

# R Programming Language

Peter Andrejko, Marek Dráb, Dávid Gavenda, Marek Klimo



# Roger D. Peng

“I don’t think anyone actually believes that R is designed to make *everyone* happy. For me, R does about 99% of the things I need to do, but sadly, when I need to order a pizza, I still have to pick up the telephone.”



# Čo je R?

- Je programovacie prostredie určené na štatistickú analýzu dát a ich zobrazenie
- Ide o implementáciu jazyka S pod slobodnou licenciou
- Poskytuje širokú škálu štatistických a grafických techník vrátane lineárneho a nelineárneho modelovanie, klasických štatistických testov, analýzy časových radov, zhlukovania
- Rozširuje sa pomocou knižníc „packages“ (15 325 k 1.1.2020)
- Používa sa v príkazovom riadku, existuje niekoľko grafických rozhraní – Jupyter, RStudio, RKWard, R Commander, Visual Studio Code

# Kompatibilita

- Verzia 1.0.0 vydaná v roku 2000
- Operačné systémy: Linux, Windows, Apple
- Programovacie jazyky: C, C++, Fortran, Java, Python
- Vlastný LaTeX dokumentový formát (online, tlačivo)

# R Core team

- Založený v 1997
- Doug Bates, Peter Dalgaard, Robert Gentleman, Kurt Hornik, Ross Ihaka, Friedrich Leisch, Thomas Lumley, Martin Mächler, Paul Murrell, Heiner Schwarte a Luke Tierney.
- Práca čisto dobrovoľná
- V 2003 vytvorené *R Foundation* pre financovanie projektu

# Jazyk S- predchodca

- Vytvorený v 1976 v Bell Laboratories
- Dátová analýza
- Knižnica pre fortran
- 1988 prepísaný do C
- Posledná verzia v 1998: S-PLUS
- Len komerčná verzia

# História

- Prvé vydanie – August 1993 - Robert Gentleman a Ross Ihaka
- 1995- vydanie pod licenciou Free Software License GNU
- Verzia 0.16 – posledná alfa verzia 1.4.1997
- Verzia 1.0.0 – prvá stabilná verzia na komerčné využitie 29.2.2000
- 1997- R-Core team

Náhrada za S

Podobná syntax

Odlišná sémantika (bližšie k Scheme)

### **Freeware**

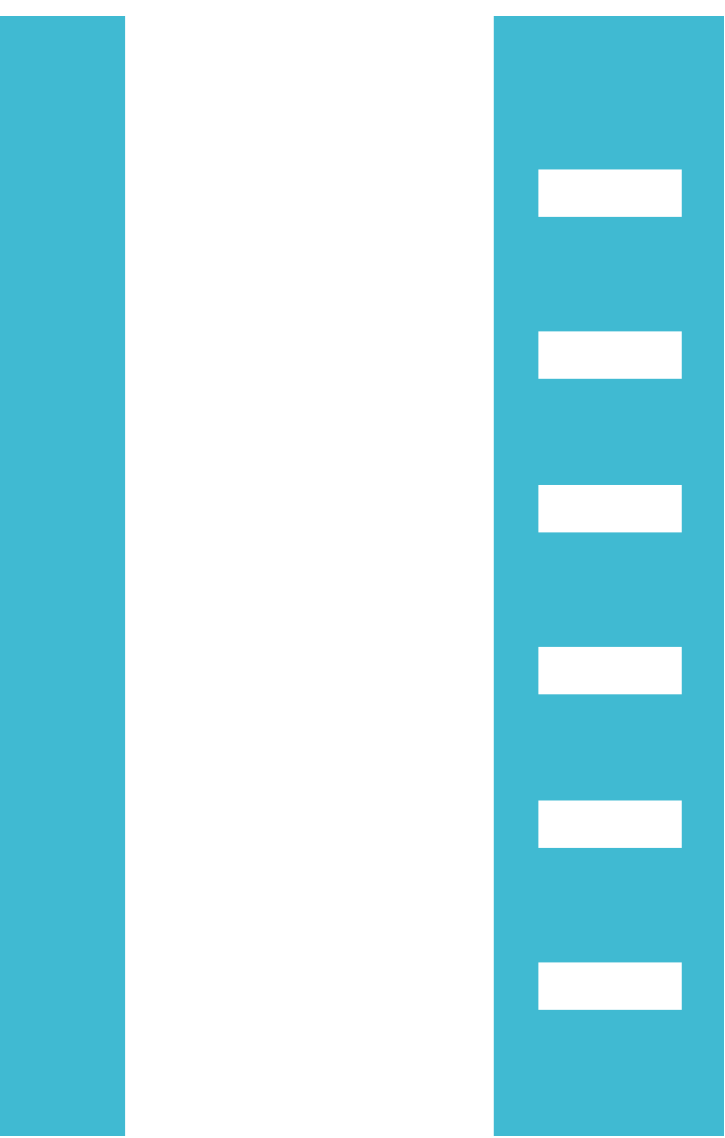
Rozdelenie na 2 časti: základné balíčky (utils, stats, datasets, graphics) a všetky ostatné

Starý jazyk: zakladá na 50-ročnej technológii, problem s pamäťou





- Open source
- Vzorová podpora pre prácu s údajmi
- Kvalitné vykresľovanie grafov
- Kompatibilita
- Nezávislosť softvéru – umožňuje closs-platform programovanie
- Machine learning – klasifikácia, regresia
- Štatistika

- 
- Slabý základ
  - Data handling
  - Bezpečnosť
  - Komplikovaný jazyk
  - Rýchlosť jazyka
  - Algoritmy v rôznych packages

