Data exploration using the palmerpenguins dataset

[***Gbadamassi G.O. Dossa***](https://scholar.google.com/citations?user=w20AkZcAAAAJ&hl=en)

last updated: 2022-06-28

# 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).

## Warning: package 'palmerpenguins' was built under R version 4.1.1

## Warning: package 'ggplot2' was built under R version 4.1.3

## Warning: package 'tibble' was built under R version 4.1.3

## Warning: package 'tidyr' was built under R version 4.1.3

## Warning: package 'readr' was built under R version 4.1.3

## Warning: package 'dplyr' was built under R version 4.1.3

## Data

The data are displayed below (first 10 rows) :

penguins %>%   
 slice(1:10) %>%  
 knitr::kable()

| 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

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:

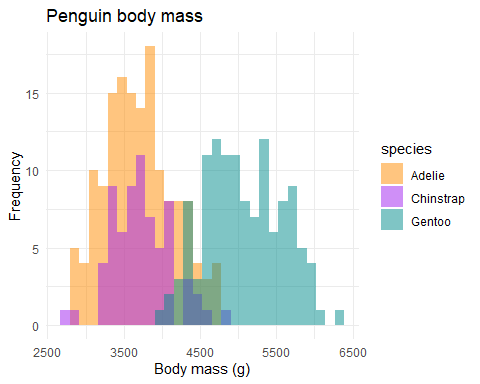
## # A tibble: 3 x 6  
## species bill\_length\_mm bill\_depth\_mm flipper\_length\_mm body\_mass\_g year  
## <fct> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Adelie 38.8 18.3 190. 3701. 2008.  
## 2 Chinstrap 48.8 18.4 196. 3733. 2008.  
## 3 Gentoo 47.5 15.0 217. 5076. 2008.

## 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, Brendan, Michael J. Malick, Gregory T. Ruggerone, Pete Rand, Milo Adkison, James R. Irvine, Robert Campbell, and Kristen Gorman. 2020. “Climate and Competition Influence Sockeye Salmon Population Dynamics Across the Northeast Pacific Ocean.” *Canadian Journal of Fisheries and Aquatic Sciences* 77 (6): 943–49. <https://doi.org/10.1139/cjfas-2019-0422>.