Berry Team : Janani Krishnamurthy, Siva Sushmitha Meduri, Lakshmi Prasanna Kumar Nalabothu

Written Homework 4 : Unsupervised Learning

Title : Exploring Global Air Pollution Trends: Unsupervised Learning Analysis of 100 Countries from 1960 to 2022

Abstract

This project focuses on applying unsupervised learning techniques to analyse air pollution data [1] from 100 countries, spanning from 1960 to 2022. The dataset includes concentrations of seven pollutants (Nitrogen Oxide, Carbon, Sulphur Dioxide, Methane, Ammonia) along with demographic information such as urban population [2], rural population [3], and total population [4] for each country. By leveraging Principal Component Analysis (PCA), k-means clustering, Singular Value Decomposition (SVD), and Hierarchical Clustering, the aim is to uncover hidden patterns and groupings within the data. These methods provide a comprehensive understanding of how different factors contribute to air pollution across various countries and time periods, facilitating the identification of significant trends and clusters in the data. Principal Component Analysis (PCA) was used to reveal 10 principal components, with the first principal component (PC1) explaining nearly 83.92% of the variance. The optimal number of clusters was identified as four using the elbow method and silhouette analysis, and this was applied in k-means clustering. The number of data points that fell into each of the four clusters was noted. For hierarchical clustering, different linkage methods were compared to determine the most effective grouping. By leveraging PCA, k-means clustering, Singular Value Decomposition (SVD), and Hierarchical Clustering, hidden patterns and groupings within the data were uncovered. These methods provided a comprehensive understanding of how different factors contribute to air pollution across various countries and time periods, facilitating the identification of significant trends and clusters in the data.

Introduction and Overview

Air pollution is a critical global issue that poses significant health and environmental challenges. Understanding the dynamics of air pollution across different countries and over time is essential for formulating effective policies and mitigation strategies. In this study, a comprehensive dataset comprising air pollution metrics for 100 countries from 1960 to 2022 was analysed. The air pollution data, sourced from ourworldindata.org, includes measurements of seven key pollutants: Nitrogen Oxide (NOx), Carbon Monoxide (CO), Sulphur Dioxide (SO2), Non -Methane volatile compounds (NMVOC), and Ammonia (NH3), Black Carbon (BC) and Organic Carbon (OC). Alongside this, demographic information such as urban population, rural population, and total population for each country, sourced from data.worldbank.org was incorporated.

Unsupervised learning techniques are particularly suited for this type of exploratory data analysis because they do not require predefined labels, allowing us to discover the intrinsic structure of the data. In this study, four main techniques to analyze the data: Principal Component Analysis (PCA), k-means clustering, Singular Value Decomposition (SVD), and Hierarchical Clustering were employed.

Principal Component Analysis (PCA) is used to reduce the dimensionality of the data while retaining most of the variance, helping us identify the most significant features and visualize the data more effectively. k-means clustering is employed to partition the data into k distinct clusters based on similarities in the features, providing insights into the natural groupings within the dataset. Singular Value Decomposition (SVD), similar to PCA, decomposes the data into its constituent components, aiding in dimensionality reduction and highlighting key patterns. Hierarchical Clustering creates a hierarchy of clusters, offering a tree-like structure that represents nested groupings of the data.

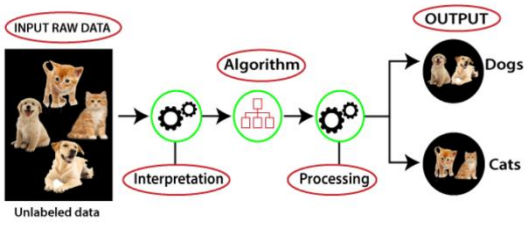
By applying these techniques, the study aim to reveal underlying patterns and relationships within the air pollution data, which can inform further research and policy development. The insights gained from this analysis will contribute to a better understanding of global air pollution trends and the factors influencing them, ultimately supporting efforts to address this pressing environmental issue.

Theoretical Background

Classification:

Unsupervised learning, also known as unsupervised machine learning, uses machine learning (ML) algorithms to analyse and cluster unlabelled data sets. These algorithms discover hidden patterns or data groupings without the need for human intervention. Unsupervised learning is a type of machine learning in which models are trained using unlabelled dataset and are allowed to act on that data without any supervision. Unsupervised learning's ability to discover similarities and differences in information make it the ideal solution for exploratory data analysis, cross-selling strategies, customer segmentation and image recognition.

Here, we have taken an unlabelled input data, which means it is not categorized and corresponding outputs are also not given. Now, this unlabelled input data is fed to the machine learning model in order to train it. Firstly, it will interpret the raw data to find the hidden patterns from the data and then will apply suitable algorithms such as k-means clustering, Decision tree, etc. Once it applies the suitable algorithm, the algorithm divides the data objects into groups according to the similarities and difference between the objects. [5]



Unsupervised learning models are utilized for three main tasks—clustering, association, and dimensionality reduction. Below each learning method and highlight common algorithms is discussed and approaches to conduct them effectively.

Dimensionality reduction

While more data generally yields more accurate results, it can also impact the performance of machine learning algorithms (e.g. overfitting) and it can also make it difficult to visualize datasets. Dimensionality reduction is a technique used when the number of features, or dimensions, in a given dataset is too high. It reduces the number of data inputs to a manageable size while also preserving the integrity of the dataset as much as possible. It is commonly used in the preprocessing data stage, and there are a few different dimensionality reduction methods that can be used, such as:

a. Principal component analysis

Principal component analysis (PCA) is a type of dimensionality reduction algorithm which is used to reduce redundancies and to compress datasets through feature extraction. This method uses a linear transformation to create a new data representation, yielding a set of "principal components." The first principal component is the direction which maximizes the variance of the dataset. While the second principal component also finds the maximum variance in the data, it is completely uncorrelated to the first principal component, yielding a direction that is perpendicular, or orthogonal, to the first component. This process repeats based on the number of dimensions, where a next principal component is the direction orthogonal to the prior components with the most variance.

b. Singular value decomposition:

Singular value decomposition (SVD) is another dimensionality reduction approach which factorizes a matrix, A, into three, low-rank matrices. SVD is denoted by the formula, A = USVT, where U and V are orthogonal matrices. S is a diagonal matrix, and S values are considered singular values of matrix A. Similar to PCA, it is commonly used to reduce noise and compress data, such as image files.

Clustering: Clustering is a data mining technique which groups unlabeled data based on their similarities or differences. Clustering algorithms are used to process raw, unclassified data objects into groups represented by structures or patterns in the information. Clustering algorithms can be categorized into a few types, specifically exclusive, overlapping, hierarchical, and probabilistic.

