White Wine Exploratory Data Analysis by Paulina Grunwald

About The Dataset

In this post i would like to share with you exploratory data analysis project (which is part of my Udacity Data Analyst Nanodegree (https://www.udacity.com/course/data-analyst-nanodegree--nd002?v=a4)) of the white wine dataset. Like many I do like to have a glass of a good quality wine every now and then but I was always wondering what are the factors that govern a specific wine's quality. The white wine dataset that i will use in my analysis is public and includes various characteristics of white wine. The details are described in P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis "Modeling wine preferences by data mining from physicochemical properties.". The dataset can be downloaded here (http://www3.dsi.uminho.pt/pcortez/wine/). My project includes white wine data analysis done by exploring the relationships between different white wine characteristics and the "Quality" rating given by wine tasters.

Input variables (based on physicochemical tests) and units:

1. fixed acidity (tartaric acid - g / dm^3) - most acids involved with wine or fixed or nonvolatile (do not evaporate readily).
2. volatile acidity (acetic acid - g / dm^3) - the amount of acetic acid in wine, which at too high of levels can lead to a vinegar taste.
3. citric acid (g / dm^3) - found in small quantities, citric acid can add 'freshness' and flavor to wines.
4. Residual sugar (g / dm^3) - the amount of sugar remaining after fermentation stops.
5. Chlorides (sodium chloride - g / dm^3 - the amount of salt in the wine.
6. Free sulfur dioxide (mg / dm^3) - the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion.
7. Total sulfur dioxide (mg / dm^3) - 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
8. Density (g / cm^3) - the density of water is close to that of water depending on the percent alcohol and sugar content.
9. 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
10. Sulphates (g / dm^3): a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant/
11. Alcohol (% by volume) - alcohol: the percent alcohol content of the wine
12. Quality - score between between 0 - very bad and 10 - very excellent

Univariate Plots Section

I would like to explore the structure of the dataset (numer of observarions, variable names, dimension of the data frame etc.) so let's display the structure of the data frame:

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

White wine dataset consists of 13 variables and 4898 observations. I will not need X value since it's the number of the observation. I will remove the x variable from my dataframe.

```
## # A tibble: 4,898 × 12
##    fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
##            <dbl>            <dbl>       <dbl>          <dbl>     <dbl>
## 1            7.0             0.27        0.36           20.7     0.045
## 2            6.3             0.30        0.34            1.6     0.049
## 3            8.1             0.28        0.40            6.9     0.050
## 4            7.2             0.23        0.32            8.5     0.058
## 5            7.2             0.23        0.32            8.5     0.058
## 6            8.1             0.28        0.40            6.9     0.050
## 7            6.2             0.32        0.16            7.0     0.045
## 8            7.0             0.27        0.36           20.7     0.045
## 9            6.3             0.30        0.34            1.6     0.049
## 10           8.1             0.22        0.43            1.5     0.044
## # ... with 4,888 more rows, and 7 more variables:
## #   free.sulfur.dioxide <dbl>, total.sulfur.dioxide <dbl>, density <dbl>,
## #   pH <dbl>, sulphates <dbl>, alcohol <dbl>, quality <int>
```

As a first step of my exploration I will display summary statistics for all variables present in white wine data frame:

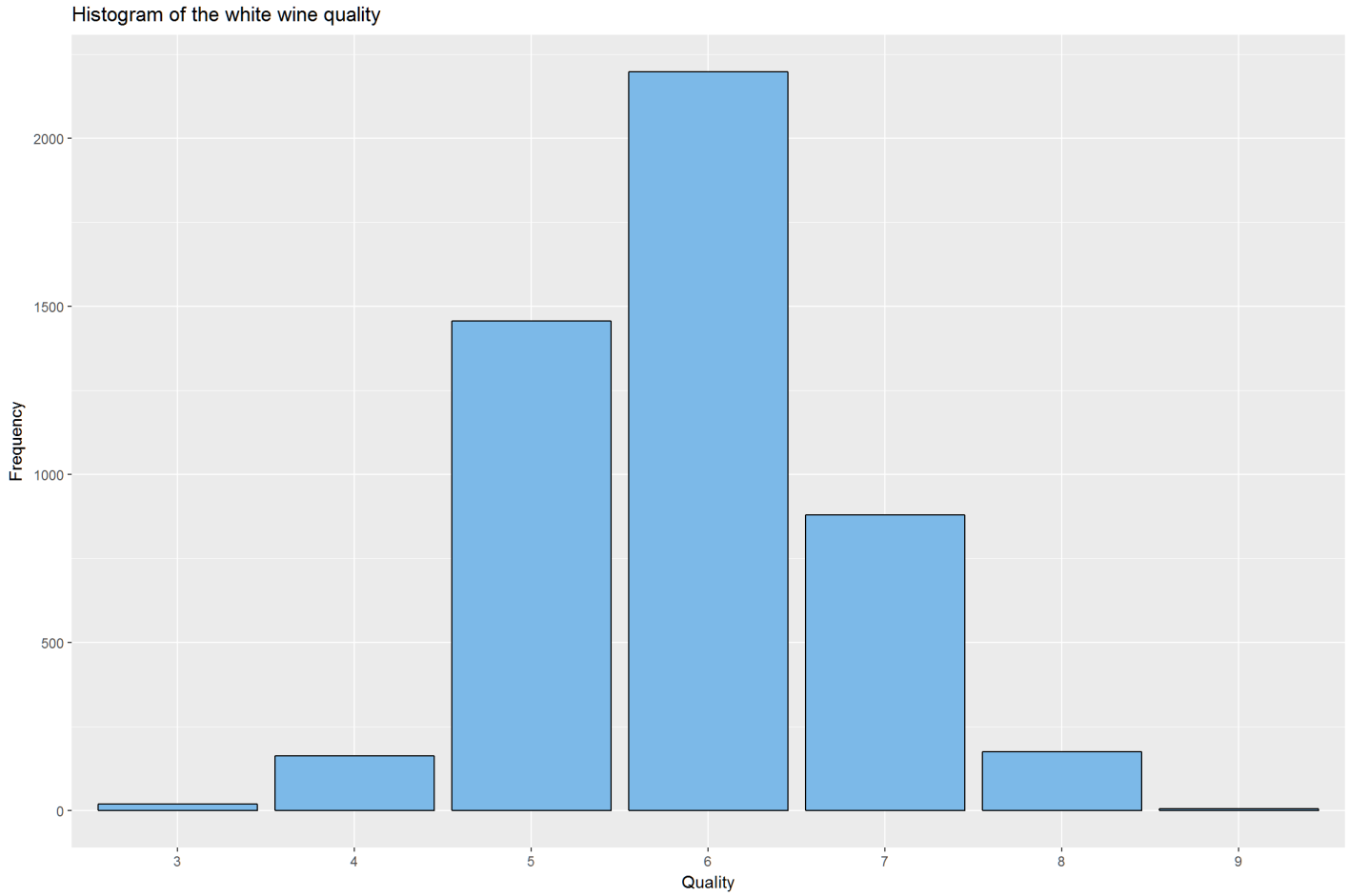
```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
##  Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
##  Mean   : 6.855   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
```

Looking at above displayed summary statistics we can observe that the quality variable range is from 3 to 9 with mean of 5.878 and median of 6. All of the variables the features have a minimum value greater than 0 except for citric acid. All of the variables have value of mean and median that are very close to each other. Most pH values fall between 3 and 3.3 which falls into reange of acidic to ultra acidic.Residual sugar has very big range as the smallest value is equal to 0.6 and largest to 65.8 g/dm^3) (which would suggest very sweet wine). Alcohol ranges between 8 and 14.2 procent. In my exploration I will try to access which properties of white wine have influence on the quality thus it is important to check what is the frequency distribution for the quality variable.

```
## # A tibble: 7 × 2
##   quality     n
##     <int> <int>
## 1       3    20
## 2       4   163
## 3       5  1457
## 4       6  2198
## 5       7   880
## 6       8   175
## 7       9     5
```

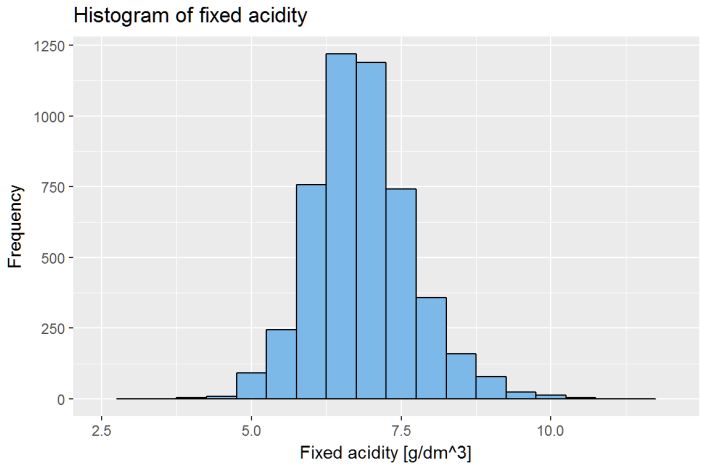
We have the highest number of observation for wines of quality 5 and 6. There are only 9 the highest quality wines (quality = 9).

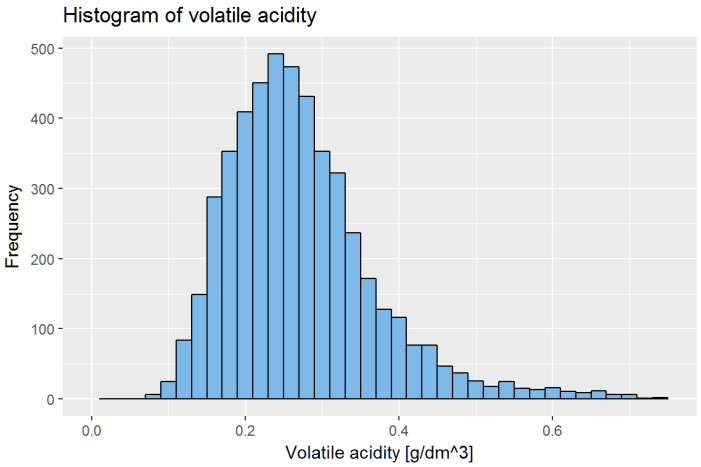
Now I will plot frequency distribution of the white wines using histogram.

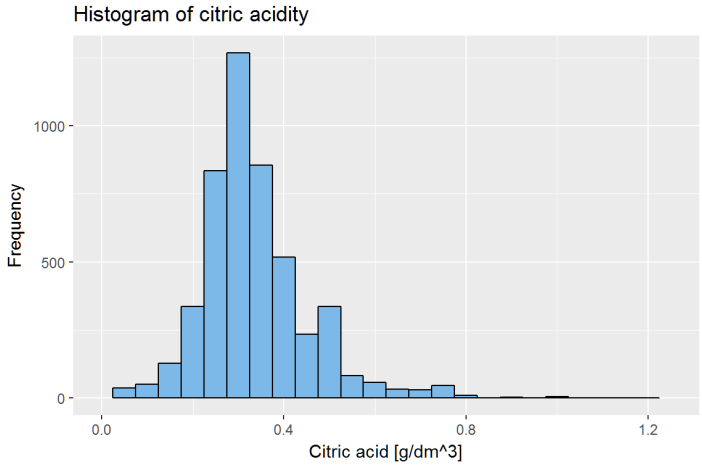


Acidity

I will create frequency distribution histograms for all variables related to the white wine acidity - fixed acidity, volatile acidity, citric acid.



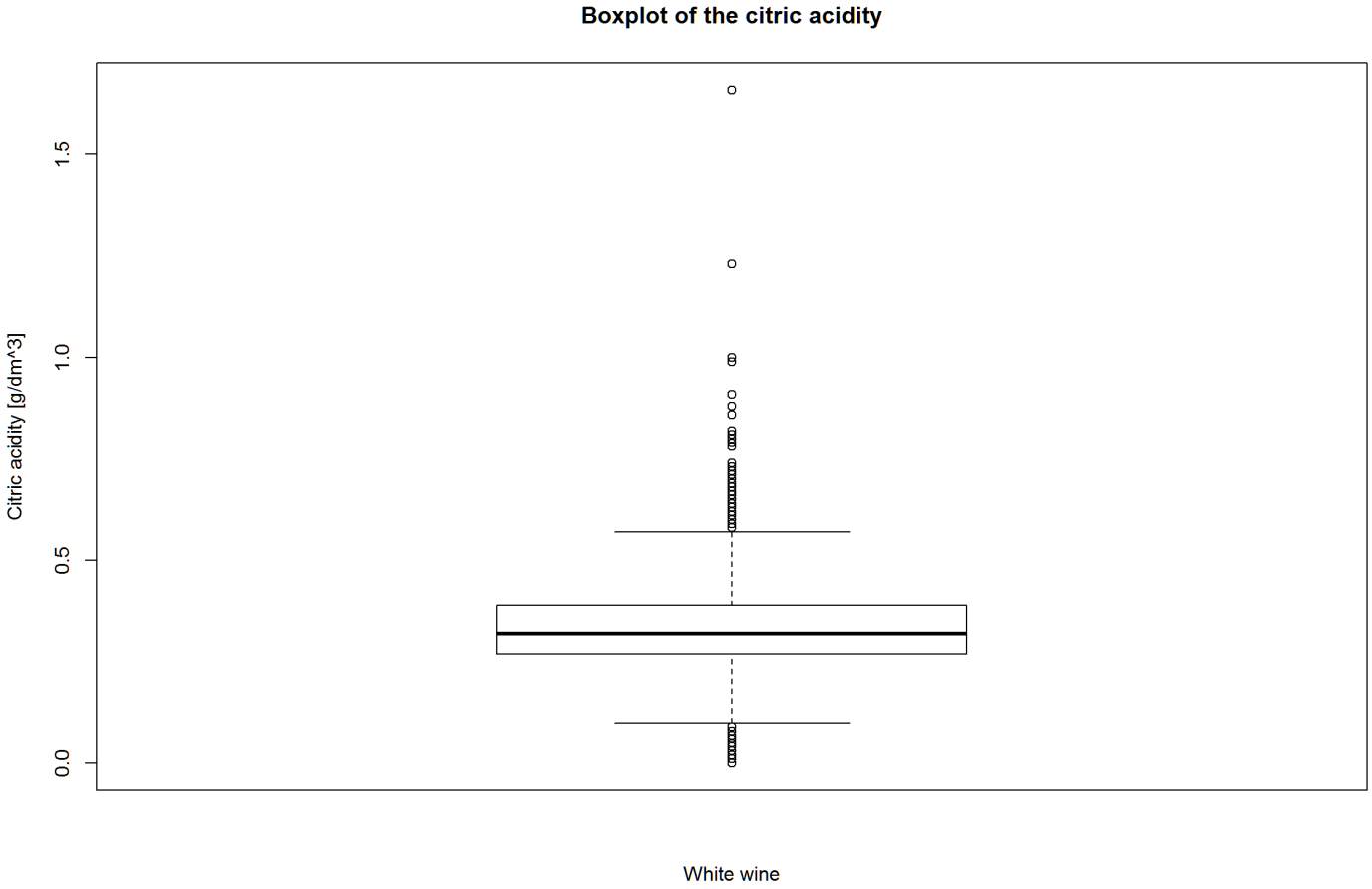




We can We observe that fixed acidity and citric acidity have a normal distribution. Volatile acidity can be characterized right-skewed distribution and has many outliers. It looks like citric acidity has many values equal to 0 and quite a number of outliers. Let's determine how many values are equal to zero:

```
## [1] 19
```

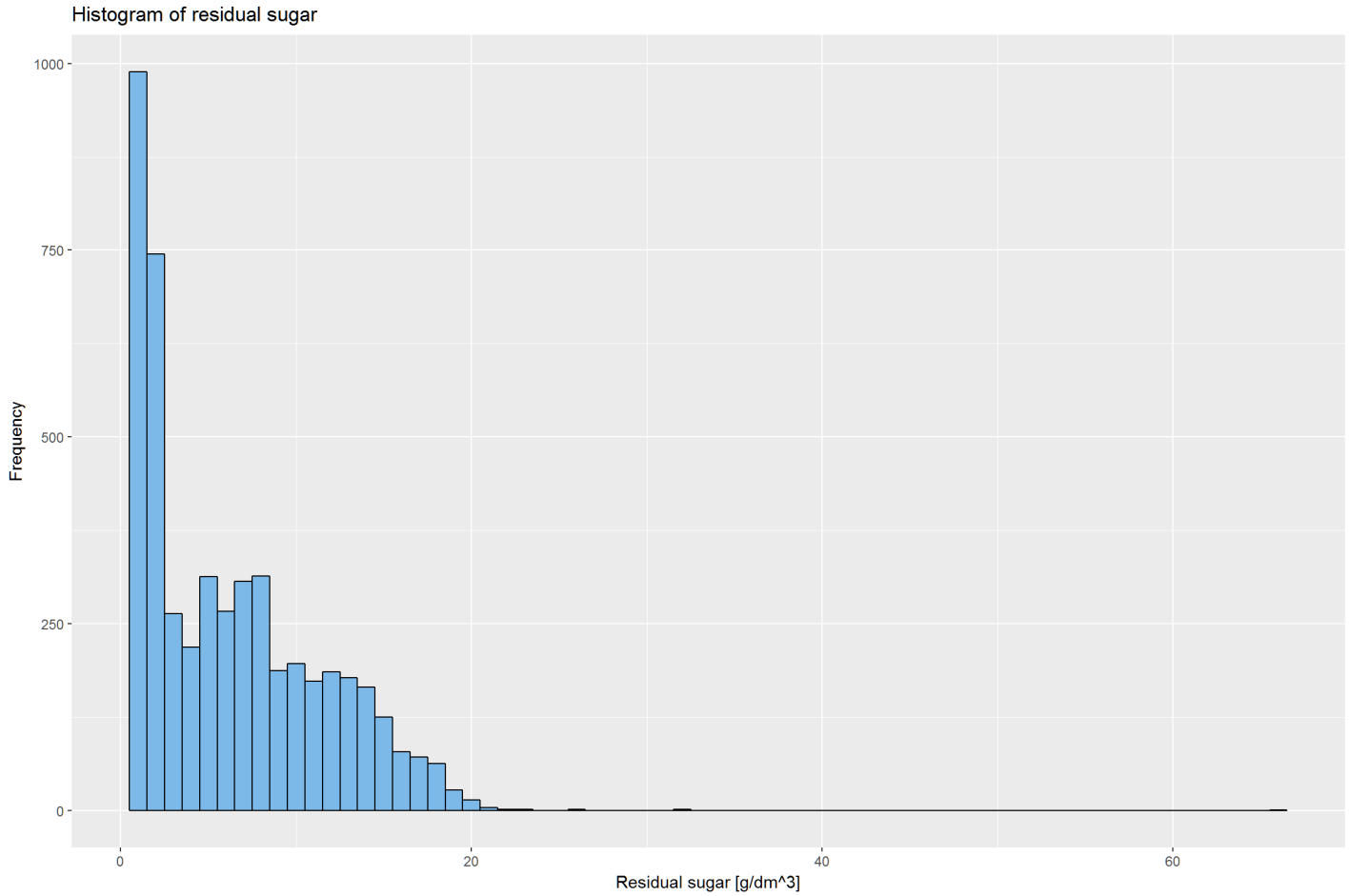
There are 19 values of citric acidity that are equal to 0. To get more insights about outliers let's have a look at boxplot of citric acidity. Now we will display the boxplot to get better feel about the outliers.



