UNIVERSITY OF CALIFORNIA

Los Angeles

An Exploratory Analysis of the Challenges of Latino Students in Two-Year Colleges and Their Perception of Challenges of Transfer to Four-Year Colleges

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Applied Statistics

by

Mariana Gonzalez Castro

2023

ABSTRACT OF THE THESIS

An Exploratory Analysis of the Challenges of Latino Students in Two-Year Colleges and Their Perception of Challenges of Transfer to Four-Year Colleges

by

Mariana Gonzalez Castro
Master of Applied Statistics
University of California, Los Angeles, 2023
Professor Chad J. Hazlett, Chair

A representative sample was collected from an established two-year Hispanic-Serving Institution to explore the academic challenges currently enrolled Latino students are facing today and their perception of potential challenges when transferring to a four-year institution. A mix of exploratory data analysis and text mining analysis resulted in overall themes of financial barriers, academic preparedness and support, family culture and support, and the perception of the transfer process. In addition, through logistic regression analysis, factors that influence the decision to attend a four-year university will be explored with coursework requirements for transfer, family financial support, and distance from home and family being significant influences.

The thesis of Mariana Gonzalez Castro is approved.

Guani Wu

Maryam Mahtash Esfandiari

Ying Nian Wu

Chad J. Hazlett, Committee Chair

University of California, Los Angeles

2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This research project was an opportunity to provide a platform for Latino students' voices to be heard. This would not have been accomplished without the support of my family, friends, and community. All of whom gave me everything they could offer to help me grow and achieve. *Para mi pueblo, seguimos avanzando.*

CHAPTER 1

Introduction

1.1 Background

To say that Latinos are generate an impact in the United States would be a vast understatement. Latinos in the U.S. contribute to the growth of the U.S. economy and population with being responsible for more than 65% of U.S. population growth and 73% of the growth of the U.S. labor force since 2010 [Cola]. Latinos account for 18.7% of the U.S. national population [Cola], while in California alone, 40% of its population is Latino [Bur]. There is a federal designation to higher education institutions called Hispanic-Serving Institution or HSI. For a higher education institution to qualify as HSI, 25% of its undergraduate student population must identify as Hispanic or Latino [Edu]. According to the Hispanic Association of College and Universities (HACU) analysis of the 2021-2022 Integrated Postsecondary Data Education System (IPEDS) Data, there are currently 572 Hispanic-Serving Institutions and 400 emerging Hispanic-Serving Institutions [Colb]. The University of California system are among these 400 emerging Hispanic-Serving Institutions.

1.2 The UC HSI Initiative

On December 2020, University of California, Los Angeles (UCLA) announced the goal to become a Hispanic-Serving Institution by the year 2025 [Cha]. As of 2022, 21.7% of UCLA's undergraduate student population is Latino. UCLA specifically formed a Chancellor Appointed HSI Task Force dedicated to this cause. On April 27, 2023, a HSI Student Town

Hall Meeting took place where updates about this initiative were reported, and the voices of undergraduate Latino students heard.

How then can we bridge this gap between four-year universities wishing to become a Hispanic-Serving Institution and Latino students striving for higher education? Do Latino students feel academically unprepared and perhaps believe that attending a four-year university is therefore unattainable? Or since in common Latino culture family has a higher value over education, Latino students feel the need to prioritize family above all else?

1.3 Research Purpose

A representative sample was collected from an established two-year Hispanic-Serving Institution to explore the academic challenges currently enrolled Latino students are facing today and their perception of potential challenges when transferring to a four-year institution. In addition, factors that influence the decision to attend a four-year university will be explored. The hope with this study is to build a perspective of the Latino academic and transfer experience to better aid four-year institutions in their efforts to increase Latino representation and inclusivity.

CHAPTER 2

Methodology

2.1 Data

2.1.1 Survey Design, Distribution, and Collection

A survey was developed to collect the data for this study. Based on prior knowledge and reading past research related to academic challenges Latino students face, the survey questions focused on financial barriers, academic preparedness and support, Latino family culture and support, and demographics.

There were 36 survey questions and were structured as multiple choice, Likert scaled, or open-ended based on the type of analysis that would follow after collecting responses. The multiple questions were either of the form ”yes” or ”no” or were given a specific set of answers. For questions with Likert scaled answers, students were asked to provide an opinion regarding a specific statement by selecting one of the five scale points: Strongly Agree, Agree, Neither Agree or Disagree, Disagree, and Strongly Disagree. To convert these survey questions into variables, the Likert scale responses were collapsed from five categories to three categories: Agree, Neutral, and Disagree. For the open-ended questions, students were asked to write a short response. The 36 questions were used in three different sections of analysis: exploratory data analysis, regression modeling, and text mining analysis. A full list of the survey questions can be found in the appendix (Table A.1).

The survey was created using the Google Forms application and a QR code generator was used to create a QR code for distribution. The survey link and QR code were shared to various two-year colleges through social media posts, word-of-mouth starting with my personal network, and canvassing at two-year colleges. After much advertising and canvassing, there were 64 respondents to the survey with 63 responses from Imperial Valley College and 1 response from Chaffey College. Since most of the responses were collected from Imperial Valley College, the data was evolved into a representative sample of Imperial Valley College.

Although it was heavily advertised that the target demographic were Latino students currently enrolled in a two-year college, it was not guaranteed that only Latino students completed the survey. Hence, a question regarding ethnicity was asked. Of the 64 respondents, 63 respondents identified as Latino and 1 respondent identified as not Latino. Table 2.1 and 2.2 depicts the number of respondents who identified as Latino and college currently enrolled in, respectively.

Latino	Total
Yes	63
No	1

Table 2.1: Respondent is Latino or Not

College	Total
Imperial Valley College	63
Chaffey College	1

Table 2.2: Respondents' Currently Enrolled College

2.2 Exploratory Data Analysis

2.2.1 Demographics

Multiple questions regarding the respondent's demographics were asked. The respondents were asked to report their age, gender, current GPA, total hours worked per week at their current job, and parents' level of education. They were also asked if they have a significant other, have children, and if English is their first or native language. Finally, they were asked if they are the first in their family to be born in the U.S. and if they are the first in their family to study in higher education in the U.S., implying whether they are first generation American and first generation college student, respectively.

According to the histograms of age and GPA, Figure 2.1 and 2.2 respectively, most respondents are between the ages of 18-21 and their GPA ranges between 3.0-4.0. Table 2.3 describes that for gender, there were 1 non-binary, 25 male, and 38 female respondents. Table 2.4 depicts the distribution of the total number of hours respondents work per week at their current job. Out of 64 respondents, 30 respondents reported that they are not working while 34 respondents reported that they are working.

Gender	Total
Female	38
Male	25
Non-Binary	1

Table 2.3: Respondents' Gender

Hours Worked	Total
40 or more hours	8
20-39 hours	15
Less than 20 hours	11
I am not working	30

Table 2.4: Hours Respondent Works at Current Job

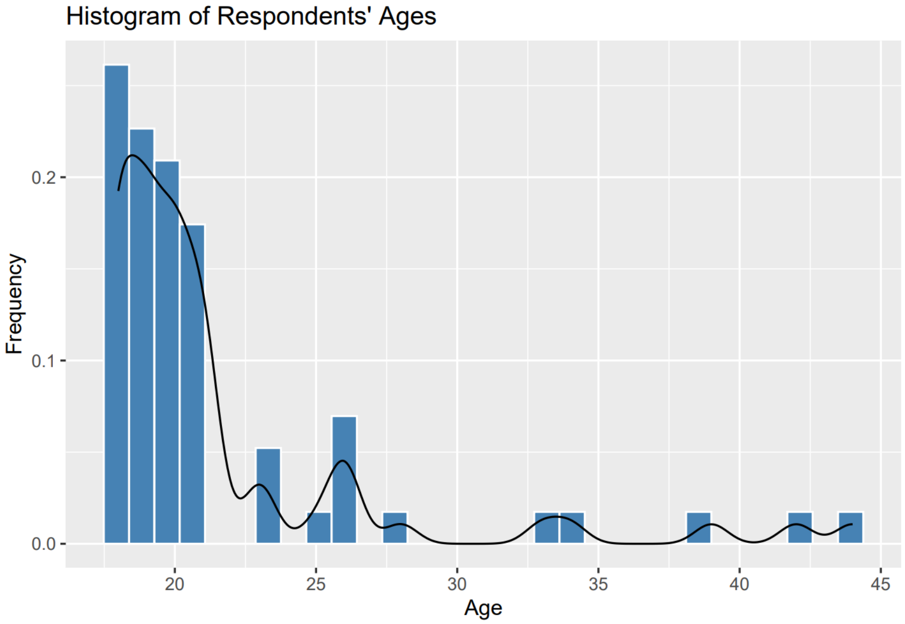

Figure 2.1: Histogram of Age

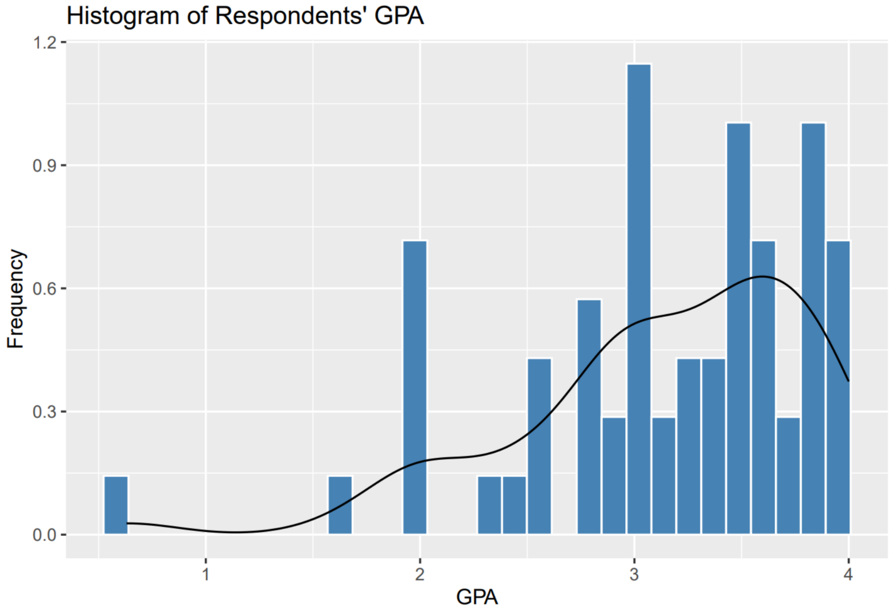

Figure 2.2: Histogram of GPA

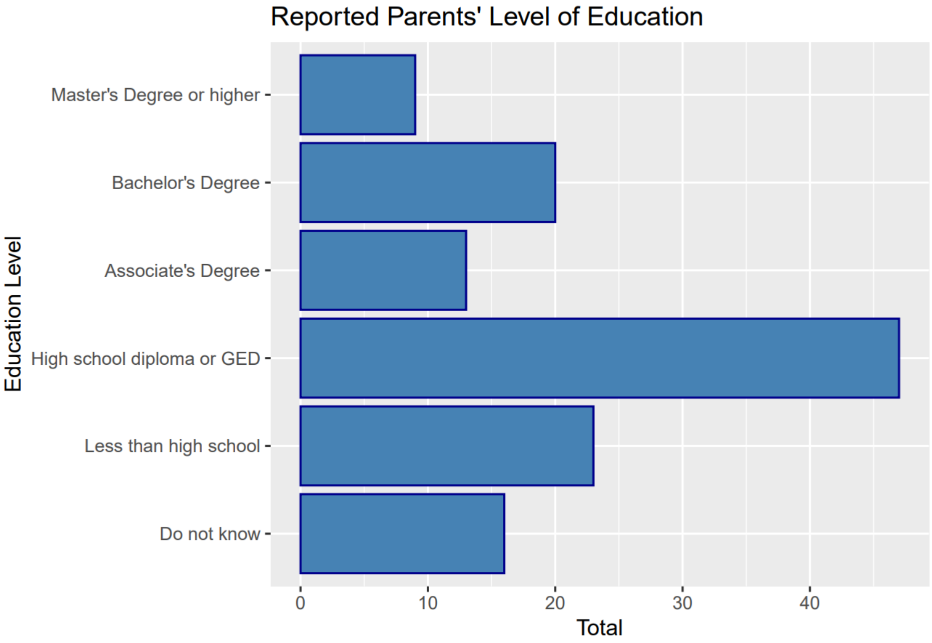

Figure 2.3: Reported Parents' Level of Education

Survey participants were also asked two questions regarding their parents' level of education. The participant was asked to report the highest level of education for Parent #1 and Parent #2. The answer choices for each question they could have chosen were: Master's Degree or higher, Bachelor's Degree, Associate's Degree, High school diploma or GED, Less than high school, and Do not know. Figure 2.3 depicts the responses of the two questions regarding parents' education combined. Of the 128 responses of reported parents' level of education, 70 parents, or 54.7%, have an education level of high school diploma or less. On the other hand, 42 reported parents, or 32.8%, have an education level of an Associate's Degree or higher. Only 9 parents, or 7%, have received a Master's Degree or higher.

