Air Quality Intelligence for NGOs Strategic Insights and Actions

Kari Hattabaugh, Jessica Davis, Manali Sudhir Chavaj, Daniel Nagy

The George Washington University

Course number: ISTM 4217 & 6217

Faith Bradley

November 11, 2025

Introduction & Objectives

Background

CleanAir Futures International (CAFI) is a global nonprofit organization established in 2014 in response to the growing need for data-driven, science-based approaches to addressing air pollution and its impacts on human health and the environment. Founded by environmental scientists, public health experts, and policy professionals, CAFI was created to bridge the gap between environmental monitoring, technological innovation, and sustainable development.

The organization's mission is to improve air quality and reduce pollution-related health risks through the integration of advanced technologies, cross-sector partnerships, and evidence-based policy solutions. CAFI focuses on using Internet of Things (IoT)-enabled sensor networks, artificial intelligence (AI), and predictive analytics to collect and analyze real-time environmental data from cities and industrial regions worldwide. This data supports governments, businesses, and communities in making informed decisions about pollution control and public health protection.

CAFI's overarching goals are to identify and mitigate environmental risks, promote equitable access to clean air, and strengthen global capacity for sustainable development. Through its research and partnerships, the organization seeks to demonstrate how technology and collaboration can drive measurable improvements in air quality, reduce health disparities, and advance long-term environmental resilience across the globe.

Air quality monitoring has evolved significantly with the use of Internet of Things (IoT) technology. Traditional monitoring methods relied on manual sampling and delayed reporting, which limited their ability to provide real-time data or predict harmful exposure levels. IoT-enabled sensors now make it possible to continuously collect detailed information on pollutants across a range of environments.(Ramadan, Ali, Khoo, Alkhedher, & Alherbawi, 2024) This approach creates a more accurate and data-driven understanding of air quality conditions.

Research in industry environments such as chrome plating facilities show how IoT systems are transforming pollutant management. By combining sensor networks with artificial intelligence models such as Long Short-Term Memory and Random Forest, these systems can forecast changes in pollutant levels with a high degree of accuracy. When elevated levels are predicted, automated responses such as activating ventilation systems or modifying operations can prevent health risks before they occur.

For CleanAir Futures International, these technological advancements highlight the value of using IoT data to connect environmental research, public health, and operational planning. Continuous monitoring allows the organization to identify patterns, evaluate community health impacts, and allocate resources to areas where they can achieve the greatest impact. This data also supports research on the link between air quality and mortality, helping decision-makers move from reactive to preventative measures

IoT sensors represent a major step forward in sustainability and health research, providing the knowledge and tools needed to build cleaner and healthier communities.

According to the U.S. Environmental Protection Agency, air pollution affects every part of life including people, animals, and the environment. Exposure to polluted air has been linked to respiratory illnesses, decreased physical activity, and premature mortality. Beyond human health, air pollutants contribute to the degradation of natural resources by contaminating lakes, rivers, streams, and by accelerating the deterioration of infrastructure such as buildings and roads. Also, air pollution weakens the ozone layer and intensifies climate change, disrupting the environmental balance that sustains life on Earth.(U.S. Environmental Protection Agency, n.d.)

There are six air pollutants that are globally recognized and as the most commonly measured and regulated substances that affect air quality, public health and the environment. The six commonly monitored air pollutants, particulate matter (PM_{10} and $PM_{2.5),}$), ground-level ozone (O_3), carbon monoxide (CO), sulfur dioxide (SO_2), nitrogen dioxide (NO_2). These standards became a model for how air quality could be measured and managed using consistent data and scientific evidence.

CleanAir Futures International applies IoT based air quality data to link environmental change directly to community health outcomes. The organization's monitoring networks identify the countries most affected by elevated levels of key air pollutants. Based on these insights, CleanAir Futures is conducting research on health remediations strategies tailored to each region, focusing on early detection of population-related illnesses, public health education, and preventative interventions in high-risk populations. Collaboration with national governments and health ministries ensures these recommendations can be integrated into existing public health frameworks. The real-time nature of IoT data enables rapid responses to pollution spikes, helping reduce exposure and related health complications. It is believed over time, this approach supports measurable declines in respiratory illness and improved community well-being. Through this intersection of technology, research, and collaboration, CleanAir Futures International demonstrates how NGOs can turn environmental insights into sustainable health and policy solutions.

Problem Definition

The research gap addressed by this project centers on the limited integration of environmental and socioeconomic data to guide NGO decision making. While air quality and health outcomes are widely studied, few analyses combine these indicators with operational and economic performance metrics. For organizations like CleanAir Futures International (CAFI), this disconnection limits the ability to assess where investments can achieve both social and environmental returns. The project bridges this gap by merging IoT-enabled air quality data with health and social economic datasets from World Health Organization (WHO).. This integration provides a multidimensional understanding of how pollutants influence mortality and economic productivity, supporting evidence-based planning.

Balancing operational efficiency with human and environmental health remains a persistent challenge for global NGOs. Efficient operations sustain impact, but cost-driven decisions risk neglecting environmental justice and community wellbeing. CAFI’s IoT sensor network and collaborators provide real-time insights into environmental conditions, enabling smarter resource allocation that aligns with both sustainability and health priorities. By using predictive analytics to anticipate risks, CAFI ensures that its operational strategies minimize harm, protect human life, and strengthen long-term environmental resilience.

Objectives & Research Questions

This project aims to generate new evidence on how air pollution contributes to variation in mortality and chronic disease across countries. The analysis will integrate sensor data with global health and economic indicators to inform investment decisions.

Measurable Objective:

1) Quantify relationships between pollutants, mortality and chronic disease rates using IoT generated air quality data and national health indicators.
2) Identify geographic areas or countries where sustainability and public health outcomes align most effectively with organizational efficiency.
3) Translate research findings into policy and operational guidance that informs CAFI's future health remediation and environmental initiatives in priority countries.

The results will fill critical knowledge gaps and support CAFI's mission to advance evidence-based strategies for cleaner air, healthier populations, and sustainable development.

II. Data Management & Preparation

This study integrates multiple datasets to analyze how air quality and environmental factors influence public health. Emphasis was placed on data integrity, reproducibility, and efficient structure for later modeling.

A. Data Sources Overview

Three complementary datasets were combined to form a multi-level analytical framework linking local air pollution with national health indicators.

Dataset	Records	Variables	Purpose
Air Quality & Health	88,489	12	City-level IoT-style readings on $PM_{2.5}$, PM_{10}, NO_2, O_3, AQI, temperature, humidity, and hospital admissions.
Demographic & Health (WHO)	594	54	National health indicators including life expectancy, mortality, sanitation, and disease metrics.
AQI & Coordinates	16,695	14	City-level AQI data with latitude/longitude, enabling spatial mapping and linkage between datasets.

The Air Quality dataset captures short-term pollution effects, the WHO dataset provides population-level health context, and the AQI–Geospatial dataset connects both through spatial references.

B. Data Ingestion and Platform Use

All data was processed using Databricks for scalability and collaboration.

