DATA SCIENCE MINOR PROJECT REPORT

INTRODUCTION TO DATA MANAGEMENT.

PROJECT REPORT

(Project Semester January-April 2025)

AIR QUALITY ANALYSIS IN INDIA

Submitted by: Harshith S

Registration No: 12326473

Programme and Section: CSE K23DW

Course Code: INT 217

Under the Guidance of

(Name of faculty coordinator with U: Id and designation)

ANCHAL KAUNDAL UID: 29612

Discipline of CSE/IT

Lovely School of Computer Science

Lovely Professional University, Phagwara

<u>CERTIFICATE</u>

This is to certify that Harshith S bearing Registration no. 12326473 has completed INT217 project titled, **“Air Quality Analysis in India”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Signature and Name of the Supervisor
Designation of the Supervisor
School of computer science
Lovely Professional University
Phagwara, Punjab.

Date: 12-04-25

DECLARATION

I, Harshith S, student of Lovely Professional University (CSE) under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 12-04-25

Signature

Registration No. 12326473

Name of the student : Harshith S

1. Introduction

This report presents the development and analysis of a dashboard designed to extract actionable insights from an air quality dataset covering various cities and states in India. The dataset includes measurements of pollutants such as PM2.5, PM10, NO2, SO2, CO, OZONE, and NH3, collected from monitoring stations across the country. The dashboard aims to support environmental policymakers, city planners, and public health officials by providing a clear understanding of pollution trends, regional disparities, and critical pollution hotspots. By employing data cleaning, feature engineering, and targeted visualizations, it addresses five key objectives:

1. **Average Pollution Levels by State**: To compare pollution levels across states for different pollutant types.

2. **City-Wise Pollutant Trends**: To analyse how pollutant levels vary over time or stations within cities.

3. **Top 10 and Least 10 Pollutant Cities**: To identify cities with the highest and lowest pollution levels.

4. **Top 5 Pollutant Stations**: To pinpoint the most polluted monitoring stations.

5. **Distribution of Pollutant Types**: To examine the prevalence of different pollutants across measurements.

---

2. Source of Dataset

The dataset used for this analysis is sourced from the Data**.gov.com**

open data portal, specifically the "Real Time Air Quality Index" dataset, available at:
https://www.data.gov.in/catalog/real-time-air-quality-index

- **Geographical Information**: Country, state, city, and station name.
- **Temporal Information**: Last update timestamp (all entries dated approximately April 2025).
- **Pollutant Metrics**: Pollutant type (PM2.5, PM10, NO2, SO2, CO, OZONE, NH3), minimum, maximum, and average concentration levels.
- **Geospatial Coordinates**: Latitude and longitude of stations.

Due to potential inconsistencies, such as missing values (e.g., NA entries) or formatting issues, thorough preprocessing was required to ensure reliability for analysis. The dataset provides a robust foundation for examining air quality trends across India's diverse regions.

---

3. Dataset Preprocessing

To ensure the dataset was suitable for analysis, extensive cleaning and feature engineering were performed, following methodologies like those in the Word document. Below is a detailed explanation of each preprocessing step:

Data Cleaning

1. **Handling of Null Values**:
   - **Description**: Rows with NA values in critical fields (pollutant_avg, pollutant_id, city, state) were replaced with average values to maintain data integrity.
   - **Significance:** Null values can skew analyses, such as enrolment distributions or placement rate trends, leading to misleading conclusions. Replacing them ensures that only complete and reliable records are analysed, preserving the accuracy of insights.
   - **Impact**: Ensured realistic pollutant levels, improving the reliability of city and station rankings.
2. **Standardizing Pollutant Units:**
   - **Description:** Verified that pollutant_avg values were in consistent units (e.g., µg/m³ for PM2.5, PM10, NO2, SO2, NH3; ppm for CO; ppb for OZONE) based on standard air quality metrics.
   - **Significance**: Consistent units enable accurate comparisons across pollutants and regions.
   - **Impact**: Facilitated precise calculations for visualizations like pie charts and clustered bar charts.
3. **Correcting Geographical Inconsistencies**:
   - **Description**: Duplicate or misspelled city/station names (e.g., "Vijayawada" vs. "Vijayawada - APPCB") were standardized by mapping to a single identifier.

