DataVerse
Africa

AIR POLLUTION AND RESPIRATORY DISEASE ANALYSIS

Internship Presentation

Presented By: HA - 3

Judd Marc and Olive Mwende

1

Problem Statement

Context

Lagos faces severe air pollution driven by rapid urbanization, industrial activity, diesel generator use, and open waste burning. PM2.5 levels exceed WHO limits by more than five times, creating major public health risks.

Issue

Air quality monitoring is fragmented and not linked with health data, limiting policymakers' ability to predict and prevent respiratory health crises. Hospitals are seeing rising respiratory cases without the benefit of predictive insights.

Impact

Delayed responses are driving higher disease burdens, healthcare costs, and productivity losses. Vulnerable groups—children, the elderly, and those with pre-existing conditions are most affected, while overall socio-economic development is at risk.

2

Project Objectives

Primary Objectives

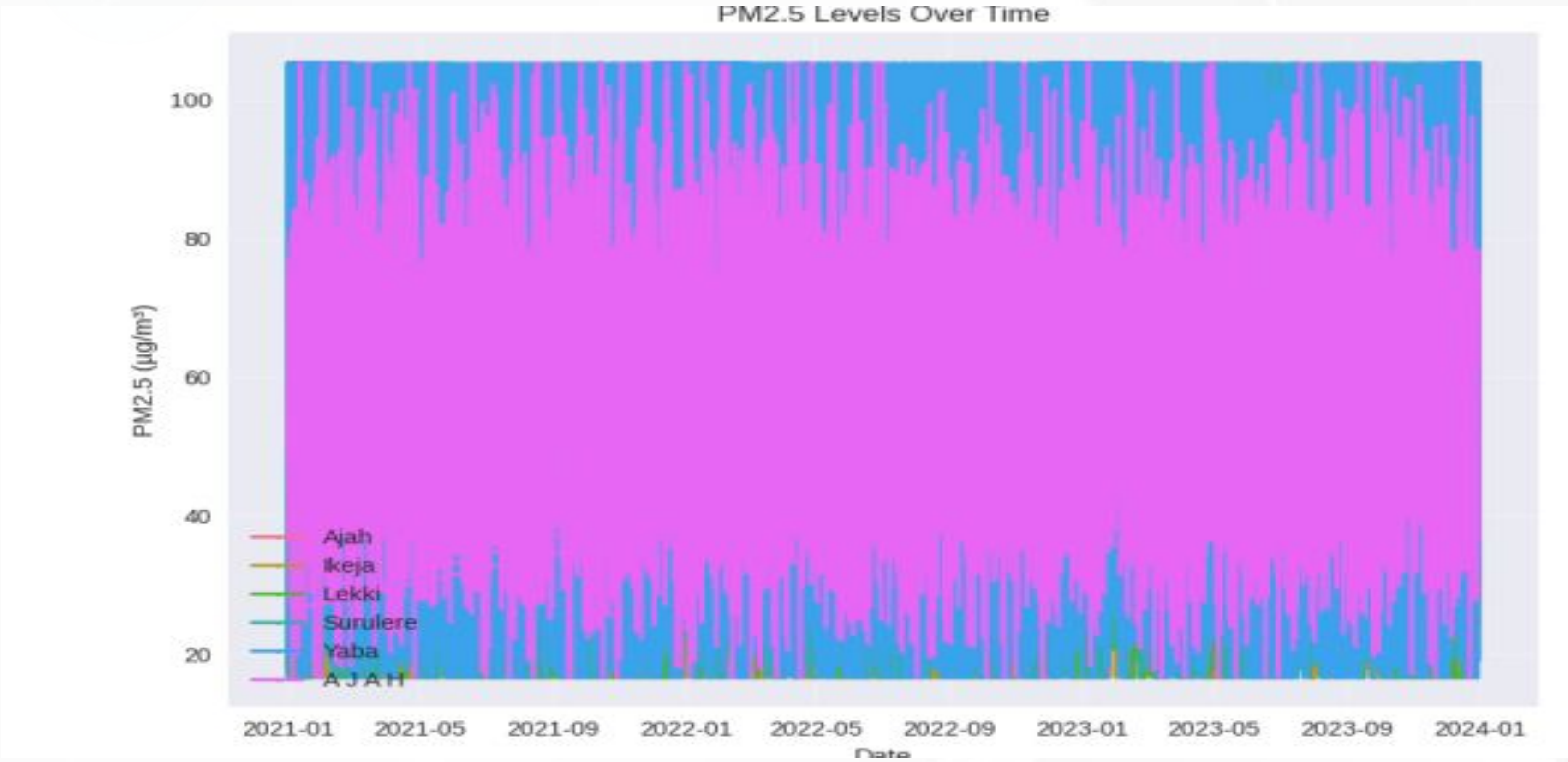
🏥 Analyze the relationship between air pollution spikes and hospital respiratory cases
📈 Predict respiratory disease surges using pollution data
⚠️ Identify high-risk cities, periods, and pollutants
🏛️ Recommend actionable public health and urban environmental policies

