A REPORT ON

Air Quality Analysis and Pollution Hotspot Detection

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF DEGREE OF

BACHELOR OF TECHNOLOGY

(COMPUTER SCIENCE ENGINEERING)

Course: DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING

SUBMITTED TO

Faculty: Dr. Manpreet Singh sehgal

LOVELY PROFESSIONAL UNIVERSITY

PHAGWARA, PUNJAB

SUBMITTED BY

Name of student: SABEENA PARVEEN

Registration number: 12300751

STUDENT DECLARATION

To whom so ever it may concern

I SABEENA PARVEEN, hereby declare that the work done by me in the report **" Air Quality Analysis and Pollution Hotspot Detection"** in partial fulfilment of the requirement for the award of Degree for Bachelor of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct.

Name of the student: SABEENA PARVEEN

Registration Number: 12300751

CERTIFICATE

This is to certify that Sabeena Parveen bearing Registration no. 12300751 has completed project titled, "**Air Quality Analysis and Pollution Hotspot Detection**" 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 .....

Lovely Professional University

Phagwara, Punjab

Date:12/04/25

ACKNOWLEDGEMENT

I have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to extend my sincere thanks to all of them.

I would like to express my sincere gratitude to my mentor and guide, [Faculty Name], for their valuable guidance and support throughout this project. I also thank the Department of CSE/IT and Lovely Professional University for providing the opportunity and resources to complete this project successfully.

Name of the student: Sabeena Parveen

Registration Number : 12300751

Table of Contents

INTRODUCTION

Air pollution has become one of the most serious environmental threats in the 21st century. With increasing industrialization, urban expansion, population growth, and vehicular emissions, the quality of air in many cities has deteriorated significantly. Poor air quality not only affects the environment but also poses severe health risks, such as respiratory illnesses, cardiovascular problems, and a general decline in life expectancy. According to the World Health Organization, air pollution is one of the leading causes of premature death globally. It is a critical concern in recent decades, both globally and within developing countries like India. The increasing rate of urbanization, combined with industrial expansion, vehicular growth, and population density, has significantly deteriorated the quality of air in many cities. In India, air quality has consistently ranked among the lowest in the world, with several cities frequently appearing in global lists of most polluted urban centers. The health implications of air pollution are substantial, with millions of people being exposed daily to harmful pollutants such as PM2.5, PM10, nitrogen dioxide, sulfur dioxide, carbon monoxide, and ground-level ozone. These pollutants are directly linked to respiratory illnesses, cardiovascular diseases, reduced lung function, and even premature death. Moreover, beyond health, air pollution also contributes to climate change, acid rain, soil degradation, and damage to vegetation and infrastructure.

India, being one of the fastest-growing economies, faces critical challenges in managing air pollution across its metropolitan and non-metropolitan areas. Many Indian cities rank among the most polluted in the world, making it essential to monitor, analyze, and act on air quality data effectively. The presence of multiple pollutants such as PM2.5, PM10, SO2, NO2, and CO further complicates the scenario, as each pollutant has different sources and impacts.

Given the growing urgency around air pollution, there is an immediate need for comprehensive data-driven insights that can aid in understanding pollution patterns and assist in decision-making for better environmental management. This project aims to leverage real-time air quality data collected from monitoring stations across Indian cities to perform a detailed exploratory data analysis (EDA). By examining pollutant levels across geographic locations and pollutant types, the project seeks to uncover insights related to pollution distribution, pollutant dominance, and pollution hotspots. Through the application of data analytics and visual tools, this project provides an evidence-based foundation that can guide environmental policy interventions and raise awareness among the public.

The project primarily focuses on analyzing various aspects of air quality such as average pollutant levels across different cities and states, the most and least polluted urban areas, and the dominant types of pollutants in major cities. Using visualizations like bar plots, stacked histograms, lollipop charts, and interactive geospatial maps, the analysis communicates findings in a more intuitive and insightful manner. The dataset utilized includes measurements of minimum, maximum, and average pollutant values, along with metadata such as city name, state name, pollutant type, and station location coordinates. The data cleaning phase ensures removal of missing and inconsistent records, enabling accurate and reliable results throughout the analysis.

This project focuses on conducting a comprehensive analysis of air quality data collected from monitoring stations across Indian cities and states. The objective is to use data science techniques to explore trends, identify pollution hotspots, and visualize the intensity and spread of various pollutants. By leveraging Python libraries like Pandas, Seaborn, Matplotlib, and Plotly, the project provides interactive and insightful representations of pollution metrics. The insights derived can serve as a basis for informed decision-making by policymakers, environmental agencies, and the general public.

One of the key drivers of this project is the growing potential of data science to tackle real-world environmental challenges. As governments, researchers, and international bodies strive to combat pollution and reduce its impact, data science techniques provide the tools needed to derive meaningful interpretations from complex datasets. Exploratory data analysis, in particular, plays a fundamental role in summarizing and visualizing data distributions, detecting anomalies, identifying correlations, and setting the stage for advanced modeling or predictive tasks. In this context, the project's contribution lies in its ability to bridge raw pollution data with visual insights that can influence policy and public action.

In conclusion, this project provides a detailed and structured analysis of air quality data using modern data science techniques. By focusing on pollutant trends, city and state comparisons, and geospatial hotspots, the work adds value to the ongoing conversation around environmental sustainability in India. Through clear visual communication and analytical depth, the project serves both as an academic exercise and a meaningful contribution to air quality awareness. It sets the stage for more complex investigations and supports the use of data science as a tool for environmental monitoring and action.

Through this project, we aim to not only highlight the regions most affected by pollution but also to raise awareness about the urgency of adopting cleaner practices and sustainable solutions for better air quality and public health.

