NDH802 - Assignment 1

Group no. 12

- The assignment includes 2 questions, 6 points each.
- Submit your assignment via Canvas before 15:00 CET, April 6, 2022.
- 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.

```r
# Add a second backslash. This is because "\\" has the special meaning of a single backslash.

setwd("C:\\Users\\Dougl\\Documents\\SSE\\Year_1\\NDH802 Data Analytics\\R_DataAnalytics\\AS1")

our_group <- 12
```

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. Hereby each and every group should have a unique `df`. Accordingly, your results should be different from other groups' and you should not be comparing them.

Data description

Variable	Description
cust.id	Unique customer id
age	Customer age in year 2021
email	If there is an email of the customer in the system
member.since	Year from which the customer become a member. They can only register at the physical stores.
distance.to.store	Distance (in km) from customer's address to the physical store they register their membership
store.trans	Total number of offline transactions the customer made in year 2021
store.spend	Total amount (in SEK) the customer spend from offline transaction in year 2021

Variable	Description
online.visits	Total number of time the customer visit (does not necessarily mean purchase) the online store in year 2021
online.trans	Total number of online transactions the customer made in year 2021
online.spend	Total amount (in SEK) the customer spend from online transactions in year 2021
points	Total loyalty points the customer accumulates since they become a member deducted by points they have used
main.format	The format in which the customer made the most transactions in year 2021

Question 1. Mean and variance

(a) Compute the `total.spend` of the customer (that includes `store.spend` and `online.spend`). Plot the histogram of `total.spend`. Imagine you will present this to your manager. Make it readable and self-explanatory (e.g., add the title for the chart and labels for the axes where needed). (1p) How do you explain the peak of the histogram, in general and in this context? (1p)

Answer.

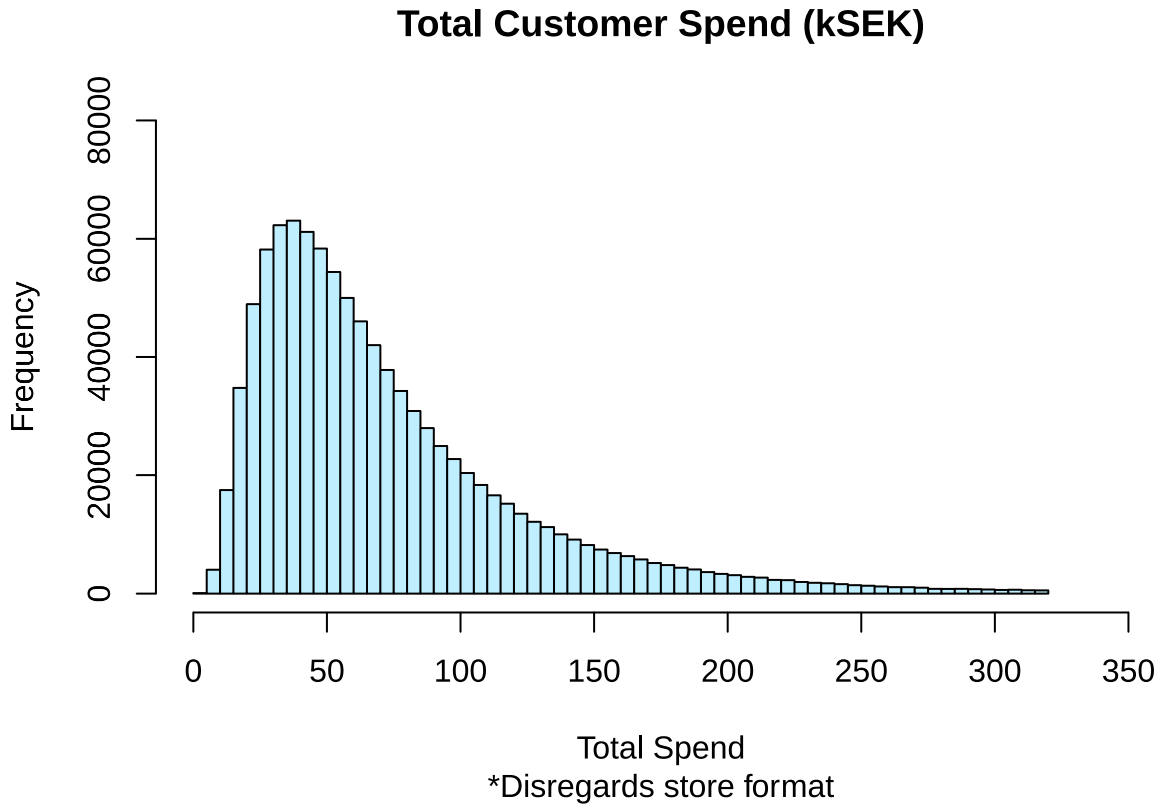
The total spend of the customers can be computed as

```
# Total spend irregardless of store format
sum(df[,"total.spend"])
```

```
## [1] 73066195431
```

