Exploratory Data Analysis and Clustering on the Palmer Penguins Dataset

Name: Md Shahedur Rahman

Student ID: 23036883

Github Repository: github.com/shahedur23036883/clustering

Introduction

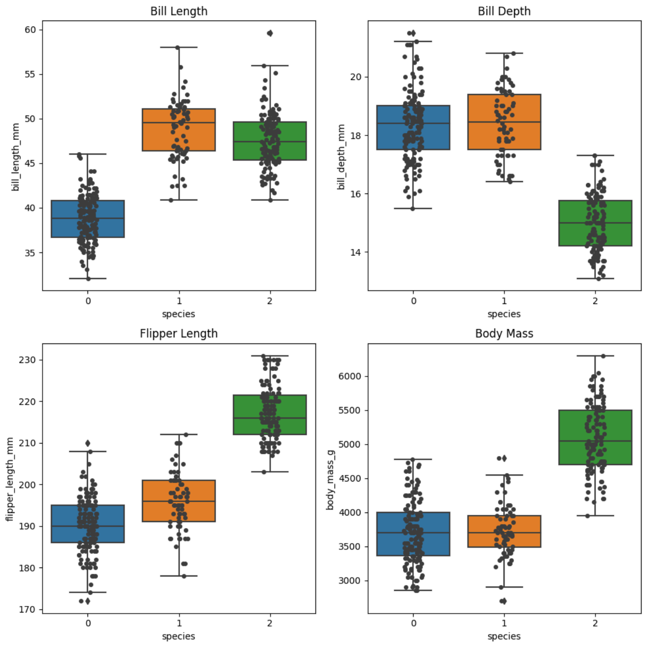
This report explores the Palmer Penguins dataset, which contains measurements and other data on three penguin species found in the Palmer Archipelago in Antarctica. The goals are to perform exploratory data visualization, cluster the penguins into groups based on their physical characteristics and fit a linear regression model to examine relationships between variables.

Data Preprocessing

The raw data was cleaned by removing rows with missing values and encoding the species names numerically. Infinite values were also replaced with NaN.

Exploratory Data Visualization

To get an initial overview of the variables and differences between species, boxplots and stripplots were created for the key measurement variables:



Bill Length

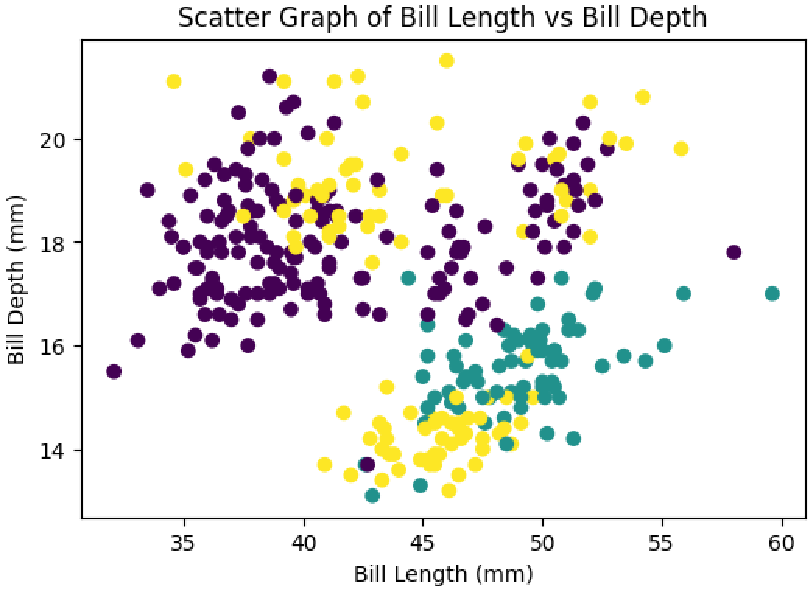
Bill Depth

Flipper Length

Body Mass

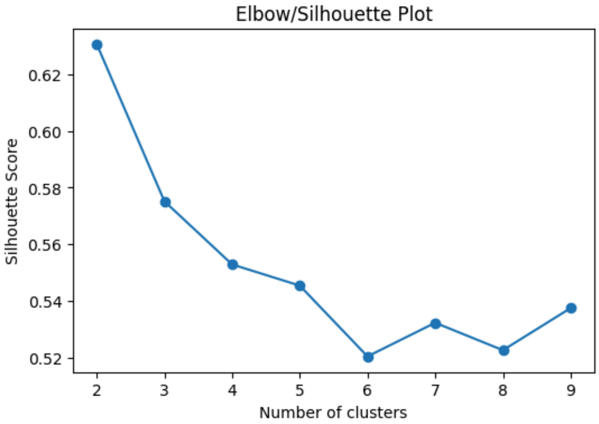
The boxplots show that the Gentoo penguins tend to be larger than the other two species across all measurement variables. There is also some overlap in the ranges between Adelie and Chinstrap penguins.

A scatter plot of bill length vs bill depth colored by cluster assignment shows potential clustering in the data:

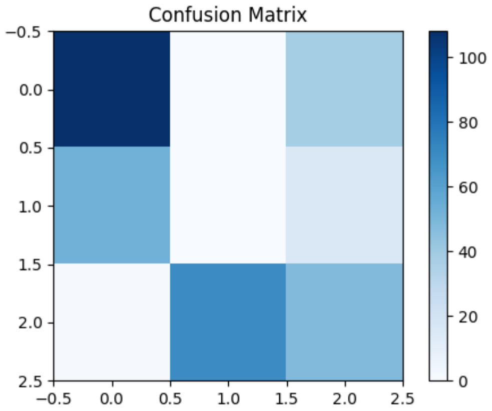


K-Means Clustering

To explore clustering in an unsupervised manner, k-means clustering was applied to the measurement variables using sklearn's KMeans class. The optimal number of clusters was determined by computing the silhouette score for 2 to 9 clusters:



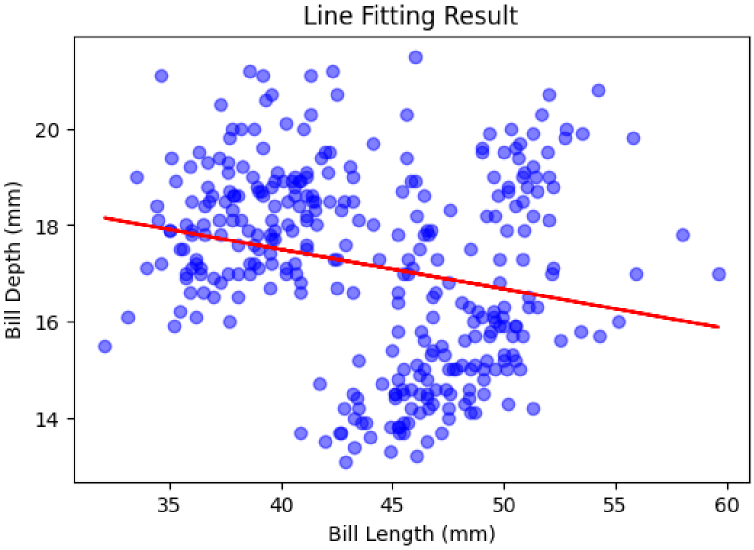
The peak silhouette score occurs at 3 clusters, corresponding to the three known species in the data. The k-means clustering assignments were compared to the true species labels via a confusion matrix:



The confusion matrix shows that while the 3 clusters separate out the Gentoo penguins well, there is some mixing between the Adelie and Chinstrap clusters.

Linear Regression

As an example of a supervised learning technique, linear regression was used to model bill depth as a function of bill length:



The regression line shows a reasonably strong positive linear relationship between bill length and depth across the combined species.

Conclusion

This exploratory analysis has uncovered some key patterns and relationships in the Palmer Penguins dataset. The different penguin species exhibit distinct distributions across the measured variables, which allow them to be clustered reasonably well, though some species show overlap. Linear regression also revealed the expected positive correlation between bill length and depth measurements. Further analysis could explore non-linear relationships, additional predictor variables, and clustering by other characteristics like island location.