MADE Project Work 4 – Data Report

Name: Souhardya Chattopadhyay | IDM ID: ow85ymut

Topic: Impact of Sectoral Energy Consumption on GHG Emissions and Lung Disease Mortality in Germany

1. Introduction

In this data engineering project, we analyse the relationship between energy consumption, greenhouse emissions, and lung diseases such as pneumonia in Germany from 2011-2021. The goal is to understand sector-wise contributions to greenhouse emissions and identify trends for future predictions. The following key questions interests us:

1. Analyse the correlation between greenhouse emissions and total energy consumption in Germany
2. Determine the impact of individual sectors on greenhouse emissions in Germany.
3. Identify trends in greenhouse emissions in relation to increases or decreases in deaths by pneumonia in Germ any.

2. Methods

2.1. Data Source Details:

Data source 1: European Data Portal - Final energy consumption by sector

Metadata URL: https://ec.europa.eu/eurostat/cache/metadata/en/nrg_bal_esms.htm
DataURL: https://ec.europa.eu/eurostat/api/dissemination/sdmx/2.1/data/ten00129?format=TSV&compressed=true
Data Type and Quality: The data type is TSV. Data is collected and updated by Eurostat which ensures high quality and reliability.
Data Descriptions : The "Final Energy Consumption by Sector" dataset details energy usage across various sectors in EU countries from 2011 to 2022, focusing on Germany's consumption patterns in commercial, household, transportation, and industrial activities.
Licence: Eurostat data is typically licensed under open-data licenses, permitting free use with proper attribution. We intend to comply with these obligations by appropriately attributing the data in our reports and analyses.

Data source 2: European Data Portal - Net greenhouse gas emissions

Metadata URL: https://ec.europa.eu/eurostat/cache/metadata/en/sdg_13_10_esmsip2.htm
DataURL: https://ec.europa.eu/eurostat/api/dissemination/sdmx/2.1/data/sdg_13_10?format=TSV&compressed=true
Data Type and Quality: The data type is TSV. Data is collected and updated by Eurostat which ensures high quality and reliability.
Data Descriptions: The "Net Greenhouse Gas Emissions" dataset covers emission patterns in EU countries from 1990 to 2022, with an in-depth analysis of Germany's data.
Licence: Eurostat data is typically licensed under open-data licenses, permitting free use with proper attribution. We intend to comply with these obligations by appropriately attributing the data in our reports and analyses.

Data source 3: European Data Portal - Death due to pneumonia, by sex

Metadata URL: https://ec.europa.eu/eurostat/cache/metadata/en/hlth_cdeath_sims.htm
DataURL: https://ec.europa.eu/eurostat/api/dissemination/sdmx/2.1/data/tps00128?format=TSV&compressed=true
Data Type and Quality: The data type is TSV. Data is collected and updated by Eurostat which ensures high quality and reliability.
Data Descriptions: The "Death Due to Pneumonia, by Sex" dataset records pneumonia-related deaths in EU countries from 2011 to 2021, highlighting how air pollution and climate change from greenhouse emissions increase respiratory infection risks, focusing on Germany.
Licence: Eurostat data is typically licensed under open-data licenses, permitting free use with proper attribution. We intend to comply with these obligations by appropriately attributing the data in our reports and analyses.

3 Data Pipeline

3.1 Overview

This data pipeline is constructed using Python and integrates various libraries such as Pandas, SQLite, and Requests. The pipeline encompasses several stages, including data retrieval, cleaning, transformation, and storage.

3.2 ETL Process

After acquiring the data from Eurostat's API, we undertake several transformation and cleaning procedures to ensure its quality and suitability for analysis. These procedures include:

Extract: Using the Requests library, we fetch datasets from Eurostat's API through the provided URLs.
Cleaning & Preprocessing: We handle missing values by replacing them with zeros and standardize column names to maintain consistency across datasets. This step is crucial for ensuring that our analysis is based on complete and uniform data.
Transform: Depending on the specific analysis requirements, we may perform additional transformations such as aggregations, filtering, or merging of datasets to derive meaningful insights.

3.3 Challenges and Solutions

Challenges: Several challenges were encountered during the implementation of the data pipeline, such as network issues during dataset downloads and inconsistencies in data formatting. The following solutions were implemented to address these challenges:
Error Handling: Error handling mechanisms were incorporated to manage network issues and retry dataset downloads in case of failures. This ensures the robustness of the pipeline and prevents data loss due to network disruptions.
Data Validation: Data validation checks were implemented to identify and correct inconsistencies or errors in data formatting, ensuring the accuracy and reliability of the analysis results.

