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
```

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"
  }
}
```

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

```
# factorise column
df$gender <- factor(df$gender, levels = c("male","female","diverse"))
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"))

#summary dataset
summary(df)
```

```
##      gender        age                                        study_programme
##  male   :44   Min.   :22.00   Medical Informatics BSc               : 1
##  female :17   1st Qu.:24.00   Software & Information Engineering BSc: 1
##  diverse: 1   Median :25.50   Technical Mathematics BSc             : 1
##               Mean   :27.45   Business Informatics MSc              : 2
##               3rd Qu.:28.00   Data Science MSc                      :56
##               Max.   :57.00   Mathematics MSc                       : 1
##    question_1      question_2          question_3
##  Min.   : 1.000   Length:62          Min.   :1.000
##  1st Qu.: 4.000   Class :character   1st Qu.:4.000
##  Median : 6.500   Mode  :character   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

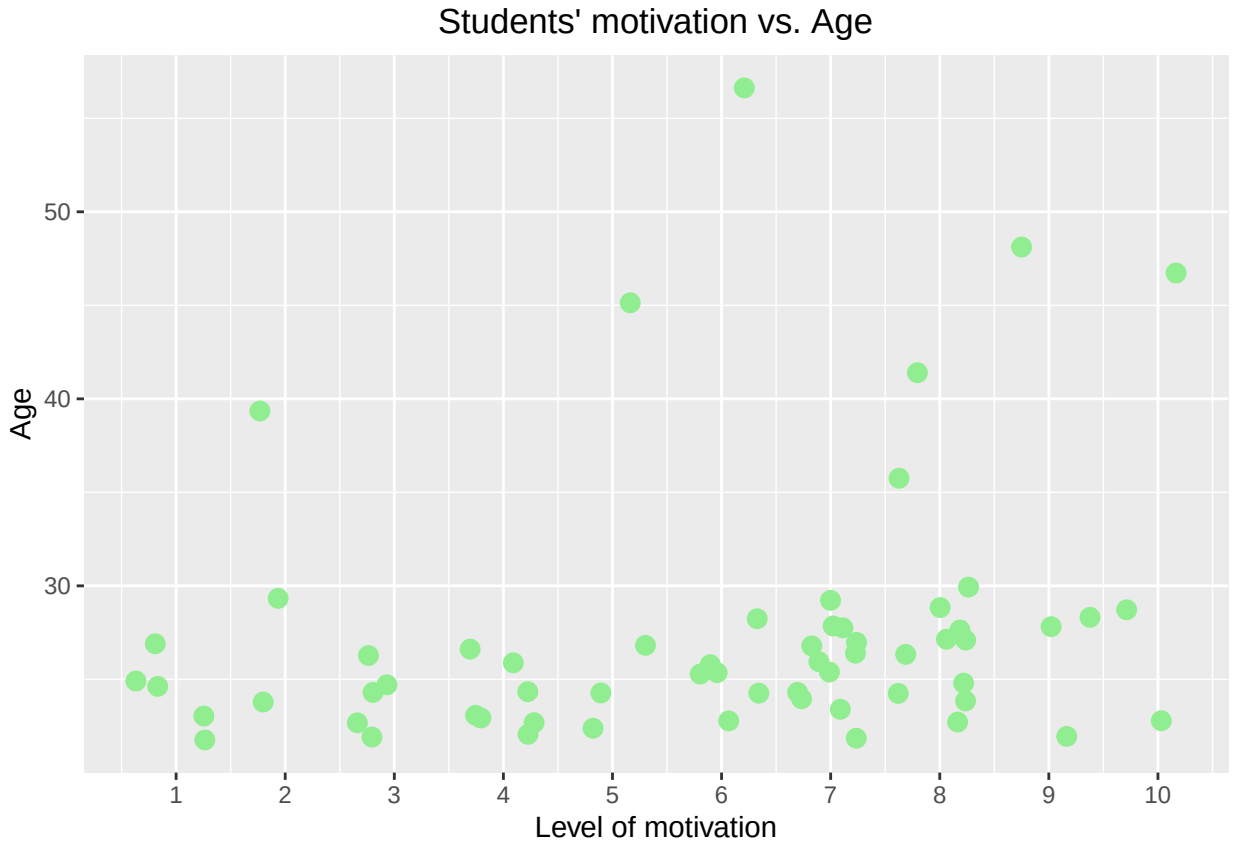
The first research question says: “Is there a correlation between the age and the motivation of students to reduce their carbon footprint?”. We wanted to find out if there is a relationship between students’ age and their motivation to reduce their carbon footprint. For answering this question we used variables “age” and “the”question_1”.

5.1 Exploratory Data Analysis

We can consider both variables as continuous variables, therefore as first step in exploratory analysis we created a jitter plot. We can see a positive correlation between these two variables, anyway, the relationship isn’t very strong.

