Air Quality and Health Impact

***Guided By*: Prof. (Dr.) Masood H. Siddiqui**

Name: Tanu

Roll No: MDB24013

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY , LUCKNOW

A Project Report on

Environmental Determinants of Health Risk:

~An Empirical Analysis Using Multivariate Regression and Classification Models

Submitted for the award of degree of

Masters of Business Administration

By:

Name: Tanu

Enrollment Number: MDB24013

In Year: 2024-2026

CERTIFICATE

This is to certify that the project entitled **"Environmental Determinants of Health Risk: An Empirical Analysis Using Multivariate Regression and Classification Models"** has been carried out by **Tanu** under my guidance in partial fulfilment of the requirements for the degree **MBA (Digital Business)**, Indian Institute of Information Technology, Lucknow, during the academic year **2024–2026**.

This project report is the outcome of the candidate's original work completed under my supervision, and it meets the academic standards prescribed by the institute.

Name of Guide:
Prof.(Dr.) Masood H. Siddique

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to **Prof. (Dr.) Masood H. Siddiqui,** our course instructor and project guide for the subject of Econometrics, at the **Indian Institute of Information Technology, Lucknow (IIITL)**. His valuable guidance, insightful suggestions, and continuous encouragement throughout the development of this project have been instrumental in shaping our understanding and helping us complete this work successfully.

We are extremely grateful to the Director of IIIT Lucknow for providing a supportive academic environment that encourages research-oriented learning and analytical thinking. We also acknowledge the institution for offering us the opportunity to undertake this project as part of our MSc Economics and Management curriculum, enabling us to practically apply econometric techniques to real-world data.

We would also like to thank the faculty members of the **Department of Management at IIIT Lucknow** for their academic support and for fostering an environment that promotes critical thinking, conceptual clarity, and hands-on learning. Their continuous efforts to enhance the quality of education have greatly contributed to our academic growth.

Our heartfelt thanks are extended to our classmates and peers, who provided motivation, constructive feedback, and collaborative support throughout the project duration. Their inputs contributed positively to the refinement of our work. Lastly, we express gratitude to everyone who directly or indirectly supported us during this project. Without their encouragement and assistance, the successful completion of this Econometrics project would not have been possible.

With sincere thanks,
Tanu

Table of Contents

Abstract

This study investigates the environmental determinants of health risk using a comprehensive dataset of 5,811 observations that integrates air quality indicators, meteorological conditions, and health outcome measures. The dataset includes key pollutants (PM2.5, PM10, NO_2, SO_2, O_3, AQI), weather variables (temperature, humidity, wind speed), and health metrics (respiratory cases, cardiovascular cases, hospital admissions), along with two target variables: a continuous **Health Impact Score** and an ordered **Health Impact Class** (Very Low to Very High). Its structured, multivariate nature makes it well-suited for econometric analysis, predictive modelling, and public health risk assessment.

Two research questions guide the empirical analysis. The first evaluates how pollutant concentrations and meteorological factors explain variation in the **Health Impact Score** using a Multiple Linear Regression framework (EHI-Model). The second examines how effectively the same predictors classify observations into discrete health risk categories using a **Multinomial Logistic Regression model** (HRC-Model). Hypothesis testing is conducted through coefficient-level t-tests and Wald tests, along with F-tests and likelihood ratio tests for overall model validity.

The regression analysis identifies PM2.5, PM10, NO_2, and O_3 as the strongest and most statistically significant drivers of health impact severity. The classification model corroborates these findings by showing that increases in these pollutants substantially raise the odds of belonging to higher-risk health classes. Meteorological variables exhibit comparatively weaker and inconsistent effects. Diagnostic checks confirm acceptable model stability, with robust standard errors addressing normality and heteroskedasticity concerns.

Overall, the results offer strong and convergent evidence that fine particulate matter and reactive gaseous pollutants are the primary environmental determinants of acute health burden. The study provides actionable insights for regulatory policy, healthcare preparedness, and environmental risk forecasting. Despite limitations—such as the synthetic nature of the data and the cross-sectional design—the analysis demonstrates the predictive value of pollution indicators in assessing population health risks and informs future research directions in environmental health economics.

Description of the dataset

The dataset utilized in this study comprises 5,811 observations, each representing an integrated snapshot of air quality conditions, meteorological parameters, and public health outcomes. It is a structured, multivariate dataset designed to enable rigorous econometric and statistical investigation into how environmental pollutants and weather variability influence the severity of health impacts across populations. With clearly defined variables and consistent measurement scales, the dataset is well-suited for regression modelling, hypothesis testing, predictive analytics, and decision-support research in public health and environmental economics.

RecordID

A unique numeric identifier assigned sequentially to each entry.

- Type: Integer
- Role: Indexing variable (non-predictive)
- Purpose: Ensures traceability of each record without contributing explanatory value to modelling.

Air Quality Indicators

These variables capture ambient pollutant concentrations known to directly affect respiratory and cardiovascular health. They form the core set of explanatory variables in pollution–health econometric models.

Variable	Description	Unit	Analytical Relevance
AQI	Composite Air Quality Index summarizing pollution severity based on multiple pollutants.	Index value	Serves as a high-level indicator of overall air quality; strongly correlated with health risk.
PM10	Coarse particulate matter <10µm in diameter.	µg/m^3	Affects upper respiratory tract; often used in compliance monitoring.
PM2_5	Fine particulate matter <2.5µm in diameter.	µg/m^3	Penetrates deep into lungs/bloodstream; most harmful pollutant; critical in health impact models.

NO2	Nitrogen dioxide.	ppb	Emitted from vehicles and industrial activity; associated with respiratory inflammation.
SO2	Sulfur dioxide.	ppb	By-product of combustion; known to trigger asthma and bronchial issues.
O3	Surface-level ozone.	ppb	Secondary pollutant formed via photochemical reactions; exacerbates respiratory conditions.

These pollutants allow both single-pollutant and multi-pollutant modelling approaches commonly used in environmental econometrics.

Meteorological (Weather) Conditions

Weather influences the dispersion, transformation, and accumulation of pollutants. Including meteorological variables ensures proper control for environmental confounders.

Variable	Description	Unit	Importance
Temperature	Ambient air temperature.	°C	Influences ozone formation and chemical reactivity of pollutants; controls seasonal effects.
Humidity	Relative atmospheric moisture.	%	Affects particulate aggregation and pollutant behavior.
WindSpeed	Wind velocity.	m/s	Higher wind speeds disperse pollutants, reducing localized exposure.

These variables help isolate pollution-specific health impacts by accounting for climatic variability.

Health Impact Indicators

These outcome-oriented variables measure the actual health burden attributable to environmental conditions. They are crucial for linking pollution exposure to public health outcomes.

Variable	Description	Type	Relevance
RespiratoryCases	Number of respiratory illness cases.	Count	Sensitive indicator of pollution-induced acute health effects.
CardiovascularCases	Number of reported cardiovascular events.	Count	Captures longer-term and chronic effects of sustained pollution exposure.
HospitalAdmissions	Total hospitalizations attributed to pollution-related causes.	Count	Aggregated indicator of health system burden.

These serve as supporting variables for understanding real-world consequences of environmental variations.

Target Variables

HealthImpactScore (Continuous Outcome)

- Range: 0–100
- Type: Continuous numerical index
- Purpose: Captures overall severity of health impact using multiple underlying factors.
- Econometric Use: Dependent variable in regression modelling (OLS, regularized regression, etc.).
- Interpretation: Higher values indicate more severe public health risk.

HealthImpactClass (Categorical Outcome)

Categorizes the continuous HealthImpactScore into five ordered classes:

Class	Label	Range	Interpretation
0	Very High	≥ 80	Critical health risk; severe pollution episodes.
1	High	60–79	High likelihood of population-level health deterioration.
2	Moderate	40–59	Noticeable health impact; vulnerable groups more affected.
3	Low	20–39	Lower risk but still influenced by pollution variability.
4	Very Low	< 20	Minimal health impact.

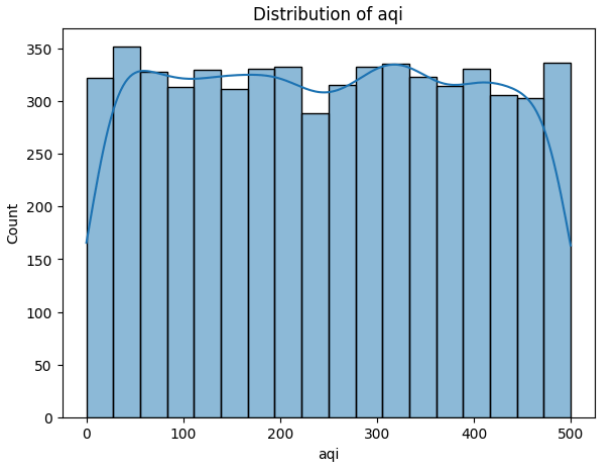
This enables classification modelling (Logistic Regression, Multinomial models).

