**DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING**

**PROJECT REPORT**

**Analyzing Real-Time Air Quality Across Indian Cities**

Submitted by

Vadisala Manoj

Registration No.12321220

Programme and Section K23GD

Course Code INT 375

Under the Guidance of

**Baljinder Kaur (UID:28968)**

Assistant Professor

**Discipline of CSE/IT**

**Lovely School of Computer Science**

**Lovely Professional University, Phagwara**

**DECLARATION**

I Vadisala Manoj , student of BTech under CSE 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: 11-04-2025 Signature:A close up of a name

AI-generated content may be incorrect.

Registration No. 12321220

**CERTIFICATE**

This is to certify that VADISALA MANOJ bearing Registration no. 12321220 has completed INT 375 project titled **“ Analyzing Real-Time Air Quality 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 …………………………………………….**

Lovely Professional University

Phagwara, Punjab.

Date:

**ACKNOWLEDGEMENT**

I would like to extend my sincere gratitude to my guide, Mrs. Baljinder Kaur, for her invaluable guidance, support, and encouragement throughout the duration of this project. Her expertise, constructive feedback, and constant motivation have played a significant role in the successful completion of this work.

I would also like to express my sincere thanks to Lovely Professional University for providing me with the opportunity to undertake this project as part of the course **Data Science Toolbox: Python Programming (INT 375)**. The university has provided all the necessary resources and knowledge that formed the foundation for this research.

Furthermore, I wish to acknowledge the support of my peers and everyone who assisted me throughout the project. Their contributions, in any form, have been deeply appreciated.

Thank you all for your continuous support and encouragement.

**Name: Vadisala Manoj  
Section:** K23GD

**Roll No. :** 63

**Reg. No. :** 12321220  
**Date:** 13-04-2025

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **S No.** | **PROJRCT** | **Page No.** |
| **1.)** | Introduction | **6** |
| **2.)** | Source of dataset | **7** |
| **3.)** | Dataset Preprocessing | **7-8** |
| **4.)** | Analysis on dataset (for each objective)   1. General Description 2. Specific Requirements 3. Analysis results 4. Visualization | **9-20** |
| **5.)** | Conclusion | **21** |
| **6.)** | Future Scope | **21** |
| **7.)** | References | **22** |

**INTRODUCTION**

The Excel sheet titled "**Analyzing Real-Time Air Quality Across Indian Cities** " contains a comprehensive dataset detailing with Air pollution is a major environmental issue, particularly in developing countries like India, where rapid urbanization, industrial growth, and vehicle emissions have worsened air quality. Poor air quality impacts public health, the environment, and the economy, contributing to respiratory illnesses, heart disease, and other long-term conditions.

The **Air Quality Index (AQI)** is a standardized measure used to describe air quality in a specific area. It is calculated using concentrations of key pollutants such as **PM2.5, PM10, NO₂, SO₂, CO,** and **O₃**, and classifies air into categories like Good, Moderate, Poor, and Severe. These classifications help the public and policymakers understand pollution levels and take appropriate actions.

This project, titled **“Air Quality Index (AQI) Analysis of Indian Cities”**, focuses on analyzing real-time AQI data from various Indian locations. The main goals include:

* Examining AQI trends across cities and states,
* Identifying the most polluted regions and common pollutants,
* Categorizing air quality levels,
* Visualizing findings through clear and simple charts.

The analysis uses **Python**, with libraries such as **Pandas** for data handling, **NumPy** for calculations, and **Matplotlib** and **Seaborn** for visualizations. The code is written to be beginner-friendly and suitable for academic use, especially for students interested in data science and environmental analytics.

This study not only provides insight into air quality trends in India but also highlights how data science can be used to address real-world challenges.

1

**Source of Dataset**

The dataset used in this project, titled **“Real-Time Air Quality Index from Various Locations,”** has been sourced from [**data.gov.in**](https://data.gov.in), the official open data portal of the Government of India. This platform provides access to high-quality datasets published by various government departments and organizations.

