The (Un)Dropout: Exploring Factors Influencing Completion Rates at 4-year US Institutions

Jennifer Pham

Executive Summary

Higher education is an important part of our society due to its promises of social mobility, financial security, and improved quality of life. We can measure the quality of an institution by various metrics such as faculty-to-student ratio, average accumulated debt, or completion rates. Completion rate is defined as the percentage of students who start a degree at a particular institution and are able to complete it. This project seeks to explore the question: what factors are associated with completion rates for four year colleges in the US?

To further explore this question, this report utilizes the College Scorecard. The College Scorecard is a widely accessible data set from the US Department of Education that contains institutional level data for all colleges in the United States and US territories. The College Scorecard has been collected every year since 1996. This report focus on data from 2013-2014. The provided data set contains 7,869 institutions and over 100 institutional level variables. Since completion rates is time dependent, this report focuses on four-year institutions. Though the College Scorecard contains many variables, I selected variables that I believed would be relevant in my analysis. The final data set utilized in this report contains ~2100 observations and 40 variables. Some limitations to this analysis may be the quality of the overall data set. There have been some critiques on the utility of the College Scorecard such as this data set should not necessarily be used to inform individual decisions. Rather, this data set can be informative in helping us understand macro-level trends within the college landscape.

The first part of this project is an exploratory data analysis of the completion rate variable. The data set consists primarily of continuous variables and a few categorical variables. Some continuous variables include cost of institution, student demographics, and undergraduate size. Some categorical variables include state, region, and control. To explore the completion rate variables, I used a variety of methods to visualize some variables. Some of the main methods I used were maps, box plots, and scatter plots. Maps allowed me to explore the state variable against continuous variables. Box plots allowed me to see the relationship between a categorical variable and a continuous variable such as cost of institution by control. Scatter

plots allowed me to observe the relationship between two continuous variables. Some initial observations I observed while exploring the completion rate variable included faculty pay, retention rates, and family income are positively associated with completion rates. On the other hand completion rate and the percentage of first generation college students are negatively associated. Based on initial observations from my exploratory data analysis and my own domain knowledge, I hypothesized higher completion rates are associated with schools that are not for-profit privates, pay their faculty more, have higher retention rates, have a lower percentage of first generation students and have higher family income

To further explore my primary hypothesis, I used linear regression to understand the effects that the different variables in the data set have on completion rates. I first split my data into two, where the first half would be to perform model selection and the second half would be to assess the effects the variables have on completion rate. My selected model was from using backward stepwise regression with interaction terms.

Some of the variables that have an impact in understanding the completion rate variable include undergraduate population, percentage of Hispanic undergraduates, percentage of NHPI undergraduates, faculty salary, retention rate, family income, and percentage of women. Since interaction terms were also included in the model, these interaction should be thought of in relationship to the variable standing alone. For instance, there is a negative effect on completion rates from the first generation and Pell Grant variables but are mitigated with a higher rates of both. Cost has an effect on completion rate but is more pronounced for public and private for-profit institutions. These results support my hypothesis that institution control, retention rate, faculty pay, percentage of first and generation students impact completion rates to some degree.

This analysis shows us what the larger college completion landscape looks like. This is an observational study that looks at the association between variables and completion rates, thus causal statements cannot be made. However, the relationships uncovered can point to possible smaller intervention points such as for first generation and low income students or financial aid programs that can improve student learning opportunties.

Introduction

College is an important aspect of society due to its promises of social mobility, financial security, and improved quality of life. Though

college is often emphasized to young people, conversations about about the actual completion of college is rarely talked about. In a 2008 Pew Research study, it found that of 100 high school graduates only 30 of them will complete college (Beach et al). Why is completion rate so low?

College completion can be defined as the proportion of first-time college enrollees that are able to finish their degree within a certain amount of time. This project seeks to answer the question: What factors are associated with completion rates across higher education institutions in the US? My hypothesis is that some of the factors that influence completion rate include family income, percentage of first generation students, retention rate, and whether an institution is private for-profit.

My work builds off of existing literature that explores completion rates. Though the discussion of completion rates is not new (Hanover Research), many sources often discuss completion rates in relationship to demographics of the individual students, such as by race, gender, and socioeconomic status. This at times implies deficit models that portray student identities as being the driving factor behind low completion rates, rather than effects of the environment on student achievement. This report seeks to be mindful of this distinction by recognizing the differences between institutional level data and student identities.

The data set that is used in this report is the College Scorecard. The data provided by the College Scorecard was collected through surveys conducted by the US Department of Education. It contains institutional level data from the 8000+ institutions across the US and US territories. Though it may be hard to recreate this data set, this data is easily accessible if one wants to recreate this analysis.

The goal of this report is to understand the factors that are associated with completion rate. First, I examined and explored the data set used in this analysis, the College Scorecard. Next, I took a deeper dive into the variable of interest, completion rate, and perform an initial exploratory data analysis which can help contextualize our hypothesis. Then, I focused on formalizing initial observations on completion rates by performing a multiple linearregression analysis. Last, I discussed the results and further implications for this work.

About the Data

The data set be utilized in this report is the US Department's College Scorecard. The College Scorecard contains aggregate data for institutions across the United States and US Territories such as enrollment, costs, student demographics and financial outcomes. It contains over

100 variables and is representative of over 8000 unique schools in the US and US territories. The College Scorecard is collected every year, and this report focuses on the 2013-2014 academic year. The data is collected by the Integrated Postsecondary Education Data System, the National Students Loan Data System, and the Department of the Treasury, so there should not be any issues regarding the collection processes or accuracy of the data.

The analysis is focused on 4-year institutions that primarily provide their instruction on a physical campus. This results in 2121 unique institutions. In addition, the College Scorecard provides hundreds of variables ranging from student enrollment, faculty salary, and post-graduate outcomes. Many of the variables are unusable (often due to discontinuation of collecting these data points), so I went through and selected ~40 variables that I believed were important to the analysis. In addition to the variables from the College Scorecard, I introduced a HBCU variable that indicates whether an institution is an HBCU or not. HBCU or historically Black colleges and universities are colleges that were founded before 1964 with the intention on serving Black students.

This data set used in this analysis contains 41 columns, 1 of which is an identifier column (name of institution) with 2121 observations. By taking a subset of the College Scorecard, I introduced some bias as I might have not included variables that are pertinent to the overall analysis. Some of the variables included in the data set include demographic, geographic, admissions, and outcome related variables. Some demographic data include undergraduate enrollment, percentage of students by race, percent of Pell Grant recipients, percentage of students above 25, and percentage of students with a loan. Some geographic variables include the state and region an institution is in. Some cost related variables include in-state/out-of-state tuition, median earning, median debt, faculty monthly salary, and net price. Some outcome related variables include 150% time completion rate, retention rates for part-time/full-time students, median-debt, etc. To see full data dictionary, see Appendix 1.

Most variables have a good amount of the data provided. For the purposes of the exploratory data analysis I did not remove missing values. Instead, for the variables I am plotting, I removed the rows with missing data for a desired variable. Many of the missing values are associated with the private for-profit institutions, such as SAT scores. This could be due to the fact that these institutions operate different and may not require or collect this data.

The goal of this report is to examine completion rate in relation to some of the aforementioned variables. Though the College Scorecard provides a completion rate variable, this is reflective of 6 years

prior, i.e. 2013 data is reflective of the 2007 entering cohort. Various papers have utilized institutional level data from the entering cohort to model completion/retention metrics, but I have also seen analysis done on the same year. For simplicity and to avoid complications with merging with multiple data sets, this report uses the completion metric from the College Scorecard for the year 2013.

There have been some critiques (The Institute for College Access and Success) on the implications of the College Scorecard such as it not being useful for an individual student to decide what college to attend. Rather, the College Scorecard should be used as a proxy in understanding what the current college landscape looks like from a macro perspective.

