Correlation between Temperature Increase and Energy Consumption in European Countries

Dennis Mustafić

Matriculation Number: 233302080

Data Report

Methods of Advanced Data Engineering

Research Questions:

1. Is there been a temperature increase in European countries since the 2000s?
2. How has energy consumption changed since the beginning of the 2000s?
3. **Main question:** does a potential temperature increase correlate with energy consumption of Europe?

1 Data Sources

1.1 Climate Change Indicators

1.1.1 Data origin and content

The data used in this project, in .csv format, is obtained from Kaggle[1], provided by user Tarun Mugesh, who scraped it using Python from the sources `worldbank.org` and `climatedata.imf.org` [1]. The World Bank Group is described as "one of the world's largest sources of funding and knowledge for developing countries" [2], offering extensive research and data on global development, which is reflected in the dataset. Additionally, `climatedata.imf.org` focuses on climate-related data and analysis, particularly on the economic implications of climate change [3].

Climate Change Indicators comprises 72 columns and 225 rows, including 10 descriptor columns such as country names, ISO2 codes (standardized country abbreviations), and units used. The remaining columns provide annual surface temperature changes for 225 countries, measured in degrees Celsius, with each column representing one year from 1961 to 2022.

1.1.2 Data quality

Accuracy Regarding the accuracy of the data, no explicit statement can be made; it relies on the trustworthiness of the platforms mentioned above.

Completeness Two European countries—Turkey and Kosovo—are not included.

Consistency Approximately 10.6% of observations are missing from the entire dataset. The European subset is missing 21.95% of its data. It's also worth noting that the dataset does not include data for the most recent year, 2023.

1.1.3 Factors behind dataset selection

This dataset was chosen for several reasons: Firstly, it holds the highest possible usability rating on Kaggle, highlighting its ease of use. Moreover, it offers extensive coverage, encompassing a broad range of countries (almost worldwide) and years of measurements. Additionally, Kaggle's own pip package and intuitive API simplify the process of accessing and downloading the data from their platform.

1.1.4 License

The type of license associated with this dataset is Creative Commons Zero (CC0). This license enables various groups of individuals, including scientists, educators, artists, and other creators or owners of copyright- or database-protected content, to relinquish their rights [4]. According to the Creative Commons website, CC0 places works as fully as possible in the public domain, allowing others to freely use, enhance, and reuse the content for any purpose without copyright or database law restrictions [4]. This ensures that the dataset can be utilized without limitations for this project.

1.2 Primary Energy Consumption EU

1.2.1 Data origin and content

The dataset[2] is distributed by Eurostat, the statistical office of the European Union. Eurostat provides independent and credible European statistics essential for decision-making, research, and public debate, supporting the work of the Commission and various sectors [5]. The data is in SDMX-CSV format. SDMX (**S**tatistical **D**ata and **M**etadata e**X**change) is a global initiative to standardize and modernize

[1] `https://www.kaggle.com/datasets/tarunrm09/climate-change-indicators`
[2] `https://ec.europa.eu/eurostat/databrowser/view/sdg_07_10/default/table`

statistical data and metadata exchange between international organizations and member countries. This format ensures consistent and efficient data sharing, recognized and utilized by major statistical agencies and organizations worldwide [8].

The dataset includes annual energy consumption measured in three distinct ways:

1. **Tonnes of Oil Equivalent per Thousand Euro of GDP**: This metric measures energy intensity, indicating how much energy is consumed to produce one thousand euros of GDP.
2. **Index, 2005=100**: This index tracks changes in energy consumption over time, with the base year set to 2005 (value = 100). For example, if the index value for 2006 is 80, it signifies a 20 percent reduction in energy consumption compared to 2005.
3. **Tonnes of Oil Equivalent (TOE) per Capita**: This metric measures the average energy consumption per person in tonnes of oil equivalent, providing insight into individual energy use.

