Homework 3
of
STAT 3355 Data Analysis for Statisticians & Actuaries

Due: 11:30 am

March 9 (Wednesday), 2022

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")
```

Problem 1 ($1 \times 6 = 6$ points)

Let's first clean the data:

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

(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)

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

(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)

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

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

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

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 199
  9 1999 2008 ...
$ 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
  " ...
```

Problem 2 ($1 \times 4 = 4$ 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) 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?

(b) 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)

(c) 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.

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