# Project

## Problem formulation

The goal of this project is to analyze global air pollution data using clustering techniques to identify patterns in air quality across countries and cities. This is an unsupervised learning task, specifically clustering, where we aim to group regions with similar air quality profiles without prior knowledge of their categories.

We focused on identifying:

* Which regions (countries/cities) have similar air pollution characteristics.
* What pollutants are dominant in different clusters.
* How different clustering algorithms perform on this dataset.

This type of analysis can support environmental policy-making, urban planning, and public health initiatives.

## Discussion of related works (optional)

What has previously been done by others on this topic?

## EDA and data preprocessing

We loaded the dataset containing 23,463 entries with columns including:

country\_name, city\_name

aqi\_value (overall air quality index)

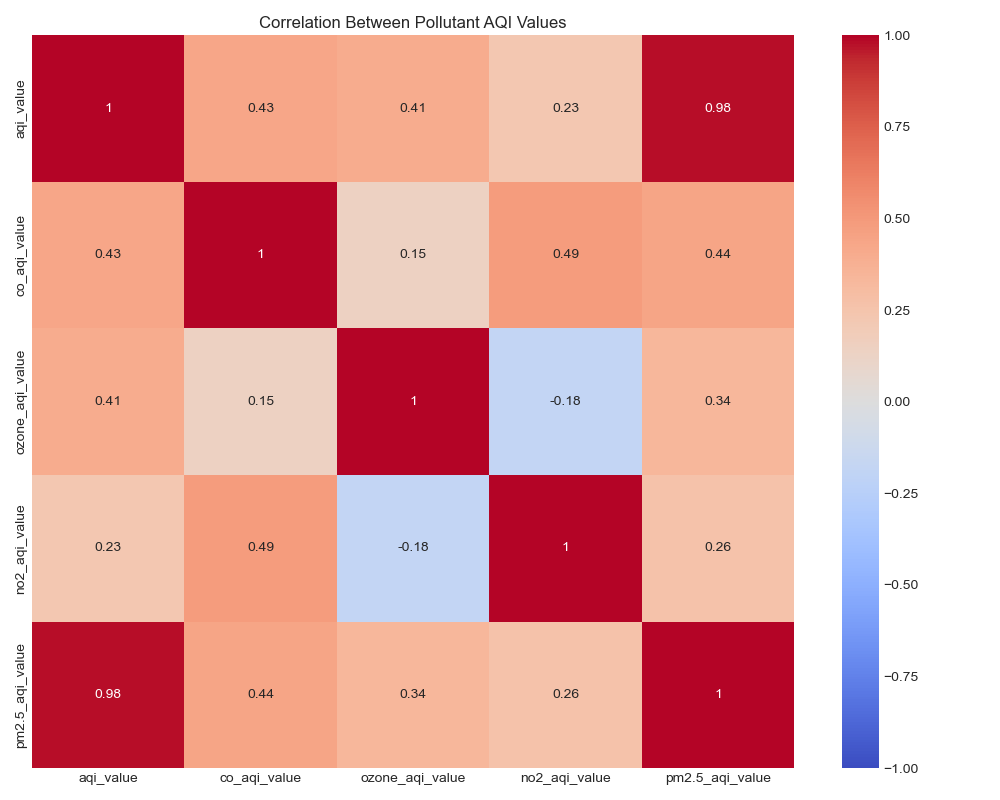
Pollutant-specific AQIs: CO, Ozone, NO₂, PM2.5

