

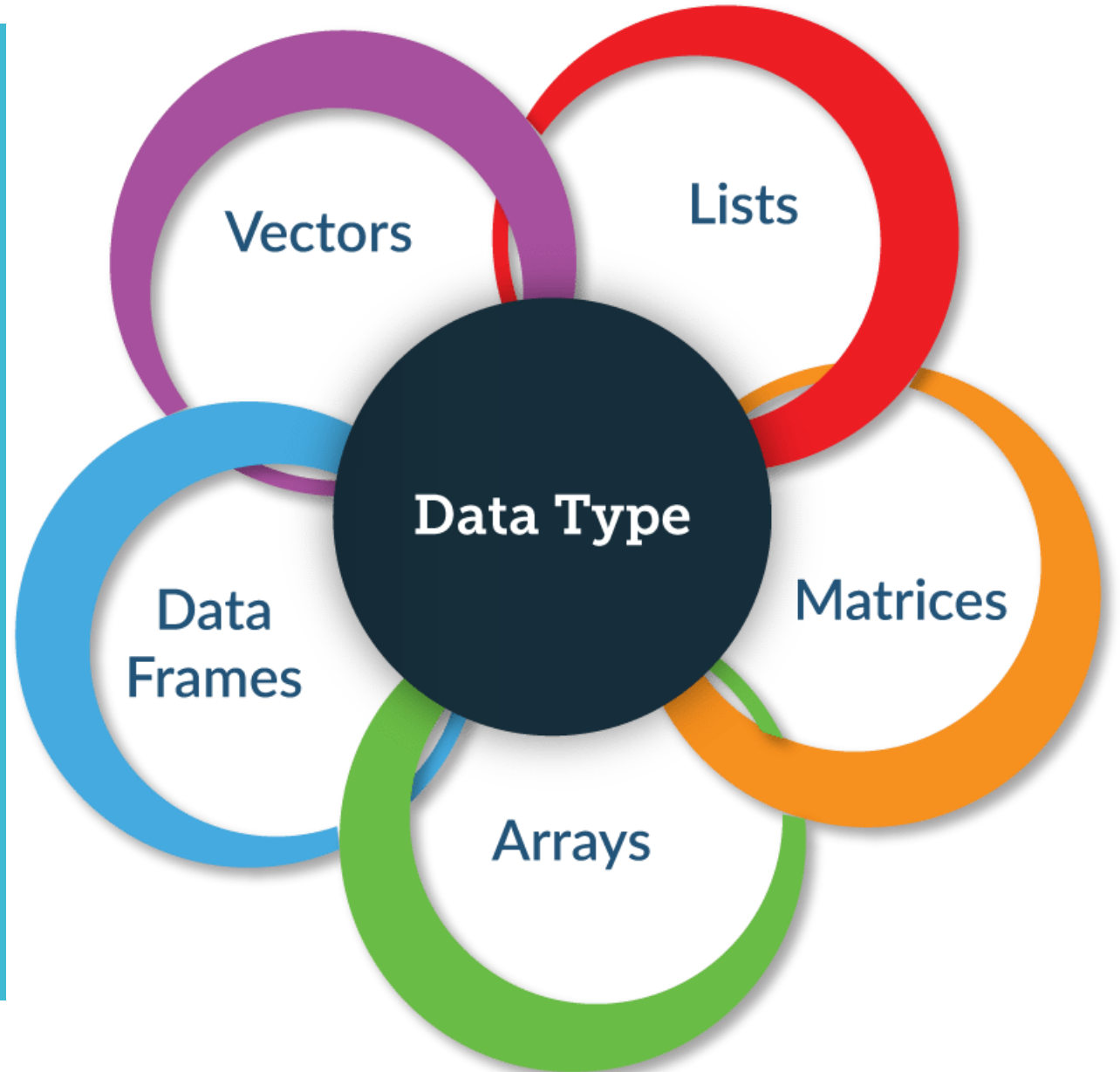
# R Programming Language

Peter Andrejko, Marek Dráb, Dávid Gavenda, Marek Klimo

“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.”



