Wine Chemistry and its Quality

Rita Sánchez

February 2022

Capstone: Wine Quality Project

This report is the last stage of the HarvardX PH125.9x course. The goal of this project is to work with several machine learning algorithms to determine wine quality using its physicochemical characteristics.

1. Introduction

Wine is part of the Mediterranean culture. There is no family gathering or meal with friends in which wine does not play an important role. In last years, there has been a boom in what can be called "wine culture": visits to wineries are organized to learn about its elaboration process or attending wine tasting workshops, in which you try to learn how to appreciate its different facets. In short, the aim is to determine what makes one wine better than another and why.

This project attempts to answer, briefly, this question by analyzing the different chemical compounds in wine. With this purpose, we use a wine quality dataset from Kaggle [1]. As predicting wine can be considered a classification problem, we will apply some machine learning classification techniques.

This work is organized as follows: in Section 2, we apply exploratory data analysis techniques; in Section 3, we briefly discuss the proposed methodology, followed by model comparison; and in Section 4, the main finding and conclusion are shown.

2. Data exploration

We load the data and we name our data base data as *wine_data*

```
wine_data <- read.csv("~/Documents/Curso_DataSience(edX)/9-Capstone Project/My own project/wine_quality
```

Now, we can check the dataset structure to know which type of data we are going to work with.

```
# any NA value to clean
any(is.na(wine_data))
# [1] FALSE  ==> (no NA in the dataset)

# a first view of the data
str(wine_data)

data.frame':    1599 obs. of  12 variables:
 $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
 $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
 $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
```

