**DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING**

**PROJECT REPORT**

(Project Semester January-April 2025)

***Air Quality Analysis & Statistical Evaluation of Pollution Trends Across Indian Cities***

Submitted by

**Harsh Porwal**

Registration No. 12311999

Programme and Section : KM007

Course Code : INT375

Under the Guidance of

**Dr. Mrinalini Rana (UID – 22138)**

**Discipline of CSE/IT**

**Lovely School of Computer Science and Engineering**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that **Harsh Porwal** bearing Registration no. **12311999** has completed . INT375 project titled, **“*Air Quality Analysis & Statistical Evaluation of Pollution Trends Across Indian Cities*”** 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 and Engineering**

Lovely Professional University

Phagwara, Punjab.

Date: 11-04-2025

**DECLARATION**

I, Harsh Porwal, student of Computer Science and Engineering 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: 11-02-2025 Signature

Registration No. 12311999 Name of the student Harsh Porwal

1. **Introduction**

This project, "Air Quality Analysis & Statistical Evaluation of Pollution Trends Across Indian Cities," aims to explore and understand the levels and trends of major air pollutants—SO₂ (Sulphur Dioxide), NO₂ (Nitrogen Dioxide), PM10 (Particulate Matter that is 10 micrometers or less in diameter), and SPM (Particulate Matter)—across various cities in India. By leveraging real-world environmental data, the project applies statistical methods and visual analytics to identify pollution patterns, inter-pollutant relationships, and potential sources. The insights gained can support evidence-based environmental policies and promote awareness of urban air quality issues

1. **Source of Dataset**

Site name: Data.gov.in

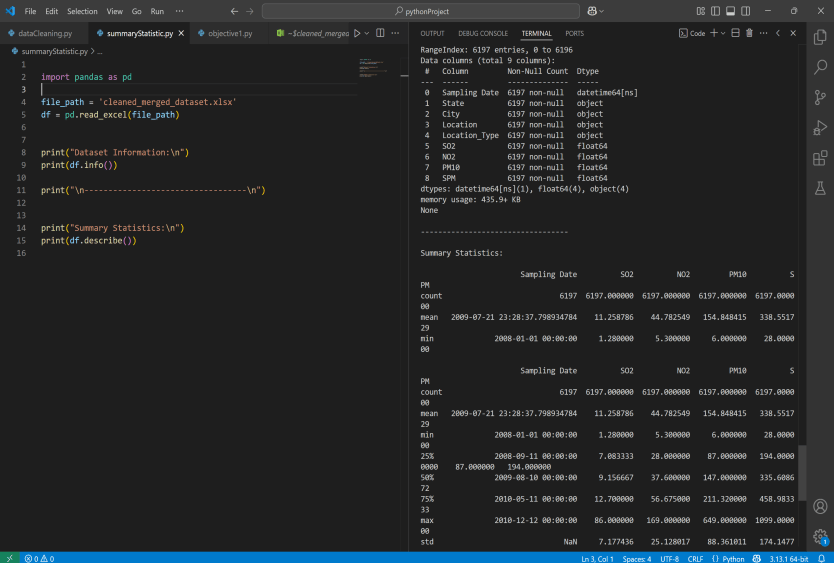
Link to site: <https://www.data.gov.in/>

Link to dataset: <https://www.data.gov.in/catalog/historical-daily-ambient-air-quality-data>

1. **EDA (Exploratory Data Analysis) process**

The EDA process began with loading and inspecting the dataset, which included pollution readings from various Indian cities. Key pollutants analyzed were **SO₂, NO₂, PM10, and SPM**. Initial steps involved cleaning and organizing the data, followed by statistical summarization to understand overall trends.

Screenshots of EDA analysis on next page.



1. **Analysis on dataset**

There are 5 analysis which we have done those are:

1. **State-wise & City-wise Pollution Analysis:**

Introduction: Analyze the average levels of SO2, NO2, PM10, and SPM across different states and cities to identify the most and least polluted areas.

