Jazyk R

II. Dáta, Grafy, Balíčky, Štatistika

IT ACADEMY

Lektor: PhDr. Ing. Mgr. Miroslav Reiter, DiS., DBA, MPA
Kontakt: miroslav.reiter@it-academy.sk

Ako Začneme?

1. **Registrácia na Jetbrains Datalore +**

Stiahni si **R** a **Anaconda Navigator**

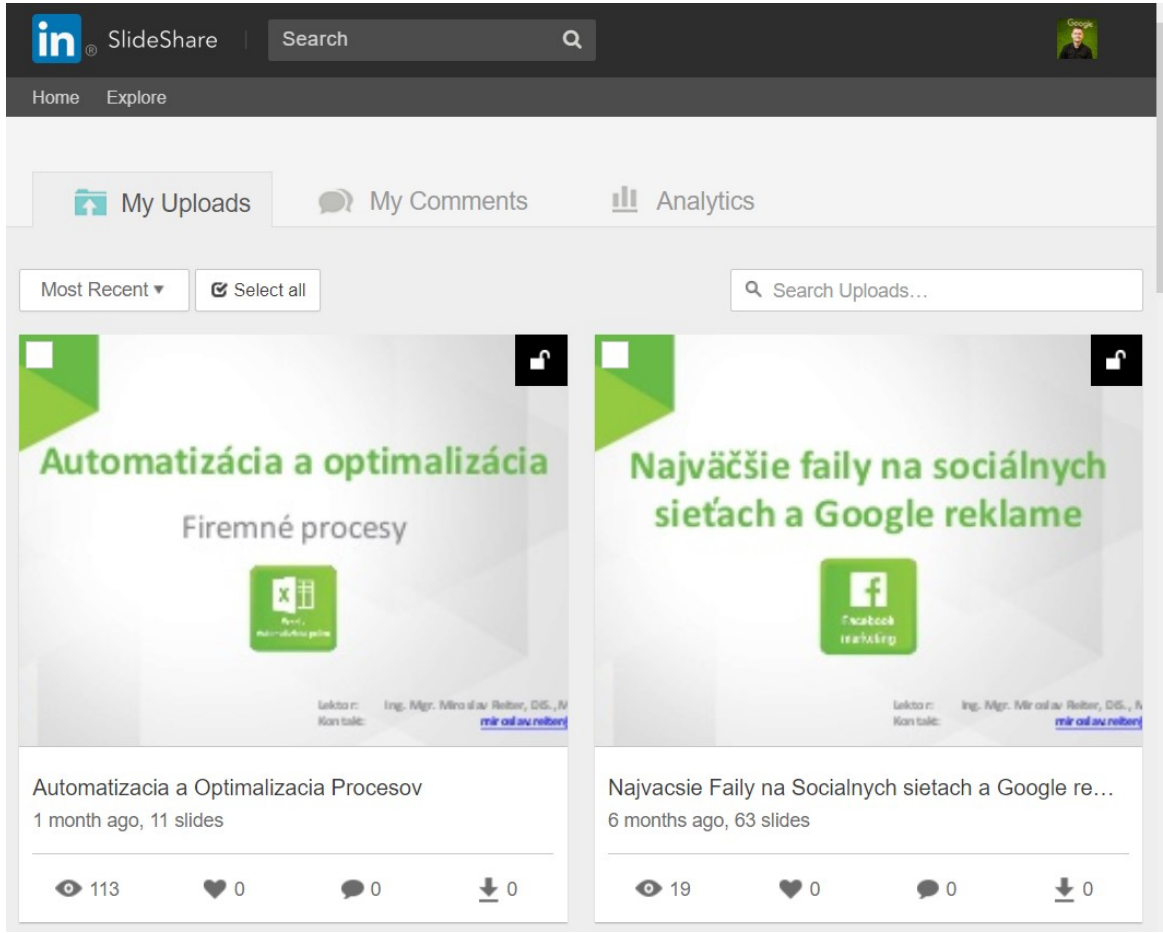
- https://datalore.jetbrains.com/
- https://cran.r-project.org/bin/windows/base/
- https://www.anaconda.com/products/

2. **Pridajte** si ma na **LinkedIn**

- www.linkedin.com/in/miroslav-reiter

3. **Prezentácia** a **materiály** po prednáške

IT ACADEMY



Štatistika v R - Spracovanie a vizualizácia dát

Kurz sa zameriava na pokročilejšie techniky práce s prostredím pre štatistickú analýzu R. Nosnou témou kurzu bude spracovanie a vizualizácia dát.

1. **Spracovanie a vizualizácia dát v R:**
   - CSV a Excel súbory
   - Koláčové grafy
   - Stĺpcové grafy
   - Krabicové grafy
   - Histogramy
   - Čiarové grafy
   - Bodové grafy

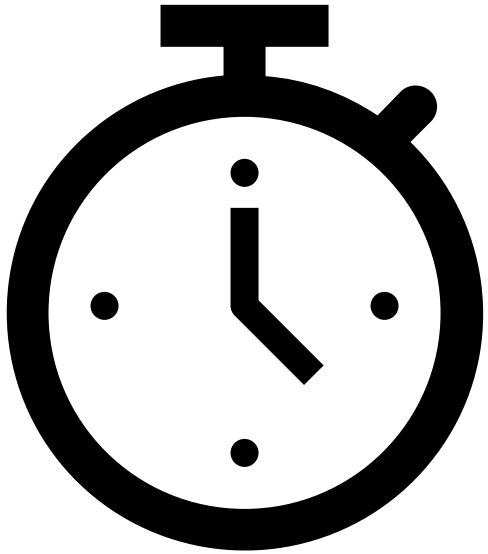
2. **Dátové štruktúry a programovanie**
   - Matice a Submatice
   - Polia
   - Faktory
   - Dátové rámce/Tabuľky
   - Balíčky
   - GUI

3. **Štatistická analýza**
   - Korelačná analýza
   - Regresná analýza
   - Analýza kontingenčných tabuliek
   - Časové rady

Úvodné informácie

- Časový rozvrh (9:00-13:30)
- Prestávky
- Mobilné telefóny a zariadenia

- Priprav si otázky a rovno sa pýtaj
- Interaktívna forma



O lektorovi - Miroslav Reiter

10000+ klientov a 500+ firiem

Programátor
Analytik
Manažér

Google
Microsoft
ISTQB tréner

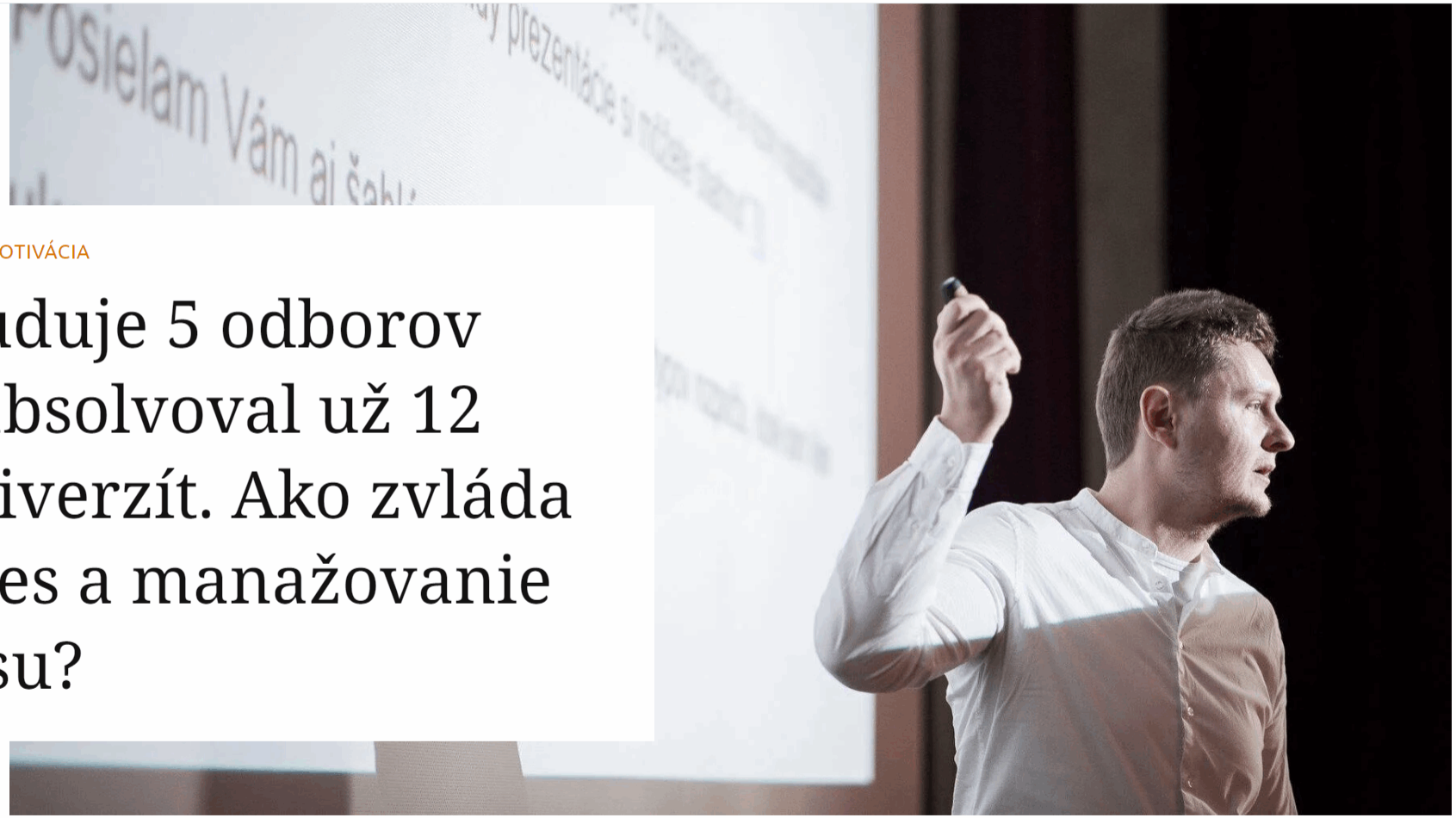
115 certifikácii

83 príručiek a publikácii

13 škôl

50+ projektov

Vlastná firma



MOTIVÁCIA

Študuje 5 odborov a absolvoval už 12 univerzít. Ako zvláda stres a manažovanie času?

Foto: Jakub Kovalík pre FMK UCM | Miroslav Reiter na prednáške Grow with Google na FMK UCM.

Nikola Kotláriková

19. júl 2022 · 8 min. čítania

Miroslav Reiter

1. PhDr. VŠM (Podnikovný manažment)
2. **Ing. STU FEI (Aplikovaná informatika)**
3. **Mgr. UK FM (Strategický manažment a marketing)**
4. **Mgr. VŠM (Manažérstvo kvality)**
5. Mgr. VŠEMVŠ (Verejná správa)
6. Mgr. DTI (Učiteľstvo ekonomických predmetov)
7. DiS. AMOS (Cestovný ruch)
8. MBA LIGS (Executive management)
9. **DBA Humanum (IT manažment)**
10. MPA IES (Verejná správa a samospráva krajov)
11. **MSc. Humanum (Bezpečnosť informačných systémov)**
12. Ing. Paed. IGIP
13. **Mgr. PEVŠ (Bezpečnosť informačných systémov)**

DIGITÁLNA KOALÍCIA.SK

SLOVENSKÁ AKADÉMIA VIED ACADEMIA SCIENTIARUM SLOVACA

IT ACADEMY

DIGITÁLNA UNIVERZITA

STU FIIT

FAKULTA MANAGEMENTU
Univerzita Komenského
v Bratislave

VITA

Vyberte si online kurz

Naučte sa programovať, tvoriť webstránky a grafiku, manažovať alebo sa zamerajte na osobný rozvoj. Všetko jednoducho vďaka našim online kurzom z pohodlia domova.

Ročné predplatné na všetky online kurzy

~~2299.99€~~

399.99€

Prístup pre Vás do všetkých aktuálnych aj pripravovaných online kurzov

12 mesačná platnosť

Kúpiť teraz

407 kurzov v ponuke

Zábavné online lekcie

Akreditované kurzy

11 rokov skúseností

Certifikovaní profesionálni lektori

Odporúčame Kurzy špeciálne pre vás

Online kurz SAP I. Začiatočník
224,00€ ~~292,50€~~

Online kurz Java I. Začiatočník
67,00€ ~~88,40€~~

Online kurz PRINCE2 Foundation
224,00€ ~~308,70€~~

Online kurz Lektor (Akreditovaný Kurz Lektor)
127,00€ ~~167,50€~~

Online kurz Copywriting I. Začiatočník
50,00€ ~~66,30€~~

Online kurz Testovanie Softvéru I. Začiatočník
140,00€ ~~183,30€~~

Moje začiatky

Miroslav Reiter

získava status

Google Certified Trainer

Automation

Google

Luigi, Mário a Yoshi

Čo robíte?

1. Študent/učiteľ
2. Zamestnanec
3. Podnikateľ
4. Nezamestnaný/materská
5. Dievča pre všetko

National competence centre for high performance computing SLOVAKIA

EURO

EuroHPC
Joint Undertaking

NATIONAL
SUPERCOMPUTING
CENTRE

Národné kompetenčné centrum pre vysokovýkonné počítanie
Národné superpočítačové centrum
Projekt EuroCC2

Vzdelávanie

Kurzy:
itkurzy.sav.sk

Propagácia

Prednášky:
https://eurocc.nscc.sk
/vzdelavanie/prednasky
/prednasky-archiv/

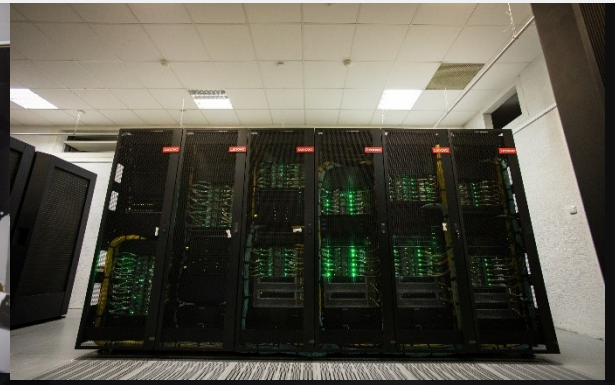
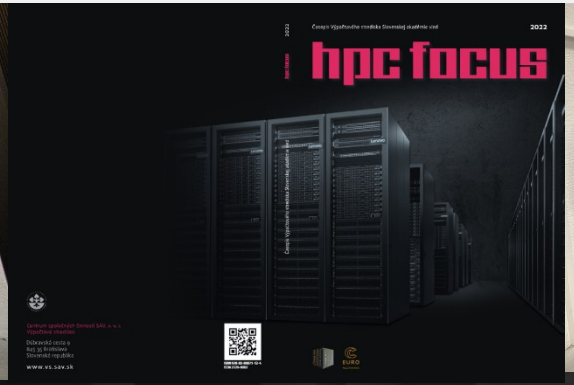
HPC služby

Prístup k výpočtovým prostriedkom

Spolupráca

Pilotné projekty

Dlhodobá spolupráca

robíme it

Qubit Conference

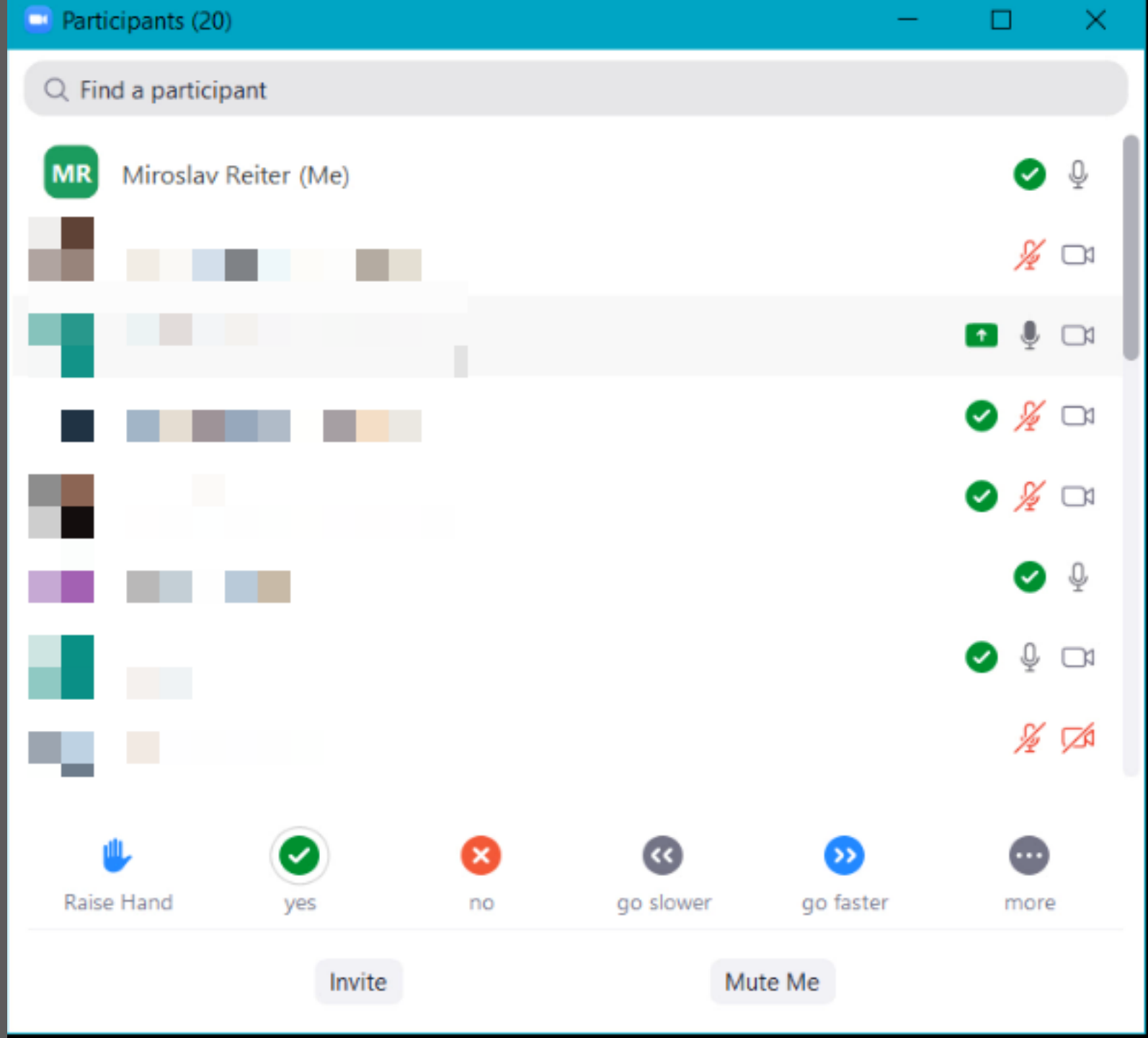
S kým spolupracujeme:

- Akademické inštitúcie, univerzity, ústavy SAV...
- Verejná správa
- Súkromné firmy
- Tretí sektor

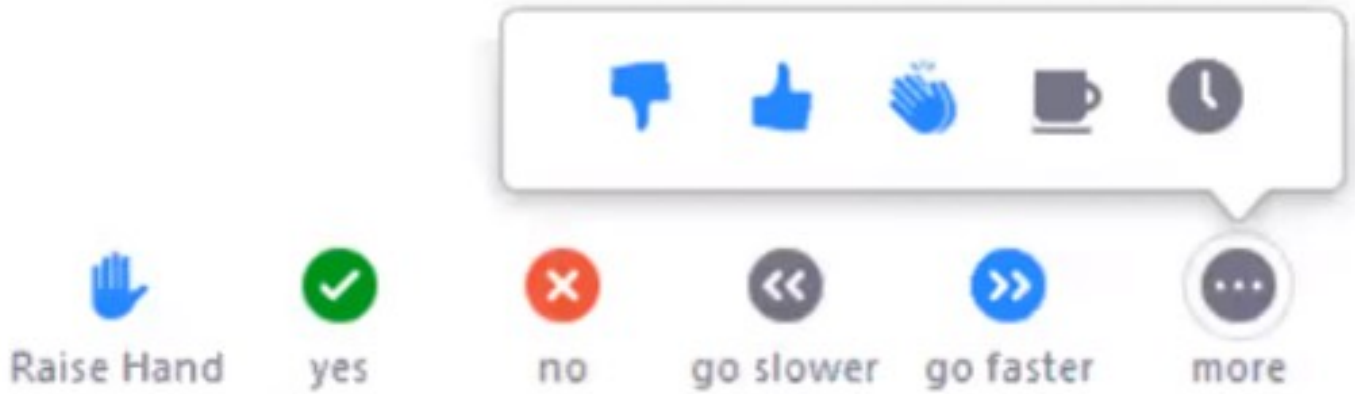
Sledujte nás na sociálnych sieťach:
#nccprehpc

Interaktívna prednáška

Aktívne používanie a zapájanie sa



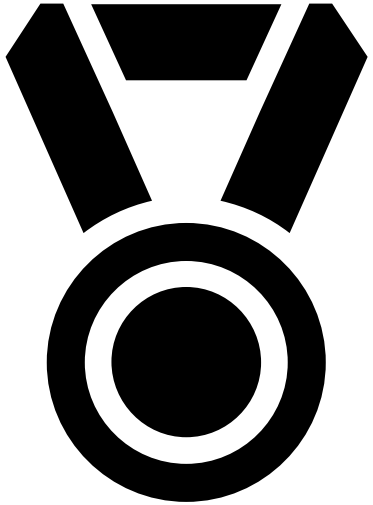
Používame Zoom



Vaše ciele a očakávania

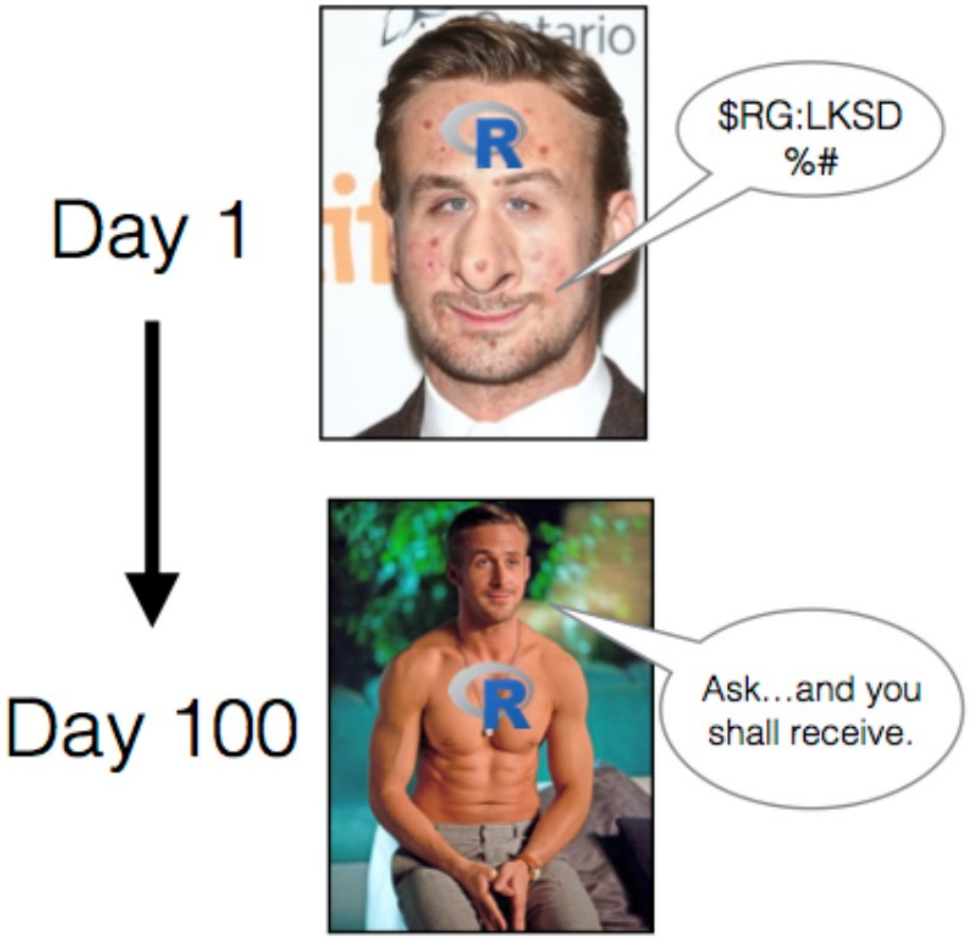
1. Základy jazyka R a Data Science
2. Základy analytického/štatistického myslenia
3. Základy programovania v R
4. Základy s knižnicami/balíčkami
5. Základy práce s vývojovými nástrojmi pre jazyk R

Zábava je v zaručená v každom bode :-)



Data Science a Jazyk R

Začiatky v R a Data Science



R a Data Science je ako vzťah...

ZVR_842 · UPDATED 9 DAYS AGO

26 | New Notebook | Download (15 kB)

Global pollution by counties

Worldwide waste by countries (inc. all types of pollution by 2010-2020)

Data | Code (1) | Discussion (0)

About Dataset

aggregate data on solid waste management from around the world. This database features the statistics collected through the effort, covering nearly all countries and over 330 cities. The metrics included cover all steps from the waste management value chain, including waste generation, composition, collection, and disposal, as well as information on user...

Classification | Data Analytics | Exploratory Data Analysis | Energy | Advanced | Global

country_level_data_0.csv (46.19 kB)

Detail | Compact | Column | 10 of 49 columns

About this file

there are 49 columns, be acccurate with columns name because it is divided by types of wasting

region_id	country_name	gdp	composition_foo...	composition_glas...	compo

Usability
9.71

License
CC BY-SA 4.0

Expected update frequency
Annually

Data Explorer
46.19 kB

country_level_data_0.csv

Vyhľadať dataset

Doprava
148 datasetov

Ekonomika a práca
37 datasetov

Infraštruktúra, výstavba a bývanie
46 datasetov

Kultúra, šport, cestovný ruch a Wifi
71 datasetov

Obyvateľstvo
123 datasetov

Politika a voľby
156 datasetov

Priestorové údaje
82 datasetov

Rozpočet, dane a zmluvy
40 datasetov

Sociálna oblasť

Veda a vzdelávanie

Zákony a spravodlivosť

Zdravie, životné prostredie a klíma

R Graphics Cookbook, 2nd edition

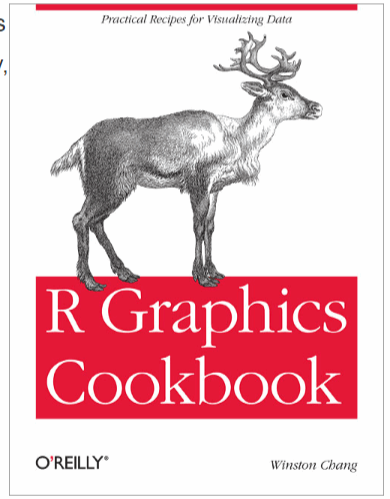
Winston Chang

2022-11-23

Welcome

Welcome to the **R Graphics Cookbook**, a practical guide that provides more than 150 recipes to help you generate high-quality graphs quickly, without having to comb through all the details of R's graphing systems. Each recipe tackles a specific problem with a solution you can apply to your own project, and includes a discussion of how and why the recipe works.

