International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, Delaware, USA
Volume 12– No. 47, July 2025 – www.ijais.org

Global Patterns of Urban Air Pollution: A Multivariate and Cluster-based Analysis Across Six Continents

Anika Rahman
Department of Computer Science and Engineering
Stamford University Bangladesh,
Dhaka, Bangladesh

Mst. Taskia Khatun, PhD
Department of Software Engineering
Daffodil International University,
Dhaka, Bangladesh

ABSTRACT
This study investigates global urban air pollution patterns across six continents using an extended version of the Global Air Pollution Data, covering major cities in Asia, Africa, Europe, Australia, North America, and South America. Building on our previously published work at the FMLDS2024 Conference, which focused exclusively on Asian cities, this research broadens the geographic scope and integrates multivariate statistical techniques alongside six clustering algorithms: K-Means, Hierarchical Clustering, DBSCAN, Gaussian Mixture Models (GMM), Agglomerative Clustering, and Spectral Clustering. Cluster performance is evaluated using multiple metrics, including Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index, WCSS, Cohesion, and Separation. The analysis identifies three distinct pollution clusters: 'High Pollution,' 'Low Pollution,' and 'Ozone-Dominated Pollution.' South Asia and East Asia exhibit the highest concentration of cities with ozone-dominated pollution, while Western Europe shows the greatest prevalence of low pollution cities. North America has the largest number of cities classified in the high pollution cluster, primarily driven by particulate matter (PM2.5) and nitrogen dioxide (NO2). Additionally, a focused analysis of capital cities provides further insight into regional urban air quality variations. This global analysis offers critical understanding of spatial disparities in urban pollution and underscores the necessity for region-specific strategies to mitigate pollution sources. The findings aim to support policymakers and environmental agencies in developing targeted air quality management plans.

Keywords
Urban Air Pollution, Clustering Algorithms, Global Patterns, Multivariate Analysis, Capital Cities.

1. INTRODUCTION
Urban air pollution is a critical global environmental challenge, posing severe health risks and contributing to climate change. The United Nations Environment Program reported that indoor and outdoor air pollution caused 6.5 million premature deaths worldwide in 2016 [1], making it one of the leading preventable causes of death and disease globally. As urbanization and industrialization accelerate, the increased consumption of fossil fuels exacerbates air quality issues, leading to more frequent occurrences of hazy weather [2-4]. Air pollution has thus emerged as the single biggest environmental risk to human health, demanding immediate attention and effective management strategies.

While many studies have focused on specific regions or pollutants, there remains a lack of comprehensive, global perspectives on urban air pollution patterns. This study seeks to bridge that gap by analyzing air quality across six continents—Asia, Africa, Europe, Oceania, North America, and South America—using an extensive dataset that includes major cities worldwide. Building upon our earlier research on air pollution in Asian cities, we aim to provide a broader understanding of pollution levels and patterns across different regions of the world. A preliminary version of this work focusing on urban air pollution patterns in Asian cities was published in the proceedings of the FMLDS2024 Conference [5]. This journal article extends the analysis to six continents and includes additional clustering algorithms, evaluation metrics, and policy discussions. Forecasting air pollution and understanding its spatial distribution is crucial for public health guidelines and government pollution management efforts [6].

This work adopts a systems-level view of global urban air pollution, where each urban region represents a component in a globally interacting environmental system. By treating cities and regions as interdependent subsystems within a broader planetary framework, we aim to understand not only individual pollutant levels but also the emergent global patterns of pollution behavior. Through multivariate analysis and clustering, we identify systemic trends and regional groupings that can inform cross-border cooperation, targeted interventions, and international policy-making strategies.

To achieve this, we employ clustering algorithms, a robust and widely used technique for exploring complex data. Clustering enables the identification of hidden patterns and natural groupings in the dataset, allowing us to group cities with similar pollution characteristics. This approach provides a data-driven framework to classify cities into meaningful clusters, facilitating the development of targeted interventions and policies. Specifically, we applied six clustering algorithms: K-Means, Hierarchical Clustering, DBSCAN, Gaussian Mixture Models (GMM), Agglomerative Clustering, and Spectral Clustering, to analyze urban air pollution patterns. These algorithms were chosen for their ability to handle diverse data characteristics and provide insights into the variability of pollution across global cities.

For each clustering algorithm, we evaluated performance using various metrics, including Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index, WCSS, Cohesion, and Separation. The four primary pollutants considered in this study are particulate matter (PM2.5), carbon monoxide (CO), nitrogen dioxide (NO2), and ozone. Our analysis identified three distinct pollution clusters—'High Pollution,' 'Low Pollution,' and 'Ozone-Dominated Pollution.' Notably, South Asia, East Asia, and parts of Africa exhibit the highest number

of cities in the 'Ozone-Dominated Pollution' cluster, driven by elevated levels of PM2.5 and NO2. In contrast, Western Europe hosts the highest number of cities in the 'Low Pollution' cluster, while North America contains the largest number of cities in the 'High Pollution' cluster.

Beyond clustering analysis, this study incorporates region mapping for each continent to elucidate the spatial distribution of air pollution. Additionally, a detailed analysis of capital cities highlights the primary pollutants contributing to air quality issues in urban centers. The study also identifies the most polluted cities within specific countries across all continents, providing valuable insights into areas requiring urgent attention. These findings emphasize the need for targeted policies and interventions to address the unique air quality challenges faced by different regions. By offering a global perspective on urban air pollution, this work aims to inform and guide policymakers and environmental agencies in mitigating the adverse effects of air pollution on public health and the environment.

2. RELATED WORK

Over recent decades, air pollution in Asia has become a major threat to food security [7] and human health [8]. Wild-fire smoke, pollen-based aeroallergens, and climate change, primarily due to greenhouse gas emissions, heavily influence PM2.5 levels. Research highlights severe impacts on health and the environment, with 37 of the world's 40 most polluted cities in South Asia [9]. Contaminated air in this region leads to millions of preventable deaths annually and harms crops essential for feeding many [10]. Biomass burning is a significant source of haze [11], accounting for up to 40–60% of haze events in Southeast Asia from 2003 to 2014 [12]. This worsening air quality presents significant challenges for sustainability and health. Considering these situations, forecasting air quality is becoming more and more important as air pollution becomes a serious environmental and social problem on a worldwide scale, preventing both health and financial damage [13]. At the moment, the majority of studies only offer air pollution projections for one or a few locations rather than data for the entire city and a collection of cities [14-15]. The distribution of air quality within different cities can vary greatly due to the intricate the structure and spatial movement of air contaminants, making precise forecasting very difficult [16].

Different algorithms and methods are used for the analysis of big data. Clustering algorithms are widely utilized in data mining and analytics [17-18]. The clustering algorithm can also be used to decompose PM2.5 concentration forecasting data. The clustering technique can classify the original data based on various air pollution conditions [19]. Cluster analysis is a popular method for dividing chaotic data into many categories with high similarity in order to evaluate internal patterns [20]. A large number of NWP samples, which have a significant influence on forecasting accuracy, are chosen as inputs to the DBN model using clustering analysis to improve the model's efficiency [21]. The clustering process can give model training examples that are highly comparable, which shortens the training period and improves generalization capacity [22]. In order to improve the data regularity, cluster analysis has been applied to air quality forecasting which is done by choosing or splitting the input variables [23].

However, in recent, other than clustering, machine learning and deep learning methods have also been used for air quality prediction. Several models have been created to forecast these potential consequences. However, making precise forecasts is very impossible. A hybrid intelligent model integrating LSTM and MVO has been created to forecast air pollution from Combined Cycle Power Plants [24]. A deep learning framework using a temporal sliding LSTM extended model has been created [25].

A model was built for buildings and pollution category labels in Beijing from 2013 to 2017 to train a convolutional neural network (CNN) [26]. Different machine learning-based models are used for air quality prediction systems by measuring the different gases present in the atmosphere [27].

3. METHODOLOGY

3.1 Data Collection

The dataset used in this study was obtained from Kaggle [28], containing over 21,294 rows and 12 columns, covering 157 countries and more than 21,000 cities globally. For the purposes of this analysis, we focused on cities from diverse regions worldwide, including Asia, Africa, Europe, Australia, North America, and South America, in order to study global air pollution patterns. The dataset includes key columns such as country name, city name, AQI value, AQI category, CO AQI value, CO AQI category, ozone AQI value, ozone AQI category, NO2 AQI value, NO2 AQI category, PM2.5 AQI value, and PM2.5 AQI category. These columns provide a comprehensive view of the air quality in urban areas, capturing key pollutant levels and their associated categories across different countries and cities. This dataset served as the foundation for performing multivariate analysis, clustering, and the subsequent pollution pattern analysis.

3.2 Data Preprocessing

To ensure the quality and usability of the dataset, several preprocessing steps were undertaken using **Python libraries** including Pandas, NumPy, and Scikit-learn.

3.2.1 Handling Missing and Invalid Data

One record with a missing city name was removed. No other missing entries were observed. All numerical attributes were converted using pd.to_numeric() with invalid entries coerced to NaN, which were then imputed with column-wise mean values.

3.2.2 Categorical Encoding

Categorical attributes (e.g., AQI category, pollutant categories) were cast to the category datatype to optimize memory and allow efficient analysis.

3.2.3 Outlier and Error Checking

The dataset was validated for negative or nonsensical values; no invalid pollutant concentrations were found.

3.2.4 Log Transformation

Log transformation was applied to pollutant AQI values to:

- Normalize skewed distributions
- Reduce the effect of outliers
- Improve algorithmic performance during clustering

3.2.5 Duplicate Removal

Duplicate entries were checked and none were found.

These steps ensured a clean and standardized dataset, suitable for statistical and machine learning-based analysis.

3.3 Determining the Optimal Number of Clusters

To determine the optimal number of clusters for our analysis, we employed two complementary methods:

3.3.1 Elbow method

We utilized the Elbow Method, which involves plotting the sum of squared distances between data points and their respective cluster centroids for varying numbers of clusters. The optimal number of clusters is identified at the "elbow" point of the curve, where the rate of decrease in the sum of squared distances slows significantly. This point indicates a balance between the number of clusters and the variance explained.

3.3.2 Silhouette analysis

In addition, we performed Silhouette Analysis, which measures how similar each data point is to its own cluster compared to other clusters. The average silhouette score was calculated for different cluster counts. The optimal number of clusters corresponds to the highest silhouette score, which indicates well-separated and cohesive clusters.

By applying both the Elbow Method and Silhouette Analysis, we ensured a robust and reliable determination of the optimal number of clusters, enhancing the stability and interpretability of the clustering results.

3.4 Clustering Algorithms

We applied several clustering algorithms to analyze the dataset, each offering a unique approach to partitioning the data.

- K-Means Clustering: This algorithm partitions the data into a predefined number of clusters by minimizing the within-cluster variance, ensuring that data points within each cluster are as similar as possible.
- Hierarchical Clustering: Using both agglomerative and divisive methods, Hierarchical Clustering creates a tree-like structure (dendrogram) that visually represents the relationship between clusters. This approach does not require specifying the number of clusters in advance.
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise): DBSCAN identifies clusters based on the density of data points, making it effective at detecting outliers and finding arbitrarily shaped clusters.
- Agglomerative Clustering: A type of Hierarchical Clustering, Agglomerative Clustering iteratively merges clusters based on a chosen distance metric, providing a flexible approach to clustering.
- Gaussian Mixture Models (GMM): GMM assumes that the data is generated from a mixture of Gaussian distributions. It provides probabilistic cluster assignments, allowing for soft clustering and accommodating clusters of varying shapes and sizes.
- Spectral Clustering: This algorithm uses eigenvalues of a similarity matrix to reduce the dimensionality of the data and identify clusters, making it suitable for capturing complex structures.

By applying these diverse clustering algorithms, we gained multiple perspectives on the dataset, improving the robustness and comprehensiveness of our analysis.

3.5 Evaluation of Clustering Algorithms

To evaluate the performance and effectiveness of the clustering algorithms, we utilized several metrics.

- Silhouette Score: This metric measures how similar each data point is to its own cluster relative to other clusters. A higher silhouette score indicates better-defined, well-separated clusters.
- Davies-Bouldin Index: This index evaluates the validity of the clusters by comparing the average similarity between clusters. A lower Davies- Bouldin Index suggests better clustering, with more distinct and well-separated clusters.
- Calinski-Harabasz Index: This metric assesses the clustering quality by comparing the dispersion between clusters with the dispersion within clusters. Higher values indicate more distinct and well-separated clusters.
- Within-Cluster Sum of Squares (WCSS):WCSS computes the sum of squared distances between data points and their respective cluster centroids. Lower values suggest more compact and cohesive clusters.
- Cohesion and Separation: These two metrics measure the compactness within clusters (cohesion) and the distance between clusters (separation). The ideal clustering configuration maximizes both cohesion and separation, indicating well-defined clusters with clear boundaries.

These metrics provided a comprehensive evaluation of the clustering algorithms, allowing us to select the most suitable algorithm for our dataset.

3.6 Additional Analysis

To deepen our understanding of global air pollution patterns, we conducted several additional analyses.

3.6.1 Region Mapping and Region-Wise Analysis

We mapped the clusters across different regions, performing a detailed analysis to identify patterns and trends in air quality across six continents: Asia, Africa, Europe, Australia, North America, and South America. This regional breakdown provided valuable insights into the variability of pollution levels and the underlying factors driving pollution in each region.

3.6.2 Capital City Analysis

We evaluated air quality in the capital cities of all countries worldwide. This analysis compared the pollution levels in capital cities with those in other cities and regions, highlighting specific urban pollution challenges faced by national capitals, which often experience higher pollution due to population density, industrialization, and vehicular emissions.

4. RESULT ANALYSIS

4.1 Optimal Numbers of Clusters

4.1.1 Elbow Method Results

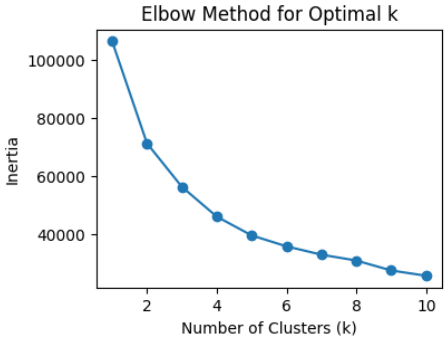


Fig 1: Elbow Method for Optimal Number of Clusters

The Elbow Method indicated that the optimal number of clusters is three. This conclusion was drawn by plotting the sum of squared distances between data points and their respective cluster centroids. As illustrated in Figure 1, the "elbow" point appears at three clusters, where the rate of decrease in variance sharply levels off. This suggests that adding more clusters does not result in significant improvements in clustering quality.

