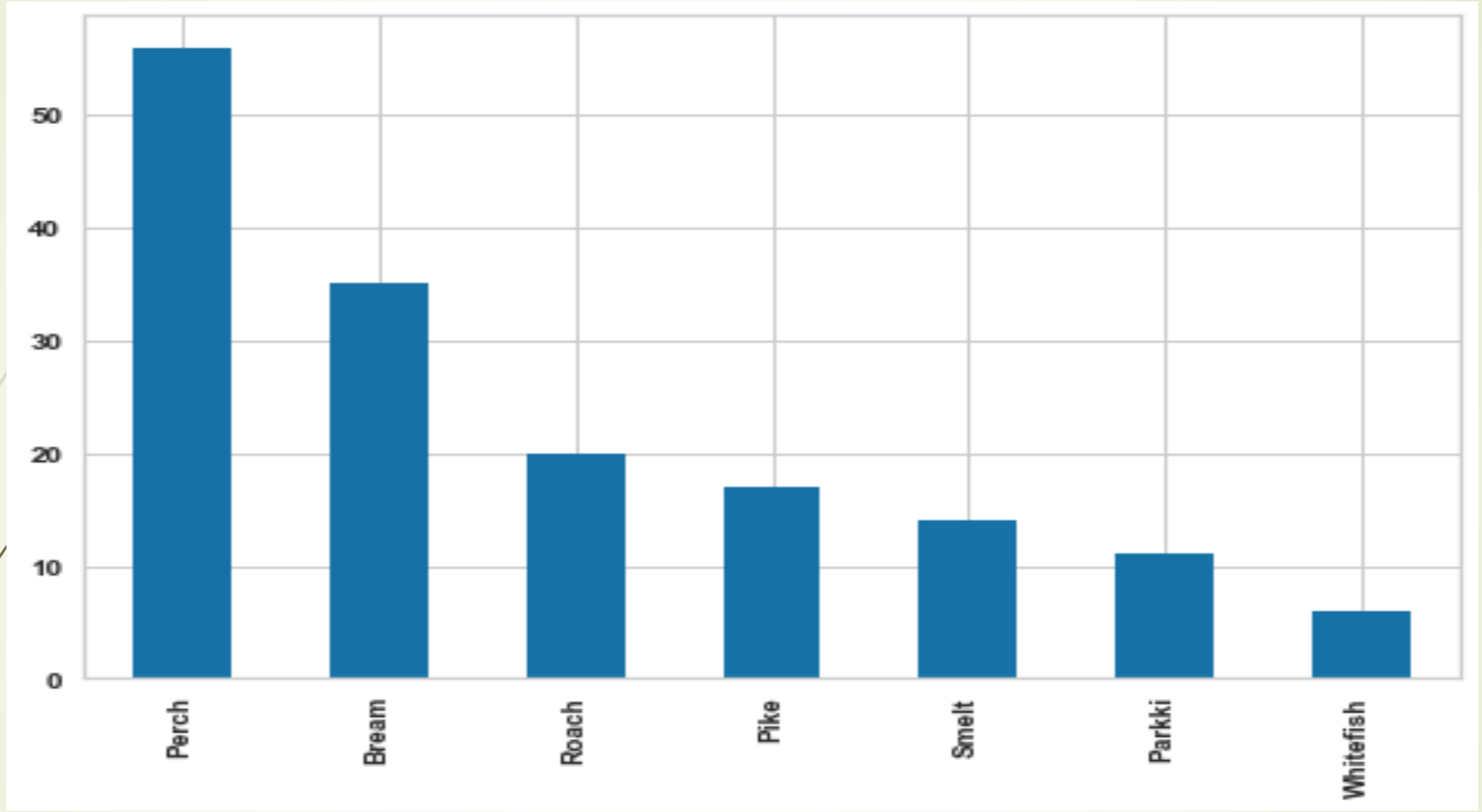
Unsupervised Capstone

by **Krishna Devabhaktuni**¶

https://www.kaggle.com/aungpyaeap/fish-market

Is it possible to classify fish species based on their physical characteristics?

