Data Acquisition Assignment 2 Report

Artur Ohanian (12239164), Artur Sogomonyan (12143554), Anastasia Cissa (11937948)

2023-06-22

Contents

Introduction

This report presents the results of a survey data analysis project. The goal of the analysis is to explore and interpret the categorical and quantitative features in the data set. By answering specific research questions, we aim to gain valuable insights from the survey data. This report provides a summary of the exploratory data analysis, descriptive inference, and analytic inference conducted. The findings and interpretations presented here contribute to our understanding of the subject matter.

Data Set Description

The data set for this assignment was created from the survey answers. It contains columns about gender, age, and academic program. The academic program column includes information not only about the program itself but also about the degree and university name. This column needs to be altered during preprocessing. Additionally, there are three answers, two of which are qualitative variables. The first is social media, indicating where the respondent spends the most time, and the second is how the respondent actually spends time on social media (activity). The qualitative variable represents the mean time in hours spent on social media per day.

Exploratory Data Analysis

Data Preprocessing

In this section, we will clean the data and prepare it for analysis. We will remove unnecessary columns, rename the columns, and create new columns. We will also remove outliers and missing values.

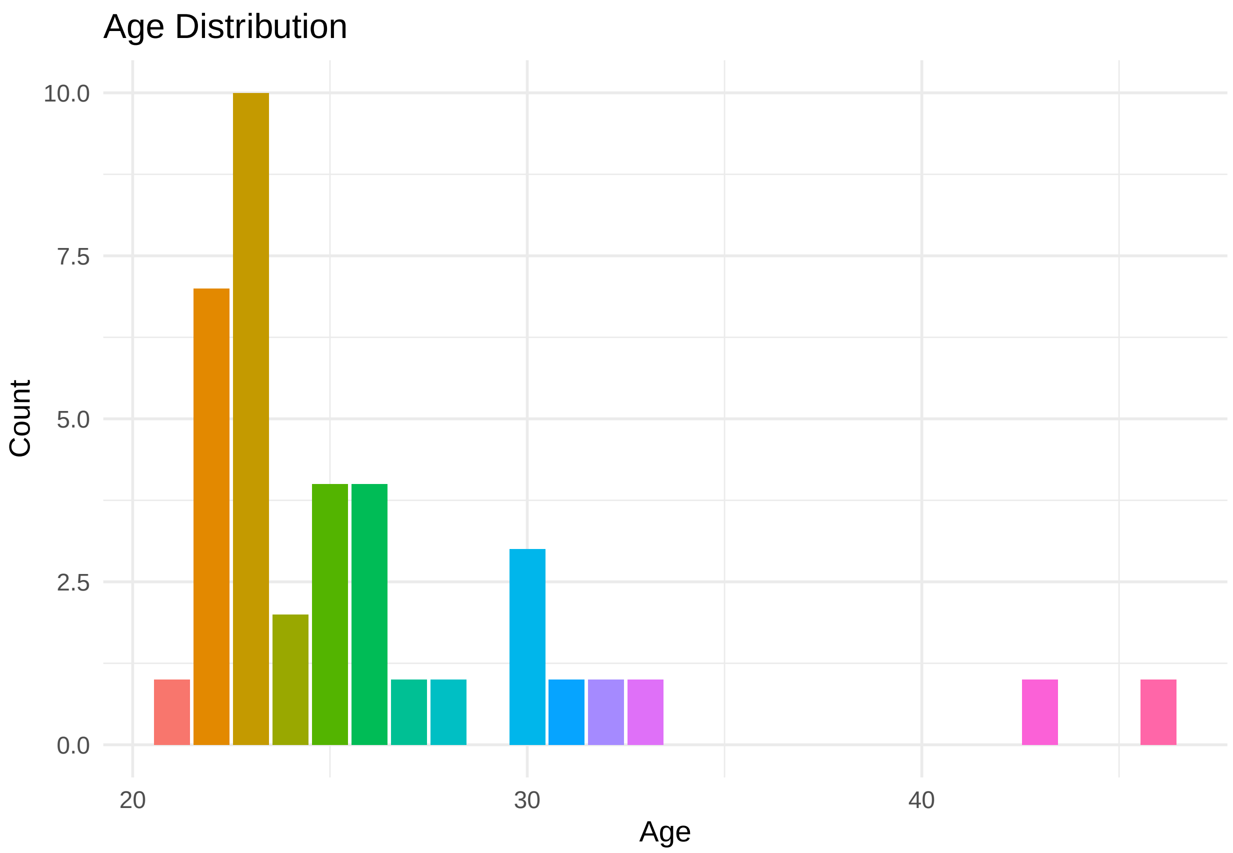
```r
library(tidyr)
library(dplyr)
library(ggplot2)
library(readr)
library(stringr)
library(AICcmodavg)
library(corrplot)

res <- read_delim("survey_results.csv", delim =";") %>%
  mutate(Program = str_extract(`Academic Program`,
  "(Data Science|Business Informatics|Statistic|
  Statistik und Wirtschaftsmathematik)"),
         Degree = str_extract(`Academic Program`,
         "[M/B][S/s]c"),
         University = str_extract(`Academic Program`,
         "(TU W[i\\I]en|University of Zagreb|Erasmus student)")) %>%
  select(-`Academic Program`) %>%
  select(Gender, Age, `Academic Program` = Program, Degree, University,
         `Social Media` = `Antwort 1`,
          `Time` = `Antwort 2`, `Activity` = `Antwort 3`)

res$Degree <- ifelse(is.na(res$Degree) &
 res$`Academic Program` == "Data Science", "MSc", res$Degree)
res$University <- ifelse(is.na(res$University) &
 res$`Academic Program` == "Data Science", "TU Wien", res$University)
res
```

```
## # A tibble: 38 x 8
##    Gender   Age ‘Academic Program‘   Degree University ‘Social Media‘  Time
##    <chr>  <dbl> <chr>                <chr>  <chr>      <chr>          <dbl>
##  1 female    23 Data Science         MSc    TU Wien    YouTube          3
##  2 male      21 Data Science         MSc    TU Wien    YouTube          6
##  3 female    23 Data Science         MSc    TU Wien    Instagram        2
##  4 female    24 Data Science         MSc    TU Wien    Instagram        1.5
##  5 male      22 Data Science         MSc    TU Wien    YouTube          2
##  6 male      25 Data Science         MSc    TU Wien    Instagram        1
##  7 male      26 Data Science         MSc    TU Wien    YouTube          5
##  8 male      43 Data Science         MSc    TU Wien    Whatsapp         1
##  9 male      23 Business Informatics BSc    TU Wien    YouTube          2
## 10 female    26 Data Science         MSc    TU Wien    Instagram        3
## # i 28 more rows
## # i 1 more variable: Activity <chr>
```

