NDH802 - Assignment 1

Group no.

- The assignment includes 2 questions, 6 points each.
- Submit your assignment via Canvas before 10:00 CET, April 12, 2021.
- Your submission should be an RMarkdown file with your solutions in words and/or R code. If you handwrite parts of your assignment, insert it as an image near the corresponding question(s). Name the files NDH802_Assignment1_GroupNumber.
- You should work in groups and contribute equally.
- You can copy my code, but make sure you understand it.
- You should not have the exact solutions and/or results with other groups.
- Results without code/justifications will not be graded.

Set things up

Set your working directory

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

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. Fill in your group number within 'set.seed()'. For example, if you are group 3, make it `set.seed(3)`. Hereby each and every group should have a unique `df`. Accordingly, your results should be different from other groups' and you should not be comparing them.
Note. If you make the wrong seed, your assignment will **not** be graded.

```
inference_dataset <- read.csv("https://cda.hhs.se/inference_dataset.csv")
set.seed(2); df <- inference_dataset[sample(1:nrow(inference_dataset),
                           size = 1000000, replace = FALSE), -1]
rm(inference_dataset)
```

Please refer to Canvas, Hand-in 1 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 and modify it (1p). Imagine you are the customer relationship manager. What would you say about this figure? (1p)

```
# boxplot(
#   your_variable_of_interest ~ the_group,
#   data = your_df,
#   ylim = c(-10, 100) # adjust the ylim that better illustrates your data
# )
```

(c) Compute the mean and variance of `value`, `deals`, `points`. Be careful, variance is **different from** 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 *made at least one offline purchase*. (1p)

Question 2. Probability theory

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

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

(E1) Being loyal

(E2) Not being loyal

(E3) Made at least one offline purchase

(E4) Made at least one online purchase

(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.

Great job!

Your work demonstrates your understanding about the (sample) mean, variance, data manipulation and visualization with R. Love your personal touches on the graphs. I also appreciate your effort in answering the questions with words, code, numbers and venn diagrams.

Something that can make your work even better:

- The histogram: You get full score for this. Yet for your future work, you may want to put the currency in the x labels. You're giving this to your boss, you want it to be as informative as possible.
- The boxplot: the bold lines in box plot represent the medians, not mean. They are not always the same. It would Your last sentence in the comment is not wrong. However from the box plot, or even the data set, we don't have enough evidence to infer the profitability. Though it is intuitive to say more visits generate more profit, we don't know that for sure. What if they always bought the products with lowest margin?
- Q1e: what you did is absolutely right. I like how to manage to solve the problem with limited resources. I share an alternative above. You are however free to do whatever you feel most comfortable with.
- Q2b: what you did is absolutely right. Notice that the code in Q2b gives you the same results as Q2a, and table() is somewhat more time-efficient to count simple things. Also, subset() give you the same operation as df[df$column == "condition",]. Again, feel free to do whichever you like best.

```
loyal = nrow(subset(df, loyal == 1))
```

- To avoid copying and pasting the results, you can look into inline code, for example 378970. If you knit it to html file, you'll see the value of loyal. This is totally optional, but I think it's one of the beauties of RMarkdown.