NAVEEN BALARAJU

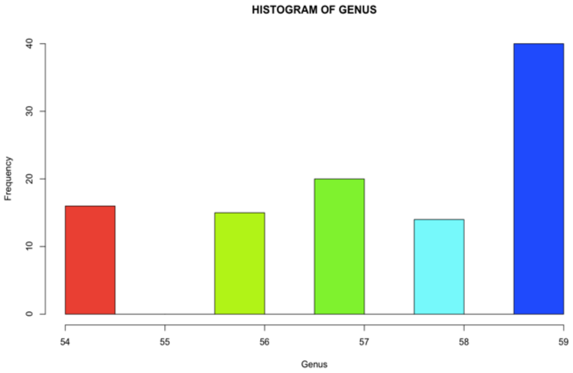
Goal: To implement Hierarchical clustering on Primate Scapulae dataset and evaluate results with the Analytical methods.

Dataset description: Primate. Scapulae is a Dataset that contains the physiological measurement of the scapula bone from 5 different genera of primates i.e
1. Gibbons (Hylobates)
2. Orangutans (Pongo)
3. Chimpanzees (Pan)
4. Gorillas(gorilla)
5. Man (Homo)

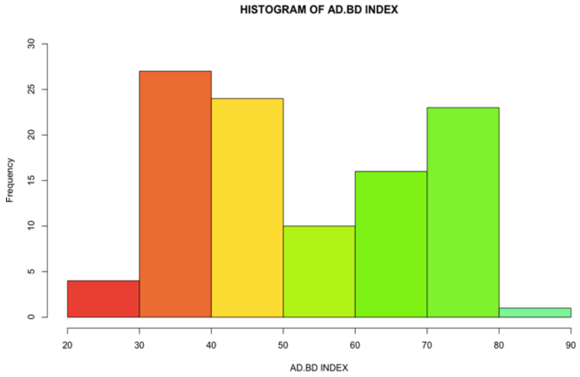
The data set consists of 11 variables and 105 observations. The variables "AD.BD","AD.CD","EA.CD" "Dx.CD", "SH.ACR" are the 5 different indices of the scapulae bones. The variable "EAD", "beta", "gamma" are the angles related to bone. Since, gamma angle is is missing for Homo, the gamma variable is deleted for the further analysis.

Exploratory Data Analysis:

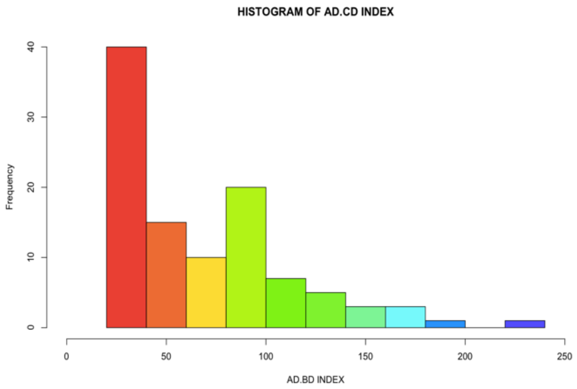
1. Analysis of "Genus" variable: Genus is a numeric vector indicating different orders of the primates.59-Homo;58-Gorilla;57-Pan;56-pongo;54-Hylobates. From, the histogram I can infer that 38% of genus is Homo,13.3% is gorilla,19% pan,14.2 % pongo and 15.5 % hlyobates.



