Final Project Details

Data Challenge 2

Objective

- The objective of the final project is to *challenge you on your knowledge of data science in R.*
- *Only the methods presented in the lectures* of "Data Challenge 2" will be evaluated. You are free to use any statistical method you know but your grade won't necessarily increase if you use more advanced techniques than the ones presented in the course.

Typical structure of the work

The project is expected to represent an example of a data science consultancy where you are both the client and the data scientist team. You are:

- *the client* because you will propose a dataset and ask some questions on the data.
- *the data scientist team* because you will
  - clean and explore the data;
  - study the feasibility of the questions;
  - propose different methods to answer the questions and study their advantages and limitations;
  - develop visualizations to explain both the data and the methods' advantages and limitations;
  - present the results in a coincise way through a presentation;
  - present the work in a more extensive way with a report.

An example structure for a final report could contain the following parts.

1. *Introduction and problem setting:* short section that introduces the data (sources, purpose for collecting, age?) and presents the main questions answered by the project. In this section plots could be used to showcase the data and anticipate the answers.
2. *Data cleaning and exploration:* present how the data were loaded, cleaned and put in a format suitable for the rest of the study. This part will show pre-eminently in the report with the code chunks needed to clean the data. Some plots could be useful here to understand the structure of the data.
3. *Exploratory data analysis:* present the data in details before modelling with plots (boxplots, scatter plots, etc.) and summary statistics (means, median, sd). This part serves as a first justification of why some models are chosen to answer the questions. For example if we are interesting in questions related to a binary variable, then we will use classification models. In this part we can show the summary statistics for the binary variable.
4. *Modelling:* for each question asked in the introduction, present one (or more) methods to answer the question. The output of each model should be explained. When more than one method is compared, advantages and limitations of those methods related to the particular case should be highlighted through plots and tables. Moreover plots should be used to showcase the power of the best method chosen.
5. *Summary of results and conclusion:* This section should summarize the main findings with reference to the questions asked in the introduction.

Evaluation

The project will be evaluated on two deliverable items:

- A short presentation (25 minutes+5-10 minutes questions) of the main findings delivered in front of your peers (online) during the last two sessions of the course (**30/05** and **06/06**). The presentation:
  - is intended as an "executive summary" of your work: you should present the data and the questions in details, but only report the main findings without going into details on how to do things (unless explicitly asked);
  - should be intended for an audience that has not seen your data before (i.e. pay attention to the presentation of the problem, sections 1,3,5 outlined above);
  - should tell a story.
- A report to hand in to the lecturer after the presentation. The report should
  - be a R Markdown report. The `Rmd` and the output file (html, PDF or Word) should be hand in;
  - provide an explanation of the data and a detailed overview of the questions answered by the report;
  - test several methods to answer the questions posed. Here it is expected that you report details on how the methods are selected, tuned and tested;
  - be self-consistent: it should either be delivered with the data or with a way to obtain the data (link, git, etc.). Moreover the code chunks in the report should run and produce output on the lecturer computer.

The **report** will account for **60%** of the final grade and the **presentation** for **40%**.

All members of the team are expected to present. The evaluation of the presentation part could slightly differ within the same team depending on the answers provided during the presentation by each member.

Final notes and suggestions

- You are a *team* on this project, so you can harness this power by diving the work into individual members.
- *Keep track, with commented Rmd files, of all the work you do.* Not all your attempts will end up in your final report, but if you keep track of all the work you can easily reuse parts in this or in future projects.
- The presentation and the report are not independent works: the plots/tables/visualizations you prepare for the report can (and should) be *reused* in the presentation. *You are not expected to work twice!*