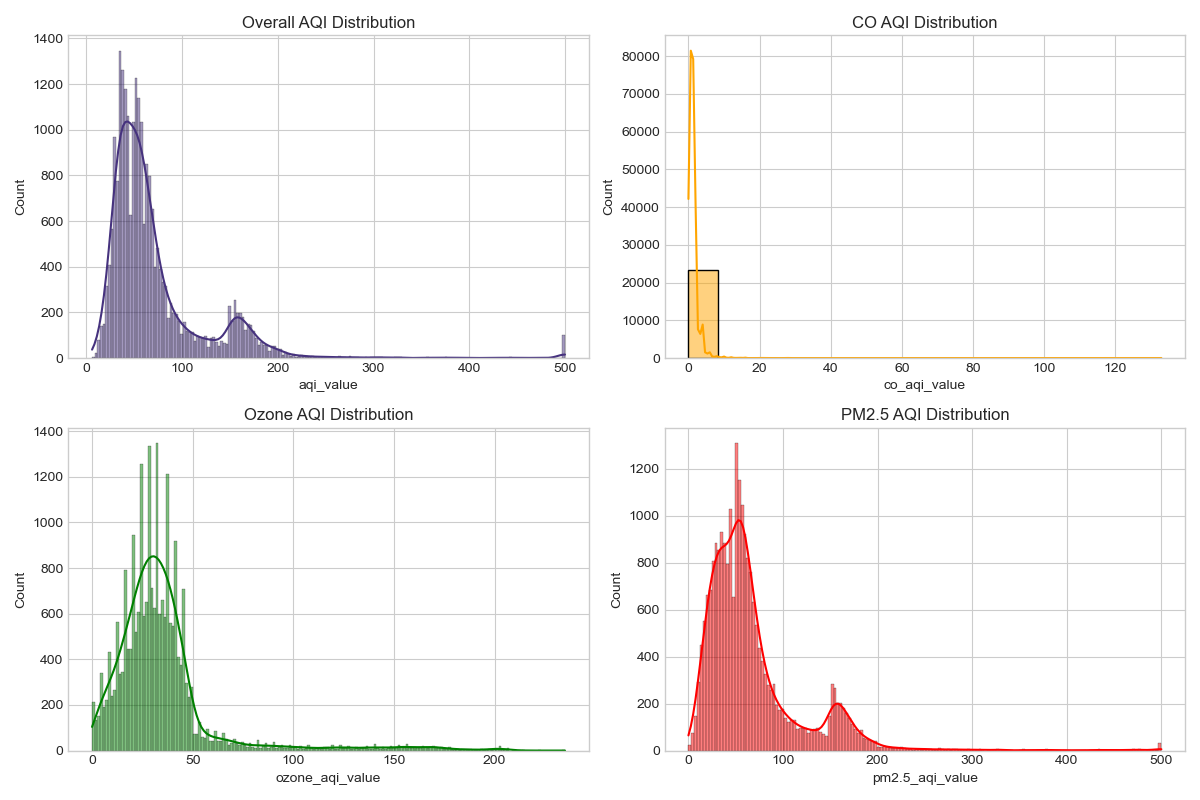
**Key findings from EDA:**

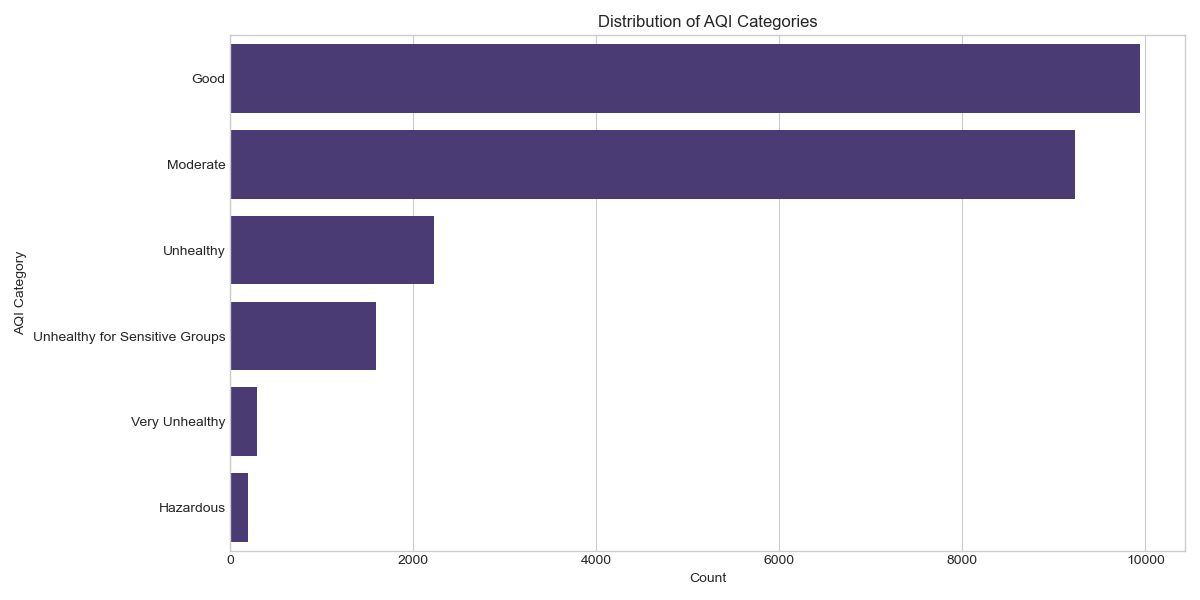
* The dataset contains missing values in country\_name (427) and city\_name (1).
* Most entries fall under "Good" or "Moderate" AQI categories.
* PM2.5 was identified as the most impactful pollutant.
* High correlation between overall AQI and PM2.5 AQI.