Secondary Objectives

🌍 Derive a composite **Pollution Index** from multiple pollutants
📊 Monitor and track trends in air pollution levels over time
☁️ Assess the influence of weather and industrial activity on pollution and health outcome

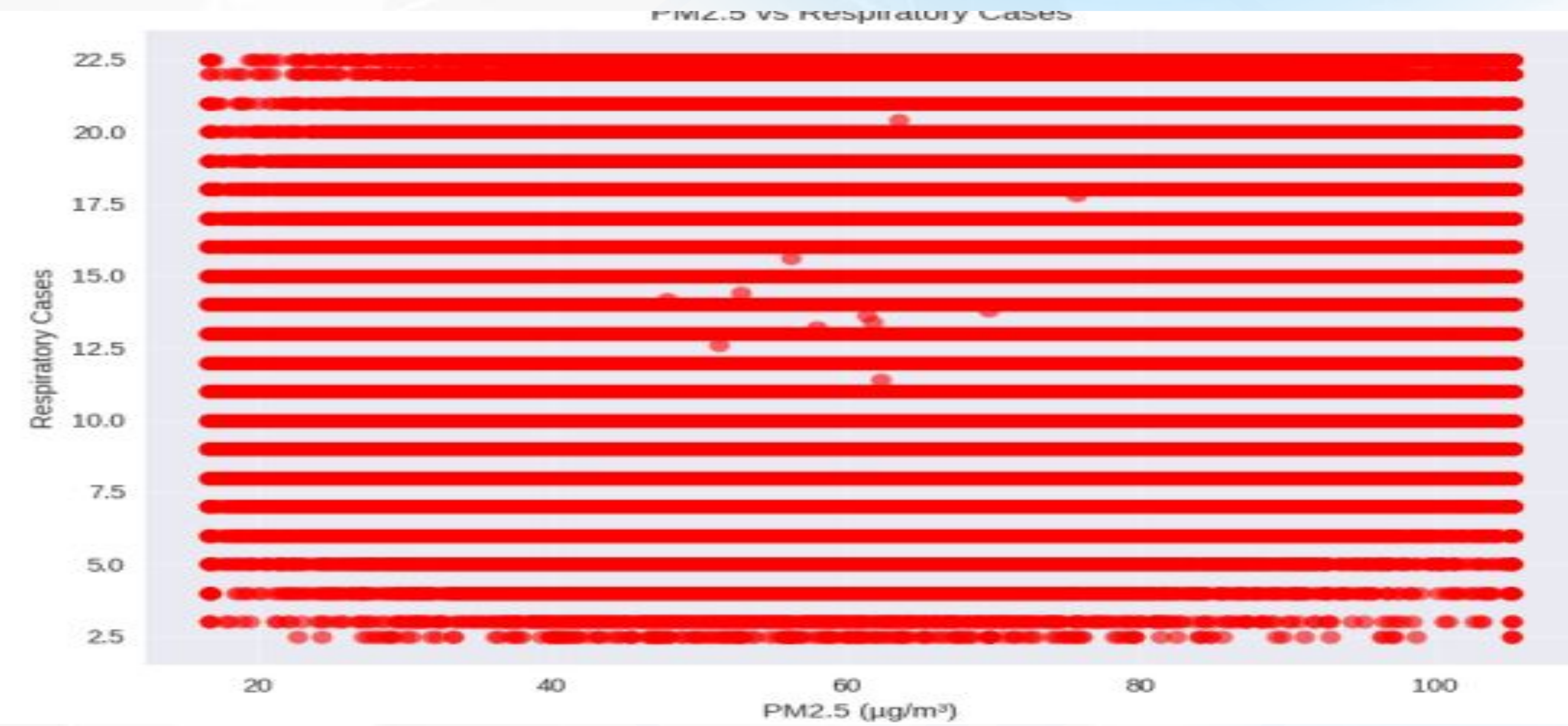
3 Data Insights

PM2.5 levels Over time



- The PM2.5 levels are consistently high, hovering around the 100 µg/m³ mark for most of the period
- Ajah is the most significant contributor to the total PM2.5 levels, as indicated by the large magenta portion of the bars. This suggests it may be a major pollution source

