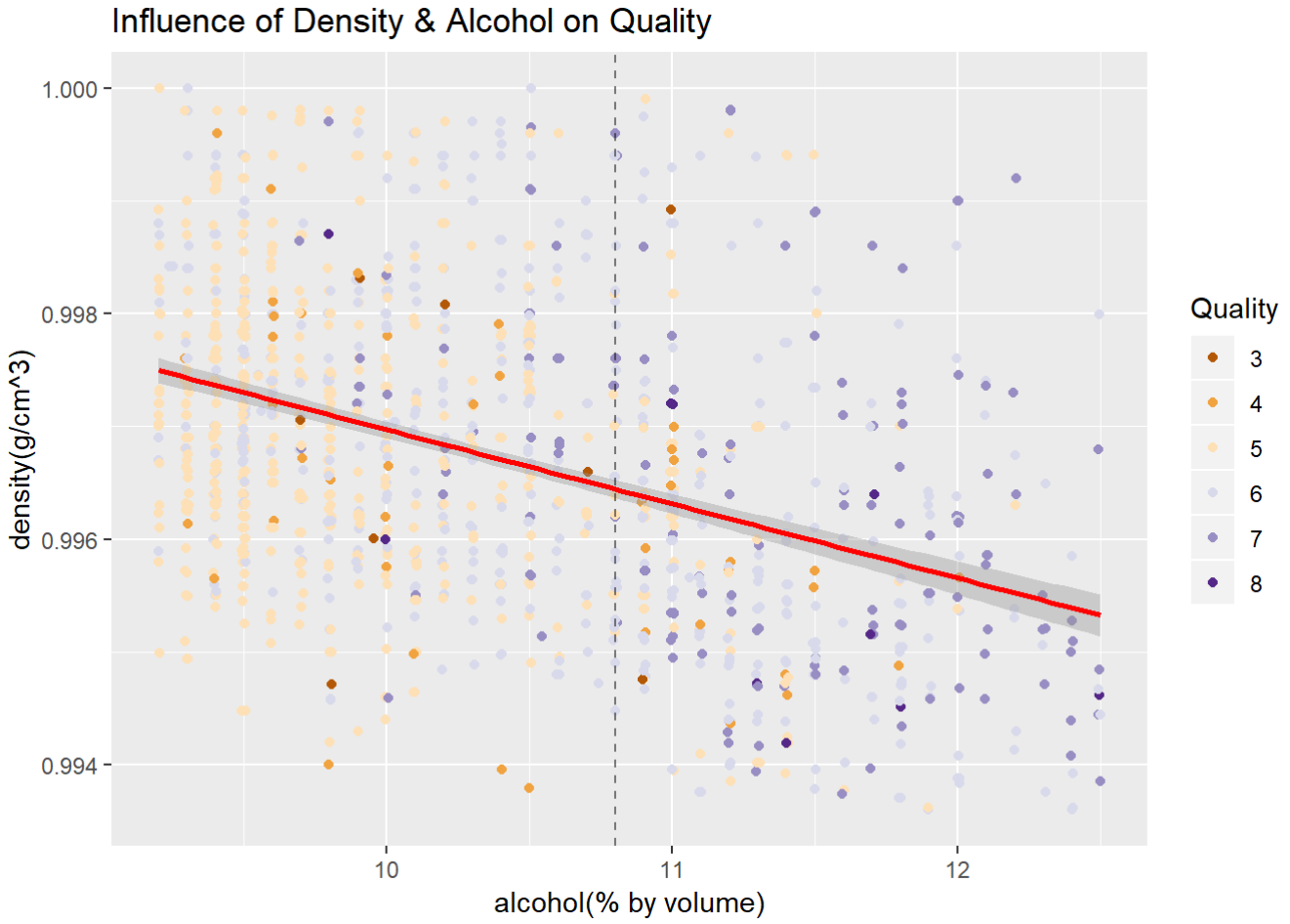
The result showed that both sulphates and total sulfur dioxide had meaningful impact on wine's quality. P-values for both variables were significant. Interesting point is that sulphates had postive coefficent, but totoal sulfur dioxide had negative with smaller size of coefficient than sulphates. What this means is that certain

amount of sulphates as additive could help wine to be more qualifed, yet too much of these additives might harm quality of wine.

Part4 - Summary of Multiavariate Analysis In this section, we have found various variables and relationships that are important to quality of wine. First of all, density and alcohol level were important with density being highest estimate(34) of all variables. Second sugar and salt should be considered simultaneously with ratio. Last one is that sulphates could help people to brew sound wine, yet too much of the ingredient can ruin the quality.

