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) @(Satchell and Dourish 2009).

## Warning: package 'tidyverse' was built under R version 4.2.2

## Warning: package 'ggplot2' was built under R version 4.2.2

## Warning: package 'tidyr' was built under R version 4.2.2

## Warning: package 'readr' was built under R version 4.2.2

## Warning: package 'purrr' was built under R version 4.2.2

## Warning: package 'dplyr' was built under R version 4.2.2

## Warning: package 'stringr' was built under R version 4.2.2

## Warning: package 'forcats' was built under R version 4.2.2

## Warning: package 'lubridate' was built under R version 4.2.2

## Data

The data are displayed below (first 10 rows) :

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

## Warning: There was 1 warning in `summarize()`.  
## ℹ In argument: `across(where(is.numeric), mean, na.rm = TRUE)`.  
## ℹ In group 1: `species = Adelie`.  
## Caused by warning:  
## ! The `...` argument of `across()` is deprecated as of dplyr 1.1.0.  
## Supply arguments directly to `.fns` through an anonymous function instead.  
##   
## # Previously  
## across(a:b, mean, na.rm = TRUE)  
##   
## # Now  
## across(a:b, \(x) mean(x, na.rm = TRUE))

## # 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.

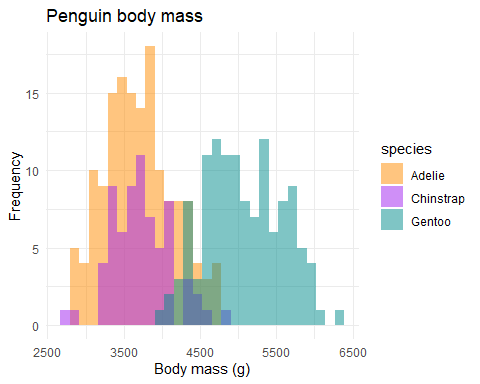
Inline code is powerful because, when the dataset chages or the number of obsevation changes, this will automatically be detected and updated. You remember there were missing values for some individuals for the sex? We can now change …

## 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")

## Warning: Removed 2 rows containing non-finite values (`stat\_bin()`).



## 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.

Satchell, C., and P. Dourish. 2009. [Beyond The User: Use And Non-Use in HCI](https://doi.org/10.1145/1738826.1738829). Proceedings of the Annual Conference of the Australian Computer-Human Interaction Special Interest Group (OZCHI ’09) 9–16.