Programming for Data Science

AIR QUALITY AND HEALTH IMPACT ANALYSIS

22BDS0394 ANUSHRI AHIR

Course: Interactive Visualization

B.Tech.

in

Computer Science and Engineering

(with specialization in Data Science)

School of Computer Science and Engineering

28 October 2025

1. Introduction

Interactive visualization is one of the most powerful ways to explore and interpret complex datasets, especially in domains like environmental health where data varies by time, geography, and population type.
This project focuses on Air Quality and Health Impact Analysis, using data on PM2.5 (fine particulate matter) levels from the World Health Organization (WHO). PM2.5 is a major pollutant known to cause serious health problems such as asthma, bronchitis, and cardiovascular diseases.
The goal of this project is to use interactive visualization to understand how air quality differs across regions, residence types, and years, and to assess its potential health impact through a derived metric called the *Health Impact Score*.
Interactive visualization is relevant here because it allows users to explore the dataset dynamically — zooming into specific countries, filtering by region, and hovering over data points to view exact PM2.5 values. This level of interaction turns static data into a live analytical environment, making trends and anomalies immediately visible.
In contrast, static visualizations are limited: they show only predefined views and require separate plots for every new question. Interactive tools, however, enable real-time queries and "what-if" analysis, which are essential for data-driven decision-making in public health and environmental policy.

2. Overview of Packages

Plotly
Plotly is a powerful R library for creating interactive and web-based graphics. It supports a wide range of charts including scatter plots, line graphs, histograms, boxplots, bar charts, heatmaps, and bubble charts.
Key features include zooming, panning, hover tooltips, and dynamic legends. Plotly integrates easily with R Markdown and Shiny dashboards, allowing users to embed interactive visuals directly in reports or web apps.
It's especially useful for exploratory data analysis where users need to drill down into specific data points or compare multiple variables simultaneously.
Rbokeh
Rbokeh is an R interface to the Bokeh visualization library. It is designed for high-quality, interactive plots that can be embedded in web browsers. Rbokeh excels in custom interactivity — for example, allowing linked brushing between two plots or adding dynamic widgets for filtering.
While less widely used than Plotly, it integrates smoothly into R workflows and can handle medium-sized datasets effectively. It's best suited for customized dashboards or when finer control over interactions is needed.
Leaflet
Leaflet is the leading R package for interactive map visualizations. It supports features such as zooming, panning, marker clustering, popups, and multi-layer control.
For this project, Leaflet was used to plot country-level PM2.5 averages using latitude and longitude centroids obtained from the rnaturalearth package. Each circle's color intensity and size represent pollution severity, making it easy to identify hotspots around the world.
Leaflet is particularly useful for geospatial data analysis, especially when working with environmental or demographic datasets that are location-based.

2.1 Comparison of Packages

Feature	Plotly	Rbokeh	Leaflet
Type	General-purpose charting	Custom interactive charts	Map-based visualization
Strength	Versatility and ease of use	Fine control and linked brushing	Spatial data and maps
Output	HTML, R Markdown, Shiny	HTML, R Markdown	Interactive maps in R Markdown or Shiny
Interactivity	Zoom, hover, selection	Widgets, linked plots	Pan, zoom, popups, layer toggle
Best For	Analytical dashboards	Custom interfaces	Location-based analysis

In this project, Plotly was used for most analytical visualizations (histograms, boxplots, scatterplots), while Leaflet handled global mapping of pollution levels. Rbokeh was included for demonstrating custom chart interactivity.

3. Background Review

Air pollution is one of the major causes of premature death globally, with PM2.5 being a key pollutant responsible for serious respiratory and cardiovascular illnesses. According to the WHO, nearly 99% of the world's population breathes air that exceeds safe quality levels. Governments and environmental agencies collect massive datasets on air quality to inform public health decisions. However, without effective visualization, such data often remains underutilized.

Existing platforms like AirVisual Earth, IQAir, and NASA's Air Quality Dashboard rely heavily on interactive visualization to show real-time pollution levels and trends. Similarly, OpenAQ uses open data and visual dashboards to allow global comparisons.

Interactive visualization is especially important in this domain because it helps non-technical users — such as policymakers or the general public — understand where pollution levels are rising, which areas are safe, and how conditions change over time. This project follows the same principle by turning WHO's raw data into an interactive visual narrative that highlights global air quality inequalities and health risks.

4. Case Explanation

4.1 Dataset Description

The dataset used is the World Health Organization's Global PM2.5 Data, containing records for 9,450 observations across 293 countries. Each record includes:

- Country and Region
- Residence Type (City, Rural, Urban, Total)
- Year (Period)
- PM2.5 Concentration (μg/m³)

Using the rnaturalearth and countrycode packages, geographic coordinates were added for each country to enable mapping.

