Quantifying Student Success

Santiago Canyon College

Cesar Pardede, Ezra Mock, Marc Sanchez, and Michael Smith

May 15, 2021

CALIFORNIA STATE UNIVERSITY

FULLERTON

Contents

1 Abstract

Student success and non-success can be measured many different ways, whether it's by percentage of units completed, percentage of courses completed, continuing enrollment, or other definitions. In cases where success and non-success are determined by a percentage of completion, the threshold percentage for completion adds another aspect of complexity to defining success. We aim to discover three main things: whether units completed or courses completed provides a better metric for measuring success and non-success, which variables are the top predictors of success and non-success, and whether a student living in a low-income ZIP code has the same predictive value on success and non-success as a student's self-reported low-income status. Additionally, we want to examine the effect of the COVID-19 pandemic on each of these goals, and whether the COVID-19 pandemic affected our results.

We found units completed provide a better metric for success and non-success, particularly at the 80% threshold; a 100% threshold is too restrictive and likely to flag students who would otherwise be considered successful as a case of non-success - this is also true during COVID terms. The top predictors for success and non-success come down to nine variables which repeatedly ranked among the most important variables using various variable importance methods and metrics - this is also true during COVID terms. A student's ZIP code provides more predictive value during non-COVID terms; within COVID terms, a student's self-reported income status provides more predictive value than the ZIP code of their residence.

2 Introduction

The Office of Institutional Effectiveness and Assessment of Santiago Canyon College (OIE), our client, has partnered with a student data science team at CSUF, to seek answers to some research questions related to student success. The OIE supports the college by making data-informed policy recommendations to increase student success.Student success is of interest because when students are successful, then so is the institution. However, defining student success is not easy, there are many possible definitions and which one is used depends on the surrounding discussion that is taking place.

The team was provided with a multi-year data file with anonymized student data, that includes variables related to student characteristics and performance data. This data file was to be used to derive new student variables that measure other types of performance of interest.

The following definitions for student success were: 1) 100% of courses passed; 2) 100% of units passed; 3) 80% or more courses passed; 4) 80% or more units passed. All these are measured for each major term where a major term is either a Spring or Fall term.

A few major goals that were established early on were: 1) which of the definitions given above for student success is the better student success metric; 2) which variables are most important for determining student success; 3) whether low-income status or low income zip code are a better low income variable for determining student success.

In addition, the team was asked to explore these goals for non-success. And, finally to determine if we get different answers when we look at pre-COVID terms (before 2020) and COVID terms (Spring & Fall 2020).

2.1 Literature Review

The City of University of New York received funding from the Center of Economic Opportunity to develop a program to increase the number of associate degrees and increase transfer rates. The program was called Accelerated Study in Associate Programs (ASAP). The main goal for the ASAP program was to have students graduate within three years with a success rate of 50% or higher. The students selected for the program all had the goal of obtaining an associate's degree in programs closely related to future employment or to transferring to four year colleges (Linderman 2009). These chosen students were given incentives to help improve retention, performance, and graduation rates, which included tuition waivers for

financial aid-eligible students, monthly Metrocards, small class size, career development and textbooks. In addition, these students made use of winter and summer classes, and "mandatory and intrusive student support services, including advice, career development, and academic supports" (Linderman 2009). This program was very successful at increasing graduation rates. ASAP obtained a graduation rate of 55% across all cohorts, whereas the control group only had a graduation rate of 23 percent. These results provide an immense amount of insight for what predictors to consider.

In a follow up paper about ASAP (Linderman 2013), authors report how the piloted students were asked what was the biggest resource provided to them, and 70% stated that receiving the financial resource was the biggest contributor to their completion and success at their college (Linderman 2013). From our personal experiences as community college students or faculty, receiving financial aid was very impactful because it gave students the ability to not have to worry about taking on more work shifts than needed. This allowed us more time to be on campus, allowing us to study more and interact with professors and classmates.

In a recent journal (Fauria and Fuller 2015), the authors explain how previous researcher found a significant difference in students transferring to a four year college if the community college had an Articulation Agreement of Transfers (AAOT). This means the courses students take at a community college would transfer and be counted towards their baccalaureate. As reported by Fauria and Fuller, major predictors for transfer students in obtaining a baccalaureate are gender, courses taken in high school, successful math remediation, expectations prior to starting college, and socioeconomic status and that student interactions with one another was a major predictor in academic success. Fauria and Fuller (2019) noted that students who created study groups with each other would make a serious impact on their overall success at the college and the courses taken. Additionally, student faculty interaction was also a predictor for student success. When students felt comfortable to ask their professors for help, they were more likely to figure out things they were struggling with in the class.

In the article by Turk (2017), the author provides troubling realities that can make it harder for some students to be successful at a community college. He points out that females were more likely to graduate over males, and minority students succeeded at much lower rates than white students. Also, students that were first-generation or low income had the lowest rates of success (Turk 2017). This is of serious concern to all community colleges. Students should all start with an even "playing field", which means community colleges need to provide resources to better help struggling students. Santiago (2013) States: "While student enrollment at community colleges has increased over the last 10 years, degree completion has not grown as quickly for Latino students". She explained that this degree completion is due to several factors such as the cost of attending college, family responsibilities, limited college knowledge, and work (Santiago 2013). In the paper she explains that many students do not realize that there are a lot of resources out there like tax credits, food assistance, and public healthcare. This might be due to a large proportion of students being first generation college students and they may not know how the system works and or lack the guidance from family members.

Eagan and Jaeger (2008) report that although having more part-time faculty is beneficial to saving cost, "Findings suggest that students tend to be significantly less likely to transfer as their exposure to part-time faculty increases". They point out that this could be affecting students who come from a lower socioeconomic status, since community colleges come as a first choice for many students looking to have a more affordable college experience. On average, most part-time faculty have other full time jobs. This could mean that part-time faculty have less time in the day to give to their students, but a full-time teacher can give more quality time to their students.

In addition, how well students do in high school can be a big predictor of success. Belfield and Crosta (2012) discovered that "high school GPAs are useful for predicting many aspects of students college performance". A general "rule-of-thumb association is that a student's college GPA tends to be one grade notch below that student's high school GPA. If success is defined as a C average in college, we expect this would be attained

by all students with at least a C+ average in high school". This means that students who struggle in high school will need more support once they enter into a community or else it might be very difficult for them to succeed. Lastly, the University of Lamar Texas, published research that showed "parents education level has a significant impact on their children's success". In their study, they found that 80 percent of students, who were raised with parents with college degrees, had encouragement to attend a 4 year college, whereas students with parents with no degree had only 29 percent. The article reports that this might be due to lacking an educational role model. This is another indication that we must find these students early and make sure they have enough support in order to be successful at their community college.

3 The Raw Data

3.1 Data Structure

This file contained 340,151 rows and 55 columns. Each row uniquely identified a student and course taken during a given semester spanning from Summer 2015 through Fall 2020 along with other historic and demographic data. The columns contain two basic categories: academic based and demographic information. The academic based data describes information about the students current workload. For example, the school, term, school year, course taken, units per course, total units taken, total courses taken, grade received, units passed, etc. The demographic data is of a different nature. Attributes falling into this category describe characteristics about the student themselves. Included here are characteristics such as age, gender, address, athlete, veteran, educational goals and other. In addition to the provided variables we will need to compute and transform additional variables. For example low_income status as defined in the FAFSA documents may not be complete or entirely accurate.

See partial sample below:

ID_Number	Status	Snapshot	School	Section_No	Course_No	Units	Grade	Tflag	Passed
278288230	N	20203ET	CC	85754	PSYC	100	3	A	3
278166790	N	20203ET	CC	85763	PSYC	100	3	A	3
278106400	N	20203ET	CC	85763	PSYC	100	3	A	3
277823810	N	20203ET	CC	85758	PSYC	100	3	A	3
277677290	N	20203ET	CC	85758	PSYC	100	3	A	3
240933550	N	20153ET	CC	1	CHEM	210	5	A	3
240823110	N	20153ET	CC	1	CHEM	210	5	A	3
238987760	N	20153ET	CC	1	CHEM	210	5	A	3
238241520	N	20153ET	CC	1	CHEM	210	5	A	3
229677580	N	20153ET	CC	1	CHEM	210	5	A	3
250485950	N	20191ET	CC	61127	MATH	N98	0.5	P	3
266470490	D	20191ET	CC	61184	MATH	N98	0.5	W	3
266064590	N	20191ET	CC	64423	MATH	N98	0.5	P	3
265903660	D	20191ET	CC	61184	MATH	N98	0.5	W	3
265433080	D	20191ET	CC	61184	MATH	N98	0.5	W	3

As mentioned previously, the raw excel file contained 340,151 rows and 55 columns. The rows identified all courses taken at Santiago Canyon College over the previous 5+ years. The covariates represented in the columns are both quantitative and categorical. Some were reported by the student, some were reported by the college, some were derived by the SCC team, and some were ultimately derived by us.

The quantitative variables come in a few different variations. Here are some examples: **Continuous** - Semester GPA, Cumulative GPA, Median income, longitude and latitute. **Discrete** - Age, Semester Units attempted, and Cumulative Units completed. **Binary** - First Generation College Student, Athlete, and Veteran.

In addition to the quantitative variables, we have many qualitative variables. Some examples of these are here: **Grade** - grades are given categorically, but we also have this in binary form as a course Passed flag. **Gender** - This vairable is less straightforward than you might expect and is subject to change. **Ethnicity** - Ethnicities are broken up into 8 major categories. **Term** - Defined as Fall, Winter, Spring, Summer, this variable will not be used in our final model but it is needed to derive other attributes.

Something interesting about the structure of this dataset is that the response variables *were not explicitly provided*. We were tasked with deriving these target variables based on information that was provided in the dataset.

3.2 Data Wrangling

The purpose of data wrangling i.e. removing outliers and noninformative information is to prepare our data for statistical analysis. Our first task was to remove a single blank row and a few covariates that were not giving us any useful information. For example, one of our columns was called "School" and every single student had the same school of SCC. When talking to our client, they explained that their district has two schools, Santiago Canyon College and Santa Ana College, but they only pulled SCC data which is why every student had the same value of SCC. The other variables we eliminated were: Location, Section Number, and Enroll Status. We then filtered out all students that were a part of the Apprenticeship program because these students weren't representative of the larger student population. We had justifiable reasons for eliminating each of these variables and the client agreed that we should remove them for better modeling.

In addition to removing unnecessary covariates, we needed to align the data in a usable form. The clients at SCC gave the variables: First Generation, Low Income (b), Low Income (All), Day Classes Only, Full Time Only, Athlete Current, Veteran. These covariates are binary and are composed of either the flagged values (Veteran, Fulltime, etc.) or NAs. We replaced the NAs with the corresponding complements (e.g. Veteran's complement was Non-Veteran) to better clean the data. In addition, our data set had some extreme outliers that were not representative of a typical student at SCC. For example, the distribution of semester units completed had a five-number summary as: minimum equals 0, first quantile equals 3, median equals 7, third quantile equals 12, and a max of 71.8 units. The 71.8 might be a typo or an extremely gifted student, but it is clearly an outlier in our dataset. Since this person is larger than the 99th quantile (20), we decided that this person will create too much "noise" in our data and will not help us understand the majority of our student population at SCC. Similarly, when looking at the distribution of cumulative semester units, there exists a student with 600 units. This would roughly equate to a student being at SCC for 25 years if they took 24 units a year on average (full-time status). We also found the 99th quantile, 105 and decided to filter out anyone that had more than 105 cumulative units. After performing these data manipulations, our data is now more representative of a typical student at SCC and this will help us build a more accurate model.

The next steps in our data wrangling adventure were to perform data transformations. Below is a list of all the data transformation we performed on this dataset. Here is a list of all our transformations: Total Number of Classes by Term, Total Number of Classes Dropped, Completed Courses Per Term, Athlete Ever, Proportion of Successful Units by Term, Math and English, Taken within First Year, Returning after Gap Year, Persistent Fall and Spring, Proportion of Units Passed, and Low Income Variable by Zip Code and median income. After all the data was cleaned and transformed, we needed to compress a student's information into one record by term. For example, here is what a student's record might look like:

ID_Number	Term	Course Success	Age	Course	Grade	GPA	$\cdots$	Total Classes
123455	Fall 2015	1	21	PSYC-100	A	3.70	$\cdots$	3
123455	Fall 2015	1	21	MATH-170	A	3.70	$\cdots$	3
123455	Fall 2015	1	21	ENG-101	B	3.70	$\cdots$	3

And this would be condensed into:

ID_Number	Term	Course Success	Age	GPA	⋯	Total Classes
123455	Fall 2015	1	21	3.70	⋯	3

We did this for each term, so every student was only appearing once per term. This gave each student per term equal weight. Next, we will show the power of our data transformations. We were able to look at relationships between our response variables and different covariates. Our first plot compares and contrasts the 80% or More Units Completed versus the 80% or More Courses Completed thresholds. Each row represents a year and the teal graph denotes the units threshold and the purple the courses threshold. In Spring 2018 the percentage of students that obtained the units threshold is 56%, which is larger than the 52.1% for the courses threshold in Spring 2018. When we compare each term, this pattern ceaselessly remains constant, that is, the units threshold continues to have a higher percentage of success.

To confirm that the units threshold is a better metric for student success, we ran a logistic model using contingency tables. Our output informed us that when going from courses to the units threshold, your odds for success are 22% higher succeeding. This means that students are more likely to be classified as successful when they are in the units threshold. Our test statistic was 20.81, which means we had an extremely small p-value. This informs one further how much of a difference there is among the two metrics. Since the units threshold performs better and the computation will be identical for courses, we will only be focusing on the units threshold from here on out.