The dataset includes real-time air quality monitoring data collected from multiple cities and states across India. It contains crucial information such as:

* **City** and **State**: Geographic location of the monitoring station.
* **Pollutant ID**: Type of air pollutant (e.g., PM2.5, PM10, NO₂, SO₂, CO, etc.).
* **Pollutant Average (pollutant\_avg)**: The average concentration of pollutants over a defined period.
* **Last Update**: Timestamp of the latest recorded data for each entry.

This dataset is valuable for conducting exploratory data analysis (EDA), identifying pollution trends, and classifying air quality levels based on national AQI standards. It supports data-driven decision-making for environmental monitoring and public health awareness.

📌 **Source:**  
 <http://www.data.gov.in/catalog/real-time-air-quality-index>

**DATASET PREPROCESSING**

Data preprocessing is an essential step in the data analysis process. It helps transform raw data into a structured format suitable for meaningful analysis. For the AQI Analysis project, various preprocessing techniques were applied to the air quality dataset to clean and organize it for further analysis and visualization. The following operations were performed:

1. **Data Loading**
   * The dataset was loaded into a pandas DataFrame using pd.read\_csv() from a CSV file.
   * The initial structure and quality of the dataset were examined using .head(), .info(), .describe(), and .isnull().sum() to identify missing values, incorrect data types, and redundant entries.
2. **Handling Missing Data**
   * Missing values in critical columns such as city and pollutant\_avg were identified.
   * Rows with null (NaN) values in these columns were dropped using data.dropna(subset=['city', 'pollutant\_avg'], inplace=True) to ensure that the analysis was based on complete records.
   * This step ensured that incomplete or unreliable data would not affect the analysis and visualizations.
3. **Type Conversion**
   * The last\_update column, which contained datetime information, was converted into a pandas datetime object using pd.to\_datetime().
   * Any invalid dates were coerced into NaT (Not a Time) using the errors='coerce' parameter.
   * Other columns, such as pollutant data, were checked and cast to the appropriate data types to ensure smooth analysis.
4. **Duplicate Removal**
   * Duplicates were identified and removed using data.duplicated().sum() and data.drop\_duplicates() to ensure each record represented unique air quality data from a specific location at a given time.
5. **Categorization of Air Quality Index (AQI)**
   * AQI values were categorized into predefined classes such as 'Good,' 'Moderate,' 'Poor,' 'Very Poor,' and 'Severe' based on thresholds defined in the classification function. This was done to make the analysis of AQI easier and more interpretable.
   * A new column, category, was created to store the AQI categories, and this information was later used in visualizations such as pie charts and bar plots.
6. **Statistical Summarization**
   * Summary statistics, including mean, standard deviation, minimum, and maximum values for the AQI data, were generated using data.describe(). These statistics helped to understand the distribution and identify any outliers in the data.
7. **Data Visualization Readiness**
   * After preprocessing, the dataset was cleaned and structured, making it suitable for generating visualizations like bar charts, pie charts, and scatter plots.
   * The dataset was now ready for meaningful analysis and presentation of trends, which provided a clearer understanding of the air quality across various regions

**Detailed Analysis Based on Project Objectives**

**Objective 1: Load and Clean the Dataset**

In this step, the dataset was loaded from the CSV file located at C:\Users\manoj\Desktop\INT375 project\Real time Air Quality Index from various locations.csv using pandas. The dataset was first examined to understand its structure, including the number of rows, columns, and types of data. The column names were cleaned by removing unnecessary spaces and converting them into lowercase for uniformity. The 'last\_update' column, which contained date-time information, was converted into the pandas datetime format for consistency. Missing values were handled by removing rows where essential columns like ‘city’ and ‘pollutant\_avg’ had missing values.

**Objective 2. AQI Analysis Across Cities**

**i. Introduction**

This analysis aims to evaluate air quality across different cities in India by calculating the average Air Quality Index (AQI) per city. The objective is to highlight the most polluted urban areas for potential intervention.

**ii. General Description**

The dataset contains city-level AQI measurements collected from multiple monitoring stations. It includes fields like city, pollutant\_avg, and last\_update.

