Assignment 2: Survey developpment

105.708 Data Acquisition and Survey Methods

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

2023-06-19

Contents

1 Introducction

Our survey was designed to delve into students' daily habits regarding their consumption of content from a variety of sources. We were primarily interested in understanding the amount of time they dedicate to watching videos, the types of content they gravitate towards, whether it be short, fast-paced videos or longer ones that require more attention, and finally, their personal opinions on how much time they should devote to this activity. Given that we were limited to three questions, we believed these to be the most appropriate to gain insights into these aspects.

1.1 Questions

- How many hours do you spend per day watching video content?
- Quantitative: (integer) number in minutes
- Do you think you spend too much time watching video content?
- Ordinal: scale from 1(not at all) to 5(I think I should send way less time) - From what category do you consume the most video content from?
- Categories: o YouTube shorts – Tiktoks – Intsagram reels o YouTube video o TV shows (series) o Movies o None of them

1.2 Hypothesis

1. "On average students spend more than a day per week watching video content."
2. "Student that identify themselves as female are more prone to consume longer video content than others."
3. "Students think they spend too much time watching video content, regardless on the time they spend doing it."

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

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

1.3.1 Preprocessing

The columnns Gender and Age look nice and don't require preprocessing. We had to clean the data for the for the academic programs and for the answer 1.

```
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
  }
}

knitr::kable(rawData[1:3,],
             caption = "Dataset cleaned")
```

Table 1: Dataset cleaned

Gender	Age	Academic.Program	answer_1	answer_2	answer_3
female	23	Data Science / MSc / TU Wien	120	4	YouTube video
male	21	Data Science / MSc / TU Wien	180	5	YouTube video
female	23	Data Science / MSc / TU Wien	90	5	YouTube shorts – Tiktoks – Intsagram reels

2 Hypothesis

2.1 hypothesis 1

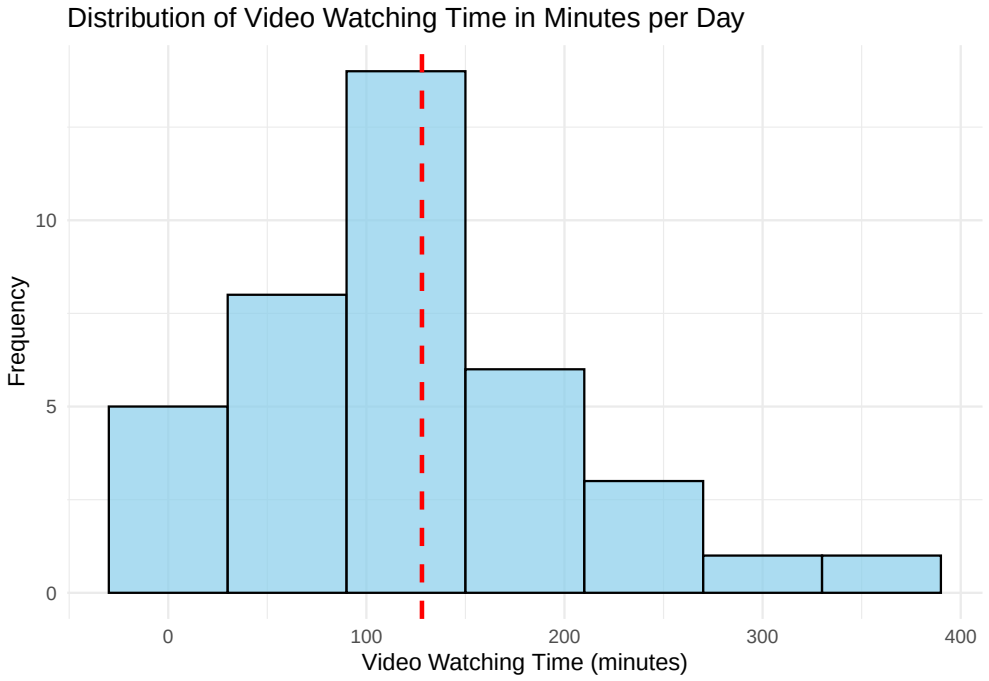
"On average students spend more than a day per week watching video content"