Furthermore, from Figure 1, in the year 2020 (Covid terms) there is a higher level of success. For example, in Spring 2018 and Spring 2019 the success rate for obtaining the 80% threshold was 56% and 57%, but in Spring 2020 the percentage of success is 69%. Similarly, we ran a logistic regression model using contingency tables to see if students were more likely to be classified as successful in Covid terms (Fall and Spring 2020) versus Non-Covid terms (Spring 2016 - Fall 2019). Our output informed us that when going from Non-Covid to Covid terms, students' odds for success are 68% higher for succeeding. This output gave a test statistic of 26.87.

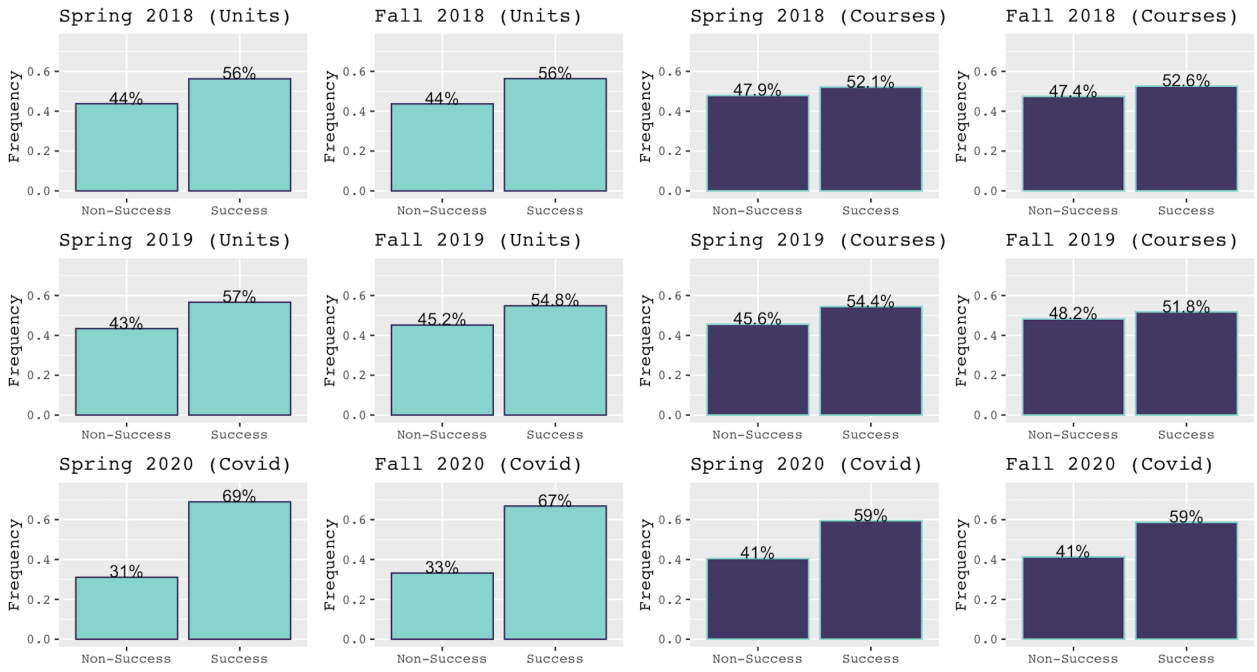

Figure 1: Academic Status (Completed 80 Percent or More Courses) in Relation to Gender

One of the more challenging data transformations we performed was to determine if a student took Math and English within their first year at SCC. Fortunately this worked paid off as the output showed a very telling story: more students obtained the 80% threshold when they took Math and English within their first year at SCC. For example, in Spring 2018 70% of students obtained the 80% threshold when they took math and English within their first versus when they did not, their success rate was only 49.74%. Looking term by term, one can see this pattern is constant.

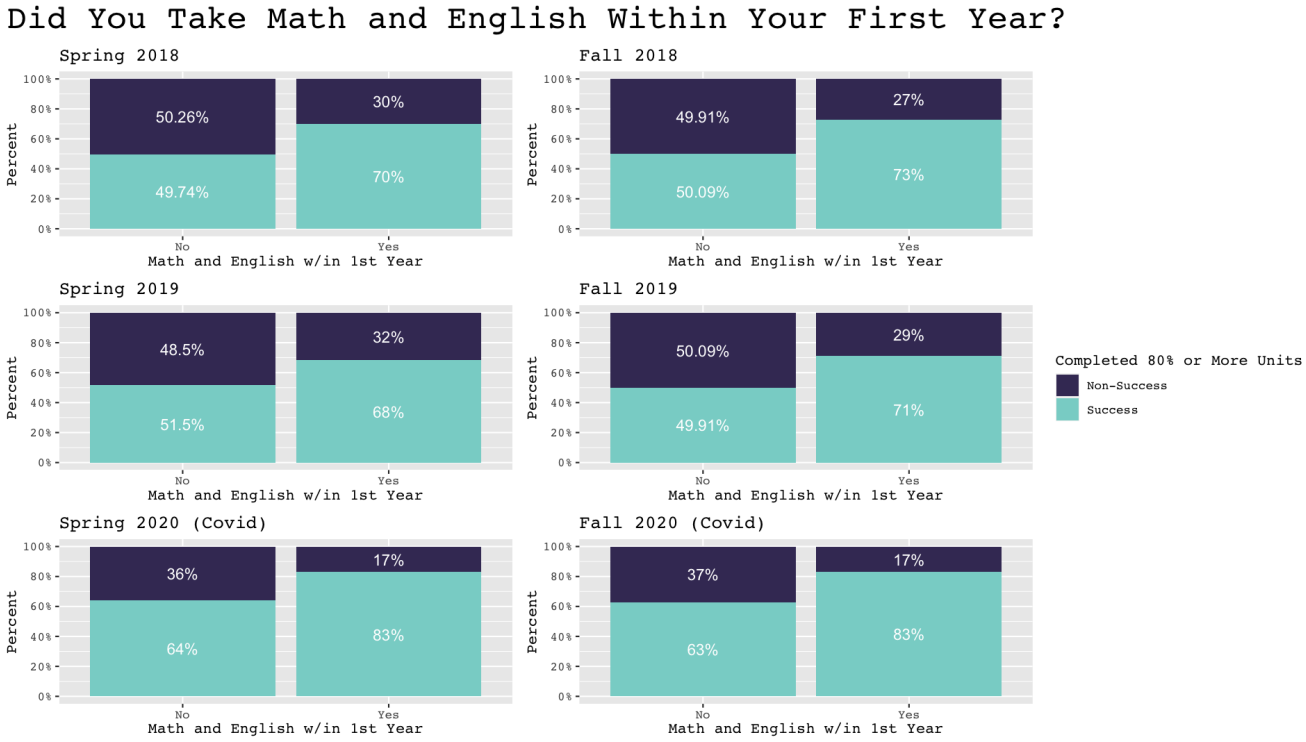

Figure 2: Academic Status (Completed 80 Percent or More Courses) in Relation to Taking Math and English Within First Year. The Yes Group Represents That a Student Took Both Math and English Within Their First Year At SCC

To support our third objective of investigating whether a student's address provides the same predictive value as their self-reported low-income status, we created a low-income binary variable based on a student's zip code. Using data from the American Community survey (ACS), (California 2021) we found the median incomes of California zip codes in the SCC data set and plotted a histogram of them in Figure 3. We defined low-income using the US Department of Housing and Urban Development's definition: not more than 80% of the state median. For our dataset, this equated to any income less than $60,000 being classified as low-income and any income above $60,000 being classified as not low-income.

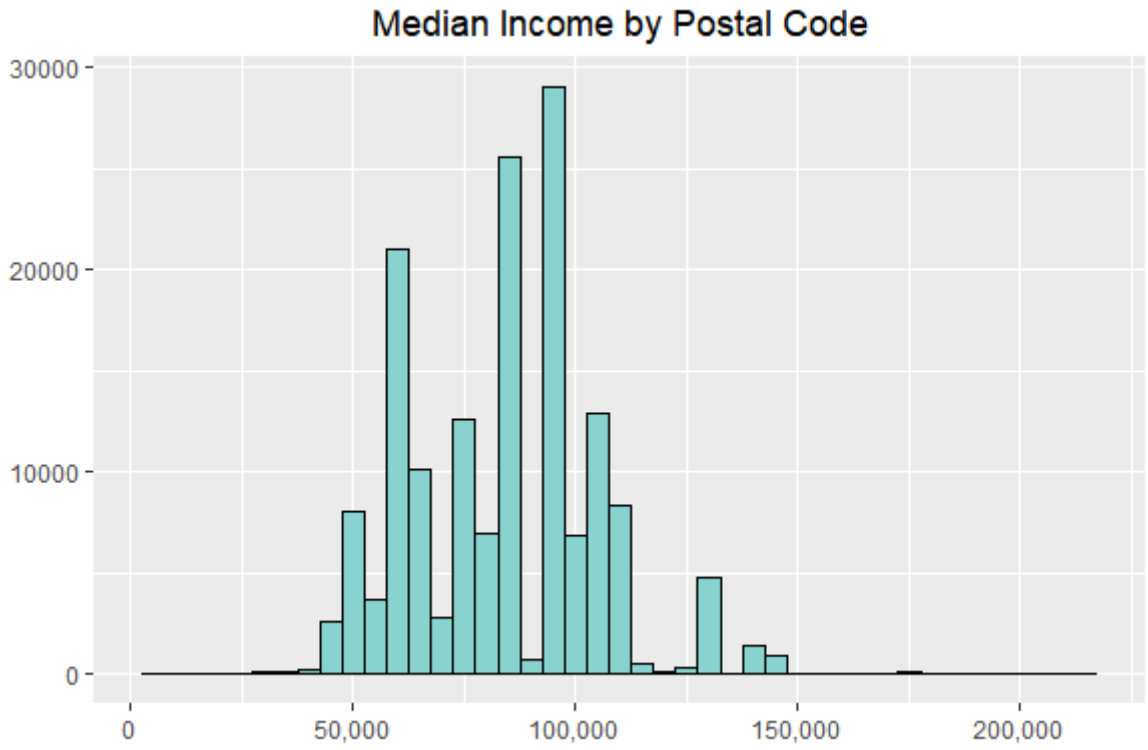

Figure 3: Histogram of median incomes for zip codes in the SCC data set.

If the median income of a student's zip code was below the threshold for low-income as defined by the US Department of Housing and Urban Development, then we flagged them as low-income in our variable. Using this transformation, we were able to determine which low-income variable, low-income by location or low-income by status (FAFSA), was better for modeling our data. The plot below shows the similarities and differences between the two variables.

Looking at Income by Zip Code graph below, on the right hand side of the 57 freeway, there is a high volume of not low-income locations (North Tustin, Yorba Linda, etc.). However, between the left-hand side of the 57 freeway and the right hand-side of the 405 freeway there are a lot more low-income locations (Santa Ana, Garden Grove, etc), and by the coast (Huntington Beach, Newport Beach). When looking at the graph on the right (Low Income Status), we can see a similarity of more not low-income on the right hand side of the 57 and more low-income on the left-hand side of the 57 freeway and on the right of the 405 freeway. However, the distinction is not as clear-cut. In this plot (Low Income Status), we can see that there is more blending happening. In the not low-income areas, there are more students that classified themselves as with low-income status and similarly some students in the low-income areas, classified themselves as not low-income status.

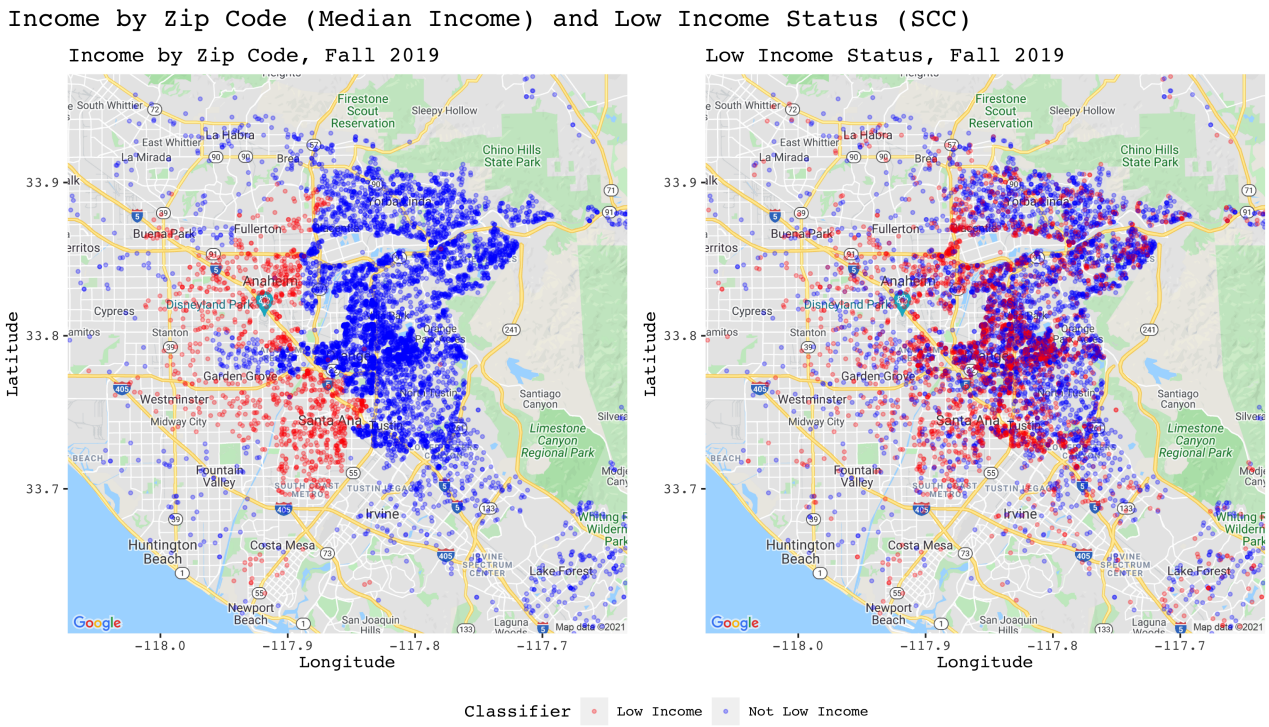

Figure 4: Low-Income by Zip Code (Left) versus Low-Income Location (Right) for 30 miles Radius Around SCC

To determine which variable has better predictability, we ran a simple logistic regression model with both predictors (low-income location and low-income status) with our response variable (completed 80% or more units). We ran our model with all terms minus Covid terms and then we ran the model with only Covid terms. In the Non-Covid terms, we found that the low-income by zip code had better predictability. On the contrary, for Covid terms only, low-income status had gave better predictability.

