White Wine Dataset Analysis

Stefan Dulman

Intro

In this project, I will focus on the white wine data set provided by udacity. This dataset contains 4898 observations of various Portuguese “Vinho Verde” white wines. Eleven different characteristics were recorded by measuring physical characteristics of the wines. Additionally, an estimate of the quality of wine by a set of experts was made available. Together with a vector assigning unique consecutive numbers to the measurements this leads to 13 different variables in our dataset:

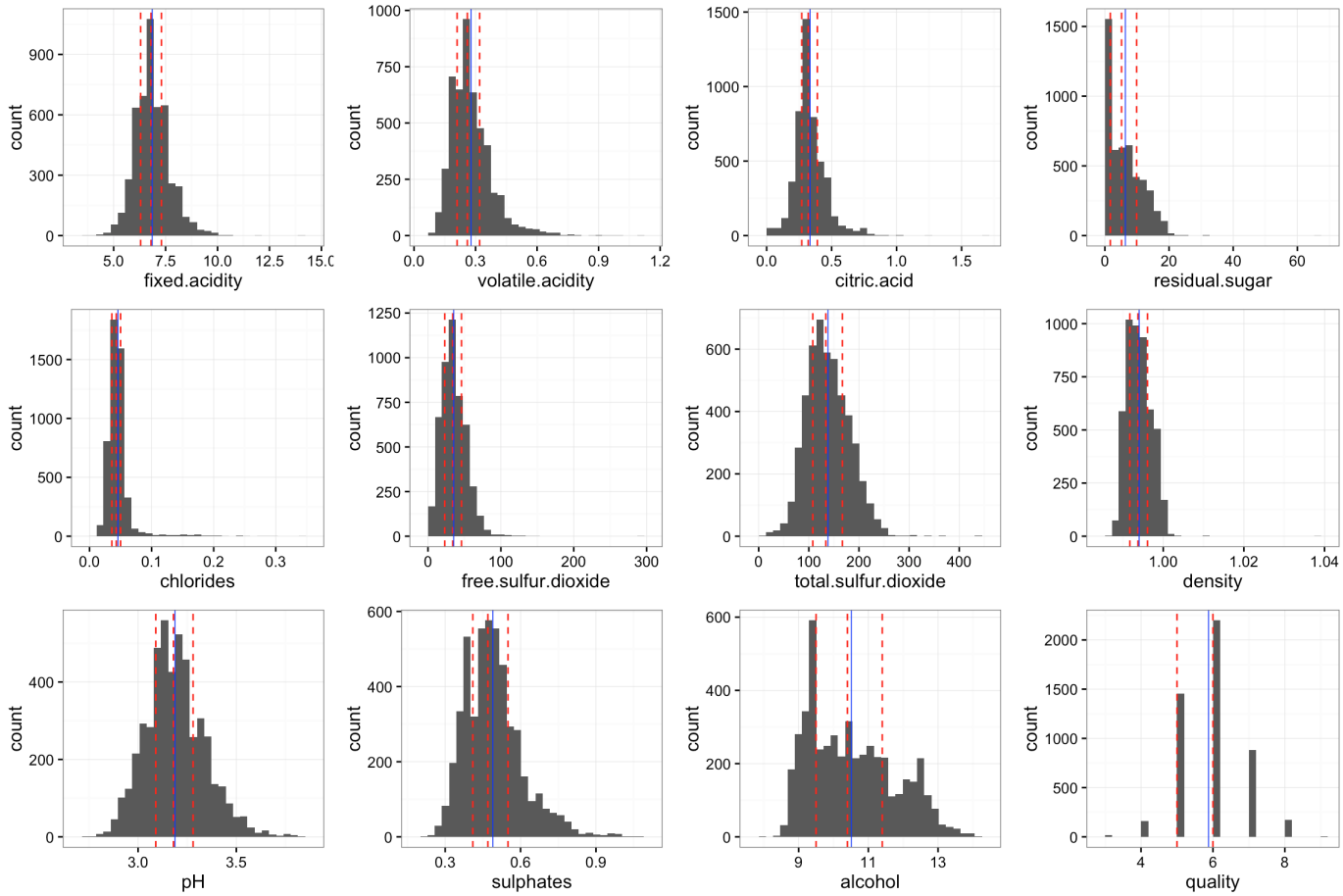
```
## 'data.frame':    4898 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  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 ...
```

Univariate Analysis

The start point of our analysis is to inspect each of the variables and observe their distributions. All the physical measurements are float numbers, while the quality of the wines is given as an integer. The statistics of these variables are:

```
##  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
```

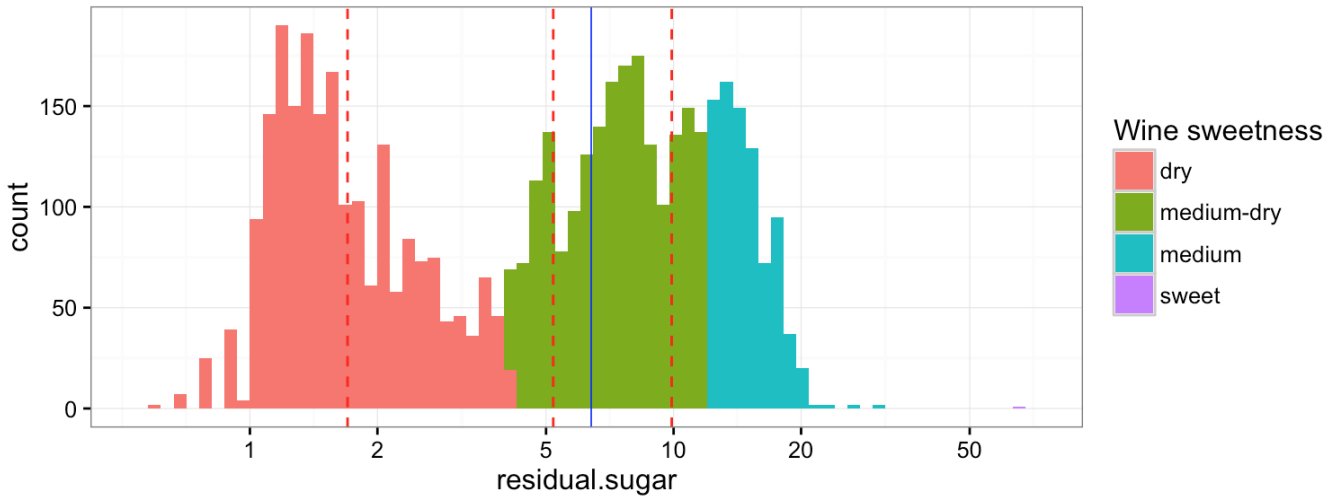
This table gives a general idea on the range of the values and the quantiles. To be noticed that the data set is clean and has no missing values. Representing the data as histograms should present a more clear picture:



These graphs show the distribution of univariate data. As expected, the 'quality' variable is the only discrete variable with just a few levels (for example, the alcohol variable is also discrete but on 103 values).

The 'quality' distribution looks normal, with few outliers (very bad wines marked with 3 and very good ones marked with 9). The large majority of wines are "normal" quality, leading to an unbalanced data set when it comes to training estimators.

Some of the distributions look skewed - I will focus on the 'residual.sugar':



Plotting the density of the residual sugar on a logarithmic x scale reveals a multimodal distribution with a clear cut point around 3.5 (the cut point between dry and medium-dry wines is usually around 4) and a second one around 10 (the cut point between medium-dry and medium wines is usually around 12). The wikipedia page for white wines [https://en.wikipedia.org/wiki/Sweetness_of_wine (https://en.wikipedia.org/wiki/Sweetness_of_wine)] mentions that it is very rare to find wines with residual sugar values of less than 1g/L. This is confirmed by our data set:

```
## [1] "percentage of wines with low residual sugar: 1.57207023274806%"
```

