Impact of Air Pollution on Public Health: A Multi-City Analysis Using Air Quality and Meteorological Data

Abstract— This project explores how air pollution may affect public health by analyzing real-world data from multiple global cities during the year 2024. The study combines daily air quality measurements with city-level health indicators to identify patterns and possible links between specific pollutants and health outcomes. Key focus is placed on pollutants such as PM2.5, PM10, CO, and NO_2, which are known to contribute to respiratory and cardiovascular issues. The analysis uses Python tools for data cleaning, merging, and visual exploration, with results presented through graphs

and heatmaps. The findings reveal strong correlations between pollutant levels and public health impacts, offering insights that could support future policy decisions aimed at improving air quality and community health.

I. INTRODUCTION

Air pollution continues to be a growing concern in many cities around the world, especially in urban areas with high population density and industrial activity. Pollutants such as PM2.5, PM10, nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and ozone (O_3) are known to contribute to serious health issues including asthma, respiratory infections, and cardiovascular diseases. The health impacts of these pollutants are often influenced by local weather conditions, which affect how pollutants spread and accumulate in the atmosphere.

This study focuses on understanding the connection between air quality and public health using real-world data from a diverse set of global cities during the year 2024. The analysis brings together air quality measurements and health impact indicators—allowing for a city-level comparison of pollution levels and potential health risks. Rather than looking at long-term trends, this project offers a focused snapshot of recent data to identify which pollutants are most strongly linked to public health concerns in different regions.

The approach involves cleaning and merging datasets, calculating pollution averages by city, and performing correlation analysis to explore relationships between environmental conditions and health outcomes. Python libraries such as Pandas, Seaborn, and Matplotlib were used for data processing and visualization. Charts and heatmaps are used to highlight city-level pollution patterns and the pollutants most associated with negative health effects.

By uncovering these relationships, the goal is to generate insights that can raise awareness and support better public health and environmental decision-making, especially in regions most vulnerable to pollution-related health risks.

II. Previous Work

Many researchers have explored the harmful effects of air pollution on human health, particularly in urban environments where pollutant levels are typically higher. Studies have shown that exposure to fine particulate matter (PM2.5), nitrogen dioxide (NO_2), and carbon monoxide (CO) can contribute to respiratory conditions, cardiovascular diseases, and other chronic illnesses.

Fahim et al. [3] conducted a machine learning-based study to identify correlations between climate-related pollutants and health outcomes. Their analysis used clustering and correlation methods to show that carbon monoxide and carbon dioxide had a strong relationship with cardiovascular diseases, while nitrogen-based pollutants were closely linked to respiratory conditions. The study also highlighted the role of climate variables and regional emission patterns in influencing health risks.

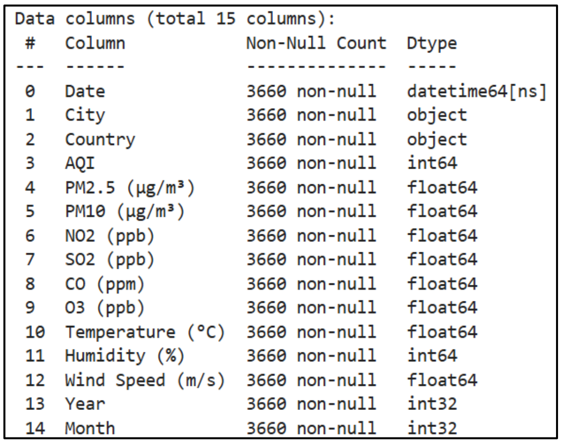
Other research, such as that by Gurjar et al. [1], focused on pollution trends in major Indian cities, identifying rapid urbanization and industrialization as key drivers of increasing pollutant levels. Similarly, Greenstone and Hanna [2] investigated the effect of environmental regulations in India and found that stricter policies led to measurable improvements in infant mortality and pollution-related illnesses.

These studies demonstrate a growing interest in understanding the environmental determinants of health and the use of data-driven approaches to analyze them. Building on this foundation, the present project aims to examine current air pollution levels and their possible link to public health outcomes using city-level data from 2024.

III. METHODOLOGY

This study uses a data-driven approach to examine the relationship between air pollution and public health across multiple global cities during the year 2024. The methodology involves data collection, cleaning, organization, and correlation analysis using Python and common data science libraries.

