K-Means Clustering Report

Objective

The goal of this analysis was to group air quality observations into clusters using the K-Means clustering algorithm and determine the optimal number of clusters for effective segmentation.

Methodology

Step 1: Data Preparation

The dataset was preprocessed to ensure quality and consistency:

- Missing values were handled appropriately.
- Features related to pollutant levels and sensor readings were selected.
- The data was standardized using `StandardScaler` to normalize feature scales.

Step 2: Determining the Optimal Number of Clusters

To identify the optimal number of clusters:
1. The Elbow Method was applied, where the inertia values (sum of squared distances of samples to their closest cluster center) were computed for different numbers of clusters.
2. A plot of inertia vs. the number of clusters was generated to identify the "elbow point," where the rate of decrease in inertia slows down.

Step 3: Clustering

K-Means clustering was performed using the optimal number of clusters derived from the Elbow Method.

Step 4: Visualization

The clustered data was visualized using a 2D scatter plot in PCA-transformed space (two principal components).

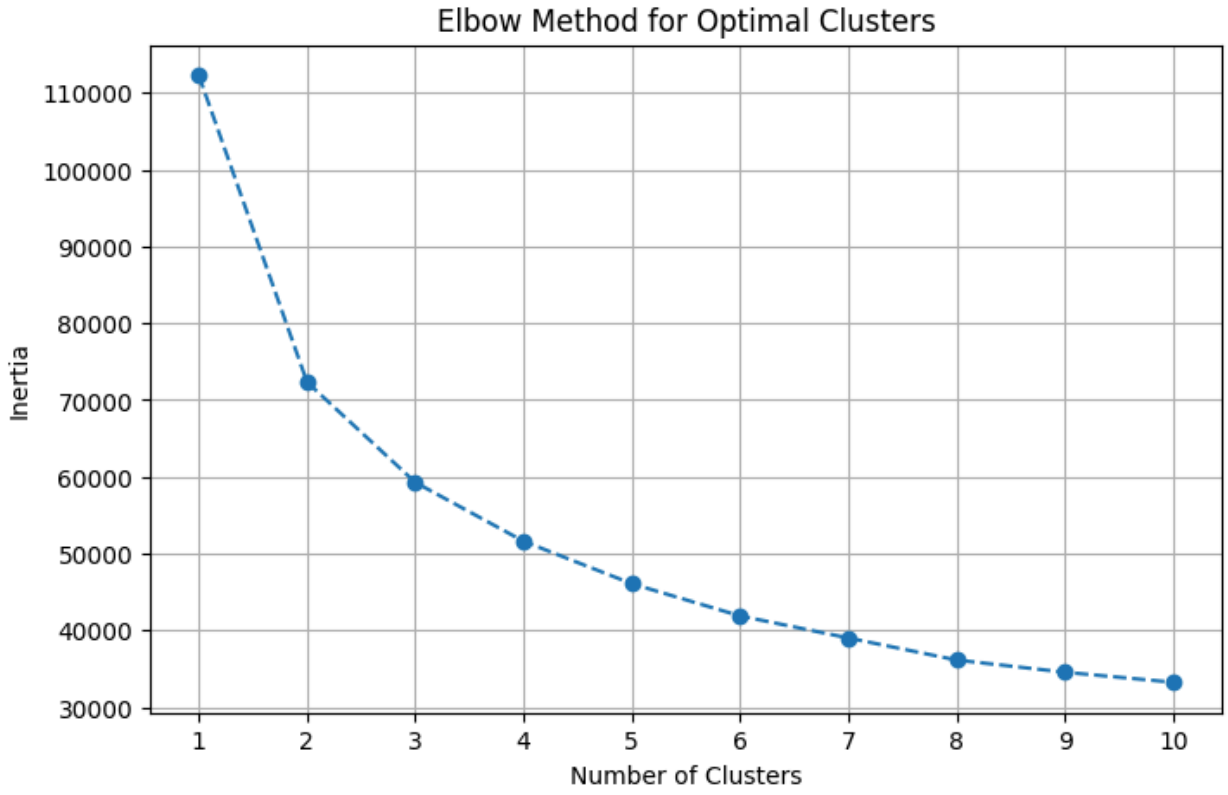
Results

Optimal Number of Clusters

Based on the Elbow Method, the optimal number of clusters was determined to be 4. This was identified as the point where the inertia plot showed a significant decrease in variance reduction.

