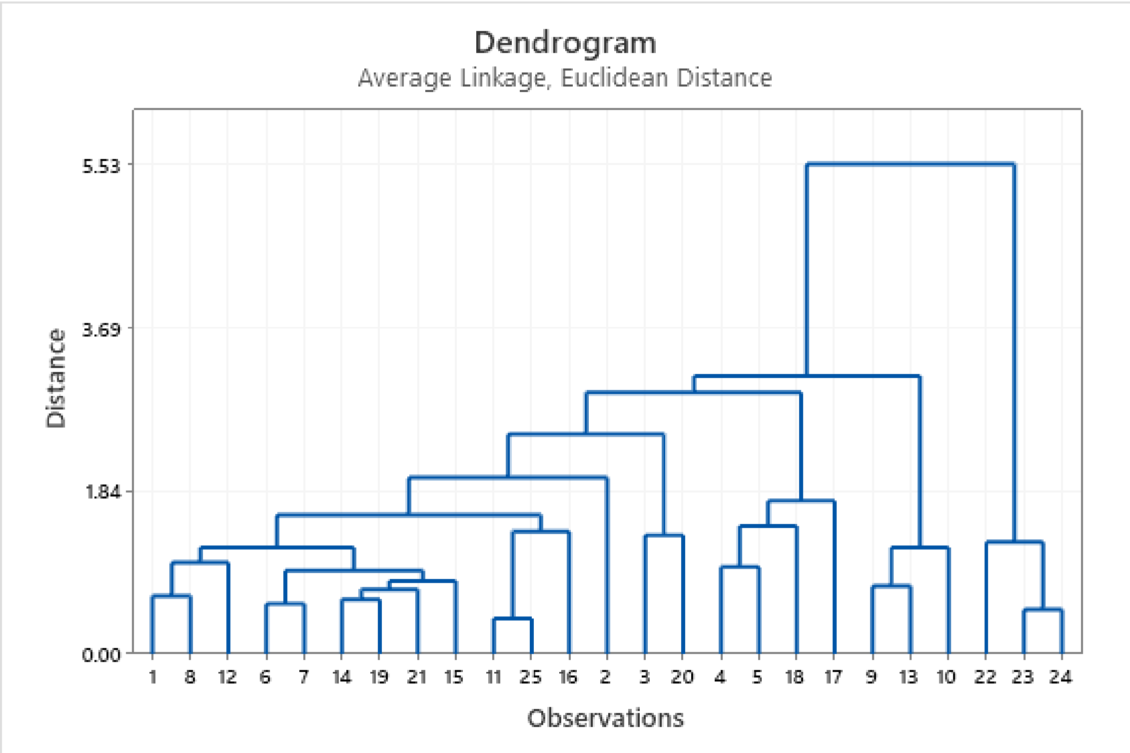
Additional Multivariate Exercise 7

Data: The goblet dataset was described in the Manly text and the question. A matrix scatterplot should be the starting point of exploration (not shown here). All variables are positively correlated, with Pearson correlations ranging from 0.25 (for x_3 and x_5) to 0.91 (for x_4 and x_6), and most of them are quite strong (with only 3 correlations below 0.45). Descriptive statistics for the 6 measurements show some variation in magnitude, with x_3 having about three times larger mean and standard deviation than x_5. For this reason, all distance calculations should be based on standardized variables. Note that this standardization across goblets does not remove the effect of size, and a different normalization is needed to focus on shape as referred to in the second part of the question.