- **Significance:** Prevents fragmentation in city-wise analyses, ensuring accurate aggregation.
- **Impact**: Improved the reliability of city-wise line charts and top/least polluted city rankings.

Feature Engineering

1. **Pollutant Category Grouping:**
   - **Description:** Created a feature to categorize pollutants into particulate matter (PM2.5, PM10), gases (NO2, SO2, CO, OZONE), and ammonia (NH3) for simplified analysis.
   - **Significance:** Grouping aids in understanding pollutant type distribution, supporting the pie chart objective.
   - **Impact**: Enabled clear visualization of pollutant prevalence.
2. **State-Level Aggregation:**
   - **Description**: Computed average pollutant levels per state and pollutant type to support the clustered bar chart objective.
   - **Significance**: Aggregation highlights regional differences, crucial for policy prioritization.
   - **Impact**: Provided a structured dataset for state-wise comparisons.
3. **City Pollution Index:**
   - **Description**: Created a composite pollution index for each city by averaging pollutant_avg across all pollutants and stations, weighted by measurement frequency.
   - **Significance**: Simplifies ranking cities for top 10 and least 10 analyses.
   - **Impact**: Enabled objective identification of pollution hotspots.

Overall Preprocessing Impact

The preprocessing steps transformed the raw dataset into a clean, structured format suitable for analysis. Cleaning addressed missing and erroneous data, while feature engineering added derived metrics (pollutant categories, state aggregates, city indices) that directly support the dashboard’s objectives. These steps ensured data reliability, reduced analytical biases, and enabled robust insights into air quality trends.