Of the 64 respondents who answered the question regarding having a significant other, 13 respondents reported that they do have a significant other while 52 respondents reported not (Table 2.5). Table 2.6 depicts how 58 respondents reported that they do not have children while 6 respondents indicted that they do have children. When asked if English is their first language, 35 respondents respondent that English is their first language while 29 reported not (Table 2.7). The 29 respondents that reported that English was not their first language were then asked what was their first language. All 29 respondents reported that Spanish was their first language.

Has Significant Other	Total
Yes	13
No	52

Table 2.5: Respondent Has a Significant Other

Has Children	Total
Yes	6
No	58

Table 2.6: Respondents Has Children

Is English First Language	Total
Yes	35
No	29

Table 2.7: Is English First Language

When asked if the respondent was the first in their family to be born in the U.S., 22 respondents answered ”yes” while 42 respondents answered ”no” (Table 2.8). The 22 respondents who answered ”yes” are then considered to be first generation Americans. Similarly, when asked if they were the first in their family to attend college in the U.S., 28 respondents answered ”yes” and 36 respondents answered ”no” (Table 2.9). These 28 respondents are considered to be first generation college students.

Is First to be Born in U.S.	Total
Yes	22
No	42

Table 2.8: Respondent is First in Family to be Born in U.S.

Is First to Attend College in U.S.	Total
Yes	28
No	36

Table 2.9: Respondent is First in Family to Attend College in U.S.

2.2.2 Survey Questions Regarding Financial Barriers

Respondents were asked to list what were their financial sources for college expenses. They selected all that applied to them from a list of sources: federal financial aid, grants/scholarships, loans, and other. The ”other” option allowed respondents to write a response that was different from the list. The varied responses were then categorized and added to the list of financial sources. Figure 2.4 depicts the responses of the survey question. Most respondents listed federal financial aid as a financial source. 16 respondents explicitly stated that they were paying for college expenses out-of-pocket and 4 respondents received aid from family.

Respondents were asked a likert scaled question regarding whether they are stressed about paying for college expenses. 36 respondents, or over half, agreed to the statement while 10 disagreed and 18 were neutral (Figure 2.5).

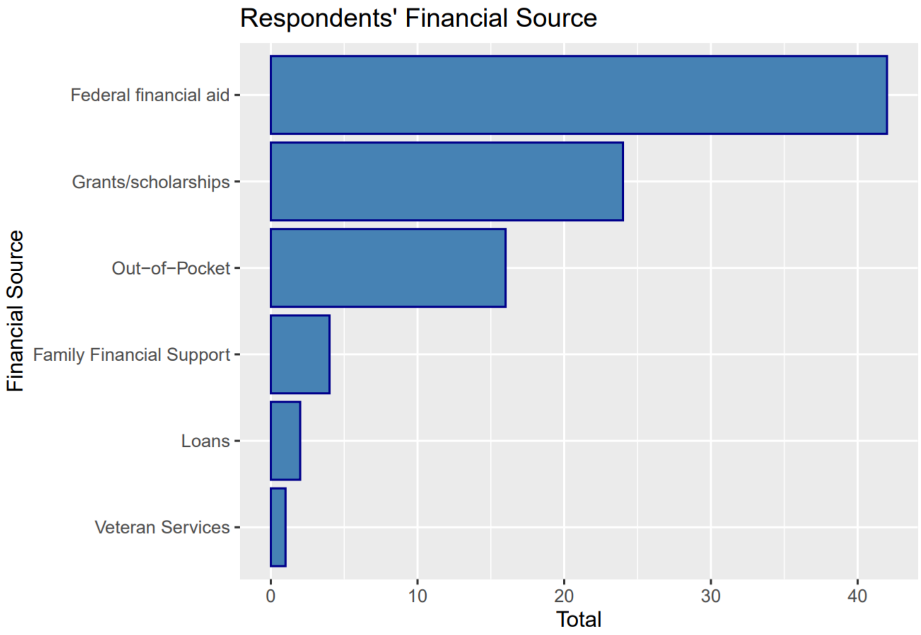

Figure 2.4: Respondents' Reported Financial Sources

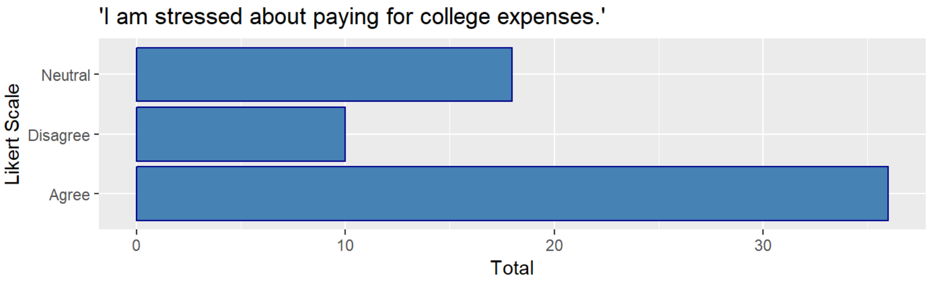

Figure 2.5: Likert Scaled Question Regarding Stress With College Expenses

2.2.3 Survey Questions Regarding Academic Preparedness and Support

Respondents were asked likert scaled questions regarding opinions of academic preparedness and support. Respondents were asked on a scale of Agree to Disagree whether they felt their high school prepared them academically for college. The responses for this question are mixed with 26 respondents agreeing to the statement, 13 disagreeing, and 25 feeling neutral. Respondents were then asked if they felt academically prepared to complete a bachelor's program. For this statement, 34 respondents agreed while 24 were neutral and 6 disagreed. The last question asked is whether the respondents were currently aware of the academic resources their college has available to them. Most respondents agreed to this statement, specifically 55 respondents are actively aware of the available academic resources. On the other hand, only 2 respondents disagreed to this statement and 7 were neutral. Figure 2.6 depicts the breakdown of responses of each question.

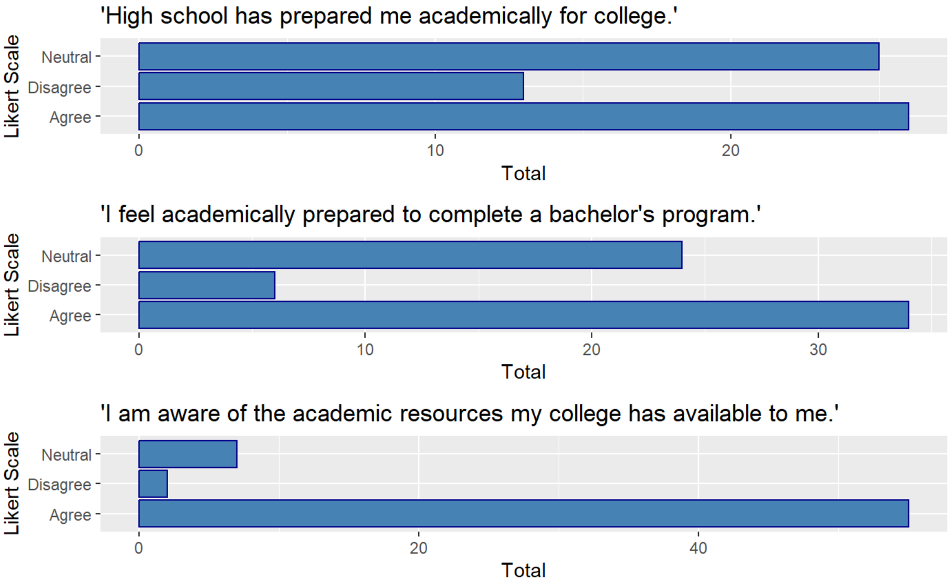

Figure 2.6: Likert Scaled Questions Regarding Academic Preparedness

Two questions were asked regarding a college support program. The respondent was first asked if they were currently enrolled in a college support program such as Extended Opportunities Program (EOP) or Disability Support Program and Services (DSPS). If the respondent answered ”yes” that they are enrolled in a college support program, they were then asked to list the name of the program. These questions were asked to gauge how many students of the study are aware of and possibly receiving services of their college's student support programs. Of the 64 respondents, only 24 respondents, or 37.5%, stated that they are enrolled in a college support program. Figure 2.7 depicts the programs the respondents are enrolled in. The programs that respondents listed were Extended Opportunities Program and Services (EOPS), TRIO-Student Support Services (TRIO-SSS), Disability Support Program and Services (DSPS), Veteran Services Program, and CalWorks Program. Of these 24 respondents, 12 of them are enrolled in Extended Opportunities Program and Services (EOPS).

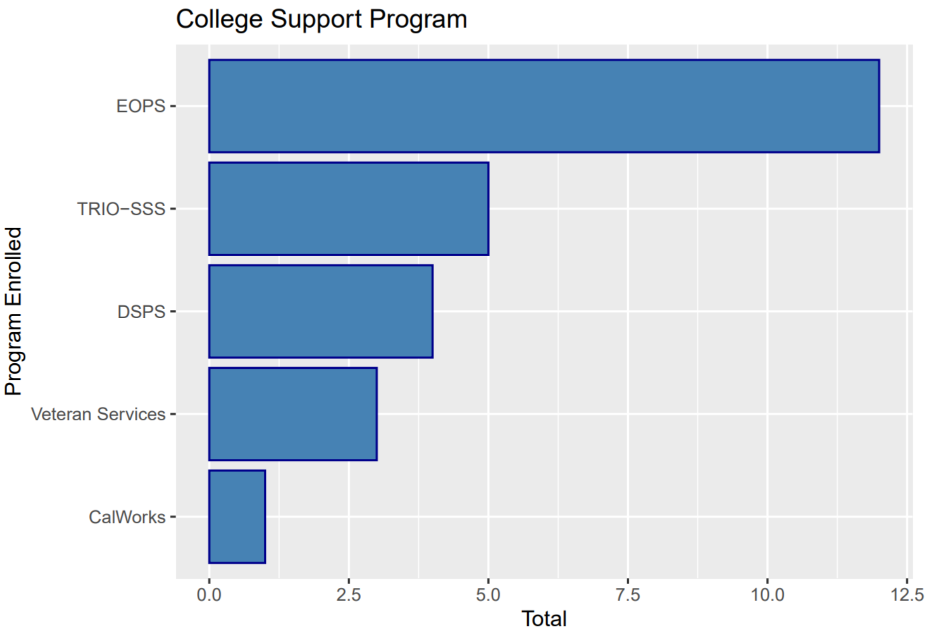

Figure 2.7: Respondents' Enrolled College Support Program

2.2.4 Survey Questions Regarding Family Culture and Support

Three likert scaled questions were asked regarding family culture and support. The respondents were first asked to select from the likert scale of agreement levels about the statement that family has a higher priority than education. Although about half of the respondents agreed to this statement, almost the other half were neutral. Respondents were then asked if their family encourages them to pursue a college education. An overwhelming 62 respondents agreed that their family encourages them to pursue a college education while 2 were neutral. Lastly, respondents were asked to select an opinion on whether their family helps them financially with their college expenses. The results are mixed with 39 respondents agreed to the statement but 18 disagreed and 7 were neutral. Figure 2.8 depicts the results of these questions.

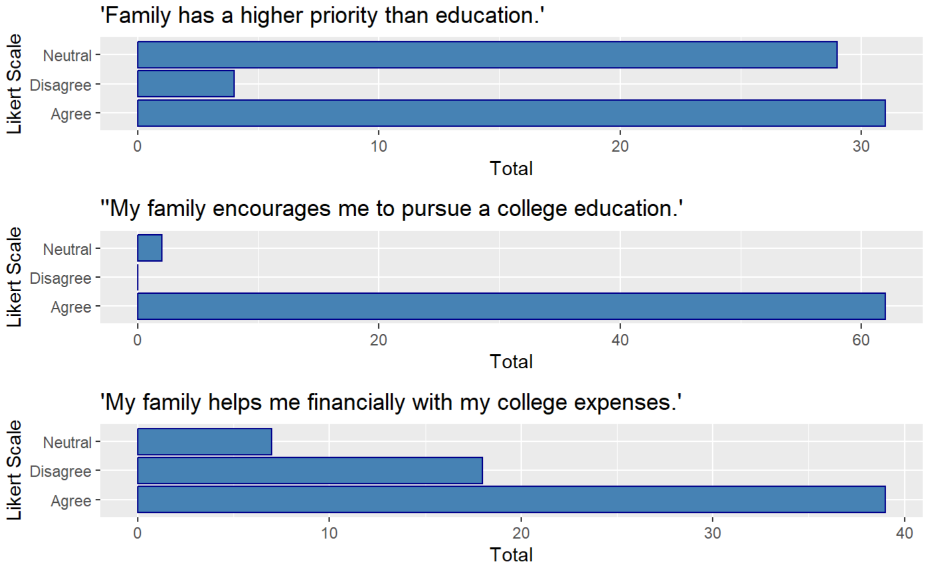

Figure 2.8: Likert Scaled Questions Regarding Family Culture and Support

2.2.5 Survey Questions Regarding Perception of Transfer Process

Respondents were asked to provide an opinion about three likert scaled questions regarding the transfer process to a four-year university. Respondents were first asked to select an opinion about whether they are aware of the transfer application process to which half the respondents agreed and the other half were split between disagreeing and being neutral about the statement. The next statement the respondents were asked to provide an opinion was whether they are taking the classes need to meet all the coursework requirements to qualify to transfer. 50 respondents stated that they agreed to the statement. The last statement was whether the academic counselors and course instructors at their college are informative about the requirements of the transfer process. The responses were mixed with over half of the respondents agreed, 11 disagreed, and 16 were neutral. Figure 2.9 depicts the results of these survey questions.