Read online here for free, or buy a physical copy on Amazon.



YaRrr! The Pirate's Guide to R

Nathaniel D. Phillips

2018-01-22

Chapter 1 Preface

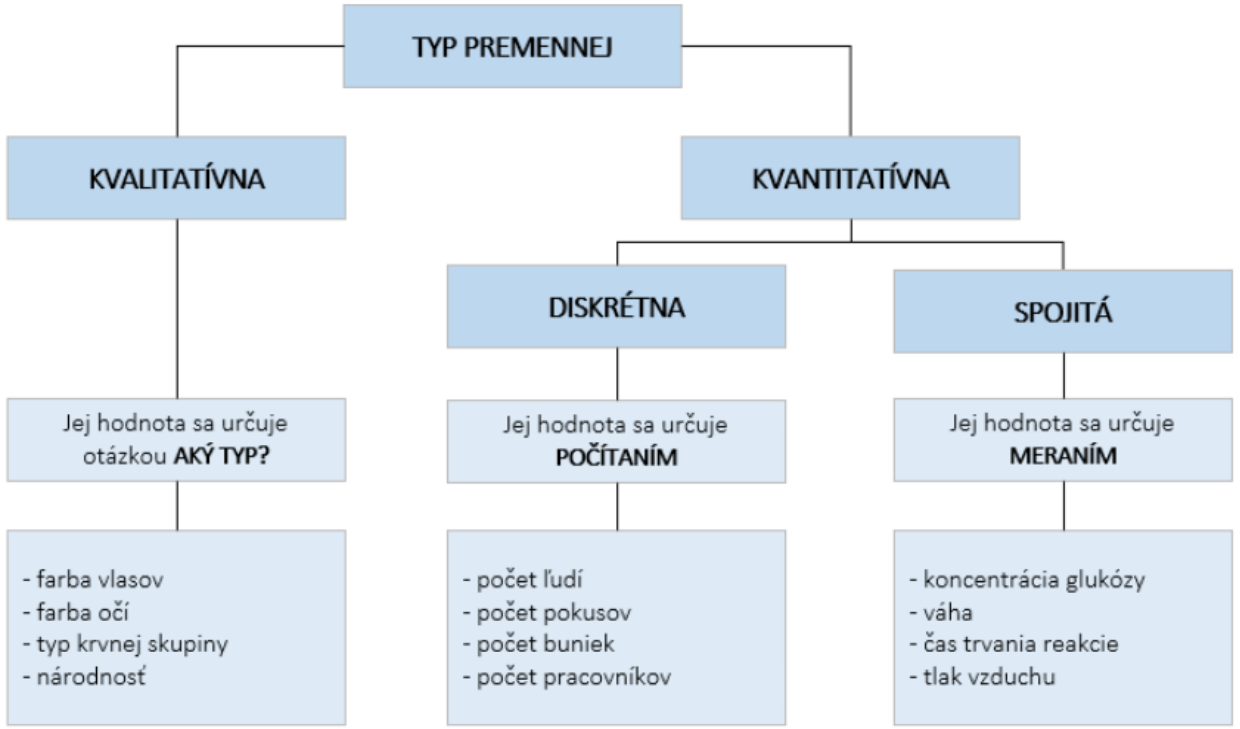
Štatistické Znaky – Premenné (Variables)

Slovné – Kvalitatívne

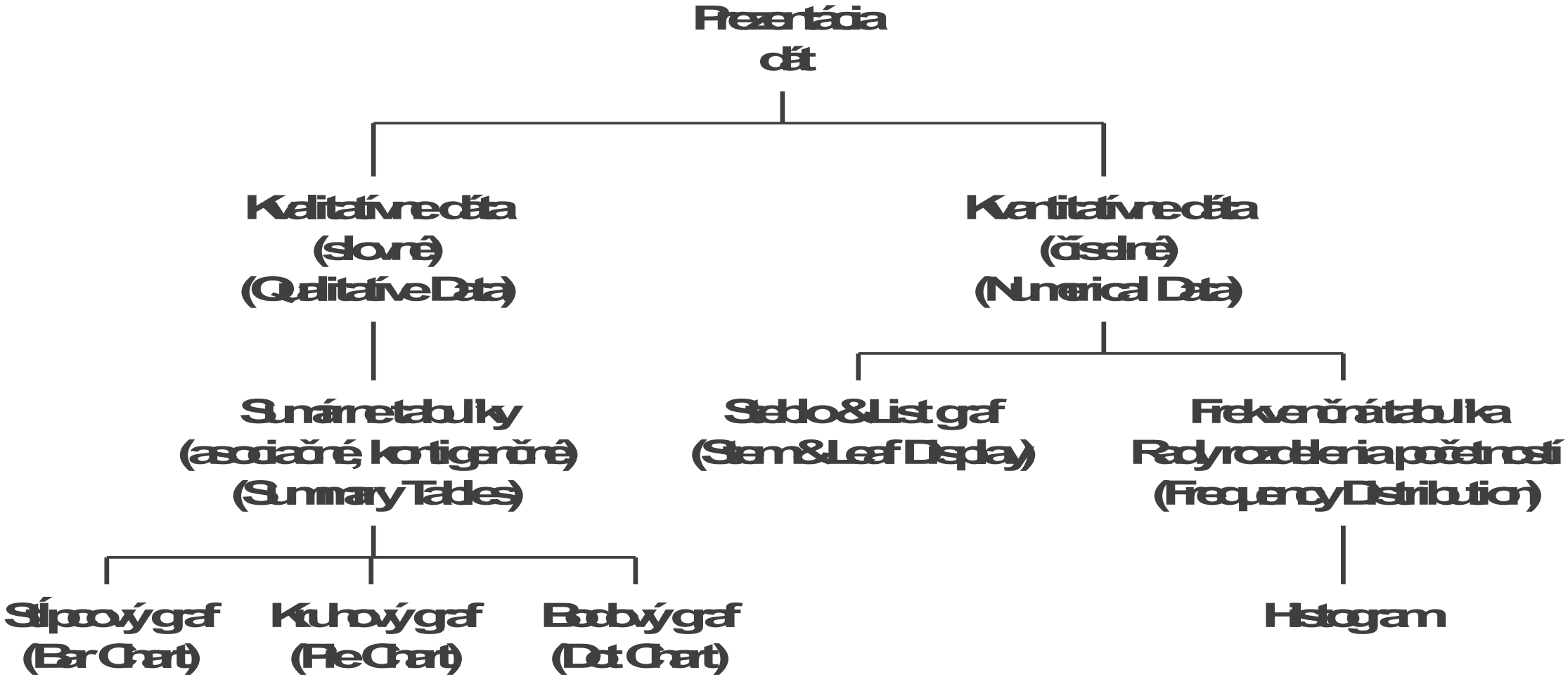
- Označenie: A, B, C, ...

Číselné – Kvantitatívne

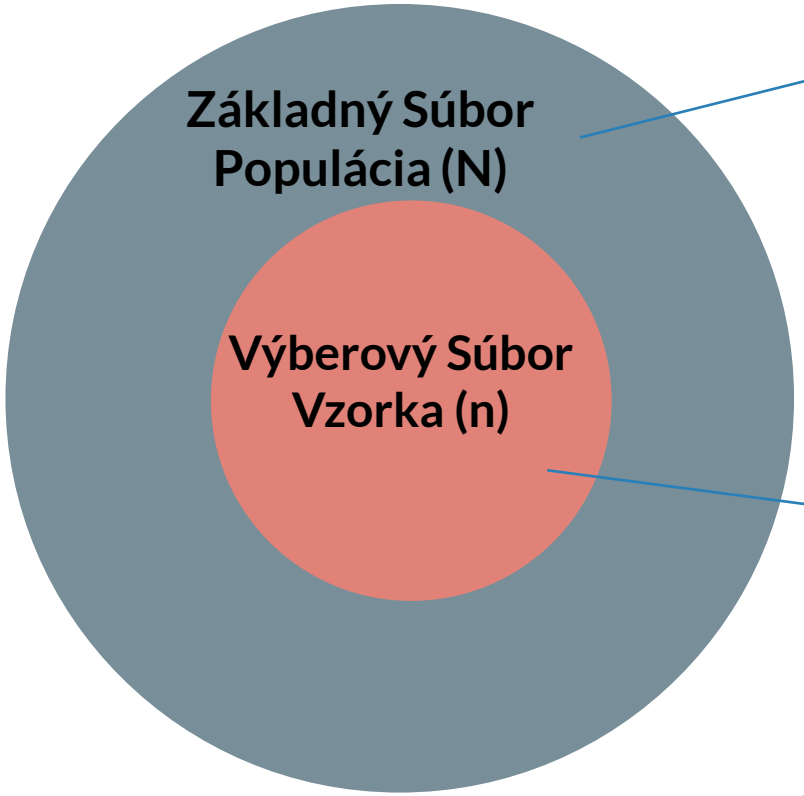
- Označenie: X, Y, Z, ...
- Diskrétne a Spojité



Spôsoby Vizualizácie (Prezentácie) Dát



Popisná (Deskriptívna) Štatistika



Vlastnosti Populácie

Matematická štatistika

Priemer, Rozptyl (Variancia),
Štandardná (Smerodajná)
Odchýlka, Modus...

Štatistické metódy

Parametrické

Neparametrické

Ukazovateľ Measure	Populácia Population	Výber Sample
Rozsah (Size)	N	n
Priemer (Mean)	μ	$\overline{x}$
Rozptyl (Variance)	σ^2	s^2
Štand.oddchýlka (Stand.Deviation)	σ	s

Priemer, Medián, Modus

MEAN GIRLS

$$\overline{GIRLS} = \frac{\sum_{i=1}^{n} GIRLS_i}{n}$$

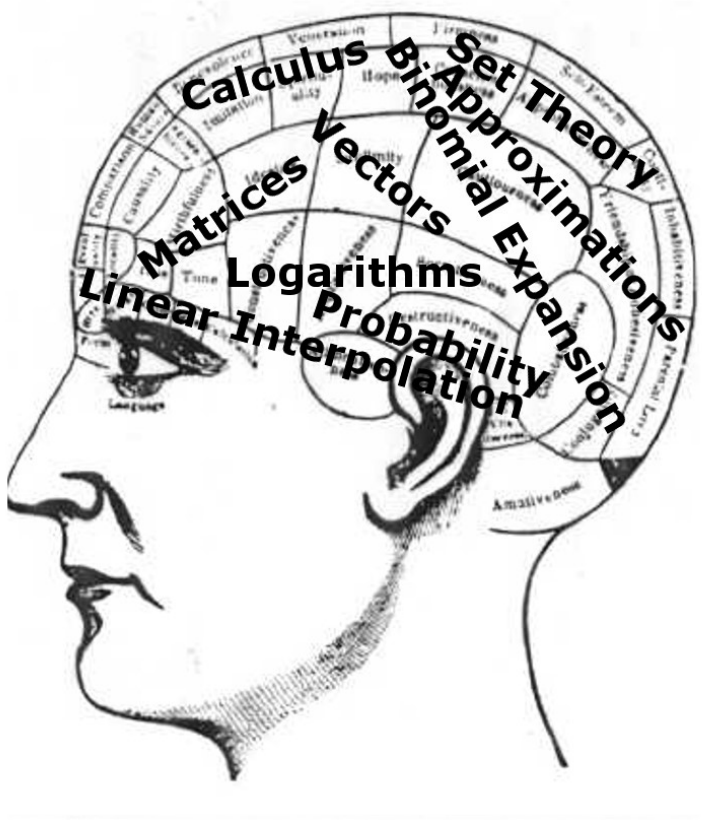
MEDIAN GIRLS

$$B_m + I.\left(\frac{\frac{n}{2} - (\sum GIRLS_i)_0}{GIRLS_m}\right)$$

MODE GIRLS

$$l + h\left(\frac{GIRLS_m - GIRLS_1}{2GIRLS_m - GIRLS_1 - GIRLS_2}\right)$$

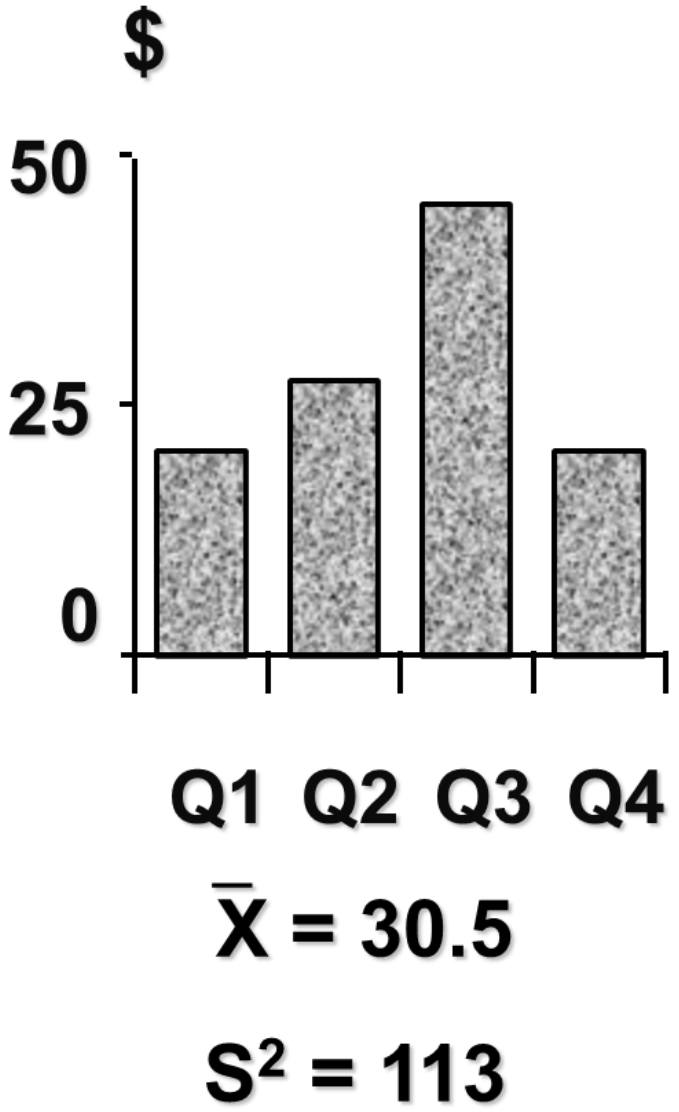
"Sometimes I wonder what goes on in that head of yours..."



Popisná (Deskriptívna) Štatistka

- **1. Obsah**
  - Zber dát
  - Prezentácia (Vizualizácia) dát
  - Charakter dát
- **2. Cieľ**
  - Popis dát

- **Spracovanie Dát**
  - **Kontrola dát** – Formálna, Vecná
  - **Triedenie** - usporiadanie dát do skupín (tried) podľa určitého štat. znaku (-ov) tak, aby čo najlepšie vynikli vlastnosti javu (-ov)
  - **Dátové Konverzie** (CSV, XLSX, XML, STATA, SAV)



MIERY

POLOHY

- Minimum
- Maximum
- Priemer
- Medián
- Modus
- Kvantily

VARIABILITY

- Štandardná odchylka
- Rozptyl
- Štandardná chyba
- Variačné rozpätie
- Medzikvartilové rozpätie
- Variačný koeficient

TVARU

- Šikmosť
- Špicatosť

MIERY

POLOHY

- Minimum
- Maximum
- Priemer
- Medián
- Modus
- Kvantily

x_{min}

x_{max}

$$r = \frac{n + 1}{2}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$$

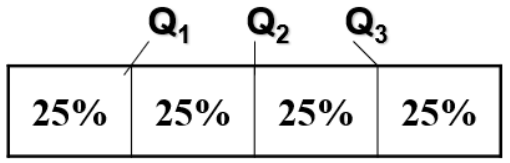
– Vážený (z frekv. tabuľky):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{m} x_i n_i = \frac{1}{n}(x_1 n_1 + x_2 n_2 + \ldots + x_m n_m)$$

Dáta: 10, 10, 11, 13, 9, 10, 10, 8
x_{Mo}=10 (1 modus)

Dáta: 10, 10, 11, 12, 12, 8, 9, 10, 12
x_{Mo}=10 a 12 (2 modusy)

Dáta: 10, 12, 8, 9, 11, 13, 7 - dáta bez modusu



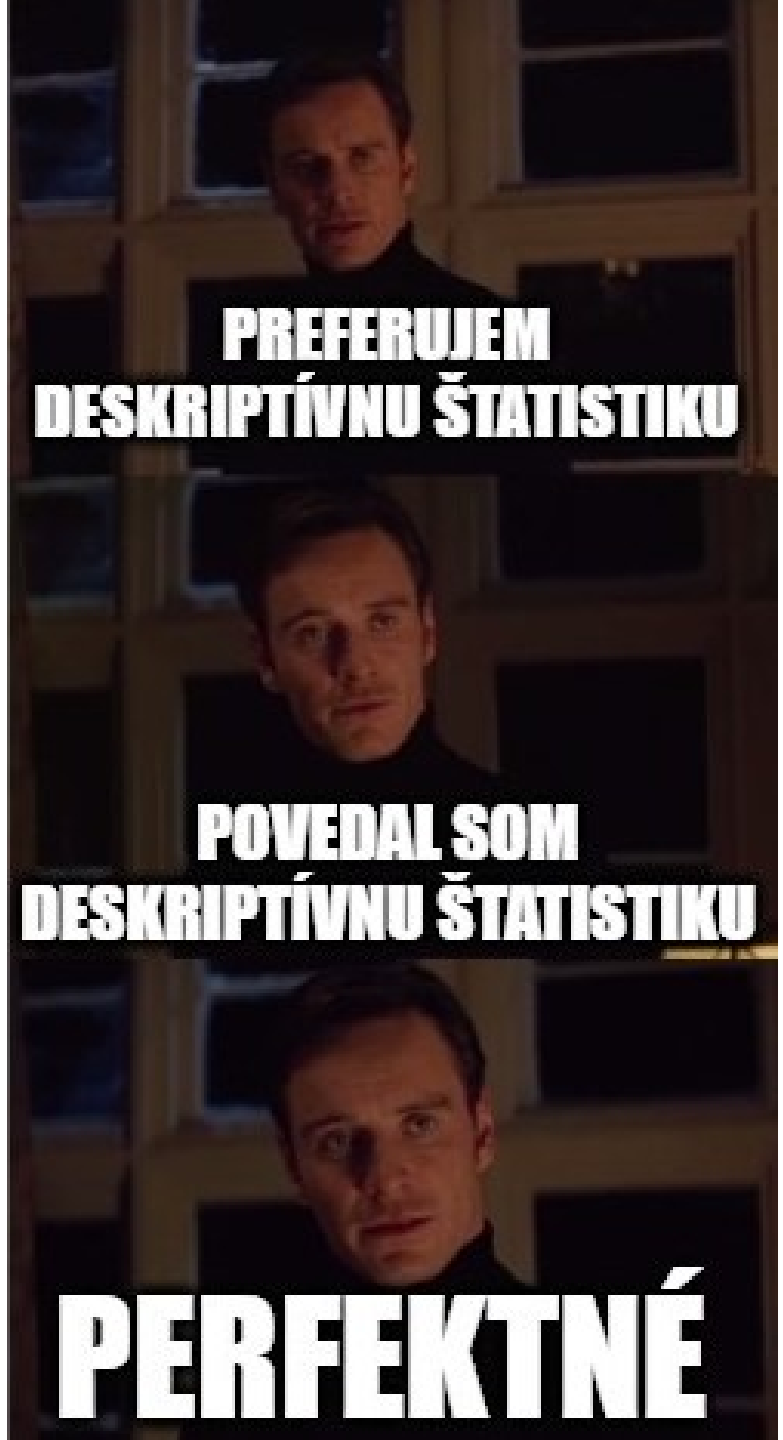
Pozícia i-ho kvartilu:

$$k_{Qi} = \frac{i(n+1)}{4}$$

	Murder	Assault
	Min. : 0.800	Min. : 45.0
	1st Qu.: 4.075	1st Qu.:109.0
	Median : 7.250	Median :159.0
	Mean : 7.788	Mean :170.8

```
USArrests

 4  Variables      50  Observations
---------------------------------------------------------------------------
Murder
       n  missing distinct     Info     Mean      Gmd      .05      .10
      50        0       43        1    7.788    5.022    2.145    2.560
     .25      .50      .75      .90      .95
   4.075    7.250   11.250   13.320   15.400

lowest : 0.8  2.1  2.2  2.6  2.7 , highest: 13.2 14.4 15.4 16.1 17.4
```

	Murder	Assault
nbr.val	50	50
nbr.null	0	0
nbr.na	0	0
min	0.8	45
max	17.4	337
range	16.6	292
sum	389.4	8538



Deskriptívna Štatistika

```
# get means for variables in data frame mydata
# excluding missing values
sapply(mydata, mean, na.rm=TRUE)
```

```
sapply(USArrests, mean)
Murder   Assault UrbanPop     Rape
 7.788   170.760   65.540   21.232
```

```
# mean,median,25th and 75th quartiles,min,max
summary(mydata)

# Tukey min,lower-hinge, median,upper-hinge,max
fivenum(x)
```

Deskriptívna Štatistika

```
> describe(USArrests)
USArrests

 4  Variables      50  Observations
---------------------------------------------------------------------------
Murder
       n  missing distinct     Info     Mean      Gmd      .05      .10
      50        0       43        1    7.788    5.022    2.145    2.560
     .25      .50      .75      .90      .95
   4.075    7.250   11.250   13.320   15.400

lowest :  0.8  2.1  2.2  2.6  2.7, highest: 13.2 14.4 15.4 16.1 17.4
---------------------------------------------------------------------------
Assault
       n  missing distinct     Info     Mean      Gmd      .05      .10
      50        0       45        1    170.8    96.44    50.25    56.90
     .25      .50      .75      .90      .95
  109.00   159.00   249.00   279.60   297.30

lowest :  45  46  48  53  56, highest: 285 294 300 335 337
```

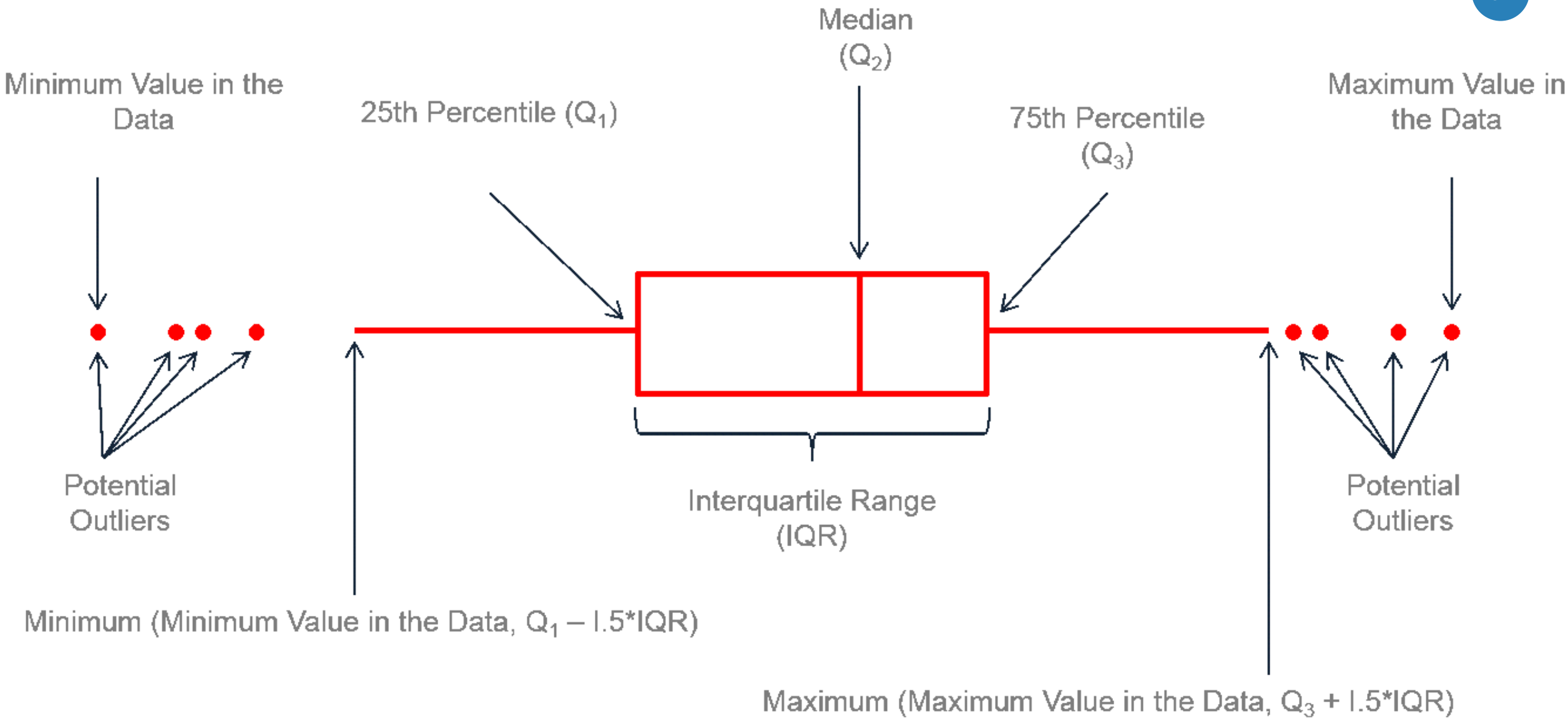
Deskriptívna Štatistika

Using the Hmisc package

```
library(Hmisc)
describe(mydata)
# n, nmiss, unique, mean, 5,10,25,50,75,90,95th percentiles
# 5 lowest and 5 highest scores
```

Using the **pastecs** package

```
library(pastecs)
stat.desc(mydata)
# nbr.val, nbr.null, nbr.na, min max, range, sum,
# median, mean, SE.mean, CI.mean, var, std.dev, coef.var
```


Median
(Q_2)
Minimum Value in the Data
25th Percentile (Q_1)
75th Percentile
(Q_3)
Maximum Value in the Data
Potential Outliers
Interquartile Range
(IQR)
Potential Outliers
Minimum (Minimum Value in the Data, Q_1 – 1.5*IQR)
Maximum (Maximum Value in the Data, Q_3 + 1.5*IQR)

