Executive Summary: Penguin Species Identification Using K-Means Clustering

The goal of this project was to assist researchers in identifying potential groups of penguins from a dataset provided by Dr. Kristen Gorman and the Palmer Station, Antarctica LTER. The dataset, collected in Antarctica, contains five columns: culmen length, culmen depth, flipper length, body mass, and sex. While the species of penguins were not recorded, it is known that at least three species—Adelie, Chinstrap, and Gentoo—are native to the region.

To identify distinct groups (potential species), the following steps were taken:

1. **Data Loading and Inspection**:
    - The dataset was loaded and basic information was examined, including the column types and non-null values.
2. **Preprocessing**:
    - Dummy variables were created for categorical features (sex), ensuring they were properly encoded for clustering.
    - Data was standardized using `StandardScaler`, which is crucial before applying clustering algorithms to ensure all variables are on the same scale.
3. **Clustering (K-Means)**:
    - The optimal number of clusters was determined using the **Elbow Method**, which showed that the ideal number of clusters was **4**.
    - The K-Means algorithm was applied with **4 clusters**, and labels were assigned to each penguin in the dataset.
4. **Visualization**:
    - A scatter plot was created to visualize the clusters, focusing on the `culmen_length_mm` feature. This allowed us to observe how penguins were grouped into distinct clusters based on this characteristic.
5. **Final Cluster Statistics**:
    - A summary DataFrame (`stat_penguins`) was generated, containing the average values of culmen length, depth, and flipper length for each of the identified clusters.

In summary, the K-Means clustering approach successfully identified 4 distinct groups of penguins based on their physical measurements. These clusters may correspond to different penguin species, although further biological analysis would be necessary to confirm this. The results offer a valuable foundation for future research into penguin species identification in Antarctica.