Final Report

Sofia Guttmann

Motivation

The sustainability industry is booming right now, and one of the largest contributors to the discourse is cars. Cars contribute excess CO2 into the environment. In order to be able to make an impact, it is important to know what types of cars can have the least impact without switching to electric vehicles. For my project, I wanted to cluster using hierarchical clusters and dendrograms. I will be spending my summer in Boston working in the sustainability sector. I would like to be able to show my work to my employers to be able to network and discuss ideas.

Data

This dataset is from Kaggle.

https://www.kaggle.com/datasets/sahirmaharajj/fuel-economy

It includes a variety of variables. Since the dataset is so large, I cleaned it to include the variables that I was going to use. The variables I isolated are "year", "engine_cylinders", "city_mpg_ft1", "tailpipe_co2_in_grams_mile_ft1." I also decided to run samples instead of working with the entire dataset to preserve memory and high-speed function. The data presents key information about efficiency and environmental harm.

```
'data.frame':   38113 obs. of  5 variables:
 $ year                          : int  1984 1984 1984 1984 1984 1984 1984 1984 1984 1984 ...
 $ engine_cylinders              : int  6 6 4 4 4 4 6 6 6 6 ...
 $ fuel_type                     : chr  "Regular" "Regular" "Regular" "Regular" ...
 $ city_mpg_ft1                  : int  17 17 18 18 18 18 13 13 15 15 ...
 $ tailpipe_co2_in_grams_mile_ft1: num  444 444 423 423 523 ...
```

Hierarchical Clustering

```
'data.frame':  38113 obs. of  4 variables:
 $ year                          : int  1984 1984 1984 1984 1984 1984 1984 1984 1984 1984 ...
 $ engine_cylinders              : int  6 6 4 4 4 4 6 6 6 6 ...
 $ city_mpg_ft1                  : int  17 17 18 18 18 18 13 13 15 15 ...
 $ tailpipe_co2_in_grams_mile_ft1: num  444 444 423 423 523 ...
```

