Project Title: Cluster Analysis for GHG Emission Profiles of Vehicles

# Introduction

The increasing concern over greenhouse gas (GHG) emissions from vehicles has necessitated the need for better analysis and categorization of vehicles based on their emission profiles. With global warming and environmental sustainability at the forefront of policy discussions, reducing vehicle emissions has become a critical goal for governments and industries worldwide. Clustering techniques help identify meaningful patterns in vehicle emissions, allowing policymakers, manufacturers, and consumers to make informed decisions about fuel efficiency and environmental impact.

Understanding vehicle emission clusters can help reduce carbon footprints, improve fuel efficiency policies, and guide consumers toward environmentally friendly vehicle choices. With transportation contributing significantly to global CO2 emissions, a structured analysis of vehicle emissions enables stakeholders to develop and implement policies that promote sustainability. By identifying clusters of vehicles based on emissions and fuel efficiency, we can pinpoint areas for improvement, encourage innovation in cleaner technologies, and support more informed consumer choices. Manufacturers can also use this information to develop better low-emission vehicles, ensuring compliance with stricter environmental regulations.

The primary objectives of this study are:

- To categorize vehicles based on their greenhouse gas emissions.
- To compare different clustering techniques and assess their effectiveness in grouping vehicles.
- To provide actionable insights for policymakers, manufacturers, and consumers based on the clustering results.

Target Audience

This analysis is intended for multiple stakeholders. Environmental policymakers can use this data to design regulations and incentives that promote low-emission vehicles, while automotive manufacturers can refine vehicle designs to develop cleaner, more efficient models. Consumers benefit by making informed purchasing decisions based on fuel efficiency and emissions data. Additionally, researchers in sustainable transportation can leverage this information to advance studies on reducing vehicle emissions through technological innovations and behavioral insights.

### Data Source and Data Metrics

The dataset used for this analysis was obtained from a Kaggle and government database on vehicle emissions, fuel consumption, and engine specifications. This dataset includes key attributes necessary for clustering vehicles based on their emission profiles. It includes fuel consumption rates, CO2 emissions, engine specifications, and vehicle class information, making it ideal for an emissions-based clustering study.  These features provide a comprehensive basis for clustering vehicles based on their environmental impact. By applying clustering algorithms, we aim to identify natural groupings among vehicles, allowing for a structured analysis of how different categories contribute to overall emissions. The insights derived from this analysis can inform fuel efficiency standards, tax policies, and the development of new technologies aimed at reducing emissions.

## Methods and Results

### Data Preparation

To ensure robust clustering results, the dataset underwent multiple preprocessing steps:

- Data Cleaning: Missing values were checked, and rows with incomplete data were removed to maintain dataset integrity.
- Feature Selection: Non-essential attributes, such as vehicle make and model, were excluded to prevent unnecessary complexity.
- Encoding Categorical Variables: Fuel type and transmission type were one-hot encoded to convert them into numerical representations suitable for clustering algorithms.

- Standardization: Continuous numerical features, such as fuel consumption and CO2 emissions, were standardized to ensure equal weighting during clustering.
- Outlier Detection and Removal: The z-score method was applied to detect extreme values, and outliers falling beyond three standard deviations from the mean were removed to prevent them from distorting cluster formations.
- Dimensionality Reduction: Principal Component Analysis (PCA) was used to reduce the dataset to three principal components for visualization while retaining most of the variance.

Modeling Approach

To ensure a comprehensive analysis of vehicle emission profiles, multiple clustering models were applied. Each clustering algorithm was selected based on its ability to capture different aspects of the dataset.

