Data exploration using the palmerpenguins dataset

Gbadamassi G.O. Dossa (dossa@xtbg.org.cn)

Updated on 2023-11-13 (created on 2021-09-13)

Acknowledgements

The content of this module are based on materials from: olivier gimenez's materials (https://oliviergimenez.github.io/)

Data exploration

Motivation

In this section, we **explore** the data from package `palmerpenguins`. A recent publication from the researcher, Dr Kristen Gorman, who shared the data is Connors et al. (2020). We will also use the package "citr" for refering to scientic citation as well. Howeve, please rember that "citr" is not hosted by CRAN (https://cran.r-project.org/) but rather on GitHub citr (https://github.com/crsh/citr). You can install it by doing `devtools::install_github("crsh/citr")`.

Data

The data are displayed below (first 10 row) :

```
penguins %>%
  slice(1:10) %>% # Slice as the verb say is to cut a tibble in piece, another of sub-setting
  knitr::kable() # Remember kable is the function that helps you to display table. It is buil
t in in knitr
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007
Adelie	Torgersen	NA	NA	NA	NA	NA	2007
Adelie	Torgersen	36.7	19.3	193	3450	female	2007
Adelie	Torgersen	39.3	20.6	190	3650	male	2007
Adelie	Torgersen	38.9	17.8	181	3625	female	2007
Adelie	Torgersen	39.2	19.6	195	4675	male	2007
Adelie	Torgersen	34.1	18.1	193	3475	NA	2007
Adelie	Torgersen	42.0	20.2	190	4250	NA	2007

Numerical exploration

Here we will make use of *inline code*. There are 344 penguins in the dataset, and 3 different species. The data were collected in 3 islands of the Palmer archipelago in Antarctica.

The mean of all traits that were measured on the penguins are:

```
## # A tibble: 3 × 6
##   species   bill_length_mm bill_depth_mm flipper_length_mm body_mass_g  year
##   <fct>              <dbl>         <dbl>             <dbl>       <dbl> <dbl>
## 1 Adelie              38.8          18.3              190.       3701. 2008.
## 2 Chinstrap           48.8          18.4              196.       3733. 2008.
## 3 Gentoo              47.5          15.0              217.       5076. 2008.
```

Inline code is powerful because, when the data set changes or the number of observations changes, this will automatically be detected and updated. You remember there were missing values for some individuals for the sex? We can now change change the inline code to only render the individual with know sex.

```
dat <- penguins %>%
# filter out missing sex
  filter(!is.na(sex)) %>%
# select variables
  select(species, island, bill_length_mm, body_mass_g)

glimpse(dat)
```

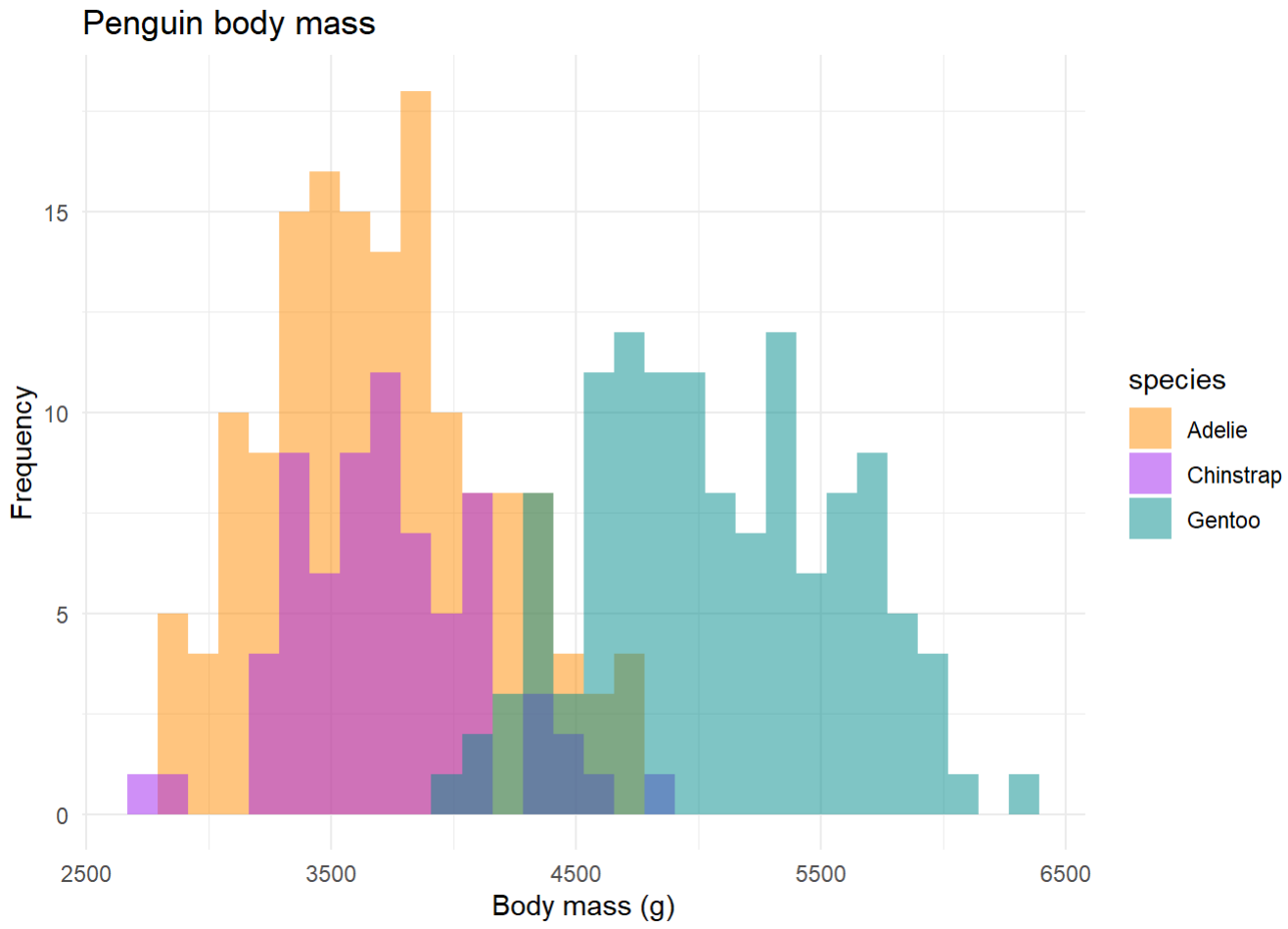
```
## Rows: 333
## Columns: 4
## $ species        <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adelie,…
## $ island         <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgersen, …
## $ bill_length_mm <dbl> 39.1, 39.5, 40.3, 36.7, 39.3, 38.9, 39.2, 41.1, 38.6, 3…
## $ body_mass_g    <int> 3750, 3800, 3250, 3450, 3650, 3625, 4675, 3200, 3800, 4…
```

Here we will make use of *inline code* but on the dataset dat which contains no missing vavlues. There are 333 penguins in the dataset, and 3 different species. The data were collected in 3 islands of the Palmer archipelago in Antarctica.

Graphical exploration

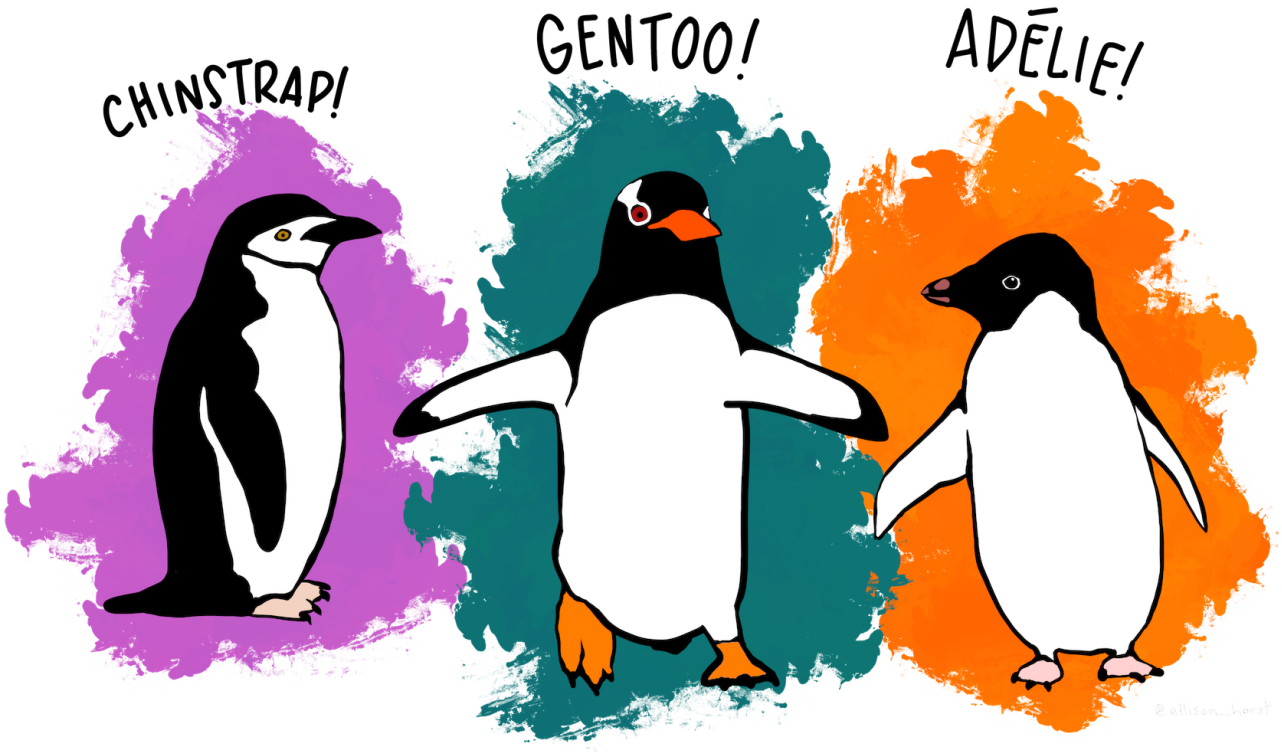
A histogram of body mass per species:

```
penguins %>%
  ggplot() +
  aes(x = body_mass_g) +
  geom_histogram(aes(fill = species),
                 alpha = 0.5,
                 position = "identity") +
  scale_fill_manual(values = c("darkorange","purple","cyan4")) +
  theme_minimal() +
  labs(x = "Body mass (g)",
       y = "Frequency",
       title = "Penguin body mass")
```



The end

The 3 species of penguins:



References

Connors, B., M. J. Malick, G. T. Ruggerone, P. Rand, M. Adkison, J. R. Irvine, R. Campbell, et al. 2020. Climate and competition influence sockeye salmon population dynamics across the Northeast Pacific Ocean (https://doi.org/10.1139/cjfas-2019-0422). Canadian Journal of Fisheries and Aquatic Sciences 77:943–949.