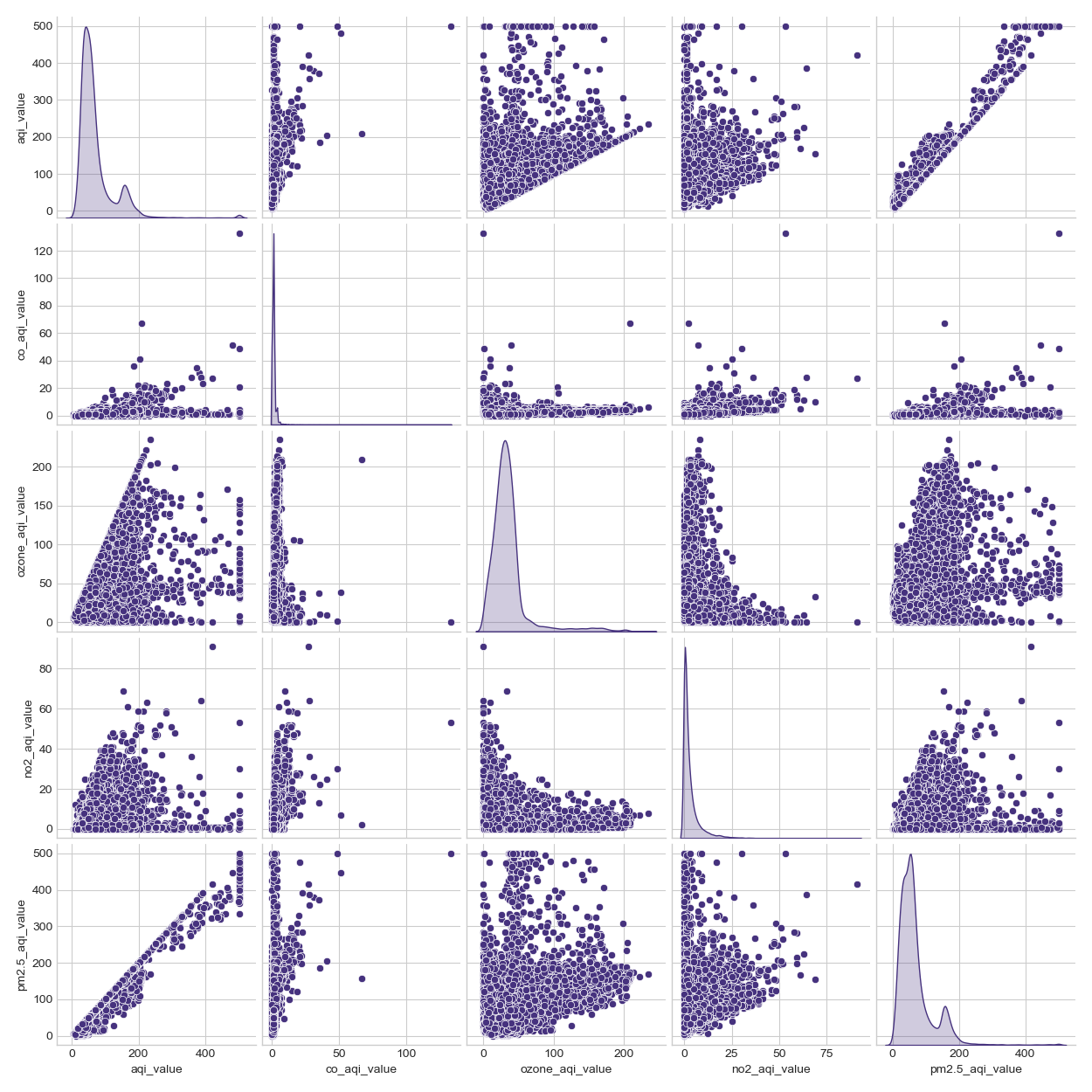
**Visualizations included:**

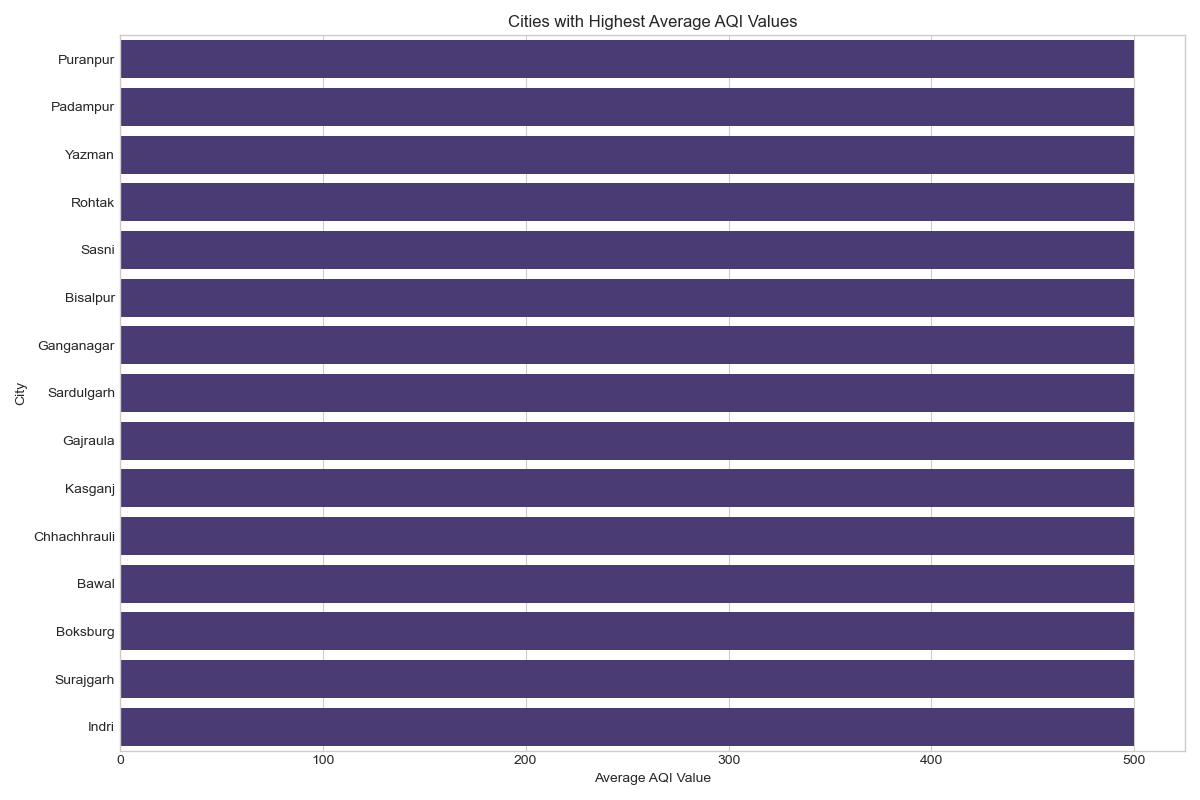
* Distribution of AQI categories
* Histograms of pollutant values
* Correlation heatmap
* Average AQI by country and city
* Pairplot of numerical features

