NDH802 - Assignment 1

Group no.

- 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.

```
#setwd("")
our_group <- 15
```

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
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

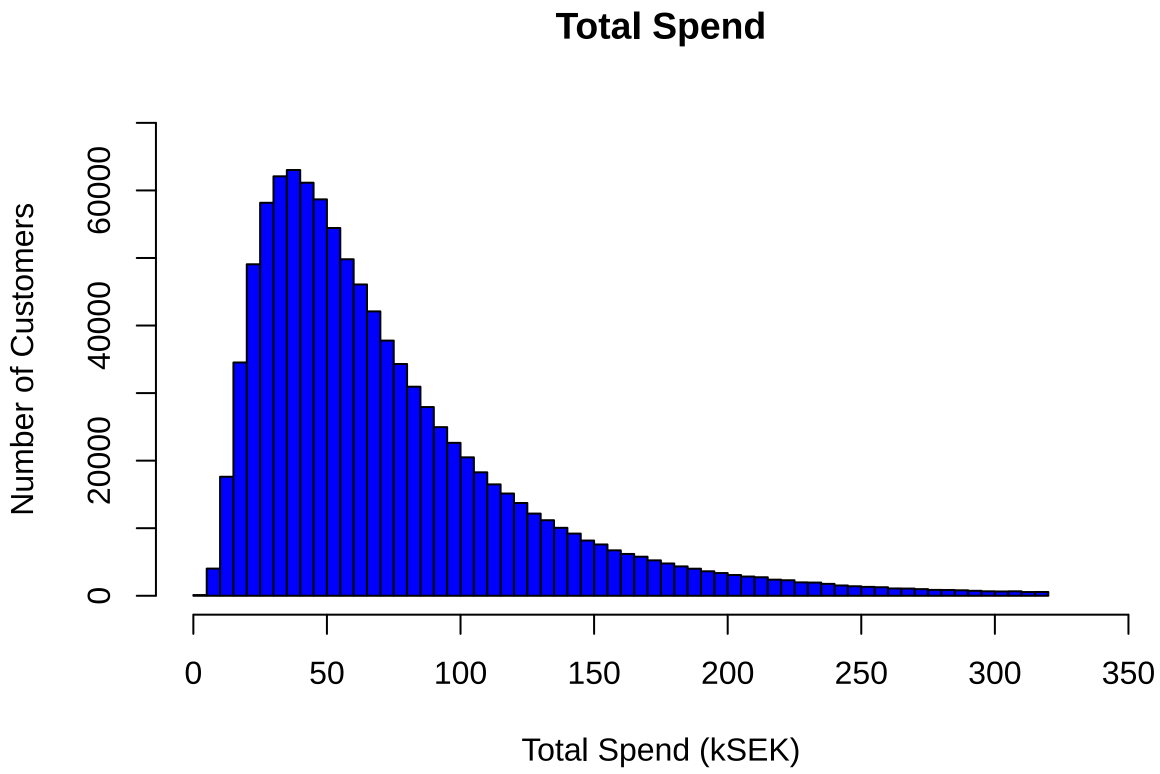
(a) Compute the `total.spend` of the customer (that includes `store.spend` and `online.spend`). Plot the histogram of `total.spend`. Imagine you will present this to your manager. Make it readable and self-explanatory (e.g., add the title for the chart and labels for the axes where needed). (1p) How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

Answer: The histogram is skewed to the right indicating that a majority of the customers spend smaller amounts. This pattern is similar to the typical distribution of income where a large proportion of the population has relatively modest incomes, but the incomes of the top earners extend over a considerable range. The peak of the histogram is the largest number of customers within the defined spend intervals. (The largest number of data points within the intervals)

```
#df[,c("total.spend", "store.spend", "online.spend")]
sum(df[,c("total.spend")])
```