In order to answer to test this hypothesis it will be necessary to look at that mean of the sample as time spent watching video content and further present it in an hourly and weekly context. We will primarily look

and plot the distribution of the answers to see how if there are some outliers which will need to be handled. There were no extremes that needed altering. However, we hod some answers in hours and some in minutes and we have processed the data above to have only minutes in our answers. In the graph we can see that out of 38 people more than 15 answered around 100-150 minutes a day and the red line represent the mean of our sample.

```
p1 <- ggplot(rawData, aes(x = answer_1)) +
  geom_histogram(binwidth = 60, fill = "skyblue", color = "black", alpha = 0.7) +
  geom_vline(aes(xintercept = mean(answer_1, na.rm = TRUE)),
             color = "red", linetype = "dashed", size = 1) +
  labs(title = "Distribution of Video Watching Time in Minutes per Day",
       x = "Video Watching Time (minutes)", y = "Frequency") +
  theme_minimal()

print(p1)
```




```
mean_answer1 <- mean(rawData$answer_1, na.rm = TRUE)
print(paste(paste("The mean time spend watching video content per day is ", mean_answer1), " in minutes
```

```
## [1] "The mean time spend watching video content per day is  128.052631578947  in minutes."
```

```
rawData <- rawData %>%
  mutate(answer_1_weekly = answer_1 * 7 / 60) # converting from minutes to hours

ttest <- t.test(rawData$answer_1_weekly, mu = 24, alternative = "greater")
ttest
```

```
##
##  One Sample t-test
```

```
##
## data:  rawData$answer_1_weekly
## t = -6.1217, df = 37, p-value = 1
## alternative hypothesis: true mean is greater than 24
## 95 percent confidence interval:
##  12.44246      Inf
## sample estimates:
## mean of x
##  14.93947
```

Here we have tested our hypothesis using t-test.The parameter mu is set to 24, which is the value that the sample mean is being compared to. The parameter alternative = "greater" specifies the alternative hypothesis, which is that the true mean is greater than mu.

The results of our t-test show a t-value of -6.1217 with 37 degrees of freedom, and the p-value is 1. The null hypothesis is that the mean weekly video watching time is equal to 24 hours. A p-value of 1 means that if the null hypothesis were true, we would expect to see data like our observed data all the time. Contrary to our hypothesis that students watch more than 24 hours of video content per week, the data does not support this. In fact, our sample's mean weekly watching time is about 14.94 hours, which is less than 24 hours.

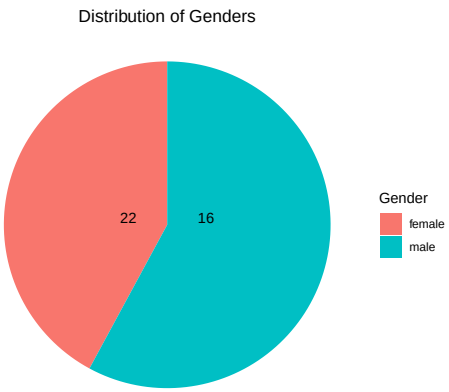
2.2 Hypothesis 2

Our second hypothesis was "Student that identify themself as female are more prone to consume longer video content than others.". For the categorical features in the data set, one can use charts to visualize the distribution of each category.

2.2.1 Distribution of Gender

In our study there where 22 males and 16 females taking part, shown in a pie chart.