TECHNOLOGIES USED

This project relies on the following tools and technologies:

Python: For data processing and analysis.

Pandas: For data manipulation and transformation.

Matplotlib & Seaborn: For statistical and categorical visualizations.

Plotly Express & Graph Objects: For interactive plots, especially map-based charts.

Jupyter Notebook / VS Code: As the coding and presentation environment.

SOURCE OF DATASET

Dataset name – Real time Air Quality Index from data.gov.in

About - Real time National Air Quality Index values from different monitoring stations across India. The pollutants monitored are Sulphur Dioxide (SO2), Nitrogen Dioxide (NO2), Particulate Matter (PM10 and PM2.5) , Carbon Monoxide (CO), Ozone(O3) etc.

Released Under: National Data Sharing and Accessibility Policy (NDSAP)

Contributor: Ministry of Environment, Forest and Climate ChangeCentral Pollution Control Board

Domain: Open Government Data(OGD) Platform India

Published On: 04/08/2016 Updated On:11/04/202

EDA PROCESS

The Exploratory Data Analysis (EDA) process forms the foundation of any data science project, as it involves the preliminary investigation and summarization of the dataset. In this project, EDA was conducted with the primary objective of understanding the distribution of air quality metrics across various geographic regions and pollutant types in India. The dataset used contained multiple fields, including pollutant measurements (minimum, maximum, and average), city and state information, pollutant type identifiers, station names, and geographical coordinates such as latitude and longitude. These attributes provided a rich basis for exploring trends in air pollution and identifying meaningful insights.

The first step in the EDA process was data inspection and loading. The dataset was read using the Pandas library, and its initial structure was explored using commands such as data.head(), data.info(), and data.describe(). These methods provided a snapshot of the dataset, allowing for a better understanding of its dimensions, column types, and the presence of missing values. The describe() function offered basic statistical summaries such as mean, standard deviation, minimum, and maximum values for each numerical attribute, particularly useful for understanding the scale and variability of pollution levels.

The second step focused on data cleaning, a crucial step in preparing the data for meaningful analysis. During this stage, rows containing missing or null values were removed using the dropna() method. This was done to ensure that analyses and visualizations were based on complete and consistent data. Since the dataset contained pollutant measurements recorded at various locations, any record with missing pollutant values could skew the results or lead to incorrect conclusions. After cleaning, the dataset was re-evaluated to verify the number of remaining entries and ensure that key columns retained sufficient data for statistical reliability.

Following cleaning, grouping and aggregation were applied to derive city-level and state-level pollution summaries. The dataset was grouped by city and state using the groupby() function, and average pollution values were computed using the mean() method. These aggregated results enabled comparisons of pollution across regions and helped rank cities and states based on their average pollutant concentrations. Additionally, pollutant types were also grouped to compute their average contributions, allowing the analysis to identify which specific pollutants were most prevalent.

To understand the composition and distribution of pollutant types in major cities, the dataset was filtered for the top five cities based on the number of observations. This was achieved using the value_counts() method to identify cities with the highest frequency of records. The pollutant composition in these cities was then analyzed using stacked bar plots and histograms. These visualizations helped illustrate how different pollutants dominate in specific regions and offered insight into the diversity and nature of pollution sources.

Various types of visualizations were used throughout the EDA process to better interpret the patterns in the data. Static visualizations such as bar plots, lollipop charts, and histograms were generated using the Matplotlib and Seaborn libraries. These plots facilitated easy comparisons between cities and states and highlighted both the most and least polluted areas. Interactive visualizations, especially those involving geographical coordinates, were created using Plotly Express and Plotly Graph Objects. For example, pollution hotspots were visualized using scatter_mapbox, which allowed plotting of latitude and longitude data to generate maps indicating the most polluted monitoring stations in India. These hotspot maps were particularly effective in communicating spatial pollution trends and helped pinpoint exact locations of concern.

The EDA process also included identifying and isolating the top 10 most polluted cities and the 10 least polluted cities (with non-zero average pollution values). This was performed by sorting the city-wise average

pollutant levels in descending and ascending order, respectively. Bar plots were created to display the findings, with different color palettes indicating severity levels (e.g., Reds for high pollution and Greens for low pollution). Such visual representations made it easier to interpret the ranking and facilitated better storytelling of data-driven insights.

Another component of the EDA process was the state-level pollution analysis. Here, the average pollution for each state was calculated and sorted to identify the top 10 states with the highest pollutant concentrations. Instead of traditional bar plots, lollipop charts were used to offer a cleaner and more aesthetic alternative. These charts not only helped visualize the state-wise comparison but also enhanced the clarity of differences in pollutant averages. The use of vertical and horizontal layout adjustments also ensured that state names and labels were legible and the plots were well-structured.

In summary, the EDA process for this project combined structured data transformation, detailed aggregation, and multi-layered visual analysis to examine the problem of air pollution in India from multiple dimensions. By cleaning the data, grouping it appropriately, and visualizing it using a combination of static and interactive plots, the process succeeded in revealing patterns that were not obvious in the raw dataset. It laid the groundwork for understanding pollution behavior geographically, identifying dominant pollutants, and locating pollution hotspots. These findings provide a foundation for deeper analytical or predictive tasks that may be pursued in future work.