4 Exploratory Data Analysis

4.1 Data Summarization and Visualization

To begin our exploratory data analysis, we looked at the distribution of grades before and during the COVID-19 pandemic (see Figure 5) and found noticeable differences. Understandably, the number of emergency withdrawals (EW) increased in 2020, but so did the number of A's awarded. The number of B's, C's, D's, F's, and W's all decreased in 2020, driving up the average GPA. Overall, it appears that maybe some of the B's, C's, D's, and F's were upgraded, and the W's were subsumed into the increased number of EW's. It's not far-fetched to believe that during this period, instructors were more forgiving with their grading to accommodate adapting to a new learning medium, and the general strain of a worldwide viral outbreak.

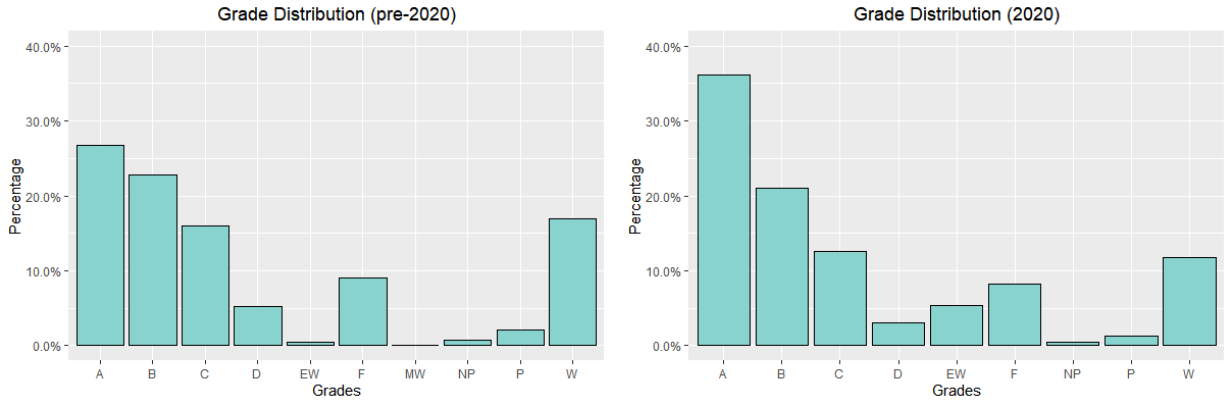

Figure 5: Distribution of grades pre-2020 and in 2020.

Next we'll discuss one of the response variables we are considering as a definition of success: the percentage of students per term who complete 80% of units they attempt. As a whole, for all students over all terms, 60% of students per term are passing 80% or more of their attempted units as shown in Figure 6. This gives us a baseline rate for which we expect students to succeed given other factors in the data set. Anything significantly higher or lower is a signal for us to further investigate that factor.

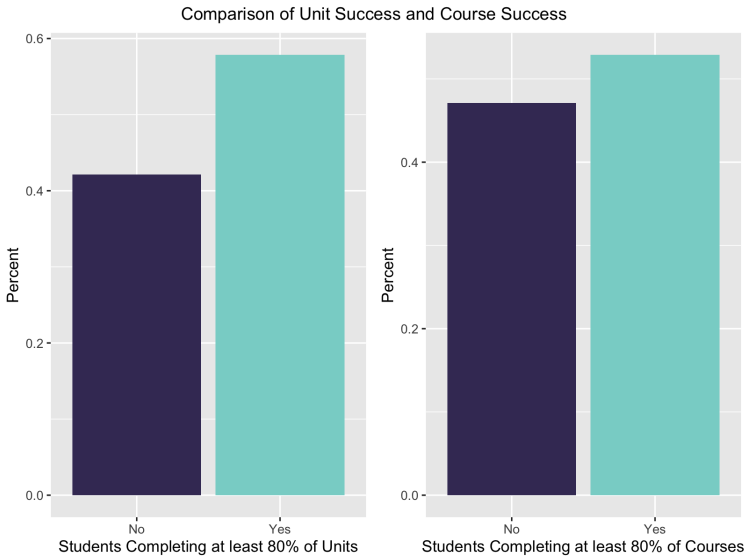

Figure 6: Percentage of students completing 80% of their attempted units.

Here we begin our bivariate EDA by plotting one of our response variables (our definitions of success) against various other factors in the SCC data set. Note we are only showing success as defined by completing 80% of classes. An example of a student characteristic that does not seem to predict success is whether a student takes classes only during the day, or otherwise (students that take at least one night class). Figure 7 contains 6 plots, one for each of the major terms between Spring 2018 and Fall 2020.

In each of these plots we are comparing the success of students taking only day classes against students taking at least one evening classes. With the exception of Fall 2020, all terms show that day only students are only slightly more successful (within 2 percentage points) than students that take at least one evening class. During Fall 2020 the difference in percentage points for success is greater than 2%. The data contains 5 other terms, major terms between Fall 2015 and Fall 2017, that are not showed during which success between the two groups stays within 2 percentage points, with the exception of Spring 2016. Since the percentage

point difference between success for students taking day only and the other group is small, it would not be surprising if the difference could be attributed to chance variation.

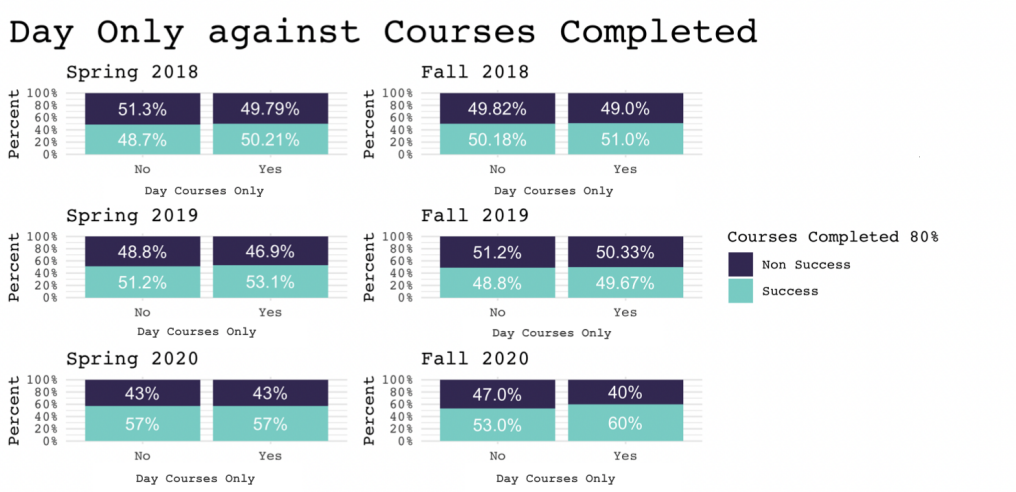

Figure 7: Whether a student takes classes exclusively during the day or not has no observable predictive value on whether they will meet our definitions of success at the 80% threholds.

The persistence variables, which track whether a student persisted from the previous Fall or Spring semester, show an interesting phenomenon. Figure 8 shows a larger proportion of students who did not persist from the previous Fall semester in success and non-success. This can be explained by the Fall semester being the traditional start of the school year. Since fewer students enroll in the Spring than in the Fall, the students who do persist contribute more to the number of students meeting our success criteria, and contribute about equally to the number of students who were met with non-success. This may be a valuable factor to consider and investigate further to inform a policy-maker when to spend more resources conducting interventions for at-risk students.

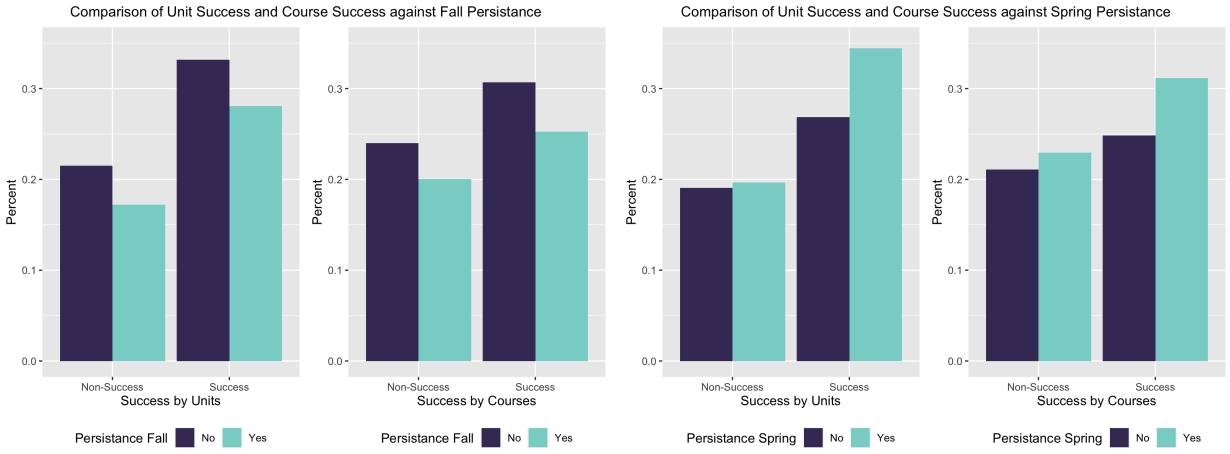

Figure 8: Students who persist from Spring to Spring are more successful than Students who persist from Fall to Fall.

We also generated plots that involved the success response variables against some of the numerical predictors. Figure 9 shows plots that involve the predictor age and the success response variables. Comparing the teal bars and salmon bars on this left plot, there are small differences and it could be that we just are seeing chance variation.The right plot is similar to the left, except that now success is measured as completing 80% or more of the courses in a term. The left and right plots are different, but they must be examined carefully to detect the differences. We could again argue that the differences, could be attributed to chance variation.

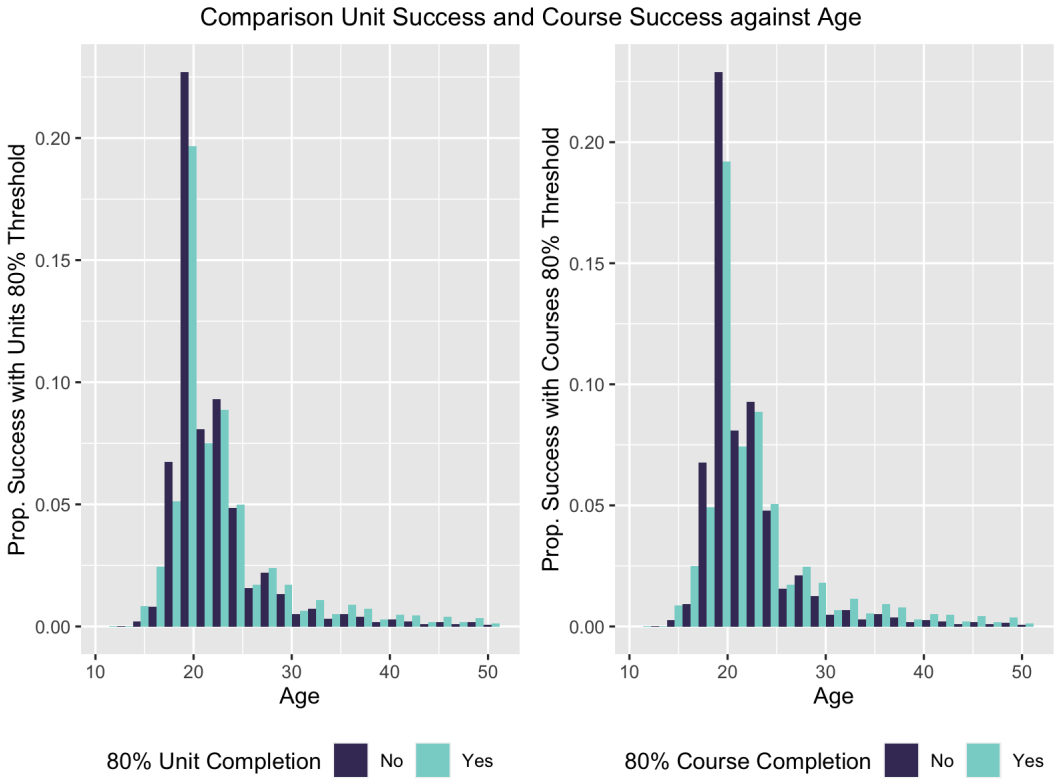

Figure 9: Histogram of Age against Success

Figure 10, shows plots where semester GPA is plotted against success. Both left and right graphs show that this predictor is doing a good job predicting success. This is not surprising given how interrelated semester GPA is with success in units and success in courses. What is surprising in this graph is to see students with a low GPA being labeled as successful, some of the observations involved students taking Pass grades instead of letter grades. There were also students with a high GPA being labeled as unsuccessful, some of these involved students dropping courses but being successful in the ones they completed.

