Main Question:

Does healthcare access impact life expectancy in the Americas, specifically between the USA and Brazil?

Data Sources

Data sources were selected for multiple reasons: open license for educational purposes and the recognition of the global authorities. In addition to the four main dimensions of quality, accuracy, completeness, consistency, and relevance. The origin of the data sources are mainly two sources: the World Health Organization (WHO) which has four datasets and the World Bank Group which has the life expectancy dataset. The datasets are structured in tabular form with Comma Separated Values (CSV) format. They meet high standards of quality, characterized by accuracy, completeness, and relevance, ensuring suitability for the intended analysis. The datasets contain the following information in 5 sheets:

- **Prevalence of Hypertension:** Percentage of adults diagnosed with hypertension aged between 30 and 79.
- **UHC Index Score:** Coverage of essential health services.
- **DTP3 Immunization:** Percentage of one-year-olds who have received three doses of the combined diphtheria, tetanus toxoid, and pertussis vaccine.
- **MCV2 Immunization:** Percentage of children who have received two doses of the measles-containing vaccine by the recommended age.
- **Life expectancy:** Expected years to live in years.

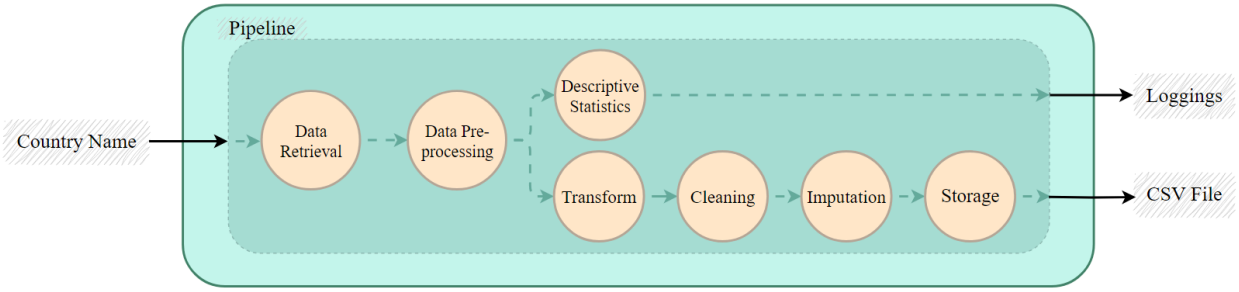
To fulfill the obligations of the licenses:

- Acknowledge the sources in all publications and analyses using the data.
- Include the required citations in any reports or presentations.

Prevalence of Hypertension sheet sample:

DIM_TIME	INDEX_N	GEO_NAME_SHORT	IND_UUID	DIM_GEO_CODE_TYPE
2005	73	Brazil	9A706FD	Country
2019	81	Brazil	9A706FD	Country

Data Pipeline



1. **High-level Description:**
    1. **Data Retrieval:** Fetching datasets from WHO and World Bank Group in a CSV format.
    2. **Preprocessing:** Cleaning and structuring the data to ensure readability and selecting the desired parts of sheets.
    3. **Descriptive statistics:** Describing the attributes of variables in each sheet
    4. **Transformation:** Merging the datasets into a single unified format based on a common column `DIM_TIME` which represents the year.
    5. **Cleaning and Imputation:** Handling missing values, duplicates and invalid entries to ensure the integrity of the final dataset.
    6. **Storage:** Storing the processed data in both CSV and SQLite formats for further use.
2. **Technologies used:**
    - Python: Programming language for scripting and data manipulation with version `3.11.7`.
    - Pandas: For efficient data processing, manipulation, and cleaning.
    - SQLite: For storing the processed data into `sqlite` format.
    - Requests and ZipFile: For downloading and extracting data.
    - Bash: To automate pipeline execution.
3. **Transformations and Cleaning Steps**
    1. **Preprocessing World Bank Data:**
        - For the World Bank Group, Skipped the first 4 columns. Dropped nulls and selected two columns for year and life expectancy to ensure the data cleanness.
        - Dropped unnecessary columns. Transposed time-series data and renamed columns for clarity. Ensured consistent data types (int64 for years and float64 for values).
    2. **Preprocessing WHO Data:**
        - For the WHO data, Filtered rows for the target country.
        - Removed unrelated columns like metadata fields.
        - Renamed columns with dataset-specific prefixes for uniqueness after merging.
    3. **Quality Check and Descriptive Statistics:**
        - Checked for duplicates and null values in each dataset before processing.
        - Generated descriptive statistics for quality assurance and to identify anomalies.

4. **Merging Data:**
    - Performed an outer join on all datasets based on the DIM_TIME column to avoid data loss while keeping all available time points.
5. **Cleaning and Imputation:**
    - Removed rows with invalid or missing values in key columns.
    - Interpolated missing numeric data using linear interpolation for time-series consistency.
    - Forward and backward filled remaining gaps to ensure a complete dataset.

4. **Problems Encountered and Solutions**
    a. The problem of inconsistent column names in variant datasets. The solution is to have renamed columns during preprocessing using descriptive prefixes.
    b. The problem of having gaps in time-series data and null values. The solution is to use interpolation and back/forward filling to impute missing values.
5. **Meta-Quality Measures and Error Handling**
    1. Applying null and duplicate checks before and after transformations.
    2. Describing the statistics of every variable are generated for all variables in all sheets to give a quick insight about the data nature.
    3. Handling missing data by using appropriate strategies (e.g., interpolation, mean imputation) to maintain data consistency.

Result and Limitations

The output of the pipeline is structured tabular data for each country stored in CSV format for simplicity and portability. The output quality dimensions are:

- **Consistency:** Uniform data format in the sheet.
- **Timeliness:** Up-to-date input data processed as per the latest availability.

Limitations

- **Imputation Effects:** Reducing accuracy and completeness by filling gaps with estimated values. As a result, the quality and accuracy measures are affected by the bias from imputation.
- **Input Dependency:** Relying on source data quality; any inaccuracies propagate into the output.

Datasources References:

World Health Organization 2024 data.who.int, United States of America [Country overview]. (Accessed on 26 November 2024)

World Health Organization 2024 data.who.int, Brazil [Country overview]. (Accessed on 26 November 2024)

World Bank Group. *Life expectancy at birth, total (years)*. Published 2022. Available at: https://data.worldbank.org/indicator/SP.DYN.LE00.IN. Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0).