Język R dla początkujących

Prowadzący: lek. Katarzyna Kubiak
Koordynator i zaliczający: dr hab. n. med Barbara Więckowska

Katedra i Zakład Informatyki i Statystyki
ul. Rokietnicka 7, 60-806 Poznań
e-mail: katarzyna.kubiak@student.ump.edu.pl, barbara.wieckowska@ump.edu.pl

Literatura

- Biecek Przemysław, *Przewodnik po pakiecie R*, GiS, 2017
- Wickham Hadley & Grolemund Garrett, *R for data science*, O'Reilly, 2017 (dostępna online pod adresem: https://r4ds.had.co.nz/)
- Wickham Hadley, *ggplot2: elegant graphics for data analysis*, Springer, 2010 (dostępna online pod adresem: https://ggplot2-book.org/index.html)

Wstęp

- R to język programowania dostępny dla użytkowników MS Windows, macOS i Linuxa
- R jest wolnym (darmowym i otwartym) środowiskiem
- R to język interpretowany, a nie kompilowany
- RStudio to zintegrowane środowisko programowania (ang. *integrated development environment*, IDE)
- Cheatsheets: https://www.rstudio.com/resources/cheatsheets/
- Biblioteki (pakiety):

```
install.packages("ggplot2")
library(ggplot2)
detach(package::ggplot2)
```

- Pomoc:

```
?mean
help(mean)
```

- Komentarz w kodzie:

```
# komentarz do końca linii
```

- Operator przypisania:

```
<- # w RStudio skrót klawiszowy ALT+-
->
= # niezalecany
```

Struktury danych

Wektor

- jednego typu: numeryczny, znakowy, zespolony, logiczny

```
(a <- 1:5)

## [1] 1 2 3 4 5

(b <- c(TRUE, FALSE, FALSE, TRUE, TRUE))

## [1]  TRUE FALSE FALSE  TRUE  TRUE

(c <- c(12, 6, NA, 17, 2))

## [1] 12  6 NA 17  2

(d <- seq(from = 1, to = 8, length.out = 5))

## [1] 1.00 2.75 4.50 6.25 8.00

e <- vector(mode = "character")
e[1] <- "R"
e[2:5] <- rep("fakultet", times = 4)
e

## [1] "R"        "fakultet" "fakultet" "fakultet" "fakultet"
```

Ramka danych

```
ramka <- data.frame(a, b, c)
View(ramka)
(ramka <- cbind(ramka, d, e))

##   a     b  c    d        e
## 1 1  TRUE 12 1.00        R
## 2 2 FALSE  6 2.75 fakultet
## 3 3 FALSE NA 4.50 fakultet
## 4 4  TRUE 17 6.25 fakultet
## 5 5  TRUE  2 8.00 fakultet

colnames(ramka)

## [1] "a" "b" "c" "d" "e"

ramka$a

## [1] 1 2 3 4 5

ramka[2,4]

## [1] 2.75
```

Tibble - nowoczesna ramka danych

```
library(tibble)
(ramka.tib <- as_tibble(ramka))
```

```
## # A tibble: 5 × 5
##       a b         c     d e
##   <int> <lgl> <dbl> <dbl> <chr>
## 1     1 TRUE     12  1    R
## 2     2 FALSE     6  2.75 fakultet
## 3     3 FALSE    NA  4.5  fakultet
## 4     4 TRUE     17  6.25 fakultet
## 5     5 TRUE      2  8    fakultet
```

```
ramka.tib$a
```

```
## [1] 1 2 3 4 5
```

```
ramka.tib[['a']]
```

```
## [1] 1 2 3 4 5
```

```
ramka.tib[[1]]
```

```
## [1] 1 2 3 4 5
```

```
ramka.tib[['e']][1]
```

```
## [1] "R"
```

Lista

```
lista <- list(a, b, c)
lista
```

```
## [[1]]
## [1] 1 2 3 4 5
##
## [[2]]
## [1]  TRUE FALSE FALSE  TRUE  TRUE
##
## [[3]]
## [1] 12  6 NA 17  2
```

```
(lista2 <- list(numer = a, logiczna = b, wartosc = c))
```

```
## $numer
## [1] 1 2 3 4 5
##
## $logiczna
## [1]  TRUE FALSE FALSE  TRUE  TRUE
##
## $wartosc
## [1] 12  6 NA 17  2
```

Wczytywanie i zapisywanie danych

```
read_table()
load()
read.csv()
read_excel()
write.table()
save()
write.csv()
write.xlsx()
```

Katalog bieżący: `getwd()`, ustalanie katalogu bieżącego: `setwd()`.

