Web scraping and basic data analysis

Greinald Pappa (12046752), Arjol Pançi(12347494)

Apr. 13, 2025 (Assignment_1.Rmd)

Contents

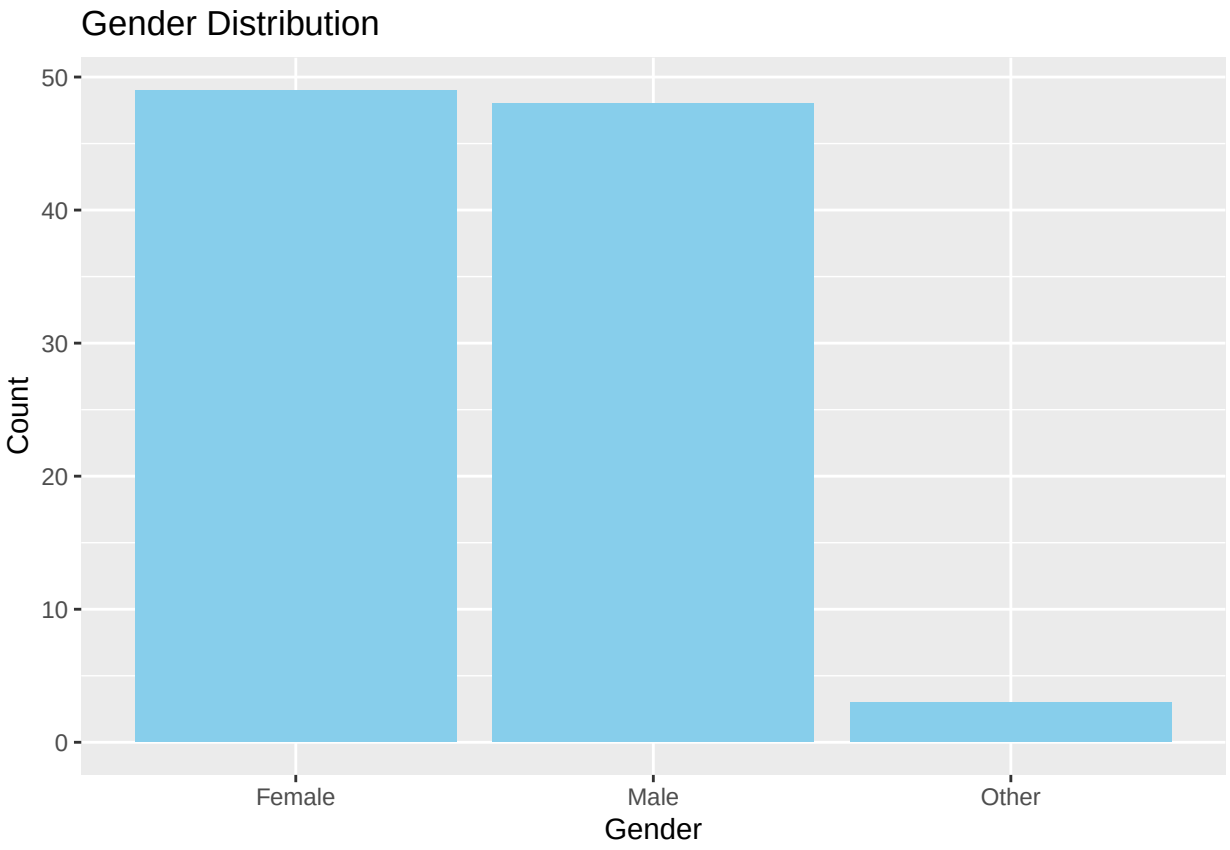
1 Load data

```
survey_data <- read.csv('survey_data.csv')
```

