Length of Commute to U of T Campuses and Academic Performance Survey

Juan Acosta

January 19, 2022

Introduction

Goal of Survey

The goal of this survey is to attempt to establish if there is a relationship between the length of a student's commute to any of the three University of Toronto campuses and their academic performance. For example, if a student's commute is long and tedious, does that correlate with a lower cumulative grade point average (CGPA) versus a student whose commute is short and straightforward? A secondary goal of this survey is also to measure if a student self-correlates their commute with a positive, negative or neutral effect on their academic standing.

This topic is objectively interesting, because it is a topic that has inspired an academic research paper on the relationship between the distance from campus and the GPA of commuter students [1], and a large-scale study of student transportation in the Greater Toronto and Hamilton Area to understand student transportation preferences and travel patterns [2].

Procedure

The target population of this survey is all commuter undergraduate and graduate students at all three University of Toronto campuses. In order to attempt to reach the most students possible, several student societies and unions at the University are asked if they can send out the link to the survey through their email newsletters, and those emails that are sent the newsletters would form my frame population. Thus, the students that choose to fill out the survey would form my sample population.

The drawbacks to only relying on student societies and unions to send out the survey link through their email newsletters is that not all of them may agree on sending out the link, and not all students are subscribed to receive those newsletters. However, these student societies and unions should have a large enough reach that the students who do choose to fill out the survey should be able to form a relatively large sample population, which is paramount to the analysis of data from the survey.

Terms of Note

In order to more neatly display some of the data and graphs, the following popular abbreviations for the campus names will be used:

- St. George (Downtown Toronto) campus will be abbreviated as UTSG.
- Mississauga campus will be abbreviated as UTM.
- Scarborough campus will be abbreviated as UTSC.

Simulated Data

The data simulated here represent the responses to the survey questions from 5167 students across all three University of Toronto campuses. The survey aims to establish if there is a relationship between the length and type of a student's commute and both their objective and perceived academic performance. Each question was simulated separately as follows:

Question 1: Which University of Toronto campus do you attend?
The responses to this question were simulated from a random sample of the three possible categorical responses - *UTSG, UTM, UTSC* - with n=5167 responses, with replacement, and with the respective probabilities of 0.676, 0.172, 0.152 for each response option. These probabilities take into account the probability that a student is from a certain campus, given the percentage of that campus's population with respect to the total University of Toronto population.

Question 2: What is your year of study (credit-wise)?
The responses to this question were simulated from a random sample of the five possible categorical responses - *First-year, Second-year, Third-year, Fourth-year, Graduate student* - with n=5167 responses, with replacement, and with the respective probabilities of 0.19575, 0.19575, 0.19575, 0.19575, 0.217 for each response option. These probabilities take into account the probability that a student is in a certain year of study, or is a graduate student, given the percentage of undergraduate and graduate students with respect to the total University of Toronto population.

Question 3: What is your CGPA?
The responses to this question were simulated using constraints based on corresponding responses to Question 2, starting off with an empty character vector, and then using a for-loop to go through each of n=5167 responses to Question 2, and if-else statements to set the constraints on the responses to Question 3 as follows:

- If the corresponding response to Q2 was *Graduate student*, then the response to this question is simulated from a random sample of three possible categorical responses - *1.7 - 2.69, 2.7 - 3.69, 3.7 - 4.0*
- Else, the response to this question is simulated from a random sample of four possible categorical responses - *0.0 - 1.69, 1.7 - 2.69, 2.7 - 3.69, 3.7 - 4.0*

These constraints on the responses take into account the fact that graduate students will tend to have a higher CGPA than the 0.0 - 1.69 range, and so the possibility of those responses is excluded.

Question 4: How far (in kilometres) do you live from campus?
The responses to this question were simulated from a random sample of the five possible categorical responses - *< 5 km, 5 - 10 km, 10.01 - 20 km, 20.01 - 30 km, > 30 km* - with n=5167 responses, with replacement.

