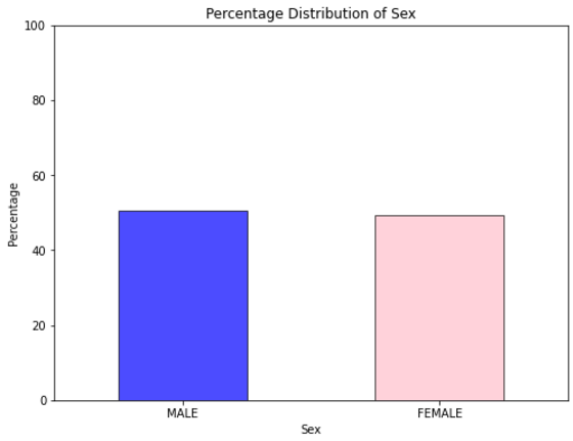
Penguin Data Analysis: Clustering and Fitting Insights

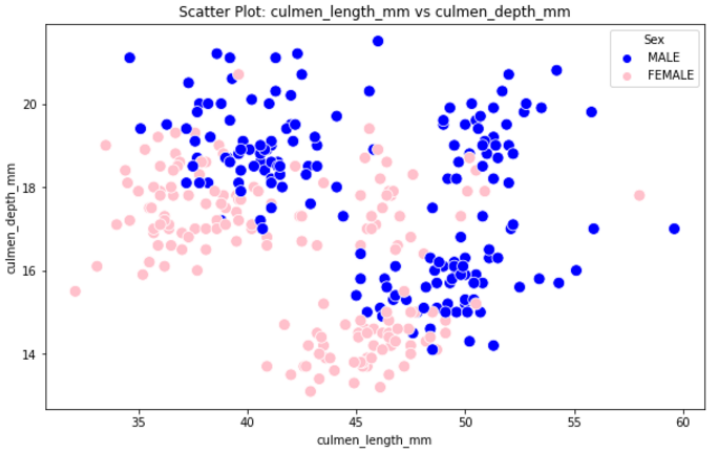
Name: Sriharshini Thatiparthi **Student ID: 23070133** **GitHub Link: Link**

A basic unsupervised machine learning method, clustering groups data points according to their commonalities. This work investigates clustering using a dataset of measurements of penguins from many species. Key elements of the dataset are culmen length and depth (in millimetres), flipper length (in millimetres), body mass (in grammes), and penguin sex. < These traits provide insightful analysis of the physical traits of the penguins and provide the basis for meaningful groupings of them. This work's main objective is to examine the dataset and find natural grouping of penguins depending on their physical characteristics. By use of clustering techniques like k-means or hierarchical clustering, underlying patterns may be found and aid differentiate between the species.

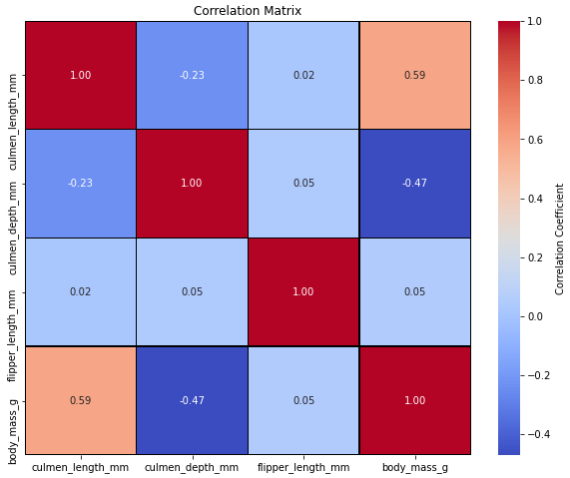
The study starts by importing necessary Python tools for data processing and visualising including pandas, matplotlib, and seaborn. Following loading the penguin's dataset, a summary showed that the dataset had 344 items including some missing values in all columns, most notably in the "sex" column with 9 missing entries. All rows with missing values were eliminated to guarantee pure data for study, therefore shrinking the dataset to 335 entries.



Computing the counts of male and female penguins revealed a about equal distribution, hence exploring the "sex" column. The percentages were then shown on a bar chart with clear, separate colours. The chart clearly shows the balanced representation of sexes, therefore providing a visual understanding of the data distribution. The bar chart displays the cleaned dataset's male and female penguin percentage distribution. Men make a somewhat higher percentage than women, however their distribution is about equal. This balance implies that the dataset is representative in terms of sex, thereby enabling objective examination of features related with gender.