4.1.2 Silhouette Analysis Results

Silhouette Analysis reinforced the choice of three clusters. The average silhouette score was highest when the number of clusters was set to three, indicating that the clusters are both well-separated and cohesive. Figure 2 presents the silhouette scores for varying cluster counts, with the peak score observed for three clusters.

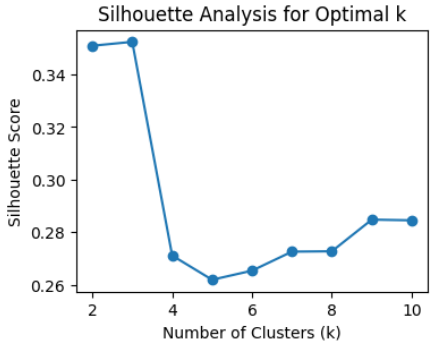


Fig 2: Silhouette Analysis for Optimal Number of Clusters

The results from both the Elbow Method and Silhouette Analysis consistently point to three clusters as the optimal configuration for this dataset, demonstrating clear distinctions in the pollution patterns across cities worldwide

4.2 Performance of Clustering Algorithms

To assess the performance of the clustering algorithms applied to our dataset, we employed a comprehensive set of evaluation metrics: Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index, Within-Cluster Sum of Squares (WCSS), Cohesion, and Separation. These metrics provide insight into the quality of the clusters formed, with each highlighting different aspects of the clustering process, such as cohesion, separation, and the overall cluster validity. The results for each algorithm are summarized in Table 1.

Table 1. Performance Metrics of Clustering Algorithms

Algorithm	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index	WCSS	Cohesion	Separation
K-Means	0.271155	1.212684	9352.368310	45935.146133	28674.504398	3.192677
Hierarchical	0.211006	1.383673	7507.943226	51735.863233	30267.741619	3.061031
DBSCAN	0.050884	1.781540	924.892078	47511.267771	26928.810227	4.464837
Gaussian Mixture Model	0.080737	2.465604	1442.023388	88489.810029	37216.996042	3.108874
Agglomerative	0.211006	1.383673	7507.943226	51735.863233	30267.741619	3.061031
Spectral	0.195279	1.555455	5058.885801	62159.420724	30997.799106	2.466136

The results indicate that K-Means consistently outperformed other algorithms in clustering the dataset. It achieved the highest Silhouette Score (0.271155), reflecting well-separated and cohesive clusters, and the lowest Davies-Bouldin Index (1.212684), which signifies distinct and nonoverlapping clusters. Additionally, K-Means had the best Calinski-Harabasz Index (9352.368310), showing strong between-cluster dispersion and low within-cluster variance. Its WCSS value was the lowest (45935.146133), highlighting compact clusters with minimal variance. K-Means also achieved excellent Cohesion (28674.504398) and a balanced Separation (3.192677), underscoring its effectiveness in clustering this dataset.

In comparison, Hierarchical and Agglomerative Clustering showed moderate performance. Both achieved a Silhouette Score of 0.211006 and a Davies-Bouldin Index of 1.383673, indicating reasonably cohesive clusters but with some overlap. Their Calinski-Harabasz Index (7507.943226) and Cohesion (30267.741619) were lower than KMeans, while their WCSS values (51735.863233) indicated higher within-cluster variance. These algorithms displayed decent separation (3.061031) but lagged behind K-Means in overall performance.

Spectral Clustering also demonstrated moderate results, with a Silhouette Score of 0.195279 and a Davies-Bouldin Index of 1.555455. Its Calinski-Harabasz Index (5058.885801) and WCSS (62159.420724) were weaker, indicating less effective clustering. The Cohesion value (30997.799106) and Separation (2.466136) were also inferior to K-Means, emphasizing its limitations for this dataset.

DBSCAN and Gaussian Mixture Models (GMM) struggled to define and separate clusters effectively. DBSCAN achieved the lowest Silhouette Score (0.050884) and a high Davies- Bouldin Index (1.781540), indicating significant overlap among clusters. Its Calinski-Harabasz Index (924.892078) and

Cohesion (26928.810227) were the lowest, though it recorded a high Separation value (4.464837). GMM had the lowest Silhouette Score among model-based approaches (0.080737) and the highest Davies-Bouldin Index (2.465604), reflecting poorly defined and highly overlapping clusters. Its Calinski-Harabasz Index (1442.023388) and WCSS (88489.810029) highlighted weak performance in cluster dispersion and compactness.

In conclusion, K-Means emerged as the most effective algorithm, forming cohesive, distinct, and compact clusters. Hierarchical, Agglomerative, and Spectral Clustering performed moderately well but did not match the quality of clusters formed by K-Means. DBSCAN and GMM were less effective for this dataset, showing significant limitations in clustering performance. These findings strongly recommend K-Means as the most suitable algorithm for clustering this dataset.

4.3 Cluster Analysis and Interpretation

The clustering analysis revealed distinct patterns in urban air pollution, categorized into three clusters: High Pollution, Low Pollution, and Ozone- Dominated Pollution. The interpretation of the clusters is based on the average values of AQI, CO, Ozone, NO2, and PM2.5 within each cluster as shown in Figure 3.

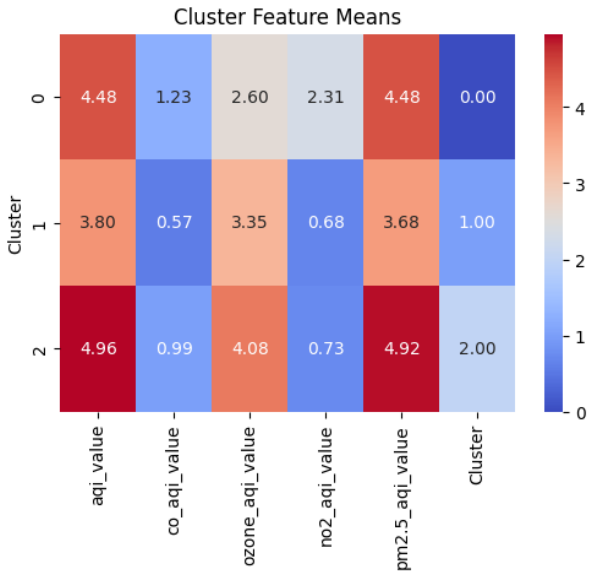


Fig 3: Cluster Feature Means

- Cluster 0 (High Pollution): This cluster is characterized by the highest average values for key pollutants, indicating severe air quality issues. The average values for AQI, CO, Ozone, NO2, and PM2.5 in this cluster are 4.48, 1.23, 2.60, 2.31, and 4.48, respectively. Regions represented in this cluster face significant challenges due to elevated levels of particulate matter and nitrogen dioxide.
- Cluster 1 (Low Pollution): Cities in this cluster exhibit the lowest pollutant levels, reflecting comparatively better air quality. The average values are 3.80 for AQI, 0.57 for CO, 3.35 for Ozone, 0.68 for NO2, and 3.68 for PM2.5. These cities are predominantly located in regions with effective air quality management practices or naturally favorable conditions.
- Cluster 2 (Ozone-Dominated Pollution): This cluster displays a unique profile with high Ozone levels dominating the pollution pattern. The average values are 4.96 for AQI, 0.99 for CO, 4.08 for Ozone, 0.73 for NO2, and 4.92 for PM2.5. This indicates that while other pollutants are relatively moderate, Ozone significantly influences the air quality in these cities.

The distribution of cities across the clusters further highlights the global air pollution disparities.

- Low Pollution: 13,871 cities (largest cluster, indicating a majority of cities experience relatively low pollution levels).
- High Pollution: 3,752 cities (regions with severe air quality challenges).
- Ozone-Dominated Pollution: 3,671 cities (locations where Ozone variability is the primary concern).

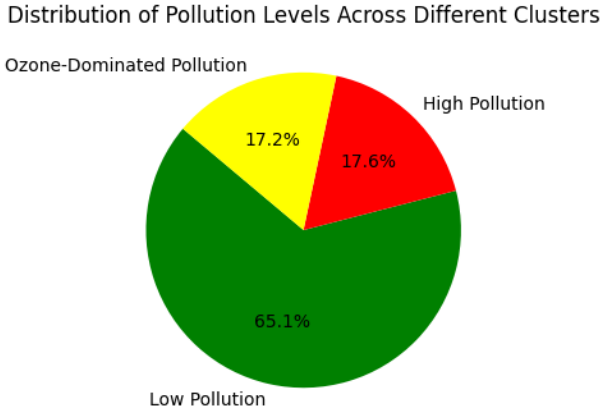


Fig 4: Distribution of Pollution Levels Across Different Clusters

A subsequent pie chart as shown in Figure 4 will visually depict the proportion of cities in each cluster, emphasizing the prevalence of low pollution but also underscoring the critical issues in high pollution and ozone-dominated regions. This analysis provides actionable insights into the geographical distribution and dominant pollution characteristics, aiding in targeted air quality management strategies.

4.4 Cluster Analysis by Continents

To provide a clearer understanding of the geographical distribution of the dataset, Figure 5 presents the percentage of unique countries from each continent. The dataset includes a total of 157 unique countries, with Africa contributing the highest proportion at 51 countries (32.5%), followed by Europe with 39 countries (24.8%), and Asia with 31 countries (19.7%). North America includes 18 countries (11.5%), South America contributes 12 countries (7.6%), and Australia has 6 countries (3.8%).

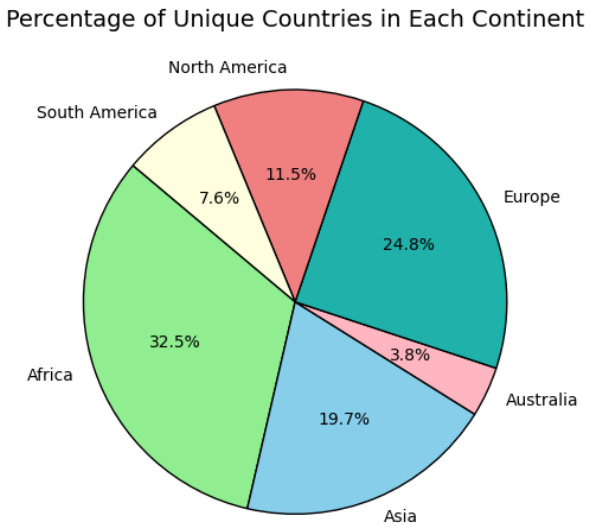


Fig 5: Distribution of Unique Countries by Continent

This distribution highlights the global scope of the study, with substantial representation from Africa, Europe, and Asia. The pie chart offers a visual representation of how each continent is represented in the dataset, providing context for the subsequent analysis of urban air pollution patterns across these diverse regions.

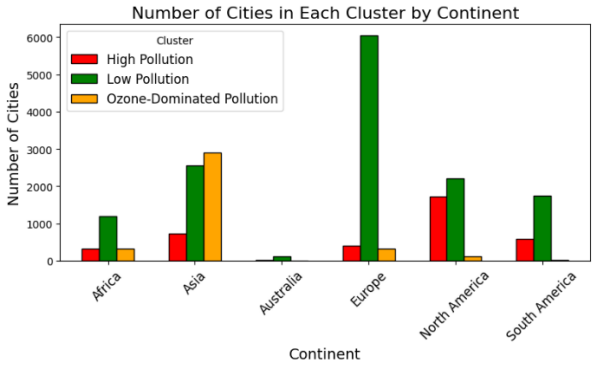


Fig 6: Distribution of Pollution Levels Across Different Continents

Figure 6 presents a bar chart illustrating the distribution of cities in each cluster across six continents, providing insights into the global air quality patterns. Africa has a total of 1,829 cities in the dataset, with 324 cities (17.71%) classified as High Pollution, 1193 cities (65.23%) falling under the Low Pollution cluster, and 312 cities (17.06%) affected by Ozone-Dominated Pollution. These findings indicate regional variability, with substantial air quality challenges in certain areas.

Asia, with 6,196 cities, has 726 cities (11.72%) in the High Pollution cluster, 2560 cities (41.32%) in the Low Pollution cluster, and the largest number of cities (2910, 46.97%) in the Ozone- Dominated Pollution cluster. This highlights Asia's unique air quality issues, particularly its susceptibility to ozone pollution in urban regions.

Australia's dataset includes 125 cities, of which the vast majority (119 cities, 95.2%) are classified as Low Pollution. Only five cities (4.0%) fall into the High Pollution category, and a single city (0.8%) is affected by Ozone-Dominated Pollution, reflecting one of the best air quality profiles globally.

Europe, with 6,748 cities, has 392 cities (5.81%) in the High Pollution cluster, 6046 cities (89.6%) in the Low Pollution cluster, and 310 cities (4.59%) in the Ozone-Dominated Pollution cluster. This distribution showcases Europe's success in maintaining better air quality, with most cities exhibiting low pollution levels.

North America has 4,058 cities in total, with 1731 cities (42.66%) classified as High Pollution, the highest proportion globally. The Low Pollution cluster contains 2209 cities (54.44%), while 118 cities (2.91%) fall into the Ozone-Dominated Pollution category, indicating a mixed urban air quality profile.

South America's 2,338 cities display a majority (1744 cities, 74.59%) in the Low Pollution cluster, with 574 cities (24.55%) in the High Pollution cluster, and 20 cities (0.86%) affected by Ozone-Dominated Pollution. This balance highlights regional variations in air quality challenges.

These findings provide a comprehensive overview of global air pollution patterns, supporting region-specific strategies to address urban region-specific pollution challenges effectively.

4.5 Cluster Distribution by Cities

In this section, we analyze the distribution of cities across different clusters within each continent based on their pollution levels. The data highlights the varying extent of high pollution, low pollution, and ozone-dominated pollution in cities from different regions. Six bar charts are included for each continent, showing the number of cities in each cluster for all countries represented in the dataset. The clusters represent different levels of air pollution, as determined through the analysis of pollutants such as PM2.5, CO, NO2, and Ozone.