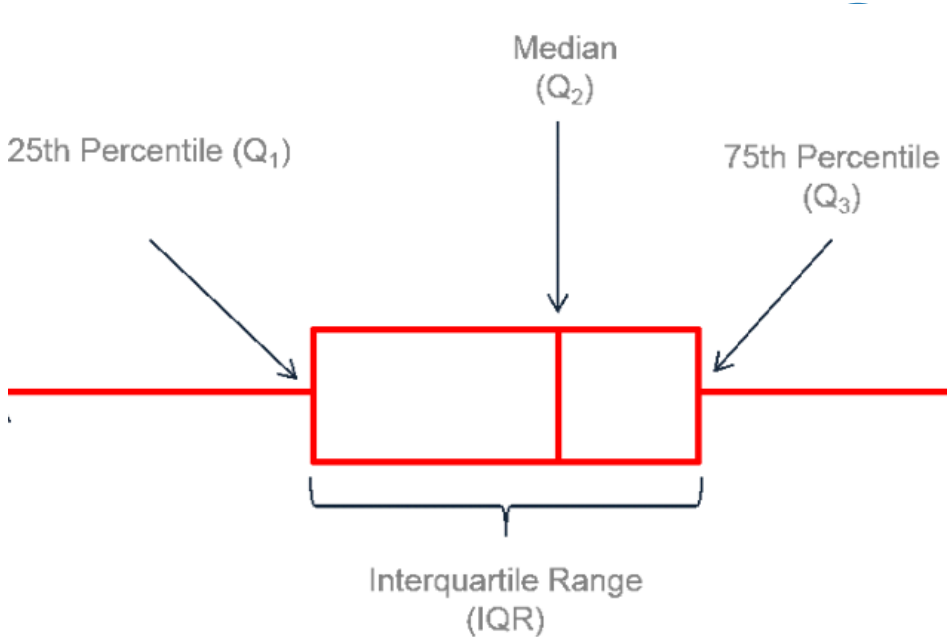
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{\left(x_1 - \bar{x}\right)^2 + \left(x_2 - \bar{x}\right)^2 + \ldots + \left(x_n - \bar{x}\right)^2}{n-1}$$

– Vážený (z frekv. tabuľky):

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2 n_i = \frac{\left(x_1 - \bar{x}\right)^2 n_1 + \ldots + \left(x_m - \bar{x}\right)^2 n_m}{n-1}$$

MIERY

VARIABILITY



$$s = \sqrt{s^2}$$

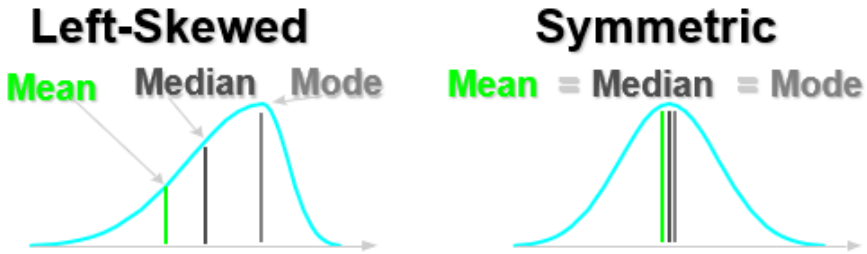
$$SE = \frac{s}{\sqrt{n}}$$

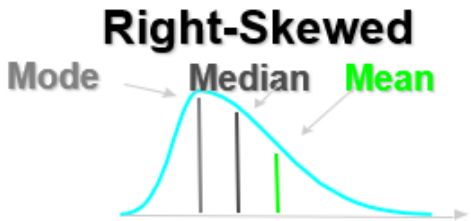
$$R = x_{max} - x_{min}$$

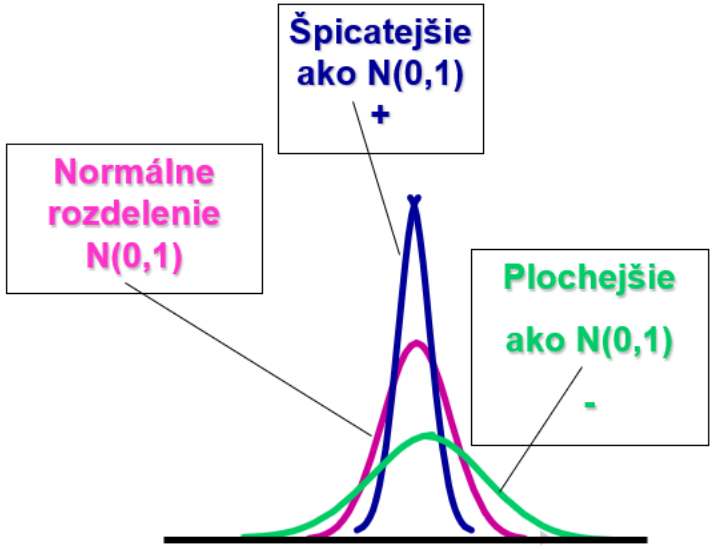
- Rozptyl
- Štandardná odchylka
- Štandardná chyba
- Variačné rozpätie
- Medzikvartilové rozpätie
- Variačný koeficient

$$V_k = \frac{s}{\bar{x}} \cdot (100)$$

MIERY







TVARU

- **Šikmosť** (Skewness)
- **Špicatosť** (Kurtosis)

```
# calculate skewness
install.packages("moments")
library(moments)
skewness(x)
kurtosis(x)
```

Tvar (Shape) & Box Plot

Koeficient Šikmosti (Skew = Skewness), označenie γ_1

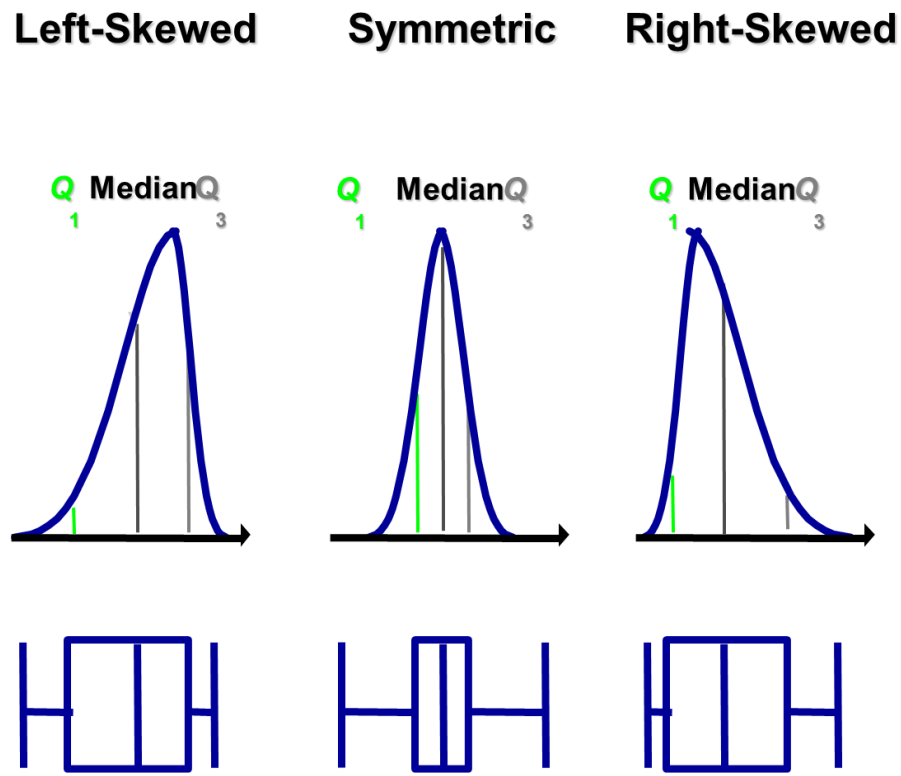
$$\gamma_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3}{s^3}$$

- Zošikmenie Doľava → Kladný
- Zošikmenie Doprava → Záporný
- Symetria → 0

Koeficient Špicatosti (Kurt = Kurtosis) označenie γ_2

$$\gamma_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^4}{s^4} - 3$$

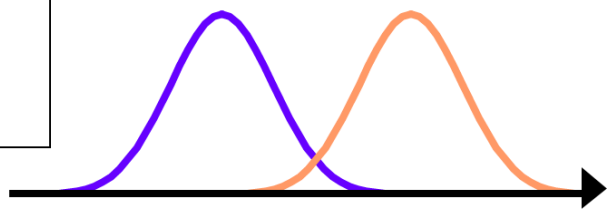
- Normálne rozdelenie → 0
- Špicatejšie → Kladný
- Plochejšie → Záporný



Ukazovatele (Miery)
Vlastnosti Kvantitatívnych Dát

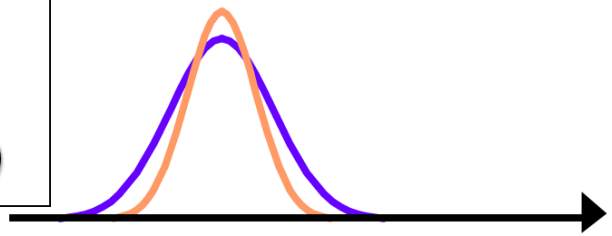
Poloha

Central Tendency (Location)



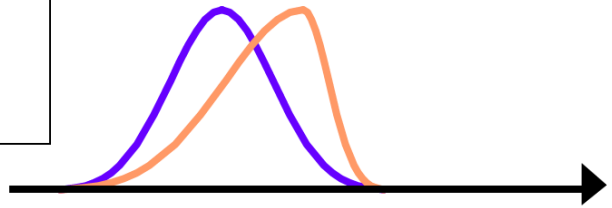
Menlivosť

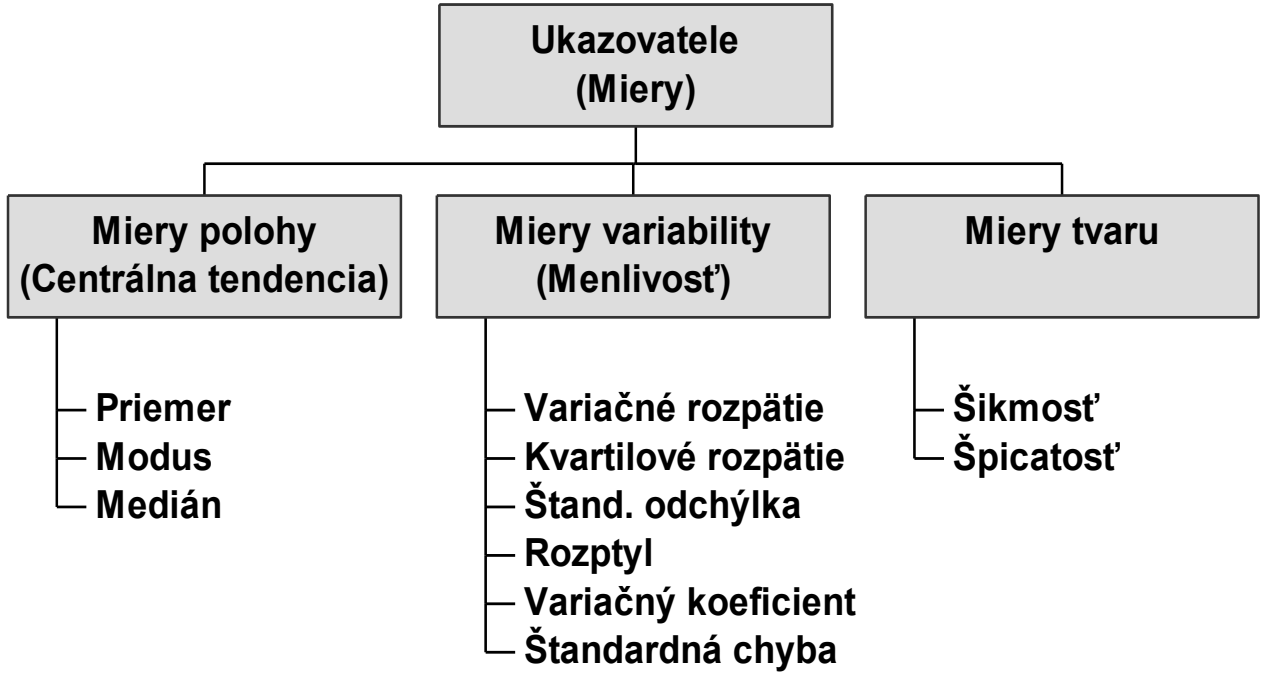
Variation (Dispersion)



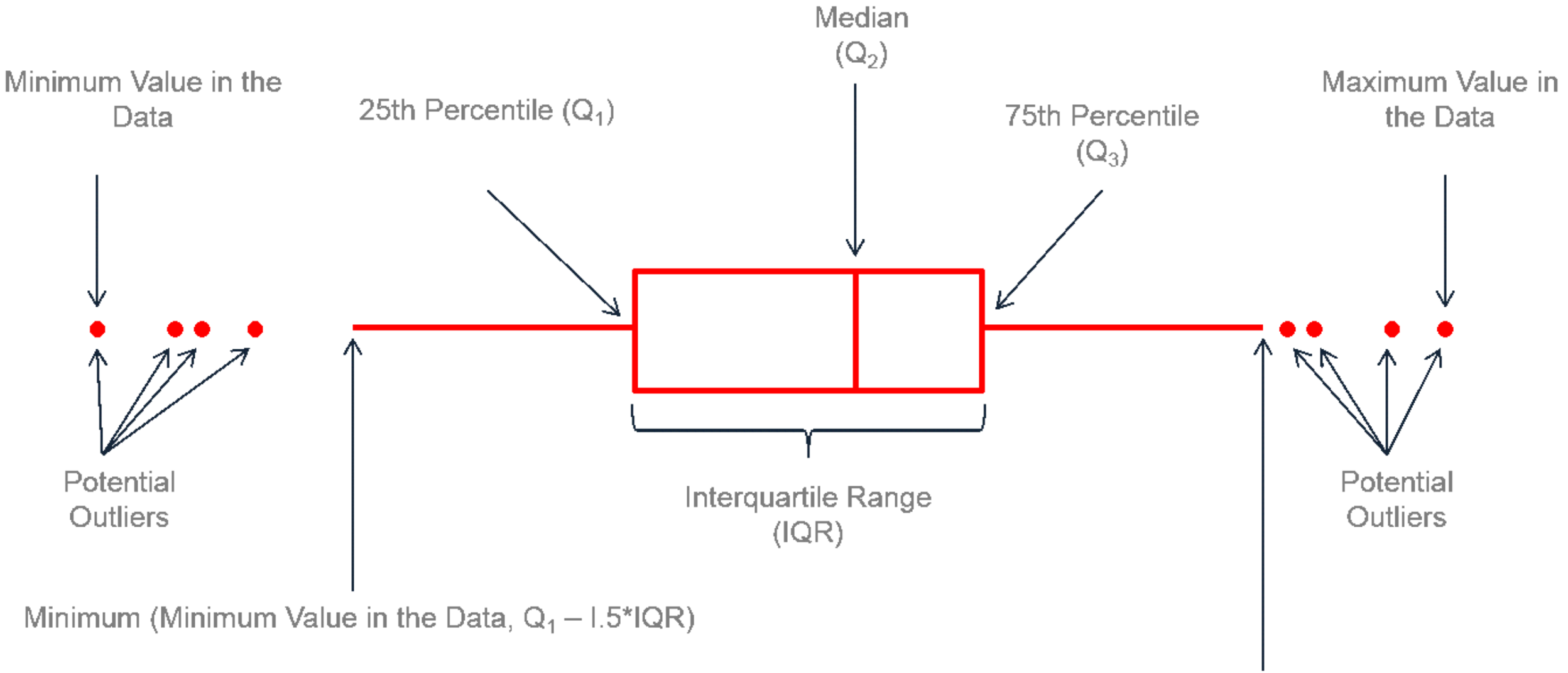
Tvar

Shape





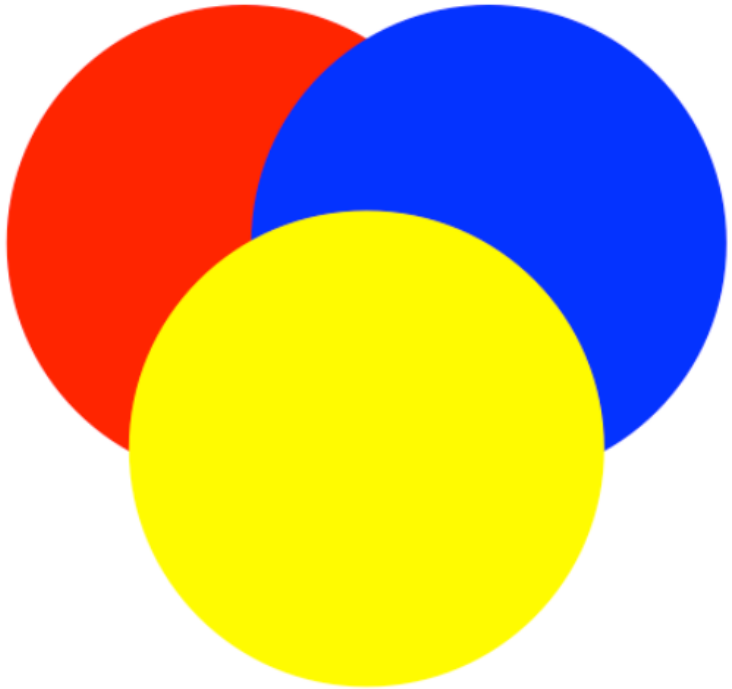
Box plot (krabicový) graf



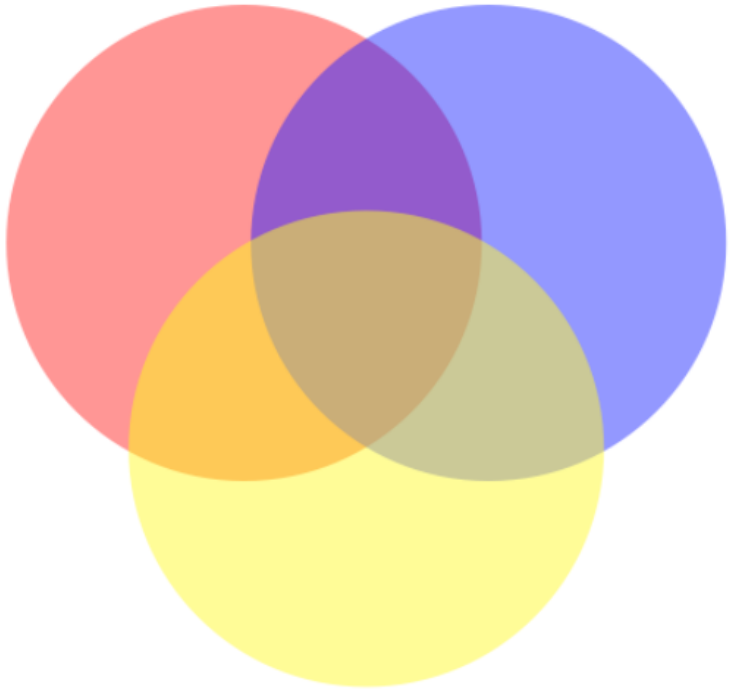
Témy a Farby v Jazyku R

alpha

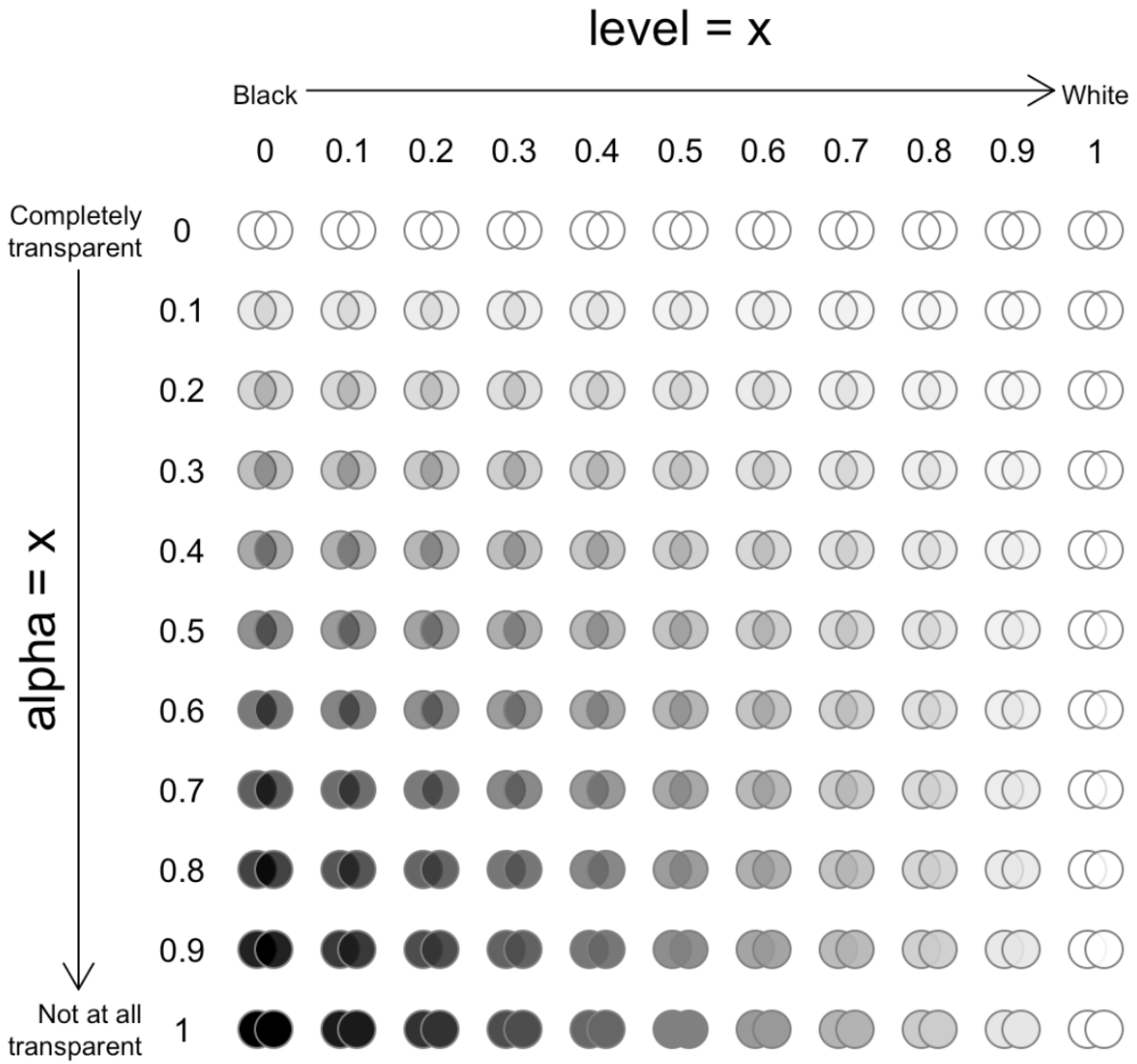
Standard



Transparent



Transparentnosť (alpha)



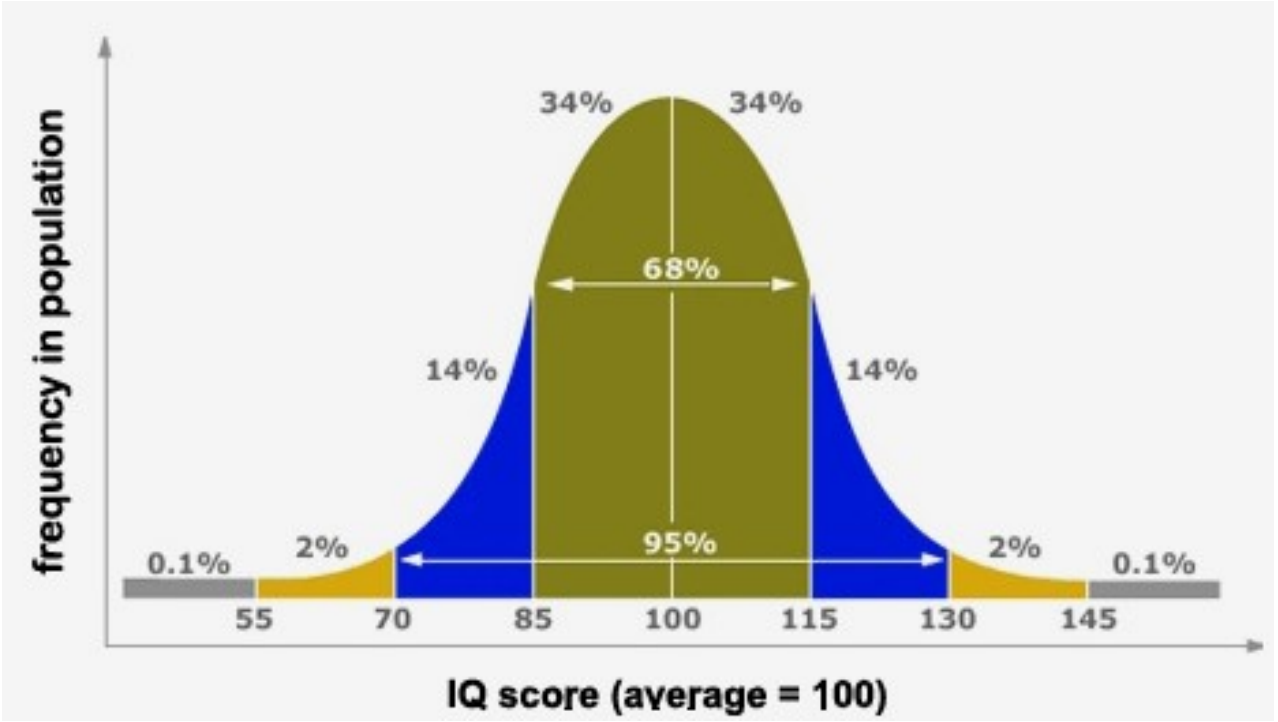
Normálne Rozdelenie (Gaussovo rozdelenie)

- Meracie prístroje (zákon chýb)
- Výrobky v továrni
- Fyzikálne a Technické Veličiny
- IQ
- Hmotnosť, výška Ľudí
- Objemy neopracovaných vzoriek
- Prirodzená Remantná Magnetizácia
- Väčšine prípadov má normálne rozdelenie, alebo rozdelenie z neho odvodené (t, F, Chi-kvadrát rozdelenie)
- Veľa premenných má normálne rozdelenie, čo sa stalo príčinou toho, že normálne rozdelenie sa považuje za všeobecnú črtu skúmanej reality
- Problém nastane, keď sa snažíme použiť takýto test na dáta, u ktorých sa normálne rozdelenie nepotvrdí

Normálne Rozdelenie

- Cieľom štatistického testovania je odhadnúť vzťah medzi premennými, teda pomer vysvetlenej varácie jednej premennej pomocou druhej voči celkovej variácii tej premennej. Teda, až do akej miery, vyjadrenej v %, sa zmena prvej premennej dá vysvetliť zmenou druhej a naopak
- Vo väčšine prípadov poznáme tvar funkcie a vieme určiť významnosť pre náš nález v príslušnej vzorke.

