

Rad s objektima i podacima u R-u

Lucija Kanjer, e-mail: lucija.kanjer@biol.pmf.hr

2024-10-07

Sadržaj praktikuma

- Uvod u rad u programskom okruženju R i osnovne funkcije, instaliranje programskih paketa
- Unos podataka u programsko okruženje R, struktura objekata
- *Rad s objektima i podacima te definiranje bioloških varijabli u R-u*
- Grafički prikaz bioloških podataka i testiranje razdiobe podataka u R-u
- Primjeri osnovnih statističkih analiza kategoričkih i numeričkih varijabli u biološkim istraživanjima u R-u
- Regresije i korelacije, linearni modeli bioloških podataka – primjeri u R-u
- Primjena parametrijskih statističkih testova bioloških podataka u R-u
- Primjena neparametrijskih statističkih testova bioloških podataka u R-u
- Primjeri multivarijatnih analize bioloških podataka u R-u - linearni modeli, klaster analize i ordinacijske analize

Sadržaj današnje vježbe

- odabir samo određenih varijabli iz seta podataka - naredba `select()`
- filtriranje uzoraka oadranih karakteristika - naredba `filter()`
- kreiranje nove varijable - naredba `mutate()`
- grupiranje rezultata po varijablama - naredba `group_by()`
- prikaz rezultata prosjeka varijabli po grupama - naredba `summarize()`
- uklanjanje nedostajućih vrijednosti - naredba `na.omit()`
- pisanje koda s pipe operatorom (`%>%`)

Uvod u Tidyverse



tidyverse

- Tidyverse je skup međusobno povezanih R paketa osmišljenih za olakšavanje **rada s podacima**.
- Osnovna filozofija Tidyverse-a je **“tidy” (uredan) oblik podataka**, gdje su podaci organizirani u tabličnom formatu (**redovi predstavljaju opažanja, a stupci varijable**).
- Omogućava intuitivno i efikasno manipuliranje, analiziranje i vizualiziranje podataka.
- Istovjetne naredbe ponekad su dostupne i u base R-u, ali tidyverse je češće korišten u praksi i pruža puno više mogućnosti za rad s podacima.

Osnovni paketi u Tidyverse-u



- **ggplot2** – Napredna i fleksibilna vizualizacija podataka.
- **dplyr** – Efikasna manipulacija podacima (filtriranje, sortiranje, agregacija).
- **tidyr** – Transformacija podataka u “tidy” format.
- **readr** – Učitavanje podataka iz tekstualnih datoteka (CSV, TSV).
- **tibble** – Poboľšani rad s tablicama, alternativa data.frame-u.

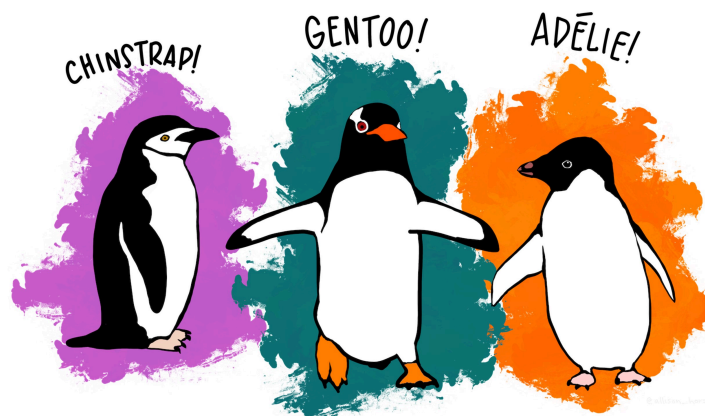
Učitajmo tidyverse u R radno okruženje!

```
# Paketi iz tidyverse-a se mogu učitati svi skupa  
library(tidyverse)
```

```
## — Attaching core tidyverse packages — tidyverse 2.0.0 —  
## ✓ dplyr      1.1.4      ✓ readr      2.1.5  
## ✓ forcats   1.0.0      ✓ stringr   1.5.1  
## ✓ ggplot2   3.5.1      ✓ tibble    3.2.1  
## ✓ lubridate 1.9.3      ✓ tidyr     1.3.1  
## ✓ purrr     1.0.2  
## — Conflicts — tidyverse_conflicts() —  
## ✗ dplyr::filter() masks stats::filter()  
## ✗ dplyr::lag()    masks stats::lag()  
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
# Paketi Tidyverse-a se mogu i zasebno učitavati, npr. ggplot2  
library(ggplot2)
```