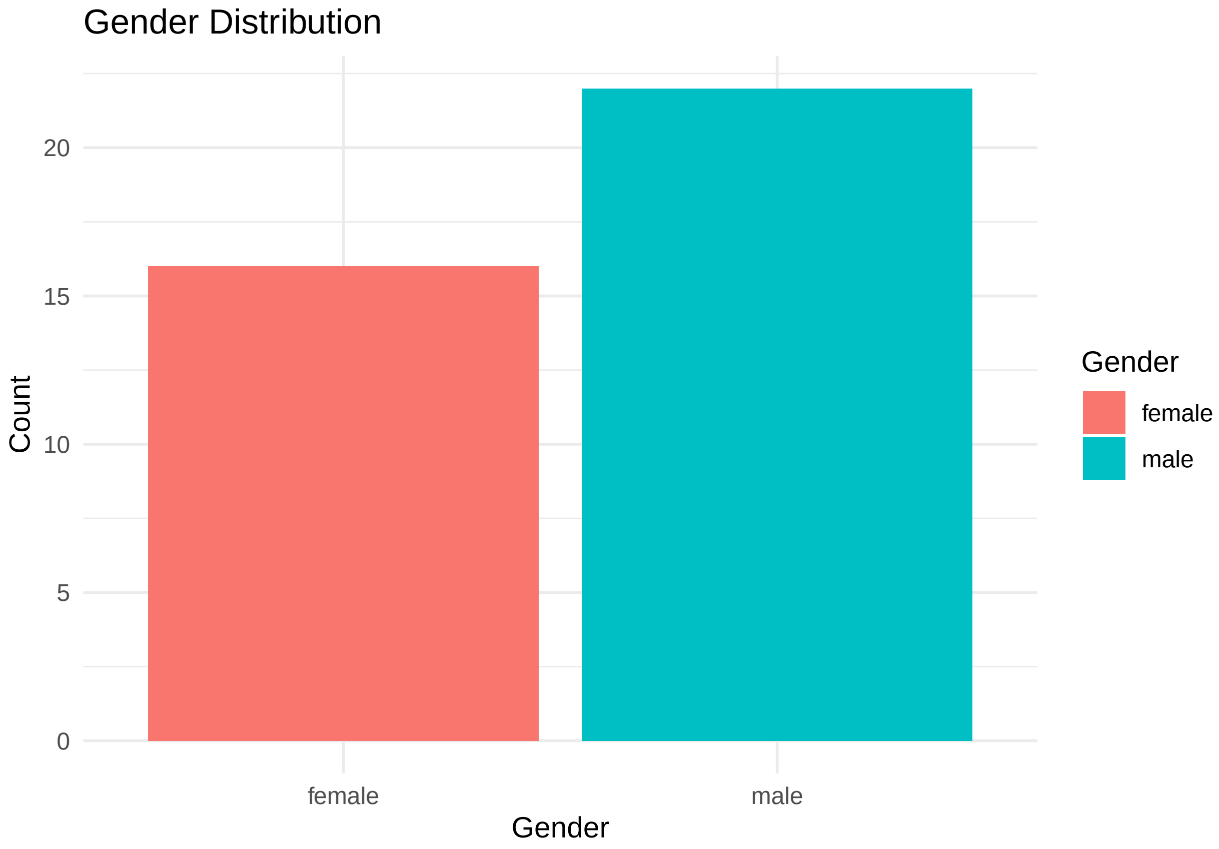
Data Analysis

Firstly, we would like to research age and gender distribution of recipients, participating in the survey.

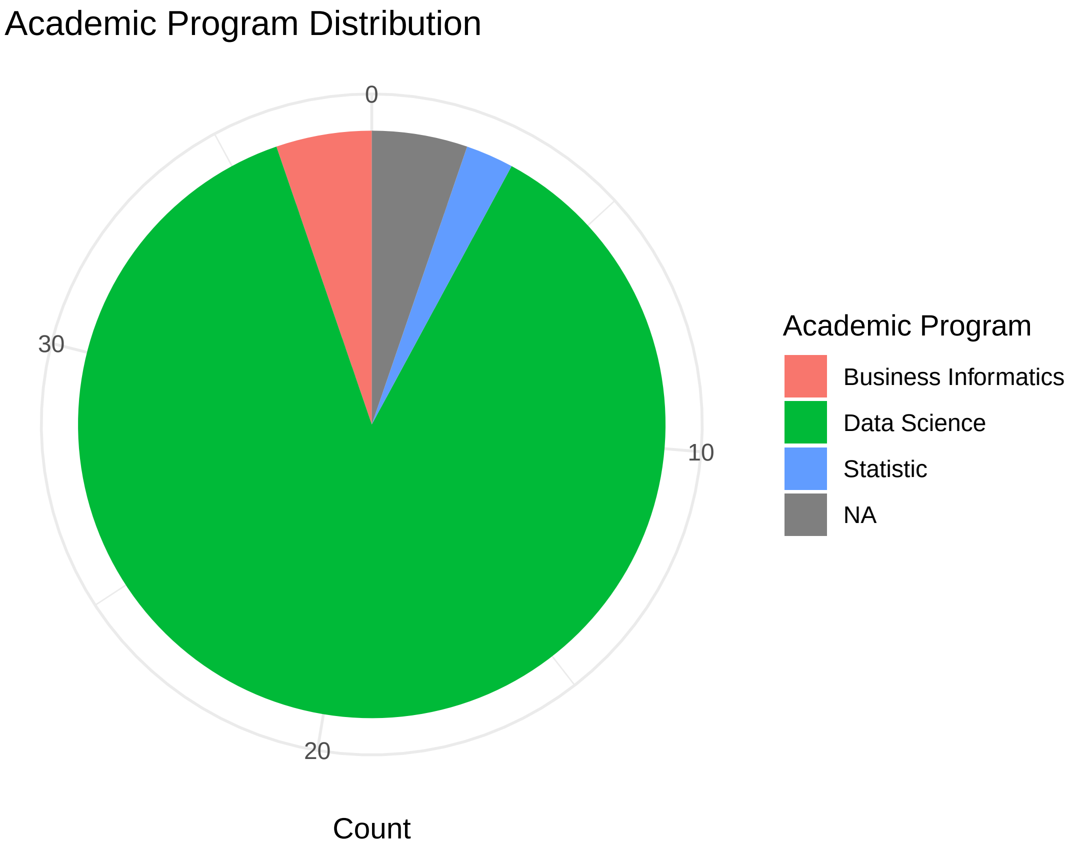
```
plot <- ggplot(res, aes(x = Age, fill = as.factor(Age))) +
  geom_bar() +
  labs(title = "Age Distribution") +
  xlab("Age") +
  ylab("Count") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme_minimal()

plot + theme(legend.position = "none")
```



```
plot1 <- ggplot(res, aes(x = Gender, fill = Gender)) +
  geom_bar() +
  labs(title = "Gender Distribution") +
  xlab("Gender") +
  ylab("Count") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme_minimal()

plot1
```

