Práce s datovými tabulkami v R

Datová tabulka (*data.frame*) je dvojrozměrným polem, které obsahuje v každém sloupci prvky stejného datového typu, ale jednotlivé sloupce se mohou datovým typem lišit. Všechny sloupce datové tabulky mají shodnou délku a samozřejmě i všechny řádky datové tabulky mají shodnou délku.

Vytvoření a úprava datové tabulky

Datovou tabulku můžeme vytvořit příkazem data.frame(), kde každý sloupec tvoří samostatný vektor (nebo faktor). Dávejte pozor, každý textový vektor je automaticky převeden na faktor, tedy hodnoty jsou převedeny do kategorií.

Úvodní informace k datovým tabulkám jsme ji uvedli v článku Datové struktury. Zde budeme pokračovat v složitějších příkladech.

```
# využijeme vestavěnou datovou tabulku "mtcars",
# jejíž data byla získána z amerického časopisu Motor Trend z roku 1974
# a která ukazuje výkon, spotřebu a další parametry 32 automobilů
data <- mtcars

# jaká jsou jména sloupců?
colnames(data)
#> [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
#> [11] "carb"

# změníme jména sloupců
# příkaz paste(…) vytváří vektor podle zadaných instrukcí,
# zde chceme mít vektor řetězců "c_1", "c_2", …
colnames(data) <- paste("c",
                        1:dim(data)[2],
                        sep = "_")

# jaká jsou jména sloupců teď?
colnames(data)
#> [1] "c_1"  "c_2"  "c_3"  "c_4"  "c_5"  "c_6"  "c_7"  "c_8"  "c_9"  "c_10"
#> [11] "c_11"

# změníme i jména řádků tabulky
rownames(data) <- paste("r", 1:dim(data)[1], sep = "_")

# náhledy
head(data)           # náhled na prvních 6 řádků
head(data, 10)       # náhled na prvních 10 řádků
tail(data)           # náhled na posledních 6 řádků
tail(data, 10)       # náhled na posledních 10 řádků

# přidáme řádek nul příkazem rbind()
# pozn.: pokud ale výsledek nepřiřadíme původní proměnné,
# tak proměnná data zůstane nezměněna
rbind(data, rep(0, dim(data)[2]))         # přidání řádku c(0, 0, ..., 0) k data.framu "data"

# přidáme sloupec nul příkazem cbind()
cbind(data, rep(0, dim(data)[1]))         # přidání sloupce c(0, 0, ..., 0) k data.framu "data"

# přidáme sloupec příkazem data.frame()
data.frame(data, "ahoj" = rep(0, dim(data)[1]))   # přidání sloupce c(0, 0, ..., 0) se jménem "ahoj" k data.framu "data"

# odebírání prvků
data[-1, ]   # odebrání 1. řádku data.framu "data"
data[, -1]   # odebrání 1. sloupce data.framu "data"
```

Indexování a adresace

Jak se odkážeme na nějaký konkrétní prvek datové tabulky, případně na konkrétní oblast? Je to podobné jako u vektorů: většinu práce zvládnou hranaté závorky za jménem proměnné.

```
# prvek na druhém řádku ve třetím sloupci
mtcars[2, 3]

# prvek na druhém řádku ve třetím sloupci: zde využíváme popisky
mtcars["Mazda RX4", "mpg"]
#> [1] 21

# celý první řádek s popisky (ve formě vektoru)
mtcars[1, ]
#>           mpg cyl disp  hp drat   wt  qsec vs am gear carb
#> Mazda RX4  21   6  160 110  3.9 2.62 16.46  0  1    4    4

# celý první řádek s popisky (ve formě vektoru) - přehledněji
# chci vypsat tu mazdu!
mtcars["Mazda RX4", ]

# celý druhý sloupec (ve formě vektoru) bez popisků
mtcars[, 2]

# sloupec s koňskými silami pomocí znaku dolaru, bez popisků
mtcars$hp
#> [1] 110 110  93 110 175 105 245  62  95 123 123 180 180 180 205 215 230  66  52
#> [20]  65  97 150 150 245 175  66  91 113 264 175 335 109

# bez popisků jsou ty sloupce nepřehledné, chci tabulku s popisky
```

```r
mtcars["hp"]

# druhý prvek sloupce s koňskými silami
mtcars$hp[2]
#> [1] 110

# prvek úplně vpravo dole, neznám-li rozměry tabulky
mtcars[dim(data)[1], dim(data)[2]]
```

Sloupcové přehledy

Někdy chceme nad všemi sloupci provést nějakou analýzu, tj. zjistit agregovaný ukazatel nad všemi sloupci.

```r
# součet jednotlivých sloupců
colSums(mtcars)
#>      mpg      cyl     disp       hp     drat       wt     qsec       vs
#>  642.900  198.000 7383.100 4694.000  115.090  102.952  571.160   14.000
#>       am     gear     carb
#>   13.000  118.000   90.000

# totéž jinak: do příkazu apply() zadám datovou tabulku,
# 2 značí sloupce (1 jsou řádky, c(1, 2) jsou řádky i sloupce),
# sum je daná funkce, kterou chci nad daty provést
apply(mtcars, 2, sum)

# průměry jednotlivých sloupců
colMeans(mtcars)

# také průměry jinak, viz poznámky k apply() výše
apply(data, 2, mean)

# ukážeme si, jak pracovat s neúplnou tabulkou,
# ve které nějaká data chybí

# přidáme řádek s NA hodnotami
mtcars <- rbind(mtcars, rep(NA, dim(mtcars)[2]))

# průměr sloupců
colMeans(mtcars)
#> mpg  cyl disp   hp drat   wt qsec   vs   am gear carb
#>  NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA

# evidentně to nefunguje, musím nastavit, aby se výpočet vyhnul
# neuvedeným hodnotám: na.rm = TRUE
colMeans(mtcars, na.rm = TRUE)

# totéž pomocí apply()
apply(data, 2, mean, na.rm = TRUE)
```