data-report

June 2, 2024

1 Data Report: Analyzing the Relationship between Renewable Energy Adoption, CO2 Emissions, and Economic Growth in Germany, Sweden, and Bulgaria (2000–2021)

1.1 Main Question

How have GDP and the adoption of renewable energy sources (total energy generation, renewable energy generation, share of generation) influenced energy consumption and CO2 emissions in Germany, Norway, and Bulgaria from 2000 to 2021?

1.2 Data Sources

To answer the question, two data sources have been selected for this project: EMBER Climate, which shows yearly information on european electricity, and The World Bank, which shows the Gross Domestic Product (GDP US$) from all countries.

1.2.1 Data source 1: European Electricity Dataset

- **Metadata URL**: European Electricity 2022
- **Data URL**: European Electricity Raw Data
- **Data Type**: CSV Directory
- **Description**: The European Electricity Dataset from Ember provides a collection of datasets related to electricity generation, CO2 emissions, net import and demand across various countries in Europe. For this project, we will remove the net import dataset and focus on the other variables (generation, CO2 emissions and demand) to analyze the adoption of renewable energy sources for Germany, Sweden, and Bulgaria from 2000 to 2021.
- **Data Structure & Quality**: The dataset is structured as a CSV directory, with seperate files containing data for different variables in tabular format (CSV files). It reflects data from various reliable sources (e.g. EIA, Eurostat, ENTSO-E, SolarPower Europe), ensuring its accuracy to real-world condition. The dataset contains all necessary information required about electricity generation, emissions, and demand across multiple countries in Europe. It has consistency in its format across different variables, countries, and years. Furthermore, The data aligns with the needs of the project. No missing, duplicate or invalid information is found in the dataset either for the chosen countries and time periods. Additional filtering is applied in the preprocessing step to refine the dataset.

- **License and Obligations**: The dataset is freely available for both non-commercial and commercial use from ember-climate.org. It is licensed under the Creative Commons Attribution 4.0 International License (CC-BY-4.0), allowing for sharing and adaptation for various purposes. Appropriate credit to EMBER and link to the License will be provided in this project to comply with the obligations.

1.2.2 Data source 2: World GDP Dataset

- **Metadata URL**: World Bank GDP Metadata
- **Data URL**: World Bank GDP Data
- **Data Type**: CSV
- **Description**: This dataset provides annual GDP figures in US dollars for countries worldwide from 1960 to 2021. For this project, we will focus the GDP data for Germany, Sweden, and Bulgaria from 2000 to 2021.
- **Data Structure & Quality**: The dataset is structured in tabular format (CSV files), with rows representing countries and columns representing years. Similar to the European Electricity dataset, The GDP dataset is sourced from a reputable institution (World Bank and OECD) and considered reliable, accurately represents GDP figures in US dollars for various countries in Europe. It contains all necessary variables/information and ensures the consistency in its format across countries and years for this project. The presentation of data fits the project's objective, focusing on economic indicators (GDP)
- **License and Obligations**: The dataset is licensed under the Creative Commons Attribution 4.0 International License (CC-BY-4.0), similar to the Data Source 1. Appropriate credit to World Bank and link to the License will be provided in this project to comply with the obligations.

1.3 Data Pipeline

The ETL (Extract [extraction.py], Transform [transformation.py], Load [saving.py]) pipeline is implemented using python to handle both data sources, each downloaded as a CSV directory within a zip archieve. This process involves extracting the right CSV file, transforming, and saving it as CSV format.

European Electricity Dataset ETL:

- **Extraction:** The dataset is fetched as a zip archieve, that contains multiple CSV files. A function *('extract_zip_data')* that sends an HTTP GET request to European Electricity Raw Data and read the content of the zip file. The desired CSV file is extracted based on the its exact name. No regular expressions are required as the CSV file name does not change.
- **Transformation:** then the fetched CSV file is read into a pandas Dataframe. A function *('transform_eer_data')* filters the data for the countries (Germany, Sweden, Bulgaria) and the years 2000-2021. Depending on type of data as function parameter, unnecessary columns are dropped. The columns are then renamed to more readable names using a dictionary.
- **Loading:** The function *('save_dataframe_to_csv')* save the transformed dataframe to a CSV file in */data* directory.

World GDP Dataset ETL:

- **Extraction:** Similar to the first dataset, The World GDP dataset is fetched from World Bank GDP Data as a zip archieve, that contains multiple CSV files. the same function is used here. But for this dataset, a regular expression is used to identify and match the CSV file due to randomly generated number in the file name each time it is downloaded. The skiprows parameter is also applied to ensure that the data begins after the first 4 rows.
- **Transformation:** Here, the countries and years are filtered just like the first dataset, but the DataFrame is then reshaped from wide format to long format using the *('pd.melt')* function from pandas. This creates a new column 'Year' and 'GDP (US$)' to ensure consistency in both datasets.
- **Loading:** The function *('save_dataframe_to_csv')* save the transformed dataframe to a CSV file in */data* directory.

Problems Encountered, Solutions, and Error Handling:

- **Dynamic file names**: The World GDP dataset's CSV file name includes a randomly generated number, making it difficult to directly identify the file. Thus Regular expressions (regex) were used to match the pattern of the file name.
- **Incorrect or inconsistent data**: during the transformation process, data cleaning steps such as filtering for relevant countries and years, dropping unnecessary columns, and renaming columns were implemented to ensure the data consistency. No missing values are found after data cleaning. Additional data validation is also applied to ensure the correct data type before analysis.
- **Error Handling**: Exception handling is implemented throughout the function to catch and handle errors, providing debugging during HTTP requests, file extraction, data reading, and transformation.

1.4 Result and Limitations

Data Ouput:

The output of the data pipeline consists of transformed datasets in CSV format. For the European Electricity dataset, the output includes separate CSV files for Electricity Demand *(Columns: [Electricity Demand (TWh),Electricity Demand per Capita (MWh)])*, Emissions *(Columns: [Emissions Intensity (gCO2/kWh),Emissions (MTcO2e)])*, and Generation *(Columns: [Fuel Code,Fuel Description,Electricity Generation (TWh),Share of Generation (%)])* data for the selected countries (Germany, Sweden, and Bulgaria) from the years 2000 to 2021. Similarly, for the World GDP dataset, the output is a CSV file containing GDP data for the same countries and time period.

Data Structure, Quality and Format:

The output datasets maintain a tabular data schema with data types for each attribute. The quality of the output is ensured for further analysis through various data cleaning and transformation steps, including filtering for relevant countries and years, dropping unnecessary columns, and renaming columns for clarity. considering where the data sources come from, the realibility also can be ensured.

Moreover, CSV (Comma-Separated Values) is chosen as final format because it is widely used, easily accessible, and compatible with various data analysis tools like pandas in Python.

Reflection and Potential Issues:

While the data pipeline successfully processed the data and no missing data was found for the given countries and years, potential outliers may still appear during the analysis. Additionally, uncertainties regarding the optimal representation of GDP ($US) values is an issue, since large numerical values may hinder readability and comprehension during analysis. Thus, further refinement of the GDP data representation, such as converting values to billion or million units, still need to be determined.

The focus on only three specific countries and the limited timeframe from 2000 to 2021 may introduce potential biases into the analysis. These biases may restrict the generalizability of conclusions. Careful interpretation is important, as the transformed dataset may not fully represent the diversity across Europe. Lastly, a thoughtful consideration and exploration on the choice of variables (e.g. Electricity Demand or Electricity Demand per Capita) is still required to ensure the robustness of analysis outcomes.