Shluková analýza

4ST512 Vícerozměrná statistika

Lubomír Štěpánek

10. prosince 2017

Obsah

1 Zadání úlohy

Východiskem je soubor `du3.sav`, který obsahuje celkem 16 proměnných (sloupců) týkajících se vybraných charakteristik automobilů a 136 pozorování (řádků) tvořených daty jednotlivých modelů automobilů. Od daného modelu jsou vždy uvedeny hodnoty parametrů nejlevnější verze. Celkově jde o modely, které k datu sestavení datasetu nepřesáhly tržní cenou 40 000 €.

Cílem je identifikace potenciálních tržních segmentů mezi automobily pomocí některé metody shlukové analýzy. Z šestnácti proměnných vyberme nejméně tři a nejvíce dvanáct podle vlastního uvážení.

(i) Výběr podmnožiny použitých proměnných racionálně zdůvodňeme.

(ii) Proveďme explorativní analýzu dat.

(iii) Vyberme si některou metodu z rodiny shlukové analýzy, výběr zdůvodněme.

(iv) Proveďme shlukování na rozumném počtu shluků, eventuálně použijme metodu, která apriorní odhad počtu shluků nevyžaduje. Shluky se pokusme slovně vystihnout, interpretovat.

Vysázeno pomocí R-ka, YAML-u, TEX-u a BibTEX-u

2 Řešení úlohy

2.1 Metodologie a analýza dat

Celá úloha byla řešena v prostředí R, které je určeno pro statistické výpočty a následné grafické náhledy [1]. Datový soubor `du3.sav` byl nahrán do prostředí R pomocí balíčku `foreign`.

Jednotlivé výpočty v rámci shlukové analýzy a další přidružené výpočty byly provedeny zejména pomocí R-kového balíčku `stats`.

2.1.1 Výběr podmnožiny použitých proměnných

Výběr provedeme dle direktivy ze zadání expertně a podložíme jej racionální argumentací, ideálně s pomocí určité doménové znalosti.

Samozřejmě by bylo možné provést výpočetně náročnější[1] exhaustivní *greedy* postup, během kterého by bylo postupně uplatněno všech $\sum_{i=3}^{12} \binom{16}{i} = 64702$ možností, jak z 16 zadaných proměnných postupně vybrat navzájem různých 3, 4, ..., 12 použitých proměnných; pro každý výběr by bylo poté možné provést shlukovou analýzu a vždy hodnotit dle některého kritéria její validitu.

2.1.2 Explorativní analýza dat

Přestože většina metod z rodiny shlukování nevyžaduje žádné silné vstupní předpoklady (v podstatě jde jen o vhodnou proporci počtu proměnných ku pozorování a numerický, nebo alespoň kategoriální charakter hodnot proměnných), explorativní analýza dat je vždy vhodnou iniciální fází pohledu na data.

Vzhledem k numerickému charakteru proměnných bude pomocí boxplotů zkoumána symetrie výběrových rozdělení a případná přítomnost odlehlých hodnot.

Pro detekci možných odlehlých pozorování byla použita metoda *inner and outer fences* (*vnitřní a vnější hradby*) [2].

Je-li $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$, pak suspektně odlehlá v rámci výběru $\boldsymbol{x}$ je ta hodnota x_i, pro kterou platí, že

$$x_i \notin \langle Q_1 - 1{,}5 \cdot (Q_3 - Q_1); Q_3 + 1{,}5 \cdot (Q_3 - Q_1) \rangle,$$

kde $i \in \{1, 2, \ldots, n\}$, n je rozsah souboru, Q_1 a Q_3 je první a třetí výběrový kvartil nad vektorem dat $\boldsymbol{x}$, respektive. Graficky suspektní odlehlost zobrazují i krabicové diagramy pomocí samostatně zobrazených pozorování mimo krabici danou kvartilovým rozpětím. Bylo by možné použít i formální testy hypotéz na jednorozměrnou odlehlost v rámci výběru $\boldsymbol{x}$, například Dixonův test nebo Grubbsův test.

[1] Algoritmus by proběhl v polynomickém čase $\Theta(n^{12})$, ke n je počet všech proměnných.

Nicméně vždy jde jen o podpůrnou metodu, konečné rozhodnutí o odlehlosti vytipované hodnoty má spíše expertní charakter. Zvláště v použitém datasetu mají všechny hodnoty charakter měřitelných údajů, proto suspektní odlehlosti nebudeme přikládat zásadní význam.

2.1.3 Shluková analýza

V této práci použijeme *hierarchickou* shlukovou analýzu, která využívá *Wardovu metodu* nad maticí nepodobností, jež je měřena čtvercem eukleidovské vzdálenosti, [3], tedy

$$d_{ij} = d(\{X_i\}, \{X_j\}) = \|X_i - X_j\|^2,$$

kde d_{ij} je vzdálenost mezi dvěma shluky a X_i, resp. X_j je i-té, resp. j-té pozorování. Pokud vyčíslíme vzdálenosti d_{ij} mezi všemi dvojicemi i-tého a j-tého pozorování pro $\forall i, j \in \{1, 2, \ldots, n\}$, kde n je počet pozorování, a uspořádáme je do čtvercové matice $\mathcal{D}^{n,n}_{i,j=1,1}$, pak $\mathcal{D}$ je matice nepodobností.

Wardova metoda je aglomerativní, tj. na počátku jsou všechna pozorování v p-rozměrném prostoru, kde p je počet použitých proměnných, považována za samostatné shluky. Poté jsou shluky v dalších krocích iterativně slučovány, čímž roste úměrně jejich velikost a klesá jejich počet. Objektivní funkcí, která je použita pro rozhodnutí, které dva shluky budou v daném kroku sloučeny v jeden, je minimalizace nárůstu celkového vnitroshlukového rozptylu, [4]. Princip Wardovy metody tedy tkví nalezení takové „architektury“ shluků, přiražení původních pozorování do těchto shluků, aby úhrnný vnitroshlukový roztpyl byl nejmenší možný (*Wardovo kritérium minimální celkové variance*).