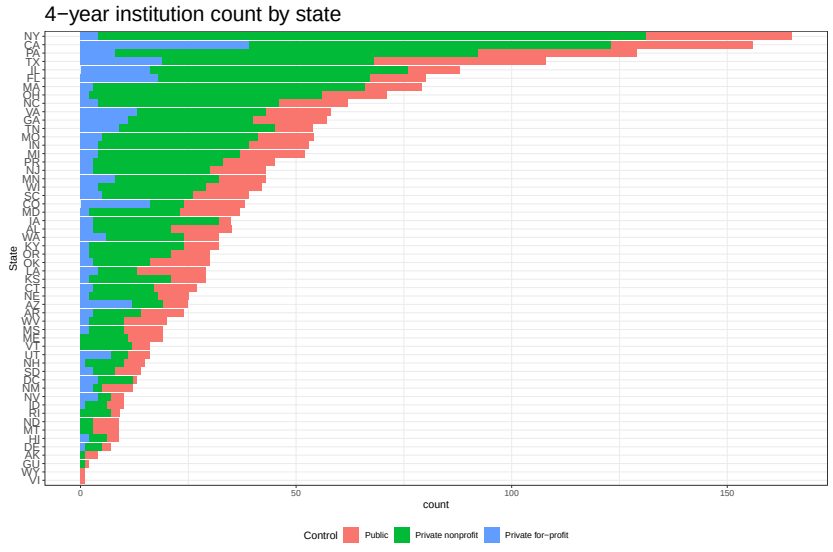
Exploratory Data Analysis

The College Landscape: At a Glance

In order to understand trends within college completion, we need to grasp the college landscape. We can begin by looking at some of the categorical variables in our data set – Control, State, and Region.

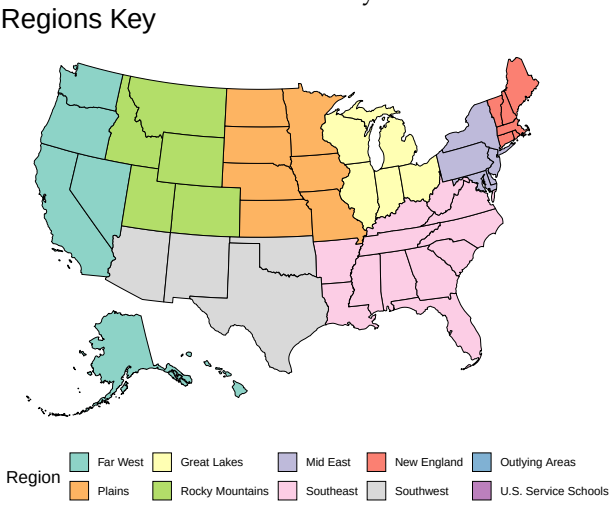
Control refers to how the institution operates and is funded, whether that be public, private nonprofit, or private for profit. This data set contains 1260 private nonprofit, 586 public, and 275 private for-profit. Control is one of the most distinct features of an institution, and is used throughout this report as a way to compare various metrics by subgroup.

How are institution distributed by state? What does the control look like by state?



This graph shows the count of institutions in each state by control. We notice the distribution of institions vary greatly. States like New York and California have over 150 four year institutions, while states/territories like Wyoming and Virgin Islands have one institution. This can be reflective of population sizes as California and New York are two of the more populous states in the US. In addition this plot also shows the distribution of Control by each state. We notice New York has the most four year institutions, but of these institutions, it has less private for-profit than many states with less institutions including California, Pennsylvania, Texas, Arizona, and others.

This data set contains a "Regions" variable which categories the states into larger geographic groupings. Since regions is based off the state variable, this report does not look at the relationship between region and state to avoid redundancy.



This map is a key to distinguish which states belong to which regions. US Service Schools and Outlying Areas do not show up on the map as distinct regions. US Service Schools include United States Air Force Academy, United States Coast Guard Academy, United States Naval Academy, United States Merchant Marine Academy, and United States Military Academy. Outlying Areas consists of Samoa, Federated States of Micronesia, Guam, Marshall Islands, Puerto Rico, Palau, and Virgin Islands. Geographic region is more than a grouping of states, but can show geographic clustering by key variables about language, gender, or wealth. For instance, the geographic distribution of race in the US is not equal. An example can be seen in Figure 1.

