Data Analytics - Project 2 / Course Project N.01: Fish

Miguel Aldaz, Bhanu Prakash

May 22, 2024

1 Introduction

This document outlines the methodology and results for the Data Analytic's Project 2. Here, a group of two students was tasked to apply concepts seen in class to describe and analyse a given data set.

1.1 Fish Species Data Analysis

This report provides a comprehensive analysis of a dataset containing measurements for various fish species. The dataset includes measurements for body length, head length, width, weight, and species. Specifically, it contains information on three fish species: goldfish, island mackerel, and Mugilidae. Each species has distinct characteristics that can be analyzed to understand their physical attributes and differences.
We were tasked to use python and additional libraries to address the following points:

- Data exploration and description
- Pre-processing the data
- Use two different clustering algorithms
- Compare the accuracy of the two models

2 Data Overview and Analysis

2.1 Dataset Overview

We are given excel file contains the dataset includes the following columns:

- body.length: The total length of the fish body in unspecified units.
- head.length: The length of the fish head in unspecified units.
- width: The width of the fish in unspecified units.
- weight: The weight of the fish in unspecified units.
- species: The species of the fish (goldfish, island mackerel, or Mugilidae).

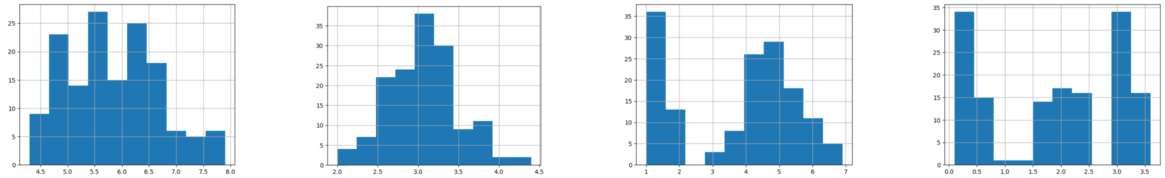
Body Length	**Head Length**	**Width**	**Weight**	**Species**
5.1	3.5	1.4	3.2	goldfish
4.9	3.0	1.4	3.2	goldfish
4.7	3.2	1.3	3.2	goldfish
.	.	.	.	.

Table 1: Sample Data of Goldfish From Excel

2.2 Dataset Analysis

2.2.1 Data Overview

To fully understand the dataset, we have inspected each column in order to extract some useful information. Firstly, the species column was easily analyzed, obtaining that we have 3 different species equally distributed, having 50 entries for each. Then, to analyze numeric columns, we have looked at their values using histograms.



(a) body length

(b) head length

(c) width

(d) Weight

Figure 1: Histograms of the different columns

First of all, the weight column presents a scenario that is catching the eye. Three groups on the graph are easily distinguished from one another: heavier than 2.75, from 1.25 to 2.75, and from 0 to 1.25. This leads us to think that the 3 species have very different weights between each other.

Then we can see also a separated group of values in the width column, making us deduce that a specie might be thinner than the other 2. Head length and body length represent normal distributions, where we can not see any outliers or unexpected values.

Once the dataset is fully understood, we can get attention into each fish to see some values.

2.2.2 Gold Fish

For the goldfish the average body length is 4.98 cm, with a standard deviation of 0.33 cm. The mean head length is 3.40 cm, with a standard deviation of 0.38 cm. The average width is 1.47 cm, with a standard deviation of 0.17 cm. The mean weight is 3.25 grams, with a small standard deviation of 0.11 grams. The low standard deviations in these measurements suggest consistency in size. The median values for all four measurements are close to their respective means, suggesting a normal distribution. The range of measurements shows that most goldfish fall within a narrow size range. The goldfish maintain proportionate growth, indicating high uniformity in their physical characteristics.

