

Uncovering Carbon Dioxide Emission Patterns: Data-Driven Insights with K-Means Clustering Across 20 Countries

Ariana Yunita^{1, 2} and Angelia Regina Dwi Kartika¹

Department of Computer Science, Universitas Pertamina¹

Center of Data Science and Automation, Universitas Pertamina²

ariana.yunita@universitaspertamina.ac.id¹, 105220039@student.universitaspertamina.ac.id²

Article Info

Article history:

Received Nov 24, 2025

Revised Dec 11, 2025

Accepted Aug 31, 2026

Keyword:

Clustering Analysis

Dimensionality Reduction

CO₂ emissions

Greenhouse gas

K-Means

ABSTRACT

Greenhouse gas (GHG) emissions constitute a global challenge, with carbon dioxide (CO) identified as a main contributor to climate change and its cascading effects on ecosystems and human societies. Despite growing recognition of the multisectoral nature of CO₂ emissions, comparative analyses across diverse national contexts remain limited, particularly those employing unsupervised machine learning approaches. This study aims to gain insights into the sectors contributing to CO₂ emissions across 20 countries, which is crucial for decision-making in emissions mitigation. This study utilized a dataset from World Development Indicators (WDI) comprising twenty countries using over a 43-year period (1974-2016). The seven sectors investigated include electricity and heat generation, gas fuel consumption, liquid fuel consumption, manufacturing and construction industries, residential and commercial sectors, solid fuel consumption, and transportation. The methods employed encompass clustering with the K-Means algorithm and dimensionality reduction using Principal Component Analysis (PCA). From PCA, we reduce 301 features to 8 Principal Components. The clustering results yielded four clusters based on CO₂ emission levels across these sectors, with PCA analysis and exploratory data analysis (EDA) findings supporting this grouping. The evaluation of the model uses the Davies-Bouldin Index (DBI), the Elbow Method, and the Silhouette Score to determine the optimal parameter for the K value in K-Means. Insights from these findings are discussed.

© This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Corresponding Author:

Ariana Yunita

Department of Computer Science

Universitas Pertamina

Jl. Teuku Nyak Arief, Jakarta, Indonesia

ariana.yunita@universitaspertamina.ac.id

1. INTRODUCTION

CO₂ is the main contributor to GHG emissions, with 37.4 gigatons emitted globally in 2023. In Indonesia, CO₂ accounts for about 71.6% of total GHG emissions [1], largely from fossil fuel combustion. This combustion is not only the primary global energy source, but also a major driver of climate change, leading to widespread impacts such as heatwaves, droughts, infrastructure damage, and coastal ecosystem threats from sea-level rise [2, 3]. The detrimental effects of CO₂ emissions on human life are significant. In 2022, Indonesia ranked seventh globally accounting for 2.26% of total global emissions [4] and is projected to increase by 2050 [5]. Each country, with its unique characteristics, contributes to CO₂ emissions differently. One country can generate emissions from transportation, while another country may contribute through its industrial activities.

To address CO₂ emissions, leveraging technology and artificial intelligence is vital for innovation and effective mitigation strategies. Research from the Grantham Research Institute highlights that advancements in AI could reduce global greenhouse gas emissions by 3.2 to 5.4 billion tonnes of CO₂-equivalent annually by 2035, through improving renewable energy efficiency, optimizing transportation, and enabling smarter climate policies [6]. AI's role in technological innovation, resource efficiency, and predictive modeling creates a clear roadmap to accelerate the net-zero transition and climate adaptation efforts, making it a crucial tool for climate change mitigation.

Clustering is a machine learning technique designed for data grouping, offering the advantage of generating recommendations aligned with cluster-specific characteristics [7]. Prior research has employed clustering methods to categorize students based on shared attributes. Similarly, this approach can identify countries with comparable emission patterns, enabling targeted strategies for emission reduction.

Some previous research related to data science and energy, such as the research by Mardani et al. [8] employs Self Organizing Map (SOM) clustering, Adaptive Neuro Fuzzy Inference System (ANFIS), and Artificial Neural Network (ANN) to analyze energy use and economic development in G20 countries, identifying links to CO₂ emissions. Namazkhan [9] utilized decision trees to understand the residential energy consumption sectors and their direct contribution to CO₂ emissions, supported by building characteristics, socio-demographics, and psychological factors. While these studies provide valuable insights into specific aspects of energy consumption and CO₂ emissions, there remains a research gap in applying PCA and K-Means to simultaneously analyze longitudinal clustering patterns across multiple sectors and countries, which is essential for understanding comprehensive cross-sectoral emission trajectories.

This study aims to apply Principal Component Analysis (PCA) to reduce the dimensionality of the CO₂ emission dataset and to extract the most significant features for subsequent clustering analysis. The K-Means clustering algorithm is then employed to identify distinct groupings of countries based on their CO₂ emission profiles across seven contributing sectors. To ensure the reliability of the clustering results, the optimal number of clusters is determined through three validation metrics, namely the Elbow Method, the Silhouette Score, and the Davies-Bouldin Index (DBI). The resulting clusters are subsequently interpreted and profiled through Exploratory Data Analysis (EDA) to uncover meaningful patterns in CO₂ emission behavior across the 20 selected countries. Ultimately, this study seeks to derive data-driven insights from the clustering outcomes that can inform evidence-based policy recommendations for emissions reduction and climate change mitigation strategies. The combination of PCA and K-means adopts the method from Yunita et al. [7] that utilized Principal Component Analysis (PCA) and K-means clustering to group students based on regional and socioeconomic factors, with PCA reducing the large dataset dimensions. The aim is to

provide a deeper understanding of the energy sector’s contribution to CO_2 emissions.

2. RESEARCH METHOD

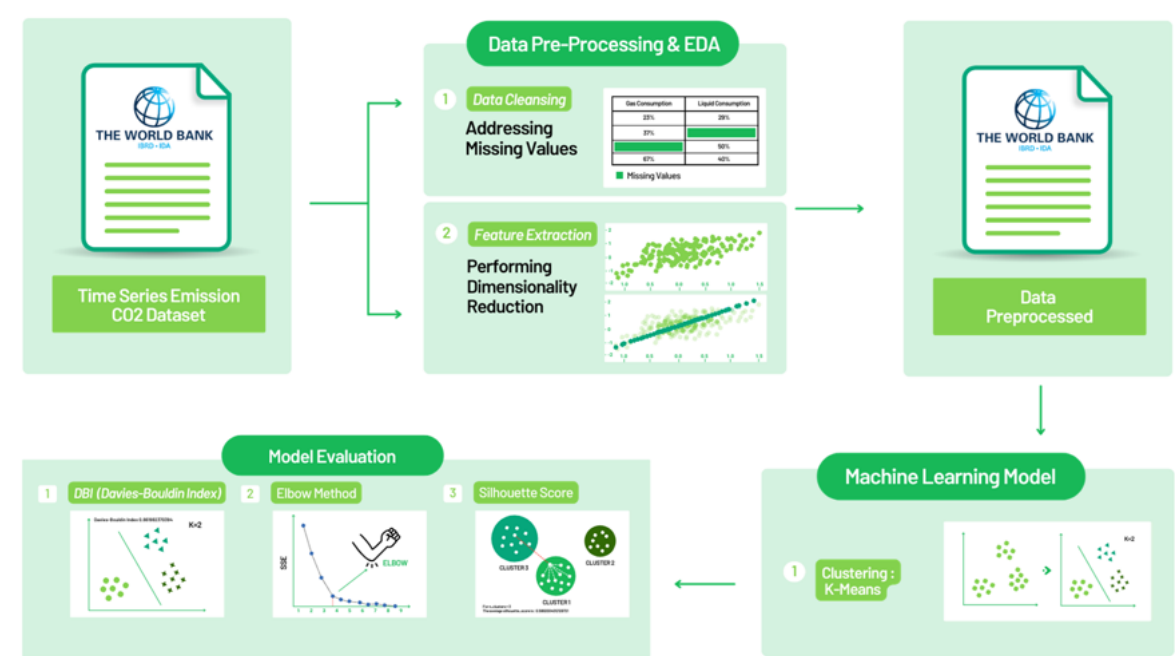


Figure 1. Research Stages

This research methodology addressed the formulated research questions by utilizing global data from the World Bank. There were four research stages as seen in Figure 1, namely data collection, data preprocessing, modeling, and evaluation. The data source was from the World Bank’s World Development Indicators (WDI) database [10], covering 20 countries over a 43-year period (1974–2016). Data collection utilized four variables from the database, namely country, series, and time. The 20 countries selected were China, the United States, India, Russia, Japan, Indonesia, Iran, Germany, South Korea, Saudi Arabia, Canada, Mexico, Turkey, Brazil, South Africa, Australia, the United Kingdom, Vietnam, Italy, and Poland. The series variable focused on the keyword CO emissions, with seven indicators related to CO emissions, including CO emissions from electricity and heat production, CO emissions from gas fuel consumption, CO emissions from liquid fuel consumption, CO emissions from the manufacturing and construction industries, CO emissions from residential and commercial buildings, CO emissions from solid fuel consumption, and CO emissions from the transportation sector. The last variable, time, referred to the research period spanning 43 years, from 1974 to 2016. Data were retrieved in March 2024.

Data preprocessing included cleaning and handling missing values through linear interpolation. Exploratory Data Analysis (EDA) was conducted using line charts and bar plots to identify data patterns. Principal Component Analysis (PCA) was applied for dimensionality reduction, retaining 95% of the data’s variance. The preprocessed data were then used to train the clustering (K-Means) models. Two experiments were conducted: one with the raw data and another with the PCA-reduced data. The K-Means algorithm clustered data based on feature similarities, and the model’s performance was assessed using the Davies-Bouldin Index (DBI), the Elbow Method,

and the Silhouette Score to determine the optimal number of clusters.

3. RESULT AND ANALYSIS

3.1. Preprocessing and EDA Results

The raw data collected from the data source consisted of 165 rows and 45 columns. These data were further processed by removing unnecessary columns, such as the country code. After the first processing step, new data comprising 140 rows (resulting from 20 countries multiplied by 7 features) and 45 columns were obtained for the study. The next step involved data reshaping by adding new features based on the Series Name feature, which contained seven emission-producing sectors of CO₂. These seven sectors were separated by the data for each year over 43 years for each country. The result of data reshaping yielded 20 rows, representing 20 countries and 301 new features. The number of new features was derived from 7 sectors multiplied by 43 years. The reshaping of the data followed the rule that each row represents an instance and each column represents a feature. Table 1 presents the details of each feature after reshaping the raw data.

Table 1. Feature Description of the Data *Reshaping* Results

No	Feature Name	Description
1	0_1974, 0_1975, ..., 0_2015, 0_2016	CO ₂ emissions data in the electricity and heat sector over a period of 42 years (1974-2016)
2	1_1974, 1_1975, ..., 1_2015, 1_2016	CO ₂ emissions data in the gaseous fuel sector over a period of 42 years (1974-2016)
3	2_1974, 2_1975, ..., 2_2015, 2_2016	CO ₂ emissions data in the liquid fuel sector over a period of 42 years (1974-2016)
4	3_1974, 3_1975, ..., 3_2015, 3_2016	CO ₂ emissions data in the manufacturing and construction sector over a period of 42 years (1974-2016)
5	4_1974, 4_1975, ..., 4_2015, 4_2016	CO ₂ emissions data in the residential and commercial sector over a period of 42 years (1974-2016)
6	5_1974, 5_1975, ..., 5_2015, 5_2016	CO ₂ emissions data in the solid fuel sector over a period of 42 years (1974-2016)
7	6_1974, 6_1975, ..., 6_2015, 6_2016	CO ₂ emissions data in the transport sector over a period of 42 years (1974-2016)

The preprocessed data were analyzed using EDA to help with clustering and labeling. This involved creating line charts to visualize data trends over 43 years for the 7 emission sectors across 20 countries, and bar plots to identify which countries had the highest and lowest CO₂ emission percentages in each sector. The visualizations in Figure 2 illustrate CO₂ emissions trends across various sectors from 1974 to 2016 for multiple countries. For the electricity and heat sector, the line chart shows annual CO₂ emissions, with South Africa, Poland, and Australia as the highest emitters, while Brazil and Vietnam have the lowest. The bar plot highlights significant variation in emissions per capita among countries. In the gaseous fuel sector, the line chart shows a sharp increase in emissions in countries like Russia and Iran. The bar plot confirms these countries as top emitters, with Brazil, Vietnam, India, China, and South Africa having the lowest emissions. The liquid fuel sector's line chart reveals fluctuating emissions trends, with the bar plot identifying the 10 highest

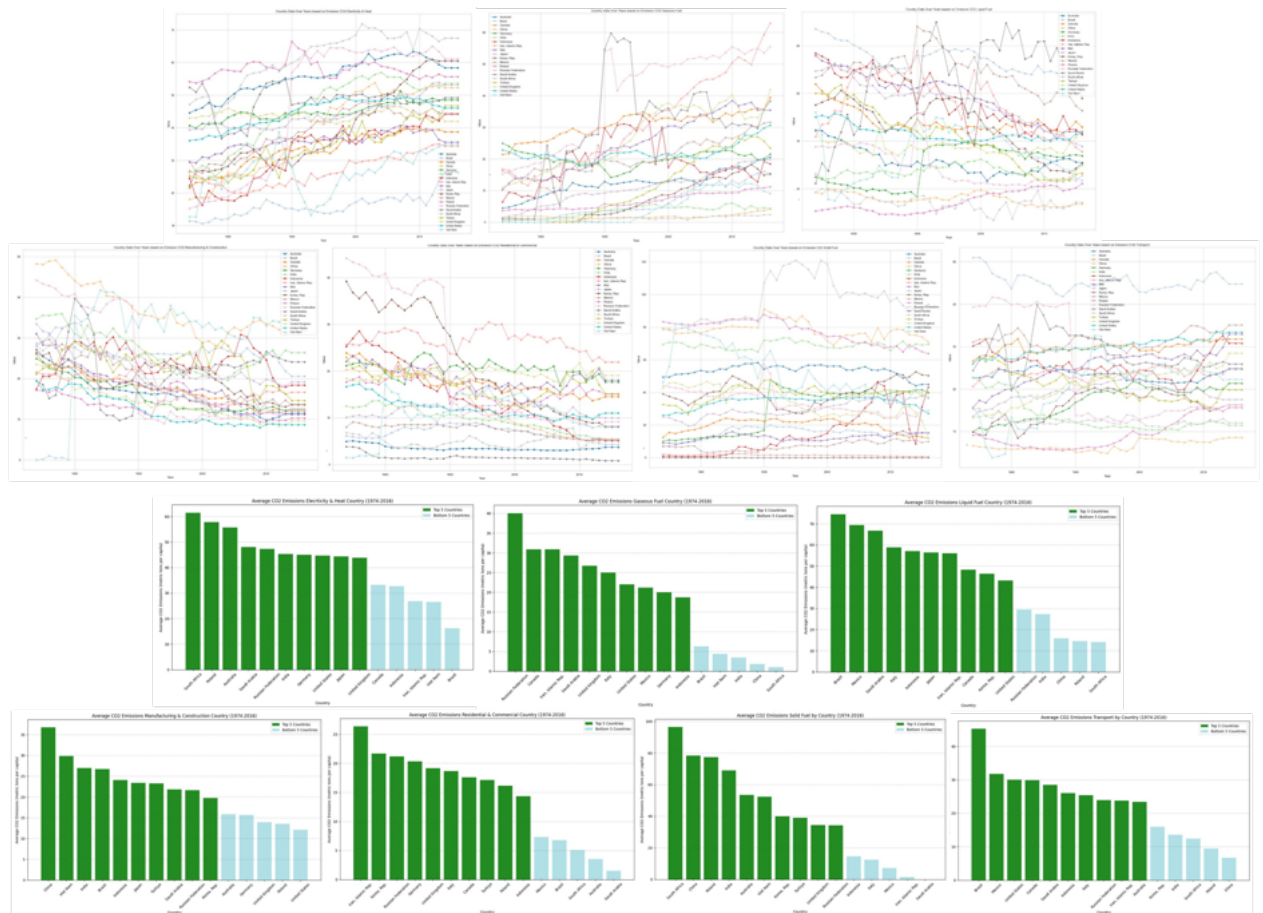


Figure 2. Results of the line chart and bar plot for 7 sector

and 5 lowest emitters.

For the manufacturing and construction sector, the line chart shows varying trends, with the bar plot highlighting the top and bottom emitters. In the residential and commercial sector, the line chart shows varied trends, with Korea showing significant decreases and Australia and Saudi Arabia exhibiting flat trends. The bar plot identifies the highest and lowest emitters. The solid fuel sector's line chart shows high emissions from South Africa and stagnant trends in Iran and Saudi Arabia, reflected in the bar plot. Finally, the transportation sector's line chart shows significant emission increases in Brazil and Mexico, steady growth in China, and fluctuating trends in the US and Canada. The bar plot indicates that China, Poland, South Africa, India, and South Korea as the lowest emitters. Overall, the visualizations highlight significant variations in CO₂ emissions across different sectors and countries over the 43-year period. The results from the EDA bar plot were used to determine the labels for the K-Means clustering results, providing a clear and useful dataset for each cluster label. Next, data analysis was performed using a correlation matrix to identify correlations among features, enabling feature selection or extraction.

Figure 3 illustrates correlations between variables related to the liquid fuel sector and other sectors. A deep red color indicates a strong positive correlation between liquid fuels (variable X) and the transportation sector (variable Y), suggesting that transportation heavily relies on liquid fuels like gasoline and diesel. Conversely, a deep blue color shows a strong negative correlation between liquid

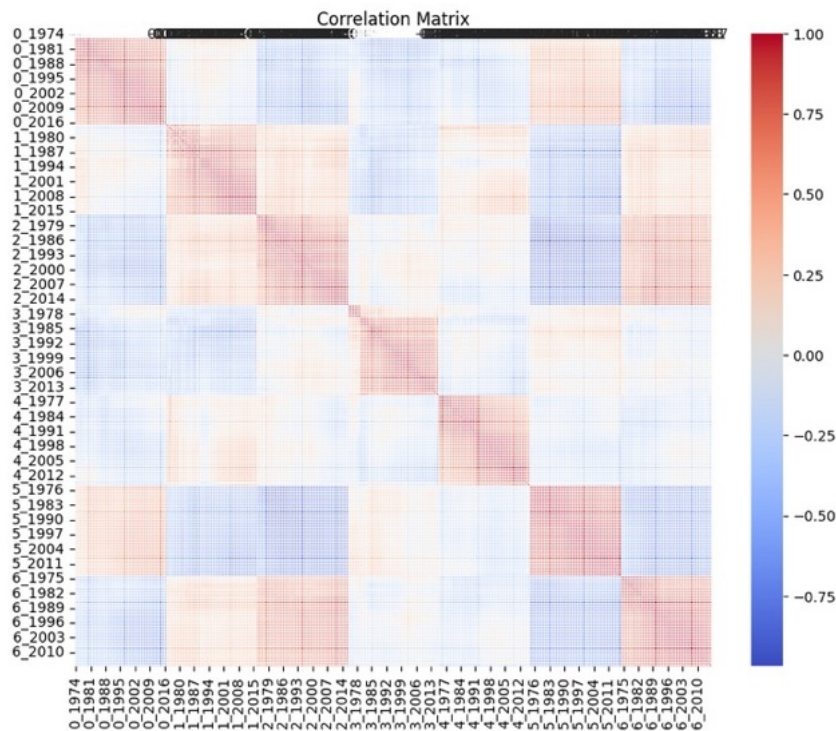


Figure 3. *Feature correlation matrix*

fuels (X) and solid fuels (Y), indicating a decrease in solid fuel consumption, possibly due to a shift in fuel preferences [11, 12]. The next step involves applying Principal Component Analysis (PCA) to reduce the dimensionality of the data from 301 features to 8 principal components, as depicted in Figure 4. PCA helps to summarize the data while retaining the most significant variance.

Based on the results obtained, the eight selected principal components have a variance ratio of approximately 0.959 or 95.1%. This variance ratio is the most suitable for representing the data well, as values above 95% indicate that the eight principal components are sufficient to represent the data effectively. This reduces the data complexity from its original dimension with minimal information loss, about 4.7% of the total variance of the original data. Each principal component (PC) is interpreted by determining the top 5 values of each PC. These values are then selected for their representation and labeled based on their significance. The interpretation of each PC is presented in Table 2. For example, PC1 represents as solid sector year in 2002, 2003, 2004, 2005, and 2013, since 5 features that have top values are 5_2013, 5_2002, 5_2003, 5_2005, and 5_2004. For PC2, this PC represents the manufacturing and construction sector of 1998, 1999, 2000, 2001, and 2005.

3.2. Clustering Results Using PCA and K-Means

After performing dimensionality reduction with PCA, the next step was to cluster using the K-Means algorithm (distance metrics = euclidean) to group features based on the reduced-dimensional data. Before running the clustering, the first step was to find the optimal value of K. This K value determines the number of clusters produced. To find the optimal K, several evaluation methods in K-Means were used, including Davies-Bouldin Index (DBI), Elbow Method, and Silhouette Score. The evaluation was conducted by systematically examining a range of K values

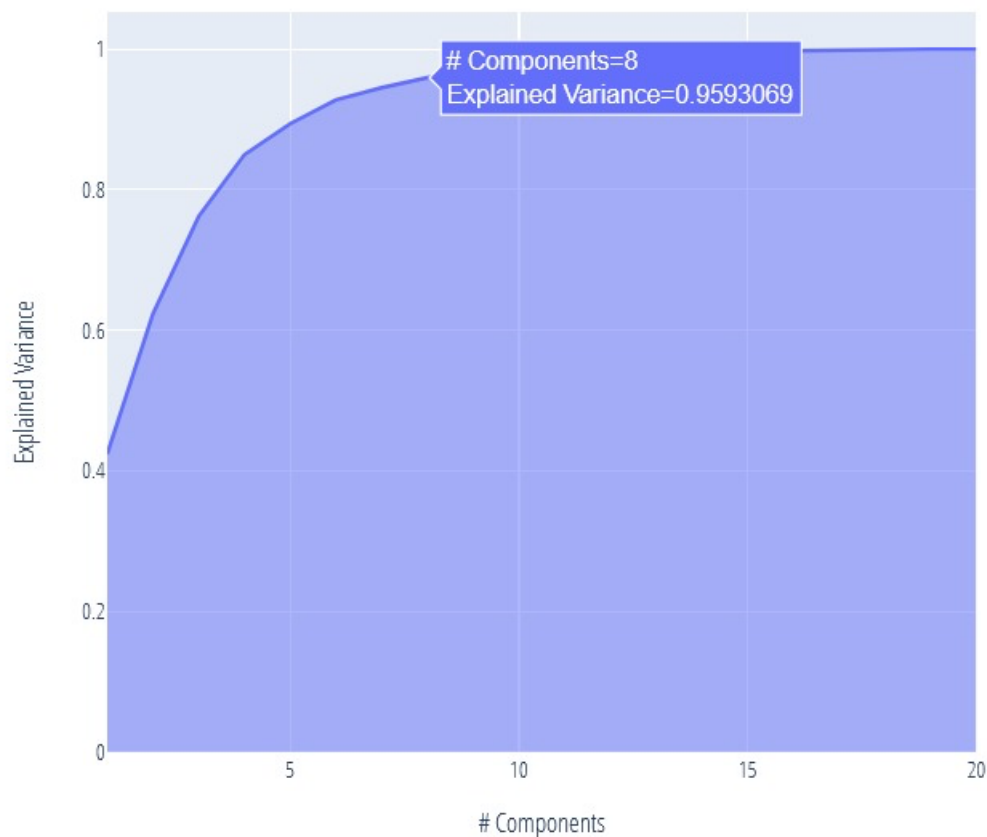


Figure 4. Variance ratio results for PCA determination

from 1 to 10 to determine the optimal number of clusters. DBI indicates that lower values are better for cluster separation. The Elbow Method examines the decrease in inertia or WCSS, and the point where this decrease starts to slow down significantly forms an "elbow" in the curve. The Silhouette Score, approaching 1, indicates better clustering quality.

The evaluation metrics provided varying suggestions for the optimal number of clusters (K). The Elbow Method indicated K=4, while the Davies-Bouldin Index (DBI) suggested K=8 (Figure 5). However, the final decision was based on the Silhouette Score, which evaluates the quality of the clustering. The Silhouette Score indicated that K=4, with a score of 0,210 (Figure 6), was the optimal choice, since it is closest to 1. In comparison, K=8 had a Silhouette Score of 0,193 (Figure 7). Therefore, K=4 was selected as the optimal number of clusters for the analysis.

The PCA-based clustering analysis demonstrates varying degrees of cluster separation. In Figure 8 reveals that the PC1 vs PC2 plot offers the best cluster separation, with Clusters 1, 2, 3, and 4 distinctly separated and showing minimal overlap. In contrast, Figure 9 highlights significant overlap among clusters in various principal component combinations. Specifically, the PC2 vs PC3 plot shows some overlap between Clusters 1 and 2, and between Clusters 2 and 3. Some indicate persistent overlapping in visualizations involving PC3 through PC8 and PC4 through PC8, making the distinction between clusters less clear. Further illustrates that overlap occurs in combinations of PC6 with PC7 through PC8 and PC7 with PC8, indicating that higher principal components also fail to provide clear separation. Overall, while PC1 vs PC2 provides the most distinct separation,

Table 2. Feature Representation for Each Principal Component (PC)

PC	Features		Label
PC1	5_2013(0.562), 5_2003(0.558), 5_2004(0.557)	5_2002(0.559), 5_2005(0.557),	Solid fuel sector year 2002, 2003, 2004, 2005, 2013
PC2	3_1999(0.790), 3_2000(0.778), 3_2005(0.774)	3_2001(0.779), 3_1998(0.775),	Manufacturing and construction sector of the year 1998, 1999, 2000, 2001, 2005
PC3	4_1991(1.037), 4_1993(0.998), 4_1994(0.978)	4_1992(1.028), 4_1990(0.995),	Residential and commercial sector of the year 1990, 1991, 1992, 1993, 1994
PC4	3_1980(0.994), 3_1982(0.986), 3_1988(0.955)	3_1981(0.988), 3_1979(0.983),	Manufacturing and construction sector of the year 1979, 1980, 1981, 1982, 1988
PC5	6_1980(0.895), 6_1981(0.841), 6_1975(0.800)	6_1974(0.848), 6_1982(0.827),	Transportation sector of the year 1974, 1975, 1980, 1981, 1982
PC6	2_1978(1.011), 2_1979(0.959), 2_1980(0.907)	2_1977(0.988), 2_1976(0.941),	Liquid fuels sector year 1976, 1977, 1978, 1979, 1980
PC7	1_1975(1.541), 1_1976(1.486), 1_1977(1.462)	1_1974(1.500), 1_1978(1.485),	Emissions from gas fuel year 1974, 1975, 1976, 1977, 1978
PC8	3_1977(1.799), 3_1974(1.682), 3_1976(1.669)	3_1978(1.693), 3_1975(1.677),	Manufacturing and construction sector year 1974, 1975, 1976, 1977, 1978

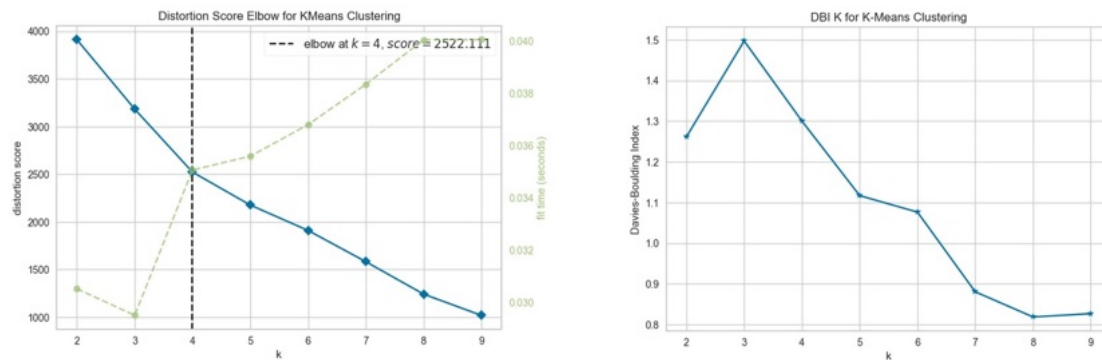
other principal component combinations reveal considerable overlap, reducing the clarity of cluster distinctions.

Figure 9 shows the box plot of each cluster versus the principal components (PCs) that provides insight into the rankings of clusters based on various emission sectors. PC1, representing the solid fuel sector, shows Cluster 1 with the highest value, followed by Cluster 4 (second lowest), Cluster 3 (third lowest) and Cluster 2 (lowest). For PC2, related to the manufacturing and construction sector, Cluster 4 has the highest value with a broad distribution, while Cluster 1 is the third lowest, Cluster 2 is the second lowest, and Cluster 3 is the lowest. In PC3, which represents the residential and commercial sector, Cluster 2 has the highest value, with Cluster 1 being the second lowest, Cluster 3 the third lowest, and Cluster 4 the lowest. PC4, also from the manufacturing and construction sector, ranks Cluster 2 highest, followed by Cluster 4 (second highest), Cluster 1 (third highest), and Cluster 3 (lowest). For PC5, representing transportation emissions, Cluster 4 has the highest value, followed by Cluster 3, Cluster 2, and Cluster 1 (lowest). PC6, related to the liquid fuel sector, shows Cluster 4 as the highest, with Cluster 2 second, Cluster 1 third, and Cluster 3 the lowest. PC7, covering the gas fuel sector, ranks Cluster 2 as the highest, followed by Cluster 3, Cluster 4, and Cluster 1 (lowest). Lastly, PC8, linked to the manufacturing and commercial sector, shows Cluster 3 with the highest value, followed by Cluster 2, Cluster 1, and Cluster 4 (lowest).

From this analysis, it can be concluded that the countries grouped for each cluster are

Table 3. Cluster Labeling Results

Cluster	Countries	Label
Cluster 1	Australia, India, Poland, South Africa	Countries with the highest CO ₂ emissions in the solid fuel sector
Cluster 2	Brazil, Indonesia, Mexico, Saudi Arabia	Countries with the highest CO ₂ emissions in the liquid fuel sector
Cluster 3	Canada, Germany, Iran, Italy, Russian Federation, United Kingdom, United States	Countries with the highest CO ₂ emissions in the gas fuel sector
Cluster 4	China, Japan, Korea Rep., Turkey, Vietnam	Countries with the highest CO ₂ emissions in the manufacturing and construction sector

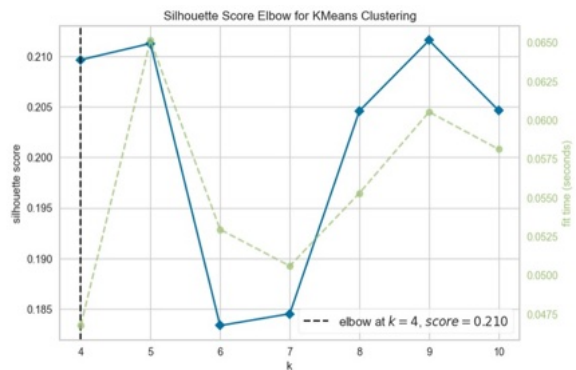
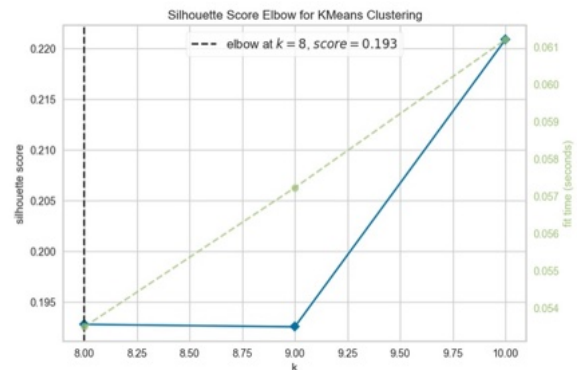
Figure 5. Evaluation metrics for K value with Davies-Bouldin Index and Elbow Method

represented as follows:

- Cluster 1: Countries with CO₂ emissions caused by the solid fuel sector.
- Cluster 2: Countries with CO₂ emissions caused by the residential and commercial sector, the manufacturing and construction sector, and the gas fuel sector.
- Cluster 3: Countries with CO₂ emissions caused by the manufacturing and commercial sector.
- Cluster 4: Countries with CO₂ emissions caused by the manufacturing and construction sector, the transportation sector, and the liquid fuel sector.

Due to the overlap in sector representation for some clusters, precise labeling remains challenging. The next step involves using the EDA visualization results to further refine the group labels, as shown in Figure 10.

Based on the analysis presented in Table 3 and Figure 10, each cluster contains countries that rank among the top 5-10 CO₂ emitters in specific emission sectors. Cluster 1 includes countries that are among the top 5 in average CO₂ emissions from the solid fuel sector. Cluster 2 comprises countries that rank among the top 5 in average CO₂ emissions from the liquid fuel sector. Cluster 3 consists of countries that are among the top 10 in average CO₂ emissions from the gas fuel sector. Finally, Cluster 4 includes countries that are among the top 10 in average CO₂ emissions from the manufacturing and construction sector. The bar graph, which shows the average CO₂ emissions

Figure 6. *Silhouette Score results for K=4*Figure 7. *Silhouette Score results for K=8*

per sector, proved to be more effective for accurate labeling than the box plot, particularly because Clusters 2 and 4 had multiple possible labels. Thus, the cluster labels were finalized on the basis of the EDA barplot results.

3.3. Insights Based on Clustering Results

As seen in Table 4, the first cluster includes Australia, India, Poland, and South Africa. These countries rely heavily on solid fuels such as coal. India and Australia are the main coal producers, with Poland generating 80% of its energy from coal. South Africa also has substantial coal reserves [13]. Coal power plants in these countries contribute significantly to CO₂ emissions in the solid fuel sector [14, 15].

The second cluster consists of Brazil, Indonesia, Mexico, and Saudi Arabia, all of which are significant producers and consumers of liquid fuels, especially oil. These countries are among the top 10 global oil consumers, with daily consumption exceeding 2 million barrels [16]. Indonesia, in particular, uses oil extensively in its transportation sector [17, 18].

The third cluster includes Canada, Germany, Iran, Italy, Russia, the United Kingdom, and the United States, focusing on gas fuels. These countries have a robust infrastructure for natural gas production and consumption, all of which are among the top 20 natural gas producers worldwide. The United States was the leading producer from 2014 to 2016 [19].

The last cluster comprises China, Japan, South Korea, Turkey, and Vietnam, which are prominent in manufacturing and construction. China leads globally in industrial production with high CO₂ emissions from its manufacturing sector. Vietnam is a key manufacturing hub, especially in electronics, textiles, and footwear. Japan and South Korea are known for their advanced manufacturing sectors, and Turkey's economic growth is driven by its large construction industry [20].

3.4. Policy/Mitigation Decisions Based on Clustering Results

The countries in the first cluster need to diversify their energy sources by investing in renewables such as solar, wind, and biomass instead of relying on coal-fired power plants. These countries are the major CO₂ emitters of solid fuels, underscoring the role of coal in their high emissions. Solid biofuels from biomass sources such as agricultural products, forestry, and wood industry waste offer a sustainable alternative [21]. Reforestation and sustainable forest management are vital for CO₂ absorption and ecosystem balance [22].

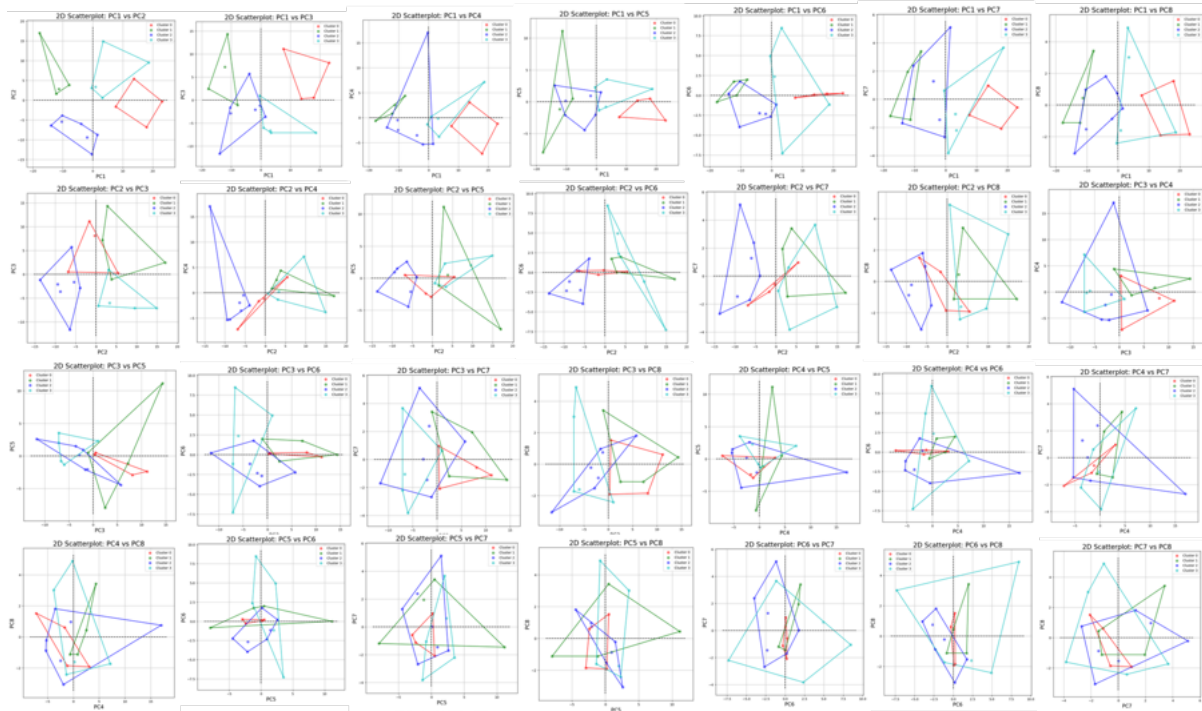


Figure 8. *Two-Dimensional Visualization Using Principal Components and Cluster Separation*

In the second cluster—Brazil, Indonesia, Mexico and Saudi Arabia—known for high oil production and consumption, the transition to electric vehicles, the development of charging infrastructure, and the promotion of biofuels are crucial. Improved public transportation policies can also help reduce the dependence on private vehicles, which might consequently reduce carbon dioxide emissions [23, 24].

Third countries should focus on developing clean natural gas technologies to reduce methane emissions and invest in green and low-emission technologies [25]. The fourth cluster—China, Japan, South Korea, Turkey, and Vietnam—requires the adoption of energy-efficient manufacturing and sustainable construction practices, including waste reduction, the use of recycled and low-carbon materials, prefabrication and renewable energy sources [26].

4. CONCLUSION AND RECOMMENDATION

This study demonstrates that a data science approach can be utilized to guide climate policy formulation. Furthermore, the findings suggest that understanding emission clusters enables the formulation of targeted climate policies and international cooperation strategies. Using a data science approach, this study also reveals that a one-size-fits-all policy is insufficient to address CO₂ emissions in diverse national contexts, because each country has their own characteristics.

Clustering analysis provides information on targeted CO₂ emission reduction strategies tailored to the characteristics of each sector. Using a data science approach for policy-making is expected to be effective in mitigating climate change impacts. Future research should expand the dataset beyond 20 countries and include additional variables such as GDP or economic growth. Other data sources also should be included for more comprehensive analysis, such as Gross Domestic

Product (GDP). This will improve policy formulation and help achieve Net Zero Emission targets by 2060. Collaboration with experts in related fields is also recommended to validate mitigation results. Related to algorithms, other clustering algorithms can also be used, such as dynamic time-series clustering or Dynamic Time Warping (DTW) algorithm. Furthermore, the mitigation strategies should also be validated by climate change policy experts.

ACKNOWLEDGEMENT

This research was partially funded by UPER-UTP Collaborative Research 2023.

REFERENCES

- [1] M. Crippa, D. Guizzardi, F. Pagani, M. Banja, M. Muntean, E. Schaaf, W. Becker, F. Monforti-Ferrario, R. Quadrelli, A. Risquez Martin, P. Taghavi-Moharamli, J. Köykkä, G. Grassi, S. Rossi, J. Brandao De Melo, D. Oom, A. Branco, J. San-Miguel, and E. Vignati, *JRC Science for Policy Report: GHG Emissions of All World Countries*. Publications Office of the European Union, 2023. [Online]. Available: [https://edgar.jrc.ec.europa.eu/booklet_GHG_emissions_of_all_world_countries_booklet_2023\\$report.pdf](https://edgar.jrc.ec.europa.eu/booklet_GHG_emissions_of_all_world_countries_booklet_2023$report.pdf)
- [2] EPA, “Impacts of climate change,” <https://www.epa.gov/report-environment/greenhouse-gases#impacts>, July 2023, diakses pada tanggal 02 April 2024.
- [3] R. Quadrelli and S. Peterson, “The energy–climate challenge: Recent trends in co2 emissions from fuel combustion,” *Energy Policy*, vol. 35, no. 11, pp. 5938–5952, 2007. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0301421507003126>
- [4] A. Ahdiat, “Emisi gas rumah kaca indonesia meningkat pada 2022, tembus rekor baru,” *Databoks*, September 2023. [Online]. Available: <https://databoks.katadata.co.id/datapublish/2023/09/29/emisi-gas-rumah-kaca-indonesia-meningkat-pada-2022-tembus-rekor-baru>
- [5] W. Eko Cahyono, Parikesit, B. Joy, W. Setyawati, and R. Mahdi, “Projection of co2 emissions in indonesia,” *Materials Today: Proceedings*, vol. 63, pp. S438–S444, 2022, 2nd International Conference on Chemical Engineering and Applied Sciences. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2214785322022258>
- [6] Grantham Research Institute on Climate Change and the Environment, London School of Economics and Political Science, “New study finds ai could reduce global emissions annually by 3.2 to 5.4 billion tonnes,” <https://www.lse.ac.uk/granthaminstitute/news/new-study-finds-ai-could-reduce-global-emissions-annually-by-3-2-to-5-4-billion-ton>, Jun. 22 2025, accessed 2025-11-24.
- [7] A. Yunita, H. Santoso, and Z. Hasibuan, “‘everything is data’: towards one big data ecosystem using multiple sources of data on higher education in indonesia,” *Journal of Big Data*, vol. 9, no. 1, 2022.
- [8] A. Mardani, H. Liao, M. Nilashi, M. Alrasheedi, and F. Cavallaro, “A multi-stage method to predict carbon dioxide emissions using dimensionality reduction, clustering, and machine learning techniques,” *Journal of Cleaner Production*, vol. 262, p. 122942, July 2020.
- [9] M. Namazkhan, C. Albers, and L. Steg, “A decision tree method for explaining household gas consumption: The role of building characteristics, socio-demographic variables, psychological factors and household behaviour,” *Renewable and Sustainable Energy Reviews*, vol. 119, p. 109542, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1364032119307506>
- [10] World Bank, “World development indicators,” World Bank Group, 2024, accessed: 2026-01-28. [Online]. Available: <https://databank.worldbank.org/source/world-development-indicators>

- [11] W. Pratama, “Konsumsi BBM besar karena ketergantungan kendaraan pribadi,” <https://ekonomi.bisnis.com/read/20230117/98/1619093/konsumsi-bbm-besar-karena-ketergantungan-kendaraan-pribadi>, 2023, diakses pada tanggal 25 Juni 2024.
- [12] C. A. Romero, P. Correa, E. A. Ariza Echeverri, and D. Vergara, “Strategies for reducing automobile fuel consumption,” *Applied Sciences*, vol. 14, no. 2, 2024. [Online]. Available: <https://www.mdpi.com/2076-3417/14/2/910>
- [13] D. Maarif, “Apa hasil tambang utama negara Afrika Selatan?” <https://tirto.id/apa-hasil-tambang-utama-negara-afrika-selatan-gVbV>, 2024, diakses pada tanggal 25 Juni 2024.
- [14] B. Imanda, “10 negara penghasil batu bara terbesar di dunia, Indonesia salah satunya,” <https://www.solarkita.com/blog/10-negara-penghasil-batu-bara-terbesar-di-dunia-indonesia-salah-satunya>, 2024, diakses pada tanggal 25 Juni 2024.
- [15] P. Guo, H. Shen, Y. Chen, H. Dai, Z. Mai, R. Xu, R. Zhang, Z. Wang, J. He, L. Zheng, H. Z. Sun, K. Ke, J. Meng, M. Liu, J. Li, W. Adalibieke, C. Wang, J. Ye, L. Zhu, G. Shen, T.-M. Fu, A. Tsang, X. Yang, A. G. Russell, C. T. Driscoll, and S. Tao, “Carbon dioxide emissions from global overseas coal-fired power plants,” *Nature Climate Change*, vol. 14, no. 11, pp. 1151–1157, 11 2024. [Online]. Available: <https://doi.org/10.1038/s41558-024-02114-y>
- [16] C. Annur, “10 negara dengan konsumsi minyak bumi terbesar dunia (2022),” <https://databoks.katadata.co.id/datapublish/2023/10/06/10-negara-konsumen-minyak-bumi-terbesar-dunia-2022-as-juara>, 2023, diakses pada tanggal 25 Juni 2024.
- [17] A. Wibowo, “Penggunaan minyak di berbagai sektor industri di Indonesia,” <https://data.goodstats.id/statistic/penggunaan-minyak-di-berbagai-sektor-industri-di-indonesia-ejrtb>, 2023, diakses pada tanggal 25 Juni 2024.
- [18] Deendarlianto, A. Widyaparaga, T. Widodo, I. Handika, I. Chandra Setiawan, and A. Lindasista, “Modelling of Indonesian road transport energy sector in order to fulfill the national energy and oil reduction targets,” *Renewable Energy*, vol. 146, pp. 504–518, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0960148119310092>
- [19] B. P. Statistik, “Produksi minyak mentah (ribu metrik ton), 2014–2016,” <https://www.bps.go.id/id/statistics-table/2/MzUzIzI=/produksi-minyak-mentah.html>, 2018, diakses pada tanggal 25 Juni 2024.
- [20] W. B. Group, “The world bank,” <https://www.worldbank.org/en/country>, 2024, diakses pada tanggal 25 Juni 2024.
- [21] T. Kalak, “Potential use of industrial biomass waste as a sustainable energy source in the future,” *Energies*, vol. 16, no. 4, 2023. [Online]. Available: <https://www.mdpi.com/1996-1073/16/4/1783>
- [22] Amri, Khairul and Ningrum, Sinta, “Sustainable forestry policy: Indonesia’s adaptation in supporting sustainable development goals (sdgs),” *E3S Web Conf.*, vol. 611, p. 03005, 2025. [Online]. Available: <https://doi.org/10.1051/e3sconf/202561103005>
- [23] L. C. Voumik, M. A. Islam, A. Rahaman, and M. M. Rahman, “Emissions of carbon dioxide from electricity production in ASEAN countries: GMM and quantile regression analysis,” *SN Business & Economics*, vol. 2, no. 9, p. 133, 8 2022. [Online]. Available: <https://doi.org/10.1007/s43546-022-00318-y>
- [24] A. Sugiyono, I. Fitriana, I. Rahardjo, and J. Santosa, “Peran kendaraan bermotor listrik berbasis
-

baterai dalam mengurangi permintaan bbm di indonesia,” *JTERA (Jurnal Teknologi Rekayasa)*, vol. 7, pp. 65–72, 2022.

- [25] C. A. T. Force, “Oil and gas methane mitigation programoil and gas methane mitigation program,” <https://www.catf.us/methane/mitigation-program/>, 2024, diakses pada tanggal 26 Juni 2024.
 - [26] M. Senjak, I. Pesko, P. M.D, and V. Mučenski, “Reduction of carbon emissions in the construction industry using lean practices,” *Journal of Computational and Applied Mathematics*, 2023.
-

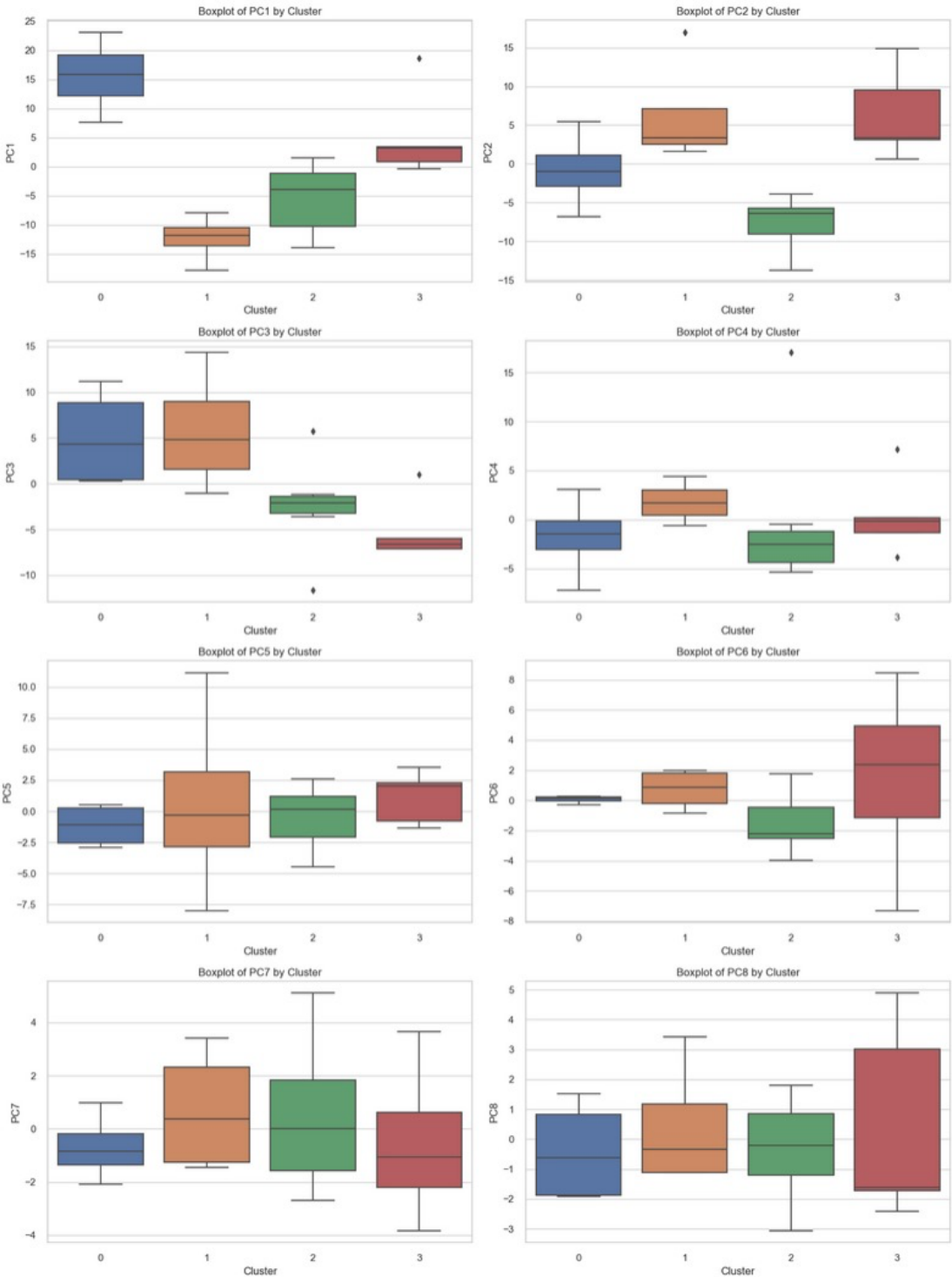


Figure 9. *Boxplot of Clustering Results*

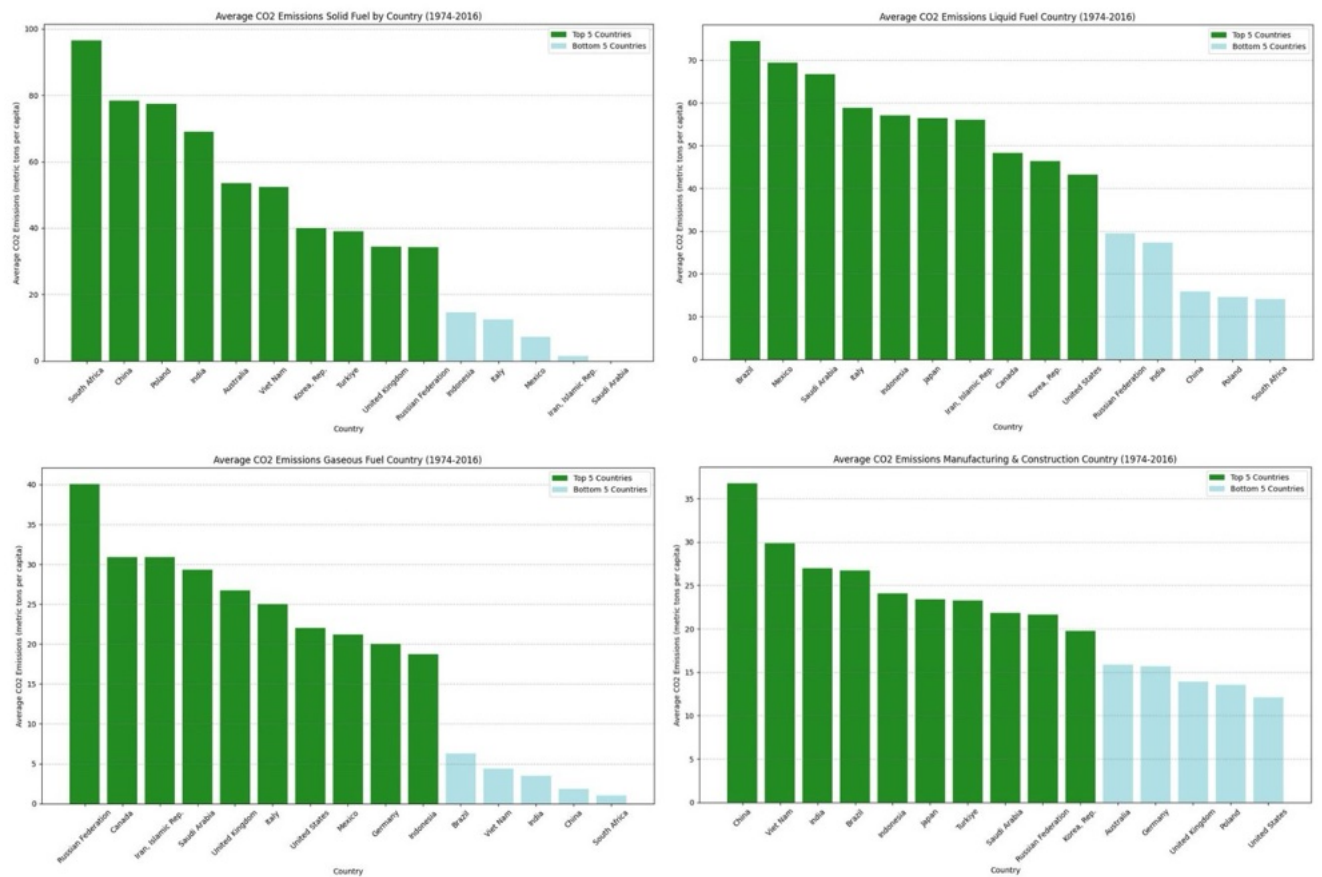


Figure 10. Barplot of Four Sectors for Cluster Labeling

Table 4. Cluster insights and mitigation strategies for CO₂ emission reduction.

Cluster & Countries	Insights & Mitigation Strategies
Cluster 1 (Solid Fuel) Australia, India, Poland, South Africa	<i>Insights:</i> • Economy relies heavily on solid fuel (coal) • India and Australia: coal producers in the world • Poland: 80% of energy comes from coal • South Africa: Significant coal reserves <i>Mitigation Strategies:</i> • Diversify energy mix: solar, wind, and biomass investment • Utilize solid biofuel • Reforestation and sustainable forest management
Cluster 2 (Liquid Fuel) Brazil, Indonesia, Mexico, Saudi Arabia	<i>Insights:</i> • Focus on liquid fuel (oil) • Among the top 10 largest oil consumers globally • Indonesia: 62% of oil is used in the transportation sector <i>Mitigation Strategies:</i> • Transition to electric vehicles: Charging infrastructure and biofuel • Transportation policy: Improving public transportation systems
Cluster 3 (Gas Fuel) Canada, Germany, Iran, Italy, Russia, United Kingdom, United States	<i>Insights:</i> • Focus on the gas fuel sector • All countries rank among the top 20 natural gas producers globally • United States: Largest producer of natural gas <i>Mitigation Strategies:</i> • Development of clean natural gas technologies, such as biogas and hydrogen • Investment in green and low-emission technologies
Cluster 4 (Manufacturing) China, Japan, South Korea, Turkey, Vietnam	<i>Insights:</i> • Focus on manufacturing and construction industries • China: World’s largest industrial producer with high manufacturing CO₂ emissions • Vietnam: Manufacturing hub for electronics, textiles, and footwear • Japan and South Korea: Advanced automotive and electronics manufacturing • Turkey: Economic growth driven by construction industry <i>Mitigation Strategies:</i> • Energy efficiency in manufacturing and construction • Sustainable construction practices: Waste reduction, use of recycled materials, and renewable energy