# Air Quality Clustering & Anomaly Detection Report

**Internship at Sapience Edu Connect Pvt Ltd**

**Name:** Muhammad Yasir

**Internship:** Data Science intern

**Duration**: 15-june-2025 To 14-july-2025

**Week-03 Task:**

**Dataset**: *https://archive.ics.uci.edu/dataset/360/air+quality*

This dataset contains information about various air pollutants (e.g., CO, NOx) recorded over time, making it suitable for unsupervised machine learning techniques.

## Cleaned and Preprocessed Dataset:

The dataset was cleaned and preprocessed as follows:  
- Removed non-numerical and irrelevant columns  
- Handled missing values by dropping incomplete rows  
- Standardized features using `StandardScaler` for consistent scaling across all variables  
- Final dataset shape: (N rows × M features) (replace with actual values)  
- Dataset saved as: cleaned\_air\_quality\_data.csv

## 2. K-Means Clustering Report:

### Objective:

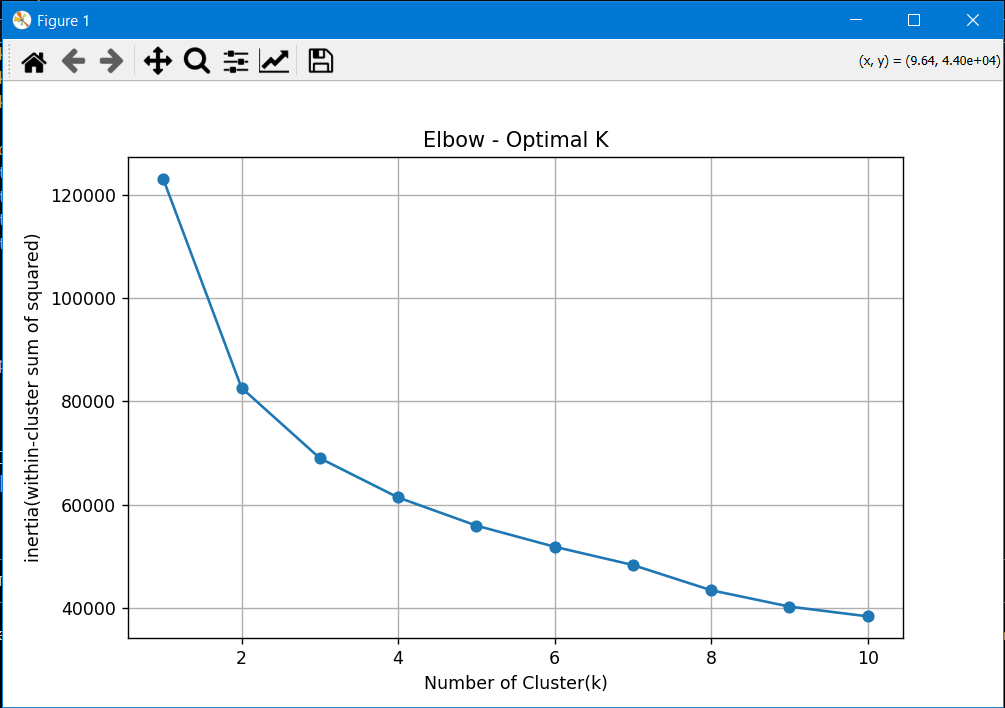
To identify distinct clusters within the air quality data.

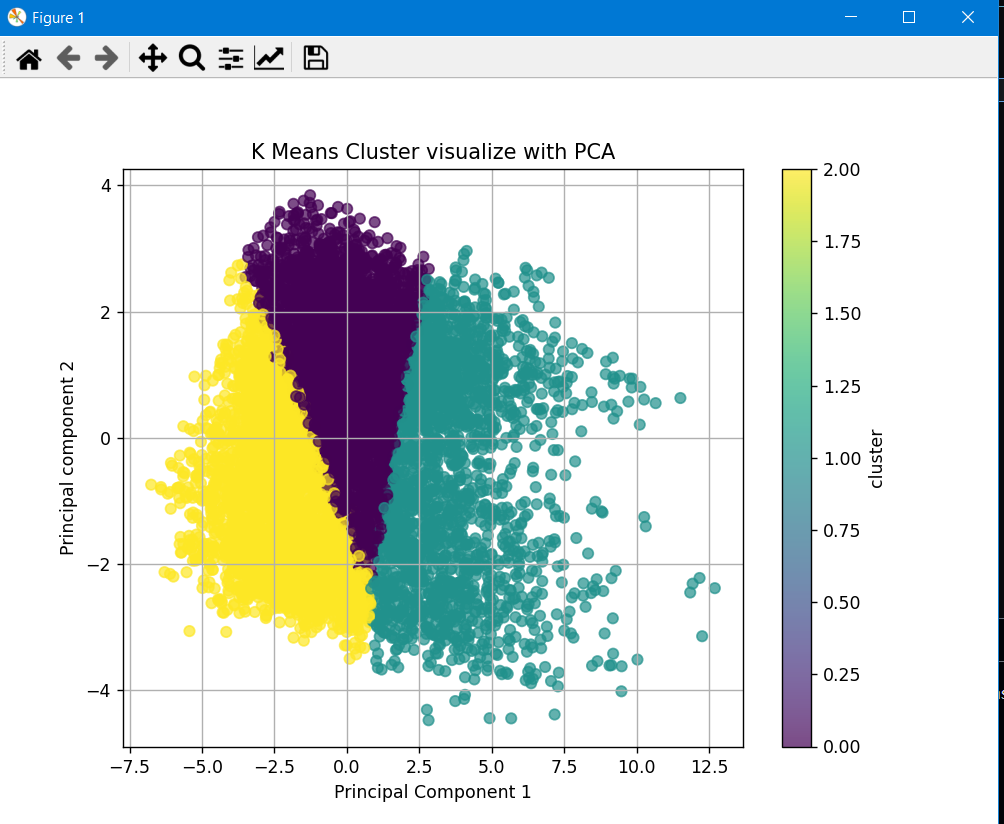
### Method:

- Applied K-Means clustering on standardized data  
- Used the Elbow method to determine optimal number of clusters  
- Optimal number of clusters: 3 (determined using inertia plot)

### Visualization:

- Cluster labels added to dataset  
- PCA plot generated showing data points colored by cluster assignment





### 

### Tools Used:

- sklearn.cluster.KMeans  
- matplotlib, seaborn, PCA from sklearn.decomposition

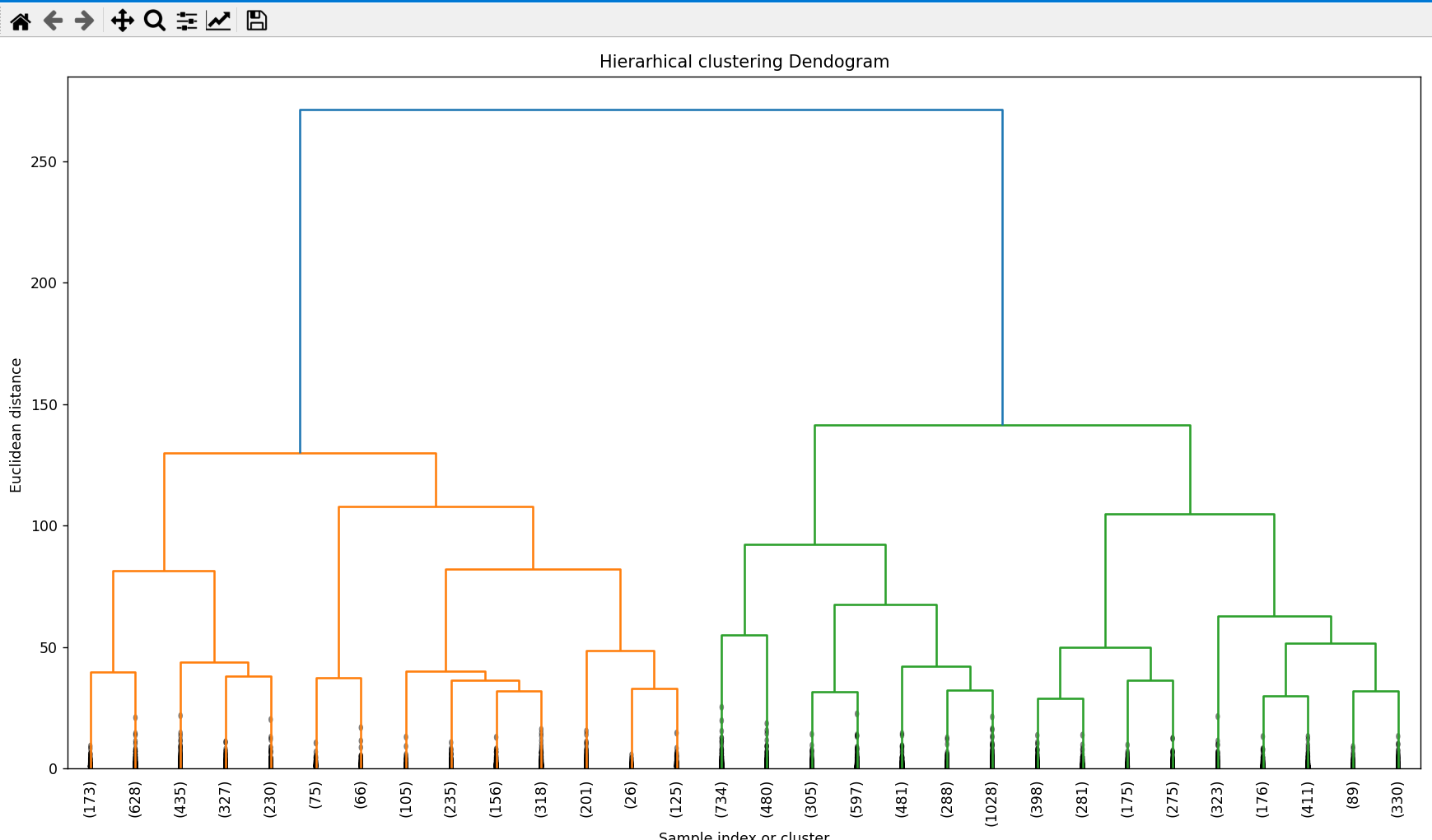
## 3. Hierarchical Clustering (Dendrogram):

