ANLY 6100
ASSIGNMENT #1: 100 PTS

The goal of this homework is to introduce you to coding in R, which also includes some review of data concepts you've seen in previous courses (histograms, correlation, etc.). Most of what you will need for this assignment in R can be found from our in-class tutorial and the "Introduction to R Coding" file, though it may not be obvious at first. You do not have to use anything we've done in class if you have (or find) another way to do it. As long as it works, you're welcome to use it—with the caveat that yes, you are required to fully complete this assignment in R! Be sure to include the code you used, so I can see how you got your answers.

Cereal Data Analysis

How bad are your cereal choices? A famous data set collected/cleaned by Petra Isenberg, Pierre Dragicevic and Yvonne Jansen includes approximately 80 different kinds of cereal from several manufacturers. The data set includes some important characteristics of the nutrition of these cereals:

- *name*: Name of cereal
- *mfr*: Manufacturer of cereal (A = American Home Food Products, G = General Mills, K = Kellogg's, N = Nabisco, P = Post, Q = Quaker Oats, R = Ralston Purina)
- *type*: hot or cold (H = Hot, C = Cold)
- *calories*: calories per serving
- *protein*: grams of protein per serving
- *fat*: grams of fat per serving
- *sodium*: milligrams of sodium per serving
- *fiber*: grams of dietary fiber per serving
- *carbo*: grams of complex carbohydrates per serving
- *sugars*: grams of sugars per serving
- *potass*: milligrams of potassium per serving
- *vitamins*: amount of vitamins and minerals present in the cereal (0%, 25%, or 100%, indicating the typical percentage of FDA recommended)
- *shelf*: display shelf (1, 2, or 3, counting from the floor)
- *weight*: weight in ounces of one serving
- *cups*: number of cups in one serving
- *rating*: a rating of the cereals (possibly from Consumer Reports?)

The data can be found in the accompanying file **cereal.csv** (posted on Canvas). Imagine you are working in the cereal industry, and you have been tasked with investigating the nutritional content of current major cereal brands. Use the data set with R to answer the following questions. For each question, include any output and the line(s) of R code you used to answer the question. For example:

0) (0 points) What are the variable names in the cereal data set?

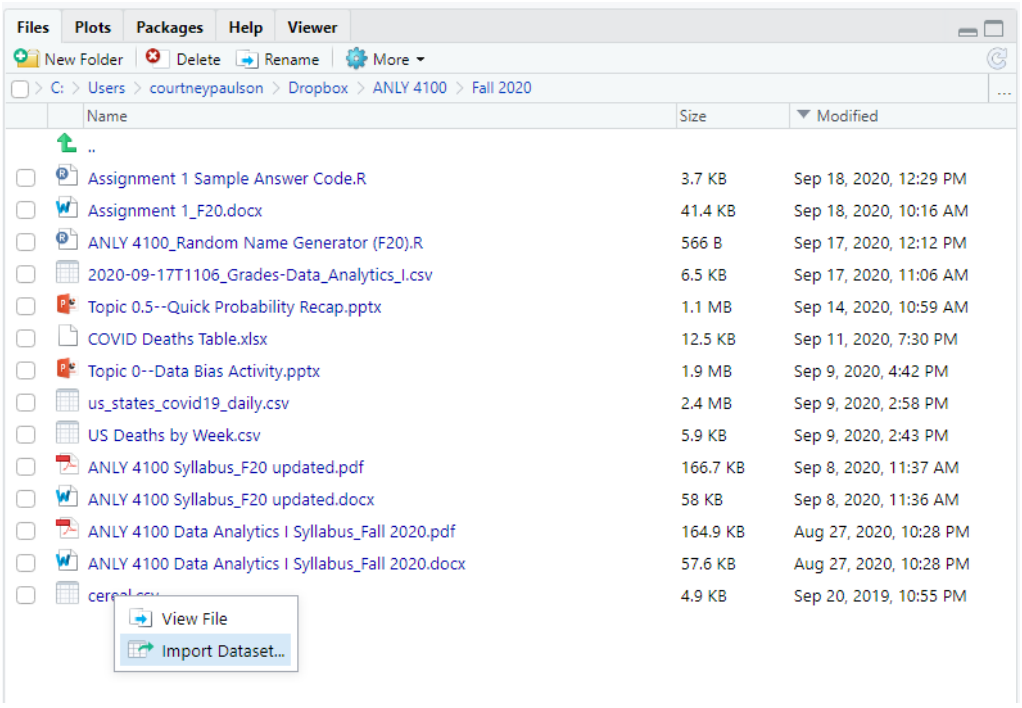
Answer: "name", "mfr", "type", "calories", "protein", "fat", "sodium", "fiber", "carbo", "sugars", "potass", "vitamins", "shelf", "weight", "cups"