4.2 Topic Description

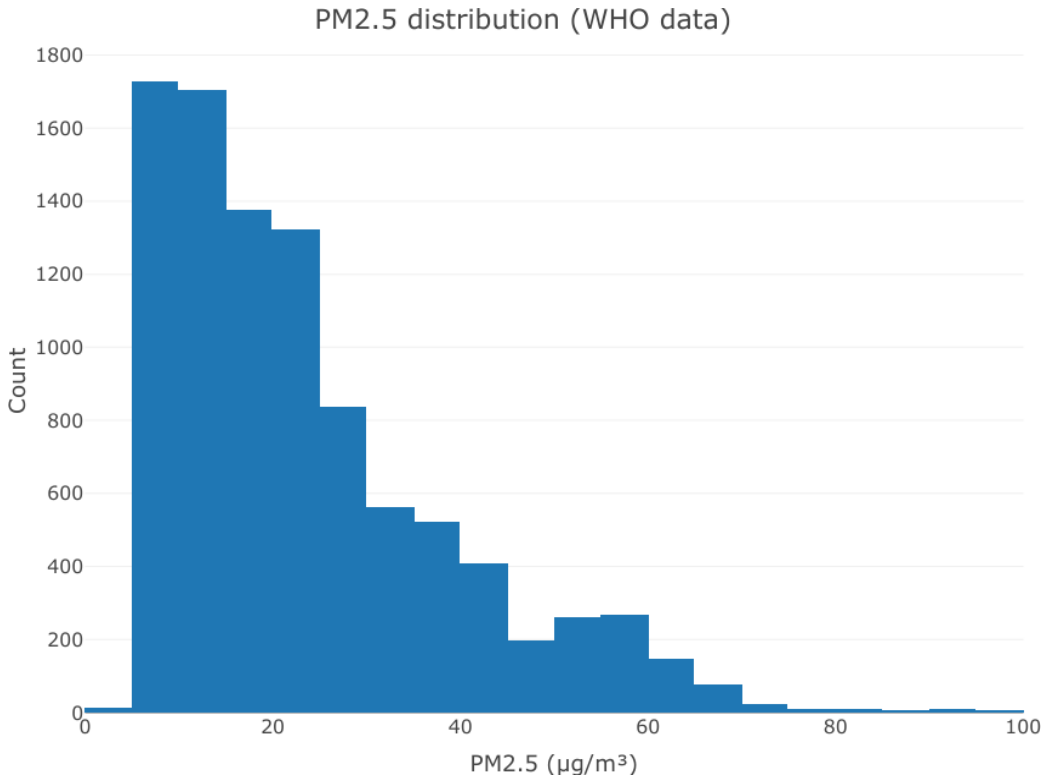
The project explores how PM2.5 levels differ across regions and residence types, and how they correlate with health impact using a derived Health Impact Score. The analysis identifies high-risk zones, compares cities versus rural areas, and visualizes how air quality has evolved over time.

4.3 Expected Outcomes

1. Identify global pollution hotspots through Leaflet maps.
2. Quantify differences in PM2.5 exposure between urban and rural regions.
3. Assess the relationship between PM2.5 and health risk levels.
4. Demonstrate how interactivity improves insight compared to static charts.

