A MULTIVARIATE STATISTICAL ANALYSIS OF MAJOR CHANGE PATTERNS AND SIGNIFICANT FACTORS THAT INFLUENCE GRADUATION RATES: A CASE STUDY AT CALIFORNIA STATE UNIVERSITY, LONG BEACH

A THESIS

Presented to the Department of Mathematics and Statistics

California State University, Long Beach

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Applied Statistics

Committee Members:

Sung Kim, Ph.D. (Chair)
Jen-Mei Chang, Ph.D.
Kagba Suaray, Ph.D.

College Designee:

Will Murray, Ph.D.

By Yale B. Quan

B.S., 2013, California State University, Long Beach

December 2020

ABSTRACT

In 2015 the California State University system launched Graduation Initiative 2025 which aims to eliminate the equity gaps in degree completion and increase the average four-year graduation rate from 19% to 40% and the average six-year graduation rate from 57% to 70%. To support CSULB in meeting these goals, this study focuses on performing a multivariate statistical analysis to determine the effects of major change and various demographic and academic factors on timely graduation. The dataset was obtained from the Department of Institutional Research and Analytics and contained academic and demographic information on first-time freshmen accepted between 2009 and 2012. Due to high multicollinearity, dimensionality reduction was performed using Factor Analysis, and the data were analyzed using a combination of hypothesis testing, correlation analysis, multinomial logistic regression, and Fishers linear discriminant analysis.

This study found that students who changed majors graduated at a significantly higher rate than students who did not change majors. Students who changed majors at least once graduated at over double the rate of students who did not change majors. The multinomial logistic model suggested that a student's academic performance, measured as a weighted linear combination of GPA, and the number of DFW, WU, and WE courses, is the most significant predictor of a student dropping out versus graduating within four or six years. STEM majors, Pell-eligible students, and local admission students are more likely to graduate as six-year graduates as opposed to four-year graduates. Fisher's Linear Discriminant Analysis provided two liner discriminants that can be used to predict a student's graduation status and uncover graduation trends within departments and colleges. The first linear discriminant can be used to

separate students who exit without graduating from students who graduate, and the second linear discriminant can be used to separate four- and six-year graduates.

ACKNOWLEDGEMENTS

I would like to thank my amazing wife Jasmine for all her support through graduate school. She was always there to listen to me discuss math and to provide a safe space when the pressure became overwhelming.

I want to thank my thesis committee: Dr. Sung Kim for agreeing to chair my committee and for helping me develop the statistical knowledge I needed for this thesis, Dr. Kagba Suaray for suggesting the master's program to me and supporting me throughout the program, and Dr. Jen-Mei Chang for introducing me to institutional research and for providing me my first teaching opportunity at CSULB. I also want to thank all the professors that I have either worked with or taken classes with. Your guidance and support have made studying and teaching at CSULB an incredible experience that I will miss dearly after I graduate.

Lastly, I want to thank my friends for their support through the thesis process and throughout the master's program. I was once told by a professor that you don't get through graduate school by yourself, and I firmly believe that without my friends I would not be here today writing and preparing to defend this thesis.

TABLE OF CONTENTS

ABSTRACT ii

ACKNOWLEDGEMENTS iv

LIST OF TABLES vi

LIST OF FIGURES viii

1. INTRODUCTION TO THE STUDY 1
2. THE DATASET 3
3. LITERATURE REVIEW 14
4. METHODOLOGY 21
5. RESULTS 38
6. CONCLUSIONS 60

APPENDICES 65

A. DESCRIPTIVE STATISTCS FOR ACADEMIC VARIABLES 66
B. TABLES OF DEMOGRAPHIC INFORMATION 73
C. TABLES USED FOR CHI-SQURE ANALYSIS 81
D. TABLES FOR TIMELY GRADUATION 83
E. VARIANCE INFLATION VALUE CALCULATIONS 85
F. CORRELATION ANALYSIS 87
G. UNIQUENESS TABLE AND RESIDUAL MATRIX 93
H. FACTOR LOADINGS 96
I. OUTPUT FROM THE MULTINOMIAL LOGISITIC MODEL 98
J. TABLES AND FIGURES RELATED TO FISHERS LDA 102

REFERENCES 106

LIST OF TABLES

1. Cohort Sample Size 4
2. Major Change Groupings 6
3. Eligibility Index Scores for Each Cohort Year 8
4. STEM Admissions for Major Change Groups 10
5. Results of the 2001 University of South Florida Study 14
6. Results of the 2016 EAB Study 16
7. Table for Calculating the Phi Coefficient of Correlation 24
8. Misclassification Costs with Two Populations 34
9. Post Hoc Testing of $\chi 2$ Test for Homogeneity of Proportions 39
10. Graduation Rates for Each Major Change Group 40
11. Timely Graduation Rates for Each Major Change Group 40
12. Binary Logit Models Used to Calculate VIF Values 42
13. Variables Retained After Correlation Analysis 43
14. Eigenvalue and Percent Variance Explained 43
15. Factor Loadings for Factor 1 45
16. Factors and Their Associated Latent Variables 45
17. First 10 Rows of Calculated Factor Scores 46
18. Variables used in the Multinomial Logistic Model 47
19. Relative Risk Ratios for Did Not Graduate Versus Four Year Graduate 50
20. Relative Risk Ratios for Six Year Graduate Versus Four Year Graduate 51
21. Confusion Matrix for The Multinomial Logistic Model 51
22. Variables Used in Fishers LDA 52

23. Confusion Matrix for Fishers LDA; Training Dataset 55

24. Confusion Matrix for Fishers LDA; Testing Dataset 56

25. Example Student Raw Data 58

26. Example Student Variables and Factor Scores 58

LIST OF FIGURES

1. Scree and cumulative scree plots 44

2. Scatterplots produced from fishers lda 57

3. Example of using LD1 and LD2 to classify a new observation 59

CHAPTER 1

INTRODUCTION TO THE STUDY

In 2015 the California State University (CSU) system launched Graduation Initiative 2025 which aims to increase the four- and six-year graduation rates for first-time freshmen and eliminate equity gaps in degree completion. Graduation Initiative 2025 set a CSU-wide goal of a 40% four-year graduation rate and a 70% six-year graduation rate (California State University 2015). To meet the graduation rate goals set by Graduation Initiative 2025 it is important to determine what factors might influence a student's time to graduation. There is a concern in the higher education community about major change patterns and how changing majors may affect timely graduation (Micceri 2001; Foraker 2012; United States Department of Education 2017). Additionally, there are socioeconomic, demographic, and academic factors to consider which might influence a student's time to graduate. California State University, Long Beach (CSULB) serves a diverse population of students who come from a variety of demographic and socioeconomic backgrounds. These students have a total of 90 different majors to select from spread across 7 different colleges.

Identifying significant academic and demographic traits that influence a student's time to graduation will allow Academic Advising to identify students who are at risk for not graduating or who might graduate in six years and offer support as needed. With this goal in mind this study aims to answer the research question: Does a change of major have a significant effect on a student's graduation status, and can this result be combined with the students' academic and demographic information to identify significant factors that influence timely graduation?

To answer the research question, a combination of multivariate statistics techniques were used including a χ^2 test for homogeneity of proportions and correlation analysis to determine if

there was a significant difference in the proportion of students who graduated between major change groupings. The dimensionality of the dataset was reduced using the Variance Inflation Factor and Factor Analysis. This reduced dataset was then used for modeling graduation rates with a Multinomial Logistic model, and classification using Fishers Linear Discriminant.

Chapter 2 presents the descriptive statistics for essential academic and demographic variables. Chapter 3 outlines the background literature and previous studies involving timely graduation and major change patterns. Chapter 4 presents the statistical methodology used in this study, and in Chapter 5 the results from the statistical analyses are presented. Chapter 6 discusses the results of the study and provides recommendations for future research.

CHAPTER 2

THE DATASET

For this study, *first-time freshmen*, will be defined as students who have never enrolled at another higher education institution before attending CSULB, and *cohort* will be defined to be the group of students who started CSULB in the same academic year. The data for this study consists of academic and demographic data on first-time freshmen who enrolled at CSULB as part of the 2009, 2010, 2011, or 2012 cohorts. These cohort years were selected because they are the most current years in which the four- and six-year graduation rates can be reliably calculated. The data were obtained from the Department of Institutional Research and Analytics at CSULB after receiving approval from the Institutional Review Board. In the initial dataset, each row represented a class a student took while enrolled at CSULB and each of the columns represented a specific variable. The dataset consisted of 181 columns that contained academic and demographic information on 16,468 first-time freshmen measured over 757,730 rows of semester information.

For this study, *time to graduation* will be defined as the number of semesters it takes for a student to graduate from CSULB. Time to graduation is traditionally binned into three categories: students who do not graduate, students who graduate within four years (8 semesters), and students whose time to graduation is longer than four years and no more than six years (12 semesters). At the time of this study there were 382 students in the dataset who are either currently enrolled or who were enrolled for longer than 12 semesters before graduating. Since this study focuses on the four- and six-year graduation rates these students will be excluded from the study. High school GPA and standardized test scores (SAT and/or ACT exams) are used

significantly throughout this study. Consequently, the 239 students who did not have a high school GPA and/or a standardized test score were removed from the study.

There are three types of non-traditional students in the dataset. These students are considered non-traditional because their initial admittance into the university was not as a degree-seeking student. These are students whose initial admission classification was as Early Entrant, Young Scholar, or as a Non-Degree Seeking student. In this dataset there were 45 Early Entrant students, 442 Non-Degree seeking students, and 34 Young Scholar students. These 521 students were excluded from this study.

The last type of student excluded from this study are students who enrolled in at least one Winter or Summer semester course. Students who were enrolled in either Winter or Summer semesters may exhibit different course taking patterns and graduation rates than students who have never enrolled in a Winter or Summer course. Additionally, Summer and Winter semester classes are usually not covered under financial aid which disenfranchises students who cannot afford to pay for their classes. These 7,355 students were excluded from the study. The final dataset used for the study consisted of 8,505 students.

TABLE 1. Cohort Sample Size

Cohort Year	Sample Size
2009	1,792
2010	2,107
2011	2,655
2012	1,951

2.1 Descriptive Statistics

This study will focus on the 24 variables that have no missing data. These variables can be separated into four distinct components: student major change patterns and graduation rates, measures of academic preparation, CSULB academic information, and demographic information. Academic preparation measures how well the student academically prepared for college, CSULB academic information measures the students' academic progress while enrolled at CSULB, and demographic information consists of variables that describe the student outside of their academic information.

2.1.1 Major Change Patterns and Graduation Rates

To begin answering the research question the four- and six-year graduation rates for each cohort year were calculated. Graduation rates are defined by the National Center for Education Statistics as the percentage of students who completed their degree within a specified time limit and are calculated using Equation 1:

$$\text{Graduation Rate} = \frac{\text{number of students who completed the program within a specified time}}{\text{number of students in the cohort or sample}} \quad (1)$$

(National Center for Education Statistics 2016).

For all cohort years the first-time freshman four-year graduation rate was between 16% -18% and the first-time freshman six-year graduation rate was between 47%-50%. The percentage of first-time freshman who did not graduate was relatively stable at around 36%. These tables can be seen in Appendix A. It is important to note that the data provided does not separate students who exited without a degree, and students who transferred to another institution. This might artificially increase the number of students classified as "Did Not Graduate."

A coding system was then developed to group students based on their number of major changes. Students in group "A" had no major changes, students in group "B" had one major change, students in group "C" had two major changes, and students in group "D" had more than two major changes. This grouping was chosen to be consistent with previous research performed by Western Kentucky University (Foraker 2012) and University of South Florida (Micceri 2001).

Table 2 contains the number of students in each major change group within each cohort along with the percentage of students in that group relative to the rest of the cohort year. Over half of the students in all cohorts switched majors at least twice. These results combined with the timely graduation percentages indicate that, for the years studied, first-time freshmen at CSULB tended to switch majors at least twice while maintaining relatively stable graduation rates.

TABLE 2. Major Change Groupings

Major Change Group	2009 Cohort	2010 Cohort	2011 Cohort	2012 Cohort
No Major Change	$612_{(34\%)}$	$634_{(30.\%)}$	$817_{(31\%)}$	$549_{(28\%)}$
One Major Change	$157_{(9\%)}$	$148_{(7\%)}$	$167_{(6\%)}$	$90_{(5\%)}$
Two Major Changes	$644_{(36\%)}$	$808_{(38\%)}$	$959_{(36\%)}$	$719_{(37\%)}$
More than Two Major Changes	$379_{(21\%)}$	$634_{(25\%)}$	$712_{(27\%)}$	$593_{(30\%)}$

Note: The number in parenthesis denotes the percentage of students in a specified group for that cohort year. Columns sum to 100%

2.1.2 Academic Preparation

Variables in the dataset that measure a students' academic preparation for college include the student's high school GPA, and their SAT /ACT math and reading scores. These variables are then used to calculate the student's Eligibility Index Score which is used to establish baseline admission eligibility.

For each cohort year the mean and standard deviation of the students' high school GPA and SAT/ACT math and reading scores were calculated. All the calculated means for both variables were within one standard deviation of each other, which indicates that there was no statistically significant difference in the average high school GPA or the student's SAT/ACT math and reading scores between cohort years. These statistics were also calculated within each major change grouping with the same results. The tables containing the statistics of the students' high school GPA and SAT/ACT math and reading scores can be seen in Appendix A.

A student's Eligibility Index score is a linear combination of the student's high school GPA and SAT/ACT math and reading scores. Students who apply to CSULB and take the ACT have their ACT score converted to a comparable SAT score (CSULB Admissions n.d.a.). In March of 2016 CollegeBoard changed how the SAT was scored (CollegeBoard SAT n.d.) and the available concordance tables, which convert ACT to SAT scores, are only available for ACT and SAT exams taken after 2016. Therefore, SAT and ACT scores provided in the dataset were converted to the 2016 scoring scale using concordance tables provided by CollegeBoard (CollegeBoard n.d.a.).

CSULB uses two formulas to determine a student's Eligibility Index based on what major the student is applying into. The general Eligibility Index formula (2) equally weights the SAT Math and SAT Critical Reading while placing the greatest weight on the student's high school GPA. CSULB students whose designated major or pre-major is in the College of Engineering, the College of Natural Sciences and Mathematics, or in the College of Liberal Arts as an Environmental Science & Policy major are considered STEM majors. The STEM Eligibility Index formula (3) weights the student's SAT Math score twice as much as their SAT Critical

Reading score and lowers the weight on the student's high school GPA by 25%. (CSULB Admissions n.d.a.)

$$800(\text{High School GPA}) + \text{SAT Critical Reading} + \text{SAT Math} \quad (2)$$

$$600(\text{High School GPA}) + \text{SAT Critical Reading} + 2(\text{SAT Math}) \quad (3)$$

Table 3 contains the mean, median and standard deviation for the Eligibility Index score in each cohort year. The average Eligibility Index score for each cohort year were all within one standard deviation of each other which indicates that was no statistically significant difference between the cohort years. The same statistics were calculated for each major change group within a cohort with the same results. These tables can be seen in Appendix A.

TABLE 3. Eligibility Index Scores for Each Cohort Year

Cohort Year	Mean Eligibility Index	Median Eligibility Index	Standard Deviation
2009	3855.66	3850	386.70
2010	3847.24	3862	362.99
2011	3826.52	3792	361.80
2012	3913.01	3912	359.45

2.1.3 CSULB Academic Information

To calculate a student's GPA, CSULB uses a weighted GPA system. Each letter grade has a weight assigned to it which is then multiplied by how many units the course is. This is called the Grade Point. To calculate the overall GPA, the earned Grade Points are divided by the total number of units the student enrolled in while at CSULB and rounded to two decimal places. The result is a maximum GPA of 4.00 and a minimum GPA of 0.00 (CSULB Student Records n.d).

The average CSULB GPA for each cohort year was approximately 2.50 with a median of approximately 2.70. The CSULB GPA variable is skewed left, and the calculated medians and Inner Quartile Values for each cohort year were approximately equal which suggests that there was no statistically significant difference between the median CSULB GPA between cohort years. The same outcome can be seen when examining CSULB GPA for each major change group using mean and standard deviation. These tables can be found in Appendix A.

One metric commonly used to measure student success is the number of D, F, W, WU and WE grades the student earned in a semester. A letter grade of D and F are calculated in the student's semester and overall GPA. A W grade indicates that the student withdrew from the class before having an official grade assigned and is not calculated in the student's GPA but does appear on their transcript. A grade of a WU is classified as an unauthorized withdrawal and is equivalent to an F with respect to calculating GPA. A student earns a WU grade by not officially withdrawing from a course and failing to complete the course requirements. A WE grade is commonly referred to as a catastrophic withdrawal. These are withdrawals granted for extenuating circumstances that are outside of the student's control. A WE grade is calculated as a W with respect to the student's GPA (CSULB Student Records n.d.).

The number of D, F or W grades a student earned were totaled and assigned to the *DFW* variable. The mean, median, and standard deviation for the number of DFW, WE, and WU courses in each cohort year were calculated. The average number of DFW, WE, and WU courses were all within one standard deviation of each other both at the cohort level and at the major change grouping level. This indicates that there was no statistically significant difference between these variables between cohorts or major change groupings. These results are also listed in Appendix A.

For each cohort year the percentage of STEM admissions is far lower than the percentage of non-STEM majors, just 20%-30% of first-time freshmen were admitted as a STEM majors. This table is in Appendix B. The number and percentage of STEM majors within each major change group was also calculated and recorded in Table 4. The major change behavior seen in Table 4 is different than what was observed in Table 2, which is the overall major change activity within each cohort year. In Table 2 there was a small percentage of students within each major change group that changed majors once, however, for STEM majors there were no students who only changed majors once.

TABLE 4. STEM Admissions for Major Change Groups

Major Change Group	2009 STEM Admissions	2010 STEM Admissions	2011 STEM Admissions	2012 STEM Admissions
No Major Change	$162_{(26\%)}$	$179_{(28\%)}$	$225_{(40\%)}$	$257_{(31\%)}$
One Major Change	$0_{(0\%)}$	$0_{(0\%)}$	$0_{(0\%)}$	$0_{(0\%)}$
Two Major Changes	$124_{(19\%)}$	$175_{(22\%)}$	$169_{(23\%)}$	$211_{(22\%)}$
More than Two Major Changes	$79_{(21\%)}$	$119_{(23\%)}$	$197_{(33\%)}$	$206_{(23\%)}$

Note: The number in parenthesis denotes the percentage of students in a specified group for that cohort year. Columns will not sum to 100%.

The last CSULB academic variable measured is the number of undeclared semesters a student had before declaring their first major and the percentage of undeclared admissions within each cohort. On average, undeclared students declared their first major within 3 or 4 semesters, and at most 20% of each cohort was admitted as undeclared. When studying the number of undeclared admissions in each major change group a few observations can be made: students who changed majors at least three times had the lowest percentage of undeclared admissions each cohort year, and there were zero undeclared admissions in the group of students that only changed majors once. The tables containing this information is in Appendix A.

2.1.4 Student Demographic Information

When applying to CSULB students have the option to self-identify their gender as male or female. For each cohort year over 50% of the students self-identified as female in each cohort and over 50% of each major change group is female as well. These tables are listed in Appendix B. Students have eight options, listed in Appendix B, when deciding what level of parent education they select when applying to CSULB. The National Center for Education Statistics classifies a student as a first generation college student if both parents have never earned a bachelor's degree (National Center for Education Statistics 2018). Using the coding system shown in Appendix B, these would be parents whose reported education level is no higher than 5 ("Some College"). After classifying each student as first-generation college or non-first-generation college, the percentage of first-generation students was calculated. For each cohort year over half of the admitted first-time freshmen were first-generation college students. When examining the distribution of first time freshmen in each major change group the same trend can be observed.

The dataset provided did not include socioeconomic information, but the student's Federal Pell Grant eligibility is provided. The Federal Pell Grant is a grant awarded by the Federal government to higher education students who display financial need, and who have not earned a bacherlor's, graduate, or professional degree. Student eligibility is determined through a variety of factors that include the students Expected Family Contribution, the cost of attendance, full or part time enrollment, and if the student is planning on attending the university for the full academic school year. Students who statisfy the requirements are eligible to receive the Pell Grant for up to 12 academic terms at their campus (Federal Student Aid n.d.).

Generally, Federal Pell Grants are awarded to students whose families have a total income below $20,000 per year (Scholarships.com n.d.).

Over half of the first-time freshmen in this study were Pell Grant eligible. The 2009 cohort had the lowest percentage of Federal Pell Grant eligible students at 51% while the 2010-2011 cohorts had the highest percentage of Federal Pell Grant eligible students at 62%. Please note that a student being eligible for the Federal Pell Grant does not imply that the student accepted the grant, only that they met the eligibility requirements. Examining the percentage of Federal Pell Eligible students in each major change group showed that, within each major change group, most students were Pell Eligible.

The final two pieces of demographic information collected are the students local admission status and their self-identified minority status. Local admission is determined by the high school the student graduated from. There are 42 different high schools in 10 school districts that are considered local for the purpose of identifying local preference. The list of eligible high schools are located in Appendix B. Applicants to CSULB who are classified as a local admission are provided preferential admission which is referred to as Local Preference. Local Preference allows students who are not accepted to their selected major to be admitted to the univeristy as undeclared. Students classified as undeclared in this manner are not allowed to pursue their initial major and must work with Academic Advising to select an alternate major (CSULB Admissions n.d.b). For each cohort year over 50% of the first-time freshman were non-local admissions. The 2010 cohort had the largest percentage of non-local admissions at 59% while the 2012 cohort had the lowest percentage of non-local admissions at 54%. A similar distribution of local admissions can be seen within each major change group. These tables can be seen in Appendix B.

