Project Work 4 - Data Report

1 Question

How did the average temperature in Alabama during the year 2020 correlate with the number of COVID-19 death cases in the state?

2 Brief Explanation on Why I Had to Change the Previous Dataset

The dataset lacked information about the location where the data was collected, which was crucial for solar radiation and weather data as these parameters can vary significantly by geographic region. Secondly, The dataset provided did not share common columns or keys with the other dataset, which was necessary for a merge. Without shared columns, finding a basis for merging was problematic.

3 Data Sources

3.1 Weather Data

Description: This dataset contains daily weather data for Alabama in the year 2020, including Date, Average Tempreture (TAVG), Minimum Tempreture (TMIN), and Maximum Temperature (TMAX).
Data Source: Meteostat
URL: https://bulk.meteostat.net/v2/daily/KGVQ0.csv.gz
License: Open data license from Meteostat
Data Quality: The dataset is structured with a clear format and includes daily records with minimal missing values. The data is reliable for time-series analysis.

3.2 COVID-19 Data

Description: This dataset includes COVID-19 cases and deaths data for counties in the USA.
Data Source: Kaggle (imdevskp/corona-virus-report)
URL: https://www.kaggle.com/imdevskp/corona-virus-report
License: Kaggle Open Data license
Data Quality: The dataset is comprehensive and contains daily records of COVID-19 cases and deaths. Some counties may have inconsistencies in reporting, which could affect data quality.

4 Data Pipeline

4.1 High-Level Description

The data pipeline involves downloading, extracting, transforming, and merging datasets, followed by loading the cleaned data into a SQLite database. Python is used for implementing the ETL (Extract, Transform, Load) process.

4.2 Transformation and Cleaning Steps

1. **Downloading and Extracting Data:**
    - Download temperature data from Meteostat and COVID-19 data from Kaggle.

2. **Data Cleaning:**
    - Rename columns for clarity.
    - Convert dates to a standard format.
    - Aggregate temperature data to calculate daily averages.
    - Summarize COVID-19 data to get daily death cases.
3. **Merging Datasets:**
    - Combine temperature and COVID-19 data based on the date.
    - Ensure alignment of date formats and handle any missing data appropriately.
4. **Loading Data into SQLite Database:**
    - Store cleaned and merged data into an SQLite database for further analysis.

4.3 Problems Encountered and Solutions

One of the main challenges was ensuring the alignment of date formats between the two datasets. This was solved by converting all date columns to a standard datetime format in Python.

4.4 Error Handling and Changing Input Data

The pipeline includes checks for data integrity and handles missing or inconsistent data points by filling missing values with appropriate measures (mean, median, or interpolation).

5 Result and Limitations

5.1 Output Data Description

The final dataset includes daily average temperatures and COVID-19 death cases for Alabama in 2020. This data is stored in a CSV format and an SQLite database for easy querying and analysis.

5.2 Data Structure and Quality

The output data is structured with columns for date, average temperature, and death cases. The quality of the data is ensured through various cleaning steps, but potential issues include missing temperature data for some dates and variations in COVID-19 reporting standards.

5.3 Output Data Format

The output data format (CSV and SQLite) is chosen for its simplicity and ease of use in further analysis.

5.4 Critical Reflection

The dataset is comprehensive and allows for a detailed analysis of the correlation between temperature and COVID-19 death cases. However, potential issues include the accuracy of reported COVID-19 cases and deaths, as well as missing temperature data for certain dates.