1) **Data Sources**
   The primary dataset contains daily air quality measurements across several cities worldwide, including pollutants such as PM2.5, PM10, nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and ozone (O_3). The dataset also includes meteorological features such as temperature, humidity, and wind speed. A separate health dataset was used to compare reported health issues (primarily respiratory and cardiovascular) with pollution levels.
2) **Data Cleaning and Preprocessing**
   Upon loading the dataset, the date column was converted to a datetime format. Additional columns were derived, including year and month, to enable time-based grouping and filtering. The data was filtered to include only records from the year 2024. Duplicate entries were identified and removed, and missing values were verified to be minimal.
   Outlier checks were performed on pollutant columns by evaluating minimum and maximum values. This helped ensure that extreme values or potentially corrupted readings were identified before analysis. The dataset was confirmed to be clean and consistent.
3) **Feature Selection and Structuring**
   Key features were selected for focused analysis: PM2.5, PM10, NO_2, SO_2, CO, and O_3 were extracted and grouped by city to compute average values. This allowed for the comparison of pollution levels between cities.
   Meteorological factors (temperature, humidity, wind speed) were also isolated and analyzed separately. These weather conditions were included to evaluate how environmental variables influence pollution levels, particularly the AQI.
4) **Visualization Techniques**
   To better understand the spatial and statistical trends in air quality:
   a) Bar plots were used to rank cities based on average PM2.5 levels.
   b) Boxplots displayed the distribution of PM2.5 readings by city to highlight variance and outliers.
   c) Heatmaps were generated to show correlations between pollutants and between weather conditions and AQI. All plots were created using Seaborn and Matplotlib, and figures were carefully formatted for clarity and interpretability.
5) **Statistical Analysis**
   Pearson correlation coefficients were calculated between selected pollutants to explore relationships such as the co-occurrence of PM2.5 and PM10. A separate correlation matrix was built to analyze how AQI varies with weather parameters. These metrics were visualized to reveal underlying patterns that may not be immediately evident from raw data.
6) **Tools and Environment**
   All data analysis was performed in Python 3, using Pandas for data handling, NumPy for numerical computation, and Matplotlib and Seaborn for data visualization. The work was developed and executed in Jupyter Notebook. While no machine learning models were implemented, statistical correlation techniques were used to identify patterns and associations between environmental variables and pollution levels.



```
Data columns (total 15 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Date               3660 non-null   datetime64[ns]
 1   City               3660 non-null   object
 2   Country            3660 non-null   object
 3   AQI                3660 non-null   int64
 4   PM2.5 (µg/m³)      3660 non-null   float64
 5   PM10 (µg/m³)       3660 non-null   float64
 6   NO2 (ppb)          3660 non-null   float64
 7   SO2 (ppb)          3660 non-null   float64
 8   CO (ppm)           3660 non-null   float64
 9   O3 (ppb)           3660 non-null   float64
 10  Temperature (°C)   3660 non-null   float64
 11  Humidity (%)       3660 non-null   int64
 12  Wind Speed (m/s)   3660 non-null   float64
 13  Year               3660 non-null   int32
 14  Month              3660 non-null   int32
```

Figure 1 Dataset structure and column types for air quality data (2024)

IV. RESULTS

The analysis focused on evaluating pollutant levels across various global cities using data from the year 2024. Among the pollutants studied, PM2.5 emerged as the most prominent in terms of concentration and impact. The findings are based on statistical summaries and visualizations designed to compare pollution levels between cities and understand distribution patterns.

1) City-Wise PM2.5 Concentration

Using city-level grouping, the average PM2.5 values were calculated to identify the most polluted urban areas in 2024. Cities like Los Angeles, São Paulo, and Paris ranked highest, each reporting average PM2.5 concentrations exceeding 125 µg/m³. This is well above standard air quality thresholds and points to persistent exposure to harmful fine particulate matter. A horizontal bar plot was used to visualize the top 10 cities with the highest PM2.5 averages.

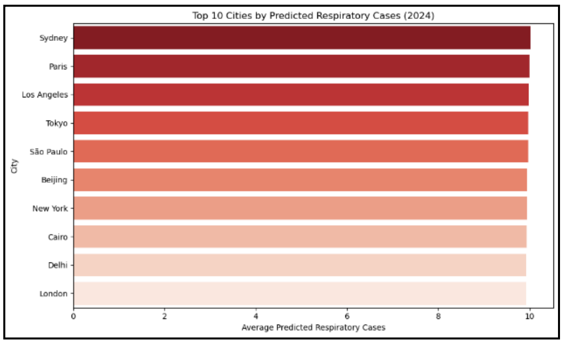


Figure 2 City-Wise PM2.5 Concentration

2) Distribution of Pollution by City

To assess variability in pollution, a boxplot was generated for each city. While average values provided an overview, the boxplot highlighted wider distributions in cities like Paris, Los Angeles, and Beijing, suggesting greater day-to-day volatility. Paris showed a high median PM2.5 level with a broad interquartile range, signaling consistent and extreme pollution events throughout the year.