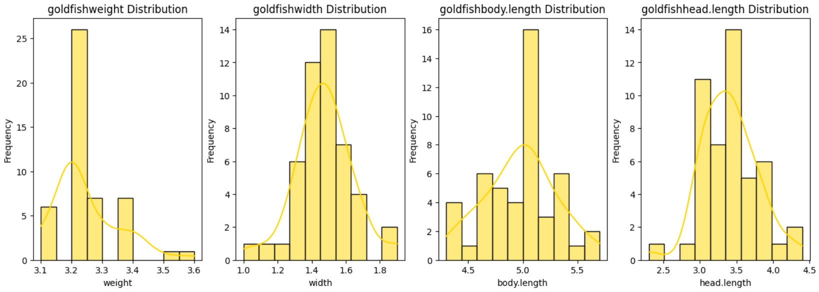


Figure 2: Gold fish Statistics

2.2.3 Island Mackerel

Then we can see the island mackerel. The average body length is 5.94 cm, with a standard deviation of 0.52 cm. The head length is 2.77 cm, with a smaller standard deviation of 0.31 cm. The width of the fish is 4.26 cm, with a standard deviation of 0.47 cm. The weight of the fish is 0.37 kg, with a higher relative standard deviation of 0.16 kg. The data suggests a relatively homogeneous population in terms of size, but with considerable diversity in weight, possibly influenced by factors such as diet, health, or environmental conditions. The fish's weight is the most variable characteristic, indicating differences in individual fish conditions or possibly varying age groups within the sample.

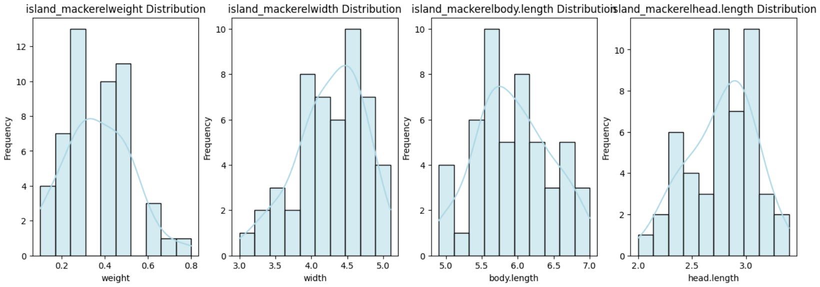


Figure 3: Island Makerel Statistics

2.2.4 Mugilidae

The data analysis of Mugilidae reveals low variability in body length, head length, width, and weight. The average body length is 4.98 units, head length and width are 3.40 and 1.47 units respectively, and weight is around 3.25 units. The interquartile range for body length, head length, width, and weight shows a level of uniformity within the Mugilidae population. Overall, Mugilidae exhibit consistent body metrics, indicating a stable and homogeneous population.

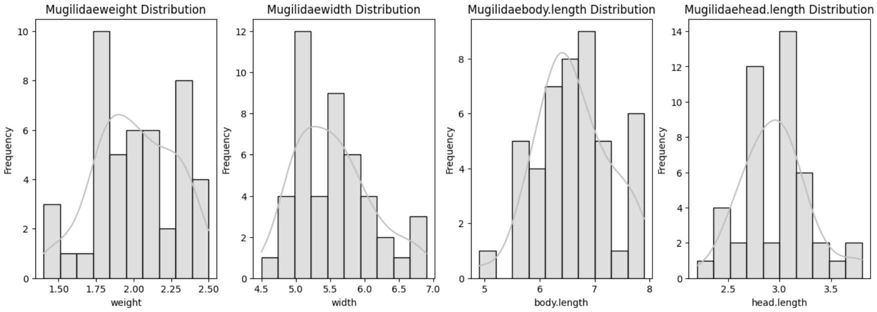
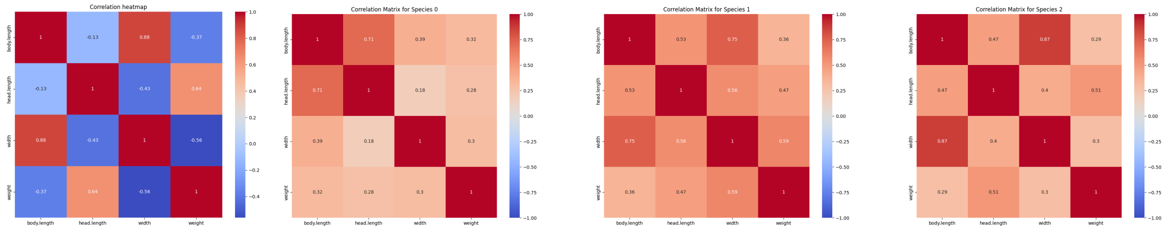


Figure 4: Mugilidae Statistics

2.3 Correlations

Firstly, we have the full correlation matrix, comparing the different columns. The only thing we can highlight on this correlation heatmap is the high relationship between body length and width, something that appears logical, as if a fish is longer, it will tend to be wider. On the contrary, we can see that it is quite impressive the low correlations weights have with the columns, as it is usually related to the size of the fish, so this makes us think that each species has a different relation between their values. Because of this, we can inspect each of the species and their internal correlations.

