Assignment 2 - Survey Data Analysis

Haider, Tobias Abraham Selenge, Tuvshin Crespí, Ahmad Antonio

2025-05-30

Contents

Preparation

Imports

```
library(knitr)
library(readr)
library(stringr)
library(tidyr)
library(pillar)
library(dplyr)
library(ggplot2)
```

Global variables

```
data_path <- "./data/survey_data_clean.csv"
```

Data loading

```
survey_data <- read_csv(data_path)
```

Data Cleaning

Manual Data Cleaning

Most of the data cleaning was done by hand. The reason for this is the variety of inconsistencies in the collected textual data. However, the manual cleaning was done to not change any meaning of the data.

Changes data were made by hand:

- Align nationalities to be the country name (Austrian -> Austria, Pakistani -> Pakistan)
- Extract desired degree and university from the study program column. If any of these were not specified, it was left blank
- Remove abbreviations in country names and study programs (A -> Austria, AUT -> Austria)
- Ensure consistent capitalisation (Data science -> Data Science)
- Remove invalid responses (age of 4 years was set to empty)
- Sleep hours that contained dashed were set to a floating point number (6-7 -> 6.5)

Data Frame Structure

To be able to use the table with all common libraries, we ensured to represent the values as their natural data type. All variables are factors, except for the daily hours of sleep.

```
names(survey_data) <- c("gender", "age", "study_program", "desired_degree", "university_name", "daily_s
```

```r
survey_data <- survey_data %>%
  mutate(
    gender = as.factor(gender),
    study_program = as.factor(study_program),
    desired_degree = as.factor(desired_degree),
    university_name = as.factor(university_name),
    contentness = factor(contentness, levels = c("Not at all", "Slightly content", "Quite content", "Ve
    primary_nationality = as.factor(primary_nationality)
  )
```

Grouping of Study Programs

While plotting, it was notable that some verbose descriptions of study programs. Luckily, most responses can be grouped into more general categories of study programs. This leads to some data loss but improves the expressiveness of all analysis done afterwards. Above all, plots are less noisy.

```r
survey_data <- survey_data %>%
  mutate(
    study_program = case_when(
      str_detect(study_program, regex("Management", ignore_case = TRUE)) ~ "Management",
      str_detect(study_program, regex("Mathematics", ignore_case = TRUE)) ~ "Mathematics",
      study_program == "Außerordentliche Studentin" ~ NA,
      TRUE ~ study_program
    )
  )
```

Data Exploration

Before performing the statistical analysis, we can take a look at the responses and visualize some of the values collected.

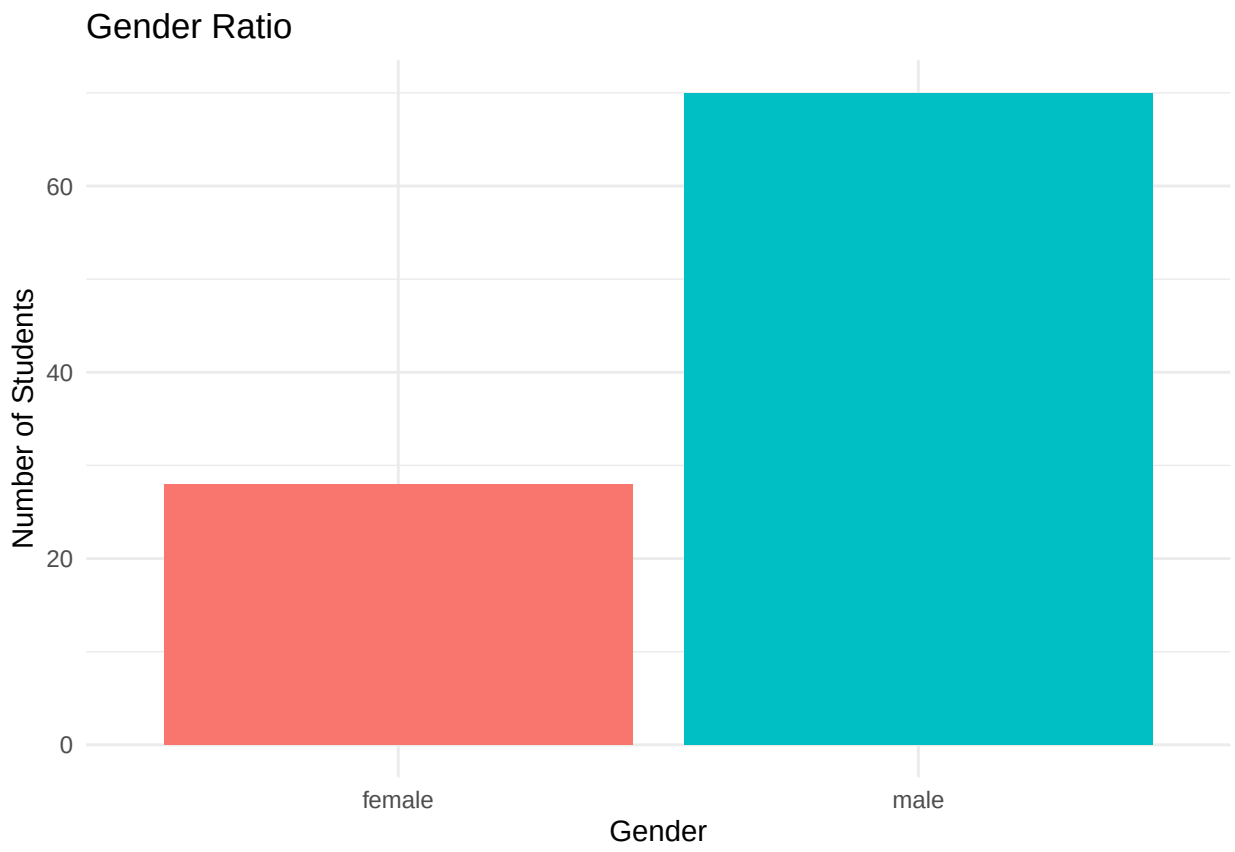
```r
glimpse(survey_data)
```

```
## Rows: 98
## Columns: 8
## $ gender              <fct> male, male, female, male, male, male, male, male, ~
## $ age                 <dbl> 23, 24, 24, 23, 22, 22, 33, 24, 26, 45, 28, 24, 22~
## $ study_program       <chr> "Data Science", "Data Science", "Data Science", "D~
## $ desired_degree      <fct> MSc, NA, NA, MSc, MSc, MSc, NA, MSc, MSc, NA, MSc,~
## $ university_name     <fct> TU Wien, NA, NA, TU Wien, TU Wien, TU Wien, NA, TU~
## $ daily_sleep         <dbl> 6.0, 8.0, 7.0, 8.0, 7.0, 8.0, 7.0, 8.0, 8.0, 7.0, ~
## $ contentness         <ord> Quite content, Slightly content, Quite content, Qu~
## $ primary_nationality <fct> India, Austria, Bulgaria, Austria, Bosnia and Herz~
```