**iii. Specific Requirements, Functions, and Formulas**

* Grouped by city using groupby('city')
* Computed mean AQI: pollutant\_avg.mean()
* Selected top 10 cities: sort\_values().head(10)
* Visualization: seaborn.barplot()

**iv. Analysis Results**

The top 10 most polluted cities were identified based on their average AQI. These cities exceeded acceptable air quality thresholds and pose serious health risks.

**v. Visualization**

A horizontal bar chart was used to display the top 10 cities with the highest AQI, making comparison straightforward.

**Objective 3. Frequent and Severe Pollutants**

**i. Introduction**

This analysis identifies which pollutants appear most frequently and which have the highest average AQI values.

**ii. General Description**

The dataset column pollutant\_id indicates the type of pollutant. Two metrics were analyzed: frequency and average AQI per pollutant.

**iii. Specific Requirements, Functions, and Formulas**

* Frequency count: np.unique()
* Severity: groupby('pollutant\_id')['pollutant\_avg'].mean()
* Visualization: pie chart for frequency and bar chart for severity

**iv. Analysis Results**

Five major pollutants were observed frequently, and some, like PM2.5 and NO2, showed higher average AQI levels, indicating a larger environmental and health impact.

**v. Visualization**

Pie chart for frequency and bar chart for severity were generated using matplotlib and seaborn.

**Objective 4. AQI Trend Within a City (Pie Chart)**

**i. Introduction**

This section categorizes AQI readings in the most polluted city into standard AQI classes.

**ii. General Description**

The AQI values were mapped to categories such as Good, Moderate, Poor, etc., for a better understanding of air quality trends in a high-risk area.

**iii. Specific Requirements, Functions, and Formulas**

* Function to classify AQI levels
* Applied using apply() on pollutant\_avg
* Counted using value\_counts()
* Visualization with pie chart

**iv. Analysis Results**

A city like Ghaziabad showed a high frequency of "Poor" and "Very Poor" AQI days, signaling urgent need for environmental reforms.

**v. Visualization**

A pie chart clearly showed the distribution of AQI categories in the selected city.

**Objective 5. Classification of Cities by AQI Category**

**i. Introduction**

Cities are grouped based on AQI category to understand the general air quality status across India.

**ii. General Description**

Average AQI was calculated per city and then categorized using the same classification logic as above.

**iii. Specific Requirements, Functions, and Formulas**

* Grouped by city and calculated mean AQI
* Categorized using the classification function
* Counted occurrences per category

**iv. Analysis Results**

A large number of cities fell into the "Moderate" and "Poor" categories, reflecting the urgent need for national-level air quality interventions.

**v. Visualization**

Bar chart representing number of cities in each AQI category.

**Objective 6. State-Level Scatter Plot Analysis**

**i. Introduction**

This analysis provides a state-wise view of air quality by aggregating average AQI values.

**ii. General Description**

State-level aggregation was done to summarize data at a broader geographic scale.

**iii. Specific Requirements, Functions, and Formulas**

* Grouped data by state and averaged AQI
* Visualized using seaborn.scatterplot()