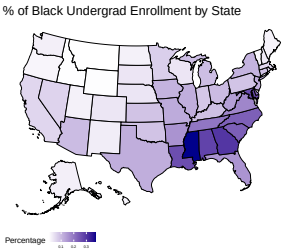
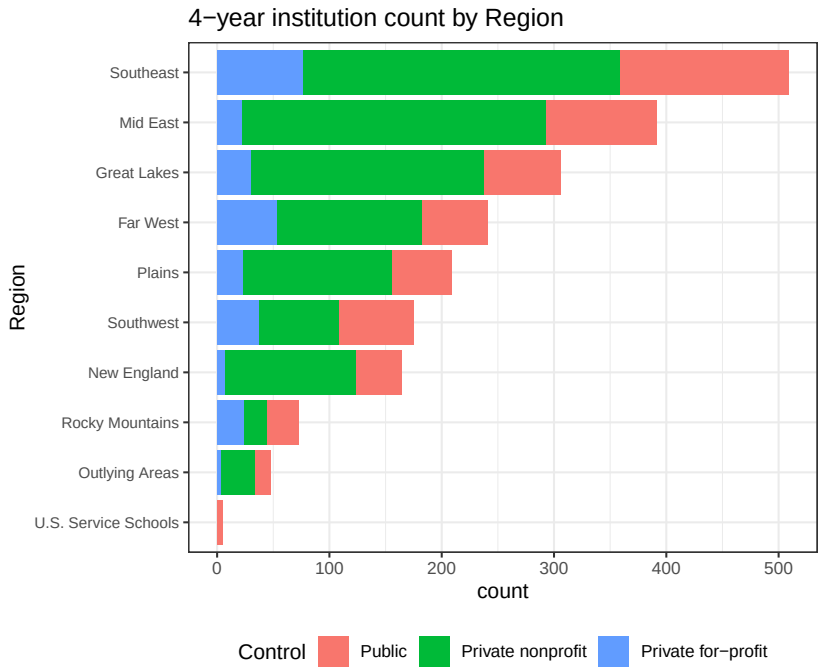


Figure 1: This can be due to historical events, urban sprawl, and immgration patterns. Take for instance the distribution of Black undergraduates. The highest concertration is in the Southeast which is motivated by US's history of slavery. Relationship between race and location was not deeply discussed in this report, but is important to consider as race and location are not isolated from each other. A further analysis is provided in Appendix 2.

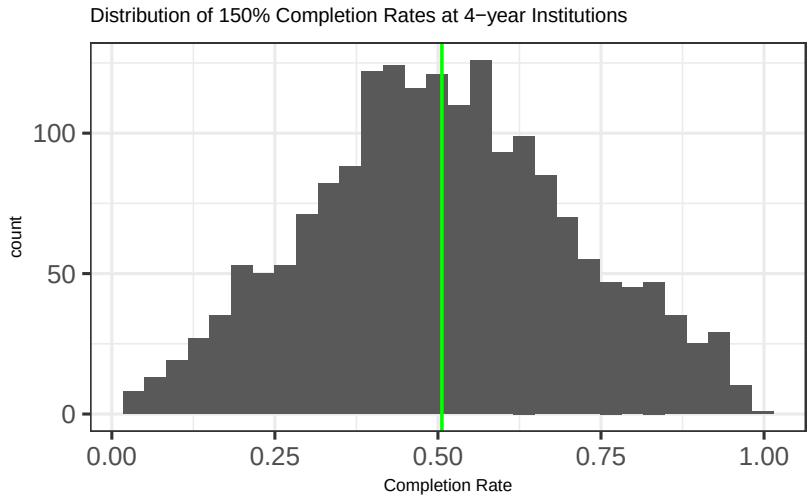


Though our previous plot that indicates New York as the state with the most four year institutions, this graph shows a majority of institutions are located in the Southeast. Most regions have a majority private nonprofit schools with exceptions in the Rocky Mountains and U.S. Service Schools.

