Bioinformatic Analysis

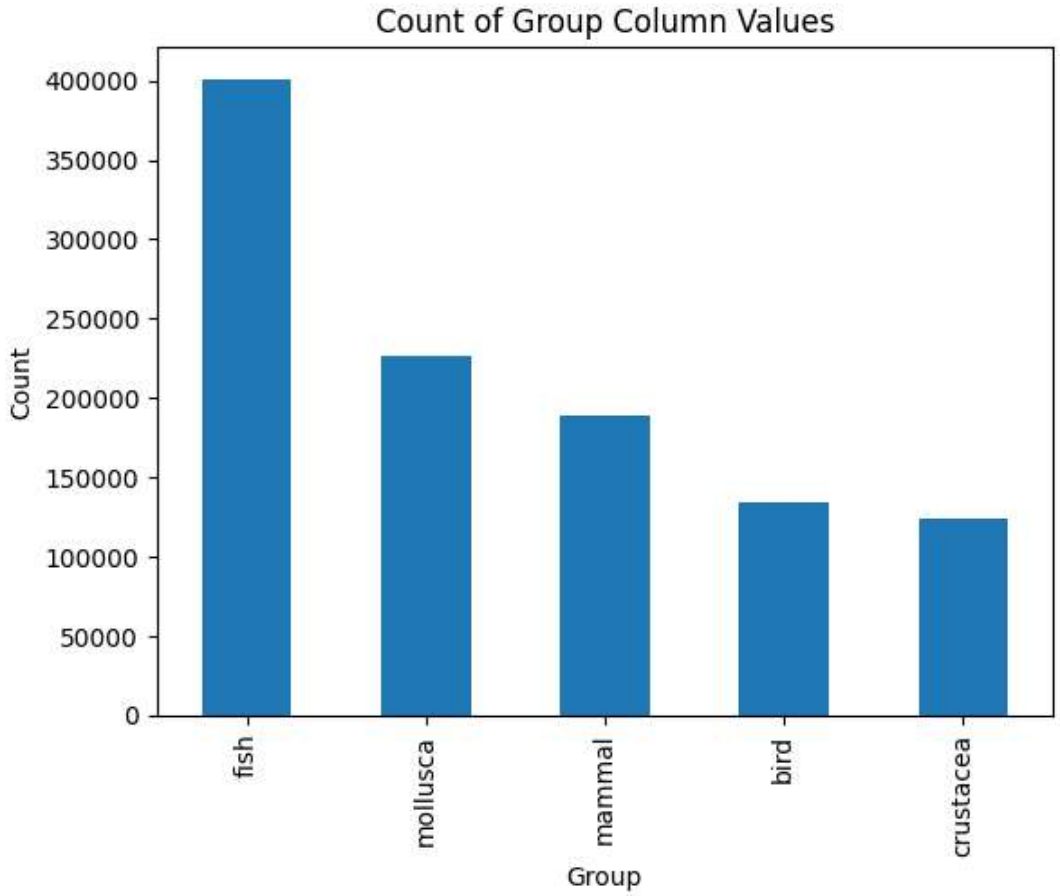
By

Shivam Goyal

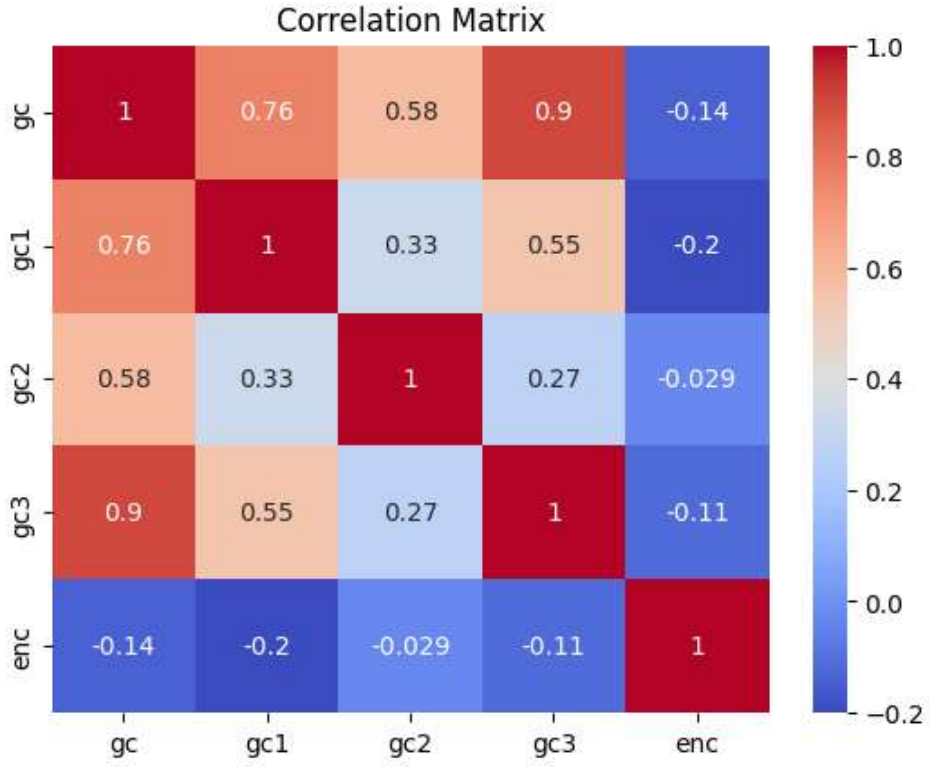
Introduction:

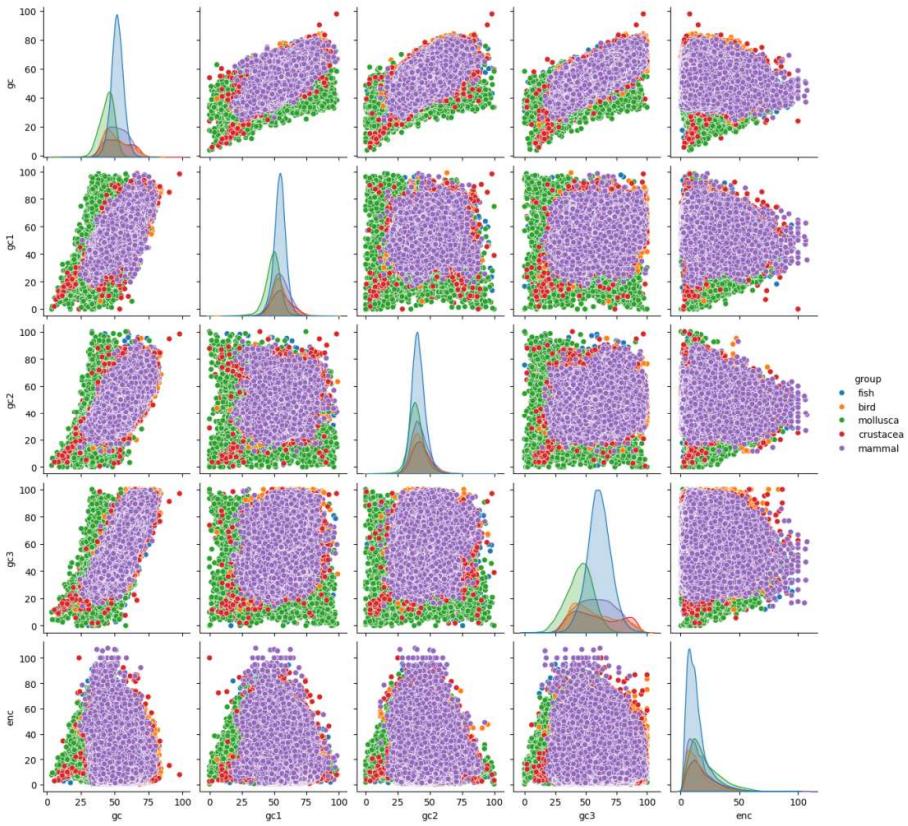
Data:

I collected data for 25 different species from the ensemble.org, categorizing them into five groups. My methodology involved comparing data points from various groups, which revealed significant differences. The fish group had the maximum number of data points, whereas the crustacean group had the fewest.



Performed a thorough examination of feature correlations, demonstrating a constant positive connection across all features. Notably, the 'enc' feature had a weak negative connection with other features.


Correlation Matrix
gc
gc1
gc2
gc3
enc
1
0.76
0.58
0.9
-0.14
0.76
1
0.33
0.55
-0.2
0.58
0.33
1
0.27
-0.029
0.9
0.55
0.27
1
-0.11
-0.14
-0.2
-0.029
-0.11
1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


