NDH802 - Assignment 1

Group no.

- The assignment includes 2 questions, 6 points each.
- Submit your assignment via Canvas before 17:00 CET, April 12, 2021.
- Your submission should include (1) an RMarkdown file with your solutions in words and/or R code, (2) a knitted pdf file and (3) your handwritten solution if you have one. Name the files NDH802__Assignment1__GroupNumber.
- You should work in groups and contribute equally, or at least understand all parts of the assignment.
- You can copy code from "Hello R!", but make sure you understand it.
- You should not have the exact solutions and/or results with other groups.

Set things up

Set your working directory

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

Load the package(s) you're going to use. If you don't use any packages, leave it as is.

```r
#library(tidyverse)
```

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, i.e., everytime you run the code, you have a new (unique) data set `df`. Accordingly, your results should be different from your friends and you should not be comparing them.

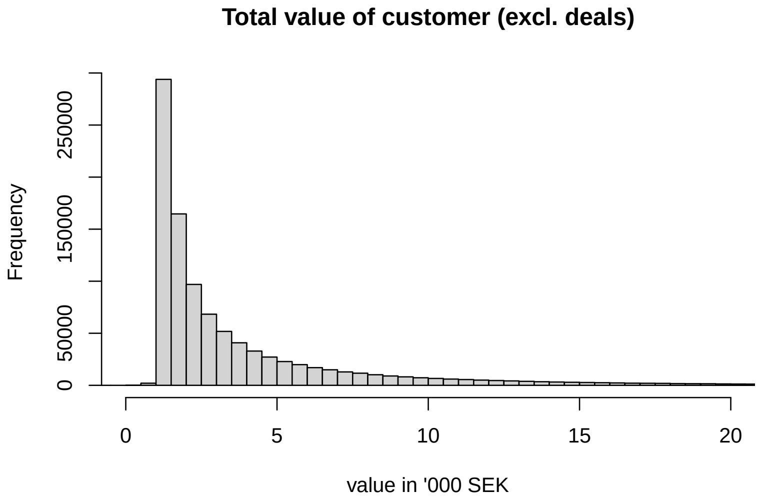
```r
inference_dataset <-
  read.csv("~/Desktop/Data/NDH802/inference_dataset.csv")
df <-
  inference_dataset[sample(1:nrow(inference_dataset),
                           size = 1000000,
                           replace = FALSE), -1]
rm(inference_dataset)
```

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)

Answer

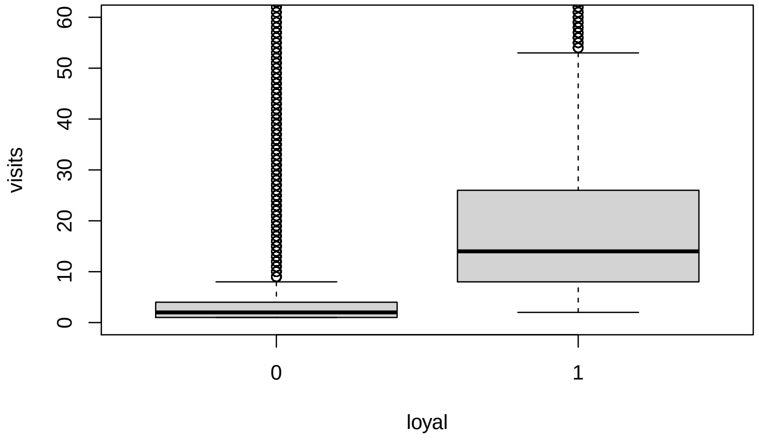
```r
hist(
  df$value / 1000,
  breaks = 1000,
  xlim = c(0, 20),
  xlab = "value in '000 SEK",
  main = "Total value of customer (excl. deals)"
)
```



(b) Make a box plot for `visits` for 2 groups, loyal and not loyal customers (1p).
Imagine you are the customer relationship manager. What would you say about this figure? (1p).

Answer

```r
boxplot(visits ~ loyal, data = df, ylim = c(0,60))
```



(c) Compute the mean and variance of `value`, `deals`, `points`. Be careful, variance is **not** standard deviation.(1p)

Answer

```r
#the more efficient way
mean_1c <- lapply(df[,2:4], mean, na.rm = TRUE)
var_1c <- lapply(df[,2:4], var, na.rm = TRUE)
print(data.frame(mean = unlist(mean_1c), variance = unlist(var_1c)))
```

```
##             mean variance
## value  4124.1005 31936711
## deals   894.9748  3248266
## points 3242.1143 31473863
```

```r
# newbies' way
mean(df$value, na.rm = TRUE)
```

```
## [1] 4124.1
```

```r
mean(df$deals, na.rm = TRUE)
```

```
## [1] 894.9748
```

```r
mean(df$points, na.rm = TRUE)
```

```
## [1] 3242.114
```

```r
var(df$value, na.rm = TRUE)
```

```
## [1] 31936711
```

```r
var(df$deals, na.rm = TRUE)
```

```
## [1] 3248266
```

```r
var(df$points, na.rm = TRUE)
```

```
## [1] 31473863
```

```r
# q1c <- df %>% summarise(
#     mean_value = mean(value, na.rm = TRUE),
#     var_value = var(value, na.rm = TRUE),
#     mean_deals = mean(deals, na.rm = TRUE),
#     var_deals = var(deals, na.rm = TRUE),
#     mean_points = mean(points, na.rm = TRUE),
#     var_points = var(points, na.rm = TRUE),
#   )
# print(q1c)
```

(d) Compute the mean and variance of `value`, `deals`, `points` of the *loyal* customers. (1p)
Answer

```r
mean_1d <- lapply(df[df$loyal == 1,2:4], mean, na.rm = TRUE)
var_1d <- lapply(df[df$loyal == 1,2:4], var, na.rm = TRUE)
print(data.frame(mean = unlist(mean_1d), variance = unlist(var_1d)))
```

```
##            mean variance
## value  7176.571 61561613
## deals  1590.597  6165782
## points 6330.204 59977672
```

```r
# newbies' way
mean(df[df$loyal == 1, "value"], na.rm = TRUE)
```

```
## [1] 7176.571
```

```r
mean(df[df$loyal == 1, "deals"], na.rm = TRUE)
```

```
## [1] 1590.597
```

```r
mean(df[df$loyal == 1, "points"], na.rm = TRUE)
```

```
## [1] 6330.204
```

```r
var(df[df$loyal == 1, "value"], na.rm = TRUE)
```

```
## [1] 61561613
```

```r
var(df[df$loyal == 1, "deals"], na.rm = TRUE)
```

```
## [1] 6165782
```

```r
var(df[df$loyal == 1, "points"], na.rm = TRUE)
```

```
## [1] 59977672
```

```r
# q1d <- df %>% filter (loyal == 1) %>%  summarise(
#     mean_value = mean(value, na.rm = TRUE),
#     var_value = var(value, na.rm = TRUE),
#     mean_deals = mean(deals, na.rm = TRUE),
#     var_deals = var(deals, na.rm = TRUE),
#     mean_points = mean(points, na.rm = TRUE),
#     var_points = var(points, na.rm = TRUE),
#   )
# print(q1d)
```

Question 2. Probability theory

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

```
table(df$loyal)
```

```
##
##      0      1
## 620625 379375
```

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

(E1) Being loyal customer

(E2) Not being loyal customer

(E3) Buy offline

(E4) Buy online

(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)

For Q2b-e, you can write the solutions using formula, 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.

Have fun and good luck!
Huong and Emelie