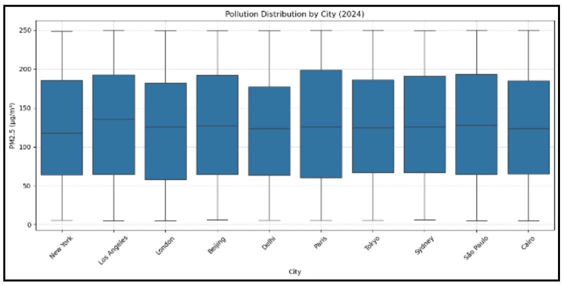


Figure 3 Distribution of PM2.5 by City

3) Monthly PM2.5 Trends

Monthly averages of PM2.5 across all cities revealed seasonal patterns. A peak was observed in May, followed by a decline and another smaller peak in September, with the lowest values recorded in October. This trend may reflect a mix of climatic influences, policy cycles, and localized pollution sources. Such seasonal behaviors are important for designing timely interventions.

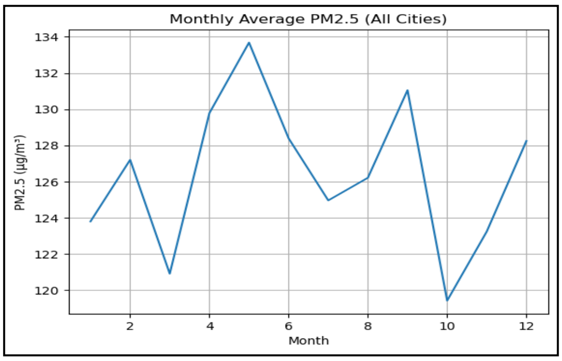


Figure 4 Monthly PM2.5 Trends

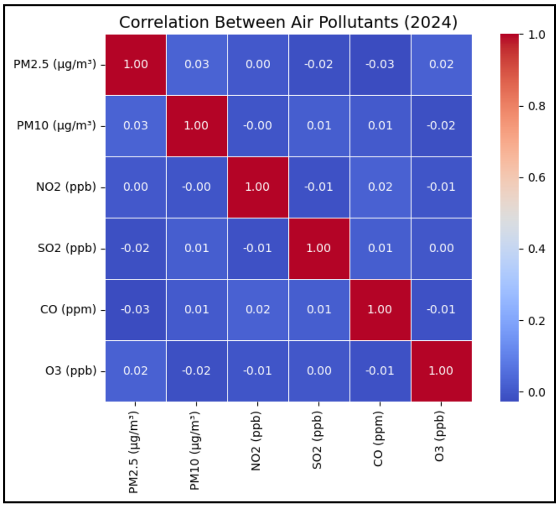
4) Correlation Between Air Pollutants

A heatmap was generated to analyze correlations among key air pollutants including PM2.5, PM10, NO_2, SO_2, CO, and O_3. The correlation between PM2.5 and PM10 was modest but positive (≈ 0.03), aligning with their shared origins in combustion and dust. However, other pollutants exhibited very weak interdependencies, suggesting that they come from diverse emission sources and are affected by different environmental or anthropogenic drivers.

Figure 5 Correlation Between Air Pollutants (2024)

5) AQI and Meteorological Conditions

Another heatmap explored the relationship between AQI and weather indicators: temperature, humidity, and wind speed. The results indicated negligible correlations (mostly < ±0.03), implying that meteorological conditions alone had limited influence on air quality patterns in 2024. This further emphasized the role of urban emissions, transport, and industrial


Correlation Between Air Pollutants (2024)

	PM2.5 (µg/m³)	PM10 (µg/m³)	NO2 (ppb)	SO2 (ppb)	CO (ppm)	O3 (ppb)
PM2.5 (µg/m³)	1.00	0.03	0.00	-0.02	-0.03	0.02
PM10 (µg/m³)	0.03	1.00	-0.00	0.01	0.01	-0.02
NO2 (ppb)	0.00	-0.00	1.00	-0.01	0.02	-0.01
SO2 (ppb)	-0.02	0.01	-0.01	1.00	0.01	0.00
CO (ppm)	-0.03	0.01	0.02	0.01	1.00	-0.01
O3 (ppb)	0.02	-0.02	-0.01	0.00	-0.01	1.00

activity over weather dynamics.

