Report

Ouassim Kiassa | Jonas Schindler

2024-06-30

Contents

Introduction

Our research topic is smoking behavior of students. We asked three questions related to the age the person started smoking, how many cigarettes they smoke per day and what they are currently studying at TU Wien.

```
# Load and preprocess the data
df <- read.csv("group13.csv")
colnames(df) <- c("Gender", "Age", "Program", "Smoke", "Age_Started_Smoking", "Amount")

# Define the function to parse the program
parse_program <- function(program) {
  # Normalize case for consistent processing
  program_lower <- tolower(program)

  # Find degree and normalize
  if (str_detect(program_lower, "\\b(bsc|bachelor|bachelorstudium)\\b")) {
    degree <- "Bachelor"
  } else if (str_detect(program_lower, "\\b(msc|master)\\b") | str_detect(program_lower, "data science"
    degree <- "Master"
  } else {
    print(paste("Unknown Degree:", program))
    degree <- "Unknown"
  }

  subject <- program_lower %>%
    str_replace_all("/msc|/bsc|/bachelor|/master|tu wien|msc|bsc|bachelor|master|\\s+at\\s+.*", "") %>%
    str_trim() %>%
    str_replace_all("^[/\\s]+|[/\\s]+$", "") %>%
    str_to_title()


  if (str_detect(subject, "data science") | str_detect(subject, "Data Science") | str_detect(subject, "
    subject <- "Data Science"
  } else if (str_detect(subject, "Mathematics")) {
    subject <- "Mathematics"
  } else if (str_detect(subject, "Medical Informatiocs")) {
      subject <- "Medical Informatics"
  } else if (str_detect(subject, "Studium Software & Information Engineering")) {
    subject <- "Software & Information Engineering"
  } else if (str_detect(subject, "Business Informatics")) {
    subject <- "Business Informatics"
  }
  else {
    print(paste("Unknown Subject:", program))
    subject <- "Unknown"
  }

  return(c(Degree = degree, Subject = subject))
}
```

```
# Apply the function and assign new columns
parsed_programs <- t(apply(df["Program"], 1, parse_program))
```

```
## [1] "Unknown Degree: TU Wien"
## [1] "Unknown Subject: TU Wien"
## [1] "Unknown Degree: TU Wien"
## [1] "Unknown Subject: TU Wien"
```

```
df <- cbind(df, as.data.frame(parsed_programs, stringsAsFactors = FALSE))

# Define the function to map gender
map_gender <- function(gender) {
  case_when(
    gender == "female" ~ "Female",
    gender == "male" ~ "Male",
    gender == "no answer" ~ "Unknown",
    gender == "diverse" ~ "Diverse",
    TRUE ~ "Unknown"
  )
}

df$Gender <- sapply(df$Gender, map_gender)
df$Smoke <- sapply(df$Smoke, function(x) ifelse(tolower(x) == "yes", TRUE, FALSE))

df = subset(df, select = -c(Program) )

# Show the first few rows of the DataFrame to confirm changes
df %>%gt()
```

Gender	Age	Smoke	Age_Started_Smoking	Amount	Degree	Subject
Female	22	FALSE	0	0.00	Master	Data Science
Female	22	FALSE	0	0.00	Master	Data Science
Female	48	FALSE	0	0.00	Master	Data Science
Male	25	FALSE	0	0.00	Master	Data Science
Male	24	FALSE	0	0.00	Master	Data Science
Male	22	FALSE	0	0.00	Bachelor	Software & Information Engineering
Male	26	FALSE	0	0.00	Master	Data Science
Female	26	FALSE	0	0.00	Master	Data Science
Male	23	FALSE	0	0.00	Master	Data Science
Female	27	FALSE	0	0.00	Master	Data Science
Male	45	FALSE	0	0.00	Master	Business Informatics
Male	23	FALSE	0	0.00	Master	Data Science
Male	22	FALSE	0	0.00	Master	Business Informatics
Male	24	FALSE	0	0.00	Unknown	Unknown
Male	41	FALSE	13	0.00	Master	Data Science
Male	57	FALSE	20	0.00	Master	Data Science
Male	27	FALSE	0	0.00	Master	Data Science
Male	29	FALSE	0	0.00	Master	Data Science
Male	27	TRUE	18	5.00	Master	Data Science
Female	32	FALSE	0	0.00	Unknown	Unknown
Female	23	FALSE	0	0.00	Master	Data Science
Male	24	FALSE	0	0.00	Master	Data Science
Male	24	FALSE	0	0.00	Master	Data Science
Male	23	FALSE	0	0.00	Master	Data Science
Male	28	FALSE	0	0.00	Master	Data Science
Female	30	FALSE	0	0.00	Master	Data Science
Female	28	FALSE	0	0.00	Master	Mathematics
Female	24	FALSE	0	0.00	Master	Data Science
Male	27	TRUE	24	2.00	Master	Data Science
Female	29	FALSE	0	0.00	Master	Data Science
Female	24	FALSE	0	0.00	Master	Data Science
Male	25	FALSE	0	0.00	Master	Data Science
Male	28	TRUE	21	15.00	Master	Data Science
Male	24	TRUE	17	8.00	Master	Data Science
Female	28	FALSE	0	0.00	Master	Data Science
Male	23	FALSE	0	0.00	Master	Data Science
Male	27	FALSE	0	0.00	Master	Data Science
Male	23	FALSE	0	0.00	Master	Data Science
Male	24	FALSE	0	0.00	Master	Data Science
Female	22	TRUE	16	6.00	Master	Data Science
Male	23	TRUE	21	5.00	Master	Data Science
Male	25	FALSE	0	0.00	Master	Data Science
Male	26	FALSE	0	0.00	Master	Data Science
Male	25	FALSE	0	0.00	Master	Data Science
Male	28	FALSE	0	0.00	Master	Data Science
Male	25	TRUE	15	5.00	Master	Data Science
Female	47	FALSE	0	0.00	Master	Data Science
Unknown	32	FALSE	0	0.00	Master	Data Science
Female	36	FALSE	0	0.00	Master	Data Science
Male	24	FALSE	0	0.00	Master	Data Science
Female	24	FALSE	0	0.00	Master	Data Science
Male	23	FALSE	0	0.00	Master	Data Science

