

**PYTHON PROJECT REPORT**

(Project Semester: January-April 2025)

**Title of the Project: Air Pollution Based On Real Time Data**

**Submitted by:**

**Anisha Kumari  
Registration No.: 12306607  
Programme and Section: B.Tech CSE (K23FD)  
Course Code: INT375**

**Under the Guidance of:  
Baljinder Kaur (UID : 27952)**

**Discipline of CSE/IT**  
**Lovely School of Computer Science & Engineering**  
**Lovely Professional University, Phagwara**

**DECLARATION**

I, **Anisha Kumari**, student of **Bachelor of Technology (B.Tech)** 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: 03-April-2025

Signature: A close-up of a signature

AI-generated content may be incorrect.  
Registration No.: 12306607  
Name of the Student: **Anisha Kumari**

# ****CERTIFICATE****

This is to certify that **Anisha Kumari** bearing Registration No. **12306607** has completed **INT375** project titled **“Air Pollution”** under my guidance and supervision. To the best of my knowledge, the present work is the result of her original development, effort, and study.

**Baljinder Kaur**  
**Assistant Professor**  
**School of Computer Science & Engineering**

**Lovely Professional University**  
**Phagwara, Punjab**

Date: **04-April-2025**

**ACKNOWLEDGMENT**

I would like to express my sincere gratitude to **Baljinder Kaur Ma’am**, my project guide, for their invaluable support, guidance, and encouragement throughout the development of this project. Their expert insights and constructive feedback have been instrumental in shaping the project's outcome.

I am also thankful to **Lovely Professional University** for providing a conducive learning environment and access to resources that made this project possible. Additionally, I extend my appreciation to my professors and peers for their continuous motivation and insightful discussions, which greatly enhanced my understanding of the subject.

Lastly, I would like to acknowledge the unwavering support of my family and friends, whose encouragement has been a source of inspiration throughout this journey.

# ****TABLE OF CONTENTS****

1. [Introduction](#_1._INTRODUCTION)
2. [Source of Dataset](#_2._SOURCE_OF)
3. [Dataset Preprocessing](#_3._DATASET_PREPROCESSING)
4. [Analysis on Dataset (for each objective)](#_4._ANALYSIS_ON)
   * [i. General Description](#_i._General_Description)
   * [ii. Specific Requirements](#_ii._Specific_Requirements)
   * [iii. Analysis Results](#_iii._Analysis_Results)
   * [iv. Visualization](#_iv._Visualization)
5. [Conclusion](#_5._CONCLUSION)
6. [Future Scope](#_6._FUTURE_SCOPE)
7. [References](#_7.REFERENCES)

# ****1. INTRODUCTION****

In the modern era of data-driven decision-making, organizations increasingly rely on data analysis to understand business trends, customer behavior, and sales performance. One of the most effective techniques for uncovering insights from raw datasets is **Exploratory Data Analysis (EDA)**. EDA plays a crucial role in identifying patterns, spotting anomalies, testing hypotheses, and checking assumptions through statistical summaries and visualizations.

This project, titled **“Air Pollution”**, leverages the power of **Exploratory Data Analysis (EDA)** to study and interpret sales data efficiently. The objective of the project is not only to build a system that displays and manages sales records but also to provide valuable insights into sales performance through data exploration. Python, with its rich ecosystem of data analysis libraries such as **Pandas**, **Matplotlib**, and **Seaborn**, offers an ideal platform for implementing EDA-based solutions.

Air quality is a critical environmental concern, impacting public health and sustainability. This project leverages Python to analyse air quality data from monitoring stations across India, focusing on pollutants like PM2.5, PM10, NO2, SO2, OZONE, and CO. The dataset, sourced from CLEANED.csv, includes measurements from various cities and states, updated as of April 3, 2025. The goal is to uncover patterns, identify pollution hotspots, and explore relationships between pollutants and geographic factors.