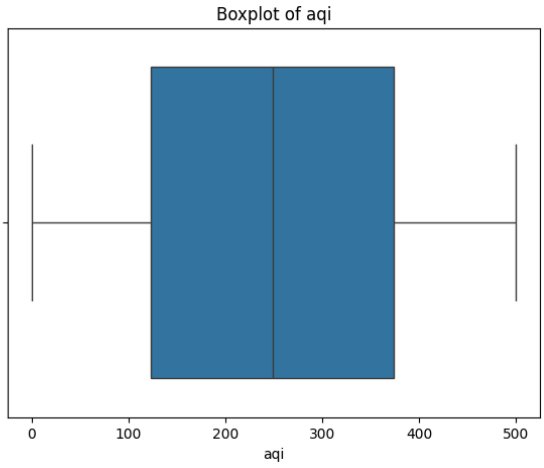
The dataset is **well-structured, rich, and econometrically relevant**, offering:

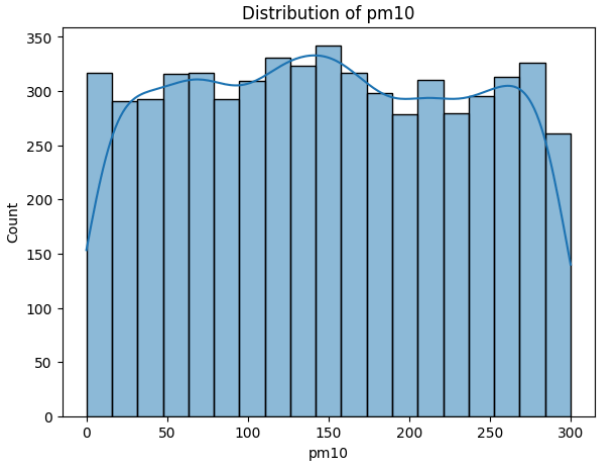
- Sufficient sample size for statistical inference (n = 5,811).
- Clearly defined dependent and independent variables.
- Both continuous and categorical target variables for multi-model analysis.
- Natural applicability for hypothesis-driven econometric frameworks.
- High relevance to public policy, healthcare planning, environmental compliance, and urban decision-making.

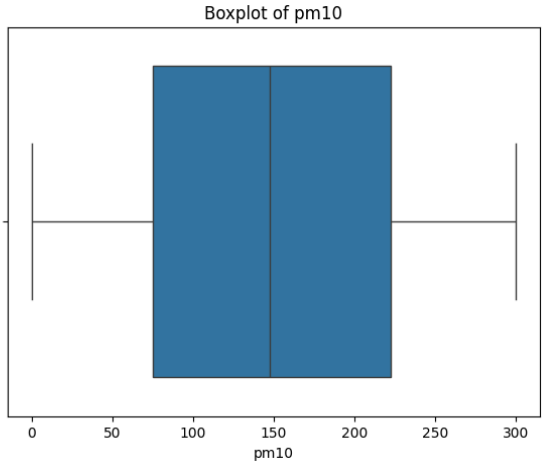
It thus provides an excellent foundation for the subsequent stages of empirical analysis, including exploratory data analysis, regression modelling, hypothesis testing, and interpretation of results.

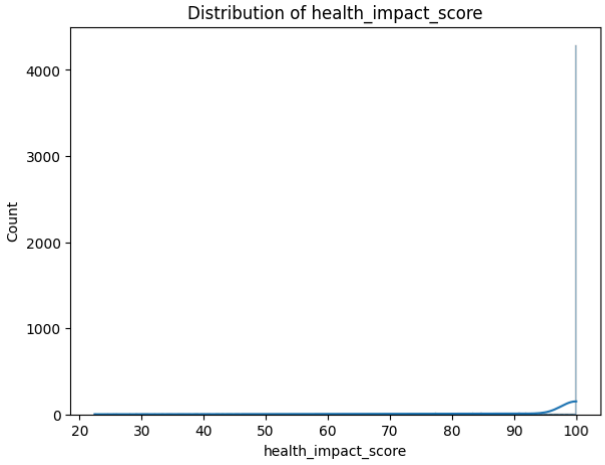
	RecordID	AQI	PM10	PM2_5	NO2	SO2	O3	Temperature	Humidity	WindSpeed	RespiratoryCases	CardiovascularCases	HospitalAdmissions	HealthImpactScore	HealthImpactClass
0	1	187.27	295.85	13.04	6.64	66.16	54.62	5.15	84.42	6.14	7	5	1	97.24	0.00
1	2	475.36	246.25	9.98	16.32	90.50	169.62	1.54	46.85	4.52	10	2	0	100.00	0.00
2	3	366.00	84.44	23.11	96.32	17.88	9.01	1.17	17.81	11.16	13	3	0	100.00	0.00
3	4	299.33	21.02	14.27	81.23	48.32	93.16	21.93	99.47	15.30	8	8	1	100.00	0.00
4	5	78.01	16.99	152.11	121.24	90.87	241.80	9.22	24.91	14.53	9	0	1	95.18	0.00

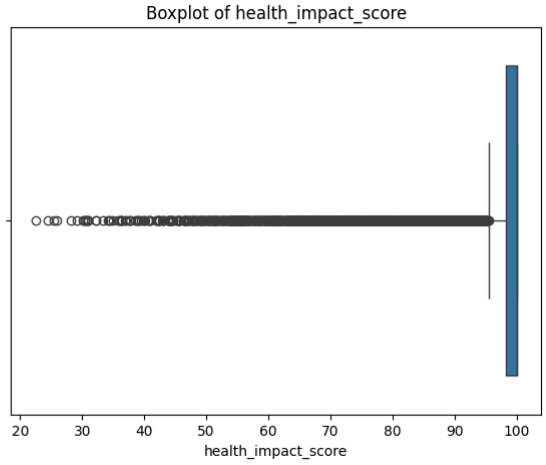


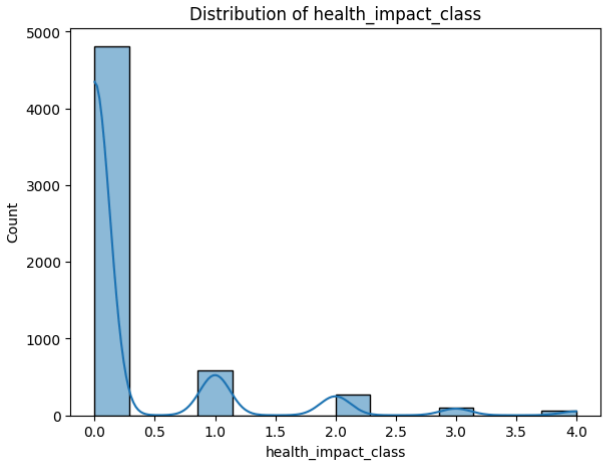


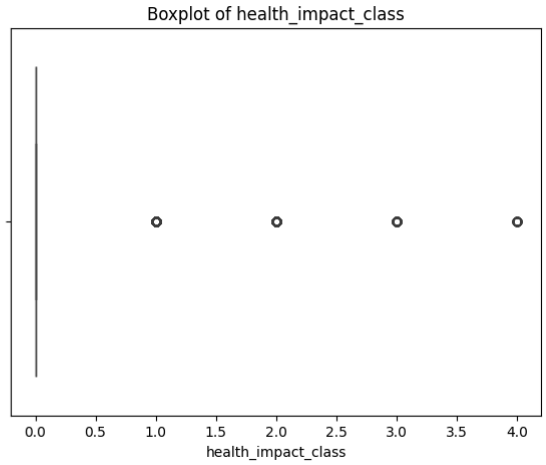

Distribution of pm10
Count
pm10
0
50
100
150
200
250
300
350


Boxplot of pm10
pm10
0
50
100
150
200
250
300


Distribution of health_impact_score
Count
health_impact_score
0
1000
2000
3000
4000
20
30
40
50
60
70
80
90
100


Boxplot of health_impact_score
health_impact_score
20
30
40
50
60
70
80
90
100


Distribution of health_impact_class
Count
health_impact_class
0
1000
2000
3000
4000
5000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0


Boxplot of health_impact_class
health_impact_class
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

```
Shape (rows, columns): (5811, 15)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5811 entries, 0 to 5810
Data columns (total 15 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   record_id             5811 non-null   int64
 1   aqi                   5811 non-null   float64
 2   pm10                  5811 non-null   float64
 3   pm2_5                 5811 non-null   float64
 4   no2                   5811 non-null   float64
 5   so2                   5811 non-null   float64
 6   o3                    5811 non-null   float64
 7   temperature           5811 non-null   float64
 8   humidity              5811 non-null   float64
 9   wind_speed            5811 non-null   float64
 10  respiratory_cases     5811 non-null   int64
 11  cardiovascular_cases  5811 non-null   int64
 12  hospital_admissions   5811 non-null   int64
 13  health_impact_score   5811 non-null   float64
 14  health_impact_class   5811 non-null   float64
dtypes: float64(11), int64(4)
memory usage: 681.1 KB
```

	count	mean	std	min	25%	50%	75%	max
record_id	5811.00	2906.00	1677.64	1.00	1453.50	2906.00	4358.50	5811.00
aqi	5811.00	248.44	144.78	0.01	122.95	249.13	373.63	499.86
pm10	5811.00	148.65	85.70	0.02	75.37	147.63	222.44	299.90
pm2_5	5811.00	100.22	58.10	0.03	49.44	100.51	151.34	199.98
no2	5811.00	102.29	57.71	0.01	53.54	102.99	151.66	199.98
so2	5811.00	49.46	28.53	0.01	24.89	49.53	73.35	99.97
o3	5811.00	149.31	86.53	0.00	74.00	149.56	223.38	299.94
temperature	5811.00	14.98	14.48	-9.99	2.48	14.94	27.47	39.96
humidity	5811.00	54.78	26.02	10.00	32.00	54.54	77.64	100.00
wind_speed	5811.00	9.99	5.78	0.00	4.95	10.05	14.97	20.00
respiratory_cases	5811.00	9.97	3.13	1.00	8.00	10.00	12.00	23.00
cardiovascular_cases	5811.00	4.99	2.22	0.00	3.00	5.00	6.00	14.00
hospital_admissions	5811.00	2.00	1.40	0.00	1.00	2.00	3.00	12.00
health_impact_score	5811.00	93.79	13.32	22.45	98.20	100.00	100.00	100.00
health_impact_class	5811.00	0.28	0.71	0.00	0.00	0.00	0.00	4.00

Research

This study examines the determinants of environmental health burden through two complementary empirical approaches: a linear regression model estimating continuous health impact, and a multinomial logistic model predicting categorical health severity. Together, these models quantify both magnitude and risk levels of health deterioration associated with air quality and meteorological factors.