If we then want to plot the numbers into a **histogram**, we can do so as

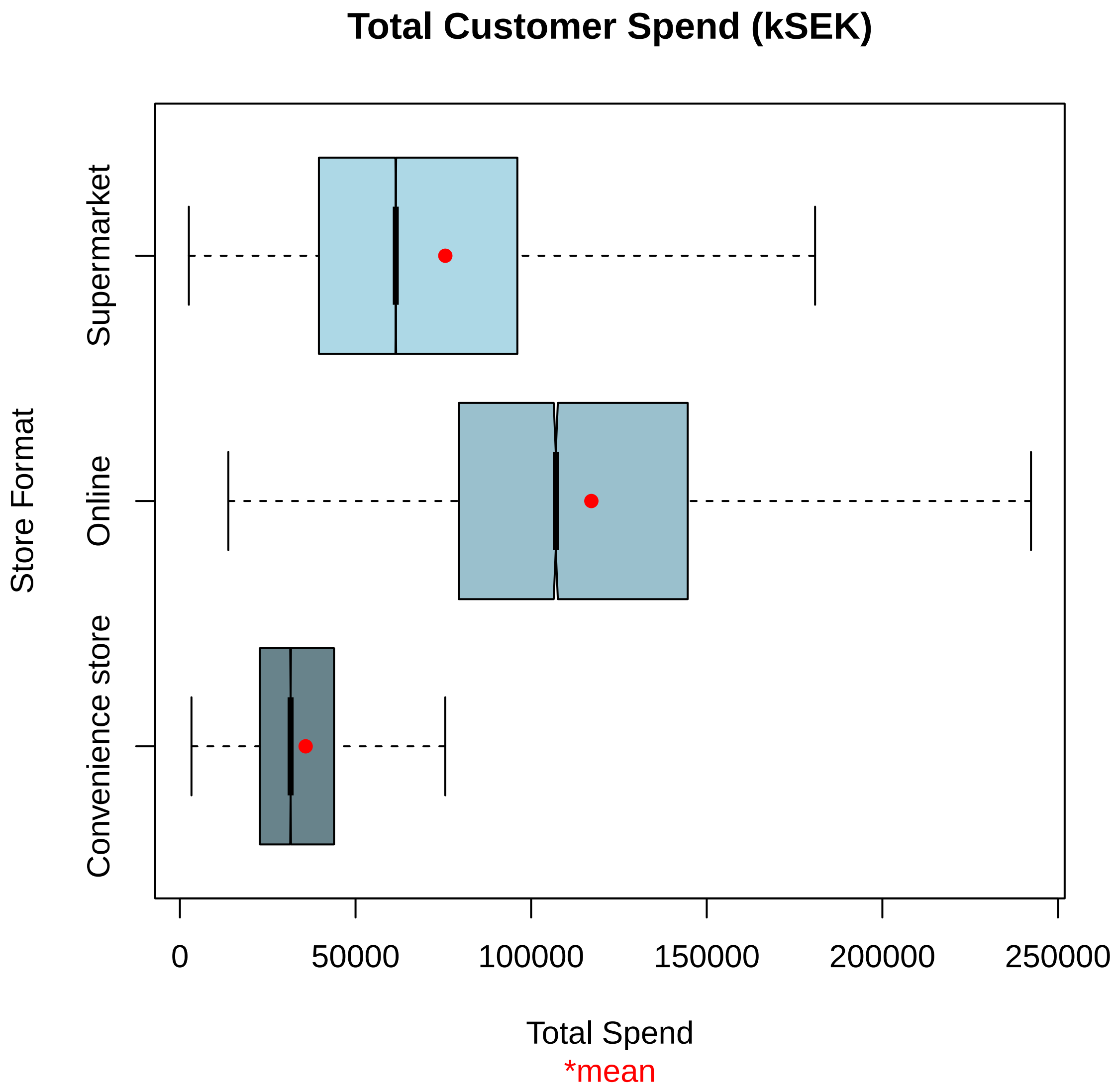
```
## Warning in plot.window(xlim, ylim, "", ...): "data" is not a graphical parameter

## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...): "data"
## is not a graphical parameter

## Warning in axis(1, ...): "data" is not a graphical parameter

## Warning in axis(2, ...): "data" is not a graphical parameter
```