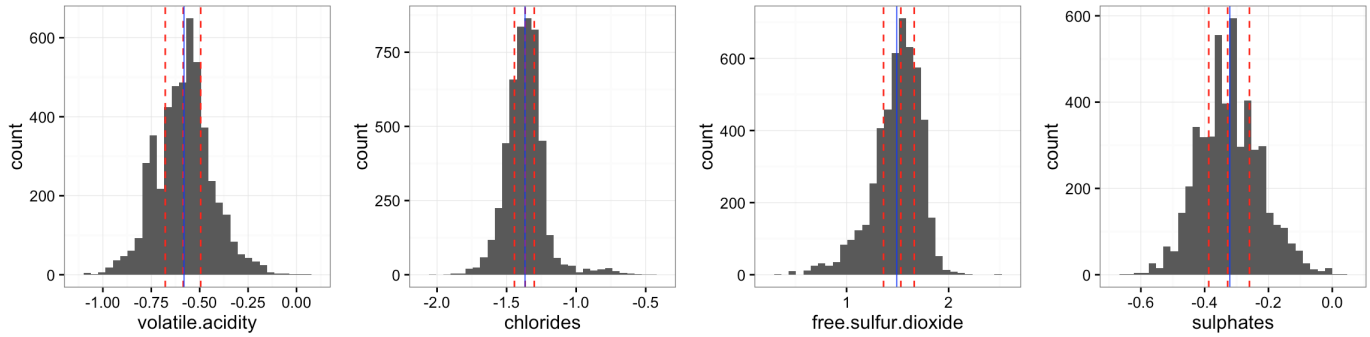
The same wikipedia page mentions that, as a general rule of thumb, wines with a residual sugar over 45g/L are considered sweet. Our data set is heavily biased from this perspective, the amount of entries in each taste category being:

```
##
##        dry medium-dry     medium      sweet
##       2097        1975        825          1
```

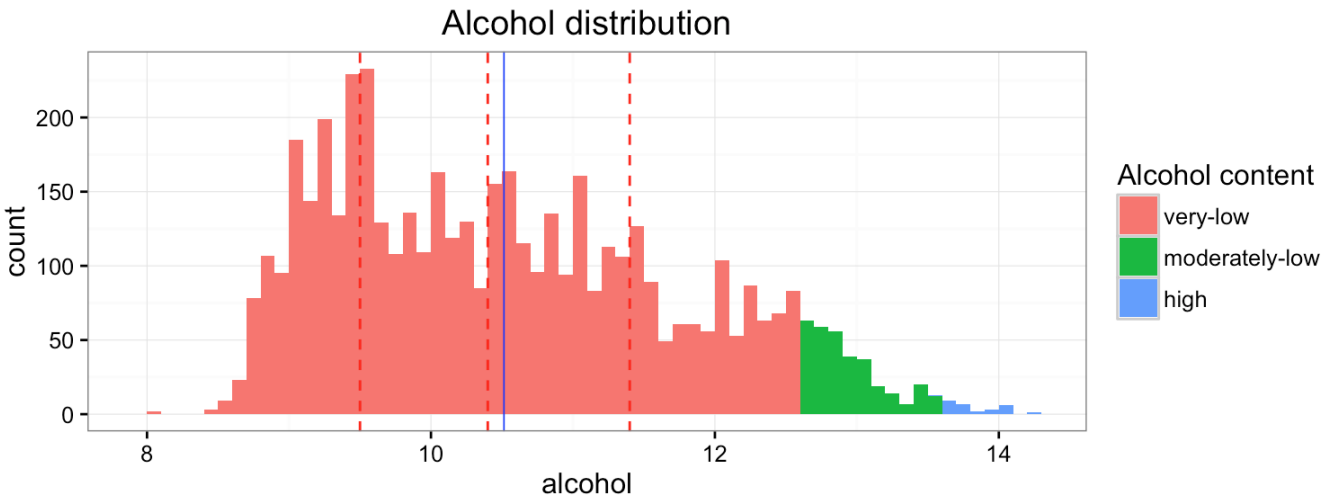
This data set has only one sweet wine entry in almost 5000 entries! Either the region/producers from which the data is collected specializes in dry wines or there might be a systematic error in the measurement of residual sugar.

We confirm that the three peaks we see in the data correspond to the rough categories described on the wikipedia page (the cut points suggested by wikipedia being 4, 12 and 35).

Using a logarithmic scale makes some of the distributions look more close to normal distributions, without revealing multimodality:



The graph of the alcohol has an interesting shape:

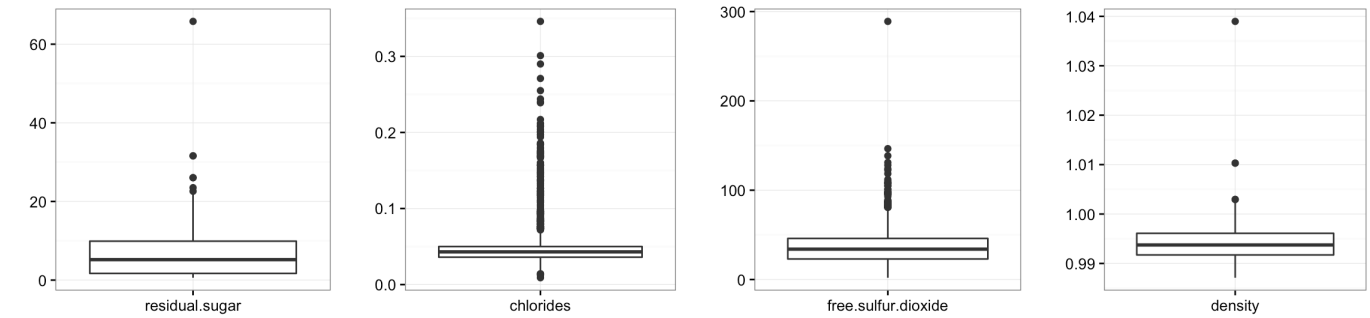


According to this link: [http://www.realsimple.com/holidays-entertaining/entertaining/food-drink/alcohol-content-wine (http://www.realsimple.com/holidays-entertaining/entertaining/food-drink/alcohol-content-wine)] wine can be classified in four categories, with the cut points at 12.5%, 13.5% and 14.5%:

```
##
##       very-low moderately-low           high      very-high
##           4543             326              29              0
```

The data set exhibits a clear bias towards the very low alcohol wines. This was to be expected as Portuguese vinho verde is a typical case of very low alcohol wine.

As a final step, I investigate the distribution of the outliers in the heavily skewed distributions:



These new graphs confirm that using logarithmic scales for chlorides and free.sulfur.dioxide is a sensible choice. Their logarithmic representation (see a previous graph) resemble more normal distributions, exhibiting few outliers.

Based on the observations above, I extended the original data set with several new columns:

- two categorical variables based on the thresholds of sugar and alcohol (thr.sugar and thr.alcohol) - for use in easier displaying of information

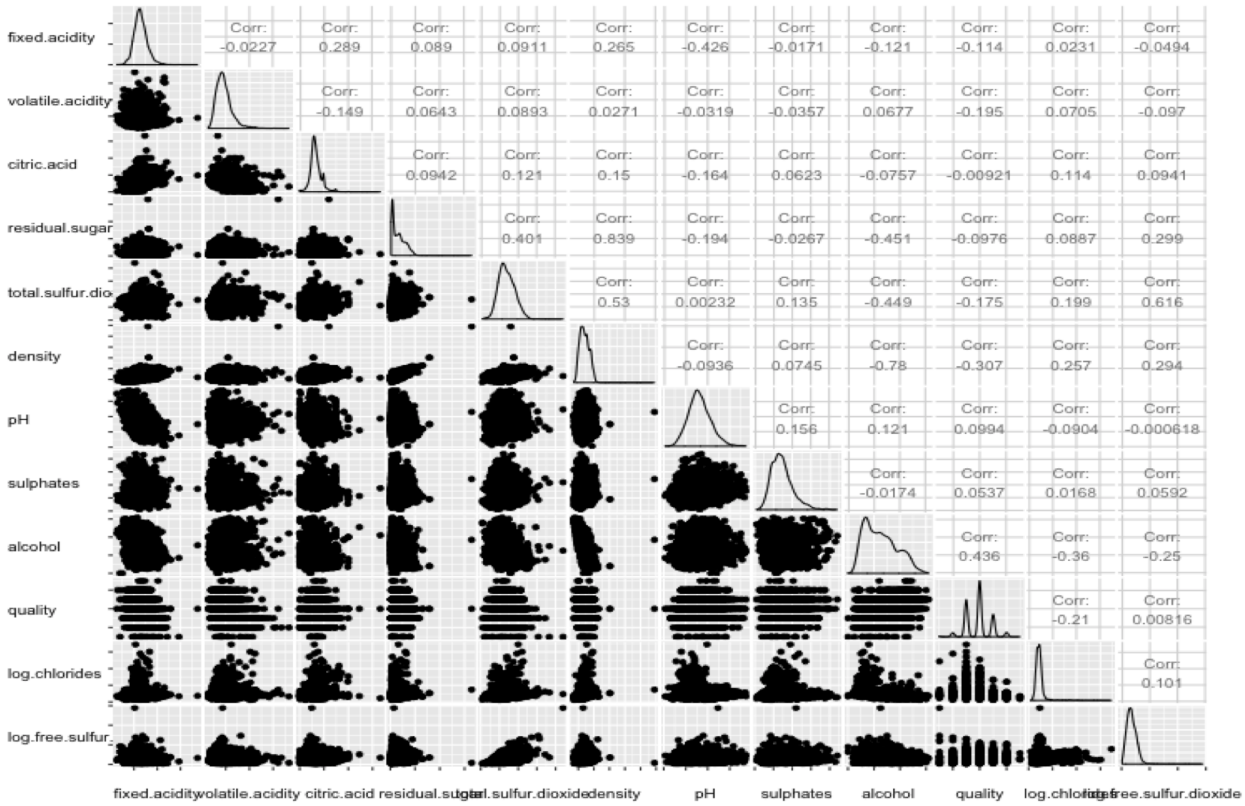
- two numerical variables based on the logarithms of chlorides and free.sulfur.dioxide (log.chlorides, log.free.sulfur.dioxide) - motivated by a significant increase in correlation with the 'quality' variable (25% and 1137% respectively)
- two quantified variables (quality, density) - for ease of displaying info

