Correlation Between Education Spending and GDP Growth in North America

(KM Rashedul Alam, 23008271)

i. Question

The main question for this project is: "**How does government expenditure on education as a percentage of GDP correlate with GDP growth in North American countries from 2016 to 2023?**"

ii. Objective

This question investigates the potential relationship between public investment in education and economic growth, providing insights into how education funding impacts financial performance from 2016 to 2023.

iii. Data Source

For this project, I have chosen two datasets that offer comprehensive data on expenditure on education and GDP growth annually.

Source 1: Government Expenditure on Education (% of GDP)

- **Source Name:** World Bank Open Data
- **Description:** This dataset contains data on government expenditure on education as a percentage of GDP, covering countries worldwide. For this project, only data from North American countries for 2016–2023 were extracted.
- **Structure:** CSV format, with columns for "Country Name," "Country Code," and annual expenditure values for each year (1960–2023).
- **Quality:** The data is generally well-structured, but it includes missing values for some years and countries. It is provided in a wide format with years as separate columns.
- **Why These Datasets Were Chosen:** This dataset was chosen because it quantifies public investment in education, a critical indicator for assessing the priority governments give to building human capital**.**

Source 2. GDP Growth (Annual %)

- **Source Name:** World Bank Open Data
- **Description:** This dataset contains GDP growth rates for countries worldwide. We focused on North American countries for the years 2016–2023.
- **Structure:** It is also like the education expenditure dataset, it is in CSV format with columns for "Country Name," "Country Code," and GDP growth rates for each year (1960–2023).
- **Quality:** Like the education dataset, this data contains some missing values and is structured in a wide format.

- **Why These Datasets Were Chosen:** This dataset provides annual economic growth rates, a key metric to evaluate the potential impact of education spending on the overall economic performance of North American countries.

iv. 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).

Obligations: Attribution is required when using the data. This is fulfilled by citing the World Bank as the source.

License Source: Data Access and Licensing

v. Data Pipeline

The data pipeline was designed to:

1. **Download:** Retrieve datasets from World Bank APIs in ZIP format.
2. **Extract:** Unzip the files and identify relevant CSVs while excluding metadata.
3. **Filter:** Focus on North American countries and the years 2016–2023.
4. **Transform:** Convert datasets from wide to long format, reshaping year columns into rows.
5. **Clean:** Address missing values by removing or interpolating them where appropriate.
6. **Store:** Save the cleaned data sets in both CSV format and as tables in an SQLite database for further analysis.

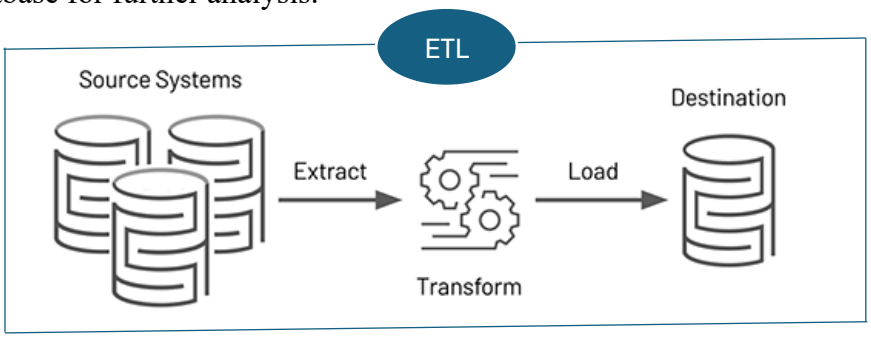

Fig 1: ETL Data Pipeline Architecture

vi. Technologies

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

vii. Transformation and Cleaning Steps:

1. **Download and Extract Data:** The pipeline downloads datasets from the World Bank's API in ZIP format. It then extracts all files into a specified directory and identifies CSV files while excluding metadata files.
2. **Filter by Country and Year:** Using a predefined list of North American countries, the pipeline filters the datasets to retain only relevant rows. It also selects columns for the years 2016–2023.
3. **Wide-to-Long Transformation:** The pipeline reshapes the data from wide format (with years as columns) to long format. This conversion creates a "Year" column and a "Value" column, making the data more suitable for analysis and storage.
4. **Save Cleaned Data to CSV:** The cleaned and reshaped data is saved as CSV files for easy inspection and further processing.
5. **Export Data to SQLite:** The pipeline stores the cleaned datasets in an SQLite database, with separate tables for GDP and education expenditure data. This allows efficient querying and integration with analytical tools.

viii. Challenges and Solutions

1. **ZIP files contained irrelevant metadata:**
   - Filtered and processed only relevant CSV files, excluding metadata.
2. **Original datasets were in wide format, unsuitable for analysis:**
   - Reshaped data using pandas.melt for a long-format structure.
3. **CSV files had inconsistent encodings:**
   - Handled encoding dynamically, with fallback to alternative encodings.
4. **Managing table updates in SQLite could cause redundancy:**
   - Used if_exists="replace" to keep tables updated without conflicts.

ix. Results and Limitations

Output Data: The cleaned datasets are stored in:

CSV files: Suitable for immediate inspection or use in analysis tools like Python.

SQLite Database: Enables efficient querying and integration with analytical applications.

Structure:

Columns: "Country Name," "Country Code," "Year," and "Value" (either GDP growth or education expenditure).

Rows: One row per country-year combination for the years 2016–2023.

Output Dataset: The cleaned datasets look like this.

Country Name	Country Code	Year	Value
Antigua and Barbuda	ATG	2016	2,25177E+14
Bahamas, The	BHS	2016	2,58661E+14
Belize	BLZ	2016	5,72171E+14
Bermuda	BMU	2016	1,91544E+14
Barbados	BRB	2016	4,68383E+14

Table 1: Government Expenditure on Education

Country Name	Country Code	Year	Value
Antigua and Barbuda	ATG	2016	4,09977E+14
Bahamas, The	BHS	2016	-9,61909E+14
Belize	BLZ	2016	-1,10344E+14
Bermuda	BMU	2016	-6,57206E+14
Barbados	BRB	2016	2,55268E+14

Table 2: GDP Growth

Data Quality: The cleaned data ensures minimal missing values and a consistent format. However, interpolated values may introduce slight inaccuracies.

Limitations: Some limitations that I have found:

- **Data Gaps:** Some countries had missing values which leads to datasets being incomplete. And interpolated values may not accurately reflect actual trends.
- **Correlation vs. Causation:** The analysis might indicate relationships but cannot prove causation between education spending and GDP growth.
- **Scope:** Only North American countries are included, which limits generalizability.

x. Conclusion

This project explores the relationship between government expenditure on education and GDP growth in North America from 2016 to 2023. By extracting high-quality datasets from the World Bank, it offers insights into how investment in education correlates with economic performance. Despite challenges such as missing data and the complexity of reshaping datasets, the project successfully produced clean, analysis-ready outputs. These findings contribute to understanding the broader implications of education funding on economic growth, providing a foundation for further research and policymaking in this area.