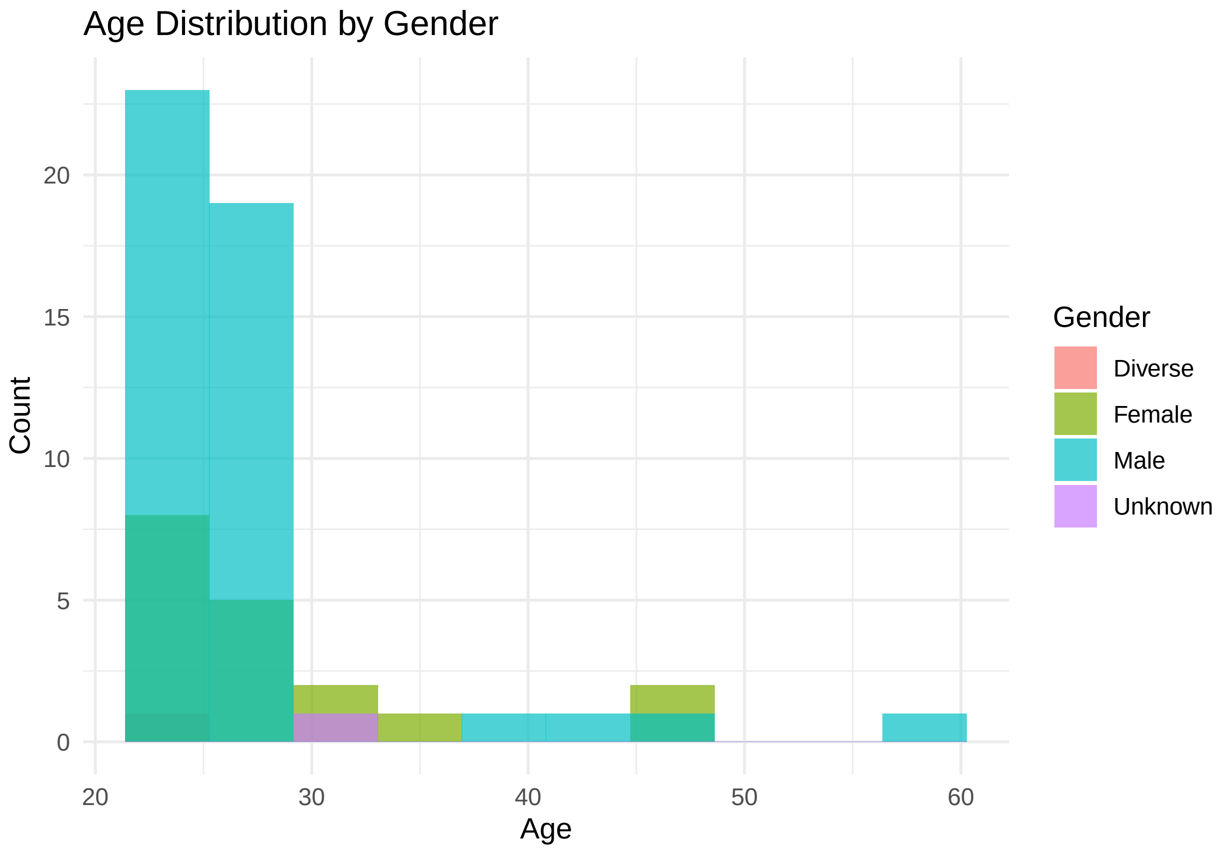
Male	27	TRUE	22	2.00	Master	Data Science
Female	22	FALSE	0	0.00	Master	Data Science
Male	25	FALSE	0	0.00	Master	Data Science
Male	39	FALSE	0	0.00	Bachelor	Medical Informatics
Male	27	FALSE	26	7.00	Master	Data Science
Male	29	FALSE	0	0.00	Master	Data Science
Male	28	TRUE	22	13.00	Master	Data Science
Male	26	FALSE	0	0.00	Bachelor	Mathematics
Diverse	23	FALSE	0	0.15	Master	Data Science
Male	26	FALSE	0	0.00	Master	Data Science
Male	26	FALSE	0	0.00	Master	Data Science
Male	25	TRUE	15	8.00	Master	Data Science
Male	27	FALSE	0	0.00	Master	Data Science
Male	29	FALSE	0	0.00	Master	Data Science

After cleaning the data we have aggregated the subjects and the study programm for every student, besides two which only entered TU Wien. Now we have a dataset with the following columns:

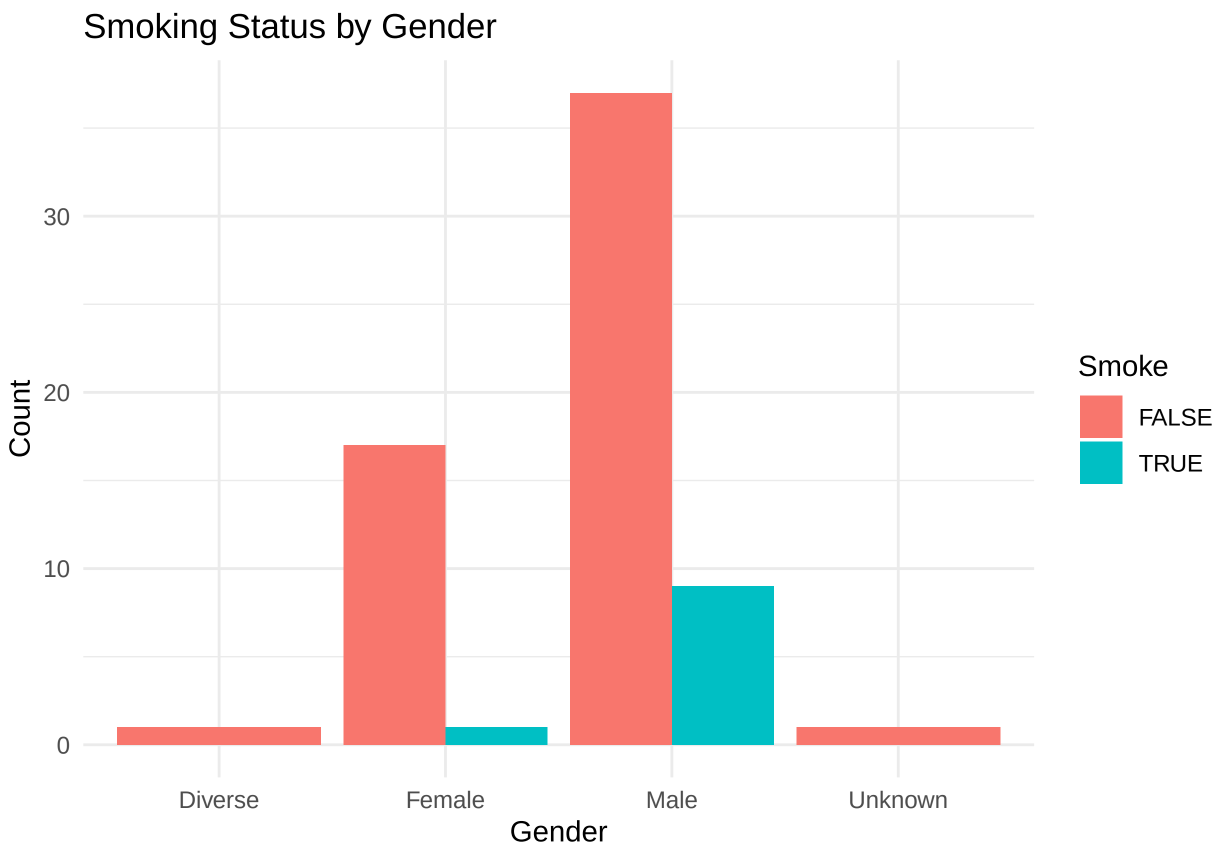
- Gender: The gender of the student, which can be "Male", "Female", "Unknown", or "Diverse".
- Age: The age of the student in years. - Smoke: A boolean value indicating whether the student smokes (TRUE) or not (FALSE).
- Age_Started_Smoking: The age at which the student started smoking, with 0 indicating non-smokers.
- Amount: The amount the student smokes per day.
- Degree: The highest academic degree obtained by the student, such as "Bachelor", "Master", or "Unknown".
- Subject: The subject area of the student's degree, including areas like "Data Science", "Software & Information Engineering", "Business Informatics", "Mathematics", "Medical Informatics", or "Unknown".

```r
p1 <- ggplot(df, aes(x = Age, fill = Gender)) +
  geom_histogram(bins = 10, alpha = 0.7, position = "identity") +
  theme_minimal() +
  labs(title = "Age Distribution by Gender", x = "Age", y = "Count")

# Plot 2: Smoking status by gender
p2 <- ggplot(df, aes(x = Gender, fill = Smoke)) +
  geom_bar(position = "dodge") +
  theme_minimal() +
  labs(title = "Smoking Status by Gender", x = "Gender", y = "Count")

# Plot 3: Age distribution of smokers vs non-smokers
p3 <- ggplot(df, aes(x = Age, fill = Smoke)) +
  geom_histogram(bins = 10, alpha = 0.7, position = "identity") +
  theme_minimal() +
  labs(title = "Age Distribution of Smokers vs Non-Smokers", x = "Age", y = "Count")

# Plot 4: Breakdown by Degree
p4 <- ggplot(df, aes(x = Degree, fill = Gender)) +
  geom_bar(position = "dodge") +
  theme_minimal() +
  labs(title = "Degree Breakdown by Gender", x = "Degree", y = "Count")

# Plot 5: Breakdown by Subject
```

