Lab 1

STA 216

The file “lab 1 - cereal.xlsx” on Blackboard consists of a dataset containing different kinds of cereal. The variables in the dataset include:

* name: Name of cereal
* mfr: Manufacturer of cereal, with values: A = American Home Food Products, G = General Mills, K = Kelloggs, N = Nabisco, P = Post, Q = Quaker Oats, R = Ralston Purina
* type: with values C = cold, H = hot
* calories: calories per serving
* protein: grams of protein
* fat: grams of fat
* sodium: milligrams of sodium
* fiber: grams of dietary fiber
* carbo: grams of complex carbohydrates
* sugars: grams of sugars
* potass: milligrams of potassium
* vitamins: vitamins and minerals – taking values 0, 25, or 100, indicating the typical percentage of daily recommended amount
* shelf: display shelf (1, 2, or 3, counting up from the floor)
* weight: weight in ounces of one serving
* cups: number of cups (unit of volume) in one serving
* rating: a rating of the cereals (not sure where this comes from)

1. After saving the file “lab 1 - cereal.xlsx” from Blackboard into a folder on your computer, read the data into SAS using PROC IMPORT. Include a comment that explains what the PROC IMPORT does.
   1. Copy and paste PROC IMPORT code (including the comment) into your writeup.
   2. Looking at the dataset in the “Output Data” tab, answer the following:
      1. How many unique kinds of cereal are there in the dataset?
      2. How many variables are there?
      3. Separate the variables into numeric variables and character variables.
2. Use PROC PRINT to:
   1. Make a data table showing all the variables for cereals manufactured by Kelloggs. Use the Snipping Tool to copy and paste this data table into your writeup.
   2. Again focusing on only Kelloggs cereals, make a data table showing only the variables name, calories, protein, fat, carbo, and sugars. (Again, copy and paste this table into your writeup.)
   3. Focusing on only Kelloggs cereals and showing the same variables as before, make the data table only contain cereals with greater than 10 g of sugar per serving. (Again, copy and paste.)
   4. Of the cereals in the table from (c), which are NOT marketed primarily to kids?
3. Before we go any further, make labels for each of the following variables so the data table would be more clear to someone unfamiliar with the variable names: mfr, type, carbo, potass, and weight.
   1. Copy and paste the code used to do this.
   2. Name one more variable to which you think it would be appropriate to add a label and why.
   3. Copy and paste the output from PROC PRINT for all 16 variables (with the labels showing up) for cereals manufactured by Post.
4. Make formats for the variables mfr, type, and shelf based on the descriptions of these variables at the beginning of the lab.
   1. Copy and paste the code that does this.
   2. Copy and paste the PROC PRINT output that displays these three variables (which should have formatted values now) and the variable name for cereals manufactured by Quaker Oats.
5. Use PROC SORT to sort the data by amount of fiber from highest to lowest. What word is common to the names of 7 of 8 of the highest fiber cereals? Copy and paste output from PROC PRINT that supports your answer.
6. Identify the only three hot cereals in the list. Copy and paste output from SAS that supports your answer.  
   1. Which cereal has the smallest number of cups per serving? Copy and paste SAS output that supports your answer.
   2. Based on (a), why is this cereal sold in smaller boxes than most others?
7. Use PROC SORT to sort by shelf and then by sugar. What shelf are the sugary kids cereals usually on? Why do you think this is?
8. First use PROC SORT to sort by rating. Focusing on fiber and sugar, describe the kinds of cereals that tend to have the lowest and highest ratings.
9. It is known that each gram of fat contains 9 calories and each gram of protein and carbohydrates (whether complex carbohydrates or sugars) contains 4 calories.
   1. Use this information to create a new variable named calc\_calories that calculates calories per serving from the variables fat, protein, carbo, and sugars.
      1. Show the SAS code used to do this.
      2. Copy and paste output from PROC PRINT that shows the variables name, fat, protein, carbo, sugars, calories, and calc\_calories for cereals manufactured by Nabisco.
      3. Looking at the above table, what seems to be causing the difference between calories and calc\_calories in for all cereals except one.
   2. Calculate another new variable named diff\_calories that is equal to calories minus calc\_calories. Copy and paste a table like in 10aii that also includes this new variable.
   3. Use PROC SORT by diff\_calories and look at the entire dataset.
      1. All of the cereals have calories within at least 20 of calc\_calories except for one. Which one is it?
      2. What observations for this cereal make the value of calc\_calories suspect?
10. I remember commercials like these from when I was growing up: <https://www.youtube.com/watch?v=f2qBxY-xm8A> (or, for a funny SNL version: <https://www.youtube.com/watch?v=Ku42Iszh9KM>).
    1. Create a variable named nbowls\_fiber that calculates how many servings of each cereal it would take to have the same amount of fiber as in the brand All-Bran with Extra Fiber. Also, on the next line of the data step, use the code   
       round\_nbowls = round(nbowls\_fiber);   
       to round the result to the closest number. Copy and paste the code that does this.
    2. Copy and paste the PROC PRINT output that contains the variables name, fiber, and these two new variables for all Kelloggs cereals.
    3. There should be two cereals that have periods the values of these two variables. In SAS, these periods stand for “missing values”. Considering how the variables were calculated, why does these two cereals have missing values?