Základy spracovania a vizualizácie dát

18. marca 2024

rozcvička

- v budúcnosti budem pri inom vypracovanom zadaní ako ste dostali dávať 0 (netýka sa tých, ktorým sa nenačítalo zadanie)
- posledný týždeň semestra bude opravná/náhradná písomka
- hoci je písomka openbook, je vhodné si prejsť ešte raz doma to, čo sa robilo na hodine, rozcvičky môžu byť časovo náročné, ak základné veci z hodiny nie sú zautomatizované

užitočné tipy

- pozor na správne načítanie dát (problém bol . vs , v desatinnom čísle), možnosti:
  - zmeniť nastavenie E
  - načítať ako text (pri dátume už často niet jednoduchej pomoci) a v E prepísať . na ,
  - zmeniť . na , v pôvodnom textovom súbore
  - inak

užitočné tipy

- prepísanie . na ,:
  - natvrdo (pri malom počte dát a dostatku času)
  - vyhľadávanie a nahradenie: Ctrl+F, Find, Replace, Replace All - užitočné aj pri programovaní

užitočné tipy

- pri kopírovaní v E sa dá kopírovať:
  - celý súbor buniek tak ako sú aj s formátovaním
  - iba hodnoty (Paste Values)
  - iba formát
  - funkcie
  - kombinácie

užitočné tipy

- pozor na filter a zobrazovanie - v tabuľke je filter iba zobrazovacia záležitosť, tabuľka sa tým nezmení/neprepíše, pri ďalšom spracovávaní vyfiltrovaných dát je potrebné ich prekopírovať alebo inak oslobodiť

užitočné tipy

- kontrolovať, kontrolovať, kontrolovať - ak niečo vychádza zvláštne, neočakávane, je vysoká pravdepodobnosť, že je niekde chyba napr.:
  - hodnoty gini indexu sa v tabuľke hýbu okolo 30-40 a priemer vyjde 60
  - medián a druhý kvartil vyjdú inak
  - boxplot ukazuje jedného outliera, výpočtom vyšlo nula
  - vo výsledkoch z Descriptive statistics z Data Analysis je $Count = 1345$ a našich dát je $n = 75$

grafy

- rýchly prenos informácie
- nezabudnúť aj na vizuálnu stránku

bar/column plot

- najčastejšie pre kategoriálne dáta
- grafy početností (percent)
- spolupracujú s pivotkami
- *Department* (*HRdata*)
- úprava textu

X Y/scatter plot

- dve číselné premenné
- sledovanie závislosti
- *Monthly Income* vs *Total Working Years*
- zmeniť popisy osí
- trendline

line chart

- časové rady - vykreslenie trendu v čase
- spurious correlations

pie chart

- často pre kategoriálne dáta s malým počtom kategórií
- pri kategoriálnych dátach niekedy treba spraviť najskôr pivotku
- 3D vie zavádzať
- *Marital Status*

histogram

- histogram = graf, ktorý vykresľuje intervalovú frekvenčnú tabuľku
- column chart je podobný
- jeden z najpoužívanejších grafov (vizuálne počiatočné zhodnotenie dát)
- porovnanie s teoretickými rozdeleniami
- *Hourly Rate*

boxplot

- zobrazuje veľa číselných charakteristík v jednom grafe
- minimum, maximum, kvartily, priemer, outlierov $(q_{0,25} - 1{,}5IQR; q_{0,75} + 1{,}5IQR)$, niekedy aj rozptyl
- pozor, *QUARTILE.EXC()*
- porovnanie mužov a žien *Hourly Rate*

iné

- area, surface, ...
- špeciálne typy dát

samostatné cvičenie

- pomocou boxplotov porovnajte *Age* pre skupiny v *BusinessTravel*
- histogramom zobrazte *MonthlyIncome* iba pre ženy
- využite stĺpcový/čiarový diagram pre porovnanie *OverTime* medzi mužmi a ženami