Gender Ratio

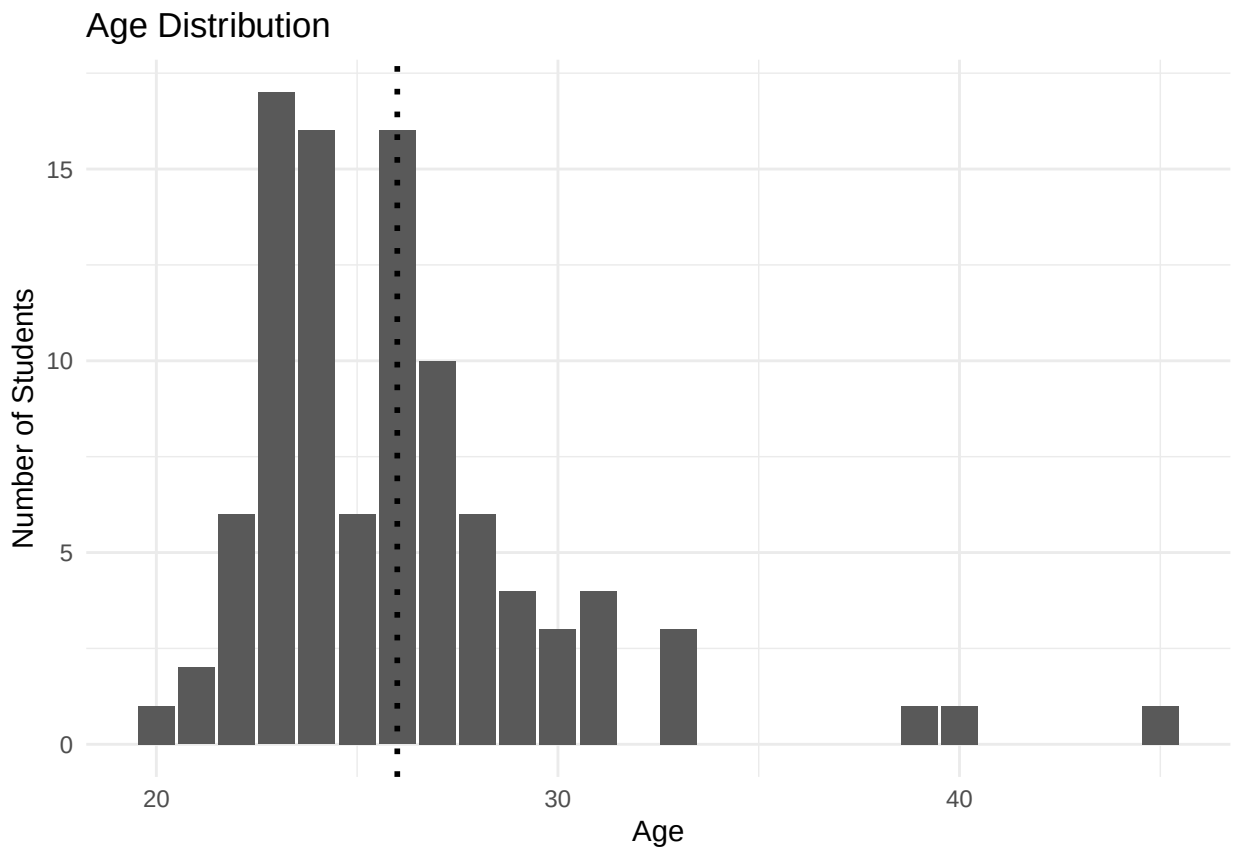
```r
ggplot(survey_data, aes(x = gender, fill = gender)) +
  geom_bar() +
```

```
  theme_minimal() +
  guides(fill="none") +
  labs(title = "Gender Ratio", x = "Gender", y = "Number of Students")
```