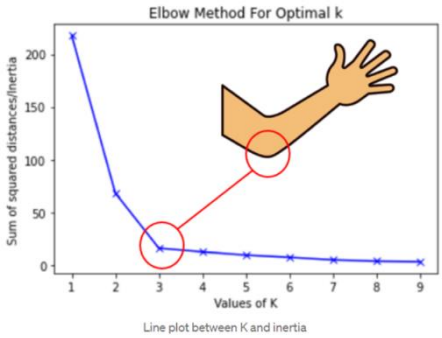
Exclusive and Overlapping Clustering: Exclusive clustering is a form of grouping that stipulates a data point can exist only in one cluster. This can also be referred to as “hard” clustering. The K-means clustering algorithm is an example of exclusive clustering.

c. K-means clustering is a common example of an exclusive clustering method where data points are assigned into K groups, where K represents the number of clusters based on the distance from each group’s centroid. The data points closest to a given centroid will be clustered under the same category. A larger K value will be indicative of smaller groupings with more granularity whereas a smaller K value will have larger groupings and less granularity. K-means clustering is commonly used in market segmentation, document clustering, image segmentation, and image compression.[6]

There are two main methods to find the best value of K. We will discuss them individually.

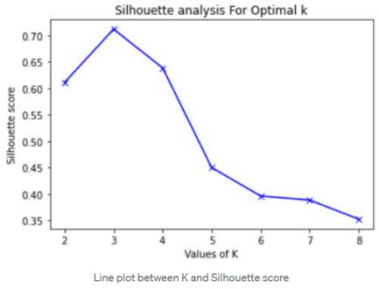
1) *Elbow Curve Method:* Recall that the basic idea behind partitioning methods, such as k-means clustering, is to define clusters such that the total intra-cluster variation [or total within-cluster sum of square (WSS)] is minimized. The total wss measures the compactness of the clustering, and we want it to be as small as possible. The

elbow method runs k-means clustering (kmeans number of clusters) on the dataset for a range of values of k (say 1 to 10) In the elbow method, we plot mean distance and look for the elbow point where the rate of decrease shifts. For each k, calculate the total within-cluster sum of squares (WSS). This elbow point can be used to determine K.



2) *Silhouette Analysis:* Silhouette analysis is a method used to evaluate the quality of clustering by measuring how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette coefficient ranges from -1 to 1, where a higher value indicates better-defined clusters. Apply a clustering algorithm (e.g., K-means, hierarchical clustering) to the dataset for different values of k (number of clusters). For each data point i: Compute the mean distance $a(i)$ between i and all other points in the same cluster (intra-cluster distance). Compute the mean distance $b(i)$ between i and all points in the nearest neighboring cluster (inter-cluster distance).The silhouette coefficient $s(i)$ for point i is defined as:
Plot the mean silhouette coefficient against different values of k.The optimal k is typically the one that maximizes the mean silhouette coefficient.

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$



The silhouette Method is used in combination with the Elbow Method for a more confident decision.

Overlapping clusters differs from exclusive clustering in that it allows data points to belong to multiple clusters with separate degrees of membership. “Soft” or fuzzy k-means clustering is an example of overlapping clustering.

d. Hierarchical clustering

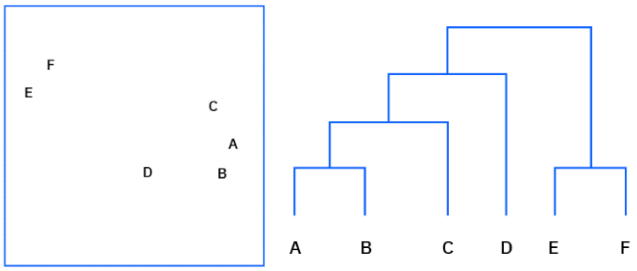
Hierarchical clustering, also known as hierarchical cluster analysis (HCA), is an unsupervised clustering algorithm that can be categorized in two ways: agglomerative or divisive.

Agglomerative clustering is considered a “bottoms-up approach.” Its data points are isolated as separate groupings initially, and then they are merged together iteratively on the basis of similarity until one cluster has been achieved. Four different methods are commonly used to measure similarity:

- *Ward’s linkage:* This method states that the distance between two clusters is defined by the increase in the sum of squared after the clusters are merged.
- *Average linkage:* This method is defined by the mean distance between two points in each cluster.
- *Complete (or maximum) linkage:* This method is defined by the maximum distance between two points in each cluster.
- *Single (or minimum) linkage:* This method is defined by the minimum distance between two points in each cluster.

Euclidean distance is the most common metric used to calculate these distances; however, other metrics, such as Manhattan distance, are also cited in clustering literature.

Divisive clustering can be defined as the opposite of agglomerative clustering; instead it takes a “top-down” approach. In this case, a single data cluster is divided based on the differences between data points. Divisive clustering is not commonly used, but it is still worth noting in the context of hierarchical clustering. These clustering processes are usually visualized using a dendrogram, a tree-like diagram that documents the merging or splitting of data points at each iteration.



Applications of unsupervised learning:

Machine learning techniques have become a common method to improve a product user experience and to test systems for quality assurance. Unsupervised learning provides an exploratory path to view data, allowing businesses to identify patterns in large volumes of data more quickly when compared to manual observation. Some of the most common real-world applications of unsupervised learning are:

- News Sections: Google News uses unsupervised learning to categorize articles on the same story from various online news outlets. For example, the results of a presidential election could be categorized under their label for “US” news.
- Computer vision: Unsupervised learning algorithms are used for visual perception tasks, such as object recognition.
- Medical imaging: Unsupervised machine learning provides essential features to medical imaging devices, such as image detection, classification and segmentation, used in radiology and pathology to diagnose patients quickly and accurately.

- Anomaly detection: Unsupervised learning models can comb through large amounts of data and discover atypical data points within a dataset. These anomalies can raise awareness around faulty equipment, human error, or breaches in security.
- Customer personas: Defining customer personas makes it easier to understand common traits and business clients' purchasing habits. Unsupervised learning allows businesses to build better buyer persona profiles, enabling organizations to align their product messaging more appropriately.
- Recommendation Engines: Using past purchase behavior data, unsupervised learning can help to discover data trends that can be used to develop more effective cross-selling strategies. This is used to make relevant add-on recommendations to customers during the checkout process for online retailers.

