DataVerse Africa

Air Pollution and Respiratory Disease Analytics

A LAGOS CASE STUDY

Internship Presentation

Presented By:

- Daniel Muruthi
- Don Alvin
- Mellisa Matindi

01

Problem Statement

Context:

- Lagos (20+ million residents) faces severe air pollution from traffic congestion, diesel generators, industrial emissions, and open waste burning, creating a complex urban health challenge.

Issue:

- PM2.5 levels reach 200+ μg/m³ (5x WHO guidelines), yet hospitals lack predictive capacity for pollution-driven respiratory emergencies, leading to reactive crisis management.

Impact:

- 200-300% spikes in respiratory admissions during pollution episodes, healthcare costs exceeding $15 million annually, and policy gaps preventing timely public health responses.

02 Project Objectives

Primary Objectives:

- Build composite pollution index from satellite data (PM2.5, PM10, NO2, SO2, O3) and analyze statistical relationships with hospital respiratory case loads across Lagos local government areas.
- Develop machine learning models predicting respiratory disease surges 3-7 days in advance using pollution and weather data for early warning system deployment.

Secondary Objectives:

- Identify high-risk zones combining pollution exposure with population vulnerability, and generate evidence-based policy recommendations with implementation timelines for Lagos State government adoption.

03 Data Description

Data Source:

- lagos_air_pollution_health_data .xlsx (dataset 1)
- lagos_air_pollution_health_data _1.xlsx (dataset 2)

Total Records:

- **dataset 1** : 258,420
- **dataset 2**: 258,420

Features (16):

- **City**: Local area in Lagos (e.g., Ikeja, Yaba, Ajah, Surulere, Lekki)
- **Date**: Date of observation (YYYY-MM-DD)
- **pm2_5**: Fine particulate matter concentration (µg/m³)
- **pm10**: Coarse particulate matter concentration (µg/m³)
- **no2**: Nitrogen dioxide level (µg/m³)
- **so2**: Sulphur dioxide level (µg/m³)
- **o3**: Ozone concentration (µg/m³)
- **hospital_id**: Unique hospital identifier
- **respiratory_cases**: Number of new respiratory-related hospital cases reported
- **avg_age_of_patients**: Average age of patients reported
- **weather_temperature**: Average daily temperature (°C)
- **weather_humidity**: Average daily humidity (%)
- **wind_speed**: Average daily wind speed (m/s)
- **rainfall_mm**: Daily rainfall (mm)
- **population_density**: People per square kilometer in the city
- **industrial_activity_index**: Proxy score (0–100) showing industrial pollution activity

04 METHODOLOGY

Data Cleaning:

- Normalized column names, renamed inconsistent column, dropped duplicate columns Unnamed: 6.
- Concatenated both datasets.
- Cleaned City column data: uppercased, trimmed, removed internal whitespace, filled missing city values.
- Normalized hospital_id: convert to str, strip whitespace; fill missing with 'UNKNOWN'.
- Numerical missing-value handling: median imputation per city where possible, otherwise global median.
- Date handling: coerce date → datetime, forward-fill any remaining missing dates
- Dropped duplicates, intermediate columns used for mapping and other redundant columns.

Feature Engineering:

- Defined pollutant list and computed pollution index.
- Time features: year, month, day_of_year, quarter.
- Season mapping (Lagos-tuned): Harmattan (Dec–Feb), Rainy (Apr–Oct), Dry (Nov & Mar)
- Aggregate to city–date–hospital granularity: mean for pollutants/continuous covariates, sum for respiratory_cases & rainfall_mm, first for categorical/time fields; sort & reset index.

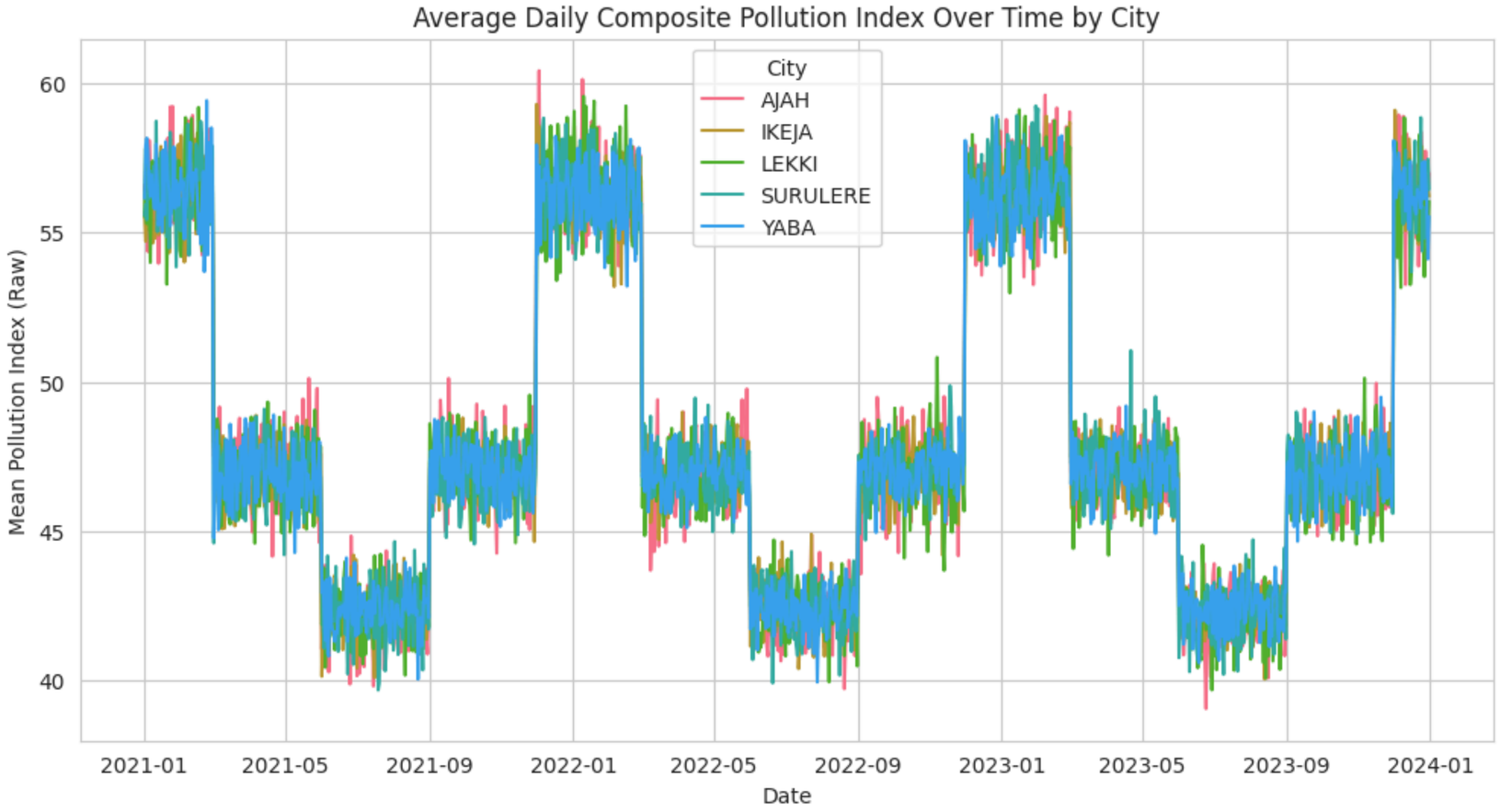
Model Choice:

