DPRPy 2021/2022

Homework assigment no. 2 (max. = 40 p.)

Maximum grade: 40 p.

Deadline: 14.01.2022, 23.59

This task is solved in groups of two or three people. Homework should be sent via the `Moodle` platform:

one archive .zip2 (for each group) named

`Last-name1_First-name1_Last-name2_First-name2_assgment_2.zip` (one directory inside: `assgment_2`), in which the following files will be placed:

- presentation (slides) containing the results of data analysis (PDF or HTML)

  the assessment will be issued mainly based on presentation;
- all the `.R` scripts / `.py` moduls and notebooks that allow to recreate results (figures, tables) contained in the presentation;
- indirect data on the basis of which the results were generated (files `.csv`,`.json` etc.); note: we don't add files containing raw data - the uploaded .zip file should be "reasonable" sizes;

1 Data

This year we will go back in time a little bit. We are going to work on the data from Data Expo 2009 **Airline on-time performance**.

> [. . .] The data consists of flight arrival and departure details for all commercial flights within the USA, from October 1987 to April 2008. This is a large dataset: there are nearly 120 million records in total, and takes up 1.6 gigabytes of space compressed and 12 gigabytes when uncompressed.

Data and its description in available at

https://community.amstat.org/jointscsg-section/dataexpo/dataexpo2009

> Please, submit the information about group memebes via the Google form
>
> https://forms.gle/gv6QyZiHC9fmqzR5A .

This homework is a data science challenge - each group creates interesting (for themselves and the audience) questions and generates answers to them.

> The challenge The aim of the data expo is to provide a graphical summary of important features of the data set. This is intentionally vague in order to allow different entries to focus on different aspects of the data, but here are a few ideas to get you started: * When is the best time of day/day of week/time of year to fly to minimise delays? * Do older planes suffer more delays? * How does the number of people flying between different locations change over time? * How well does weather predict plane delays? * Can you detect cascading failures as delays in one airport create delays in others? > * Are there critical links in the system?

The projects that meet the following criteria will receive at least satisfactory grade ($> 50\%$, i.e., 20 p.):

1. uses data concerning at least 3 years;
2. contain the code that alows to read data and generate at least three interesting results (answers to 'research' questions in the form of charts / tables / etc.),
3. present the obtained results (10 min. long presentation).

Each additional analysis or non-trivial technique used will have a positive impact on the assessment (e.g. interactive charts, animations, web applications, maps, algorithms and data structures own implementations enabling the improvement of the speed of the analyzes performed methods known from the literature (with author's modifications), etc.). In particular, the maximum grade (very good) only works that really stand out.