

# Rad s objektima i podacima u R-u

Lucija Kanjer, e-mail: [lucija.kanjer@biol.pmf.hr](mailto:lucija.kanjer@biol.pmf.hr)

2024-10-28

# 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

## Excel

- pretvaranje neuredne u (*untidy*) u urednu (*tidy*) tablicu

## R

- 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 (`%>%`)

# Tipovi tablica u analizi podataka

Tablice za vizualizaciju podataka - koriste se u radovima, izvještajima itd.

- Lako su čitljive ljudima, ne sadrže mnogo podataka, ne više od jedne stranice.
- Poruka tablice jasno je vidljiva, npr. prosjek jedne grupe je veći od ostalih.
- Sve vrijednosti jedne varijable su prikazane u istim mjernim jedinicama i s istim brojem decimalnih mjesta.

Tablice za analizu podataka

- Nazivaju se još *raw table*, *dataset*.
- Služe kao **uredno** spremište svih podataka na jednom mjestu.
- Koriste se kao input za računalne analize pa moraju biti napravljenije da ih programi koje korisrimo mogu čitati.
- U analizi podataka ovakve tablice se slažu u tzv. *tidy* tj. urednom formatu

# Untidy tablica

- Otvorite datoteku "kornjace\_untidy.xlsx"

SampleID	Turtle_ID	TurtleName	Age	length_cm	width(cm)	SamplingDate	rescue_centre	Location	DNA concentration
S1	TB175	Maksimus	Juvenile	45.2	44.2	28.7.2020.	Aquarium Pula	Porer island 44.113065, 15.232353	31,46 good
S2	TB181	CC_VIS_2001		42	37	29.6.2020.	Blue World Institute	Vis island 42.777562, 16.901862	27,11 good
S3	TB183	CC_VIS_2002		31	29	11.8.2020.		Vis island 42.959317, 17.141558	20,73 good
S4	TB185	CC_VIS_2003		31	28	20.8.2020.		Vis island 44.758325, 13.890561	100,06 good
S5	TB189	Valbiska		30	27.5	5.11.2020.		Krk island 44.983781, 13.754311	87,61 good
S6	TB195	FS35		45	41	10.12.2020.		Susak island lat: 42.90401 lon: 16.03579	19,58 poor
S7	TB201	Apox		36	34	14.1.2021.		Mali Lošinj lat: 42.9546 lon: 15.85258	27,26 good
S8	TB203	CC_LOŠINJ_2102		55	53.5	25.1.2021.		Lošinj island lat: 42.9536, lon: 16.09005	15,76 poor
S9	TB205	Zlata		35.5	33	3.2.2021.		Mali Lošinj 45.023144, 14.578583	15,35 poor
S10	TB207	Noemi		40.5	38	16.2.2021.		Veli Lošinj 44.506341, 14.285180	26,74 good
S11	TB209	Sanjin	Sub-adult	32	29	16.2.2021.	Aquarium Pula	Mali Lošinj 44.506341, 14.285180	26,66 good
S12	TB211	CC_LOŠINJ_2106		40	38	8.3.2021.		Lošinj island 44.506341, 14.285180	36,07 good
S13	TB217	Oliver Raul		26	24.2	23.4.2021.		Medulin 44.506341, 14.285180	19,57 good
S14	TB219	Martin		37.8	34.8	23.4.2021.		Mali Lošinj 44.546785, 14.447440	22,06 good
S15	TB221	CC_LOŠINJ_2112		31	28	25.5.2021.	Blue World Institute	Mali Lošinj 44.572978, 14.408607	21,01 good
S16	TB227	Calimero		52	52	6.4.2021.		Susak island 44.542943, 14.441804	22,37 good
S17	TB229	Marijana		26	23	6.4.2021.		Trstenika island 44.522182, 14.508105	25,94 good
S18	TB231	CC_LOŠINJ_2111		27	25.5	21.04.2021.		Mali Lošinj 44.546785, 14.447440	31,05 good
S19	TB159	Ella		62	58	30.6.2020.	Aquarium Pula	Zadar N 44.61342, E 14.40082	34,99 good
S20	TB163	Huanita		68	66	20.3.2020.		Lastovo island 43.886247, 15.193743	72,93 good
S21	TB167	Maro		67	63.2	21.7.2020.		Korčula island 44.807439, 13.937336	31,41 good
S22	TB177	Špela		68	67	3.8.2020.		Barbariga 44.534310, 14.478769	32,43 good
S23	TB191	FS94	Adult	67.5	66	30.11.2020.	Blue World Institute	Susak island N 44.56873, E 14.42382	24,68 good
S24	TB197	FS25		70	67	19.12.2020.		Susak island N 42.991, E 16.051	21,64 good
S25	TB199	FS60		73	68	19.12.2020.	Aquarium Pula	Susak island 44.519761, 14.181472	25,69 good
S26	TB215	Karlo Albano		73	73	25.3.2021.		Dugi island 44.671560, 14.580641	16,83 poor
S27	TB223	Bova		70	64	8.6.2021.	Blue World Institute	Vis island N 44.54774, E 14.43973	19,08 poor

# Sredite tablicu u *tidy* format!

- 1 varijabla = 1 stupac tablice
- 1 uzorak (opažanje) = 1 redak tablice
- 1 vrijednost = 1 kućica tablice

Pazite! U biologiji je često 1 uzorak = 1 jedinka, ali i ne mora biti tako! Ako np. uzorkujemo jedinku više puta, onda svako uzorkovanje = 1 opažanje i predstavlja 1 redak u tablici.

country	year	cases	population
Afghanistan	1999	745	15557071
Afghanistan	2000	2566	20595360
Brazil	1999	37737	17206362
Brazil	2000	80488	174504898
China	1999	210258	1272915272
China	2000	210766	128042583

variables

country	year	cases	population
Afghanistan	1999	745	15557071
Afghanistan	2000	2566	20595360
Brazil	1999	37737	17206362
Brazil	2000	80488	174504898
China	1999	210258	1272915272
China	2000	210766	128042583

observations

country	year	cases	population
Afghanistan	1999	745	15557071
Afghanistan	2000	2566	20595360
Brazil	1999	37737	17206362
Brazil	2000	80488	174504898
China	1999	210258	1272915272
China	2000	210766	128042583

values

# Najčešće greške

- Imena stupaca nisu varijable, nego vrijednosti.
- spajanje dvije varijable u isti stupac.
- spajanje ćelija istih vrijednosti u tablici.
- spajanje naziva stupaca/redova.
- puštanje praznih redova i stupaca.
- kreiranje više tablica u istom dokumentu.

# Preporuke

- maknuti sve boje
- maknuti svo formatiranje (podebljani font, kurziv, crte između tablica)
- ne pisati razmak u imenima stupaca (varijabli)
- pisati imena varijabli istim stilom
- uobičajeni stilovi pisanja varijabli u R-u: `ime_varijable` i `ImeVarijable`
- pišite puna imena varijabli, a ne skraćenice (npr. M i F može značite *male* i *female*, a može značiti i *mother* i *father*)



# Rezultat - *tidy* tablica kornjača!

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	sample_ID	turtle_ID	turtle_name	age	length_cm	width_cm	sampling_date	rescue_centre	location	latitude	longitude	DNA_concentration	DNA_quality
2	S1	TB175	Maksimus	Juvenile	45.2	44.2	28.7.2020.	Aquarium Pula	Porer island	15.23235	44.11306	31,46	good
3	S2	TB181	CC_VIS_2001	Juvenile	42	37	29.6.2020.	Blue World Institute	Vis island	16.90186	42.77756	27,11	good
4	S3	TB183	CC_VIS_2002	Juvenile	31	29	11.8.2020.	Blue World Institute	Vis island	17.14156	42.95932	20,73	good
5	S4	TB185	CC_VIS_2003	Juvenile	31	28	20.8.2020.	Blue World Institute	Vis island	13.89056	44.75832	100,06	good
6	S5	TB189	Valbiska	Juvenile	30	27.5	5.11.2020.	Blue World Institute	Krk island	13.75431	44.98378	87,61	good
7	S6	TB195	FS35	Juvenile	45	41	10.12.2020.	Blue World Institute	Susak island	16.03579	42.90401	19,58	poor
8	S7	TB201	Apox	Juvenile	36	34	14.1.2021.	Blue World Institute	Mali Lošinj	15.85258	42.9546	27,26	good
9	S8	TB203	CC_LOŠINJ_2102	Juvenile	55	53.5	25.1.2021.	Blue World Institute	Lošinj island	16.09005	42.9536	15,76	poor
10	S9	TB205	Zlata	Juvenile	35.5	33	3.2.2021.	Blue World Institute	Mali Lošinj	14.57858	45.02314	15,35	poor
11	S10	TB207	Noemi	Juvenile	40.5	38	16.2.2021.	Blue World Institute	Veli Lošinj	14.28518	44.50634	26,74	good
12	S11	TB209	Sanjin	Juvenile	32	29	16.2.2021.	Blue World Institute	Mali Lošinj	14.28518	44.50634	26,66	good
13	S12	TB211	CC_LOŠINJ_2106	Juvenile	40	38	8.3.2021.	Blue World Institute	Lošinj island	14.28518	44.50634	36,07	good
14	S13	TB217	Oliver Raul	Juvenile	26	24.2	23.4.2021.	Aquarium Pula	Medulin	14.28518	44.50634	19,57	good
15	S14	TB219	Martin	Juvenile	37.8	34.8	23.4.2021.	Aquarium Pula	Mali Lošinj	14.44744	44.54678	22,06	good
16	S15	TB221	CC_LOŠINJ_2112	Juvenile	31	28	25.5.2021.	Blue World Institute	Mali Lošinj	14.40861	44.57298	21,01	good
17	S16	TB227	Calimero	Juvenile	52	52	6.4.2021.	Blue World Institute	Susak island	14.4418	44.54294	22,37	good
18	S17	TB229	Marijana	Juvenile	26	23	6.4.2021.	Blue World Institute	Trstenika island	14.50811	44.52218	25,94	good
19	S18	TB231	CC_LOŠINJ_2111	Juvenile	27	25.5	21.04.2021.	Blue World Institute	Mali Lošinj	14.44744	44.54678	31,05	good
20	S19	TB159	Ella	Sub-adult	62	58	30.6.2020.	Aquarium Pula	Zadar	14.40082	44.61342	34,99	good
21	S20	TB163	Huanita	Sub-adult	68	66	20.3.2020.	Aquarium Pula	Lastovo island	15.19374	43.88625	72,93	good
22	S21	TB167	Maro	Sub-adult	67	63.2	21.7.2020.	Aquarium Pula	Korčula island	13.93734	44.80744	31,41	good
23	S22	TB177	Špela	Sub-adult	68	67	3.8.2020.	Aquarium Pula	Barbariga	14.47877	44.53431	32,43	good
24	S23	TB191	FS94	Sub-adult	67.5	66	30.11.2020.	Blue World Institute	Susak island	14.42382	44.56873	24,68	good
25	S24	TB197	FS25	Adult	70	67	19.12.2020.	Blue World Institute	Susak island	16.051	42.991	21,64	good
26	S25	TB199	FS60	Adult	73	68	19.12.2020.	Blue World Institute	Susak island	14.18147	44.51976	25,69	good
27	S26	TB215	Karlo Albano	Adult	73	73	25.3.2021.	Aquarium Pula	Dugi island	14.58064	44.67156	16,83	poor
28	S27	TB223	Bova	Adult	70	64	8.6.2021.	Blue World Institute	Vis island	14.43973	44.54774	19,08	poor
29													

# 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** – Poboľjšani rad s tablicama, alternativa data.frame-u.

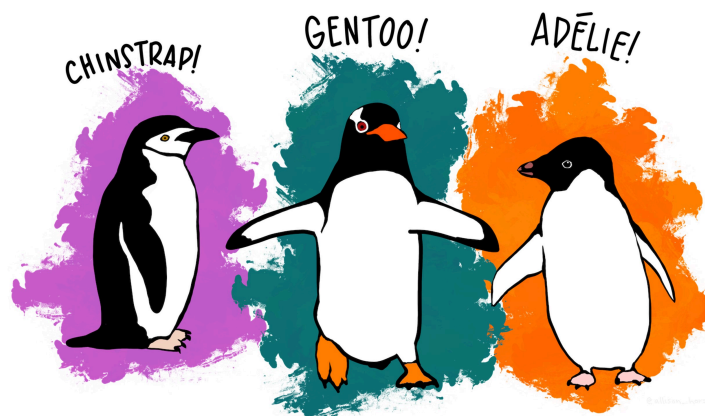
# Učitajmo tidyverse u R radno okruženje!

```
# 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)
```

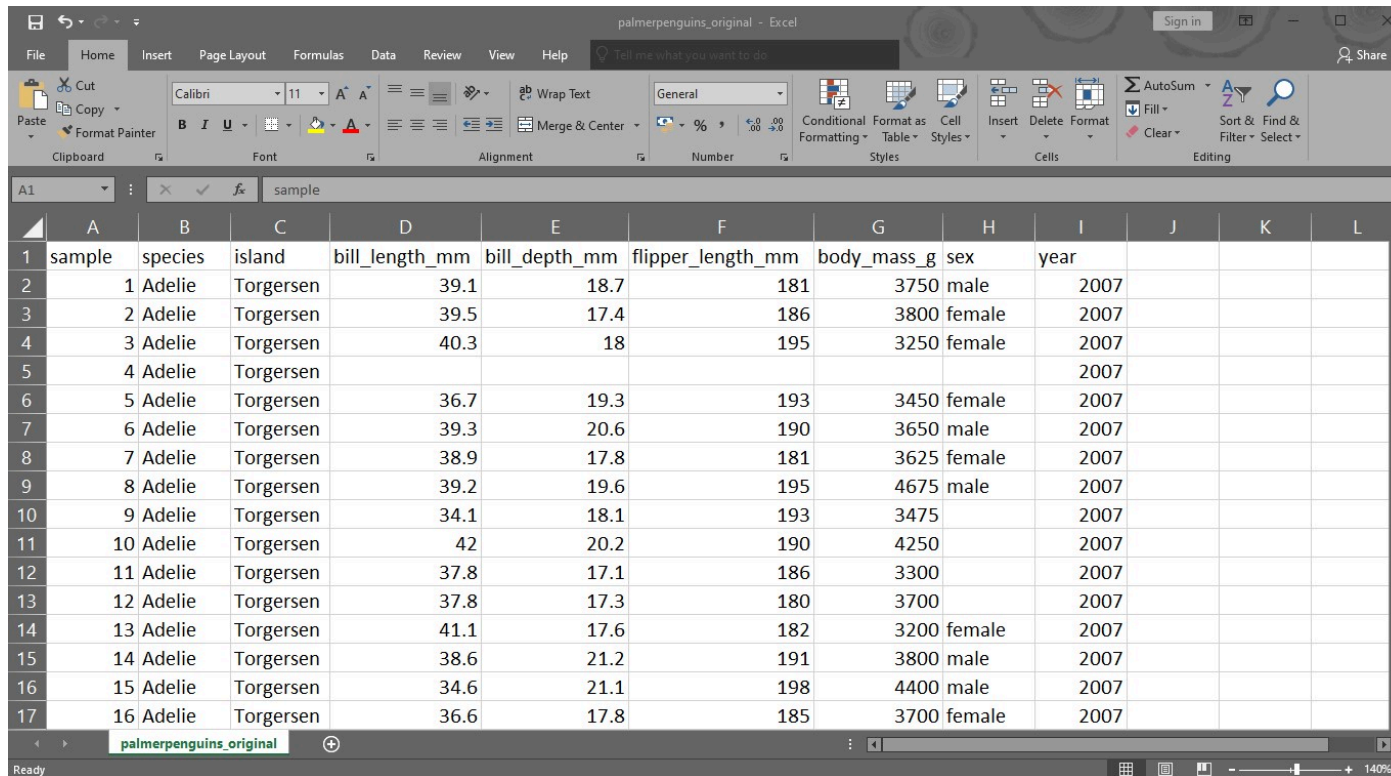
# Set podataka o Palmer pingvinima



- 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()
```

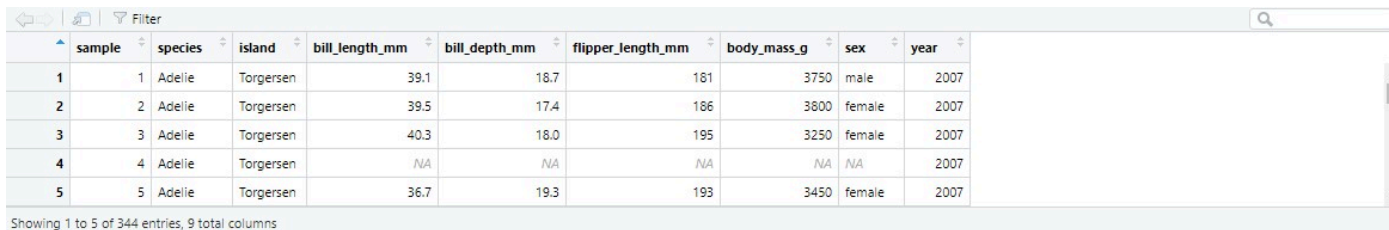
```
## [1] "C:/Users/Hrvoje/Documents/APUBI/03_Rad_s_podacima"
```

```
# 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
```

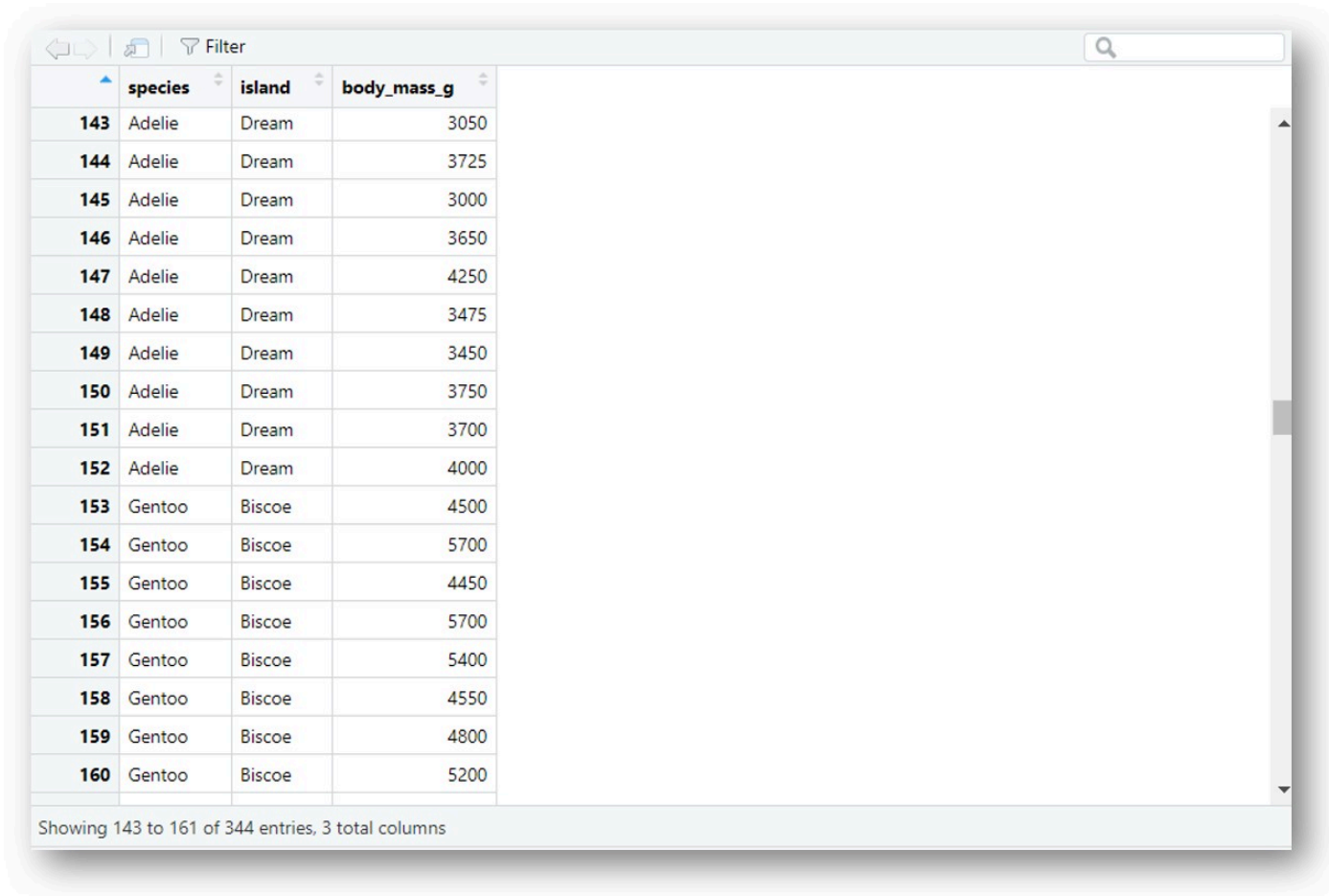
```
## # i 334 more rows
```

# 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.



	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

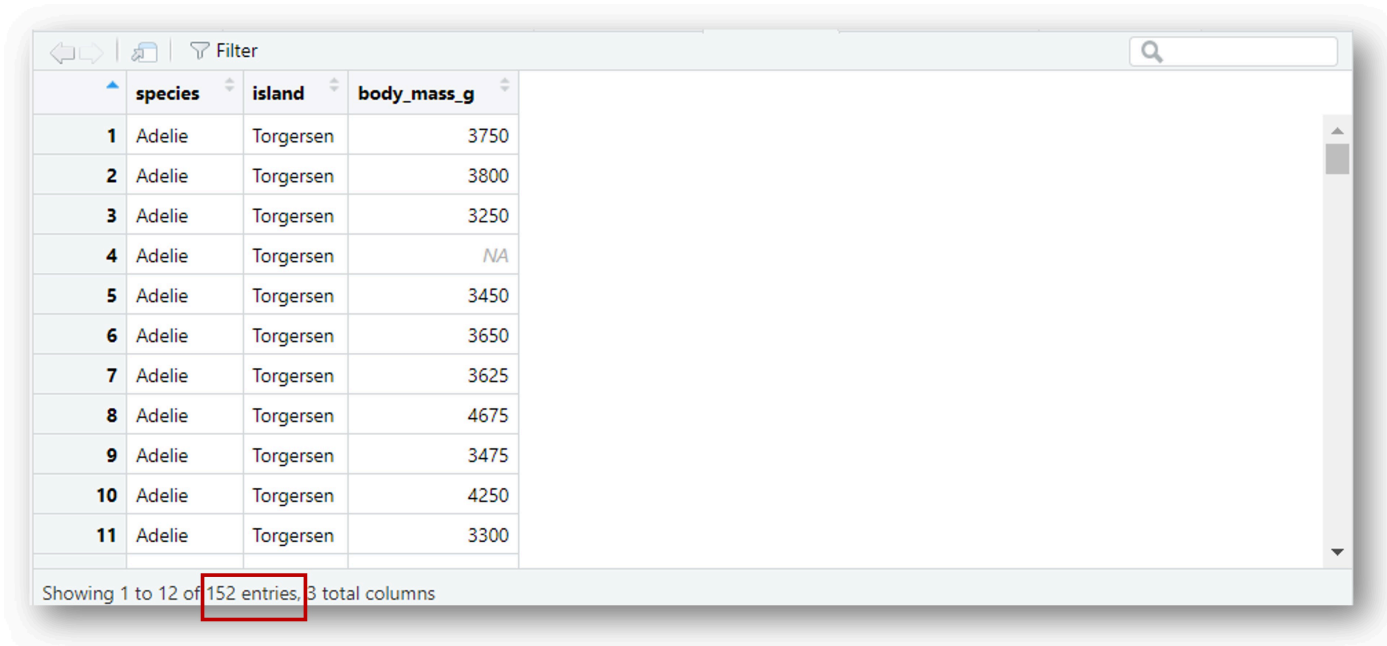
# Funkcija `filter()`

- **`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

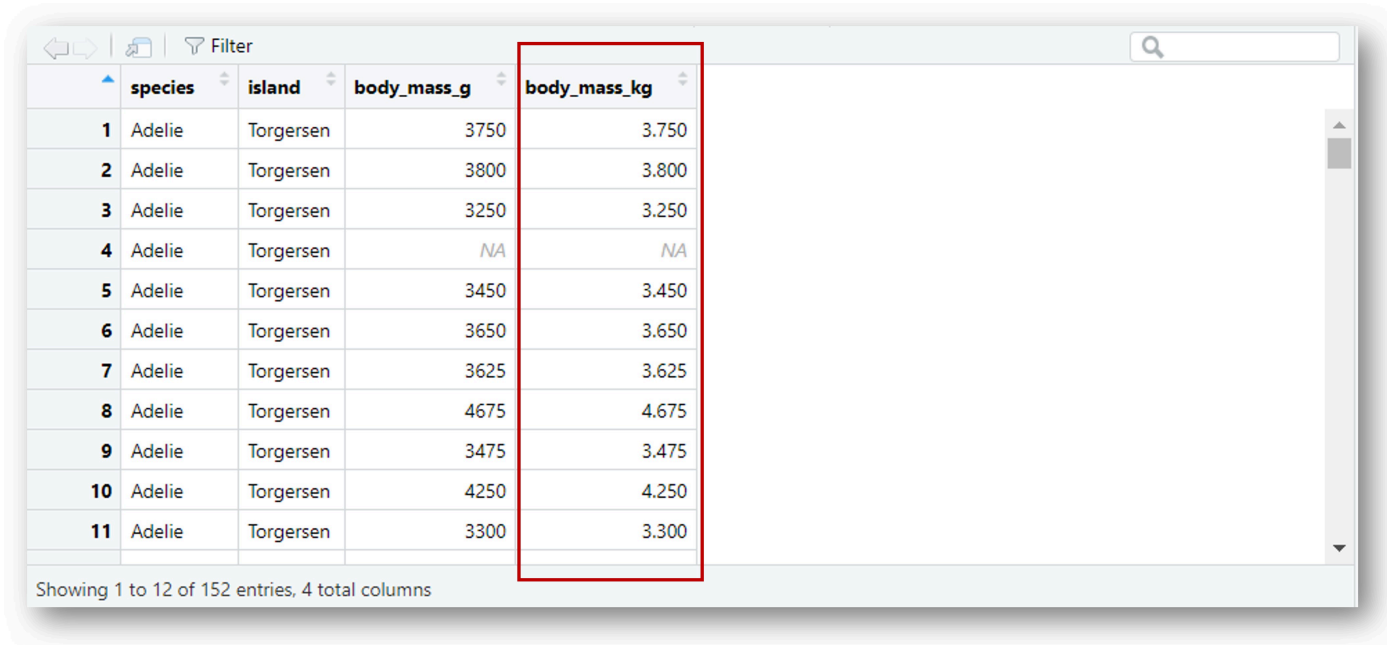


# Funkcija `mutate()`

- **`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.



The screenshot shows a data viewer window with a table of penguin data. 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 displays 11 rows of data, all for 'Adelie' penguins from 'Torgersen' island. The 'body\_mass\_g' column contains values ranging from 3250 to 4675, with one 'NA' value. The 'body\_mass\_kg' column contains the corresponding values in kilograms, ranging from 3.250 to 4.675, with one 'NA' value. 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)
```