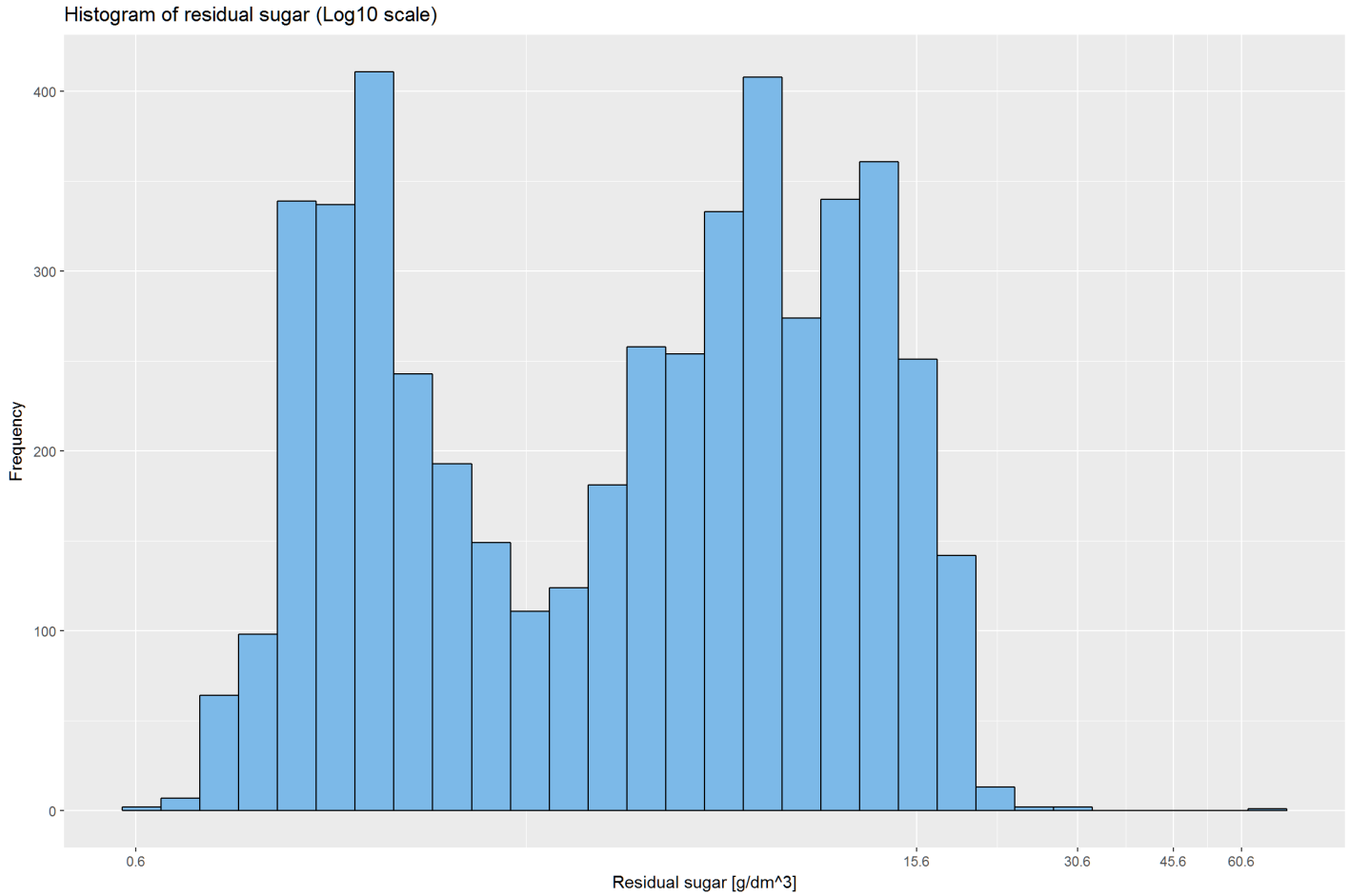
From above displayed boxplot we can observe that spread of citric acid data is quite big and there are a lot of outliers.

Residual sugar

Now I will display frequecy distribution of residual sugar.

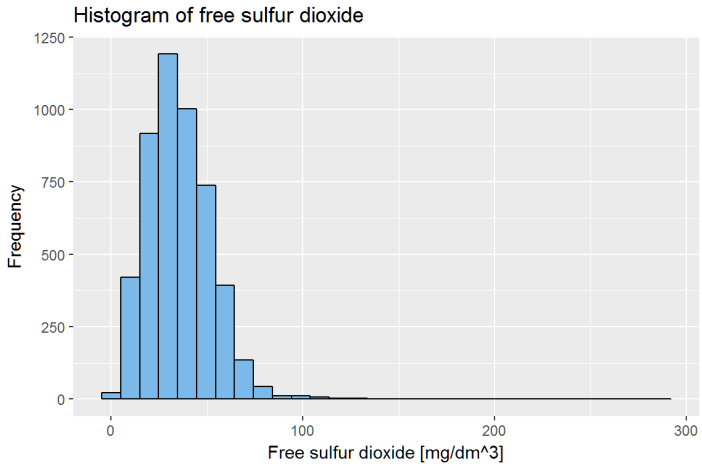


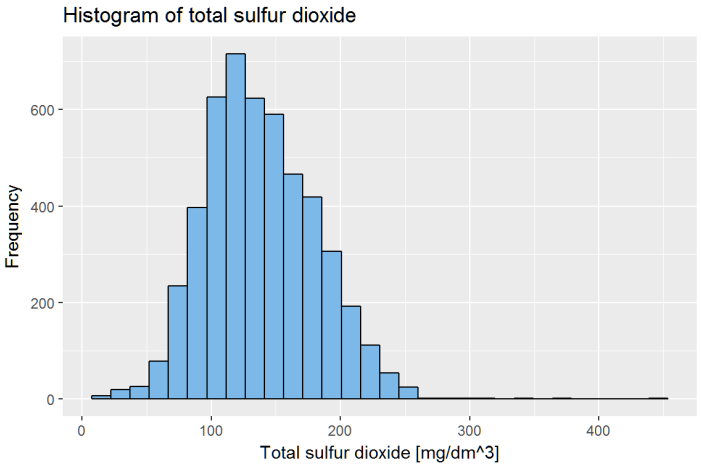
Residual Sugar has right-skewed distribution. We have see a lot of outliers and the highest values are over 60g/l. According to the EU regulations winez that have over 45g/l residual sugar are clasified as sweet. I will perform log10 transformation to the x axis in order to better see distribution of this variable.

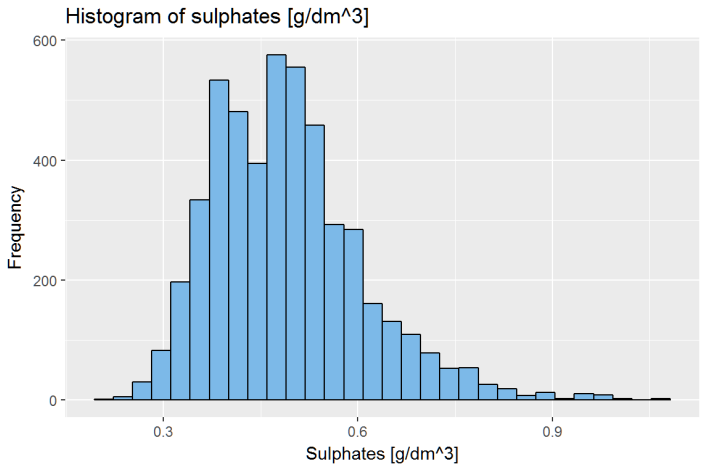


Residual sugar has bimodal distribution as it seems like two different bell in the distribution. I will try to explore more this interesting quality later on in my analysis.

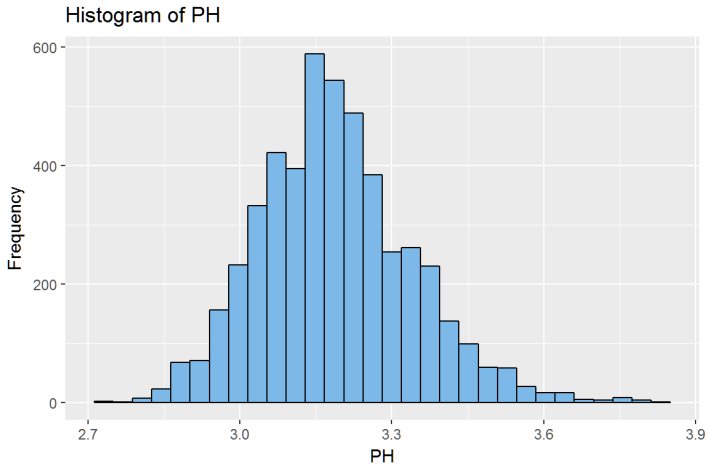
Sulfur dioxide and sulphates

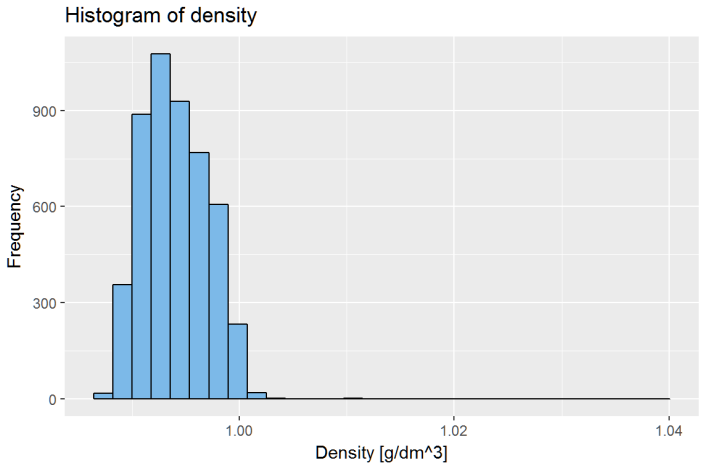


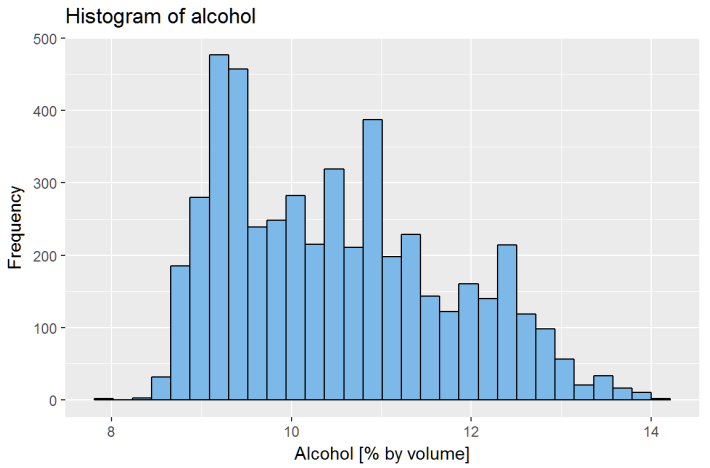




All three variables have normal distribution. Free sulfur dioxide and total sulfur dioxde have outliers and the maximum value8 and 3 times respectively bigger than the mean.

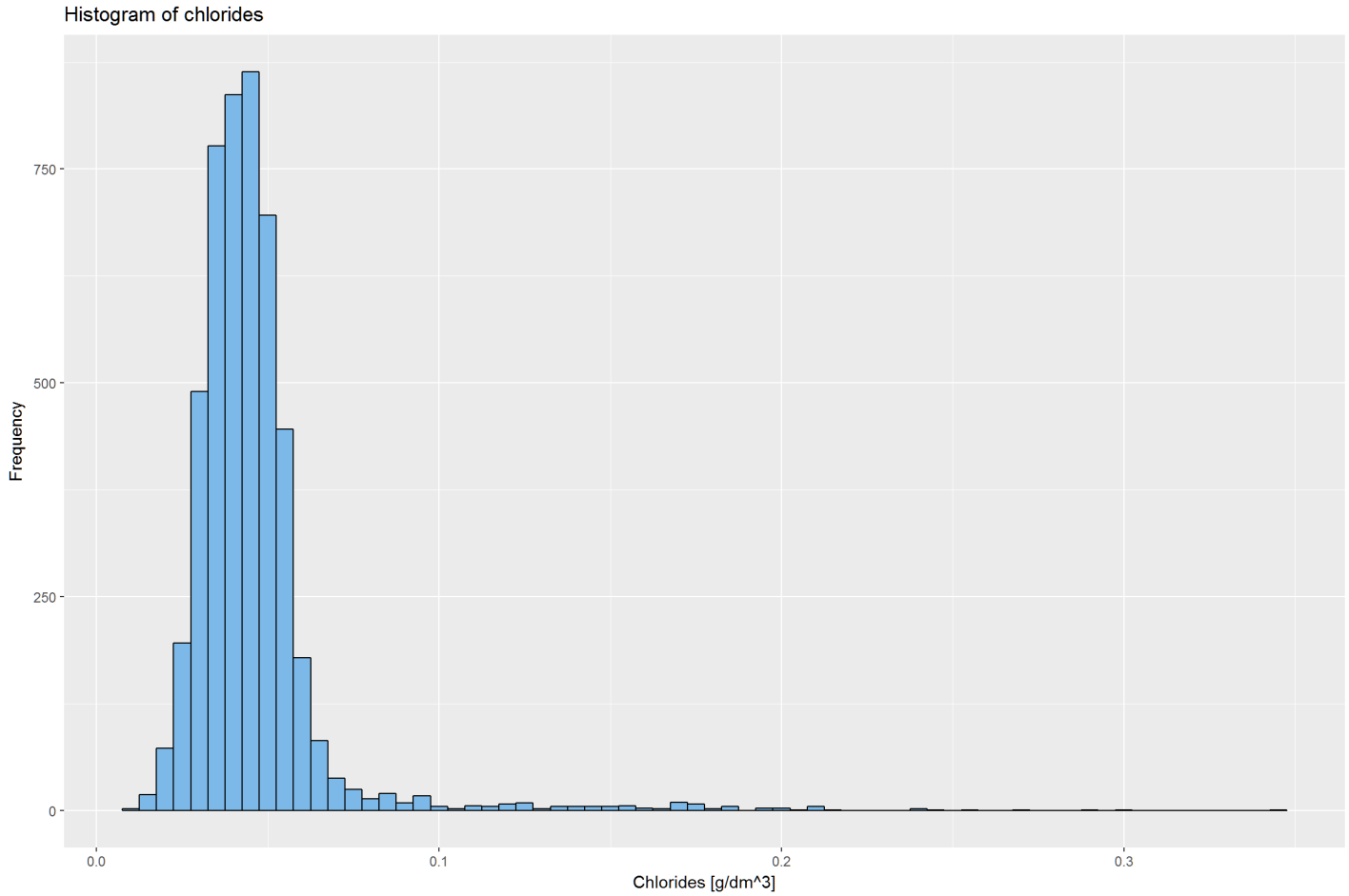




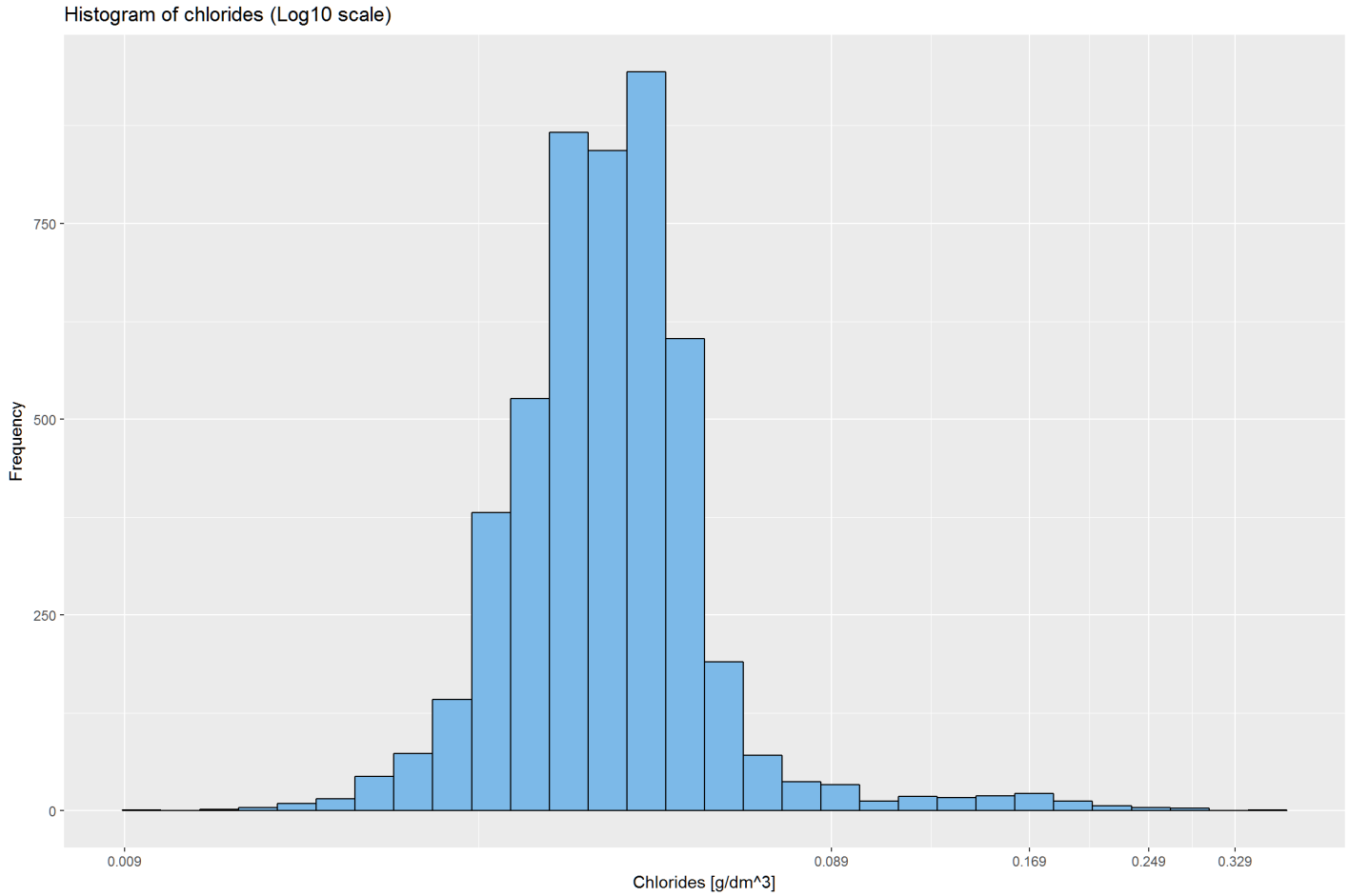


pH and density have also normal distribution but pH has many more outliers in comparision to density. Density variable has only 4 outliers. Alcohol has normal distribution and no outliers.

Chlorides



Chlorides are highly right skewed. I will use log10 transformation on x axis.



We can see that outliers shift a bit the bell curve of the chlorides frequency distribution.

Create new variables

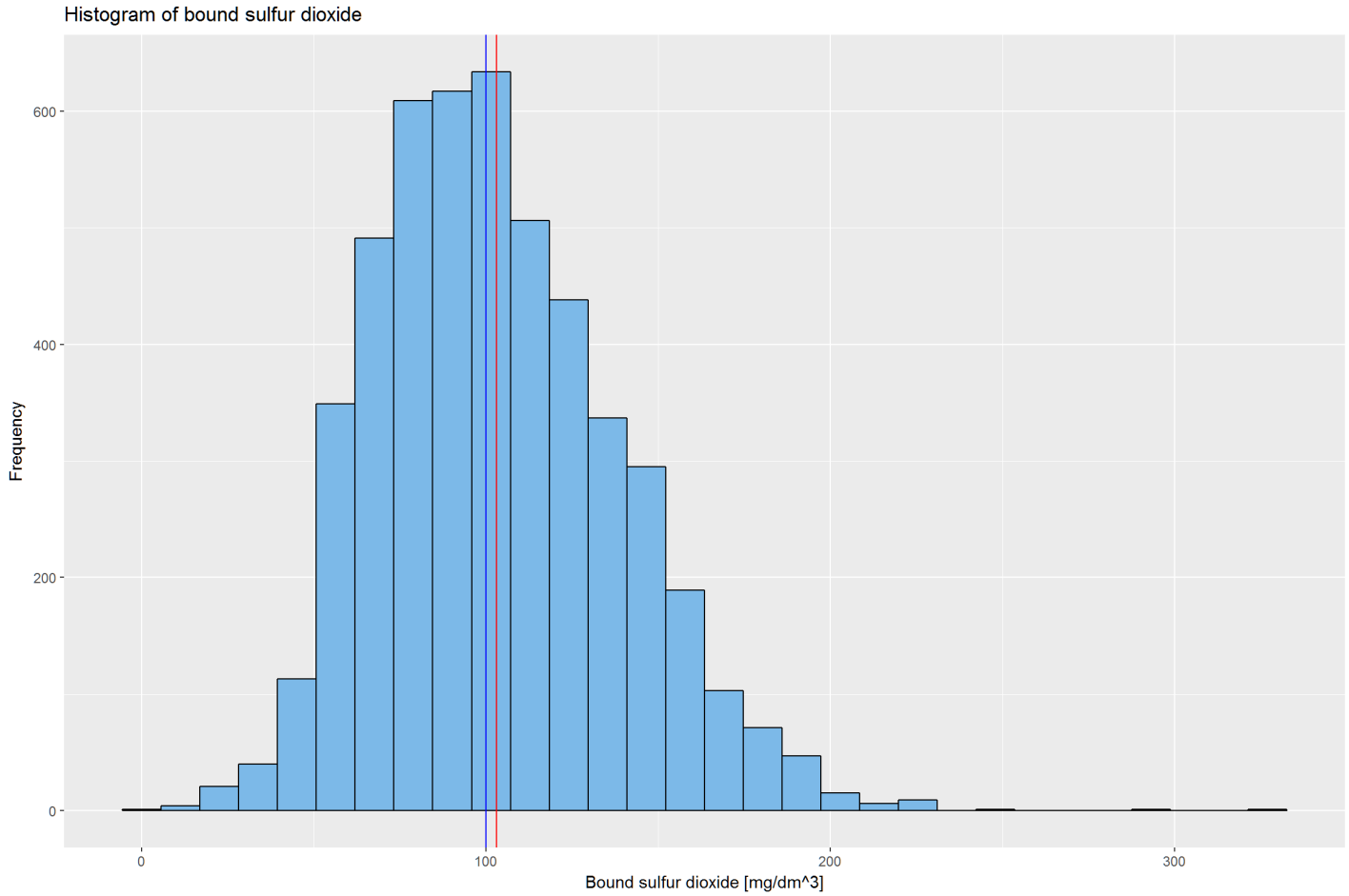
I will create following new variables: 1. Bound sulfur dioxide 2. Total acidity 3. Wine quality score 4. Wine strenght 4. Residual sugar range

1. Bound sulfur dioxide

I have decided to create bound sulfur dioxide variable.

Bound sulfur dioxide = total sulfur dioxide - free sulfur dioxide.