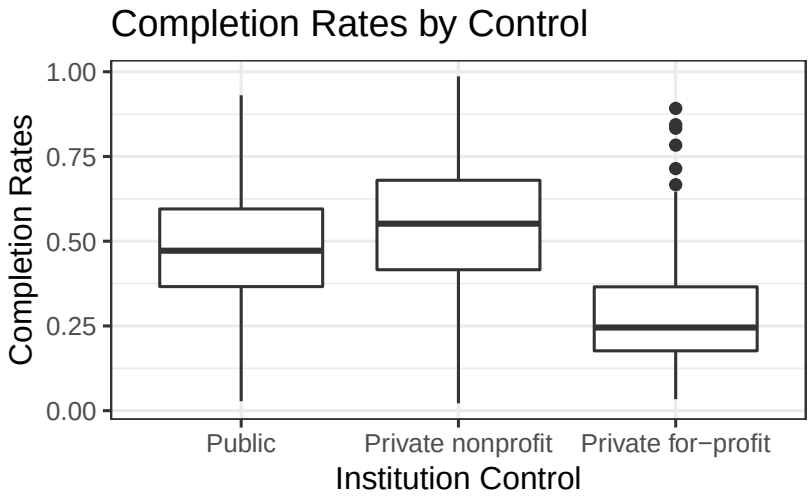
Looking at Completion Rates

The focus of this report is to better understand completion rates. This section is an exploratory analysis of the completion rate variable and other variables in relationship to it. The College Scorecard defines completion rates as "The proportion of full-time, first-time, degree/certificate-seeking undergraduates who completed a degree or certificate at the institution within 150 percent of normal time." Since the focus of this analysis is four year institutions, completion rate is defined as finishing a program in six years.

In our data set, the completion rate variable had outlying values of 0% and 100%, which were removed since it is improbable that an institution have 0% or 100% completion rates (Resulting in a data set with 1859 institutions). Upon further investigation, some of these institutions are no longer operating or do not seem to be reliable. Some of the issues could be due to utilizing data from nearly a decade ago which could be addressed by using more recent data. This issue in outlying values can point to potential issues with the College Scorecard's reliability or even the institution need to provide accurate data.

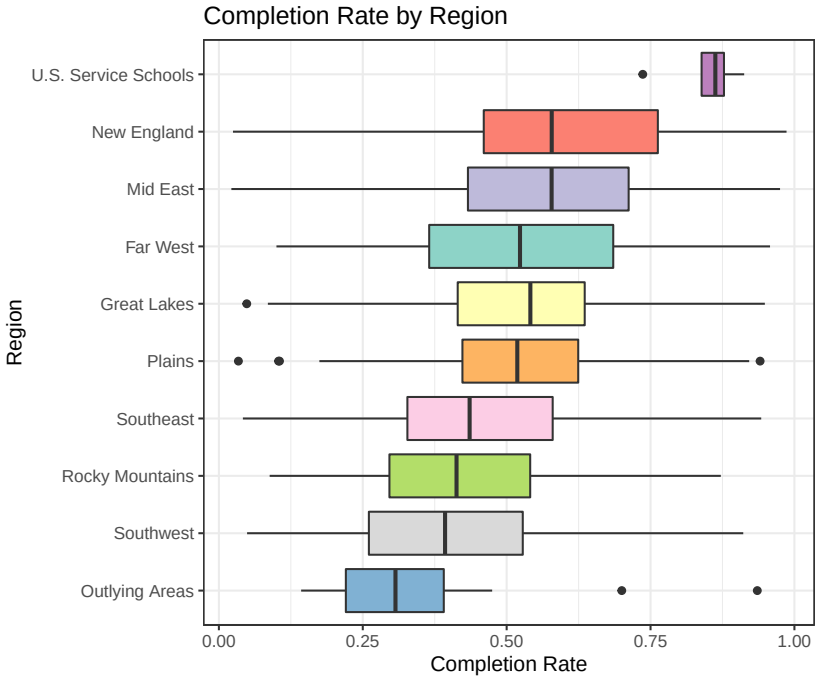


This histogram shows the distributions of completion rates for the institutions in the data set. Completion rates follow a roughly normal distribution that is centered around the mean which is just over 50% (50.6%). The median of the completion rate variable is also around 50%. Though this begins to tell us the story, do completion rates change as we look across different subgroups?

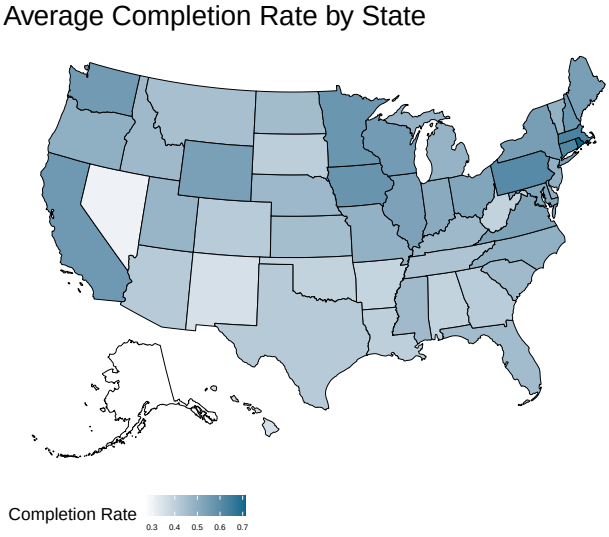


When we take a look at completion rate by control we notice differences between different control types. This box plot shows the distribution of completion by public, private nonprofit, and private for-profit. We notice the median for public and private nonprofit hovers just below and above 50%. The whiskers of the public and private nonprofit are fairly similar as they range from both low and high completion rates. On the other hand, private for-profit institutions have a substantially lower median at about 25%, with outlying values on the higher end. This can show that private for profit schools

