DANA 4840 - Research Question 2

Aryan Mukherjee, Maryam Gadimova, Patricia Tating, Roman Shrestha

How can we identify and interpret similarities in features within clustering results to enhance the understanding of cluster formation and structure?

Loading the Libraries

```
library("tidyverse")
library("factoextra")
library("dendextend")
library("cluster")
library("gridExtra")
```

Reading the Data

```
mtcars <- read.csv("data/mtcars.csv", header = T, sep = ",")
head(mtcars)
```

```
##               model  mpg cyl disp  hp drat    wt  qsec vs am gear carb
## 1         Mazda RX4 21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## 2     Mazda RX4 Wag 21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## 3        Datsun 710 22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## 4    Hornet 4 Drive 21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## 5 Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## 6           Valiant 18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Checking Data Structure

```
dim(mtcars)
```

```
## [1] 32 12
```

```
str(mtcars)
```

```
## 'data.frame':    32 obs. of  12 variables:
##  $ model: chr  "Mazda RX4" "Mazda RX4 Wag" "Datsun 710" "Hornet 4 Drive" ...
##  $ mpg  : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl  : int  6 6 4 6 8 6 8 4 4 6 ...
```

```
##  $ disp : num  160 160 108 258 360 ...
##  $ hp   : int  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat : num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt   : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec : num  16.5 17 18.6 19.4 17 ...
##  $ vs   : int  0 0 1 1 0 1 0 1 1 1 ...
##  $ am   : int  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear : int  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb : int  4 4 1 1 2 1 4 2 2 4 ...
```

We can see that our data comprises 32 observations of different car models and 12 automobile features of mixed (numeric, integer, character) data types.

Data Pre-processing

```
mtcars_categorical <- data.frame(
  cyl = mtcars$cyl,
  vs = mtcars$vs,
  am = mtcars$am,
  gear = mtcars$gear,
  carb = mtcars$carb
)

mtcars_numerical <- data.frame(
  mpg = mtcars$mpg,
  disp = mtcars$disp,
  hp = mtcars$hp,
  drat = mtcars$drat,
  wt = mtcars$wt,
  qsec = mtcars$qsec
)

mtcars_numerical_scaled <- data.frame(scale(mtcars_numerical))

mtcars_joined <- cbind(mtcars_numerical_scaled, mtcars_categorical)
rownames(mtcars_joined) <- mtcars$model
mtcars <- mtcars_joined
head(mtcars)
```

```
##                           mpg        disp         hp       drat           wt
## Mazda RX4           0.1508848 -0.57061982 -0.5350928  0.5675137 -0.610399567
## Mazda RX4 Wag       0.1508848 -0.57061982 -0.5350928  0.5675137 -0.349785269
## Datsun 710          0.4495434 -0.99018209 -0.7830405  0.4739996 -0.917004624
## Hornet 4 Drive      0.2172534  0.22009369 -0.5350928 -0.9661175 -0.002299538
## Hornet Sportabout -0.2307345  1.04308123  0.4129422 -0.8351978  0.227654255
## Valiant           -0.3302874 -0.04616698 -0.6080186 -1.5646078  0.248094592
##                         qsec cyl vs am gear carb
## Mazda RX4         -0.7771651   6  0  1    4    4
## Mazda RX4 Wag     -0.4637808   6  0  1    4    4
## Datsun 710         0.4260068   4  1  1    4    1
## Hornet 4 Drive     0.8904872   6  1  0    3    1
## Hornet Sportabout -0.4637808   8  0  0    3    2
```