Red line: mean Blue line: median

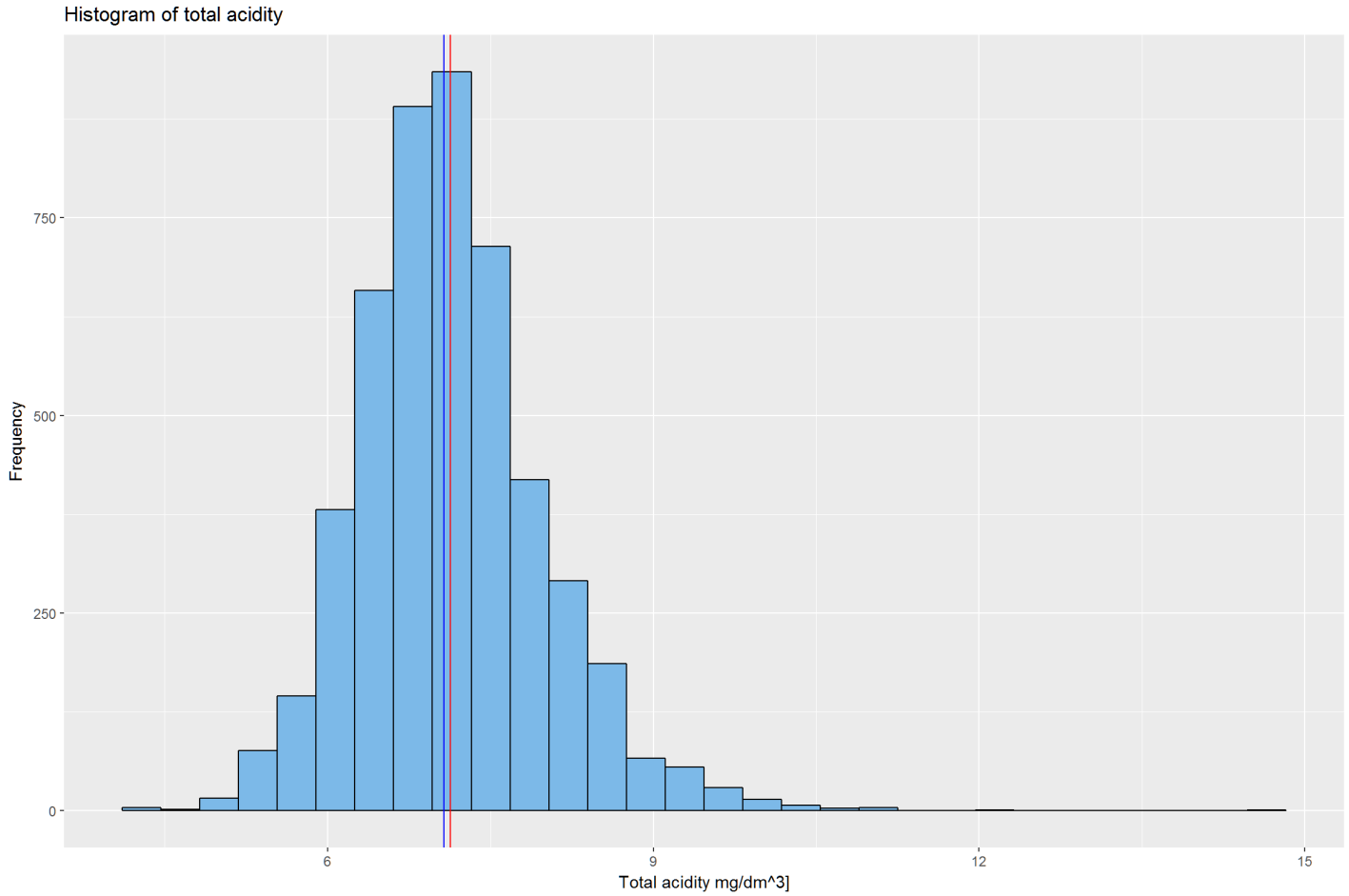


Let's look at the summary statistics for the bound sulfur dioxide.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     4.0    78.0   100.0   103.1   125.0   331.0
```

The range is quite large but mean and median are close to each other.

2. Total acidity I have calculated total acidity by adding volatile acidity to fixed acidity.

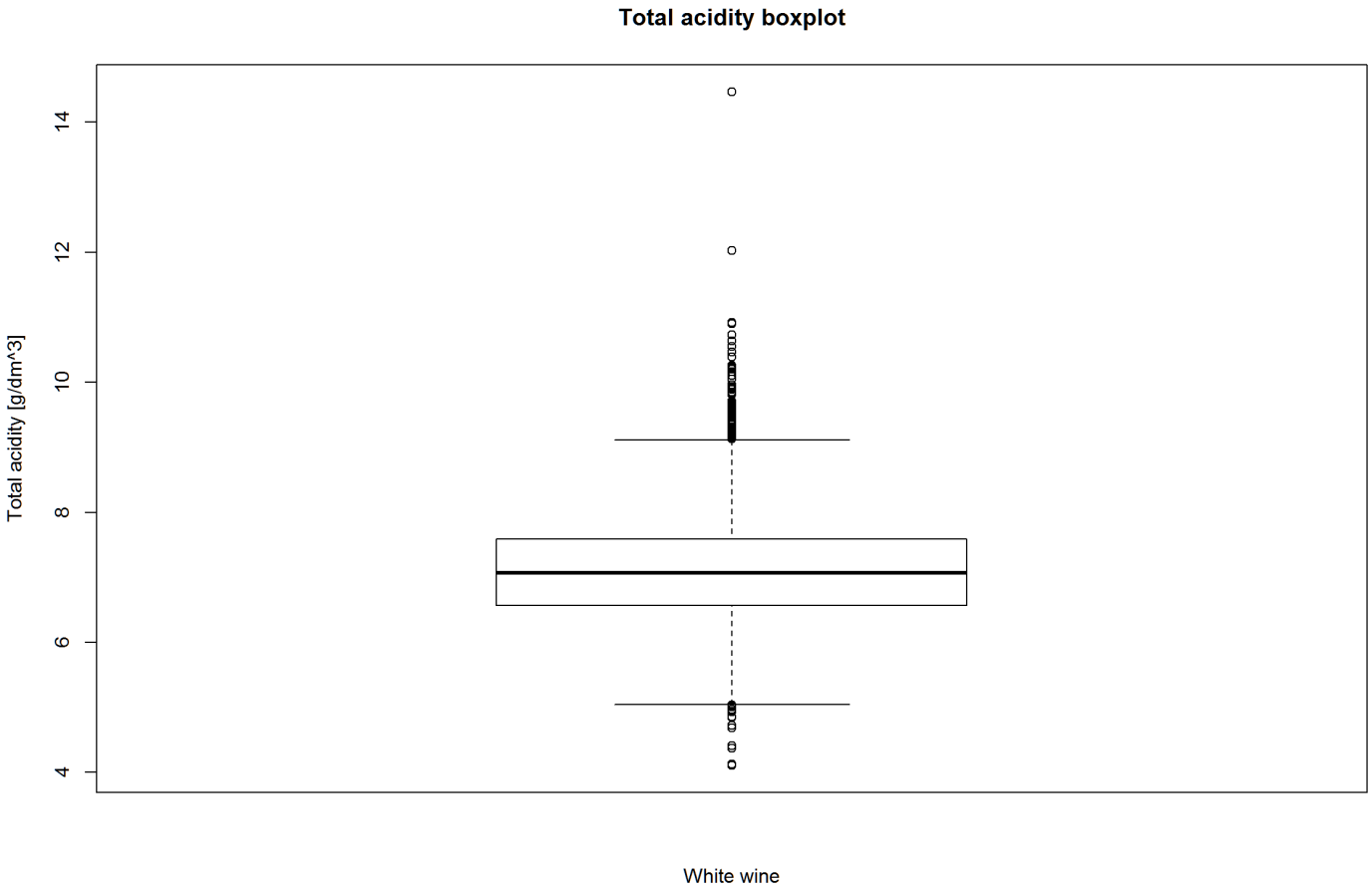



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.110   6.570   7.070   7.133   7.590  14.470
```

Let's see how many values of total acidity are equal or bigger than 10.

```
##    Mode   FALSE    TRUE    NA's
## logical    4878      20       0
```

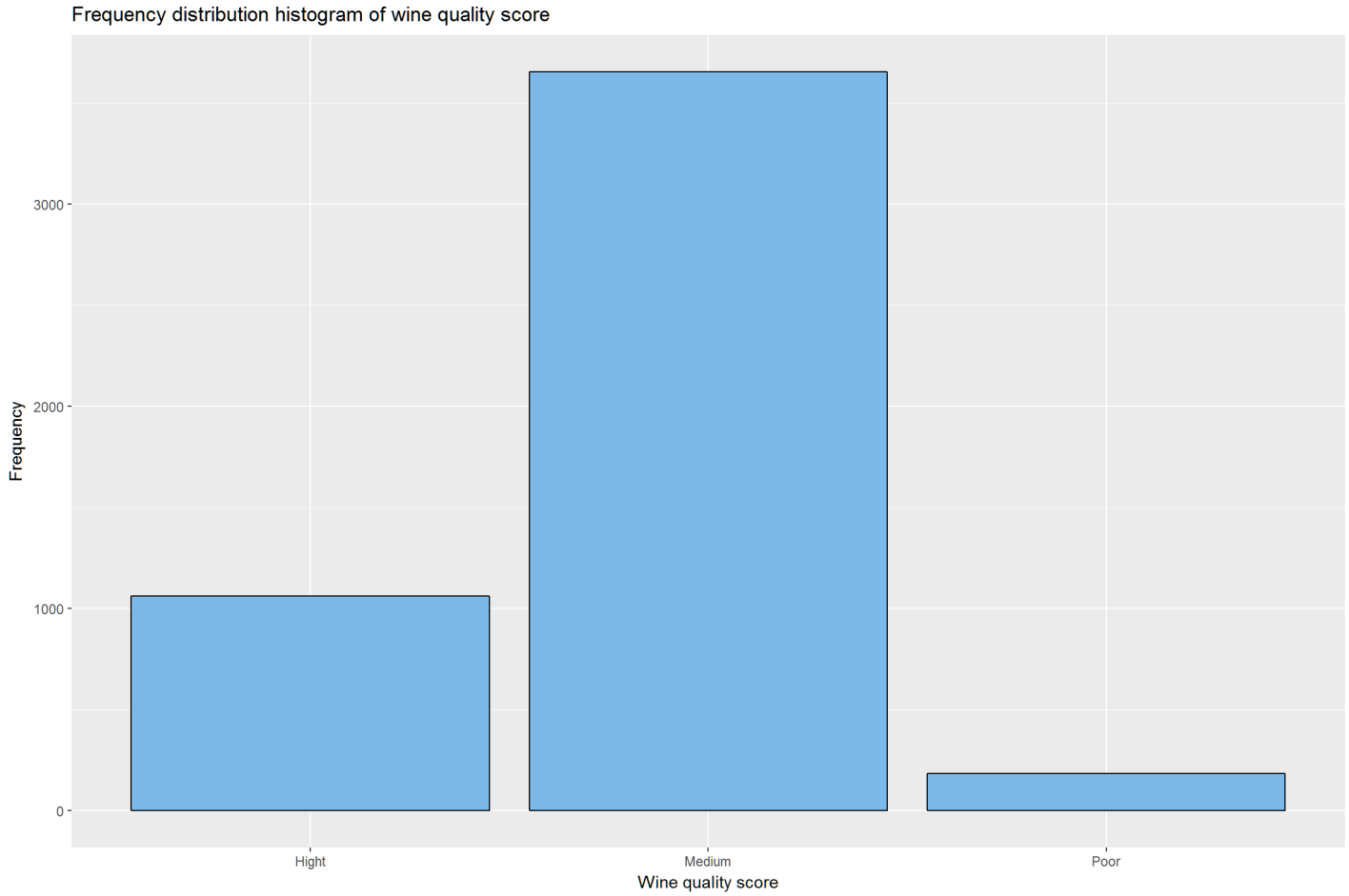
There are 20 values of total acidity that are bigger than 10. We could consider those as outliers but for clarity i will plot a boxplot.



Boxplot confirms that we have quite a number of outliers mostly. As anticipated the values of approx.higher than 10 are outliers.

3. Wine quality score

I'm going to create new quality variable quality. Quality scores: - <= 4: poor - >4 and <=6: Medium - >6 and <10: High

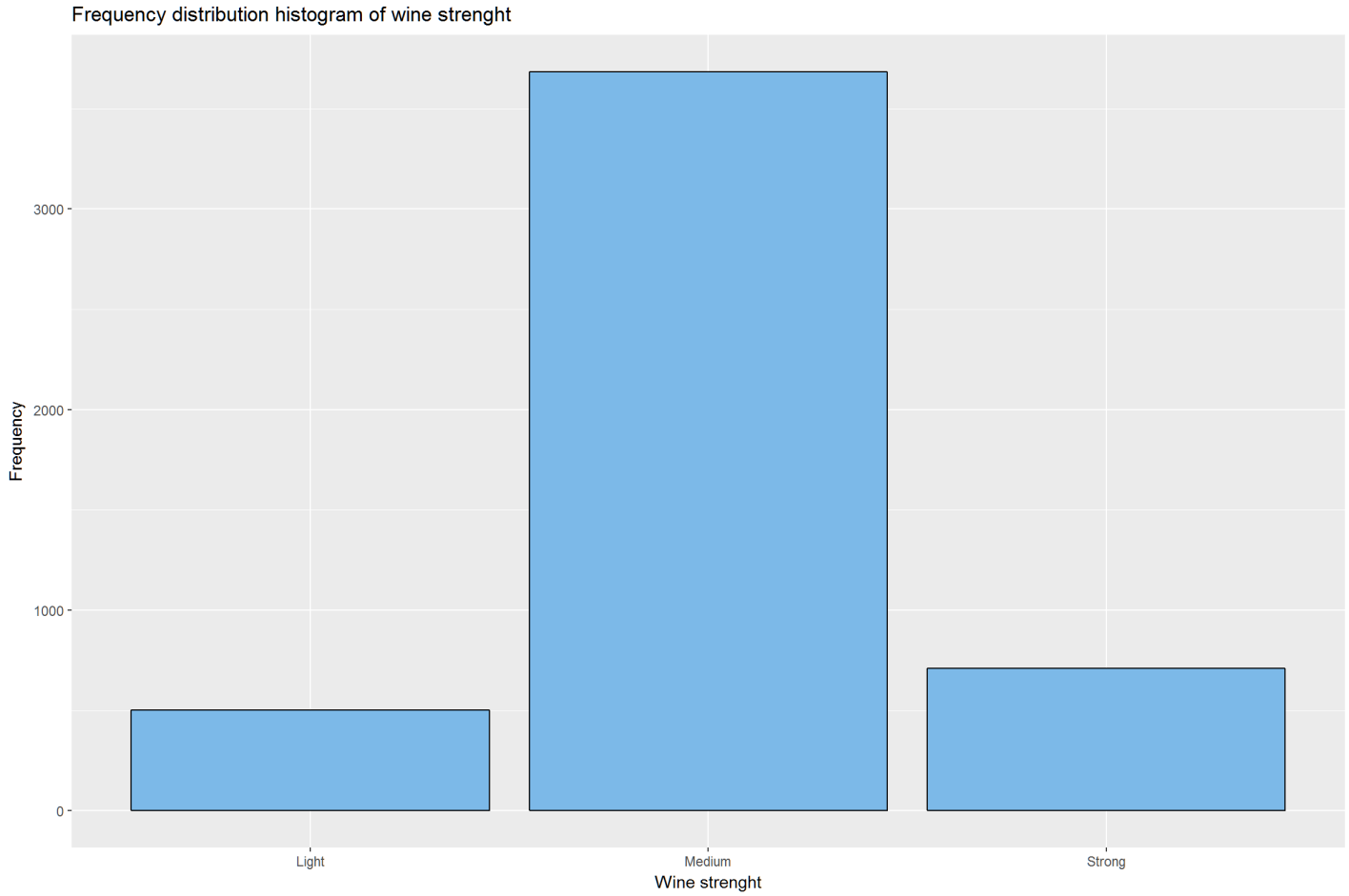


Let’s display count of high, medium and poor wine quality score.

```
##        x freq
## 1  Hight 1060
## 2 Medium 3655
## 3   Poor  183
```

3. Wine strenght

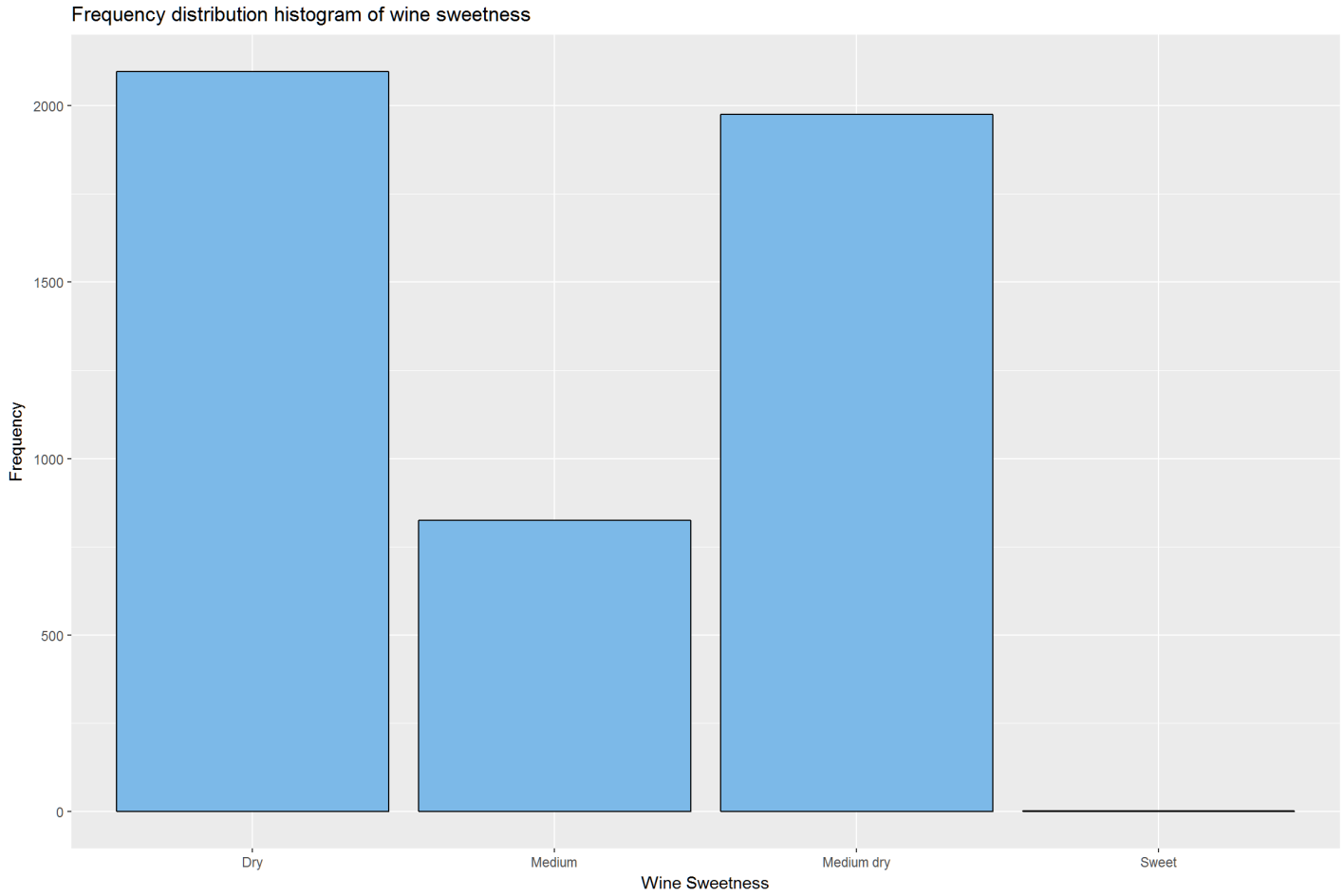
I will split alcohol content of the wines into following categories: * Low: <= 9 * Medium: >9 and <= 12 * High: > 12



4.Residual sugar score

According to EU regulation 753/2002, the following terms may be used on the labels of table wines and quality wines: - Dry: up to 4 g/l of residual sugar - Medium dry: up to 12 g/l of residual sugar - Medium: up to 45 g/l of residual sugar - Sweet: more than 45 g/l of residual sugar

Let’s create frequency distribution histogram for wine sweetness.