Description: The dataset contains pollution readings for various Indian cities, including concentrations of key pollutants: SO₂, NO₂, PM10, and SPM. Each entry includes the state, city, pollution levels, and sampling date. The goal is to evaluate how pollution levels vary geographically across states and cities.

Specific Requirements, Functions, and Formulas:

Grouping & Aggregation: Data was grouped by State and City to calculate:

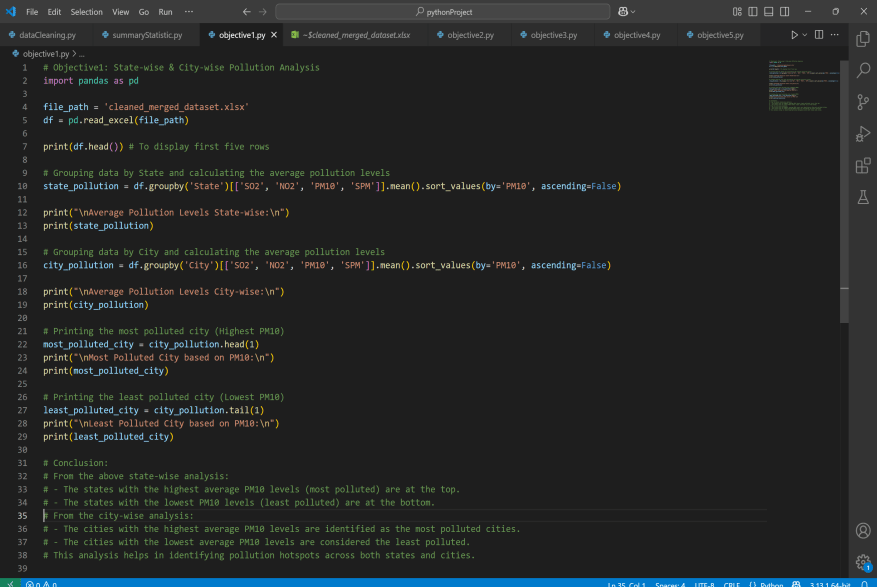
Average levels of each pollutant

Maximum and minimum pollutant levels

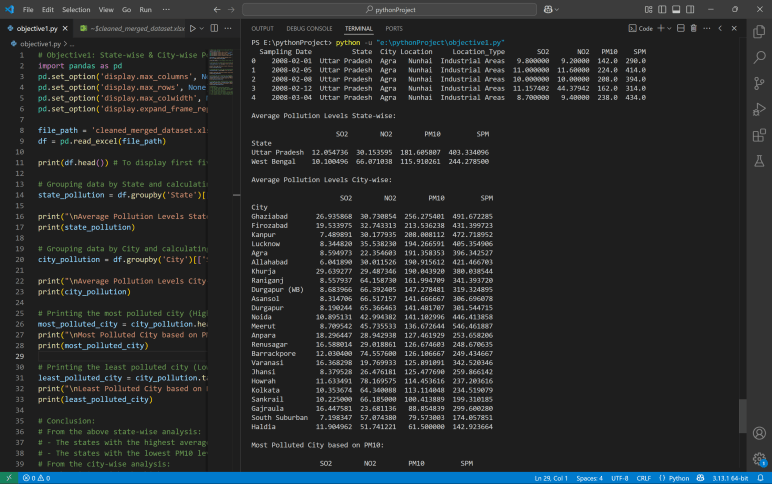
Pandas functions used: groupby(), mean(), sort\_values()