Although there were not any strong correlations between all four fields, there were two correlations we could draw from each species. For the goldfish species, the head length and body length have a value of 0.71, suggesting a proportional growth pattern compared to other parameters. In island mackerel, there are moderately positive connections between body length, width, head length, and weight, showing a degree of interdependence. Mugilidae species have moderate to significant connections between body length and width but small correlations with the other characteristics.



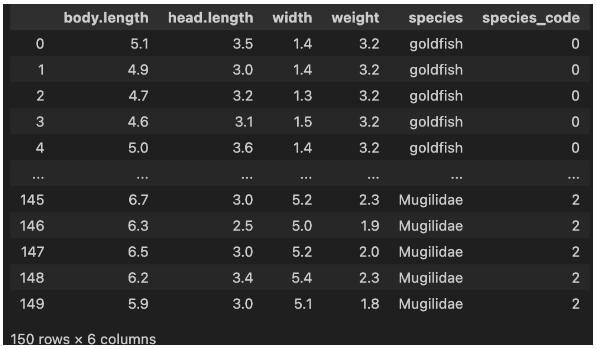
(a) Correlation of All Data (b) Correlation Goldfish (c) Correlation for Island Mackerel (d) Correlation for Mugilidae

Figure 5: Correlation MAtrices for each species

3 Data Pre-Processing

For the given data set, there were four parameters represented by a float value, however, it was clear that there are three types of species listed in the species column to categorize them. We have given each species a unique code to easily identify them using the numbers.

After this, we inspected the NaNs values. We can first see that the important column, assigning the species, has no empty values. Then all the other columns have 1 or 2 empty values that have to be investigated and corrected. As an easy solution, we have thought that filling those values with the mean of the species might be an appropriate and easily applicable value so that, if a goldfish does not have the height, the mean of the goldfish heights will be calculated and assigned so that it is representative.



	body.length	head.length	width	weight	species	species_code
0	5.1	3.5	1.4	3.2	goldfish	0
1	4.9	3.0	1.4	3.2	goldfish	0
2	4.7	3.2	1.3	3.2	goldfish	0
3	4.6	3.1	1.5	3.2	goldfish	0
4	5.0	3.6	1.4	3.2	goldfish	0
...	...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	Mugilidae	2
146	6.3	2.5	5.0	1.9	Mugilidae	2
147	6.5	3.0	5.2	2.0	Mugilidae	2
148	6.2	3.4	5.4	2.3	Mugilidae	2
149	5.9	3.0	5.1	1.8	Mugilidae	2

150 rows × 6 columns

Figure 6: Data frame of Processed Data

4 Clustering Algorithm - Methodology

We have chosen two types of clustering for the given dataset.

4.1 K-Means Clustering

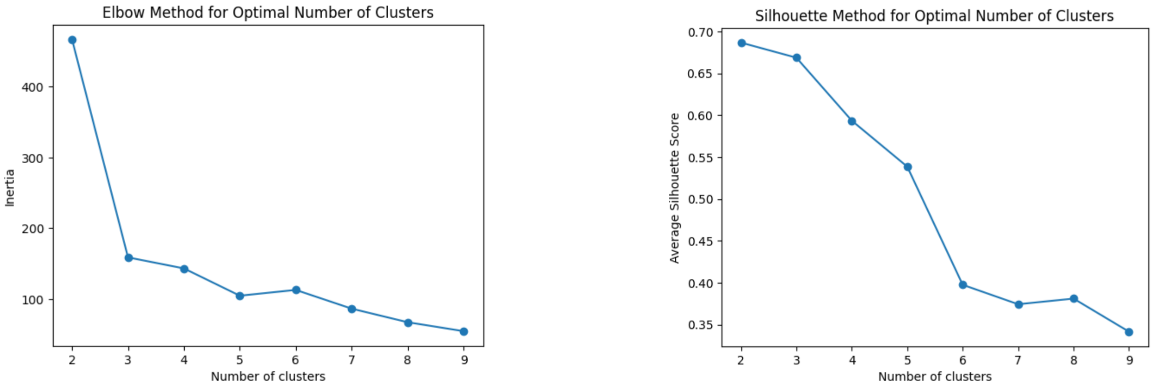
K-Means is an iterative clustering algorithm that partitions a dataset into K distinct, non-overlapping clusters. It involves initialization, assignment, update, and iteration. Initialization involves randomly selecting K centroids from data points. Assignments are made based on Euclidean distance, forming K clusters. Updates are calculated using the mean of data points in each cluster. Iterations are repeated until the centroids no longer change significantly or a predefined number of iterations are reached.