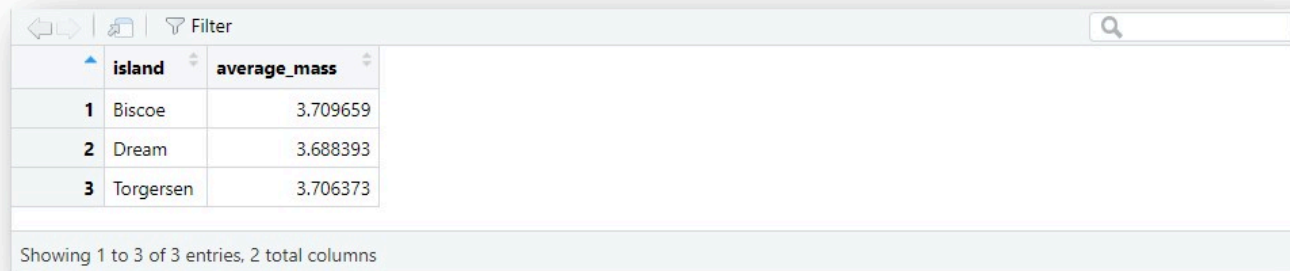
```
# Korak 5: Kreiranje finalne sumariziranje tablice rezultata
penguins_result <- summarise(penguins_grouped, average_mass = mean(body_mass_kg))

# 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 search bar and a filter icon. 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
```