Analyzing the Relationship Between Current Health Expenditure and GDP Growth in south America

Md Abdullah Al Mahmud Khosru – 23070520

This project dives into the relationship between Current Health Expenditure as a Percentage of GDP and GDP Growth in South American countries. By analyzing data from the World Bank, we aim to uncover whether increased investment in health correlates with economic growth across the reg- ion. The data pipeline automates the extraction, cleaning, and reshaping of data, integrating health expenditure and GDP growth data for 12 South American countries from 2014 to 2023. Through this analysis, we'll identify trends and patterns that can provide valuable insights into the impact of health investment on economic performance.

I. QUESTION

How does health expenditure as a percentage of GDP influence economic growth in South American countries between 2014 and 2023?

II. DATA SOURCES

I have chosen two datasets that provide comprehensive annual data on health expenditure and GDP growth. The first dataset tracks current health expenditure as a percentage of GDP for South American countries, while the second focuses on annual GDP growth rates across the same countries from 2014 to 2023. The primary reasons for selecting these datasets because together, these datasets offer a robust foundation for the project. They span the same time period (2014-2023) and cover South American countries, ensuring consistency in analysis. Combining health expenditure with GDP growth allows for a nuanced understanding of how financial priorities in health care affect economic progress.The datasets directly address the central question of whether health expenditure influences GDP growth in South America.

B. Data Structure

Health expenditure dataset contains data on current health expenditure (WorldBank Open Data) as a percentage of GDP, covering South American countries. For this project, the data from South American countries for the years 2014–2023 was extracted. And The data is in CSV format, with the following columns: Country Name, Country Code, Year, Value.

GDP growth (WorldBank Open Data) dataset contains data on GDP growth (annual percentage change in GDP) for South American countries. For this project, the data for South American countries from 2014 to 2023 was extracted. The data is in CSV format, with the following columns: Country Name, Country Code, Year, Value

C. Data Quality

The health expenditure dataset has several key dimensions of data quality. Accuracy is ensured as the data accurately reflects real-world. However, there were some instances where certain countries did not report data for all years. These missing values were handled by removing the records for those years, ensuring that the analysis only included complete data for each country. The health expenditure dataset was successfully cleaned and reshaped to include only the relevant countries (South

American countries) and the desired years (2014–2023)

Similarly, the GDP growth dataset also demonstrates robust data quality. However, there were some missing data points for specific years, which were cleaned by focusing on the relevant countries. These missing data points were identified and removed, ensuring that only complete data points were used in the analysis, thus maintaining the integrity of the dataset.

D. Licenses: Both datasets are under the World Bank Terms of Use for Open Data (source). These datasets are licensed under the Open Data Commons Attribution License (ODC-BY 1.0).

E. Obligations: Attribution is required when using the data. This is fulfilled by citing the World Bank as the source. License Source is Data Access and Licensing

III. DATA PIPELINE

1. Data Extraction
A. Download the Data: The pipeline fetches data from the World Bank API using HTTP requests.For each dataset (Health Expenditure and GDP Growth), the CSV files are downloaded in ZIP format. Health Expenditure and GDP Growth

B. Extract ZIP Files: The pipeline uses Python's zipfile library to extract the CSV files from the downloaded ZIP archives. The files are saved in the specified data directory (./data).

2. Data Cleaning and Transformation
A. Filter Data for South American Countries: The dataset is filtered to include only data for South American countries (e.g., Argentina, Brazil, Chile , etc.)

B. Select Relevant Columns: For both datasets, only the following columns are retained: Country Name, Country Cod, Year, Value.

C. Reshape Data from Wide to Long Format: The data is transformed from a wide format (where each year is a separate column) to a long format (where each year is a row). This allows easier analysis across multiple years. The reshaped data contains three columns: Country Name, Country Code, Year, and Value (Health Expenditure or GDP Growth)

3. Data Storage:
A. Store Cleaned Data: The cleaned and reshaped data is stored in two formats. (i) CSV Files: The cleaned data is saved as CSV files for easy sharing and further processing. (ii) SQLite Database: The cleaned data is stored in an SQLite database for efficient querying and long-term storage.
B.Database: data_cleaned_south_america.db
C.Tables created: gdp_data_south_america and health_expenditure_data_south_america

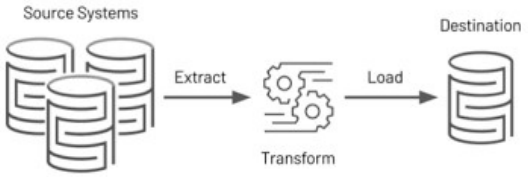

Fig 1: ETL Data Pipeline Architecture

4.Technologies

- ✓ Programming Language: Python
- ✓ Libraries: pandas (data processing), requests (data retrieval), sqlite3 (database integration), zipfile (file extraction).