Each dataset was stored in the Databricks File System (DBFS) and imported via Spark's read.csv() function with schema inference:

```
df = spark.read.csv("/FileStore/tables/air_quality_health_dataset.csv", header=True,
inferSchema=True)
display(df)
```

Databricks' shared notebooks enabled multi-user access, live visualization, and version control for reproducibility. Spark's distributed engine supported large-scale transformations efficiently.

Google Colab and Jupyter Notebooks were used for validation, ensuring consistency across environments.

C. Data Cleaning and Transformation

A structured pipeline was applied to ensure analytical readiness:

- Missing Values: Imputed using mean or median; variables with >90% missing data were excluded.
- Normalization: Pollutant measures standardized via z-scores for cross-city comparison.
- Standardization: Dates converted to ISO format; city/country names trimmed and lowercased for consistent joins.
- Geospatial Verification: Latitude–longitude pairs validated against national boundaries.
- Filtering: Retained only relevant pollution, weather, and health variables.
- Quality Checks: Outliers capped using WHO limits; all transformations logged in Databricks for traceability.

D. Integration Schema

A relational star-schema was used to unify data around spatial and temporal dimensions.

The Air Quality table (fact data) connects to AQI–Geospatial data via standardized city names and to WHO Health Indicators via country and year.

This design allows flexible queries by city, country, or year and supports scalable analyses like AQI vs. Life Expectancy or Pollution vs. Hospital Admissions.

III. Data Processing & Feature Engineering

Following integration, datasets were refined and enhanced to support modeling.

A. Merging and Joining Logic

Joins were performed in two stages:

1. City-level join between Air Quality and AQI–Geospatial data (city_std).
2. Country-level join between the merged city data and WHO indicators (Country).

Field types were validated, duplicates checked, and record counts compared before and after merging to confirm consistency. Minor naming mismatches were corrected through manual cleaning and fuzzy matching.

B. Feature Engineering

New derived features were created to strengthen model interpretability:

- Pollution Index: Mean of normalized $PM_{2.5}$, PM_{10}, NO_2, and O_3 values.
- Health Impact Ratio: Hospital admissions normalized by population density.
- Environmental Stress Score: Combines AQI, temperature, and humidity extremes.
- Aggregated Indicators: Country-level yearly averages for pollution and health variables.

These features support regression and classification tasks examining pollution’s influence on health outcomes.

C. Validation

Final validation ensured reliability:

- Descriptive statistics confirmed plausible ranges post-cleaning.

- Correlation plots validated logical relationships between pollution and health indicators.
- Outliers were visualized and removed when exceeding 3σ thresholds.
- Null audits confirmed completeness after imputation.

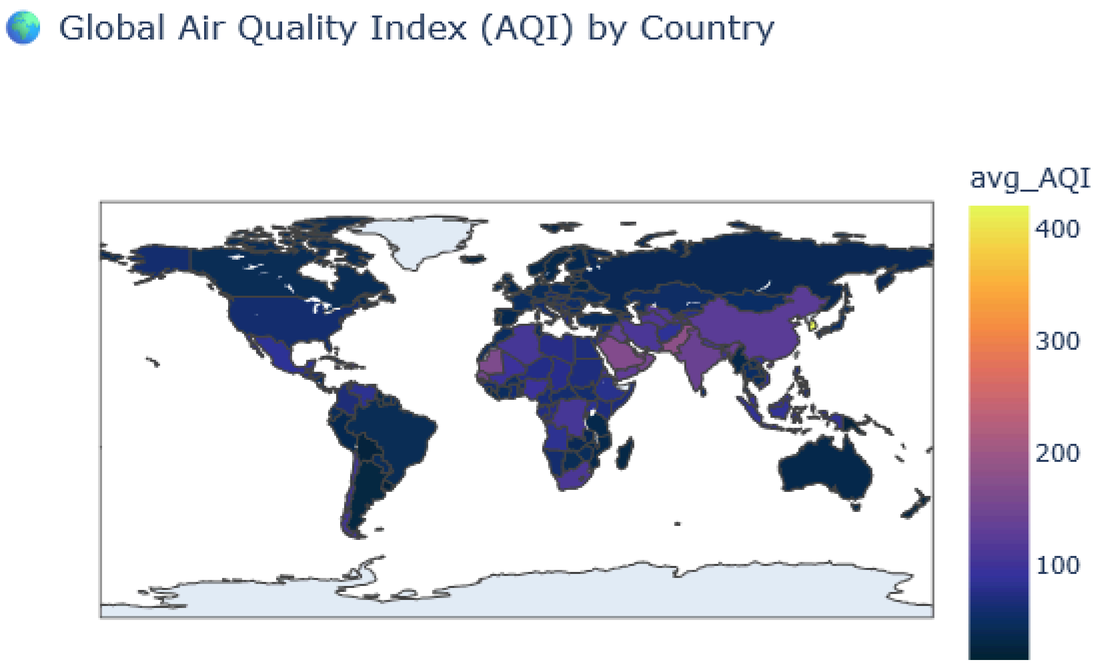
The resulting dataset is clean, consistent, and ready for analytical modeling.

Exploratory Data Analysis

The exploratory data analysis (EDA) phase provides a comprehensive overview of air quality trends, health correlations, and model performance across multiple countries. This state was designed to uncover underlying relationships between AQI values, pollutant exposure, and population health indicators while assessing the reliability of our predictive modeling techniques. A combination of visual analytics were used where CAFI identified key geographic and health patterns that reveal the global burden of air pollution. The EDA results show how regions with high AQI levels also experience higher mortality rates and lower life expectancy,These findings form the analytical foundation for CADI's ongoing work to integrate IoT based environmental data with global health metrics, enabling more targeted and better informed health interventions.