Challenges of unsupervised learning:
While unsupervised learning has many benefits, some challenges can occur when it allows machine learning models to execute without any human intervention. Some of these challenges can include:

- Computational complexity due to a high volume of training data
- Longer training times
- Higher risk of inaccurate results
- Human intervention to validate output variables
- Lack of transparency into the basis on which data was clustered

Methodology

Data clean up or pre-processing:

The air pollution data initially contained information from 1750 to 2022, while the population data (urban, rural, and total) spanned from 1960 to 2022. To ensure consistency, data from 1960 to 2022 was considered for both datasets. The population dataset had years as columns, whereas the pollution dataset had years as different rows. Consequently, the population dataset was melted to bring the years as different rows for each country. Additionally, there were discrepancies between the countries listed in the population and pollution datasets. To address this, only the data(1960 to 2022) of top 100 countries with the highest total population as of 2022 were included in the analysis.

Models:

For conducting Principal Component Analysis (PCA) on the dataset, the important variables such as emissions data (Nitrogen oxide (NOx), Sulphur dioxide (SO_2), Carbon monoxide (CO), Organic carbon (OC), Non-methane volatile organic compounds (NMVOC), Black carbon (BC), Ammonia (NH_3)) and population metrics (Urban Population, Rural Population, Total Population) were selected. Standard Scaler was applied to standardise the chosen columns. Then, PCA was performed on the normalised data to decrease its dimensions while preserving a majority of the variance. The first two principal components of the PCA results were graphed in a scatter plot, with country names labelled on the plot and loading vectors included to explain their impact on the principal components. This visualisation aided in grasping the connections between variables and pinpointing patterns in the data.

The Elbow Method and Silhouette Analysis were employed to find the best number of clusters, and the optimum cluster quantity was established. Following that, K-means

clustering was used on the normalised data to place each data point into the appropriate cluster. The data points were plotted in the space of the first two principal components from PCA to visualise the clustering results, with distinct colours indicating each cluster. Furthermore, the central features of each cluster is found using the centroids which were then plotted and data points distribution among the clusters were evaluated. This clustering procedure assisted in discovering unique categories in the data, exposing significant trends linked to emissions and population measurements.

Singular Value Decomposition (SVD) was conducted on the standardised data to break it down into three matrices: U, s, and Vt. Though the data set has very minimal to no missing values, SVD was conducted just for comparison purpose. This breakdown helped to assess the variance attributed to each principal component. The percentage of variance was charted for each component in order to assess their importance. Furthermore, the influence of initial features of the first two principal components was investigated by plotting their values from the Vt matrix. A scatter plot in the first two principal components' space was generated, with each data point coloured based on their cluster labels from earlier K-means clustering, to illustrate the data's distribution and connections. This examination offered understanding on the fundamental organisation of the information and the significance of each characteristic in elucidating variability.

For hierarchical clustering, initially the pairwise distances between all observations were determined by applying the Euclidean distance metric on the standardised data. Agglomerative hierarchical clustering was performed using different linkage methods: 'ward,' 'single,' 'complete,' and 'average.' Each method defines how the distances between clusters are calculated during the merging process. For each linkage method, a dendrogram was generated to visualize the hierarchical clustering results. The dendrogram displayed the merging of clusters at different distance levels, with the height of each merge representing the dissimilarity between the merged clusters. The dendrograms were cut at a specific distance threshold or number of clusters to assign cluster labels to each observation. The fcluster function is used to determine the cluster assignments based on the chosen criterion. The resulting clusters are analysed by examining the countries and years that belong to each cluster. This analysis aims to identify the characteristics that define each cluster in terms of air pollution and population trends.

Computational Results

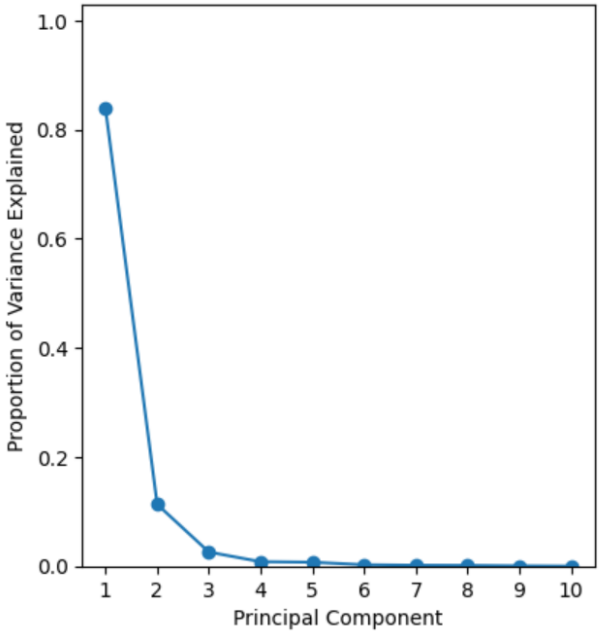
The computation results are as below for each unsupervised learning method as below.

