Cluster analysis

Homework Week 9

The Solution

06 April 2019

```
pacman::p_load(tidyverse, GGally, vegan)
```

Data preparation, identification

```
 load("C:/Users/Devan.McGranahan/GoogleDrive/Teaching/Classes/Analysis of Ecosystems/compiled notes/AoE
str(mtcars2)
```

```
## 'data.frame':    32 obs. of  16 variables:
##  $ make.model: Factor w/ 32 levels "AMC Javelin",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ make      : Factor w/ 20 levels "AMC","Cadillac",..: 1 2 3 4 5 6 16 7 8 8 ...
##  $ origin    : Factor w/ 2 levels "domestic","foreign": 1 1 1 1 2 1 1 2 2 2 ...
##  $ country   : Factor w/ 6 levels "Germany","Italy",..: 6 6 6 6 3 6 6 2 2 2 ...
##  $ continent : Factor w/ 3 levels "Asia","Europe",..: 3 3 3 3 1 3 3 2 2 2 ...
##  $ am        : Factor w/ 2 levels "Automatic","Manual": 1 1 1 1 2 1 1 2 2 2 ...
##  $ vs        : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 1 2 2 ...
##  $ gear      : Factor w/ 3 levels "3","4","5": 1 1 1 1 2 1 1 3 2 2 ...
##  $ carb      : Factor w/ 6 levels "1","2","3","4",..: 2 4 4 4 1 2 4 5 1 1 ...
##  $ cyl       : Factor w/ 3 levels "4","6","8": 3 3 3 3 1 3 3 2 1 1 ...
##  $ disp      : num  304 472 350 440 108 318 360 145 78.7 79 ...
##  $ hp        : int  150 205 245 230 93 150 245 175 66 66 ...
##  $ drat      : num  3.15 2.93 3.73 3.23 3.85 2.76 3.21 3.62 4.08 4.08 ...
##  $ wt        : num  3.44 5.25 3.84 5.34 2.32 ...
##  $ qsec      : num  17.3 18 15.4 17.4 18.6 ...
##  $ mpg       : num  15.2 10.4 13.3 14.7 22.8 15.5 14.3 19.7 32.4 27.3 ...
```

Analysis

Univariate relationships

```
mtcars2 %>%
  select(., .data$disp :.data$mpg ) %>%
      ggpairs() + theme_bw()
```

Distance matrix

```
car.em <-
  mtcars2 %>%
    select(., .data$disp :.data$mpg ) %>%
      vegdist(., method = "euc")
```