This histogram, with an asymmetrical skewed-right distribution, shows the spending patterns among the customers in the sample, where the total spend is plotted against the x-axis and the frequency, i.e. how many people spent roughly the same amount, is plotted against the y-axis. Judging by this, most of the customers spent less than 100 kSEK with a peak at around 40 kSEK.

(b) Make a box plot for `total.spend` for 3 groups, customers whose main format is supermarket, convenience store and online. Refer to the code provided and modify it (1p). From this figure, would you conclude that online is the format that contributes the most to the `total.spend`? Why/why not? (1p)

Answer.



The box plot is a graphical representation of the *five-number representation* (min<Q1<median<Q3<max). All formats show signs of having asymmetrical skewed-right distributions, since the *mean* is greater than the *median* in all three box plots. Here, the outliers (not shown) can play different roles both depending on what data set is analyzed and what story we want to tell. In this case, outliers are omitted from the visual representation to place greater emphasis on the *five-number* representation.

However, the box plot is but a mean to exhibit the distribution of the data, which is *not* to say that the further to the right the more is spent in total. A quick calculation of the totals can be computed as

```
# Total spend irrespective of format
TotSpend <- sum(df$total.spend)

# Total spend per store format
TotSpendSup <- sum(df[df$main.format == "Supermarket", "store.spend"])

TotSpendCon <- sum(df[df$main.format == "Convenience store", "store.spend"])

TotSpendOnl <- sum(df[,"online.spend"])

# # Results
# TotSpend
# # -------------------
# TotSpendSup
# TotSpendCon
# TotSpendOnl

# Share of total spend
ShareSup <- TotSpendSup / TotSpend
ShareCon <- TotSpendCon / TotSpend
ShareOnl <- TotSpendOnl / TotSpend
```

```
# Results, shares
ShareSup
```

```
## [1] 0.7489328
```

```
ShareCon
```

```
## [1] 0.02902941
```

```
ShareOnl
```

```
## [1] 0.2042755
```

As we can see, online contributes to about a fifth of the total spend, whereas the format Supermarket dominates.

(c) Compute the mean and variance of `distance.to.store` of the customers whose `main.format` is supermarket and of the customers whose `main.format` is convenient store (1p). Comment on the difference between the means of the two groups; and the difference between the variances of the two groups (1p).

Answer.

```
# Mean distance to supermarket
mean(
  df[df$main.format == "Supermarket","distance.to.store"],
  na.rm = TRUE
)
```

```
## [1] 4.769085
```

```r
# Var distance to supermarket
var(
  df[df$main.format == "Supermarket","distance.to.store"],
  na.rm = TRUE
)
```

```
## [1] 33.45829
```

```r
# Mean distance to convenience store
mean(
  df[df$main.format == "Convenience store", "distance.to.store"],
  na.rm = TRUE
)
```

```
## [1] 1.331416
```

```r
# Var distance to convenience store
var(
  df[df$main.format == "Convenience store", "distance.to.store"],
  na.rm = TRUE
)
```

```
## [1] 0.3842234
```

The sample mean, $\bar{x}$, distance to a Supermarket or a Convenience store is computed by summing the total distance each customer has to either store, and then divides that distance by the total number of customers, n.

The sample variance, s^2, is the sum of the squared differences between each observation and the sample mean divided by the sample size minus 1, $n-1$. In this example, the variance tells us that people generally don't have very far to the nearest convenience store. The same can not be said about supermarkets.

Question 2. Probability theory

Consider the following events:

(A) Made at least one offline transaction

(B) Made at least one online transaction

(a) Compute $P(A)$ and $P(B)$. (1p)

Answer.

