Data Acquisition and Survey Methods - Group Project

Assignment 2: Survey Analysis Report, Group 1

Anna Stonek, Jakub Kotala and Meiqi Liu

2024-06-21

Contents

1 Basic setup

```
# Set the code chunks visible throughout the document
knitr::opts_chunk$set(echo = TRUE)
```

2 Loading necessary libraries

```
# List of required packages
required_packages <- c("knitr", "tidyverse", "kableExtra", "lessR", "infer")

# Function to check if the package is installed on local environment,
# If not, install first, then load to the library
check_and_install_packages <- function(packages) {
  for (pkg in packages) {
    if (!require(pkg, character.only = TRUE)) {
      install.packages(pkg)
      library(pkg, character.only = TRUE)
    }
  }
}

check_and_install_packages(required_packages)
```

3 Introduction

In this group assignment, we are interested on the attitude of students taking this course towards environmental issues.

We know that we will be given basic demographic profile of the students on age, gender and their academic program.

Therefore, we posed three research questions:

1. Is there a correlation between the age and the motivation of students to reduce their carbon footprint?
2. How do students in different Master's degrees feel about the measures TU has implemented to reduce the environmental impact?
3. Will gender affect the importance students place on environmental issues?

Based on the research questions, we developed a survey contains following questions:

1. On a scale from 1 to 10, how motivated are you to take steps towards reducing your carbon footprint, realising that it may raise the cost of living?
    - For example taking an expensive train rather to take a cheap Ryanair flight. 1: Not motivated at all, 10: Very motivated.
2. Do you think TU has already implemented enough measures to reduce their environmental impact?

- - Yes/No.

3. On a scale of 1-5, how important do you think the environmental issues are?
    - 1: there is something else more urgent we need to take care, 5: climate change is the top issue we need to solve now.

4 Loading and cleaning data

Firstly, we need to load the data.

We then examine the data structure, find there is a need to clean the data, since it contains some non-sensible values (for example: "no answer" on gender).

```
df <- read.csv("group1.csv")
str(df)
```

```
## 'data.frame':    66 obs. of  6 variables:
##  $ DemographicAnswer.1: chr  "female" "female" "female" "male" ...
##  $ DemographicAnswer.2: int  22 22 48 25 24 22 26 26 23 27 ...
##  $ DemographicAnswer.3: chr  "Data Science / MSc / TU Wien" "Data Science / MSc / TU Wien" "Data Sci
##  $ Answer.1           : int  4 1 9 1 6 7 6 7 6 4 ...
##  $ Answer.2           : chr  "No" "No" "No" "Yes" ...
##  $ Answer.3           : int  3 1 5 1 4 5 4 4 4 5 ...
```

```
colnames(df) <- c("gender", "age", "study_programme",
                  "question_1", "question_2", "question_3")
```

Since this only concerns one data point, and we have 66 observations in total. This is only 1.5% of the total sample size, we decided to exclude it from the analysis.

Also, two data point only contains the value "TU Wien" for the study programme, so we also decided to exclude this data point from further analysis.

This also applies to one NA value regards the question 1.

As a result, we cleaned 4 entry rows, and now the sample size is 62.

```
# drop NA and non-valuable values
df = df[!df$gender %in% "no answer",]
df = df[!df$study_programme %in% "TU Wien",]
df <- df %>% filter(!is.na(question_1))

# check sample size
nrow(df)
```

```
## [1] 62
```

Furthermore, we decided to bin the age values into groups of even sizes, based on the summary statistics, dividing the groups to: [minimum - 1st quantile], [1st quantile - median], [median - 3rd quantile], [3rd quantile - maximum]

