Exploring Quality of White Wine

1. Outline

In this report, we use the dataset of white wines (white variants of the Portuguese "Vinho Verde" wine) available here and carry out exploratory data analysis. We will analyze what chemical properties influence the quality of the white wines. For some details of the dataset, please refer to Cortez et.al., Decision Support Systems Vol.47(4) 547-553 (2009). The structure of this report is as follows:

2. Univariate Plots Section

The goal of this part is to make some univariate plots from the dataset and grasp some profiles of the features in it.

2.1. Summary of Data

Before plotting the data, we start with basic summaries of the dataset (after cleaning as we will explain shortly) to be considered throughout this report.

```
## 'data.frame':    4898 obs. of  13 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             : Factor w/ 7 levels "3","4","5","6",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ quality3            : Factor w/ 3 levels "low","mid","high": 2 2 2 2 2 2 2 2 2 2 ...
```

```
##  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  quality3
##  3:  20   low :1640
##  4: 163   mid :2198
##  5:1457   high:1060
##  6:2198
##  7: 880
##  8: 175
##  9:   5
```

The original dataset contains 4898 observations and 13 features. We then carried out the following cleanings:

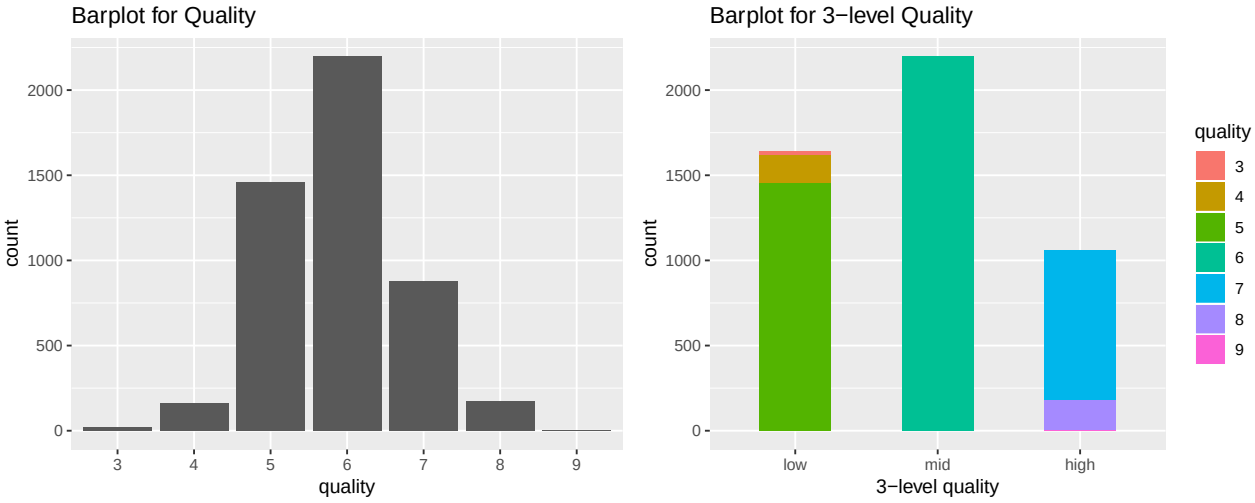
- We have removed the feature `X` in the original dataset since it just stores the indices for the observations (1,2,3, ..., 4898).
- We have converted the type of the feature `quality` to a factor (it was an integer originally), since in most of this report we will treat this as a factor. This `quality` can take a value in {0,1,2,3,4 ..., 10} by definition, but in the dataset, we can see the observations with `quality` = 3,4,5, ..., 9 only.
- The dataset contains only a small number of the observations corresponding to very high (`quality`>= 8) and very low (`quality`<= 4) quality wines. It is therefore useful to introduce a feature which classifies the quality of the wines into a smaller number of categories. For this purpose, we have added a new feature `quality3` which classifies the qualities of the wines into three categories (we call this classification by *3-level quality* here):
  - low: `quality` <= 5
  - mid: `quality` = 6
  - high: `quality` >= 7

Thus, in the end, there are 13(= 13 - 1 + 1) features in the dataset after the cleaning. No missing value can be seen in any observations in the dataset. We thus can use all the observations (= 4898 observations) for our exploratory data analysis (some outliers will be removed depending on which feature(s) we analyze, as we will explain later).

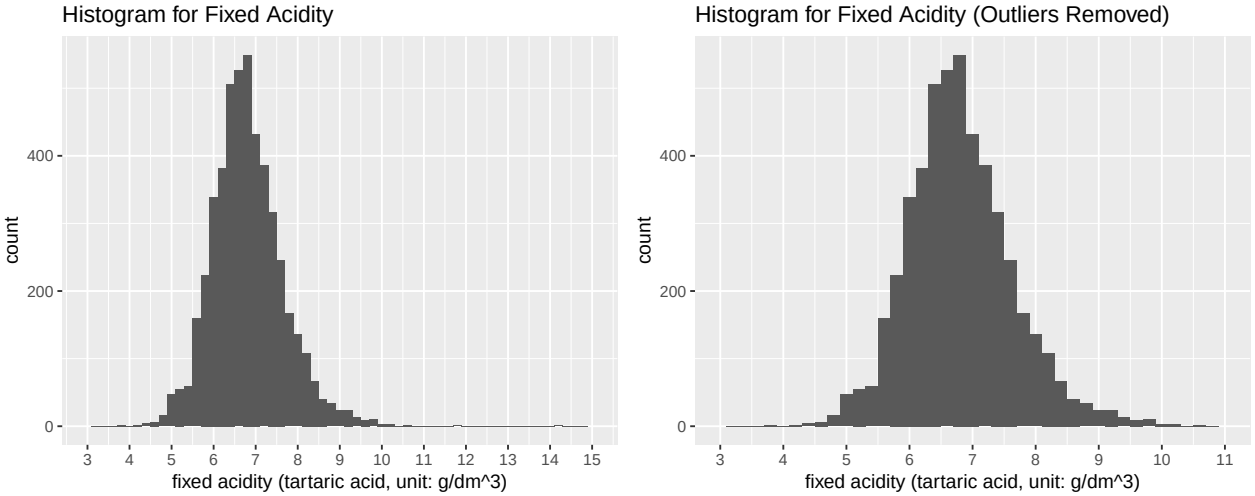
The goal of this report is to carry out the exploratory data analysis to understand how the categorical features (`quality` and `quality3`) are affected by the numerical ones (`fixed.acidity`, `volatile.acidity`, `citric.acid`, `residual.sugar`, `chlorides`, `free.sulfur.dioxide`, `total.sulfur.dioxide`, `density`, `pH`, `sulphates`, `alcohol`).

2.2. Univariate Plots

In this part, we provide histograms/bar plots for each feature to visually see the distribution of it.

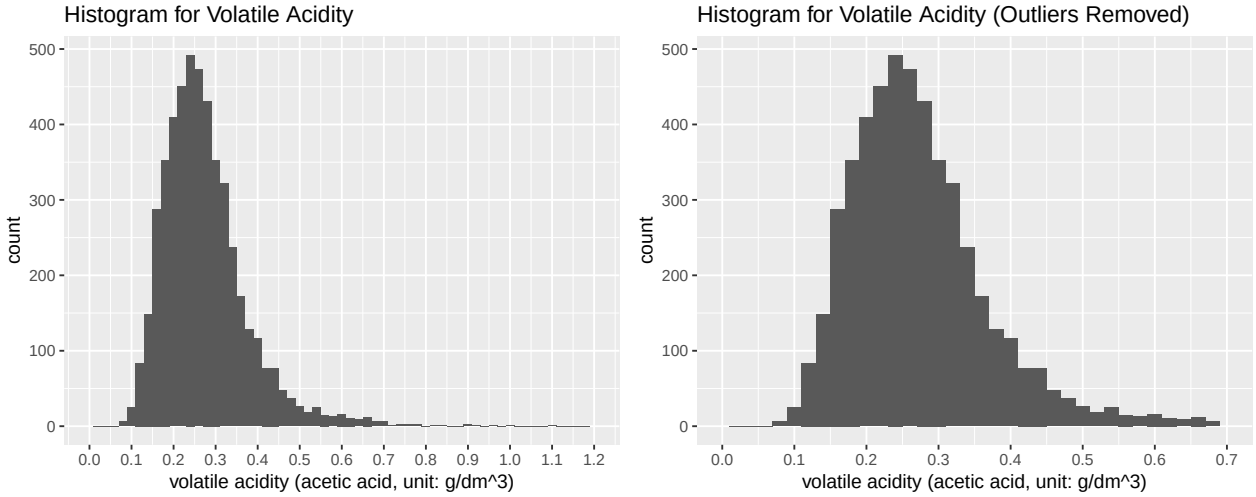


We note that `quality` can by definition take a value in $\{0, 1, 2, 3, \ldots, 10\}$ but, as can be seen in the left plot, the dataset does not contain any observations with `quality` $= 0, 1, 2, 10$. In the right plot, the data with different qualities are colored with different colors (exactly speaking, this is an bivariate plot, but we put it here for comparison with the barplot for `quality`). From these plots, we can see that the dataset contains only a small number of observations with very low (`quality` $<= 4$) or very high quality (`quality` $>= 8$) compared to those in between.



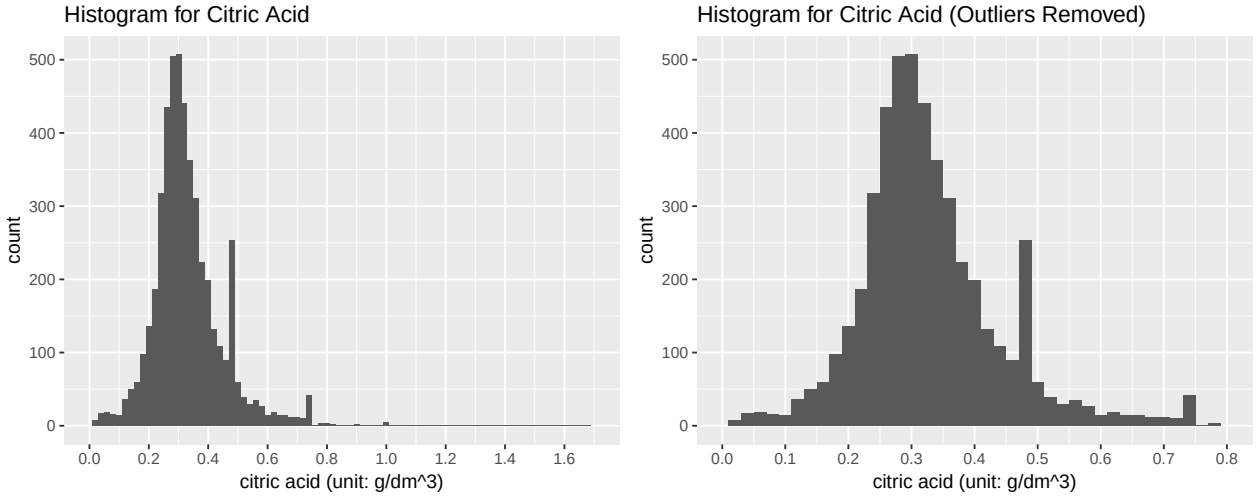

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.800   6.300   6.800   6.855   7.300  14.200
```

For the readers' convenience, the summary of the feature `fixed.acidity` is again displayed just below the plot. (We will also add this when we create the histograms of the other features below.) In the left plot, all the observations are used, while some outliers (`fixed.acidity` > 11) are removed in the right plot. The distribution is close to the normal distribution but a little bit right-skewed.



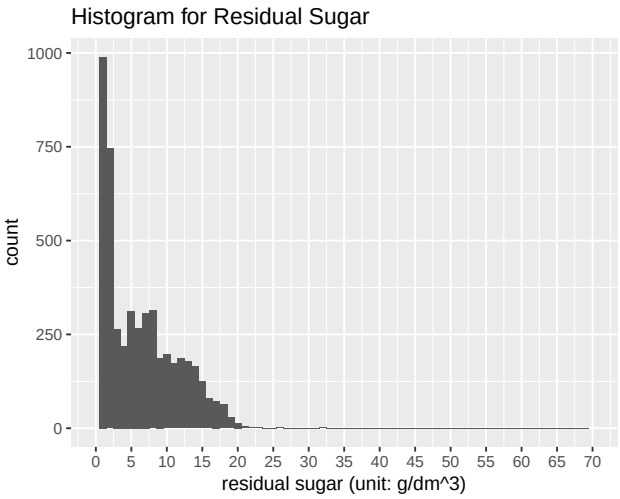

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.2100  0.2600  0.2782  0.3200  1.1000
```

