Project Work 4 - Data Report

Neelima Kawatra

June 5, 2024

1 Question Addressed using this project

Are there any trends in net greenhouse gas emissions with the use of energy from renewable sources?

2 Data Sources

The data for this project was sourced from the Eurostat database, a comprehensive source of statistical information provided by the European Union. Two datasets were selected:

1. **Net Greenhouse Gas Emissions**: This dataset provides information on net greenhouse gas emissions of various countries, which is crucial for understanding the environmental impact of various sectors, from years 1990-2022.
2. **Share of Energy from Renewable Sources**: This dataset contains information on the percentage share of energy generated from renewable sources, from years 2013-2022.

Both datasets are publicly available from the Eurostat API in TSV format and were chosen due to their relevance to the project question. The data structure of both datasets is tabular, with rows representing observations for different time periods and columns representing variables such as country, year, and emission values. The quality of the data is generally reliable, although missing values were present, which were handled during the data cleaning process. The datasets are licensed under the European Union Public License (EUPL), which allows for the reuse of the data under certain conditions. We are allowed to use the data for analysis and reporting purposes, provided we attribute Eurostat as the data source. I plan to fulfill our obligations by including appropriate citations and acknowledgments in our final report.

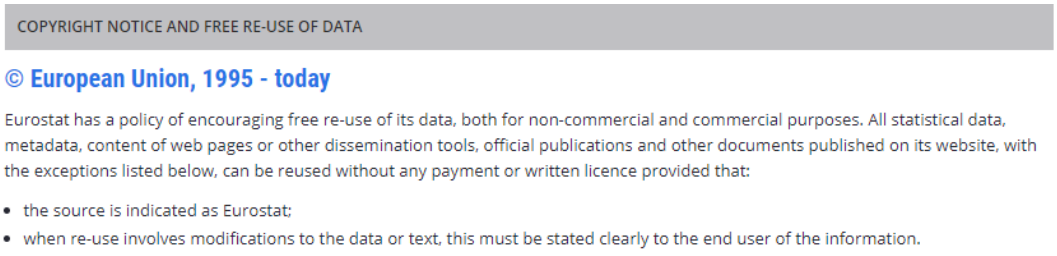

Figure 1: License Information

3 Data Pipeline

The data pipeline implemented for this project is an automated process that involves several steps:

1. **Data Download**: The datasets were downloaded from the Eurostat API using Python's `requests` library.
2. **Data Cleaning**: Missing values in both datasets were filled with zeros, and column names were standardized by converting them to lowercase and removing leading/trailing whitespace.

3. **Data Storage**: Cleaned datasets were saved as CSV files and SQLite databases for further analysis and reporting.

The main technologies used for implementing the data pipeline include Python (for data download, cleaning, and analysis), pandas (for data manipulation), SQLite (for data storage), and requests (for HTTP requests). During the implementation of the data pipeline, we encountered some challenges, such as handling missing values and ensuring data consistency across different sources. These issues were addressed by carefully reviewing the data and applying appropriate cleaning techniques. The pipeline is designed to handle errors gracefully by using exception handling mechanisms and logging to track any issues that may arise during data processing.

4 Result and Limitations

- The data pipeline creates a file in SQLite3 format, which contains all the necessary data for our project. We make sure the data is good by cleaning and organizing it properly.
- We choose SQLite3 because it's easy to work with and can be used in many different programs. This is helpful because our project is based on Python, and SQLite3 works well with it.
- One problem we might face is that the data we get comes from different places, and they might not fit together perfectly. This means we might have trouble making all the data work together smoothly. We'll need to spend time making sure everything matches up correctly.

5 Data Pipeline Overview

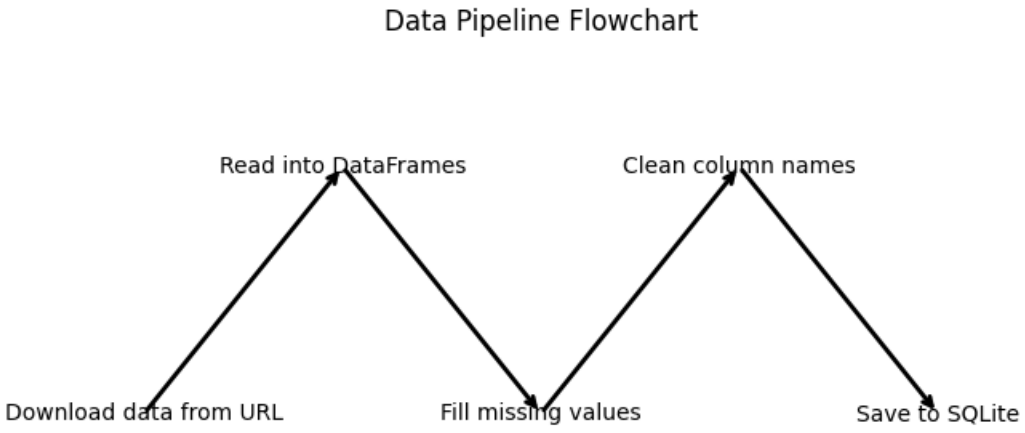

Figure 2: Data Pipeline