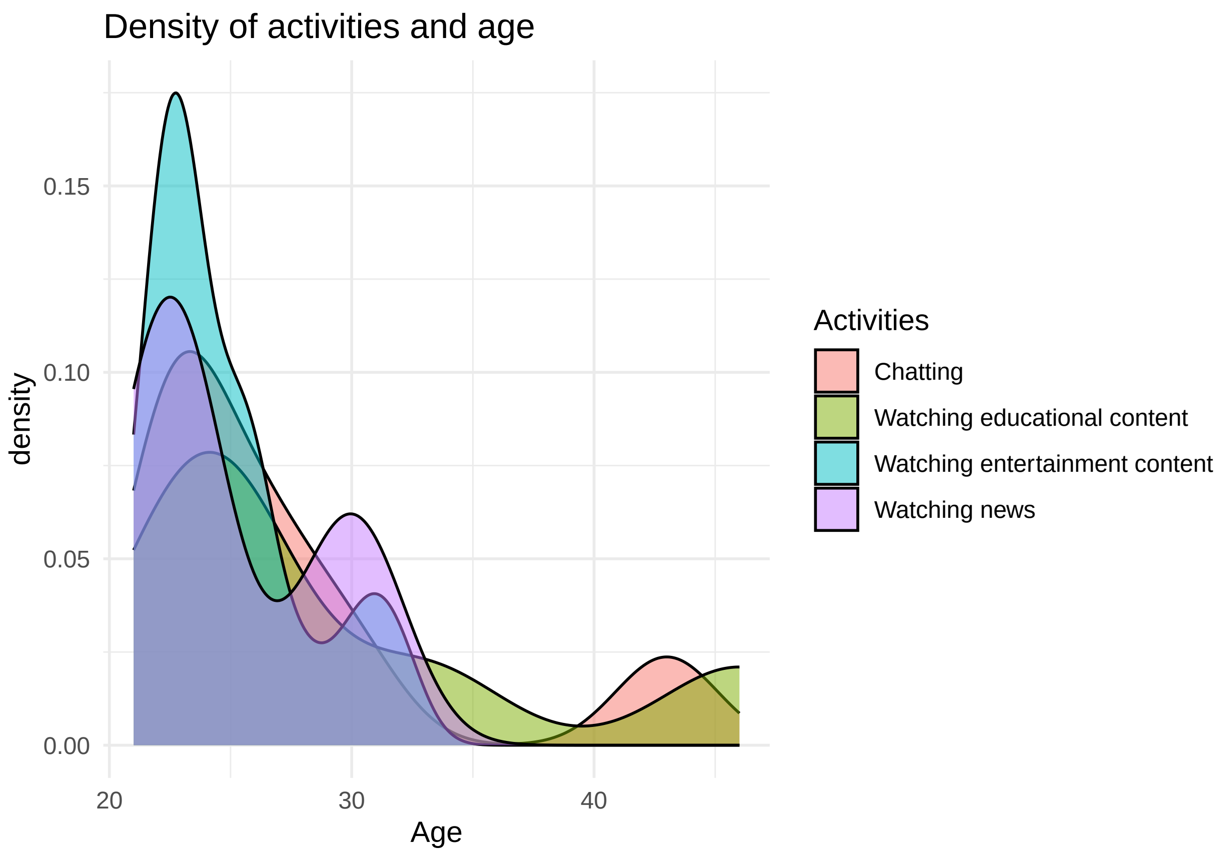


```
program_counts <- res %>%
  count(`Academic Program`)

ggplot(program_counts, aes(x = "", y = n, fill = `Academic Program`)) +
  geom_bar(stat = "identity") +
  labs(title = "Academic Program Distribution") +
  xlab("") +
  ylab("Count") +
  coord_polar("y", start = 0) +
  theme_minimal()
```



