NDH802 - Assignment 1

Group no.

- The assignment includes 2 questions, 6 points each.
- Submit your assignment via Canvas before 15:00 CET, April 6, 2022.
- 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 and fill in your group number. For example, if you are group 3, make it `our_group <- 3`. If you don't fill in your group number or fill in the wrong number, your assignment will **not** be graded.

```
#setwd("")
our_group <- 27
```

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. 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.

Data description

Variable	Description
cust.id	Unique customer id
age	Customer age in year 2021
email	If there is an email of the customer in the system
member.since	Year from which the customer become a member. They can only register at the physical stores.
distance.to.store	Distance (in km) from customer's address to the physical store they register their membership
store.trans	Total number of offline transactions the customer made in year 2021
store.spend	Total amount (in SEK) the customer spend from offline transaction in year 2021
online.visits	Total number of time the customer visit (does not necessarily mean purchase) the online store in year 2021
online.trans	Total number of online transactions the customer made in year 2021
online.spend	Total amount (in SEK) the customer spend from online transactions in year 2021
points	Total loyalty points the customer accumulates since they become a member deducted by points they have used
main.format	The format in which the customer made the most transactions in year 2021

Question 1. Mean and variance

(a) Compute the `total.spend` of the customer (that includes `store.spend` and `online.spend`). Plot the histogram of `total.spend`. Imagine you will present this to your manager. Make it readable and self-explanatory (e.g., add the title for the chart and labels for the axes where needed). (1p) How do you explain the peak of the histogram, in general and in this context? (1p)

(b) Make a box plot for `total.spend` for 3 groups, customers whose main format is supermarket, convenience store and online. Refer to the code provided and modify it (1p). From this figure, would you conclude that online is the format that contributes the most to the `total.spend`? Why/why not? (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 `distance.to.store` of the customers whose `main.format` is supermarket and of the customers whose `main.format` is convenient store (1p). Comment on the difference between the means of the two groups; and the difference between the variances of the two groups (1p).

Question 2. Probability theory

Consider the following events:

(A) Made at least one offline transaction

(B) Made at least one online transaction

(a) Compute $P(A)$ and $P(B)$. (1p)

(b) What is the complement of B? Formally, define event $\bar{B}$ and compute $P(\bar{B})$. (1p)

(c) Compute $P(B \cap A)$ and $P(B \mid A)$. (1p)

(d) Are A and B independent events? Why/why not? (1p)

(e) Compare and explain (with formula) the similarities/differences among $P(B)$, $P(B \cap A)$ and $P(B \mid A)$ (2p).

Have fun and good luck!
Huong and Emelie