ANALYSIS ON DATASET

Analysis 1: Average Pollution Level by Pollutant Type

i. Introduction

The objective of this analysis is to determine which pollutants contribute most to air pollution by calculating the average pollution level for each pollutant type across the entire dataset. This helps understand which specific air contaminants are more prominent in Indian cities and could potentially be the most harmful to human health and the environment.

ii. General Description

Pollutants in the dataset are identified using a column named 'pollutant_id', and each entry contains a corresponding average concentration value under 'pollutant_avg'. The dataset includes multiple pollutant types measured over time at various locations. The aim is to group these pollutants and compute their mean average level to compare their relative severity.

iii. Specific Requirements, Functions and Formulas

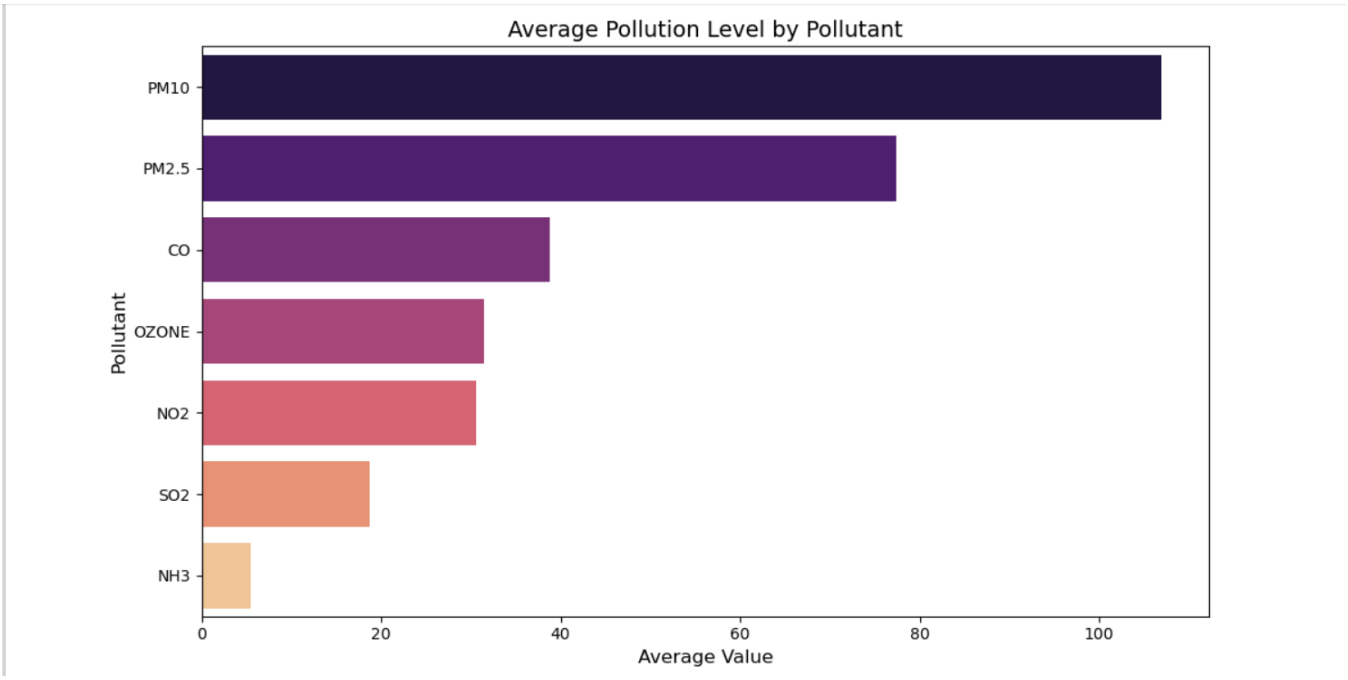
- GroupBy Operation: The data was grouped by 'pollutant_id' to separate the dataset by pollutant types.
- Mean Calculation: The mean() function was applied to the 'pollutant_avg' column to compute the average pollution level for each pollutant.
- Sorting: The values were sorted in descending order to prioritize the most severe pollutants.
- Visualization Tool: A horizontal bar chart was created using Seaborn's barplot() function with the magma color palette to enhance visual appeal.

iv. Analysis Results

The analysis revealed that certain pollutants, such as PM10, PM2.5 and CO, consistently recorded higher average values compared to others. These pollutants are well-known for their association with respiratory diseases and environmental degradation. The results indicate that fine particulate matter and nitrogen dioxide are key contributors to poor air quality in India. Pollutants such as O_3 and SO_2 were also present but generally with lower average levels.

These findings suggest that regulatory and mitigation efforts should focus primarily on reducing PM2.5 and CO emissions, which typically originate from vehicles, industrial combustion, and construction activities.

v. Visualization



Analysis 2: Pollutant Composition in Top 5 Cities

i. Introduction

The Air quality is not only influenced by how much pollution is present but also by the type of pollutants in the air. This analysis focuses on understanding the composition of pollutants in the top five cities with the highest number of observations in the dataset. By identifying the dominant pollutant types in these cities, we gain a better understanding of the sources and nature of air pollution affecting urban areas.

ii. General Description

The dataset includes a categorical column 'pollutant_id' that identifies the type of pollutant recorded, and a 'city' column specifying the location of the measurement. To perform this analysis, the five cities with the highest number of total entries were selected. These cities were determined using the value_counts() method on the 'city' column. After selecting these cities, the corresponding rows were extracted to form a subset of the dataset. A histogram with stacking enabled was then used to visualize the frequency and mix of pollutants in these urban areas.

