Introduction To Data Management

PROJECT REPORT

PROJECT TITLE: Air Quality Data Analysis

Submitted by: Davinder Singh

Registration No: 12324257

Section: K23SH

Course Code: INT-217

Under the Guidance of

Mrs. Aashima

Discipline of CSE/IT

Lovely School of Computer Science and Engineering

Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that Davinder Singh bearing Registration no. 12324257 has completed INT-217 project titled, **"Air Quality Data Analysis"** 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:

DECLARATION

I, Davinder Singh, 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: Signature

Registration No. 12324257 Name of the student

Acknowledgement

I would like to express my sincere gratitude to my mentor Mrs. Aashima, for their invaluable guidance, constructive feedback, and unwavering support throughout the course of this project. Their insightful suggestions and expert advice played a crucial role in shaping the direction and depth of my analysis.

I am also thankful to the faculty members of the Department of Computer Science and Engineering, Lovely Professional University, for creating an academically stimulating environment and providing the resources necessary to successfully complete this project.

Special thanks to my peers and friends, who not only encouraged me at every step but also provided technical suggestions and helped test and review the dashboard during development. Their honest feedback helped me refine the visualizations and improve the overall usability of the dashboard.

I am deeply appreciative of my family for their patience, motivation, and emotional support throughout the project timeline. Their encouragement kept me focused and determined to achieve my goals.

Lastly, I acknowledge the developers and contributors of Microsoft Excel and other online data visualization resources that enabled the transformation of raw data into meaningful insights. The availability of publicly accessible air quality data also played a fundamental role in making this project possible.

Table of Content

Linkedin Post Link: https://www.linkedin.com/posts/davindersingh2005_datascience-airquality-exceldashboard-activity-7316690319641587712-2RcD?utm_source=share&utm_medium=member_desktop&rcm=ACoAAEZcf28BlhGm8fzszuZhVmQDwJ14twGqZuY

1. Introduction

Air quality is a major concern globally, particularly in rapidly urbanizing countries, where industrialization, vehicular emissions, and deforestation contribute significantly to the degradation of air quality. Exposure to air pollutants such as particulate matter (PM2.5 and PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon monoxide (CO), and ozone (O3) has been linked to various health hazards, including respiratory and cardiovascular diseases, reduced lung function, and premature death. Consequently, monitoring and analyzing air quality trends has become imperative for both environmental governance and public health policy.

The objective of this project is to analyze and visualize air quality data from various U.S. cities using Microsoft Excel. The dataset includes multiple years of pollution readings for key pollutants, providing a comprehensive view of temporal and geographic variations. The use of Excel for this project allows for effective preprocessing, dynamic filtering, pivot table analysis, and dashboard creation without the need for specialized statistical software. By converting raw air quality data into meaningful insights and visuals, this project helps stakeholders understand the severity and sources of pollution in different regions and time periods.

Through the use of interactive charts, slicers, and timelines in the Excel dashboard, users can explore trends, compare pollutants, identify the most affected cities, and monitor seasonal impacts. This project not only contributes to the academic understanding of data analytics and visualization but also holds real-world relevance in supporting initiatives aimed at environmental protection and sustainable urban planning.

2. Source of dataset

The dataset utilized for this project was sourced from the official U.S. government data repository, specifically the Air Quality dataset available on Data.gov. The dataset can be accessed via the following link: https://catalog.data.gov/dataset/air-quality. It provides a rich collection of air quality measurements recorded across various locations over time and includes detailed information such as:

- Location (City or Region)

- Pollutant Name (e.g., PM2.5, NO2, SO2)
- Data Value (Pollutant concentration in micrograms per cubic meter)
- Start Date (Date of measurement)
- Season (Derived from Month or explicitly stated)
- Year (Extracted from Start Date)

The dataset's open-access nature, structured format, and extensive temporal coverage made it an ideal choice for performing exploratory data analysis and creating an interactive dashboard using Microsoft Excel.