Age distribution plot

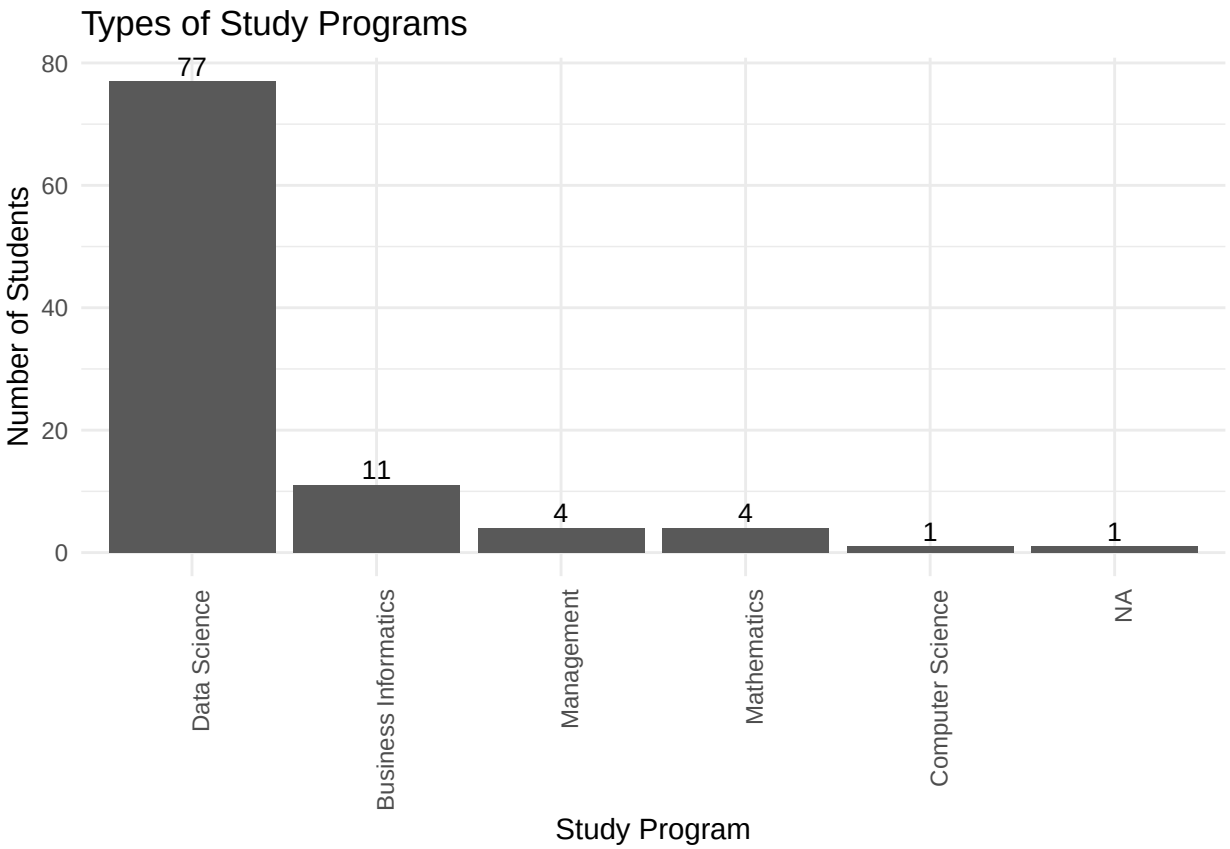
```
ggplot(survey_data, aes(x = age)) +
  geom_bar() +
  geom_vline(aes(xintercept = median(age, na.rm = TRUE)), linetype = "dotted", size = 1) +
  theme_minimal() +
  labs(title = "Age Distribution", x = "Age", y = "Number of Students")
```



The median of the student age is 26, while almost all studens are between the age of 20 and 30 years.

