TD 4

Outils pour l'exploration de données - Report: Data visualization and summarization

1 Data set

The data available for this TD is stored in the file `AlgaeAnalysis.txt` available on CELENE. Such a file contains data regarding 200 water samples[1] collected at different European rivers (see TD4 for the references). Each water sample is described by 11 variables. Three of these variables are nominal (qualitative): they describe, respectively, the season of the year when the sample was collected, the size of the river, and the water speed of the river. The eight remaining variables are the values of different chemical parameters measured in the water sample. More specifically, the variable measured are:

- Season of the year
- Size of the river
- Water speed
- Maximum pH value
- Minimum value of O_2 (Oxygen)
- Mean value of Cl (Chloride)
- Mean value of NO_3^- (Nitrates)
- Mean value of NH_4^+ (Ammonium)
- Mean of PO_4^{3-} (Orthophosphate)
- Mean of total PO_4 (Phosphate)
- Mean of Chlorophyll

The last seven values of each observation (example) are the distribution of different kinds of algae (`a1` to `a7`). No information is given regarding which specific type algae were identified.

Just recalling that for the context of this data set we have a *regression problem*. In a simplified manner, we could say that we have seven regression problems, each one concerning one of the the variables `a1` to `a7`.

2 Data visualization and summarization

For each of the task below you should write at least one paragraph with your answer to the questions raised. The analysis you are developing in this TD might be important for decisions/analysis you will have to take/perform in the next TDs. Important: several variables of the data set can have missing values (`NA`). This might cause problem in the use of different functions (e.g., `var`). In order to avoid this, one can set the parameter `na.rm` to `T`. There is also the function `is.na()` that gets as parameter a vector `x` and produces a vector of boolean values (true or false). An element of this vector is true when a value in `x` is a NA.

Task 1: supposing that the data set is in your current working directory, load it into R using:

```
algae <- read.table("AnalysisH.txt", header=TRUE)
```

[1] Here I used the word "sample" meaning "observation". In this context, writing 200 water "observations" would sound strange.

Examine its structure. Identify each variable from the list given in section 1. Note its data type.

Task 2: Summarize (`summary()`) the contents of the `algae` data set. To save typing, you can first `attach` the `algae` data frame. This makes the field names in the data frame visible in the outer R environment. For example, when we type `season`, this field of the attached frame is accessed; otherwise we would have had to type `algae$season`.

- What evidence does the summary give about the symmetry and spread of the distributions of each numeric variables (`mxPH`, `mnO2`, `Cl`, `NO3`, `NH4`, `oPO4`, `PO4` and `Chla`)? Support your arguments by observing, among other things, the difference between medians and means, as well as the inter-quartile range (3rd quartile minus the 1st quartile). Can you observe extreme values?

  Basically, with the exceptions of the variables `mxPH` and `mnO2` , all the others present a large difference between the mean and median (as well as a large inter-quartile range). These give an indication that the distribution of these variables are not symmetric (otherwise mean and median would be close to each other) and can be rather spread around the mean. With respect to presence of extreme values, observing the minimum and maximum values of each variable, as well as the median/mean, with the exception of `mxPH`, all the other variables present a rather large range what may indicate the presence of extreme values. For example, for the variable `NH4` the minimum value observed is `5.0` and the maximum is `24064.00`, with the median in `103.17`. That is, the value `24064.00` is strong "candidate" to be a case of extreme value. However, a more sounded analysis using, for example, box plot is needed.

- Create a function to implement the Yule coefficient (asymmetry) and apply it to the numeric variables. Observe the results.
  With exception of `mxPH` and `mnO2` which present a negative skewness, all the other variable are positively skewed. That is, they tend to present relatively fewer large values.

Task 3 Visualize, with a histogram and a density plot, the distributions of the contents of, respectively, `mxPH`, `mnO2` and `Cl`.

- Does the distributions look symmetric? skewed? peaked?