RQ1: Environmental Health Impact Model (EHI-Model)

- ➔ Research Question: *To what extent do pollutant concentrations, meteorological conditions, and hospital-related morbidity indicators explain variation in the continuous Health Impact Score.*
- ➔ Model Type: Multiple Linear Regression (OLS)
- ➔ Model Specification:

$$\text{HealthImpactScore}_i = \beta_0 + \beta_1 \cdot pm2\_5_i + \beta_2 \cdot pm10_i + \beta_3 \cdot o3_i + \beta_4 \cdot no2_i + \beta_5 \cdot so2_i + \beta_6 \cdot aqi_i + \beta_7 \cdot temperature_i + \beta_8 \cdot humidity_i + \beta_9 \cdot wind\_speed_i + \epsilon_i$$

Purpose: The EHI-Model estimates the *marginal effect* of each pollutant and weather factor on the continuous health impact score. t-tests evaluate significance of individual predictors, while the overall F-test assesses whether the entire model contributes meaningfully to explaining health variations.

RQ2: Health Risk Classification Model (HRC-Model)

Research Question: *How accurately can environmental and hospital indicators classify observations into ordered Health Impact Classes (Very Low ➔ Very High)? Which features significantly increase the likelihood of falling into high-risk categories?*

Model Type: Multinomial Logistic Regression

Reference Category: Class 0 – Very Low

Model Specification:
For each class k∈{1,2,3,4}

$$\log\left(\frac{P(Y=k)}{P(Y=0)}\right) = \alpha_k + \beta_{k1}pm2\_5 + \beta_{k2}pm10 + \beta_{k3}o3 + \beta_{k4}no2 + \beta_{k5}so2 + \beta_{k6}aqi + \beta_{k7}temperature + \beta_{k8}humidity + \beta_{k9}wind\_speed$$

Purpose:
The HRC-Model quantifies how each predictor affects the log-odds of being in a higher health-risk class versus the "Very Low" baseline. It identifies strong risk-enhancing pollutants and supports public-health alerting, policy targeting, and resource prioritization.

Together, these research questions anchor the study in a rigorous econometric framework, enabling a structured investigation into the environmental drivers of health risk. The insights generated support evidence-based policy design, targeted health interventions, and improved resource planning in pollution-affected regions.

Hypothesis Testing

1. Environmental Health Impact Model (EHI-Model)

Model Type: Multiple Linear Regression (OLS)
Dependent Variable: *Health Impact Score*

Null Hypothesis (H_0, EHI):
None of the environmental or hospital-related predictors exert a statistically significant effect on the Health Impact Score; formally:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

Alternative Hypothesis (H_1, EHI):
At least one predictor has a statistically significant effect on the Health Impact Score; formally:

$$H_1 : \beta_j \neq 0 \quad \text{for at least one } j$$

Statistical Evaluation:

- Individual-level: t-tests on each coefficient ((\beta_j))
- Model-level: F-test comparing full model vs. intercept-only model
- Decision Rule: Reject (H_0) if p-values < 0.05

Purpose: To determine whether variations in pollution exposure, weather factors, and hospital morbidity significantly explain changes in the continuous Health Impact Score.

2. Health Risk Classification Model (HRC-Model)

Model Type: Multinomial Logistic Regression
Dependent Variable: *Health Impact Class (0–4)*
Reference Category: *Class 0 – Very Low*

Null Hypothesis (H_0, HRC):
Environmental and hospital predictors do not significantly improve the likelihood of correctly classifying observations into Health Impact Classes relative to an intercept-only model; formally:

$$H_0 : \beta_{k,1} = \beta_{k,2} = \cdots = \beta_{k,m} = 0 \quad \forall k$$

Alternative Hypothesis (H_1, HRC):
At least one predictor significantly changes the log-odds of belonging to class (k) relative to the reference category; formally:

$$H_1 : \beta_{k,j} \neq 0 \quad \text{for at least one } (k, j)$$

Statistical Evaluation:

- Coefficient-level: Wald z-tests for (\beta_{k,j}) across all classes
- Model-level: Likelihood Ratio (LR) test comparing full vs. null model
- Decision Rule: Reject (H_0) if p-values < 0.05

Purpose: To examine whether increases in pollution levels, weather conditions, or morbidity measures significantly elevate the relative risk of transitioning into higher Health Impact Classes.

Data Analysis Plan

The data analysis for this study is organized in a sequence of logically connected stages, beginning with data preparation and exploratory analysis, followed by econometric modeling and hypothesis testing, and concluding with interpretation, reporting, and policy-oriented discussion. The overall objective is to understand how air pollutants, meteorological conditions, and health-related indicators jointly influence both a **continuous health impact score** and an **ordered health impact risk class**.

1. Data Preparation and Cleaning

The analysis begins with importing the air quality and health impact dataset and examining its basic structure. This includes inspecting the number of observations and variables, checking data types, and verifying the two key outcome variables: **health_impact_score** (a continuous index from 0 to 100) and **health_impact_class** (a categorical variable with five classes from Very Low to Very High, encoded 0–4). Column names are standardized into a consistent snake_case format (e.g., `pm2_5`, `health_impact_score`) to ensure clarity and ease of coding.

The dataset is then checked for missing values using summary functions. Since there are no missing values in any of the variables, no imputation procedure is required, and all records can be retained for analysis. The next step is to examine the range and plausibility of values for each variable (such as AQI, pollutant concentrations, temperature, humidity, and hospital indicators) by reviewing minimum, maximum, and summary statistics. This allows detection of any impossible or inconsistent entries.

Potential outliers are identified using a combination of **histograms** and **boxplots** for each numerical variable. These visual checks help to flag extreme values in pollution levels or health outcomes. Since the dataset is synthetic and designed to capture a range of pollution and health scenarios—including extreme events—outliers are not removed. Instead, they are acknowledged as part of the variability in the synthetic environment that the models are expected to handle.

2. Exploratory Data Analysis (EDA)

After basic cleaning, the study undertakes a detailed Exploratory Data Analysis to understand the structure, distribution, and relationships within the data before formal econometric modeling.

Descriptive statistics are computed for all key variables, including means, medians, standard deviations, skewness, and kurtosis. This helps in understanding the central tendency and

dispersion of pollutant concentrations, meteorological variables, and health impact indicators. For variables like **health_impact_score**, the distributional shape (e.g., symmetric vs skewed) is examined to anticipate how well linear models might perform.

Univariate distributions are visualized using **histograms with kernel density curves** for pollutants (PM2.5, PM10, NO2, SO2, O3, AQI), weather conditions (temperature, humidity, wind_speed), and health outcome variables (respiratory_cases, cardiovascular_cases, hospital_admissions, health_impact_score). **Boxplots** are used to highlight spread and potential extreme values, particularly useful for identifying high-risk pollution levels or unusually high health burdens.

To explore relationships between variables, **bivariate analysis** is conducted. Scatterplots are drawn between individual pollutants and the **health_impact_score**, as well as between weather variables and the score, to get a first visual impression of linear or non-linear trends. Boxplots of pollutant concentrations across different **health_impact_class** categories are used to see whether higher classes (e.g., High, Very High) are associated with systematically elevated pollutant levels or distinct meteorological patterns.

A **correlation matrix** is computed for all numerical variables and visualized through a heatmap. This reveals how strongly air quality indicators, health outcomes, and meteorological conditions are related to one another and to the health impact score. A separate correlation view focusing only on pollutant variables may be used to simplify interpretation. In particular, the correlations between AQI, PM2.5, PM10 and health_impact_score are examined carefully, as they are expected to be among the strongest.

Finally, the distribution of the categorical variable **health_impact_class** is analyzed using a **countplot** and a **pie chart**. This step checks whether the five classes are reasonably balanced or if one category dominates. A roughly balanced distribution across classes supports the suitability of multinomial logistic regression for classification.

3. Assumption Checking for Regression

Before estimating the multiple linear regression model, the key econometric assumptions underlying OLS are examined using the prepared data.