## Oblasti využitia

Finančné služby (Banky, Poistovne)

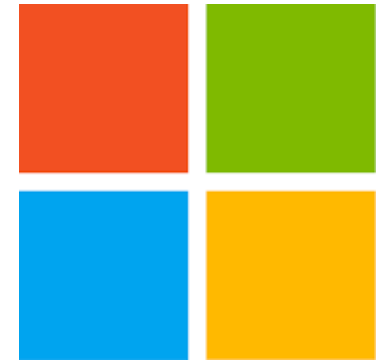
Akademické inštitúcie

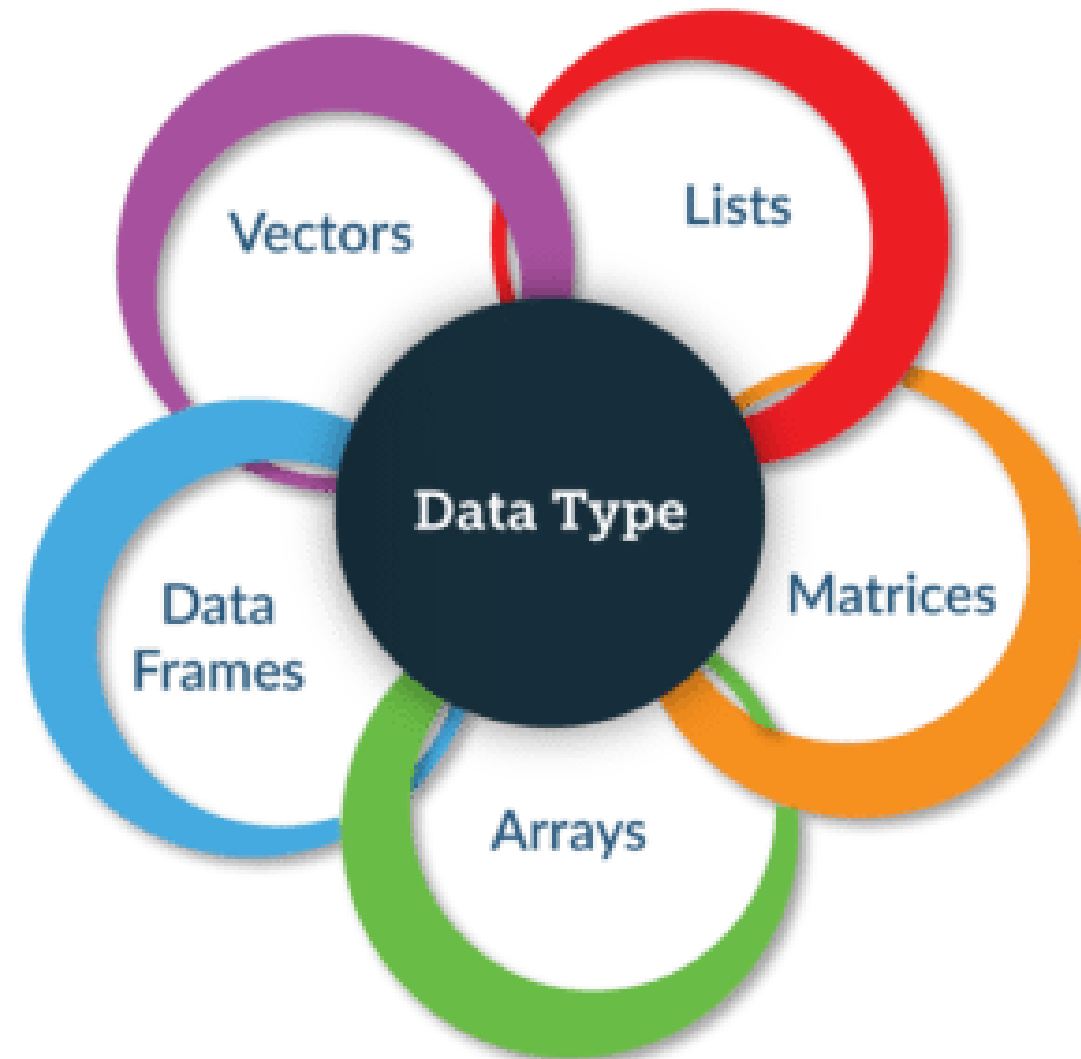
Vládne inštitúcie

Sociálne media

Zdravotníctvo

Firmy





# Syntax- základy



- Vstupné hodnoty

```
> x <- 1
> print(x)
[1] 1
> x
[1] 1
> msg <- "hello"
```

- Komentáre

```
x <- ## Incomplete expression
```

## Relational

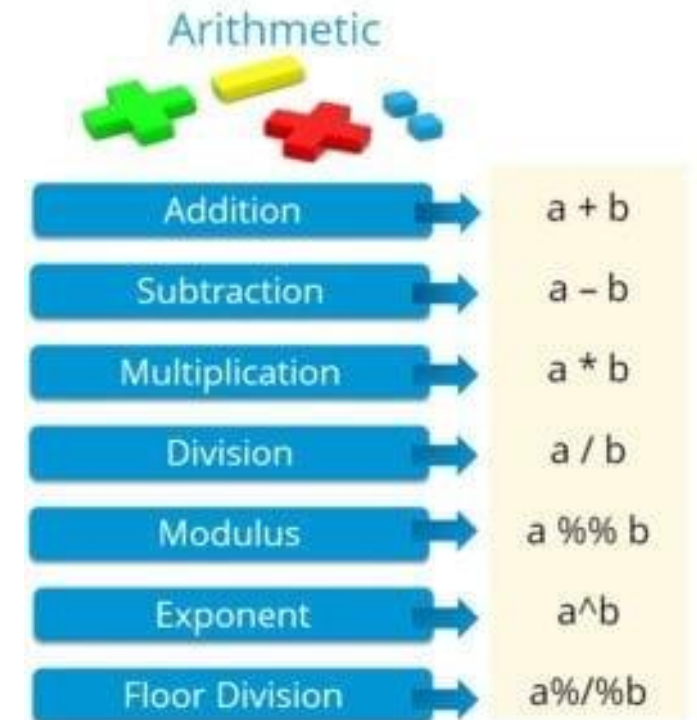
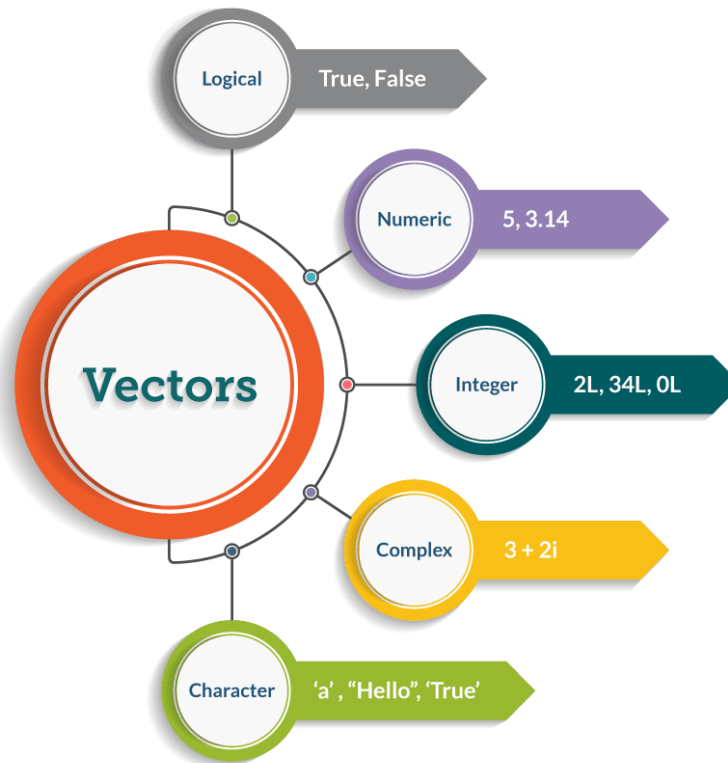


Equal To	→	a == b
Not Equal To	→	a != b
Greater Than	→	a > b
Less Than	→	a < b
Greater Than Equal To	→	a >= b
Less Than Equal To	→	a <= b

Typy objektov (Character, numeric, integer, complex, logical (True/False))

Numeric: dve desatinné miesta, integer sa musí špecifikovať suffixom 1L

Existuje nekonečná hodnota (iba kladná)



# Základná syntax- Vektory

- Základné pravidlo rovnaký typ objektu
- Nie vždy pravda
- Kombinovanie numeric objektu a character objektu- character vector
- Možná reprezentácia čísel ako stringu

```
> x <- c(0.5, 0.6)      ## numeric
> x <- c(TRUE, FALSE)   ## logical
> x <- c(T, F)          ## logical
> x <- c("a", "b", "c") ## character
> x <- 9:29              ## integer
> x <- c(1+0i, 2+4i)     ## complex
```