CSULB applicants are asked to self-identify their race from nine pre-defined options. The full table of races which students can self-identify as can be found in Appendix B. Students were classified as a non-minority admission if they self-identified as White, Asian, Native Hawaiian or Other Pacific Islander, or Two or More Races (California State University, n.d.). Students who did not select those options were classified as a minority admission. The majority of admissions for the 2009 – 2012 cohorts were non-minority students, and a similar distribution can be seen within each major change group. The cohort with the largest percentage of minority admissions was the 2011 cohort with 49.94%.

In all four cohorts the majority of students changed majors at least twice and students who changed majors graduated at a higher rate than students who did not change majors. Students who changed majors either once or twice were more likely to graduate within four years. There was no statistically significant difference in the students academic preperation and CSULB academic information at the cohort level or at the major change grouping level. Based on these results, the four cohorts were combined before continuing the analysis.

CHAPTER 3

LITERATURE REVIEW

There are many topics that come up when discussing how to improve the four- and six-year graduation rates. Topics include the number of major changes a student has, course taking patterns, academic readiness, and demographic influence. This research paper will focus on major change and its effects on the four- and six-year graduation rates along with identifying significant factors that influence timely graduation.

In 2001, University of South Florida published a study that compared the average times to graduation between students who did not change majors, students who changed majors one, students who changed majors twice, and students who changed majors at least three times. The sample studied consisted of 9,110 students who were enrolled for longer than two semesters from 1991 to 1998. For each major change group, the average time to degree completion was calculated. Based on a numeric comparison the study concluded that successive major changes did not significantly impact the average time to graduation at University of South Florida. No advanced statistical methodology was used in this study (Micceri, 2001). The results of this study are recorded in Table 5.

TABLE 5. Results of the 2001 University of South Florida Study

Number of Major Changes	Average Time to Degree Completion (years)
No major changes	4.80
One major change	4.82
Two major changes	4.88
More than two major changes	5.03

A 2012 study performed by Matthew Foraker at Western Kentucky University expanded on the study performed by the University of South Florida by attempting to answer two research

questions: does the number of major changes have an effect on graduation rates, and are students who finalize their major before two years more successful than students who finalize their major after two years? To determine if the number of major changes had an effect on graduation rates Foraker divided the sample of 7,009 first-time freshmen into three groups: students who did not change majors, students who changed majors once, and students who changed majors more than once. The graduation rates were then calculated for each group. Foraker concluded that students who started as undeclared and declared a major by their second year were the most successful with 83.4% of those students graduating within 6 years. Students who did not change majors had a 72.8% graduation rate, students who changed majors once had a 71.70% graduation rate, and students with multiple major changes had a 70% graduation rate.

To determine if students who finalize their major before two years were more successful (higher graduation rate) the sample was divided into two groups: students who finalized their major before two years, and students who finalized their major after two years. The corresponding graduation rates were calculated for each group. Foraker noted that the graduation rate decreased from around 80% to 72% for students who had not finalized their major after two years. Foraker concluded that major changes that occurred within the first two years did not have a negative impact on student success while major changes after the second year was correlated with an increased time to graduation and lower graduation rates (Foraker 2013).

One of the most recent studies that looked at major change and its impact on timely graduation and graduation rates was performed by EAB, formerly the Education Advisory Board, in 2016. EAB's research focused on the timing of the final major switch and the associated time to graduation and graduation rates. Their sample consisted of academic

information of students from ten higher education institutions for a total sample size of over 78,000 students.

EAB reported the percentage of students who graduated within six years. EAB found that students who never changed majors had an average graduation rate of 78.45%, and students who changed majors at least once within the first ten semesters (five years) had a higher graduation rate than students who never changed majors. EAB also reported that students finalized their major before their sixth semester (third year) saw no negative impact on time to gradation. After the sixth semester students who changed majors experienced a longer time to graduation and a decrease in their graduation rate. However, the increased time to graduation was not significant and the EAB study concluded that there was very little evidence to support the theory that major switching increases the time to graduation and decreases graduation rates (EAB 2016).

TABLE 6. Results of the 2016 EAB Study

Semester of final Major Declaration	1	2	3	4	5	6	7	8	9	10	11	12
Average Graduation Rate (Percentage)	78.45	82	83	83	84	83	82	83	80	80	74	70
Median Time to Graduation (Semesters)	NA	8	8	8	8	9	10	9	11	12	14	14

Comparing the studies performed by EAB, Western Kentucky University, and University of South Florida shows that determining if major change influences graduation rates is difficult. University of South Florida reported that major changes had no effect on time to graduation and graduation rates, while both Western Kentucky University and EAB reported that late major change activity was correlated with increased time to graduation and lower graduation rates.

When modeling academic performance, such as timely graduation, the dependent variable is generally categorical with more than two mutually exclusive groups (i.e., A-F grades). Traditional modeling techniques such as logistic regression cannot be used with this type of dependent variable since binary logistic regression can only model a dependent variable with two mutually exclusive outcomes. To address this problem Park and Kerr (1990) suggest using a multinomial logit or multinomial probit model to model academic performance. In their 1990 study Park and Kerr applied both a multinomial logistic and multinomial probit model to determine which factors are statistically significant in determining a student's grade in an undergraduate money and banking course. Park and Kerr concluded that while both a multinomial logistic and multinomial probit model could be applied to academic data, the multinomial logistic model was able to provide more details than the multinomial probit model. Park and Kerr (1960) also concluded that if the goal was to predict the probability of a student earning a specific letter grade, the multinomial probit model would be preferable, but the multinomial logistic model provided a better understanding of the relationship between the multiple mutually exclusive outcomes.

In 2010, White and Huesman from the Office of Institutional Research at University of Minnesota – Twin Cities (UM-TC) performed a similar study by comparing the results from a binary logit, a multinomial logit, and a multinomial probit model to determine the best approach to model academic data. The goal of the White and Huesman study was to identify "factors leading to success across the different paths students take through higher education" (White and Huesman 2010). The binary logit model was implemented with the dependent variable as: 1 = graduated from UM-TC within six years, and 0 = did not graduate within six years. The binary logit model concluded that at a significance group of $\alpha = 0.05$ student's academic

preparation, financial need, and geography were significant in determining if a student will complete their degree within six years without transferring.

When fitting the multinomial logit and multinomial probit models a dependent variable was created with four mutually exclusive outcomes: (1) baccalaureate degree from UM-TC, (2) baccalaureate degree from another higher education institution, (3) associate degree/certificate award from another institution, or (4) student failed to obtain a degree in the six-year period examined. For both the multinomial logit and multinomial probit models the reference category was outcome (4). White and Huesman concluded that when modeling academic data with a categorical dependent variable with more than two outcomes, implementing a multinomial logit or a multinomial probit model will provide more information than a binary logit model.

Predicting if a student will graduate in four years, six years, or if they will exit the university before graduating is important for academic advisors. One of the techniques suggested by Marr and Hume (1996) for identifying these students is Discriminant Analysis (DA). Marr and Hume suggest that using DA with educational data results in a predicted category membership such as pass/fail or graduate/did not graduate, and these memberships can be useful when evaluating intervention techniques. The other common technique used in institutional research is multiple regression, however, most variables used in institutional research are not continuous variables which violates one of the assumptions of multiple regression. Marr and Hume suggest that while DA was not specifically intended to work with categorical variables, the technique works quite well with a mixture of categorical, discrete, and continuous data.

In their study Marr and Hume used DA to study and classify students at the Georgia Institute of Technology who took an Electricity and Magnetism course. The Electricity and

Magnetism course had a 70% pass rate and students who do not pass the class are required to repeat the course. Marr and Hume used a sample of 1,622 students for training and 534 students for testing. The training dataset was split into two groups: at-risk students who earned a D or F, and not at-risk students. Marr and Hume then applied DA to the students' academic information and achieved an 82% classification risk of not at-risk students. Using the produced linear discriminate Marr and Hume attempted to classify 534 students to determine how well the DA model can predict student grades. Marr and Hume achieved an 84% success rate when classifying not at-risk students, however, the DA model they produced had difficulty classifying students who earned a C. Students who earn a C might have similar academic information to students who earned a B or D which could cause the student to be misclassified. Students who earned an A, B, D, or F were more likely to be classified correctly.

Tripp and Duffey (1981) used Discriminant Analysis to classify students from 1970-1979 who applied to the Nursing Master's degree at the University of Kansas. The sample consisted of 228 students who were pre-classified into three groups: students who graduated, students who dropped out, and students who applied and were not accepted. After controlling for highly correlated variables Tripp and Duffey found that using discriminate function 1 explained 98% of the total variance within the discriminate space. Discriminate function 1 measured a linear combination of the student's baccalaureate GPA, and the students GRE Verbal and Quantitative scores. Using discriminate function 1, the 228 students were classified into their respective groups. Using a single discriminate function resulted in an accuracy rate of less than 50% for students who graduated or did not graduate and an accuracy rate of 72.3% for students who were not accepted. Tripp and Duffey concluded that included categorical variables might increase the

effectiveness of their model and that separating the GRE scores might have resulted in a higher error rate.

It is important to note that both the studies performed by Marr and Hume (1996) and Tripp and Duffey (1981) use datasets that violate the assumptions of Discriminate Analyses. The datasets for both datasets are skewed right and not distributed as multivariate normal. Additionally, the assumption of the equality of covariance matrices was not met in the Marr and Hume study. These results are expected when using datasets with mixed data types.

Research performed by Ethel Gilbert (1968) compared Fishers Linear Discriminate Function (LDF) compared to the Cochran and Hopkins procedure, MLE estimates of β, and minimum logit χ^2 estimates of β when using a mixture of qualitative, discrete, and continuous variables. After combining the discrete and continuous variables to create a dataset of only categorical variables, 100 samples of size 100 and 100 samples of size 500 were randomly generated using Monte Carlo simulations. Then the average misclassification probabilities and the variance of the estimators was calculated for each method. Gilbert concluded that using 6 or more variables with a sample size of at least 500 resulted in very little difference in the average variance of the estimated coefficients or average misclassification rate of all the functions except for the Cochran and Hopkins procedure. Gilbert concluded that since Fishers LDF relies on bivariate marginals and is simple to interpret, Fishers LDF is the superior model when using categorical variables.

CHAPTER 4

METHODOLOGY

4.1 Chi-Square Test for Homogeneity of Proportions

The chi-square test for homogeneity of proportions is outlined in Tanbakuchi (2009) and tests the null hypothesis that different populations or groups have the same proportion of some characteristic. The alternative hypothesis is that at least one group does not have the same proportion of some characteristic. The null and alternative hypothesis can be written as:

$H_0: p_1 = \cdots = p_k$

H_A: At least one proportion is not the same.

where k is the number of populations or samples, and p is the proportion of the characteristic of interest represented in the population or sample. The chi-square test statistic (4) is calculated as:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \tag{4}$$

Where r is the number of rows in the chi-square table, c is the number of columns in the chi-square table, O_{ij} is observed count in the ij^{th} cell and E_{ij} is expected count in the ij^{th} cell

$$E_{ij} = \frac{(i^{th} \text{ row total})(j^{th} \text{column total})}{\text{Overall total}}$$

If the resulting chi-square test statistic is significant at a specified α level, then it can be concluded that at least one population or sample proportion is different from the others. Post-hoc testing using pairwise chi-square tests for homogeneity of proportions should be performed whenever the contingency table is larger than 2×2.

It is important to note that when multiple hypothesis tests are conducted on the same dataset the probability of committing a Type I error (rejecting the null hypothesis when it is true) increases. To control for this increase, the Bonferroni correction can be used to adjust the

significance level of the test. The Bonferroni correction (5), as defined in McDonald (2014), is calculated by dividing the original significance level by the number of pairwise comparisons being performed. This quotient is referred to as Bonferroni's Alpha and will be denoted as α_B. It can be shown that as the number of pairwise comparisons increase the value of α_B approaches zero.

$$\alpha_B = \frac{\alpha}{\text{number of pairwise comparisons made}} \tag{5}$$

4.2 Methodology for Calculating Correlation

Correlation is commonly used to measure the relationship between two random variables and to identify significant variables that can be used for classification and predictive modeling. Three types of correlation measures were selected based on the type of random variable being analyzed: Point-Biserial Correlation, Pearson's-Product Moment Correlation Coefficient, and the Phi Coefficient of Correlation.

The Point-Biserial Correlation is defined in Tate (1954) and is used when calculating the correlation between a continuous or discrete random variable and a binary random variable. The Point-Biserial Correlation, r_{pb}, is a value between -1 and 1. Values of r_{pb} between 0 and 1 indicate that larger values of Y are more commonly associated with the 1 group of X, and values between 0 and -1 indicate that smaller values of Y are associated with the 0 group of X.

Let Y be a continuous random variable and let X be a dichotomous random variable with values of 0 and 1. Assume that there are n many paired observation notated as $(X_i, Y_i), i = 1,2, \dots, n$. Then the Point-Biserial Correlation between X and Y is calculated using equation 6:

$$r_{pb} = \frac{\overline{Y_1} - \overline{Y_0}}{S_y} \sqrt{\frac{np_0(1-p_0)}{n-1}} \tag{6}$$

$$S_y = \sqrt{\frac{\sum_{k=1}^{n}(Y_k - \bar{Y})^2}{n-1}}, \bar{Y}_j = \frac{1}{n}\sum_{k=1}^{n} Y_k \cdot 1_{\{Y_k = j\}}, j = 0, 1$$

$$p_1 = \frac{\sum_{k=1}^{n} X_k}{n}, \quad p_0 = 1 - p_1$$

As defined in Mendenhall and Sincich (2012), the Pearson Product Moment Correlation Coefficient, commonly referred to as the Pearson Correlation Coefficient (7) and denoted as r, measures the linear relationship between two continuous variables. The Pearson Correlation Coefficient is a value between -1 and 1. A value of -1 indicates that the coordinate pairs $(X_i, Y_i), i = 1,2, \ldots, n$ lie on the line that implies that the random variables *X* and *Y* have a negative relationship, and a value of 1 indicates that the random variables *X* and *Y* have a positive relationship. Values close to zero indicate that *X* and *Y* are not linearly related.

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} \tag{7}$$

The Phi Coefficient of correlation, as explained in Zyso (1998), is derived from the Pearson Correlation Coefficient and is used when examining the correlation between two binary random variables. The Phi Coefficient is expressed as the quotient of two frequencies. The numerator being the difference between the products of the outcomes, and the denominator being the corresponding expected value. When the Phi Coefficient is calculated using a 2×2 contingency table, its interpretation is the same as the Pearson Correlation coefficient. The assumptions of the Phi Coefficient are that all counts are greater than 1 and no more than 20% of the expected counts are less than 5. The Phi Coefficient is calculated using Table 7 and Equation 8 below.

TABLE 7. Table for Calculating the Phi Coefficient of Correlation

	0	1	Totals
0	a	b	n_i
1	c	d	u_i
Totals	n_j	u_j	N

$$r_\phi = \frac{ad - bc}{\sqrt{n_i u_i n_j u_j}} \tag{8}$$

4.3 Multicollinearity

Prior to fitting a predictive or classification model, the data should first be checked for multicollinearity. Multicollinearity exists when two independent variables are correlated with each other. High multicollinearity values will increase the likelihood of rounding errors in the analysis and can influence the regression results and make them difficult to interpret.

A common technique used to detect multicollinearity involves calculating the *variance inflation factor* or VIF. The interpretation of VIF used in this study can be found in An Introduction to Statistical Learning with Applications in R (James, Witten, and Hastie 2017). VIF is used to numerically describe how much the variance of a predicted regression coefficient is inflated by the existence of correlation between the predictor variables. A VIF_j value of 1 indicates that there is no correlation between the j^{th} predictor variable and the remaining predictor variables while larger values indicate higher levels of correlation. Generally, a VIF_j value larger than 4 should be reviewed and VIF values of at least 10 should be corrected. The VIF for the j^{th} predictor variance can be calculated using equation 9 below.

$$VIF_j = \frac{1}{1 - R_j^2} \tag{9}$$

R^2, as outlined in Mendenhall and Sincich (2012), is commonly referred to as the multiple coefficient of determination and is a measure of how well the proposed linear model fits the data. R^2 can be calculated using equation 10 below.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}, \qquad 0 \leq R^2 \leq 1 \tag{10}$$

where $\hat{y}_i$ is the predicted value of y_i from the multiple regression model. Values of R^2 closer to 1 indicate a good fit while values closer to 0 indicate the suggested model is not a good fit for the data.

4.4 Multinomial Logistic

A multinomial logistic model is used to explain how the multinomial response variable Y depends on k many predictor variables. The multi-population Multinomial Logit model is derived in Maximum Likelihood Estimation of Logistic Regression Models (Czepiel n.d.) and was adapted for this single population study. Let J represent the number of mutually exclusive outcomes of the dependent variable and let $J > 2$. Let Z be a random variable that can take on one of the values of J. If each observation of Z is independent of each other, then Z is said to be a multinomial random variable. Let K be the number of independent predictor variables in the dataset, and N represent the total sample size.

Let $\boldsymbol{Y}$ be a $N \times (J-1)$ matrix where y_{ij} represents the counts of the j^{th} value of Z_i. $\boldsymbol{\pi}$ is a matrix with the same dimensionality as $\boldsymbol{Y}$ where each element, π_{ij}, is the probability of observing the j^{th} value of the dependent variable at any given observation in the i^{th} row. The design matrix, $\boldsymbol{X}$, has N rows and $K+1$ columns. Note that the first element of each row of $\boldsymbol{X}$, namely x_{i0}, is equal to 1 and is called the intercept. Let $\boldsymbol{\beta}$ be a $(K+1) \times (J-1)$ matrix

where each element, β_{kj}, is the parameter estimation for the k^{th} predictor variable and the j^{th} value of the dependent variable.

The J^{th} category is considered the reference or baseline category and is omitted from the model. The logits of the remaining $J-1$ categories are constructed with the reference category in the denominator. Traditionally the reference category is the last category in the dependent variable. Using this notation, the multinomial logit function is written as:

$$\log\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = \log\left(\frac{\pi_{ij}}{1-\sum_{j=1}^{J-1}\pi_{ij}}\right) = \sum_{k=0}^{K} x_{ik}\beta_{kj}\,, \quad i = 1 \dots N \;\; j = 1 \dots J-1 \tag{11}$$

where

$$\pi_{ij} = \frac{e^{\sum_{k=0}^{K} x_{ik}\beta_{kj}}}{1 + \sum_{j=1}^{J-1} e^{\sum_{k=0}^{K} x_{ik}\beta_{kj}}}, \quad j < J$$

$$\pi_{iJ} = \frac{1}{1 + \sum_{j=1}^{J-1} e^{\sum_{k=0}^{K} x_{ik}\beta_{kj}}}$$

For each row, the dependent variable will follow a multinomial distribution with J many groups. The joint probability density function (PDF) can be written as:

$$f(\boldsymbol{y}|\beta) = \prod_{i=1}^{N}\left[\frac{n_i!}{\prod_{j=1}^{J} y_{ij}!} \cdot \prod_{j=1}^{J} \pi_{ij}^{y_{ij}}\right]$$

When calculating the likelihood, the terms that do not contain a π_{ij} can be treated as constants. Therefore, the likelihood can be written as:

$$L(\beta|\boldsymbol{y}) \cong \prod_{i=1}^{N}\prod_{j=1}^{J} \pi_{ij}^{y_{ij}}$$

Replacing the J^{th} term allows the likelihood to be written as:

$$L(\beta|\boldsymbol{y}) = \prod_{i=1}^{N}\prod_{j=1}^{J-1} \pi_{ij}^{y_{ij}} \cdot \frac{\pi_{iJ}^{ni}}{\prod_{j=1}^{J-1}\pi_{iJ}^{y_{ij}}}$$

By grouping terms that are raised to the power of y_{ij}:

$$L(\beta|\boldsymbol{y}) = \prod_{i=1}^{N}\prod_{j=1}^{J-1}\left(\frac{\pi_{ij}}{\pi_{iJ}}\right)^{y_{ij}} \cdot {\pi_{ij}}_{i}^{n}$$

Substituting the values for π_{ij} and π_{iJ}, the likelihood becomes:

$$L(\beta|\boldsymbol{y}) = \prod_{i=1}^{N}\prod_{j=1}^{J-1}\left(e^{\sum_{k=0}^{K} x_{ik}\beta_{kj}}\right)^{y_{ij}} \cdot \left(\frac{1}{1+\sum_{j=1}^{J-1} e^{\sum_{k=0}^{K} x_{ik}\beta_{kj}}}\right)^{n_i}$$

Finally, simplifying the likelihood results in equation 12:

$$L(\beta|\boldsymbol{y}) = \prod_{i=1}^{N}\prod_{j=1}^{J-1} e^{y_{ij}\sum_{k=0}^{K} x_{ik}\beta_{kj}} \cdot \left(1+\sum_{j=1}^{J-1} e^{\sum_{k=0}^{K} x_{ik}\beta_{kj}}\right)^{-n_i} \quad (12)$$

Applying the natural log to the likelihood function and simplifying provides the log likelihood function (13) for the multinomial logistic mode:

$$l(\beta) = \sum_{i=1}^{N}\sum_{j=1}^{J-1} y_{ij}\sum_{k=0}^{K} x_{ik}\beta_{kj} - n_i \log\left(1+\sum_{j=1}^{J-1} e^{\sum_{k=0}^{K} x_{ik}\beta_{kj}}\right) \quad (13)$$

By maximizing the log likelihood with respect to β, parameter estimation can be performed.