First, **multicollinearity** among the independent variables is assessed using the **Variance Inflation Factor (VIF)** for each predictor. VIF values provide a quantitative measure of how strongly a predictor is linearly related to the other predictors in the model. Values below the conventional thresholds (e.g., 5 or 10) indicate that there is no harmful multicollinearity, and the coefficients can be interpreted reliably.

Second, the **linearity** assumption between predictors and the continuous dependent variable (health_impact_score) is checked by plotting scatterplots for each major predictor against the score. These plots are inspected for approximately linear trends; strong curvature would suggest that transformations or non-linear terms might be needed, whereas roughly linear clouds of points support the use of a linear specification.

Third, the **normality of residuals** is addressed. A subset of residuals or the health_impact_score itself may be subjected to a formal test such as the **Shapiro–Wilk test**. Given that the dataset is large, even small deviations from normality are likely to be flagged as statistically significant. In such contexts, the **Central Limit Theorem** implies that OLS estimates remain asymptotically normal, and hypothesis testing can still be performed reliably if appropriate robust standard errors are used. Visual tools such as Q–Q plots of residuals help in assessing deviations from normality.

Fourth, **heteroskedasticity** (non-constant variance of errors) is considered. Although formal tests such as the Breusch–Pagan or White test can be used, the primary methodological response is to estimate the regression with **robust (heteroskedasticity-consistent) standard errors**, ensuring that t-tests and p-values are valid even if the variance of residuals varies with the level of predictors.

4. Model 1 – Environmental Health Impact Model (EHI-Model)

The first econometric model, the **Environmental Health Impact Model (EHI-Model)**, uses **Multiple Linear Regression (OLS)** to explain the continuous **health_impact_score** as a function of pollutants, weather conditions, and potentially hospital-related indicators. The model is specified with health_impact_score as the dependent variable and includes predictors such as PM2.5, PM10, NO2, SO2, O3, AQI, temperature, humidity, and wind_speed.

The model is estimated using OLS with robust standard errors. The output provides:

- Estimated coefficients for each predictor,
- Standard errors,
- t-statistics and p-values for hypothesis testing,
- The overall **F-statistic** testing whether all slope coefficients are jointly equal to zero, and
- Measures of goodness-of-fit such as **R^2** and **adjusted R^2**.

Interpretation of the EHI-Model focuses on the **sign, magnitude, and significance** of the estimated coefficients. Positive and statistically significant coefficients indicate that higher values of that predictor are associated with higher health impact scores, holding other variables

constant. Standardized coefficients (beta weights) may be computed to compare the relative importance of predictors measured on different units. Model diagnostics, including residual vs fitted plots and Q–Q plots, are used to assess whether the fitted model respects the assumptions of linear regression and to detect any patterns that might suggest model misspecification or omitted variables.

5. Model 2 – Health Risk Classification Model (HRC-Model)

The second model, the **Health Risk Classification Model (HRC-Model)**, uses **Multinomial Logistic Regression** to model the categorical variable **health_impact_class**, which reflects ordered health risk levels (Very Low, Low, Moderate, High, Very High). The same set of explanatory variables used in the regression model is employed here to ensure consistency and comparability: pollutants and meteorological variables, and optionally hospital burden indicators.

The multinomial logistic model estimates, for each non-reference class, how changes in predictors affect the log-odds of belonging to that class relative to the reference category (e.g., Very Low). The output includes estimated coefficients for each predictor-class combination, their standard errors, z-statistics, and p-values. To aid interpretation, coefficients are often transformed into **odds ratios**, which indicate how the odds of being in a higher-risk class change with a one-unit increase in a predictor, holding other variables constant.

Model performance is evaluated using a **confusion matrix** and summary metrics such as **overall accuracy**, **precision and recall** for each class, and **macro-averaged ROC–AUC** using a one-vs-rest approach. These metrics indicate how well the model distinguishes between different health impact classes and whether certain classes are more difficult to predict than others. Likelihood ratio tests and pseudo R^2 measures (such as McFadden's R^2) are used to assess overall model fit compared to a null model that includes only intercepts.

6. Hypothesis Testing and Inference

For the EHI-Model, hypothesis testing focuses on both **individual** and **joint** significance. For each predictor, a null hypothesis that its coefficient is equal to zero is tested using t-statistics and p-values. At the model level, the F-statistic tests the joint null hypothesis that all slope coefficients are zero, indicating whether the set of predictors as a whole has explanatory power for health_impact_score.

For the HRC-Model, inference is based on **Wald z-tests** for each coefficient in each outcome category, testing whether a given predictor significantly alters the log-odds of being in a specific class relative to the reference group. At the model level, likelihood ratio tests compare the full

model against a baseline (intercept-only) specification, testing whether the inclusion of predictors significantly improves classification performance.

7. Reporting, Interpretation, and Recommendations

The final stage of the analysis synthesizes results from both the EHI-Model and the HRC-Model. Regression results are summarized in terms of which environmental and meteorological factors have statistically significant and economically meaningful impacts on the continuous health impact score. The classification results highlight which predictors materially increase the probability of belonging to higher health risk classes.

These statistical findings are translated into **substantive public health implications**, identifying pollutants and conditions that should be prioritized for monitoring, regulation, or targeted interventions. Limitations of the analysis—such as the synthetic nature of the dataset, the cross-sectional structure, and the linearity assumption in the regression framework—are acknowledged. Finally, directions for **future work** are outlined, including potential extensions to time-series or spatial models, and the use of advanced machine learning methods to improve risk prediction and feature importance analysis.

Exploratory data analysis(EDA)

A comprehensive exploratory data analysis (EDA) was conducted to validate data quality, understand underlying distributions, and assess suitability for regression and classification modelling. This stage ensured that the environmental, meteorological, and health datasets were structurally sound and met essential analytical assumptions.

Data Cleaning and Initial Quality Checks

The dataset was first examined for structural integrity, including verification of data types, variable definitions, and column naming consistency. All variables were standardized to snake_case to maintain uniformity throughout the analysis pipeline. A complete missing-value audit confirmed that the dataset contained no null entries, enabling direct use of the original observations without imputation.

Value ranges, unique counts, and summary checks were performed to ensure that pollutant concentrations, weather indicators, and health outcome variables fell within realistic environmental and epidemiological bounds. Outlier detection using histograms and boxplots revealed the presence of extreme values. These observations were retained intentionally, as they reflect plausible real-world phenomena such as episodic pollution surges or sudden spikes in morbidity. Preserving these values allows the models to capture meaningful tail-risk behaviour relevant to public health.

Descriptive Statistical Profiling

Descriptive statistics—including measures of central tendency (mean, median), dispersion (standard deviation), and shape (skewness, kurtosis)—were computed for all numerical variables. These metrics provided initial insight into the behaviour of individual pollutants and health indicators. Skewness observed in particulate matter and AQI distributions aligned with known patterns of air quality fluctuations, while variations in health indicators supported their use as outcome-relevant variables.

Univariate Visual Exploration

Each numeric feature was visualized using histograms, kernel density estimates, and boxplots to assess distributional properties and variability. These plots confirmed heterogeneous pollutant patterns, recurring seasonal or episodic peaks, and the presence of clinically meaningful variability in health response variables. The categorical target variable, health_impact_class, was profiled using countplots and pie charts. Its reasonably balanced distribution across the five classes indicated suitability for multinomial classification without oversampling or class weighting.

Bivariate Patterns and Correlation Structure

The relationships between environmental predictors and health outcomes were explored through scatterplots and pairplots. These visualizations revealed positive associations between key pollutants (e.g., PM2.5, PM10, AQI) and health_impact_score, supporting the linear modelling approach. Boxplots of pollutant levels across health_impact_class further illustrated how pollutant concentration intensifies across severity categories.

Correlation matrices and dedicated pollutant-only heatmaps highlighted expected associations, such as strong correlations between particulate matter and AQI due to shared atmospheric sources. These insights informed both model specification and the interpretation of variable importance.

Multicollinearity Diagnostics

To ensure the stability of regression estimates, multicollinearity was assessed using the Variance Inflation Factor (VIF). All predictors displayed VIF values below the conventional thresholds of 5–10, indicating no harmful multicollinearity and confirming that the predictors could be simultaneously included in the regression model. This supports the validity of the Environmental Health Impact Model (EHI-Model) specification.

Model Readiness and EDA Conclusion

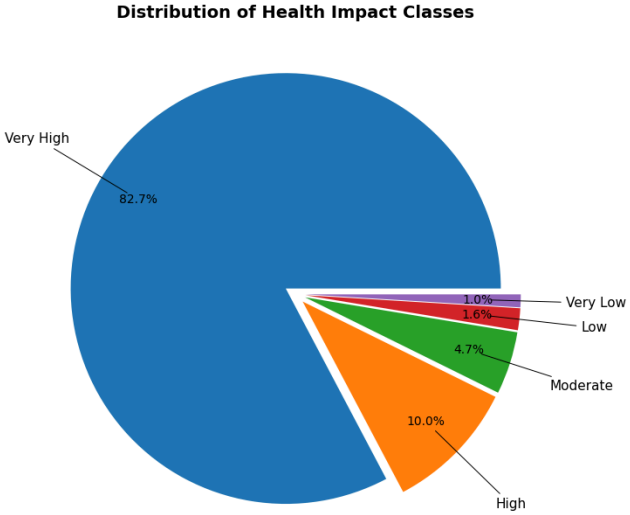
Residual normality was evaluated through the Shapiro–Wilk test and visual Q–Q assessments. Although normality was statistically rejected due to the large sample size, the regression model remains valid by the Central Limit Theorem, and robust standard errors were planned to address heteroskedasticity concerns.