```
ggplot(df, aes(x = question_1, y = age)) +
  geom_jitter(color = "light green", size = 3) +
  labs(title = "Students' motivation vs. Age",
       x = "Level of motivation",
```

```
       y = "Age") +
theme(plot.title = element_text(hjust=0.5)) +
  scale_x_continuous(breaks = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10),
                     labels = c("1", "2", "3", "4", "5", "6", "7", "8", "9", "10"))
```



In the next part we will compute a correlation coefficient. To find out what kind of correlation coefficient to use, we had to inspect the normality of variables. One possibility could be to look at the histograms, but we focused more on the Shapiro-Wilk test of normality. For both variables we got a very small p-value, therefore we rejected the null hypothesis -> they aren't normally distributed.

```
shapiro.test(df$age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$age
## W = 0.66704, p-value = 0.000000000139
```

```
shapiro.test(df$question_1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$question_1
## W = 0.93531, p-value = 0.002783
```

5.2 Descriptive Inference

Both variables aren't normally distributed, so we computed Spearman correlation coefficient. The resulting number is 0.3204. The value of the coefficient is higher than we expected, but we cannot yet say there is a relationship. At first we have to check it with a statistical test.

```
cor(df$age, df$question_1, method = "spearman")
```

```
## [1] 0.3204118
```

5.3 Analytic Inference

For testing our hypothesis that correlation coefficient isn't zero, we used Spearman's variation of correlation test. The resulting p-value is smaller than 0.05 and we can finally conclude there is a relationship between students' age and their motivation to reduce their carbon footprint.