The likelihood ratio test can be used to determine if an explanatory variable is statistically significant by testing if the k^{th} variable belongs in the model by testing the hypothesis:

$$H_0: \beta_{kj} = 0 \quad \text{vs} \quad H_A: \beta_{kj} \neq 0$$

Performing the likelihood ratio test involves calculating the log likelihood of the saturated model, $L_s(\hat{\theta})$, then dropping the k^{th} predictor variable and calculating the new log likelihood value, $L_g(\hat{\theta})$. The likelihood ratio test statistic (14) is calculated as:

$$\Lambda_k = -2\log\left(\frac{L_s(\hat{\theta})}{L_g(\hat{\theta})}\right) \tag{14}$$

Wilks theorem states that under the null hypothesis, the asymptotic distribution of the likelihood ratio test statistic is chi-square distribution with the degrees of freedom equal to the difference between the dimensionality of $L_s(\hat{\theta})$ and $L_g(\hat{\theta})$. For multinomial logistic regression this will result in a chi-square distribution with degrees of freedom equal to two (Queen Mary University of London n.d.; O'Halloran n.d.). Subsequent models will be fitted with the $K + 1^{st}$ predictor variable removed, and the log likelihood will be calculated to test the null hypothesis that the removed estimated coefficient is equal to zero. This will be repeated until all predictor variables have been checked.

McFadden's pseudo-R^2 (15), as defined in Cameron and Windmeijer (1996), is used to determine how well the proposed multinomial logistic model fits the data. McFadden's pseudo-R^2, treats the log likelihood of the intercept model as the total sums of squares and the log likelihood of the fitted model as the total sum square errors. A MFR^2_{adj} value close to zero indicates that the model has no predictive ability while values closer to 1 indicates that the fitted model does a good job explaining the data. When comparing multinomial logistic regression models the model with a higher McFadden's pseudo-R^2 better explains the data.

$$MFR^2_{adj} = 1 - \frac{\log\hat{L}(M_{fitted})}{\log\hat{L}(M_{intercept})}, \quad 0 \leq MFR^2_{adj} < 1 \tag{15}$$

The interpretation of the multinomial logistic model uses techniques found in Starkweather (2018) and Korosteleva (2018). Recall that the multinomial logistic regression model estimates $J - 1$ models where J is the total number of mutually exclusive values of the dependent variable. The J^{th} outcome is referred to as the reference category, to which all other outcomes are compared. The *relative risk ratio coefficient* indicates how the risk of the

observation belonging to the j^{th} outcome compares to the risk of the observation belonging to the reference category, J, changes with the variable in question.

A relative risk ratio larger than 1 indicates that the observation is more likely to belong to the j^{th} category then the reference category, while a relative risk ratio less than 1 indicates that the observation is more likely to belong to the reference category. The relative risk ratio coefficient can be calculated by exponentiating the coefficient estimates of the multinomial logistic model.

$$\text{Relative Risk Ratio} = e^{\beta_{kj}}, \quad \beta = 1 \dots K, j = 1 \dots J \tag{16}$$

The estimated regression coefficients obtained from multinomial logistic model are in the form of $\beta_{kj} \cdot \boldsymbol{X}_i = \log\left(\frac{\pi_{ij}}{\pi_{iJ}}\right)$ where $\boldsymbol{X}_i$ is the i^{th} row vector of predictor variables, β_{kj} is the vector of the maximum likelihood estimates and π_{ij} is the probability that the i^{th} observation belongs to the j^{th} category. Each observation in the dataset will have J many probabilities that correspond with the probability of that observation belonging to the j^{th} category.

Solving for π_{ij}, i.e. the probability that the i^{th} observation belongs to the j^{th} outcome:

$$\pi_{ij} = \left(\pi_{iJ}\right)\left(e^{\beta_{kj}\boldsymbol{X}_i}\right)$$

Applying the law of total probability, the probability that the i^{th} outcome belongs to the J^{th} category can be expressed as:

$$\pi_{iJ} = \frac{1}{1 + \sum_{j=1}^{J-1}\left(e^{\beta_{kj}\cdot\boldsymbol{X}_i}\right)}$$

Using this value for π_{iJ} the other probabilities can be calculated:

$$\pi_{ij} = \frac{e^{\beta_{kj}\boldsymbol{X}_i}}{1 + \sum_{j=1}^{J-1}\left(e^{\beta_{kj}\cdot\boldsymbol{X}_i}\right)}$$

4.4 Factor Analysis Using the Principal Component Method

Consider a dataset with p many random variables, $\boldsymbol{Y}^{\mathrm{T}} = [y_1, y_2, y_3, \dots, y_p]$, with a mean column vector $\boldsymbol{\mu}^T = [\mu_1, \mu_2, \mu_3, \dots, \mu_p]$ and variance/covariance matrix $\boldsymbol{\Sigma}(\underline{Y})$. The objective of factor analysis is to explain the covariance structure of the p many random variables in m many terms of unobservable variables called factors, notated as $f_1, f_2, f_3, \dots, f_m$ where $p \gg m$. To determine the size of *m,* the eigenvalues for $\boldsymbol{\Sigma}$ must first be calculated $\boldsymbol{\Lambda}^{\mathrm{T}} = [\lambda_1, \lambda_2, \dots, \lambda_p]$. Using the calculated eigenvalues, two of the most common methods for determining m are Kaiser's Stopping Rule, often abbreviated as the K1 rule, and the Percentage of Variance Criterion.

Kaiser's (1960) Stopping Rule states that *m* is the number of eigenvalues that are larger than 1. However, Guttman (1954) suggests that Kaiser's Stopping Rule should only be used as a lower bound for *m*. Further research performed by Humphreys (1964) and Mote (1970) also confirms that Kaiser's Stopping Rule can be too restrictive and suggests that Kaiser's Stopping Rule be used as a suggested lower bound for *m*. The Percentage of Variance Criterion (Johnson and Wichern 2007) states that *m* is the number of eigenvalues that explain a sufficient percentage of the total variance. The total variance explained (17) is the proportion of total variance explained using m many factors and the trace of the variance/covariance matrix.

$$\sum_{m=1}^{p} \frac{\lambda_m}{trace\{\boldsymbol{\Sigma}(\mathbf{Y})\}} \tag{17}$$

If performing factor analysis on a correlation matrix, $\boldsymbol{\rho}(\mathbf{Y})$, the trace of the matrix is the number of variables in the dataset. Therefore, the proportion of total variance explained is calculated as $\sum_{i=1}^{p} \frac{\lambda_i}{p}$. Note that both $\boldsymbol{\Sigma}(\mathbf{Y})$ and $\boldsymbol{\rho}(\mathbf{Y})$ are square $p \times p$ matrices.

The following derivation and interpretation of the factor model was adapted from material provided in Johnson and Wichern (2007), Kim (2019b), and Grimm and Yarnold (1995). The factor model (18) has three main assumptions and is traditionally expressed using matrix notation (19).

1. $\boldsymbol{f}_k \sim iid(0, \boldsymbol{I})$ *where* $k = 1, 2, \ldots, m$
2. $\boldsymbol{\nu} \sim (\mathbf{0}, \psi)$
3. $cov(\boldsymbol{f}_k, \boldsymbol{\nu}) = 0$

$$Y_j = \mu_j + \lambda_{j1} f_1 + \lambda_{j2} f_2 + \cdots + \lambda_{jm} f_m + \nu_j, \text{where } j = 1, 2, \ldots, p \tag{18}$$

$$\boldsymbol{Y}^{p\times 1} = \boldsymbol{\mu}^{p\times 1} + \boldsymbol{\Lambda}^{p\times m}\boldsymbol{f}^{m\times 1} + \boldsymbol{\nu}^{p\times 1} \tag{19}$$

Using the matrix notation of the factor model, the factor equation can be constructed.

$$\boldsymbol{\Sigma}(\mathbf{Y}) = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}} + \boldsymbol{\psi} \tag{20}$$

The estimated loading matrix $\boldsymbol{\Lambda}$ is calculated as $\boldsymbol{\Lambda^T} = [\sqrt{\widehat{\lambda_1}}\widehat{e_1}, \sqrt{\widehat{\lambda_2}}\widehat{e_2}, \ldots, \sqrt{\widehat{\lambda_m}}\widehat{e_m}]$ where $(\widehat{\lambda_1}, \widehat{e_1})$, $(\widehat{\lambda_2}, \widehat{e_2})$, $\ldots, (\widehat{\lambda_p}, \widehat{e_p})$ are the eigenvalue-eigenvector pairs found through spectral decomposition of $\boldsymbol{\Sigma}(\boldsymbol{Y})$. $\boldsymbol{\psi}$ is found by subtraction, $\boldsymbol{\psi} = \boldsymbol{\Sigma}(\boldsymbol{Y}) - \boldsymbol{\Lambda}\boldsymbol{\Lambda^T}$. Note that when performing factor analysis on the correlation matrix $\boldsymbol{\rho}(\boldsymbol{Y})$, the loading matrix $\boldsymbol{\Lambda}$ is the correlation between

$z_j = \frac{Y_j - \mu_j}{\sqrt{\sigma_{jj}}}$ and f_k, which is the correlation between each variable and its respective component.

Recall that $\boldsymbol{\Lambda}$ is the factor loading and is calculated using the eigenvalue-eigenvector pairs of $\boldsymbol{\Sigma}(\underline{Y})$. *Communality* notated as h_j^2, is the proportion of the variance in Y_j that is explained by its factors. Communality when performing factor analysis on $\boldsymbol{\Sigma}(\mathbf{Y})$ is calculated as $h_j^2 = \Sigma_{i=1}^{m} l_{ji}$ where $l_{j1}^2, l_{j2}^2, \ldots, l_{jm}^2$ correspond to the factor loadings of Y_j. Larger values of h_j^2

indicate that the variables' variance are well explained by their factors. The ratio of variance explained by h_j^2 can be calculated as $\frac{h_j^2}{\sigma_{jj}}$.

Communality when performing factor analysis on the correlation matrix, $\boldsymbol{\rho}(\boldsymbol{Y})$, is also calculated as $h_j^2 = \Sigma_{i=1}^{m} l_{ji}$ where $l_{j1}^2, l_{j2}^2, \ldots, l_{jm}^2$ correspond to the factor loadings of Y_j. The value for communality is between 0 and 1 where values of h_j^2 that are closer to 1 indicate that the variance for the j^{th} variable is well explained by its factors. The ratio of variance explained by h_j^2 can be calculated as $\frac{h_j^2}{p}$ where p is the original number of variables in the dataset. The specific variance of Y_j , ψ_j, is calculated as $\sigma_{jj}^2 - h_j^2$, or $1 - h_j^2$. A high value for ψ might indicate that the variance for Y_j is not well explained by its factors.

The VARIMAX rotation was developed by Kaiser (1958) and is used to find a rotation, or linear combination, of the original factor loadings with the goal of maximizing the variance between loadings. The process begins by calculated the scaled factor loadings which will be denoted as l_{ij}^*.

$$l_{ij}^* = \frac{l_{ij}}{h_i} \tag{21}$$

After the factor loadings are scaled the VARIMAX function is maximized.

$$V = \frac{1}{p}\sum_{j=1}^{m}\left\{\sum_{i=1}^{p}(l_{ij}^*)^4 - \frac{1}{p}\left(\sum_{i=1}^{p}(l_{ij}^*)^2\right)^2\right\} \tag{22}$$

Thurstone (1935) introduced the concept of using least squares regression to calculate factor scores where the independent variables in the regression equation are the original observed values. These predictor variables are then weighted by a regression coefficient which is the inverse of the correlation matrix $\boldsymbol{\rho}(\boldsymbol{Y})$ multiplied by the factor loading matrix $\boldsymbol{\Lambda}$. The factor score

is the dependent variable in the regression equation and is standardized with a mean of zero and a standard deviation of 1. The factor score is a numeric value that indicates how an individual is affected by a factor. These factor scores can be used to replace the initial p many columns with the m many factor score columns.

4.5 Fishers Linear Discriminate Analysis

The following derivation and interpretation of Fishers Linear Discriminate Analysis is based on material found in Kim (2019a), and Johnson and Wichern (2007). The following derivation will first establish Discriminate Analysis for 2 populations (or samples) and then will be extended to g many populations (or samples).

Consider a dataset with 2 mutually exclusive populations (or groups) labeled π_1 and π_2. Let $\boldsymbol{X}$ be a column vector of p many random variables in the dataset such that $\boldsymbol{X}^T = [X_1, X_2, \dots, X_p]$. Let $f_{1(\boldsymbol{x})}$ and $f_{2(\boldsymbol{x})}$ be probability density functions associated with a $p \times 1$ random vector of $\boldsymbol{X}$, representing the population of $\boldsymbol{x}$ values that belong to π_1 and π_2 respectively. Let Ω represent the entire sample space, and R_1 be the set of $\boldsymbol{x}$ values that are classified into π_1. Then $R_2 = \Omega - R_1$, and is the set of set of $\boldsymbol{x}$ values that are classified into π_2.

The conditional probability of misclassifying on object into π_2 when it belongs in π_1 is found using equation 23 below.

$$P(2|1) = P(\boldsymbol{X} \in R_2|\pi_1) = \int_{R_2=\Omega - R_1} f_1(\boldsymbol{x})d\boldsymbol{x} = 1 - \int_{R_1} f_1(\boldsymbol{x})d\boldsymbol{x} \qquad (23)$$

Likewise, the conditional probability of misclassifying on object into π_1 when it belongs in π_2 can be calculates using equation 24 below.

$$P(1|2) = P(\boldsymbol{X} \in R_1|\pi_2) = \int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x} = 1 - \int_{R_2} f_2(\boldsymbol{x})d\boldsymbol{x} \qquad (24)$$

Let p_1 and p_2 be the prior probabilities of π_1 and π_2 such that $p_1 + p_2 = 1$. Then the associated probabilities of correctly classifying an object into π_1 (26), correctly classifying an object into π_2 (27), misclassifying an object into π_1 (28), and misclassifying an object into π_2 (29) can be seen below:

$$P(\boldsymbol{X} \in R_1|\pi_1)P(\pi_1) = P(1|1)p_1 \tag{25}$$

$$P(\boldsymbol{X} \in R_2|\pi_2)P(\pi_2) = P(2|2)p_2 \tag{26}$$

$$P(\boldsymbol{X} \in R_1|\pi_2)P(\pi_2) = P(1|2)p_2 \tag{27}$$

$$P(\boldsymbol{X} \in R_2|\pi_1)P(\pi_1) = P(2|1)p_1 \tag{28}$$

The associated cost of misclassifying an object is expressed as c and can be seen in Table 8. There is a cost of 0 when an object is classified into the correct population, a cost of $c(2|1)$ when an observation from π_1 is classified into π_2, and a cost of $c(1|2)$ when an observation from π_2 is classified into π_1.

TABLE 8. Misclassification Costs with Two Populations

		Classified Population	
		π_1	π_2
True Population	π_1	0	$c(2\|1)$
	π_2	$c(1\|2)$	0

The Expected Cost of Misclassification (ECM) is the average, or expected, percentage of misclassifications after applying the classification rule. The ECM for two populations is calculated by multiplying the off diagonal entries in Table 8 by their associated probabilities of occurring. A reasonable classification rule should attempt to minimize the ECM (29) as much as possible.

$$\begin{aligned} \text{ECM} &= c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2 \\ &= p_1 c(2|1)\left[1 - \int_{R_1} f_1(\boldsymbol{x})d\boldsymbol{x}\right] + p_2 c(1|2)\int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x} \\ &= \int_{R_1} [p_2 c(1|2) f_2(\boldsymbol{x}) - p_1 c(2|1) f_1(\boldsymbol{x})]\, d\boldsymbol{x} + p_1 c(2|1) \end{aligned} \tag{29}$$

From (29) the ECM is minimized for regions R_1 and R_2 is such that $p_2 c(1|2) f_2(\boldsymbol{x}) \leq 0$. This is generally expressed using the inequalities in Equations 30 and 31.

$$R_1 : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \tag{30}$$

$$R_2 : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \tag{31}$$

To extend the ECM model to multiple populations let $f_i(\boldsymbol{x})$ be the density associated with population $\pi_i, i = 1, 2, \ldots, g$. Then, let p_i be the prior probability of population π_i, $c(k|i)$ is the cost of allocating an item to π_k when it belongs to π_i where $k = 1, 2, \ldots g$ and $k \neq i$. Lastly, let R_k be the set of $\boldsymbol{x}$'s classified as π_k and $P(k|i) = P(\text{classifying as } \pi_k|\pi_i) = \int_{R_k} f_i(\boldsymbol{x})d\boldsymbol{x}$ with $P(i|i) = 1 - \sum_{k=1}^{g} P(k|i), k \neq i$. The ECM of misclassifying an $\boldsymbol{x}$ from π_1 into $\pi_2, \pi_3, \ldots, \pi_g$ is found through equation 32:

$$ECM(1) = P(2|1)c(2|1) + P(3|1)c(c|1) + \cdots + P(g|1)c(g|1) = \sum_{k=2}^{g} P(k|1)c(k|1) \tag{32}$$

Similarly, the ECM costs for misclassifying an $\boldsymbol{x}$ from π_2 into $\pi_3, \pi_4, \ldots, \pi_g$ can be found and calculated as $\text{ECM}(2)$. Repeating this process for all g populations provides $\text{ECM}(1), \text{ECM}(2), \ldots, \text{ECM}(g)$. Multiplying each conditional ECM by its prior probability and summing provides the overall error rate.

Fisher's Linear method with multiple populations does not require that the g populations be Multivariate Normally distributed; however, the method does assume that the covariance

matrices are equal and full rank. Let $\boldsymbol{\bar{\mu}}$ denote the mean vector of the combined populations and $\boldsymbol{B_{\mu}}$ be the between group sums of cross products.

$$\boldsymbol{B_{\mu}} = \sum_{i=1}^{g}(\boldsymbol{\mu_i} - \boldsymbol{\bar{\mu}})(\boldsymbol{\mu_i} - \boldsymbol{\bar{\mu}})^T \text{ where } \boldsymbol{\bar{\mu}} = \frac{1}{g}\sum_{i=1}^{g}\boldsymbol{\mu}_i \tag{33}$$

Consider the linear combination $Y = \boldsymbol{a}^T\boldsymbol{X}$ and $E(Y) = \boldsymbol{a}^T E(\boldsymbol{X}|\pi_i) = \boldsymbol{a}^T\boldsymbol{\mu}_i$ for population π_i and $Var(Y) = \boldsymbol{a}^T Cov(\boldsymbol{X})\boldsymbol{a} = \boldsymbol{a}^T\boldsymbol{\Sigma a}$ for all populations. With a changing expected value, the overall mean, $\hat{\mu}_Y$, is defined as:

$$\hat{\mu}_Y = \frac{1}{g}\sum_{i=1}^{g}\mu_i Y = \frac{1}{g}\sum_{i=1}^{g}\boldsymbol{a}^T\boldsymbol{\mu}_i = \boldsymbol{a}^{T\left(\frac{1}{g}\sum_{i=1}^{g}\boldsymbol{\mu}_i\right)} = \boldsymbol{a}^T\boldsymbol{\bar{\mu}} \tag{34}$$

The overall variability between groups relative to the common variability is measured as the ratio of the sum of squared distance from each population to the overall mean of Y and the variance of Y.

$$\frac{\boldsymbol{a}^T\left(\sum_{i=1}^{g}(\boldsymbol{\mu_i} - \boldsymbol{\bar{\mu}})(\boldsymbol{\mu_i} - \boldsymbol{\bar{\mu}})^T\right)\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{\Sigma a}} \; or \; \frac{\sum_{i=i}^{g}(\mu_{iY} - \bar{\mu}_Y)^2}{\sigma_Y^2} = \frac{\boldsymbol{a}^T\boldsymbol{B_{mu}a}}{\boldsymbol{a}^T\boldsymbol{\Sigma a}} \tag{35}$$

Using a well-defined training sample with correctly classified observations, the sample mean vector (36), the overall average vector (37), the sample between groups matrix (38), and the estimate for $\boldsymbol{\Sigma}$ (39) can be calculated.

$$\boldsymbol{\bar{x}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}x_{ij} \tag{36}$$

$$\boldsymbol{\bar{x}} = \frac{1}{g}\sum_{i=1}^{g}\boldsymbol{\bar{x}}_i \tag{37}$$

$$\boldsymbol{B} = \sum_{i=1}^{g}(\boldsymbol{\bar{x}}_i - \boldsymbol{\bar{x}})(\boldsymbol{\bar{x}}_i - \boldsymbol{\bar{x}})^T \tag{38}$$

$$\boldsymbol{W} = \sum_{i=1}^{g}(n_i - 1)\boldsymbol{S}_i = \sum_{i=1}^{g}\sum_{j=1}^{n_j}(\boldsymbol{x}_{ij} - \boldsymbol{\bar{x}}_i)(\boldsymbol{x}_{ij} - \boldsymbol{\bar{x}}_i)^T \tag{39}$$

To derive Fisher's sample linear discriminate, let $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_s > 0$ denote the

$s \leq \min(g-1, p)$ nonzero eigenvalues of $\boldsymbol{W^{-1}B}$. Let $\hat{\boldsymbol{e}}_1, \hat{\boldsymbol{e}}_2, \ldots, \hat{\boldsymbol{e}}_s$ be the corresponding eigenvectors that are scaled so $\hat{\boldsymbol{e}}\boldsymbol{S}_{pooled}\hat{\boldsymbol{e}} = 1$. Then the vector of coefficients, $\hat{\boldsymbol{a}}$, that maximizes the ratio in (40) is given by $\hat{\boldsymbol{a}}_1 = \hat{\boldsymbol{e}}_1, \hat{\boldsymbol{a}}_2 = \hat{\boldsymbol{e}}_2, \ldots, \hat{\boldsymbol{a}}_k = \hat{\boldsymbol{e}}_k$. Where $\hat{\boldsymbol{a}}^T{}_1\boldsymbol{x}$ is the sample first discriminate, $\hat{\boldsymbol{a}}^T{}_2\boldsymbol{x}$ is the sample second discriminate, up to $\hat{\boldsymbol{a}}^T{}_k\boldsymbol{x}$ which is the k^{th} sample discriminate.

$$\frac{\hat{\boldsymbol{a}}^T\boldsymbol{B}\hat{\boldsymbol{a}}}{\hat{\boldsymbol{a}}^T\boldsymbol{W}\hat{\boldsymbol{a}}} = \frac{\hat{\boldsymbol{a}}^T\left(\sum_{i=1}^{g}(\overline{\boldsymbol{x}}_i - \overline{\boldsymbol{x}})(\overline{\boldsymbol{x}}_i - \overline{\boldsymbol{x}})^T\right)\hat{\boldsymbol{a}}}{\hat{\boldsymbol{a}}^T\left[\sum_{i=1}^{g}\sum_{j=1}^{n_j}(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_i)^T\right]\hat{\boldsymbol{a}}} \tag{40}$$

Using Fisher's classification procedure a new observation $\boldsymbol{x}$ is classified into π_k using equation 41 where r is the number of discriminates being used , $\tilde{y}_{kj} = \hat{\boldsymbol{a}}^T{}_j\overline{\boldsymbol{x}}_{\mathrm{i}}$ and $\hat{\boldsymbol{a}}^T{}_j$ is found from (40).

$$\sum_{j=1}^{r}\left(\hat{y} - \tilde{y}_{kj}\right)^2 = \sum_{j=1}^{r}\left[\hat{\boldsymbol{a}}^T{}_j(\boldsymbol{x} - \overline{\boldsymbol{x}}_k)\right]^2 \leq \sum_{j=1}^{r}\left[\hat{\boldsymbol{a}}^T{}_j(\boldsymbol{x} - \overline{\boldsymbol{x}}_i)\right]^2 \quad \forall\, i \neq k \tag{41}$$

CHAPTER 5

RESULTS

This study aims to answer two main research questions: Does major change have a significant effect on a student's graduation status, and can this result be combined with the students' academic and demographic information to identify significant factors that influence timely graduation?