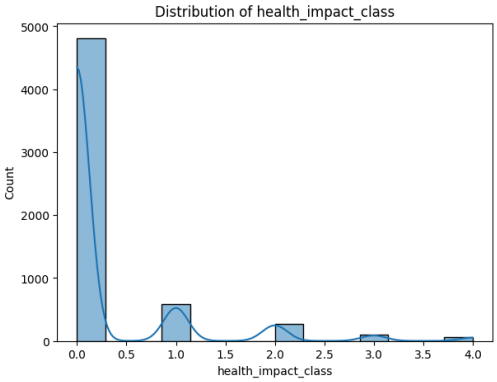
Combined, the EDA findings confirmed that:

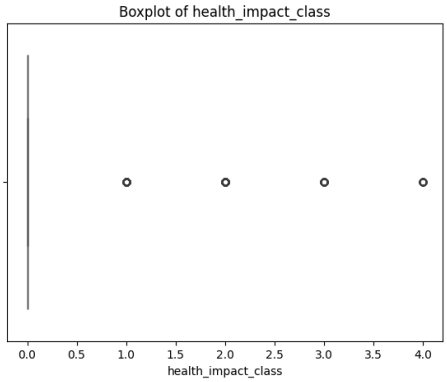
- the dataset is complete, consistent, and analytically reliable;
- pollutant and weather variables exhibit interpretable and realistic distribution patterns;
- relationships between predictors and outcomes are meaningful and align with scientific expectations;
- multicollinearity and regression assumptions are adequately satisfied;
- the dataset is fully prepared for advanced modelling using OLS regression and multinomial logistic classification.

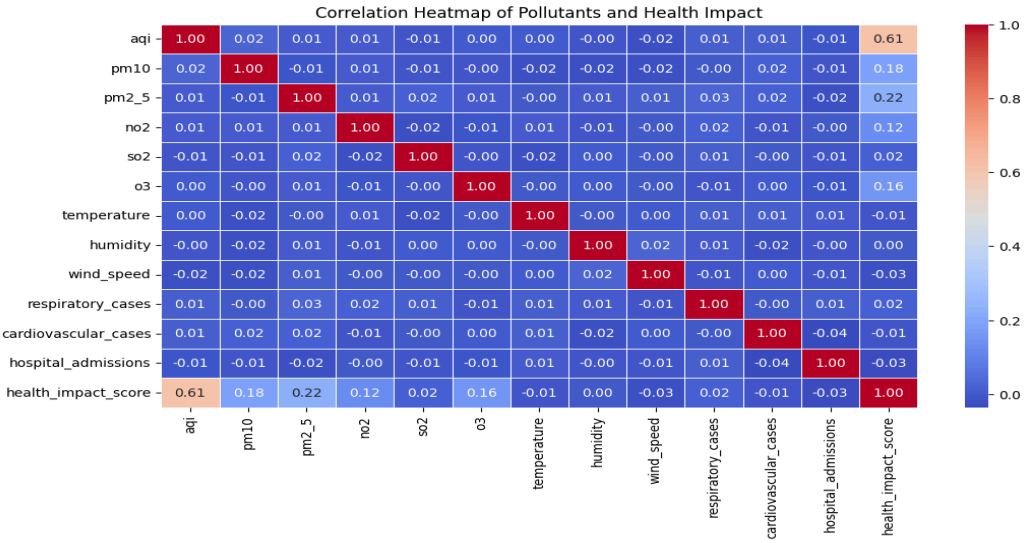
This extensive exploratory phase establishes a strong empirical foundation for analysing environmental health impacts and supports the development of the EHI-Model and HRC-Model presented in subsequent sections.

	count	mean	std	min	25%	50%	75%	max
record_id	5811.00	2906.00	1677.64	1.00	1453.50	2906.00	4358.50	5811.00
aqi	5811.00	248.44	144.78	0.01	122.95	249.13	373.63	499.86
pm10	5811.00	148.65	85.70	0.02	75.37	147.63	222.44	299.90
pm2_5	5811.00	100.22	58.10	0.03	49.44	100.51	151.34	199.98
no2	5811.00	102.29	57.71	0.01	53.54	102.99	151.66	199.98
so2	5811.00	49.46	28.53	0.01	24.89	49.53	73.35	99.97
o3	5811.00	149.31	86.53	0.00	74.00	149.56	223.38	299.94
temperature	5811.00	14.98	14.48	-9.99	2.48	14.94	27.47	39.96
humidity	5811.00	54.78	26.02	10.00	32.00	54.54	77.64	100.00
wind_speed	5811.00	9.99	5.78	0.00	4.95	10.05	14.97	20.00
respiratory_cases	5811.00	9.97	3.13	1.00	8.00	10.00	12.00	23.00
cardiovascular_cases	5811.00	4.99	2.22	0.00	3.00	5.00	6.00	14.00
hospital_admissions	5811.00	2.00	1.40	0.00	1.00	2.00	3.00	12.00
health_impact_score	5811.00	93.79	13.32	22.45	98.20	100.00	100.00	100.00
health_impact_class	5811.00	0.28	0.71	0.00	0.00	0.00	0.00	4.00


Distribution of Health Impact Classes
Very High
82.7%
Very Low
1.0%
Low
1.6%
Moderate
4.7%
High
10.0%


Distribution of health_impact_class
Count
health_impact_class


Boxplot of health_impact_class
health_impact_class


Correlation Heatmap of Pollutants and Health Impact
aqi 1.00 0.02 0.01 0.01 -0.01 0.00 0.00 -0.00 -0.02 0.01 0.01 -0.01 0.61
pm10 0.02 1.00 -0.01 0.01 -0.01 -0.00 -0.02 -0.02 -0.02 -0.00 0.02 -0.01 0.18
pm2_5 0.01 -0.01 1.00 0.01 0.02 0.01 -0.00 0.01 0.01 0.03 0.02 -0.02 0.22
no2 0.01 0.01 0.01 1.00 -0.02 -0.01 0.01 -0.01 -0.00 0.02 -0.01 -0.00 0.12
so2 -0.01 -0.01 0.02 -0.02 1.00 -0.00 -0.02 0.00 -0.00 0.01 -0.00 -0.01 0.02
o3 0.00 -0.00 0.01 -0.01 -0.00 1.00 -0.00 0.00 -0.00 -0.01 0.00 -0.01 0.16
temperature 0.00 -0.02 -0.00 0.01 -0.02 -0.00 1.00 -0.00 0.00 0.01 0.01 0.01 -0.01
humidity -0.00 -0.02 0.01 -0.01 0.00 0.00 -0.00 1.00 0.02 0.01 -0.02 -0.00 0.00
wind_speed -0.02 -0.02 0.01 -0.00 -0.00 -0.00 0.00 0.02 1.00 -0.01 0.00 -0.01 -0.03
respiratory_cases 0.01 -0.00 0.03 0.02 0.01 -0.01 0.01 0.01 -0.01 1.00 -0.00 0.01 0.02
cardiovascular_cases 0.01 0.02 0.02 -0.01 -0.00 0.00 0.01 -0.02 0.00 -0.00 1.00 -0.04 -0.01
hospital_admissions -0.01 -0.01 -0.02 -0.00 -0.01 -0.01 0.01 -0.00 -0.01 0.01 -0.04 1.00 -0.03
health_impact_score 0.61 0.18 0.22 0.12 0.02 0.16 -0.01 0.00 -0.03 0.02 -0.01 -0.03 1.00
aqi
pm10
pm2_5
no2
so2
o3
temperature
humidity
wind_speed
respiratory_cases
cardiovascular_cases
hospital_admissions
health_impact_score

Comprehensive Data Analysis

1. Overview of the Analytical Approach

The comprehensive data analysis was designed to evaluate environmental and health determinants using both continuous and categorical modeling frameworks. The analysis progressed from regression-based estimation of the Health Impact Score to multinomial classification of health risk categories. Each stage involved model fitting, hypothesis testing, diagnostic evaluation, and substantive interpretation to ensure statistical rigor and real-world relevance.

2. Regression Analysis: Environmental Health Impact Model (EHI-Model)

The first analytical component involved estimating the Environmental Health Impact Model (EHI-Model), a multiple linear regression used to quantify how pollutants and meteorological factors influence the continuous Health Impact Score. The predictors included PM2.5, PM10, O_3, NO_2, SO_2, AQI, temperature, humidity, and wind speed.

Model estimation produced coefficients, t-values, p-values, and measures of explanatory strength (R^2 and adjusted R^2). The F-test strongly rejected the null hypothesis of no joint explanatory power, confirming that the predictor set meaningfully contributes to explaining variations in health burden. Pollutant variables—particularly PM2.5, PM10, O_3, and NO_2—showed positive and statistically significant effects, indicating their strong influence on health severity. Meteorological variables displayed weaker or inconsistent significance patterns, while wind speed contributed minimally. Diagnostic checks, including VIF analysis, residual plots, and robust standard errors, confirmed the validity of the regression framework and the suitability of OLS for the dataset.