generally have lower completion rates and potentially show some initial differences in completion rates between nonprofit schools and for-profit schools. School control can be an indicator on student completion rates, but can differences in external factors such as location show us differences between completion rates?



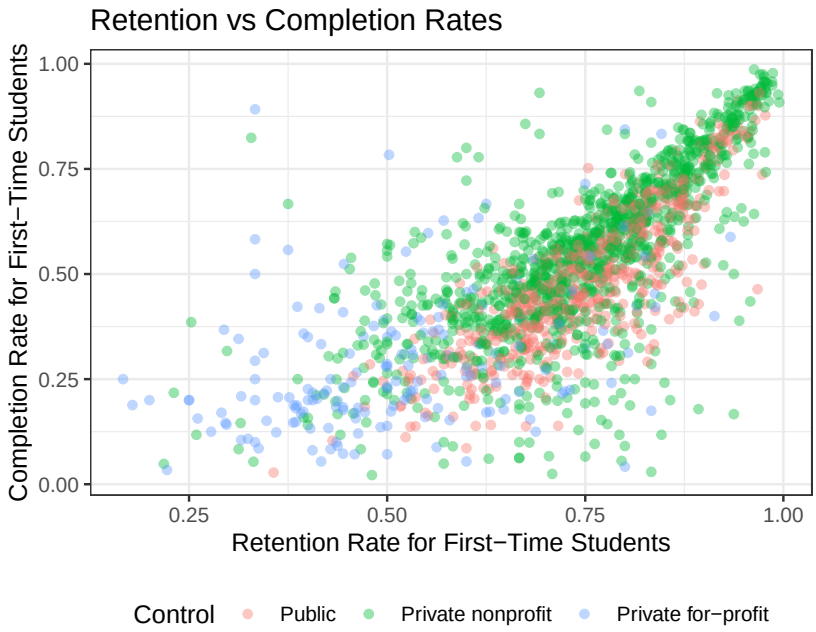
Taking a look at completion rate across the 10 regions provided in this data set, there are some obvious differences in completion rates. This box plot shows that regions such as the Southeast, Rocky Mountains, Southwest, and Outlying Areas have medians below the population median. On the other hand regions such as New England and Mid East have higher completion rates with very comparable medians. Since New England and Mid East are geographically close in proximity to each other, this could potentially point to a geographic trend in completion rate in these regions (though further testing may needed to further show this). Though this plot shows us macro trends of completion rate in a region, this glosses over what is happening on the institutional or even state level.



Further, to examine completion rates by geographic location, this map paints a different, more specific picture on completion rates. This map shows us the average completion rate by state. Already we notice geographic differences across regions despite the observations we made in the box plot. Take the Far West region consisting of states such as Washington, Oregon, Nevada, California, Alaska, and Hawaii, for instance. Even though it has the 4th highest median completion rate, as seen by the previous plot, this region contains states such as Alaska and Nevada which have two of the lower average completion rates in the United States. Initial differences across regions and states could possibly point to an association between location and completion rates.

Retention Rate – 1 year "Completion Rate"

This analysis defines completion rates as completion of a four year program in at most six years. To further explore completion rate, this section looks at a closely related metric: retention rate. Retention rate is defined as the proportion of first-year students who complete their first year of their program. Completion is dependent on retention since students cannot complete six years without finishing their first.



This scatter plot shows us there is a positive linear relationship between retention rate and completion rates (r=0.74). The association is particularly stronger for schools with retention rates greater than 60%. We notice the approximate slope of the line is flatter, indicating that retention rates are the upper bound for completion rates which makes sense since students need to persist through their first year in order to complete college. Further exploration of retention rates can be seen in Figure 2.

