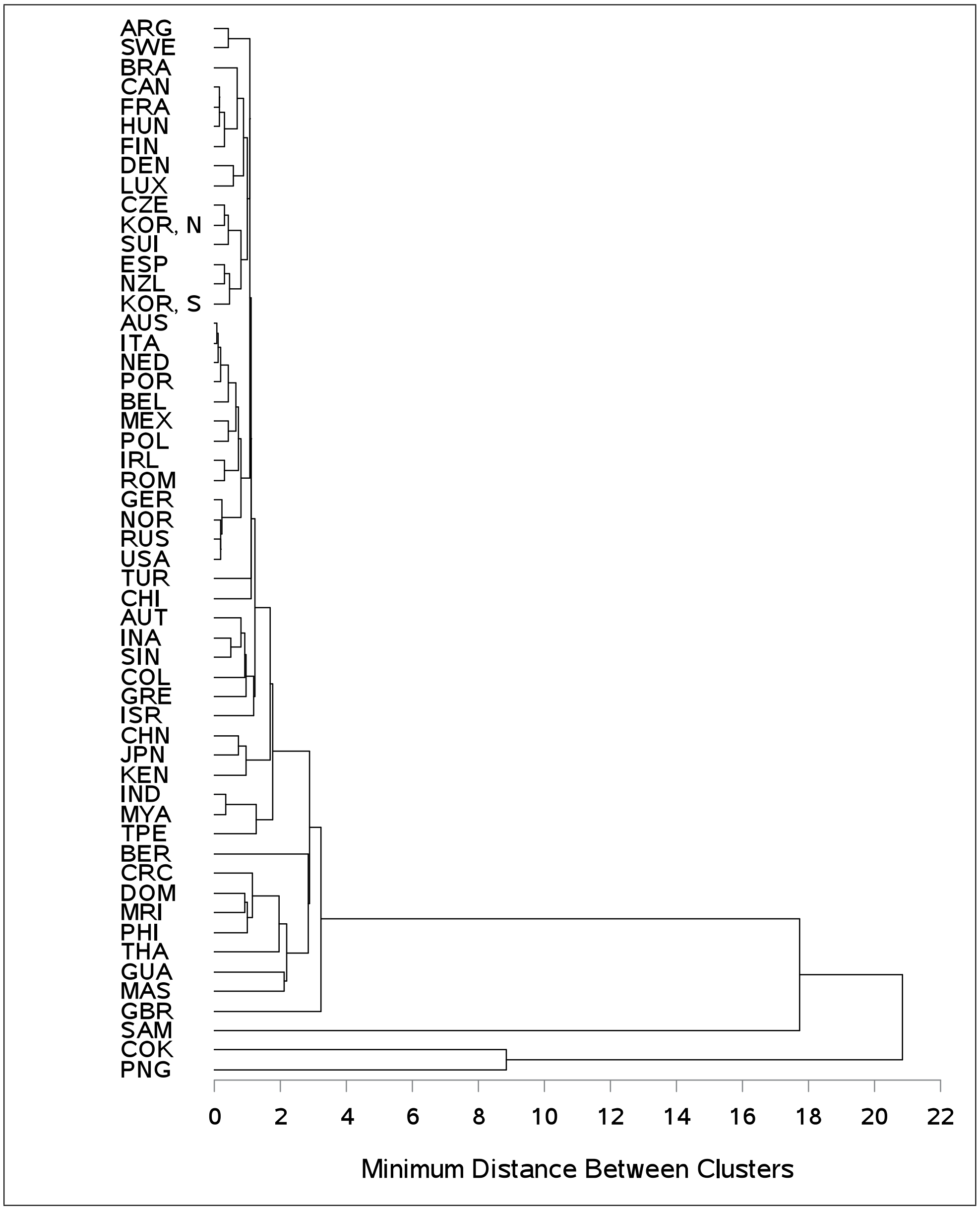
