# Penguins Dataset Analysis Report

## 1. Dataset Description

The dataset used in this analysis is the 'Penguins' dataset from the Seaborn library. This dataset contains measurements of penguins from different species collected from various islands in the Palmer Archipelago, Antarctica.

* - \*\*species\*\*: The species of the penguin (Adelie, Chinstrap, or Gentoo).
* - \*\*island\*\*: The island where the penguin was observed (Biscoe, Dream, or Torgersen).
* - \*\*bill\_length\_mm\*\*: The length of the penguin's bill in millimeters.
* - \*\*bill\_depth\_mm\*\*: The depth of the penguin's bill in millimeters.
* - \*\*flipper\_length\_mm\*\*: The length of the penguin's flipper in millimeters.
* - \*\*body\_mass\_g\*\*: The body mass of the penguin in grams.
* - \*\*sex\*\*: The sex of the penguin (Male or Female).

There were some missing values in 'bill\_length\_mm', 'bill\_depth\_mm', 'flipper\_length\_mm', 'body\_mass\_g', and 'sex'. These were handled during the data cleaning process.

## 2. Data Cleaning

The following steps were performed to clean the dataset:

* - Checked for missing values using `isnull().sum()` and filled numerical column NaNs with the column mean.
* - Removed rows where 'sex' was missing using `dropna(subset=['sex'], inplace=True)`.
* - Identified duplicate rows using `df.duplicated().sum()` (though no duplicates were found).
* - Renamed column 'island' to 'Island' and 'species' to 'Species' for standardization.
* - Removed outliers from numerical columns using the IQR method.

## 3. Exploratory Data Analysis (EDA)

The cleaned dataset was analyzed using the following steps:

* - \*\*Load Data\*\*: Loaded the dataset using `sns.load\_dataset('penguins')`.
* - \*\*Univariate Analysis\*\*: Displayed summary statistics (`describe()`) for numerical columns.
* - \*\*Frequency Distribution\*\*: Counted the occurrences of species and sex categories.
* - \*\*Visualizing Distribution\*\*: Used histograms and boxplots to visualize numerical distributions.
* - \*\*Bivariate Analysis\*\*: Computed correlation between numerical variables and plotted scatter plots.
* - \*\*Grouping & Aggregation\*\*: Calculated the mean and median values of 'body\_mass\_g' and 'flipper\_length\_mm' for each species.
* - \*\*Combined Visualizations\*\*: Used heatmaps and pair plots to study relationships among numerical variables.
* - \*\*Box & Violin Plots\*\*: Compared distributions of 'body\_mass\_g' and 'bill\_length\_mm' across species and sex.