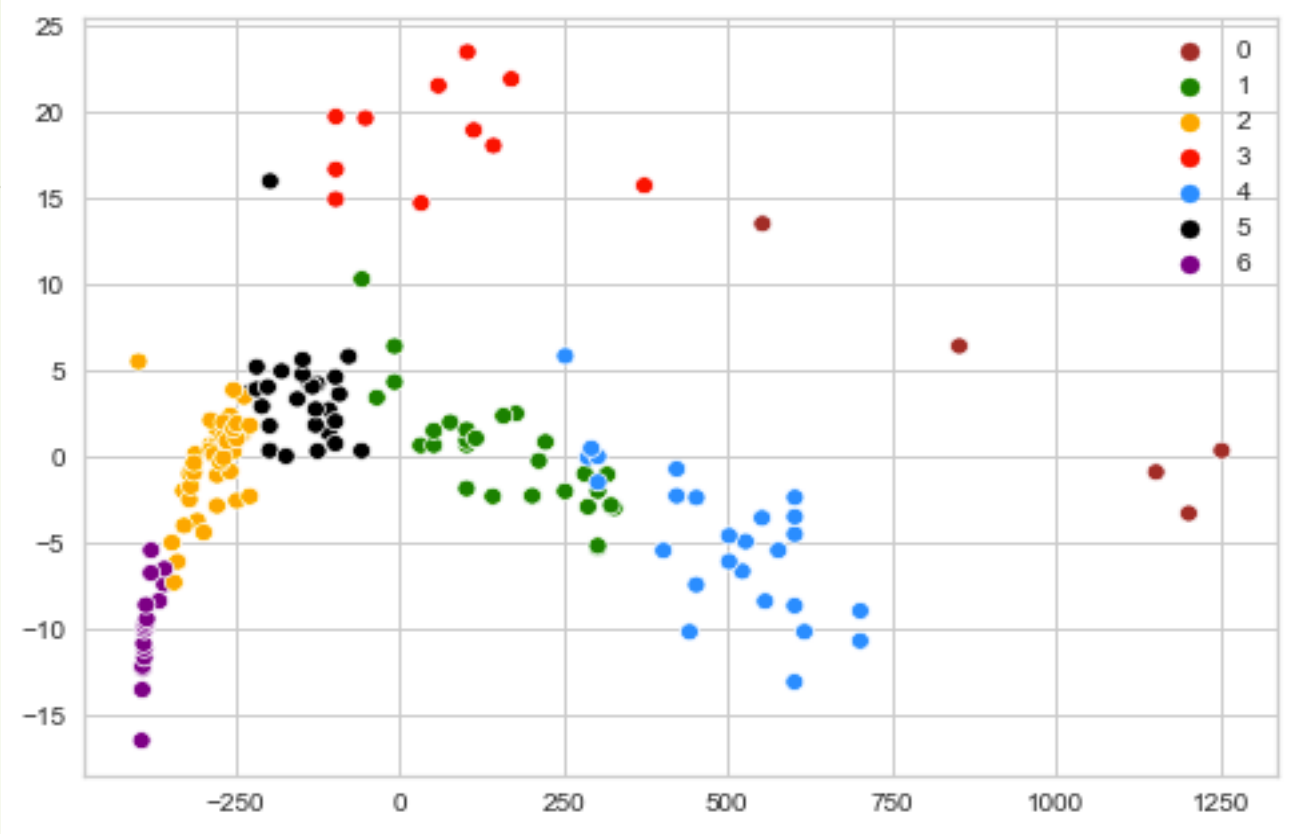
1	Species	Weight	Length1	Length2	Length3	Height	Width
2	Bream	242	23.2	25.4	30	11.52	4.02
3	Bream	290	24	26.3	31.2	12.48	4.30
4	Bream	340	23.9	26.5	31.1	12.38	4.69
5	Bream	363	26.3	29	33.5	12.73	4.45