**iv. Analysis Results**

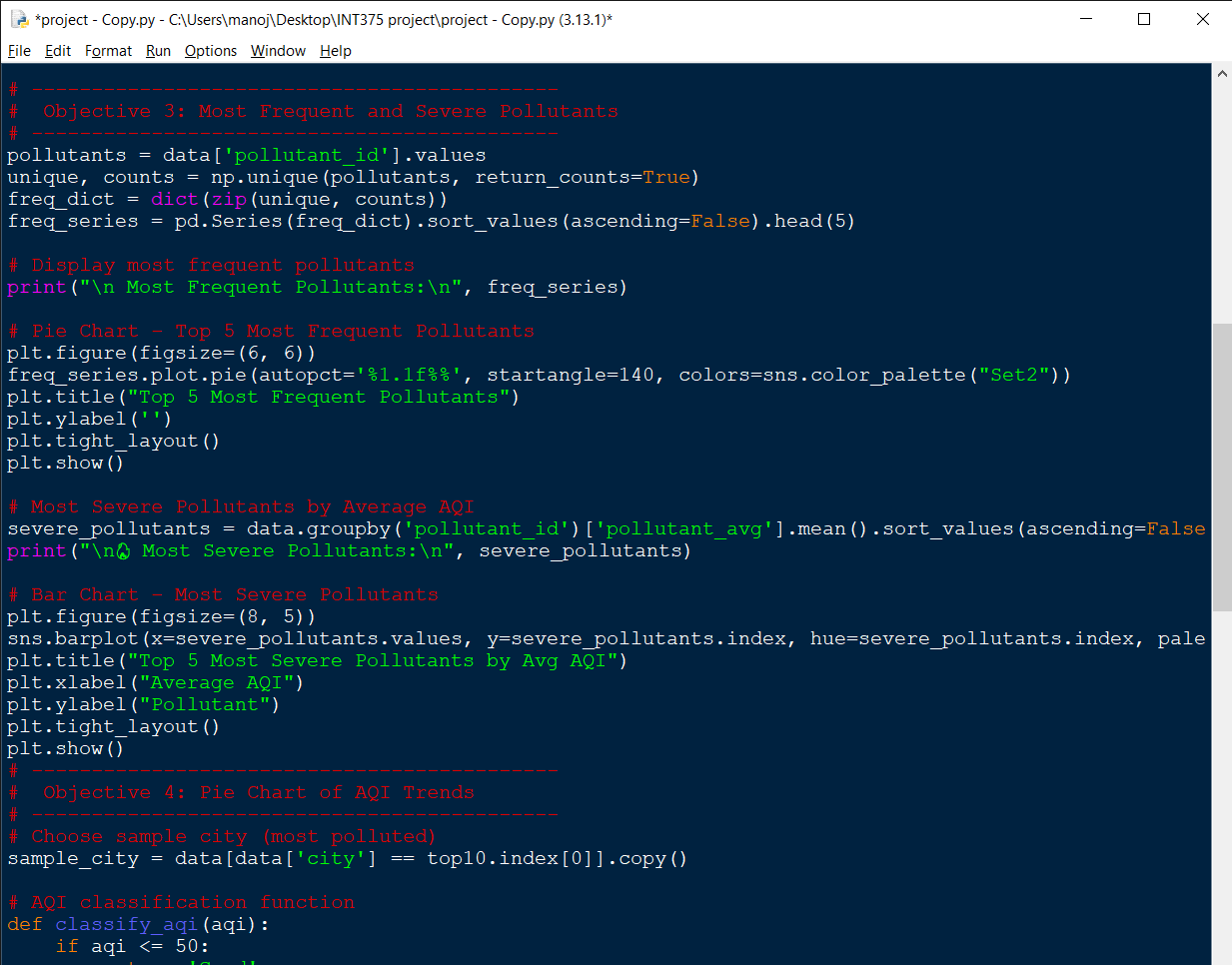
States like Delhi and Uttar Pradesh displayed higher average AQIs, confirming well-known pollution concerns.