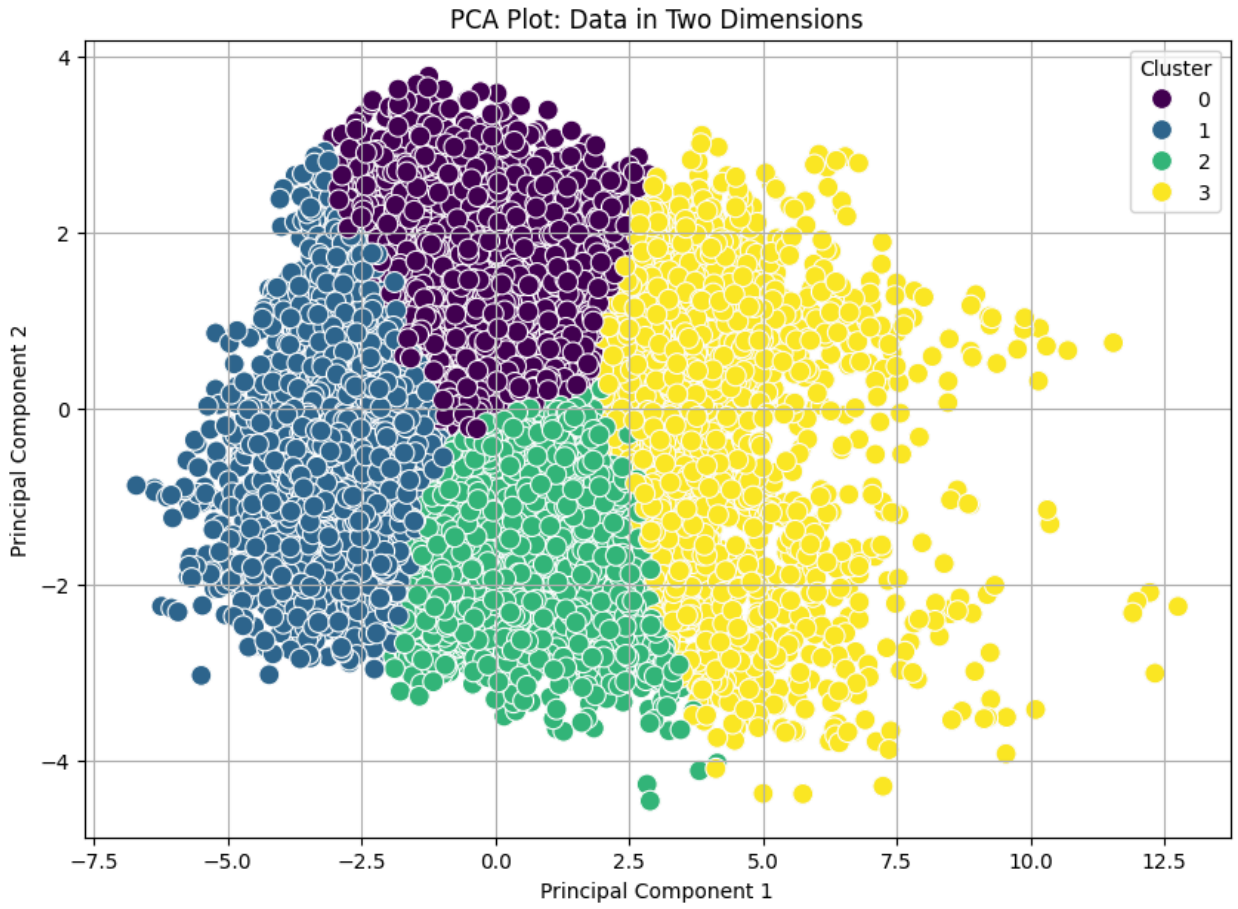
Visualizations

Inertia Plot (Elbow Method)



- A line plot was generated showing inertia values for 1 to 10 clusters.
- The elbow point was observed at 4 clusters, indicating the most suitable segmentation.