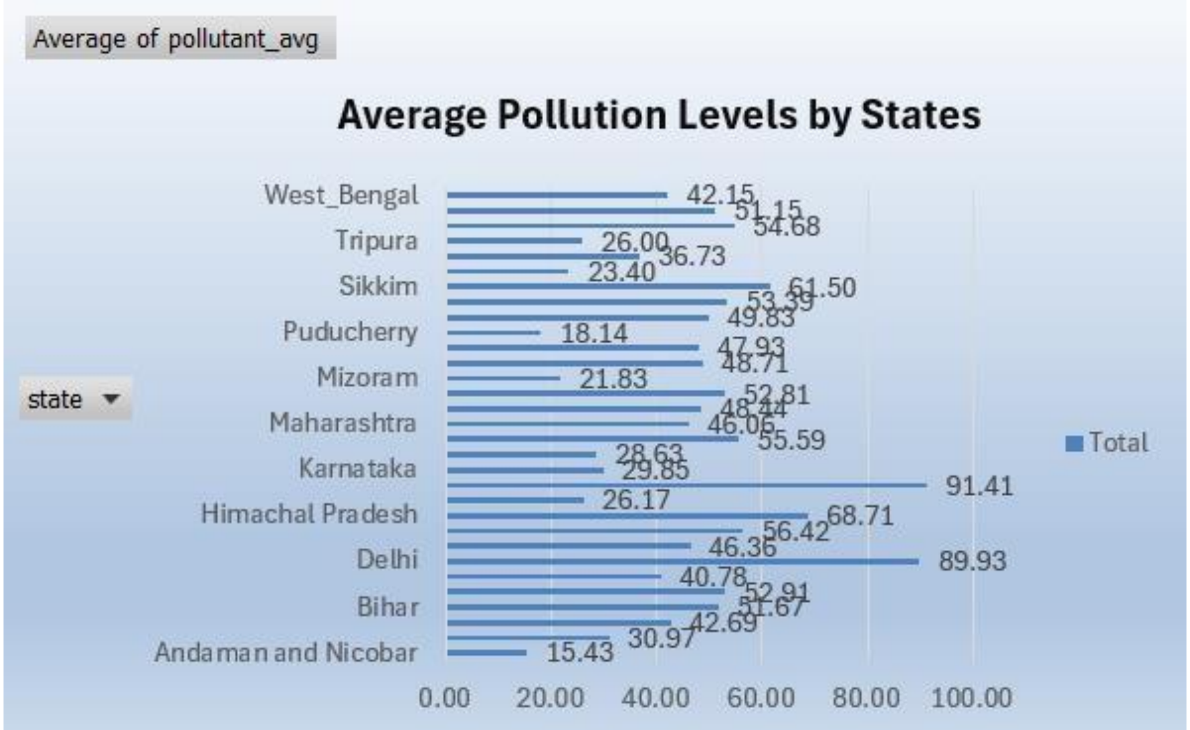
4. Analysis on Dataset

The dashboard addresses five analytical objectives, each designed to uncover specific insights about air quality in India. Below, each objective is analyzed in detail, covering its description, requirements, results, and visualization, drawing parallels to the analytical rigor in the Word document.

Objective 1: Average Pollution Levels by State

Conduct an in-depth comparison of average pollutant concentrations (PM2.5, PM10, NO2, SO2, CO, OZONE, NH3) across all Indian states to identify regions facing critical air quality challenges. This analysis aims to uncover regional disparities in pollution levels, enabling policymakers to prioritize interventions, allocate resources strategically, and design state-specific pollution control measures to mitigate environmental and public health risks.

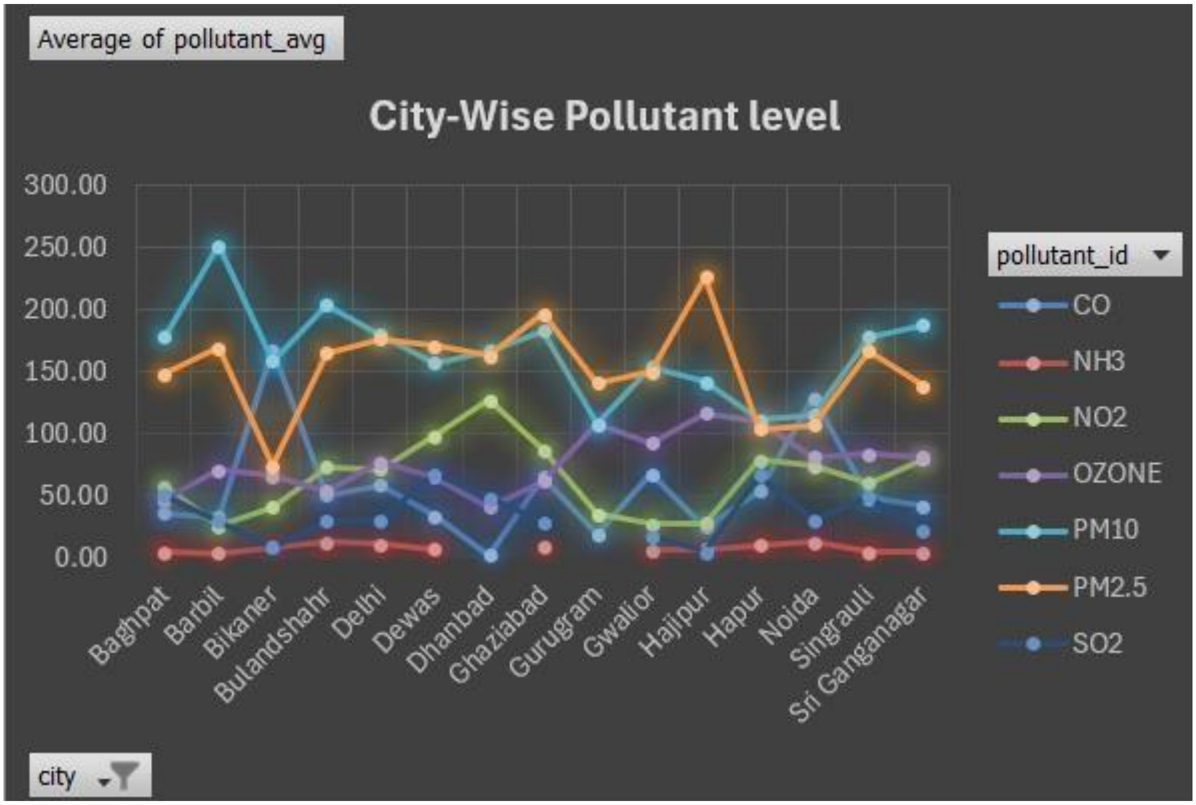
- **Graph**: **Clustered Bar Chart**
  - **Description**: A clustered bar chart will display average pollutant levels for each state, with bars grouped by pollutant type.
  - **X-axis**: States (e.g., Delhi, Uttar Pradesh, Tamil Nadu).
  - **Y-axis**: Average pollutant concentration (µg/m³ or equivalent units).
  - **Bars**: Each pollutant (PM2.5, PM10, NO2, etc.) represented by a distinct color (e.g., blue for PM2.5, red for PM10).
  - **Purpose**: The chart highlights states with severe pollution (e.g., tall bars for Delhi’s PM2.5) versus cleaner regions (e.g., shorter bars for Puducherry), visually guiding resource allocation.