Moreover, we had to clean and generalize the values for the study programme, since this was a free-text field in the survey and students put very different strings there, though meaning the same programme.

```
for (i in (1:length(df$study_programme))){
  if (grepl("Data Science", df$study_programme[i]) |
      grepl("Data science", df$study_programme[i]) |
      grepl("DataScience", df$study_programme[i])){
    df$study_programme[i] = "Data Science MSc"
  }
  if (grepl("Business Informatics", df$study_programme[i])){
    df$study_programme[i] = "Business Informatics MSc"
  }
  if (grepl("Medical", df$study_programme[i])){
    df$study_programme[i] = "Medical Informatics BSc"
  }
  if (grepl("Mathematics / MSc", df$study_programme[i])){
    df$study_programme[i] = "Mathematics MSc"
  }
  if (grepl("Software", df$study_programme[i])){
    df$study_programme[i] = "Software & Information Engineering BSc"
  }
  if (grepl("Bachelor for", df$study_programme[i])){
    df$study_programme[i] = "Technical Mathematics BSc"
  }
  if (df$age[i] < 25){
    df$age[i] = "22-24"
  } else if (df$age[i] >= 25 & df$age[i] < 27){
    df$age[i] = "25-26"
  } else if (df$age[i] >= 27 & df$age[i] < 29){
    df$age[i] = "27-28"
  } else if (df$age[i] >= 29){
    df$age[i] = "29-57"
  }
}
```

Next, based on the column, we categorised the data.

```
# factorise column
df$gender <- factor(df$gender, levels = c("male","female","diverse"))
df$age <- factor(df$age, levels = c("22-24","25-26","27-28","29-57"))
df$study_programme <- factor(df$study_programme,
                             levels = c("Medical Informatics BSc",
                                        "Software & Information Engineering BSc",
                                        "Technical Mathematics BSc",
                                        "Business Informatics MSc",
                                        "Data Science MSc",
                                        "Mathematics MSc"))
df$question_2 <- factor(df$question_2, levels = c("Yes","No"))

#summary dataset
summary(df)
```

```
##      gender      age                                    study_programme
##  male   :44   22-24:24   Medical Informatics BSc               : 1
##  female :17   25-26:13   Software & Information Engineering BSc: 1
##  diverse: 1   27-28:13   Technical Mathematics BSc             : 1
```

```
##               29-57:12   Business Informatics MSc             : 2
##                          Data Science MSc                     :56
##                          Mathematics MSc                      : 1
##    question_1     question_2   question_3
##  Min.   : 1.000   Yes:24     Min.   :1.000
##  1st Qu.: 4.000   No :38     1st Qu.:4.000
##  Median : 6.500              Median :4.000
##  Mean   : 5.839              Mean   :4.065
##  3rd Qu.: 8.000              3rd Qu.:5.000
##  Max.   :10.000              Max.   :5.000
```

