Lauren Butcher

DSC680: Applied Data Science

June 23, 2024

Milestone 2- Draft White Paper

# Air Quality Analysis: Identifying Pollution Trends and Informing Policy

**Business Problem**

The primary goal of this project is to analyze air quality data from New York City and Beijing to identify specific pollutants, understand their temporal trends, assess their impact on respiratory health, and recommend targeted policies to reduce pollution and improve public health outcomes. Specifically, the business question is: How do specific air pollutants in New York City and Beijing impact respiratory health, and what targeted policies can be implemented to reduce these pollutants and improve air quality?

**Background/History**

Air pollution is a critical public health issue that affects millions of people worldwide. Urban areas, in particular, face severe pollution problems due to industrial activities, vehicle emissions, and other anthropogenic sources. Previous studies have linked poor air quality to a range of health issues, including asthma, bronchitis, and cardiovascular diseases. This project focuses on two major urban centers, New York City and Beijing, which are known for their high pollution levels and significant health impacts.

**Data Explanation**

Datasets Used:

* World Air Quality Data 2024 (Updated)
* Air Quality Data - NYC
* Beijing Multi-Site Air-Quality Data Set
* 2023 Air Quality Data for CBSAs
* Air Quality Data in India (2015-2020)

Data Preparation:

* Cleaning: Handling missing values, correcting erroneous entries, and ensuring consistency.
* Integration: Combining data from different sources to create a comprehensive dataset.
* Feature Engineering: Creating new variables such as pollution averages, peak pollution times, and seasonal trends.

Data Dictionary:

* PM2.5: Particulate matter less than 2.5 micrometers.
* PM10: Particulate matter less than 10 micrometers.
* NO2: Nitrogen Dioxide levels.
* SO2: Sulfur Dioxide levels.
* CO: Carbon Monoxide levels.
* O3: Ozone levels.
* AQI: Air Quality Index.

**Methods**

The analysis will involve:

* Exploratory Data Analysis (EDA): To understand the data distribution and identify patterns.
* Time Series Analysis: To examine trends over time and identify peak pollution periods.
* Predictive Modeling: Using machine learning models to forecast future pollution levels.
* Geospatial Analysis: To visualize the spatial distribution of pollutants.
* Health Impact Analysis: Correlating pollution data with public health records to assess health impacts.

**Analysis**

**Visualization 1: Trend of PM2.5 Levels Over Time in Beijing**

**A graph of blue lines

Description automatically generated**

**Visualization 2: Comparison of PM2.5 and PM10 Levels in New York City**

**A graph with blue dots and green dots

Description automatically generated**

**Visualization 3: Average AQI Levels by City**

**A purple bars with text on it

Description automatically generated with medium confidence**

**Visualization 4: Distribution of NO2 Levels in New York City**

**A graph of a number of levels

Description automatically generated**

**Visualization 5: Correlation Heatmap of Pollutants in Beijing**

**A screenshot of a graph

Description automatically generated**

Detailed statistical analysis will be conducted to determine the relationship between pollution levels and health outcomes. This will include:

* Descriptive Statistics: Summarizing the data to understand the central tendency and dispersion.
* Correlation Analysis: Identifying relationships between different pollutants and health indicators.
* Regression Analysis: To predict health outcomes based on pollution levels.
* Visualization: Graphs and maps to illustrate key findings.

**Conclusion**

1. Trend Analysis:
   * The PM2.5 levels in Beijing show significant seasonal variations with peaks corresponding to specific times of the year, indicating the impact of factors like weather conditions and human activities.
   * In New York City, the Ozone (O3) and Sulfur Dioxide (SO2) levels have shown some fluctuations over the observed period, reflecting changes in emissions and regulatory impacts.
2. Average AQI Levels:
   * The average AQI levels vary significantly across different cities, indicating localized pollution issues and highlighting areas that require more stringent pollution control measures.
3. Pollutant Distribution:
   * The distribution of NO2 levels in New York City shows that most measurements fall within a certain range, with occasional spikes indicating higher pollution levels during specific events or periods.
4. Correlation Analysis:
   * The correlation heatmap for Beijing shows strong correlations between certain pollutants, indicating common sources or related atmospheric processes affecting multiple pollutants simultaneously.

