

Analyzing Greenhouse Gas (CO₂) Emissions and Temperature Trends in the Rathaus Area of Konstanz, Germany

Introduction:

The Rathaus area of Konstanz, like many other regions is experiencing the effects of climate change, which are evident in altering weather patterns and rising temperatures. Among the primary contributors to these changes are greenhouse gas emissions, particularly carbon dioxide (CO₂). This data science project aims to explore the relationship between greenhouse gas (CO₂) emissions and regional temperature trends in the Rathaus area of Konstanz. By analyzing historical data on CO₂ levels and temperature records, we seek to identify patterns and correlations that can provide insights into how human activities influence local climate changes.

Now, the Question that interests us is: How do Greenhouse Gas (CO₂) emissions affect climate change in the Rathaus area of Konstanz?

Data Sources Details: To answer the question I have used two open source data. The data were hosted on the "Offene Daten Konstanz" portal, which typically uses open data licenses such as the [Creative Commons Attribution 4.0 International \(CC BY 4.0\)](https://creativecommons.org/licenses/by/4.0/) license.

Datasource 1: Greenhouse Data

- **Metadata URL:** <https://offenedaten-konstanz.de/dataset/co-werte-konstanz>
- **Data URL:** https://offenedaten-konstanz.de/sites/default/files/CO2_Werte_Rathaus_1.csv
- **Data Type:**

The data is tabular and fits the description of structured data, making it easy to query and analyze. The CSV format used is a general-purpose format suitable for tabular data. The data aligns with the dimensions of data quality (accuracy, completeness, consistency, timeliness, and relevancy).

- **Description:**

This dataset contains information on greenhouse gas emissions, specifically CO₂, in the Rathaus area of Konstanz. It helps analyze trends in greenhouse gas emissions over time in this locality.

Datasource 2: Temperature Data

- **Metadata URL:** <https://offenedaten-konstanz.de/dataset/temperaturwerte-konstanz>
- **Data URL:** https://offenedaten-konstanz.de/sites/default/files/Temperaturwerte_Rathaus_0.csv
- **Data Type:**

The temperature dataset is structured tabular fitting well into the category of structured data. The CSV format used is a general-purpose format suitable for tabular data. Evaluating the dataset against accuracy, completeness, consistency, timeliness, and relevancy ensures high-quality data.

- **Description:**

This dataset contains information on temperature trends in the Rathaus area of Konstanz. It is essential for analyzing how temperature patterns have changed over time in correlation with CO₂ levels.

License Permissions and Obligations:

- **License:** The both datasets are available under an open-data license, which allows for use, distribution, and modification. The obligations include proper attribution and ensuring that any derived datasets also remain open. These obligations will be fulfilled by including appropriate citations in all published materials.

Data Pipeline:

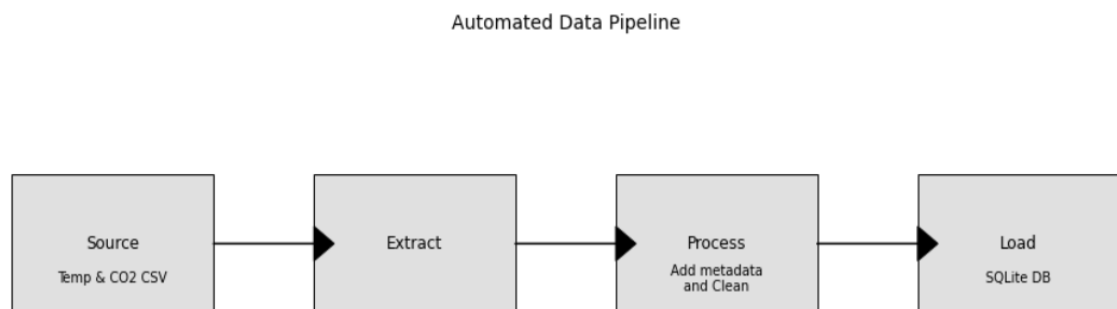


Figure: Data Pipeline

The data pipeline is designed to load, process, merge, and store datasets related to temperature and CO₂ emissions in the Rathaus area of Konstanz. The pipeline is implemented using Python, leveraging the pandas library for data manipulation and analysis, re for regular expressions to validate data formats, and sqlite3 for database operations. The overall workflow involves reading CSV files, cleaning and transforming the data, merging the datasets, storing the merged data into an SQLite database, and saving the final dataset as a CSV file.

Transformation and cleaning steps:

1. **Data Loading:** The pipeline reads data from CSV files hosted online.
2. **Data Validation:** A function checks for invalid formats in the value columns (t for temperature and eco2 for CO₂), filtering out rows that don't match the expected format (numeric values with up to three decimal places). This step ensures that only correctly formatted data is used in further analysis.
3. **Date and Time Separation:** The Date column, which contains combined date and time information, is split into separate Date and Time columns. The trailing 'Z' in the time values is also removed to standardize the format.
4. **Data Merging:** The cleaned temperature and CO₂ datasets are merged on the Date column to create a comprehensive dataset that includes both temperature and CO₂ data for each date.
5. **Data Storage:** The merged dataset is stored in an SQLite database, which allows for efficient querying and data management.

6. **Exporting Data:** Finally, the merged dataset is exported to a CSV file for easy sharing and further analysis.

Problems I encountered and solved:

- **Invalid Data Formats:** Some data values did not conform to the expected numeric format, which could affect analysis. This was solved by implementing a regular expression check to filter out invalid rows.
- **Combined Date and Time:** The Date column contained both date and time information, which needed to be split for better analysis. This was handled using string operations to separate the date and time into distinct columns.
- **Trailing Characters in Time Values:** The 'Z' character at the end of the time values needed to be removed. This was accomplished using string replacement methods.

Pipeline dealing with Errors:

- **Data Validation:** The pipeline includes a function to check and filter out invalid data formats. This ensures that only correctly formatted data is processed and analyzed.
- **Scalability:** By using SQLite for data storage, the pipeline can efficiently manage and query large datasets. The modular design allows for easy updates and modifications to accommodate new data or additional cleaning steps as needed.

Result and Limitations:

The output dataset contains merged information on temperature and CO₂ levels in the Rathaus area of Konstanz. It is Structured Tabular and well-organized for easy querying and analysis. The data is saved in CSV format due to its universality, simplicity, and compatibility with various tools.

The data has been cleaned and structured effectively, but several potential issues may arise. These include missing data, temporal misalignment, data source consistency, and the presence of outliers. While the pipeline efficiently processes and merges the data, careful consideration of these potential issues is crucial to ensure accurate and reliable analysis in the final report.