- Väčšina týchto funkcií súvisí s funkciou nazývanou 'normálna'
- Jej krivka, nazývaná aj **Gaussova krivka**, má tvar **zvona** a je funkciou iba 2 parametrov: **priemeru** a **štandardnej odchýlky**
- Normálne rozdelenie (distribúcia) reprezentuje jednu z empiricky verifikovaných právd o 'všeobecnej povahe reality'
- O normálnom rozdelení platí: **68 % populácie má hodnotu meraného znaku ležiacu v intervale priemer ± 1 krát štandardná odchýlka**, **95 % v intervale priemer ± 2 krát štandardná odchýlka**
- Inými slovami, hodnota, ktorá neleží v tomto intervale má relatívnu frekvenciu 5% alebo menej



Aké IDE použiť?

Integrated Development Environment

Chceme úplne všetko!

ANACONDA®

Individual Edition is now

ANACONDA DISTRIBUTION

The world's most popular open-source Python distribution platform

Anaconda Distribution

Download

For Windows

Python 3.9 • 64-Bit Graphical Installer • 510 MB

Get Additional Installers

Open Source

Access the open-source software you need for projects in any field, from data visualization to robotics.

User-friendly

With our intuitive platform, you can easily search and install packages and create, load, and switch between environments.

Trusted

Our securely hosted packages and artifacts are methodically tested and regularly updated.

Collaboration

Flexibility

Code
Completion

Search notebooks

MR

MR **Home**

- Your notebooks
- Recent activity
- Shared with you
- Your favorites
- Trash

- Workspace files
- Reports
- Scheduled notebooks
- Databases
- Cloud storages
- Environment variables

- Settings

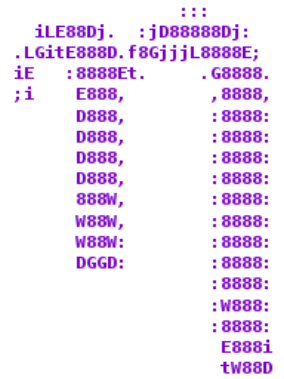
+ New notebook

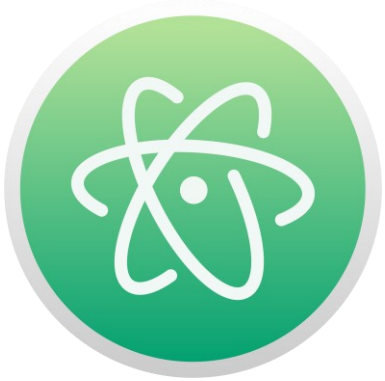
Your notebooks

Name ↓	Last modified	Author
Python II. Mierne Pokrocily	26 Oct 2022 00:34	Miroslav Reiter
Python I. Zaciatocnik	26 Oct 2022 00:34	Miroslav Reiter
Python 5 Automatizácia Úloh	09 May 2023 12:56	Miroslav Reiter
Python 4 Databázy	25 Apr 2023 23:15	Miroslav Reiter
Python 3 SVD	26 Oct 2022 00:32	Miroslav Reiter
Python 2 OOP	26 Oct 2022 00:33	Miroslav Reiter
Kurz SAV R Základy	10 Nov 2022 10:09	Miroslav Reiter
Kurz SAV R Spracovanie Udajov	13 Apr 2023 00:59	Miroslav Reiter
Kurz SAV Python Zaklady	15 Apr 2023 11:39	Miroslav Reiter
Getting started tutorial	29 Sep 2022 22:42	Miroslav Reiter


```
print("Sorry, we are down for maintenance")
print("We'll be back shortly")
```

Aký Editor Použiť?





IDE ≠ Editor

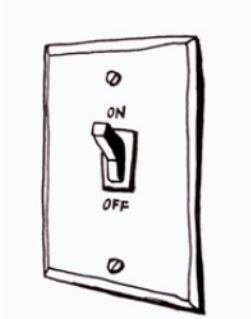
Balíčky (Inštalácia a Načítanie)

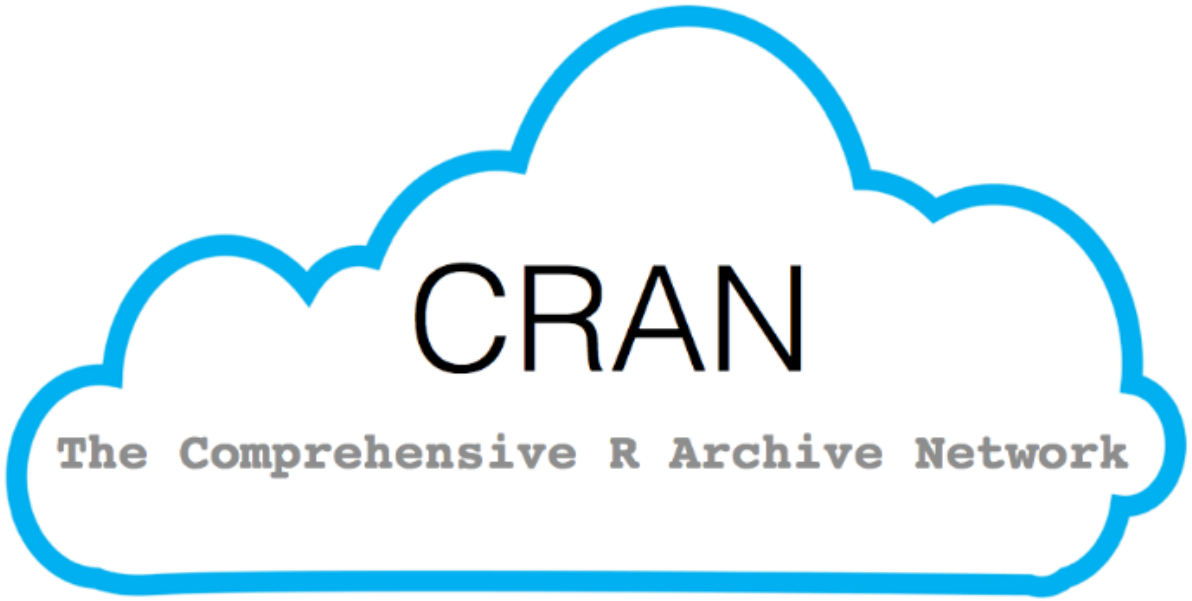
Installing a package

```
install.packages('my.package')
```

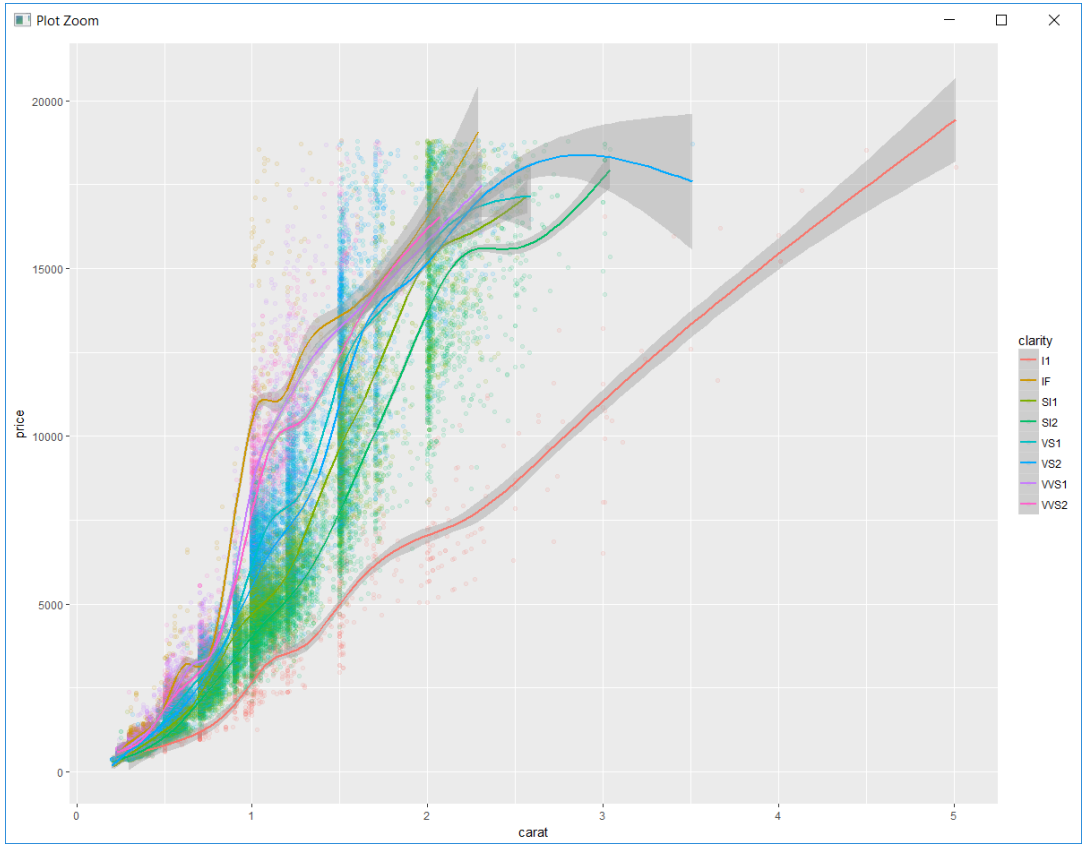
Loading a package

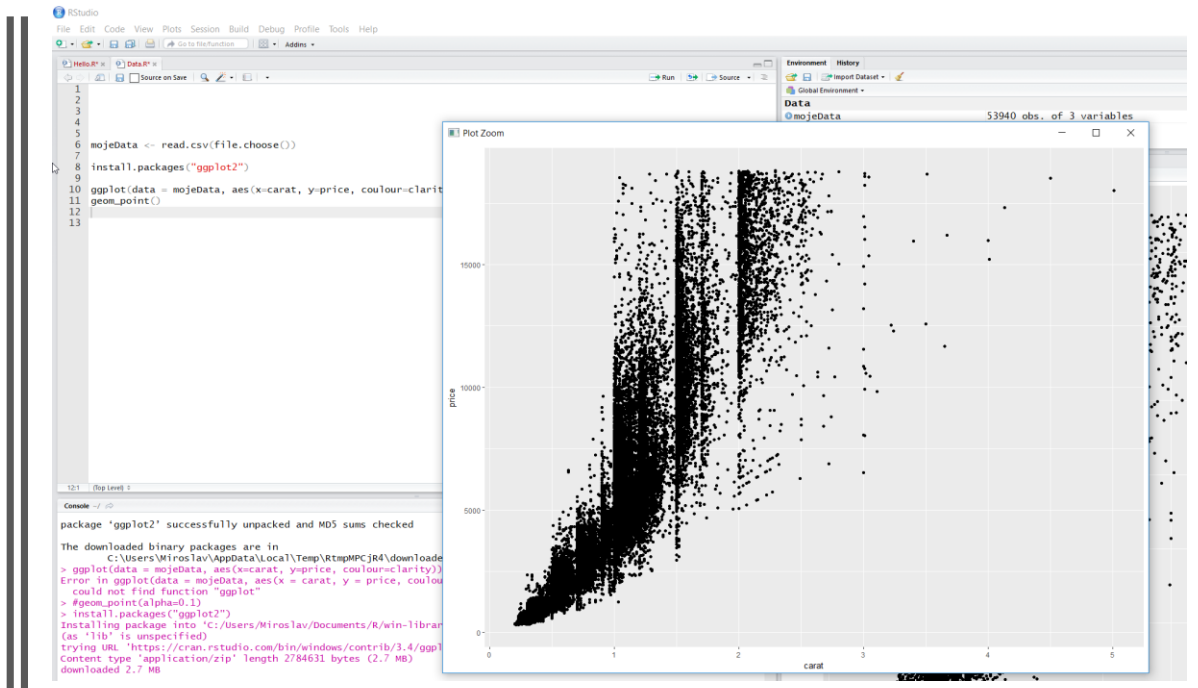
```
library('mypackage')
```



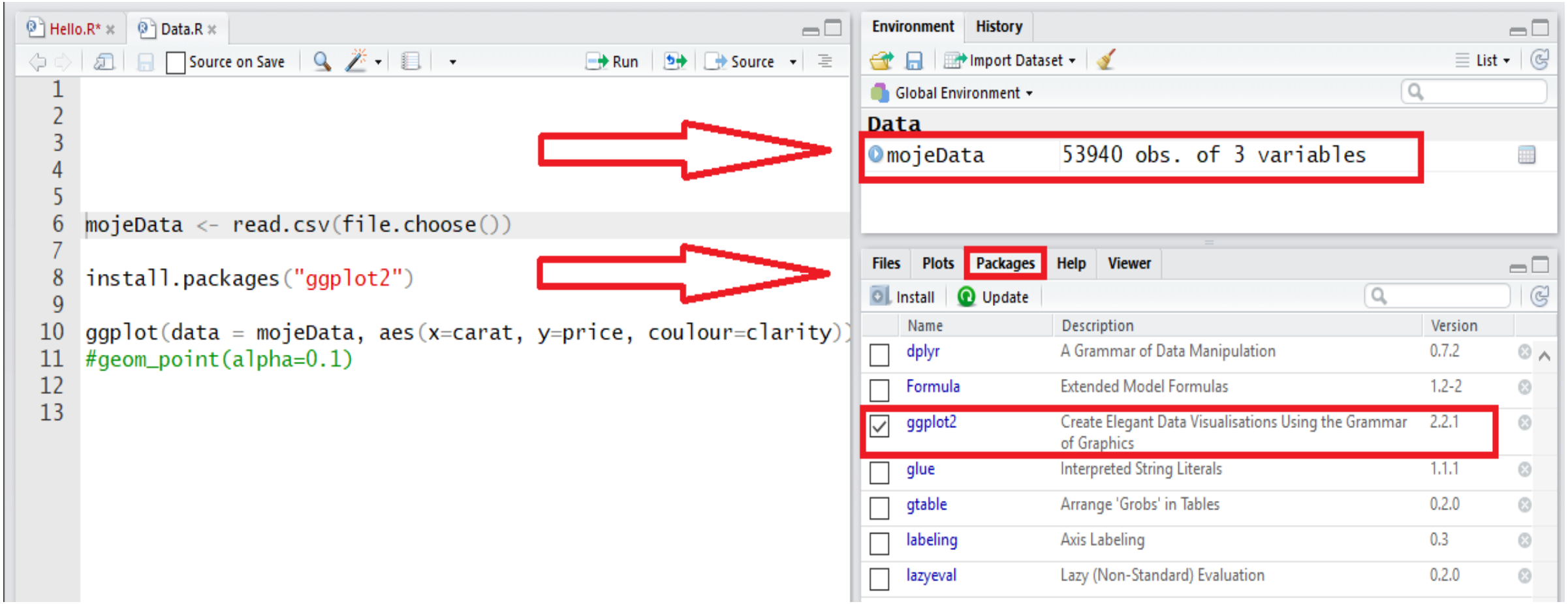


Inštalácia Balíčkov (packages)





Import Dát a Balíčkov



dplyr

Overview

dplyr is a grammar of data manipulation, providing a consistent set of verbs that help you solve the most common data manipulation challenges:

- `mutate()` adds new variables that are functions of existing variables
- `select()` picks variables based on their names.
- `filter()` picks cases based on their values.
- `summarise()` reduces multiple values down to a single summary.
- `arrange()` changes the ordering of the rows.

These all combine naturally with `group_by()` which allows you to perform any operation "by group". You can learn more about them in `vignette("dplyr")`. As well as these single-table verbs, dplyr also provides a variety of two-table verbs, which you can learn about in `vignette("two-table")`.

LINKS

View on CRAN

Browse source code

Report a bug

Learn more

COMMUNITY

Contributing guide

Code of conduct

Getting help

CITATION

Citing dplyr

DEVELOPERS

Hadley Wickham

Author, maintainer

Výber Dát (Select)

Dplyr.R

Source on Save

```
1  # Import dplyr
2  library(dplyr)
3  data(infert)
4
5  infert %>%
6    select(education,age,parity) %>%
7    head()
8
9
```

Selecting Columns of Data

When you're dealing with large datasets, it can be confusing to look at many different columns of data at once. Here, select the `President`, `Date`, and `Approve` columns from the `approval_polls` dataset, and pipe the result to the `head()` function to display the results.

INSTRUCTIONS 100 XP

- Using the `select()` function from the `tidyverse` package, extract the `President`, `Date`, and `Approve` columns from the `approval_polls` dataset.

Take Hint (-30 XP)

```
1  # Import dplyr
2  library(dplyr)
3
4  # Select President, Date, and Approve from approval_polls
5  approval_polls %>%
6    select(President,Date,Approve) %>%
7    head()
```

Run Code | Submit Answer

R CONSOLE | SLIDES

```
> # Import dplyr
> library(dplyr)
>
> # Select President, Date, and Approve from approval_polls
> approval_polls %>%
    select(President,Date,Approve) %>%
    head()
  President       Date Approve
1     Trump 12/12/2017      36
2     Trump  12/9/2017      36
3     Trump  12/6/2017      37
4     Trump  12/3/2017      35
5     Trump 11/30/2017      34
6     Trump 11/27/2017      37
```

Atomické Vektory

- Commonly knows as vectors
- Homogeneous data structure

Character vector	Numeric vector	Integer vector	Logical vector	Complex vector
{"A","c"}	{ 4.36, 7.42 }	{ 3,5 }	{ True, False }	{ 1 + 7i , 8 - 2i }

Zoznam premenných

```
> ls(all.names = T)
 [1] ".Random.seed"    "a"               "A"               "B"
 [5] "C"               "jeSkladom"       "komplexneCislo"  "meno"
 [9] "menoSpolocnosti" "mzdaZamestnanca" "priezvisko"      "sprava"
[13] "vekKlienta"      "vysledok"        "x2"              "x3"
> ls(all.names = T,sorted = T)
 [1] ".Random.seed"    "a"               "A"               "B"
 [5] "C"               "jeSkladom"       "komplexneCislo"  "meno"
 [9] "menoSpolocnosti" "mzdaZamestnanca" "priezvisko"      "sprava"
[13] "vekKlienta"      "vysledok"        "x2"              "x3"
> ls(all.names = T,sorted = F)
 [1] "sprava"          "jeSkladom"       "A"               "B"
 [5] "C"               "vysledok"        "a"               "komplexneCislo"
 [9] "meno"            ".Random.seed"    "mzdaZamestnanca" "priezvisko"
[13] "vekKlienta"      "menoSpolocnosti" "x2"              "x3"
> ls(all.names = T,sorted = T, pattern = "a")
[1] ".Random.seed"    "a"               "jeSkladom"       "mzdaZamestnanca"
[5] "sprava"          "vekKlienta"
> ls(all.names = T,sorted = T, pattern = "m*")
 [1] ".Random.seed"    "a"               "A"               "B"
 [5] "C"               "jeSkladom"       "komplexneCislo"  "meno"
 [9] "menoSpolocnosti" "mzdaZamestnanca" "priezvisko"      "sprava"
[13] "vekKlienta"      "vysledok"        "x2"              "x3"
```

Príkaz ls() | objects()

Zoznam Objektov a Funkcií

```
> ls(all.names = T)
 [1] ".Random.seed"    "a"               "A"               "B"
 [5] "C"               "jeSkladom"       "komplexneCislo"  "meno"
 [9] "menoSpolocnosti" "mzdaZamestnanca" "priezvisko"      "sprava"
[13] "vekKlienta"      "vysledok"        "x2"              "x3"
> ls(all.names = T,sorted = T)
 [1] ".Random.seed"    "a"               "A"               "B"
 [5] "C"               "jeSkladom"       "komplexneCislo"  "meno"
 [9] "menoSpolocnosti" "mzdaZamestnanca" "priezvisko"      "sprava"
[13] "vekKlienta"      "vysledok"        "x2"              "x3"
> ls(all.names = T,sorted = F)
 [1] "sprava"          "jeSkladom"       "A"               "B"
 [5] "C"               "vysledok"        "a"               "komplexneCislo"
 [9] "meno"            ".Random.seed"    "mzdaZamestnanca" "priezvisko"
[13] "vekKlienta"      "menoSpolocnosti" "x2"              "x3"
> ls(all.names = T,sorted = T, pattern = "a")
[1] ".Random.seed"    "a"               "jeSkladom"       "mzdaZamestnanca"
[5] "sprava"          "vekKlienta"
> ls(all.names = T,sorted = T, pattern = "m*")
 [1] ".Random.seed"    "a"               "A"               "B"
 [5] "C"               "jeSkladom"       "komplexneCislo"  "meno"
 [9] "menoSpolocnosti" "mzdaZamestnanca" "priezvisko"      "sprava"
[13] "vekKlienta"      "vysledok"        "x2"              "x3"
```

Príkaz builtins() | ls(baseenv(), all = TRUE)

Dátové Štruktúry

Homogénne typy

- Rovnaké typy položiek
  - Atomické Vektory (Vectors)
  - Matice (Matrices)
  - Polia (Arrays)

Heterogénne typy

- Rôzne typy položiek
  - Zoznamy (Lists)
  - Dátové Rámce (DataFrames)

Ďalšie Dátové Typy (Dátové Štruktúry)

Matice alebo Polia	• Sú viacrozmerné zovšeobecnenia vektorov • V skutočnosti sú to vektory, ktoré môžu byť indexované dvomi alebo viacerými indexmi
Faktory	• Kompaktné spôsoby spracovania kategorických údajov
Zoznamy	• Všeobecná forma vektora, v ktorej rôzne prvky nemusia byť rovnakého typu a sú často samotné vektory alebo zoznamy
Dátové Rámce	• Sú maticové štruktúry, v ktorých môžu byť stĺpce rôznymi typmi • Premýšľajte o dátových rámcoch ako o "dátových maticiach„ tabuľkách
Funkcie	• Sú samotné objekty v R, ktoré môžu byť uložené v pracovnom priestore projektu

Konverzie dát

	to one long vector	to matrix	to data frame
from vector	c(x,y)	cbind(x,y) rbind(x,y)	data.frame(x,y)
from matrix	as.vector(mymatrix)		as.data.frame(mymatrix)
from data frame		as.matrix(myframe)	

Dátumy (Dates)

```
# use as.Date( ) to convert strings to dates
mydates <- as.Date(c("2007-06-22", "2004-02-13"))
# number of days between 6/22/07 and 2/13/04
days <- mydates[1] - mydates[2]
```

```
# print today's date
today <- Sys.Date()
format(today, format="%B %d %Y")
"June 20 2007"
```

Formátovanie Dátumy

Symbol	Meaning	Example
%d	day as a number (0-31)	01-31
%a %A	abbreviated weekday unabbreviated weekday	Mon Monday
%m	month (00-12)	00-12
%b %B	abbreviated month unabbreviated month	Jan January
%y %Y	2-digit year 4-digit year	07 2007

default format yyyy-mm-dd

Konverzie Dátumy

```
# convert date info in format 'mm/dd/yyyy'
strDates <- c("01/05/1965", "08/16/1975")
dates <- as.Date(strDates, "%m/%d/%Y")
```

```
mydates <- as.Date(c("2007-06-22", "2004-02-13"))
```

```
# convert dates to character data
strDates <- as.character(dates)
```

default format yyyy-mm-dd

Generovanie dát

Sekvencie	Replikácie/opakovanie	Permutácie a vzorové dáta
• 1:10, • seq(10), • seq(1, 10, 2)	• rep(1:4, 2), • rep(1:4, each = 2)	• sample(1:20, 9)

as.Date("1998-12-17")

Matematické funkcie

Aritmetické funkcie

- log(x)
- logb()
- log10()
- log2()
- exp()
- expm1()
- log1p()
- sqrt()

Goniometrické funkcie

- cos()
- sin()
- tan()
- acos()
- asin()
- atan()
- atan2()

Hyberbolické funkcie

- cosh()
- sinh()
- tanh()
- acosh()
- asinh()
- atanh()

Množinové operácie

- union()
- intersect()
- setdiff()
- setequal()

Frekvenčné a Krížové Tabuľky

```
# 2-Way Frequency Table
attach(mydata)
mytable <- table(A,B) # A will be rows, B will be columns
mytable # print table

margin.table(mytable, 1) # A frequencies (summed over B)
margin.table(mytable, 2) # B frequencies (summed over A)

prop.table(mytable) # cell percentages
prop.table(mytable, 1) # row percentages
prop.table(mytable, 2) # column percentages
```

```
# 3-Way Frequency Table
mytable <- table(A, B, C)
ftable(mytable)
```

Frekvenčné a Krížové Tabuľky 3-way

```
# 3-Way Frequency Table
mytable <- table(A, B, C)
ftable(mytable)
```

```
# 3-Way Frequency Table
mytable <- xtabs(~A+B+c, data=mydata)
ftable(mytable) # print table
summary(mytable) # chi-square test of indepedence
```

Frekvenčné a Krížové Tabuľky

```
# 2-Way Cross Tabulation
library(gmodels)
CrossTable(mydata$myrowvar, mydata$mycolvar)
```

Testy Nezávislosti

Test Chi-Square

- Pre dvojcestné tabuľky môžete použiť test chisq.test (mytable) na overenie nezávislosti premennej riadka a stĺpca
- Predvolene je hodnota p vypočítaná z asymptotickej distribúcie štatistických údajov testu
- Prípadne môže byť hodnota p odvodená pomocou simultánnej funkcie Monte Carlo