```
## Valiant            1.3269868   6  1  0    3    1
```

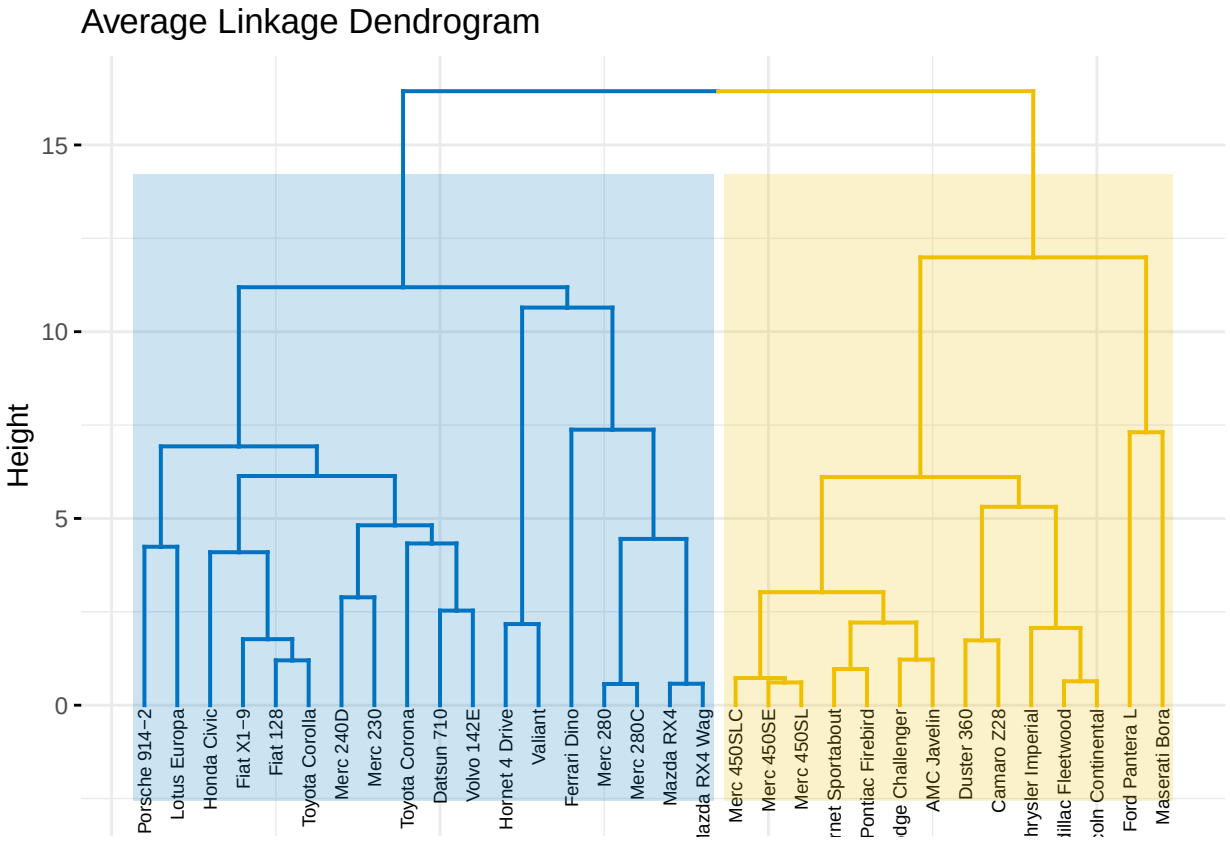
Our numerical features have different scale of measurements, so we standardized the data to ensure each variable contributes equally to the distance calculations, preventing variables with larger scales to have more weight in the clustering results. We also do not standardize the categorical data.

Average Linkage Method

```
res.dist <- dist(mtcars, method = "manhattan")

hc_average <- hclust(d = res.dist, method = "average")
grp_average <- cutree(hc_average, k = 2)

average_cluster <- fviz_cluster(
  list(data = mtcars, cluster = grp_average),
  palette = "jco",
  geom = "point",
  ellipse.type = "convex",
  show.clust.cent = FALSE,
  main = "Average Linkage Cluster",
  ggtheme = theme_minimal()
)

average_dendrogram <- fviz_dend(
  hc_average,
  cex = 0.5,
  k = 2,
  k_colors = "jco",
  rect = TRUE,
  rect_border = "jco",
  rect_fill = TRUE,
  label_cols = "black",
  label_cex = 0.5,
  main = "Average Linkage Dendrogram",
  ggtheme = theme_minimal()
)
```

```
## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none" instead as
## of ggplot2 3.3.4.
## i The deprecated feature was likely used in the factoextra package.
##   Please report the issue at <https://github.com/kassambara/factoextra/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
average_dendrogram
```



Clustering Results

