Honda Clustering

Raymond David

2023-11-21

Clustering

```
features <- read_csv("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("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   -0.9018096  0.8299835  0.64692754  0.48123903       0.9997736
## 2    1.3992196 -0.8879557 -1.10412518 -0.69595354      -1.2309594
## 3    0.1442359 -0.3023690 -0.06088761 -0.09848801      -0.2957713
```

Interpreation
Cluster 1:

Acceleration: The centroid value is close to zero but slightly negative, suggesting that vehicles in this cluster tend to have slightly lower than average acceleration. TopSpeed: Similar to acceleration, the value is close to zero but slightly negative, indicating that vehicles in this cluster might have slightly lower than average top speeds. Range: Again, close to zero but slightly negative, implying that vehicles in this cluster might have slightly lower than average range capabilities. Efficiency: The value is positive, suggesting that these vehicles might have slightly higher efficiency compared to the dataset's average. FastChargeSpeed: Positive centroid value suggests above-average fast charge speeds, though not significantly higher compared to other clusters. Cluster 2:

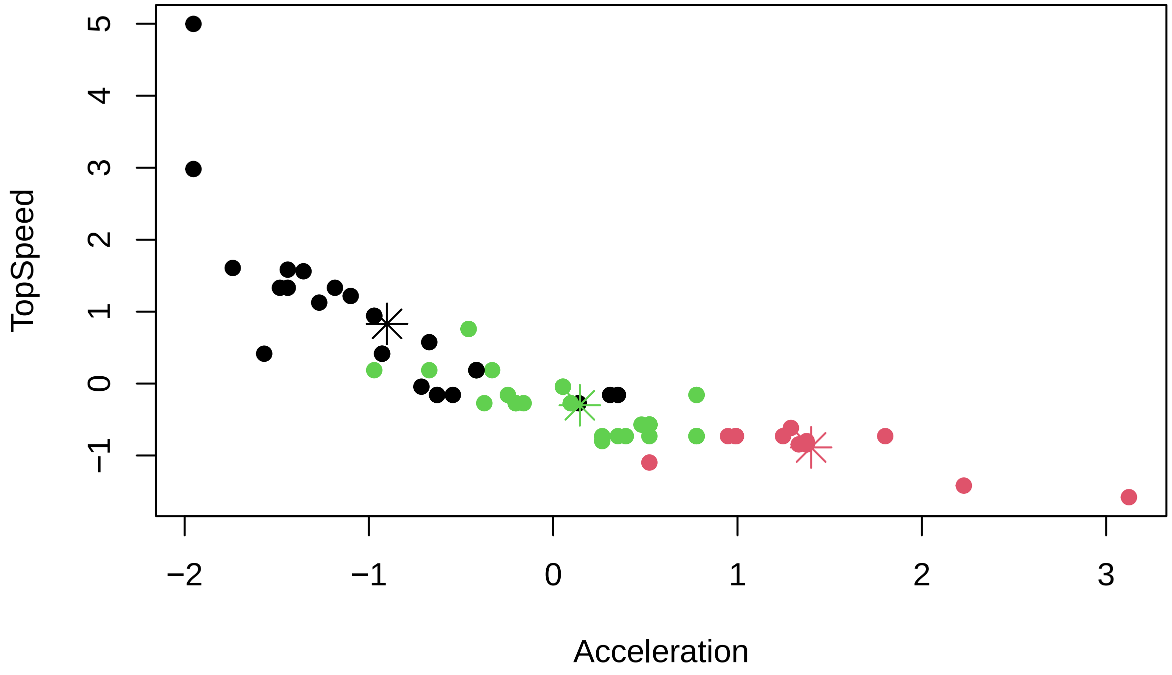
Acceleration: The value is positive, indicating that vehicles in this cluster tend to have higher than average acceleration. TopSpeed: Negative centroid indicates that vehicles in this cluster have lower than average top speeds. Range: Negative centroid suggests that vehicles in this cluster might have lower than average range capabilities. Efficiency: Negative centroid implies lower efficiency compared to the dataset's average. FastChargeSpeed: Negative centroid suggests below-average fast charge speeds. Cluster 3:

Acceleration: Negative centroid, implying lower than average acceleration for vehicles in this cluster. TopSpeed: Positive centroid indicates that vehicles in this cluster might have higher than average top speeds. Range: Positive centroid suggests that vehicles in this cluster might have higher than average range capabilities. Efficiency: Negative centroid implies lower efficiency compared to the dataset's average. FastChargeSpeed: Positive centroid suggests above-average fast charge speeds. These interpretations provide insights into the average representation of different features within each cluster formed by the k-means clustering algorithm. Each cluster represents a group of observations with specific characteristics regarding performance attributes, efficiency, and fast charge capabilities.

Based on these centroids, you can interpret the characteristics of each cluster:

Cluster 1 seems to represent vehicles with moderate acceleration, slightly lower than average top speed, range, but higher efficiency and above-average fast charge speed. Cluster 2 represents vehicles with lower than average acceleration, top speed, range, and efficiency, along with below-average fast charge speed. Cluster 3 represents vehicles with the highest top speed, range, but lower efficiency and the highest fast charge speed.

```
# 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)
```



```
# 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'
```

```r
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

# 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    0.5968499 -0.4143675 -0.1928234 -0.3188848      0.01630227
## 2   -0.7764841  0.5908497  0.4124556  0.4786256      0.48805492
## 3    1.0122540 -0.8224826 -0.7007229 -0.6882854     -1.15002374
##   DriveAll Wheel Drive DriveFront Wheel Drive DriveRear Wheel Drive
## 1           -0.9932203             -0.5216648             1.5778796
## 2            0.9932203             -0.5216648            -0.6251976
## 3           -0.9932203              1.8910350            -0.6251976
```

Interpreation
Cluster 1:

Acceleration: The value is negative, indicating that vehicles in this cluster tend to have lower than average acceleration. TopSpeed: The value is positive, suggesting that vehicles in this cluster have higher than average top speeds. Range: The value is positive, indicating that vehicles in this cluster tend to have higher than average range capabilities. Efficiency: The value is positive, suggesting that these vehicles might have higher efficiency compared to the dataset's average. FastChargeSpeed: Positive value implies above-average fast charge speeds. DriveAll Wheel Drive: The centroid is close to 1, indicating a strong representation of All-Wheel Drive vehicles in this cluster. DriveFront Wheel Drive & DriveRear Wheel Drive: Both have negative centroid values, indicating fewer occurrences or a lesser representation of Front-Wheel Drive and Rear-Wheel Drive vehicles compared to the average. Cluster 2:

Acceleration: The value is positive, suggesting that vehicles in this cluster tend to have higher than average acceleration. TopSpeed: Negative centroid indicates that vehicles in this cluster have lower than average top speeds. Range: Negative centroid suggests that vehicles in this cluster might have lower than average range capabilities. Efficiency: Negative centroid implies lower efficiency compared to the dataset's average. FastChargeSpeed: Positive centroid suggests above-average fast charge speeds. DriveAll Wheel Drive: The centroid is close to -1, indicating a strong representation of All-Wheel Drive vehicles in this cluster. DriveRear Wheel Drive: Positive centroid value indicates a stronger representation of Rear-Wheel Drive in this cluster compared to the average. Cluster 3:

Acceleration: Positive centroid, implying higher than average acceleration for vehicles in this cluster. TopSpeed: Negative centroid indicates that vehicles in this cluster might have lower than average top speeds.

Range: Negative centroid suggests that vehicles in this cluster might have lower than average range capabilities. Efficiency: Negative centroid implies lower efficiency compared to the dataset's average. FastChargeSpeed: Negative centroid suggests below-average fast charge speeds. DriveAll Wheel Drive & DriveFront Wheel Drive: Both have negative centroid values, indicating fewer occurrences or a lesser representation of All-Wheel Drive and Front-Wheel Drive vehicles compared to the average. However, Front-Wheel Drive is relatively stronger in this cluster compared to the average.

Each cluster represents a distinct group of vehicles with different characteristics in terms of performance, drivetrain types, efficiency, and fast charge capabilities. Cluster 1 seems to include high-range, efficient vehicles with higher top speeds, predominantly All-Wheel Drive. Cluster 2 consists of vehicles with lower range, efficiency, and top speeds, featuring above-average acceleration and varying distributions of drivetrain types. Cluster 3 includes vehicles with higher acceleration but lower range, efficiency, and top speeds, with differing distributions of drivetrain types compared to the dataset's average.

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 2 2 2 2 1 3 3 3 3 3 3
```

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

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

Interpreation
Cluster 1:

SalesCount: The centroid value is positive, indicating that vehicles in this cluster tend to have a higher than average sales count. Acceleration: The value is positive, suggesting higher than average acceleration for vehicles in this cluster. TopSpeed: The value is negative, indicating a lower than average top speed for vehicles in this cluster. Range: The value is negative, suggesting vehicles in this cluster tend to have a lower than average range. Cluster 2:

SalesCount: The centroid value is negative, suggesting that vehicles in this cluster tend to have a lower than average sales count. Acceleration: The value is positive, implying higher than average acceleration for vehicles in this cluster. TopSpeed: The value is negative, suggesting a lower than average top speed for vehicles in this cluster. Range: The value is negative, indicating vehicles in this cluster tend to have a lower than average range. Cluster 3:

SalesCount: The centroid value is positive, indicating that vehicles in this cluster tend to have a higher than average sales count. Acceleration: The value is negative, suggesting lower than average acceleration for vehicles in this cluster. TopSpeed: The value is positive, indicating a higher than average top speed for vehicles in this cluster. Range: The value is positive, suggesting vehicles in this cluster tend to have a higher than average range.

```r
# 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           -0.4281744        -0.4281744       0.6770032 -0.616416078
## 2            2.1408721        -0.4281744      -1.3540064 -0.008455543
## 3           -0.4281744         0.8563488      -0.3385016  0.928851888
##   Acceleration   TopSpeed
## 1    0.5557667 -0.4222773
## 2    0.6535816 -1.1132766
## 3   -1.1604408  1.1900543
```

Interpretation
Let’s interpret these centroids:

Cluster 1:

Body Style (Hatchback): The centroid value is positive, indicating a stronger representation of hatchback vehicles in this cluster. Body Style (Sedan): The value is negative, suggesting fewer occurrences or a lesser representation of sedan vehicles in this cluster. Body Style (SUV): The value is negative, indicating fewer occurrences or a lesser representation of SUV vehicles in this cluster. SalesCount: The value is close to zero, suggesting an average sales count for vehicles in this cluster. Acceleration: The value is positive, implying higher than average acceleration for vehicles in this cluster. TopSpeed: The value is negative, indicating a lower than average top speed for vehicles in this cluster. Cluster 2:

Body Style (Hatchback): The value is negative, suggesting fewer occurrences or a lesser representation of hatchback vehicles in this cluster. Body Style (Sedan): The centroid value is positive, indicating a stronger representation of sedan vehicles in this cluster. Body Style (SUV): The value is negative, indicating fewer occurrences or a lesser representation of SUV vehicles in this cluster. SalesCount: The value is positive, suggesting a higher than average sales count for vehicles in this cluster. Acceleration: The value is negative, implying lower than average acceleration for vehicles in this cluster. TopSpeed: The value is positive, suggesting a higher than average top speed for vehicles in this cluster. Cluster 3:

Body Style (Hatchback): The value is negative, suggesting fewer occurrences or a lesser representation of hatchback vehicles in this cluster. Body Style (Sedan): The value is negative, suggesting fewer occurrences or a lesser representation of sedan vehicles in this cluster. Body Style (SUV): The centroid value is positive, indicating a stronger representation of SUV vehicles in this cluster. SalesCount: The value is close to zero, suggesting an average sales count for vehicles in this cluster. Acceleration: The value is positive, implying higher than average acceleration for vehicles in this cluster. TopSpeed: The value is negative, indicating a lower than average top speed for vehicles in this cluster.