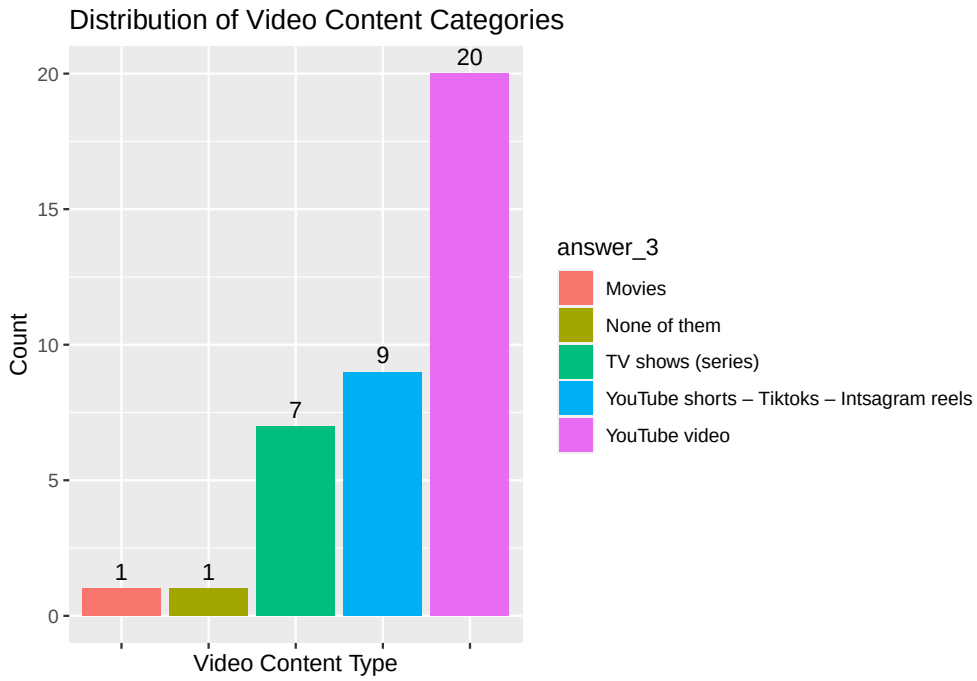
```
ggplot(rawData, aes(x = "", fill = Gender)) +
  geom_bar(width = 1) +
  coord_polar(theta = "y") +
  labs(x = NULL, y = NULL, fill = "Gender") +
  ggtitle("Distribution of Genders") +
  theme_void() +
  theme(plot.title = element_text(hjust = 0.5, vjust = -0.3)) +
  geom_text(stat = "count", aes(label = after_stat(count)), vjust = -7)
```



2.2.2 Distribution of type of Video Content

The third question of our study was “From what category do you consume the most video content from?”, where the participants could choose between five categorical values: “Movies”,“TV shows (series)”, “YouTube shorts – Tiktoks – Intsagram reels”, “Youtube videos” or “None of them”. In the following bar chart one can visually see the distriubtion between the five values with “Youtube Video” beeing the most popular category, because it was chosen 20 times. On the other hand only one person watches mainly movies.

```
ggplot(rawData, aes(x = answer_3, fill = answer_3)) +
  geom_bar(show.legend = TRUE) +
  labs(x = "Video Content Type", y = "Count") +
  theme(axis.text.x = element_blank()) +
  geom_text(stat = "count", aes(label = after_stat(count)), vjust = -0.5) +
  ggtitle("Distribution of Video Content Categories")
```



2.2.3 Descriptive Inference

In general one can see by this table that females taking part in our study in average watch less video content than male participants.

```
summary_stats <- rawData %>%
  group_by(Gender) %>%
  summarize(average_duration = mean(answer_1),
            median_duration = median(answer_1))

summary_table <- data.frame(Gender = summary_stats$Gender,
                            Average_Duration = summary_stats$average_duration,
                            Median_Duration = summary_stats$median_duration)
```

```
knitr::kable(summary_table,
             caption = "Average and meadian daily consuption time of video")
```

Table 2: Average and meadian daily consuption time of video

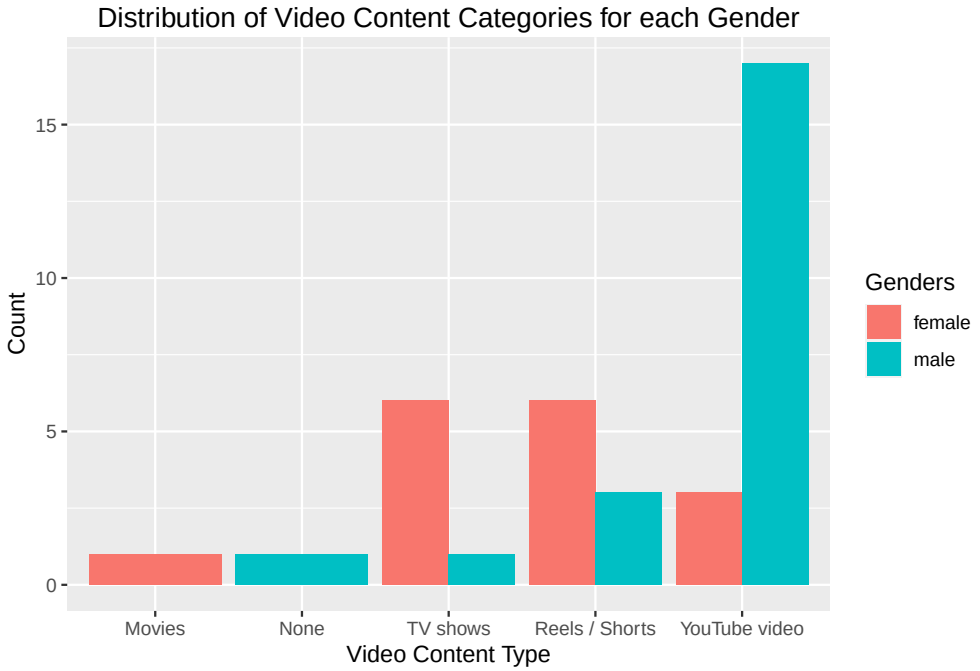
Gender	Average_Duration	Median_Duration
female	103.1250	120
male	146.1818	120

```
rawData[, "Gender"] <- ifelse(rawData[, "Gender"]=="male",1,0)
rawData$answer_3 <- factor(rawData$answer_3, levels = c( "Movies", "None of them", "TV shows (series)",
                                                         "YouTube shorts - Tiktoks - Intsagram reels",
rawData$answer_3 <- as.integer(rawData$answer_3)
rawData$Gender <- as.numeric(rawData$Gender)
```