**Implications for Public Health and Policy**

* Health Impact: The identified trends and pollutant levels have direct implications for public health, particularly in terms of respiratory conditions such as asthma and bronchitis.
* Policy Recommendations:
  + Implement stricter emission standards to control industrial and vehicular emissions.
  + Promote green spaces in urban areas to help absorb pollutants and improve air quality.
  + Increase public awareness about the sources and health impacts of air pollution to encourage community actions and support for pollution control measures.

**Assumptions**

* Data Accuracy: Assuming the datasets are accurate and representative of actual conditions.
* Consistency: Assuming pollution levels and health impacts are consistently measured across different datasets.

**Limitations**

* Data Gaps: Incomplete or missing data points may affect the analysis.
* Generalizability: Findings may not be generalizable to other cities or regions.
* Model Limitations: Predictive models may not capture all factors influencing air quality.

**Challenges**

* Data Integration: Combining datasets from different sources.
* Handling Missing Data: Imputing or excluding missing values without introducing bias.
* Model Accuracy: Ensuring the predictive models are accurate and reliable.

**Future Uses/Additional Applications**

* Extended Analysis: Applying the methodology to other cities or regions.
* Policy Development: Using the findings to inform broader policy initiatives.
* Public Health Interventions: Developing targeted health interventions based on the analysis.

**Recommendations**

Based on the analysis, specific policy recommendations will be made, such as:

* Stricter Emission Standards: Implementing stricter controls on industrial emissions.
* Increased Green Spaces: Promoting urban green spaces to improve air quality.
* Public Awareness Campaigns: Educating the public on the health impacts of air pollution and how to mitigate them.

**Implementation Plan**

* Data Collection: Ongoing monitoring and data collection.
* Policy Development: Working with policymakers to develop and implement recommendations.
* Community Engagement: Engaging with the community to raise awareness and promote action.

**Ethical Assessment**

* Data Privacy: Ensuring data privacy and security, especially for health data.
* Accuracy and Honesty: Presenting data and findings accurately to avoid misinformation.
* Equity: Ensuring recommendations benefit all segments of the population, particularly vulnerable groups.

**Appendix**

Supporting Documentation

* Methodology: Detailed explanation of the methods used.
* Additional Graphs and Tables: Supplementary analysis and visualizations.
* Supplementary Analysis: Additional findings not included in the main report.

**10 Questions an Audience Might Ask**

1. What specific pollutants have the most significant impact on health in New York City and Beijing?
2. How do pollution levels vary seasonally in these cities?
3. What are the primary sources of pollution in these urban areas?
4. How reliable are the predictive models used in the analysis?
5. What health outcomes are most strongly correlated with pollution levels?
6. What are the limitations of the data used in the analysis?
7. How can policymakers use the findings from this project?
8. What are some potential interventions to reduce pollution in these cities?
9. How does air quality in New York City and Beijing compare to other major cities?
10. What additional data would improve the accuracy of the analysis?

**APA References**

1. World Health Organization. (n.d.). Air quality guidelines - global update 2005. Retrieved from <https://www.who.int/phe/health_topics/outdoorair/outdoorair_aqg/en/>
2. United States Environmental Protection Agency. (n.d.). Air quality data. Retrieved from <https://www.epa.gov/outdoor-air-quality-data>
3. Rao, R. (2015). Air Quality Data in India (2015-2020). Kaggle. Retrieved from <https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india>
4. Perry, N. (2023). 2023 Air Quality Data for CBSAs. Kaggle. Retrieved from <https://www.kaggle.com/datasets/nikkiperry/2023-air-quality-data-for-cbsas>
5. Kaggle. (2022). Air Quality Data - NYC. Kaggle. Retrieved from <https://www.kaggle.com/datasets/kaggleprollc/air-quality-data-nyc>
6. Kanchana. (2024). World Air Quality Data 2024 (Updated). Kaggle. Retrieved from <https://www.kaggle.com/datasets/kanchana1990/world-air-quality-data-2024-updated>
7. Sid. (2017). Beijing Multi-Site Air-Quality Data Set. Kaggle. Retrieved from <https://www.kaggle.com/datasets/sid321axn/beijing-multisite-airquality-data-set>