Data-report

Md Mainul Haque, 23003992

1.1 Question

How have the number and types of weather phenomena in Germany changed over time?

1.2 Overview

To answer the question, the historical data sets of 998 different stations of the German Weather Service are loaded. The data of each station is provided in a structured text file. The text files contain comma separated values with ";" as delimiter.

1.3 Data sources

Source 1: GDP Data

- **Description**: The GDP data is sourced from the World Bank, specifically the dataset containing GDP growth (annual %) for all countries.
- **Data URL:** World Bank GDP Data
- **Metadata URL:** World Bank Metadata
- **Why Chosen**: GDP growth is a key indicator of economic performance, and analyzing trends over recent years can provide valuable insights into economic development in Latin America.
- **Data Structure**: The dataset contains rows for each country and columns for various years, along with country codes and indicator code, GDP growth etc.
- **Data Quality**:
    - The dataset contains some missing values for certain years or countries.
    - Metadata files included in the ZIP archive are not relevant to the analysis and were excluded.
- **License**: The data is publicly available under the World Bank open-data license. Obligations under the license include providing attribution, which will be included in any reports or publications.
    - **License Source**: World Bank Open Data Terms

Source 2: Renewable Energy (RE) Data

- **Description**: Renewable energy data is sourced from the International Renewable Energy Agency (IRENA). It includes the renewable energy share of electricity generation and capacity for all countries.

- **Data URL**: IRENA Renewable Energy Data
- **Metadata URL**: IRENA Metadata
- **Why Chosen**: Understanding renewable energy trends is critical for evaluating sustainability in Latin American countries.
- **Data Structure**: Contains columns for Region/country/area, Year, RE share of electricity generation (%), and RE share of electricity capacity (%).
- **Data Quality**:
    - Dataset is well-structured but uses a mix of regional and country-specific entries, requiring filtering.
    - UTF-8 encoding issues were resolved by switching to latin1 encoding when necessary.
- **License**: The dataset is available under an open-data license that allows reuse for non-commercial purposes with attribution.
    - **License Source**: IRENA Data Terms

1.3.1 Data Pipeline

Overview

An automated data pipeline was developed in Python, employing:

- **Libraries**: **pandas** for data manipulation, **requests** for file downloads**, zipfile** for extracting archives, and s**qlite3** for database storage.
- **Goal**: To Download and clean GDP and Renewable Energy datasets for Latin American countries (2018–2022).

Steps

1. **Data Retrieval**:
    - Downloaded GDP data as a ZIP file and extracted relevant CSV files.
    - Downloaded Renewable Energy data directly as a CSV file.
2. **Data Cleaning**:

    GDP Data:
    - Excluded metadata files in the ZIP archive.
    - Filtered rows for Latin American countries and columns for years 2018–2022.
    - Reshaped the data from wide to long format for consistent structure.

    Renewable Energy Data:
    - Filtered rows for Latin American countries and years 2018–2022.
    - Resolved encoding issues by switching from UTF-8 to latin1 where necessary.

3. **Error Handling**:

- Dynamically identified the correct GDP file by excluding metadata files and selecting valid CSVs.
- Skipped rows with missing or invalid data and logged processing errors.

4. **Output Generation**:
   - Cleaned datasets saved as CSV files.
   - Exported data to an SQLite database for structured storage and efficient querying.

Challenges

1. **Changing GDP File Names**:
   - The main GDP data file name within the ZIP archive changes frequently. This was addressed by dynamically identifying the correct CSV file based on its structure and excluding metadata files.
2. **Encoding Issues**:
   - Addressed UTF-8 decoding errors by using latin1 encoding fallback.
3. **Inconsistent Data Formats**:
   - Both datasets required significant filtering and restructuring to ensure consistency.

1.4 Results and Limitations

GDP Dataset:

- Structure: Contains Country Name, Country Code, Year, and GDP Growth (%) for Latin American countries from 2018–2022.
- Format: Saved as gdp_data_cleaned.csv and in the SQLite table gdp_data.

Country Name	Country Code	Indicator Name	Indicator Code	Year	GDP Growth (%)
Argentina	ARG	GDP growth (annual %)	NY.GDP.MKTP.KD.ZG	2018	-2,6174E+14
Bolivia	BOL	GDP growth (annual %)	NY.GDP.MKTP.KD.ZG	2018	4,22362E+14
Brazil	BRA	GDP growth (annual %)	NY.GDP.MKTP.KD.ZG	2018	1,78367E+12
Chile	CHL	GDP growth (annual %)	NY.GDP.MKTP.KD.ZG	2018	3,99003E+14

Renewable Energy Dataset:

- Structure: Contains Region/country/area, Year, RE share of electricity generation (%), and RE share of electricity capacity (%) for Latin American countries from 2018–2022.
- Format: Saved as re_data_cleaned.csv and in the SQLite table renewable_energy_data.

Region/country/area	Year	RE share of electricity generation (%)	RE share of electricity capacity (%)
Argentina	2021	2444	3446
Argentina	2022	2918	3466
Brazil	2018	8157	8317
Brazil	2019	8142	8330

Data Structure and Quality

- Missing values were addressed by excluding non-relevant rows and columns.
- Data was normalized into long-format tables, enabling easier aggregation and analysis.

Potential Issues

Incomplete Coverage:

- In renewable energy data set missing data for 2023 that's why I can not work with latest data set.
- Some Latin American countries may have missing data for certain years.

License Obligations:

- Attribution must be provided in any derived works or publications.

Reflection

The pipeline is robust and can handle minor changes to input data structure, but reliance on external sources means future data updates may require manual adjustments. Despite these limitations, the cleaned datasets provide a reliable foundation for further analysis.