NDH802 - Assignment 1

Group no.4

- 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
setwd("~/Desktop/R/R files/Assignment 1")
our_group <- 4
```

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

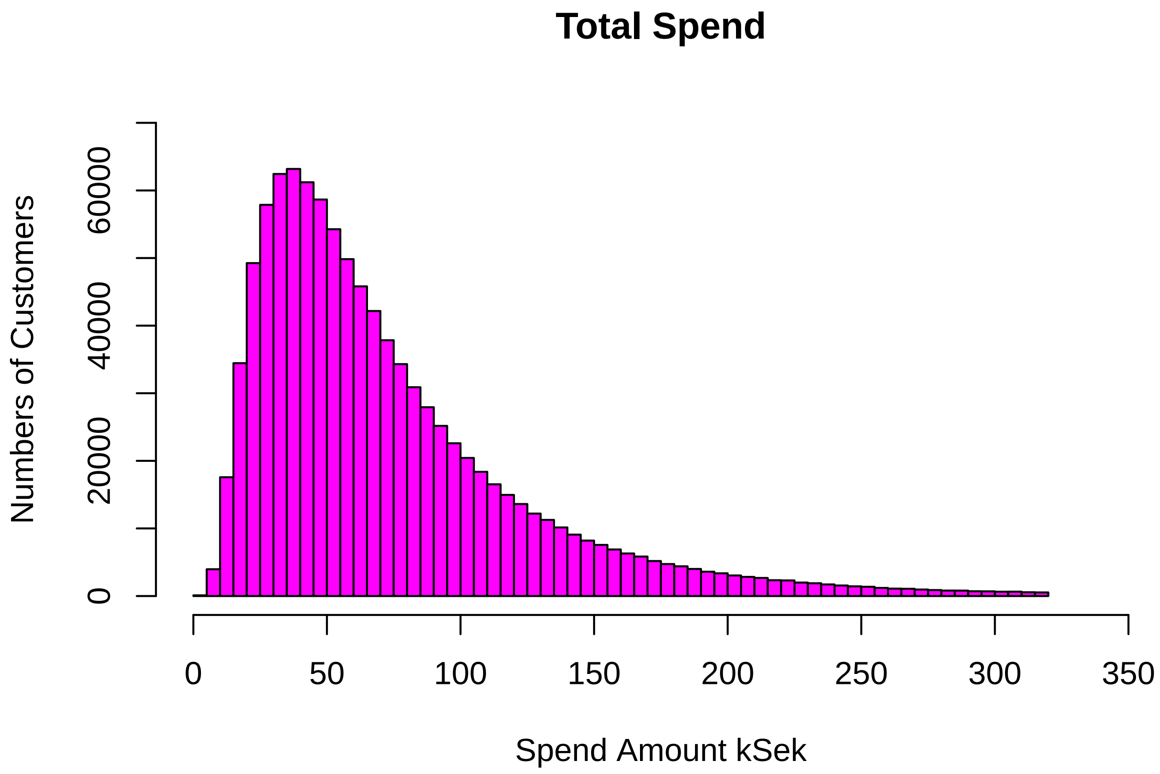
Variable	Description
cust.id	Unique customer id
age	Customer age in year 2021
email	If there is an email of the customer in the system
member.since	Year from which the customer become a member. They can only register at the physical stores.
distance.to.store	Distance in km from customer's address to 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

Variable	Description
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)

```
t_s = sum(df[,c("total.spend")]) #This is the sum of all sales
hist( #Histogram
  df[,"total.spend"]/1000, #Divided total spend by 1000 to get values in kSek
  main = "Total Spend",
  breaks = 100,
  ylab = "Numbers of Customers",
  ylim = c(0,70000),
  xlab = "Spend Amount kSek",
  xlim = c(0, 350),
  col = "magenta"
)
```



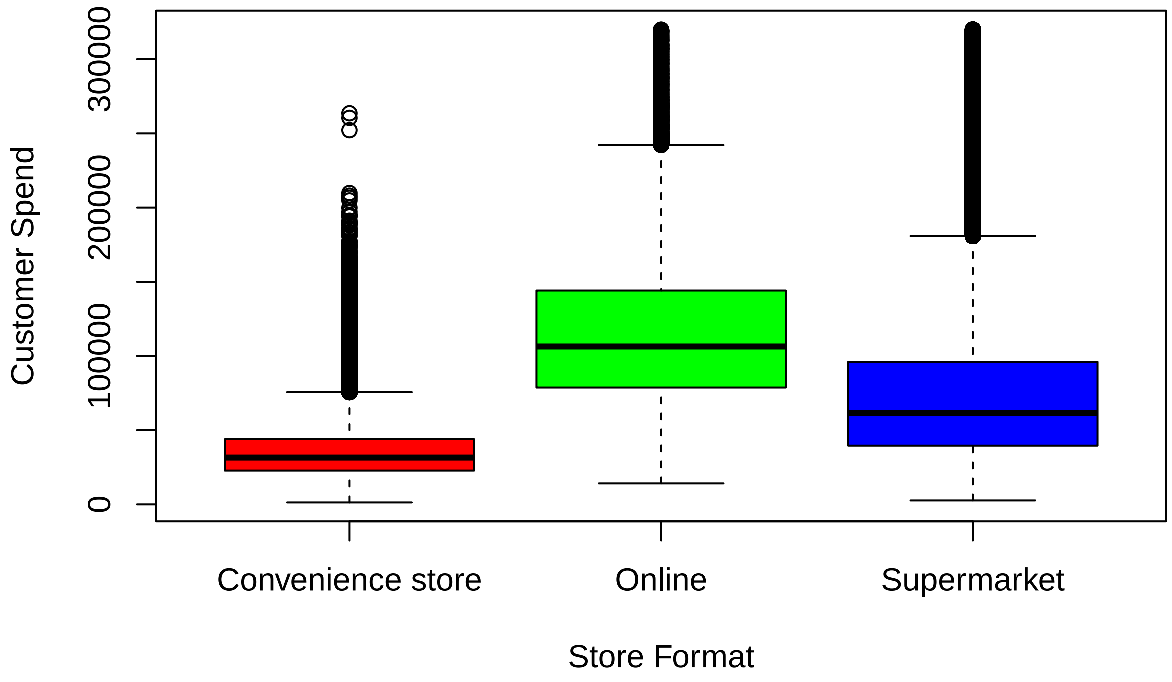
Answer Q1 a)

In general, the peak in the histogram shows which range of a bar that has the most data points in it. If the data has a normal distribution it is also within this range that the mean is found.

The total spend of customers is 7.305873×10^{10}. By looking at the histogram it is evident to say that it is a Skewed-Right (Or positively skewed) distribution which means that the most amount of customers make purchases on the lower-end of the spend amount. This can be expected since it reflects the income of population which is generally also skewed-right. The height of each bar corresponds to the total amount of customers making purchases in the chosen baskets.

(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(
  total.spend ~ main.format,
  data = df,
  xlab = "Store Format",
  ylab = "Customer Spend",
  col = rainbow(3),
  varwidth = FALSE, #With Variable width on the boxes are drawn with widths
  #proportional to the square-roots of the number of observations in the groups.
  outline = TRUE #Change to FALSE to remove Outliers due cluttering
)
```



```
s_o = sum(df[df$main.format == "Online", "total.spend"])
s_s = sum(df[df$main.format == "Supermarket", "total.spend"])
s_c = sum(df[df$main.format == "Convenience store", "total.spend"])
```