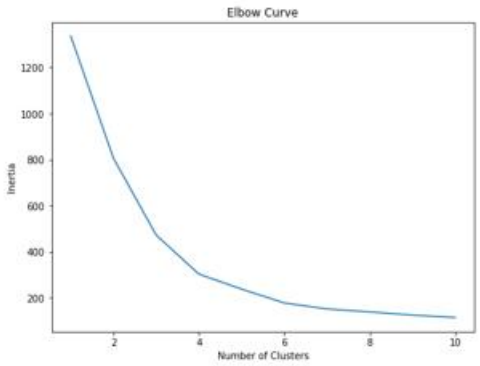


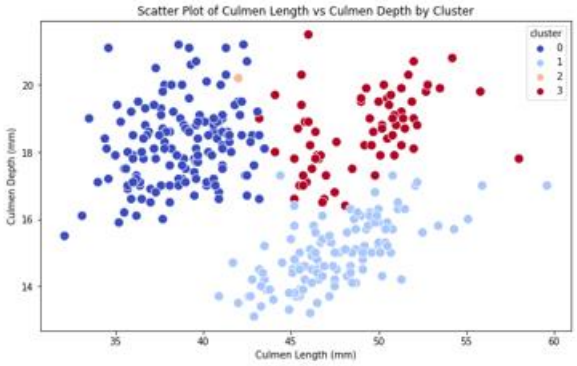
Using colours (blue for men and pink for women), I was able to see in this scatter plot the association between culmen_length_mm and culmen_depth_mm while separating male from female penguins. With obvious grouping patterns, the graphic clearly shows how these two factors differ across sexes. Although culmen depth displays overlaps between the sexes, males usually seem to have greater culmen lengths than women.



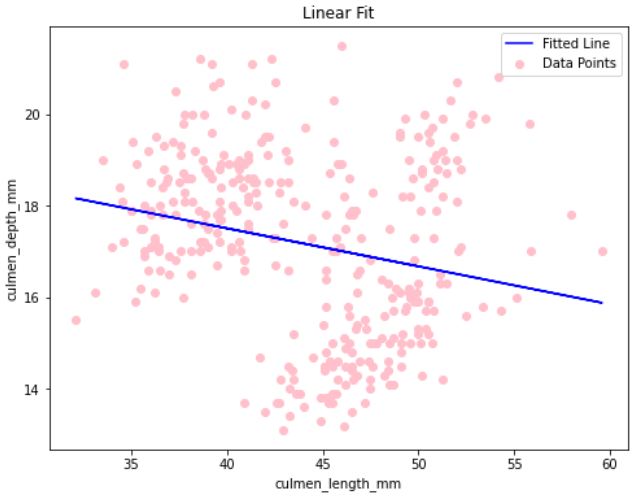
To see the numerical feature correlation matrix in the dataset, I produced a heat map. Colours show the strength and direction of the correlations, therefore highlighting the links between variables. Body mass-g, for example, is inversely connected with culmen's depth-mm but favourably correlated with culmen's length-mm.

StandardScaler helped me scale the numerical variables—culmen_length_mm, culmen_depth_mm, flipper_length_mm, and body_mass_g—to guarantee equal weighting. I calculated inertia for cluster counts between 1 and 10 and plotted the Elbow Curve to find the best number of clusters.





At four clusters, the curve exposed a clear "elbow," indicating this as the best option. For quicker convergence, I used a well-initialized centroids approach (k-means++) using the K-Means model with four clusters. The data will enable me to investigate trends across penguin species or sizes; yet, I must confirm significant separation by means of domain knowledge or silhouette analysis. I arranged by clusters the culmen length against culmen depth. Although overlaps point to certain restrictions, the well-separated clusters demonstrate that K-Means efficiently caught differences.



To examine the association between culmen depth and culmen length, I ran a basic linear regression. Appropriate a linear model to the data by using culmen depth as the dependent variable and culmen length as the independent variable. The fitted regression line is blue; the scatter plot displays pink data points. Although the trend hints to a slight negative association, data point distribution shows considerable variance. This implies that the connection could benefit from further investigation as it seems to be not substantially linear.