Data Wrangling

Sofia

```r
# load packages
library(tidyverse)
library(kableExtra)
library(robotstxt)
library(rvest)
library(purrr)
library(readr)
library(tidyr)
library(sf) #for reading shape files

# set code chunk defaults
knitr::opts_chunk$set(tidy = F, # display code as typed
                      size = "small", # slightly smaller code font
                      message = FALSE,
                      warning = FALSE,
                      comment = "\t")

# set black & white default plot theme
theme_set(theme_classic())

# improve digit and NA display
options(scipen = 1, knitr.kable.NA = '')
```

Data Wrangling and Cleaning

Data Source 1: Kaggle- Fuel Data

```
# Read csv file obtained from Database
fuel <- read.csv("fuel.csv")
```

Cleaning

```
# Subset the dataset to include only the specified variables
Fuel_clean <- fuel[, c("year", "engine_cylinders", "fuel_type", "city_mpg_ft1", "tailpipe_

# View the structure of the cleaned dataset
str(Fuel_clean)
```

```
'data.frame':   38113 obs. of  5 variables:
 $ year                          : int  1984 1984 1984 1984 1984 1984 1984 1984 1984 198
 $ 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 ...
```

Hclust

```
# Select only the specified columns
Fuel_hclust <- Fuel_clean[, c("year", "engine_cylinders", "city_mpg_ft1", "tailpipe_co2_in

# View the structure of the cleaned dataset
str(Fuel_hclust)
```

```
'data.frame':   38113 obs. of  4 variables:
 $ year                          : int  1984 1984 1984 1984 1984 1984 1984 1984 1984 198
 $ 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 ...
```

```r
# Select only the specified columns
Fuel_hclust <- Fuel_clean[, c("year", "engine_cylinders", "city_mpg_ft1", "tailpipe_co2_in

# View the structure of the cleaned dataset
str(Fuel_hclust)
```

```
'data.frame':   38113 obs. of  4 variables:
 $ year                         : int  1984 1984 1984 1984 1984 1984 1984 1984 1984 198
 $ 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 ...
```

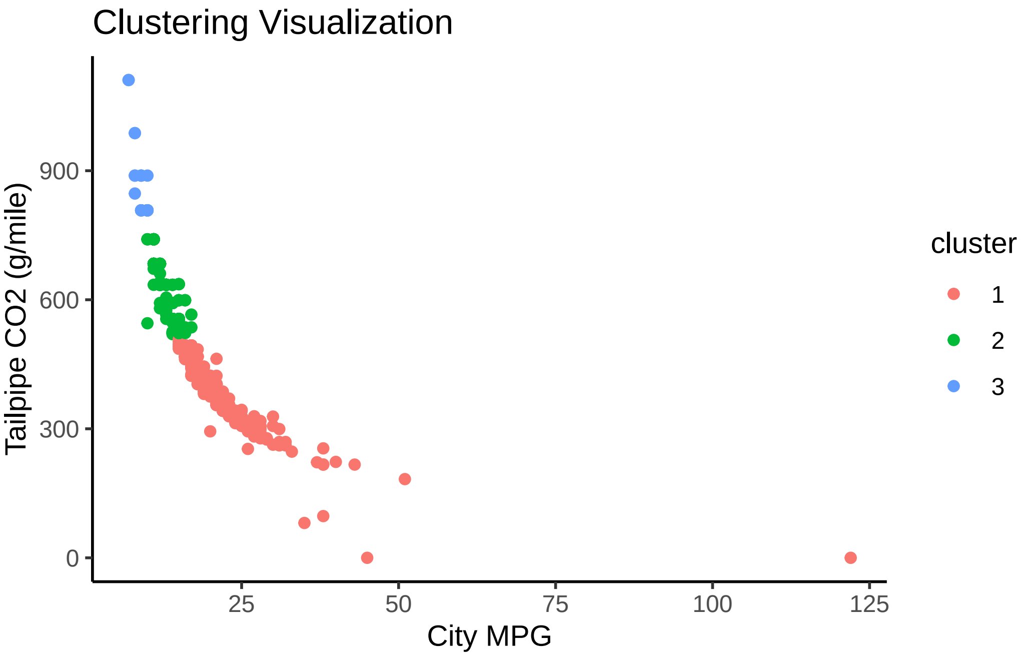
```r
# Sample a subset of rows from your dataset
sampled_data <- Fuel_hclust[sample(nrow(Fuel_hclust), 1000), ]

# Scale the sampled data
scaled_data <- scale(sampled_data[, "tailpipe_co2_in_grams_mile_ft1"])

# Perform hierarchical clustering on the scaled data
hclust_result <- hclust(dist(scaled_data), method = "complete")

# Cut the dendrogram to obtain clusters
num_clusters <- 3  # Adjust the number of clusters as needed
cluster_assignment <- cutree(hclust_result, k = num_clusters)

# View cluster assignments
print(cluster_assignment)
```