Final Plots and Summary

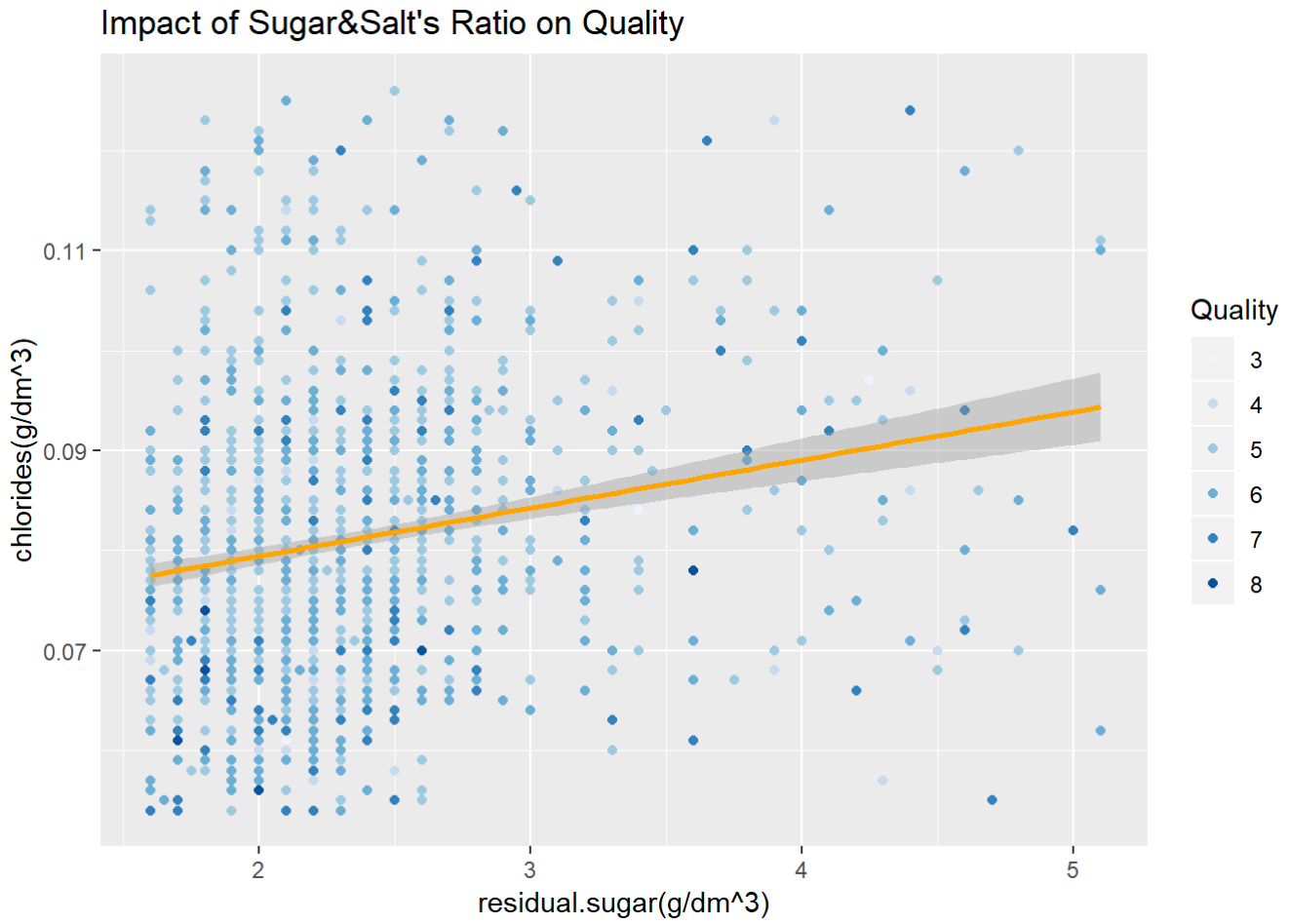
Plot One



Description One

First point for the summary, there was a certain correlation between alcohol and density. To go one step further, I have included quality in multivariate analysis to see what kind of relationship that I could find. There were two interesting points. Relatively low quality dots whose color is bright brown were concentrated in left part of dashed line in the graph, whereas purple color which stands for higher quality were densily populated in the right side of dashed line. This intuition gave me some idea to form an model. Since negative correlation between alcohol and density might directly influence the wine's quality. In order to prove this point, I have established a multiple regression model as follows. " fit <- lm(quality ~ alcohol + density, data = df) " " summary(fit) " The model result showed that both alcohol and density were significant facotr for determining quality of wine. Especially coefficient for density being the highest among all coefficients.