[1] https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009/version/2 P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

```
$ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
$ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
$ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
$ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
$ density             : num  0.998 0.997 0.997 0.998 0.998 ...
$ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
$ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
$ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
$ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

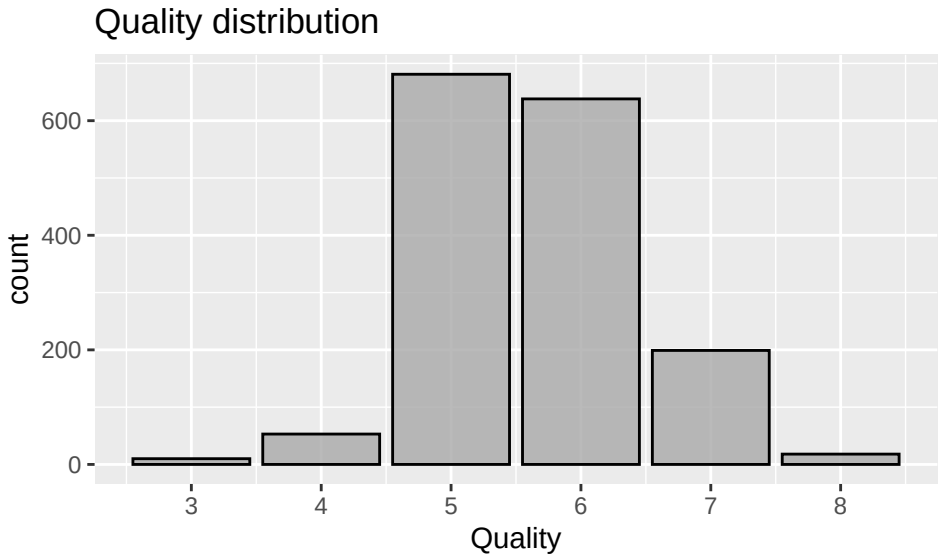
So, our data set is a tidy data frame with 1599 observation in each of the 12 variables and with no NA values.

What attribute is each variable?

- *Fixed acidity*: The fixed acidity is the set of natural acids in wine and it starts in the vineyard, already. Our mouths react instinctively to acidity levels. Hold your mouth open after you sip. If you begin salivating, your mouth is reacting to the acid. The more saliva, the more acid.The predominant fixed acids found in wines are tartaric, malic, citric, and succinic. Their respective levels found in wine can vary greatly but in general one would expect to see 1 to 4 g/dm^3 tartaric acid; 0 to 8 g/dm^3 malic acid; 0 to 0,5 g/dm^3 citric acid; and 0,5 to 2 g/dm^3 succinic acid. (g/dm^3)
- *Volatile acidity*: amount of acetic acid in wine(g/dm^3), that is considered a fault at higher levels (1,4 g/dmˆ3 in red and 1,2 g/dmˆ3 in white) and can smell sharp like nail polish remover or vinegar. Long fermentation (1 month or more), accumulate higher levels of volatile acidity.
- *Citric acid*: found in small quantities can add some freshness and flavor (g/dm^3)
- *Residual sugar*: amount of sugar remaining after wine fermentation/production (g/dm^3). Accordingly to the dataset documentation, it's rare to find wines with less than 1 g/dm^3, while dry wines range from 1 - 3 g/dm^3 . On the other hand, wines with more than 45 g/dm^3 of sugar are categorized as sweet.
- *Chlorides*: amount of salt in the wine (g/dm^3)
- *Free sulfur dioxide*: free forms of S02, prevents microbial growth and the oxidation of wine (mg/dm^3)
- *Total sulfur dioxide*: amount of free and bound forms of S02 (mg/dm^3)
- *Density*: the density of water depending on the percent alcohol and sugar content (g/dm^3)
- *PH*: describes how acidic or basic a wine is on a scale 0-14 (very acidic: 0, very basic: 14); most wines are between 3-4 on the pH scale. The pH level tells us, also, how intense the acids taste: the higher the ph, the lower the acidity of wine, and vice versa.
- *Sulphates*: salts derived from sulfuric acid, which are basically used in fertilizers, pesticides and pesticides (potassium sulphate in g/dm^3). The maximum acceptable limit in wine recommended by the International Organisation of Vine and Wine is 1 g/L
- *Alcohol*: the percent alcohol content of the wine (% of volumen)
- *Quality*: target variable (based on sensory data, score between 0 and 10)

2.1. A look at each attribute

Let start with *quality* , the attribute we employ to define how good or bad is a wine.

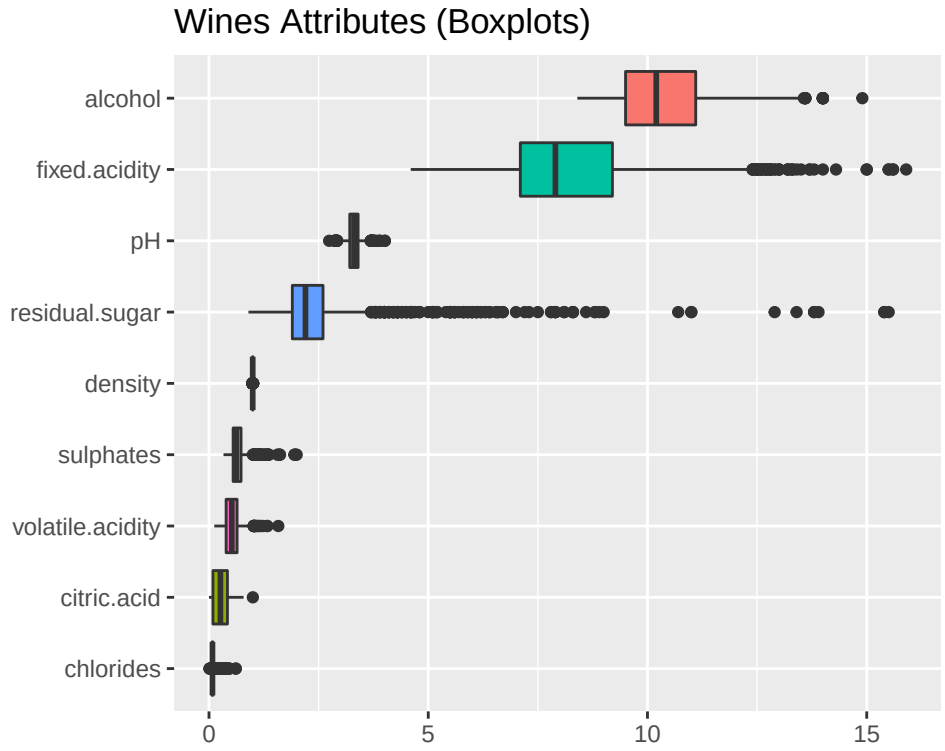


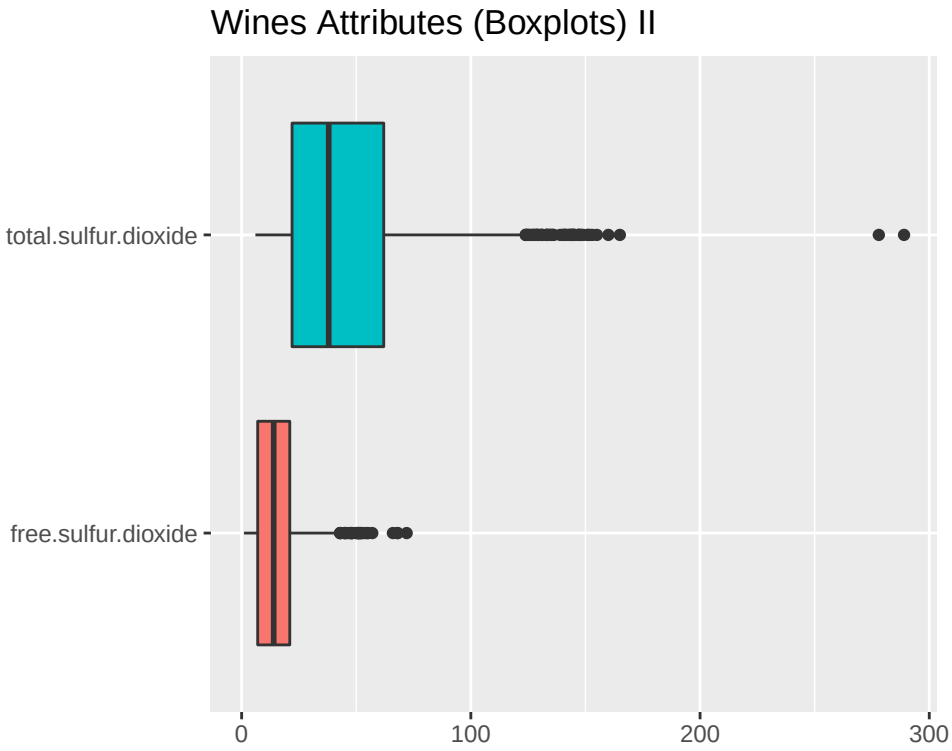
One interesting thing we observe is that are very few wines considered as *bad wines* (i.e., with a quality below 4), a bunch that can be defined as *really good* (with a 7 quality) and is almost anecdotal to find an excellent wine (i.e., with a quality above 8). The vast majority of the wine in this dataset can be considered as *normal* (5 in quality) or *good* (6 in quality).

Considering that quality is the target attribute, we will try to find some relationship with the rest of the attributes in our dataset. So, first, let's take a look on the distribution of the other dataset variables.

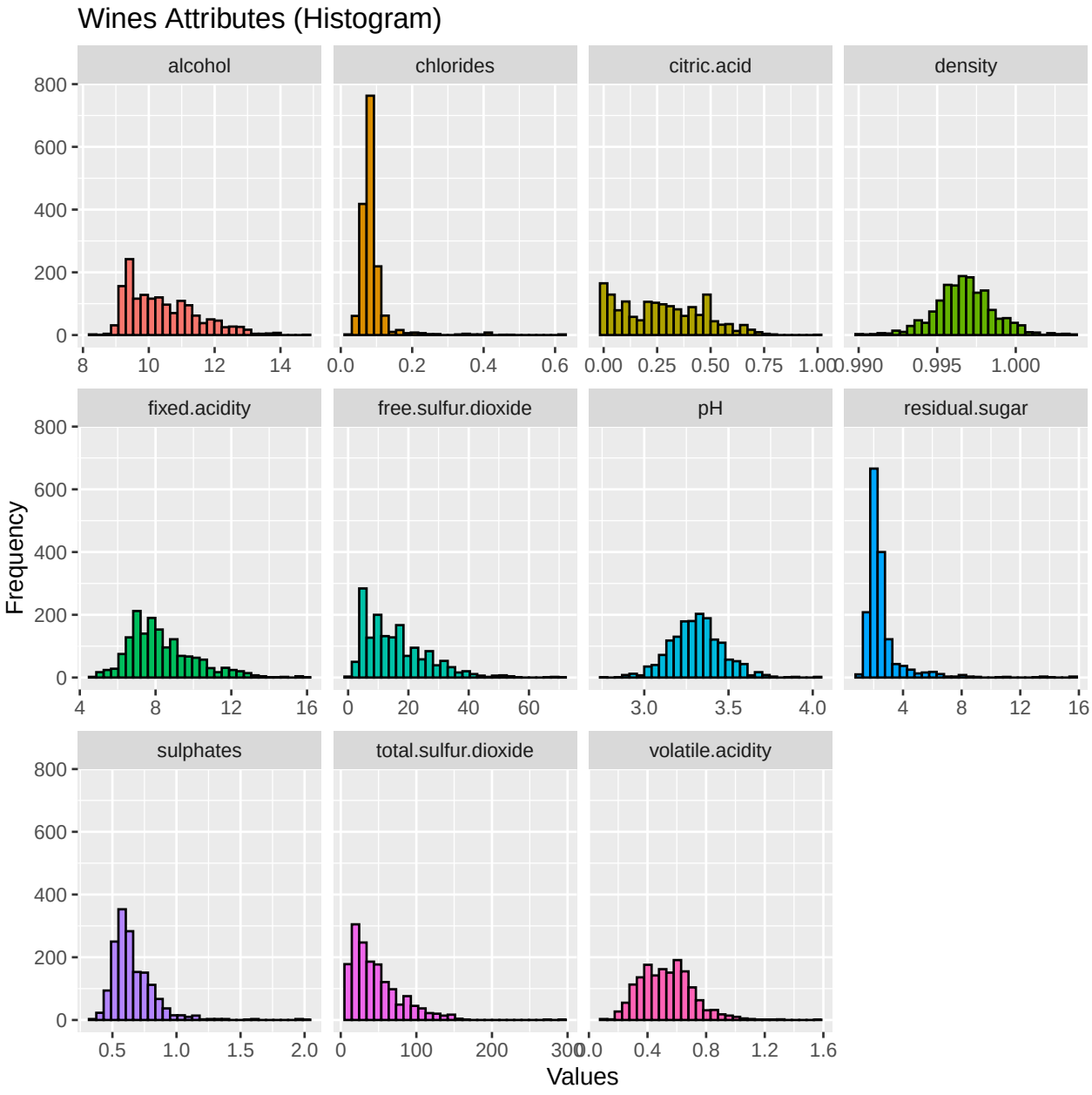
Summary report

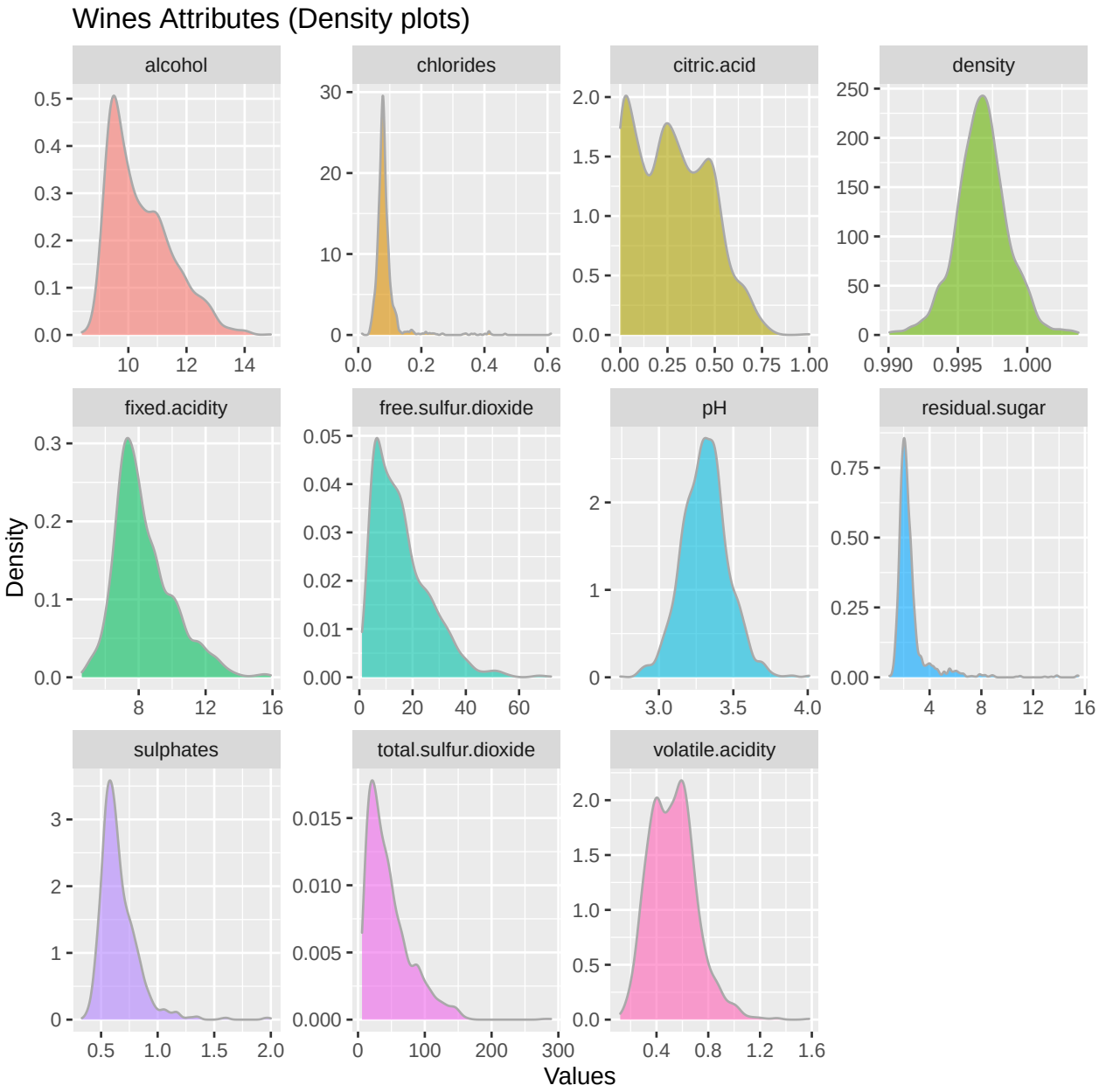
```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean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