For Africa, as shown in Figure 7, the distribution of cities reveals that the majority of cities are classified under Low Pollution, with countries like South Africa (116 cities) and Nigeria (62 cities) showing high pollution levels. In contrast, only a few cities in Democratic Republic of the Congo (27 cities) and Egypt (31 cities) fall under the Ozone-Dominated Pollution cluster. This pattern suggests that while low pollution is prevalent across much of Africa, high pollution is particularly concentrated in certain countries, especially South Africa and Nigeria, contributing to air quality concerns in these regions

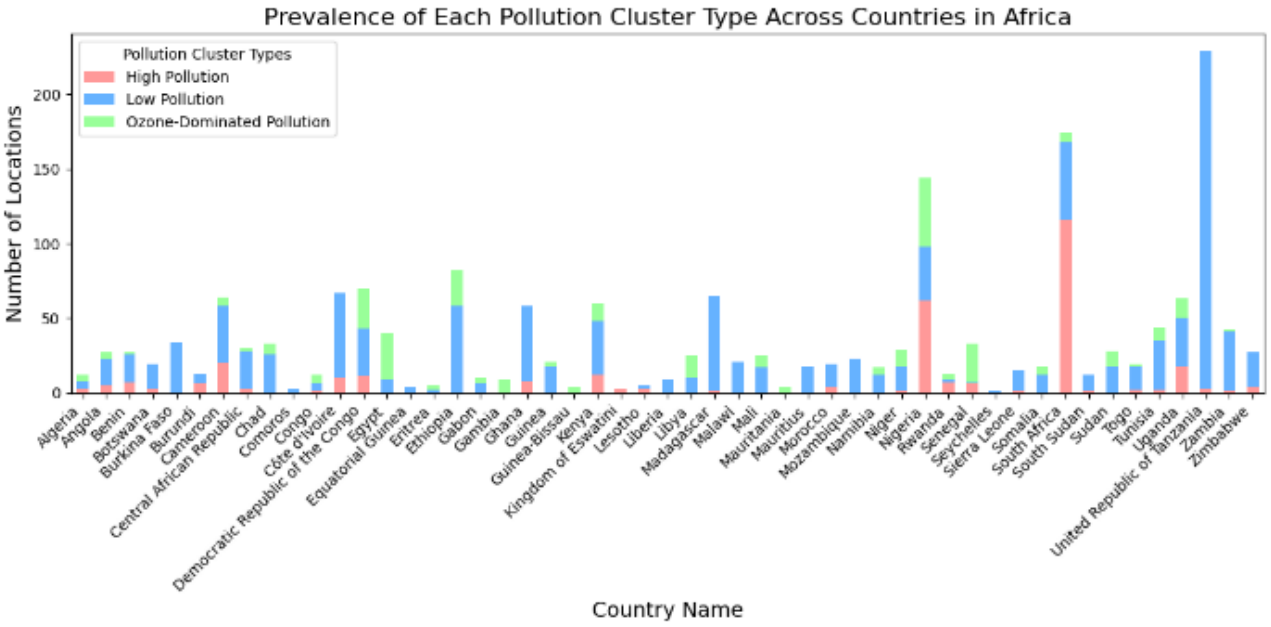


Fig 7: Cluster Distribution by Cities Across Africa

Figure 8 shows the situation differs significantly in Asia. A substantial number of cities in India (1687 cities), China (504 cities), and Bangladesh (46 cities) are in the Ozone- Dominated Pollution cluster. Countries such as China and India also have considerable numbers of cities in the High Pollution cluster, with China leading with 168 high pollution cities. The spread of pollution in Asia highlights the severity of air quality issues, particularly related to ozone levels, especially in urban centers.

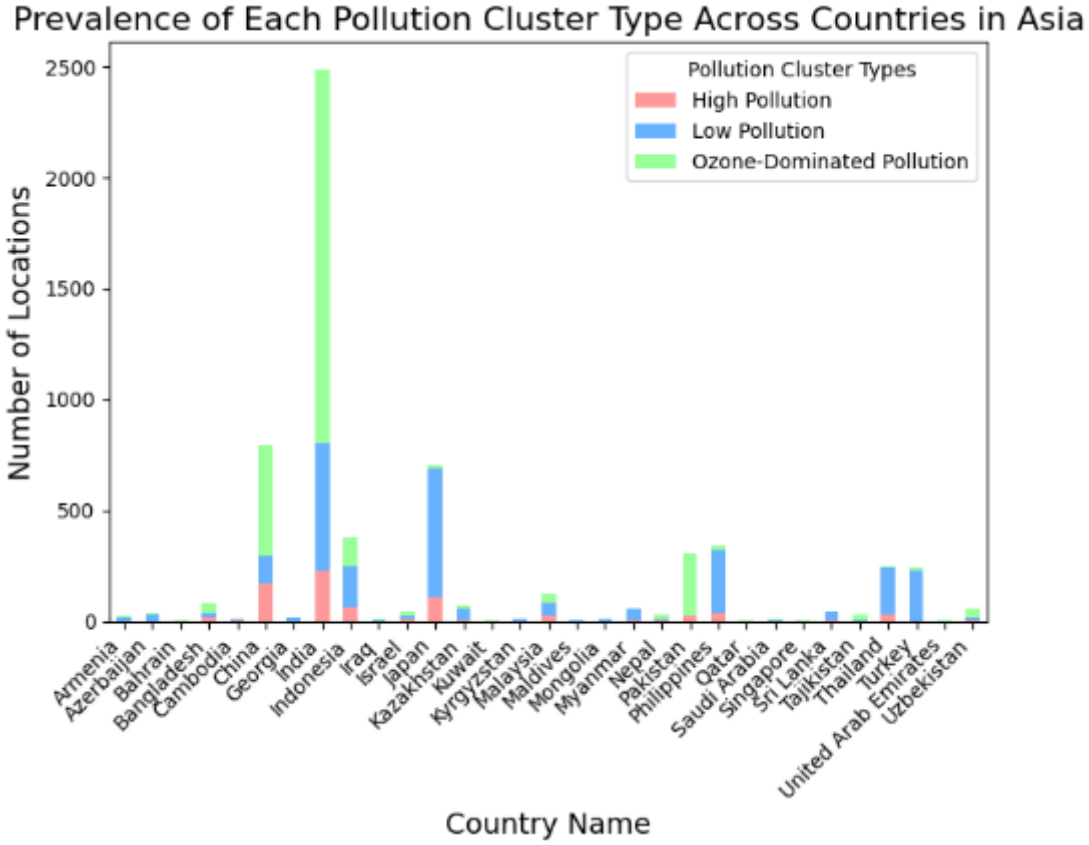
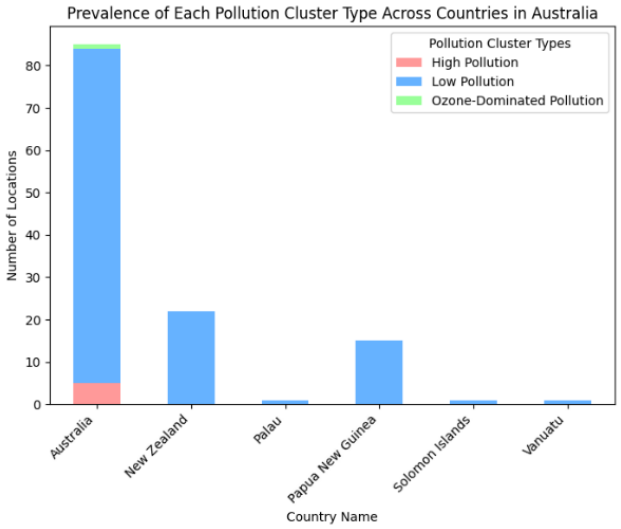


Fig 8: Cluster Distribution by Cities Across Asia



Australia, with countries like Australia and New Zealand, shows a distinct profile in Figure 9. The majority of cities fall under the Low Pollution category, with only a small number in the Ozone-Dominated Pollution cluster. This suggests that air quality is relatively good in this continent compared to others, with only a minimal number of cities facing significant pollution challenges.

The European continent also has a dominant presence of Low Pollution cities, especially in countries like Germany, France, and the United Kingdom, with large urban populations exhibiting low pollution levels as shown in Figure 10. However, countries such as Italy (127 cities), Spain (15 cities), and Poland (35 cities) show significant numbers of cities in the Ozone-Dominated Pollution cluster. The prevalence of low pollution cities in Europe can be attributed to the robust environmental policies and regulations in place across many European countries, which have led to improved air quality in urban areas.

Fig 9: Cluster Distribution by Cities Across Australia

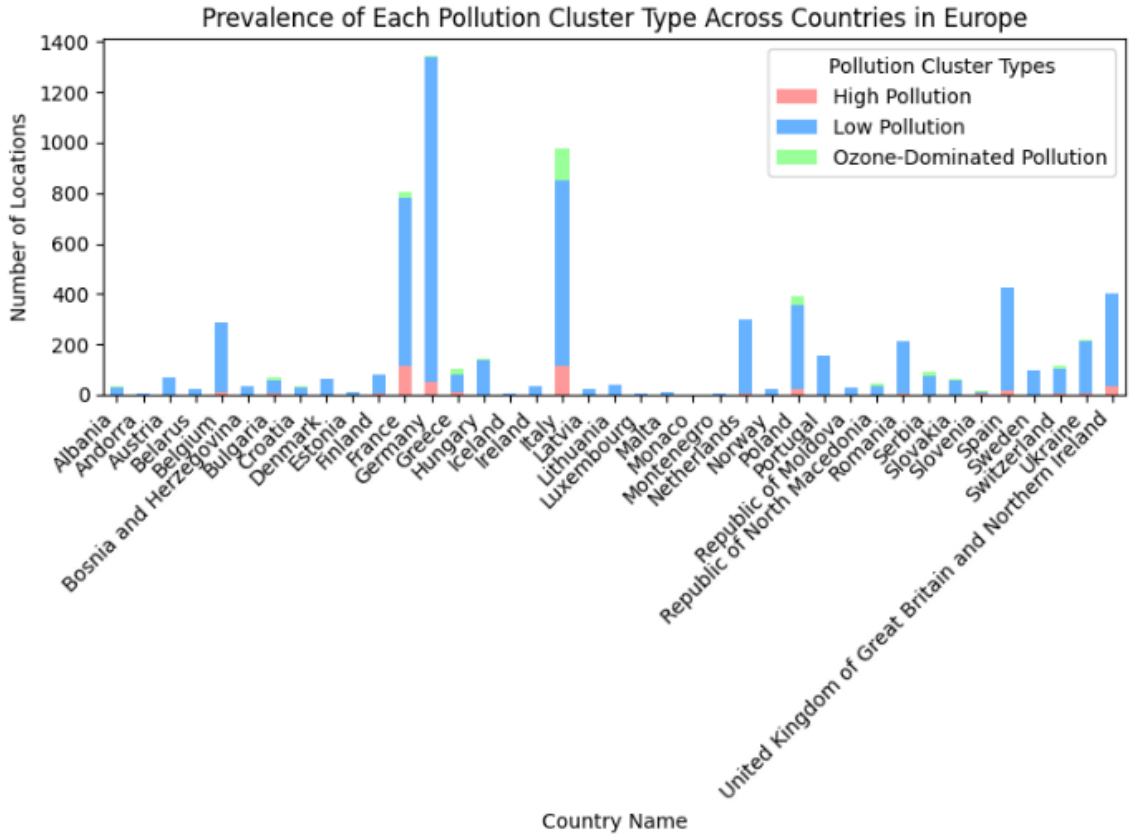


Fig 10: Cluster Distribution by Cities Across Europe

North America is home to a large number of High Pollution cities, with the United States (1212 cities) and Mexico (273 cities) contributing the highest numbers. Despite this, the Low Pollution cluster still contains a significant portion of cities, with countries like Canada and Costa Rica showing relatively better air quality in Figure 11. The Ozone-Dominated Pollution cluster in North America, while present, is notably smaller in comparison to the other two clusters.