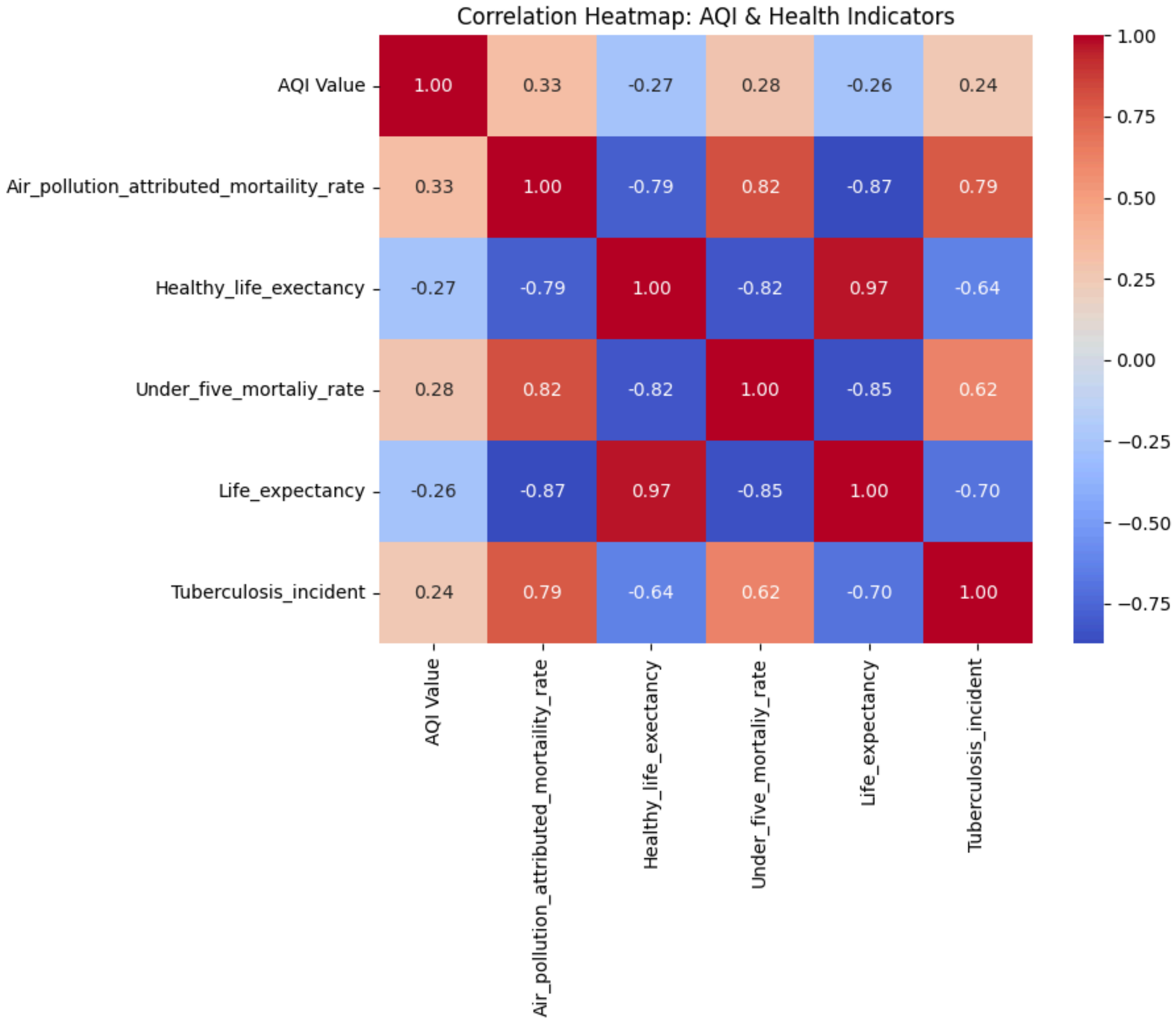
Visual Analytics

The world map visualizes AQI averages geographically. The Global Air Quality Index (AQI) map highlights severe regional disparities in pollution exposure. Countries such as the Republic of Korea, Bahrain, Pakistan, Mauritania, and Qatar consistently record very unhealthy or hazardous air quality, while North America and parts of Europe maintain cleaner conditions.



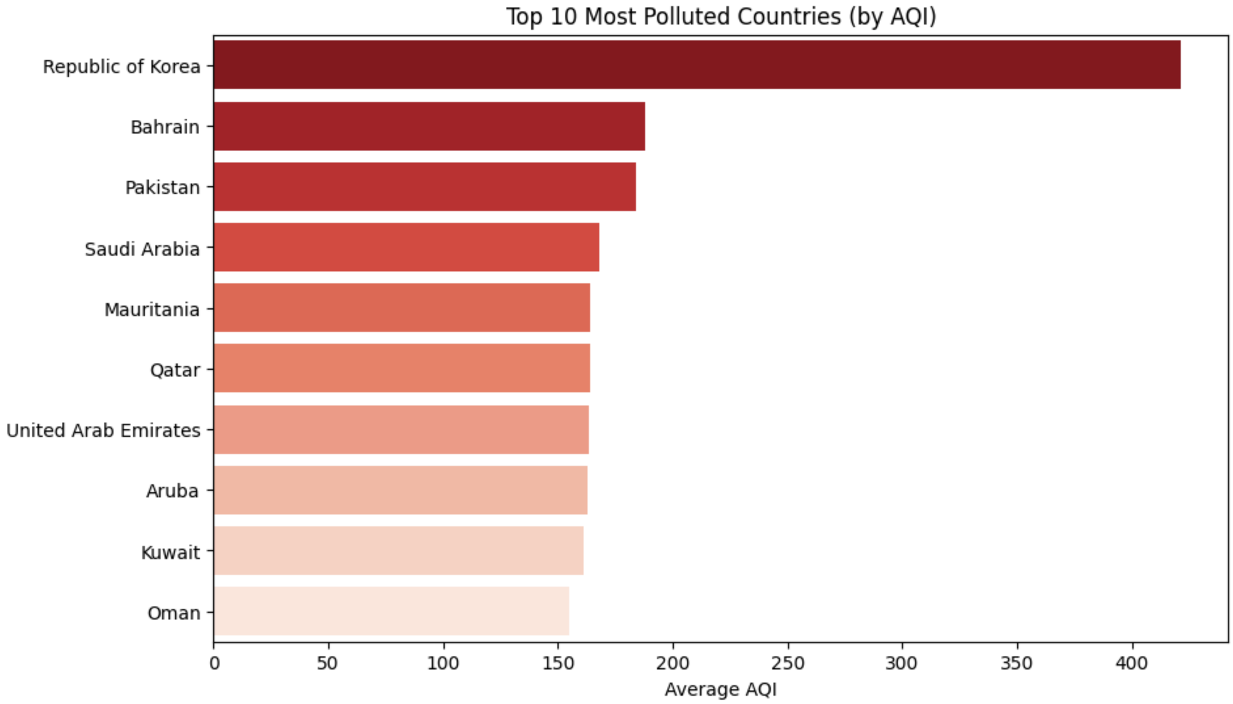
Correlation HeatMap: AQI and Health Indicators

The heatmap illustrates the relationship between AQI values and health metrics. A positive correlection between AQI and the air pollution attributed mortality rate (0.33) confirms that poorer air quality is associated with higher mortality. Conversely, negative correlations between AQI and healthy life expectancy (-0.27) and life expectancy (-0.26) suggest that increased pollution levels reduce longevity and quality of life. The strong inverse relationships between mortality rates and life expectancy (-0.87) further validate the significant health burden linked to poor air quality.