Note: the attributes total.suful.dioxide and free.sulfur.dioxide appear in a separate plot due to the presence of very high values that do not allow them to be represented together with the rest of the parameters in the dataset.


Wines Attributes (Histogram)
alcohol
chlorides
citric.acid
density
fixed.acidity
free.sulfur.dioxide
pH
residual.sugar
sulphates
total.sulfur.dioxide
volatile.acidity
Frequency
Values



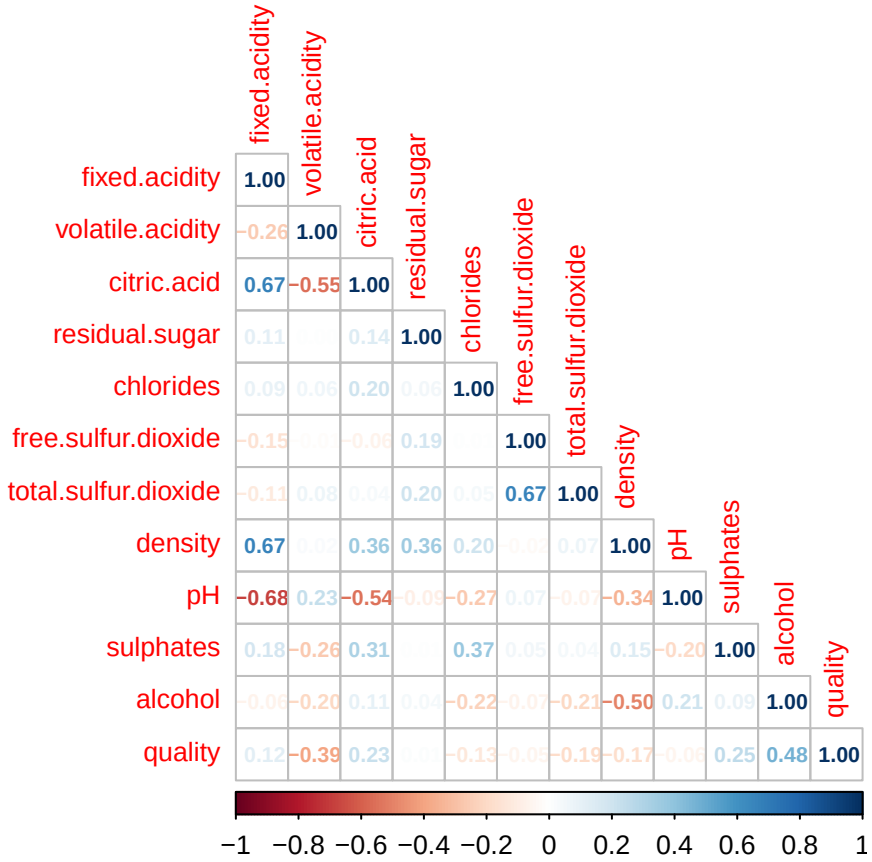
The main findings from the summary and the plots above are summarized as follows:

- *Alcohol*: few outliers and with a distribution concentrated on 9,0%.
- *Sulphates*: distribuited around the mean but showing some skewness, indicating a high number of outliers.
- *Sulfur dioxide*: this component can be found in its free-form or in bounded forms with other molecules.In the plots above, it is observed that both variable *free.sulfur.dioxide* and *total.sulfur.dioxide* follow the same patterns. In addition, the summary suggests there must be big outliers on data.That's why, these attributes are plotted in a separate boxplot from the rest.
- *Density*: very narrow range variation... in fact, the difference between min and max is less than 0,01... maybe too small to be perceived by an ordinary human.
- *Chlorides*: very skewed distribution with the majority of the values below 0,1 and with outliers that "squash" the density function.

- *Residual sugar*: Considering that our density plot shows that almost 90% of the distribution is below 4,0, our dataset refers mainly to dry wines, but there are certain outliers.
- *Citric acid*: from the plots above, we can conclude that there are a wide range of "freshness" in our wine selection.
- *Volatile acidity*: The plots show that above 95% of the distribution is below 1,2, so the majority of our dataset are unoaked white wines
- *Fixed acidity*: presents values in a higher scale that the other two "acidity" attributes, as well as more outliers.
- *PH*: plots are consistent with a quasi-normal distribution. They also show that it varies between 3,0 and 3,6, with few exceptions. These values corresponding with the different types of white wine.

2.2. Searching relationships among variables

One we have explore the characteristics of each variable, now we are interested in exploring the relationships among all variables in the dataset, directing the analysis to find out which characteristics are more related to the quality score. A correlation matrix will give this information.



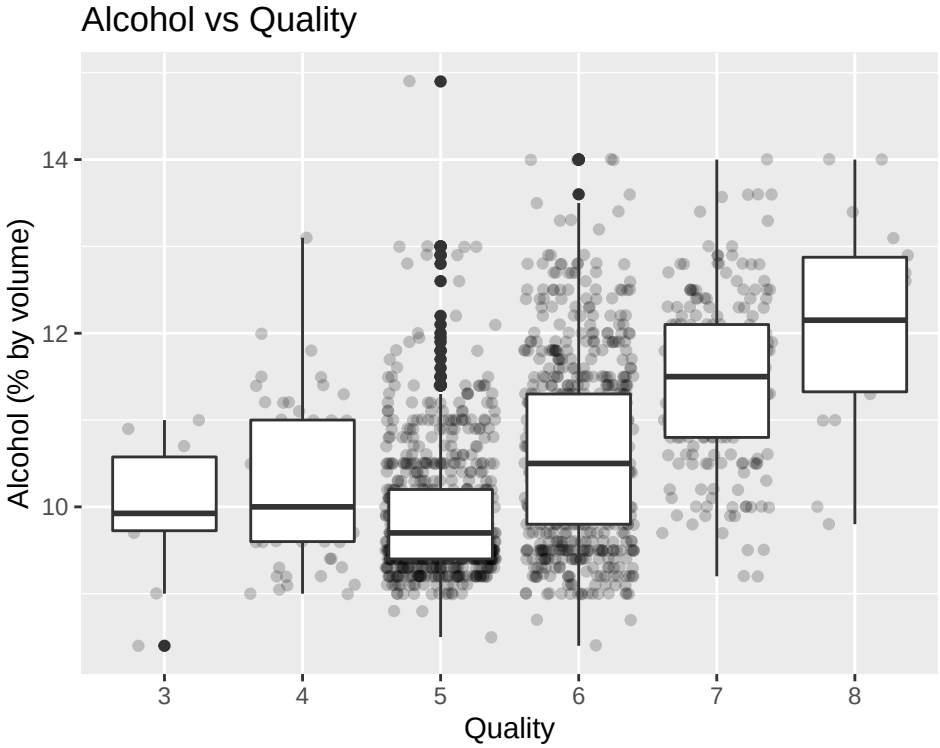
From the correlation matrix we notice the most relevant attributes to quality are **alcohol** and the **volatile acidity**. We also observe some other relevant correlation among other variables and, although they don't contribute directly to the quality, it will be interesting to have a quick look on them.

Analyzing the correlation matrix and considering the previous analysis for each attribute, we can draw some assumptions:

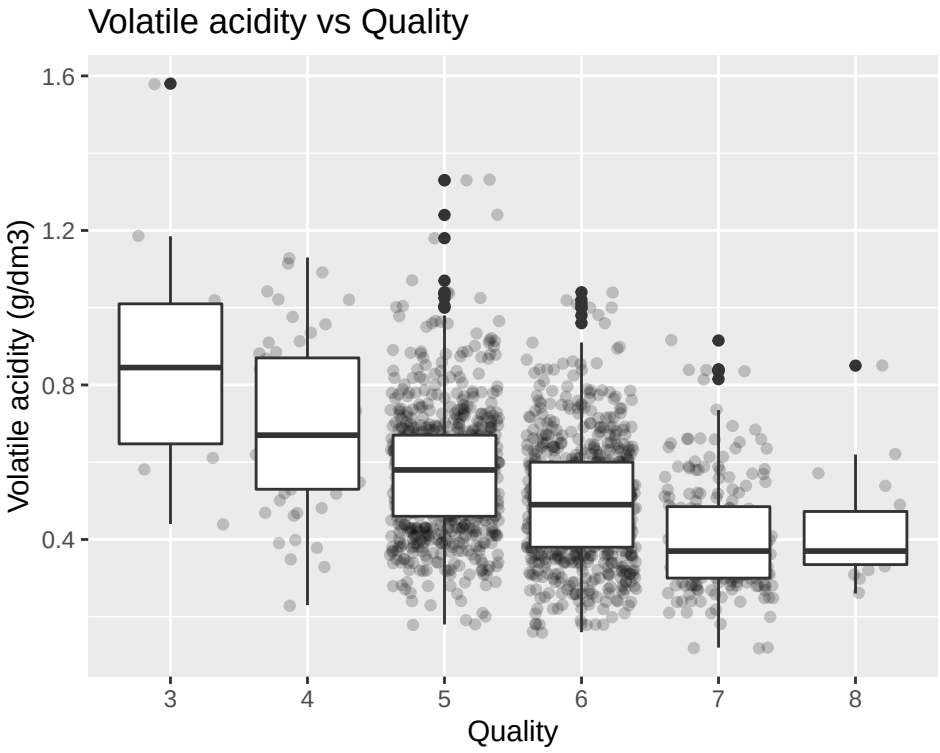
- The wine quality is positively correlated to alcohol and negatively correlated to volatile acidity. It makes sense with our previous analysis, since the volatility acidity is responsible for an unpleasent taste in the wine.
- It is observed a significant correlation between density and other attributes such as (from high to low) fixed acidity, alcohol, citric acid and residual sugar.
- Ph is negatively correlated to acidity, since lower values in the ph scale means a higher acidity. As it is a natural fact, it has no sense to explore here this relationships.
- We detect some correlation between sulphates and chlorides, but none of them presents a relevant correlation with quality.
- Obviously, free and total forms of sulfur dioxide will be related to each other and the same stands for the different acidity types. Regarding this last, we will consider only the volatile acidity in next section due to its high correlation to quality.

