APPLIED DATA SCIENCE 1 -CLUSTERING AND FITTING

Penguin Size Analysis

Name: Daswadayalan Myladumparai Deenadayalan

Student Id: 23068427

GitHub repository:https://github.com/MDD25-Web-data-science/cluster_fitting_analysis

Introduction:

The Palmer Archipelago is located in the icy water of antarctica is home for penguins. The birds were attracted the nature enthusiasts and scientists. In this dataset we going to analyse our feathered inhabitants, exploring their physical characteristics and insights.

Let us introduce our feathered friends Adélie, Chinstrap, Gentoo in this datasert we are going to analyse and predict.

Dataset Overview:

The dataset contains info about adelie, Chinstrap, Gentoo penguins. Key column include species, island, culmen length(mm),culmen depth(mm),flipper length(mm),body mass (g) and sex.

Descriptive Statistics:

Culmen length mm: There is a slight shift to the left which is indicated by the negative skewness. The kurtosis value is the same as that of a normal distribution, but it suggests a slightly heavy tail. **culmen depth mm:** The data set is positively skewed, and the tail extends out to the right beyond the mass of the data. The kurtosis is close to zero, showing a distribution that is close to normal**. flipper length mm:** There is a slight shift to the left which is indicated by the negative skewness**.** The kurtosis suggests a lighter tail than a normal distribution.**body_mass_g:** The data is unusually distributed with the left long tail, while the kurtosis is very close to 0 that all concentrate at ±1, i.e., equal tails, which generally lead to easily interpreted charts.

Correlation

Negative relationship: Increasing the value of one of the variables results in a decrease in the other one (e.g. culmen_length_mm and culmen_depth_mm).**Positive relationship**: When one variable goes up, the other one also goes up (e.g. culmen_depth_mm and flipper_length_mm).**Low correlation**: The correlation coefficient is near zero, which means there is a weak linear relationship between two variables.

High correlation: The coefficient is close to -1 or 1, which means a very strong linear relationship.

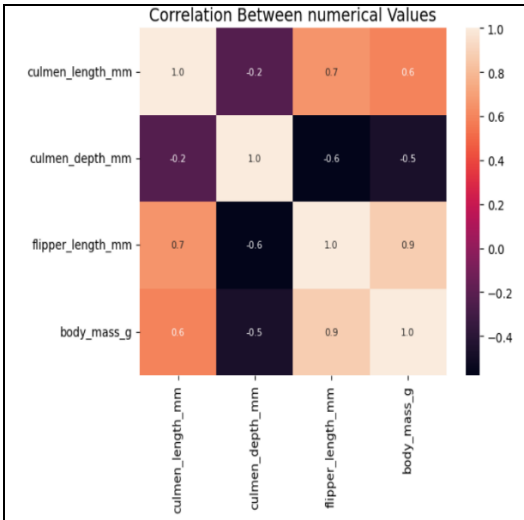

Fig1.1 correlation between numeric values

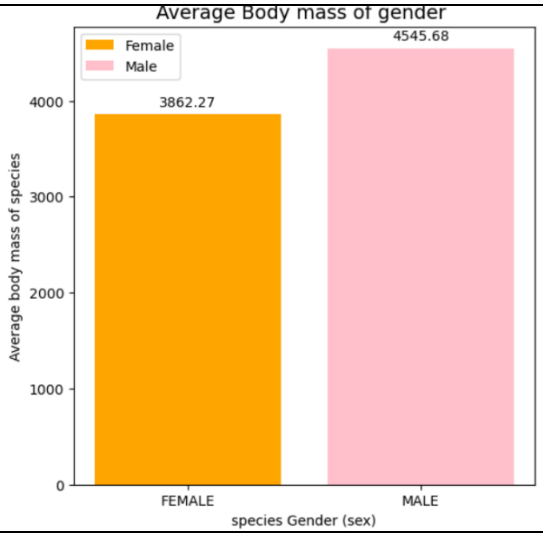

Fig 1.2 Average Body mass of gender

Categorical Graph:

The categorical is dataset where the sex is grouped and the body mass of the penguins were averaged. The male gender has a more body mass than the female gender penguins.

Relational Graph:

The Relational graph is relation between two columns. In this dataset I have taken the culmen_length_mm and culmen_depth_mm and then plotted with scattered plot.