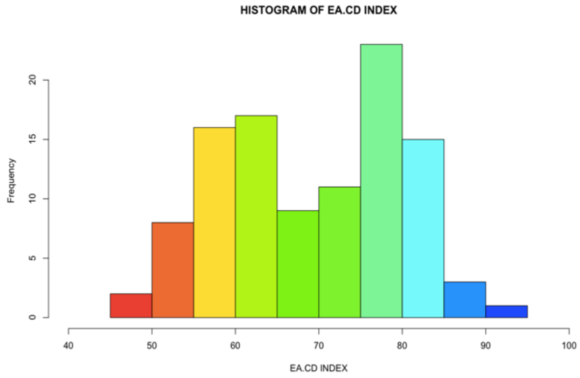
2. Analysis of "AD.BD" index: From the histogram, I can infer that 4% of the observations have AB.BD index in the range of 20-30, 26% between 30-40, 23% between 40-50, 10% between 50-60, 14 % between 60-70,22% between 70-80 and only 1% above 80.



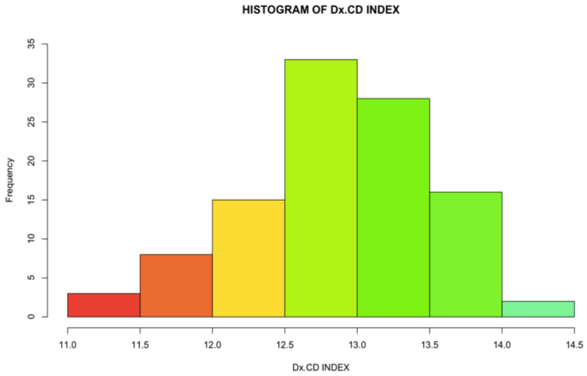
3. Analysis of “AD.CD” index: From the histogram, I can infer that about 52% of observations have AD.CD index values between 23-50, 28% between 50-100 and the rest 20% have more than 100.



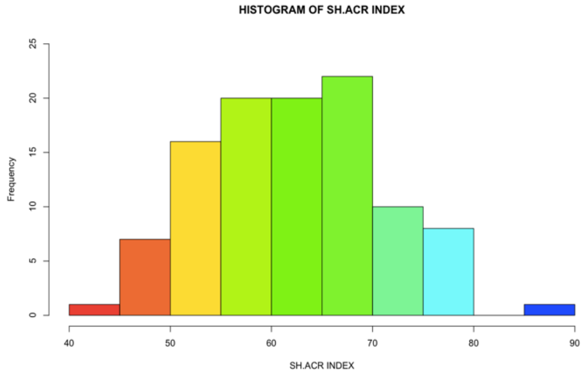
4. Analysis of “EA.CD” index: From the histogram, I can infer that about 25% of the observations have EA.CD index value between 45-60, 35% between 60-75 and 40% above 75.



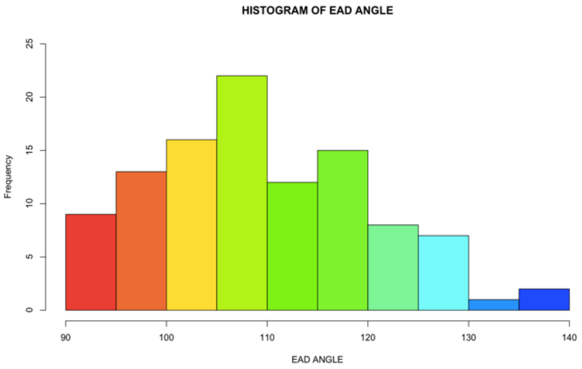
5. Analysis of “Dx.CD” index: From the histogram, I can infer that about 24% of the observations have Dx.CD index between 11-12.5, 57% between 12.5-13.5 and 19% above 13.5.



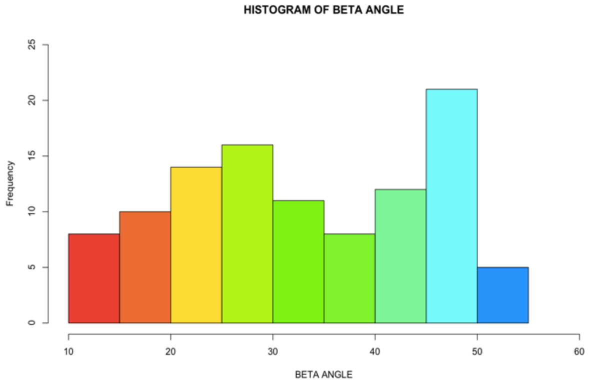
6. Analysis of “SH.ACR” index: From the histogram, I can infer that about 8% of the observations have SH.ACR INDEX value between 40-50, 34% between 50-60, 40% between 60-70 and 18% above 70.



