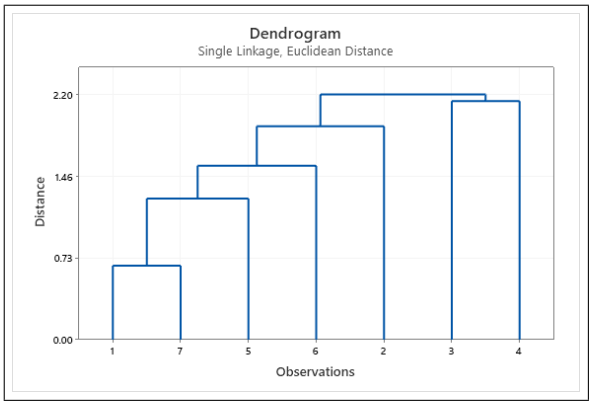
Additional Multivariate Exercise 4

Data: The prehistoric dogs data were described in the Manly textbook; we will here use the summarized version of Example 1.4. For each of 7 species, 6 lower jaw measurements are available.

Hierarchical cluster analysis (single linkage): The output from this analysis, based on standardized variables and Euclidean distances, performed in Minitab are shown below.



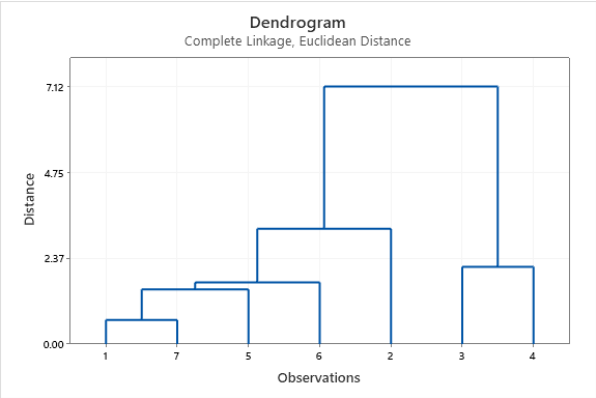
Amalgamation Steps

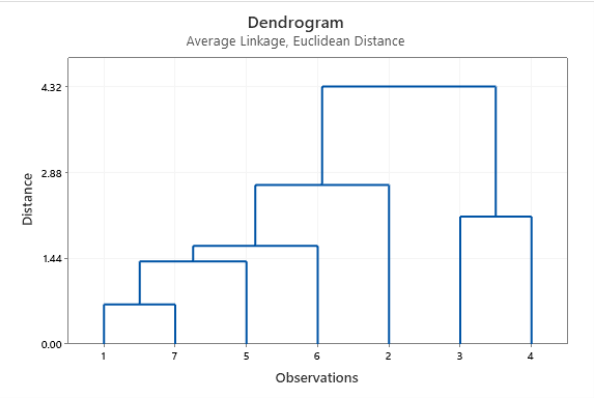
Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of obs. in new cluster
1	6	90.6842	0.66324	1	7	1	2
2	5	82.2415	1.26432	1	5	1	3
3	4	78.1022	1.55901	1	6	1	4
4	3	73.1401	1.91229	1	2	1	5
5	2	69.9611	2.13861	3	4	3	2
6	1	69.1466	2.19660	1	3	1	7

It is seen that the distances do not quite match with those shown in Table 9.6 of the Manly text. For example, the distance at which the first cluster is formed (between the Modern and Prehistoric dogs) equals 0.66, in agreement with the computed distance matrix (Table 5.2 of Manly and lecture 9). In Table 9.6 this value is given as 0.72, for unknown reasons. The R code accompanying the book also reproduces a distance of 0.66. Another confusing feature of Table 9.6 is its duplicate entries for many of the steps where clusters are formed; it is not described what these are, but they could be the corresponding distances with different linkages (below).

The dendrogram however has exactly the same structure. Note that in Minitab it is possible to get the species labels displayed in the graph instead of the observation number, by using the variable containing the species name as "Case labels". Two types of conclusions can be made. First from the bottom of the dendrogram, the closest distance is between the two dogs (observations 1 and 7; this we had observed already), and the closest species to the two dogs is the cuon (observation 5; joined at a distance of 1.26, nearly twice the distance between the two dogs). Next from the top of the dendrogram and down, the first split of the data into two clusters is obtained by combining the two wolves (Chinese and Indian) into one cluster, and contrasting them with the other species. This makes sense from the distance matrix despite that the two wolves are not that close to each other, because their distances to the other species are quite large across the board. The suggested formation of three clusters would split the two wolves from each other (at their distance of 2.14), and keep the remaining species as a single cluster. The following step then separates the golden jackal from the other species (at a distance of 1.91) to make it a cluster on its own, together with the two wolves clusters and the remaining species cluster.

Hierarchical cluster analysis (other linkages):
We next consider other linkages; on the next page, dendrograms are shown for complete and average linkages (others could be explored as well). These two linkages lead to the exactly same structure of the dendrogram, differing only from that of the single linkage by the order of the splits involving the wolves and the golden jackal. These dendrograms will, from the top, separate out the golden jackal before splitting the wolves. The difference does not affect our interpretations much.





K-means clustering: When this analysis is run in Minitab with $K = 2$ without any initial partition column specified, the partitioning puts the golden jackal in a separate cluster. According to the Minitab documentation, the default initial partioning uses the order of the rows of the worksheet and does not involve any randomness. So in order to get different results, one needs to either reorder the rows or specify an initial column with values representing the desired number of clusters. With an initial column separating out the two wolves into a cluster, a different final partitioning was obtained (equal to that initial partitioning). The within-cluster sum-of- squares is much lower for this partitioning (10.7 versus 24.3). Therefore the default result in this instance happened to be pretty useless. Even if it is not too difficult to create a random initial partition (e.g. using the discrete distribution in the "Random Data" menu for generating random numbers), it still seems that for practical applications, one might be better off using another software than Minitab for K-means partitioning.

In Stata, 10 K-means runs with $K = 2$ and random starting configurations all produced the second local minimum above. The same solution is obtained as the best among 100 random starting configurations in the R implementation. For $K = 3$, 10 runs in Stata gave four different solutions, of which the one with lowest within-cluster sum-of-squares (5.40) appeared six times. This solution had split the golden jackal from the 5-species cluster remaining after the wolves had been separated, i.e. in agreement with the dendrograms obtained from complete and averages linkages. Among 100 runs in R with random starting configurations the same solution was obtained.

In conclusion, once performed appropriately the K-means clustering algorithm for these data produces results that are both sensible and in agreement with those from hierarchical clustering algorithms.