Fisher Exact Test

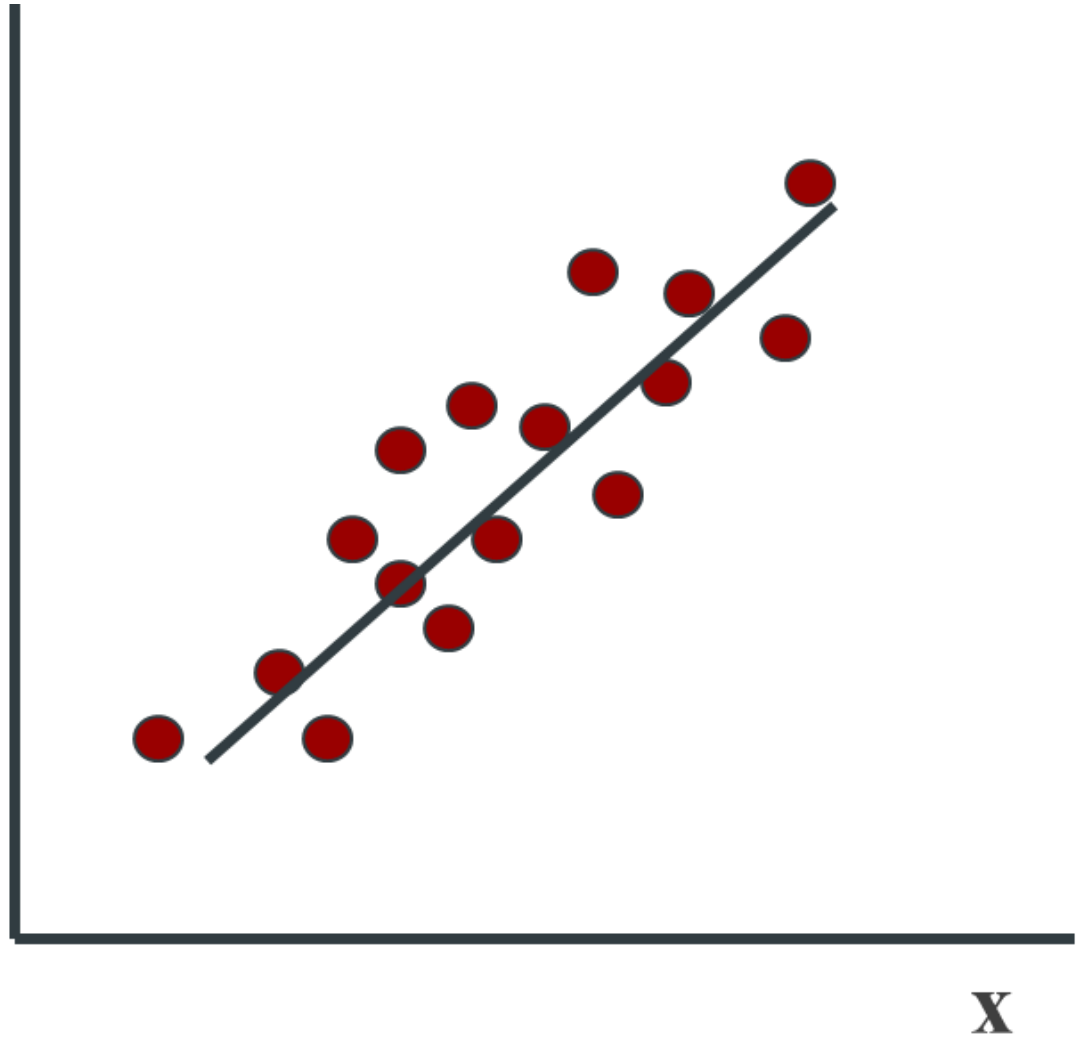
- fisher.test (x) poskytuje presný test nezávislosti
- x je dvojrozmerná kontingenčná tabuľka v maticovej forme

```
> chisq.test(zlocinci)

        Pearson's Chi-squared test

data:  zlocinci
X-squared = 714.85, df = 147, p-value < 2.2e-16

Warning message:
In chisq.test(zlocinci) : Chi-squared approximation may be incorrect
>
```

IT ACADEMY

Korelačná Analýza

- Overenie vypovedacej schopnosti kvantifikovaných regresných modelov ako celku, aj jeho častí
- Výpočet **číselných charakteristík**, ktoré v koncentrovanej forme popisujú **kvalitu vypočítaných modelov**
- Požadujeme od nich, aby sa pohybovali v pevne ohraničenom intervale
- V rámci intervalu rástli s vyššou silou závislosti
- Porovnanie 2 prípadov závislosti



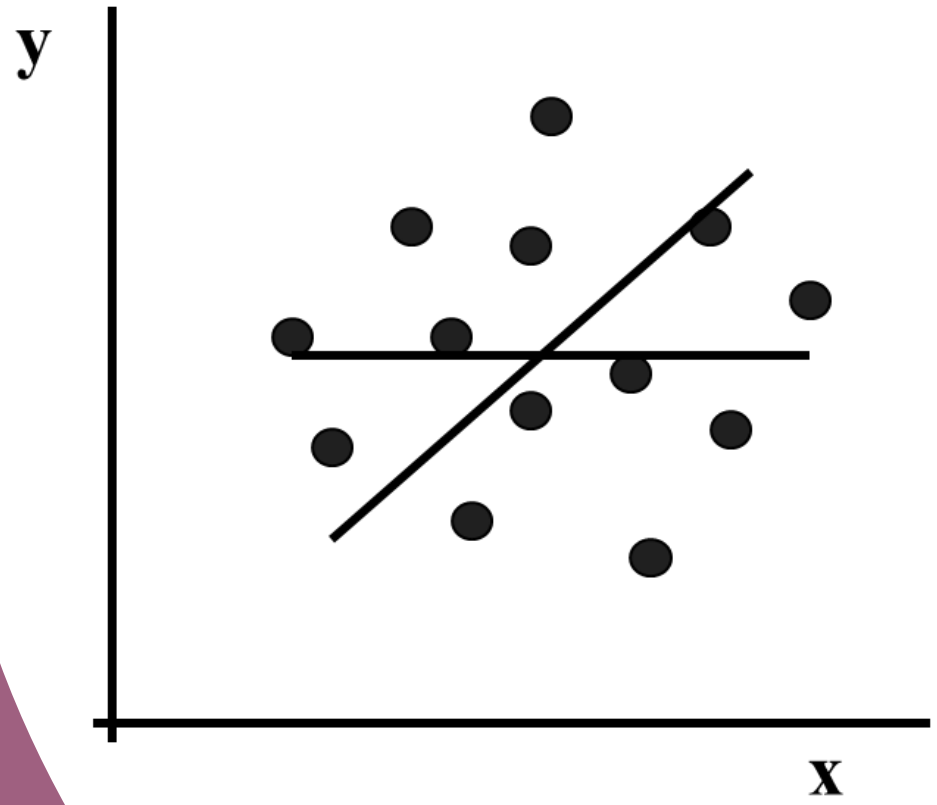
Korelačná Analýza

- Po úvodnom **grafickom preskúmaní** nastupuje **fáza hľadania presných štatistík**, ktoré potvrdia odhady z grafov
- Korelačné štatistiky zisťujú **či medzi premennými existuje korelácia aká je sila korelácie**
- **Koreláciou nazývame vzájomný lineárny vzťah – závislosť 2 premenných**
- Tento vzťah môže byť:
  1. **Priamy** – s rastúcimi hodnotami 1. premennej rastú aj hodnoty 2. premennej
  2. **Nepriamy** – s rastúcimi hodnotami 1. premennej klesajú hodnoty 2. premennej
  3. Ak medzi hodnotami 2 premenných neexistuje ani priama ani nepriama lineárna závislosť, hovoríme, že sú **nekorelované.**

Korelačná Analýza

Miery tesnosti štatistickej závislosti:

- **Kovariancia** – cov_{yx}
  - len pre lineárnu závislosť
- **Koeficient Korelácie** r_{yx}
  - len pre lineárnu závislosť
- **Koeficient Determinácie** r_{yx}^2
  - len pre lineárnu závislosť
- index korelácie i_{yx}
- index determinácie i_{yx}^2



Kovariancia

- **Miera**, ktorá sa používa, aby sme **potvrdili** alebo **vyvrátili existenciu lineárnej závislosti (korelácie)**
- Zo spôsobu výpočtu možno odvodiť, kedy **potvrdzuje existenciu pozitívnej**, **negatívnej korelácie** a **kedy nekorelovanosti**
- **Ak kovariancia** potvrdí **neexistenciu lineárneho vzťahu**, medzi premennými **môže existovať nelineárny vzťah**
- **Ak kovariancia potvrdí existenciu lineárneho vzťahu**, môžeme **merať** jeho **intenzitu**

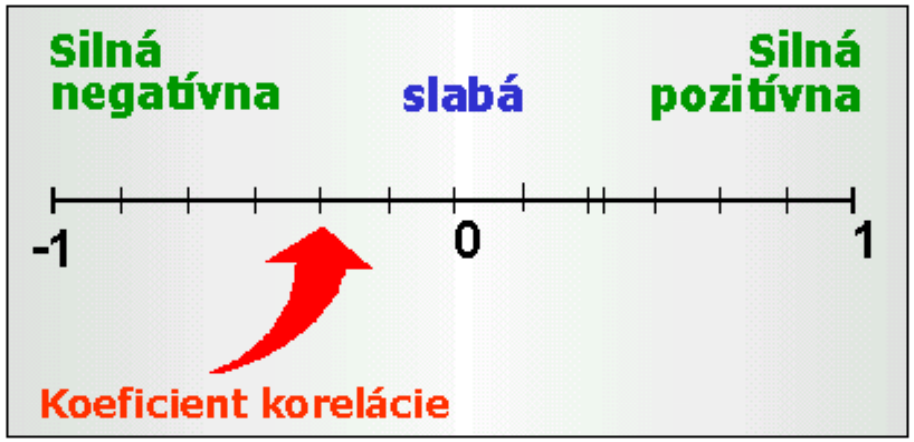
$$\operatorname{cov}\, xy = \frac{1}{n}\sum \left(x_i - \bar{x} \right) \left(y_i - \bar{y} \right)$$

cov xy = 0, medzi premennými **nie je lineárny vzťah**
cov xy > 0, medzi premennými **je priamy** lin. vzťah
cov xy < 0, medzi premennými **je nepriamy lin. vzťah**

Korelačná Analýza

- **Koeficient korelácie - r_{yx}**

$$r_{yx} = \frac{\mathrm{COV}_{yx}}{s_x s_y}$$



- Hodnoty sa pohybujú v intervale: –1, 1
  - r_{yx} = -1 – silná negatívna závislosť
  - r_{yx} = 0 – bez závislosti
  - r_{yx} = 1 – silná pozitívna závislosť
- Koeficient determinácie r_{yx}^2
- Hodnoty sa pohybujú v intervale: 0,1
- Udáva % vysvetlenej variability závisle premennej

Index Korelácie a Index Determinácie

- V ZS I_{yx} odhadom z výberových údajov je i_{yx} est $I_{yx} = i_{yx}$.
- Princíp spočíva v rozklade variability závisle premennej Y

$$\sum_{j=1}^{n}(y_j - \bar{y})^2 = \sum_{j=1}^{n}(y_j' - \bar{y})^2 + \sum_{j=1}^{n}(y_j - y_j')^2$$

Celková variabilita závisle premennej

Variabilita závisle premennej vysvetlená regresnou funkciou

Variabilita nevysvetlená regresnou funkciou – reziduálna variabilita

Index korelácie a index determinácie

- **Index korelácie i_{yx}**

$$i_{yx} = \sqrt{\frac{\sum_{j=1}^{n}(y_j' - \bar{y})^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2}} = \sqrt{\frac{V}{C}}$$

- **Index determinácie i_{yx}^2**

$$i_{yx}^2 = \frac{C - N}{C} = 1 - \frac{N}{C} = 1 - \frac{\sum_{j=1}^{n}(y_j - y_j')^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2}$$

Index korelácie a Index determinácie

Index Korelácie

- Hodnoty sa pohybujú v intervale od (0,1)
- Čím sa hodnota indexu blíži k 1, tým je tesnosť závislosti vyššia a opačne

Index Determinácie

- Nadobúda hodnoty z intervalu 0 až 1
- Čím viac sa hodnota indexu blíži k 1, tým väčšia časť celkovej variability je modelom vysvetlená a naopak
- Ak sa index determinácie blíži k 0, tým menšia časť celkovej variability je vysvetlená modelom

Index determinácie

- **Kritérium** pri **rozhodovaní** o **voľbe konkrétneho tvaru regresnej funkcie**
- **Volíme ten model**, ktorý má **vyšší koeficient determinácie** (**vyššie % vysvetlenej variability**)
- Ak však majú regresné funkcie rôzny počet parametrov, je potrebné upraviť index determinácie do korigovanej podoby v tvare →
- **Výrazný rozdiel medzi** i^2 a $i^2_{adj.}$ indikuje, že do modelu bolo zahrnutých **príliš veľa premenných**

$$i^2_{korig} = 1 - \frac{(n-1)\sum_{j=1}^{n}\left(y_j - y'_j\right)^2}{(n-p)\sum_{j=1}^{n}\left(y_j - \bar{y}\right)^2}$$

Korelácia

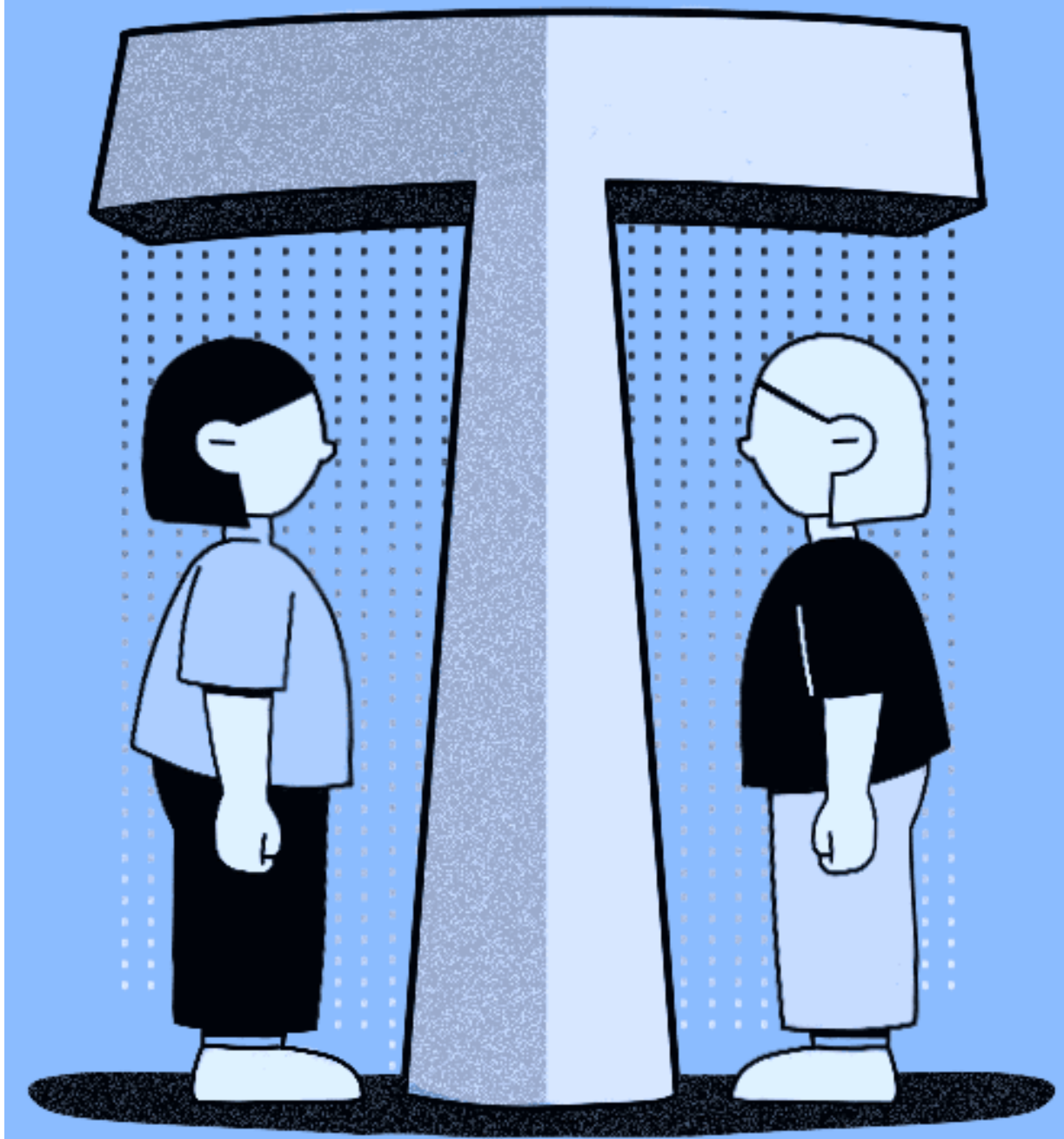
Option	Description
x	Matrix or data frame
use	Specifies the handling of missing data. Options are **all.obs** (assumes no missing data - missing data will produce an error), **complete.obs** (listwise deletion), and **pairwise.complete.obs** (pairwise deletion)
method	Specifies the type of correlation. Options are **pearson**, **spearman** or **kendall**.

```
# Correlations/covariances among numeric variables in
# data frame mtcars. Use listwise deletion of missing data.
cor(mtcars, use="complete.obs", method="kendall")
cov(mtcars, use="complete.obs")
```

Korelácia

```
# Correlations with significance levels
library(Hmisc)
rcorr(x, type="pearson") # type can be pearson or spearman

#mtcars is a data frame
rcorr(as.matrix(mtcars))
```

```
# Correlation matrix from mtcars
# with mpg, cyl, and disp as rows
# and hp, drat, and wt as columns
x <- mtcars[1:3]
y <- mtcars[4:6]
cor(x, y)
```



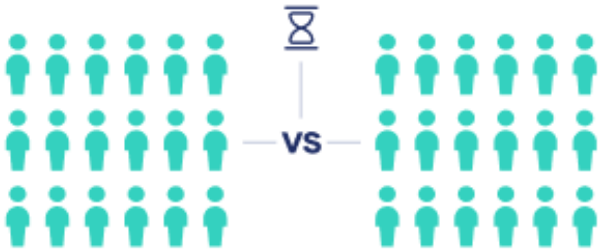
T-Test

['tē-'test]

A statistical test used to compare the means of two groups of data.

Typy t-testov

Paired-samples t test

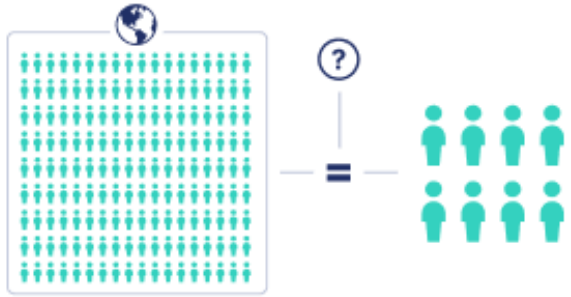


Investigate whether there's a difference within a group between two points in time (within-subjects).

Independent-samples t test

Investigate whether there's a difference between two groups (between-subjects).

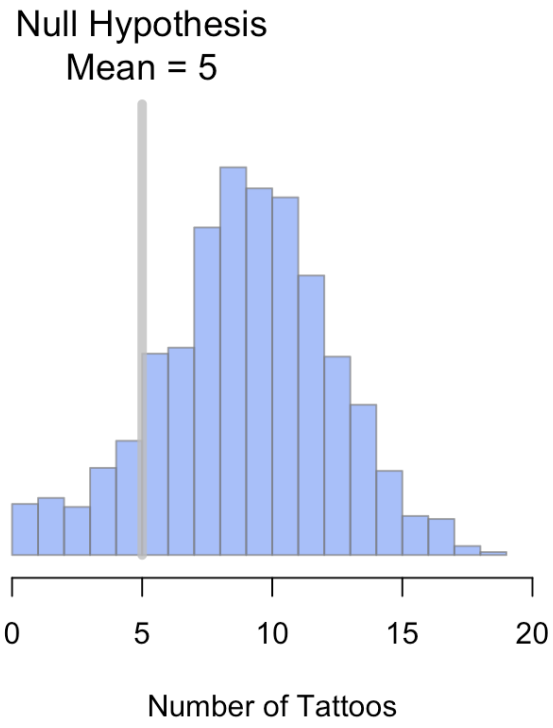
One-sample t test



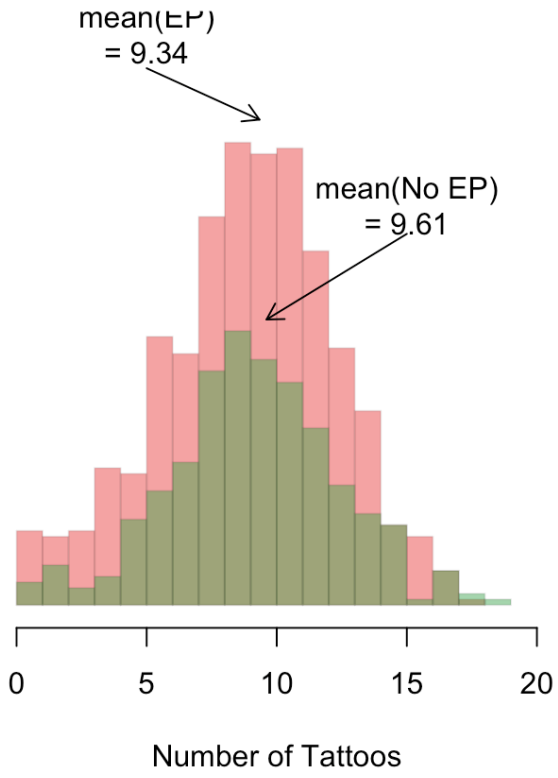
Investigate whether there's a difference between a group and a standard value or whether a subgroup belongs to a population.

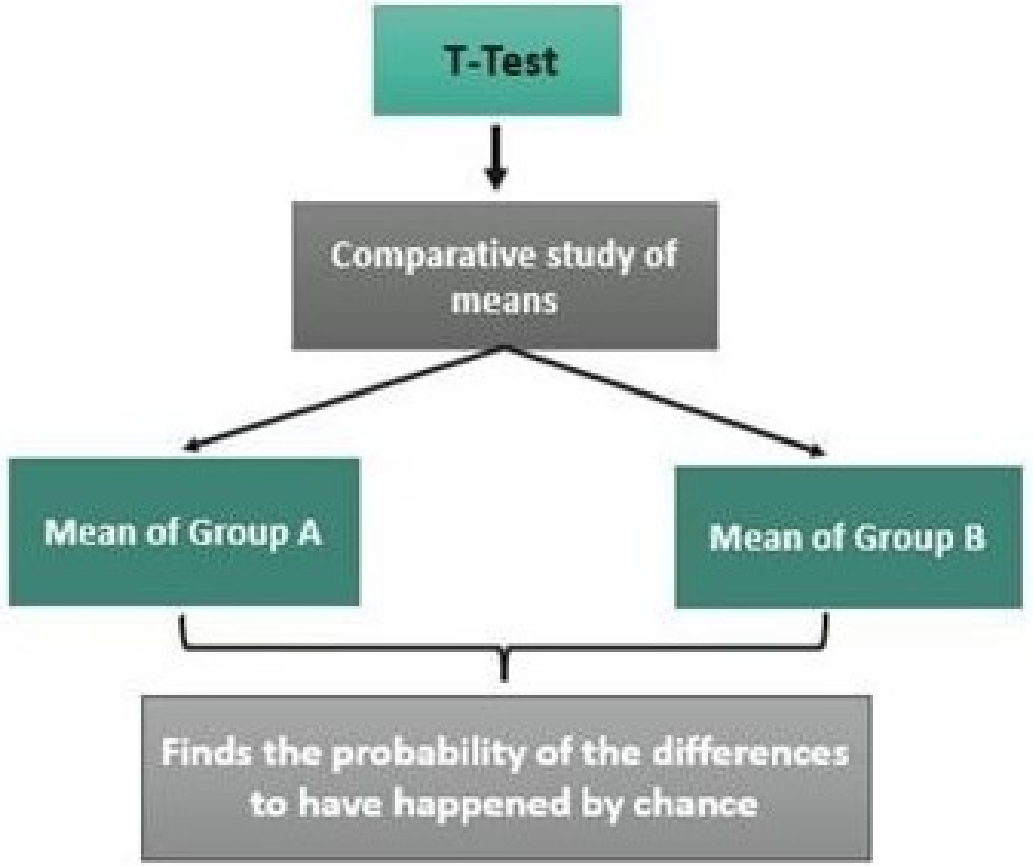
Typy t-testov

1-Sample t-test



2-Sample t-test





T-testy

```
# independent 2-group t-test
t.test(y~x) # where y is numeric and x is a binary factor
```

$$t = \frac{m_A - m_B}{\frac{\sqrt{s^2}}{\sqrt{n_A}} + \frac{\sqrt{s^2}}{\sqrt{n_B}}}$$

```
# independent 2-group t-test
t.test(y1,y2) # where y1 and y2 are numeric
```


```
# paired t-test
t.test(y1,y2,paired=TRUE) # where y1 & y2 are numeric
```

$$t = \frac{m}{s/\sqrt{n}}$$

```
# one sample t-test
t.test(y,mu=3) # Ho: mu=3
```

$$t = \frac{m - \mu}{s/\sqrt{n}}$$

var.equal = TRUE

Old Shatterhand a T-Test (1-výberový)

T-test 2 sample

- `t.test(Petal.Length ~ Species, data = flower.data)`
- Hodnota t: -33,719. Všimnite si, že je negatívny; toto je fajn! Vo väčšine prípadov nás zaujíma len absolútna hodnota rozdielu alebo vzdialenosť od 0. Nezáleží na tom, ktorým smerom.
- Stupne voľnosti: 30,196. Stupne voľnosti súvisia s veľkosťou vašej vzorky a ukazujú, koľko „bezplatných" údajových bodov je dostupných vo vašom teste na porovnávanie. Čím väčšie sú stupne voľnosti, tým lepšie bude fungovať váš štatistický test.
- Hodnota p: 2,2e-16 (t.j. 2,2 s 15 nulami). Toto popisuje pravdepodobnosť, že by ste náhodou videli takú veľkú hodnotu t, ako je táto.
- Vyjadrenie alternatívnej hypotézy (Ha). V tomto teste Ha znamená, že rozdiel nie je 0.
- 95 % interval spoľahlivosti. Toto je rozsah čísel, v ktorom bude skutočný rozdiel v priemeroch 95 % času.
- Toto sa dá zmeniť z 95 %, ak chcete väčší alebo menší interval, ale veľmi bežne sa používa 95 %.
- Priemerná dĺžka okvetných lístkov pre každú skupinu.

```
 Welch Two Sample t-test

data:  Petal.Length by Species
t = -33.719, df = 30.196, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.331287 -3.836713
sample estimates:
   mean in group setosa mean in group virginica
                  1.456                   5.540
```

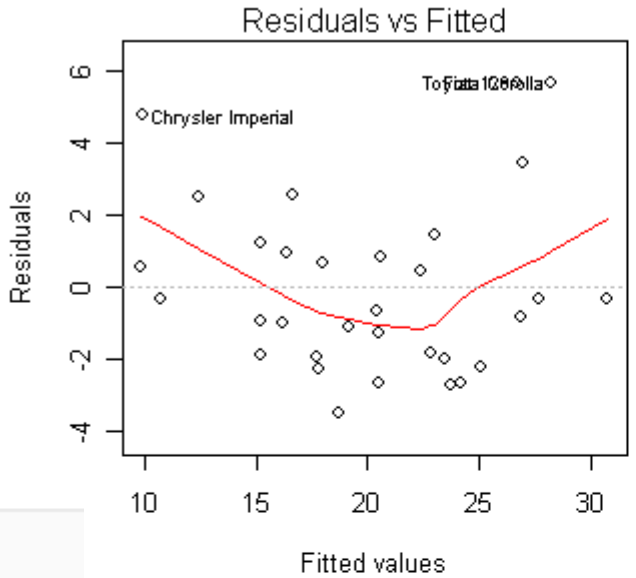
Viacnásobná (Lineárna) Regresia

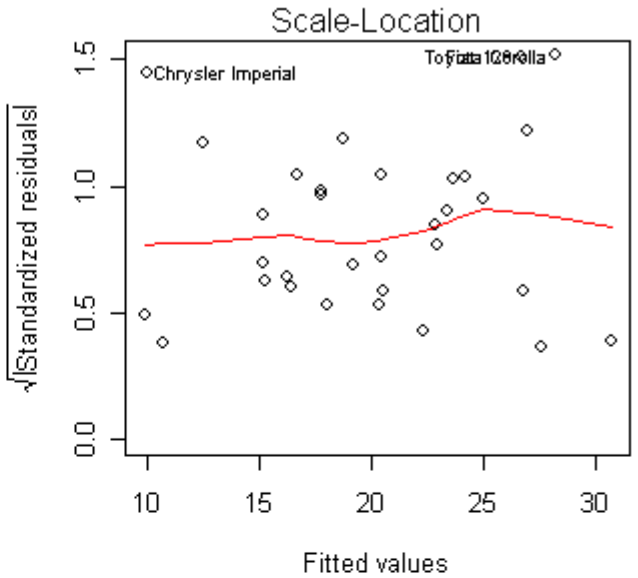
```
# Multiple Linear Regression Example
fit <- lm(y ~ x1 + x2 + x3, data=mydata)
summary(fit) # show results
```