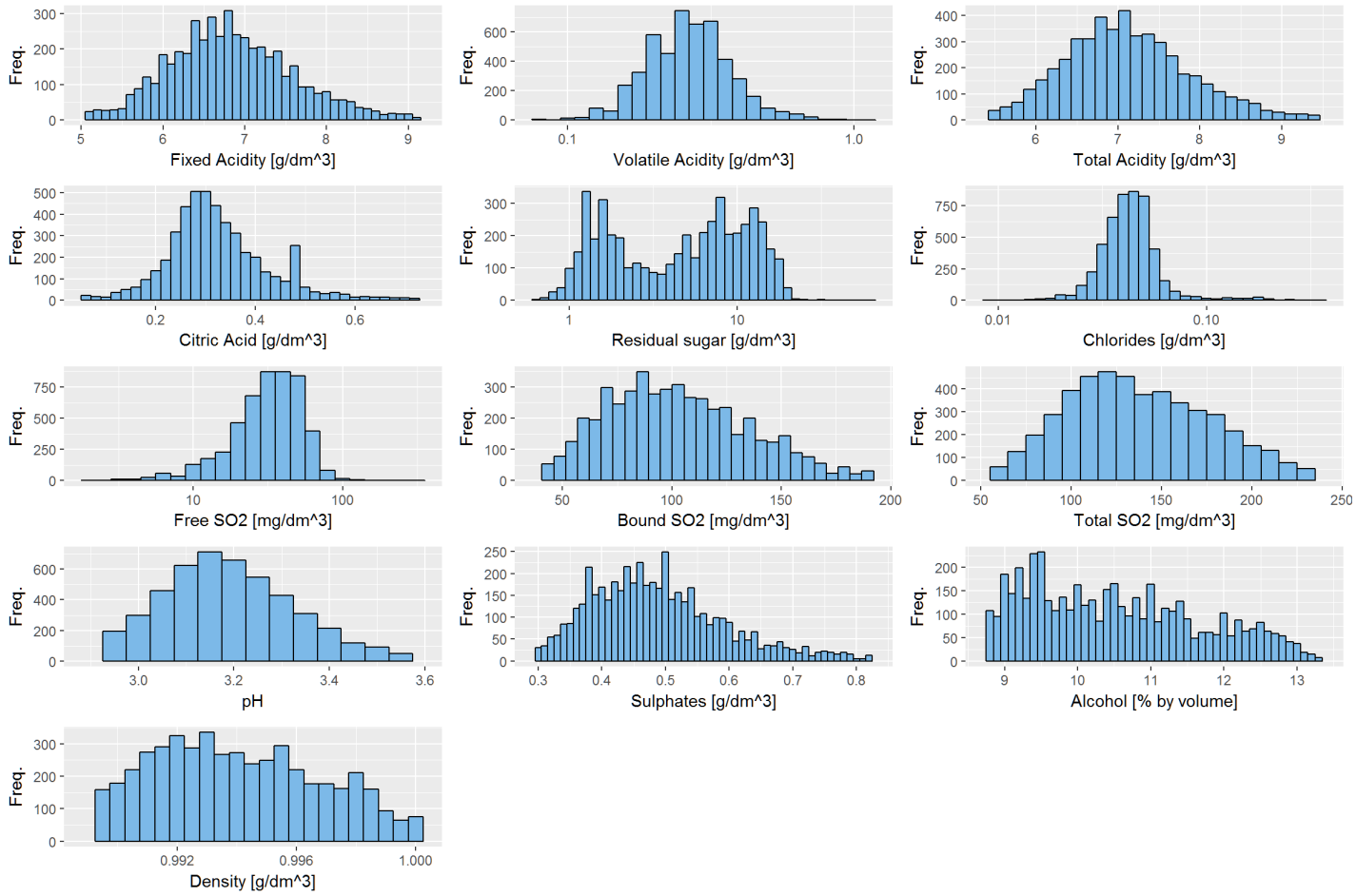


Let's see count of wine sweetness variable:

```
## # A tibble: 4 × 2
##   wine.sweetness     n
##            <chr> <int>
## 1            Dry  2097
## 2         Medium   825
## 3     Medium dry  1975
## 4          Sweet     1
```

There is only wine sweet wine which has more than 45g/l residual sugar. We have the highest number of dry and medium dry wines.

As a last step i would like to do create the plots with the bottom and top 1% removed from each variable. I want to do this to get a clearer view for the shape of the histogram distribution.I will use Log10 for residual sugar, chlorides and volatile acidity variables.



Univariate Analysis

What is the structure of your dataset?

My dataset is related to the quality of white wine. Each observation relateds to the chemical characteristic of the white wine. There are 4898 observations and 13 different variables (X, fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol and quality). I have removed X variable, as I will not use it in further investigation. All of the relevant variables are numerical except the quality variables which are integers. By looking at the dataset reference file, we learn that the quality score is graded within a 0 to 10 interval but in provided dataset we have only quality values from 3 to 9 (with median of 6 and mean of 5.878).

Following observation can be done regarding other variables: 1. pH ranges from 2.72 to 3.82 (with mean of 3.188) - all wines lie on the acidic side of the pH spectrum and most range from 2.5 to about 4.5 2. Only citric acid has values equal 0 (19 values of citric acidity is equal to 0). 3. Residual sugar histogram after transformaton of x axis to log10 scale shows bimodal pattern. This could sugest that we could have two different groups of wines in our dataset - with more residual sugar and with less residual sugar. 4. Volatile Acidity which is slightly right skewed but after applying log10 transformation on x axsis of the volatile acidity histogram has more bell-shaped signature. 5. Free sulfur dioxide has very high max. value which is 8 times bigger than the mean. 6. For all the mean and median have similar values.

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

In my evaluation i would like to access what wine characteristics have influence on wine quality thus wine quality will be my main feature. I will also explore influence of alcohole content, pH, acidity on wine quality.

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

In principle any of the 11 variables could influence the quality of the wine but from my research about white wine I found out that most likely alcohol and residual sugar would play significant role. Possibly density could be an influencing factor.

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

I have created following new variables: 1. Bound sulfur dioxide 2. Total acidity 3. Wine quality score 4. Wine strenght 5. Wine sweetness

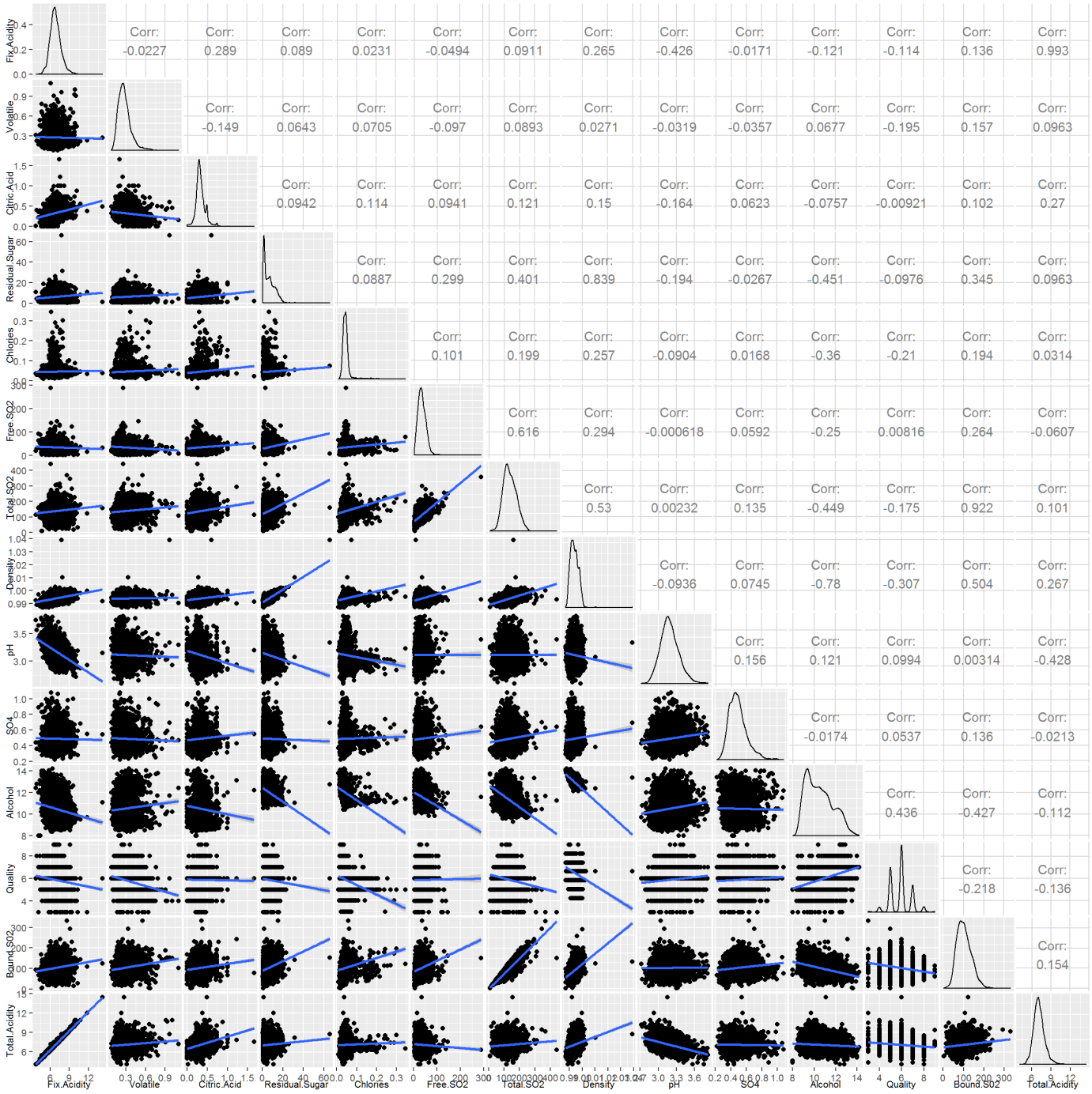
From chemical stand point total SO2 is a sum of free SO2 and bound S02. In our data set we have values for total and free SO2. The bound sulfur dioxide was derived from following formula: total SO2 = free SO2 + bound SO2. Total acidity of a wine is the combined sum of fixed(titratable) and volatile acids present. I have also decided to add 3 qualitative variables: wine quality score (separated in Poor, Medium and High), wine strenght(separated in light, medium and strong) and wine sweetness (Dry, Medium Dry, Medium, Sweet). I think those new variables will help me in accessing what influences white wine quality. I will include those variable in my analysis.

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

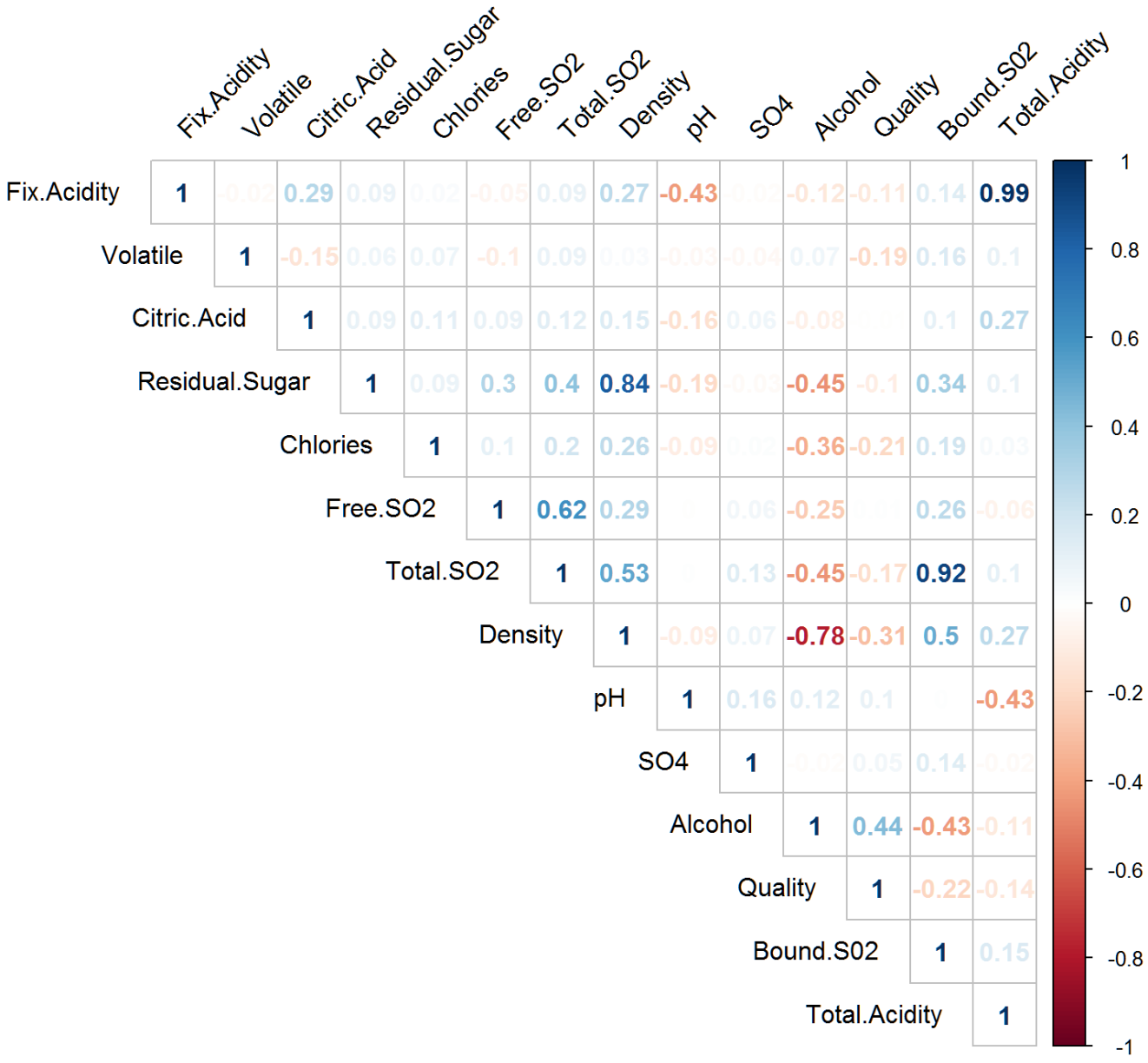
I have made number of adjustments to my datasets. I have changed the data type of quality variable from integer to ordered factor. I have applied as well x-axis log10 transformation to three variables: chlorides, residual sugar and volatile acidity. After the transformation the residual sugar turned out to have bimodal distribution and other normal. As a final adjustment i have displayed all the variables using frequency distribution histogram after removing 1% of top and bottom data which helped me to get better overview of the variables distribution.

Bivariate Plots Section

At the beginning og my bivariate analysis i would like to have a look at the correlation plots for all the variables.



And also corrplot which shows more clearly cor. values:

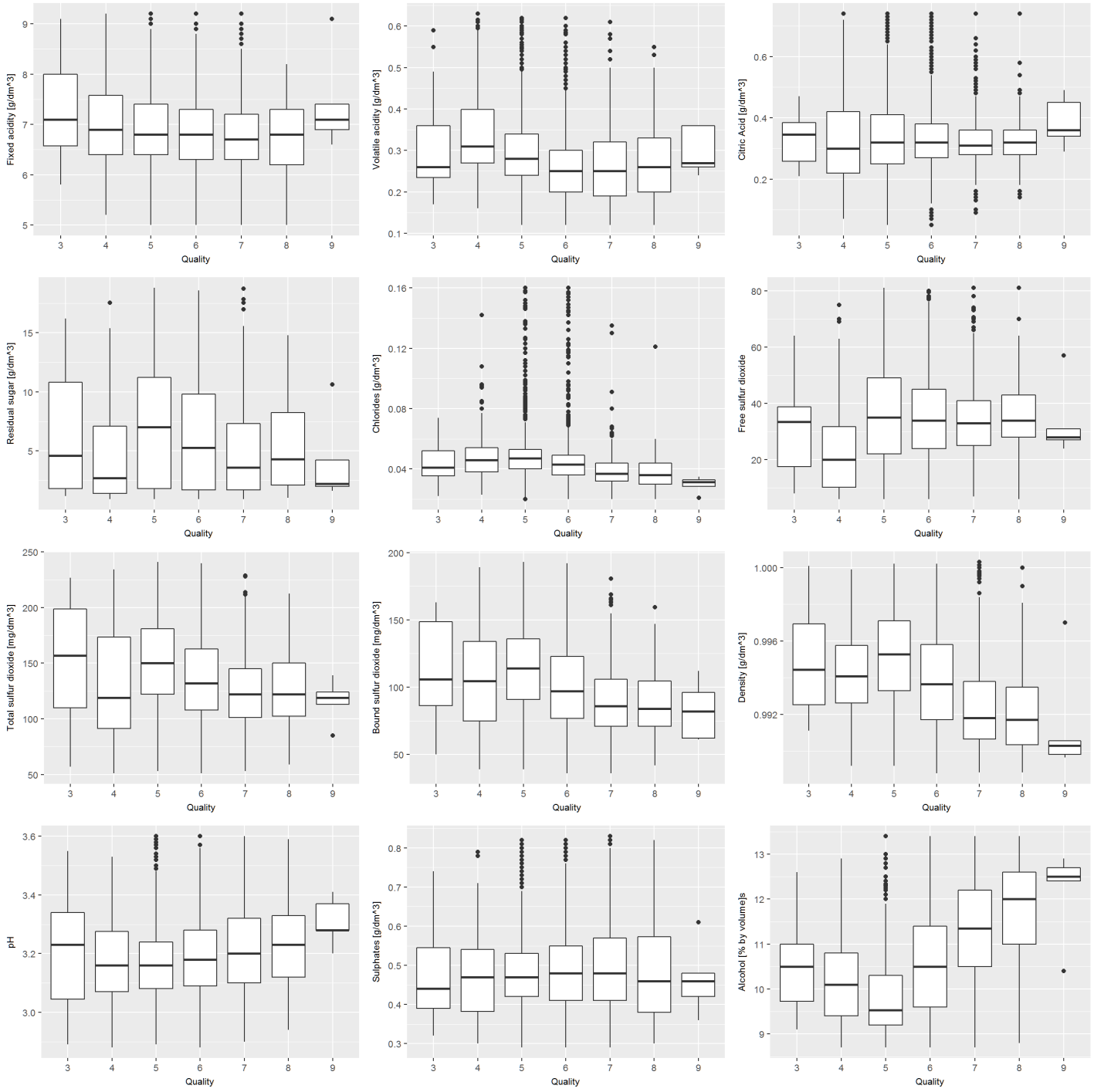


From above displayed plot we can notice that following variables have very good correlation:

- quality and alcohol: 0.44
- quality and density: -0.31
- residual sugar and density: 0.84
- residual sugar and alcohol: -0.45
- residual sugar and bound sulfur dioxide: 0.34
- fixed acidity and pH: -0.43
- Total acidity and fixed acidity: 0.99 (*not a suprise as the variables are derived from one another!*)
- chlorides and alcohol: -0.36
- free sulfur dioxide and density and total sulfur dioxide: 0.62
- density and bound sulfur dioxide/total sulfur dioxide: -0.58/0.27
- density and alcohol: -0.78

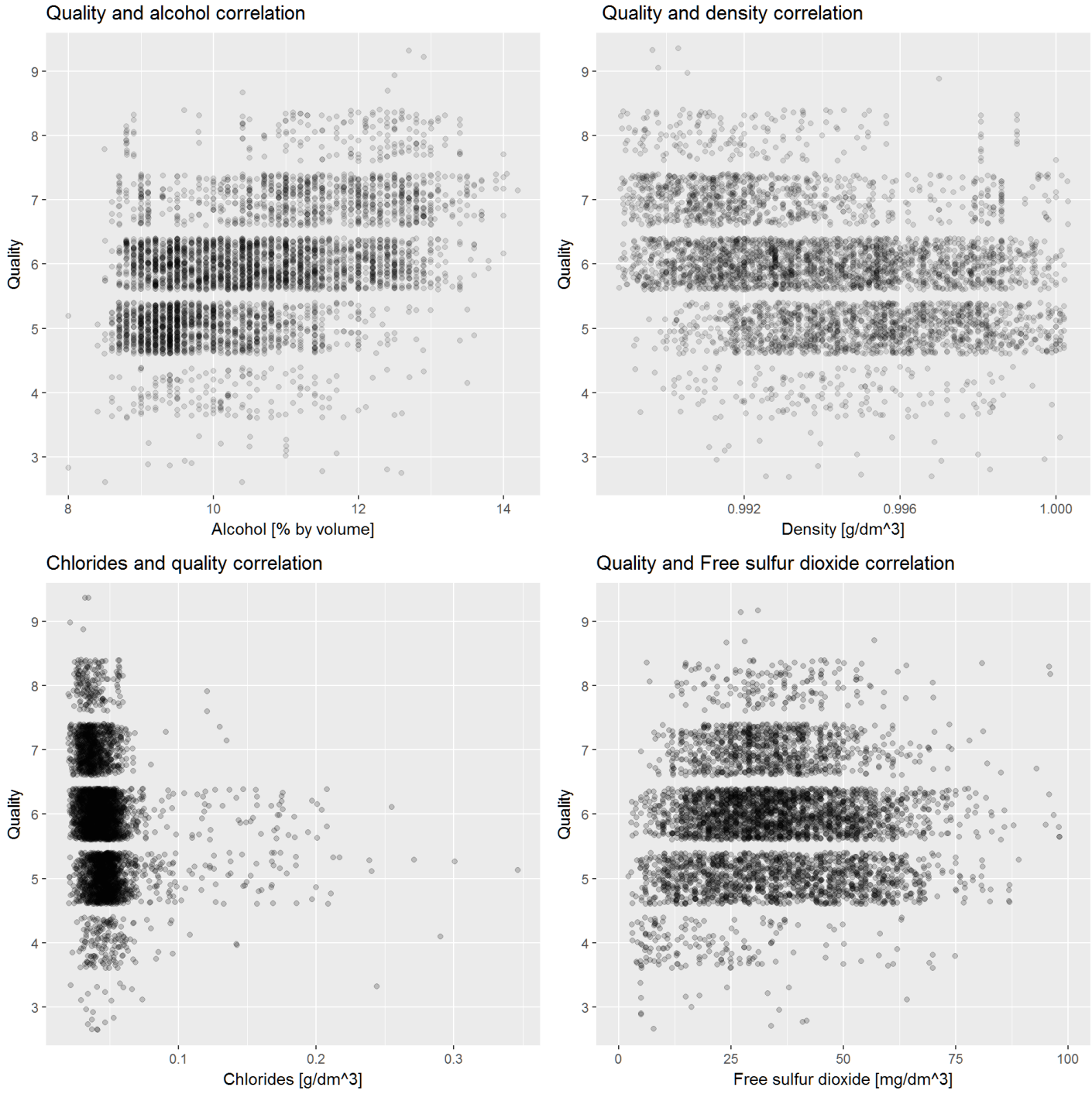
In my exploration i will apply following techniques to get better insight from my data: - add jitter and transparency (helps with overplotting of the data)for that scatterplot data - adjusting the axes (limits, log10 etc.), - remove top and bottom 1% of outliers, - I will use scatterplots to visualize the data.

I would like continue my analysis with a quick look on the boxplot visualization of our variables against the quality factor to get even better insights of the correlation between quality and other variables. I will remove to a



1. Quality and alcohol, density, chlorides, free sulfur dioxide

Let's first explore relationship between quality and few other variables. In quality vs. density and in quality vs. chlorides plots i will remove 1% of top and bottom values for the x axis to get better view of the trend of the data.

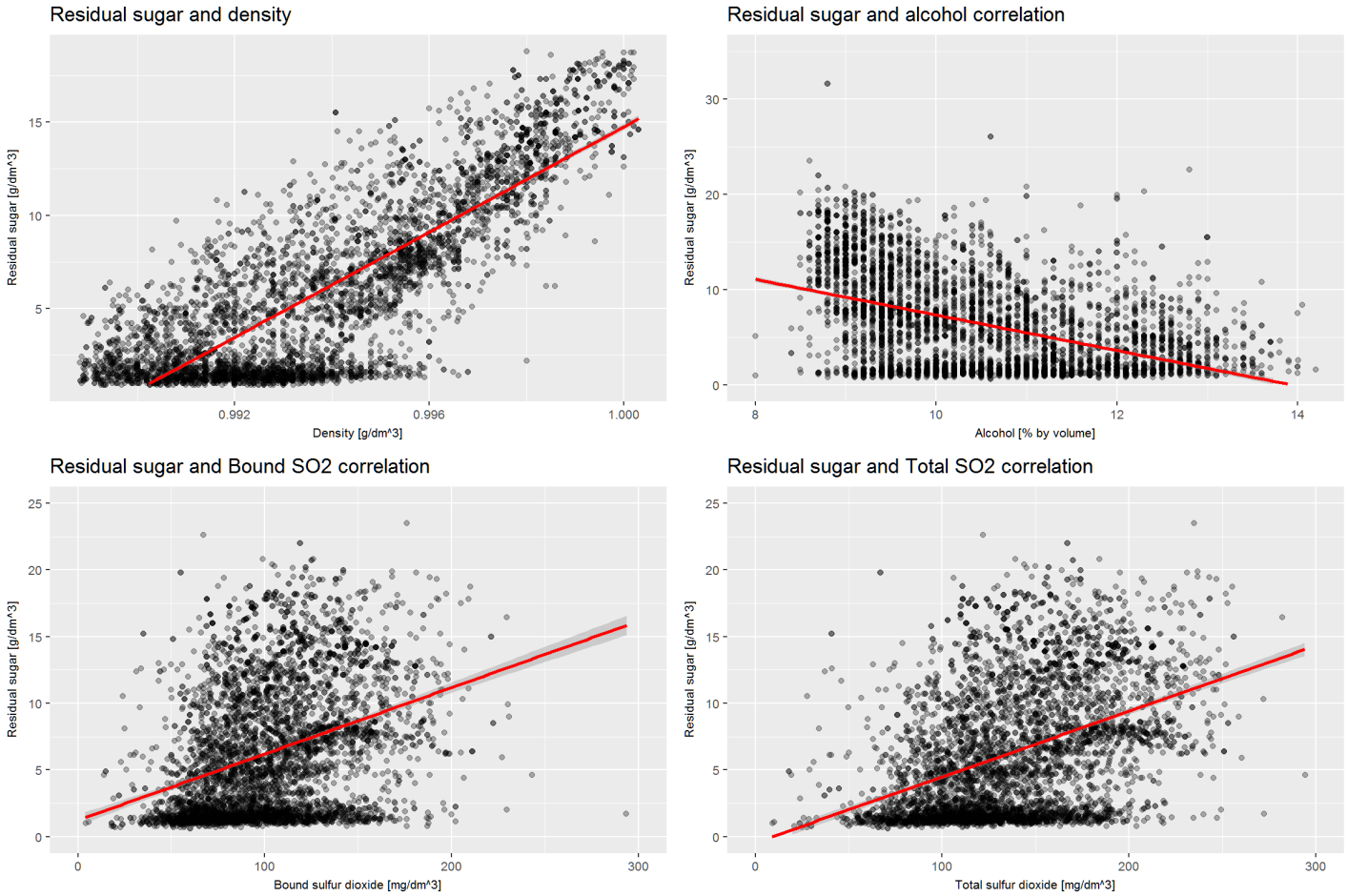


Quality has the strongest correlation with density (-0.31), alcohol (-0.43), chlorides (-0.21), bound sulfur dioxide (-0.22). From above displayed plots we can observe: - quality is well correlated to alcohol - with increase content of alcohole quality rises - downward tren can be observed forcorrelation between quality and density - density decreases with increase of quality. In this plot i have removed 1% of top and bottom values. - there is rather week correlation (-0.21) between quality and chlorides. - there is very week correlation between quality and free sulfur dioxide.

We can conclude that quality ahs the strongest correlation with alcohole. In the multivariate analysis I will preform additional investigations between those variables which probably would help me to uncover other trends.

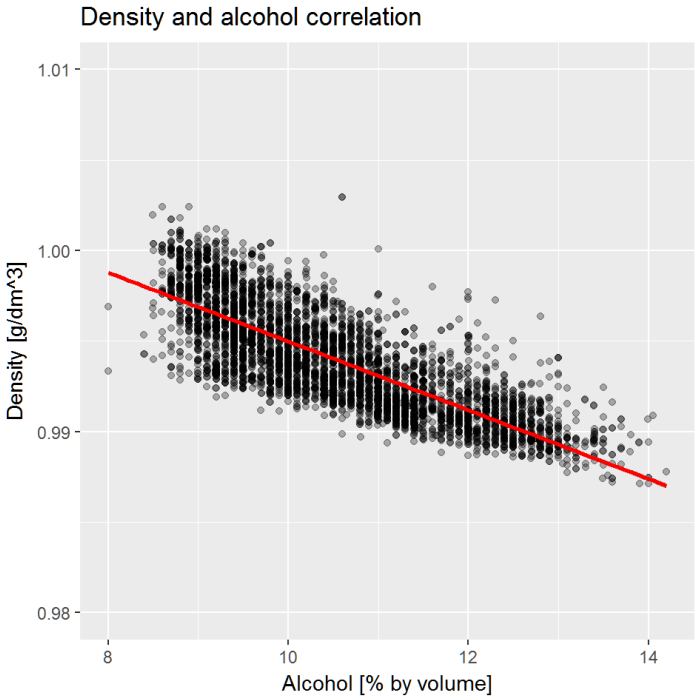
2. Residual sugar and density, alcohole,bound sulfur dioxide

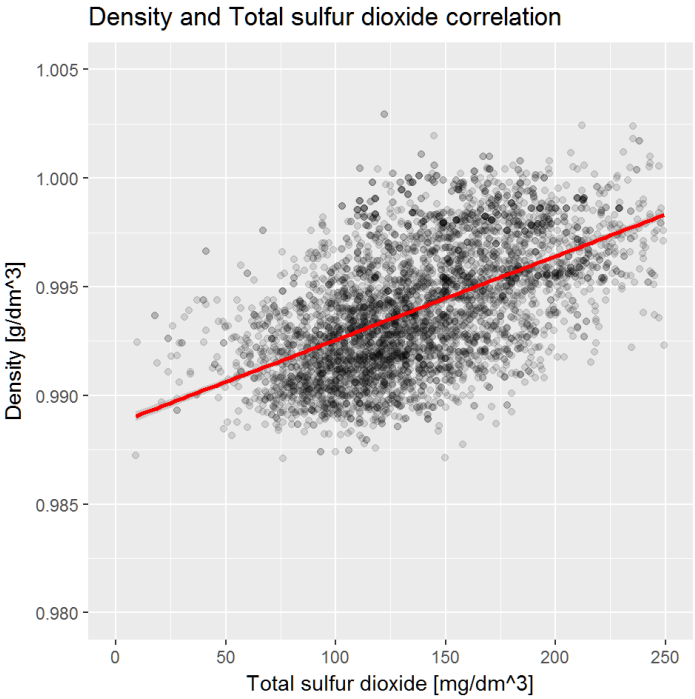
Note that in the first plot *density vs. residual sugar) I will remove 1% of top and bottom values for both x and y axis to remove outliers and to get better feel for the correlation between the variables.

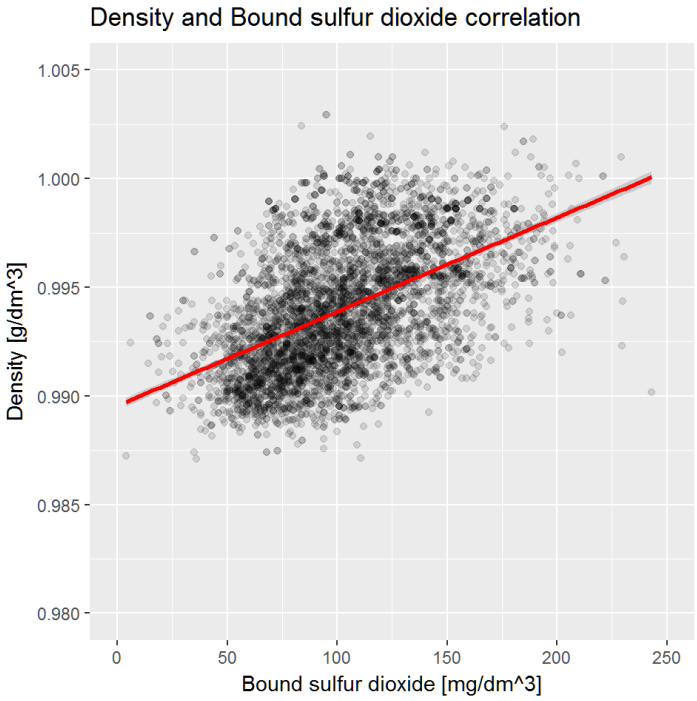


Residual sugar has good correlation with density (0.84), Total sulfur dioxide (0.4), Alcohole (-0.45), Bound sulfur dioxide (0.34). Residual sugar and density has the strongest correlation between all investigated variables (except Bound sulfur dioxide and Total sulfur dioxide which have 0.92 correlation but those variables are derived from each other so very strong correlation is not a suprise).

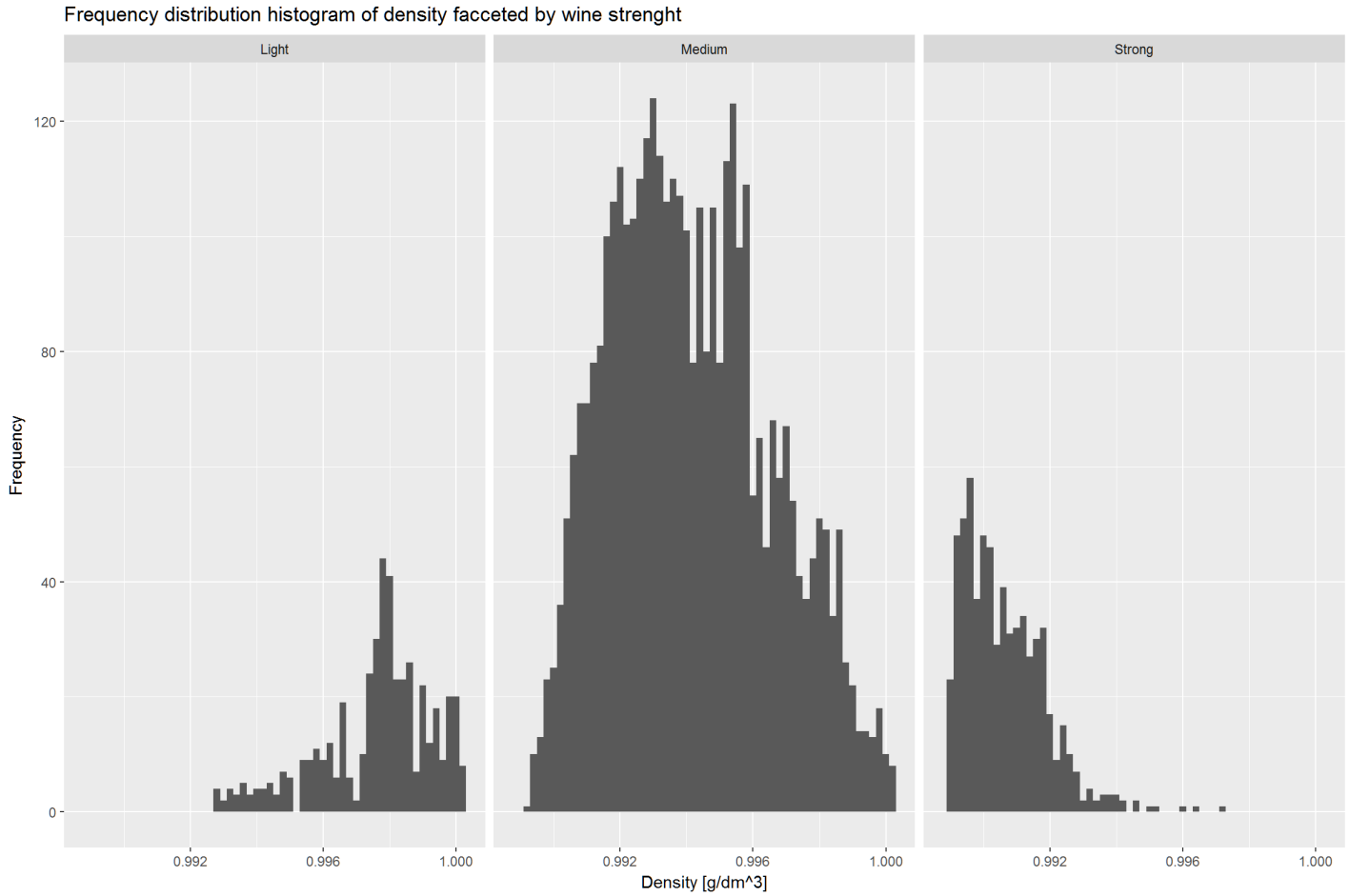
4. Density and alcohol, total sulfur dioxide, bound sulfur dioxide



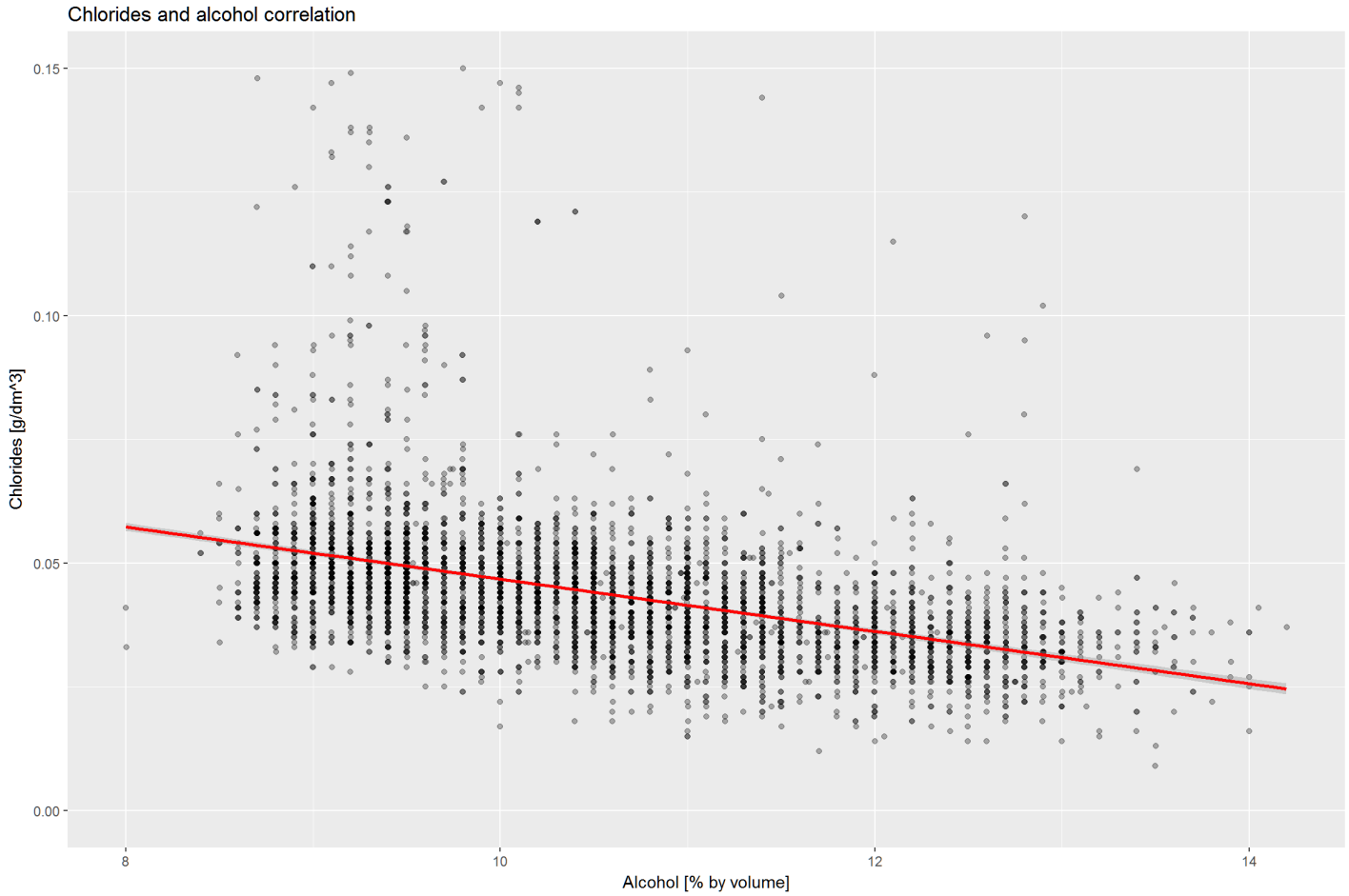




Let's have a quick look at the frequency distribution of density facceted by wine strenght (we have already observed in above plot that density has strong correlation with the alcohol). As in previous plots I will remove 1% of top and bottom values for both x and y axis for the density variable.

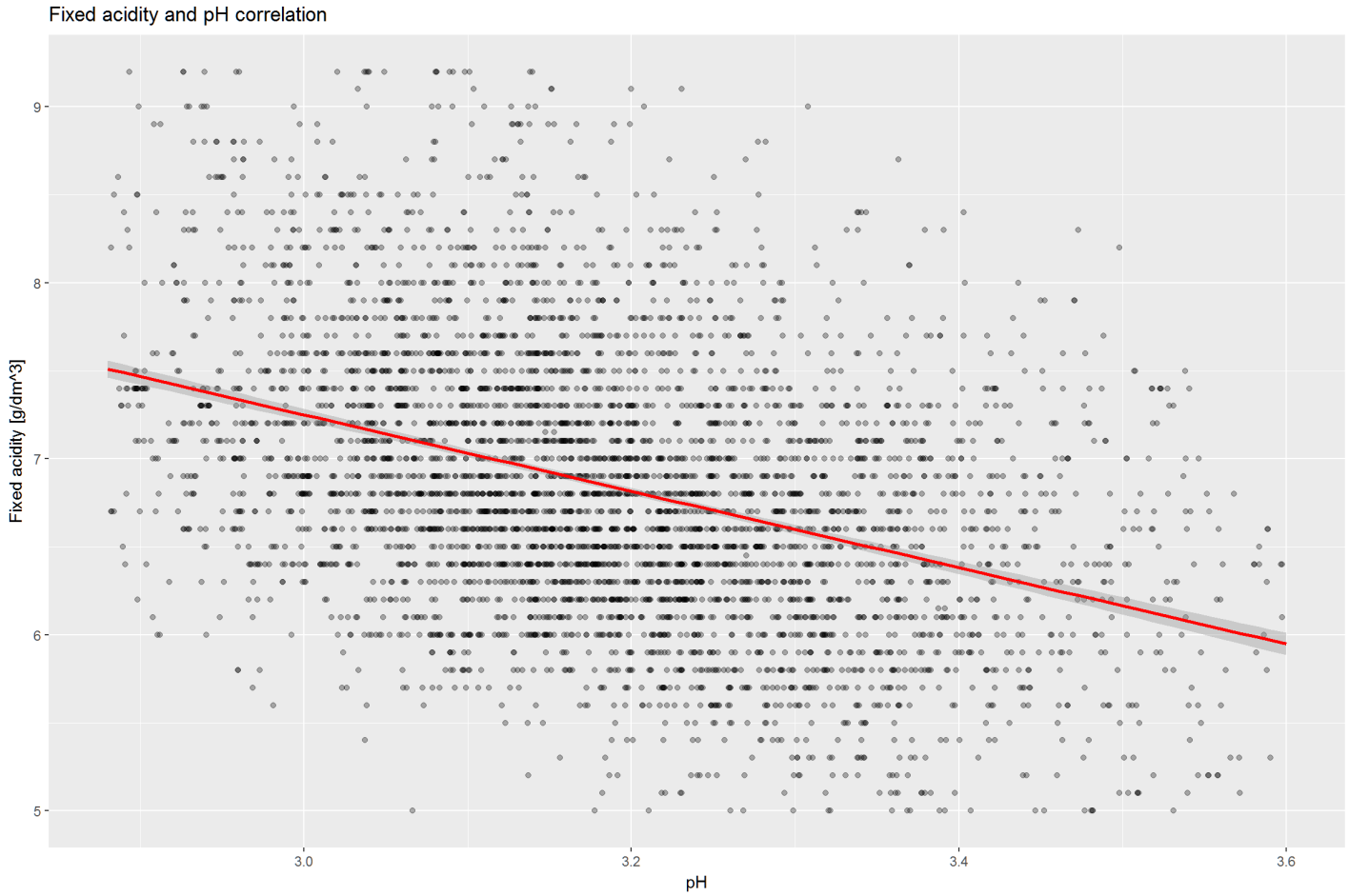


5. Chlorides and alcohol correlation (-0.36)



8. Correlation between pH and fixed acidity (-0.43)

Now i will create scatter plot of fixed acidity and pH. I will remove 1% of top and bottom values for both x and y axis to get better feel for the correlation between the variables.



We can observe that as fixed acidity increases, the pH value becomes more acidic.

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?

As anticipated there is a strong correlation is between quality and alcohol. There is a slight upward trend that can be observed on the plot once jitter is added. This also confirmed by alcohol boxplot - for wines that have quality equal or higher than 5 the median and quartiles increase. This shows that alcohol is significant factor for determining wine quality. Density has slight downward trend which means that the higher the quality of wine the lower the density. There is a relatively weak correlation between quality and chlorides and quality and sulfur dioxide. Residual sugar correlates very well with the density. With increase of residual sugar the density inceases. Moreover, I have observed that the stronger the wine is the less residual sugar it contains. Weak correlation was observed between residual sugar and bound/total sulfur dioxide. Density has very strong correlaton with alcohol - with increase of strenght of the wine the density decreases. Total sulfur dioxide and Bound sulfur dioxide content in the wine increases with increase of wine density. Alcohol is well correlated with chlorides - with increase of alcohol content the chlorides decrease so downward trend can be observed.

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

I have observed relationship between residual sugar, alcohol, density (described above). I also determined relationship between chlorides and alcohol. It might be worth to explore combinations of those variables in multivariate analysis.

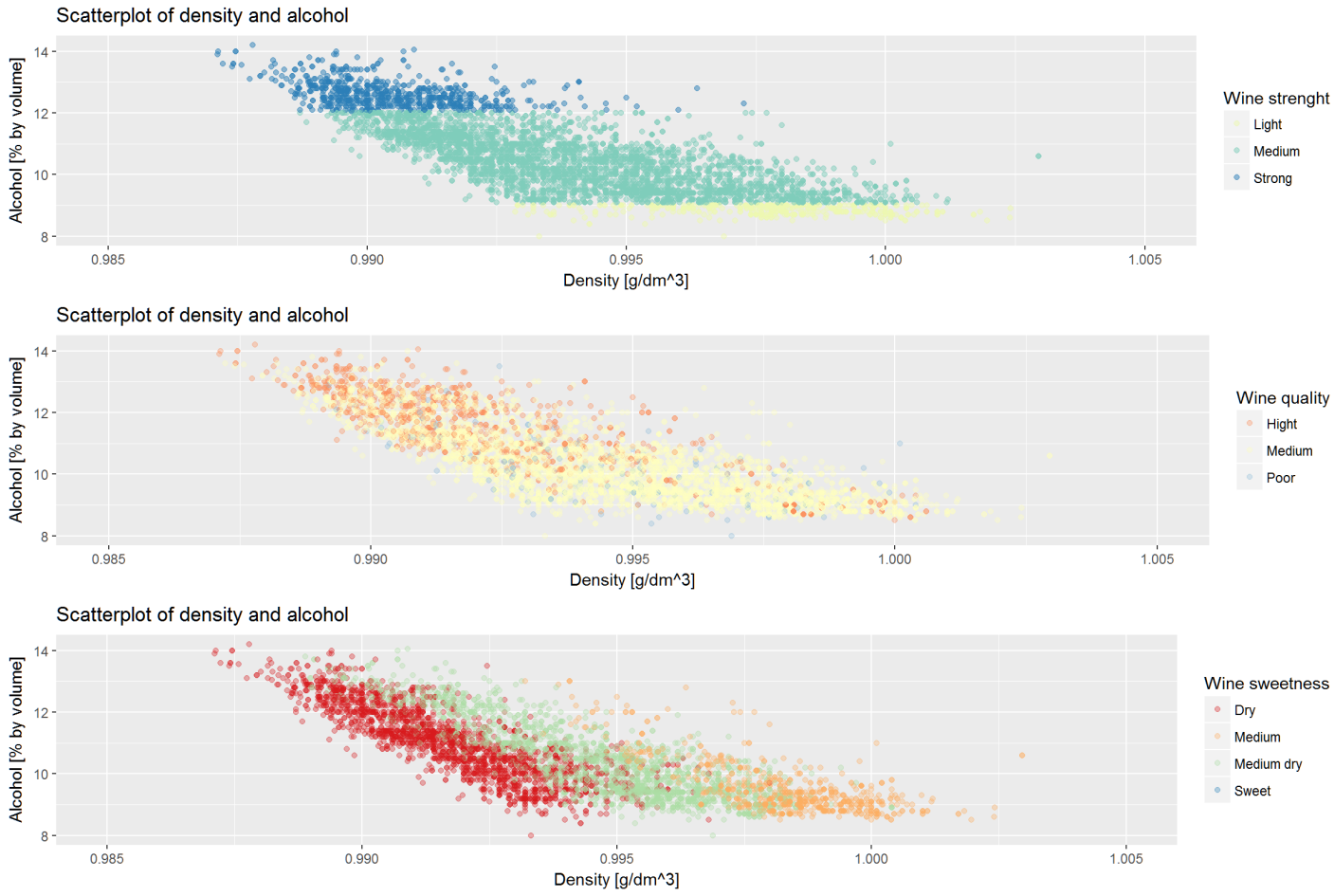
What was the strongest relationship you found?

The strongest relationships I found was between total and bound sulfur dioxide, because the correlation was 0.922 and between fixed acidity and total acidity (0.99). This is not a suprizing finding as they are derived form each other so strong correlation was expected. The the strongest relationships was observed also between density and residual sugar and density and alcohol.

Multivariate Plots Section

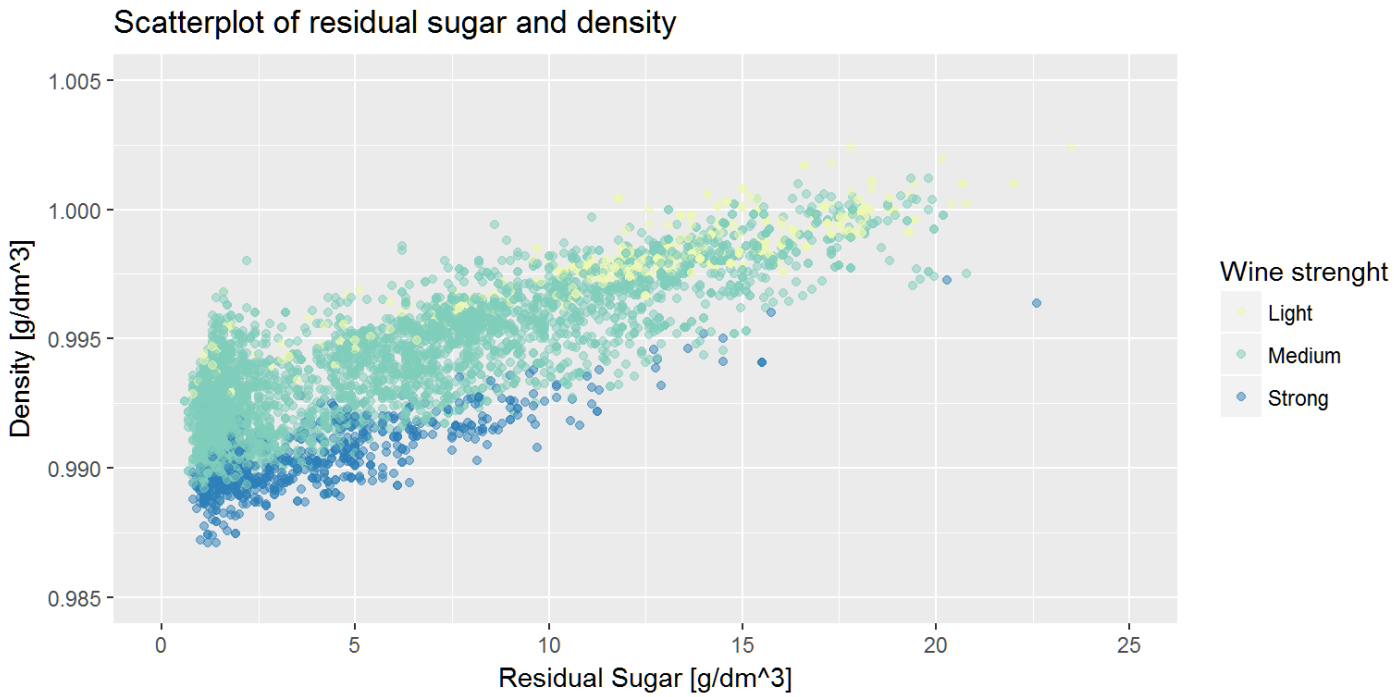
I want to use multivariate analysis methods in order to evaluate the relationship between density, alcohol, and residual sugar and chlorides. I will use the previously created categorical variables (wine strenght, wine quality score and wine sweetness) for further analysis. I will try to plot 3 variables on one plot to find connection between various wine characteristics.

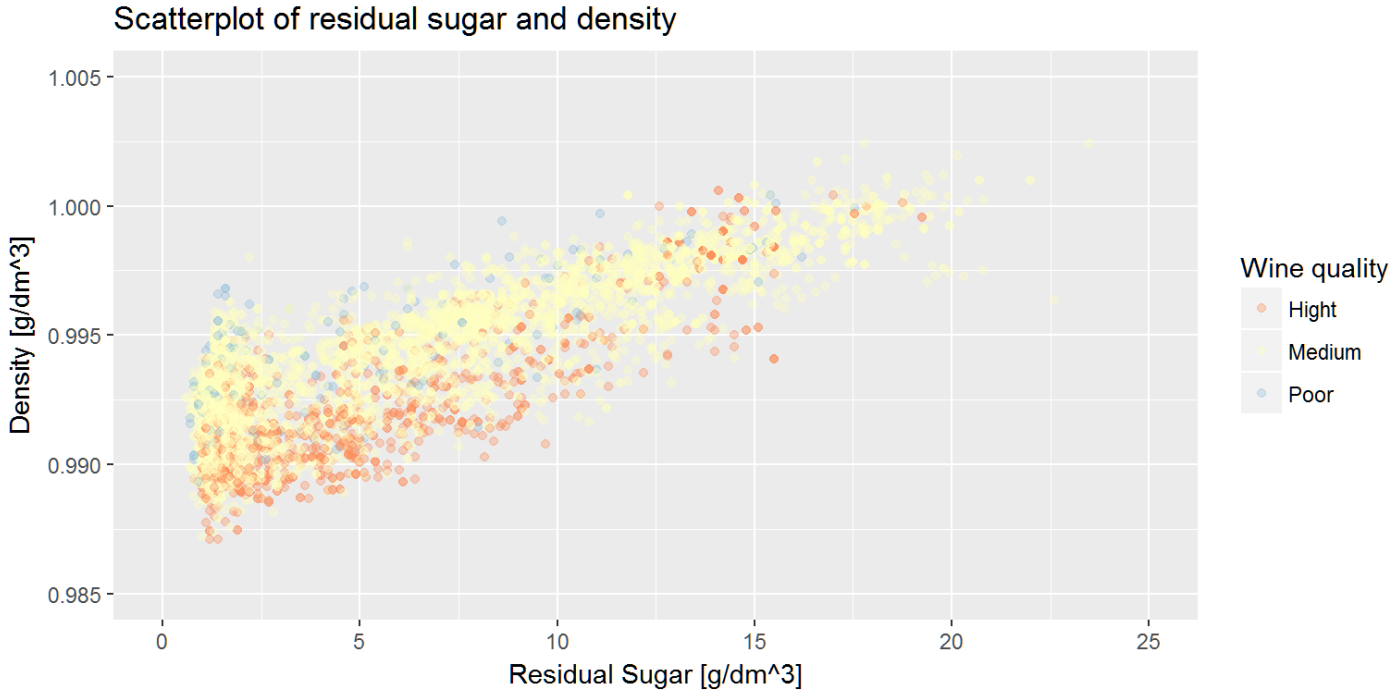
Density versus alcohol



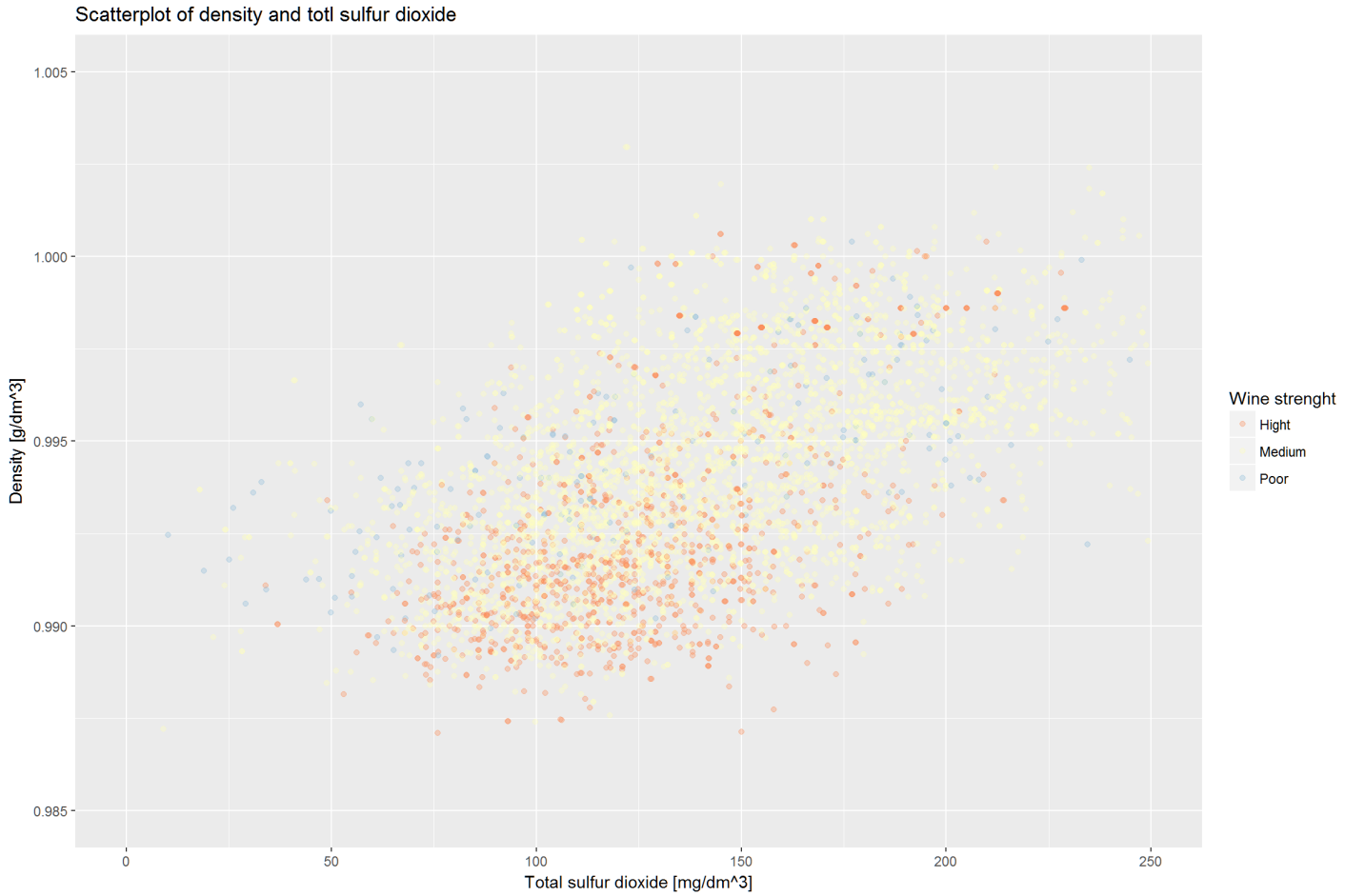
On the first plot we can observe that density of the wine decreases with increase of the alcohol. The wines with high (>12%) content of alcohol have lower density and they correspond to high quality wines. Those wines are either dry or medium dry (wine sweetnes). Wine sweetnes reflect content of residual sugar. It is a bit suprising that high quality wines are dry and medium dry. Medium sweetness wines have medium quality wine score.

Residual sugar and density/Alcohol

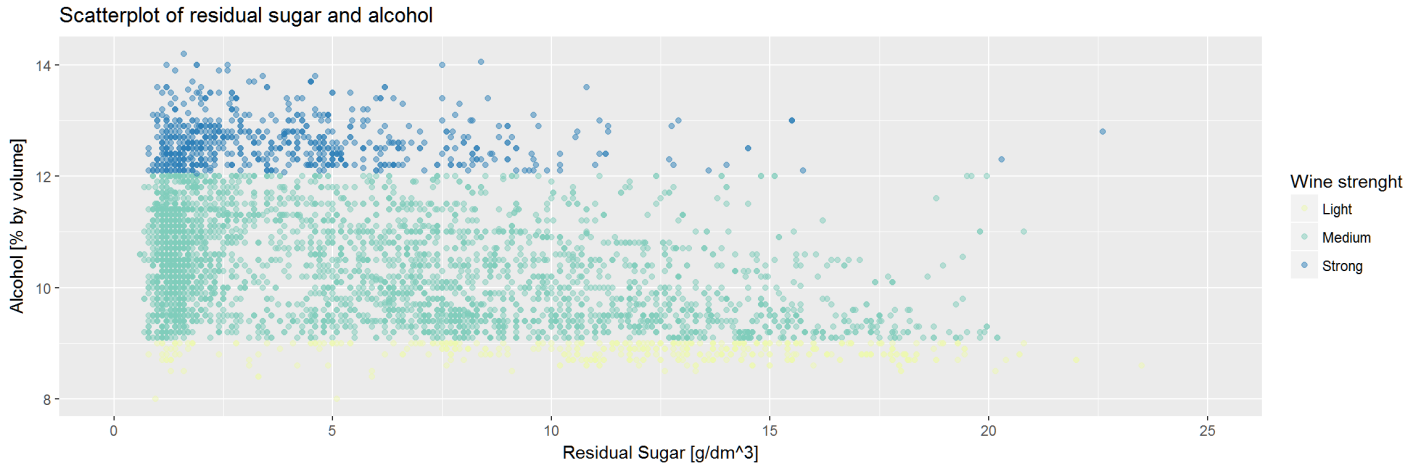


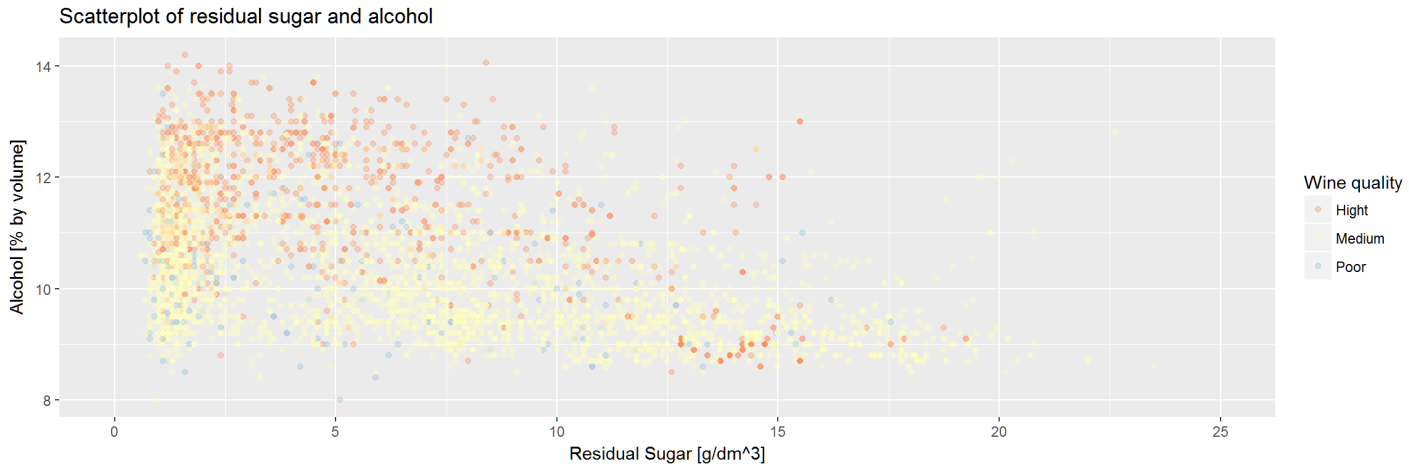


We can observe that density increases with the increase of residual sugar. Strong wines have content ff residual sugar between aprpox. 0.6 and 10 g/dm^3. Those strong wines correspond to high quality wines.



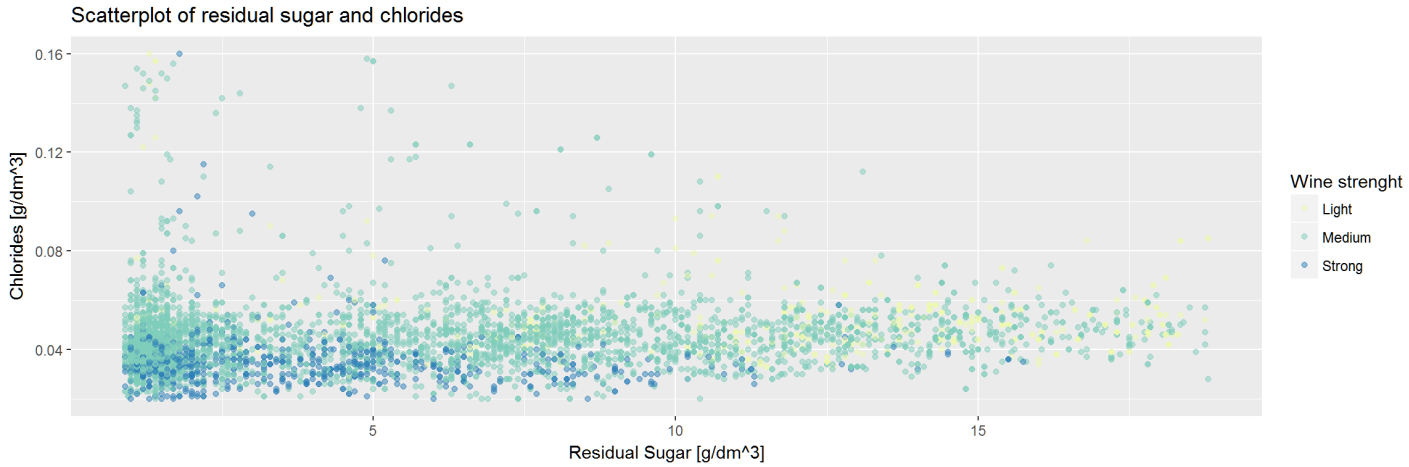
It seems that high qyality wines also have total sulfur dioxide content between 75 and 175.

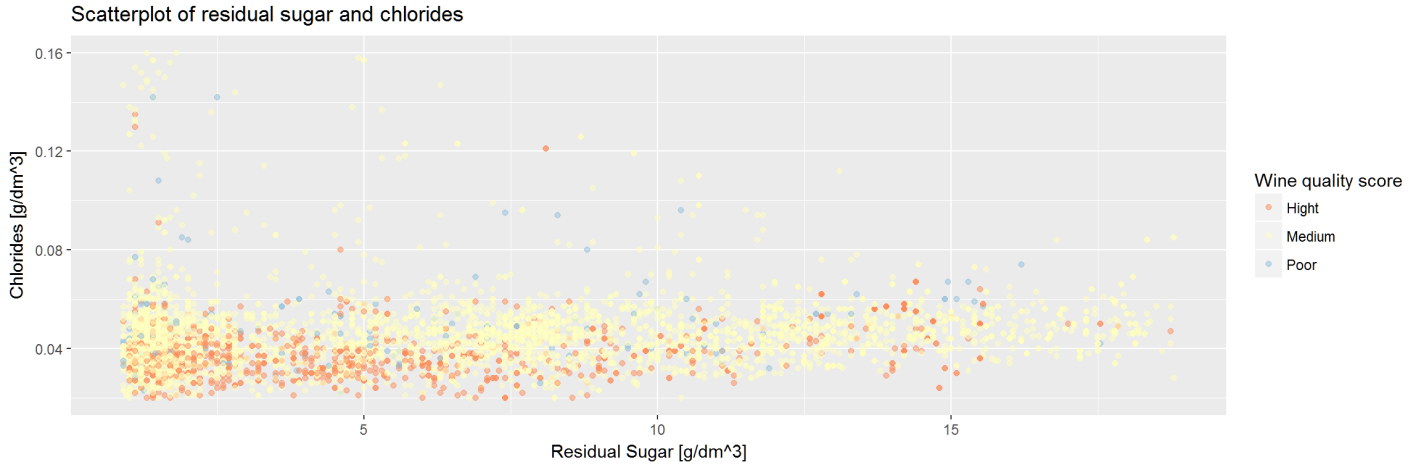




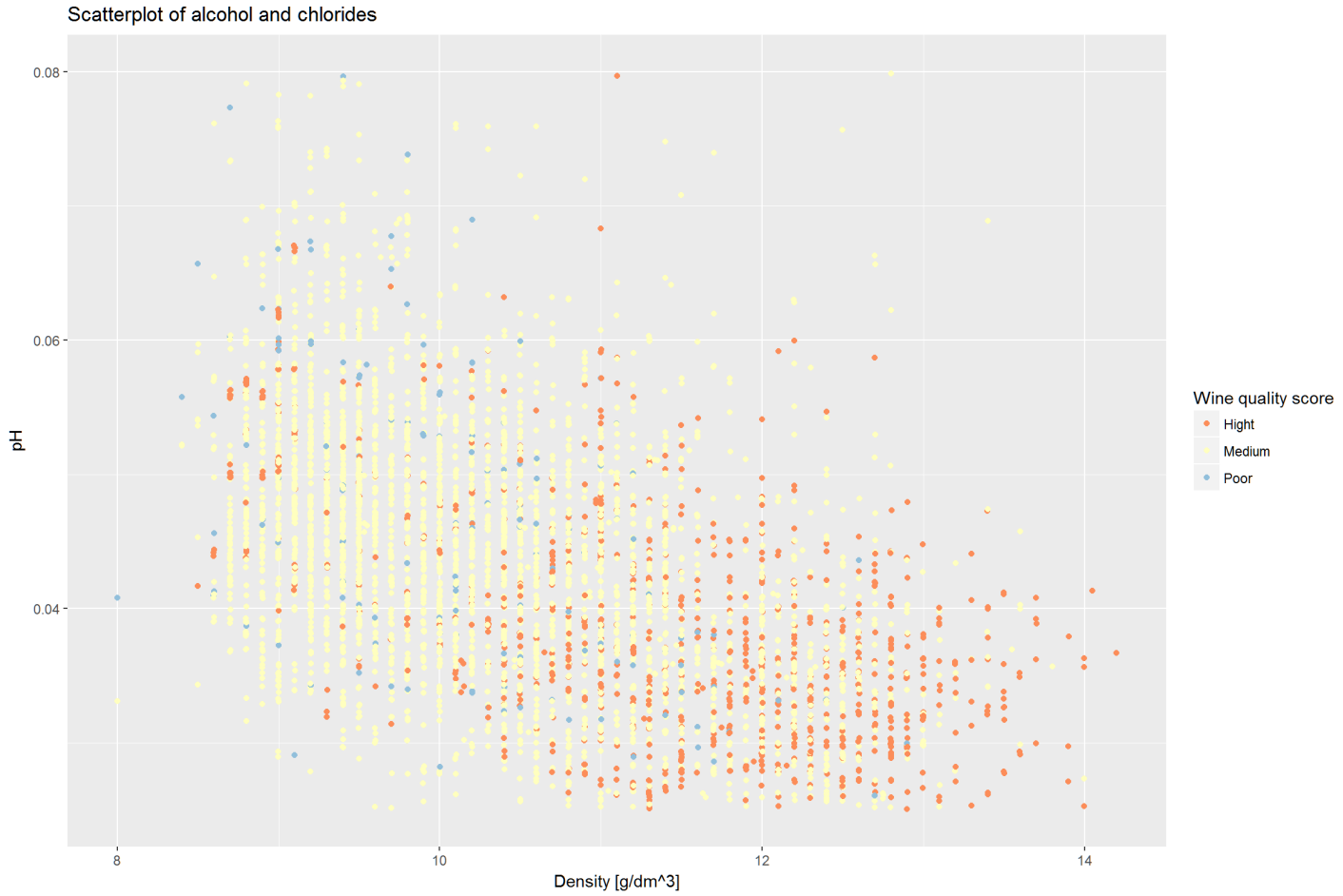
Residual sugar and chlorides

Now i will create scatter plots of residual sugar and chlorides.I will remove 1% of top and bottom values for both x and y axis to get better feel for the correlation between the variables. Moreover, in the first plot the values will be colored by the wine strenght and the second by the wine quality score.

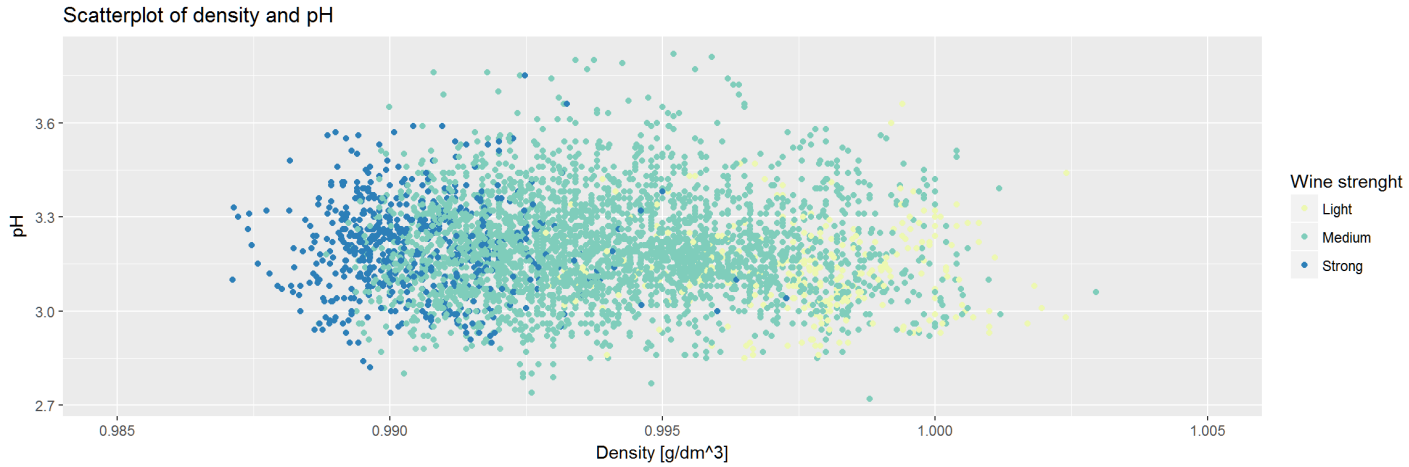


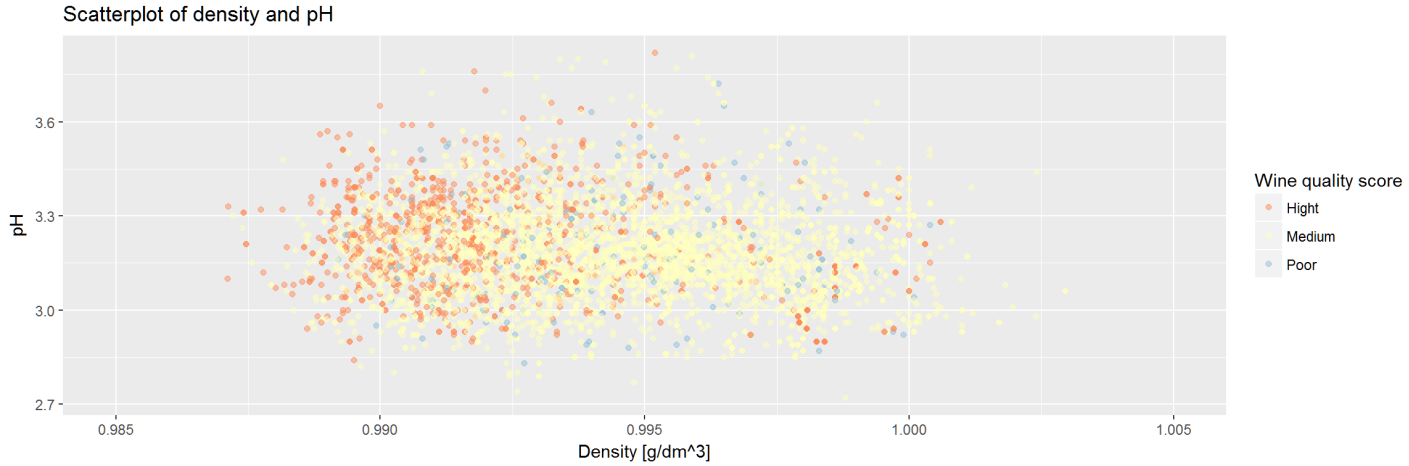


Alcohol and chlorides



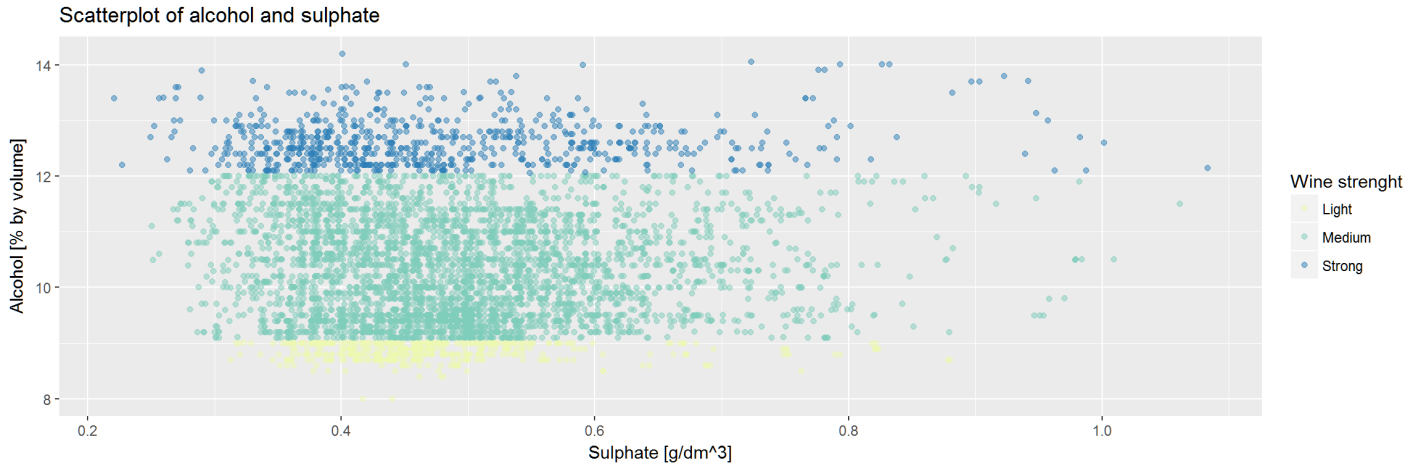
There is positive correlation between chlorides and alcohol. From above displayed scatterplot we can observe that high quality wines have in general high content of alcohol lower content of chloride. I believe that chlorides also affect wine quality.

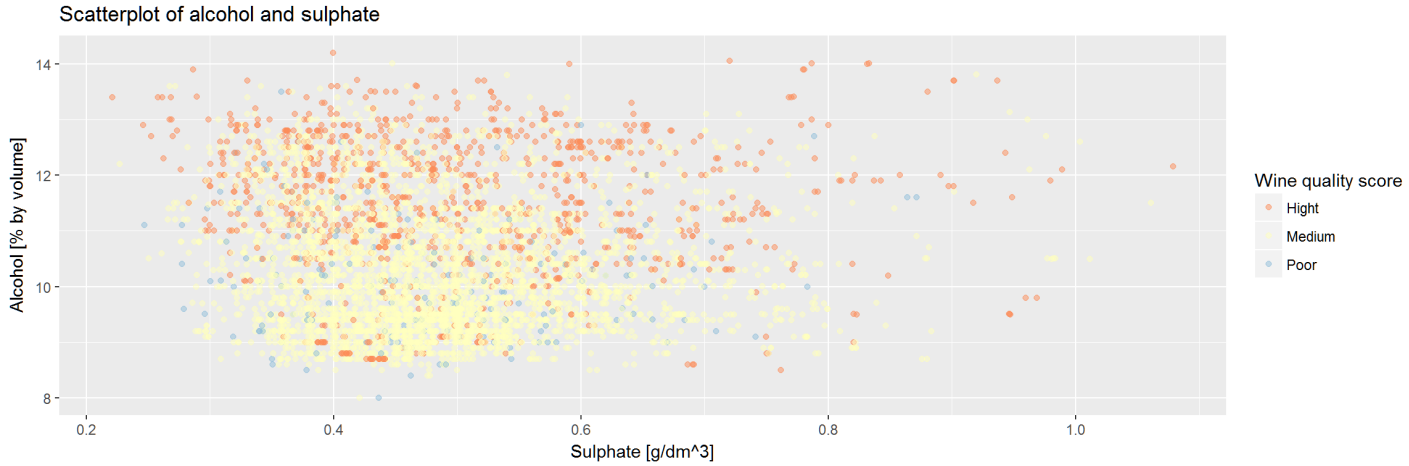




pH by itself it's not a good characteristing for determining high quality wines since high, medium and poor quality wines have pH between approx. 2.7 and 3.9.

Sulphates and alcohol





Looking at above displayed plots i can conclude that sulphate content has not so much influence on wine quality. Sulfates values have no correlation with the wine quality. Taking this into the consideration, i will not explore sulphate variable any further.

Now i will create a model with lm function

```
##
## Calls:
## m1: lm(formula = quality ~ alcohol, data = df_ww)
## m2: lm(formula = quality ~ alcohol + density, data = df_ww)
## m3: lm(formula = quality ~ alcohol + density + chlorides, data = df_ww)
## m4: lm(formula = quality ~ alcohol + density + chlorides + log(residual.sugar),
##     data = df_ww)
## m5: lm(formula = quality ~ alcohol + density + chlorides + log(residual.sugar) +
##     fixed.acidity, data = df_ww)
## m6: lm(formula = quality ~ alcohol + density + chlorides + log(residual.sugar) +
##     fixed.acidity + citric.acid, data = df_ww)
## m7: lm(formula = quality ~ alcohol + density + chlorides + log(residual.sugar) +
##     fixed.acidity + citric.acid + volatile.acidity, data = df_ww)
## m8: lm(formula = quality ~ alcohol + density + chlorides + log(residual.sugar) +
##     fixed.acidity + citric.acid + volatile.acidity + total.sulfur.dioxide,
##     data = df_ww)
## m9: lm(formula = quality ~ alcohol + density + chlorides + log(residual.sugar) +
##     fixed.acidity + citric.acid + volatile.acidity + total.sulfur.dioxide +
##     pH, data = df_ww)
## m10: lm(formula = quality ~ alcohol + density + chlorides + log(residual.sugar) +
##     fixed.acidity + citric.acid + volatile.acidity + total.sulfur.dioxide +
##     pH + sulphates, data = df_ww)
## m11: lm(formula = quality ~ alcohol + density + chlorides + log(residual.sugar) +
##     fixed.acidity + citric.acid + volatile.acidity + total.sulfur.dioxide +
##     pH + sulphates + free.sulfur.dioxide, data = df_ww)
## m12: lm(formula = quality ~ alcohol + density + chlorides + log(residual.sugar) +
##     fixed.acidity + citric.acid + volatile.acidity + total.sulfur.dioxide +
##     pH + sulphates + free.sulfur.dioxide + bound.sulfur.dioxide,
##     data = df_ww)
##
## ================================================================================================
=============================================================================
##                             m1          m2          m3          m4          m5          m6
      m7          m8          m9         m10         m11         m12
## ------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------
##   (Intercept)            2.582***  -22.492***  -21.150***   48.139***   33.193**    36.209***
   23.650*     25.924*     42.639***   52.595***   47.757***   47.757***
##                         (0.098)     (6.165)     (6.162)     (9.703)    (10.587)    (10.603)
  (10.255)    (10.343)    (11.284)    (11.432)    (11.437)    (11.437)
##   alcohol                0.313***    0.360***    0.343***    0.272***    0.289***    0.284***
    0.331***    0.333***    0.308***    0.294***    0.296***    0.296***
##                         (0.009)     (0.015)     (0.015)     (0.017)     (0.018)     (0.018)
   (0.017)     (0.017)     (0.019)     (0.019)     (0.019)     (0.019)
##   density                           24.728***   23.671***  -45.593***  -30.337**   -33.350**
  -20.603*    -22.981*    -40.902***  -50.976***  -46.226***  -46.226***
##                                     (6.079)     (6.074)     (9.646)    (10.571)    (10.587)
  (10.240)    (10.337)    (11.414)    (11.563)    (11.567)    (11.567)
##   chlorides                                    -2.382***   -2.003***   -2.074***   -2.292***
   -0.960      -1.000      -0.808      -0.796      -0.818      -0.818
##                                               (0.558)     (0.554)     (0.554)     (0.556)
   (0.542)     (0.542)     (0.544)     (0.543)     (0.541)     (0.541)
##   log(residual.sugar)                                      0.196***    0.171***    0.174***
```

```
    0.189***    0.186***    0.226***    0.248***    0.232***    0.232***
##                                                          (0.021)     (0.022)     (0.022)
   (0.022)     (0.022)     (0.024)     (0.025)     (0.025)     (0.025)
##   fixed.acidity                                                      -0.052***   -0.065***
   -0.064***   -0.062***   -0.029      -0.022      -0.014      -0.014
##                                                                     (0.015)     (0.015)
   (0.015)     (0.015)     (0.017)     (0.017)     (0.017)     (0.017)
##   citric.acid                                                                    0.368***
    0.056       0.043       0.075       0.055       0.037       0.037
##                                                                                 (0.098)
   (0.096)     (0.096)     (0.097)     (0.096)     (0.096)     (0.096)
##   volatile.acidity
   -2.113***   -2.132***   -2.101***   -2.076***   -1.953***   -1.953***
##
   (0.111)     (0.112)     (0.112)     (0.112)     (0.114)     (0.114)
##   total.sulfur.dioxide
               0.001       0.000       0.000      -0.001*     -0.001*
##
              (0.000)     (0.000)     (0.000)     (0.000)     (0.000)
##   pH
                           0.332***    0.306***    0.317***    0.317***
##
                          (0.090)     (0.090)     (0.090)     (0.090)
##   sulphates
                                       0.493***    0.502***    0.502***
##
                                      (0.099)     (0.099)     (0.099)
##   free.sulfur.dioxide
                                                   0.004***    0.004***
##
                                                  (0.001)     (0.001)
## ---------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
##   R-squared                  0.2         0.2         0.2         0.2         0.2         0.2
      0.3         0.3         0.3         0.3         0.3         0.3
##   adj. R-squared             0.2         0.2         0.2         0.2         0.2         0.2
      0.3         0.3         0.3         0.3         0.3         0.3
##   sigma                      0.8         0.8         0.8         0.8         0.8         0.8
      0.8         0.8         0.8         0.8         0.8         0.8
##   F                       1146.4      583.3       396.3       323.4       261.8       221.1
    255.1       223.6       200.8       184.1       170.8       170.8
##   p                          0.0         0.0         0.0         0.0         0.0         0.0
      0.0         0.0         0.0         0.0         0.0         0.0
##   Log-likelihood         -5839.4     -5831.1     -5822.0     -5780.1     -5773.9     -5766.8
   -5592.4     -5591.0     -5584.2     -5571.7     -5557.8     -5557.8
##   Deviance                3112.3      3101.8      3090.2      3037.8      3030.2      3021.4
    2813.6      2812.0      2804.3      2790.0      2774.2      2774.2
##   AIC                    11684.8     11670.3     11654.0     11572.1     11561.8     11549.7
   11202.7     11201.9     11190.4     11167.5     11141.6     11141.6
##   BIC                    11704.3     11696.2     11686.5     11611.1     11607.3     11601.7
   11261.2     11266.9     11261.8     11245.4     11226.1     11226.1
##   N                        4898        4898        4898        4898        4898        4898
    4898        4898        4898        4898        4898        4898
```

```
## ============================================================================================
================================================================================
```

The result of of my model did not result in a good prediction for wite wine quality. The max. R value was 0.3. Possibly the linear model is not the best choice for white wine dataset.