5 Research Question 1

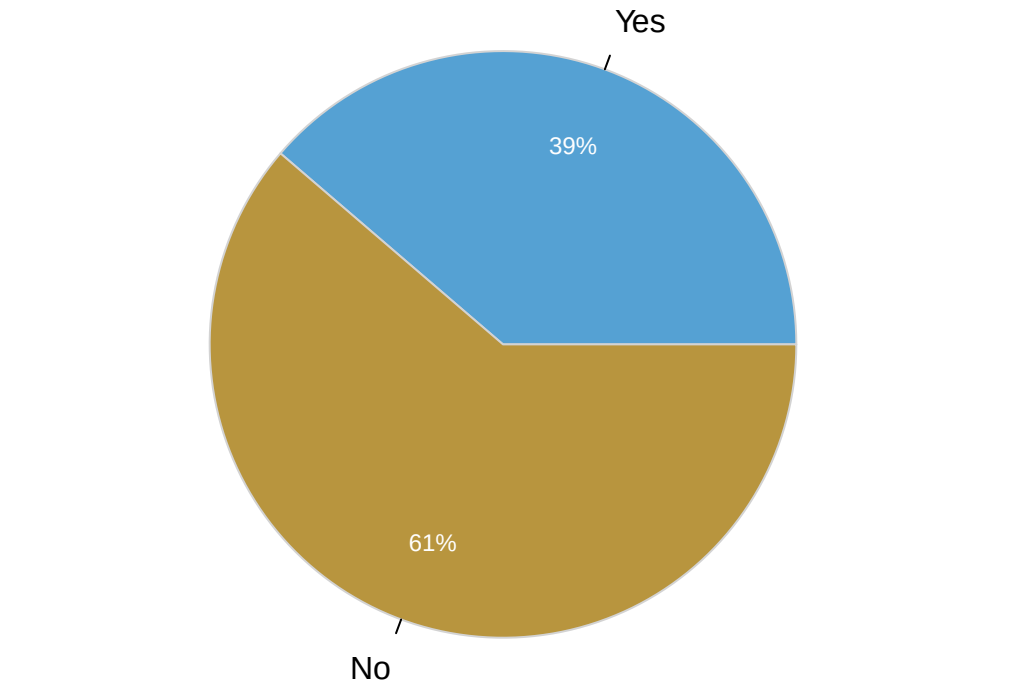
6 Research Question 2

The second research question we wanted to investigate was "How do students in different Master's degrees feel about the measures TU has implemented to reduce the environmental impact?" We wanted to see if there were any differences between the students of different Master's degrees regarding the students's opinion towards the measures TU has implemented to reduce the environmental impact. When looking at the data, we can see that there are not only Master's students in the course, so we will include students from Bachelor programmes as well in our analysis.

6.1 Exploratory Data Analysis

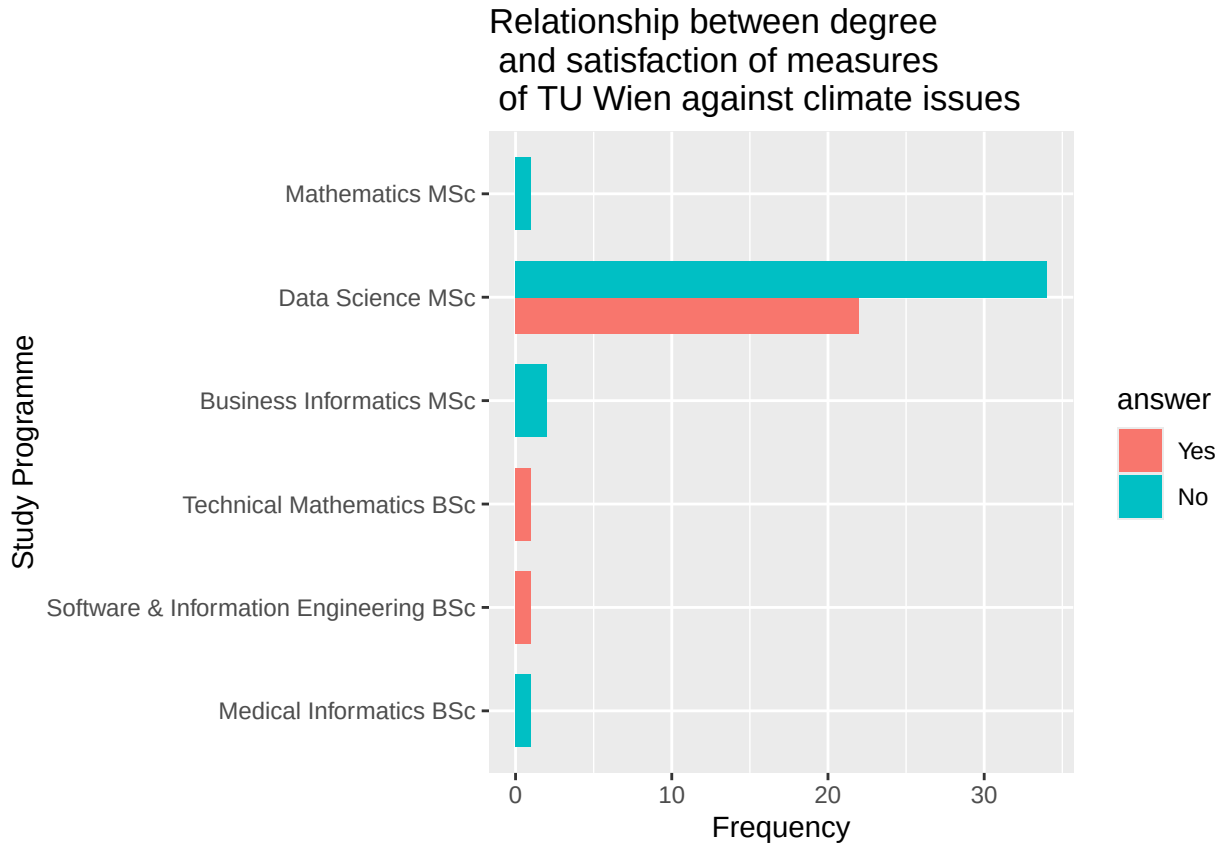
```
q = table(df$question_2)
PieChart(q, hole = 0, main="Do you think TU has already implemented \n enough measures to reduce their e
```

```
## >>> Note: q is not in a data frame (table)
## >>> Note: q is not in a data frame (table)
```




```
## >>> suggestions
## PieChart(q, hole=0)  # traditional pie chart
## PieChart(q, labels="%")  # display %'s on the chart
## PieChart(q)  # bar chart
## Plot(q)  # bubble plot
## Plot(q, labels="count")  # lollipop plot
##
## --- q ---
##
##                  Yes     No    Total
## Frequencies:      24     38       62
## Proportions:   0.387  0.613    1.000
##
## Chi-squared test of null hypothesis of equal probabilities
##   Chisq = 3.161, df = 1, p-value = 0.075
```

```
ggplot(data=df, aes(factor(study_programme), fill=question_2)) +
  geom_bar(stat="count", width=0.7, position=position_dodge()) +
  coord_flip() +
  ggtitle("Relationship between degree \n and satisfaction of measures \n of TU Wien against climate iss
  ylab("Frequency") +
  xlab("Study Programme") +
  labs(fill='answer')
```