5.1 Chi-Square Hypothesis Testing

Recall that students were grouped by the number of major changes where students in group "A" had no major changes while enrolled, students in group "B" had only one major change, students in group "C" had two major changes, and students in group "D" had at least 3 major changes. To test if the proportion of students who graduated is the same across all major change groups a Chi-Square test for homogeneity of proportions will be used.

Prior to applying the test, the assumptions of the test were checked by creating a contingency table for each cohort, which can be seen in Appendix C. After verifying the assumption that all expected counts were larger than 5, the chi-square test for homogeneity of proportions was applied at a significant level of $\alpha = 0.01$ with the following hypothesis:

H_0: The proportion of students who graduated was the same across all major change groups H_A: The proportion of students who graduated was not the same across all major change groups. A p-value less than 0.01 indicates that there is a statistically significant difference in the proportion of students who graduated between major change groupings.

The results of the chi-square test for homogeneity of proportions had a p-value less than 0.001 which indicated that there was a statistically significant difference between the proportions of students who graduated between at least one major change groupings. Post-hoc analysis using

pairwise chi-square tests for homogeneity of proportions was performed using a Bonferroni correction of $\alpha_B = \frac{0.01}{6} = 0.0017$. The results of the post hoc analysis is summarized in Table 9.

TABLE 9. Post Hoc Testing of χ^2 Test for Homogeneity of Proportions

χ^2 Test for Homogeneity of Proportions Categories	p-value
No Major Change vs One Major Change	$p < 0.001^*$
No Major Change vs Two Major Changes	$p < 0.001^*$
No Major Change vs More than Two Major Changes	$p < 0.001^*$
One Major Change vs Two Major Changes	$p = 0.63$
One Major Change vs More than Two Major Changes	$p < 0.001^*$
Two Major Changes vs More than Two Major Changes	$p < 0.001^*$

Note: An * indicates significance at $\alpha_B = 0.0017$

From the *p*-values shown in Table 9, it can be concluded that, at a significance level of $\alpha_B = 0.0017$, students who did not change majors graduated at a statistically significant different rate then students who changed majors. It is important to note that the chi-square test for homogeneity of proportions is a two-tailed test and cannot provide a clear answer on the directionality of the difference. Therefore, a correlation analysis was performed to determine how major change and graduation rates are related.

5.2 Correlation and Graduation Rates Analysis

The Phi coefficient of correlation (equation 8) was calculated between major change and student's graduation status at a significance level of $\alpha = 0.01$. For the major change variable, students who changed majors were coded as 1; for the graduated variable, students who graduated were coded as 1. At a significance level of $\alpha = 0.01$, there was a medium statistically

significant ($p < 0.001$) positive correlation ($r_\phi = 0.47$) between a student changing majors and their graduation status. In other words, students who changed majors are more associated with graduating than students who did not change majors.

The graduation rate and timely graduation rates were calculated for each major change grouping and recorded in Tables 10 and 11. From these tables, it can be concluded that students who changed majors at least once graduated at a substantially higher rate than students who did not change majors. The group of students who changed majors twice had the highest percentage of four-year graduates followed by the group of students who changed majors once. The group of students who changed majors at least three times had the highest percentage of six-year graduates. It is important to note that the data provided does not separate students who exited without a degree, and students who transferred to another institution. This might artificially increase the number of students classified as "Did Not Graduate". Additional tables containing information on graduation rates by cohort year can be found in Appendix D.

TABLE 10. Graduation Rates for Each Major Change Group

Major Change Group	Did not Graduate	Graduated Within Six Years
Did Not Change Majors	69%	31%
Changed Majors Once	23%	77%
Changed Majors Twice	24%	76%
Changed Majors More than Twice	14%	86%

TABLE 11. Timely Graduation Rates for Each Major Change Group

Major Change Group	Did not Graduate	Four-Year Graduate	Six-Year Graduate
Did Not Change Majors	69%	12%	19%
Changed Majors Once	23%	17%	60%
Changed Majors Twice	24%	21%	55%
Changed Majors More than Twice	14%	15%	71%

From the exploratory data analysis, it can be concluded, that for the sample of first-time freshman used in this study, major change activity does not reduce the overall graduation rate. First-time freshmen who changed majors at least once graduated at a higher rate than students who do not change majors. However, major change is associated with a longer time to graduation. Students who changed majors at most twice had the highest percentage of four-year graduates, while students who changed majors more than twice were more likely to graduate as six-year graduates. Based on these results it will be very difficult for a student to graduate within four years if the student changes majors more than twice.

5.3 Dimensionality Reduction

With high dimensional data, it is reasonable to assume that there exists a high degree of multicollinearity in the dataset as well as underlying factors that need to be explained before modeling and classification can occur. This section will focus on detecting multicollinearity using the variance inflation factor and examining the correlation structure of the data. Factor Analysis using the Principal Component method will be used to reduce the overall dimensionality of the dataset. The dataset created after reducing the dimensionality will be used as an input dataset for modeling and classification. With these goals in mind, 24 complete variables (no missing data) were chosen for creating the dataset.

To calculate the variance inflation factor (VIF), 3 binary logistic models need to be fit to the data. The dependent variables for the binary logistic models were created based on the results of the exploratory data analysis. From the Chi-Square test for homogeneity of proportions it was concluded that students who do not change majors graduate at a statistically significant different rate than students who do. Additionally, over half the students in each cohort changed majors at least twice. Using these results, the binary logistic models in Table 12 were created.

TABLE 12. Binary Logit Models Used to Calculate VIF Values

Model Number	Binary Dependent Variable
1	1 = No major change, 0 = Changed majors
2	1 = Changed majors only once 0 = Either no major change or more than 1 major change
3	1 = Changed majors more than once, 0 = Either no major change or 1 major change

The tables for the calculated VIF values can be found in Appendix E. A high VIF value ($VIF \geq 4$) indicates that the variance in the predicted coefficients are inflated by the existence of correlation between the predictor variables. The variables high school GPA, number of semesters enrolled, math standardized test score, reading standardized test score, eligibility index score, GE credits, and non-GE credits all had a VIF value larger than 4.

To investigate the correlation structure, the Pearson-Correlation Coefficient, Point-Biserial Correlation Coefficient, and the Phi-Coefficient of Correlation were used when appropriate to calculate the correlations between the variables used in the VIF analysis. The full table with all 190 correlations is in Appendix F. A heatmap of the lower triangle of the correlation values was also generated and can be seen in Appendix F. There were 25 pairs of variables with a correlation magnitude larger than 0.30. These correlations were examined, and the following 17 variables were retained.

TABLE 13. Variables Retained After Correlation Analysis

Categorical Variables	Continuous or Discrete Variables
STEM Admission	High School GPA
Student Gender	Overall CSULB GPA
Pell Eligibility	Math Standardized Test Score
Local Admission	Reading Standardized Test Score
Minority Admission	Eligibility Index Score
First Generation Student	GE or Non-Major Credits
Major Change Grouping	Major Credits
	Total Number of DFW Courses
	Total Number of WU Courses
	Total Number of WE Courses

Factor Analysis using the Principal Component method will only accept continuous variables as an input. Therefore, the 10 continuous variables in Table 13 were used for factor analysis. To determine the m many factors to extract, the eigenvalue/eigenvector pairs and percentage of explained variance were calculated for each of the 10 variables. A Scree Plot of the eigenvalues and a Cumulative Variance Scree Plot were created to determine the number of factors to extract. These results can be seen in Table 14 and Figure 1. Using the Percentage of Variance Criterion with a target of 80% explained variance, 5 factors will be extracted from the model.

TABLE 14. Eigenvalue and Percent Variance Explained

Eigenvalue Number	Eigenvalue	Cumulative Eigenvalue	Percent of Variance Explained	Cumulative Variance Explained
1	3.41	3.41	0.34	34%
2	1.72	5.13	0.17	51%
3	1.42	6.54	0.14	65%
4	0.97	7.51	0.10	75%
5	0.90	8.42	0.09	84%
6	0.55	8.97	0.06	90%
7	0.49	9.47	0.05	95%
8	0.37	9.84	0.04	98%
9	0.14	9.98	0.01	100%
10	0.02	10.00	0.00	100%

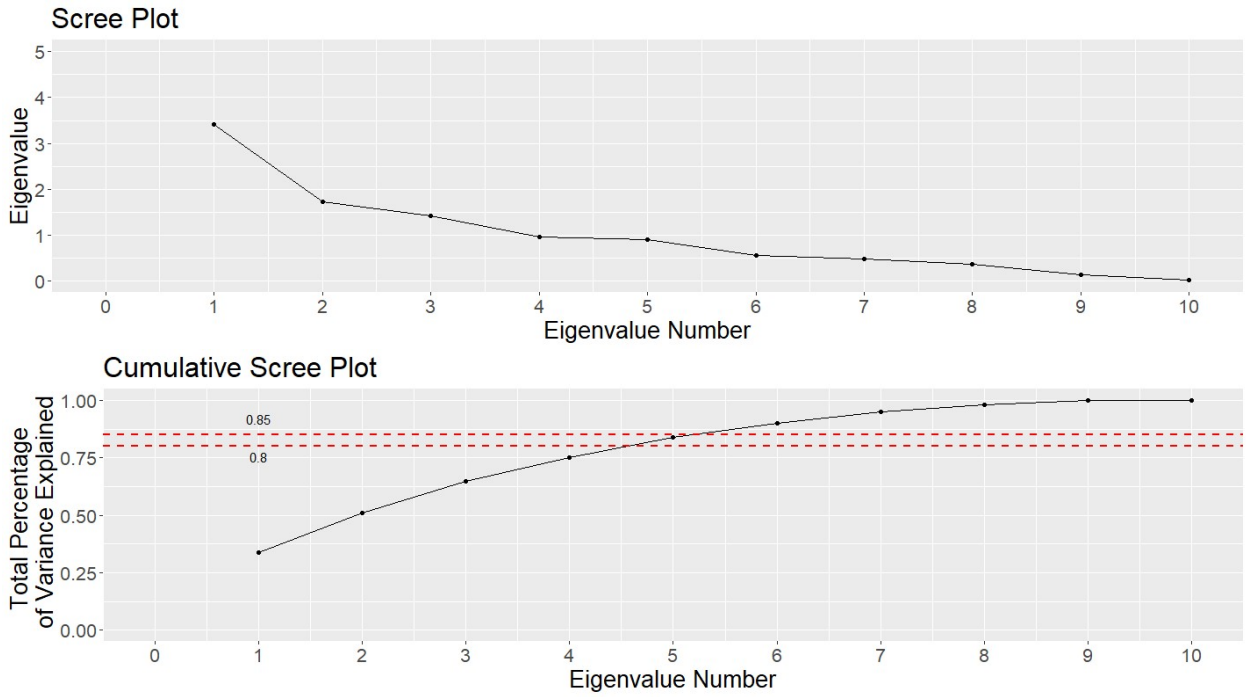


FIGURE 1. Scree and cumulative scree plots.

Using the "factanal" function inside the "stats" library in R 3.6.3, factor analysis was performed on the correlation matrix with $m = 5$. The VARIMAX rotation was performed to maximize the variance in the resulting factor loadings. Before observing the generated factor loadings, the uniqueness vector and residual matrix must be checked. Variables with high values of uniqueness indicate that these variables do not have information in common with other variables, while variables with low values of uniqueness might be grouped together when determining latent variables. The residual matrix is calculated as $S - \Lambda\Lambda^{-1} + \Psi$ where S is the sample correlation matrix, Λ is the matrix of factor loadings, and Ψ is a vector containing the diagonal of the uniqueness matrix. There is a single value close to 1 in the uniqueness table and only two values in the residual matrix have a value where $\varepsilon_{ij} > |0.05|, i = 1 \dots m, j = 1 \dots m$. From these results it can be concluded that using $m = 5$ is acceptable. The full residual matrix and uniqueness vector can be seen in Appendix G.

By examining the generated factor loadings, latent variables are created using the highest loading values. For example, consider Table 15 which contains the factor loadings for Factor 1. When compared to the rest of the factors, Factor 1 had the highest loadings on High School GPA and Eligibility Index Score. From this it can be said that Factor 1 explains the latent variable *Academic Preparation*. Using similar methodology, the latent variables for the other 4 factors were determined. The 5 factors extracted are summarized in Table 16 below. The full output of factor loadings can be seen in Appendix H.

TABLE 15. Factor Loadings for Factor 1

Variable Name	Factor Loading for Factor 1
High School GPA	0.98
Overall CSULB GPA	0.26
Math Standardized Test Score	0.20
Reading Standardized Test Score	0.20
Eligibility Index Score	0.86
Major Credits	0.12
Total Number of DFW Courses	-0.19

TABLE 16. Factors and Their Associated Latent Variables

Factor Number	Latent Variable Explained
Factor 1	Academic Preparation
Factor 2	CSULB Academic Information
Factor 3	Credits Taken
Factor 4	Math Standardized Test Score
Factor 5	Reading Standardized Test Score

After the latent variables were identified, the *Factor Score* was calculated using least squares regression as outlined by Thurstone (1935). These factor scores will be used to replace the initial 10 columns with the 5 extracted latent variables. The first 10 rows of the factors

scores are provided in Table 17 for reference. It is important to note that the extracted factors are linearly independent which addresses the multicollinearity that was detected using the VIF.

TABLE 17. First 10 Rows of Calculated Factor Scores

Observation	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
1	0.38	−0.15	0.02	1.14	1.99
2	1.60	1.66	−0.76	2.47	−3.32
3	−0.35	0.34	−3.13	−0.60	−1.30
4	0.25	0.40	1.11	−0.48	1.40
5	−0.08	1.73	1.14	0.18	0.52
6	−0.84	1.05	1.55	0.67	−1.32
7	0.44	1.75	−2.14	−0.75	−0.63
8	−1.12	−1.82	−0.98	−0.42	−0.22
9	−0.97	−0.54	−0.44	0.47	0.58
10	1.05	−0.33	−0.08	−0.63	−0.94

5.4 Multinomial Logistic Model

A multinomial logistic model is used to describe the relationship between predictor variables and a multi-level dependent variable whose outcomes are mutually exclusive. For this study, the multinomial logistic model was used to determine factors that influence a student's time to graduation. The multinomial dependent variable was created based on a student's timely graduation status $\underline{Y}$ = [1: Did not Graduate, 2: Four-Year Graduate, 3: Six-Year Graduate] with the reference category set to $Y = 2$. The major change grouping variable was replaced with the number of major changes a student had. This was done to avoid the high multicollinearity between the major change groupings and timely graduation. Table 18 contains the variables used in the Multinomial Logistic model.

TABLE 18. Variables Used in the Multinomial Logistic Model

Categorical Variables	Continuous and Discrete Variables
STEM Admission	Factor 1 - Academic Preparation
Student Gender	Factor 2 - CSULB Academic Information
Pell Eligibility	Factor 3 - Credits Taken
Local Admission	Factor 4 - Math Standardized Test Score
Minority Admission	Factor 5 - Reading Standardized Test Score
First Generation Student	Number of Major Changes

The multinomial logistic model does not accept categorical variables as an input; therefore, the categorical variables were converted to binary indicator variables. These variables are often called "dummy" variables in the sense that the 0 or 1 values are artificial and do not convey any meaningful information on their own. For the STEM admission variable, students who were admitted as a STEM major were coded as 1 and students who were either accepted as undeclared or into a non-STEM major were coded as 0. The student gender variable was coded as 1 for students who self-identified as female and 0 for students who self-identified as male.

The Pell eligibility variable, used in this study as socioeconomic status, was coded with 1 for students admitted who were Pell eligible and 0 for students admitted who were not Pell eligible. To indicate local admission, students were coded with 1 if they were considered a local admission and 0 if they were a non-local admission. The racial/ethnic minority indicator variable was coded as 1 for students who self-identified as a minority and 0 for students who self-identified as a non-minority.

To fit the multinomial logistic model, the package "nnet" and the function "multinom" were used in R 3.6.3. To reduce the complexity of the final model and to prevent overfitting, the *p*-values for the estimated coefficients were calculated using the likelihood ratio test after each model was fitted. The estimated coefficient with the largest *p*-value greater than or equal to 0.05

was removed and the model was re-fitted. This was repeated until all estimated coefficients had *p*-values less than 0.05. This technique is commonly referred to as backward elimination.

After 1 fitting, the estimated coefficients for the predictor variable First Generation Status was found to be not statistically significantly different from zero, and all other predictor variables had a *p*-values less than 0.05. The final model had a McFaddens Adjusted R^2 value of 0.51 which indicates that the model is a reasonably good fit. It is important to note that multinomial logistic model will return $J - 1$ many models where J is the number of mutually exclusive outcomes the dependent variable can take on. For the complete model fitting analysis, please see Appendix I.

The first model, equation 42, generated by the multinomial logistic model represents the relationship between a student being classified as "Did not Graduate" versus the student being classified as a "Four Year Graduate." The second model, equation 43, generated by the multinomial logistic model represents the relationship between a student being classified as a "Six Year Graduate" versus the student being classified as a "Four Year Graduate."

$$\ln\left(\frac{P(Y = \text{Did not Graduate})}{P(Y = \text{Four Year Graduate})}\right) \tag{42}$$

$$\begin{aligned} &= 2.43 + 0.99(\text{STEM ADMISSION}) + 0.16\ (\text{STUDENT GENDER}) \\ &\quad + \ 0.12(\text{PELL ELIGIBILITY}) - 0.18(\text{LOCAL ADMISSION STATUS}) \\ &\quad - 0.47(\text{MINORITY STATUS}) - 0.76(\text{FACTOR 1}) + 3.03(\text{FACTOR 2}) \\ &\quad - 3.76(\text{FACTOR 3}) - 0.47(\text{FACTOR 4}) - 0.90(\text{FACTOR 5}) \\ &\quad - 0.42(\text{NUMBER OF MAJOR CHANGES}) \end{aligned}$$

$$\ln\left(\frac{P(Y = \text{Six Year Graduate})}{P(Y = \text{Four Year Graduate})}\right) \tag{43}$$

$$= 1.01 + 0.36\,(\text{STEM ADMISSION}) - 0.36(\text{STUDENT GENDER})$$
$$+ 0.30(\text{PELL ELIGIBILITY}) + 0.21(\text{LOCAL ADMISSION STATUS})$$
$$- 0.01(\text{MINORITY STATUS}) - 0.22(\text{FACTOR 1}) + 0.79(\text{FACTOR 2})$$
$$+ 0.73(\text{FACTOR 3}) - 0.19(\text{FACTOR 4}) - 0.57(\text{FACTOR 5})$$
$$+ 0.12(\text{NUMBER OF MAJOR CHANGE}$$

By exponentiating the estimated coefficients, the relative risk ratios can be calculated for each of the modeled outcomes. Tables 19 and 20 contains the relative risk ratios for the multinomial logistic model. A relative risk ratio larger than 1 indicates that the parameter is associated with a student in the comparison category. A relative risk ratio close to 1 indicates that the parameter has no effect on the student's risk of belonging to either category. A relative risk ratio less than 1 indicates the parameter is associated with a student in the reference category.

The most significant risk factor that contributed to students not graduating when compared to graduating in four years was the student's value in the latent variable CSULB Academic Information. The CSULB Academic Information variable measures the student's CSULB GPA, and the number of DFW, WU, and WE courses the student had while enrolled at CSULB. High values in the CSULB Academic Information variable indicates that the student performed poorly while enrolled at CSULB. A one-unit increase in the students CSULB Academic Information variable corresponds to an approximate decrease of 0.70 in the students CSULB GPA, and an approximant one course increase in the number of the students DFW, WU, and WE courses. Female students were slightly more likely to exit CSULB before earning a degree when compared to earning a degree in four years.

Students who were admitted as STEM Majors are 1.44 times more likely to graduate as a six-year graduate instead of a four-year graduate, Pell Eligible and Local students are, respectively, 1.36 and 1.24 times more likely to graduate as six-year graduates instead of four-year graduates. When comparing four- and six-year graduates, a one-unit increase in the students' CSULB Academic Information are 2.20 times more likely to finish as six-year graduates instead of four-year graduates.