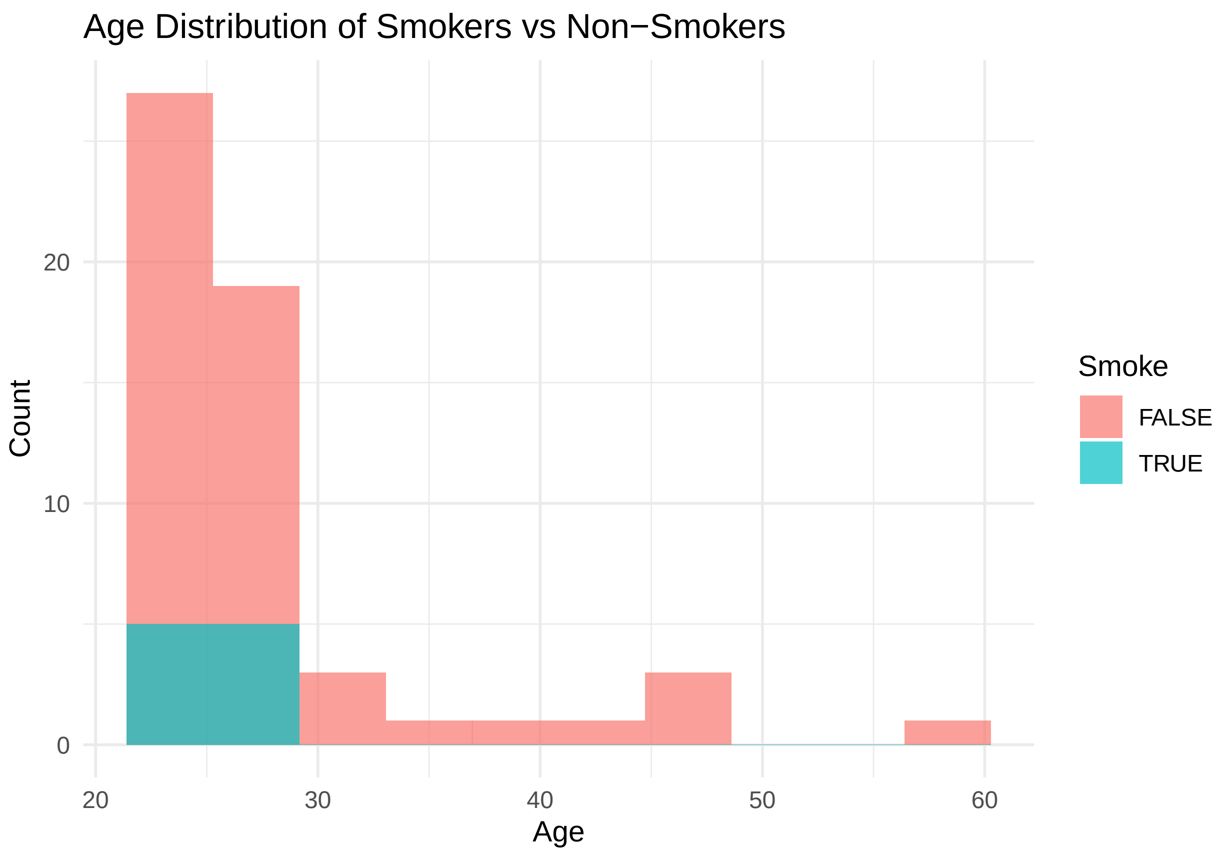
```
p5 <- ggplot(df, aes(x = Subject, fill = Gender)) +
  geom_bar(position = "dodge") +
  theme_minimal() +
  labs(title = "Subject Breakdown by Gender", x = "Subject", y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

print(p1)
```



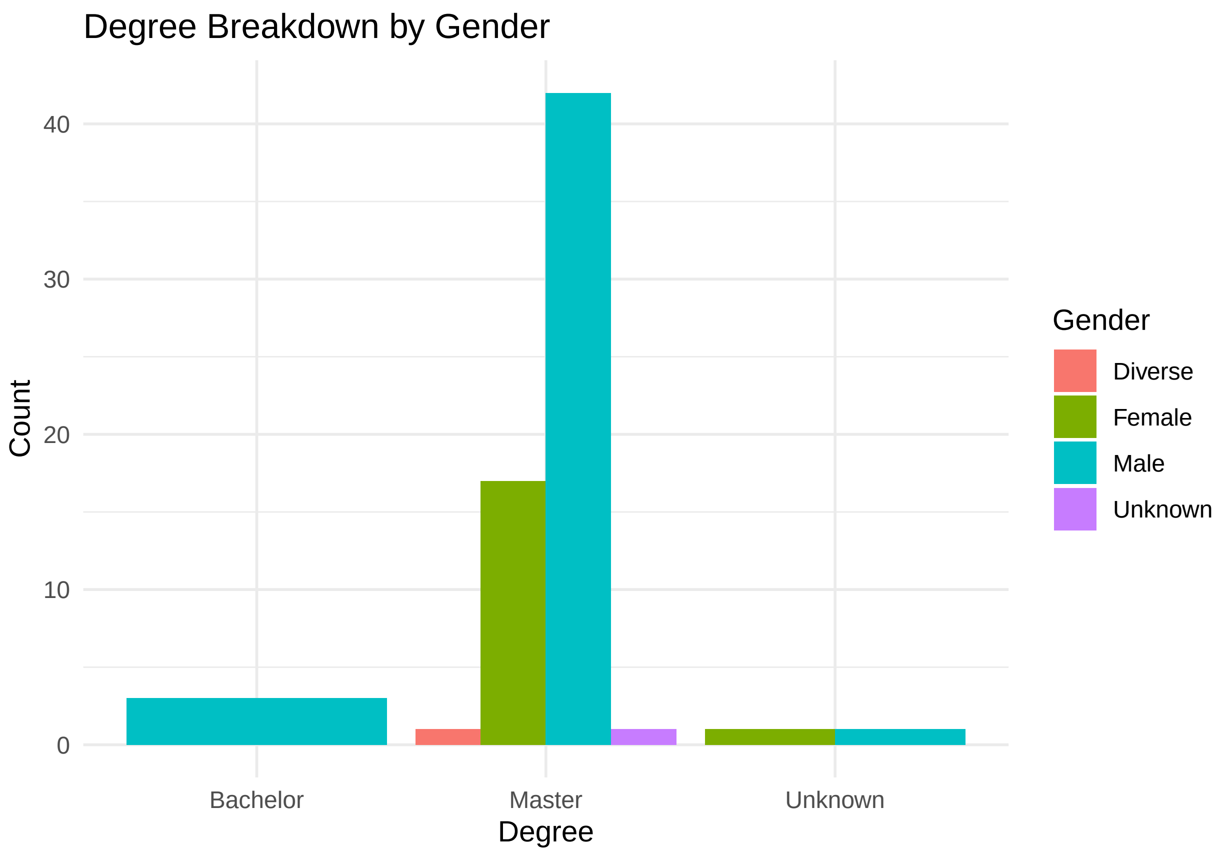
```
print(p2)
```


Smoking Status by Gender
Count
0
10
20
30
Diverse
Female
Male
Unknown
Gender
Smoke
FALSE
TRUE