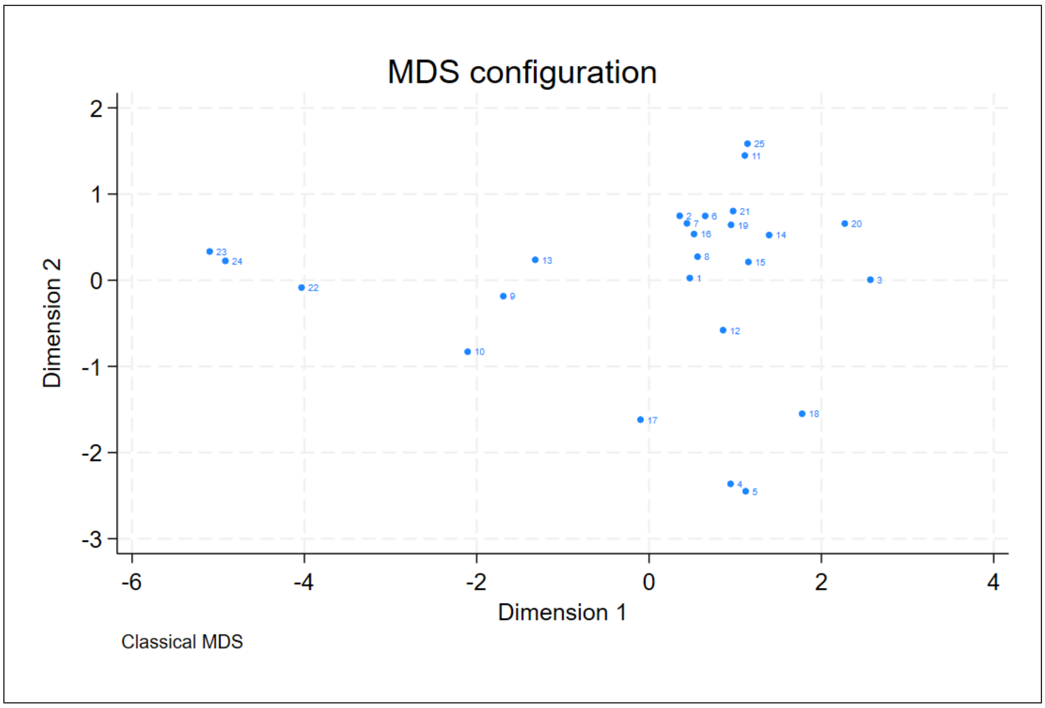
Hierarchical cluster analysis (average linkage): The results from single and average linkage indicate similar clusters but the scales for the distances are different (ranging from 0.4 to 2.17 for single linkage, contrasting the dendrogram below for average linkage).



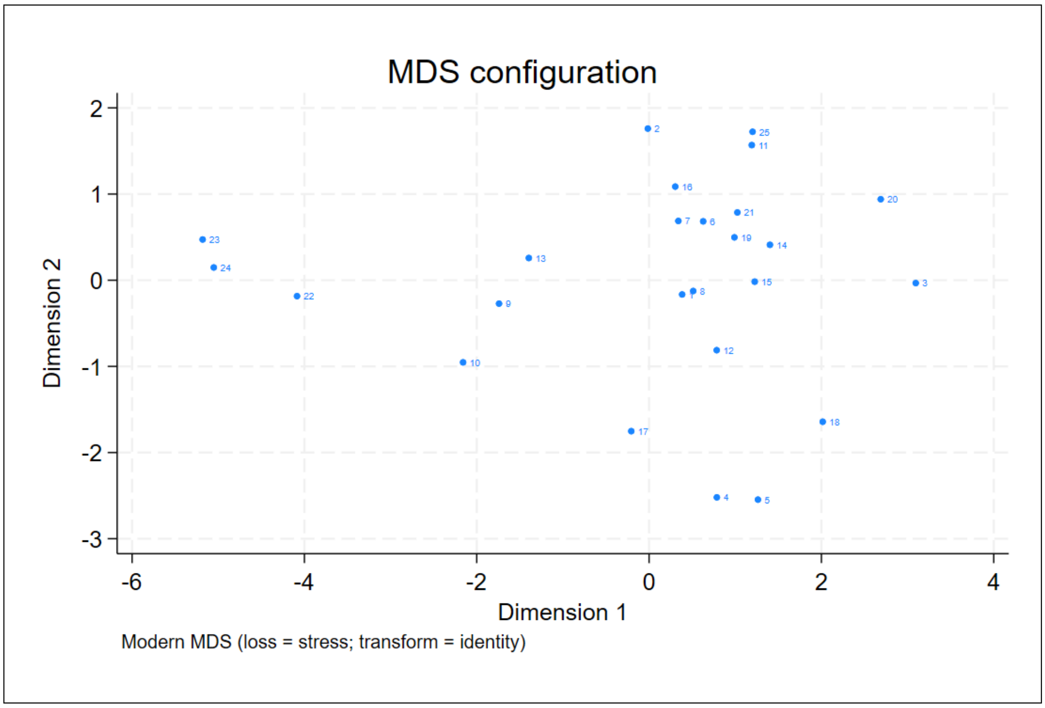
Looking at the clusters determined from the top down, the first one, and strongest in terms of distance, separates out the three clearly smallest goblets (22−24). The next two clusters separated consist of goblets (9,10,13) and (4,5,17,18). The first of these seems to be characterized by a sizeable value on x_1 but relatively low (but not too extreme) values on x_2, x_4, x_6 without all values being low. The second of these seems to be characterized by high values on x_1 and x_5, without all other values being high as well. If one wants to consider further clusters, the next splits are for the goblets (3,20) and 2 alone, but the interpretations become more and more difficult. At the lower end of the dendrogram, some pairs of goblets are seen to be very similar (e.g. (23,24) and (11,25)), which is also visible directly in the data matrix.