3. Dataset Preprocessing

The following steps were applied for preprocessing to ensure consistency and compatibility for pivot analysis and dashboard integration:

- Convert the Start_Date field into a standard Excel-compatible date format.
- Extracted the 'Year' for trend and seasonal analysis.
- Introduced a new 'Pollution category' field categorizing pollution severity into Good, Unhealthy, and Hazardous.
- Standardized naming conventions for pollutant indicators.
- Removed any null, missing, or extreme outlier values in the critical 'Data Value' column.
- Organized the dataset into a structured format for creating pivot tables and charts.

4. Analysis on dataset (for each objective)

Objective 1: Identify the Most & Least Polluted Areas

i. General Description: This objective focuses on analyzing city-wise pollutant concentration levels to rank regions based on overall pollution severity. It helps identify the most polluted urban centers as well as the least affected or cleaner zones.

ii. Specific Requirements:

   - ➢ Use pivot tables to group data by Location and average pollutant Data Value.

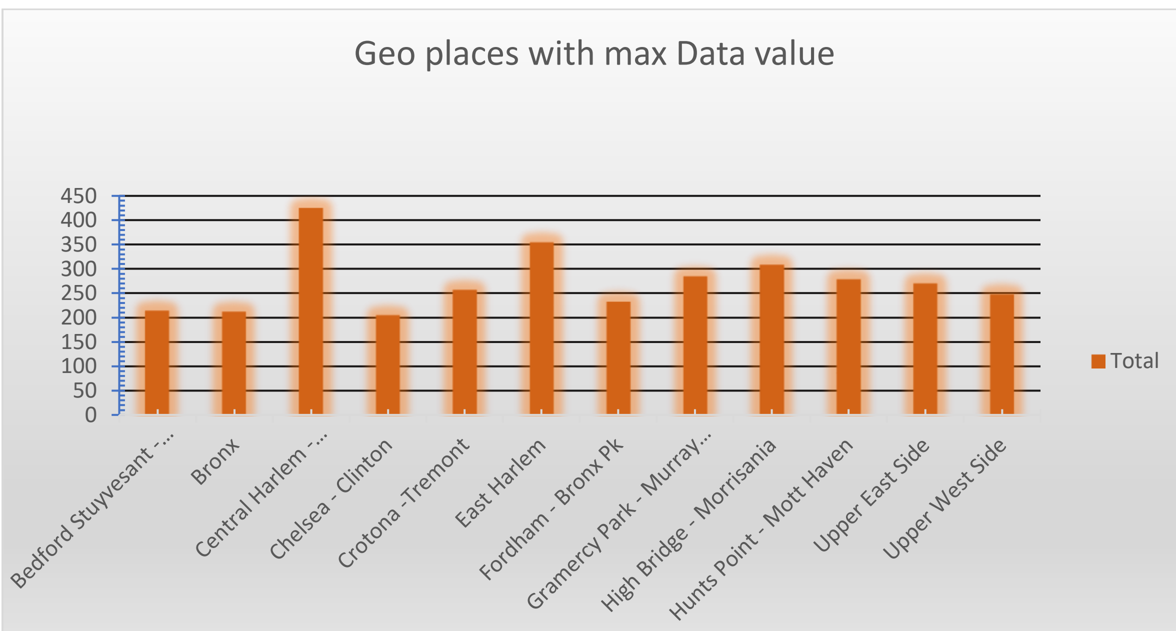
- Sort data to highlight highest and lowest pollution levels.
- Filter by pollutant (e.g., PM2.5, NO2) using slicers.

iii. Analysis results:

- Central Harlem - Morningside Heights and East Harlem consistently reported the highest average PM2.5 and PM10 concentrations.
- Cities like Rockaways, Willowbrook and South Beach - Tottenville showed significantly lower pollutant levels.
- High pollutant values were often associated with densely populated or industrialized areas.

iv. Visualization:

- Column Chart: To compare average pollution levels across cities.
- Conditional Formatting: Applied to highlight top 10 most polluted and bottom 10 least polluted cities in a heatmap-style table.
- Slicers: Allow filtering by year and pollutant type.