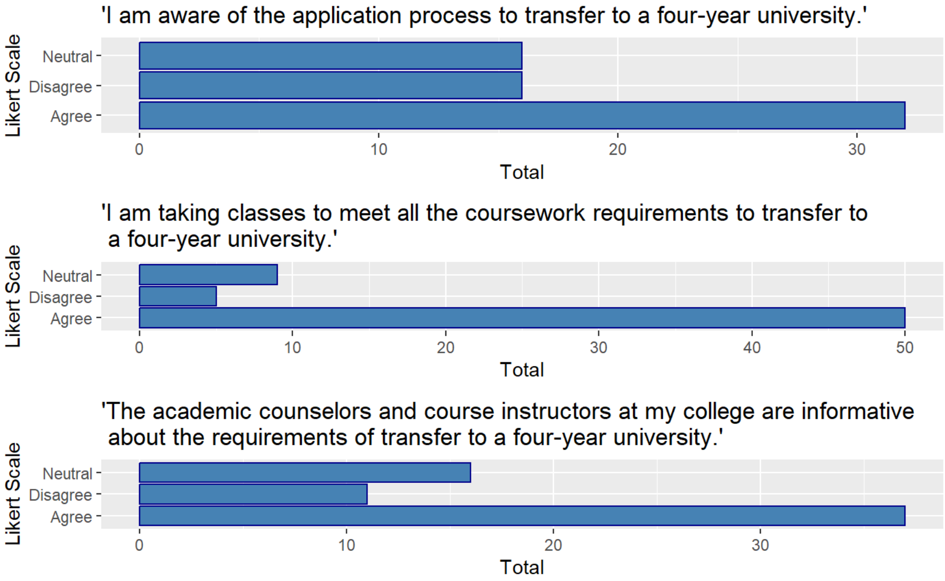

Figure 2.9: Likert Scaled Questions Regarding Perception of Transfer Process

Two likert scaled questions were asked about some factors one might consider when transferring to a four-year university. The respondent was asked to select an opinion about the belief that distance between home and a four-year university is an important factor in the decision to transfer to a four-year university. Out of 64 respondents, over half of the respondents agree that distance is indeed an important factor to consider. Respondents were also asked to provide an opinion about the belief that classes at a four-year university are more rigorous compared to classes they are currently taking at their college. Not one respondent disagreed to this statement but rather 49 respondents agreed and 15 were neutral about the statement. Figure 2.10 depicts the results of these two questions.

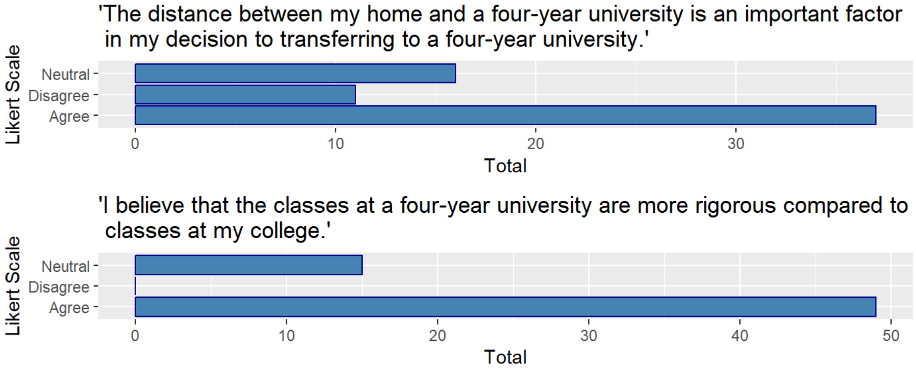

Figure 2.10: Likert Scaled Questions Regarding Factors to Consider When Transferring

2.3 Text Mining Analysis

Four open-ended questions were asked so that text mining analysis could be applied to the responses. Each question contains 64 responses at most and since the respondent was asked to provide a short answer, the length of each of the 64 responses are less than 100 words. The 64 responses were tokenized into individual terms and aggregated into one document. In that one document, the text mining techniques are applied in order to determine common themes, and thus, common answers of the survey question. On each question, word frequencies were calculated and depicted as a word cloud first to determine possible themes of the responses at a glance.

Normally in text mining analysis, a deep cleaning of the text is required prior to applying any text mining techniques. The text is cleared of all punctuation and stopwords as well all terms are converted to lower case letters. One must also stem the document which means that we remove any inflection of words, such as verb tenses, and return it to its root form. For example, words such as ”excite” and ”exciting” are interpreted by the computer as two separate terms although an individual knows that these two words are the same conceptually. When stemming the document, the word ”exciting” will be reduced to its root form, ”excite”. However, in this study, deep cleaning was not initially applied due to the low amount of terms. Instead, a Term Frequency-Inverse Document Frequency, or TF-IDF, was calculated as a way to extract the most important words of the document without risking loss of any terms. The TF-IDF takes the term frequency and the inverse document frequency, and multiplies it together to return a score. This score is considered a weight of importance.

The TF-IDF score is calculated to identify words that are the most important in a document. A unique word will be given more weight since it implies that it must be important to the document if it only appeared in said document.

After calculating the TF-IDF scores of the document, bigrams, or a two-word sequence of words, were then analyzed. When keeping the context of the study in mind, some words paired together may contain a greater significance compared to looking at terms individually. For example, if a respondent describes that a major academic challenge they face is that they experience test anxiety, the bigram ”test anxiety” is more significant together than analyzed separately as ”test” and ”anxiety”. Thus, the document of responses were stemmed and tokenized into bigrams and bigram frequencies were calculated to determine the most common bigrams found in the document. Bigrams with stop words were also removed from the set of bigrams since common phrases that are used in everyday language and sentence structure will always appear the most. Thus, we remove the stop words so we can pinpoint the commonly used bigrams or even unique bigrams. The most common bigrams will imply the topics that were most commonly found within the survey responses. Bigram correlation plots were also drawn which depicts any positive correlation between any terms found in the document. In other words, the correlation plot will depict any words that will most likely appear together as a bigram. Finally, sentiment analysis of the responses were applied. In the tokenized set of words, each word was classified with a sentiment of positive or negative, as well as one of eight emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. Determining an overall sentiment of the survey responses will provide insight of the overall sentiment of the 64 respondents.

The open-ended questions are as follows:

- ”What is your main motivation for attending college? Please explain.”
- ”What are your major academic challenges you are facing right now? Please provide at least one example.”
- ”What do you believe would be the greatest obstacle(s) for you once you attend a four-year university? Please explain.”
- ”What resources do you feel your college should have to better aid you in your path toward attending a four-year university?”

2.4 Regression Modeling

2.4.1 Research Question

Regression was applied to explore the research question: Is there a relationship between current academic challenges, perception of the transfer process, or demographics with the decision to attend a four-year university? The response variable in this study is the participant's response to whether they plan to attend a four-year university. The possible responses to this survey question are ”yes” or ”no”. Table 2.10 depicts that of the 64 responses, 52 participants responded ”yes” to the question of planning to attend a four-year university while 12 participants responded ”no”. Since there are only two possible responses, the response variable is binary. With a binary response variable, logistic regression modeling is applied to determine the model that best fits the data.

Plan to Attend	Total
Yes	52
No	12

Table 2.10: Response Variable: Plan to Attend 4-Year College

2.4.2 Exploratory Data Analysis for Modeling

Since the survey consists of 36 questions, there are 36 potential variables to consider. However, not all questions were designed to be converted to a variable. Of the 36 variables, there are 21 categorical explanatory variables and 2 numeric explanatory variables. Although the list of variables narrows down to 24 total potential explanatory variables, there are still too many variables to consider for regression modeling. Thus, dimension reduction methods should be applied to narrow down the number of variables for the model.

First, a Chi-Square Test for Significance was applied between all the categorical explanatory variables and the response variable to determine if any of the categorical explanatory variables have a significant relationship with the response variable. Table 2.11 depicts the variables that have a p-value less than 0.05, which would indicate that there may be a significant relationship with the response variable. The categorical explanatory variables with a significant relationship with the response variable are *likert-family-financial*, *likert-classes-require*, *likert-college-informative*, *likert-distance*, and gender.

Explanatory Variable	X^2-Value	P-Value
likert_family_financial	8.8463	0.0120
likert_classes_require	11.998	0.0025
likert_college_informative	6.3867	0.0410
likert_distance	6.9673	0.0307
gender	10.154	0.0062

Table 2.11: Chi-Square Test for Significance with Response Variable

A Random Forest Classification was then applied for a more efficient and accurate representation of significant explanatory variables. According to the Random Forest generated in Figure 2.11, it appears that the following are the most significant predictors: *likert_classes_require*, *likert_family_financial*, age, gender, GPA, *likert_college_informative* and *likert_distance*. This is consistent with the results of the Chi-Square Tests for Significance Between Response Variable and Categorical Explanatory Variables done previously. The modeling will focus on this list of explanatory variables.

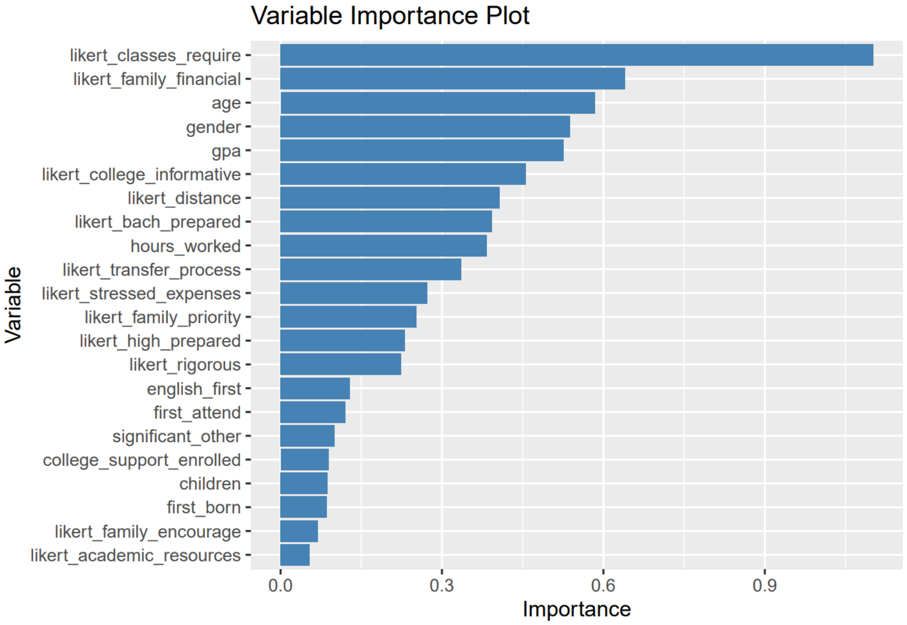

Figure 2.11: Random Forest Variable Importance Plot

Before proceeding with the modeling, a Chi-Square Test for Independence was performed between the four categorical explanatory variables to determine if these variables are independent of each other. Table 2.12 depicts the results of the Chi-Square Tests for Independence. According to the results of the Chi-Square Test some variables are not independent of each other and so there is a possibility of multicollinearity existing in the final model. Of the list of categorical variables, some variables may not be included in the final model due to collinearity, so we proceed with caution with the modeling.

Variable 1	Variable 2	X^2-Value	P-Value
likert_classes_require	likert_family_financial	9.6417	**0.0469**
likert_classes_require	gender	4.5758	0.3337
likert_classes_require	likert_college_informative	6.7946	0.1472
likert_classes_require	likert_distance	11.8060	**0.0189**
likert_family_financial	gender	6.6514	0.1555
likert_family_financial	likert_college_informative	13.3390	**0.0097**
likert_family_financial	likert_distance	9.5050	**0.0496**
gender	likert_college_informative	3.5378	0.4722
gender	likert_distance	4.2908	0.3681
likert_college_informative	likert_distance	10.552	**0.0321**

Table 2.12: Chi-Square Test for Independence Between Categorical Explanatory Variables

For the numeric explanatory variables, a correlation matrix was made to determine if there was any correlation between age and GPA. According to the correlation matrix in Table 2.13, age and GPA have a correlation value of 0.11. This is a low correlation between age and GPA and should not be of concern.

	age	gpa
age	1.0000	0.1159
gpa	0.1159	1.0000

Table 2.13: Correlation Matrix of Numeric Explanatory Variables

Also, boxplots between the numeric explanatory variables and the response variable were created to determine if there are any skewness or outliers that would need to be addressed in the final model analysis. According to the boxplot between GPA and the response variable in Figure 2.13, there is no skewness and only one outlier with the respondents who answered "yes" in their plan to attend a four-year university. In the boxplot between age and the response variable in Figure 2.12, there is also no major skewness but there are multiple outliers with respondents who answered "yes". We would need to conduct outlier and influential points tests if the age or GPA variables are included in the final model.

