STAT 475-575 **Assignment 1** **Name**_____

Suggested Reading: Everitt and Hothorn, Chapters 1 and 2.

Written Assignment: Due Friday, September 17, 2020 by 11:59pm.

Although you may work together on written assignments to review notes, review matrix algebra, or to implement statistical software, you should each submit an individual solution. The interpretation of data and results of analyses should be your own

1. Consider the following data matrix in which information on six response variables is reported for each of seven subjects.

$$X = \begin{bmatrix} 2 & 5 & 1 & 3 & 0 & 6 \\ -1 & 1 & 0 & -3 & 2 & -3 \\ 6 & 9 & 5 & 7 & 8 & 5 \\ -2 & 2 & 0 & -4 & -1 & -5 \\ 5 & 8 & 4 & 1 & 9 & 6 \\ 1 & 5 & -2 & 0 & 7 & 3 \\ 6 & 8 & 3 & 1 & 6 & 5 \end{bmatrix}$$

   (a) Report the values of n, the number of rows in X, and p, the number of columns in X,

   (b) Report the value of x_{32} the entry in the third row and second column of X.

   (c) Evaluate the vector of sample means for the six variables, $\mathrm{X}=\begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_6 \end{bmatrix}$.

   (d) Evaluate the sample covariance matrix S.

   (e) Evaluate the matrix of sample correlations R.

   (f) Evaluate X', the transpose of the data matrix X.

   (g) Evaluate Z the matrix of standardized scores for these data.

   (h) Evaluate the covariance matrix for Z. How is it related to the correlation matrix for X? How is it related to the correlation matrix for Z?

2. The following data consist of measurement made on the levels of three liver enzymes (U/L): aspartate aminotransferase (X_1), alanine aminotransferase (X_2), and glutamate dehydrogenase (X_3) in n=10 patients diagnosed with aggressive chronic hepatitis.

Patient	X_1	X_2	X_3
1	31	63	4
2	32	56	6
3	50	59	9
4	56	72	7
5	39	87	9
6	46	95	8
7	29	57	5
8	40	50	3
9	29	44	4
10	24	42	3

These data are part of a larger set of data reported by Plomteux (1980, Clin. Chem. 26, 1897-1899). They are posted on Canvas in the file **liver_enzymes.csv**.

(a) Evaluate the sample mean vector $\underset{\sim}{\bar{X}} = (\bar{X}_1, \bar{X}_2, \bar{X}_3)'$.

(b) Evaluate the sample covariance matrix S.

(c) Evaluate the sample correlation matrix R. Which variables have significant correlations?

(d) Construct a scatterplot matrix for these data. Does it appear that all of the pairwise relationships are linear? Explain.

(e) Compute the generalized sample variance, |S|. If the three variables have a multivariate normal distribution, explain why |S| is reasonable measure of overall variation in the values of the three variables.

(f) Compute the total sample variance, trace(S). Trace(S) is also a possible measure of overall variance in the three variables. Which of the measures of overall variability, trace(S) or |S|, ignores the correlations between the variables?

4. The data set **cereal.csv**, available in the data folder on the course Canvas page, contains nutritional information for 77 brands of breakfast cereal. More information about the variables on the data file is presented on pages 4 and 5, at the end of this problem.

a. Create a table that shows the names of the manufacturers and how many cereals are produced by each manufacturer. (Note the manufacturers are represented by their first initial: A=American Home Food Products, G=General Mills, K=Kelloggs, N=Nabisco, P=Post, Q=Quaker Oats, R=Ralston Purina.)

b. Data on the number of cups per serving is negative 1 for some cereals, possibly because that information was not provided in the nutritional information printed on the cereal box. Missing information is indicated by a negative number. Use the subset function in R, which was illustrated in Lab 2, to create a new data set containing only cereals with no missing data (this will be used for the rest of the problem). Report the number of cereals for which information is missing and the proportion of missing values.

c. There are several ways to make the nutritional information comparable across cereals. The information provided at the end of this problem suggests that a reasonable way to do this is by volume instead of by weight, because people just tend to fill their cereal bowls. Scale the values of the calories, protein, fat, sodium, fiber, carbs, sugar, potassium, and vitamins variables to put them in units of per cup of cereal, i.e., divide the measurements by the number of cups in a serving. Use the head( ) function in R to display the rescaled values for the first six cereals in the data frame and include that information in this part of your report.

d. Apply R and ggplot package to the rescaled data from part (c) to answer the following questions.

   i) Which cereal has the highest number of calories per cup? Which has the lowest?

