Grafički prikaz bioloških podataka u R-u

Lucija Kanjer

2024

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 biloških 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

- ggplot2 paket - slojevi i estetika
- Različiti načini grafičkih prikaza - base R vs. ggplot2
- Grafički prikazi po tipu varijabli

R Paket: ggplot2

Opis

- ggplot2 je jedan od najpopularnijih paketa za vizualizaciju podataka u R-u.
- Zasnovan je na *Grammar of Graphics*, što omogućava korisnicima da sloje grafičke elemente i prilagođavaju ih na različite načine.
- Fleksibilan je za stvaranje složenih grafika uz relativno jednostavan i intuitivan kod.

Ključne karakteristike ggplot2 paketa

- Jednostavnost u izradi prilagođenih vizualizacija.
- Podržava različite tipove grafikona: histogrami, scatterplot-ovi, boxplot-ovi, line grafikoni i mnogi drugi.
- Omogućava kombiniranje više plotova u jedan prikaz.
- Visok stupanj prilagodbe: boje, naslovi, osi, oznake, veličina i drugi estetski elementi.

Učitavanje paketa

```
library(readxl) # excel tablice
library(ggplot2) # paket za crtanje grafova
library(dplyr) # paket za manipulaciju tablicama
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Postavljanje radnog direktorija

```
# Postavljanje  radnog direktorija
getwd()
```

```
## [1] "C:/Users/lucij/Documents/APUBI/04_Grafički_prikaz"
```

```
setwd("C:/Users/lucij/Documents/APUBI/04_Grafički_prikaz")
```

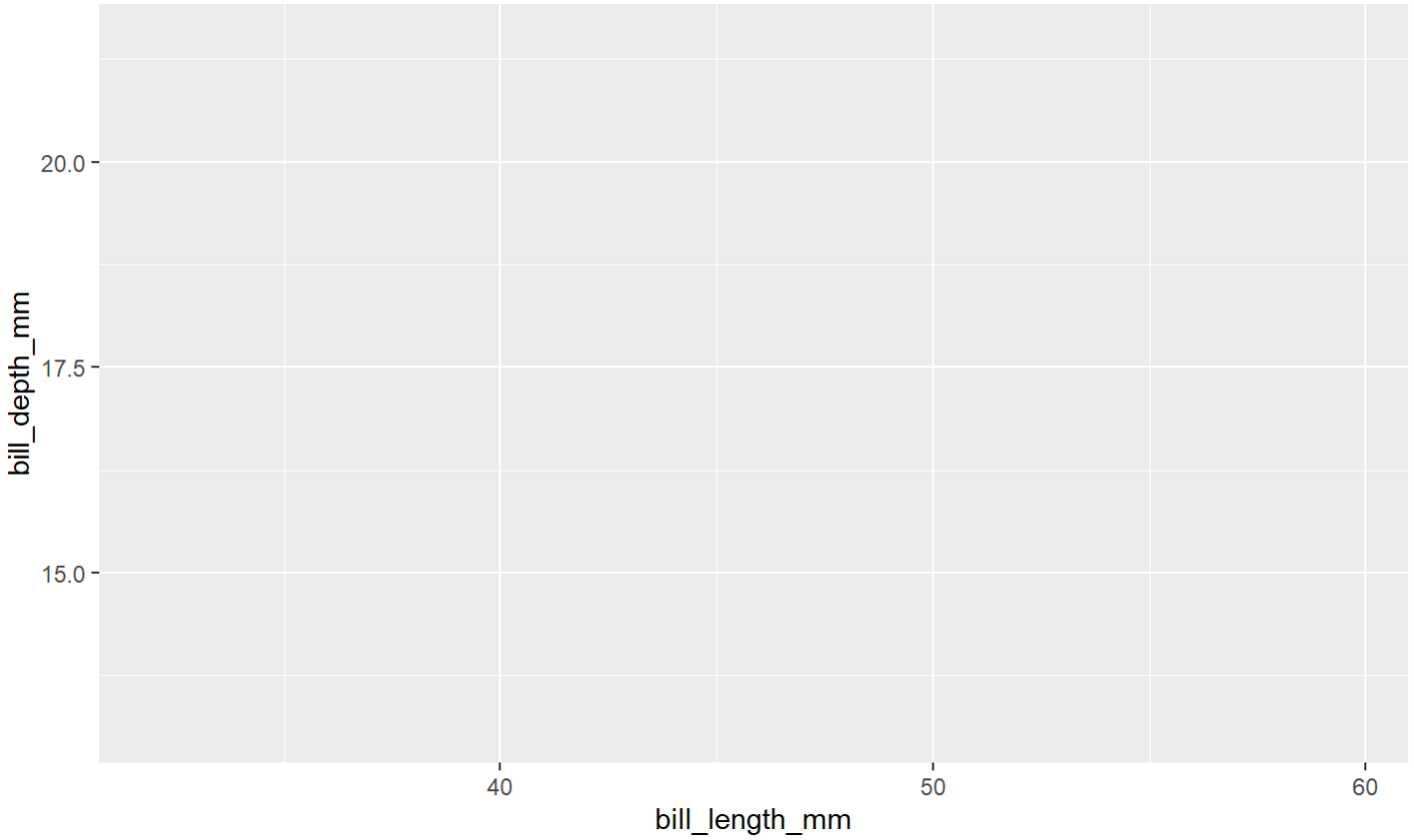
Učitavanje seta podataka o pingvinima

```
penguins <- read_excel("pingvini.xlsx")

View(penguins)
```

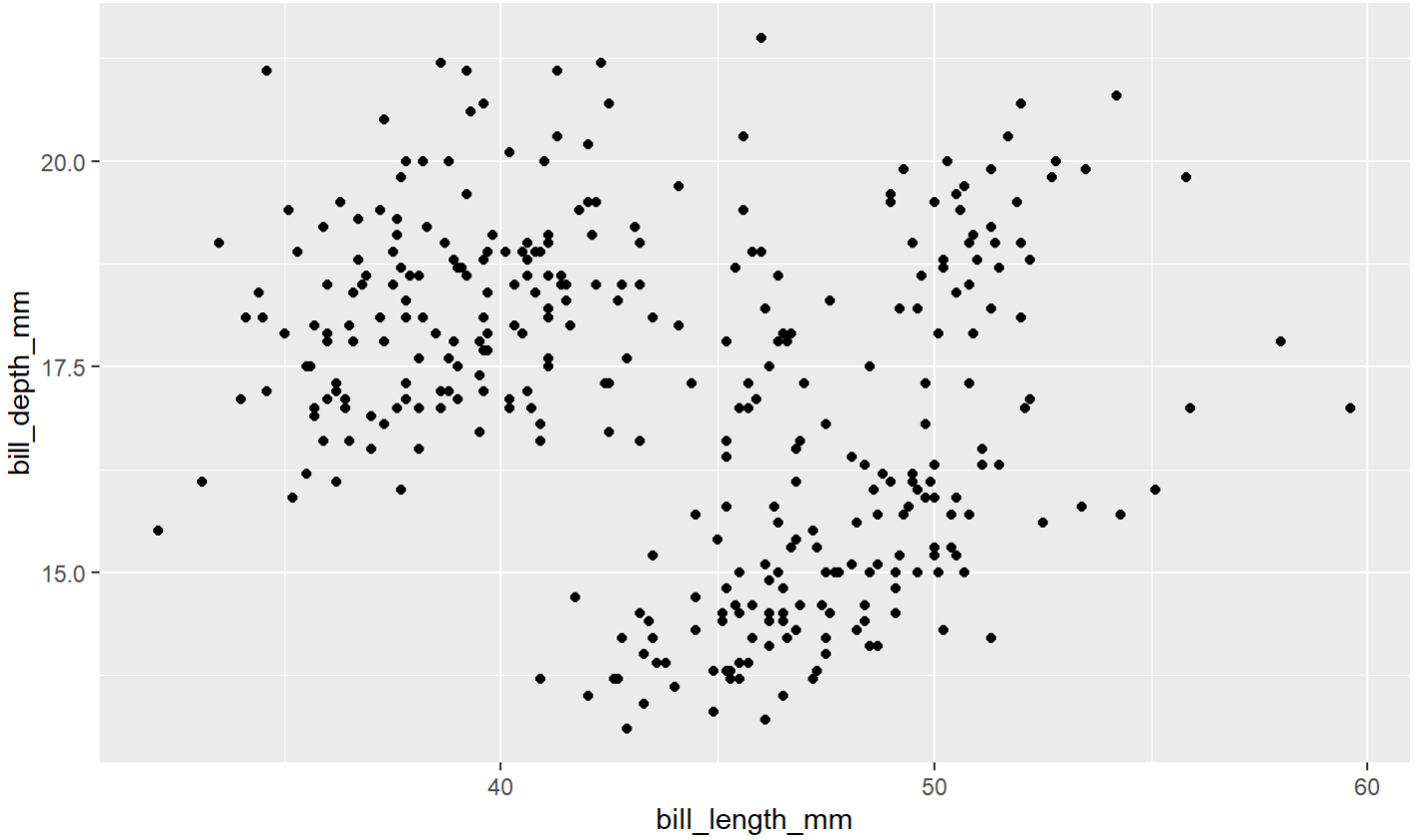
ggplot2 paket - slojevi i estetika

```
# 1. Osnovni graf bez slojeva

ggplot() # osnovna naredba
```

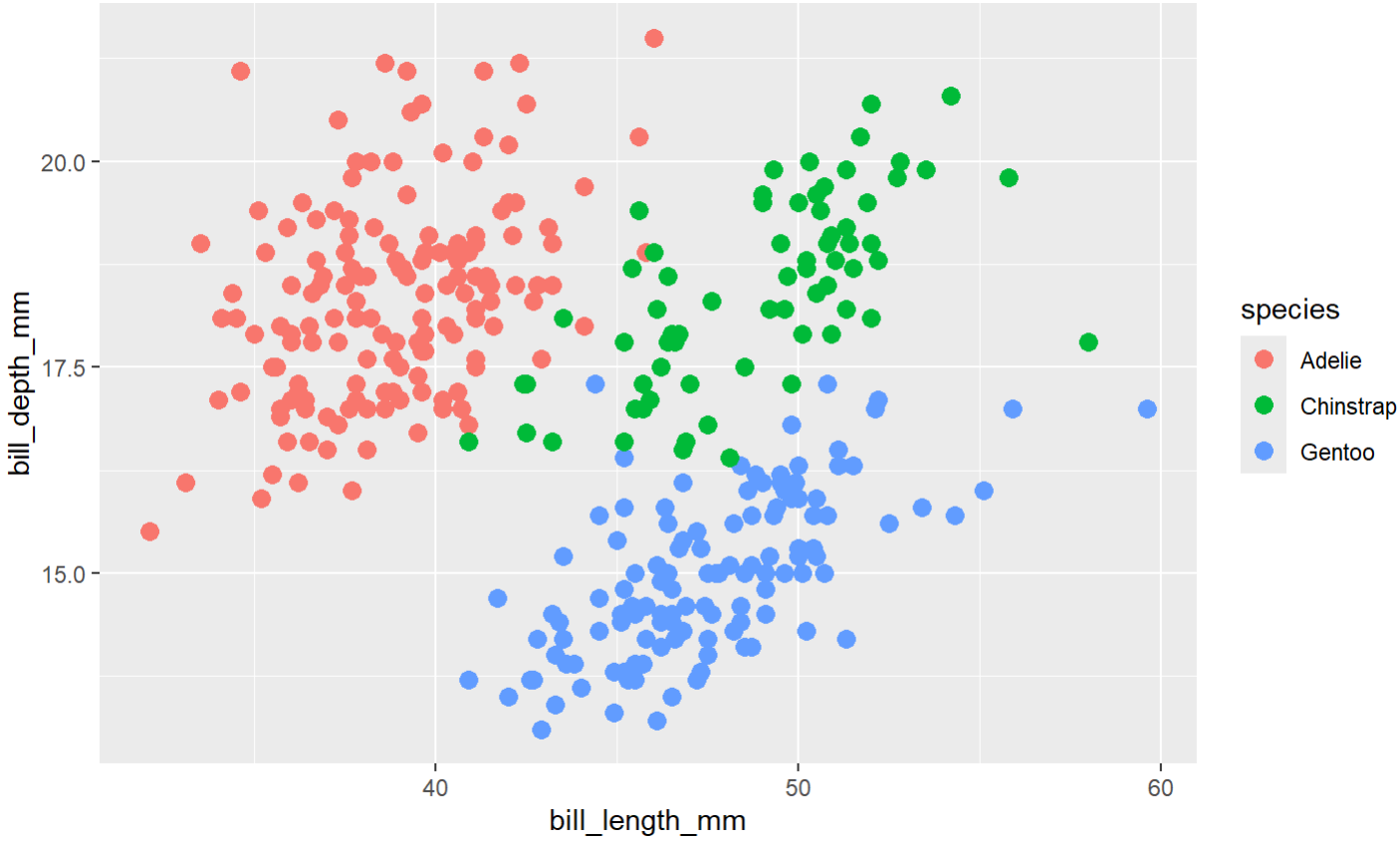
```r
# Osnovni grafikon - postavljamo estetiku, ali bez sloja
ggplot(data = penguins, aes(x = bill_length_mm, y = bill_depth_mm))
```