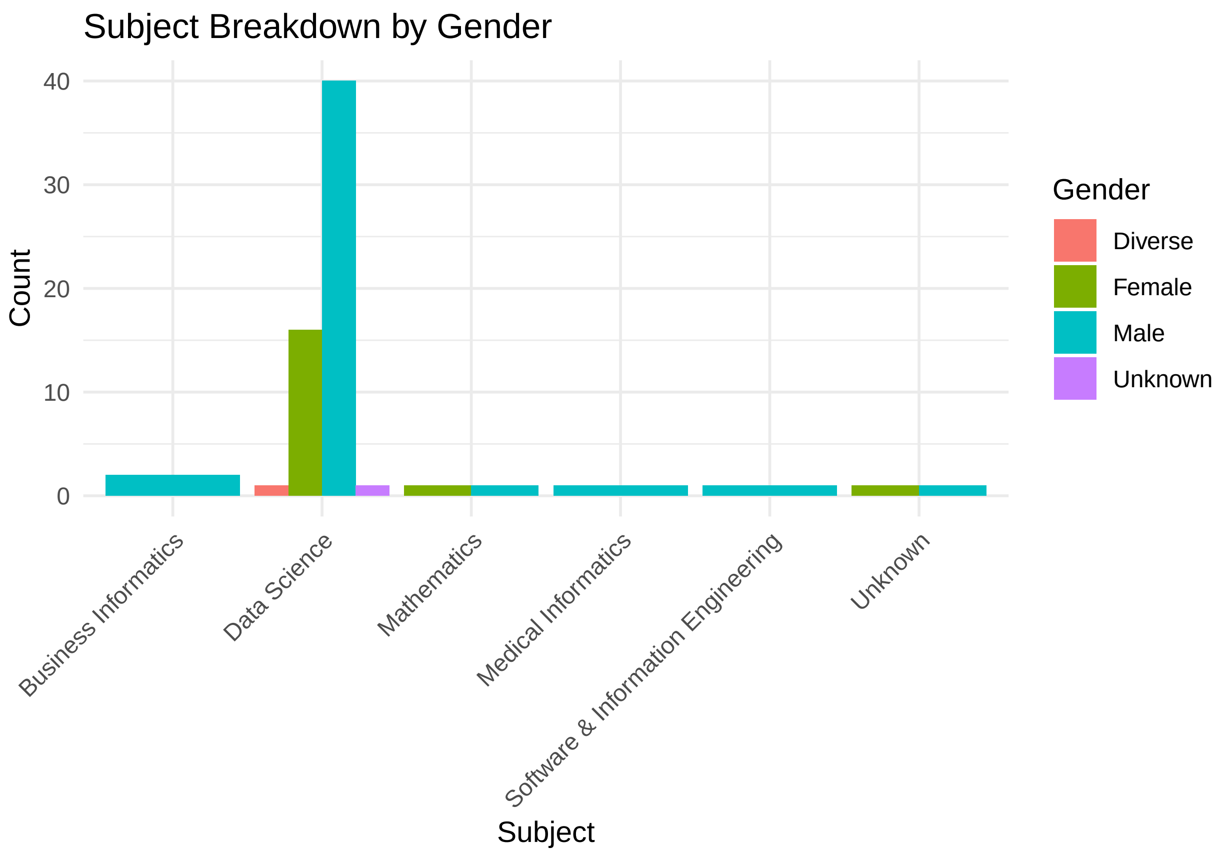
```r
print(p3)
```


Age Distribution of Smokers vs Non−Smokers
Count
Age
0
10
20
20
30
40
50
60
Smoke
FALSE
TRUE

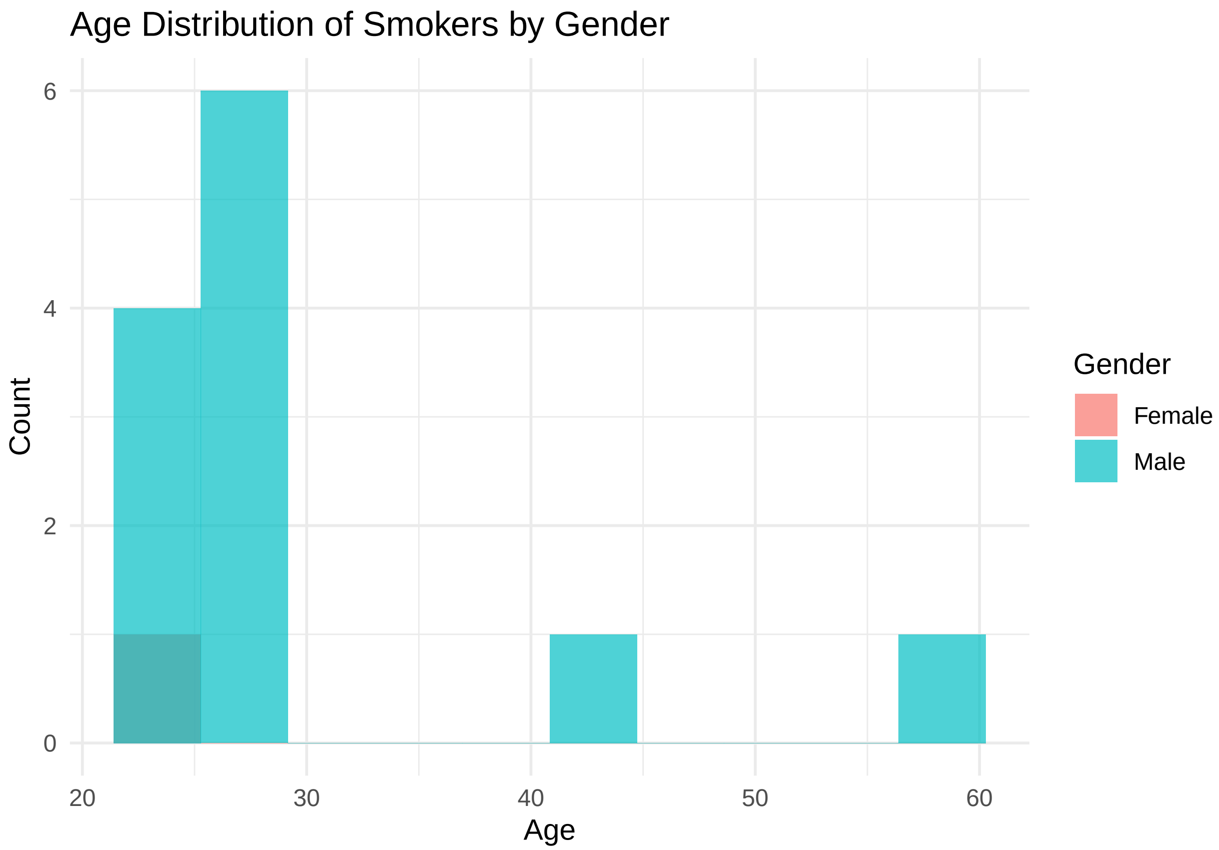
```
print(p4)
```


Degree Breakdown by Gender
Count
40
30
20
10
0
Bachelor
Master
Unknown
Degree
Gender
Diverse
Female
Male
Unknown