Summary

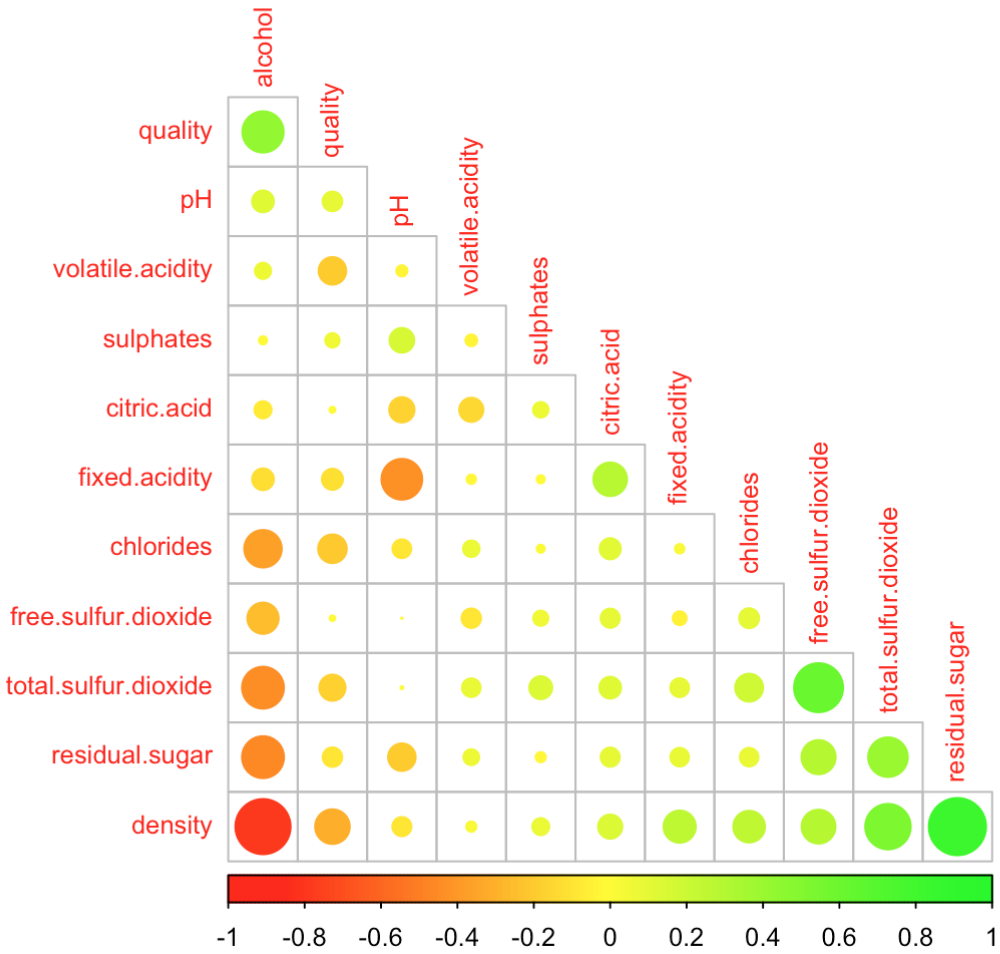
- **Structure of the dataset:** 4898 observations of 13 variables
- **Main features of interest:** the quality of the wine which is a function of several characteristics (alcohol, density)
- **Additional features of interest:** several other variables are loosely correlated with quality (residual.sugar, volatile.acidity)
- **New variables created:** log.chlorides, log.free.sulphur.dioxide (log versions of original variables); thr.sugar, thr.alcohol (tresholded versions of the original variables), thr.quality, thr.density
- **Unusual distributions:** residual.sugar is a multimodal distribution
- **Data tidy, adjust or change operations:** apart from the newly introduced variables, no other modifications were performed.

Bivariate Analysis

I started by exploring the correlation between all the pair-wise variables:



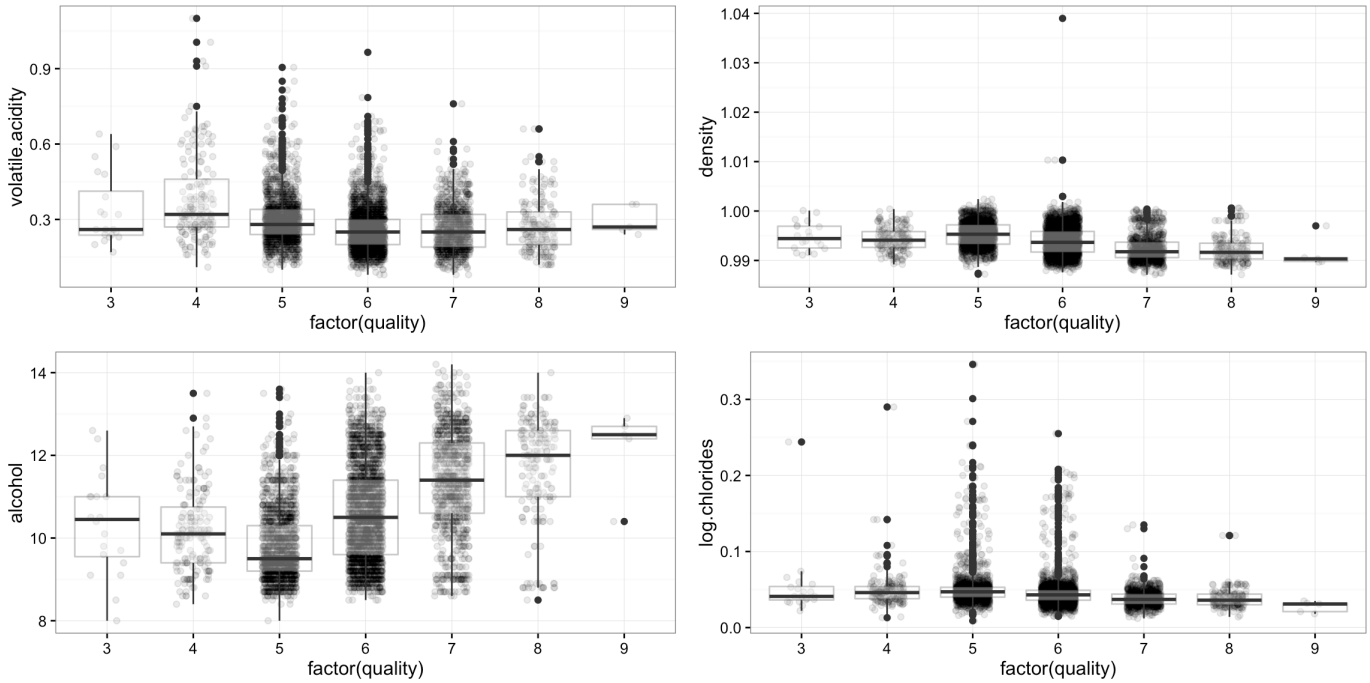
Most of the correlation coefficients in the above image are close to zero, with a few notable exceptions. A better look at the correlations is provided by:



This graph is intended as a helping tool for training classifiers. I have color and size coded the correlations between the variables in the original data set. Several observations can be made:

- alcohol is correlated with roughly half of the features in the set
- the acidity variables are correlated among themselves (pH with fixed.acidity which in turn correlates with citric.acid)
- the sulfur variables also match (total.sulfur.dioxide and free.sulfur.dioxide)
- quality is one of the weakest correlated variables - only alcohol seems to match it

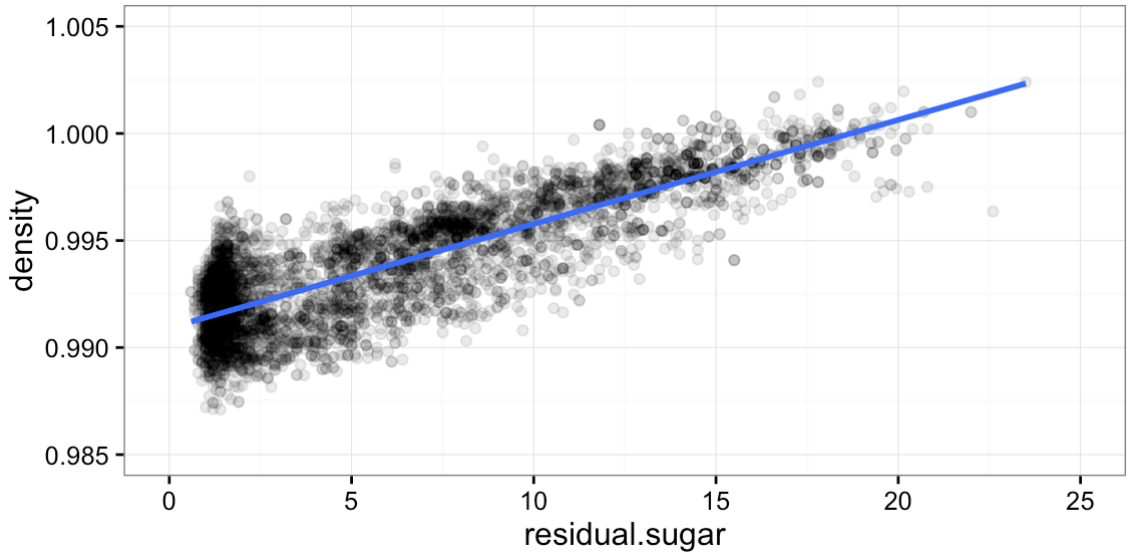
I will try to get a better understanding of how quality is related to other variables:

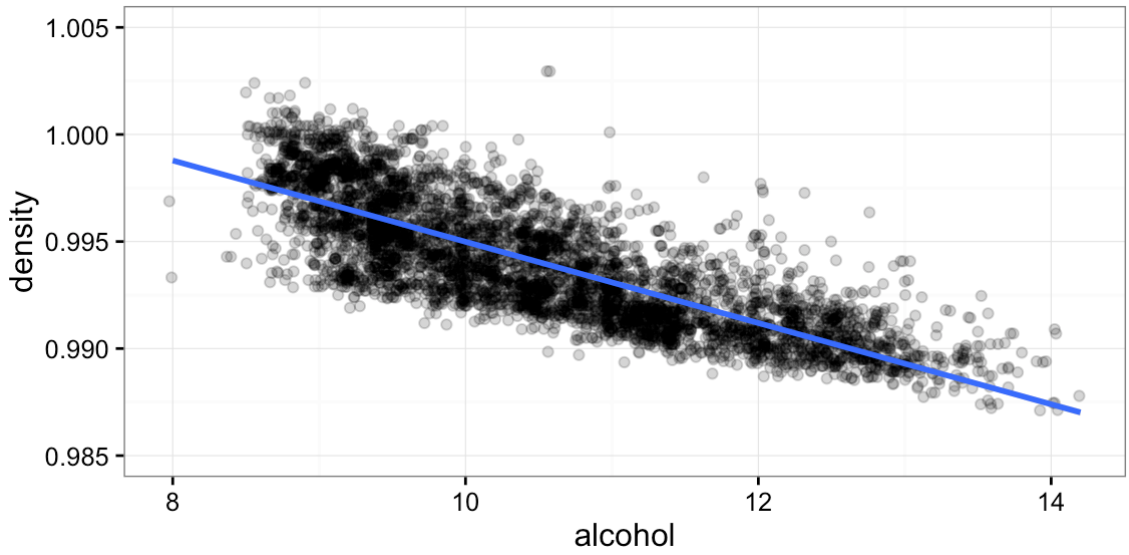


With maybe the exception of alcohol (higher values correlate with higher quality), no clear relationship related to quality stands out from these graphs. This is justified also by the small correlation found between the variables: (-0.195 - volatile.acidity, -0.3 - density, 0.436 - alcohol and 0.21 - log.chlorides).

Alcohol seems to be correlated to most of the variables in the dataset. It is exhibiting the largest correlation coefficients with the other variables (density -0.78, residual.sugar -0.451, total.sulphur.dioxide -0.449, log.chlorides -0.36).

Density correlates strongly with residual.sugar (0.839) and alcohol (-0.78):



