Assignment 2: Survey developpment

105.708 Data Acquisition and Survey Methods

Group 3: Gjorgieva Aleksandra(), de Lambertye Grégoire(12202211), Loos Annika Katharina()

2023-06-14

Contents

I Introducction

topic

blabla

Hypothesis

blabla

question

blaba

Data exploration

```
file_path <- "./data/group_3.csv"

rawData <- read.csv(file_path, sep=";")
# rename the column in a english friendly format
colnames(rawData)[4] <- "answer_1"
colnames(rawData)[5] <- "answer_2"
colnames(rawData)[6] <- "answer_3"

print(head(rawData))
```

```
##   Gender Age             Academic.Program answer_1
## 1 female  23 Data Science / MSc / TU Wien      120
## 2   male  21 Data Science / MSc / TU Wien      180
## 3 female  23 Data Science / MSc / TU Wien       90
## 4 female  24 Data Science / MSc / TU Wien      140
## 5   male  22 Data Science / MSc / TU Wien      100
## 6   male  25                 Data Science       50
##                                   answer_2
## 1                                       4
## 2 5 (I think I should send way less time)
## 3 5 (I think I should send way less time)
## 4                                       4
## 5                                       4
## 6                                       4
##                                      answer_3
## 1                              YouTube video
## 2                              YouTube video
## 3 YouTube shorts - Tiktoks - Intsagram reels
## 4                          TV shows (series)
## 5                              YouTube video
## 6                              YouTube video
```

```
print(paste("The data set dimensions are ", dim(rawData)[1], " x ", dim(rawData)[2]))
```

```
## [1] "The data set dimensions are  38  x  6"
```

Preprocessing

The colummns Gender and Age look nice and don't requier preprocessing.

Academic programm

```
print(unique(rawData[3]))
```

```
##                                   Academic.Program
## 1                     Data Science / MSc / TU Wien
## 6                                     Data Science
## 9              Business Informatics / BSc/ TU Wien
## 10                                MSc Data Science
## 11                     Data Science / MSc / TU WIen
## 12                                Data Science Msc
## 17 Statistik und Wirtschaftsmathematik BSc TU Wien
## 23                               Data Science MSc.
## 25                                             MSc
## 28                       Erasmus student, Statistic
## 34       Data Science / MSc / University of Zagreb
## 36             Business Informatics / BSc / TU Wien
```

As we see the format differs for the academic program answer. Where we have missing information we associate them by default to Data Science / MSc / TU Wien. To say we change “Data Science”, “MSc Data Science”, “MSc”, “Data Science Msc”, ” Data Science MSc.” to “Data Science / MSc / TU Wien”.

```
row_index <- which(rawData$Academic.Program %in% c("Data Science",
                                                   "Data Science / MSc / TU Wien",
                                                   "MSc Data Science",
                                                   "Data Science / MSc / TU WIen",
                                                   "MSc",
                                                   "Data Science Msc",
                                                   "Data Science MSc."))

rawData[row_index, "Academic.Program"] <- "Data Science / MSc / TU Wien"
```

```
print(unique(rawData[3]))
```

```
##                                  Academic.Program
## 1                      Data Science / MSc / TU Wien
## 9              Business Informatics / BSc/ TU Wien
## 17 Statistik und Wirtschaftsmathematik BSc TU Wien
## 28                       Erasmus student, Statistic
## 34       Data Science / MSc / University of Zagreb
## 36             Business Informatics / BSc / TU Wien
```

```
rawData$answer_2 <- str_sub(rawData$answer_2, end = 2)
rawData$answer_2 <- as.numeric(rawData$answer_2)
```

```
for (i in 1:length(rawData$answer_1)){
  if(rawData$answer_1[i] <= 5){
    rawData$answer_1[i] <- rawData$answer_1[i] * 60
  }
}
```

```
head(rawData)
```

```
##   Gender Age             Academic.Program answer_1 answer_2
## 1 female  23 Data Science / MSc / TU Wien      120        4
## 2   male  21 Data Science / MSc / TU Wien      180        5
## 3 female  23 Data Science / MSc / TU Wien       90        5
## 4 female  24 Data Science / MSc / TU Wien      140        4
## 5   male  22 Data Science / MSc / TU Wien      100        4
## 6   male  25 Data Science / MSc / TU Wien       50        4
##                                     answer_3
## 1                              YouTube video
## 2                              YouTube video
## 3 YouTube shorts - Tiktoks - Intsagram reels
## 4                          TV shows (series)
## 5                              YouTube video
## 6                              YouTube video
```

