Research on Factors Influencing Statistics Students' Engagement in Extracurricular Activities

Affan Rana, Shawn Le, Fred Nguyen, Victoria Tran, Danny Le

May 7, 2025

STA304H5: Surveys, Sampling and Observational Data

Professor: Luai Al Labadi

Teaching Assistant: Anna Ly

Group 5: Halal Vietnamese Food

Contents

1 Abstract

Extracurricular activities in university often present new opportunities for students, foster a sense of community, and enhance the school's dynamic. However, not every student has the chance to engage in such activities due to various factors that may be associated with participation rates. To understand which factors are related to students' involvement, several variables were measured to identify which ones may be significant.

This study aims to explore whether factors such as study time, work commitments, travel time, personal preferences, and the supportiveness of friends and family are related to student participation in extracurricular activities. In a third-year statistics survey course (STA304) at the University of Toronto Mississauga, we asked students about their perceptions of participating in extracurricular activities, as well as information related to our hypotheses regarding these factors.

The results show that approximately 76% of students participate in at least one extracurricular activity, with academic-related activities being the most popular choice. The most common reasons for participating in extracurricular activities are stress relieve and such findings help us dive further into other aspects regarded. Based on these findings, we aim to investigate the factors influencing extracurricular participation and explore future related projects to provide foundational resources to suggest to academic institution where not only benefiting students of STA304 but also a broader population within the University of Toronto Mississauga.

2 Introduction

As students ourselves, we know firsthand how extracurricular activities can significantly shape our personal and educational experiences. These activities—ranging from sports and music to clubs and community service—provide opportunities for personal growth, social connections, and the development of valuable skills outside the traditional classroom setting. This study seeks to explore the various factors that may help explain a student's decision to participate in extracurricular activities.

This study aims to analyze the factors that are potentially associated with student participation in extracurricular activities at the University of Toronto Mississauga. In October 2024, our team deployed an online survey via Google Forms and posted it via an announcement on Quercus to collect data on student participation, student preferences, demographic factors, workload, and their thoughts on extracurricular activities, as well as the supportiveness of close individuals regarding their choices. In addition, we performed simple random sampling by randomly select a number of student's survey responses that were collected as part of the entire dataset. The exact sample size will be calculated in the analysis section of this report. For this specific study, we aim to examine the following research questions:

2.1 Research Questions

1. (RQ1) **Do personal preferences influence participation in extracurricular activities?**

   - *Null hypothesis:* Personal preferences have no association with student's extracurricular activities participation.
   - *Alternative hypothesis:* Personal preferences are correlated with certain extracurricular activities.

2. (RQ2) **To what extent does schoolwork and/or professional commitments (such as study time, off-campus work, internships, commuting, etc.) in-**

fluence a student's participation in extracurricular activities?

   - *Null hypothesis:* Schoolwork and/or professional commitments are not associated with student's participation in extracurricular activities.
   - *Alternative hypothesis:* Schoolwork and/or professional commitments are correlated with student's participation in extracurricular activities.

3. (RQ3) **How does the importance of having extracurricular activities among domestic and international students influence their choice of participating?** (Thematic analysis)

   - *Null hypothesis:* There is no association between a student's opinion on the importance of extracurricular activities and their participation among domestic and international students.
   - *Alternative hypothesis:* A student's opinion on the importance of extracurricular activities is correlated with their choice of participation among domestic and international students.

4. (RQ4) **Do the opinions of friends and/or family influence students' views on participating in extracurricular activities?**

   - *Null hypothesis:* The opinions of friends and/or family have an association with students' participation in extracurricular activities.
   - *Alternative hypothesis:* The opinions of friends and/or family are not correlated with students' participation in extracurricular activities.

2.2 Paper Structure

The structure of the paper is as follows: Section 2 outlines our data collection methods, including our sampling approach and survey design. Section 3 presents our quantitative analysis, showcasing relevant statistics and graphical visualizations. Section 4 discusses our results, addressing each research question and interpreting our findings. Section 5 covers the limitations of our study, while Section 6 concludes our analysis and offers suggestions for future research. Finally, an appendix containing our R code and additional data follows.

3 Methodology

In October 2024, a Google Form survey was launched through an announcement on the Quercus website for STA304 students in the Fall 2024 semester at the University of Toronto Mississauga. The survey link was also pinned on the course Piazza platform. The survey remained open from October 1st to October 20th, 2024. Our target population included approximately 240 students, with around 120 students enrolled in the two sections, LEC0101 and LEC0102, representing both domestic and international students.

To minimize selection bias and gather sufficient responses, we shared the survey link in person with students in both LEC0101 and LEC0102 lectures, offering participation through a QR code during the Thursday (LEC0102) and Friday (LEC0101) sessions. This provided an option for students to participate voluntarily. Additionally, the survey link was prominently placed on the STA304 course Piazza page to increase accessibility, emphasizing voluntary participation. Once all answers are collected, we will randomly select n participants by using simple random sampling to create a subset dataset of the original dataset based on the sample size calculation for each analysis test.

The survey included 11 questions. These questions captured students' course information for grading purposes, as well as closed-ended and open-ended questions focused on topics such as time commitments, types of extracurricular activities, personal motivations, and whether or not the students perceived support from family and friends.

4 Results

4.1 Simple Test: Two-Sample Test for Proportion

The study estimates population statistics with a bound on the error equal to 0.05 ($B = 0.05$). The population consists of 121 students who participated in the survey.

To answer **RQ2**, we perform two Two-Sample Tests for Proportion to compare the difference in proportion of the following demographics:

- Students who participate in any extracurricular activities and whether or not they are working.
- Students who participate in any extracurricular activities and whether or not they spend a lot of time studying (with a threshold of 4 hours of study per day to qualify as spending a lot of time).