Interestingly, as the number of major changes increased students became less at risk of not graduating, and more at risk of graduating in six years. This is consistent with what was concluded after hypothesis testing and correlation analysis. Students who were more academically prepared for college, measured in Factors 1, 4, and 5, were more likely to graduate in four years instead of dropping out or graduating in six years.

TABLE 19. Relative Risk Ratios for Did Not Graduate Versus Four Year Graduate

Variable Name	Y = Did Not Graduate vs Four Year Graduate
STEM Admission	1.10
Student Gender	1.18
Pell Eligibility	1.14
Local Admission	0.84
Minority Admission	0.62
Factor 1 - Academic Preparation	0.47
Factor 2 - CSULB Academic Information	20.64
Factor 3 - Credits Taken	0.02
Factor 4 – Math Standardized Test Score	0.63
Factor 5 - Reading Standardized Test Score	0.41
Number of Major Changes	0.66

Note: Values larger than 1 indicate the student is more likely to not graduate than be a four-year graduate.

TABLE 20. Relative Risk Ratios for Six Year Graduate Versus Four Year Graduate

Variable Name	Y = Six Year Graduate vs Four Year Graduate
STEM Admission	1.44
Student Gender	0.70
Pell Eligibility	1.36
Local Admission	1.24
Minority Admission	0.99
Factor 1 - Academic Preparation	0.80
Factor 2 - CSULB Academic Information	2.20
Factor 3 - Credits Taken	2.08
Factor 4 – Math Standardized Test Score	0.83
Factor 5 - Reading Standardized Test Score	0.56
Number of Major Changes	1.12

Note: Values larger than 1 indicate the student is more likely be a six-year graduate than a four-year graduate.

After fitting the model, the data were separated into a testing and training dataset to assess how accurately the model can classify students based on the estimated coefficients. The data were randomly split with 70% of the data reserved for training and the remaining 30% used for testing the accuracy of the model. The training data contained 6909 observations, the testing data contained 1596 observations. After training, the model had an overall accuracy rate of 75.46%. The confusion matrix can be seen in Table 21.

TABLE 21. Confusion Matrix for The Multinomial Logistic Model

	Predicted Four-Year Graduate	Predicted Did Not Graduate	Predicted Six-Year Graduate
True Four-Year Graduate	106	29	315
True Did Not Graduate	63	662	124
True Six-Year Graduate	51	44	1157

5.5 Classification Using Fishers Linear Discriminant

The finish answering the research question, Fishers Linear Discriminant Analysis (LDA) was used to develop a classification function for timely graduation. Fishers LDA was used to develop a classification functions between the three graduation statuses used in the multinomial

logistic model: “Did Not Graduate”, “Four-Year Graduate”, and “Six-Year Graduate”. The variables used for Fishers LDA can be seen in Table 22

TABLE 22. Variables Used in Fishers LDA

Categorical Variables	Continuous or Discrete
STEM Admission	Number of Major Changes
Student Gender	Factor 1 - Academic Preparation
Undeclared Admission	Factor 2 - CSULB Academic Information
Pell Eligibility	Factor 3 - Credits Taken
Local Admission	Factor 4 - Math Standardized Test Score
Minority Admission	Factor 5 - Reading Standardized Test Score
First Generation Student	
Time to Graduation Grouping	

Prior to implementing Fishers LDA the data should be checked to determine if response groups are multivariate normally distributed with approximately equivalent covariance matrices. However, with mixed data types these assumptions cannot be satisfied. Research performed by Ethel Gilbert (1968) concluded that using Fishers LDA with qualitative variables and a sample size larger than 500 is acceptable. Therefore, this section of the study was performed without checking if the data satisfy the assumptions of Fishers LDA.

The Fishers LDA model was fit using the package “MASS” and function “lda” in R 3.6.3. The three populations were set as $\pi_1 =$ “Did Not Graduate”, $\pi_2 =$ “Four-Year Graduate”, and $\pi_3 =$ “Six-Year Graduate”. Prior to implementing Fishers LDA the data were split into a testing and training dataset with 80% used for training and 20% used for training. Without any a priori knowledge about the distribution of graduation rates for first-time freshman at CSULB, the prior probabilities for the testing data were set to the proportion of each population with respect to the entire test dataset. Therefore, $p_1 = 0.35357$, $p_2 = 0.1611$, and $p_3 = 0.4854$ with $n_{\pi_1} = 2406, n_{\pi_2} = 1096,$ and $n_{\pi_3} = 3303$.

After applying Fishers LDA to the testing dataset, the proportion of variance explained by each linear discriminant was calculated. The first linear discriminant (LD1) explained 90.86% of the between group variance, with the remaining 9.14% of the between group variance explained by the second linear discriminant (LD2). The coefficients of the linear discriminates are then standardized to have a mean of 0 and standard deviation of 1. The magnitude of the coefficients indicates how strong the discriminating variable effects the discriminate score. For LD1 the variables with the largest magnitude are: Factor 3 – Credits Taken (1.0944) and Factor 2 - CSULB Academic Information (0.7131). For LD2 the variables with the largest magnitude are: Undeclared Admission (0.5943), Factor 2 - CSULB Academic Information (0.5219), Factor 5 – Reading Standardized Test Score (0.5139) and Stem Admission (0.4355). The full output from Fishers LDA can be seen in Appendix J.

The equations for LD1 and LD2 are given below:

$$\begin{aligned} LD1 = {} & 0.5033(\text{STEM Admission}) - 0.1157(Student\ Gender) \\ & + 0.1355(Undeclared\ Admission) + 0.0662(Pell\ Eligibility) \\ & + 0.1112(Local\ Admission) + 0.1722(Minority\ Admission) \\ & + 0.0130(First\ Generation\ Student) \\ & + 0.2955(Number\ of\ Major\ Changes) + 0.2198(Factor\ 1) \\ & - 0.7131(Factor\ 2) + 1.0944(Factor\ 3) + 0.0874(Factor\ 4) \\ & + 0.1205(Factor\ 5) \end{aligned}$$

$$LD2 = 0.4335(\text{STEM Admission}) - 0.2910(Student\ Gender) + 0.5943(Undeclared\ Admission) + 0.2684(Pell\ Eligibility) + 0.2374(Local\ Admission) - 0.0856(Minority\ Admission) + 0.0835(First\ Generation\ Student) + 0.2159(Number\ of\ Major\ Changes) + 0.2859(Factor\ 1) + 0.5219(Factor\ 2) + 0.2711(Factor\ 3) - 0.2126(Factor\ 4) + 0.5139(Factor\ 5)$$

After determining the equation for each linear discriminate, the discriminate score for LD1 and LD2 was calculated for each observation. The population of students who were pre-classified as "Did Not Graduate" had the lowest LD1 scores with an average score of -1.5594. Students who were pre-classified as "Four-Year Graduate" and "Six-Year Graduate" had essentially the same distribution of LD1 scores based on their mean and standard deviations.

These results are similar to what was seen in the multinomial logistic model. The multinomial logistic model found that high values in the students CSULB Academic information variable, and a value of 1 in the student gender variable increased the student's risk of dropping out. The same results can be seen in LD1. The only negative coefficients in the LD1 score equation are the CSULB Academic information variable, and the student gender variable. Therefore, negative LD1 scores correspond to high values in the students CSULB Academic information variable, and a value of 1 in the student gender variable.

The second linear discriminant separates students who were pre-classified as "Four-Year Graduate" and "Six-Year Graduate". "Four-Year Graduate" students had a mean LD2 score of -0.7913 while "Six-Year Graduate" students had a mean LD2 score of 0.2606. These results again match the risk factors described in the multinomial logistic model. A negative LD2 score

is associated with high academic preparation, and high standardized test scores and these variables were associated with students who were more likely to graduate as a "Four-Year Graduate" in the multinomial logistic model.

The scatterplot in Figure 3A displays the boundary lines between the three populations. After applying the decision boundaries to the training dataset, a scatterplot was created (Figure 3B) to visualize the misclassification rate. In this scatterplot each observation was plotted and classified using their LD1 and LD2 scores. The observations were then color coded based on their true classified value. There is very little overlap between students who did not graduate (red) and the other graduation statuses; however, there is substantial overlap between four-year graduates (green) and six-year graduates (blue). The classification of the training dataset had an accuracy rate of 75.42%. The corresponding confusion matrix is in Table 23

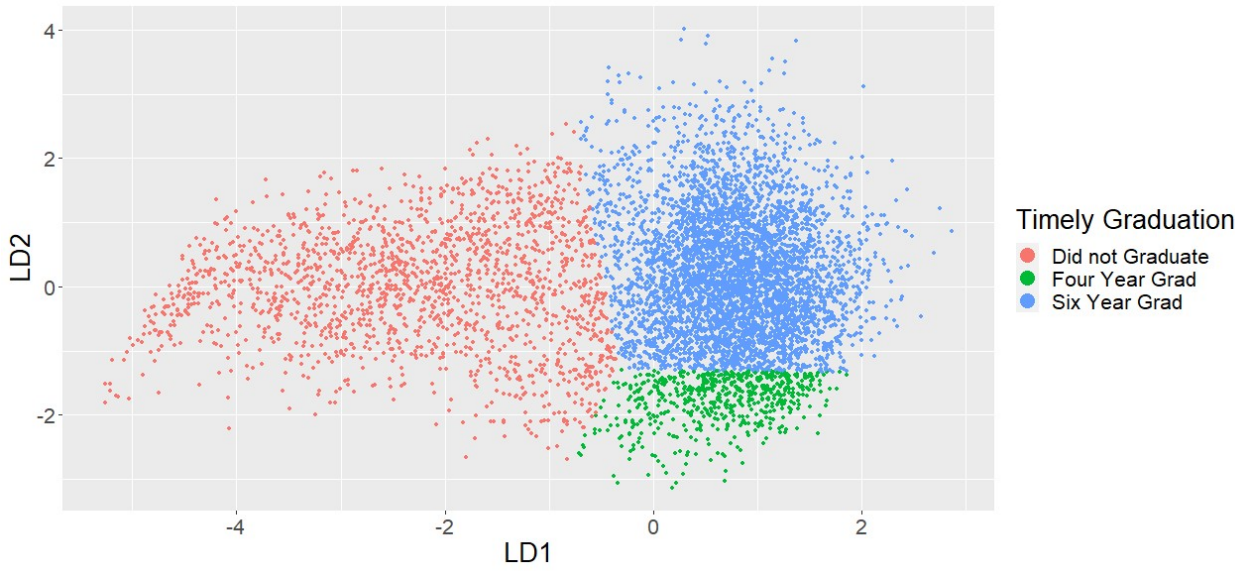
The discriminant scores were then calculated for the testing dataset and observations were plotted and classified using their LD1 and LD2 scores. Each observation was then color coded based on their true classified value (Figure 3C). Using the decision boundaries to classify new observations in the testing dataset resulted in an accuracy rate of 74.65%. The corresponding confusion matrix is in Table 24. In the testing dataset students who did not graduate were well classified, while four- and six-year graduates were not as well separated.

TABLE 23. Confusion Matrix for Fishers LDA; Training Dataset

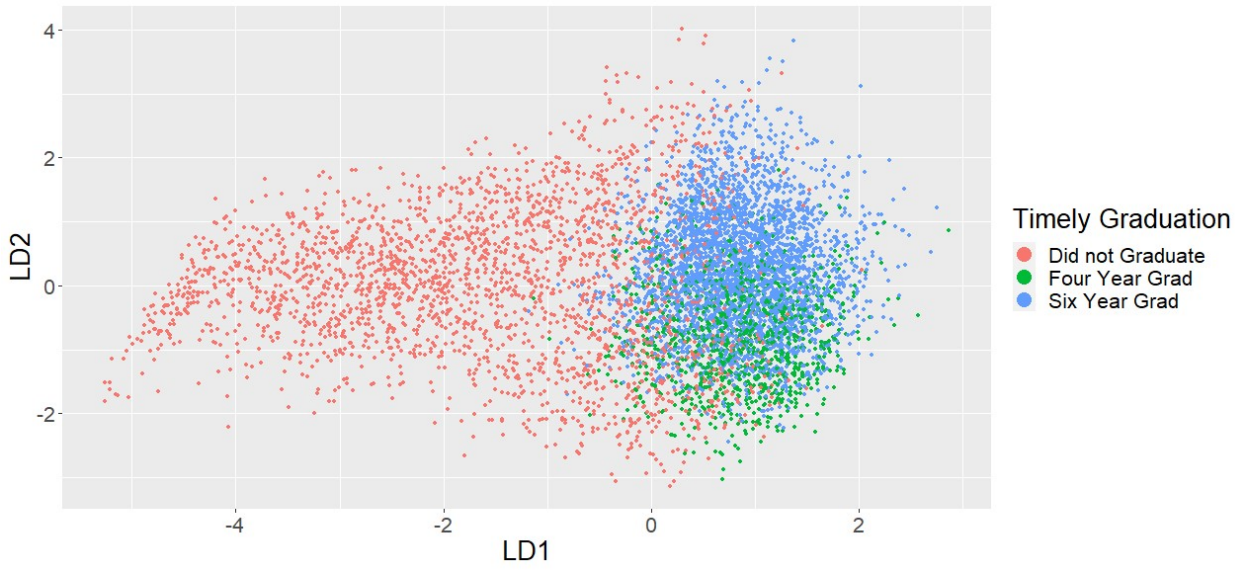
	Predicted Four-Year Graduate	Predicted Did Not Graduate	Predicted Six-Year Graduate
True Four-Year Graduate	1691	11	23
True Did Not Graduate	181	292	131
True Six-Year Graduate	534	793	314

TABLE 24. Confusion Matrix for Fishers LDA; Testing Dataset

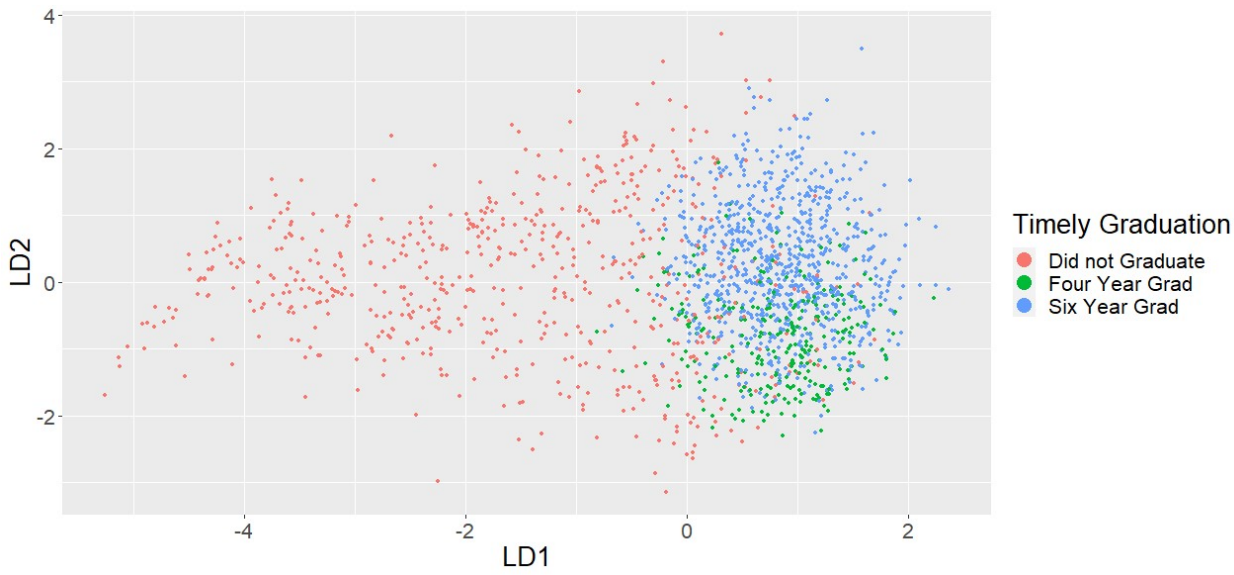
	Predicted Four-Year Graduate	Predicted Did Not Graduate	Predicted Six-Year Graduate
True Four-Year Graduate	409	3	3
True Did Not Graduate	49	72	34
True Six-Year Graduate	143	199	788



A: Classification Boundaries



B: Classification of the Training Data



C: Classification of the Testing Data

FIGURE 2. Scatterplots produced from Fishers LDA.

Using Fishers Linear Discriminant to classify new student is very straightforward. First, the factor scores need to be calculated and then, the LD1 and LD2 values can be calculated. The student can then be plotted on the decision boundary plot and classified. Table 25 below contains an example student and their associated academic and demographic information. After calculating the students factor scores, Table 26 was created which contains the variables used for Fishers Linear Discriminant Analysis. Using the linear discriminant functions, the student's linear discriminant scores were calculated as the ordered pair (1.2071, 1.5173). Using the ordered pair, the student is plotted on the boundary classification. From Figure 3, by plotting (1.2071, 1.5173), this student would be classified as a 6-year graduate.

TABLE 25. Example Student Raw Data

Categorical Variables	Variable Value
STEM Admission	0
Student Gender	0
Pell Eligibility	1
Local Admission	0
Minority Admission	0
First Generation Student	0
Major Change Grouping	C (2 Major Changes)
High School GPA	3.63
Overall CSULB GPA	3.26
Math Standardized Test Score	710
Reading Standardized Test Score	760
Eligibility Index Score	4374
GE or Non-Major Credits	64
Major Credits	65
Total Number of DFW Courses	0
Total Number of WU Courses	0
Total Number of WE Courses	0
Timely Graduation	Six-Year Graduate

TABLE 26. Example Student Variables and Factor Scores

Categorical Variables	Continuous and Discrete Variables
STEM Admission	0
Student Gender	0
Pell Eligibility	1
Local Admission	0
Minority Admission	0
First Generation Student	0
Factor 1 - Academic Preparation	0.3834
Factor 2 - CSULB Academic Information	-0.1500
Factor 3 - Credits Taken	0.0178
Factor 4 - Math Standardized Test Score	1.1365
Factor 5 - Reading Standardized Test Score	1.9898
Number of Major Changes	2

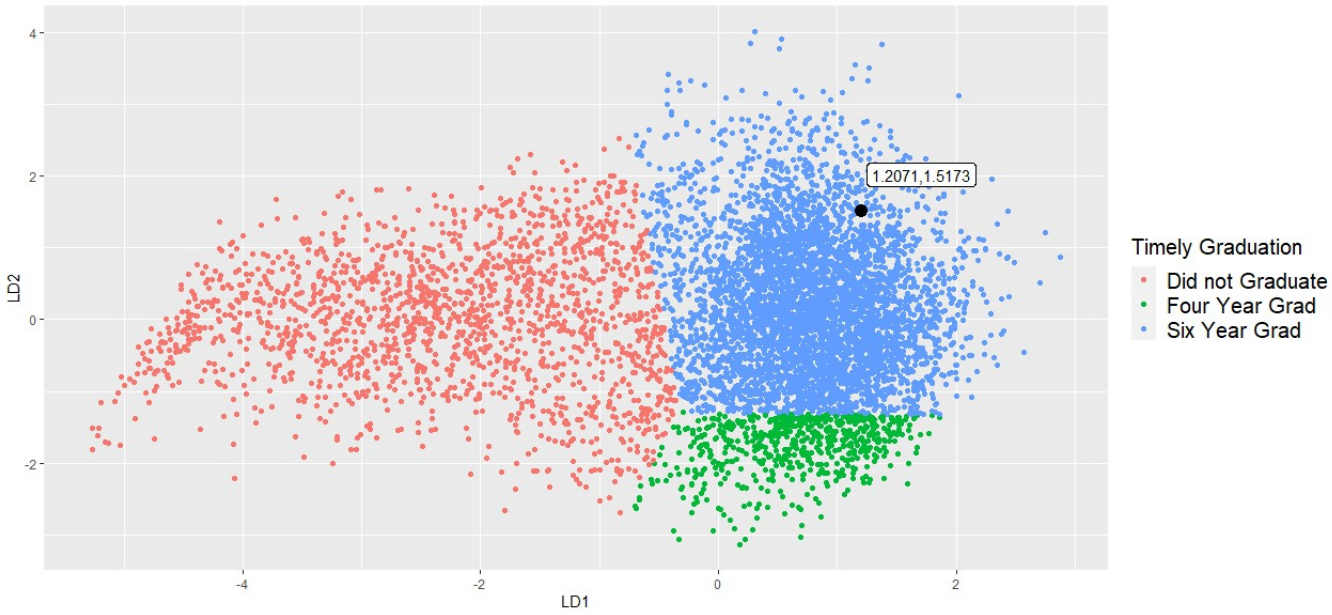


FIGURE 3. Example of using LD1 and LD2 to classify a new observation.

CHAPTER 6

CONCLUSIONS

6.1 Summary

This study aims to answer the question Does major change have a significant effect on a student's graduation status, and can this result be combined with the students' academic and demographic information to identify significant factors that influence timely graduation? An initial sample of 16,468 first time freshmen who were admitted between 2009 and 2012 was obtained from the CSULB Department of Institutional Research and Analytics after obtaining approval from Institutional Review Board. After controlling for students who did not fit the scope of this study and performing exploratory data analysis which included dimensionality reduction, a sample of 8,505 first-time freshmen and 17 variables were used in the study.

Hypothesis testing and correlation analysis were used to determine if there was any significant relationship between the number of major changes a student had and timely graduation. Dimensionality reduction using the Variance Inflation Factor and Factor Analysis was performed to reduce the complexity of the data and uncover latent variables in the data. This reduced dataset was then passed into a Multinomial Logistic model to determine statistically significant variables and their associated relative risk ratios. Lastly, the reduced dataset was used to develop a classification rule that can be used to predict a student's timely graduation status using Fishers Linear Discriminant.