7. Analysis of “EAD Angle”: From the histogram, I can infer that about 21% of the observation have EAD angle between 90-100, 37% between 100-110,28% between 110-120 and 14% greater than 120.



8. Analysis of “Beta Angle”: From the histogram, I can infer that about 17% of the observations have Beta angle between 10-20, 29% between 20-30,18% between 30-40 and 36% above 40.

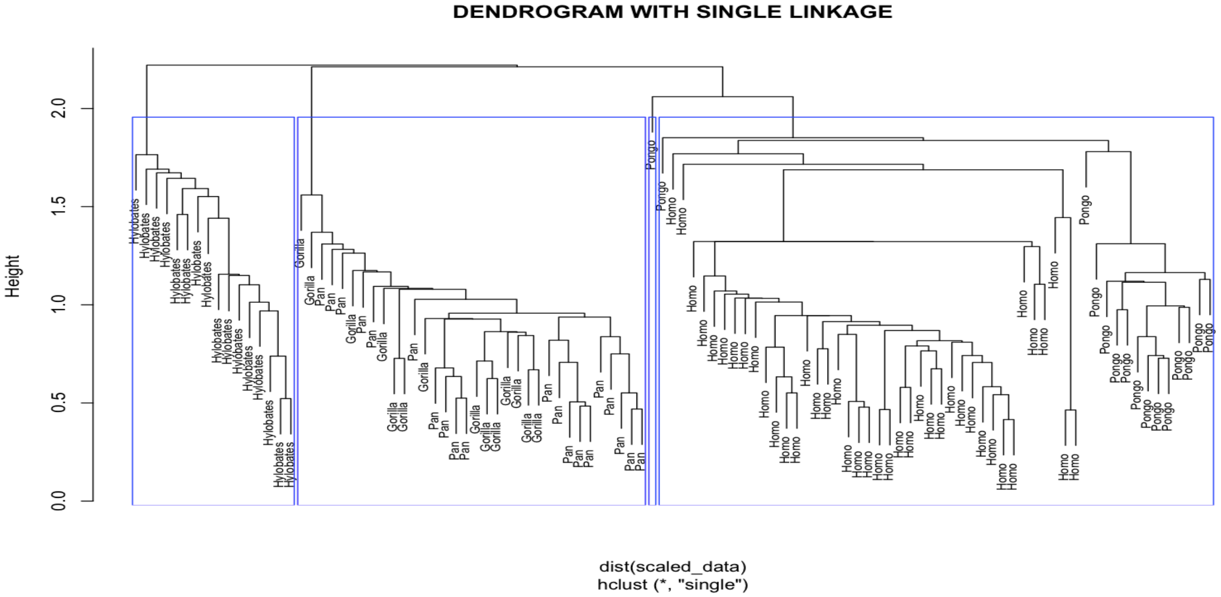


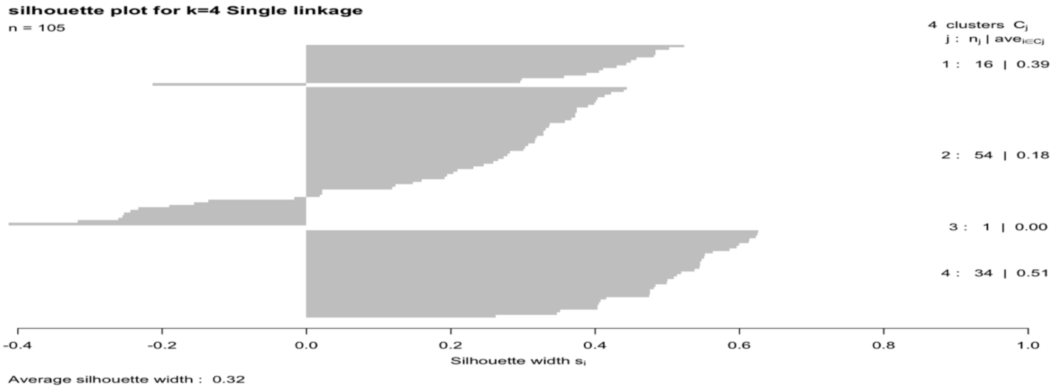