Answer Q1 b)

By just looking at the box plot we are not able to state clearly that the online format is contributing the most to the total spend as the box plot is presenting us only with the range of each variable and the patterns of value of each format. We can also see outliers in the data. In this case the sum of online sales is 3.6219253×10^9 compared to the sum of supermarket sales 6.6022236×10^{10}.

(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_s = mean(df[df$main.format == "Supermarket", "distance.to.store"])
m_c = mean(df[df$main.format == "Convenience store", "distance.to.store"])
v_s = var(df[df$main.format == "Supermarket", "distance.to.store"])
v_c = var(df[df$main.format == "Convenience store", "distance.to.store"])
```

Answer Q1 c)

Store Format	Mean	Variance
Supermarket	4.7703945	33.5089996
Convenience Store	1.3306193	0.3833665

As presented in the table above, the mean distance to the convenience store is shorter than the mean distance to the supermarkets, which can be expected as that is the purpose of a convenience store. The difference in variance is most likely due to the fact that there are generally fewer stores and people will travel longer distances to them.

Question 2. Probability theory

Consider the following events:

(A) Made at least one offline transaction

(B) Made at least one online transaction

```
n_st = nrow(df[df[,"store.trans"] != "0",])
#Amount of customer with at least 1 offline transaction
n_ot = nrow(df[df[,"online.trans"] != "0",])
#Amount of customer with at least 1 online transaction
N = nrow(df) #Total amount of customers
```

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

```
prob_a = n_st/N #Probability of A
prob_b = n_ot/N #Probability of B
```

Answer Q2 a)

The $P(A)$ is 1 and the $P(B)$ is 0.959849.

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

```
comp_b = 1-prob_b #Complement of B as defined by the Complement Rule
```

Answer Q2 b)

In general the event $\bar{B}$ is defined as all the cases within the data set that are not within event B . So for this example the event $\bar{B}$ is defined as all people who have not made at least one online transaction. In reality everyone who has made 0 online transactions, because you can only make a whole number of transactions and you can not make a negative amount of transactions.

The definition of Complement B is $P(\bar{B}) = 1 - P(B)$ and the $P(\bar{B})$ is 0.040151

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

```
b_int_a = nrow(df[df[,"store.trans"] !="0" & df[,"online.trans"] !="0",])/N #Probability of B intersect
b_giv_a = b_int_a/prob_a #B given A
```

Answer Q2 c)

The $P(B \cap A)$ is 0.959849

$P(B \mid A) = \frac{P(B \cap A)}{P(A)} = 0.959849$

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

```
stat_ind = setequal(b_int_a,prob_a*prob_b) #A and B are independent If and only If this function is TRU
```

Answer Q2 d)

A and B are statistically independent if and only if $P(B \cap A) = P(A)P(B)$ which is TRUE. In other words if a costumer chooses to buy online will not affect offline purchases.

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

Answer Q2 e)

$P(B) = \frac{N_B}{N}$ We can find the probability of event B by looking at how many times that event will occur in the sample space.

$P(B \cap A)$ The probability of all basic outcomes in the sample space that belong to both B and A.

$P(B \mid A) = \frac{P(B \cap A)}{P(A)}$ B given A will show how many times the event B will occur if event A occurs.

In our case we can see that the event A covers the entire sample space which means that any event B will be included in event A and they will intersect. Likewise all events of B given A will occur. This is why all the values will we the same in this case.