Carbon Dioxide Emissions and Mean Sea Level Pressure Analysis

Climate change is a real phenomenon whose impacts are felt both in the natural environment and in people's lives. It is crucial to analyze trends in carbon dioxide emissions and mean sea level pressure to adequately develop risk management strategies. This project deals with the assessment of historical data on Carbon Dioxide emissions in Europe up to the year 2022 and Global mean Sea Level Pressure as different variables to predict the trends and probable effects.

Main Questions of this project:

1. What are the historical trends in carbon dioxide emissions in Europe from 1850 to 2022?
2. How has the mean sea level pressure changed globally over time?

Data Sources

Datasource 1: Data on CO2 and Greenhouse Gas Emissions by Our World in Data

Metadata URL: https://github.com/owid/co2-data/blob/master/owid-co2-codebook.csv
Data URL: https://nyc3.digitaloceanspaces.com/owid-public/data/co2/owid-co2-data.csv
File Format: CSV
License Type: CC BY 4.0

Description:

The data includes extensive reports of CO2 emissions for a number of countries over a period of time ranging from 1850 to 2022.The reason for selecting this dataset which is that, it provides extensive year-by-year data required for understanding historical trends of carbon emissions. Drawn from **Our World in Data** which is a highly regarded platform, it contains data on CO2 emissions from fossil fuels and cement production, and gas flaring which makes it a rich database for environmental and climate analysis. Nonetheless, some gaps may exist in the collection of data due to either errors in the process or because certain historical archives were not well documented. The data is used under the **license Creative Commons Attribution 4.0 International (CC BY 4.0)**, which provides permission to use it wisely while citing the source, providing the metadata link and the source data link, and mentioning modifications if any.

Datasource 2: Worldwide Sea Level Pressure

Metadata URL:
https://opendata.dwd.de/climate_environment/CDC/observations_global/CLIMAT/monthly/qc/mean_sea_level_pressure/historical/
Data URL:
https://opendata.dwd.de/climate_environment/CDC/observations_global/CLIMAT/monthly/qc/mean_sea_level_pressure/historical/01001_195101_202112.txt

File Format: TXT
License Type: CC BY 4.0
License Permission: The right to distribute, use, and modify the licensed work is granted by this license.
Citation: The Deutscher Wetterdienst. It is a public institution.

More information about the License can be found here:
https://opendata.dwd.de/climate_environment/CDC/Terms of_use.pdf

Description:

The dataset is a time series of monthly mean sea level measurements spanning from 1951 to 2015. Each row represents a year, with columns for the sea level measurements for each month (January through December). The data is in a text format, which includes the year and the corresponding sea level values for each month. This dataset was chosen for its relevance in studying climate change, sea level rise, and their potential impacts on coastal regions. The dataset is also under the Creative Commons Attribution 4.0 International (CC BY 4.0) license like **Data Source 1**.

Data Pipeline:

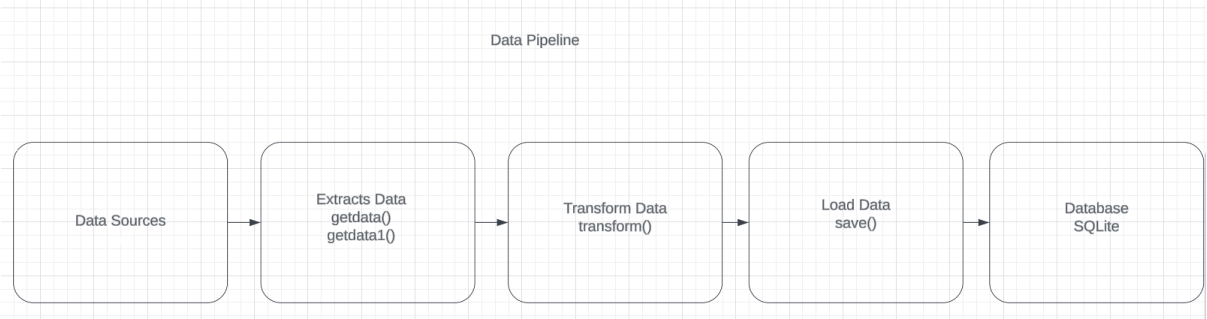

Figure: Automated Data Pipeline

These data pipeline pathways associated with this project are aimed at making efficient extraction, transformation, and loading of data on CO2 and Sea level. The pipeline starts with the Request library, aimed to retrieve the datasets from given URLs**. get_data()** method, Downloads a CSV file that includes the CO2 data from a repository in the GitHub platform. If successful, it stores the data on a Pandas Data Frame. And **get_data1()** Creates a Pandas Data Frame to store sea level mean data from a URL provided, replaces missing values with the last known, and writes it to the SQLite database. In the transformation phase of the pipeline, the **transform()** method is applied to the CO2 dataset to obtain only the record from Europe, dropping the rows with missing data in CO2 and finally, the user selects the year, population, CO2 levels, change in temperature from CO2, and the growth rate of percentage of CO2 from these exploration steps. An attempt is made to deal with missing values: the backward fill technique is used on the population and temperature change parameters. The pre-processed data is finally imported into the SQLite database using **save()** to structure the data in an optimum manner for the next analysis stage. Some of the problems that the pipeline tackles in this process include the possibility of missing values where the pipeline uses the backward fill mechanism to handle it, and places where downloads go wrong

wherein the pipeline includes methods that check HTTP status codes and print appropriate messages. The ability of the loading phase to load data and use the **replace** option, in particular, means that the pipeline is ready to deal with such modifications of input data – overwrite tables without problem. In this way, this firm-united approach guarantees data cogency and sustainability which makes the pipeline the important structural part of the project's data management system.

Result and Limitations:

The data pipeline takes datasets on CO2 emissions and sea level to generate information on how the two factors are related. The output data is also presented in a tabular form as year, country, CO2 emissions, and sea level mean. This format makes analysis and visualization easy and simple. The data is relatively clean, with proper sources used to gather information, and outliers are dealt with during data cleaning. The completeness is still quite good, but there are some cases when the missing values are filled by imputation methods.

CSV format was chosen as output data storage as it is widely used, easy to read, and compact to store. However, there are certain concerns that may be considered, which are discussed in the following sections. Approximations resulting from missing values may cause biases that would influence accuracy, even when imputed. Logistical constraints or time restrictions might also limit the study area to collect data for the analysis. That is why the process of data processing may lead to certain biases, and the results may not be fully representative if some areas of the world are underrepresented. Lastly, it was possible that the outcomes could be somehow affected by inconsistencies in data reported in the original data sources. They ought to be critically discussed in the final report to facilitate transparency of the results.