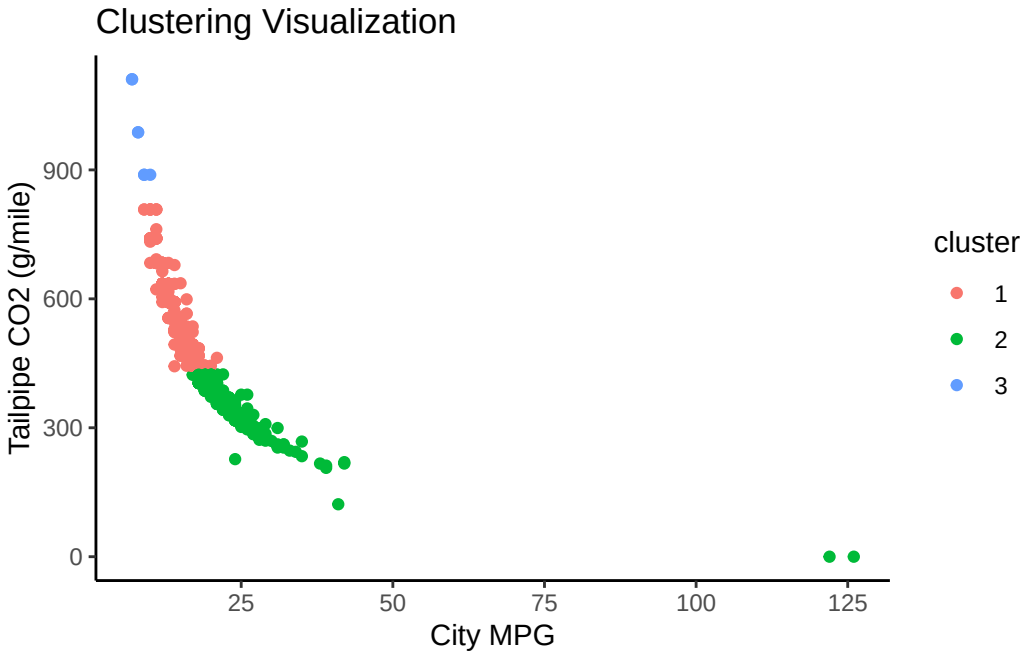
```
'data.frame':  38113 obs. of  4 variables:
 $ year                          : int  1984 1984 1984 1984 1984 1984 1984 1984 1984 1984 ...
 $ engine_cylinders              : int  6 6 4 4 4 4 6 6 6 6 ...
 $ city_mpg_ft1                  : int  17 17 18 18 18 18 13 13 15 15 ...
 $ tailpipe_co2_in_grams_mile_ft1: num  444 444 423 423 523 ...
```

```
   [1] 1 2 2 1 1 1 1 2 1 1 2 1 1 1 2 2 3 1 1 1 1 2 2 2 2 2 1 1 1 2 1 2 2 1 1 1 2
  [38] 1 1 2 2 2 2 1 1 2 1 1 1 1 2 2 2 1 1 2 2 1 1 2 1 1 2 1 1 1 2 1 1 1 1 2 1 2
  [75] 2 1 1 1 2 1 1 1 2 1 1 2 1 2 2 2 1 2 3 1 2 1 1 1 2 1 1 2 1 1 1 1 1 2 1 2 1
 [112] 1 2 1 2 1 1 1 2 2 1 1 2 1 2 2 2 1 1 2 1 1 2 1 1 1 2 1 2 2 1 2 1 2 2 2 1 2
 [149] 1 1 1 1 2 2 2 1 1 2 1 1 1 1 2 2 2 2 1 2 1 2 1 1 2 2 1 1 2 1 1 2 1 1 1 1 2
 [186] 1 2 1 1 2 2 1 2 2 2 1 1 2 2 1 3 2 2 1 2 1 1 2 1 2 2 1 1 2 1 1 1 2 1 1 2 2
 [223] 1 1 1 1 2 2 1 1 1 1 1 2 2 1 2 3 2 1 2 2 1 1 1 1 2 2 1 2 1 2 1 1 2 2 1 2 2
 [260] 2 1 1 2 1 1 1 2 1 2 1 1 1 1 1 1 1 2 2 1 1 1 1 1 2 2 2 2 2 2 2 2 2 1 1 2 1
 [297] 1 1 1 2 1 2 2 1 1 2 1 2 1 1 1 2 1 1 2 1 1 2 1 1 1 2 1 1 1 2 2 1 2 2 1 1 1
 [334] 1 1 2 1 2 1 1 1 1 2 1 2 1 2 1 2 1 2 2 1 2 2 1 1 1 2 1 2 2 1 1 2 1 1 2 1 1
 [371] 1 2 1 1 1 1 1 1 2 2 1 1 1 2 1 2 2 2 2 2 2 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 1
 [408] 2 2 1 1 1 2 2 2 1 2 1 2 2 1 1 1 1 1 1 2 1 2 2 1 1 1 2 1 1 1 1 2 1 2 2 1 1
 [445] 1 1 1 1 1 2 2 1 2 2 1 1 1 1 1 2 2 2 2 1 1 1 1 1 1 1 2 1 1 2 2 1 2 1 1 1 1
 [482] 1 1 1 1 1 2 2 1 1 1 2 1 1 1 1 2 2 2 1 1 2 1 1 1 1 1 2 2 1 1 2 1 1 1 2 1 2
 [519] 1 1 1 1 1 1 2 1 1 1 2 1 1 2 1 2 1 1 1 2 2 1 2 2 2 1 1 1 1 1 2 1 1 2 1 1 1 2
 [556] 1 1 2 2 1 1 1 2 2 1 1 1 1 1 2 1 1 2 1 1 1 2 2 1 2 2 2 2 2 2 2 2 1 2 2 2 2
 [593] 2 2 1 2 2 1 1 1 2 1 2 2 2 1 2 2 1 2 2 2 1 1 2 2 2 1 1 1 2 2 1 2 1 2 2 1 1
 [630] 1 3 2 1 1 1 2 2 1 2 1 1 1 2 1 2 1 1 1 1 2 1 2 1 1 1 1 2 1 2 1 2 1 2 1 1 1
 [667] 1 1 2 2 2 1 1 2 1 1 1 2 1 1 2 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 2 2 2 2 3 1 1
 [704] 1 1 2 2 1 1 1 1 1 1 1 1 2 1 2 1 2 1 2 1 1 3 1 1 2 1 2 1 1 1 2 1 2 1 2 1 1
 [741] 1 2 1 2 1 2 1 1 1 1 2 1 1 1 2 1 2 1 2 2 1 2 2 2 1 2 1 2 2 2 1 2 1 1 2 1 1
 [778] 2 1 1 2 2 2 1 1 1 1 2 1 1 2 2 2 1 1 1 1 2 2 2 1 1 1 2 2 2 1 2 2 1 1 2 1 1
 [815] 1 2 2 2 2 2 1 2 1 2 1 2 2 1 1 2 2 2 2 2 1 2 2 2 2 1 2 1 1 2 1 2 1 2 1 1 1
 [852] 2 2 2 2 2 2 1 1 2 1 1 2 2 2 1 1 2 1 2 2 1 2 2 1 1 1 2 1 1 1 1 1 2 2 1 1 2
 [889] 1 2 1 1 1 1 1 1 1 1 2 1 2 1 1 2 1 2 2 2 2 2 1 2 1 1 2 2 2 1 2 1 1 1 1 2 2