Question 5: What is your primary mode of transportation to and from campus?
The responses to this question were simulated using constraints based on corresponding responses to Question 4, starting off with an empty character vector, and then using a for-loop to go through each of n=5167 responses to Question 4, and if-else statements to set the constraints on the responses to Question 5 as follows:

- If the corresponding response to Q4 was *< 5 km*, then the response to this question is simulated from a random sample of four possible categorical responses - *Public transit (bus, subway, train, streetcar, LRT, Scarborough RT), Taxi or rideshare (Uber, Lift, etc.), Cycling, Skateboard, or Scooter, Walking*
- If the corresponding response to Q4 was *5 - 10 km*, then the response to this question is simulated from a random sample of five possible categorical responses - *Driving (by yourself), Carpool, Public transit (bus, subway, train, streetcar, LRT, Scarborough RT), Taxi or rideshare (Uber, Lift, etc.), Cycling, Skateboard, or Scooter*

- If the corresponding response to Q4 was *10.01 - 20 km*, then the response to this question is simulated from a random sample of five possible categorical responses - *Driving (by yourself), Carpool, Public transit (bus, subway, train, streetcar, LRT, Scarborough RT), Taxi or rideshare (Uber, Lift, etc.), Cycling, Skateboard, or Scooter*
- If the corresponding response to Q4 was *20.01 - 30 km*, then the response to this question is simulated from a random sample of four possible categorical responses - *Driving (by yourself), Carpool, Public transit (bus, subway, train, streetcar, LRT, Scarborough RT), Taxi or rideshare (Uber, Lift, etc.)*
- If the corresponding response to Q4 was *> 30 km*, then the response to this question is simulated from a random sample of four possible categorical responses - *Driving (by yourself), Carpool, Public transit (bus, subway, train, streetcar, LRT, Scarborough RT), Taxi or rideshare (Uber, Lift, etc.)*
- Else, the response to this question is simulated from a random sample of six possible categorical responses - *Driving (by yourself), Carpool, Public transit (bus, subway, train, streetcar, LRT, Scarborough RT), Taxi or rideshare (Uber, Lift, etc.), Cycling, Skateboard, or Scooter, Walking*

These constraints on the responses are to ensure that for an individual simulated respondent, the relation between the responses for Question 4 and Question 5 make sense. For example, this ensures that an individual simulated respondent who answered that they live greater than 30 kilometres from campus does not answer that they walk as their primary mode of transportation to and from campus. (While a pair of responses such as this are not impossible, they are highly improbable.)

Question 6: In total, how long (in minutes) do you commute in a day?
The responses to this question were simulated using constraints based on corresponding responses to Question 4, starting off with an empty character vector, and then using a for-loop to go through each of n=5167 responses to Question 4, and if-else statements to set the constraints on the responses to Question 6 as follows:

- If the corresponding response to Q4 was *< 5 km*, then the response to this question is simulated from one possible categorical response - *< 15 minutes*
- If the corresponding response to Q4 was *5 - 10 km*, then the response to this question is simulated from a random sample of three possible categorical responses - *< 15 minutes, 15 - 30 minutes, 31 - 60 minutes*
- If the corresponding response to Q4 was *10.01 - 20 km*, then the response to this question is simulated from a random sample of three possible categorical responses - *31 - 60 minutes, 61 - 90 minutes, 91 - 120 minutes*
- Else, the response to this question is simulated from a random sample of four possible categorical responses - *31 - 60 minutes, 61 - 90 minutes, 91 - 120 minutes, > 120 minutes*

These constraints on the responses are to ensure that for an individual simulated respondent, the relation between the responses for Question 4 and Question 6 make sense. For example, this ensures that an individual simulated respondent who answered that they live greater than 30 kilometres from campus does not answer that their commute is less than 15 minutes in total in a day.