If we look at the piechart of the overall answers, we can see that about 40% of the interviewed students think that TU already has implemented enough measures against their climate impact and 60% if the students think that this is not the case.

In the bar plot we can clearly see that the group of Data Science Students is the biggest one among the study programmes. It seems like the students in study programmes like Medical Informatics BSc, Mathematics MSc and Business Informatics MSc rather do not think that TU has implemented enough measures, but since the groups are very small (< 5) we cannot make a statement about any possible correlation.

It also seems like the students of Technical Mathematics BSc and Software & Information Engineering BSc rather agree to the statement, that TU has already implemented enough measures, but again the groups are too small to make a confident statement about any possible correlation between the variables.

6.2 Descriptive Inference

```
df_help = df %>%
  group_by(study_programme) %>%
  count(df$question_2)
r = df_help$n/c(2, 57,57,1,1,1,1)
df_help$r = r
df_help
```

```
## # A tibble: 7 x 4
## # Groups:   study_programme [6]
##   study_programme                         `df$question_2`     n       r
```

```
##   <fct>                                  <fct>           <int>   <dbl>
## 1 Medical Informatics BSc                No                  1  0.5
## 2 Software & Information Engineering BSc Yes                 1  0.0175
## 3 Technical Mathematics BSc              Yes                 1  0.0175
## 4 Business Informatics MSc               No                  2  2
## 5 Data Science MSc                       Yes                22 22
## 6 Data Science MSc                       No                 34 34
## 7 Mathematics MSc                        No                  1  1
```

Again, we can see here that the group of Data Science students makes the biggest portion of the surveyed students, and the other group are very small with only 1 or 2 members, so we cannot make any statement about a possible correlation. We can also see that the distribution of "yes" and "no" answers is very similar for the Data Science group, compared to the overall distribution, meaning that the other groups hardly contribute anything to the distribution.

6.3 Analytic Inference

To test our hypothesis H0: The proportions of "Yes" answers are equal in all groups of students, we will conduct a proportional t-test.

```
SF<-df[df$study_programme=="Data Science MSc",]
SM<-df[df$study_programme=="Mathematics MSc",]
ST=df[df$study_programme=="Business Informatics MSc",]
SA=df[df$study_programme=="Medical Informatics BSc",]
SB=df[df$study_programme=="Software & Information Engineering BSc",]
SC=df[df$study_programme=="Technical Mathematics BSc",]

x1<-sum(SF$question_2=="Yes")
n1<-length(SF$question_2)
x2<-sum(SM$question_2=="Yes")
n2<-length(SM$question_2)
x3=sum(ST$question_2=="Yes")
n3=length(ST$question_2)
x4=sum(SA$question_2=="Yes")
n4=length(SA$question_2)
x5=sum(SB$question_2=="Yes")
n5=length(SB$question_2)
x6=sum(SC$question_2=="Yes")
n6=length(SC$question_2)
prop.test(c(x1,x2,x3,x4,x5,x6),c(n1,n2,n3,n4,n5,n6))
```

```
##
##  6-sample test for equality of proportions without continuity correction
##
## data:  c(x1, x2, x3, x4, x5, x6) out of c(n1, n2, n3, n4, n5, n6)
## X-squared = 5.7008, df = 5, p-value = 0.3364
## alternative hypothesis: two.sided
## sample estimates:
##    prop 1    prop 2    prop 3    prop 4    prop 5    prop 6
## 0.3928571 0.0000000 0.0000000 0.0000000 1.0000000 1.0000000
```

As we can see, the p-value for this test is 0.3384 and therefore not smaller than the significance level 0.05. Hence we cannot reject the null hypothesis. Since some of the groups are very small though, this statistical result should be checked again with a bigger sample.

