Analyzing the Relationship Between GDP, Public Education Expenditure, and Illiteracy Rates, Education Level, and Economic Activity Participation in Brazil, Colombia and Peru (2002–2020)

1. Introduction:

Main Question:

How do GDP and public expenditure on education relate to selected educational outcomes in Brazil, Colombia and Peru from 2002 to 2020, specifically in terms of illiteracy rates, population with 13 years or more of education, and economic activity participation among the population with 13 years or more of education*?*

Motivation:

Education is a cornerstone of societal and economic development, and its interplay with economic conditions has long been a topic of interest for policymakers and researchers alike. Despite efforts to enhance educational access and quality in Latin America, disparities in educational and economic outcomes persist. This project explores how GDP and public expenditure on education correlate with key educational outcomes in Brazil, Colombia, and Peru between 2002 and 2020. By focusing on illiteracy rates, the population attaining 13 or more years of education, and their economic activity participation, this study seeks to uncover valuable insights into the dynamics between economic investments and educational opportunities in these countries.

2. Data Sources:

For this project, I utilized five datasets from two different sources. Due to significant data gaps in Latin American countries, I narrowed the scope to focus on Brazil, Colombia, and Peru within the years 2002 to 2020 to maximize the use of real data and avoid reliance on artificial data. Despite this restriction, there were still some missing values in the selected datasets.

Licenses:

I selected these datasets due to their relatively low number of missing values compared to other available options. Both datasets are provided under open data licenses, ensuring their suitability for academic and educational use. The data from CEPAL (United Nations Economic Commission for Latin America and the Caribbean) is made available under the CEPAL Open Data policy, which allows for public use with proper attribution. While there is no specific direct link regarding CEPAL data usage, the datasets are accessible through the open data portal, making them suitable for public use in academic research. Similarly, the World Bank data is governed by its Open Data Initiative, providing free access to its datasets under a Creative Commons Attribution 4.0 International License. As this report is for academic purposes and includes proper attribution to the original sources, I ensure compliance with the licensing terms of both CEPAL and the World Bank by referencing to the datasets used. This approach guarantees adherence to their data use guidelines while maintaining the integrity of this research.

CEPAL Open Data Related Link: API datos abiertos - CEPALSTAT Bases de Datos y Publicaciones Estadísticas

World Bank Data License Link : Data Access And Licensing | Data Catalog

Data Source Description:

In this project, all datasets are structured data provided in Excel format. The data required for analysis are largely complete, with only a minimal number of missing values. As historical data spanning 2002 to 2020, the datasets are timely and highly relevant to the main research question.

Overall, there are 5 datasets in 2 main groups:

Data Source 1: CEPAL Statistics and Indicators. This source provides four key datasets used in the analysis:

- Public expenditure on education
- Illiteracy rates
- Population aged 15 years and older, categorized by years of education
- Population aged 15 years and older participating in economic activity, categorized by years of education

These datasets offer detailed insights into education-related indicators across Brazil, Colombia, and Peru. The data are structured and organized by year, ensuring relevance to the study's timeline and research objectives.

Metadata URL: CEPALSTAT   Data URL: Not Available

Data Source 2: World Bank - GDP per capita (current US$). This dataset provides GDP per capita data for Brazil, Colombia, and Peru, measured in current US dollars, from 2002 to 2020. The dataset is structured by year and offers crucial information on the economic conditions of these countries.

