Group Project: Research skills Programming with R, BLOCK 2 2019-2020

Welcome to the Group Project, which makes up 30% of your final grade. It is due **Friday, January 10th 2020, 21:00**. Each group must be made up of 2-5 students; only one of you should submit the final report on Canvas, and you will all receive a single grade. Contact me if there are teamwork problems.

In this assignment, you must use R for a small research project. First, find an existing, well-documented data set (dimensions of at least 400 x 15), that has not been used in the course, and formulate one or two relevant research questions. Then, use R and your recently acquired skills in mastering this software package, to obtain insight into the data and *to contribute to the answer* of the research questions. Thirdly and finally, you will have to write a short, lucid and convincing research paper, that reports on all relevant steps that have been taken (1250 words ±10%).

Depending on the data that you found, you will have to do some preprocessing with R (cleaning, merging and exploring the data). Next, you will have to (further) analyze the data, and fit models using the classifiers *knn* and *logistic regression* from the general R machine learning package "caret" and interpret the results. In addition, you will apply a third machine learning / statistical learning **technique of your own choice**. This could well be a visualization tool that sheds new light on the problem, but you may also apply for instance linear discriminant analysis, random forest, principal component analysis, etc. to extend your analysis.

Your "core analysis" should be based on what you have learned throughout the course**: base R, dplyr, ggplot2 and caret**. It is permitted to use other packages as well, but only with respect to the "third technique of your own choice". You should do this if the overall quality of the analysis benefits from it. Showing off recently found "exotic" R-packages does not contribute to this aim.

Each group should appoint an editor / representative who actually makes the submission on behalf of the others. Only one student per group should submit, but all group members will receive the same grade.

The assignment will be graded as follows:

- Introduction, choice and presentation of data, research questions, motivation: 15 points
- Preprocessing / Exploratory Data Analysis: 15 points
- Modeling / Graphical Depiction / Interpretation of results: 40 points
- Conclusion: 20 points
- Overall quality and presentation; 10 points

The **structure of the research paper** should reflect the first four categories of the grading procedure:

- Introduction + RQs
- Method Section related to Preprocessing and EDA
- Method Section explaining the models you built, the steps you took and the interpretation of the main findings
- Conclusion, that wraps up your findings

Further Instructions for the editor

Your submission should contain the data set in csv- or txt- format as well as **two separate documents: an R-script and a pdf.**

Regarding the R-script:

- the R script should contain all code in order to guarantee reproducible research
- use `Group_Assigment_DemoScript.R`, from Canvas, as the basis of this script
- clarify your code with small comments, using ##
- use any function from "base R", "dplyr", "ggplot2" and "caret"
- name your script `lastname_u-number_group_project.R`
- include your name and u-number at the top of your script
- also include the full names of the other group members

Regarding the pdf:

Create **one single pdf** that comprises:

- a research paper of 1250 words (±10%)
- an appendix A containing the most relevant visualizations you created
- an appendix B with other relevant output, such as confusion matrices, evaluation metrics, or tables with regression coefficients.
- In the paper you may refer to Appendices A and B
- name your pdf `lastname_u-number_group_project.pdf`
- include your name and u-number at the top of your pdf
- also include the full names of the other group members

This is a group assignment: You are supposed to complete it within the group; directly **sharing code, plots or answers with other groups or individuals is strictly prohibited**. Submissions will be checked for potential plagiarism. Suspected plagiarism (both parties) will be referred to the Exam Board. Good luck!