# Pollution Cluster Analysis Report

**Prepared by:** Rizan Shrestha

## Introduction

This project delves into the complex world of air pollution, aiming to categorize different regions based on their unique pollution characteristics. By employing K-Means clustering, we've identified four distinct groups, providing a clearer picture of the environmental challenges we face. Our analysis hinges on a variety of factors, including:

* Temperature and humidity levels
* Concentrations of PM2.5 and PM10 particulates
* Levels of gases such as NO₂, SO₂, and CO
* Proximity to industrial zones
* Population density within these areas

Through the lens of K-Means clustering and Principal Component Analysis (PCA) for visualization, this report uncovers distinct pollution profiles. These insights are crucial for developing effective environmental planning strategies and implementing targeted interventions to combat air pollution.

## Key Questions Addressed

* What are the prevailing pollution patterns across the analyzed regions?
* Which environmental and demographic factors most significantly influence these patterns?
* Which areas are most in need of immediate pollution control measures?

## Cluster Insights

### Cluster 0: Hot & Moderately Polluted

* **Avg Temp:** 29.9°C
* **PM2.5:** 35.8, **PM10:** 46.1 (high)
* **Humidity, NO₂, SO₂, CO:** Moderate levels
* **Population Density:** 502 (medium-high)

**What this means:** These are warm areas with moderate humidity, characterized by significant particulate pollution. They likely represent city outskirts or industrial belts that also have a moderate residential population, suggesting a mix of industrial and residential impact on air quality.

### Cluster 1: Hot, Dense, High CO Pollution

* **Avg Temp:** 36.2°C (highest)
* **Humidity:** 83.3% (high)
* **PM2.5/PM10:** Moderate levels
* **NO₂:** 34.5, **SO₂:** 15.3 (highest)
* **CO:** 2.02 (highest)
* **Population Density:** 619 (high)

**What this means:** This cluster typically represents densely populated urban centers. The high levels of CO, NO₂, and SO₂ strongly indicate that vehicle emissions are the primary source of pollution in these areas. The high temperature and humidity further exacerbate the pollution impact.

### Cluster 2: Mild Pollution, Warm, Moderate Density

* **Avg Temp:** 30.5°C
* **PM2.5:** 8.9, **PM10:** 19.3 (low)
* **CO:** 1.55 (moderate)
* **Population Density:** 504 (moderate)

**What this means:** These areas are characterized by mild pollution, warm temperatures, and moderate population density. They could be small towns or peri-urban areas that offer a good environment with relatively clear skies, making them suitable for balanced urban expansion.

### Cluster 3: Cool, Clean, Near Industries

* **Avg Temp:** 25°C (lowest)
* **Humidity:** 59.9% (lowest)
* **PM2.5:** 8.8, **PM10:** 14.0 (lowest)
* **CO:** 1.00 (lowest)
* **Industrial Proximity:** 11.9 (highest)
* **Population Density:** 397 (low)

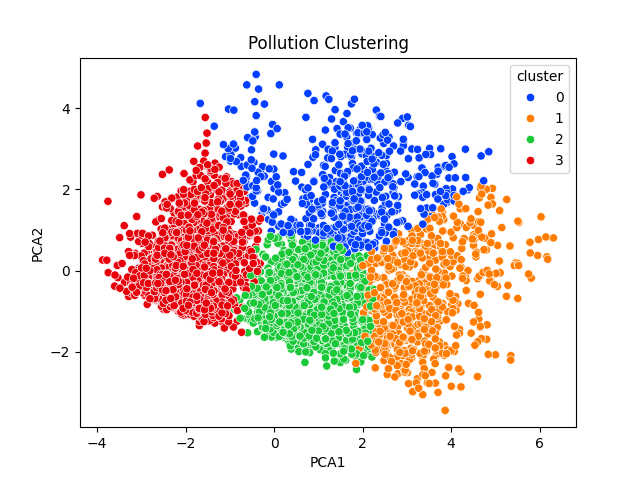
**What this means:** These are cooler regions with the lowest pollution levels, despite their proximity to industries and a higher industrial proximity score. This suggests that effective environmental regulations or industrial isolation might be in place. However, continuous monitoring is recommended due to their industrial proximity.

Image representing Cluster 3: An industrial site with clear surroundings.

## Visual Highlights

Our analysis is supported by several visualizations:

* **PCA Scatterplot:** Clearly separates the identified clusters, aiding in strategic planning.



* **Heatmaps Key Observations:**

1. **Strong Correlation between PM2.5 and PM10 (0.92):**
   * These two pollutants are highly linked, meaning where PM2.5 is high, PM10 usually is too. This suggests they often come from similar sources like traffic or industry.
2. **CO (Carbon Monoxide) has high correlation with several pollutants**
   * CO correlates strongly with **NO2 (0.65)** and **SO2 (0.64)**, indicating shared emission sources such as vehicles and industrial combustion.
3. **Population Density moderately correlates with air pollution**
   * Shows moderate positive correlation with **NO2 (0.51)** and **CO (0.51)**, meaning denser areas tend to have more pollution due to traffic and activity.
4. **Industrial Areas negatively correlate with pollutants**
   * Interestingly, **Industrial\_Areas** shows a strong negative correlation with most pollutants (e.g., **CO: -0.73**)—this might imply industrial zones are located away from monitoring stations or residential/populated zones.
5. **Temperature positively correlates with most pollutants**
   * Especially with **CO (0.63)** and **NO2 (0.51)**. Warmer temperatures might increase pollutant reactions or emissions.

A screenshot of a graph

AI-generated content may be incorrect.

* **CO KDE Plot:** Illustrates that Cluster 1 exhibits the highest distribution of Carbon Monoxide (CO) pollution.

A graph of a number of colored lines

AI-generated content may be incorrect.

* **Population Density Barplot:** Shows that Cluster 3 has low population density but covers extensive areas.

A graph of a number of people

AI-generated content may be incorrect.

* **Boxplots:** Indicate higher variability in population density for Clusters 1 and 0, while Cluster 3 consistently shows low density.

A diagram of a cluster

AI-generated content may be incorrect.

## Conclusion

This clustering analysis has successfully delineated four distinct pollution patterns, offering valuable insights for environmental management.

* **Clusters 1 and 0** require immediate and focused pollution control interventions due to their high pollution levels.
* **Cluster 2** presents an opportunity for planned urban development with minimal pollution impact.
* **Cluster 3**, while currently clean, necessitates ongoing compliance monitoring due to its industrial proximity.

Future research could expand on this by incorporating seasonal variations and real-time monitoring data. This will allow for more refined policy interventions and a better tracking of pollution mitigation progress.