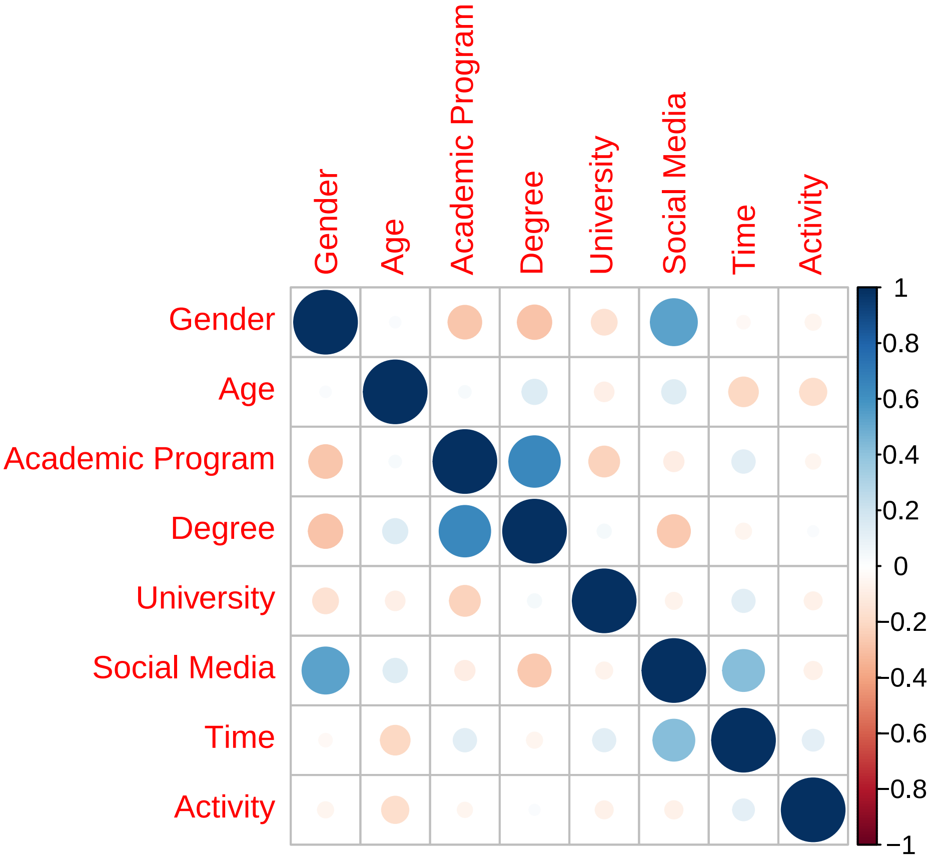
The pie chart shows the distribution of students across academic programs, with Data Science being the largest group. Some others are from Business Informatics and Statistics. NA are some Erasmus students.

```r
plot2 <- ggplot(res, aes(x = Age, fill = as.factor(Activity))) +
  geom_density(alpha = 0.5) +
  ggtitle("Density of activities and age") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(fill = "Activities") +
  theme_minimal()

plot2
```



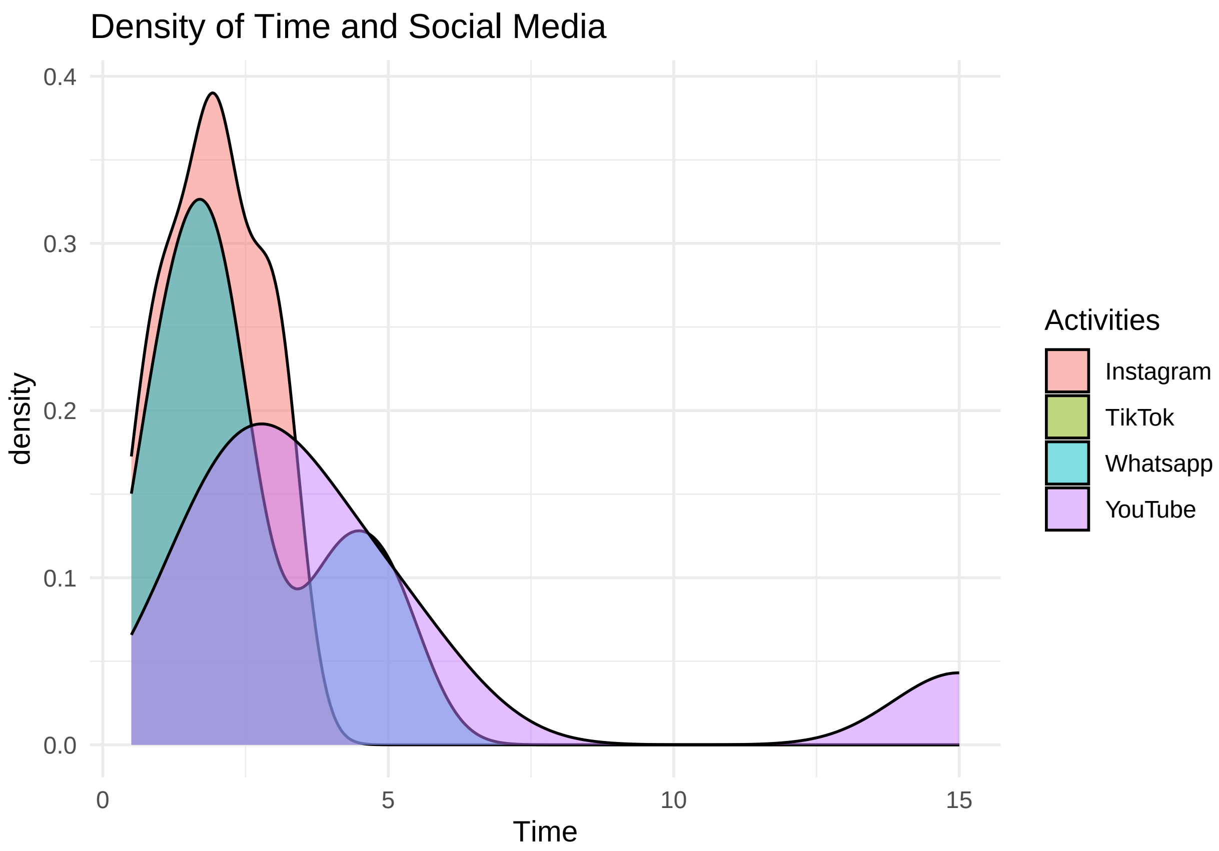
This density plot displays activities by age. Younger individuals tend to watch entertainment content, while older individuals prefer chatting and educational content.

```
corrplot(cor(res %>%
mutate_all(~as.numeric(factor(.))) %>%
   mutate(across(everything(), ~replace_na(., median(., na.rm = TRUE))))))
```



