Use cluster analysis to explore and analyze the given dataset as follows:

a. Use only the numerical variables (1 to 9) to cluster the 21 firms. Justify the various choices made in conducting the cluster analysis, such as weights for different variables, the specific clustering algorithm(s) used, the number of clusters formed, and so on.

b. Interpret the clusters with respect to the numerical variables used in forming the clusters.

c. Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters)

d. Provide an appropriate name for each cluster using any or all of the variables in the dataset.

15.3 **Customer Rating of Breakfast Cereals.** The dataset *Cereals.csv* includes nutritional information, store display, and consumer ratings for 77 breakfast cereals.

Data Preprocessing. Remove all cereals with missing values.

a. Apply hierarchical clustering to the data using Euclidean distance to the normalized measurements. Compare the dendrograms from single linkage and complete linkage, and look at cluster centroids. Comment on the structure of the clusters and on their stability. *Hint*: To obtain cluster centroids for hierarchical clustering, compute the average values of each cluster members, using the *aggregate()* function.

b. Which method leads to the most insightful or meaningful clusters?

c. Choose one of the methods. How many clusters would you use? What distance is used for this cutoff? (Look at the dendrogram.)

d. The elementary public schools would like to choose a set of cereals to include in their daily cafeterias. Every day a different cereal is offered, but all cereals should support a healthy diet. For this goal, you are requested to find a cluster of "healthy cereals." Should the data be normalized? If not, how should they be used in the cluster analysis?

15.4 **Marketing to Frequent Fliers.** The file *EastWestAirlinesCluster.csv* contains information on 3999 passengers who belong to an airline's frequent flier program. For each passenger, the data include information on their mileage history and on different ways they accrued or spent miles in the last year. The goal is to try to identify clusters of passengers that have similar characteristics for the purpose of targeting different segments for different types of mileage offers.

a. Apply hierarchical clustering with Euclidean distance and Ward's method. Make sure to normalize the data first. How many clusters appear?

b. What would happen if the data were not normalized?

c. Compare the cluster centroid to characterize the different clusters, and try to give each cluster a label.

d. To check the stability of the clusters, remove a random 5% of the data (by taking a random sample of 95% of the records), and repeat the analysis. Does the same picture emerge?

e. Use k-means clustering with the number of clusters that you found above. Does the same picture emerge?

f. Which clusters would you target for offers, and what types of offers would you target to customers in that cluster?