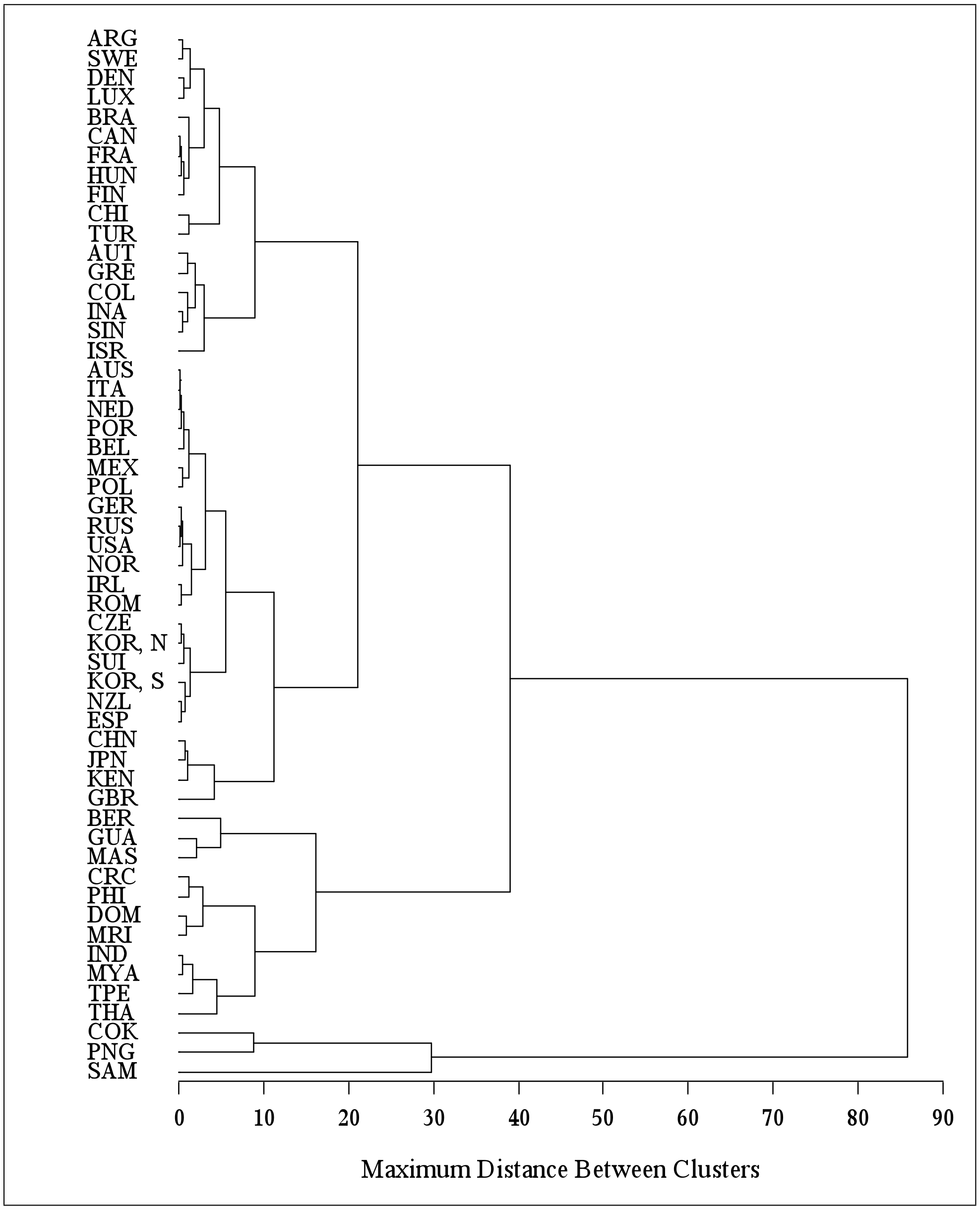
STAT505 Assessment #12