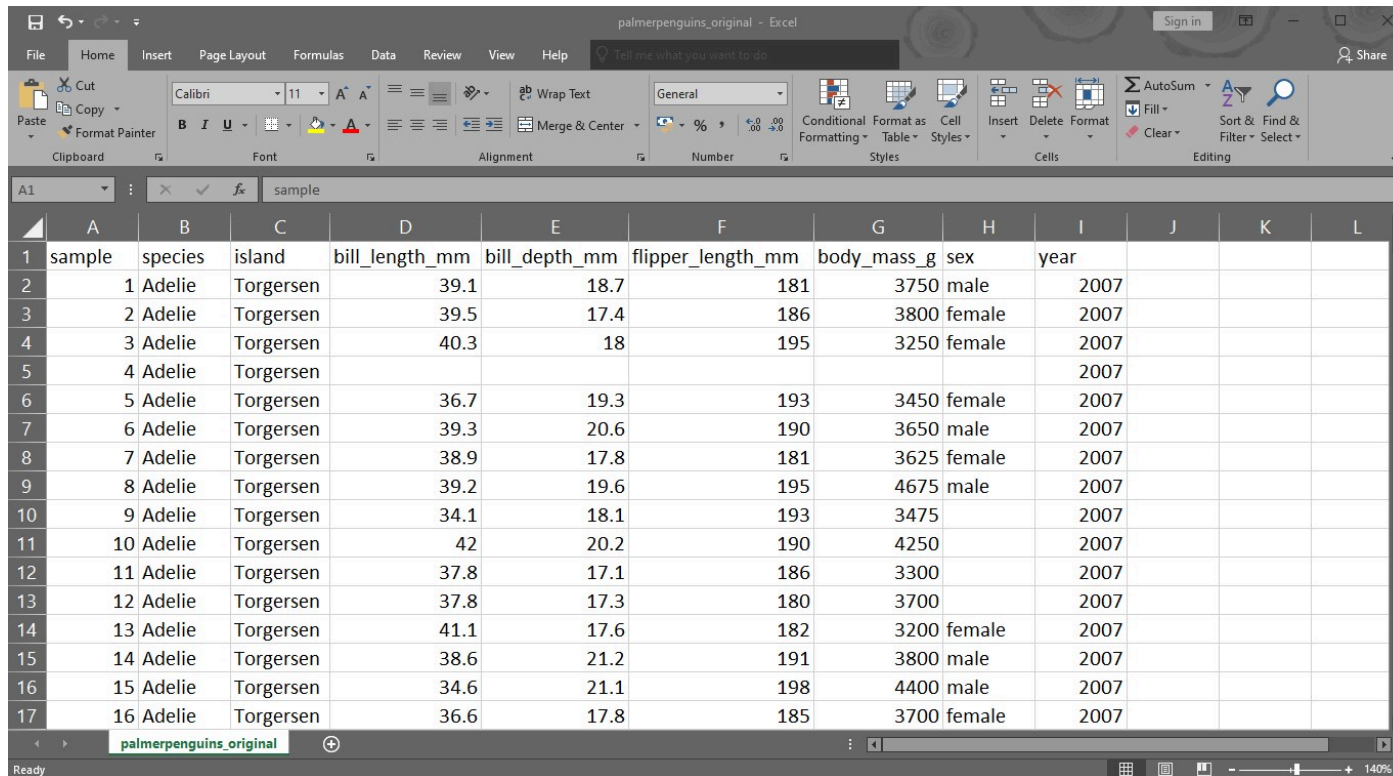
Set podataka o Palmer pingvinima



- Za ovu vježbu koristit ćemo set proširenu verziju podataka **Palmer penguins**.
- Podaci o pingvinima arhipelaga Palmer sadrže mjerenja veličine za **tri vrste pingvina** (Adelie, Chinstrap i Gentoo) promatrane na **tri otoka** (Torgersen, Dream, Biscoe) u arhipelagu Palmer na Antarktici.
- Ove je podatke prikupila dr. Kristen Gorman u sklopu dugoročnih američkih ekoloških istraživanja stanice Palmer. Podaci su uvezeni izravno s podatkovnog portala Inicijative za podatke o okolišu (Environmental Data Initiative - EDI) i dostupni su za korištenje uz CC0 licencu ("Bez pridržanih prava") u skladu s Politikom podataka Palmer Station.
- prošireni set podataka sadrži dodatne varijable i dostupan je na <https://www.kaggle.com/datasets/samybaladram/palmers-penguin-dataset-extended/data>

Tablica s podacima o pingvinima

Otvorite tablicu palmerpenguins_extended.xlsx u Excelu.