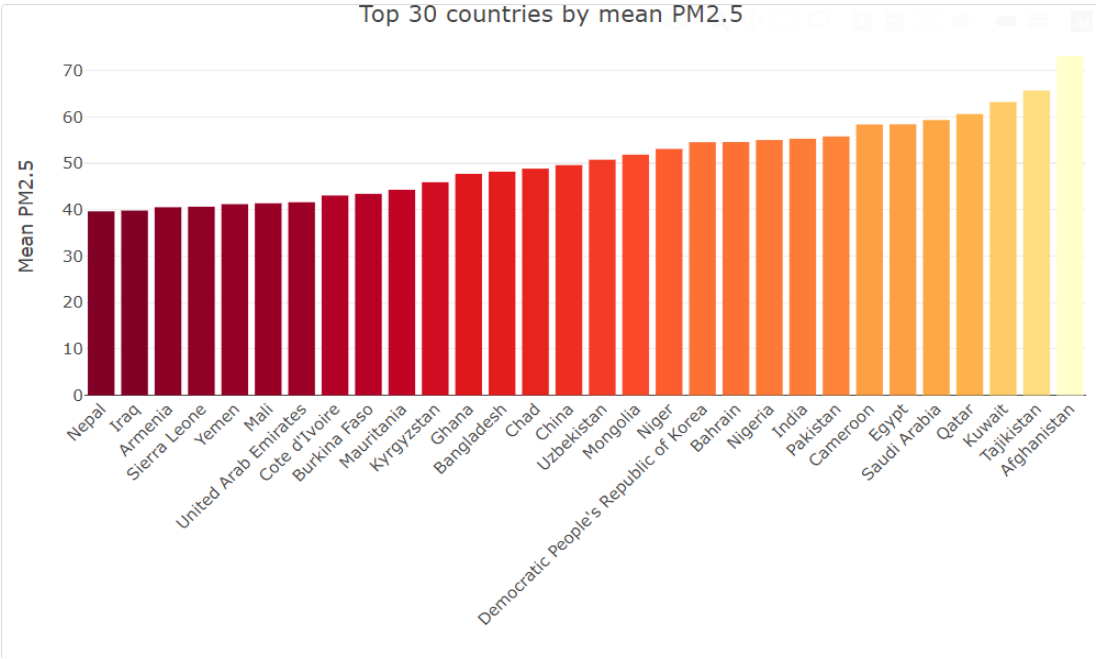
4.4 Interactive Visualizations

1. Global PM2.5 Distribution (Histogram – Plotly)
   Displays frequency distribution of PM2.5 across all countries.



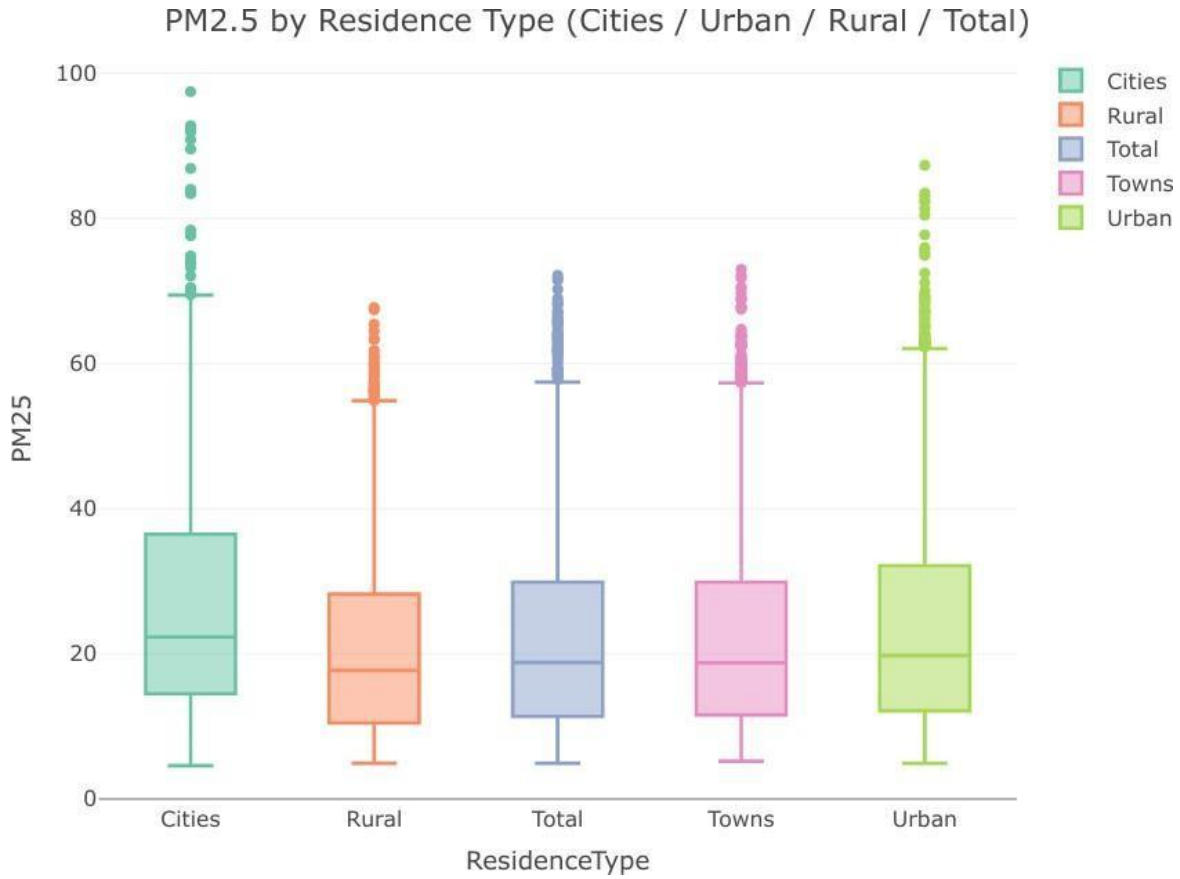
   Insight: Shows how global PM2.5 values are distributed. Most countries fall under low exposure, with a long-left tail representing low-pollution zones.

2. Top 30 countries by mean PM2.5 (bar chart)



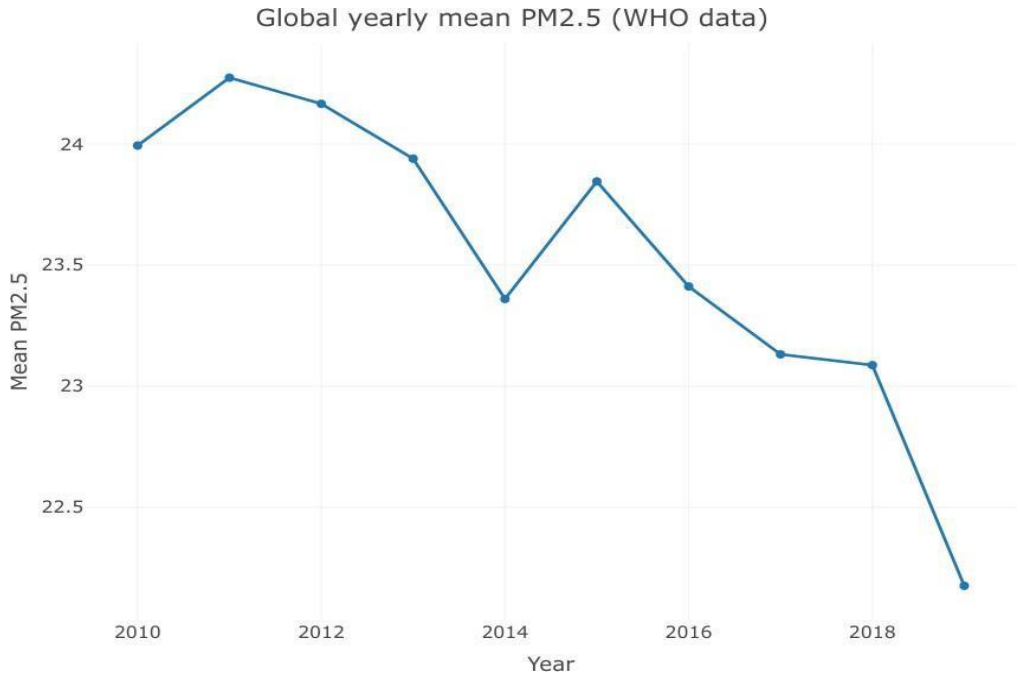
Insight: Highlights the most polluted countries globally, using YlOrRd color scale for intuitive severity display.