# ****2. SOURCE OF DATASET****

The dataset, <https://www.data.gov.in/catalog/real-time-air-quality-index>, contains air quality measurements from monitoring stations across India, recorded on April 3, 2025. It includes the following columns:

* **Country**: India (uniform across entries).
* **State**: Administrative regions (e.g., Delhi, Uttar Pradesh, Karnataka).
* **City**: Urban centers or towns (e.g., Visakhapatnam, Guwahati, Patna).
* **Station**: Specific monitoring locations (e.g., “Gandhi Nagar, Guwahati”).
* **Last Update**: Timestamp of data collection (e.g., “2025-04-03”).
* **Latitude, Longitude**: Geographic coordinates of stations.
* **Pollutant ID**: Identifies pollutants (PM2.5, PM10, NO2, SO2, OZONE, NH3, CO).
* **Min Pollution, Max Pollution, Avg Pollution**: Concentration metrics (assumed in µg/m³ for consistency with standard air quality reporting).

The dataset spans diverse geographies, capturing both metropolitan hubs and smaller towns, making it ideal for regional and urban-rural comparisons. While the exact source isn’t specified in the provided files, the structure suggests it’s a curated dataset, likely aggregated from regulatory bodies like the Central Pollution Control Board (CPCB), State Pollution Control Boards, or open-source platforms like OpenAQ. For this report, I assume it’s a reliable public dataset, pre-cleaned to some extent but requiring further preprocessing for analysis.

# ****3. DATASET PREPROCESSING****

High-quality data is the foundation of meaningful analysis. The CLEANED.csv dataset, while structured, required preprocessing to address common issues like missing values, outliers, and inconsistencies. The following steps were executed using Python’s pandas library in a Jupyter Notebook environment:

* **Missing Values**:
  + Inspected for nulls in critical columns (state, city, pollutant\_id, avg\_pollution).
  + [Placeholder: e.g., “Found X% missing avg\_pollution values. Dropped rows where pollutant\_id or avg\_pollution was null (Y rows), as these were essential. For minor gaps in latitude/longitude, imputed with state-level averages.”]
  + Rationale: Dropping essential nulls ensures accuracy, while imputation preserves geographic context.
* **Outlier Detection and Treatment**:
  + Calculated z-scores for avg\_pollution per pollutant to identify anomalies.
  + [Placeholder: e.g., “Capped values beyond 3 standard deviations (e.g., PM2.5 > 500 µg/m³) to the 99th percentile, affecting Z rows, to mitigate sensor errors or extreme weather effects.”]
  + Rationale: Outliers can skew statistical measures like means, so capping preserves data integrity.
* **Data Type Validation**:
  + Ensured avg\_pollution, min\_pollution, and max\_pollution were float64 for numerical analysis.
  + Converted state, city, station, and pollutant\_id to strings for categorical consistency.
  + Verified last\_update as a datetime object (parsed to “2025-04-03”).
  + [Placeholder: e.g., “Corrected N rows where avg\_pollution was misread as strings due to formatting errors.”]
* **Duplicate Handling**:
  + Checked for duplicate station-pollutant-timestamp combinations.
  + [Placeholder: e.g., “Removed X duplicate entries (same station, pollutant\_id, and last\_update) to prevent overrepresentation, reducing dataset by Y%.”]
  + Rationale: Duplicates can bias averages, especially for state-level aggregates.
* **Standardization**:
  + Unified naming conventions (e.g., “Uttar Pradesh” vs. “UP”, “O3” to “OZONE”).
  + Corrected typos in city/station names (e.g., “Gawhati” to “Guwahati”).
  + [Placeholder: e.g., “Standardized Z station names using fuzzy matching, resolving N inconsistencies.”]
* **Derived Columns** (if applicable):
  + For Objective 4.7 (urban vs. rural CO), created an urban\_rural column.
  + [Placeholder: e.g., “Classified stations as ‘urban’ if city population > 1 million or name included ‘Industrial’/’City’; else ‘rural’. Used external population data or station name patterns, affecting X stations.”]
  + Rationale: Enables urban-rural comparisons without external datasets.
* **Verification**:
  + Post-cleaning, validated dataset shape and summary statistics.
  + [Placeholder: e.g., “Final dataset had X rows, Y stations, Z states, and N pollutants, ready for analysis.”]