4.1.1 Two-Sample Test 1

The sample size calculation for the population proportion p with simple random sampling is given by:

$$n = \frac{Npq}{(N-1)D + pq}$$

where:

- p: proportion of students who participate in extracurricular activities and are working, we assume p = 0.5
- q: proportion of students who participate in extracurricular activities and are not working $= 1 - p$
- N: total number of students who participated in the survey $= 121$
- $D = \frac{B^2}{4} = \frac{(0.05)^2}{4}$

Thus, we sampled 94 students for the first proportion test to satisfy the following assumptions and ensure result reliability:

- The data was randomly sampled, and the responses from students are independent of each other.
- The response variable (**activities**) is binary (yes or no) to ensure the population follows a binomial distribution.

4.1.2 Hypotheses for the First Proportion Test

Let:

- p_1: Proportion of students who attend extracurricular activities and are not working.
- p_2: Proportion of students who attend extracurricular activities and are working.

$$\begin{aligned} \text{Null Hypothesis:} \quad & H_0 : p_1 - p_2 = 0 \\ \text{Alternative Hypothesis:} \quad & H_A : p_1 - p_2 \neq 0 \end{aligned}$$

The intuition behind this test is that if the p-value of this test is less than $\alpha = 0.05$ (95% confidence level), then it is safe to reject the null hypothesis and conclude that working does affect extracurricular attendance.

4.1.3 Result of The First Proportion Test

estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high
0.7619048	0.6438356	0.5637584	0.4527499	1	-0.1253124	0.3614507

method	alternative
2-sample test for equality of proportions with continuity correction	two.sided

Table 1: Results of the first two-sample test for equality of proportions.

The test yields a p-value larger than $\alpha = 0.05$ (95% confidence level). Hence, the null hypothesis is not concluded, and there is no significant evidence to prove that working status affects the attendance of extracurricular activities.

4.1.4 Two-Sample Test 2

We used the same sample size as above and sampled 94 students for the second proportion test to satisfy the following assumptions and ensure result reliability:

- The data was randomly sampled, and the responses from students are independent of each other.
- The response variable (**activities**) is binary (yes or no) to ensure the population follows a binomial distribution.

4.1.5 Hypotheses for the Second Proportion Test

Let:

- p_1: Proportion of students who attend extracurricular activities and study less than 4 hours a day.
- p_2: Proportion of students who attend extracurricular activities and study more than 4 hours a day.

$$\text{Null Hypothesis:} \quad H_0 : p_1 - p_2 = 0$$
$$\text{Alternative Hypothesis:} \quad H_A : p_1 - p_2 \neq 0$$

The intuition behind this test is that if the p-value of this test is less than $\alpha = 0.05$ (95% confidence level), then it is reasonable to reject the null hypothesis and conclude that hours spent on study affect extracurricular attendance.

4.1.6 Contingency Table Summary

	Study less than 4 hours	**Study at least 4 hours**
Do not participate in any extracurricular activities	17 students	6 students
Participate in any extracurricular activities	48 students	23 students

4.1.7 The Result of The Second Proportion Test:

estimate1	**estimate2**	**statistic**	**p.value**	**parameter**	**conf.low**	**conf.high**
0.7391304	0.6760563	0.0957649	0.7569724	1	-0.1755964	0.3017446

method	**alternative**
2-sample test for equality of proportions with continuity correction	two.sided

Table 2: Results of the second two-sample test for equality of proportions.

The test yields a p-value larger than 0.05 (95% confidence level). Hence, the null hypothesis is not concluded, and there is no significant evidence to prove that study hours affect the attendance of extracurricular activities.

4.2 Simple Test: Testing for Independence

The relationship between personal preferences and participation in extracurricular activities can be best examined using the Chi-Square test for Independence. This test is most effective when used to assess whether there is an association between two categorical variables.

The two categorical variables of choice are **activities whether a student participates in any extracurricular activity or not** and their **favorite activity**. For convenience, only the first response of the **activities** variable is retained. For example, if a student participates in more than one activity, such as “Sport and Fitness” (SF) and “Arts and Culture” (AC), only the first response (SF) is kept.

For simplicity, we use the sample data from the initial proportion test since the responses collected are independent.

4.2.1 Hypothesis

Let:

- A denote activities (type of activities that a student is participating in),
- B denote favorite activities (student preferences).

$$\text{Null Hypothesis:} \quad A \perp B$$
$$\text{Alternative Hypothesis:} \quad A \not\perp B$$

The intuition behind this test is that if the p-value of this test is less than or equal to 0.05 (95% confidence level), then it is reasonable to reject the null hypothesis and conclude that there is a correlation between the type of activities attended and activity preference.

4.2.2 Result of the Chi-Square Test

statistic	**p.value**	**parameter**	**method**
83.02536	0.00000000004719795	16	Pearson's Chi-squared test

Table 3: Results of Chi-squared test.

Since the p-value is less than $\alpha = 0.05$ (95% confidence level), we reject the Null Hypothesis and conclude that there is a relationship between **activities** and **favourite_ activities**.