  Looking at the graphics of `mxPH` (Figure 1), although it is negatively skewed, visually one can observe that the distribution is almost symmetric, resembling a normal distribution. In contrast, the distribution for `Cl` (Figure 2) is quite skewed (right), presenting several small values. (I am not analyzing `mnO2`).

Task 4: Show the distribution of `oPO4` using a box plot:

```
>result <- boxplot(oPO4,ylab='Orthophosphate (oPO4)')
>rug(jitter(oPO4),side=2)
>abline(h=mean(oPO4,na.rm=T),lty=2)
```

The analysis of this graphic shows that the variable `oPO4` has a distribution of the observed values skewed to the right. That is, in most of the water samples, the value of `oPO4` is small, but there are several observations with large values. Indeed, some values are extremely large. These observations are shown as *box plot outliers*, i.e., they are more than 1.5 times the inter-quartile range (width of the box) larger than the 3rd quartile. This is a technical measure: they are box plot outliers, but this does not necessarily mean that they are part of a different population.

- Often, when we there are outliers in the data, we are interested in inspecting the observations that have these values. In the case of `oPO4`, out of a total 200 observations, which are exactly these unusual observations?

The observations are 2,20,21,32,43,44,88,89,91,119,120,157,171 and 172. These observations should be inspect, together with the domain expert, to check if they could present possible errors.

- Now, remove from `oPO4` and `NH4` the outliers you have identified. Next, draw the box plots and histograms for each of these variables. What can you observe when compare these graphics (without the outliers) with to the previous one you generated to these variables?

  In this item, I just want you to realize that the presence of extreme values can distort completely any attempt to visualize the data. For example, in the case of `NH4`, originally the values span from 5.00 to 24060.00, that is, several orders of magnitude. After removing the extreme values, for the purpose of display, we can now have a better visualization of the distribution of `NH4`. (Compare Figures 3 and 4).

- Create a function that, given a numeric vector `x`, remove the *extreme* outliers (any value 3 times the inter-quartile range).

Task 5: Using a box plot, show the distribution of algae `a1` with respect to `size`. Do the same thing for algae `a3` with respect to `season`.

- In each case, can you obverse any influence of the qualitative variable (respectively, `size` and `season`) on the distribution of the algae (respectively, `a1` and a3)? If yes, which type of influence have you found?

  For the case of algae `a1` with respect to `size`, Figure 5 allows us to observe that higher frequencies of alga `a1` are expected in smaller rivers, which can be valuable knowledge.