- K-Means Clustering: This method was chosen for its efficiency in identifying compact, well-separated clusters. The number of clusters was determined using the elbow method, which suggested that three clusters provided a good balance between accuracy and interpretability.
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise): DBSCAN was used to detect clusters of arbitrary shapes and to identify outliers in the dataset. Unlike K-Means, DBSCAN does not require specifying the number of clusters in advance and is particularly effective at detecting noise in the data.
- Gaussian Mixture Model (GMM): This probabilistic approach was applied to model the data as a mixture of Gaussian distributions. It was useful for identifying overlapping clusters and allowed for a more flexible representation of the data.
- Hierarchical Clustering: This method was used to explore nested relationships between vehicles with similar emission profiles. The dendrogram visualization helped assess potential sub-clusters that might not be evident through other clustering techniques.

Each algorithm provided unique insights into the dataset, with K-Means and GMM offering a clear separation of vehicle categories, DBSCAN effectively identifying outliers, and hierarchical clustering providing a more granular perspective of vehicle groupings.

Evaluation Metrics

To assess the clustering performance, the following metrics were used:

- Silhouette Score: This metric measures how well-separated clusters are by evaluating how similar a data point is to its assigned cluster compared to other clusters. A higher silhouette score indicates better-defined clusters with minimal overlap.
- Davies-Bouldin Index: This metric assesses cluster compactness and separation by analyzing the ratio of intra-cluster distances to inter-cluster distances. A lower Davies-Bouldin index indicates more distinct and well-separated clusters.

## Clustering Results

K-Means clustering successfully identified three distinct clusters based on vehicle emissions and fuel consumption. The algorithm grouped vehicles with similar characteristics efficiently, with cluster centroids clearly differentiating between low, moderate, and high-emission vehicles. The PCA-reduced visualization confirmed well-separated clusters, and silhouette scores indicated fairly good cohesion and separation, though some overlap was observed in the moderate category.

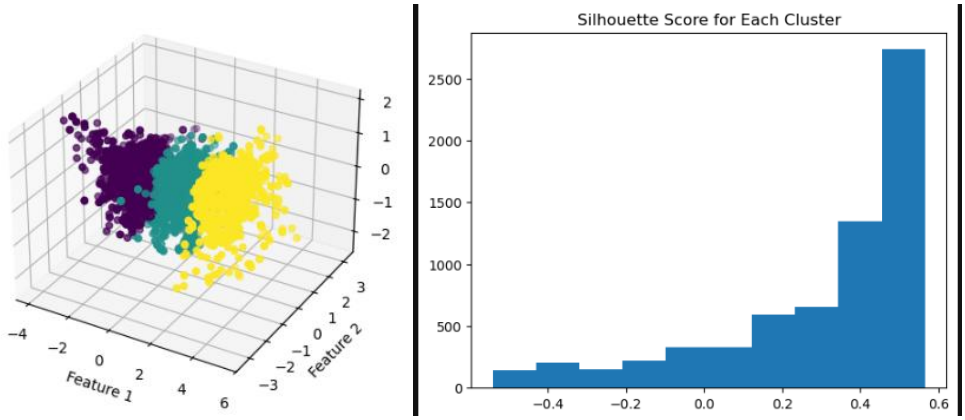


Figure 1: 3D Scatter Plot of Clusters (Left) and Histogram of Silhouette Scores (Right)

The clear separation between clusters suggests strong clustering performance, with minimal overlap ensuring meaningful differentiation. The three principal components effectively separate the clusters, emphasizing distinct boundaries and relationships within the data, as seen in Figure 1 above. Most clusters have positive silhouette scores, indicating good separation, while some clusters exhibit lower scores, signaling areas for improvement. Additionally, a small number of negative scores suggest potential misclassifications in a few cases.

DBSCAN revealed multiple clusters and effectively identified outliers that did not fit into well-defined categories. As seen below in figure 2, the main cluster (Cluster 0) included most vehicles with standard emission levels, while smaller clusters (e.g., Clusters 2, 5, 7, 10, and 12) captured outliers such as high-performance or alternative-fuel vehicles. The advantage of DBSCAN was its ability to detect outliers without requiring a predefined number of clusters.