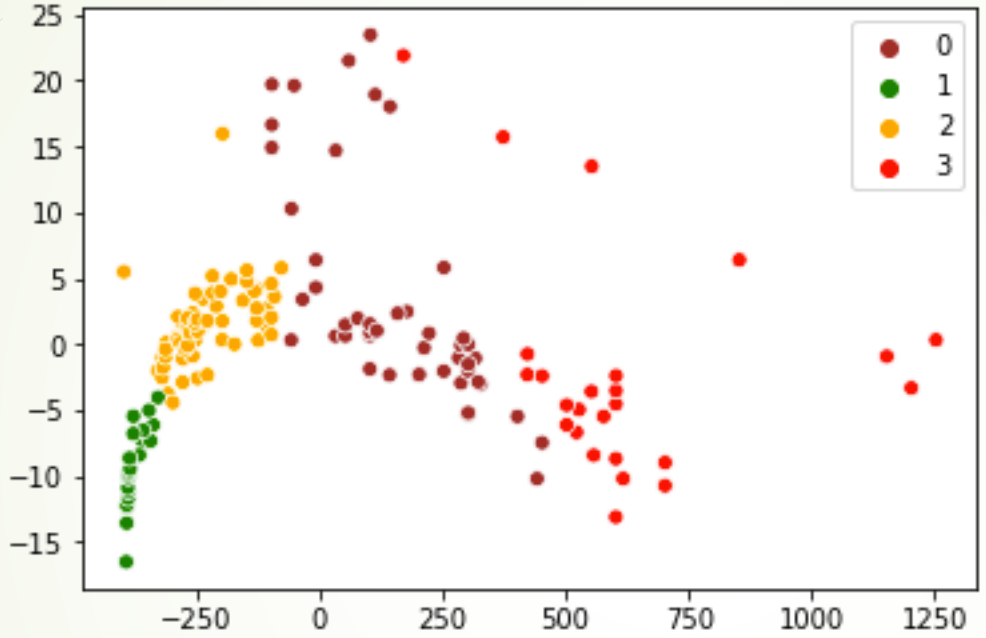
FISH SPECIES

DATA DICTIONARY

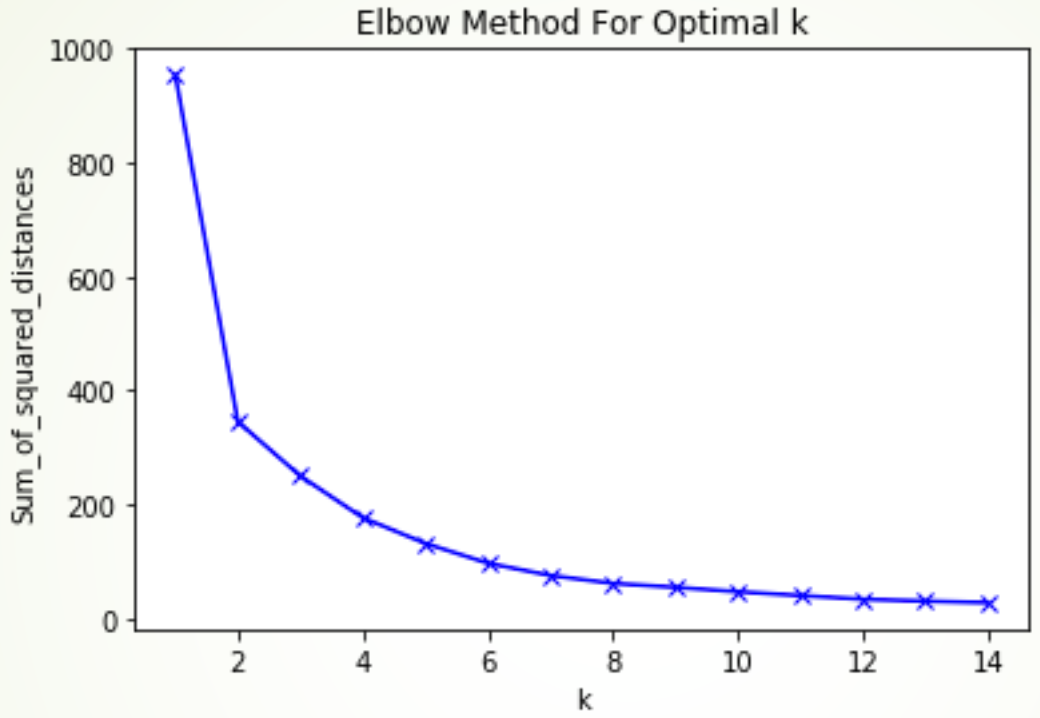
Species	:	Species name of a fish
Weight	:	Weight of a fish in Grams
Length1	:	Vertical length in CM
Length2	:	Diagonal length in CM
Length3	:	Cross length in CM
Height	:	Height in CM
Width	:	Diagonal width in CM