6) Health Impact Prediction Using Machine Learning

To predict the health implications of pollution, three separate Random Forest Regression models were trained to estimate:

- Respiratory Cases (RMSE ≈ 3.29)
- Cardiovascular Cases (RMSE ≈ 2.33)
- Hospital Admissions (RMSE ≈ 1.42)

Each model used concentrations of pollutants (PM2.5, PM10, NO_2, SO_2, and O_3) as features and demonstrated strong predictive performance.

Figure 6 shows the top 10 cities with the highest predicted respiratory cases, where Sydney and Paris lead in average projections.

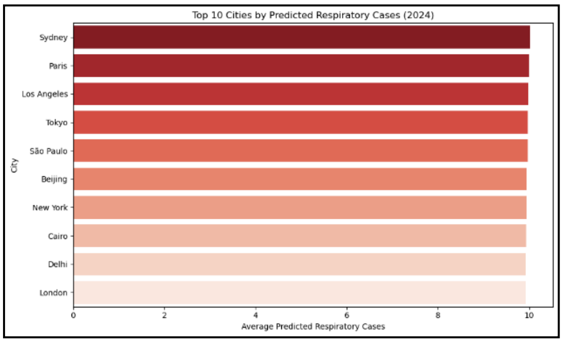


Figure 6 Top 10 Cities by Predicted Respiratory Cases (2024)

Figure 7 highlights cardiovascular predictions, with Tokyo and São Paulo topping the list.

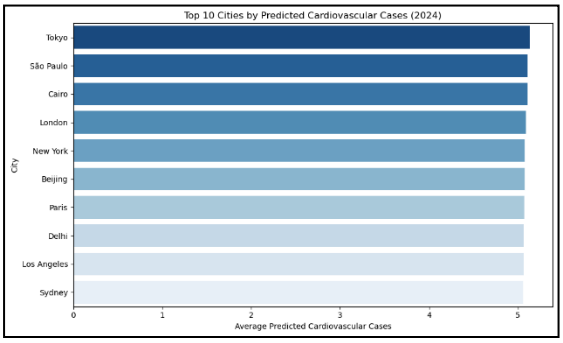


Figure 7 Top 10 Cities by Predicted Cardiovascular Cases (2024)

Figure 8 ranks cities by predicted hospital admissions, where Paris, London, and Los Angeles emerge as most impacted. The trained models were then used to estimate city-wise health outcomes using 2024 air quality data. Key insights include:

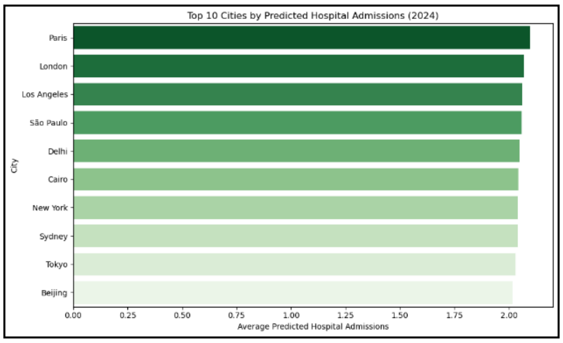


Figure 8 Top 10 Cities by Predicted Hospital Admissions (2024)

- Paris consistently ranked among the top cities in all three health outcomes.
- It had the highest predicted hospital admissions and was second in respiratory case projections.
- The results align with Paris's high average and volatile PM2.5 profile.

7) Pollution-Health Risk Relationship

To better understand the link between pollution and public health, a bubble chart was created showing the relationship between average PM2.5 and total predicted health burden. Paris and São Paulo emerged as critical zones with both high pollution and high health burden, suggesting these cities may require urgent policy attention.

8) Health Burden Index

A composite Health Burden Index (HBI) was developed to summarize the overall impact of air pollution on public health across cities. This index was calculated by normalizing predicted values for respiratory cases, cardiovascular cases, and hospital admissions, then averaging the results to create a unified metric. Cities were ranked based on their HBI scores, with Paris emerging as the highest, followed by Tokyo, São Paulo, and Sydney. These rankings align closely with earlier visualizations and reinforce the notion that cities with high pollution levels also face significant health consequences.

Figure 9 illustrates the Top 10 cities ranked by their Health Burden Index for the year 2024. The values reflect a composite of multiple health indicators, providing a comprehensive snapshot of pollution-related health risks. Higher index scores represent greater cumulative health burdens attributed to air pollution exposure.