4.2 Hierarchical Clustering

Hierarchical clustering is a method that creates a hierarchy of clusters, either agglomerative (bottom-up) or divisive (top-down). The agglomerative approach involves initializing each data point as a single cluster, merging the closest pairs based on distance metrics, updating them to form a new cluster, and iterating until all data points are merged into a single cluster or a desired number of clusters is achieved. The resulting dendrogram represents the nested grouping of clusters and can be adjusted to obtain different cluster solutions.

4.3 K value

To search for the optimum k value we have used two approaches, the elbow method and the silhouette method.



(a) Elbow Method

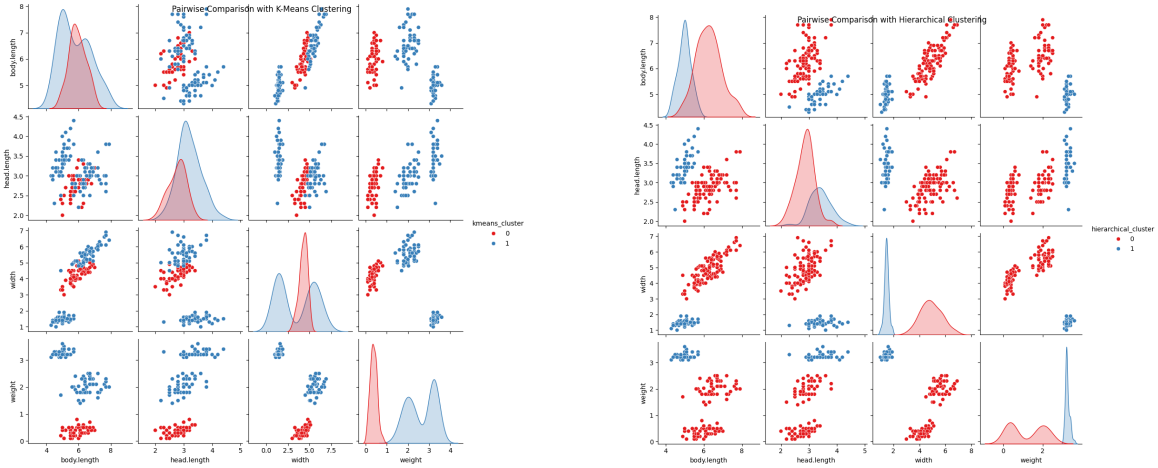
(b) silhouette method

Figure 7

As we can see, the elbow method indicates a clear value of 3 as the most optimum one to be tried. On the silhouette method, we can also see a high value in 2 clusters, and 4 has also a high value. As a result we will try 2,3 and 4 as k values for the evaluation and results.

5 Model Evaluation and Results

For 2 Group Clustering Following our two techniques, we were able to group data into 1 big group of 100 fish and a small group of 50. Looking into data, two species were grouped together and another one was set apart. This aggregation has been made differently between both algorithms but effectively.