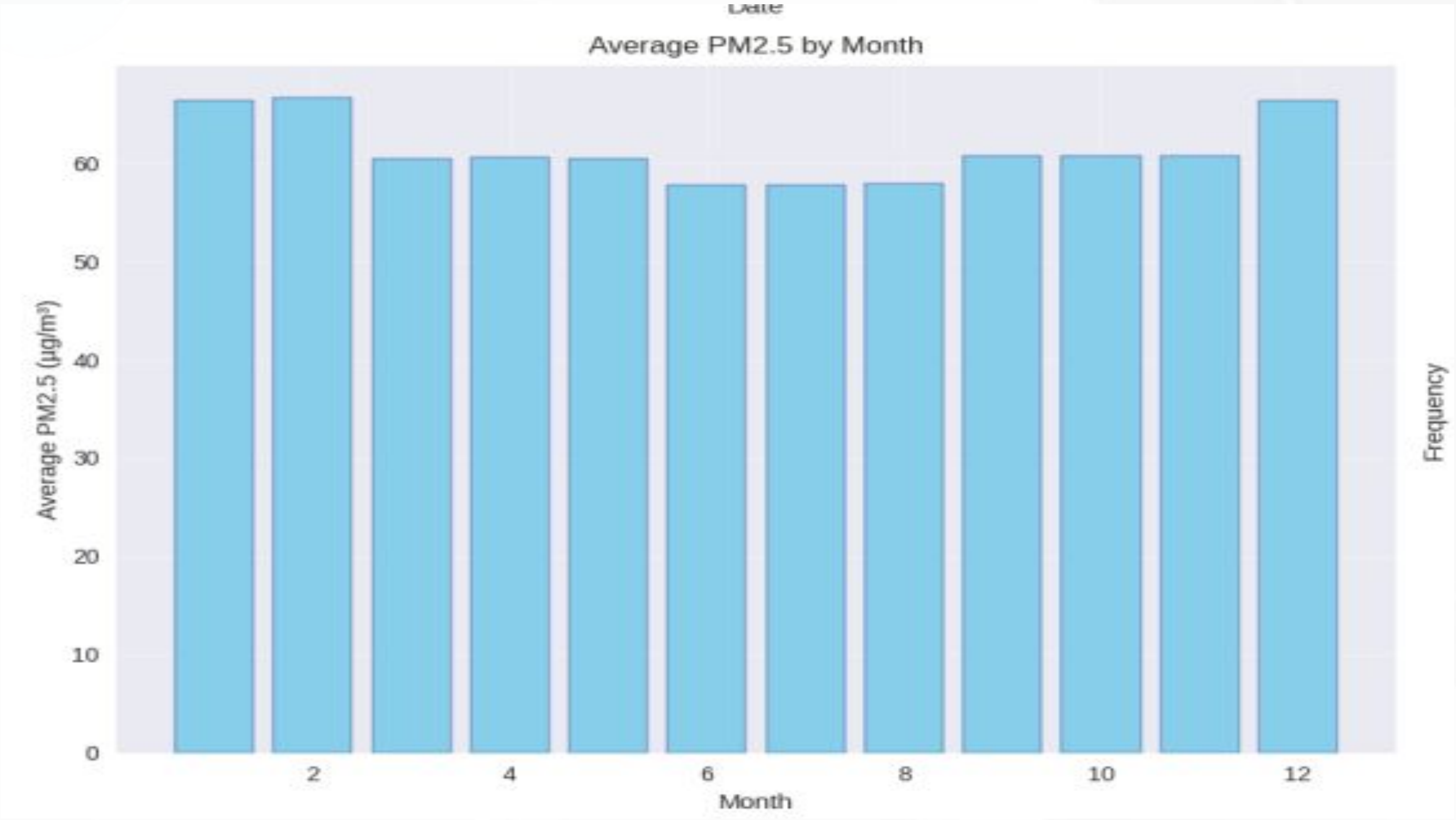
PM2.5 levels vs Respiratory Cases



- There is a strong positive relationship between the two variables. As the PM2.5 levels increase, the number of respiratory cases also increases.

4 Data Insights

Average PM2.5 by month



- PM2.5 levels are highest during the early months of the year (January and February) and again at the end (December).

