Air Quality Fluctuations, Demographics, and Hospital Admissions in the Punjab Region

Group 10: Aiden Kim, Colby Ogrin, Shantanu Patil and Martina Veit Acosta
CS 5821
Dr. Robert Makin

Abstract

India has the largest pollution-related death toll in the world (Hayward, 2021). We hypothesized that air pollution factors, such as AQI, PM2.5, and PM10, along with demographic attributes like age, contribute to variations in the duration of hospital stays in the region of Punjab, India. A variety of regression models were trained and examined to pick the best one, with Random Forest Regressor slightly outperforming the others. Unfortunately, the trained model cannot predict duration of stay with high certainty. Critical variables such as patient medical history, specific diagnoses, and hospital capacity were missing, and multicollinearity among air quality metrics also affected performance negatively. We found that air-related metrics showed very strong correlations with each other, but ultimately admission age had the greatest importance to the response. Ultimately, our findings underscore the need for comprehensive datasets and integrated approaches in environmental health research.

Introduction

In this paragraph, we will discuss how the chosen factors influence air pollution. AQI (Air Quality Index) is a composite measure providing an overall indicator of air quality and pollution levels, combining multiple pollutants into a single value (AirNow). PM2.5 refers to fine particles that can enter the lungs and bloodstream, causing health issues. PM10 refers to larger particles that can still cause respiratory issues. NO2 is a major pollutant from vehicle emissions and industrial sources, contributing to respiratory problems and ground-level ozone formation. NH3 contributes to particulate matter formation and can harm both human health and the environment. SO2 is a harmful gas from burning fossil fuels and industrial processes, causing lung irritation and contributing to acid rain. CO is a colorless, odorless gas produced by incomplete combustion of fuels. Ozone is a component of smog, which can irritate the respiratory system and exacerbate asthma. Temperature affects the formation of pollutants like ozone and influences their dispersion and concentration. Finally, high humidity can influence the formation and persistence of pollutants like ozone and particulate matter, and affect the health impacts of pollution. For this research, we cleaned pre existing datasets to obtain a new one containing the following columns:

Column Name	Column Description
Date	The date of the pollution measurement and admissions/mortality data.
AQI	Air Quality Index values from the pollution dataset.
PM2.5	Average PM2.5 values from the pollution dataset.
PM10	Average PM10 values from the pollution dataset.
NO2	Average NO2 values from the pollution dataset.
Age_x	The average age of individuals from the mortality data, when the pollution and mortality data overlap by date.
Gender_x	Gender data from the mortality dataset, for matching rows.

Age_y	The average age of individuals admitted to hospitals on the same date and location as the pollution and mortality records.
Gender	The most common gender of admitted individuals for the matching dates.
Duration of stay	The average duration of hospital stays after admission.

Table 1. Final dataset columns

Methods and Results

Once we obtained this new dataset, we performed a few machine learning approaches to analyze the data. Our objective was to predict the duration of hospital stay based on features like AQI, individual pollution levels, and demographics. We used linear regression, random forest regressor, and a tuned version of random forest. The evaluation metrics used were Mean Squared Error (MSE) and R squared.

Our first method was Linear Regression. The MSE indicates how far the predicted values deviate from the actual values on average. The R squared represents the proportion of variance in the target variable explained by the features. Our second method is Random Forest Regressor. This is a tree-based ensemble model that often outperforms linear regression for nonlinear relationships. Results showed improved performance compared to Linear Regression, especially if relationships between features and target are complex. Lastly, we applied cross-validation, which gives a more robust evaluation of model performance by averaging errors across multiple data splits. Regression results, refinements, and relationships can be seen in the data/figures below.

Method	**MSE**	**R-squared**
Linear Regression	4.8457	-0.0450
Random Forest	3.0955	0.3324
Tuned Random Forest	3.1022	0.3310

Table 2. MSE and R-squared results

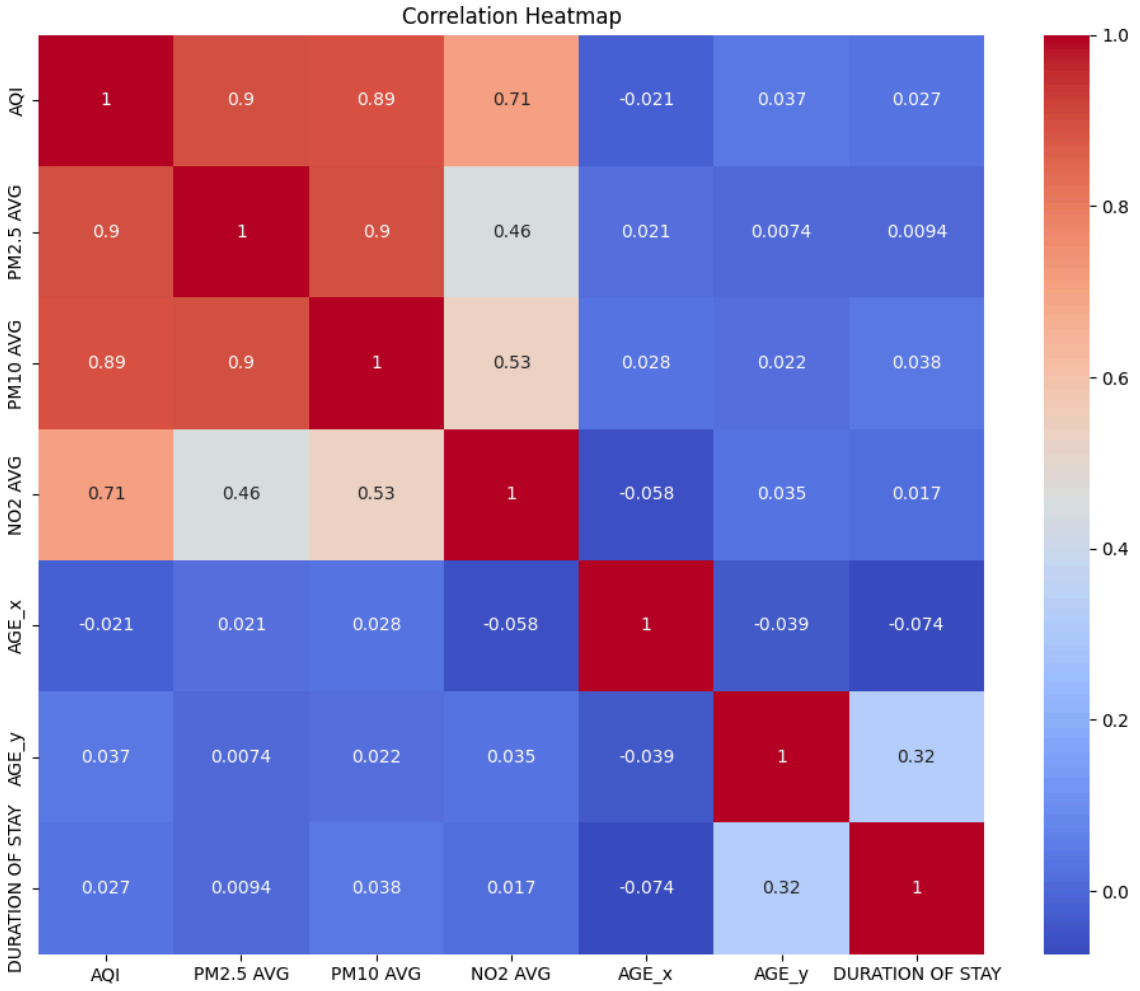

Figure 1. Correlation Heat Map. Shows relationship strength between several attributes.

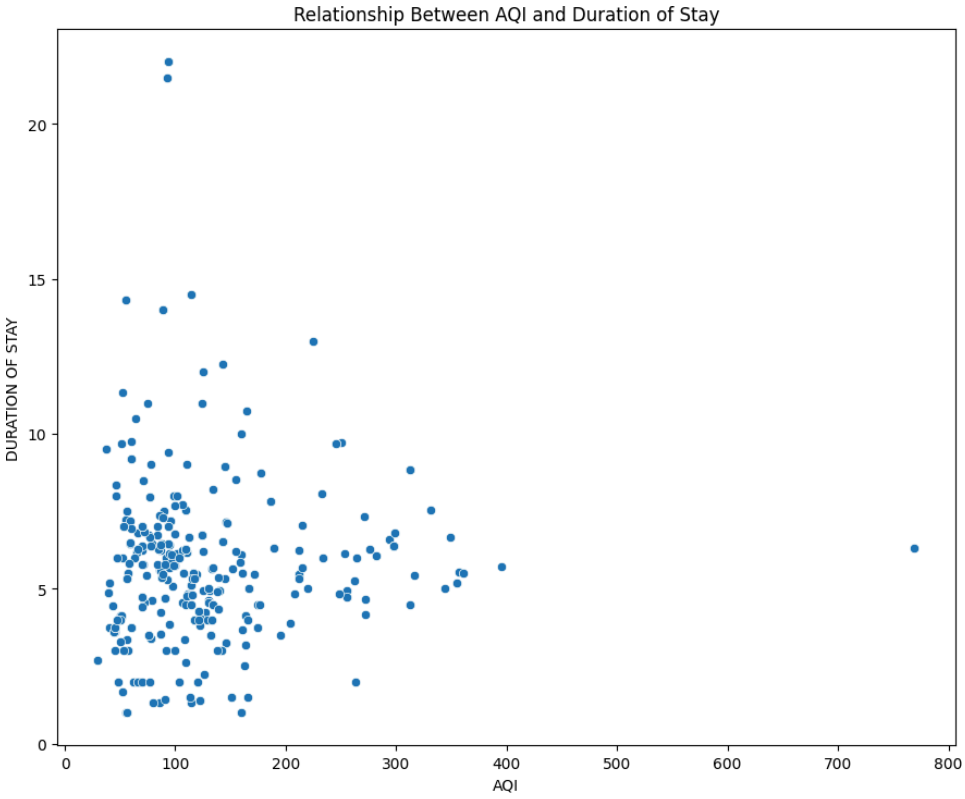

Figure 2. AQI vs. Duration of Stay. 0-50 Good, 51-100 Moderate, 101-150 Unhealthy for Some, 151-200 Unhealthy, 201-300 Very Unhealthy, 301+ Hazardous (AirNow).

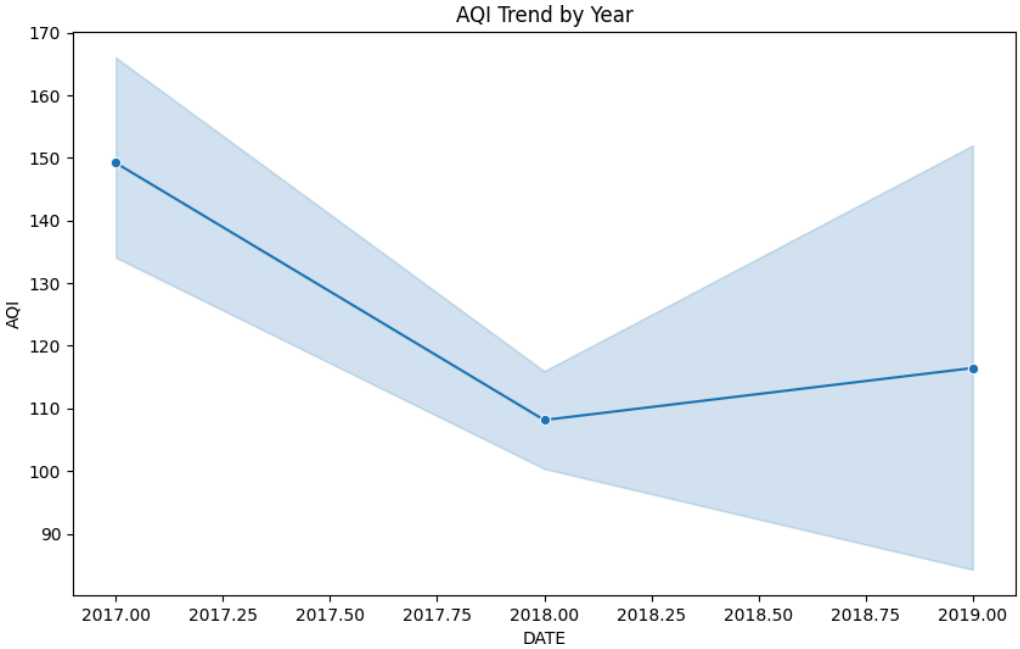

Figure 3. AQI Trend by Year.

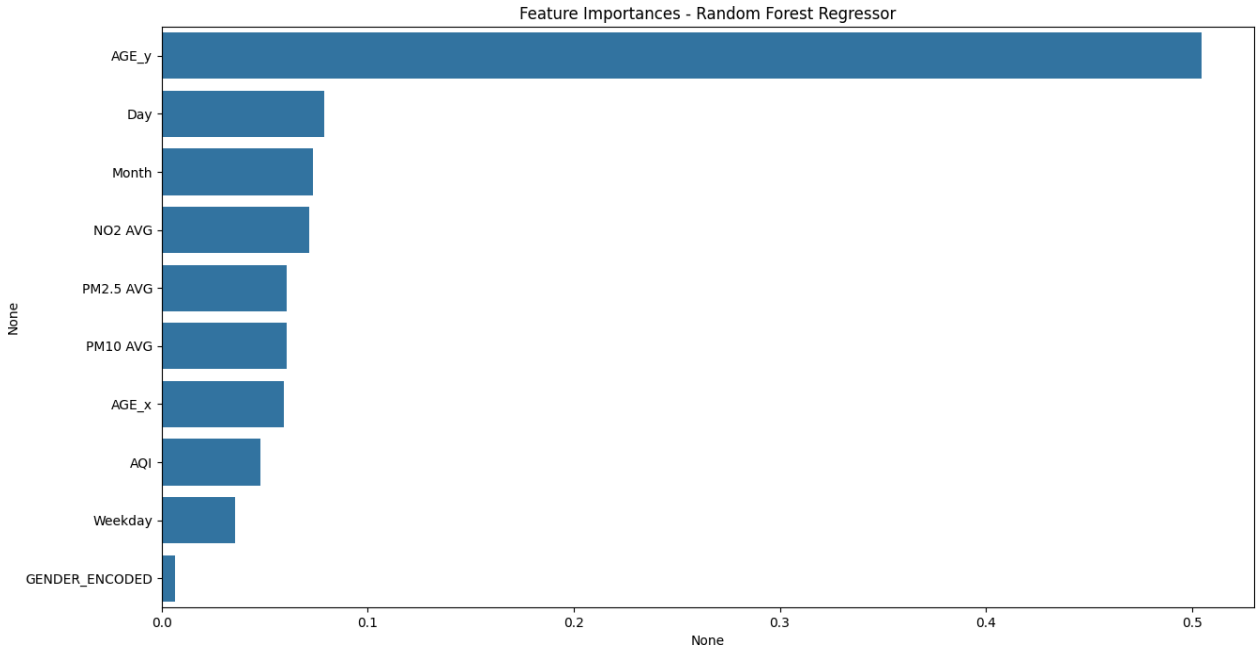

Figure 4. Normalized feature importance values that sum up to 1.

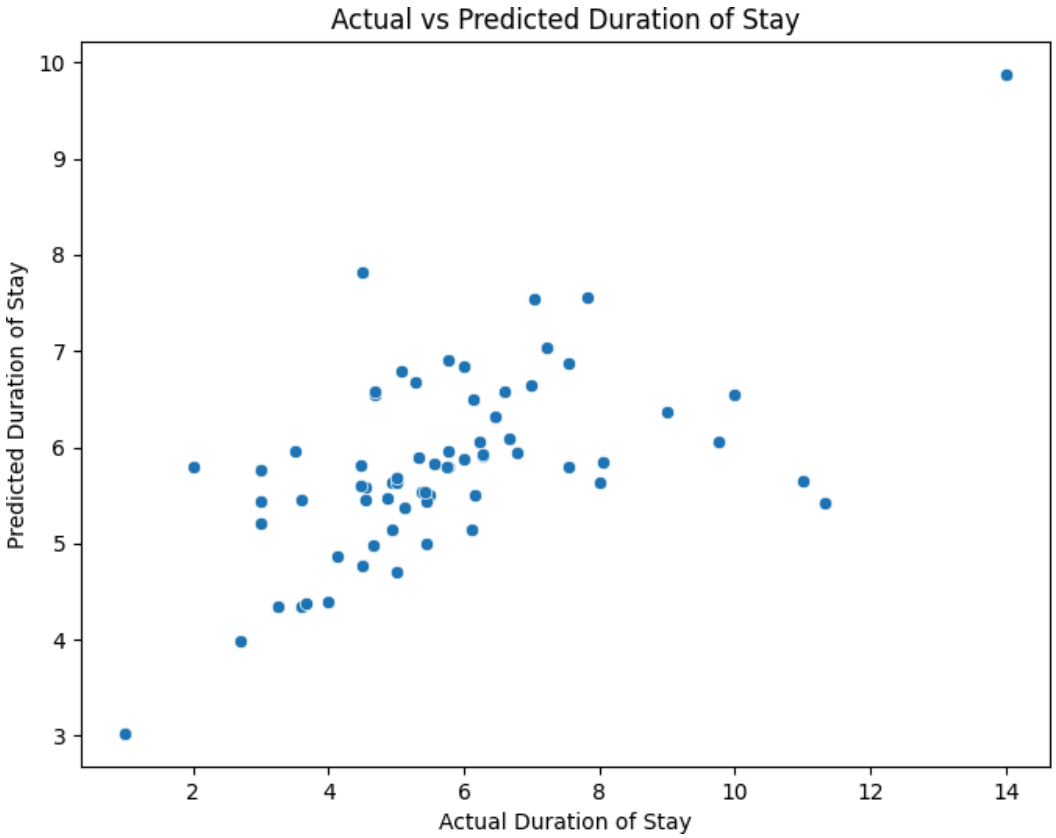

Figure 5. Predicted Durations vs. Actual Duration. Trained using Random Forest.

Key Findings

By analysing the different visualizations and analytical measures, we were able to uncover a few important insights. We found a high correlation among the key variables using correlation heatmap (Figure 1). AQI, PM2.5 AVG, and PM10 AVG described strong positive correlations (coefficients in the order of 0.9) indicating that these pollutants tend to fluctuate in a correlated fashion. Moderate positive correlation was shown for NO2 with other air quality metrics, but the correlation was weaker than the PM correlation. A moderate positive correlation (0.32) between AGE_y and DURATION OF STAY was observed, that is, patients had longer hospital stays the older they are.

The air quality measurements distribution versus Duration of Stay (Figure 2) showed the scatter plot for AQI versus Duration of Stay. A slight positive trend exists but the distance between points shows that AQI by itself is not a very strong predictor for stay duration. Most data points were below 150 AQI (unhealthy for sensitive groups threshold) which reflect mostly moderate to poor air quality conditions.

From 2017 to 2018-2019, the AQI trend analysis (Figure 3) shows that mean AQI decreased approximately 40 units. Therefore, this substantial decrease may reflect the possibility of environmental intervention or policy change in this period.

Unsurprising, feature importance analysis (Figure 4) yielded unexpected results. All other features, including air quality metrics, had surprisingly low importance scores of less than 0.1, while patient age (AGE_y) emerged as the dominant predictor with an importance score of 0.5. Gender had little or no effect; surprisingly, strength, air-related components and temporal features had little importance. Actual vs Predicted duration plot (Figure 5) showed our best model's (Random Forest Regressor, R^2= 0.3324) performance limitations. It shows a positive correlation between predicted and actual values with the significant scatter indicating substantial prediction errors.

Significance

Our results have great consequences for understanding the relationship between air pollution and healthcare outcomes in Punjab. Strong correlations between the air quality metrics appear to suggest the redundancy of pollution measurements and a need for additional diverse environmental indicators. Age has a moderate correlation with stay duration, which demonstrates the role of the demographic factors in healthcare planning.

While air quality metric predictions for hospital stays tend to follow expected patterns, age dominates the equation and disconfirms direct assumptions of pollution's effect on healthcare utilization. This implies that air pollution is still a significant public health issue, although its influence on hospital stay duration was not as simple as expected.

While the improvement in air quality from 2017 to 2019 is positive, it is not well correlated with hospital stay duration, suggesting that healthcare outcomes are at least partially dependent on other factors than environmental conditions. Our best prediction model (R^2 = 0.3324), however, does not perform remarkably well, suggesting that large factors that have

major effects on hospital stays may not be identifiable by the hospital stays we are currently capturing.

Conclusion

In this study we investigated the relationship between fluctuations in air quality, demography and hospital admissions in Punjab and found results contrary to what we initially expected. We had hypothesized that increased air pollution metrics will correlate strongly with longer hospital stays, but we found that our analysis showed a more complicated correlation based on demographic factors, primarily patient age.

However, there were several limitations in our study. Model results may have been affected by multicollinearity among air quality metrics. Our dataset was also missing variables that were extremely crucial including patient medical history, specific diagnosis information and hospital capacity metrics. The temporal misalignment and granularity issue of available datasets made the original plan of including GDP data infeasible.

We suggest for future research that mortality rates could be used as an alternative response variable and may provide different insights from those using air quality as a surrogate for health impacts. By a modified study design based on regional averages, macroeconomic factors such as GDP could better be incorporated. Different age groups or seasons might separate out more nuanced patterns. In addition, non linear relationships and time series analysis would provide better insights to the complexity of environmental health impacts.

These findings underscore the importance of integrated healthcare planning of which the environment and demography are unified. While air quality improved over the study period, the health impacts of air appear more complicated than correlations between air quality and health would suggest. Future studies should aim at collecting additional comprehensive data to help understanding and predicting healthcare outcomes of regions with severe air quality.

References

Datasets:

1. Bollepalli, Sandeep Chandra, et al. “An Optimized Machine Learning Model Accurately Predicts In-Hospital Outcomes at Admission to a Cardiac Unit.” Diagnostics, vol. 12, no. 2, 1 Feb. 2022, p. 241, www.mdpi.com/2075-4418/12/2/241, https://doi.org/10.3390/diagnostics12020241. Accessed 30 Nov. 2024.
2. “Hero DMC Heart Institute.” Herodmc.com, 2022, www.herodmc.com/. Accessed 30 Nov. 2024.

Others:

1. Hayward, Ed. “The Human Toll of Air Pollution in India.” www.bc.edu, Jan. 2021, www.bc.edu/bc-web/bcnews/nation-world-society/international/air-pollution-in-inda.html
2. “Air Quality Index (AQI) Basics.” AirNow, https://www.airnow.gov/aqi/aqi-basics/. Accessed 4 Dec. 2024.