Kilka przydatnych funkcji

```
View()
typeof()
str()
summary()
head()
tail()
is.na()
seq()
rep()
which()
round()
cbind()
rbind()
sample()
```

Zadania - wstęp

Zadanie 0 Otwórz RStudio. Utwórz nowy skrypt i zapisz go jako `wstep_do_R.R`.
Zadanie 1 Utwórz wektor złożony z liczb: 1, 15, 8, 13, 6, 4, 11. Uporządkuj elementy wektora od największego do najmniejszego.
Zadanie 2 Utwórz wektor złożony z kolejnych liczb naturalnych od 10 do 50. Odwróć kolejność elementów wektora.
Zadanie 3 Utwórz wektor złożony z kolejnych parzystych liczb naturalnych od 2 do 50. Usuń trzeci, piąty i szósty element wektora. Jaka jest długość tych wektorów?
Zadanie 4 Utwórz wektor złożony z kolejnych nieparzystych liczb naturalnych postaci: 99, 97, …, 3, 1. Następie usuń z tego wektora liczy 11 i 13.
Zadanie 5 Utwórz wektor złożony z 4 powtórzeń sekwencji liczb (4, 6, 8). Zastąp pierwsze powtórzenie brakami w danych (`NA`).
Zadanie 6 Wygeneruj wektor `wiek` z wartościami z przedziału od 20 do 60 lat dla 30 osób. Następnie wyznacz następujące statystyki opisowe: średnia, wariancja, odchylenie standardowe, mediana, rozstęp międzykwartylowy, minimum, maksimum.
Zadanie 8 Utwórz wektor `imie` zawierający imiona piątki twoich znajomych oraz wektor `wzrost` zawierający ich wzrost w `cm`. Utwórz ramkę danych `osoby` złożoną z wektorów `imie` i `wzrost`.
Zadanie 9 Wczytaj dane z pliku `Table1.txt`. Zmień nazwy kolumn na `Imie, Wiek,`

`Wzrost, Masa, Plec`.
Zadanie 10 Zainstaluj i załaduj bibliotekę `Przewodnik`. Otwórz zbiór danych `koty_ptaki`. Zbadaj jego strukturę. Wyświetl w konsoli jego pierwsze i ostatnie wiersze. Wybierz pierwsze 3 kolumny i wiersze pierwszy, piąty i ósmy.
Zadanie 11 Zbiór danych `daneSoc` z pakietu `Przewodnik` zawiera dane socjodemograficzne i wartości ciśnienia tętniczego 204 osób. Wyznacz statystyki opisowe dla wybranej zmiennej ilościowej. Wyznacz tablicę liczebności (kontyngencji) dla wykształcenia. Wyznacz tablicę liczebności dla wykształcenia z podziałem na płeć.
Zadanie 12 Funkcja `barplot()` służy do tworzenia wykresów słupkowych. Narysuj wykres słupkowy dla wykształcenia ze zbioru danych `daneSoc`. Ustaw słupki poziomo. Zmień kolor słupków. Narysuj wykres z podziałem na płeć. Dodaj do wykresu legendę (uzyj funkcji `legend()`).
Zadanie 13 Funkcja `hist()` służy do tworzenia histogramów. Narysuj histogram dla wieku ze zbioru danych `daneSoc`. Dodaj tytuł wykresu i nazwy osi. Zmień kolor słupków i ich obramowania. Zmień oś pionową, aby pokazywała częstości zamiast liczebności.
Zadanie 14 Funkcja `boxplot()` służy do tworzenia wykresów pudełkowych. Narysuj wykres pudełkowy dla wartości ciśnienia tętniczego z pudełkami w poziomie. Narysuj wykres pudełkowy dla wieku w grupach wykształcenia.
Zadanie 15 Funkcja `scatterplot()` z biblioteki `car` służy do tworzenia wykresu kropkowego (punktowego, rozrzutu). Narysuj wykres rozrzutu ciśnienia skurczowego i rozkurczowego ze zbioru danych `daneSoc`.

dplyr

Pakiet `dplyr` służy do wydajnego manipulowania danymi. Jest częścią większej grupy pakietów zwanej `tidyverse`.

Niektóre funkcje pakietu `dplyr`:

```
filter()    # wybiera wskazane wiersze
select()    # wybiera wskazane kolumny
arrange()   # sortuje wiersze według wskazanych kolumn
mutate()    # dodaje nową kolumnę lub zmienia istniejącą
group_by()  # grupuje dane względem wskazanych czynników
summarise() # wyznacza wskazane podsumowania w każdej grupie
count()     # liczy w grupach
```

Potoki, ang. *streams*, skrót: `SHIFT+CTRL+M`:

```
%>%
```

Introduction to `dplyr` dostępne pod https://dplyr.tidyverse.org/articles/dplyr.html.