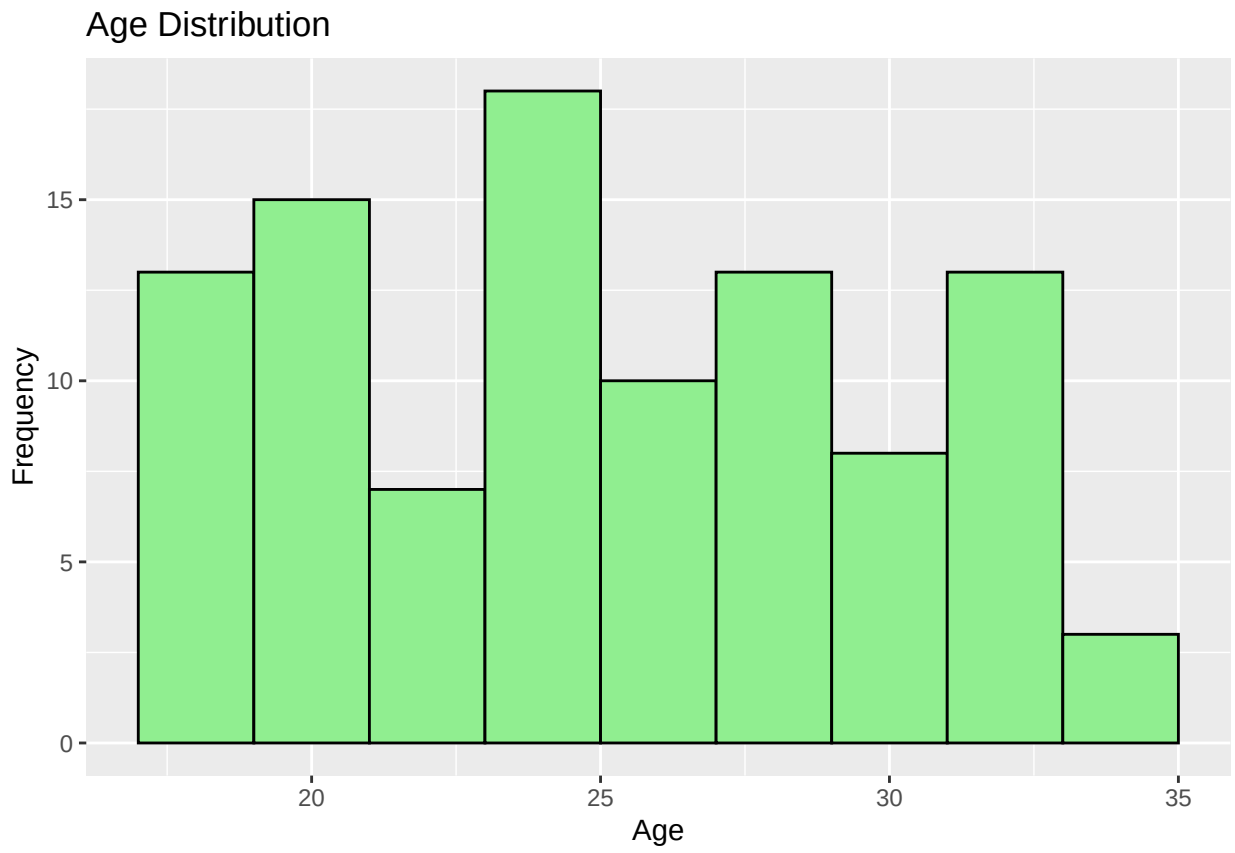
1.1 1. Exploratory Data Analysis (EDA)

