DEPARTMENT OF
DATA SCIENCE

DS311
Exploratory_Data_Analysis

Introduction to EDA

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** (H0: 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()`).

1. Identify Variables

```
library(palmerpenguins)
```

Independent variable:

`species` → because we are comparing different groups of penguins (Adelie, Chinstrap, Gentoo).

Dependent variable:

`body_mass_g` → because this is the measurement to analyze to see if it changes depending on species.

2. Summary Statistics:

`body_mass_g` summary statistics:

```
summary(penguins$body_mass_g)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    2700    3550    4050    4202    4750    6300       2
```

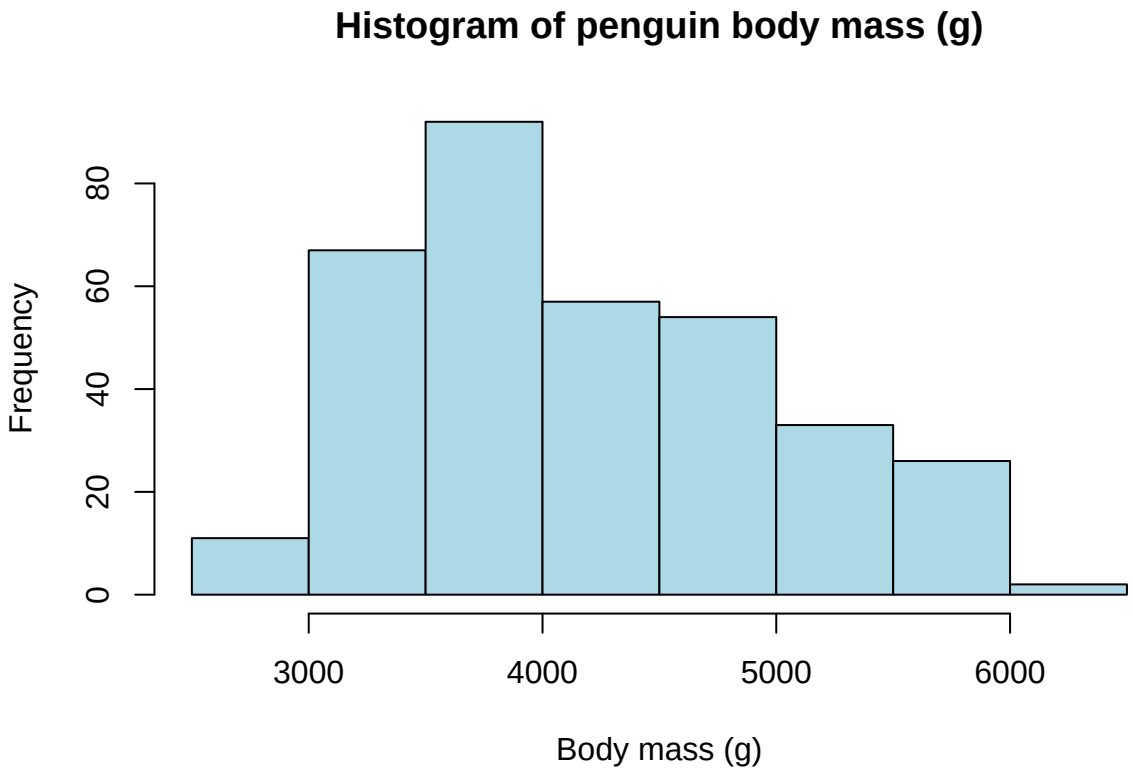
Standard deviation:

```
## [1] 801.9545
```

3. Category Counts:

```
##
##    Adelie Chinstrap    Gentoo
##       152        68       124
```

4. Histogram:



The distribution of penguin body mass is roughly **symmetric**, with a slight **right skew**, since there are few penguins with much higher body masses.

5. Hypothesis Test:

Test whether the **average penguin body mass** is **4500 grams** (H0: mean = 4500) at the 5% significance level.

```
##
##  One Sample t-test
##
## data:  penguin_body_mass
## t = -6.8776, df = 341, p-value = 2.917e-11
## alternative hypothesis: true mean is not equal to 4500
## 95 percent confidence interval:
##  4116.458 4287.050
## sample estimates:
## mean of x
##  4201.754
```

Conclusion

- The p-value is extremely small (< 0.50), so we fail to accept the null hypothes *Ho*.

- There is strong statistical evidence that the true mean body mass of penguins is different from 4500 grams.