6.2 Conclusions

Exploratory data analysis showed that students who changed majors at least once graduated at over double the rate of students who did not change majors. Additionally, the group of students who changed majors either once or twice had the highest percentage of four-year

graduates, and students who changed majors at least three times had the highest percentage six-year graduates. The group of students who did not change majors had the highest percentage of students who exited without completing a degree. However, the data provided does not separate students who exited without a degree, and students who transferred to another institution. This might artificially increase the number of students classified as "Did Not Graduate".

To determine if these differences were statistically significant, a Chi-Square test for the homogeneity of proportions was applied at a significance level of $\alpha = 0.01$. Post-hoc testing was performed by applying pairwise Chi-Square tests with Bonferroni's Correction at a significance level of $\alpha_B = 0.0017$. The tests confirmed that the graduation rates of students who did not change majors was statistically significantly different than the graduation rates of students who did change majors. This conclusion was reinforced by calculating the Phi Coefficient of Correlation between major change and graduation status for all cohort years. The Phi Coefficient of Correlation showed a strong positive correlation between major change and graduation status. Students who changed majors were more associated with graduating than students who did not change majors.

After determining that major change activity was associated with a higher graduation rate, a multinomial logistic model was fit to the data to determine the most significant risk factors that contributed to a student not graduating or graduating in six years versus graduating in four years. At a significance level of $\alpha = 0.05$, the most significant factor that increased the student's risk of not graduating was the student's value in the latent variable CSULB Academic Information. This variable is a weighted linear combination of the students CSULB GPA, and the number of DFW, WU, and WE courses the student had while enrolled at CSULB. High values in this variable suggests that the students did not perform well while enrolled at CSULB.

Students with a high value in this variable were 20 times more likely to exit without completing a degree than to graduate within four years. Students with lower values in the latent variable CSULB Academic Information were twice as likely to graduate in six years instead of four years.

Students admitted as STEM Majors, Pell Eligible students, and local admission students are more likely to graduate as six-year graduates instead of four-year graduates. STEM majors were 1.44 times more likely to graduate within six years. Pell Eligible students were 1.36 times more likely to graduate within six years, and local students were 1.24 times more likely to graduate within six years instead of four. The multinomial logistic model also suggested that as the number of major changes increased students became less at risk of not graduating, and more at risk of graduating in six years. This is consistent with what was concluded after hypothesis testing and correlation analysis.

Lastly, Fishers LDA was applied to the dataset to develop a classification method that can be used to predict a student's graduation status. Fishers LDA with 3 populations ("Did Not Graduate", "Four-Year Graduate", and "Six-Year Graduate") resulted in a classification model that correctly classified 75.42% of the training dataset and 74.65% of the training dataset. The first linear discriminate (LD1) was used to separate students who did not graduate from those who did.

Students who did not graduate had an average LD1 score of -1.5594 and LD1 placed the highest negative weight on the student's value in the latent variable CSULB Academic Information. The second linear discriminant (LD2) was used to separate four-and six-year graduates. Four-year graduates had an average LD2 score of -0.7913 while six-year graduates had an average LD2 score of 0.2606. Students who were classified as four-year graduates had

higher levels of academic preparation and have higher standardized test scores than students who graduated as six year graduates.

These results suggest that applying Fishers LDA to the student population and calculating each student's discriminant score each semester will provide meaningful information to Academic Advising. Students will enroll at CSULB with a value of zero in both linear discriminants, and their discriminate scores can be updated each semester. If the student's LD1 score develops a negative trend, Academic Advising can contact these students and provide needed support to help retain these students. The student's average LD2 score can be calculated to determine if a student is on track to graduate within four years. Additionally, the average LD1 and LD2 scores can also be calculated for each college and major within CSULB to create a profile of students within each major. Stakeholders can then focus on colleges and majors where students traditionally have negative LD1 scores and positive LD2 scores.

6.3 Comments for Future Research

In the provided dataset 7,355 students were excluded based on Summer or Winter semester enrollment. Future research should study these students and determine if there is a statistically significant relationship between Summer or Winter semester enrollment and timely graduation. It also might prove beneficial to determine if there is a relationship between Summer or Winter enrollment and major change. The dataset does not separate students who exited by transferring to another institution or who exited and did not continue their college education. Future studies should attempt to identify students who transferred to other institutions and attempt re-work the models to incorporate this parameter.

The choice to implement "classical" statistics in this study over current machine learning techniques was a deliberate choice. Machine learning is often referred to as "black box analysis"

which refers to the automation and complexity of the algorithms. Many popular machine learning techniques such as a Support Vector Machine (SVM), or Gradient Boosted Decision Trees (XG Boost) can be used to classify observations into two or more groups; however, the methodology behind these algorithms is very difficult to understand and the results can be difficult to interpret. Therefore, for this pilot study, "classical" statistics was used to ensure that the methodology implemented is easily understood and the results are interpretable and reproduceable. Future studies should implement and compare the results from both SVM and XG Boost to the results obtained in this study to determine if machine learning algorithms offer any benefit over "classical" statistics.

Fishers LDA uses pre-classified data to determine how to group observations together, and to develop a system for classifying new observations. It might be beneficial to use a classification method such a Partitioning Around the Medoids (Kaufman and Rosseeuw 1990) that does not rely on pre-classified data. Partitioning Around the Medoids is a clustering algorithm primarily used to cluster mixed datatypes and has been shown to work well with educational data (Zeidenberg and Scott 2011) when using Gowers Distance (Gower 1971) to determine how similar observations are to each other. This method will allow the patterns within the data to be discovered without influencing the results through pre-classification.

From Fall 2020 to Fall 2024, the University of California system will not use either the SAT or ACT test scores in their admission process (Gordon 2020). This will likely result in the University of California system placing more emphasis on a student's high school GPA, extracurricular activities, and letters of recommendations. In anticipation of the California State University system implanting the same policy, the multinomial logistic model, and Fishers LDA should be reworked to exclude SAT/ACT scores along with the students Eligibility Index score.

APPENDICES

APPENDIX A

DESCRIPTIVE STATISTCS FOR ACADEMIC VARIABLES

GRADUATION RATE FOR EACH COHORT YEAR

Cohort Year	Did Not Graduate Rate	Four-Year Graduation Rate	Six-Year Graduation Rate
2009	36.60%	16.46%	46.93%
2010	35.60%	14.57%	49.83%
2011	36.12%	15.78%	48.06%
2012	32.85%	17.89%	49.26%

GRADUATION RATES FOR 2009 COHORT YEAR BY MAJOR CHANGE LEVEL

Major Change Level	Did Not Graduate Rate	Four-Year Graduation Rate	Six-Year Graduation Rate
A	68.95%	11.27%	19.77%
B	24.84%	15.29%	59.87%
C	22.20%	23.14%	54.65%
D	13.72%	13.98%	72.30%

GRADUATION RATES FOR 2010 COHORT YEAR BY MAJOR CHANGE LEVEL

Major Change Level	Did Not Graduate Rate	Four-Year Graduation Rate	Six-Year Graduation Rate
A	72.71%	11.36%	15.93%
B	24.32%	16.22%	59.46%
C	22.52%	16.71%	60.77%
D	13.73%	14.70%	71.57%

GRADUATION RATES FOR 2011 COHORT YEAR BY MAJOR CHANGE LEVEL

Major Change Level	Did Not Graduate Rate	Four-Year Graduation Rate	Six-Year Graduation Rate
A	69.40%	10.65%	19.95%
B	20.96%	16.77%	62.28%
C	26.28%	21.38%	52.35%
D	14.89%	13.90%	71.21%

GRADUATION RATES FOR 2012 COHORT YEAR BY MAJOR CHANGE LEVEL

Major Change Level	Did Not Graduate Rate	Four-Year Graduation Rate	Six-Year Graduation Rate
A	63.75%	13.48%	22.77%
B	24.44%	18.89%	56.67%
C	26.15%	22.25%	51.60%
D	13.66%	16.53%	69.81%

HIGH SCHOOL GPA FOR EACH COHORT YEAR

Cohort Year	Mean HS GPA	Median HS GPA	Standard Deviation
2009	3.440	3.450	0.395
2010	3.440	3.460	0.365
2011	3.417	3.420	0.372
2012	3.475	3.510	0.396

SAT/ACT MATH SCORES FOR EACH COHORT YEAR

Cohort Year	Mean SAT/ACT Math Score	Median SAT/ACT Math Score	Standard Deviation
2009	562.36	560	85.58
2010	559.03	560	84.36
2011	556.76	560	83.06
2012	584.60	580	72.72

SAT/ACT READING SCORES FOR EACH COHORT YEAR

Cohort Year	Mean SAT/ACT Reading Score	Median SAT/ACT Reading Score	Standard Deviation
2009	558.58	560	82.84
2010	556.07	560	85.00
2011	553.98	560	82.78
2012	569.55	560	76.87

ELIGIBILITY INDEX SCORES FOR EACH COHORT YEAR

Cohort Year	Mean SAT/ACT Reading Score	Median SAT/ACT Reading Score	Standard Deviation
2009	3852.39	3850	386.49
2010	3842.24	3862	362.69
2011	3818.56	3792	361.45
2012	3904.83	3912	359.36

ELIGIBILITY INDEX SCORES FOR EACH MAJOR CHANGE GROUP

2009 Cohort

Major Change Group	Mean Eligibility Index	Median Eligibility Index	Standard Deviation
A	3818.07	3810	407.10
B	3768.60	3744	325.93
C	3908.85	3924	381.65
D	3848.33	3848	372.47

2010 Cohort

Major Change Group	Mean Eligibility Index	Median Eligibility Index	Standard Deviation
A	3817.01	3838	382.17
B	3814.34	3809	357.34
C	3859.50	3893	346.80
D	3855.28	3848	363.97

2011 Cohort

Major Change Group	Mean Eligibility Index	Median Eligibility Index	Standard Deviation
A	3806.85	3772	371.63
B	3771.55	3730	343.60
C	3838.14	3830	362.46
D	3818.59	3786	352.61

2012 Cohort

Major Change Group	Mean Eligibility Index	Median Eligibility Index	Standard Deviation
A	3883.72	3890	365.10
B	3906.60	3903	394.17
C	3945.08	3964	360.84
D	3876.62	3866	343.13

CSULB GPA

Cohort Year	Mean HS GPA	Median HS GPA	IQR
2009	2.51	2.72	1.18
2010	2.51	2.72	1.17
2011	2.46	2.68	1.21
2012	2.66	2.86	1.09

CSULB GPA FOR EACH MAJOR CHANGE GROUP

2009 Cohort

Major Change Group	Mean CSULB GPA	Median CSULB GPA	Standard Deviation
A	2.02	2.16	1.19
B	2.62	2.68	0.72
C	2.79	2.90	0.71
D	2.78	2.80	0.62

2010 Cohort

Major Change Group	Mean CSULB GPA	Median CSULB GPA	Standard Deviation
A	2.02	2.00	1.14
B	2.73	2.90	0.74
C	2.70	2.83	0.76
D	2.76	2.79	0.60

2011 Cohort

Major Change Group	Mean CSULB GPA	Median CSULB GPA	Standard Deviation
A	1.95	1.95	1.19
B	2.66	2.86	0.81
C	2.66	2.80	0.79
D	2.73	2.74	0.60

2012 Cohort

Major Change Group	Mean CSULB GPA	Median CSULB GPA	Standard Deviation
A	2.21	2.48	1.14
B	2.83	2.93	0.75
C	2.84	2.99	0.76
D	2.82	2.87	0.58

NUMBER OF DFW COURSES FOR EACH COHORT YEAR

Cohort Year	Mean DFW COURSES	Median DFW COURSES	Standard Deviation
2009	3.12	2	3.54
2010	3.12	2	3.54
2011	3.41	2	3.68
2012	3.14	2	3.63

NUMBER OF WE COURSES FOR EACH COHORT YEAR

Cohort Year	Mean WE COURSES	Median WE COURSES	Standard Deviation
2009	0.03	0	0.20
2010	0.03	0	0.21
2011	0.03	0	0.19
2012	0.03	0	0.20

NUMBER OF WU COURSES FOR EACH COHORT YEAR

Cohort Year	Mean WU COURSES	Median WU COURSES	Standard Deviation
2009	0.27	0	0.64
2010	0.25	0	0.61
2011	0.26	0	0.61
2012	0.26	0	0.61

UNDECLARED ADMISSION FOR EACH COHORT YEAR

Cohort Year	Undeclared Admission	Mean Number of Undeclared Semesters	Median Number of Undeclared Semesters	Standard Deviation
2009	$346_{(19\%)}$	2.85	2	1.55
2010	$334_{(15.85\%)}$	3.05	3	1.64
2011	$366_{(13.79\%)}$	3.02	3	1.60
2012	$266_{(13.63\%)}$	2.93	3	1.61

Note: The number in parenthesis denotes the percentage of students at a specified group for that cohort year.

UNDECLARED ADMISSIONS FOR EACH MAJOR CHANGE GROUP

Major Change Group	2009 Cohort	2010 Cohort	2011 Cohort	2012 Cohort
A	$69_{11.27\%}$	$61_{9.62\%}$	$77_{9.42\%}$	$46_{8.38\%}$
B	$0_{0\%}$	$0_{0\%}$	$0_{0\%}$	$0_{0\%}$
C	$86_{13.35\%}$	$84_{10.40\%}$	$89_{9.28\%}$	$99_{13.77\%}$
D	$34_{8.97\%}$	$41_{7.93\%}$	$33_{4.63\%}$	$31_{5.23\%}$

Note: The number in parenthesis denotes the percentage of students at a specified group for that cohort year.

APPENDIX B

TABLES OF DEMOGRAPHIC INFORMATION

NUMBER OF FEMALE AND MALE STUDENTS FOR EACH COHORT YEAR

Cohort Year	Number of Male Students	Number of Female Students
2009	$733_{(41\%)}$	$1059_{(59\%)}$
2010	$900_{(43\%)}$	$1207_{(57\%)}$
2011	$1127_{(42\%)}$	$1528_{(58\%)}$
2012	$914_{(47\%)}$	$1037_{(53\%)}$

NUMBER OF FEMALE AND MALE STUDENTS IN EACH MAJOR CHANGE GROUPS

2009 Cohort

Major Change Group	Number of Male Students	Number of Female Students
A	$257_{(42\%)}$	$355_{58\%)}$
B	$64_{(41\%)}$	$93_{(59\%)}$
C	$263_{(41\%)}$	$381_{(59\%)}$
D	$281_{(44\%)}$	$353_{(56\%)}$

Note: The number in parenthesis denotes the percentage of students at a specified group for that cohort year.

2010 Cohort

Major Change Group	Number of Male Students	Number of Female Students
A	$284_{(44\%}$	$353_{(56\%)}$
B	$69_{(47\%)}$	$79_{(53\%)}$
C	$340_{(42\%)}$	$468_{(58\%)}$
D	$210_{(41\%)}$	$307_{(59\%)}$

Note: The number in parenthesis denotes the percentage of students at a specified group for that cohort year.

2011 Cohort

Major Change Group	Number of Male Students	Number of Female Students
A	$393_{(48\%)}$	$424_{(52\%)}$
B	$68_{(41\%)}$	$99_{(59\%)}$
C	$387_{(40\%)}$	$572_{(60\%)}$
D	$279_{(39\%)}$	$433_{(61\%)}$

Note: The number in parenthesis denotes the percentage of students at a specified group for that cohort year.

2012 Cohort

Major Change Group	Number of Male Students	Number of Female Students
A	$298_{(54\%}$	$251_{(46\%}$
B	$38_{(42\%)}$	$52_{(58\%)}$
C	$339_{(47\%)}$	$380_{(53\%)}$
D	$239_{(40\%)}$	$354_{(60\%)}$

Note: The number in parenthesis denotes the percentage of students at a specified group for that cohort year.

CODING FOR PARENT EDUCATION GROUPS

Coded Values	Parent Education Group
8	Postgraduate
7	4-Year College Graduate
6	2-Year College Graduate
5	Some College
4	High School Graduate
3	Some Highschool
2	No Highschool
1	No Response

PERCENTAGE OF FIRST-GENERATION STUDENTS

Cohort Year	Not First- Generation	First-Generation
2009	$844_{(47\%)}$	$948_{(53\%)}$
2010	$887_{(42\%)}$	$1220_{(58\%)}$
2011	$1146_{(43\%)}$	$1509_{(57\%)}$
2012	$883_{(45\%)}$	$1068_{(55\%)}$

Note: The number in parenthesis denotes the number of students at a specified group for that cohort year

PERCENTAGE OF FIRST-GENERATION STUDENTS FOR MAJOR CHANGE GROUPS

Major Change Group	2009 Cohort	2010 Cohort	2011 Cohort	2012 Cohort
A	$47\%_{(285)}$	$57\%_{(363)}$	$57\%_{(465)}$	$54\%_{(298)}$
B	$43\%_{(68)}$	$69\%_{(102)}$	$65\%_{(108)}$	$62\%_{(56)}$
C	$50\%_{(319)}$	$58\%_{(468)}$	$56\%_{(533)}$	$55\%_{(393)}$
D	$45\%_{(172)}$	$56\%_{(287)}$	$57\%_{(403)}$	$54\%_{(321)}$

Note: The number in parenthesis denotes the number of students at a specified group for that cohort year

PELL ELIGIBILITY PERCENTAGE FOR EACH COHORT YEAR

Cohort Year	Pell Eligible Students	Non-Pell Eligible Students
2009	$51\%_{(914)}$	$49\%_{(848)}$
2010	$62\%_{(1310)}$	$38\%_{(797)}$
2011	$62\%_{(1654)}$	$38\%_{(1001)}$
2012	$60\%_{(1167)}$	$40\%_{(784)}$

Note: The number in parenthesis denotes the number of students at a specified group for that cohort year

PELL ELIGIBILITY PERCENTAGE FOR MAJOR CHANGE GROUPS

Major Change Group	2009 Cohort	2010 Cohort	2011 Cohort	2012 Cohort
A	$46\%_{(279)}$	$58\%_{(370)}$	$59\%_{(480)}$	$55\%_{(300)}$
B	$56\%_{(88)}$	$66\%_{(97)}$	$74\%_{(123)}$	$70\%_{(63)}$
C	$50\%_{(321)}$	$62\%_{(504)}$	$60\%_{(580)}$	$60\%_{(428)}$
D	$60\%_{(226)}$	$66\%_{(339)}$	$66\%_{(471)}$	$63\%_{(376)}$

Note: The number in parenthesis denotes the number of students at a specified group for that cohort year.

HIGH SCHOOLS ELIGIBLE FOR PREFERENTIAL LOCAL ADMISSION

City	School Name
Artesia	EPHS
Cerritos	Jordan
Gahr	Lakewood
Whitney	Millikan
Cypress	McBride
Oxford Academy	Polytechnic
Bellflower	Reid
Mayfair	Renaissance
Centennial	Sato Academy
Compton	Wilson
Dominguez	Beach High
Downey	Los Alamitos
Warren	Paramount
Edison	Brethren Christian
Fountain Valley	Calvary Chapel
Huntington Beach	Opportunities for Learning
Marina	Parkridge
Ocean View	Pius X/St. Matthias
Westminster	St. Anthony
Avalon	St. Joseph
Cabrillo	St. John Bosco
CAMS	Valley Christian

NUMBER OF LOCAL ADMISSIONS IN EACH COHORT YEAR

Cohort Year	Local Admission	Non-Local Admission
2009	$914_{(51\%)}$	$848_{(49\%)}$
2010	$1310_{(62\%)}$	$797_{(38\%)}$
2011	$1654_{(62\%)}$	$1001_{(38\%)}$
2012	$1167_{(60\%)}$	$784_{(40\%)}$

Note: The number in parenthesis denotes the number of students at a specified group for that cohort year.

PERCENTAGE OF LOCAL ADMISSION FOR MAJOR CHANGE GROUPS IN EACH COHORT YEAR

Major Change Group	2009 Cohort	2010 Cohort	2011 Cohort	2012 Cohort
A	$40\%_{(245)}$	$37\%_{(236)}$	$37\%_{(301)}$	$43\%_{(237)}$
B	$48\%_{(76)}$	$38\%_{(56)}$	$51\%_{(85)}$	$44\%_{(40)}$
C	$47\%_{(301)}$	$43\%_{(345)}$	$45\%_{(431)}$	$46\%_{(331)}$
D	$46\%_{(173)}$	$46\%_{(237)}$	$45\%_{(319)}$	$48\%_{(285)}$

Note: The number in parenthesis denotes the number of students at a specified group for that cohort year

RACIAL/ETHNIC OPTIONS STUDENTS CAN SELF-IDENTIFY AS AT ADMISSION

Racial/Ethnic Options	Minority Status
Hispanic/Latino	Minority
Black/African American	Minority
American Indian	Minority
Alaskan Native	Minority
White	Non-Minority
Asian	Non-Minority
Native Hawaiian	Non-Minority
Other Pacific Islander	Non-Minority
Two or More Races	Non-Minority

PERCENTAGE OF MINORITY ADMISSIONS IN EACH COHORT YEAR

Cohort Year	Minotiry Admission	Non-Minority Admission
2009	$42.75\%_{(766)}$	$57.25\%_{(1026)}$
2010	$46.99\%_{(990)}$	$53.01\%_{(117)}$
2011	$49.94\%_{(1326)}$	$50.06\%_{(1329)}$
2012	$45.05\%_{(879)}$	$54.95\%_{(1072)}$

Note: The number in parenthesis denotes the number of students at a specified group for that cohort year

PERCENTAGE OF MINORITY ADMISSION FOR MAJOR CHANGE GROUPS IN EACH COHORT YEAR

Major Change Group	2009 Cohort	2010 Cohort	2011 Cohort	2012 Cohort
A	$42\%_{(256)}$	$48\%_{(307)}$	$53\%_{(433)}$	$47\%_{(259)}$
B	$46\%_{(72)}$	$51\%_{(75)}$	$54\%_{(90)}$	$52\%_{(47)}$
C	$42\%_{(272)}$	$45\%_{(362)}$	$47\%_{(450)}$	$43\%_{(307)}$
D	$44\%_{(166)}$	$48\%_{(246)}$	$50\%_{(353)}$	$45\%_{(266)}$

Note: The number in parenthesis denotes the number of students at a specified group for that cohort year.