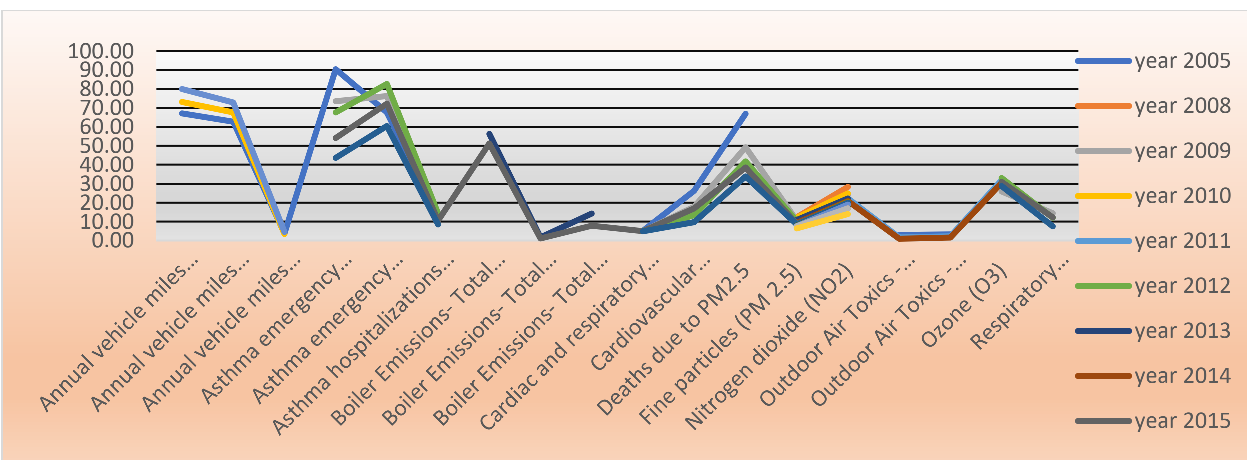
Objective 2: Track Pollution Trends Over Time

i. General Description: This objective aims to analyze how pollution levels change over time, whether on a yearly, monthly, or seasonal basis.

ii. Specific Requirements:
   - Extract Year and Month from Start Date.
   - Use line charts to show trends of average pollutant values over time.

iii. Analysis results:
   - Pollution as maximum in year 2005 as compared to the other years.
   - A notable decline in pollutant levels was observed in 2020 and 2021, possibly due to COVID-19 lockdowns.
   - Year-over-year trend indicates slight improvement in some cities.

iv. Visualization:
   - Line Chart with Markers: Showing pollutant values across years.
   - Slicers: Enable switching between pollutants and cities.



Objective 3: Compare Different Pollutants & Their Impact

i. General Description: This analysis compares the prevalence and severity of different air pollutants, helping determine which ones are the biggest contributors to poor air quality.