The dataset covers the years from 2000 to 2022 and includes 38 countries. Notably, it includes countries like Bosnia and Herzegovina, which has been granted candidate status by the European Council for EU membership [6].

1.2.2 Data quality

Accuracy Although no explicit statement about the accuracy of the data can be made, it is important to note that the data originates directly from the statistical office of the European Union. This source is known for its high credibility, which suggests that the data is reliable.

Completeness The data exhibits a high degree of completeness, with only a few missing data points. Specifically, for Tonnes of oil equivalent per thousand euro GDP, approximately 0.72% of data points are missing. Similarly, for TOE per Capita, around 0.80% of data points are absent.

Consistency One issue with Eurostat data is the inconsistency in the number of values for each year depending on the variable. For instance, in the case of TOE per Capita, there are varying numbers of observations for different years, potentially leading to challenges in data analysis and interpretation. For instance, in the TOE per Capita dataset, there are 38 values for the year 2018, 37 values for 2019, 35 values for 2021, and 28 values for 2022, illustrating the inconsistency in the number of observations across different years. Also, the dataset does not cover the most recent year, 2023.

1.2.3 Factors behind dataset selection

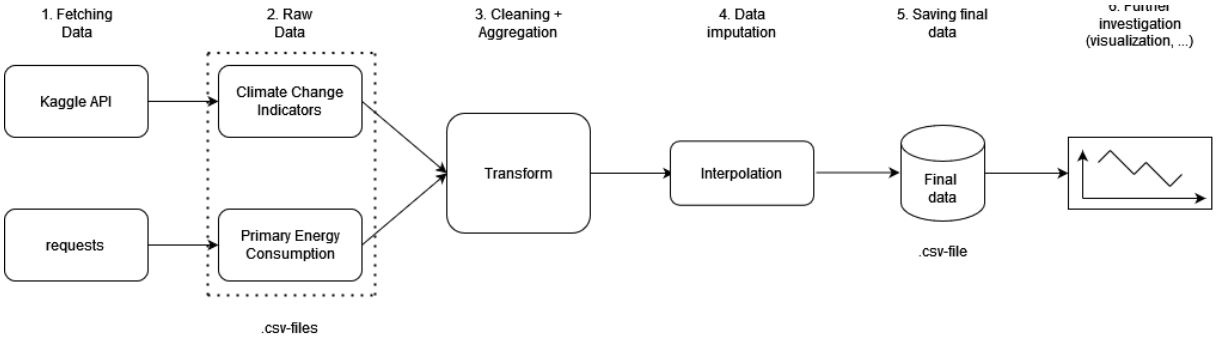
The data was chosen since it has high credibility because it comes directly from Eurostat. Additionally, as previously mentioned, it covers a broad range of countries, providing extensive insights into energy consumption patterns across Europe.

1.2.4 License

The dataset is distributed under an Open Data License. This permits unrestricted use, as outlined on the Eurostat website. According to the EU's definition of open data, this license ensures the legal accessibility of the data, allowing everyone to access, reuse, and distribute it freely [7].

2 Pipeline

Fetching Data The first step in the pipeline involves fetching the necessary data from different sources. For the climate change indicators, the Kaggle API is utilized. Kaggle provides a convenient pip package that allows data to be downloaded with a Python script in just a few steps. An API key is required



to access the Kaggle API. For the primary energy consumption data, the Python requests package is used. A `GET` request is sent to the provided URL, downloading the data and saving it with a specified filename. Both methods are implemented in the `DataRetriever` class, saved in `project/downloader.py`. These methods take a `dataset_name` string as an argument, and the `download_eurostat_data()` method additionally requires the URL. The fetched data is then saved in the `data/` folder located in the root directory.