```
grp <- cutree(hc_average, k = 2)
head(grp, n = 32)
```

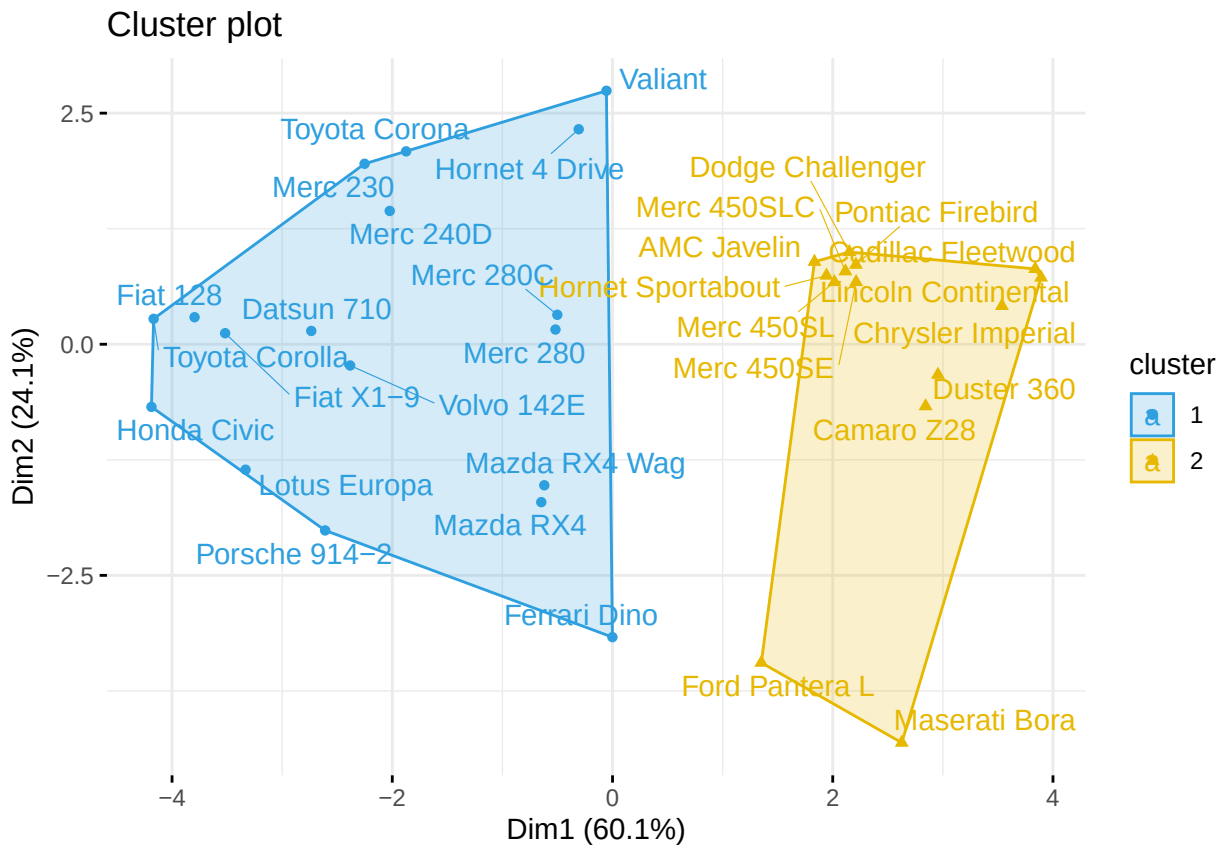
```
##           Mazda RX4       Mazda RX4 Wag          Datsun 710      Hornet 4 Drive
##                   1                   1                   1                   1
##   Hornet Sportabout             Valiant          Duster 360           Merc 240D
##                   2                   1                   2                   1
##            Merc 230            Merc 280           Merc 280C          Merc 450SE
##                   1                   1                   1                   2
##          Merc 450SL         Merc 450SLC  Cadillac Fleetwood Lincoln Continental
##                   2                   2                   2                   2
##   Chrysler Imperial            Fiat 128          Honda Civic       Toyota Corolla
##                   2                   1                   1                   1
##       Toyota Corona     Dodge Challenger          AMC Javelin          Camaro Z28
##                   1                   2                   2                   2
##    Pontiac Firebird           Fiat X1-9       Porsche 914-2         Lotus Europa
##                   2                   1                   1                   1
##      Ford Pantera L         Ferrari Dino       Maserati Bora           Volvo 142E
##                   2                   1                   2                   1
```

After applying the average linkage method to cluster the dataset, we obtained the following cluster assignments for each car model. Below is the number of members for each clusters:

```r
table(grp)
```

```
## grp
##  1  2
## 18 14
```

```r
fviz_cluster(list(data = mtcars, cluster = grp),
             palette = c("#2E9FDF", "#E7B800"),
             ellipse.type = "convex", #Concentration ellipse
             repel = TRUE, #Avoid label overplotting (slow)
             show.clust.cent = FALSE, ggtheme = theme_minimal()
             )
```



The cluster plot clearly shows how the car models are grouped together. However, it lacks insight into the specific feature similarities within each cluster that led to these groupings. To make our clustering results valuable, we need to interpret these clusters to make actionable decisions, particularly in marketing. By understanding the key features that define each cluster, we can tailor our marketing strategies to target specific customer segments more effectively. For example, if one cluster predominantly includes fuel-efficient cars, we can market these models to environmentally conscious consumers. Similarly, if another cluster consists of high-performance cars, we can focus our marketing efforts on car enthusiasts and performance-focused buyers.

There are several methods to identify and interpret clusters:

1. Descriptive Statistics

- Objective: Summarize and describe the features of each cluster.

- Implementation: Mean and Median, Standard Deviation and Range
- Use Case: In customer segmentation, calculate the average age, income, and spending score for each cluster.

2. Visualization

- Objective: Provide a graphical representation of feature similarities and differences within clusters.
- Implementation: Box Plots, Heatmaps, Pairwise Scatter Plots
- Use Case: In customer segmentation, visualize the average age, income, and spending score for each cluster.

3. Feature Importance Analysis

- Objective: Determine the most significant features contributing to cluster formation.
- Implementation: Decision Trees, Feature Importance Scores, ANOVA
- Use Case: In a marketing dataset, use a decision tree to identify the key demographic features driving customer segmentation.

4. Cluster Profiles

- Objective: Create detailed profiles for each cluster based on feature similarities.
- Implementation: Centroids, Cluster Summaries
- Use Case: In healthcare data, create profiles for patient clusters based on average age, medical history, and treatment outcomes.

5. Correlation Analysis

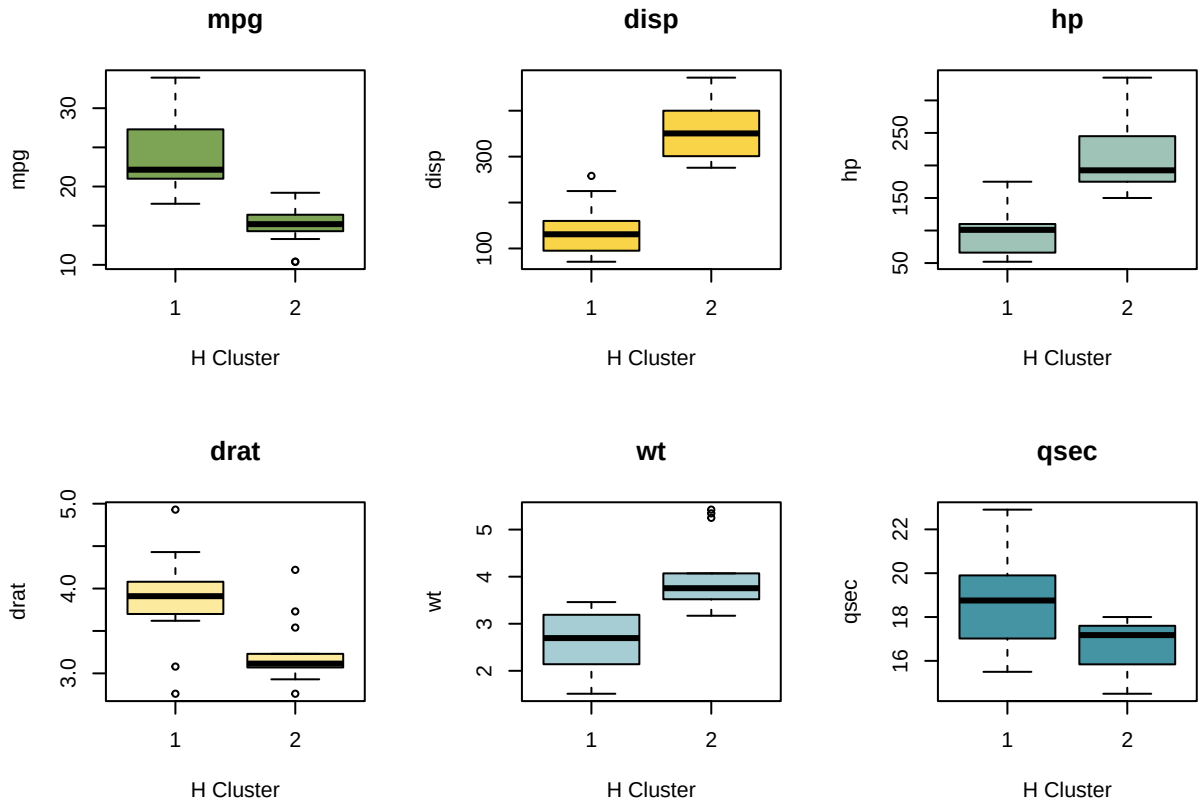
- Objective: Assess the relationships between features within clusters.
- Implementation: Correlation Coefficients, Correlation Matrices
- Use Case: In financial data, analyze the correlation between different financial indicators within investor clusters.

Visualization on Cluster Features