	A	B	C	D	E	F	G	H	I	J	K	L
1	sample	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year			
2	1	Adelie	Torgersen	39.1	18.7	181	3750	male	2007			
3	2	Adelie	Torgersen	39.5	17.4	186	3800	female	2007			
4	3	Adelie	Torgersen	40.3	18	195	3250	female	2007			
5	4	Adelie	Torgersen						2007			
6	5	Adelie	Torgersen	36.7	19.3	193	3450	female	2007			
7	6	Adelie	Torgersen	39.3	20.6	190	3650	male	2007			
8	7	Adelie	Torgersen	38.9	17.8	181	3625	female	2007			
9	8	Adelie	Torgersen	39.2	19.6	195	4675	male	2007			
10	9	Adelie	Torgersen	34.1	18.1	193	3475		2007			
11	10	Adelie	Torgersen	42	20.2	190	4250		2007			
12	11	Adelie	Torgersen	37.8	17.1	186	3300		2007			
13	12	Adelie	Torgersen	37.8	17.3	180	3700		2007			
14	13	Adelie	Torgersen	41.1	17.6	182	3200	female	2007			
15	14	Adelie	Torgersen	38.6	21.2	191	3800	male	2007			
16	15	Adelie	Torgersen	34.6	21.1	198	4400	male	2007			
17	16	Adelie	Torgersen	36.6	17.8	185	3700	female	2007			

Rad s podacima u R-u

Podsjetimo se: pregled trenutnog i postavljanje novog radnog direktorija.

```
# pregled trenutnog radnog direktorija  
getwd()
```

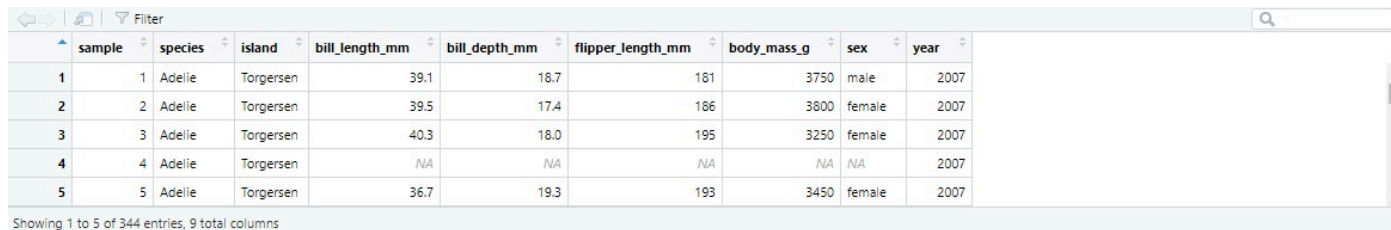
```
## [1] "C:/Users/Hrvoje/Documents/APUBI/03_Rad_s_podacima"
```

```
# postavljanje novog radnog direktorija  
setwd("C:/Users/Hrvoje/Documents/APUBI/03_Rad_s_podacima")
```

Učitavanje podataka iz Excel tablice

```
# Učitavanje potrebnog paketa
library(readxl)
# Učitavanje podataka iz Excel tablice u objekt
penguins <- read_excel("palmerpenguins_original.xlsx")
```

View(penguins) # ili klik na objekt u environmentu



	sample	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
1	1	Adelie	Torgersen	39.1	18.7	181	3750	male	2007
2	2	Adelie	Torgersen	39.5	17.4	186	3800	female	2007
3	3	Adelie	Torgersen	40.3	18.0	195	3250	female	2007
4	4	Adelie	Torgersen	NA	NA	NA	NA	NA	2007
5	5	Adelie	Torgersen	36.7	19.3	193	3450	female	2007

Showing 1 to 5 of 344 entries, 9 total columns

Provjera strukture tablice i tipa podataka.

```
# Provjera tipa i strukture objekta  
str(penguins)
```

```
## tibble [344 × 9] (S3: tbl_df/tbl/data.frame)  
## $ sample      : num [1:344] 1 2 3 4 5 6 7 8 9 10 ...  
## $ species     : chr [1:344] "Adelie" "Adelie" "Adelie" "Adelie" ...  
## $ island      : chr [1:344] "Torgersen" "Torgersen" "Torgersen" "Torgersen" ...  
## $ bill_length_mm : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...  
## $ bill_depth_mm : num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...  
## $ flipper_length_mm: num [1:344] 181 186 195 NA 193 190 181 195 193 190 ...  
## $ body_mass_g   : num [1:344] 3750 3800 3250 NA 3450 ...  
## $ sex          : chr [1:344] "male" "female" "female" NA ...  
## $ year         : num [1:344] 2007 2007 2007 2007 2007 ...
```

Pitanje na koje želimo odgovor je:

“Koja je prosječna masa pingvina vrste Adelie u kilogramima na svakom od otoka?”