Preprocessing was iterative, with checks at each step to ensure no data loss compromised the objectives. The cleaned dataset was saved as a new CSV or retained in memory for analysis, ensuring reproducibility.

# ****4. ANALYSIS ON DATASET****

The analysis addresses nine objectives, each structured with a general description, specific requirements, placeholder results, and visualization descriptions (without code, per your request). The process utilized pandas for data wrangling, NumPy for numerical computations, and Matplotlib/Seaborn for visualizations, executed in a Jupyter Notebook environment. Below, I detail the methodology and anticipated outputs for each objective, leaving room for your specific findings.

**4.1 Distribution of PM2.5 Levels Across India**

**i. General Description**

PM2.5, fine particulate matter less than 2.5 micrometers, is a leading air quality concern due to its ability to penetrate deep into the respiratory system, causing asthma, heart disease, and premature mortality. This analysis examines the nationwide distribution of PM2.5 concentrations to understand its prevalence, variability, and whether most areas meet World Health Organization guidelines (e.g., <10 µg/m³ annual mean). By studying the spread, we can identify the extent of hazardous exposure and prioritize mitigation efforts.

**ii. Specific Requirements**

* Selected all records where the pollutant identifier was PM2.5.
* Calculated descriptive statistics for the average pollution metric, including mean, median, standard deviation, minimum, and maximum.
* Analyzed the shape of the distribution to determine if PM2.5 levels are uniform, skewed, or clustered.
* Created a visualization to display the frequency of PM2.5 concentrations across stations.

**iii. Analysis Results**

[Placeholder: e.g., “Across X monitoring stations, PM2.5 concentrations ranged from Y µg/m³ to Z µg/m³, with a mean of A µg/m³ and a median of B µg/m³. The standard deviation was C µg/m³, reflecting significant variability. The distribution appeared right-skewed, with most stations reporting moderate levels (20–60 µg/m³) but a notable tail of extreme values (>150 µg/m³), likely in urban or industrial hubs. This suggests widespread exposure to unsafe PM2.5, with pockets of severe pollution requiring urgent action.”]

**iv. Visualization**

* **Type**: Histogram with Kernel Density Estimate (KDE).
* **Description**: The visualization features a histogram, where bars represent the frequency of PM2.5 concentrations within discrete intervals (bins), typically 30 for granularity. Overlaid is a KDE curve, which smooths the distribution to highlight its shape—whether normal, skewed, or multimodal. The x-axis shows PM2.5 levels in µg/m³, and the y-axis indicates frequency or density. A light grid was added for readability, and the plot was titled “Distribution of PM2.5 Levels Across India (April 3, 2025)” to contextualize the data. The histogram likely reveals a concentration of moderate values with outliers, making the KDE essential for trend clarity.
* **Figure**: [Placeholder: Insert saved plot, e.g., “Figure\_4.1.png”.]

**4.2 Average PM2.5 Levels by State**

**i. General Description**

PM2.5 levels vary regionally due to differences in urbanization, industrial activity, and environmental factors like wind patterns. This analysis ranks states by their average PM2.5 concentrations to identify pollution hotspots, enabling policymakers to target high-risk areas with measures like stricter vehicle emission standards or industrial regulations. It also sheds light on geographic disparities, such as whether northern states face worse air quality than southern ones.

**ii. Specific Requirements**

* Filtered the dataset for PM2.5 records.
* Grouped data by state and computed the mean of the average pollution metric for each state.
* Sorted states in descending order to highlight those with the highest PM2.5 levels.
* Developed a visualization to compare state averages side by side.

