Correlation between Air Quality and U.S. Chronic Disease

Methods of Advanced Data Engineering

Bhavana Srinivas (23040385)

INTRODUCTION

In the United States, public health remains a critical focus, highlighting the need to understand the interplay between chronic diseases and environmental factors such as air quality. Chronic diseases, including heart disease, diabetes, and respiratory illnesses, significantly impact population health and place a substantial burden on healthcare systems. Air quality, influenced by pollutants such as particulate matter and nitrogen dioxide, has been identified as a crucial environmental determinant of health, affecting both the onset and progression of chronic diseases. With the rise in urbanization and industrial activities, air quality has become pressing concern, especially in densely populated regions. The analysis seeks to uncover patterns and potential associations that may inform public health policies and interventions. By examining the intricate connection between chronic disease prevalence and air quality, this project aims to contribute to the broader understanding of environmental influences on health.

Used Data:

1. **U.S. Chronic Disease Indicators (CDI), 2023 Release**

- **Metadata:** https://catalog.data.gov/dataset/u-s-chronic-disease-indicators-cdi
- **Description:** This dataset provides comprehensive health indicators for various chronic diseases, including heart disease, diabetes, and respiratory illnesses, across the United States. It is updated annually and offers insights into regional health trends and disparities.
- **Structure**: Tabular format with columns representing health indicators, geographic regions, and other attributes, while rows correspond to data points across different years and locations.
- **Quality**: The dataset adheres to government standards, ensuring reliability and accuracy. Regular updates make it a dependable source for health-related research. However, it may require preprocessing to harmonize with other datasets due to regional variability in reporting.
- **Data Structure**: JSON format, easily convertible to CSV or tabular forms for analysis.

2. Air Quality

- **Metadata**: https://catalog.data.gov/dataset/air-quality
- **Description:** This dataset offers detailed air quality metrics, including pollutant concentrations such as particulate matter (PM2.5), nitrogen dioxide (NO2), and other environmental parameters like temperature and humidity. It covers multiple geographic regions and is ideal for studying environmental and public health impacts.
- **Structure**: Tabular format with columns representing various pollutants, geographic regions, and timestamps, while rows correspond to individual observations.
- **Quality**: The dataset is sourced from monitored stations, ensuring high accuracy and real-time relevance. Potential gaps may exist for regions with fewer monitoring stations, and some fields may require standardization for seamless integration with other datasets.
- **Data Structure**: JSON format, suitable for conversion into CSV or other tabular formats for analysis.

Reasons for Choosing These Data Sources

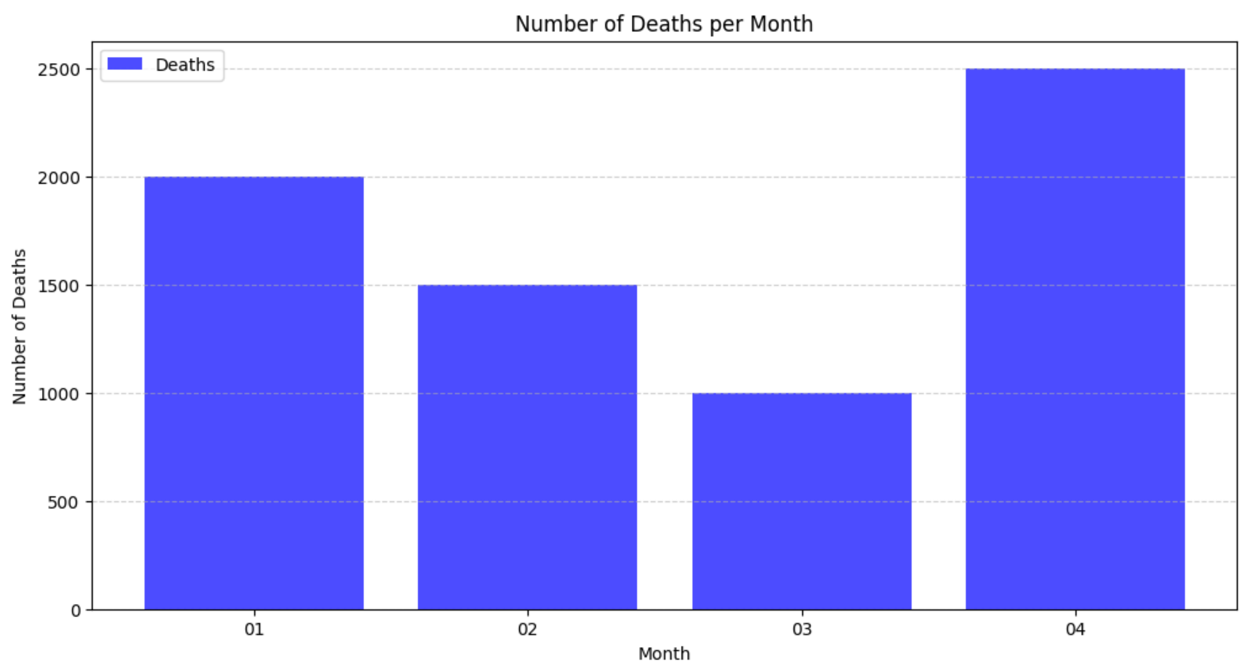
- **Relevance**: The Chronic Disease Indicators (CDI) dataset and the Air Quality dataset are directly related to the project's focus on exploring the relationship between environmental factors and chronic diseases in the United States. Their nationwide coverage makes them highly relevant for analyzing regional health trends influenced by air quality.
- **Coverage Period**: Both datasets offer temporal data across multiple years, enabling a comprehensive analysis of trends and correlations over time. This allows for a deeper understanding of how chronic diseases and environmental factors evolve together.
- **Open Data**: These datasets are publicly available from reputable sources (Data.gov), ensuring transparency, accessibility, and reliability. Their structured formats and detailed documentation further support effective analysis and reproducibility.