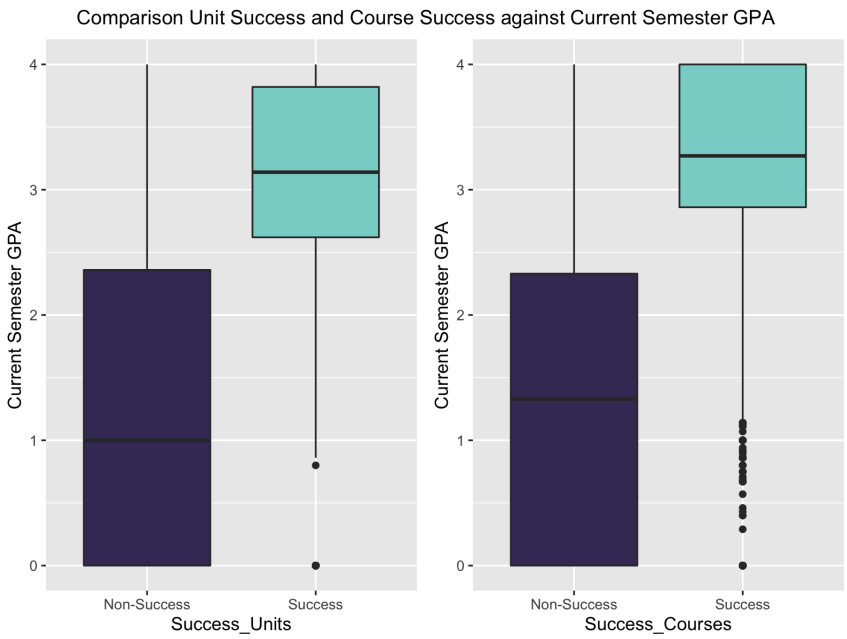

Figure 10: Boxplot of semester GPA against Success

Figure 11, shows a density plot where cumulative GPA is plotted against course success at the 80% level for 4 consecutive major terms. The interesting thing about this graph is that cumulative GPA is partially separating success from non-success consistently throughout the chosen terms. This graph is also suggesting that we should explore success at levels different from the 80% level, this may provide a better separation of success and non-success.

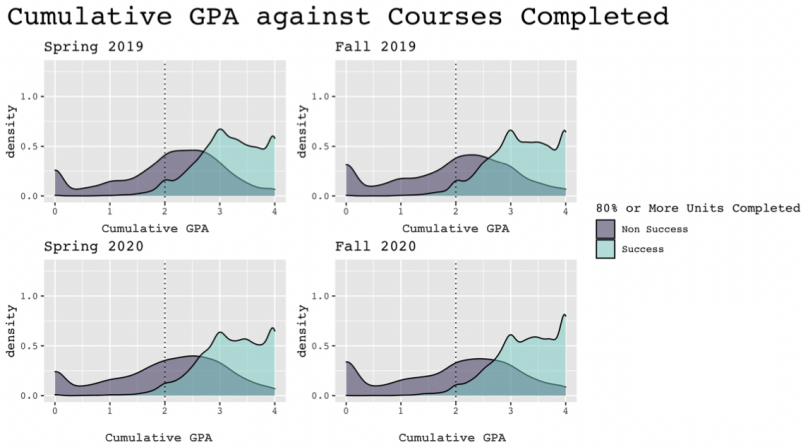

Figure 11: Density plot of cumulative GPA against Success

The data visualization has served the team in two ways, through the generated plots it has helped refine the data clean up by identifying cases that involve an excessive number of cumulative units and semester units, unusual number of units, and more. It has also helped identify some variables that may have some predictive value for the defined measures of success.

4.2 Clustering Data

We conduct a clustering algorithm because they are most useful at partitioning a dataset into homogeneous clusters (groups). That is, each cluster is constructed such that objects within the same clusters are more similar to each other than objects in different clusters. We are hopeful that this will lead to groups where we can best identify which students are more likely to be non-successful versus those who are more likely to be successful. In this section, we will be focusing on three aspects of clustering: (1) Discussing the clustering theory, (2) Exploratory Data Analysis via Plots, and (3) Using clusters to predict future students.

In the first part, we will be talking about the theory of Clustering using K-Prototypes. Since our data is a combination of both numerical and categorical data, we decided to use an R package called **clustMixType** (Szepannek 2009). Szepannek R package is an extension of (Huang 1998) k-prototypes algorithm, a combination of k-means and k-modes for mixed data. The algorithm randomly selects k clusters, whose centers are called prototypes. Using an iterative process, prototypes and clusters are reallocated until the total within-cluster variation is minimized. This leads to clusters that share common themes and thus will allow us to better understand why certain students are successful or not.

The formula for commuting the clusters is given by:

$$P(W, \mathbf{Q}) = \sum_{l=1}^{k} \Big[\sum_{i=1}^{n} w_{i,l} \sum_{j=1}^{m_r} (x_{i,j}^r - q_{l,j}^r)^2 + \gamma \sum_{i=1}^{n} w_{i,l} \sum_{j=m_r+1}^{m_c} \delta(x_{i,j}^c, q_{l,j}^c) \Big] \qquad (*)$$

x_r and x_c are numerical and categorical data, m_r and m_c number of numerical and categorical data, $\sum_{i=1}^{n} w_{i,l} = 1$ (weights), γ a tuning parameter. Furthermore,

$$\delta(x_{i,j}^c, q_{l,j}^c) = \begin{cases} 0 & x_{i,j}^c = q_{l,j}^c \\ 1 & x_{i,j}^c \neq q_{l,j}^c \end{cases}$$

where $q_{l,j}^r$ and $q_{l,j}^c$ are the centroids (mean) and modes of clusters l on a given iteration.From here there is an iterative process. Each iteration compute k-prototypes until the total within cluster distance is minimized. The iteration goes as follows: (i) Choose the number of clusters k, (ii) Select k random points from the data as prototypes, (iii) assign all points to the closest clustering prototype using $(*)$, (iv) recompute the prototype of the newly formed clusters, (v) repeat (iii) and (iv) until within group sum of squared errors (WGSS) is minimized.

Furthermore, the tradeoff between the categorical and numerical values is controlled by the tuning parameter γ. This formula is a combination of k-means (numerical data) and k-modes (categorical data). When γ is small, this leads to a model more heavy on k-means and when γ is larger, this favors k-modes. (Huang 1998) took γ to be the average standard deviation of the numerical values and in some applications $\frac{1}{3}\sigma \leq \gamma \leq \frac{2}{3}$. The paper talks about other ways to tune this parameter.

The next step for our team was to determine the optimal number of clusters (k). To do this, we used an Elbow Plot. The Elbow Plot has the number of clusters (k) on the x-axis and on the y-axis the within group sum of squared error (WGSS). The theory behind this plot is based on mathematical optimization. We search for the point where diminishing returns are no longer worth the additional cost. That is, increasing the number of clusters will lower our WGSS, but at some point we will be overfitting the model. The elbow reflects where this over-fitting will start to happen. From our plot below, we can see that a safe range of values falls between $[5, 7]$.

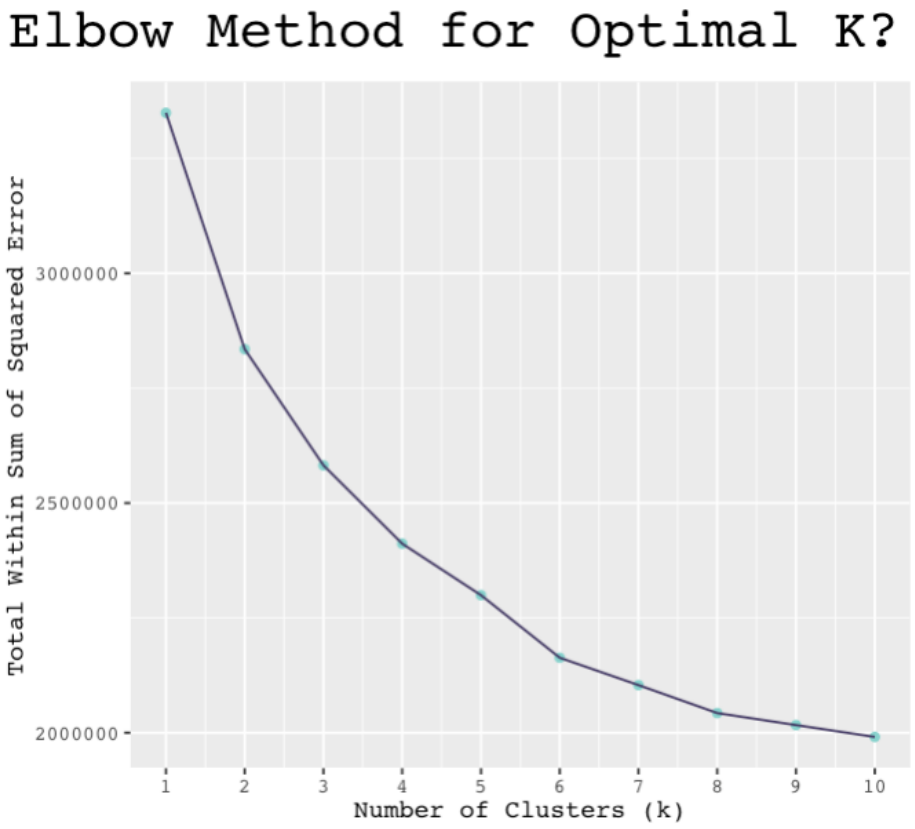

Figure 12: Plot showing elbow method for optimal number of clusters. K = 6 seems to produce reasonable clusters.

After plotting different values of k (5,6,7), we decided it would be best to use $k = 6$. The k=prototypes clustering output gave the following statistics:

Cluster	Cluster Size	Percentage of Data	WGSS
1	13,285	15%	443352.7
2	12,938	15%	385247.3
3	16,118	19%	109756.5
4	14,722	17%	353763.5
5	12,112	14%	390439.1
6	17,477	20%	443814.2

Now that we have obtained our clusters, we can use exploratory data science to see trends within each cluster. The goal is to determine which clusters consistently have successful students versus non-successful students. We can now make predictions to determine whether students will be more likely to be classified as successful or non-successful. SCC can then provide non-successful students with the necessary resources to be more successful. Lastly, we wanted to determine if using the 80% threshold was an optimal cutoff for student success. Below we can see the various clusters with our first response variable.

When investigating these plots in Figure 13, the graphs and clusters on the left hand side show more non-success students and the graphs on the right show higher level of success. We will continue this theme to see if similar patterns emerge with other variables. Here one can see that the order of the cluster numbers are placed in a ranking system. Note, cluster 5 has more students being identified as successful, but the percentage is not as high as clusters 4,2, and 3, so we placed it on the left hand side.

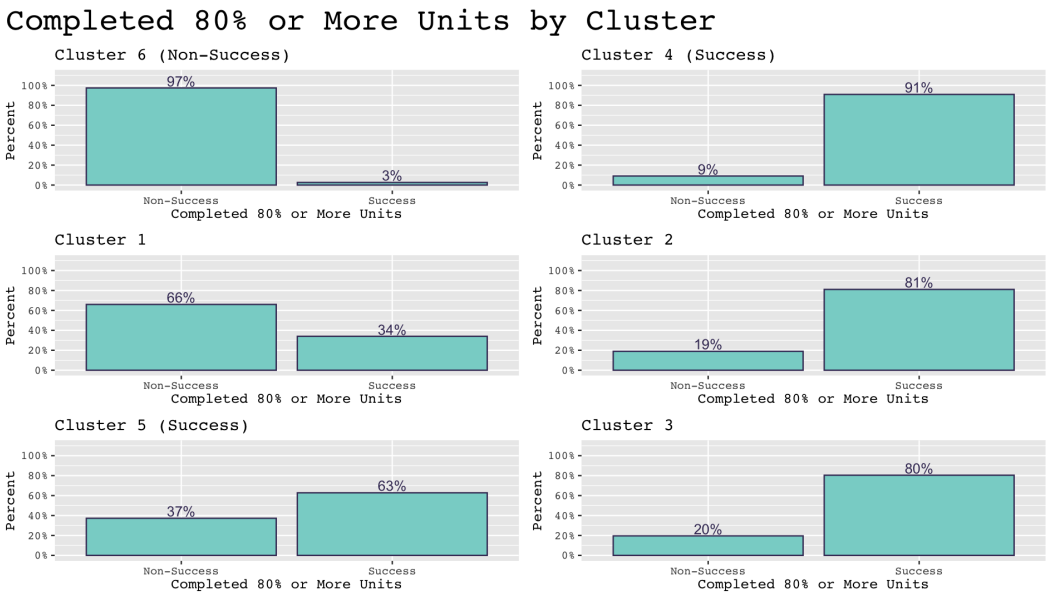

Figure 13: Completed 80 Percent or More Units Clusters

Moving forward, we will only focus on the top two clusters for non-success and success, hoping to not get lost in too many plots. This next plots shows the distribution of GPA and cumulative GPA over all clusters.