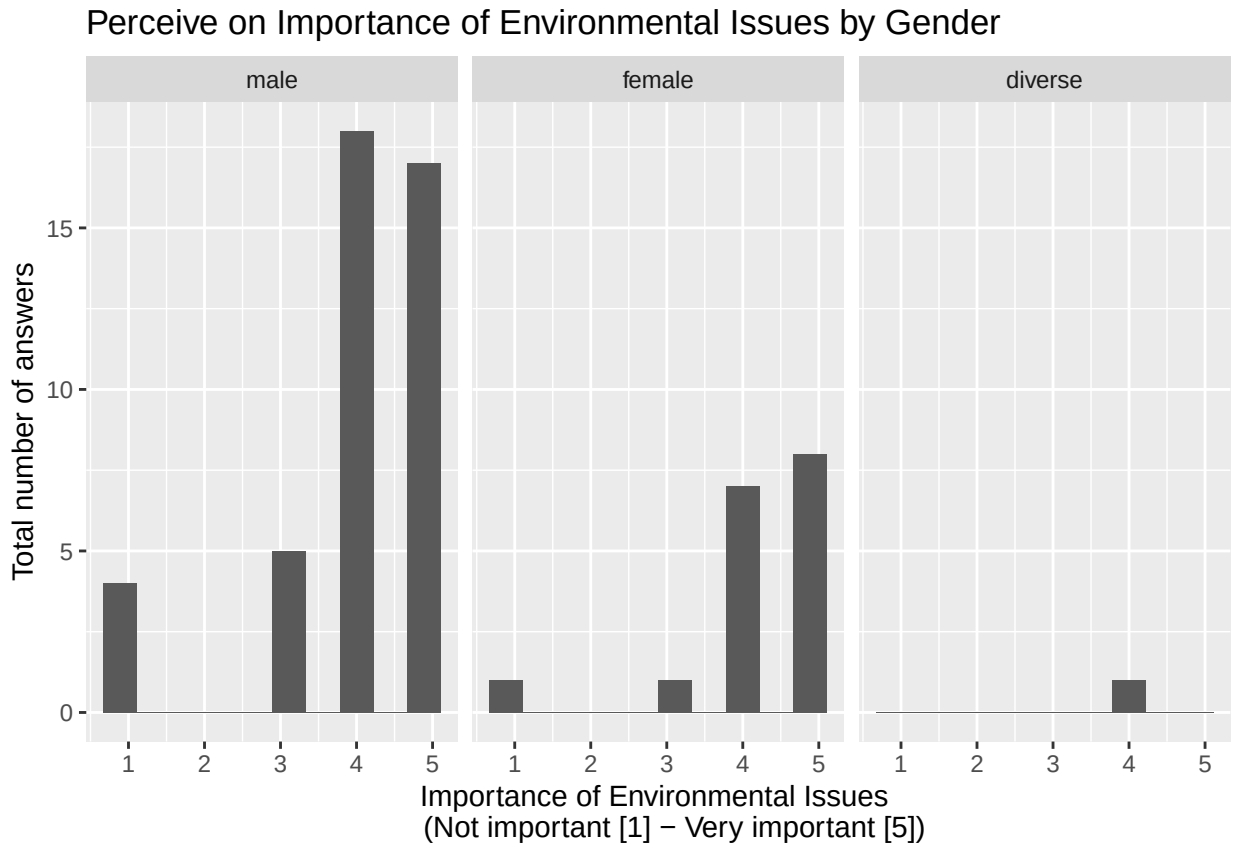
7 Research Question 3

In the third research question we want to know will gender affect the importance students place on environmental issues, and we gather answer on a scale of 1-5, with higher number indicating higher importance the student perceive this issue.