1. Principal Component Analysis:

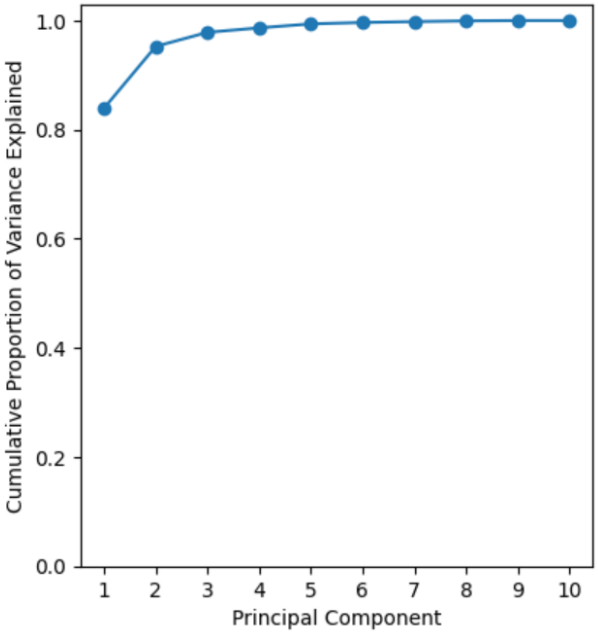
Number of Principal Components : 10

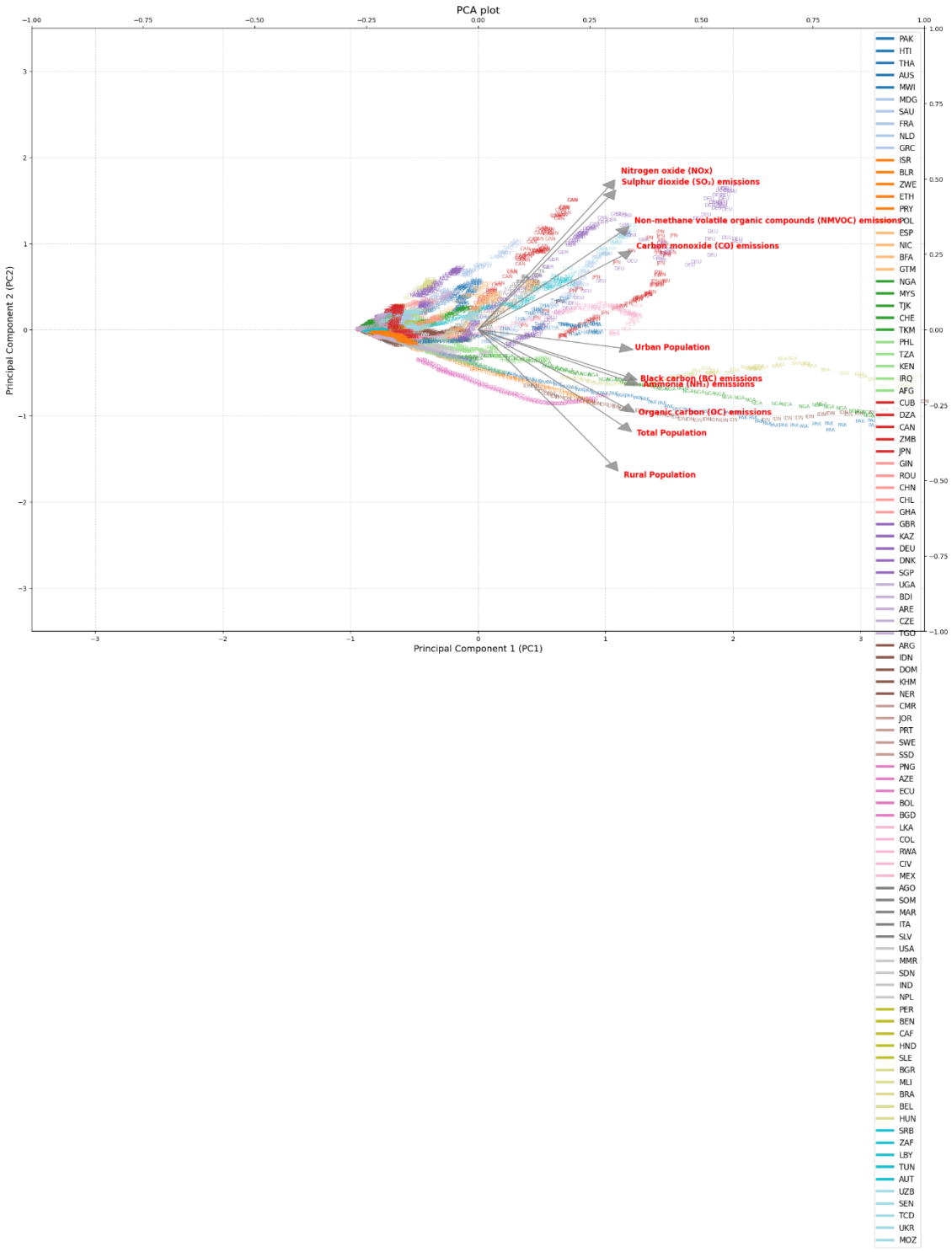
	Principal Components									
	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**	**PC10**
Proportional of Variance Explained	83.92%	11.33%	2.58%	0.82%	0.72%	0.25%	0.16%	0.13%	0.05%	0.00%