```
cor.test(df$age, df$question_1, method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  df$age and df$question_1
## S = 26987, p-value = 0.01112
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.3204118
```

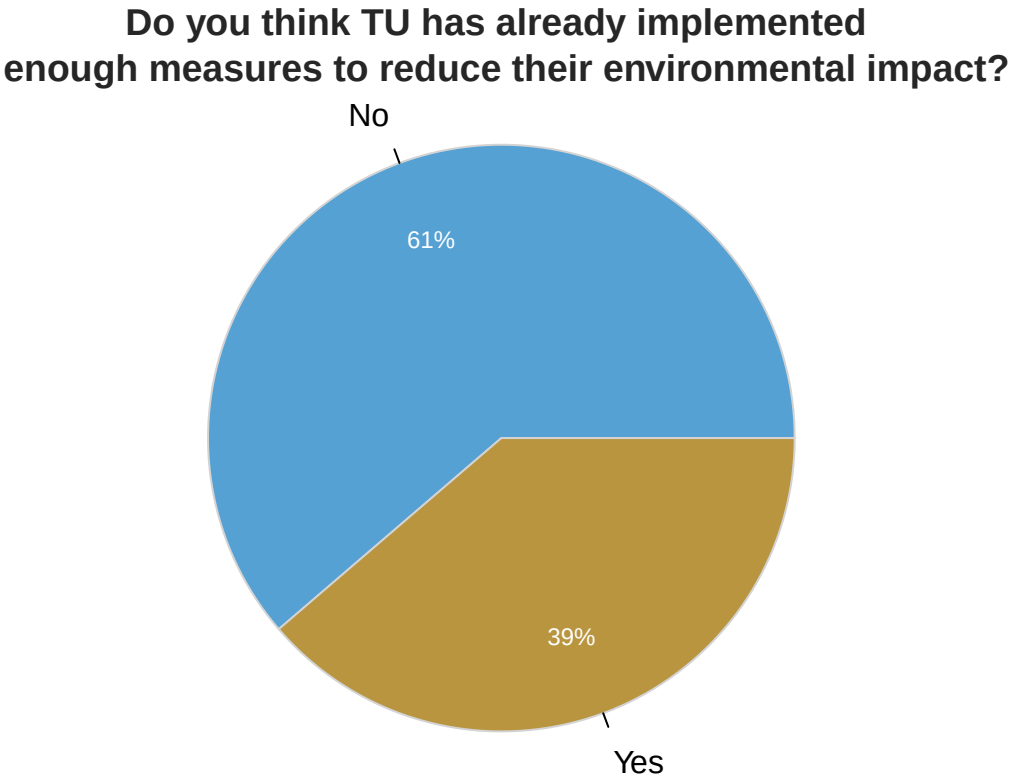
6 Research Question 2

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

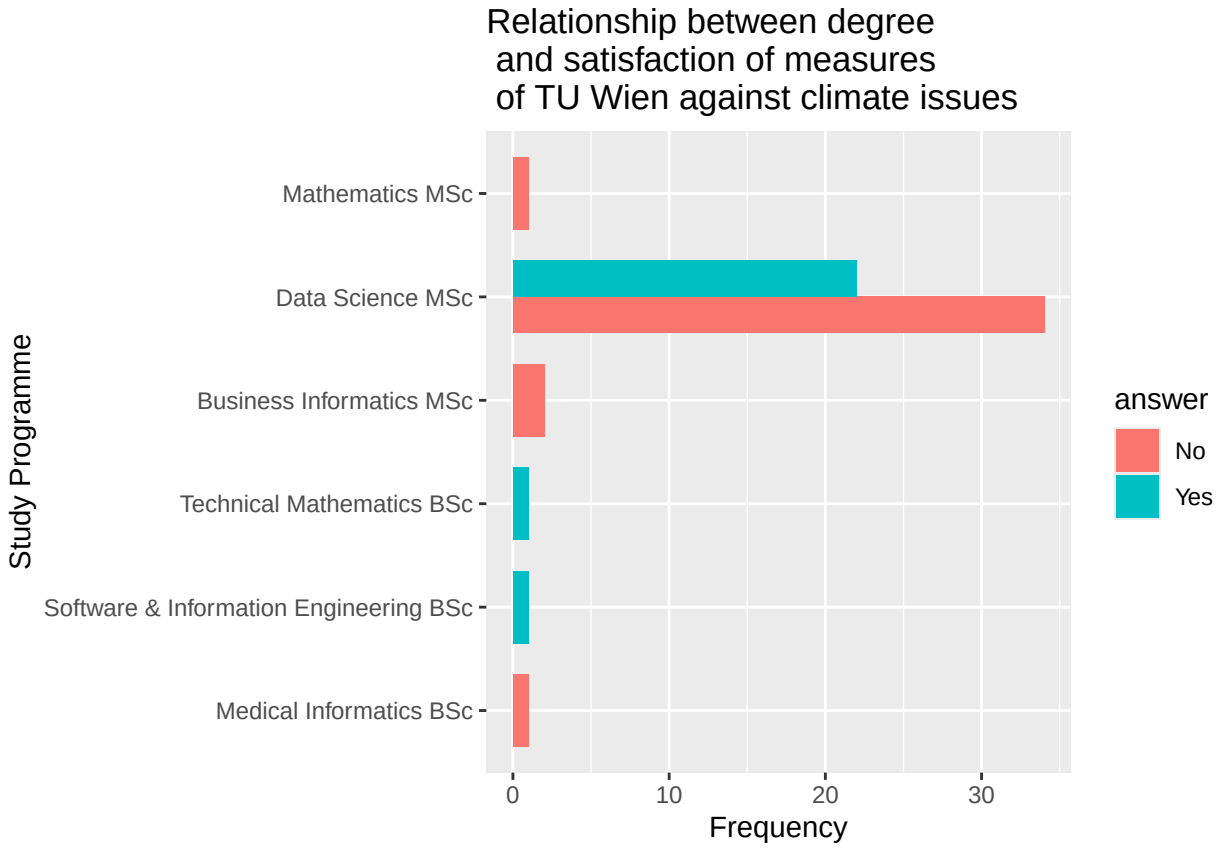
```
q = table(df$question_2)
PieChart(q, hole = 0, main="Do you think TU has already implemented \n enough measures to reduce their
```

```
## >>> 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 ---
##
##                   No    Yes    Total
## Frequencies:      38     24       62
## Proportions:   0.613  0.387    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=as.factor(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 programs. So one thing we can deduct from this first look at the data is that the groups other than Data Science are too small to make any statistically relevant statement. To remedy this, we are going to pool all the small groups together and compare them to the group of Data Science students.

6.2 Descriptive Inference

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

```
## # A tibble: 7 x 4
## # Groups:   study_programme [6]
##   study_programme                          `df$question_2`     n     r
##   <fct>                                    <chr>           <int> <dbl>
## 1 Medical Informatics BSc                  No                  1 1
## 2 Software & Information Engineering BSc   Yes                 1 1
## 3 Technical Mathematics BSc                Yes                 1 1
## 4 Business Informatics MSc                 No                  2 1
## 5 Data Science MSc                         No                 34 0.596
```

```
## 6 Data Science MSc                        Yes                22 0.386
## 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. As mentioned before, to remedy this for the statistical test, the other groups will be pooled together to get more meaningful results.

6.3 Analytic Inference

To test our hypothesis H0: The frequency of "Yes" answers are equal in all groups of students meaning group "Data Science" and group "Others", we will conduct a Chi-Square test.

```
helpy = ifelse(df$study_programme == "Data Science MSc", 1, 0)
table(helpy, df$question_2)
```

```
##
## helpy No Yes
##     0  4   2
##     1 34  22
```

```
chisq.test(helpy, df$question_2)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  helpy and df$question_2
## X-squared = 0.00000000000000000000000000040458, df = 1, p-value = 1
```

