PROBLEMS

4.1 **Breakfast Cereals.** Use the data for the breakfast cereals example in Section 4.8 to explore and summarize the data as follows:

a. Which variables are quantitative/numerical? Which are ordinal? Which are nominal?

b. Compute the mean, median, min, max, and standard deviation for each of the quantitative variables. This can be done through R's *sapply()* function (e.g., *sapply(data, mean, na.rm = TRUE)*).

c. Use R to plot a histogram for each of the quantitative variables. Based on the histograms and summary statistics, answer the following questions:

i. Which variables have the largest variability?

ii. Which variables seem skewed?

iii. Are there any values that seem extreme?

d. Use R to plot a side-by-side boxplot comparing the calories in hot vs. cold cereals. What does this plot show us?

e. Use R to plot a side-by-side boxplot of consumer rating as a function of the shelf height. If we were to predict consumer rating from shelf height, does it appear that we need to keep all three categories of shelf height?

f. Compute the correlation table for the quantitative variable (function *cor()*). In addition, generate a matrix plot for these variables (function *plot(data)*).

i. Which pair of variables is most strongly correlated?

ii. How can we reduce the number of variables based on these correlations?

iii. How would the correlations change if we normalized the data first?

g. Consider the first PC of the analysis of the 13 numerical variables in Table 4.11. Describe briefly what this PC represents.

4.2 **University Rankings.** The dataset on American college and university rankings (available from www.dataminingbook.com) contains information on 1302 American colleges and universities offering an undergraduate program. For each university, there are 17 measurements that include continuous measurements (such as tuition and graduation rate) and categorical measurements (such as location by state and whether it is a private or a public school).

a. Remove all categorical variables. Then remove all records with missing numerical measurements from the dataset.

b. Conduct a principal components analysis on the cleaned data and comment on the results. Should the data be normalized? Discuss what characterizes the components you consider key.

4.3 **Sales of Toyota Corolla Cars.** The file *ToyotaCorolla.csv* contains data on used cars (Toyota Corollas) on sale during late summer of 2004 in the Netherlands. It has 1436 records containing details on 38 attributes, including Price, Age, Kilometers, HP, and other specifications. The goal will be to predict the price of a used Toyota Corolla based on its specifications.

a. Identify the categorical variables.

b. Explain the relationship between a categorical variable and the series of binary dummy variables derived from it.