2.2.1. Correlation to quality

As we have seen in the correlation matrix, *alcohol* and *volatile acidity* are the characteristic with the highest correlation to *quality*. Let's see these relationships with the corresponding boxplots.



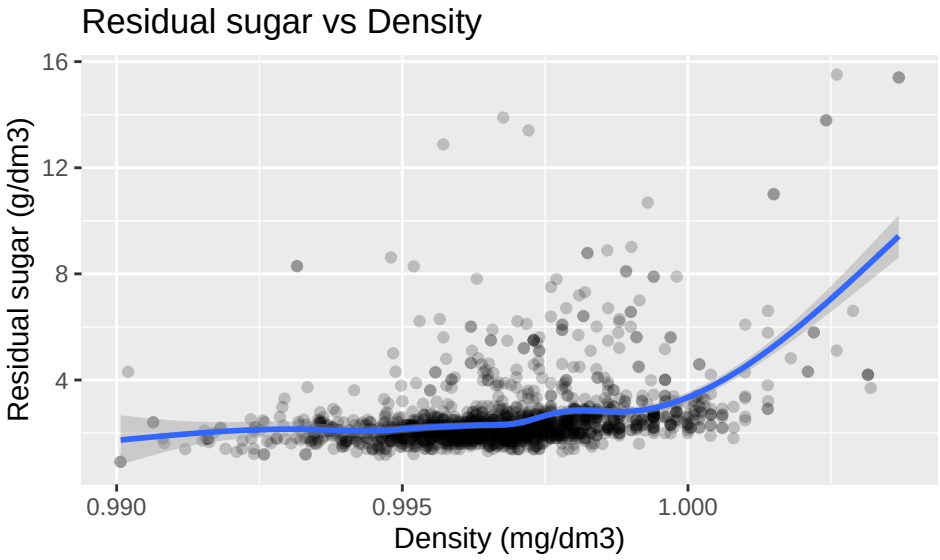
According to this plot, up to a quality level of 6, the average alcohol percentage is around 10%. It is the higher quality levels (7 and 8) that show a significant increase in their average alcohol percentage. Thus, we can point out that relevant correlation between alcohol and quality arises only when we are dealing with the bordering up values.

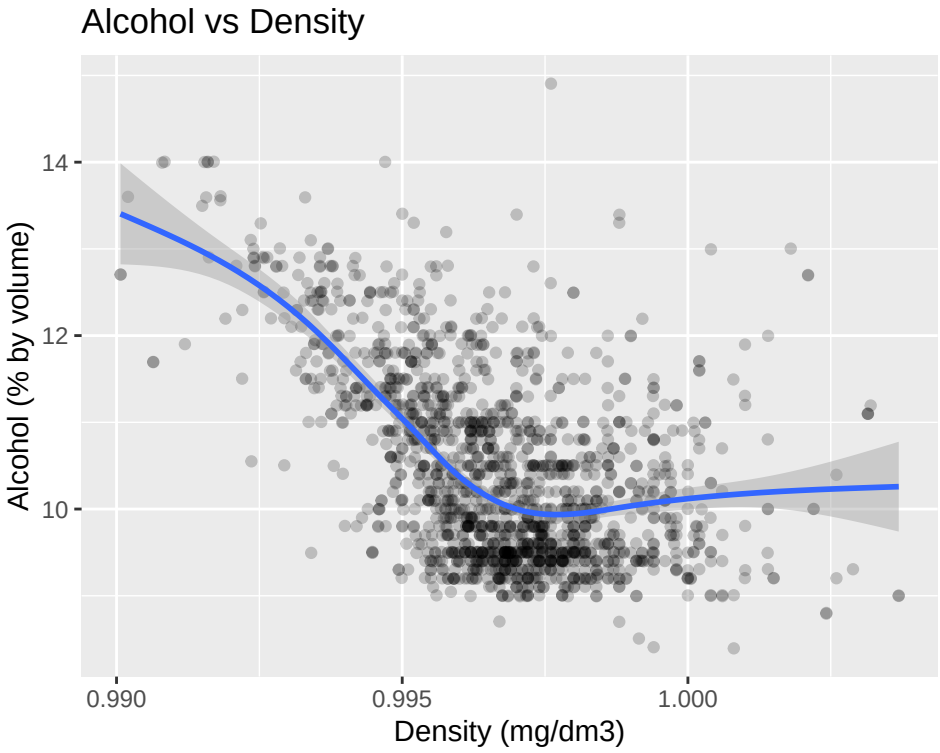


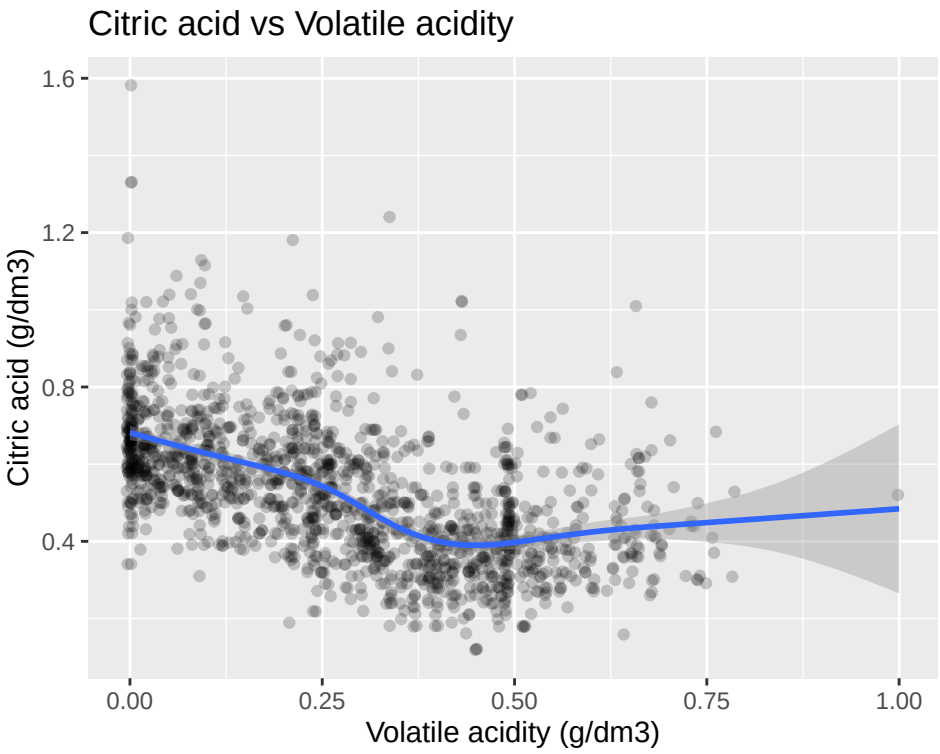
This boxplot shows that the higher the quality, the lower the acidity. Nevertheless, it is not possible to discriminate between the two highest level of quality on the basis of volatile acidity.