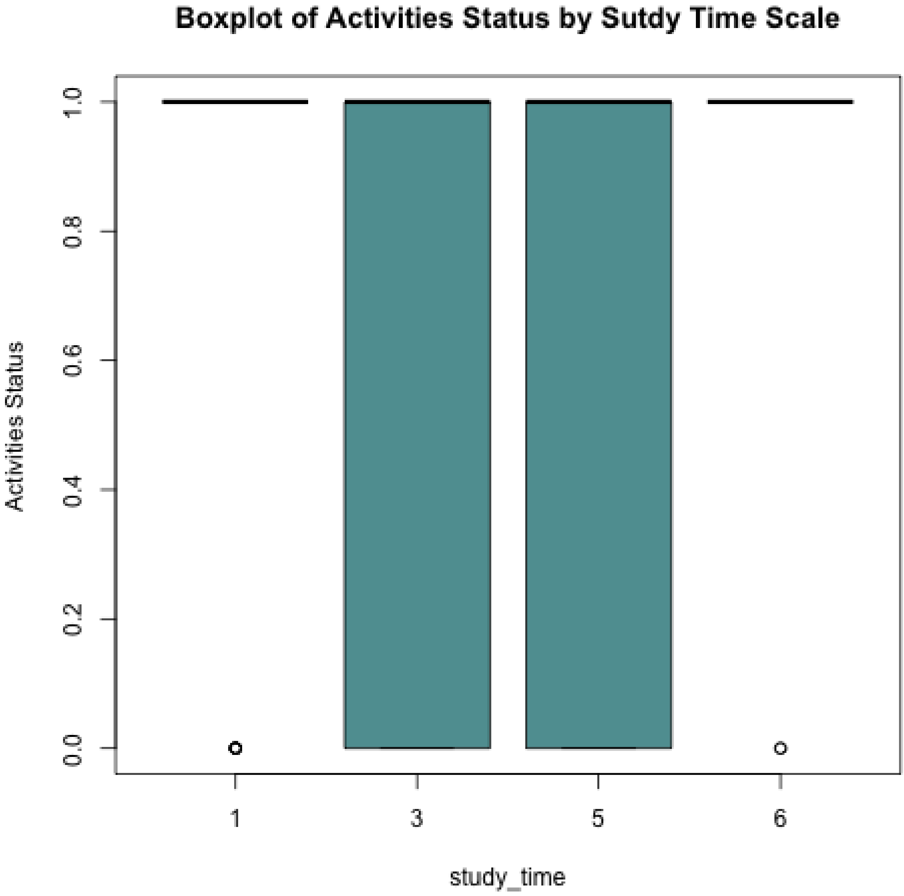
However, it is important to note that the expected values between the two categorical variables do not strictly equal 5. Therefore, the result of the test may not be as reliable as it would be if this assumption were satisfied.

Activity versus Favorite Activity	AA	AC	CV	N	SF
A	1.27659574	1.91489362	0.85106383	4.8936170	11.0638298
D	0.31914894	0.47872340	0.21276596	1.2234043	2.7659574
M	1.40425532	2.10638298	0.93617021	5.3829787	12.1702128
P	0.06382979	0.09574468	0.04255319	0.2446809	0.5531915
S	2.93617021	4.40425532	1.95744681	11.2553191	25.4468085

Table 4: Activity vs Favorite Activity Table.

4.3 Advanced Test: Logistic Regression

During the proposal phase of the study, we proposed using ANOVA to examine the relationship between attendants in extracurricular activities versus different potential predictors. However, the collected response variable is binary. This binary nature of the response variable make it an inappropriate data type to use ANOVA.

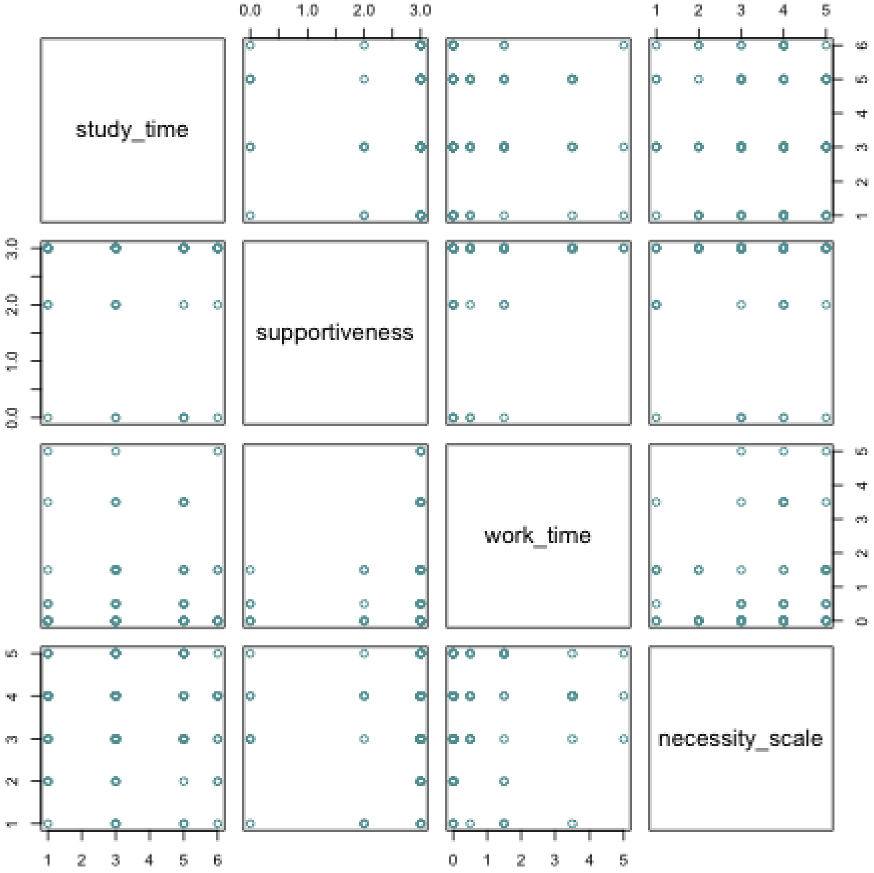


This can be confirm by plotting a simple box plot. The plot above shows that the variable response is not continuous as it only takes 1 and 0 as student's participation status across different level of study time. In addition, we can observe the skewness in distribution of data from the dot presentation of students whom spent more than 6 hours (study-time = 6) and between 0 and 2 hours (study-time = 1) to study . This means that most students spend between 2 to 6 hours to study (refer to codebook for proper description of the dataset).

Hence, Logistic Regression is best suited in this case. Prior to implementing a Logistic Regression predicting model we first need to ensure there is no multicollinearity between the potential predictor variables. In the following section we will explore the relationship between **activities_status** and each of **necessity_scale** and **supportiveness**, as well as both predictors together.

4.3.1 No Multicollinearity Assumption (Backed by Two Methods)

Scatterplot Matrix of Predictor Variables



The scatterplot matrix shows no sign of collinearity between predictor variables as data points are evenly spread out with no patterns.

4.3.2 Variance Inflation Factors (VIF) for Predictors

names	x
study_time	1.022866
supportiveness	1.014622
work_time	1.023584
necessity_scale	1.017183

Table 5: Variance Inflation Factors (VIF) for Predictors.

The VIF of all predictors are almost 1 suggesting complete no multicollinearity between potential predictor variables..