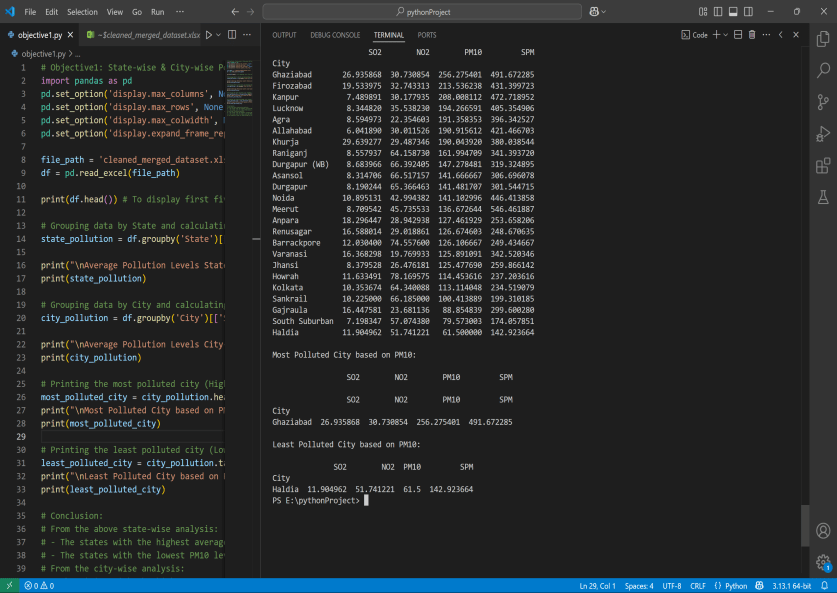
Implementation:

Code:



Output:





Analysis Results:

From the above state-wise analysis:

The states with the highest average PM10 levels (most polluted) are at the top.

The states with the lowest PM10 levels (least polluted) are at the bottom.

From the city-wise analysis:

The cities with the highest average PM10 levels are identified as the most polluted cities.

The cities with the lowest average PM10 levels are considered the least polluted.

This analysis helps in identifying pollution hotspots across both states and cities.

Visualization:

Not for this analysis.

1. **Time Series Analysis of Pollution Trends**

Introduction: This section focuses on the Time Series Analysis of pollution trends to understand how air quality has changed over time.

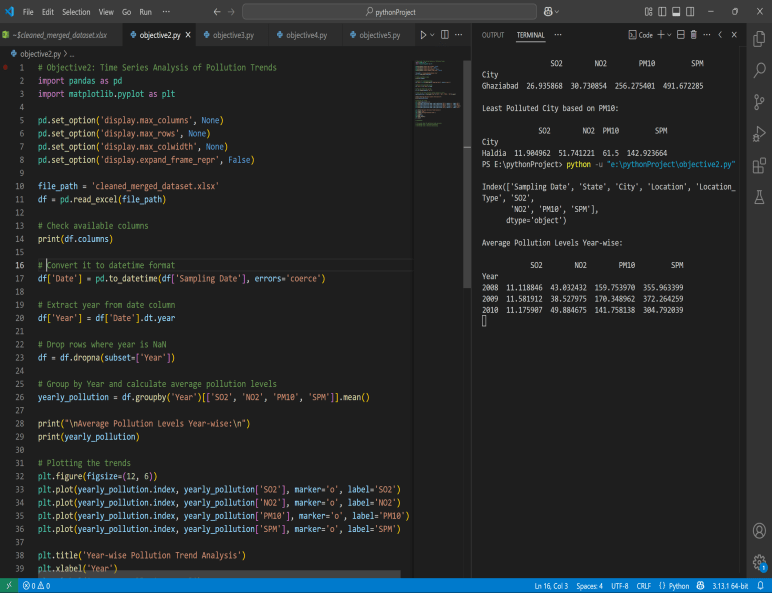
General Description: The dataset includes pollution readings collected over several years from various Indian cities. Each record contains a sampling date along with pollutant levels. This time-based data enables the study of pollution trends monthly, seasonally, or yearly to identify peaks, declines, or consistent behaviors in air quality over time.