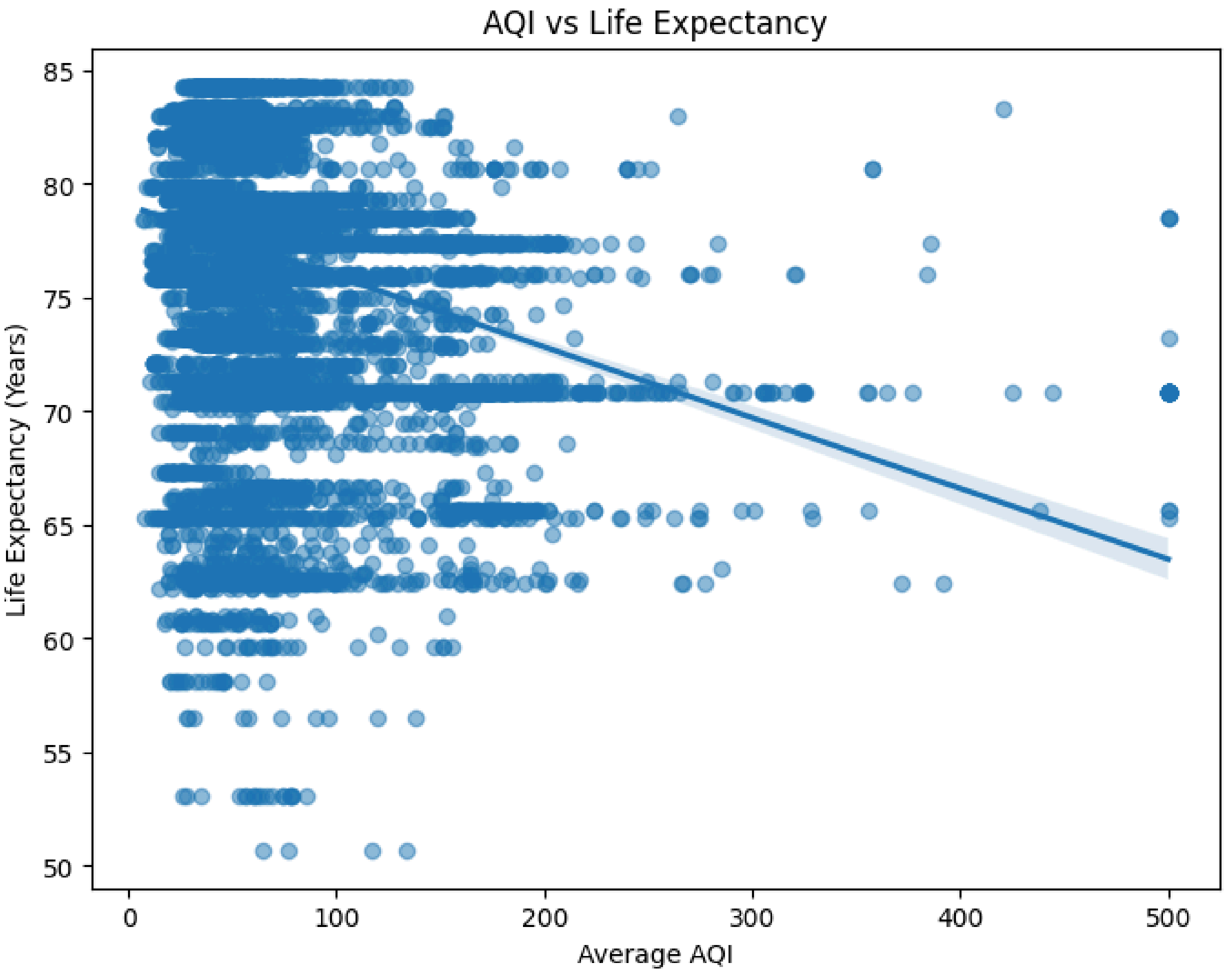
Top 10 Most Populated Countries by AQI

The bar chart ranks the ten countries with the highest average Air Quality Index. The Republic of Korea shows the highest pollution level, significantly exceeding 400 AQI, indicating hazardous air quality. Other countries, including Bahrain, Pakistan, Saudi Arabia, Mauritania, and Qatar, also report high pollution levels, suggesting chronic exposure to particulate matter and industrial emissions. These results highlight regional disparities in air quality and point to the need for targeted interventions in industrial and urban centers within Asia and the Middle East.



AQI versus Life Expectancy

The scatter plot examines the relationship between average AQI levels and life expectancy across multiple countries and reveals a predictable pattern that as AQI rises the life expectancy falls.This pattern suggests that higher exposure to air pollution is associated with reduced longevity and overall population health. Countries with AQI values above 200 show the steepest decline in life expectancy, reinforcing the global health implications of sustained air pollution. The distribution of data also reveals that nations with cleaner air (AQI below 100) generally maintain higher and more consistent life expectancy rates. This finding supports the broader conclusion that reducing particle matter and gas emissions can lead to measurable gains in public health outcomes.



Residuals versus Predicted AQi

The residual plot evaluates the accuracy and consistency of the predictive model used to estimate AQI. Each point represents the difference between observed and predicted AQI values. Most data points cluster around the zero line, indicating that the model provides reliable predictions for the majority of observations. However, the spread of residuals at higher predicted AQI values suggests greater variability in extreme pollution conditions. This could reflect regional differences in emission sources or limitations in available sensor data. Overall the model demonstrates strong performance for moderate air quality ranges, supporting its use in large-scale predictive analysis and forecasting applications for environmental and health planning.

Insights from EDA

The analysis of air quality and health indicators reveals several important patterns across countries. The global AQI map and the top ten ranking highlight that regions such as East Asia, the Middle East, and North Africa consistently experience the highest pollution levels. The correlation heatmap further confirms a strong relationship between poor air quality and adverse health outcomes. Higher AQI values correspond with higher mortality rates attributed to air pollution, as well as lower life expectancy and healthy life years. The scatter plot between AQI and life expectancy reinforces this relationship, showing a clear downward trend where higher pollution levels are associated with shorter lifespans.

The heatmap of model prediction accuracy and the residuals plot also provide early insight into data consistency. The confusion matrix demonstrates that predictions are most reliable for moderate air quality conditions, while extreme AQI levels introduce greater variability. The residuals plot supports this finding, showing more dispersion at higher AQI values. Together, these results confirm that air quality is a strong determinant of health outcomes and that reliable modeling requires robust, high-quality environmental data across pollution levels.

These observed relationships formed the basis for developing predictive models capable of capturing complex, non-linear patterns between pollutants and health outcomes. CleanAir Futures International advanced this research by integrating IoT-enabled sensor data with global health socioeconomic datasets to produce a more dynamic and regionally specific understanding of risks. The goal was to build models that not only estimate AQI with greater accuracy but also predict its impact on mortality and life expectancy in real time.

