Homework 4
of
STAT 3355 Introduction to Data Analysis

Question 1 (0 points)

Download the Mobile Price Classification data set (train.csv). Read the data in its original format (.csv) by using the function `read.csv()` in to the data frame **mobile_data**. In this dataset, there are 2000 observations with 21 variables.

The variables are listed as they appear in the data file.

Variable Name	Description
battery_power	energy charge that a battery will hold and how long a device will run before the battery needs recharging.
blue	"1" for phone has bluetooth and "0" for phone doesn't have bluetooth
clock_speed	speed at which a single microprocessor core executes instructions
dual_sim	"1" for phone that can handle 2 sim cards simultaneously and "0" for phone that can only handle 1 sim card at a time
fc	The mega pixels that the front camera can support
four_g	"1" for 4G capability on phone and "0" for no 4G capability on phone
int_memory	Internal Memory of the phone in Gigabytes
m_depth	Mobile Depth in cm
mobile_wt	Weight of mobile phone
n_cores	Number of cores in the phone's microprocessor
pc	The mega pixels that the primary camera can support
px_height	Pixel Resolution Height

px_width	Pixel Resolution Width
ram	Random Access Memory in Megabytes
sc_h	Screen height of phone in cm
sc_w	Screen width of phone in cm
talk_time	the total time a battery can power a phone while the phone is used to receive or perform a call
three_g	"1" for 3G capability on phone and "0" for no 3G capability on phone
touch_screen	"1" for touchscreen capability on phone and "0" for no touchscreen capability on phone
wifi	"1" for wireless network connection capability on phone and "0" for no wireless connection capability on phone
price_range	"0" for low cost phones, "1" for medium cost phones, "2" for high cost phones, and "3" for very high cost phones

Let's work on the Mobile Price Classification dataset using the package `ggplot2`. You can use the following code to install this package. Use `ggplot2` to make all the required plots and data visualizations.

```
# Install the package if you never did
install.package("ggplot2")

# Load the package
library(ggplot2)
```

(a) Make a scatter plot between the variables battery_power vs ram. Add colors based on price_range.

(b) Recreate the plot from Part a, and add the trend lines for each price range separately.

(c) Make density curves of the ram where the 4 price ranges are in one plot.

(d) Make box plots of the ram where the 4 price ranges are in one plot.

(e) Make a violin plot of the ram where the 4 price ranges are in one plot.

(f) Make a stacked bar plot to show the relationship between price range and $\log_2(\text{ram})$.

Problem 2 (0 points)

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 mpg dataset
data("UScereal")
```

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" ...
```

(d) 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.

(e) 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?

(f) 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.

(g) 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.

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

(i) 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?

(j) 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)