Project I
Descriptive analysis of demographic data

The International Data Base (IDB) of the U.S. Census Bureau contains various demographic data from 1950 to the present (and projections until 2060) for over 200 countries with a population of 5,000 or more. The underlying data are based on the respective countries' official demographic statistics—e.g. in form of censuses, surveys, administrative records—or on estimates and projections by the U.S. Census Bureau itself.

The data set in the file `census_2022_2002.csv` contains a small extract from the IDB. It includes life expectancy at birth and total fertility rates for 228 countries from 2002 and 2022. For the exact definitions of these variables see `https://www.census.gov/programs-surveys/international-programs/about/glossary.html`. Life expectancy is stratified by sex. The countries are divided geographically into 5 regions and 21 subregions. For further details regarding data collection see `https://www.census.gov/programs-surveys/international-programs/about/idb.html`.

Tasks:

For tasks 1–3, consider only the year 2022.

1. Describe the frequency distributions of the variables. Consider also the differences between the sexes.

2. Describe the dependence structure of the involved variables. Are there hints towards monotonic relationships between variables? Are there hints towards functional relationships that are monotone but not linear?

3. Describe the variables' variabilities within and between subregions.

4. How have the values of the variables changed from 2002 to 2022?

You are expected to use appropriate statistical measures and graphical methods for the analysis in all parts of the project. Be sure to present your results in the appropriate form following the reporting guidelines.

Submission

Submission of the report and the corresponding program code (executable and commented) is due on **Friday, 6 May 2022 at 08:30 a.m.** All relevant files must be uploaded to Moodle.