DAT565/DIT407
Introduction to Data Science and AI

Assignment 1
Deadline: 2024-09-12 08:00

Problem 1: Dependency ratio

The *dependency ratio* is a demographic statistic that represents the ratio of *dependent* population and the *productive* population. The most commonly used definition (such as the one used by the World Health Organization (WHO) [2]), is as follows.

The population is divided into three classes by the age: people aged 0–14 are considered *children*, people aged 15–64 are considered *labor force*, and people aged 65+ are considered the *elderly*. The children and the elderly make up the dependent population, whereas the labor force makes up the productive population.

The dependency ratio is usually expressed in units of the number of dependent people per 100 people in labor force, that is,

$$\begin{aligned}\text{Dependency ratio} &= 100 \cdot \frac{\#\text{children} + \#\text{elderly}}{\#\text{labor force}} \\ &= 100 \cdot \frac{(\#\text{people aged 0–14}) + (\#\text{people aged 65–})}{\#\text{people aged 15–64}}\,.\end{aligned}$$

This statistic is of macroeconomical interest in understanding the pressure exhibited by the demographic structure of a country on national economy. The underlying assumption is that the productive population must sustain the dependent population through labor.

The file `swedish_population_by_year_and_sex_1860-2022.csv`, available on Canvas, contains the data downloaded from SCB [1]. The data contains census information about the number of people in Sweden by the age and sex from 1860 to 2022. The data has been slightly preprocessed for easier access, by changing the character encoding and translating the names and values of certain fields from Swedish to English.

Your task is the following:

(i) Use Python tools, such as Pandas, NumPy, and Matplotlib, to read the dataset, and plot a figure that shows the dependency ratio of Sweden from 1860 to 2022,

(ii) Plot another figure that shows the fraction of the children, the elderly, and the total dependent population of the total Swedish population from 1860 to 2022, and

(iii) Discuss the development of the Swedish population in light of these figures; how have the Swedish demographics changed over the years and why, and relate this to what you know (or can find out) about general trends of population among industrialized countries.

Hints

- Please get acquainted with the documentation of the required libraries (Pandas, NumPy, matplotlib); they often contain the answers to your technical questions.
- The data has been supplied in *wide* format (also known as a pivot table in spreadsheet vocabulary); it may be useful to *unpivot* the data (convert it to *long* format) by using `pd.melt`. See Pandas documentation on how to use this function.
- You can select subsets of rows by condition with `df[`*`condition`*`]`.
- The conditions can be created by using Boolean expressions on `Series` objects, or by using functions of the `Series` such as `isin`.
- There are many ways to compute the required quantities, using the function `groupby` can be useful.
- Some functions expect function arguments, *lambda expressions* are a useful way to supply anonymous functions.
- Array functions, implemented in C under the hood, are often an order of magnitude faster than writing loops manually in Python. *You are expected* to compute, e.g., sums, differences, products, and quotients using array functions instead of iterating over the rows of the data manually.
- Remember to label the axes of your figures (including units when applicable); *you are expected* to do this.
- You are also expected to present your figures in numbered `figure` environments, with descriptive captions, and refer to the figures (preferably with `\ref`) in your text when discussing them.

Returning your report

Write a report, typeset in LaTeX, that answers *all* questions above. Include all your Python code in your report as an appendix, preferably using the `listings` package. Your report should be legible even without having a look at your code.

If you refer to outside sources, remember to add an appropriate literature reference (including websites) in references by `\cite`ing the references. It is recommended that you use the package `biblatex` to manage citations.

Place your figures in numbered `figure` environments, with descriptive captions and `\ref` to the figures in your discussion. Likewise, place your tables in numbered `table` environments with desciprive captions and `\ref` to the tables in your discussion.

After grading, you will be given another attempt to revise your report according to TA comments if it is not considered acceptable.

The deadline is *hard*. Late submissions will not be read at all and are considered failed. This means you will not get any feedback for the first round and the submission is considered a revision; there will be no third attempt, so if a late submission is failed, you will need to participate in a later iteration of the course for a re-attempt.

References

[1] Statistiska centralbyrån. *Folkmängden efter ålder och kön. År 1860 - 2022.* Retrieved 2023-10-20. 2023. URL: https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__BE__BE0101__BE0101A/BefolkningR1860N/.

[2] World Health Organization. *The Global Health Observatory Indicator Metadata Registry List: Dependency Ratio.* Retrieved 2023-10-25. 2023. URL: https://www.who.int/data/gho/indicator-metadata-registry/imr-details/1119.