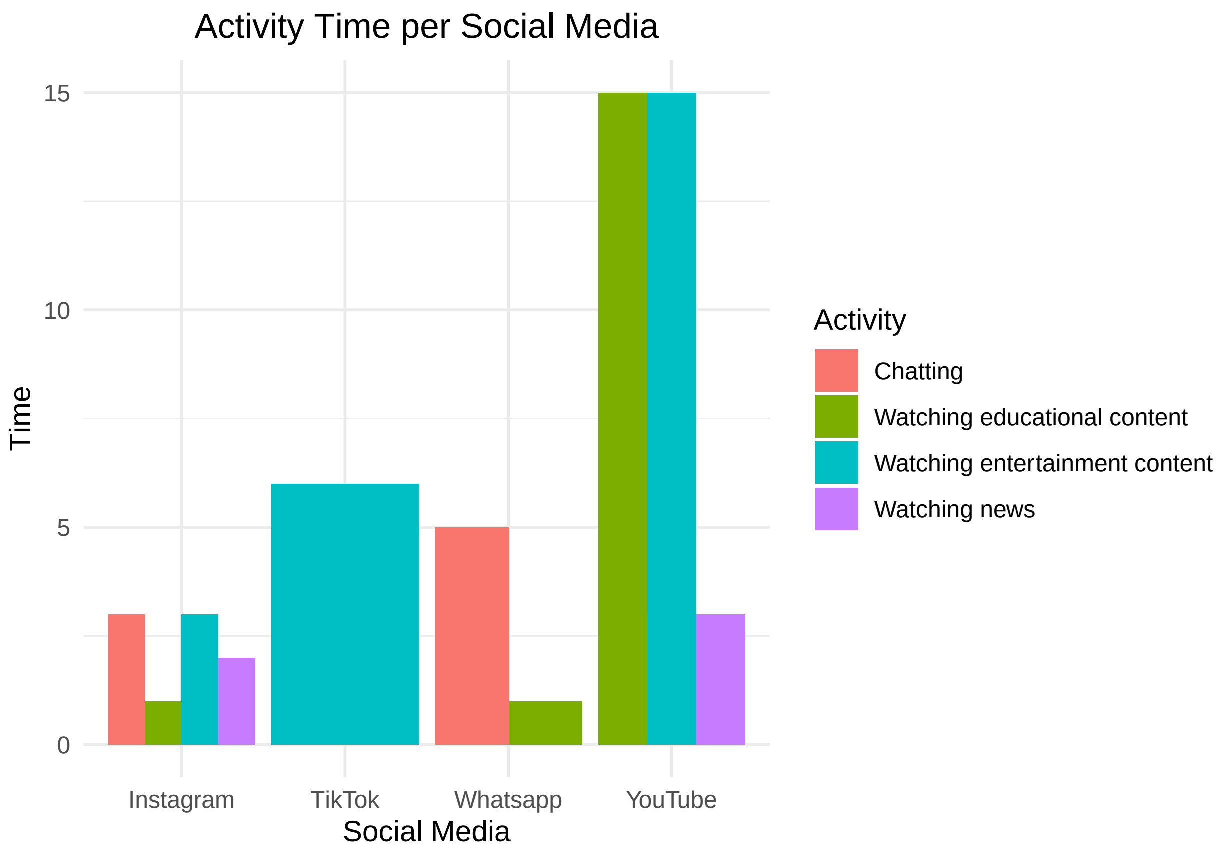
In terms of correlations, we observe a strong relationship between gender and social media type. We also see a correlation between time spent and social media type.

```
plot3 <- ggplot(res, aes(x = Time, fill = as.factor(`Social Media`))) +
  geom_density(alpha = 0.5) +
  ggtitle("Density of Time and Social Media") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(fill = "Activities") +
  theme_minimal()

plot3
```



From this plot we can conlude that most people spent up to 5 hours per day. But someone somehow spends up to 15 hours per day watching YouTube.

```r
plot4 <- ggplot(res, aes(x = `Social Media`, y = Time, fill = Activity)) +
  geom_bar(stat = "identity", position="dodge") +
  ggtitle("Activity Time per Social Media") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))

plot4
```



By this plot we can see how people allocate time on different social media platforms. Instagram here is the most universal platform, where people do all kinds of activities, but least time. TikTok as expected is only for entertainment and more time consumption, WhatsApp also as expected mostly for chatting and a bit for educational purposes and YouTube takes the most time mostly for educational and entertainment content and for watching news.

Descriptive Inference

```
summary(res)
```

```
##     Gender               Age        Academic Program      Degree
##  Length:38          Min.   :21.00   Length:38          Length:38
##  Class  :character   1st Qu.:23.00   Class  :character   Class  :character
##  Mode   :character   Median :24.00   Mode   :character   Mode   :character
##                      Mean   :25.97
##                      3rd Qu.:26.75
##                      Max.   :46.00
##   University         Social Media            Time          Activity
##  Length:38          Length:38          Min.   : 0.500   Length:38
##  Class  :character   Class  :character   1st Qu.: 2.000   Class  :character
##  Mode   :character   Mode   :character   Median : 2.000   Mode   :character
##                                          Mean   : 3.289
##                                          3rd Qu.: 3.750
##                                          Max.   :15.000
```

In terms of descriptive inference, we can say that our dataset is representative of a young, primarily Data Science student population. The majority of the individuals surveyed spend their time on social media for entertainment, followed by education and chatting. The time spent on these activities does not significantly differ based on the student's age or academic program. However, as always, these conclusions are made under the assumption that the data is representative and accurate, and further research with a larger and more diverse sample size might yield different results.

For our hypothesises let's make needed data wranglings.

