Data Mining: Project 4

Akash Shanmugam

11/20/24

Dataset link:

https://www.kaggle.com/datasets/taweilo/fish-species-sampling-weight-and-height-data/data

Data Introduction:

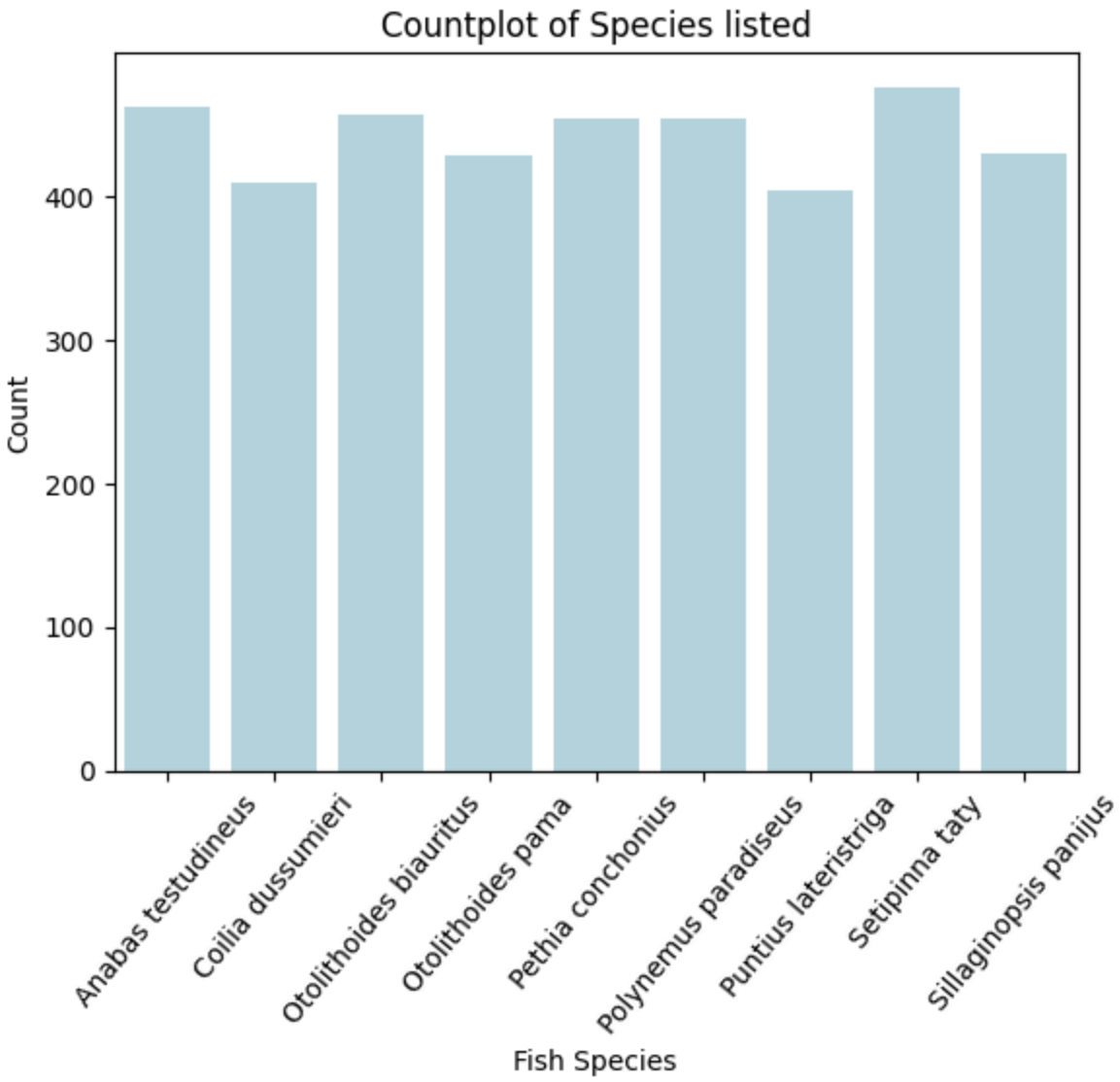
For the purposes of this clustering project I specifically looked up datasets on Kaggle that were based on clustered data that could represent the different clustering algorithms clearly. The dataset I ended up using was the “Fish sampling dataset” which contains the length and width of multiple different recorded fish that fit in the category of 9 different fish species. The dataset is overall very simple in terms of features with only length, width, and the ratio for each fish with 4080 rows of data overall.

