Data Report: Analyzing Climate Patterns in Milan Using an Automated Data Pipeline

Question

The primary question of this project is to analyze and compare the winter and summer climate patterns in Milan. Specifically, main aim to determine:

1. How does the heating demand vary during the winter season between different locations in Milan?
2. How many days during the summer season reach or exceed the threshold for physiological discomfort and physiological danger?

Data Sources

Choice of Data Sources

For this analysis, two datasets were chosen, each representing the climate data for Milan during the winter and summer thermal seasons. These datasets were selected due to their comprehensive coverage of climate indicators and their relevance to the project's objectives.

1. **Winter Thermal Season Data**
    - **URL**: Winter Data
    - **Metadata URL:** https://data.europa.eu/en
    - **Description**: This dataset includes climate indicators such as temperature, humidity, precipitation, wind, and radiation for seven monitoring stations in Milan during the winter season.
2. **Summer Thermal Season Data**
    - **URL**: Summer Data
    - **Metadata URL:** https://data.europa.eu/en
    - **Description**: This dataset contains similar climate indicators for the summer season, including metrics for days of physiological discomfort and danger based on the Humidex index.

Data Structure and Quality

Both datasets are structured as CSV files with climate metrics reported for different areas of Milan. The data quality is generally high, with detailed records from multiple monitoring stations ensuring comprehensive coverage of the city's climate conditions.

Licensing and Usage

Data Pipeline

Overview

The data pipeline was implemented to automate the extraction, transformation, and loading (ETL) of the climate data. The technologies used include Python, Pandas, Requests, and SQLAlchemy.

ETL Process

1. **Extraction**
    - Data is downloaded from the provided URLs using HTTP requests.
    - The raw CSV data is read into Pandas DataFrames.
2. **Transformation**
    - Columns are renamed for consistency to make metrics for both data sets.
    - DataFrames are unpivoted to a long format.
    - The winter and summer datasets are merged on common metrics and stations.
    - The merged data is pivoted to create separate columns for each metric.
3. **Loading**
    - The final DataFrame is saved to a SQLite database for easy analysis.

Error Handling and Data Quality

The pipeline includes error handling to manage issues during data download and transformation. Logging is implemented to track the process and capture any errors. The pipeline ensures that missing or inconsistent data entries are handled appropriately.

Results and Limitations

Output Data

The output of the data pipeline is a SQLite database containing the transformed climate data for Milan. This format was chosen for its ease of use and integration with analytical tools. The winter and summer datasets are merged on common metrics and stations that makes it easier to read and understand the data.

Data Structure and Quality

The output data maintains high quality, with clearly labeled metrics and stations. The data structure allows for easy querying and analysis, facilitating the project's objectives.

Potential Issues

One potential issue is the accuracy and completeness of the raw data, which depends on the original data collection methods. Additionally, changes in the format or structure of the source data could require adjustments to the pipeline.

Figure 1: Data Pipeline Structure

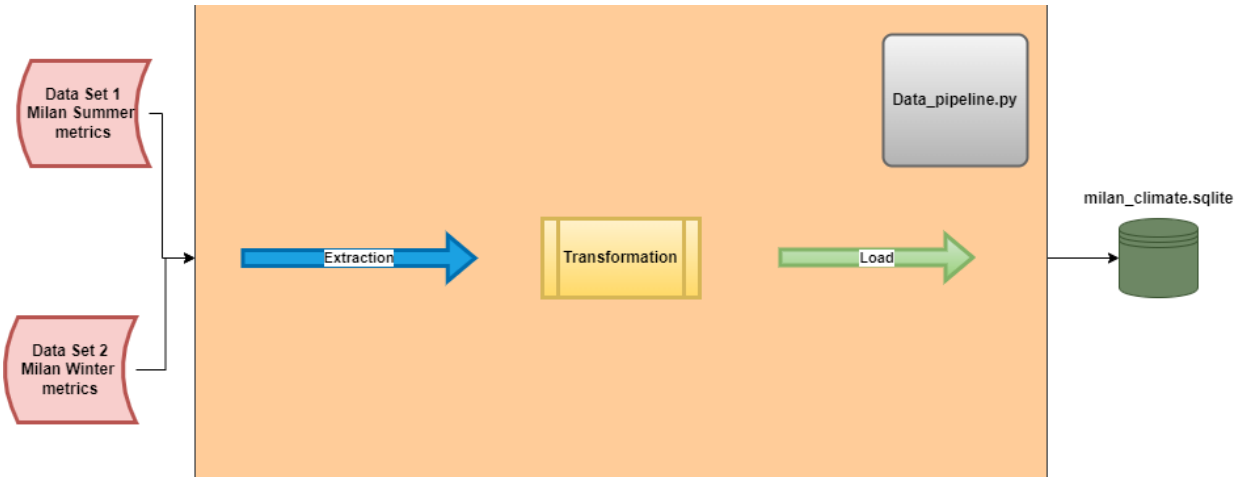

Table 1: Example of Summer transformed Data

Station	Average Days (Physio Discomfort)	Summer Thermal Degree Days
Milano Bicocca	26	344.3
Milano Bocconi	25.2	346.1
Milano Bovisa	25	331.1
Milano Centro	21	353.9
Milano Citta' Studi	27	319.5

Conclusion

The automated data pipeline significantly improved the efficiency and accuracy of processing the climate data for Milan. By automating the ETL process, tried to ensure consistent and reliable data handling, allowing for comprehensive analysis and comparison of winter and summer climate patterns.