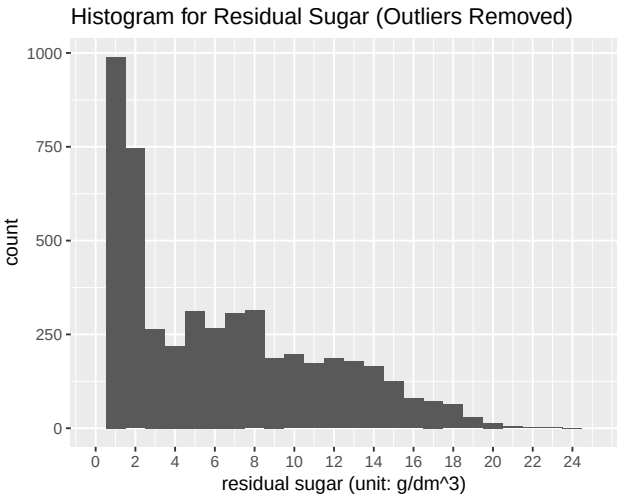
In the left plot, all the observations are used, while some outliers (`volatile.acidity` > 0.7) are removed in the right plot. The distribution is close to the normal distribution but right-skewed.

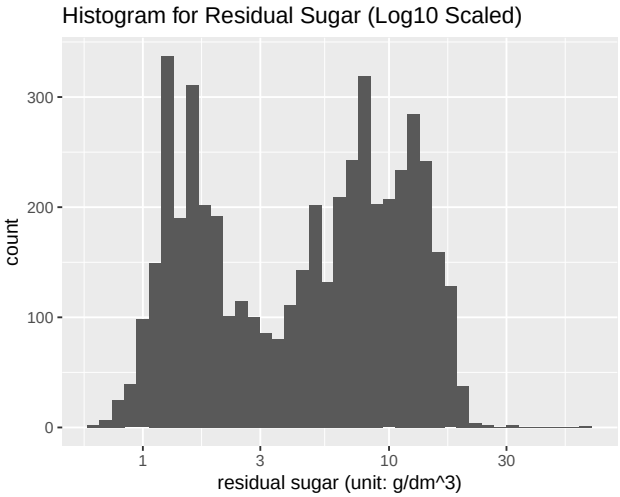



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2700  0.3200  0.3342  0.3900  1.6600
```

In the left plot, all the observations are used, while some outliers (`citric.acid` > 0.8) are removed in the right plot. The distribution is close to the normal distribution but a little bit right-skewed. We can also see two very localized peaks just below `citric.acid` $= 0.5, 0.75$.

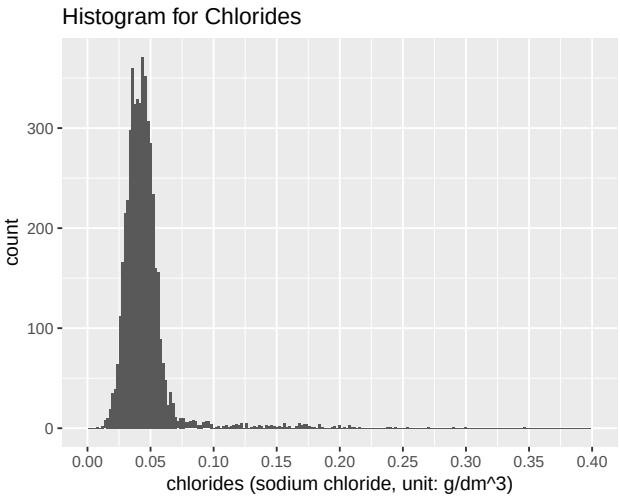


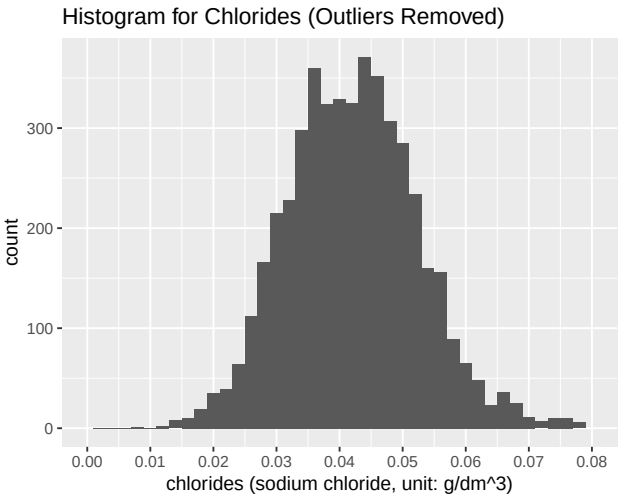





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.600   1.700   5.200   6.391   9.900  65.800
```

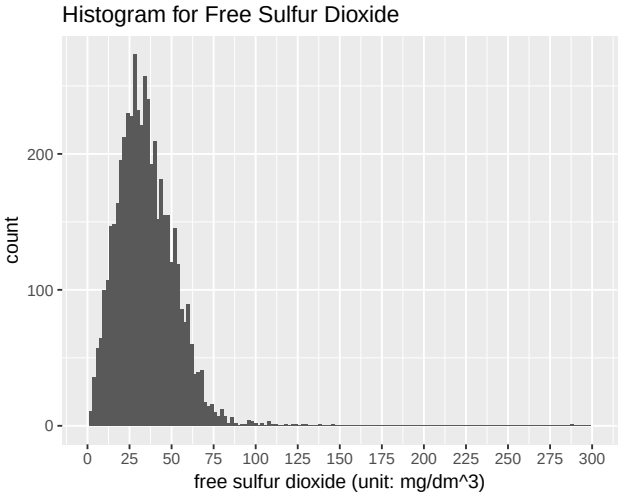
In the top-left plot, all the observations are used, while some outliers (`residual.sugar` > 25) are removed in the top-right plot. The distribution seems to be monotonically decreasing but we can also see a small peak around `residual.sugar` $= 8$. To see this more in detail, we set the x-axis to the log10 scale and plotted (all the observations used) again (bottom-left plot). We can now see two peaks with similar heights.

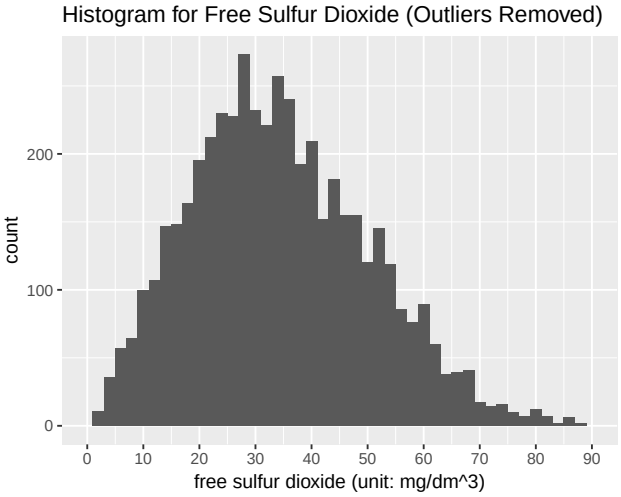





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00900 0.03600 0.04300 0.04577 0.05000 0.34600
```

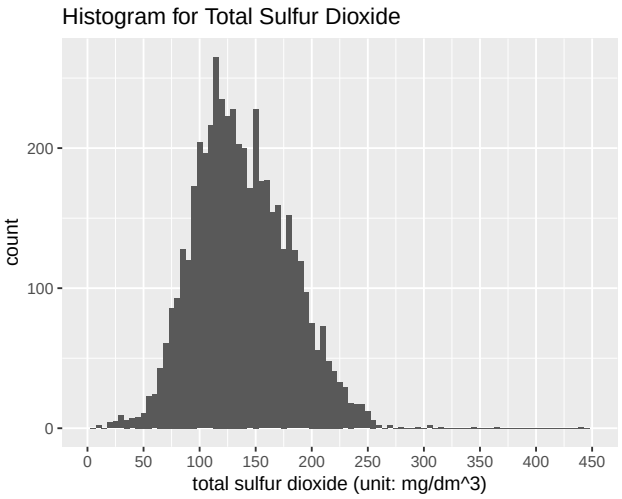
In the left plot, all the observations are used, while some outliers (`chlorides` > 0.08) are removed in the right plot. The distribution is close to the normal distribution.

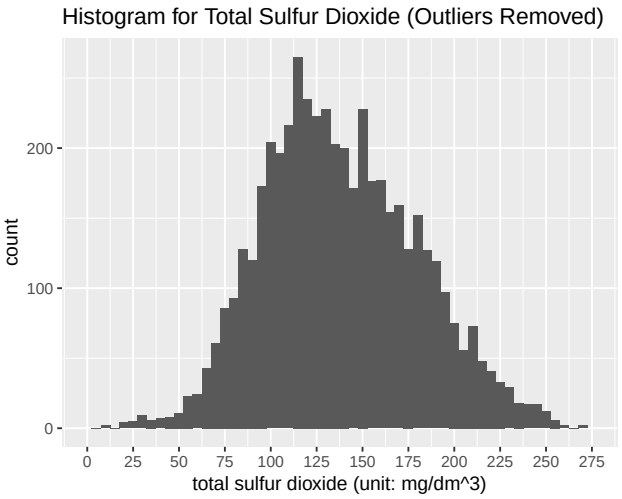





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   23.00   34.00   35.31   46.00  289.00
```

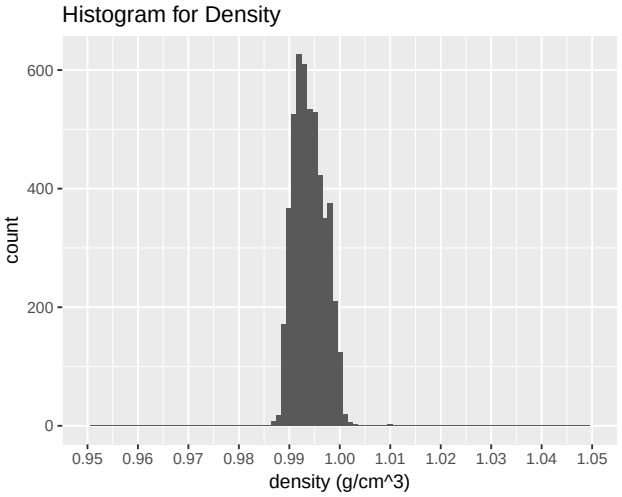
In the left plot, all the observations are used, while some outliers (`free.sulfur.dioxide` > 90) are removed in the right plot. The distribution is close to the normal distribution but a little bit right-skewed.