**Roger D. Peng**



# Čo je R?

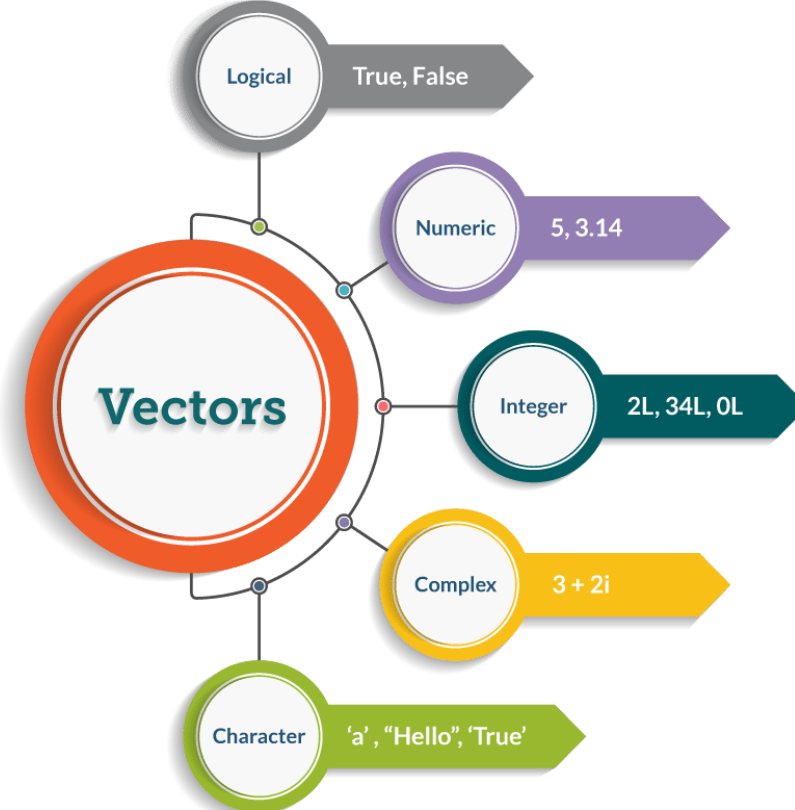
- Je programovacie prostredie určené na štatistickú analýzu dát a ich zobrazenie
- Ide o implementáciu jazyka S pod slobodnou licenciou
- Poskytuje širokú škálu štatistických a grafických techník vrátane lineárneho a nelineárneho modelovanie, klasických štatistických testov, analýzy časových radov, zhlukovania
- 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

# História

- Prvé vydanie – August 1993 - Robert Gentleman a Ross Ihaka
- Verzia 0.16 – posledná alfa verzia 1.4.1997
- Verzia 1.0.0 – prvá stabilná verzia na komerčné využitie 29.2.2000
- Vývojár – R Core Team



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



#### 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	$a <= b$

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



Arithmetic

Addition	$a + b$
Subtraction	$a - b$
Multiplication	$a * b$
Division	$a / b$
Modulus	$a \% b$
Exponent	$a^b$
Floor Division	$a // b$