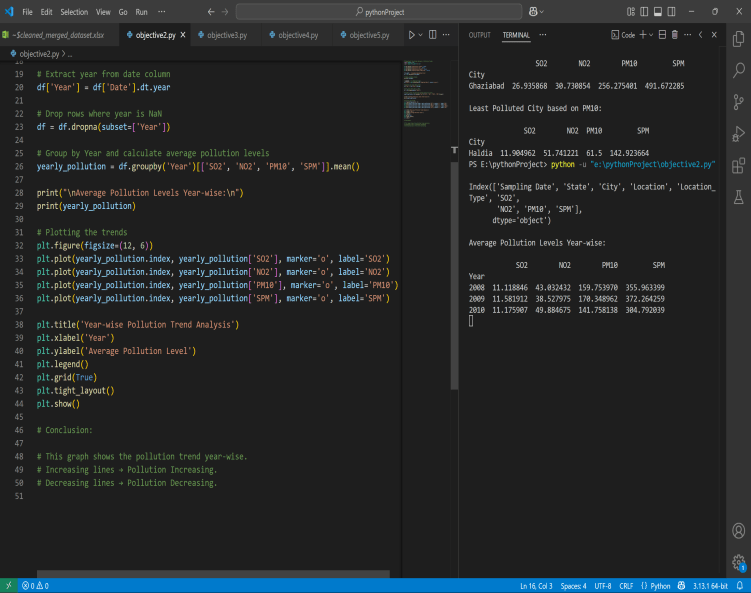
Specific Requirements, Functions, and Formulas: Data Conversion, TimebasedGrouping and usage of modules like pandas, matplotlib.

Implementation:

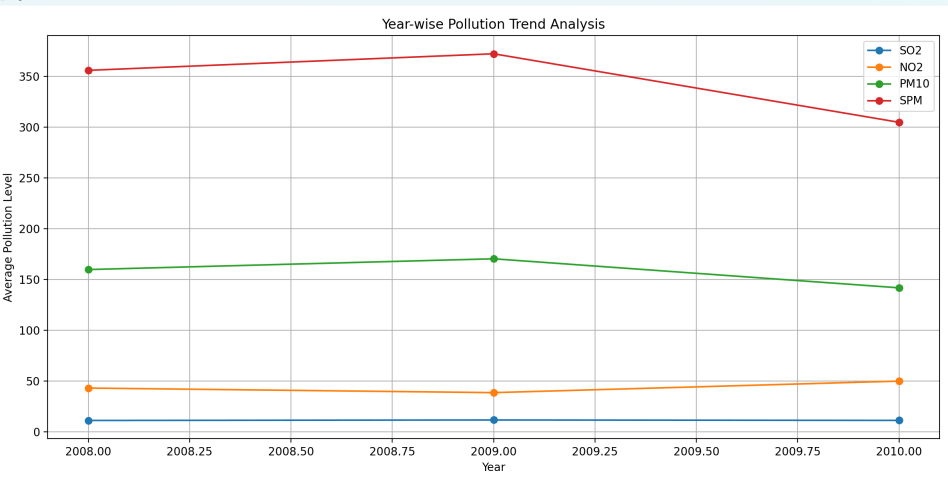
Code with output:

On next page-





Visualization:



Analysis Result:

This graph shows the pollution trend year-wise.

Increasing lines → Pollution Increasing.

Decreasing lines → Pollution Decreasing.

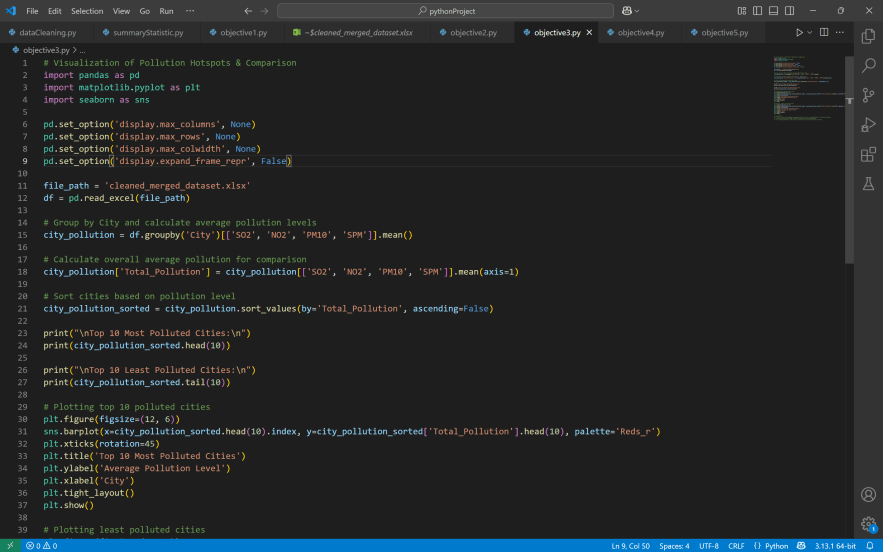
1. **Visualization of Pollution Hotspots & Comparison**