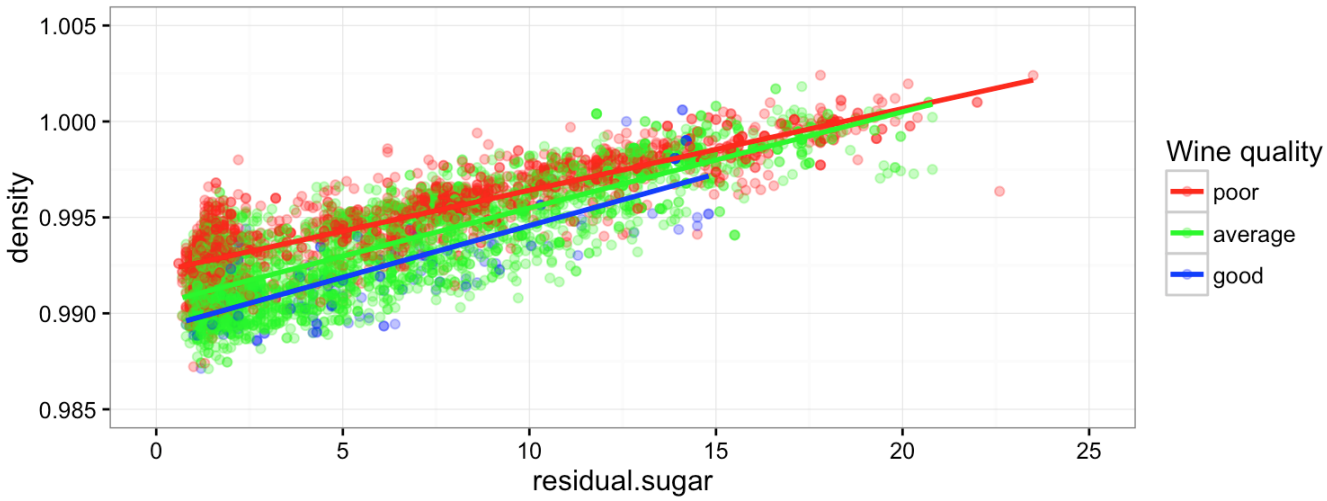


Summary

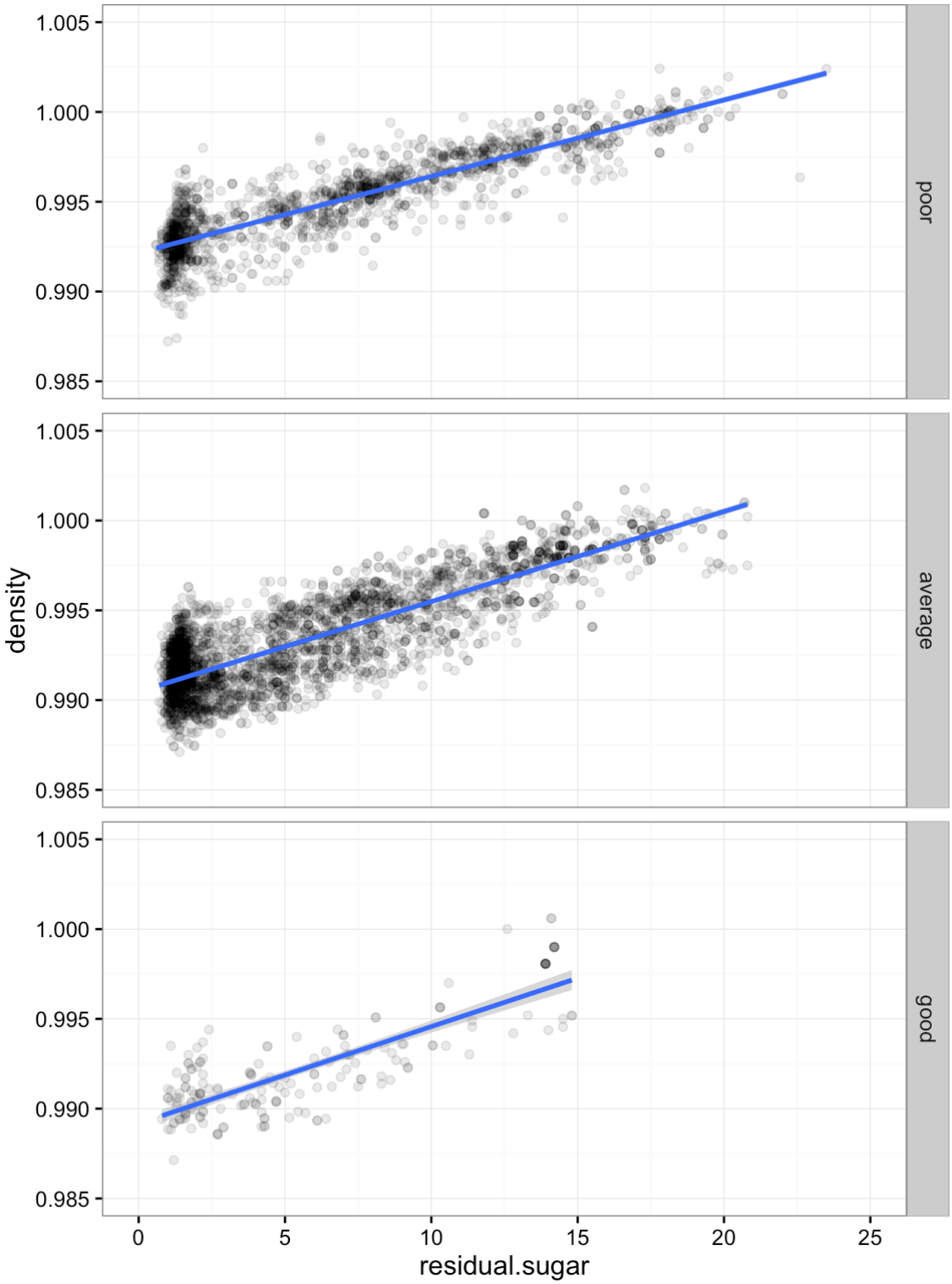
- **Relationships observed** - the main feature of interest (the quality of the wine samples) correlates weakly with the given variables.
- **Interesting relationships between other features** - several other relations are visible from the matrix of graphs. For example, density correlates with residual.sugar and with volatile.acidity.
- **Strongest relationship found** - residual.sugar and density with a correlation value of 0.839.

Multivariate Analysis

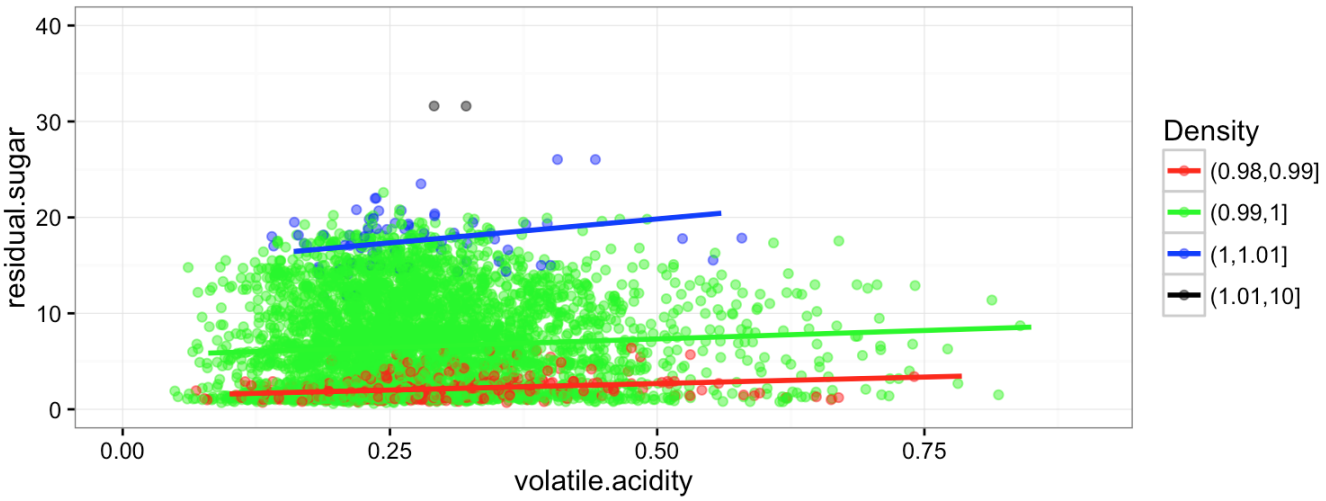
As shown in the previous section, density correlated with residual.sugar. In the next graph, I am using color code to explore if quality is also affected by this relation. For ease of visualization, I will be using a quantified version of quality on three levels:

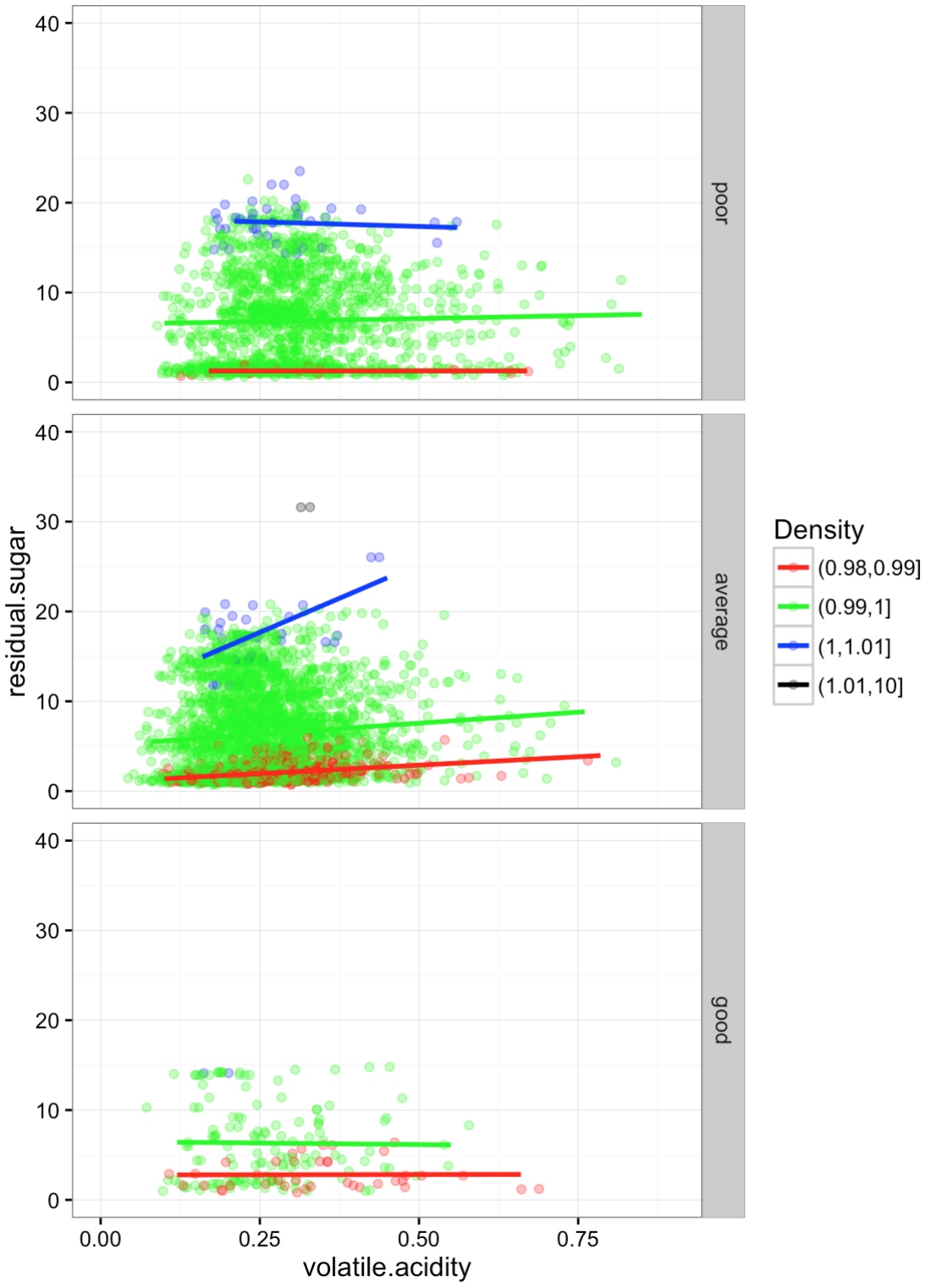

1.005
1.000
0.995
0.990
0.985
density
0
5
10
15
20
25
residual.sugar
Wine quality
poor
average
good