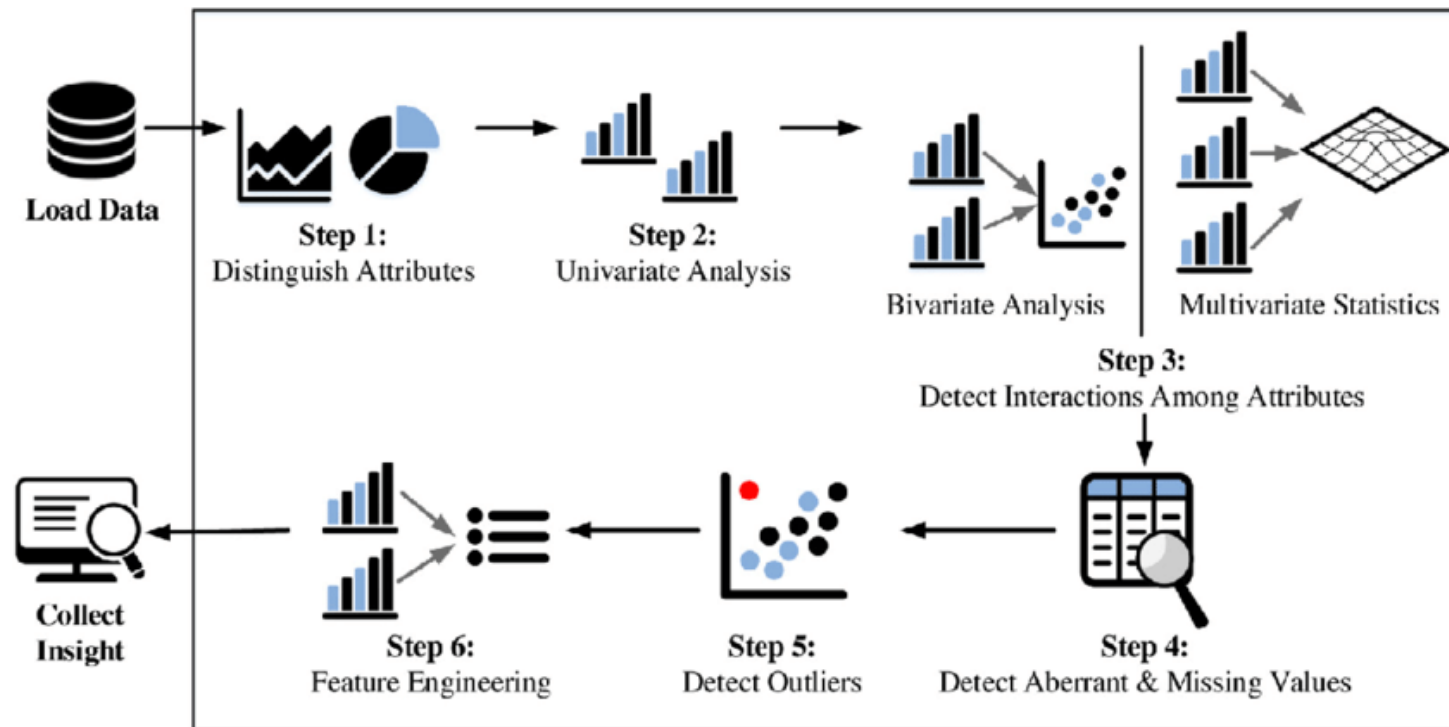
# Exploratory Data Analysis

- 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



RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

app.r\* bike\_buyers\_clean\* bike\_buyers\* analysis.R\*

Source on Save Run

Search: AGe Next Prev All Replace Replace All

☐ In selection ☐ Match case ☐ Whole word ☐ Regex ☒ Wrap

```
1 library('ggvis')
2 library('tidyverse')
3 library('ggplot2')
4
5 bike_buyers = read.csv('C:/Users/pandr/Documents/R/bike_buyers.csv', header=T, na.strings='')
6 head(bike_buyers)
7
8 class(bike_buyers)
9 str(bike_buyers)
10 summary(bike_buyers)
11 levels(bike_buyers$Gender)
12
13 bike_buyers$Marital.Status <- as.factor(bike_buyers$Marital.Status)
14 bike_buyers$Gender <- as.factor(bike_buyers$Gender)
15 bike_buyers$Home.Owner <- as.factor(bike_buyers$Home.Owner)
16 bike_buyers$Purchased.Bike <- as.factor(bike_buyers$Purchased.Bike)
17
18 str(bike_buyers)
19 colSums(is.na(bike_buyers))
```

107:1 (Top Level) R Script

Console Terminal Background Jobs

```
R 4.2.2 ~/\> p5 <- ggplot(bike_buyers, aes(x = Age, y = Purchased.Bike))
> p5 + geom_line(aes(color = Age))
>
> (p5 <- p5 + geom_line() +
+   facet_wrap(~Gender, ncol = 10))
>
> boxplot(bike_buyers$Income, main = 'Income Boxplot')
> boxplot(bike_buyers[,c(1,4)], main='Multiple Box plots')
>
> OutVals = boxplot(bike_buyers$Income)$out
> print(OutVals)
[1] 160000 170000 170000 150000 160000 150000 160000 150000 170000 150000
>
> which(bike_buyers$Income %in% OutVals)
[1] 7 13 44 122 179 260 322 357 830 994
>
> x = bike_buyers$Income [!(bike_buyers$Income %in% OutVals) ]
> boxplot(x)
> |
```

Environment History Connections Tutorial

Import Dataset 359 MiB

R Global Environment

Data

bike_buyers	1000 obs. of 13 variables
bike_buyers_clean	1000 obs. of 13 variables
p3	List of 9
p4	List of 9
p5	List of 9

Values

counts	'table' int [1:6, 1:2] 127 135 4 162 33 28 111 132 5 180 ...
OutVals	num [1:10] 160000 170000 170000 150000 160000 150000 160000 ...
x	num [1:990] 40000 30000 80000 70000 30000 10000 40000 20000 ...