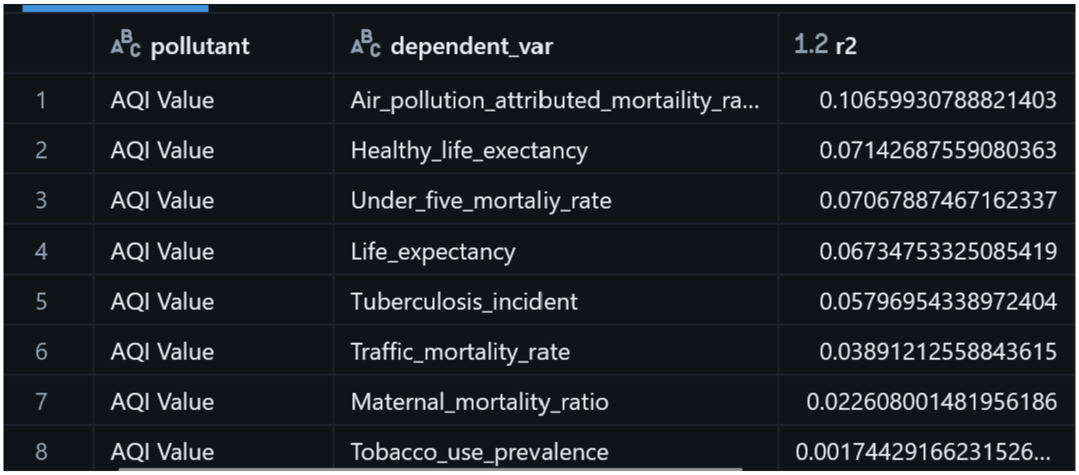
Modeling & Analysis

Modeling Strategy

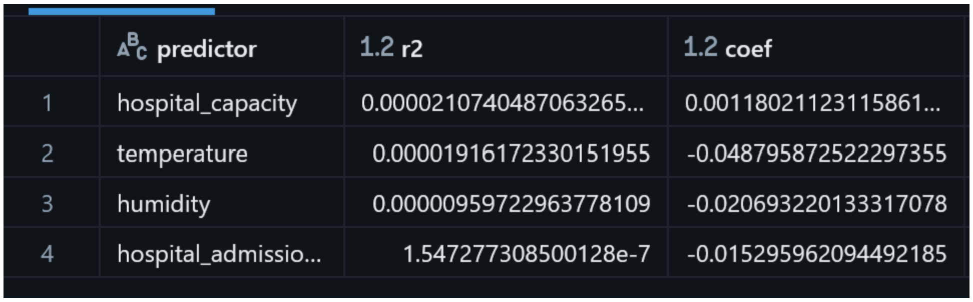
We initially used linear regression to understand the relationships between the key data points. This allowed us to understand which attributes would be most aligned with our independent variable, AQI. This allows us to understand some of the key correlations between different health data from countries with poor air quality. Furthermore, once we tested the various linear regression models based on their complexity, we took the strongest performing feature combinations into our classification model. The classification model is intended to classify the air quality based on the relevant environmental, health, and demographic data. For all our models, we used a standard 30% of reserved data for the model test data and 70% of the dataset for our model training data.

Data Selection

We had two key data sets to compare health data with our air sensor IoT data. These compared attributes included a variety of environmental, health, and demographic data. With the use of linear regression, we were able to identify 5 key features to examine further: Air_pollution_attributed_mortaility_rate, Healthy_life_exectancy, Under_five_mortaliy_rate, Life_expectancy, and Tuberculosis_incident. These were identified as having a statistically significant relationship through the r^2 values. The values with an r^2 value greater than 0.05 were carried into our first model.



	pollutant	dependent_var	1.2 r2
1	AQI Value	Air_pollution_attributed_mortaility_ra...	0.10659930788821403
2	AQI Value	Healthy_life_exectancy	0.07142687559080363
3	AQI Value	Under_five_mortaliy_rate	0.07067887467162337
4	AQI Value	Life_expectancy	0.06734753325085419
5	AQI Value	Tuberculosis_incident	0.05796954338972404
6	AQI Value	Traffic_mortality_rate	0.03891212558843615
7	AQI Value	Maternal_mortality_ratio	0.022608001481956186
8	AQI Value	Tobacco_use_prevalence	0.00174429166231526...



	predictor	1.2 r2	1.2 coef
1	hospital_capacity	0.0000210740487063265...	0.00118021123115861...
2	temperature	0.00001916172330151955	-0.048795872522297355
3	humidity	0.00000959722963778109	-0.020693220133317078
4	hospital_admissio...	1.547277308500128e-7	-0.015295962094492185

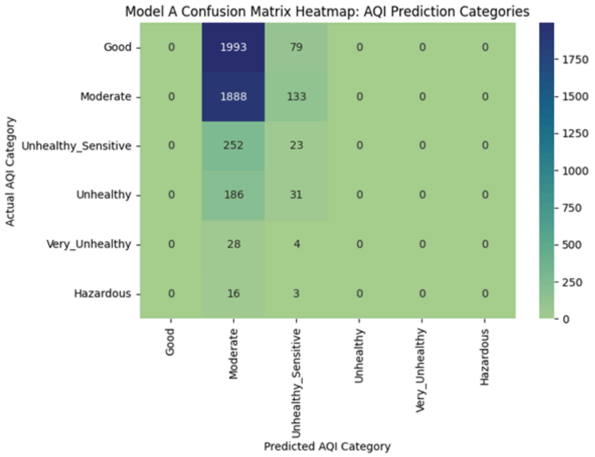
Predictive Model

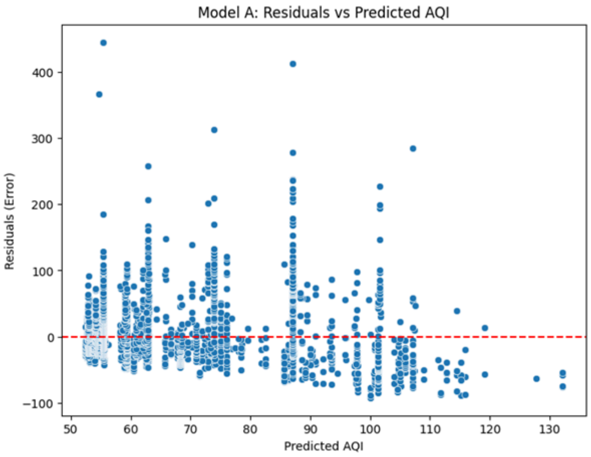
We evaluated 3 key predictive models to identify the strongest performing model based on the number of features present in the model. We used the Root Mean Squared Error (RMSE) value to determine the strongest performance across Models A, B, and C. Model C scored the lowest RMSE value and the highest R^2 values.

	Features	R^2	RMSE

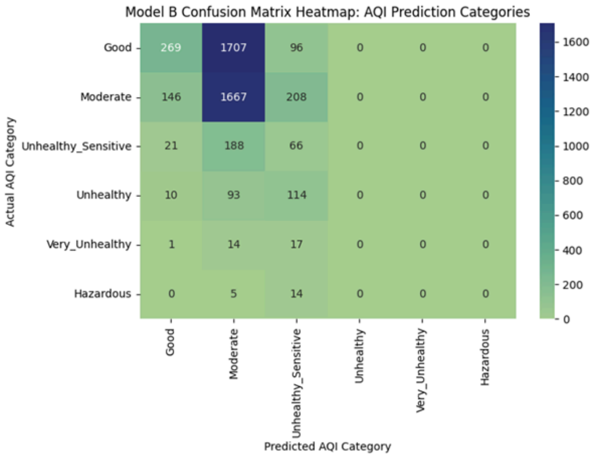
Model A	3	0.1017	40.5868
Model B	4	0.1361	39.8021
Model C	5	0.1372	39.7774