 [926] 1 1 2 1 2 1 1 1 2 1 2 1 1 1 1 2 1 2 1 1 1 1 2 1 2 2 1 1 1 2 2 2 2 2 2 2 1
 [963] 1 2 1 1 1 1 2 1 1 1 2 2 2 1 1 2 1 1 1 1 1 2 1 1 1 2 2 1 2 1 2 2 1 2 1 1 2
[1000] 1
```

```
   [1] 1 2 2 1 1 1 1 2 1 1 2 1 1 1 2 2 3 1 1 1 1 2 2 2 2 2 1 1 1 2 1 2 2 1 1 1 2
  [38] 1 1 2 2 2 2 1 1 2 1 1 1 1 2 2 2 1 1 2 2 1 1 2 1 1 2 1 1 1 2 1 1 1 1 2 1 2
  [75] 2 1 1 1 2 1 1 1 2 1 1 2 1 2 2 2 1 2 3 1 2 1 1 1 2 1 1 2 1 1 1 1 1 2 1 2 1
 [112] 1 2 1 2 1 1 1 2 2 1 1 2 1 2 2 2 1 1 2 1 1 2 1 1 1 2 1 2 2 1 2 1 2 2 2 1 2
 [149] 1 1 1 1 2 2 2 1 1 2 1 1 1 1 2 2 2 2 1 2 1 2 1 1 2 2 1 1 2 1 1 2 1 1 1 1 2
 [186] 1 2 1 1 2 2 1 2 2 2 1 1 2 2 1 3 2 2 1 2 1 1 2 1 2 2 1 1 2 1 1 1 2 1 1 2 2
```

```
 [223] 1 1 1 1 2 2 1 1 1 1 1 2 2 1 2 3 2 1 2 2 1 1 1 1 2 2 1 2 1 2 1 1 2 2 1 2 2
 [260] 2 1 1 2 1 1 1 2 1 2 1 1 1 1 1 1 1 2 2 1 1 1 1 1 2 2 2 2 2 2 2 2 2 1 1 2 1
 [297] 1 1 1 2 1 2 2 1 1 2 1 2 1 1 1 2 1 1 2 1 1 2 1 1 1 2 1 1 1 2 2 1 2 2 1 1 1
 [334] 1 1 2 1 2 1 1 1 1 2 1 2 1 2 1 2 1 2 2 1 2 2 1 1 1 2 1 2 2 1 1 2 1 1 2 1 1
 [371] 1 2 1 1 1 1 1 1 2 2 1 1 1 2 1 2 2 2 2 2 2 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 1
 [408] 2 2 1 1 1 2 2 2 1 2 1 2 2 1 1 1 1 1 1 2 1 2 2 1 1 1 2 1 1 1 1 2 1 2 2 1 1
 [445] 1 1 1 1 1 2 2 1 2 2 1 1 1 1 1 2 2 2 2 1 1 1 1 1 1 1 2 1 1 2 2 1 2 1 1 1 1
 [482] 1 1 1 1 1 2 2 1 1 1 2 1 1 1 1 2 2 2 1 1 2 1 1 1 1 1 2 2 1 1 2 1 1 1 2 1 2
 [519] 1 1 1 1 1 2 1 1 1 2 1 1 2 1 2 1 1 1 2 2 1 2 2 2 1 1 1 1 1 2 1 1 2 1 1 1 2
 [556] 1 1 2 2 1 1 1 2 2 1 1 1 1 1 2 1 1 2 1 1 1 2 2 1 2 2 2 2 2 2 2 2 1 2 2 2 2
 [593] 2 2 1 2 2 1 1 1 2 1 2 2 2 1 2 2 1 2 2 2 1 1 2 2 2 1 1 1 2 2 1 2 1 2 2 1 1
 [630] 1 3 2 1 1 1 2 2 1 2 1 1 1 2 1 2 1 1 1 1 2 1 2 1 1 1 1 2 1 2 1 2 1 2 1 1 1
 [667] 1 1 2 2 2 1 1 2 1 1 1 2 1 1 2 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 2 2 2 2 3 1 1
 [704] 1 1 2 2 1 1 1 1 1 1 1 1 2 1 2 1 2 1 2 1 1 3 1 1 2 1 2 1 1 1 2 1 2 1 2 1 1
 [741] 1 2 1 2 1 2 1 1 1 1 2 1 1 1 2 1 2 1 2 2 1 2 2 2 1 2 1 2 2 2 1 2 1 1 2 1 1
 [778] 2 1 1 2 2 2 1 1 1 1 2 1 1 2 2 2 1 1 1 1 2 2 2 1 1 1 2 2 2 1 2 2 1 1 2 1 1
 [815] 1 2 2 2 2 1 2 1 2 1 2 2 1 1 2 2 2 2 2 1 2 2 2 2 1 2 1 1 2 1 2 1 2 1 1 1 1
 [852] 2 2 2 2 2 1 1 2 1 1 2 2 2 1 1 2 1 2 2 1 2 2 1 1 1 2 1 1 1 1 1 2 2 1 1 2 2
 [889] 1 2 1 1 1 1 1 1 1 2 1 2 1 1 2 1 2 2 2 2 2 1 2 1 1 2 2 2 1 2 1 1 1 1 2 2 2
 [926] 1 1 2 1 2 1 1 1 2 1 2 1 1 1 1 2 1 2 1 1 1 1 2 1 2 2 1 1 1 2 2 2 2 2 2 2 1
 [963] 1 2 1 1 1 1 2 1 1 1 2 2 2 1 1 2 1 1 1 1 1 2 1 1 1 2 2 1 2 1 2 2 1 2 1 1 2
[1000] 1
```