A - "Made at least one offline transaction"

B - "Made at least one online transaction"

Classical probability suggests that

$$P(A) = \frac{N_A}{N}$$

Therefore,

```r
n_A <- sum(df[,"store.trans"] > 0, na.rm = TRUE)
n_B <- sum(df[,"online.trans"] > 0, na.rm = TRUE)
n <- nrow(df)

P_A <- n_A/n
P_B <- n_B/n

# Results
P_A
```

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

```r
P_B
```

```
## [1] 0.960029
```

The probability that a randomly chosen customer has made at least one offline transaction as per `store.trans`, denoted $P(A)$, is 100%, meaning *all* customers have made at least one offline transaction in the sample space. Furthermore, the probability that a randomly chosen customer has made at least one online transaction as per `offline.trans`, denoted $P(B)$, is 96%, which means that most but not all customers have made at least one online transaction.

(b) What is the complement of B? Formally, define event $\bar{B}$ and compute $P(\bar{B})$. (1p)

Answer.

The complement of B (defined previously as: "Made at least one online transaction") is any constituted by all customers in the sample space who have *not* made any online transactions. From the `df`, these subjects are identified through `online.trans` when it equals zero.

```r
Bbar <- sum(df[,"online.trans"]==0, na.rm=TRUE)

P_Bbar <- Bbar/n

# Results
P_Bbar
```

```
## [1] 0.039971
```

Similarly, the complement to B can also be computed if we know the probability of event B, $P(B)$, and subtracting it from the set of the total basic outcomes in the sample space, 1, as such:

```r
P_Bbar <- 1 - P_B

# Results
P_Bbar
```

```
## [1] 0.039971
```

(c) Compute $P(B \cap A)$ and $P(B \mid A)$. (1p)

Answer.

To compute the intersection of events A and B, we need to first see how many observations fulfill both events and then divide that number by the total number of basic outcomes, n. Using equations, that is

$$P(A \cap B) = \frac{n_{A \cap B}}{n}$$

In this case, we can compute the probability of the intersection of event A and B, $P(A \cap B)$, as

```
n_AnB <- sum(df$store.trans>0 & df$online.trans>0)

# Number of outcomes that satisfies event A and B
n_AnB
```

```
## [1] 960029
```

```
# Total number of basic outcomes
n
```

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

```
# The probability of the intersection
P_AnB <- n_AnB / n

# Results
P_AnB
```

```
## [1] 0.960029
```

The *conditional probability* of $P(A|B)$, read: "the probability of event A, given that event B has occurred." is derived from

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \text{ provided that } P(B) > 0$$

and

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \text{ provided that } P(A) > 0$$

The given conditional probability $P(B|A)$ is therefore

```
P_BgA<- P_AnB/P_A

# Results
P_BgA
```

```
## [1] 0.960029
```

(d) Are A and B independent events? Why/why not? (1p)

Answer.

Yes, the events A and B are *statistically independent events*, since the following equation holds:

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

```
P_AnB == P_A * P_B
```

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

```
# The same test can be carried out using: setequal()
```

Their statistical independence can also be proven by the **Multiplication Rule of Probabilities**, that is

$$P(A|B) = P(A) \text{ provided that P(B) } > 0$$

and,

$$P(B|A) = P(B) \text{ provided that P(A) } > 0$$

Since these three conditions hold, the events are statistically independent.

(e) Compare and explain (with formula) the similarities/differences among $P(B)$, $P(B \cap A)$ and $P(B \mid A)$ (2p).

Answer.

The first mentioned probability, $P(B)$, expresses the probability of event B occurring, in this case defined as a randomly chosen customer in the sample space having "made at least one online transaction." The definition of $P(B)$ is as follows

$$P(B) = \frac{N_B}{N}$$

where N_B refers to the number of outcomes where event B occurs and the total number of basic outcomes is denoted N.

The second probability mentioned, $P(B \cap A)$ (which is the same as $P(A \cap B)$), refers to the *intersection* of events A and B. The intersection of events translates to both events occurring. In this case, since the events were proven *statistically independent*, the intersection of the two events equals the product of them. If, say, the events instead were *mutually exclusive*, then the intersection would be an *empty set*.

The last mentioned probability, $P(B|A)$, is a *conditional probability*. That is to say, in plain English: "What is the probability of event B given that event A has occurred?"