Types of Study Programs

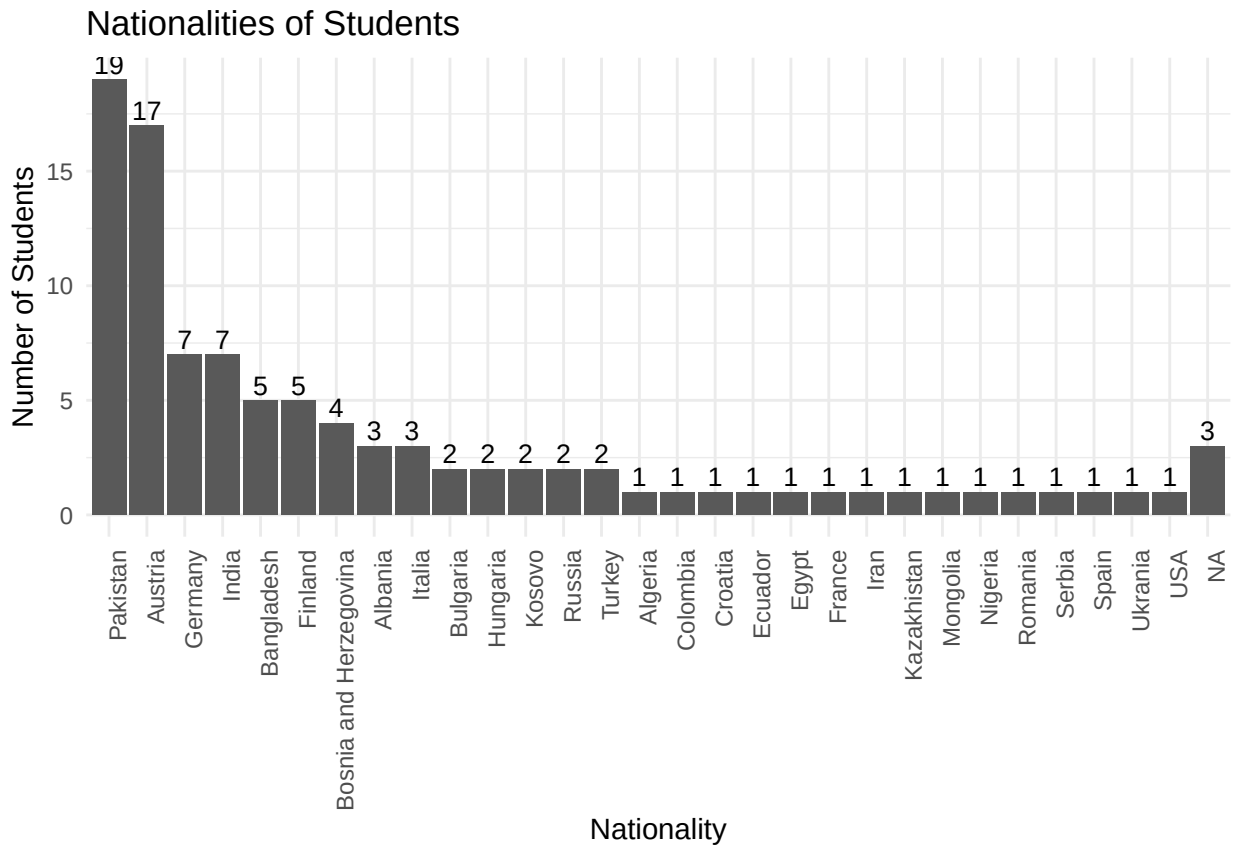
```
ggplot(survey_data, aes(x = reorder(study_program, study_program, function(x) - length(x)))) +
  geom_bar() +
  geom_text(stat = "count", aes(label = after_stat(count)), vjust = -0.3, size = 3.5) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(title = "Types of Study Programs", x = "Study Program", y = "Number of Students")
```



This plot clearly shows that Data Science is the study program with most students attending the course.

Different Nationalities

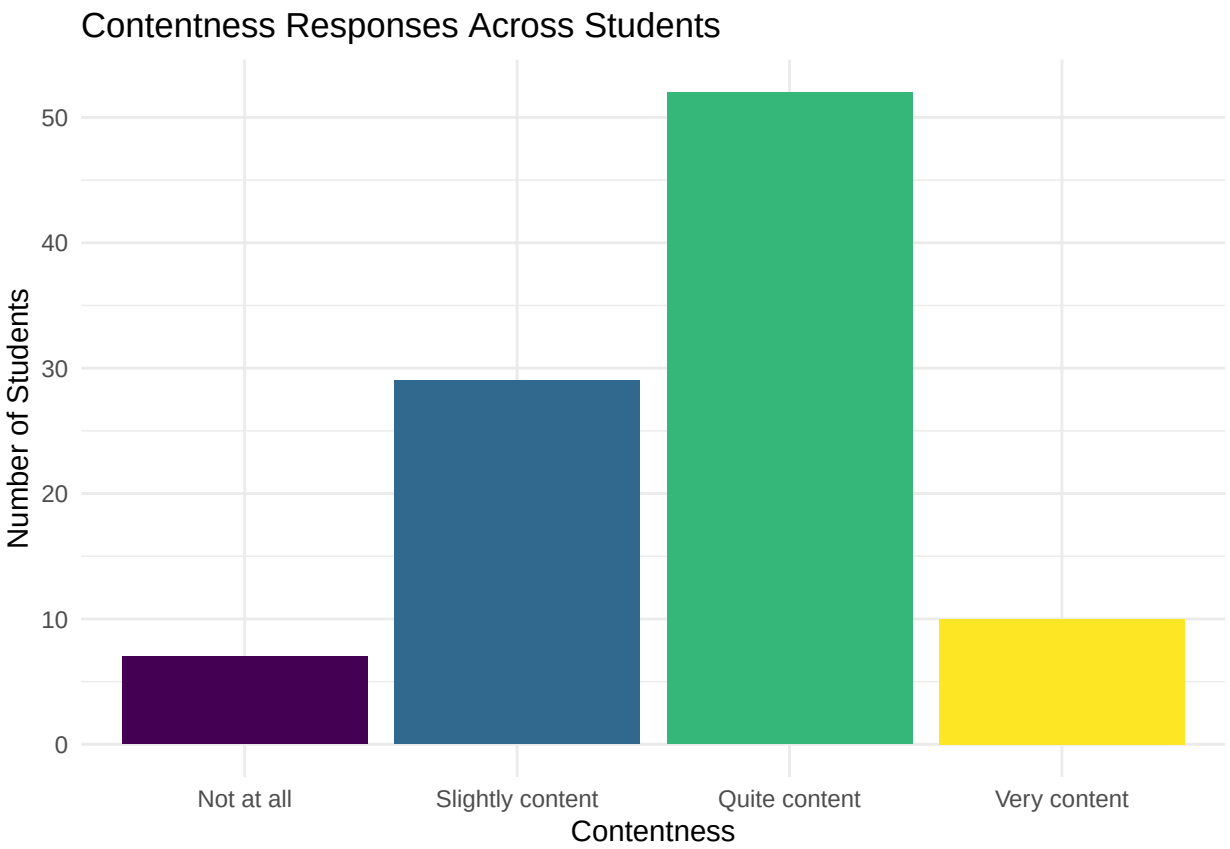
```
ggplot(survey_data, aes(x = reorder(primary_nationality, primary_nationality, function(x) - length(x)))
  geom_bar() +
  geom_text(stat = "count", aes(label = after_stat(count)), vjust = -0.3, size = 3.5) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(title = "Nationalities of Students", x = "Nationality", y = "Number of Students")
```



When ordering the number of students per nationality, it can be seen that Austria is not the top country. Many students from Pakistan attend this course. The majority of students have nationalities from Central and Eastern Europe.

