Temperature Anomaly and Natural Disasters Analysis in Germany

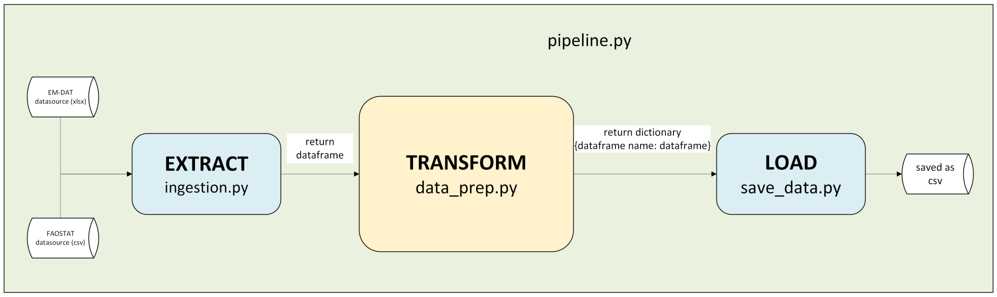
May 29, 2024

This report aims to analyze temperature anomaly trends in Germany from 1961 to 2023 and investigate natural disasters in Germany from 2001 to 2023.

0.1 Questions this project aims to address

1. What are the temperature anomaly trends over the years 1961 to 2023 in Germany?
2. Are the temperature anomaly trends similar to previous reports seen across the world?
3. What natural disasters have struck Germany from 2001 to 2023 and what is the impact?

0.2 Data Sources



0.2.1 1. EM-DAT - The International Disaster Database

- **Metadata URL**: https://public.emdat.be/data
- **Data URL**: https://public.emdat.be/data
- **Data Type**: Microsoft Excel (Structured and Tabular Dataset)

EM-DAT contains data on the occurrence and impacts of over 26,000 mass disasters worldwide from 1900 to the present day. The Centre for Research on the Epidemiology of Disasters (CRED) distributes the data in open access for non-commercial use. The terms and conditions of the dataset allow utilization of the datasets for academic, non-commercial purposes, as long as it is not reproduced, distributed, or a derivative of the databases is created in any unauthorized manner. All of these conditions are met and the source is cited here.

The dataset for Germany was subdivided into meteorological and hydrological disasters, as these two types of disasters have been found to be the most frequent. The timeline for recorded hydrological disasters ranges from 2002-2021, and for meteorological disasters from 2001-2023. Thus, the dataset appears to be appropriate to answer the questions this project aims to solve.

Citation: EM-DAT, CRED / UCLouvain, Brussels, Belgium – www.emdat.be

More information about disaster classification can be found here: Disaster Classification System.

More information on the legal use of this dataset can be found here: Terms of Use.

0.2.2 2. Food and Agriculture Organization of the United Nations (FAO)

- Metadata URL: https://www.fao.org/faostat/en/#data/ET
- Data URL: https://bulks-faostat.fao.org/production/Environment_Temperature_change_E_All_Data.zip
- Data Type: CSV (Structured and Tabular Dataset)

The FAO is a specialized agency of the United Nations that leads international efforts to defeat hunger. Access, downloading, creating copies and re-disseminating datasets are subject to the following terms: "Unless specifically stated otherwise, all datasets disseminated through the databases below are licensed under the Creative Commons Attribution-Noncommercial-Share Alike 3.0 IGO (CC BY-NC-SA 3.0 IGO)." The terms and conditions are met and the source has been cited here.

The FAOSTAT Temperature Change on Land domain disseminates statistics on mean surface temperature changes by country, with annual updates. Statistics are available for monthly, seasonal, and annual mean temperature anomalies, i.e., temperature changes concerning a baseline climatology corresponding to the period 1951–1980.

The data is complete and consistent, with all temperature values present and measured in degrees Celsius. This project aims to study only the temperature anomaly of Germany, making the dataset relevant to the problem at hand.