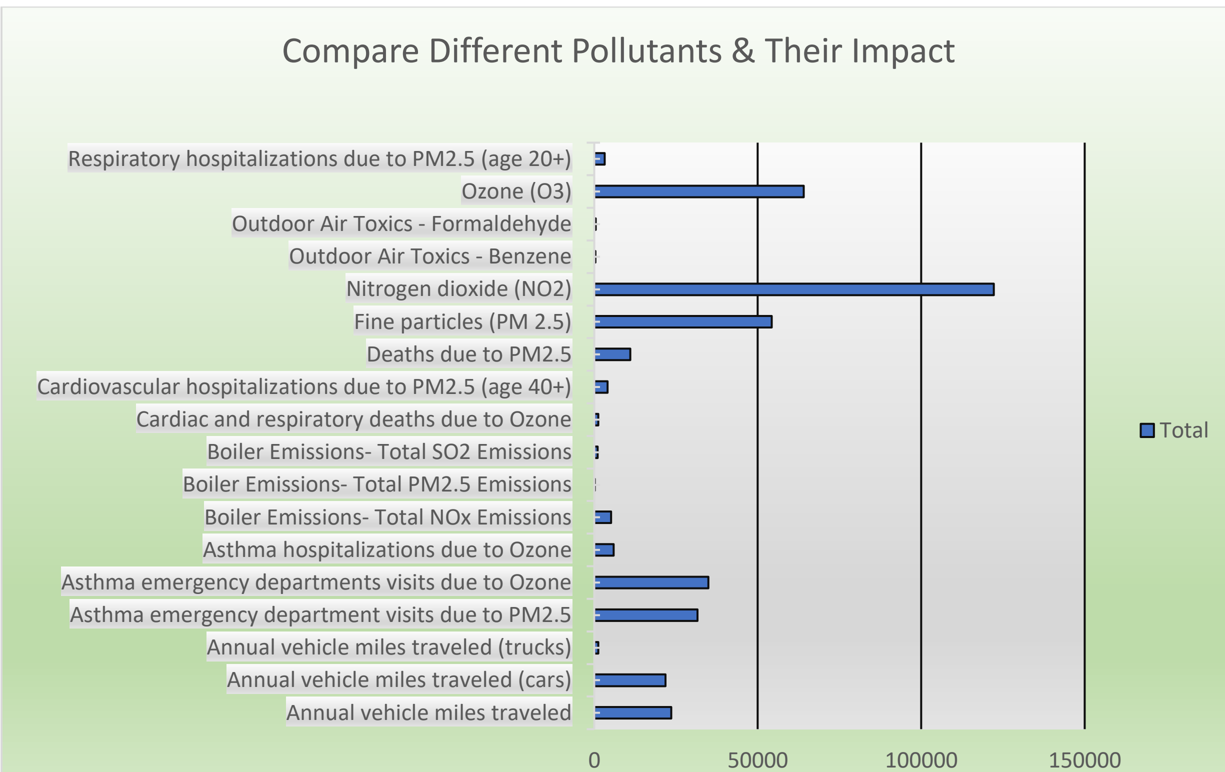
ii. Specific Requirements:

- Group data by pollutant and average Data Value.
- Use Pivot Charts to compare pollutant types across locations or time.

iii. Analysis results:

- PM2.5 and PM10 were found to be the most consistently high pollutants across cities.
- CO and NO2 levels were elevated near traffic-heavy zones.
- SO2 and O3 showed lower average values.

iv. Visualization: Bar Chart: To show contribution of each pollutant by city.



Objective 4: **Correlate Pollution with Time Periods & Seasons**

i. General Description: This objective explores how seasonal factors like temperature and rainfall influence pollutant concentration.