1.1.1 General Demographics

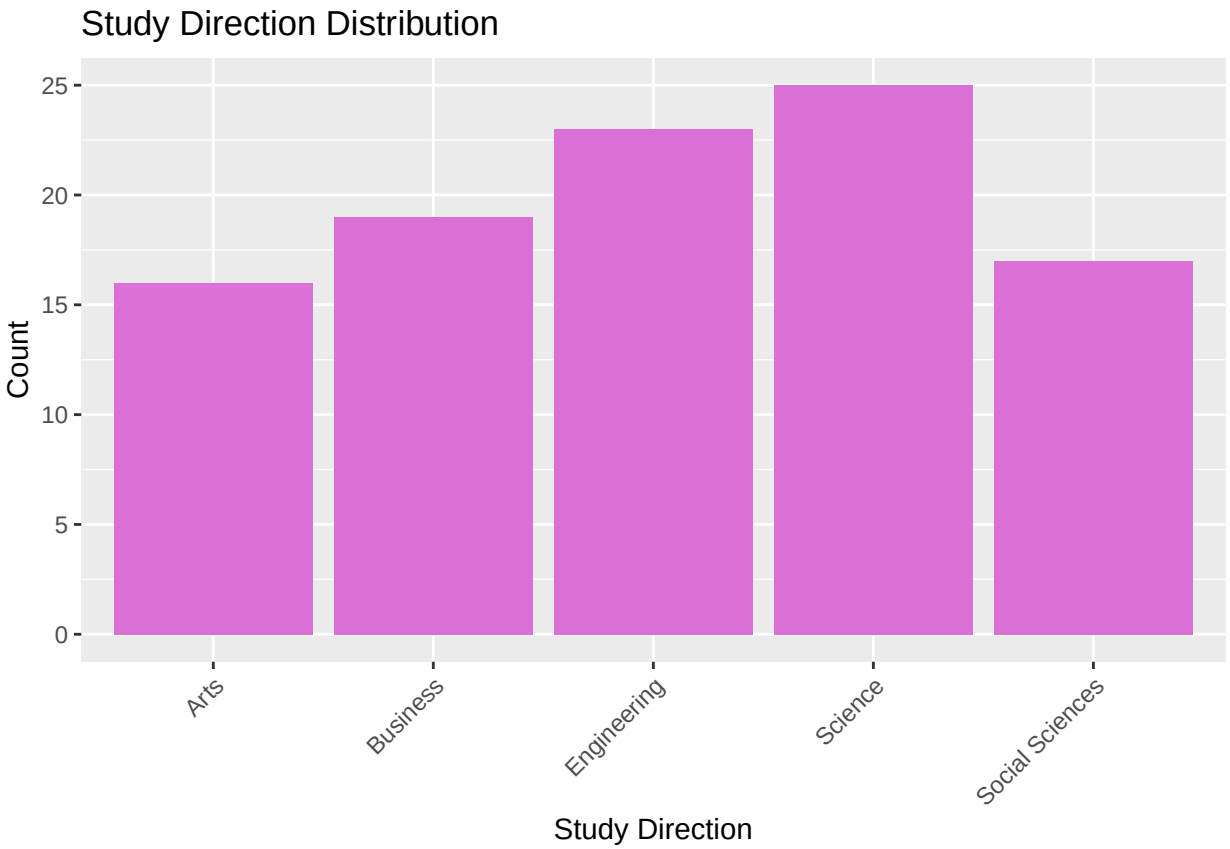
```
ggplot(survey_data, aes(x = Gender)) +
  geom_bar(fill = 'skyblue') +
  labs(title = "Gender Distribution", x = "Gender", y = "Count")
```



```r
# Age distribution
ggplot(survey_data, aes(x = Age)) +
  geom_histogram(binwidth = 2, fill = 'lightgreen', color = 'black') +
  labs(title = "Age Distribution", x = "Age", y = "Frequency")
```