For the first research question let's filter the dataset res for rows where Activity equals "Watching entertainment content". It then groups the filtered data by Social Media. Then, it calculates the mean value of the Time column for each group of Social Media. Each row in H1 represents a different social media platform, with the calculated average time that respondents spent watching entertainment content on that platform.

```
H1 <- res %>%
  filter(`Activity` == "Watching entertainment content") %>%
  group_by(`Social Media`) %>%
  summarize(`Average Time ` = mean(`Time`))
```

For the second research question let's filter the dataset res, but it keeps only the rows where Age is greater than 24. It then groups this filtered data by Activity, and calculates the mean value of the Time column for each group of Activity. Each row in H2 represents a different activity, with the calculated average time that respondents older than 24 spend on that activity.

```
H2 <- res %>%
  filter(Age > 24) %>%
  group_by(`Activity`) %>%
  summarize(`Average Time` = mean(`Time`))

print(H2)
```

```
## # A tibble: 4 x 2
##   Activity                       ‘Average Time‘
##   <chr>                                   <dbl>
## 1 Chatting                                 2
## 2 Watching educational content             5.25
## 3 Watching entertainment content           3.06
## 4 Watching news                            3
```

For the third research question we need to group the dataset res by Academic Program. Then, we calculated the mean value of the Time column for each group of Academic Program.

```
H3 <- res %>%
  group_by(`Academic Program`) %>%
  replace_na(list(`Academic Program` = "Others")) %>%
  summarize(`Average Time ` = mean(`Time`))

print(H3)
```

```
## # A tibble: 4 x 2
##   ‘Academic Program‘   ‘Average Time ‘
##   <chr>                          <dbl>
## 1 Business Informatics            2
## 2 Data Science                    3.33
## 3 Others                          3.5
## 4 Statistic                       4
```

Analytic Inference

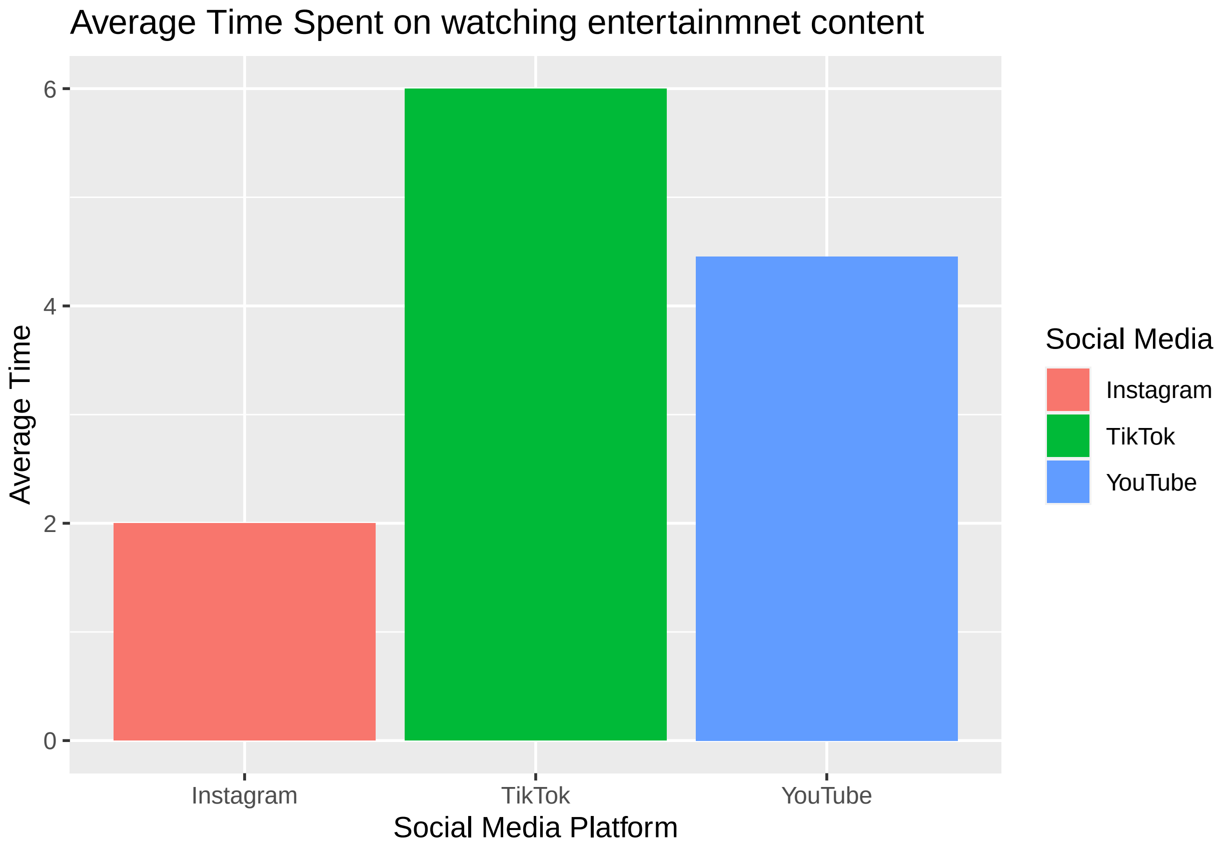
Based on the result data that we received, we put several hypothesis that we tried to prove with visualizations and later with test statistics. Because of that we presented the hypothesis the following way: null hypothesis represents opposite of the hypothesis and alternative is the target statement that we want to get.

Research Question 1

H0: YouTube is not the most time-consuming social media where students watch entertainment content the most. H1: YouTube is the most time-consuming social media where students watch entertainment content the most.

In order to see the result based on the survey we build bar plot that shows Average time spent on Social Media to watch Entertainment content. Initial result was that TikTok is the most watched social media. The result disagrees with the alternative hypothesis.

```r
#Hypothesis 1
ggplot(H1, aes(x = `Social Media`, y = `Average Time `,
               fill = `Social Media`)) +
  geom_bar(stat = "identity") +
  labs(x = "Social Media Platform", y = "Average Time") +
  ggtitle("Average Time Spent on watching entertainmnet content")
```



```r
#Hypothesis 1
H1_data <- res %>%
  filter(`Activity` == "Watching entertainment content") %>%
  group_by(`Social Media`)
  # summarize(`Average Time ` = mean(`Time`))
H1_data
```

```
## # A tibble: 21 x 8
## # Groups:   Social Media [3]
##    Gender   Age ‘Academic Program‘   Degree University ‘Social Media‘  Time
##    <chr>  <dbl> <chr>                <chr>  <chr>      <chr>          <dbl>
##  1 female    23 Data Science         MSc    TU Wien    YouTube          3
##  2 male      21 Data Science         MSc    TU Wien    YouTube          6
##  3 female    23 Data Science         MSc    TU Wien    Instagram        2
##  4 female    24 Data Science         MSc    TU Wien    Instagram        1.5
##  5 male      22 Data Science         MSc    TU Wien    YouTube          2
##  6 male      26 Data Science         MSc    TU Wien    YouTube          5
##  7 male      23 Business Informatics BSc    TU Wien    YouTube          2
##  8 female    26 Data Science         MSc    TU Wien    Instagram        3
##  9 female    22 Data Science         MSc    TU WIen    Instagram        3
## 10 male      28 Data Science         Msc    TU Wien    YouTube          3
## # i 11 more rows
## # i 1 more variable: Activity <chr>
```