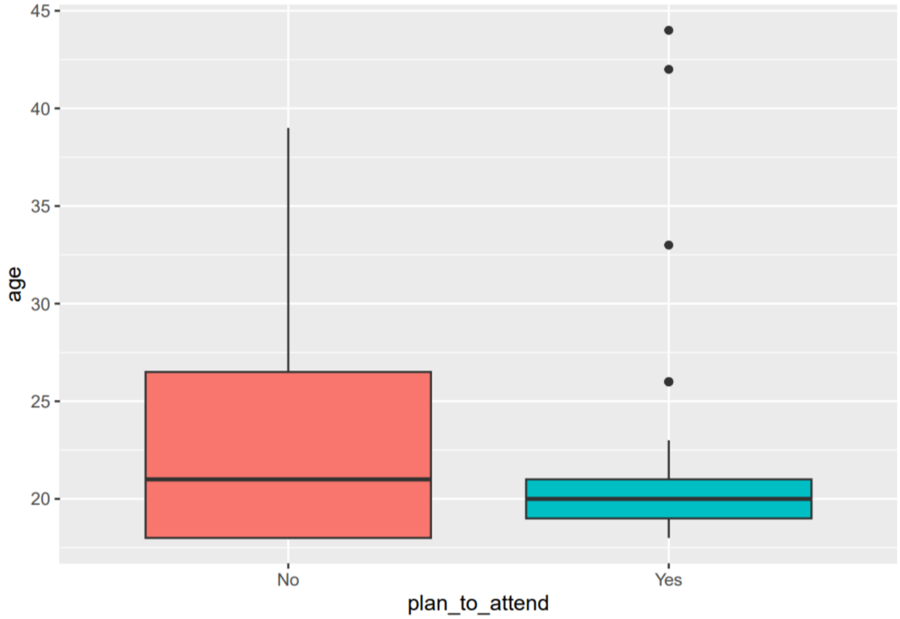

Figure 2.12: Boxplot of Age vs Response Variable

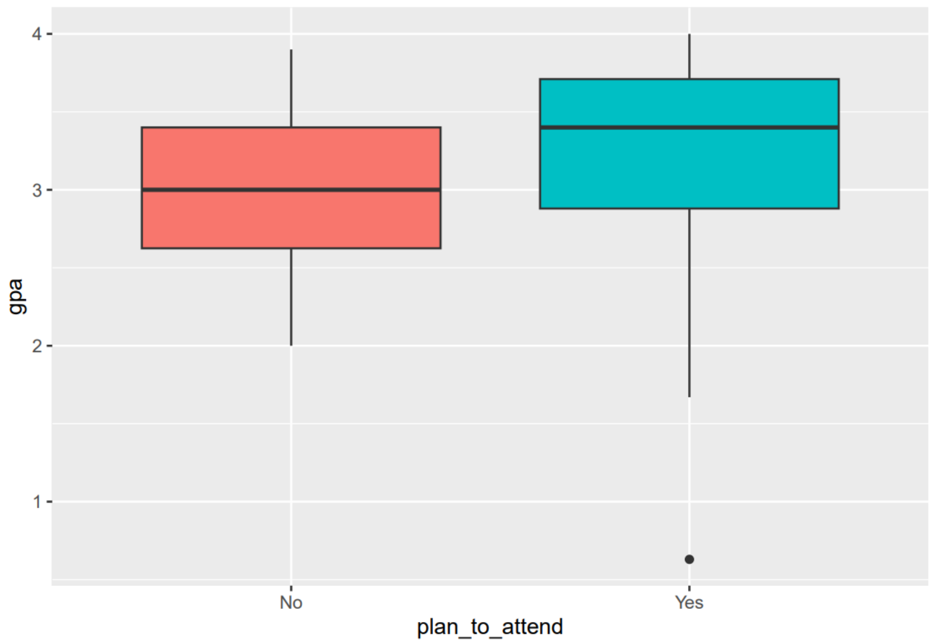

Figure 2.13: Boxplot of GPA vs Response Variable

2.4.3 Logistic Regression Model

The modeling began with an initial model that contains all the significant explanatory variables with respect to the response variable, Model 2.1.

$$\begin{aligned}\boldsymbol{plan\text{-}to\text{-}attend} = {} & \boldsymbol{likert\text{-}classes\text{-}require} + \boldsymbol{age} + \boldsymbol{gpa} \\ & + \boldsymbol{likert\text{-}family\text{-}financial} + \boldsymbol{gender} \\ & + \boldsymbol{likert\text{-}college\text{-}informative} \\ & + \boldsymbol{likert\text{-}distance}\end{aligned} \tag{2.1}$$

According to the initial model summary in Table 2.14, it appears that there are no significant predictors with all p-values being greater than 0.05. We then calculate the Variance Inflation Factor value or VIF value. If the VIF value is greater than 5, then there is a presence of multicollinearity. After calculating VIF, we see that there is a high collinearity with *likert-classes-require*, *likert-family-financial*, *likert-college-informative*, and *likert-distance*. We don't need to be concerned with age, GPA, and gender. This, however, is consistent with the results of the correlation matrix with age and GPA (Table 2.13). However, we should remove either *likert-classes-require* or *likert-family-financial* from the model due to the extremely inflated Variance Inflation Factor values found in Table 2.15.

Variable	Coefficient Estimate	P-Value
Intercept	20.6603	0.9969
likert_classes_requireDisagree	-45.0797	0.9966
likert_classes_requireNeutral	-22.4134	0.9967
age	0.2325	0.3633
gpa	-0.3292	0.8417
likert_family_financialDisagree	-17.764	0.9974
likert_family_financialNeutral	-20.7367	0.9969
genderMale	-3.5733	0.0551
genderNon Binary	-45.6116	0.9992
likert_college_informativeDisagree	64.818	0.9964
likert_college_informativeNeutral	-3.1452	0.3016
likert_distanceDisagree	-2.9163	0.2631
likert_distanceNeutral	2.0125	0.6047

Table 2.14: Initial Model Summary

Variable	VIF Value
likert_classes_require	27396554.01
age	2.41
gpa	2.63
likert_family_financial	31550990.07
gender	1.64
likert_college_informative	10.47
likert_distance	11.43

Table 2.15: Initial Model Variance Inflation Factor

2.4.4 Final Model

This process of elimination was applied to narrow down the variables for the model. One variable was removed and then the VIF values were calculated to determine if multicollinearity was still present. This process continued until there was no high multicollinearity. Interactions between the variables of the model were considered but none of the interaction terms were significant to the response variable. After this process, the finalized model is the following:

$$\begin{aligned}\boldsymbol{\textit{plan-to-attend}} = \boldsymbol{\textit{likert-family-financial}} + \boldsymbol{\textit{likert-distance}} \\ + \boldsymbol{\textit{likert-classes-require}}\end{aligned} \tag{2.2}$$

Table 2.16 depicts the summary for the final model, Model 2.2. The intercept, *likert-family-financial*, and *likert-classes-require* have significance toward the response variable, with p-values less than 0.05.

Variable	Coefficient Estimate (Not Transformed)	P-Value
Intercept	3.1880	**0.000645**
likert_family_financial Disagree	0.2637	0.8504
likert_family_financial Neutral	-3.5310	**0.0231**
likert_distance Disagree	-0.9209	0.4606
likert_distance Neutral	0.4727	0.7370
likert_classes_require Disagree	-2.2144	0.0889
likert_classes_require Neutral	-3.7111	**0.0105**

Table 2.16: Final Model Summary

As seen with Table 2.17, the Variance Inflation Factors of the variables are less than 5 and thus, multicollinearity is no longer an issue.

Variable	VIF Value
likert_family_financial	2.3680
likert_distance	2.3263
likert_classes_require	2.4722

Table 2.17: Final Model Variance Inflation Factor

Table 2.18 depicts the confusion matrix of the final model. Based on the confusion matrix, the accuracy was calculated to be 0.85.

	Actual		
		No	Yes
Prediction	No	6	5
	Yes	4	45

Table 2.18: Final Model Confusion Matrix

2.4.5 Logistic Regression with LASSO Model

An alternative logistic regression method was applied to address with the large amount of variables involved. Instead of manually narrowing down the variables with a random forest classifier, a logistic regression model with Least Absolute Shrinkage and Selection Operator, or LASSO, was applied. LASSO regression is a penalized form of regression that gives weights to variables based on variable importance. This model absorbs all the variables as inputs and generates a model with the most significant variables with larger weights. Through cross validation, a tuning parameter to control bias-variance tradeoff, known as lambda, is estimated to provide the model with the best fit coefficients. Table 2.19 displays the LASSO logistic model's selected variables with its corresponding non-transformed coefficients.

Variable	Coefficient (Not Transformed)
Intercept	2.4618
likert_family_financial Neutral	-1.2686
gender Male	-0.4845
gender Non-Binary	-1.8308
age	-0.0110
children Yes	-0.0485
hours_worked I am not working	-0.0963
likert_classes_require Disagree	-0.6777
likert_classes_require Neutral	-1.6044
likert_college_informative Neutral	-0.3587

Table 2.19: LASSO Logistic Model Coefficients

CHAPTER 3

Results

3.1 Logistic Regression Results

3.1.1 Final Model

The LASSO logistic model is too complex to interpret, so my final logistic regression model is still Model 2.2. Table 3.1 depicts the coefficient estimates exponentially transformed so it can be interpreted. Based on the exponentiated coefficients, respondents who selected the neutral option of the statement that family supports the respondent financially have 0.03 times the odds of planning to attend a four-year university. Similarly, respondents who were neutral about the statement that they are currently taking the classes required to transfer to a four-year have 0.02 times the odds of planning to attend a four-year university. Meanwhile, respondents who **agreed** to the statements that their family provides financial support, they are taking the courses required to transfer to a four-year university, **and** distance is an important factor when considering to attend a four-year university, have **24 times** the odds of planning to attend a four-year university. The respondents who selected agree to these three statements by far have higher odds to plan to attend a four-year university.

Variable	Transformed Coefficient Estimate	P-Value
Intercept	24.2402	**0.000645**
likert_family_financial Disagree	1.3017	0.8504
likert_family_financial Neutral	0.02928	**0.0231**
likert_distance Disagree	0.3981	0.4606
likert_distance Neutral	1.6043	0.7370
likert_classes_require Disagree	0.1092	0.0889
likert_classes_require Neutral	0.0244	**0.0105**

Table 3.1: Final Model Summary with Transformed Coefficients

3.2 Text Mining Results

3.2.1 Survey Question 1: What is your main motivation for attending college?

3.2.1.1 TF-IDF Analysis

Figure 3.1 depicts the most common words found in the 64 responses without cleaning the document or calculating the TF-IDF scores. At first glance of the responses of this question, the words ”education”, ”career”, ”degree”, ”programs”, and ”stable” are appearing as the most common. When calculating the TF-IDF scores of each term, the words ”nurse”, ”md”, ”feel”, ”opportunities”, and ”parents” are appearing to contain the most weight and are deemed the most important in the survey responses. The term ”md” refers to the designation of Medical Doctor or ”M.D.”. Figure 3.2 depicts the top 15 TF-IDF scores of the survey questions regarding the respondents’ motivation for attending college. The terms that respondents are using to describe their main motivation for attending college are related to higher education, career opportunities, and family influence.

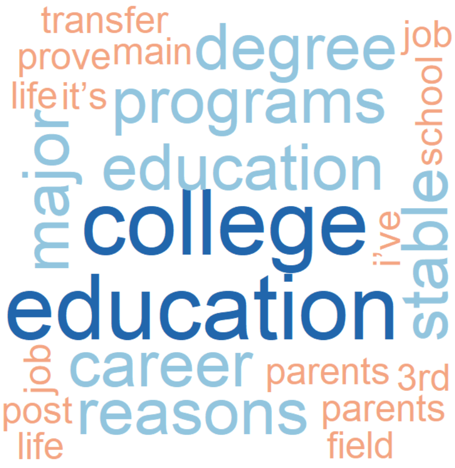

Figure 3.1: Word Cloud About College Motivation

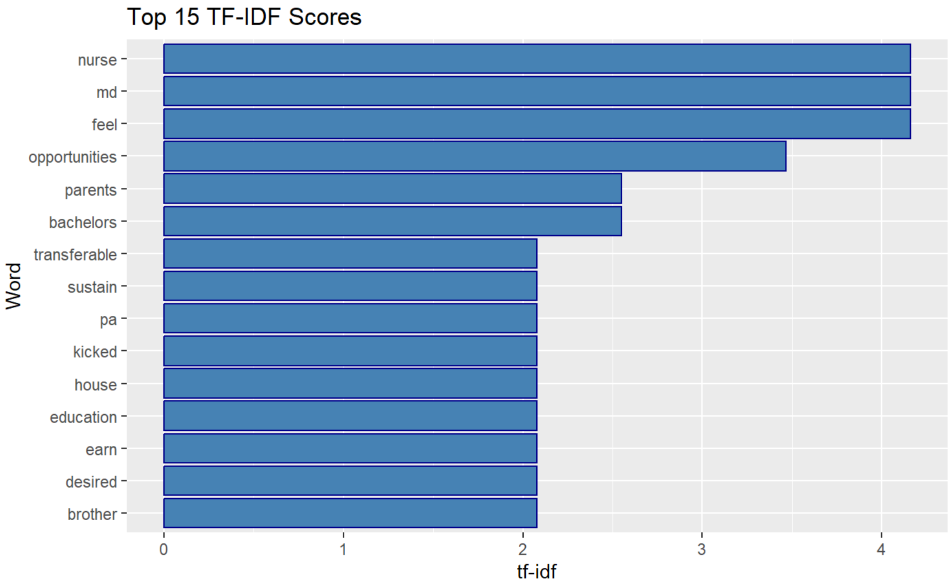

Figure 3.2: Top 15 TF-IDF Scores for College Motivation Question

3.2.1.2 Bigram Analysis

The responses of this question were stemmed and tokenized into bigrams. After cleaning the bigrams by removing all stop words, the most common bigrams to appear in the responses are "attending college" , "college education", "career field", and "bachelors degree" (Figure 3.3). Furthermore, according to the bigram correlation plot in figure 3.4, there appears to be three large clusters of words that will most likely appear together as bigrams. One cluster depicts the word "stable" being connected to "future", "chosen", "permanent", and "paying", possibly implying a theme of job stability. Another cluster appears to describe possible family influence since the word "parents" is connected to "family" and "family" is connected to the words "achieve", "obtain", "apply", "spend", "earning", and "expand". The third cluster contains words related to education with the word "college" being connected to "bachelors", "transferable", "entry", and "elementary".