Pre-Processing the Data:

I was able to check the data for any null values and quickly found that there were no major issues with the data by itself. The next thing I checked for was duplicated values and saw that the data had a sum of 109 duplicated fish values and dropped them for the final data. I then checked the

distribution of the species to make sure there were no major outliers where a species of fish could have too many or too little fish entries and affect how the clusters are predicted through the algorithm.

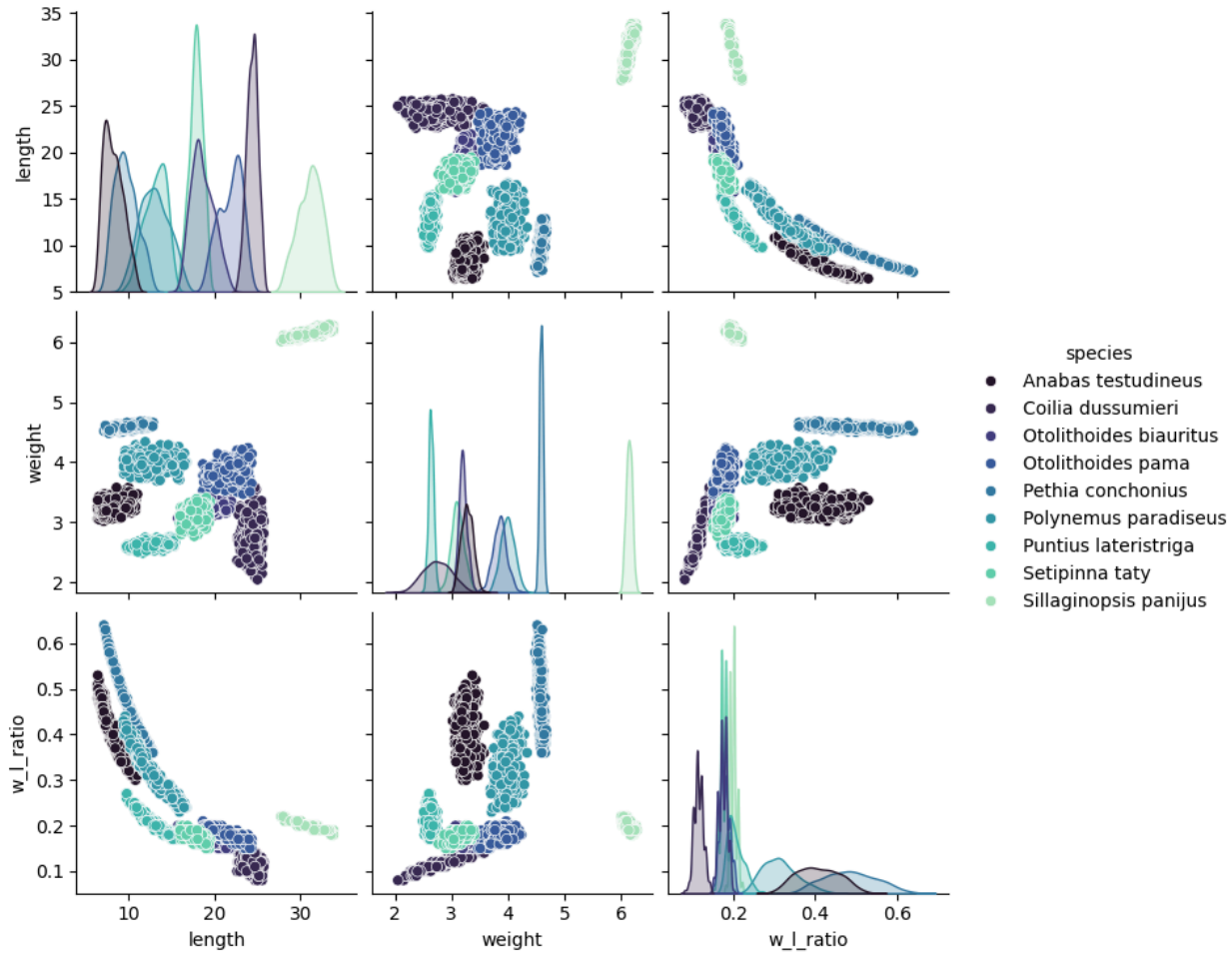
```
[4]: df.isna().sum()

[4]: species      0
     length       0
     weight       0
     w_l_ratio    0
     dtype: int64

[5]: print(df.shape)

     (4080, 4)

[6]: df.duplicated().sum()

[6]: np.int64(109)

[7]: df=df.drop_duplicates()

[8]: df.info()

     <class 'pandas.core.frame.DataFrame'>
     Index: 3971 entries, 0 to 4079
     Data columns (total 4 columns):
      #   Column     Non-Null Count  Dtype
     ---  ------     --------------  -----
      0   species    3971 non-null   object
      1   length     3971 non-null   float64
      2   weight     3971 non-null   float64
      3   w_l_ratio  3971 non-null   float64
     dtypes: float64(3), object(1)
     memory usage: 155.1+ KB
```

