ΔΠΜΣ ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

ΜΑΘΗΜΑ: *Προγραμματιστικά Εργαλεία και Τεχνολογίες για Επιστήμη Δεδομένων*

ΔΙΔΑΣΚΩΝ: *Δημήτρης Φουσκάκης*

ΑΚΑΔΗΜΑΙΚΟ ΕΤΟΣ: *2021-2022*

Home Assignment
02/12/2021
Title: Exploratory Data Analysis using R

In the `coronavirus` package in R you will find the `covid19_vaccine` dataset. In the following link you will find information about the dataset

https://cran.r-project.org/web/packages/coronavirus/coronavirus.pdf

Read carefully the package documentation, from the above link, and update the dataset. Create two new variables in the dataset named `fully_vaccinated_ratio` and `partially_vaccinated_ratio` by dividing the values of the variables `people_fully_vaccinated` and `people_partially_vaccinated` respectively by the values of the variable `population`. By keeping the same names of the two new variables created, convert the relative frequencies to percentages and keep only one decimal place. Finally keep only the data on the country level (including world data) and remove the information on provinces.

Then your task is to perform exploratory data analysis in order to visualize the data, make comparisons (for example between countries, between continents, between time seasons, etc.…) and draw conclusions using the two main variables of interest: `fully_vaccinated_ratio`, `partially_vaccinated_ratio`.

Your conclusions can be drawn on specific time periods that you choose and/or on the latest day of your dataset. In addition, you can choose specific countries of your interests. Your aim is to perform appropriate ranks and/or aggregations and plots in order to reveal hidden structures in your data, using possibly values from several variables at the same time. All tables and plots should be labeled appropriately and cited in the main body of your paper.

The data are updated daily, so in your final report **state clearly which is the latest day of the dataset that you used**.

Instructions:

1. **Assignment submission deadline**: **26 January 2022 at 13:00.** Please send me your paper at fouskakis@math.ntua.gr. Please note that no assignments will be acceptable after this date and time.
2. **Your paper should be written in Latex.** You have to submit the **pdf output**. Your pdf file should be named using the following format: Surname-Name.pdf (replace with your details in English; for example Fouskakis-Dimitris.pdf). Your file should start with a cover page in which you will include your details (title of the assignment, your name, your surname, your email, your student number and if you are an MSc or PhD student). The maximum length of your file should be 15 pages. You are free to write your report in Greek or in English.
3. You should try to explore the data using appropriate tables and plots. It is **compulsory** for your plots to use the R library `ggplot2` and for your tables the R library `data.table`. For each table and plot you produce, it is important to explain your findings, in a compact way, as simple as possible, extracting all the information. Your R code should be included in the main body of your report, i.e., not as an appendix.
4. It is important that your work reflects your knowledge rather than it being simply an accumulation of information. The assignment should be well structured and easy to read.