**METHODS OF ADVANCED DATA ENGINEERING WS24/25**

I. QUESTION:

How does educational access influence economic mobility in various regions across the Americas?

II. DATA SOURCES

For the purpose of conducting this study, 2 different datasets were considered from ***World Bank Open Data***. The details of these datasets are given below:

| S. NO. | DATA NAME | DATA RANGE | Metadata URL |
|---|---|---|---|
| 1 | School enrollment, primary (% gross) | 1970 - 2023 | Click Here |
| 2 | GDP per capita (current US$) | 1960 - 2023 | Click Here |

Both of the data sources listed above are licensed under **CC BY-4.0.** This allows us to copy, modify, and use the data for any purpose, including commercial, as long as proper attribution is provided and modifications are disclosed. Information on the same can be accessed through the URL above.

To comply, the project will:

- Credit the World Bank as the source in the report and derived materials.
- Document any transformations made to the dataset.

Data Source 1:

**Definition**: Gross enrollment ratios indicate the capacity of each level of the education system, but a high ratio may reflect a substantial number of overage children enrolled in each grade because of repetition or late entry rather than a successful education system. The net enrollment rate excludes overage and underage students and more accurately captures the system's coverage and internal efficiency. Differences between the gross enrollment ratio and the net enrollment rate show the incidence of overage and underage enrollments. (Source)

**Limitations and Exceptions:** Enrollment indicators are based on annual school surveys, but do not necessarily reflect actual attendance or dropout rates during the year. Also, the length of education differs across countries and can influence enrollment rates, although the International Standard Classification of Education (ISCED) tries to minimize the difference. Moreover, age at enrollment may be inaccurately estimated or misstated, especially in communities where registration of births is not strictly enforced.

**Relevance in project**: This data source provides information on enrollment statistics in primary schools. This is crucial in understanding how education levels correlate with economic performances.

**Structure**: The dataset is extracted as a csv, which includes Country Name, Country Code, Gross Enrollment Ratio across the years.

Data Source 2

**Definition**: GDP per capita is gross domestic product divided by midyear population. GDP is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in current U.S. dollars.

**Limitations and Exceptions**: GDP per capita does not account for income inequalities and variations in the cost of living between countries. Also, since the data is presented in US dollars, it may introduce exchange rate volatility as a factor in comparisons. Moreover, it does not reflect sustainability or well-being.

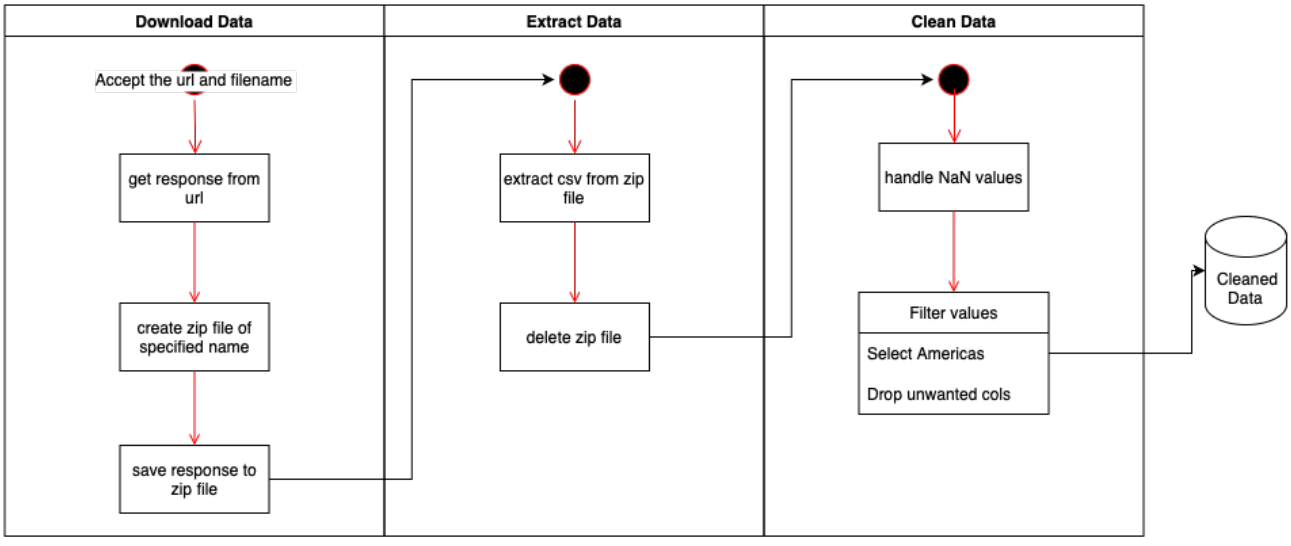
**Relevance to project:** This data source provides critical indications to assess how correlation between economic performance and primary education enrollment rates.