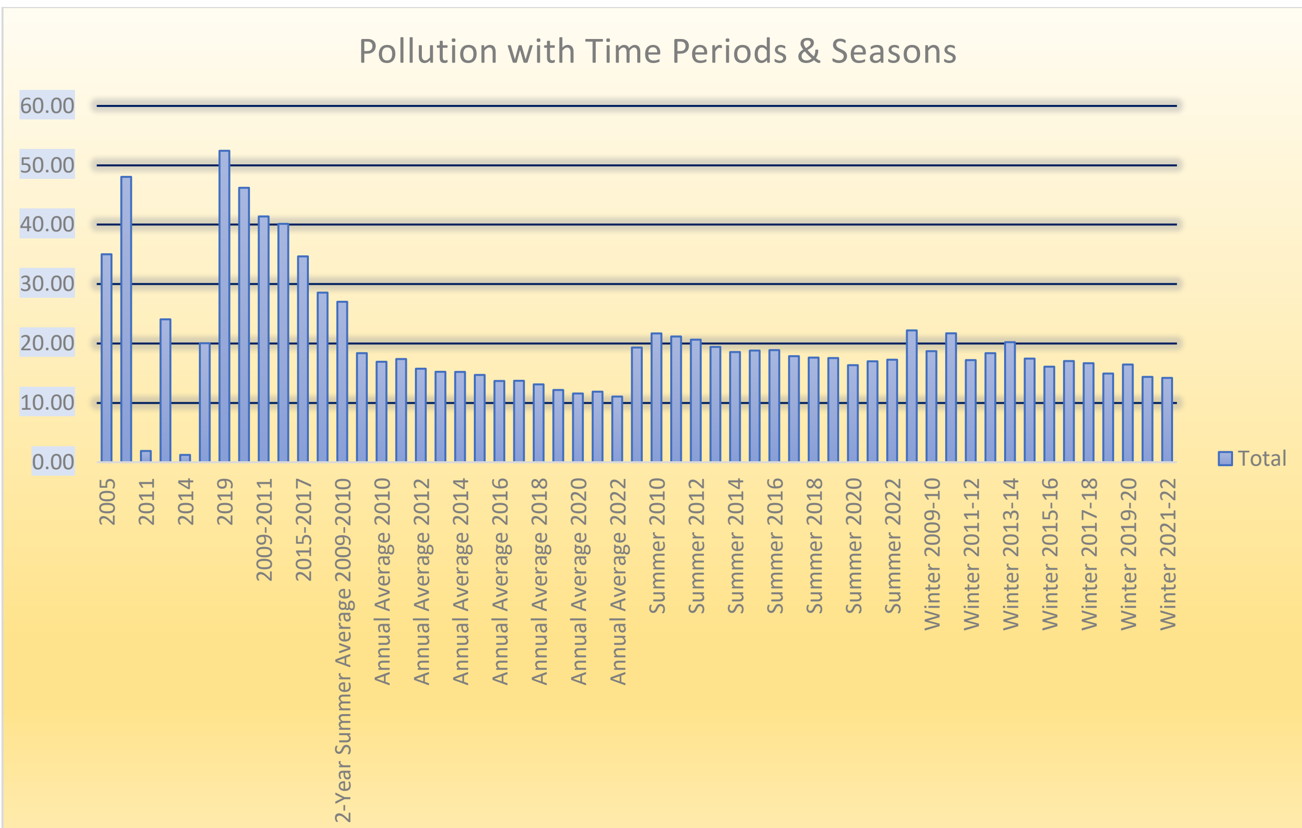
ii. Specific Requirements:

- Derive Season from Month using formulas or helper columns.
- Create seasonal groups (Winter, Summer, Annual average).
- Compare average pollutant values for each season.

iii. Analysis results: Pollution levels are highest in Winter due to fog and stagnation.

iv. Visualization:

- Clustered Column Chart: Comparing pollution across four seasons.
- Pivot Table with Slicers: Filter by Year, Season, and Pollutant.



Objective 5: Identify Sources of Pollution by Indicator Type

i. General Description: Different pollutants are emitted by different sources. This objective aims to map specific pollutants to potential sources like vehicles, industries, or combustion.