3.4 Error Handling

The pipeline maintains reliability and integrity by implementing robust error handling mechanisms, including multiple retries during dataset downloads and data validation checks to detect and address any errors or inconsistencies in the input data.

3.5 Licenses

Eurostat data is typically licensed under open-data licenses, permitting free use with proper attribution. We intend to comply with these obligations by appropriately attributing the data in our reports and analyses.

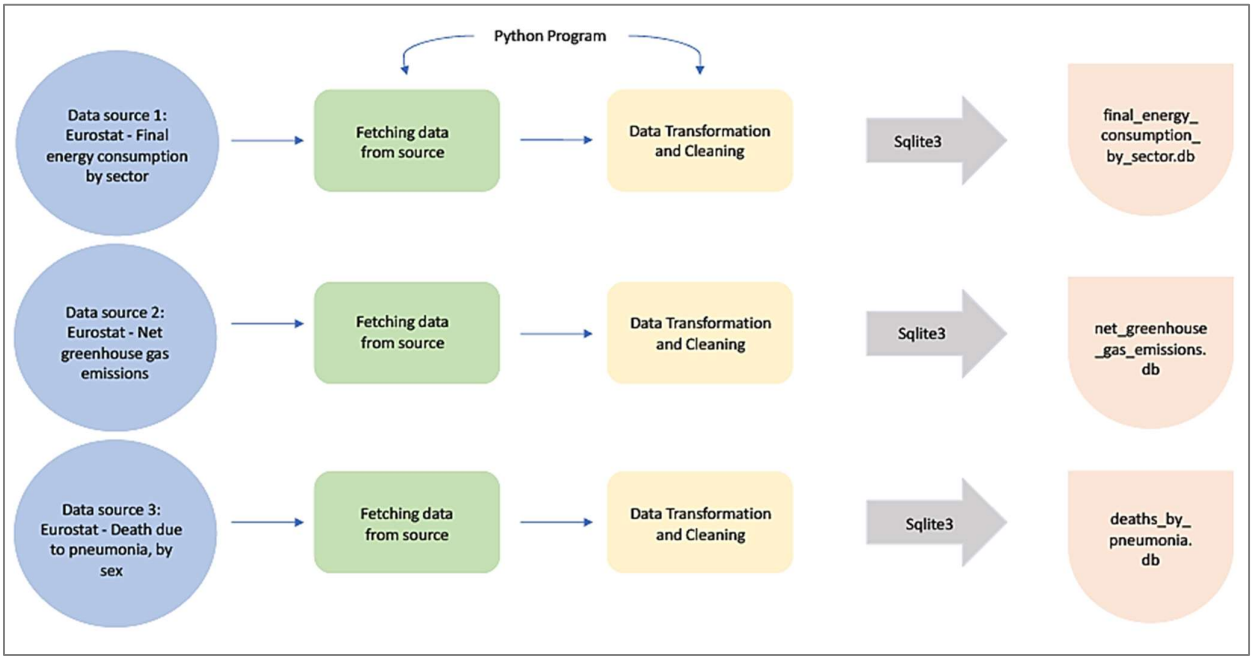

Fig 1: Data Pipeline

4 Results and Limitations

4.1 Processed Data

The output produced by our data pipeline includes cleaned datasets stored in CSV format and SQLite databases. These datasets contain the transformed and standardized data retrieved from Eurostat's API, making them ready for further analysis and exploration.

4.2 Data Format and Integrity

The resulting data maintains the tabular format of the original datasets, with rows representing different entities (such as countries and sectors) and columns representing various attributes (like energy consumption, emissions, and expenditure). The quality of the output is high, as missing values have been addressed and column names standardized during the cleaning process. However, the quality is contingent on the original source data and the completeness of the information provided by Eurostat.

4.3 Output Format

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

4.4 Evaluation and Constraints

While our pipeline effectively transforms and cleans datasets from Eurostat's API, there are potential issues and limitations to consider for our final report. These include data completeness, the accuracy and reliability of Eurostat's source data, the limited scope of analysis to Eurostat's API data, and potential biases in interpretation. Recognizing these limitations helps in taking proactive measures to ensure the validity and reliability of our final report.