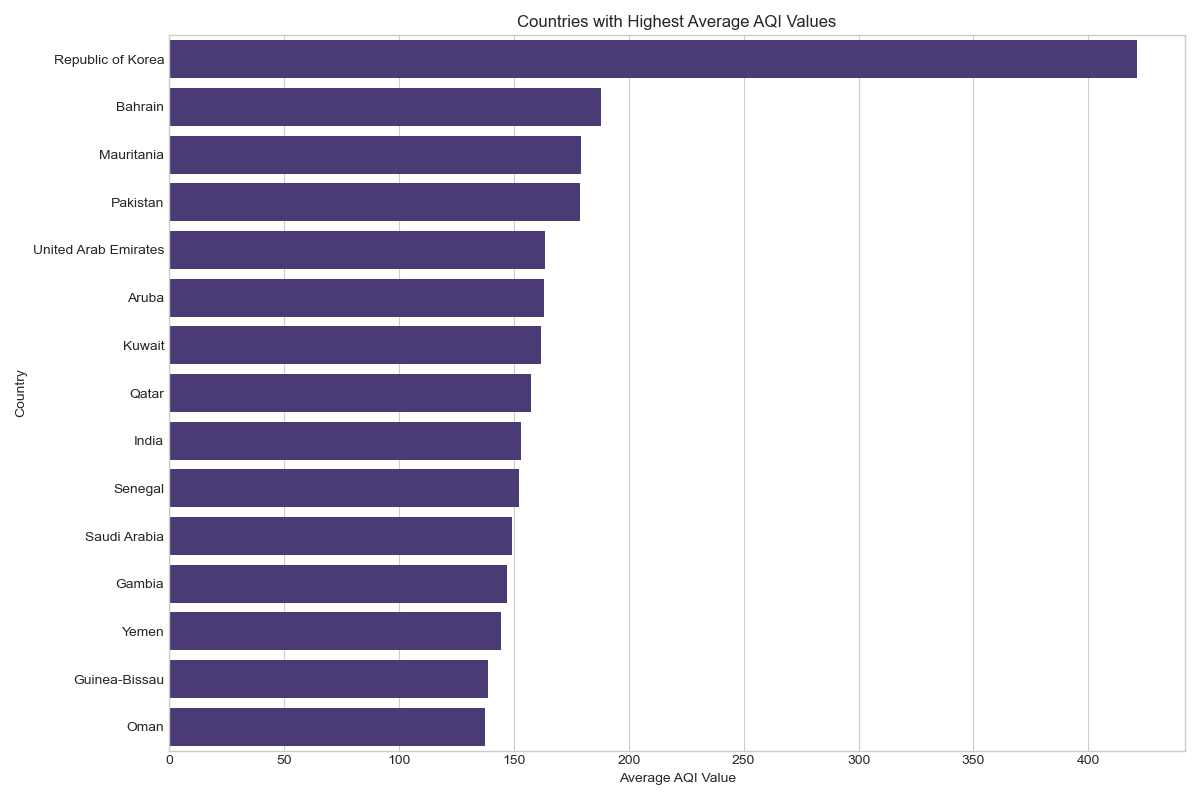












**Data Preprocessing Steps**

1. Feature Selection: Selected numerical AQI values for clustering: aqi\_value, co\_aqi\_value, ozone\_aqi\_value, no2\_aqi\_value, pm2.5\_aqi\_value.Missing Value Handling: Filled missing values with column means.