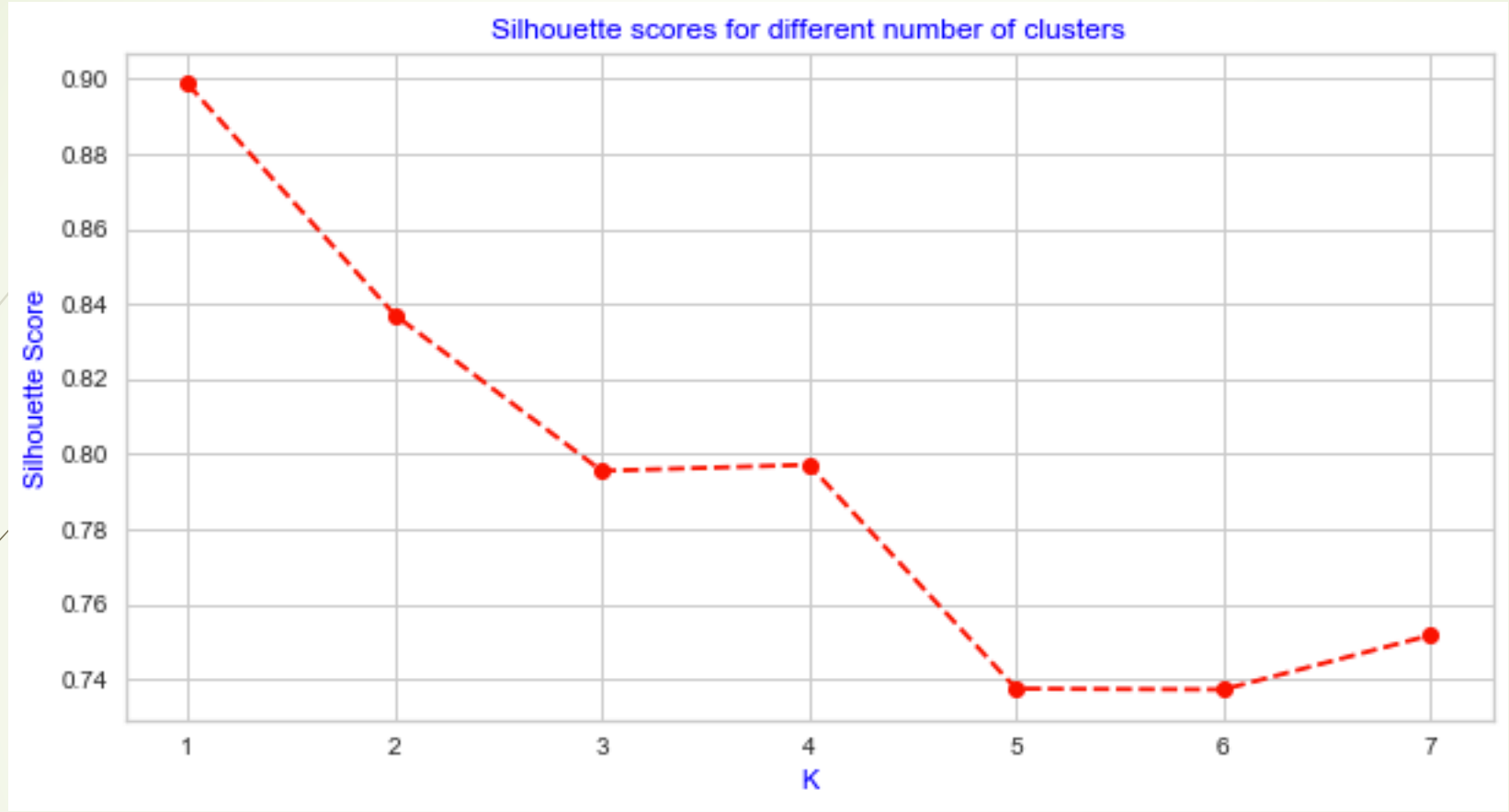
Adjusted Rand Index of the Kmeans Solution for 7 clusters is 0.27
Silhouette Score of the Kmeans Solution for 7 clusters is 0.37