**iii. Analysis Results**

[Placeholder: e.g., “Delhi recorded the highest average PM2.5 at X µg/m³ across Y stations, followed by Uttar Pradesh (A µg/m³) and Haryana (B µg/m³). In contrast, southern states like Tamil Nadu and Karnataka reported lower averages (C and D µg/m³, respectively), indicating cleaner air. The top five states exceeded the national annual standard of 40 µg/m³, underscoring the need for regional interventions. Northern states dominated the upper ranks, possibly due to dense urbanization and seasonal factors like crop burning.”]

**iv. Visualization**

* **Type**: Bar plot.
* **Description**: The bar plot displays each state along the y-axis, with horizontal bars extending along the x-axis to represent average PM2.5 concentrations in µg/m³. States are ordered from highest to lowest average, making it easy to spot outliers like Delhi or cleaner regions like Kerala. A color palette (e.g., shades of green to purple) differentiates bars, with longer bars indicating worse pollution. A grid along the x-axis helps estimate values, and the title “Average PM2.5 Levels by State (April 3, 2025)” anchors the context. The plot likely highlights a stark north-south divide, with northern bars visibly longer.
* **Figure**: [Placeholder: Insert “Figure\_4.2.png”.]

**4.3 Pollutant Correlations**

**i. General Description**

Pollutants often share emission sources—PM2.5 and NO2 from vehicles, SO2 from coal plants—leading to correlations that reveal underlying patterns. This analysis quantifies relationships between PM2.5, NO2, SO2, OZONE, and CO to determine which pollutants co-occur, aiding in source attribution. For example, a strong PM2.5-NO2 link could point to traffic, while OZONE-NO2 interactions reflect atmospheric chemistry, informing holistic pollution control strategies.

**ii. Specific Requirements**

* Aggregated average pollution values for PM2.5, NO2, SO2, OZONE, and CO across stations.
* Calculated Pearson correlation coefficients (ranging from -1 to 1) to measure linear relationships between pollutant pairs.
* Ensured correlations accounted for station-level variations to avoid aggregation bias.
* Visualized the correlation matrix to highlight strong positive or negative relationships.

**iii. Analysis Results**

[Placeholder: e.g., “PM2.5 and NO2 exhibited a strong positive correlation (r = X), suggesting common sources like vehicular exhaust and industrial emissions. PM2.5 and SO2 showed a moderate correlation (r = Y), indicating some overlap in industrial zones. OZONE had a negative correlation with NO2 (r = Z), consistent with ozone depletion in high-NO2 urban environments due to chemical reactions. CO correlated weakly with others (r < A), reflecting unique sources like biomass burning or traffic congestion. These patterns suggest targeted controls for traffic-related pollutants in urban areas.”]

**iv. Visualization**

* **Type**: Heatmap.
* **Description**: The heatmap presents a grid where rows and columns represent the five pollutants (PM2.5, NO2, SO2, OZONE, CO). Each cell’s color indicates the correlation coefficient, with red for positive values, blue for negative, and intensity reflecting strength (e.g., dark red for r ≈ 1). Numerical r values are annotated in each cell for precision. The color scale ranges from -1 to 1, centered at 0 (white or neutral) to distinguish no correlation. Titled “Correlation Between Air Pollutants (April 3, 2025),” the plot likely emphasizes PM2.5-NO2 as a bright red cell, with OZONE-NO2 in blue, guiding source-focused policies.
* **Figure**: [Placeholder: Insert “Figure\_4.3.png”.]

**4.4 Ozone Levels Across States**

**i. General Description**

Ozone, a secondary pollutant formed by sunlight-driven reactions of NO2 and volatile organic compounds, poses health risks even at low concentrations, irritating lungs and exacerbating asthma. Its levels depend on regional factors like solar intensity, traffic, and industrial emissions. This analysis explores state-wise ozone distributions to identify areas with elevated concentrations, informing measures like precursor emission controls or public health advisories during high-ozone days.

**ii. Specific Requirements**

* Selected records where the pollutant identifier was OZONE.
* Grouped data by state and computed the mean and standard deviation of average pollution to capture central tendency and variability.
* Analyzed the range of ozone levels within each state to assess consistency.
* Visualized the distribution across states to compare variability and outliers.

**iii. Analysis Results**

[Placeholder: e.g., “Rajasthan reported the highest average ozone concentration at X µg/m³, with a standard deviation of Y µg/m³, indicating wide fluctuations across stations. Gujarat followed with A µg/m³, while southern states like Kerala and Tamil Nadu averaged B and C µg/m³, respectively, with lower variability (std = D). High ozone in western states may reflect intense sunlight and industrial precursors, whereas southern stability suggests fewer emission sources. Some states showed outlier stations with peaks above E µg/m³, warranting localized investigation.”]

**iv. Visualization**

* **Type**: Violin plot.
* **Description**: The violin plot displays states along the x-axis, with each “violin” representing the distribution of ozone concentrations for that state. The y-axis shows ozone levels in µg/m³. Each violin’s width reflects the density of data points (wider where more stations report that level), and its height spans the range of values. A central line or box may indicate the median or quartiles. The title “Ozone Levels by State (April 3, 2025)” clarifies the scope, and rotated state labels prevent overlap. The plot likely shows fat violins for states like Rajasthan, indicating variability, and slimmer ones for southern states, suggesting consistency.
* **Figure**: [Placeholder: Insert “Figure\_4.4.png”.]

**4.5 Pollutant Variability in Major Cities**

**i. General Description**

Major cities, as economic and population hubs, face unique pollution challenges driven by traffic congestion, construction, and industrial zones. This analysis compares PM2.5 and NO2 levels across prominent cities (e.g., Delhi, Mumbai, Bengaluru, Chennai) to highlight variations linked to local factors, such as vehicle density or coastal geography. Understanding these differences can guide city-specific policies, like low-emission zones or public transit expansion.

**ii. Specific Requirements**

* Identified major cities based on population, economic significance, or dataset prominence (e.g., Delhi, Mumbai, Bengaluru, Chennai).
* Filtered for PM2.5 and NO2 records in these cities.
* Calculated average pollution metrics for each pollutant per city.
* Visualized the relationship between PM2.5 and NO2 to explore co-occurrence and city-specific trends.

**iii. Analysis Results**

[Placeholder: e.g., “Delhi led with PM2.5 averaging X µg/m³ and NO2 at Y µg/m³, reflecting intense traffic and industrial activity. Mumbai followed with A µg/m³ PM2.5 but lower NO2 (B µg/m³), possibly due to coastal winds dispersing pollutants. Bengaluru reported moderate levels (C µg/m³ PM2.5, D µg/m³ NO2), benefiting from green spaces, while Chennai’s values (E and F µg/m³) suggested balanced sources. PM2.5 and NO2 showed city-specific alignment, with Delhi’s high values indicating a strong traffic link.”]

**iv. Visualization**

* **Type**: Pair plot.
* **Description**: The pair plot is a grid with histograms on the diagonal and scatter plots off-diagonal. The histograms show the distribution of PM2.5 and NO2 across selected cities, revealing whether levels are clustered or spread. The scatter plot (PM2.5 vs. NO2) displays each station as a dot, with axes in µg/m³, highlighting correlations within cities. Titled “PM2.5 and NO2 Variability in Major Cities (April 3, 2025),” the plot likely shows Delhi’s points clustered at high values, Bengaluru’s at lower ones, and a positive trend line for most cities, reflecting shared sources.
* **Figure**: [Placeholder: Insert “Figure\_4.5.png”.]

**4.6 SO2 Levels by Region**

**i. General Description**

Sulfur dioxide, emitted primarily by coal-fired power plants, smelters, and heavy industries, contributes to acid rain and respiratory irritation. Its distribution varies by industrial intensity, making state-wise analysis critical for identifying emission hotspots. This objective examines SO2 levels across states to pinpoint regions needing stricter industrial regulations or cleaner energy transitions.

**ii. Specific Requirements**

* Filtered the dataset for SO2 records.
* Grouped by state and calculated the average pollution metric.
* Analyzed individual station measurements to capture within-state variability.
* Visualized SO2 levels to display state differences and outliers.

**iii. Analysis Results**

[Placeholder: e.g., “Chhattisgarh recorded the highest average SO2 at X µg/m³, driven by coal-based industries and power plants. Odisha followed with Y µg/m³, while coastal states like Tamil Nadu and Andhra Pradesh averaged Z and A µg/m³, reflecting minimal heavy industry. Within-state variability was high in industrial states, with some stations reporting peaks above B µg/m³, possibly near factories. Lower SO2 in southern and northeastern states suggests cleaner energy or geographic dispersion.”]

**iv. Visualization**

* **Type**: Strip plot.
* **Description**: The strip plot places states along the x-axis and SO2 concentrations (µg/m³) on the y-axis. Each station’s measurement is a dot, slightly offset (jittered) to avoid overlap, revealing the spread within states. The title “SO2 Levels by State (April 3, 2025)” sets the context, and rotated state labels ensure clarity. A horizontal grid aids value estimation. The plot likely shows dense clusters of high SO2 dots for states like Chhattisgarh and sparse, low-value dots for Tamil Nadu, highlighting industrial disparities.
* **Figure**: [Placeholder: Insert “Figure\_4.6.png”.]

**4.7 CO Levels in Urban vs. Rural Areas**

**i. General Description**

Carbon monoxide, a byproduct of incomplete combustion from vehicles, industries, and biomass burning, poses risks like dizziness and heart strain at high levels. Urban areas, with dense traffic and industry, may have higher CO than rural ones, where sources are fewer. This analysis compares CO concentrations by station type to test this hypothesis, informing urban planning and rural emission controls.

**ii. Specific Requirements**

* Classified stations as urban or rural, possibly using city population (e.g., >1 million for urban) or station name cues (e.g., “Industrial” vs. “Village”).
* Filtered for CO records.
* Calculated average pollution metrics for urban and rural groups.
* Visualized CO distributions to contrast the two categories.

**iii. Analysis Results**

[Placeholder: e.g., “Urban stations averaged X µg/m³ CO, significantly higher than rural stations at Y µg/m³. Cities like Delhi and Mumbai drove urban highs, with some stations exceeding Z µg/m³, linked to traffic congestion. Rural areas, such as stations in Arunachal Pradesh or Himachal Pradesh, showed lower variability (std = A), reflecting sparse sources like household burning. The urban-rural gap suggests vehicles and industry as primary CO contributors, with rural air relatively cleaner.”]

**iv. Visualization**

* **Type**: Swarm plot.
* **Description**: The swarm plot positions “Urban” and “Rural” categories on the x-axis, with CO concentrations (µg/m³) on the y-axis. Each station’s measurement is a dot, arranged to avoid overlap, showing the distribution within each group. A distinct color palette (e.g., green for rural, orange for urban) differentiates types. Titled “CO Levels in Urban vs. Rural Areas (April 3, 2025),” the plot likely shows a dense cloud of higher CO dots for urban areas and a sparser, lower cluster for rural ones, emphasizing the urban burden. A grid supports value comparison.
* **Figure**: [Placeholder: Insert “Figure\_4.7.png”.]
* **Note**: Assumes urban\_rural classification. If undefined, urban could include Delhi, Mumbai, etc., and rural could cover smaller towns like Kolar. Clarification needed for accuracy.

**4.8 Pollutant Trends by Latitude**

**i. General Description**

Latitude influences air quality through climate (e.g., sunlight for ozone), population density (urban north vs. rural south), and industrial patterns. This analysis explores whether pollutants like PM2.5 and OZONE vary systematically with latitude, testing ideas like higher PM2.5 in northern urban belts (25–30°N) versus cleaner southern regions (10–15°N). Such trends can guide geographically tailored policies.