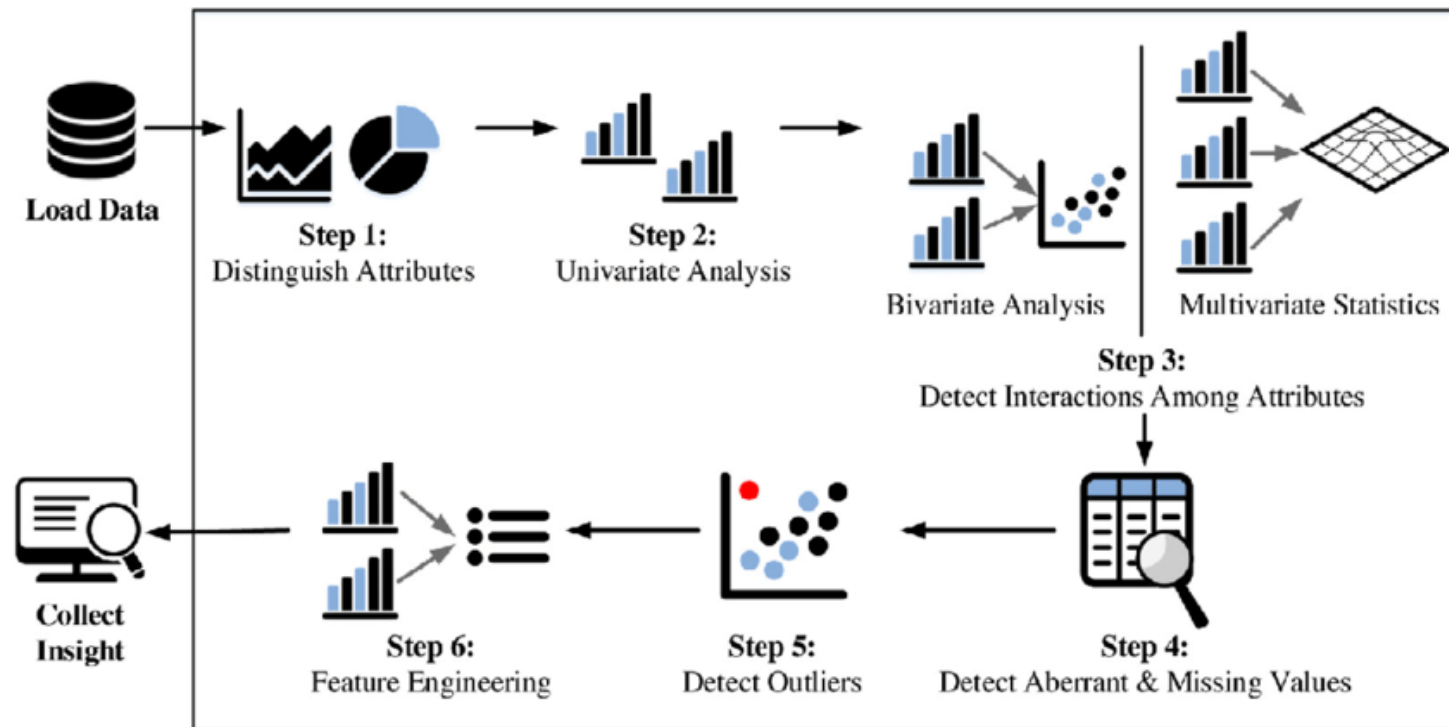
# Exploratory Data Analysis

- Kritický proces prvotného rozboru na dátach

## Ciele:

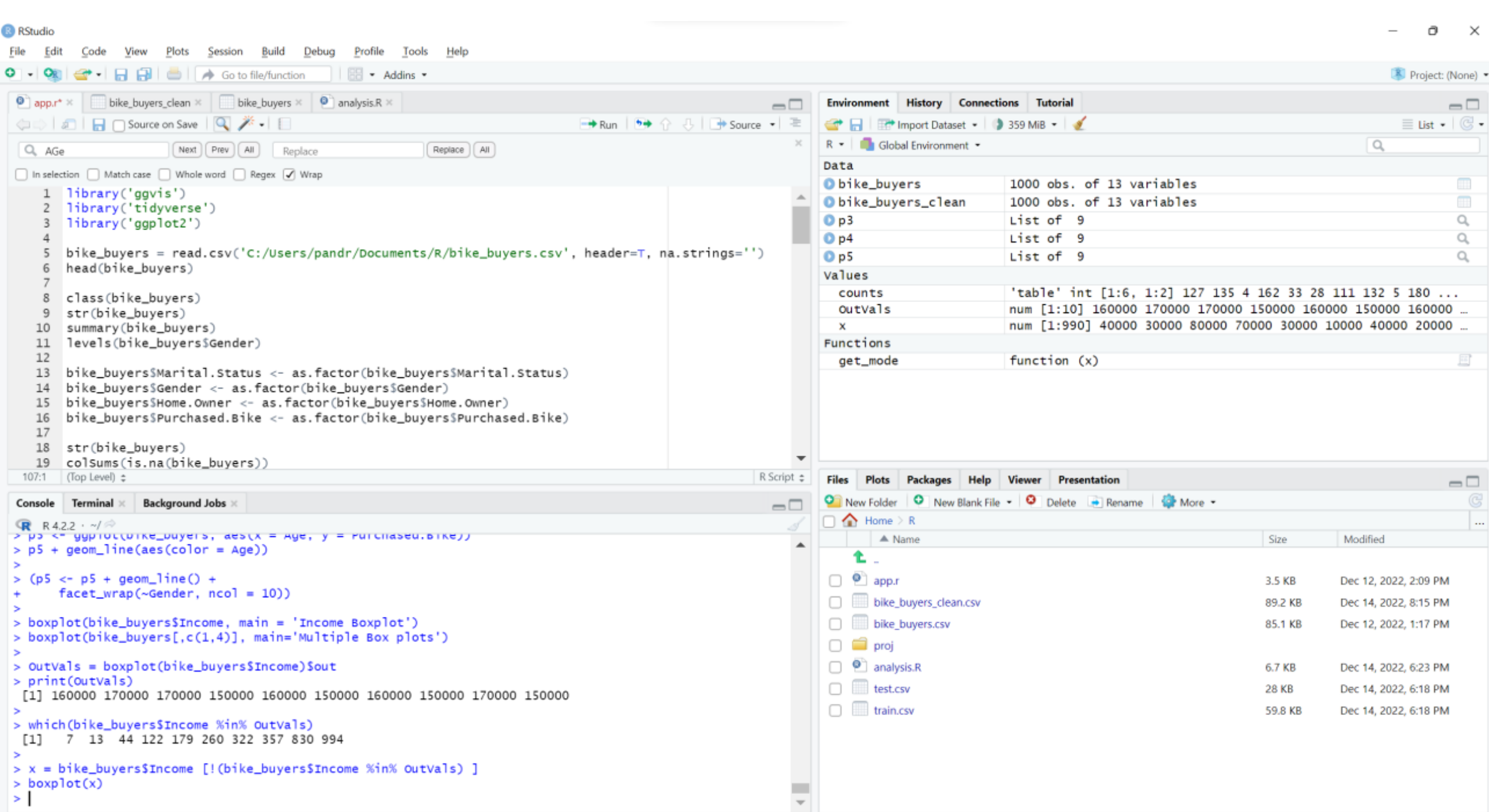
- objaviť vzorce
- objaviť anomálie
- otestovať hypotézy s pomocou súhrnu štatistík a grafických reprezentácií

# EDA



The background of the image is composed of three solid-colored rectangular blocks. On the left is a vertical grey bar. To its right is a large teal block containing the text 'RStudio'. Below the teal block is a dark grey horizontal bar.