2.2.2. Other correlations

Now, we plot other attributes that may have some impact on quality in "second round": (i) residual sugar is correlated with density and, in turn, (ii) density is highly correlated with alcohol; and (iii) citric acid is correlated to volatile acidity.




Alcohol vs Density
Alcohol (% by volume)
14
12
10
0.990
0.995
1.000
Density (mg/dm3)


Citric acid vs Volatile acidity
Citric acid (g/dm3)
1.6
1.2
0.8
0.4
0.00
0.25
0.50
0.75
1.00
Volatile acidity (g/dm3)

3. Modelling

During the exploratory data analysis it has seen that alcohol and volatile acidity are the best attributes to predict wine quality. In this section we discuss several modeling approaches with machine learning to predict wine quality. For that,we generate our training and test set by randomly splitting the data using the following code:

```
set.seed(123)
n <- nrow(wine_data)
trainIndex <- sample(1:n, size = round(0.7*n),replace=FALSE)

training <- wine_data[trainIndex,]
testing <- wine_data[-trainIndex,]
```

So, our training set consist on the 70% of our total dataset (*wine_data*) and the remaining 30% is our test set. Now we can develop different algorithms to predict the wine quality.

To evaluate the performance of the models we will use the followings metrics:

- **Mean Absolute Error (MAE)**. It represents the average of the absolute difference between the actual and predicted values in the dataset. It measures the average of the residuals in the dataset.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$

- **Root Mean Squared Error (RMSE)**. It represents the squared root of the average of the squared difference between the original and predicted values in the data set. It measures the standard deviation of residuals.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

- **R-squared ($\mathbf{R^2}$)**. It represents the proportion of the variance in the dependent variable which is explained by the regression model. The value of R square will be less than one.

$$R^2 = \frac{\sum_{i=1}^{N}(\hat{y} - \bar{y})^2}{\sum_{i=1}^{N}(y - \bar{y})^2}$$

Where, $\hat{y}$ is the predicted value of y, and $\bar{y}$ is the mean value of y

The lower value of MAE and RMSE implies higher accuracy of a model. However, a higher value of R square is considered desirable as well.

3.1. Linear regression

We developed a linear regression model to predict wine quality (y) as a function of the differents attributes in the dataset. Since the variables have very different magnitudes, it will be appropriate to normalize these variables in order not to distort the results. Once we have the normalized variables, we estimate the following equation.

$$y = \beta * X$$

where X is a matrix with 11 columns, one for each attribute analyzed in the previous section normalized.

The results obtained are summarized below:

```
##
## Call:
## lm(formula = quality ~ ., data = training_sc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.98580 -0.48419 -0.05934  0.56417  2.43999
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.00555    0.02465   0.225   0.8219
## fixed.acidity          0.01198    0.06939   0.173   0.8630
## volatile.acidity      -0.22794    0.03322  -6.862 1.12e-11 ***
## citric.acid           -0.04277    0.04418  -0.968   0.3332
## residual.sugar         0.01132    0.03271   0.346   0.7295
## chlorides             -0.12056    0.02904  -4.152 3.55e-05 ***
## free.sulfur.dioxide    0.05069    0.03432   1.477   0.1400
## total.sulfur.dioxide -0.14531    0.03715  -3.912 9.72e-05 ***
## density               -0.02852    0.06200  -0.460   0.6455
## pH                    -0.11474    0.04483  -2.560   0.0106 *
## sulphates              0.18602    0.02983   6.236 6.39e-10 ***
## alcohol                0.38992    0.04308   9.052  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8229 on 1107 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3534
## F-statistic: 56.55 on 11 and 1107 DF,  p-value: < 2.2e-16
```

The summary above shows that alcohol have a strong positive relationship with quality, implying that more alcohol will translate into a higher quality of wine. Reversely, there is a strong negative relationship between volatile acidity and quality that means that lower volatile acidity levels corresponding with higher quality of wine. The coefficients for the remaining attributes are in line with the results obtained in the correlation matrix. In short, these results are consistent with the conclusions obtained from the exploratory analysis of our dataset.

The metrics to evaluate the model are:

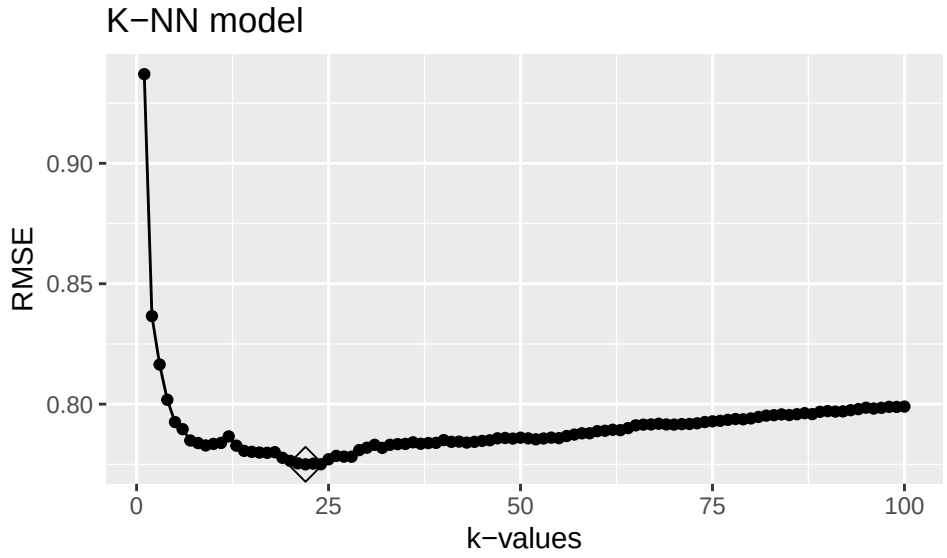
metric	value
RMSE	0.59
MAE	0.76
R_2	0.40

3.2. K Nearest Neighbors Regressor

K-NN algorithm is a supervised machine learning algorithm than can solve classification and regression tasks. The k-NN algorithm uses feature similarity to predict the values of any new data points. It classifies each new data in the corresponding group, depending on whether it has k neighbors closer to one group or another. In other words, it calculates the distance of the new element to each of the existing ones, and sorts these distances from smallest to largest to select the group to which it belongs. This group will be, thus, the one with the highest frequency and the smallest distances.

Before implementing the k-NN Regressor, we need to scale the features. With the k-NN algorithm, we

measure the distance between the pair of samples that are influenced by the measurement unit. To avoid this, we should normalize the data before implementing the algorithm.