Analysis

hypothesis 1

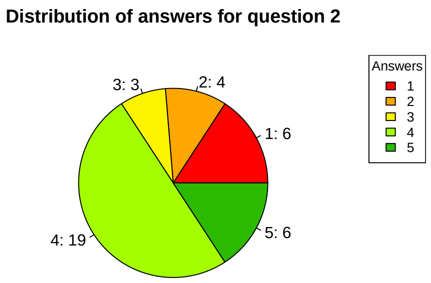
hypothesis 2

hypothesis 3

The 3rd hypothesis was - *Students think they spend too much time watching video content, regardless on the time they spend doing it.*. In other word, we are looking for the correlation between the time spend watching video and the answered provided for the question 2. Let's have first a deeper look to those columns. The following table show the answer repartition.

```
pie( table(rawData$answer_2),
    main = "Distribution of answers for question 2",
    col = c("#ff0000", "#ffa700", "#fff400", "#a3ff00", "#2cba00"),
    labels = paste0(names(table(rawData$answer_2)), ": ", table(rawData$answer_2)))

legend("topright",
       legend = names(table(rawData$answer_2)),
       fill = c("#ff0000", "#ffa700", "#fff400", "#a3ff00", "#2cba00"),
       cex = 0.8,
       title = "Answers")
```



A majority of the answers are between 4 and 5 (66%). This let us think that people have a generaly a bad opinion on the time they spend watching video content. But how is it correlated to the real time they spend?

```
correlation <- cor(rawData$answer_1, rawData$answer_2)
print(paste('The correlation is ', correlation))
```

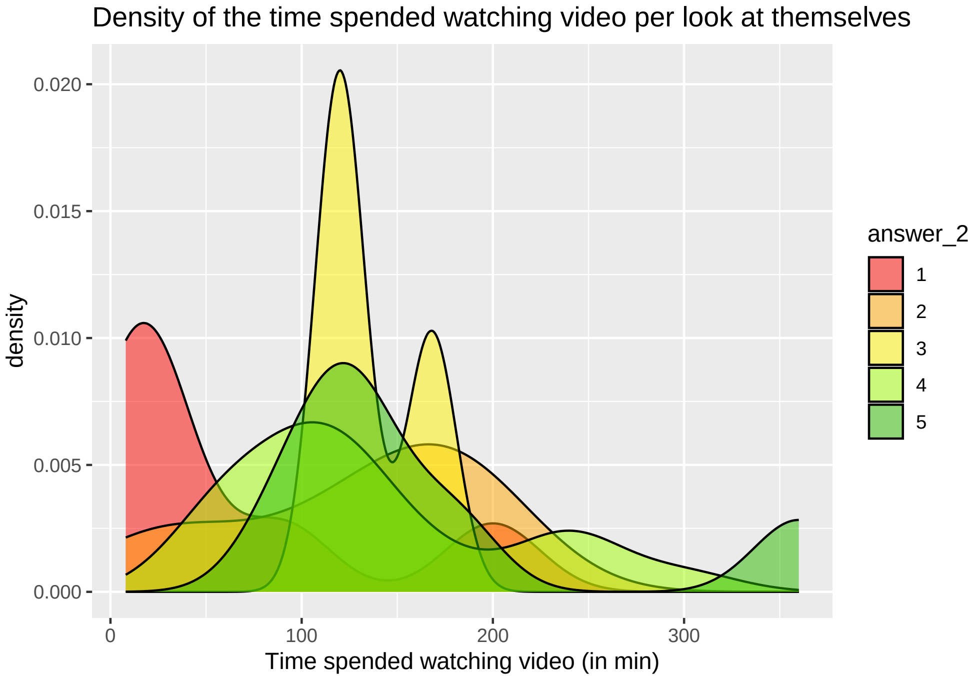
```
## [1] "The correlation is  0.378687920940175"
```

The correlation between the 2 column is 0.38, a strong correlation is close to 1 and a correlation between 0.3 and 0.5 is generally considered as low.