Functions

get_mode	function (x)
----------	--------------

Files Plots Packages Help Viewer Presentation

New Folder New Blank File Delete Rename More

Home > R

	Name	Size	Modified
	..		
	app.r	3.5 KB	Dec 12, 2022, 2:09 PM
	bike_buyers_clean.csv	89.2 KB	Dec 14, 2022, 8:15 PM
	bike_buyers.csv	85.1 KB	Dec 12, 2022, 1:17 PM
	proj		
	analysis.R	6.7 KB	Dec 14, 2022, 6:23 PM
	test.csv	28 KB	Dec 14, 2022, 6:18 PM
	train.csv	59.8 KB	Dec 14, 2022, 6:18 PM

# Vstupný dataset

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

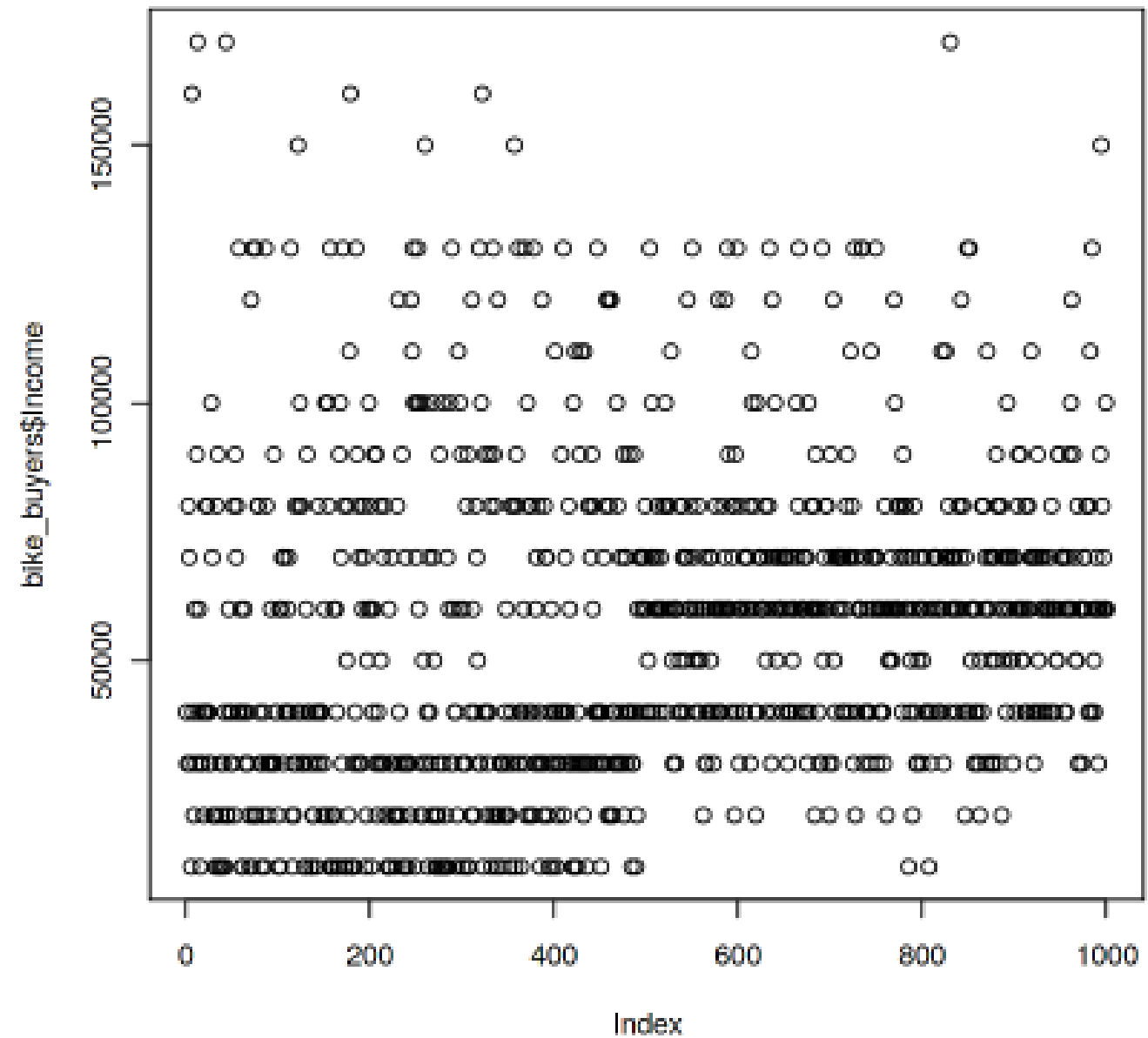
	ID	Marital.Status	Gender	Income	Children	Education	Occupation	Home.Owner	Cars	Commute.Distance	Region	Age	Purchased.Bike
1	12496	Married	Female	40000	1	Bachelors	Skilled Manual	Yes	0.000000	0-1 Miles	Europe	42	No
2	24107	Married	Male	30000	3	Partial College	Clerical	Yes	1.000000	0-1 Miles	Europe	43	No
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