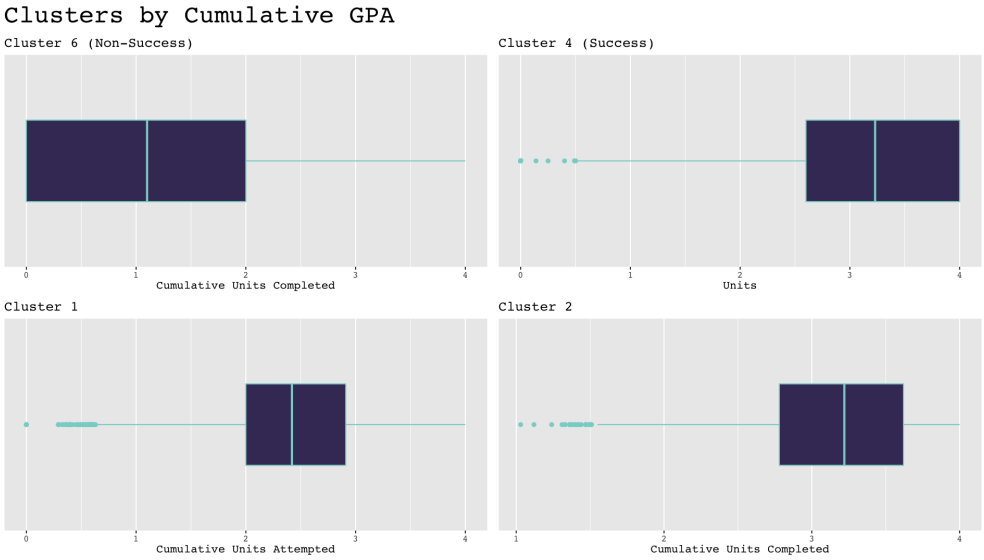

Figure 14: Semester GPA and Cumulative GPA by Clusters

We can see that on average the plots on the right hand side have a higher cumulative GPA, on average, and students on the left have lower GPAs, on average, indicating more at-risk outcomes. This is consistent with our response variable, success on the right hand side and non-success on the left. Note, cluster 3 has a high percentage of students with high GPAs, but this is because it could have been placed on the right hand side with more successful students falling in this group.

The next plot is showing the distribution of ethnicity across the top non-success and success clusters. Recall in our literature review that Ms. Santiago reported that "While student enrollment at community colleges has increased over the last ten years, degree completion has not grown as quickly for Latino students." We can see that overall, SCC has roughly 51% Latino students. When we look at the top successful student cluster (4), the percentage is (34.87%), which is much lower than our population percentage. Similarly, when we dive into the top at-risk cluster (6), the percentage of Latino students is 58.034%, which is much higher than the population parameter at 51%. This informed us that us that Latino students at SCC are facing similar problems that Ms. Santiago pointed out.

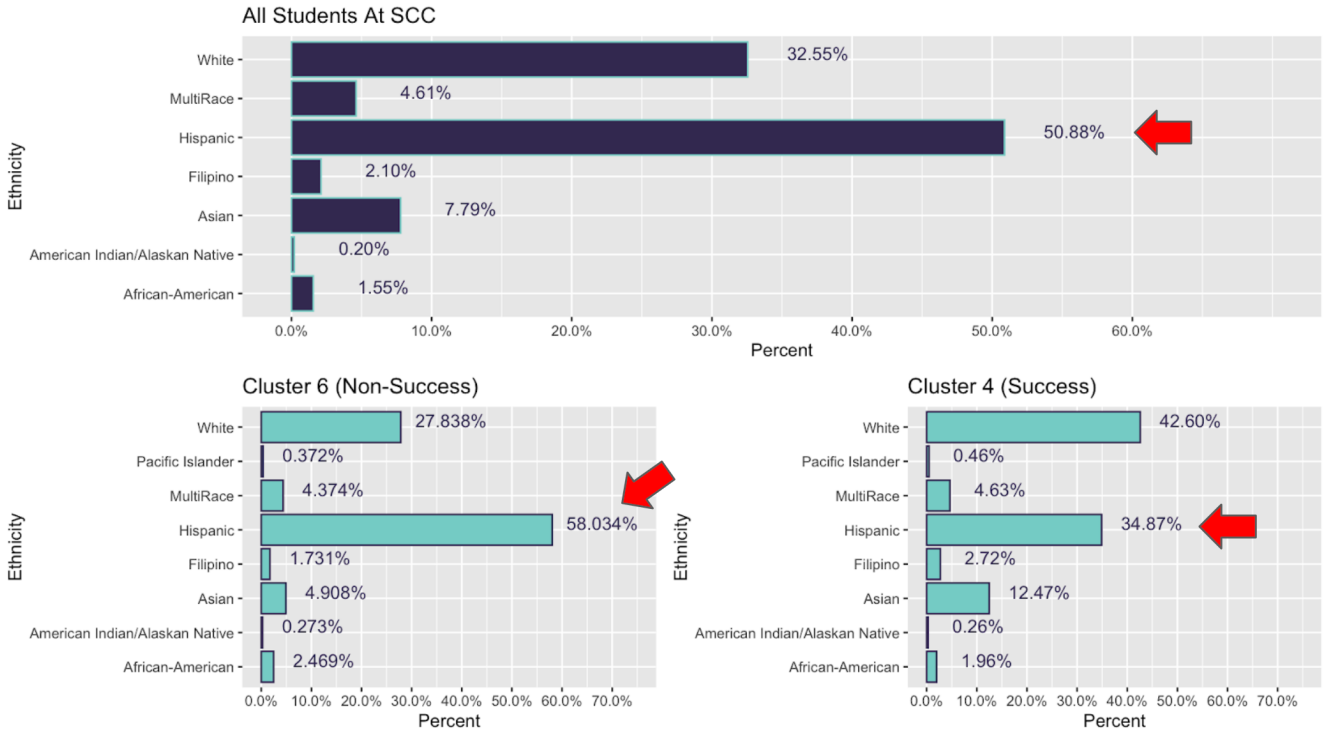

Figure 15: Success Rates by Ethnicity within Clusters

When we look at Cluster 4 (highest success cluster), the percentage of Latinos is 34.87%, which is much lower than our population percentage at 51%. For the top non-success cluster (6), the percentage of Latino students is 58.034%, which is much higher than the population parameter at 51%.

The next step is to look at the plot below (Percentage of Units Completed by Clusters) and decide if the 80% threshold for our response variables is optimal.

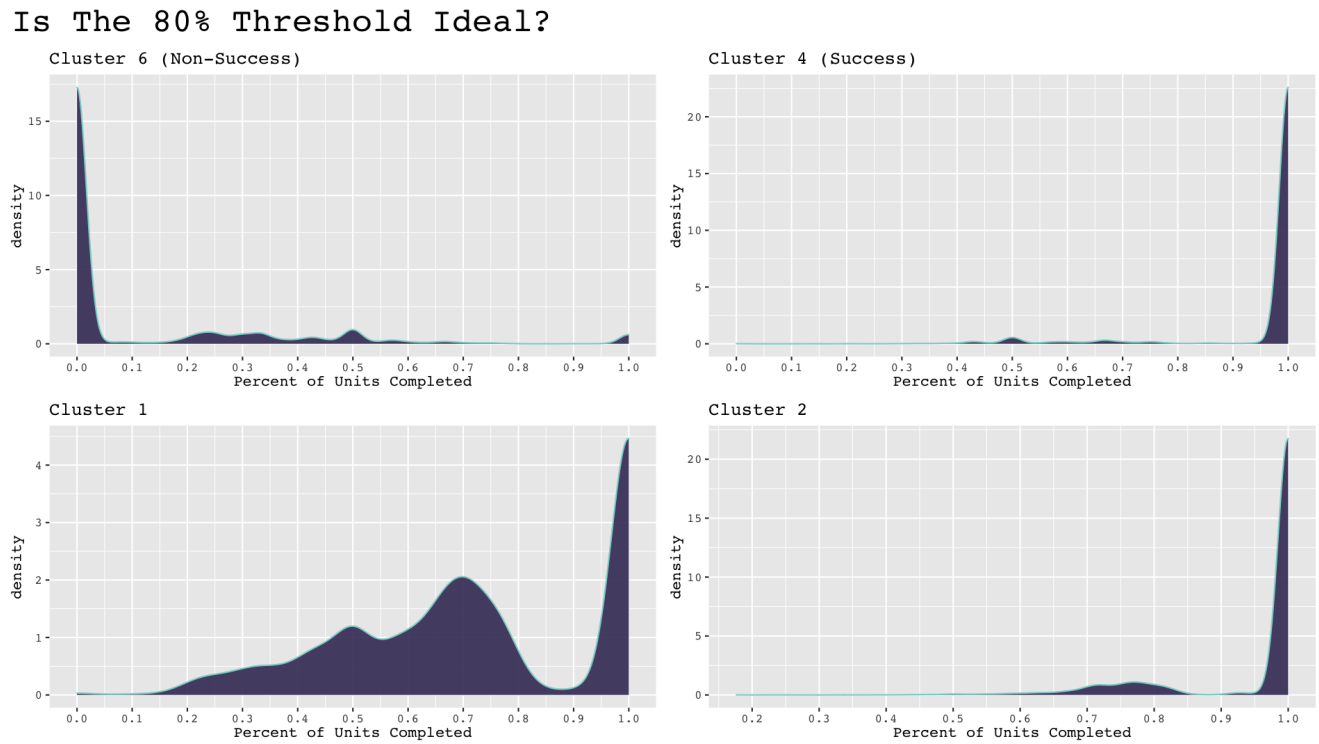

Figure 16: This graph shows the distribution of percentage of units completed for the top 4 clusters, where the top two non-success clusters are on the left and the top two success clusters are on the right.

From the plot above, we can see that the successful clusters (4 and 2) have the majority of their density at 95% and above. If we were to take the threshold at 100%, a lot of students at 95% would be flagged as non-successful. Based on the successful clusters, any choice from $[70\%, 80\%]$ would suffice. We can also see that the majority of students being classified as non-successful in Cluster 6 was due to them having only completed 5% and below or $[20\%, 60\%]$. Cluster 1 had 66% non-success and 34% successful students. We can see from the Cluster 1 density plot (Percentage of Units Completed) that the density from 0.8 and above represents the 34% and 0.8 percent and below much represent the 66%. It is hard to save exactly where the threshold should be classified at optimal, however, it is clear that the 100% threshold is too stringent.

5 Methods

5.1 Variable Selection

After clustering showed there are inherent groups of successful and non-successful students within the data, we used regularization methods and random forests with various importance metrics to find which variables have the greatest impact on success and non-success. We provide a quick summary of each method and metric below.

Regularization is the method of fitting a linear regression model with a penalty term to avoid overfitting and allow a more flexible model than ordinary least squares regression. They aim to minimize or completely eliminate the effect of collinear and uncorrelated predictor variables. For a general two-dimensional example of regularization, we have provided an example in Appendix: Regularization from The Elements of Statistical Learning (Hastie 2009).

For our analyses, we removed semester-specific variables (courses attempted per term, units attempted per term, courses dropped per term, courses attempted per term, term GPA, etc.), because those variables were used in calculating the response variables. Since those variables were used to calculate the response variables, if we left those variables in for consideration, then the regularization methods we chose would undoubtedly select those variables as important. Then, we built lasso, ridge, and elastic net models using all terms (COVID and pre-COVID). We extracted the important variables based on the absolute value of the variable coefficients in the model, where a variable coefficient with a greater magnitude indicates a more important variable.

Considering both response variables (unit-based success/non-success and course-based success/non-success) four variables continually ranked among the largest coefficients in each regularization model: cumulative GPA, cumulative units attempted, cumulative units attempted, and the student's full-time status. Although these four variables appeared in different rank order depending on the model, each of these four variables did appear in the top four most important variables of each model based on the magnitude of their accompanying coefficients.

Table 5: Top Ranked Variables Using Regularization Methods for Response: >80% Units Passed

Rank	Lasso	Ridge	Elastic Net
.	abs(coef)	abs(coef)	abs(coef)
1	Cum. Units Att.	Full-time Status	Cum. Units Att.
2	Cum. Units. Comp.	Cum. GPA	Cum. Units. Comp.
3	Full-time Status	Cum. Units Att.	Full-time Status
4	Cum. GPA	Cum. Units. Comp.	Cum. GPA

Table 6: Top Ranked Variables Using Regularization Methods for Response: >80% Courses Passed

Rank	Lasso	Ridge	Elastic Net
.	abs(coef)	abs(coef)	abs(coef)
1	Cum. GPA	Cum. Units Att.	Cum. Units Att.
2	Full-time Status	Cum. Units. Comp.	Cum. Units. Comp.
3	Cum. Units Att.	Cum. GPA	Cum. GPA
4	Cum. Units. Comp.	Full-time Status	Full-time Status

Next, we used Random Forests (Breiman 2001) to rank the importance of all variables excluding the semester-specific ones previously mentioned. A random forest is a collection of classification trees with splits on random variables at values that provide the best probability of correct classification. Random forests use the majority vote of classification trees in a forest to classify students as a success or non-success. While classification trees individually are a weak predictor (not accurate by itself), as a collection of trees in a random forest, they are a strong predictor. For an example of a classification tree and further discussion on random forests, refer to Appendix: Classification Trees and Random Forests.