7.1 Exploratory Data Analysis

Since answer to question 3 is collected as quantitative data, we use histogram here to show the count and distributions of the answer, with gender shown in facet.

```
ggplot(df, aes(question_3)) +
  geom_histogram(bins = 10) +
  facet_wrap( ~ gender) +
  labs(
    title = "Perceive on Importance of Environmental Issues by Gender",
    x = "Importance of Environmental Issues
    (Not important [1] - Very important [5])",
    y = "Total number of answers"
  )
```



We can see that there is only 1 answer under gender group "diverse", so no distribution can be found with this category. While for the "male/female" group, the distribution of answer looks indeed are not so identical. Male respondents gave higher answer count in scale 4 than scale 5, compared to female respondents where more people gave answer in scale 5.

7.2 Descriptive Inference

```
# explore the gender data
summary_gender <- df %>% group_by(gender) %>%
  summarise(count = n()) %>% #total number of observations
  mutate(percentage = count/sum(count) * 100)

kable(summary_gender, caption =
        "Summary statistics on the gender of sample group")
```

Table 1: Summary statistics on the gender of sample group

gender	count	percentage
male	44	70.967742
female	17	27.419355
diverse	1	1.612903

```
# explore the answer we got for question 3
summary_q3 <- df %>% group_by(gender) %>%
  summarise(
    count = n(),
    mean = mean(question_3),
    median = median(question_3),
    sd = sd(question_3)
  )

kable(summary_q3, caption =
        "Summary statistics on the answer to importance of environmental issues by gender")
```

Table 2: Summary statistics on the answer to importance of environmental issues by gender

gender	count	mean	median	sd
male	44	4.000000	4	1.161395
female	17	4.235294	4	1.032558
diverse	1	4.000000	4	NA

In summary, in the total 62 sample we collected, there are 71% male respondent, 27% female respondent and 2% non-binary respondent.

Besides the only 1 non-binary respondent who gave answer in a scale of 4, both the answers from male and female respondent show a median value of 4, while mean from female group (4.2) slightly higher than the male group (4).

This indicates in general, our students in this survey have a tendency of perceive the environmental issue as "Important".

The difference from basic statistics and different sample size for each group can't tell us whether there is a difference on the answer among different gender groups clearly.

Therefore, We will further test our hypothesis in the next section.

Before that, since there is only 1 non-binary respondent and no meaningful inference can be made from this, I will exclude it from the dataset.

```
# exclude one non-binary answer
df_q3 <- df %>% filter(gender != "diverse")

#drop the level for the factorised column
df_q3$gender <- droplevels(df_q3$gender)
```

7.3 Analytic Inference

Now we made a null hypothesis (H0):

There is no difference between gender and student's perceive of importance on environmental issue.

We then test this hypothesis with our sample, since the gender is categorical, answers to questions 3 is ordinal/numerical, we adpoted a Kruskal-Qallis test on the sample.

```
kruskal_test_q3 <- kruskal.test(question_3 ~ gender, data = df_q3)

print(kruskal_test_q3)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  question_3 by gender
## Kruskal-Wallis chi-squared = 0.59508, df = 1, p-value = 0.4405
```

The result show us p-value for the hypothesis is 0.4405, which is not significant.