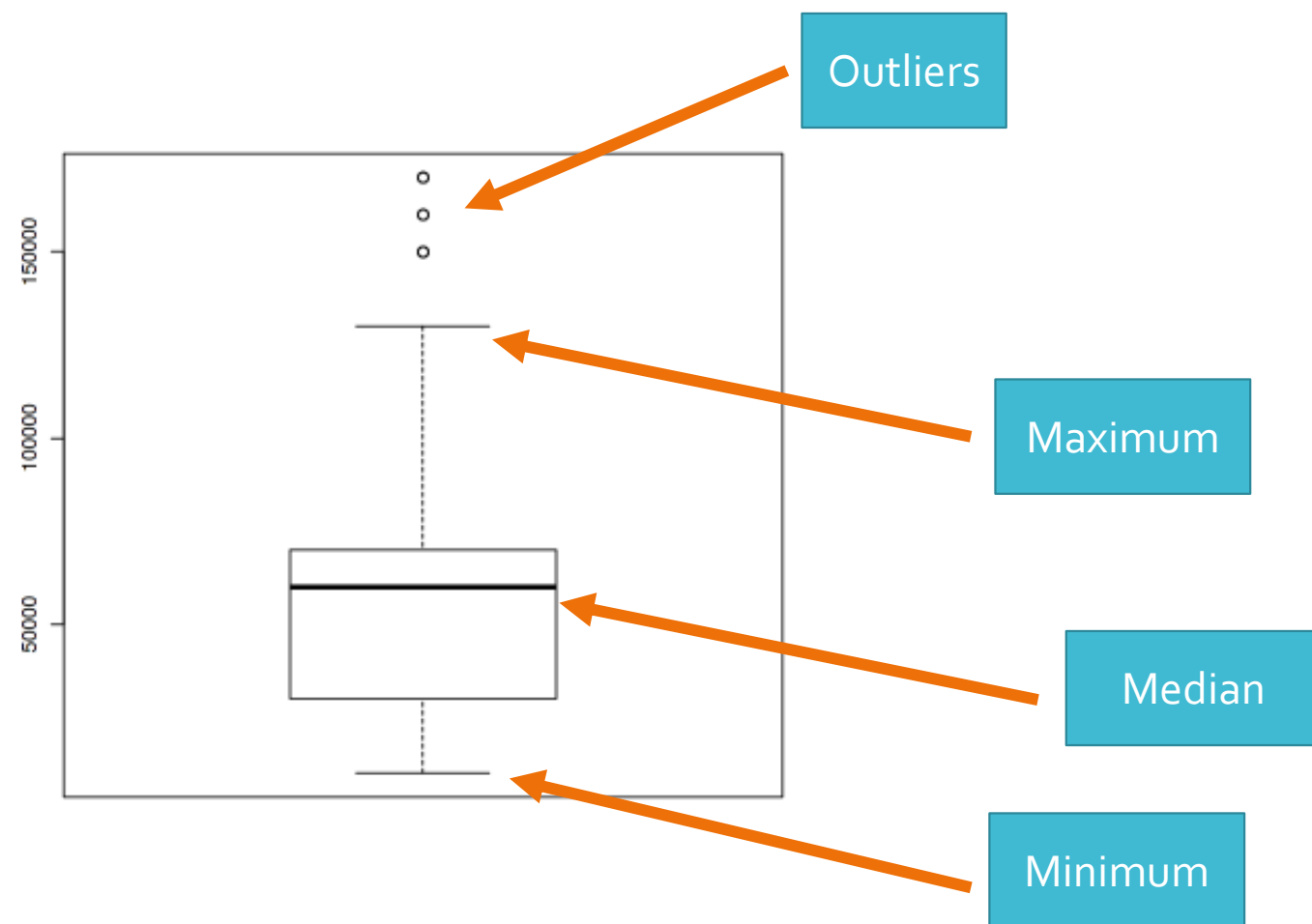


# Outliers pre príjem

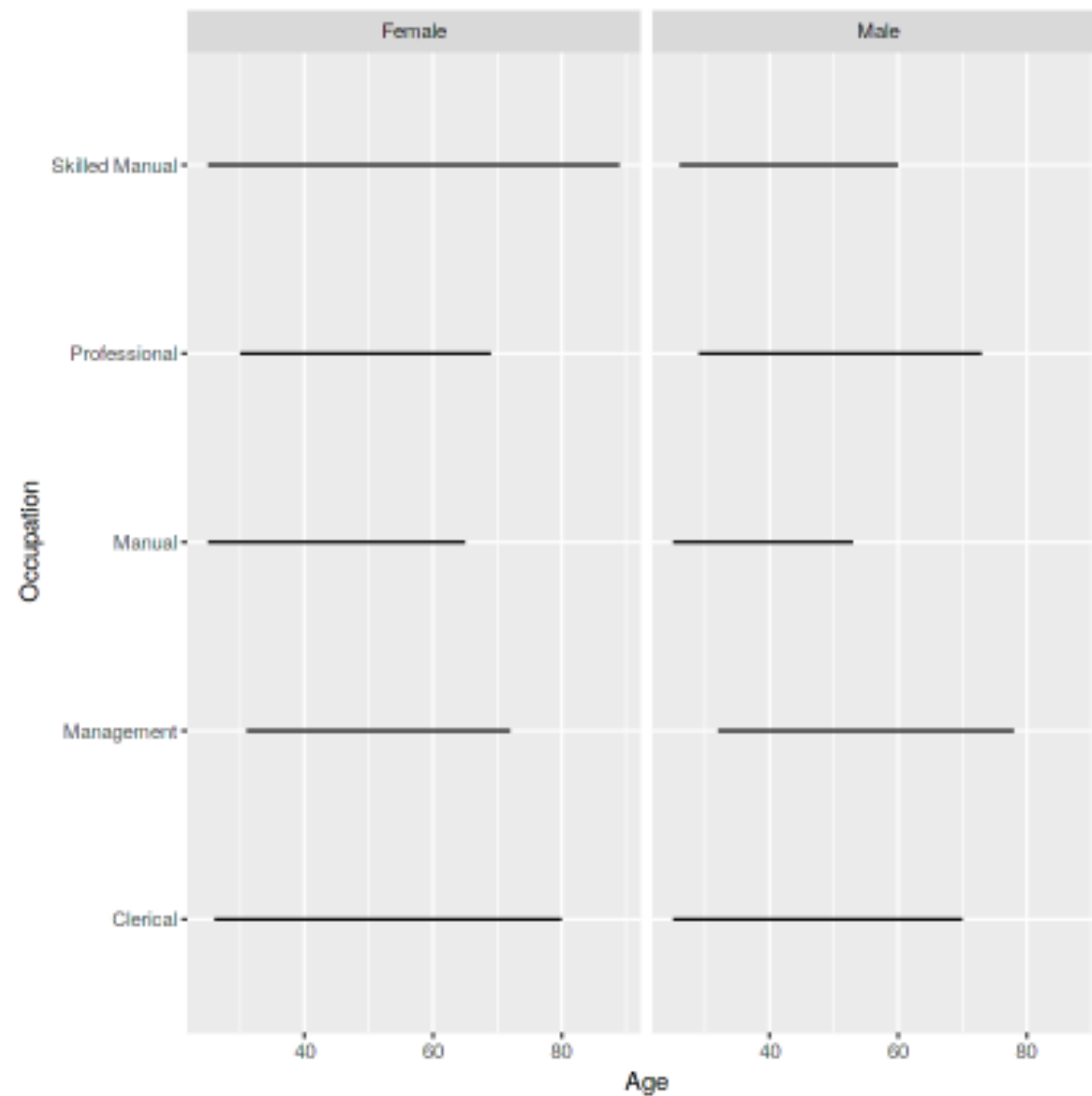