We explore three different type of random forest implementations: Breiman's original paper on random forests (Breiman 2001); ranger, a fast implementation of random forests suited for high-dimensional data (Wright 2017); and regularized random forests (RRF), which discourage splitting on a new feature unless there is substantial information to be gained from that split (Deng 2012). When available in each package, we use three different measures of variable importance: mean decrease in Gini impurity, mean decease in adjusted Gini impurity, and permutation error accuracy.

Gini impurity measures the homogeneity of a group, that is, it measures how "mixed" a group is. A smaller Gini impurity value indicates a more pure group - a group with a large number of elements within a fewer

categories. A larger Gini impurity value indicates a more mixed group - a group containing many categories with a few elements in each group. In our case, Gini impurity would measure how well a split on a variable at some value divides the dataset into cases of success and non-success. Then, a mean decrease in Gini measures the mean decrease in impurity in all trees on the given variable - a larger mean decrease in Gini indicates a more important variable (Nembrini 2018). Adjusted Gini impurity is an attempt to correct bias in the Gini impurity value - a larger adjusted Gini impurity value indicates a more important variable as well (Sandri 2008). Permutation error accuracy measures the affect on accuracy of shuffling the values for a variable - a larger permutation error accuracy indicates a more important variable (Breiman 2001).

We built random forests models using each algorithm then ranked the importance of each variable in each model. We did this for each available importance metric available to each method, using all terms and for both response variables. Five variables continually ranked as the most important variables regardless of the method or metric: cumulative GPA, cumulative units completed, cumulative units attempted, the student's full-time status, and the student's academic standing. This closely matches the results of variable importance that the regularization methods arrived at with the addition of student's academic standing variable. We provide a sample of the top-ranked variables for each method and metric below.

Table 7: Top Ranked Variables Using Random Forests for Response: >80% Units Passed

Rank	Classic RF	Ranger (Impurity)	Ranger (Permutation)	RRF
.	Mean Decrease Gini	Impurity Corrected	Permutation Error	Mean Decrease Gini
1	Cum. GPA	Cum. GPA	Cum. Units. Comp.	Cum. GPA
2	Cum. Units Comp.	Cum. Units Comp.	Cum. GPA	Cum. Units. Comp.
3	Cum. Units Att.	Full-time Status	Cum. Units Att.	Cum. Units Att.
4	Full-time Status	Cum. Units Att.	Full-time Status	Academic Program
5	Student Standing	Student Standing	Cum. Units Cat.	Geo. Latitude

Table 8: Top Ranked Variables Using Random Forests for Response: >80% Courses Passed

Rank	Classic RF	Ranger (Impurity)	Ranger (Permutation)	RRF
.	Mean Decrease Gini	Impurity Corrected	Permutation Error	Mean Decrease Gini
1	Cum. GPA	Cum. GPA	Cum. GPAC	Cum. GPA
2	Cum. Units Comp.	Cum. Units Comp.	um. Units. Comp.	Cum. Units Att.
3	Cum. Units Att.	Student Standing	Cum. Units Att.	Cum. Units Comp.
4	Student Standing	Full-time Status	Full-time Status	Geo. Latitude
5	Geo. Longitude	Cum. Units Att.	Sessions Comp.	Academic Program

We repeated the process for discovering the top predictors for success to discover the top predictors for non-success, and to investigate whether the results changed during the Spring 2020 and Fall 2020 COVID affected terms. The top predictors for non-success were very similar results to the top predictors to success. The variables which ranked most important to success also ranked highly in importance to non-success. During COVID-19 affected terms, the same variables again ranked highly important - the only noticeable difference being that the order of importance of some variables may be shuffled around depending on the method, importance measurement, or response variable used. In some cases, the cumulative units attempted by a student would rank higher for success and non-success during COVID terms.

Beyond the four or five important variables determined by regularization methods or random forests, several variables commonly ranked highly in variable importance across various methods and importance measurements. Among all methods of variable selection and measures of importance, the nine variables that frequently appeared among the most important are shown in the table below. A full table of variable importance is available in the "variable_importance" spreadsheet.

Table 9: Most Common Important Variables Among All Methods and Measures

Rank	Variable
1	Cumulative GPA
2	Cumulative Units Completed
3	Cumulative Units Attempted
4	Full-time Status
5	Age
6	Student Standing
7	Session (Terms) Completed
8	Geographic Latitude
9	Geographic Longitude

5.2 Model Justification and Selection

Model selection is loosely defined as selecting a "best" model among several competing candidate models. This section will serve as a brief refresher of things to consider when choosing the appropriate techniques when developing candidate models. Below are a few questions any team should ask themselves during this process.

Q: What is the structure of the data that has been provided?
A: As noted in previous sections, our data is mixed i.e. a combination of numerical and categorical preditors.

Q: What is the form of the response variable or variables?
A: We have several response variables representing measures of educational success, all are binary {0,1}.

Q: How large is the dataset you will be working with in terms of dimensionality?
A: The initial datset (prior to any data reduction efforts) was 340,141 x 55, after it's 86,650 x 34.

Q: Are there any foreseeable computational or technological constraints?
A: Our models will be built locally by a geographically distributed team and this can present challenges.

Q: Is interpretability an important consideration or requirement for your project?
A: Achieving clear and concise model interpretability is a primary goal of this project.

Q: Does our model need to be self-contained, deliverable, or reusable by the client?
A: Ideally, a model could be rerun by the client at the beginning / middle / end of each academic term.

Having sufficiently answered these questions we proceeded to build our candidate models where we could undergo a champion/challenger situation and evaluate individual model performance using variables deemed to be important or useful based on the analysis in the previous sections. Those candidate models are LASSO Regression, Logistic Regression, and Random Forest.

5.3 Model Evaluation

After identifying the top nine predictors for success and non-success, we developed LASSO regression, logistic regression, and random forest models to quantify how well our top predictors actually predict success or non-success. We chose a LASSO regression model, because it is flexible, interpretable, and we already found some success using LASSO regression in previous sections of this project. Due to the binary nature of our response variable, we felt logistic regression was a natural candidate for modeling. Then, chose to develop a random forest model because we wanted to explore a non-parametric method that could easily handle both categorical and numeric variables, in addition to the previous success we had using random forests in determining variable importance.

We randomly split the data into training and testing sets, where 80% of the data went to training the models using only the nine predictors we identified, and 20% went to test the models' accuracy in predicting success with an 80% units completed threshold. To compare the performance of each classification model, we plotted ROC curves for each model on all terms in the figure below.

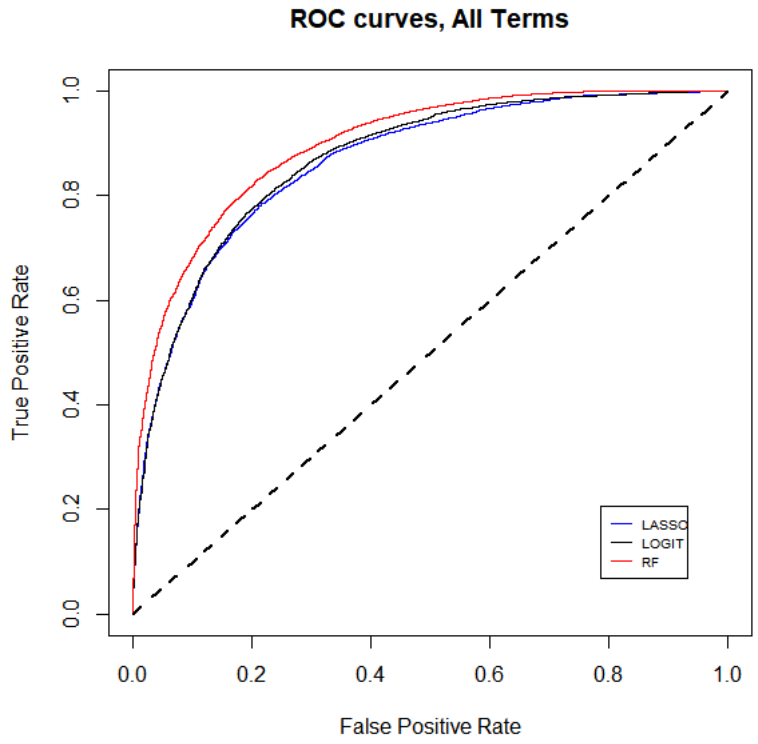

Figure 17: Academic Status (Completed 80 Percent or More Courses) in Relation to Gender

Classifiers with ROC curves closer to the upper left corner indicate better performance. Under ideal conditions a perfect classifier would have a false positive rate of 0% and true positive rate of 100% represented by the upper left-hand corner of this graph. As a baseline, the dashed diagonal y = x line signifies no predictive value, and would have an AUC of 0.5. In figure 17, notice how the LASSO and LOGIT plots are close to each other showing that the models perform similarly, with random forest doing slightly better. The performance of each model is evaluated by their accuracy, and their area under the curve (AUC) in the ROC plots. We provide a summary table below that quantifies the results shown in the ROC curve plots. The closer these numbers are to one the better their performance. LASSO and Logistic regression perform similarly to one another with accuracy rates of 0.787 and 0.796 respectively, random forests perform slightly better than both with an accuracy of 0.814. LASSO and Logistic regression also perform similarly when evaluating performance with AUC with values of 0.867 and 0.873 respectively, random forests again perform slightly better than both with an AUC of 0.898. Overall, our models were able to predict success or non-success 80% of the time, and achieve AUC's between 0.867 and 0.898 using only the nine important predictor variables we identified.

Table 10: Performance Metrics for Models Using Top Predictors

Model	Accuracy	AUC
LASSO	0.787	0.867
Logistic	0.796	0.873
Random Forests	0.814	0.898

6 Conclusion

6.1 Summary of Key Findings

In this section we'll provide a brief overview of the key findings of this project as they relate to the goals of this project. For goal number 1 we established that the best threshold for success on all terms, including COVID terms, was an 80% threshold on units completed. To accomplish this task, we incorporated contingency tables and developed a logistic regression model in the data wrangling section of this paper. We found that a 100% threshold, in general, on units and courses, was too stringent a threshold and would falsely flag students for non-success despite completing the term with a high GPA or good grades. Additionally, we found thresholding on the number of courses was too strict because it did not allow student's much flexibility to drop courses and still be considered successful; if a student takes fewer than six courses and drops one, then even if a student passes all their classes, they would still be considered a non-success in that term. For these reasons, we determined that an 80% threshold on the number of units attempted the most appropriate response variable for determining success, on all terms including COVID terms

For goal number 2 we determined that living in a low income area provided more predictive value than the low income status as reported on student's financial aid documentation. Interestingly the opposite was true when looking only at terms that occurred during the pandemic (Fall 2020 and Spring 2020). This determination was made by again building a simple logistic regression model using only the provided and derived income variables and then compared the test statistics. This process gave a clear picture of how these variables interacted with the response over time.

Finally in goal number 3, using a rigorous cross-validation method involving multiple random forest implementations, we discovered that the same nine variables were frequently found to be important predictors for success and for non-success regardless of whether the terms were consider COVID or non-COVID. Many of the top predictors are historical student data such as cumulative GPA, or cumulative units attempted indicating that best predictors for future student success or non-success would be previous student performance. Other top predictors include a student's full-time status, age, academic standing, number of terms completed, and their home location. Besides, a student's academic history, these would be the other variables one would want to closely monitor, collect, or investigate to predict student's success or non-success.

6.2 Future Work

Taking a forward looking view, there are other avenues of research we would have liked to explore. Given additional bandwidth, we would have certainly investigated how other statistical and machine learning models performed when applied to this problem. For example, neural networks could be an interesting and effective choice. Another approach we would have liked to investigate given more time would be to run k-prototypes clustering and perform variable importance methods within each cluster to determine which variables are most prominent in the successful clusters and non-successful clusters, and would be unaffected by correlated variables.

For a predictive engine like those discussed here to truly perform as an early warning system it would need to be executed early/mid semester. This would require the elimination of current semester data like final grades, G.P.A.s, etc.; and the inclusion of intra-term data like current grades. We might also consider tracking additional variables such as time utilizing campus resources, extra-curricular activities, and instructor information (full-time, part-time, number of courses, taught, cumulative number of courses taught, etc.). Though we did determined the 100% threshold was too restrictive. Perhaps 80% is not ideal either. It's likely the best response variable would be derived from a threshold somewhere between 80% and 100%.

Finally, when running our model for variable importance, we had to remove some variables due to high collinearity (drop courses, semester units attempted, etc.) Instead, we could have run k-prototypes clustering and within each cluster one could run a LASSO model to determine which variables were most important. This would yield which variables are most prominent in the successful clusters and non-successful clusters, and would be unaffected by correlated variables.

7 References

Belfield, Clive and Crosta, Peter M. 2012. *Predicting Success in College: The Importance of Placement Tests and High School Transcripts.* Teacher College, Columbia University. CCRC Working Paper No. 42

Breiman, Leo. 2001. *Random Forests.* Machine Learning. Volume 45: 5-32.

California Median Household Income Zip Code Rank. USA.com. World Media Group, LLC., 2021. http://www.usa.com/rank/california-state--median-household-income--zip-code-rank.htm.