3. PM2.5 by Residence Type (Boxplot – Plotly)
Compares pollution levels among Cities, Urban, Rural, and Total.



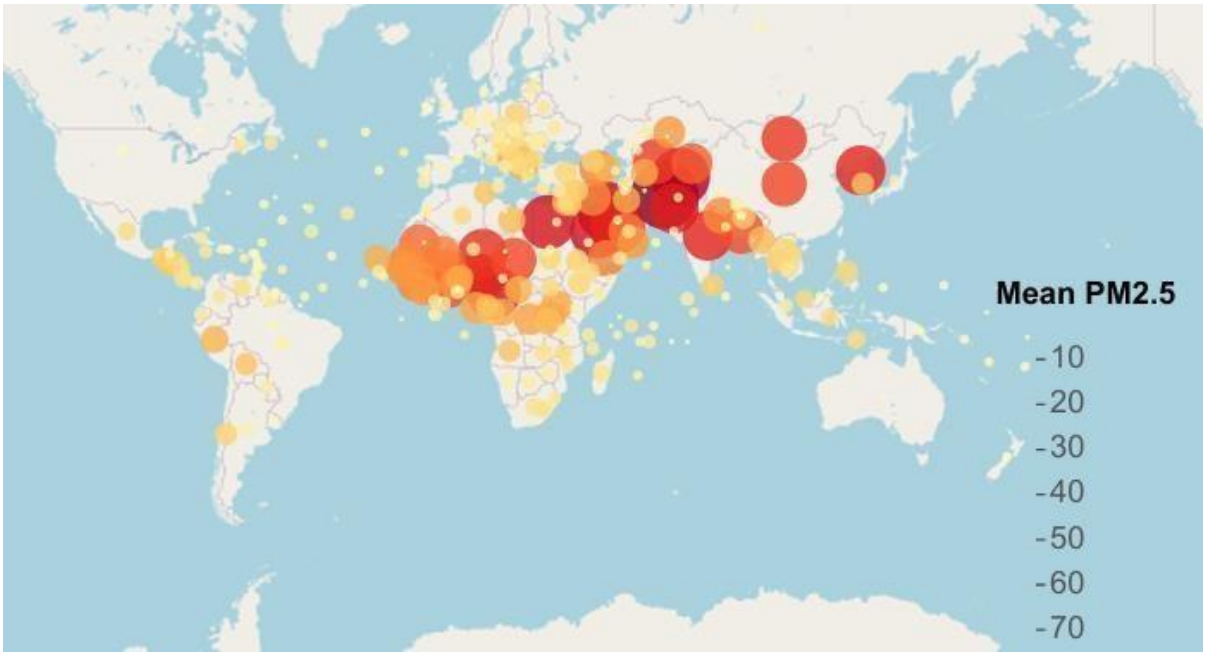
Insight: Cities consistently record higher PM2.5 concentrations.

4. PM2.5 time trend (Yearly mean)
   Compares regional pollution variability.



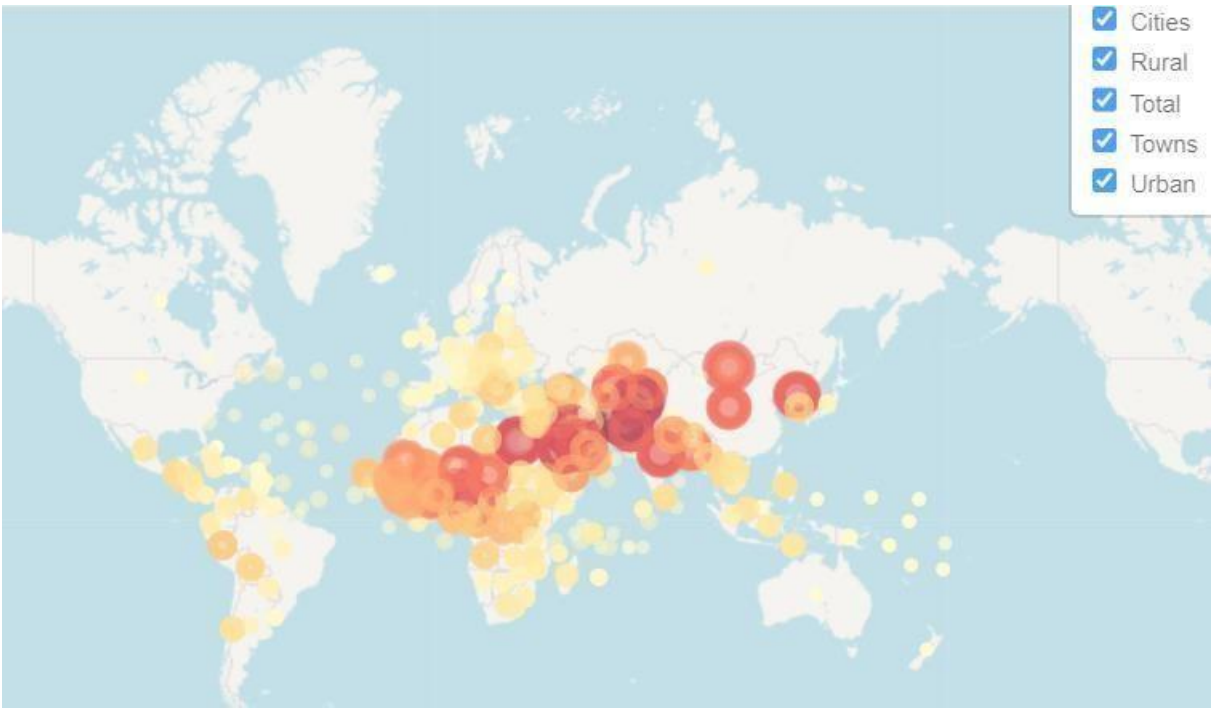
Insight: Displays global PM2.5 trends over time. Used to assess if pollution levels have decreased or stabilized in recent years.