1.5x väčšie

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

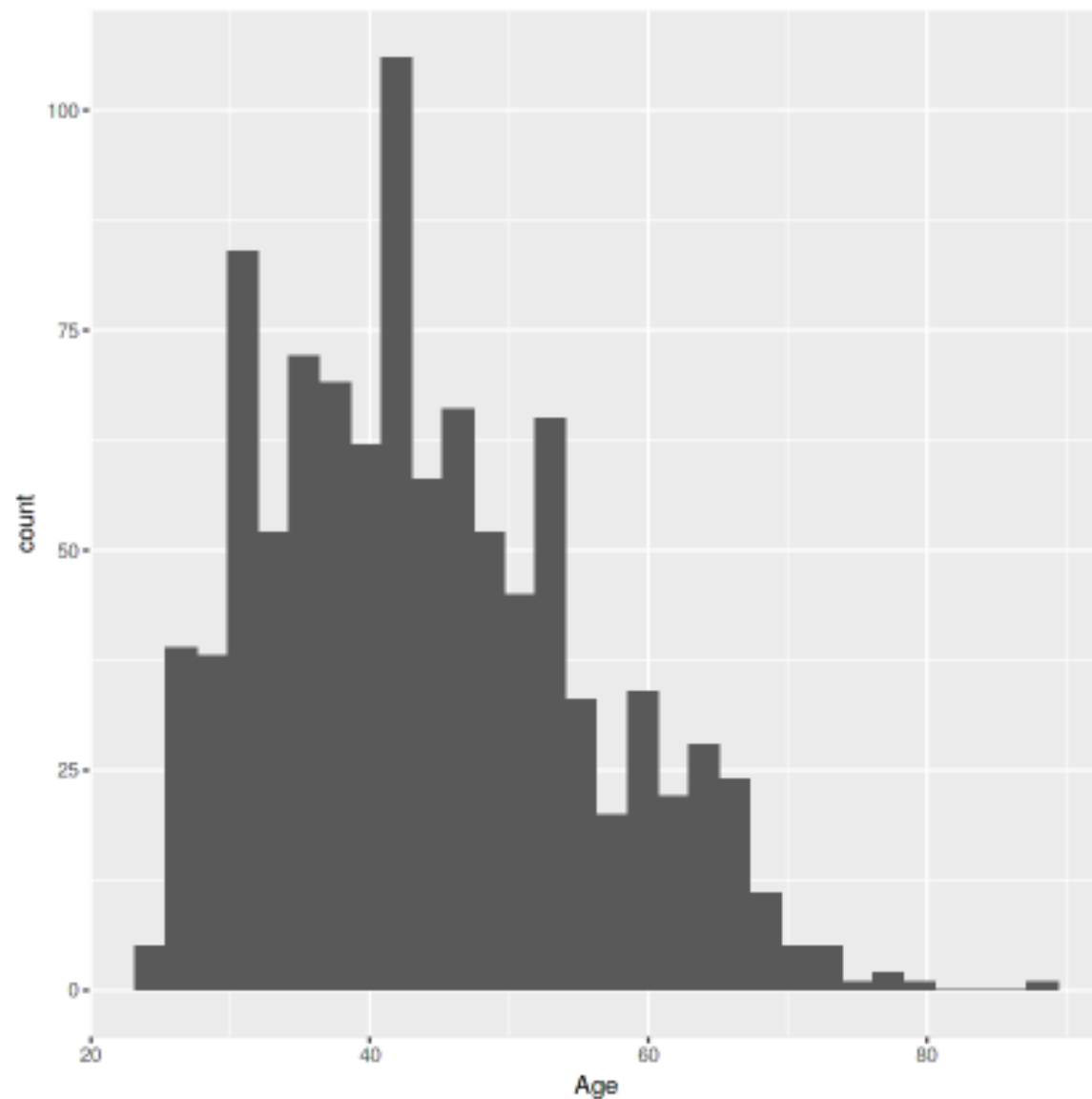
```
7 · 13 · 44 · 122 · 179 · 260 · 322 · 357 · 830 · 994
```