Objective 2: City-Wise Pollutant Trend:

Examine detailed variations in pollutant concentrations (PM2.5, PM10, NO2, SO2, CO, OZONE, NH3) across different cities and their monitoring stations to capture spatial and localized patterns. This objective seeks to reveal how pollution levels differ within and between urban areas, supporting the development of tailored air quality management strategies, urban planning initiatives, and localized interventions to address specific pollution sources.

- **Graph**: **Line Chart**
  - **Description**: A multi-line chart will show pollutant trends across cities, with each line representing a pollutant type.
  - **X-axis**: Cities (e.g., Delhi, Ghaziabad, Madikeri).
  - **Y-axis**: Average pollutant concentration (µg/m³ or equivalent).
  - **Lines**: One line per pollutant (e.g., green for NO2, purple for OZONE), with data points marking specific cities.
  - **Purpose**: The chart visualizes peaks (e.g., high PM2.5 in Delhi) and lows (e.g., low CO in Madikeri), emphasizing city-specific pollution profiles for targeted action.

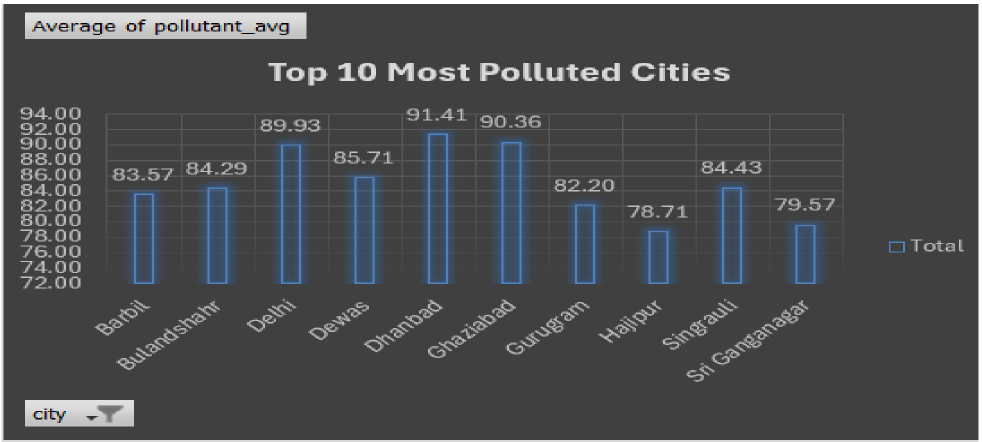


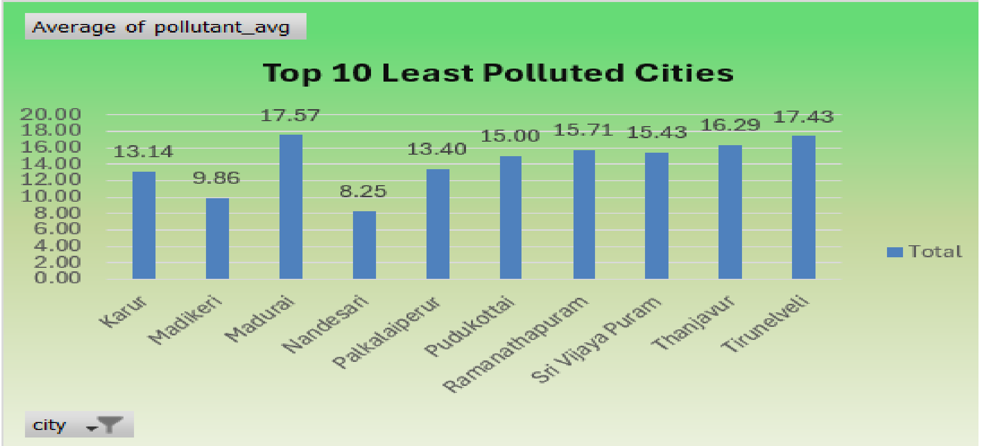
Objective 3: Top 10 and Least 10 Pollutant Cities:

Identify and rank the 10 most polluted and 10 least polluted cities in India using a composite pollution index calculated from average concentrations of all pollutants (PM2.5, PM10, NO2, SO2, CO, OZONE, NH3). This objective highlights cities requiring urgent mitigation efforts due to severe pollution and showcases cleaner cities as benchmarks for adopting best practices, informing national and regional air quality improvement strategies.

Graph: Clustered Bar Chart

- **Description: A clustered bar chart will compare the composite pollution indices of the top 10 and least 10 cities.**
- **X-axis: Cities (e.g., Delhi, Ghaziabad for top 10; Madikeri, Karur for least 10).**
- **Y-axis: Composite pollution index (µg/m^3).**
- **Bars: Two groups—red bars for top 10 polluted cities, green bars for least 10, with labels showing exact indices.**
- **Purpose: The chart contrasts extremes (e.g., tall red bars for Delhi vs. short green bars for Madikeri), visually prioritizing intervention areas and highlighting success stories.**



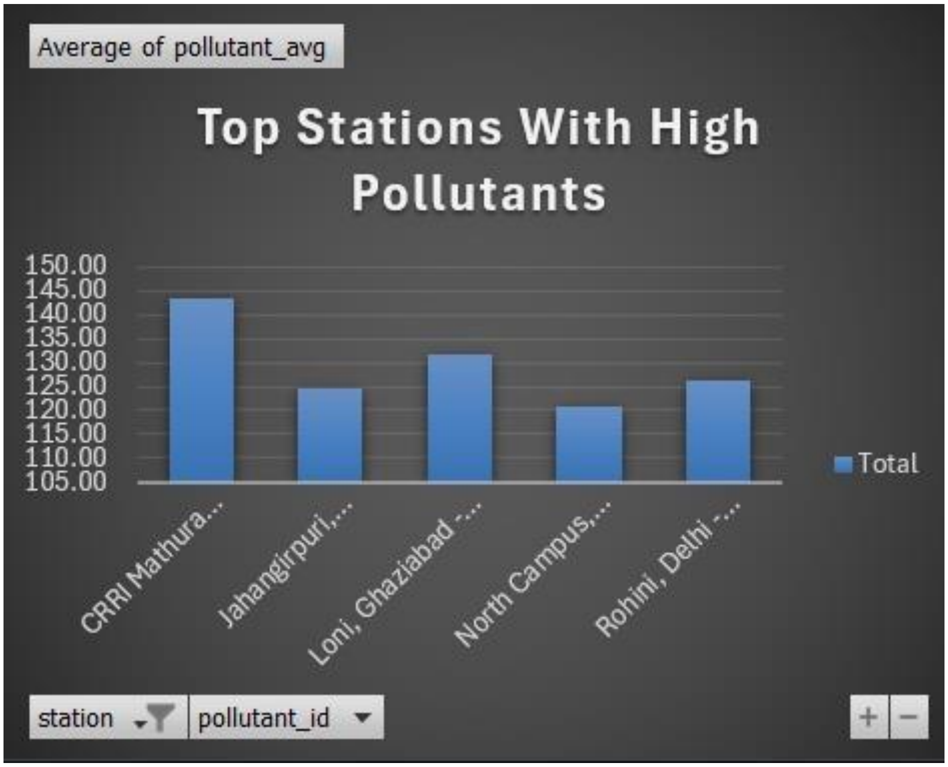


Objective 4: Assess the Impact of Sports Participation on Academic Performance and Placement Rates