```
fisher.test(helpy, df$question_2)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  helpy and df$question_2
## p-value = 1
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   0.1679467 15.3814338
## sample estimates:
## odds ratio
##   1.288887
```

As we can see, the p-value for this test is 1 and therefore not smaller than the significance level 0.05. To further test the hypothesis, we also conducted a fisher test, which led to the same result. Hence we cannot reject the null hypothesis, so we cannot prove that students from different Master's programs have a different opinion on the topic of measures against environmental impact that have already been implemented by TU Wien.

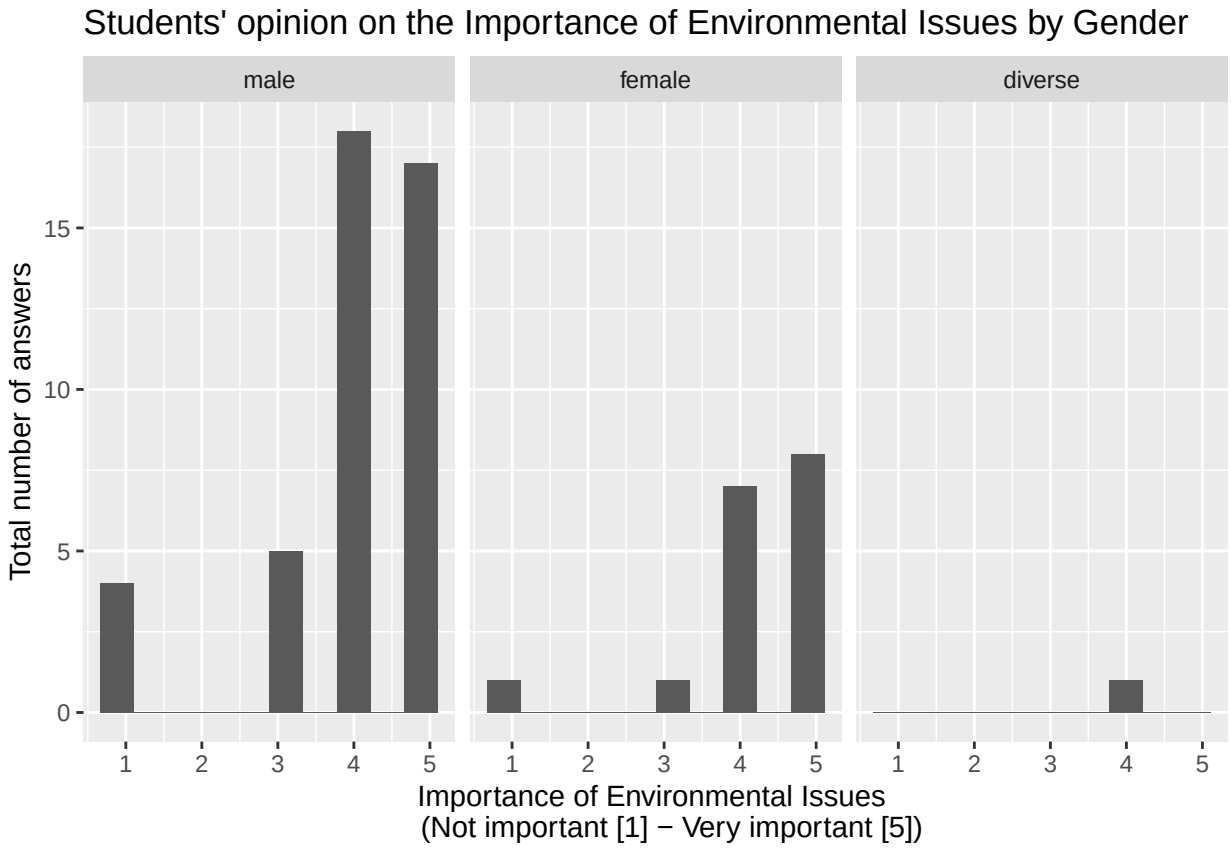
7 Research Question 3

In the third research question we want to know: Will gender affect the importance students place on environmental issues? And we gather answers on a scale of 1-5, with a higher number indicating higher importance the students 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 = "Students' opinion on the 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 students' opinion on the importance of environmental issues.

We then test this hypothesis with our sample, since the gender is categorical, we used two test method to the sample:

- Kruskal- Wallis test, where we treat answer as ordinal data, and in the result we can compare medians of the groups, to see if the medians of male/female for this question is different.
- Chi-Square test, where we treat answer as categorical data (1,2,3,4,5), and in the result we can compare the independence between the answers and the factor (gender).

```
# kruskal wallis test
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
```

```
# chi-square test
# change answer to categorical first
df_q3$question_3_cate <- as.factor(df_q3$question_3)