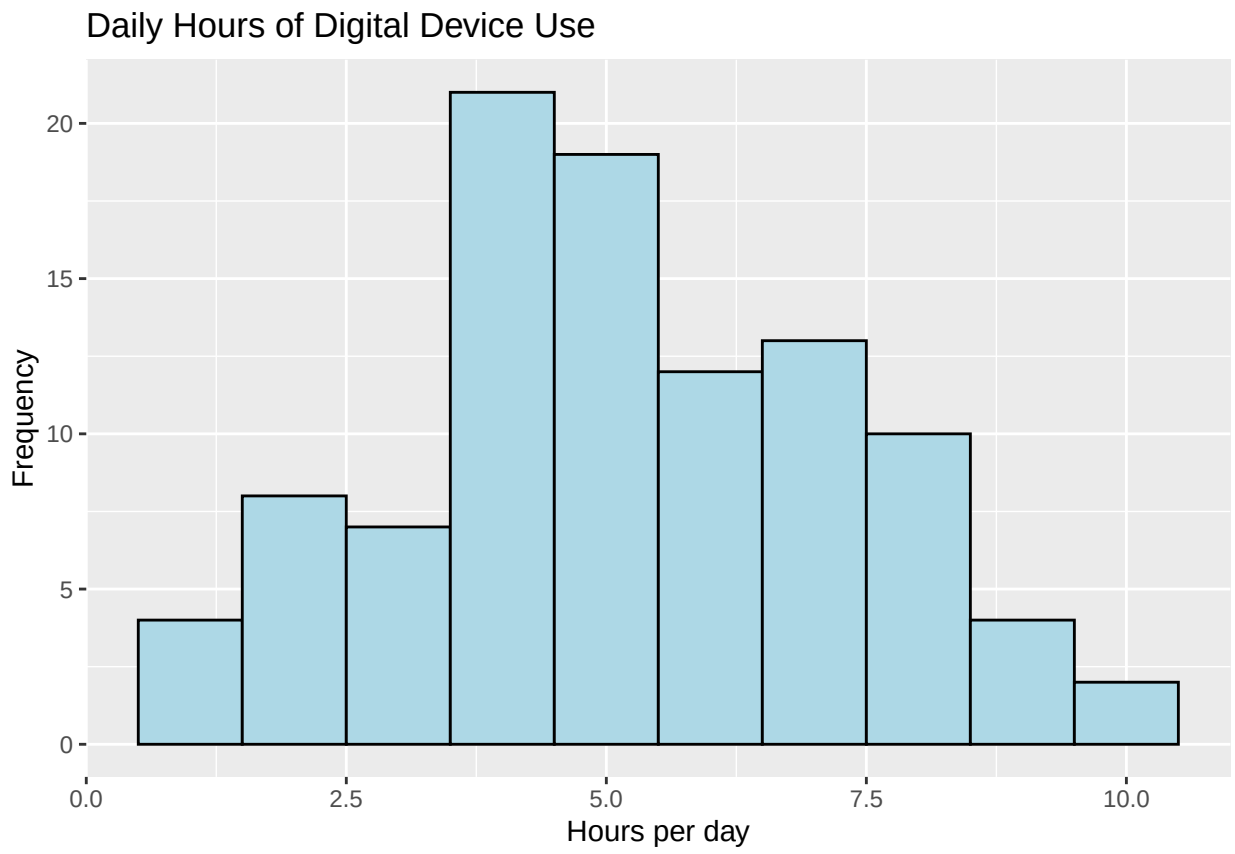

```r
# Study direction distribution
ggplot(survey_data, aes(x = Study_Direction)) +
  geom_bar(fill = 'orchid') +
  labs(title = "Study Direction Distribution", x = "Study Direction", y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



1.1.2 Research Question 1

"How many hours per day do you typically use digital devices?"

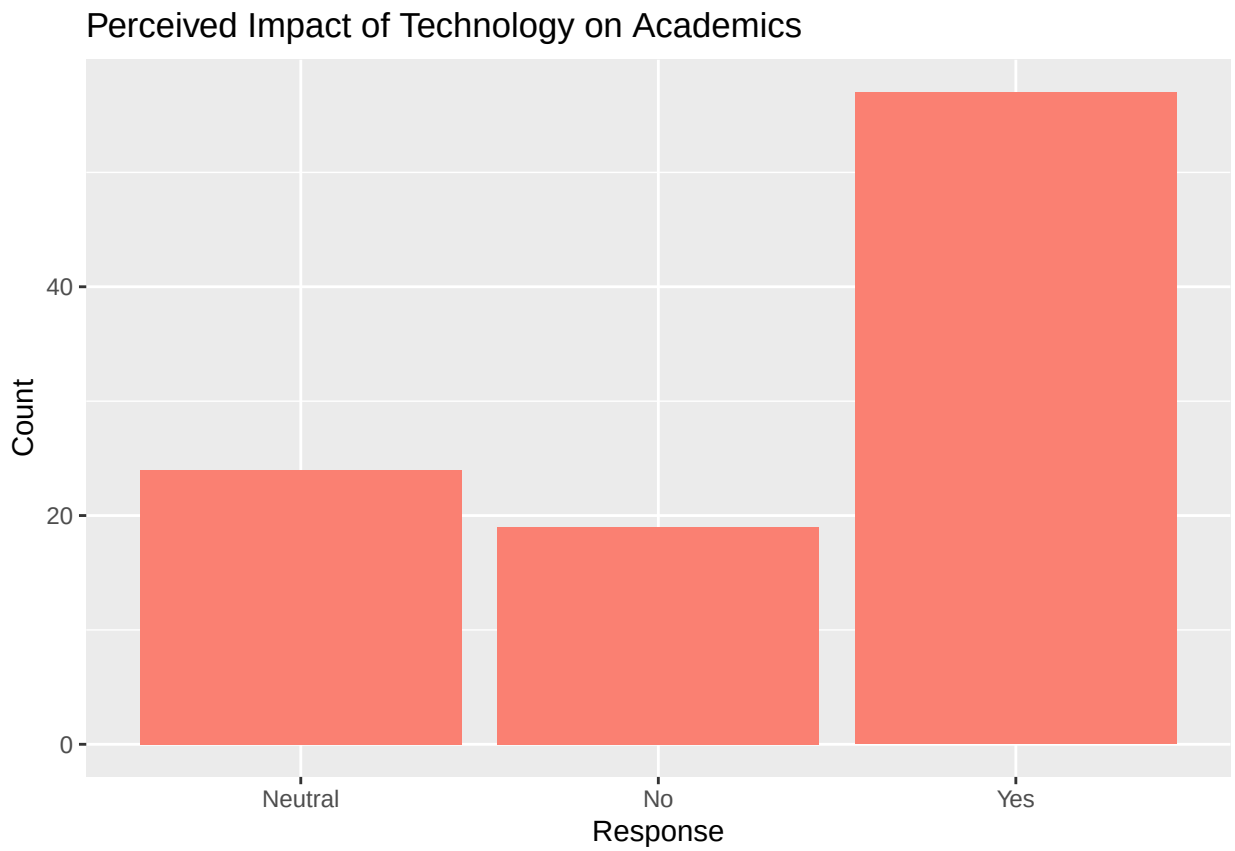
```
ggplot(survey_data, aes(x = Hours_Digital_Use)) +
  geom_histogram(binwidth = 1, fill = 'lightblue', color = 'black') +
  labs(title = "Daily Hours of Digital Device Use", x = "Hours per day", y = "Frequency")
```



1.1.3 Research Question 2

"Has technology improved your academic performance?"