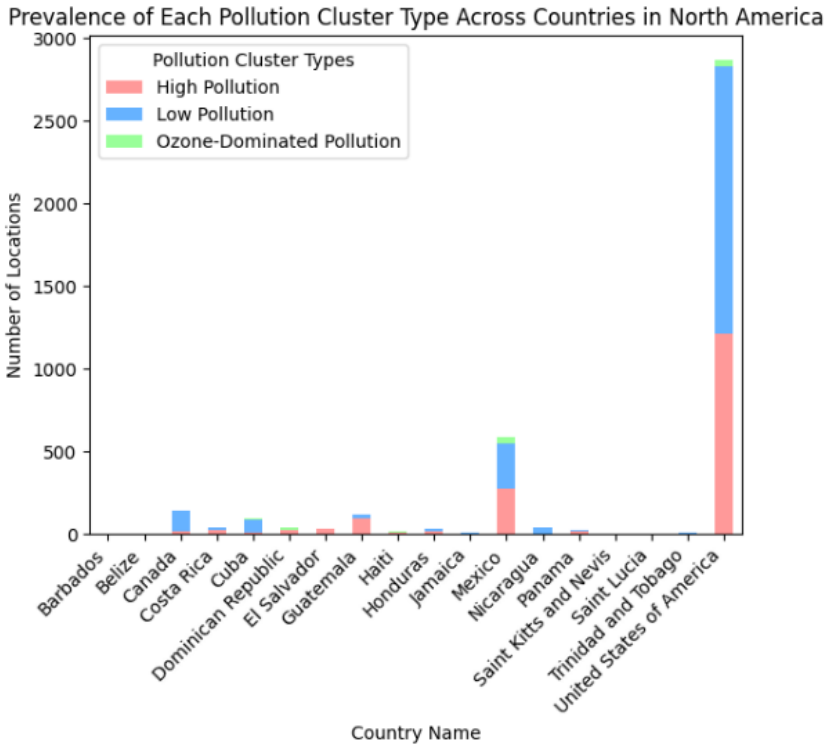


Fig 11: Cluster Distribution by Cities Across North America

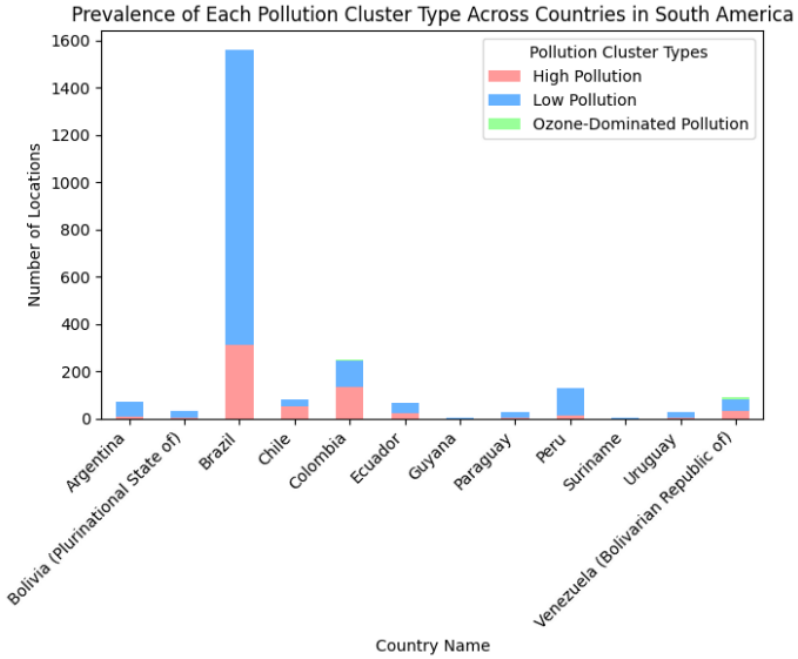


Fig 12: Cluster Distribution by Cities Across South America

In South America, the Low Pollution cluster is also the most prevalent, with countries like Brazil (1249 cities) and Argentina (61 cities) contributing to this trend. However, Brazil (3 cities) and Colombia (8 cities) have a few cities in the Ozone-Dominated Pollution cluster as shown in Figure 12, highlighting regional challenges related to specific pollutants. The relatively low number of cities in the High Pollution cluster in South America suggests that the continent does not face the same level of air pollution challenges as others.

Overall, the cluster distribution across continents reveals both global and regional disparities in air quality. Low Pollution remains the most dominant cluster across all continents, but Asia, Africa, and North America face notable challenges with High Pollution, while Asia is particularly impacted by Ozone-Dominated Pollution. These findings underscore the need for targeted policy interventions that consider the regional variations in pollution levels, with a focus on reducing high and ozone-related pollution in urban centers across continents.

4.6 Region Mapping and Findings

In this section, we present the regional mapping of air pollution data across six continents, categorizing cities based on pollution levels. The regions in six continents have been defined based on geographical proximity and political boundaries, and the clustering results are summarized.

4.6.1 Asia

The Central Asia region, characterized by countries such as Kazakhstan and Kyrgyzstan, has a significant number of cities in the High Pollution and Ozone-Dominated Pollution clusters, with 9 and 83 cities in each category, respectively. East Asia, which includes countries like China, Japan, and Korea, has a large number of cities in all three clusters, with 278 cities in the High Pollution cluster, 714 in Low Pollution, and 515 in Ozone-Dominated Pollution. The Middle East region, which encompasses countries like Saudi Arabia, Iran, and the UAE, shows a notable concentration of cities in the Low Pollution category (245), with a smaller number of cities in the High Pollution (11) and Ozone-Dominated Pollution (68) clusters. South Asia, which includes densely populated countries like India, Bangladesh, and Pakistan, displays a high concentration of cities in the Ozone-Dominated Pollution category, with 276 cities in High Pollution, 642 in Low Pollution, and a substantial 2037 in Ozone-Dominated Pollution. The South Caucasus region, including Armenia, Azerbaijan, and Georgia, shows fewer cities in the Ozone-Dominated Pollution cluster (10), while the Southeast Asia region, including countries like Thailand, Indonesia, and Malaysia, has 152 cities in the High Pollution cluster, 810 in Low Pollution, and 197 in Ozone-Dominated Pollution. Table 2 summarizes the pollution levels across different regions of Asia.

Table 2. Cluster Distribution by Region in Asia

Region	High Pollution	Low Pollution	Ozone-Dominated Pollution
Central Asia	9	79	83
East Asia	278	714	515
Middle East	11	245	68
South Asia	276	642	2037
South Caucasus	0	70	10
Southeast Asia	152	810	197

This regional mapping provides valuable insights into the geographical distribution of air pollution across Asia, highlighting regions that require more focused air quality management interventions.

Figure 13 shows the distribution of High Pollution, Low Pollution, and Ozone-Dominated Pollution across different regions and countries in Asia. The data covers regions such as Central Asia, East Asia, South Asia, Southeast Asia, South Caucasus, and the Middle East. The pollution levels are classified into three categories:

- High Pollution: Indicates regions and countries with significant pollution levels.
- Low Pollution: Represents areas with relatively low pollution.
- Ozone-Dominated Pollution: Highlights regions where ozone is a major contributor to air pollution.

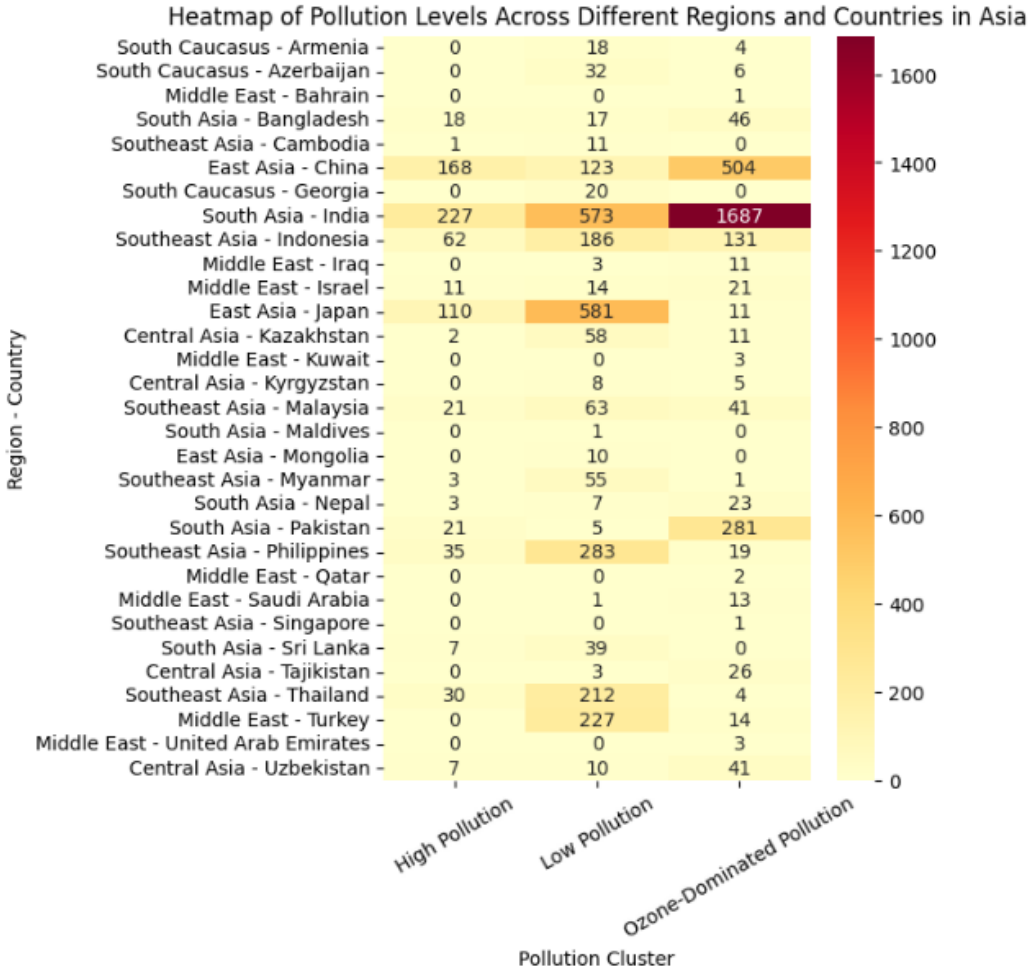


Fig. 13 Heatmap of Pollution Level Distribution Across Different Regions and Countries in Asia

South Asia emerges as a major region with high pollution levels, particularly India, which stands out with a significantly high level of high pollution (227) and ozone-dominated pollution (1687). Other South Asian countries such as Bangladesh and Pakistan also show notable pollution levels. East Asia has a varied distribution of pollution, with China exhibiting high levels of both high pollution (168) and ozone-dominated pollution (504). Japan also shows a considerable level of ozone-dominated pollution (11) despite a relatively lower level of high pollution (110).

Southeast Asia has a mix of high and low pollution levels. For instance, Thailand (30 for high pollution, 4 for ozone-dominated pollution) and Indonesia (62 for high pollution, 131 for ozonedominated pollution) display a significant level of both high and ozone-dominated pollution. On the other hand, countries like Singapore and Bahrain have minimal pollution. Middle Eastern countries such as Saudi Arabia, Qatar, and United Arab Emirates report low to zero pollution levels across all categories, indicating a better air quality situation compared to other regions.

Central Asia countries like Kazakhstan, Kyrgyzstan, and Uzbekistan show varying pollution levels, with some countries like Kazakhstan exhibiting a relatively higher level of high pollution (2) and ozone-dominated pollution (11), while others show minimal pollution levels. South Caucasus countries, including Armenia and Georgia, have relatively low levels of high pollution, but still experience some ozone-dominated pollution, particularly Azerbaijan (6 for ozone-dominated pollution).

The heatmap visualizes these findings, helping identify regions with significant pollution and those in need of further monitoring and management. Figure: Heatmap of Pollution Level Distribution Across Different Regions and Countries in Asia.

These observations suggest that South Asia, particularly India, and East Asia, notably China, require more attention in terms of pollution control and management. Countries in the Middle East and South Caucasus, however, exhibit relatively lower pollution levels and may be on a more sustainable path in terms of air quality.

4.6.2 Africa

This section presents a detailed analysis of pollution distribution across Africa. Table 3 provides a region-wise summary of pollution levels across Africa. Southern Africa exhibits the highest levels of High Pollution (125), with South Africa as a significant contributor (116). West Africa follows with a notable level of high pollution (97), largely driven by Nigeria (62). East Africa has the highest count of Low Pollution (568), indicating better air quality across most of the region. However, some countries like Uganda and Kenya show moderate levels of High Pollution (18 and 12, respectively).

Central Africa and North Africa display moderate pollution levels, with North African countries like Egypt and Libya showing higher contributions to Ozone-Dominated Pollution (31 and 15, respectively).

Table 3. Cluster Distribution by Region in Africa

Region	High Pollution	Low Pollution	Ozone-Dominated Pollution
Central Africa	40	154	56
East Africa	52	568	63
North Africa	10	101	69
Southern Africa	125	82	11
West Africa	97	288	113

The data is categorized into High Pollution, Low Pollution, and Ozone-Dominated Pollution, providing insights into pollution trends across different regions and countries. The information is summarized in Figure 14 to visualize these patterns effectively.

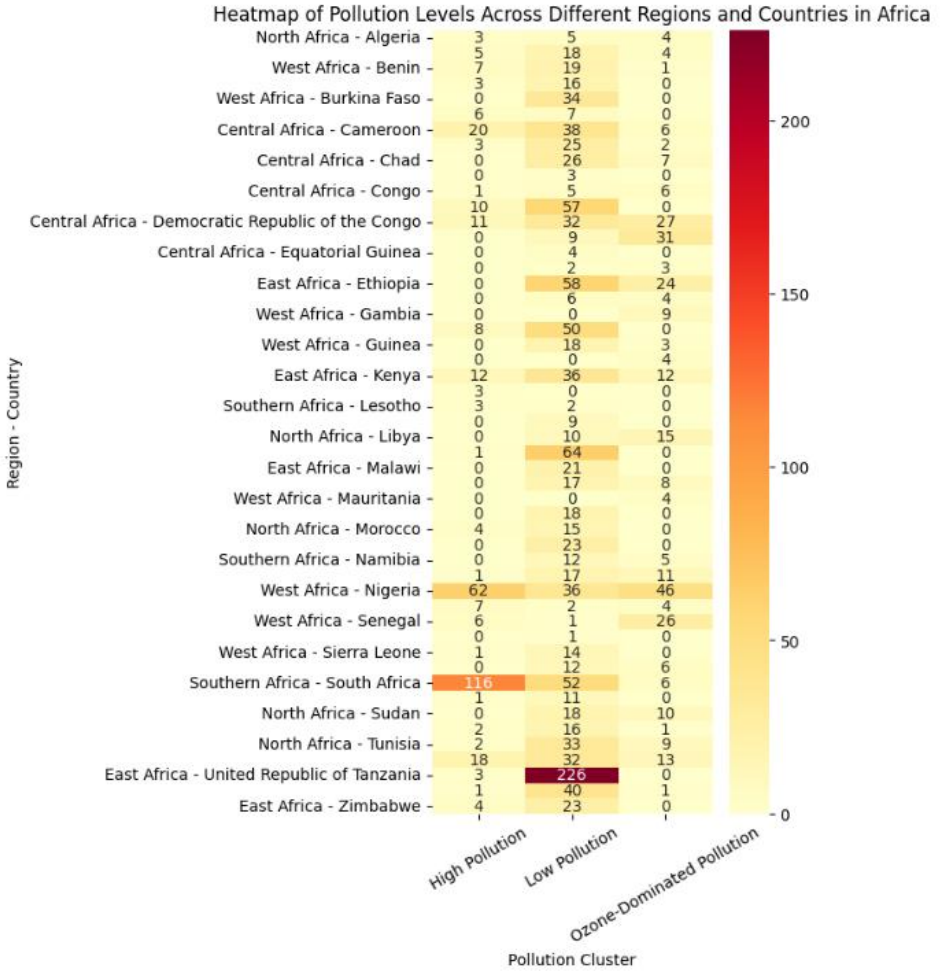


Fig 14: Heatmap of Pollution Level Distribution Across Different Regions and Countries in Africa

South Africa has the highest level of High Pollution in Africa (116), reflecting the industrialization and urbanization challenges in the region.

