NDH802 - Assignment 1

Group no. 11

Setting Working Directory and naming the group

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.

```
our_group <- 11
setwd("/Users/emelieolsson/Downloads")
getwd()
```

```
## [1] "/Users/emelieolsson/Downloads"
```

Loading data into R environment

Summary of the data

```
summary(df)
```

```
##     cust.id             age           email            member.since
##  Min.   :      3   Min.   :19.00   Length:1000000     Min.   :2019
##  1st Qu.: 500692   1st Qu.:26.00   Class :character   1st Qu.:2019
##  Median :1000674   Median :32.00   Mode  :character   Median :2019
##  Mean   :1000489   Mean   :32.44                      Mean   :2020
##  3rd Qu.:1500014   3rd Qu.:38.00                      3rd Qu.:2020
##  Max.   :1999998   Max.   :75.00                      Max.   :2021
##  distance.to.store  store.trans      store.spend     online.visits
##  Min.   :  0.040   Min.   :  13.0   Min.   :  1832   Min.   :  0.00
##  1st Qu.:  1.390   1st Qu.:  77.0   1st Qu.: 25362   1st Qu.: 15.00
##  Median :  2.730   Median :  97.0   Median : 42969   Median : 38.00
##  Mean   :  4.478   Mean   : 104.6   Mean   : 58145   Mean   : 57.19
##  3rd Qu.:  5.350   3rd Qu.: 124.0   3rd Qu.: 74393   3rd Qu.: 79.00
##  Max.   :113.980   Max.   :1028.0   Max.   :319806   Max.   :835.00
##   online.trans     online.spend     total.spend        points
##  Min.   :  0.00   Min.   :     0   Min.   :  2394   Min.   :  29.0
##  1st Qu.:  5.00   1st Qu.:  3157   1st Qu.: 36932   1st Qu.: 451.0
##  Median : 13.00   Median :  8528   Median : 58676   Median : 722.0
##  Mean   : 21.38   Mean   : 14955   Mean   : 73100   Mean   : 885.5
##  3rd Qu.: 28.00   3rd Qu.: 19459   3rd Qu.: 93650   3rd Qu.:1148.0
##  Max.   :365.00   Max.   :278143   Max.   :319998   Max.   :5860.0
##  main.format
##  Length:1000000
##  Class :character
##  Mode  :character
```

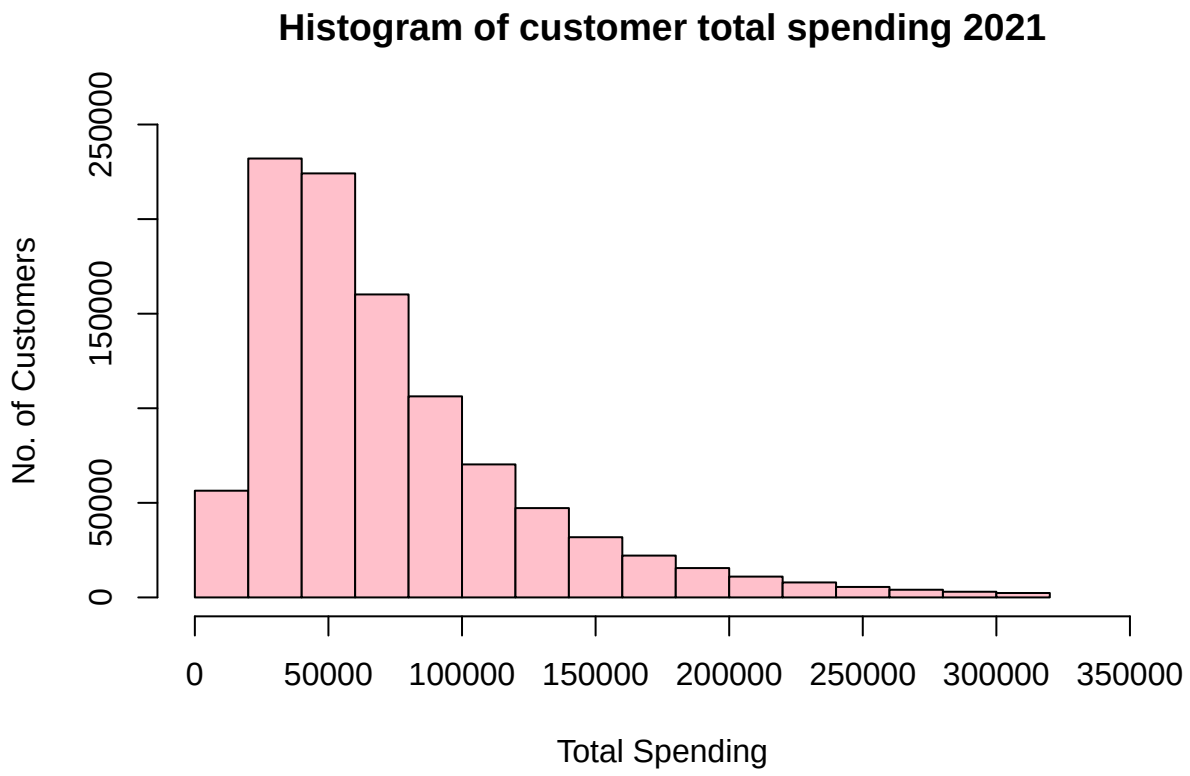
##
##
##

Question 1A

Calculating the total spending of each customer by combining and summarizing store spending and online spending into a single column named total.spend

```
df$total.spend <- df$store.spend + df$online.spend
```

```
hist(df$total.spend,
main = "Histogram of customer total spending 2021",
xlab = "Total Spending",
ylab = "No. of Customers",
border = "black",
col = "pink",
xlim = c(0, 350000),
ylim = c(0, 250000),
breaks = 20)
```




```
mode <- function(x) unique(x)[which.max(tabulate(match(x, unique(x))))]
mode(df$total.spend)
```

```
## [1] 57980.43
```