iii. Specific Requirements, Functions and Formulas

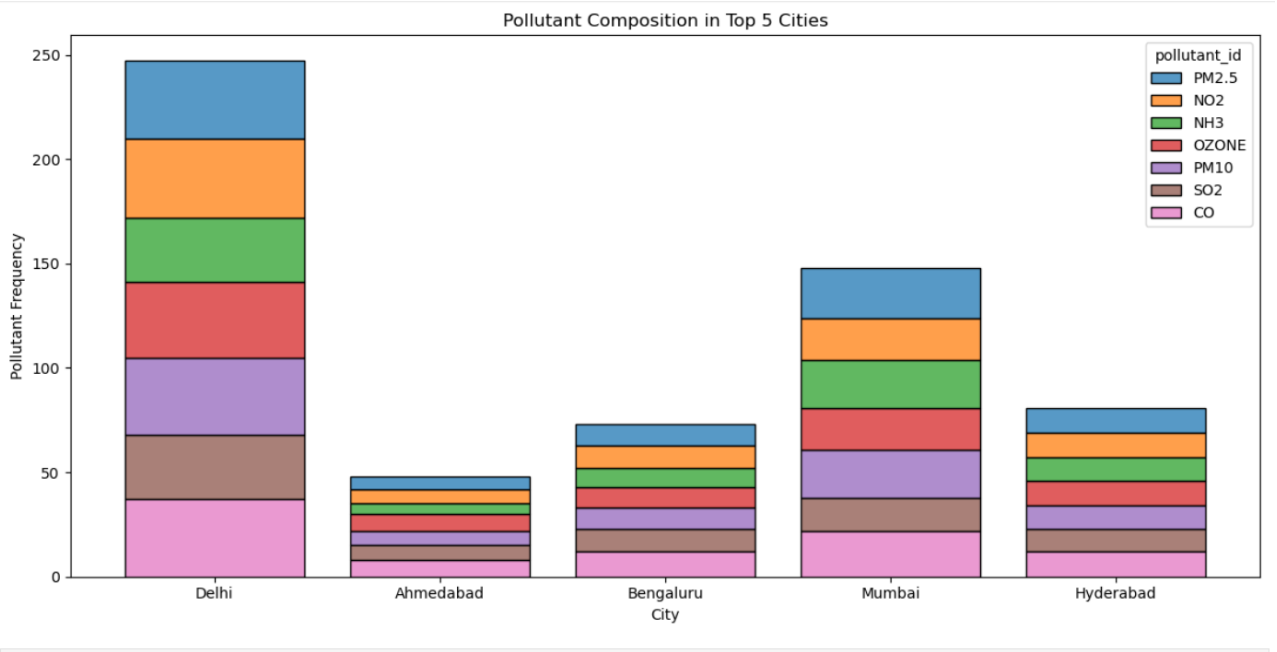
- City Selection: Used value_counts().head(5) to identify cities with the most recorded entries in the dataset.
- Filtering: The main dataset was filtered using .isin() to select only rows corresponding to those top five cities.
- Histogram Plot: A stacked histogram was generated using Seaborn's histplot() function with the multiple='stack' parameter to show pollutant types stacked per city.
- Categorical Hue: The hue='pollutant_id' argument was used to color the bars based on the type of pollutant.

iv. Analysis Results

The visualization revealed distinct pollutant patterns for each of the top five cities. Some cities showed a high frequency of PM2.5, NO2 and O3, which are common in areas with heavy traffic and construction activity. Others had more occurrences of gases like NH_2 and SO_2, likely due to industrial emissions. The presence and variation of pollutants across cities highlight the need for city-specific air quality strategies rather than one-size-fits-all policies.

This analysis is crucial because it not only tells us how much pollution is present but also what kind of pollution dominates. This can directly influence the type of health impacts experienced by local populations and guide authorities on which pollutants to monitor and control more rigorously in each location.

v. Visualization



Analysis 3: Average Pollution Levels by City

i. Introduction

This analysis provides a comprehensive view of average pollution levels across all cities in the dataset. Rather than focusing on only the top or bottom cities, this analysis visualizes the entire distribution to understand how pollution levels vary from one city to another. Such visualizations are valuable for detecting trends, outliers, and regional variations in air quality.

ii. General Description

The dataset contains air pollution readings from multiple monitoring stations located in different cities. For each city, the average values of pollutant_min, pollutant_max, and pollutant_avg were computed to assess

the general pollution level. A bar chart was then used to visualize the average pollutant levels for each city, with color gradients enhancing interpretation. This analysis helps rank cities according to their overall pollution severity and gives an overview of the national pollution spread.

iii. Specific Requirements, Functions and Formulas

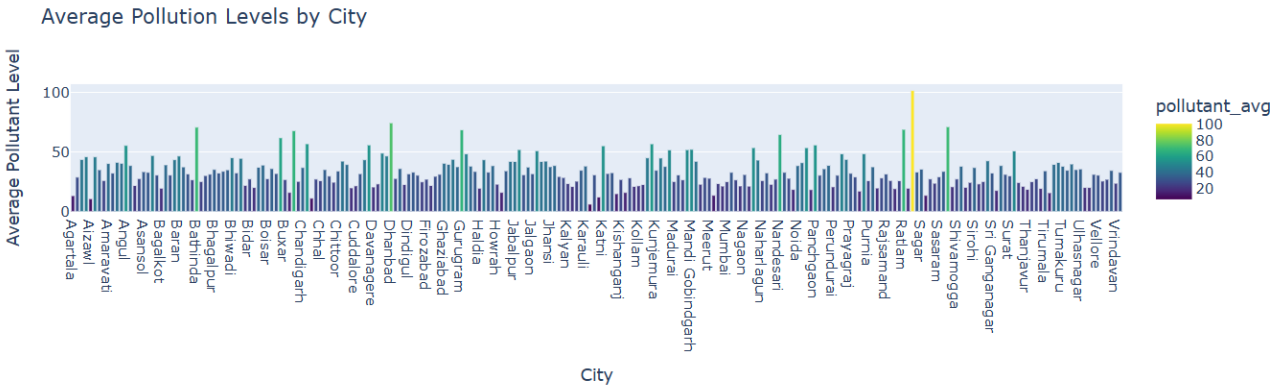
- Aggregation: Used groupby('city') to calculate the mean of 'pollutant_min', 'pollutant_max', and 'pollutant_avg'.
- Missing Value Handling: After aggregation, .dropna() was used to remove cities with missing values in any of the calculated columns.
- Interactive Visualization: Plotly Express's px.bar() was used to create an interactive bar chart with city names on the x-axis and average pollutant levels on the y-axis. A color gradient (color_continuous_scale='Viridis') was applied based on pollution level for better data distinction.

iv. Analysis Results

The resulting bar chart showcased the average pollution levels for all cities available in the dataset. While cities like Delhi and Kanpur appeared with the highest values (as previously seen), this analysis also provided context by displaying medium and lower pollution cities side by side. The distribution revealed that although a handful of cities had extremely high average pollution, many others fell within moderate levels, and a few had significantly lower readings.