Nigeria stands out in West Africa, with significant contributions to both High Pollution (62) and Ozone-Dominated Pollution (46). Countries like Egypt and Libya in North Africa show higher levels of Ozone-Dominated Pollution, indicating a specific type of pollution concern in the region.

Many countries in East Africa and Central Africa report low levels of High Pollution, with nations like Eritrea, Madagascar, and Mozambique showing minimal pollution across all categories. Small island nations such as Mauritius, Comoros, and Seychelles exhibit low pollution levels, suggesting a better environmental state in these regions.

Regions such as North Africa and West Africa have noticeable ozone-dominated pollution levels, with Egypt (31) and Nigeria (46) standing out as significant contributors. Southern Africa has minimal ozone-dominated pollution (11), possibly reflecting the specific industrial and urbanization patterns in the region.

These findings emphasize the need for targeted pollution control measures in Southern Africa, particularly in South Africa, and West Africa, with a focus on Nigeria. The relatively cleaner regions, such as East Africa and Central Africa, offer examples of potentially effective environmental policies and

practices. The heatmap further provides a visual guide to policymakers for prioritizing regions and countries for pollution management efforts.

4.6.3 Europe

Table 4 presents the pollution levels across European regions, categorized into High Pollution, Low Pollution, and Ozone-Dominated Pollution. The heatmap provides additional insights into country-specific distributions.

Table 4. Cluster Distribution by Region in Europe

Region	High Pollution	Low Pollution	Ozone-Dominated Pollution
Eastern Europe	27	1043	68
Northern Europe	5	360	0
Southern Europe	142	1588	199
Western Europe	218	3055	43

Table 4 highlights the pollution distribution in European regions. Western Europe shows the highest count of both high and low pollution cases, whereas Southern Europe leads in ozone-dominated pollution. Conversely, Northern Europe has the lowest levels of high pollution. These patterns are further detailed in the heatmap presented in Figure 15. The heatmap showcases pollution level distributions across European countries, categorized by High Pollution, Low Pollution, and Ozone-Dominated Pollution.

Western Europe dominates with both the highest low pollution counts, notably in countries like France, Germany, and the United Kingdom. It also exhibits notable levels of high pollution, particularly in France (111) and Germany (53). Southern Europe has significant ozone-dominated pollution, with Italy (127) and Greece (26) contributing prominently. This region also records considerable high pollution levels, notably in Italy (112) and Spain (15).

Eastern Europe reflects moderate levels of high pollution, with Poland (21) standing out. The region also has relatively high ozone-dominated pollution counts, such as Poland (35). Northern Europe demonstrates the lowest pollution levels overall, with minimal contributions to high pollution. Countries like Finland, Sweden, and Ireland primarily report low pollution cases, emphasizing the region's cleaner air quality.

The pollution patterns in Europe highlight the stark contrasts between its regions. Western and Southern Europe grapple with significant high and ozone-dominated pollution levels, reflecting industrial activities and urbanization. In contrast, Northern Europe exemplifies successful environmental management with its predominantly low pollution levels. These findings underscore the need for tailored strategies to address region-specific pollution challenges and promote sustainable practices across Europe.

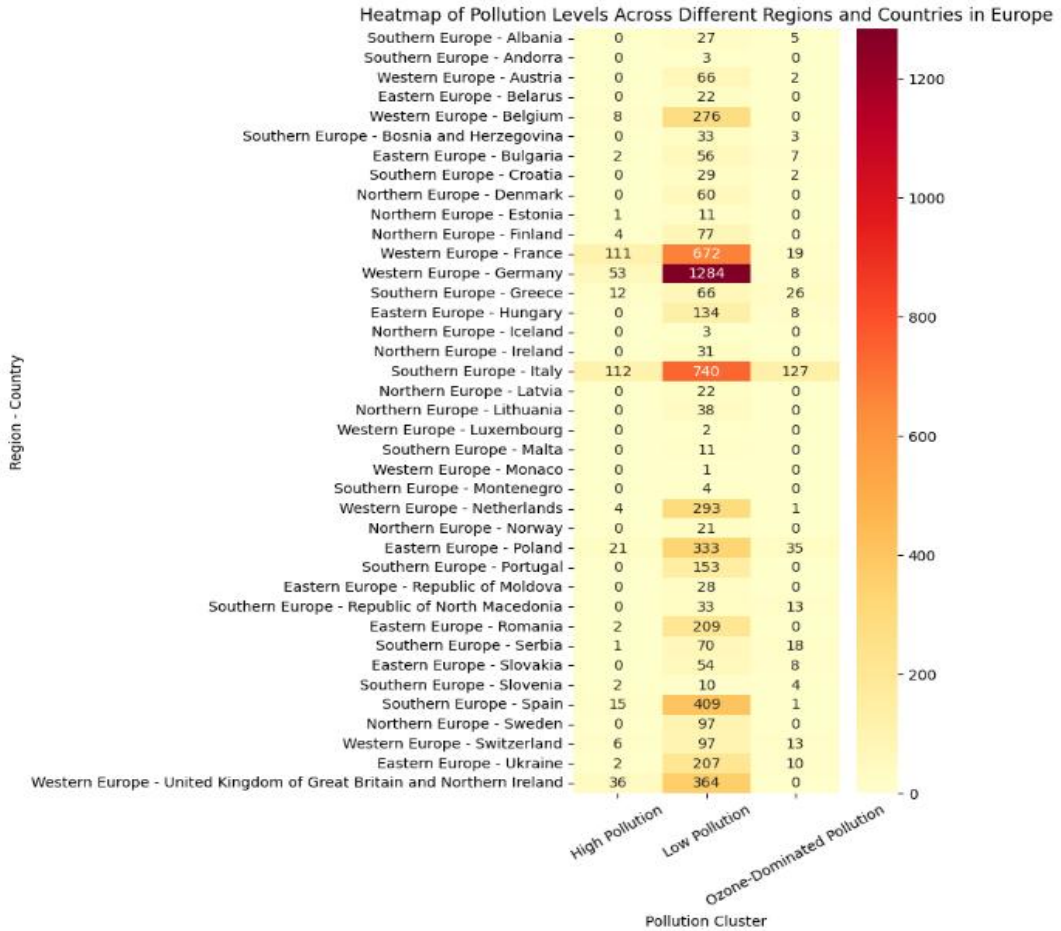


Fig 15: Heatmap of Pollution Level Distribution Across Different Regions and Countries in Europe

4.6.4 Australia

The Oceania region mapping covers six countries: Australia, New Zealand, Palau, Papua New Guinea, Solomon Islands, and Vanuatu. This diverse group of nations includes large developed countries and smaller island states, reflecting a mix of urbanized and rural areas.

The heatmap analysis in Figure 16 reveals that Australia is the dominant contributor to pollution within Oceania, with the highest count of high pollution cases (5) and the only recorded instance of ozone-dominated pollution (1). In contrast, the remaining countries predominantly report low pollution cases, with New Zealand showing a notable count of low pollution (22), followed by Papua New Guinea (15). Smaller island nations like Palau, Solomon Islands, and Vanuatu report only isolated instances of low pollution, emphasizing their generally pristine air quality.

The Oceania region demonstrates overall low pollution levels, with Australia being an outlier due to its higher pollution contributions. This analysis underscores the need for localized interventions in Australia to address pollution concerns, while efforts should also be made to preserve the exceptional air quality in the smaller island nations.

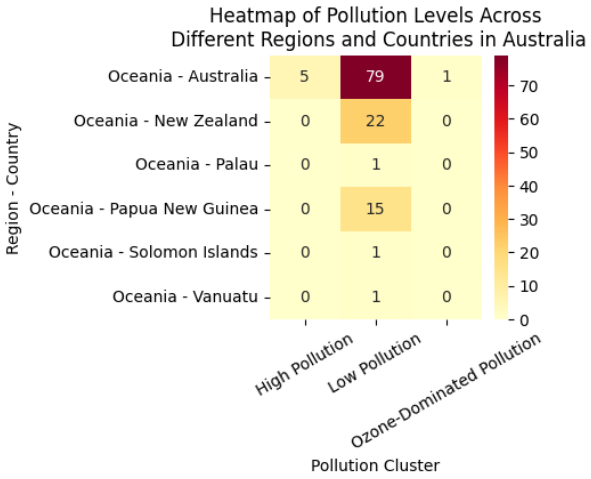


Fig 16: Heatmap of Pollution Level Distribution Across Different Regions and Countries in Australia

4.6.5 North America

Table 5 highlights the pollution distribution across three major regions in North America: the Caribbean, Central America, and North America. It shows the counts for three pollution types: high pollution, low pollution, and ozonedominated pollution. North America leads in both high and low pollution levels, reflecting significant urban and industrial contributions. The Caribbean region exhibits notable ozonedominated pollution, whereas Central America shows moderate high pollution with relatively lower ozone-dominated cases.

Table 5. Cluster Distribution by Region in North America

Region	High Pollution	Low Pollution	Ozone-Dominated Pollution
Caribbean	50	88	33
Central America	179	106	2
North America	1502	2015	83

The heatmap in Figure 17 provides a detailed breakdown of pollution levels across countries within North America.

In Caribbean, Cuba (11) and the Dominican Republic (23) stand out for high pollution levels, while ozone-dominated pollution is significant in Cuba (10) and the Dominican Republic (12). Other nations like Haiti (10) contribute marginally to high pollution levels but exhibit notable ozone-dominated pollution (7).

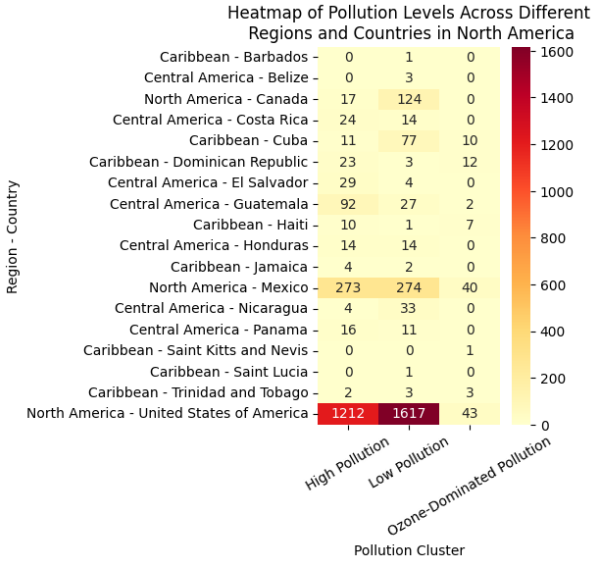


Fig 17: Heatmap of Pollution Level Distribution Across Different Regions and Countries in North America

In Central America, Guatemala leads high pollution levels with 92 cases, followed by El Salvador (29) and Costa Rica (24). The region records minimal ozone-dominated pollution, with Guatemala contributing marginally (2). Low pollution cases are distributed across countries, with Nicaragua (33) and Guatemala (27) showing the highest counts. In North America, The United States and Mexico dominate high pollution counts, with 1212 and 273, respectively. Canada has relatively lower high pollution (17). Ozone-dominated pollution is notable in the United States (43) and Mexico (40), reflecting industrial and urban contributions. Low pollution levels are highest in the United States (1617), followed by Mexico (274).

North America demonstrates significant diversity in pollution patterns. The United States dominates pollution metrics, with high levels across all categories. Mexico contributes notably to high and ozone-dominated pollution. The Caribbean region showcases moderate high pollution but a significant share of ozone-dominated pollution, particularly in Cuba and the Dominican Republic. Central America reflects a balanced distribution, with Guatemala leading high pollution counts.

4.6.6 South America

Table 6 presents the pollution distribution in South America across four main regions: the Andean Region, the Caribbean, Northern South America, and the Southern Cone. Southern Cone exhibits the highest counts of low pollution, reflecting cleaner air quality, with Brazil contributing significantly. Conversely, it also has the highest high pollution levels, indicating localized industrial or urban influences. The Andean Region shows moderate pollution levels across all categories, while Northern South America displays notable high and low

pollution. The Caribbean records minimal pollution metrics, maintaining relatively pristine air quality.

Table 6. Cluster Distribution by Region in South America

Region	High Pollution	Low Pollution	Ozone-Dominated Pollution
Andean Region	217	332	9
Caribbean	0	7	0
Northern South America	33	49	8
Southern Cone	324	1356	3

The heatmap in Figure 18 gives a detailed view of pollution levels in South American countries. In Southern Cone, Brazil dominates high pollution levels (310), followed by smaller contributions from Argentina (8) and Paraguay (5). The region leads in low pollution, with Brazil (1249) standing out, reflecting its vast geographic area and urban-industrial variations. In Andean Region, Colombia (135) and Chile (50) show significant high pollution levels. Peru (117) and Colombia (107) contribute notably to low pollution levels, while Colombia (8) and Chile (1) exhibit moderate ozone-dominated pollution.

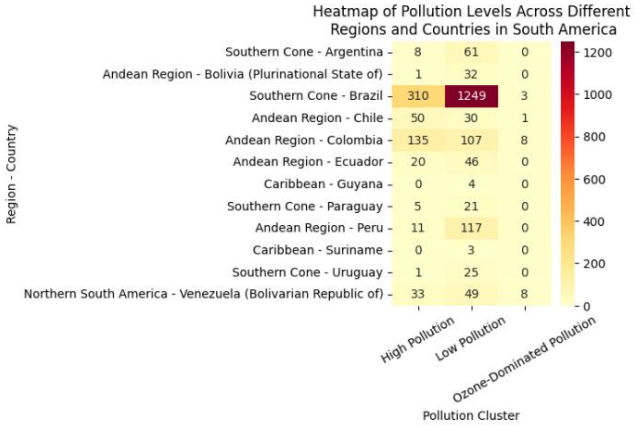


Fig 18: Heatmap of Pollution Level Distribution Across Different Regions and Countries in South America

In Northern South America, Venezuela records all high pollution cases (33) in this region, along with notable ozone-dominated pollution (8). It also reflects moderate low pollution levels (49). Caribbean, this region, represented by countries like Guyana and Suriname, shows minimal pollution cases, indicating relatively cleaner air quality. South America's pollution levels reveal contrasting regional patterns. The Southern Cone, dominated by Brazil, shows significant high and low pollution levels due to its large urban and industrial hubs. The Andean Region reflects diverse pollution types, with Colombia and Chile being the primary contributors. Northern South America and the Caribbean maintain relatively lower pollution metrics, emphasizing cleaner environments, though Venezuela shows some significant pollution levels. Overall, the region exhibits notable variability in air quality, influenced by geography and industrialization.