Scree Plot



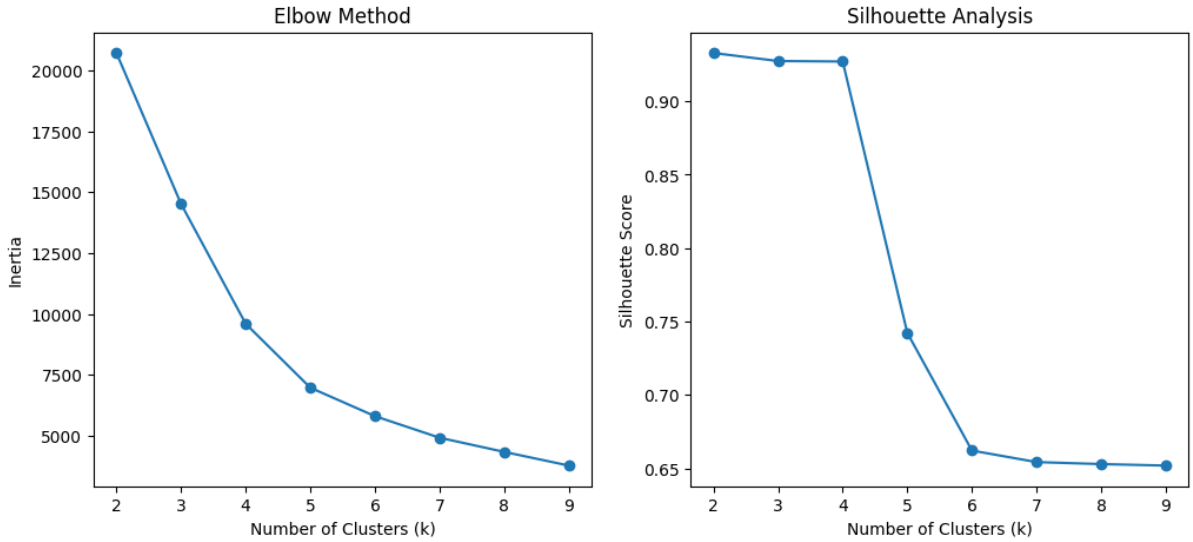
Proportion Of Variance



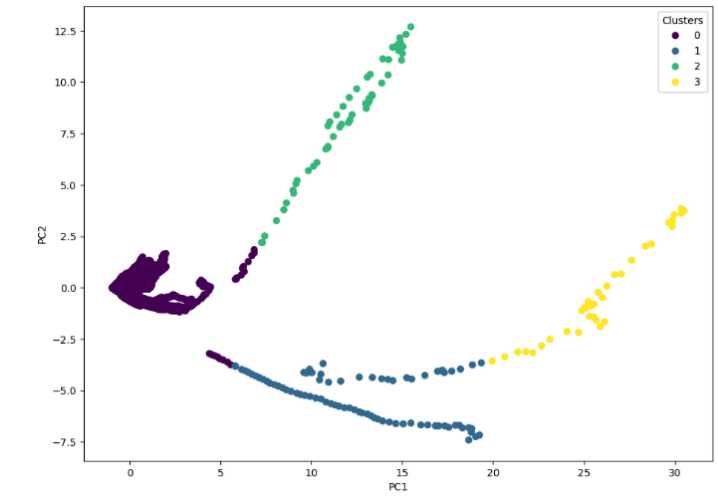


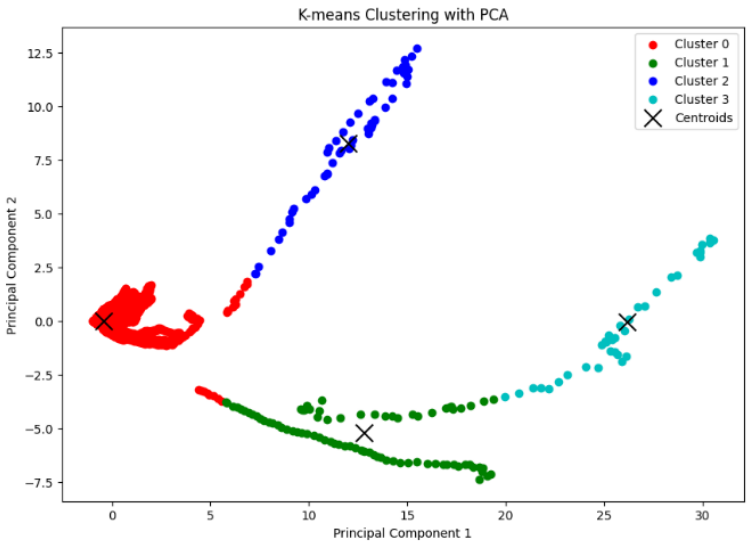
2. K-means clustering:

K-means clustering was performed to group countries based on air pollution and population data. The optimal number of clusters was determined to be 4 using the elbow method and silhouette analysis.



Cluster centroids indicate the average values for each feature in each cluster.





Number of data points in each cluster	
Cluster	**Number of Data points**
0	6130
1	81
2	52
3	37

Cluster	Countries	Number of Years
Cluster 0	Afghanistan	63
	Algeria	63
	Angola	63
	Argentina	63
	Australia	63
	Austria	63
	Azerbaijan	63
	Bangladesh	63
	Belarus	63
	Belgium	63
	Benin	63
	Bolivia	63
	Brazil	63
	Bulgaria	63
	Burkina Faso	63
	Burundi	63
	Cambodia	63
	Cameroon	63
	Canada	63
	Central African Republic	63
	Chad	63
	Chile	63
	Colombia	63
	Cote d'Ivoire	63
	Cuba	63
	Czechia	63
	Denmark	63
	Dominican Republic	63
	Ecuador	63
	El Salvador	63
	Ethiopia	63
	France	63
	Germany	63
	Ghana	63
	Greece	63

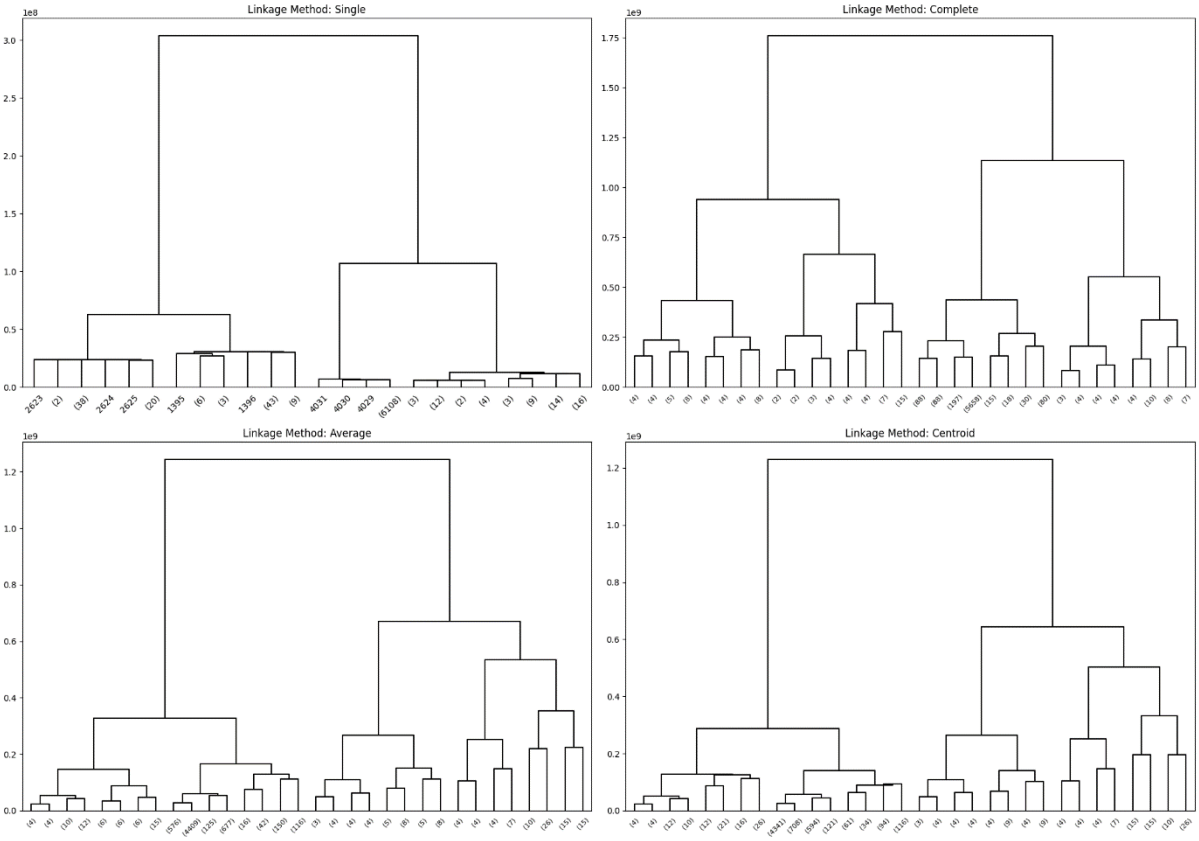
	Guatemala	63
	Guinea	63
	Haiti	63
	Honduras	63
	Hungary	63
	India	8
	Indonesia	63
	Iraq	63
	Israel	63
	Italy	63
	Japan	63
	Jordan	63
	Kazakhstan	63
	Kenya	63
	Libya	63
	Madagascar	63
	Malawi	63
	Malaysia	63
	Mali	63
	Mexico	63
	Morocco	63
	Mozambique	63
	Myanmar	63
	Nepal	63
	Netherlands	63
	Nicaragua	63
	Niger	63
	Nigeria	63
	Pakistan	63
	Papua New Guinea	63
	Paraguay	63
	Peru	63
	Philippines	63
	Poland	63
	Portugal	63
	Romania	63
	Rwanda	63
	Saudi Arabia	63
	Senegal	63
	Serbia	63
	Sierra Leone	63
	Singapore	63
	Somalia	63
	South Africa	63
	South Sudan	63
	Spain	63