Measure of Contentness

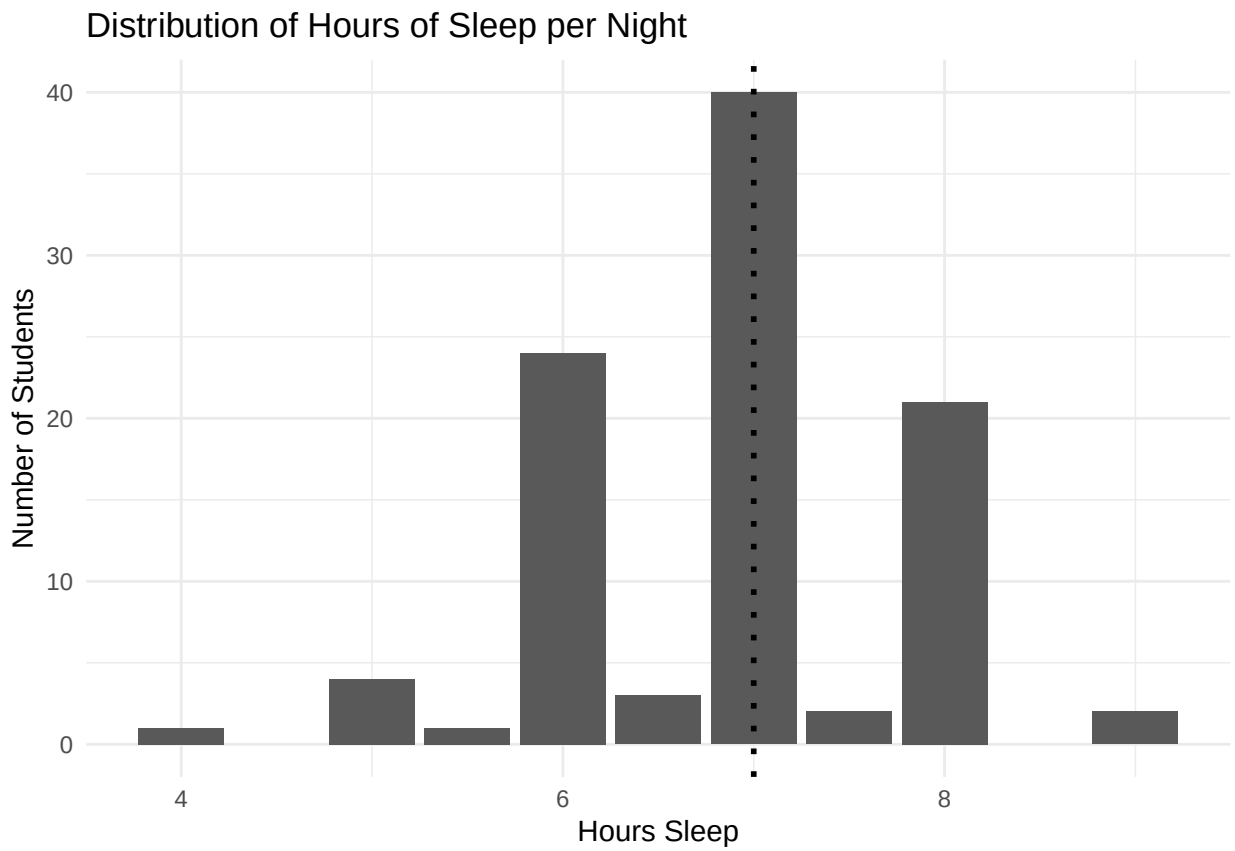
```
ggplot(survey_data, aes(x = contentness, fill = contentness)) +
  geom_bar() +
  theme_minimal() +
  guides(fill="none") +
  labs(title = "Contentness Responses Across Students", x = "Contentness", y = "Number of Students")
```



When looking at the contentness isolated from all other variables, it becomes apparent that studens are overall content. Around two thirds are quite or very content and one third is only slightly or not content at all.

Distribution of Sleep Hours per Night

```
ggplot(survey_data, aes(x = daily_sleep)) +
  geom_bar() +
  geom_vline(aes(xintercept = median(daily_sleep, na.rm = TRUE)), linetype = "dotted", size = 1) +
  theme_minimal() +
  labs(title = "Distribution of Hours of Sleep per Night", x = "Hours Sleep", y = "Number of Students")
```



