Homework 3
of
STAT 3355 Introduction to Data Analysis

Question 1

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

Answer:

(a) Turn the variable price_range into a factor variable with levels: “0” for low, “1” for medium, “2” for high, and “3” for very_high.

Answer:

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

Answer:

(c) Find the Pearson correlation between the variables ram and battery power.

Answer:

(d) Create four separate data sets by sub-setting the “mobile_data” using the variable price_range as “priceLow”, “priceMedium”, “priceHigh” and “priceVeryhigh”.

Answer:

(e) Calculate the Pearson correlation coefficient between the variable pair (ram , battery_power) separately for each price range. Explain any correlations you might find in terms of how a cellphone operates. Why is this result so much different from the one that we found in Part c?

Answer:

(f) Recreate the plot from Part b, and using the lm() function add the trend lines for each price range separately.

Answer:

(g) Find the average and the medium clock speed of the mobile phones which has 4, 6 and 8 cores in their processors. Round your answer to two decimal places. Explain why the average and median clock speed doesn't change.

Answer:

(h) Using the density() function make density curves of the ram where the 4 price ranges are in one plot and describe their shapes

Answer:

(i) Make box plots of the ram where the 4 price ranges are in one plot and describe their shapes respectively.

Answer:

(j) Make a stacked bar plot to show the relationship between price range and $\log_2(\text{ram})$. (Hint: use different colors to indicate different ram types).

Answer:

Problem 2

Let's work on the mpg dataset in the package ggplot2. You can use the following code to load the data. Use necessary code to read the description of the dataset, which contains 234 samples and 11 variables.

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

# Load the pacakge
library(ggplot2)

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

Let's first clean the data:

(a) Turn the variable cyl to an ordered factor variable with levels "4", "5", "6", and "8"

Answer:

(b) Turn the variable trans to a factor variable, of which unique values are "auto" and "manu" (Hint: use the function substr() to extract substrings in a character vector before converting to a factor vector)

Answer:

(c) Turn the variable drv to an ordered factor variable with levels "f", "r", and "4",

Answer:

(d) Turn the variable fl to a factor variable, of which unique values are "gasoline", "diesel", and "other" (Hint, "other" should include "e" and "c" in the original variable: "e" for E85, which is an ethanol fuel blend of 85% ethanol fuel and 15% gasoline and "c" for compressed natural gas)

Answer:

(e) Turn the variable class to an ordered factor variable with levels "2seater", "subcompact", "compact", "midsize", "suv", "minivan", and "pickup"

Answer:

(f) Create a new variable of `country` to indicate the manufacturer base location (Hint: You can refer to the following tables)

Country	Manufacturer
United States	Chevrolet, Dodge, Ford, Jeep, Lincoln, Mercury, Pontiac
Japan	Honda, Nissan, Subaru, Toyota
Germany	Audi, Volkswagen
South Korea	Hyundai
Great Britain	Land Rover

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

```
$ manufacturer: chr  "audi" "audi" "audi" "audi" ...
$ model       : chr  "a4" "a4" "a4" "a4" ...
$ displ       : num  1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
$ year        : int  1999 1999 2008 2008 1999 1999 2008 1999 1999 200
  8 ...
$ cyl         : Factor w/ 4 levels "4","5","6","8": 1 1 1 1 3 3 3 1 1
   1 ...
$ trans       : Factor w/ 2 levels "auto","manu": 1 2 2 1 1 2 1 2 1 2
   ...
$ drv         : Factor w/ 3 levels "f","r","4": 1 1 1 1 1 1 1 3 3 3
  ...
$ cty         : int  18 21 20 21 16 18 18 18 16 20 ...
$ hwy         : int  29 29 31 30 26 26 27 26 25 28 ...
$ fl          : Factor w/ 3 levels "diesel","gasoline",..: 2 2 2 2 2
  2 2 2 2 2 ...
$ class       : Factor w/ 7 levels "2seater","subcompact",..: 3 3 3 3
   3 3 3 3 3 3 ...
$ country     : chr  "germany" "germany" "germany" "germany" ...
```

Answer:

(g) Draw a bar plot of the variable `country` and arrange the country in decreasing order in terms of the number of samples. Which country has the most samples in this dataset? Which has the least?

Answer:

␣

(h) Summarize what a typical U.S. car looks like, in terms of engine displacement (i.e. `displ`), number of cylinders (i.e. `cyl`), type of transmission (i.e. `trans`), drive type (i.e. `drv`), fuel type (i.e. `fl`), and type of car (i.e. `class`)? (Hint: Use the function `table()` to find the mode for each of the above discrete univariate data)

Answer:

(i) Make a boxplot of the combined miles per gallon (i.e. (`cty` + `hwy`)/2) of U.S. cars and Japan cars, respectively, and report their means, medians, standard deviations, and IQRs.

Answer:

```







                                                                  ␣










```

(j) Make a histogram of the engine displacement (i.e. `displ`) of U.S. cars and Japan cars, respectively, and describe their shapes.

Answer:

```
                                                             ␣
                                                                  ␣

                                                               ␣
                                                                    ␣
```