# Domácí úkol 3

## Shluková analýza

Marie Melínová

Dle zadání má datový soubor obdahovat údaje o 136 modelech automobilů vybraných značek s určitými specifiky. Úkolem je tedy vybrat určitý počet proměnných k identifikaci tržních segmentů pomocí vybrané metody shlukové analýzy.

Ze všeho nejdřív si opět načteme potřebná data, se kterými budeme následně pracovat. Jelikož má každý sloupec v datech jinou jednotku, přeškálujeme si data na hodnotu mezi 0 a 1.

Při analýze jsem si také všimla, že řádek 42 s hodnotou "Honda accord" byl duplicitní. Proto jsem ho z analýzy vynechala.

```
library(foreign)
data0 = read.spss("du3.sav",to.data.frame=TRUE)
data0 <- data0[-42,]

data <- as.data.frame(sapply(data0[,2:17], function(x) {(x-min(x))/(max(x)-min(x))}))
row.names(data) <- data0$model
colnames(data) <- c("Cena",  "ObjemValcu", "Vykon", "MaxRychlost", "Zrychleni",
                    "SpotrebaMesto", "SpotrebaMimoMesto", "SpotrebaKombinovana",
                    "Emise", "Hmotnost", "Delka", "Vyska", "Vaha", "RozvorKol",
                    "UzitnaHmotnost", "ObjemZavazProst")

dataT <- t(data)
dataT[,sample(1:135, 2)]
```

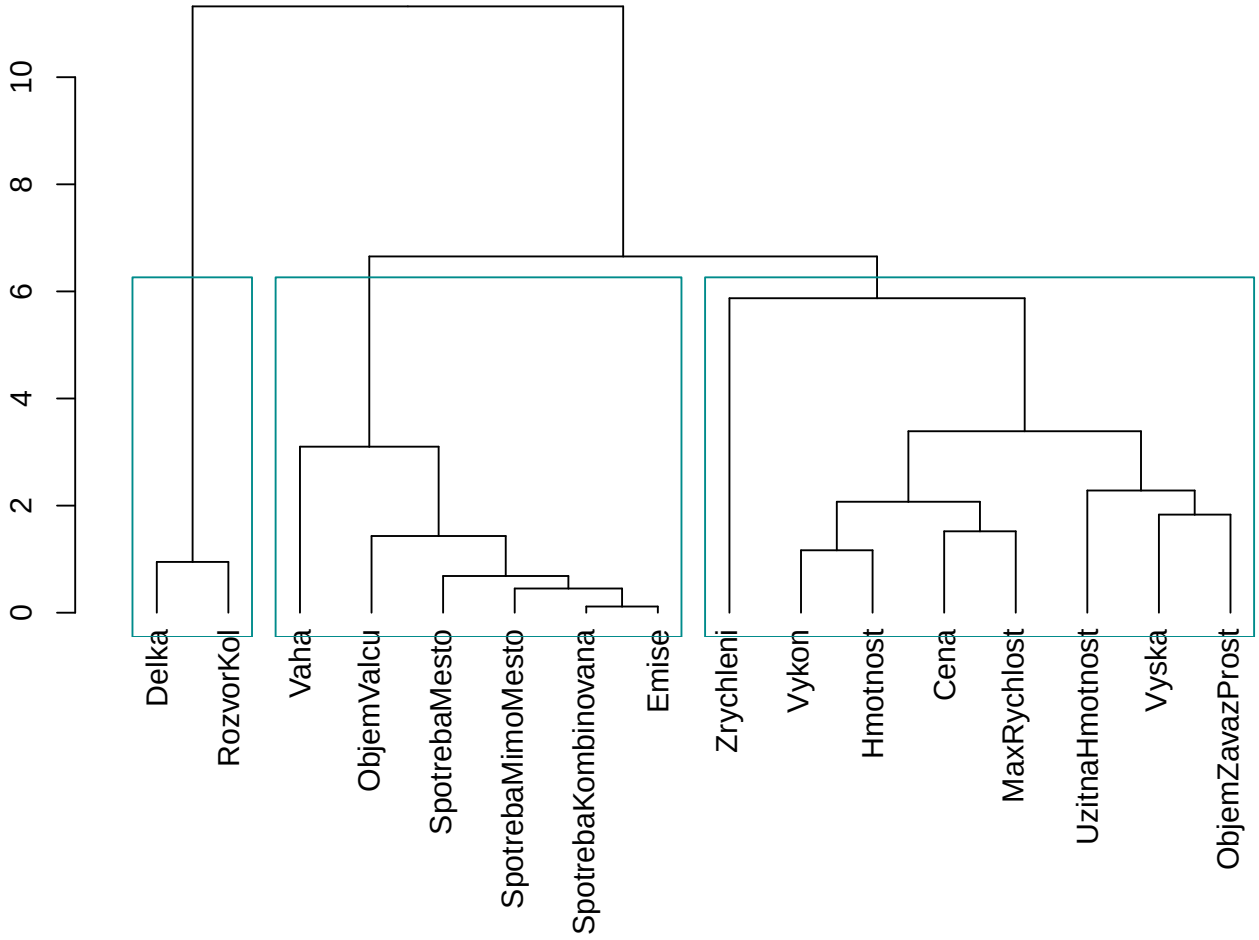
```
##                     Peugeot 308 SW            Seat Exeo
## Cena                     0.2734772             0.43813452
## ObjemValcu               0.1509221             0.22544223
## Vykon                    0.2142857             0.25000000
## MaxRychlost              0.3529412             0.49411765
## Zrychleni                0.5196078             0.44117647
## SpotrebaMesto            0.3165468             0.41007194
## SpotrebaMimoMesto        0.2083333             0.27777778
## SpotrebaKombinovana      0.2500000             0.33333333
## Emise                    0.2597403             0.32900433
## Hmotnost                 0.3995726             0.37037037
## Delka                    0.7801236             0.86302781
## Vyska                    0.5650000             0.45750000
## Vaha                     0.3596674             0.08108108
## RozvorKol                0.7704026             0.69858542
## UzitnaHmotnost           0.4163265             0.39795918
## ObjemZavazProst          0.6737235             0.53300125
```

Nyní si z dat vytvoříme dendogram a pokusíme se v něm vyznačit rozumné množství shluků. Pro shlukovou analýzu jsem vybrala hierarchické shlukování, euklidovskou vzdálenost a Wardovu metodu.

Hierarchické shlukování jsem zvolila pro jeho schopnost vytvořit dendogram, který umožňuje jednoduše vizualizovat vztahy mezi proměnnými a identifikovat ty proměnné, které si jsou mezi sebou podobné.

```
D_matrix <- dist(dataT, method = 'euclidean')
clust <- hclust(D_matrix, method = "ward.D")

dendogram <- as.dendrogram(clust)
par(mar=c(10,4,2,2))
plot(dendogram)
rect.hclust(clust, k = 3, border = "darkcyan")
```




Shlukovou analýzou jsme proměnné rozdělili do tří shluků. První shluk, který obsahuje proměnné `Delka` a `RozvorKol` bych zkráceně nazvala jako **velikost**.

Druhý shluk, obsahující proměnné `Vaha`, `ObjemValcu`, `SpotrebaMesto`, `SpotrebaMimoMesto`, `SpotrebaKombinovana` a `Emise`, bych nazvala jednotně jako **spotřeba**.

Poslednímu shluku, který obsahuje proměnné `Zrychleni`, `Vykon`, `Hmotnost`, `Cena`, `MaxRychlost`, `UzitnaHmotnost`, `Vyska` a `ObjemZavazProstoru`, bych ponechala název **rychlost** nebo **výkon**.