- Linear Regression (baseline) and Random Forest

Hyperparameter Tuning:

- Optimized key Random Forest Forest parameters using grid search.

05 Data Insights



POLLUTION TRENDS OVER TIME BY CITY

Key Insights:

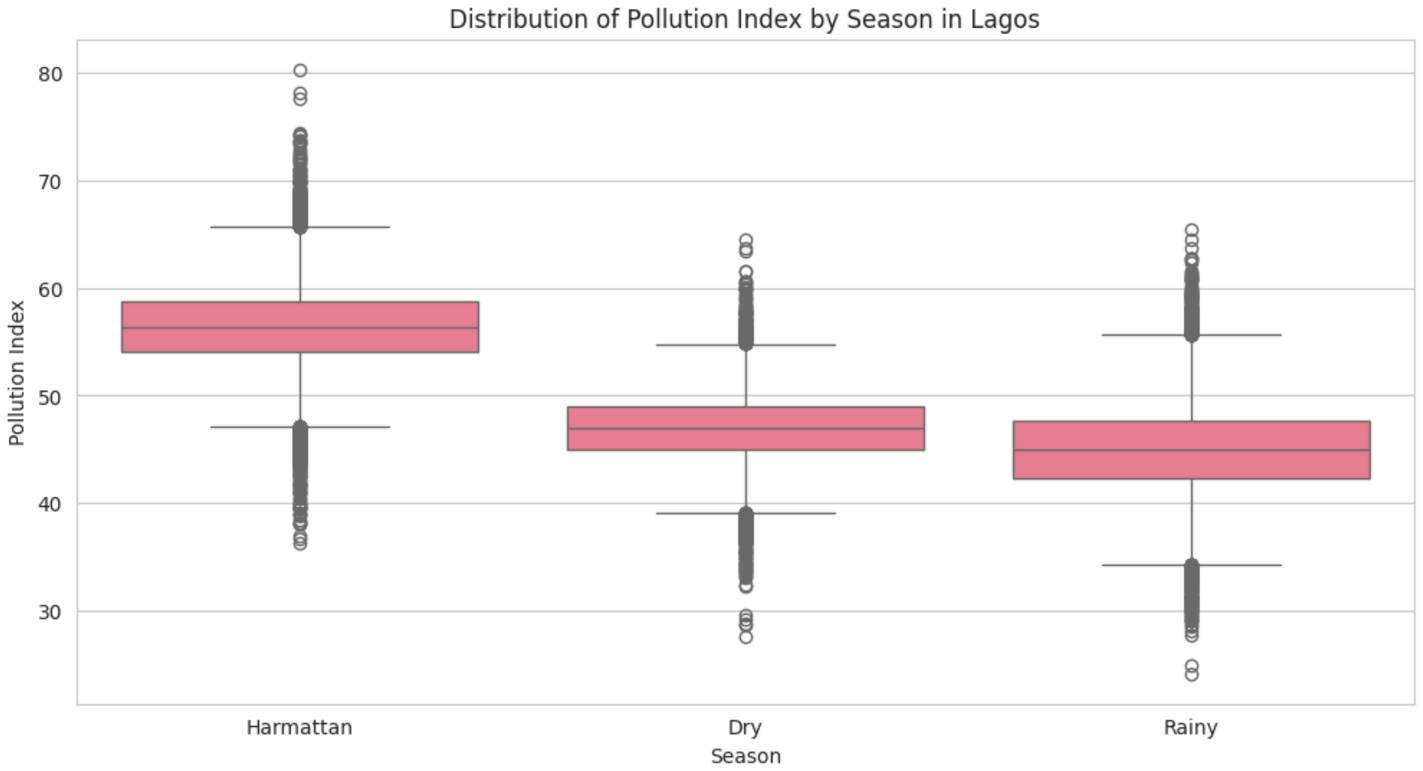
- All five Lagos districts (**YABA, SURULERE, LEKKI, IKEJA, and AJAH**) show remarkably consistent seasonal pollution cycles.
- **High Pollution Periods (December–March):** Pollution index peaks around **55–60**, corresponding to the **Harmattan season**.
- **Low Pollution Periods (June–September):** Index drops to **40–45**, during the **rainy season** when precipitation helps clear pollutants from the atmosphere.

