Main Question

1. How has the temperature changed in Germany over the years?
2. How has the global CO2 emission changed over the years?
3. How has the temperature change and global CO2 emission affected the precipitation in Germany?
4. How does the global CO2 emission, temperature change and precipitation affect the soil condition?

Data sources

For this project, we have selected 5 data sources, which are global CO2, German regional monthly temperature, German regional monthly precipitation, German monthly soil condition and soil station. The detail of each data source can be found below.

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
- Data Type: CSV
- License Type:CC BY 4.0

The data comes from Our World in Data, and is licensed under CC BY 4.0 DEED, which is allowed to copy,redistribute the material and remix, transform and build upon material for any purpose, even commercially. The only one restriction is to given appropriate credit for the data. It can be illustrated in readme file.

The datasource consists of CO2 emissions of different countries from 1850 to 2022. The data source records detailed emission source, from oil, gas, industry ...etc. Moreover, emission per capita and per GDP is also accounted.

The data source is accurate as it is collected from reliable data sources. The data source is somehow complete, as some countries' data is missing. And it is also consistent, and timely, and relevant.

Datasource 2: Regional average value for historical monthly mean temperature

- Metadata URL: https://www.dwd.de/EN/ourservices/cdc/cdc_ueberblick-klimadaten_en.html;jsessionid=31CF9807245E6D5DC250829F80D5EF75.live21071?nn=495490#doc725352bodyText4:~:text=4.-,Average%20values%20for%20the%20individual%20federal%20states%20and%20for%20Germany%20as%20a%20whole,-The%20mean%20values
- Data URL: https://opendata.dwd.de/climate_environment/CDC/regional_averages_DE/monthly/air_temperature_mean/
- Data Type: CSV directory
- License Type: CC BY 4.0

The datasource is licensed under CC BY 4.0 , the same with datasource 1, is allowed to use.

The datasource consists of German regional average values for historical monthly mean temperature. The dataset dates back to 01.1881 until 01.2024.

The data source is accurate as it's from authorities. Since no missing information is found, it's complete. It's also consistent, timely and relevant.

Datasource 3: Regional average value for historical monthly mean precipitation

- Metadata URL: https://www.dwd.de/EN/ourservices/cdc/cdc_ueberblick-klimadaten_en.html;jsessionid=31CF9807245E6D5DC250829F80D5EF75.live21071?nn=495490#doc725352bodyText4:~:text=4.-,Average%20values%20for%20the%20individual%20federal%20states%20and%20for%20Germany%20as%20a%20whole,-The%20mean%20values
- Data URL: https://opendata.dwd.de/climate_environment/CDC/regional_averages_DE/monthly/precipitation/
- Data Type: CSV
- License Type: CC BY 4.0

The datasource is licensed under CC BY 4.0 , the same with datasource 1, is allowed to use.

The datasource consists of German regional average values for historical monthly mean precipitation. The dataset dates back to 01.1881 until 01.2024.

The data source is accurate as it's from authorities. Since no missing information is found, it's complete. It's also consistent, timely and relevant.

Datasource 4: Historical monthly soil condition data in Germany

- Metadata URL: https://opendata.dwd.de/climate_environment/CDC/derived_germany/soil/monthly/historical/DESCRIPTION_derivgermany_soil_monthly_historical_en.pdf
- Data URL: https://opendata.dwd.de/climate_environment/CDC/derived_germany/soil/monthly/historical/
- Data Type: GZ file directory
- License Type: CC BY 4.0

The datasource is licensed under CC BY 4.0 , the same with datasource 1, is allowed to use.

The datasource consists of historical monthly soil conditions in Germany at different stations, from 01.1991 to 12.2023. Soil Properties, like soil moisture,soil temperatures ..etc are included.

The data source is accurate as it's from authorities. Since no missing information is found, it's complete. It's also consistent, timely and relevant.