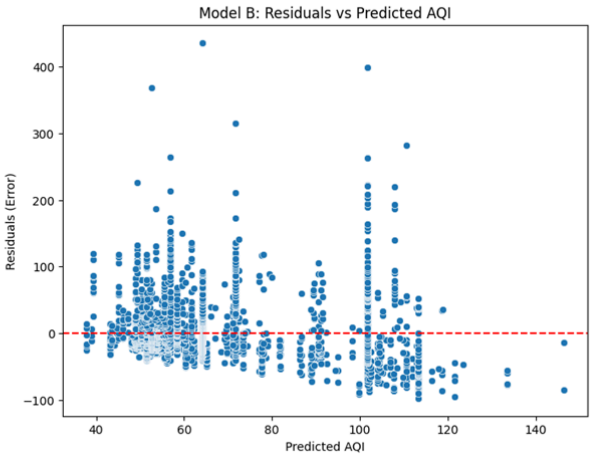
Further we used a confusion matrix to visualize the performance of the models. The confusion matrix assesses the accuracy of the AI model predicting AQI categories (Good, Moderate, Unhealthy Sensitive, Unhealthy, Very Unhealthy, Hazardous). We are primarily concerned with our model having high rates of false positives, as this prevents our organization from providing the assistance we strive for in these communities affected by poor air quality. We also examined the Residuals and predicted AQIs to understand the density of where our greatest errors were coming in.



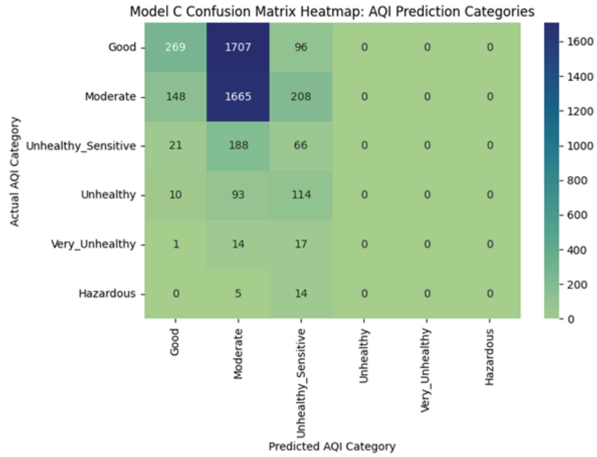


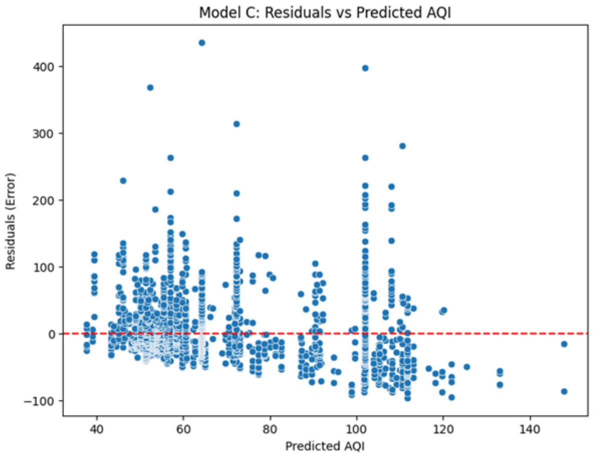
For Model A, we see a higher rate of missed predictions of the Good vs Moderate air quality. The greatest concern being the model's inability to predict any of the unhealthy, very unhealthy, and hazardous environments. These seem to be a significant amount of Moderate predictions and potentially over-fitting of the model. We can also see there is a higher rate of error in the lower air quality, higher AQI, value.





With Model B, we see stronger performance in the Good and unhealthy-sensitive air quality. The greatest concern being the model's inability to predict any of the unhealthy, very unhealthy, and hazardous environments. These seem to be a significant amount of Moderate predictions and potentially over-fitting of the model. We can also see there is a higher rate of error in the lower air quality, higher AQI, values. However, Model B does demonstrate lower frequency of error in predicting the air quality.





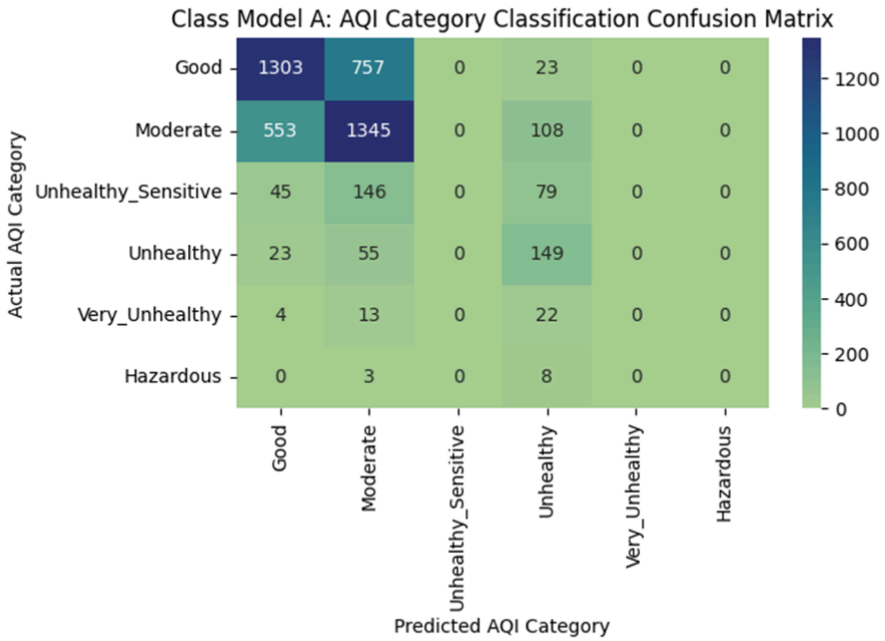
With Model C, we see a stronger performance in correctly classifying Good, and unhealthy-sensitive air qualities, however continues to fail to predict any of the unhealthy, very unhealthy, and hazardous environments. The high diagonal values for the Good and Moderate categories indicate strong model performance for lower pollution levels. However, smaller counts in Unhealthy and Hazardous classes suggest limited data or underrepresentation of extreme pollution events. Overall, the model demonstrates high reliability in classifying common air quality ranges but may benefit from additional training data to improve predictions for severe pollution conditions.

Model C performed the strongest with utilizing all 5 features to predict the AQI values. This provides us with insight into the relationship between higher rates of pollution leading to higher rates of health issues in a few key areas: life expectancy, a toddler's ability to thrive, lung disease, and pollution related deaths. To further develop our model, we developed a classification model to better predict the general state of a country's air quality.

Classification Model:

We utilized a Random Forest classification model to predict the classification of the air quality. We used a 10 tree model with a maximum depth of 5. We found that this produced an accuracy of 60.33% for the model.