To assess the relationship between extracurricular attendance and student perception of the importance of extracurricular activities, we fit a simple logistic regression model with **activities_status**, student's extracurricular participation (Yes or No) as the response variable and **necessity_scale**, student's perception on the importance of participating in extracurricular as the predictor variable. It is important to note that, instead of fitting the model on a random sample, we applied the regression to the entire dataset. This approach does not require a smaller sample, and including more observations could improve the model's accuracy in predicting the response variable, despite the potential risk of over fitting.

4.4 Results Summaries

4.4.1 Summary of the Logistic Regression Model 1

term	estimate	std.error	statistic	p.value
(Intercept)	-1.461377	0.6960697	-2.099469	0.0357755497
necessity_scale	0.771211	0.2043068	3.774769	0.0001601558

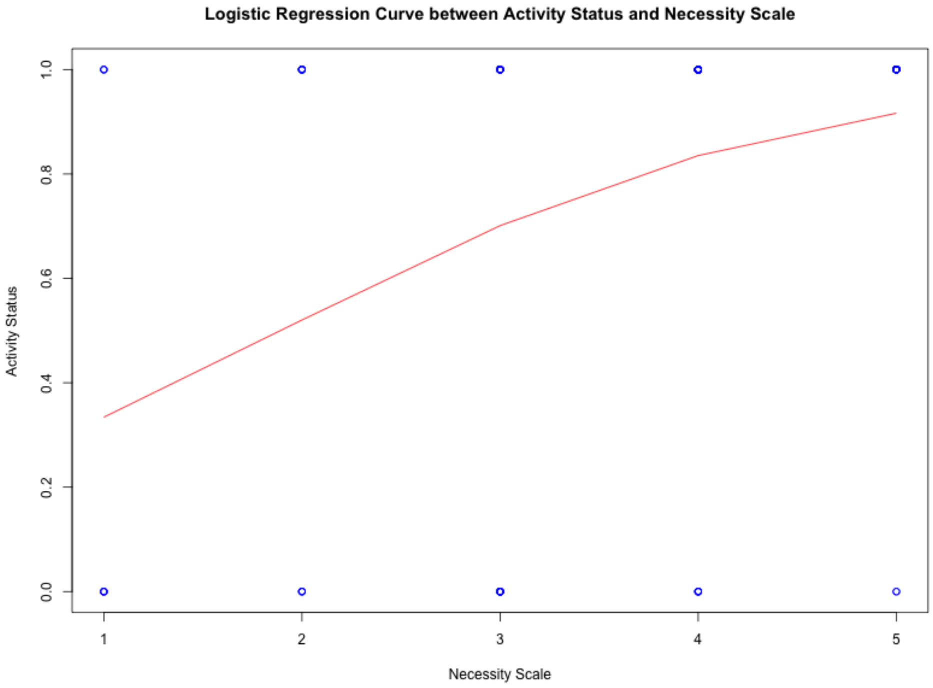
Table 6: Summary of the Logistic Regression Model 1.

From the model, we can see that $\beta_{\text{necessity scale}} \neq 0$ and the p-value is less than $=$ 0.05 (95% confidence level). This indicates that there is significant statistical evidence supporting the relationship between attendance in extracurricular activities and students' own perception of importance.

To interpret the result of the predicting model, we need to convert the result from log odds into probability. The results are as follows:

	Probability that a student will participate in any extracurricular activities across various necessity levels
necessity_scale = 1	0.3339962
necessity_scale = 2	0.5202502
necessity_scale = 3	0.7010402
necessity_scale = 4	0.8352727
necessity_scale = 5	0.9164206

Table 7: Probability of Student Participation Across Necessity Levels.



Looking at the regression curve of the predicting model, we can observe an upward fitted curve. The upward trend as necessity scale increases support the positive relationship between necessity scale and student's chances of participating in extracurricular activity.

Next, we fit a logistic regression model to assess the relationship between **activities_status** and **supportiveness**. For this model, we created a helper variable (**supportiveness_2**) that identifies whether a student received any support from friends/ family. We performed this transformation on the predictor because we only care about whether support from friends or family influences student attendance in extracurricular activities.

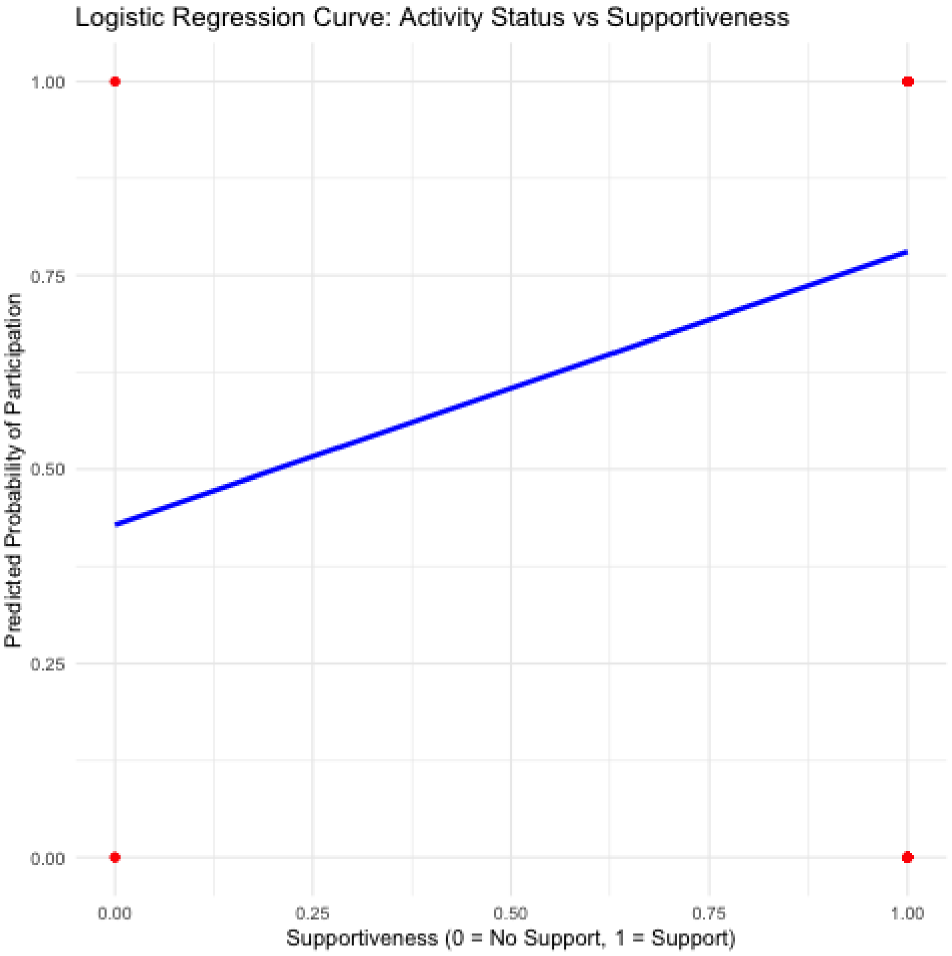
4.4.2 Summary of the Logistic Regression Model 2