	Sri Lanka	63
	Sudan	63
	Sweden	63
	Switzerland	63
	Tajikistan	63
	Tanzania	63
	Thailand	63
	Togo	63
	Tunisia	63
	Turkmenistan	63
	Uganda	63
	Ukraine	63
	United Arab Emirates	63
	United Kingdom	63
	United States	11
	Uzbekistan	63
	Zambia	63
	Zimbabwe	63
Cluster 1	India	55
	China	26
Cluster 2	United States	52
Cluster 3	China	37

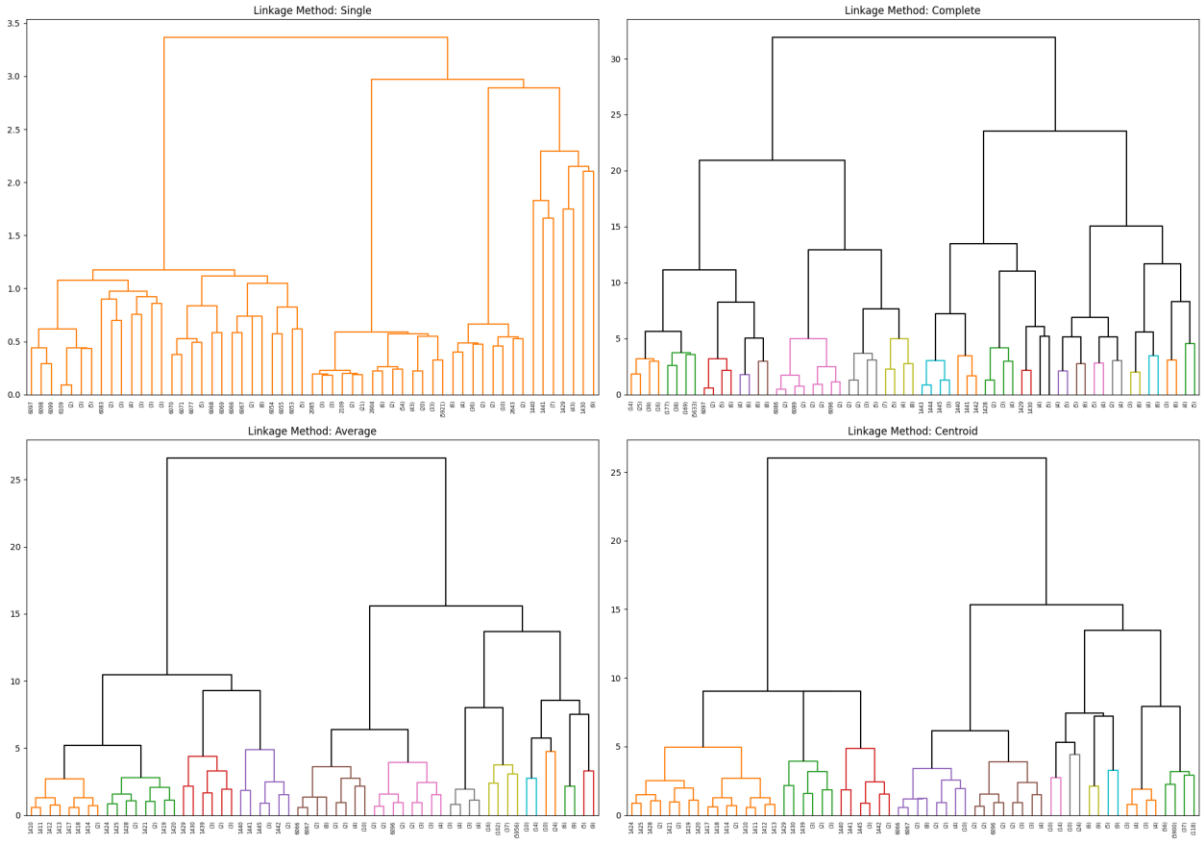
3. **Hierarchical clustering:**

Hierarchical clustering using the four methods are applied on original data and scaled data, and dendrograms were created to visualize the clusters.

Hierarchical Clustering with orginal Data



Hierarchical Clustering with Scaled Data



Discussion

The scree plot illustrates the proportion of variance explained by each principal component. In this analysis, the first principal component (PC1) explains approximately 84% of the variance, while PC2 explains 11%. The remaining principal components each explain less than 3% of the variance. The shape of the scree plot, with a steep drop-off after the first two components, suggests that most of the information in the original dataset can be effectively summarized by these two components. This indicates that the data may have a relatively simple underlying structure that can be captured by a reduced number of dimensions. Also, from the scree plot considering elbow method PC3 can be considered as optimal number of principal components.

Rotation Matrix(x and rotation from PCA) was plotted as below

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Nitrogen oxide (NOx)	0.291509	0.471876	0.270905	-0.148467	-0.244109	-0.352596	-0.556249	-0.266809	0.180030	0.000000e+00
Sulphur dioxide (SO_2) emissions	0.292245	0.438684	-0.279940	-0.668343	0.286654	0.203773	0.242473	0.097800	-0.071055	1.948788e-16
Carbon monoxide (CO) emissions	0.322644	0.244584	-0.373491	0.462540	0.047952	-0.300558	-0.041282	0.186718	-0.594506	-2.468901e-16
Organic carbon (OC) emissions	0.327470	-0.256739	-0.137220	-0.022197	-0.255984	0.682328	-0.418871	-0.189571	-0.255024	5.127110e-16
Non-methane volatile organic compounds (NMVOC) emissions	0.318389	0.321705	0.026496	0.377284	-0.426306	0.262230	0.517734	-0.058731	0.360680	4.093984e-16
Black carbon (BC) emissions	0.330459	-0.154628	-0.274412	0.277295	0.621398	-0.011311	-0.147013	-0.263786	0.487066	7.683746e-17
Ammonia (NH_3) emissions	0.336048	-0.171163	0.179327	-0.020809	-0.020106	0.011495	-0.198743	0.848327	0.255913	-3.404961e-15
Urban Population	0.319109	-0.060799	0.710973	0.061698	0.343496	0.071925	0.211622	-0.135283	-0.316327	-3.144399e-01
Rural Population	0.297363	-0.444539	-0.250220	-0.265590	-0.314668	-0.389431	0.190977	-0.132343	0.045731	-5.235087e-01
Total Population	0.323299	-0.318027	0.116895	-0.151082	-0.071631	-0.228893	0.210286	-0.141211	-0.095375	7.918751e-01

Each row corresponds to one of the original variables(pollutants and population variables). Each column represents a principal component (PC1, PC2, PC3, etc.). These are new, uncorrelated variables derived from the original data. The values in the table are the loadings, which indicate the strength and direction of the relationship between each original variable and each principal component. The magnitude, represented by the absolute value of the loading, shows how strongly the variable contributes to the principal component. Larger absolute values mean stronger contributions. The sign (positive or negative) indicates the direction of the relationship. A positive loading means that as the principal component score increases, the variable also increases, while a negative loading indicates that the variable decreases as the principal component score increases. For different principal components, the table below shows the key variables contributing.

