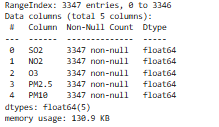
**Analysis of Air Pollution Data**

The dataset used for this project contains five key air pollutants that play critical roles in air quality assessment: SO₂ (Sulphur Dioxide), NO₂ (Nitrogen Dioxide), O₃ (Ozone), PM2.5 (Particulate Matter less than 2.5 microns), and PM10 (Particulate Matter less than 10 microns). Each pollutant is known for its specific sources and impacts on human health and the environment. For example, PM2.5 and PM10 are closely linked to respiratory and cardiovascular issues, while NO₂ and SO₂ primarily originate from industrial and vehicular emissions. Ozone, as a secondary pollutant, forms through complex atmospheric chemical reactions. This dataset serves as an excellent foundation for understanding pollution patterns and grouping regions with similar pollution characteristics through clustering.

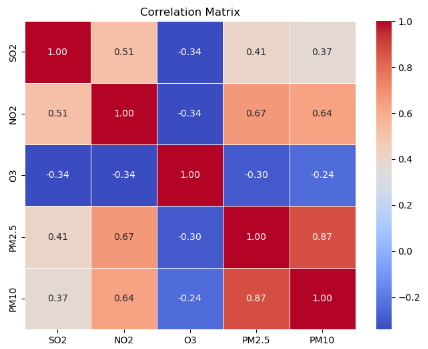


Before proceeding with clustering, the data was pre-processed to ensure it was clean, standardized, and compatible with machine learning algorithms. First, all variables were normalized using the StandardScaler to adjust for differences in scale and magnitude, as k-means clustering is sensitive to variable ranges. Outlier detection was performed using Z-scores and boxplots to identify extreme values that could distort clustering results. These outliers were removed, preserving the overall data integrity while minimizing noise.

A table with numbers and numbers

Description automatically generated

Exploratory Data Analysis (EDA) was conducted to understand the dataset, including the computation of descriptive statistics to capture the range, mean, and distribution of each pollutant. This robust preprocessing ensured that the dataset was ready for further analysis and helped improve the interpretability of the results.

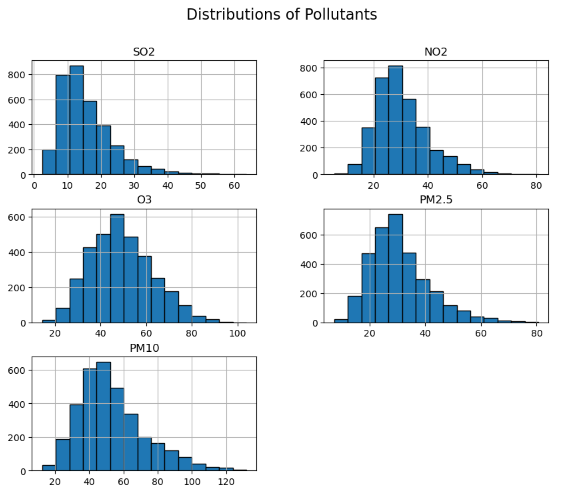


To explore relationships among the pollutants, a correlation analysis was conducted. The correlation matrix revealed a strong positive relationship between PM2.5 and PM10, with a correlation coefficient of

𝑟 = 0.87

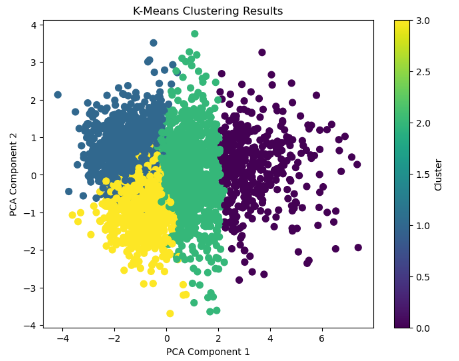
This finding suggests that these two particulate pollutants share common sources or environmental behaviour, such as emissions from combustion processes or atmospheric interactions. Moderate correlations were also observed between NO₂ and PM2.5 (𝑟 = 0.67) and between NO₂ and PM10 (𝑟 = 0.64), likely stemming from overlapping emission sources such as vehicle exhaust and industrial activities. SO₂ exhibited weaker correlations with other pollutants, reflecting its distinct sources, including coal combustion. These insights were visualized using a heatmap, providing a clear depiction of pollutant interdependencies and emphasizing the importance of targeting correlated pollutants for effective mitigation strategies.

The distributions of each pollutant were analysed using histograms.



SO₂ and NO₂ showed right-skewed distributions, indicating that their values were concentrated at the lower end with occasional spikes. Ozone followed a more normal distribution, with values clustering around the mean of 45–50, consistent with its formation as a secondary pollutant influenced by sunlight and atmospheric conditions. PM2.5 and PM10 exhibited heavy right skewness, highlighting their episodic nature, with certain regions or periods experiencing elevated concentrations due to localized pollution events. Understanding these distributions provides context for interpreting the clustering results and the behaviours of specific pollutants.

The optimal number of clusters for k-means was determined using the Elbow Method. A plot of inertia (the sum of squared distances of samples to their nearest cluster centre) versus the number of clusters revealed a noticeable "elbow" at 𝑘 = 4. This indicated that four clusters provided the best balance between simplicity and the ability to capture meaningful groupings within the data. Consequently, the k-means clustering algorithm was applied with 𝑘 = 4, producing four distinct clusters based on pollution profiles.



Some overlap between clusters was observed, suggesting that certain data points exhibited transitional features or shared characteristics, indicating a gradient rather than sharply defined boundaries.

In conclusion, this project successfully utilized k-means clustering to analyse air pollution data, revealing significant patterns and actionable insights. The combination of rigorous preprocessing, detailed correlation and distribution analysis, and robust clustering methodology ensured accurate and interpretable results. Future work could incorporate additional variables, such as meteorological data, or explore advanced clustering techniques, such as hierarchical clustering, to refine the analysis further. By identifying and understanding pollution patterns, this study highlights the potential of machine learning to support informed decision-making and promote cleaner, healthier environments.