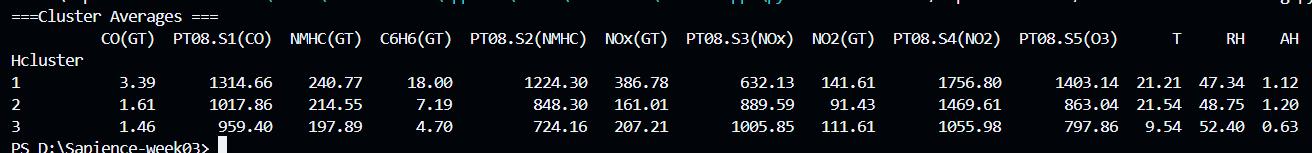
### Objective:

To visualize the nested relationships between air quality readings.

### Output:

- Dendrogram plot shows how data points merge into clusters  
- Helped validate the number of clusters visually (aligned with KMeans = 3)





### Tools Used:

- scipy.cluster.hierarchy

## 4. PCA Dimensionality Reduction:

### Purpose:

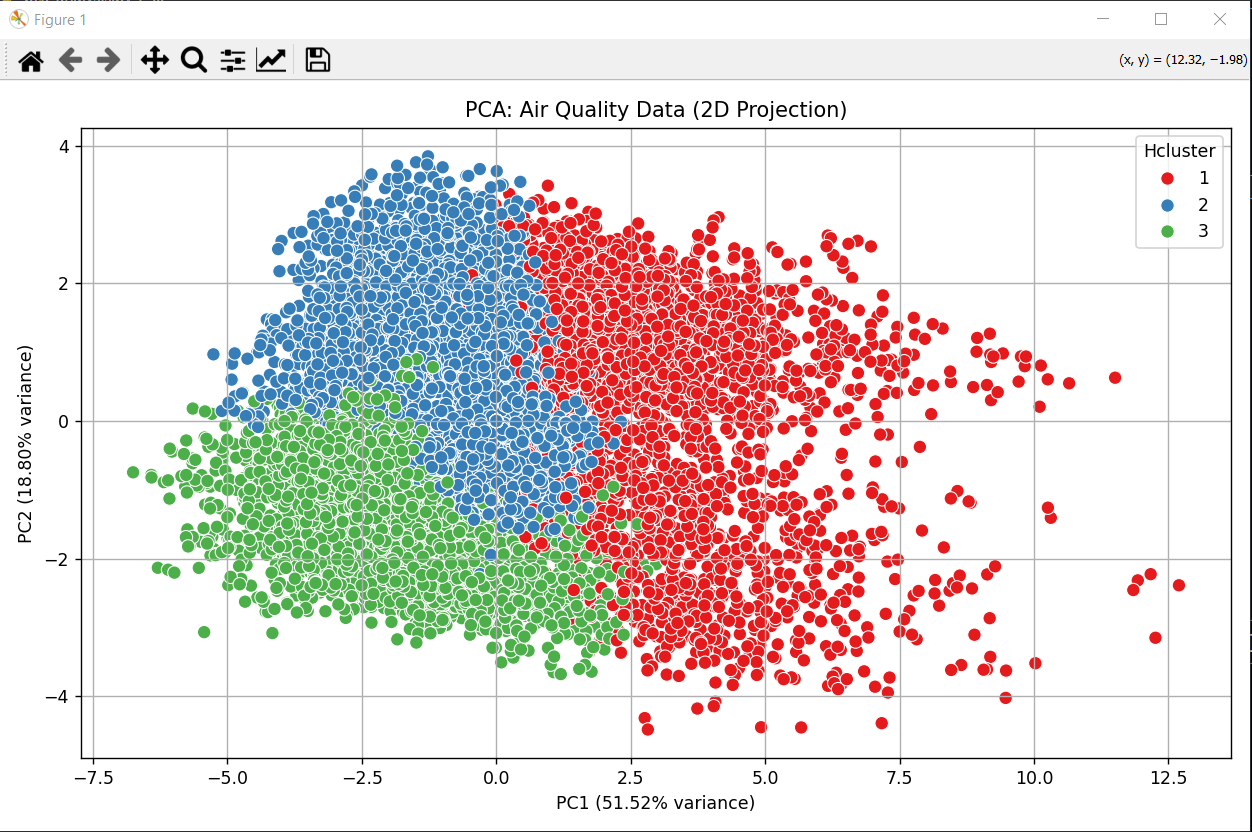
To reduce dimensionality and visualize high-dimensional data in 2D.