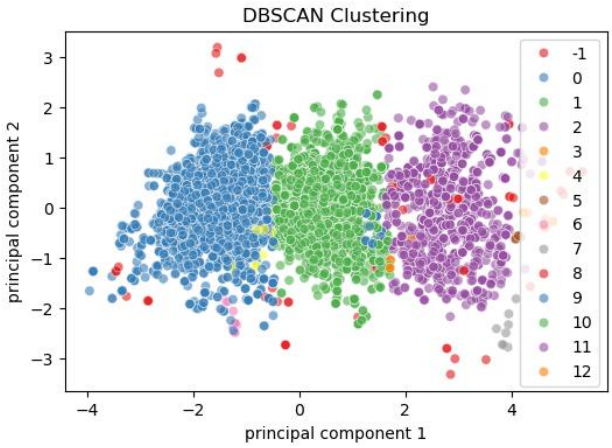


Figure 2: Scatter plot for DBSCAN clustering

GMM clustering also identified three primary clusters. This can be seen below in figure 3, the results aligned with K-Means but provided a more probabilistic interpretation of cluster membership. Vehicles with overlapping characteristics were assigned soft cluster memberships rather than hard classifications.

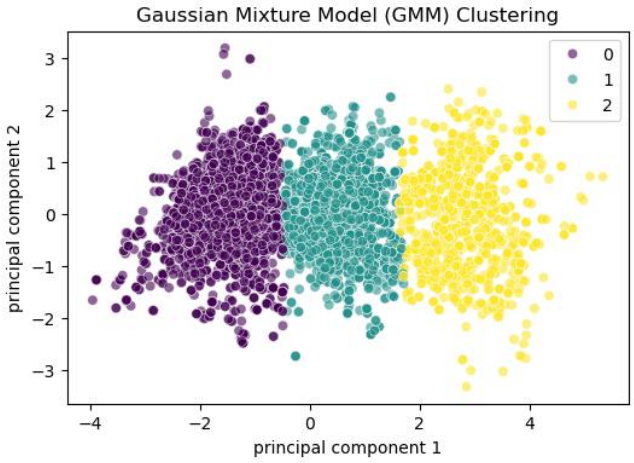


Figure 3: Scatter plot for GMM clustering

Hierarchical clustering provided additional insights into the relationships between vehicle categories. The dendrogram allowed for a hierarchical breakdown of clusters, revealing potential subcategories within high-emission vehicles. This approach was useful for understanding nested structures but was computationally intensive for large datasets.

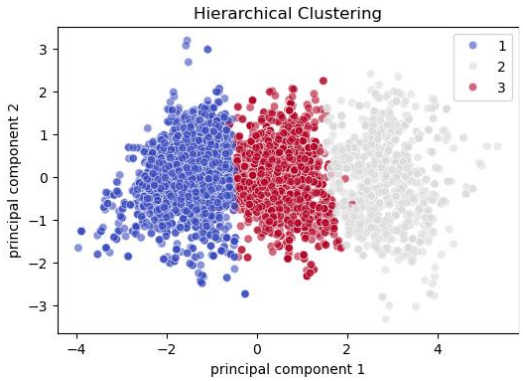


Figure 4: Scatter plot for Hierarchical clustering

When we compare the above models, each clustering method exhibited unique strengths. K-Means provided clear separation but was sensitive to outliers, while DBSCAN excelled at identifying outliers but required careful parameter tuning. GMM performed well with Gaussian-distributed data but was less effective at detecting outliers. Hierarchical clustering was useful for understanding hierarchical relationships but proved computationally expensive for large datasets.

**Indicators of CO2 emissions**

A correlation analysis was also performed, providing a visual representation of the correlation matrix between different variables. This allowed for a clearer understanding of the relationships between the variables and helped identify any strong or weak correlations.