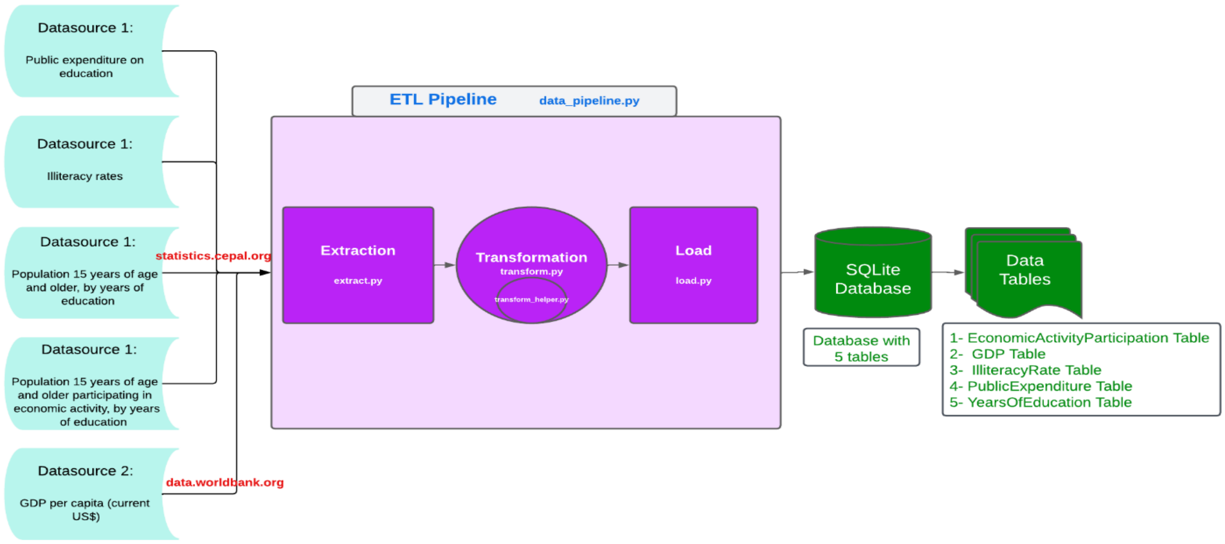
Metadata URL: GDP per capita Data URL: Not Available

3. Data Pipeline:

An ETL pipeline in Python was developed to process these datasets and integrate them into one SQLite database. The script will automatically extract data, transform it, and load it into the database while ensuring consistency across all the datasets. The missing values were dealt with using linear interpolation on the transformation step. All required Python packages are listed in the requirements.txt file, simplifying the installation process. This approach allows keeping the data efficiently managed and promotes easy analysis of the integrated data.

Pipeline:

As depicted in the diagram below, the project utilizes a structured ETL (Extract, Transform, Load) pipeline approach.



Extraction:

In this step, all sources of data will be gathered from different online platforms through a web automation approach. The shared links of data do not allow downloading the data directly; hence, web automation for the extraction of data will be implemented using the Helium library. Each step of the process is logged for traceability, and in case of errors, the Retry library is used to retry up to three times with 10-second intervals between attempts. This ensures a robust and smooth process for data collection.

Transformation:

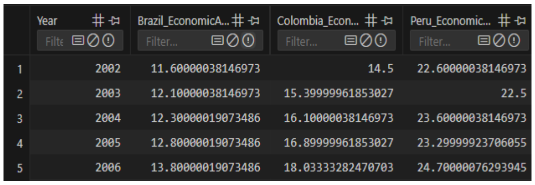
In this project, the transformation phase is organized in a structured manner, yet managing different datasets in terms of common workflow and specific needs. In general, the process of cleaning, restructuring, and refining the data involves preparing it for analysis. The general processes involve dropping unnecessary columns, renaming key fields for consistency, and filtering records of interest over relevant data points-such as countries (Brazil, Colombia, and Peru) and years of interest (2002-2020). With each dataset, further steps are taken according to the characteristics of that particular dataset. This includes pivoting columns to restructure data for comparison; appending country-specific suffixes to ensure clarity in the final dataset. Where applicable, missing values are interpolated to enhance completeness. Notably, the GDP dataset applies a melting operation to deal with its temporal structure, whereas other datasets have a much simpler pattern of filtering and reshaping. In addition, reliability in the transformation phase is ensured by robust error handling mechanisms using retry logic along with detailed logging. Finally, all datasets are standardized into the same format, ensuring a uniform DataFrame structure with consistent columns. This uniformity across datasets facilitates seamless integration and comparison in subsequent analysis phases.

Load :

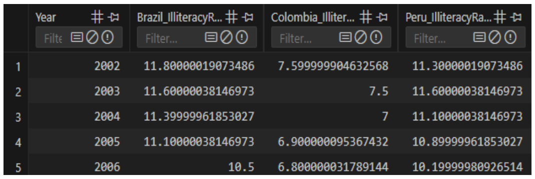
In this phase, five datasets are saved into the "educationAndEconomy_BrazilColombiaPeru.db" database as five separate tables.