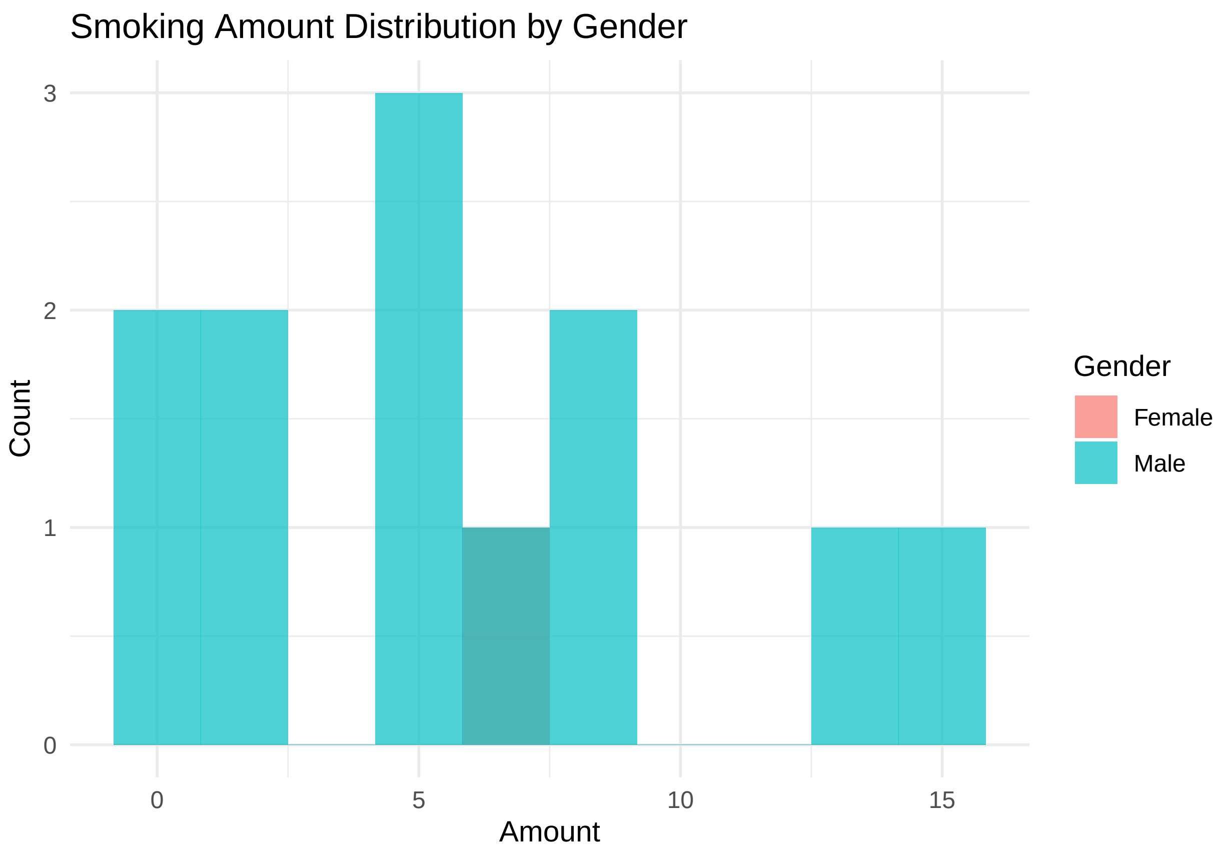
```r
print(p5)
```



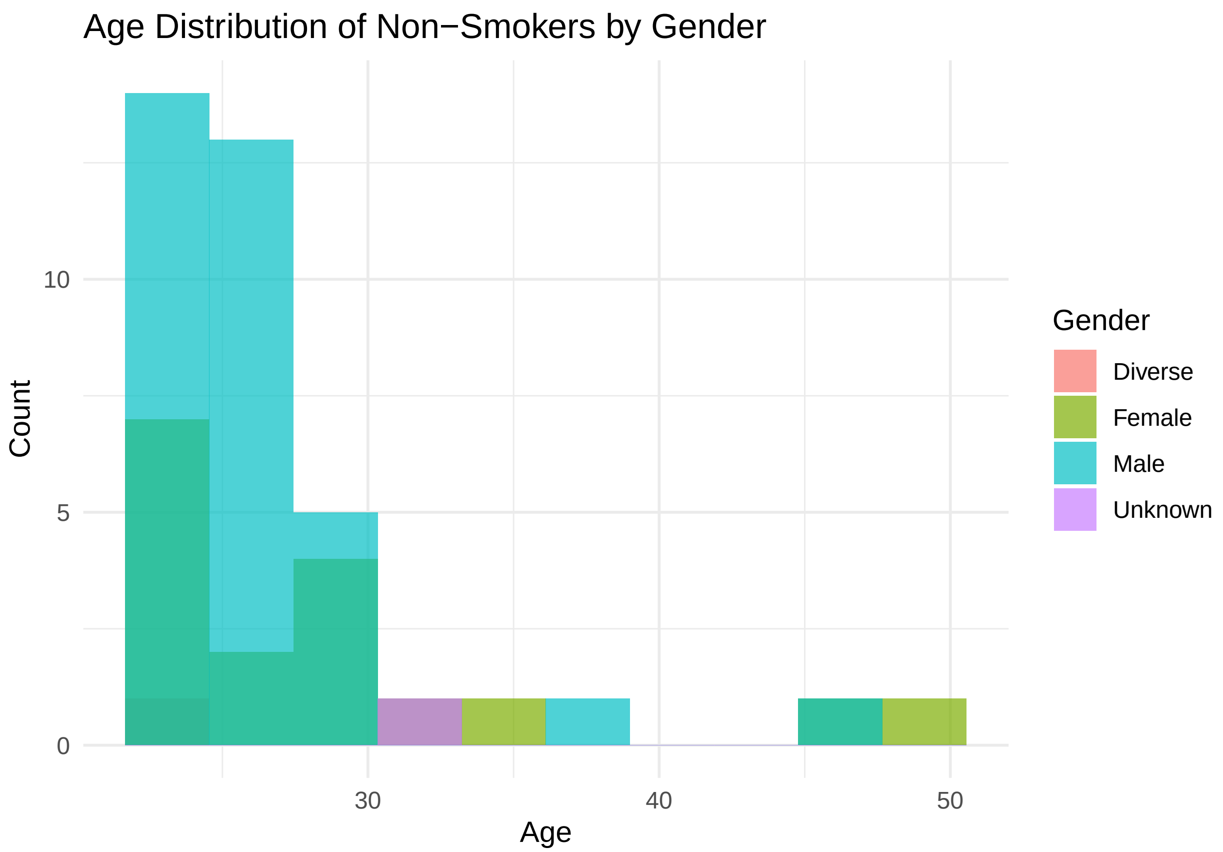
```r
smokers <- df %>% filter(Age_Started_Smoking != 0)
non_smokers <- df %>% filter(Age_Started_Smoking == 0)

# Some more plots for smokers and non smokers
# Plot 6: Age distribution of smokers by gender
p6 <- ggplot(smokers, aes(x = Age, fill = Gender)) +
  geom_histogram(bins = 10, alpha = 0.7, position = "identity") +
  theme_minimal() +
  labs(title = "Age Distribution of Smokers by Gender", x = "Age", y = "Count")

# Plot 7: Smoking amount distribution by gender
p7 <- ggplot(smokers, aes(x = Amount, fill = Gender)) +
  geom_histogram(bins = 10, alpha = 0.7, position = "identity") +
  theme_minimal() +
  labs(title = "Smoking Amount Distribution by Gender", x = "Amount", y = "Count")

# Visualizations for Non-Smokers

# Plot 8: Age distribution of non-smokers by gender
p8 <- ggplot(non_smokers, aes(x = Age, fill = Gender)) +
  geom_histogram(bins = 10, alpha = 0.7, position = "identity") +
  theme_minimal() +
  labs(title = "Age Distribution of Non-Smokers by Gender", x = "Age", y = "Count")

# Plot 9: Degree distribution of non-smokers by gender
p9 <- ggplot(non_smokers, aes(x = Degree, fill = Gender)) +
  geom_bar(position = "dodge") +
  theme_minimal() +
  labs(title = "Degree Distribution of Non-Smokers by Gender", x = "Degree", y = "Count")

print(p6)
```