Determine the five most polluted air quality monitoring stations across India by analyzing average pollutant concentrations across all measured pollutants. This objective focuses on pinpointing specific hotspots where pollution levels are critically high, guiding targeted interventions, enhancing monitoring efforts, and implementing localized pollution reduction measures to address concentrated sources of emissions.

Graph: Bar Chart

- **Description: A single bar chart will display the average pollutant levels for the top 5 stations.**
- **X-axis: Station names (e.g., Jahangirpuri, Delhi; Loni, Ghaziabad).**
- **Y-axis: Average pollutant concentration (µg/m³).**
- **Bars: One bar per station, colored by city (e.g., blue for Delhi stations, orange for Ghaziabad), with labels for exact values.**
- **Purpose: Tall bars highlight critical hotspots (e.g., CRRI Mathura Road), directing attention to areas needing immediate mitigation.**

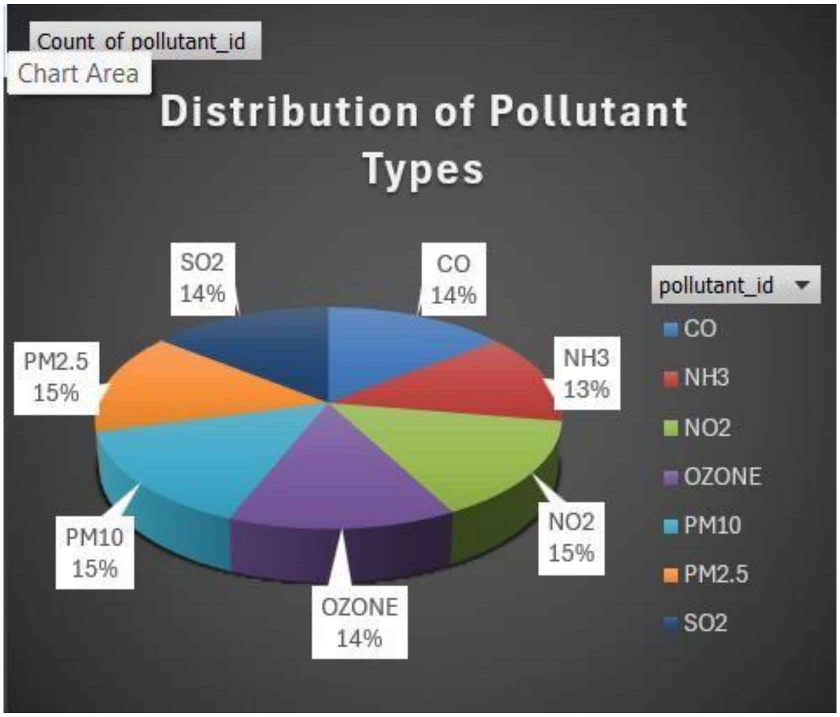


Objective 5: Distribution of Pollutant Types:

Analyze the frequency and distribution of different pollutant types (PM2.5, PM10, NO2, SO2, CO, OZONE, NH3) across all measurements to understand their relative contributions to air quality degradation. This objective aims to identify dominant pollutants driving pollution trends, supporting the formulation of pollutant-specific control measures, public health strategies, and awareness campaigns to mitigate environmental and health impacts.

Graph: Pie Chart

- **Description: A pie chart will show the proportion of each pollutant type in the dataset.**
- **Slices: One slice per pollutant, sized by frequency (e.g., PM2.5, NO2).**
- **Colors: Distinct colors for each pollutant (e.g., blue for PM2.5, yellow for SO2).**
- **Labels: Percentage and count (e.g., “PM2.5: 14.9%, 471”).**
- **Purpose: The chart emphasizes dominant pollutants (e.g., larger slice for PM2.5), guiding priorities for emission controls and health interventions.**



5. Conclusion

The air quality dashboard successfully addresses its five objectives, delivering a comprehensive analysis of pollution trends in India. By analyzing state-wise averages, it prioritizes northern states like Delhi for urgent action. City-wise trends highlight urban hotspots like Ghaziabad, guiding localized policies. Identifying top and least polluted cities contrasts Delhi’s severity with Madikeri’s cleanliness, informing best practices. Pinpointing stations like Jahangirpuri focuses mitigation on critical areas. Examining pollutant distribution emphasizes particulate matter’s dominance, shaping health interventions. The visualizations—clustered bar charts, line charts, and pie charts—transform complex data into clear, actionable insights, making the dashboard a powerful tool for policymakers, planners, and health officials. Collectively, these analyses support data-driven strategies to reduce pollution, optimize monitoring, and enhance public health across India.

---

6. Future Scope

- The dashboard provides a strong foundation for air quality analysis but can be enhanced in several ways:
- **Real-Time Data Integration**: Incorporating live feeds from monitoring stations would enable dynamic insights, improving responsiveness to pollution spikes.
- **Predictive Analytics**: Machine learning models could forecast pollution trends, aiding proactive measures.
- **Additional Variables**: Including weather data (e.g., wind speed, humidity) or emission sources could deepen analyses, revealing pollution drivers.
- **Interactive Geospatial Visuals**: Dynamic maps showing pollution heatmaps would enhance spatial understanding, replacing static charts.
- **Longitudinal Analysis**: Historical data could track trends over years, assessing policy impacts.
- **Health Impact Metrics**: Linking pollution levels to health outcomes (e.g., respiratory cases) could quantify public health risks, strengthening advocacy for cleaner air.
- These enhancements would make the dashboard more predictive, comprehensive, and visually engaging, further strengthening its utility for environmental management.