```
mtcars_num_results <- cbind(
  mtcars_numerical,
  hclust_average = grp
)
```

```
par(mfrow = c(2, 3))

color_palette <- c("#4494a4", "#7ca454", "#f9d448", "#9fc4b7", "#fcea9e", "#a6ccd4")

mtcars_results_col <- colnames(mtcars_num_results)[-length(colnames(mtcars_num_results))]

for (i in seq_along(mtcars_results_col)) {
  column_name <- mtcars_results_col[i]

  boxplot(
```

```
    mtcars_num_results[[column_name]] ~ mtcars_num_results$hclust_average,
    xlab = "H Cluster",
    ylab = column_name,
    main = paste(column_name),
    col = color_palette[i %% length(color_palette) + 1]
  )
}
```

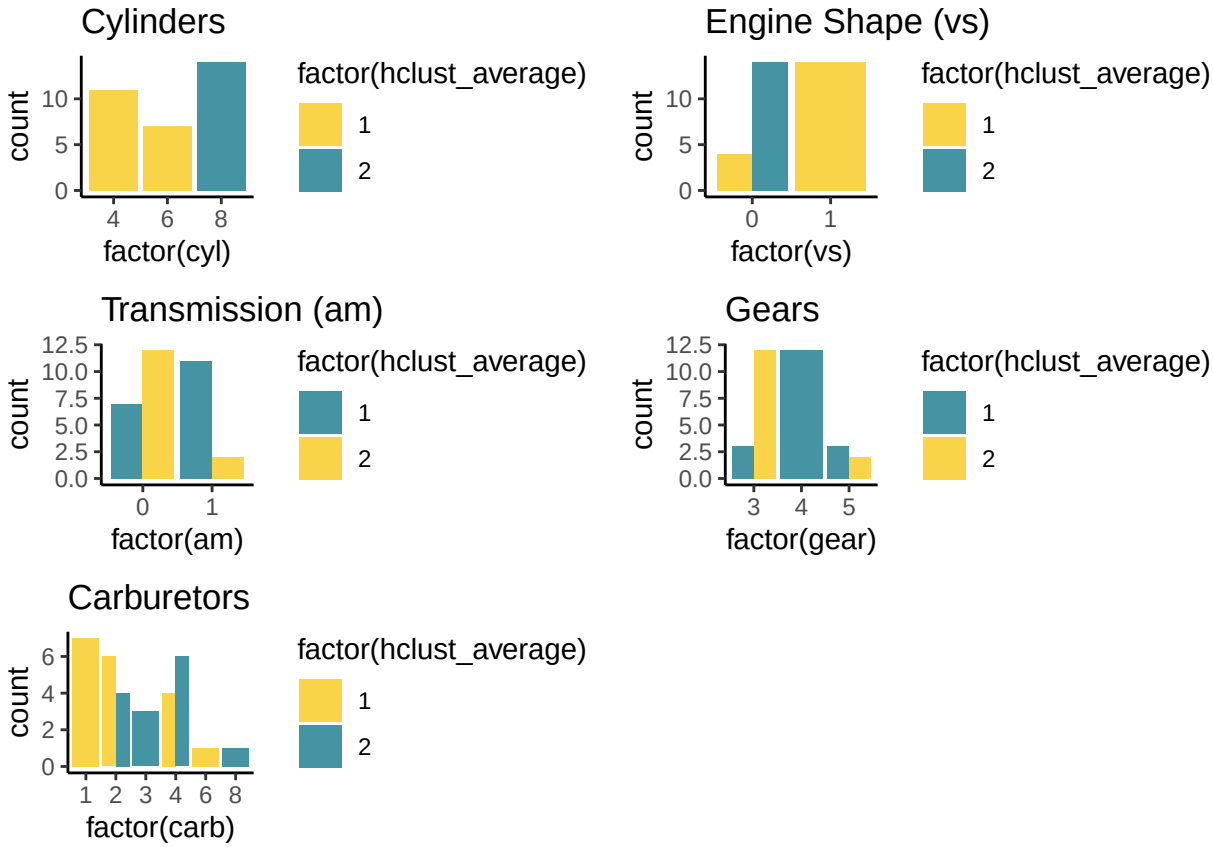