V prostředí R je shlukování pomocí Wardovy metody nad vhodným datasetem `x` implementováno funkcí a argumenty `hclust(dist(x, method = 'euclidean'), method = 'ward.D2')`, [5].

K zobrazení jednotlivých shluků a jejich složení poslouží *dendrogram*, který je v podstatě „mapou“ iterativního procesu shlukování – pro každý okamžik, kdy počet shluků klesne o jedna (a vznikne jeden nový, větší), je takové sloučení v dendrogramu naznačeno na vodorovné pozici ukazující vzdálenost (nepodobnost) mezi dvěma původními shluky, které daly v daném kroku vznik zmíněnému jednomu novému, většímu.

2.2 Výsledky

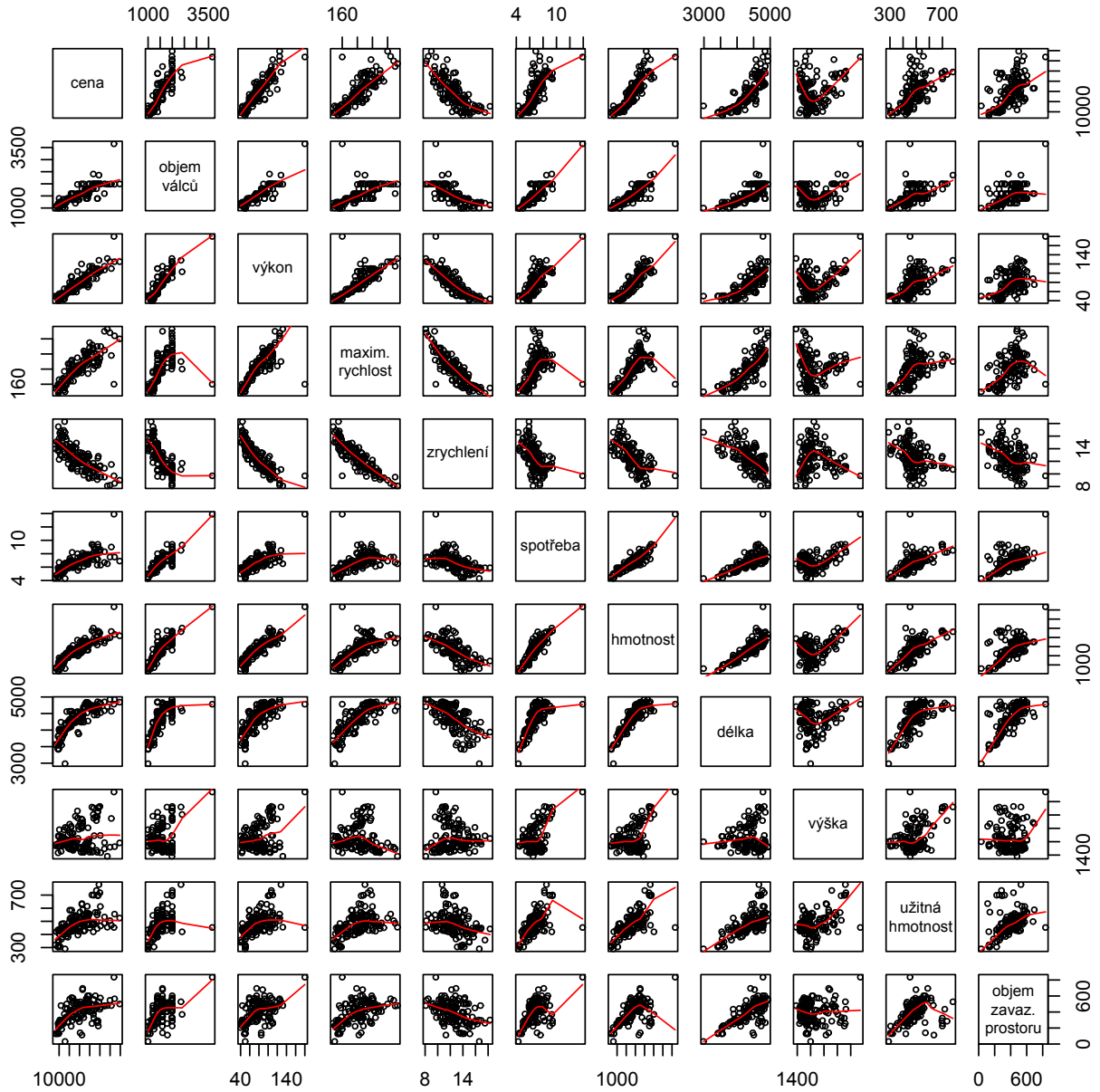
Výběr podmnožiny použitých proměnných. Vzhledem k tomu, že shluková analýza má napovědět tržní segmentaci automobilů, pokusme se expertně rozhodnout, které proměnné je možné z původní šestnáctice vynechat, aby byla informace o modelech automobilů stále ještě dostatečně vysoká.

- Omezme informaci o spotřebě pouze na kombinovanou [l], tj. nadále nepoužívejme proměnné `spotřeba - město (l)` a `spotřeba - mimo město (l)`.
- Předpokládejme, že až na některé výjimky jsou emise všech modelů nějakým způsobem legislativně regulovány a neměl by mezi nimi být nápadný rozdíl (takže na tržní segmentaci nepředpokládáme tak výrazný vliv); vynechme tedy proměnnou `emise (g/km)`.

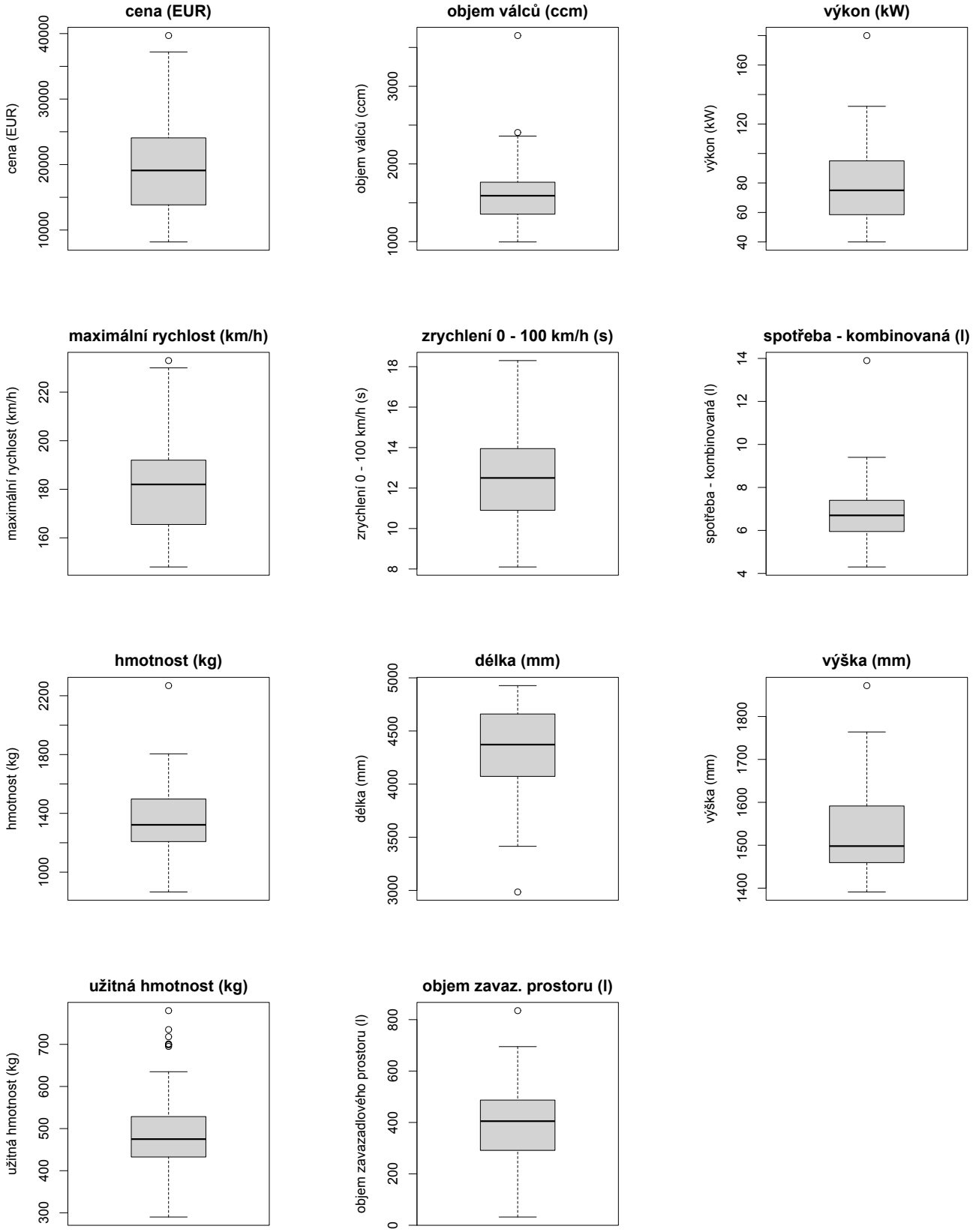
- Z rozměrových parametrů může hrát roli délka a výška vozu; naopak `šířka (mm)` a `rozvor kol (mm)` jsou shora omezeny běžnými šířkami části jednoho pruhu vozovky a opět předpokládejme, že trh nebudou zásadně segmentovat.

Ostatní proměnné použijme, tedy nadále počítejme s `cena (EUR)`, `objem válců (ccm)`, `výkon (kW)`, `maximální rychlost (km/h)`, `zrychlení 0 - 100 km/h (s)`, `spotřeba - kombinovaná (l)`, `hmotnost (kg)`, `délka (mm)`, `výška (mm)`, `užitná hmotnost (kg)` a `objem zavazadlového prostoru (l)`.

Explorativní analýza dat. Na obrázku 1 je matice bodových diagramů.



Obrázek 1: Matice bodových diagramů s naznačeným lokálně-váženým (LOESS) regresním trendem (červeně).



Obrázek 2: Krabicové diagramy pro výběry všech proměnných zájmu. Prázdná kolečka značí odlehlé hodnoty.

Pohled na „dvourozměrné“ regresní závislosti mezi všemi (numerickými) dvojicemi proměnných může napovědět, které z nich spolu pozitivně, a které negativně korelují. To lze nakonec využít i během interpretace významu jednotlivých shluků.

Na obrázku 2 jsou krabicové diagramy postupně pro všechny výběrové proměnné našeho zájmu. I přes přítomnost některých suspektně odlehlých hodnot (v krabicových diagramech jako prázdná kolečka) je za skutečně odlehlé nepovažujeme, neboť se jedná u všech proměnných o měřená a dohledatelná data. Při exploraci však byla objeveno duplicitní pozorování – model *Honda Accord* byla ve vstupním datasetu obsažena dvakrát, oba řádky však vyly identické. Jeden byl tedy před další analýzou odstraněn. Dataset neobsahuje žádné chybějící hodnoty.