term	estimate	std.error	statistic	p.value
(Intercept)	-0.2876821	0.7637626	-0.3766642	0.70642313
supportiveness2	1.5574426	0.7965985	1.9551161	0.05056937

Table 8: Summary of the Logistic Regression Model 2.

From the model, we can see that $\beta_{\text{supportiveness}} \neq 0$ and the p-value is less than = 0.06 (94% confidence level). This indicates strong statistical evidence supporting the relationship.

To interpret the result of the predicting model, we need to convert the result from log odds into probability. The result indicates that students who receive support from at least friends or family have a probability of 0.7807018 of participating in extracurricular activities, whether with or without additional support from their social circle.



Looking at the plot of the regression model, we can see the same result from the sharp increase of probability in student's participation between receiving and not receiving any social support

4.4.3 Summary of the Logistic Regression Model 3

term	estimate	std.error	statistic	p.value
(Intercept)	-2.9081300	1.096208	-2.652899	0.0079803684
supportiveness_categoryY	1.5524521	0.859755	1.805691	0.0709665804
necessity_scale	0.7748652	0.208547	3.715542	0.0002027684

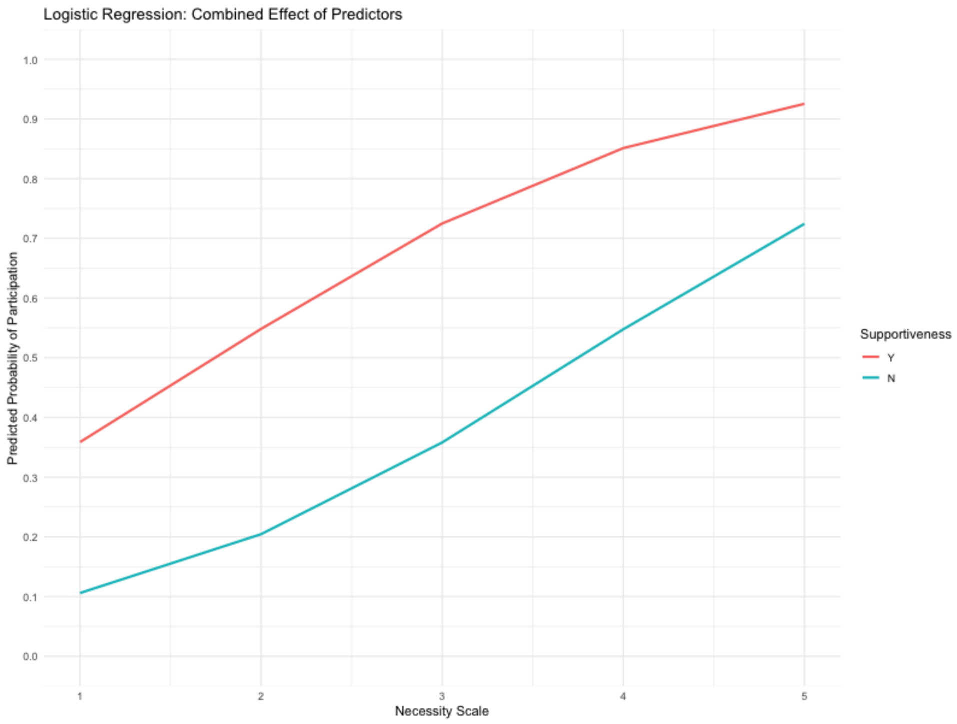
Table 9: Summary of the Logistic Regression Model 3.

From the model, we can observe that both estimations for $\beta_{\text{necessity scale}}$ and $\beta_{\text{supportiveness category (Y)}} \neq 0$. However, the p-value for the category predictor fails the test; this variable is still significant if we allow a higher error level (92% or 91% confidence interval). Under a lower confidence interval, the prediction of this model is:

	Supportive (Y)	Non Supportive (N)
Necessity Scale = 1	0.3587456	0.1059055
Necessity Scale = 2	0.5483615	0.2045005
Necessity Scale = 3	0.7249037	0.3581197
Necessity Scale = 4	0.8511669	0.5476873
Necessity Scale = 5	0.9254388	0.7243606

Table 10: Comparison of Supportive (Y) and Non Supportive (N) Groups by Necessity Scale.

We can observe a significant drop in predicted probability among students who do not receive any support across all levels of perceived importance. This signifies the impact of the categorical variables on the model, despite the trade-off in accuracy from lowering the confidence interval.



The plot of the predicting model shows necessity scale is a strong factor influencing student's participation in extracurricular activities. We can see the increase in probability of participating increases as necessity scale increases. Furthermore, the plot shows the amplification between social supportiveness and student's participation. We can observe students receive social support (redline) have higher probability of participation than those without any social support (blue line).

5 Discussion

RQ1: Do personal preferences influence participation in extracurricular activities?

