Methods of Advanced Data Engineering Project Report

Kimia Sedighi

ID: 23455613

Introduction

In this project, we investigated the impact of climate change on global carbon dioxide levels and surface temperature changes, aiming to explain their interconnected dynamics. Using datasets on World Monthly Atmospheric Carbon Dioxide Concentrations and Annual Surface Temperature Change, sourced from reputable organizations, the study explores temporal trends and potential correlations between these critical climate indicators. By employing advanced statistical techniques, the analysis seeks to uncover patterns, anomalies, and potential causal links influenced by climate change. Insights gleaned from this investigation could contribute to a deeper understanding of climate change's effects on atmospheric carbon dioxide concentrations and surface temperature changes, informing policy-making efforts aimed at mitigating its adverse impacts.

Main Question: How has climate change influenced global carbon dioxide levels in the atmosphere and surface temperature changes over time?

Data Sources

Dataset 1: World Monthly Atmospheric Carbon Dioxide Concentrations

- **Data:** Link
- **Data URL:** World Monthly Atmospheric Carbon Dioxide Concentrations
- **Data Type:** GeoJSON
- **Description:** This dataset presents the concentration of carbon dioxide in the atmosphere on a monthly and yearly basis, dating back to 1958.

Dataset 2: Annual Mean Global Surface Temperature

- **Data:** Link
- **Data URL:** Annual Mean Global Surface Temperature
- **Data Type:** GeoJSON

- **Description:** This dataset presents the mean surface temperature change during the period 1961-2023, using temperatures between 1951 and 1980 as a baseline.

Source and License: Both datasets are provided by the International Monetary Fund (IMF) and are licensed for personal, noncommercial use, including educational and research purposes. Detailed license information can be found here.

Justification for Data Selection

These datasets were chosen due to their high quality and extensive coverage. They provide comprehensive data on atmospheric carbon dioxide and temperature changes, which are crucial indicators of climate change. Analyzing these datasets allows us to understand the long-term impacts of climate change on the environment.

Data Structure and Quality

World Monthly Atmospheric Carbon Dioxide Concentrations

- **Structure:** The dataset is in GeoJSON format, containing fields for date, CO2 concentration levels, and geographic information.

- **Quality:** The dataset is highly reliable, as it is maintained by the IMF and includes continuous monthly data from 1958 to the present.

Annual Mean Global Surface Temperature

- **Structure:** The dataset is in GeoJSON format, containing fields for date, mean temperature change, and baseline information.

- **Quality:** The dataset is highly reliable, as it is maintained by the IMF and includes annual data from 1961 to 2023.

Data Pipeline

The data pipeline was implemented using Python and several key libraries, including requests, pandas, geojson, csv, and os. The pipeline involves the following main steps:

Data Ingestion

Data is fetched from the URLs provided by the IMF using the requests library.

Data Cleaning

- **Dropping Irrelevant Columns:** Columns such as 'ISO2', 'ISO3', and others not relevant to the main research question were removed to simplify the dataset.

- **Ensuring Numeric Values:** CO2 values and temperature data were checked to ensure they are numeric, and rows with invalid data were dropped.
- **Handling Missing Data:** Rows with missing values in critical columns ('Date' or 'Value' for CO2 data, and temperature columns for the temperature dataset) were dropped to ensure data integrity.

Error Handling

The pipeline includes error handling for data fetching, ensuring that any issues are logged, and the pipeline can continue running without crashing.

Output Data

The data pipeline outputs cleaned and transformed CSV files for both datasets. CSV format was chosen for its simplicity, compatibility with various data analysis tools, and wide support, making it ideal for storing the cleaned data.

Data Structure and Quality

- **Structure:** The cleaned data is structured with relevant columns retained, ensuring a tidy dataset suitable for analysis.
- **Quality:** High quality is maintained by ensuring numeric values are valid and dropping rows with missing or invalid data.

Critical Reflection

While the data pipeline effectively cleans and transforms the data, some potential issues remain:

- **Data Completeness:** Dropping rows with missing values might lead to a loss of potentially valuable data.
- **Changing Data Structures:** Future changes in the data structure of the source datasets might require updates to the pipeline.

Conclusion

The automated data pipeline successfully processes high-quality climate data, preparing it for detailed analysis and visualization. The insights gained from this analysis can inform policy-making efforts to mitigate the adverse impacts of climate change.