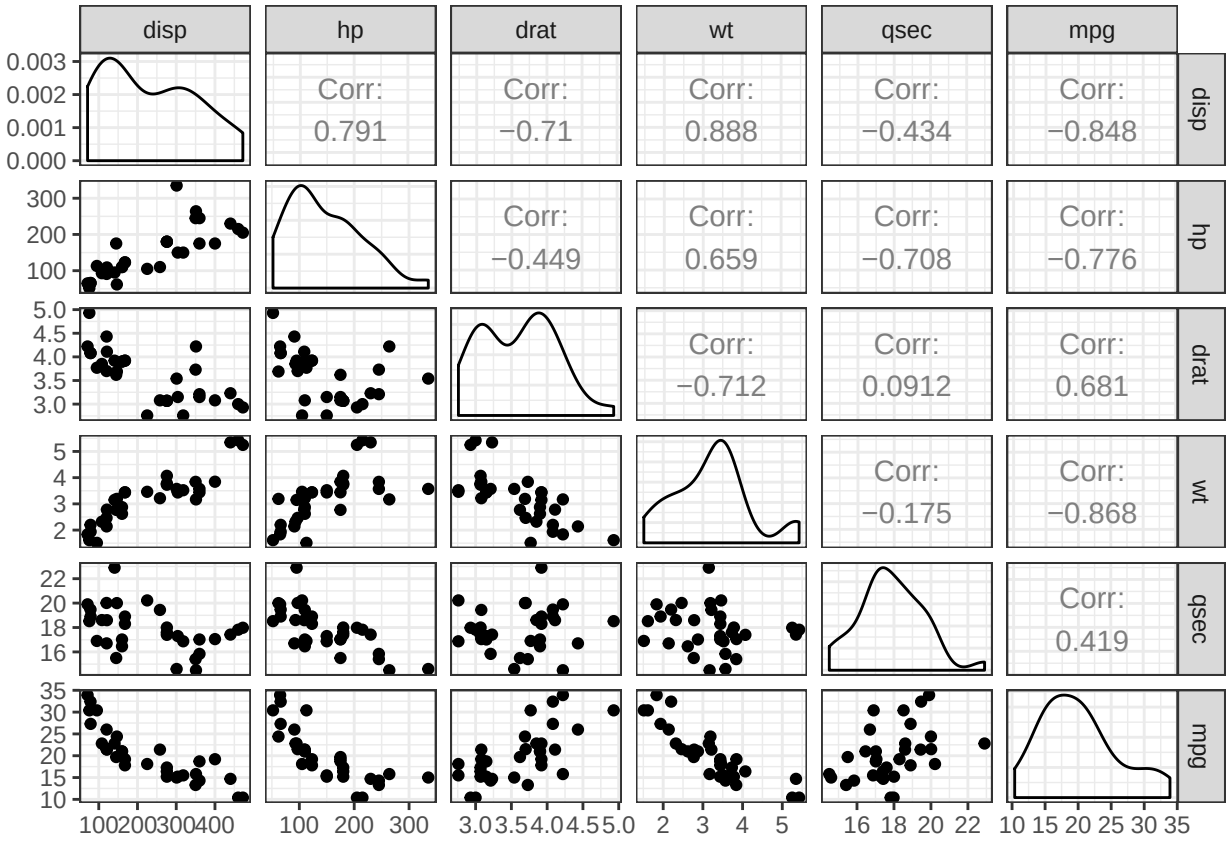


Figure 1: Scatterplot matrix of continuous variables in `mtcars2` dataset.

```
# Maximum distance
  max(car.em)
```

```
## [1] 425.3117
```

```
# Some R matrix work will show us which pair is the maximum:
  car.m <- as.matrix(car.em)
  colnames(car.m) <- mtcars2$make.model
  rownames(car.m) <- mtcars2$make.model
  car.m[car.m == 0] <- 1 # Replace zeros in diagonal
# Find maximum pair
  which(car.m == max(car.m), arr.ind = TRUE) %>%
    row.names()
```

```
## [1] "Toyota Corolla"     "Cadillac Fleetwood"
```

Cluster analysis

Cluster diagram

```
car.clust <- hclust(car.em, method="average")
car.clust$labels <- mtcars2$make.model
par(mar=c(5,5,0,10))
plot(as.dendrogram(car.clust),
     horiz=TRUE,
     xlab="Make and model",
     ylab="Euclidean distance", las=1)
```

- Which three cars are the most similar, based on these data? How do you know, and what might account for their similarity?

  The three Mercedes 450s are very similar to each other, with very low branches representing minimal Euclidean distance (Fig. 2). Their similarity is likely due to the fact that they are different trim packages of the same make and model.

- Which car or cars are the most unique? How do you know?

  The Toyota Corolla and the Cadillac Fleetwood are the most unique *pair*, as they have the maximum Euclidean distance (425.3).
  There's also an argument for the Maserati Bora being the most unique individual car, as it has a high-level branch all to itself and the only relationship it really has with the other cars is that it is simply in the dataset. The clustering analysis couldn't group it with any other car or cars.

- Which car is more similar to the Toyota Corolla: the Porsche 914-2, or the Duster 360? How do you know?

  The Porsche 914-2 is more similar to the Toyota Corolla, as they share the same high-level cluster (Fig. 2).