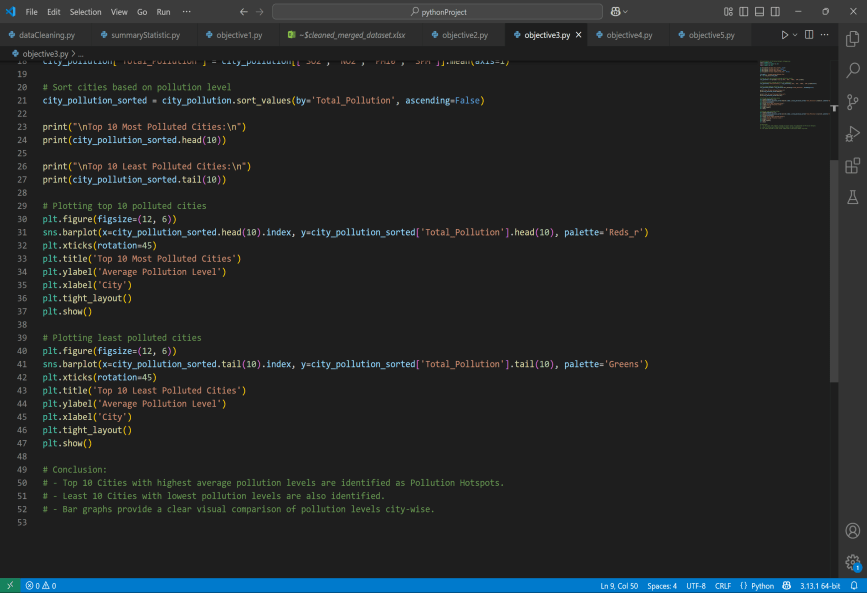
Introduction: This section aims to identify pollution hotspots—locations with consistently high pollutant levels—and compare pollution intensity across different regions

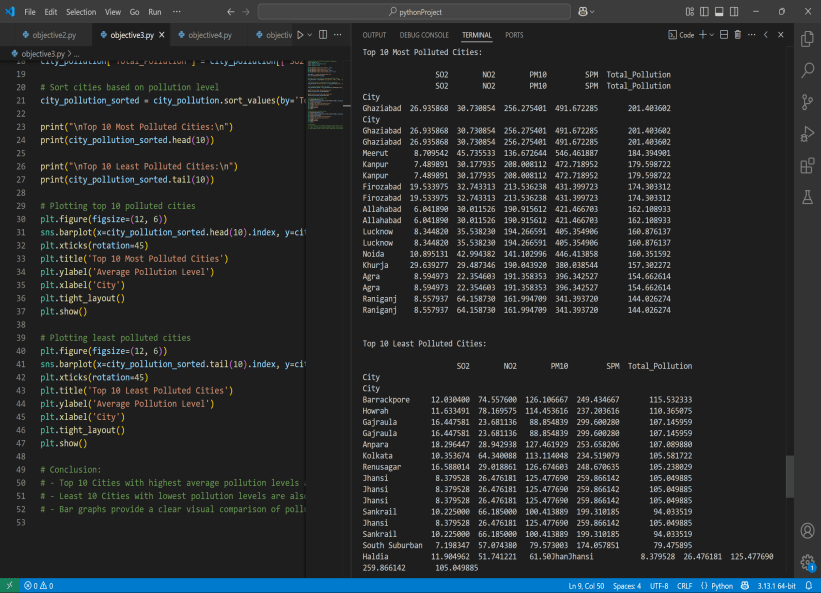
General Description: Using the dataset’s geographical components—State, City, and Location—we analyzed pollution readings (SO₂, NO₂, PM10, SPM) to determine which areas have the highest pollution concentrations.