The analysis showed a clear connection between students' favorite activities and the extracurricular they choose to join. A chi-squared test supported this link, yielding a chi-squared value of 83.02536 and a p-value of 4.719795×10^{-11}. This result suggests that students are more likely to participate in activities that align with their personal interests, choosing options that reflect what they genuinely enjoy. This finding makes sense, as students tend to gravitate toward activities that match their own passions. Based on this, we reject the null hypothesis, indicating a relationship between personal preferences and extracurricular choices.

RQ2: To what extent do schoolwork and/or professional commitments influence a student's participation in extracurricular activities?

The data analysis showed no significant link between schoolwork, work commitments, or study hours and students' participation in extracurricular activities. The two-sample proportion test comparing participation across different levels of study and work time yielded a high p-value ($p = 1$), indicating no significant association. This finding implies that students' involvement in activities may not be heavily influenced by their academic or work schedules as much as one might assume. Other factors, such as personal interest or social support, may be more influential in driving their decisions to participate, as these commitments alone don't seem to stand in the way of joining extracurriculars. We fail to reject the null hypothesis, indicating that school and work commitments are not associated with extracurricular participation.

RQ3: How does the perceived importance of extracurricular activities among domestic and international students influence their choice of participating?

The logistic regression models suggest a strong relation between attendance in extracurricular activities and the predictor variables, both individually and together as a group. We also observe the importance of having support from friends and family in logistic regression model 3, as indicated by the large drop in predicted probability for students who have no support.

RQ4: Do the opinions of friends and/or family influence students' views on participating in extracurricular activities?

From the second logistic regression showed a strong connection between support from friends and/or family and joining extracurricular, with a positive coefficient of 1.5574 for supportiveness and a p-value of 0.0506, which is just above the usual significance level. This suggests that students with family or social support are more inclined to participate in extracurricular. While the predictor is only somewhat significant, it hints that family encouragement could play a small role in motivating students to get involved. This aligns with the idea that social support can encourage students to engage in activities beyond academics. We fail to reject the null hypothesis at the standard significant level $\alpha = 0.05$ (95% confidence level) but this is still a significant result if we interpret the statistically result at a lower significant level.

6 Limitations

For questions about time spent commuting, studying, or working, we should let people enter a number instead of a set of range to collect rigorous data. Using ranges caused the graph to show large jumps between each range, making it harder to see the correlation. We also had to assign midpoint values within each range (e.g., 0–2 hours as 1), which could misrepresent data since some students might want to specify 0 hours. When interpreting the result of this report, it is worth noticing the following limitations:

More obscure results arise from generalizing the data. For example, while fitting the second and third logistic regression, we perform a simple transformation on the predictor (supportiveness) to another one that only counts whether students receive support or not, rather than breaking it down by family or friend's support. While this adjustment improves the predictability of our models, it does so by altering/simplifying what we are analyzing. Instead of assessing and comparing the impact of getting support from friends and/or family we only assess the impact of getting support or not.

In addition, Logistic Regression Models 2 and 3 show higher inaccuracy and cannot be used at the standard 95% confidence level. However, they can be considered usable if the confidence level is slightly lowered. For example, Model 2 achieves 94% confidence, while Model 3 achieves 91%. Although the reduction in confidence may appear minor, it indicates a trade-off in reliability, making these models less dependable than what is typically acceptable. Further details can be found on pages 13 and 15, respectively.

Less reliable result from the Chi-Square Test, in the analysis we acknowledge that the expectation frequencies are not strictly greater than 5 and the Independence test relies on this assumption. However, even though the results are not accurate we can improve the accurateness by grouping categories. The process and whether or not this process is appropriate for this study will be discussed in the final report.

7 Conclusion

This study explored the factors related to students' decisions to join extracurricular activities, including their personal interests, school and work schedules, how much they value extracurricular activities, and the level of family support they receive.

For **RQ1**, we found that students are more likely to participate in activities that align with their interests, indicating that personal preferences play a significant role in their choices. For **RQ2**, there was no strong connection between school or work commitments and participation, suggesting that academic or work schedules do not necessarily prevent students from joining extracurricular activities. **RQ3** showed that while domestic and international students view extracurriculars as valuable, this does not always lead to more involvement, as practical factors like time and workload might be more influential. For **RQ4**, family support had a slight connection to participation, hinting that family encouragement may help, though it's not a major factor.

Our findings show several important connections, though certain limitations shaped our results and highlight the need for more planning ahead. Using ranges instead of exact numbers for time spent on activities created data gaps, making it harder to spot clear patterns, and assigning midpoints for ranges may have led to some misrepresentation. We also had to group certain responses, like support from family or friends, into broader categories, limiting our ability to see how specific types of support connect with student experiences.

In conclusion, this study provides valuable insight into the factors that influence the participation of students in extracurricular activities. By identifying these factors, we aim to support efforts to improve student engagement and well-being through customized programs and initiatives. We hope that this research will serve as a stepping stone for future studies, encouraging academic institutions to develop strategies that foster greater involvement in extracurricular activities, benefiting not only students enrolled in STA304, but also the wider student population at the University of Toronto Mississauga.