4.7 Region Mapping and Findings

4.7.1 Asia

The analysis of air pollution patterns across Asian capital cities reveals significant insights into the clustering of urban air quality. The cities are grouped into three distinct clusters based on pollution characteristics: High Pollution, Low Pollution, and Ozone-Dominated Pollution. Each cluster exhibits unique average feature values for air quality index (AQI), particulate matter (PM2.5), carbon monoxide (CO), nitrogen dioxide (NO2), and ozone, reflecting the varying pollution dynamics across the region.

The High Pollution cluster is characterized by elevated levels of PM2.5, CO, and NO2, indicative of cities facing severe air quality challenges. Figure 19 represents Islamabad and Tokyo belong to this cluster, with Islamabad displaying an AQI value of 5.19, substantially above the cluster average of 4.48. The city's most deviating feature is ozone, exceeding the average by 1.17 units. Similarly, Tokyo shows a notable deviation in ozone, highlighting its significant contribution to the city's pollution profile.

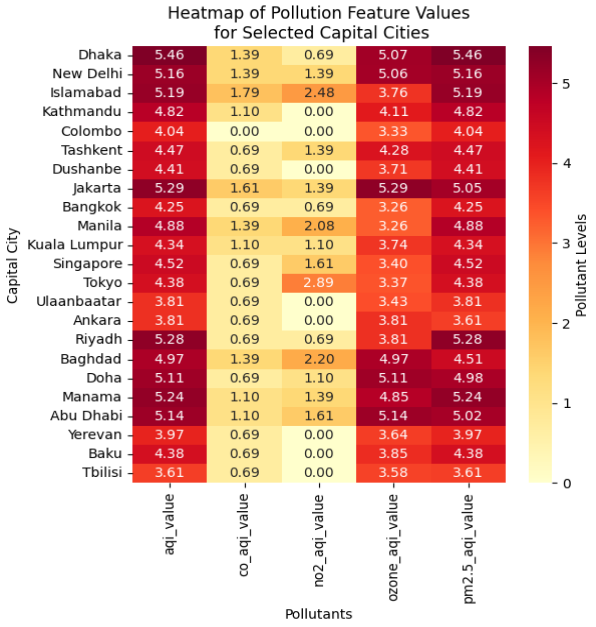


Fig 19: Heatmap of Air Quality Indicators for Asian Capital Cities

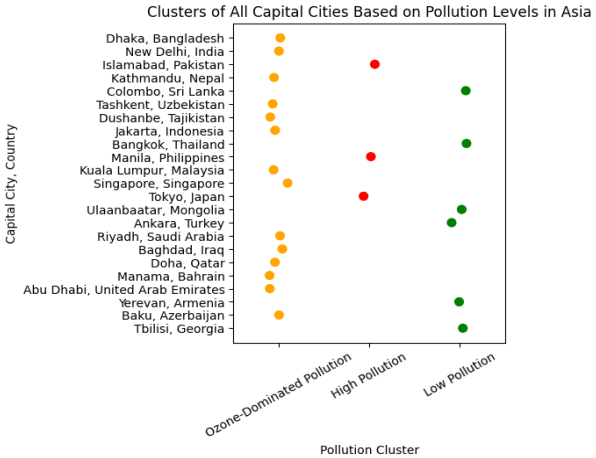


Fig 20: Cluster Plot of Asian Capital Cities

In contrast, cities in the Low Pollution cluster, such as Colombo, Bangkok, and Ulaanbaatar as shown in Figure 20, exhibit relatively lower pollutant concentrations. Colombo, for example, has an AQI of 4.04 and almost negligible NO2 levels, deviating by -0.68 units from the cluster average. Ulaanbaatar and Ankara share similar patterns, with NO2 being the most underrepresented feature, reflecting the lower industrial and vehicular emissions in these cities.

The Ozone-Dominated Pollution cluster includes cities like Dhaka, New Delhi, and Jakarta, where ozone and PM2.5 are the predominant pollutants. Dhaka's AQI is notably high at 5.46, with ozone deviating from the cluster average by 0.99 units. Jakarta, similarly, shows significant deviations in ozone, emphasizing its role as the primary pollutant. These deviations suggest strong contributions from photochemical reactions, which are influenced by sunlight and emissions of volatile organic compounds (VOCs) and NO2.

Notably, some cities lack sufficient data to be included in the clustering analysis, such as Mal′e, Bishkek, and Naypyidaw. Addressing these gaps is crucial for a comprehensive understanding of pollution patterns across Asia. Additionally, certain cities like Baghdad exhibit a mixed pollution profile, where NO2 exceeds the cluster average by 1.47 units, indicating localized industrial or traffic related emissions as significant contributors.

This clustering-based analysis provides a nuanced understanding of urban air pollution in Asia, highlighting regional differences and specific pollutant contributions. Such insights can inform targeted air quality management strategies, focusing on the predominant pollutants and their sources to mitigate health and environmental impacts in these urban centers.

4.7.2 Africa

The clustering analysis for African capital cities revealed three distinct pollution patterns: High Pollution, Low Pollution, and Ozone-Dominated Pollution. The clusters were analyzed based on AQI values for PM2.5, CO, NO2, and Ozone. Key observations for each cluster are summarized below:

Cities in High Pollution cluster exhibited elevated AQI values for PM2.5 and NO2. The average AQI for PM2.5 was approximately 4.48, while NO2 had an average AQI of 2.31. Figure 21 shows Algiers had the highest deviation in NO2 AQI, with a value of 4.25 compared to the cluster average of 2.31, highlighting NO2 as the major contributor to its overall pollution.

Low Pollution cluster represented cities with relatively lower AQI values across pollutants. The average AQI for PM2.5 was 3.68, while ozone levels were moderately low at 3.35. CO and NO2 levels were significantly lower, with averages of 0.57 and 0.68, respectively. For example, Tripoli showed a distinct pattern with an ozone AQI of 3.81, slightly above the cluster average of 3.35, making it a primary contributor.

Ozone-Dominated Pollution cluster was characterized by high ozone AQI levels averaging 4.08, along with a relatively high PM2.5 AQI of 4.92. NO2 and CO levels were low, averaging 0.73 and 0.99, respectively. For example, Nouakchott had a notable deviation in ozone levels, with an AQI of 3.14, slightly below the cluster average, but PM2.5 remained high at 5.12, indicating a significant impact of particulate matter in the city's pollution profile.

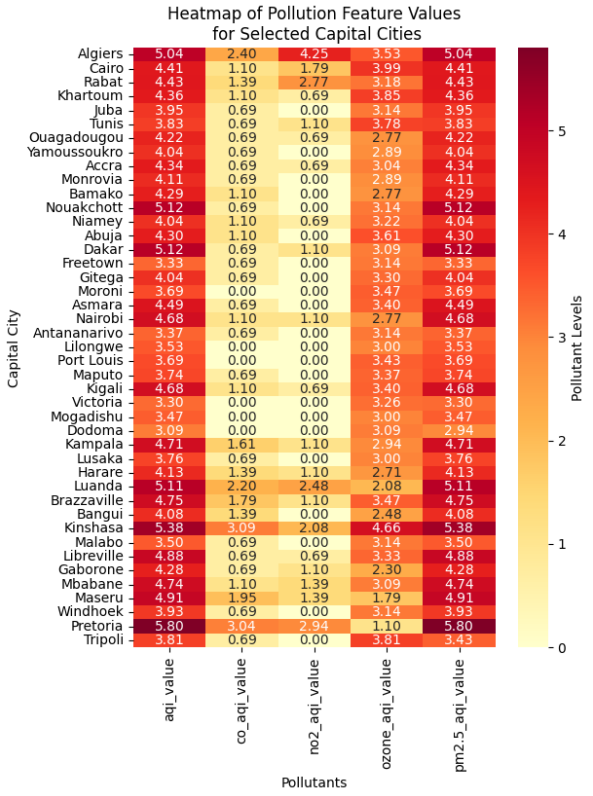


Fig 21: Heatmap of Air Quality Indicators for African Capital Cities

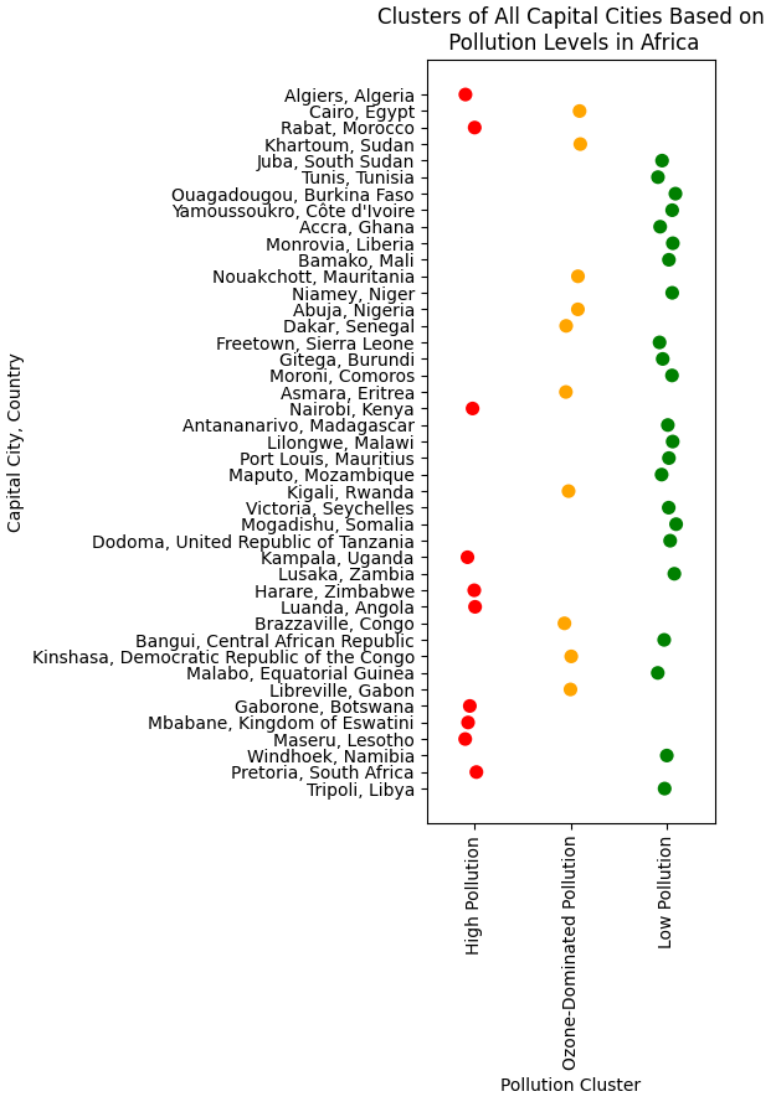


Fig 22: Cluster Plot of African Capital Cities

In Algiers (High Pollution) showed in Figure 22, NO2 contributed significantly to pollution, with a deviation of 1.94 from the cluster average. In Cairo (Ozone-Dominated Pollution), NO2 levels were slightly elevated with an AQI of 1.79, deviating by 1.06 from the cluster average. In Tripoli (Low Pollution), NO2 AQI was 0.00, far below the cluster average of 0.68, making it a distinctive feature. In Rabat (High Pollution), Ozone levels deviated positively by 0.58, indicating an emerging contribution to the pollution profile.

The results showed a diverse pollution profile across African capitals, with PM2.5 consistently contributing to elevated AQI levels in all clusters. Ozone pollution emerged as a significant factor in many cities, particularly in the Ozone-Dominated Pollution cluster. Meanwhile, CO and NO2 levels were relatively lower across clusters, with some exceptions where localized contributions caused significant deviations. This analysis provides a foundation for targeted interventions, highlighting cities and pollutants requiring immediate attention for air quality management.

4.7.3 Europe

The clustering analysis for European cities reveals three distinct clusters: High Pollution, Low Pollution, and Ozone-Dominated Pollution. The High Pollution cluster exhibits an average AQI value of 4.48 as shown in Figure 23, characterized by elevated contributions from PM2.5 (average AQI of 4.48) and significant NO2 levels (average AQI of 2.31). The Low Pollution cluster, with an average AQI value of 3.80, shows lower values for most pollutants, particularly CO (average AQI of 0.57) and NO2 (average AQI of 0.68). In contrast, the Ozone-Dominated Pollution cluster, with the highest average AQI value of 4.96, is primarily influenced by ozone levels (average AQI of 4.08) and PM2.5 (average AQI of 4.92).

