Data Report

June 3, 2024

1 Data Science Project : Data Report

1.1 *How has the adoption of renewable energy sources impacted greenhouse gas emissions across Europe over the last two decades?*

1.1.1 About Datasources:

- **Datasource 1: European Environment Agency (EEA)** This data source is provided by the European Environment Agency, which oversees environmental processes across Europe. This dataset provides comprehensive data on greenhouse gas emissions across European countries, broken down by sector and year.Datasource
- **Datasource 2: Eurostat** This data source is provided by Eurostat, which provides high-quality statistics and data in Europe. This dataset provides comprehensive data on the share of renewable energy sources in total energy production in each country across Europe.Datasource
- **Datasource 3: Eurostat** This data source is provided by Eurostat, which provides high-quality statistics and data in Europe. This dataset provides information regarding the total energy balance per country in Europe based on energy from different sources.Datasource

1.1.2 Detailed Information about data source licensing, structure, and quality:

- **Data Quality**: The data sources used in this project satisfy all the dimensions of data quality:
  >Accuracy, Completeness, Consistency, Timeliness, Relevancy:
- **Data Structure**: The data sources are available in a structured data format (CSV format) with a given schema, and they are available in batch format.
- **Licensing**: All the data sources used in this project are available under the Standard Open Data License (Creative Commons) CC which encourages the use of the data for both commercial and non-commercial purposes. However, we must indicate the source from which the data is gathered.

1.1.3 Data Pipeline :

The Data Pipeline is created in a high-level programming language like Python. The pipeline utilizes various libraries available in Python to create the data pipeline. Some of the libraries used are Pandas (for storing and manipulating data sources), Requests (to fetch the data from online sources), Matplotlib (for visualization of data sources), and SQLite3 (for storing the result of the data pipeline).

- **Data Cleaning & Transformation:** Basic transformation and cleaning are performed, such as replacing NaN values with 0, dropping duplicate columns, and renaming columns for better representation.
- **Problems:** The main problem faced during data pipelining was filtering out ambiguous data in the context of the problem statement and determining which columns are necessary for the task.
- **Solutions:** To overcome the problem of ambiguous data, I performed data cleaning and data transformation. To filter out irrelevant columns, I incorporated metadata information into the setting, which helped me in filtering out the columns.

1.1.4 Result and Limitations:

- The output of this data pipeline is an SQLite3 database file that contains all the required data and necessary information for this project. The quality of the data is improved through various transformation and cleaning processes.
- The data pipeline produces a SQLite-based database file. The reason for using SQLite3 is its flexibility and ease of use across various applications. SQLite3 allows for data access and management across different platforms, and its libraries are well-supported in Python. This compatibility is particularly beneficial for this project, which is based on Python.
- One potential issue that may arise in the future is the compatibility of data sources with each other, given that the data comes from different providers. Establishing relationships between these data sources can be complicated, as the metrics used in each source may differ.

1.1.5 Overview of Data Pipeline: