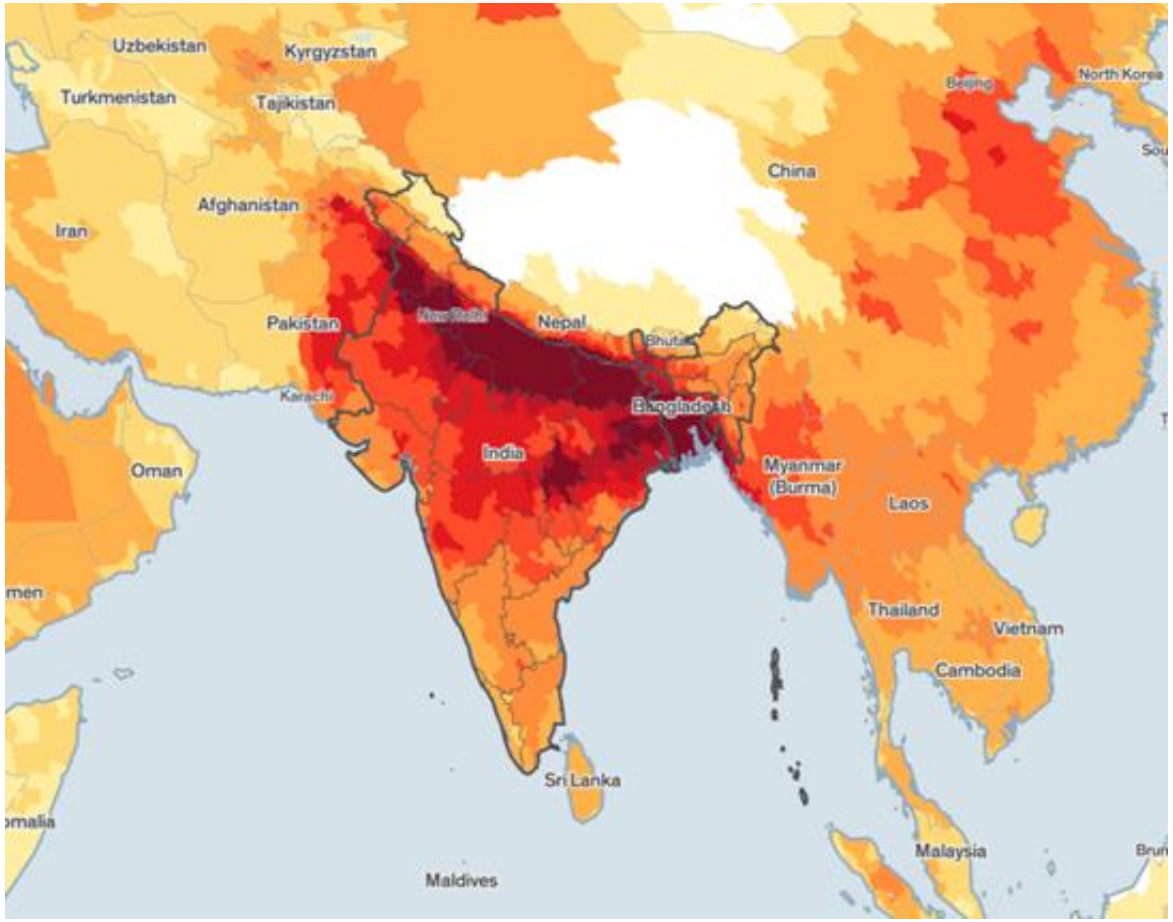
Group 10: Aiden Kim, Colby Ogrin, Shantanu Patil and Martina Veit Acosta

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



Introduction to AQI

- Air Quality Index combines multiple pollutants into an overall indicator value.
- *PM2.5:* particles fine enough to enter bloodstream via lungs
- *PM10:* large particles that can still cause respiratory issues
- *NO2:* major pollutant from vehicle/industrial emissions; promotes ozone formation

AQI Basics for Ozone and Particle Pollution

Daily AQI Color	Levels of Concern	Values of Index	Description of Air Quality
Green	Good	0 to 50	Air quality is satisfactory, and air pollution poses little or no risk.
Yellow	Moderate	51 to 100	Air quality is acceptable. However, there may be a risk for some people, particularly those who are unusually sensitive to air pollution.
Orange	Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is less likely to be affected.
Red	Unhealthy	151 to 200	Some members of the general public may experience health effects; members of sensitive groups may experience more serious health effects.
Purple	Very Unhealthy	201 to 300	Health alert: The risk of health effects is increased for everyone.
Maroon	Hazardous	301 and higher	Health warning of emergency conditions: everyone is more likely to be affected.

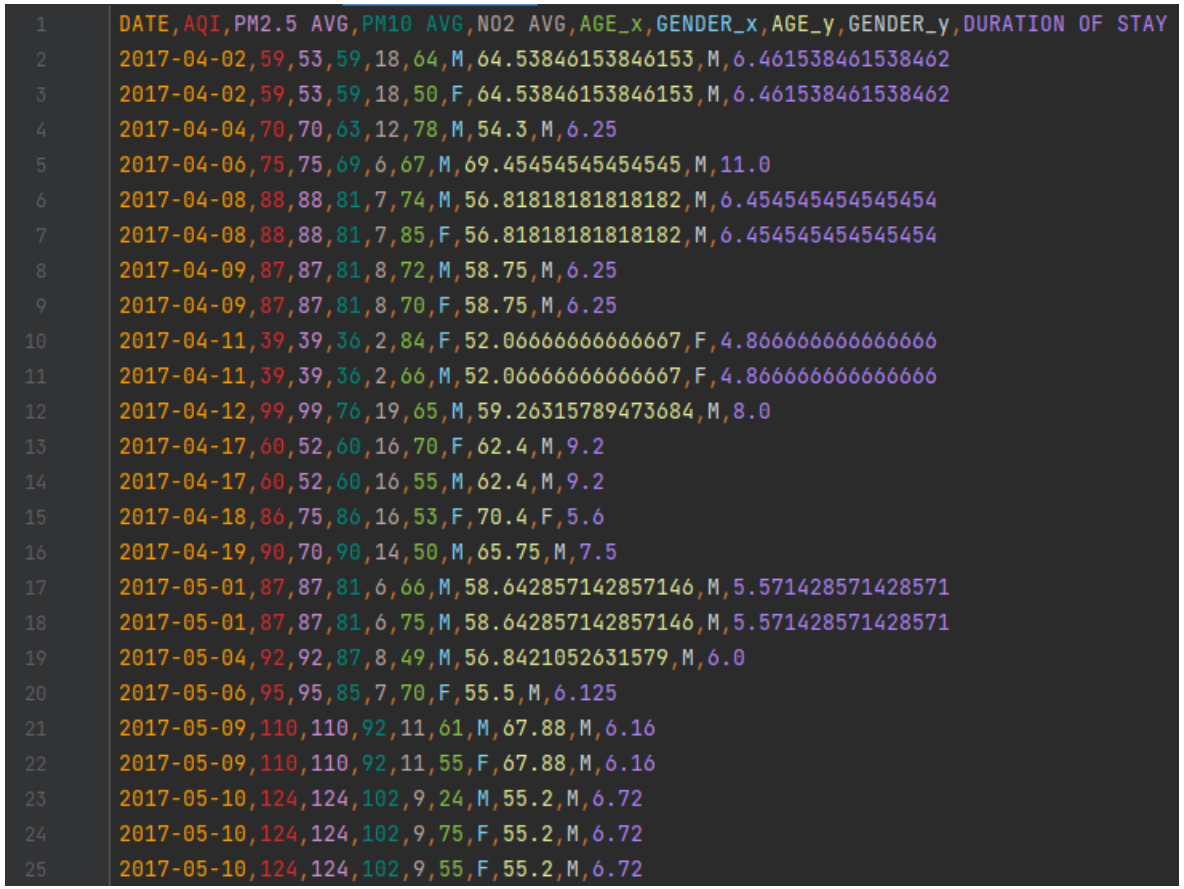
Introduction to AQI (Continued)

- *NH3:* contributes to particulate matter formation, harming human/environmental health
- *SO2:* gas emission that causes lung irritation and acid rain
- *CO:* colorless, odorless gas from incomplete fuel combustion
- *Ozone:* smog component; asthma aggravation

Hypothesis

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 for respiratory diseases in Punjab, India.

Hospital in Ludhiana, Punjab from where we got our data



```
DATE,AQI,PM2.5 AVG,PM10 AVG,NO2 AVG,AGE_x,GENDER_x,AGE_y,GENDER_y,DURATION OF STAY
2017-04-02,59,53,59,18,64,M,64.53846153846153,M,6.461538461538462
2017-04-02,59,53,59,18,50,F,64.53846153846153,M,6.461538461538462
2017-04-04,70,70,63,12,78,M,54.3,M,6.25
2017-04-06,75,75,69,6,67,M,69.45454545454545,M,11.0
2017-04-08,88,88,81,7,74,M,56.81818181818182,M,6.454545454545454
2017-04-08,88,88,81,7,85,F,56.81818181818182,M,6.454545454545454
2017-04-09,87,87,81,8,72,M,58.75,M,6.25
2017-04-09,87,87,81,8,70,F,58.75,M,6.25
2017-04-11,39,39,36,2,84,F,52.06666666666667,F,4.866666666666666
2017-04-11,39,39,36,2,66,M,52.06666666666667,F,4.866666666666666
2017-04-12,99,99,76,19,65,M,59.26315789473684,M,8.0
2017-04-17,60,52,60,16,70,F,62.4,M,9.2
2017-04-17,60,52,60,16,55,M,62.4,M,9.2
2017-04-18,86,75,86,16,53,F,70.4,F,5.6
2017-04-19,90,70,90,14,50,M,65.75,M,7.5
2017-05-01,87,87,81,6,66,M,58.642857142857146,M,5.571428571428571
2017-05-01,87,87,81,6,75,M,58.642857142857146,M,5.571428571428571
2017-05-04,92,92,87,8,49,M,56.8421052631579,M,6.0
2017-05-06,95,95,85,7,70,F,55.5,M,6.125
2017-05-09,110,110,92,11,61,M,67.88,M,6.16
2017-05-09,110,110,92,11,55,F,67.88,M,6.16
2017-05-10,124,124,102,9,24,M,55.2,M,6.72
2017-05-10,124,124,102,9,75,F,55.2,M,6.72
2017-05-10,124,124,102,9,55,F,55.2,M,6.72
```

Methods

- Preexisting datasets (pollution, admissions, mortality) cleaned and combined by date.
- Linear Regression
- Random Forest Regressor
- Tuned Random Forest Regressor
- Cross-Validation

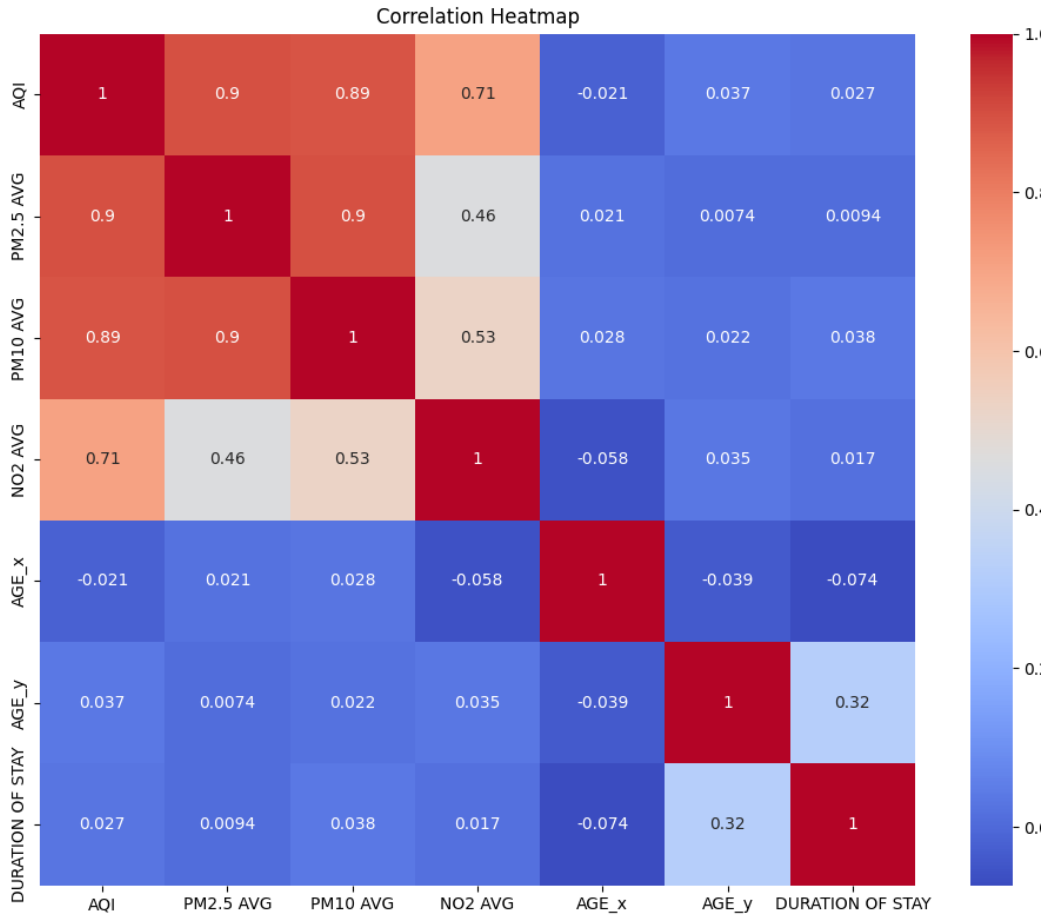
Results

Method	MSE	R^2
Linear Regression	4.8457	-0.0450
Random Forest	3.0955	0.3324
Tuned Random Forest	3.1022	0.3310

Table 1. MSE and R-squared results

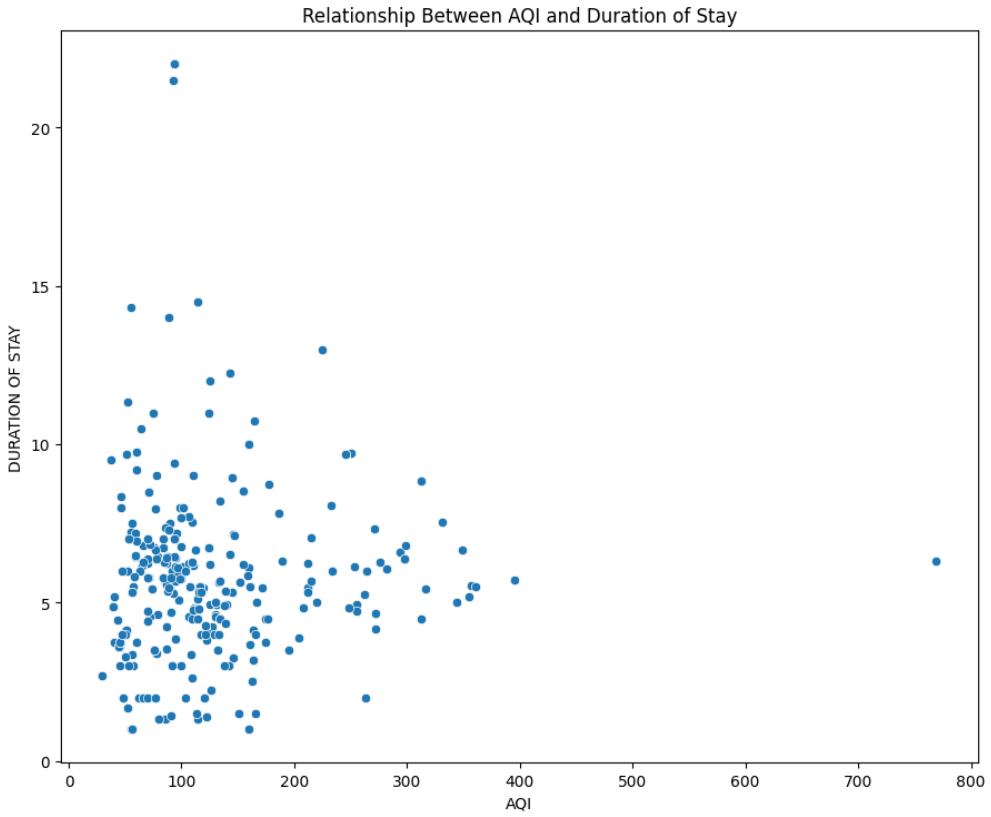
Correlation Heat Map

- *High Positive Correlations:* AQI, PM2.5 AVG, and PM10 AVG (~0.9); potential multicollinearity issues
- *Notable Positive Correlations:* NO2 + other air quality; AGE_y and DURATION OF STAY
- *Insignificant Correlations:* AGE_x and AGE_y with other variables

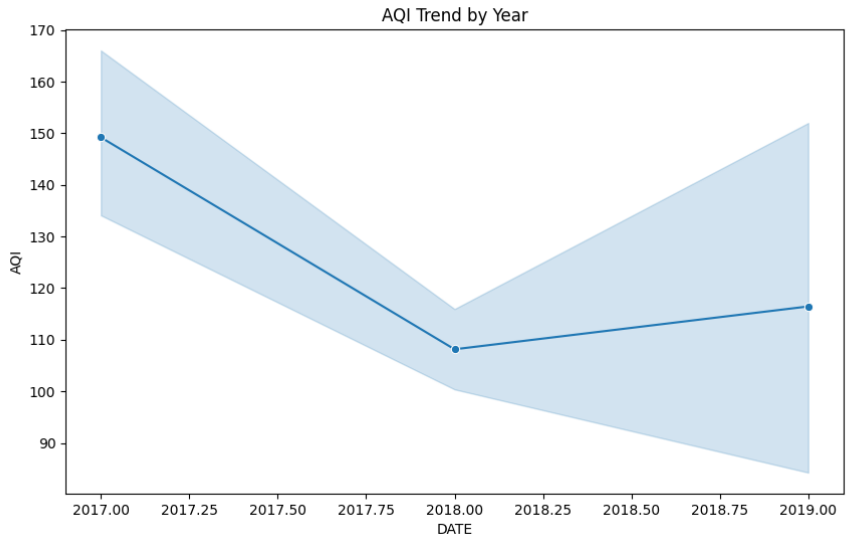


AQI vs. Duration of Stay

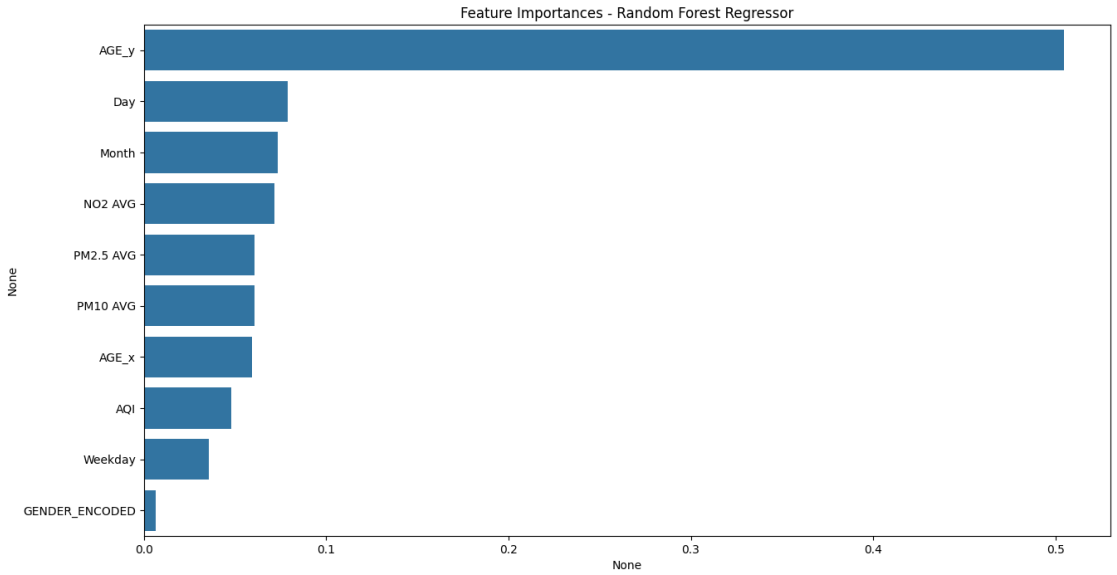
- 0-150 Ranges from Good to Unhealthy for Some
- 200+ Ranges from Unhealthy to Hazardous
- loose distribution indicates a non-linear relationship, or attribute is dominated by others



AQI Trend by Year



- *Points:* mean AQI for each year
- *Shading:* AQI Range
- 2017 roughly forty units higher compared to 2018 & 2019
- May explain clustering of points under 150 AQI in previous graph.