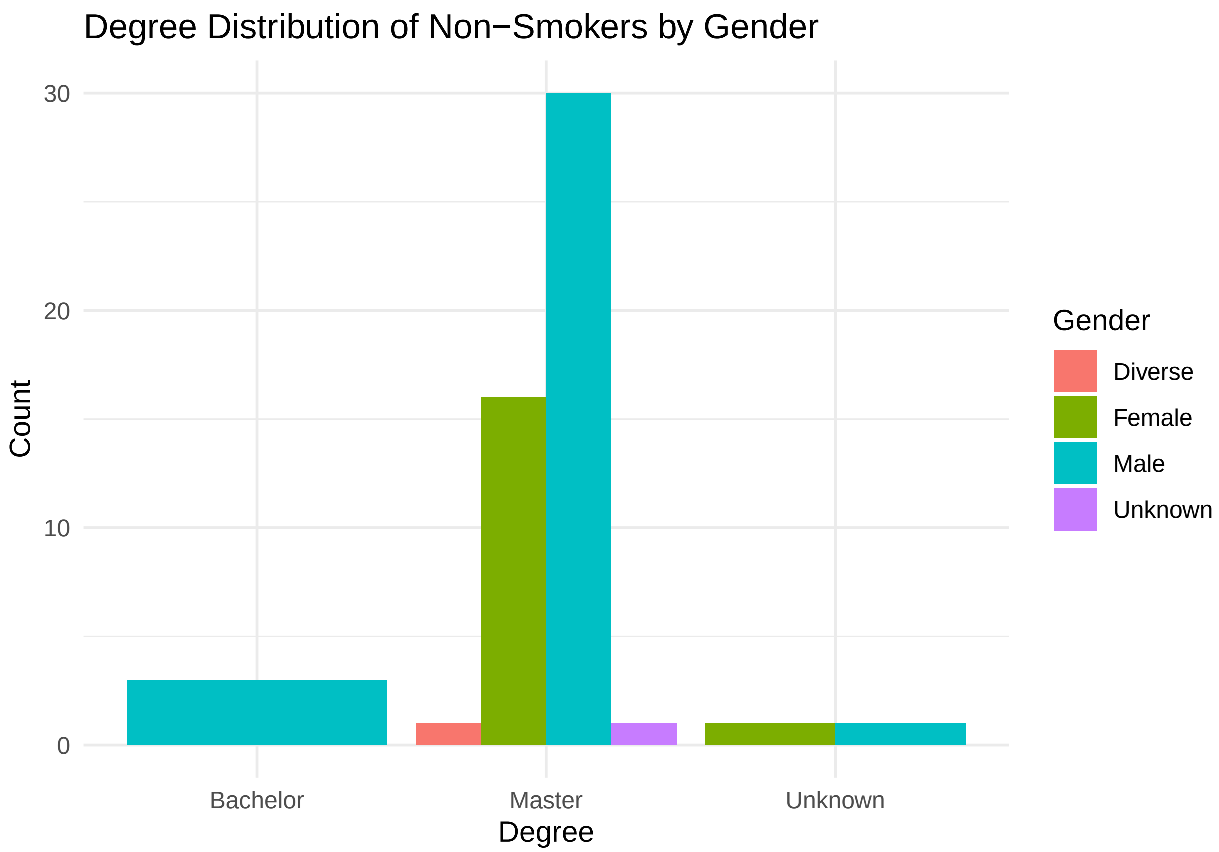
```
print(p7)
```


Smoking Amount Distribution by Gender
Count
3
2
1
0
0
5
10
15
Amount
Gender
Female
Male

```r
print(p8)
```


Age Distribution of Non−Smokers by Gender
Count
10
5
0
30
40
50
Age
Gender
Diverse
Female
Male
Unknown

```
print(p9)
```



```r
perform_chi_square_test <- function(dataset, response_col, predictor_cols) {
  for (predictor_col in predictor_cols) {
    # Create contingency table
    contingency_table <- table(dataset[[response_col]], dataset[[predictor_col]])

    # Perform chi-square test
    chi_square_result <- chisq.test(contingency_table)

    # Print results
    print(paste("Chi-square test for", response_col, "and", predictor_col))
    print(chi_square_result)
  }
}


# Example usage of chi-square test function
perform_chi_square_test(df, "Smoke", c("Gender", "Degree", "Subject"))
```

```
## Warning in chisq.test(contingency_table): Chi-squared approximation may be
## incorrect
```

```
## [1] "Chi-square test for Smoke and Gender"
##
##  Pearson's Chi-squared test
##
```

```
## data:  contingency_table
## X-squared = 2.3435, df = 3, p-value = 0.5042

## Warning in chisq.test(contingency_table): Chi-squared approximation may be
## incorrect

## [1] "Chi-square test for Smoke and Degree"
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 0.96604, df = 2, p-value = 0.6169

## Warning in chisq.test(contingency_table): Chi-squared approximation may be
## incorrect

## [1] "Chi-square test for Smoke and Subject"
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 1.6256, df = 5, p-value = 0.8981
```

exploratory-data-analysis

The age distribution of the dataset reveals that the majority of participants are in their early to mid-20s, with a prominent peak observed in the age range of 22 to 25. This suggests that the primary demographic for this dataset is younger adults, with the concentration of participants diminishing as age increases. A smaller subset of older participants is also present, with the eldest individual being 57 years old. The overall distribution is right-skewed, indicating a higher prevalence of younger individuals among the participants.

In addition to the age distribution, the dataset exhibits a pronounced gender imbalance, with a significant overrepresentation of male participants. This is noteworthy because males constitute the dominant group of smokers within the sample. This demographic trend is crucial for understanding the behavioral patterns observed in the data, particularly in relation to smoking habits.

Moreover, the educational background of the participants predominantly indicates enrollment in master's degree programs. This is in stark contrast to the number of individuals with a bachelor's degree or those whose educational level is unspecified. The focus on master's degree students may reflect the specific population targeted by the study, which could influence the generalizability of the findings to broader educational contexts.

Overall, the dataset is characterized by a young, predominantly male demographic, with most individuals pursuing advanced education at the master's level. These factors combined suggest specific trends and patterns that are important for interpreting the study's outcomes, particularly in relation to age, gender, and educational attainment.

Research-question-1

Are older persons more likely to have started smoking at a younger age than younger students?

Research-question-2

Research Question 2: There is no impact of education level on smoking -> Only data science students in there masters

Research-question-3

Is there a relationship between gender and the amount smoked?

Inference

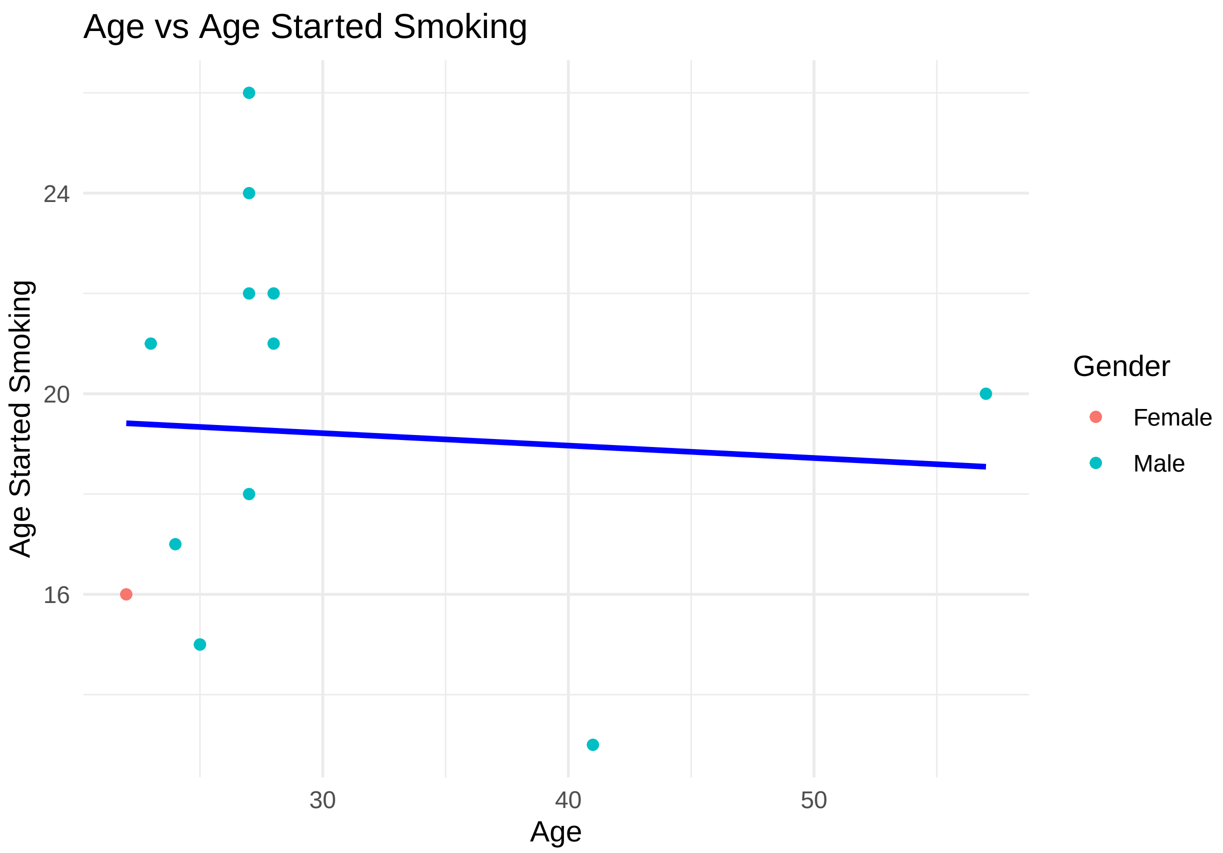
Now we will address each research question.

