NOVA IMS INFORMATION MANAGEMENT SCHOOL

Exercises: Exploring Data

Statistics for Data Science

Master Program in Advanced Analytics

2025/2026

Consider the file `penguins.xlsx`, available in moodle. Before starting any analysis or file importation, you are going to install some useful packages:

- `openxlsx`, for data importation,
- `tidyverse`, for data transformation and data tidying..
- `ggplot2`, for data visualization.

A lot of useful functions will not be covered in this class, since the focus of this course is not to teach R programming. However, it is expected for you to have some basic skills. We advice you to read the recommended bibliography (i.e., R for Data Science (2e)) and to solve the exercises of the following practical classes in R on your own (it will be useful later on for the project!).

Let's begin...

1. Import the dataset `penguins.xlsx` using the function `read.xlsx` from the package `openxlsx`.
   Before using any function, you need to load the package you want to use by typing: `library(openxlsx)`.
   You need to do this for every new package you plan to use and everytime you initiate R.
   Check help to see the arguments required for the function `read.xlsx` to import the file in R.
2. Calculate the mean and standard deviation of flipper length of the penguins in the dataset.
   Hint: Check the slides from the theoretical class.
3. Calculate the number and proportion of penguins for each island.
   Hint: Check the slides from the theoretical class.
4. Calculate the correlation between body mass and flipper length. Comment the result.
   Hint: Use the function `na.omit` to remove the missing values.
5. In the previous points, the functions used are from R base. The package `tidyverse` includes a set of useful functions to transform and tidy data. The code below gives the average mean of flipper length per island:

```
library(tidyverse)
library(palmerpenguins)
penguins %>%
  group_by(island) %>%
  summarise(mean = mean(flipper_length_mm, na.rm=TRUE))

## # A tibble: 3 x 2
##   island     mean
##   <fct>     <dbl>
## 1 Biscoe     210.
## 2 Dream      193.
## 3 Torgersen  191.
```

   The symbol "%>%" is called "the pipe" and it allows to connect several functions in R, sequentially. This is useful if you have a set of transformations you want to perform. Particularly, the code above can be read as "We have penguins dataset, **then** we group by island, **then** we summarise by calculating the mean of flipper length".
   Knowing this, find the mean body mass per species.

6. Consider the code below:

```
library(ggplot2)
ggplot(penguins) +
  geom_bar(aes(x=island))
```

This allows to obtain a bar plot of the number of penguins per island. The package `ggplot2` is a system for creating graphics. The syntax is always the following:

```
ggplot(<dataset-name>) +
  <geom_function>(mapping=aes(<mappings>))
```

The <geom-function> will determine the type of plot, the <mappings> will determine the content of the plot (by defining x and y axis, for example).
Knowing this, create a scatter plot of flipper length against body mass.
Hint: Check the slides from the theoretical class.