Principal Components	Key Variables
PC1	OC Emission BC Emission NH3 Emission Urban Population
PC2	NO Emission SO2 Emission
PC3	Urban Population
PC4	NMVOC Emission CO Emission

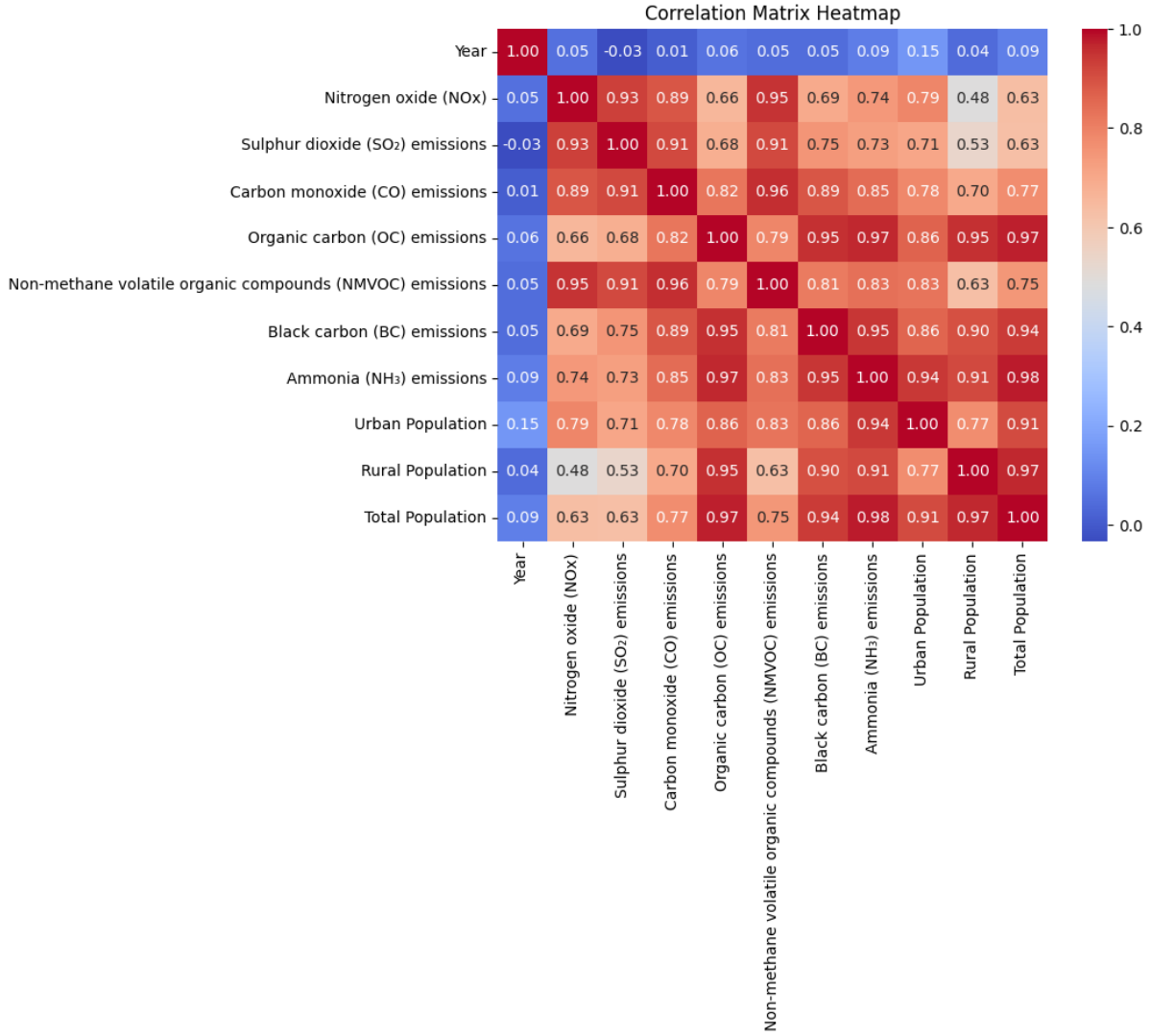
PC5	BC Emission
PC6	OC Emission
PC7	NH3 Emission
PC8	Rural Population
PC9	CO Emission
PC10	Total Population

Considering PC2 for instance, from the rotation matrix it's evident that Nitrogen Oxide (NOx): 0.471876 and Sulphur Dioxide (SO_2) emissions: 0.438684 have the highest positive loadings. So, these two variables are the most contributing in explaining the variance captured by PC2.

The PCA plot of the data in the first two principal components (PC1 and PC2) is a way to visualize the data in a lower-dimensional space while retaining as much of the original information as possible. The X-axis represents the first principal component, which accounts for the largest amount of variance in the dataset. The Y-axis represents the second principal component, accounting for the second largest amount of variance.