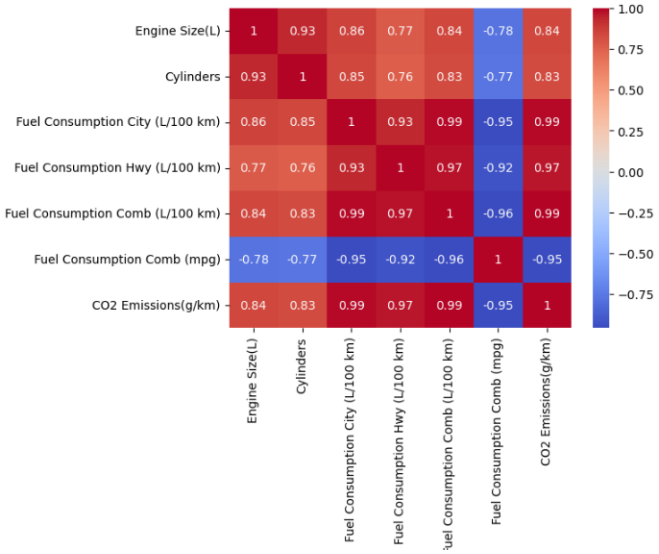


Figure 5: correlation heatmap

As shown in Figure 5 above, several key correlations are observed in the analysis. There is a strong positive correlation between CO2 emissions (g/km) and fuel consumption (L/100 km), which is expected, as burning more fuel results in higher CO2 emissions. Engine size and the number of cylinders also show moderate positive correlations with both fuel consumption and CO2 emissions, as larger engines tend to consume more fuel and produce higher emissions. These strong correlations indicate that one variable could potentially be used to predict another, such as predicting CO2 emissions based on fuel consumption.

**Brand Emission Cluster**

As seen in figure 6, the clustering analysis reveals distinct brand distributions across clusters, highlighting differences in vehicle characteristics. MERCEDES-BENZ and CHEVROLET have a high concentration in Cluster 2, possibly indicating similarities in fuel efficiency, engine size, or emissions. PORSCHE stands out with a significantly high presence in Cluster 1, suggesting a unique profile, likely due to its high-performance vehicle lineup. CHEVROLET, FORD, and MINI are predominantly found in Cluster 0, which may suggest common features such as compact or fuel-efficient models. FORD, however, has a notable presence in both Clusters 0 and 1, indicating a diverse range of vehicles. Similarly, BMW is highly concentrated in Clusters 1 and 2, reflecting variations within its lineup, possibly distinguishing between luxury sedans and performance-focused models. These clustering patterns provide valuable insights into how different automobile brands group based on key attributes such as emissions, fuel consumption, and engine specifications.