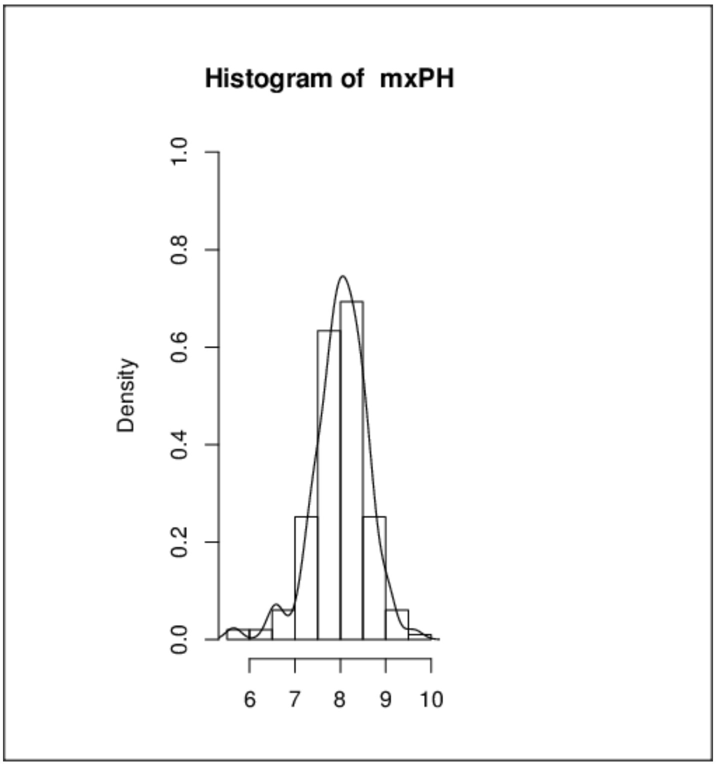

Figure 1: Histogram/Density mxPH

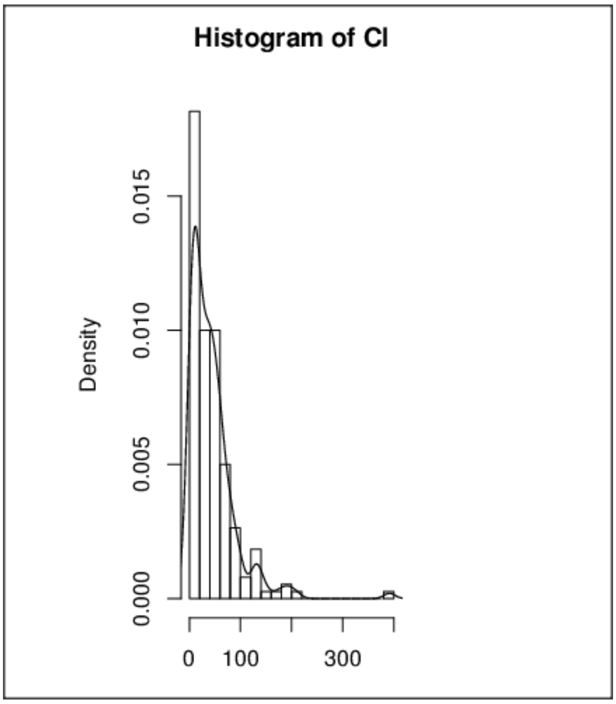

Figure 2: Histogram/Density Cl

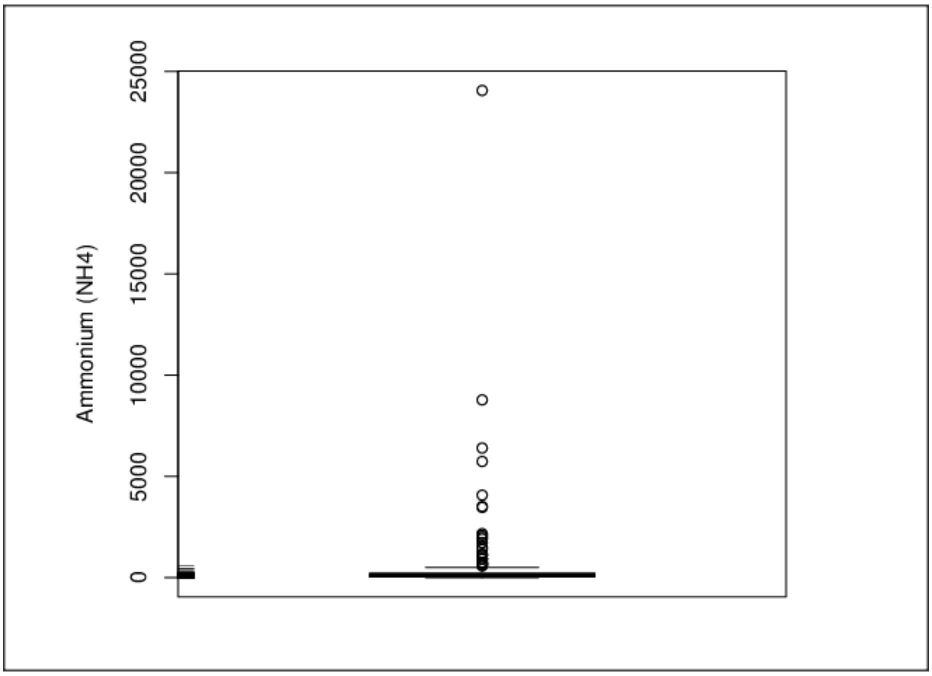

Figure 3: Box plot of the original values of NH4