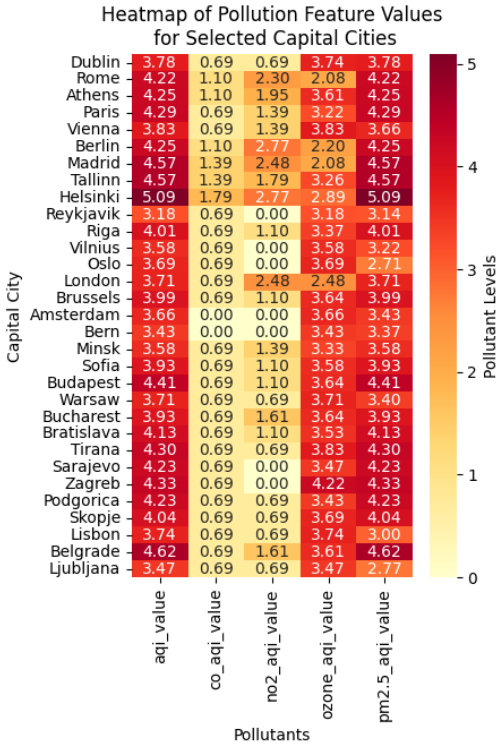


Fig 23: Heatmap of Air Quality Indicators for European Capital Cities

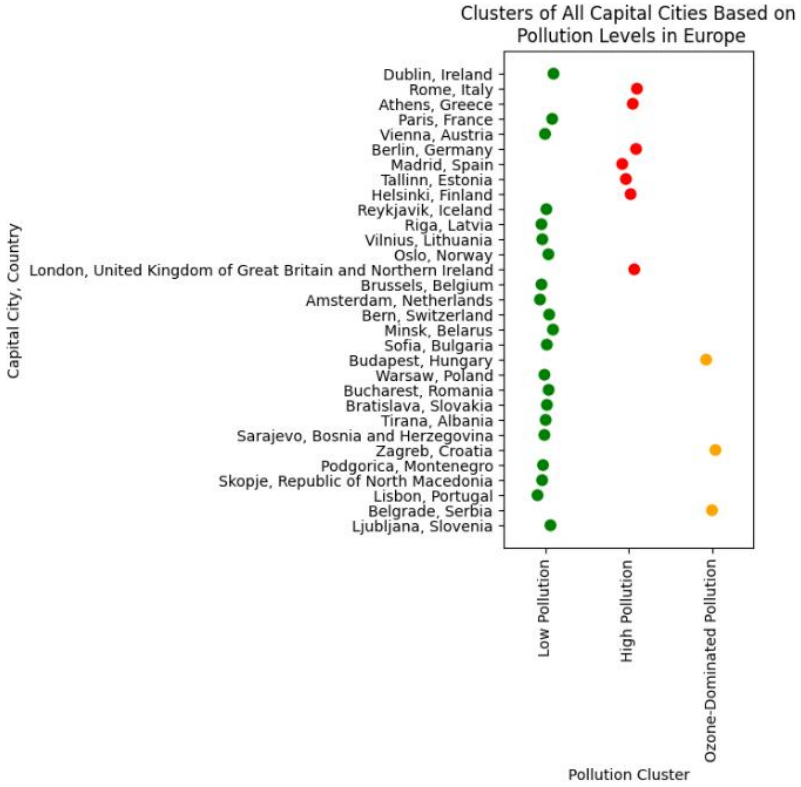


Fig 24: Cluster Plot of European Capital Cities

In Tallinn, classified in the High Pollution cluster as shown in Figure 24, the city's most contributing feature is the ozone AQI value, which deviates by 0.66 from the cluster average, indicating a localized influence of ozone. Helsinki, another city in this cluster, stands out for its PM2.5 AQI value, which exceeds the cluster average by 0.62, emphasizing significant particulate matter pollution. Conversely, London, although part of the High Pollution cluster, has an overall AQI value 0.76 below the cluster average, reflecting comparatively lower contributions from multiple pollutants.

Cities like Reykjavik, Dublin, Riga, and Vilnius are part of the Low Pollution cluster. Reykjavik shows a notable deviation in its NO2 AQI value, which is 0.68 below the cluster average, indicating minimal nitrogen dioxide pollution. Similarly, Vilnius and Amsterdam also exhibit significantly lower NO2 AQI values, aligning with the region's cleaner air profile. On the other hand, Vienna and Paris show higher NO2 contributions, deviating by 0.71 above the cluster average, highlighting localized nitrogen dioxide sources.

Budapest, representing the Ozone-Dominated Pollution cluster, has slightly lower AQI values for ozone and PM2.5 compared to the cluster averages. Still, it emphasizes the prominence of ozone in shaping pollution dynamics. This cluster's distinct signature underscores the role of regional and seasonal factors in elevating ozone levels.

The analysis underlines the diverse pollution patterns across European cities, driven by local sources and regional influences. Policymakers can leverage these findings to address specific pollutants dominating each city's air quality profile and develop targeted strategies for sustainable urban management.

4.7.4 Australia

The analysis of air pollution patterns in Australia reveals significant regional variations across its cities. Three distinct clusters emerge: High Pollution, Low Pollution, and Ozone-Dominated Pollution. Figure 25 shows that the High Pollution cluster has an average AQI of 4.48, heavily influenced by elevated PM2.5 levels (average AQI of 4.48) and significant NO2 contributions (average AQI of 2.31). The Low Pollution cluster, with an average AQI of 3.80, is marked by lower overall pollutant levels, particularly CO (average AQI of 0.57) and NO2 (average AQI of 0.68). The Ozone-Dominated Pollution cluster has the highest average AQI of 4.96, driven primarily by ozone (average AQI of 4.08) and PM2.5 (average AQI of 4.92).

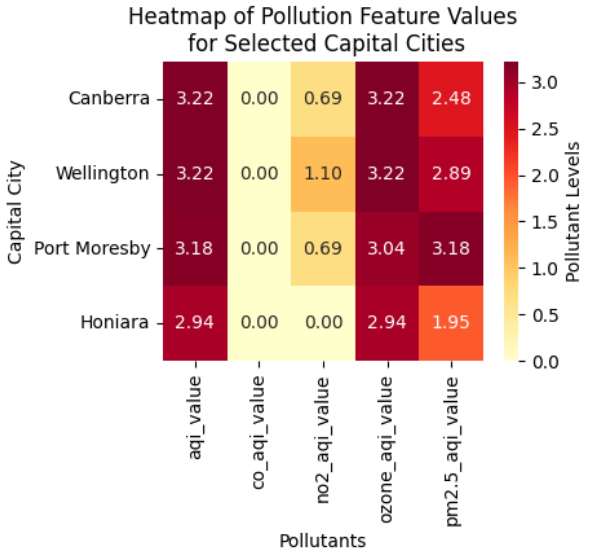


Fig 25: Heatmap of Air Quality Indicators for Australian Capital Cities

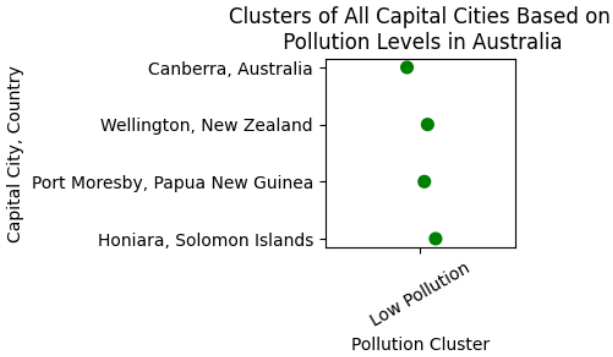


Fig 26: Cluster Plot of Australian Capital Cities

Among the cities analyzed, Canberra, classified under the Low Pollution cluster in Figure 26, demonstrates a unique profile. Its AQI value of 3.22 is lower than the cluster average of 3.80, primarily due to a PM2.5 AQI value of 2.48, which deviates significantly by -1.20 from the cluster average. Wellington, also in the Low Pollution cluster, exhibits a similar pattern with a PM2.5 AQI value of 2.89, deviating by -0.79. These deviations highlight relatively low particulate matter levels in these cities compared to the cluster norms.

Port Moresby, another city in the Low Pollution cluster, shows a notable deviation in its overall AQI value of 3.18, which is 0.62 below the cluster average. This deviation underscores generally low pollution levels across all contributing pollutants. Honiara, also part of this cluster, has an AQI value of 2.94, significantly lower than the cluster average, driven by a PM2.5 AQI value of 1.95, deviating by -1.74. This sharp contrast points to exceptionally clean air in the region, with minimal particulate pollution.

Cities like Ngerulmud and Port Vila lack sufficient data, which limits their inclusion in the clustering analysis. Nonetheless, the findings underscore the varied pollution dynamics across Australian cities, where local sources and atmospheric conditions strongly influence air quality. By addressing the specific pollutants prevalent in each city, policymakers can implement targeted measures to maintain and improve air quality, ensuring sustainable urban development in the region.

4.7.5 North America

In North America, the clustering analysis reveals distinct patterns of air pollution characterized by three main clusters: High Pollution, Low Pollution, and Ozone-Dominated Pollution. Each cluster demonstrates unique air quality profiles, reflecting the diverse environmental and urban conditions across the continent. The High Pollution cluster is dominated by elevated levels of PM2.5, with an average AQI value of approximately 4.48. This cluster also exhibits higher contributions from NO2 (average AQI 2.31) and CO (average AQI 1.23), while ozone levels remain moderate at around 2.60 as shown in Figure 27. Cities like San Salvador, Guatemala City, and Managua fall into this cluster, with notable deviations. For instance, San Salvador exhibits a slightly higher ozone AQI (+0.45), whereas Guatemala City and Managua show negative deviations in ozone and CO AQI, respectively.

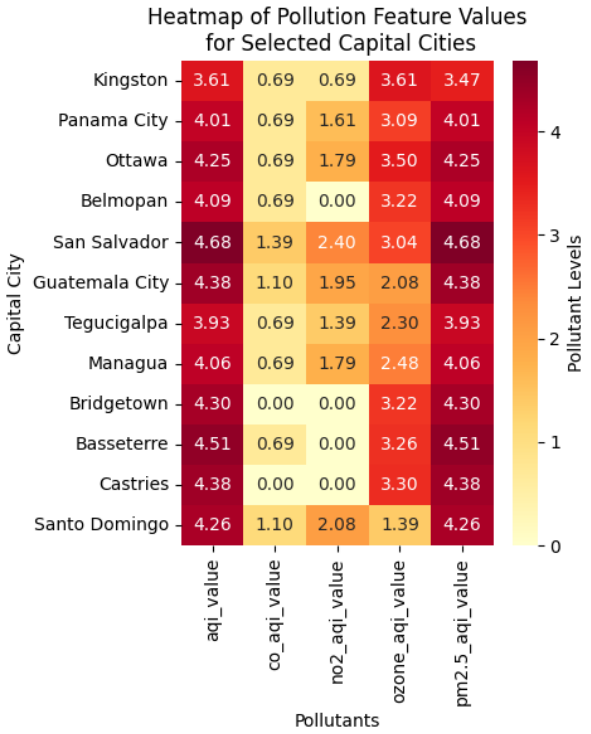


Fig 27: Heatmap of Air Quality Indicators for North American Capital Cities

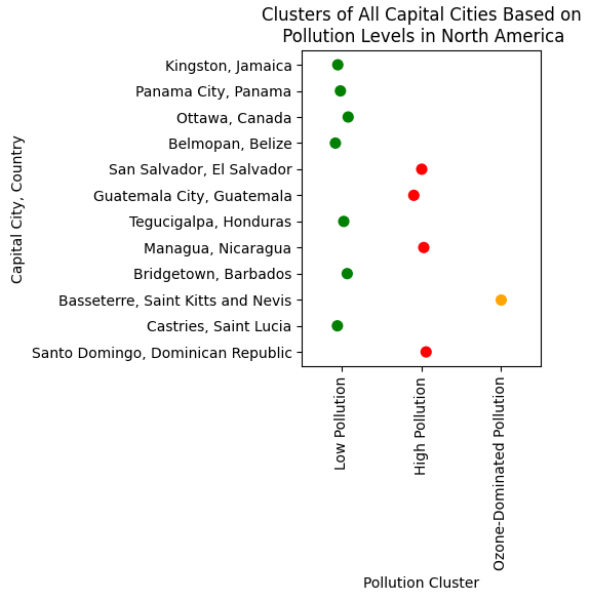


Fig 28: Cluster Plot of North American Capital Cities

Figure 28 shows the Low Pollution cluster, in contrast, is marked a significantly elevated NO2 AQI (+1.11) compared to the cluster average, while Tegucigalpa has a notably lower ozone AQI (-1.05). Such deviations highlight localized factors influencing air quality in these regions.

The Ozone-Dominated Pollution cluster represents cities with the highest ozone contributions, averaging around 4.08, coupled with significant PM2.5 levels (4.92). Other pollutants, such as CO and NO2, are less pronounced in this cluster. Basseterre is an example of a city in this category, with its ozone AQI falling below the cluster average (-0.82), suggesting variability even within this distinct pollution profile.

The analysis of individual cities provides further insights into pollution dynamics. For example, Santo Domingo, in the High Pollution cluster, exhibits a lower-than-average ozone AQI (-1.21), emphasizing localized reductions in specific pollutants. Similarly, Castries, part of the Low Pollution cluster, demonstrates a positive deviation in PM2.5 AQI, reinforcing its contribution to the city's air quality index.

In summary, North America showcases a diverse range of air pollution profiles, with clusters defined by varying levels of pollutants such as PM2.5, ozone, NO2, and CO. The High Pollution cluster is primarily driven by particulate matter and nitrogen dioxide, the Low Pollution cluster balances modest ozone levels with reduced contributions from other pollutants, and the Ozone-Dominated Pollution cluster highlights significant ozone impacts. These findings underline the importance of targeted air quality management strategies tailored to the specific needs of each region and pollutant profile.

4.7.6 South America

In South America, the analysis of air pollution data reveals three primary clusters: High Pollution, Low Pollution, and Ozone-Dominated Pollution. Each cluster reflects distinct pollutant profiles that provide insight into the continent's air quality dynamics. The High Pollution cluster is characterized by elevated PM2.5 levels, with an average AQI of 4.48. This cluster also shows significant contributions from NO2 (average AQI 2.31) and CO (average AQI 1.23) in Figure 29, while ozone levels are moderate at around 2.60. Cities such as Santiago, Lima, and Quito fall into this cluster. Santiago demonstrates a lower-than-average CO AQI (-0.54), suggesting localized reductions in this pollutant. Lima shows slightly higher CO levels (+0.38) than the cluster average, emphasizing its contribution to the overall air quality index. Quito, on the other hand, exhibits a notable deviation in ozone AQI (-2.60), highlighting the variability in pollutant levels even within this high-pollution category.