Naredba `select()`

- Kako bi odgovorili na to pitanje, najlakše je stvoriti novi tablicu u kojoj ćemo **odabrati** samo one varijable koje su nam potrebne za izračun: `species`, `island` i `body_mass_g`.
- Naredba **`select()`** je funkcija iz `dplyr` paketa koja služi za odabir (selektiranje) specifičnih stupaca iz data frame-a. Pomaže u fokusiranju samo na one varijable (stupce) koje su potrebne za analizu, a ignorira ostatak podataka.
- Primjer: **`select(podaci, varijabla1, varijabla2, ...)`**

Naredba select()

Korak 1: Odabir relevantnih varijabli (stupaca)

```
select(penguins, # podaci
       species, # varijabla 1
       island, # varijabla 2
       body_mass_g)# varijabla 3
```

```
## # A tibble: 344 × 3
```

```
##   species island   body_mass_g
```

```
##   <chr>    <chr>         <dbl>
```

```
## 1 Adelie  Torgersen      3750
```

```
## 2 Adelie  Torgersen      3800
```

```
## 3 Adelie  Torgersen      3250
```

```
## 4 Adelie  Torgersen         NA
```

```
## 5 Adelie  Torgersen      3450
```

```
## 6 Adelie  Torgersen      3650
```

```
## 7 Adelie  Torgersen      3625
```

```
## 8 Adelie  Torgersen      4675
```

```
## 9 Adelie  Torgersen      3475
```

```
## 10 Adelie Torgersen      4250
```

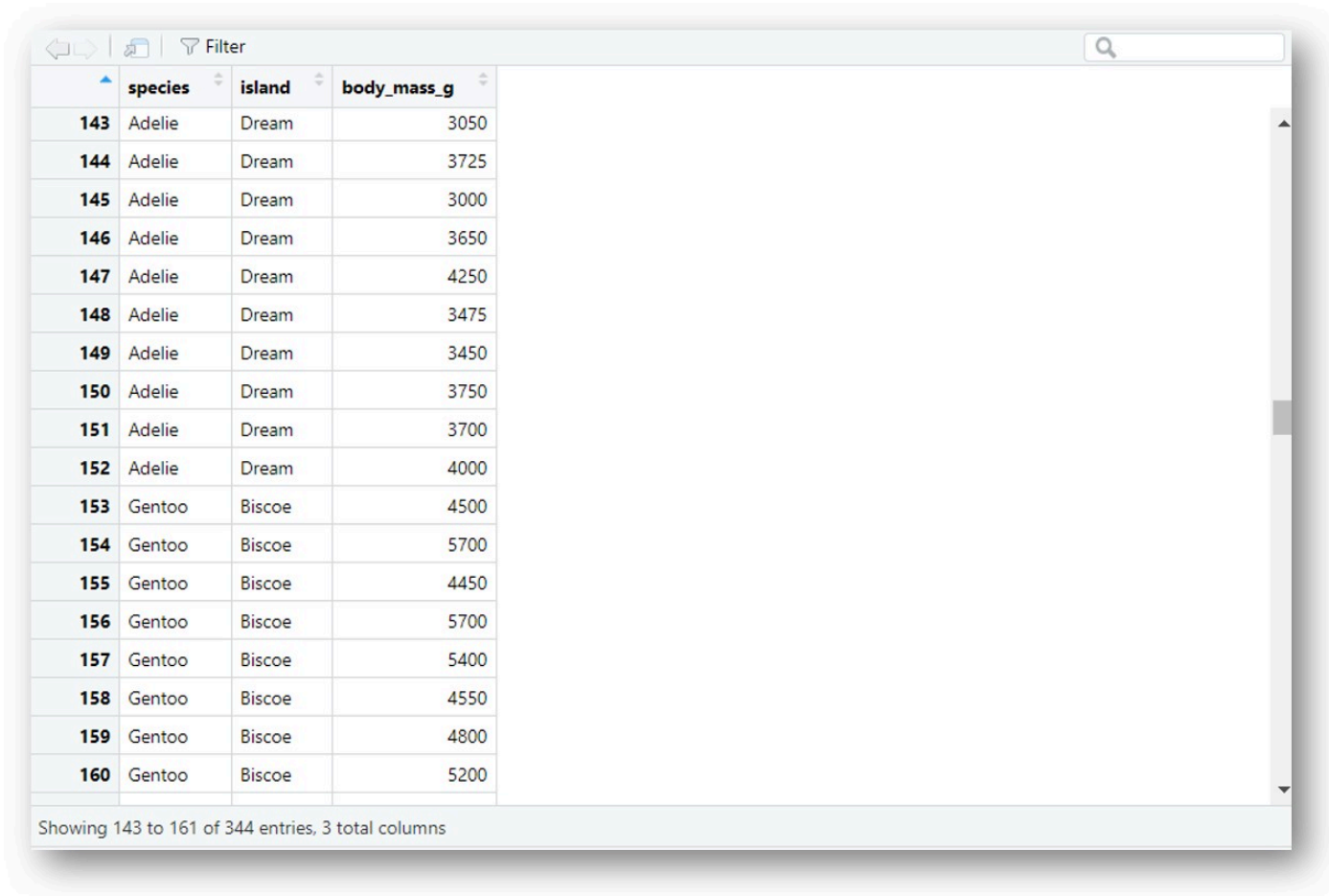
```
## # i 334 more rows
```

Gdje je objekt? Zašto nije u environmentu?