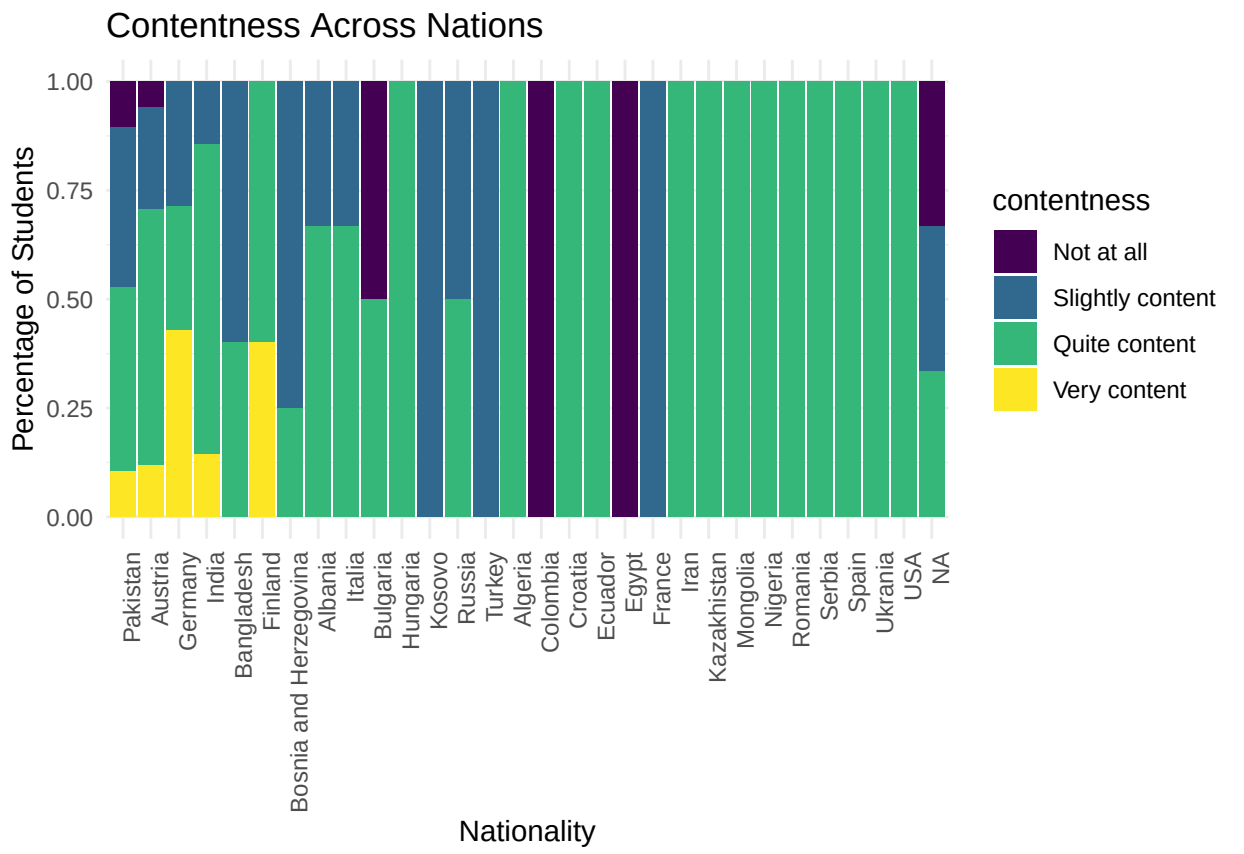
Students sleep around 7 hours per night. There is some variation present in the data. 6, 7, and 8 are the most common response.

Visualization of Hypotheses

Next, we represented our hypotheses in the form of plots to see whether there are some obvious patterns.

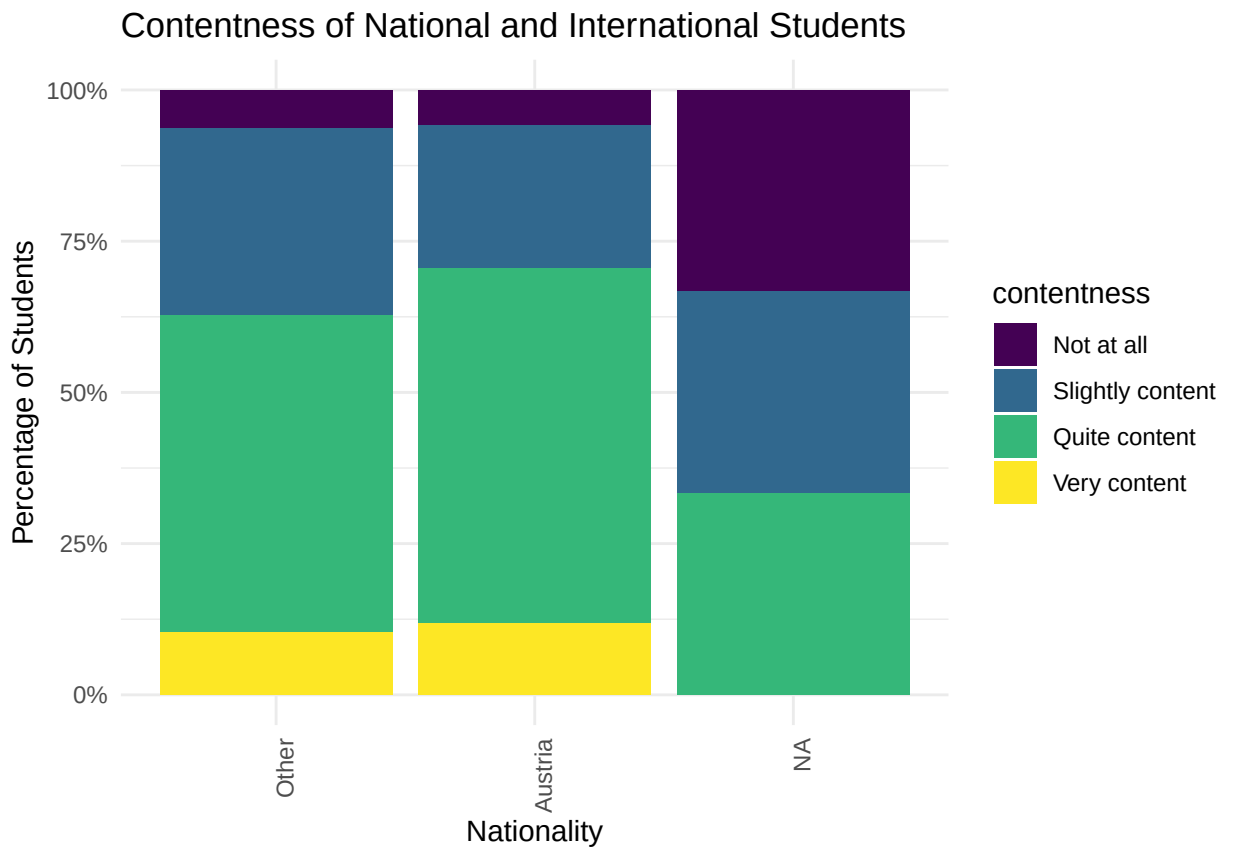
Visualization of Relation Between Nationality and Contentness

```
ggplot(survey_data, aes(x = reorder(primary_nationality, primary_nationality, function(x) - length(x)),
  geom_bar(position = "fill") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(title = "Contentness Across Nations", x = "Nationality", y = "Percentage of Students")
```