**v. Visualization**

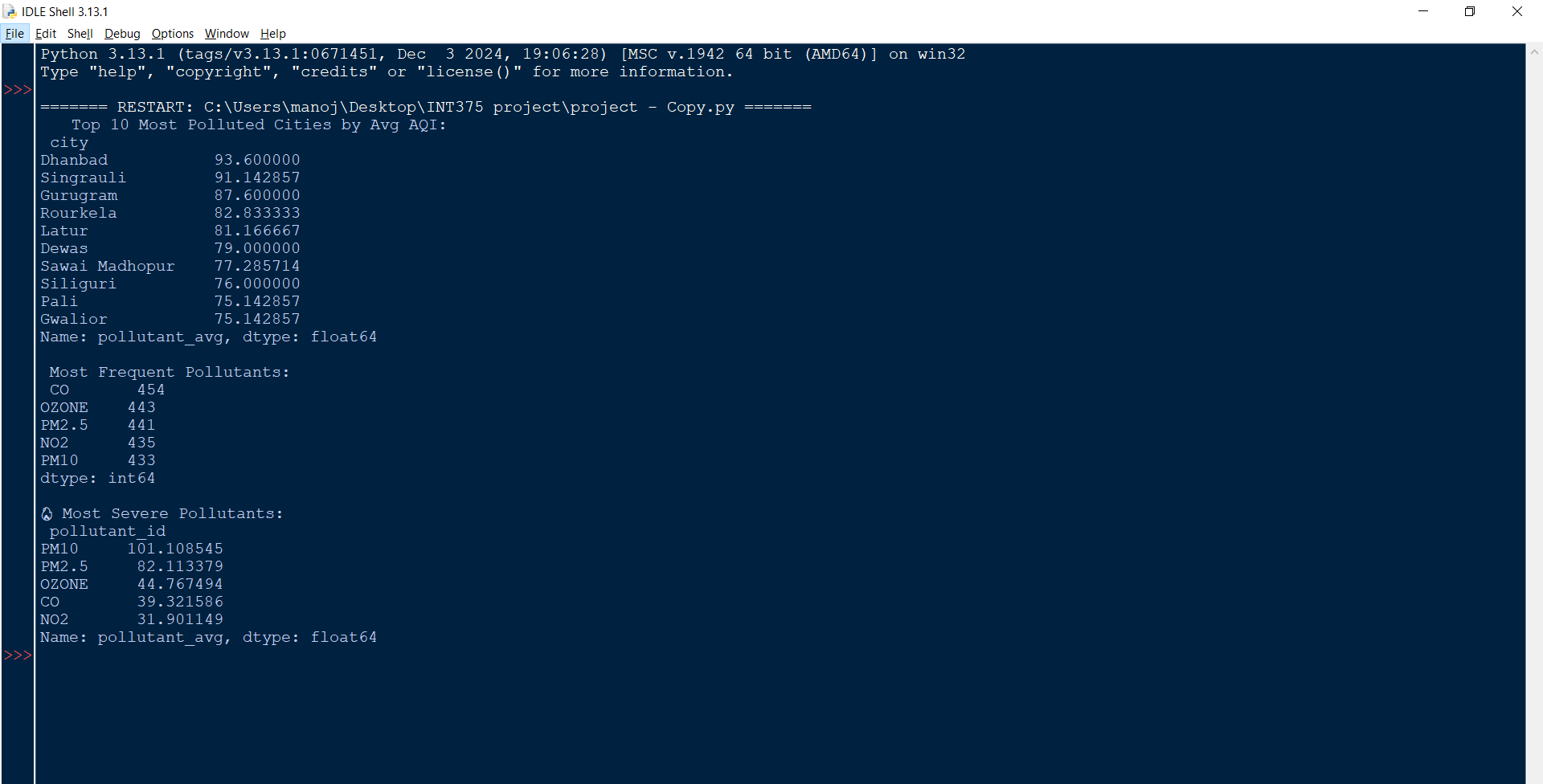
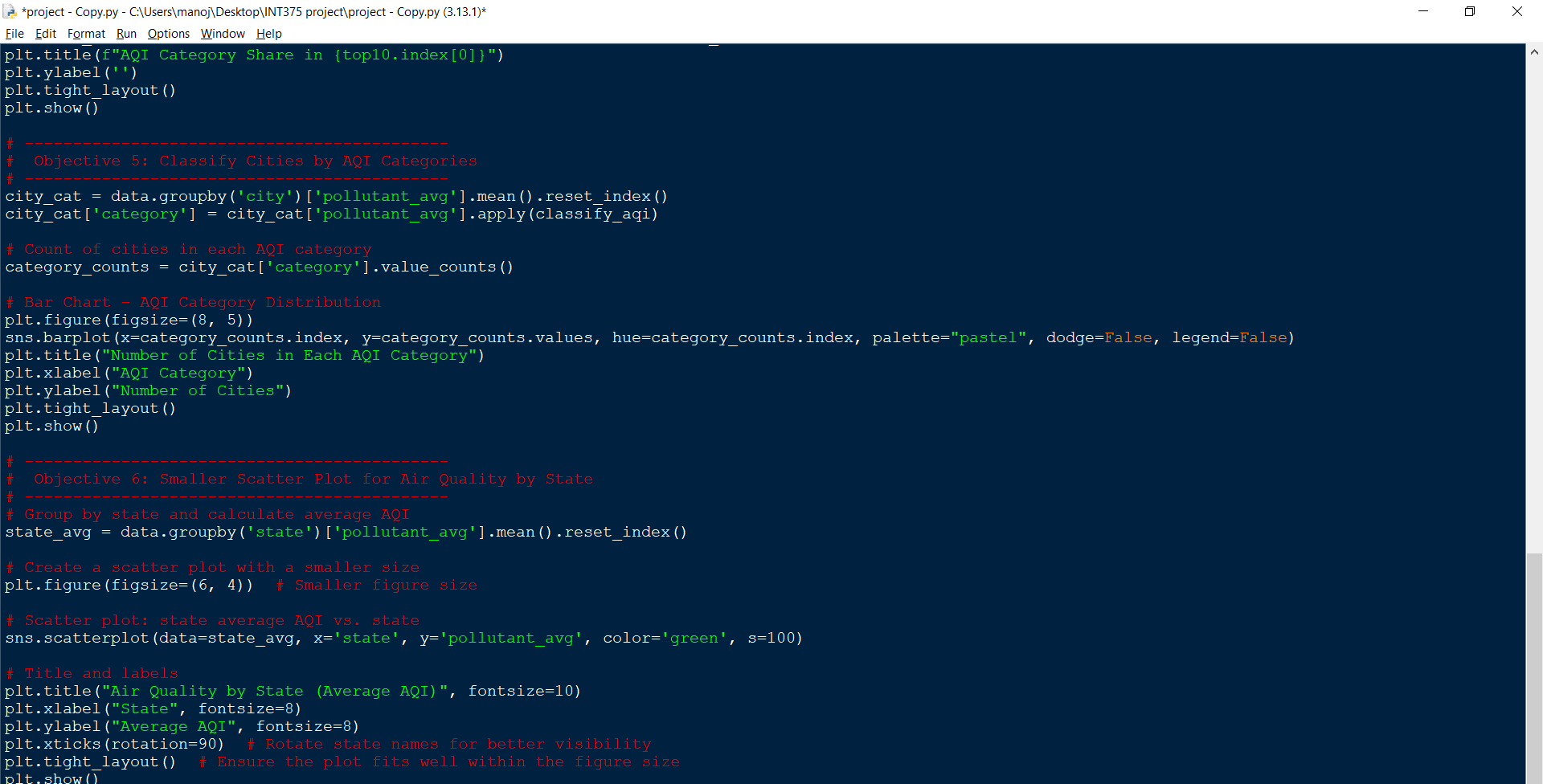
A simple scatter plot with rotated x-tick labels was used for space-efficient and informative display.