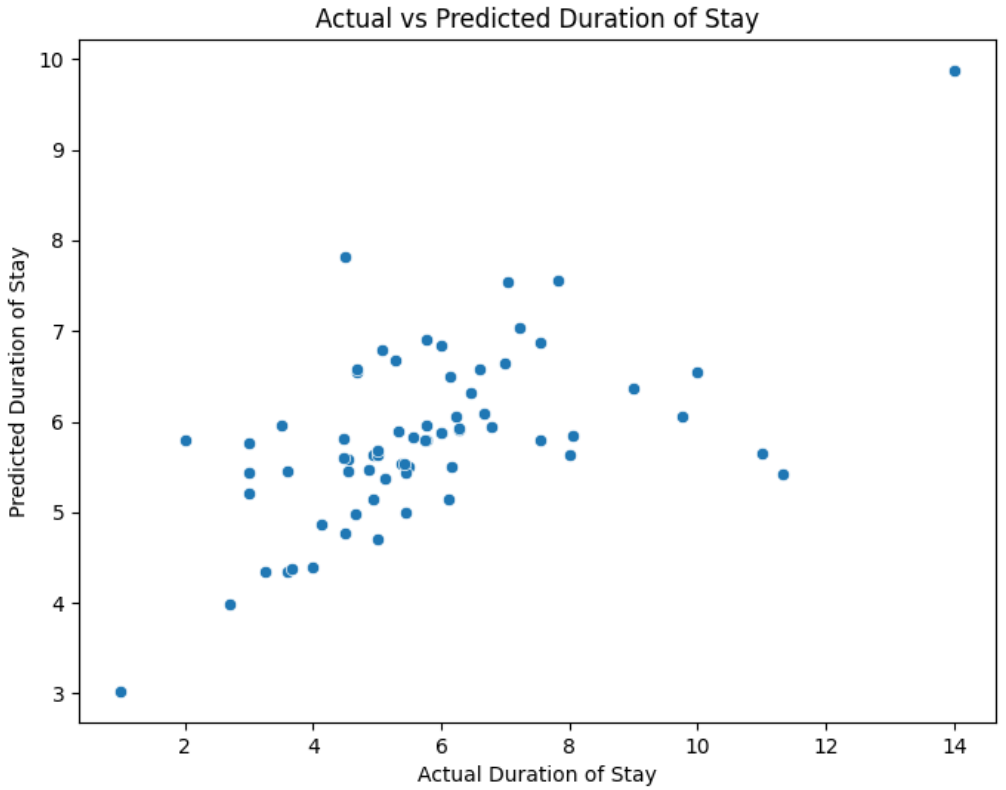


Feature Importances

- Normalized feature importance values sum up to 1
- AGE_y highest by far with 0.5
- Time/air-related components under 0.1
- Gender has almost no effect
- Further explains AQI vs. Duration anomalies

Predicted Durations vs. Actual Durations

- Random Forest Regression selected as best model
- Trained data used to predict the duration of patient stay
- Graph shows loose dispersion of points with many outliers
- Cannot predict DURATION OF STAY with high certainty



Key Statistical Findings 📊

Strong Air Quality Correlations

AQI, PM2.5, PM10 show 0.9 correlation

Age & Hospital Stay: 0.32 correlation

NO2 moderately linked to other pollutants

Model Performance

Random Forest: Best performer (R^2 = 0.3324)

Linear Regression: Poor fit (R^2 = -0.0450)

Feature Importance Hierarchy

Patient Age: Dominant (50% importance)

Air Quality: Surprisingly low (<10%)

Gender: Minimal impact

Environmental Trends

40-point drop in Mean AQI (2017-2019)

Most readings below 150 AQI threshold

Key Insights from Analysis

- **Age Impact**
  - Strongest predictor of stay duration
  - More influential than environmental factors
  - Suggests age-specific healthcare planning needed
- **Complex Relationships**
  - Non-linear patterns detected
  - Multiple influencing factors
  - Environmental impacts more subtle than expected
- **Data Patterns**
  - Clustered AQI measurements
  - Significant prediction scatter
  - Strong metric interdependencies
- **Model Limitations**
  - Only 33% variance explained
  - Poor linear regression fit
  - Potential multicollinearity issues