**ii. Specific Requirements**

* Focused on PM2.5 and OZONE for their health and environmental significance.
* Extracted average pollution metrics and corresponding latitude values for each station.
* Examined visual patterns between pollution levels and latitude, without assuming linear relationships.
* Visualized the data to assess geographic trends.

**iii. Analysis Results**

[Placeholder: e.g., “PM2.5 concentrations peaked around 25–30°N, averaging X µg/m³, corresponding to urban centres like Delhi, Lucknow, and Patna. Southern latitudes (10–15°N, e.g., Kerala, Tamil Nadu) averaged Y µg/m³, indicating cleaner air. OZONE showed scattered highs (e.g., Z µg/m³ at 20°N in Gujarat), with no clear latitudinal gradient, possibly due to localized precursors. The northern PM2.5 bias suggests urban-industrial density, while OZONE’s variability reflects complex atmospheric factors.”]

**iv. Visualization**

* **Type**: Scatter plot.
* **Description**: The scatter plot maps latitude (°N) on the x-axis and average pollution (µg/m³) on the y-axis. Each station is a dot, with PM2.5 and OZONE distinguished by color (e.g., blue for PM2.5, red for OZONE). Transparency prevents overlap clutter, and a legend identifies pollutants. Titled “Pollutant Levels vs. Latitude (April 3, 2025),” the plot likely shows a cluster of high PM2.5 dots in the 25–30°N range and scattered OZONE dots across latitudes, hinting at regional rather than linear trends. A grid enhances readability.
* **Figure**: [Placeholder: Insert “Figure\_4.8.png”.]

**4.9 Pollutant Composition by State**

**i. General Description**

Each state’s air quality is shaped by its unique mix of pollutants, reflecting local sources—PM2.5 from traffic, SO2 from industry, OZONE from photochemistry. This analysis quantifies the contribution of PM2.5, NO2, SO2, OZONE, and CO to each state’s pollution profile, revealing dominant pollutants for targeted interventions, like NO2 controls in traffic-heavy states or SO2 limits in industrial ones.

**ii. Specific Requirements**

* Grouped data by state and pollutant identifier (PM2.5, NO2, SO2, OZONE, CO).
* Calculated the mean average pollution metric for each pollutant per state.
* Ensured all five pollutants were represented, handling any missing data appropriately.
* Visualized the proportion of each pollutant to compare state profiles.

**iii. Analysis Results**

[Placeholder: e.g., “Delhi’s pollution was dominated by PM2.5 (X µg/m³, A% of total), followed by NO2 (Y µg/m³, B%), reflecting urban traffic and dust. Gujarat showed elevated OZONE (Z µg/m³, C%), driven by sunlight and precursors, with lower PM2.5 (D%). Southern states like Karnataka had balanced profiles, with PM2.5, NO2, and OZONE each around E µg/m³, suggesting diverse but moderate sources. Industrial states like Chhattisgarh had high SO2 contributions (F%), highlighting coal reliance.”]

**iv. Visualization**

* **Type**: Stacked bar plot.
* **Description**: The stacked bar plot places states along the x-axis, with each bar’s height representing the total average pollution (µg/m³) from PM2.5, NO2, SO2, OZONE, and CO. Each bar is segmented, with a distinct color for each pollutant (e.g., blue for PM2.5, red for NO2), showing relative contributions. The y-axis measures concentration, and a legend identifies pollutants. Titled “Pollutant Composition by State (April 3, 2025),” the plot likely shows tall, PM2.5-heavy bars for Delhi and balanced, shorter bars for southern states. Rotated labels and a grid ensure clarity.
* **Figure**: [Placeholder: Insert “Figure\_4.9.png”.]

# ****5. CONCLUSION****

This Python-based analysis of India’s air quality, using the CLEANED.csv dataset, delivers a comprehensive view of pollution patterns as of April 3, 2025. The nine objectives illuminated critical insights:

* [Placeholder: e.g., “PM2.5 levels are alarmingly high in northern states like Delhi, averaging well above safe thresholds, while southern states enjoy cleaner air. Strong correlations between PM2.5 and NO2 point to traffic and industry as key culprits. Ozone varies widely, with western states like Rajasthan at risk, and urban areas bear a heavier CO burden than rural ones, underscoring urbanization’s toll.”]
* The visualizations—ranging from histograms and heatmaps to violin and swarm plots—transformed complex data into intuitive narratives, enabling stakeholders to grasp trends, disparities, and priorities at a glance.
* Preprocessing ensured data reliability, tackling missing values, outliers, and inconsistencies with methodical care, laying a solid foundation for analysis.

The project highlights Python’s prowess in environmental data science, blending technical rigor with real-world impact. It identifies urgent needs—stronger emission controls in northern cities, industrial regulation in SO2-heavy states, and ozone monitoring in variable regions—while offering a replicable framework for future studies. The findings, visualizations, and methodology are documented for sharing, available at [Placeholder: e.g., “github.com/your-repo” or “upon request”]. This work not only advances understanding of India’s air quality crisis but also empowers action toward cleaner air and healthier communities.

# ****6. FUTURE SCOPE****

The analysis, while thorough, opens numerous pathways for deeper exploration, leveraging additional data or advanced methods:

* **Temporal Analysis**: Incorporating multi-day, monthly, or seasonal data could reveal dynamic patterns, such as winter PM2.5 spikes in Delhi due to stubble burning or monsoon-related pollutant washout. Time-series techniques would quantify these cycles.
* **Health Impact Modelling**: Linking pollutant levels to health outcomes (e.g., hospital admissions for respiratory issues) could estimate human costs, using epidemiological data or regression models to inform public health policies.
* **Predictive Analytics**: Machine learning models, like random forests or neural networks, could forecast pollution levels based on weather, traffic, or industrial activity, aiding proactive measures like smog alerts.
* **Expanded Pollutant Scope**: Including emerging pollutants, such as volatile organic compounds (VOCs), black carbon, or ultrafine particles, would provide a fuller air quality picture, requiring new datasets.
* **Intra-City Granularity**: Analysing pollution at the neighbourhood level (e.g., Delhi’s Connaught Place vs. outskirts) could guide urban planning, like green corridors or traffic rerouting, using finer-grained monitoring data.
* **Policy Impact Simulation**: Modelling scenarios—e.g., 50% electric vehicle adoption or coal plant phase-outs—could estimate pollution reductions, employing atmospheric models to quantify benefits.
* **Public Engagement Tools**: Developing interactive dashboards with tools like Polly or Streamlet could share findings with citizens, fostering awareness and advocacy for cleaner air.
* **Global Contextualization**: Comparing India’s air quality to other nations (e.g., China, USA) could benchmark progress, requiring harmonized international datasets.

These extensions would amplify the project’s scope, bridging data science with policy, health, and community action, while building on the robust foundation established here.

# ****7.REFERENCES****

 [Placeholder: e.g., “Central Pollution Control Board, India. National Air Quality Monitoring Programme, 2025. Accessed via public data portal.”]

 Pandas Development Team. (2025). Pandas Documentation. Available at: <https://pandas.pydata.org/>

 Seaborn Development Team. (2025). Seaborn Documentation. Available at: <https://seaborn.pydata.org/>

 Matplotlib Development Team. (2025). Matplotlib Documentation. Available at: <https://matplotlib.org/>

 Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3), 90–95.

 Waskom, M. (2021). Seaborn: Statistical Data Visualization. *Journal of Open-Source Software*, 6(60), 3021.

GITHUB LINK: <https://github.com/MERCYFX/INT375-PROJECT>