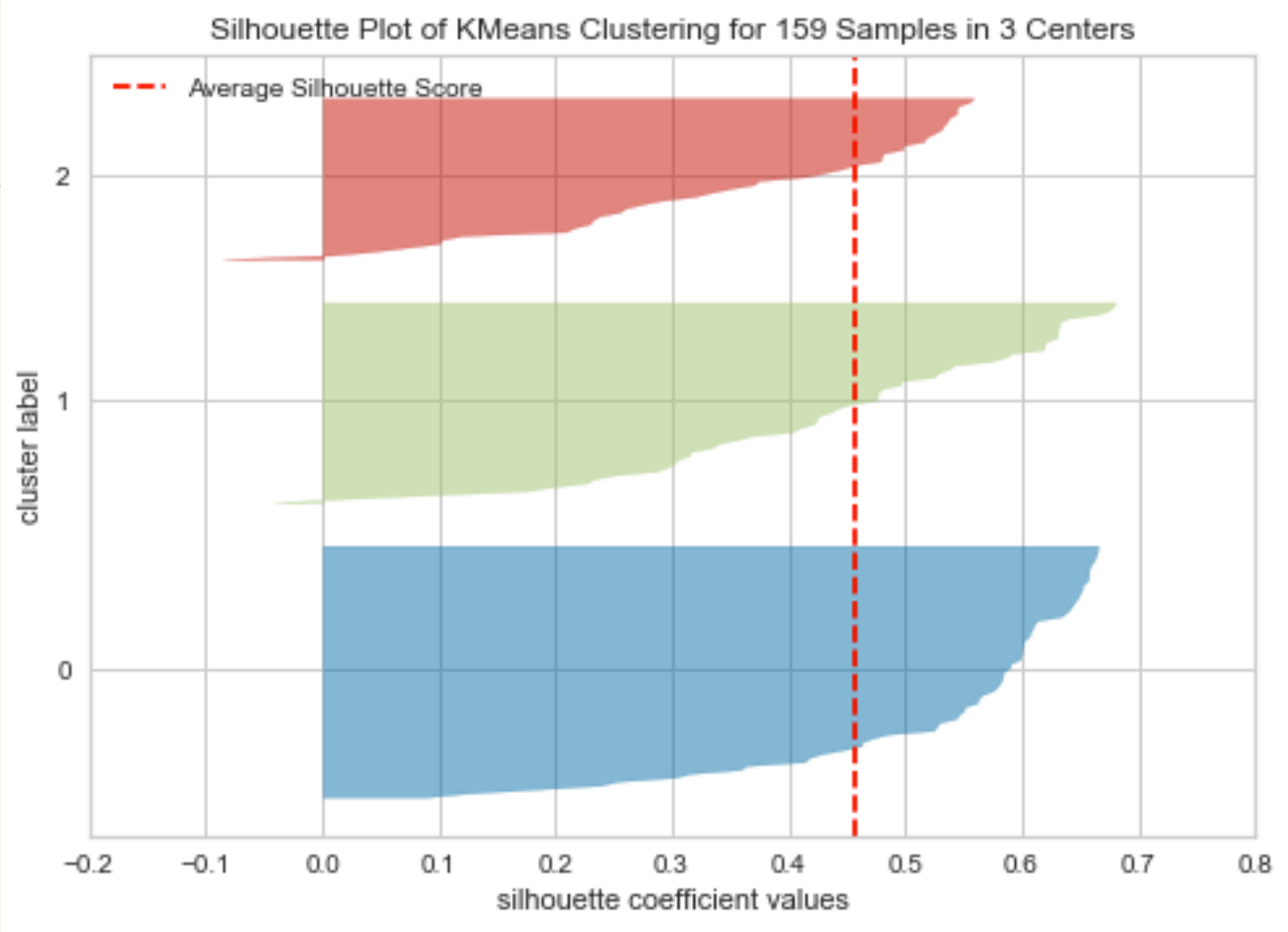
Adjusted Rand Index of the Kmeans solution for 4 clusters: 0.2209785047386537
The silhouette score of the KMeans solution for 4 clusters: 0.4404678727453739 7