Question 7: Do you study on your commute?
The responses to this question were simulated using constraints based on corresponding responses to Question 5, starting off with an empty character vector, and then using a for-loop to go through each of n=5167 responses to Question 5, and if-else statements to set the constraints on the responses to Question 7 as follows:

- If the corresponding response to Q5 was either *Driving (by yourself)*, *Cycling, Skateboard, or Scooter* or *Walking*, then the response to this question is simulated from one possible categorical response - *Never*
- Else, the response to this question is simulated from a random sample of three possible categorical responses - *Always, Sometimes, Never*

These constraints on the responses are to ensure that for an individual simulated respondent, the relation between the responses for Question 5 and Question 7 make sense. It is improbable that someone commuting by car, bicycle or walking will be reading a textbook or working on homework.

Question 8: Have you ever been late for or missed a lecture, tutorial, test or exam, primarily because of traffic congestion or a delay in public transit?
The responses to this question were simulated using constraints based on corresponding responses to Question 5, starting off with an empty character vector, and then using a for-loop to go through each of n=5167 responses to Question 5, and if-else statements to set the constraints on the responses to Question 8 as follows:

- If the corresponding response to Q5 was either *Cycling, Skateboard, or Scooter* or *Walking*, then the response to this question is simulated from one possible categorical response - *No*
- Else, the response to this question is simulated from a random sample of two possible categorical responses - *Yes, No*

These constraints on the responses are to ensure that for an individual simulated respondent, the relation between the responses for Question 5 and Question 8 make sense. If the individual simulated respondent does not drive or take public transit, then the response to this question is logically no.

Question 9: Have you ever arrived on campus much earlier than necessary, or left campus much later than necessary, primarily to avoid traffic congestion or a potential delay in public transit?
The responses to this question were simulated using constraints based on corresponding responses to Question 5, starting off with an empty character vector, and then using a for-loop to go through each of n=5167 responses to Question 5, and if-else statements to set the constraints on the responses to Question 9 as follows:

- If the corresponding response to Q5 was either *Cycling, Skateboard, or Scooter* or *Walking*, then the response to this question is simulated from one possible categorical response - *No*
- Else, the response to this question is simulated from a random sample of two possible categorical responses - *Yes, No*

These constraints on the responses are to ensure that for an individual simulated respondent, the relation between the responses for Question 5 and Question 9 make sense. If the individual simulated respondent does not drive or take public transit, then the response to this question is logically no.

Question 10: Was ease of commuting one of the reasons you chose to attend U of T?
The responses to this question were simulated from a random sample of the two possible categorical responses - *Yes, No* - with n=5167 responses, with replacement.

Question 11: Do you feel that your commute time has had a positive, negative, or neutral effect on your academic performance?
The responses to this question were simulated from a random sample of the three possible categorical responses - *Positive, Negative, Neutral* - with n=5167 responses, with replacement.

Important Variables

The important variables are the following:

- **q1_responses, q2_responses, . . . , q11_responses**
  - These variables store the vectors with the response options for each question as their elements. From these vectors, the random samples drew the responses for each of the n=5167 simulated respondents.
- **Q1, Q2, . . . , Q11**
  - These variables store the vectors that contain the responses of each question for each of the n=5167 simulated respondents.
- **my_data**
  - This variable stores the table formed by combining the corresponding entries of each question response vector (Q1, Q2, . . . , Q11) to form 5167 entries which correspond to the survey responses of n=5167 simulated respondents.

Numerical Summaries

Each numerical summary is a marginal proportional table.

The first table measures the proportion of students from each campus that have missed class or an important assessment due to traffic congestion or transit delays. The percentages are as follows:

	No	Yes
UTM	0.60	0.40
UTSC	0.56	0.44
UTSG	0.60	0.40

The second table measures the proportion of students from each year of study / graduate students that have a positive, negative or neutral view on their commute's impact on their academic performance. The percentages are as follows:

	Negative	Neutral	Positive
First-year	0.32	0.34	0.34
Fourth-year	0.33	0.32	0.35
Graduate student	0.31	0.34	0.35
Second-year	0.30	0.37	0.33
Third-year	0.31	0.34	0.35

The third table measures the proportion of students from each campus who have a particular primary mode of transportation to and from campus. The percentages are as follows:

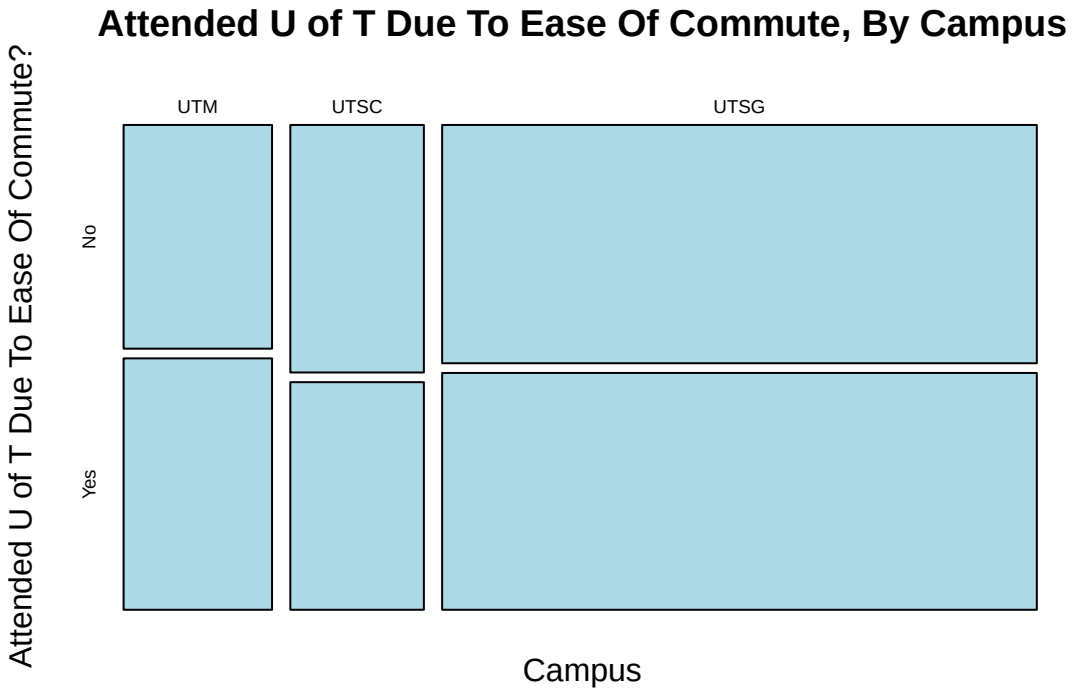
	Carpool	Cycling, Skateboard, or Scooter	Driving (by yourself)	Public transit (bus, subway, train, streetcar, LRT, Scarborough RT)	Taxi or rideshare (Uber, Lift, etc.)	Walking
UTM	0.19	0.14	0.17	0.24	0.21	0.06
UTSC	0.16	0.11	0.17	0.24	0.25	0.06

