RISHIKA RAVICHANDRAN

DATA SCIENCE INTERN

SAPIENCE EDU CONNECT PVT. LTD.

WEEK 3

**ANALYSIS REPORT**

1. **K-Means clustering**

**Data Preprocessing**

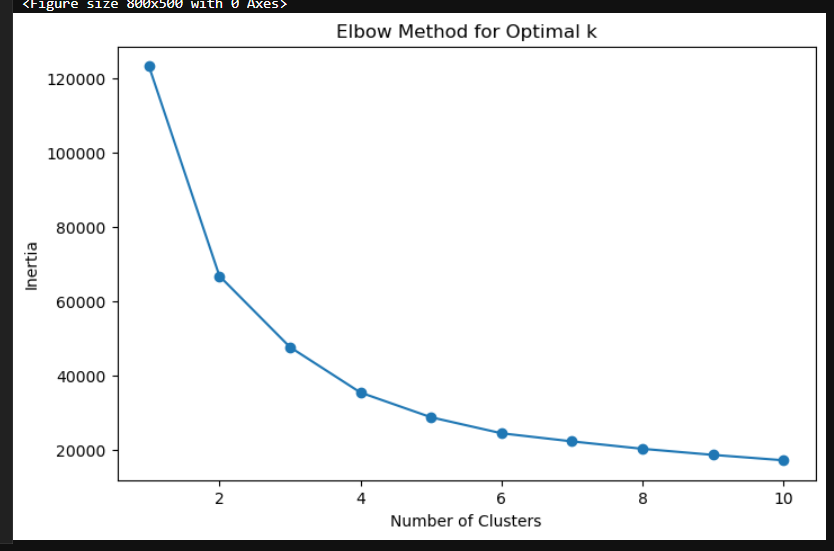
Before applying clustering, the dataset was cleaned and preprocessed:

* Non-numeric columns (such as date and time) were removed.
* Missing values were handled by replacing them with the median of the respective columns.
* Relevant features, including pollutant levels and sensor readings, were selected.
* Standardization was applied to normalize the feature values.