Future Directions 

Study Limitations

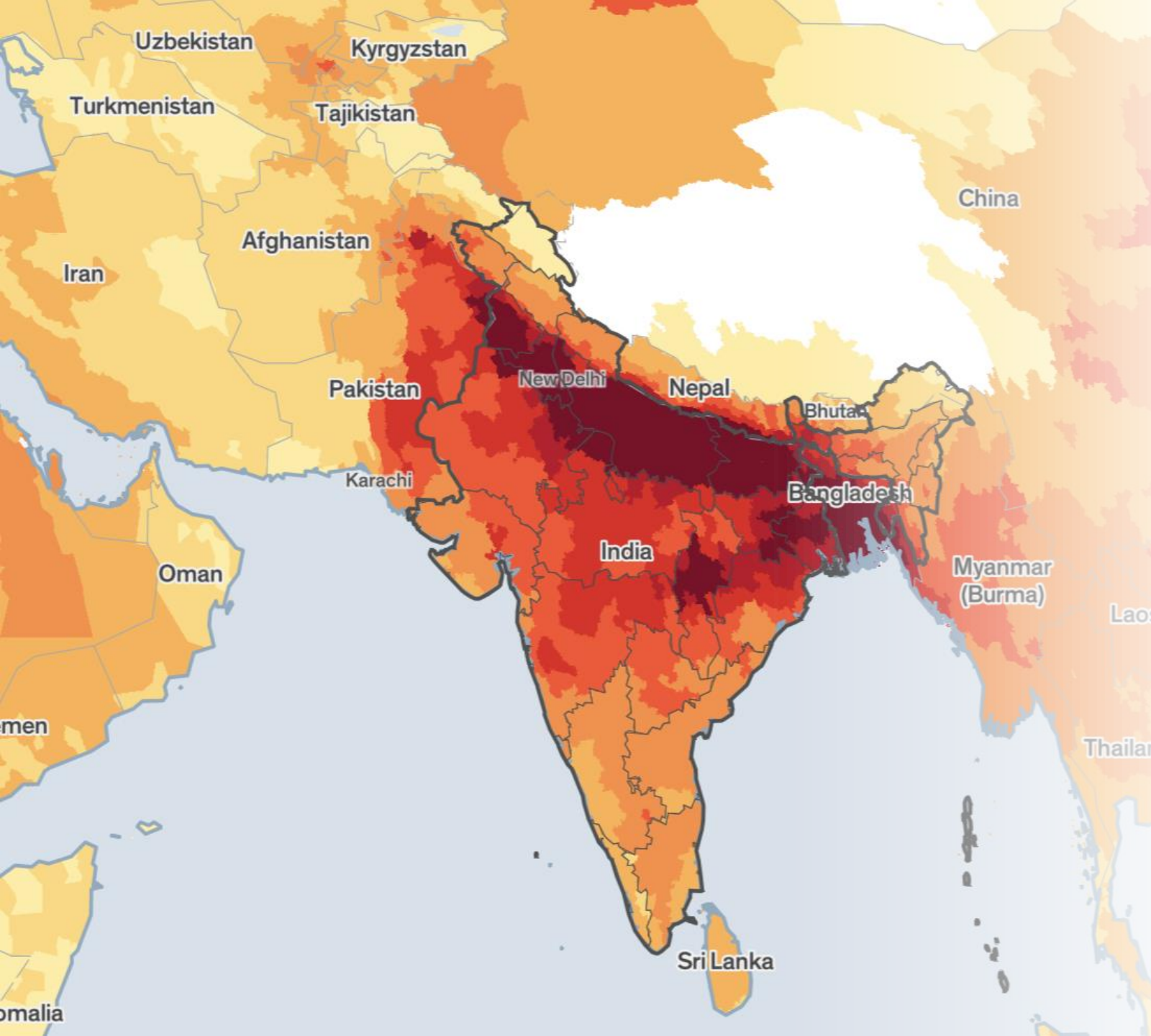
- Missing patient medical history
- No socioeconomic data
- Limited temporal scope
- Overlapping air quality metrics

Proposed Research

- Mortality rate analysis
- Regional economic integration
- Age-group specific studies
- Time-series investigation

Required Data Improvements

- Patient diagnosis details
- Hospital capacity metrics
- Socioeconomic indicators
- Extended time period.



Conclusion

- **Critical Air Quality Zone**
  - Punjab region shows severe pollution (dark red)
  - Part of larger North Indian air quality crisis
  - Chronic exposure affects entire population
- **Research Context**
  - High baseline pollution across region
  - Limited variation in air quality exposure
  - May explain complex relationship with hospital stays
- **Health Impact Complexity**
  - Entire population chronically exposed
  - Age emerged as stronger predictor than AQI
  - Suggests long-term rather than acute effects
- **Research Implications**
  - Need for mortality rate studies
  - Long-term health impact assessment crucial
  - Regional approach to solutions needed

Sources

Datasets

- 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.
- “Hero DMC Heart Institute.” Herodmc.com, 2022, www.herodmc.com/. Accessed 30 Nov. 2024.

Others

- 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
- “Air Quality Index (AQI) Basics.” AirNow, https://www.airnow.gov/aqi/aqi-basics/. Accessed 4 Dec. 2024.