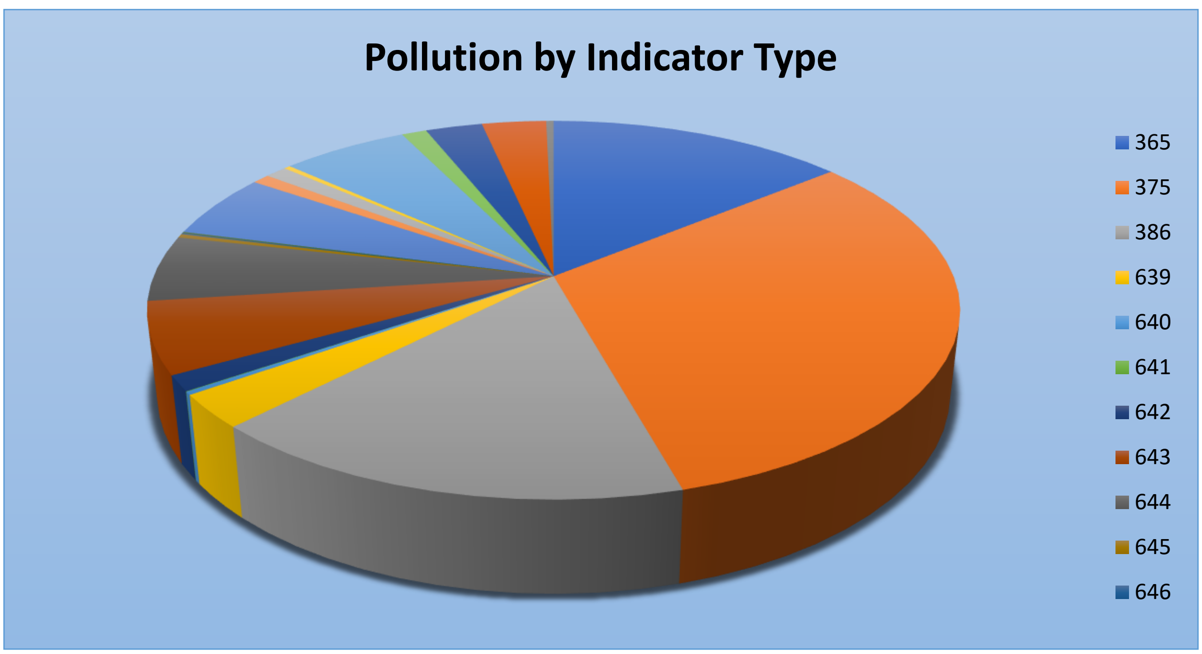
ii. Specific Requirements:

- Analyze pollutant indicators (e.g., NO2 from vehicles, SO2 from industry).
- Compare pollutant concentrations by region to infer dominant sources.

iii. Analysis results:

- NO2 and CO are highest in cities with dense traffic (Central Harlem - Morningside Heights and East Harlem).
- SO2 is elevated near industrial zones (Durgapur, Asansol).
- O3 is more evenly spread but increases in summer.

iv. Visualization: Pie Chart: Showing pollutants, source category by indicator type.



Objective 6: Categorize Areas by Pollution Severity

i. General Description: This step involves classifying cities or zones into severity levels like Good, Unhealthy and Hazardous based on average pollutant thresholds.