Hierarchical clustering: Since genus, class and class digits have specific values of each genera of primates, choosing these variables in calculating dissimilarity measure is not a good choice.

genus	class	classdigits
54	hylobates	1
56	pongo	2
57	pan	3
58	gorilla	4
59	Homo	5

For the further analysis, I have not considered class, classdigits.

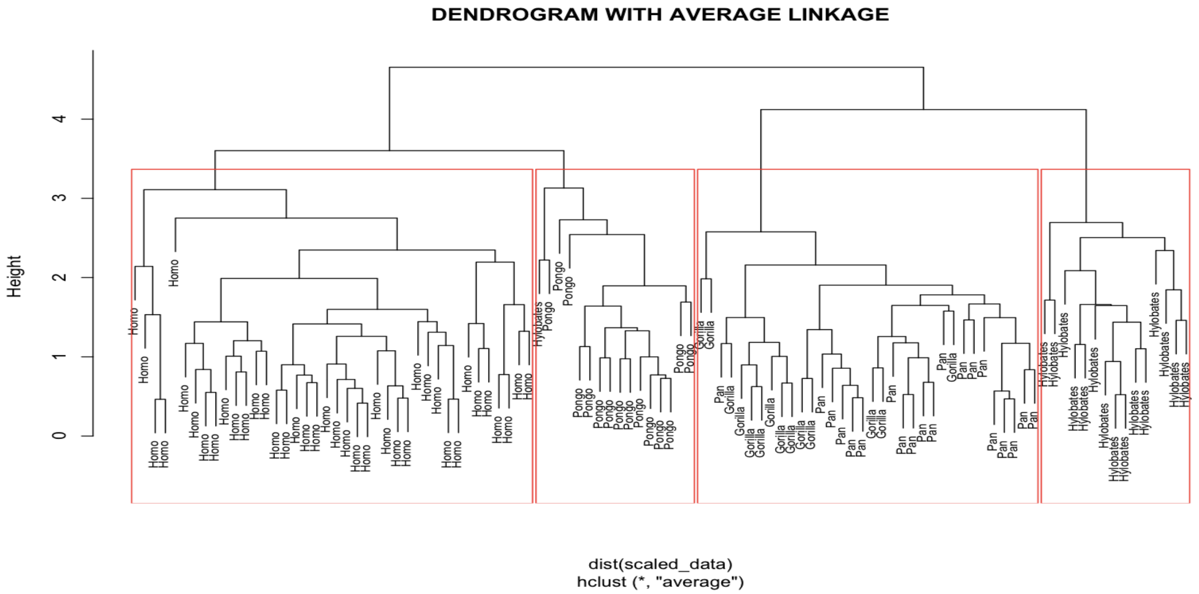
1. Hierarchical clustering with single linkage using Euclidean distance as the dissimilarity measure