R Code: `colnames(cereal)`

As a possibly helpful hint, you can load data into R in a variety of ways. If you are using regular R (not RStudio), the most common way is to use the read.csv() function:

read.csv("C:/Users/Me/ANLY 4100/cereal.csv")

You can also use that in RStudio, but RStudio gives you an additional option. Find the file wherever it is stored on your computer in the "Files" window, then click on the file to choose the "Import Dataset" option:



Another possibly helpful reminder/tip: You can access a particular variable in R by either subsetting (for example, the first variable in this data set could be accessed using cereal[,1]), or you can call the variable by name (using cereal$name).

Assignment

Please answer all questions in the dedicated space and upload on Canvas. Please ensure that your numbering of questions matches those below. Include any R code you used to answer each question with your response. If you are asked to include output, please include the output with the appropriate question. Remember: you are allowed to work with others in the class on this assignment, but don't forget to include their names in the last question!

1) **(5 points) Remove an existing variable**: The last variable in the data set, *rating*, is of unknown origin. Since we don't know who was rating the cereals or where this rating came from for sure, remove it from the data set *without creating an entirely new data set*.

   R Code:

2) **(15 points) Create new variables**:
   a. Create a new variable, *cpc*, where cpc is the calories per cup of each cereal
   b. Create a new variable, *mfr2*, where mfr2 is "Kelloggs" if the manufacturer of a cereal is Kellogg's, "GM" if the manufacturer is General Mills, and "Other" if the manufacturer is not Kellogg's or General Mills.

   R Code:

3) **(10 points) Plot a histogram in R**: Start by plotting a histogram of the calories of all cereals. Make sure the histogram:
   a. Is dark blue
   b. Has the x-axis labeled "Number of Calories"
   c. Has the y-axis labeled "Number of Cereals"
   d. Is titled "Calories in Common Cereals" in dark red
   e. Includes labels of how many cereals are in each bar of the histogram

   Histogram:

   R Code:

4) **(10 points) Plot a boxplot in R**: Create a boxplot that separates out the number of complex carbohydrates in a serving (y-axis) by the type of cereal (the x-axis). One of your fellow cereal researchers believes hot cereals have more carbs per serving than cold cereals do. Does your boxplot support this? Explain your answer.

   Answer:

   Boxplot:

   R Code:

5) **(15 points) Plot a more difficult boxplot in R**: Create a boxplot that separates out the calories per cup (y-axis) by type of cereal (the x-axis). Make sure the boxplot:
    a. Is adjusted for width (that is, the width of the boxplot indicates how many cereals are hot vs. cold)
    b. Has the x-axis labeled "Type of Cereal (C = Cold, H = Hot)"
    c. Has the y-axis labeled "Calories Per Cup"
    d. Only shows the cereals with calories per cup between 0 and 300

    Boxplot:

    R Code:

6) **(25 points) Answer the following questions about the data set. For each question, include some evidence gathered from R that supports your conclusion. Note this can be any evidence, as long as it was generated by R and agrees with your answer! For example, if you were asked to find the maximum value of a variable called "Z", you could use the R code `max(Z)` or you could show the summary of the variable Z where the maximum value is listed.**

    a. Which cereal has the highest number of calories per cup?

        Answer:
        Evidence:

    b. How many of the cereals in this data set were manufactured by Post?

        Answer:
        Evidence:

    c. What percentage of the cereals in this data set were not manufactured by either Kellogg's or General Mills?

        Answer:
        Evidence:

    d. How many cereals in this data set contain no fiber?

        Answer:
        Evidence:

    e. Which variables in this data set were not loaded into R as numeric variables?

        Answer:
        Evidence:

7) **(20 points) Conditional Probability**: Answer the following questions about probability using the cereal data set.

   a. What is the probability a random cereal (in this data set) has at least 100 milligrams of potassium per serving?

      Answer:
      R Code/Evidence:

   b. What is the probability a random cereal (in this data set) has at least 100 milligrams of potassium per serving given it was manufactured by Post?

      Answer:
      R Code/Evidence:

   c. Consider your answers to parts (a), (b), and 6(b) above. Do you think Post is better than other manufacturers about including potassium in their cereals or not? Explain your answer.

      Answer:

      R Code (and/or Evidence):

8) **Did you work with anyone on this assignment? If so, include their name(s) here.**