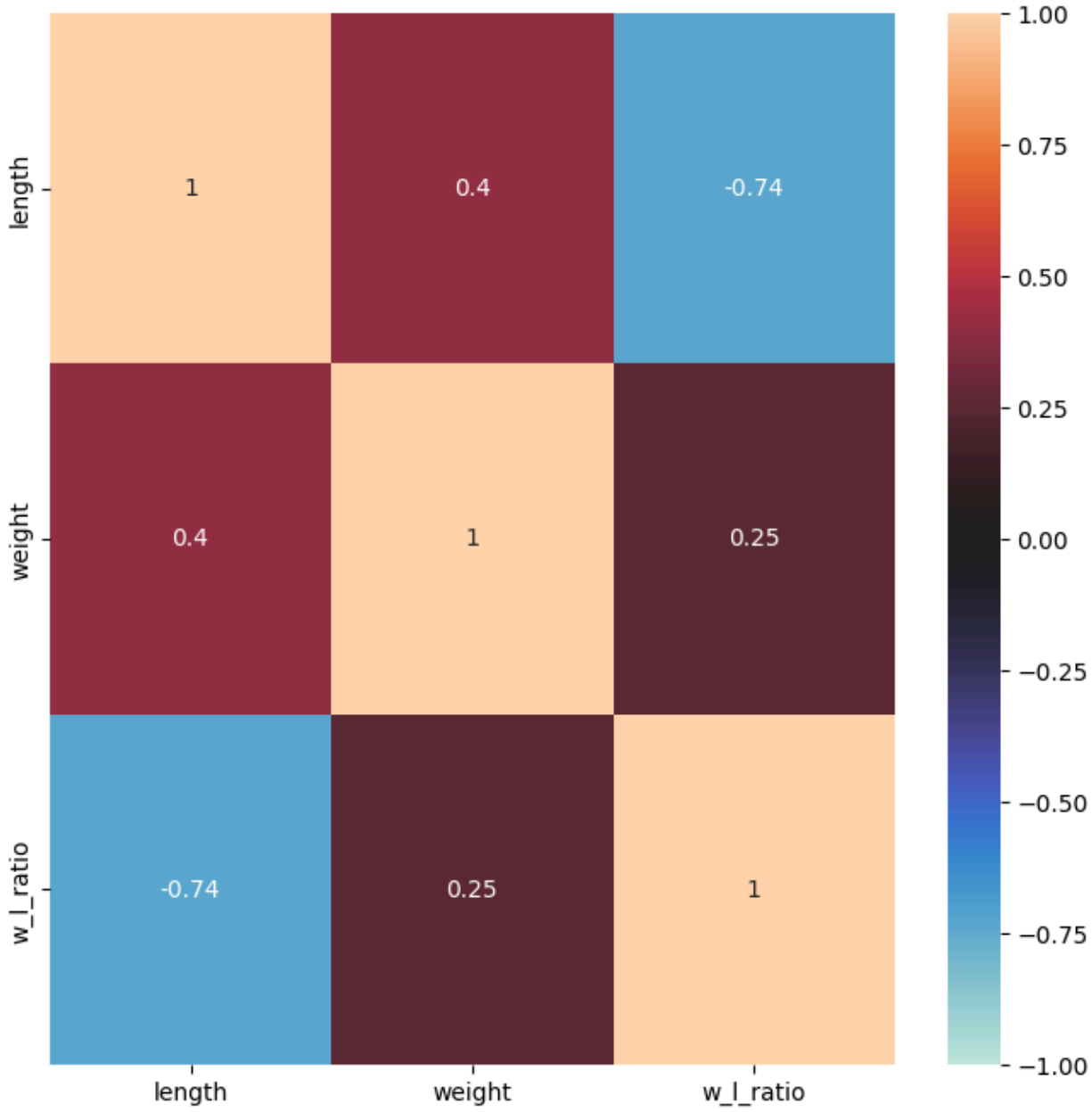