group
fish
bird
mollusca
crustacea
mammal

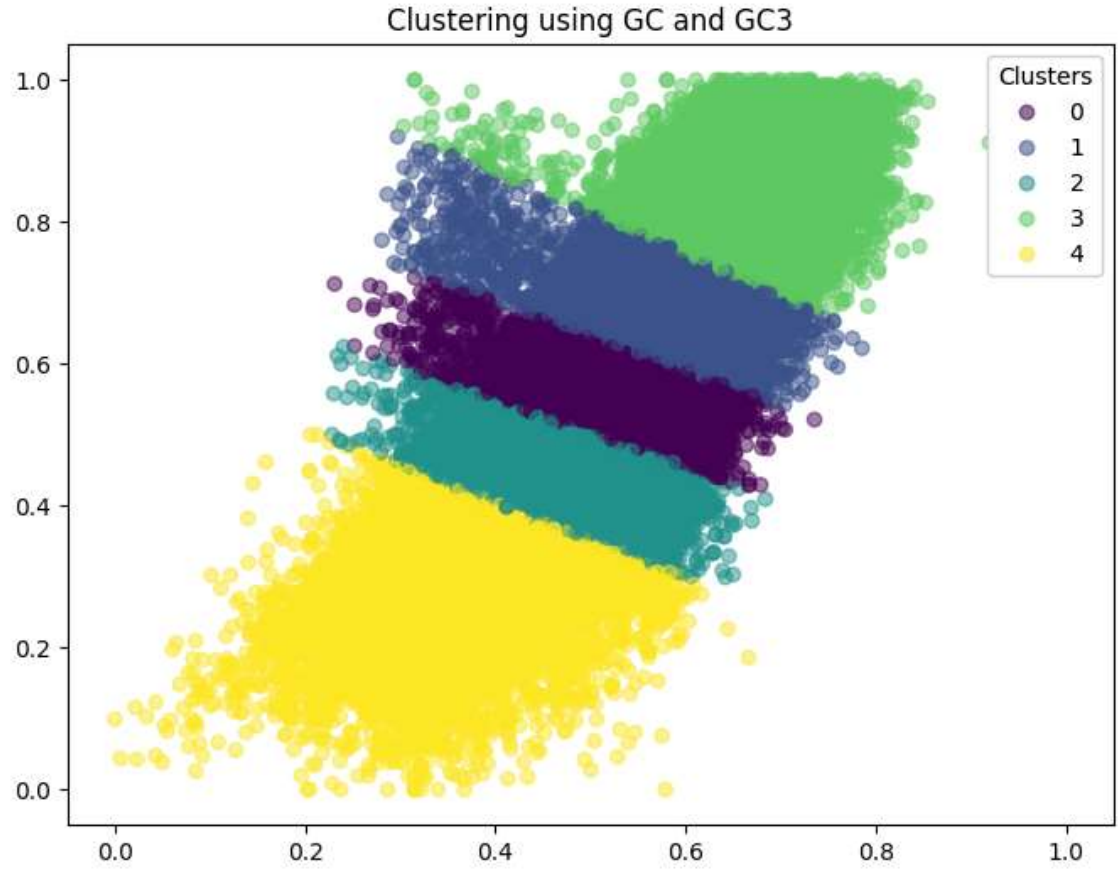
Procedure:

I started the investigation by standardizing the dataset using min-max scaling. Subsequently, we investigated other feature combinations, using various clustering techniques such as K-Means and Gaussian Mixture, to reveal distinct patterns in the data.

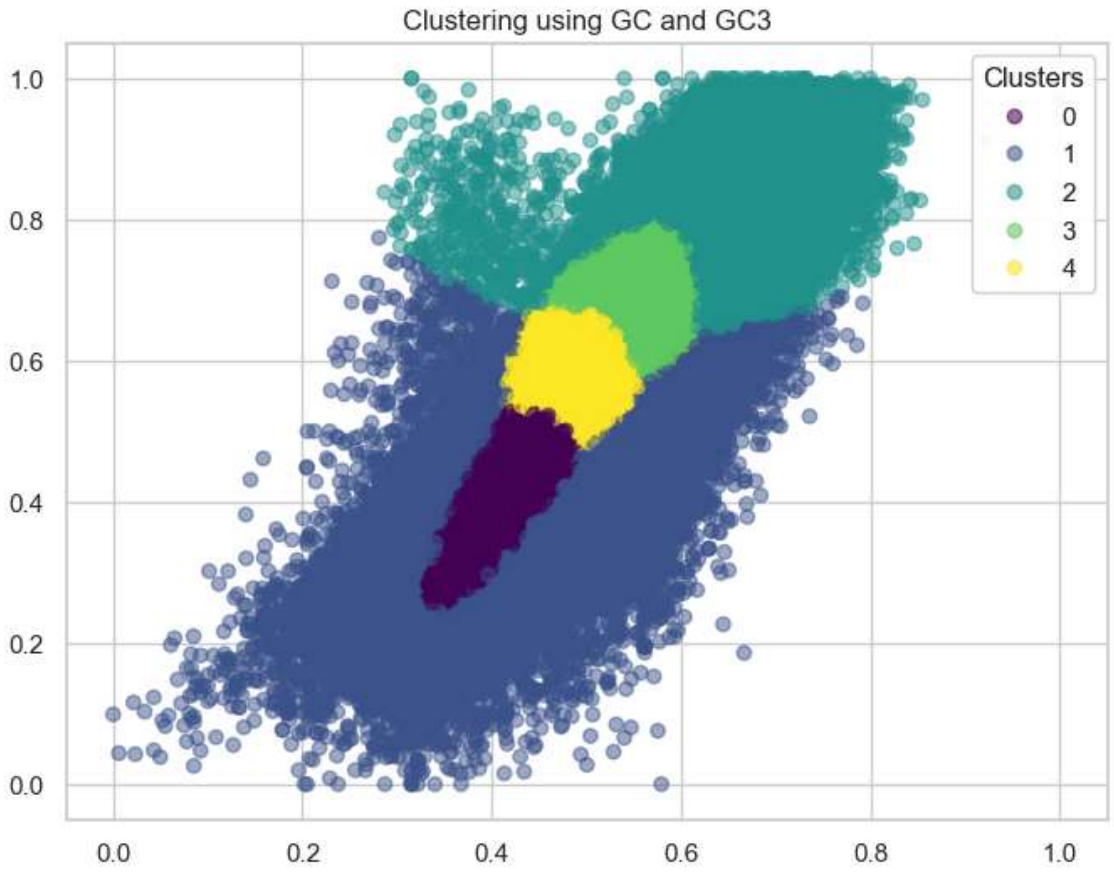
Clustering:

GC and GC3:

I used the K-Means clustering approach to create five separate clusters based on the attributes GC and GC3.



K-Means clustering efficiently established cluster boundaries by using GC and GC3 as features. However, when compared to the scatterplots of GC and GC3 based on the categories, there were inconsistencies in cluster patterns. As a result, for each group's data, I constructed five centroids and examined their placements on the graph.



Unlike the K-Means the pattern of the Gaussian Mixture clusters is elliptical cluster.

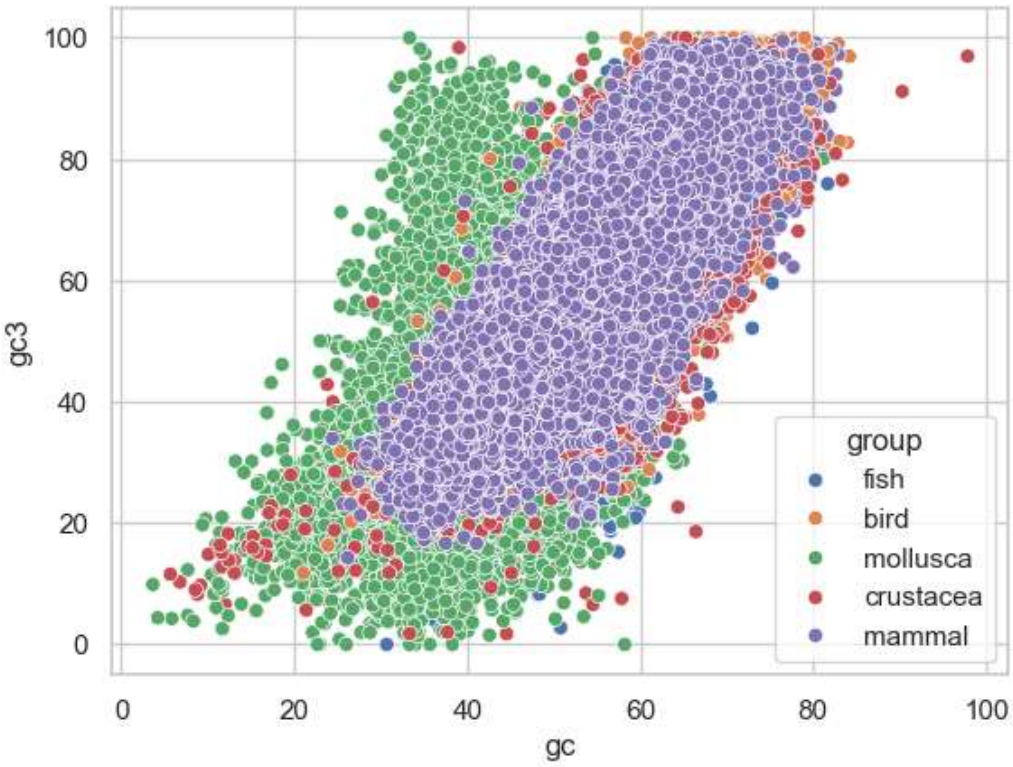


Fig: Scatterplot of the gc and gc3

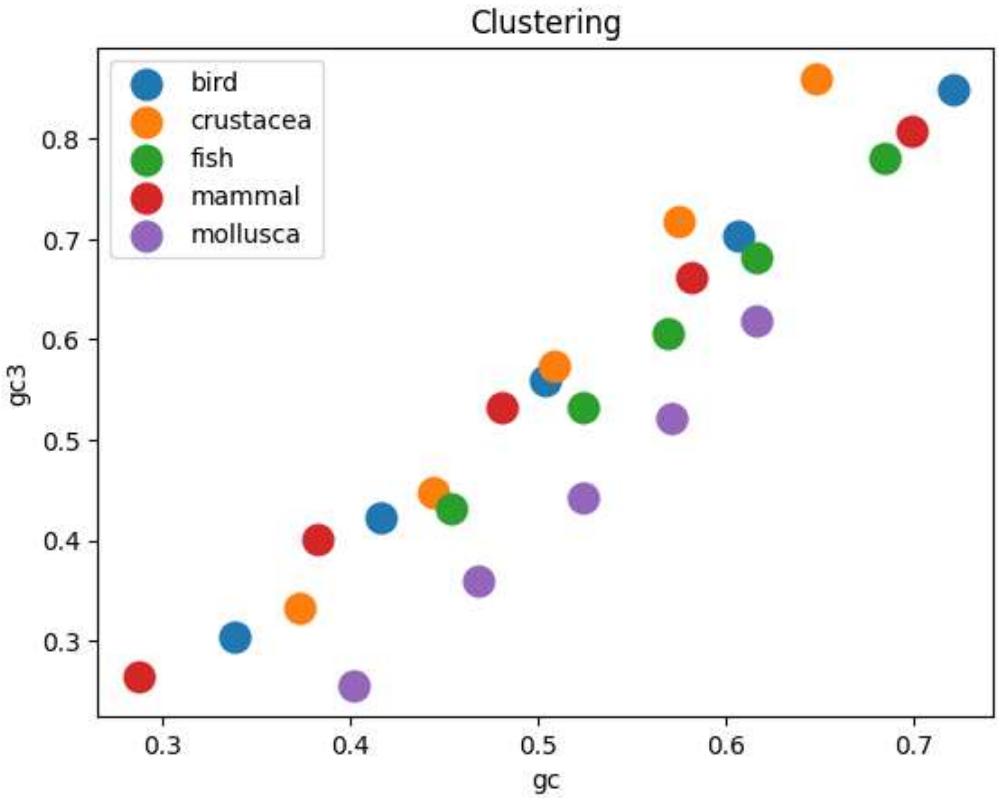


Fig : Centroid of the groups for KMeans

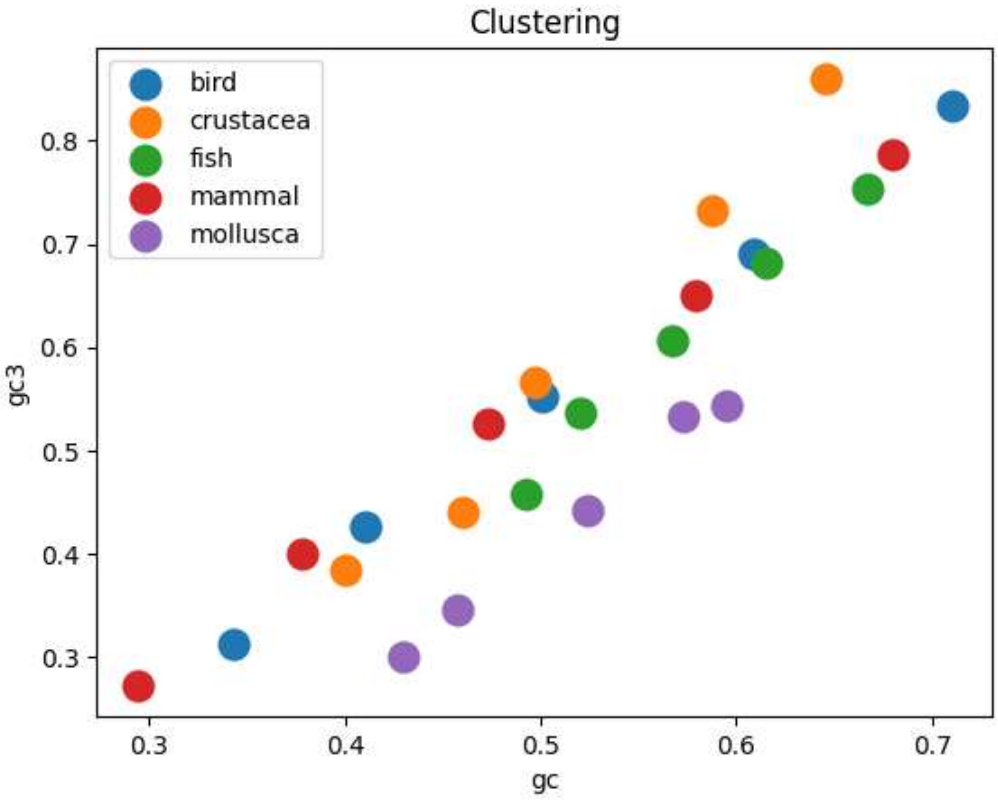


Fig: Centroid of the groups for Gaussian Mixture

The centroid of the various groups follows the data patterns in the scatterplot fig.

As we can see, the mollusca group has a greater gc/gc3 ratio, and its data points are more densely packed than the other groups.

(GC1+GC2)/2 and GC3:

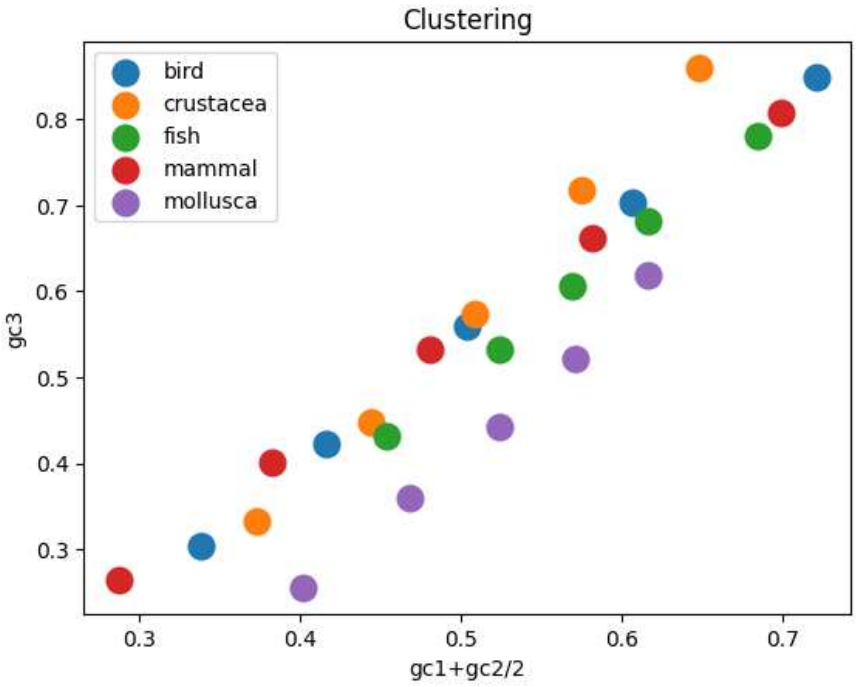


Fig: KMeans

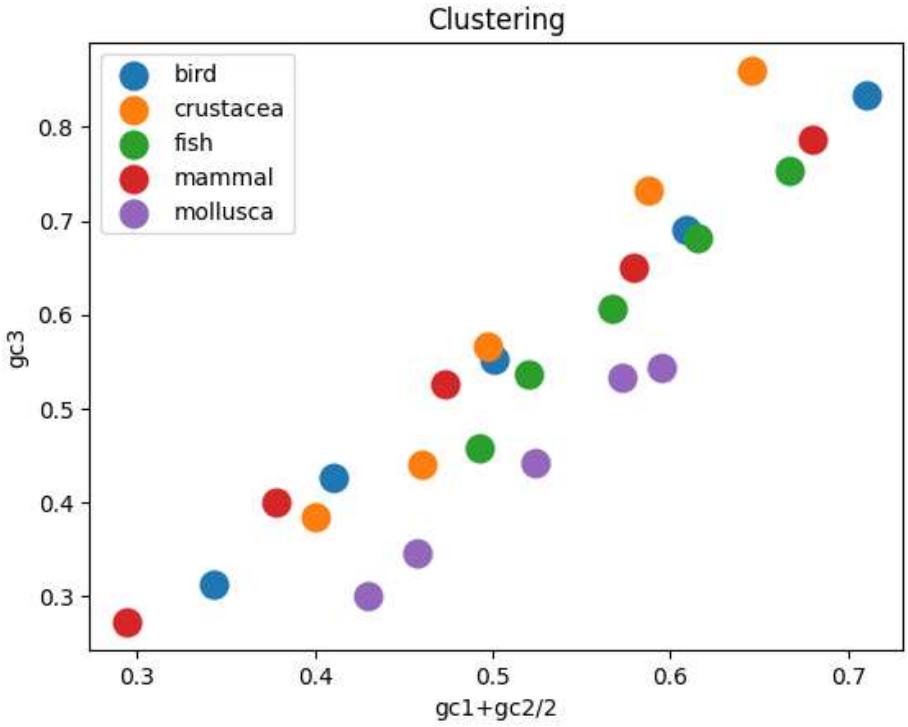


Fig: Gaussian Mixture

GC3 and ENC:

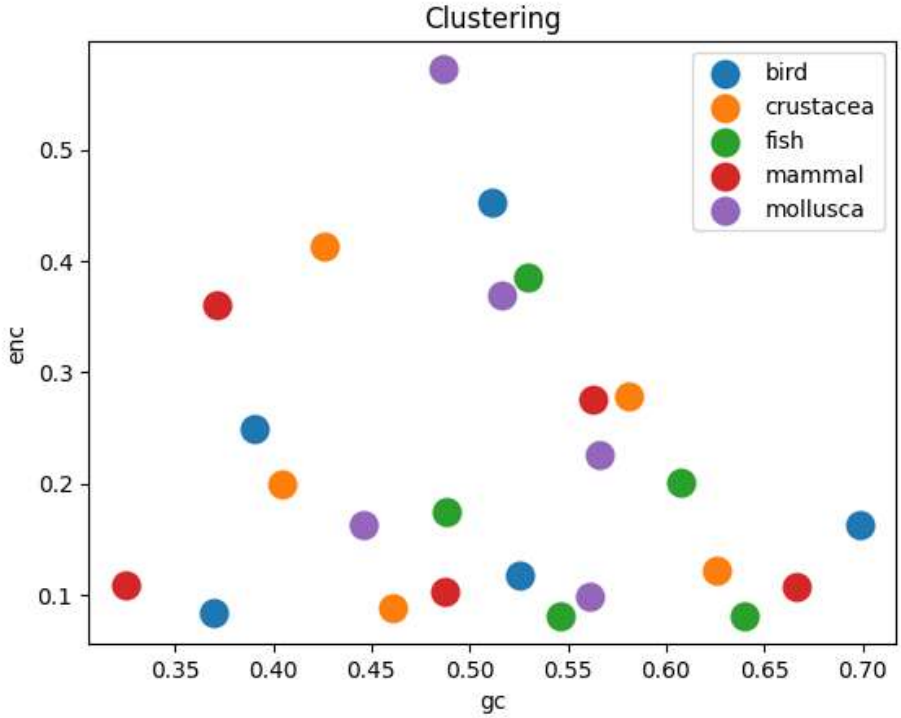


Fig: KMeans

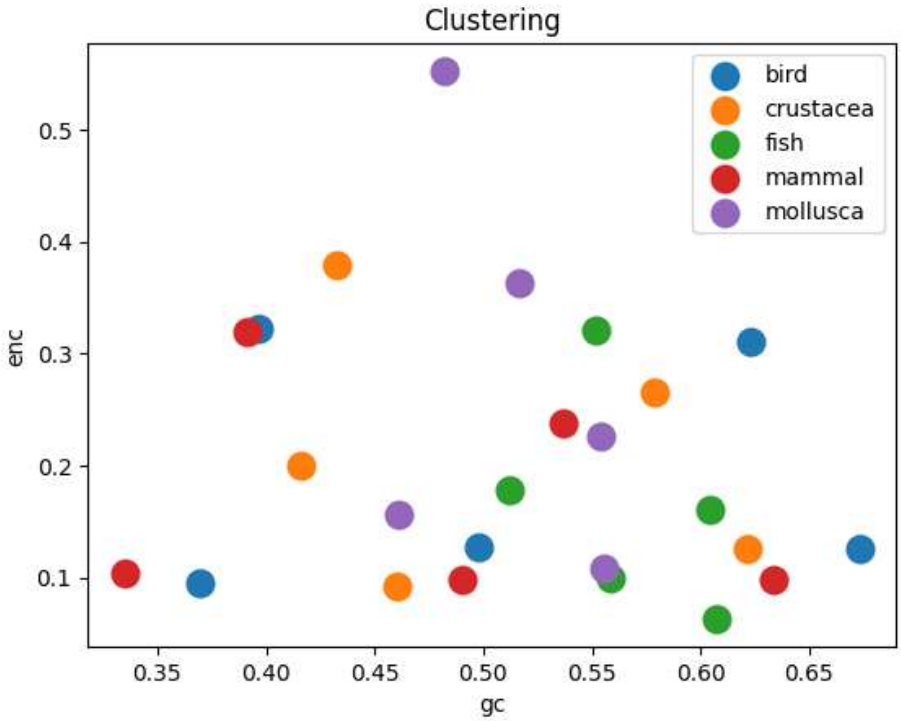


Fig: Gaussian Mixture

GC and ENC:

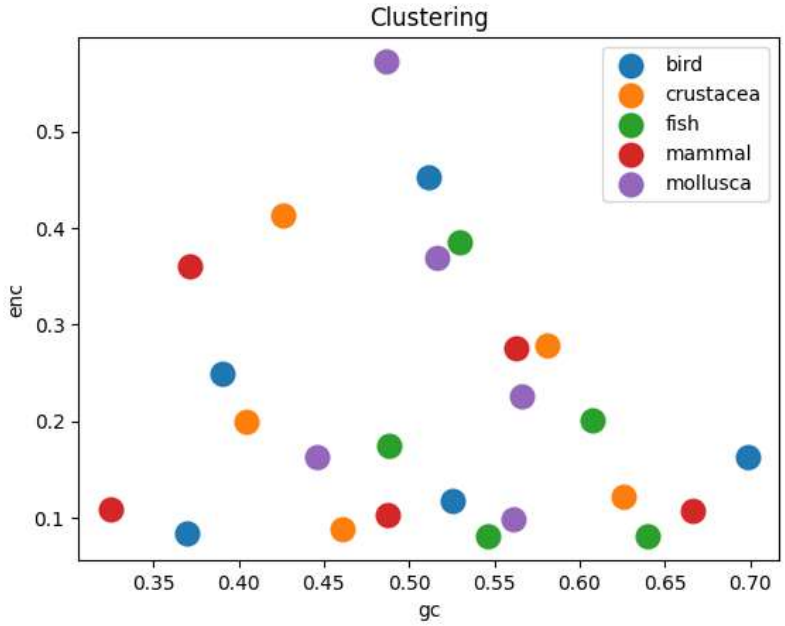


Fig: KMeans

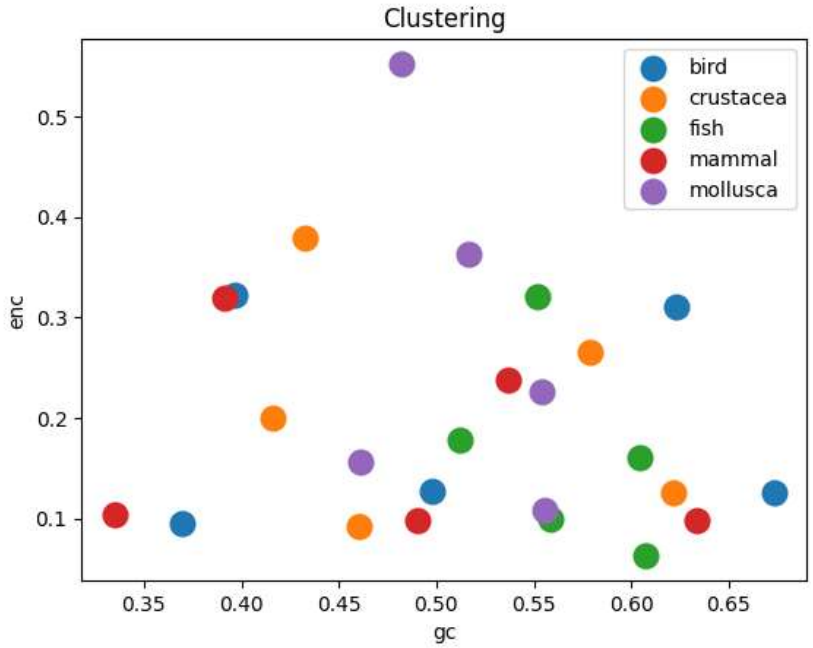


Fig: Gaussian Mixture

Silhouette Score:

Silhouette score for the K-means:

Features/ Group	Bird	Crustacea	Fish	Mammal	Mollusca
GC + GC3	0.41697630484 9112	0.43964720643 927757	0.35963539862 743854	0.38499646423 562683	0.35837638709 772013
(GC1+GC2)/2 and GC3	0.41271038096 8572	0.43512958759 352655	0.08746661140 364118	0.37197004660 037813	0.35678361731 7447
GC3 and ENC	0.40608819985 730576	0.39002069229 04657	0.31358324430 3344	0.34077041038 22914	0.33792927491 94637
GC and ENC	0.33334849367 60932	0.27124152161 41311	0.33239029137 543813	0.36508398523 86045	0.28216125739 5115

Silhouette score for the Gaussian Mixture:

Features/ Group	Bird	Crustacea	Fish	Mammal	Mollusca
GC + GC3	0.12969889719 551925	0.31505103789 890077	0.2694732013 134906	0.01119270675 3703565	0.03326857404 528402
(GC1+GC2)/2 and GC3	0.39120977000 73723	-0.07259534616 660462	0.1220356476 682629	0.37618839982 895125	0.07576739075 386127
GC3 and ENC	0.33852266595 155217	0.32350040937 74181	0.3151403028 975161	0.32649106652 522514	0.34095142627 14156
GC and ENC	0.30306568510 30655	0.22699226895 62331	0.3305345628 179873	0.28232930659 924366	0.26507162988 71848

As per Silhouette score

For Bird The best clustering is K- Means with features GC and GC3

For Crustacea The best clustering is K- Means with features GC and GC3

For Fish The best clustering is K- Means with features GC and GC3

For Mammal The best clustering is K- Means with features GC and GC3

For Mollusca The best clustering is K- Means with features GC and GC3

According to the Silhouette score and the graph, the best clustering can be achieved with K means using features GC and GC3