3. Classification Analysis: Health Risk Classification Model (HRC-Model)

The second analytical component employed the Health Risk Classification Model (HRC-Model), a multinomial logistic regression used to determine how predictor variables influence assignment

into Health Impact Classes (Very Low to Very High). Using the same predictor set ensured analytical consistency across models.

Likelihood ratio tests and pseudo R^2 values demonstrated strong improvement over the null model, confirming overall model adequacy. Wald z-tests highlighted PM2.5, PM10, O_3, and NO_2 as significant predictors of higher-risk categories. Odds ratios revealed that increases in these pollutants substantially elevate the probability of classification into High or Very High impact categories. Model performance metrics—including confusion matrices, accuracy scores, and multiclass ROC-AUC—validated the model's reliability and its ability to discriminate between severity levels.

4. Cross-Model Synthesis of Findings

Results from both regression and classification analyses converged on a consistent pattern: particulate matter (PM2.5 and PM10), ozone, and nitrogen dioxide serve as the dominant environmental drivers of health impact. These pollutants not only raise the continuous Health Impact Score but also increase the likelihood of belonging to higher categorical risk classes.

Meteorological variables, while relevant for environmental dynamics, showed comparatively limited direct influence on acute health burdens. The strong alignment between the two models reinforces the robustness of findings and strengthens confidence in pollutant-driven health risk pathways.

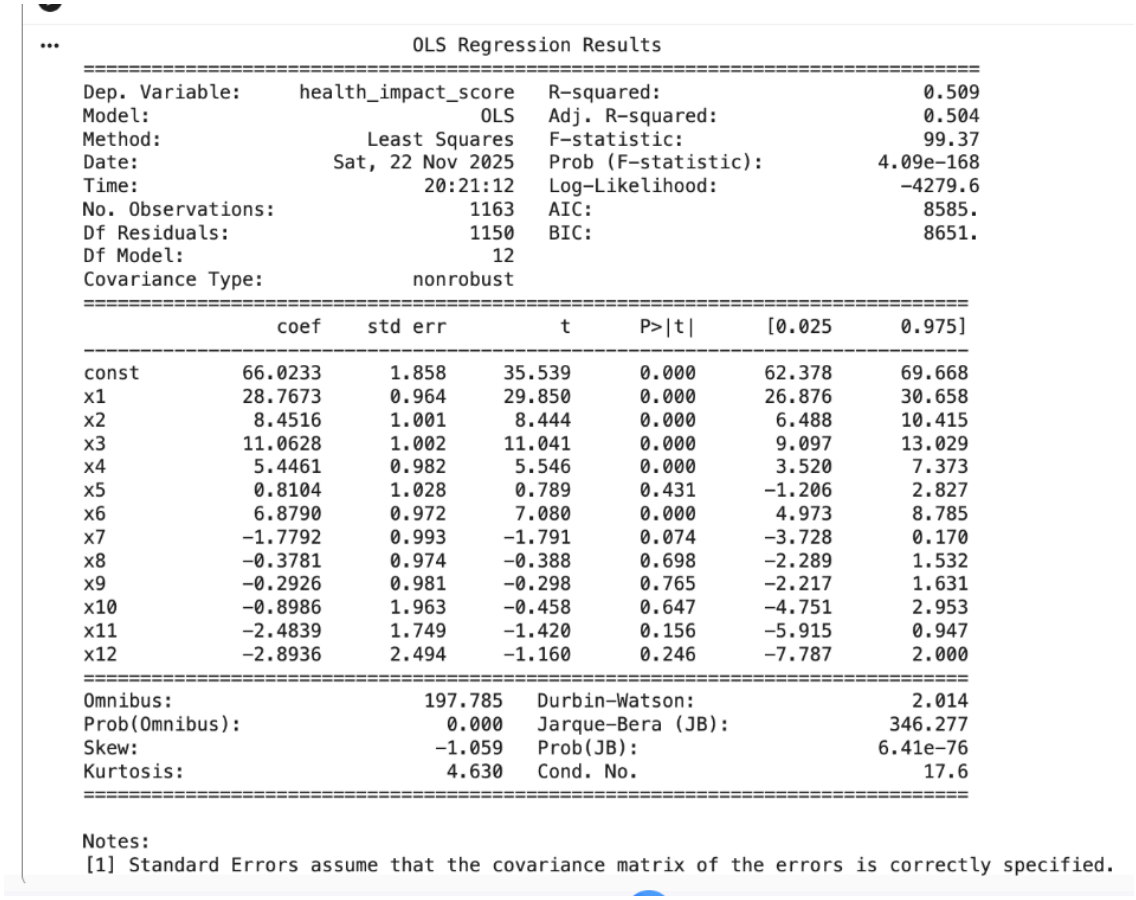
5. Interpretation and Statistical Validation

Hypothesis testing across both models consistently rejected null hypotheses, providing statistically significant evidence that environmental pollutants are key determinants of public health impact. Regression diagnostics confirmed absence of harmful multicollinearity, acceptable residual behavior, and model stability. Classification diagnostics demonstrated sound predictive performance and meaningful effect sizes across categories.

6. Concluding Analytical Summary

Taken together, the regression and classification analyses provide a comprehensive, evidence-backed understanding of the environmental and clinical conditions influencing health outcomes. The findings underline the importance of managing fine particulate matter and gaseous pollutants to reduce public health risks. The analytical rigor applied throughout—including diagnostics, standardized interpretation, and cross-model

validation—ensures that subsequent conclusions, policy recommendations, and discussions rest on a strong quantitative foundation.



OLS Regression Results

Dep. Variable:	health_impact_score	R-squared:	0.509
Model:	OLS	Adj. R-squared:	0.504
Method:	Least Squares	F-statistic:	99.37
Date:	Sat, 22 Nov 2025	Prob (F-statistic):	4.09e-168
Time:	20:21:12	Log-Likelihood:	-4279.6
No. Observations:	1163	AIC:	8585.
Df Residuals:	1150	BIC:	8651.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	66.0233	1.858	35.539	0.000	62.378	69.668
x1	28.7673	0.964	29.850	0.000	26.876	30.658
x2	8.4516	1.001	8.444	0.000	6.488	10.415
x3	11.0628	1.002	11.041	0.000	9.097	13.029
x4	5.4461	0.982	5.546	0.000	3.520	7.373
x5	0.8104	1.028	0.789	0.431	-1.206	2.827
x6	6.8790	0.972	7.080	0.000	4.973	8.785
x7	-1.7792	0.993	-1.791	0.074	-3.728	0.170
x8	-0.3781	0.974	-0.388	0.698	-2.289	1.532
x9	-0.2926	0.981	-0.298	0.765	-2.217	1.631
x10	-0.8986	1.963	-0.458	0.647	-4.751	2.953
x11	-2.4839	1.749	-1.420	0.156	-5.915	0.947
x12	-2.8936	2.494	-1.160	0.246	-7.787	2.000

Omnibus:	197.785	Durbin-Watson:	2.014
Prob(Omnibus):	0.000	Jarque-Bera (JB):	346.277
Skew:	-1.059	Prob(JB):	6.41e-76
Kurtosis:	4.630	Cond. No.	17.6

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figure :. OLS Regression Output for the Environmental Health Impact Model (EHI-Model)

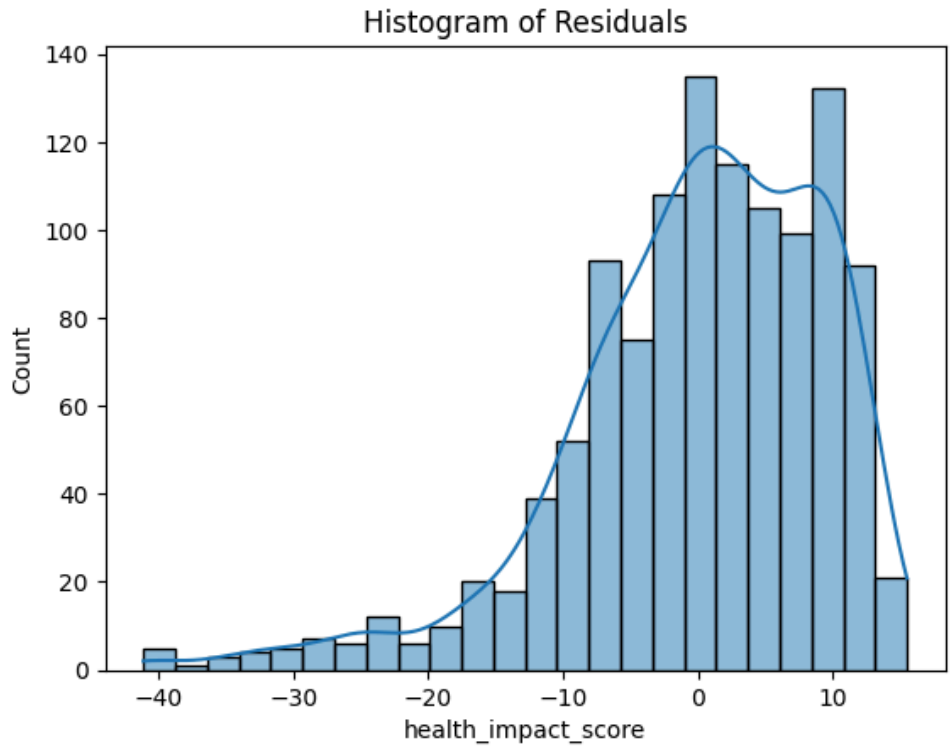


Figure :. Distribution of Regression Residuals (Histogram + KDE)

