DATA SCIENCE

Working with data

Practical 2

M.José Ramírez-Quintana
José Hernández-Orallo
MITSS
Universitat Politècnica de València

- Exercise 1: Inspection of data.
  The "titanic.csv" file (available from the course documentation page at poliformat) contains data on the sinking of the Titanic. More concretely, the file contains data for 891 of the real Titanic passengers. Copy the file in your working directory. Then, go to R and use the command

  ```
  titanic <- read.csv(file.choose(),header=TRUE, sep=',')
  ```

  and choose the relevant `csv` file. You may write the file name (including the path to the file) instead of `file.choose()`, as shown in the R seminar and used in the Practical Work 1, i.e. 'titanic.csv'. Notice how you can change the field separator character according to what is used in the csv file, so that the file is interpreted in the correct way. Show the names of the columns. Observe that the first column (whose name is `"PassengerId"`) is redundant (it denotes the identifier of each instance) so it could be removed. To do this, we can use the `select` function of the `dplyr` package:

  ```
  library(dplyr) titanic<-select(titanic,-1)
  ```

  After that, we inspect the resulting data frame:

  ```
  > head(titanic)
  > summary(titanic)
  ```

  Which variables are quantitative and which variables are categorical?

- Exercise 2: Looking at the protected attributes.

  The data in the titanic.csv dataset is an example of biased data. Answer the following questions:

  1. Identify which attributes correspond to personal data. Are there any that we can rule out (that is, disregard) for data analysis?
  2. Identify which attributes are protected (sex, ...).
  3. Use table() to analyse the class distribution (attribute Survided) for each protected attribute, and comment the results.

- Exercise 3: Data transformation I.

  Download the file “airquality.csv” from poliformat. This dataset contains some New York air quality measurements. Solve the following exercises:

  1. Discretise the Ozone column into five bins (‘bin1’, ‘bin2’, ...) of equal width and a sixth bin (‘binNA’) for NA.
  2. Discretise the Solar column into four bins of equal frequency and a fifth bin for NA.

- Exercise 4: Data transformation II

  Download the file “titanic2.csv” from poliformat (a simplified version of the original titanic dataset), and solve the following exercises:

  1. Numerise the ‘class’ column, where Crew=4, 1st=3, 2nd=2 and 3rd=1.
  2. Transform the titanic2 data frame into a new data frame (titanic3) with as many examples as passengers using the Freq column. In other words, there should be no rows for those for which Freq=0 and there should be 35 replicated rows for those with Freq=35.

- Exercise 5: Data selection.

  1. Using the data frame ‘air’, perform a simple random sampling of 50 examples.
  2. Using the data frame ‘air’, perform a stratified random sampling of 5 examples of each month.