Specific Requirements, Functions, and Formulas: groupby() function, sorting, and libraries like pandas, matplotlib etc.

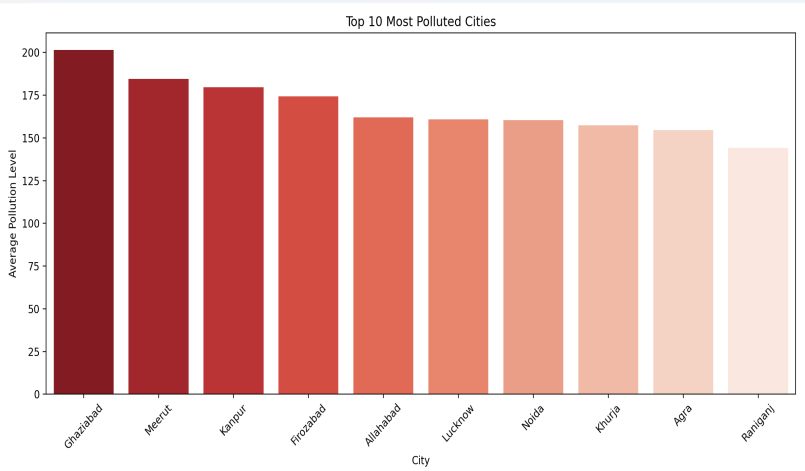
Implementation:

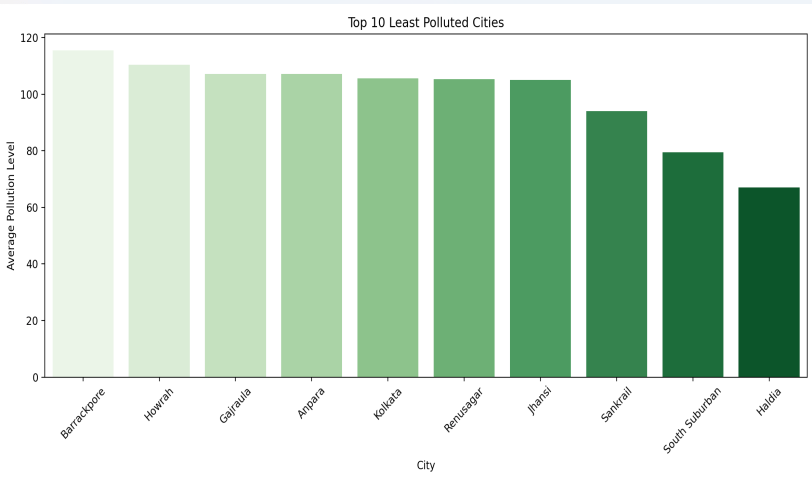






Visualization:





Result Analysis:

Top 10 Cities with highest average pollution levels are identified as Pollution Hotspots.

Least 10 Cities with lowest pollution levels are also identified.

Bar graphs provide a clear visual comparison of pollution levels city-wise.

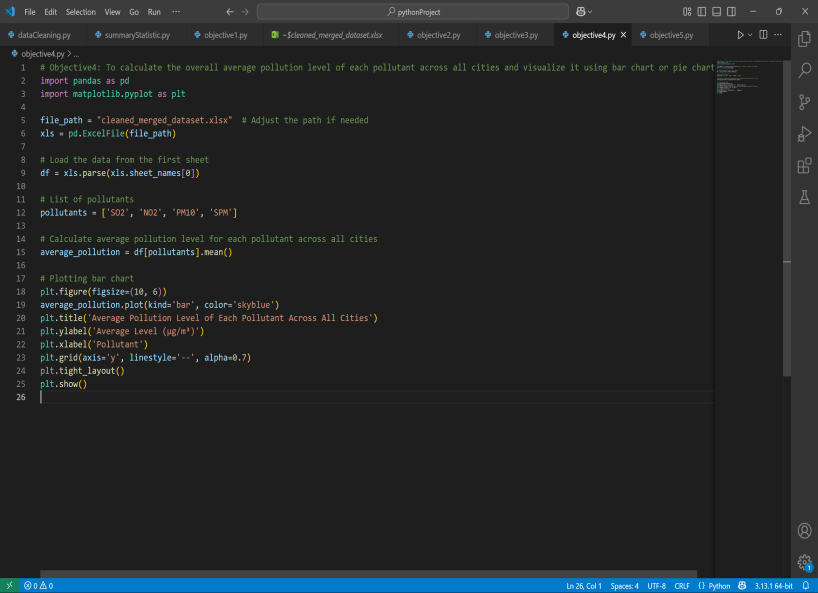
1. **To calculate the overall average pollution level of each pollutant across all cities and visualize it using bar chart**

