COMS5027A – HEALTH ANALYTICS

K-Means Clustering on Air Pollution Data

1156794 – Mothapo Regina

Introduction

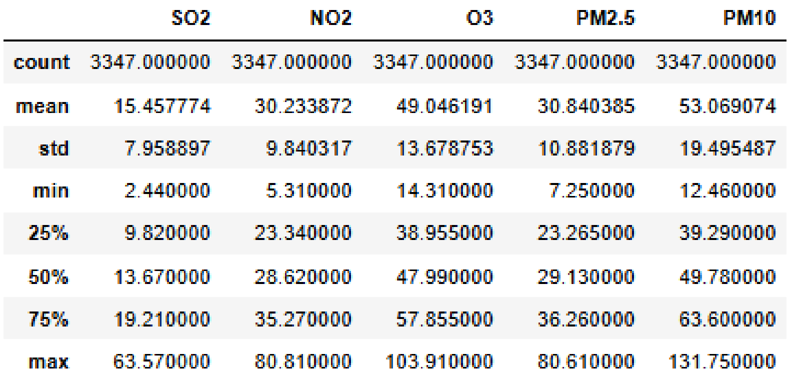
This report is based on an air pollution dataset with the aim to create a k-means clustering model to find the ideal number of distinct clusters for the pollutants, based on the similarities they share. The pollutants explored for the report are Sulphur Dioxide (SO2), Nitrogen Dioxide (NO2), Ozone (O3), Particulate matter with a diameter < 2.5 µm (PM2.5) and Particulate matter with a diameter < 10 µm (PM10). Air pollution affects humans and the environment negatively and may result in serious conditions like heart disease, lung disease and many other health conditions. It would therefore be important to monitor the air and identify areas of concern to mitigate the risk posed by air pollution (Kingsy, et al., 2016).

Principal Component Analysis (PCA) is a dimensionality reduction method explored in this report. PCA determines the importance of the variables within the data and assists to visually verify adequate separation within the clusters (Riches, et al., 2022). The PCA for this report is explored to capture the essence of the whole dataset and plot the clusters in 2-dimesion (2D) from the five pollutants.

Descriptive Statistics

Table 1 below shows the descriptive statistics for the air pollution dataset. As can be seen from the table, there are no missing records for all the pollutants. The dataset has 3347 records and 5 columns. The average concentration level for SO2 is 15.5 µg/m^{3}, NO2 is 30.2 µg/m^{3} and O3 is 49.0 µg/m^{3}. The average level for particulate matter is 30.8 µg/m^{3} for PM2.5 and 53.1 µg/m^{3} for PM10. Standard deviation shows the variability, with PM10 showing the highest variability at 19.5 µg/m^{3}, followed by O3 with 13.7 µg/m^{3}. The minimum recorded concentration of all the pollutants is 2.4 µg/m^{3} for SO2. The highest concentration is 131.8 µg/m^{3} for PM10 suggesting severe particulate pollution. For all the pollutants, the median is close to the mean, suggesting an almost symmetrical distribution of the data.

Table 1: Descriptive Statistics



	SO2	NO2	O3	PM2.5	PM10
count	3347.000000	3347.000000	3347.000000	3347.000000	3347.000000
mean	15.457774	30.233872	49.046191	30.840385	53.069074
std	7.958897	9.840317	13.678753	10.881879	19.495487
min	2.440000	5.310000	14.310000	7.250000	12.460000
25%	9.820000	23.340000	38.955000	23.265000	39.290000
50%	13.670000	28.620000	47.990000	29.130000	49.780000
75%	19.210000	35.270000	57.855000	36.260000	63.600000
max	63.570000	80.810000	103.910000	80.610000	131.750000

Data Visualizations

Figure 1 shows the correlation matrix heatmap for the pollutants SO2, NO2, O3, PM2.5 and PM10. As can be seen from the figure, the strongest positive correlation is between PM2.5 and PM10, with a correlation coefficient of 0.87, suggesting they increase or decrease together. There is a moderate positive correlation between NO2 and PM2.5 with a correlation coefficient of 0.67, suggesting high levels of PM2.5 particles increase with high levels of NO2. The other positive correlation relationship is between NO2 and SO2, which may indicate common sources. NO2 sources include the industrial burning of fossil fuels, coal, vehicle exhaust and gas combustion. SO2 sources on the hand include burning of fossil fuels and domestic heating (Jion, et al., 2023). There is a negative correlation relationship between O3 and both NO2 and SO2 with a correlation coefficient of -0.34.

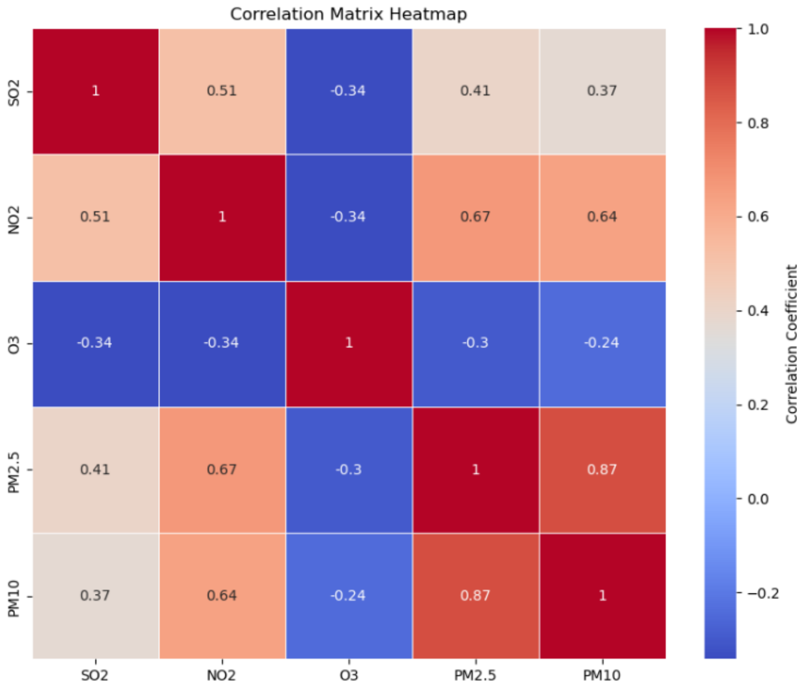


Figure 1: Correlation matrix heatmap

Figure 2 on the next page is a box plot illustrating the distribution of the 5 pollutants. As can be seen from figure 2, the spread of SO2 is moderate but there are high outliers over 30 μg/m^3. NO2 shows a significant number of outliers over 50 μg/m^3. O3 shows fewer outliers compared to other pollutants, suggesting stability in the O3 concentration levels. PM2.5 shows numerous outliers over 50 μg/m^3, suggesting frequent spikes in the concentration levels. PM10 shows frequent high levels of outliers from 100 μg/m^3 to over 120 μg/m^3. PM10 and PM2.5 show extreme and frequent outliers than all the other pollutants, raising concerns for particulate matter pollution.

