Statistics 407 Homework 2

Due date: Wednesday, September 26, 2012, in class.
Although you may work together, you should hand in an individual solution.

The dataset cereal-data-expo-cold.csv is available on the course web page, contains nutritional data for 74 brands of breakfast cereal. Read the information sheet to find out more about the data.

1. Make a summary of the cereal data. Which variables have missing values? Remove the cases that have any missing values (not something usual to do, but will work best for this data.) How many cereals remain in the data?

2. There are several ways to make the nutritional information comparable. The information sheet suggests that a reasonable way to do this is by volume, because people tend to just fill their cereal bowls. Scale the values so that they are measurements per cup, ie divide the measurements by the number of cups.

3. Load the resulting data set into ggobi, and answer these questions.

   (a) Which cereal has the highest number of Calories? Lowest?

   (b) Which cereal has the highest number of Vitamins? Lowest?

   (c) From a scatterplot matrix, which cereals are outliers in calories and carbs?

   (d) In the parallel coordinate plot, which cereal is outlying on sodium and potassium?

   (e) Describe how Kelloggs cereals (k) differ from General Mills cereal (G), if at all.

4. Conduct a principal component analysis of the nutritional data, and answer the following questions.

   (a) Explain why you are working with the correlation or covariance matrix. Summarize the results.

   (b) How many PCs would you recommend using? Justify your answer by using a scree plot, proportion of variance explained and interpreting the coefficients of the eigenvectors.

   (c) Summarize the PCA, to your selection of number of PCs, by tabulating the eigenvectors, variance and cumulative proportion of variance explained.