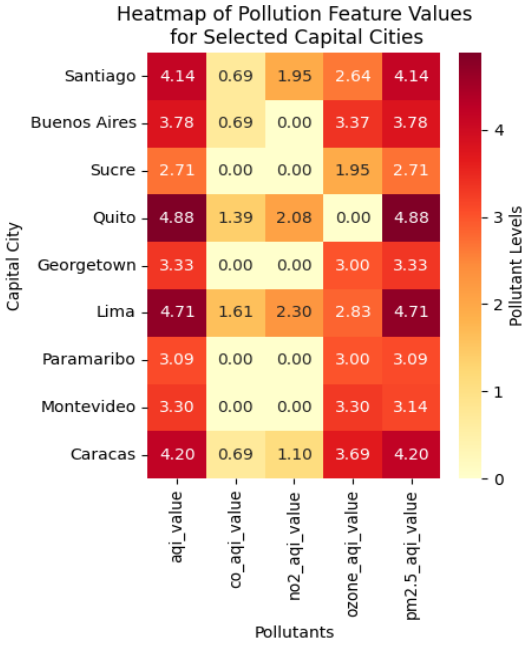


Fig 29: Heatmap of Air Quality Indicators for South American Capital Cities

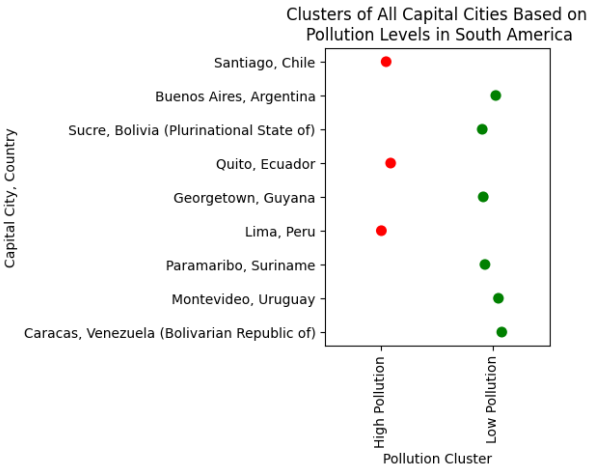


Fig 30: Cluster Plot of South American Capital Cities

Figure 30 represents the Low Pollution cluster cities with reduced pollutant levels, maintaining an average AQI of 3.80. PM2.5 levels are relatively low (3.68), while ozone levels are moderately high (3.35). CO and NO2 levels are minimal, averaging 0.57 and 0.68 AQI, respectively. Cities such as Buenos Aires, Sucre, Georgetown, Caracas, Paramaribo, and Montevideo are part of this cluster. Buenos Aires shows a significantly low NO2 AQI (-0.68), reflecting minimal nitrogen dioxide pollution. Sucre also exhibits a substantial reduction in ozone AQI (-1.41), indicating localized improvements in air quality. Caracas, however, has a higher PM2.5 AQI (+0.52), contributing to its slightly elevated pollution levels within this cluster.

The Ozone-Dominated Pollution cluster features cities with the highest ozone contributions, averaging 4.08 AQI, along with notable PM2.5 levels (4.92). CO and NO2 levels are lower, averaging 0.99 and 0.73 AQI, respectively. However, specific cities belonging to this cluster were not highlighted in the current dataset, leaving room for further exploration.

Individual city profiles provide further nuances. For example, Montevideo exhibits a significant reduction in NO2 AQI (-0.68), aligning with the low nitrogen dioxide levels observed in the cluster. Paramaribo, with a lower overall AQI (-0.71), reflects minimal contributions from pollutants like CO and NO2. Meanwhile, Caracas shows a balanced yet slightly higher PM2.5 AQI, reinforcing its influence on the city's air quality.

In summary, South America exhibits diverse air pollution patterns across its clusters. The High Pollution cluster emphasizes particulate matter and nitrogen dioxide, the Low Pollution cluster balances moderate ozone levels with minimal contributions from other pollutants, and the Ozone-Dominated Pollution cluster underscores the impact of elevated ozone levels. These findings emphasize the need for region-specific air quality management strategies to address the unique challenges posed by different pollutants in South American cities.

5. DISCISSION

This study provides a comprehensive analysis of global urban air pollution patterns by examining air quality indicators, including PM2.5, CO, NO2, and Ozone, across cities from six continents. The analysis identified three distinct pollution clusters: High Pollution, Low Pollution, and Ozone-Dominated Pollution, which were observed across various regions, including Asia, Africa, Europe, North America, South America, and Australia.

The High Pollution cluster, dominated by elevated levels of PM2.5 and NO2, was found in regions like Asia, Africa, and North America, signaling urgent air quality management needs. On the other hand, the Low Pollution cluster, with generally lower pollutant levels, was more common in Europe, Australia, and parts of South America. The Ozone-Dominated Pollution cluster, characterized by high ozone levels, was notably present in cities in Asia and Africa, highlighting the significant role of photochemical reactions in certain regions.

While these findings provide valuable insights into regional disparities in air quality, there are several limitations to this study. One of the key challenges is the absence of data for several major capital cities, which could further enrich the analysis and provide a more complete picture of global air pollution trends. In future studies, efforts will be made to include more cities from these regions to address this gap and improve the robustness of the analysis. Further research can focus on longitudinal studies to track changes in air quality over time and assess the effectiveness of pollution control measures in different regions. Additionally, the integration of real-time data and the incorporation of other pollutants, such as Sulfur Dioxide (SO2) and Volatile Organic Compounds (VOCs), could provide a more comprehensive understanding of urban air pollution. Finally, expanding the clustering approach to include not only capital cities but also major industrial hubs and smaller cities would allow for a more nuanced understanding of air quality across diverse urban environments.

In conclusion, while this study highlights important regional pollution patterns, future work will build on these findings by incorporating more cities, exploring the temporal dynamics of pollution, and refining the clustering methodology to provide deeper insights into global air quality challenges.

6. CONCLUSION

This study offers a comprehensive analysis of urban air pollution across six continents, revealing the varied contributions of PM2.5, NO2, CO, and Ozone to the overall air quality in capital cities worldwide. Through clustering, we identified distinct pollution patterns, which help to understand how different cities group together based on similar pollutant profiles. The clustering approach provided valuable insights into the similarities and differences between cities, offering a clear picture of regional pollution trends and helping to highlight cities that may require targeted intervention strategies.

Cities in high-pollution clusters, particularly in Asia, Africa, and North America, are in urgent need of comprehensive air quality management strategies that target multiple pollutants, especially PM2.5, NO2, and Ozone. In contrast, cities in the Low Pollution Clusters across Europe, Australia, and South America provide valuable lessons in achieving cleaner air through strong policies and investments in sustainable infrastructure, including public transportation, emissions controls, and green spaces.

The role of ozone in certain regions underscores the importance of addressing the unique challenges posed by photochemical reactions, which are influenced by both environmental conditions and emissions. In this context, targeted policies that address local sources of Ozone and PM2.5 are crucial for

reducing pollution in affected cities.

This research emphasizes the need for region-specific approaches to air quality management and highlights the critical role of international collaboration in addressing the global challenge of urban air pollution. The use of clustering not only facilitated the identification of key pollution drivers but also provided a framework for creating effective, regionally tailored solutions. As cities worldwide continue to grow and industrialize, it is imperative that policymakers implement more effective and regionally tailored solutions to safeguard public health and the environment.

7. LIMITATION

Despite the comprehensive nature of this study, several limitations must be considered. First, the dataset used in this analysis relies on available air pollution data from urban areas, which may not fully represent pollution levels in smaller or rural cities. The lack of consistent and up-to-date data from some countries, particularly in regions with limited monitoring infrastructure, could affect the robustness of the clustering results. Additionally, missing data from certain cities, such as Maldives and Kyrgyzstan, may have influenced the overall patterns observed.

Second, while clustering provides valuable insights into pollution trends, it is important to note that the method is inherently dependent on the selected variables. Other factors, such as seasonal variations, specific local sources of pollution, and socioeconomic conditions, were not considered in this study but could have provided further context to the observed patterns. Lastly, the study focuses mainly on urban areas and the findings may not fully capture the dynamics of rural pollution, which may differ significantly from those in urban centers.

Future research could expand on this analysis by incorporating more diverse geographical data and additional environmental and health-related variables.

8. REFERENCES

[1] AirVisual, I., 2018. World Air Quality Report: Region & City PM 2.5 Ranking. *IQAir AirVisual*.

[2] International Day of Clean Air for Blue Skies, Accessed: 2024-12-31.

[3] Mo, X., Li, H., Zhang, L. and Qu, Z., 2021. Environmental impact estimation of PM2. 5 in representative regions of China from 2015 to 2019: policy validity, disaster threat, health risk, and economic loss. *Air Quality, Atmosphere & Health*, *14*(10), pp.1571-1585.

[4] Mo, X., Li, H. and Zhang, L., 2022. Design a regional and multistep air quality forecast model based on deep learning and domain knowledge. *Frontiers in Earth Science*, *10*, p.995843.

[5] Rahman, A. and Khatun, M.T., 2024, November. Multivariate Analysis of Urban Air Pollution: Clustering and Patterns Across Major Asian Cities. In *2024 IEEE International Conference on Future Machine Learning and Data Science (FMLDS)* (pp. 462-467). IEEE.

[6] Mu, L., Su, J., Mo, X., Peng, N., Xu, Y., Wang, M. and Wang, J., 2021. The temporal-spatial variations and potential causes of dust events in Xinjiang Basin during 1960–2015. *Frontiers in Environmental Science*, *9*, p.727844.

[7] Li, X., Jin, L. and Kan, H., 2019. Air pollution: a global problem needs local fixes. *Nature*, *570*(7762), pp.437-439.

[8] Greenstone, M., Nilekani, J., Pande, R., Ryan, N., Sudarshan, A. and Sugathan, A., 2015. Lower pollution, longer lives: life expectancy gains if India reduced particulate matter pollution. *Economic and Political Weekly*, pp.40-46.

[9] Ghude, S.D., Chate, D.M., Jena, C., Beig, G., Kumar, R., Barth, M.C., Pfister, G.G., Fadnavis, S. and Pithani, P., 2016. Premature mortality in India due to PM2. 5 and ozone exposure. *Geophysical Research Letters*, *43*(9), pp.4650-4658.

[10] Kumar, R., Barth, M.C., Pfister, G.G., Delle Monache, L., Lamarque, J.F., Archer-Nicholls, S., Tilmes, S., Ghude, S.D., Wiedinmyer, C., Naja, M. and Walters, S., 2018. How will air quality change in South Asia by 2050?. *Journal of Geophysical Research: Atmospheres*, *123*(3), pp.1840-1864.

[11] Lee, H.H., Iraqui, O., Gu, Y., Yim, S.H.L., Chulakadabba, A., Tonks, A.Y.M., Yang, Z. and Wang, C., 2018. Impacts of air pollutants from fire and non-fire emissions on the regional air quality in Southeast Asia. *Atmospheric Chemistry and Physics*, *18*(9), pp.6141-6156.

[12] Lee, H.H., Bar-Or, R.Z. and Wang, C., 2017. Biomass burning aerosols and the low-visibility events in Southeast Asia. *Atmospheric Chemistry and Physics*, *17*(2), pp.965-980.

[13] Guo, R., Zhang, Q., Yu, X., Qi, Y. and Zhao, B., 2023. A deep spatio-temporal learning network for continuous citywide air quality forecast based on dense monitoring data. *Journal of Cleaner Production*, *414*, p.137568.

[14] Elbaz, K., Hoteit, I., Shaban, W.M. and Shen, S.L., 2023. Spatiotemporal air quality forecasting and health risk assessment over smart city of NEOM. *Chemosphere*, *313*, p.137636.

[15] Abdul Jabbar, S., Tul Qadar, L., Ghafoor, S., Rasheed, L., Sarfraz, Z., Sarfraz, A., Sarfraz, M., Felix, M. and Cherrez-Ojeda, I., 2022. Air quality, pollution and sustainability trends in South Asia: a population-based study. *International journal of environmental research and public health*, *19*(12), p.7534.

[16] Jacob, D.J. and Winner, D.A., 2009. Effect of climate change on air quality. *Atmospheric environment*, *43*(1), pp.51-63.

[17] Guha, S., Rastogi, R. and Shim, K., 2000. ROCK: A robust clustering algorithm for categorical attributes. *Information systems*, *25*(5), pp.345-366.

[18] Singha, S.P., Hossain, M.M., Rahman, M.A. and Sharmin, N., 2024. Investigation of graph-based clustering approaches along with graph neural networks for modeling armed conflict in Bangladesh. *International Journal of Data Science and Analytics*, *18*(2), pp.187-203.

[19] Liu, H., Long, Z., Duan, Z. and Shi, H., 2020. A new model using multiple feature clustering and neural networks for forecasting hourly PM2. 5 concentrations,

and its applications in China. *Engineering*, *6*(8), pp.944-956.

[20] Yan, R., Liao, J., Yang, J., Sun, W., Nong, M. and Li, F., 2021. Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering. *Expert Systems with Applications*, *169*, p.114513.

[21] Wang, K., Qi, X., Liu, H. and Song, J., 2018. Deep belief network based k-means cluster approach for short-term wind power forecasting. *Energy*, *165*, pp.840-852.

[22] Li, D.H. and Cao, Y.J., 2005, November. SOFM based support vector regression model for prediction and its application in power system transient stability forecasting. In *2005 International power engineering conference* (pp. 765-770). IEEE.

[23] Chen, S., Wang, J.Q. and Zhang, H.Y., 2019. A hybrid PSO-SVM model based on clustering algorithm for short-term atmospheric pollutant concentration forecasting. *Technological Forecasting and Social Change*, *146*, pp.41-54.

[24] Heydari, A., Majidi Nezhad, M., Astiaso Garcia, D., Keynia, F. and De Santoli, L., 2022. Air pollution forecasting application based on deep learning model and optimization algorithm. *Clean Technologies and Environmental Policy*, pp.1-15.

[25] Mao, W., Wang, W., Jiao, L., Zhao, S. and Liu, A., 2021. Modeling air quality prediction using a deep learning approach: Method optimization and evaluation. *Sustainable Cities and Society*, *65*, p.102567.

[26] Chakma, A., Vizena, B., Cao, T., Lin, J. and Zhang, J., 2017, September. Image-based air quality analysis using deep convolutional neural network. In *2017 IEEE international conference on image processing (ICIP)* (pp. 3949-3952). IEEE.

[27] Chaturvedi, P., 2024. Air Quality Prediction System Using Machine Learning Models. *Water, Air, & Soil Pollution*, *235*(9), p.578.

[28] Dataset: Global Air Pollution Data, Accessed: 2024-12-31.