5. Map — mean PM2.5 by country (Leaflet)



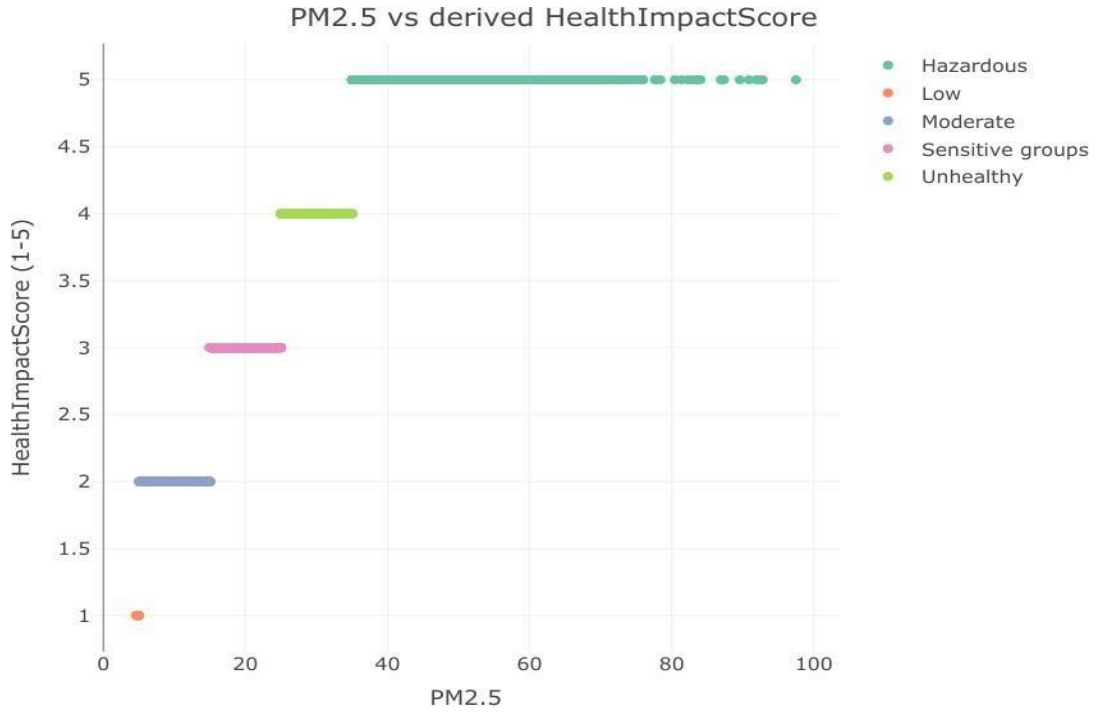
Insight: Provides a spatial overview of pollution hotspots; larger red markers indicate higher PM2.5.

6. Leaflet multi-layer view (ResidenceType filter via groups)



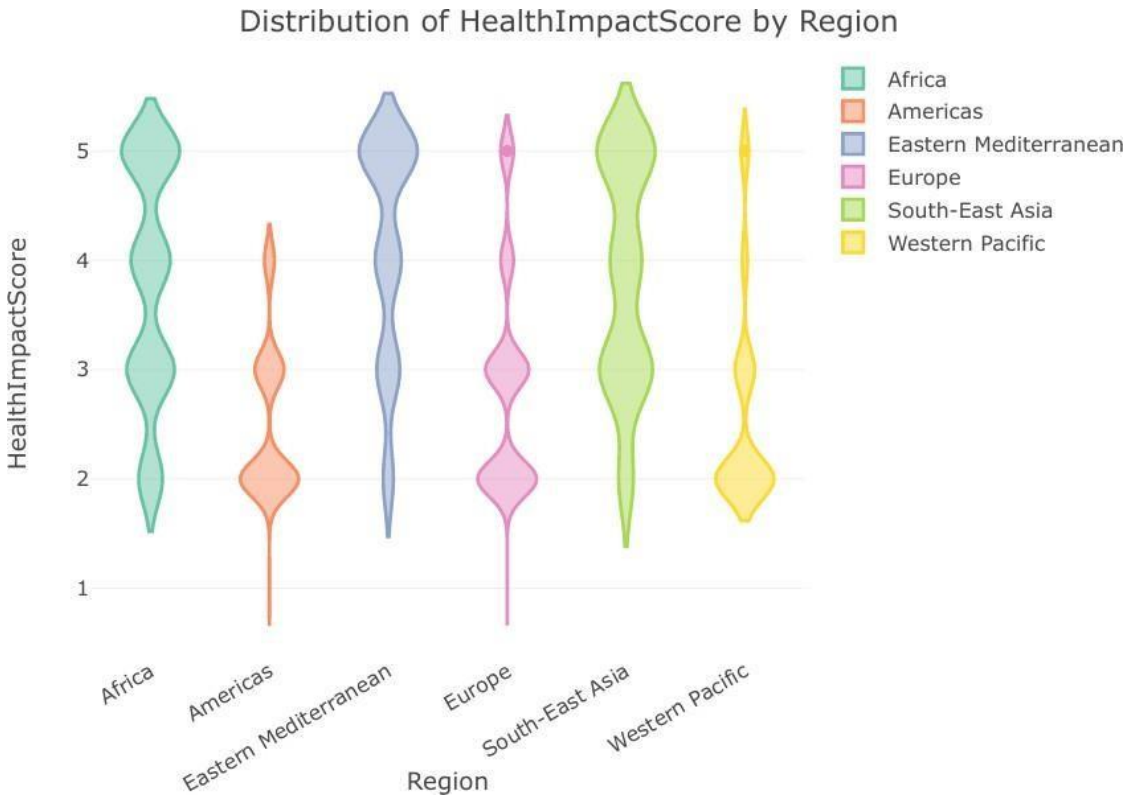
Insight: Enables filtering of pollution levels by residence category — interactive comparison between population zones.