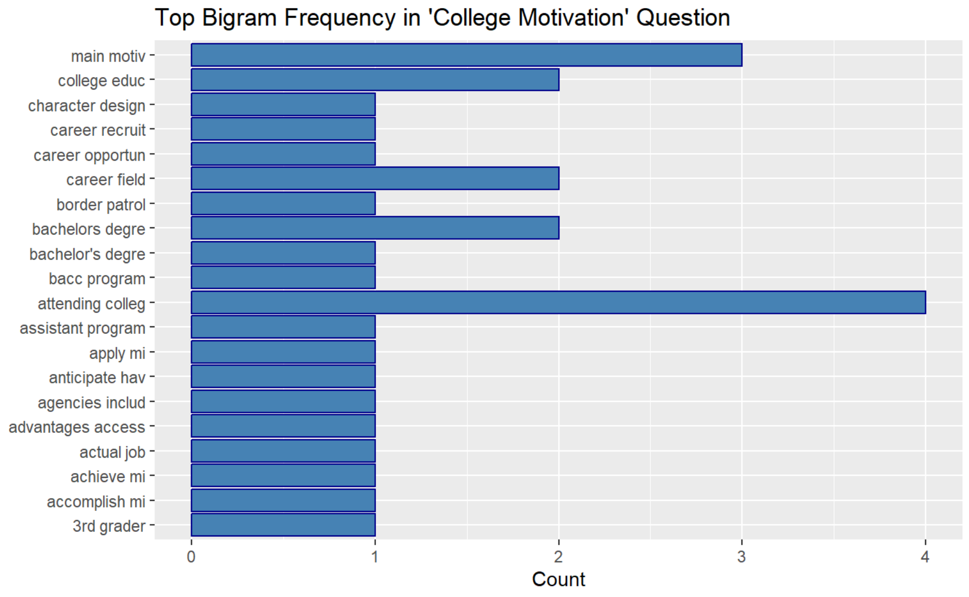

Figure 3.3: Top Bigram Frequency for College Motivation Question

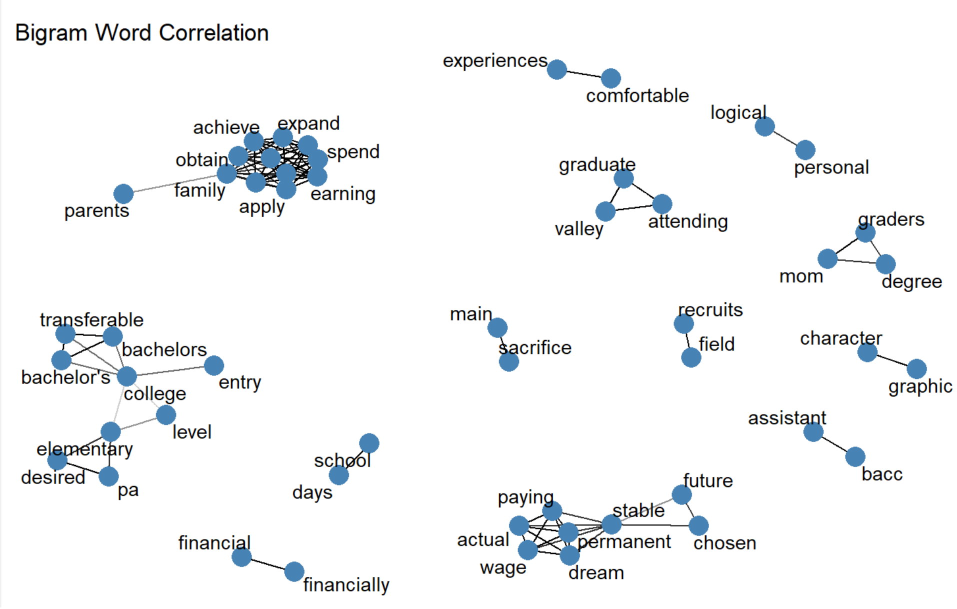

Figure 3.4: Bigram Correlation Plot College Motivation Question

3.2.1.3 Sentiment Analysis

For sentiment analysis, an overall sentiment of positive and negative were calculated on the words. In this question, the words used in the responses were mostly positive (Figure 3.5. Figure 3.6 depicts a breakdown of the positive and negative words that appeared in the question regarding college motivation. The only negative words found in the responses were ”fall”, ”enforcement”, and ”criminal”. However, adding context to this question, the words ”enforcement” and ”criminal” are related to the terms ”law enforcement” and ”criminal justice”, respectively, which are both career paths of fields of study.

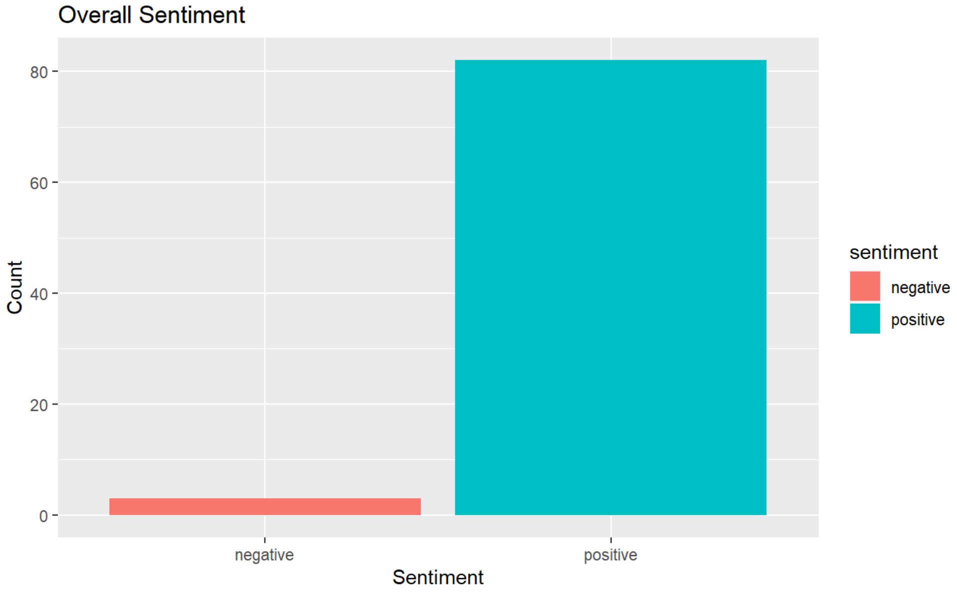

Figure 3.5: Overall Sentiment of College Motivation Question

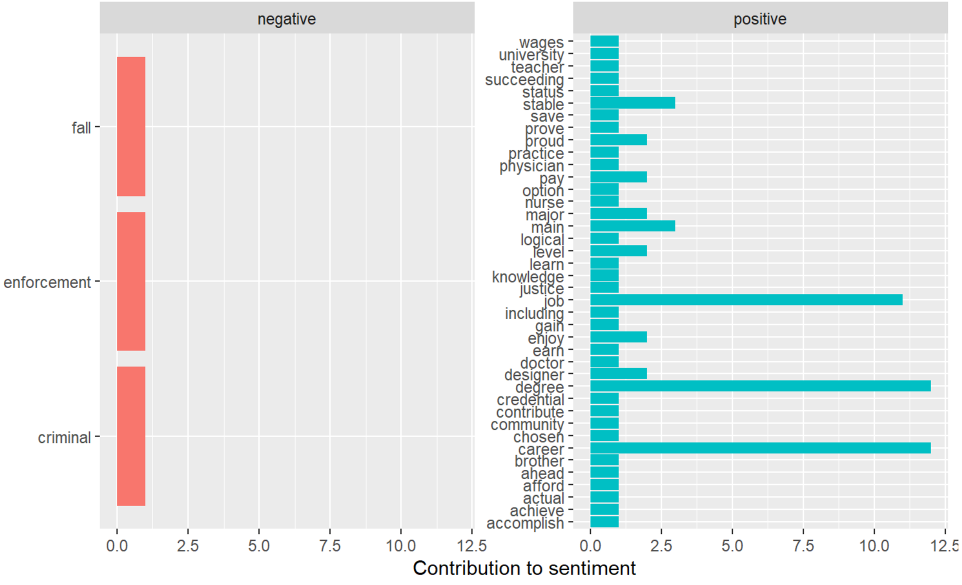

Figure 3.6: Positive and Negative Words in College Motivation Question

3.2.2 Survey Question 2: What are your major academic challenges you are facing right now?

3.2.2.1 TF-IDF Analysis

When it comes to current academic challenges, the words ”concepts”, ”classes”, ”online”, and ”transfer” are appearing to be the most common words found in the survey responses (Figure 3.7). According to the TF-IDF values in Figure 3.8, the words transportation”, ”statistics”, ”disability”, ”concentrate”, and ”amount” contain the most weight and appear to be the most important terms in the document. It appears that these are terms related to a respondent’s coursework and struggles related to that such as ”procrastination” and ”anxiety”.

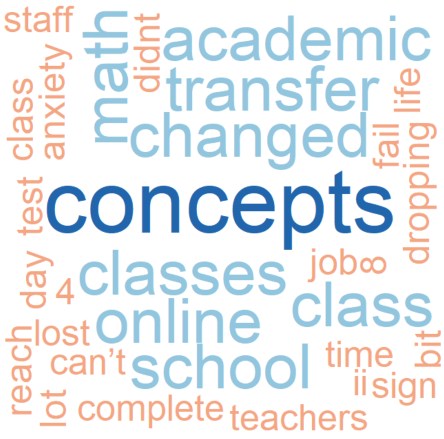

Figure 3.7: Word Cloud About Current Challenges

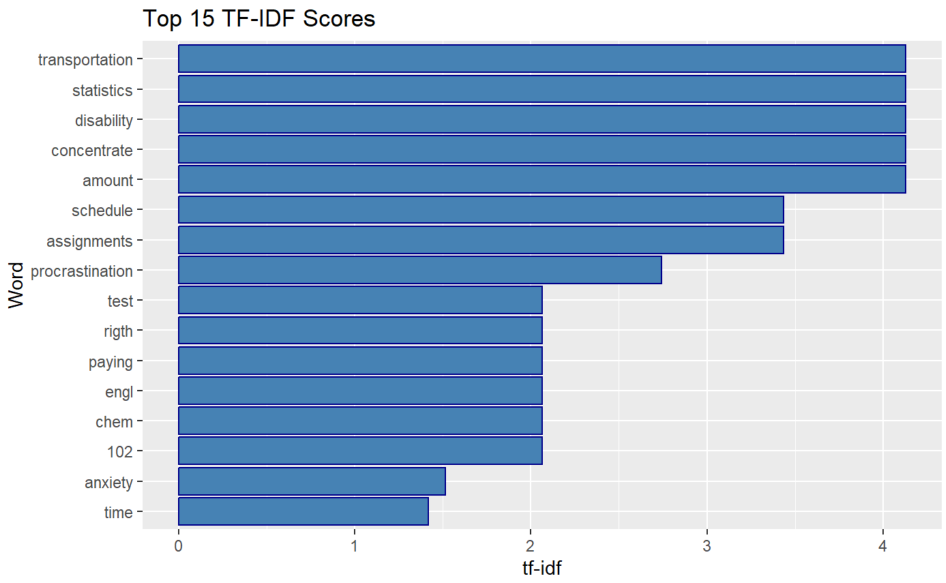

Figure 3.8: Top 15 TF-IDF Scores for Current Challenges Question

3.2.2.2 Bigram Analysis

In Figure 3.9, the top relevant bigram found in the responses of the question related to current academic challenges is the bigram ”time management”. Other bigrams that also appears often were ”online class”, ”mental health”, and ”balancing school”. These bigrams describe some of the academic challenges the respondents are stating they face today. In the bigram correlation plot (figure 3.10), there is one cluster that is most prominent. The word ”online” is connected to the words ”english”, ”politics”, ”math”, and ”chem”, and well as uniquely connected to the word ”balancing”. It appears here that respondents are also describing the classes they may be currently taking as a current academic challenge. Another interesting cluster pertains to the word ”stress”. In this small cluster, ”stress” is uniquely connected to the words ”test” and ”time”. Since ”stress” has a positive correlation and will more likely appear with these two words, it is possible that the respondents are describing stress with tests or time management.

