k-means clustering

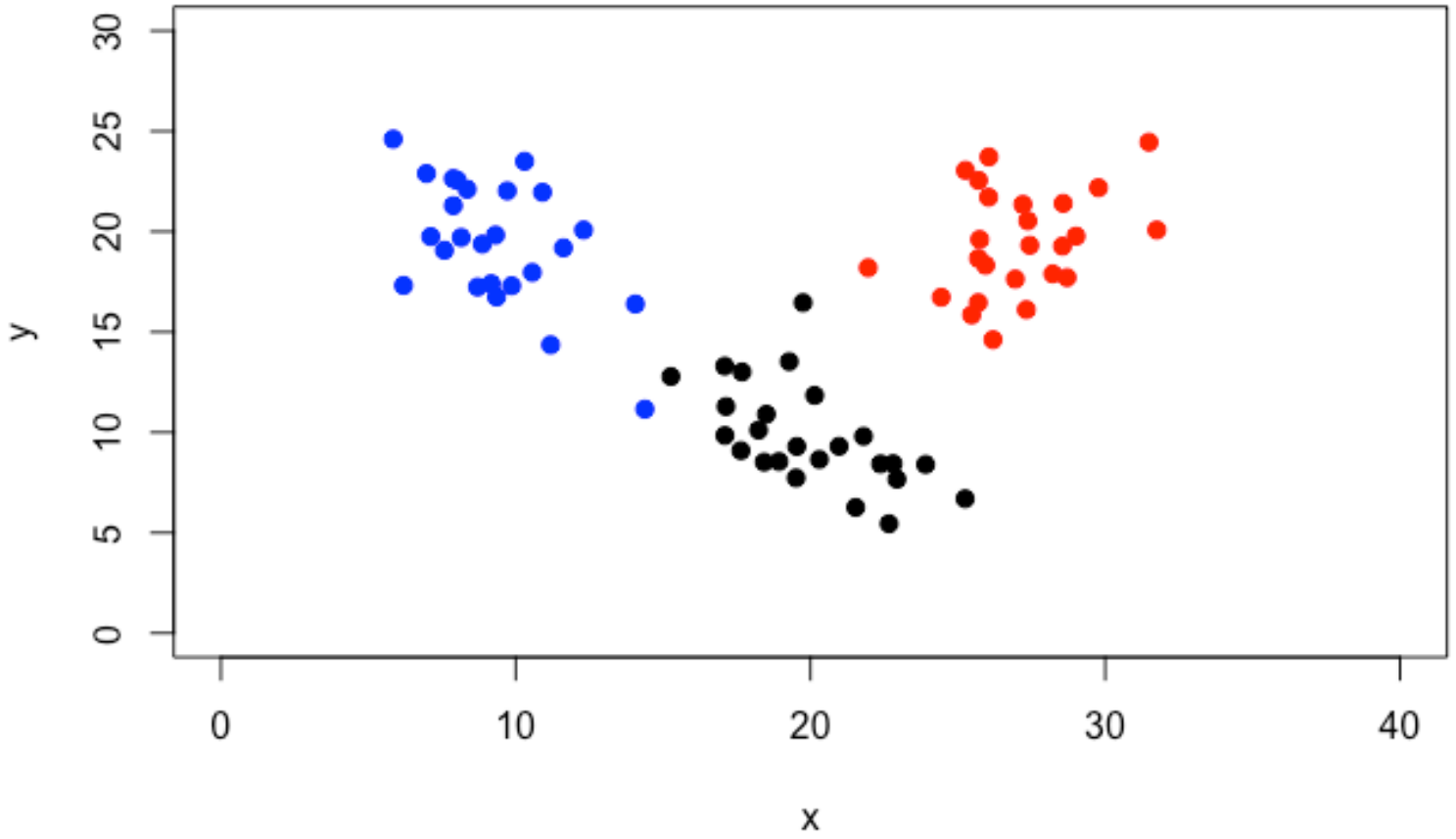
A non-hierarchical clustering scheme

to subdivide the observations into *k* groups

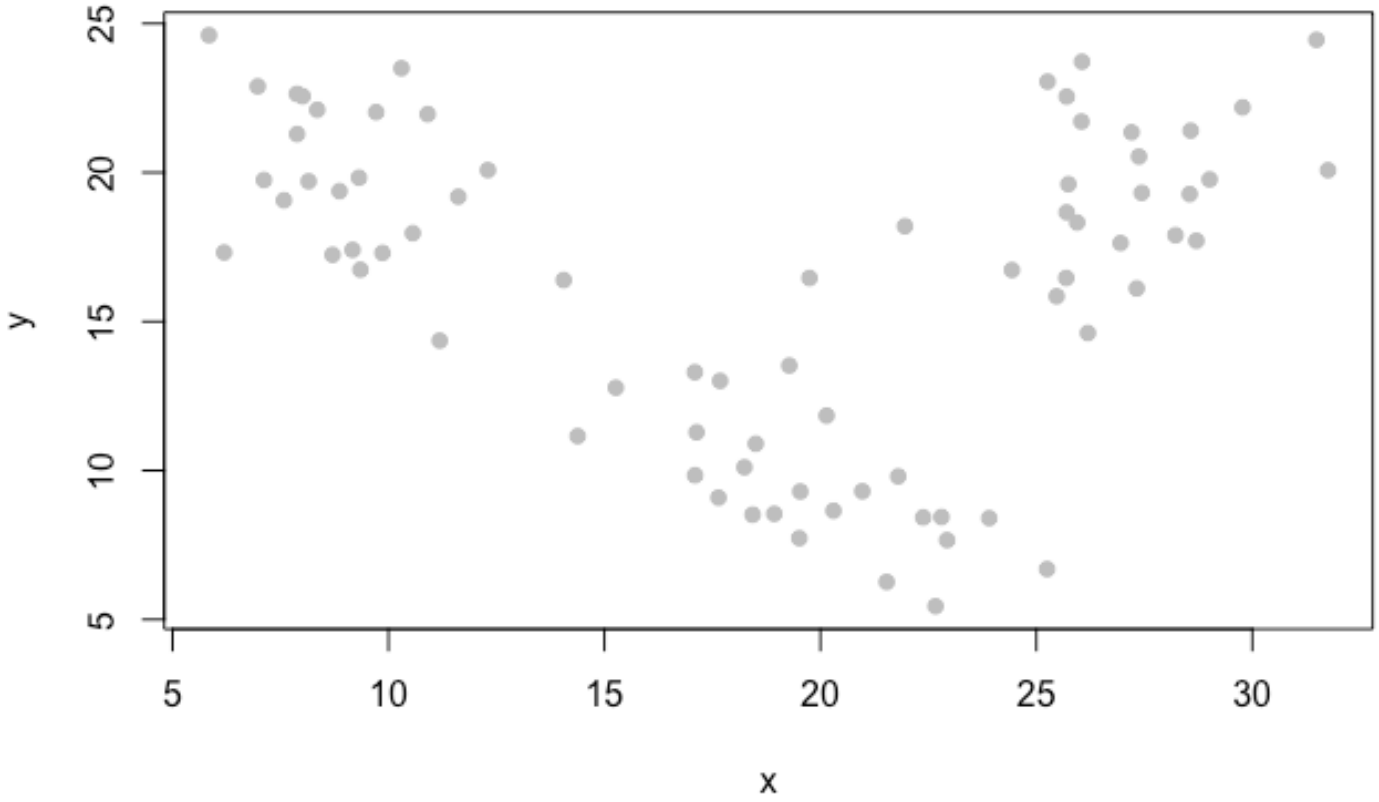
The *k* means algorithm

1. Subdivide the complete data into *k* initial groups.
2. Compute the centroids (mean vector) for each group.
3. Sequentially go through the data reassigning each case to the group with the closest centroid.
4. After reassigning a case to a new group recalculate the centroid for the original group and the new group to which it is a member.
5. Continue until there are no new reassignments of cases.

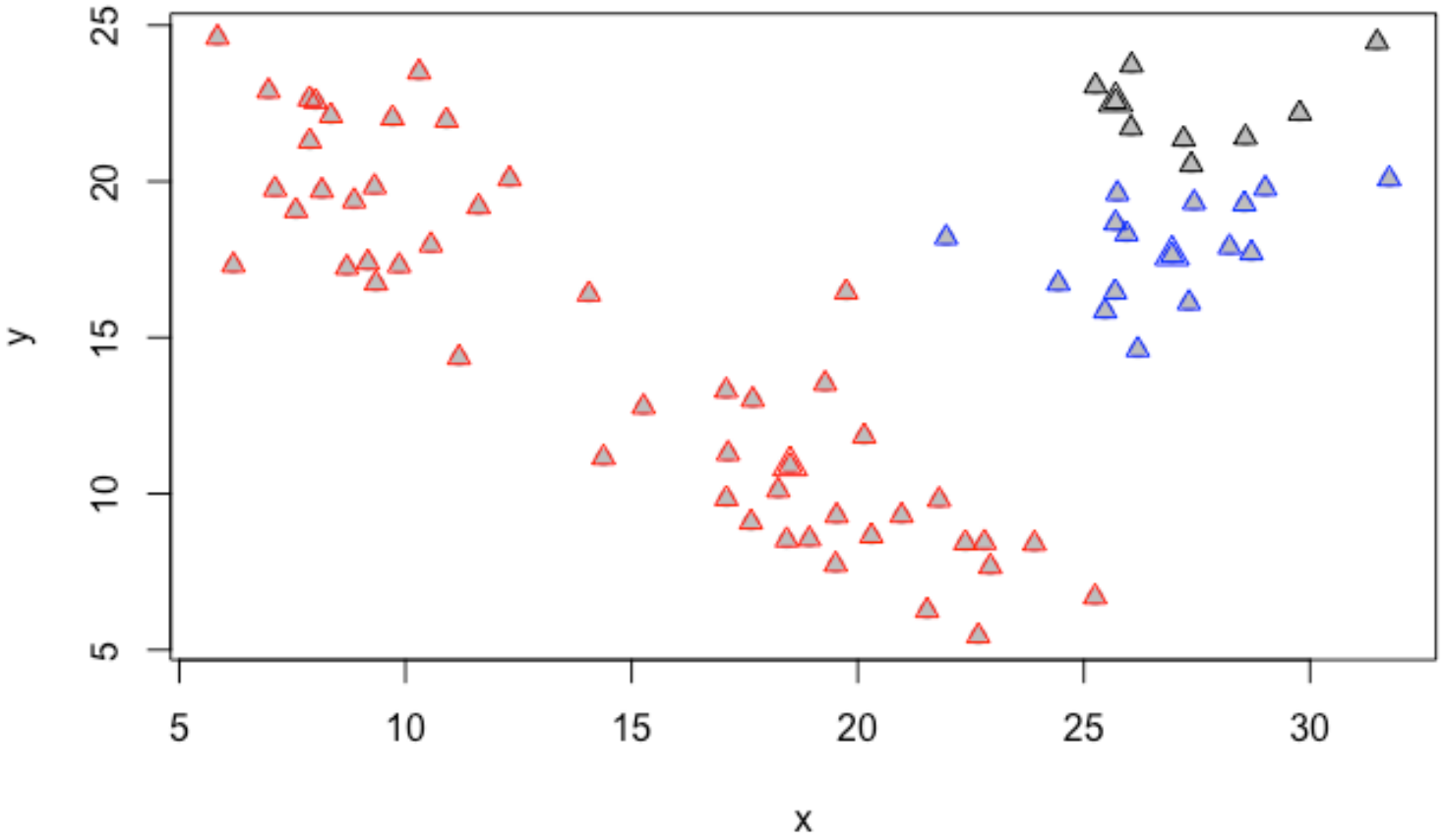
Example 1: *Simulated data*



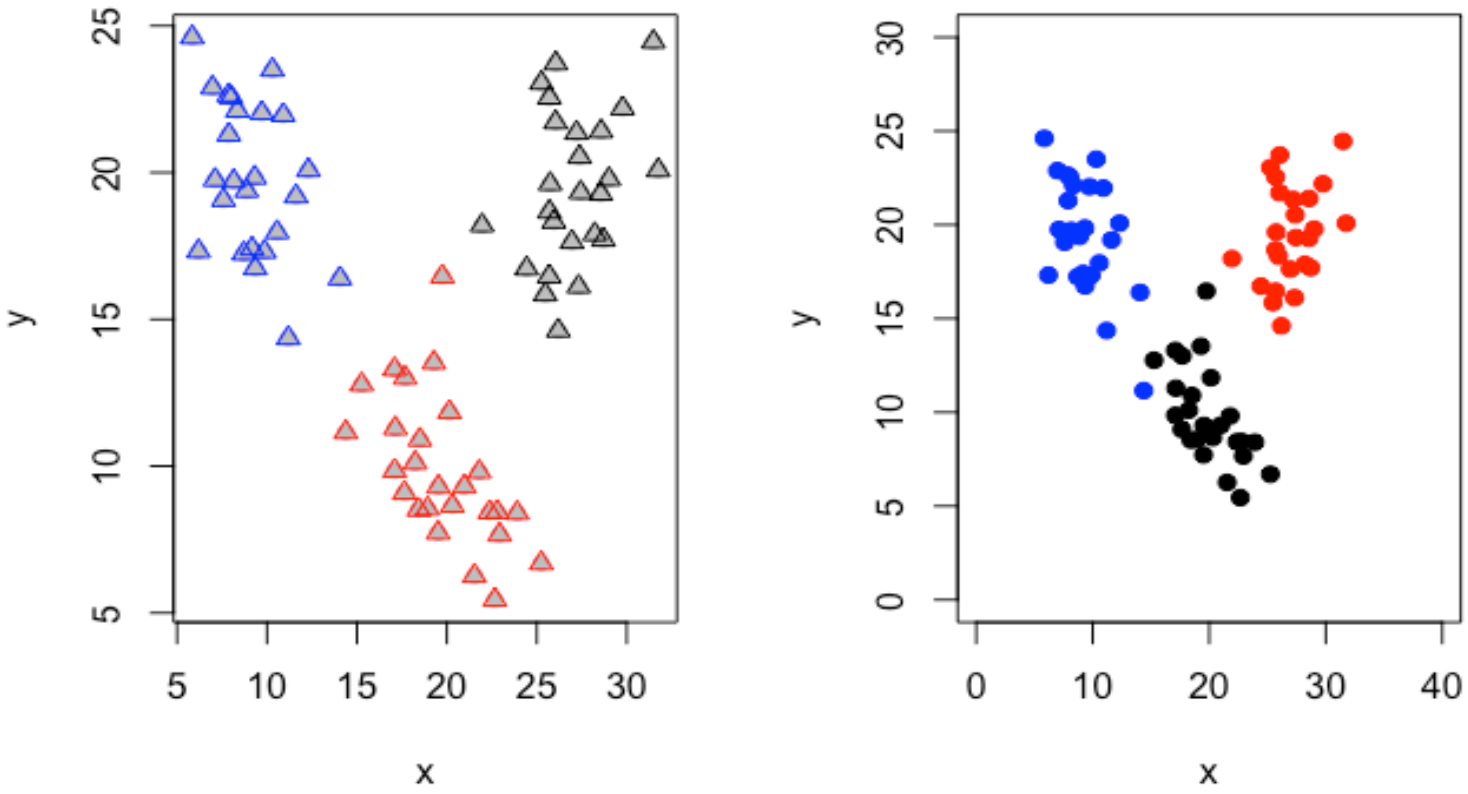
Scatterplot of the data



Graph: Initial Clustering



Graph: Final Clustering vs. True



A Comprehensive Example:
Cluster Analysis → MANOVA → Discriminant Analysis

- Example 15.4.1a on p. 532 of Rencher