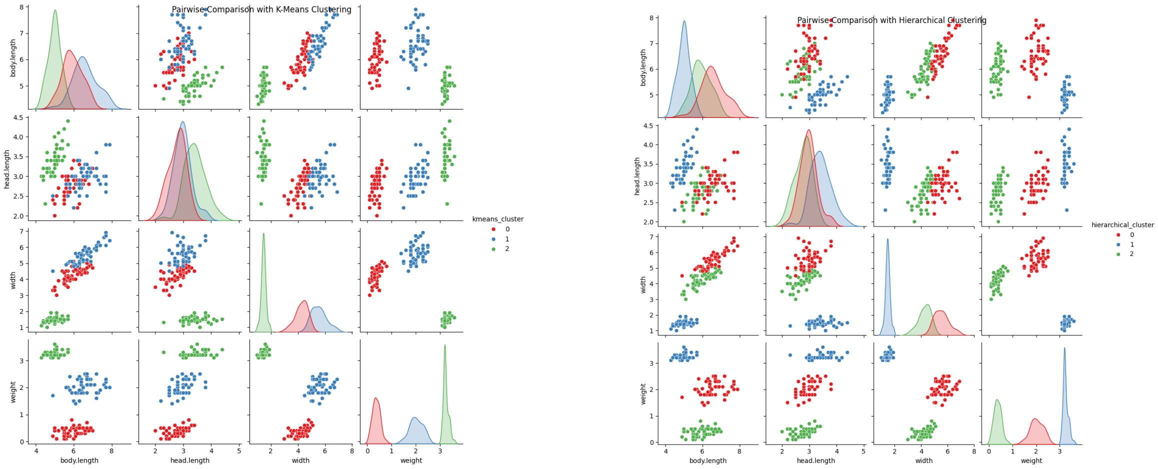


(a) K-Means Clustering

(b) Hierarchical Clustering

Figure 8: Pairwise Plots for Two Approaches with K = 2

For 3 Group Clustering Upon calculating value three, the two methods have divided each species into distinct groups, obtaining 3 clearly differentiated groups as seen in the graphics. The weight clearly separates values into 3 clusters, and the other columns also have differentiable results, only seeing superposition in the graph comparing body and head length.

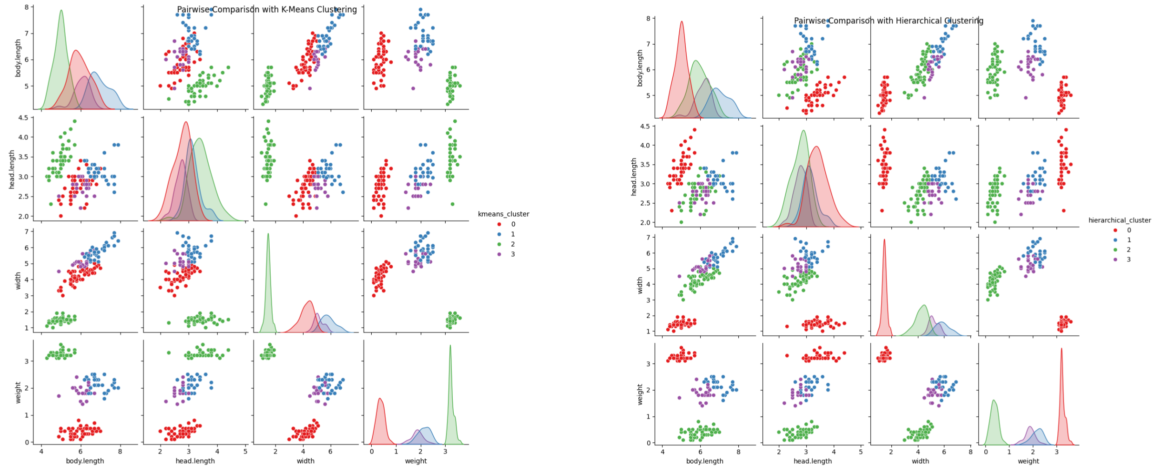


(a) K-Means Clustering

(b) Hierarchical Clustering

Figure 9: Pairwise Plots for Two Approaches with K = 3

For 4 Group Clustering With a k of 4, the two techniques have resulted in clearly dividing one of the 3 clusters obtained in the k=3 into 2 subgroups. Both algorithms took the Island Meckerel as the one that could be easily divided in 2. The other 2 species are set apart in different groups.



(a) K-Means Clustering

(b) Hierarchical Clustering

Figure 10: Pairwise Plots for Two Approaches with K = 4

After analyzing the results, we have concluded that the most accurate and appropriate number of clusters is 3, as the groups are clearly differentiated. This makes sense as there are 3 species, however, it could have easily been a different value. Talking about the most accurate algorithm, we have found that both work similarly, not having big difference in the results.

We think it has been a very interesting project, where we have learn how to analyze data, deal with categorical values, treat NaNs and investigate the most optimal value of clusters in order to use different clustering algorithms, to finally analyze the results.