Peak of the histogram explained

The histogram is skewed to the right, which means that the peak of the graph lies to the left side of the center. As computed, the mode is lower than the mean and median, hence the histogram shows a **positively skewed distribution**.

Thus, in general terms, the higher the bar, the higher the number of observations in each selected interval.

In our case, the value that appears most frequently in the data set shows that the most customers spend around SEK 50 000.

Question 1B

Calculating the total spending for different store formats through box plot showing the median, the upper-and lower extreme values and the upper- and lower quartile

```
boxplot(
  df$total.spend ~ df$main.format,
  data = df,
  ylim = c(0,350000),
  main = "Box plot for total spending in different store formats 2021",
  xlab = "Store Formats",
  ylab = "Total Spending",
  border = "black",
  col = "pink")
```

Explaining whether online contributes most to total spending or not

A box plot shows how all distributions relates to the median for the respective store but it does not say how many observations have occured. Thus, we can not interpret the total spending from the box plots. We need to know how many observations to draw such a conclusion.

However, the box plots shows which store generates the highest revenue on average *per* customer, as it has the highest median, which is online.

Question 1C

Calculating the mean and variance of the distance to the preferred physical store format (Supermarket and Convenience store)

```
mean(df[which(df$main.format == 'Supermarket'), 5])
```

```
## [1] 4.769982
```

```
mean(df[which(df$main.format == 'Convenience store'), 5])
```

```
## [1] 1.330919
```

```
var(df[which(df$main.format == 'Supermarket'), 5])
```

```
## [1] 33.50062
```

```
var(df[which(df$main.format == 'Convenience store'), 5])
```

```
## [1] 0.3833695
```

Explaining the differences in mean and variance of the different store formats

Comparing the mean between Supermarket and Convenience store, we can interpret that customers are closer to Convenience stores than to Supermarket. This is shown since the mean, i.e. distance, is lower for Convenience.

Considering the variance, Supermarket has a larger variance compared to Convenience. A large variance indicates that numbers in the set are far from the mean and far from each other. A small variance indicates the opposite. With regards to this, the customers who have chosen their supermarket both lives close and far away from it, i.e. a large spread between the distance of customers to store and thus a large variance. Regarding convenience, most customer selecting their store live close to it as the variance is low.

Question 2

Events

```
(A) Made at least one offline transaction
(B) Made at least one online transaction
```

Question 2A

Calculating probability of A and B

```
n_offline <- nrow(df[df$store.trans >0,])
n_offline
```

```
## [1] 1000000
```

```
sum_customers <- nrow(df)
sum_customers
```

```
## [1] 1000000
```

```
pa <- n_offline/sum_customers
pa
```

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

$$P(A) = 1.0$$

```
n_online <- nrow(df[df$online.trans >0,])
n_online
```

```
## [1] 960006
```

```
pb <- n_online / sum_customers
pb
```

```
## [1] 0.960006
```

$$P(B) = 0.960006$$

Question 2B

Computing the complement of B

```
#The complement of B (B') is customers who did *not* make at least one online transaction. In this case
```

```
n_onlinecomp <- nrow(df[df$online.trans == 0,])
n_onlinecomp
```

```
## [1] 39994
```

```
pbcomp <- n_onlinecomp / sum_customers
pbcomp
```

```
## [1] 0.039994
```

```
n_intersect <- nrow(df[df$online.trans >0 & df$store.trans >0,])
n_intersect
```

```
## [1] 960006
```

```
pintersect <- n_intersect / sum_customers
pintersect
```

```
## [1] 0.960006
```

$$P(A \cap B) = 0.960006$$

Question 2C

Computing the probability of the intersection of A and B and the conditional probability of B and A

```
pcond <- pintersect / pa
pcond
```

```
## [1] 0.960006
```

$$P(B \mid A) = \frac{P(A \cap B)}{P(B)} = 0.960006$$

Question 2D

Interpret whether A and B are independent events

```
#Statistically independent if and only if the probability of both A and B is the same as the probabilit
pintersect
```

```
## [1] 0.960006
```

```
x <- pa * pb
x
```

```
## [1] 0.960006
```

The condition is fulfilled, hence the events are statistically independent

Question 2E

Interpret the similarities between probability of B, intersection of A and B and the conditional probability of B given A

Given above calculations, we can interpret that the intersection of A and B and of B given A is always equal to the probability of B as the probability of A is 100%, thus the sample never differs depending on the event.

$$P(A \cap B) = P(B \mid A) * P(A) > P(B \mid A) = \frac{P(A \mid B)}{P(B)} > P(A \cap B) = P(B \mid A) = P(B)$$