We found an optimum model at $k = 22$, where the RMSE is minimum. This is the k value we use to validate the test set. The evaluation of model performance is as follow:

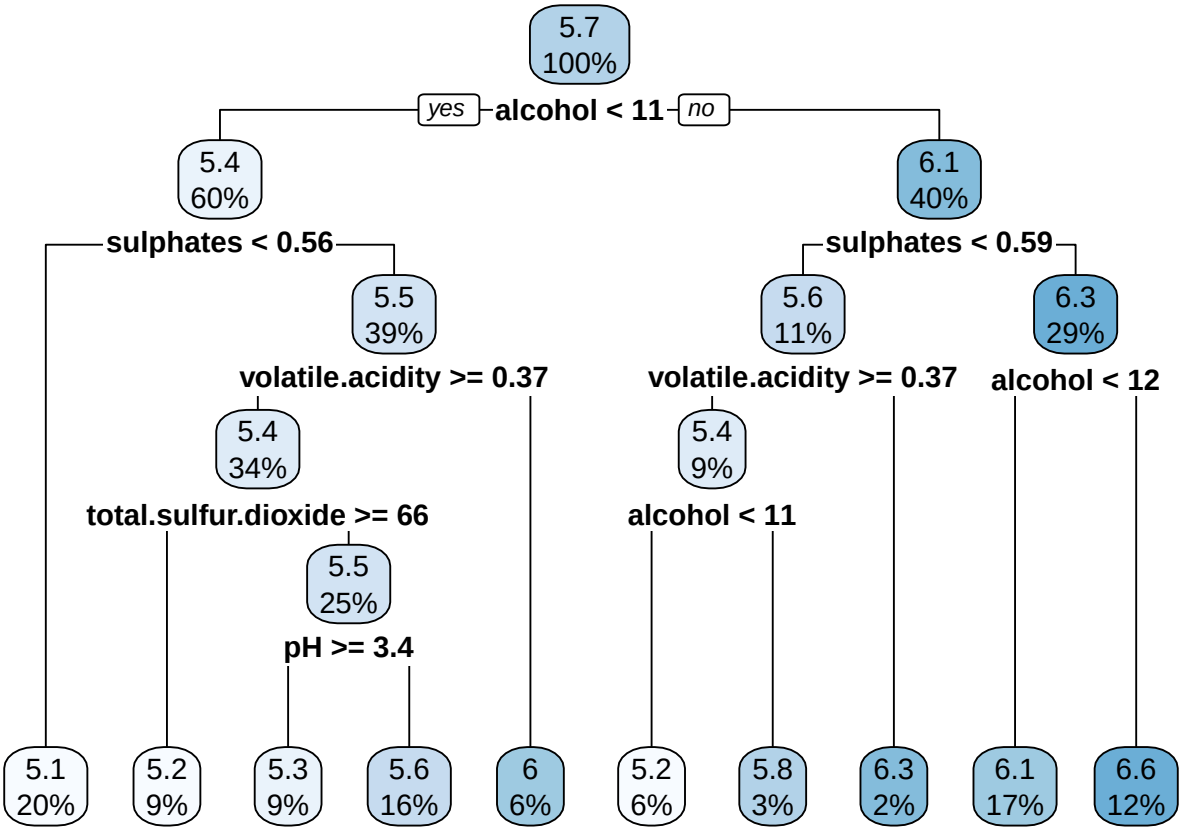
```
##            Reference
## Prediction   3   4   5   6   7   8
##          3   0   0   0   0   0   0
##          4   0   0   0   0   0   0
##          5   3  10 176  81  12   0
##          6   0   4  40 109  27   2
##          7   0   0   1   5   9   1
##          8   0   0   0   0   0   0
```

From the result above we notice that this method is not too good in covering the extreme wine quality qualifications. The validation metrics are as follow:

metric	value
RMSE	0.63
MAE	0.79
R_2	0.10

3.3. Regression Tree

The general idea is to build a decision tree and, at the end of each node, obtain a predictor.



From the tree above we notice that our predictions barely covering the extreme qualities (both lowest or highest). This idea is reinforced with a look at the summary statistics:

```
##      actual
## pred   3  4  5  6  7  8
##   5.1  0  4 69 23  0  0
##   5.2  1  3 57 22  0  0
##   5.3  0  0 16 18  1  0
##   5.6  2  4 46 42  4  0
##   5.8  0  1  4  5  1  0
##   6    0  0  7 14  4  0
##   6.1  0  2 14 48 12  2
##   6.3  0  0  0  8  2  0
##   6.6  0  0  4 15 24  1
```

Summary statistics of the quality predicted

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.089   5.155   5.643   5.610   6.058   6.576
```

Summary statistics of the true quality

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.583   6.000   8.000
```

And here, the metrics to evaluate the model

metric	value
RMSE	0.48
MAE	0.65
R_2	0.40

3.4. Random Forest

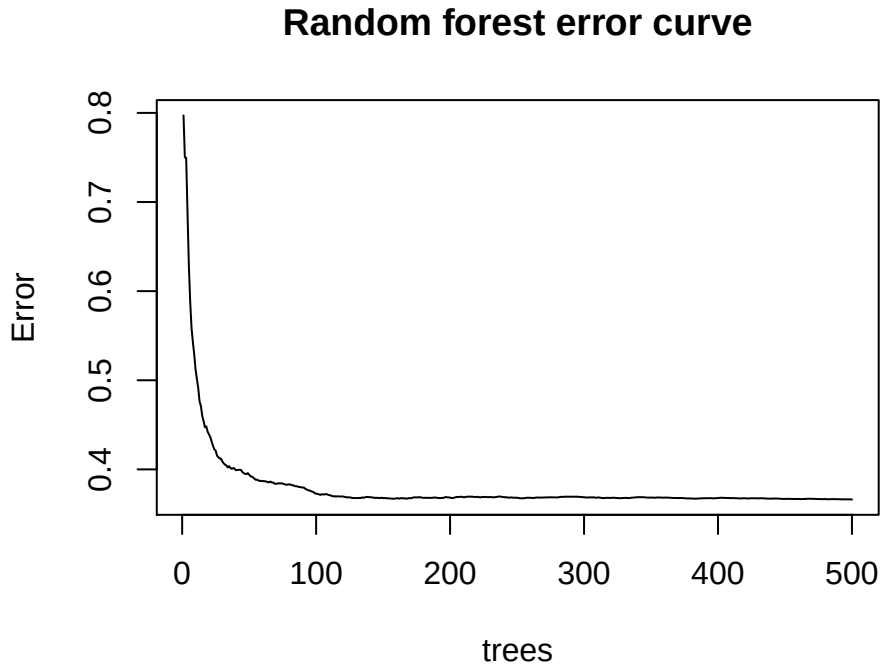
This algorithm is a very popular machine learning approach that addresses the shortcomings of decision trees. The goal is to improve prediction performance and reduce instability by *averaging* multiple decision trees. Each decision trees is trained with a random sample drawn from the original training data by bootstrapping. This means that each tree is trained on slightly different data.

The general idea is to generate many predictor, each using regression (or classification trees) and then forming a final prediction based on the average prediction of all these trees.

