NDH802 - Assignment 2

Group no.

- The assignment includes 2 questions, 6 points each.
- Submit your assignment via Canvas before 10:00 CET, April 12, 2021. Late submissions will receive 5 points at most even if you obtain ≥ 5 points.
- 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_Assignment2_Groupnumber.
- You should work in a group of 4-5 students and contribute equally, or at least understand all parts of the assignment.
- You should not have the exact solutions and/or results with other groups.
- You can copy code from my illustration, but make sure you understand it.

Set things up

Set your working directory

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

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

```
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.

```
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. Discrete distribution

(a) What are the probabilities of loyal and not loyal customers in your data set? Formally, compute $P(\text{loyal} = 1)$ and $P(\text{loyal} = 0)$. (1p)
Answer

```
prop.table(table(df$loyal))
```

```
##
##       0       1
## 0.62029 0.37971
```

```
#Prob loyal
q2a_1 = df %>% filter(loyal == 1) %>% count() /
        df %>% count()
#Prob not
q2a_0 = df %>% filter(loyal == 0) %>% count() /
        df %>% count() #or
q2a_0 = 1-q2a_1
```

(b) What is the probability of loyal customers, given the customers who only shop offline? Formally, compute $P(\text{loyal} = 1 \mid \text{channel type} = \text{offline only})$. (1p)
Answer

```
q2b = prop.table(table(df$loyal, df$channeltype), margin = 2)
print(q2b[2,2])
```

```
## [1] 0.2559162
```

```
#this way look more like the mathematical formula
q2b_2 = df %>% filter(loyal == 1 & channeltype == "offlineonly")%>% count() /
        df %>% filter(channeltype == "offlineonly")%>% count()
```

(c) What is the probability of shopping only offline, given the loyal customers? Formally, compute $P(\text{channel type} = \text{offline only} \mid \text{loyal} = 1)$. (1p)
Answer

```
q2c = prop.table(table(df$loyal, df$channeltype), margin = 1)
print(q2c[2,2])
```

```
## [1] 0.38753
```

```
#this way look more like the mathematical formula
q2c_2 = df %>% filter(loyal == 1 & channeltype == "offlineonly")%>% count() /
        df %>% filter(loyal == 1)%>% count()
```

(d) What is the probability of customers who are both loyal and only shop offline? Formally, compute $P(\text{channel type} = \text{offline only}, \text{loyal} = 1)$. (1p)

```
q2d = df %>% filter(loyal == 1 & channeltype == "offlineonly")%>% count() /
      df %>% count()
```

(e) Compare the results of (b), (c) and (d). Are they similar/different? Should they? Why/why not? You can use mathematical formula, words, venn diagrams or the combination of them, whichever expresses your rationales the best. If you find handwriting is more convenient, feel free to do so and attach a photo of it in the submission.(2p)

Answer They are/should be different because:

$$P(\text{channel type} = \text{offline only}, \text{loyal} = 1) = \frac{P(\text{loyal} = 1, \text{channel type} = \text{offline only})}{1}$$

$$P(\text{channel type} = \text{offline only} \mid \text{loyal} = 1) = \frac{P(\text{loyal} = 1, \text{channel type} = \text{offline only})}{P(\text{loyal} = 1)}$$

$$P(\text{loyal} = 1 \mid \text{channel type} = \text{offline only}) = \frac{P(\text{loyal} = 1, \text{channel type} = \text{offline only})}{P(\text{channel type} = \text{offline only})}$$

Question 2. Continuous distribution

Have fun and good luck!