# create a contingency table
contingency_table_q3 <- table(df_q3$gender, df_q3$question_3_cate)

# perform chi-square test
chi_square_test_q3 <- chisq.test(contingency_table_q3)

print(chi_square_test_q3)
```

```
##
##  Pearson's Chi-squared test
##
## data:  contingency_table_q3
## X-squared = 0.74102, df = 3, p-value = 0.8635
```

The result show us:

- p-value from Kruskal-Wallis test for the hypothesis is 0.4405, which is not significant.Showing that the median answers on the importance of environmental issues does not differ among male/female respondent groups.
- p-value from Chi-Square test for the hypothesis is 0.8635, which is also not significant. Showing that the answer of question 3 is not dependent on the gender.

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

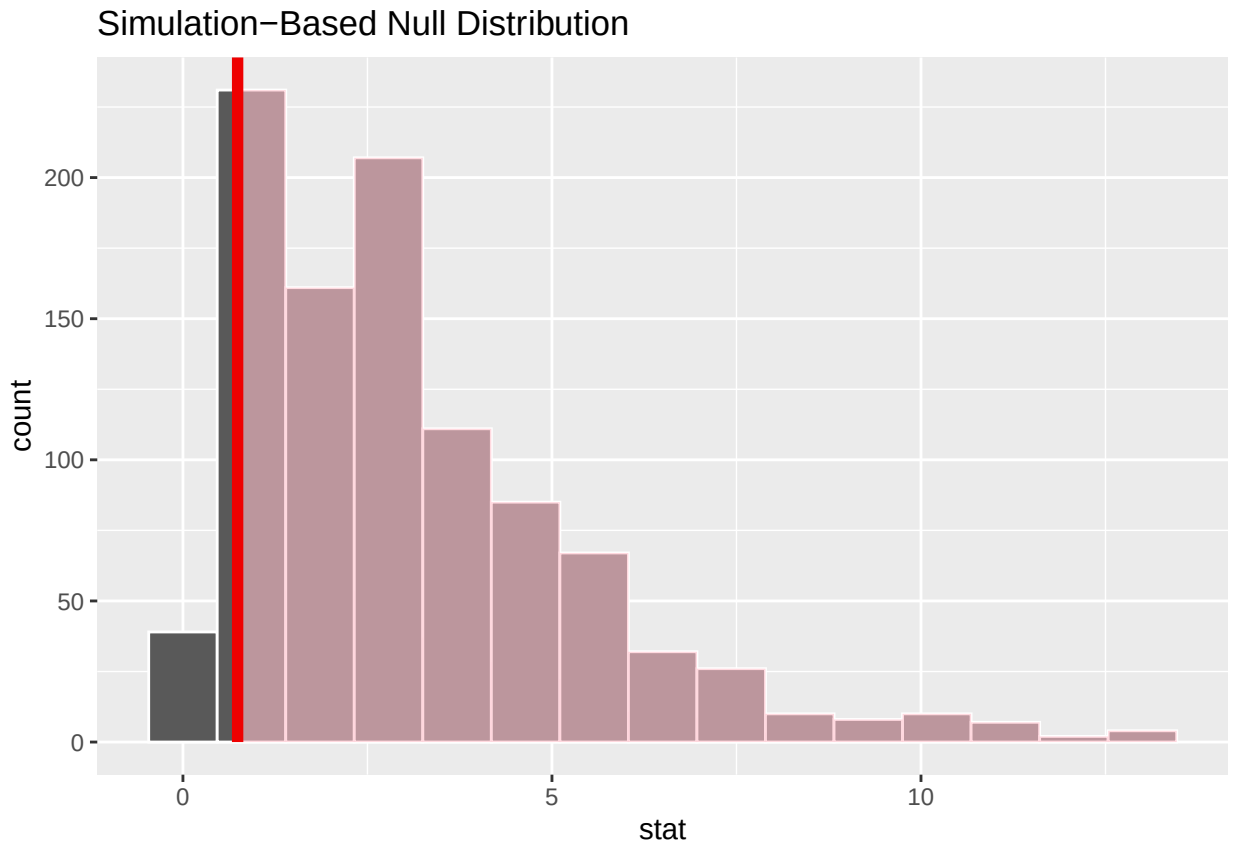
Furthermore, since propably not all students in this class have responded to the survey questionaire, by using *infer* package, we can re-sample our dataset and test the hypothesis using non-parametric permutation method, to make the result more reliable.

```
# specify the hypothesis
hypothesis_q3 <- df_q3 %>%
```

```r
  # specify answer to q3 as response, gender as explanatory variable
  specify(question_3_cate ~ 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 = "Chisq")

# calculate the observed statistics
obs_q3 <- df_q3 %>%
  specify(question_3_cate ~ gender) %>%
  calculate(stat = "Chisq")

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




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

print(p_value_q3)
```

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

```
## 1   0.857
```

The visualisation shows the simultion-based distribution of null hypothesis for the chi-square test 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 p-value from permutation is 0.882, which is the proportion of permuted test statistics as extreme or more extreme than the observed statistics (original data), and we can also see on the graph that most permuted tests are larger than our observed statistics. The observed chi-square statistic (red line) lies well within the central part of the null distribution.

Therefore, both p_value and graph suggests that there is no strong evidence against the null hypothesis of independence between gender and the importance students placed on the environmental issues.

8 Conclusion

- The older the students are, the more motivation they have to reduce their carbon footprint.
- The Master's programme of students does not have an effect on the opinion of students on the measures TU has implemented to reduce their environmental impact.
- Gender does not affect students' opinions on the importance of environmental issues.