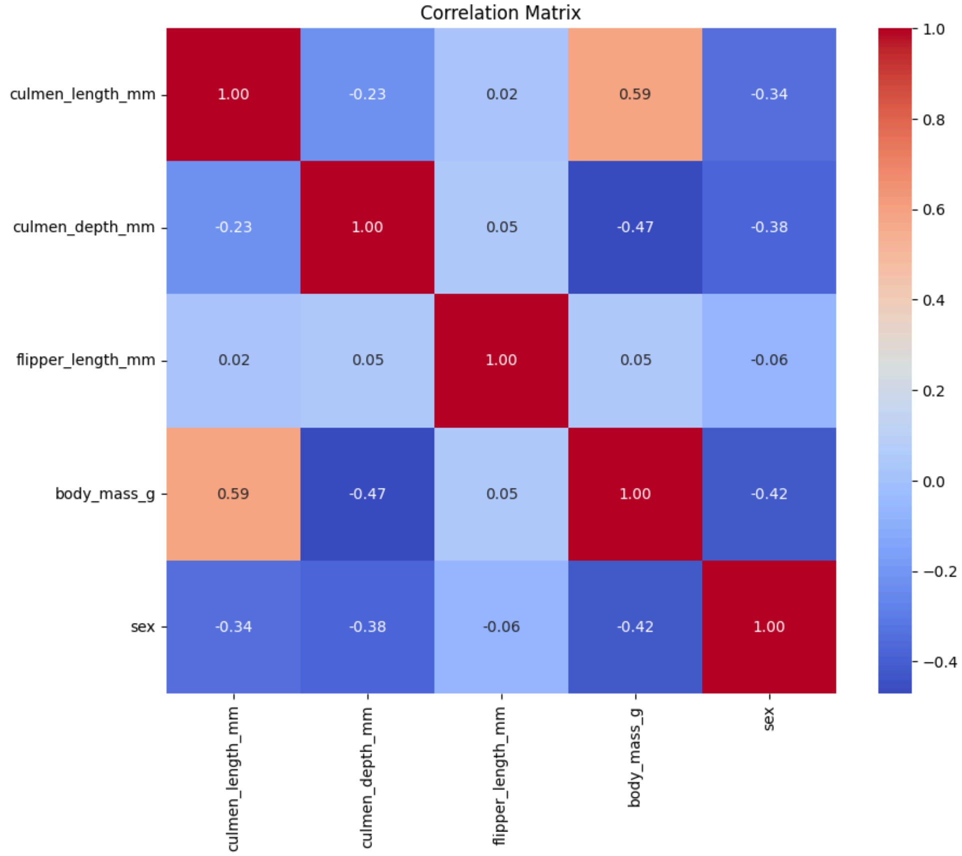
Clustering and Fitting Analysis Report

Introduction:
This report presents a comprehensive analysis of the "Penguins Dataset" using clustering and fitting techniques. The dataset contains measurements of various attributes of penguins, such as culmen length, culmen depth, flipper length, body mass, and the sex of the penguins. The analysis aims to uncover insights into the dataset's structure, identify natural groupings within the data, and model the relationship between variables using linear regression.

Data Preprocessing:
The dataset underwent thorough preprocessing to handle missing values and encode categorical variables. Rows with missing values were dropped, and the categorical variable "sex" was encoded into numeric values to facilitate analysis.
(https://www.kaggle.com/datasets/youssefaboelwafa/clustering-penguins-species/data)

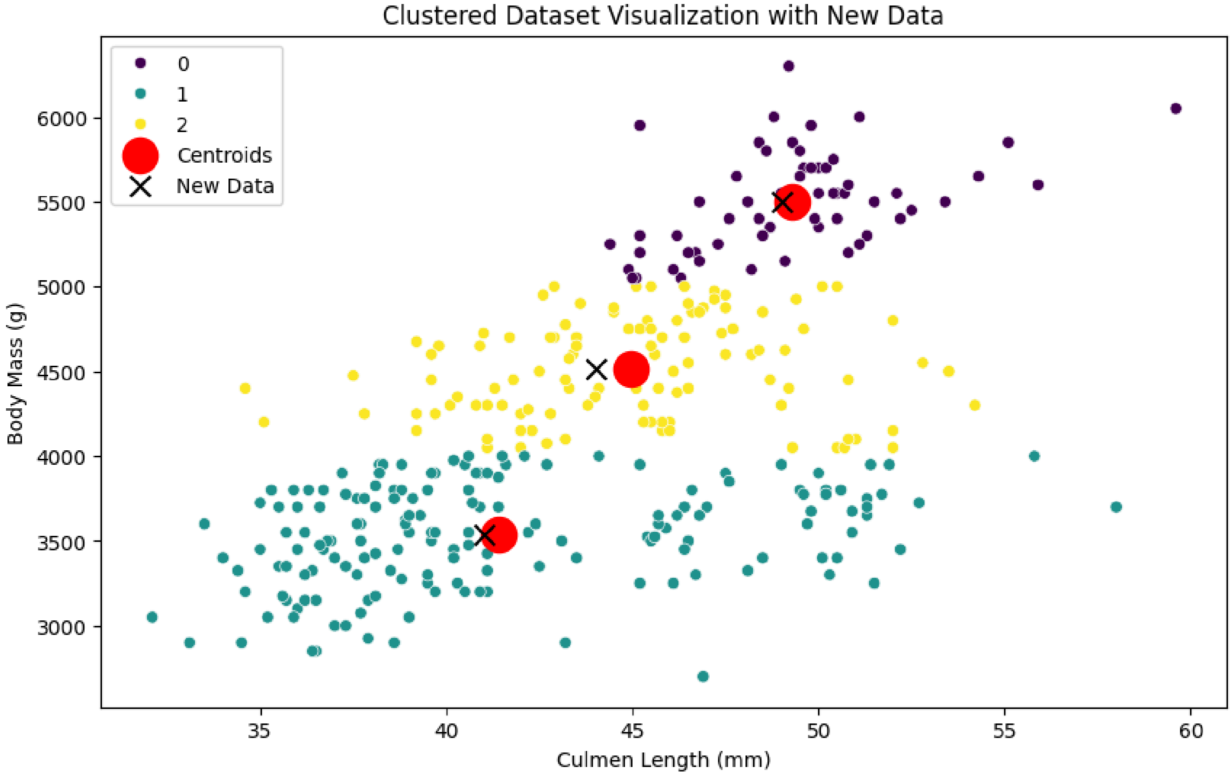
Exploratory Data Analysis:
Correlation Patterns: The correlation matrix heatmap revealed relationships between different features, helping us understand how they are related to each other. For example, we observed correlations between culmen length, body mass, and flipper length.