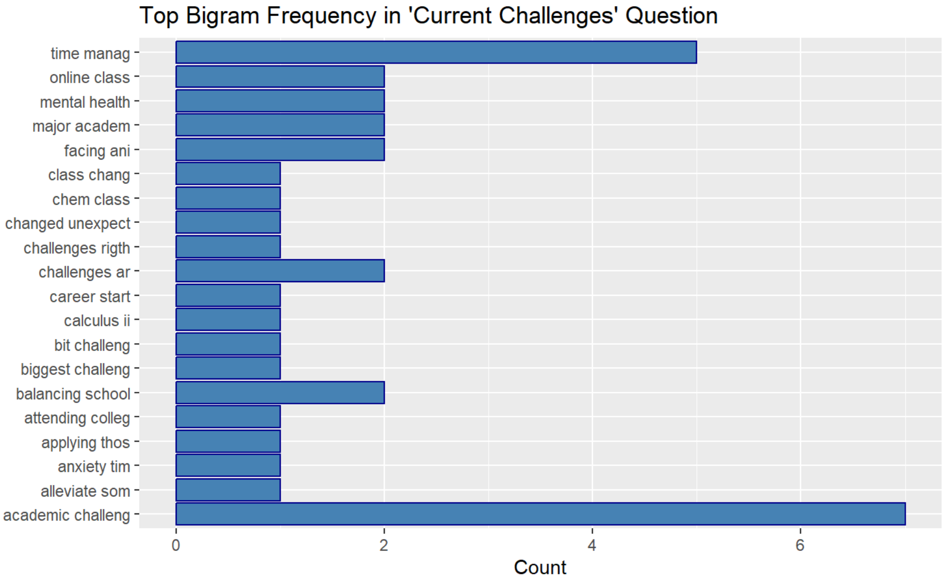

Figure 3.9: Top Bigram Frequency for Current Challenges Question

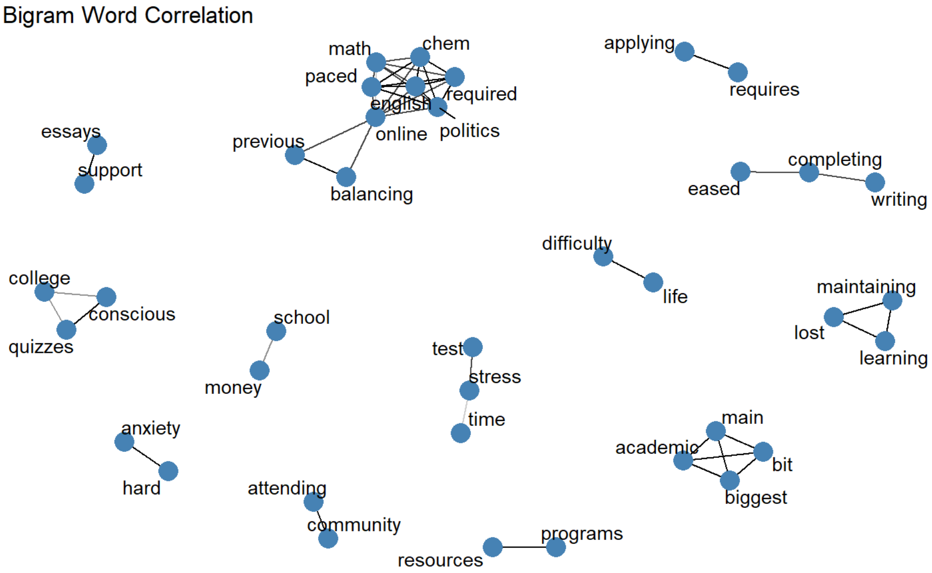

Figure 3.10: Bigram Correlation Plot for Current Challenges Question

3.2.2.3 Sentiment Analysis

For the question regarding current academic challenges, the responses have an overall positive sentiment (Figure 3.11). However, more negative words appear in these responses. It is logical that more negative words would appear in the responses of this question, since the question is about describing challenges the respondents are struggling with at the moment. Upon taking a closer look in Figure 3.12, the negative words that were used the most were ”procrastination”, ”anxiety”, ”stress”, and ”struggle”.

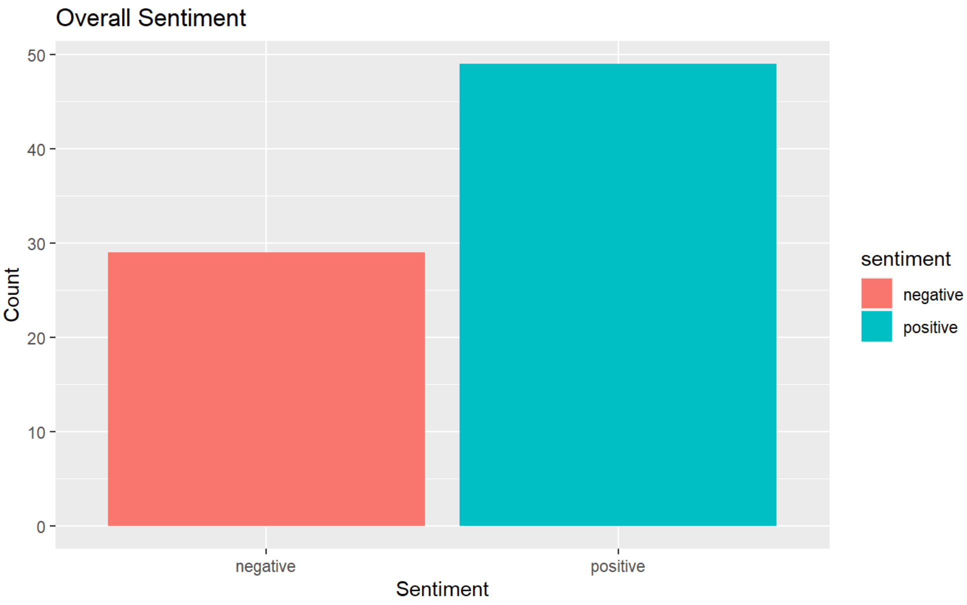

Figure 3.11: Overall Sentiment for Current Challenges Question

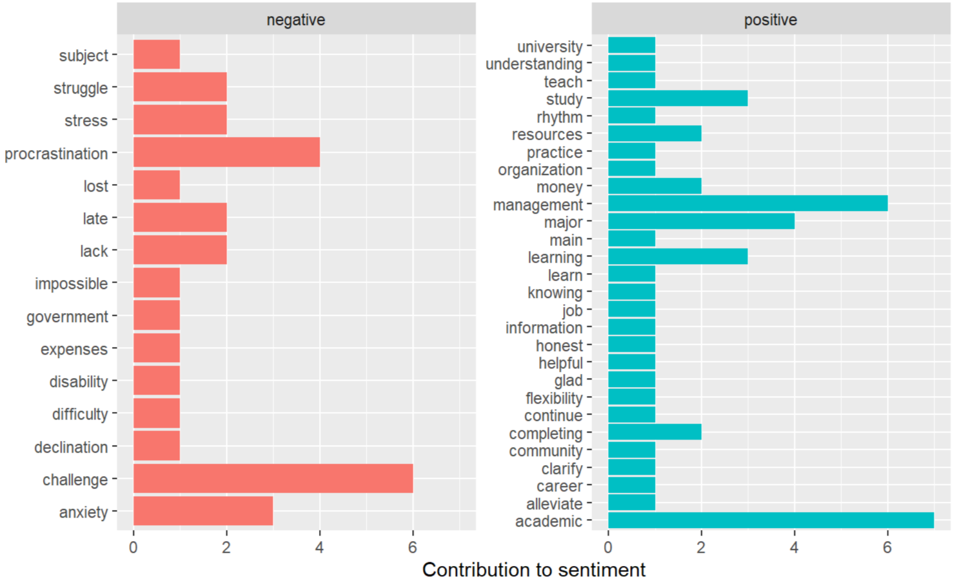

Figure 3.12: Positive and Negative for Current Challenges Question

3.2.3 Survey Question 3: What do you believe would be the greatest obstacle(s) for you once you attend a four-year university?

3.2.3.1 TF-IDF Analysis

The most common words found in the question regarding the greatest obstacle once a student attends a four-year university, as seen in figure 3.13, are the words ”transfer”, ”housing”, ”expenses”, ”home”, and ”parents”. The words with the highest TF-IDF score are ”homesickness”, ”finishing”, ”disability”, ”procrastination”, and ”motivation” (figure 3.14). It seems some words from the survey question regarding current academic challenges are appearing again such as ”procrastination”. It appears personal challenges are being described with homesickness, disability, and procrastination being prominent challenges according to the respondents.

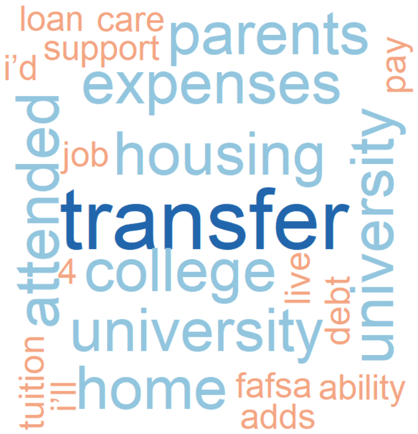

Figure 3.13: Word Cloud About Greatest Obstacles

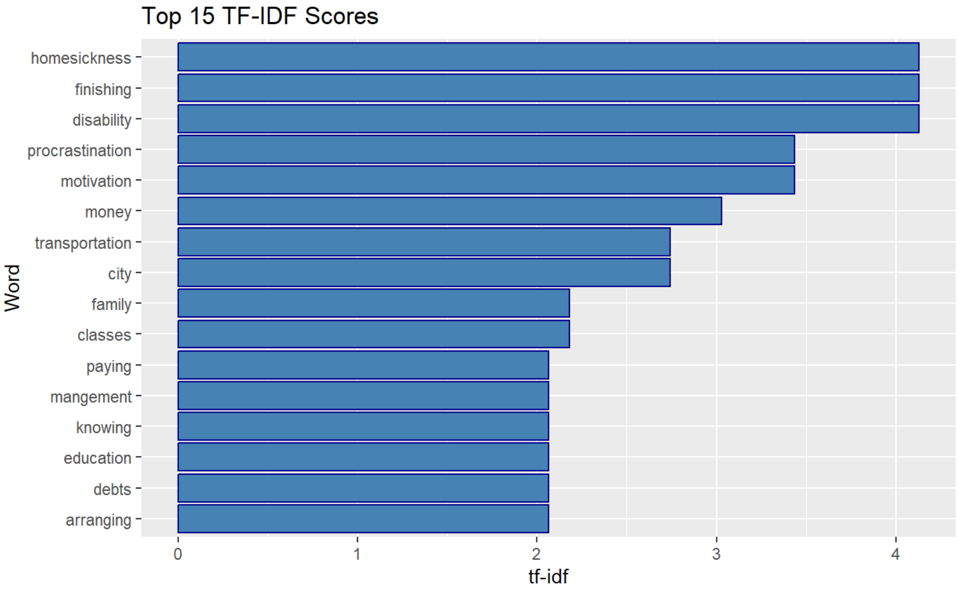

Figure 3.14: Top 15 TF-IDF Scores for Greatest Obstacles Question

3.2.3.2 Bigram Analysis

Figure 3.15 depicts the top bigrams that appeared the most in the responses to this question. The top relevant bigrams are ”time management” and ”financial aid”. Other bigrams that are also appearing are ”book expenses”, ”balancing school”, and ”attending class”. It appears that the respondents are describing obstacles related to the classes they would need to take while attending the four-year university. According to the bigram correlation plot in figure 3.16, there are a lot of small clusters. One interesting cluster is the word ”student” being connected unique to the words ”loan” and ”school”, perhaps indicated something related to student loans. Another cluster depicts the words ”travel”, ”living”, and ”book” all being connected to each other. This possibly indicates a topic related to expenses since these words serve as types of expenses a student would normally consider when attending college.