PCA Plot

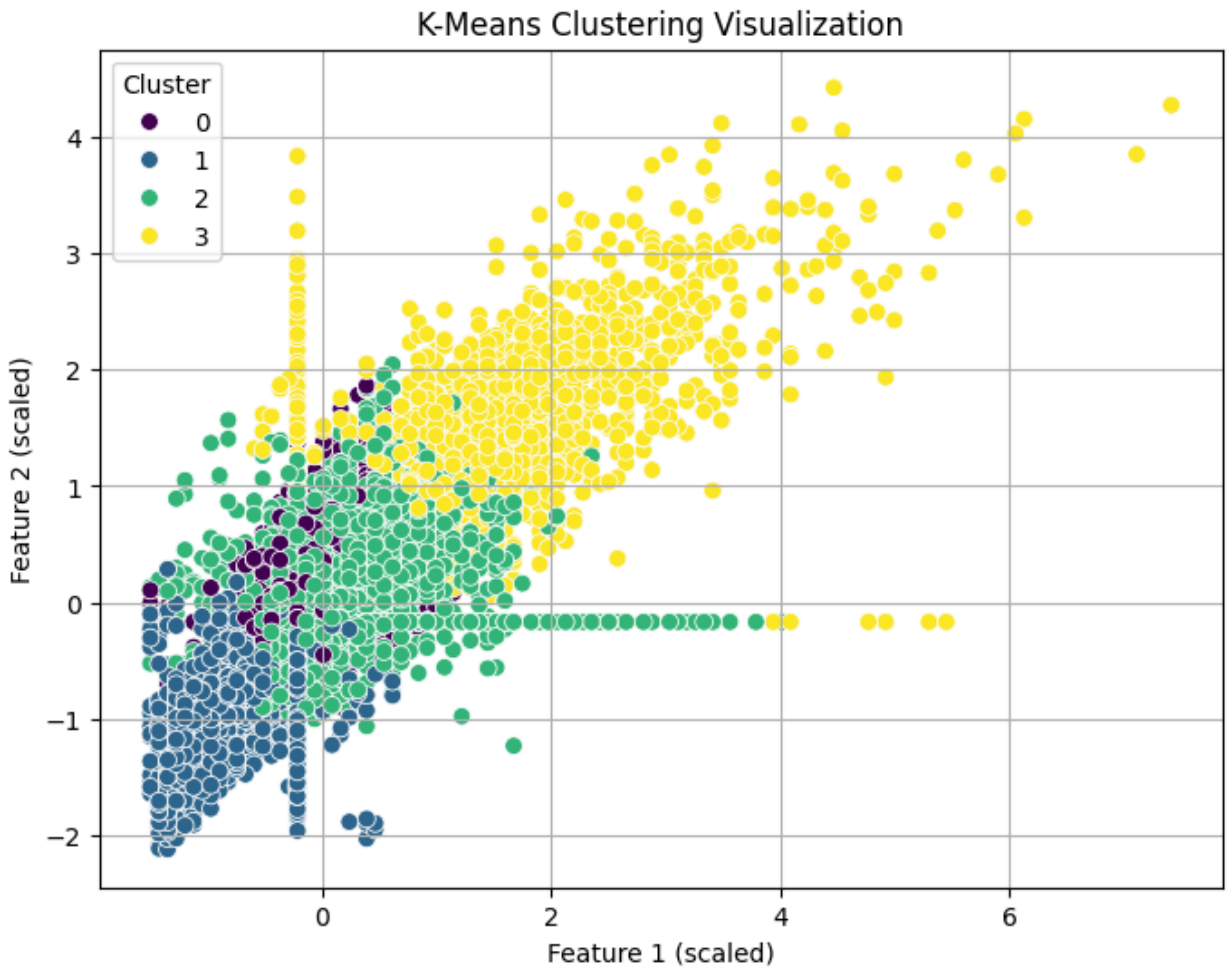


Summary of the clustering results:

- Cluster Centers: Four distinct centroids were formed.
- Cluster Distribution: Data points were grouped into 4 clusters, visualized in a 2D PCA plot.

PCA-Transformed Scatter Plot

The data points were visualized in PCA-reduced space, showing clear groupings of clusters. Each cluster was represented with a unique color.



Conclusion

The K-Means clustering successfully segmented the air quality data into four distinct clusters. These clusters can be analyzed further for insights into air quality patterns and anomalies.