DATA SCIENCE

Working with data

Practical 2

M.José Ramírez-Quintana
José Hernández-Orallo
Régel González-Usach

ETSINF
Universitat Politècnica de València

- Exercise 1: Inspection of data.
  The “titanic.csv” file (available from the resources folder at poliformat) contains data on the sinking of the Titanic. 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 Section 12 explains, 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 "X") is redundant (it denotes the identifier of each instance) so it could be removed. To do this, use the subset command as follows (consult the help if it is needed):

  ```
  titanic<-subset(titanic,select=-X)
  ```

  Now try the following commands:

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

  Which variables are quantitative and which variables are categorical? How can we know it?

- Exercise 2: Working with basic graphics.
  Download the file “cars.csv” from poliformat. This file contains information about the speed and stopping distances of cars.

2.1 Make a plot of the distance field in terms of the speed field (use the $ syntax).

2.2 Make a histogram of the distance variable.

2.3 Make a histogram of the speed variable.

2.4 Modify the previous plots to show the name of the variables ("speed" or "distance") as the title of the axis. Change the title of the three graphics, and also use colours for the histograms and titles. Save the new graphics as pdf files.

- Exercise 3: Transformations of variables and datasets.
  Remove the first column of the cars data frame. Now, assume that data from two more cars are made available:

speed	dist
21	47
34	87

3.1 Construct a new data frame with the above data.

3.2 Add the constructed data frame to the cars data frame.

3.3 Sort the data in the resulting dataset by column speed (ascending). There is two ways to do it: using the order() command or combining the with and the order() commands. (Suggestion: to search on the internet "how to sort a data frame by columns").

- Exercise 4: Data manipulation. Download the file "airquality.csv" from PoliformaT. This dataset contains some New York air quality measurements. Solve the following questions:

1. Extract the first 2 rows of the data frame and print them to the console. What does the output look like?
2. How many observations (i.e., rows) there are in this data frame?
3. What is the value of Ozone in the 40th row?
4. How many missing values there are in the Ozone column of this data frame?
5. What is the mean of the Ozone column in this dataset? Exclude missing values (coded as NA) from this calculation.
6. Extract the subset of rows of the data frame where Ozone values are above 31 and Temp values are above 90. What is the mean of Solar.R in this subset?

- Exercise 5: Data transformation (2).

  With the data frame“airquality.csv” 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 size and a fifth bin for NA.
  3. Create a new column AbsDay from the columns Month and Day such that counts the number of days passed from Month=5 and Day=1.

- Exercise 6: Data transformation (3).

  With the data frame “titanic” solve the following exercises:

  1. Numerise the class column, where Crew=4, 1st=3, 2nd=2 and 3rd=1.
  2. Transform the titanic data frame into a new data frame (titanic2) 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.
  3. Compare the plots of the original titanic data frame with the new one.

- Exercise 7: Data selection.

  1. Calculate a correlation matrix for the air dataset. Do you see a pair of attributes that are redundant?
  2. Calculate a correlation matrix for the cars dataset. Do you see a pair of attributes that are redundant?
  3. Using the data frame ‘air’, perform a simple random sampling of 50 examples.
  4. Using the data frame ‘air’, perform a stratified random sampling of 5 examples of each month.

- Exercise 8: Data aggregation.

  Download the file “sales.txt” from PoliformaT. This dataset contains information about sales transactions from various stores.

  1. Calculate the total sales per store. Use the aggregate() function.
  2. Find the average sales amount for each product category.
  3. Group the data by store and product category, and calculate the total sales for each combination.
  4. Plot the total sales per store in a bar chart. Make sure to label the axes and title the plot appropriately.
  5. Save the bar chart as a PDF.

- Exercise 9: Merging Datasets.

  Download the files “customers.txt” and “orders.txt” from PoliformaT. These files contain customer data and order data, respectively.

  1. Load both datasets into R.
  2. Merge the datasets by the "customer_id" column. Use an inner join to keep only the records that are present in both datasets.
  3. How many unique customers are there in the merged dataset?
  4. Find the total number of orders placed by each customer. Use the table() function or similar.
  5. Save the merged dataset as a new CSV file called "customer_orders.csv."