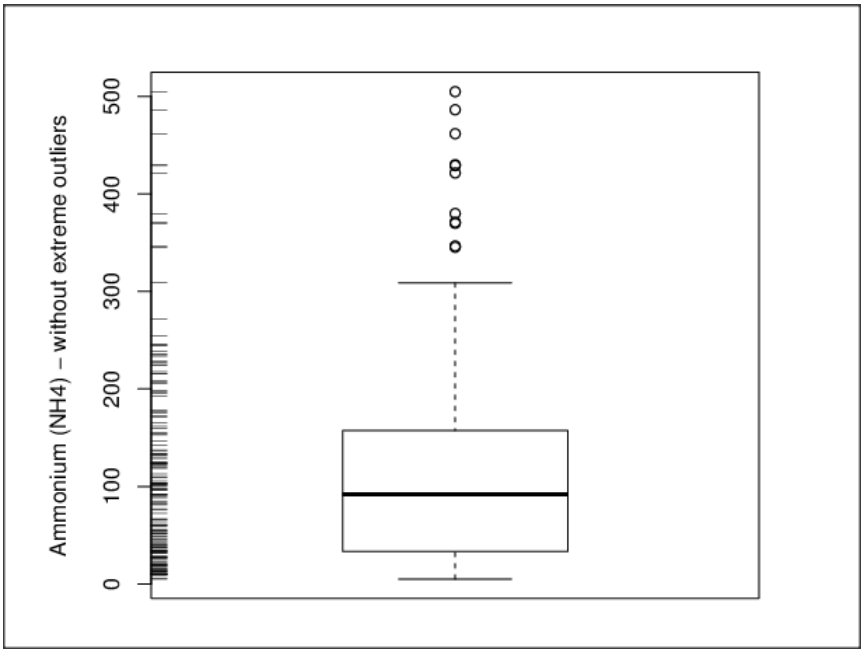

Figure 4: Box plot of the values of NH4 without the "outliers" in figure 3

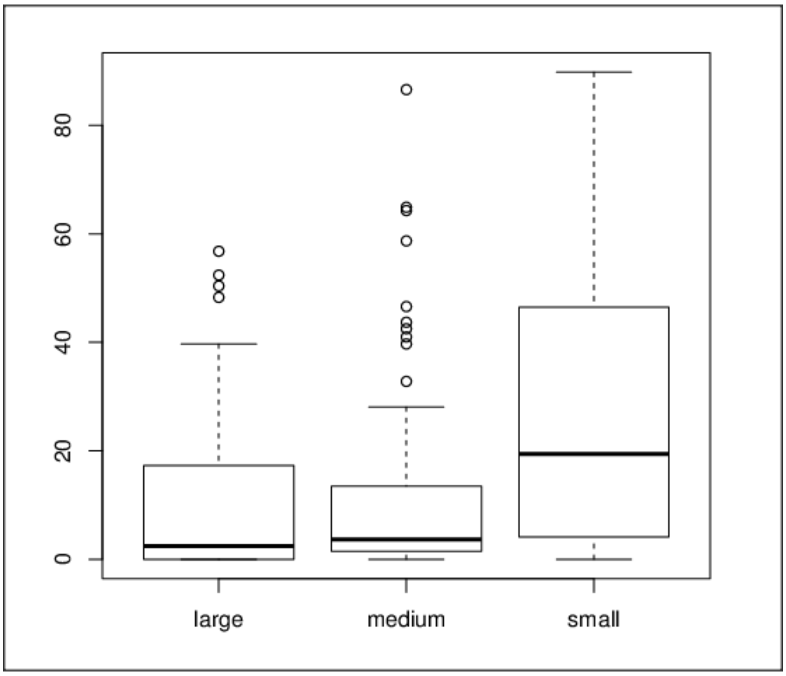

Figure 5: Conditional box plot: a1 versus size