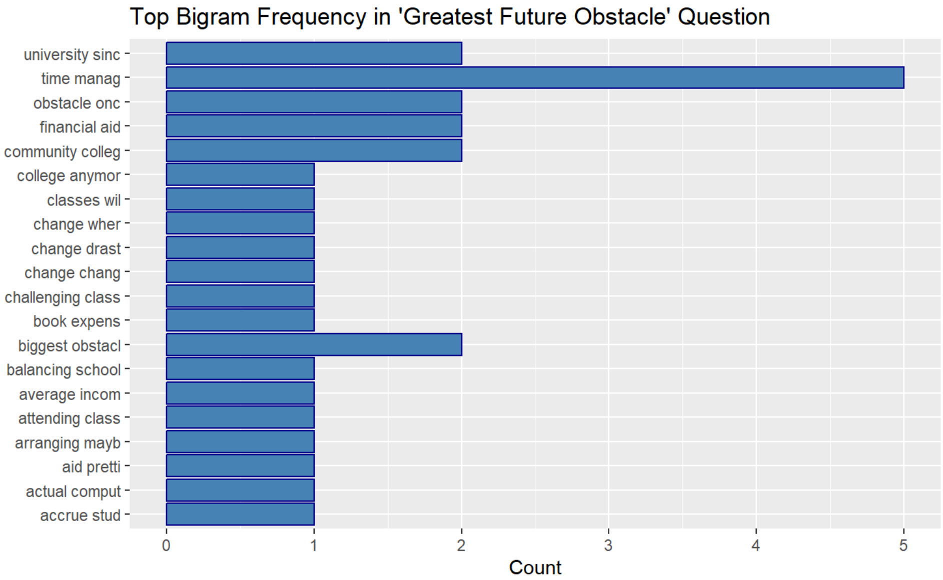

Figure 3.15: Top Bigram Frequency for Greatest Obstacles Question

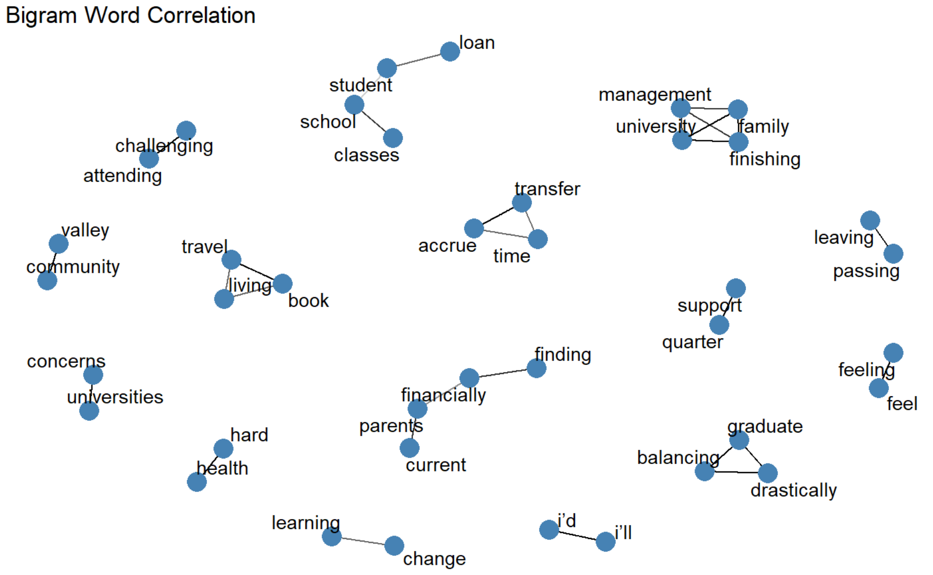

Figure 3.16: Bigram Correlation Plot for Greatest Obstacles Question

3.2.3.3 Sentiment Analysis

For this question, there is an overall positive sentiment in the responses (figure 3.17). However, in figure 3.18, there are some negative words that respondents used that is related to the obstacles they are describing. The most common negative words that are appearing are ”expenses”, ”procrastination”, ”income”, ”hardship”, ”disability”, and ”debt”.

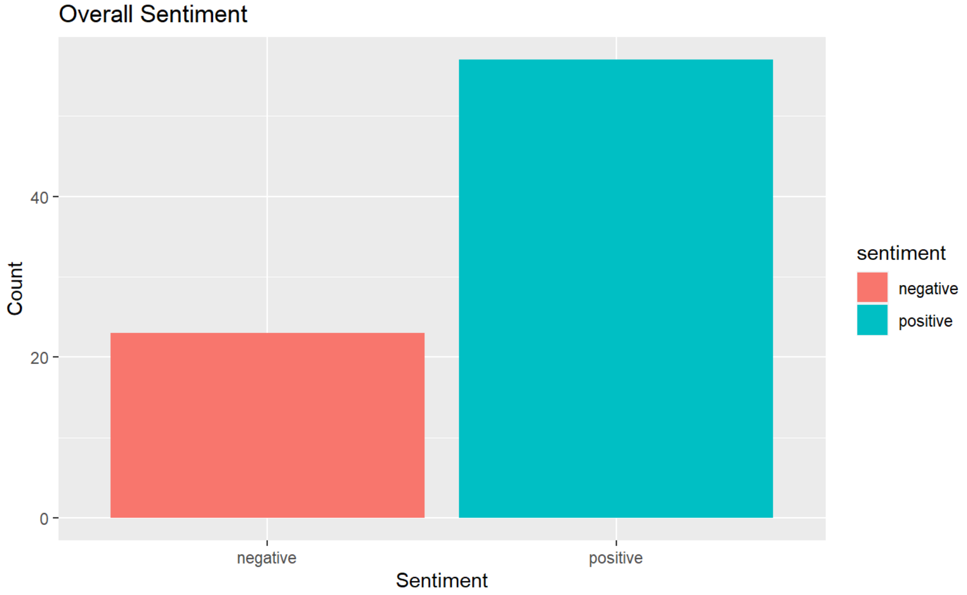

Figure 3.17: Overall Sentiment for Greatest Obstacles Question

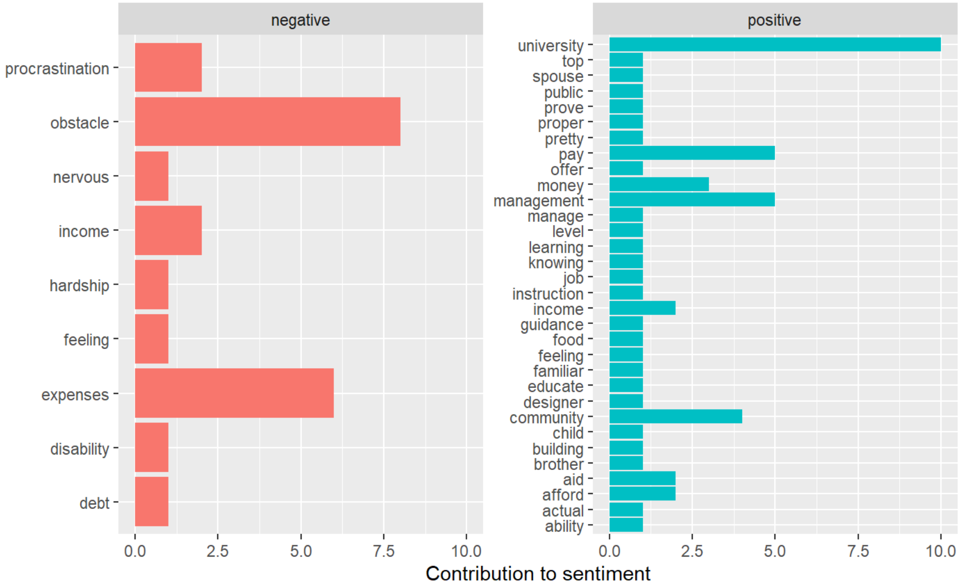

Figure 3.18: Positive and Negative for Greatest Obstacles Question

3.2.4 Survey Question 4: What resources do you feel your college should have to better aid you in your path toward attending a four-year university?

3.2.4.1 TF-IDF Analysis

The words most commonly found in the responses to the question regarding what resources should be added by the respondent's college are "college", "counseling", "career", "plan", and "time" (Figure 3.19). The words with the top TF-IDF scores were "transport", "scholarships", "internships", "counseling", "money", and "information" (Figure 3.20). Based on these words that were given the most weight and therefore are deemed the most important in the document, the respondents' are describing resources related to financial aid, career experience opportunities, and counseling.

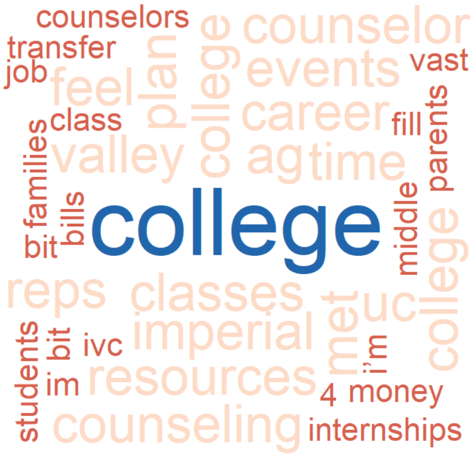

Figure 3.19: Word Cloud About Better Resources

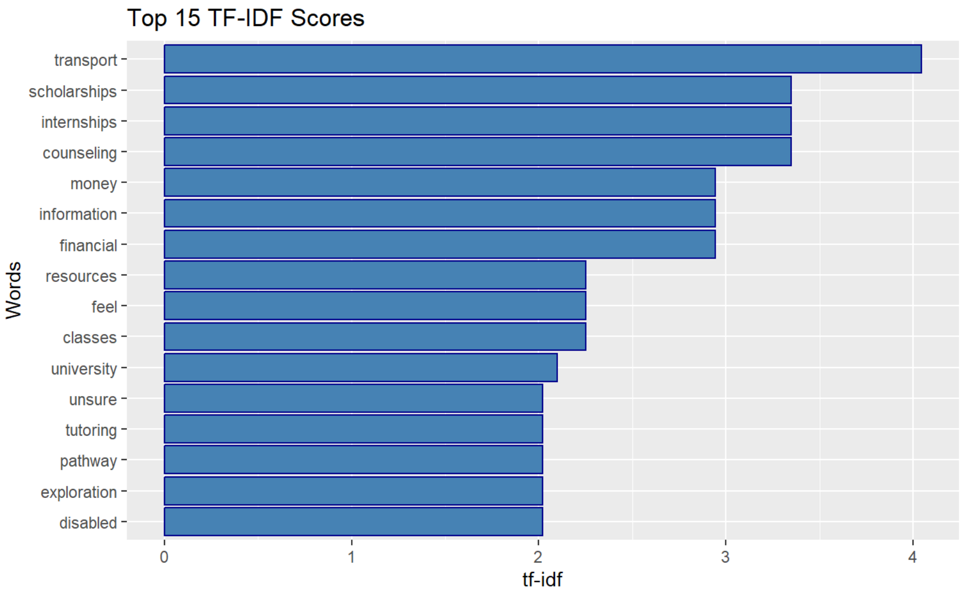

Figure 3.20: Top 15 TF-IDF Scores for Better Resources Question

3.2.4.2 Bigram Analysis

Furthermore, the top bigrams found in the responses of this question were ”financial aid”, ”educational plan”, and ”college rep” (Figure 3.21). The word ”college rep” perhaps refers to college representatives, of whom make visits to various colleges to extend information about their college in hopes of encouraging students to apply to said college. According to the bigram correlation plot in figure 3.22, there are some interesting clusters. One such cluster is the word ”transfer” being connected to the words ”guidance” and ”application”. These words may be connected to possibly describe the application process to transfer to a four-year university.