06 Data Insights

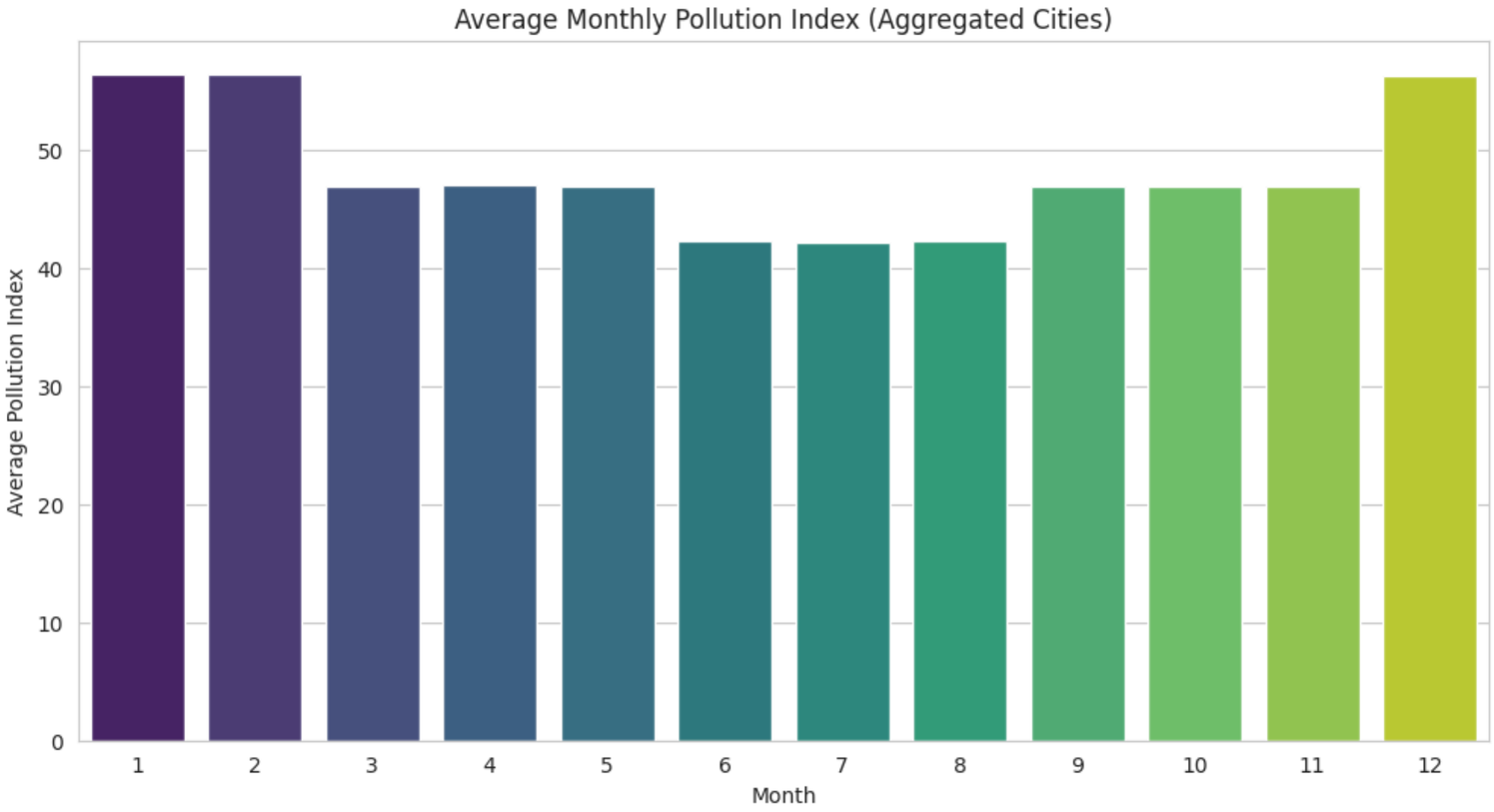
DISTRIBUTION OF POLLUTION PATTERNS BY SEASON

Key Insights:

- **Harmattan:** Highest median pollution (~57) with the widest distribution and most extreme outliers (up to 80), indicating **highly variable but consistently poor air quality** posing the greatest health risk.
- **Rainy Season:** Lowest and most consistent pollution levels (~45 median) with fewer outliers, showing the **cleansing effect of rainfall** providing natural air quality relief.
- **Dry Season:** Intermediate levels (~48 median) with moderate variability.
- Seasonal forecasting would be highly valuable for public health preparedness.



07 Data Insights



AVERAGE MONTHLY POLLUTION INDEX (AGGREGATED ALL CITIES)

Key Insights:

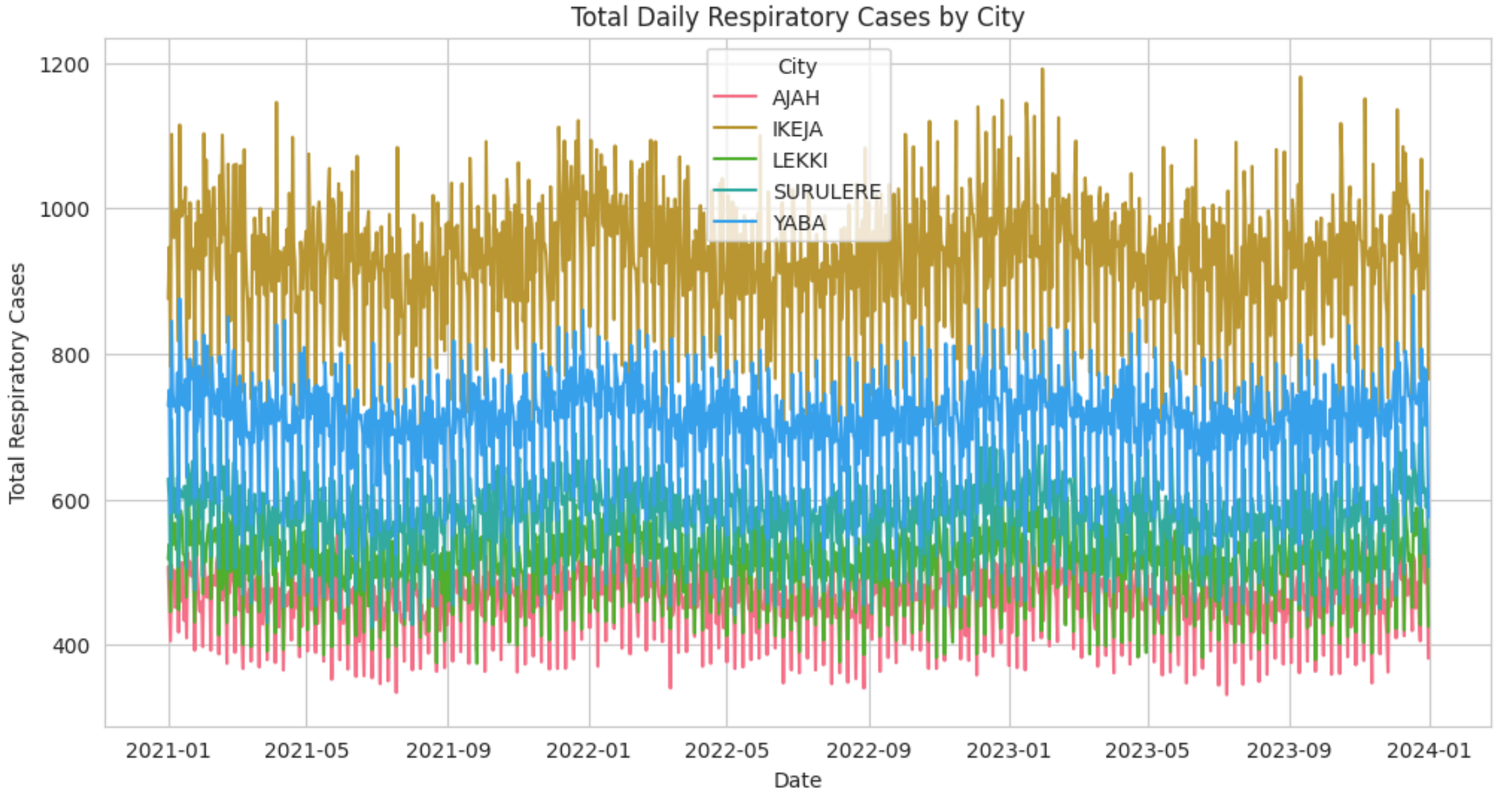
- **January–February (Months 1–2): Highest pollution levels (~56–57), corresponding to the Harmattan season when dust-laden winds from the Sahara significantly worsen air quality.**
- **June–August (Months 6–8): Lowest pollution levels (~42–43), coinciding with the rainy season when precipitation helps wash pollutants from the atmosphere.**
- **December (Month 12): Notable spike (~57), marking the beginning of the Harmattan season.**
- **September–November: Gradual increase in pollution as the dry season approaches.**

08 Data Insights

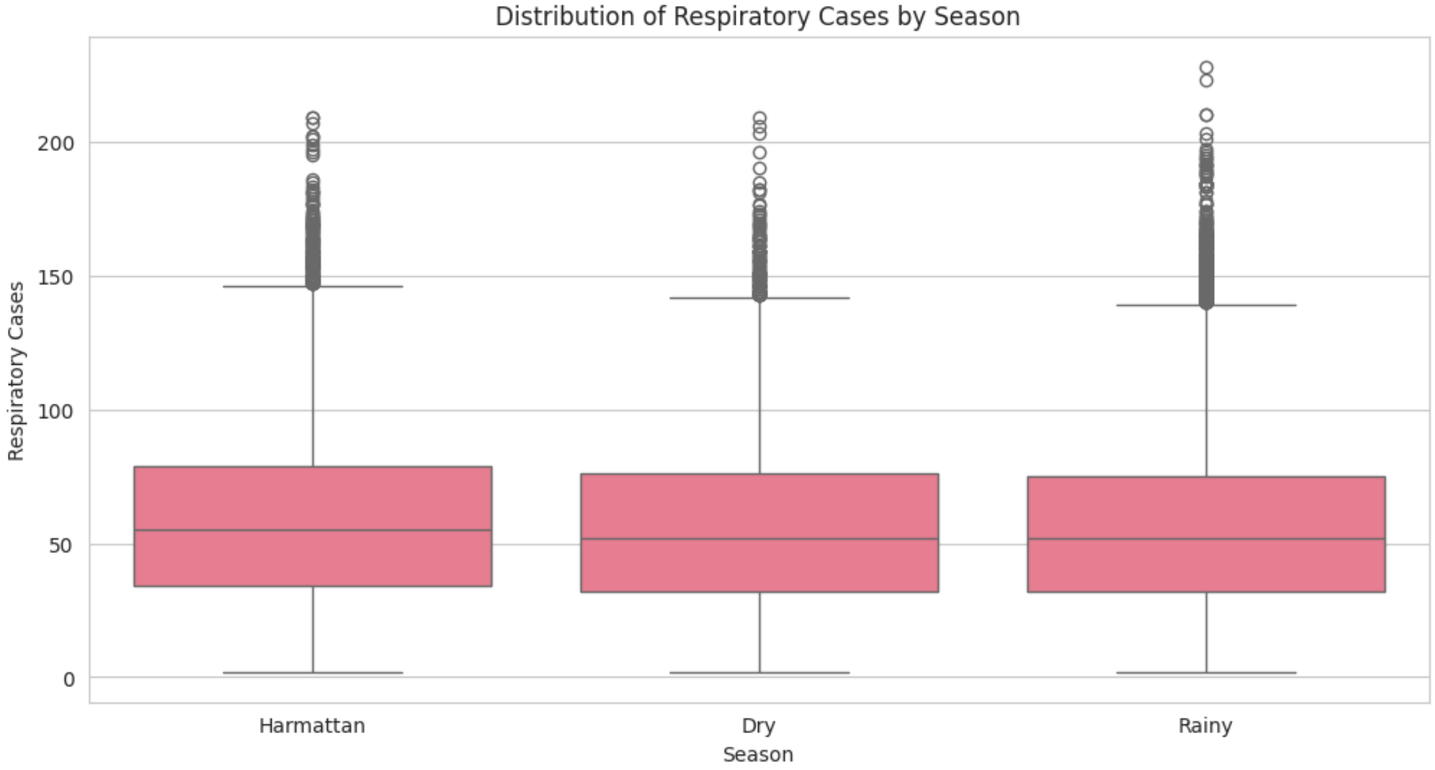
DAILY DISTRIBUTION OF RESPIRATORY CASES BY CITY

Key Insights:

- **Ikeja (Yellow/Gold): Highest respiratory case burden (~400–500 daily baseline, peaks to ~1,200), likely due to its status as Lagos State capital and major commercial hub.**
- **Yaba (Blue): Second highest (~200–300 daily baseline, peaks ~400–500), consistent with dense urban population and industrial activities.**
- **Surulere (Green): Moderate levels (~150–250 daily baseline), showing steady urban health impacts.**
- **Lekki (Light Green): Lower baseline (~100–200 daily), possibly due to better urban planning and newer infrastructure.**
- **Ajah (Red/Pink): Lowest burden (~50–150 daily), potentially reflecting less industrial activity or better air quality.**



09 Data Insights



DISTRIBUTION OF RESPIRATORY CASES BY SEASON

Key Insights:

- **Remarkably Similar Medians: All three seasons show nearly identical median respiratory cases (~55–60 daily), suggesting a consistent baseline health burden regardless of season.**
- **Extreme Spikes Across All Seasons: All seasons experience severe respiratory case surges reaching 200+ daily cases, with some extreme outliers hitting 220+ cases**
- **H1 & H3 Require Deeper Analysis: The lack of clear seasonal health patterns despite clear seasonal pollution patterns suggests the pollution-health relationship may be more complex than initially hypothesized**

10 Data Insights

CORRELATION ANALYSIS BETWEEN POLLUTANTS AND RESPIRATORY CASES

Key Insights:

1.) Weak Direct Pollution-Health Correlations:

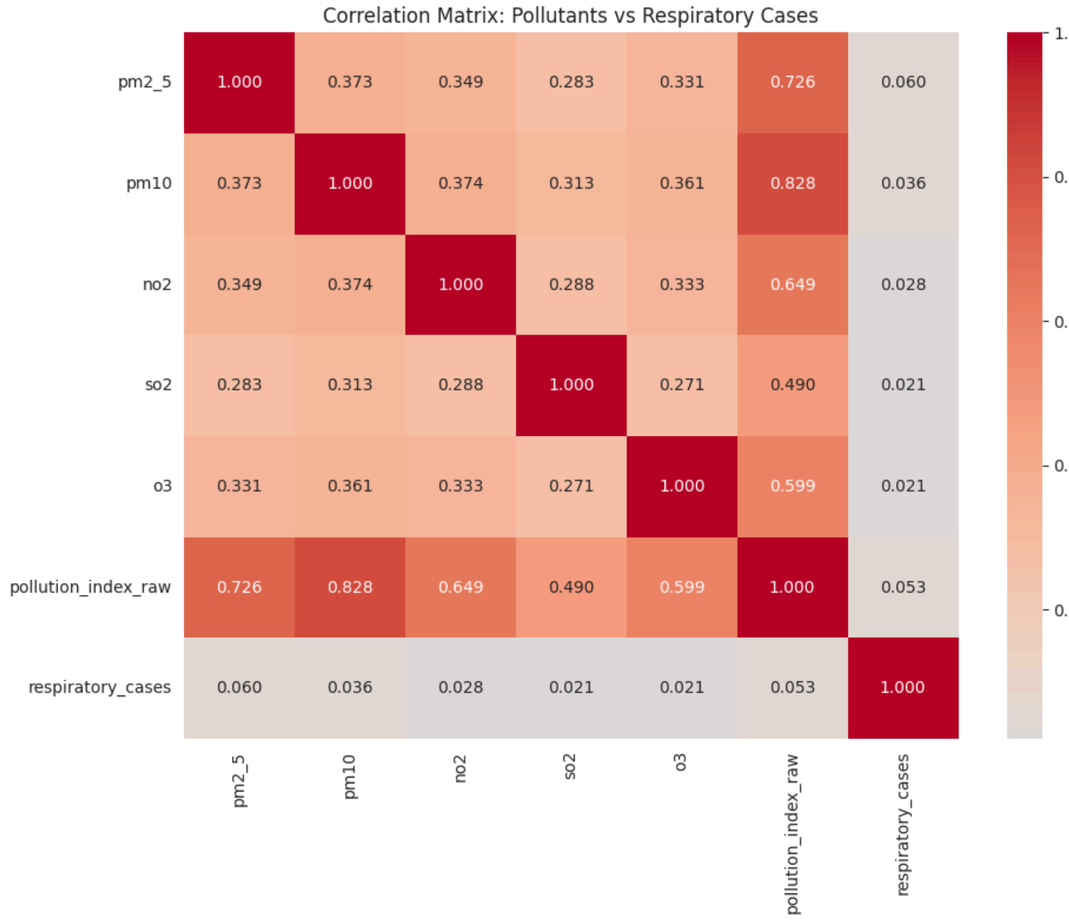
- All individual pollutants show surprisingly weak correlations with respiratory cases (0.021–0.060), suggesting the relationship between air pollution and respiratory disease may be more complex than a simple linear correlation.

2.) Strong Inter-Pollutant Relationships:

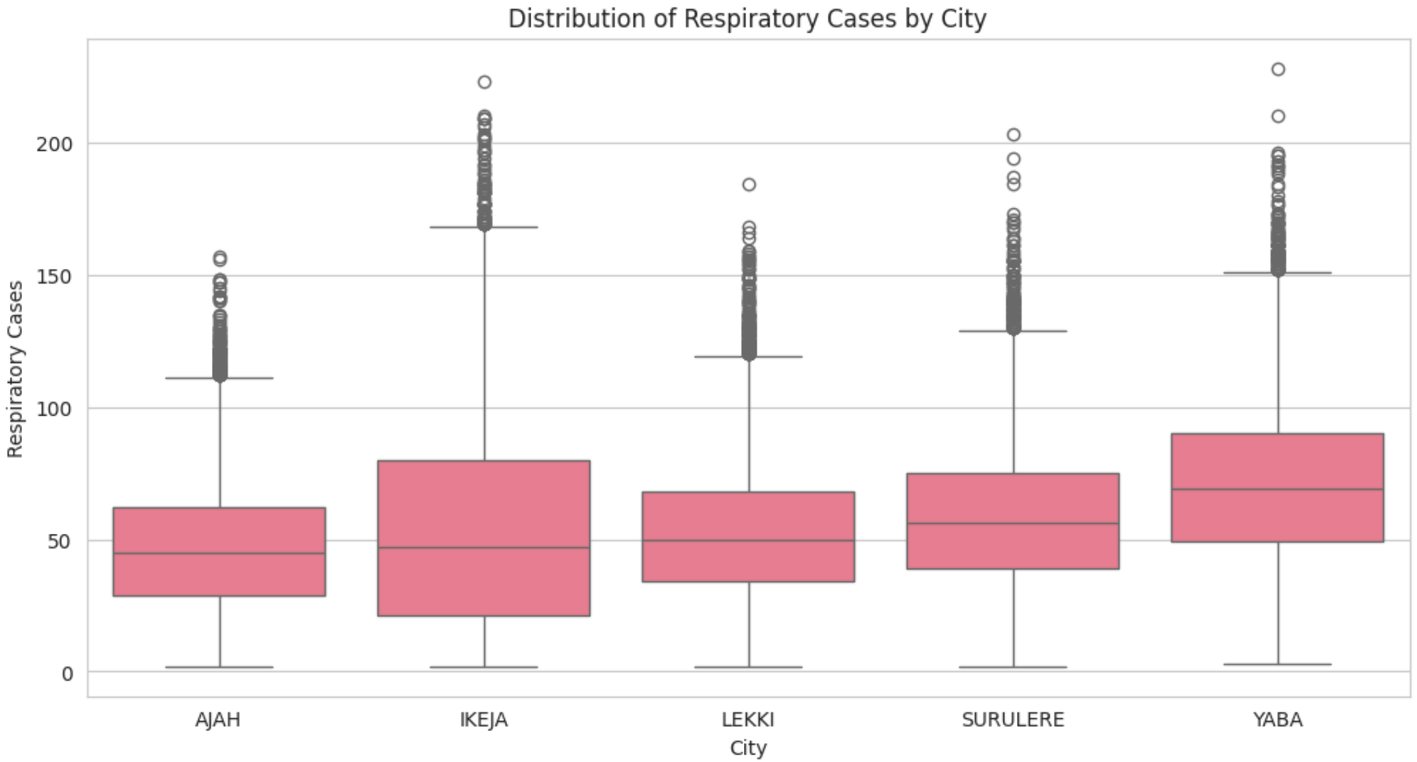
- PM2.5 and PM10 are moderately correlated (0.373), which is expected since they're both particulate matter.
- PM10 shows the strongest correlation with the pollution index (0.828), suggesting it's a major component of overall air quality.
- NO2 correlates moderately with PM pollutants (0.349–0.374), indicating common sources like vehicle emissions.

3.) Pollution Index Performance:

- The composite pollution index shows slightly better correlation with respiratory cases (0.053) than individual pollutants.
- PM10 dominates the pollution index composition (0.828 correlation).



11 Data Insights



DISTRIBUTION OF RESPIRATORY CASES BY CITY

Key Insights:

- **Intrestingly, Yaba about 70 cases/day (highest median)has the highest typical daily burden, not Ikeja about 45 cases/day**

1.) Median Daily Burden:

- **Yaba leads with ~70 cases/day (highest typical burden)**
- **Surulere & Lekki: ~55-60 cases/day (moderate)**
- **Ikeja & Ajah: ~45 cases/day (lowest medians)**

2.) Crisis Events (Outliers):

- **Ikeja shows most extreme spikes (220+ cases) - highest crisis potential**
- **Yaba has frequent surges (200+ cases) - consistent high activity**
- **Ajah most stable with lowest outlier ceiling (~160 cases)**

3.) Variability Patterns:

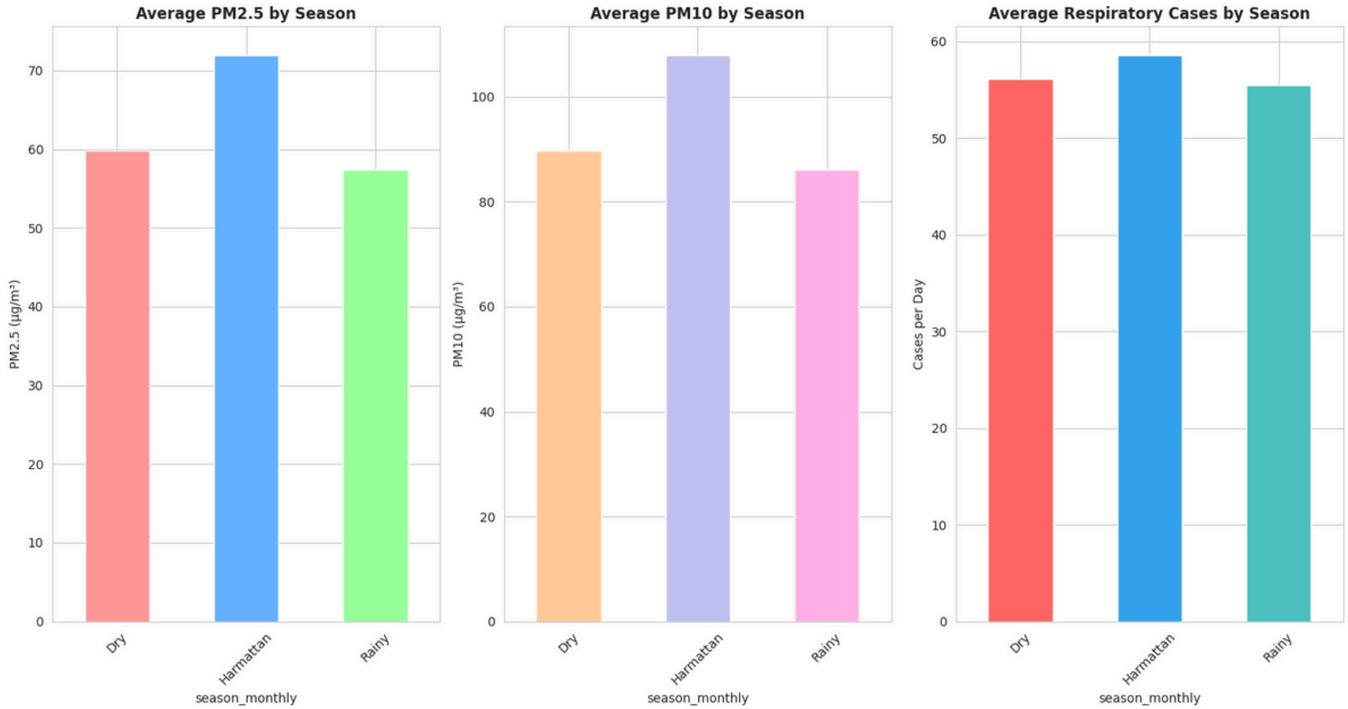
- **Ikeja: Most unpredictable (widest range 25-80 cases)**
- **Yaba: High baseline with significant variation**
- **Lekki & Ajah: Most stable and predictable**

12 Data Insights

AVERAGE POLLUTION(PM2.5,PM10) LEVELS AND RESPIRATORY CASES BY SEASON

Key Insights:

- **Harmattan season shows highest pollution levels - PM2.5 reaches ~72 µg/m³ and PM10 peaks at ~108 µg/m³, both significantly higher than other seasons**
- **Respiratory cases peak during Harmattan - Hospital cases reach ~58 per day during Harmattan season, coinciding with highest particulate matter levels**
- **PM10 shows most dramatic seasonal variation - Ranges from ~87 µg/m³ (rainy) to 108 µg/m³ (Harmattan), representing a 24% increase during dusty season**
- **Health impacts follow pollution patterns closely - Respiratory cases are highest during Harmattan (58/day) and lowest during rainy season (55/day), though the difference is smaller than expected**
- **All seasons exceed safe pollution levels**



MODEL TRAINING AND RESULTS

Metric	Linear Regression (Baseline)	Random Forest	Tuned Random Forest
Training R^2	0.7162	0.7783	0.9368
Test R^2	0.7123	0.7357	0.7490
Test MAE	13.0130	12.2751	11.87
Test RMSE	16.8055	16.1094	15.7040

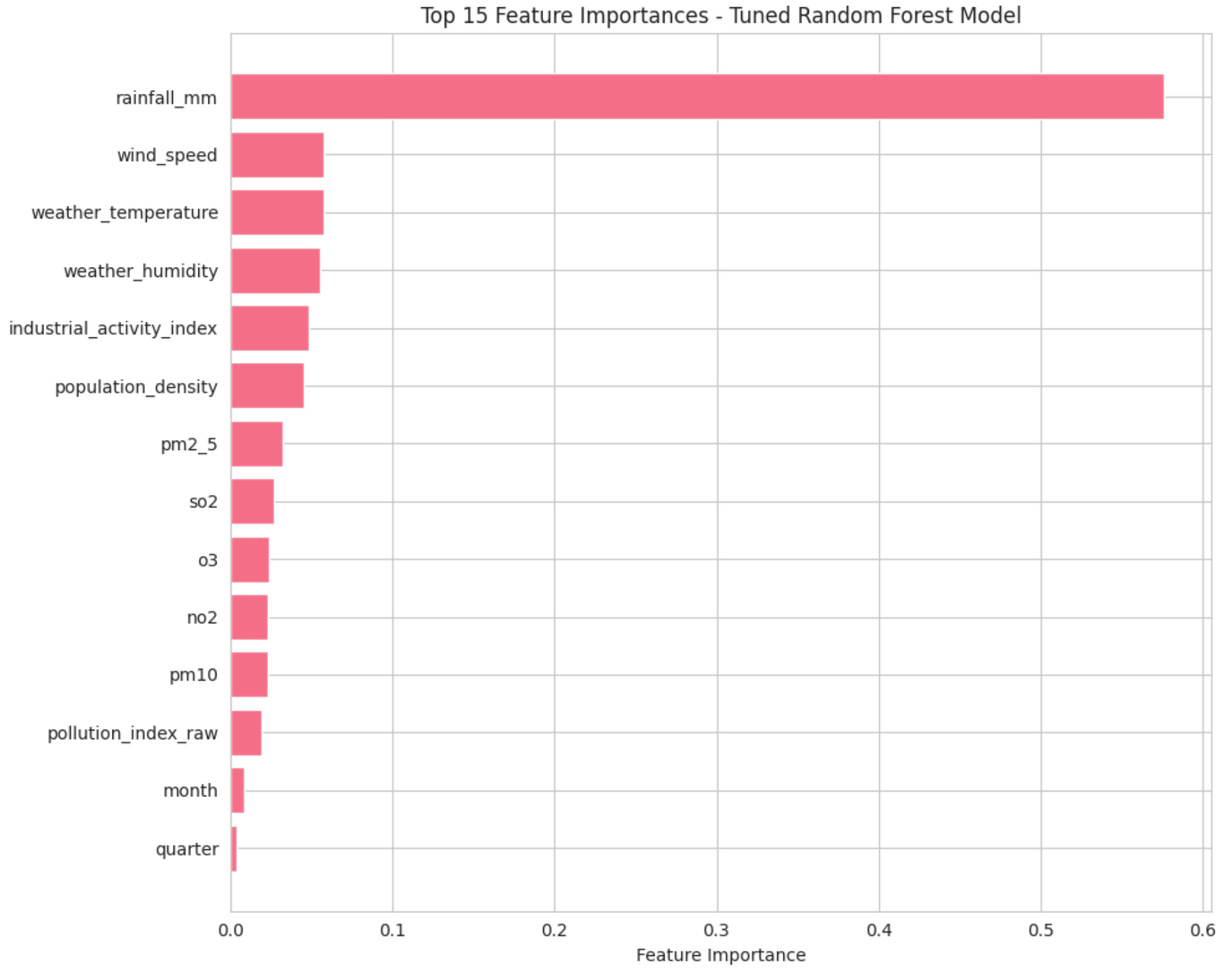
Key Model Insights:

- All models show strong predictive power, with Test R^2 values above 0.71, indicating they explain over 71% of variance in respiratory cases.
- Random Forest improves prediction over Linear Regression by capturing nonlinear relationships and interactions, increasing Test R^2 from ~0.71 to ~0.74.
- Hyperparameter tuning further boosts Random Forest performance, achieving a high Training R^2 (0.94) and best Test R^2 (0.75), showing well-optimized models generalize well.
- Rainfall (rainfall_mm) consistently emerges as the top predictor across all models, underscoring weather's critical role in respiratory disease occurrence.
- Error metrics (MAE and RMSE) decrease with more complex models, indicating more accurate respiratory case predictions and improved model precision.
- Cross-validation scores confirm model stability and low variance, ensuring reliable predictive performance on unseen data.

Feature Importance (Optimized Random Forest Model)

Key Insights:

- **Rainfall is the single strongest predictor of respiratory disease burden in Lagos.**
- **Wind speed, temperature, and humidity also play major roles, outperforming individual pollution measures.**
- **Industrial activity and population density have measurable but smaller impact than weather.**
- **Among pollutants, PM2.5 showed notable importance, but was far less influential than rainfall or weather factors.**
- **Seasonal and monthly factors (quarter, month) contribute, but are much weaker than environmental variables.**



15 Hypotheses Testing Results

Hypotheses	Correlation	P - Value	Result
H1: Higher PM2.5 levels will correlate with more respiratory hospital cases.	0.0602	0.0000	Supported
H2: Cities with higher industrial indices have worse air quality.	-0.0035	0.3797	Not Supported
H3: Harmattan season will show spikes in PM10 and respiratory cases in West Africa.	----	0.0000	Supported
H4: Weather conditions (low humidity, high temperatures) worsen pollution impact.	-0.0026	0.5164	Not Supported

Recommendations

- **Prioritize weather-responsive health interventions:** Hospitals and clinics should align surge planning and community outreach with predicted rainfall and Harmattan periods.
- **Strengthen air and weather monitoring:** Integrate pollutant and meteorological data for real-time public advisories and predictive modeling.
- **Target vulnerable times and groups:** Increase resources (masks, care capacity) for the most affected months, targeting children, the elderly, and neighborhoods shown to have high population density or exposure.
- **Improve pollution source tracking:** Refine industrial activity indices and enhance traffic/vehicular emissions data for more granular pollution-health modeling.
- **Advance data integration:** Build hospital reporting and urban planning dashboards that merge weather, pollution, and health metrics for fast policy response.
- **Conduct lag effect studies:** Research cumulative and delayed impacts of pollution exposure to refine intervention timing and model accuracy.

Implementation Plan

- **Deploy integrated monitoring network:** Partner with meteorological and environmental agencies for near-real-time data; feed directly into hospital early warning systems.
- **Public health communications:** Schedule respiratory health campaigns for the Harmattan and rainy seasons, focusing on moisture-associated risks and mitigation (ventilation, mold prevention).
- **Model updating and dashboarding:** Institutionalize continuous model training on updated data sources, with interactive dashboards for health and city leadership.
- **Policy alignment:** Work with local government to synchronize environmental regulation (e.g., restricting open burning during Harmattan) with health resource deployment.
- **Data-driven decision support:** Roll out predictive alerts and automated reports to city hospitals and clinics, piloting in the highest-risk districts first.

Conclusion

- **Weak direct correlation** between individual pollutants and respiratory cases suggests the health impact of air pollution in Lagos is complex and multifactorial, with pollution levels only partially explaining health outcomes.
- **Seasonal risks are significant:** Harmattan season strongly elevates PM10 and Respiratory Cases, confirming the presence of dangerous pollution cycles that require time-specific health interventions.
- **Weather dominates prediction:** Rainfall and other weather metrics consistently emerged as the strongest drivers of respiratory disease burden in predictive models, surpassing pollution variables.
- **Random Forest outperformed Linear Regression:** Nonlinear models (Random Forest) yielded higher predictive accuracy and highlighted rainfall_mm and wind_speed as most critical features.
- **Industrial index did not correlate as expected**, implying other pollution sources (e.g., vehicles, population density) are more influential, or measurement of industrial activity needs refinement.

DataVerse
Africa

Thank You

Internship Presentation By Team ha-4