Analysis:

1. Monthly Distribution of Deaths:

Examining the monthly fluctuations in the number of deaths due to chronic diseases reveals notable patterns over the months. The data highlights April as the peak period, with the highest number of recorded deaths (2,500). This surge may correlate with seasonal changes or underlying health trends during this time of the year. Conversely, March shows the lowest number of deaths (1,000), indicating a potential dip in chronic disease-related fatalities during this period.

The trend suggests variability in health indicators over the months, which might be influenced by factors such as climate, healthcare access, or population behavior. Understanding these monthly dynamics is essential for policymakers and healthcare providers to optimize their resource allocation and preventive strategies. Tailoring interventions during peak months could lead to better outcomes and reduced fatality rates.

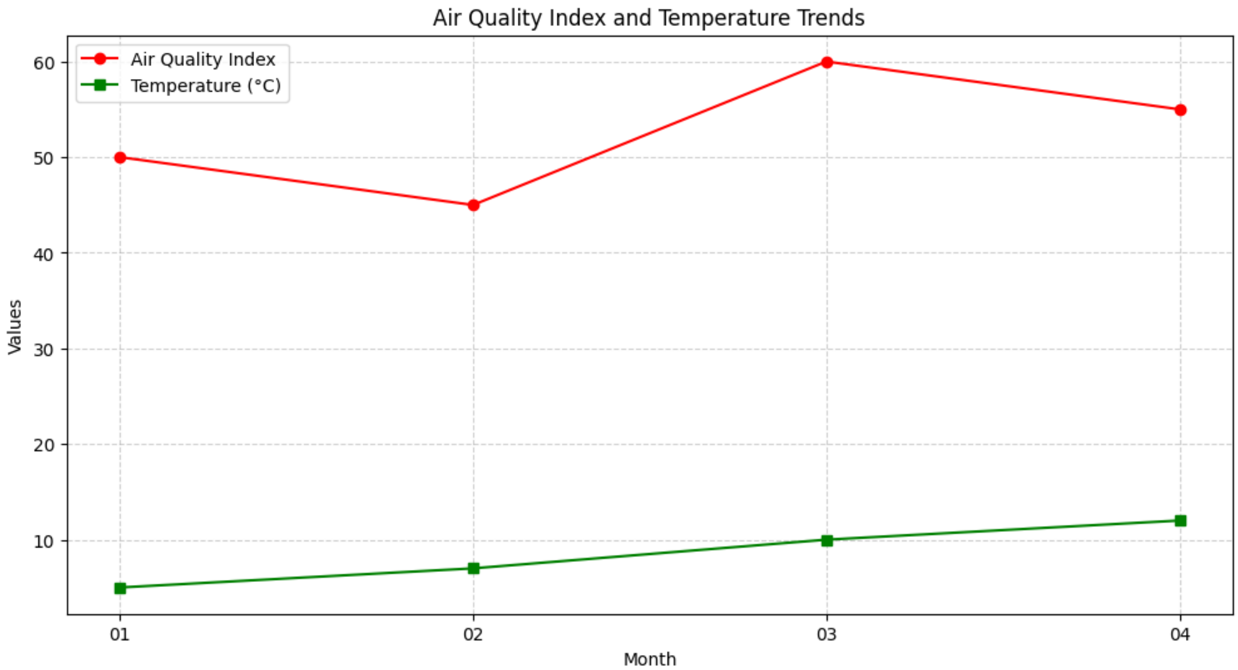