```
par(mfrow = c(1, 1))
```

```
mtcars_cat_results <- cbind(
  mtcars_categorical,
  hclust_average = grp
)
```

```
mtcars_cat_results$hclust_average <- as.factor(mtcars_cat_results$hclust_average)

cols_palette <- c("#4494a4", "#f9d448")
variables_to_plot <- colnames(mtcars_cat_results)[colnames(mtcars_cat_results) != "hclust_average"]

plot_cyl <- ggplot(mtcars_cat_results, aes(x = factor(cyl), fill = factor(hclust_average))) +
  geom_bar(position = "dodge") +
  scale_fill_manual(values = sample(cols_palette)) +
  theme_classic() +
```

```
  ggtitle("Cylinders")

plot_vs <- ggplot(mtcars_cat_results, aes(x = factor(vs), fill = factor(hclust_average))) +
  geom_bar(position = "dodge") +
  scale_fill_manual(values = sample(cols_palette)) +
  theme_classic() +
  ggtitle("Engine Shape (vs)")

plot_am <- ggplot(mtcars_cat_results, aes(x = factor(am), fill = factor(hclust_average))) +
  geom_bar(position = "dodge") +
  scale_fill_manual(values = sample(cols_palette)) +
  theme_classic() +
  ggtitle("Transmission (am)")

plot_gear <- ggplot(mtcars_cat_results, aes(x = factor(gear), fill = factor(hclust_average))) +
  geom_bar(position = "dodge") +
  scale_fill_manual(values = sample(cols_palette)) +
  theme_classic() +
  ggtitle("Gears")

plot_carb <- ggplot(mtcars_cat_results, aes(x = factor(carb), fill = factor(hclust_average))) +
  geom_bar(position = "dodge") +
  scale_fill_manual(values = sample(cols_palette)) +
  theme_classic() +
  ggtitle("Carburetors")

# Arrange the plots in a grid
grid.arrange(plot_cyl, plot_vs, plot_am, plot_gear, plot_carb, ncol = 2)
```



We can use box plots and bar plots to visualize the distinct characteristics of each cluster based on the features. This visualization clearly highlights the differences between the clusters. However, we'll have to validate this with statistics as sometimes visualization can be subjective.

Descriptive Statistics on Clusters

```
summary(mtcars_num_results[mtcars_num_results$hclust == 1, 1:5])
```

```
##       mpg             disp             hp             drat      
##  Min.   :17.80   Min.   : 71.10   Min.   : 52.00   Min.   :2.760  
##  1st Qu.:21.00   1st Qu.: 98.33   1st Qu.: 72.25   1st Qu.:3.717  
##  Median :22.15   Median :130.90   Median :101.00   Median :3.910  
##  Mean   :23.97   Mean   :135.54   Mean   : 98.06   Mean   :3.882  
##  3rd Qu.:26.98   3rd Qu.:160.00   3rd Qu.:110.00   3rd Qu.:4.080  
##  Max.   :33.90   Max.   :258.00   Max.   :175.00   Max.   :4.930  
##        wt        
##  Min.   :1.513  
##  1st Qu.:2.155  
##  Median :2.695  
##  Mean   :2.609  
##  3rd Qu.:3.180  
##  Max.   :3.460
```