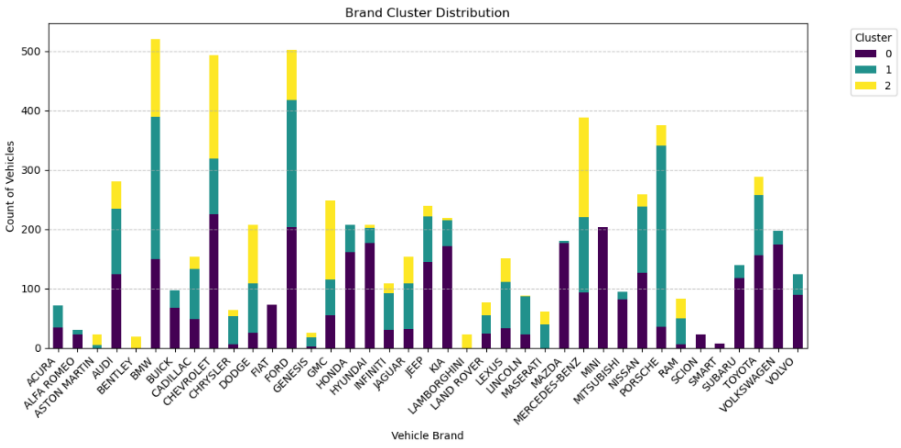


Figure 6: Brand cluster distribution

## Conclusion and Recommendations

The clustering analysis of vehicle emissions has provided valuable insights into the structure of fuel consumption and CO2 emissions data. The results indicate that the data naturally forms distinct clusters, enabling meaningful categorization of vehicles based on their environmental impact. K-Means and GMM effectively identified well-separated groups, while DBSCAN excelled in detecting outliers and anomalies. Hierarchical clustering provided a layered understanding of the data, revealing potential subcategories within emissions groups. These findings emphasize the importance of selecting the appropriate clustering algorithm depending on the analysis objective, whether for general categorization, outlier detection, or exploring hierarchical relationships.

The identified clusters can serve as a foundation for targeted policy-making and industry advancements. Policymakers can leverage these insights to design regulations, such as stricter fuel efficiency standards for high-emission vehicles and incentives for low-emission alternatives. Consumers can benefit from clearer vehicle categorization, aiding in the selection of environmentally friendly options. Automotive manufacturers should utilize this clustering information to develop more efficient and sustainable vehicle technologies, aligning with evolving environmental standards.

The clustering analysis effectively categorized vehicles into high, medium, and low-emission groups, offering clear distinctions in their environmental impact. Fuel consumption emerged as the strongest predictor of emission levels, highlighting its critical role in determining a vehicle's carbon footprint. The analysis identified three key clusters: Cluster 0, consisting of the most environmentally friendly vehicles, such as hybrid, or highly efficient gasoline-powered cars; Cluster 1, representing vehicles with moderate fuel consumption and emissions, likely standard gasoline-powered models balancing efficiency and performance; and Cluster 2, which includes high-consumption, high-emission vehicles such as high-performance cars. Addressing these variations is essential for advancing sustainability in the automotive industry.

Recommendations

- ✓ Encourage Sustainable Choices: Promote low-emission vehicles through incentives like tax breaks, rebates, and priority parking while discouraging high-emission vehicles with higher taxes.
- ✓ Improve Fuel Efficiency: Support research and development of advanced fuel-saving technologies across all vehicle categories.
- ✓ Implement Regulations: Enforce stricter fuel efficiency and emissions standards, especially for high-consumption vehicles.
- ✓ Educate Consumers: Raise awareness about the environmental impact of different vehicle types and encourage eco-friendly driving practices.
- ✓ Promote Alternative Fuels: Invest in the development and adoption of alternative fuel sources, such as electric, hydrogen, or biofuels, to reduce reliance on fossil fuels.

**Model Deployment and Ethical Considerations**

While the model provides meaningful insights, real-world deployment requires continuous validation with updated datasets. Ethical concerns include potential biases in the dataset, such as underrepresentation of electric vehicles, and ensuring fairness in policy implementation.

**Mitigating Ethical Concerns**

Future analyses should incorporate diverse datasets that represent a wide range of vehicle technologies and global regions to ensure comprehensive insights. Additionally, when proposing policy changes, it is crucial to consider socioeconomic factors to prevent disproportionately impacting lower-income groups and ensure fair and effective implementation.

**Future Work**

Future research should incorporate real-time emission data to improve model accuracy and provide more dynamic insights. Expanding the study to include electric vehicles will ensure a more comprehensive understanding of emission patterns across different technologies. Additionally,

enhancing the interpretability of cluster results will help policymakers and stakeholders make more informed decisions regarding sustainability and regulatory measures.

References


1. International Energy Agency. (2021). *Global Energy and CO2 Status Report 2021*.
2. U.S. Environmental Protection Agency. (2022). *Greenhouse Gas Emissions from a Typical Passenger Vehicle.*
3. European Environment Agency. (2021). *Monitoring CO2 emissions from passenger cars and vans in Europe.*
4. Data Science from Scratch. (2015). Clustering. Page 225-238.