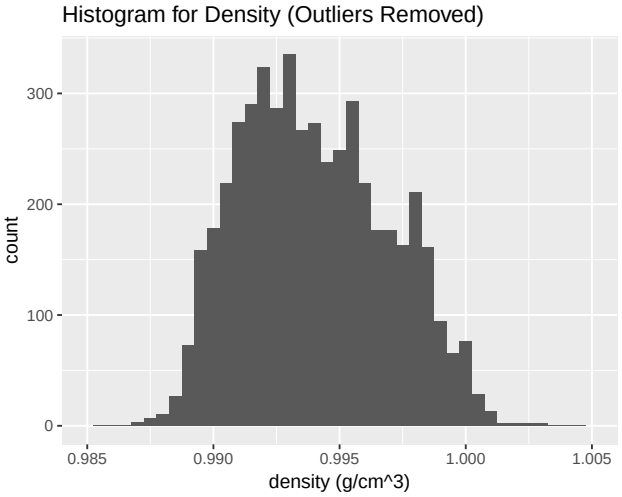





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.0   108.0   134.0   138.4   167.0   440.0
```

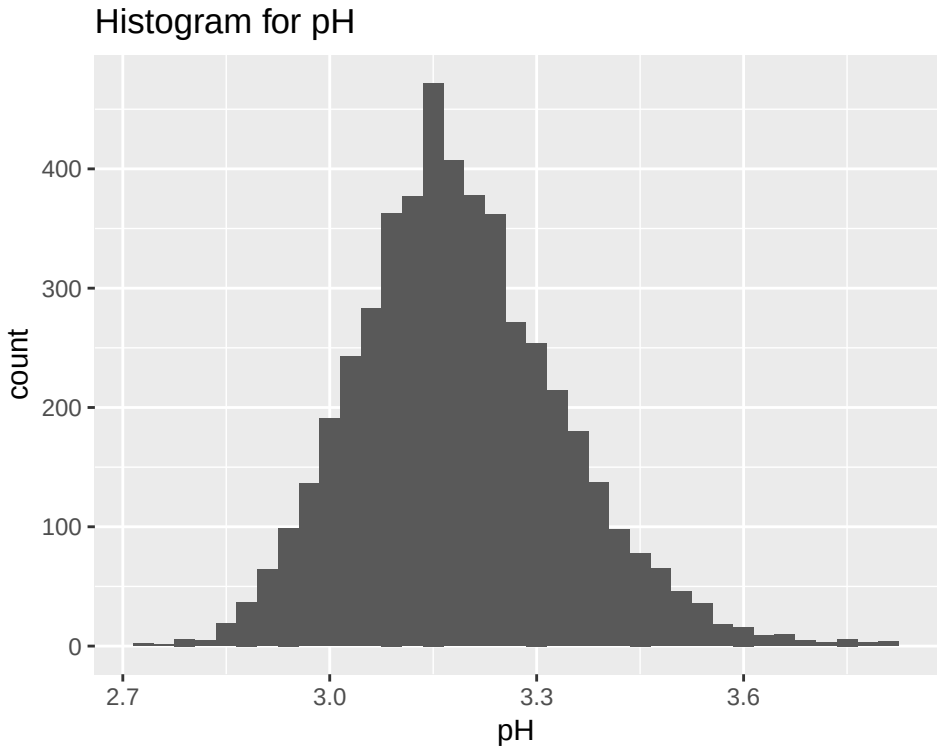
In the left plot, all the observations are used, while some outliers (`total.sulfur.dioxide` > 275) are removed in the right plot. The distribution is close to the normal distribution but a little bit right-skewed.





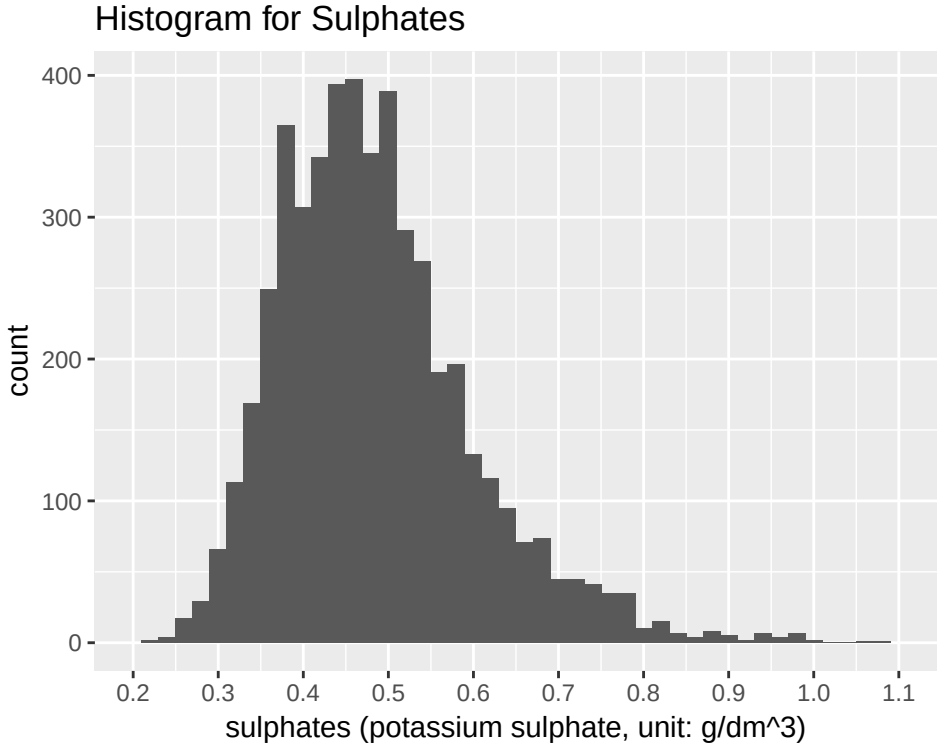

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9917  0.9937  0.9940  0.9961  1.0390
```

In the left plot, all the observations are used, while some outliers (`density` > 1.005) are removed in the right plot. The distribution is close to the normal distribution but a little bit right-skewed.



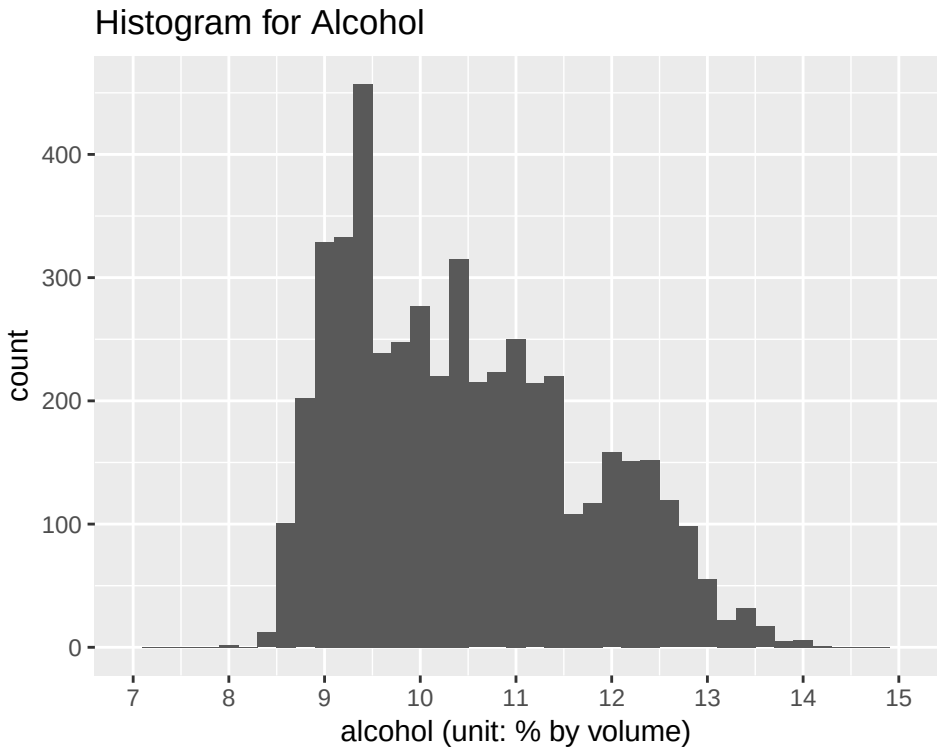

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.720   3.090   3.180   3.188   3.280   3.820
```

In this plot, all the observations are used. The distribution is close to the normal distribution.




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

In this plot, all the observations are used. The distribution is close to the normal distribution but a little bit right-skewed.




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

In this plot, all the observations are used. Roughly, we can see two peaks in the distribution.

2.3. Univariate Analysis

In this part, based on the univariate plots we made above, we carry out some analysis to understand the features in the dataset.

- **Structure of dataset** Number of Observations: 4898, Number of Features: 13 We note that (1) the feature `X` in the original dataset is removed, (2) a new feature `quality3` is added so that we can classify the quality of the white wines into three categories, low, mid and high (see above for the detail), (3) among the features, `quality` and `quality3` are categorical, while the other 11 features are numerical.
- **Main feature(s) of interest** `quality` and `quality3` In the rest part of this report, we will use both of them as main features depending on which numerical features we consider at the same time. The former is useful for analyzing in detail, while the latter is convenient to more roughly see how the quality of the wines is influenced by their chemical properties.
- **Other features useful for investigating the main feature(s)** `density`, `residual.sugar` From the univariate plots, we can see that these features show non-normal distributions. We therefore analyze these variables more in detail below by using bivariate and multivariate plots. (We however note that the univariate analysis is not enough to identify which numerical features are really useful to investigate the main features. We thus analyze other features as well through the bivariate plots.)
- **New feature(s) created by ourselves** `quality3` We have introduced this new feature to classify the quality of the wines into three categories, high, mid, and low (see Section 2.1 for more details of their definitions). As we will see, this feature is useful to grasp how the quality of the wines is affected

by the other features roughly and easily (the classification by `quality` is sometime too detailed to grasp some tendency).

- **Features with unusual distributions** `density`, `residual.sugar` The distribution of `density` after log10 scaling shows two peaks clearly. The distribution of `residual.sugar` (without any scaling) shows two peaks, too.

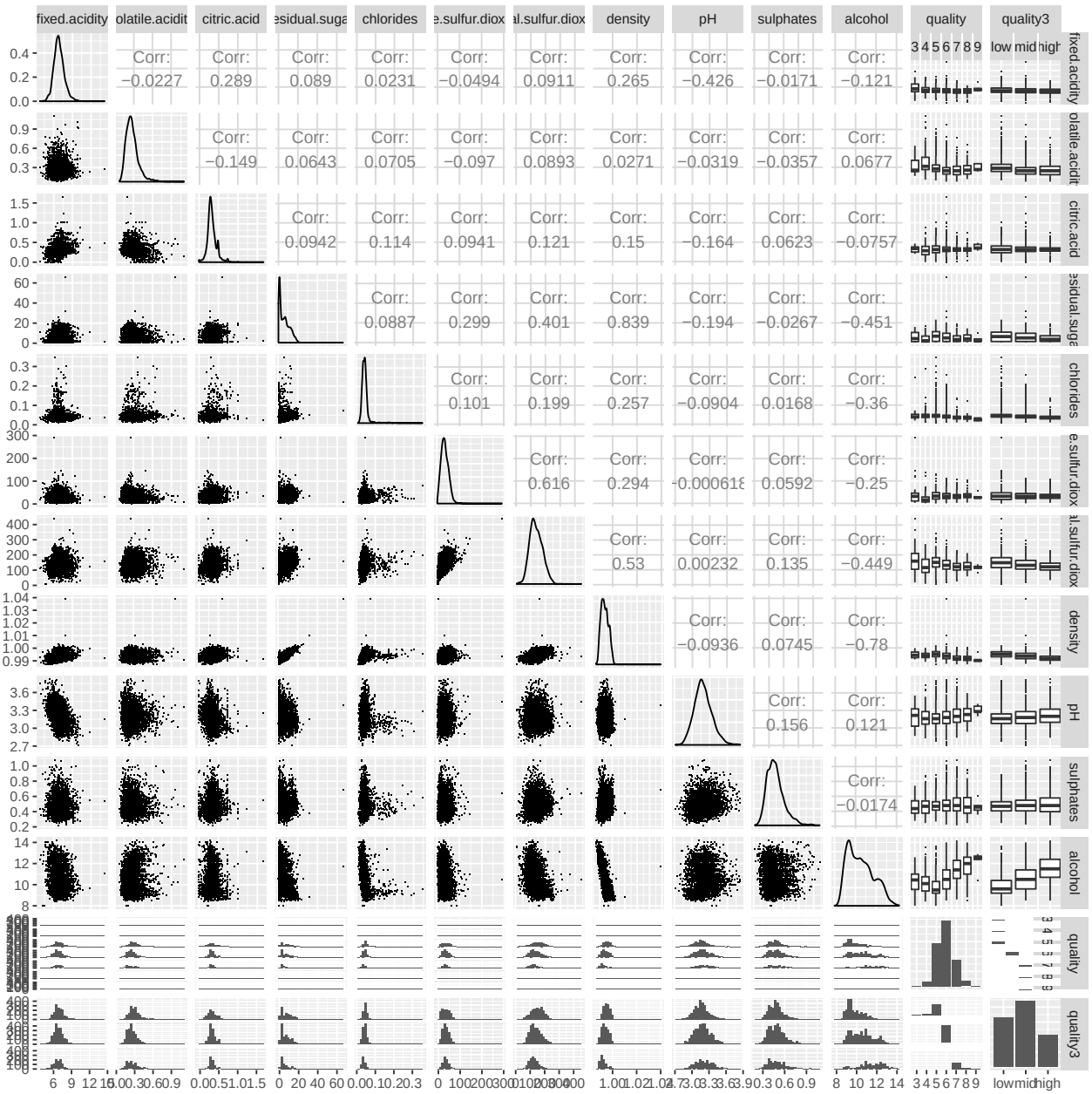
3. Bivariate Plots Section

As a next step, here we make some bivariate plots and investigate relationships between a pair of the features.

3.1. Bivariate Plots

In this part, we provide some bivariate plots.

Before looking into the details, let us start with a matrix plot of the dataset:



We also summarize the correlations between `quality` and the numerical features (we note that here `quality` must be treated as an integer, not a factor):

```
##        fixed.acidity     volatile.acidity          citric.acid
##         -0.113662831         -0.194722969         -0.009209091
##       residual.sugar            chlorides  free.sulfur.dioxide
##         -0.097576829         -0.209934411          0.008158067
## total.sulfur.dioxide              density                   pH
##         -0.174737218         -0.307123313          0.099427246
##            sulphates              alcohol
##          0.053677877          0.435574715
```

From these, we can see that pairs of features with high correlations are:

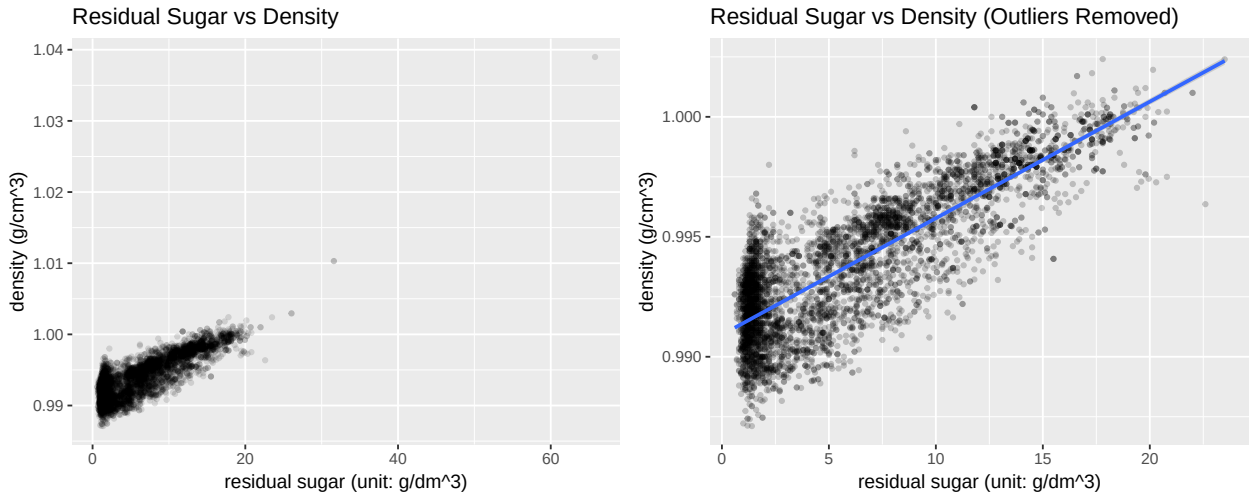
- (`residual.sugar`, `density`) with |correlation| = 0.839
- (`density`, `alcohol`) with |correlation| = 0.78

- (`free.sulfur.dioxide`, `total.sulfur.dioxide`) with |correlation| = 0.616

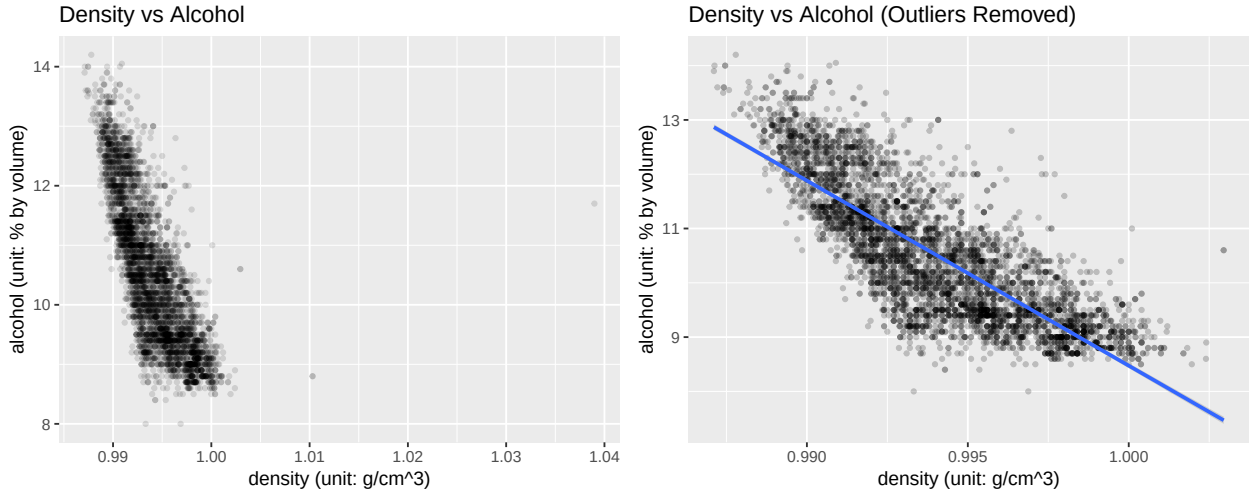
We also note that the first four features which have relatively large correlations with the main feature `quality` are:

- `alcohol` (|correlation| = 0.436)
- `density` (|correlation| = 0.307)
- `chlorides` (|correlation| = 0.210)
- `volatile.acidity` (|correlation| = 0.195)

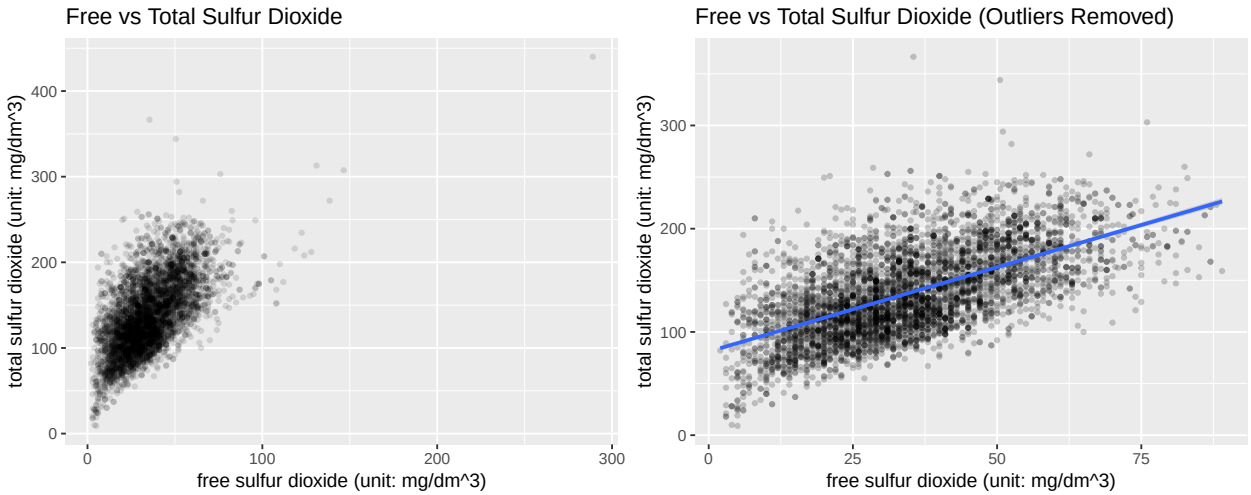
As a next step, by taking into account the result of the matrix plot and correlations, we now look at the bivariate plots with high correlations more in detail.



In the left scatterplot the full data is used, while in the right one the outilers are removed (based on the value of the x-axis, the threshold is the same as in the univariate plots). The blue line in the right plot is the result of the linear regression. (We will do the same analysis in the scatter plot for the other pairs of variables just below.) From these plots, we can see that more residual sugar indicates higher density. This seems to be similar to sugar water which has higher density than water itself.

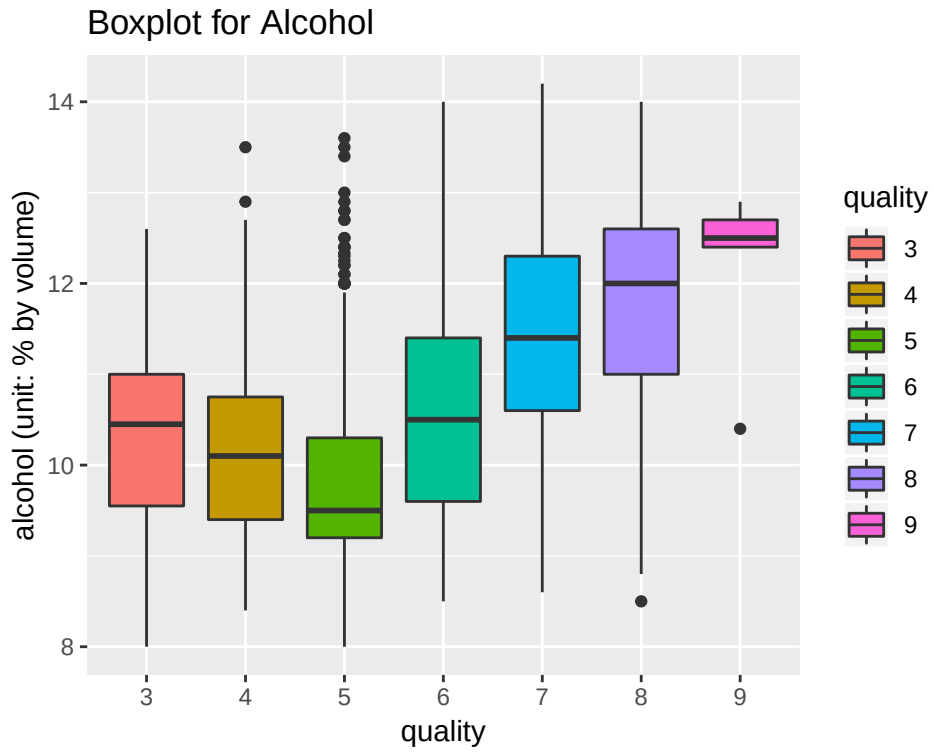


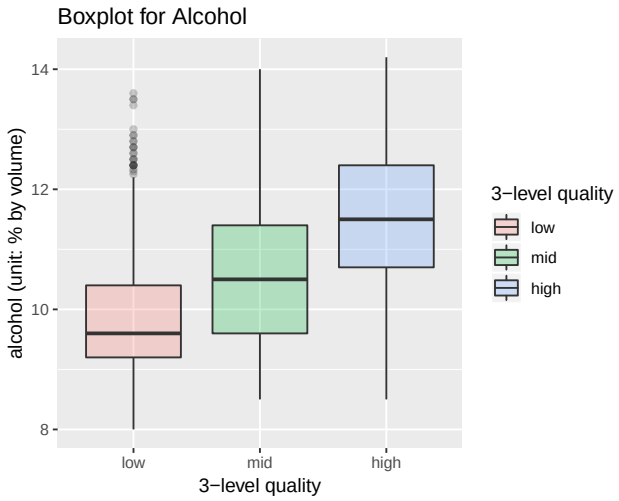
We can see that as the amount of alcohol increases, density decreases. We note that the density of alcohol at room temperature (20 degree Celsius) is 0.7893 g/cm^3, while the density of water (which is most part of wine) is 0.9982 g/cm^3. Thus this result seems to be reasonable.
(See Wikipedia: Ethanol and Wikipedia: Water for densities of alcohol and water.)

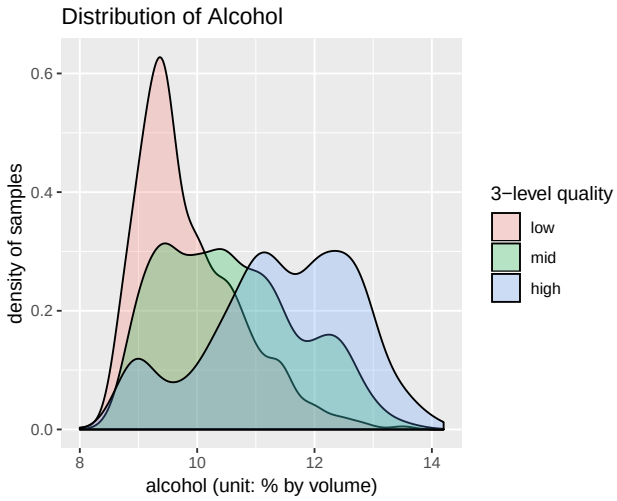


We can see that as the amount of the free form of sulfur dioxide increases, the total amount of sulfur dioxide also increases. This large correlation is natural since the free form of sulfur dioxide is a part of the total sulfur dioxide (= sum of free and bound forms of sulfur dioxide).

Now we consider the numerical features which have relatively high correlations with `quality` and make boxplots and density plot for them. We plot the observations with different value in `quality` (or `quality3`) separately.

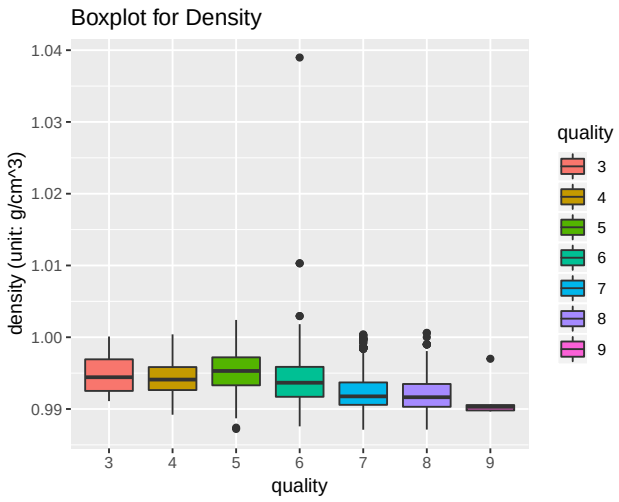


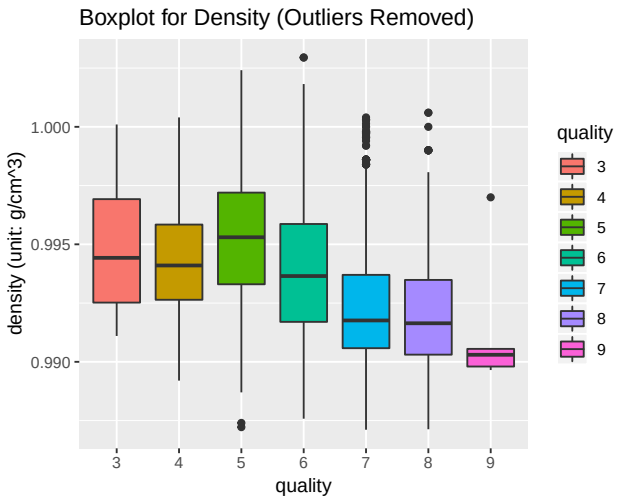


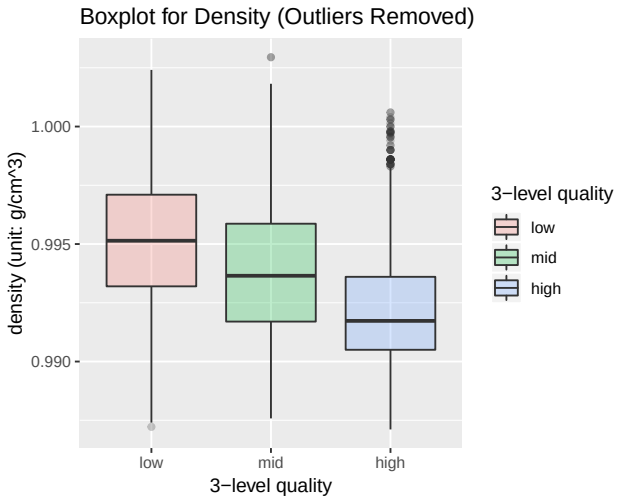


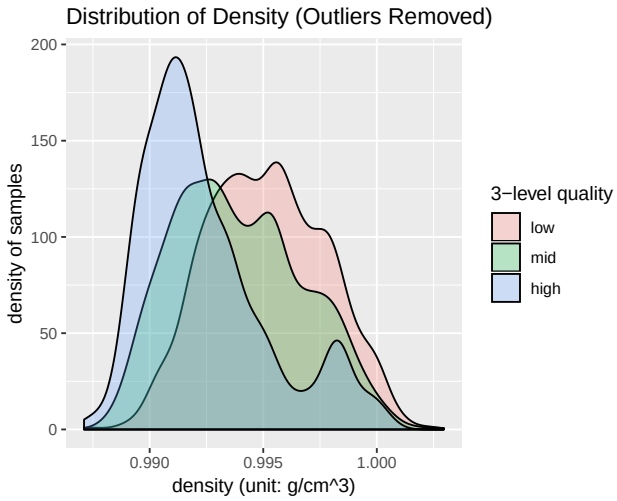

```
## $low
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.20    9.60    9.85   10.40   13.60
##
## $mid
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50    9.60   10.50   10.58   11.40   14.00
##
## $high
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50   10.70   11.50   11.42   12.40   14.20
```

Below the plot, we also provided the summary for `alcohol` in which observations with different values in `quality3` are treated separately (all the observations are used for the computation). There seems to be a tendency that as the percentage of alcohol increases, the quality of wine becomes better. We can see this clearly through the bottom-left and bottom-right plot in which `quality3` is used instead of the original feature `quality` to classify the observations.



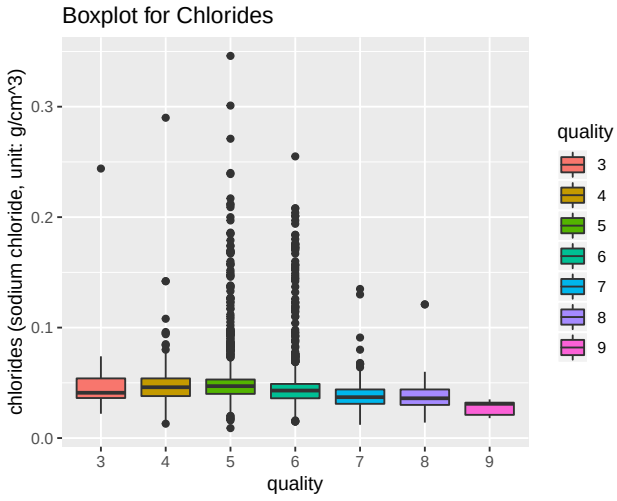


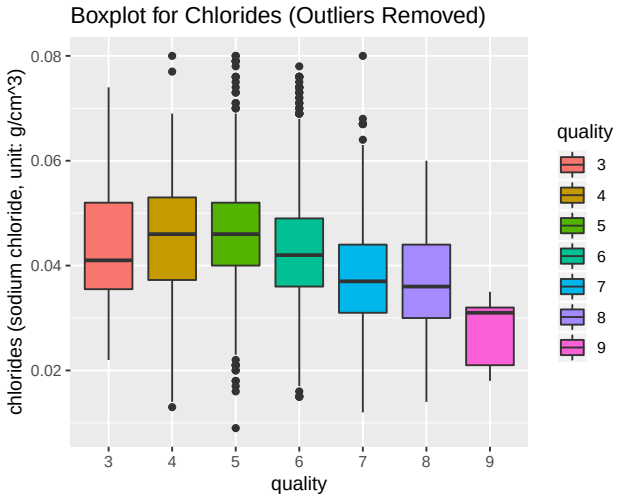


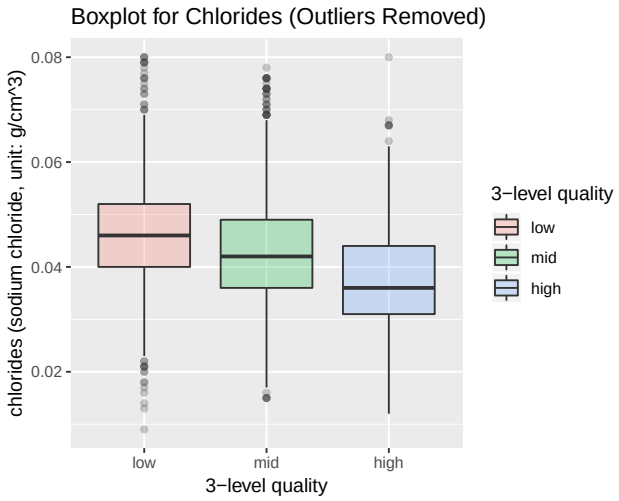


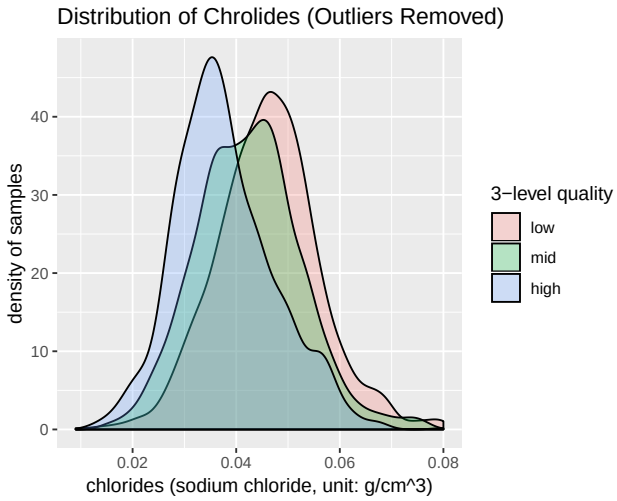

```
## $low
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9872  0.9932  0.9951  0.9952  0.9971  1.0024
##
## $mid
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9876  0.9917  0.9937  0.9940  0.9959  1.0390
##
## $high
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9905  0.9917  0.9924  0.9936  1.0006
```

We note that, as in the univariate plots, we have removed some outliers when plotting the top-right, bottom-left and bottom-right plots. (We will do the same thing when we make boxplots and density plots of `chlorides` and `volatile.acidity` just below.) These plots, especially the two bottom plots indicate that the white wines with higher quality have smaller density.



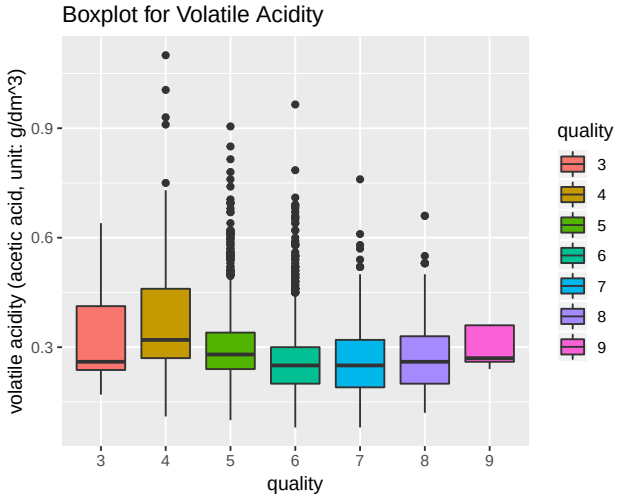


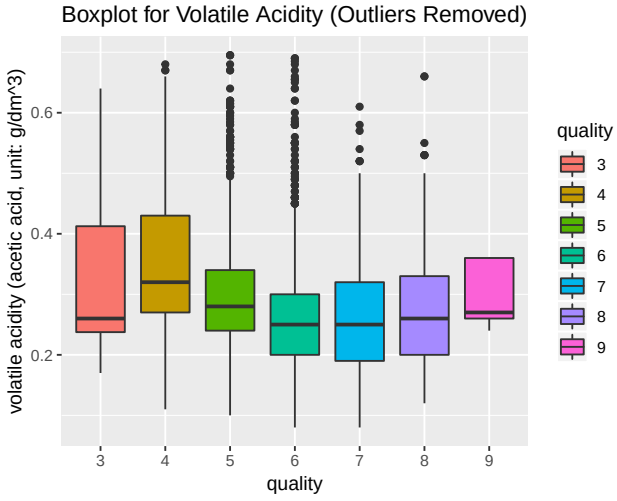


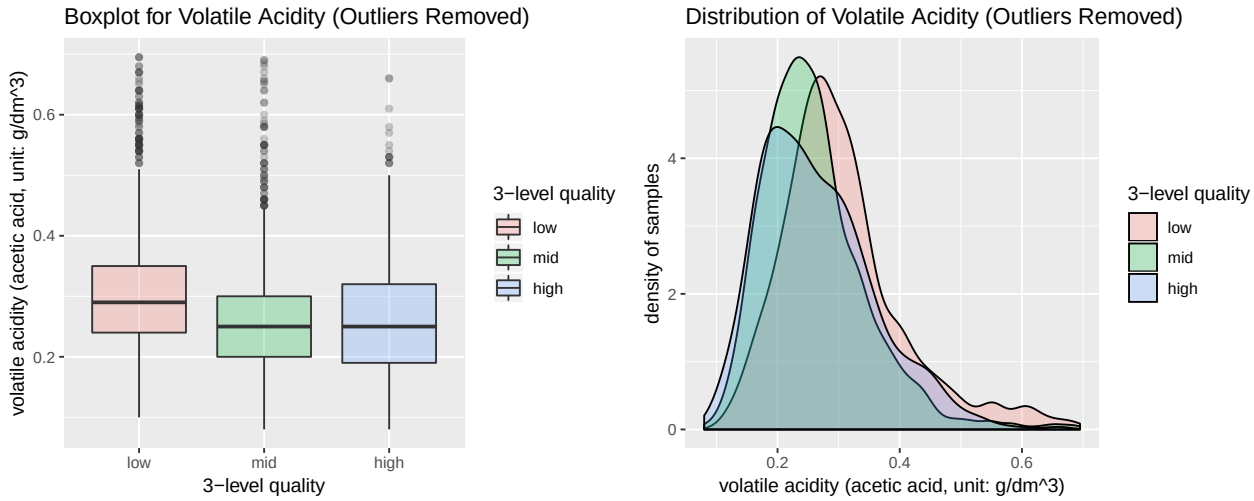



```
## $low
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00900 0.04000 0.04700 0.05144 0.05300 0.34600
##
## $mid
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01500 0.03600 0.04300 0.04522 0.04900 0.25500
##
## $high
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.03100 0.03700 0.03816 0.04400 0.13500
```

These plots imply that there is a tendency that the white wines with higher quality contain less chlorides.