Hierarchical clustering is a method used in unsupervised machine learning to group similar data points into clusters. It builds a hierarchy of clusters by successively merging or splitting them based on their similarity or dissimilarity. Agglomerative Hierarchical Clustering starts with each data point as a separate cluster and then iteratively merges the most similar clusters until only one cluster remains. I sampled a subset of rows from your dataset to make the clustering process computationally feasible. I scaled the sampled data to ensure that all variables have the same scale, which is important for distance-based methods like hierarchical clustering. I performed hierarchical clustering on the scaled data using the **`hclust()`** function, which calculates the pairwise distances between data points and then applies a linkage method to determine the distance between clusters. Complete linkage is a method used in hierarchical clustering to measure the distance between two clusters. In complete linkage, the distance between two clusters is defined as the maximum distance between any two points in the two clusters. Finally, I interpreted the clusters by examining the characteristics of each cluster, such as the mean values of the variables within each cluster, to understand the differences between them.



In this diagram, we can see that as Mile Per Gallon increases, the Tailpipe CO2 variable decreases. This makes intuitive sense because as engines are more efficient, they release less pollution into the air. Cluster 3 produces the most emission with the lowest Mile Per Gallon rate, Cluster 1 is in the middle, and Cluster 2 has the lowest rates of emission with the highest MPG.

```
# A tibble: 3 x 4
  cluster mean_co2 mean_city_mpg mean_engine_cylinders
    <int>    <dbl>         <dbl>                 <dbl>
1       1     498.          16.2                  6.03
2       2     324.          25.9                 NA
3       3     794.          10.1                  8.20
```