So far we saw that a majority of the people think they spend too much time watching video content and this is weakly correlated to the real time. We could accept our hyothis but the indicators are not straightforward. A more visual approach is to look at the distribution of the time regarding the "look at themself". This have been done in the following graph.

```
plot <- ggplot(rawData, aes(x = answer_1, fill = as.factor(answer_2))) +
  geom_density(alpha = 0.5) +
  scale_fill_manual(values =
                      c("#ff0000", "#ffa700", "#fff400", "#a3ff00", "#2cba00")) +
  labs(fill = "answer_2") +
  xlab("Time spended watching video (in min)") +
  ggtitle("Density of the time spended watching video per look at themselves")

plot
```

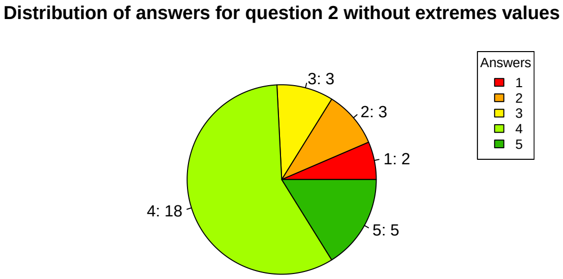


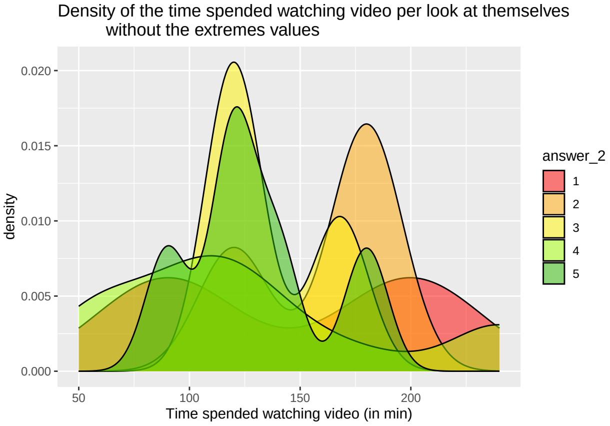
On this graph we can see that the density for answer 1 ("I don't think I spend too much time") has 2 spikes, one small (time < 50 min) and one around 200 mins. For the opposite answer (5, I think I should spend less time watching video content) we have a spike around 120 minutes and a second one around time > 320 minutes. For the other distribution they are pretty similar and centered around 120 minutes. We have to take in account that some densities are computed on very few data (as for answer 3 where only 3 peoples selected this answer).

If our hypothesis were exact we should only have people with answer 4 or 5 and the same densities for the answers. But here we clearly see that when people watch video less than 30 min per day they have a good opinion on it (1 or 2). On the opposite every person that watch video more than 300 minutes per day think he/she should watch less video. So **we reject our third hypothesis**.

Diving a bit deeper in the analysis made us consider the values without the “outliers”, to say without the values under 30 minutes (5 peoples) and more than 300 minutes a day (2 peoples). We recomputed the correlation and redraw the previous plot.

```
## [1] "In the new dataset, the correlation is  -0.154480818505668"
```





Without the extreme values, the correlation drastically drops and is even negative that mean the more people watch video the better they feel with it. In this region people’s look on themselves is independent to the time they spend watching video.

In conclusion we can say that **we reject our hypothesis** but our idea seems valid outside of extreme values. We also have to stress out the fact that this survey have been done on few people only.

Conclusion