   ii) Which cereal has the highest number of vitamins per cup? Which has the lowest?

   iii) Create a scatterplot matrix of the per cup values of the calories, carbs, fat, fiber, sodium, and potassium variables. Which cereals are "outliers" with respect to either calories or carbs per cup or both? Which pairs of variables have nearly linear associations, which do not? Which pairs of variables have positive associations, which have negative associations?

   iv) Create and interpret side-by-side box plots for the sugar content variable, with one box plot for each shelf position, low, middle, or high, in grocery stores. How do the distributions of the per cup sugar content of cereals differ with respect shelf position?

   v) Create and interpret side-by-side box plots for the fiber content variable, with one box plot for each shelf position, low, middle, or high, in grocery stores. How do the distributions of the per cup fiber contents of cereals differ with respect shelf position?

   vi) Create a parallel coordinate plot using per cup values of calories, protein, fat, sodium, fiber, carbs, sugar, potassium, and vitamins. Color code the profiles on the plot with respect to position on the shelf (top, middle, bottom) in grocery stores What insights into the cereal data, if any, are provided by the parallel coordinate plot?

   vii) Describe how Kelloggs cereals (K) differ from General Mills cereals (G), if at all. You may want to use the subset function in R to create a new data frame with just the Kelloggs and General Mills cereals. Apply ggplot2 functions, making good use of color. Insert important plots in your report to support your conclusions.

==========

Breakfast Cereal Information

The breakfast cereal data set contains nutritional information for 77 different breakfast cereals. It was used for the 1993 Statistical Graphics Exposition as a challenge data set. We retrieved this data from StatLib at CMU. These data were obtained from the nutritional labels on the cereal boxes and is in CSV format. Information on calories, protein, sodium, dietary fiber, complex carbohydrates, sugars, potassium, vitamins, and weight are recorded on a per serving basis.

The variables are:

- Cereal name;
- manufacturer (e.g.,K for Kellogg's);
- type (cold/hot);
- calories (number);
- protein (g);
- fat (g);
- sodium (mg);
- dietary fiber (g);
- complex carbohydrates (g);
- sugars (g);
- display shelf (1, 2, or 3, counting from the floor);
- potassium (mg);
- vitamins and minerals (0, 25, or 100 percent, respectively);
- weight (in ounces) of one serving (serving size);
- cups per serving.

The data set has information on seventy-seven commonly available breakfast cereals that was obtained from government mandated FDA food labels. What are you getting when you eat a bowl of cereal? Can you get a lot of fiber without many calories? Are there differences in cereals that are displayed on the high, low, and middle shelves of grocery stores? The good news is that none of the cereals in the data file has any cholesterol, and manufacturers rarely use artificial sweeteners and colors, nowadays. However, there is still a lot of information for the consumer to understand while choosing a good breakfast cereal.

Cereals vary considerably in their densities and listed serving sizes. Thus, the serving sizes listed on cereal labels (in weight units) translate into different amounts of nutrients in your bowl. Most people simply fill a cereal bowl (resulting in constant volume, but not weight). The information about serving size in cups provides another way to make information comparable across cereals.

Here are some facts about nutrition that might help you in your analysis. Nutritional recommendations are from the references at the end of this document:

* Adults should consume between 20 and 35 grams of dietary fiber per day.

* The recommended daily intake (RDI) for calories is 2200 for women and 2900 for men.

* Calories come in three food components. There are 9 calories per gram of fat, and 4 calories per gram of carbohydrate and protein.

* Overall, in your diet, no more than 10% of your calories should come from simple carbohydrates (sugars), and no more than 30% should come from fat. The RDI of protein is 50 grams for women and 63 grams for men. The balance of calories should be consumed in the form of complex carbohydrates (starches).

* The average adult with no defined risk factors or other dietary restrictions should consume between 1800 and 2400 mg of sodium per day.

* The type and amount of milk added to cereal can make a significant difference in the fat and protein content of your breakfast.

References:

National Research Council, 1989a. "Diet and Health: Implications for Reducing Chronic Disease Risk". National Academy Press, Washington, D.C.

National Research Council, 1989b. "Recommended Dietary Allowances, 10th Ed." National Academy Press, Washington, D.C.

National Cancer Institute, 1987. "Diet, Nutrition, and Cancer Prevention: A Guide to Food Choices," NIH Publ. No. 87-2878. National Institutes of Health, Public Health Service, U.S. Department of Health and Human Service, U.S. Government Printing Office, Washington, D.C.