Citation: FAO. [Environment_Temperature_change_E_Europe_NOFLAG]. License: CC BY-NC-SA 3.0 IGO. Extracted from: FAO Database. Date of Access: 20-05-2024.

0.3 Data Pipeline

0.3.1 Running the Pipeline

The pipeline can be run using the bash script `pipeline.sh` present in the project directory. It runs the `pipeline.py`. The `pipeline.py` script runs three Python scripts present inside the **components** folder.

1. **Ingestion** The first script is **ingestion.py**, which is responsible for fetching the datasets. There are two functions for ingestion:

   - `ingestion_from_url()`: Directly fetches the dataset from the provided URL.
   - `authenticate_and_ingest()` : Although the EM-DAT dataset is available for download, the user must be logged into the EM-DAT website. To achieve this programmatically, the Chromedriver for the respective system must be installed in the PATH. By using Selenium which can handle JavaScript and interact with the browser like a human user, the authentication is executed successfully. The username and password are fetched from the `.env` file, this file must be present in the **components** folder. The environment variables are named `EM_DAT_USERNAME` and `EM_DAT_PASSWORD`.

   The creation of two different methods of fetching the datasets posed a challenge in the data ingestion phase. The ingestion step returns the dataset in CSV format to the pipeline.py.

2. **Data Preparation** The next step is data preparation, accomplished by the `data_cleaning()` function present in the script **data_prep.py**. The function takes an ID associated with the dataset to identify which dataset it is supposed to clean. The datasets being different require different cleaning steps:

   - For the 'FAO Temperature Change on Land dataset', only temperature anomaly data from 1961-2023 and Monthly(Jan-Dec)/Seasonal/Meteorological columns are filtered and rest of the columns are dropped. For the final analysis, the dataset is divided into six datasets:`temp_change_annual.csv`, `temp_change_seasonal.csv`, `temp_change_met.csv`, `std_dev_annual.csv`, `std_dev_met.csv`, `std_dev_seasonal.csv`. This function returns a dictionary, with the key being the name of the dataframe and the value being the dataset itself.

   - In the 'EM-DAT disaster dataset', there are two categories of disasters: natural and technological. This project focuses only on natural disasters. The columns 'Start Year', 'Start Month', and 'Start Day' are combined to form a full date in the format YYYY-MM-DD named 'Start Date' and converted to a DateTime object. The original columns are dropped. Similar steps are taken to create the 'End Date' column. In categorical columns, the value 'Unknown' is filled where None values are present. For the numerical columns, NaN values are filled with zeros. Finally, using the 'Meteorological' and 'Hydrological' categories of 'Disaster Subgroup' column, two dataframes are created for further analysis. The function returns a dictionary, the key is the name of the dataframe and the value being the dataframe.

   The main challenge in this step of the pipeline was deciding on a strategy to handle the incomplete data. Also, combining year, month, and day columns to form complete datetime objects and ensuring the correct handling of missing day values by substituting the first day of the month where necessary was tricky.

3. **Saving Dataframes** The last step of the pipeline is to save the returned dataframes as separate CSV files. The goal is to perform Exploratory Data Analysis (EDA) on the datasets to answer the questions posed by this project. CSV is found to be a suitable format to save the datasets.

0.4 Results and Limitations

0.4.1 Chosen Data Format for the Output of the Pipeline

The pipeline output is saved in CSV format due to its simplicity, compatibility, and efficiency. CSV files are easy to read/write, widely supported across tools, and sufficient for managing structured datasets without unnecessary complexity.

0.4.2 Critical Reflection on the Data and Potential Issues

- **Temperature Anomaly Dataset**
  - **Issues**: The dataset is of high quality with no missing values. However, temperature anomalies alone may not fully capture the complexities of climate change impacts.
- **Natural Disasters Dataset**
  - **Issues**: The dataset had missing values that were filled with zeros or 'Unknown', which could introduce bias or inaccuracies. Additionally, the economic and human impact data might be underreported due to incomplete records or reporting discrepancies.