Students from Germany, Finland, Austria and Pakistan seem to be the happiest amongst all. However, we have to note that most contries do not have sufficient responses to actually compare them.

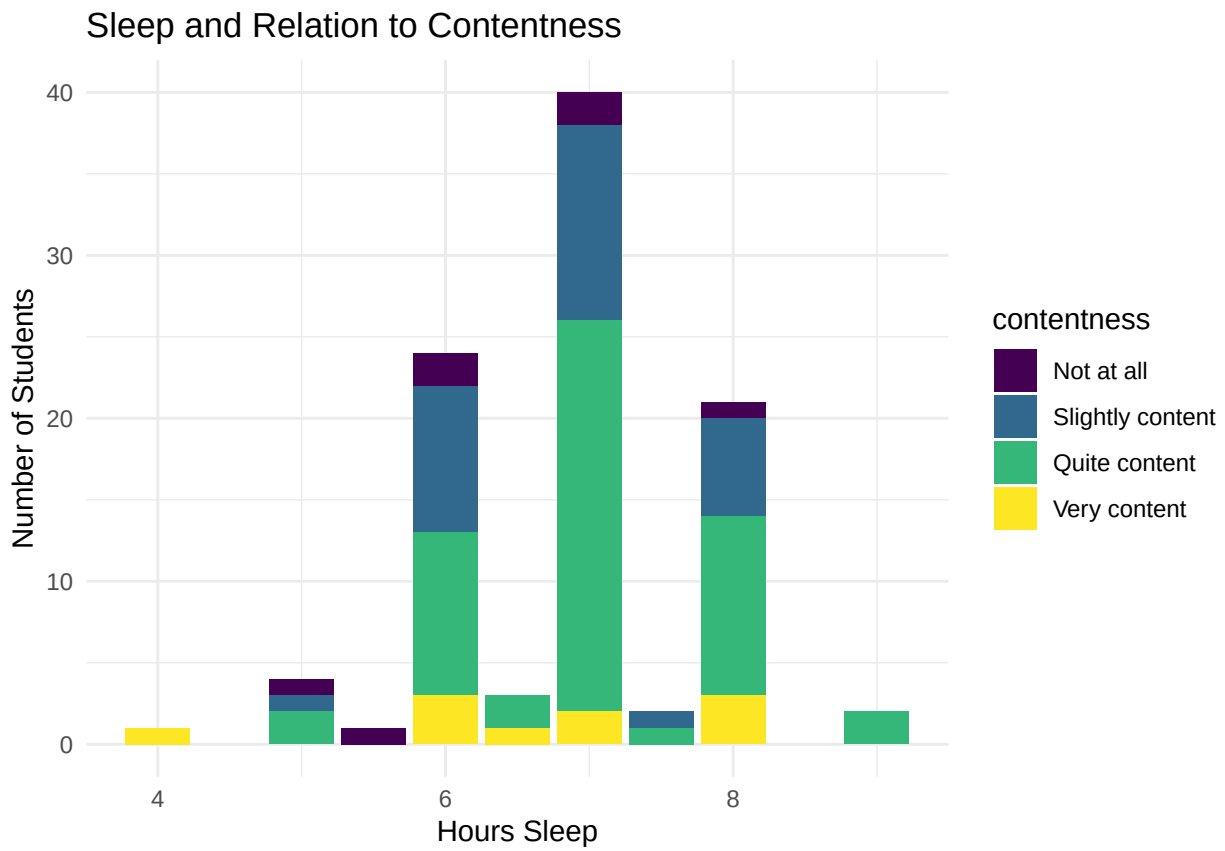
```
survey_data %>%
  mutate(
    primary_nationality = ifelse(primary_nationality == "Austria", "Austria", "Other")
  ) %>%
  ggplot(aes(x = reorder(primary_nationality, primary_nationality, function(x) - length(x)), fill = con
  geom_bar(position = "fill") +
  scale_y_continuous(labels = scales::percent_format()) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(title = "Contentness of National and International Students", x = "Nationality", y = "Percentage
```



This plot compares the contentness of Austrians specifically to international students. According to the data, Austrians feel slightly more content than international students. It is notable, that unhappy students are more likely to decline to disclose their nationality. Because of this, there is some bias in the collected data.