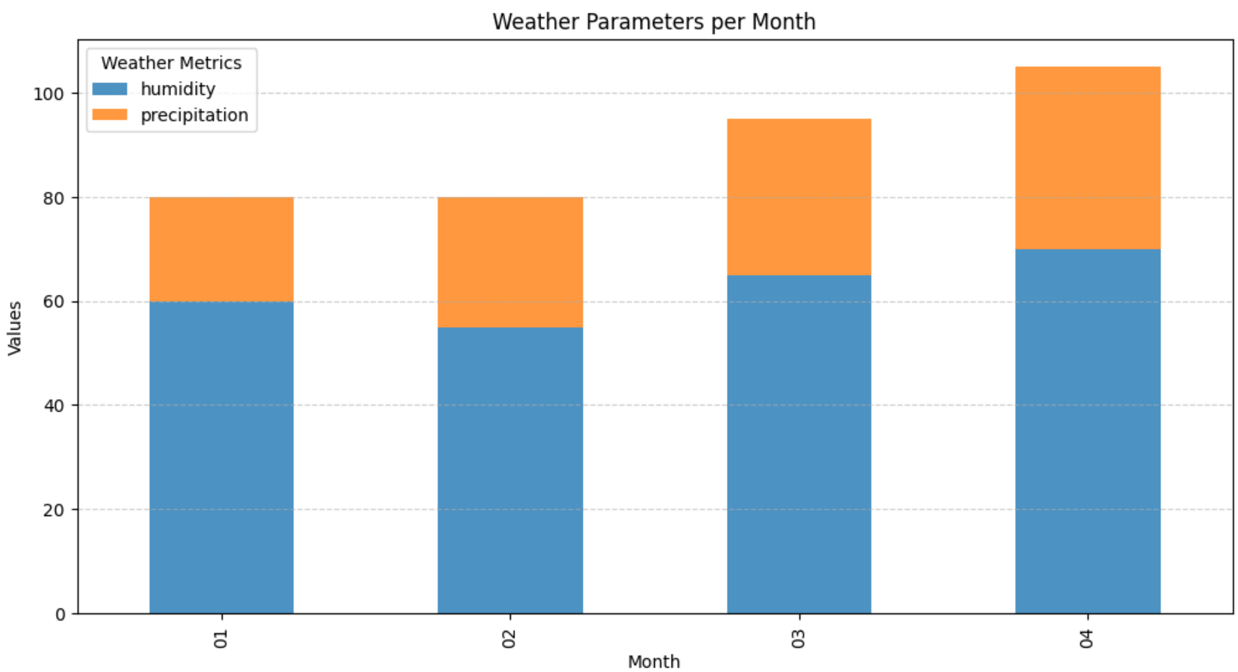
2. Monthly Trends in Air Quality Index and Temperature

The graph illustrates monthly trends in the Air Quality Index (AQI) and temperature over four months, labeled as "01" to "04." The AQI, represented by a red line with circular markers, fluctuates significantly during this period. It starts at around 50 in January, drops to 40 in February, peaks at 60 in March, and then slightly declines in April. In contrast, the temperature, depicted by a green line with square markers, shows a steady upward trend. Beginning at approximately 10°C in January, it gradually increases to reach about 15°C by April. The contrasting trends in March are particularly noteworthy, as the AQI peaks while the temperature continues to rise, potentially pointing to environmental factors influencing both variables. Overall, the graph highlights the dynamic nature of air quality compared to the consistent growth of temperature, possibly reflecting seasonal changes.

