NDH802 - Assignment 1

Group no.

- The assignment includes 2 questions, 6 points each.
- Submit your assignment via Canvas before 17:00 CET, April 12, 2021.
- Your submission should include (1) an RMarkdown file with your solutions in words and/or R code, (2) a knitted pdf file and (3) your handwritten solution if you have one. Name the files NDH802_Assignment1_GroupNumber.
- You should work in groups and contribute equally.
- You can copy code from "Hello R!", but make sure you understand it.
- You should not have the exact solutions and/or results with other groups.

Set things up

Set your working directory

```
#setwd("")
```

Load the package(s) you're going to use. If you don't use any packages, leave it as is.

Run this code chunk to load data into your R Environment. The command will randomly select 1,000,000 rows of data from the original data set, i.e., everytime you run the code, you have a new (unique) data set `df`. Accordingly, your results should be different from your friends and you should not be comparing them.

```
# inference_dataset <- read.csv("https://cda.hhs.se/inference_dataset.csv")
# df <- inference_dataset[sample(1:nrow(inference_dataset),
#                                size = 1000000,
#                                replace = FALSE), -1]
# rm(inference_dataset)
```

Please refer to Canvas for more details about the data set.

Question 1. Mean and variance

(a) Plot the histogram of customer's `value`. Imagine you will present this to your boss at work. Make it readable and self-explanatory (e.g., add the title for the chart and labels for the axes where needed). (1p)
(b) Make a box plot for `visits` for 2 groups, loyal and not loyal customers. Refer to the code provided (1p)
Imagine you are the customer relationship manager. What would you say about this figure? (1p)

```
# boxplot(
#   sal_expected ~ program,
#   data = salaries,
#   ylim = c(0, 120)
# )
```

(c) Compute the mean and variance of `value`, `deals`, `points`. Be careful, variance is **not** standard deviation. (1p)
(d) Compute the mean and variance of `value`, `deals`, `points` of the *loyal* customers. (1p)
(e) Compute the mean and variance of `value`, `deals`, `points` of the *loyal* customers who *shop offline*. (1p)

Question 2. Probability theory

(a) How many loyal and not loyal customer do you have in your `df`? Formally, compute N_{loyal} and N_{not_loyal}. (1p)

For Q2b-f, consider the following events in your `df`:

- (E1) Being loyal customer
- (E2) Not being loyal customer
- (E3) Being an offline customer
- (E4) Being an online customer

(b) Are E1 and E2 mutually exclusive? Why/why not? (1p)
(c) Are E1 and E2 collectively exhaustive? Why/why not? (1p)
(d) Are E3 and E4 mutually exclusive? Why/why not? (1p)
(e) Are E3 and E4 collectively exhaustive? Why/why not? (1p)
(f) Are E1 and E4 mutually exclusive? Why/why not? (1p)

For Q2b-f, you can write the solutions using formulas, words, Venn diagrams, code, numbers or the combination of them, whichever expresses your rationales the best. If you find handwriting more convenient, feel free to do so and attach a photo of it in the submission.

Have fun and good luck!
Huong and Emelie