Project Work 4 - Data Report

By Sonu Pathak

1 Question

How does final energy consumption by sector correlate with net greenhouse gas emissions and climate-related expenditure?

2 Data Sources

2.1 Data Source Descriptions

Data Source 1: Final Energy Consumption by Sector

- **Source:** Eurostat API
- **URL:** `https://ec.europa.eu/eurostat/api/dissemination/sdmx/2.1/data/ten00124?format=TSV&compressed=false`
- **Description:** Provides data on final energy consumption by sector, allowing analysis of energy usage patterns.
- **Data Structure:** Tabular format with columns representing different sectors and rows representing time periods.
- **Data Quality:** The data is collected and maintained by Eurostat, ensuring high quality and reliability. However, it may contain occasional missing values or inconsistencies that need to be addressed during preprocessing.
- **License:** Eurostat data is usually licensed under open-data licenses, allowing free use with proper attribution. We plan to fulfill our obligations by appropriately attributing the data in our reports and analyses.

Data Source 2: Net Greenhouse Gas Emissions

- **Source:** Eurostat API
- **URL:** `https://ec.europa.eu/eurostat/api/dissemination/sdmx/2.1/data/sdg_13_10?format=TSV&compressed=false`
- **Description:** Contains information on net greenhouse gas emissions, facilitating analysis of environmental impact.
- **Data Structure:** Similar to the first data source, in tabular format with columns representing different variables and rows representing time periods or geographic regions.
- **Data Quality:** Maintained by Eurostat, the data is generally of high quality, but may require cleaning and preprocessing to address any anomalies or inconsistencies.
- **License:** As with other Eurostat data, this dataset is typically licensed under open-data licenses, allowing free use with proper attribution. We will ensure compliance with the licensing terms by providing appropriate attribution in our reports and analyses.

Data Source 3: Climate-related Expenditure

- **Source:** Eurostat API

- **URL:** `https://ec.europa.eu/eurostat/api/dissemination/sdmx/2.1/data/sdg_13_50?format=TSV&compressed=false`
- **Description:** Provides data on climate-related expenditure, enabling assessment of financial contributions to environmental initiatives.
- **Data Structure:** Similar to the previous datasets, in tabular format with columns representing different variables and rows representing time periods or geographic regions.
- **Data Quality:** Maintained by Eurostat, the data is generally reliable but may contain errors or missing values that require preprocessing.
- **License:** Eurostat data is typically available under open-data licenses, allowing free use with proper attribution. We will ensure compliance with licensing terms by providing appropriate attribution in our reports and analyses.

3 Data Pipeline

3.1 Pipeline Overview

This data pipeline is implemented using Python, leveraging various libraries such as Pandas, SQLite, and Requests. The pipeline consists of several steps, including data retrieval, cleaning, transformation, and storage.

3.2 Data Pipeline Diagram

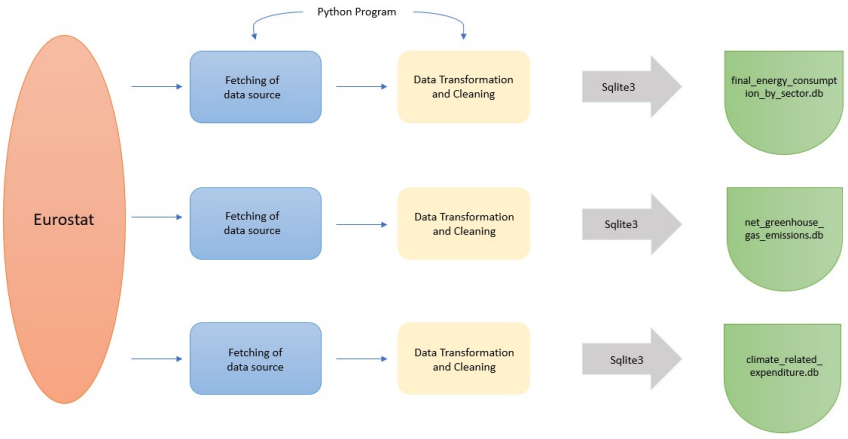

Figure 1: Data Pipeline Diagram

3.3 Transformation and Cleaning

Upon retrieving the data from Eurostat's API, we perform several transformation and cleaning steps to ensure its quality and usability for analysis. These steps include:

- **Downloading Datasets:** We utilize the Requests library to fetch the datasets from Eurostat's API using the provided URLs.
- **Cleaning Data:** We address missing values by filling them with zeros and standardize column names to ensure consistency across datasets. Cleaning the data ensures that our analysis is based on complete and consistent information.

- **Data Transformation:** Depending on the specific analysis requirements, we may perform additional data transformations such as aggregations, filtering, or merging of datasets to derive meaningful insights.

3.4 Challenges and Solutions

During the implementation of the data pipeline, we encountered several challenges, including network issues during dataset downloads and inconsistencies in data formatting. To address these challenges, we implemented the following solutions:

- **Error Handling:** We incorporated error handling mechanisms to handle network issues and retry dataset downloads in case of failures. This ensures the robustness of our pipeline and prevents data loss due to network disruptions.
- **Data Validation:** We implemented data validation checks to identify and correct any inconsistencies or errors in the data formatting. This ensures the accuracy and reliability of our analysis results.

3.5 Error Handling

This pipeline ensures reliability and integrity by implementing robust error handling mechanisms, including multiple retries during dataset downloads and data validation checks to identify and handle any errors or inconsistencies in the input data.

3.6 Licenses

The Eurostat data is usually licensed under open-data licenses, allowing free use with proper attribution. We plan to fulfill our obligations by appropriately attributing the data in our reports and analyses.

4 Result and Limitations

4.1 Output Data

The output of our data pipeline consists of cleaned datasets stored in CSV format and SQLite databases. These datasets contain the transformed and standardized data obtained from Eurostat's API, ready for further analysis and exploration.

4.2 Data Structure and Quality

The output data maintains the tabular structure of the original datasets, with rows representing different entities (e.g., countries, sectors) and columns representing various attributes (e.g., energy consumption, emissions, expenditure). The quality of the result is high, as we have addressed missing values and standardized column names during the cleaning process. However, the quality may vary depending on the original source data and the completeness of the information provided by Eurostat.

4.3 Output Format

We chose CSV and SQLite formats as the output of our pipeline due to their simplicity, ease of access, and compatibility with common data analysis tools and databases. CSV format allows for easy sharing and manipulation of data, while SQLite databases provide a structured storage solution for more complex datasets.

4.4 Critical Reflection

While our data pipeline effectively cleans and transforms the Eurostat API datasets, potential issues and limitations for our final report include data completeness, accuracy reliant on Eurostat's source data quality, limited scope of analysis to Eurostat's API data, and potential interpretation biases. Acknowledging these limitations enables proactive measures to ensure the validity and reliability of our final report.