Assignment 1

In this assignment, you will both *visualize* data and *analyze* data using linear regression 'plus plus'. That is, linear regression analysis on a dataset with many predictors which makes use of subset selection or shrinkage methods, and it is tested how well the model fits the data. The assignment is to be worked on in threes or fours, for which you will receive the group division from your lab instructor. You will decide yourself which data you will use, and formulate a research question that you want to answer. Below you will find a step-by-step walk through for the assignment.

The deadline for handing in Assignment 1 is on Thursday **May 27th before your lab meeting**. The assignment should be emailed to your lab instructor as a zip file, which contains the following:

- a .Rproj file;
- a raw Rmarkdown file (.Rmd);
- a .html compiled Rmarkdown file;
- a folder containing the data you used.

Step 1:

Find yourself a suitable data set. The dataset should be

- suitable for linear regression. That is, the dataset should contain a variable or variables that is/are interesting to predict using the other variables in the dataset, and the outcome variable(s) should be on a continuous scale
- call for a data science approach. That is, it makes sense to apply subset selection or shrinkage methods in your analysis.

Suitable datasets can be found for example in/on:

- R packages
- Kaggle
- Our world in data
- The European Social Survey
- ...

Step 2:

Explore and learn about the structure of your data by constructing visualizations. Select a minimum of 2 and maximum of 3 graphs to illustrate your data in your final report.

Step 3:

Based on the content of your data and the visualizations you constructed, formulate 1 research question that you will investigate using linear regression - data science style. That is, in the linear regression, make use of either best subset selection or shrinkage methods (Ridge regression or Lasso). Select the best linear model appropriately. To ensure reproducibility of your findings, please make sure to use `set.seed()` in your `R` code.

Step 4:

Present your results in a R markdown file. In the R markdown file, the visualizations and analysis work that you did are presented in a logical order and are combined with a description of the data, a description of the

steps you have taken, the research questions you formulated and your results and conclusions regarding the research question.

In the compiled R markdown file, show both your used R code (that is, include the R code chunks) and the output. Show all your work in the Rmarkdown file, this includes any preprocessing of the used data (e.g., steps you have taken to be able to work with the data).

Grading

Your grade will be determined by:

- Originality and fit of the dataset for linear regression with model selection / regularization;
- Quality and appropriateness of the data visualizations;
- Formulation of a fitting and interesting research question;
- Quality of the performed regression analysis and analysis choices made;
- Quality of the interpretation of the model results and drawn conclusions;
- Quality of the `R` code;
- Overall quality of the report and presentation of the results.