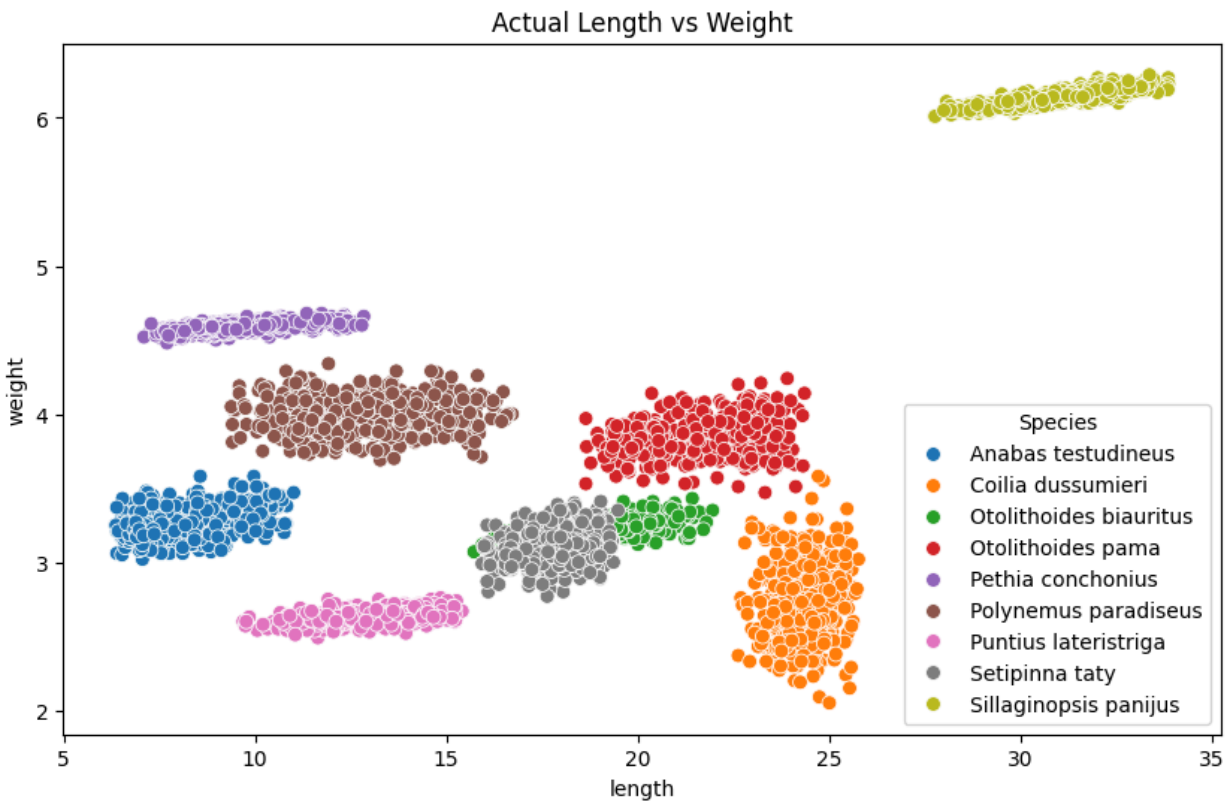
Data Understanding:

The immediate thing that was needed to check the general distribution of data was a simple pairplot that I have used for all other prior projects for a full showcase of the data based on species. From this, there was an easily identifiable group in the weight by length plot for each species which was a good sign and showed that the clusters would be able to be distinguished when it came time to implement the clustering algorithms. Another interesting point was the

weight per length ratio mostly staying around the 0.2 mark for the data in the fish recordings. To elaborate more on the data, I looked at the correlation matrix chart to make sure there was no hidden correlation between the features that existed in the data and found nothing beyond what was already expected for the height and weight.

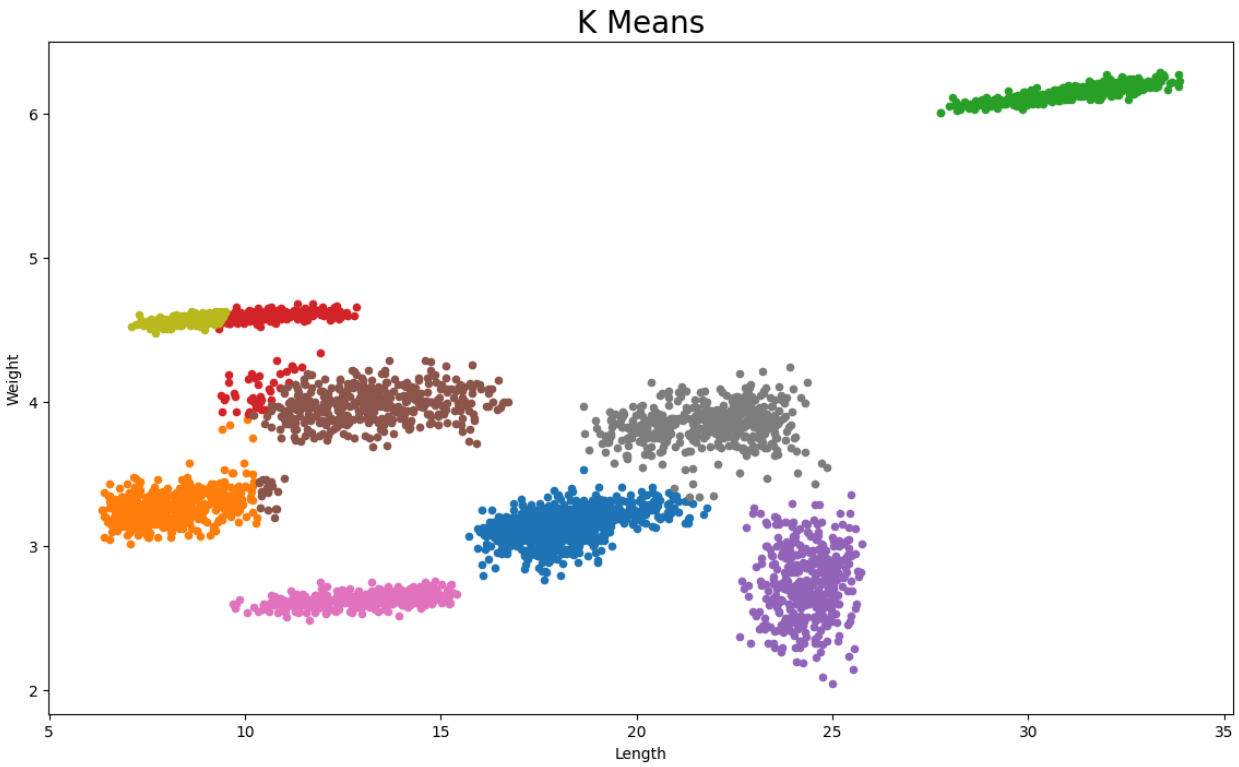



length
weight
w_l_ratio
1
0.4
-0.74
0.4
1
0.25
-0.74
0.25
1
length
weight
w_l_ratio
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00