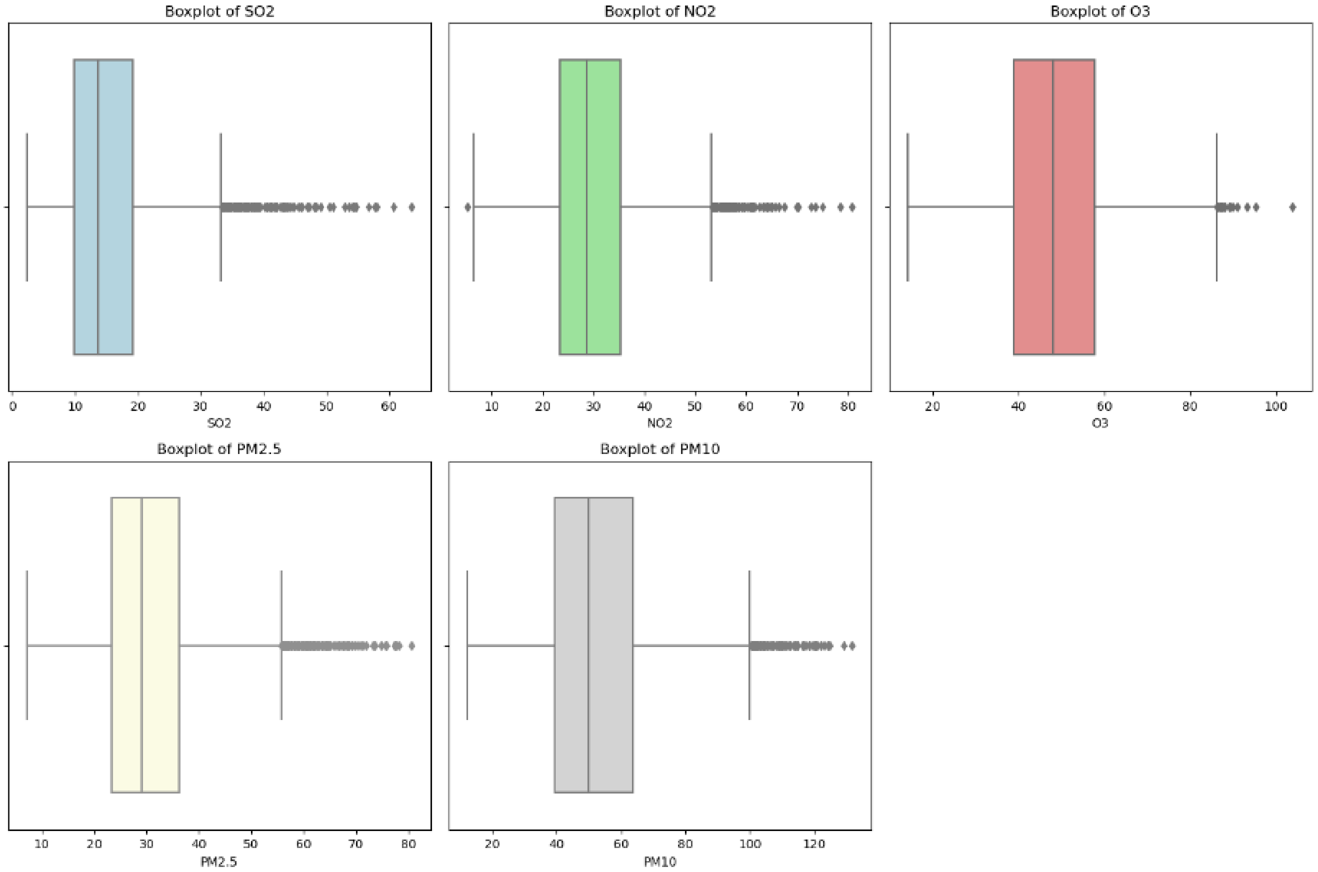


Figure 2: Distribution of the pollutants

K-Means Clustering

K-means clustering groups data into distinct and useful clusters based on the similarities in the data, for this report, the Scikit learn cluster library was used. The air pollution data was preprocessed with StandardScaler to scale the data to ensure that all features have equal weight, rather than high values dominating in the clusters. The elbow method was used, together with kneelocator library from kneed to automatically detect the elbow point that indicates the optimal number of clusters for the dataset. The elbow method uses the Sum of Squared Errors (SSE) which sums the squared distances between each point and its centroid. Lower SSE suggests the points are closer to the centroid. The optimal number of clusters was found to be 3.

Dimensionality Reduction

To plot and visualize the pollutants in the form of clusters, PCA was used as a dimensionality reduction method to capture the essence of the five pollutants, displayed in 2D. Figure 3 shows the k-means clustering results, reduced to two principal components, PCA1 and PCA2. As can be seen from the figure there a three-color coded clusters with their respective centroids marked with an x. The three clusters can serve as guide for further analysis and research for varying exposure levels and regions.

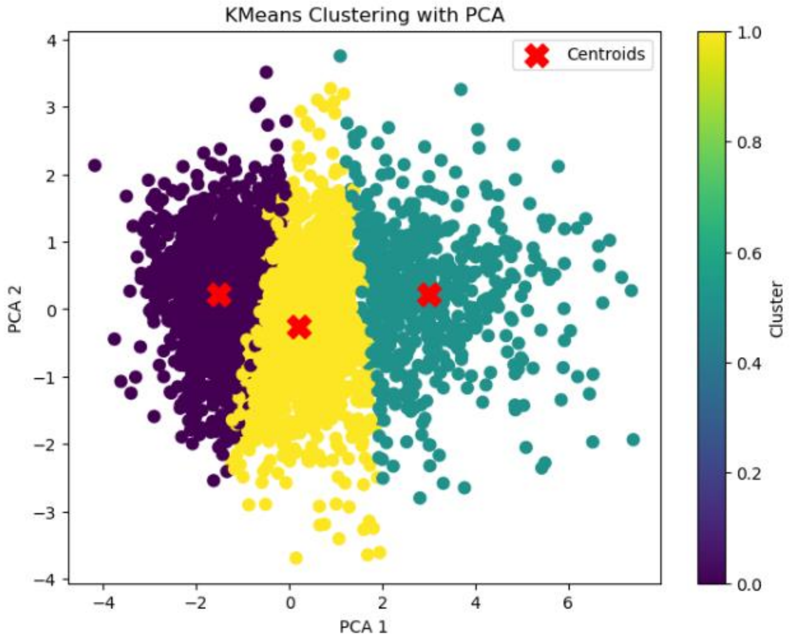


Figure 3: K-means clusters after PCA

Conclusions

The main objective of this report was to perform k-means clustering on air pollution data to segment the data into meaningful and optimal number of clusters. The summary of the descriptive statistics, correlation heatmap matrix, distribution boxplots, and k-means clustering with PCA offered a comprehensive overview of the air pollution data. The statistics results revealed that pollutants such as PM10 and O3 display high variability indicating occasional spikes in the concentration levels. The heatmap displayed a positive strong correlation between PM10 and PM2.5. The boxplots revealed that presence of outliers in all pollutants and PM10 outliers concentrated above 100 μg/m^3. Three clusters were obtained as the ideal number and displayed with the help of PCA and dimensionality reduction. The insights gained may be utilized to manage air pollution in the areas of high concentration to mitigate health risks. PM10, PM2.5 and O3 are the three main pollutants presenting more issues in concentration levels, requiring further research for health purposes.

References

1. Jion, M. M. F. et al., 2023. A critical review and prospect of NO2 and SO2 pollution over *Asia: Hotspots, trends, and sources. Science of The Total Environment, Volume Volume 876*.

2. *Kingsy, G. et al., 2016. Air pollution analysis using enhanced K-Means clustering algorithm for real time sensor data. s.l., s.n.*

3. *Riches, N. O., Gouripeddi, R., Payan-Medina , A. & Facelli , J. C., 2022. K-means cluster analysis of cooperative effects of CO, NO2, O3, PM2.5, PM10, and SO2 on incidence of type 2 diabetes mellitus in the US. Environmental Research, Volume Volume 212, Part B.*