```
## $low
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1000  0.2400  0.2900  0.3103  0.3500  1.1000
##
## $mid
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.2000  0.2500  0.2606  0.3000  0.9650
##
## $high
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.1900  0.2500  0.2653  0.3200  0.7600
```

We can not see any significant relationship between quality and volatile acidity.

3.2. Bivariate Analysis

In this part, we will analyze the relationships between a pair of features based on the bivariate plots above.

- **Relationships between the main and other features** From the bivariate boxplots and density plots we made above, we can see the following relationships between the main features, `quality` and `quality3`, and some of the other numerical features:
  - Wines with higher quality contain higher percentage of alcohol.
  - Wines with higher quality have less density.
  - Wines with higher quality contain less chrolides.
- **Other relationshipes observed** From the bivariate scatter plots we made in the previous part, we can see the following relations among some of the numerical features:
  - Wines with larger amount of residual sugars has higher density.
  - Wines with higher density contains less percentage of alcohol.
  - Wines with larger amount of the free form of sulfurdioxide contains larger amount of sulfur dioxide in total.
- **Strongest relationship found** Among the relationships we have just mentioned, as can be seen from the boxplot above, we can especially see the strong relation between `quality3` and `alcohol`: The white wines with higher quality contain higher percentage of alcohol.

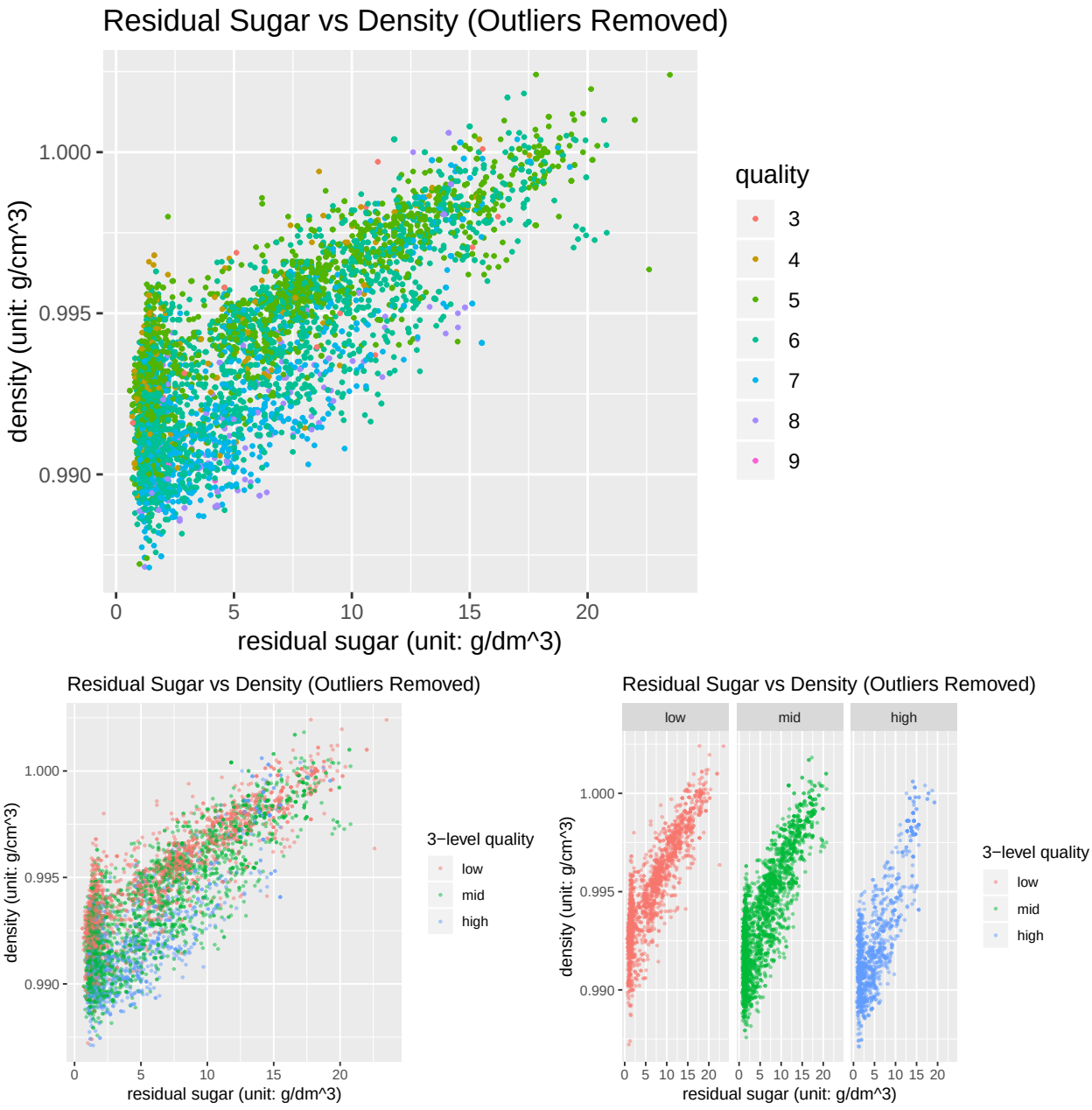
Motivated by the bivariate analysis, we will analyze the relationships between `residual.sugar` and `density` as well as `density` and `alcohol` further in the next part of this report.

4. Multivariate Plots Section

Now we consider more than two features at the same time. We make some multivariate plots and analyze them. We also discuss the statistical model building based on the linear regression for explaining how the chemical properties of the white wines can influence their quality.

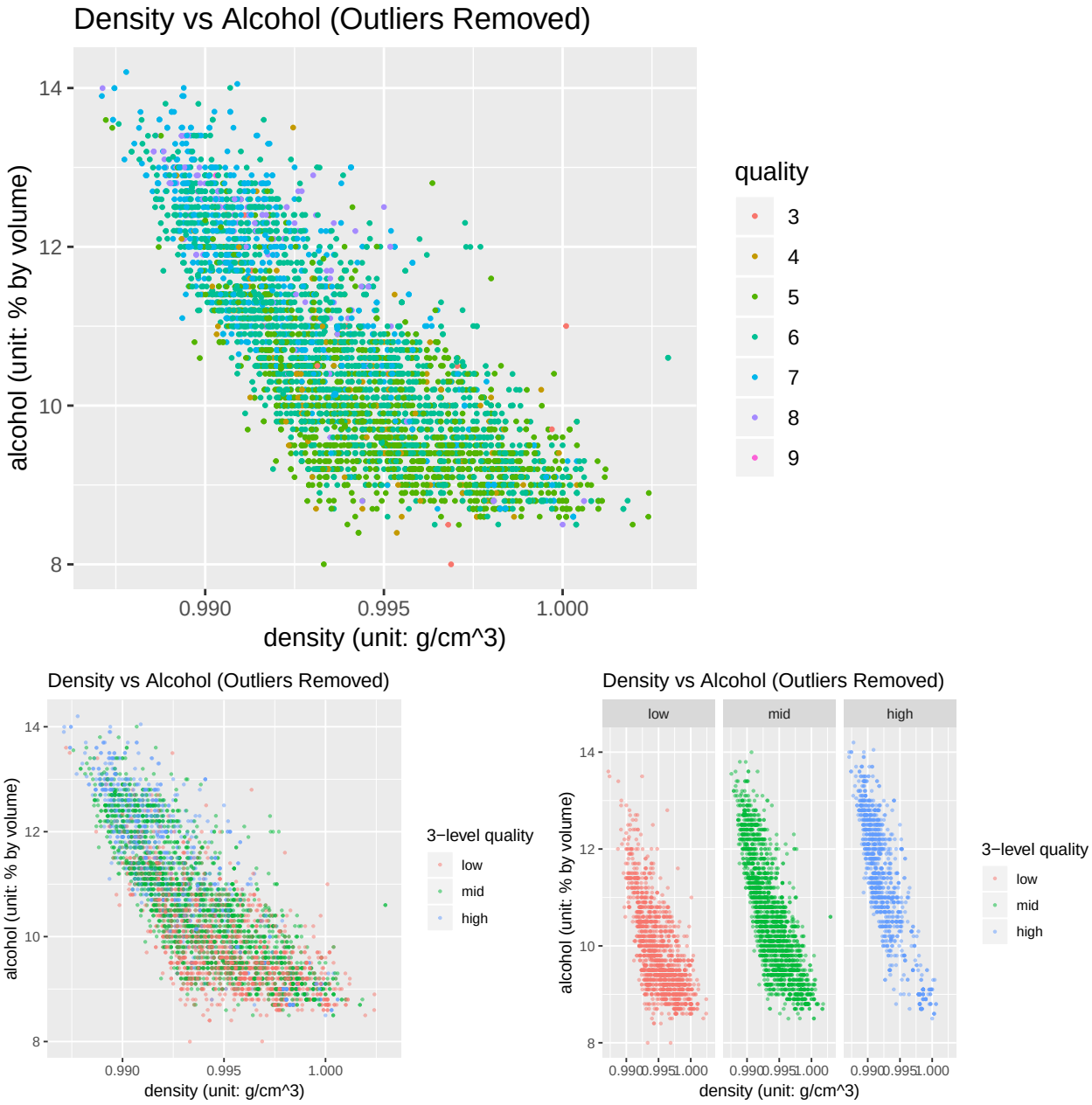
4.1. Multivariate Plots

Taking into account the analysis we did in the previous parts, here we create some multivariate plots.



Here we have removed some outliers as we did in the univariate plot part (same for `density` vs `alcohol` plot below). In the top plot, `quality` is used to classify the observations, but it is difficult to capture some potential relationships. From the two bottom plots in which `quality3` is used for the classification, on the

other hand, we can see that, for a fixed density, the white wines with low quality tend to contain smaller amount of residual sugar than those with high quality. (On the other hand, the amount of residual sugar varies widely for the wines with mid quality. It is hard to see any specific tendency compared to high/low quality wines.)



Again, in the top plot, `quality` is used for the classification of the observations but it is difficult to capture some potential relationships. From the two bottom plots in which `quality3` is used for the classification, on the other hand, we can see that the high quality wines are located in the low density and high alcohol range, while the low quality one are in the high density and low alcohol range. (On the other hand, the mid quality wines spread in a wider range. No specific tendency can be seen compared to high/low quality wines.)

4.2. Model Building

As a simple analysis, let us treat the variable `quality` as an integer type (not factor) and carry out the linear regression. By taking into account the high correlation between some numerical variables, we use `alcohol`, `residual.sugar`, `chlorides` and `volatile.acidity` as confounders and carry out the linear regression (since `density` has a high correlation with `residual.sugar` we did not add it as a confounder):

```
##
## Call:
## lm(formula = as.integer(as.character(quality)) ~ alcohol + residual.sugar +
##     chlorides + volatile.acidity, data = df_wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3867 -0.4976 -0.0362  0.4668  3.0269
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.464220   0.131013  18.809   <2e-16 ***
## alcohol           0.367897   0.010743  34.245   <2e-16 ***
## residual.sugar    0.026198   0.002433  10.767   <2e-16 ***
## chlorides        -0.906807   0.540285  -1.678   0.0933 .
## volatile.acidity -2.086006   0.109696 -19.016   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7627 on 4893 degrees of freedom
## Multiple R-squared:  0.259,  Adjusted R-squared:  0.2583
## F-statistic: 427.5 on 4 and 4893 DF,  p-value: < 2.2e-16
```

This linear regression model gives a very low R-squared value (around 0.26). Thus this modes does not explain how the chemical properties affect the quality of the wines well.

Next, we carry out the linear regression by selecting the confounders by using AIC (Akaike Information Criterion). (For the detail of AIC, see, for example, Wikipedia:AIC):

```
##
## Call:
## lm(formula = quality ~ fixed.acidity + volatile.acidity + residual.sugar +
##     free.sulfur.dioxide + density + pH + sulphates + alcohol,
##     data = df_wine[original_var])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8246 -0.4938 -0.0396  0.4660  3.1208
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.541e+02  1.810e+01   8.514  < 2e-16 ***
## fixed.acidity        6.810e-02  2.043e-02   3.333 0.000864 ***
## volatile.acidity    -1.888e+00  1.095e-01 -17.242  < 2e-16 ***
## residual.sugar       8.285e-02  7.287e-03  11.370  < 2e-16 ***
## free.sulfur.dioxide  3.349e-03  6.766e-04   4.950 7.67e-07 ***
## density             -1.543e+02  1.834e+01  -8.411  < 2e-16 ***
## pH                   6.942e-01  1.034e-01   6.717 2.07e-11 ***
## sulphates            6.285e-01  9.997e-02   6.287 3.52e-10 ***
```