K-means clustering: We briefly review the results for $K = 2, 3, 4, 5$ clusters. For 2 clusters, the goblets (9,10,13,22,23,24) are split out from the rest, at 56.0% of variation explained. A third added cluster consists of goblets (4,5,17,18), at 70.5%. In the next step, the first 6-goblet cluster is split into (9,10,13) and (22−24), at 80.5%. Finally, the goblets (3,20) at split from the main cluster, at 86.3%. The clusters are seen to agree very well with those from above.

Classical MDS: The classical MDS from Euclidean standardized distances corresponds to a PCA for the correlation matrix. The first two eigenvalues explain 89% of the variation, so it is natural to focus on the first two components only. For display, we prefer the MDS plot because it allows the goblet numbers to be overlaid directly on the plot (below). The PCA eigenvectors (see table on page 3) show high loadings on the first component for all six variables (range $0.296 - 0.462$), a similar situation as for the sparrow data. Thus the first component simply reflects goblet size. The second eigenvector loads negatively on x_1 and x_5 and positively on x_3 and to some extent on x_6. These loadings seem to represent a measure of the height versus the width of the goblet, thus a shape feature. The MDS configuration (or score) plot shows the same clusters of goblets as previously noted.



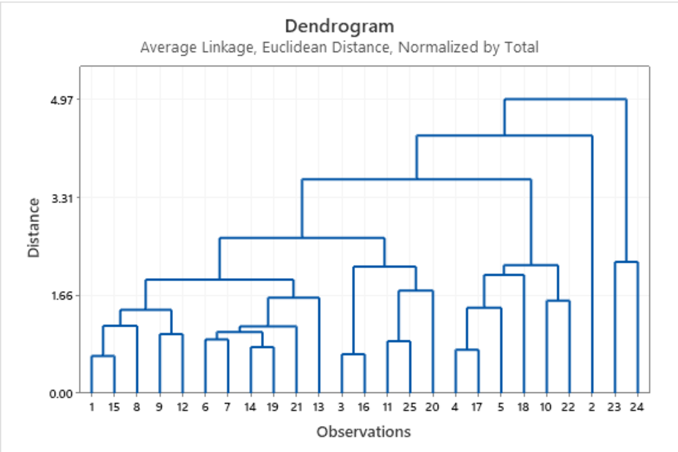
Modern MDS minimizing the "stress-1" loss function does not seem to change the position of points substantially, but it does spread out the main cluster of points. The resulting stress value is quite low (0.0836), indicating that an acceptable adaptation to the actual distances has been found.