Research-question-1

To address the research question, we analyzed the relationship between students current age and the age at which they started smoking. A scatter plot with a regression line shows this relationship, and the Pearson correlation coefficient is calculated to quantify it.

```r
# Scatter plot with regression line
p1 <- ggplot(smokers, aes(x = Age, y = Age_Started_Smoking)) +
  geom_point(aes(color = Gender)) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  theme_minimal() +
  labs(title = "Age vs Age Started Smoking", x = "Age", y = "Age Started Smoking")

# Print the plot
print(p1)
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



```
# Calculate correlation
correlation <- cor(smokers$Age, smokers$Age_Started_Smoking)
print(paste("Correlation between Age and Age Started Smoking:", correlation))
```

```
## [1] "Correlation between Age and Age Started Smoking: -0.0604858138049794"
```

The correlation between the current age and the age at which individuals started smoking is -0.0605. This weak negative correlation shows that, in this dataset, there is a tendency for older individuals to have started smoking at a younger age. However, the correlation is very close to zero, indicating that the relationship is weak and not statistically significant. Therefore, we cannot conclusively say that older individuals are more likely to have started smoking at a younger age based on this correlation metric alone.

```
perform_chi_square_test(smokers, "Age_Started_Smoking", c("Age"))
```

```
## Warning in chisq.test(contingency_table): Chi-squared approximation may be
## incorrect
```

```
## [1] "Chi-square test for Age_Started_Smoking and Age"
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 76.375, df = 63, p-value = 0.12
```

To further investigate the relationship between age and the age at which individuals started smoking, we performed a chi-square test. The test compared the age of individuals and the age they started smoking. The chi-square test returned a p-value of 0.12 with an X-squared value of 76.375 and 63 degrees of freedom, suggesting no significant association between the two variables at the common significance level of 0.05.

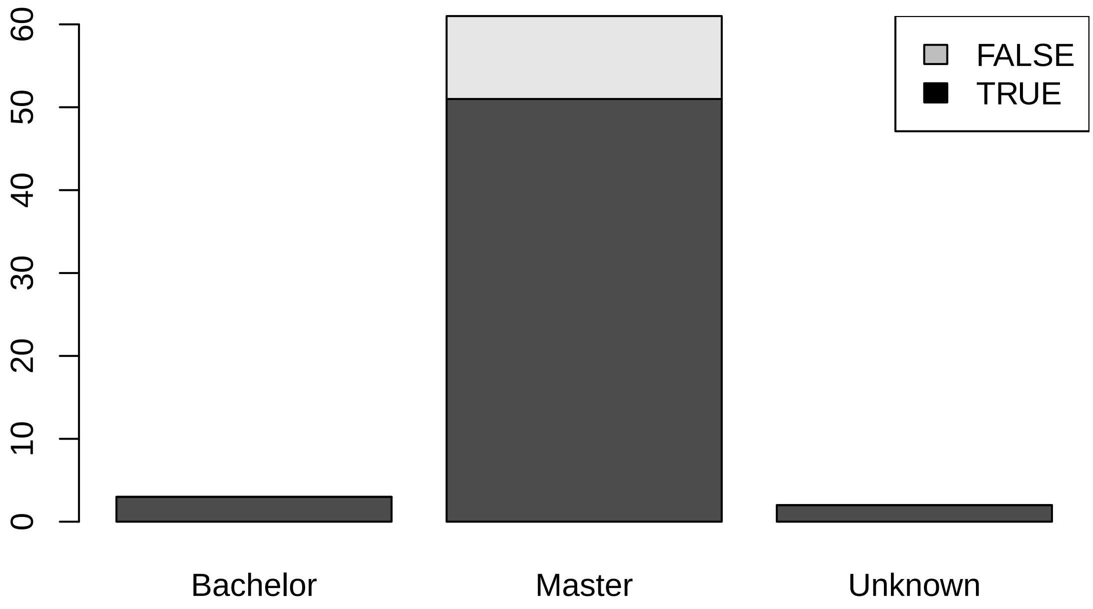
However, the warning "Chi-Quadrat-Approximation kann inkorrekt sein" indicates inaccuracies due to low frequencies in the contingency table. This warning suggests that the chi-square test results may not be reliable, which could be due to the sparsity of the data. This issue arises from the low number of smokers in the dataset.

Both the correlation analysis and the chi-square test do not provide strong evidence to support the hypothesis that older individuals are more likely to have started smoking at a younger age. The weak correlation and potential inaccuracies in the chi-square test due to low frequencies imply that these results should be interpreted with caution.

Research-question-2

To address the research question, we analyzed the relationship between students current Degree and the current status of there smoking habits.

```r
smoke_vs_degree <- table(df$Smoke, df$Degree)

barplot(smoke_vs_degree)

## add legend to the plot
legend("topright", legend = rownames(smoke_vs_degree), fill = c("grey", "black"))
```



The provided bar plot illustrates the distribution of participants across three educational levels: Bachelor, Master, and Unknown, with each bar segmented by boolean values (TRUE and FALSE). The x-axis represents these educational categories, while the y-axis denotes the participant count, incrementing from 0 to 70. The legend differentiates TRUE (dark shade) from FALSE (light shade). The Bachelor category shows an almost negligible count, entirely marked as FALSE. In stark contrast, the Master category stands out with the highest count, slightly over 60, predominantly categorized as TRUE, with a smaller fraction marked as FALSE. The Unknown category, slightly more populated than Bachelor, still has a low count, exclusively marked as TRUE. These observations suggest a significant focus on individuals with a Master's degree within the dataset, this indicate a very concentrated dataset, with a majority of participants in Master's degree.