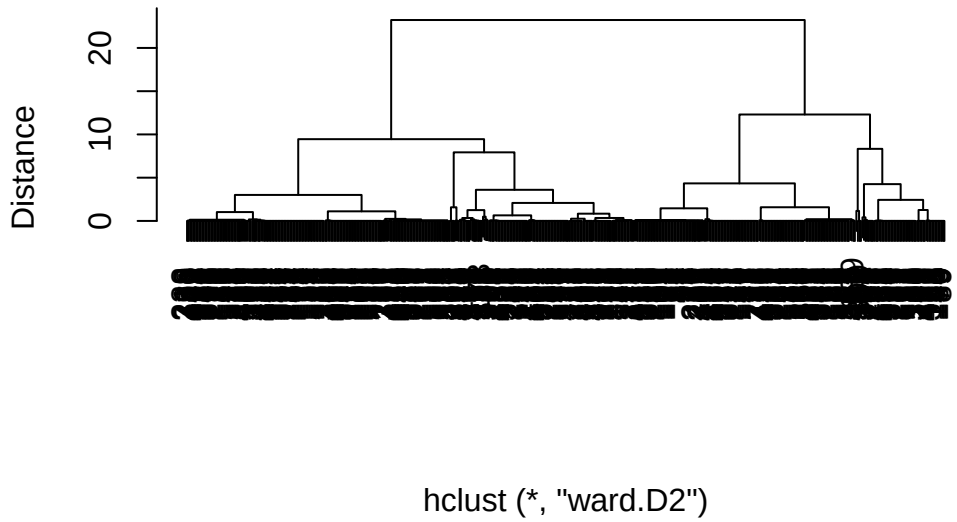
This summary suggests that the data has been clustered into three groups based on the mean values of three variables: CO2 emissions, city miles per gallon (MPG), and engine cylinders. Cluster 1 has the highest mean CO2 emissions, moderate city MPG, and a moderate number of engine cylinders. Cluster 2 has the lowest mean CO2 emissions, highest city MPG, and the number of engine cylinders is not available for this cluster. Cluster 3 has high mean CO2 emissions, low city MPG, and the highest number of engine cylinders among the clusters.

These clusters indicate distinct patterns or groups within the data based on the characteristics of the vehicles represented by these variables. For example, Cluster 2 could represent more fuel-efficient vehicles with lower CO2 emissions, while Cluster 3 could represent less fuel-efficient vehicles with higher CO2 emissions and more powerful engines.

Dendrograms

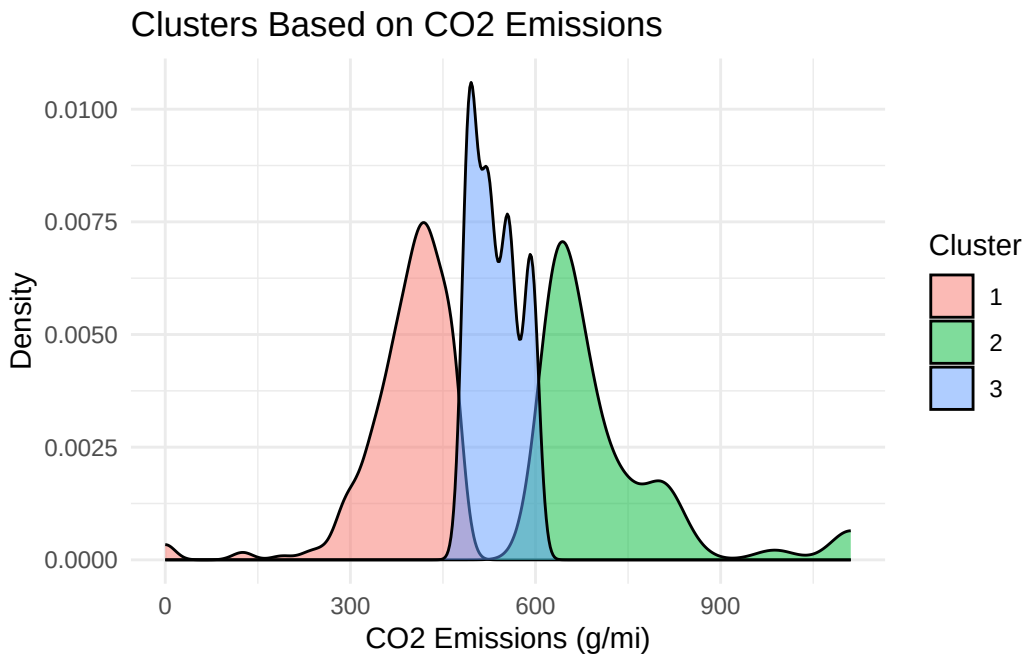
```r
# Load the required libraries
library(readr)
library(dplyr)
library(ggplot2)
```

```r
# Select relevant variable
variable <- "tailpipe_co2_in_grams_mile_ft1"

# Sample the dataset (optional)
set.seed(123)  # Set seed for reproducibility
sampled_data <- Fuel_hclust %>% sample_n(500)  # Adjust the number of samples as needed

# Normalize the data (optional)
scaled_data <- scale(sampled_data[[variable]])

# Compute the distance matrix
distance_matrix <- dist(scaled_data)

# Perform hierarchical clustering
hierarchical_clusters <- hclust(distance_matrix, method = "ward.D2")

# Plot the dendrogram
plot(hierarchical_clusters, main = "Dendrogram of Hierarchical Clustering", xlab = "", ylab = "
```




```r
# Cut the dendrogram to get clusters
num_clusters <- 3  # You can adjust this based on the dendrogram
clusters <- cutree(hierarchical_clusters, k = num_clusters)

# Add cluster labels to the original dataset
sampled_data$cluster <- clusters

# Visualize the clusters
ggplot(sampled_data, aes(x = tailpipe_co2_in_grams_mile_ft1, y = ..density.., fill = factor(clu
  geom_density(alpha = 0.5) +
  labs(x = "CO2 Emissions (g/mi)", y = "Density", title = "Clusters Based on CO2 Emissions") +
```

```
  scale_fill_discrete(name = "Cluster") +
  theme_minimal()
```




```
library(dplyr)
```