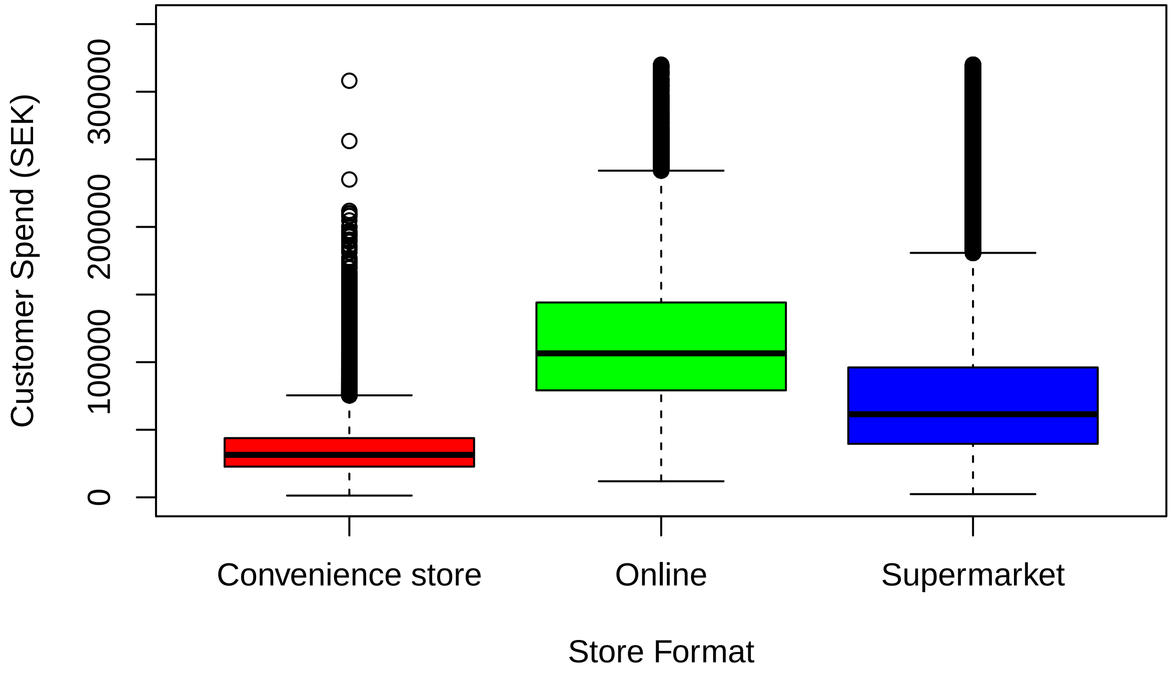
```
## [1] 73073885185
```

```
hist(df[,"total.spend"]/1000, #Define data to be collected
     breaks = 100, #Define breakes for accurate data representation
     xlim = c(0, 350), #Define x-axis limitations
     ylim = c(0, 70000), #Define y-axis limitations
     ylab = "Number of Customers", #Define y-axis name
     xlab = "Total Spend (kSEK)", #Define x-axis name
     col = "blue", #Define graph-colour
     main = "Total Spend") #Define main head
```



(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)

```
boxplot(
  df[,"total.spend"]~df[,"main.format"], #Define data to be collected
  xlab = "Store Format", #Define x-axis name
  ylab = "Customer Spend (SEK)", #Define y-axis name
  col = rainbow(3), #Define box-colour
  ylim = c(-5, 350000)) # adjust the ylim that better illustrates your data
```



```
S_O = sum(df[df$main.format == "Supermarket", "total.spend"]) #Total contribution supermarket
S_ON = sum(df[df$main.format == "Online", "total.spend"]) #Total contribution Online
S_C = sum(df[df$main.format == "Convenience store", "total.spend"]) #Total contribution Convenience Sto
```

Answer: The mean purchase amount for online is higher than both supermarket and convenience store. However, this does not give an indication of the total contribution due to the absence of the total number of translations thus the total contribution. When evaluating the sum of spend for each given category one can identify the actual contribution. In this company the Supermarket is by far the greatest economic contributor, see table below. There is an option to remove the outliers in order to make the visual presentation between the minimum and maximum values easier to read (outline=false). However, for this boxplot we decided to include all the data provided.

