Funding Health, Extending Life – Data Report

Question

Do higher health expenses per capita increase the life expectancy at birth in Latin American and Caribbean countries?

Data Sources

Data source 1: Life expectancy at birth

This data set was chosen because it contains a lot of data for many countries in the LAC region (42 countries in total). The data spans from 1960 to 2022 for most countries so there is a lot of data available. The data set is available on World Bank and was published by the United Nations Population Division [1].

The data set is available under the CC BY-4.0 license which is an open data license [2]. This license “allows users to copy, modify and distribute data in any format for any purpose [...]. Users are only obligated to give appropriate credit (attribution) and indicate if they have made any changes, including translations” [2]. I plan to give appropriate credit by citing the source of the data and indicate if I made any changes and what I changed.

Data source 2: Current health expenditure per capita

This data set contains data for most countries included in data set 1 (33 countries in total). However, it only contains data between 2000 and 2021 which limits the available data. It still provides sufficient data for analysis. The data set is available on World Bank and is part of the World Health Organization Global Health Expenditure database [3].

The data set is also available under the CC BY-4.0 license which is an open data license [2]. The same conditions as for data set 1 apply. I plan to give appropriate credit by citing the source of the data and indicate if I made any changes and what I changed.

Data Pipeline

The data pipeline was implemented using Python. It consists of three steps: Downloading the data, cleaning the data and writing both datasets into one sqlite

database. If an error occurs while downloading the data, the download is retried a maximum of two times. After three unsuccessful attempts, the pipeline is aborted. If data cleaning or writing to the database fails, the pipeline stops as well without retries.

Data cleaning is performed on both data sets separately. Data cleaning includes:

- Filtering out all countries which are not in the LAC region using the Metadata – Countries Sheet of the Excel file. With this sheet a list of appropriate Country Codes can be obtained, and the data sheet can be filtered using these codes.
- Dropping unnecessary columns like the Indicator Name and the Indicator Code, which contain information about the category of the data (life expectancy or health expenditure) and the years 1960 to 1999 as data set 2 does not contain data for these years. The indicator information is dropped because the indicators are stored in separate tables.
- Filtering the data sets so they contain data about the same countries (e.g. there was no data for Aruba in data set 2 so the corresponding rows are dropped in both data sets).
- Filling in missing values with ‘0’.

Results and Limitations

After data cleaning, both data sets are stored in the same sqlite database in separate tables. The sqlite database allows for easy data access and filtering with SQL during the analysis.

The tables contain data from 33 different countries in the LAC-region from 2000 to 2021. There are no missing values as they were filled with ‘0’. One possible problem with the data is the limited timeframe that can be analyzed.

Sources

[1] United Nations Population Division, „Life expectancy at birth, total (years) - Latin America & Caribbean,“ 2022. [Online]. Available: https://data.worldbank.org/indicator/SP.DYN.LE00.IN?locations=ZJ. [Zugriff am 16 11 2024].

[2] The World Bank Group, „Data Access And Licensing,“ 2022. [Online]. Available: https://datacatalog.worldbank.org/public-licenses#cc-by.