```
ggplot(survey_data, aes(x = Tech_Improved_Academics)) +
  geom_bar(fill = 'salmon') +
  labs(title = "Perceived Impact of Technology on Academics", x = "Response", y = "Count")
```



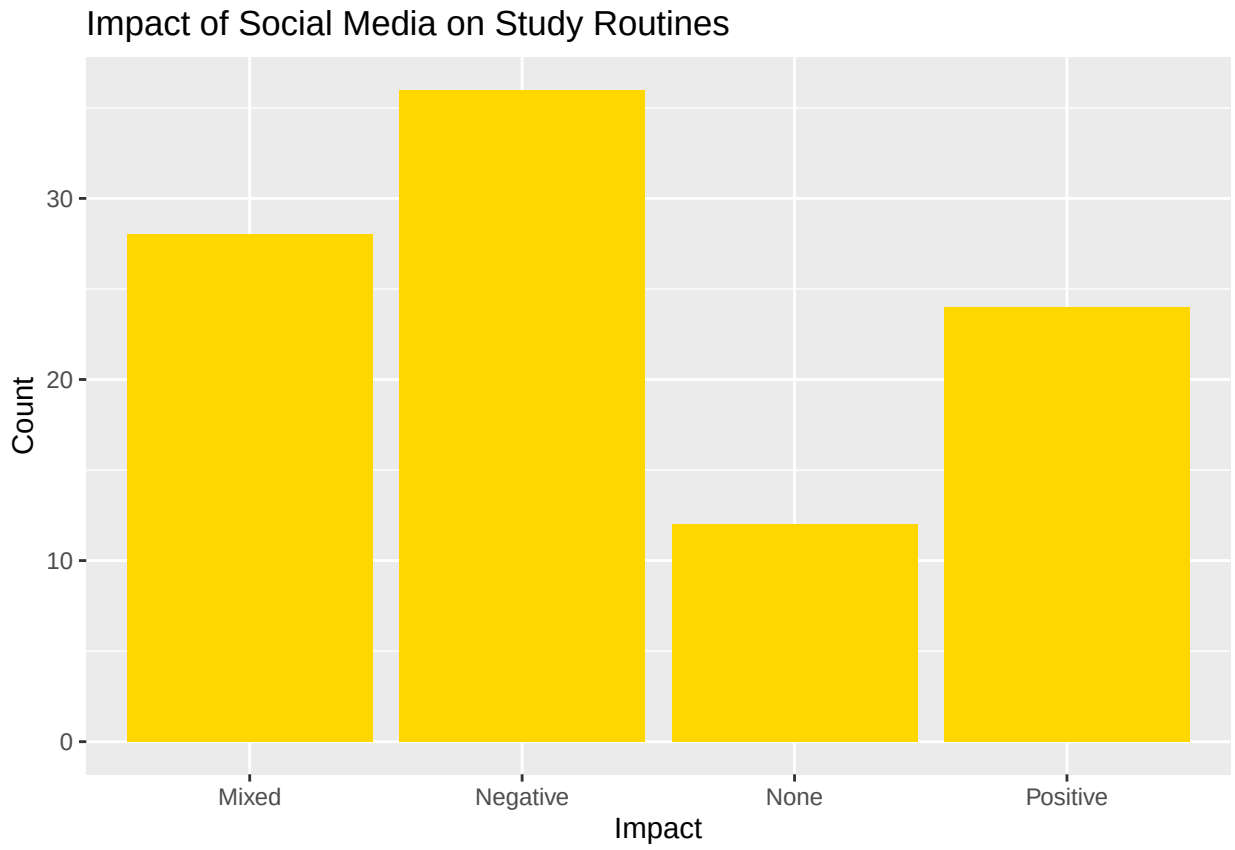
1.1.4 Research Question 3

"How has social media impacted your study routines?"

