Honda Clustering

Raymond David, Hyoju Kang, Angel Hsu

2023-11-21

Clustering

```
features <- read_csv("/Users/hyoju/Desktop/Honda Competition/Datasets/feature.csv")
```

```
## Rows: 74 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (11): Make, Model, Subtitle, Acceleration, TopSpeed, Range, Efficiency, ...
## dbl  (1): NumberofSeats
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
merged <- read_csv("/Users/hyoju/Desktop/Honda Competition/Datasets/merged.csv")
```

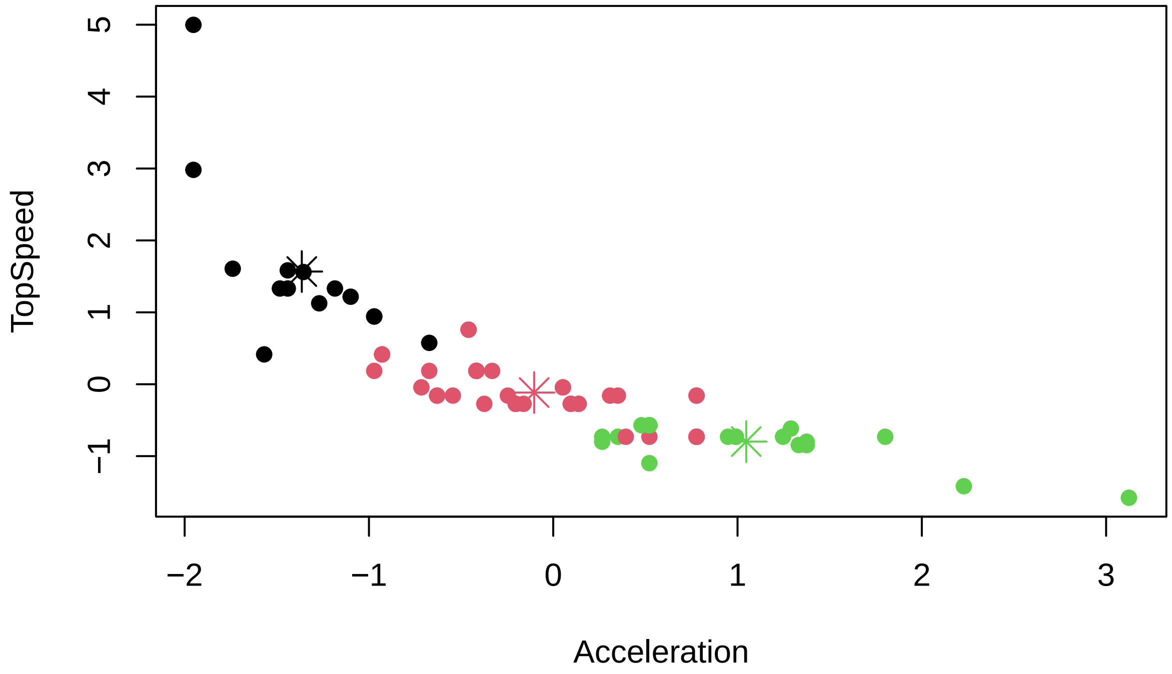
```
## Rows: 12 Columns: 14
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): Make, Model, Body Style, Drive, PriceinGermany
## dbl (8): Sales Count, Acceleration (sec), TopSpeed (km/h), Range (km), Effic...
## num (1): PriceUS ($)
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
library(stats)

# Function to extract numeric values
extract_numeric <- function(x) {
  as.numeric(gsub("[^0-9.]", "", x))
}

# Select columns for clustering
selected_columns <- data.frame(
  Acceleration = extract_numeric(features$Acceleration),
  TopSpeed = extract_numeric(features$TopSpeed),
  Range = extract_numeric(features$Range),
  Efficiency = extract_numeric(gsub("[^0-9.]", "", features$Efficiency)),
```

```r
  FastChargeSpeed = extract_numeric(features$FastChargeSpeed)
)

# Normalize the data
normalized_data <- scale(selected_columns)

# Determine the number of clusters (k value)
# For demonstration purposes, let's assume k = 3
k <- 3

# Perform k-means clustering
kmeans_result <- kmeans(normalized_data, centers = k)

# View the cluster assignments
cluster_assignments <- kmeans_result$cluster
#print(cluster_assignments)

# View the centroids of each cluster
centroids <- kmeans_result$centers
print(centroids)
```

```
##   Acceleration   TopSpeed       Range Efficiency FastChargeSpeed
## 1   -1.3643853  1.5672824  1.23831303 -0.1610230       1.1170278
## 2   -0.1035798 -0.1157187 -0.02034549  0.4628905       0.2012213
## 3    1.0471557 -0.7974838 -0.75287518 -0.6970689      -1.0583999
```

```r
# Assuming kmeans_result is your kmeans clustering result
plot(normalized_data, col = kmeans_result$cluster, pch = 19)
points(kmeans_result$centers, col = 1:k, pch = 8, cex = 2)
```



```r
# K-means + One Hot Encoding
# Extract numerical values from columns
extract_numeric <- function(x) {
  as.numeric(gsub("[^0-9.]", "", x))
}

# Select numerical columns for clustering
selected_columns <- data.frame(
  Acceleration = extract_numeric(features$Acceleration),
  TopSpeed = extract_numeric(features$TopSpeed),
  Range = extract_numeric(features$Range),
  Efficiency = extract_numeric(gsub("[^0-9.]", "", features$Efficiency)),
  FastChargeSpeed = extract_numeric(features$FastChargeSpeed)
)

# One-hot encode categorical column 'Drive'
drive_column <- model.matrix(~ Drive - 1, data = features)
selected_columns <- cbind(selected_columns, drive_column)

# Normalize the data
normalized_data <- scale(selected_columns)

# Determine the number of clusters (k value)
# For demonstration purposes, let's assume k = 3
k <- 3
```

```r
# Perform k-means clustering
kmeans_result <- kmeans(normalized_data, centers = k)

# View the cluster assignments
cluster_assignments <- kmeans_result$cluster
#print(cluster_assignments)

# View the centroids of each cluster
centroids <- kmeans_result$centers
print(centroids)
```

```
##   Acceleration    TopSpeed       Range Efficiency FastChargeSpeed
## 1   -0.3954477 -0.01470931 -0.08249112  0.9241744       0.1401197
## 2   -1.3643853  1.56728242  1.23831303 -0.1610230       1.1170278
## 3    0.7942261 -0.59969251 -0.42657210 -0.5534962      -0.5278128
##   DriveAll Wheel Drive DriveFront Wheel Drive DriveRear Wheel Drive
## 1            0.9104519             -0.5216648            -0.5334027
## 2            0.9932203             -0.5216648            -0.6251976
## 3           -0.9932203              0.5506462             0.5987342
```

