Lab 1: Cereals

The dataset cereal.xlsx contains data on 80 Breakfast Cereals (https://www.kaggle.com/code/jeandsantos/breakfast-cereals-data-analysis-and-clustering/data)

Variables in the dataset:

Name: Name of cereal

mfr: Manufacturer of cereal (A = American Home Food Products, G = General Mills, K = Kelloggs, N = Nabisco, P = Post, Q = Quaker Oats, R = Ralston Purina)
type: (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 - 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)

Questions:

(i) Load the dataset into R Studio. Create a frequency table and barchart of the variable mfr (manufacturer). Which Manufacturer has the most number of cereals?

(ii) Replace any -1 in the dataset with NA (which the term R uses for missing data)

(iii) Create histograms of fiber, carbo and ratings and suggest appropriate summary statistics for these variables.

(iv) Create a matrix of scatterplots of the variable calories, protein, carbo, sugars, fat, rating. Create a matrix of Pearson correlations for these variables. Which variable appears to have the strongest linear relationship with ratings? Interpret this relationship.

(v) Create a new dataset with Cold Cereals only (i.e. select type = “C”).

(vi) Fit and interpret a multiple regression model for ratings in Cold Cereals using the predictor variables calories, fiber, fat, protein, sugar and potassium.

(vii) Create residual plots to check if the regression model in (vi) satisfies the assumptions of regression