```
ggplot(survey_data, aes(x = Social_Media_Impact)) +
  geom_bar(fill = 'gold') +
  labs(title = "Impact of Social Media on Study Routines", x = "Impact", y = "Count")
```

Table 1: Summary Statistics of Digital Device Use (hours/day)

Mean_Hours	Median_Hours	SD_Hours	Min_Hours	Max_Hours
5.227	5.1	2.046226	1	9.8



1.2 2. Descriptive Inference

1.2.1 Quantitative Variables Summary

```
summary_stats <- survey_data %>%
  summarise(
    Mean_Hours = mean(Hours_Digital_Use),
    Median_Hours = median(Hours_Digital_Use),
    SD_Hours = sd(Hours_Digital_Use),
    Min_Hours = min(Hours_Digital_Use),
    Max_Hours = max(Hours_Digital_Use)
  )

knitr::kable(summary_stats, caption = "Summary Statistics of Digital Device Use (hours/day)")
```

Table 2: Gender Frequency Table

Var1	Freq
Female	49
Male	48
Other	3

Table 3: Technology Impact on Academics Frequency Table

Var1	Freq
Neutral	24
No	19
Yes	57

1.2.2 Categorical Variables Frequency Tables

```
table_gender <- table(survey_data$Gender)
table_academics <- table(survey_data$Tech_Improved_Academics)
table_social_media <- table(survey_data$Social_Media_Impact)

knitr::kable(table_gender, caption = "Gender Frequency Table")
```

```
knitr::kable(table_academics, caption = "Technology Impact on Academics Frequency Table")
```

```
knitr::kable(table_social_media, caption = "Social Media Impact Frequency Table")
```

1.3 3. Analytic Inference

1.3.1 Hypothesis Test 1

Hypothesis: *Is there a significant difference in daily digital device use between genders?*

```
test1 <- t.test(Hours_Digital_Use ~ Gender, data = survey_data %>% filter(Gender %in% c("Male", "Female
print(test1)
```

Table 4: Social Media Impact Frequency Table

Var1	Freq
Mixed	28
Negative	36
None	12
Positive	24

```
##
##  Welch Two Sample t-test
##
## data:  Hours_Digital_Use by Gender
## t = 1.586, df = 94.092, p-value = 0.1161
## alternative hypothesis: true difference in means between group Female and group Male is not equal to
## 95 percent confidence interval:
##  -0.1664529  1.4881366
## sample estimates:
## mean in group Female   mean in group Male
##             5.579592             4.918750
```

1.3.2 Hypothesis Test 2

Hypothesis: *Is there an association between study direction and perceived academic improvement due to technology?*

```
chisq_test <- chisq.test(table(survey_data$Study_Direction, survey_data$Tech_Improved_Academics))
print(chisq_test)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(survey_data$Study_Direction, survey_data$Tech_Improved_Academics)
## X-squared = 11.118, df = 8, p-value = 0.1951
```

1.3.3 Correlation Analysis

Hypothesis: *Correlation between age and hours of digital device usage.*

```
cor_test <- cor.test(survey_data$Age, survey_data$Hours_Digital_Use)
print(cor_test)
```

```
##
##  Pearson's product-moment correlation
##
## data:  survey_data$Age and survey_data$Hours_Digital_Use
## t = 2.6915, df = 98, p-value = 0.008365
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.06952419 0.43629449
## sample estimates:
##       cor
## 0.2623596
```