- Jer ga nismo spremili kao novi objekt!
- Kreirajmo novi objekt naziva “**penguins_selected**” u koji će se spremiti izabrane varijable.

```
# Ponovimo korak 1, ali kreirajmo novi objekt u koji će se spremiti  
penguins_selected <- select(penguins, species, island, body_mass_g)
```

View(penguins_selected) ili klik na objekt u environmentu za vizualizaciju nove tablice.



Filter

	species	island	body_mass_g
143	Adelie	Dream	3050
144	Adelie	Dream	3725
145	Adelie	Dream	3000
146	Adelie	Dream	3650
147	Adelie	Dream	4250
148	Adelie	Dream	3475
149	Adelie	Dream	3450
150	Adelie	Dream	3750
151	Adelie	Dream	3700
152	Adelie	Dream	4000
153	Gentoo	Biscoe	4500
154	Gentoo	Biscoe	5700
155	Gentoo	Biscoe	4450
156	Gentoo	Biscoe	5700
157	Gentoo	Biscoe	5400
158	Gentoo	Biscoe	4550
159	Gentoo	Biscoe	4800
160	Gentoo	Biscoe	5200

Showing 143 to 161 of 344 entries, 3 total columns

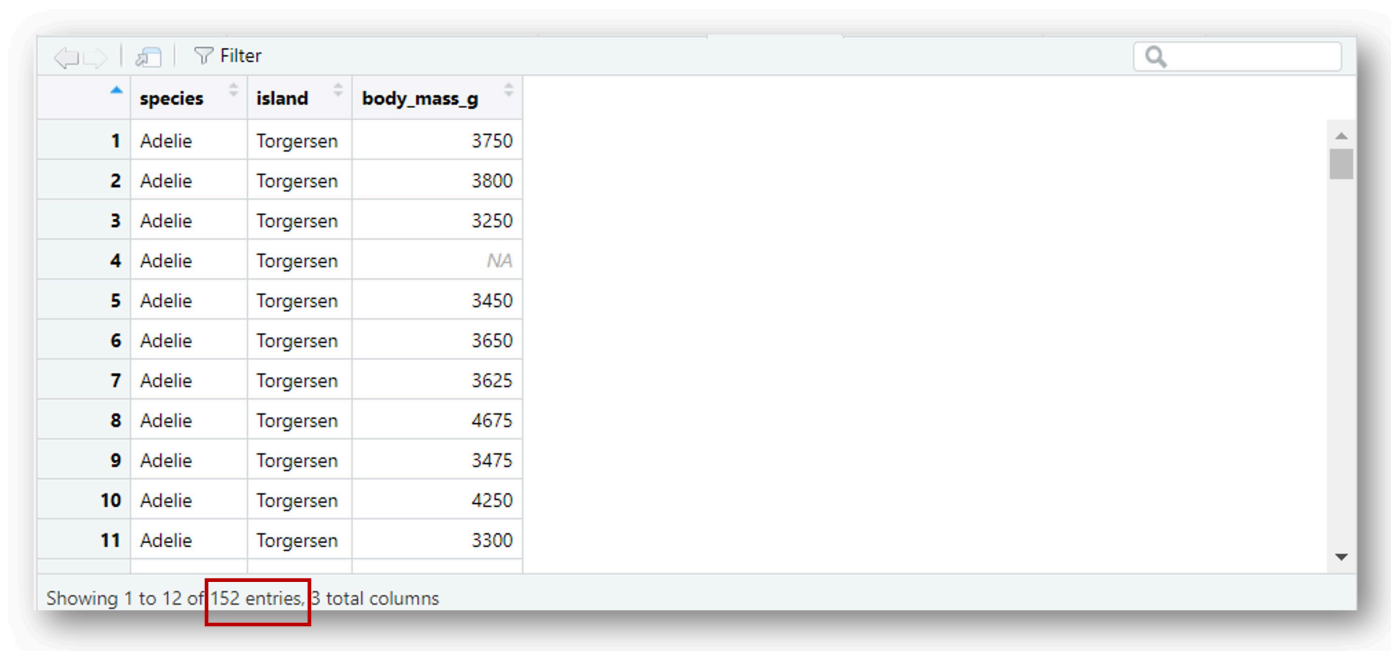
Funkcija `filter()`

- **`filter()`** je funkcija iz dplyr paketa koja služi za filtriranje redova u *data frame*-u.
- Zadržava samo one redove koji zadovoljavaju specificirane uvjete.
- Čitljivost - Jasno izražava uvjete u kodu.
- Fleksibilnost - Moguće kombinirati više uvjeta korištenjem logičkih operatora (&, |). Primjer:

Naredbom **`filter()`** želimo od svih redova s vrstama pingvina zadržati samo pripadnike vrste Adelie.

```
# Korak 2: Filtriranje uzoraka (redaka) vrste "Adelie"  
penguins_adelie <- filter(penguins_selected, # podaci  
                          species == "Adelie") # uvjet filtriranja
```

View(penguins_adelie) ili klik na objekt u environmentu za vizualizaciju nove tablice.



	species	island	body_mass_g
1	Adelie	Torgersen	3750
2	Adelie	Torgersen	3800
3	Adelie	Torgersen	3250
4	Adelie	Torgersen	NA
5	Adelie	Torgersen	3450
6	Adelie	Torgersen	3650
7	Adelie	Torgersen	3625
8	Adelie	Torgersen	4675
9	Adelie	Torgersen	3475
10	Adelie	Torgersen	4250
11	Adelie	Torgersen	3300

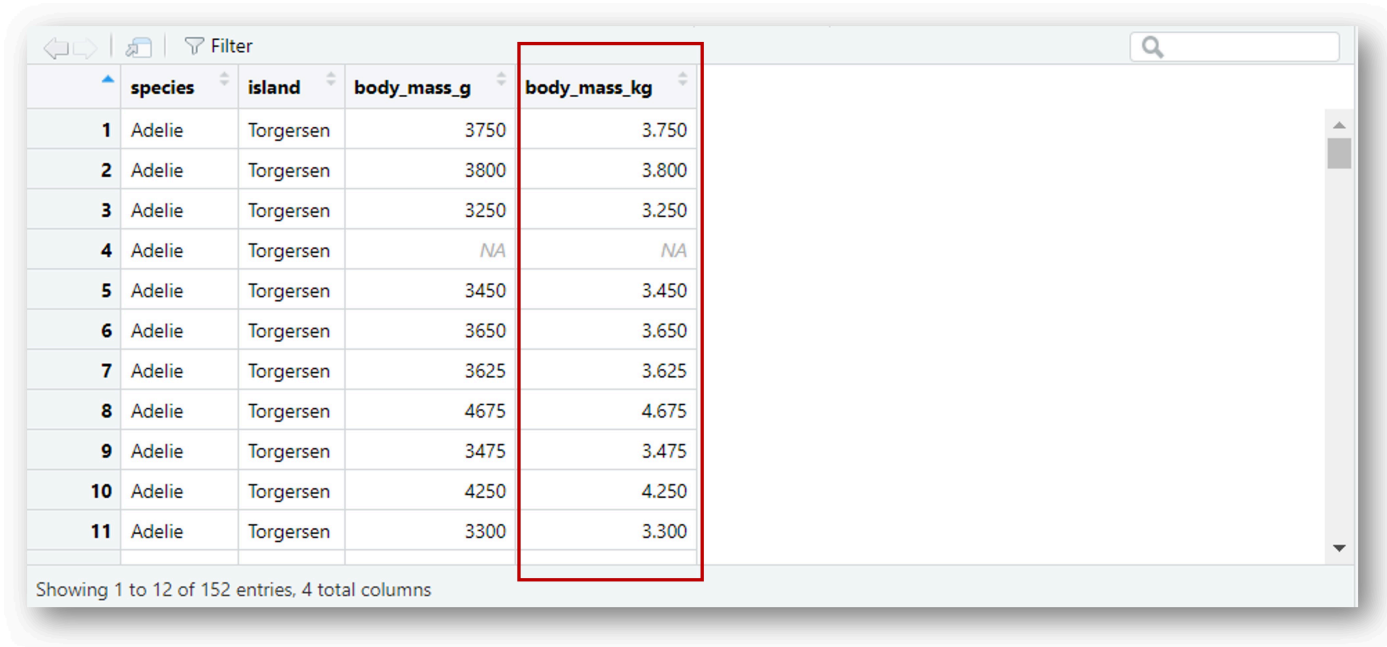
Showing 1 to 12 of 152 entries, 3 total columns

Funkcija `mutate()`

- **`mutate()`** je funkcija iz dplyr paketa koja služi za kreiranje novih stupaca (varijabli) ili modifikaciju postojećih unutar data frame-a.
- Pomaže u dodavanju izmjenjenih varijabli bez potrebe za kreiranjem novog data frame-a.
- koristit ćemo funkciju **`mutate()`** kako bi kreirali novu varijablu koja prikazuje masu pingvina u kilogramima umjesto u gramima.

```
# Korak 3: Kreiranje nove varijable koja sadrži masu izraženu u kilogramima  
penguins_mass_kg <- mutate(penguins_adelie, # podaci  
                           body_mass_kg = body_mass_g / 1000) # kreiranje nove varijable
```

View(penguins_mass_kg) ili klik na objekt u environmentu za vizualizaciju nove tablice.