Deng, Houtao and Runger, George. 2012. *Feature Selection via Regularized Trees.* [online] arXiv.org. Available at https://arxiv.org/pdf/1201.1587.pdf [Accessed April 25, 2021]

Eagan, M. Kevin Jr and Jaeger, Audrey J. 2009. *Effects of Exposure to Part Time Faculty on Community College Transfer.* Research in Higher Education. Volume 50:2 168-188

Fauria, Renee M and Fuller, Matthew B. 2015. *Transfer Student Success: Educationally Purposeful Activities Predictive Of Undergraduate GPA.* Research and Practice in Assessment. Volume 10. 39-52

Gareth, James et al.. 2013. *An Introduction to Statistical Learning: With Applications in R.* New York: Springer.

Hastie, Trevor, Robert, Tibshirani and Friedman, Jerome. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.

Huang, Zhexue. 1998. *Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values.* Data Mining and Knowledge Discovery, 2: 283-304

Lamar University. 2019. *Correlation Between Parents' Education Level and Children's Success.* [Accessed March 7, 2021]. https://degree.lamar.edu/articles/undergraduate/parents-education-level-and-childrens-success/

Linderman, Donna and Kolenovic, Zineta. 2009. *Early Outcomes Report For City University Of New York (CUNY) for Accelerated Study In Associate Programs (ASAP).* CUNY/NYC Center for Economic Opportunity

Linderman, Donna and Kolenovic, Zineta. 2013. *Moving the Completion Needle.* Change: The Magazine of Higher Learning. Volume 45:5 43-50.

Nembrini, Stefano, Konig, Inke R., and Wright, Marvin N.. 2018. *The revival of Gini importance* Bioinformatics. Volume 34, Issue 21: 3711-3718.

Santiago, Deborah and Stettner, Andrew. 2013. *Supporting Latino Community College Students.* Excelencia in Education.

Sandri, Marco and Zuccolotto, Paolo. 2008. *A Bias Correction Algorithm for the Gini Variable Importance Measure in Classification Trees.* Journal of Computational and Graphical Statistics. Volume 17, Issue 3:

611-628.

Speiser, Jaime Lynn et al.. 2019. *A comparison of random forest variable selection methods for classification prediction modeling.* Expert Systems with Applications. Volume 134: 93-101.

Stanford University. 2014. *Glmnet Vignette.* [Accessed April 2021]. http://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html

Szepannek, Gero. 2018. *clusMixType: User-Friendly Clustering of Mixed-Type Data in R.* The R Journal, Vol. 10/2 (December): 200-208. https://journal.r-project.org/archive/2018/RJ-2018-048/RJ-2018-048.pdf

Turk, Jonathan M. 2020. *Identifying Predictors Of Credential Completion Among Beginning Community College Students.* Center for Policy Research and Strategy, American Council on Education

Wright, Marvin N., and Ziegler, Andreas. 2017. *ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R.* Journal of Statistical Software. Volume 77:1.

8 Appendices

8.1 Regularization Visualization

On the axes are coefficients β_1 and β_2 for two variables. The ordinary least square solution, the value of (β_1, β_2) that minimizes the sum of squared error, is represented by $\hat{\beta}$. Regularization methods allow a flexible solution depending on the amount of error permitted training a model represented by the red ellipses. The blue fields show the constrained region where a solution exists. The point where the red ellipses contacts the blue fields denotes a solution and set of (β_1, β_2) values.

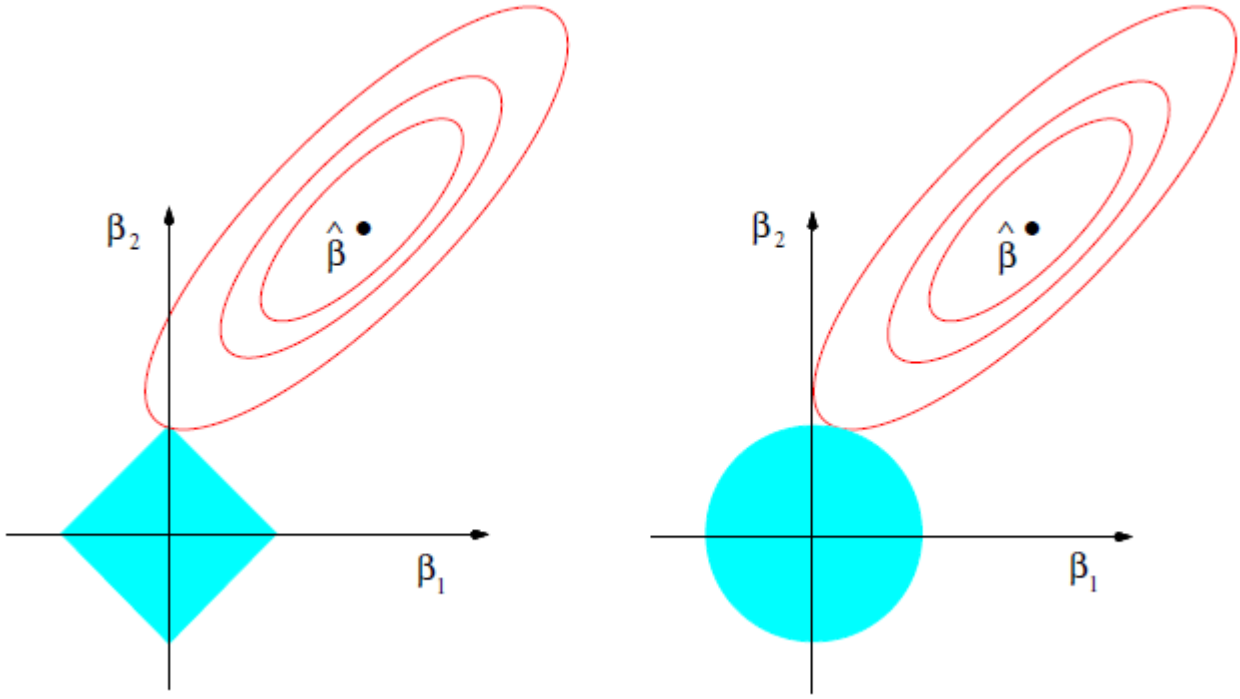

Figure 18: Visual Representation of Lasso and Ridge Regression in Two Dimensions

On the left side of the figure is an example of lasso regression where the constrained solution space is visualized as a diamond. As demonstrated in the figure, the red ellipses hit the blue diamond at a corner indicating a solution where $\beta_2 = 0$, eliminating the variable associated with the coefficient β_2 as an important factor. Lasso regression has the potential to perform dimensionality reduction, where variables are eliminated from a model, because of its diamond-shaped solution space.

On the right side of the figure is an example of ridge regression which performs similarly to lasso regression, but instead has a circular constrained solution space. Ridge regression will still find a vector of coefficients for variables in the model just like lasso regression. However, unlike lasso, since ridge regression uses a circular solution space it is much less likely to eliminate a variable, but will instead return coefficients close to zero for unimportant variables.

Another regularization method not shown here is an elastic net, which combines the characteristics of both lasso and ridge regression. The shape of the constrained solution space would look like a compromise of the lasso and ridge regression solution spaces, similar to a bloated diamond with sharp edges on the axes and curves connecting edges.

8.2 Classification Trees and Random Forests

An example of a classification tree is given below. As you go down the classification tree, it splits the data based on some condition related to the variables. In this example, if a student had a cumulative GPA greater than or equal to 2.4 in a term, the classification tree classifies them as a success. If they had a cumulative GPA less 2.4, but greater than or equal to 1.9, the classification tree checks if the student attempted more or less than 5.7 units in a term, then whether they were in a full-time status or not.

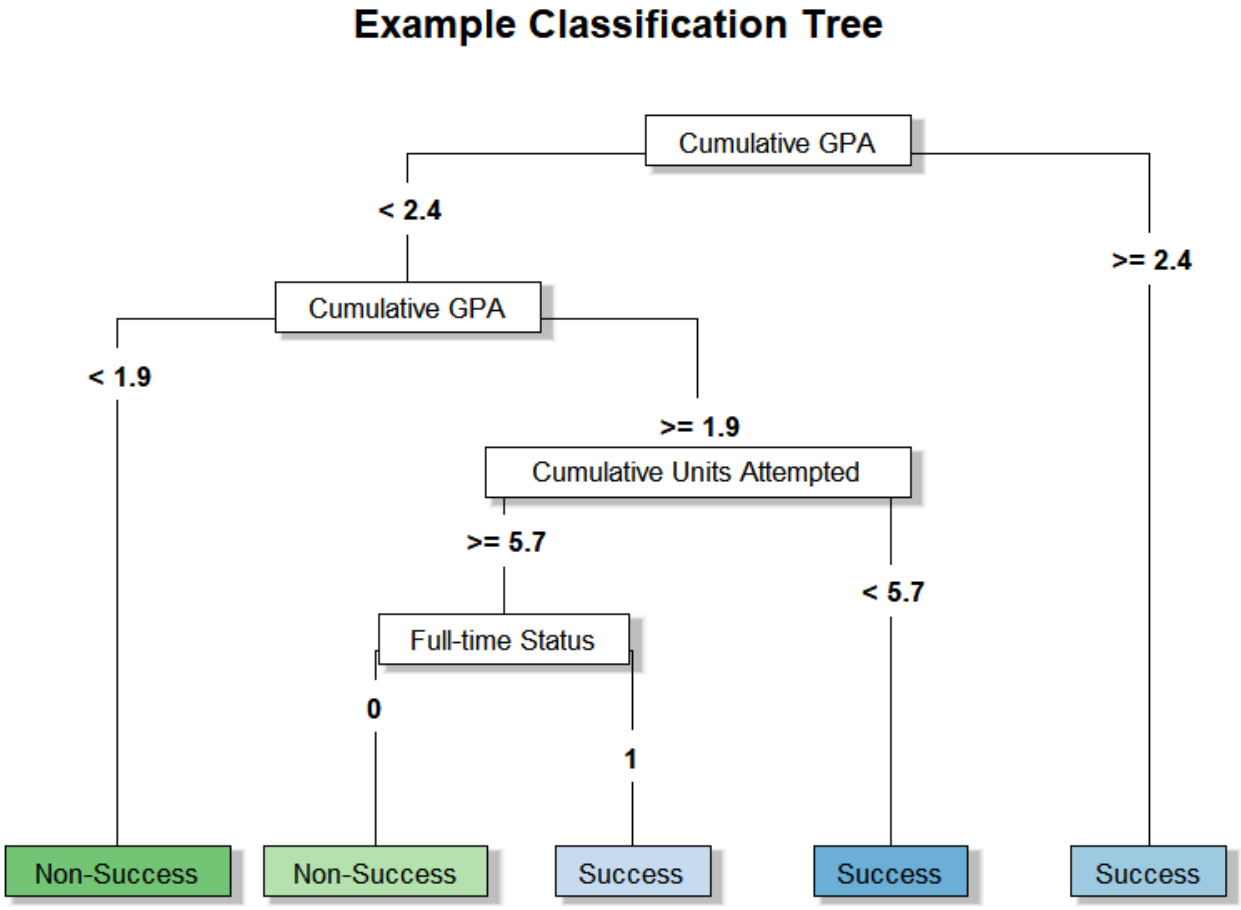

Figure 19: An Example of a Classification Tree; Random Forests are Collections of Classification Trees

Random forests are comprised of many (500 specifically for our analysis) of these random forests. The classification trees are created by splitting at random variables, then random forests use the majority prediction or "vote" of all the trees in a forest to classify a student in a term as a success or non-success. Although a classification tree by itself is a weak predictor (not high accuracy), a collection of random classification trees in a random forest creates a strong predictor with higher accuracy.

8.3 Additional Visualizations

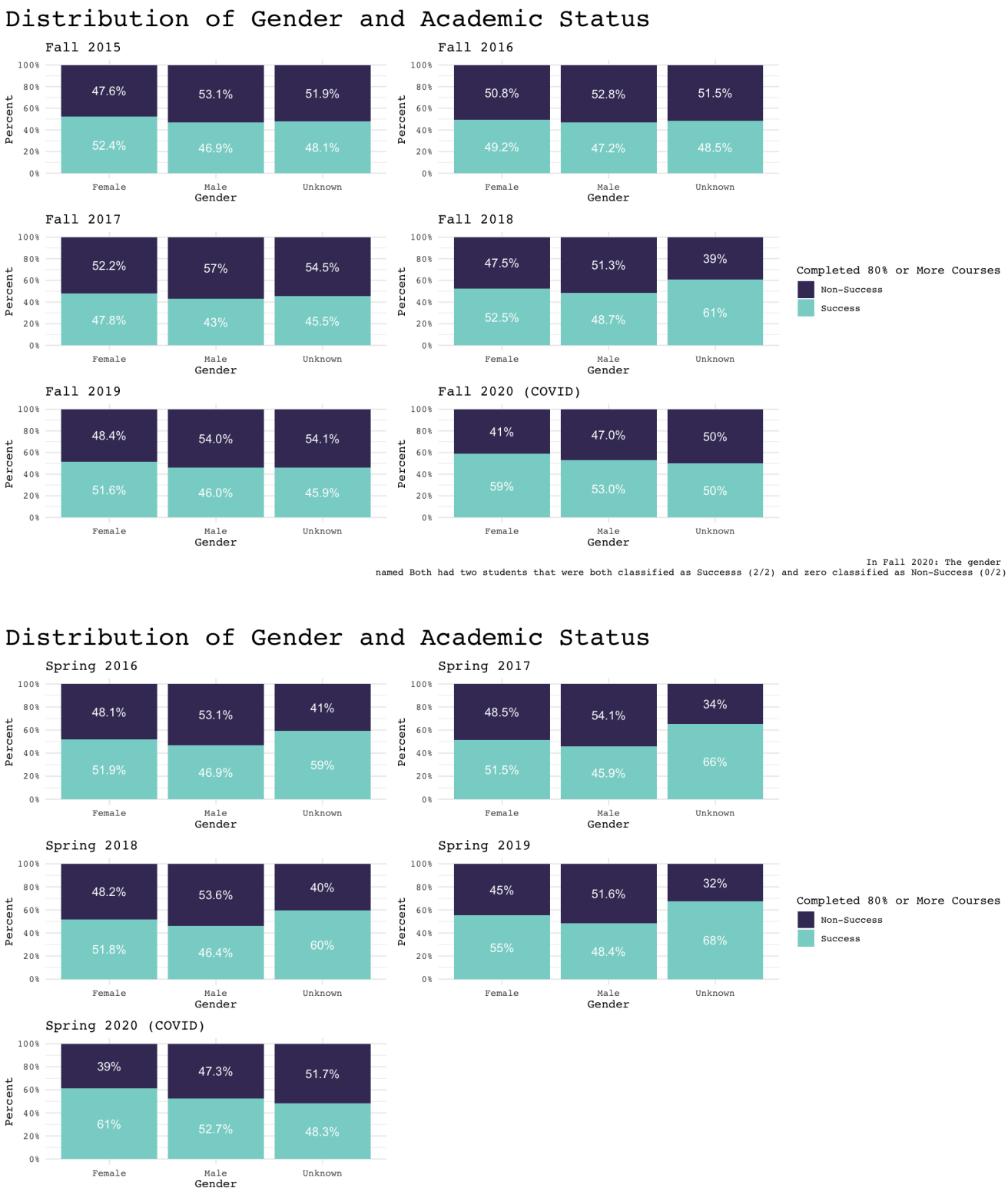

Figure 20: Here are the graphs of both Fall and Spring Terms in association with Gender

