Exercise: Data Preprocessing and Visualization with the Penguins Dataset

Overview

In this exercise, you will explore the **Palmer Penguins dataset**, a real-world dataset describing physical characteristics of penguins from three species observed in Antarctica. You will perform data preprocessing and create exploratory visualizations focused exclusively on **numerical features**, mirroring the techniques used in the California Housing dataset analysis.

Objectives

By completing this exercise, students will be able to:

- Load and inspect a structured dataset
- Identify and handle missing data
- Select and isolate relevant numeric features
- Create meaningful visualizations to analyze trends and relationships
- Interpret patterns and correlations among variables

Dataset Summary

The **Palmer Penguins dataset** includes observations for three penguin species: Adelie, Chinstrap, and Gentoo. For the purpose of this exercise, you will work exclusively with the following **numerical attributes**:

- **Bill Length (mm)**
- **Bill Depth (mm)**
- **Flipper Length (mm)**
- **Body Mass (g)**

Instructions

1. Data Familiarization

- Load the Palmer Penguins dataset.
- Display a preview of the dataset.
- Review the structure of the data, summary statistics, and data types.
- Identify and quantify missing values.

2. Data Preprocessing

- Remove any rows that contain missing values.
- Create a new dataset that includes only the selected numeric features.
- Ensure all variables are properly formatted for numerical analysis.

3. Data Visualization

Using a statistical plotting library of your choice, create the following visualizations:

- **Histogram** showing the distribution of penguin body mass
- **Scatter plot** to analyze the relationship between flipper length and body mass
- **Correlation heatmap** to reveal relationships between all numeric variables

Ensure all visualizations are clearly labeled with appropriate titles and axis labels.

4. Analysis and Interpretation

Based on your findings:

- Identify which variables show strong positive or negative correlation
- Describe the nature of the relationship between body mass and flipper length
- Note any notable trends or outliers observed in the distributions

Summarize your insights in 2–3 well-written paragraphs or a bullet point list.

Deliverables

- A well-structured Jupyter notebook containing:
    - Data inspection and cleaning steps
    - Clearly formatted and labeled visualizations
    - A summary section with your key insights