ΔΠΜΣ ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

ΜΑΘΗΜΑ: *Προγραμματιστικά Εργαλεία και Τεχνολογίες για Επιστήμη Δεδομένων*

ΔΙΔΑΣΚΩΝ: *Δημήτρης Φουσκάκης*

ΑΚΑΔΗΜΑΙΚΟ ΕΤΟΣ: *2020-2021*

Home Assignment
20/11/2020
Title: Exploratory Data Analysis using R

In the following links, you can find up-to-date Covid-19 data from John Hopkins CSSE.

https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv

https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_global.csv

The first file has information about the cumulative confirmed cases and the second file has information about the cumulative number of deaths from Covid-19, on different countries and dates.

Initially, perform the following tasks in R (in both datasets):

1. Remove columns with names `Province`, `State`, `Lat` and `Long`.
2. Convert data from wide to long format.
3. Rename variable `Country.Region` to `Country`.
4. Name the variable with the cumulative confirmed cases as `confirmed` and the variable with the cumulative number of deaths as `deaths`.
5. Convert the variable `date` from character to a date object (check the `mdy()` function in R). In the initial datasets for example X1.22.20 refers to 22/1/2020.
6. Group by `country` and `date`.
7. Merge the two datasets into one.
8. Calculate counts (`confirmed` and `deaths`) for the whole world.
9. Sort (again) by `country` and `date`.
10. Create two extra variables: `confirmed.ind` and `deaths.inc` with the daily confirmed cases and daily deaths respectively (check the `lag()` function in R).

Then your task is to perform exploratory data analysis in order to visualize the data, make comparisons (for example between countries, between continents, between seasons, etc.…) and draw conclusions using the four main variables of interest: `confirmed`, `deaths`, `confirmed.ind` and `deaths.inc`, or any additional ones that you have created based on these. If needed you can convert back your dataset from long to wide format and work with any of the two formats you wish.

You are free to select specific time periods if you wish, specific countries of your interest, create additional variables, if needed, drop variables if you wish, combine variables if so, and perform appropriate 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, but you are free to use the data that you will originally download.

Instructions:

1. **Assignment submission deadline**: **18 January 2021 at 13:00.** Please send me your paper at fouskakis@math.ntua.gr. Please note that no assignment 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). 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 codes should be included in your report, not as an appendix, but in the main body of your work.
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.