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

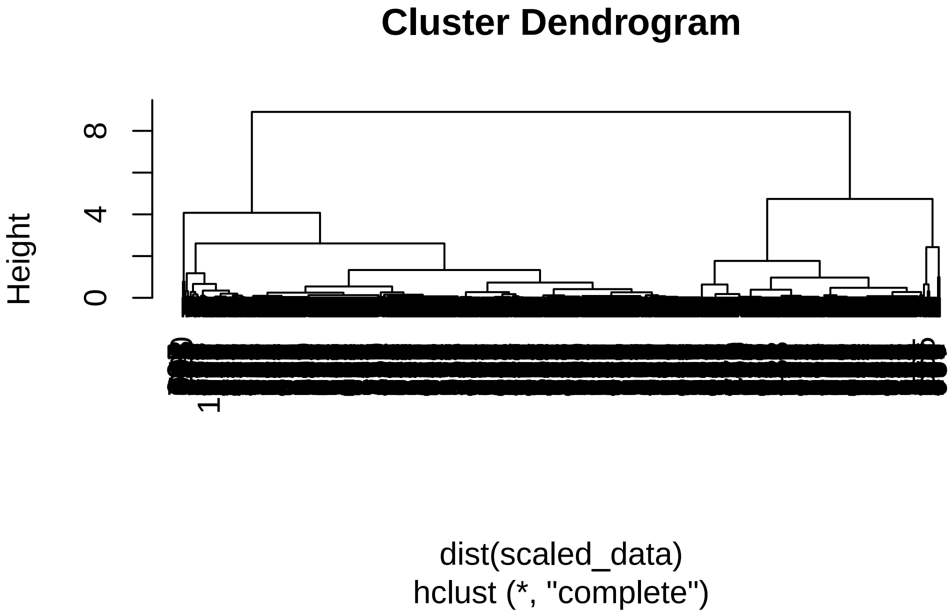
```r
# Perform hierarchical clustering with complete linkage
hclust_result <- hclust(dist(scaled_data), method = "complete")

# View cluster assignments
cluster_assignments <- cutree(hclust_result, k = 3)  # Adjust k as needed
print(cluster_assignments)
```

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

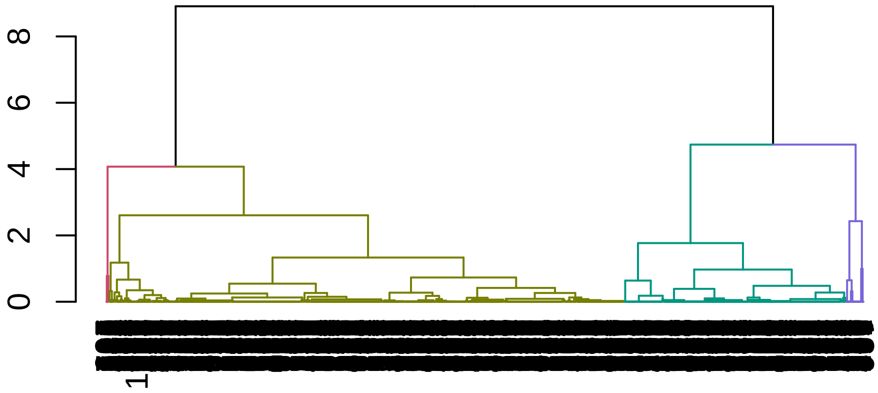
```r
library(ggplot2)

# Add cluster assignments to the sampled data
sampled_data$cluster <- factor(cluster_assignments)

# Plot the data points with cluster assignments
ggplot(sampled_data, aes(x = city_mpg_ft1, y = tailpipe_co2_in_grams_mile_ft1, color = clu
  geom_point() +
  labs(title = "Clustering Visualization", x = "City MPG", y = "Tailpipe CO2 (g/mile)")
```