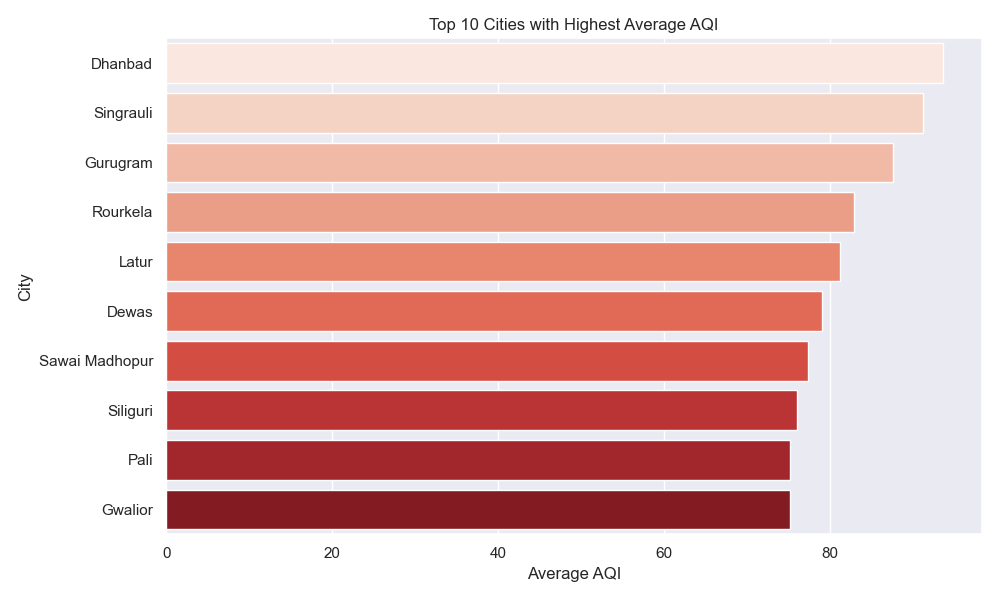
A screenshot of a computer program

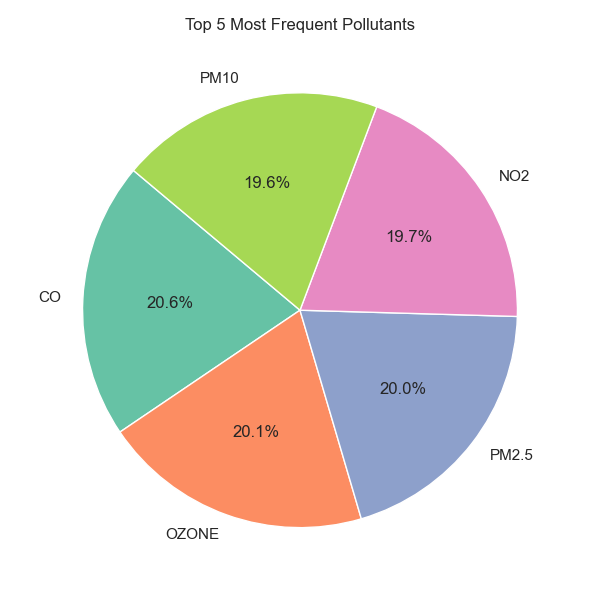
AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.





A graph of a number of pollutants

AI-generated content may be incorrect.A blue circle with a number of percentages

AI-generated content may be incorrect.A graph showing a number of cities

AI-generated content may be incorrect.A graph with green dots

AI-generated content may be incorrect.

**CONCLUSION**

This project aimed to analyze real-time air quality data across various Indian cities and states to gain valuable insights into pollution trends, pollutant severity, and geographical AQI variations. Through comprehensive data cleaning, preprocessing, and visualization using Python libraries such as Pandas, Matplotlib, and Seaborn, we were able to classify cities and states based on AQI levels, identify the most frequent and severe pollutants, and observe category-wise distributions.

The results indicate that certain pollutants, such as PM2.5 and PM10, are both frequent and contribute significantly to severe air quality. A handful of cities consistently reported high average AQI values, indicating poor air quality conditions. Additionally, the classification of AQI categories across cities revealed a substantial number falling into the 'Moderate' to 'Poor' categories, highlighting ongoing air quality concerns.

Overall, the project not only provided a clear and structured view of India’s air quality landscape but also emphasized the importance of continuous monitoring, localized mitigation strategies, and policymaking based on data-driven insights.

**6. Future Scope**

The AQI analysis project provides a strong foundation for further enhancements. Future work can involve integrating real-time air quality APIs for live data visualization and forecasting models using machine learning to predict pollution trends. Interactive dashboards with drill-down capabilities could help policymakers and the public better understand air quality dynamics. Additionally, expanding the analysis to include meteorological factors (e.g., wind speed, humidity, temperature) could improve the accuracy and depth of insights. More granular insights at ward or zone levels within cities can also support hyperlocal action plans.

**REFERENCES**

**Python Libraries Used in Obj.py**:

**Libraries and Tools:**

* Pandas: https://pandas.pydata.org/
* NumPy: <https://numpy.org/>
* Matplotlib: <https://matplotlib.org/>
* Seaborn: https://seaborn.pydata.org/

**Dataset Source:**  
 <https://www.data.gov.in/catalog/real-time-air-quality-index>

* LINKED IN LINK : <https://www.linkedin.com/posts/vadisalamanoj_datascience-airquality-python-activity-7317153881669414914-GQWy?utm_source=social_share_send&utm_medium=member_desktop_web&rcm=ACoAAEgOVrcBkG4UdvMabxr1JIk0vVpa9uL1NhQ>
* GitHub link: <https://github.com/Vadisalamanoj/-Analyzing-Real-Time-Air-Quality-Across-Indian-Cities>