EDSD 2025: Computer Programming E140
Contact: dudel@demogr.mpg.de
Deadline: 17. October 2025

1*: Obtain the practice data set of the German Socio Economic Panel (SOEP). You can download it here: `https://www.diw.de/documents/dokumentenarchiv/17/diw_01.c.412698.de/soep_lebensz_en.zip`. To learn more about the SOEP you can have a look at this paper: `https://doi.org/10.1515/jbnst-2018-0022`. The practice data set is a small subset of the respondents and variables covered in the SOEP.

1a: The data is stored in Stata's dta-format. Load the data set into R. You can use, for instance, the `read.dta` function of the `foreign` package.

1b: As the SOEP is a longitudinal data set, the same individuals appear several times in the data. Each individual has a unique identification number (ID). For each observation (set of measurements) in the practice data set the variable `id` has the ID of the corresponding individual. How many unique individuals are included in the practice data set? You can use the functions `unique` and `length` to answer this question.

1c: For each observation, the year of measurement is given in the variable `year`. Tabulate the number of observations per year.

1d: Restrict the data to the most recent year. Answer the following questions using the resulting, restricted data set and the functions we covered in the course: What is the proportion of females in this subset of the data? Is the average subjective health higher for men or for women? For this, you can use `by` from base-R or `group_by` from the tidyverse. Subjective health is captured by the variable `health_org` and can take on values of 1 (bad health) to 5 (very good health). Note that the variable might behave strange if you have not specified the `convert.factors` argument if you use `read.dta` to load the data.

2*: Register with the Human Mortality Database (`https://www.mortality.org/`) if you do not have an account yet. Download the data on life expectancy by gender for a country of your choice.

2a: You can load the data using the `read_csv` command (or similar), and you will likely need to use several of the arguments of the function you use. Alternatively, you can use the package `HMDHFDplus`.

2b: Visualize the trend in life expectancy at birth in the data you chose. How the visualization looks is up to you. The only requirement is that it includes the data.

2c: Visualize how the gender gap in life expectancy at birth has developed over time in your country of choice. Again, how the visualization looks is up to you. For this exercise we define the gender gap for a given year as the life expectancy of females minus the life expectancy of males.