RStudio



# Vstupný dataset

- 1000 ľudí rôzneho pozadia
- 13 stĺpcov informácií (vrátane ID)
- Obsahuje informáciu o kúpe bicykla
- Obsahuje NA hodnoty, ktoré je dobré ošetriť

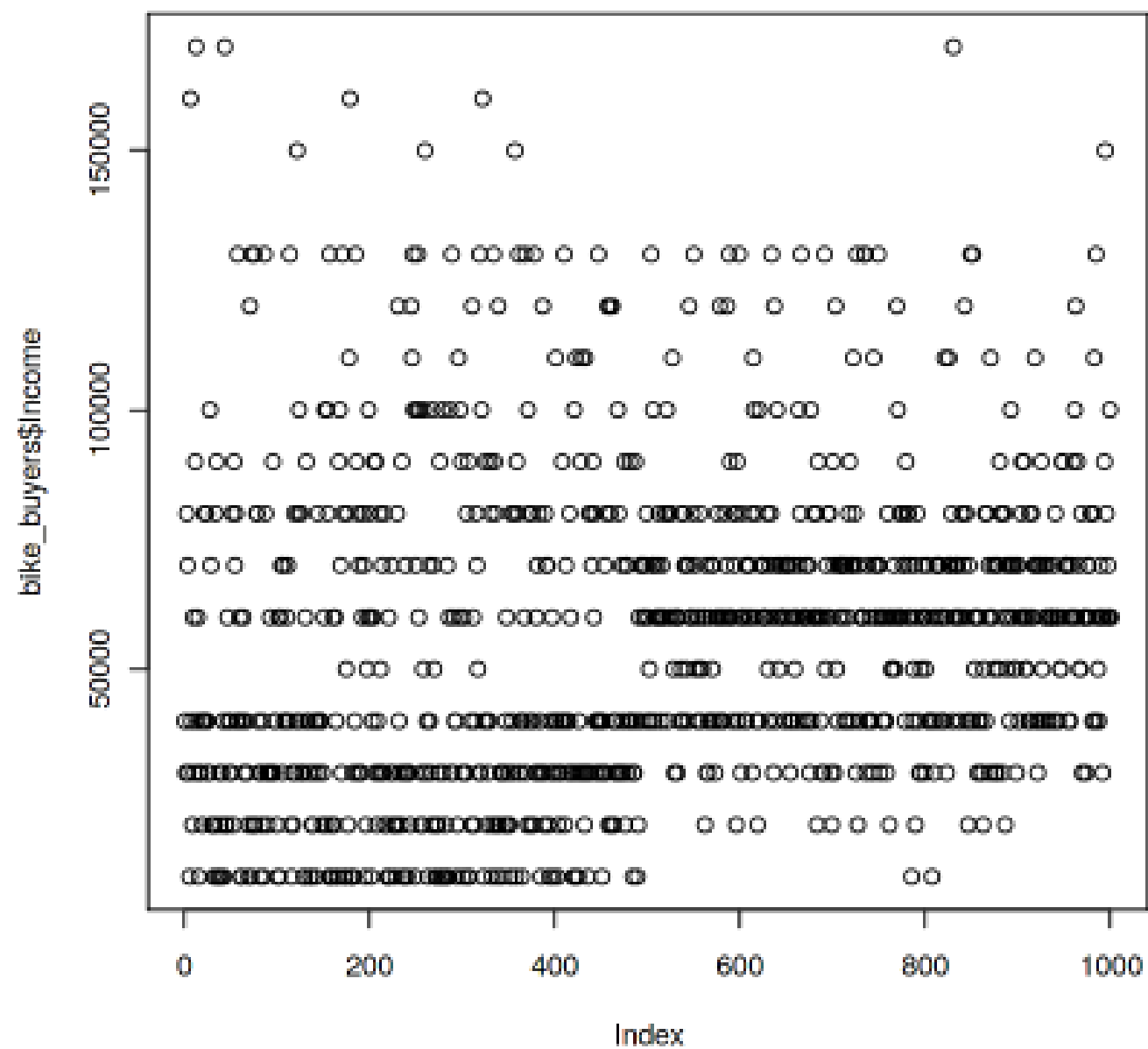
	ID	Marital.Status	Gender	Income	Children	Education	Occupation	Home.Owner	Cars	Commute.Distance	Region	Age	Purchased.Bike
1	12496	Married	Female	40000	1	Bachelors	Skilled Manual	Yes	0.000000	0-1 Miles	Europe	42	No
2	24107	Married	Male	30000	3	Partial College	Clerical	Yes	1.000000	0-1 Miles	Europe	43	No
3	14177	Married	Male	80000	5	Partial College	Professional	No	2.000000	2-5 Miles	Europe	60	No
4	24381	Single	Male	70000	0	Bachelors	Professional	Yes	1.000000	5-10 Miles	Pacific	41	Yes
5	25597	Single	Male	30000	0	Bachelors	Clerical	No	0.000000	0-1 Miles	Europe	36	Yes
6	13507	Married	Female	10000	2	Partial College	Manual	Yes	0.000000	1-2 Miles	Europe	50	No
7	27974	Single	Male	160000	2	High School	Management	Yes	4.000000	0-1 Miles	Pacific	33	Yes
8	19364	Married	Male	40000	1	Bachelors	Skilled Manual	Yes	0.000000	0-1 Miles	Europe	43	Yes
9	22155	Married	Male	20000	2	Partial High School	Clerical	Yes	2.000000	5-10 Miles	Pacific	58	No
10	19280	Married	Male	60000	2	Partial College	Manual	Yes	1.000000	0-1 Miles	Europe	43	Yes