```r
# 2. Dodavanje prvog geometrijskog sloja: scatter plot
ggplot(data = penguins, aes(x = bill_length_mm, y = bill_depth_mm)) +
  geom_point()
```

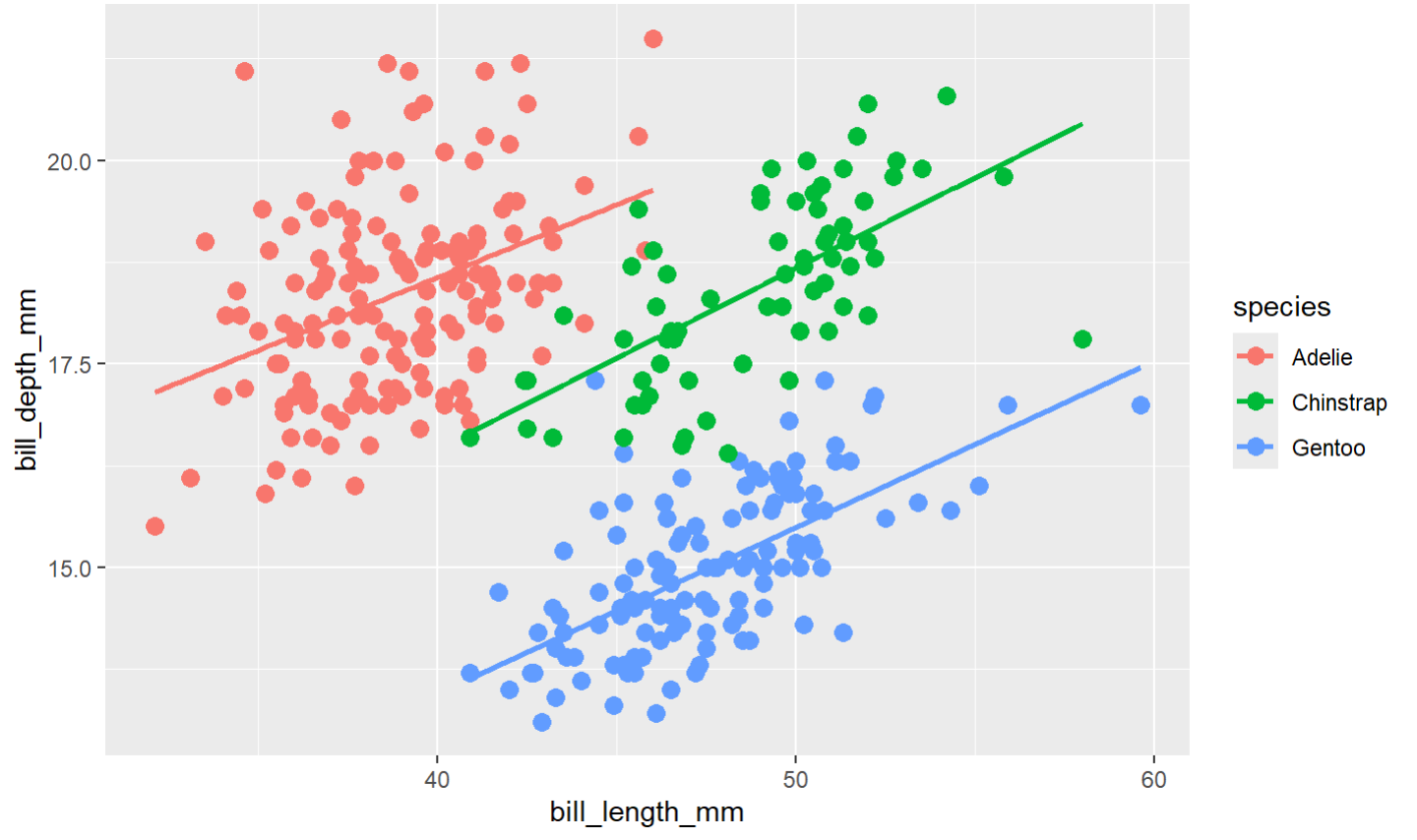


```r
# 3. Dodavanje boje kao estetike i većih točaka
ggplot(data = penguins, aes(x = bill_length_mm, y = bill_depth_mm, color = species)) +
  geom_point(size = 3)
```



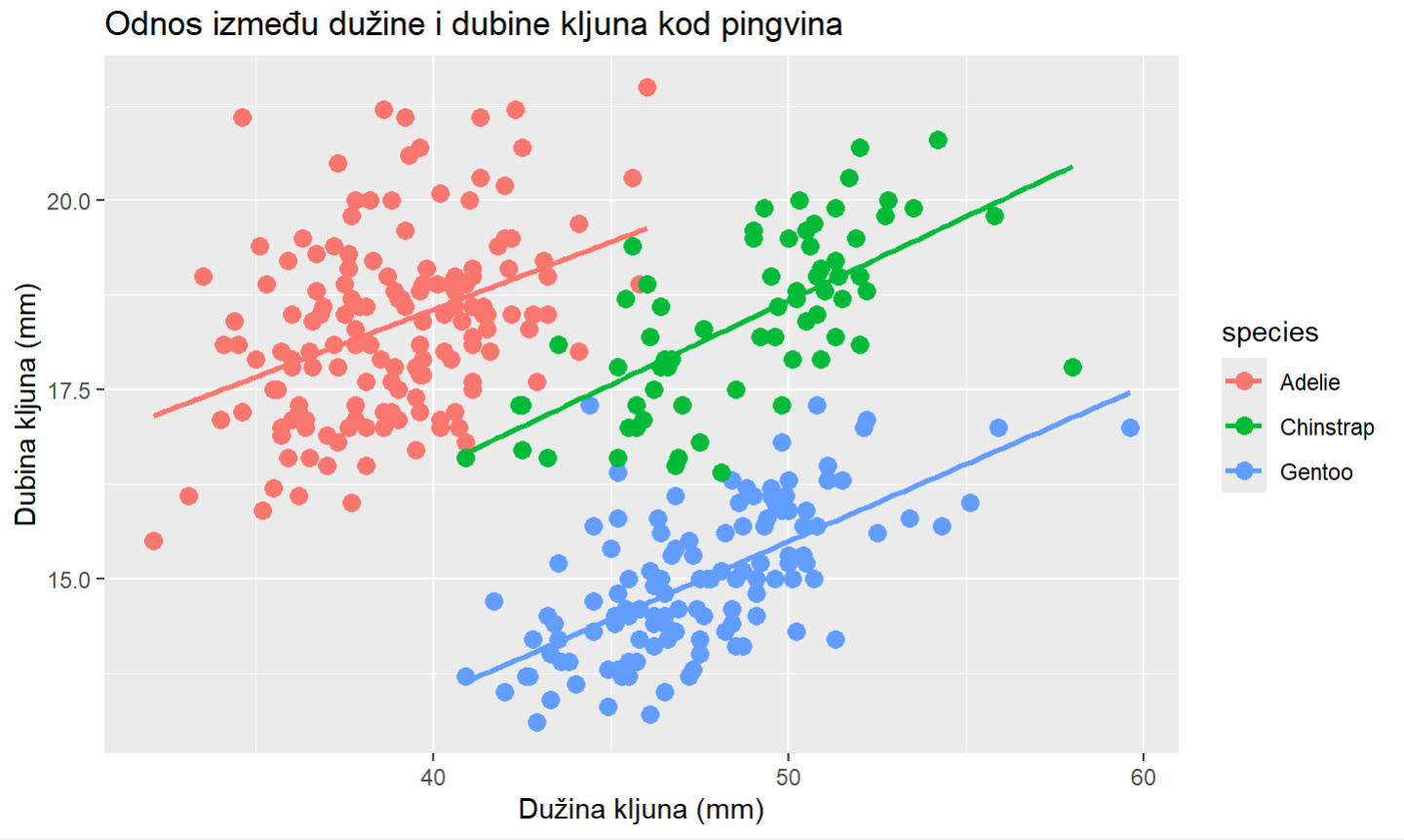
```r
# 4. Dodavanje trenda sa geom_smooth()
ggplot(data = penguins, aes(x = bill_length_mm, y = bill_depth_mm, color = species)) +
  geom_point(size = 3) +
  geom_smooth(method = "lm", se = FALSE)  # linearna regresija bez prikaza greške
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



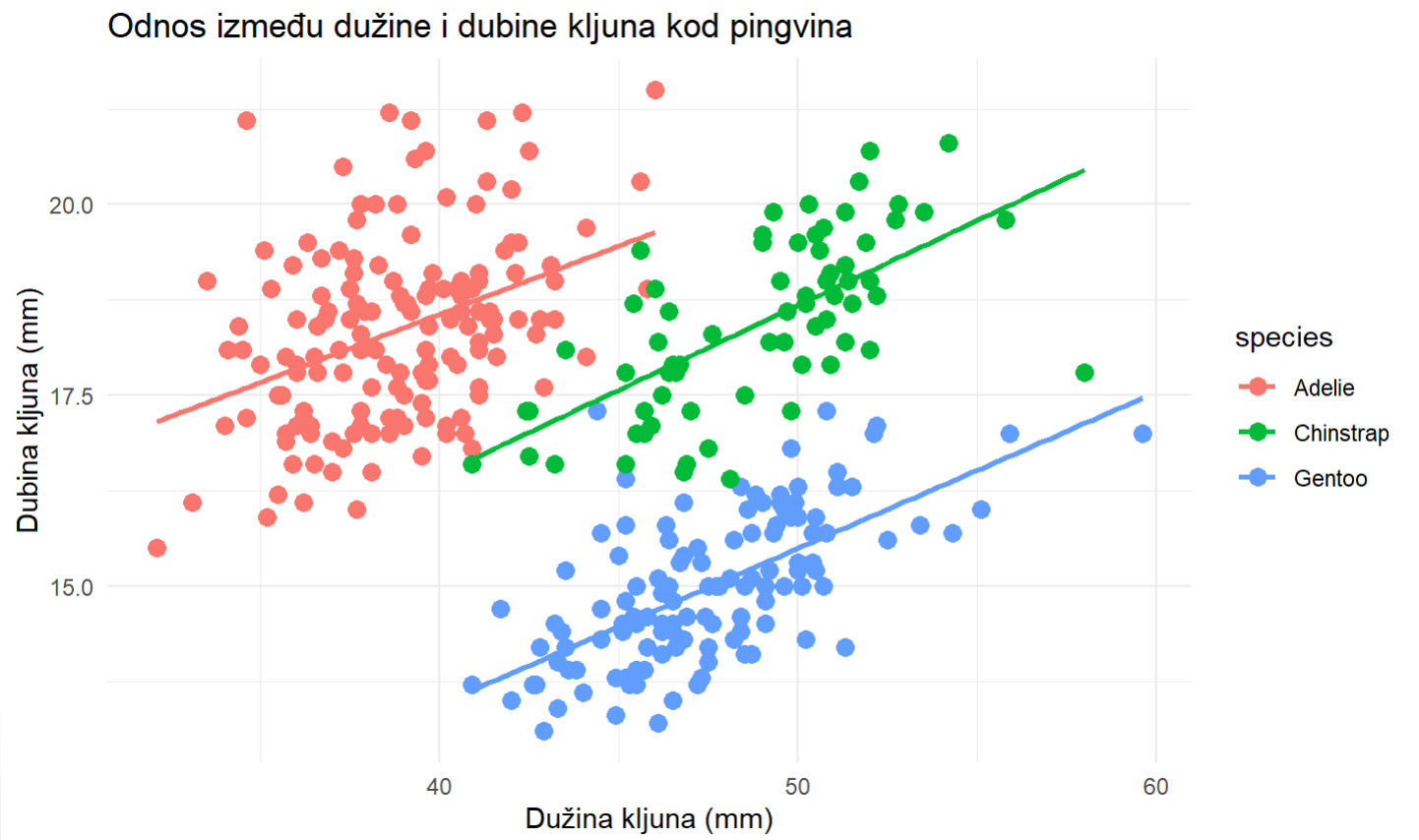
```r
# 5. Dodavanje naslova i oznaka osi sa slojem labs()
ggplot(data = penguins, aes(x = bill_length_mm, y = bill_depth_mm, color = species)) +
  geom_point(size = 3) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Odnos između dužine i dubine kljuna kod pingvina",
       x = "Dužina kljuna (mm)", y = "Dubina kljuna (mm)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



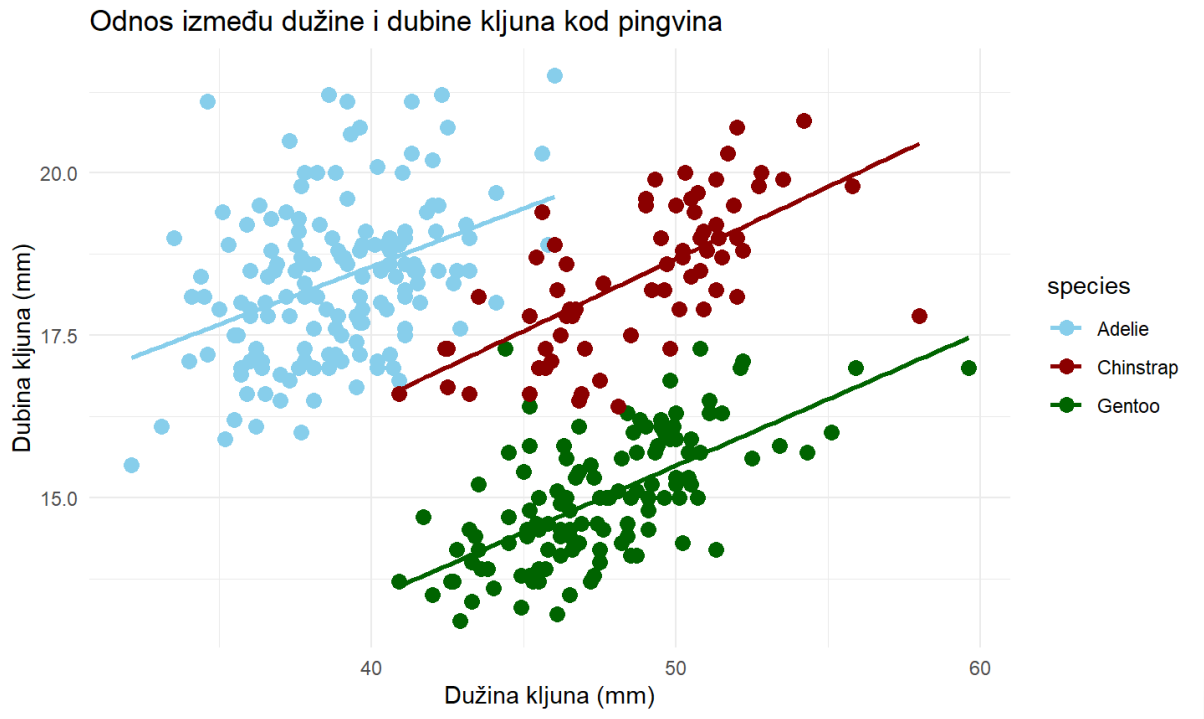
```r
# 6. Podešavanje tema sa slojem theme()
ggplot(data = penguins, aes(x = bill_length_mm, y = bill_depth_mm, color = species)) +
  geom_point(size = 3) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Odnos između dužine i dubine kljuna kod pingvina",
       x = "Dužina kljuna (mm)", y = "Dubina kljuna (mm)") +
  theme_minimal()  # Minimalna tema za čist izgled
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```r
# 7. Finalno prilagođavanje: promjena skale boja
ggplot(data = penguins, aes(x = bill_length_mm, y = bill_depth_mm, color = species)) +
  geom_point(size = 3) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Odnos između dužine i dubine kljuna kod pingvina",
       x = "Dužina kljuna (mm)", y = "Dubina kljuna (mm)") +
  scale_color_manual(values = c("Adelie" = "skyblue", "Chinstrap" = "darkred", "Gentoo" = "darkgreen")) +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Zadatak