In the following chart one can see that males watch more video content of the category “YouTube video” than females. The rest is fairly balanced.

```
Genders <- factor(rawData$Gender, labels = c("female", "male"))

ggplot(rawData, aes(x=factor(answer_3),fill = Genders)) +
  geom_bar(position = "dodge") +
  labs(x = "Video Content Type", y = "Count") +
  ggtitle("Distribution of Video Content Categories for each Gender") +
  theme(plot.title = element_text(hjust = 0.5, vjust = -0.3)) +
  scale_x_discrete(labels = c("Movies", "None", "TV shows", "Reels / Shorts", "YouTube video"))
```



2.2.4 Analytic Inference

```
contingency_table <- table(rawData$Gender, rawData$answer_3)

chi_squared_test <- chisq.test(contingency_table)
chi_squared_test
```

```
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 15.818, df = 4, p-value = 0.003273
```

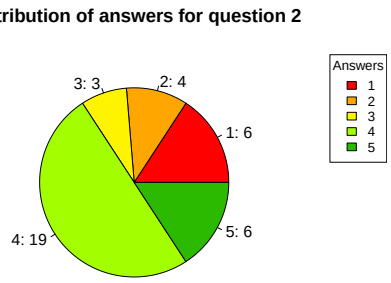
The p-value is a measure of the statistical significance of the association. In this case, the p-value is 0.003273. A p-value less than the chosen significance level (e.g., 0.05) indicates that the association between gender and the categorical value in the fifth column is statistically significant. One can see that more males watch Youtube Videos which can be put into a category of longer content and more females watch video content of the category "YouTube shorts – Tiktoks – Intsagram reels". Therefore the hypothesis of "Student that identify themself as female are more prone to consume longer video content than others" can be rejected.

2.3 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))
```

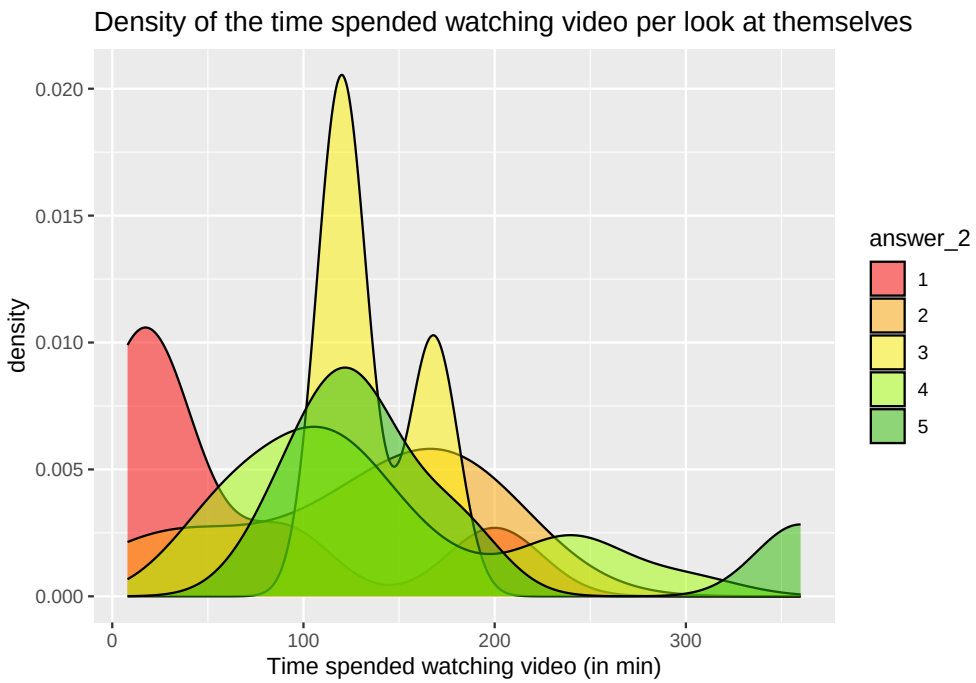
```
## [1] "The correlation is  0.378687920940175"
```

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
```