8 Appendix

Listed below is the R code:

```
###loading the dataset

‘‘‘{r}
# Change the directory to where the dataset is stored when running the code
library(readr)
data <- read_csv("/Users/phamhieu/Desktop/UTMCOURSES/STA’S/STA304/Project/survey_re
spond_clnd.csv")
‘‘‘

### Presentation of Data and Statistical Analysis

‘‘‘{r}
#Sample Size Calculation - Simple Random Sampling
library(dplyr)
library(broom)
library(flextable)

N <- 121
B = 0.05
D = (B^2)/4

#filter out students that participates in any extracurricular activities and  dont work
student_extra_wrk <- nrow(subset(data, activities_status == 1 & work_time != 0))

#proportion of students that participates in any extracurricular
activities and work, assume p=0.5
p <- 0.5
q <- 1-p

#sample size calculation for p
n <- ceiling((N*p*q)/((N-1)*D+p*q))

#Drawing Simple Random Samples

#Sample 1 - Set a sample seed so that it saves the sample data
set.seed(002)
sample_data1 <- sample_n(data,n)
head(sample_data1)

#Two Sample Proportion Test between activity status and work time
contingency_table <- table(sample_data1$activities_status, sample_data1$work_time2)
print(contingency_table)
prop.test(contingency_table)
prop_test_sum <- tidy(prop.test(contingency_table))
prop_test1_table <- flextable(prop_test_sum) %>%
  set_caption(caption = "Two Sample Proportion Test between Activity
  Status and Work Time")%>%
  autofit()%>%
  theme_vanilla()
print(prop_test1_table)

#fail to reject Null hypotheis


‘‘‘

‘‘‘{r}
#Sample Size Calculation - Simple Random Sampling
library(dplyr)
library(broom)
library(flextable)

#Drawing Simple Random Samples

#Sample 2 - Set a sample seed so that it saves the sample data
set.seed(212)
sample_data2 <- sample_n(data,n)
head(sample_data2)


#Two Sample Proportion Test between activity status and study time
contingency_table2 <- table(sample_data2$activities_status, sample_data2$study_time2)
print(contingency_table2)
prop.test(contingency_table2)

prop_test_sum <- tidy(prop.test(contingency_table2))
prop_test2_table <- flextable(prop_test_sum) %>%
  set_caption(caption = "Two Sample Proportion Test between Activity
  Status and Study Time")%>%
  autofit()%>%
  theme_vanilla()

print(prop_test2_table)

#fail to reject null hypothesis

#conclude: RQ2 can’t conlude that studytime and worktime affect activities status
‘‘‘

### Simple Test 1 - Simple Linear Regression (Failed gotta do another one T.T)

‘‘‘{r}

#Response Variable
barplot(table(data$activities_status), main="Bar plot of Activities Status",
        xlab = "Activities Status", ylab = "Frequency",border =NA,col =
```

~~~~
        c("#FF6666","#6699FF"))

#box plot
data2 <- data.frame(activities_status = data$activities_status, study_time
= data$study_time)
boxplot(activities_status ~ study_time, data = data2, main = "Boxplot of
Activities Status by Sutdy Time Scale",xlab = "study_time", ylab =
"Activities Status", col = "cadetblue")

#box plot2
data3 <- data.frame(activities_status = data$activities_status,
necessity_scale = data$necessity_scale)
boxplot(activities_status ~ necessity_scale, data = data3, main = "Boxplot
of Activities Status by Necessity Scale",xlab = "necessity_scale", ylab =
"Activities Status", col = "cadetblue")


```

### Advanced Test: Logistic Regression

```{r}
#No Multicolinearity

#Matrix of potential predictors
plot(data[,c("study_time","supportiveness","work_time","necessity_scale")],col="cadetblue")

#VIF factors for each predictor

library(broom)
library(car)
model_log0 = glm(activities_status ~ study_time + supportiveness +
work_time + necessity_scale, data = data, family = "binomial")
vif_value <- vif(model_log0)
vif_df <- tidy(vif_value)

library(flextable)
vif_table <- flextable(vif_df) %>%
  set_caption(caption = "VIF Values for each Predictor")%>%
  autofit()%>%
  theme_vanilla()

print(vif_table)

#Conclusion since all VIF values almost equal to 1, no correlation between any predictor

