DEPARTMENT OF
DATA SCIENCE

DS311
Exploratory Data Analysis

Make Sense of Data & Univariate Analysis

Section 3: Lab Activity

Let's put it all together with a hands-on **case study**. We'll use the **Palmer Penguins** dataset (`penguins` data frame from the `palmerpenguins` package). This real dataset contains measurements of three penguin species (Adelie, Chinstrap, Gentoo) in Antarctica. The variables include species, island, sex, and numerical measurements like flipper length, bill dimensions, and body mass.

Scenario: Imagine we are exploring this penguin data to understand their physical characteristics. We will focus on the **body_mass_g** (body mass in grams) of the penguins and how it varies by species and overall.

Tasks:

1. **Identify Variables:** We want to investigate if there are differences in penguin body mass between species. In this context, what would be the independent and dependent variables? (State which variable is independent and which is dependent for this question.)
2. **Summary Statistics:** Calculate basic summary statistics for the penguins' body mass (the `body_mass_g` variable). Specifically, find the **mean**, **median**, and **standard deviation** of body mass in the entire dataset. (Ignore any missing values in calculations.)
3. **Category Counts:** How many penguins of each **species** are in the dataset? Provide the count for each species (Adelie, Chinstrap, Gentoo).
4. **Histogram:** Plot a histogram of the penguins' body mass. Describe the distribution shape you observe (for example, is it symmetric, skewed, bimodal, etc.?).
5. **Hypothesis Test:** Perform a one-sample t-test on `body_mass_g`. Test whether the **average penguin body mass** is **4500 grams** (H_0: mean = 4500) at the 5% significance level. State the null and alternative hypotheses in words, and conclude whether the data provide evidence that the true mean body mass is different from 4500 g.

Note: For any computations, you may use either base R or tidyverse approaches. When plotting, feel free to use base `hist()` or `ggplot2::geom_histogram()`. Make sure to handle missing data where appropriate (e.g., `na.rm=TRUE` in calculations or using `na.omit()`).

Take your time to write out not just the results but also a short interpretation for each answer. Good luck!