STAT40730 Data Programming with R
Assignment 1

Isabella Gollini

Instructions

- This assignment is due on **Monday 23rd October 2023** at 11:59pm.
- You should submit it to the "Assignment 1" assignment object in Brightspace.
- You should submit 2 or 3 files only depending on your final format:
  1. Qmd file detailing the commented code you used to obtain your answers.
  2. Rendered document in either pdf or HTML showing all your code and your answers.
  3. **Only if you are rendering to HTML:** a pdf file obtained by converting the HTML to pdf.
     - You can use Google Chrome. File > Print > Destination [Change. . .] > select Save as PDF.
- Remember that if you decide to produce an HTML output you must have the following on your YAML header:

```
format:
  html:
    embed-resources: true
```

- You may submit it multiple times before the deadline, but only the last version will be marked.
- There is a maximum of 19 marks for this assignment. This assignment is worth 19% of your final grade.
- The marks available for each question are shown in brackets.
- Late submissions will score 0, unless a "Late Submission of Coursework" form is submitted.
- Assignment 1 consists of 3 tasks: data manipulation, analysis, and creativity.
- This assignment covers up to the material in Topic 4. Topic 4 is only needed to complete the last question of Task 2 and Task 3.
- You may have to discover and learn some new functions. Use `help()` and `help.search()` to find what you need.
- A couple of suggestions: create an RStudio Project for this assignment and save the qmd file and the data set in the same folder. Render your document frequently to fix errors.

Assignment 1

The dataset `EurostatCrime2021.xlsx` records offences (values per hundred thousand inhabitants) by offence category in 41 European Countries in 2021. Full information on the dataset is available at: https://ec.europa.eu/eurostat/cache/metadata/en/crim_off_cat_esms.htm.

- **Write a scientific report** by completing the three tasks below. [2.5]
  - Complete your assignment using Quarto, check that all the code and output are correctly shown in your final document.
  - Clearly indicate in each code chunk which question it is referring to.
  - In tasks 1 and 2 you must use base R and the packages that we have used in class up to topic 4 only. You are free to use functions from other packages for task 3 if you wish.
  - Do not print the full dataset, it makes the document very hard to read.
  - Save the data file in the same folder as the .Qmd file, so that you don't have to specify the file path that is specific of the computer you are using (and we would not be able to run your code without changing it).

Task 1: Manipulation

1. Load the dataset `EurostatCrime2021.xlsx`. Notice that the data starts in row 6, with row 7 containing the variable names, and that the missing values are represented by `":"`. [Note: do not modify the original file `EurostatCrime2021.xlsx` directly (by opening it with Excel, for example) as we need to be able to render your .Qmd file and reproduce your final document using the original dataset.] [1]
2. What is the size (number of rows and columns) and the structure of this dataset? [0.5]
3. Remove the columns **Fraud** and **Money laundering** (they contain no data). [0.5]
4. For some countries **Theft** includes also **Theft of a motorized vehicle or parts thereof**, **Burglary**, and **Burglary of private residential premises**, in others they are recorded separately. To compare different countries, remove the columns involving theft and burglary:
   - **Theft**,
   - **Theft of a motorized vehicle or parts thereof**,
   - **Burglary**,
   - **Burglary of private residential premises** [0.5]
5. Add a column containing the overall record of offences for each country (per hundred thousand inhabitants) [Hint: there is a function in `base` R that allow you to do this]. [2]
6. Work with the dataset you just created, and write some code to list the countries that contain any missing data. [2]
7. Remove the countries with missing data. [0.5]
8. How many observations and variables are in this new dataset? [0.5]

Task 2: Analysis

Work with the dataset produced at the end of Task 1.

1. Which country has the highest overall record of offences in 2021 (per hundred thousand inhabitants)? To get full marks you must also provide the R code that returns that country name only. [1]
2. Produce a table showing the countries and the proportion of the overall crimes due to acts against computer system, sort the rows in ascending order of proportions, and display only the first three decimal digits. [2]
3. Create a plot displaying the relationship between robbery and unlawful acts involving controlled drugs or precursors. Make the plot "nice" i.e., show country names, change size of the plot, change axis labels, etc. [2]

Task 3: Creativity

Do something interesting with these data! Create **two plots** showing something we have not discovered above already and outline your findings. For this Task you can decide if you want to use the original dataset from Task 1 Question 1 or the modified one. [4]

END OF ASSIGNMENT 1

Few tips for troubleshooting

- Be aware that a common error is to give the same label to two different code chunks!

````
```{r}
#| label: cars
summary(cars)
```
````

````
```{r}
#| label: cars
plot(cars)
```
````

You can fix this by changing the label to one of them:

````
```{r}
#| label: fig-cars
plot(cars)
```
````

- In case of a code error that you can't fix in time for your submission.

Add the option `error: TRUE` into the R chunk to run the code, show the error message on the rendered file. For example:

````
```{r}
#| error: true
x <- "a"
sum(a)
```
````

Or you can add the option in your YAML header to work on the full document:

```
execute:
  error: true
```