Side Datasource: Data of soil station in Germany

- Data URL: https://opendata.dwd.de/climate_environment/CDC/derived_germany/soil/monthly/historical/derived_germany_soil_monthly_historical_stations_list.txt
- Data Type: CSV
- License Type: CC BY 4.0

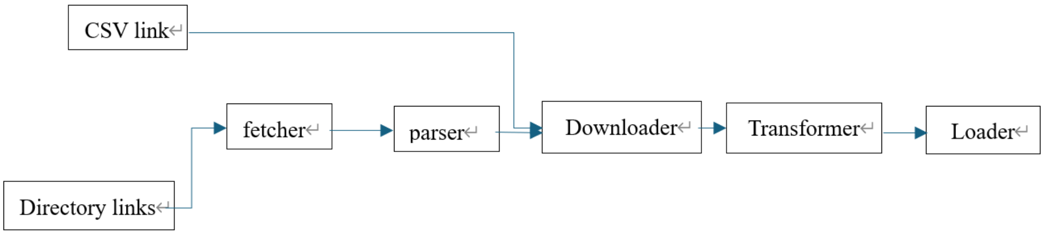
The datasource is licensed under CC BY 4.0 , the same with datasource 1, is allowed to use.

The datasource is assistance datasource, mainly cooperating with datasource 4, can be used for station details by station Index provided in datasource 4. The datasource consists of data Index, station name, station position...etc.

Data pipeline

There are two types of data sources in thi pipeline, which are CSV link and directory link respectively. For CSV links, it's directly passed to Downloader. For Directory links, it's first passed to fetcher that fetches HTML content, then passes to parser that extracts the required links from the HTML content. Downloader takes a link, downloads the corresponding file and saves it locally. The transformer does manipulation on the raw data. After transformation, the loader loads the processed data to CSV file.

The fetcher and Downloader is implemented with requests library, and parser is implemented with beautifulsoup. Both Transformer and Loader are implemented with Pandas.



Transformation in data source 1

For extracted data, we apply filter on the column 'country', only selecting the world data. Then we remove the rows that has null co2 value, afterwards we select 3 columns [year,co2,co2_growth_prct], sort data by year, and set the first row's co2_growth_prct to 0.

Transformation in data source 2

Since data source 2 is directory link, the extracted data is a list of pandas dataframe. For each dataframe, we remove the last column, since it's the error column. Then we concate these dataframes, sort by Jahr and Monat, and drop any null rows.

Transformation in data source 3

After extraction, we should obtain a list of dataframes. For these dataframes, we first drop the last column, since it's the error column. Then concate the dataframes, sort by columns Jahr and Monat, and finally drop any row with null value.

Transformation in data source 4

After extractions, we should get a list of dataframes. First, we remove last two columns, since they are null and error column. Then we concate these dataframes and drop any null rows. Since the column Monat is 'YYYYMM' formatted, we create a new column Jahr that extract years from Monat, and correct the Monat column. Finally, we sort the concated dataframe by Stationsindex,Jahr and Monat.

Transformation in data source 5

After extraction, we apply remove empty spaces on the second and third columns, and then drop any null rows. Finally, select columns Stationindex,Name and Bundesland from it.

The pipelines described above is flexible when the input data changes. If the data source 1 changes, the transformation in pipeline 1 still guarantees the output data is complete and up-to-date(If the fetched data is up-to-date). If the data source 2 changes, the transformation in pipeline 2 works similar as in pipeline 1. The same for data source 3 and 5. For data source 4, if the Monat column changes its format, we may fail this pipeline. Furthermore, if the columns of the data source is modified,(e.g reducing specific columns), all these 5 pipelines could fail. But in terms of non-structural change(i.e entry change), the pipelines still hold.

Result and Limitations

All the pipelines output the data in a CSV file with headers. Since the CSV is lightweight and easy to load and interpret, it's used. Data source 1 might be inaccurate, since it's collected and computed from multiple data sources. Some of the results are estimated, leading to some bias. And because the publisher is an environmentalism NGO, the exaggeration of CO2 emission can occur, which is biased.

One potential issue of the final report is the weak correlation between CO2 and soil condition. The soil condition depends on too many factors. Solely CO2 emission,temperature and precipitation might not contribute too much. Therefore, a weak correlation may occur. And in the data source of CO2, it only indicates how much it emits annually with ignoring the amount absorbed by nature. Low amount of CO2 in the atmosphere can happen, which can bias the analysis.