MADE-Datareport

Question

How does the life expectancy of populations in different countries in America compare with gross domestic product per capita?

Data Sources

Life Expectancy Data (1):

(https://genderdata.worldbank.org/en/indicator/sp-dyn-le00-in?gender=total**)**

This data set was selected because it gives me the life expectancies for countries in the years 1960 to 2022. It is differentiated whether men, women or both are considered. It is also linked to the country code. It is organised and has a fixed schema, as the data is available as a table in csv.

The licence type is CC BY-4.0 (https://datacatalog.worldbank.org/public-licenses#cc-by)

Obligations:
Users are only required to give appropriate credit (attribution) and indicate when they have made changes, including translations.

Gross domestic product per capita (2):

(https://data.worldbank.org/indicator/NY.GDP.PCAP.CD**)**

The next data set deals with the gross domestic product per capita in US dollars in the years 1960 to 2023. Here, too, the values are linked to country names to connect them with the first data set. The data set is structured in the same way as the first one, although in some countries there is no data available for GDP per capita in early years.

The licence type is CC BY-4.0 (https://datacatalog.worldbank.org/public-licenses#cc-by)

Obligations:
Users are only required to give appropriate credit (attribution) and indicate when they have made changes, including translations.

ISO-3166-Countries-with-Regional-Codes (3):

(https://github.com/lukes/ISO-3166-Countries-with-Regional-Codes/blob/master/all/all.csv**)**

The last data set contains information on the country names and country codes in connection with the continent on which these countries are located. This dataset is elementary as it allows to extract only the data on the other datasets that are in America. The dataset is structured without gaps and is available as .csv.

License: Creative Commons Attribution-ShareAlike 4.0 International License:
(https://github.com/lukes/ISO-3166-Countries-with-Regional-Codes/blob/master/LICENSE.md)
Obligations:
- User must give appropriate credit, provide a link to the license, and indicate if changes were made.

- User who remix, convert or build on the material must distribute their contributions under the same license as the original.

Data Pipeline

Description:

I am using Python 3.10.12, pandas and SQLite to implement my data pipeline. The first step of the pipeline is to download all data records. Some of these are available as a zip file. This is extracted and then all of them are assigned to a pandas dataframe.

In the next step, the third data set is used to obtain only countries and country codes that are in America. From this list of country codes, the two other data sets are then filtered so that only life expectancy and GDP data from America are available. It is then ensured that the same countries are contained in both data sets. If there are more countries in one data set because they were not present in the other, these are also discarded. After the filtering and transformations, the data is provided in a SQLite file with two tables, one for gross domestic product per capita and one for life expectancy.

Transformations / Cleaning:

- Remove unimportant columns (e.g. 'Indicator Name')
- Delete all rows that have no information on countries in America
- Unknown values that are not a number are replaced with zero for both economic data and life expectancy

Problems:

There were various problems. Firstly, there were several files in the zip files. The challenge was to extract only the correct cvs file. This was solved by searching for character strings in the names to have only one file name left.

Another problem was to make sure that both GDP per capita (1) and life expectancy data sets (2) contained the same countries. This was solved by filtering only those countries that are not in America according to dataset 3 and then selecting only those countries from GDP per capita that match these. This is a subset of the countries from dataset 3. This subset was used to query the dataset for life expectancy. If there is also a subset here, it is subtracted from the GDP per capita data set. This ensures that the same countries are included in both data sets.

Meta-quality measures:

In order to find errors, it was checked that GDP per capita cannot be negative and that the life expectancy is between 0 and 100. If the correct CSV file cannot be found in the zip, an error is displayed indicating that the string for comparing the file name must be changed. To prevent connection problems, two more attempts are made to download the data if the download fails.

Results and Limitations

The output of my data pipeline is a structured SQLite format to which I can easily send queries. All fields contain valid values that can be used for calculations. However, it should be noted that in some years there was no data available in some countries and therefore the entry is zero.

I have only left data that is relevant to the project and that answers my question. These are the country names, the code and the GDP per capita and the life expectancy for women, men and the total for all years that were available to me in the data set.

Possible problems could be that some countries are so small that the results could be falsified due to the small population. It is also important to realise that data is not available for some years in some countries.