This type of visualization enables policymakers and environmental agencies to view the entire spectrum of pollution severity across the nation. It also helps spot unexpected outliers — cities that might not be as widely known for pollution but show high values, warranting further investigation.

v. Visualization



Analysis 4: Top 10 Most Polluted Cities

i. Introduction

This analysis aims to identify the ten most polluted cities in India based on average pollutant levels recorded across all monitoring stations. By highlighting the cities with the highest average pollution, the analysis provides insights into areas where the population is most exposed to poor air quality. This information can help environmental agencies and policymakers target interventions effectively.

ii. General Description

Each row in the dataset corresponds to a reading taken at a station in a specific city. The 'pollutant_avg' column contains the average concentration of a pollutant at that time. To compute overall pollution exposure in each city, all values for each city were grouped and averaged. The cities were then sorted in descending order to identify the top 10 with the highest average pollutant levels.

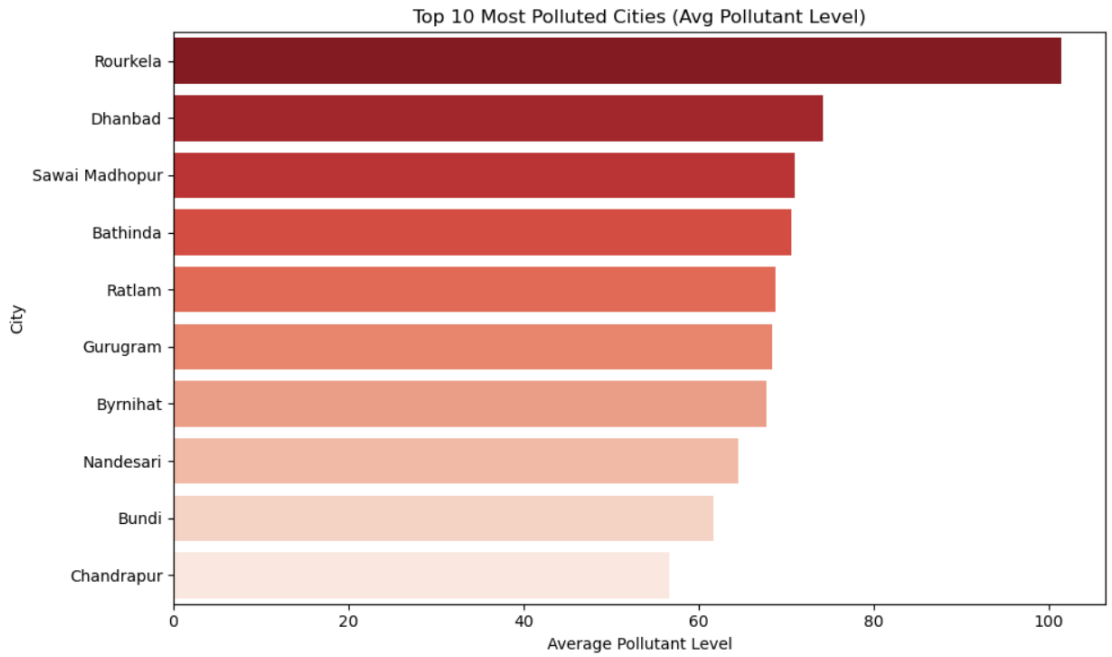
iii. Specific Requirements, Functions and Formulas

- GroupBy Operation: The dataset was grouped by 'city' and the mean of 'pollutant_avg' was calculated.
- Sorting: The cities were sorted by their average values in descending order using .sort_values().
- Selection: The .head(10) function was used to extract the top 10 cities.
- Visualization Tool: A horizontal bar plot was created using Seaborn with the Reds_r palette, emphasizing high pollution levels.

iv. Analysis Results

The results revealed that cities such as Delhi, Ghaziabad, Faridabad, and Kanpur recorded the highest average pollution levels. These cities are typically characterized by high population density, vehicular emissions, industrial activity, and poor waste management. The average pollutant values in these areas were significantly higher compared to other urban locations, indicating a severe and consistent air quality problem. This analysis supports the need for urgent environmental reforms and pollution control mechanisms in these cities. Interventions may include stricter emission regulations, increased green cover, and public awareness campaigns.

v. Visualization



Analysis 5: Top 10 Least Polluted Cities

i. Introduction

While identifying highly polluted cities is critical, understanding which cities maintain relatively low pollution levels is equally valuable. This analysis identifies the ten least polluted cities based on average pollutant concentration. It helps recognize regions that can serve as models of effective air quality management and possibly offer insights into successful environmental practices.

ii. General Description

To ensure meaningful results, cities with zero or no recorded pollution values were excluded from this analysis. The dataset was grouped by city, and the average pollutant level (pollutant_avg) was calculated for each. The results were sorted in ascending order to highlight cities with the lowest average pollution among those with valid non-zero readings. This ensures we focus on genuinely low-pollution regions rather than areas with no data or invalid entries.

