Project I
Descriptive analysis of demographic data

The *International Data Base* (*IDB*) of the *U.S. Census Bureau* contains various demo- graphic data (currently from 1950 to 2060) on all states and regions of our world that are recognized by the US Department of State and have a population of 5000 or more. The sources of the database are information from state institutions, such as censuses, surveys or administrative records, as well as estimates and projections by the U.S. Census Bureau itself.

The dataset in the file census_2021_2001.csv contains a small extract from the IDB. It includes life expectancy at birth and total fertility rates for 228 countries from 2001 and 2021. 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:

1. Describe the frequency distributions of the variables. Consider also the differences between the sexes.
2. Are there bivariate correlations between the variables?
3. Are the values of the individual variables comparatively homogeneous within subregions and heterogeneous between different subregions? To answer this question, first compare the variability of the values in the individual subregions and then compare the values between different subregions.
4. How have the values of the variables changed over the last 20 years, i.e. comparing 2001 with 2021?

For tasks 1–3, consider only the year 2021. This project serves to practice the use of explorative and descriptive methods. Therefore, use appropriate statistical measures and graphical methods for the analysis in all parts of the project.