A screenshot of a data viewer window, likely from RStudio, displaying a table of penguin data. The window has a header bar with navigation icons and a search bar. The table has four columns: 'species', 'island', 'body_mass_g', and 'body_mass_kg'. The 'body_mass_kg' column is highlighted with a red box. The table shows 11 rows of data, all for 'Adelie' species on 'Torgersen' island. The 'body_mass_g' column contains values in grams, and the 'body_mass_kg' column contains the corresponding values in kilograms. The status bar at the bottom indicates 'Showing 1 to 12 of 152 entries, 4 total columns'.

	species	island	body_mass_g	body_mass_kg
1	Adelie	Torgersen	3750	3.750
2	Adelie	Torgersen	3800	3.800
3	Adelie	Torgersen	3250	3.250
4	Adelie	Torgersen	NA	NA
5	Adelie	Torgersen	3450	3.450
6	Adelie	Torgersen	3650	3.650
7	Adelie	Torgersen	3625	3.625
8	Adelie	Torgersen	4675	4.675
9	Adelie	Torgersen	3475	3.475
10	Adelie	Torgersen	4250	4.250
11	Adelie	Torgersen	3300	3.300

Showing 1 to 12 of 152 entries, 4 total columns

Funkcija `group_by()` u R-u (dplyr)

- **`group_by()`** je funkcija iz dplyr paketa koja omogućava grupiranje podataka prema jednoj ili više varijabli.
- Koristi se često u kombinaciji s funkcijama poput **`summarise()`** za izvođenje agregatnih operacija unutar svake grupe.

```
# Korak 4: Zadavanje grupiranja i prikaza rezultata po otocima  
penguins_grouped <- group_by(penguins_mass_kg, # podaci  
                             island) # varijabla po kojoj želimo grupirati
```

Funkcija summarise()

- **summarise()** ili **summarize()** je funkcija iz dplyr paketa koja se koristi za sažimanje podataka na temelju agregatnih operacija.
- Najčešće se koristi u kombinaciji s **group_by()** kako bi se izračunale sumirane statistike unutar grupa.

```
# Korak 5: Kreiranje finalne sumirane tablice rezultata
penguins_result <- summarise(penguins_grouped, # podaci
                             average_mass = mean(body_mass_kg)) # nova varijabla za prosjek
```

```
# Ispis konačnog rezultata
print(penguins_result)
```

```
## # A tibble: 3 × 2
##   island    average_mass
##   <chr>         <dbl>
## 1 Biscoe         3.71
## 2 Dream          3.69
## 3 Torgersen      NA
```

Zašto nam se ne prikazuju podaci za Torgersen otok?

- Jer nismo uklonili nedostajuće vrijednosti!
- Koristiti funkciju `na.omit()`.

Funkcija `na.omit()`

- `na.omit()` funkcija iz *base* R-a koja se koristi za **uklanjanje redaka s nedostajućim vrijednostima (NA)** iz data frame-a ili vektora.
- Vraća filtrirani data frame bez redaka s NA vrijednostima.

```
# Kako bi mogli izračunati rezultat za otok Torgersen moramo ukloniti nedostajuće podatke  
# Uklanjanje uzoraka s nedostajućim podacima  
penguins_cleaned <- na.omit(penguins_mass_kg)  
  
# Ponovimo korake 4 i 5 s novom tablicom  
# Korak 4: Zadavanje grupiranja i prikaza rezultata po otocima  
penguins_grouped <- group_by(penguins_cleaned, island)
```



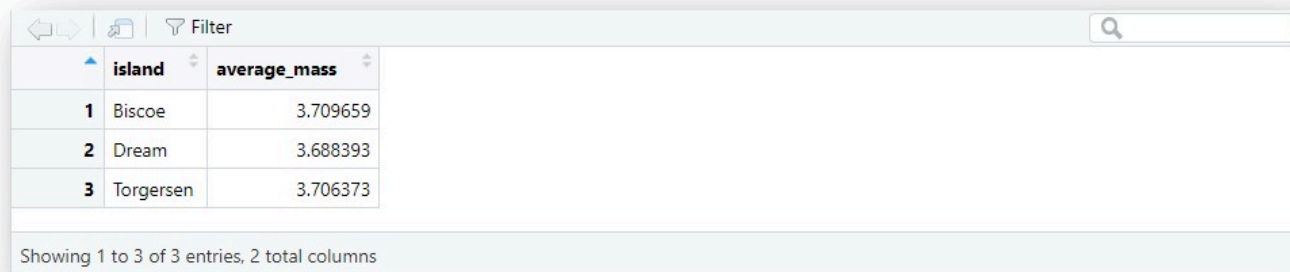
```
# Korak 5: Kreiranje finalne sumariziranje tablice rezultata
penguins_result <- summarise(penguins_grouped, average_mass = mean(body_mass_kg))

# Ispis konačnog rezultata
print(penguins_result)
```

```
## # A tibble: 3 × 2
##   island    average_mass
##   <chr>         <dbl>
## 1 Biscoe       3.71
## 2 Dream        3.69
## 3 Torgersen    3.71
```

Odgovor na postavljeno pitanje pitanje s početka:

“Prosječna masa pingvina vrste Adelie na otoku Biscoe i Torgersen iznosila je 3.71 kg, a na otoku Dream 3.69 kg.”