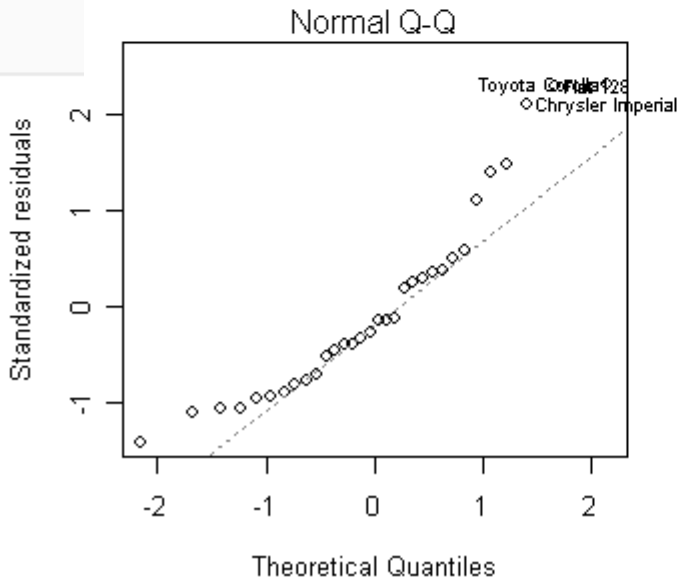
```
# Other useful functions
coefficients(fit) # model coefficients
confint(fit, level=0.95) # CIs for model parameters
fitted(fit) # predicted values
residuals(fit) # residuals
anova(fit) # anova table
vcov(fit) # covariance matrix for model parameters
influence(fit) # regression diagnostics
```

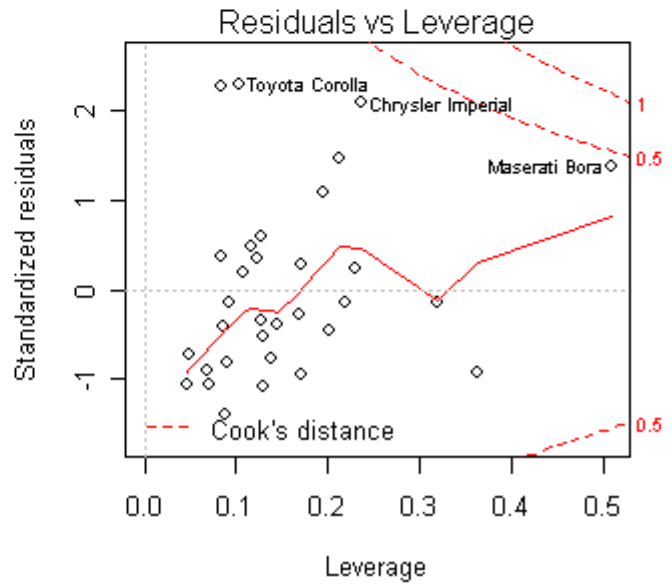
Diagnostické grafy

```
# diagnostic plots
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(fit)
```









IT ACADEMY

Porovnanie a Validácia modelov

```
# compare models
fit1 <- lm(y ~ x1 + x2 + x3 + x4, data=mydata)
fit2 <- lm(y ~ x1 + x2)
anova(fit1, fit2)
```

```
# K-fold cross-validation
library(DAAG)
cv.lm(df=mydata, fit, m=3) # 3 fold cross-validation
```

IT ACADEMY

Regresná Diagnostika

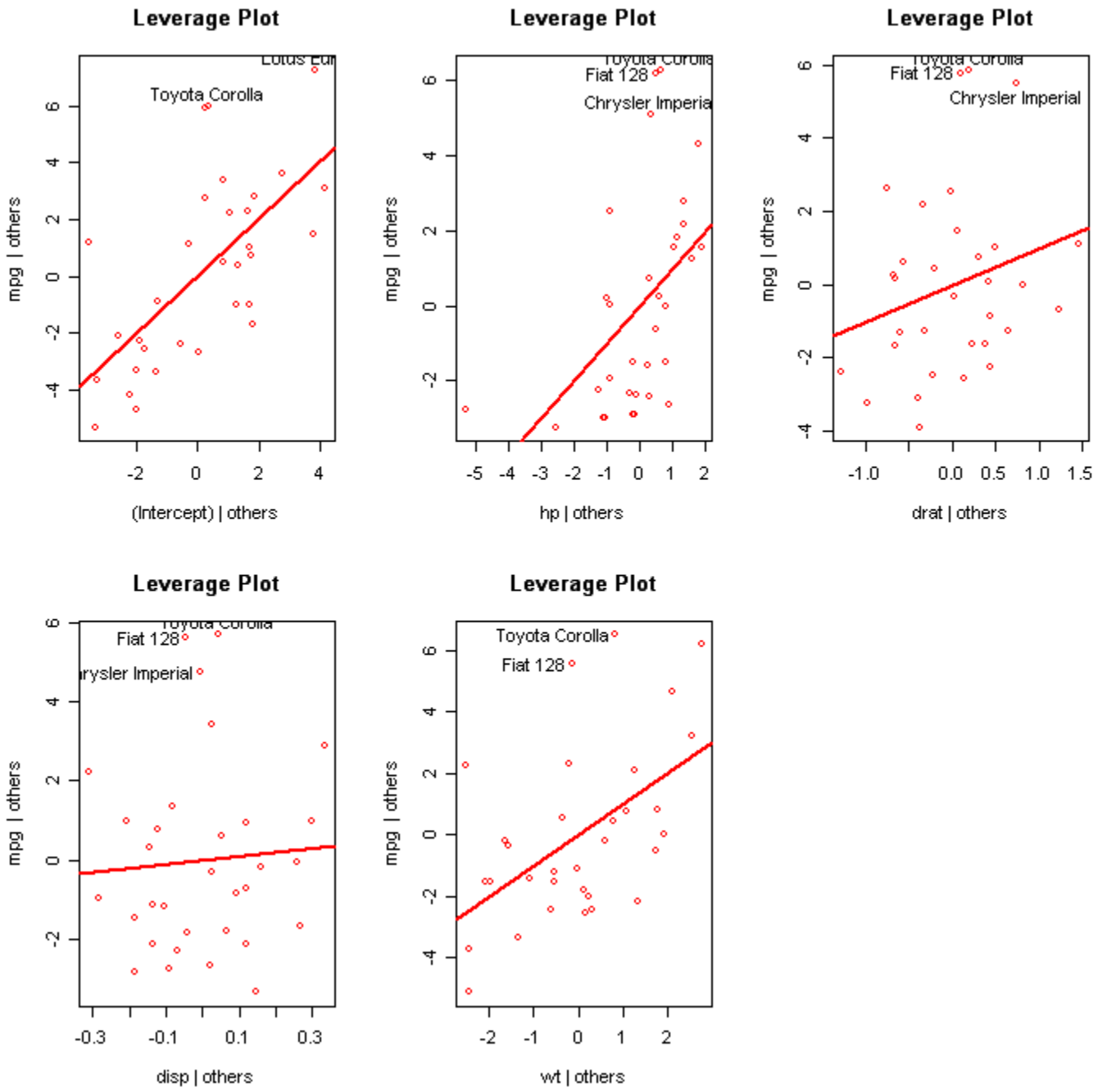
```
# Assume that we are fitting a multiple linear regression
# on the MTCARS data
library(car)
fit <- lm(mpg~disp+hp+wt+drat, data=mtcars)
```

IT ACADEMY

Instal package car

Regresná diagnostika

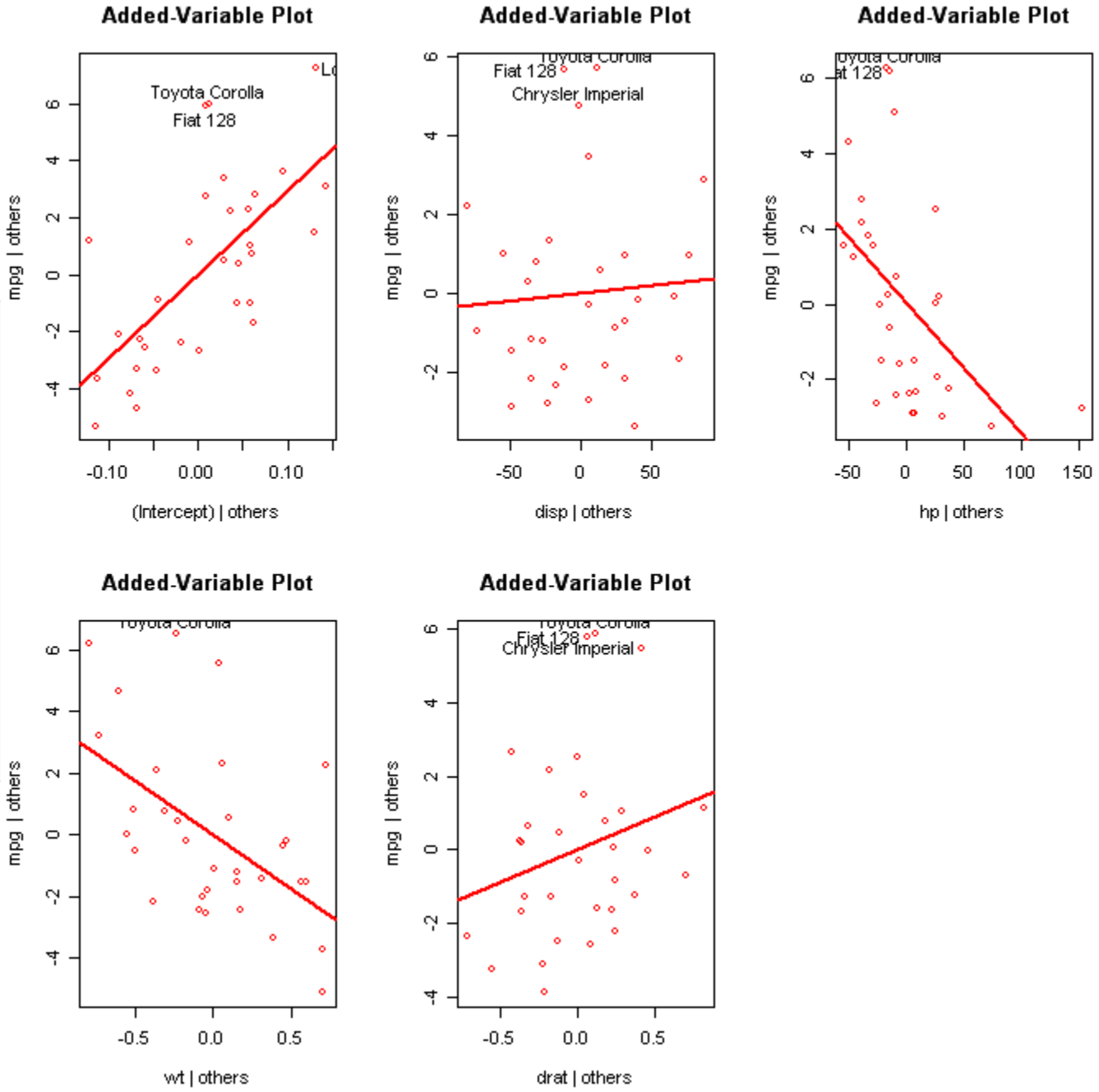
```
# Assessing Outliers
outlierTest(fit) # Bonferonni p-value for most extreme obs
qqPlot(fit, main="QQ Plot") #qq plot for studentized resid
leveragePlots(fit) # leverage plots
```



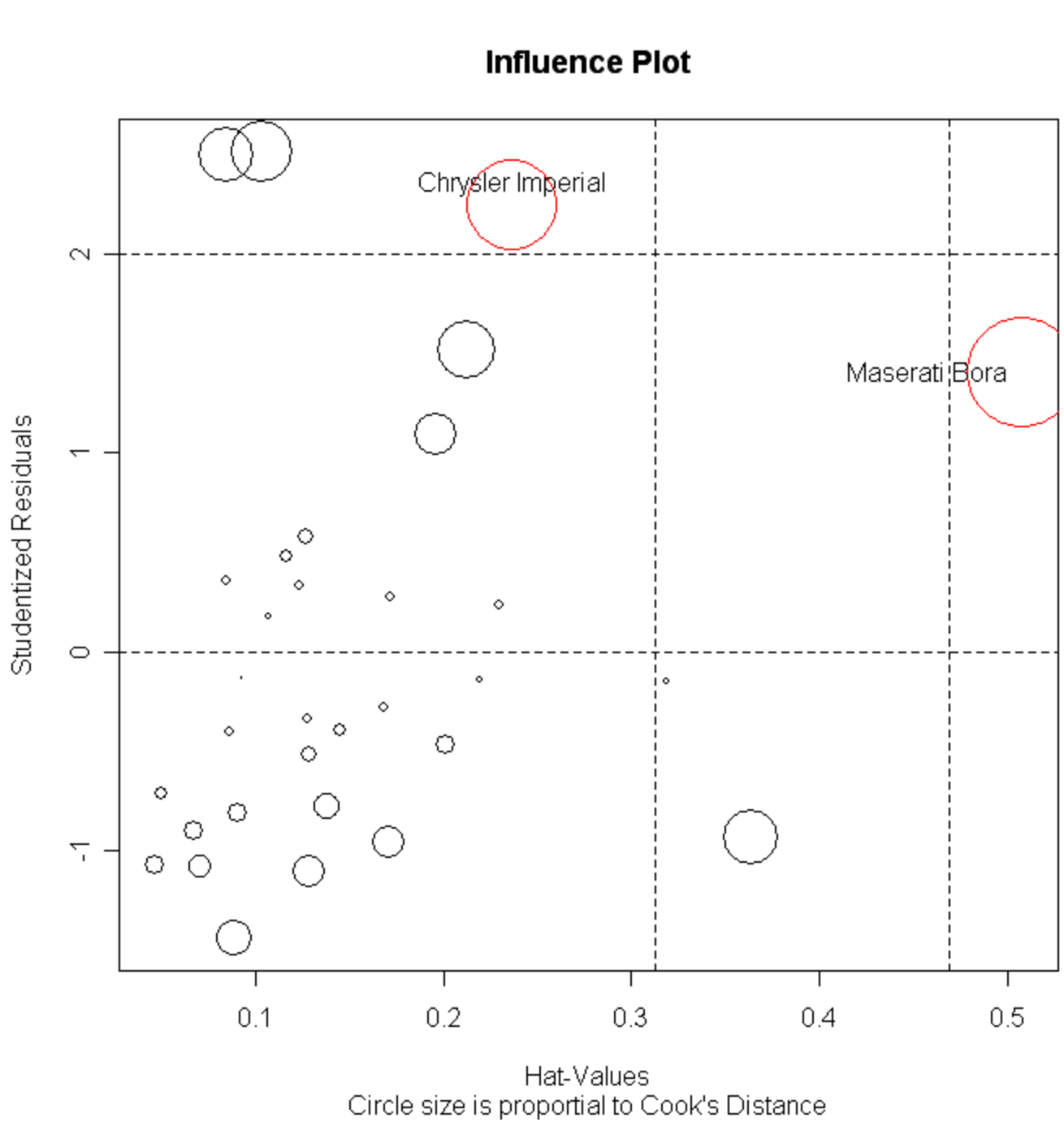
IT ACADEMY

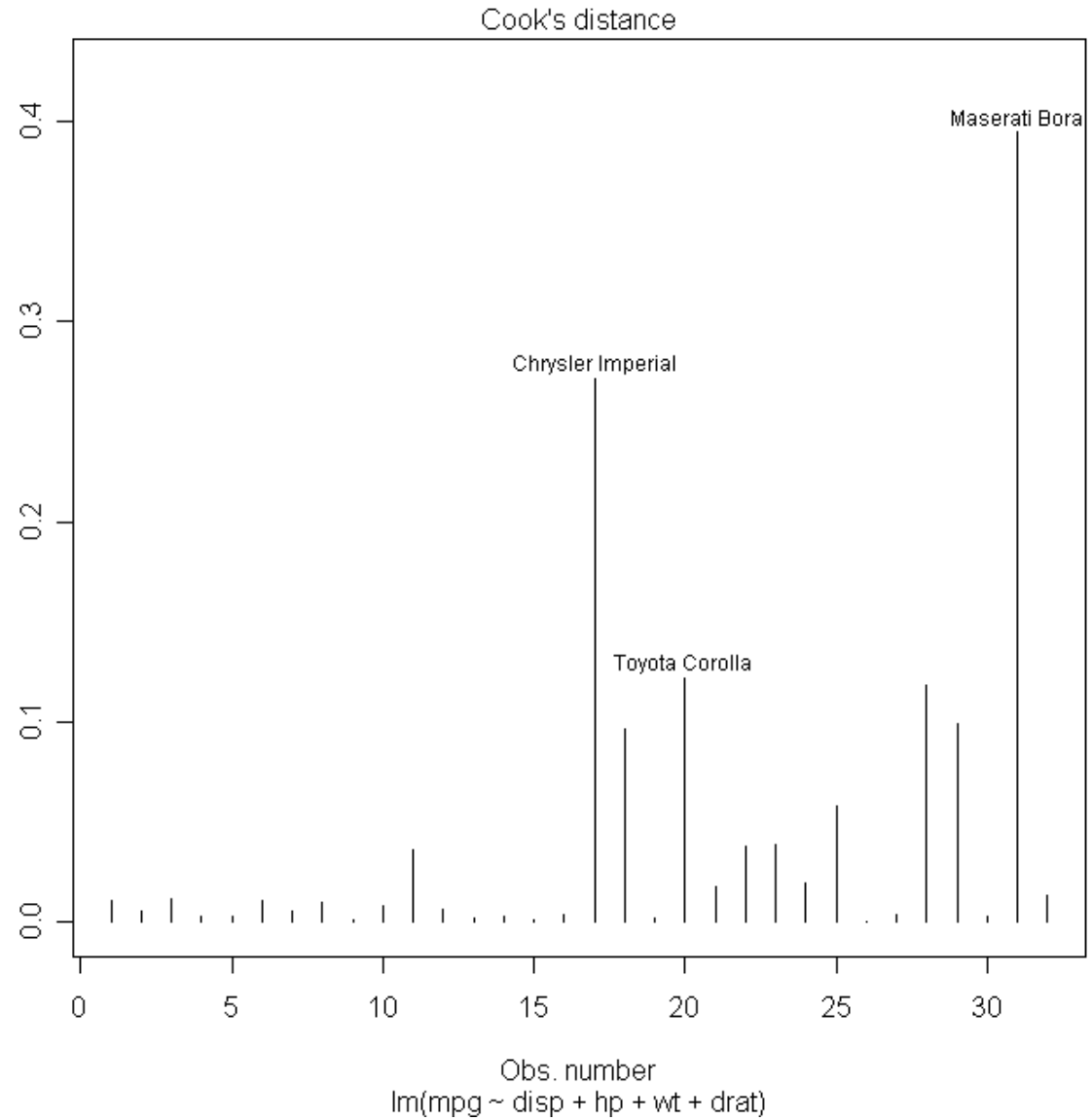
Vplyvné pozorovania

```
# Influential Observations
# added variable plots
av.Plots(fit)
# Cook's D plot
# identify D values > 4/(n-k-1)
cutoff <- 4/((nrow(mtcars)-length(fit$coefficients)-2))
plot(fit, which=4, cook.levels=cutoff)
# Influence Plot
influencePlot(fit, id.method="identify", main="Influence Plot",
sub="Circle size is proportial to Cook's Distance" )
```



IT ACADEMY

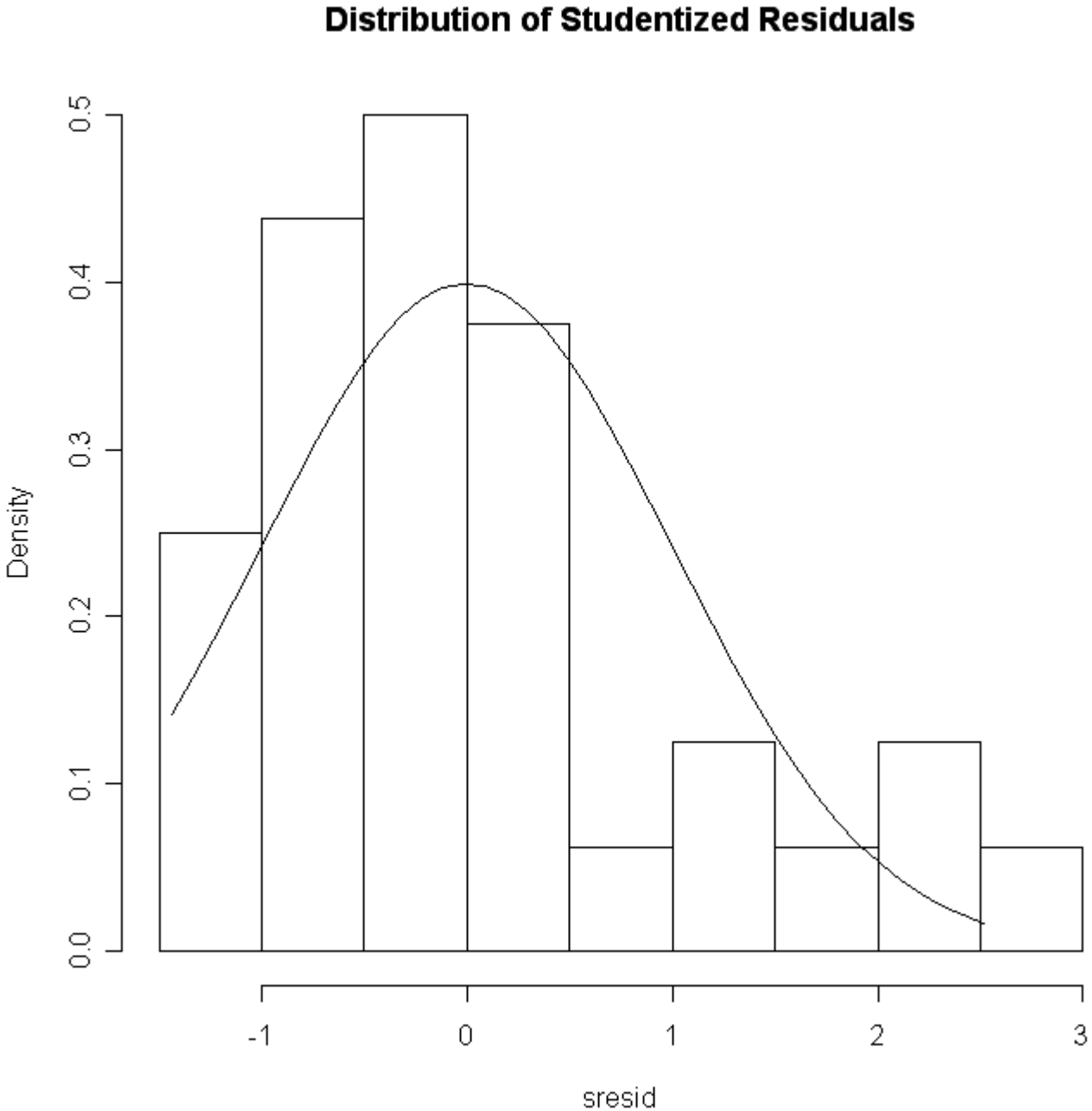

Influence Plot
Chrysler Imperial
Maserati Bora
Studentized Residuals
2
1
0
-1
0.1
0.2
0.3
0.4
0.5
Hat-Values
Circle size is proportial to Cook's Distance

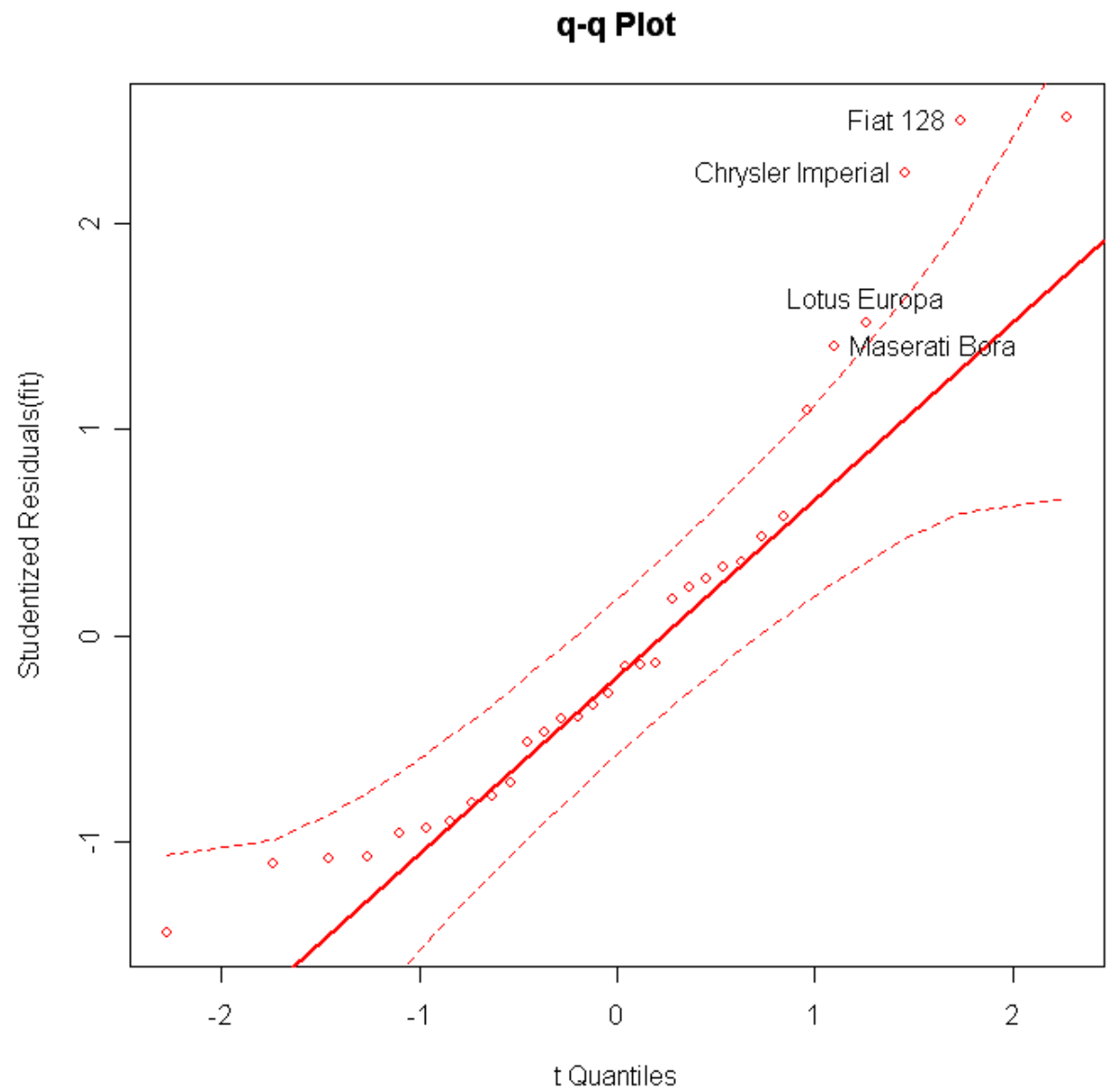

Cook's distance
Maserati Bora
Chrysler Imperial
Toyota Corolla
0.4
0.3
0.2
0.1
0.0
0
5
10
15
20
25
30
Obs. number
lm(mpg ~ disp + hp + wt + drat)