7. PM2.5 vs Health Impact Score (Scatter Plot – Plotly)
   Shows correlation between PM2.5 and derived health risk.



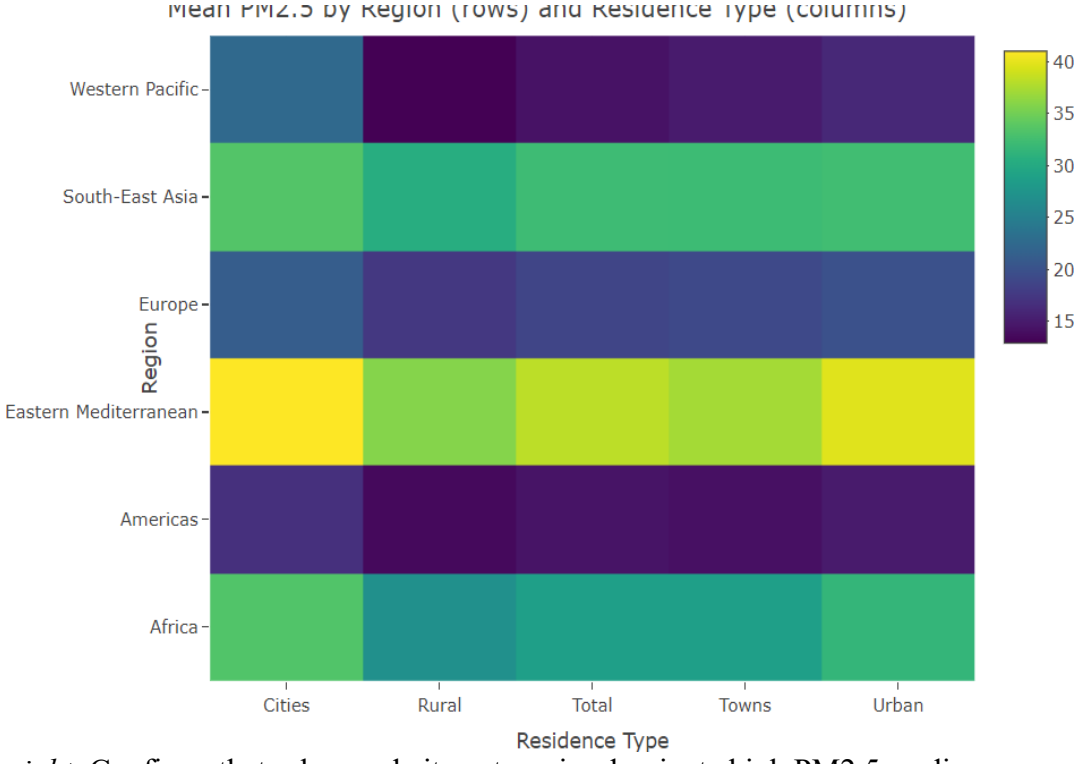
Insight: Strong positive trend validates that PM2.5 drives health impact.

