Data Analytics - R Exam

Ester Bernadó

5th november 2020

The file **accidents.csv** collects the information of several accidents that occurred in the city of Barcelona during year 2017. Upload the dataset and look carefully at the information that it contains, to see the structure and contents of it.

Please, before you start, read carefully these guidelines:

- You need to write R instructions to answer the questions. Visual inspection is not permitted. The solution needs to be provided by code.
- Your instructions should always print the result to screen.
- When you write your code, please specify clearly to which question it belongs. Do this with a comment, like:

  #1.

  read.csv(...)

  #2.

  print("hello world")
- If you don't answer a question, write the question number and leave the answer blank.
- You need to submit the .R file.
- Include your name and surname inside the .R file as a comment.
- If anyone cheats, all the implied students will get a 0 in the full exam. Remind that cheating is very easy to identify, since there are so many ways in which the code can be solved that an exact code is very easy to be identified as cheating.

Answer the following questions:

1. How many different districts there are in the dataset (column **DistN**)?
2. List the names of the different districts in increasing alphabetical order.
3. Every row refers to an injured person in the accident. Thus, one accident may have more than one row. In any case, calculate how many injured people there are for each district.
4. Then, plot the name of the district with the greatest number of injured people.
5. Calculate how many pedestrians ("Vianant"; look at column PersT) are involved, and how many of them resulted in "Ferit greu" (as reported in column Vict).
6. Calculate the mean age of the injured people that died in the accident. To know who died in the accident, look at column **Vict**.
7. Print a plot that shows how many accidents (injured people) there are per hour. Column **Hora** stores the hour in which the accident happened.
8. Plot a pie chart that shows the proportion of men and women in the dataset. You need to use column **Sexo**.
9. Delete column X.1 from the dataset

10. Search all accidents that occur from 0 until 6 in the morning (both included) and delete them from the dataset.

11. Imagine that we want to calculate the distance to a hospital that is located in X=431510.63 and Y=4584212.20. Calculate the distance of every point to this hospital, by applying this formula:

$$dist = \sqrt{(x_i - X)^2 + (y_i - Y)^2}$$

where X, Y are the coordinates given, and x_i and y_i are the location of every accident. Once you have the distance calculated for every accident, add a column in the dataset named *dist* with this information.

12. Now, calculate the accident that occurred closest to this hospital. Print its location (column *BarN*)