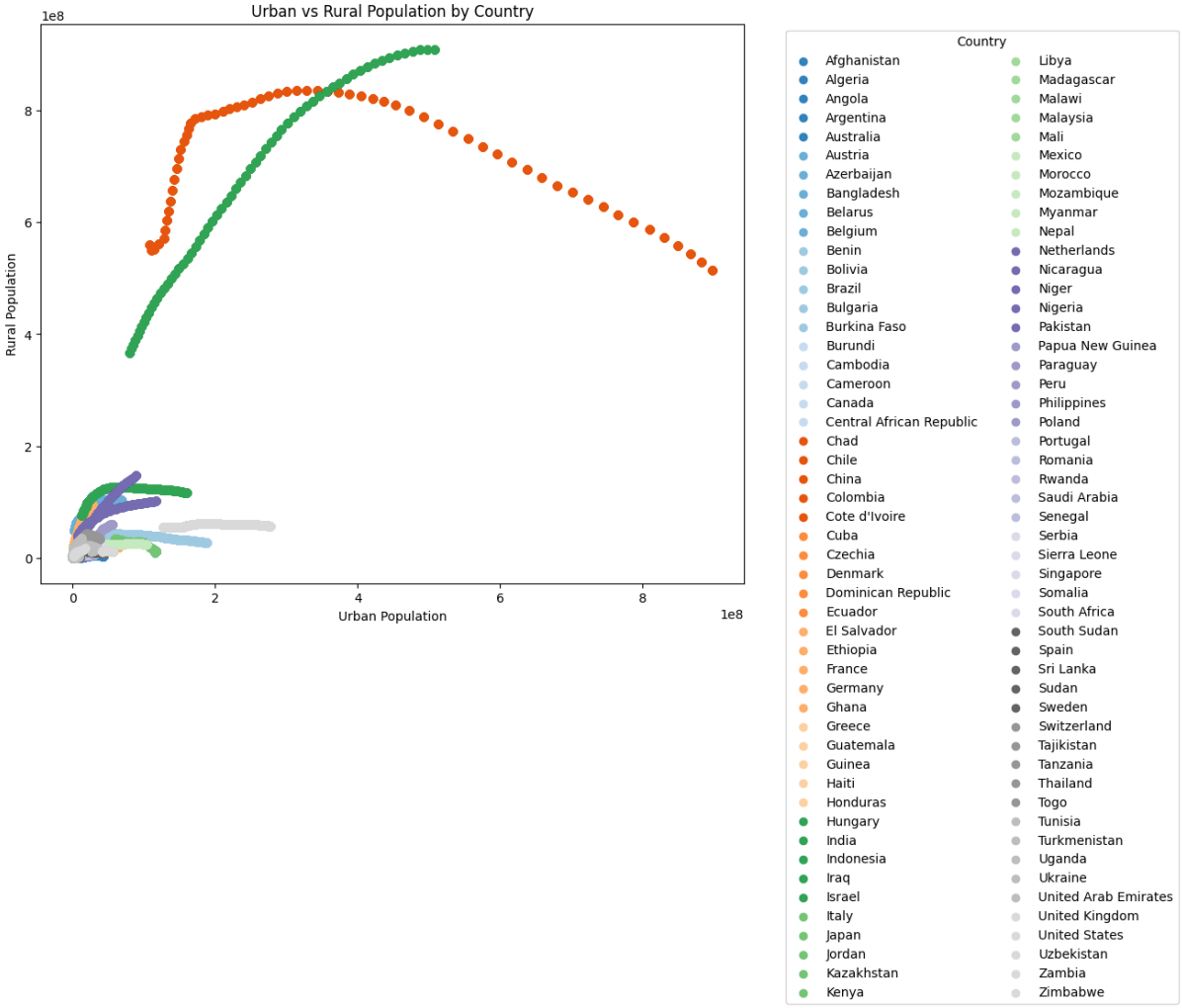
When plotting data using two original features, the plot reflects the relationship between those specific features without considering the overall variance in the data. Original feature plots can show direct relationships (e.g., correlation) between specific pairs of features.

Below is the correlation matrix of all the original features.



PCA combines information from multiple features to maximize the variance captured in the first few components, providing a more comprehensive view of the structure within the data. It reduces dimensionality and helps identify underlying patterns that may not be visible when considering only two original features.

For example, features like "Nitrogen oxides (NOx)" and "Sulphur dioxide (SO2) emissions" have long vectors pointing in similar directions, indicating they are strongly correlated and contribute similarly to the variance in the data. Similarly, all three population variable vectors are pointing toward the same direction which is as expected. While considering carbon related pollutants vectors it is observed that organic carbon(OC) and Black Carbon(BC) are in same direction with population vectors. This reveals a pattern of countries which are densely populated have higher carbon emissions.



Cluster 0: This cluster is characterised by low population and low emissions. Most of the countries during 1960 to 2022 fall under this cluster except few countries like India, China, United States of America which fall under other clusters for most of the years.

Cluster 1: This cluster is characterised by high population and high levels of carbon and ammonia emissions. India was categorized in this cluster for 55 years(1968-2022) and rest of the years in Cluster 0(1960-1967). While China falls under this for 26 years(1960-1985).

Cluster 2: This cluster is characterised as increasing with Sulphur Dioxide(SO2),Nitrogen Oxide(NO) and Carbon Monoxide(CO) and Non -Methane volatile compounds (NMVOC) with these emissions. United States of America falls under this category for 52 years(1960-2011), rest of the years (2012-2022) it was under Cluster 0.

Cluster 3: This cluster is characterised by increasing Urban population and Black Carbon(BC) and Ammonia(NH3) emission. China falls under this cluster for 37 years(1986-2022).

One pattern that is observed is for countries which has drastic increase in population like India and China, reveals the impact of urbanization and industrialization on air quality.

The dendrograms produced from the original and scaled data emphasize the impact of various linkage methods and the necessity of scaling in hierarchical clustering. In the initial data, single linkage is very sensitive to noise and outliers, complete linkage is impacted by clusters with different densities, and average linkage provides a more even clustering method. On the other hand, when data is scaled, the dendrograms show clusters that are more evenly spread out, suggesting that scaling lessens the influence of variables with different scales and improves the accuracy of clustering. These results highlight the importance of choosing the right linkage methods and utilizing scaling techniques for accurate and meaningful clustering outcomes.

Conclusion

This study applied unsupervised learning techniques to analyse air pollution data from 100 countries over the period from 1960 to 2022, integrating data from ourworldindata.org for air pollution and data.worldbank.org for population metrics. Principal Component Analysis (PCA) revealed 10 significant principal components, with the first principal component (PC1) explaining nearly 83.92% of the variance, enabling a more manageable and interpretable dataset. K-means clustering, optimized through the elbow method and silhouette analysis, identified four distinct clusters, providing insights into the natural groupings of countries based on their pollution and population profiles. Hierarchical clustering was performed using various linkage methods, allowing for an effective comparison of groupings. This multi-faceted analysis demonstrated the utility of unsupervised learning techniques in uncovering hidden patterns and relationships within complex datasets. The findings offer valuable insights into global air pollution trends and their association with demographic factors, which can inform policy development and targeted interventions aimed at mitigating air pollution. Overall, this study highlights the power of unsupervised learning in environmental data analysis and lays the groundwork for future research in this critical area.

Bibliography/References

1. https://ourworldindata.org/explorers/air-pollution?time=earliest..2022&uniformYAxis=0&country=USA~CHN~IND~GBR~OWID_WRL~OWID_NAM&Pollutant=All+pollutants&Sector=From+all+sectors+%28Total%29&Per+capita=false
2. https://data.worldbank.org/indicator/SP.URB.TOTL
3. https://data.worldbank.org/indicator/SP.RUR.TOTL
4. https://data.worldbank.org/indicator/SP.POP.TOTL
5. https://www.javatpoint.com/unsupervised-machine-learning
6. https://www.ibm.com/topics/unsupervised-learning

Appendix: