Explorační analýza dat v R

Explorační analýzu dat (*exploratory data analysis*, *EDA*) můžeme chápat jako aplikovanou popisnou statistiku: cílem je zkrátka prozkoumat získaná data, odfiltrovat chyby a chybějící data, zobrazit několik základních grafů a získat přehled, s jakým datasetem vlastně pracuji. To pak usnadní rozmýšlení, jakou statistickou metodu a jaké analýzy mohu nad daty provádět.

V tomto článku se pokusíme ukázat základní nástroje explorační analýzy v jazyce a prostředí R.

Pro správnou funkci kódu v tomto článku bude potřeba nainstalovat balíček *lattice*:

```
# balicky pro tuto kapitolu
library(lattice)
```

Čísla

Základním nástrojem může být např. zobrazení minima, prvního a třetího kvartilu, mediánu, průměru a maxima.

```
# vyrobim si nahodnou velicinu
a <- rnorm(100)

# a vypisu souhrn
summary(a)

#>    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
#> -2.65361 -0.62364  0.06343  0.05238  0.75077  2.60554
```

Podobného výsledku bez průměru můžeme dosáhnout i jinou funkcí.

```
# stejna velicina, jiny souhrn
fivenum(a)

#> [1] -2.65361146 -0.63345254  0.06343441  0.75899874  2.60553666
```

Grafy

Stem and leaf

Zajímavým přechodem mezi čísly (textem) a grafy je *stem-and-leaf diagram*. Ve své jednoduché podobě ho vídáme na zastávkách hromadné dopravy. Vlevo od vertikály je "kořen" a vpravo je "větev" nebo "list". U jízdních řádů jsou vlevo hodiny a vpravo minuty, každý autobus je zmíněn samostatně (i kdyby jely dva v jeden čas). Stejné je to u jiných diagramů, vzniká vlastně jakýsi textový histogram převrácený na bok.

```
# vykon automobilu v konich
stem(mtcars$hp)

#>  The decimal point is 2 digit(s) to the right of the |
#>
#>  0 | 5677799
#>  1 | 0011111122
#>  1 | 55888888
#>  2 | 123
#>  2 | 556
#>  3 | 4
```

Histogram

Zde je *histogram* se stejnou veličinou jako u předchozího zobrazení.

```
hist(x    = mtcars$hp,
     main = "",
     xlab = "Vykon vozu v hp")
```

Krabicový graf

Krabicový graf také ukazuje minimum, maximum, první a třetí kvantil a průměr. Můžeme si nechat zobrazit i odlehlé hodnoty.

```
# pouziju vygenerovanou nahodnou velicinu z prvniho prikladu
boxplot(x    = a,
        xlab = "Velicina 'a'",
        ylab = "Hodnota")
```

Sloupcový graf

Sloupcový graf ukazuje složení souboru pomocí výšky sloupců.

```
barplot(height = table(mtcars$cyl),
        xlab   = "Pocet valcu v motoru",
        ylab   = "Cetnost")
```

Mozaikový graf

Mozaikový graf ukazuje vztah mezi vícero kvalitativními proměnnými.

```
mosaicplot(x    = apply(HairEyeColor, c(1, 2), sum),
           main = "Vztah mezi barvou oci a barvou vlasu")
```

Lattice

Dobré možnosti nabízí balíček *lattice*, který obsahuje funkce tvořící pokročilé grafy.

```
lattice::dotplot(weight ~ feed,
                 data = chickwts)
```

```
lattice::bwplot(weight ~ feed,
                data = chickwts)
```

```
lattice::xyplot(Petal.Length ~ Petal.Width | Species,
                data = iris)
```

Příjemná je také možnost barvení podle kategorií.

```
lattice::xyplot(Petal.Length ~ Petal.Width,
                data  = iris,
                group = Species)
```

Odkazy

Použitá literatura

- OLDŘICH, Neubauer. *Základy statistiky.* - vydání. Grada Publishing a.s., 2012. 236 s. ISBN 9788024742731.
- KERNS, G Jay. *Introduction to Probability and Statistics Using R.* 1. vydání. IPSUR, 2018. ISBN 978-1726343909.

Použité balíčky R

- SARKAR, Deepayan. *Lattice: Multivariate Data Visualization with R.* 1. vydání. New York : Springer, 2008. ISBN 978-0-387-75968-5.