Introduction: This section aims to compute the overall average concentration of each major air pollutant—SO₂, NO₂, PM10, and SPM—across all cities in the dataset. This helps in understanding which pollutants are most dominant nationwide and guides priorities for pollution control.

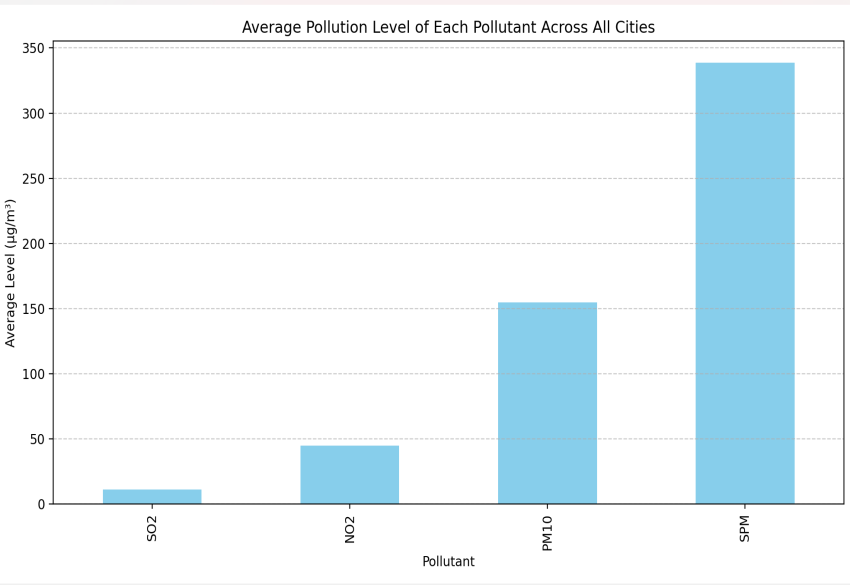
General Description: The dataset contains air quality measurements for various Indian cities. Each record includes the pollutant levels and city-wise data. By aggregating these values, we obtain the national average for each pollutant, which offers a high-level view of air pollution across urban India.

Specific Requirements, Functions, and Formulas: Libraries used matplotlib, pandas.

Implementation:



Visualization:



Result Analysis:

* Among all pollutants, PM10 and SPM had the highest average concentrations, indicating that particulate pollution is the most severe concern across cities.
* NO₂ had moderate levels, while SO₂ was relatively lower, reflecting cleaner fuel adoption and reduced industrial sulfur emissions in some areas.

1. **To create a correlation heatmap between various pollutants (SO₂, NO₂, PM10, SPM) to check their relationship with each other**