2. Outlier Detection & Capping: Used IQR method to detect and cap outliers (e.g., 12.51% of AQI values were outliers).
3. Standardization: Scaled all features using StandardScaler.
4. Dimensionality Reduction: Applied PCA for visualization; explained variance ratio was [55%, 32%].

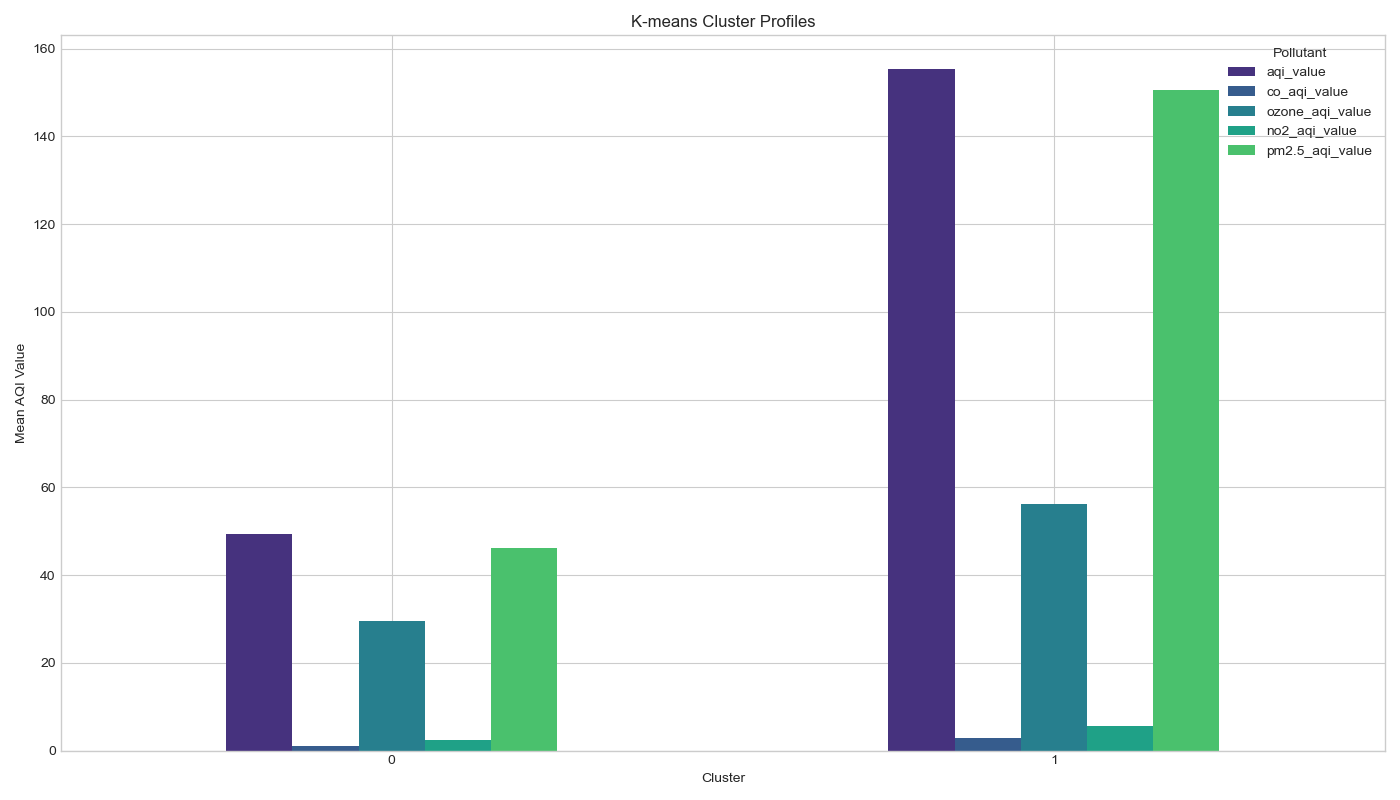
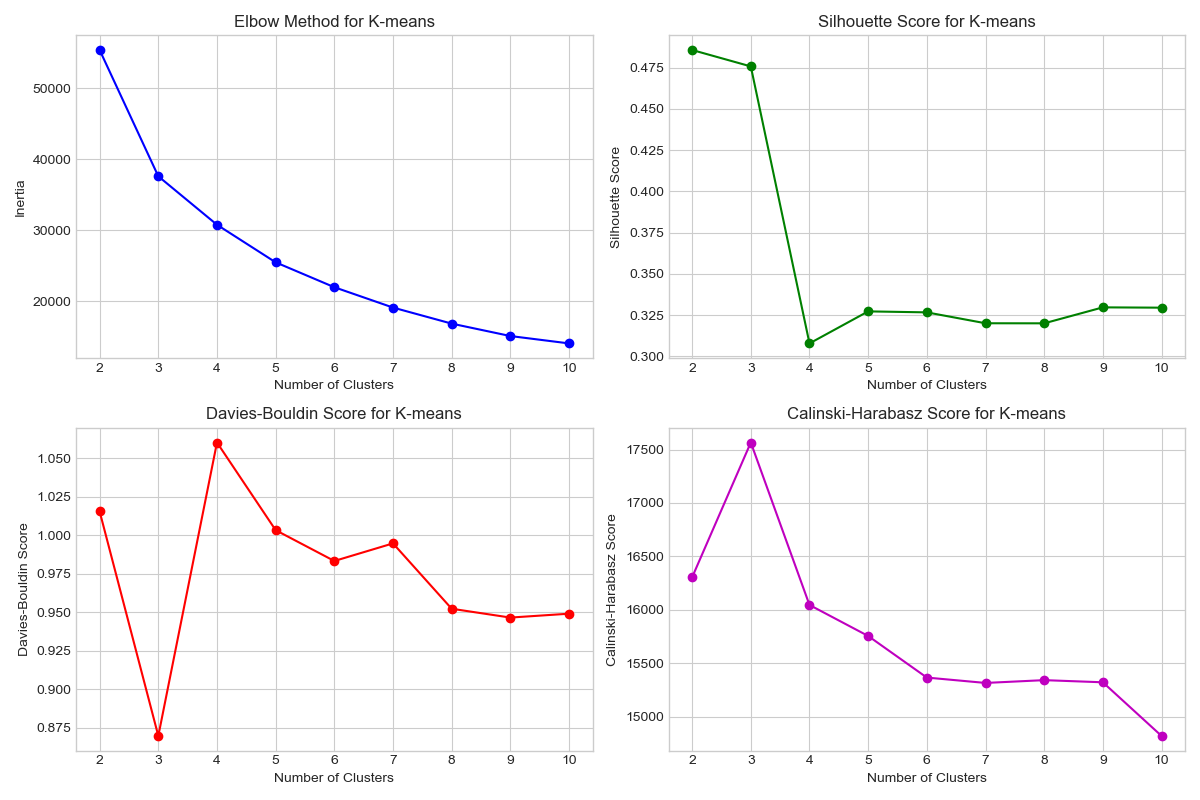
## Mining