Data Transformation Once the data is fetched, it must be transformed to ensure it is properly structured for analysis. This is done using the `DataPreprocessor` class in `project/preprocessing.py`, which takes file paths for both datasets as arguments. Two methods are implemented that return pandas DataFrames, one for Kaggle preprocessing and the other for Eurostat preprocessing. The method `_preprocess_kaggle()` drops obsolete columns, transposes the data so countries are columns and years are rows, and then uses the `.melt()` function to reshape the DataFrame into a row-format. `_preprocess_eurostat()` uses `.pivot_table()` to organize the DataFrame and fill missing years with NaN values, enabling interpolation in the subsequent step. The data is then transformed using `.melt()` for later concatenation. Lastly, both datasets get aggregated. A comparison determines which countries are missing in each dataset, and only the intersection of both datasets is used. This dynamic decision-making allows the data to adapt if new measurements are added.

Data Imputation The `clean_and_interpolate_data()` method in the `DataPreprocessor` class performs data imputation. It takes a DataFrame and a column name, returning the processed DataFrame. If more than 10 values are missing for a specific column in a country, that country is removed from the dataset. If 10 or fewer values are missing, the values are interpolated. The threshold of 10 balances the need to maintain countries with accurate data approximation. If values are missing at the beginning of the DataFrame, those countries are also excluded from the final dataset. As mentioned in 1.1.2 and 1.2.2, several observations are missing, especially for the Kaggle dataset. After performing data imputation, there are no missing data points for the entire dataset, thus enhancing the data quality significantly.

3 Results and Limitations

The resulting dataset layout is as follows:

TIME_PERIOD	ISO2	COUNTRY	CHANGE_INDICATOR	MTOE	TOE_HAB
2000	AL	Albania	1.065	1.8	0.58
2001	AL	Albania	1.400	0.8	0.99
...	...	...	...	...	...

The new dataset contains three columns: surface temperature changes in degrees (CHANGE_INDICATOR), tonnes of oil equivalent per thousand euros of GDP (MTOE), and tonnes of oil equivalent

(TOE_HAB) per capita. Notably, Index, 2005=100 (referenced in 1.2.1) is not utilized in the final dataset. The enhanced dataset covers 32 countries and comprises a total of 736 rows of measurements. This comprehensive dataset serves as a robust foundation for analyzing climate indicator and energy consumption across Europe. One of the main limitations is that Kosovo and Turkey were missing from the beginning in the Climate Change Indicator Data. Additionally, during the data imputation process, three countries were removed: Bosnia and Herzegovina due to having more than 10 missing values, and Serbia and Montenegro due to missing values at the beginning, making interpolation impossible.

References

[1] Kaggle. (2024) Climate change Indicators. Retrieved from https://www.kaggle.com/datasets/tarunrm09/climate-change-indicators

[2] World Bank. (2024). Retrieved from https://www.worldbank.org/en/who-we-are; Last accessed on May 26, 2024.

[3] Climate Data. (2024). Retrieved from https://climatedata.imf.org/; Last accessed on May 26, 2024.

[4] Creative Commons. (2024). Retrieved from https://creativecommons.org/public-domain/cc0/; Last accessed on May 26, 2024.

[5] Eurostat. (2024). Working arrangements November 2022. Retrieved from https://ec.europa.eu/eurostat/documents/10186/15478239/Working+arrangements+Nov+2022.pdf/8e170bc0-6729-22b2-1e6a-fd39acc86922?t=1669391049507; Last accessed on May 26, 2024.

[6] Euronews. (2024). Brussels recommends opening EU membership talks with Bosnia and Herzegovina. Retrieved from https://www.euronews.com/my-europe/2024/03/12/brussels-recommends-opening-eu-membership-talks-with-bosnia-and-herzegovina; Last accessed on May 26, 2024.

[7] Data Europa. (2024). Open data licensing. Retrieved from https://data.europa.eu/en/academy/open-data-licensing; Last accessed on May 26, 2024.

[8] SDMX. (2024). Retrieved from https://en.wikipedia.org/wiki/SDMX; Last accessed on May 26, 2024.