Cluster using Merged Dataset

```r
# Extract numerical values from columns
extract_numeric <- function(x) {
  as.numeric(gsub("[^0-9.]", "", x))
}

# Select columns for clustering
selected_columns <- data.frame(
  SalesCount = merged$`Sales Count`,
  Acceleration = extract_numeric(merged$`Acceleration (sec)`),
  TopSpeed = extract_numeric(merged$`TopSpeed (km/h)`),
  Range = extract_numeric(merged$`Range (km)`)
)

# Normalize the data
normalized_data <- scale(selected_columns)

# Determine the number of clusters (k value)
# For demonstration purposes, let's assume k = 3
k <- 3

# Perform k-means clustering
# Convert data.frame to matrix as kmeans() requires a matrix input
kmeans_result <- kmeans(as.matrix(normalized_data), centers = k)

# View the cluster assignments
cluster_assignments <- kmeans_result$cluster
print(cluster_assignments)
```

```
##  [1] 3 1 1 1 1 2 3 3 3 3 3 3
```

```
# View the centroids of each cluster
centroids <- kmeans_result$centers
print(centroids)
```

```
##   SalesCount Acceleration  TopSpeed      Range
## 1  0.9288519   -1.1604408  1.190054  0.8686131
## 2  0.5847253    0.6802584 -1.444380 -1.8380046
## 3 -0.6143047    0.5659293 -0.473691 -0.2337783
```

```
# K-means + One-hot Encoding
# Extract numerical values from columns
extract_numeric <- function(x) {
  as.numeric(gsub("[^0-9.]", "", x))
}

# Manually encode 'BodyStyle' column
encoded_body_style <- model.matrix(~ `Body Style` - 1, data = merged)

# Select numerical columns for clustering
selected_columns <- cbind(
  encoded_body_style,
  SalesCount = merged$`Sales Count`,
  Acceleration = extract_numeric(merged$`Acceleration (sec)`),
  TopSpeed = extract_numeric(merged$`TopSpeed (km/h)`)
)

# Normalize the data
normalized_data <- scale(selected_columns)

# Determine the number of clusters (k value)
# For demonstration purposes, let's assume k = 3
k <- 3

# Perform k-means clustering
kmeans_result <- kmeans(normalized_data, centers = k)

# View the cluster assignments
cluster_assignments <- kmeans_result$cluster
# print(cluster_assignments)

# View the centroids of each cluster
centroids <- kmeans_result$centers
print(centroids)
```

```
##   ‘Body Style‘Hatchbag ‘Body Style‘Sedan ‘Body Style‘SUV   SalesCount
## 1            2.1408721        -0.4281744      -1.3540064 -0.008455543
## 2           -0.4281744        -0.4281744       0.6770032 -0.616416078
## 3           -0.4281744         0.8563488      -0.3385016  0.928851888
##   Acceleration   TopSpeed
## 1    0.6535816 -1.1132766
## 2    0.5557667 -0.4222773
## 3   -1.1604408  1.1900543
```

US Vehicle Feature Dataset

Cluster1:

Vehicles exhibit well-balanced performance attributes, slightly leaning towards enhanced efficiency and fast charging capabilities. Although All-Wheel Drive stands prevalent, it does not assert exclusive dominance within this cluster.

Cluster2:

In contrast, the second cluster encompasses vehicles with diverse performance traits, markedly emphasizing speed and range. The presence of various drive types underscores a broad spectrum of vehicles, each possessing distinct performance features

Cluster3:

Lastly, the third cluster suggests vehicles with more modest performance attributes, demonstrating a slight inclination towards acceleration while exhibiting lower efficiency, moderately reduced top speed, and range. The cluster portrays a mixed representation of vehicles, with a notable focus on front and rear-wheel drives

Washington EV Population Dataset

Cluster1:

The initial group predominantly comprises hatchback cars with sales figures that align with the average. These vehicles notably exhibit higher acceleration but relatively lower top speeds, highlighting a trade-off between these performance metrics

Cluster2:

The second cluster presents a blend of SUVs, fewer hatchbacks, and sedans. Vehicles in this cluster generally showcase higher acceleration but demonstrate lower top speeds, indicating a distinctive category of cars encompassing diverse body styles yet sharing analogous performance attributes

Cluster3:

The third cluster primarily consists of sedans boasting sales counts higher than the average. These vehicles display lower acceleration but boast higher top speeds, possibly targeting a market segment prioritizing velocity over rapid acceleration