Homework 4
of
STAT 3355 Data Analysis for Statisticians & Actuaries

Due: 11:30 am

March 21 (Monday), 2022

Let's work on the `UScereal` dataset in the package `UsingR`. You can use the following code to load the data. Use necessary code to read the description of the dataset, which contains 65 samples and 11 variables.

```
# Install the package if you never did
install.packages("UsingR")

# Load the pacakge
library(UsingR)

# Load the UScereal dataset
data("UScereal")
```

Problem 1 ($1 \times 3 = 3$ points)

Let's first clean the data:

(a) Replace the levels of the factor variable `mfr` to their full names, i.e. "G" for General Mills, "K" for Kelloggs, "N" for Nabisco, "P" for Post, "Q" for Quaker Oats, and "R" for Ralston Purina. (Hint: use the function `levels()`)

(b) Turn the variable `shelf` to a factor variable, of which levels are "1" for low, "2" for middle, and "3" for upper

(c) Create a new variable named `Product` for the product name. (Hint: use the function `rownames()`) to access the name for each sample

Hint: You should get the following response after applying the function `str()` on the cleaned dataset

```
'data.frame':  65 obs. of  11 variables:
$ mfr      : Factor w/ 6 levels "General Mills",..: 3 2 2 1
    2 1 6 4 5 1 ...
```

```
$ calories : num  212 212 100 147 110 ...
$ protein  : num  12.12 12.12 8 2.67 2 ...
$ fat      : num  3.03 3.03 0 2.67 0 ...
$ sodium   : num  394 788 280 240 125 ...
$ fibre    : num  30.3 27.3 28 2 1 ...
$ carbo    : num  15.2 21.2 16 14 11 ...
$ sugars   : num  18.2 15.2 0 13.3 14 ...
$ shelf    : Factor w/ 3 levels "Lower","Middle",..: 3 3 3
  1 2 3 1 3 2 1 ...
$ potassium: num  848.5 969.7 660 93.3 30 ...
$ vitamins : Factor w/ 3 levels "100%","enriched",..: 2 2 2
   2 2 2 2 2 2 2 ...
$ product  : chr  "100% Bran" "All-Bran" "All-Bran with
  Extra Fiber" "Apple Cinnamon Cheerios" ...
```

Problem 2 ($1 \times 7 = 7$ points)

Let's analyze the data. For each figure, please make it complete/readable, in other words, it should include all the label information, title, and legend if necessary.

(a) Calculate the Pearson correlation coefficient between `calories` and each of the seven nutrition facts, `protein`, `fat`, `sodium`, `fibre`, `carbo`, `sugars`, and `potassium`, and show their numbers.

(b) Make a bar plot of the resulting correlations in (a) and arrange the nutrition facts in decreasing order in terms of their correlation with the `calories`. Which nutrition fact has the highest values?

(c) Make a scatter plot where y axis represents `calories` and x axis represents the nutrition fact with the largest Pearson correlation coefficient to `calories` in (b). Add a trend line and interpret the meanings of intercept and slope in this context.

(d) The main cereal manufactures are General Mills and Kelloggs, since they have larger numbers of samples than any others. Make density curves of `calories` to compare these two manufactures in one plot and describe their shapes, respectively.

(e) Are `calories` significant different between these two main manufacturers, i.e. General Mills and Kelloggs? Answer this question via showing an appropriate plot.

(f) Make seven side-by-side box plots to compare each of the seven nutrition facts, including `protein`, `fat`, `sodium`, `fibre`, `carbo`, `sugars`, and `potassium`, among the six manufactures. Discuss which manufacture aims for a better healthy diet?

(g) Make a stacked bar plot to show the relationship between manufacture (i.e. `mfr`) and shelf placement (i.e. `shelf`). (Hint: use meaningful or your favorite colors to indicate different manufactures, you may consider to use the color tones in their logos)


Kellogg's
NABISCO
Post CONSUMER BRANDS
QUAKER
EST? 1877