# Lab 4 Task

# Penguins dataset

**Dataset:** `penguins.csv` (Palmer Penguins — columns: species, island, bill_length_mm, bill_depth_mm, flipper_length_mm, body_mass_g, sex, year)
**Tools:** Pandas, Seaborn, Matplotlib

## Tasks

1. Load the dataset and show first 10 rows. Identify datatypes and count of missing values per column.
2. Visualize missingness (heatmap or bar chart) and decide which columns (if any) to drop because of excessive missingness — explain your decision in one sentence.
3. Calculate and display the correlation matrix for numeric features (`bill_length_mm`, `bill_depth_mm`, `flipper_length_mm`, `body_mass_g`) and plot it as a heatmap.
4. Create a pairplot of the numeric features, colored by `species`. Comment (one line) on which feature pairs best separate species.
5. Draw boxplots of `body_mass_g` for each `species` and for each `island` (two separate plots). Note any islands with systematically heavier/lighter penguins.
6. Make a scatter plot of `bill_length_mm` vs `bill_depth_mm` colored by species; add `sex` as marker shape if available.
7. Using groupby, compute the mean and standard deviation of `flipper_length_mm` for each combination of `species` and `island`. Show results as a tidy table.
8. Handle missing values: choose one reasonable imputation strategy for numeric columns (explain why) and apply it; then show before/after missing counts.
9. Create a histogram of `body_mass_g`, faceted by `species` (use seaborn `FacetGrid` or `displot` with `col=species`).
10. Short modelling preparation: create a new dataframe with only numeric features and encode `species` to numeric labels — save it as `penguins_for_model.csv`.

## Exercise:

A. Which two numerical features seem to be most predictive of `species`? Justify with visuals/tables.
B. Is there significant sexual dimorphism (difference between sexes) in `body_mass_g` or `flipper_length_mm`? Show supporting plot(s).
C. Write a 5–7 line conclusion: two most important biological factors affecting penguin size or species differences, based on your EDA.