```

```{r}
#Logistic Regression on activity status and necessity scale
model_log = glm(activities_status ~  necessity_scale, data = data, family = "binomial")
~~~~

```
summary(model_log)
library(broom)
library(flextable)
model_log_sumry <- tidy(model_log)
logreg_table <- flextable(model_log_sumry) %>%
  set_caption(caption = "Logistic Regression between Activity Status and
  Necessity Scale")%>%
  autofit()%>%
  theme_vanilla()

print(logreg_table)

#Logistic Regression Curve

newdata <- data.frame(necessity_scale = seq(1,5,1))
newdata$prob <- predict(model_log, newdata, type = "response")
plot(activities_status~necessity_scale, data = data, col = "blue",main
="Logistic Regression Curve between Activity Status and Necessity Scale",
     xlab = "Necessity Scale", ylab = "Activity Status")
lines(prob~necessity_scale, data = newdata, col = "red")


#model coefficients
coef_intercept = as.numeric(coefficients(model_log)[1])
necessity_scale_coef = as.numeric(coefficients(model_log)[2])

#probability across necessity scale

#necessity scale = 1
odds = exp(coef_intercept+necessity_scale_coef*1)
probs = odds/(1+odds)
probs

#necessity scale = 2
odds = exp(coef_intercept+necessity_scale_coef*2)
probs = odds/(1+odds)
probs

#necessity scale = 3
odds = exp(coef_intercept+necessity_scale_coef*3)
probs = odds/(1+odds)
probs

#necessity scale = 4
odds = exp(coef_intercept+necessity_scale_coef*4)
probs = odds/(1+odds)
probs

#necessity scale = 5
odds = exp(coef_intercept+necessity_scale_coef*5)
probs = odds/(1+odds)
```

```
probs
‘‘‘


‘‘‘{r}
#Probability of participating in extracurricular activities given necessity scale

# Logistic Regression on activity status and supportiveness
model_log2 = glm(activities_status ˜ supportiveness2, data = data, family = "binomial")
summary(model_log2)
library(broom)
library(flextable)
model_log2_sumry <- tidy(model_log2)
logreg_table2 <- flextable(model_log2_sumry) %>%
  set_caption(caption = "Logistic Regression between Activity Status and
  Supportiveness")%>%
  autofit()%>%
  theme_vanilla()

print(logreg_table2)

#Plotting the Logistic Regression Curve
library(ggplot2)
# Generate predicted probabilities
new_data <- data.frame(supportiveness2 = seq(0, 1,1))
new_data$predicted_probs <- predict(model_log2, newdata = new_data, type = "response")

ggplot(new_data, aes(x = supportiveness2, y = predicted_probs)) +
  geom_line(color = "blue", size = 1.2) +
  geom_point(data = data, aes(x = supportiveness2, y = activities_status),
  color = "red") +
  labs(title = "Logistic Regression Curve: Activity Status vs
  Supportiveness",x = "Supportiveness (0 = No Support, 1 = Support)",y =
  "Predicted Probability of Participation") +theme_minimal()

#model coefficients
coef_intercept = as.numeric(coefficients(model_log2)[1])
supportiveness_scale = as.numeric(coefficients(model_log2)[2])
coef_intercept
supportiveness_scale

#Probability across all supportiveness levels

#supportiveness = 0
odds = exp(coef_intercept+supportiveness_scale*0)
probs = odds/(1+odds)
probs

#supportiveness = 1
odds = exp(coef_intercept+supportiveness_scale*1)
probs = odds/(1+odds)
```

```
probs

‘‘‘

‘‘‘{r}
library(broom)
library(flextable)
model_log3 = glm(activities_status ~ supportiveness_category +
necessity_scale, data = data, family = "binomial")
summary(model_log3)
model_log3_sumry <- tidy(model_log3)
mltpl_logreg_table <- flextable(model_log3_sumry) %>%
  set_caption(caption = "Logistic Regression between Activity Status
  versus Supportiveness and Necessity Scale")%>%
  autofit() %>%
  theme_vanilla()

print(mltpl_logreg_table)

#Plotting the Multiple Logistic Regression Curve


library(ggplot2)

new_data <- expand.grid(necessity_scale =
seq(1,5,1),supportiveness_category = unique(data$supportiveness_category))
new_data$predicted_probs <- predict(model_log3, newdata = new_data, type =
"response")
ggplot(new_data, aes(x = necessity_scale, y = predicted_probs, color =
supportiveness_category)) +
  geom_line(size = 1) +  # Smooth logistic regression curves
  scale_y_continuous(limits = c(0, 1), breaks = seq(0, 1, 0.1)) +  #
  Ensure y-axis is between 0 and 1
  labs(
    title = "Logistic Regression: Combined Effect of Predictors",
    x = "Necessity Scale",
    y = "Predicted Probability of Participation",
    color = "Supportiveness"
  ) +
  theme_minimal()



#model coefficients
coef_intercept = as.numeric(coefficients(model_log3)[1])
supportiveness_Y = as.numeric(coefficients(model_log3)[2])
necessity_scale_coef = as.numeric(coefficients(model_log3)[3])


#Probability of participating in extracurricular activities given
necessity scale and supportiveness
```

```
#If student family is supportive(Y) and necessity scale = 1
odds = exp(coef_intercept+supportiveness_Y*1+necessity_scale_coef*1)
probs = odds/(1+odds)
probs

#If student family is not supportive(N) and necessity scale = 1
odds = exp(coef_intercept+supportiveness_Y*0+necessity_scale_coef*1)
probs = odds/(1+odds)
probs

#If student family is supportive(Y) and necessity scale = 2
odds = exp(coef_intercept+supportiveness_Y*1+necessity_scale_coef*2)
probs = odds/(1+odds)
probs

#If student family is not supportive(N) and necessity scale = 2
odds = exp(coef_intercept+supportiveness_Y*0+necessity_scale_coef*2)
probs = odds/(1+odds)
probs

#If student family is supportive(Y) and necessity scale = 3
odds = exp(coef_intercept+supportiveness_Y*1+necessity_scale_coef*3)
probs = odds/(1+odds)
probs

#If student family is not supportive(N) and necessity scale = 3
odds = exp(coef_intercept+supportiveness_Y*0+necessity_scale_coef*3)
probs = odds/(1+odds)
probs

#If student family is supportive(Y) and necessity scale = 4
odds = exp(coef_intercept+supportiveness_Y*1+necessity_scale_coef*4)
probs = odds/(1+odds)
probs

#If student family is not supportive(N) and necessity scale = 4
odds = exp(coef_intercept+supportiveness_Y*0+necessity_scale_coef*4)
probs = odds/(1+odds)
probs

#If student family is supportive(Y) and necessity scale = 5
odds = exp(coef_intercept+supportiveness_Y*1+necessity_scale_coef*5)
probs = odds/(1+odds)
probs

#If student family is not supportive(N) and necessity scale = 5
odds = exp(coef_intercept+supportiveness_Y*0+necessity_scale_coef*5)
probs = odds/(1+odds)
probs
```

```
‘‘‘
### Pearson’s Chi Square Test
‘‘‘{r}

# Create a new column that contains only the first activity
data_activities_short <- substr(sample_data2$activities, 1, 2)

# Create a contingency table for the two columns
contingency_table3<- table(sample_data2$favorite_activity,data_activities_short)
print(contingency_table3)

# Perform the chi-squared independence test
chi_squared_test <- chisq.test(contingency_table3)
chi_squared_test$expected



# Print the results
library(broom)
library(knitr)

#Chi-Square Test Result
chi_squared_test_sum <- tidy(chi_squared_test)

chi_sqr_tst_tble <- flextable(chi_squared_test_sum) %>%
  set_caption(caption = "Pearson’s Chi Square Test between Favorite
  Activity and Activity")%>%
  autofit()%>%
  theme_vanilla()

print(chi_sqr_tst_tble)

#Frequency Table
frqncy_sum <- as.data.frame(chi_squared_test$expected)
frqncy_sum <- cbind(Activity_vs_Favorite_Activity =
rownames(chi_squared_test$expected), frqncy_sum)

frqncy_table <- flextable(frqncy_sum) %>%
  set_caption(caption = "Expected Values for Pearson’s Chi Square Test
  between Favorite Activity and Activity")%>%
  autofit()%>%
  bold(part = "header")

print(frqncy_table)

#Success to reject the Null Hypothesis
‘‘‘


### Stacked bar-chart to summarize the open responses
‘‘‘{r}
```

```
library(ggplot2)

# Stacked bar chart
ggplot(data, aes(x = open_reason, fill = factor(necessity_scale))) +
  geom_bar(position = "fill") +
  labs(y = "Proportion", fill = "Likert Scale") +
  theme_minimal()

library(tidyr)

summary_table <- data %>%
  group_by(open_reason, necessity_scale) %>%
  summarise(count = n()) %>%
  spread(key = necessity_scale, value = count, fill = 0)

print(summary_table)
```

‘‘‘