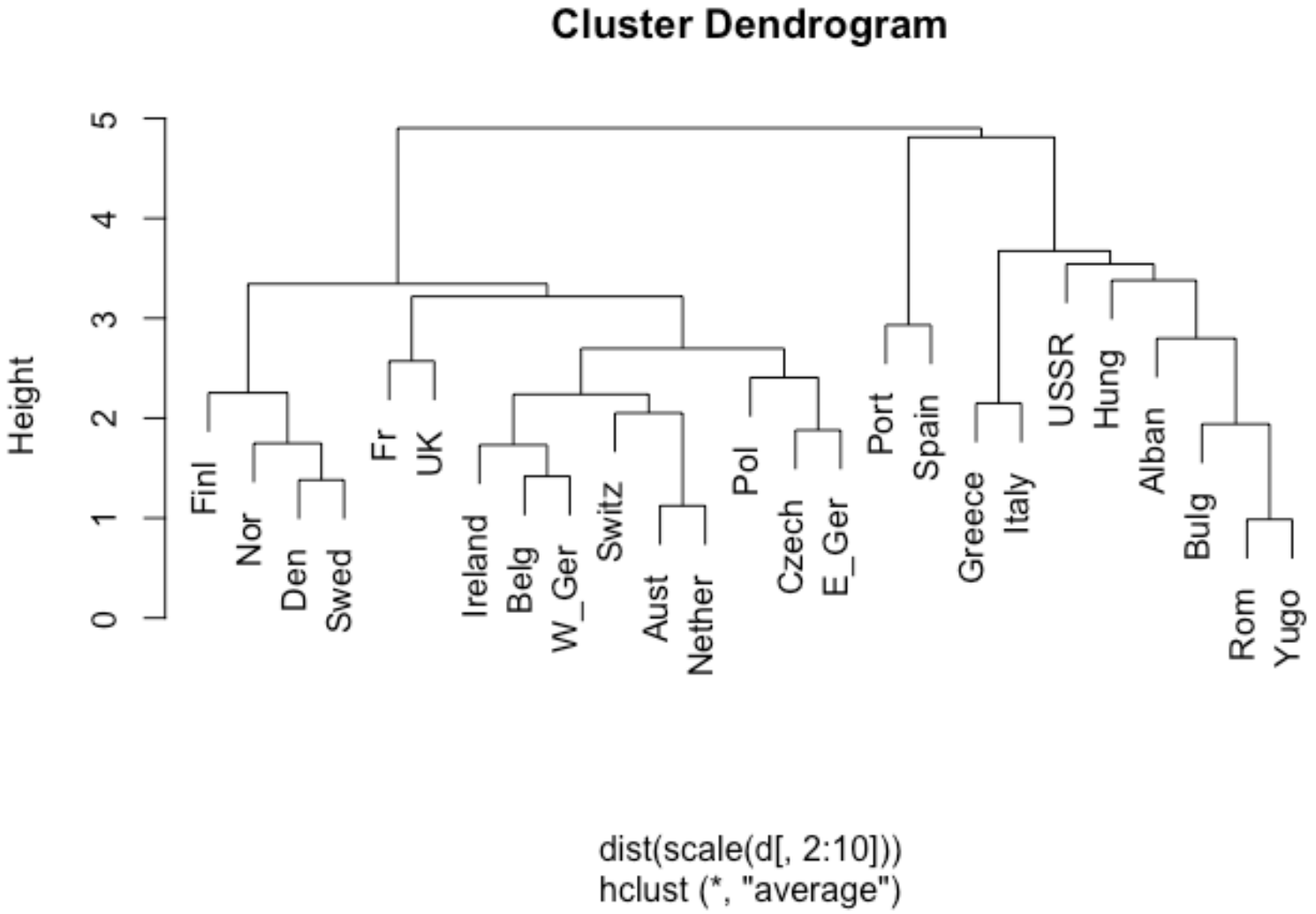
- Protein consumption in 25 European countries
- Nine food groups
- Hand et al. 1994, p. 298

The Analysis

The first step in the analysis is to perform cluster analysis to see if there are any subgroups of interest:

Both hierarchical and partitioning method (K-means) approaches were used for the cluster analysis.

Average Linkage after Rescaling



The Analysis

Let's use 4 or 5 clusters for more detailed analysis

The *k*-means method was then used (with $k = 4$ and $k = 5$) to identify members of these clusters.

The result clearly identifies different cuisines:

Mediterranean, Scandinavian, Balkans, Western, Central European

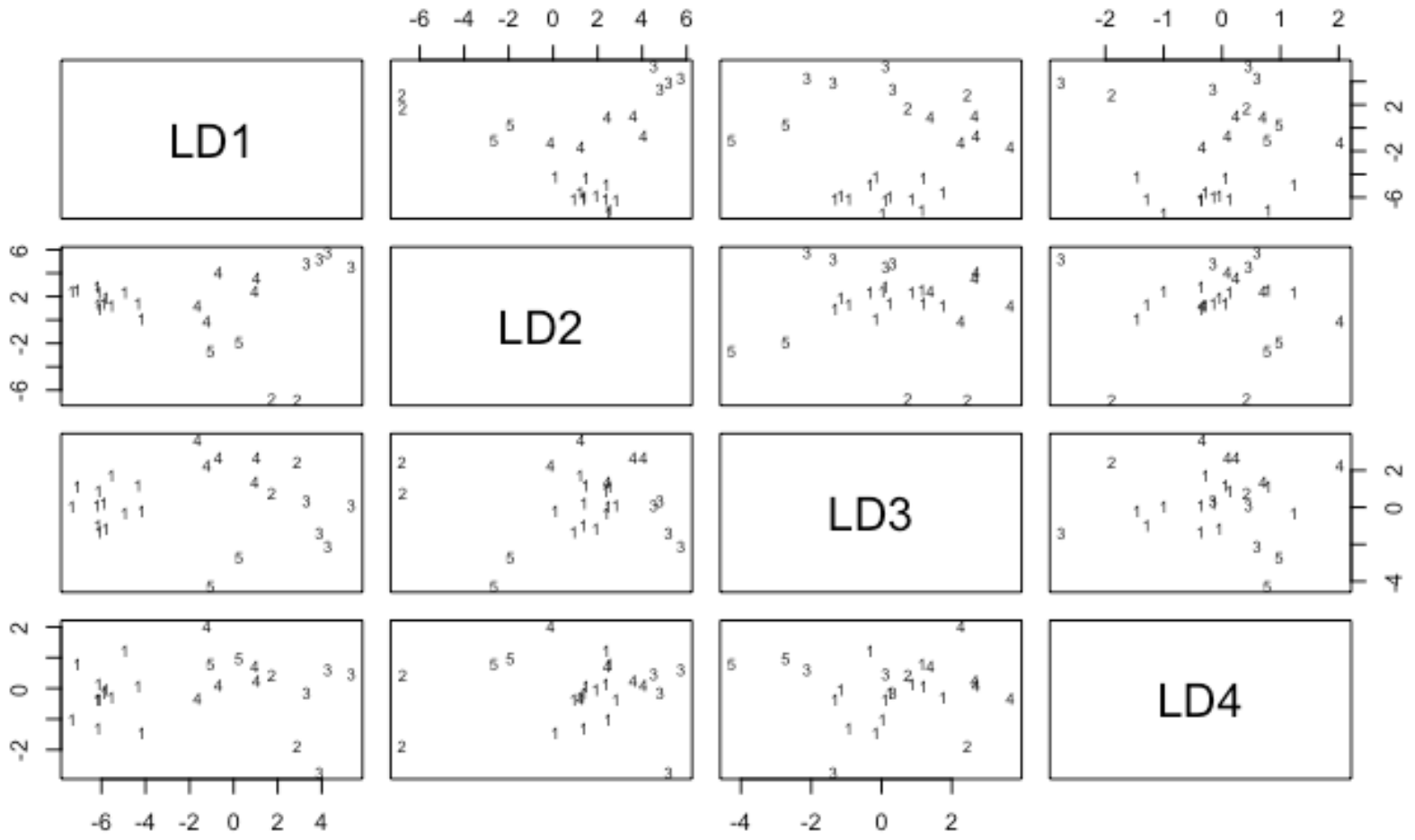
One-way MANOVA was then utilized to test for significant differences between the clusters

It was also used to identify the variables on which the differences between the clusters were most significant.

A discriminant function analysis was done to predict cluster membership and to attempt to identify the minimal set of variables used to identify cluster separation for the countries.

Coefficients of linear discriminants:

	LD1	LD2	LD3	LD4
RedMeat	-0.9040233	1.09976803	-0.05055129	-0.43472170
WhiteMeat	-1.0906298	1.42170445	0.58477156	-0.09845386
Eggs	-0.6291439	-0.89035138	-1.11157017	0.86389368
Milk	-2.0153619	0.09463257	-0.66325804	0.53222526
Fish	-0.6051214	-0.32981488	-0.51141421	-0.40115286
Cereals	0.8904206	1.95578435	-0.56552367	0.93888906
StarchyFoods	0.2648148	-0.23545731	1.74944256	0.19199041
Nuts	-0.2836815	-0.72740745	-0.25176002	-0.21220839
FruitVeg	0.1642081	-1.89289316	-0.24453012	0.73141403


LD1
LD2
LD3
LD4
-6 -4 -2 0 2 4 6
-2 -1 0 1 2
-6 -2 2 6
-6 -4 -2 0 2
-2 0 2
-4 -2 0 2
-6 -2 2
-4 -2 0 2

Discrimination performance

	Actual				
Predicted	1	2	3	4	5
1	12	0	0	0	0
2	0	2	0	0	0
3	0	0	4	0	0
4	0	0	0	5	0
5	0	0	0	0	2