Clustering Analysis:

Cluster Analysis: Through KMeans clustering, we identified distinct clusters within the dataset based on culmen length and body mass. This allowed us to group penguins with similar physical characteristics together, providing insights into potential species or population differences.

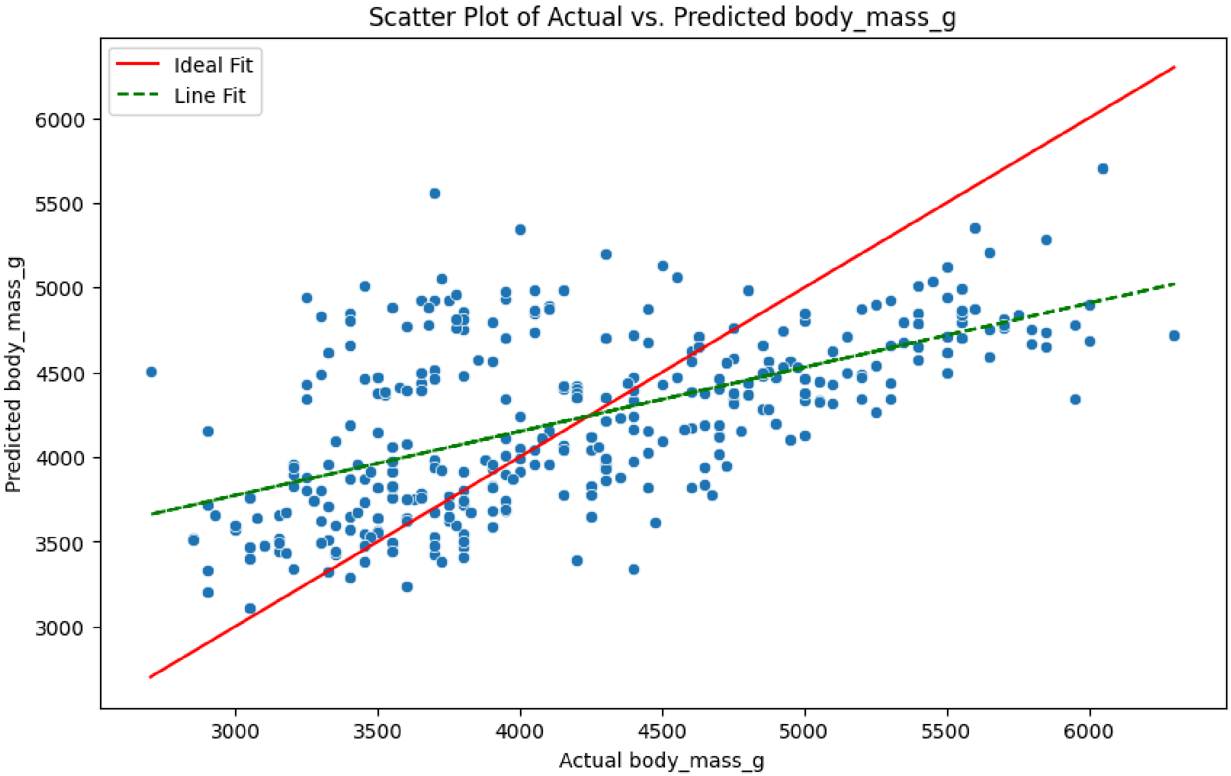
New Data Prediction: By predicting clusters for new data points, we were able to classify them into the existing clusters. This prediction can be useful for categorizing new penguin observations based on their culmen length and body mass.



Fitting Analysis:

Regression Analysis: The linear regression model fitted to the data revealed a relationship between culmen length and body mass.

Model Evaluation: The mean squared error calculated for the regression model allowed us to assess its predictive accuracy. Lower error values indicate better model performance, providing confidence in the relationship between culmen length and body mass.



Overall, these insights enhance our understanding of the penguin dataset and can be valuable for various applications such as species classification, ecological studies, and conservation efforts.

Name: Sravanth Baratam
StudentID:23001152
Code: https://github.com/sravanth-space/ads_custering_and_fitting