Non-normality

```
# Normality of Residuals
# qq plot for studentized resid
qqPlot(fit, main="QQ Plot")
# distribution of studentized residuals
library(MASS)
sresid <- studres(fit)
hist(sresid, freq=FALSE,
   main="Distribution of Studentized Residuals")
xfit<-seq(min(sresid),max(sresid),length=40)
yfit<-dnorm(xfit)
lines(xfit, yfit)
```

IT ACADEMY


Distribution of Studentized Residuals
Density
0.0
0.1
0.2
0.3
0.4
0.5
-1
0
1
2
3
sresid


q-q Plot
Studentized Residuals(fit)
-1
0
1
2
-2
-1
0
1
2
t Quantiles
Fiat 128
Chrysler Imperial
Lotus Europa
Maserati Bora

Skúmanie závislostí medzi premennými (javmi, znakmi)

Symbolika:

- ✓ ZP – závislá premenná Y (resp. Y=A), resp. vysvetľovaná, modelovaná, cieľová, angl. dependent variable, response variable, target
- ✓ NZP – nezávislá premenná X, (resp. A, B, ...), resp. vysvetľujúca, angl. independent variable, input variable, exploration variable

1. Skúmanie závislostí medzi premennými (javmi, znakmi) – A a B

ZP	NZP	Názov závislosti	Štat. metóda/y	Jazyk R
A	B	asociácia	• Analýza kontingenčnej tabuľky • Chi2 test závislosti (asociácie)	Describe – Table Analysis

2a. Skúmanie závislostí medzi premennými (javmi, znakmi) – Y a A

ZP	NZP	Názov záv.	Štat. metóda/y	Jazyk R
Y: Y_{A1} a Y_{A2}	**A – faktorová premenná** (Classification variable) len binárna: A_1 a A_2 k = 2		**t-test** (predtým ešte F-test): • pre nezávislé výbery • pre závislé výbery - párový	**ANOVA – t-test:** • Two Sample (F-test je súčasť výstupu) • Paired

2b. Skúmanie závislostí medzi premennými (javmi, znakmi) – Y a A

ZP	NZP	Názov záv.	Štat. metóda/y	Jazyk R
Y: Y_{A1} až Y_{Ak}	**A – faktorová premenná** (Classification variable) množná: A_1 až A_k k > 2		**ANOVA:** 1. a) jednofaktorová b) viacfaktorová 2. a) parametrická b) neparametrická 3. a) vyvážená b) nevyvážená	**ANOVA:** 1. a) len 1-faktorová (One-Way ANOVA) b) 1 aj viacfaktorová (Linear Models) 2. a) parametrická (One-Way ANOVA, Linear Models) b) neparametrická (Nonparametric One-Way ANOVA) 3. a) vyvážená (One-Way ANOVA) b) nevyvážená (Linear Models)

3a. Skúmanie závislostí medzi premennými (javmi, znakmi) – Y a X

ZP	NZP	Názov záv.	Štat. metóda/y	Jazyk R
Y	X	**Korelácia jednoduchá**	**1. výpočet mier závislosti (korelácie):** • cov(x,y) ϵ (-∞,∞) • r(x,y) ϵ (-1,1) **2. odhad regresného modelu** jednoduchého: • lineárneho alebo • nelineárneho	1. Multivariate – **Correlations** 2. **Regression:** • Linear Regression • Nonlinear Regression

3b. Skúmanie závislostí medzi premennými (javmi, znakmi) – Y a X

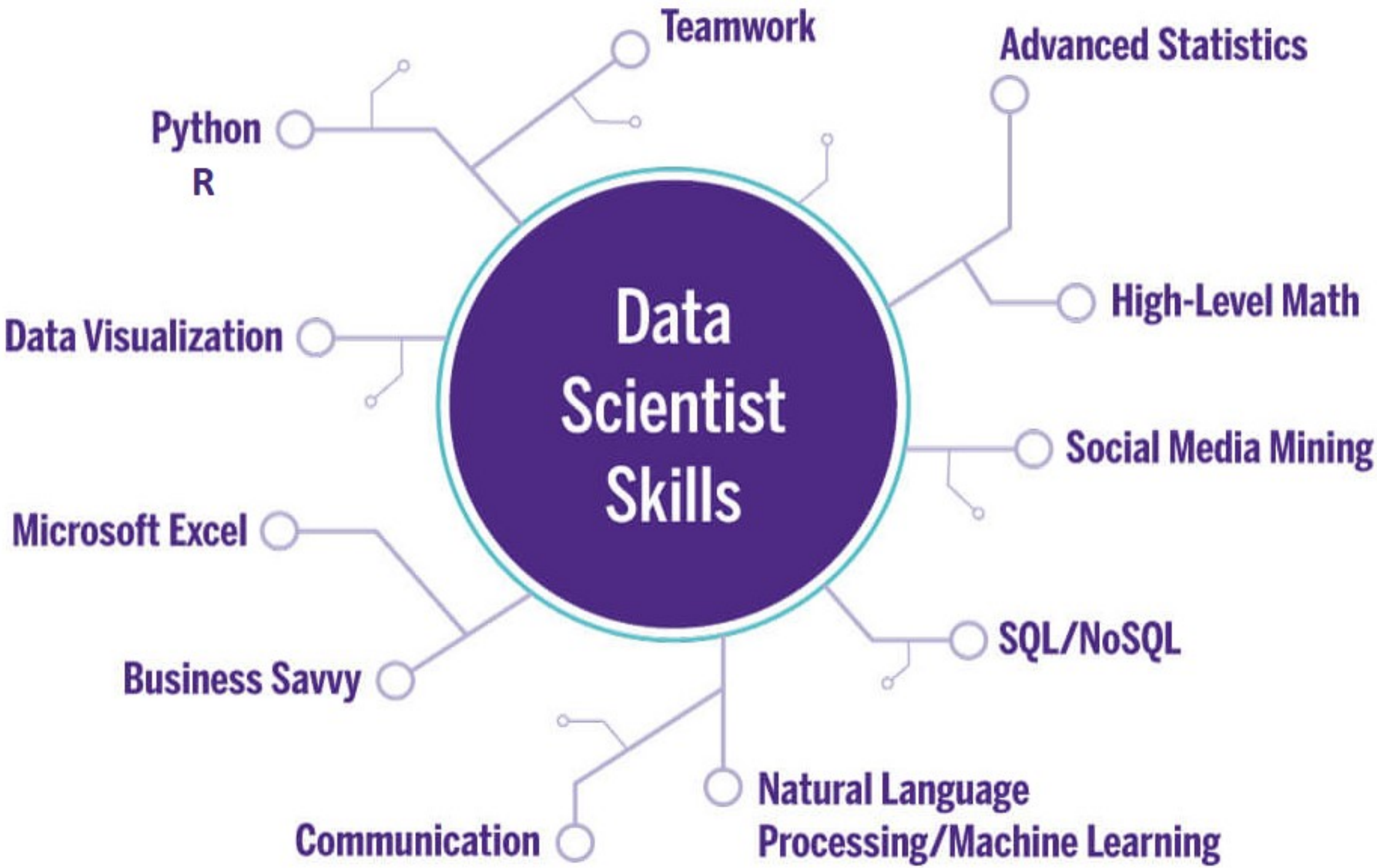
ZP	NZP	Názov záv.	Štat. metóda/y	Jazyk R
Y	X_1, X_2, X_3, ..., X_k	**Korelácia viacnásobná**	**1. výpočet mier závislosti (korelácie):** • kovariančná matica (**S**) • korelačná matica (**R**) **2. odhad regresného modelu viacnásobného:** • lineárneho alebo • nelineárneho	1. Multivariate – **Correlations** 2. **Regression:** • Linear Regression • Nonlinear Regression

4. Skúmanie závislostí medzi premennými (javmi, znakmi) – Y a t (čas)

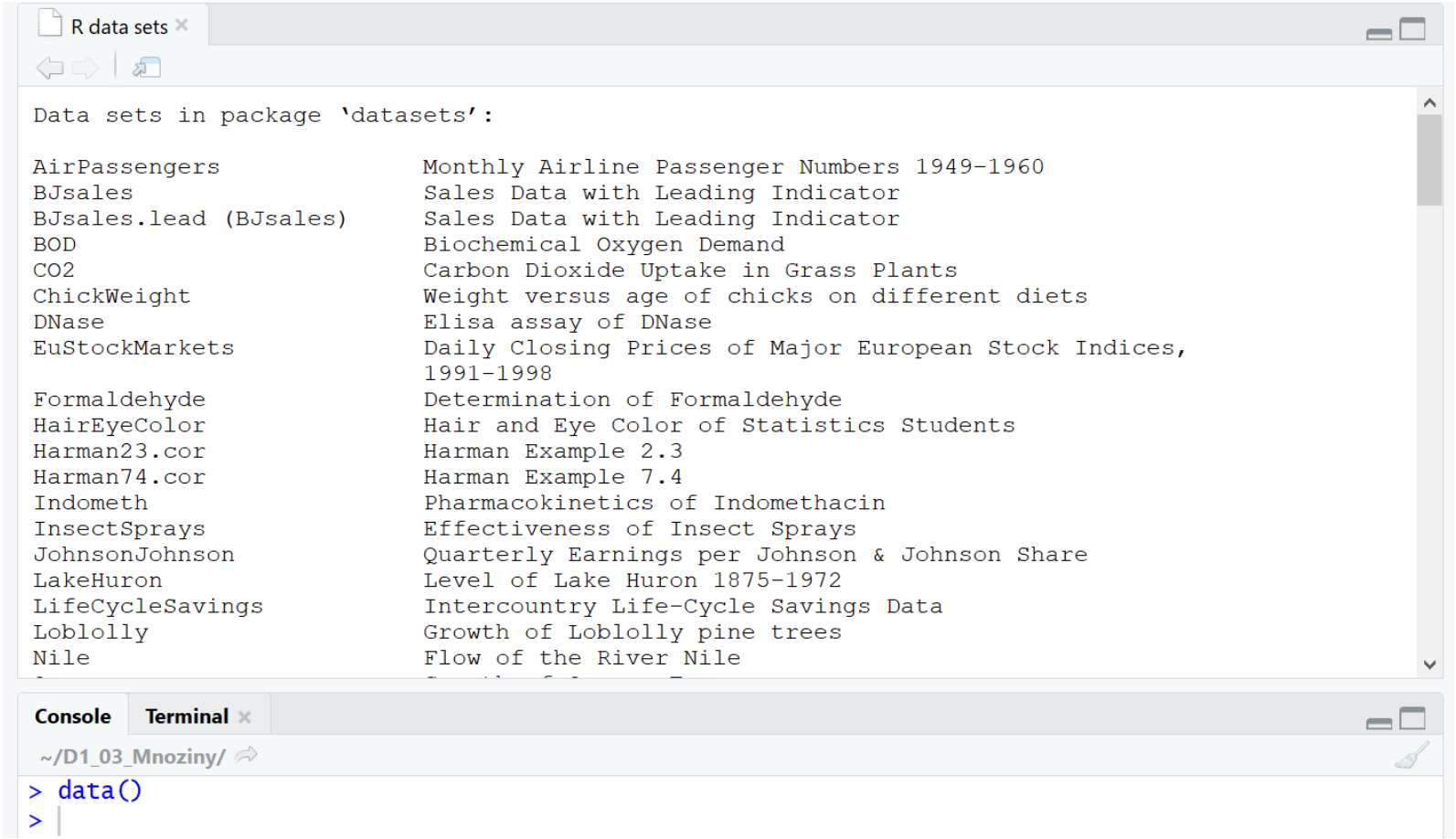
ZP	NZP	Názov záv.	Štat. metóda/y	Jazyk R
Y	t (čas)		Analýza časových radov	Time Series Analysis

5. Skúmanie závislostí medzi premennými (javmi, znakmi) – $Y=A$ a X_i, B_j

ZP	NZP	Názov záv.	Štat. metóda/y	Jazyk R
$Y=A$ $P(Y=A_1/X_i, B_j)$	$X_1, X_2, X_3, ..., X_k$ $B_1, B_2, B_3, ..., B_q$		Logistická regresia	Regression – Logistic Regression


Data Scientist Skills
Teamwork
Advanced Statistics
Python
R
High-Level Math
Data Visualization
Social Media Mining
Microsoft Excel
SQL/NoSQL
Business Savvy
Natural Language Processing/Machine Learning
Communication

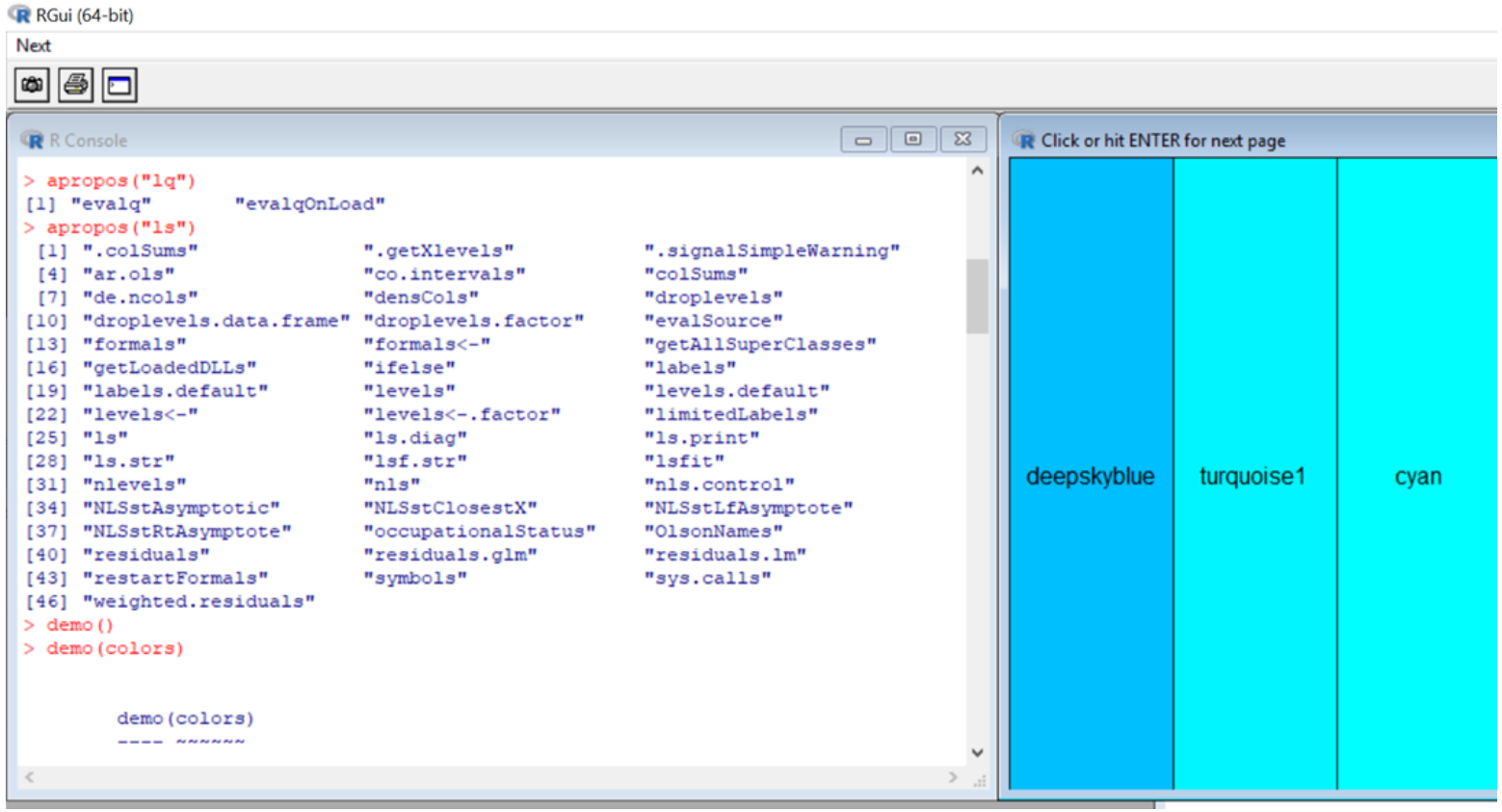
Zabudované Datasety



Data()

Apropos a Demo

- Príkaz **apropos ("nazov")** zobrazíme príbuzné príkazy k príkazu názov, pre príklady k nejakej téme použijeme príkaz example (tema)
- K dispozícii sú tiež **PDF manuály**, ktoré obsahujú základné manuály a referencie o funkciách
- **Príkaz demo()** vypíše zoznam všetkých dostupných demonštrácií.



Zlúčenie/Merging Údajov

```
# merge two data frames by ID
total <- merge(data frameA,data frameB,by="ID")
```

```
# merge two data frames by ID and Country
total <- merge(data frameA,data frameB,by=c("ID","Country"))
```

```
total <- rbind(data frameA, data frameB)
```

Čo sa Oplatí Prečítať?

Slovensko a Česko

- Albatrosmedia
- Kopp
- Grada
- Wolters Kluwer
- BEN
- Veda

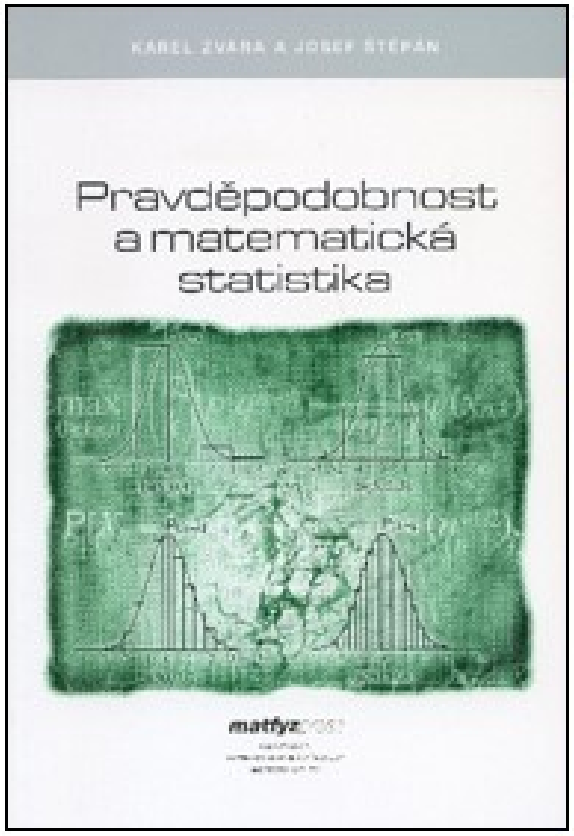
Zahraničie

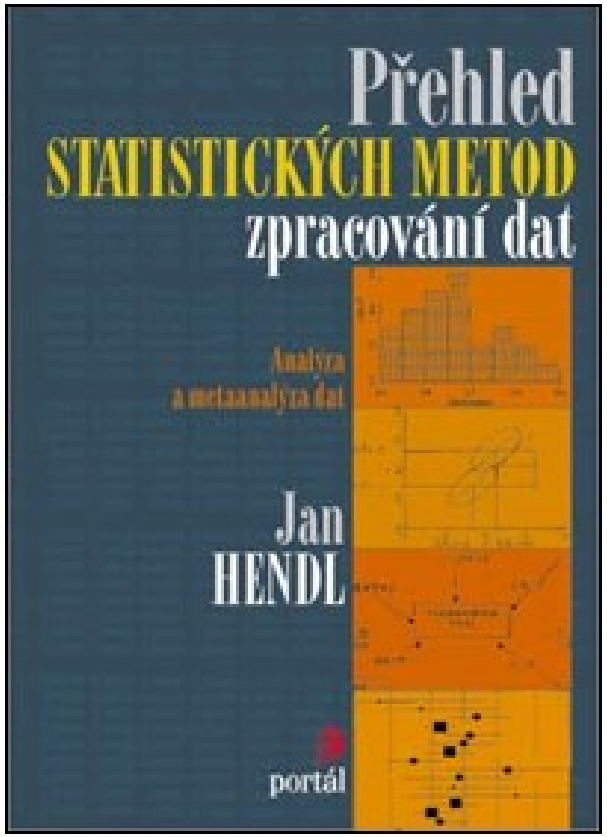
- O'Reilly
- Manning
- Packt
- Apress
- Wiley
- No Starch Press

YouTube Tutorialy

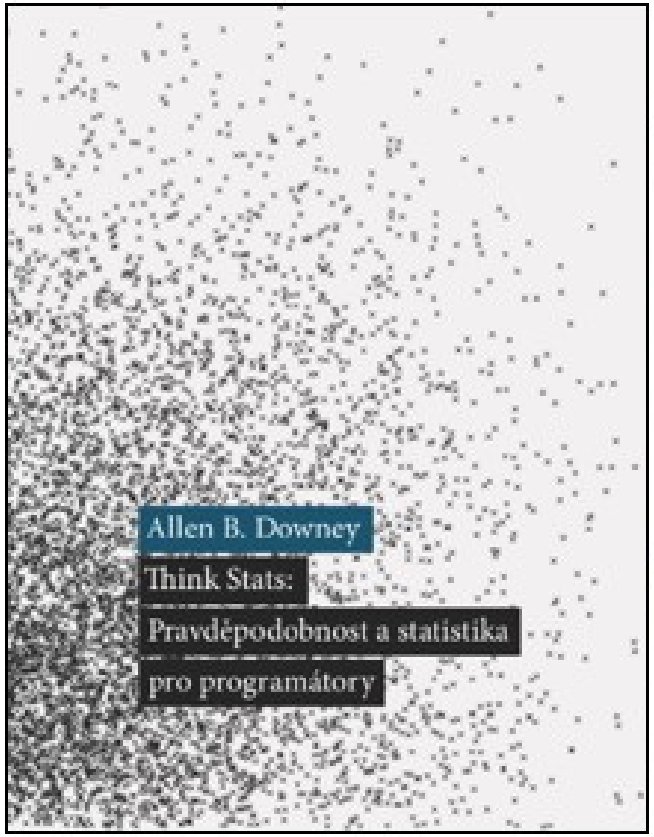
- IT Academy

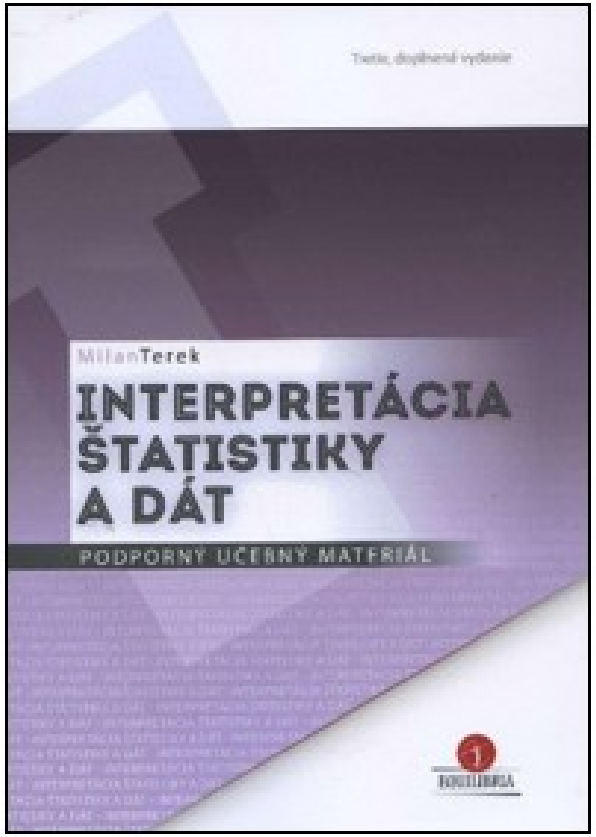
IT Academy

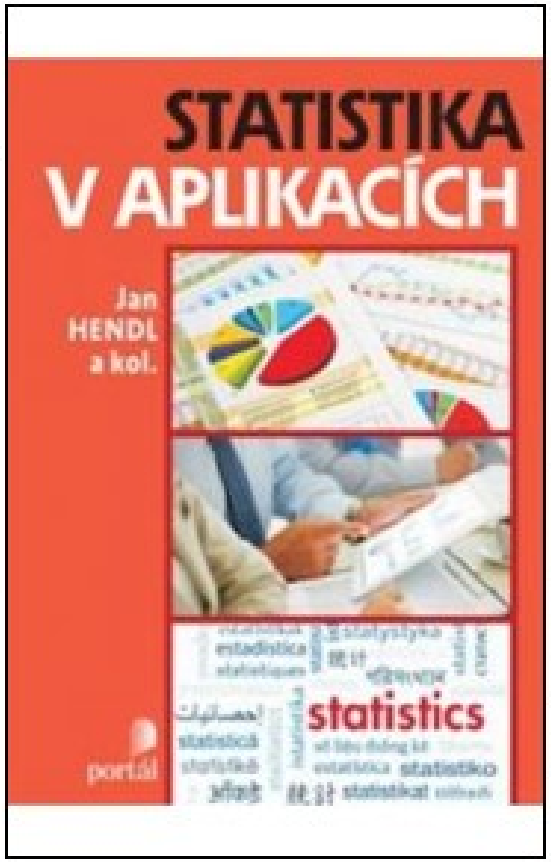


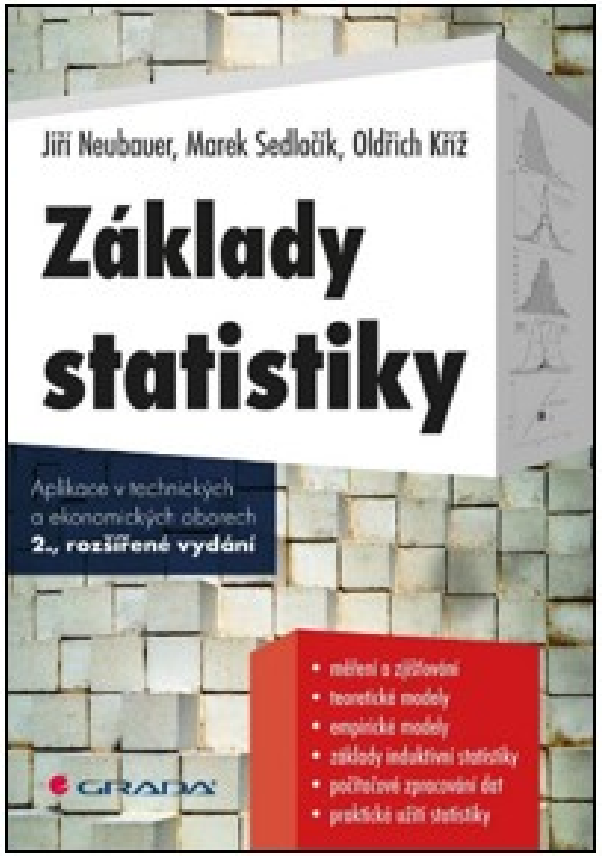


Čo sa oplatí/neoplatí prečítať SK/CZ?