- Lowest Pollution Months: The lowest average PM2.5 levels occur during the middle of the year, particularly from March to September.

Population index distribution.

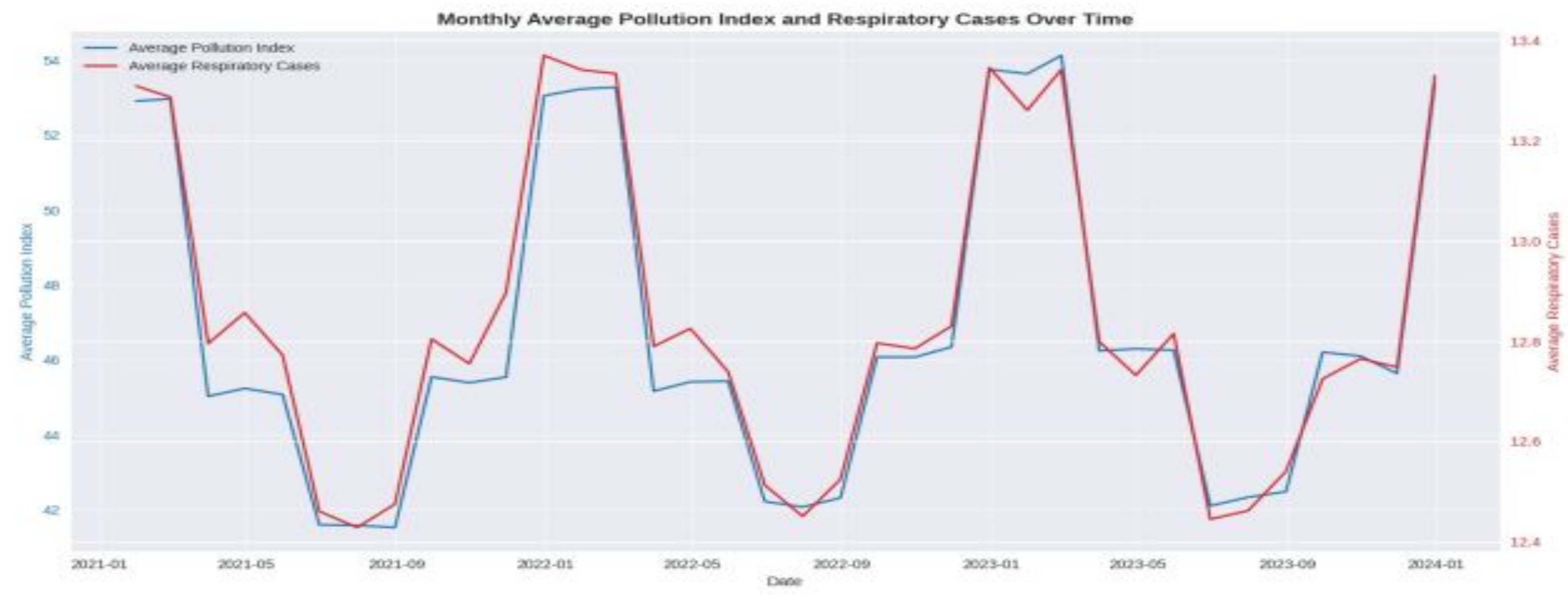
- The majority of the pollution index readings are clustered around the average (the peak of the curve), while readings that are significantly higher or lower than the average occur much less frequently.