**Algorithm Selection**

We applied three clustering algorithms:

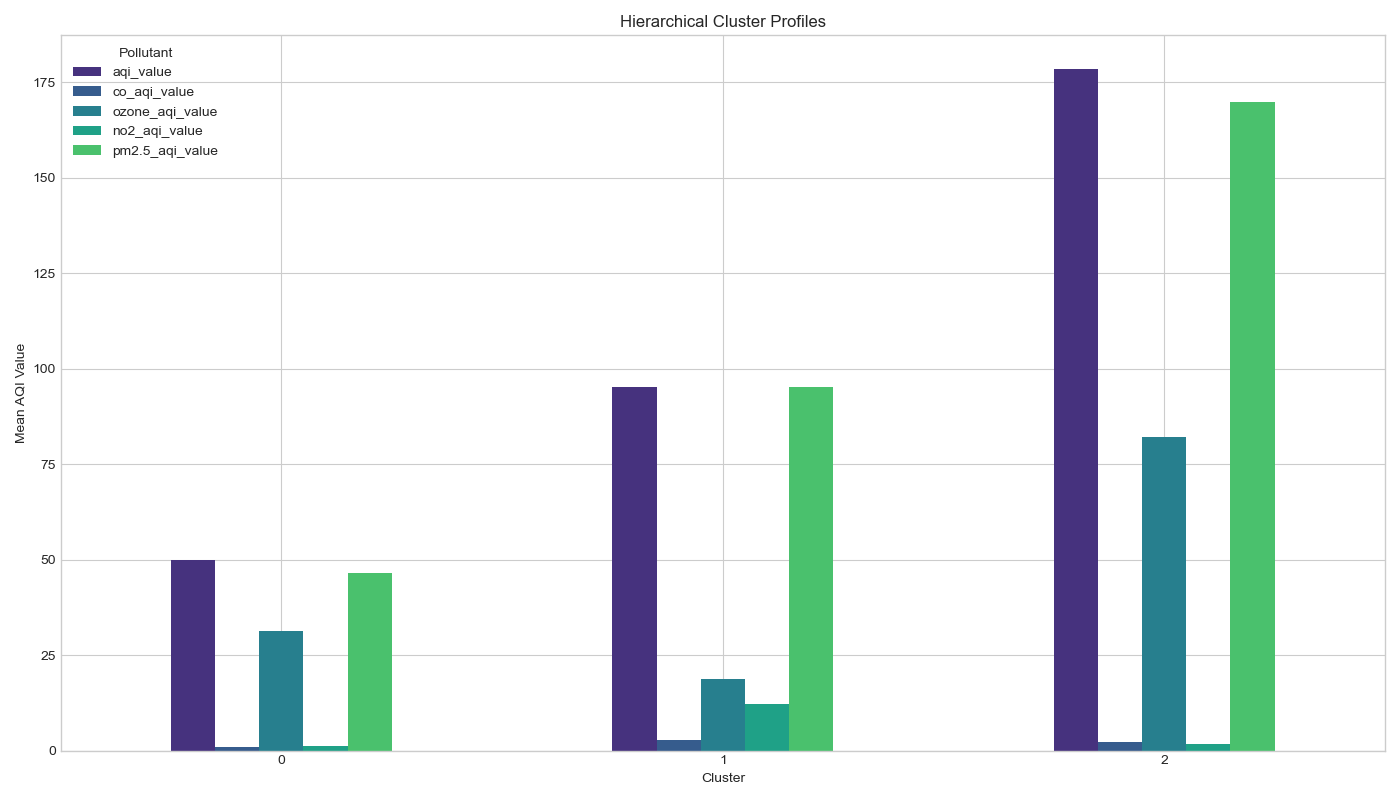
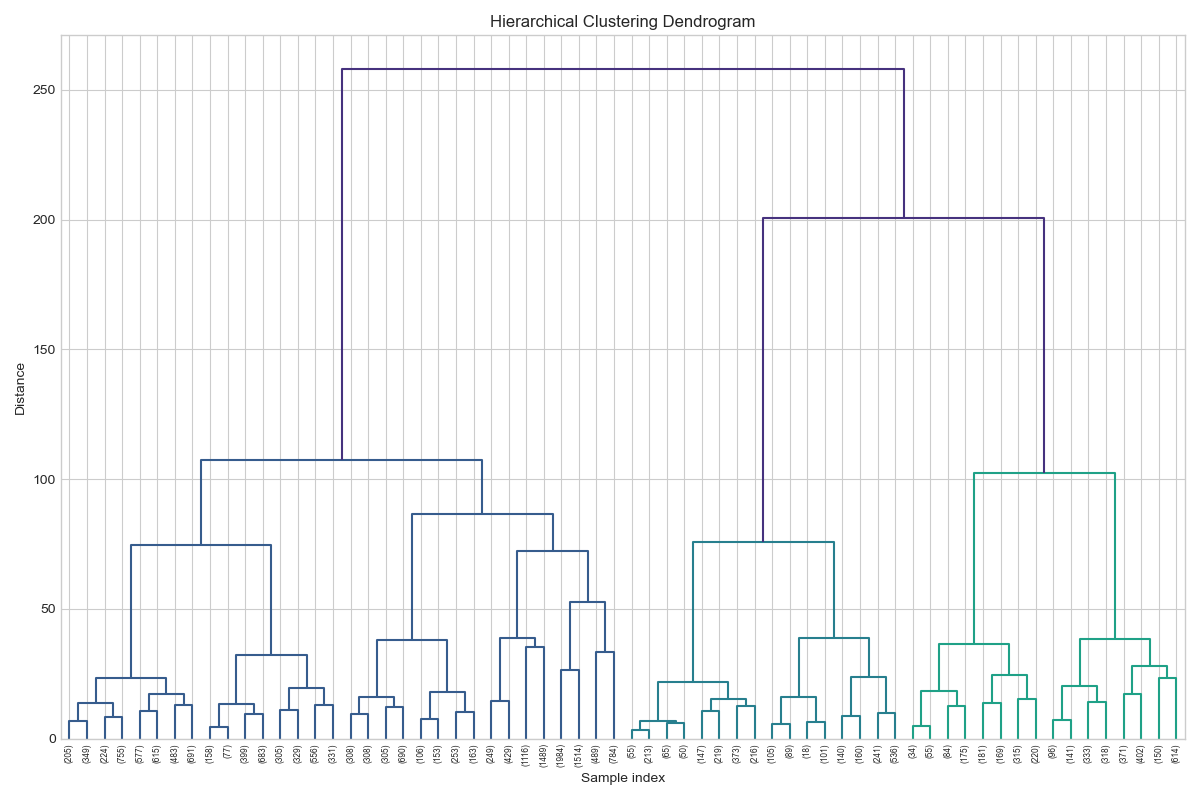
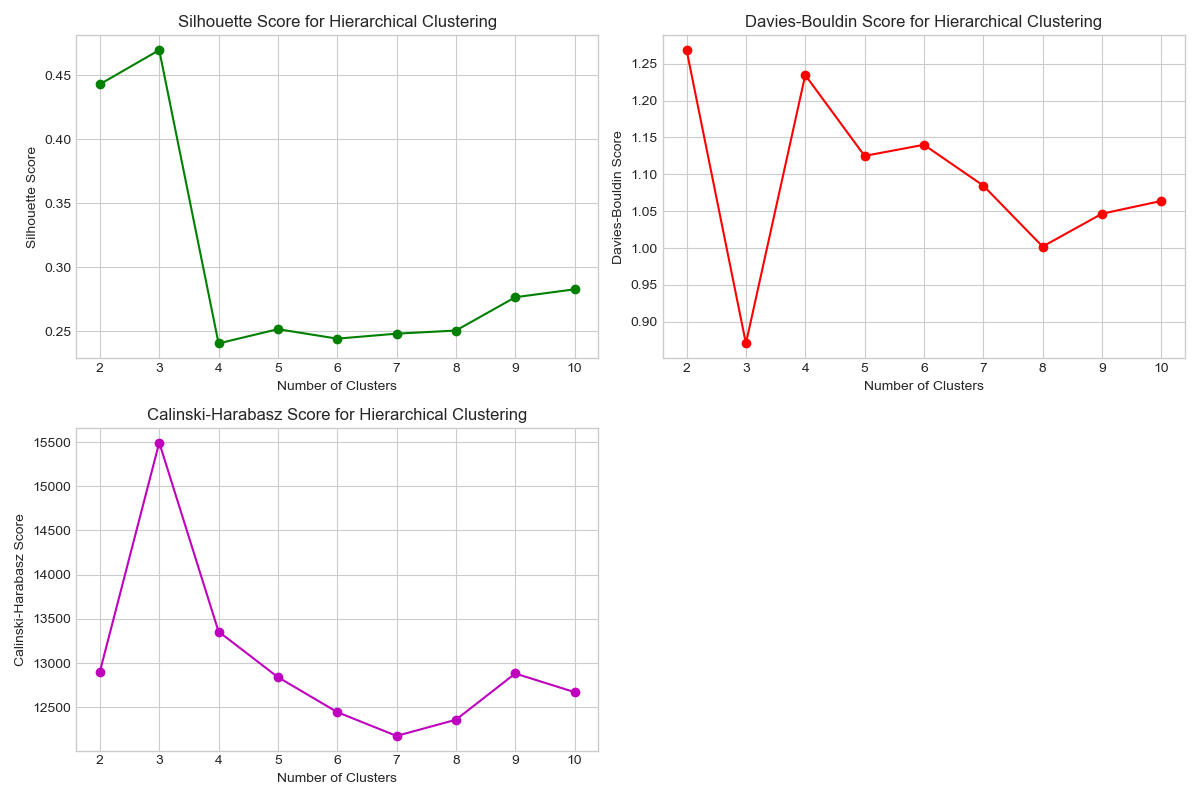
1. K-means Clustering

* Used Elbow method and Silhouette Score to determine optimal k.
* Final model had 2 clusters.