1. J&W Exercise 12.16 (parts b and c only). The data are given in "trackw.dat". Columns correspond to country, 100m, 200m, 400m, 800m, 1500m, 3000m, and marathon. Note that the data is delimited by tabs ☹.

   (b) In both cases, the countries of COK, PNG, and SAM are the most distinct from the others; their times generally seem less impressive in comparison with other countries. The complete linkage method does tend to produce more clusters earlier in its algorithm than does the single linkage method, which tends to add to existing clusters. In either case, there is arguably three to five clusters potentially.



single linkage



complete linkage

(c) With $k = 5$, the countries of COK, PNG, and SAM are once again distinguished from the other countries. Upon inspection, many of the countries clustered together using the linkage methods above are also clustered together using the k-means procedure here; they seem reasonably consistent with each other. Below are some relevant summary information for each of $k = 3$ and $k = 5$. Note that in both cases, the cluster with three countries contains COK, PNG, and SAM.

Cluster	Frequency	RMSSD	Distance between cluster centroids
1	34	1.5168	16.5934
2	17	2.5201	16.5934
3	3	5.7660	46.9501

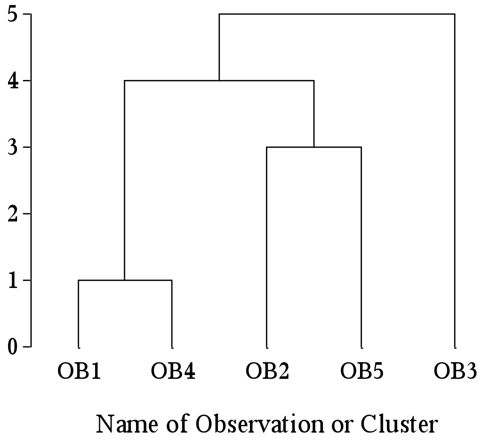
Cluster	Frequency	RMSSD	Distance between cluster centroids
1	8	1.4608	11.9480
2	17	0.8990	6.3028
3	10	0.9133	7.7494
4	16	0.7392	6.3028
5	3	5.7660	40.8247

2. Distances between pairs of five items are given in the following matrix. For example, the distance between the third and second observations is nine, found in the third row, second column.

$$\begin{array}{cccccc} & 1 & 2 & 3 & 4 & 5 \\ 1 & 0 & & & & \\ 2 & 4 & 0 & & & \\ 3 & 6 & 9 & 0 & & \\ 4 & 1 & 7 & 10 & 0 & \\ 5 & 6 & 3 & 5 & 8 & 0 \end{array}$$

(a) Create clusters using a single linkage hierarchical procedure. Describe the progression from all points as separate clusters to all points in a single cluster. At each step, list the clusters that exist at that stage.

The first iteration joins observations 1 and 4 at a distance of 1, then observations 2 and 5 at a distance of 3. The next pairing is between these clusters and occurs, by single linkage, at a distance of 4, which is the distance between observations 2 and 1. Finally, observation 3 is added to the others.



(b) Create clusters using a complete linkage hierarchical procedure. Describe the progression from all points as separate clusters to all points in a single cluster. At each step, list the clusters that exist at that stage.

Progression here proceeds similar to that above except after the first two clusters are formed (1 with 4 and 2 with 5), they are joined at a greater distance by complete linkage and corresponds to 8, the distance between observations 4 and 5. Finally, observation 3 is joined with the others.

SAS code:

```
options ls=78;
data track;
   infile "v:\trackw.dat" delimiter=’09’x;
   input country $ d100 d200 d400 d800 d1500 d3000 marathon;
   d100=d100/60;
   d200=d200/60;
   d400=d400/60;
  run;
proc cluster data=track method=complete nosquare nonorm outtree=clust1;
  var d100 d200 d400 d800 d1500 d3000 marathon;
  id country;
  run;
proc tree horizontal data=clust1 nclusters=5 out=clust2;
  run;
proc fastclus data=track maxclusters=3 maxiter=100
    out=clust replace=random;
  var d100 d200 d400 d800 d1500 d3000 marathon;
  id country;
  run;
data dendro (type=distance);
  input _type_ $  x1-x5;
  cards;
  distance 0 . . . .
  distance 4 0 . . .
  distance 6 9 0 . .
  distance 1 7 10 0 .
  distance 6 3 5 8 0
  ; run;
PROC CLUSTER DATA=dendro (type=distance) METHOD=complete
    nosquare nonorm OUTTREE=tree;
  RUN;
proc tree;
   run;
```