data-report

November 27, 2024

1 Data Report

Question: How have North America and Latin America contributed to renewable energy adoption and climate mitigation efforts, and how do these trends compare from 1960 to 2023?

Data Source For this project, I have considered the datasets of North America and Latin America and Caribbean regions' from World Bank Group. These datasets provide a comprehensive view of different indicators namely Climatic Change, Education, Environment, Gender, Health, Infrastructure, Poverty, etc. The dataset also includes indicators such as renewable energy output, fossil fuel consumption, greenhouse gas emissions, and carbon dioxide emissions. It covers energy efficiency and access to clean fuels. It spans multiple countries, enabling both regional and country-level analysis. The data spans from 1960 to 2023.

Region	Data Source
North America	Download
Latin America & Caribbean	Download
Metadata URL	Link
Data Type	csv

Each indicator is categorized by country, ensuring clarity and organization. The data quality is high, meaning that, it is well organized with standardized methods and periodic updates by the World Bank.

Data License The World Bank strives to enhance public access to and use of data that it collects and publishes. The data are organized in datasets listed in The World Bank Data Catalog (the "Datasets").

Data Pipeline

- **Overview**

  The data pipeline involved integrating economic and environmental indicators for both North America and Latin America & Caribbean from the World Bank API. The steps included the following:

1. Data Ingestion: Download raw CSV files from the World Bank.
2. Data Extraction: Extract the data from the zip files. The zippped files contained three csv files out of which only two were necessary. One with the main data and the other contained the metadata of the indicators.
3. Data Preparation: There were altogether 4 CSV files, two of each regions. The `Unnamed` columns and `NaN` values had to removed. Also the missing values were replaced with 0 or interpolated where needed to ensure continuity in time series data.
4. Data Integration: The idea was to merge the data and metadata into one dataframe and for this the column in the metadata had to be renamed to `Indicator Code` for consistency.
5. Storage: By now there were only two CSV files created after merging for each regions. Then columns of these files were renamed as below to keep it consistent while storing in SQLite.

```
"Country Name": "CountryName",
"Country Code": "CountryCode",
"Indicator Name": "IndicatorName",
"Indicator Code": "IndicatorCode",
"SOURCE_NOTE": "SourceNote",
"SOURCE_ORGANIZATION": "SourceOrganization",
```

- **Technologies Used**
  - Programming: Python (libraries: `pandas`, `requests`, `zipfile`, `sqlalchemy`)
  - Database: SQLite for structured, light weight and queryable data storage.
  - File Handling: Python's I/O functions for file extraction, management and deletion.

Challenges and Solutions

1. **Inconsistent Column Names Across Files**
   - Problem: The main data files and metadata files had inconsistent column naming conventions (e.g., `Country Name` vs. `CountryName`). This caused issues during the merging process.
   - Solution: Standardized column names during the data cleaning step to ensure compatibility. For example, all spaces in column names were replaced with underscores, and column names were matched to the database schema.
2. **Missing and Inconsistent Data**
   - Problem: Many countries within the datasets had missing values for certain years, especially in earlier decades when data collection was not consistent across all regions.
   - Solution: Applied interpolation or filled missing values with `0` based on the indicator's nature.
3. **Temporary File Management**
   - Problem: Downloaded zip files and extracted CSV files occupied disk space, which could become an issue if left unmanaged, especially when working with large datasets.
   - Solution: Created a temporary directory to store the downloaded and extracted files. After successfully storing the cleaned data in SQLite, the temporary directory and its contents were deleted using `shutil.rmtree`.This ensured efficient use of disk space.

4. **Data Integration Issues**
    - Problem: Merging the main data files with metadata required ensuring that keys like Indicator Code were consistent. Mismatched keys resulted in incomplete merges.
    - Solution: Preprocessed both datasets to align their keys by trimming whitespace, correcting case sensitivity, and ensuring unique identifiers were consistent across files.

Results and Limitations

- **Output Data**
    - *Format:* Data is stored in an SQLite database with a schema matching the cleaned data structure. Each region has a separate table (NAC for North America and LCN for Latin America Caribbean).
    - *Structure:* Tables include columns for CountryName, CountryCode, IndicatorName, IndicatorCode, yearly values (1960–2023), and metadata such as source notes.
- **Data Quality**
    - *Strengths:*
        * Comprehensive time-series data with minimal missing values for core indicators.
        * Metadata ensures interpretability of each indicator
        * Standardized format improves usability for analysis.
    - *Weaknesses:*
        * Interpolated missing values may introduce inaccuracies in trends.
        * Some indicators may have limited data coverage, reducing comparability across countries.
- **Output Format**
    - Why SQLite?
        * Allows structured queries for analysis.
        * Compact and efficient for storage and data access.
        * Easily portable for integration with analysis tools like Python and R.

Critical Reflection

1. **Strengths:**
    - The pipeline ensures high data integrity and aligns with analysis goals (e.g., exploring climate trends).
    - Standardized processing makes it adaptable for other datasets or future updates.
2. **Limitations:**
    - Inconsistent Start Years for Data Collection
        - Not all countries within a region began collecting data at the same time. Some countries may have data available starting from the 1960s, while others might only have records beginning in the 1990s or later. This results in missing values for earlier years in certain indicators, leading to potential inconsistencies.
        - *Impact*:
            * Temporal analysis might be biased, as early averages or trends for a region may reflect only a subset of countries.
            * Interpolation or filling missing values might introduce artificial trends that do not represent the reality for all countries.
        - *Example:*
            * Renewable energy data might only be available for developed countries initially,

which could distort early averages for the entire region.

3. **Mitigation Measures**
   - *Data Availability Filters:* Analyzing trends only for years where most countries have complete data, ensuring a more balanced representation.
   - *Separate Analysis for Early and Late Adopters:* Grouping countries by the year they started collecting data to provide insights into early versus late adopters of climate-related policies.