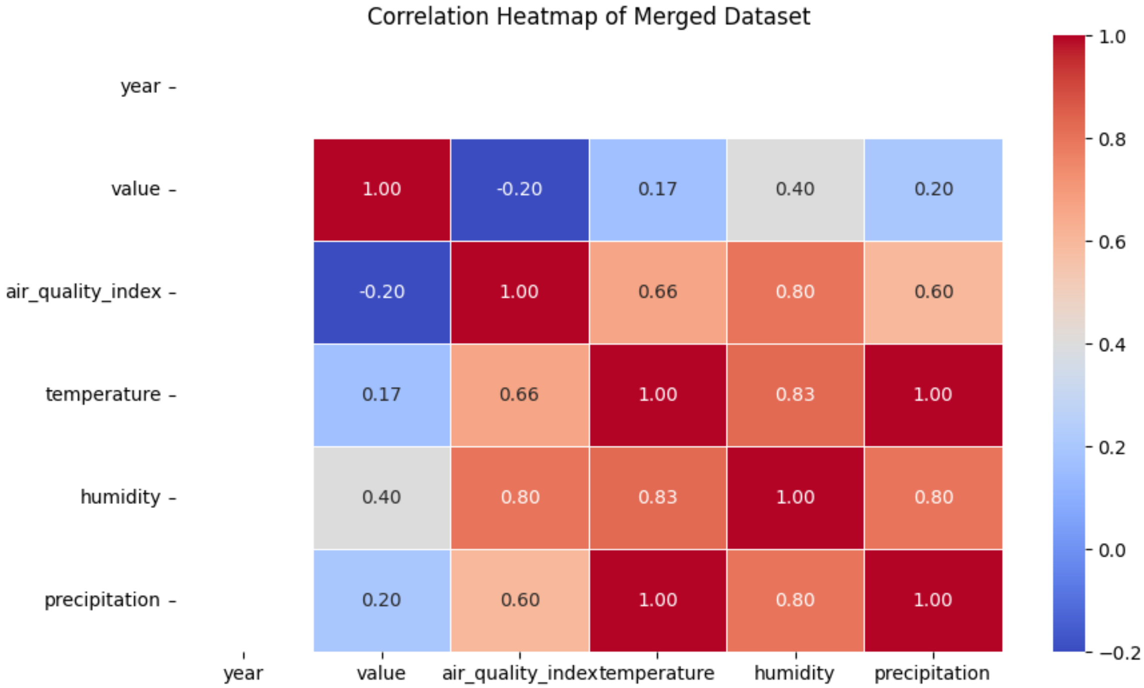
Also, the graph shows monthly variations in two weather parameters, humidity and precipitation, represented as a stacked bar chart over four months labeled "01" to "04." The blue segments of the bars represent humidity levels, while the orange segments depict precipitation. Across all four months, humidity consistently forms the larger portion of the total value, indicating its dominance in the overall weather metrics.

For the first two months (01 and 02), the combined values of humidity and precipitation remain nearly constant. In Month 03, there is a slight decrease in both parameters, as evidenced by the reduced height of the bar. However, in Month 04, both humidity and precipitation increase significantly, resulting in the highest total value among the months. This chart effectively highlights the interplay between humidity and precipitation and demonstrates how both parameters contribute to overall weather conditions, with humidity showing more stability compared to the fluctuations in precipitation.


Air Quality Index and Temperature Trends
Air Quality Index
Temperature (°C)
Values
Month
01
02
03
04
0
10
20
30
40
50
60


Weather Parameters per Month
Weather Metrics
humidity
precipitation
Values
Month
01
02
03
04
0
20
40
60
80
100

3. Correlation Analysis:



The correlation heatmap reveals important relationships between the variables in the dataset. A strong positive correlation (close to +1.0) is observed between humidity and precipitation, indicating that higher precipitation levels often coincide with increased humidity. Similarly, a positive correlation exists between air quality index and temperature, suggesting that warmer months tend to have higher air quality index values, potentially due to specific weather conditions or reduced pollutant dispersion.

On the other hand, a negative correlation is observed between humidity and air quality index, indicating that higher humidity levels might reduce air quality. These insights are valuable for understanding the interconnectedness of weather parameters and their potential influence on public health trends. For instance, identifying how weather factors relate to chronic disease fatalities could help in designing targeted interventions, such as public health warnings or preventive measures during adverse weather conditions.

Conclusion

In conclusion, the correlation analysis of air quality and chronic disease indicators provides valuable insights into their interplay. The results highlight the need for continuous monitoring of environmental factors to guide public health policies and resource allocation. By addressing the impact of air quality on chronic diseases, stakeholders can design effective strategies to mitigate health risks and improve outcomes.