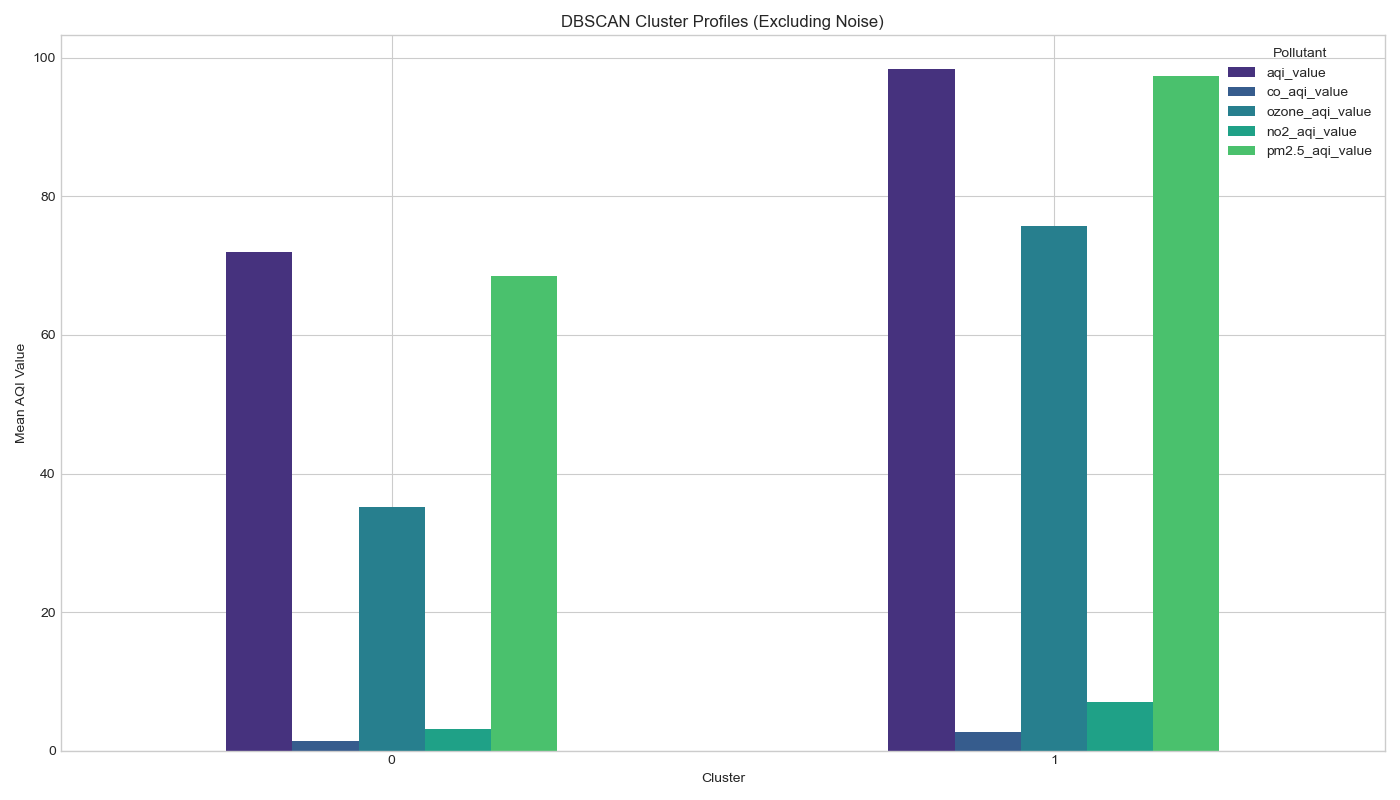
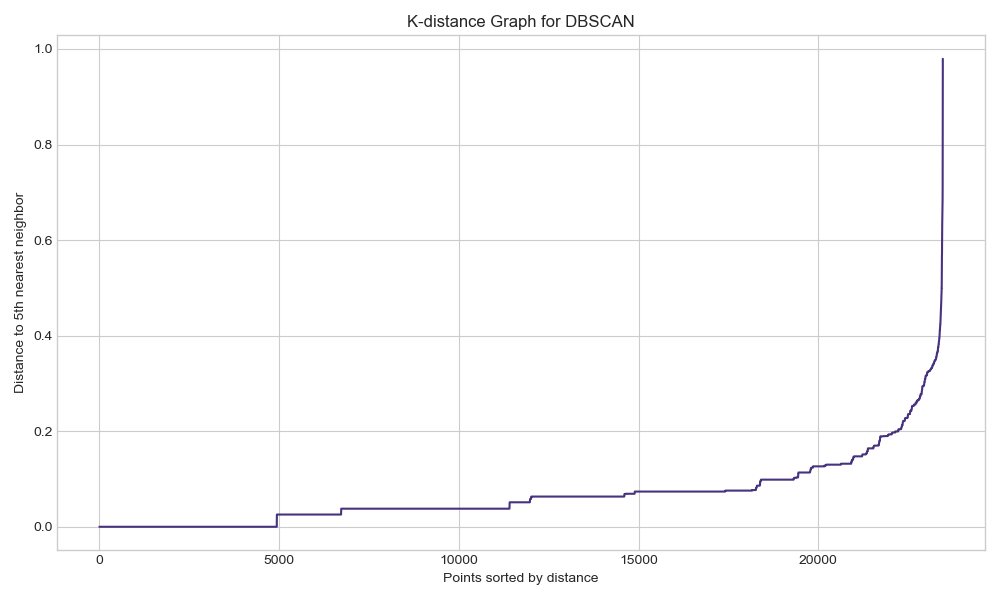
1. Hierarchical Clustering (Agglomerative)

* Dendrogram suggested natural grouping.
* Optimal number of clusters = 3.



1. DBSCAN

* Determined optimal parameters via k-distance graph.
* Best configuration: eps=0.52, min\_samples=3.



**Evaluation Metrics**

We evaluated models using:

* Silhouette Score (best for K-means: 0.486)
* Davies-Bouldin Index
* Calinski-Harabasz Index