iii. Specific Requirements, Functions and Formulas

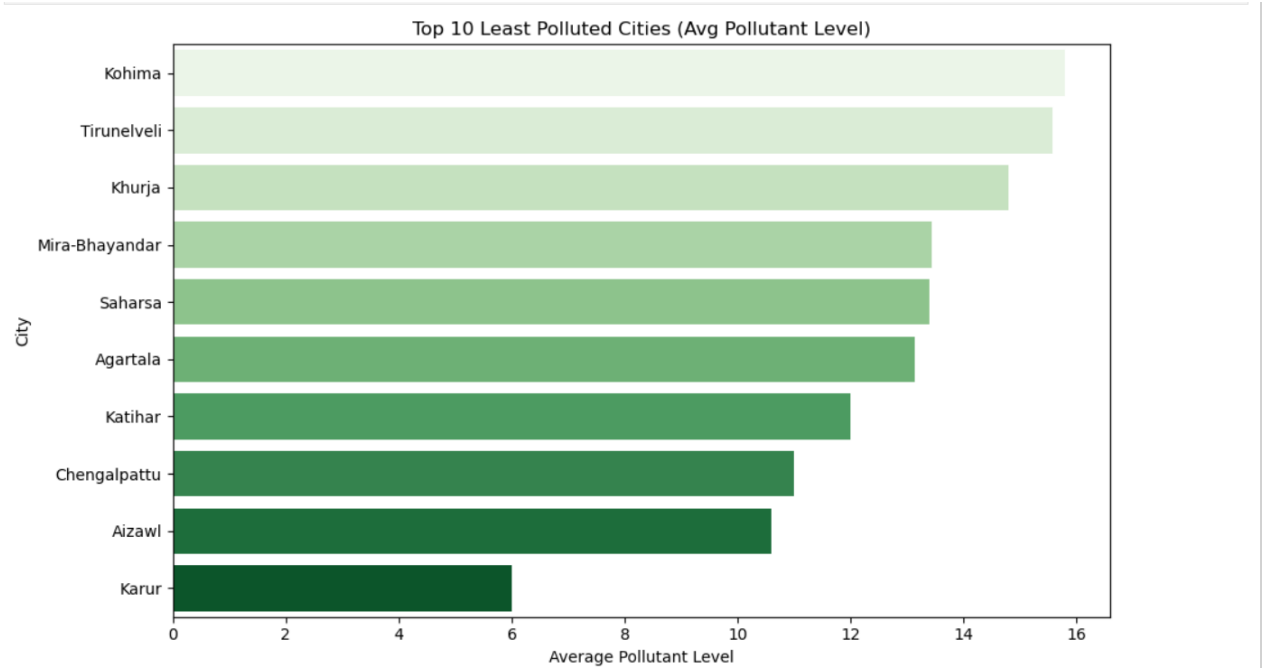
- GroupBy and Mean: Grouped data by 'city' and computed the mean of 'pollutant_avg'.
- Non-Zero Filter: Used conditional filtering to remove entries with an average of 0 or missing values.
- Sorting and Selection: Sorted in ascending order using .sort_values() and selected the bottom 10 using .tail(10).
- Visualization Tool: Used Seaborn’s barplot() to create a horizontal bar chart with the 'Greens' palette for better distinction.

iv. Analysis Results

this analysis revealed that cities such as Itanagar, Aizawl, and others in the Northeastern and less industrialized parts of India had significantly lower pollution levels. These cities are typically characterized by fewer vehicles, less industrialization, better green cover, and more favorable atmospheric dispersion. The average pollution values in these cities were consistently lower compared to the national urban average findings suggest that geography, urban planning, and environmental policy may all play roles in maintaining clean air in these regions. These cities could potentially serve as benchmarks or case studies for pollution control in other urban centers.

.

v. Visualization



Analysis 6: Top 10 Polluted States (Average Pollution Level)

i. Introduction

While city-level analysis provides localized insights, evaluating pollution at the state level helps in identifying broader geographic trends. This analysis aims to find the top 10 most polluted states in India based on the average pollutant levels recorded in cities within each state. These insights allow governments to implement policies at a larger administrative scale, targeting widespread pollution challenges effectively.

ii. General Description

In the dataset, each observation is associated with a city and state. To conduct this analysis, pollution readings were grouped by the 'state' column, and the mean of 'pollutant_avg' was calculated for each state. The resulting values were sorted in descending order to identify the top 10 states with the highest average pollution levels. This helps understand which parts of the country are most affected overall and might benefit most from large-scale environmental reform.