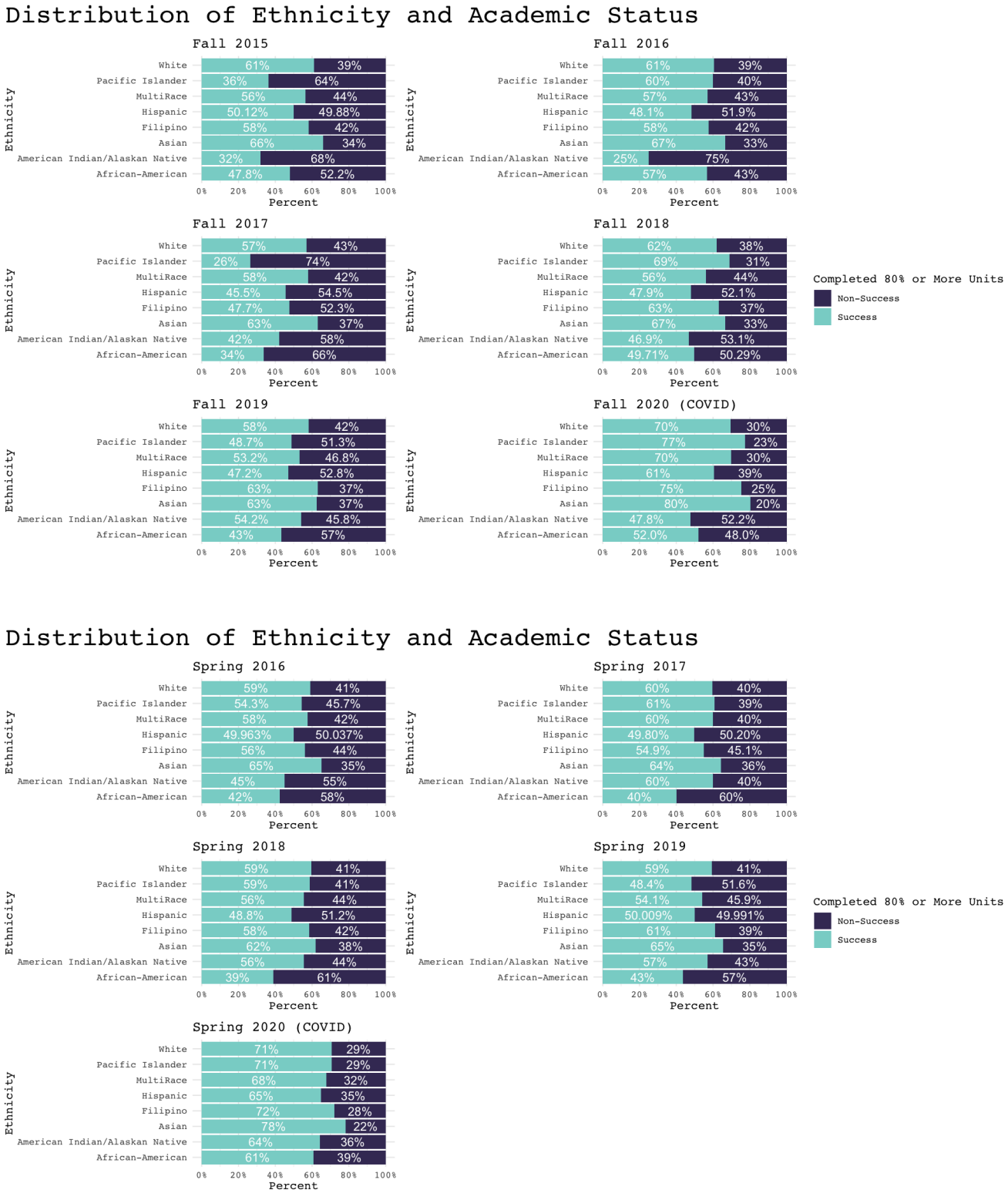

Figure 21: Here are the graphs of both Fall and Spring Terms in association with Ethnicity

Figure 22: Academic Status (Completed 80 Percent or More Courses) in relation to Taking Math and English within First Year

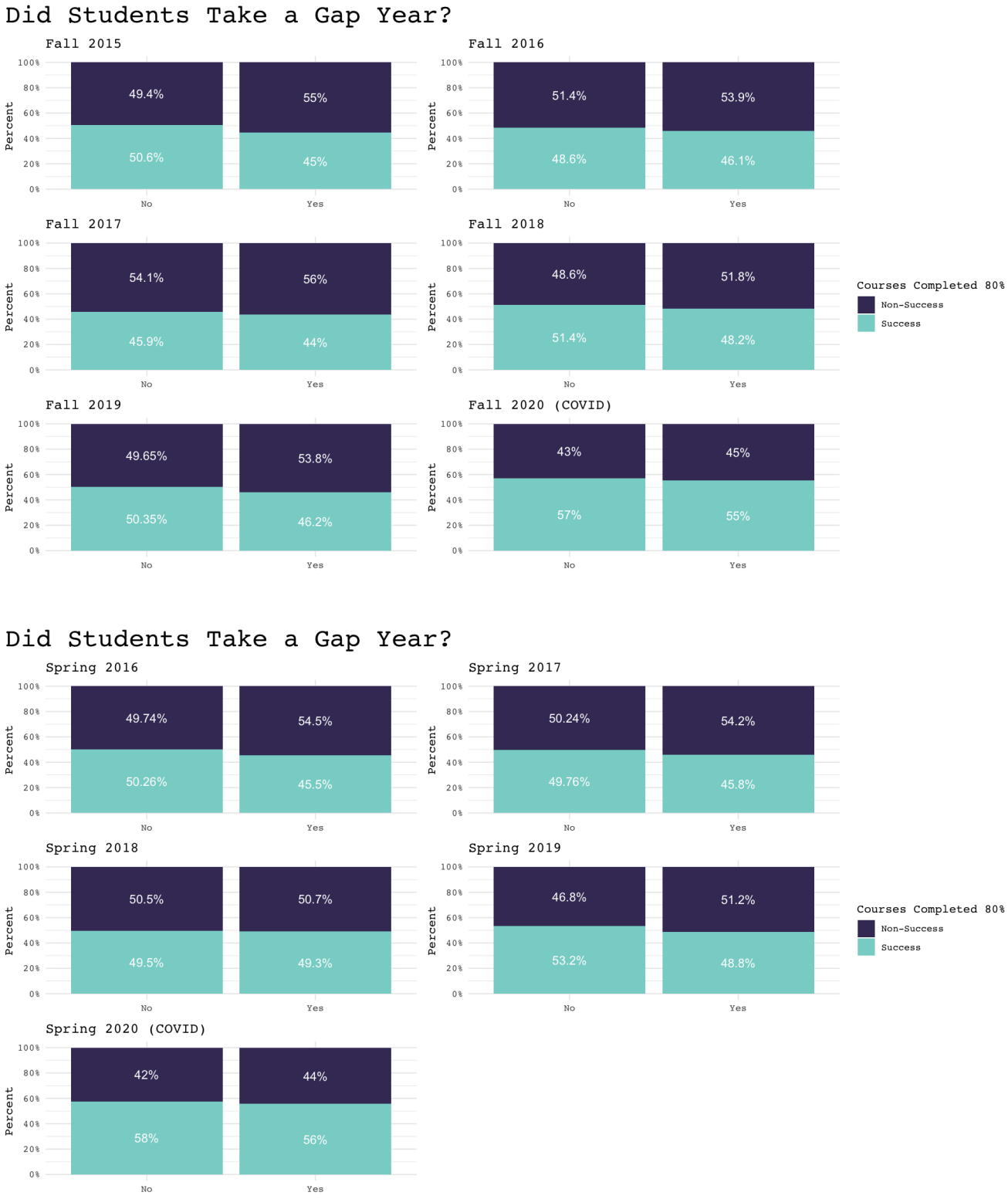

Figure 23: Academic Status (Completed 80 Percent or More Courses) in Relation to Taking a Gap Year

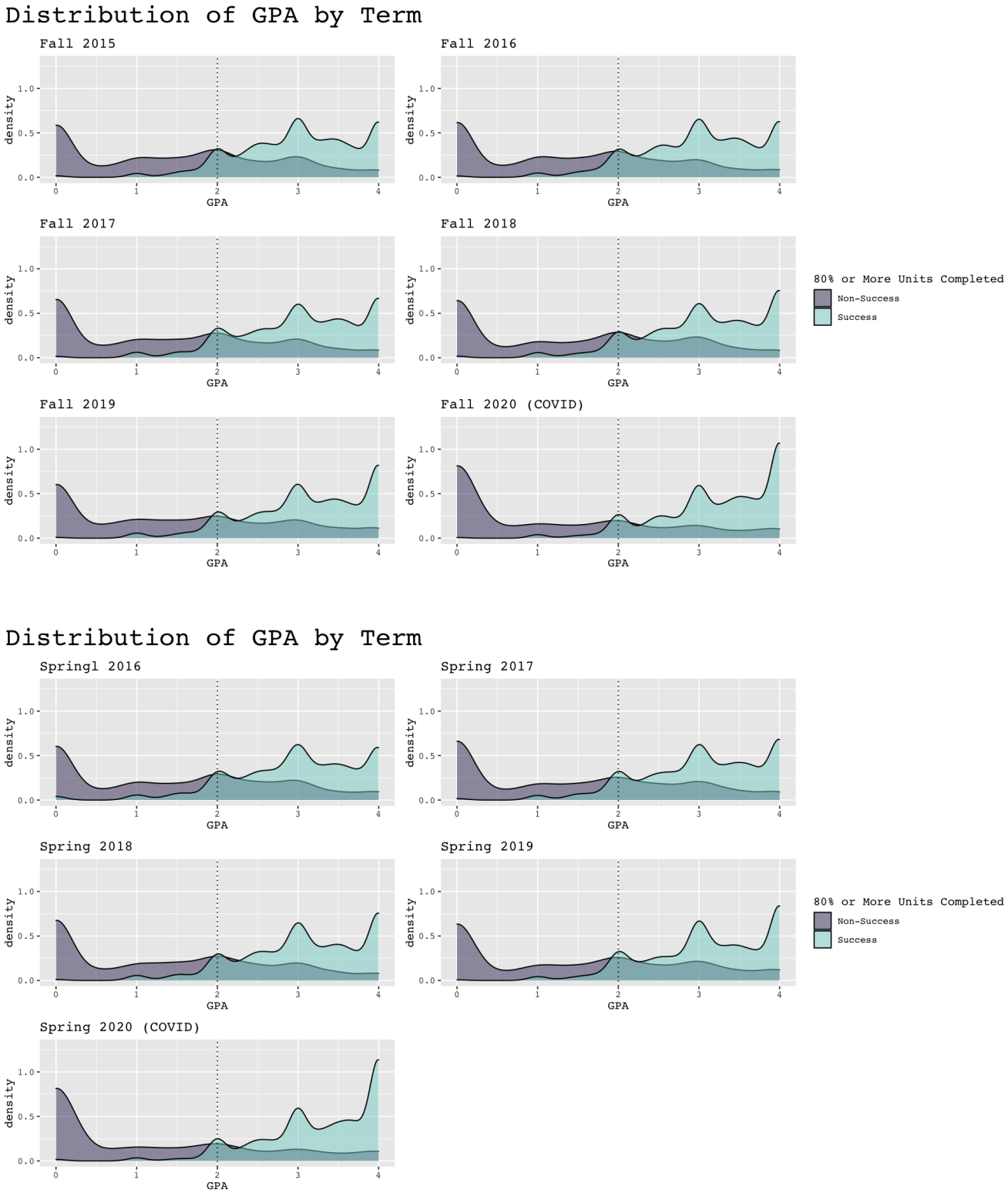

Figure 24: Academic Status (Completed 80 Percent or More Units) in Relation to GPA