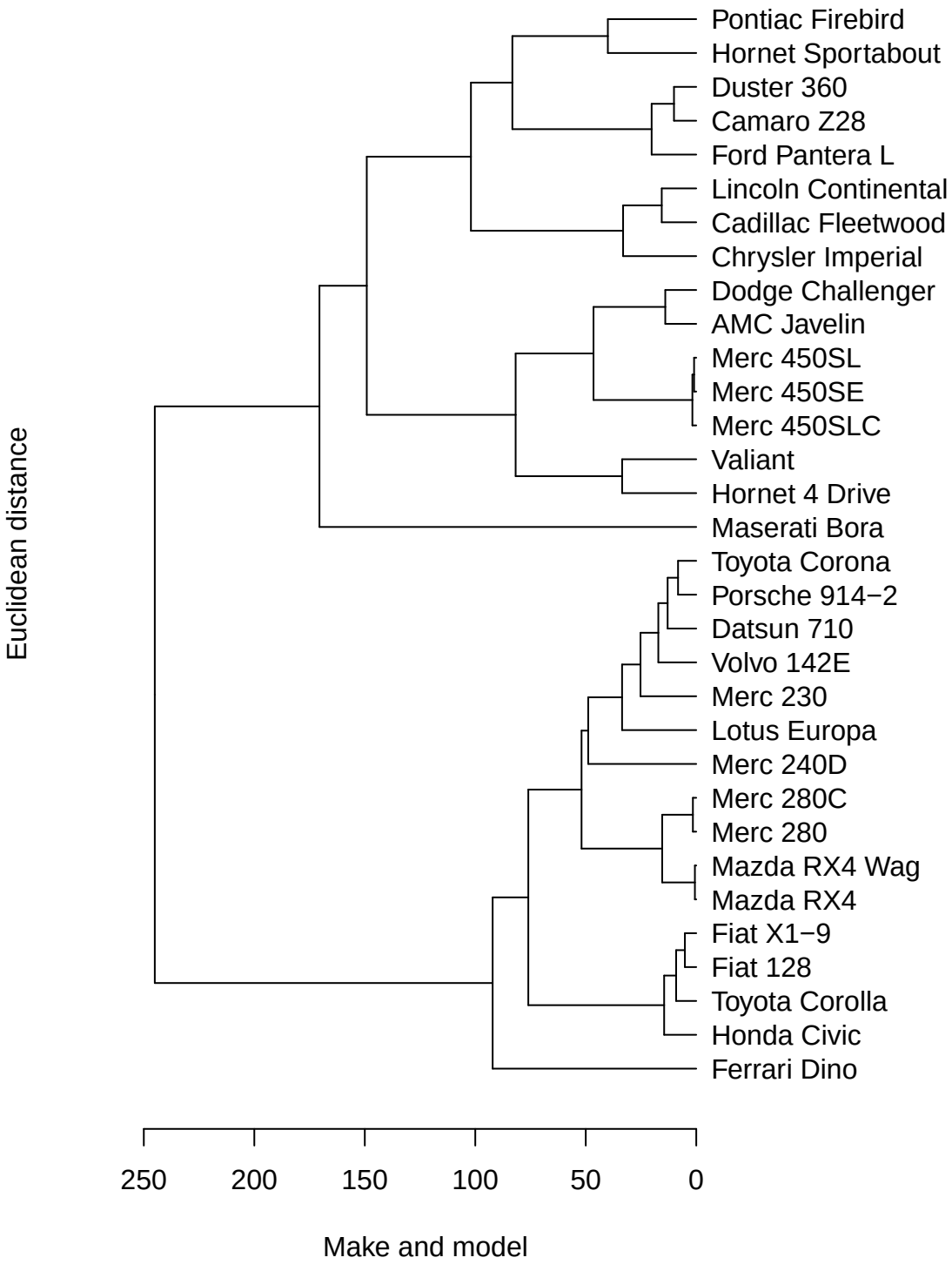


Figure 2: Cluster diagram of the `mtcars2` dataset based on the Euclidean distance measure.

Visualize clusters

```
par(mar=c(1,5,0,0))
plot(car.clust, labels=mtcars2$make.model,
         main=" ",
       xlab="Make and model",
       ylab="Euclidean distance", las=1)
  rect.hclust(car.clust, 2, border="red")
par(mar=c(1,5,0,0))
plot(car.clust, labels=mtcars2$make.model,
       main=" ",
       xlab="Make and model",
       ylab="Euclidean distance", las=1)
  rect.hclust(car.clust, 3, border="blue")
```

- At which Euclidean distance do the two-group and three-group clustering scenarios cut the tree?

  k=2 cuts the tree at 200, while k=3 cuts at 150.

- How many clusters would be formed if one were to cut the tree at 150?

  As stated above, 3. But in the spirit of the question, say one cut the tree at *125*, there would be 4 groups.

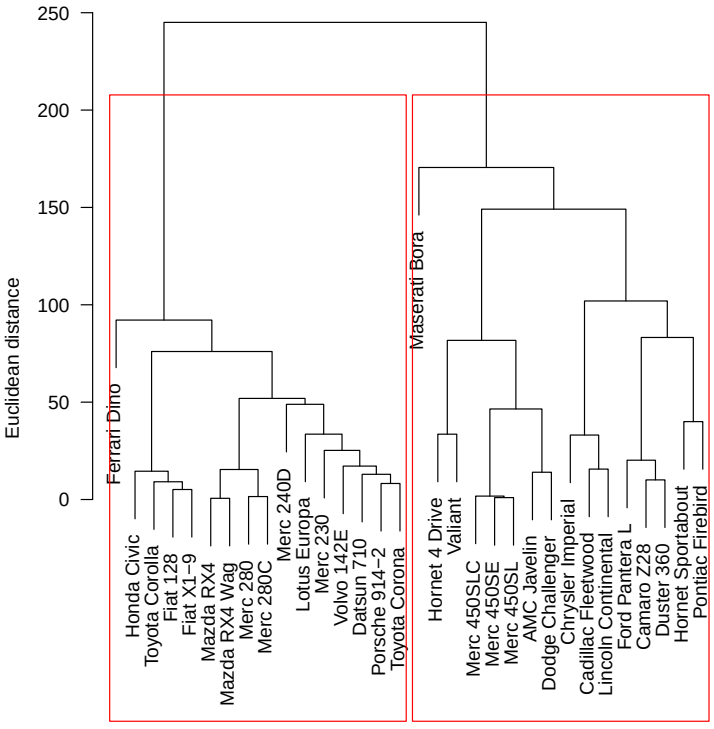
k-means clustering

Determine best number of clusters

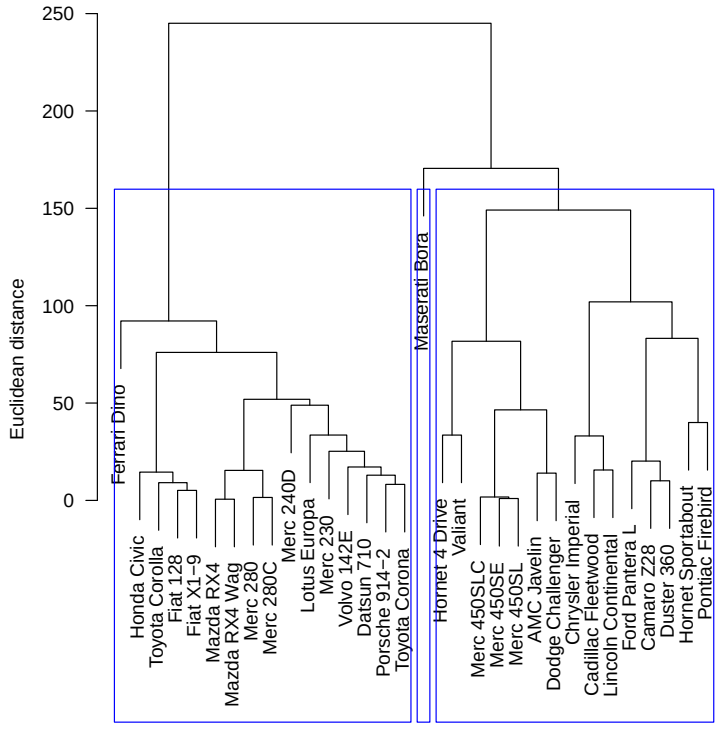
```
try.clusters <- 1:8

k.groups <- data.frame(matrix(
  ncol=length(try.clusters),
  nrow=length(mtcars2$make.model)))
colnames(k.groups) <- paste0("clstrs",
                             c(try.clusters))

k.ss <-data.frame(matrix(nrow=length(try.clusters), ncol=2))
colnames(k.ss) <- c("clusters",
                    "diff.SumSquares")

for(i in 1:length(try.clusters)) {
  cl <- kmeans(car.em, i)
  k.groups[i] <- cl$cluster
  k.ss[i,] <- cbind(i, with(cl, totss-betweenss))}

plot(diff.SumSquares ~ clusters, k.ss, type="b")
```

Test clusters

```
car.clusts <- data.frame(select(mtcars2, .data$make.model :.data$cyl),
                             k.groups)
```



(a) Groups formed with k=2 clusters.



(b) Groups formed with k=3 clusters.

Figure 3: Groups formed by k=2 and k=3 clusters.

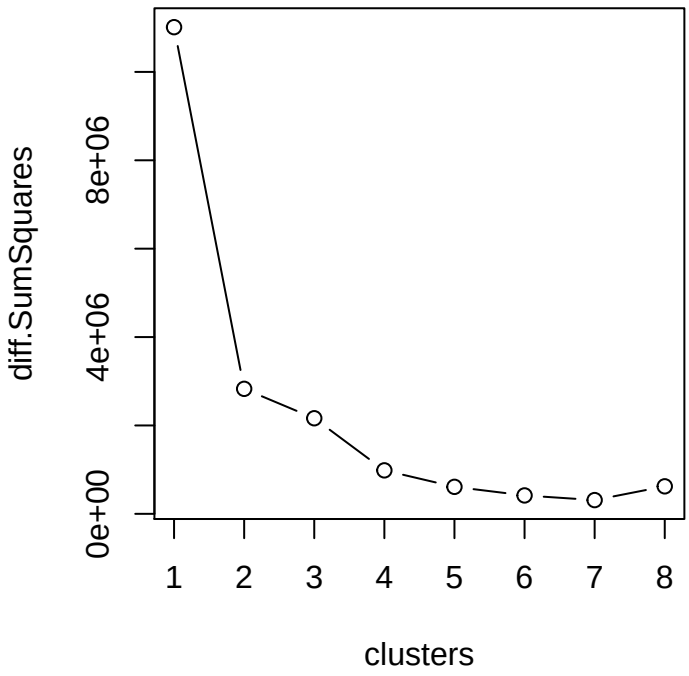


Figure 4: Results of k-means clustering showing the lowest residual sums of squares in the two-cluster solution.

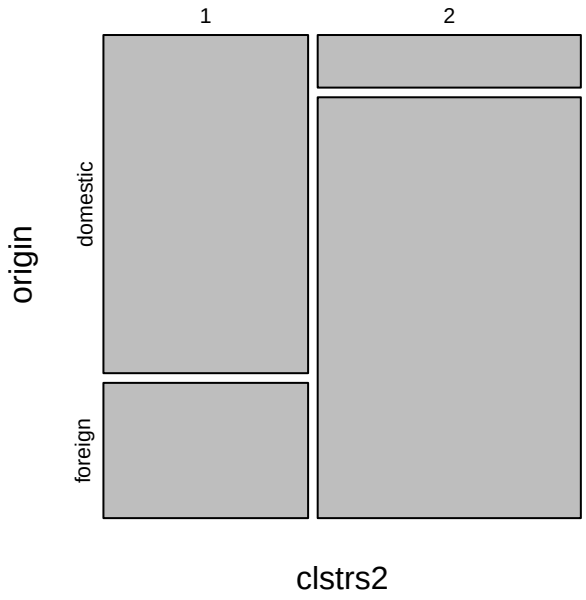


Figure 5: Mosaic plot showing proportion of each car cluster of USA or foreign origin.

```
tabOR <- with(car.clusts, table(clstrs2, origin))
plot(tabOR, main=" ")
```

```
summary(tabOR) %>%
  pander::pander("Chi square test on association between
                 car origin and two-cluster solution.")
```

Number of cases in table: 32 Number of factors: 2

Table 1: Chi square test on association between car origin and two-cluster solution.

Chisq	df	p.value
12.22	1	0.0004717

Based on these results, we can say there is a significant difference between the two clusters in terms of their origins: one cluster is primarily domestic, US automobiles, while the other is predominantly foreign.

Bonus round

The Toyota Corona and Porsche 914 are not very similar cars (Fig. 6).

```
knitr::include_graphics(c('1973ToyotaCorona', '1973Porsche914'))
```

(a) A 1973 Toyota Corona.

(b) A 1973 Porsche 914.

Figure 6: Even when they are the same color they do not look much alike.