NDH802 - Assignment 3

Group no.

- The assignment includes 3 questions, 4 points each.
- Question 1 is purely theoretical (no data). Question 2 and 3 are empirical (based on the data provided).
- Your submission should include (1) an RMarkdown file with your solutions in words and/or R code, (2) a pdf file. If you handwrite parts of your assignment, include a photo of it in your pdf file. Name the files NDH802_Assignment3_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(1); df <- inference_dataset[sample(1:nrow(inference_dataset),
                           size = 1000000,
                           replace = FALSE), -1]
rm(inference_dataset)
```

Please refer to Canvas, Hand-in 3 for more details about the data set.

Question 1. Sampling theory

Let X (measured in cm) denote the height of all the Swedes in 2021. Assume $X \sim N(\mu = 175, \sigma^2 = 20)$.

(a) Assume you randomly meet a Swede in the street. What is the probability that he/she is from 173cm to 177cm? (1p)
(b) Assume you randomly meet 2 Swedes in the street. What is the probability that *both of them* are from 173cm to 177cm tall? (1p)
(c) Assume you randomly meet 3 Swedes in the street. What is the probability that *your sample mean* is from 173cm to 177cm? (1p)
(d) Assume you randomly meet 30 Swedes in the street. What is the probability that *your sample mean* is from 173cm to 177cm? (1p)

Question 2. Hypothesis testing

(a) Imagine you are the customer relationship manager. One of your colleagues argues that **the average `value` of the customers is 4150**. Do you reject this claim at 95% confidence level based on your sample data (i.e., `df`)? Justify your answer. (1p)
(b) Now you want to investigate only the customers who return more than 30 items. Your colleague argues that **their average `value` is less than 40000**. Do you reject this claim at 95% confidence level based on your sample data (i.e., `df`)? Justify your answer. (1p)
(c) Consider these two groups of customers:

- Group 1: Loyal customers whose points are lower than 6500
- Group 2: Not loyal customers whose points are higher than 1000

Which group is more valuable (based on their `value`)?
Perform the test at 99% confidence level (1p) and justify your choice (1p).

Question 3. Correlation and regression

(a) Your objective for 2021 is to increase the customer `value`. In order to do that, you first aim to understand which factors (variables) are the most influential. Write your own linear regression equation (modify the one below), perform the estimation, print out the model estimation.

$$value = \beta_0 + \beta_1 var_1 + \ldots + \beta_n var_n + \epsilon$$

You are free to include any variables in `df` (other than `channeltype`) as your dependent variables. (1p)

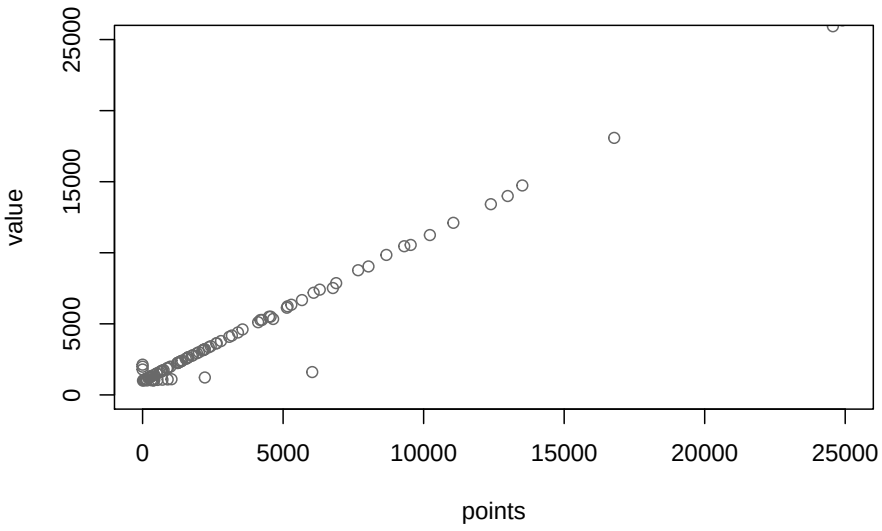
(b) You need to explain the results from (a) to your boss who has not taken the course "Data analytics". How would you do that? (1p)

(c) Imagine you have three new customers:

```
##   deals points items returns stores visits loyal channeltype
## 1     0    565    23       9      5     17     0       multi
## 2  1062    465    11       0      2      3     1  onlineonly
## 3  1293    500     5       0      2      2     0 offlineonly
```

Based on your linear regression model, which customer (choose one) do you think is the most valuable? Justify your choice. (1p)

(d) Looking at the figure below, your colleague suggests that you should give the customers more `points` to increase their `value`. What do you think? (1p)



Note: For illustration purpose, the plot only includes 100 customers. One little circle represents one customer.

Have fun and good luck!
Huong and Emelie