- prisjetite se gradive prošle vježbe - manipulacija tablicom i napravite 3 grafa po uzoru na gornji
- svaki graf neka pokazuje odnos duljine i dubine kljuna za 1 vrstu!

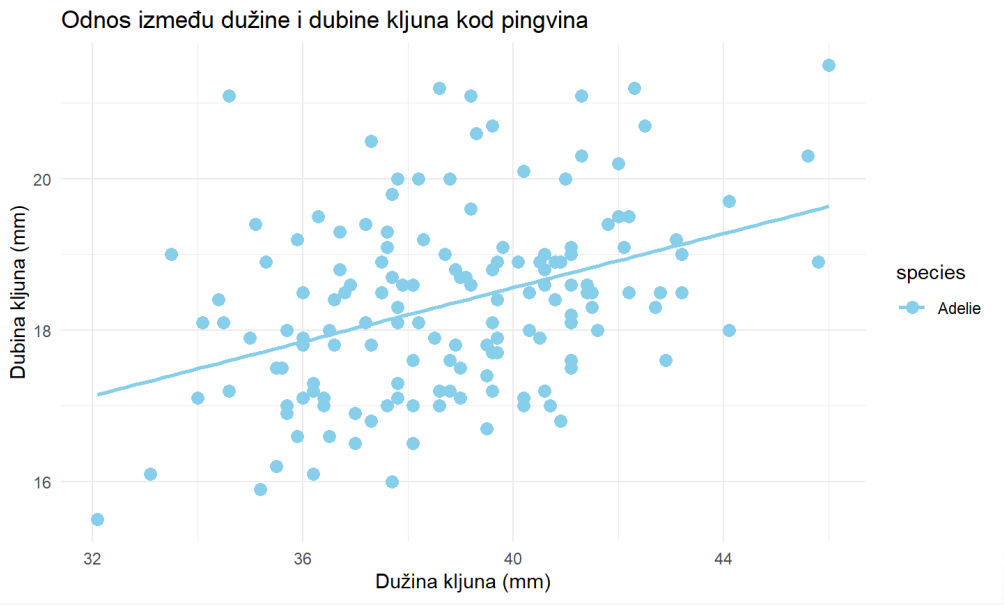
Rješenje

```r
# Izrada tablice za svaku vrstu naredbom filter()
adelie <- filter(penguins, species == "Adelie")
chinstrap <- filter(penguins, species == "Chinstrap")
gentoo <- filter(penguins, species == "Gentoo")
```

Rješenje

```
# adelie graf
ggplot(data = adelie, aes(x = bill_length_mm, y = bill_depth_mm, color = species)) +
  geom_point(size = 3) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Odnos između dužine i dubine kljuna kod pingvina",
       x = "Dužina kljuna (mm)", y = "Dubina kljuna (mm)") +
  scale_color_manual(values = c("Adelie" = "skyblue")) +
  theme_minimal()
```

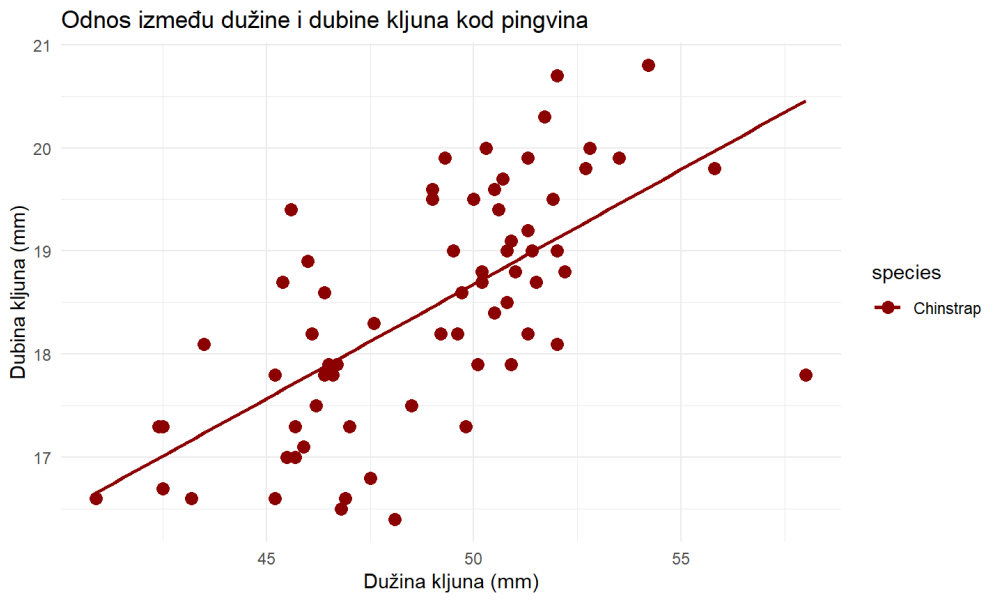
```
## `geom_smooth()` using formula = 'y ~ x'
```



Rješenje

```
# chinstrap graf
ggplot(data = chinstrap, aes(x = bill_length_mm, y = bill_depth_mm, color = species)) +
  geom_point(size = 3) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Odnos između dužine i dubine kljuna kod pingvina",
       x = "Dužina kljuna (mm)", y = "Dubina kljuna (mm)") +
  scale_color_manual(values = c("Chinstrap" = "darkred")) +
  theme_minimal()
```

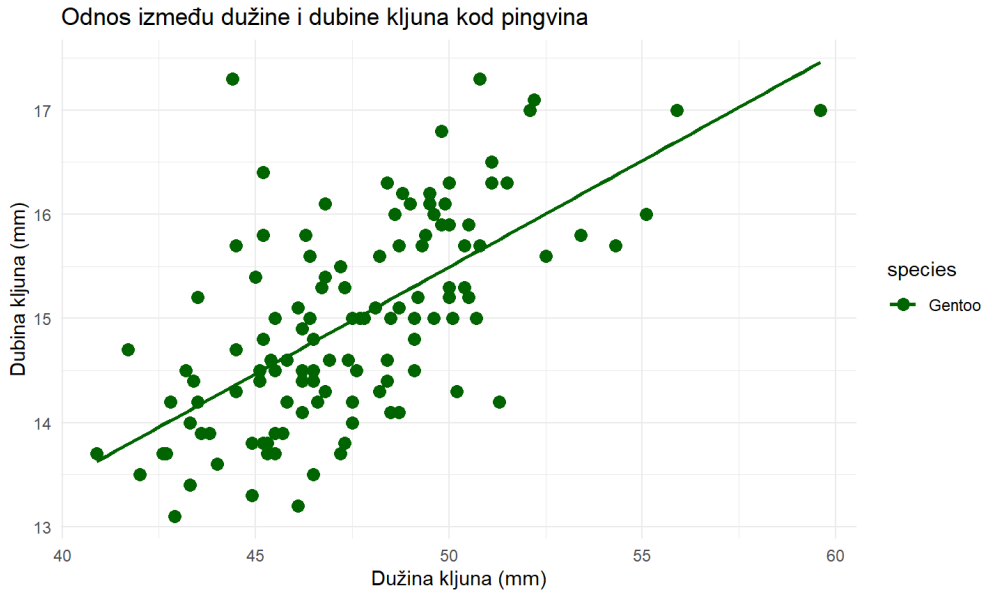
```
## `geom_smooth()` using formula = 'y ~ x'
```