```r
summary(mtcars_num_results[mtcars_num_results$hclust == 2, 1:5])
```

```
##       mpg             disp             hp             drat
##  Min.   :10.40   Min.   :275.8   Min.   :150.0   Min.   :2.760
##  1st Qu.:14.40   1st Qu.:301.8   1st Qu.:176.2   1st Qu.:3.070
##  Median :15.20   Median :350.5   Median :192.5   Median :3.115
##  Mean   :15.10   Mean   :353.1   Mean   :209.2   Mean   :3.229
##  3rd Qu.:16.25   3rd Qu.:390.0   3rd Qu.:241.2   3rd Qu.:3.225
##  Max.   :19.20   Max.   :472.0   Max.   :335.0   Max.   :4.220
##        wt
##  Min.   :3.170
##  1st Qu.:3.533
##  Median :3.755
##  Mean   :3.999
##  3rd Qu.:4.014
##  Max.   :5.424
```

- **Cluster 1:** Contains cars that are generally more fuel-efficient, lighter, and have smaller engine displacements and lower horsepower. This cluster likely represents smaller, more economical vehicles.
- **Cluster 2:** Contains cars with lower fuel efficiency, heavier weights, larger engine displacements, and higher horsepower. This cluster likely represents larger, more powerful vehicles.

```r
summary(mtcars_cat_results[mtcars_num_results$hclust == 1, 1:5])
```

```
##       cyl             vs               am              gear        carb
##  Min.   :4.000   Min.   :0.0000   Min.   :0.0000   Min.   :3   Min.   :1.000
##  1st Qu.:4.000   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:4   1st Qu.:1.000
##  Median :4.000   Median :1.0000   Median :1.0000   Median :4   Median :2.000
##  Mean   :4.778   Mean   :0.7778   Mean   :0.6111   Mean   :4   Mean   :2.278
##  3rd Qu.:6.000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:4   3rd Qu.:3.500
##  Max.   :6.000   Max.   :1.0000   Max.   :1.0000   Max.   :5   Max.   :6.000
```

```r
summary(mtcars_cat_results[mtcars_num_results$hclust == 2, 1:5])
```

```
##       cyl          vs          am              gear            carb
##  Min.   :8   Min.   :0   Min.   :0.0000   Min.   :3.000   Min.   :2.00
##  1st Qu.:8   1st Qu.:0   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.25
##  Median :8   Median :0   Median :0.0000   Median :3.000   Median :3.50
##  Mean   :8   Mean   :0   Mean   :0.1429   Mean   :3.286   Mean   :3.50
##  3rd Qu.:8   3rd Qu.:0   3rd Qu.:0.0000   3rd Qu.:3.000   3rd Qu.:4.00
##  Max.   :8   Max.   :0   Max.   :1.0000   Max.   :5.000   Max.   :8.00
```

- **Cluster 1:** Contains cars with fewer cylinders (mostly 4, some 6), a mix of engine shapes (more straight engines), a higher proportion of manual transmissions, typically 4 gears, and fewer carburetors. These cars are generally lighter and more fuel-efficient, indicating they might be more economical or everyday vehicles.
- **Cluster 2:** Consists exclusively of cars with 8 cylinders, V-shaped engines, mostly automatic transmissions, typically 3 gears, and more carburetors. These cars are heavier and less fuel-efficient, indicating they might be more powerful or performance-oriented vehicles.

Cluster Interpretation

Cluster 1:

Cars in this cluster typically have:

- Lower horsepower (hp)
- Higher fuel efficiency (mpg)
- Lower weight (wt)
- Smaller engine displacement (disp)
- Fewer cylinders (cyl)

These characteristics suggest that cars in Cluster 1 are generally more economical and smaller, possibly including compact and subcompact cars.

Cluster 2:

Cars in this cluster typically have:

- Higher horsepower (hp)
- Lower fuel efficiency (mpg)
- Higher weight (wt)
- Larger engine displacement (disp)
- More cylinders (cyl)

These characteristics suggest that cars in Cluster 2 are generally more powerful and larger, possibly including sports cars, muscle cars, and luxury cars.