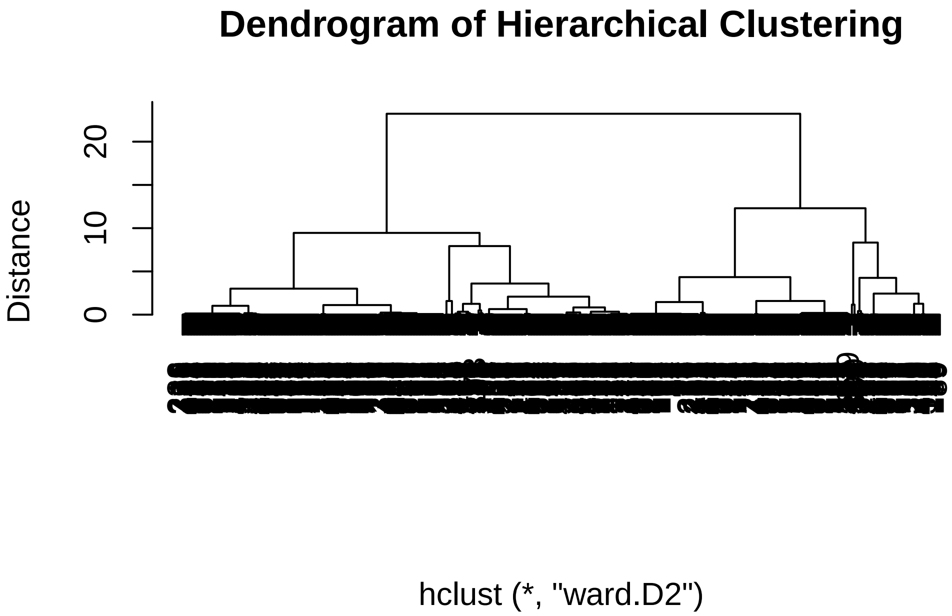
```
# Plot the dendrogram
plot(hclust_result)
```



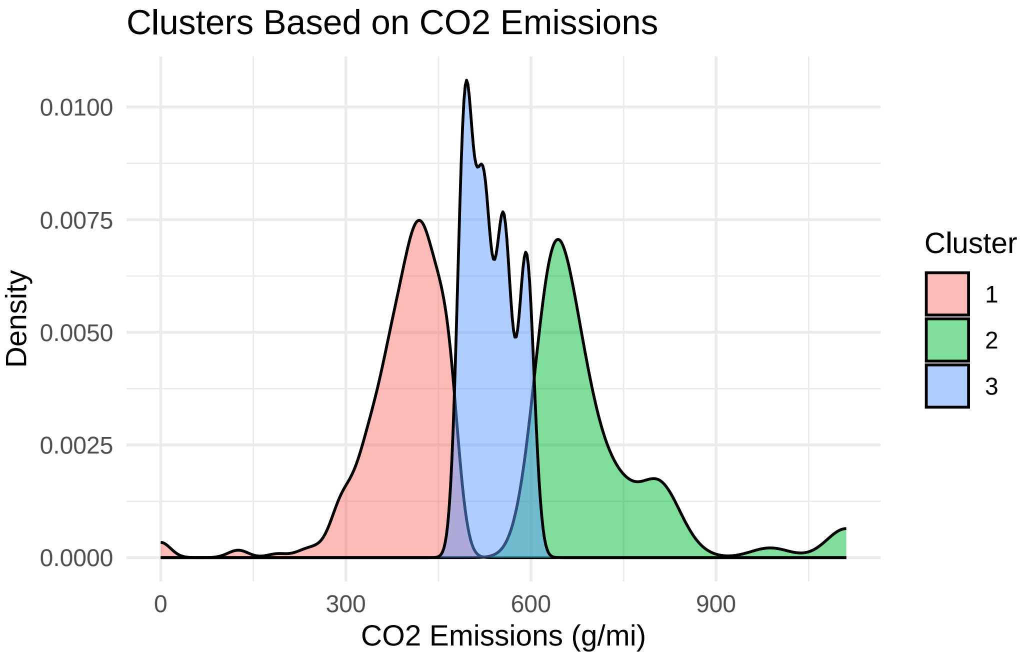
```
suppressPackageStartupMessages(library(dendextend))
avg_dend_obj <- as.dendrogram(hclust_result)
avg_col_dend <- color_branches(avg_dend_obj, h = 3)
plot(avg_col_dend)
```



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


# 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 = "", yla
```



```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 = facto
  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()
```



```r
library(dplyr)
```