NDH802 - Assignment 3

Group no.

- The assignment includes 3 questions, 4 points each.
- 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, namely `df`, `sample_A` and `sample_B` into your R Environment. Refer to Assignment 1 for data description.

```
inference_dataset <- read.csv("https://cda.hhs.se/NDH802data.csv")
set.seed(our_group); df <- inference_dataset[sample(1:nrow(inference_dataset),
                           size = 1000000,
                           replace = FALSE), -1]; rm(inference_dataset)
sample_A = sample(df$online.spend/df$online.trans, 25)
sample_B = sample(df$online.spend/df$online.trans, 2500)
```

Question 1. Sampling theory and confidence intervals

To investigate the average value of online baskets, you sample some online baskets from the population. Assume the *true* average value of online baskets follows normal distribution. `sample_A` contains the average online baskets values of 25 customers and `sample_B` contains the average online baskets values of 2500 customers.

```
#sample
```

(a) Based on `sample_A`, what is the 90% confidence interval (CI) for the population mean? (0.5p) How would you interpret the 90% CI, in general and in this case? (0.5p)

(b) Based on `sample_B`, what is the 90% confidence interval (CI) for the population mean? (0.5p) Why does this CI different from what you found in (a)? (0.5p)

(c) If we repeatedly and independently draw 25 customers from the population, would 90% of the population means fall into the CI you found in (b)? Why/why not? (1p)

(d) Assuming sample standard deviation stay unchanged, what are the possible ways to reduce the margin of error? Justify your answer and discuss the pros and cons of each of the ways. (1p)

Question 2. Hypothesis testing

In order to get full score for this question, you need to formulate the null and alternative hypotheses, perform the tests at the significance level of your choice and explain your results in details. Note that we would like to generalize the finding to the whole customer base (the population) and not only the observations in your sample `df`. Therefore, computing and comparing the means with our eyes is not substantial. Code without motivations will not be graded.

(a) Imagine you are the customer relationship manager. One of your colleagues argues that the average `store.spend` of the customers is 58080. Based on your sample data `df`, would you reject this claim at 95% confidence level? (1p)

(b) Suppose you want to know if there is strong evidence that the average `store.spend` is higher than the average `online.spend`. How would you test it and what would you conclude? (1p)

(c) Suppose you want to know if there is strong evidence that the customers whose `main.format` is offline live nearer to the store as compared with the customers whose `main.format` is online. How would you test it and what would you conclude? (1p)

(d) What is Type I error in general and in the test you perform in (b)? What can we do to reduce Type I and what are the consequences? (1p)

Question 3. Correlation and regression

(a) Your objective for 2022 is to increase the customer `total.spend`. In order to do that, you first aim to understand which factors (independent variables - IVs) are the most influential. Write your own linear regression equation (modify the one below), explain your choice of IVs [1], perform the estimation, print out the model estimation.

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

Heads up:

- `main.format` and `member.since` are categorical variable, which is a little more difficult to work with.
- The more IVs do not not always guarantee the better model.

(b) Interpret the results, in both statistical and business language. (0.5p)
Discuss a strategic plan to increase customer `total.spend` based on your model. (0.5p)

(c) Imagine you have three new customers. Based on your linear regression model, which customer (choose one) do you think will have highest `total.spend` and why? (1p)

```
##   age email member.since distance.to.store store.trans online.visits
## 1  58    no         2020              0.18         276             4
## 2  23   yes         2020              0.48         161           417
## 3  35   yes         2021             10.37          77           173
##   online.trans points       main.format
## 1            1    855       Supermarket
## 2          104   2099 Convenience store
## 3          114   1595            Online
```

(d) What is the key underlying assumption of linear regression model? When is linear regression model useful and when is it not? (1p)

Have fun and good luck!
Huong and Emelie

[1] You are free to include any variables in 'df' as your IVs