The screenshot shows a data table interface with a header bar containing navigation icons, a 'Filter' button, and a search box. The table has two columns: 'island' and 'average_mass'. It displays three rows of data. Below the table, a status bar indicates 'Showing 1 to 3 of 3 entries, 2 total columns'.

	island	average_mass
1	Biscoe	3.709659
2	Dream	3.688393
3	Torgersen	3.706373

Showing 1 to 3 of 3 entries, 2 total columns

Zadatak

- Koristeći gore naučene funkcije za manipulaciju podacima, kreirajte data frame koji će dati odgovor na pitanje:

“Koja je posječna masa u kilogramima pingvina vrste Gentoo mužjaka, a koja ženki?

Rješenje

Korak 1: Selektiranje relevantnih varijabli

```
penguins_selected_2 <- select(penguins, species, sex, body_mass_g)
```

Korak 2: Filtriranje uzoraka (redaka) vrste "Gentoo"

```
penguins_gentoo <- filter(penguins_selected_2, species == "Gentoo")
```

Korak 3: Kreiranje nove varijable koja sadrži masu izraženu u kilogramima

```
gentoo_mass_kg <- mutate(penguins_gentoo, body_mass_kg = body_mass_g / 1000)
```

Korak 4: Uklanjanje nedostajućih vrijednosti

```
gentoo_cleaned <- na.omit(gentoo_mass_kg)
```

Korak 4: Zadavanje grupiranja i prikaza rezultata po spolu

```
gentoo_grouped <- group_by(gentoo_cleaned, sex)
```

Korak 5: Kreiranje finalne sumiranja tablice rezultata

```
gentoo_result <- summarise(gentoo_grouped, average_mass = mean(body_mass_kg))
```

```
# Ispis konačnog rezultata  
print(gentoo_result)
```

```
## # A tibble: 2 × 2  
##   sex      average_mass  
##   <chr>      <dbl>  
## 1 female      4.68  
## 2 male       5.48
```

Odgovor: "Prosječna masa pingvina vrste Gentoo ženki iznosila je 4.68 kg, a mušjaka 5.48 kg."

Kako smanjiti količinu napisanog koda?

Pipe operator (%>%)

- Pipe operator (%>%) dolazi iz magrittr paketa (dio Tidyverse-a) i koristi se za povezivanje više funkcija na čitljiviji način.
- Omogućuje prosljeđivanje rezultata iz jedne funkcije kao ulaz u sljedeću funkciju bez potrebe za ugnježđivanjem.

Prednosti:

- Čitljivost – Kod je linearan i lakši za razumijevanje.
- Modularnost – Lako povezivanje različitih operacija bez pretrpavanja.
- Fleksibilnost – Može se koristiti s većinom funkcija.

Primjer pisanja koda pomoći pipe operatora

```
# Korištenje pipe operatora za smanjenje količine koda
adelie_result <- penguins %>% #podaci
  select(species, island, body_mass_g) %>% #odabir relevantnih varijabli
  filter(species == "Adelie") %>% #filtriranje samo pingvina vrste Adelie
  mutate(body_mass_kg = body_mass_g/1000) %>% #kreiranje nove varijable
  na.omit() %>% #uklanjanje nedostajućih vrijednosti
  group_by(island) %>% #grupiranje po otocima
  summarise(average_mass = mean(body_mass_kg)) #sumariziraj kao prosjek
print(adelie_result)
```

```
## # A tibble: 3 × 2
##   island      average_mass
##   <chr>         <dbl>
## 1 Biscoe         3.71
## 2 Dream          3.69
## 3 Torgersen      3.71
```

Rješenje zadatka pomoći pipe operatora

```
gentoo_result <- penguins %>%  
  select(species, sex, body_mass_g) %>%  
  filter(species == "Gentoo") %>%  
  mutate(body_mass_kg = body_mass_g/1000) %>%  
  na.omit() %>%  
  group_by(sex) %>%  
  summarise(average_mass = mean(body_mass_kg))  
  
print(gentoo_result)
```

```
## # A tibble: 2 × 2  
##   sex      average_mass  
##   <chr>         <dbl>  
## 1 female         4.68  
## 2 male           5.48
```