Let's see how the relationship holds for each of the three classes:



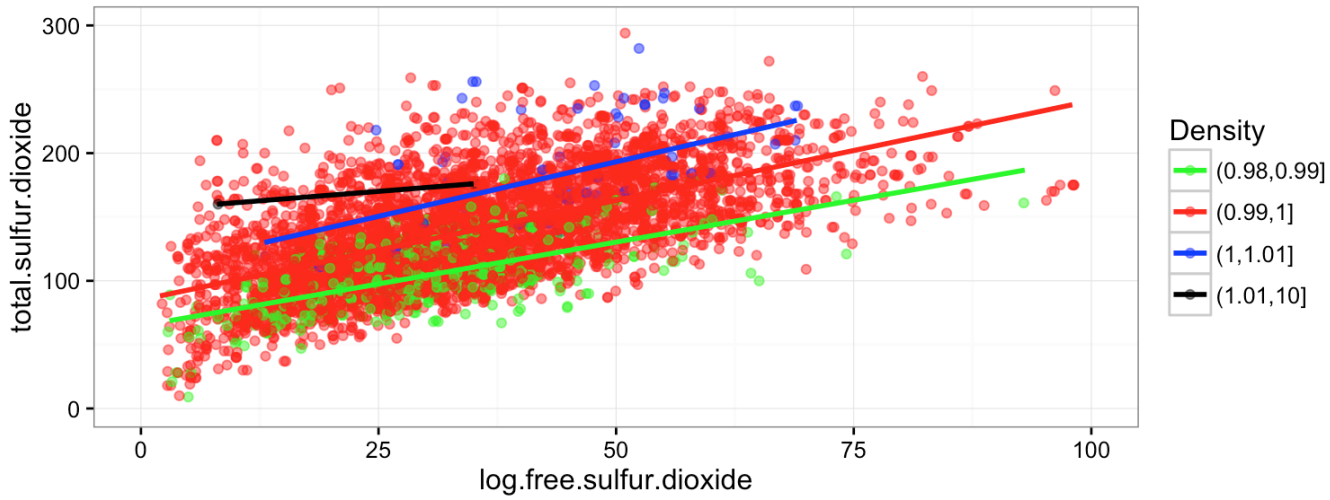
Even at this large level of magnification on the y axis a linear relationship is obvious. Additionally, the quality seems to be directly linked also with the spread of values on the x axis.

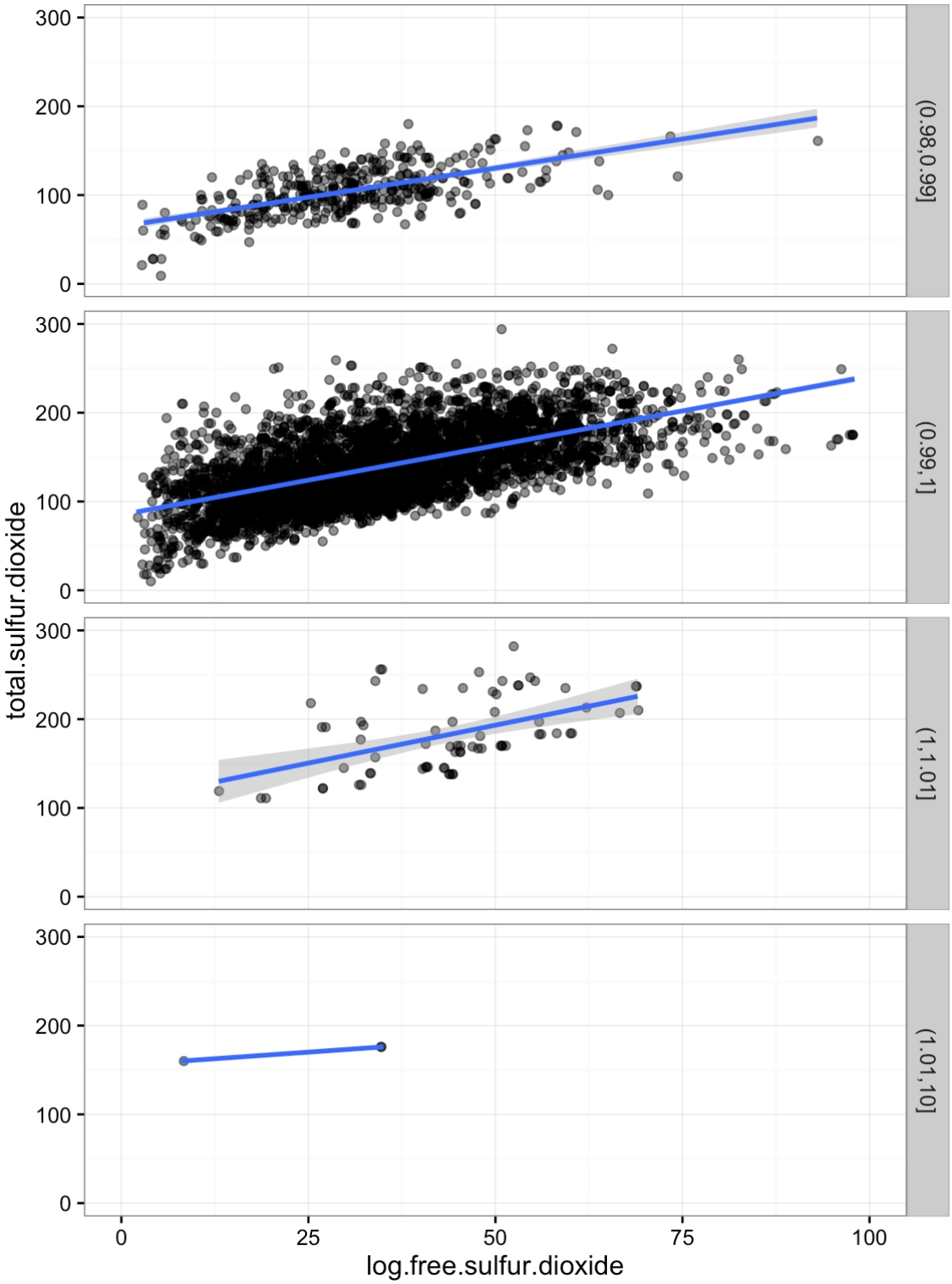

residual.sugar
volatile.acidity
Density
(0.98,0.99]
(0.99,1]
(1,1.01]
(1.01,10]


poor
average
good
residual.sugar
volatile.acidity
Density
(0.98,0.99]
(0.99,1]
(1,1.01]
(1.01,10]

The volatile.acidity adds little information - the spread on the x axis is reduced for good-quality wines and spread to the maximum for the poor wines. The relationship between residual.sugar and density is strongly visible even for the quantified version of density.