APPENDIX C

TABLES USED FOR CHI-SQURE ANALYSIS

CONTINGENCY TABLES FOR MAJOR CHANGE GROUPING

	No major changes (A)	One major change (B)	Two major changes (C)	At least 3 major changes (D)	Total
Did not Graduated	$1800_{(923.49)}$	$132_{(168.70)}$	$765_{(1106.63)}$	$310_{(778.18)}$	3007
Graduate	$812_{(1688.51)}$	$430_{(363.30)}$	$2365_{(2023.67)}$	$1891_{(1422.82)}$	5498
Total	2612	562	3130	2201	8505

Note: The numbers in parentheses denote the expected count for the respective cell

APPENDIX D

TABLES FOR TIMELY GRADUATION

TIMELY GRADUATION FOR EACH MAJORS CHANGE GROUPED BY COHORT YEAR

2009 Cohort

Major change Group	Did not Graduate	Four Year Graduate	Six Year Graduate
A	$68.95\%_{(422)}$	$11.27\%_{(69)}$	$19.77\%_{(121)}$
B	$24.84\%_{(39)}$	$15.29\%_{(24)}$	$59.87\%_{(94)}$
C	$22.20\%_{(143)}$	$23.14\%_{(149)}$	$54.66\%_{(352)}$
D	$13.72\%_{(52)}$	$13.98\%_{(23)}$	$72.30\%_{(274)}$

Note: The numbers in parentheses denote the number of students in that group

2010 Cohort

Major change Group	Did not Graduate	Four Year Graduate	Six Year Graduate
A	$72.71\%_{(461)}$	$11.36\%_{(72)}$	$15.93\%_{(101)}$
B	$24.32\%_{(36)}$	$16.22\%_{(24)}$	$59.46\%_{(88)}$
C	$22.52\%_{(182)}$	$16.71\%_{(135)}$	$60.77\%_{(491)}$
D	$13.73\%_{(71)}$	$14.70\%_{(76)}$	$71.57\%_{(370)}$

Note: The numbers in parentheses denote the number of students in that group

2011 Cohort

Major change Group	Did not Graduate	Four Year Graduate	Six Year Graduate
A	$69.40\%_{(567)}$	$10.65\%_{(87)}$	$19.95\%_{(163)}$
B	$20.96\%_{(35)}$	$16.77\%_{(28)}$	$62.28\%_{(104)}$
C	$26.28\%_{(252)}$	$21.38\%_{(205)}$	$52.35\%_{(502)}$
D	$14.89\%_{(106)}$	$13.90\%_{(99)}$	$71.21\%_{(507)}$

Note: The numbers in parentheses denote the number of students in that group

2012 Cohort

Major change Group	Did not Graduate	Four Year Graduate	Six Year Graduate
A	$63.75\%_{(350)}$	$13.48\%_{(74)}$	$22.77\%_{(125)}$
B	$24.44\%_{(22)}$	$18.89\%_{(17)}$	$56.67\%_{(51)}$
C	$26.15\%_{(188)}$	$22.25\%_{(160)}$	$51.60\%_{(371)}$
D	$13.66\%_{(81)}$	$16.53\%_{(98)}$	$69.81\%_{(414)}$

Note: The numbers in parentheses denote the number of students in that group

APPENDIX E

VARIANCE INFLATION VALUE CALCULATIONS

BINARY LOGIT MODELS USED TO CALCULATE VIF VALUES

Model Number	Binary Dependent Variable
1	1 = No major change, 0 = Changed majors
2	1 = Changed majors only once 0 = Either no major change or more than 1 major change
3	1 = Changed majors more than once, 0 = Either no major change or 1 major change

VIF Values

Variable Name	VIF Model 1	VIF Model 2	VIF Model 3
High School GPA	52.44	10.41×10^8	54.69
Number of Semesters Enrolled	21.93	16.70	19.79
STEM Entrant	3.69	2.54	4.05
STEM Graduate	2.75	1.15	3.08
CSULB GPA	4.57	4.30	4.43
Max ACT/SAT MATH	8.44	67.45×10^5	8.49
Max ACT/SAT Reading	6.02	70.92×10^6	6.15
Eligibility Index Score	83.96	14.54×10^8	85.89
Gender	1.20	1.19	1.21
GE Credits	8.00	7.12	7.64
Non-GE Credits	10.02	7.45	8.90
Number of Undeclared Semesters	4.26	1.40	4.81
Undeclared Admission	4.34	1.43	4.67
Total Number of DFW Courses	2.52	2.64	2.47
Total Number of WU Courses	1.42	1.43	1.40
Total Number of WE Courses	1.07	1.09	1.06
Pell Eligibility Status	1.27	1.32	1.28
Local Admission	1.16	1.14	1.16
Minority Admission	1.31	1.33	1.31
First Generation Student	1.32	1.33	1.32

APPENDIX F

CORRELATION ANALYSIS

CORRELATION VALUES

Variable 1	Variable 2	Correlation
Eligibility Index Score	High School GPA	0.89
Undeclared Admission	Number of Undeclared Semesters	0.86
STEM Graduate	STEM Admission	0.86
GE Credits Taken	Semesters Enrolled	0.83
Non-GE Credits Taken	Semesters Enrolled	0.83
Reading Standardized Test Score	Math Standardized Test Score	0.59
Eligibility Index Score	Math Standardized Test Score	0.58
Eligibility Index Score	Reading Standardized Test Score	0.57
Non-GE Credits Taken	CSULB GPA	0.56
CSULB GPA	Semesters Enrolled	0.53
Eligibility Index Score	CSULB GPA	0.50
Non-GE Credits Taken	GE Credits Taken	0.46
Total WU Courses	Total DFW Courses	0.44
CSULB GPA	High School GPA	0.44
GE Credits Taken	CSULB GPA	0.42
First Generation Student	Pell Eligible Admission	0.38
First Generation Student	Minority Admission	0.36
Reading Standardized Test Score	CSULB GPA	0.32
Minority Admission	Pell Eligible Admission	0.30
Math Standardized Test Score	CSULB GPA	0.29
Math Standardized Test Score	High School GPA	0.24
Total DFW Courses	GE Credits Taken	0.24
Reading Standardized Test Score	High School GPA	0.24
Non-GE Credits Taken	Eligibility Index Score	0.21
Non-GE Credits Taken	High School GPA	0.20
Math Standardized Test Score	STEM Graduate	0.20
Math Standardized Test Score	STEM Admission	0.19
Student Gender	High School GPA	0.16
Pell Eligible Admission	Total DFW Courses	0.15
Number of Undeclared Semesters	GE Credits Taken	0.15
Minority Admission	Total DFW Courses	0.14
Non-GE Credits Taken	Math Standardized Test Score	0.14
Pell Eligible Admission	GE Credits Taken	0.14
First Generation Student	Total DFW Courses	0.13
Local Admission	GE Credits Taken	0.12
Pell Eligible Admission	Semesters Enrolled	0.12
Student Gender	CSULB GPA	0.12
Local Admission	Total DFW Courses	0.11
Total DFW Courses	STEM Admission	0.10
Total DFW Courses	Semesters Enrolled	0.10
Semesters Enrolled	High School GPA	0.10

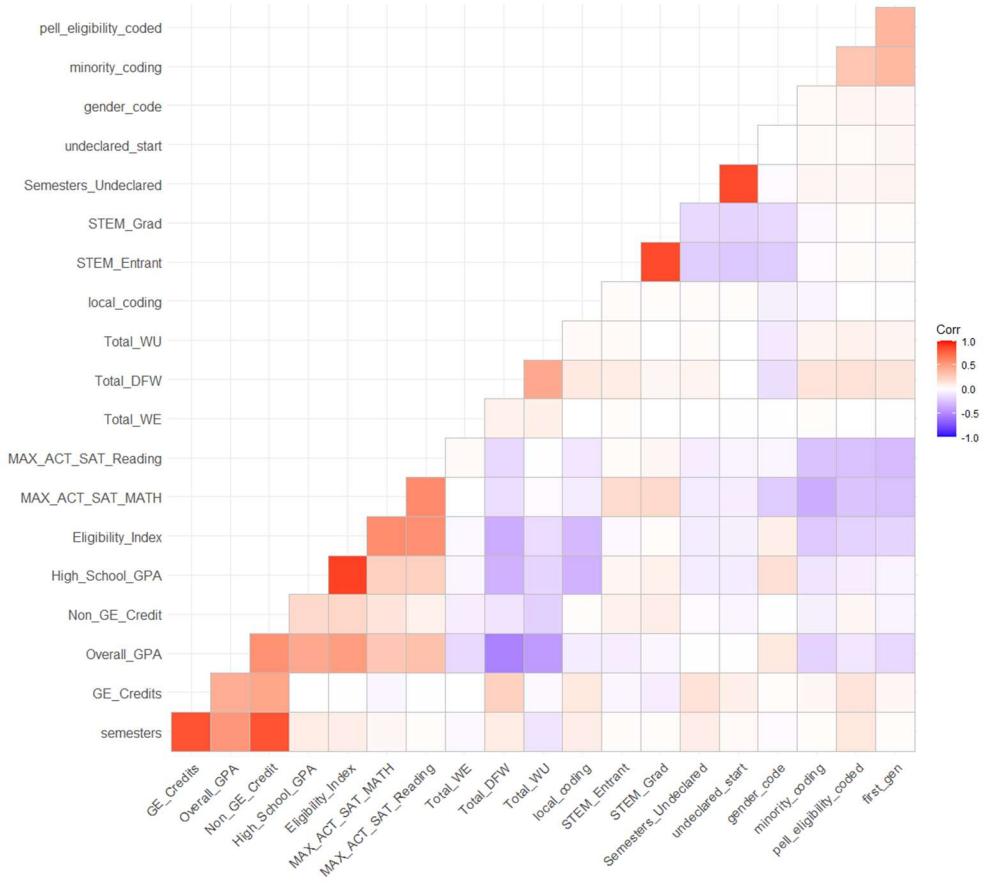
Local Admission	Semesters Enrolled	0.09
Number of Undeclared Semesters	Semesters Enrolled	0.09
Non-GE Credits Taken	STEM Graduate	0.09
Eligibility Index Score	Semesters Enrolled	0.09
Total WE Courses	Total WU Courses	0.08
Student Gender	Eligibility Index Score	0.08
Undeclared Admission	GE Credits Taken	0.08
Non-GE Credits Taken	Reading Standardized Test Score	0.07
Pell Eligible Admission	Total WU Courses	0.07
Total WE Courses	Total DFW Courses	0.07
STEM Graduate	High School GPA	0.07
Non-GE Credits Taken	STEM Admission	0.07
Minority Admission	Total WU Courses	0.06
First Generation Student	Total WU Courses	0.06
Total DFW Courses	Number of Undeclared Semesters	0.06
First Generation Student	Number of Undeclared Semesters	0.06
Minority Admission	Number of Undeclared Semesters	0.05
Pell Eligible Admission	Student Gender	0.05
STEM Admission	High School GPA	0.05
Pell Eligible Admission	Number of Undeclared Semesters	0.04
Pell Eligible Admission	Non-GE Credits Taken	0.04
First Generation Student	Undeclared Admission	0.04
Total DFW Courses	STEM Graduate	0.04
First Generation Student	Student Gender	0.04
Reading Standardized Test Score	STEM Graduate	0.04
Minority Admission	GE Credits Taken	0.04
First Generation Student	GE Credits Taken	0.04
Math Standardized Test Score	Semesters Enrolled	0.04
Undeclared Admission	Semesters Enrolled	0.03
Minority Admission	Student Gender	0.03
Total WE Courses	Reading Standardized Test Score	0.03
Minority Admission	Undeclared Admission	0.03
Total WU Courses	STEM Admission	0.03
Local Admission	Total WU Courses	0.03
Local Admission	Number of Undeclared Semesters	0.02
STEM Admission	Semesters Enrolled	0.02
First Generation Student	STEM Admission	0.02
Pell Eligible Admission	Undeclared Admission	0.02
Total WU Courses	Number of Undeclared Semesters	0.02
Pell Eligible Admission	STEM Admission	0.02
Reading Standardized Test Score	STEM Admission	0.02
Local Admission	STEM Admission	0.02
First Generation Student	Semesters Enrolled	0.01
Minority Admission	Total WE Courses	0.01

GE Credits Taken	Student Gender	0.01
First Generation Student	STEM Graduate	0.01
Local Admission	Undeclared Admission	0.01
Local Admission	STEM Graduate	0.01
Reading Standardized Test Score	Semesters Enrolled	0.01
STEM Graduate	Semesters Enrolled	0.01
Local Admission	Non-GE Credits Taken	0.01
Eligibility Index Score	STEM Graduate	0.01
Total WE Courses	STEM Admission	0.01
Pell Eligible Admission	STEM Graduate	0.01
Minority Admission	Semesters Enrolled	0.01
Total DFW Courses	Undeclared Admission	0.00
Local Admission	Pell Eligible Admission	0.00
Total WU Courses	Undeclared Admission	0.00
GE Credits Taken	High School GPA	0.00
Total WE Courses	Number of Undeclared Semesters	0.00
Total WE Courses	GE Credits Taken	0.00
Local Admission	Total WE Courses	0.00
Undeclared Admission	Student Gender	0.00
Total WE Courses	STEM Graduate	0.00
Total WE Courses	Math Standardized Test Score	0.00
Pell Eligible Admission	Total WE Courses	0.00
Total WU Courses	STEM Graduate	0.00
Number of Undeclared Semesters	CSULB GPA	−0.01
First Generation Student	Total WE Courses	−0.01
Total WE Courses	Student Gender	−0.01
GE Credits Taken	Eligibility Index Score	−0.01
First Generation Student	Local Admission	−0.01
Undeclared Admission	CSULB GPA	−0.01
GE Credits Taken	Reading Standardized Test Score	−0.01
Total WE Courses	Undeclared Admission	−0.01
Total WU Courses	Reading Standardized Test Score	−0.01
Non-GE Credits Taken	Student Gender	−0.01
Total WU Courses	Math Standardized Test Score	−0.02
Student Gender	Semesters Enrolled	−0.02
Minority Admission	STEM Admission	−0.02
Number of Undeclared Semesters	Non-GE Credits Taken	−0.02
Number of Undeclared Semesters	Student Gender	−0.02
Total WE Courses	Eligibility Index Score	−0.03
Total WE Courses	Semesters Enrolled	−0.03
Minority Admission	STEM Graduate	−0.03
Total WU Courses	GE Credits Taken	−0.03
Eligibility Index Score	STEM Admission	−0.03
Student Gender	Reading Standardized Test Score	−0.04

Undeclared Admission	Non-GE Credits Taken	-0.04
CSULB GPA	STEM Graduate	-0.04
Total WE Courses	High School GPA	-0.04
GE Credits Taken	Math Standardized Test Score	-0.04
GE Credits Taken	STEM Admission	-0.04
Undeclared Admission	Reading Standardized Test Score	-0.05
Minority Admission	Local Admission	-0.05
First Generation Student	Non-GE Credits Taken	-0.05
First Generation Student	High School GPA	-0.05
Local Admission	Student Gender	-0.06
Minority Admission	Non-GE Credits Taken	-0.06
Undeclared Admission	Eligibility Index Score	-0.06
Number of Undeclared Semesters	Reading Standardized Test Score	-0.07
Undeclared Admission	Math Standardized Test Score	-0.07
CSULB GPA	STEM Admission	-0.07
Total WE Courses	Non-GE Credits Taken	-0.07
Pell Eligible Admission	High School GPA	-0.07
GE Credits Taken	STEM Graduate	-0.07
Local Admission	CSULB GPA	-0.08
Number of Undeclared Semesters	Eligibility Index Score	-0.08
Local Admission	Math Standardized Test Score	-0.08
Number of Undeclared Semesters	High School GPA	-0.08
Undeclared Admission	High School GPA	-0.08
Number of Undeclared Semesters	Math Standardized Test Score	-0.08
Total WU Courses	Student Gender	-0.09
Local Admission	Reading Standardized Test Score	-0.10
Pell Eligible Admission	CSULB GPA	-0.10
Minority Admission	High School GPA	-0.11
Total WU Courses	Semesters Enrolled	-0.11
Total DFW Courses	Non-GE Credits Taken	-0.11
Total DFW Courses	Student Gender	-0.13
Total DFW Courses	Math Standardized Test Score	-0.14
Total WU Courses	Eligibility Index Score	-0.15
Total WE Courses	CSULB GPA	-0.16
Number of Undeclared Semesters	STEM Graduate	-0.16
Total DFW Courses	Reading Standardized Test Score	-0.17
Student Gender	STEM Graduate	-0.17
First Generation Student	CSULB GPA	-0.17
Total WU Courses	High School GPA	-0.18
First Generation Student	Eligibility Index Score	-0.18
Undeclared Admission	STEM Graduate	-0.18
Pell Eligible Admission	Eligibility Index Score	-0.19
Minority Admission	CSULB GPA	-0.19
Total WU Courses	Non-GE Credits Taken	-0.20

Number of Undeclared Semesters	STEM Admission	−0.21
Student Gender	STEM Admission	−0.22
Student Gender	Math Standardized Test Score	−0.22
Minority Admission	Eligibility Index Score	−0.23
Undeclared Admission	STEM Admission	−0.24
Pell Eligible Admission	Math Standardized Test Score	−0.25
Minority Admission	Reading Standardized Test Score	−0.26
First Generation Student	Math Standardized Test Score	−0.27
Pell Eligible Admission	Reading Standardized Test Score	−0.27
First Generation Student	Reading Standardized Test Score	−0.29
Local Admission	Eligibility Index Score	−0.30
Local Admission	High School GPA	−0.33
Total DFW Courses	High School GPA	−0.34
Total DFW Courses	Eligibility Index Score	−0.35
Minority Admission	Math Standardized Test Score	−0.35
Total WU Courses	CSULB GPA	−0.43
Total DFW Courses	CSULB GPA	−0.54

HEATMAP OF CORRELATION VALUES



APPENDIX G

UNIQUENESS TABLE AND RESIDUAL MATRIX

Uniqueness Table $m = 5$

Variable Name	Uniqueness
High School GPA	0.01
Overall CSULB GPA	0.03
Math Standardized Test Score	0.01
Reading Standardized Test Score	0.14
Eligibility Index Score	0.02
GE or Non-Major Credits	0.29
Major Credits	0.57
Total Number of DFW Courses	0.23
Total Number of WU Courses	0.71
Total Number of WE Courses	0.97

Factor Analysis Residual Matrix $m = 5$

	High School GPA	Overall CSULB GPA	Math Standardized Test Score	Reading Standardized Test Score	Eligibility Index Score	GE or Non-Major Credits	Major Credits	Total Number of DFW Courses	Total Number of WU Courses	Total Number of WE Courses
High School GPA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Overall CSULB GPA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.01
Math Standardized Test Score	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Reading Standardized Test Score	0.00	0.00	0.00	0.00	0.00	-0.01	0.01	0.00	0.00	0.01
Eligibility Index Score	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GE or Non-Major Credits	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	-0.01	0.05
Major Credits	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	-0.01	0.01
Total Number of DFW Courses	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	-0.04
Total Number of WU Courses	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01	0.01	0.00	-0.01
Total Number of WE Courses	0.00	-0.01	0.00	0.01	0.00	0.05	0.01	-0.04	-0.01	0.00

APPENDIX H

FACTOR LOADINGS

FACTOR LOADINGS

Variable Name	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
High School GPA	0.98	-0.19	-	-	-
Overall CSULB GPA	0.26	-0.70	0.58	0.14	0.24
Math Standardized Test Score	0.20	-	-	0.92	0.32
Reading Standardized Test Score	0.20	-	-	0.31	0.85
Eligibility Index Score	0.86	-0.16	-	0.31	0.35
GE or Non-Major Credits	-	-	0.84	-	-
Major Credits	0.12	-0.25	0.58	-	-
Total Number of DFW Courses	-0.19	0.82	0.21	-	-0.13
Total Number of WU Courses	-	0.53	-	-	-
Total Number of WE Courses	-	0.16	-	-	-

APPENDIX I

OUTPUT FROM THE MULTINOMIAL LOGISITIC MODEL

Parameter	STEM Admission	Student Gender	First Generation Student	Factor 1 - Academic Preparation	Factor 2 - CSULB Academic Information	Factor 3 - Credits Taken	Factor 5 - Reading Test Score
P-Value	$p < 0.01$	$p < 0.01$	$p < 0.01$	$p < 0.01$	$p < 0.01$	$p < 0.01$	$p < 0.01$
RR: One Major Change	0.00	1.02	1.74	0.75	1.10	1.97	0.73
RR: More Than One Major Change	0.32	1.37	1.14	0.82	1.46	1.06	0.73
Coeff. Estimate: One Major Change	-0.55	-12.97	0.02	0.55	-0.29	0.09	0.68
Coeff. Estimate: More Than One	$\mathbf{1.95}$	$\mathbf{-1.15}$	$\mathbf{0.32}$	$\mathbf{0.13}$	$\mathbf{-0.19}$	$\mathbf{0.38}$	$\mathbf{0.06}$