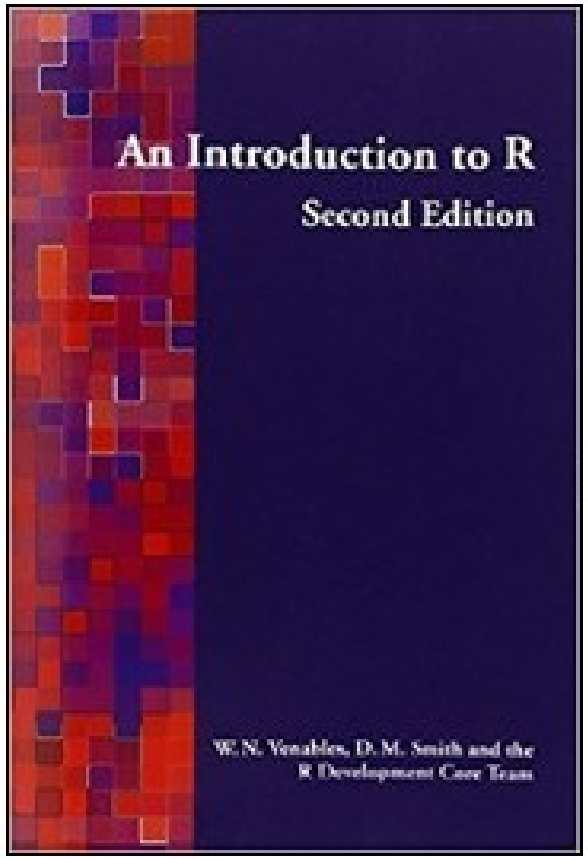


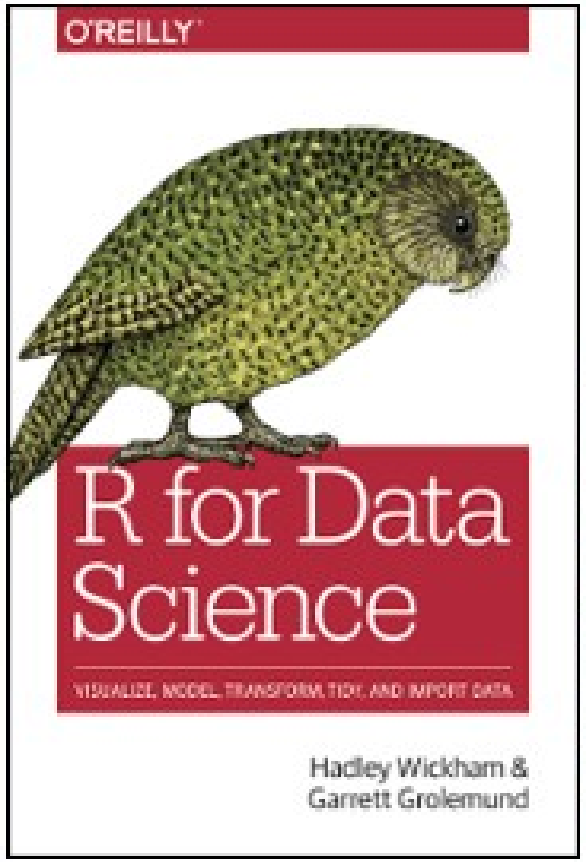


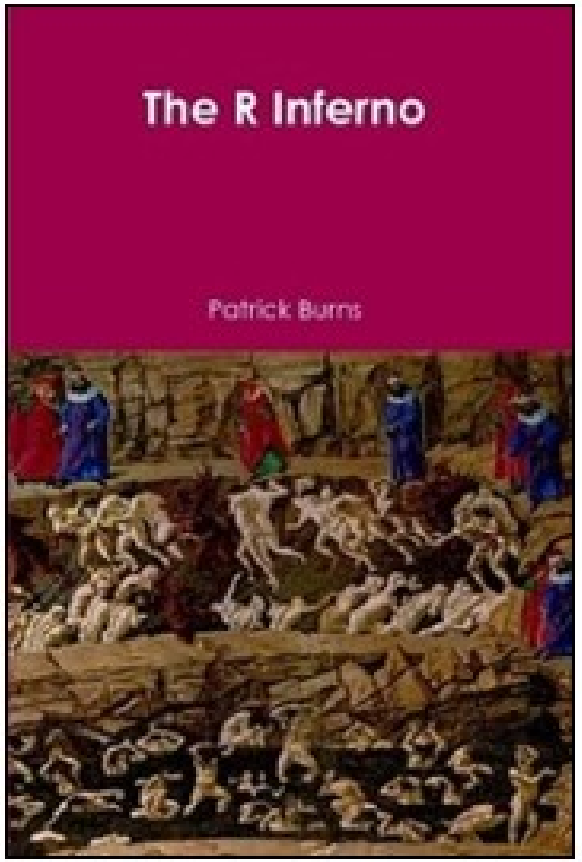


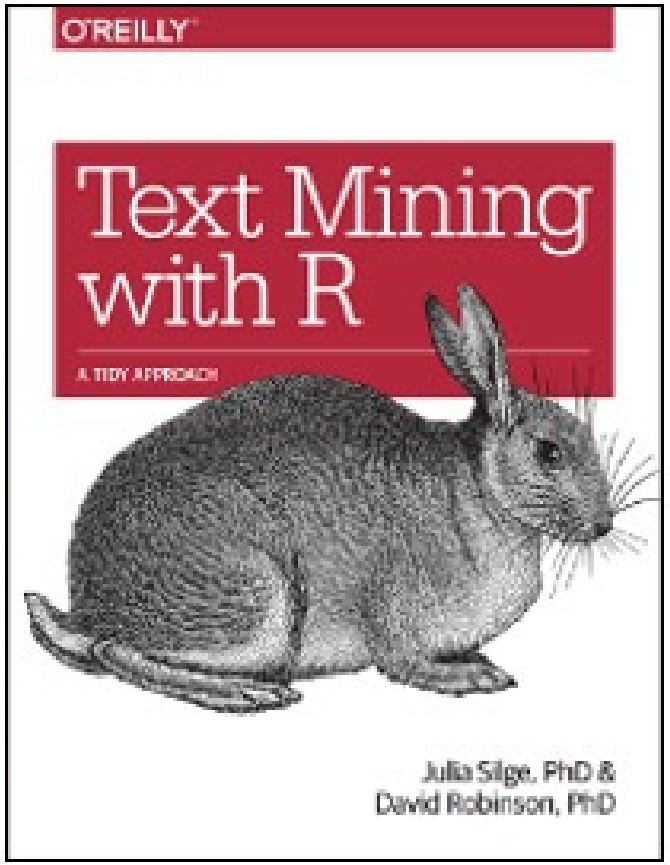


Čo sa oplatí/neoplatí prečítať SK/CZ?

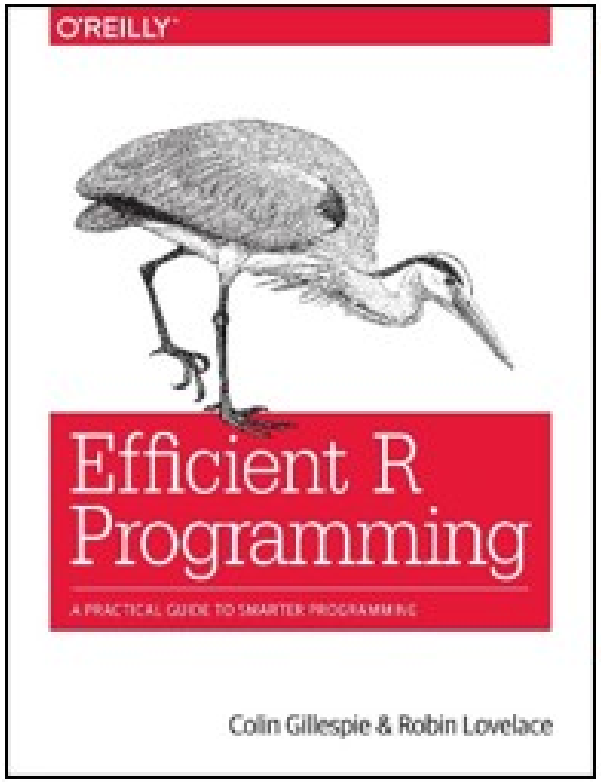


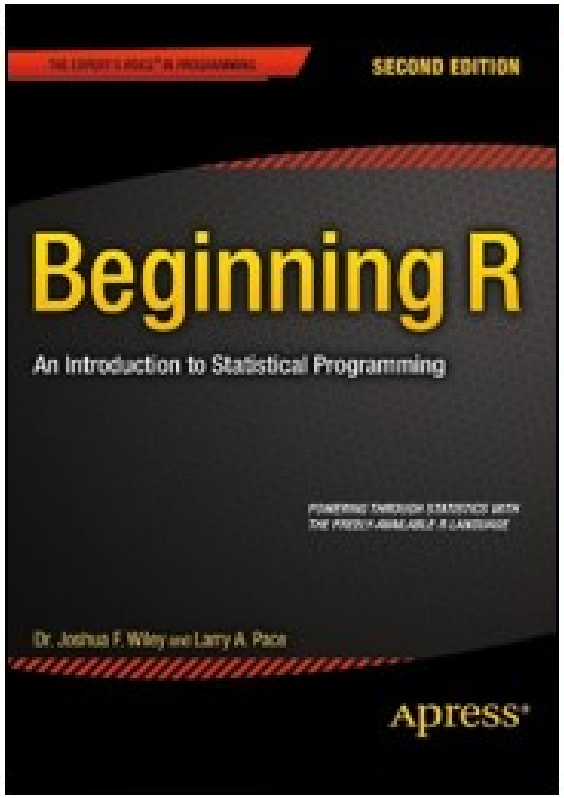


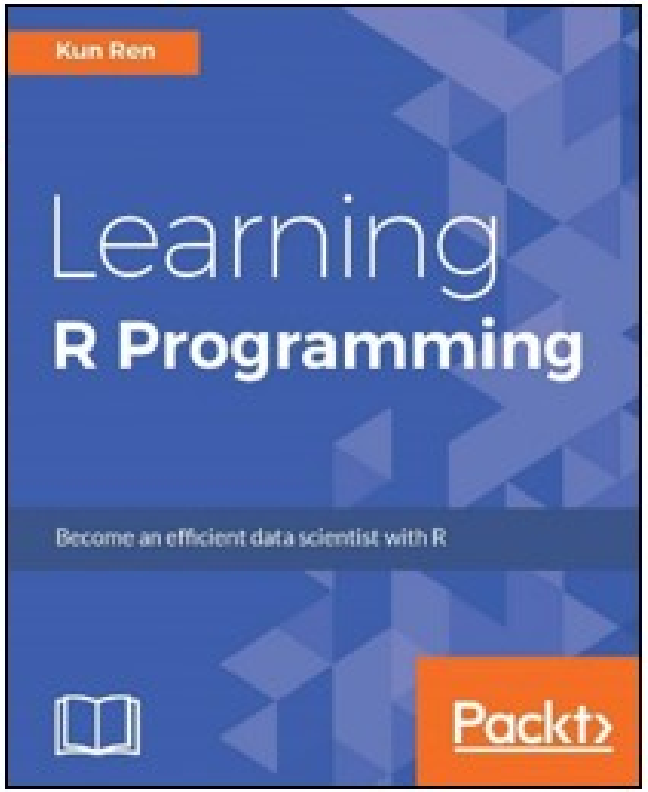


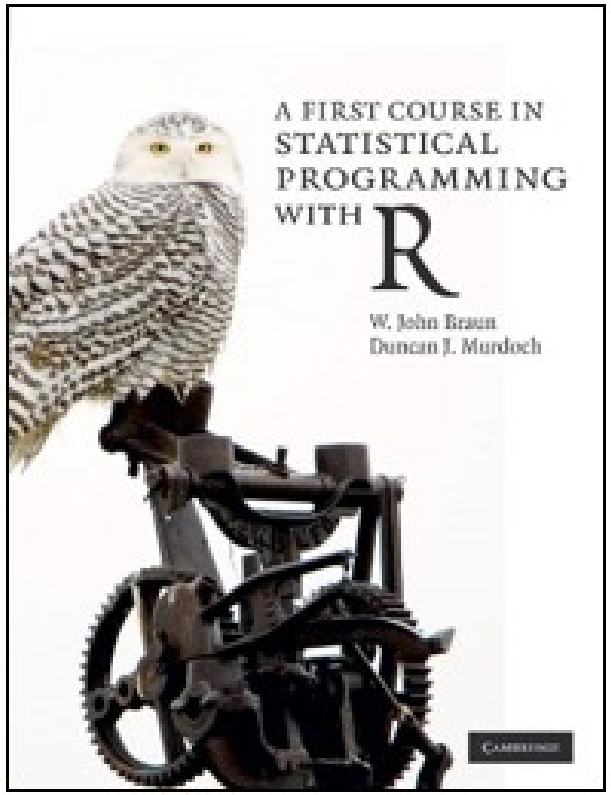


Čo sa oplatí/neoplatí prečítať EN









Čo sa oplatí/neoplatí prečítať EN

Home

PUBLIC
Questions
Tags
Users
Companies

COLLECTIVES
Explore Collectives

TEAMS
Create free Team

Looking for your Teams?

Tags

A tag is a keyword or label that categorizes your question with other, similar questions. Using the right tags makes it easier for others to find and answer your question.

Show all tag synonyms

r

Popular | Name | New

r 489303 questions 113 asked today, 800 this week	javascript 2493694 questions401 asked today, 2384 this week	android 1403892 questions 159 asked today, 856 this week
jquery 1033870 questions 46 asked today, 229 this week	reactjs 454618 questions 219 asked today, 1319 this week	arrays 412869 questions 34 asked today, 227 this week
ruby-on-rails 336768 questions 16 asked today, 102 this week	sql-server 329798 questions 45 asked today, 255 this week	angular 295326 questions 72 asked today, 444 this week
angularjs 262780 questions 7 asked today, 27 this week	regex 257806 questions 25 asked today, 162 this week	ruby 227846 questions 11 asked today, 77 this week

Dáta, analýzy a štatistika

Verejná skupina · 1,5 tis. členov

+ Pozvať

Informácie | Diskusia | Členovia | Podujatia | Médiá | Súbory

Create a public post...

Fotka/video | Označiť ľudí | Pocit/aktivita

Nová aktivita

Pavol Skapik zdieľa odkaz.
Administrátor · 29. januára o 19:18 ·

Marek Krajčí vyhlásil, že po dopočítaní oneskorených údajov sa podiel pozitívnych testov v Bratislave dostal až na 1,3 percenta

BSK: Výsledky testovania v mestských častiach Bratislavy hovoria o podiele pozitívnych 0,81%, resp. všetko čo je k dispozícii je 0,88.

Čím si prosím vysvetľujete rozdiel? Ďakujem... **Zobraziť viac**

Informácie

Skupina je len súkromným projektom, v žiadnom prípade zverejnené názory a príspevky nie sú oficiálnymi výstupmi.

Verejná
Členov skupiny a ich príspevky bude vidieť ktokoľvek.

Viditeľná
Túto skupinu nájde ktokoľvek

Bratislava, Slovakia

Skupina o Všeobecné

Populárne témy v príspevkoch

census (1) | demograf...

Data Analysts, Data Engineers & Data Scientists - Czech&Slovak Group

Verejná skupina · 2,3 tis. členov

+ Pozvať

Informácie | Diskusia | Oznámenia | Miestnosti | Témy | Členovia | Podujatia | Médiá

Create a public post...

Fotka/video | Označiť ľudí | Pocit/aktivita

Oznámenie · 1

Zobraziť všetko

Vojta Roček zdieľa odkaz.
Administrátor · 21. apríla 2017 · Prague, Praha, Česká republika ·

Zajimave veci, na ktere jste prisli a chcete je posdilet. Do popisu klidne piste, proc vam to pripada zajimave. Otazky, odpovedi, rady. Pridejte dalsi datalidi, a ptejte se na co chcete! (Ale až po prostudování https://www.hash.cz/inferno/otazky.html).

Nabidky a poptavky po praci jsou v sesterske groupe - https://www.facebook.com/groups/1788236724824404/

Pokud se chystate mistnich 1600 lidi oslovit zdarma s nabidkou na svou komercni akci (spam), tak to udelejte to dobre (j... **Zobraziť viac**

Zobraziť preklad

Informácie

Dataflow.cz Dataflow
Twitter: https://twitter.com/dataflowcz
Web: http://dataflow.cz... **Zobraziť viac**

Verejná
Členov skupiny a ich príspevky bude vidieť ktokoľvek.

Viditeľná
Túto skupinu nájde ktokoľvek

Prague, Czech Republic

Skupina o Všeobecné

Nedáve médiá


Čo ti odporúčam si pozrieť?

1. https://google.github.io/styleguide/Rguide.xml
2. https://www.statmethods.net/input/exportingdata.html
3. http://www.sr.bham.ac.uk/~ajrs/R/r-function_list.html
4. https://www.statmethods.net/management/index.html
5. https://cran.r-project.org/web/packages/stringr/vignettes/regular-expressions.html

Šup do záložiek

Najdôležitejšie Klávesové Skratky

Práca s IDE

- Ctrl + D — Delete zmaž riadok
- **Ctrl + Space** — Asistent kódu
- **Ctrl + /** — Komentáre
- Ctrl + A — Označ všetko
- **Alt + /** — Dokonči slovo
- Ctrl + F — Hľadanie a náhrady
- Ctrl + Shift + F — Kompakt režim
- Ctrl + Shift + S — Ulož všetko

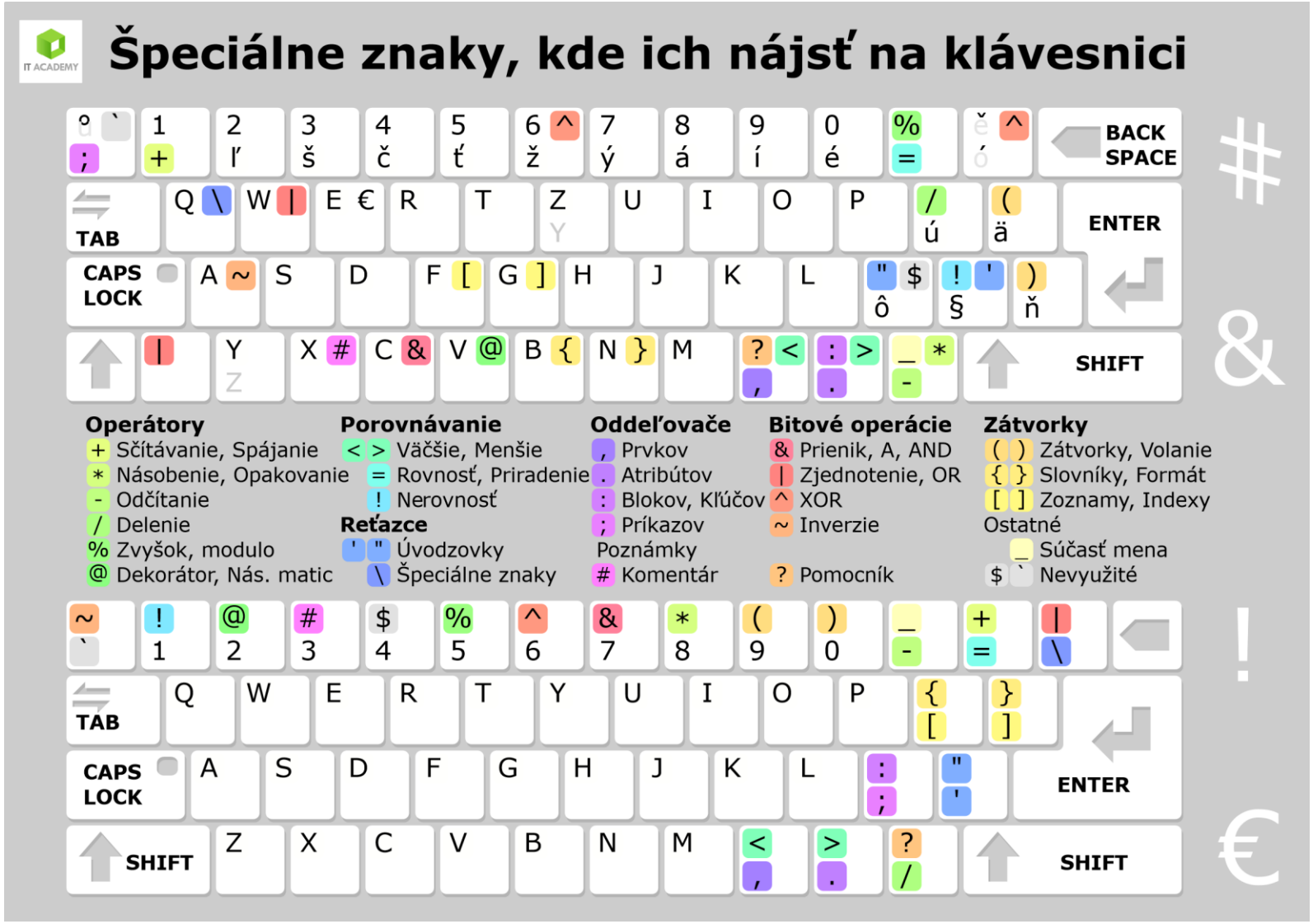
Práca s Browserom

- Ctrl + T — Vytvor nový tab
- Ctrl + W — Zatvor aktuálny tab
- Ctrl + Shift + W — Zatvor všetky taby
- **Ctrl + Shift + T** — Otvor posledný tab
- Ctrl + Shift + J/F12 — Web console
- **F11** — Fullscreen

IT ACADEMY

F5 nie je Spustenie, ale Rehresh

Efektívne Používanie Klávesnice



Ako sa s nami spojiť?

ADRESA: IT Academy, s. r. o.

Budova KOLOSEO prízemie
Tomášikova 50/A

831 04 Bratislava

WEB: www.it-academy.sk

E-MAIL: info@it-academy.sk

TELEFÓN: +421 917 095 406

Ako vieme pomôcť?

#Copywriting #Školenia #Zamestnanci #Pomáhame #Rast

#Projekty #Certifikácie #Kurzy #Tréningy #Vzdelávanie

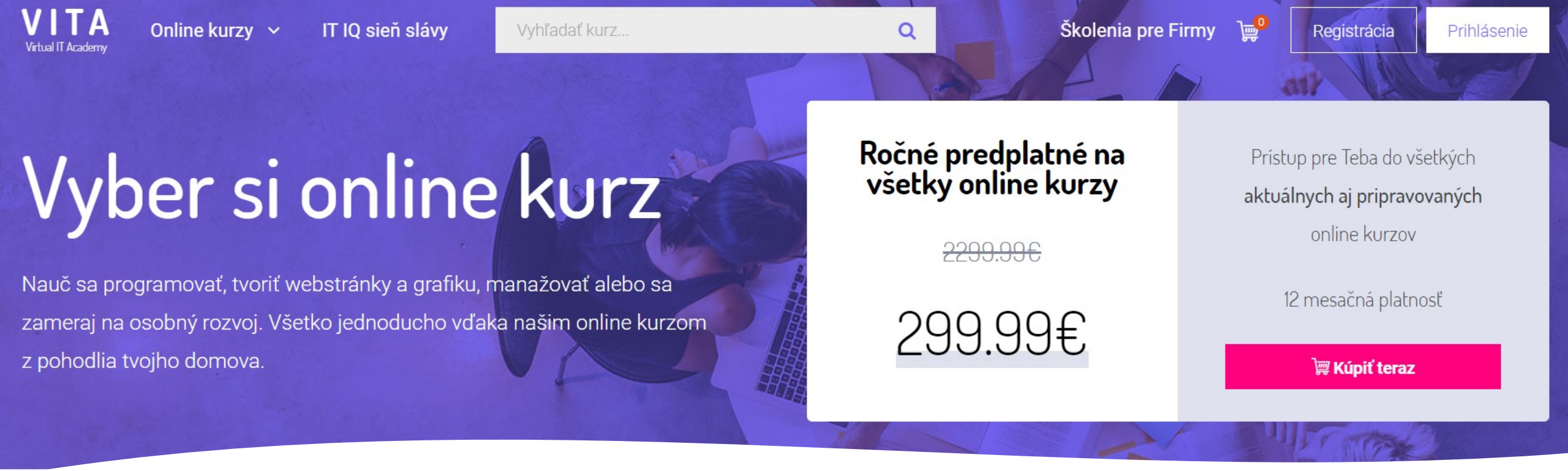
#PPC Kampane #Elearning #Mentoring #Konzultácie #Online

#Programovanie #Vývoj #Marketing #Reklama #Prenájom Techniky

Mrkni na náš YouTube kanál a daj odber

➡ WWW.YOUTUBE.COM/C/IT-ACADEMYSK ⬅

IT ACADEMY



Zadarmo

A. Ďalšie kurzy NSPC
B. Grow with Google

1. NESTRAŤ PRÁCU (Zamestnaní, SZČO, na materskej)
2. KOMPAS/REPAS (Nezamestnaní)
3. YouTube Kanál (IT Academy)

Platené

1. Moje kurzy na www.vita.sk