MUNI SCI

M9DM2 Data mining 2

Práce s datovými tabulkami v *R*

Ondřej Pokora

Ústav matematiky a statistiky, Přírodovědecká fakulta, Masarykova univerzita

1. 10. 2019

Strategie analýzy dat *split — apply — combine*

1. rozdělení datové tabulky do více homogenních částí
2. použití nějaké funkce (např. výpočet modelu, statistik) na každou část
3. sloučení dílčích výsledků do jedné tabulky

Typické příklady:

- výpočet stejného modelu pro podmnožiny datové tabulky
- výpočty sumárních statistik pro jednotlivé podskupiny dat
- skupinové transformace (např. škálování, standardizace)

Organizace dat v datových tabulkách

1. formát ***wide***
   - horizontální struktura
   - každá veličina má vlastní sloupec
   - kategoriální data jsou seskupená
   - v praxi běžný způsob zaznamenávání dat
2. formát ***long***
   - vertikální struktura
   - sloupec pro všechny typy proměnných a sloupec pro hodnoty těchto veličin
   - každý řádek odpovídá jednomu pozorávní jednotlivých kategorií
   - formát vhodný pro počítačové zpracování
   - nutný pro složitější výpočty (typicky opakovaná pozorování) a vizualizace

Změna formátu

Knihovna `tidyr`:

- *wide* $\longrightarrow$ *long*: `gather`
- *long* $\longrightarrow$ *wide*: `spread`
- `separate`

Knihovna `reshape2`:

- *wide* $\longrightarrow$ *long*: `melt`
- *long* $\longrightarrow$ *wide*: `dcast`

`dplyr`: moderní knihovna pro práci s datovými tabulkami, rozšíření knihovny `plyr`

Knihovna `dplyr`

- doplňuje datový typ `data.frame` o třídu `tbl_df`
- Striktní organizace dat: měření v řádcích, proměnné ve sloupcích
- `%>%`: operátor řetězení *(pipe)*
- `glimpse`: lepší varianta `str`
- `select`, `pull`: výběr proměnných (sloupců)
- `filter`: výběr řádků podle logických podmínek
- `slice`: výběr řádků podle indexů
- `arrange`: seřazení tabulky podle zadaných proměnných
- `mutate`: změna nebo přidání nových proměnných pomocí operací se stávajícími proměnnými
- `mutate_at`: změna zvolených proměnných
- `transmute`: jako `mutate`, ale odstraní nepotřebné sloupce
- `summarise`: výpočet agregovaných proměnných
- `summarise_at`: výpočet agregovaných proměnných pro zvolené proměnné
- `group_by`: agregování do skupin podle zadaných proměnných
- `bind_rows`, `bind_cols`: spojování tabulek po řádcích nebo sloupcích