Experiment 1: KMeans

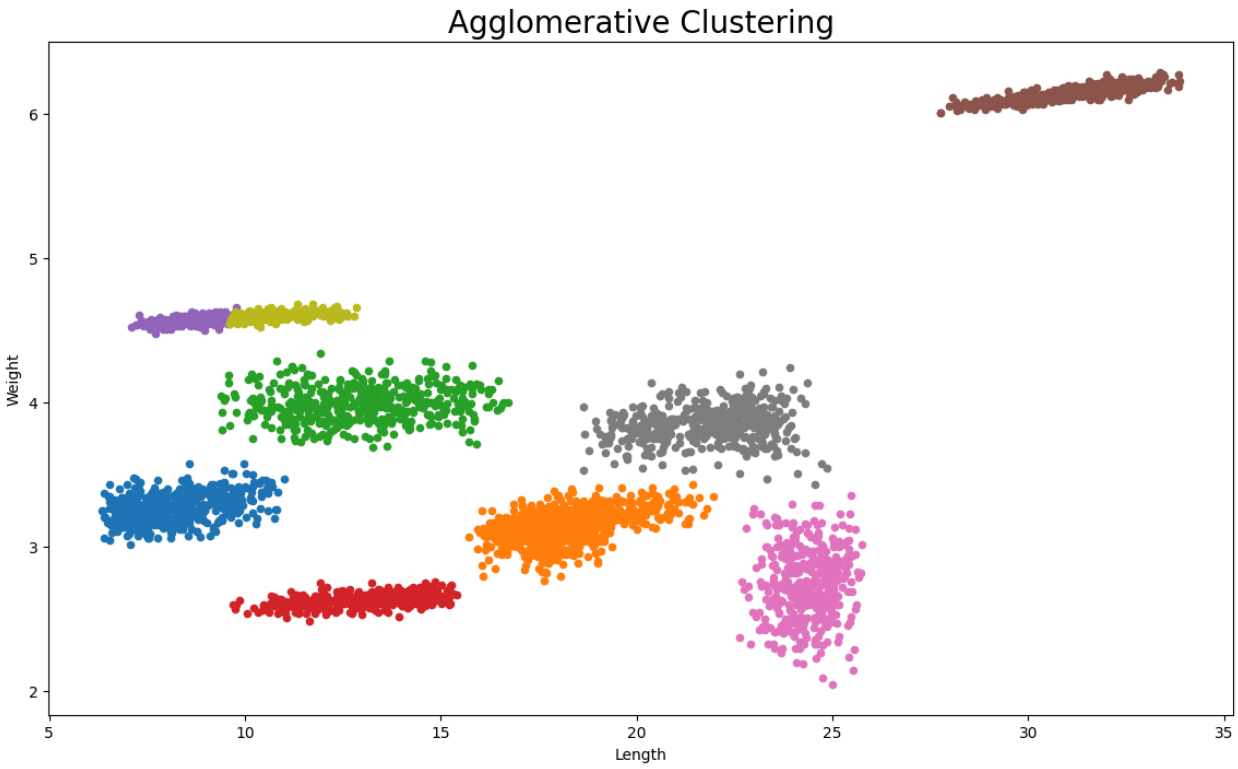
For my first clustering model I chose to use K-means as it is general purpose and seems to fit this type of dataset by having most of the data already in clear clusters that fit around the K-means type of categorization. By first fitting the data and scaling it, the K-means algorithm was able to get a Rand Index score of 0.809 and had a final predicted data plot that closely resembled the actual graph of the data and had the clusters in the same areas that they are distinguished by in the scatter representation.



Experiment 2: Agglomerative Clustering

The next type of data clustering model I used was agglomerative clustering as it is the next best general purpose model and specifically fits the types of datasets that contain large amounts of data as well as reducing the amount of variance overall. Utilizing the model from Scikit as well as setting the appropriate amount of clusters, the Rand Index score that was achieved from this data was 0.839 which is a marked improvement over the previous K-means model. This better accuracy is reflected in the graphical representation and the delineation of the different groups is much more accurate with less oddly placed groups especially on the left side of the