Parameter	STEM Admission	Student Gender	Pell Eligibility	Local Admission	Minority Admission	Factor 1 - Academic Preparation
P-Value	$p < 0.01$	$p < 0.01$	$p < 0.01$	$p < 0.01$	$p < 0.01$	$p < 0.01$
RR: Did Not Graduate	1.1	1.18	1.14	0.84	0.62	0.47
RR: Six-Year Graduate	1.44	0.70	1.36	1.24	0.99	0.80
Coeff. Estimate: Did Not Graduate	0.10	0.10	0.10	0.10	0.10	0.10
Coeff. Estimate: Six-Year Graduate	**0.37**	**0.37**	**0.37**	**0.37**	**0.37**	**0.37**

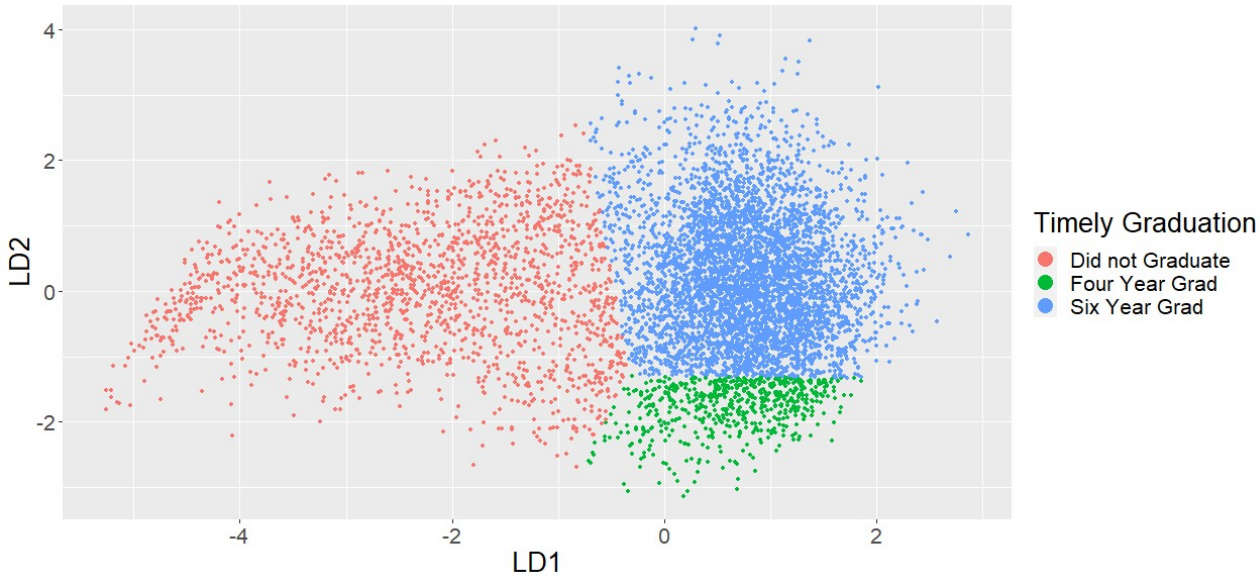
Parameter	Factor 2 - CSULB Academic Information	Factor 3 - Credits Taken	Factor 4 – Math Standardized Test Score	Factor 5 - Reading Standardized Test Score	Number of Major Changes
P-Value	$p < 0.01$	$p < 0.01$	$p < 0.01$	$p < 0.01$	$p < 0.01$
RR: Did Not Graduate	2.20	2.08	0.83	0.56	1.12
RR: Six-Year Graduate	2.20	2.08	0.83	0.56	1.12
Coeff. Estimate: Did Not Graduate	3.03	3.03	3.03	3.03	3.03
Coeff. Estimate: Six-Year Graduate	0.79	0.79	0.79	0.79	0.79

APPENDIX J

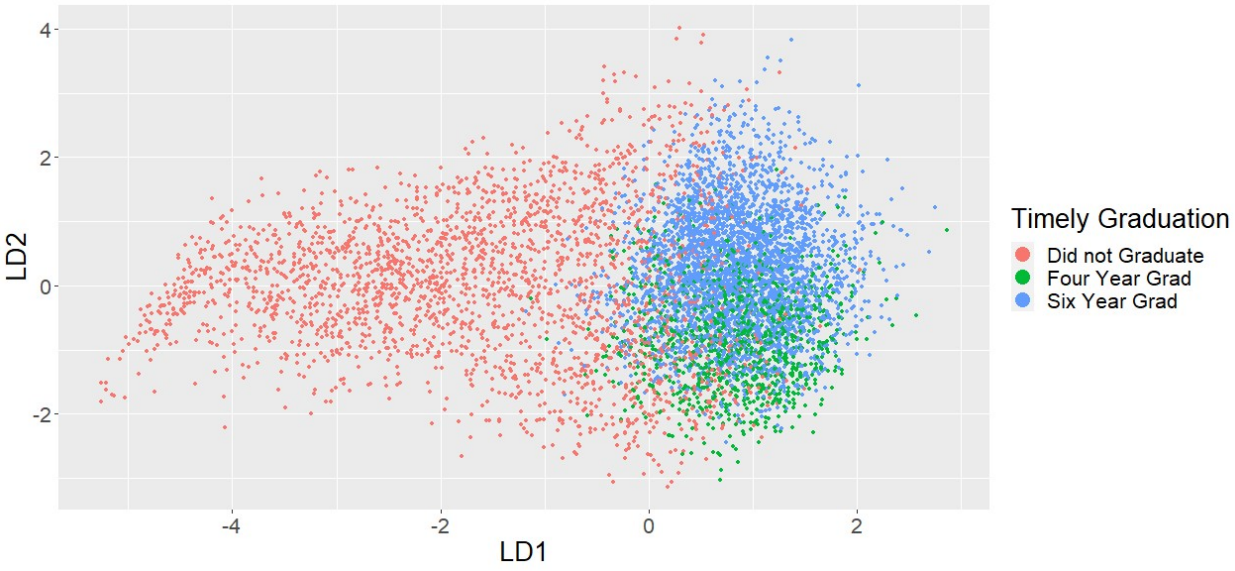
TABLES AND FIGURES RELATED TO FISHERS LDA

COEFFECIENTS OF LINEAR DISCRIMINANTS

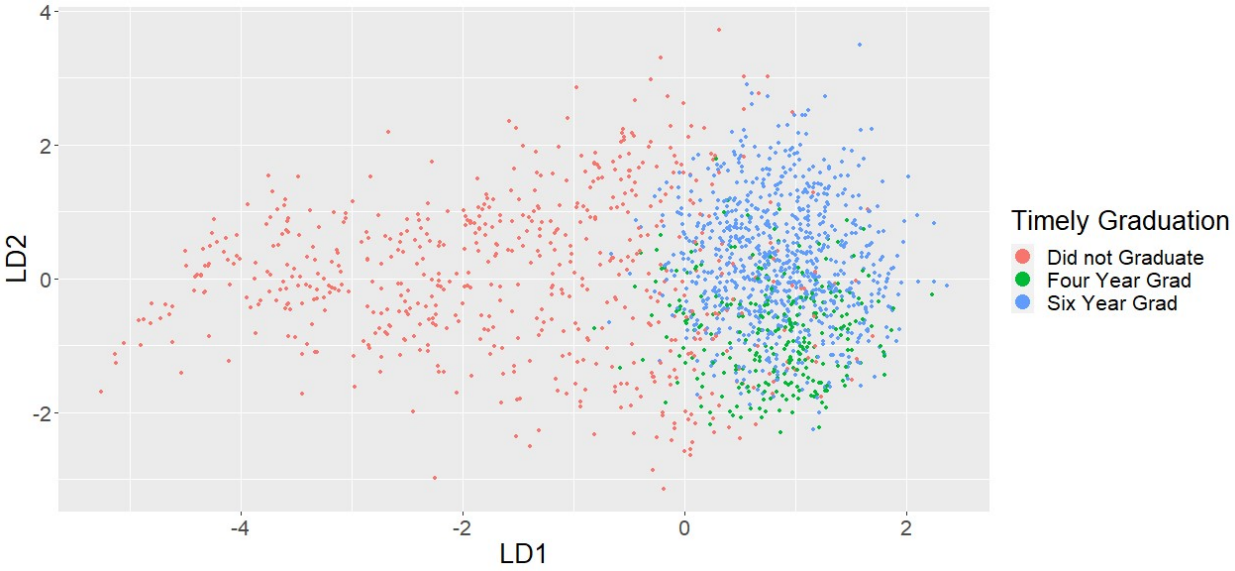
Variable Name	Coefficient LD1	Coefficient LD2
STEM Admission	0.0533	0.4355
Student Gender	−0.1157	−0.2910
Undeclared Admission	0.1355	0.5943
Pell Eligibility	0.0662	0.2684
Local Admission	0.1112	0.2374
Minority Admission	0.1722	−0.0856
First Generation Student	0.0130	0.0835
Number of Major Changes	0.2955	0.2159
Factor 1 - Academic Preparation	0.2198	−0.2859
Factor 2 - CSULB Academic Information	−0.7131	0.5219
Factor 3 - Credits Taken	1.0944	0.2711
Factor 4 - Math Standardized Test Score	0.0874	−0.2126
Factor 5 - Reading Standardized Test Score	0.1205	−0.5139



Decision Boundaries

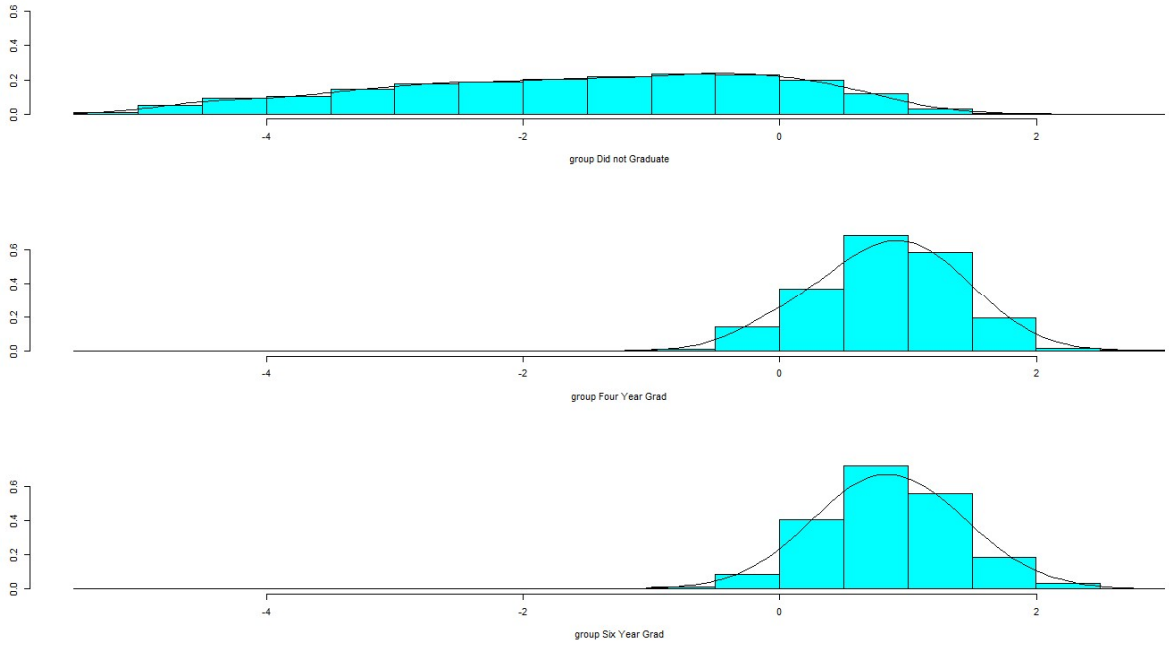


Classification of the Original Dataset

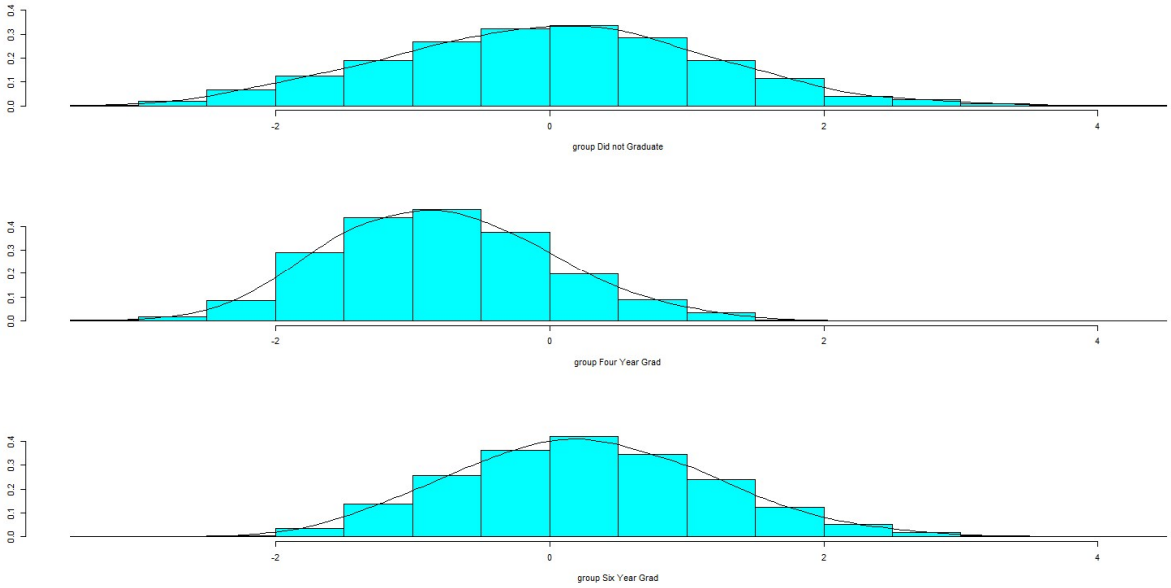


Classification of the Testing Dataset

Distribution of LD1 Values Within Each Population



Distribution of LD2 Values Within Each Population



REFERENCES

REFERENCES

California State University. n.d. “Diversity.” Accessed September 2020. https://www2.calstate .edu/impact-of-the-csu/diversity

California State University. 2015. “Graduation Initiative 2025.” Accessed August 2020. https://www2.calstate.edu/csu-system/why-the-csu-matters/graduation-initiative-2025

Cameron, A. Colin, and Frank A. G. Windmeijer. "R-Squared Measures for Count Data Regression Models with Applications to Health-Care Utilization." *Journal of Business & Economic Statistics* 14, no. 2 (1996): 209-20. https//doi:10.2307/1392433.

CollegeBoard SAT. n.d. “Compare SAT Specifications.” Accessed October 2020. https://collegereadiness.collegeboard.org/sat/inside-the-test/compare-old-new-specifications.

CollegeBoard. n.d.a. “Concordance.” Accessed October 2020. https://collegereadiness. collegeboard.org/educators/higher-ed/scoring/concordance.

CSULB Admissions. n.d.a. “Eligibility Index.” Accessed October 2020. https://www.csulb. edu/admissions/freshmen-eligibility-index.

CSULB Admissions. n.d.b. “Local Preference Admission Consideration.” Accessed October 2020. http://www.csulb.edu/admissions/local-preference-admission-consideration.

CSULB Student Records. n.d. “Understanding Grades and Grading.” Accessed October 2020. http://www.csulb.edu/student-records/understanding-grades-and-grading

Czepiel, Scott A. n.d. “Maximum Likelihood Estimation of Logistic Regression Models.” Accessed July 2020. https://czep.net/stat/mlelr.pdf

Education Advisory Board. 2016. “How Late Is Too Late? Myths and Facts About the Consequences of Switching College Majors.” Accessed October 2020. https://eab.com /technology/whitepaper/student-success/how-late-is-too-late/.

Federal Student Aid. n.d. “Understanding Aid - Federal Pell Grants.” Accessed October 2020. https://studentaid.gov/understand-aid/types/grants/pell#how-apply.

Foraker, Matthew. 2012. “Does Changing Majors Really Affect the Time to Graduate? The Impact of Changing Majors on Student Retention, Graduation, and Time to Graduate.” Accessed October 2020. https://www.wku.edu/instres/documents/air_major_change.pdf.

Gilbert, Ethel. 1968. “On Discrimination Using Qualitative Variables”. *Journal of the American Statistical Association* (63): 1399-1412

Grimm, Laurence G., and Paul R. Yarnold., Eds 1995. *Reading and Understanding Multivariate Statistics*. Washington DC: American Psychological Association.

Gordon, Larry. 2020. "In Historic Action, UC Moves to Drop SAT/ACT and Develop a Replacement Exam for Admissions." Accessed May 2020. https://edsource.org/2020/in-historic-action-uc-moves-to-drop-sat-act-and-develop-a-replacement-exam-for-admissions/632174.

Gower, J.C. 1971. "A General Coefficient of Similarity and Some of Its Properties." *Biometrics*, 27(4): 857-871.

Guttman, L. 1954. "A New Approach to Factor Analysis: the Radex." In *Mathematical Thinking in the Social Sciences*, edited by Paul F. Lazarsfeld, 258-348. New York: Free Press.

Humphreys, Lloyd G. 1964. "Number of Cases and Number of Factors: An Example Where N is Very Large." *Educational and Psychological Measurement* 24 (3): 457-466. https://doi.org/10.1177/001316446402400302

James, Gareth, Daniela Witten, and Trevor Hastie. 2017. *An Introduction to Statistical Learning with Applications in R*. New York: Springer.

Johnson, Richard A., and Dean W. Wichern. 2007. *Applied Multivariate Statistical Analysis*. New Jersey: Pearson Prentice Hall.

Kaiser, Henry. F. 1960. "The Application of Electronic Computers to Factor Analysis." *Educational and Psychological Measurement* 20(1): 141–151. https://doi.org/10.1177/001316446002000116

------. 1958. "The Varimax Criterion for Analytic Rotation in Factor Analysis." *Psychometrika* 23: 187–200. https://doi.org/10.1007/BF02289233

Kaufman, Leonard, and Peter J. Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

Kim, Sung. 2019a. "Discriminate Analysis." Class Lecture. California State University, Long Beach, Long Beach, CA.

------. 2019b. "Factor Analysis." Class Lecture. California State University Long Beach, Long Beach, CA.

Korosteleva, Olga. 2018. *Advanced Regression Models with SAS and R Applications*. 1st ed. London: Chapman and Hall. ISBN: 11380490

Marr, Jackson., and Robert M. Hume. 1996. "Using Discriminant Analysis to Identify Students at Risk." Accessed September 2020. https://ieeexplore.ieee.org/document/569940

McDonald, John H. 2014. *Handbook of Biological Statistics.* 3rd ed. Baltimore: Sparky House Publishing. http://www.biostathandbook.com/multiplecomparisons.html#bonferroni

Mendenhall, William and Terry Sincich. 2012. *A Second Course in Statistics: Regression Analysis*. Boston: Prentice Hall.

Mote, Thomas A. 1970. An Artifact of the Rotation of too Few Factors: Study Orientation vs. Trait Anxiety. *Interamerican Journal of Psychology* 4: 3-4.

Micceri, Ted. 2001. "Change Your Major and Double Your Graduation Chances." Paper presented at the AIR Annual Forum, Long Beach, CA, June 3-6, 2001. Accessed October 2020. https://eric.ed.gov/?id=ED453756.

National Center for Education Statistics. 2018. "First-Generation Students." Accessed August 2020. https://nces.ed.gov/pubs2018/2018421.pdf

------. 2016. "Graduation Rates." Accessed September 2020. https://nces.ed.gov/pubs2017/2017046.pdf.

O'Halloran, Sharyn. n.d. "Logistical Regression II – Multinomial Data". Class Lecture. Columbia University, New York City, New York. http://www.columbia.edu/~so33/SusDev/Lecture_10.pdf

Park, H. Kang, and Peter M. Kerr 1990. "Determinants of Academic Performance: A Multinomial Logit Approach" *The Journal of Economic Education* 21(2): 101-111.

Queen Mary University of London. n.d. "Hypothesis Testing." Accessed September 2020. http://www.maths.qmul.ac.uk/~bb/MS_Lectures_23and24.pdf

Scholorships.com. n.d. "Federal Pell Grants." Accessed October 2020. https://www.scholarships.com/financial-aid/federal-aid/federal-pell-grants/.

Tanbakuchi, Anthony. 2009. "Test of Independence and Homogeneity." Class Lecture. Arizona State University. Accessed July, 7 2020. http://www.u.arizona.edu/~kuchi/Courses/MAT167/Files/LH_LEC.0640.HypTest.IndepHomog.pdf

Tripp, A. and Duffey, M. 1981. "Discriminant Analysis to Predict Graduation – Nongraduation in a Master's Degree Program in Nursing." *Research in Nursing Health* 4: 345-353. https//doi:10.1002/nur.4770040403

Thurstone, L. L. 1935. *The Vectors of Mind: Multiple-Factor Analysis for the Isolation of Primary Traits*. Chicago: University of Chicago Press. https://doi.org/10.1037/10018-000

United States Department of Education. 2017. "Beginning College Students Who Change Their Majors Within 3 Years of Enrollment." Accessed October 2020. https://nces.ed.gov/pubs2018/2018434.pdf.

White, Daniel, and Ronald Huesman. 2010. “Redefining Student Success: Applying Different Multinomial Regression Techniques for the Study of Student Graduation Across Institutions of Higher Education” *Research in Higher Education* 51: 154–174

Zeidenberg, Matthew, and Mark Scott. 2011. “The Content of Their Coursework: Understanding Course-Taking Patterns at Community Colleges by Clustering Student Transcripts.” Community College Research Center, Columbia University. Accessed October 2020. https://ccrc.tc.columbia.edu/publications/course-taking-patterns-clustering.html.

Zyso, Peter V. 1998. “The Modification of the Phi Coefficient Reducing its Dependence on the Marginal Distributions.” Accessed November 2020. https://www.dgps.de/fachgruppen/methoden/mpr-online/issue2/art5/zysno.pdf.