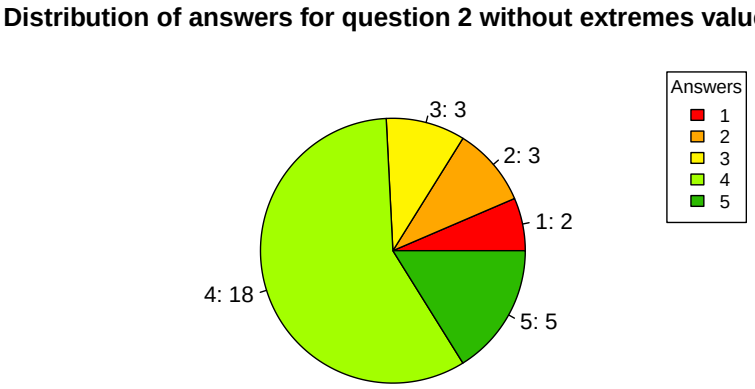
On this graph we can see that the density for answer 1 ("I don't think I spend too much time") has 2 spikes, one small (time < 50 min) and one around 200 mins. For the opposite answer (5, I think I should spend less time watching video content) we have a spike around 120 minutes and a second one around time > 320 minutes. For the other distribution they are pretty similar and centered around 120 minutes. We have to

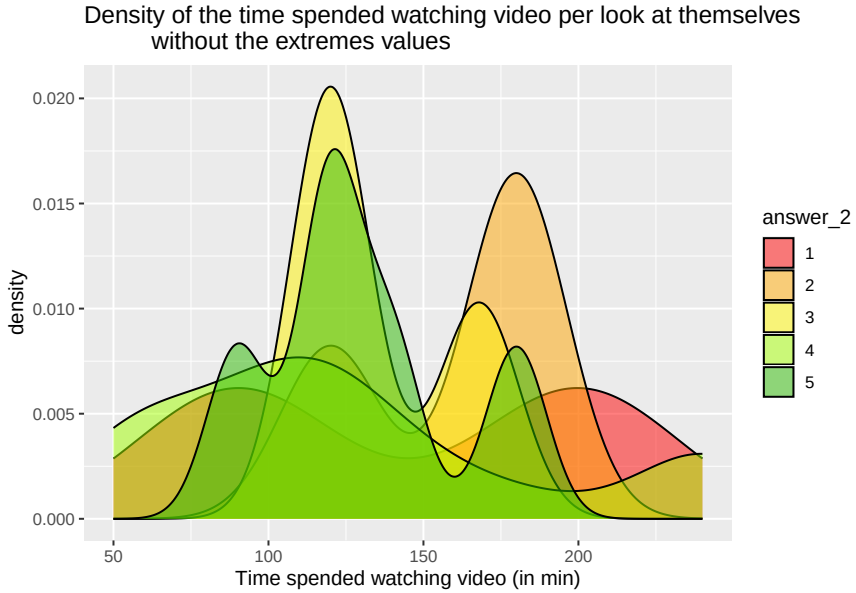
take in account that some densities are computed on very few data (as for answer 3 where only 3 peoples selected this answer).

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.

3 Conclusion

Thanks to our survey questions we could answer our hypothesis. We could gain a lot of insights ont the participant's video content consumption habits. Based on the collected data We reject all our hypothesis