```
## alcohol              1.932e-01  2.408e-02   8.021 1.31e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7512 on 4889 degrees of freedom
## Multiple R-squared:  0.2818, Adjusted R-squared:  0.2806
## F-statistic: 239.7 on 8 and 4889 DF,  p-value: < 2.2e-16
```

AIC selects 8 variables for the linear regression, but this model still has a low R-squared value (less than 0.30). We thus conclude that the linear regression models (without polynomial features at least) do not explain the quality of a given white wine from its chemical properties. I expect that more sophisticated multi-label classification model (random forest etc.) will be better to explain the quality of the wines from their chemical properties.

4.3. Multivariate Analysis

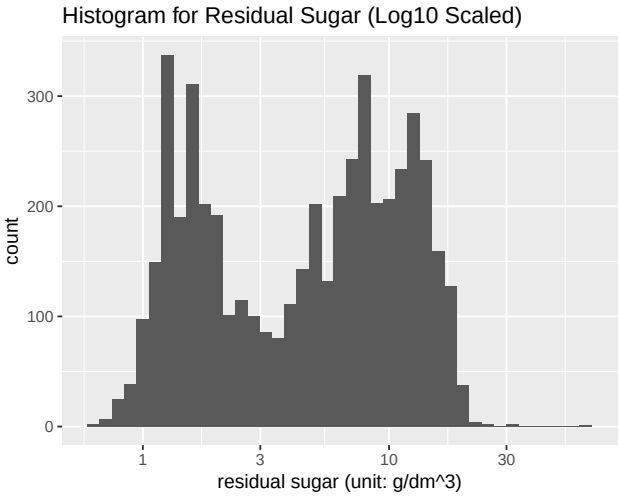
- **Relationships observed** We have further analyzed the relations between `residual.sugar` and `density` as well as `density` and `alcohol` by treating the observations with different qualities separately. We found that
  - For a fixed value of density, high quality wines have more amount of residual sugar than low quality ones.
  - High quality wines are located in the low density and high alcohol range, while low quality ones are in the high density and low alcohol range.
- **Interesting/surprising interactions between features** It is interesting that for a fixed value of density, we can clearly see that high quality wines and low quality ones contain different amounts of residual sugar. This difference allows us to distinguish high quality wines and low quality ones relatively easily. On the other hand, for a fixed value of density, the amount of residual sugar for the mid quality wines distributes in a wider range, making it difficult for us to distinguish mid quality wines from the high/low quality ones.
- **On statistical model** As a trial to explain quantitatively how the quality of the wines is influenced by their chemical properties, here we considered two statistical models based on the linear regression: (1) A linear regression model with `alcohol`, `residual.sugar`, `chlorides` and `volatile.acidity` as confounders, (2) A linear regression model with AIC used to determine the combination of confounders. However, the R-squared values of these models are very low (less than 0.30 for both of the two). I thus expect that more complicated/sophisticated statistical models (such as random forest etc.) can be useful to explain the quality of the white wines from their chemical properties.

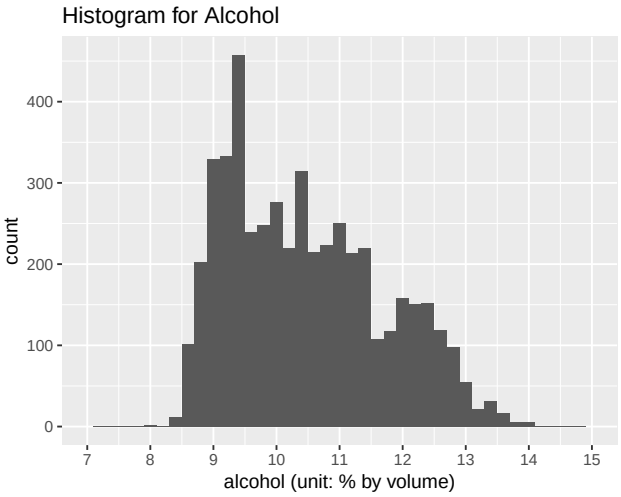
---

5. Final Plots and Summary

Among the univariate, bivariate and multivariate plots we made in the previous sections, here we pick up three key plots and give some summaries of our findings from these plots.

5.1. Plot One

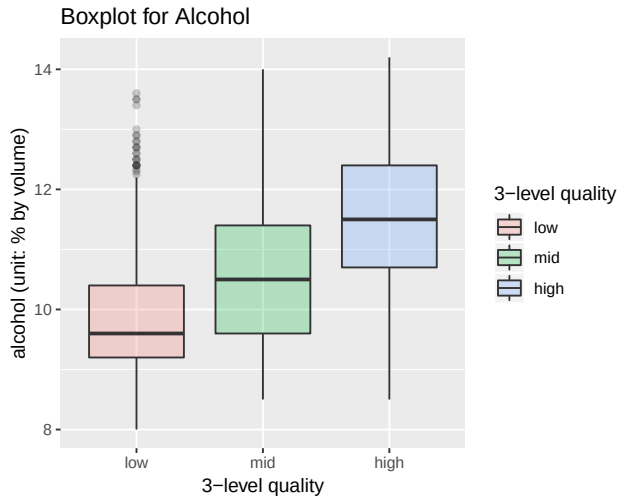


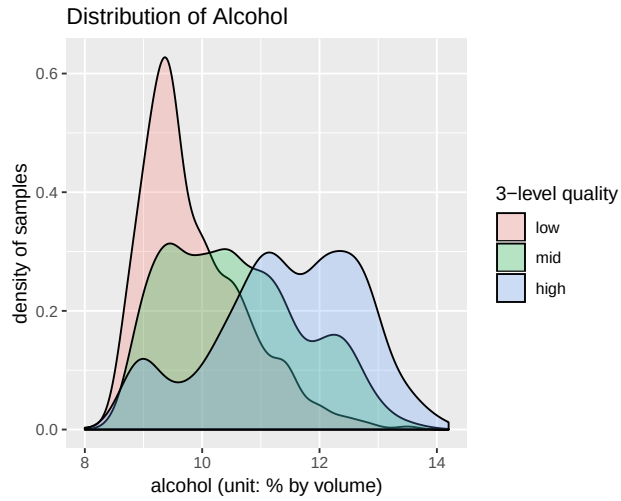


5.2. Description One

These histograms for `residual.sugar` and `alcohol` in Plot One (x-axis is log10 scaled for `residual.sugar`) show non-Gaussian distributions of these features. We can see two peaks in these distributions. This non-trivial profile motivates us to analyze `residula.sugar` and `alcohol` further by using bivariate and multivariate plots.

5.3. Plot Two



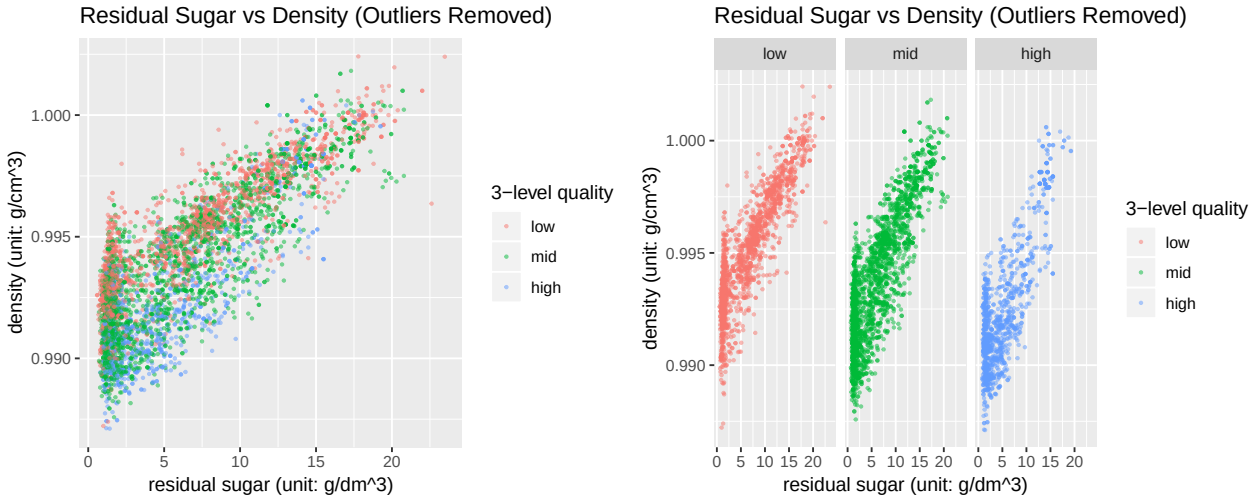


5.4. Description Two

As we have seen through the matrix plot, `alcohol` has the highest correlation with `quality`. Plot Two gives boxplot and density plot of `alcohol`. In these plots, instead of `quality` itself, we used the feature `quality3` which we introduced to classify the observations into simple three categories (low, mid, high based on the values of `quality`). We then plotted the observations with different values in `quality3` separately. This `quality3` allows us to see how the quality is influenced by `alcohol` more clearly than the plot(s) with `quality` used. Through Plot Two, we have thus confirmed that (1) the percentage of `alcohol` increases as

the quality of the white wine increases as well as (2) the usefulness of the classification of the observations based on `quality3`.

5.5. Plot Three



5.6. Description Three

Plot Three is the 2d scatter plots of the pair (`residual.sugar`, `density`) which has the highest correlation among the pair of numerical features. Here the observations with different values in `quality3` are plotted separately (in the same panel in the left plot, while in the separate three panels in the right plot). We note that some outliers (`residual.sugar` > 25) in the dataset are removed in these plots. These plots show that for a fixed value of density, high quality wines tend to have more residual sugars than low quality ones. On the other hand, the amount of the residual sugar for mid quality wines varies widely for a fixed density, making us hard to distinguish the mid quality wines with the high/low quality ones.

---

6. Reflection

- **Insights from exploratory data analysis** In this report, we carried out the exploratory data analysis on how the chemical properties of the white wines influence their quality. We started with the dataset which contains 4898 observations with 13 variables. Among them, one variable `X` (which just stores the indices of the observations) was removed. We then added a new variable `quality3` which classifies the observations into 3 categories (high, mid, low) depending on the values of `quality`. This variable allows us see the difference in the profiles of high/low quality wines relatively easily. By using this cleaned dataset, we made some univariate, bivariate as well as multivariate plots to understand the properties of the features and their relationships visually. The followings are main insights we obtained from our analysis:
  - The amount of the residual sugar and percentage of alcohol distribute in a non-Gaussian manner (two peaks),
    while the other numerical features do in a nearly Gaussian manner. (log10 scaling is needed for the residual sugar.)
  - The percentage of alcohol influences the quality of wines: High quality wines contains higher percentage of alcohol than Low quality ones.

- The following pairs of chemical properties have high correlations: (density, amount of residual sugar), (density, alcohol), (amount of total sulfur dioxides, amount of free sulfur dioxides).
- When the density is fixed, there is a tendency that high quality wines contain more amount of residual sugars than low quality ones.
- Judging from the linear regression we carried out (which results in the models with small R-squared values), it seems to be difficult to explain the quality of the white wines by using the linear regression models (polynomial features might work though). I expect that it will be worthwhile to use more complicated/sophisticated methods such as random forests for this purpose.

- **Struggles, Success and Surprise in exploratory data analysis**
  - One of the biggest difficulties in the course of the exploratory data analysis was that the number of the observations corresponding to very high (`quality` >= 8) and very low (`quality` <= 4) quality wines is very small compared to those in between.
    This make it difficult to visually figure out how the chemical properties affects the quality of the white wines by using `quality`. To see this more clearly, we have introduced a simpler feature, `quality3`, for the classification of quality. This worked well, allowing us to see the difference of high/low wines more easily.
  - Another difficulty is that there is no numerical features which have large correlations with `quality`. (The one with the highest correlation is `alcohol` but the correlation is less than 0.50). This indicates that many features can influence the quality of the wines at the same time . We therefore took a look at various features carefully to see which ones can affect the quality of the wines somehow significantly.
  - It was very surprising for me that the high quality wines tend to have high percentage of alcohol. I have never looked at the percentage of alcohol when purchasing white wines...

- **Some remarks for future analysis** Here we end up this report with some comments and remark for future investigation:
  - To improve the analysis, an interesting direction is to take into account more features. For example, the vintage of wines and species of grapes are interesting to be taken into account. The dataset we used here is for vinho verde. Some of these wines are weakly carbonated. This feature might be worthwhile to be taken into account.
  - Since our linear regression analysis does not find a reasonable statistical model. It is worthwhile to investigate sophisticated/complicated methods to construct better statistical models.
  - Another interesting question is "is taste of a high quality wine really good?". In the dataset used in this report, as described in Cortez et.al., the inter-professional organization called CVRVV decided the quality of the wines. It is interesting to ask to normal people to judge if the taste of each wines in the dataset is good or not. This will clarify the (un)gap between what the professional people think is nice and what normal people do.