Fatemeh Sadat Hashemi Pour

Data Report: Correlation Between CO_2 Emissions and COVID-19 Deaths

Question

How much does CO_2 emission correlate with COVID-19 death levels per state in the USA?

This analysis investigates whether states with higher CO_2 emissions exhibit higher COVID-19 death levels. By combining environmental and public health datasets, this project seeks to uncover potential relationships and trends between pollution and health outcomes.

Data Sources

1. COVID-19 Deaths by State

- Source: The New York Times COVID-19 Dataset
- Why Chosen: This dataset provides live, state-level data on confirmed COVID-19 cases and deaths in the U.S., essential for analyzing health impacts across states.
- Content:
  - Fields: state, cases, deaths.
  - Structure: CSV format with live updates.
  - Quality: High coverage of states and consistent updates. However, potential missing or null values for deaths are handled in the pipeline.
- License: Creative Commons Attribution 4.0 International (CC BY 4.0).
  - Obligations: Proper attribution to The New York Times in any derived work.
  - Plan to Fulfill: Clearly cite the source in this report and any published results.

2. Energy-Related CO_2 Emissions Data

- Source: EIA Energy-Related CO_2 Emissions Data
- Why Chosen: The dataset provides annual state-level CO_2 emissions, a key variable for assessing environmental impact.
- Content:
  - Fields: State, Carbon Dioxide Emissions (million metric tons).
  - Structure: Excel format; annual data with structured rows and headers.
  - Quality: Comprehensive but includes footnotes and irrelevant rows that require cleaning.

- License: Public domain data from the U.S. Energy Information Administration (EIA).
    - Obligations: Acknowledge EIA as the source in this report and analysis outputs.

Data Pipeline

Overview

The pipeline automates the data acquisition, cleaning, integration, and analysis processes to produce a merged dataset for correlation analysis. The pipeline was implemented in Python using libraries like pandas, requests, and sqlite3.

Steps

1. Data Acquisition
    - The CO_2 emissions data (Excel) and COVID-19 deaths data (CSV) are downloaded using requests from their respective sources.
2. Data Cleaning
    - CO_2 Data:
        - Removed footnotes and irrelevant rows.
        - Renamed columns for consistency (State → state, Carbon Dioxide Emissions → co2_emissions).
    - COVID-19 Data:
        - Handled missing values by filling null deaths and cases with zero.
        - Standardized state names for compatibility.
3. Data Integration
    - Merged both datasets on the state column using a case-insensitive match.
    - Ensured consistent naming conventions and data types.
4. Analysis
    - Computed Pearson's correlation coefficient to assess the relationship between CO_2 emissions and COVID-19 death levels.
5. Data Storage
    - Cleaned and merged data were stored in an SQLite database (eia_covid_data) for efficient querying.

Challenges and Solutions

1. Challenge: Mismatched state names between datasets.
   Solution: Used a mapping function to standardize state names before merging.
2. Challenge: Extra footnotes and irrelevant rows in CO_2 data.
   Solution: Skipped rows during data loading and dynamically identified relevant columns.
3. Challenge: Live updates in COVID-19 data causing discrepancies.
   Solution: Focused analysis on a snapshot of the data to ensure consistency.

Result and Limitations

Output Data

- Structure:
  - Fields: state, co2_emissions, deaths.
  - Data Type: Tabular data stored in SQLite database (eia_covid_data).
- Quality:
  - Cleaned and integrated dataset, with consistent state names and no missing values.
  - Ready for further analysis and visualization.

Output Format

- SQLite:
  - Efficient for structured data storage and querying.

Limitations

1. Granularity:
   - CO_2 data is annual, while COVID-19 data is updated live. Temporal mismatches may affect correlation strength.
2. Confounding Factors:
   - Other variables (e.g., population density, healthcare quality) that influence COVID-19 deaths are not included.
3. Causality:
   - Correlation does not imply causation. The analysis cannot establish that CO_2 emissions directly influence COVID-19 deaths.