```r
perform_chi_square_test(df, "Smoke", c("Degree"))
```

```
## Warning in chisq.test(contingency_table): Chi-squared approximation may be
## incorrect
```

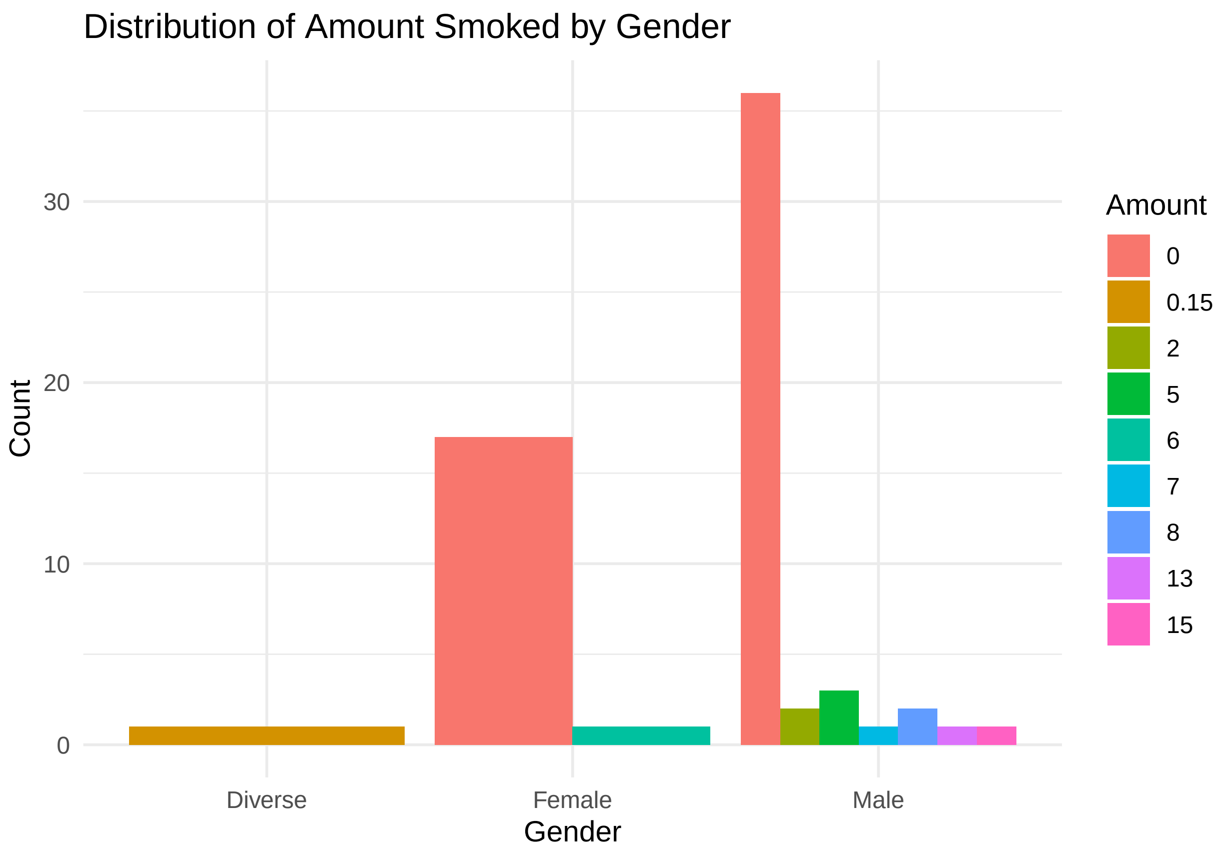
```
## [1] "Chi-square test for Smoke and Degree"
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 0.96604, df = 2, p-value = 0.6169
```

The chi-square test, with a p-value of 0.6169, indicates no significant association between educational levels and boolean values in this dataset. However, the results must be interpreted with caution due to the unequal distribution of participants, predominantly skewed towards those with a Master's degree. The insufficient data in the Bachelor and Unknown categories limits the ability to detect potential associations. For more conclusive results, it would be advantageous to obtain a more balanced dataset, ensuring adequate representation across all educational categories.

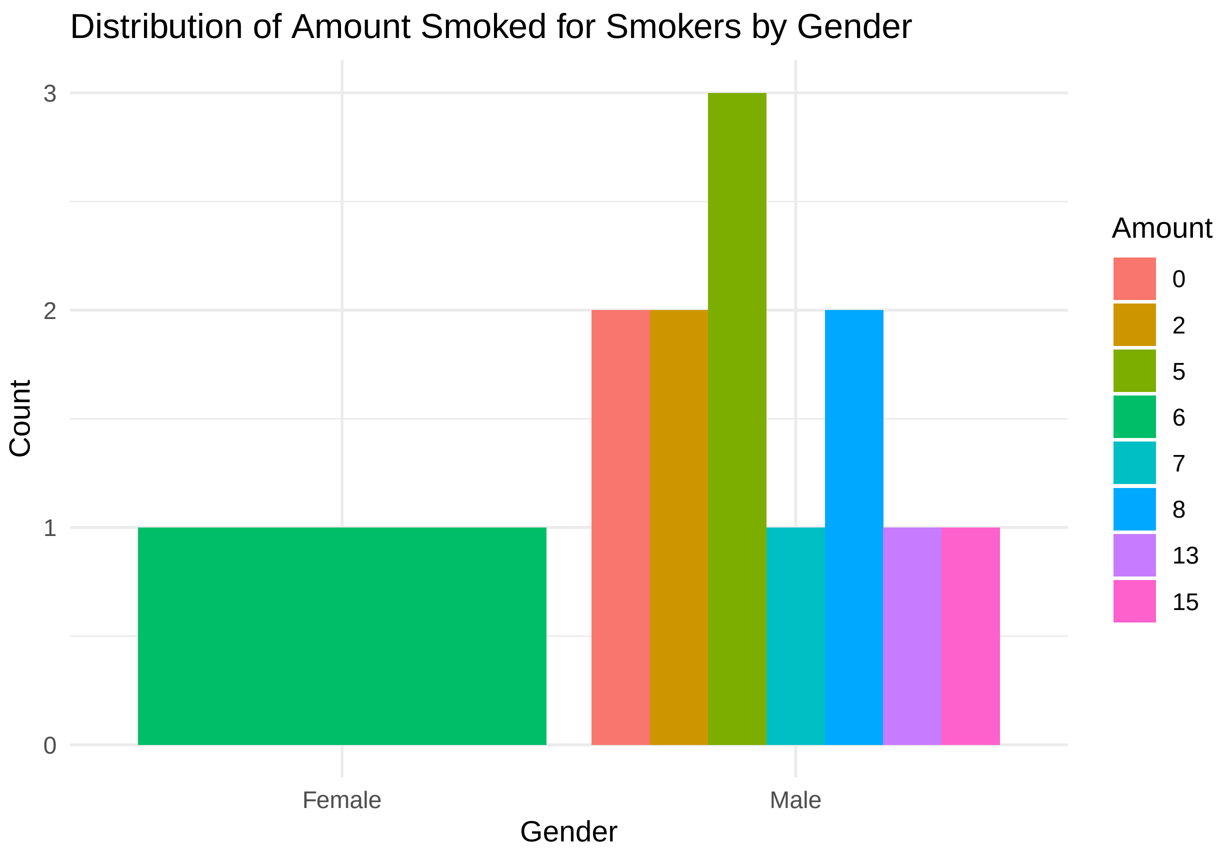
Research-question-3

To investigate the relationship between gender and the amount smoked, we performed a chi-square test using filtered datasets where gender was specified and the amount smoked per day was known (not na). We test if there is a significant association between gender and the distribution of smoking amounts. We filtered the dataset to include only students with known gender and specified smoking amounts, and similarly filtered the smokers dataset to include students with known genders. This allowed us to analyze both datasets separately. We visualized the distribution of smoking amounts among genders using bar plots. The first plot shows the overall distribution, while the second focuses specifically on individuals who smoke.

```r
filtered_data <- df %>% filter(!(is.na(Amount)) & Gender != "Unknown")
filtered_smokers <- smokers %>% filter(Gender != "Unknown")

# Create a bar plot
ggplot(filtered_data, aes(x = Gender, fill = factor(Amount))) +
  geom_bar(position = "dodge") +
  labs(x = "Gender", y = "Count", fill = "Amount") +
  ggtitle("Distribution of Amount Smoked by Gender") +
  theme_minimal()
```



```
# Create a bar plot
ggplot(filtered_smokers, aes(x = Gender, fill = factor(Amount))) +
  geom_bar(position = "dodge") +
  labs(x = "Gender", y = "Count", fill = "Amount") +
  ggtitle("Distribution of Amount Smoked for Smokers by Gender") +
  theme_minimal()
```



```
perform_chi_square_test(filtered_data, "Amount", c("Gender"))
```

```
## Warning in chisq.test(contingency_table): Chi-squared approximation may be
## incorrect
```

```
## [1] "Chi-square test for Amount and Gender"
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 71.985, df = 16, p-value = 4.47e-09
```

```
perform_chi_square_test(filtered_smokers, "Amount", c("Gender"))
```

```
## Warning in chisq.test(contingency_table): Chi-squared approximation may be
## incorrect
```

```
## [1] "Chi-square test for Amount and Gender"
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 13, df = 7, p-value = 0.07211
```

We conducted chi-square tests for both datasets to evaluate the association between gender and smoking amounts.

- For the whole dataset, the chi-square test yielded a p-value of 4.47e-09, indicating a highly significant association between gender and the amount smoked.
- For the smokers dataset, the chi-square test returned a p-value of 0.08838, suggesting no significant association between gender and the amount smoked among smokers specifically. However, the warning "Chi-Quadrat-Approximation kann inkorrekt sein" suggests caution due to potential inaccuracies from low frequencies.

Based on these findings, there is strong evidence from the general dataset that gender influences the distribution of smoking amounts. However, among individuals who smoke, the association between gender and smoking amount is not statistically significant at the significance level of 0.05, although this conclusion should be interpreted cautiously due to the warning from the chi-square test.

Conclusion

- Our dataset where heavily skewed towards master students, which made it hard to draw conclusions about the impact of education level on smoking habits.
- We found no significant association between age and the age at which individuals started smoking, suggesting that older individuals are not more likely to have started smoking at a younger age.
- gender influences the distribution of smoking amounts in the general dataset, but not among smokers specifically, and statistical significance was not observed in the latter case.
- the weak correlation and potential inaccuracies in the chi-square test due to low frequencies indicate that these results should be interpreted with caution.

Further comment:

We learned that we should be conscious about those potentioal bias and therefor we need to try to ask questions that are more general and not so specific to a certain group to get as much information as possible to conduct a proper surveys.

Reference

Function/Variable	Description
parse_program	Parse program name for degree (Bachelor/Master) and subject categorization.
parsed_programs	to apply the previous function.
perform_chi_square_test	this function is used to perform a chi-square test between two columns of a dataset
ggplot	to create the plots
correlation	to do pearson correlation between two columns