**Information about the project assignment**

General information

In groups of two or three students, you will conduct a statistical study on a topic of your choice. The main analysis methods should be the methods that have been handled during the course, that is PCA, exploratory FA, CCA, cluster analysis, GMM, discriminant analysis. Group choice will be up to you. It is important that you work together on all aspects of the study. You will co-author a written report for which all group members will receive the same grade. You will also make a roughly 15-minute oral presentation to the class about your study. In the presentation you should answer to the following questions: a) What was your project/study about? b) Which analyses did you run? c) What did you find?

**Study proposal with a short data description and analysis plan should be handed in due to May 11 (max 1-2 pages). The proposal should include the names of all group members.**

Assignment

Pose a problem that you find interesting and that may be addressed (at least in part) through the analysis of data with methods learned in the class. Think of the population you want to describe and how to obtain relevant data. You should be able to find data that is relevant to your topic. You may also want to refine the research question according to the data that you found.

Some places to look for data:

* Gapminder: [www.gapminder.org](http://www.gapminder.org)
* Data.gov: <https://catalog.data.gov/dataset>
* StatLib at Carnegie Mellon: <http://lib.stat.cmu.edu/DASL/>
* World Value Survey: <http://www.worldvaluessurvey.org/wvs.jsp>
* University of Edinburgh: <http://www.ed.ac.uk/information-services/library-museum-gallery/finding-resources/library-databases/databases-a-z>
* UCI’s Machine Learning Repository: <https://archive.ics.uci.edu/ml>

Examples from the previous projects: face recognition using PCA (see p. 209 in Izenman), comparison of different methods for classifying handwritten digits, discriminant analysis for separating persons with chronic kidney disease from healthy persons.

Final submitted components: study report, R-code file, data file (.txt or .csv or .rds)

1. Written report (about 5-6 pages) should include

* a description of your data set
* a description of the statistical methods and analyses
* explanations of your results and findings
* a summary of your findings and discussion about possible limits because of the data and techniques used (recall that many of the methods considered in this course are exploratory)

NB! Tables and figures you add to the report should be commented.

1. R-code file should contain the *minimal* set of R-code that is necessary to understand your written report in full. Thus, you are not allowed to submit all the R-code that you wrote throughout the process of working on this data analysis task.