### Results:

- Explained Variance:  
 - PC1: ~51.5%  
 - PC2: ~18.8%  
 - Total (PC1 + PC2): ~70% of total variance

### Visualization:

- Scatter plot of PCA1 vs PCA2  
- Color-coded by cluster



### Tools:

- sklearn.decomposition.PCA  
- matplotlib, seaborn

## 5. Anomaly Detection Report:

### Objective:

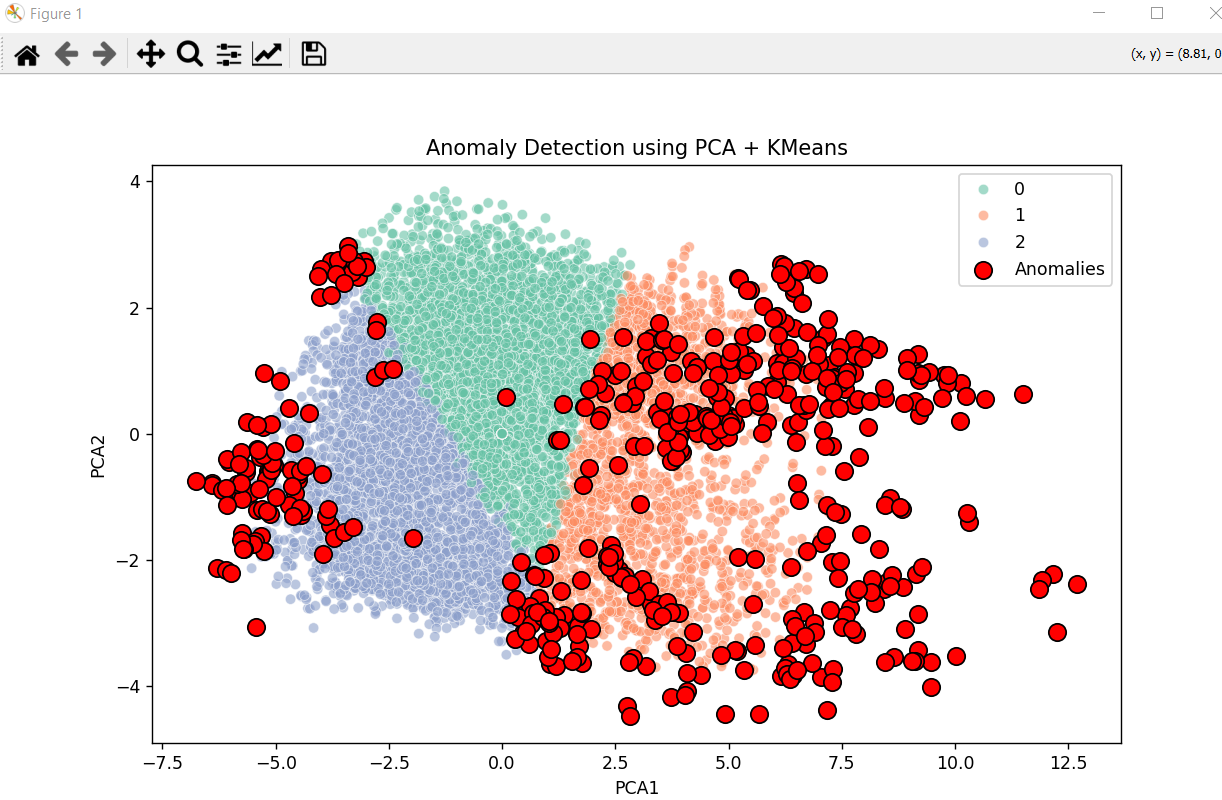
Identify data points that are far from their cluster centers and analyze them.

### Method:

- Calculated Euclidean distance of each point from its cluster centroid (KMeans)  
- Marked top 5% as anomalies based on distance threshold  
- Total anomalies detected: 474

### Visualization:

- Anomalies highlighted in red on the PCA plot



### Analysis:

Anomalies could be caused by:  
- Extreme pollution events  
- Sensor malfunction or calibration issues  
- Abnormal weather patterns  
- Data entry or logging errors  
  
Despite being part of clusters, some anomalies overlap visually in PCA due to projection limitations. They are still valid outliers in full feature space.