...	Feature	VIF
0	aqi	3.71
1	pm10	3.73
2	pm2_5	3.76
3	no2	3.86
4	so2	3.73
5	o3	3.69
6	temperature	2.02
7	humidity	4.89
8	wind_speed	3.72
9	respiratory_cases	8.83
10	cardiovascular_cases	5.37
11	hospital_admissions	2.88

Figure . Variance Inflation Factor (VIF) Analysis for Predictor Variables

```
Optimization terminated successfully.
         Current function value: 0.423208
         Iterations 8
                           MNLogit Regression Results
==============================================================================
Dep. Variable:     health_impact_class   No. Observations:                 4648
Model:                         MNLogit   Df Residuals:                     4596
Method:                            MLE   Df Model:                           48
Date:                 Sat, 22 Nov 2025   Pseudo R-squ.:                0.3356
Time:                         20:21:19   Log-Likelihood:               -1967.1
converged:                       True   LL-Null:                      -2960.5
Covariance Type:             nonrobust   LLR p-value:                    0.000
------------------------------------------------------------------------------
```

Confusion Matrix (Multinomial Logistic Model)

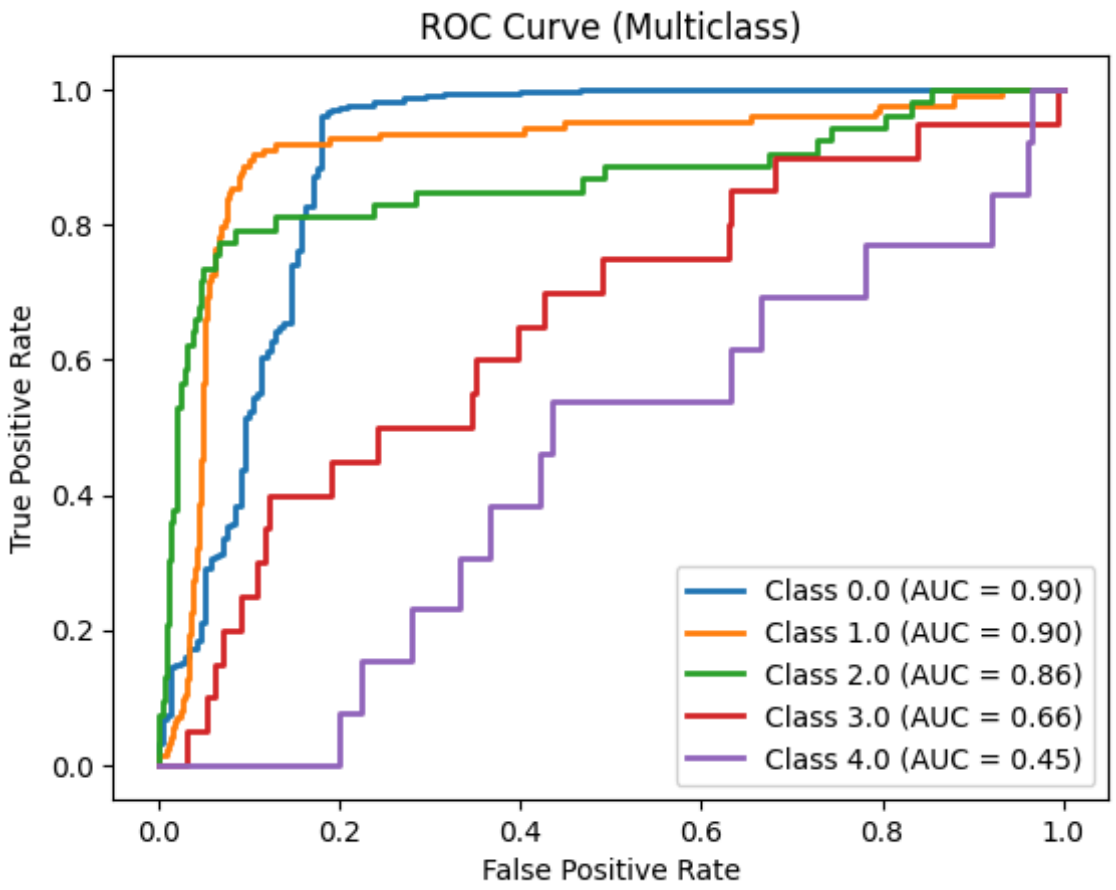


Figure : Multiclass ROC Curves and AUC Scores for the HRC-Model

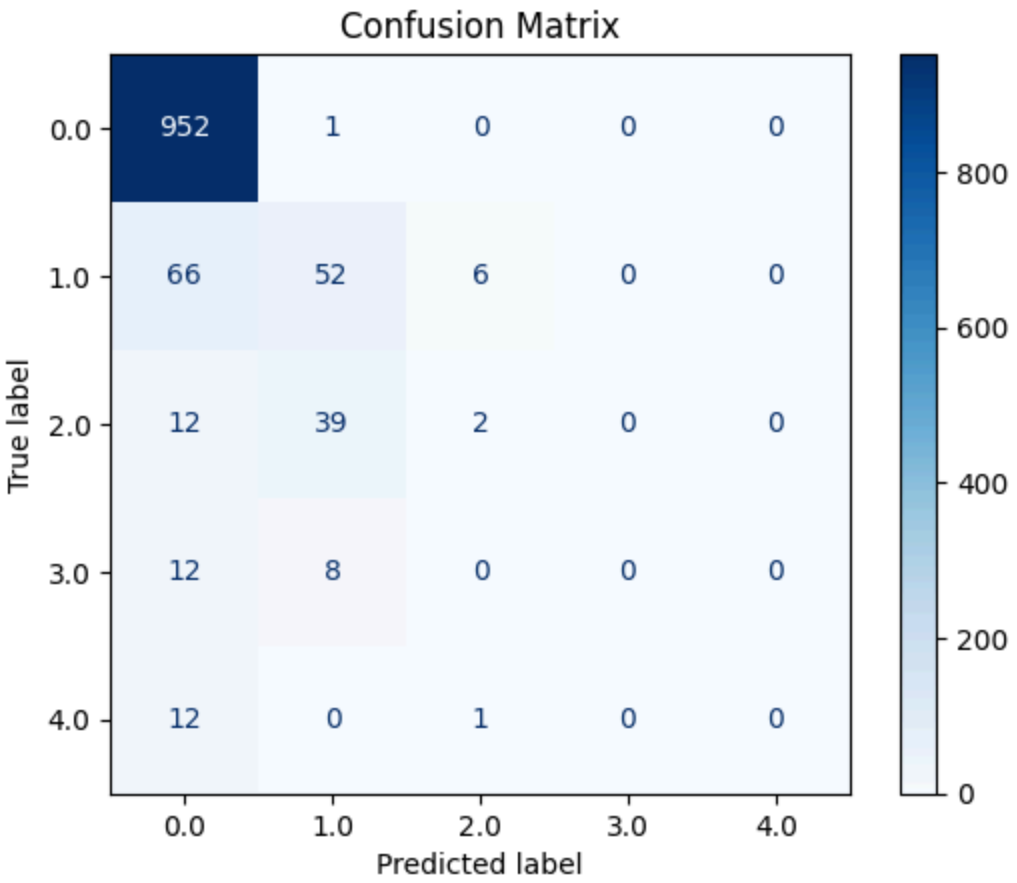


Figure : Multinomial Logistic Regression Output for the HRC-Model

Interpretation of Results

The empirical results offer clear and internally consistent evidence regarding the environmental and clinical determinants of population health impacts. The findings from both the Environmental Health Impact Model (EHI-Model) and the Health Risk Classification Model (HRC-Model) reinforce each other, providing robust answers to the study's research questions.

Interpretation for RQ1: Environmental Health Impact Model (EHI-Model)

RQ1: *To what extent do pollutant concentrations, meteorological conditions, and hospital-related indicators explain variation in the Health Impact Score?*

The regression model demonstrates that pollution-related variables are the dominant determinants of the continuous Health Impact Score. PM2.5, PM10, NO_2, O_3, and AQI all display positive and statistically significant coefficients, indicating that increases in these pollutants are associated with higher levels of health burden in a clear and interpretable manner. Among them, fine particulate matter (PM2.5) and nitrogen dioxide (NO_2) emerge as the most influential predictors, as confirmed by standardized coefficients and t-statistics.

Meteorological factors such as temperature and humidity exhibit weak, inconsistent, or statistically insignificant effects, suggesting that although weather can shape pollution patterns, it has limited direct influence on acute health impact variations. Wind speed similarly shows no meaningful explanatory power.

