Child mortality database

The goal of this exercise is to download a dataset from UNICEF, in particular, the one that I chose was on neonatal mortality, cleaning it, and then saving it in a SQLite database.

I started with importing pandas and sqlite3 to manage a relational database. The dataset was uploaded as a CSV file, specifying the ";" separator and skipping the first 14 lines that are not relevant for the analysis. I also cleaned the column names to delete possible spaces and to avoid errors during the selection and manipulation of the variables. I selected only the variables that I considered relevant for the analysis, in particular: countries' code and names, regional and subregional SDG and UNICEF classifications, and annual data on the neonatal mortality from 1989 to 2023.

The choice of selecting an annual date was taken following a completeness criterion because I included only the columns that contain an effective value, excluding those with NaN.

The only column that I kept, even if it presented some missing values, was the SDG.Subregion in which the NaNs were substituted with the term "Not classified", to preserve all the observations.

After cleaning the dataset, I exported it to an SQLite database called "neonatal_mortality_data.db", and a table called "child_mortality" was created and automatically overwritten with the cleaned data.

Finally, I made a control query to verify the correct uploading of the data in the database, visualising the first six rows of the table.

In conclusion, the data cleaning and structuring allowed us to obtain a reliable dataset that can support future analysis on the neonatal mortality trend on a global level.