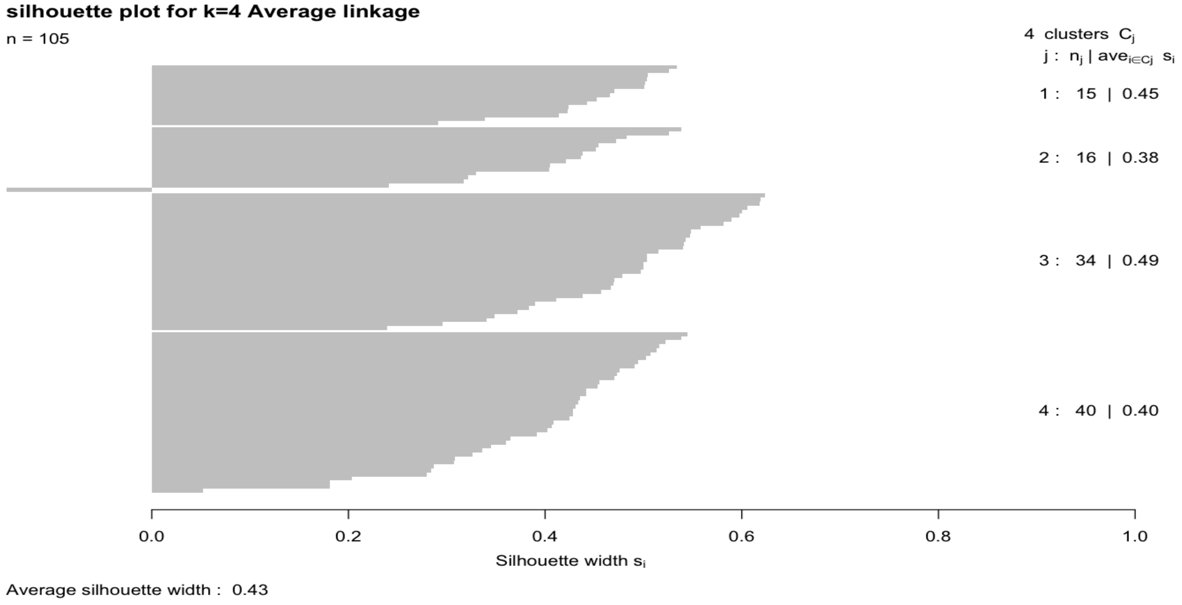




The dendrogram for single linkage shows that Homo, Hylobates and pongo forms separate clusters, while pan and gorilla are combined into one cluster showing that these two variables are correlated and hence I have chosen k= 4 and for the further analysis, I have assumed pan and gorilla as one cluster. The dendrogram generated by single linkage show many singleton clusters that are less compact so choosing Sigle linkage is not a good choice. This method performs the worst with more intracluster and intercluster distances and also low silhouette values 0.32.

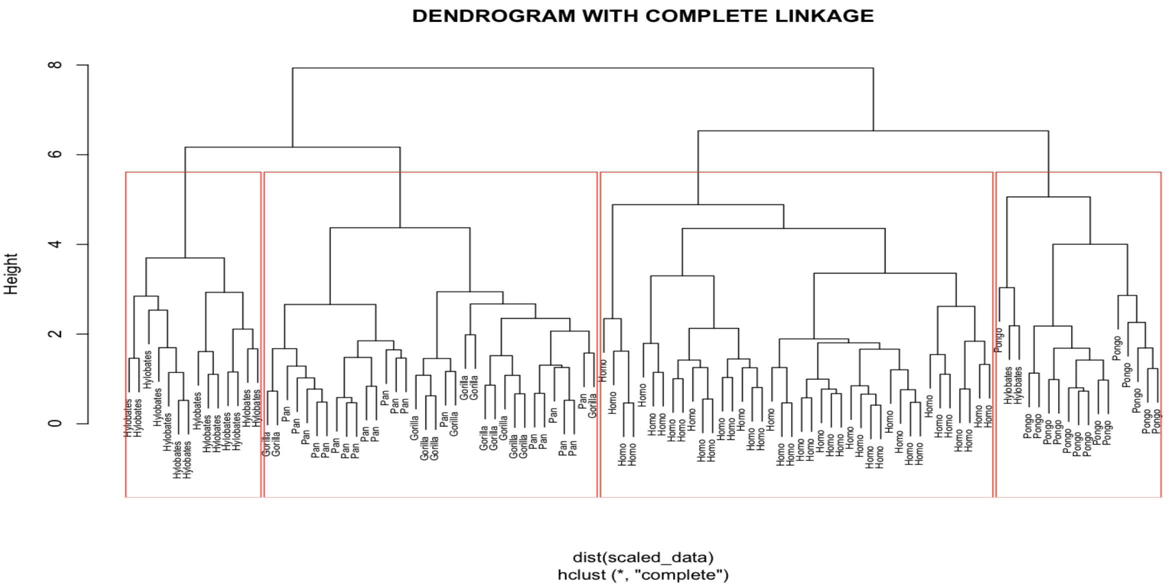
2. Hierarchical clustering with Average linkage using Euclidean distance as the dissimilarity measure.