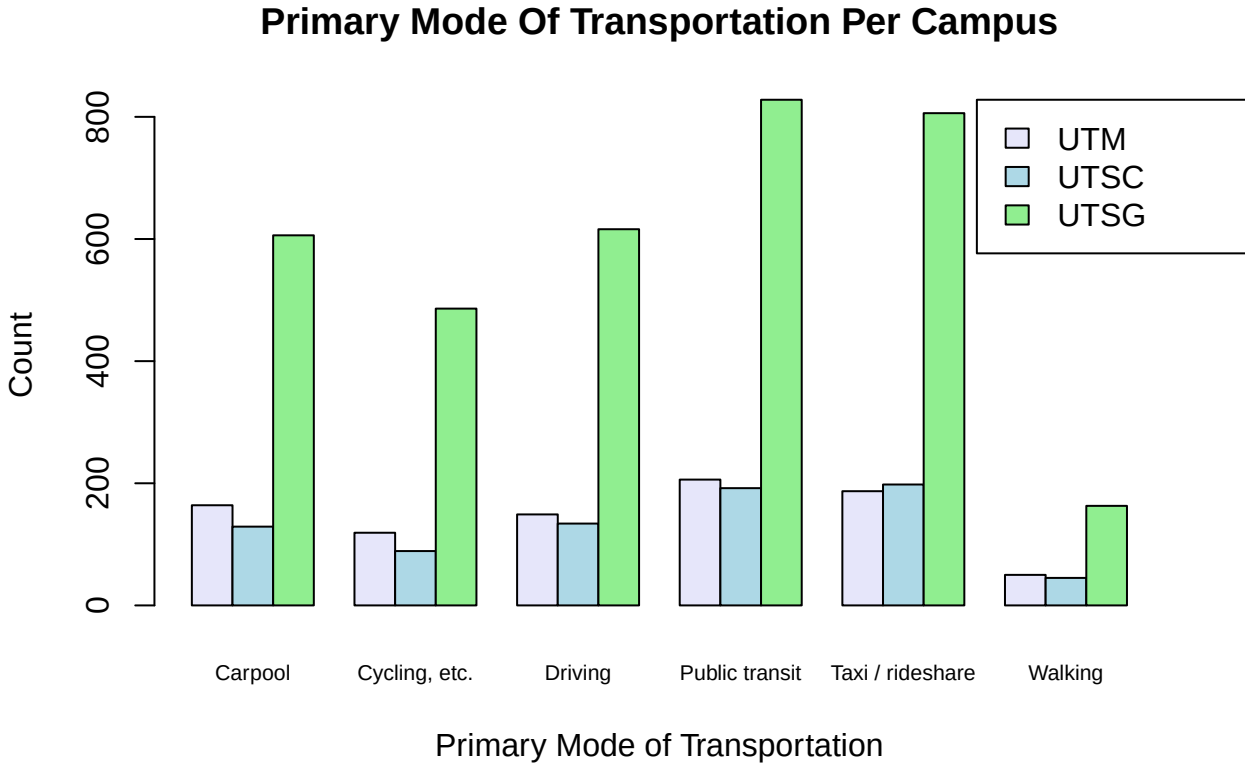
	Carpool	Cycling, Skateboard, or Scooter	Driving (by yourself)	Public transit (bus, subway, train, streetcar, LRT, Scarborough RT)	Taxi or rideshare (Uber, Lift, etc.)	Walking
UTSG	0.17	0.14	0.18	0.24	0.23	0.05

The fourth table measures the proportion of students from each campus that added time to their commute to avoid traffic congestion or public transit delays. The percentages are as follows:

	No	Yes
UTM	0.59	0.41
UTSC	0.62	0.38
UTSG	0.60	0.40

Plots


Attended U of T Due To Ease Of Commute, By Campus
Attended U of T Due To Ease Of Commute?
UTM
UTSC
UTSG
No
Yes
Campus


Primary Mode Of Transportation Per Campus
Count
0
200
400
600
800
UTM
UTSC
UTSG
Carpool
Cycling, etc.
Driving
Public transit
Taxi / rideshare
Walking
Primary Mode of Transportation

The first plot is a mosaic plot [3] that visually compares the proportions of how many students attended U of T because of its ease of commuting relative to which campus the student is registered at. The lengths of the squares represent how many respondents from each campus responded either *Yes* or *No* as a proportion of the total population of U of T.

The second plot is a side-by-side bar chart that compares the number of students per campus that responded with a particular primary mode of transportation. This not only helps compare the popularity of the modes of transportation against each other, but it also helps to compare the proportions between campuses.

All analysis for this report was programmed using `R version 4.1.1`.

Methods

The methodology used here involves both a hypothesis test of the relation between the length of a student's commute and if it affects their academic performance, and a confidence interval for the *Yes* responses to the question that asked whether a student had been late or missed a lecture, tutorial, test or exam primarily because of traffic congestion or a delay in public transit.