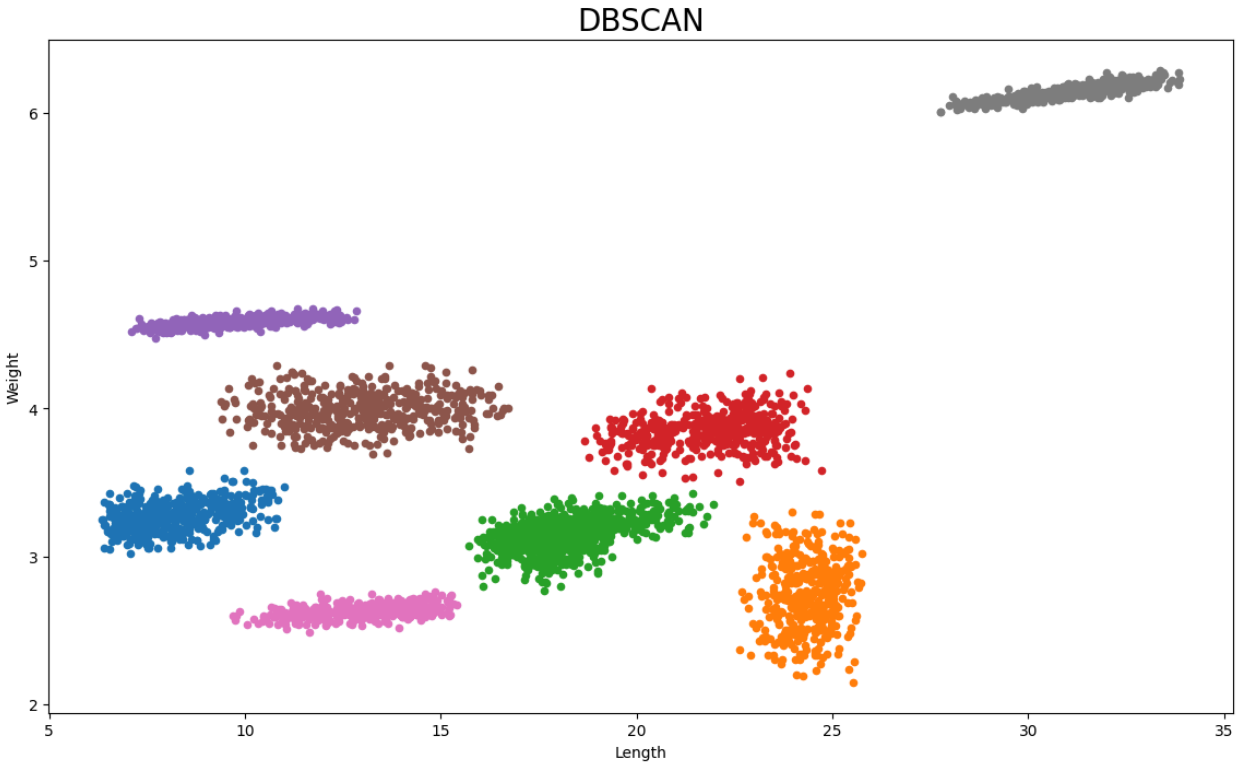
representation with lower length. Both of the models made also have failed to distinguish any different between the Otolithoides biauritus and the Setipinna taty species likely as they overlap but also split the Pethia conchonius into two different groups.



Experiment 3: DBSCAN

The third and final type of model I used for this experiment was from the Scikit-learn list of clustering models and was DBSCAN. This model is good at handling uneven shapes of clusters

and non-flat geometry which made sense to fix the specific problem of the clusters such as Pethia conchonius which had very varying length but extremely similar weight while also being very close geometrically on the graph to other clusters. Despite these factors, the Rand Index score after all of the parameter tweaking ended up at around 0.870 which was at the very least in the same ballpark as the rest of the models. The scatterplot however, managed to show that DBSCAN overcame the issue with splitting up elongated clusters and had a clean output of all the species (likely due to having no specific cluster number) despite not having the ability to separate the overlapping clusters.



Impact:

For this project, the impact can be summarized as pretty straightforward with the focus on identifying different species of fish. This has plenty of practical applications such as being able to quickly identify and manage different species of fish based on weight and length alone which could be used for something like quickly sorting large quantities of various fish during factory processing or similar situations.

References:

https://www.kaggle.com/datasets/taweilo/fish-species-sampling-weight-and-height-data/data