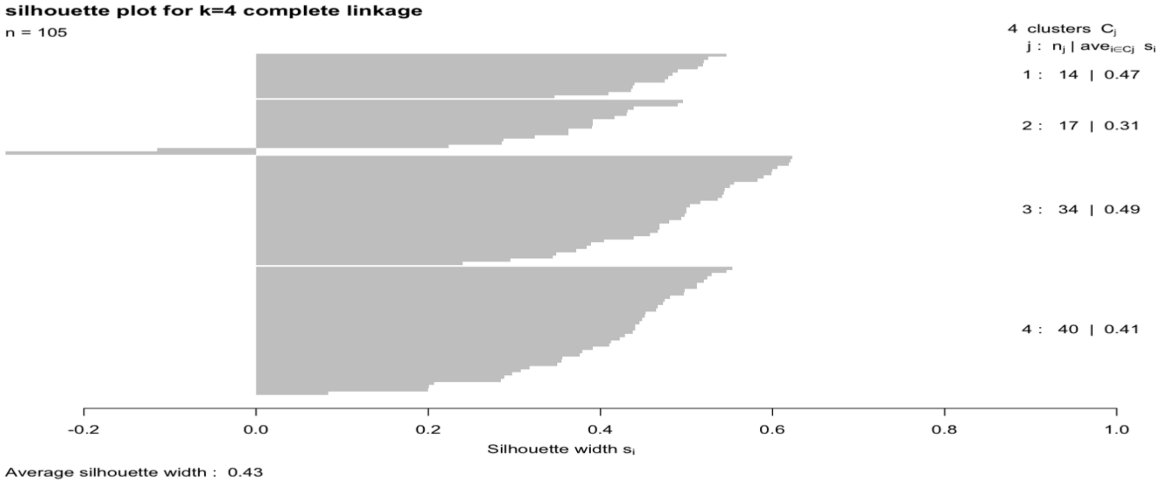




The dendrogram for Average linkage shows that Homo, Hylobates and pongo forms separate clusters, while pan and gorilla are combined into one cluster showing that these two variables are correlated and hence I have chosen k= 4. The dendrogram produced by average linkage shows 4 compact clusters, with less intracluster and intercluster distances. This method performs the best as it mis classifies one Hylobates point and have a silhouette value 0.43.

3. Hierarchical clustering with Complete linkage using Euclidean distance as the dissimilarity measure.





The dendrogram for complete linkage shows that Homo, Hylobates and pongo forms separate clusters, while pan and gorilla are combined into one cluster showing that these two variables are correlated and hence I have chosen k= 4. The dendrogram produced by average linkage shows 4 compact clusters, with less intracluster and intercluster distances. This method also has a silhouette value 0.43, which is similar to the average linkage but mis classifies two hylobates points as pongo and hence I infer hierarchical clustering with average linkage is the best suited for this data set. The clustering results did not match my expectations as there were 5 different group of primates, it gave me an impression that there are 5 clusters in the data set, but after the analysis it was found that there are 4 clusters with pan and gorilla combined in a single cluster.