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 the multivariate part of the project i explored three variables in one plot which helped me to get better insight into the data. I managed to confirm strong correlation between alcohol, density and residual sugar. Especially, the alcohol content is strongly correlated with the wine quality - as the alcohol content increases, the quality increases as well. The maximum level of the acohole in our white wine dataset is approx. 14% which according to my reaserch is maximum alcohol level for white wines. I suspect that if the white wine would have alcohol content above that the quality would drop but I do not have data to back up my hypothesis. I have also observed (which confirmed previous expectations) that density of white wine decreases with increasing quality. As we know the wine sweetnes reflect content of residual sugar. In my investigation i have observed that residual sugar increases with the increase of wine density.I have also observed that high quality wines have in general high content of alcohol lower content of chloride. I believe that chlorides also affect wine quality.

Analyzing the data i can come up the following main conclusion:

- Mostly frequent quality value of white wine are 5 and 6.
- Alcohol is a main factor correlated to the wine quality. The data strongly suggest that the higher the alcohol content the better wine quality. Wine with high alcohol percentage has quality level 7, wine with less alcohol percentage is quality level 5.
- When alcohol percentage increases, density decreases.
- The residual suger content increases with increase of density and decrease of alcohol content.
- Total sulfur dioxide and level of residual sugar are positively correlated. Correlation shows higher value with white wine.
- White wine density and residual sugar level have positive correlation.
- High quality wines have high content of alcohol and are either dry or medium dry (wine sweetness).

I believe the quality score given by the jugdges is linked with their personnal perference or i could depend on other variables which were not provided in white wine dataset.

Were there any interesting or surprising interactions between features?

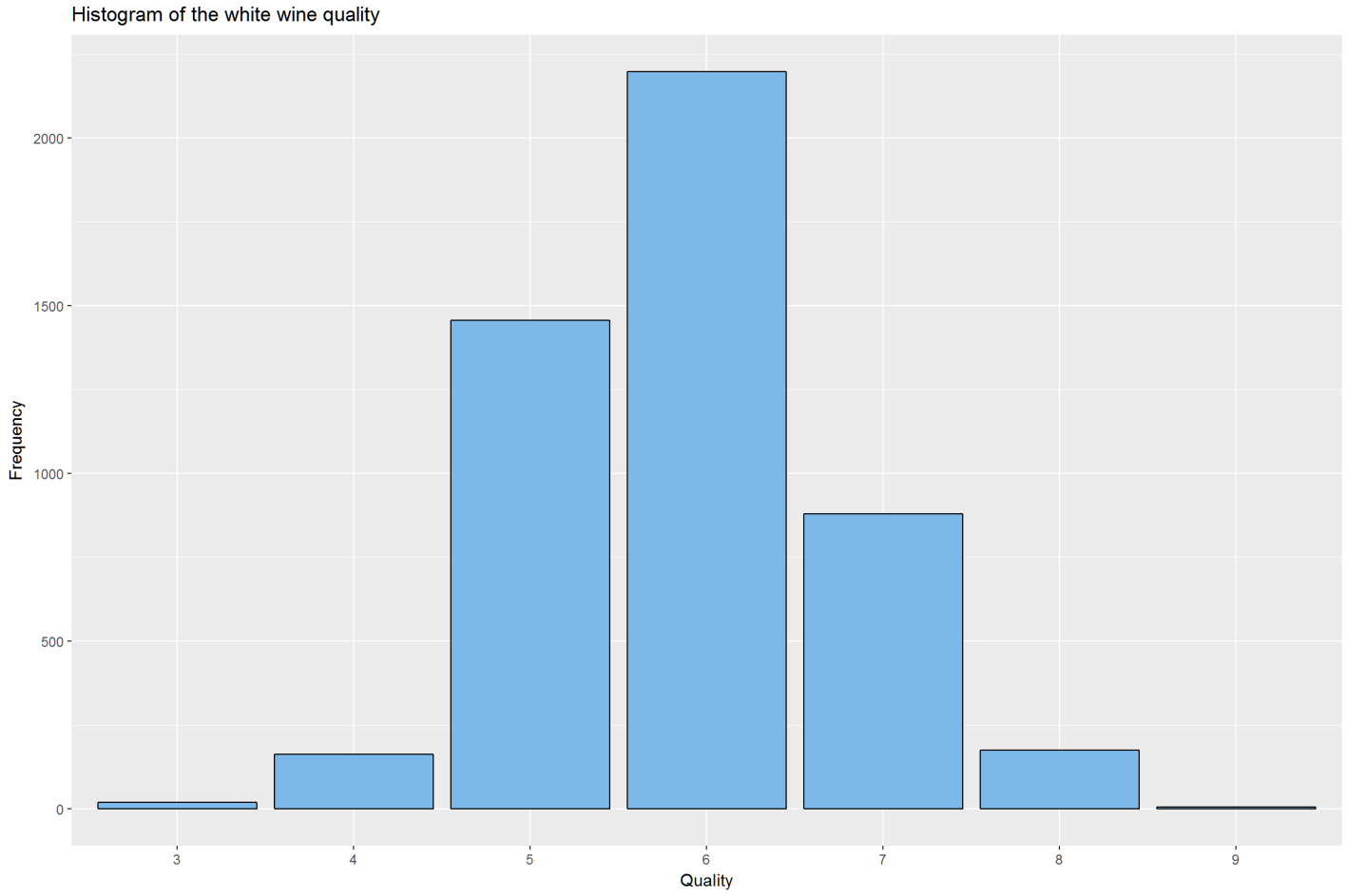
I also found the relationship between sugar, density, and alcohol to be interesting and I explored that in detail. The suprising thing for me was that the high quality wines can be characterized by two types of sweetness: medium dry and dry. Actually it’s also suprizing in the end that alcohol is very well correlated with wine quality. I would expect that not all the wines with high content of alcohole would be high quality wines. Since the wine quality jugment can be very personal it is possible that juges that gave the scores for the wine in this dataset enjoyed white wines with higher content of alcohol.

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

Yes, i did create liniar model. I wanted to predict which characteristic of the white wine have influence on it's quality. Unfortunatelly, the R value suggest that with the current data it is not possible to accurately predict white wine quality. I hope in future i would be able to create more complex models and possibly integrate more wine data.

Final Plots and Summary

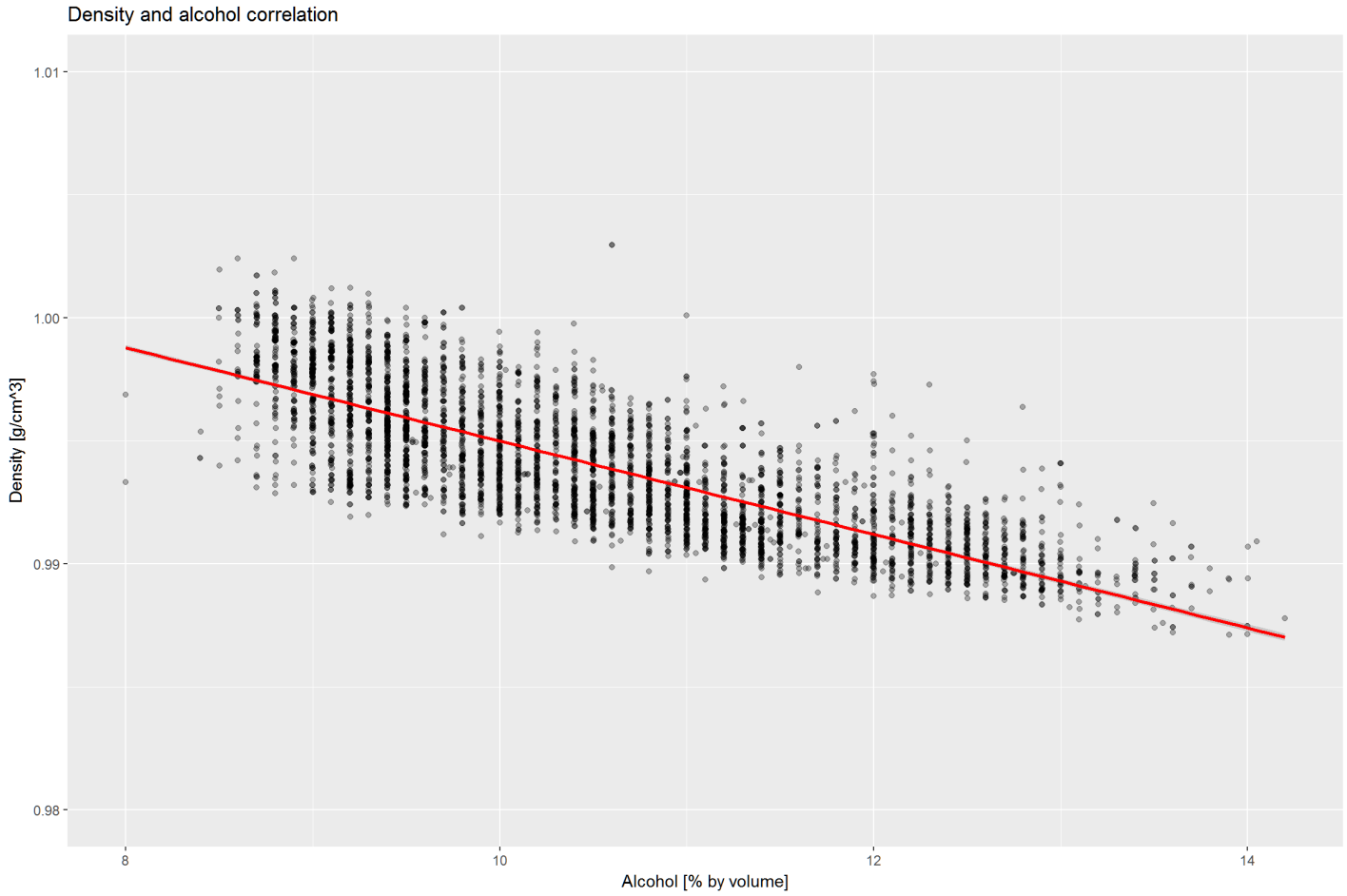
Plot One



Description One

Plot one shows the distribution of wine quality. It has many count for medium quality white wines(with quality score of 5, 6), but much fewer count on low (grade 3-4) and high (grade 7-9) quality wine. The quality variable is normally distributed.

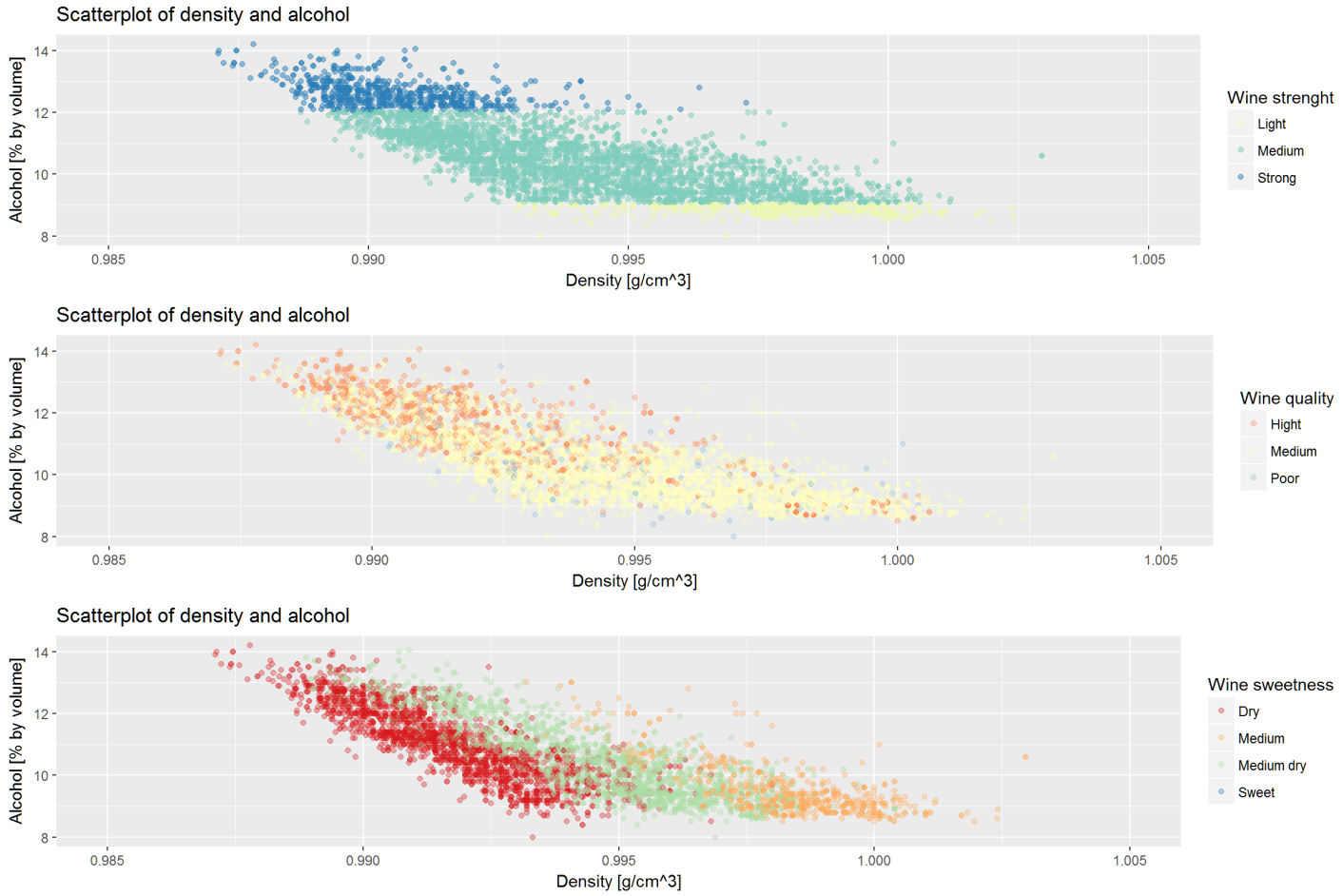
Plot Two



Density and Alcohol have the strongest correlation among all white wine parameters and it is equals to -0.78. Density would be one of the main parameters that influences wine quality.

Description Two

Plot Three



Description Three

This plot section i wanted to show the relationship between density, alcohol, and residual sugar content. For this plot i have removed top and bottom 1% of values for density variable (the same as was done in all the density plots before). This helped me to get better view of the correlation. We can observe that that density of the wine decreases with increase of the alcohol. The residual sugar (thus sweetness of the wine) increases with increase of density. Strong wines correspond to high quality wines. White wine density and residual sugar level have positive correlation. The strong wines can be characterized be either dtry or medium dry (wine sweetness). We can conclude that the best quality wines have high alcohol content (above 12%), low density and residual sugar content between approx 0.6 and 10 g/dm3. High quality wines have high content of alcohol and are either dry or medium dry (wine sweetness). I believe the quality score given by the jugdges is linked with their personnal perference or i could depend on other variables which were not provided in white wine dataset.

Reflection

In this project i have learnt how to perform exploratory data analyses using R with help of various libraries. I have used various advanced functions as ggpairs, corplot etc. I have learnt how to preform statistical analysis in R as well how to create histograms, scatterplots, boxplots. I have used various techniques for adjusting look of the graphs - jitter, transparency, alpha. ColorBrewer came in handy when i was adjusting color scale of my plots. At the end of my project i have used R markdown to create professional-looking report.

I found this project to be quite challenging. Since I was new to R it took me a lot of time to learn the syntax and basic code. I found R to be much easier to use than Python especially for the statistical analysis and plotting. I like creating plots in R since it's quite straightforward and possibilty of adding various layers in ggplot makes it more efficient - especially when you want to experiment with the layout your your graph, color code etc.

The most difficoult part of me from the whole project was multivariate section. I created a lot of graphs to find the correlations between various variables. This task seemed to be very tidious but necessary for achievement of my goal. Since my project was growing bigger and bigger it was necessary to go back and forth between univariate, bivariate and multivariate sections to make sure that the final conclusions were consistent. This was very time consuming.

This project made me quite interested in the wine quality. I would be very interested to do similar experiment either with red wine dataset or beer dataset. For future exploration of white wine dataset I would pick one category of wine (for example, quality level 5-7 or 8-9) to try to find more patterns. I would also like to get more white wine observation or maybe even include red white wines for comparision.

Resources

- http://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity (http://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity)
- https://en.wikipedia.org/wiki/Acids_in_wine (https://en.wikipedia.org/wiki/Acids_in_wine)
- https://winobrothers.com/2011/10/11/sulfur-dioxide-so2-in-wine/ (https://winobrothers.com/2011/10/11/sulfur-dioxide-so2-in-wine/)
- http://www.calwineries.com/learn/wine-chemistry/wine-acids/citric-acid (http://www.calwineries.com/learn/wine-chemistry/wine-acids/citric-acid)
- http://stackoverflow.com/questions/34594641/dplyr-summary-table-for-multiple-variables (http://stackoverflow.com/questions/34594641/dplyr-summary-table-for-multiple-variables)
- http://stackoverflow.com/questions/6286313/remove-an-entire-column-from-a-data-frame-in-r (http://stackoverflow.com/questions/6286313/remove-an-entire-column-from-a-data-frame-in-r)
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4120293/ (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4120293/)
- https://cran.r-project.org/web/packages/stargazer/vignettes/stargazer.pdf (https://cran.r-project.org/web/packages/stargazer/vignettes/stargazer.pdf)
- http://www.realsimple.com/holidays-entertaining/entertaining/food-drink/alcohol-content-wine (http://www.realsimple.com/holidays-entertaining/entertaining/food-drink/alcohol-content-wine)
- https://lembra.wordpress.com/2010/03/12/adding-new-column-to-a-data-frame-in-r/ (https://lembra.wordpress.com/2010/03/12/adding-new-column-to-a-data-frame-in-r/)
- https://www.r-statistics.com/2013/05/log-transformations-for-skewed-and-wide-distributions-from-practical-data-science-with-r/ (https://www.r-statistics.com/2013/05/log-transformations-for-skewed-and-wide-distributions-from-practical-data-science-with-r/)
- http://stackoverflow.com/questions/35085261/how-to-use-loess-method-in-ggallyggpairs-using-wrap-function (http://stackoverflow.com/questions/35085261/how-to-use-loess-method-in-ggallyggpairs-using-wrap-function)