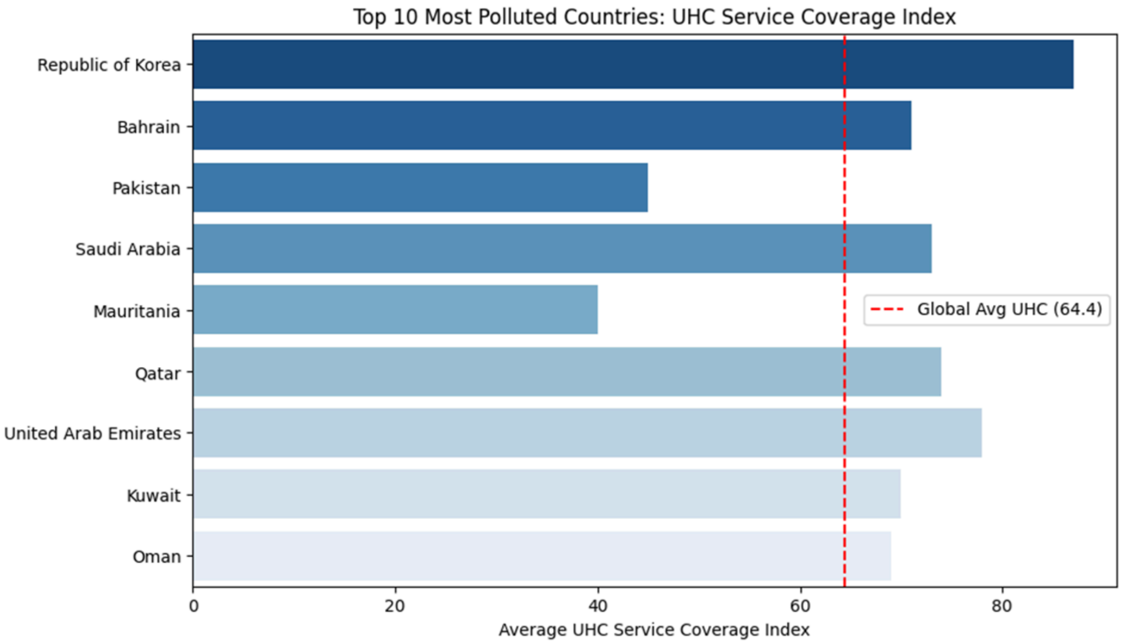
	Number of Trees	Depth	Accuracy
Class Model A	10	5	60.33%



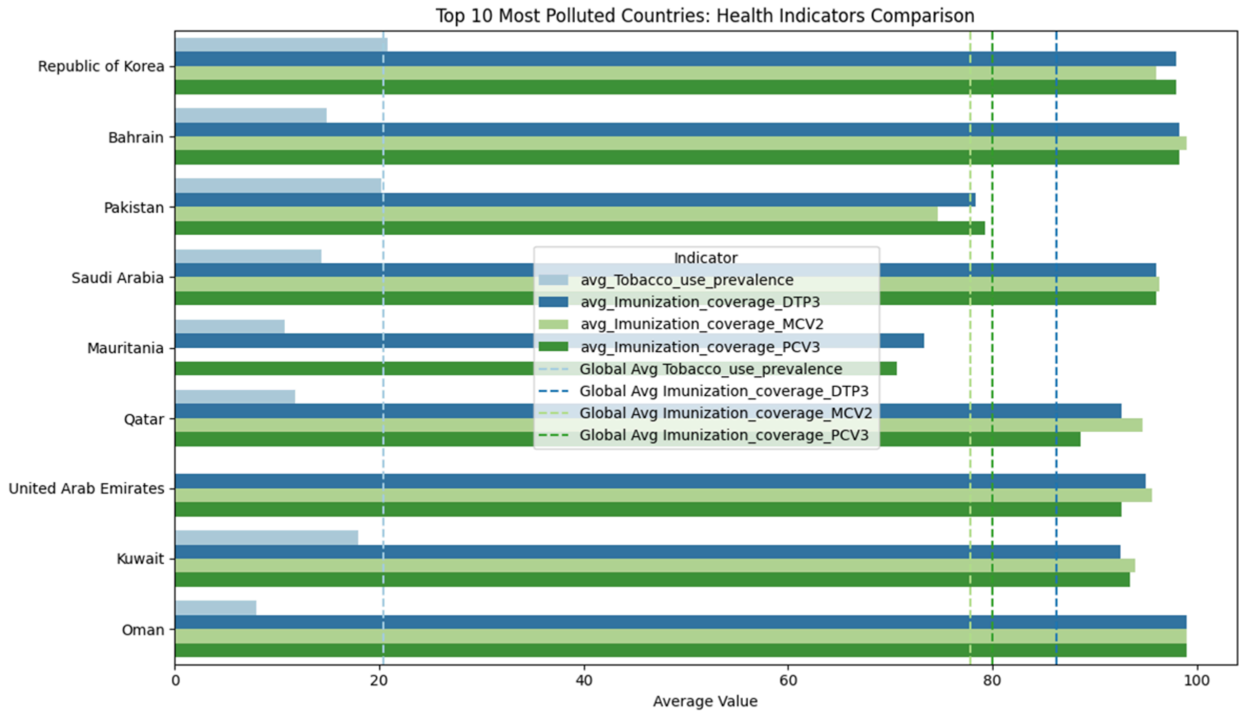
However, it is valuable to understand how those errors are spread. We found this model was able to predict good, moderate, and unhealthy air quality conditions at a significantly higher accuracy than our previous models did. However, we do have a significantly smaller sample size for the very unhealthy and hazardous air quality conditions. Due to this, it would be difficult to effectively train this model to have significantly higher accuracy without incorporating more data points into the model training and testing.

Further Analysis

With the insights provided to us from our models, we wanted to take a closer look at each of the countries that scored the lowest on their air quality. We compared these countries' UHC Service Coverage index, tobacco use prevalence, and immunization coverage. We wanted to examine these key characteristics, as they can have impacts on our dependent variables outside of air pollution.



Each country's UHC Service Coverage Index is a metric used by the World Health Organization to track the coverage of essential health services in each country. The average global score is 64.4 for this period. We only had two countries that scored below the global average, which was Pakistan and Mauritania. Pakistan has the 3rd poorest air quality and Mauritania has the 5th poorest air quality. This may indicate that there are some health coverage-related challenges faced in these countries that may interfere with our model accuracy.



We also examined vaccination coverage and tobacco use in these 10 countries. We were able to find that there seems to be identification that most of these countries have a report below the global average for tobacco use prevalence. This likely indicates there is little to no impact on our algorithm from tobacco use skewing our results. Additionally, we looked at the immunization coverage for these countries. We specifically looked at three vaccines: DTP3, MCV2, and PCV3. The DTP vaccine immunizes against diphtheria, whooping cough, and tetanus. MCV2 is the second dose for the measles vaccine. PCV3 vaccine protects against the different types of pneumococcal bacteria, which is a common cause of death in children under 5 years old. Only two countries of the ten consistently scored below the global average for this immunization coverage, which was Pakistan and Mauritania. Pakistan was just below the global average, while Mauritania was significantly below the global average. It does warrant that while these two counties do suffer from poor air quality, they also have other factors that are likely to impact their health data.