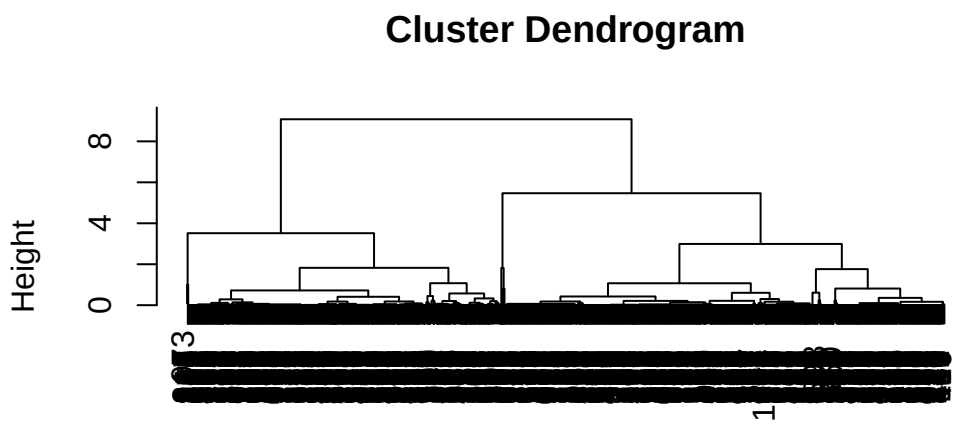
In the context of hierarchical clustering, a dendrogram is a diagram that illustrates the arrangement of clusters created during the clustering process. It's a tree-like structure where the leaves represent individual data points, and the branches represent the merging of clusters as the algorithm progresses. Each merge in the dendrogram corresponds to a level of similarity or dissimilarity between clusters. The height of each fusion in the dendrogram reflects the distance or dissimilarity between the clusters being merged. The longer the branch, the less similar the clusters are. The dendrogram shows that clusters are mainly conglomerated between two large clusters with a relatively large distance.

The second plot visualizes the clusters based on CO2 emissions. It consists of multiple density curves overlaid on the same axis. Each density curve represents a different cluster, distinguished by color. The x-axis represents CO2 emissions measured in grams per mile (g/mi), while the y-axis represents the density of data points within each range of CO2 emissions.

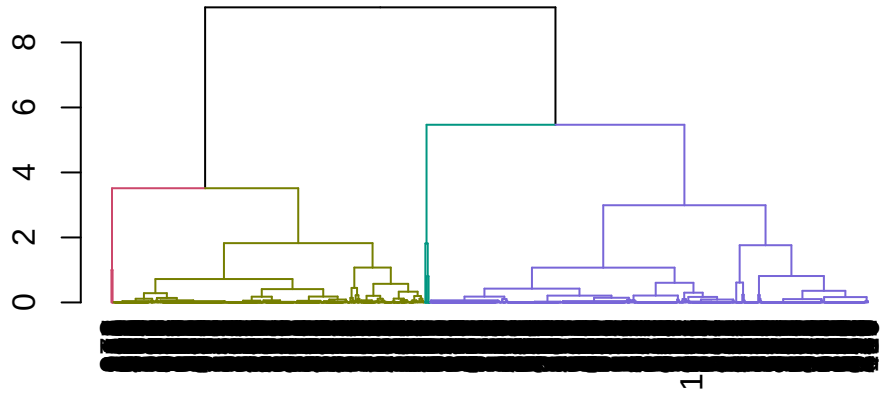
The overlapping density curves provide insights into the distribution of CO2 emissions across the clusters. Areas where the density curves overlap indicate regions of similarity in CO2 emissions between clusters, while distinct peaks or valleys suggest differences in emission levels. The transparency of the curves allows for better visualization of overlapping regions.