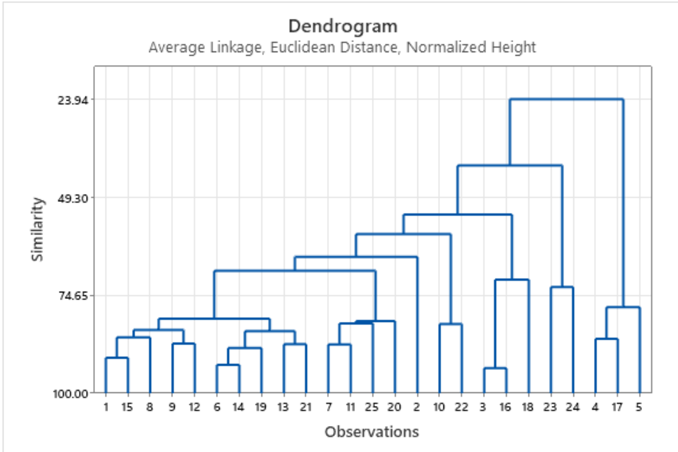


Size-adjusted variables: Two suggestions are offered for a size normalization, by the total of observations or the goblet height x_3:

$$
\begin{aligned}
x_{ki}^{(t)} &= x_{ki}/(x_{1i}+x_{2i}+x_{3i}+x_{4i}+x_{5i}+x_{6i}), \quad \text{or} \\
x_{ki}^{(h)} &= x_{ki}/x_{3i}, \quad \text{where } k=1,2,4,5,6;\ i=1,\ldots,25.
\end{aligned}
$$

Because the $x_{ki}^{(t)}$ variables add up to 1, one may decide to remove one variable from the set. We will however for ease of comparison with the unadjusted analysis keep all the variables in. Another matrix scatterplot would be good. The new variables should also be standardized (most critical for $(x_k^{(h)})$). Dendrograms for normalized variables, still using average linkage are shown below.





The clustering is seen to be strongly affected by the normalizations, both when comparing to the unadjusted dendrogram shown earlier and when comparing these two dendrograms. Results for the "total" normalization still show some similar features to the unadjusted dendrogram: the separation of goblets (23,24) (even if now without 22), the grouping of goblets (4,5,17,18) into a cluster (even if now at a lower distance) and the uniqueness of goblet 2 (even if this is much stronger now). On the other hand, the previous grouping (9,10,13) is no longer present. After the normalization, the closest goblets are the pairs (1,15) and (3,16), but neither of these were that close earlier.

The results for the "height" normalization retain some of the clusters, e.g. (4,5,17) (without 18), and (23,24) and 2, but the fourth cluster division establishes a new cluster combining (3,16,18) and (23,24), even if these are separated in subsequent steps. The cluster (10,22) appears with both normalizations, but was not present without normalization. In summary, it is clear that both normalizations affected the clusters determined by hierarchical clustering. For this solution, we omit the corresponding analyses by K-means clustering.

MDS for size-adjusted variables: For brevity, we will focus on the results of classical MDS (or PCA) for the two versions of the normalization; the table below gives (PCA) eigenvectors and eigenvalues for the first two components.

Variable/	No normalization		Total normalization		Height normalization	
Component	1	2	1	2	1	2
x_1	0.366	-0.486	0.474	-0.151	0.495	0.311
x_2	0.452	0.034	0.356	0.515	0.490	0.330
x_3	0.411	0.441	-0.379	0.502	–	–
x_4	0.462	0.115	-0.419	-0.466	0.473	-0.412
x_5	0.296	-0.683	0.298	-0.489	0.505	0.057
x_6	0.438	0.298	-0.491	-0.058	0.190	-0.788
Eigenvalue	4.27	1.09	3.04	1.62	3.04	1.28
Prop. variance	0.71	0.18	0.51	0.27	0.61	0.26

The two components still explain a large part of the variation (the normalization makes a direct comparison difficult, also because the height normalization only has five variables). The loadings of the first component for the height normalization are all positive, and thus represent a similar size (relative to height) feature. The score plots are plots show the clusters identified in the hierarchical cluster analysis, but their positions have shifted quite a bit between the different plots. Goblet 2 seems to most affected by the choice of normalization; with the height normalization it no longer forms a distinct cluster.