The hypothesis test consists of both the null hypothesis (no correlation between the two variables) and the alternative hypothesis (correlation between the two variables). The null and alternative hypotheses are as follows:

- *Null Hypothesis*: There is no correlation between the length of a student's commute and their academic performance.
- *Alternative Hypothesis*: There is a significant correlation between the length of a student's commute and their academic performance.

The hypothesis test will run at a significance level of p=0.05. That is, if the p-value from the hypothesis test is less than or equal to 0.05, then we reject the null hypothesis. However, if the p-value is greater than 0.05, then the null hypothesis is not rejected, and no correlation can be established.

I will invoke a 1-sample proportions test to generate the 95% confidence interval for the *Yes* responses to Question 8; that is, the responses that students had missed a lecture, tutorial, test or exam primarily due to traffic congestion or a delay in public transit.

Results

The hypothesis test was conducted using a chi-squared test. The test returned a p-value of 0.07456, and since this p-value is greater than the significance level of p=0.05, that means the null hypothesis is not rejected, and no correlation can be established between the length of a student's commute and their academic performance. This result seems reasonable, because the academic performance of a student is a subjective response, and might be made by the student given that it is in fact not their commute that has a strong influence on their academic performance.

The 95% confidence interval is (0.3980139, 0.4250332). The confidence interval establishes that the underlying probability of success (that is, a *Yes* response for Question 8) is nestled between approximately 39.80% and 42.50%. This result seems reasonable, given as the true percentage of success is approximately 41.15%.

Bibliography

1. Nelson, D., Misra, K., Sype, G. E., & Mackie, W. (n.d.). *An Analysis Of The Relationship Between Distance From Campus And GPA Of Commuter Students.* Journal of International Education Research, 12(1), 37–46. https://files.eric.ed.gov/fulltext/EJ1088600.pdf. (Last Accessed: October 1, 2021)

2. Mitra, R., Habib, K. N., Siemiatycki, M., Keil, R. and Bowes, J. (2020) *StudentMoveTO - From Insight to Action on Transportation for Post-Secondary Students in the GTHA: 2019 Transportation Survey Findings.* StudentMoveTO. http://www.studentmoveto.ca/wp-content/uploads/2020/10/StudentMoveTO-2019-Report-Final-5-Updated-October-15-2020.pdf. (Last Accessed: October 1, 2021)

3. Finnstats. (2021, August 16). *How to plot categorical data in R-quick guide: R-bloggers.* R-bloggers. https://www.r-bloggers.com/2021/08/how-to-plot-categorical-data-in-r-quick-guide/. (Last Accessed: October 1, 2021)

Appendix

Here is a glimpse of the data set simulated/surveyed:

```
## Rows: 5,167
## Columns: 11
## $ Q1  <chr> "UTSG", "UTM", "UTSC", "UTM", "UTM", "UTM", "UTSG", "UTM", "UTSC",~
## $ Q2  <chr> "Graduate student", "First-year", "Graduate student", "First-year"~
## $ Q3  <chr> "1.7 - 2.69", "3.7 - 4.0", "3.7 - 4.0", "1.7 - 2.69", "2.7 - 3.69"~
## $ Q4  <chr> "10.01 - 20 km", "5 - 10 km", "< 5 km", "> 30 km", "10.01 - 20 km"~
## $ Q5  <chr> "Cycling, Skateboard, or Scooter", "Cycling, Skateboard, or Scoote~
## $ Q6  <chr> "91 - 120 minutes", "< 15 minutes", "< 15 minutes", "> 120 minutes~
## $ Q7  <chr> "Never", "Never", "Sometimes", "Never", "Never", "Always", "Someti~
## $ Q8  <chr> "No", "No", "Yes", "No", "No", "No", "No", "Yes", "Yes", "No", "Ye~
## $ Q9  <chr> "No", "No", "Yes", "No", "No", "Yes", "Yes", "No", "Yes", "No", "N~
## $ Q10 <chr> "Yes", "No", "No", "Yes", "Yes", "No", "No", "No", "No", "Yes", "N~
## $ Q11 <chr> "Negative", "Neutral", "Negative", "Neutral", "Neutral", "Negative~
```