Further, we decided that we need to conduct significance testing with a two-sample t-test in order to confirm results. We received p-value equal to 0.1877 which is greater than 0.05 so we reject the initial hypothesis (in our case alternative hypothesis) and confirm the result we got before with graph that YouTube is not the most time-consuming social media where students watch entertainment content the most.

```r
youtube_time <- res$Time[res$`Social Media` == "YouTube"
                         & res$`Activity` == "Watching entertainment content"]
other_time <- res$Time[res$`Social Media` != "YouTube"
                       & res$`Activity` == "Watching entertainment content"]

t_test_result <- t.test(youtube_time, other_time)

t_test_result
```

```
##
##  Welch Two Sample t-test
##
## data:  youtube_time and other_time
## t = 1.6611, df = 13.422, p-value = 0.1199
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.6089418  4.7180328
## sample estimates:
## mean of x mean of y
##  4.454545  2.400000
```

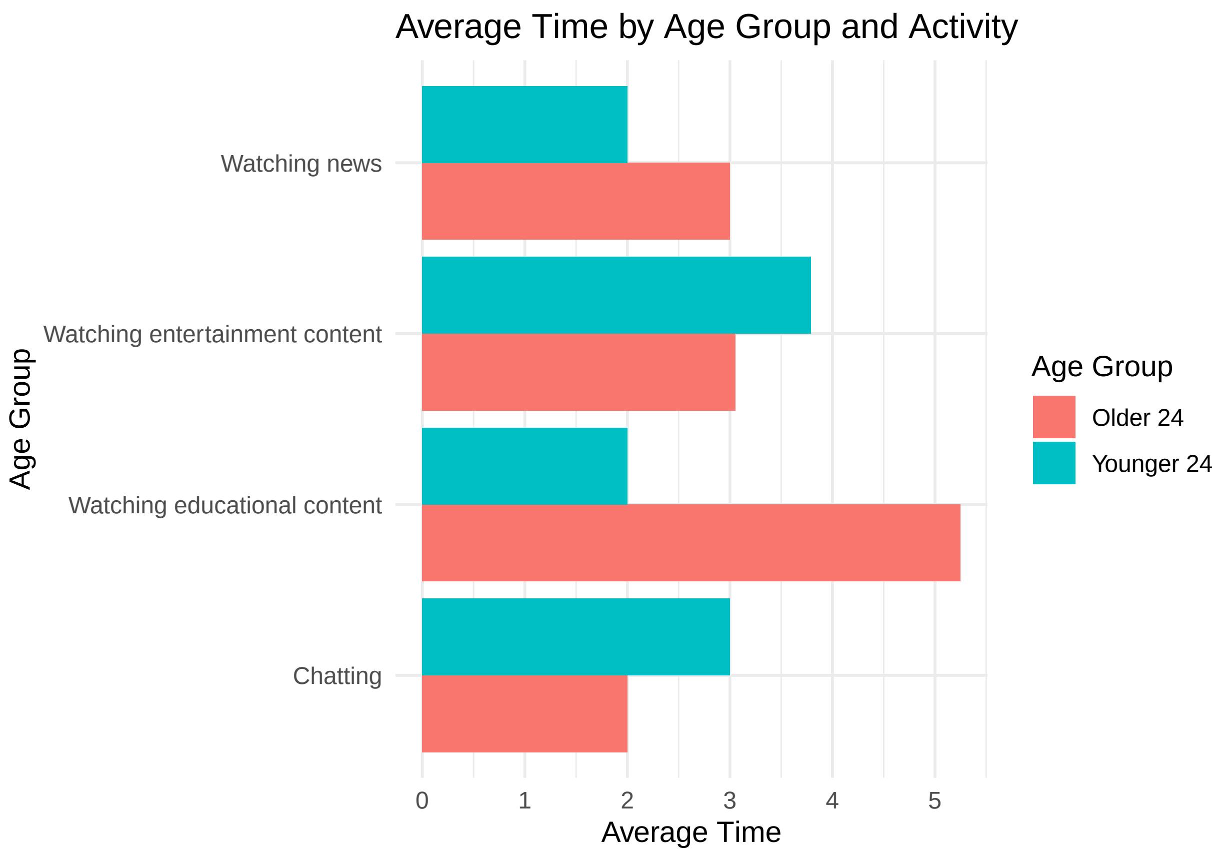
Research Question 2

H0: Students of 24 and under 24 are more likely to watch the news/chat. H1: Students over 24 are more likely to watch the news/chat.

```
#Hypothesis 2
H2_A <- res %>%
  group_by(`Age Group` = ifelse(Age <= 24, "Younger 24", "Older 24"),
           `Activity`) %>%
  summarize(`Average Time` = mean(`Time`))

print(H2_A)
```

```
## # A tibble: 8 x 3
## # Groups:   Age Group [2]
##   ‘Age Group‘ Activity                        ‘Average Time‘
##   <chr>       <chr>                                    <dbl>
## 1 Older 24    Chatting                                  2
## 2 Older 24    Watching educational content              5.25
## 3 Older 24    Watching entertainment content            3.06
## 4 Older 24    Watching news                             3
## 5 Younger 24  Chatting                                  3
## 6 Younger 24  Watching educational content              2
## 7 Younger 24  Watching entertainment content            3.79
## 8 Younger 24  Watching news                             2
```

```
ggplot(data = H2_A, aes(fill = `Age Group`, x = `Average Time`,
                        y = `Activity`)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(y = "Age Group", x = "Average Time",
       title = "Average Time by Age Group and Activity") +
  theme_minimal()
```