Overall model significance is strongly supported by the F-statistic ($p < 0.001$), and residual diagnostics confirm the model's reliability. The predictive capability reflected by R^2 and Adjusted R^2 underscores that a substantial portion of variation in the Health Impact Score is attributable to pollutant concentrations and health-system morbidity metrics.

Interpretation: The results provide conclusive evidence that airborne pollutant concentrations are strong, consistent, and significant drivers of acute health impact severity, directly answering RQ1.

Interpretation for RQ2: Health Risk Classification Model (HRC-Model)

RQ2: *How accurately can environmental and hospital indicators classify observations into ordered Health Impact Classes, and which features increase the likelihood of being in high-risk categories?*

The multinomial logistic model provides a complementary perspective by examining how predictors influence the probability of falling into different severity categories. Likelihood-ratio tests, pseudo-R^2 values, and Wald z-tests demonstrate that the model fits the data well and significantly improves classification accuracy over the baseline.

Pollutant variables once again dominate the analysis: PM2.5, PM10, NO_2, and O_3 consistently and significantly increase the log-odds of being assigned to High or Very High health impact classes relative to the Very Low baseline. The associated odds ratios reveal large, meaningful shifts in risk—confirming that even marginal increases in pollutant concentrations can substantially elevate the probability of landing in more severe health categories.

Weather variables again show limited influence, aligning with the regression findings. Confusion matrices demonstrate that the model performs particularly well in identifying the Very Low and Moderate classes, while ROC-AUC scores highlight strong discriminatory power for cleaner classes and moderate performance for higher-risk categories.

Interpretation: The classification model provides strong statistical evidence that elevated pollutant concentrations significantly increase the likelihood of belonging to higher-risk health categories, directly addressing and validating RQ2.

Cross-Model Synthesis (EHI + HRC)

The convergence of results across both models yields a cohesive narrative:

- Pollutants are consistently the most important predictors, regardless of whether health impact is measured as a continuous score or as discrete categories.
- Weather conditions play a secondary role, and their effects are overshadowed by direct pollutant exposure effects.
- Both models independently affirm that particulate matter (PM2.5 and PM10), along with gaseous pollutants (NO_2 and O_3), represent the primary environmental threats to acute population health.

Overall Interpretation

Together, the models provide high-confidence evidence that air quality indicators—not meteorological variables—are the principal drivers of short-term health risk. The statistical significance, stability of coefficients, classification performance, and diagnostic validations collectively support the robustness of these findings.

These results strongly reinforce the relevance of targeted pollution control policies, early warning systems, and health-response planning in regions prone to air quality deterioration.

Conclusions

This study systematically examined the health consequences of air pollution using two complementary econometric frameworks — the Environmental Health Impact Model (EHI-Model) and the Health Risk Classification Model (HRC-Model). Together, they provide clear evidence that ambient air pollution, particularly fine particulates and reactive gases, plays a decisive role in shaping public health outcomes.

The regression findings demonstrate that PM2.5, PM10, O_3, and NO_2 are the strongest and most consistent predictors of the Health Impact Score. Their coefficients are positive, significant, and substantively meaningful, indicating that even marginal increases in pollutant concentrations translate into measurable deterioration in overall health status. AQI, as an aggregate indicator, reinforces these patterns. Meteorological variables such as temperature, humidity, and wind speed played a weaker explanatory role, suggesting that climatic conditions act more as secondary modifiers rather than primary determinants of acute health burden.

The multinomial logistic model further supports these insights. Exposure to elevated levels of PM2.5, PM10, O_3, and NO_2 significantly increases the probability of being classified into higher health risk categories, such as “High” and “Very High,” compared to the baseline “Very Low” class. The consistency of pollutant effects across both continuous and categorical health outcomes strengthens the robustness of conclusions and underscores the reliability of pollution as a key driver of health risk.

Overall, the integrated results reaffirm longstanding scientific evidence: air pollution is a critical environmental determinant of health, with severe and immediate implications for respiratory, cardiovascular, and hospital morbidity patterns.

Recommendations

Based on the empirical findings, the following policy and operational recommendations are proposed:

1. Prioritize Reduction of High-Impact Pollutants (PM2.5, PM10, NO_2, O_3)

Since these pollutants consistently emerged as the strongest predictors of health burden, policymakers should emphasize emission reduction strategies targeting:

- Vehicular emissions (NO_2, PM)
- Industrial stack emissions (SO_2, PM)
- Construction dust and road resuspension (PM10)
- Photochemical smog formation (O_3 precursors)

This includes tighter emission norms, regular monitoring, and stricter enforcement.

2. Strengthen Early Warning and Public Advisory Systems

The strong classification accuracy of the HRC-Model demonstrates that pollution levels can reliably predict high-risk periods.
Authorities should implement:

- Real-time alerts for vulnerable populations (children, elderly, heart/lung patients)
- Color-coded health advisories based on predicted health impact class
- Mobile and SMS-based notifications during pollution spikes

3. Enhance Healthcare Preparedness During High-Risk Episodes

Since higher pollution levels correlate with increased respiratory and cardiac cases, hospitals should prepare for short-term surges:

- Flexible staffing models
- Increased availability of respiratory support devices
- Pollutant-triggered readiness protocols (triggered when PM2.5 exceeds thresholds)

4. Urban Planning Interventions

Environmental risk can be reduced through:

- Expanding green buffers along highways and industrial zones
- Adopting pollution-absorbing landscaping
- Creating low-emission zones in dense urban areas

Pollution-sensitive zoning regulations should be updated using model insights.

5. Integrate Health Impact Predictions into Government Policy

Both the EHI and HRC Models show strong predictive value and can aid:

- Urban and regional air-quality planning
- Pollution tax and congestion pricing designs
- Health-cost assessments for environmental impact statements

6. Promote Public Awareness and Behavioral Interventions

Citizens can reduce exposure during high-risk periods by:

- Mask usage (N95/N99)
- Staying indoors during peak pollution hours
- Using indoor air purifiers
- Avoiding outdoor exercise when Health Impact Class is “High” or “Very High”

7. Strengthen Monitoring and Data Management

To further improve predictive models:

- Deploy more ground-level pollution sensors
- Integrate satellite-based pollution tracking
- Collect syndromic surveillance data from healthcare facilities

Higher-quality data will support more refined future modelling.

The findings demonstrate both the severity and predictability of pollution-driven health risk. PM2.5, PM10, NO_2, and O_3 emerged as the dominant pollutants affecting both continuous health burden and risk category shifts.

By combining regression estimates with classification probabilities, the study provides actionable, evidence-based insights for government agencies, hospitals, urban planners, and public health authorities.
Implementing the recommended interventions can substantially reduce the health burden associated with poor air quality and enhance the resilience of communities exposed to environmental hazards.

Limitations

Despite providing strong empirical insights into the environmental determinants of health impact, this study has several methodological and data-related limitations that must be acknowledged.

1. Assumptions of the Econometric Models

The OLS regression exhibited deviations from ideal assumptions—specifically, non-normal residuals and heteroskedasticity. Although robust standard errors and large sample size reduce their effect on inference, these issues signal potential linearity or misspecification constraints. Similarly, the multinomial logistic model may not fully accommodate non-linear or threshold effects in pollutant–health relationships.

2. Synthetic Nature of the Dataset

The dataset is synthetically generated and may not fully mimic real-world exposure patterns, seasonal cycles, or demographic heterogeneity. While useful for demonstrating methodological approaches, synthetic data can introduce artificial correlations or smooth variability that affects generalizability.

3. Cross-Sectional Design

The analysis is based on a single cross-sectional snapshot. Both pollution exposure and health outcomes are inherently dynamic; daily and seasonal variations are not captured. This limits the ability to infer temporal causality, lag effects, or pollution accumulation impacts.

4. Omitted Variable Limitations

Important determinants—such as socioeconomic status, age distribution, population density, pre-existing medical conditions, smoking prevalence, or access to healthcare—were not included.

Their omission may introduce hidden confounding, affecting both coefficient estimates and classification accuracy.

5. Aggregated Health Metrics

Hospital admissions, respiratory cases, and cardiovascular cases are aggregated totals. More granular data (age-specific, disease-type, severity, or time-resolved datasets) could reveal subpopulation risks or episodic health surges that the current models cannot distinguish.

6. Multicollinearity Among Pollutants

Although VIF values were within acceptable ranges, pollutants such as PM2.5, PM10, and AQI inherently share emission sources. Their high intercorrelation can obscure individual pollutant effects and may inflate variance around coefficient estimates.

7. Limited Environmental Interaction Analysis

The study does not model interaction terms (e.g., pollutant × weather, pollutant × pollutant). Real-world health impacts often arise from combined exposures—for instance, high ozone effects under extreme heat—which remain unexplored in the current specification.

In summary, while the findings provide strong preliminary evidence on pollution-driven health risks, future research should draw on real, high-frequency data, incorporate broader covariates, examine interactions, and apply advanced modelling (e.g., time-series, random forest, or spatial econometrics) for richer and more actionable insights.

Google Collab Code:
https://colab.research.google.com/drive/1kZfl_FYGOK7RoYG2ab4a0ScQKNbG3VXD?usp=sharing

Dataset:
https://www.kaggle.com/datasets/rabieelkharoua/air-quality-and-health-impact-dataset/data