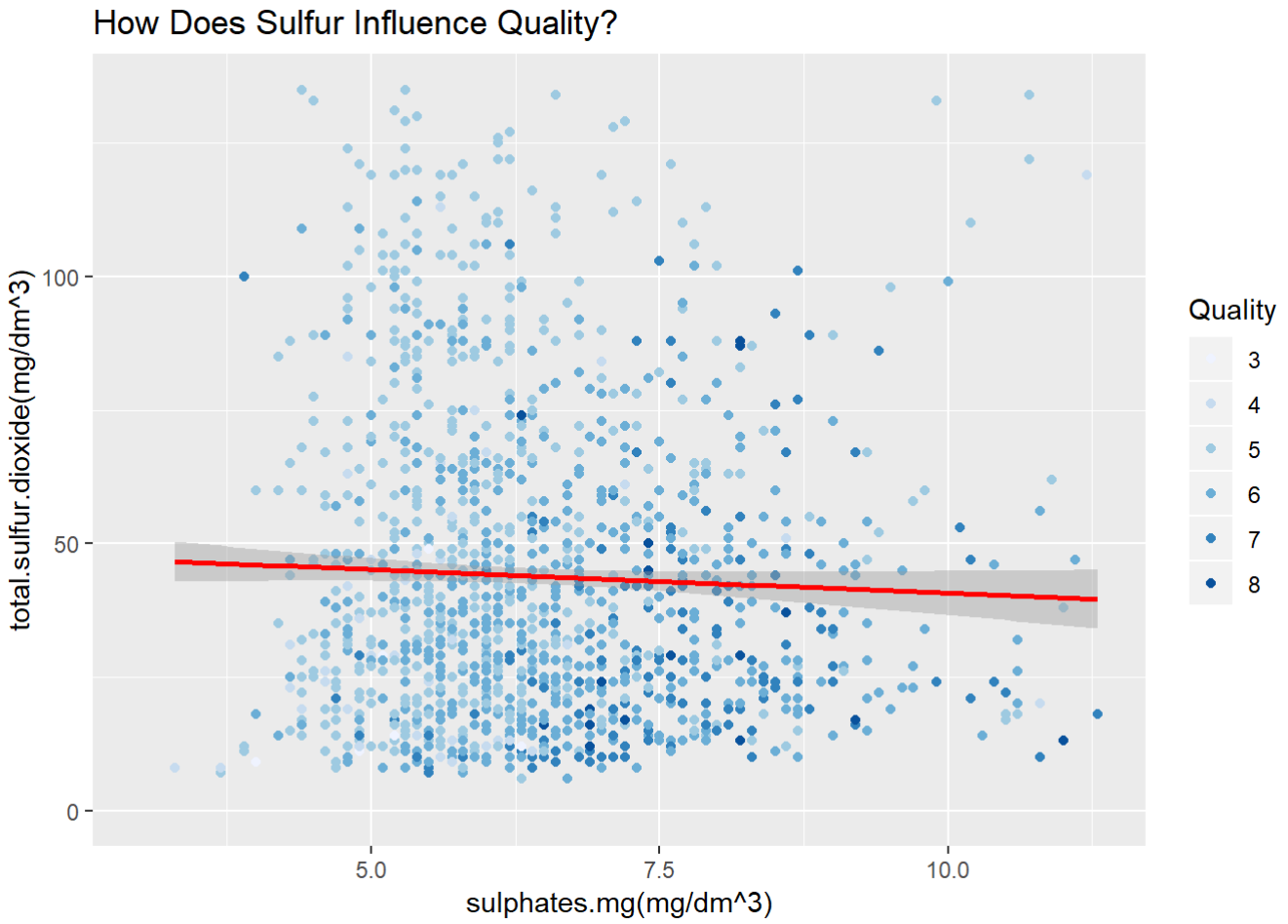
Plot Two



Description Two

From above graph, I have found that there would be certain relationship sugar and salt. In order to analyze these two variables more deeply, quality was included in multivariate anlaysis. In the model construction, independent variable was manipulated into ratio scale and conducted linear regression. The result revealed that salt sugar raito was a siginificant factor for wine's quality.

Plot Three



Description Three

Last point for the summary is about relationship with sulphates and total sulfur dioxide. Based on the multiple regression model, interesting point was that sulphates had postive coefficent, but totoal sulfur dioxide had negative with smaller size of coefficient than sulphates. This meant that sulphates could be a positive ingredient for wine quality, yet if there is too much of this material, there would be side-effects.

Reflection

In order to investigate which factor determines wine quality, several analyses were conducted. At first, univariate analysis with plots were done so that people could know basic features of variables in the data. In addition, variables were manipulated when it was essential such as total acidity or sugar salt ratio. Then, we have inspected relationship of two variables to look more deeply into the relationships between variables. For the final step, based on the findings from the first and second step, quality factor was added in multivariate analyses with proper plots.

Main analysis tool was based on descriptive analysis with ggplot as visualisation, and regression model for inferential statistics analysis. Of course, correlationtests were used in bivariate anlaysis step.

However, I still believe that there are room for improvement in this report. For example, I could have used mediation analysis with regression model in sulphates sector. Sulphates as an additive material might influence total sulfur dioxide level, and this affected total sulfur dioxide is the one that influences quality. I have actually tried to do this analysis, yet to fail due to lack of proper coding knowledge.

When it comes to predicting quality, logit regression might have been useful. Since the quality is categorical scale, technically simple regression or multiple regression might not be a perfect tool for analysis. With these things in mind, I will develop my statistical insight next time.