# Predicting the Quality of Wine

## 1. Getting Data

### Data source

We downloaded the dataset “winequalityN.csv” from <https://www.kaggle.com/rajyellow46/wine-quality>.

### Loading the data

In a first step the dataset is imported to R and stored in the data.frame *d.wine*:

d.wine <- read.csv("winequalityN.csv", header=TRUE)

### Describing the dataset

str(d.wine)

## 'data.frame': 6497 obs. of 13 variables:  
## $ type : Factor w/ 2 levels "red","white": 2 2 2 2 2 2 2 2 2 2 ...  
## $ 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 ...

The dataset contains content information of different red and white wines in 6497 observations of 13 columns. In the following, the individual attributes will be explained:

* **type**: categorial predictor with 2 levels white/red that describes whether the wine is a red or white wine.
* **fixed.acidity**: continous predictor that describes the amount of acids that are solid and do not evaporate easily.
* **volatile.acidity**: continous predictor that describes the amount of acids that can lead to a vinegar like taste.
* **citric.acid**: continous predictor that describes the amount of acids that can add freshness and flavor to wines.
* **residual.sugar**: continous predictor that describes the amount of sugar remaining after fermentation. Wines with greater than 45 grams/liter are considered sweet.
* **chlorides**: continous predictor that describes the amount of salt in the wine.
* **free.sulfur**.dioxide: continous predictor that describes the amount of the free form of sulphur dioxide (SO2). It prevents microbial growth and the oxidation of wine.
* **total.sulfur.dioxide**: continous predictor that describes the amount of the free and the bound form of sulphur dioxide (S02). A concentration greater than 50 ppm becomes evident in nose and mouth.
* **density**: continous predictor that describes the density of the water in the wine.
* **pH**: continous predictor that describes how acidic or basic a wine is on a scale of 0 (very acidic) and 14 (very basic). Most wines have a pH value between 3 and 4.
* **sulphates**: continous predictor that describes the amount of the wine additive which can contribute to sulfur dioxide gas (S02) levels.
* **alcohol**: continous predictor that describes the percent alcohol content of the wine.
* **quality**: categorical response variable with 10 levels between 0 and 10 that describes the wine quality.

### Checking the data

head(d.wine)

## type fixed.acidity volatile.acidity citric.acid residual.sugar chlorides  
## 1 white 7.0 0.27 0.36 20.7 0.045  
## 2 white 6.3 0.30 0.34 1.6 0.049  
## 3 white 8.1 0.28 0.40 6.9 0.050  
## 4 white 7.2 0.23 0.32 8.5 0.058  
## 5 white 7.2 0.23 0.32 8.5 0.058  
## 6 white 8.1 0.28 0.40 6.9 0.050  
## free.sulfur.dioxide total.sulfur.dioxide density pH sulphates alcohol  
## 1 45 170 1.0010 3.00 0.45 8.8  
## 2 14 132 0.9940 3.30 0.49 9.5  
## 3 30 97 0.9951 3.26 0.44 10.1  
## 4 47 186 0.9956 3.19 0.40 9.9  
## 5 47 186 0.9956 3.19 0.40 9.9  
## 6 30 97 0.9951 3.26 0.44 10.1  
## quality  
## 1 6  
## 2 6  
## 3 6  
## 4 6  
## 5 6  
## 6 6

tail(d.wine)

## type fixed.acidity volatile.acidity citric.acid residual.sugar chlorides  
## 6492 red 6.8 0.620 0.08 1.9 0.068  
## 6493 red 6.2 0.600 0.08 2.0 0.090  
## 6494 red 5.9 0.550 0.10 2.2 0.062  
## 6495 red 6.3 0.510 0.13 2.3 0.076  
## 6496 red 5.9 0.645 0.12 2.0 0.075  
## 6497 red 6.0 0.310 0.47 3.6 0.067  
## free.sulfur.dioxide total.sulfur.dioxide density pH sulphates alcohol  
## 6492 28 38 0.99651 3.42 0.82 9.5  
## 6493 32 44 0.99490 3.45 0.58 10.5  
## 6494 39 51 0.99512 3.52 NA 11.2  
## 6495 29 40 0.99574 3.42 0.75 11.0  
## 6496 32 44 0.99547 3.57 0.71 10.2  
## 6497 18 42 0.99549 3.39 0.66 11.0  
## quality  
## 6492 6  
## 6493 5  
## 6494 6  
## 6495 6  
## 6496 5  
## 6497 6

As it looks like the data set was imported completely. In row No 6494 there is an missing value (not avaiable, NA) in the **sulphates** column. Probably this is not the only one. Therefore we count the number of NAs in the data set.

sum(is.na(d.wine))

## [1] 38

mean(is.na(d.wine))

## [1] 0.0004499118

The complete data set contains 38 NA. These make up about 0.04% of the data set. We decide to delete the incomplete rows.

d.wine <- na.omit(d.wine)  
sum(is.na(d.wine))

## [1] 0

The data set now contains only complete observations. Now we are ready for the further analysis steps.

## 2. Graphical Analysis