Let's build the model and take a look at it:

```
##
## Call:
##  randomForest(formula = quality ~ ., data = training)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##           Mean of squared residuals: 0.3662558
##                     % Var explained: 46.33
```

We can see that 500 trees were built and the model randomly sampled 3 predictors at each split. In the following plot, we can see also the number of trees that minimizes the error



Now we test the model on the test data set:

```
##      actual
## pred   3  4  5  6  7  8
##   4.6  0  0  1  0  0  0
##   4.7  1  0  2  0  0  0
##   4.9  0  0  6  0  0  0
##   5    0  3 24  1  0  0
##   5.1  0  3 49 14  0  0
##   5.2  0  1 27  6  0  0
##   5.3  1  2 23  3  0  0
##   5.4  0  2 17 13  0  0
##   5.5  1  2 23  9  0  0
##   5.6  0  0 11 16  0  0
##   5.7  0  0 11 24  0  0
##   5.8  0  0 11 25  4  0
##   5.9  0  0  3 25  3  0
##   6    0  0  3 17  2  0
##   6.1  0  0  2 10  2  0
##   6.2  0  0  0  8  4  0
##   6.3  0  1  1 11  5  0
##   6.4  0  0  3  7  6  0
##   6.5  0  0  0  3  5  1
##   6.6  0  0  0  2  5  0
##   6.7  0  0  0  1  2  1
##   6.8  0  0  0  0  3  0
##   6.9  0  0  0  0  3  0
##   7    0  0  0  0  4  0
##   7.2  0  0  0  0  0  1
```

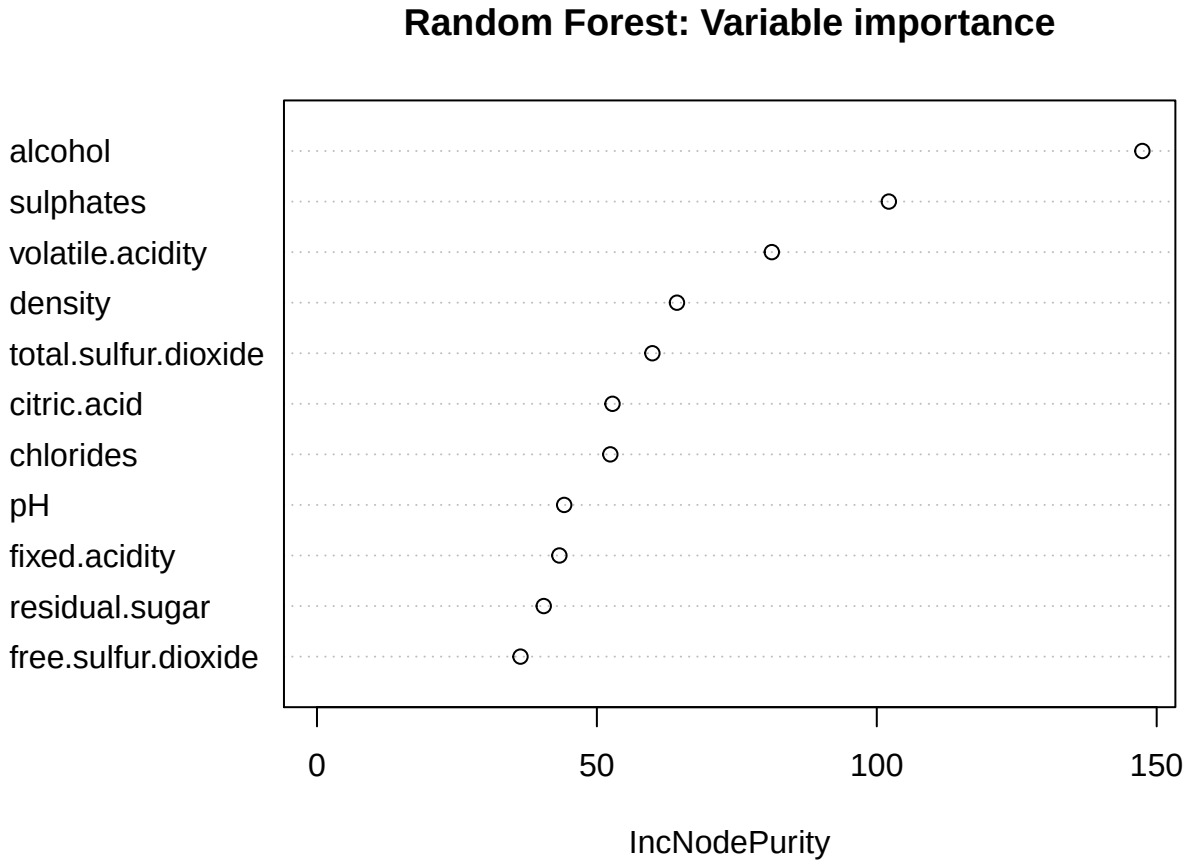
Summary statistics of the quality predicted

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.625   5.204   5.573   5.631   5.926   7.242
```

Summary statistics of the actual quality

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.583   6.000   8.000
```

From the above results we can see that the model now captures the higher quality a bit better than the regression tree, but still does not detect any wines from the lower or the highest quality. However, with this method we can identify the important variables to check when choosing a wine.



It is interesting to note that with this method, the attribute *sulphates* is among the three most relevant for wine quality. This was not observed in the correlation matrix, although in the linear regression model it was a significant variable but with a lower coefficient than the one estimated for the *alcohol* and the *volatility.acidity*.

Finally, these are the metrics to evaluate the model

metric	value
RMSE	0.41
MAE	0.55
R_2	0.42

3.5. Comparing result

After trying different algorithms to determine the wine quality based on its chemical composition, let's compare the main validation metrics

model	RMSE	MAE	R_2
Linear Regression	0.76	0.59	0.40
Knn	0.79	0.63	0.10
Regression Tree	0.65	0.48	0.40
Random Forest	0.55	0.41	0.42

From the table above we can conclude that the more effective algorithm to determine the wine quality trough its chemical composition is the *Random Forest*: it shows the lowest value in *RMSE* and *MAE* and its R^2 is also the highest, but at a very low level.

In all cases, the algorithms fail to detect the most extreme quality categories, both the highest and the lowest, although with the random forest we slightly improve the ability to detect wines of superior quality, but not those that could be described as "exceptional".

4. Conclusion

In this project we have worked with some simple machine learning algorithms trying to approximate the quality of a wine based on its chemical characteristics.

Along the process we have observed significant limitations to detect wines with the lowest qualities (rated 3 or 4) or with the highest qualities (rated 7 or 8). This may be due to the specific nature of the database. In the Section 2, we have noted that out of 1599 data, 681 had quality *5* and 638 had quality *6*. In other words, almost 85.2% of the total sample is concentrated in just two quality categories. This limits the learning ability of the algorithms to detect wines in the most extreme quality categories.

What is certain is that winemaking processes have improved substantially in recent decades, which explains why there are a large number of wines in the market with a more than acceptable quality. At this point, it is revealing the advice given to me by a sommelier friend: "the best wine is the one you like the most".

All that remains is to open a bottle and enjoy a glass of our favorite wine, the one to which we will always assign the highest quality, regardless of its chemical composition. Cheers!