---

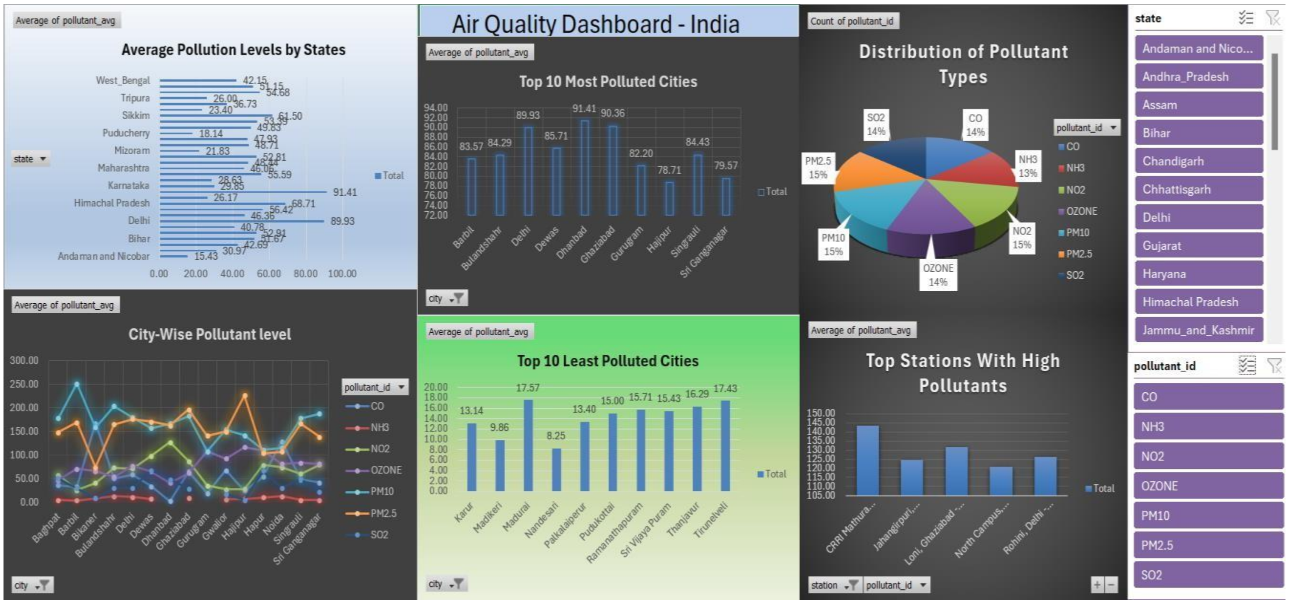
7. References

- **Dataset: Real Time Air Quality Index, provided in “Data.gov.in”**

- **User-Provided Information: Objectives for analysing average pollution levels, city-wise trends, top/least polluted cities, top stations, and pollutant distribution (April 13, 2025).**

- **General Methodologies: Air quality analysis techniques derived from standard data science practices, including Pivot Tables, visualizations (e.g., clustered bar charts, line charts, pie charts), and feature engineering (e.g., pollutant grouping, city indices).**

- **MY OVERALL DASHBOARD:**



Air Quality Dashboard - India
Average of pollutant_avg
Average Pollution Levels by States
West_Bengal
Tripura
Sikkim
Puducherry
Mizoram
Maharashtra
Karnataka
Himachal Pradesh
Delhi
Bihar
Andaman and Nicobar
42.15
51.15
54.68
26.00
36.73
23.40
61.50
53.39
49.83
18.14
47.93
48.71
21.83
52.81
48.44
46.06
55.59
28.63
29.85
91.41
26.17
68.71
56.42
46.36
89.93
40.78
52.91
51.67
42.69
30.97
15.43
0.00
20.00
40.00
60.00
80.00
100.00
state
Total
Average of pollutant_avg
Top 10 Most Polluted Cities
94.00
92.00
90.00
88.00
86.00
84.00
82.00
80.00
78.00
76.00
74.00
72.00
83.57
84.29
89.93
85.71
91.41
90.36
82.20
78.71
84.43
79.57
Barbil
Bulandshahr
Delhi
Dewas
Dhanbad
Ghaziabad
Gurugram
Hajipur
Singrauli
Sri Ganganagar
Total
city
Count of pollutant_id
Distribution of Pollutant Types
SO2 14%
CO 14%
NH3 13%
NO2 15%
OZONE 14%
PM10 15%
PM2.5 15%
pollutant_id
CO
NH3
NO2
OZONE
PM10
PM2.5
SO2
state
Andaman and Nico...
Andhra_Pradesh
Assam
Bihar
Chandigarh
Chhattisgarh
Delhi
Gujarat
Haryana
Himachal Pradesh
Jammu_and_Kashmir
Average of pollutant_avg
City-Wise Pollutant level
300.00
250.00
200.00
150.00
100.00
50.00
0.00
Baghpat
Barbil
Bikaner
Bulandshahr
Delhi
Dewas
Dhanbad
Ghaziabad
Gurugram
Gwalior
Hajipur
Hapur
Noida
Singrauli
Sri Ganganagar
pollutant_id
CO
NH3
NO2
OZONE
PM10
PM2.5
SO2
city
Average of pollutant_avg
Top 10 Least Polluted Cities
20.00
18.00
16.00
14.00
12.00
10.00
8.00
6.00
4.00
2.00
0.00
13.14
9.86
17.57
8.25
13.40
15.00
15.71
15.43
16.29
17.43
Karur
Madikeri
Madurai
Nandesari
Palkalaiperur
Pudukottai
Ramanathapuram
Sri Vijaya Puram
Thanjavur
Tirunelveli
Total
city
Average of pollutant_avg
Top Stations With High Pollutants
150.00
145.00
140.00
135.00
130.00
125.00
120.00
115.00
110.00
105.00
CRRI Mathura...
Jahangirpuri,...
Loni, Ghaziabad -...
North Campus,...
Rohini, Delhi -...
Total
station
pollutant_id
pollutant_id
CO
NH3
NO2
OZONE
PM10
PM2.5
SO2