Rješenje

```
# gentoo graf
ggplot(data = gentoo, aes(x = bill_length_mm, y = bill_depth_mm, color = species)) +
  geom_point(size = 3) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Odnos između dužine i dubine kljuna kod pingvina",
       x = "Dužina kljuna (mm)", y = "Dubina kljuna (mm)") +
  scale_color_manual(values = c("Gentoo" = "darkgreen")) +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Različiti načini grafičkih prikaza - base R vs. ggplot2

Base R plotovi

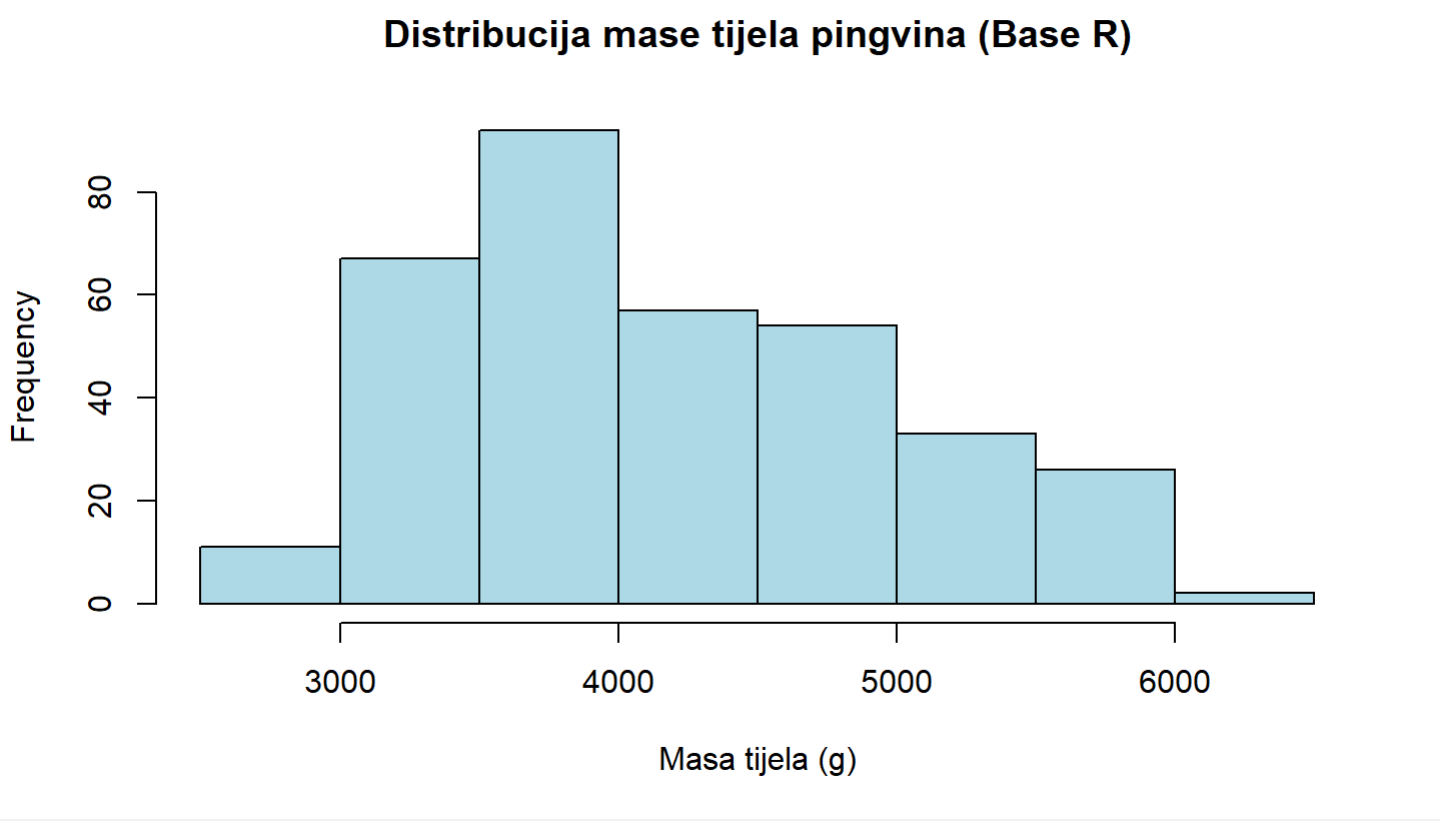
- nije potrebna instalacija ni učitavanje dodatnih paketa
- jednostavne funkcije
- najčešće korišteno za brzinsku provjeru izgleda podataka i rezultata statistike
- slabija i kompliciranija mogućnost prilagodbe grafa

ggplot2 grafovi

- potrebno instalirati i učitati paket ggplot2
- vrlo fleksibilan za prilagodbu
- dio tidyverse skupa paketa koji su prilagođeni za manipulaciju podacima
- mnogi statistički paketi kompatibilni s ggplot2
- najčešće korišten za izradu finalnog grafa za publikacije

Primjer 1: Histogram

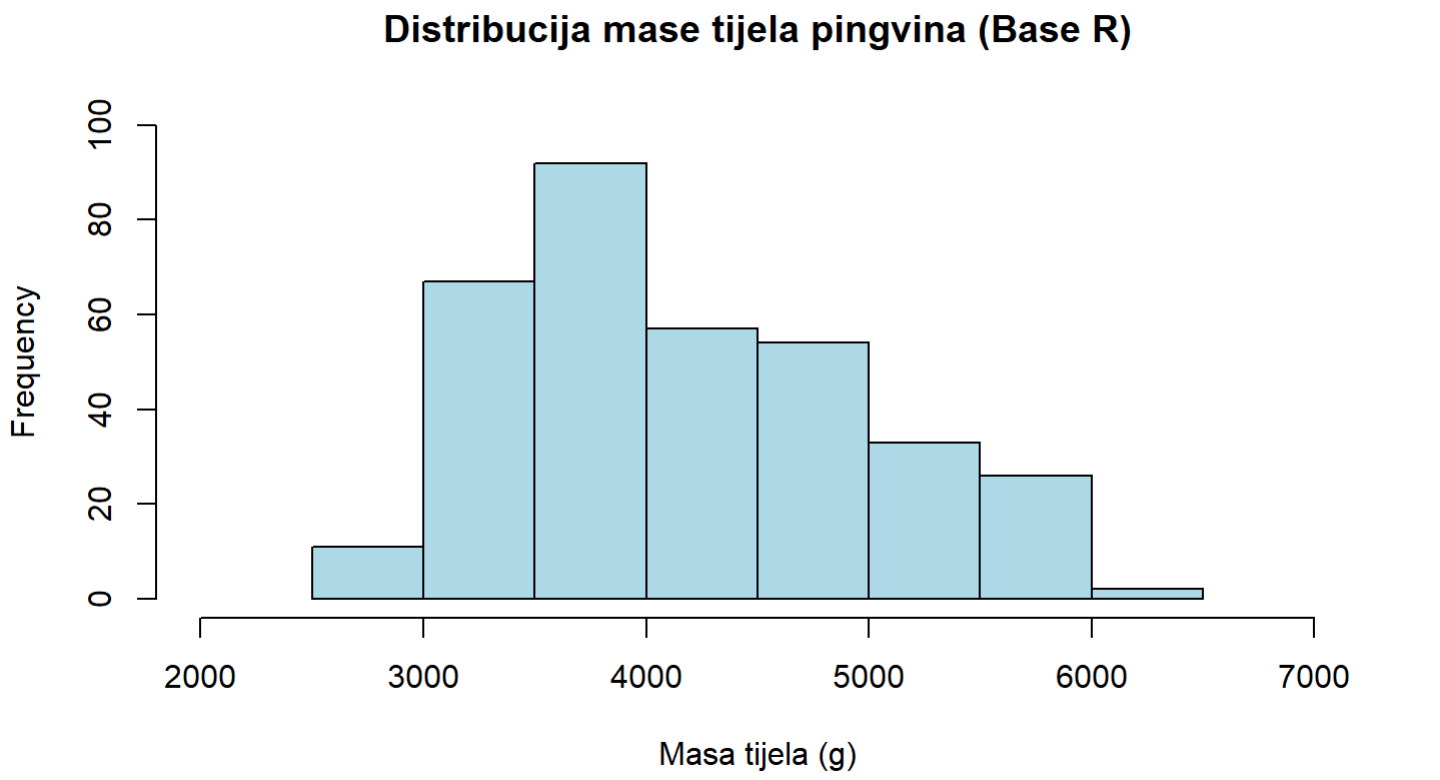
```
# Base R histogram
hist(penguins$body_mass_g,
     main = "Distribucija mase tijela pingvina (Base R)",
     xlab = "Masa tijela (g)",
     col = "lightblue",
     border = "black")
```



Koju grešku uočavate na prikazu ovog grafa?

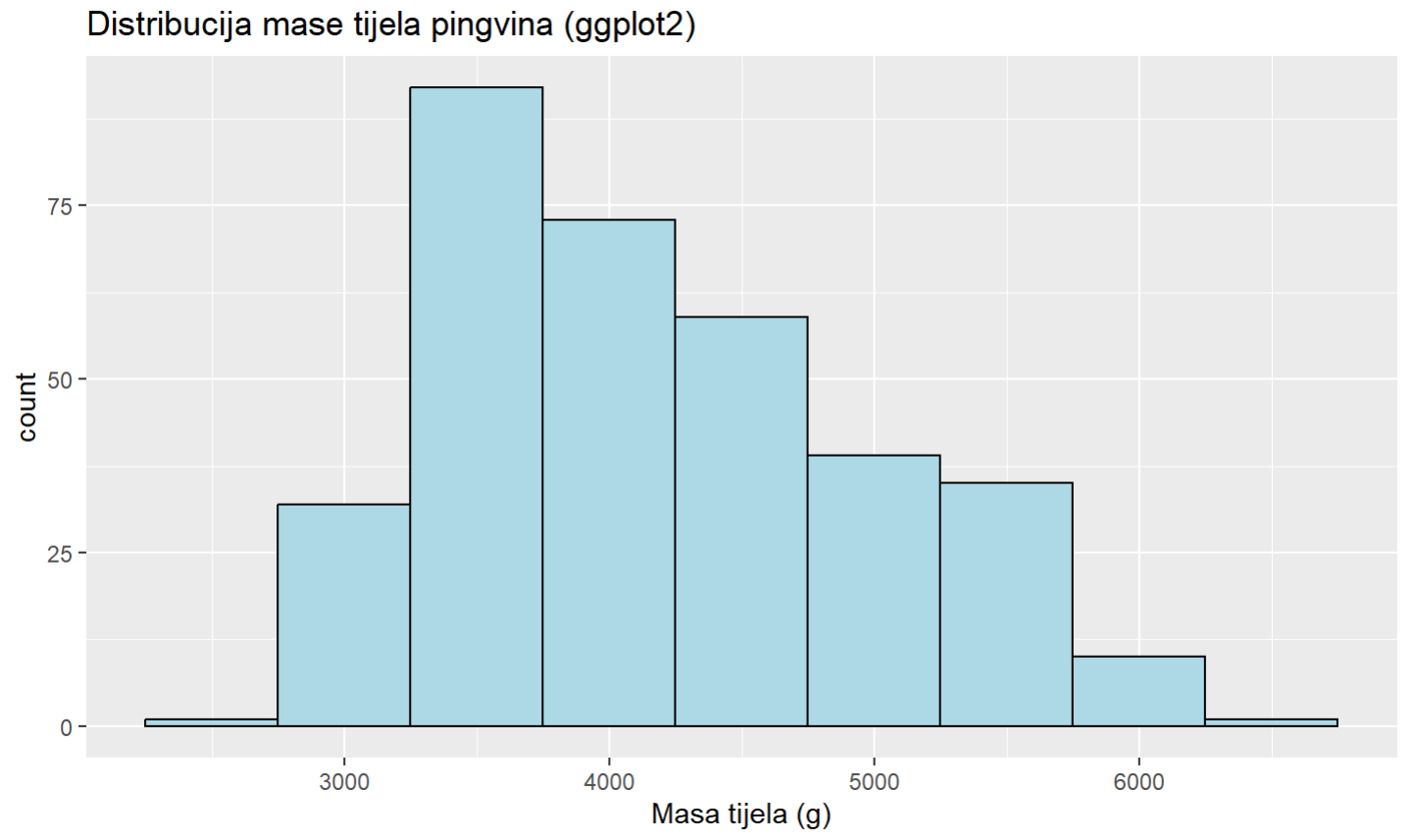
Probajte na internetu naći rješenje kako to ispraviti!

```r
# Proširivanje vrijednosti x i y osi
hist(penguins$body_mass_g,
     main = "Distribucija mase tijela pingvina (Base R)",
     xlab = "Masa tijela (g)",
     col = "lightblue",
     border = "black",
     xlim = c(2000, 7000), ylim = c(0, 100))
```



Izrada histograma pomoću ggplota

```
# ggplot2 histogram
ggplot(penguins, aes(x = body_mass_g)) +
  geom_histogram(binwidth = 500, fill = "lightblue", color = "black") +
  labs(title = "Distribucija mase tijela pingvina (ggplot2)",
       x = "Masa tijela (g)")
```

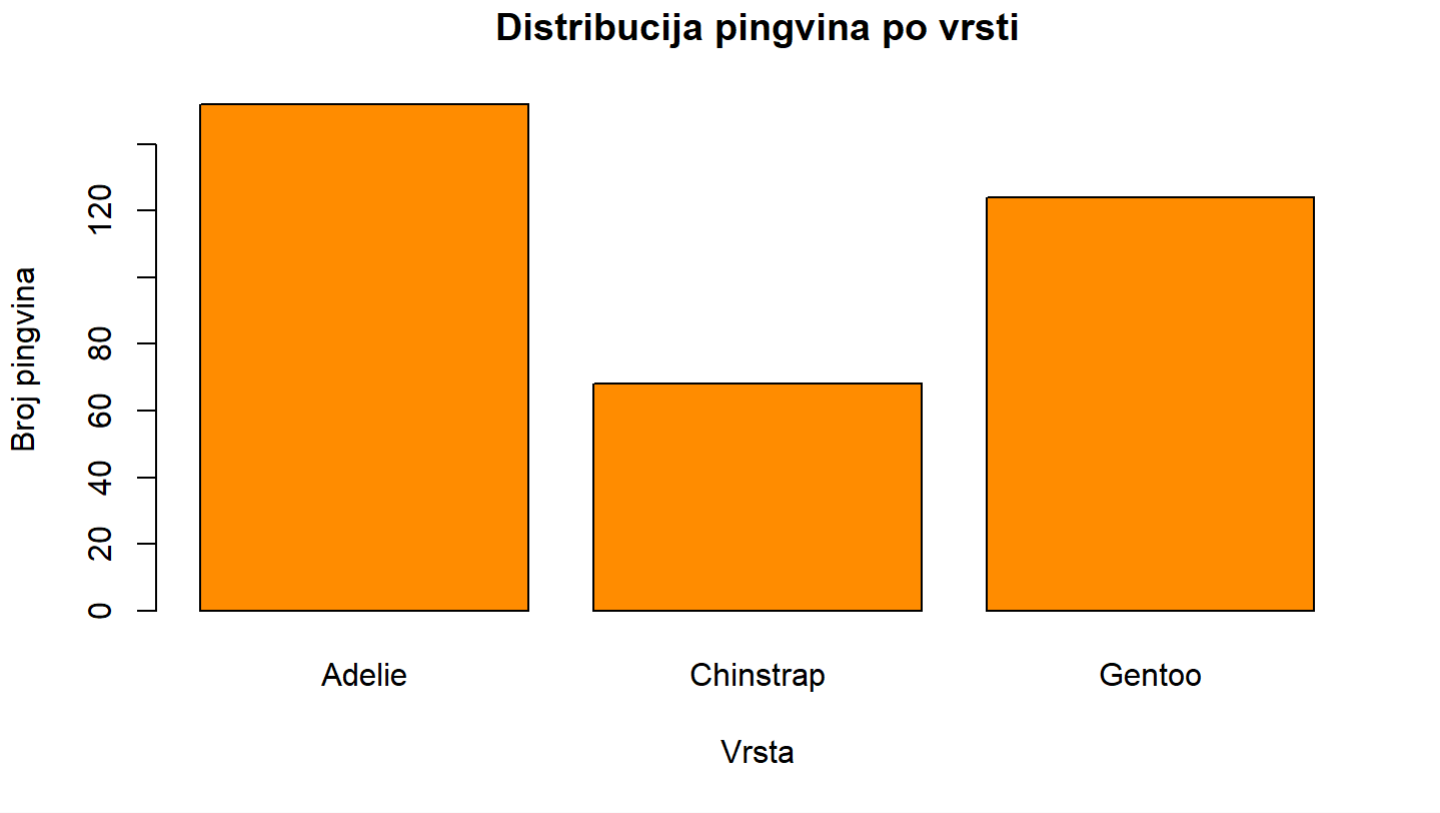


Primjer 2: Stupičasti dijagram (*bar plot*)

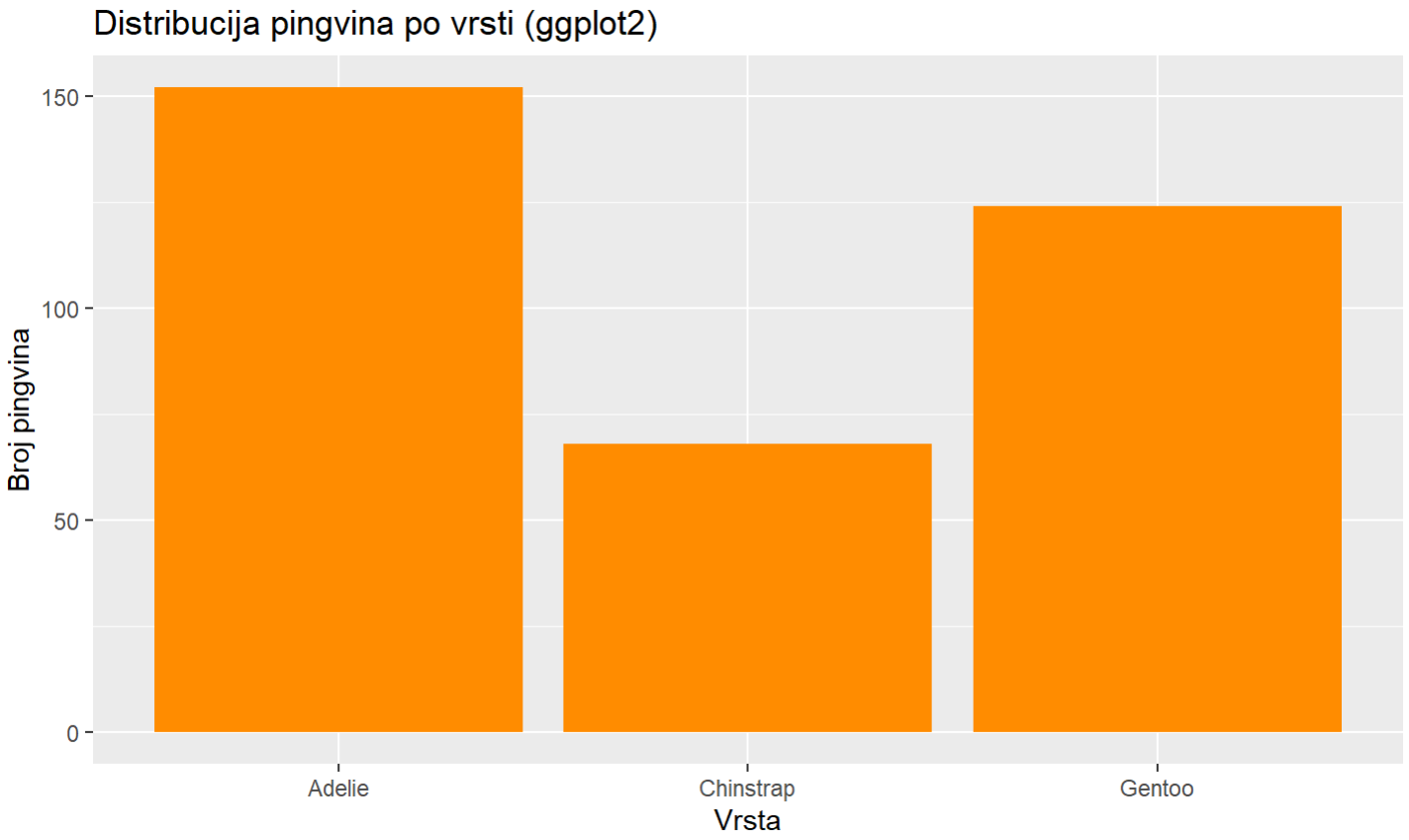
```
# Primjer 2: Stupičasti dijagram (bar plot)

# Prvo: Kreiranje tablice za broj pingvina po vrsti
species_count <- table(penguins$species)
print(species_count)
```

```
##
##    Adelie Chinstrap    Gentoo
##       152        68       124
```

```r
# Drugo: izrada stupičastog dijagrama
# Base R barplot
barplot(species_count,
        main = "Distribucija pingvina po vrsti",
        xlab = "Vrsta",
        ylab = "Broj pingvina",
        col = "darkorange")
```

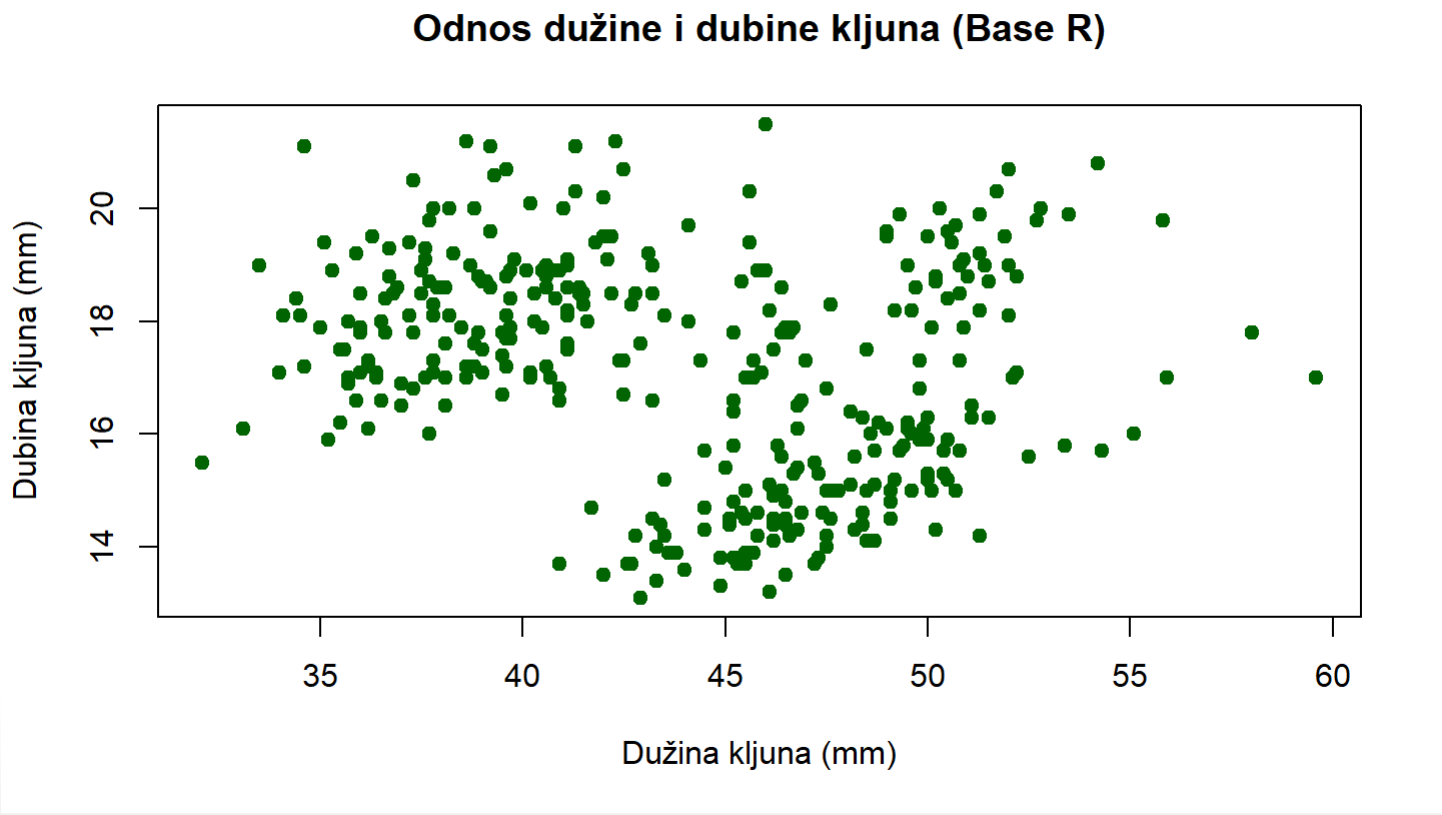


```r
# ggplot2 barplot
ggplot(penguins, aes(x = species)) +
  geom_bar(fill = "darkorange") +
  labs(title = "Distribucija pingvina po vrsti (ggplot2)",
       x = "Vrsta",
       y = "Broj pingvina")
```

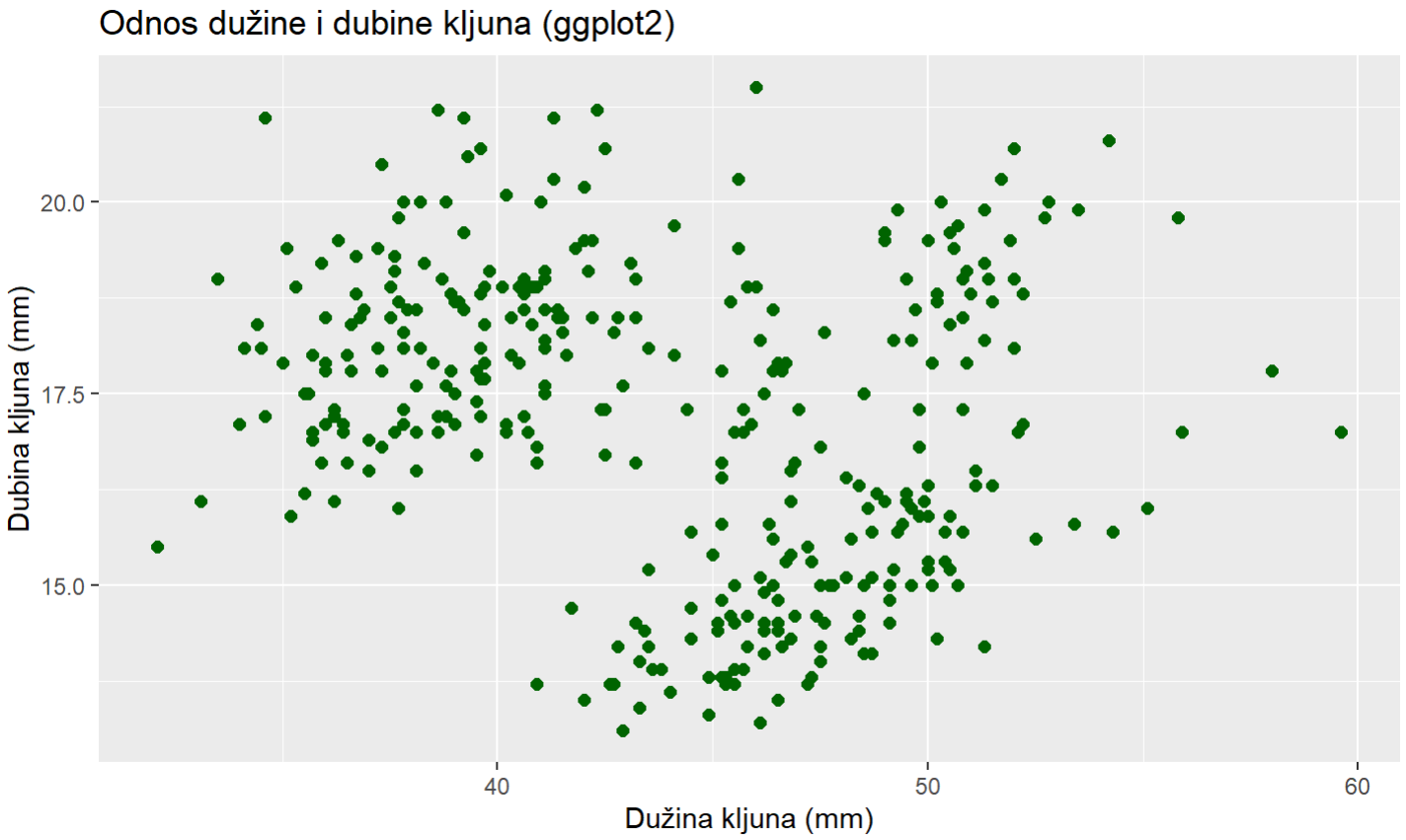


Primjer 3: Točkasti dijagram (*scatter plot*)

```r
# Base R scatter plot
plot(penguins$bill_length_mm, penguins$bill_depth_mm,
     main = "Odnos dužine i dubine kljuna (Base R)",
     xlab = "Dužina kljuna (mm)",
     ylab = "Dubina kljuna (mm)",
     col = "darkgreen", pch = 19)
```



```r
# ggplot2 scatter plot
ggplot(penguins, aes(x = bill_length_mm, y = bill_depth_mm)) +
  geom_point(color = "darkgreen", size = 2) +
  labs(title = "Odnos dužine i dubine kljuna (ggplot2)",
       x = "Dužina kljuna (mm)",
       y = "Dubina kljuna (mm)")
```

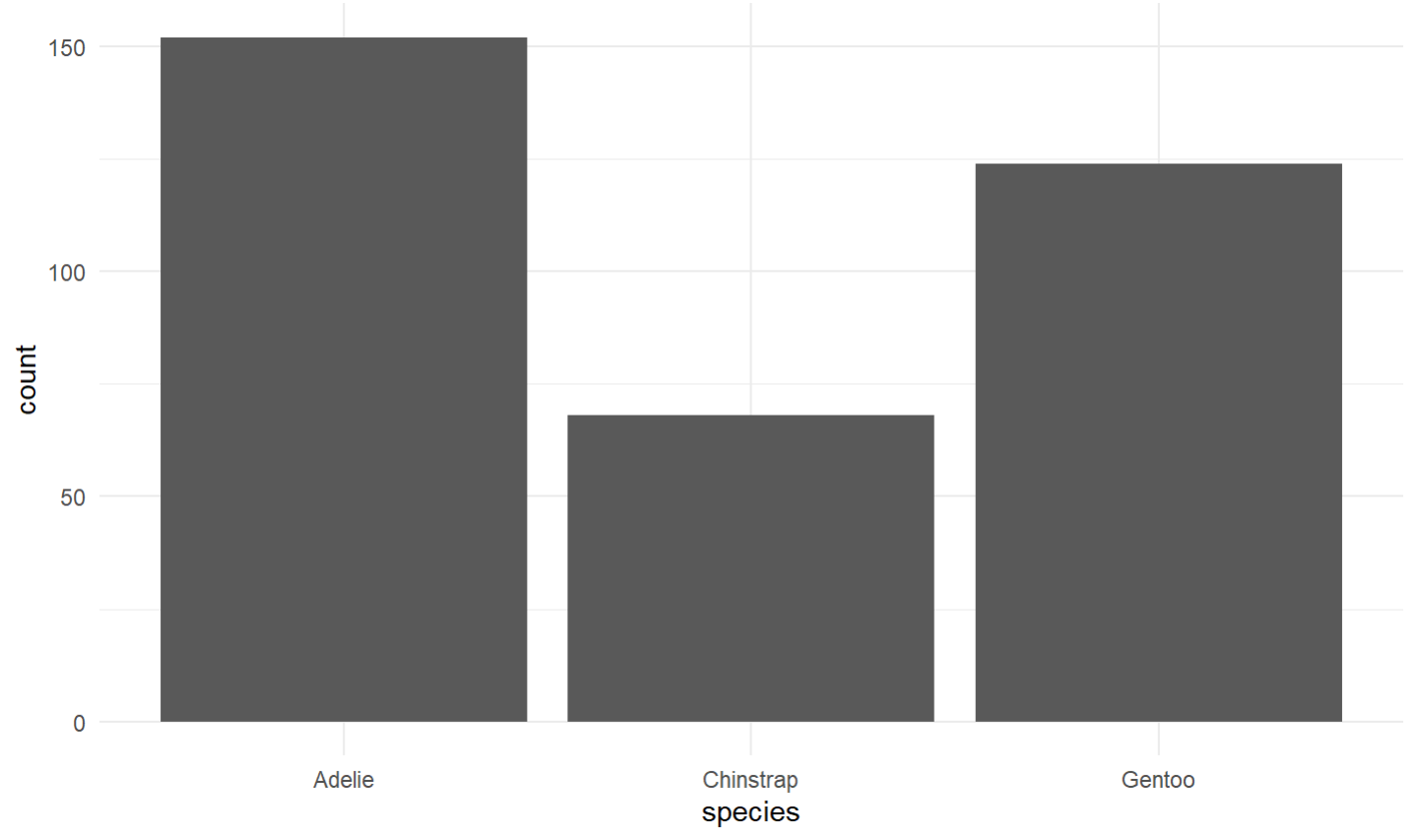


Grafički prikazi po tipu varijabli

1. kategoričke varijable
2. numeričke varijable
3. dvije kategoričke
4. dvije numeričke
5. odnos numeričke i kategoričke

1. Grafički prikazi kategoričkih varijabli

```
# 1.1 Barplot za prikaz distribucije pingvina po vrstama.
ggplot(penguins, aes(x = species)) +
  geom_bar() + theme_minimal()
```



Kako prikazati kategorije vrste od najbrojnije do najmanje brojne?

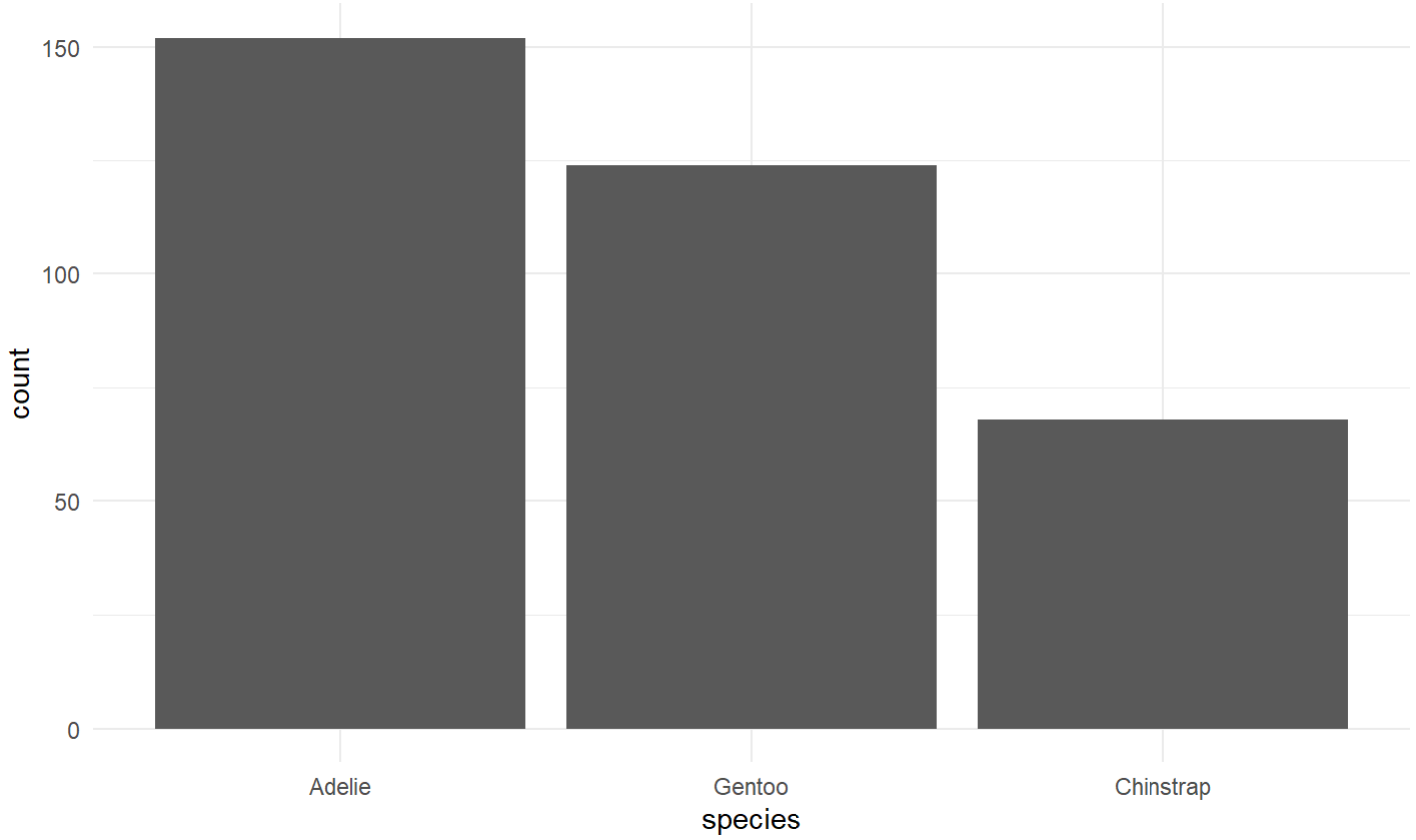
Moramo pretvoriti kategoričku varijablu “species” u faktor!

- Faktori su poseban način zapisivanja kategoričkih varijbli u R-u.
- Funkcioniraju tako da vrijednosti zapišu u nivoe tj. levele i tako olakšavaju prikaz vrijednosti kategoričke varijable.

U ovom primjeru koristimo funkciju **`factor()`** kako bi poredali varijablu “species” po silaznom (*decreasing*) poretku.

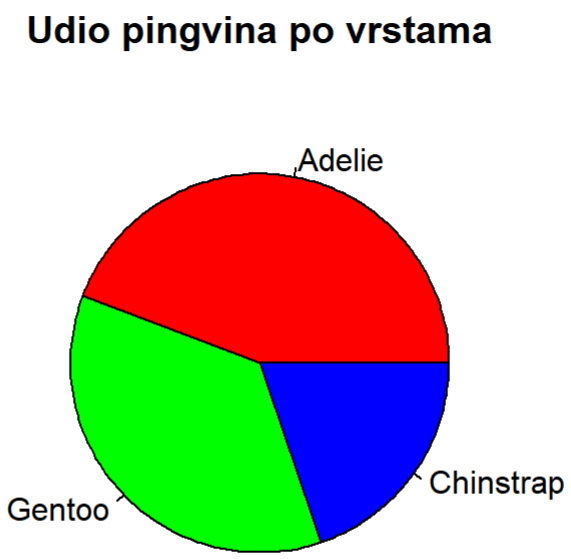
```
penguins$species <- factor(penguins$species,
                           levels = names(sort(table(penguins$species), decreasing = TRUE)))
```

```r
# Sad ponovo nacrtajte barplot!
ggplot(penguins, aes(x = species)) +
  geom_bar() + theme_minimal()
```



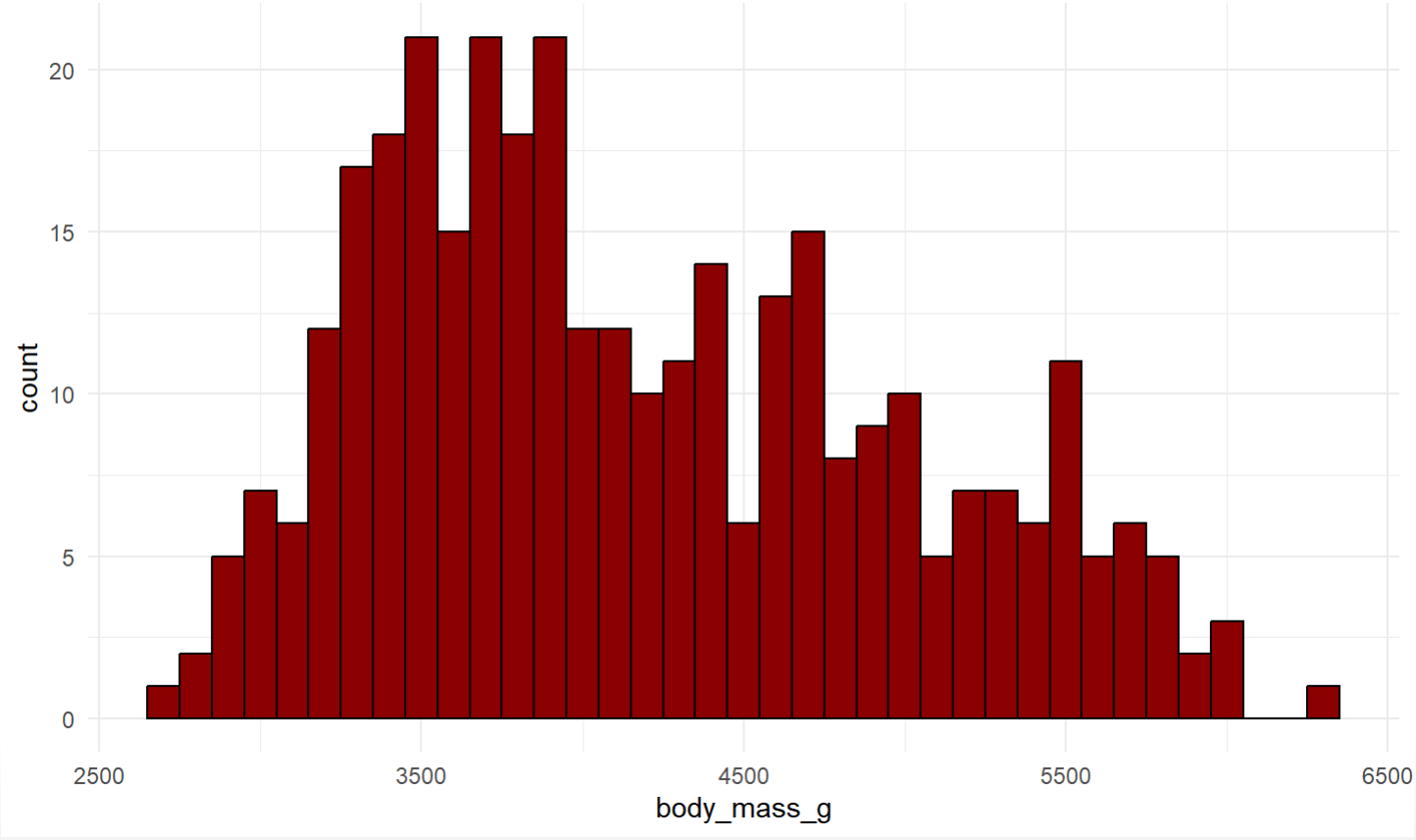
1.2 *Pie chart* za prikaz udjela pingvina po vrstama

```
species_count <- penguins %>% count(species)

pie(species_count$n, labels = species_count$species,
    main = "Udio pingvina po vrstama",
    col = rainbow(length(species_count$species)))
```



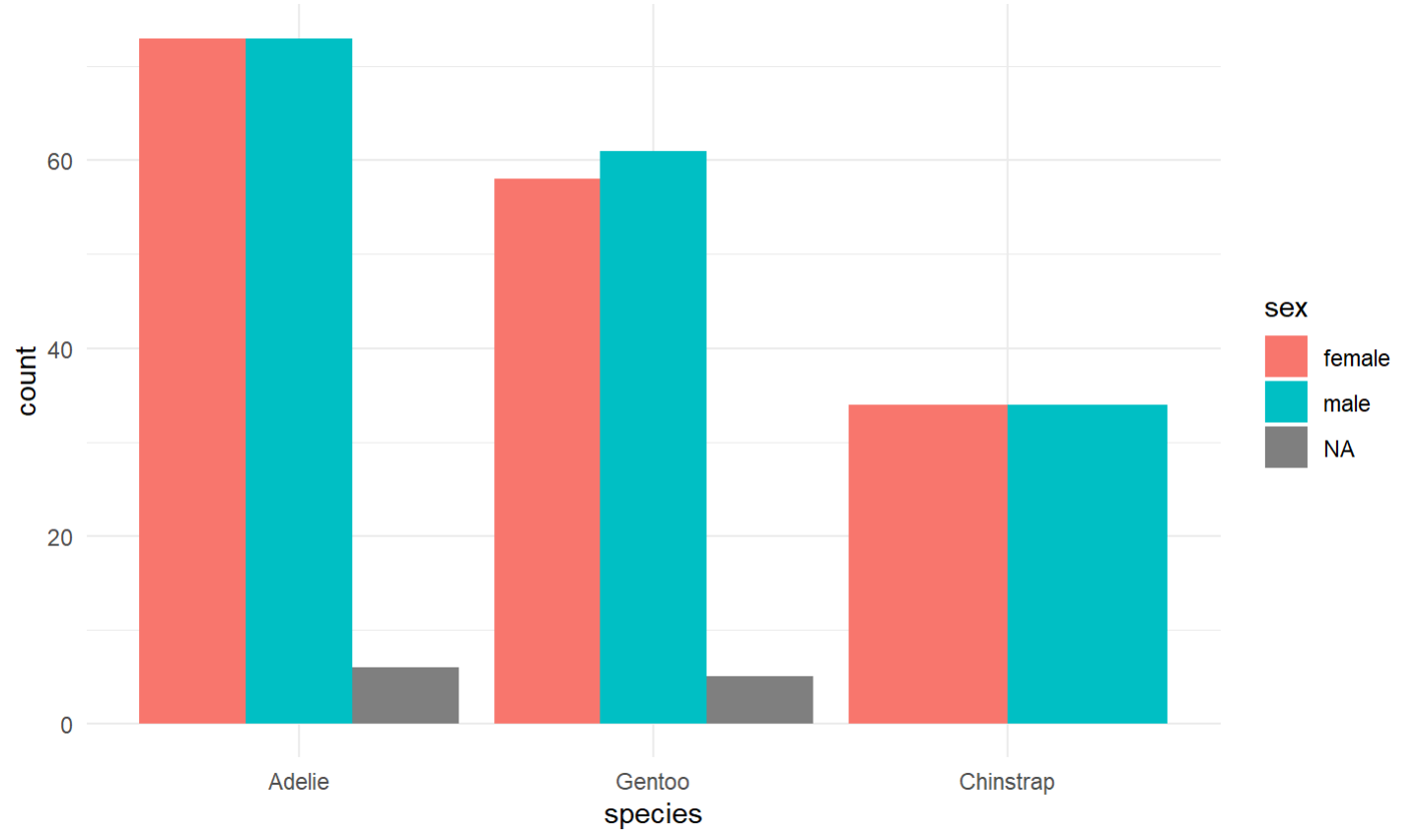
2. Grafički prikazi numeričkih varijabli

```
# 2.1 Histogram za prikaz distribucije mase tijela pingvina.
ggplot(penguins, aes(x = body_mass_g)) +
  geom_histogram(binwidth = 100, color = "black", fill = "darkred") +
  theme_minimal()
```

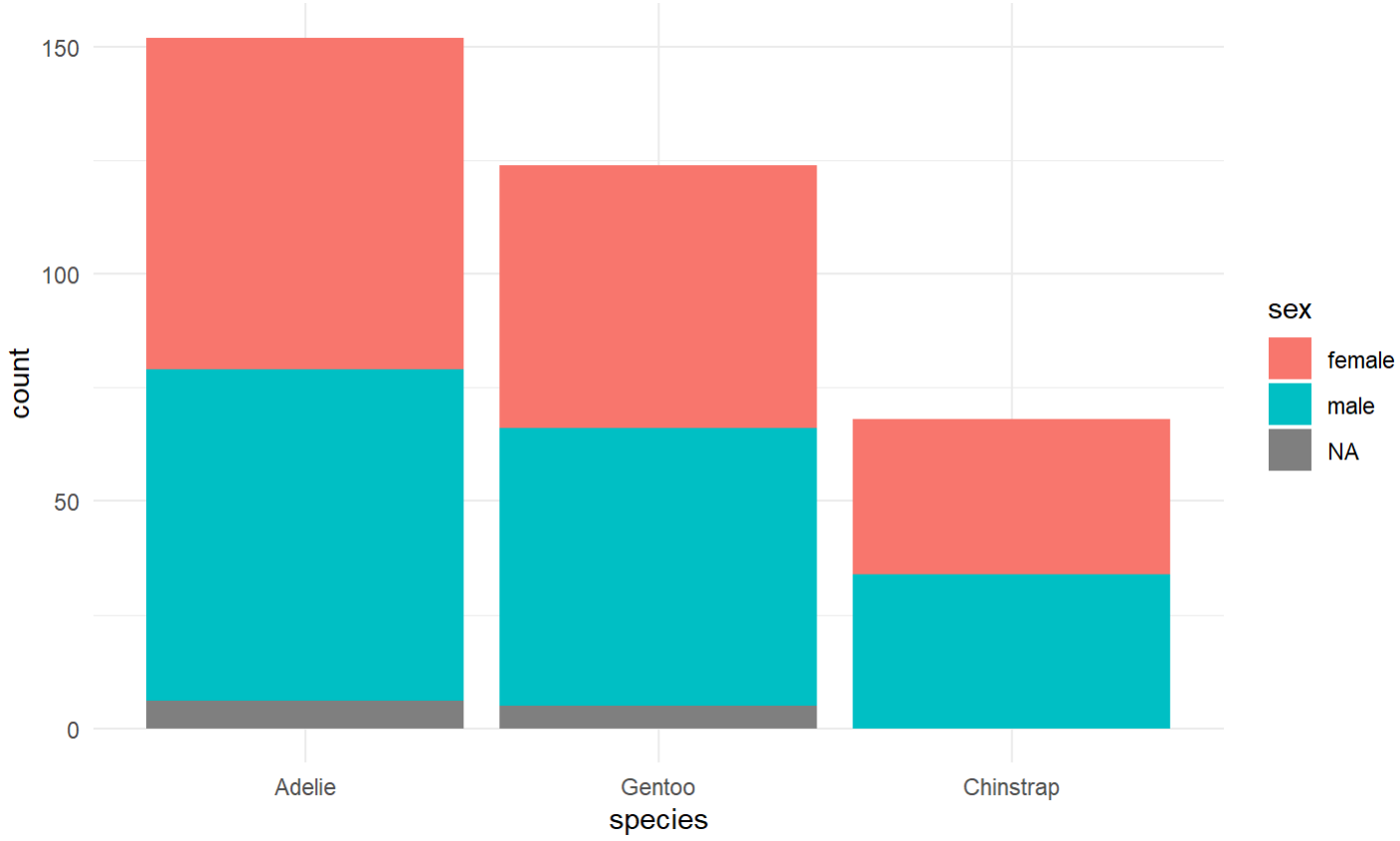


3. Prikaz odnosa dvije kategoričke varijable

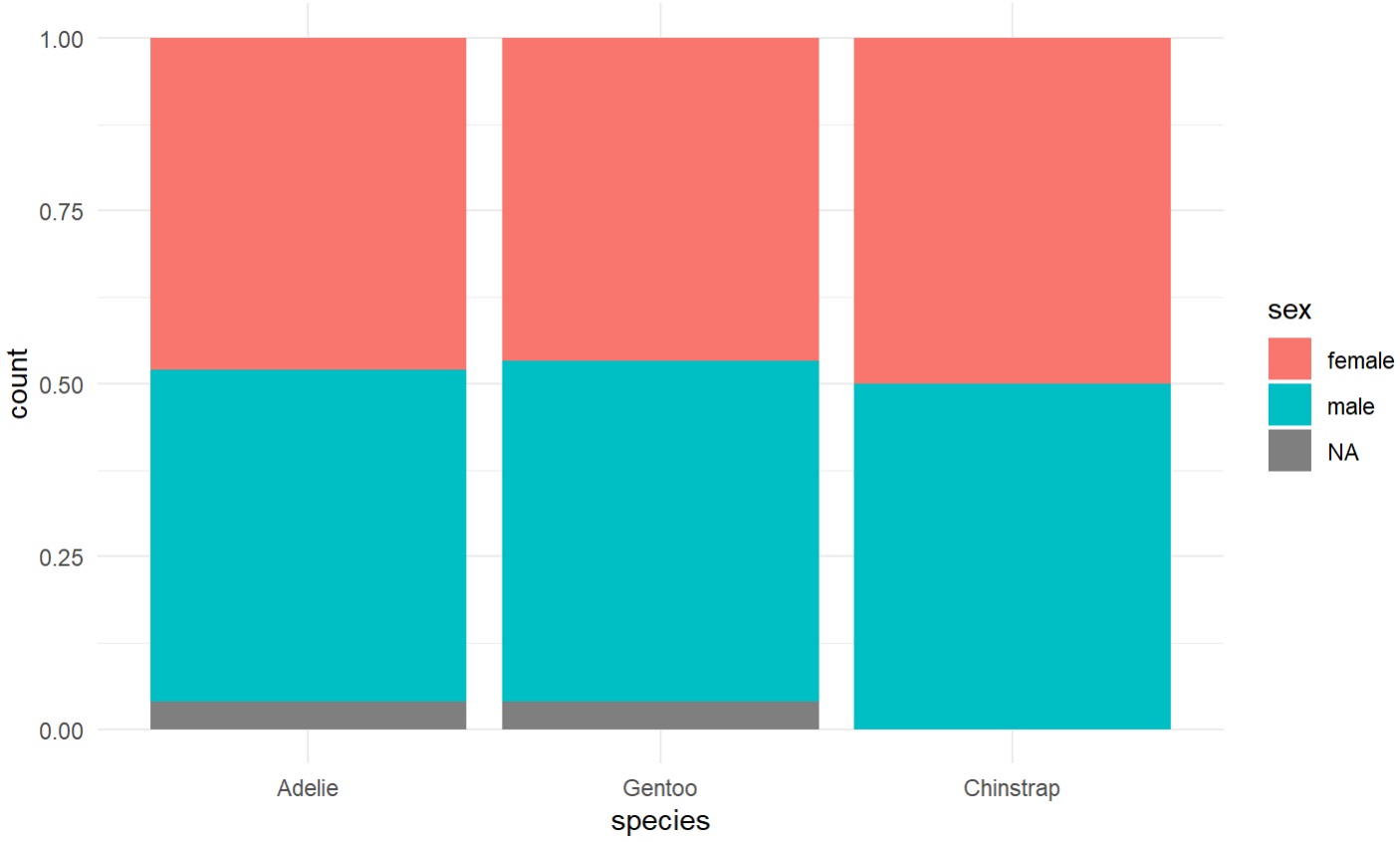
```
# 3.1 Bar plot
ggplot(data = penguins, aes(x = species, fill = sex)) +
  geom_bar(position = "dodge") +
  theme_minimal()
```



```r
# 3.2 Stacked bar plot
ggplot(data = penguins, aes(x = species, fill = sex)) +
  geom_bar() +
  theme_minimal()
```

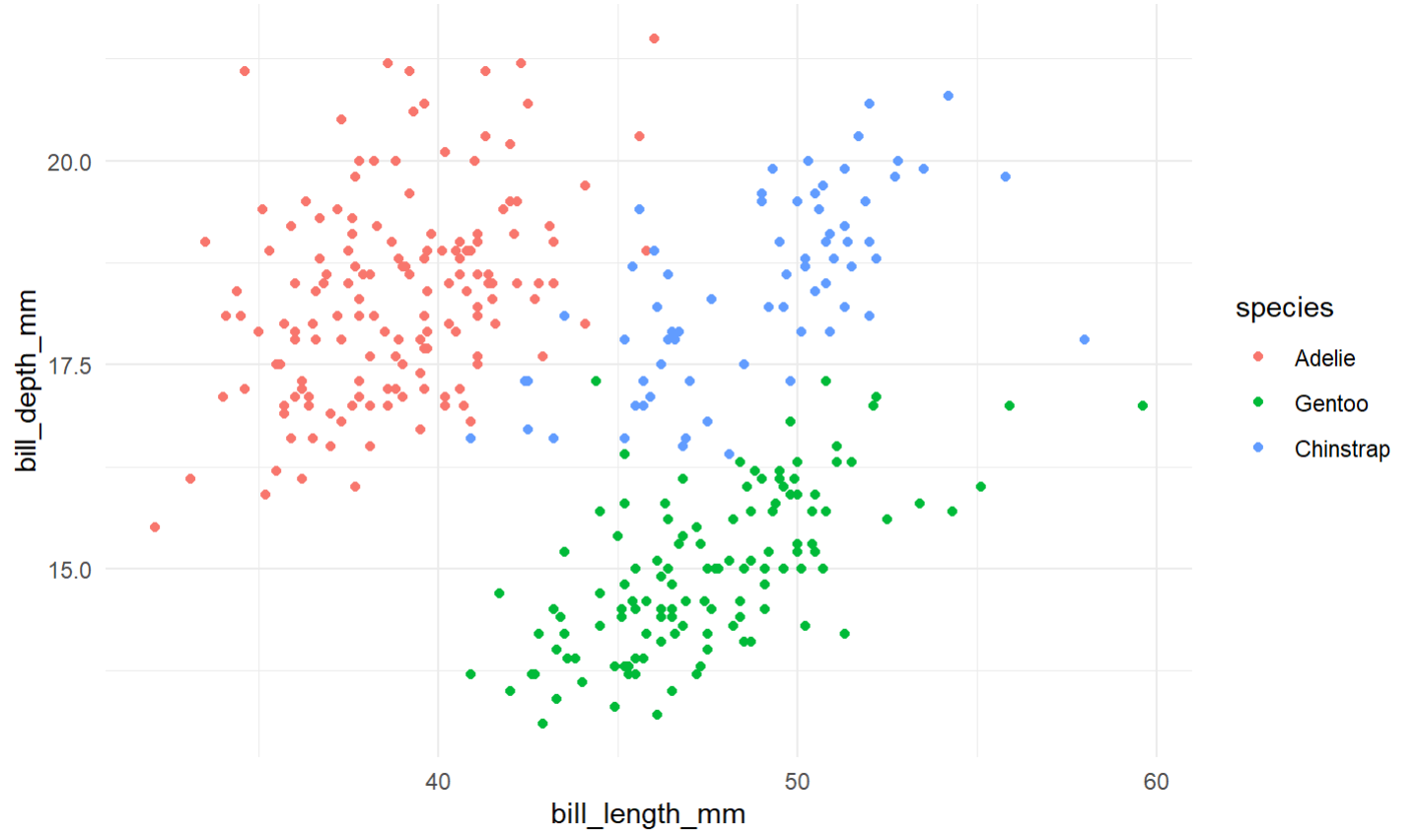


```r
# Relativni odnos
ggplot(data = penguins, aes(x = species, fill = sex)) +
  geom_bar(position = "fill") +
  theme_minimal()
```



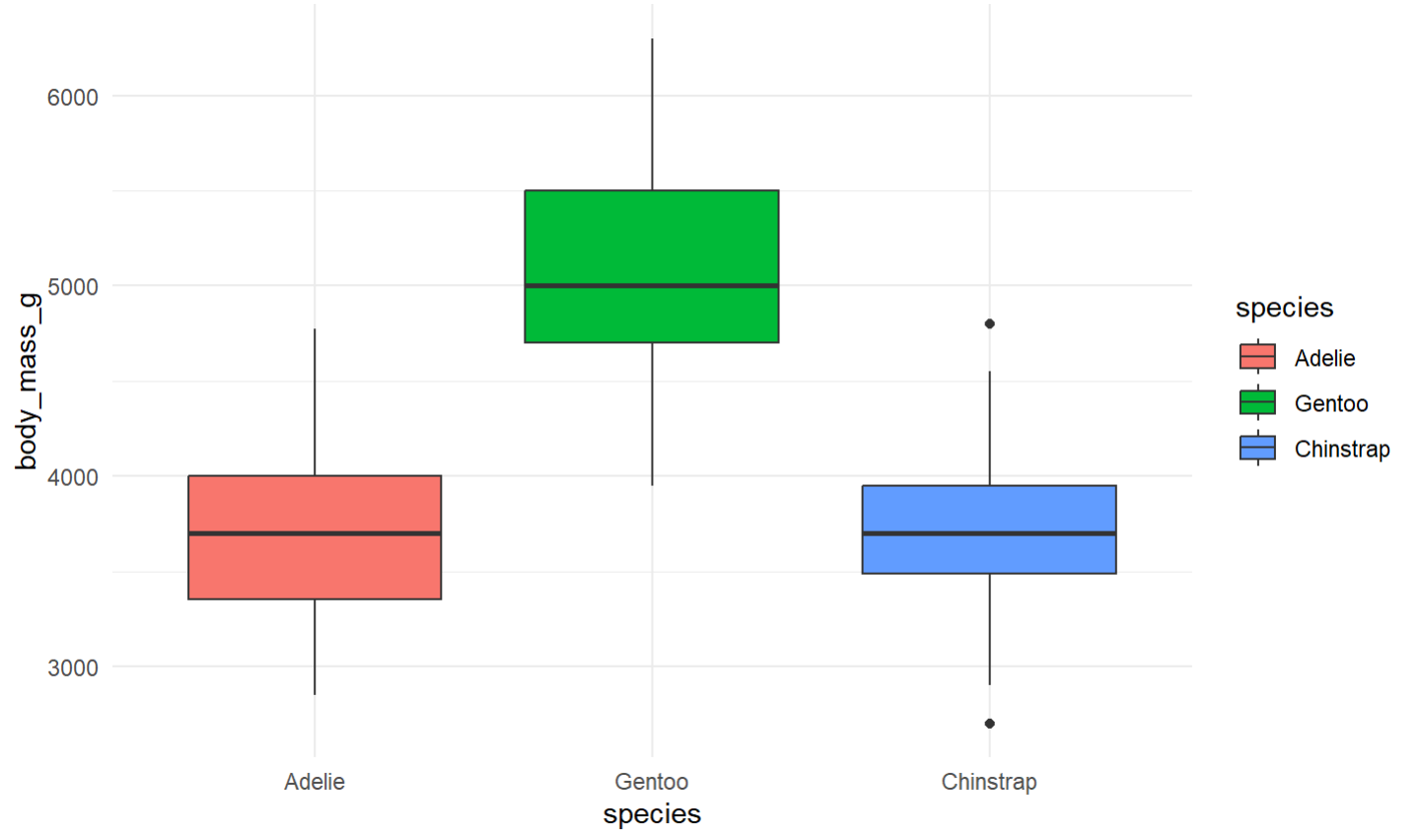
4. Prikaz odnosa dvije numeričke varijable

```
# 4.1 Točkaskti graf (scatter plot)
ggplot(data = penguins, aes(x = bill_length_mm, y = bill_depth_mm, color = species)) +
  geom_point() +
  theme_minimal()
```

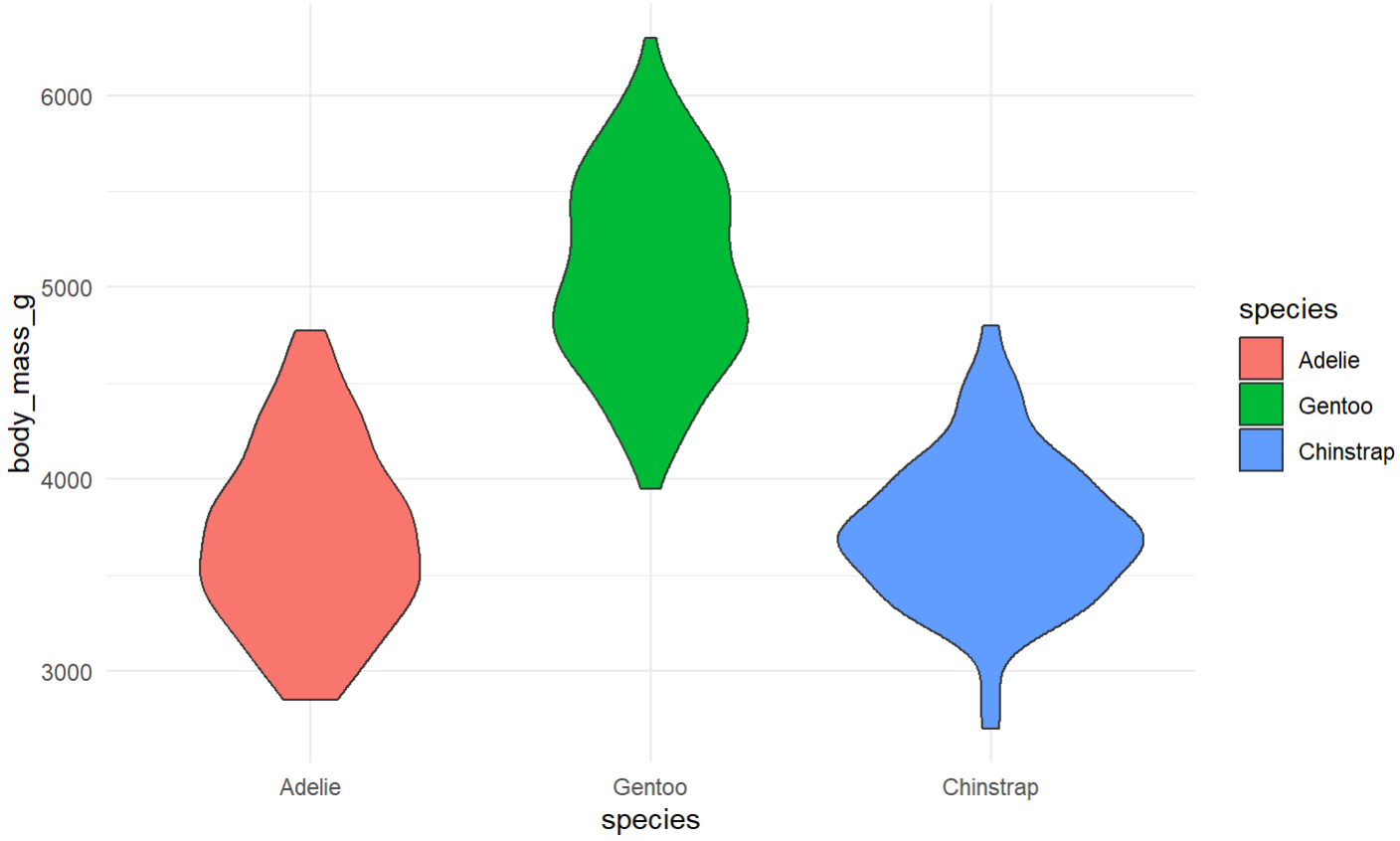


5. Prikaz odnosa numeričke i kategoričke varijable

```
# 5.1 Box plot
ggplot(data = penguins, aes(x = species, y = body_mass_g, fill = species)) +
  geom_boxplot() +
 theme_minimal()
```

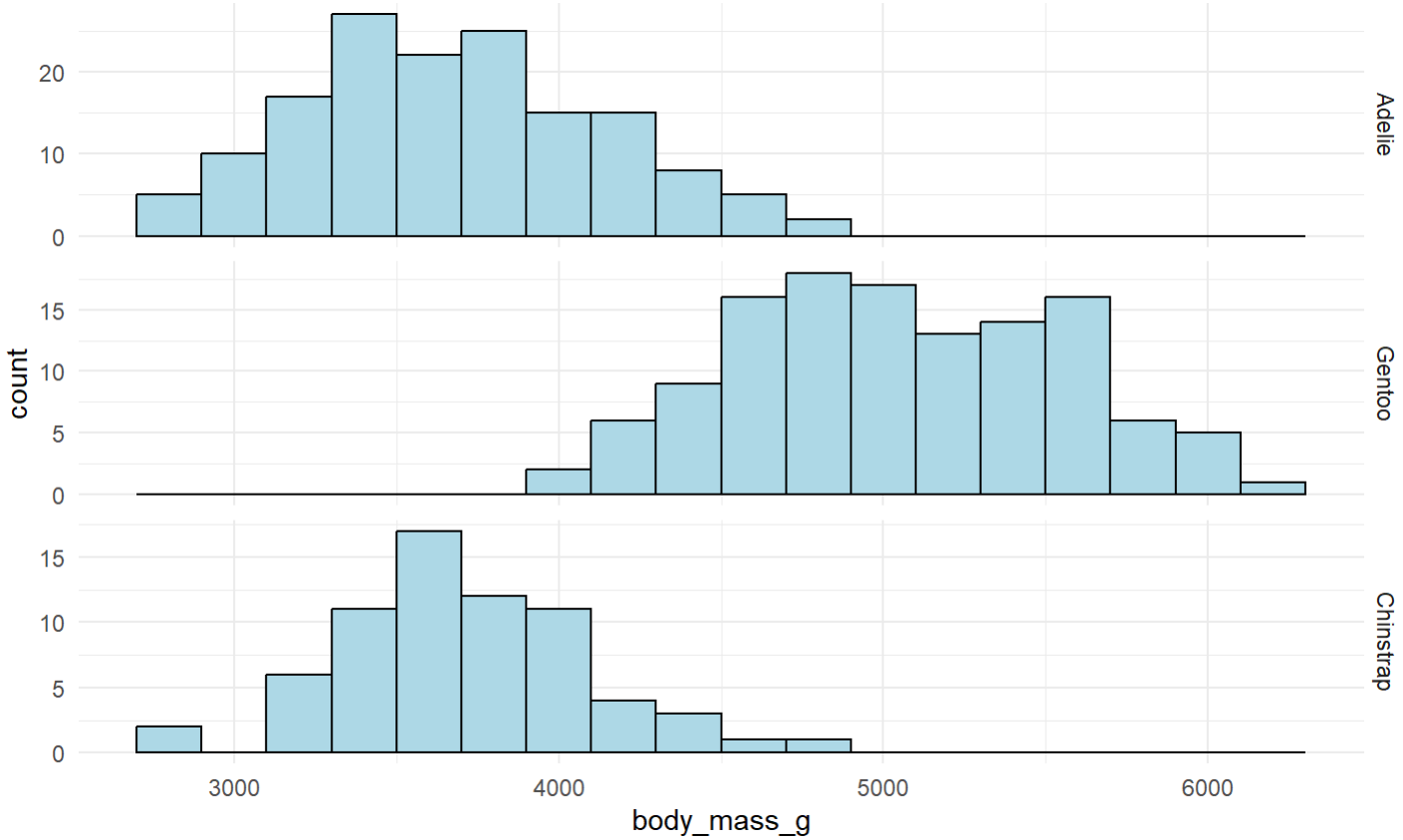


```r
# 5.2 Violin plot
ggplot(data = penguins, aes(x = species, y = body_mass_g, fill = species)) +
  geom_violin() +
  theme_minimal()
```



```r
# 5.3 Strip chart
ggplot(data = penguins, aes(x = species, y = body_mass_g, color = species)) +
  geom_jitter() +
  theme_minimal()
```

```r
# 5.4 Višestruki histogrami
ggplot(data = penguins, aes(x = body_mass_g)) +
  geom_histogram(binwidth = 200, color = "black", fill = "lightblue") +
  facet_wrap( ~ species, ncol = 1, scales = "free_y", strip.position = "right") +
  theme_minimal()
```



Zadatak

1. Učitajte tablicu proširenog dataseta s pingvinima: palmerpengiuns_extended.
2. Napravite grafove za:
    1. jednu numeričku varijablu po izboru,
    2. jednu kategoričku varijablu po izboru,
    3. odnos dvije numeričke varijable po izboru,
    4. odnos dvije kategoričke varijable po izboru,
    5. odnos kategoričke i numeričke varijable po izboru.

Poigrajte se s temama i bojama kako bi dobili graf koji najbolje prikazuje informaciju koju želite prikazati!