Exercise 1

Complete this exercises as a group and save your answers in a well-commented R script. Be sure to discuss each question amongst all members of the group, and also feel free to explore multiple solutions to the same problem. I expect you will need to consult class notes and R help files. You are welcome to use any printed or web-based resources, as well. Groups will be randomly selected to present their answers in the next class meeting.

1. *Exploring human diets*
   Data was collected on the proportion of human subjects' diet that fell within 6 categories. Using these data ("dietComposition.txt") address the following questions or task.
   a. What association metric would you chose for these data and why?
   b. What are the minimum and maximum similarities/distances observed across all subjects?
   c. How many diet classes or sub-groups would you define amongst the subjects?
   d. Based upon the diet content of the diet classes you've identified give each class a name.

2. *Drivers of inter-region compositional differences in insects*
   You've been given log-transformed insect species abundance data from two islands ("insectComposition1.txt" and "insectComposition2.txt"). The study was designed to compare and contrast composition across two regions (1=low altitude, 2= high altitude) on each island. The retiring entemology researcher that gave you the data mentioned that she thought very different processes were determining the differences in community composition between the two regions on the two islands. She mumbled something about abundances versus presence-absence, but then walked off. When you approached her again about the data she was too busy to help. Use what we have discussed in class and the little information you have gotten from the entomologist to explore the two datasets. Do you think different processes are driving community composition and the inter-region differences in community composition on the two islands? What patterns do you see or not see in the data from the two islands? Any ideas on the ecological processes that might generate any patterns you see?

3. *Sources of methodological variation*
   We've talked in class about how data transformation (e.g. quantitative vs. binary), association metric, and clustering algorithm can all alter our interpretation of multivariate data. Which of these are our interpretations most sensitive to? Write down an initial ranking of the importance of 1) quantitative vs. binary data, 2) choice of association metric, and 3) choice of clustering algorithm for determining the patterns we see in multivariate data. Now evaluate the provided dataset ("observations.txt") with a variety of combinations of quantitative vs. binary, appropriate association metrics, and clustering algorithms. Which analysis choices most influence what we see when looking at a dendrogram of the observations? Why? Was this the same or different than your expectations?