ΔΠΜΣ ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

ΜΑΘΗΜΑ: *Προγραμματιστικά Εργαλεία και Τεχνολογίες για Επιστήμη Δεδομένων*

ΔΙΔΑΣΚΩΝ: *Δημήτρης Φουσκάκης*

ΑΚΑΔΗΜΑΙΚΟ ΕΤΟΣ: *2023-2024*

Home Assignment
13/12/2023

Title: Exploratory Data Analysis using R

PISA stands for "Program for International Student Assessment" and it is applied to 15-year-old students across the world to assess their performance in Math, Reading and Science.

In the course website you will find the "Pisa mean performance scores 2015 Data.csv" dataset with 1161 rows of data and the following variables:

Variable Name	**Variable Meaning**
Country Name	The name of the country
Country Code	The coding name of the country
Series Name	Explanation of the outcome in the last column of the file
Series Code	Coding of the explanation in the previous column
2015	The outcome (performance score) in the year 2015

In the course website you will also find the "Pisa mean performance scores 2013 - 2015 Definition and Source.csv" file with helpful definitions and the source from where the dataset was extracted.

The purpose of this exercise is to investigate different features that affect student's performance. Check if gender, countries, regions affect the student's performance in the three disciplines.

Read the data into R and make sure that each variable has the correct type. Your task is to perform exploratory data analysis in the dataset, to give highlights to the above research question, and relevant ones. You should create tables and plots and comment on your findings. You are free to create additional variables, if needed, to drop variables if you wish, to combine variables if so, and to perform appropriate aggregations and plots to reveal hidden structures in your data, using possibly information of several variables at the same time. All tables and plots should be labeled appropriately and cited in the main body of your report. Beware that missing values in your csv file are denoted by the symbol "..".

Instructions:

1. **Assignment submission deadline**: **26 January 2024 at 13:00.** Please submit your paper in https://helios.ntua.gr/mod/assign/view.php?id=41495. Please note that no assignments will be accepted after this date and time.
2. **Your paper should be written in Latex.** You have to submit only 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 and in which program). The maximum length of your file should be 20 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 simply as possible, extracting all the information. Your R code should be included in the main body of your report and 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.