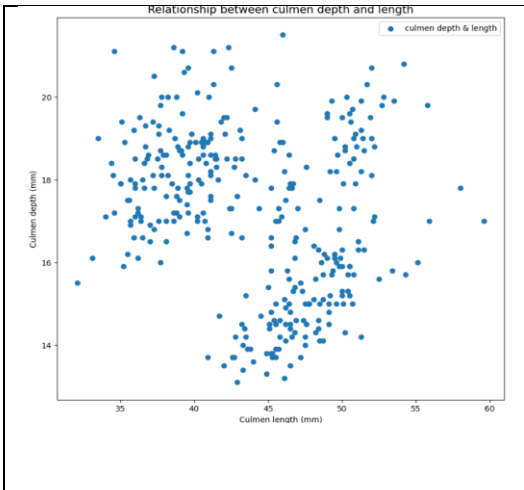

Fig 1.3 Relationship between depth and length of culmen

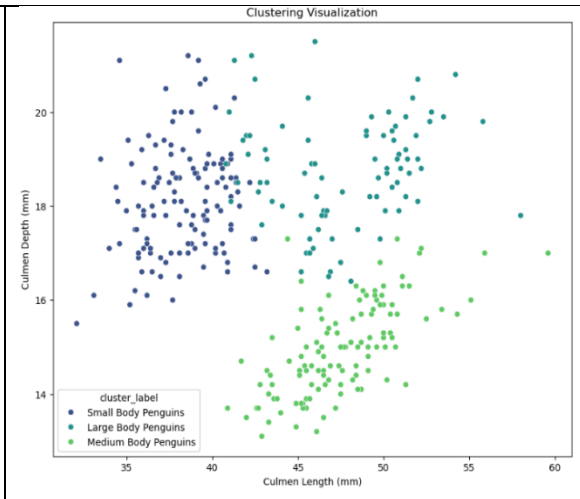

Fig 1.4 Clustering visualization

Kmean clustering:

The K-Means algorithm is a method of data separation that divides the data into a pre-specified number of clusters by variance(inertia) minimizing. The sum of squared trajectories of each point to the centroid of its cluster. Lower inertia among other things (tighter clustering). The Elbow Method also involves plotting the inertia versus the number of clusters (k)to search for the "elbow point" where the speed of decreasing inertia evens out. It is recommended by the elbow point to take into account the optimal number of clusters.

Interpretation of the Plot:

With smaller k values, the inertia is more prominent as the model spends clusters, trying to make everything fit together. Beyond the correct elbow point, the inertia becomes tumultuous that points at the fact that the extra classes capture almost no information.

Optimal Number of Clusters (k=3):

The elbow point is the optimal k in most of the cases. For example, when the plot displays a sharp turn at k=3k=3k=3, this indicates that three clusters are the best for the dataset.

Interpretation of the Fitted Line:

The regressive line demonstrates data trend as the regression line indicates the trend of the data. **Positive Slope**: If the line goes up, this suggests that culmen depth as culmen length increases is likely to increase. **Negative Slope**: If the line points downwards, it points out a reverse relationship. **Flat Line**: Means that there is no significant relationship between the two variables.

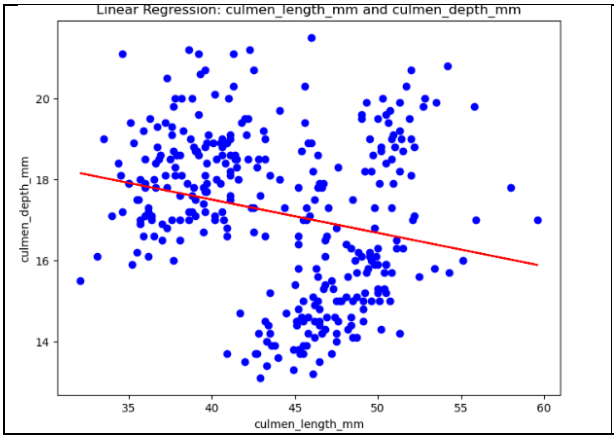

Fig1.5 line regression

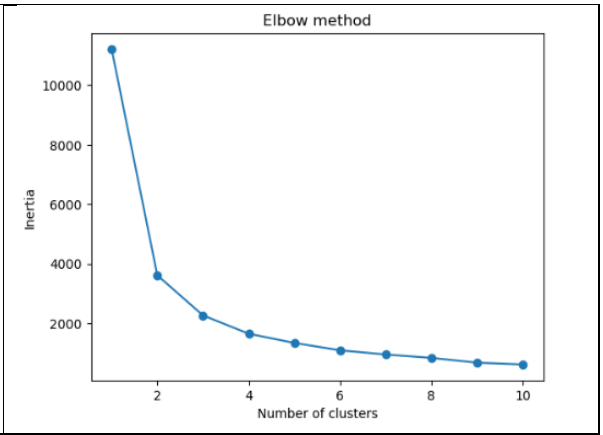

Fig 1.6 Elbow Method

Conclusion:

The model achieved an **accuracy score** of accuracy on the test data. This indicates that approximately accuracy * 100 of the test samples were correctly classified by the model.