<https://www.kaggle.com/datasets/heeraldedhia/bike-buyers>

# Výstupy

- Čisté dáta – Nahradené za MODE
  - Marital Status - Nahradené za MODE
  - Gender - Nahradené za MODE
  - Children - Nahradené za MODE
  - Home Owner - Nahradené za MODE
  - Cars - Nahradené za PRIEMER
- 
- MODE – funkcia, ktorá má najväčšiu pravdepodobnosť aby sa tam vyskytla

Príjem

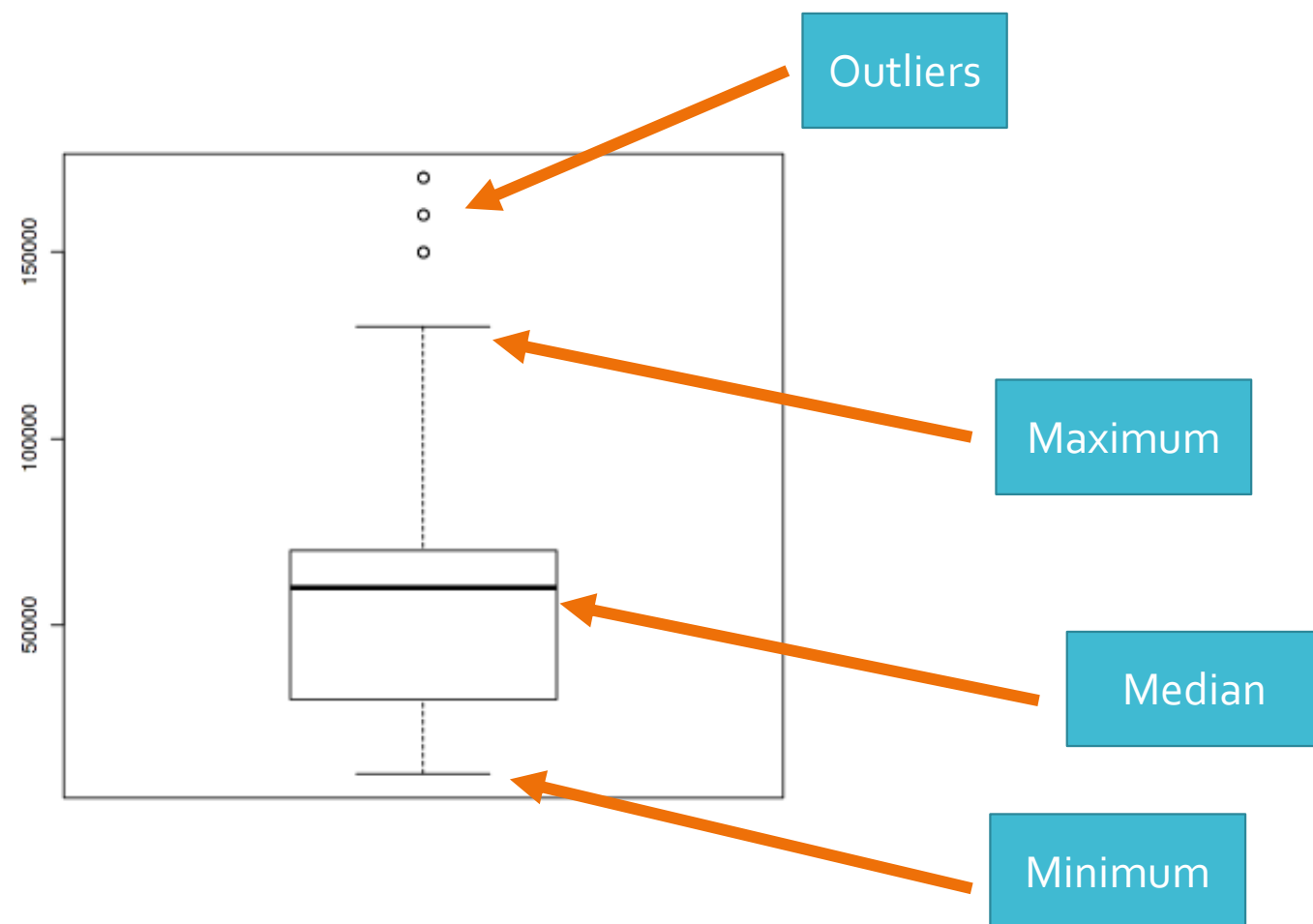


# Outliers pre príjem

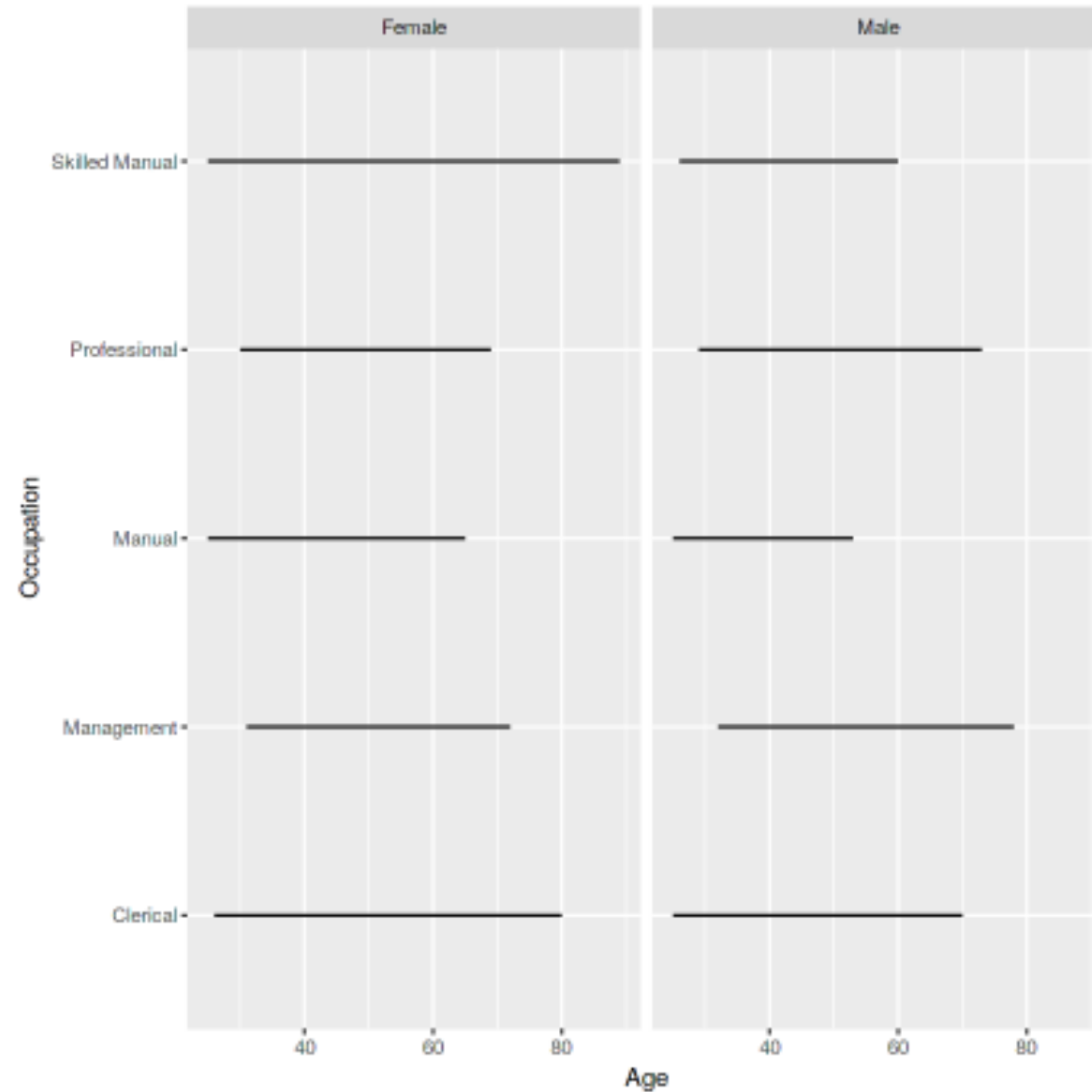
1.5x väčšie

```
[1] 160000 170000 170000 150000 160000 150000 160000 150000 170000 150000
```

```
7 · 13 · 44 · 122 · 179 · 260 · 322 · 357 · 830 · 994
```