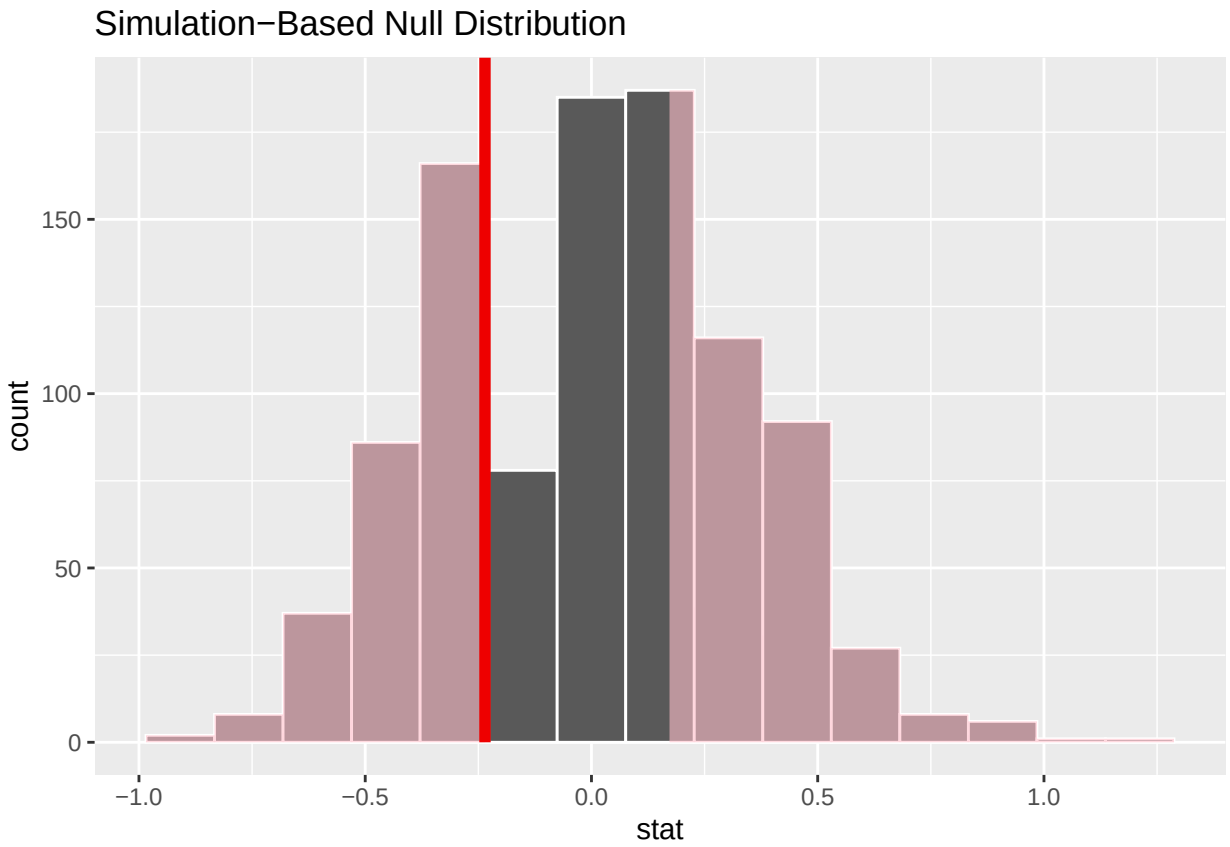
We can see that gender will not affect the importance students place on environmental issues.

Furthermore, since propably not all students in this class have responded to the survey questionaire, by using *infer* package, we can re-sample our dataset and test the hypothesis using non-parametric permutation method, to make the result more reliable.

```
# specify the hypothesis
hypothesis_q3 <- df_q3 %>%
  # specify answer to q3 as response, gender as explanatory variable
  specify(question_3 ~ gender) %>%
  # set the hypothesis as independent to question
  hypothesise(null = "independence") %>%
  # permute 1000 times
  generate(reps = 1000, type = "permute") %>%
  # calculate the diff in means
  calculate(stat = "diff in means", order = c("male", "female"))

# calculate the observed statistics
obs_q3 <- df_q3 %>%
  specify(question_3 ~ gender) %>%
  calculate(stat = "diff in means", order = c("male", "female"))

# visualise the null distribution and observed statistic
visualize(hypothesis_q3) +
  shade_p_value(obs_q3, direction = "two-sided")
```




```
# get p_value
p_value_q3 <- hypothesis_q3 %>%
  get_p_value(obs_q3, direction = "two_sided")

print(p_value_q3)
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.598
```

The visualisation shows the simultion-based distribution of null hypothesis for the difference in means between "male/female" regarding the importance we placed on environmental issues. Each bar in the histogram represents the count of times a particular value of the test statistic was obtained from the permuted samples.

The distribution shows us the differences in means we might expect to see from the answer of question 3 is mostly ranged between [-0.5, 1], while the red vertical line showing our observed test statistics is not out positioned in unusual range, suggests that the observed differenece is likely to occur by a random chance.

The p-value from permutation is 0.56, which is the proportion of permuted test statistics as extreme or more extreme than the observed statistics (original data). A value of 0.56 suggests that we can not reject the null hypothesis, and there is no significant difference observed between gender and importance on environmental issues.

8 Conclusion

- For the third research question, we found gender will not affect the importance students place on environmental issues.