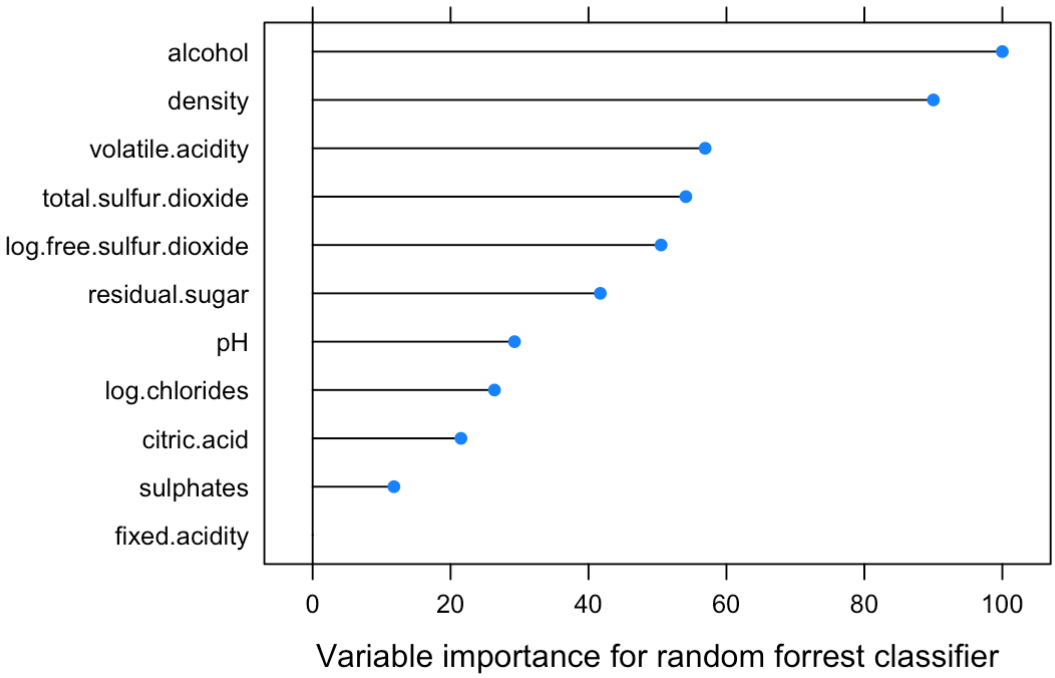
total.sulfur.dioxide is correlated with log.free.sulfur.dioxide as can be seen in the following graphs. Over-imposing the density shows a clear correlation only with the total.sulphur.dioxide. I added a linear interpolation for each of the classes:

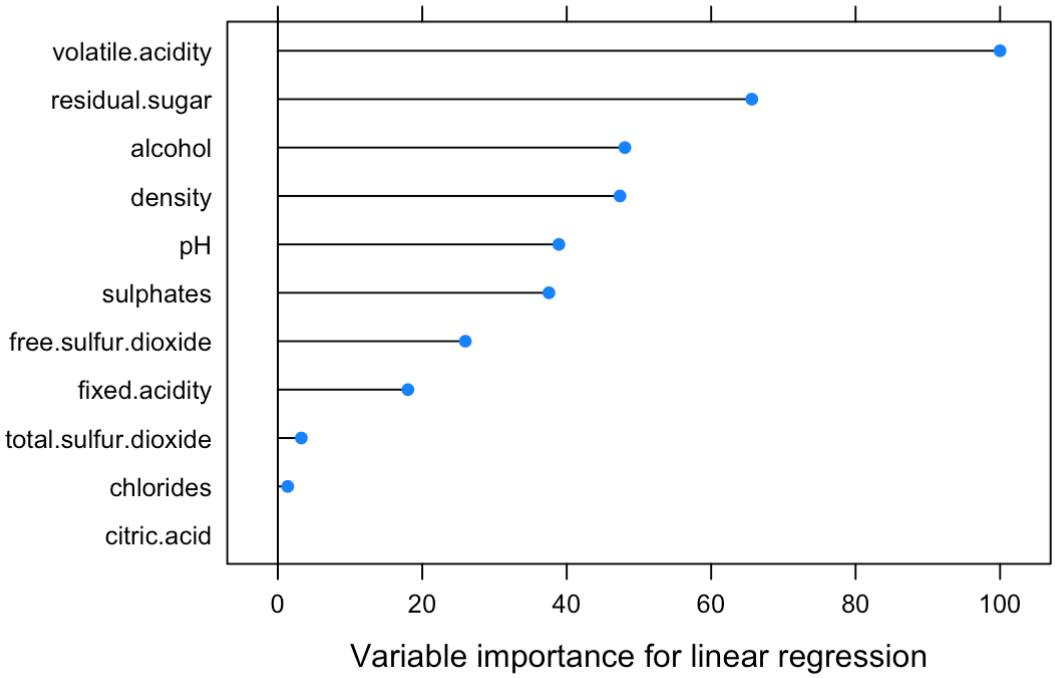




As a final step in this section, I am interested which are the important features in the dataset from a classification perspective. As the correlations with the target variable are small, I expect a different ranking for different classifiers. The graphs below show the normalized feature importances for linear regression and random forrest. As the goal of this project is not to find the best classifier available, I left all the parameters to default values. Cross validation was used to estimate the machine learning performance.

As expected, the two sets of features have a different ordering. The interesting fact is that the random forrest classifier is slightly better then the proposed SVM classifier in the original paper (relative improvement of ~10%):





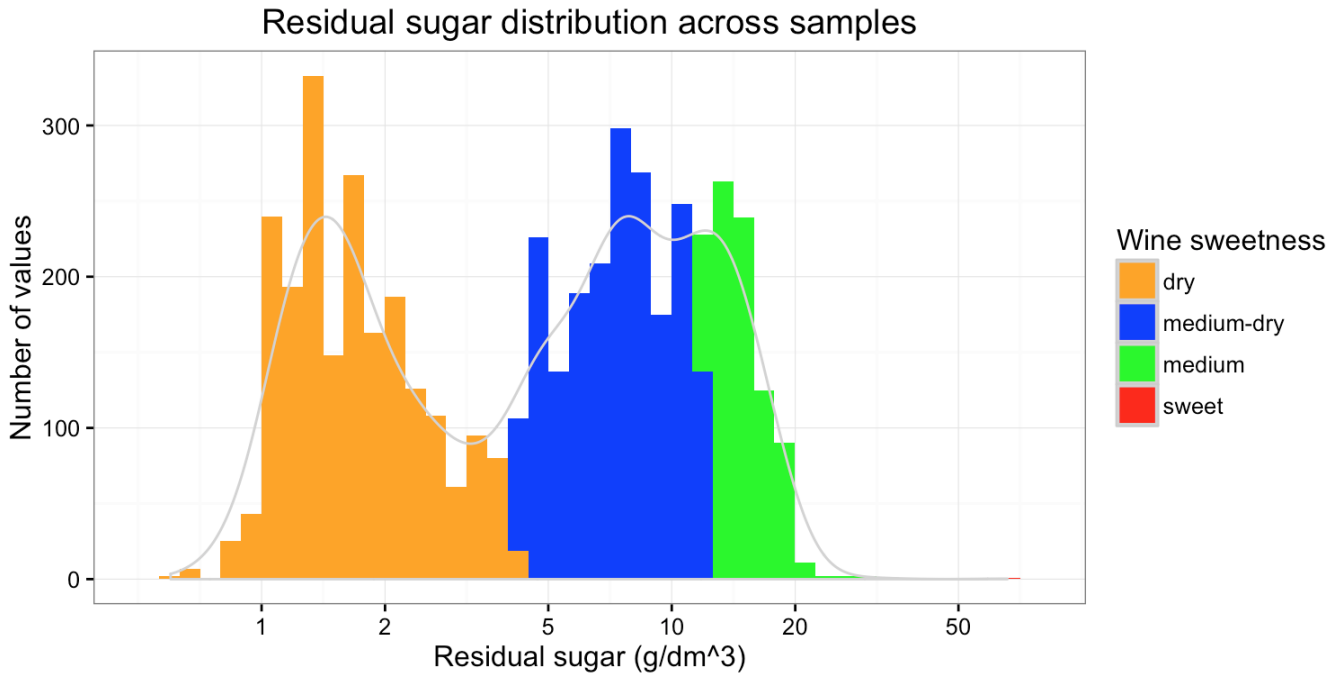
Summary

- **Relationships observed in this part of the investigation:** After surveying groups of three variables I could not find novel major interactions apart from the ones shown in the bivariate section.
- **Surprising interactions between features:** not really. All combinations showed overlapped clusters - no clear novel insight gained.
- **Models with the dataset** - yes, I trained a linear regression and a random forrest classifier. Both of them were trained for the purpose of exploring which were the most interesting features in the dataset. As expected, the ordering of the features was slightly different in the two cases. Also, as expected, at closer inspection, the random forrest exhibits a similar pattern as the SVM model in the original paper: no data is correctly classified for the classes of quality 3 and 9. The explanation is that this kind of ensemble estimator performs bad for the border classes. The averaging operation tends to bin more

values in the middle classes. Also, the classes are not equally represented and I took no steps of balancing the dataset.