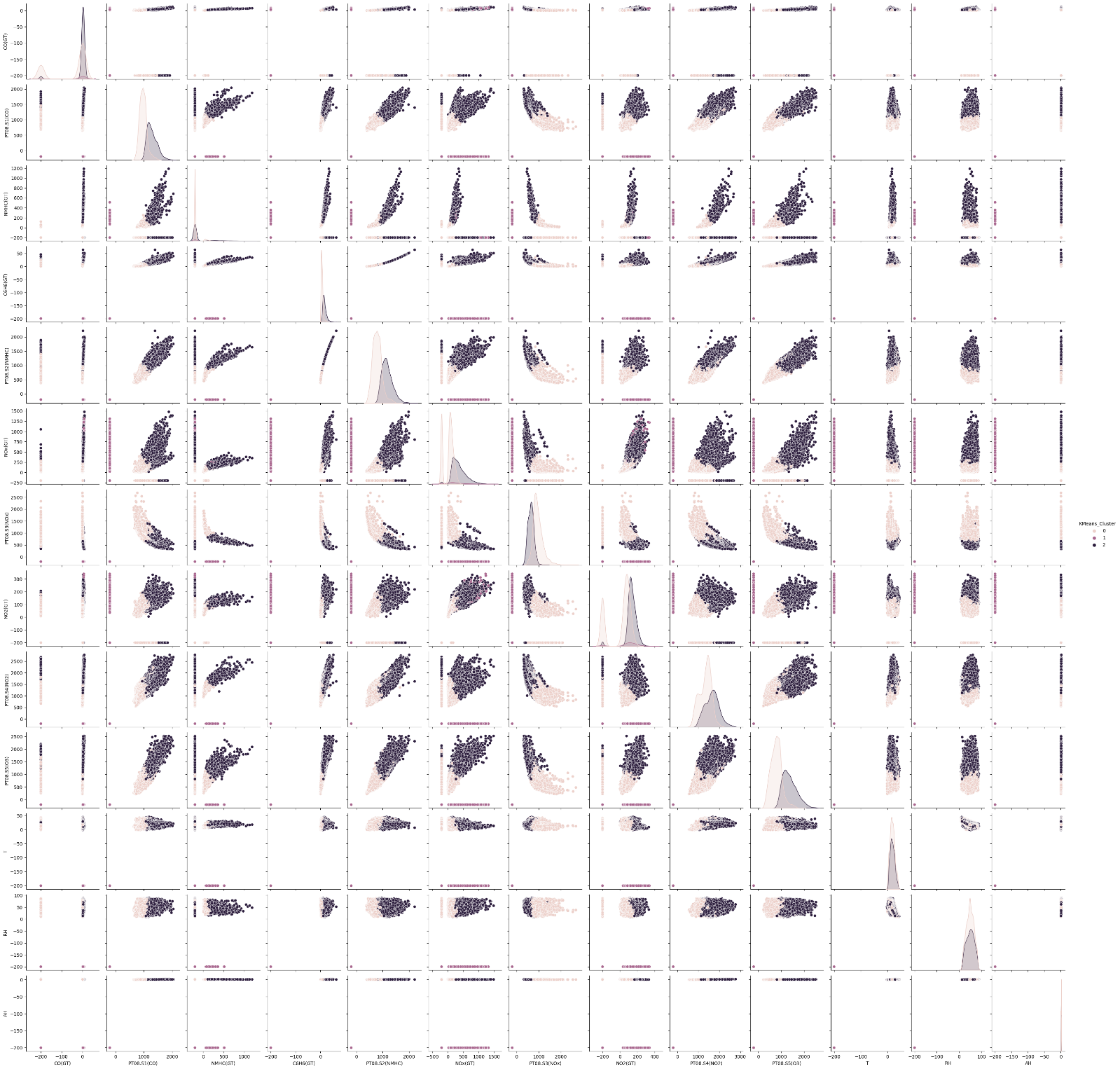
**Optimal Number of Clusters**

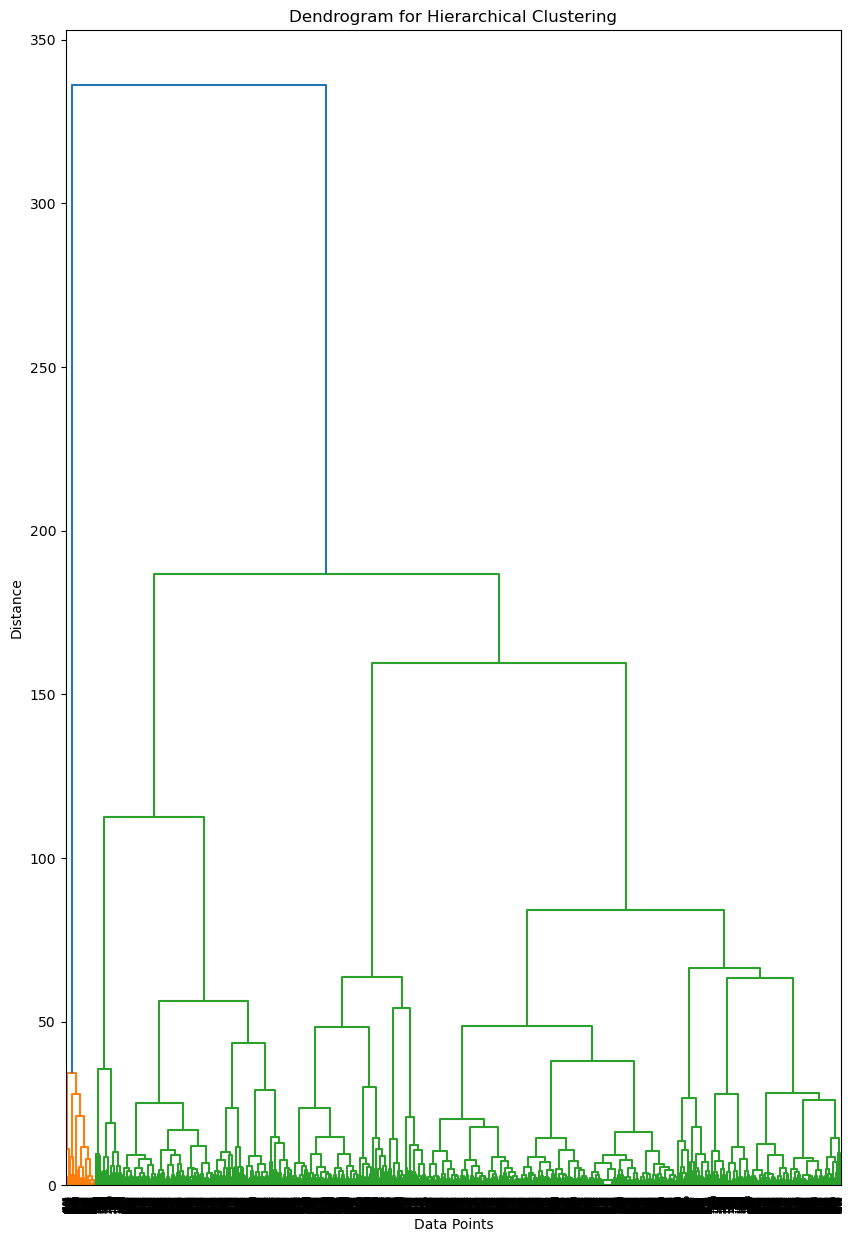
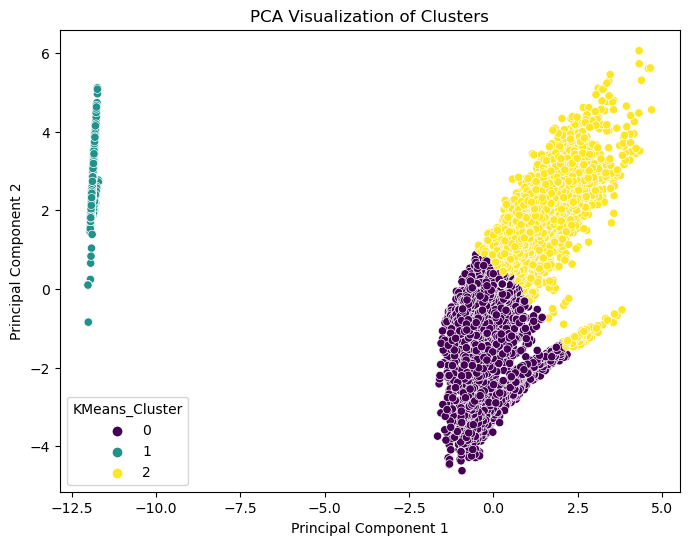
The optimal number of clusters was determined using the **Elbow Method**:

* The inertia i.e., the sum of the squared distances to the nearest cluster center, was calculated for cluster numbers ranging from 1 to 10.
* The Elbow Method plot was used to identify the point where inertia starts to decrease at a slower rate.



* The optimal number of clusters : 3
* One cluster represented good air quality with low pollutant levels.
* Another cluster contained moderate pollution with mid-range pollutant values.
* The third cluster indicated poor air quality, characterized by high levels of pollutants.
* A pair plot is used to demonstrate how different features correlated within the clusters, revealing relationships between pollutants.

**Pairplot:**

1. **Hierarchical clustering dendrogram**
2. **PCA plot**
3. **Anomaly Detection**

* Anomalies in air quality data represent unusual or extreme pollution levels that deviate significantly from normal observations. These anomalies may indicate sensor errors, or sudden environmental changes
* The dataset was transformed using PCA to reduce the number of features and enable better visualization.
* The dataset was grouped into clusters to identify patterns and compute distances between data points and their cluster centroids.
* Points farthest from their assigned cluster centroids were flagged as anomalies. The threshold for anomalies was determined based on the top percentile of distances.
* 474 anomalous data points were detected.
* A significant portion of anomalies was located at the edges of the cluster distributions.
* Some anomalies appeared in isolated regions, suggesting extreme pollution events or potential sensor faults.
* The anomalies were highlighted in red in the PCA-transformed scatter plot.
* Further investigation is needed to classify whether anomalies are true pollution spikes or erroneous readings.