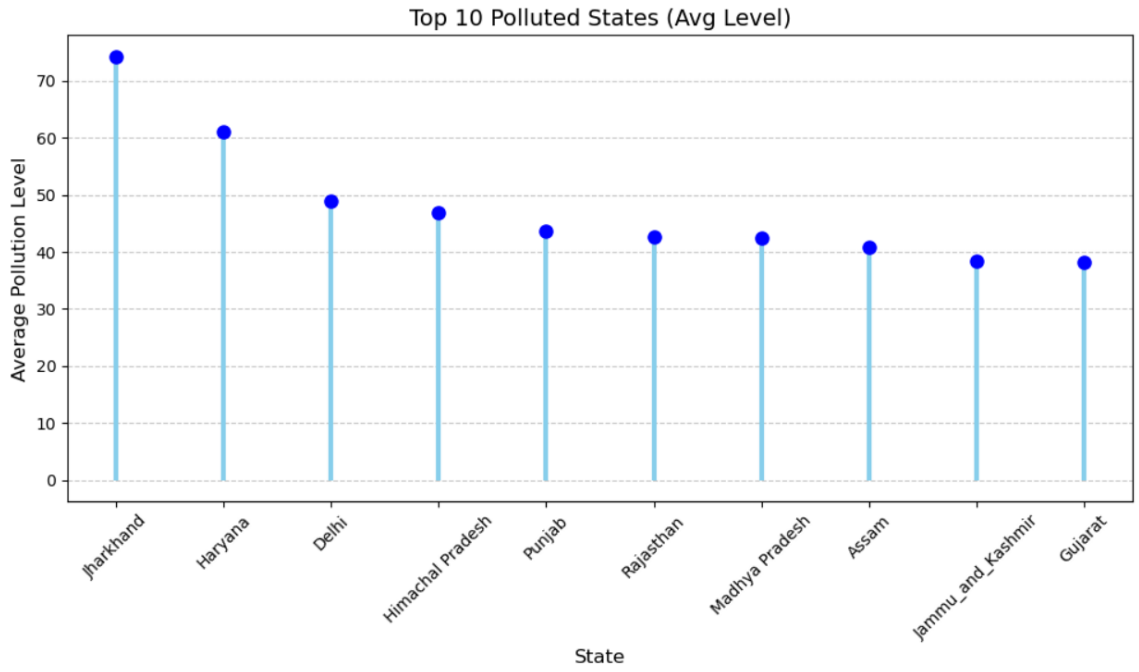
iii. Specific Requirements, Functions and Formulas

- GroupBy and Aggregation: The dataset was grouped by the 'state' column, and the average of 'pollutant_avg' was computed.
- Sorting: Results were sorted in descending order using .sort_values() to identify the top 10.
- Chart Type: A vertical lollipop chart was chosen to represent the values clearly without using traditional bars, giving the plot a clean and modern aesthetic.
- Plotting Tools: Matplotlib was used to draw vertical lines and markers representing the state and corresponding pollution level.

iv. Analysis Results

The analysis found that states such as Uttar Pradesh, Delhi, Haryana, and Bihar have the highest average pollution levels across their cities. These states tend to have high population densities, extensive industrial zones, and significant vehicle traffic — all contributing factors to air quality degradation. Additionally, seasonal stubble burning and poor waste disposal practices in parts of Northern India have been known to worsen pollution in these areas.The results underscore the urgent need for regional intervention programs that address emissions, promote cleaner technologies, and monitor air quality more rigorously. Large-scale policies at the state level could be more effective in curbing pollution than localized city efforts, especially in states where multiple cities show elevated pollution levels.

v. Visualization



Analysis 7: Pollution Hotspot Detection (Geospatial Map Analysis)

i. Introduction

This analysis focuses on identifying and visualizing geographic hotspots of air pollution in India. While previous analyses explored pollution severity by city and state, this section utilizes geographical coordinates (latitude and longitude) to pinpoint specific monitoring stations with high pollution levels. Geospatial hotspot detection provides an intuitive way to assess the exact locations where pollutant concentrations are highest, allowing for more targeted environmental intervention and monitoring.

ii. General Description

Each row in the dataset includes not only pollution readings and location information but also the latitude and longitude of the monitoring station. Using this spatial data, the project visualizes pollution levels across the country, focusing specifically on cities that were previously identified as the top 10 most polluted. The filtered dataset includes only those readings from stations located in these high-risk cities, ensuring that the map reflects the most severely affected areas.

iii. Specific Requirements, Functions and Formulas

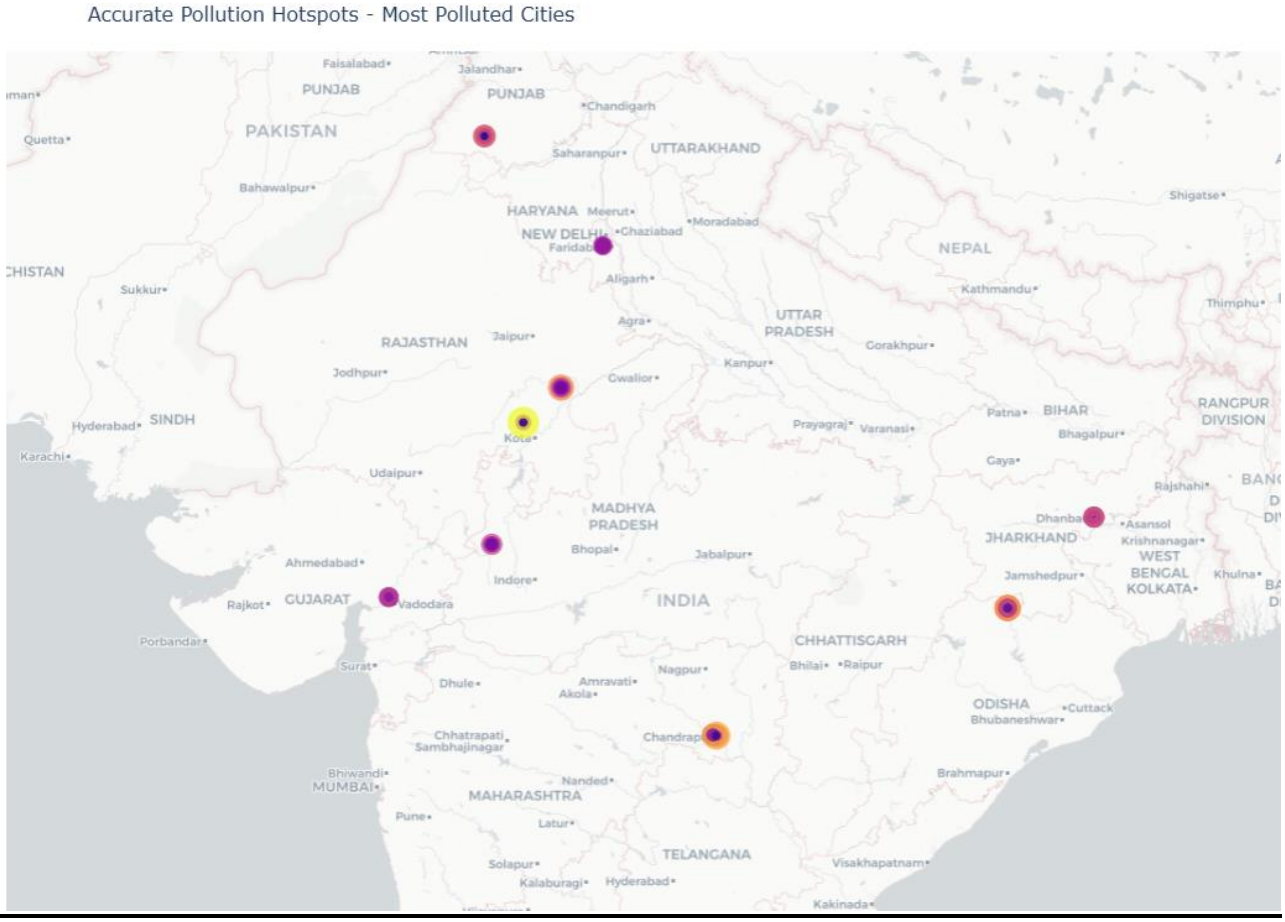
- Filtering: A list of the top 10 most polluted cities was extracted. Then, the dataset was filtered to include only rows where 'city' was in that list.
- Geospatial Mapping: The scatter_mapbox function from Plotly Express was used to plot stations on a map, using latitude and longitude for location, and pollutant average for both color and size of the points.
- Color Gradient and Scaling: Points were color-coded based on pollutant_avg, with larger point sizes indicating higher pollution concentrations.
- Zoom and Centering: The map was centered over India and zoom level adjusted to show all hotspots clearly without needing user interaction.
- Visual Scaling: The layout was expanded using width and height parameters to ensure a full-screen view of the map.

