Exploring the Influence of Income Inequality and Healthcare Access on Life Expectancy in the Americas

Md Obaidullah Haque (23413413)

1. Introduction

In the Americas, the ability to live a long and healthy life is often shaped by geographic location and socioeconomic status. From rural communities in South America with limited healthcare access to urban populations in North America grappling with income inequality, the challenges are diverse yet interconnected. This project aims to explore how these factors influence life expectancy across the regions, examining the interplay between income inequality and healthcare access. The insights gained from this analysis are intended to inform policymakers and healthcare professionals on strategies to address these disparities.

1.1 Analytical Questions

How do income inequality and healthcare access impact life expectancy in North and South America?

How does the level of education (e.g., literacy rate or school enrollment) influence life expectancy across different regions?

What is the relationship between GDP per capita and life expectancy across different countries?

Increase in CO2 emissions decrease life expectancy? Is it significant?

2. Datasets

This dataset on Life Expectancy at Birth provides valuable insights into the health and longevity of populations worldwide. It tracks the number of years a newborn would live if current mortality patterns persist throughout their life, serving as a key measure of public health. Despite global advancements in life expectancy—doubling to an average of 70 years since industrialization—inequalities persist, particularly between high- and low-income countries. This dataset highlights the variation in life expectancy across countries and suggests that factors such as healthcare access, education, environmental management, and economic growth play crucial roles in shaping life expectancy. Further analysis is needed to explore how the distribution of a nation's wealth, through investments in key sectors like healthcare and education, impacts overall population health outcomes.

- **Datasource1: Life Expectancy and Socio-Economic Data (World Bank)**
  - **Metadata URL**: Kaggle Metadata
  - **Data URL**: https://www.kaggle.com/datasets/mjshri23/life-expectancy-and-socio-economic-world-bank

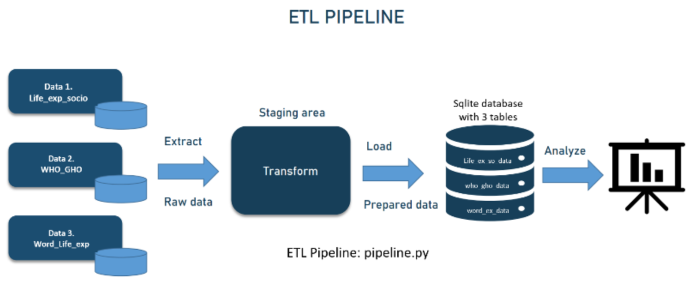
- **Description**: This dataset contains socio-economic indicators from the World Bank, including life expectancy, healthcare access, income inequality, and other relevant data for countries globally.

License: These Datasets are provided by under a Creative Commons Attribution 4.0 International License (CC BY 4.0), with the additional terms below. The basic terms may be accessed here.

- **Datasource2: WHO Global Health Observatory Data Repository**
  - **Metadata URL**: WHO GHO Metadata
  - **Data URL**: https://apps.who.int/gho/data/view.main.HEALTHEXPCAP
  - **Description**: Contains healthcare access and expenditure data across countries, providing indicators such as healthcare facilities per capita and health spending per capita.
- **Datasource3: World Bank Life Expectancy Data**
  - **Metadata URL**: World Bank Life Expectancy
  - **Data URL**: https://data.worldbank.org/indicator/SP.DYN.LE00.IN
  - **Description**: Life expectancy at birth data for countries worldwide, useful for analyzing health outcomes in relation to income inequality and healthcare investments across North and South America.

3. Data Pipeline

I implemented an automated data pipeline using Jayvee, a language designed for building data pipelines. The pipeline is composed of several key steps to extract, clean, and transform the raw data into a usable format for analysis.



3.1 Technology Used:

Jayvee for designing the pipeline, SQL for querying the processed data, and SQLite for storing the results. We used Python scripts for additional data validation and transformation when necessary.

3.2 Transformation and Cleaning Steps:

The pipeline involves the following steps: Data Extraction: Download the data directly from the World Bank's open data portal using a direct URL.

Data Interpretation and Transformation: Extract life expectancy data from the raw dataset and transform it into the required format, including renaming columns and filtering rows.

Data Validation: I validate critical fields such as life expectancy, GDP per capita, and healthcare access. Invalid data points are removed.

Data Aggregation: The data is then split into two main tables: one for life expectancy and the other for socioeconomic indicators like GDP and healthcare access.

Complex Data Extraction: Extracting specific sheets and columns from large Excel files required creating custom extraction logic to avoid irrelevant data.

3.3 Meta-Quality Measures

Error Handling: We implemented error handling at each step to ensure that if any part of the pipeline fails (e.g., an API request fails), it logs the error and continues with the next task.

Data Integrity: Automated checks for missing or invalid data ensure that only high-quality data is processed. Any missing data is flagged and handled based on the context (e.g., skipping rows or imputing missing values).

3.4 Data Structure and Quality

The data is structured as relational tables, with clear column types (e.g., TEXT for country codes, FLOAT for numeric values). After cleaning and validation, the data is of high quality, but limitations remain due to potential missing or inconsistent data across countries.

3.5 Data Format

CSV: Csv format data is chosen for its simplicity, accessibility, and versatility, making it a practical option for many data processing tasks, especially when working with tabular data.

SQLite: This was chosen because it is a lightweight, file-based database that is easy to manage and query for further analysis.

4. Critical Reflection

The quality of input data is generally high, but there are always challenges with incomplete data or non-standard reporting by different countries.

Limitations: This pipeline relies heavily on the World Bank dataset, and any changes or updates to the data source could cause issues in future analyses. Additionally, regional differences in data reporting standards could lead to inconsistencies that may not be fully addressed by the pipeline.

4.1 Conclusion

This pipeline provides a structured and automated process for transforming, cleaning, and loading World Bank life expectancy data into an SQLite database. By doing so, it helps explore the key relationships between life expectancy, income inequality, healthcare access, and other socioeconomic factors across North and South America.