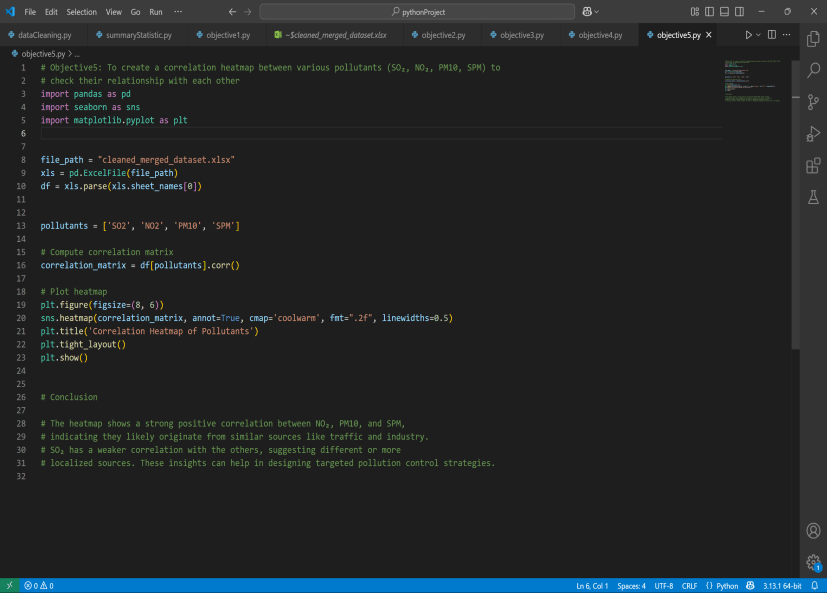
Introduction: This section aims to explore the relationships between major air pollutants—SO₂, NO₂, PM10, and SPM—by creating a correlation heatmap.

General Description: The dataset includes city-wise pollutant readings over time. By analyzing the pairwise correlation between pollutant concentrations, we can assess whether certain pollutants tend to rise or fall together, suggesting shared emission sources like traffic or industry.

Specific Requirements, Functions, and Formulas: Correlation Calculation,

Usage of libraries like pandas, matplotlib, seaborn

Implementation:



Visualization:

Result Analysis:

The heatmap shows a strong positive correlation between NO₂, PM10, and SPM,

indicating they likely originate from similar sources like traffic and industry.

SO₂ has a weaker correlation with the others, suggesting different or more

localized sources. These insights can help in designing targeted pollution control strategies.

**5. Conclusion:**

This project provides a detailed analysis of air pollution trends across Indian cities using statistical evaluation and visual exploration. The study reveals that particulate matter (PM10 and SPM) consistently shows the highest average concentration, making it the most critical pollutant affecting air quality nationwide.

City-wise and state-wise analyses highlight major pollution hotspots such as Kanpur, and Varanasi, driven largely by urbanization, traffic emissions, and industrial activity. Time series analysis uncovers seasonal trends, with pollution levels often peaking during winter months.

The correlation heatmap indicates strong relationships between NO₂, PM10, and SPM—suggesting shared sources like vehicular emissions—while SO₂ behaves more independently, pointing to different origins such as industrial fuel burning.

Overall, the findings underscore the need for region-specific pollution control strategies, stricter emission regulations, and increased public awareness to mitigate the health and environmental impacts of air pollution in India.

**6. Future Scope:**

**Health Impact Analysis:**  
Extend the analysis by correlating pollution data with public health records (e.g., respiratory diseases, hospital admissions) to study impacts on population health.

**Policy Evaluation:**  
Assess the effectiveness of pollution control measures and environmental regulations over time using pre- and post-policy data.

**7. Reference:**

1.Central Pollution Control Board (CPCB), India: <https://cpcb.nic.in>

2. Open Government Data (OGD) Platform India: <https://data.gov.in>

3. World Health Organization (WHO) – Air Pollution: <https://www.who.int/health-topics/air-pollution>

**Linkedin post :** [**https://www.linkedin.com/posts/harsh-porwal-hp01\_datascience-airpollution-python-activity-7316469762568437762-aTpQ?utm\_source=share&utm\_medium=member\_desktop&rcm=ACoAAEfKjmMBrGOboU4vgp\_pVm58CkuTXb1lMG8**](https://www.linkedin.com/posts/harsh-porwal-hp01_datascience-airpollution-python-activity-7316469762568437762-aTpQ?utm_source=share&utm_medium=member_desktop&rcm=ACoAAEfKjmMBrGOboU4vgp_pVm58CkuTXb1lMG8)