The legend on the plot identifies each cluster by color, facilitating interpretation of the density curves. The title of the plot, "Clusters Based on CO2 Emissions," provides context for the analysis, indicating that the clusters were derived from CO2 emission data. Overall, the plot enables viewers to compare the distribution of CO2 emissions across different clusters and identify potential patterns or trends within the data.

```
# Plot the dendrogram
plot(hclust_result)
```




```
suppressPackageStartupMessages(library(dendextend))
avg_dend_obj <- as.dendrogram(hclust_result)
avg_col_dend <- color_branches(avg_dend_obj, h = 3)
plot(avg_col_dend)
```



This code segment first plots a dendrogram resulting from hierarchical clustering (‘hclust_result‘). Then, it loads the ‘dendextend‘ package, which provides functions for manipulating and visualizing dendrogram objects in R. Next, it converts the hierarchical clustering result (‘hclust_result‘) into a dendrogram object (‘avg_dend_obj‘) using the ‘as.dendrogram()‘ function. Afterward, it colors the branches of the dendrogram based on a specified height threshold (‘h = 3‘) using the ‘color_branches()‘ function. Finally, it plots the dendrogram with colored branches (‘avg_col_dend‘). With the addition of the color branches function, we can see the dispersion of the data. Again, this solidifies our understanding of the clusters as two of the clusters being largely close.

Results

To preface my results, I want to make a note of a few things. First, the nature of sampling invites the inability to make specific acknowledgements about the data. These results are overall generalizations. Not

every single datapoint was factored into the clustering process due to scale. The clustering of variables based on CO2 emissions, city miles per gallon (MPG), and engine cylinders provides insights into the relationships between these variables and their impact on CO2 emission levels and fuel efficiency.

1. CO2 Emissions and Fuel Efficiency: The clusters reveal distinct patterns in the relationship between CO2 emissions and fuel efficiency (represented by city MPG). For example, Cluster 2, characterized by low CO2 emissions and high city MPG, suggests a group of vehicles that are more fuel-efficient and emit lower levels of CO2 compared to the other clusters. Conversely, Cluster 3, with high CO2 emissions and low city MPG, indicates less fuel-efficient vehicles emitting higher levels of CO2.

2. Engine Cylinders and CO2 Emissions: The clusters also shed light on the relationship between engine cylinders and CO2 emissions. Cluster 3, which exhibits the highest number of engine cylinders, corresponds to vehicles with elevated CO2 emissions. This suggests a potential correlation between engine size (as indicated by the number of cylinders) and CO2 emissions, with larger engines typically resulting in higher emissions.

3. Complex Relationships: Additionally, the presence of Cluster 1, which shows moderate CO2 emissions, city MPG, and engine cylinders, indicates a more nuanced relationship between these variables. This cluster may represent a diverse group of vehicles with varying engine sizes and fuel efficiency levels, resulting in moderate CO2 emissions.

The clustering analysis highlights how different combinations of variables contribute to variations in CO2 emissions and fuel efficiency among vehicles. It underscores the importance of considering multiple factors, such as engine size and fuel efficiency, when assessing environmental impacts and performance metrics like CO2 emissions.

Companies should diversify their product portfolio to cater to different customer preferences and market segments identified through clustering analysis. They should allocate resources towards the development of fuel-efficient vehicles, especially those with low CO2 emissions and high city MPG, which align with the characteristics of Cluster 2. By tailoring marketing messages to resonate with the identified clusters, companies can highlight the fuel efficiency, environmental benefits, and unique features of each vehicle category to appeal to specific customer segments. This type of technology can be applied to develop online tools or mobile apps that allow customers to compare vehicles based on their CO2 emissions, fuel efficiency ratings, and other relevant factors. Enable customers to make informed decisions that align with their values and preferences. By incorporating these recommendations into their business strategies, companies can leverage the insights from the data analysis to enhance their competitiveness, strengthen brand reputation, and contribute to the transition towards a more sustainable automotive industry.

Source:

Kaggle.com