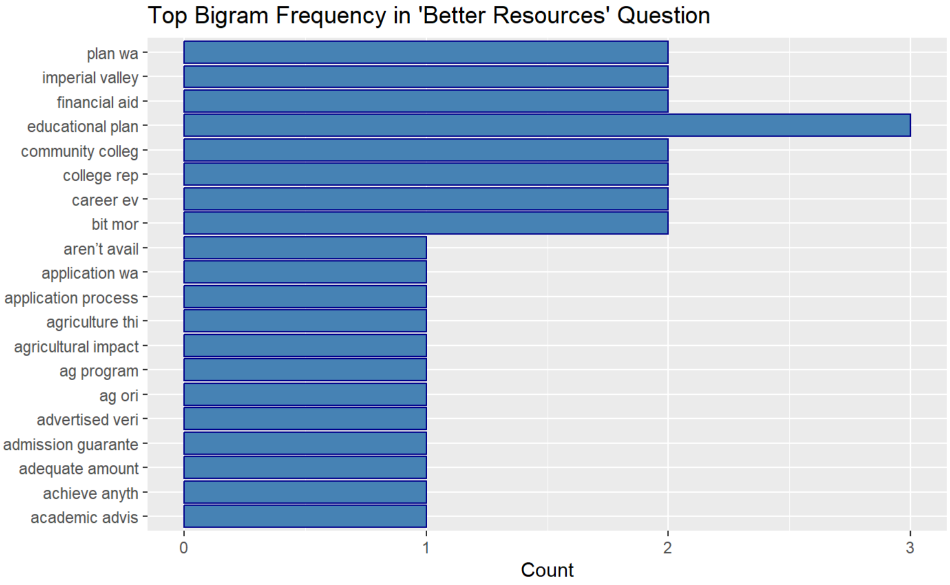

Figure 3.21: Top Bigram Frequency for Better Resources Question

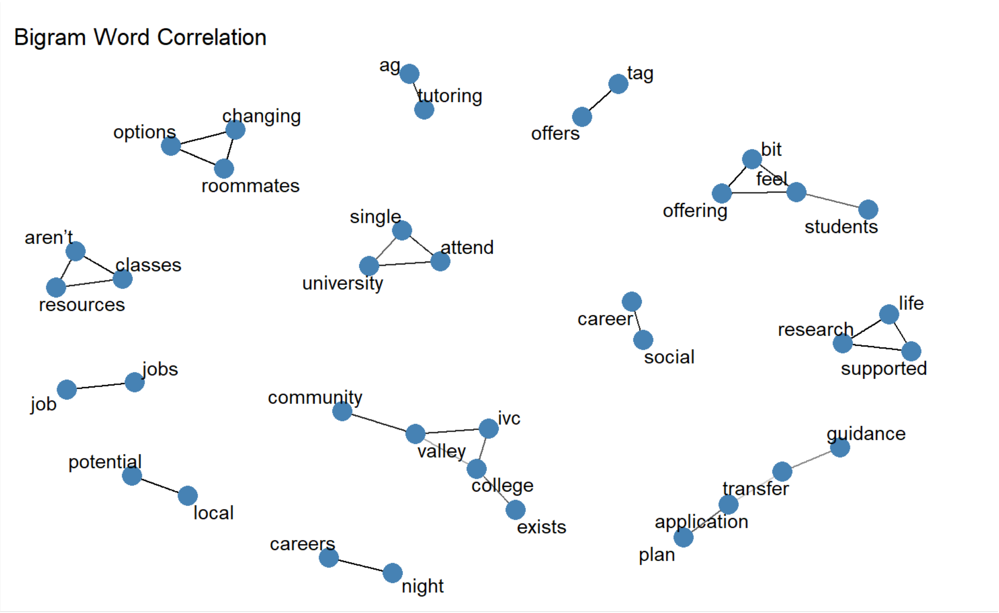

Figure 3.22: Bigram Correlation Plot for Better Resources Question

3.2.4.3 Sentiment Analysis

For the sentiment analysis portion, it appears that there is an overall positive sentiment in the responses (Figure 3.23). The only negative words that were used were ”disabled” and ”confusion” (Figure 3.24). Perhaps the respondents who used these words were referring to add resources that could address these concerns. Since respondents are describing resources that would better aid their academic journey, mostly positive words are used to describe the hope of having these resources on their campus.

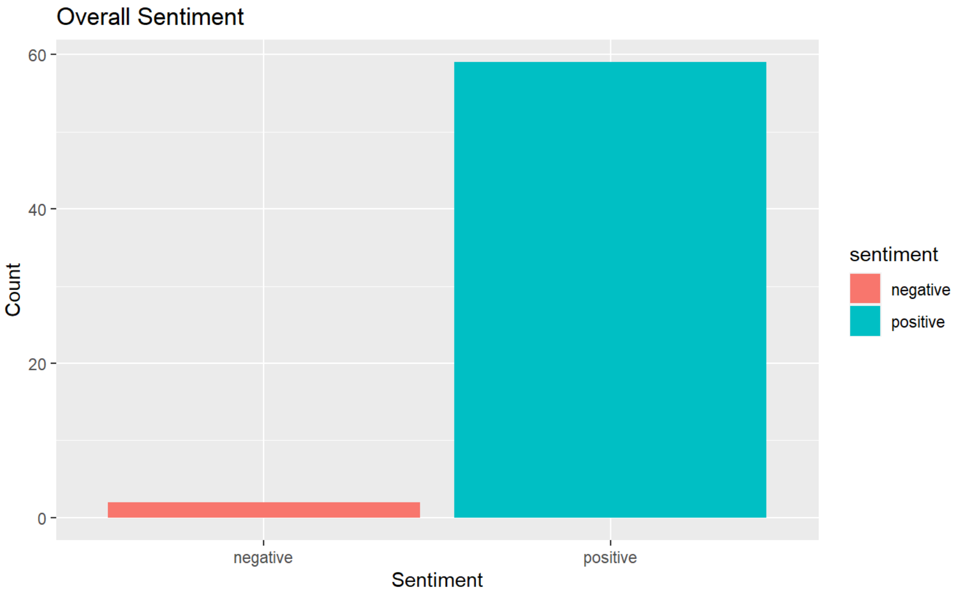

Figure 3.23: Overall Sentiment for Better Resources Question

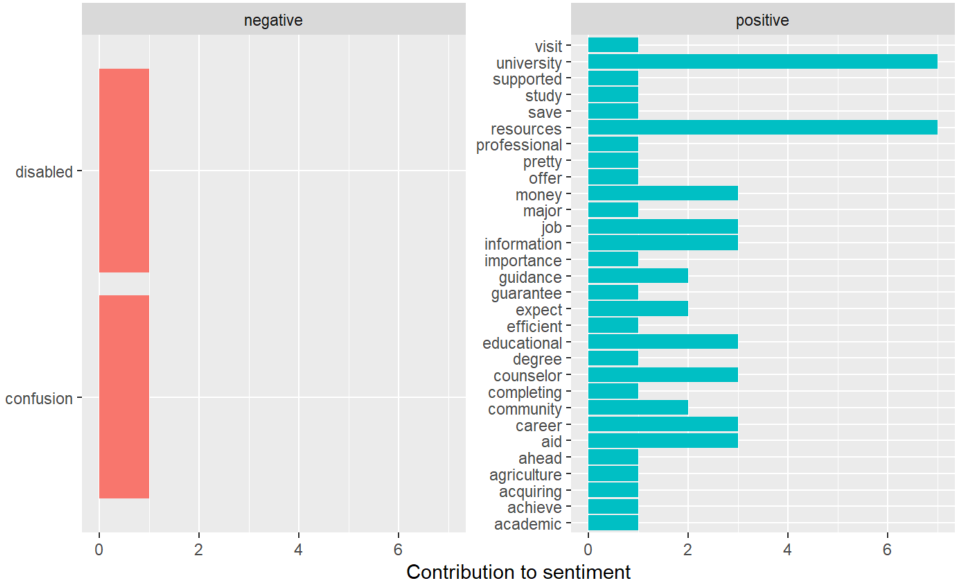

Figure 3.24: Positive and Negative for Better Resources Question

CHAPTER 4

Discussion

4.1 Financial Barriers

Some respondents expressed concern about college expenses. 36 out of 64 respondents agree to the sentiment that they are stressed about paying for college expenses. Not only that, 16 respondents reported that one of their financial sources to pay for college expenses is out-of-pocket sources. It appears this stress of paying for college expenses continues over to a four-year university since many respondents stated that college expenses will be a potential obstacle when transferring to a four-year university.

4.2 Academic Preparedness and Support

Most respondents feel academically prepared to attend college with over half feeling academically prepared to complete a bachelor's program. Despite 55 respondents acknowledging that they are aware of their college's academic resources available to them, only 24 respondents are currently enrolled in a college support program with half of those respondents being enrolled in the Extended Opportunities Program and Services (EOPS). However, this is not to say the respondents are not taking advantage of the college resources since college support programs are not the only academic resource offered by their college. Not only that, there appears to be a positive reception with the resources their college has available to them, since there is an overall positive sentiment when asked what resources can be added to better aid their academic journey.

4.3 Family Culture and Support

There appears to be influence of family in students' perception of higher education and the decision to attend a four-year university. Most respondents either stated or provided sentiment that they are receiving family support either financially or through encouragement. For example, four respondents reported that they receive family support through aiding as a financial source. Half of the respondents agreed that family has a higher priority than education and some respondents stated that family serves as a primary motivation in attending college. One respondent even stated in their answer to the question regarding the greatest obstacle once a student attend a four-year university that they would experience homesickness due to moving away from their home and family. Despite this, there is still encouragement and motivation from family to attend college.

4.4 Perception of Transfer Process

Although half of the respondents agreed to the statement that they are aware of the application process to transfer to a four-year university, the other half of the respondents show uncertainty with the application process. Not only that, almost half of the respondents felt disagreement or neutrality when asked whether their academic counselors and course instructors are informative about the transfer process. Despite the uncertainty, the majority of respondents are on track to transferring with taking courses that will meet the coursework requirements to transfer to a four-year university.

4.5 Factors Influencing Decision to Attend a Four-Year University

Among the common responses to the question of what would be the greatest obstacle once a student attends a four-year university, challenges related to handling courses such as procrastination and time management, are mentioned. This could possible be related to the almost unanimous belief that the courses at a four-year university are more rigorous than the courses at respondents' college. Another obstacle and factor to consider would be financial barriers, with growing stress of paying for college tuition and expenses as well as debt. It is more likely for a respondent to plan to attend a four-year university if they agreed to the statement that they are currently talking the classes required to attend a four-year university. Respondents also heavily consider distance between home and a four-year university as a factor in the decision to transfer, with over half believing that distance is an important factor. However, this does not deter these respondents since it is more likely for a respondent to plan to attend a four-year university if they agreed to the statement that distance is an important factor to consider.

4.6 Recommendations

Although there was mostly a positive reception when it comes the current academic resources, some resources respondents are wishing to have are more internships, financial aid opportunities, and networking events. Despite feeling academically prepared for college, it appears that there is wariness with the transfer application process and thus this information should be more readily available or at least advertised more to the students.

4.7 Limitations and Further Research

As with any research project, there are a few limitations to the study and recommendations for further research. Despite the data being a representative sample and thus a large dataset is not required, a slightly larger sample size may provide more flexibility with performing the regression modeling and allow for reproducibility. A recommendation for further research would be to explore ridge logistic regression model as a potentially more effective alternative in addressing the strong multicollinearity issue.

APPENDIX A

Data Dictionary

Survey Question	Variable Name	Variable or Analysis Type
Do you plan to apply to a four-year university?	plan_to_attend	response variable
I am stressed about paying for college expenses.	likert_stressed_expenses	categorical explanatory
Highschool has prepared me academically for college.	likert_high_prepared	categorical explanatory
I feel academically prepared to complete a bachelor's program.	likert_bach_prepared	categorical explanatory
I am aware of the academic resources my college has available to me.	likert_academic_resources	categorical explanatory
Family has a higher priority than education.	likert_family_priority	categorical explanatory

My family encourages me to pursue a college education.	likert_family_encourage	categorical explanatory
My family helps me financially with my college expenses.	likert_family_financial	categorical explanatory
I am aware of the application process to transfer to a four-year university.	likert_transfer_process	categorical explanatory
I am taking classes to meet all the coursework requirements to transfer to a four-year university.	likert_classes_require	categorical explanatory
The academic counselors and course instructors at my college are informative about the requirements of transfer to a four-year university.	likert_college_informative	categorical explanatory
The distance between my home and a four-year university is an important factor in my decision to transferring to a four-year university.	likert_distance	categorical explanatory
I believe that the classes at a four-year university are more rigorous compared to classes at my college.	likert_rigorous	categorical explanatory
How are you paying for college?	college_paying	categorical explanatory
Are you currently enrolled in a college support program?	college_support_enrolled	categorical explanatory

What is your gender?	gender	categorical explanatory
Do you have a significant other?	significant_other	categorical explanatory
Do you have children?	children	categorical explanatory
How many hours of work are you working per week?	hours_worked	categorical explanatory
Are you and/or your siblings the first in your family to attend college in the U.S.?	first_attend	categorical explanatory
Are you and/or your siblings the first in your family to be born in the U.S.?	first_born	categorical explanatory
Is English your first language?	english_first	categorical explanatory
On a 4.0 scale, what is your current GPA?	gpa	numeric explanatory
How old are you?	age	numeric explanatory
What college support program are you enrolled in?	college_support_name	exploratory data analysis
What college are you currently attending?	college_attending	exploratory data analysis
What major do you plan to pursue in the four-year university?	four_year_major	exploratory data analysis

Are you Latino/Hispanic?	latino	exploratory data analysis
For Parent #1, what is their highest level of education?	parent1_edu	exploratory data analysis
For Parent #2, what is their highest level of education?	parent2_edu	exploratory data analysis
What is your first language?	no_english_first	exploratory data analysis
If you do not plan to attend a four-year university, what is your academic goal with currently attending your college?	academic_goal	exploratory data analysis
What is your main motivation for attending college? Please explain.	college_motivation	text mining analysis
What are your major academic challenges you are facing right now?	current_challenges	text mining analysis
What do you believe would be the greatest obstacle(s) for you once you attend a four-year university?	greatest_obstacle	text mining analysis
What resources do you feel your college should have to better aid you in your path toward attending a four-year university?	better_resources	text mining analysis

Table A.1: Data Dictionary

REFERENCES

[Bur] U.S. Census Bureau. “Quick Facts on California.” https://www.census.gov/quickfacts/CA.

[Cha] UCLA Chancellor. “Becoming a Hispanic-Serving Institution by 2025.” https://chancellor.ucla.edu/messages/becoming-hispanic-serving-institution-2025/.

[Cola] Latino Donor Collaborative. “2022 LDC U.S. Latino GDP Report.” https://www.latinodonorcollaborative.org/original-research/2022-ldc-u-s-latino-gdp-report.

[Colb] Hispanic Association of Colleges & Universities. “Emerging Hispanic-Serving Institution (HSIs) 2021-2022.” https://www.hacu.net/images/hacu/OPAI/2023_EmergingHSILists.pdf.

[Edu] U.S. Department of Education. “White House Initiative on Advancing Educational Equity, Excellence, and Economic Opportunity for Hispanics.” https://sites.ed.gov/hispanic-initiative/hispanic-serving-institutions-hsis/.