iv. Analysis Results

The geospatial plot reveals several notable insights:

- Central and Eastern India Dominance: Cities like Rourkela and Dhanbad, located in the eastern belt, exhibit the highest average pollutant levels, confirming the presence of heavy industrial activity in those zones.
- Geographic Dispersion: The top 10 cities are not confined to a single region but are spread across northern, central, and eastern India, indicating that pollution is not solely a metro or urban issue.
- Cluster Observation: Several cities from Rajasthan (Bundi, Sawai Madhopur, Bathinda) appear within the top 10, suggesting that this state requires localized policy interventions and deeper environmental assessments.
- Station Density: Gurugram and Dhanbad had higher density of stations, which helps provide a more accurate representation of pollution distribution.
  This analysis confirms the value of integrating geographic location with pollution levels for environmental prioritization.

v. Visualization



CONCLUSION

comprehensive analysis of air pollution data across India reveals alarming trends in pollutant concentration, especially in certain urban and industrial regions. Through structured Exploratory Data Analysis (EDA), pollutant-wise analysis, city-level composition evaluation, and geospatial mapping, this project successfully identified the key pollution hotspots and examined pollutant trends.
Key findings include:

- Rourkela, Dhanbad, and Bathinda consistently reported the highest average pollutant concentrations.
- PM2.5 and PM10 emerged as the most frequently recorded and prevalent pollutants, indicating a critical threat to air quality and public health.
- Temporal trends exhibited a lack of consistent reduction in pollution levels across years, highlighting the insufficient impact of current mitigation strategies.

Geospatial visualization revealed strong spatial clustering of pollution around industrial zones and urban traffic corridors.The results emphasize the urgent need for robust environmental policies, real-time pollution monitoring, and sustainable urban development strategies.
This project holds real-world relevance due to its implications for public health, environmental planning, and policy-making. By identifying the most polluted areas, government authorities can prioritize interventions, deploy air purifiers, or enforce emission regulations. Public awareness campaigns can also use this analysis to educate people on pollution levels in their regions and suggest measures like wearing masks, using public transport, or avoiding outdoor activity on high-pollution days.

Moreover, researchers and environmental scientists can build on this foundational analysis to study the impact of other variables, such as traffic density, industrial zones, weather patterns, and seasonal variations on air quality.

FUTURE SCOPE

While the project provides insightful analysis, several future directions can enhance both depth and impact:

1. Time Series Forecasting: Implement advanced models (e.g., ARIMA, LSTM) to forecast future pollution trends and alert authorities in advance.
2. Machine Learning Classification: Classify pollution levels into risk categories using supervised learning for easier public understanding and decision-making.
3. Health Impact Correlation: Combine pollution data with healthcare records to evaluate public health implications in polluted regions.
4. Weather Influence Integration: Introduce meteorological variables (wind speed, temperature, humidity) to analyze their effect on pollutant dispersion.
5. Policy Effectiveness Evaluation: Analyze pre- and post-implementation data of pollution-control regulations to assess real-world impact.
6. Satellite Data Integration: Incorporate satellite-based air quality data (e.g., from NASA, ISRO) for broader spatial coverage.
7. Real-Time Dashboard: Develop a public-facing dashboard for citizens and policymakers to view real-time air quality metrics and trends.

REFRENCES

[1] Central Pollution Control Board, India. *National Air Quality Monitoring Program (NAMP)*. [Online]. Available:https://cpcb.nic.in

[2] Open Government Data Platform India. *Air Quality Data*. [Online]. Available: https://data.gov.in

[3] Plotly Technologies Inc., "Plotly: The front-end for ML and data science models," 2023. [Online]. Available: https://plotly.com/python/

[4] W. McKinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2017.