Final Plots and Summary

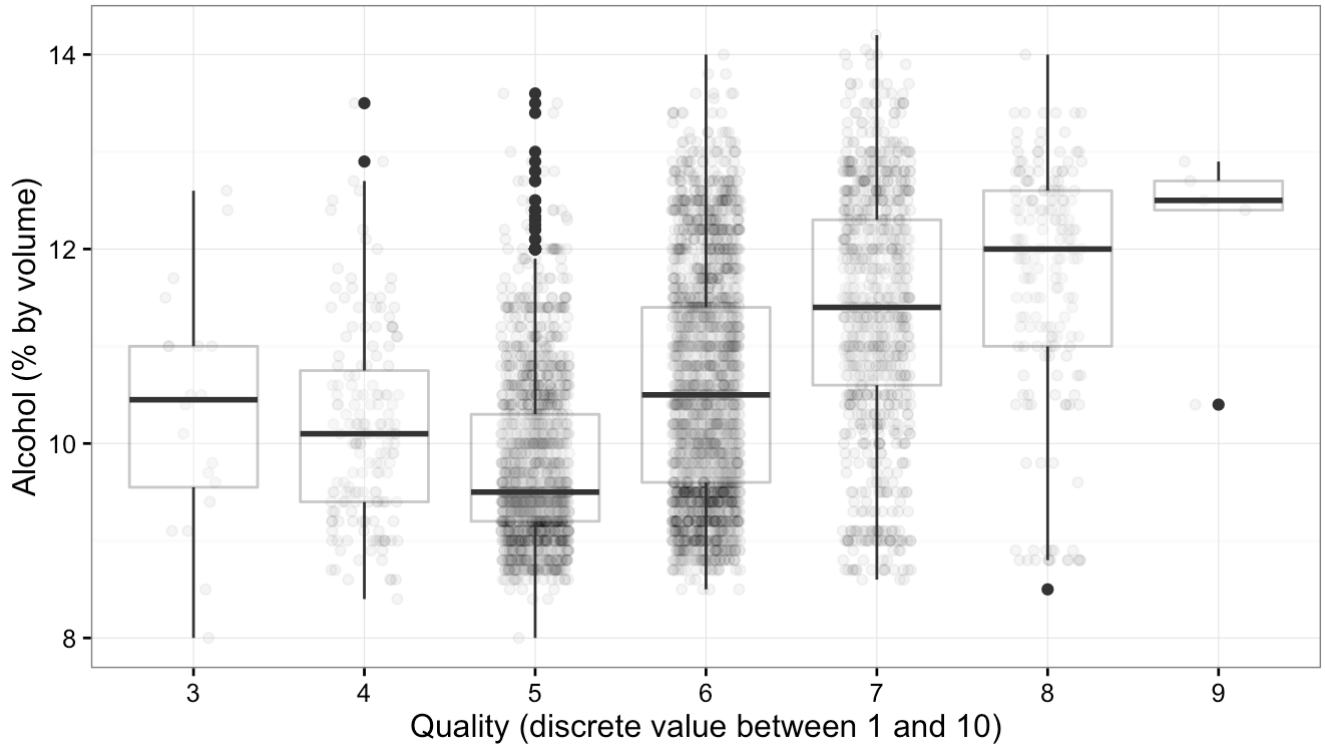
Plot One



This plot showcases the histogram and the scaled distribution of the residual sugar variable, on a logarithmic scale. Several things can be noticed:

- first, the class of sweet wines is virtually not present in the dataset (only one instance in almost 5000 samples).
- second, the thresholds taken from literature map very well the density peaks - leading to the conclusion that the thresholds for the categories were not randomly chosen. Experts certainly have a deeper understanding of how much sugar can actually be present in wines.
- third, the dry wines are a majority in this data set (even visible on the logarithmic scale).

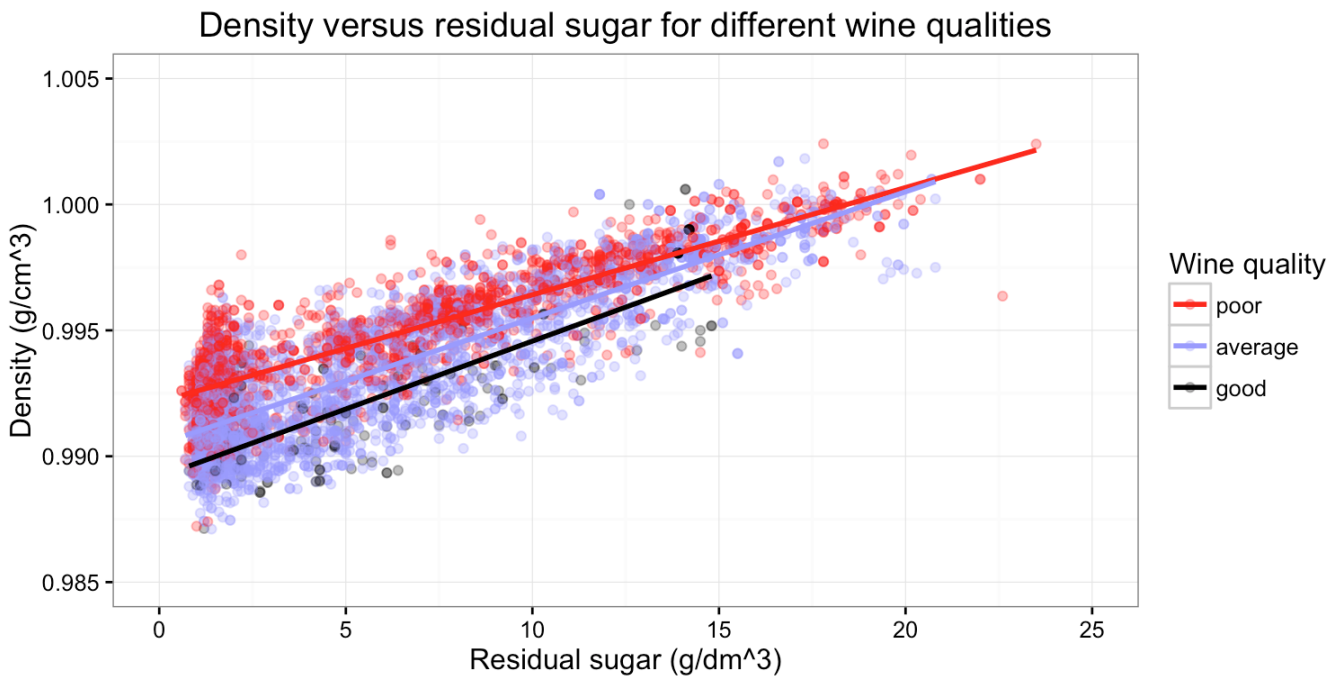
Plot Two



This plot shows the distribution of alcohol values across various quality thresholds. The graph supports the general opinion that, in general, wines with higher quantity of alcohol are perceived as being of higher quality.

The graph indirectly shows the large variance in the distribution of the samples across the classes, where for quality 3 and 9 very few samples are available. It also shows that the correlation between alcohol and quality is moderate - a large number of outliers is present in the figure and almost all quality bins span across the whole alcohol range.

Plot Three



In this plot I am trying to show the strong correlation between density and residual sugar. Although the graph shows quite some variation, the y axis does not start at 0 - changing it would result in an almost straight line, making this point stronger. By dividing the wine quality in three categories and using a linear regression model

for each class, we notice a clear ordering of the cluster points from the three classes, with better wines having a lower density and less residual sugar.

Reflection

In this project I have analyzed the white wines dataset provided by udacity. Visualization of the variables in the dataset helped a lot grasping more understanding of the information hidden within.

My major struggle was to find a direct correlation between the quality of the wine and the given variables. Looking at the correlation plots, it seems that the quality is an insignificant variable. My first instinct was to dismiss it with an excuse such as: “tasting wines is as subjective as any human action can be”. Then I paid a bit of attention on how data is collected: the opinions of three reviewers are averaged. So, the actual human observations are modified by two quantification steps (original marks given by reviewers and the average transformed to an integer mark) and one averaging step (which removes samples from the border categories). Once this was clear, then I understood why the authors of the paper were forced to further average the quality marks in order to boost the performance of their classifier.

The other major struggle was to find combinations of two variables that show clear clusters of quality-related data. This struggle was not fully satisfied - alhough clustering is to be found in the graphs in the multivariate section, clear distinctions between the classes is not present. Peeking again at the original paper, it seems that the authors also failed to identify any at all.

General qualitative observations from wikipedia state that, while accounting for exceptions, superior wines have high percentage of alcohol and low levels of sugar. This observation matches somewhat the data. Also, it is noted that the perceived sweetness can be modified for example by acidity (I also wonder about temperature, given that we deal with white wine). All in all, some chemistry knowledge would probably help creating a better variable for perceived sweetness. As future work, I would recommend searching the literature on how the chemicals in the data set modify perceived sweetness and build a variable based on the findings. I would be astonished if the popular knowledge is not verified by the data.