8. Regional Health Impact Distribution (Violin Plot – Plotly)
   Visualizes how health scores vary by region.



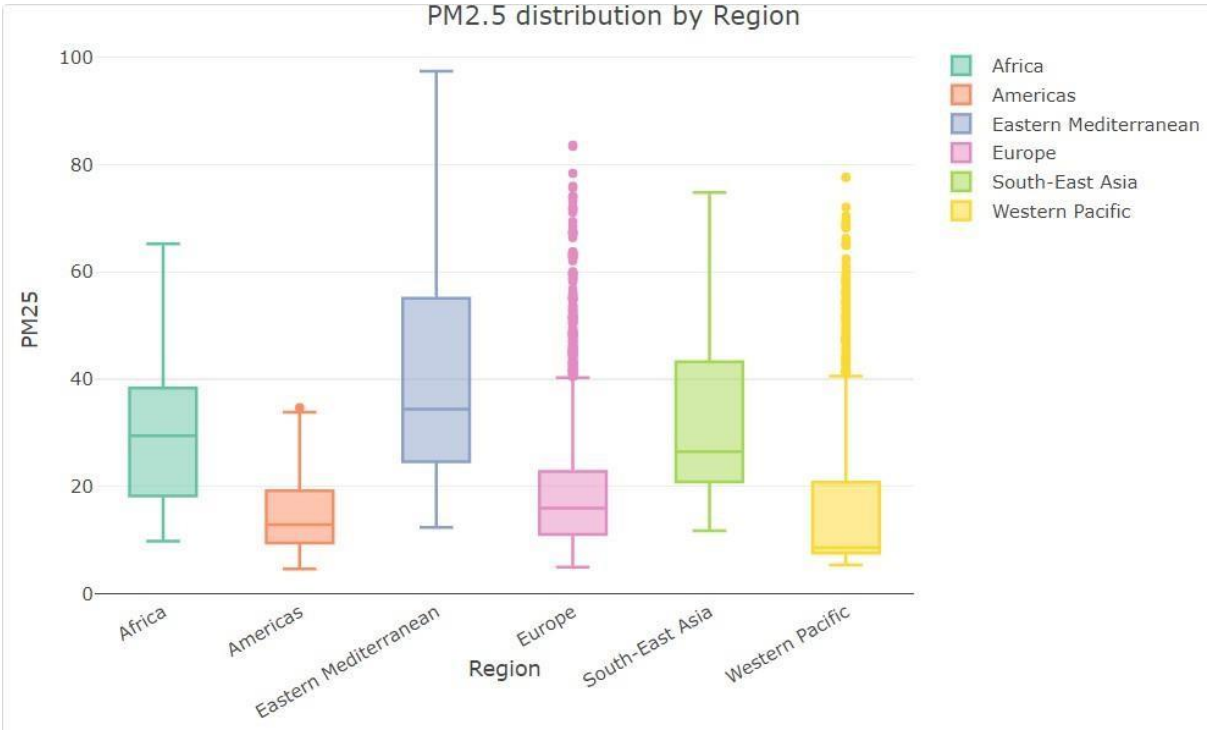
Insight: Demonstrates variability in health impact distribution across regions. Regions like Africa, south east Asia show broader, higher-risk distributions.

9. Heatmap: Region × Residence Type (Plotly)
Cross-analyzes mean PM2.5 between region and residence type.



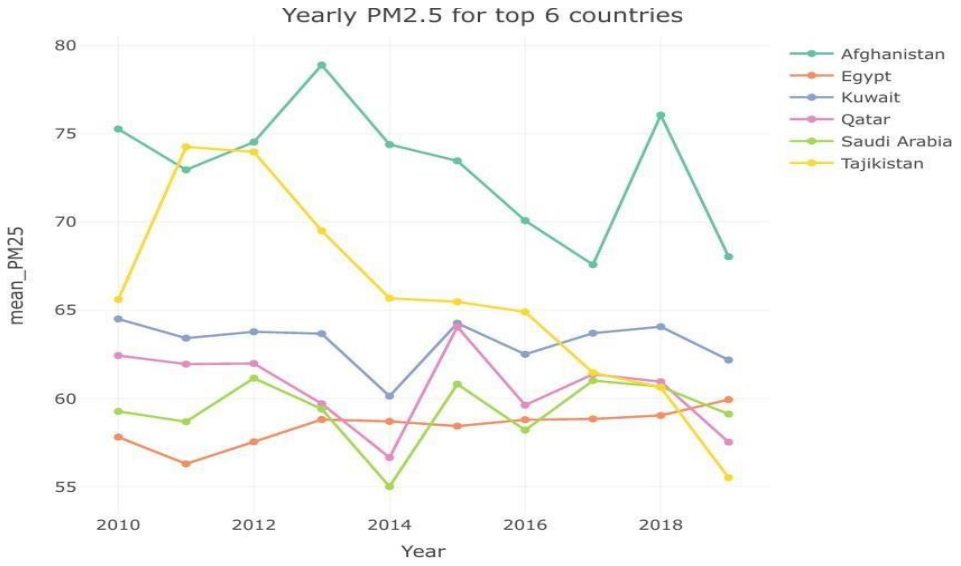
Insight: Confirms that urban and city categories dominate high PM2.5 readings.

10. Boxplot: PM2.5 by Region



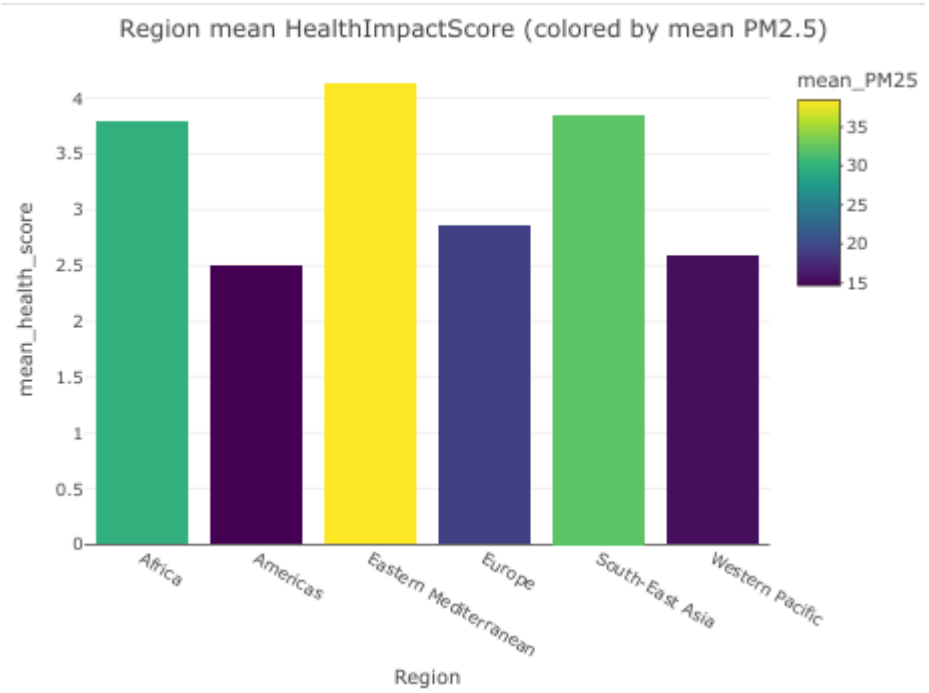
Insight: Confirms inter-regional disparity — Eastern Mediterranean and South-East Asia dominate the upper quartiles.

