NDH802 - Assignment 3

Group no.

- The assignment includes 3 questions, 4 points each.
- Question 1 is based on `sample`. Question 2 and 3 are based on `df`.
- 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. Fill in your group number next to `our_group_number =`. For example, if you are group 3, make it `our_group_number = 3`. Hereby each and every group should have a unique `sample` for questions 1 and a unique `df` for question 2 and 3. Accordingly, your results should be different from other groups' and you should not be comparing them.
Note. If you fill in the wrong group, your assignment will **not** be graded.

```
our_group_number = 16

inference_dataset <- read.csv("https://cda.hhs.se/inference_dataset.csv")
set.seed(our_group_number); df <- inference_dataset[sample(1:nrow(inference_dataset),
                           size = 1000000,
                           replace = FALSE), -1]; rm(inference_dataset)
sample = read.csv("https://bit.ly/3dLzoju")[,our_group_number]
```

Question 1. Sampling theory and confidence intervals

The maximum score for the course NDH802 exam is 70. Out of curiosity, you talk to a random sample of 25 students from previous years to learn about the scores and record them in the `sample`.

```
#sample
```

(a) Based on your `sample`, 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) You find an (imaginary) report that says the exam scores are normally distributed with the standard deviation of 12. With this new information, what is the 90% CI for the population mean? (0.5p) Is this CI different from the one you got from (a)? Why/why not? (0.5p)

(c) If you knew the grades of all the students who have ever taken the NDH802 exams, would 90% of the (population) grades fall into the CI you found in (b)? Why/why not? (1p)

(d) Assume that you want to increase the confidence level to 99%, but you do not want the CI to be wider, i.e., you want the same CI that you found in (b). What would you do? Justify your answer with empirical evidence(s). For example, if you propose to adjust something, please indicate why and by/to how much. (1p)

Note: All assumptions in this question are hypothetical.

Question 2. Hypothesis testing

Q2a, c and d: In order to get full score, you need to formulate the null and alternative hypotheses, perform the tests and explain your results in details. Code without motivations will not be graded.

(a) Imagine you are the customer relationship manager. One of your colleagues argues that the average `value` of the customers is 4129. Based on your sample data `df`, do you reject this claim at 95% confidence level? (1p)

(b) Your colleague understand that p-value is the probability that the null hypothesis is true, given your sample mean of `value` [1]? Formally, s/he thinks that $p - value = P(H_0 \text{ is true} \mid \text{sample mean})$. Do you agree/disagree? Why? (1p)

(c) Your colleague now proposes that the probability of loyal customers who has never shop offline is 23%. You think it should be higher. Perform the hypothesis testing and with the confidence level of your choice. Explain why you choose that confidence level and the results. (1p)

(d) 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`)? Note that we would like to generalize this 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. (1p)

[1] what you computed in Q1c, Assignment 1

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 (independent variables - IVs) are the most influential. Write your own linear regression equation (modify the one below), explain your choice of IVs [2], perform the estimation, print out the model estimation.

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

Heads up:

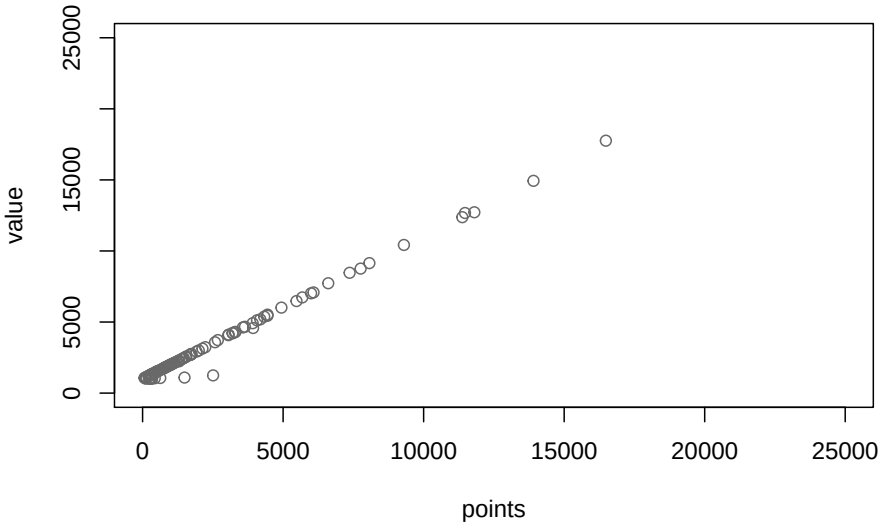
- `channeltype` is categorical variable, which is a little more difficult to work with.
- The more IVs do not not always guarantee the better model.

(b) Interpret the results, in both statistical and business language. (0.5p)
Discuss a strategic plan to increase customer `value` 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 is the most valuable and why? (1p)

```
##   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
```

(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

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