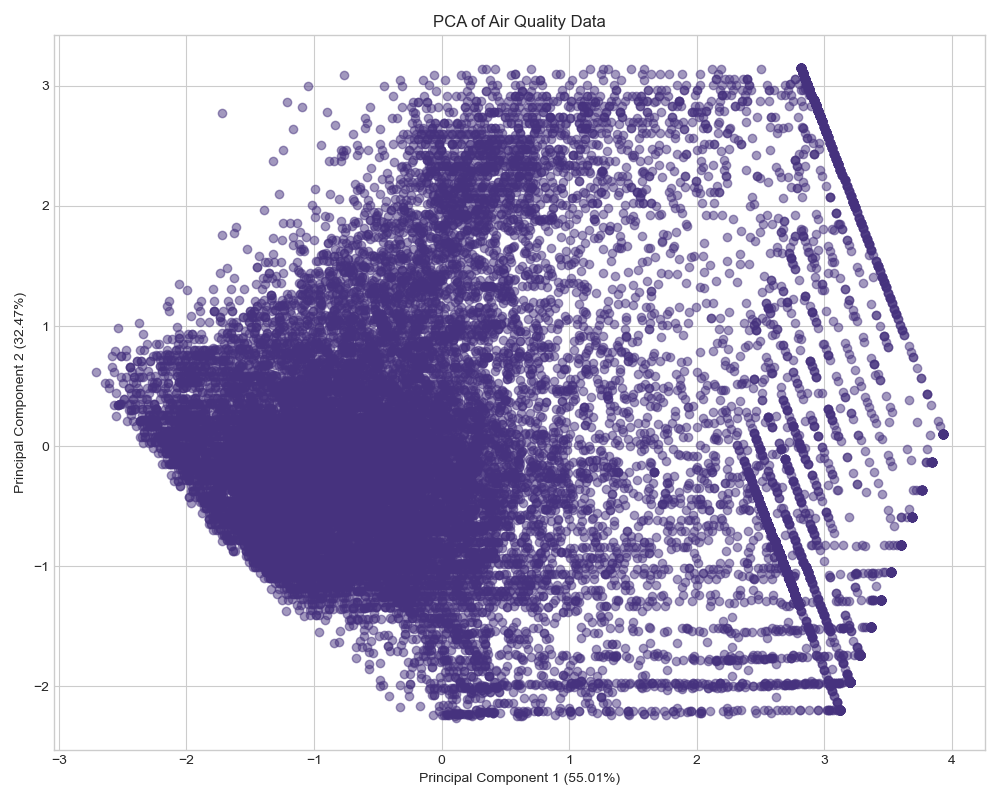
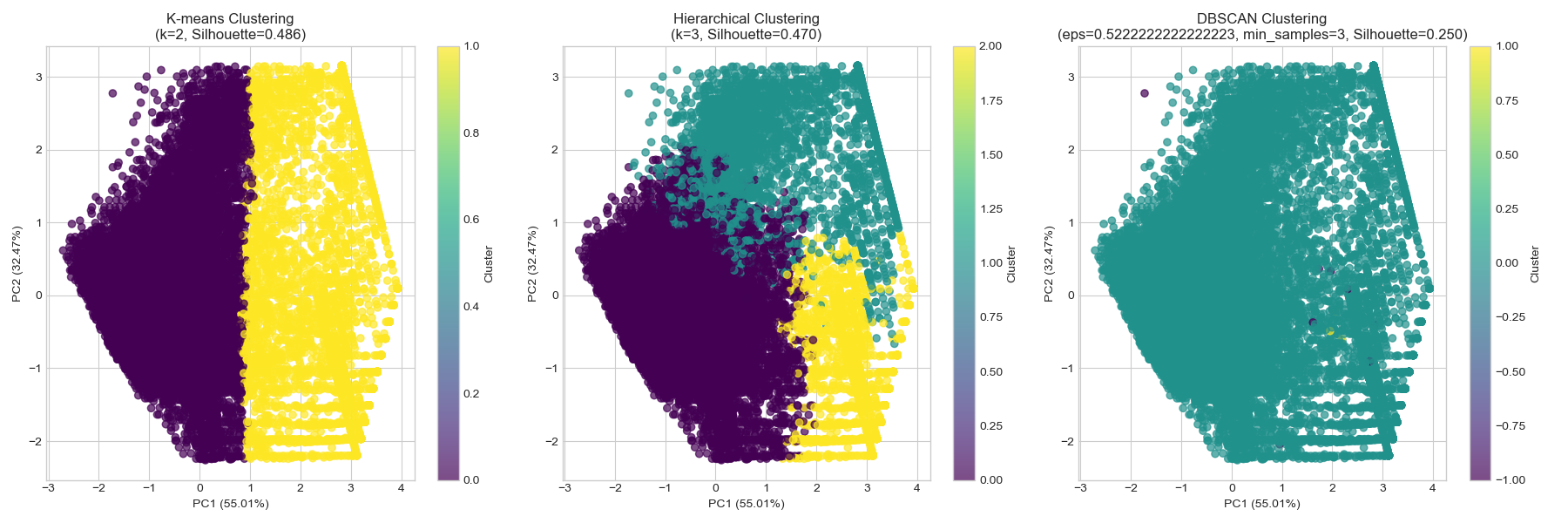
## Experiments

We ran multiple experiments to compare clustering performance:

1. Elbow Method vs Silhouette Score for K-means optimization.
2. Dendrogram-based selection for hierarchical clustering.
3. Grid search over DBSCAN parameters (eps, min\_samples).

**Visualizations**

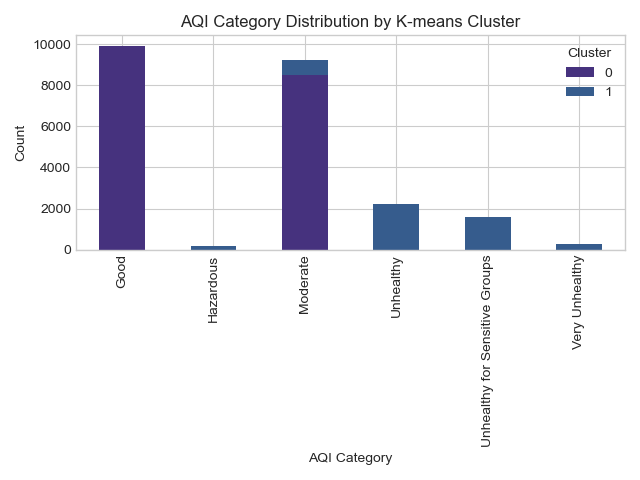
* PCA plot of data distribution
* Elbow plots, Silhouette score curves
* Dendrogram for hierarchical clustering
* K-distance graph for DBSCAN
* Cluster profile bar charts
* Geographic distribution of clusters



## Discussion of results

**Key Findings**

* K-means outperformed other algorithms, indicating that global air quality patterns can be modeled effectively using centroid-based clustering.
* Two main groups emerged:
  + Cluster 0: Low pollution, dominated by developed countries (USA, Germany, Brazil).
  + Cluster 1: High pollution, dominated by developing nations (India, China, Pakistan).
* PM2.5 was the primary pollutant in both clusters, highlighting its significance in air quality assessments.
* There is a clear geographic pattern in air pollution levels, which could inform targeted interventions.



**Limitations**

* Time-based trends were not analyzed due to lack of temporal data.
* Some noise points were identified by DBSCAN but not further explored.