11. Small multiples: 6-country yearly trends (if Year exists)
Tracks changes in air quality for the six worst-affected countries.



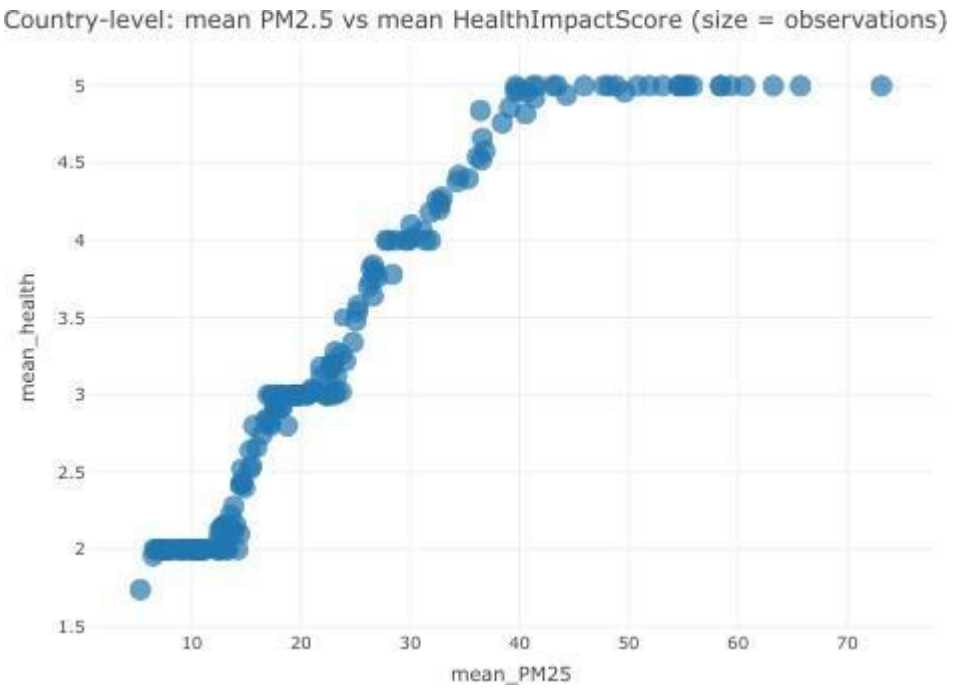
Insight: Shows if heavily polluted countries are improving or worsening over time.

12. Regional PM2.5 vs Health Impact Score (Bar Chart – Plotly)
Compares average regional PM2.5 and mean health scores side by side.



Insight: Confirms the causal link between pollution and health risk.

13. Bubble Chart (PM2.5 vs Health Score – Plotly)
    Displays mean PM2.5 (x-axis), mean health score (y-axis), bubble size = data count.



Insight: Large, high-positioned bubbles indicate countries with high risk and strong data reliability.

Summary of Insights

- PM2.5 concentration is the most influential parameter affecting health.
- Urbanization strongly correlates with pollution levels.
- Regions such as Africa and South-East Asia consistently exceed WHO limits.

- Interactive visualization allows multi-angle analysis — spatial, categorical, and temporal — leading to richer, evidence-based conclusions.

5. Conclusion

This project demonstrates that interactive visualization is essential for exploring and communicating environmental and health data effectively. By integrating WHO air quality data with advanced R visualization tools, the study uncovers clear global disparities in pollution levels and associated health risks.
Key insights include:

- PM2.5 is the primary driver of air-quality-related health risk.
- Urban areas consistently exhibit higher pollution than rural ones.
- Interactive tools (Plotly, Leaflet) provide deeper insight than static graphs.

Future Extensions

- Add other pollutants (NO_2, SO_2, CO, O_3) for multivariate analysis.
- Integrate real-time AQI data from APIs.
- Include hospital admission and mortality data for causal health modeling.
- Deploy as a R Shiny Dashboard for public access and live data monitoring.

6. References (APA Format)

World Health Organization. (2022). *Ambient Air Pollution Database (PM2.5)*. Retrieved from https://www.who.int/data/gho/data/themes/air-pollution
Plotly Technologies Inc. (2024). *Plotly R Open Source Graphing Library*. Retrieved from https://plotly.com/r/
Leaflet for R. (2024). *Interactive Maps with R*. RStudio. Retrieved from https://rstudio.github.io/leaflet/
Bokeh Developers. (2023). *Rbokeh: An R Interface to Bokeh Visualization Library*. Retrieved from https://hafen.github.io/rbokeh/
R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
WHO Global Health Observatory. (2022). *Air Pollution and Health Impact Indicators*. Retrieved from https://www.who.int/data/gho