V. Insights, Implications & Business Relevance

A. Synthesis of Key Findings

After conducting our analysis, one of the main takeaways is that local factors have a low influence. The IoT regressions show that temperature, humidity, hospital admissions, and capacity each have R^2 values below 0.002 for city level AQI. After combining these variables, the model still produced an R^2 close to 0. This indicates that daily air quality is driven by broader environmental and industrial drivers, not local weather or hospital load variations. On a country-level analysis, we see that higher AQI is associated with higher air-pollution-attributed mortality and under-five mortality and with lower life expectancy. Each one-point rise in AQI corresponds to roughly 0.425 extra pollution-related deaths per 100 000 people and 0.028 years less healthy life expectancy. The $R^2 \leq 0.11$ show modest correlation but are consistent across variables. These findings indicate correlation rather than causation and highlight that air quality is one of many determinants of population health.

In our multivariable country-level regression (Model C), which uses five health metrics (pollution-mortality, healthy life expectancy, under-five mortality, overall life expectancy, and tuberculosis incidence), we explain around 14% of the variation in AQI across countries.. Positive coefficients on pollution mortality and under-five mortality and negative coefficients on healthy life expectancy underscore the directional link, but the low R^2 signals large residual uncertainty and suggests that these demographic indicators alone explain only a small share of

AQI variation. Additionally, extreme pollution is geographically concentrated. Countries, like the Republic of Korea, Bahrain, Pakistan, and Middle Eastern nations, exhibit markedly higher average AQI than the rest of the examined countries. This suggests NGOs should focus on these concentrated areas to make the most impact.

B. Strategic Implications and policies recommendations

Our findings and insights can be translated through a business lens into valuable knowledge to make data-driven resource allocation, operation optimization, investment strategies and to address environmental justice implications.

The country level regressions show that higher AQI is systematically associated with higher air pollution attributed mortality and higher under five mortality, and with lower life expectancy. Even with R^2 values of 0.07–0.11, this is enough to prioritize countries where both AQI and mortality are high. The ranked table of average AQI by country and the joint AQI health summary identify a set of countries like Republic of Korea, Bahrain, Pakistan, Mauritania, and Guinea-Bissau with higher average AQI and, in many cases, worse child and pollution-related mortality. These should be treated as tier-one countries for technical assistance and support programs. Within these countries NGOs can target larger cities to create a more focused impact.

This recommendation aligns with our random forest classification model with 10 trees and a maximum depth of 5 achieved an overall accuracy of 60.33 %. It classifies Good and Moderate air quality conditions relatively well but often miscategorises unhealthy, very unhealthy, and hazardous environments into the moderate category. This is reflecting both higher error and fewer training examples at the extreme AQI levels. NGOs must think about AQI and mortality metrics as a screening tool and then layer in feasibility to build an intervention portfolio. Prioritizing healthier zones could be beneficial for the NGOs based on our model. As our model poorly predicts areas with unhealthy or hazardous conditions, NGOs can fund their fixed infrastructure like offices and warehouses on sites with better air quality and serve the hazardous area by deploying mobile clinics and other short term but high impact solutions.

VI. Reflection, Limitations & Future Work - Daniel

A. Project Limitations

In this section we address key limitations of our project. One of our concerns is that AQI definitions can be slightly different across countries. In many systems, AQI is calculated as the maximum of individual pollutants, so it combines many pollutants into a single number and uses category thresholds that do not change in a smooth, linear way. Our analysis also combines data from different sources and levels. The IoT dataset provides city-level daily AQI and the WHO health data is focused on the country level and often averaged over several years. Turning city data into country averages and linking them to national mortality and life expectancy can hide big differences within countries. Sensor coverage can also be another issue as some countries have many monitors, while others have only a few in major cities. This can result in bias as country averages can shift toward better monitored locations.

When it comes to the accuracy of the measurement, several of the particulate matter sensors used in IoT networks can be sensitive to humidity and temperature, causing lower accuracy and gradual drift over time as these sensors age. If calibration differences are not fully corrected, AQI results may contain systematic error.

On the modeling side, our regressions have modest explanatory power (R^2 values around 0.10–0.14), and the Random Forest classification model achieves 60.33 % accuracy with clear class imbalance: Unhealthy, Very Unhealthy, and Hazardous categories are under-represented and often misclassified as Moderate. This limits the reliability of our predictions, particularly in the most polluted environments where NGOs are most concerned.

B. Lessons Learned

This project provided us with valuable technical and teamwork lessons. On the technical side, working in Databricks with PySpark showed us the importance of understanding our data before doing any modeling. We had to standardize city and country names, solving the issue with dates, turn string columns to numeric and handle invalid column names. When joining IoT sensor outputs to WHO tables we also had to be precious and carefully pick prime keys and foreign keys. We also learned that low R^2 values can still be used for decision making when used correctly. The low R^2 values in our regressions and the 60.33% accuracy in our classification forced us to focus less on precise prediction and more on ranking and screening.

From a team perspective, we picked a divide and concur approach, where we build on each member’s strength. Some were responsible for finding clean datasets, others did coding and writing. The project taught us that a good data science project that creates impact is not only about code but also about how we communicate our findings, what logic we follow and how we interpret our results.

C. Future Research

Future work could address the limitations mentioned above and extend our approach in several directions. Most importantly, future work should include the integration of real time AQI, health metrics and pollutant data via APIs. This would allow NGOs to move from static country rankings to dynamic monitoring. Collecting daily or hourly AQI with more granular health outcomes like emergency visits or respiratory admissions where available could strengthen short term risk assessments. Further research should include varaóiables that show income distribution, urbanization rates, industrial composition and weather factors could improve model performance and provide more detail.

VII. Conclusion

This project set out to help CAFI use IoT-based air quality data and global health indicators to make data driven, smart decisions about where to deploy scarce resources. We cleaned and integrated city level sensor readings with WHO country health metrics in Databricks. After that we built regression and classification models around AQI, mortality, and life expectancy. We created a framework that links environmental conditions to population health in a way that NGOs can easily utilize the data .

Our results show that local, daily factors such as temperature and humidity explain almost none of the variation in AQI. However, countries with persistently high AQI tend to have higher air pollution attributed mortality, higher under five mortality, and lower life expectancy. Extreme pollution is concentrated in a small group of countries, which can be treated as tier one priorities. Our models are best used as screening tools to rank countries and inform where to place fixed infrastructure versus mobile, high-impact interventions.

References

Ramadan, M. N. A., Ali, M. A. H., Khoo, S. Y., Alkhedher, M., & Alherbawi, M. (2024). *Real-time IoT-powered AI system for monitoring and forecasting of air pollution in industrial environment. Ecotoxicology and Environmental Safety, 283,* 116856. https://doi.org/10.1016/j.ecoenv.2024.116856

U.S. Environmental Protection Agency. (n.d.). *Air*. Retrieved November 3, 2025, from https://www.epa.gov/report-environment/air