Store Format	Total Contribution
Supermarket	6.6050494×10^{10}
Online	3.6229154×10^{9}
Convenience Store	3.4004761×10^{9}

(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).

```
M_Super = mean(df[df$main.format == "Supermarket", "distance.to.store"], na.rm = TRUE)
M_Con = mean(df[df$main.format == "Convenience store","distance.to.store"], na.rm = TRUE)
```

```
V_Super = var(df[df$main.format == "Supermarket","distance.to.store"], na.rm = TRUE)
V_Con = var(df[df$main.format == "Convenience store","distance.to.store"], na.rm = TRUE)
```

	Mean	Variance
Supermarket	4.7752787	33.5813441
Convenience Store	1.3313827	0.3838053

Answer: When evaluating the mean distance to store for the two categories one can see that Supermarkets are located approximately four times the distance from the customer compared to the Convenience Stores. This seems very reasonable given the purpose of the different formats (Supermarkets; larger purchases, ex weekly shopping) (Convenience Stores; on the go purchases, greater accessibility, local). The low variance figure at the Convenience Stores indicate that they are attracting customers with similar distances to stores (in this case locally), while the Supermarkets with a high variance are attracting customers next-doors as well as customers from distant locations.

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)

```
#Basic calculations for Q2a-e
N = nrow(df)
N_OnT = nrow(df[df[,"online.trans"] !="0",])
N_OffT = nrow(df[df[,"store.trans"] !="0",])
N_Intersect = nrow(df[df[,"store.trans"] !="0" & df[,"online.trans"] !="0",])

P_A = N_OffT/N
P_B = N_OnT/N
P_BintA = N_Intersect/N
P_BunionA = P_A+P_B-P_BintA
P_BgivenA = P_BintA/P_A
```

Answer:

$P(A) = 1$

$P(B) = 0.959859$

See calculations of N and P values. P(x) equals the probability of event "x". $P(A) = \frac{N_A}{N}$

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

```
#Calculation for Q2b
P_Bcomp = 1-P_B
```

$P(\bar{B}) = 1 - P(B)$ (Complement Rule)

Let B be an event in the sample space, S. The set of basic outcomes of a random experiment belonging to S but not to B is called the complement of B and is denoted by $\bar{B}$.

Answer: 1 - 0.959859 = 0.040141

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

There are two different routes to be taken in this question. The first route can be calculated by extracting the intersection through our dataset (see previous calc on $P(B \cap A)$). The second route can be calculated given that they are collectively exhaustive ($P(A) = 1$ giving $P(B \cup A) = 1$)

Answer:

$P(B \cup A) = P(A) + P(B) - P(B \cap A)$

$1 = 1 + 0.959859$ - 0.959859

$P(B \mid A) = \frac{P(B \cap A)}{P(A)} = 0.959859$

0.959859 / 1 = 0.959859

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

Answer:

A and B are two events. These events are said to be **statistically independent** if and only if:

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

They are independent events since the outcome of A has no effect on the outcome of B, see below.

```
#Statistically independent if and only if the following is TRUE:
setequal(P_BintA,P_A*P_B)
```

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

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

Answer:

$P(B)$ - Probability of event "B" (number between 0 and 1) within the sample space S. $P(B) = \frac{N_B}{N}$

$P(B \cap A)$ - Probability of both B and A happening. The following formula is used to calculate the intersection if and only if the events are statistically independent. $P(B \cap A) = P(A) * P(B)$

$P(B \mid A)$ - Probability of event B happening given knowledge of A. $P(B \mid A) = \frac{P(B \cap A)}{P(A)}$ The previous two formulas are used to calculate this probability. In this assignment $P(B \mid A)$ gave the same answer as $P(B)$ due to $P(A)$ covering the entire sample space S.

Group 15

Good job!

Thank you for all the extra effort you made to produce such a nicely written report, also all the comments you made along the way. I hope you find them fun and helpful.

Q1a. The histogram and comment on the peak is correct. However, you are not encouraged to comment on something without back up. For example, in our data set we know nothing about the income distribution hence, in order to make such comment, you may want to include other sources of reliable data. 2p

Q1b. Well motivated answer, I really love it. 2p

Q1c. Great calculation, presentation, reasoning. 2p

Q2a. Correct. 1p

Q2b. I am looking for a definition in this context. 0.5p

Q2c. I highly value you take 2 routes to calculate the same quantity. Does it feel good that they produce the same result? 1p

Q2d. You have proven that $P(B \mid A) = P(B)$, which is great! 1p. But I have not seen the reason why either of them are equal to $P(B \cap A)$.