5

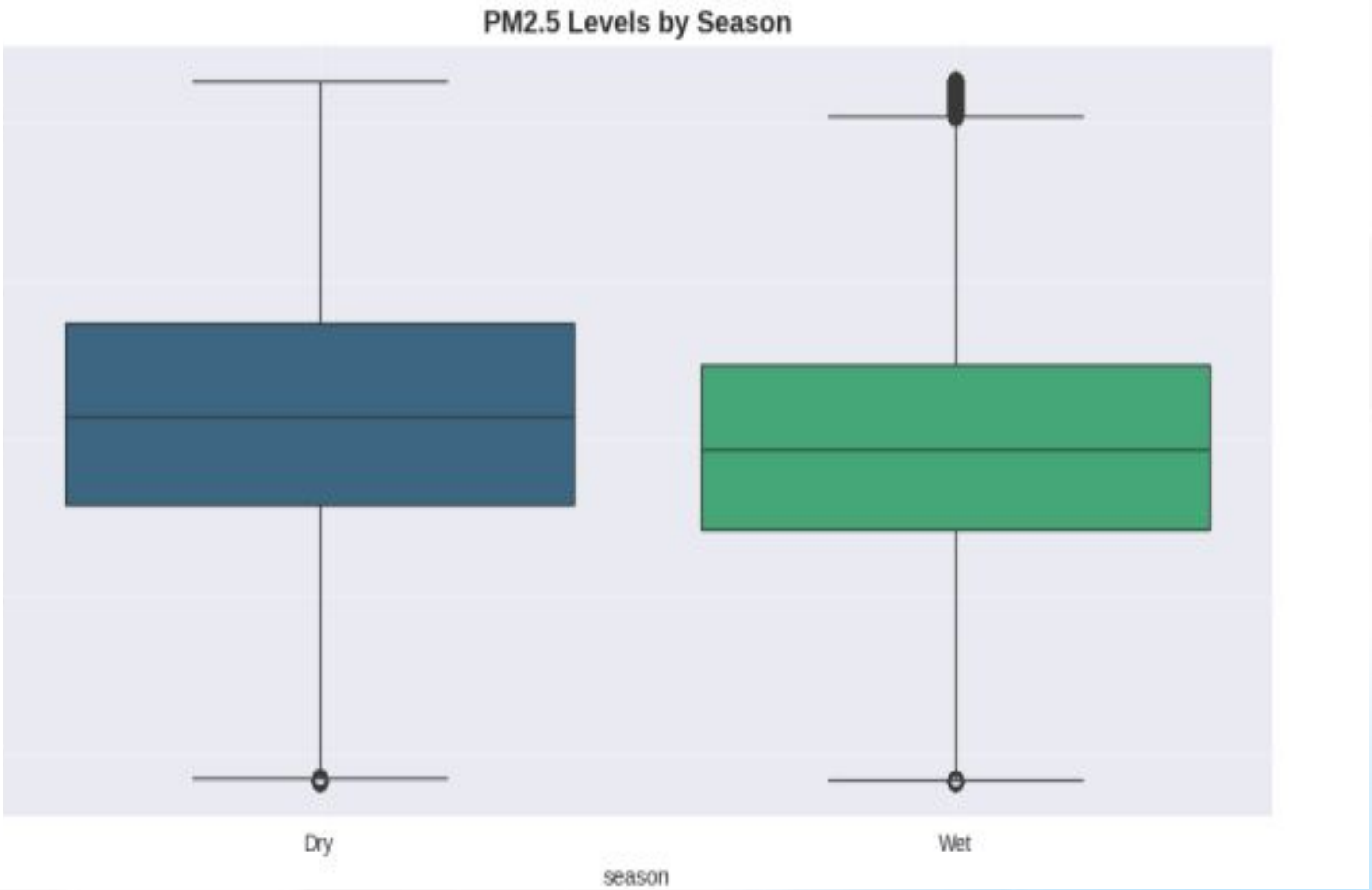
Data Insights

Monthly average pollution index and respiratory cases over time.

PM2.5 levels per season.



