**The Impact of Air Pollution on Respiratory Health and Economic Costs: A Multi-Sector Analysis**

# Introduction to the Problem

Air pollution, especially particulate matter less than 2.5 microns in diameter **(PM2.5),** is a critical public health concern. PM2.5 particles can penetrate deep into the lungs and bloodstream, causing respiratory and cardiovascular diseases. The Air Quality Index **(AQI)** is a metric used to quantify the level of air pollution; higher AQI values indicate worse air quality. This project models the relationship between PM2.5 levels and healthcare expenditures measured as a percentage of Gross Domestic Product **(GDP)**, using real-world data, linear regression, and visual analytics to predict the economic burden of pollution. The goal is to help policymakers make data-informed decisions.

Air pollution, especially PM2.5, is a critical public health concern. This project models its relationship with health expenditures, using real-world data, linear regression, and visual analytics to predict economic burdens.

# Relevant Work

This study builds on foundational work by the World Health Organization (WHO), the Centers for Disease Control and Prevention (CDC), and the Environmental Protection Agency (EPA). WHO has linked air pollution to more than 7 million premature deaths annually. The Harvard T.H. Chan School of Public Health found that even slight increases in PM2.5 can raise COVID-19 mortality rates. Despite this, many studies have not quantified the direct economic impact. This project bridges that gap by connecting air quality directly to national healthcare expenditure.

This work is based on research from WHO, Harvard, and government health data. It bridges the gap by linking air quality directly to **GDP-based healthcare spending**, a connection often overlooked in previous models.

# Description of the Project

This project uses three data sets that were manually collected and stored locally:

* **airq2.csv**: Contains PM2.5 levels by country from international air monitoring agencies.
* **data1.csv**: Healthcare expenditure (% of GDP) from the World Bank (2017–2021).
* **res.csv**: Hospital respiratory visit percentages from national health statistics.  
    
  **Subsections:**  
  1. Data Collection: Datasets were downloaded and preprocessed for consistency.  
  2. Data Cleaning: Missing values and anomalies were handled.  
  3. Merging: Data was joined on the country field.  
  4. Feature Engineering: Average healthcare expenditure over 5 years was calculated.  
    
  **Below are figures generated from this implementation.**

The project uses data from WHO, World Bank, and health agencies. After cleaning and merging, the model calculates correlation and uses regression for predictions. Below are key visuals and methodology results:

A graph with blue dots and a red line

AI-generated content may be incorrect.

Figure 1: Shows the relationship between PM2.5 levels and healthcare expenditure.

A graph with blue lines

AI-generated content may be incorrect.

Figure 2: Respiratory illness trends indicate peaks aligned with higher pollution years.

A graph with a line going up

AI-generated content may be incorrect.

Figure 3: Global health spending trend over time based on economic data.

A graph with a line

AI-generated content may be incorrect.

Figure 4: Forecasted health expenditure across different hypothetical PM2.5 levels.

# Results and Evaluation

The results confirm a moderate to strong relationship between PM2.5 levels and healthcare costs.  
  
**Subsections:**  
- Correlation Analysis: Pearson correlation showed a negative relationship between PM2.5 and GDP-based health expenditure.  
- Regression Metrics:  
 - Coefficient: Indicates the change in health expenditure with each unit change in PM2.5.  
 - R² (R-squared): Proportion of variance in health costs explained by PM2.5 levels.  
 - RMSE (Root Mean Squared Error): Average prediction error.  
  
The regression line in Figure 1 shows this trend clearly. Figures 2 and 3 further support the findings by visualizing temporal trends in health issues and spending. Figure 4 forecasts healthcare costs based on hypothetical AQI values.

The model achieved a moderate R² score, indicating a clear correlation between pollution and spending. Visualizations confirmed trend patterns. RMSE provided a benchmark for prediction accuracy.

# Conclusion and Future Work

This study successfully demonstrated a quantifiable link between PM2.5 air pollution and national health expenditure. Using basic regression, we modeled this relationship and validated it with real-world trends.  
  
**Future work should consider:**  
- Including additional features like population density, income levels, and industrialization rates.  
- Using advanced models such as Random Forest or Gradient Boosting.  
- Collecting city-level and monthly data for finer granularity.  
- Connecting with APIs for real-time data and forecasting.

We conclude that PM2.5 pollution has economic consequences. Future work includes deeper variable modeling and finer-grained data for more precision.

# References

[1] WHO. (2021). Air Pollution and Health.  
[2] Harvard T.H. Chan School of Public Health. (2020).  
[3] EPA. (2022). Air Quality Index.  
[4] CDC. (2021). Environmental Health.  
[5] Python Libraries: pandas, seaborn, matplotlib, scikit-learn