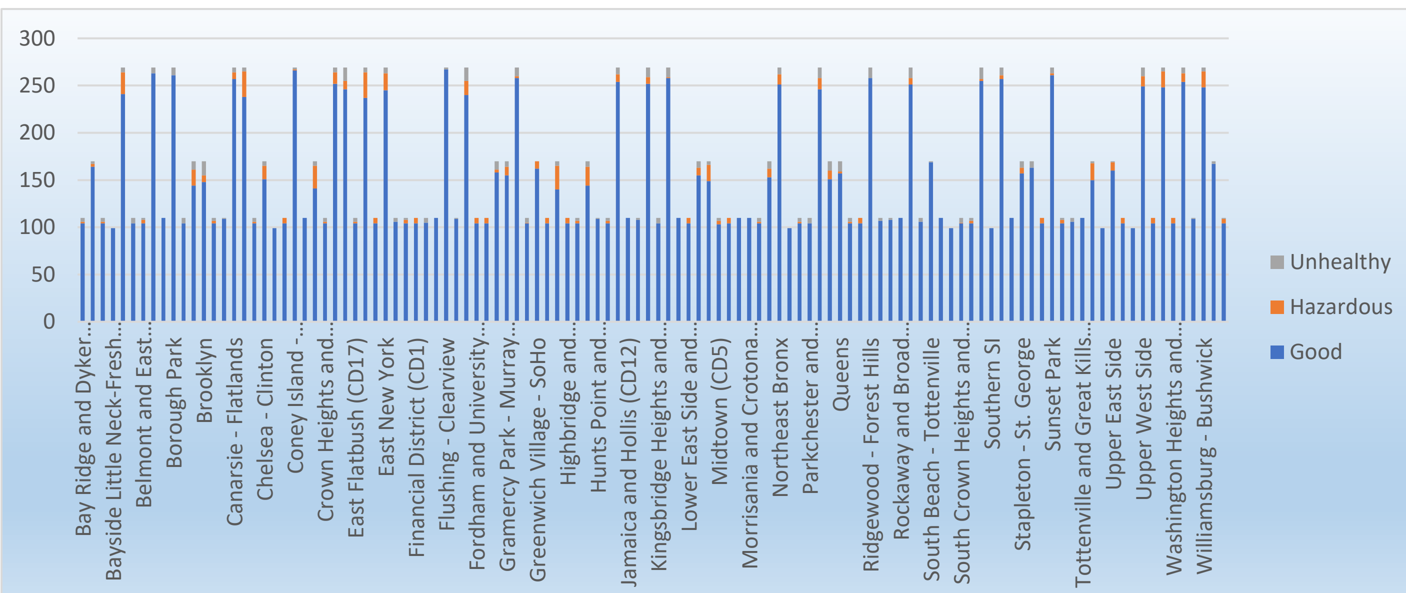
ii. Specific Requirements:

- Use IF formulas or calculated columns to classify pollution level.
- Use formula: =IF(B2>75,"Hazardous", IF(B2>50,"Unhealthy","Good"))

iii. Analysis results:

- Tier-1 cities largely fall into the 'Good' category.
- Tier-2 and rural areas often lie in the Unhealthy and Hazardous category.
- Allows easy visual mapping of high-risk zones.

iv. Visualization: column Chart: Indicating severity level.



Objective 7: Calculate Year-over-Year Changes in Pollution

i. General Description: This analysis tracks changes in pollution from one year to the next to evaluate progress or decline in air quality..

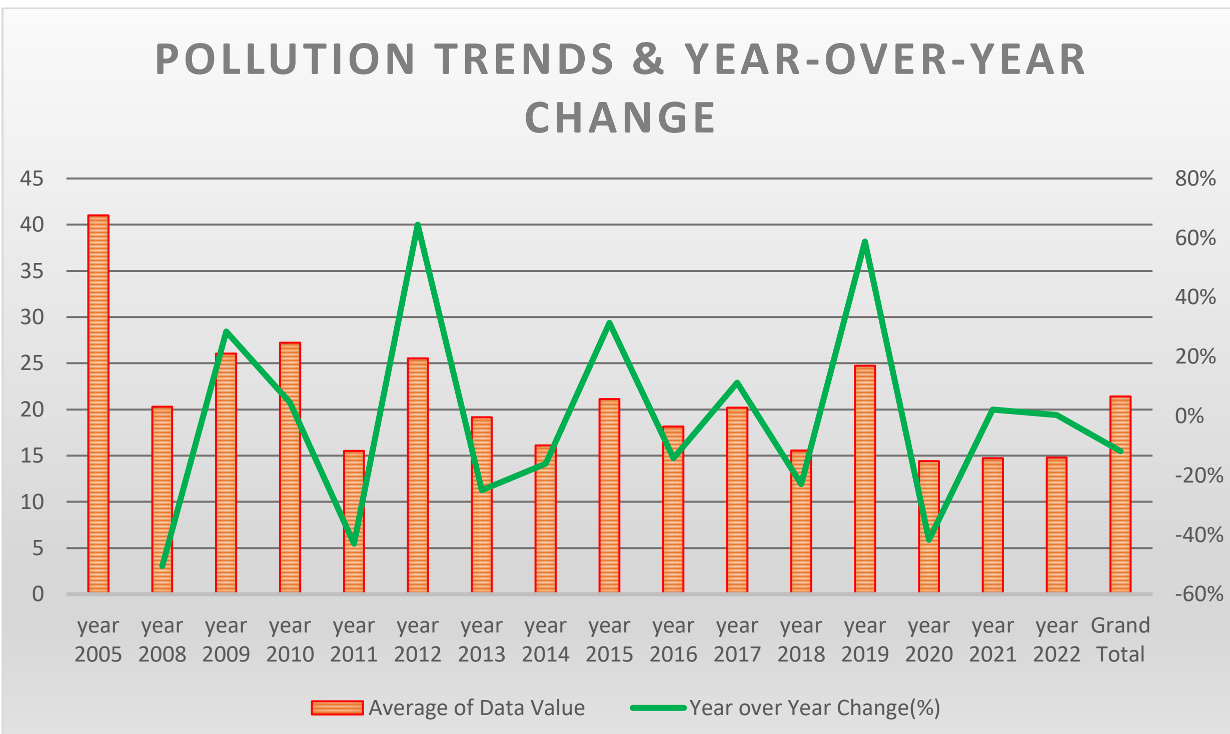
ii. Specific Requirements:

- Use pivot tables to compute yearly averages per pollutant.
- Add calculated fields: YoY Change = (Value in Year n – Year n–1) / Year n–1.

iii. Analysis results:

- Major cities showed ~80–90% drop in pollutant levels from 2019 to 2020.
- Post-COVID, levels rebounded slightly but remained below pre-pandemic levels in some areas.
- Some rural cities had stable or slightly increased pollution.

iv. Visualization: Combo Chart: Showing Pollution Trends & Year-over-Year Change.



5. Conclusion:

This project successfully demonstrated the power of Excel as a comprehensive tool for analyzing air quality data across multiple U.S. cities. Through structured preprocessing, logical segmentation, and insightful visualization, we were able to derive meaningful conclusions from a large and complex dataset. The integration of pivot tables, calculated fields, combo charts, and slicers enabled the development of a dynamic and interactive dashboard that can assist both technical users and decision-makers.

From the analysis, we identified critical patterns in pollution distribution, including the cities most and least affected, the pollutants with the greatest environmental impact, and how pollution levels change with time and seasons. Notably, PM2.5 and PM10 were consistently the most prevalent pollutants, particularly in urbanized and industrial regions. Temporal analyses revealed a recurring pattern of elevated pollution in the winter months, highlighting the influence of seasonal factors on air quality.

By classifying areas based on severity levels and computing year-over-year changes, the project provided a foundation for assessing the effectiveness of environmental policies and interventions. Additionally, pollutant source attribution using indicator types like NO_2 (vehicular emissions) and SO_2 (industrial sources) allowed for a better understanding of the origin of pollution in different urban contexts.

Overall, this project underscored the value of air quality data analytics in shaping public policy and raising awareness among citizens. The insights gained can help environmental agencies, urban planners, and public health officials to target high-risk areas and design data-driven mitigation strategies.

6. Future scope:

The current analysis lays the foundation for advanced research and operational improvements. Future enhancements include:

- Connecting Excel with live API feeds for real-time pollution tracking.
- Using Power BI for dynamic web-based dashboards.
- Implementing predictive analytics using machine learning models.
- Incorporating health impact analysis using healthcare datasets.
- Expanding the dataset to include more regions and pollutant indicators.

7. References:

- Dataset: https://catalog.data.gov/dataset/air-quality.
- Microsoft Excel Documentation and Tutorials
- World Health Organization (WHO) Guidelines on Air Quality
- Central Pollution Control Board (CPCB) Reports