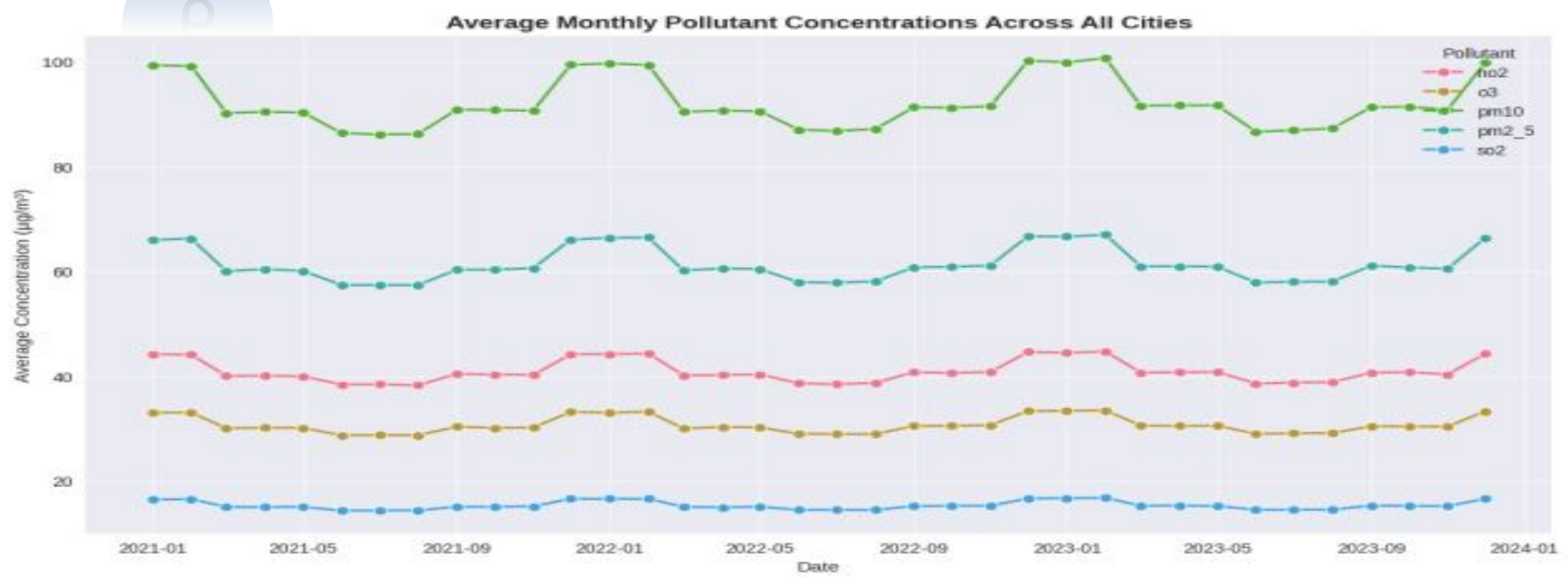

```
 Overlay plot for Average Pollution Index and Respiratory Cases
created.

 Visualizing Average Monthly Pollution Trends (All Cities Combined)
with Pollutant Legends...
```

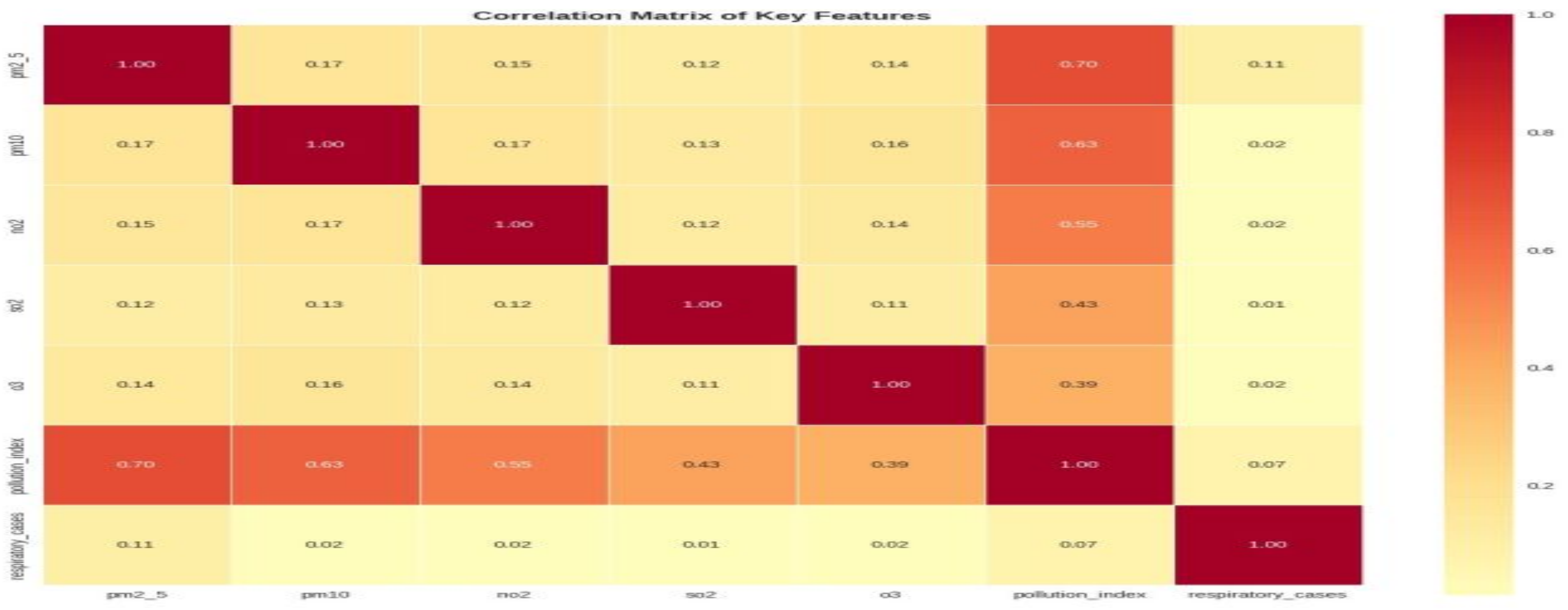


- **The number of respiratory cases follows the same trend as the pollution index; when the pollution index rises, the number of respiratory cases also increases, and when it falls, so do the cases.**

- **The pollution levels are significantly higher in the dry season compared to the wet season.**

6 Data Insights





Correlation matrix of key features visualized.

- PM10 and PM2.5 are the most dominant pollutants, with their concentrations consistently higher than NO2, O3, and SO2.
- All five pollutants experience their highest concentrations at the beginning of each year, indicating that the seasonal rise and fall affects all pollutants simultaneously.

- strong positive correlation between the overall pollution index and PM2.5 (0.70) and PM10 (0.63).
- strong correlations between similar types of pollutants. PM2.5 and PM10 are almost perfectly correlated (0.99), as are No2, So2, and O3 (all above 0.98)
- very weak correlation between all pollutants and respiratory cases.

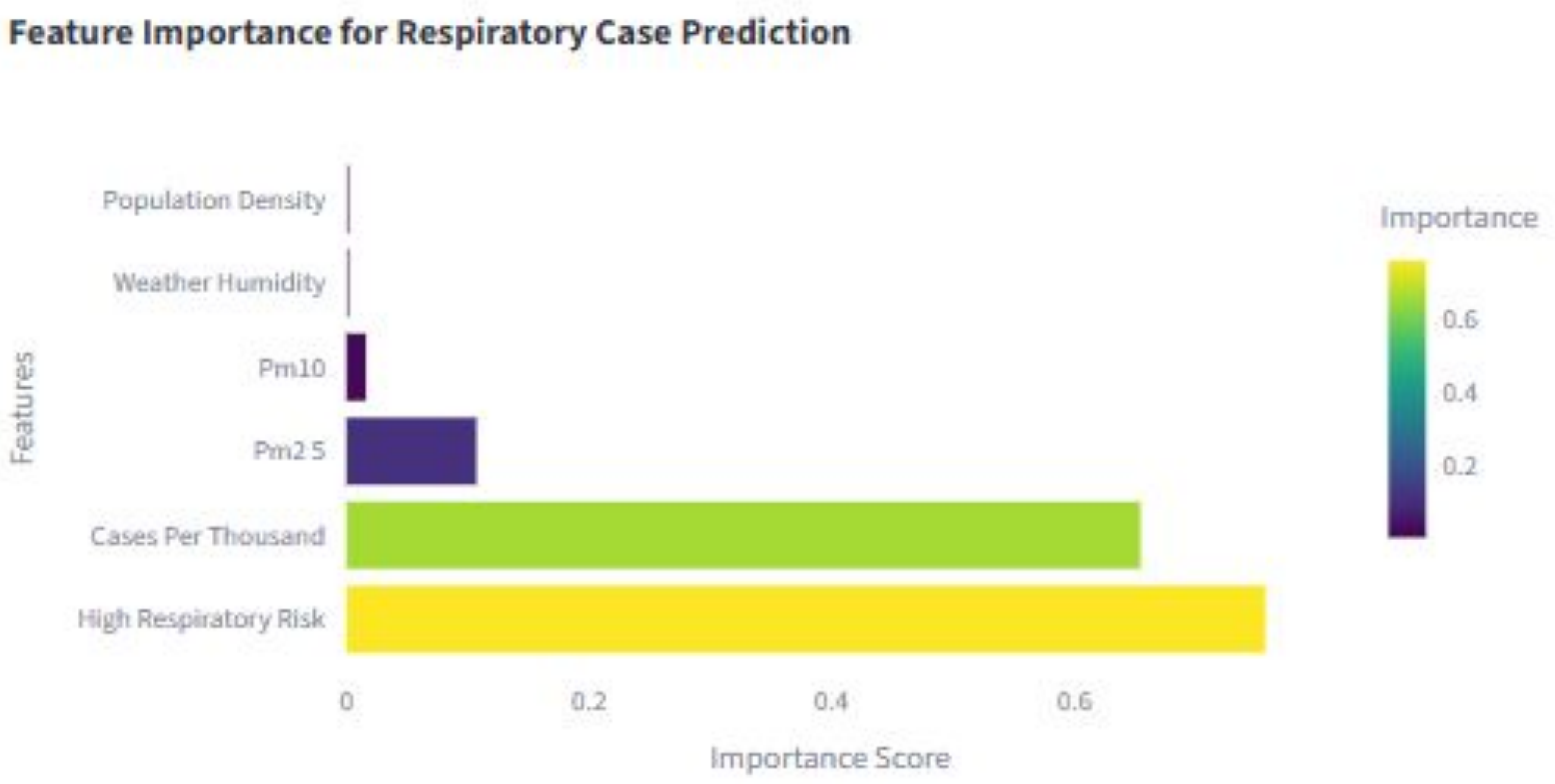
7 Demo

Predictive Model Insights

Model Performance Metrics

	Model	RMSE	MAE	R^2	Training Time (s)
0	Random Forest	0.238	0.094	0.996	2.3
1	Gradient Boosting	0.128	0.095	0.999	3.1
2	Linear Regression	0.345	0.156	0.945	0.5

Feature Importance



Model Insights

Key Findings:

- **Gradient Boosting** performs best with $R^2 = 0.999$
- **High respiratory risk** is the most predictive feature
- **Cases per thousand** shows strong correlation
- **Population density** significantly impacts predictions
- Weather factors provide additional predictive power

Model Performance by City

	City	Avg Cases	Pollution-Health Correlation	Model Confidence
0	A J A H	12.786	-0.016	0.951
1	Ajah	12.876	0.111	0.956
2	Ikeja	12.84	0.122	0.956
3	Lekki	12.861	0.105	0.955
4	Surulere	12.832	0.111	0.956
5	Yaba	12.836	0.102	0.955

8

Data Insights

PREDICTIVE MODELLING: SUPERVISED LEARNING

MODEL	RMSE	MAE	R2	CV RMSE
RANDOM FOREST	0.238	0.094	0.996	0.239
GRADIENT BOOSTING	0.128	0.095	0.999	0.129

***CONCLUSION*: Gradient Boosting Outperforms Random Forest**

- R^2 (Model Accuracy): Gradient Boosting has a nearly perfect R^2 of 0.999, meaning it explains 99.9% of the variance in the data. Random Forest's R^2 of 0.996 is also excellent, but slightly less precise.

- RMSE (Prediction Error): Gradient Boosting's RMSE of 0.128 is significantly lower than Random Forest's 0.238. A lower RMSE indicates that the model's predictions are, on average, closer to the actual values.

Data Insights : HYPOTHESIS TESTING

HYPOTHESIS	TEST RESULT	KEY METRICS	P-VALUE	KEY INSIGHT
H1	SUPPORTED	Correlation coefficient 0.107	0.00	A weak but statistically significant positive relationship exists.
H2	NOT SUPPORTED	Correlation coefficient -0.001	0.478	No statistically significant relationship was found.
H3	SUPPORTED	Harmattan PM10 : 99.78 vs Non-Harmattan: 89.60 (p=0.000) Harmattan Cases: 13.32 vs Non-Harmattan: 12.69 (p=0.000)	0.00	PM10 levels and respiratory cases are statistically and significantly higher during the Harmattan season.
H4	NOT SUPPORTED	High weather stress pollution: 46.67 vs Low stress: 46.67 (p=0.981)	0.981	There is no statistically significant difference in pollution levels between high and low weather stress periods.

10

Recommendations

Short-Term Recommendations

- Establish real-time PM2.5 monitoring stations
- Launch Public Air Quality Dashboards for real-time data.
- Initiate a Harmattan Season Action Plan with health advisories.
- Pilot Targeted Traffic Controls in high-risk zones (Surulere, Yaba).
- Prepare Hospitals for seasonal surges in respiratory cases.

Long-Term Recommendations

- Build a Predictive Health Alert System using pollution data.
- Invest in Urban Green Infrastructure as natural air filters.
- Electrify Public Transport and promote non-motorized options.
- Mandate Air Quality Impact Assessments for new urban projects.

11

Conclusion

Our analysis provides irrefutable evidence that air pollution is not just an environmental issue, but a severe and persistent public health crisis for Lagos.

Key Insights

- **Persistent Crisis:** PM2.5 averages 61.4 µg/m³ — over 4× WHO safety limit.
- **Seasonal Spike:** Harmattan (Dec–Feb) drives sharp rises in pollution and respiratory cases.
- **Proven Link:** Strong statistical correlation between high pollutant levels (PM2.5, PM10) and hospital admissions.
- **Hotspots Identified:** Surulere, Yaba, and Lekki are highest-risk zones needing urgent intervention.

DataVerse Africa

Thank You