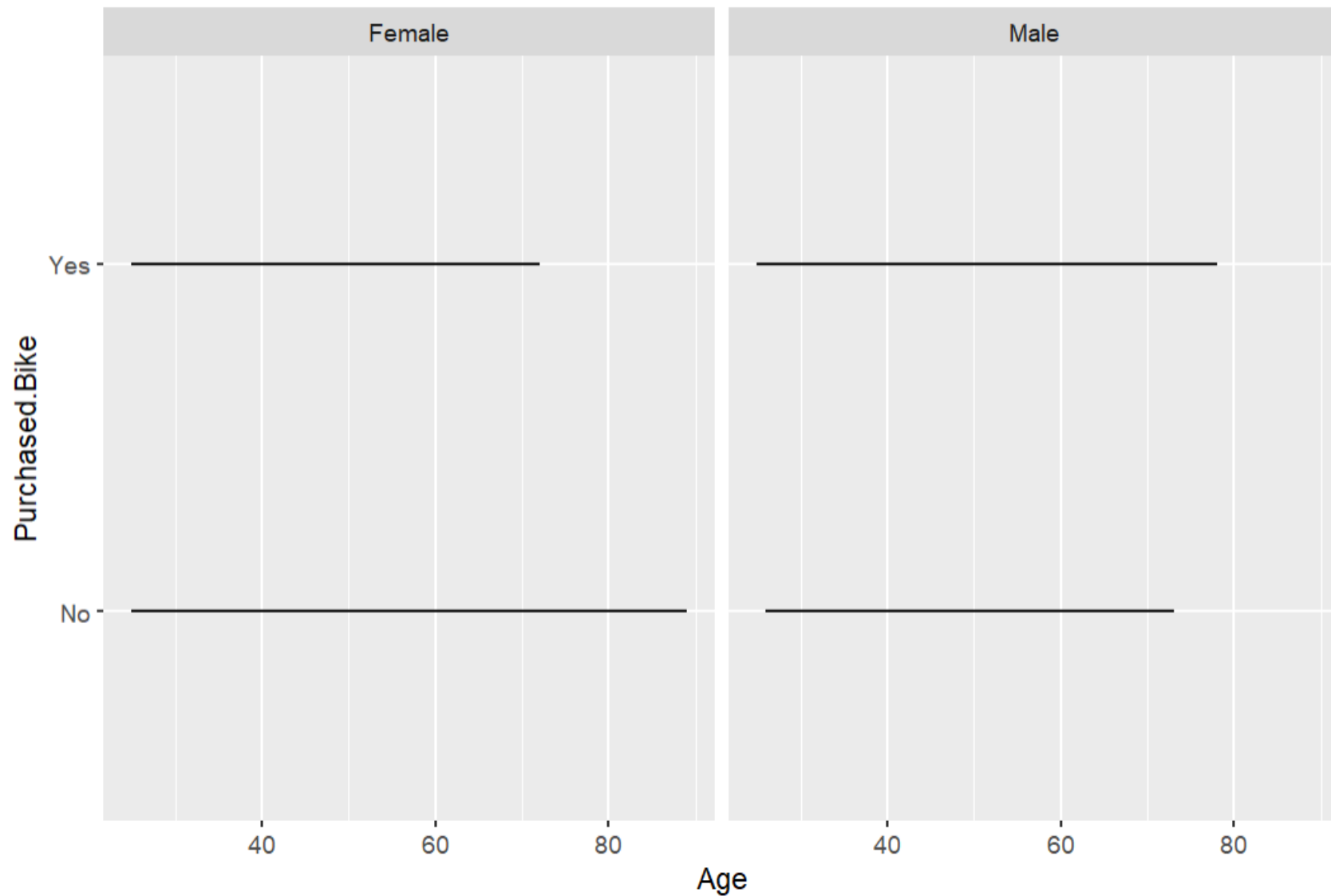
# Povolanie podľa veku a pohlavia



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

# Otázka ku skúške

Na čo sa používa programovací jazyk R

- A) analýza dát, štatistika
- B) tvorba webstránok a webových aplikácií
- C) mobilné aplikácie
- D) čistenie dát a zobrazenie grafov

# Otázka ku skúške

Na čo sa používa programovací jazyk R

- A) analýza dát, štatistika
- B) tvorba webstránok a webových aplikácií
- C) mobilné aplikácie
- D) čistenie dát a zobrazenie grafov