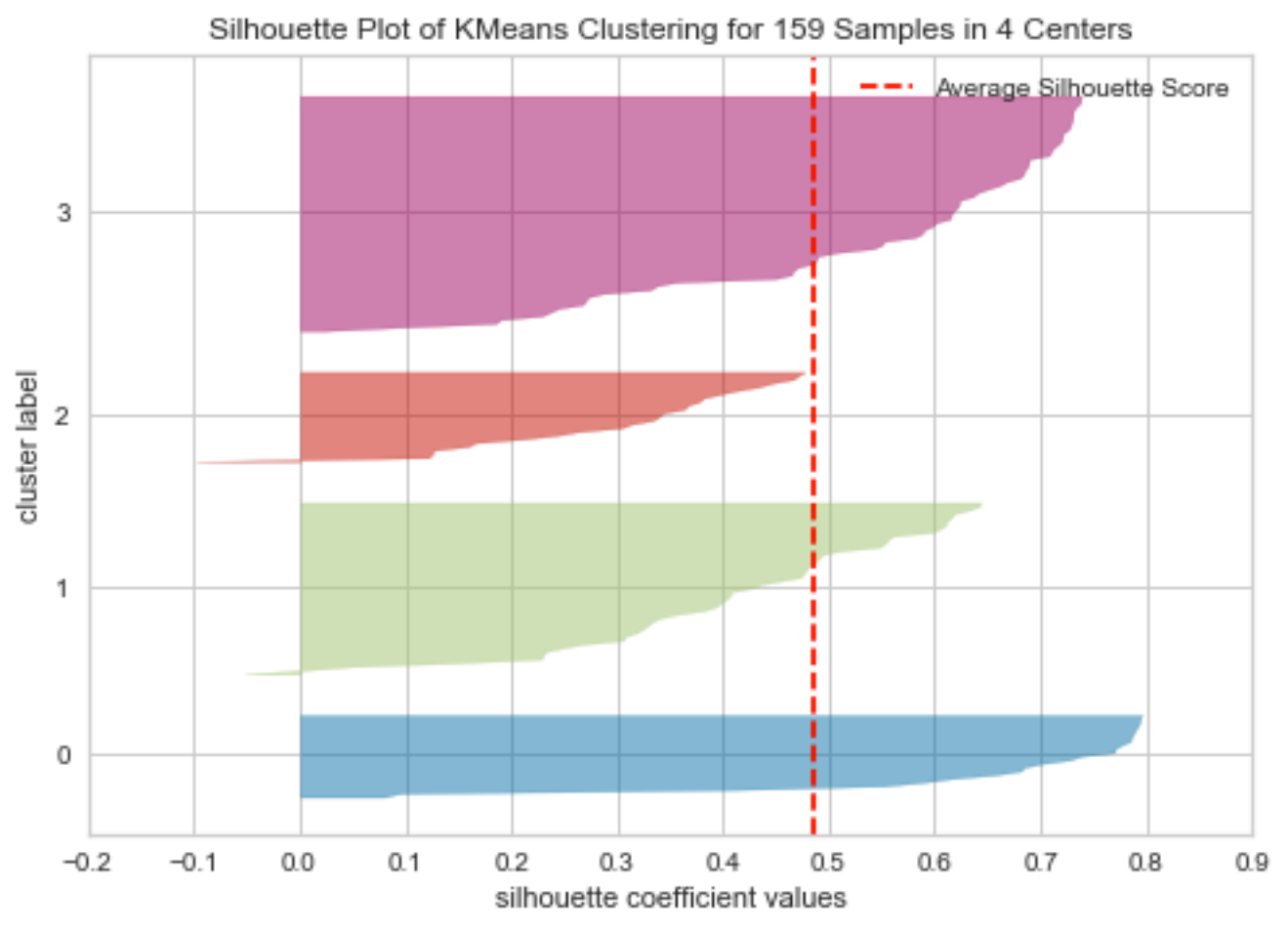


From the Elbow plot we can see that clusters beyond 4 have little value

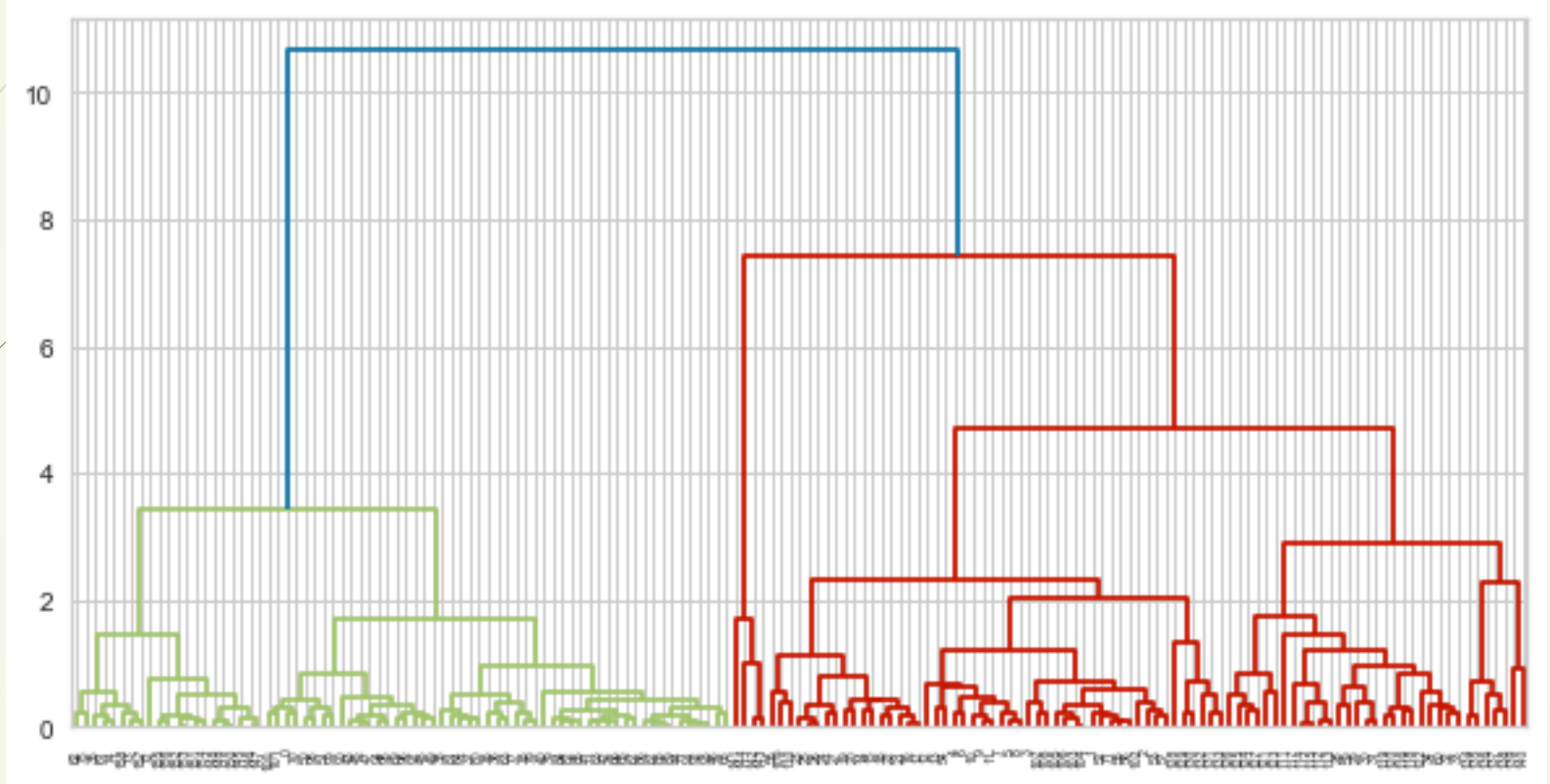


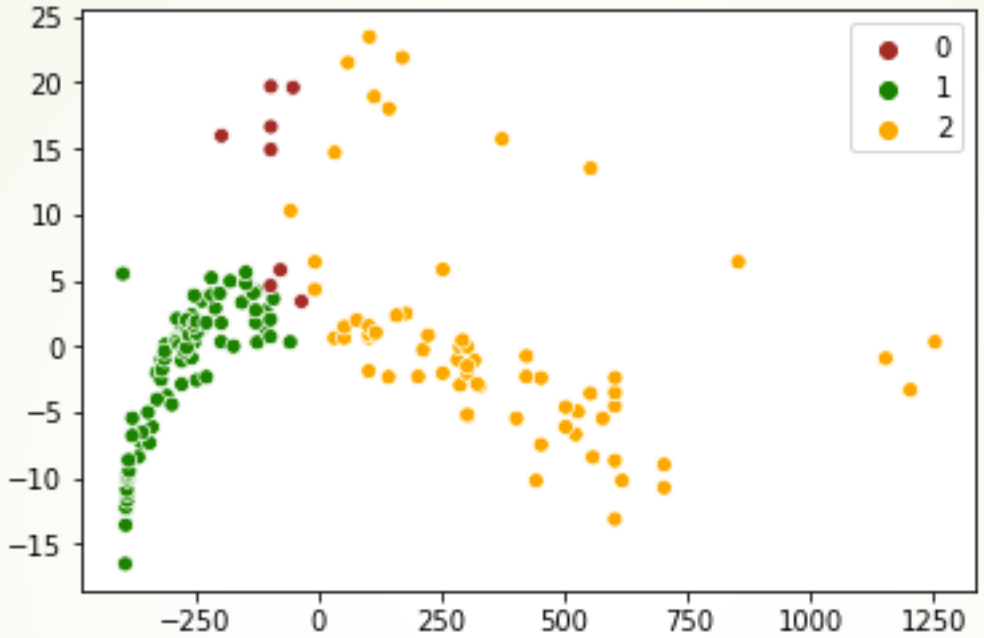
Silhouette scores after 4 clusters have


Silhouette Plot of KMeans Clustering for 159 Samples in 3 Centers
Average Silhouette Score
cluster label
2
1
0
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
silhouette coefficient values



We can observe cluster imbalance with 4 clusters and 3 clusters seems to be an ideal choice


10
8
6
4
2
0



Adjusted Rand Index of the Hierarchical cosine solution: 0.145
The silhouette score of the Hierarchical cosine solution: 0.837

Conclusion:

From the plots above, we can see that the clusters beyond 3 have little value and we can also see that the different type of fish in our dataset have high similarities in their physical characteristics and due to those similarities we are not able to differentiate all the seven species.

Next Steps:

Collect more data: We need to collect more data because we only have few rows of data for example whitefish has only 6 rows of data.