Data Pipeline Challenge: Weather Data Analysis

Deliverables

1. Python script(s) that accomplish the above tasks
2. SQLite database file with the processed data
3. CSV report file
4. README1 file with instructions on how to run the code and any dependencies required
5. README2 file containing documentations of the pipeline

Evaluation Criteria

- Code quality and organization
- Proper use of Python data structures and libraries
- Error handling and edge cases
- Documentation and code comments
- Efficiency of data processing

Background

You've been tasked with creating a simple data pipeline to process and analyze weather data from multiple cities. The goal is to extract insights from the raw data and prepare it for further analysis.

Requirements

1. Data Ingestion:
    - Download weather data for 5 major cities from a public API (e.g., OpenWeatherMap API)
    - The data should include daily temperature, humidity, and precipitation for the last 30 days
2. Data Processing:
    - Clean the data by handling any missing values or outliers

- Convert temperature from Kelvin to Celsius
- Calculate daily average temperature, humidity, and precipitation for each city

3. Data Analysis:
    - Identify the city with the highest and lowest average temperature
    - Calculate the overall average temperature across all cities
    - Determine the city with the most rainy days (precipitation > 0)
4. Data Storage:
    - Store the processed data in a SQLite database
    - Create appropriate tables to store city information and daily weather data
5. Reporting:
    - Generate a simple CSV report with the following information:
        - City name
        - Average temperature
        - Average humidity
        - Total precipitation
        - Number of rainy days

Bonus Points (Optional)

- Implement error handling and logging
- Create a simple visualization (e.g., line chart of temperature trends) using a library like Matplotlib
- Write unit tests for key functions
- Dockerize the application