**Structure**: The dataset is extracted as a csv, which includes fields such as Country Name, Country Code and GDP per capita values for multiple years.

| QUALITY DIMENSIONS | DATA SOURCE 1 | DATA SOURCE 2 |
|---|---|---|
| ACCURACY | Captures real-world enrollment numbers from verified education sources. | Based on official economic data, ensuring reliable figures. |
| COMPLETENESS | Provides enrollment ratios by country and year, but doesn't include attendance or dropout details. | Offers GDP per capita for all countries and years, though it doesn't account for inflation or purchasing power. |
| CONSISTENCY | Data is well-organized with a standardized structure across all years and countries. | Follows a consistent format, with clear fields like country names, codes, and yearly GDP values. |
| TIMELINELESS | Updated regularly, offering both current and historical enrollment data. | Frequently updated, covering both present and past economic data. |
| RELEVANCE | Essential for understanding how education relates to economic performance. | Directly relevant for analyzing economic trends and their connection to education. |

## III. DATA PIPELINE

The pipeline processes two datasets: education data and economic data. This is done by automating data download, extraction, cleaning and preparation for further analysis.



**Technology**: The data pipeline was implemented in Python. Additional libraries like requests for downloading data, zipfile for extraction, and pandas for data manipulation were also utilized.

**Transformation and Cleaning**:

- The dataset was filtered on a predefined list of countries in the Americas
- Missing numerical values were replaced with the mean for that column
- Relevant columns such as Country Name, Country Code and data from 1970-2023 were retained
- Values were standardized to ensure columns had consistent data types

**Problems and solution**:

- Direct download from link produced a zip file with multiple files. The code was adjusted to extract only relevant csv file
- There were significant gaps in fields for specific years. This was addressed by substituting to preserve data usability
- Columns were being interpreted as object types due to mixed content. This was standardized using infer_objects before data transformation

**Meta-quality measures**:

- Incorporated try-except blocked to manage runtime issues
- Pipeline can handle changing filenames
- Implemented checks to ensure cleaned dataset contains expected columns and valid entries after transformations
- Modular design allows new datasets to be added without significant code restructuring

## IV. RESULT AND LIMITATIONS

The pipeline produced two cleaned CSV files, one from each of the data sources. These datasets were filtered to include only relevant countries and cleaned to fill NaN values with column-wise mean, ensuring no gaps remain in the temporal series.

**Structure and Quality:** Each dataset is tabular with rows representing the countries, and columns include the country name, code and yearly data. The dataset was ensured for consistency by removing irrelevant countries and handling missing values. However, the substitution method may introduce fsome assumptions which may not represent real-world trends accurately. This may result in bias.

**Output Data Format**: The pipeline outputs the data as sepearte CSV files. This is easier to realize as Python DataFrames, which allows for simple data transformation and analysis.

**Critical Reflection**:

- Data is cleaned, standardized and limited to relevant countries and years
- CSV ensures ease of use and interoperability
- NaN substitution may obscure trends in enrollment or GDF if missing values are not randomly distributed
- Since the data sources contain aggregated metrics, they may miss nuances within countries such as regional disparities or population demographics
- Comparability of data may be limited due to differences in reporting practices