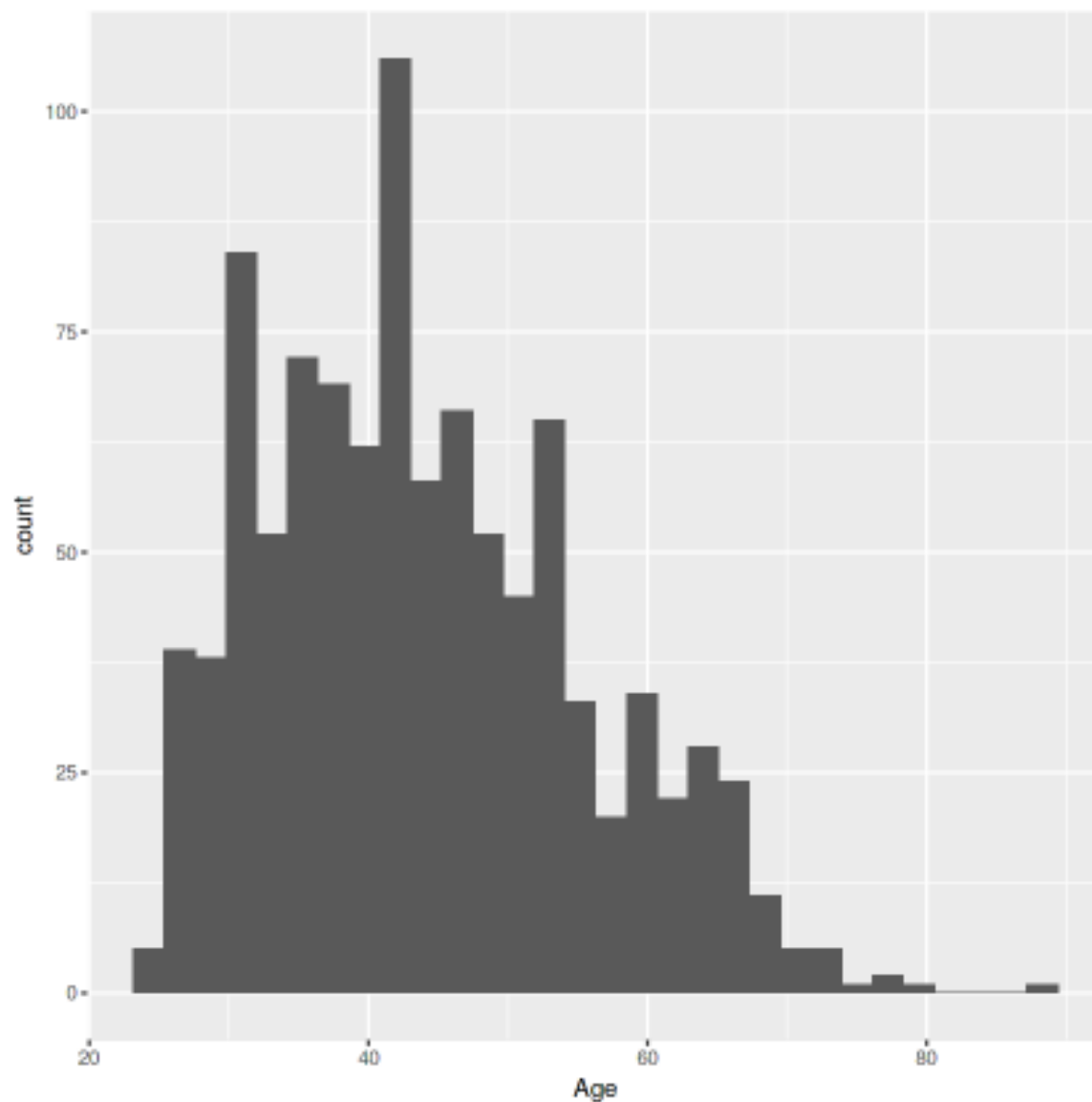


# Povolanie podľa veku a pohlavia

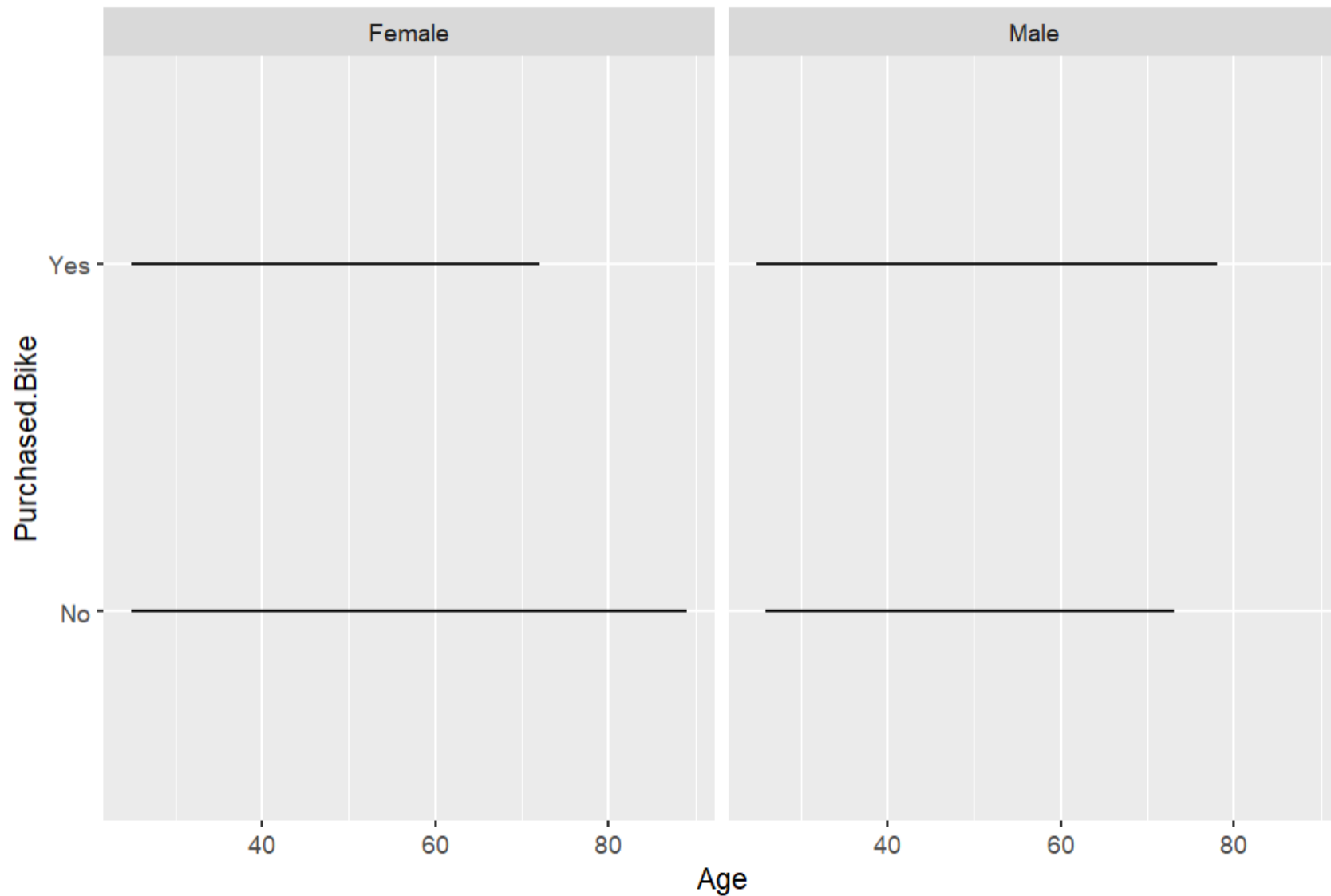




Počet ľudí  
podľa vekovej  
kategórie



Vekový rozsah  
ľudí, ktorí si  
kúpili bicykel  
na základe  
pohlavia



# Zhrnutie

- Použite radšej Python
- Inštalácia packageov dokáže byť cancer
- RStudio vyzerá ako NetBeans
- Cesty v R sú kostrbaté
- T F pre lenivých