**ΔΠΜΣ ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ**

**ΜΑΘΗΜΑ:** *Προγραμματιστικά Εργαλεία και Τεχνολογίες για Επιστήμη Δεδομένων*

**ΔΙΔΑΣΚΩΝ:** *Δημήτρης Φουσκάκης*

**ΑΚΑΔΗΜΑΙΚΟ ΕΤΟΣ:** *2024-2025*

**Home Assignment**
**11/12/2024**
**Title: Exploratory Data Analysis using R**

In the course website you will find the "insurance" dataset which contains medical information and costs billed by health insurance companies. It contains 1338 rows of data and the following variables:

| Variable Name | Variable Meaning |
|---|---|
| age | age of primary beneficiary in years |
| sex | insurance contractor gender; a factor with levels female, male |
| bmi | body mass index of primary beneficiary in $Kg/m^2$ |
| children | number of children covered by health insurance |
| smoker | if the primary beneficiary smokes; a factor with levels yes, no |
| region | the beneficiary's residential area in the US;<br>a factor with levels northeast, southeast, southwest, northwest |
| charges | individual medical costs billed by health insurance for a year in dollars |

The purpose of this exercise is to look into different features to observe their relationship, and to monitor medical expenses differences between the available variables that have been considered in this study.

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, in order to give highlights to the 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 in order 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.

**Instructions:**

1. **Assignment submission deadline**: **26 January 2025 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.