5.Challenges and Solutions: The project faced several challenges, including irrelevant metadata in ZIP files, which was filtered out, and wide format data, which was reshaped into a long format using Pandas' melt(). Inconsistent encodings in CSV files were handled with dynamic encoding and fallback options.

Managing SQLite table updates could cause redundancy, but using the if_exists="replace" parameter resolved this, ensuring data was updated without conflicts. These solutions kept the data pipeline smooth and accurate.

IV. RESULT

The datasets for health expenditure and GDP growth were cleaned and reshaped, though some countries had missing data, which was handled by removing incomplete records. While health expenditure is crucial for public health, its direct impact on economic growth is likely influenced by other factors such as political stability, economic policies, and foreign investment.

A.CSV files: Suitable for immediate inspection or use in analysis tools like Python.
B.SQLite Database: Enables efficient querying and analytical applications. Columns: "Country Name," "Country Code," "Year," and "Value" (either GDP growth or health expenditure).

C. Correlation: The figure shows a weak positive correlation between health expenditure (as a percentage of GDP) and GDP growth in South America from 2014 to 2023. While higher health spending is generally linked to slightly higher GDP growth, the correlation is not strong enough to confirm that increased health expenditure directly drives economic growth. The scatter plot's dispersed data points suggest other factors also influence GDP growth.

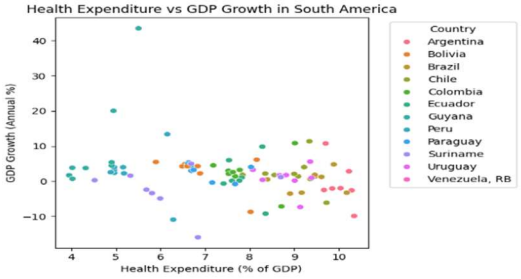

Fig2: Health Expenditure vs. GDP Growth in South America (2014–2023)

D.Output Dataset: The cleaned datasets look like

GDP Growth

	A	B	C	D
1	Country Nar	Country Co	Year	Value
2	Argentina	ARG	2014	-2.512615321
3	Bolivia	BOL	2014	5.460569506
4	Brazil	BRA	2014	0.50395574
5	Chile	CHL	2014	1.792649474
6	Colombia	COL	2014	4.499030001

Table 1: GDP Growth

Health Expenditure

	A	B	C	D
1	Country Name	Country Cod	Year	Value
2	Argentina	ARG	2014	9.67129993
3	Bolivia	BOL	2014	5.89613628
4	Brazil	BRA	2014	8.39644146
5	Chile	CHL	2014	7.84150124
6	Colombia	COL	2014	7.18553209

Table 2: Health Expenditure

V.LIMITATIONS:
Here are some limitations that I observed: Some years or countries may have missing or incomplete data, which could affect the accuracy of the analysis. The dataset only includes current health expenditure, excluding long-term investments or capital spending, limiting the scope of health investment analysis. The analysis shows a correlation but does not establish causality between health expenditure and GDP growth.

VI. CONCLUSION
In this project, we analyzed the relationship between current health expenditure as a percentage of GDP and GDP growth in South American countries from 2014 to 2023. Using an automated data pipeline, we successfully extracted, cleaned, reshaped, and stored the relevant data. The analysis highlighted potential trends between health expenditure and economic growth, though the relationship is complex and influenced by factors beyond health spending alone. While the data was cleaned and processed efficiently, limitations such as missing values and inconsistencies were noted. Future work could include incorporating additional variables like education or infrastructure spending to gain a deeper understanding of how health expenditure impacts GDP growth. Overall, the project provides valuable insights for policymakers on the role of health investment in economic development.

REFERENCES:[1] “Data Pipeline Architecture - A Deep Dive — StreamSets,” Software AG. (accessed 28-Nov, 2024)