Shluková analýza. Wardova metoda hierarchického shlukování je obhajitelná, neboť počet pozorování ($n = 135$) není příliš velký.

Na obrázku 2 však vidíme, že variabilita jednotlivých proměnných je značně odlišná, před samotným shlukováním je tedy nutná standardizace. Ta byla provedena projekcí[2] hodnot každé proměnné na interval $\langle 0, 1\rangle$. Tím se variabilita všech proměnných stává navzájem porovnatelnou.

Na obrázku 3 vidíme dendrogram, který je výsledkem iterativního procesu hierarchického shlukování podle Warda. Na ještě interpretovatelné úrovni jsme schopni odlišit šest nápadných shluků, které se postupně spojují ve větší shluky na úrovni nepodobnosti cca 1,0 až 3.0 ($\times$100000); tyto shluky jsou vykresleny postupně na obrázcích 4 až 9.

Interpretace jednotlivých shluků je relativně nesnadná, je třeba využít (ideálně) doménovou znalosti a hodnoty vybraných proměnných vždy pro některého či některé reprezentanty daného shluku.

V prvním shluku (obrázek 4) je relativně málo modelů a jedná se obecně o drahé, luxusní vozy, mnohdy větších rozměrů než typu *sedan*, mnohdy jde o modely typu *SUV* s větším úložným prostorem a větší hmotností. Některé z nich jsou však sedany a ty jsou pak intuitivně menší, s menším zavazadlovým prostorem a rychlejší. Jsou zde případně i vysoce luxusní modely (Mercedes C, BMW 5), typické pro vozoparky top manažerů.

Ve druhém shluku (obrázek 5) jsou sdruženy opět drahé a luxusní vozy, spíše běžných sedanových rozměrů (ale není to podmínkou, některé jsou typu SUV), některé modely jsou i sportovní, rychlé (Audi A4, Alfa Romeo 159 Sportwag).

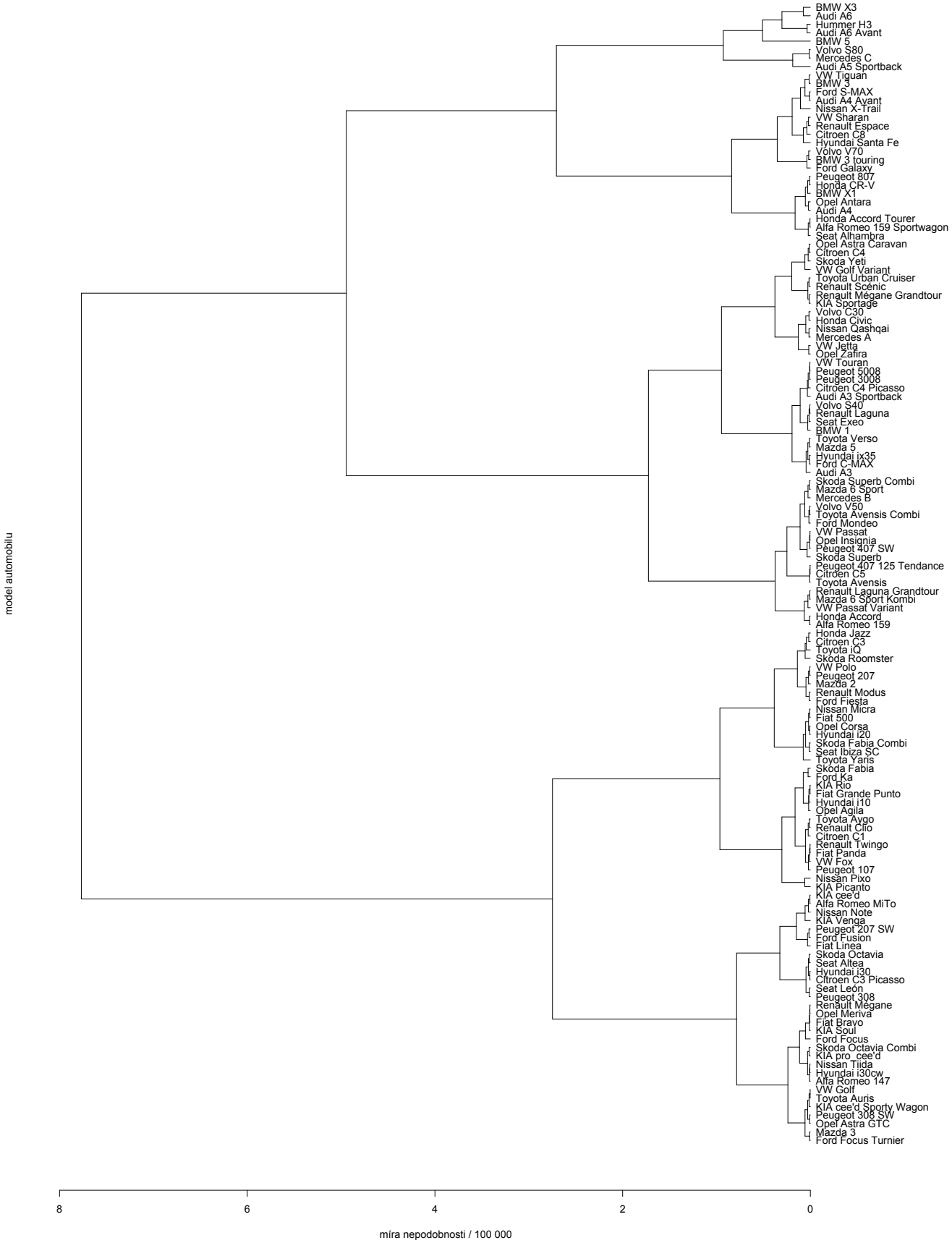
Ve třetím shluku (obrázek 6) nacházíme spíše dražší než středně drahé vozy „městského“, resp. rodinného charakteru. Obvykle zaujímají ve sledovaných proměnných průměrné hodnoty.

Zajímavý je čtvrtý shluk (obrázek 7), který se od předchozího liší v podstatě jen tím, že jde často o modely typu *combi*, tedy s větší délkou vozu.

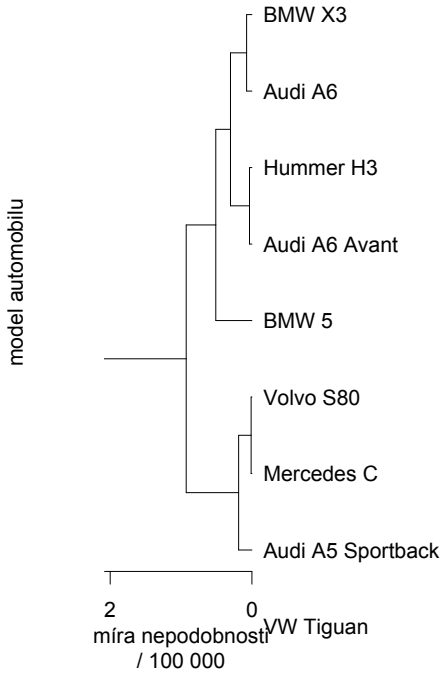
Pátý shluk (obrázek 8) modelů je relativně heterogenní, lze zde vysledovat „podshluk“ malých vozů (s krátkou délkou vozu), např. Fiat Punto, Renault Clio, KIA Picanto, které jsou i relativně levné (a mají i adekvátně menší hmotnost, menší úložný prostor apod.) a „pomalé“. Některé modely se z předchozí charakteristiky vymaňují, např. Škoda Roomster je spíše větší model téměř typu SUV.

Šestý shluk (obrázek 9) pak sdružuje rozšířené, oblíbené modely, spíše lacinější, běžné velikosti a výkonu, velmi pravděpodobně masově prodávané.

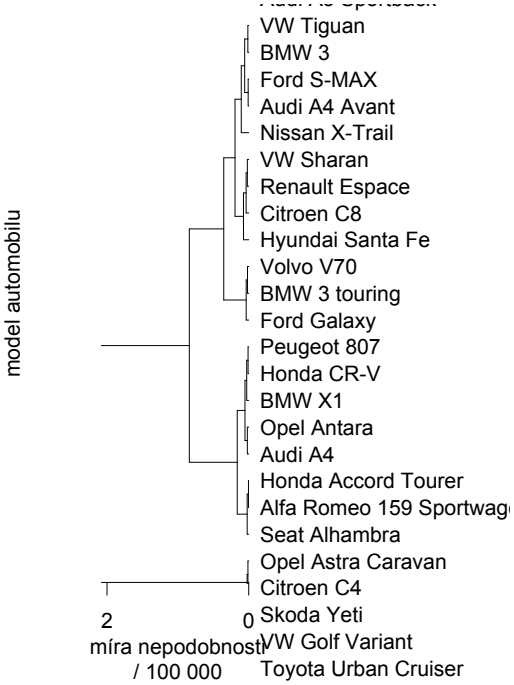
[2]Pro každou proměnnou $\boldsymbol{X} = (x_1, x_2, \ldots, x_n)^T$ byla provedena standardizace na proměnnou $\boldsymbol{X}' = (x_1', x_2', \ldots, x_n')^T$ tak, že $x_i' = \frac{x_i - \min\{\boldsymbol{X}\}}{\max\{\boldsymbol{X}\} - \min\{\boldsymbol{X}\}}$ pro $\forall i \in \{1, 2, \ldots, n\}$.



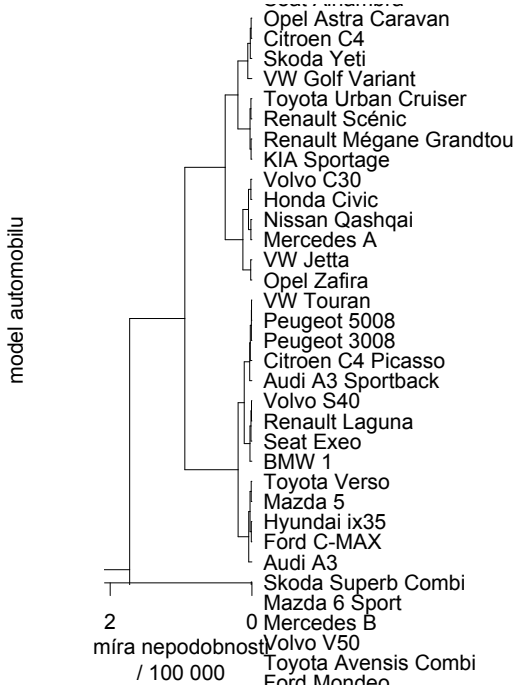
Obrázek 3: Dendrogram naznačující shlukové uspořádání jednotlivých modelů automobilů



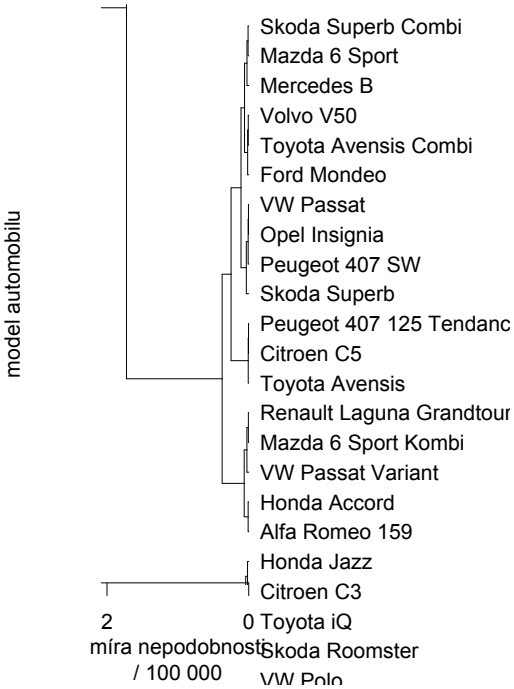
Obrázek 4: Velmi drahé, špičkové modely



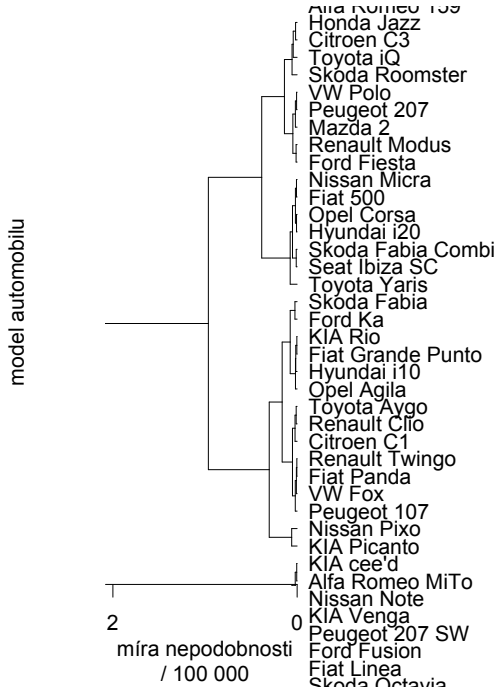
Obrázek 5: Drahé, luxusní vozy



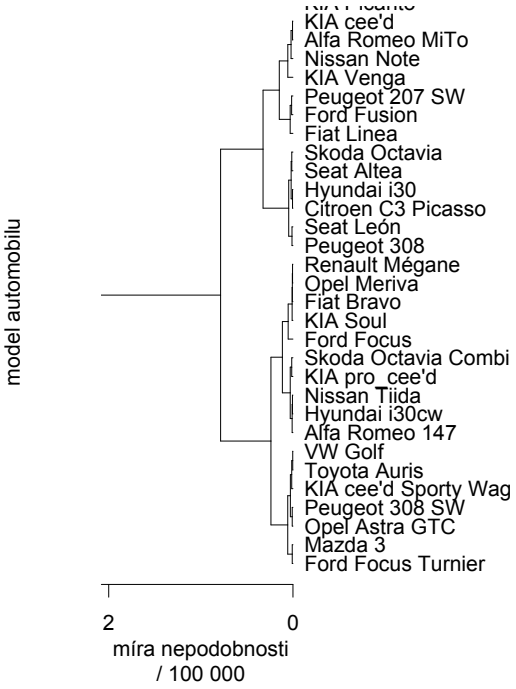
Obrázek 6: Klasické městské vozy, spíše dražší



Obrázek 7: Dostupné městské vozy, často *combi*



Obrázek 8: Heterogenní skupina vozů, spíše levnější, mnohdy menších rozměrů



Obrázek 9: Spíše levnější a užitkové vozy, rozšířené, oblíbené

2.3 Závěr

Pro analýzu byly expertně zvoleny některé proměnné, které dle autora mohou mít vliv na tržní segmentaci populace automobilů.

Explorativní grafická analýza dat ukázala některé očekávané závislosti mezi sledovanými proměnnými, např. že spotřeba roste úměrně hmotnosti modelu apod. Do jisté míry lze tyto „naivní“ znalosti použít i při interpretaci významu jednotlivých shluků.

Byla provedena shluková hierarchická analýza založená na Wardově metodě. Velikost datasetu je pro tuto metodu vhodná. Výsledný dendrogram naznačil, že rozumně interpretovatelné množství shluků je nejspíše šest. I přesto je interpretace typického zástupce každého shluku relativně obtížná a ne vždy jednoznačná.

Důležitým závěrem je, že shluky se mezi sebou mimo jiné liší průměrnou tržní cenou, za kterou jsou modely ve shluku prodávány. To odpovídá očekávání, kdy cena je obvykle rozumnou „výslednicí“ ostatních parametrů (výkonu, spotřeby, velikosti, přepychovosti) modelu. Ostatní interpretační prvky je nutné odvodit ze znalosti domény či manuálním prozkoumáním číselných hodnot parametrů v daném shluku.

Hierarchická shluková analýza je relativně rychlou a výpočetně nenáročnou metodou, dendrogram je uživatelsky velmi přívětivým grafickým výstupem.

3 Apendix

Zde je uveden kód v jazyce R, ve kterém byly zpracovávány veškeré výpočty a rovněž generovány diagramy.

```
################################################################################
################################################################################
################################################################################

## instaluji a loaduji balíčky ------------------------------------------------

invisible(
    lapply(
        c(
            "xtable",
            "openxlsx",
            "foreign"
        ),
        function(my_package){

            if(!(my_package %in% rownames(installed.packages()))){

                install.packages(
                    my_package,
                    dependencies = TRUE,
                    repos = "http://cran.us.r-project.org"
                )

            }

            library(my_package, character.only = TRUE)

        }
    )
)


## ----------------------------------------------------------------------------

################################################################################

## nastavuji handling se zipováním v R ----------------------------------------

Sys.setenv(R_ZIPCMD = "C:/Rtools/bin/zip")


## ----------------------------------------------------------------------------

################################################################################
```

```
## nastavuji pracovní složku ------------------------------------------------

while(!"__domaci_ukol_3__.R" %in% dir()){
    setwd(choose.dir())
}

mother_working_directory <- getwd()


## ------------------------------------------------------------------------

################################################################################

## vytvářím posložky pracovní složky ----------------------------------------

setwd(mother_working_directory)

for(my_subdirectory in c("vstupy", "vystupy")){

    if(!file.exists(my_subdirectory)){

        dir.create(file.path(

            mother_working_directory, my_subdirectory

        ))

    }

}


## ------------------------------------------------------------------------

################################################################################

## loaduji data --------------------------------------------------------------

setwd(
    paste(mother_working_directory, "vstupy", sep = "/")
)

my_data <- read.spss(

    file = "du3.sav",
    to.data.frame = TRUE

)
```

```r
setwd(mother_working_directory)


## ------------------------------------------------------------------------

################################################################################

## (pre)processing dat ---------------------------------------------------------

#### doplňuji jména proměnným v datasetu "my_data" ----------------------------

for(i in 1:dim(my_data)[2]){

    if(unname(attr(my_data, "variable.labels"))[i] != ""){

        colnames(my_data)[i] <- unname(
            attr(
                my_data,
                "variable.labels"
            )
        )[i]

    }

}


#### odstaňuji z datasetu duplicity -------------------------------------------

my_data <- my_data[!duplicated(my_data$model), ]


#### z hodnot proměnné "model" vytvářím názvy řádků --------------------------

rownames(my_data) <- as.character(my_data[, "model"])

my_data <- my_data[, setdiff(colnames(my_data), "model")]


## ------------------------------------------------------------------------

################################################################################

## omezuji dataset na proměnné zájmu -------------------------------------------

#### ponechávám následující ---------------------------------------------------

# cena (EUR),
# objem válců (ccm),
```

```
# výkon (kW),
# maximální rychlost (km/h),
# zrychlení 0 -- 100 km/h (s),
# spotřeba - kombinovaná (l),
# hmotnost (kg),
# délka (mm),
# výška (mm),
# užitná hmotnost (kg) a
# objem zavazadlového prostoru (l)

#### odstraňuji následující ----------------------------------------------------

# spotřeba -- město (l)
# spotřeba -- mimo město (l)
# emise (g/km)
# šířka (mm)
# rozvor kol (mm)

my_data <- my_data[

    ,
    setdiff(
        colnames(my_data),
        c(
            "spotřeba - město (l)",
            "spotřeba - mimo město (l)",
            "emise (g/km)",
            "šířka (mm)",
            "rozvor kol (mm)"
        )
    )

]


## ------------------------------------------------------------------------

################################################################################

## exploративní analýza dat ----------------------------------------------------

#### vytvářím boxploty --------------------------------------------------------

setwd(paste(mother_working_directory, "vystupy", sep = "/"))

for(my_variable in colnames(my_data)){

    cairo_ps(
        file = paste(
```

```
            gsub(
                "[ /\\(\\)]",
                "_",
                iconv(my_variable, to = "ASCII//TRANSLIT")
            ),
            "_boxplot.eps",
            sep = ""
        ),
        width = 5,
        height = 6,
        pointsize = 18
    )

    par(mar = c(4.1, 4.1, 2.1, 0.1))

    boxplot(
        x = my_data[, my_variable],
        col = "lightgrey",
        ylab = my_variable,
        main = if(my_variable == "objem zavazadlového prostoru (l)"){
            "objem zavaz. prostoru (l)"
        }else{
            my_variable
        }
    )

    dev.off()

}


#### dummy prázdný čtverec pro potřeby sazby dokumentu ------------------------

cairo_ps(
    file = "empty_square.eps",
    width = 5,
    height = 6,
    pointsize = 18
)

par(mar = c(4.1, 4.1, 2.1, 0.1))

plot(0, type = 'n', axes = FALSE, ann = FALSE)

dev.off()


setwd(mother_working_directory)
```

```
#### ukládám matici scatterplotů ------------------------------------------

setwd(paste(mother_working_directory, "vystupy", sep = "/"))

cairo_ps(
    file = "scatterplot_matrix.eps",
    width = 8,
    height = 8,
    pointsize = 14
)

par(mar = c(0.1, 0.1, 0.1, 0.1))

temp_data <- my_data
#colnames(temp_data) <- gsub("(.*)( \\(.*)", "\\1", colnames(my_data))
colnames(temp_data) <- c(
    "cena",
    "objem\nválců",
    "výkon",
    "maxim.\nrychlost",
    "zrychlení",
    "spotřeba",
    "hmotnost",
    "délka",
    "výška",
    "užitná\nhmotnost",
    "objem\nzavaz.\nprostoru"
)

pairs(
    temp_data,
    panel = "panel.smooth",
    cex = 0.6#0.75
)

dev.off()

setwd(mother_working_directory)


## ------------------------------------------------------------------------

################################################################################

## shluková analýza ------------------------------------------------------

#### standardizace proměnných na interval <0, 1> ----------------------------

for(i in 1:dim(my_data)[2]){
```

```
    my_data[, i] <- (
        my_data[, i] - min(my_data[, i])
    ) / (
        max(my_data[, i]) - min(my_data[, i])
    )

}


#### hierarchická shluková analýza, Wardova metoda ----------------------------

my_hclust <- hclust(dist(x, method = 'euclidean'), method = 'ward.D2')



#### ukládám dendrogram -------------------------------------------------------

setwd(paste(mother_working_directory, "vystupy", sep = "/"))

cairo_ps(
    file = "dendrogram.eps",
    width = 18,
    height = 24,
    pointsize = 14
)

par(mar = c(4, 4, 0, 10))

plot(
    as.dendrogram(my_hclust),
    horiz = TRUE,
    xlab = "míra nepodobnosti / 100 000",
    ylab = "model automobilu",
    xaxt = "n",
    axes = FALSE
)

axis(side = 1, at = seq(0, 8e5, by = 2e5), labels = seq(0, 8, by = 2))

dev.off()

setwd(mother_working_directory)


#### subdendrogramy -----------------------------------------------------------

setwd(paste(mother_working_directory, "vystupy", sep = "/"))
```

```
#### spíše levnější a užitkové vozy, rozšířené, oblíbené ----------------------

cairo_ps(
    file = "subdendrogram_1.eps",
    width = 6,
    height = 8,
    pointsize = 18
)

par(mar = c(4, 4, 0.5, 8.2))

plot(
    as.dendrogram(my_hclust),
    horiz = TRUE,
    xlab = "míra nepodobnosti\n/ 100 000",
    ylab = "model automobilu",
    xaxt = "n",
    axes = FALSE,
    xlim = c(2e5, 0),
    ylim = c(1, 29)
)

axis(side = 1, at = seq(0, 8e5, by = 2e5), labels = seq(0, 8, by = 2))

dev.off()


#### heterogenní skupina vozů, spíše levnější, mnohdy menších rozměrů ---------

cairo_ps(
    file = "subdendrogram_2.eps",
    width = 6,
    height = 8,
    pointsize = 18
)

par(mar = c(4, 4, 0.5, 8.2))

plot(
    as.dendrogram(my_hclust),
    horiz = TRUE,
    xlab = "míra nepodobnosti\n/ 100 000",
    ylab = "model automobilu",
    xaxt = "n",
    axes = FALSE,
    xlim = c(2e5, 0),
    ylim = c(30, 60)
)
```

```
axis(side = 1, at = seq(0, 8e5, by = 2e5), labels = seq(0, 8, by = 2))

dev.off()


#### dostupné městské vozy, často combi ---------------------------------------

cairo_ps(
    file = "subdendrogram_3.eps",
    width = 6,
    height = 8,
    pointsize = 18
)

par(mar = c(4, 4, 0.0, 10))

plot(
    as.dendrogram(my_hclust),
    horiz = TRUE,
    xlab = "míra nepodobnosti\n/ 100 000",
    ylab = "model automobilu",
    xaxt = "n",
    axes = FALSE,
    xlim = c(2e5, 0),
    ylim = c(61, 79)
)

axis(side = 1, at = seq(0, 8e5, by = 2e5), labels = seq(0, 8, by = 2))

dev.off()


#### klasické městské vozy, spíše dražší --------------------------------------

cairo_ps(
    file = "subdendrogram_4.eps",
    width = 6,
    height = 8,
    pointsize = 18
)

par(mar = c(4, 4, 0.5, 10))

plot(
    as.dendrogram(my_hclust),
    horiz = TRUE,
    xlab = "míra nepodobnosti\n/ 100 000",
    ylab = "model automobilu",
    xaxt = "n",
```

```
    axes = FALSE,
    xlim = c(2e5, 0),
    ylim = c(80, 106)
)

axis(side = 1, at = seq(0, 8e5, by = 2e5), labels = seq(0, 8, by = 2))

dev.off()


#### drahé, luxusní vozy ---------------------------------------------------

cairo_ps(
    file = "subdendrogram_5.eps",
    width = 6,
    height = 8,
    pointsize = 18
)

par(mar = c(4, 4, 0.0, 10))

plot(
    as.dendrogram(my_hclust),
    horiz = TRUE,
    xlab = "míra nepodobnosti\n/ 100 000",
    ylab = "model automobilu",
    xaxt = "n",
    axes = FALSE,
    xlim = c(2e5, 0),
    ylim = c(107, 127)
)

axis(side = 1, at = seq(0, 8e5, by = 2e5), labels = seq(0, 8, by = 2))

dev.off()


#### velmi drahé, špičkové modely ------------------------------------------

cairo_ps(
    file = "subdendrogram_6.eps",
    width = 6,
    height = 8,
    pointsize = 18
)

par(mar = c(4, 4, 0.0, 10))

plot(
```

```
    as.dendrogram(my_hclust),
    horiz = TRUE,
    xlab = "míra nepodobnosti\n/ 100 000",
    ylab = "model automobilu",
    xaxt = "n",
    axes = FALSE,
    xlim = c(2e5, 0),
    ylim = c(128, 135)
)

axis(side = 1, at = seq(0, 8e5, by = 2e5), labels = seq(0, 8, by = 2))

dev.off()


#### ------------------------------------------------------------------------

setwd(mother_working_directory)


## ------------------------------------------------------------------------

################################################################################
################################################################################
################################################################################
```

4 Reference

[1] R CORE TEAM. *R: A Language and Environment for Statistical Computing* [online]. Vienna, Austria: R Foundation for Statistical Computing, 2016. Dostupné z: https://www.R-project.org/

[2] CHAMBERS, J. M., W. S. CLEVELAND, B. KLEINER a P. A. TUKEY. *Graphical Methods for Data Analysis*. B.m.: Wadsworth & Brooks/Cole, 1983.

[3] HARTIGAN, J. A. *Clustering Algorithms*. New York: Wiley, 1975.

[4] WARD JR, J.H. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*. 1963, **58**(301), 236–244. ISSN 0162-1459.

[5] MURTAGH, Fionn a Pierre LEGENDRE. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *J. Classif.* [online]. 2014, **31**(3), 274–295. ISSN 0176-4268. Dostupné z: doi:10.1007/s00357-014-9161-z