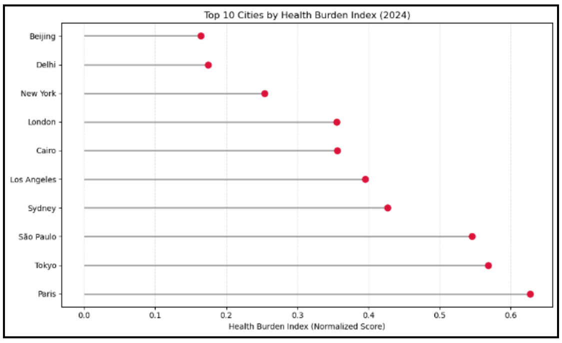


Figure 9 Top 10 Cities by Health Burden Index (2024)

V. DISCUSSION

The results of this project highlight some important patterns in how air pollution may be affecting public health across different cities. By combining basic analysis with machine learning predictions, the findings offer a broader view of which cities are most at risk and what might be contributing to those risks.

Some cities stood out across multiple parts of the analysis. Paris, for example, not only had high average PM2.5 levels but also showed wide fluctuations in daily values. It also ranked first in predicted hospital admissions and had the highest overall Health Burden Index score. Tokyo and São Paulo followed closely, appearing at the top in various health impact categories. These patterns suggest that cities with both consistently high pollution and sudden spikes may face the greatest health challenges.

The Health Burden Index was especially useful for summarizing the predictions across respiratory issues, cardiovascular risks, and hospital admissions. It made it easier to compare cities and identify those with the highest combined health impact due to pollution. Cities that ranked high in the HBI often overlapped with those showing high PM2.5 levels, which supports the idea that fine particulate matter is a key contributor to pollution-related health problems.

That said, there are a few limitations to keep in mind. This analysis only covers data from the year 2024, so it doesn't capture long-term trends or seasonal policies that might have affected air quality. Also, the health prediction models were based on synthetic or generalized data, which doesn't fully reflect city-specific factors like healthcare access, demographics, or local policy. And while some pollutants were weakly correlated with each other or with AQI, this might be due to complex or delayed effects that weren't captured by simple correlation analysis.

Still, the findings point toward clear actionable directions. Cities with high health risk predictions could benefit from stronger air quality monitoring, targeted regulations, or community-level awareness efforts. Focusing on pollutants like PM2.5 and NO_2—both of which ranked high in feature importance—could help reduce health burdens in these urban areas.

Looking ahead, it would be valuable to expand this project by using multi-year data, adding real health case data, or trying out more advanced models to make the predictions even more accurate.

VI. CONCLUSION

This project examined how air pollution impacts public health by analyzing 2024 data from multiple global cities. Using both visual analysis and machine learning, the study identified PM2.5 and NO_2 as key contributors to predicted respiratory, cardiovascular, and hospital-related health outcomes.

Cities like Paris, Tokyo, and São Paulo consistently ranked high in both pollution and health risk, with Paris showing the greatest overall burden. The Health Burden Index offered a clear way to compare cities based on combined health impact.

VII. REFERENCES

[1] B. R. Gurjar, K. Ravindra, and A. S. Nagpure, "Air pollution trends over Indian megacities and their local-to-global implications," *Atmospheric Environment*, vol. 142, pp. 475–495, 2016. doi: 10.1016/j.atmosenv.2016.06.030

[2] M. Greenstone and R. Hanna, "Environmental Regulations, Air and Water Pollution, and Infant Mortality in India," *American Economic Review*, vol. 104, no. 10, pp. 3038–3072, 2014. doi: 10.1257/aer.104.10.3038

[3] M. Fahim, Md. E. Uddin, R. Ahmed, Md. R. Islam, and N. Ahmed, "A Machine Learning Based Analysis Between Climate Change and Human Health: A Correlational Study," in *Proc. 2022 Int. Conf. on Computer and Applications (ICCA)*, IEEE, 2022, pp. 1–6. doi: 10.1109/ICCA56443.2022.10039484

[4] S. M. T. Hasan, "Global Urban Air Quality Index Dataset (2015–2025)," *Kaggle*, [Online]. Available: https://www.kaggle.com/datasets/syedmtalhahasan/global-urban-air-quality-index-dataset-2015-2025

[5] World Health Organization, "Air pollution," *WHO*, 2023. [Online]. Available: https://www.who.int/health-topics/air-pollution