Report title: Lab Assessment 5: K-Means Clustering Analysis on Air Pollution Data

Palesa Ka-Mbonane

Course Name: Health Analytics (FAMH4004A/COMS5027A)

Submission Date: 17 November 2024

1. Introduction

Air pollution is one of the major environmental concerns worldwide. Air pollution has been linked to multiple serious health conditions such as eye infections, irritation of the nose, throat, and eyes. It also results in serious disease such as heart disease, pneumonia, bronchitis, lung cancer and severe coughing caused by asthma.

Data mining refers to the mining or discovery of latest information based on patterns and rules from vast amounts of data. Unsupervised machine learning uses machine learning algorithms to analyze and cluster unlabeled data sets. These algorithms discover hidden patterns or data groupings without the need for human intervention. Unsupervised learning models are utilized for three main tasks—clustering, association, and dimensionality reduction. In this report we focus on clustering of k-means. The practice of grouping data objects into a collection of distinct classes known as clusters is referred to as clustering. Algorithms for clustering are used to sort unclassified, raw data objects into groups based on information patterns or structure. The K-means clustering method is a clustering that separates data into K groups. It is more prominent in classifying massive data rapidly and efficiently. The data points closest to a given centroid will be clustered under the same category. A lesser K number will indicate broader groupings and less granularity, a larger K value will indicate smaller groupings with more precision. Market segmentation, document clustering, image segmentation, and image compression are all prominent applications for K-means clustering.

The objective of this report is to analyse the air pollution dataset using k-mean clustering and identify the and group pollutants into meaningful clusters. In health analytics, the common benefits of k-means clustering are to help identify trends and diseases patterns. Data clustering is used to expedite diagnosis by grouping similar symptoms or trends, thus enhancing patient care and diagnosis.

2. Methodology and Data Description

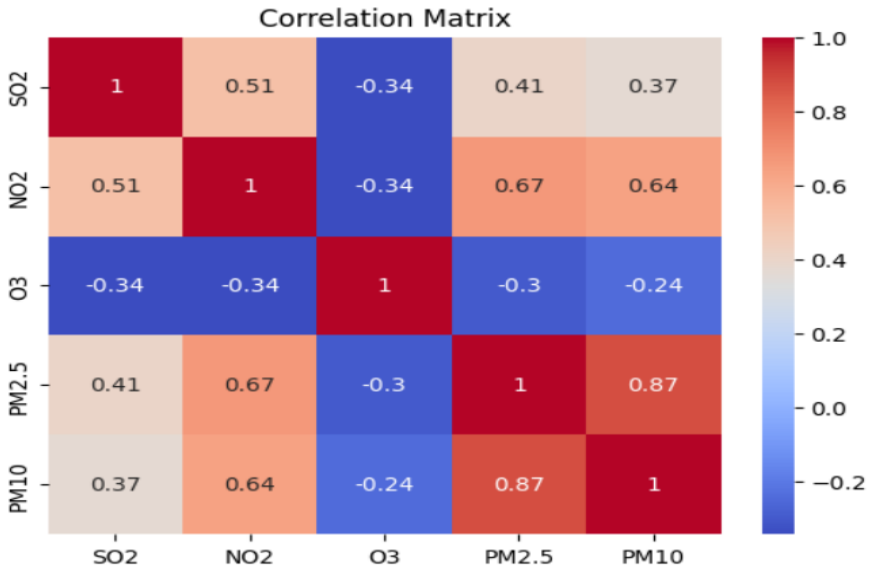
The Air pollution dataset is used in this report provided by the University of Witwatersrand. A summary of the data we used in the study is as follows, we looked at 3,347 rows and five types of information about air pollution: SO2, NO2, O3, PM2.5, and PM10. These pollutants have a significant impact on the environment and human health.

Data pre-processing includes data validation, to ensure that the dataset was in good condition. Missing values were handles by applying data imputation of the mean. In this report data processing was carried out using the Air population dataset. The features used in the dataset is data normalization to ensure that k- means clustering is carried out is carried out accurately and efficiently.

3. Exploratory Data Analysis

We consider the relationship between the variables in the dataset, a correlation matrix is computed and visualised using a heatmap as observed in **Fig.1**. Featuring values ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation), the matrix shows the degree of linear relationship between each pair of variables. Strong correlations are observed in **Fig.1**, PM2.5 and PM10 showed a correlation of 0.87 which demonstrates that they increase together. Weak correlations of -0.34 are observed between, NO2 and O3 as well as PM10 and O3 with a correlation of -0.24. The negative correlation suggests that there is no linear relationship between these variables.

Figure 1: Correlation Matrix of air pollution dataset



The **distribution** of each pollutant is important in understanding the distribution and observing any skewness, trends, and outliers in the dataset. For **Fig. 2** and **Fig. 3, Fig.5** and **Fig.6** we observe a slight skewness to the left that suggests that they are outliers among pollutants PM2.5, PM10, SO2 and NO2. The distribution observed in **Fig.4** indicates normal distribution with a slight skewness to the left which indicates that the values are concentrated around the mean, with a few higher values.

Figure 2: Distribution of Particular Matter.

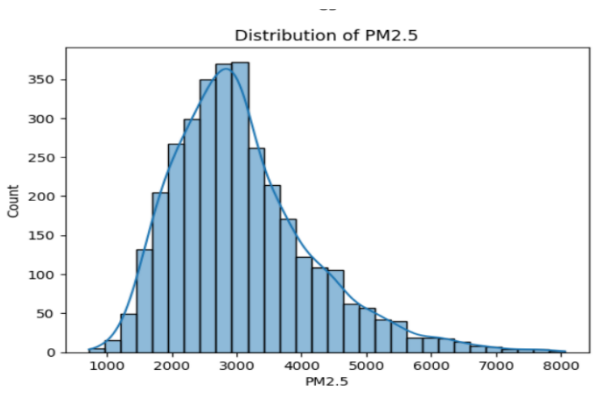
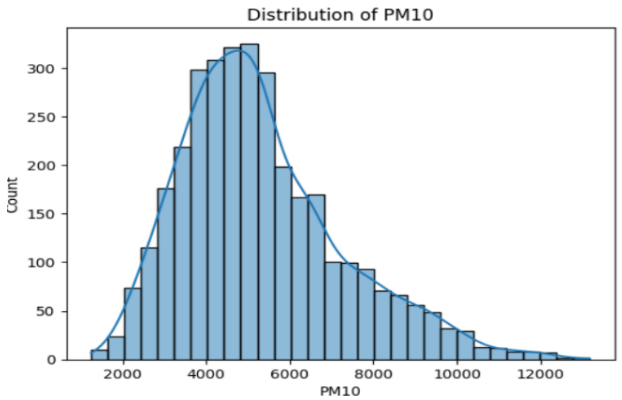


Figure 3: Distribution of Particular Matter.



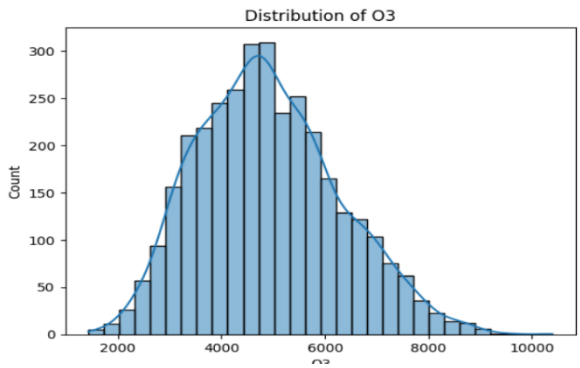


Figure 4: Distribution of Ozone.

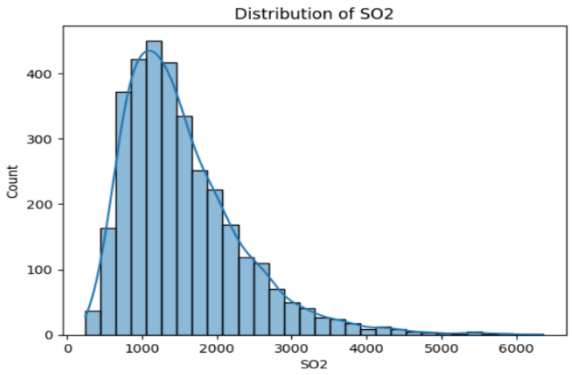


Figure 5: Distribution of Sulphur Dioxide.

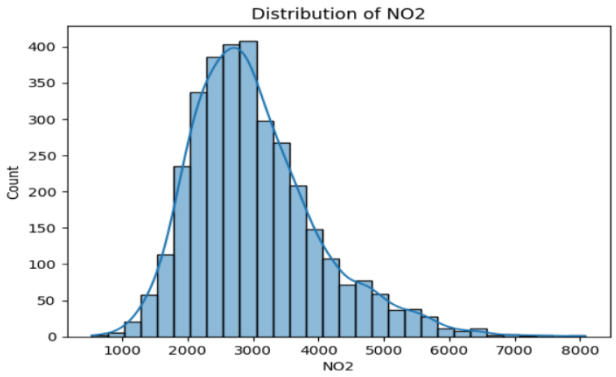


Figure 6: Distribution of Nitrogen Dioxide.

4. Optimal Number of Clusters

In order to determine the optimal number of clustered for the k-means algorithm two methods were used, which are the Elbow method and Silhouette Analysis. As observed in **Fig.7** as the k value increases the inertia decreases this is because as we add more clusters it reduces the distance between points and their respective cluster centers. The elbow point is identified as the stage where the rate of decrease in inertia significantly slowed, suggesting that adding more clusters does not meaningfully enhance the clustering results. From Fig.7, the elbow point is observed at k=2, which suggests that two clusters balance the model's complexity.

The silhouette score evaluates how well each data point fits within its assigned cluster compared to other clusters. We observe a line plot in **Fig.8** that shows the relationship between k and the silhouette score, calculated with k values from 1 to 10. In Fig.8 the highest silhouette score is observed at K = 2, which is also observed in the elbow method. This suggests that k=2 is the optimal number of clusters.

Figure 7: Inertia vs Number of Clusters.

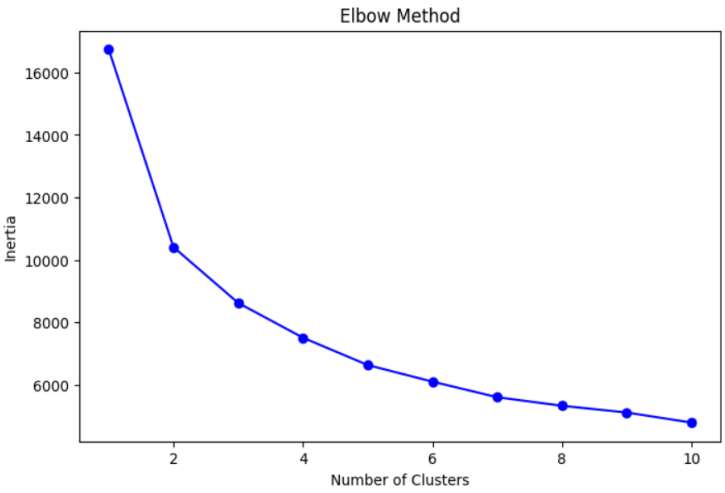
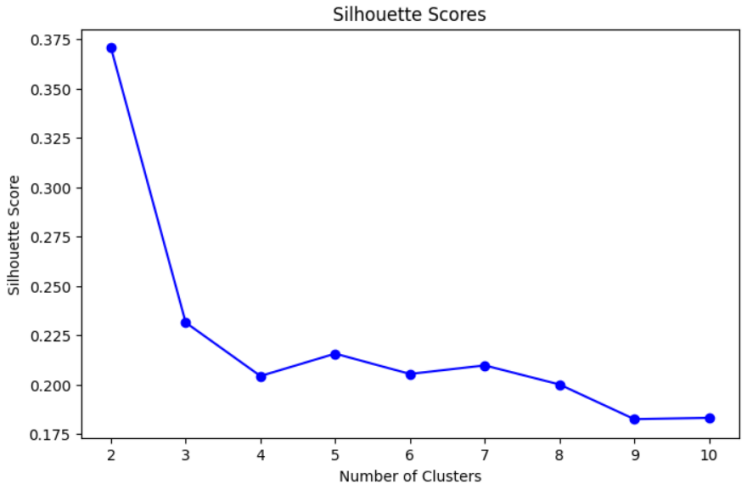


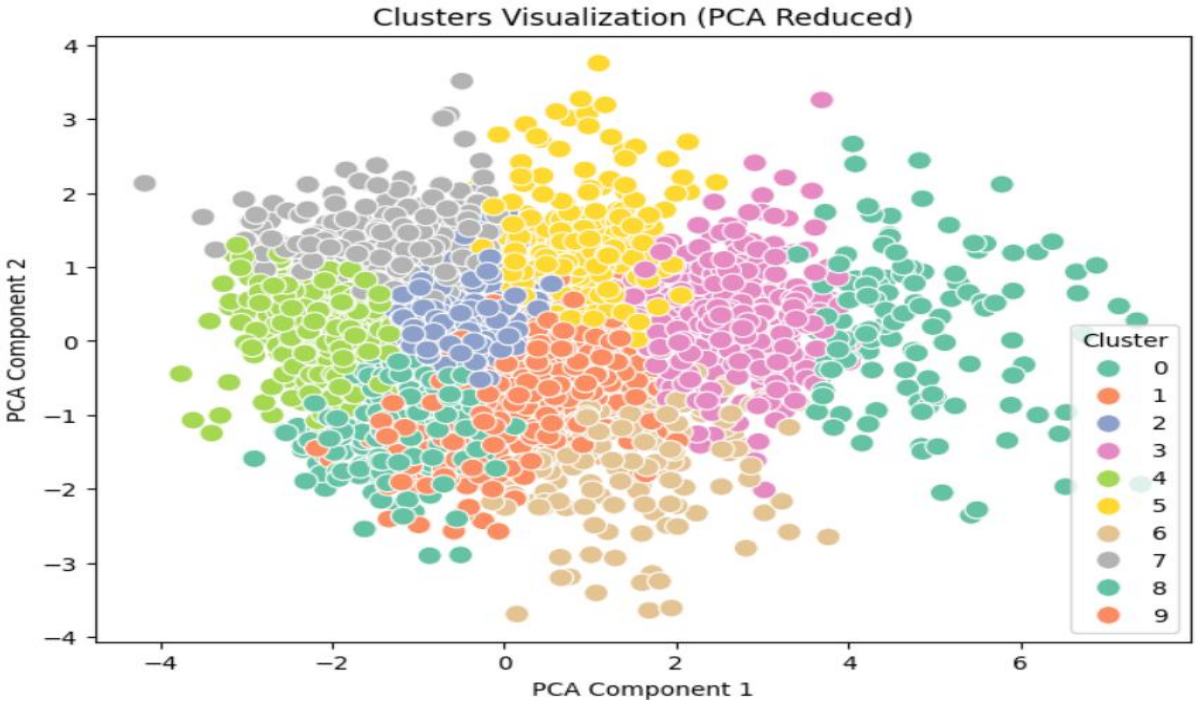
Figure 8: Silhouette score v Number of Clusters.



5. Implementation

The air pollution dataset is normalized using Standard Scaler t=this ensures that all variables were in the same scale. The optimal number of clusters (K=2) was determined using the Elbow method and silhouette analysis. The data points were assigned to one of the two clusters and labels were appended as a new column in the dataset. The clustered dataset has been exported to an Excel file, which is attached to this report for further analysis.

Figure 9: CA-Based Clusters Visualization



Principal component analysis (PCA) is used to visualize the normalized clusters as observed **in Fig.9.** Distinct groupings highlighted in the plot confirm the effectiveness of the clustering.

6. Results and Discussion

The k-means clustering analysis divided the dataset into two distinct clusters, which represent the pollutant levels. In cluster one we find higher average concentration of pollutants such as PM10, PM1.5 and NO2. These clusters are in places that experience high levels of pollution. In cluster two all variables have moderate pollution levels; there are no extreme outliers. likely representative of mixed rural-urban areas or urban areas with pollution control systems in place.

7. Conclusion

The K-means algorithm can be used to cluster and classify harmful pollutants. The results of clustering help to determine which regions are highly affected by air pollution. The analysis successfully divided the data into two distinct concentrations. Cluster one represents areas with severe air pollution levels, that are demonstrated by higher concentrations of pollutants. While cluster two defines cleaner regions with low pollutant levels. Decision-makers can better control air quality and reduce related health risks by adjusting environmental and public health policies to these unique groups. K-means is an efficient clustering tool as it effectively grouped regions with similar air pollutants, allowing a better understanding of environmental trends.

The limitations of k-mean clustering are such that data standardization was necessary due to sensitivity to scale. The Elbow Method and Silhouette Analysis demonstrate the need for a thorough assessment of the predetermined number of clusters. Work in the future to improve the algorithms include exploring more advanced clustering techniques. The report highlights the value of actionable insights for sustainable development and better public health outcomes through using clustering algorithms to support data-driven approaches in environmental policy and health analytics.

References

1. Hu, J.; Huang, L.; Chen, M.; Liao, H.; Zhang, H.; Wang, S.; Zhang, Q.; Ying, Q. Premature Mortality Attributable to Particulate Matter in China: Source Contributions and Responses to Reductions. Environ. Sci. Technol. 2017, 51, 9950–9959.
2. Lv, B.; Liu, Y.; Yu, P.; Zhang, B.; Bai, Y. Characterizations of PM2.5 Pollution Pathways and Sources Analysis in Four Large Cities in China. Aerosol Air Qual. Res. 2015, 15, 1836–1843.
3. A. V. Vidhyapeetham, “Crime Analysis and Prediction using Optimized K-Means Algorithm 1 1,” no. Iccmc, pp. 915–918, 2020.
4. Kim, K.-H.; Jahan, S.A.; Kabir, E. A review on human health perspective of air pollution with respect to allergies and asthma. Environ. Int. 2013, 59, 41–52.
5. Margaret H. Dunham, Data Mining- Introductory and Advanced Concepts, Pearson Education,2006.