Zadania `dplyr`

Zadanie 0 Załaduj biblioteki `PogromcyDanych`, `dplyr` i `tidyr`. Wczytaj zbiór danych `auta` z pakietu `Przewodnik` za pomocą wywołania `data(auta)`. Wyświetl zbiór danych `auta`.
Zadanie 1 Zadanie dotyczy zbioru danych `auta`.
a) Wybierz wiersze dotyczące Opla Astra. Spośród nich wybierz auta z silnikiem Diesla i

przebiegiem poniżej 100 tys. km. Zawęź poszukiwania do aut z roku 2011, 2010 lub 2008.
b) Wybierz kolumny z marką, modelem i ceną.
c) Utwórz nową kolumnę z wiekiem auta.
d) Posortuj każdy model względem ceny malejąco.
e) Wyznacz cenę minimalną, maksymalną i średnią dla całego zbioru danych.
f) Wyznacz cenę minimalną, maksymalną i średnią dla każdej marki osobno.
Zadanie 2 Jaki przebieg, cenę i rok produkcji mają 4 najtańsze auta marki Kia ze zbioru danych `auta`?
Zadanie 3 Zmień język na angielski za pomocą wywołania `setLang(lang = 'eng')`.
Zadanie dotyczy zbioru danych `auta2012` z pakietu `PogromcyDanych`.
a) Która marka samochodów jest najpopularniejsza?
b) Wybierz auta marki `Toyota`. Utwórz tabelę liczności dla zmiennej `Type.of.fuel`. Wyniki posortuj malejąco.
c) Ile aut marki `Volkswagen` jest napędzane benzyną? Jaki to procent wszystkich aut (zaokraglij do 2 miejsc po przecinku)?
d) Wyznacz średnią cenę dla każdej marki. Wybierz 10 najtańszych i 10 najdroższych marek i zapisz je w osobnych ramkach danych. e) Wybierz auta marki `Ford`. Wyznacz pierwszy i trzeci kwartyl ceny.
f) Wybierz modele `Passat`, `Golf` i `Amarok` marki `Volkswagen`. Dla każdego modelu wyznacz średnią i medianę ceny oraz średni przebieg. Wyniki posortuj względem średniego przebiegu rosnąco.

ggplot2

Pakiet ggplot2 służy do wizualizacji danych. Opiera się na gramatyce grafiki (ang. *the grammar of graphics*), co oznacza, że każdy wykres składa się z tych samych elementów: zbioru danych, układu współrzędnych i geometrii, która przedstawia dane. Jest częścią tidyverse.

```
ggplot(dane, aes(x, y, color, shape, size, label))
geom_point()
geom_text()
geom_line()
geom_ribbon()
geom_smooth()
geom_boxplot()
geom_histogram()
geom_bar()
xlim()
ylim()
ggtitle()
xlab()
ylab()
theme()
```

Zadania ggplot2

Zadanie 1 (Zadanie w oparciu o książkę *Przewodnik po pakiecie R* Przemysława Biecka.) Wyświetl zbiór `countries` z pakietu `Przewodnik` (dane o współczynnikach narodzin i zgonów na 1000 mieszkańców dla różnych krajów z różnych kontynentów). Zbadaj jego strukturę.
a) Utwórz szkielet wykresu współczynnika zgonu od współczynnika narodzin. Dodaj wartwę z punktami. Dodaj warstwę z linią trendu.
b) Zmodyfikuj wykres z punktu a): zmapuj zmienne population i continent na atrybuty graficzne warstwy `geom_point()`: zmienną `population` na wielkość punktów, a zmienną `continent` na kolor punktów.
c) Utwórz szkielet wykresu współczynnika narodzin dla kontynentów. Dodaj warstwę z wykresem pudełkowym. Powtórz dla współczynnika zgonu.
d) Przedstaw na wykresie słupkowym liczbę wystąpień każdego kontynentu.
e) Narysuj wykresy kropkowe współczynnika zgonu od współczynnika narodzin dla każdego kontynentu osobno. Użyj funkcji `facet_grid()` i `facet_wrap()`.
f) Narysuj wybrany wykres z poprzednich podpunktów w stylu `theme_minimal()`.
Zadanie 2 Załaduj pakiet `ggthemes`. Narysuj wybrane wykresy z poprzednich zadań w stylach: `theme_excel()`, `theme_excel_new()`, `theme_economist()`.
Zadanie 3 Ze zbioru danych `auta2012` z pakietu `PogromcyDanych` wybierz auta marki `Skoda` z 2006 roku. Narysuj histogram ceny z podziałem na modele.
Zadanie 4 Zaproponuj własny wykres do zbioru danych `auta` z pakietu `Przewodnik`.
Zadanie 5 Narysuj wykresy z Zadań 12-15 z części `Zadania - wstęp` za pomocą biblioteki `ggplot2`.