Visualization of Relation Between Sleep and Contentness

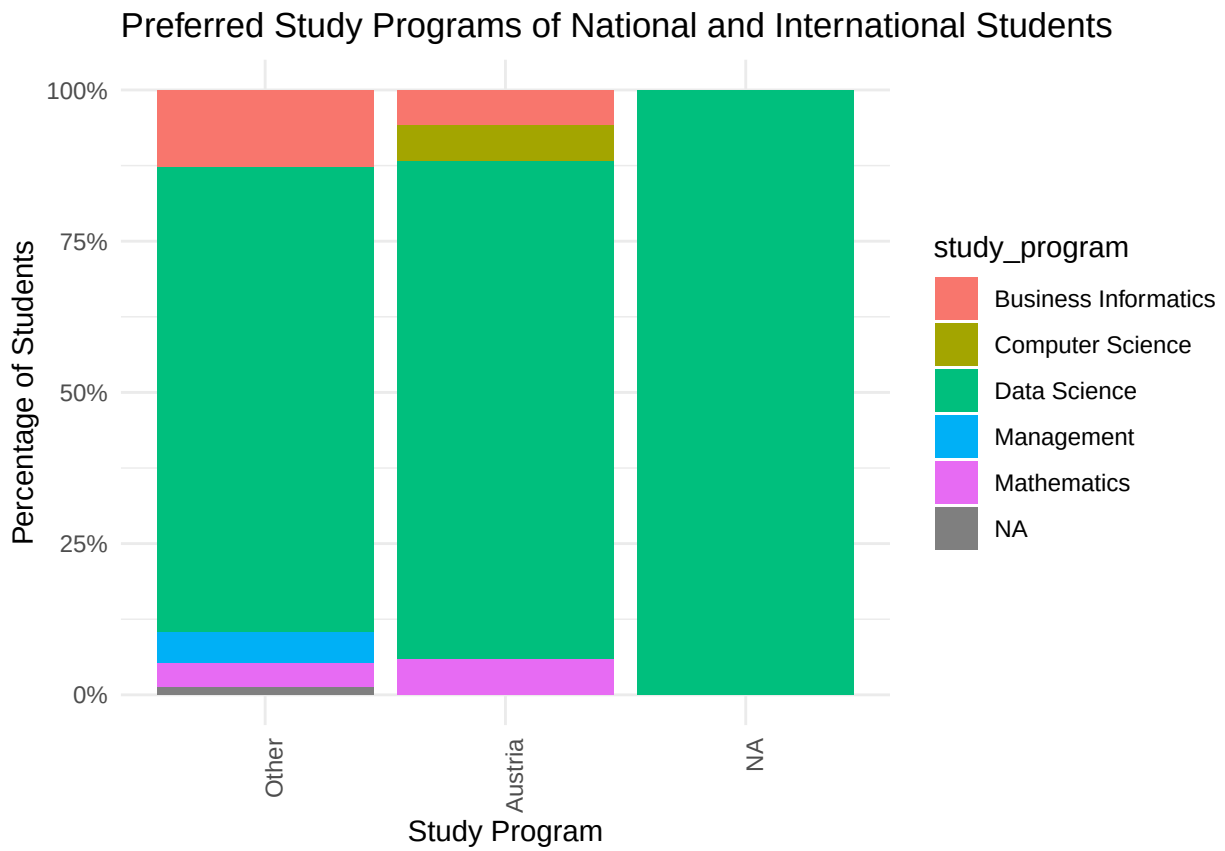
```
ggplot(survey_data, aes(x = daily_sleep, fill = contentness)) +
  geom_bar() +
  theme_minimal() +
  labs(title = "Sleep and Relation to Contentness", x = "Hours Sleep", y = "Number of Students")
```



There is no clear pattern regarding the hours of sleep and contentness. We see an increase between the ratio of quite to slightly content when the hours of sleep is higher. The change could also be completely random.

Visualization of Relation Between Nationality and Study Program

```
survey_data %>%
  mutate(
    primary_nationality = ifelse(primary_nationality == "Austria", "Austria", "Other")
  ) %>%
  ggplot(aes(x = reorder(primary_nationality, primary_nationality, function(x) - length(x)), fill = stu
  geom_bar(position = "fill") +
  scale_y_continuous(labels = scales::percent_format()) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(title = "Preferred Study Programs of National and International Students", x = "Study Program",
```



Interestingly, there is not a single computer science student with a nationality other than Austrian. Overall, the plot is not displaying anything surprising. Data Science is the biggest group of students attending this cours across all nationalities.

Descriptive Statistics

Hours of Sleep

```
summary(survey_data$daily_sleep)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.000   6.000   7.000   6.878   7.375   9.000
```

Analytical Statistics

Test for Relation Between Nationality and Contentness

Test for Relation Between Sleep and Contentness

```r
anova_result <- aov(daily_sleep ~ contentness, data = survey_data)
summary(anova_result)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## contentness  3   3.35  1.1172    1.37  0.257
## Residuals   94  76.68  0.8157
```

Test for Relation Between Nationality and Study Program

Conclusion