To examine these hypothesis we decided to conduct a chi-square test. We think it is more applicable, because the data to analyze is categorical. T-test can work only with numeric data. The resulting p-value is 0.2745. Based on it, we fail to reject the null hypothesis.

```
H2_A <- H2_A %>% unite(`Age_Activity`, `Age Group`, Activity, sep = " ")

chisq_results <- chisq.test(H2_A$`Age_Activity`, H2_A$`Average Time`)

chisq_results
```

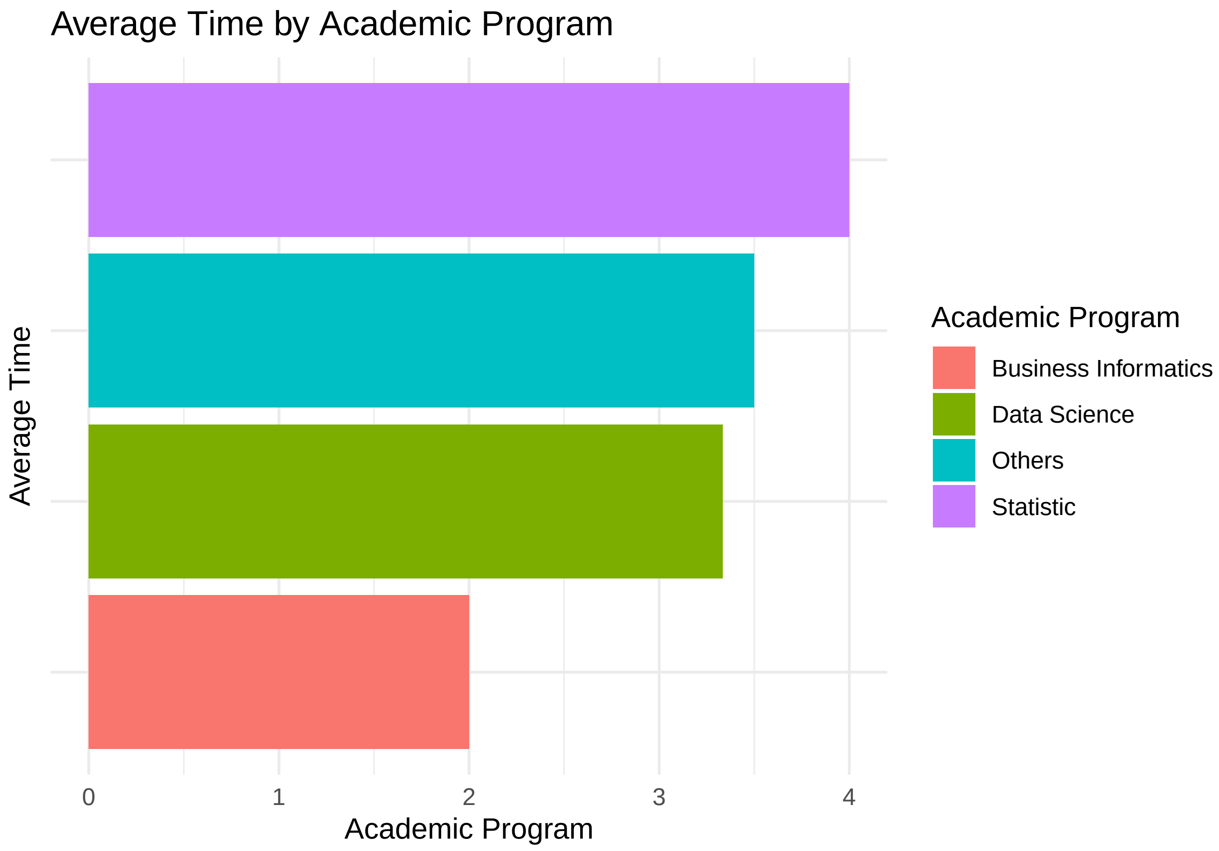
```
##
##  Pearson's Chi-squared test
##
## data:  H2_A$Age_Activity and H2_A$‘Average Time‘
## X-squared = 32, df = 28, p-value = 0.2745
```

In conclusion, the findings suggest that age does not play a significant role in determining the likelihood of students watching the news/chat.

Research Question 3

Two hypotheses were selected for this study: Students from programs other than Data Science spend more time on social media and DS students spend more time on social media.

```r
#Hypothesis 3

ggplot(data = H3, aes(y = `Academic Program`, x = `Average Time `,
                      fill=`Academic Program`)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Academic Program", y = "Average Time",
       title = "Average Time by Academic Program") +
  theme_minimal() +
  theme(axis.text.y = element_blank())
```



```r
cs_ds_time <- res$Time[res$`Academic Program` %in% c("Data Science")]
other_time <- res$Time[!res$`Academic Program` %in% c("Data Science")]

t_test_result <- t.test(cs_ds_time, other_time)


p_value <- t_test_result$p.value

print(p_value)
```

```
## [1] 0.7040155
```

```r
t_test_result
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  cs_ds_time and other_time
## t = 0.38879, df = 12.419, p-value = 0.704
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.527733  2.194400
## sample estimates:
## mean of x mean of y
##  3.333333  3.000000
```

To test these hypotheses, a t-test analysis was conducted, resulting in a p-value of 0.704.

The main findings of the study are: - The null hypothesis could not be rejected. - There was no significant difference in social media usage between DS students and students from other programs.

However, it is important to consider the study's limitations, such as a small sample size and reliance on self-reported data.

In conclusion, the findings suggest that the choice of program does not have a noticeable impact on the amount of time students spend on social media.

Conclusions

YouTube is not the most time-consuming social media platform where students watch entertainment content the most. This conclusion was made after conducting a t-test and visualizing the average time spent on social media to watch entertainment content.

Age does not play a significant role in determining the likelihood of students watching the news/chat. This conclusion was drawn after conducting a chi-square test and visualizing the average time spent by students of different age groups on various activities.

The academic program a student is enrolled in does not have a significant impact on the amount of time they spend on social media. This conclusion was made after conducting a t-test and visualizing the average time spent on social media by students from different academic programs.