4. Result and Limitations :

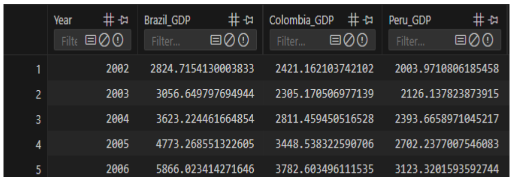
The final result depicts the outline of the five tables of SQLite database:



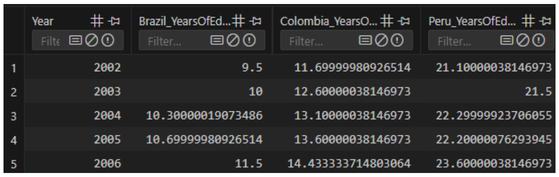
	Year	Brazil_EconomicA…	Colombia_Econ…	Peru_Economic…
1	2002	11.60000038146973	14.5	22.60000038146973
2	2003	12.10000038146973	15.39999961853027	22.5
3	2004	12.30000019073486	16.10000038146973	23.60000038146973
4	2005	12.80000019073486	16.89999961853027	23.29999923706055
5	2006	13.80000019073486	18.03333282470703	24.70000076293945



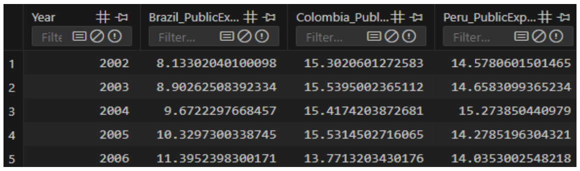
	Year	Brazil_IlliteracyR…	Colombia_Illiter…	Peru_IlliteracyRa…
1	2002	11.80000019073486	7.599999904632568	11.30000019073486
2	2003	11.60000038146973	7.5	11.60000038146973
3	2004	11.39999961853027	7	11.10000038146973
4	2005	11.10000038146973	6.900000095367432	10.89999961853027
5	2006	10.5	6.800000031789144	10.19999980926514



	Year	Brazil_GDP	Colombia_GDP	Peru_GDP
1	2002	2824.7154130003833	2421.162103742102	2003.9710806185458
2	2003	3056.649797694944	2305.170506977139	2126.137823873915
3	2004	3623.224461664854	2811.459450516528	2393.6658971045217
4	2005	4773.268551322605	3448.538322590706	2702.2377007546083
5	2006	5866.023414271646	3782.603496111535	3123.3201593592744



	Year	Brazil_YearsOfEd…	Colombia_YearsO…	Peru_YearsOfEd…
1	2002	9.5	11.69999980926514	21.10000038146973
2	2003	10	12.60000038146973	21.5
3	2004	10.30000019073486	13.10000038146973	22.29999923706055
4	2005	10.69999980926514	13.60000038146973	22.20000076293945
5	2006	11.5	14.433333714803064	23.60000038146973



	Year	Brazil_PublicEx…	Colombia_Publ…	Peru_PublicExp…
1	2002	8.13302040100098	15.3020601272583	14.5780601501465
2	2003	8.90262508392334	15.5395002365112	14.6583099365234
3	2004	9.6722297668457	15.4174203872681	15.273850440979
4	2005	10.3297300338745	15.5314502716065	14.2785196304321
5	2006	11.3952398300171	13.7713203430176	14.0353002548218

I chose SQLite for its lightweight nature, making it ideal for handling this dataset. While I initially aimed to conduct an analysis over a broader geographic and temporal range, I encountered significant data gaps, especially with respect to certain regions and years. Unfortunately, there are limited sources that provide data in this specific area. Although I considered supplementing the data artificially, I opted to work with actual data as much as possible. As a result, I had to narrow the scope to minimize missing values and used linear interpolation to fill in the few remaining gaps. In terms of data quality, I believe the accuracy and consistency of the data are satisfactory, given that the data types and ranges align with expectations. However, the current limitations suggest that with more robust data availability in the future, a more comprehensive and insightful analysis could be conducted.