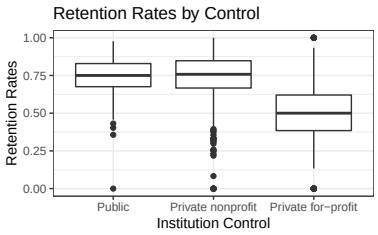
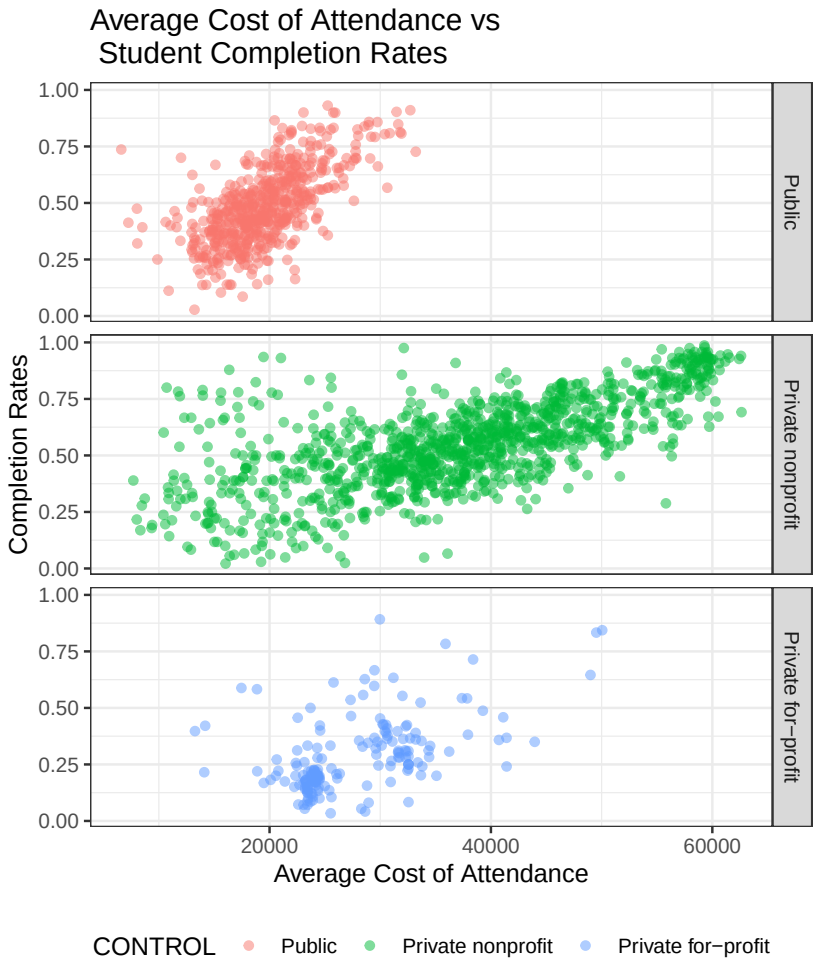


Figure 2: This box plot shows the distribution of retetion rates by control. Similar to completion rates, there is a stark difference between public/private nonprofit and private for-profit institutions. This could point to similar underylying mechanisms at nonprofit and for-profit institutions that have an affect on student persistance and completion.

Institutional Level Variables

Though completion rate is an based on an individual metric, whether or not a student completes their program, how do larger institutional factors, many of which are out of a student's control impact completion rates? This section examines the relationship between cost of attendance and faculty pay to completion rates.



This plot shows the relationship between cost and completion faceted by control. We notice that average cost of attendance is generally associated with higher completion rates across all three controls. There appears to be a stronger relationship between cost of attendance and completion rates for public and private nonprofit schools. Since public schools are generally cheaper than private schools as seen in Figure 3, a change in cost corresponds to a greater change in completion rate than in the private nonprofit case. This overall positive association between cost and completion could point to more pressure to complete school at more costly institutions to ensure students are getting their money's worth.

Further, completion rates are generally motivated by other factors such as the resources a student has available at their given institution. Some of these resources can include course selection, program offerings, and quality of teaching. Though we are limited by the variables available by the College Scorecard, teacher pay has been discussed alongside quality of instruction (Will).

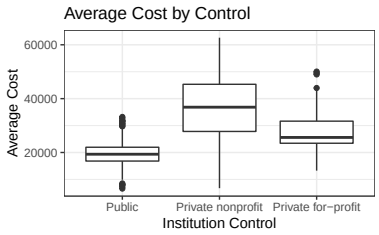
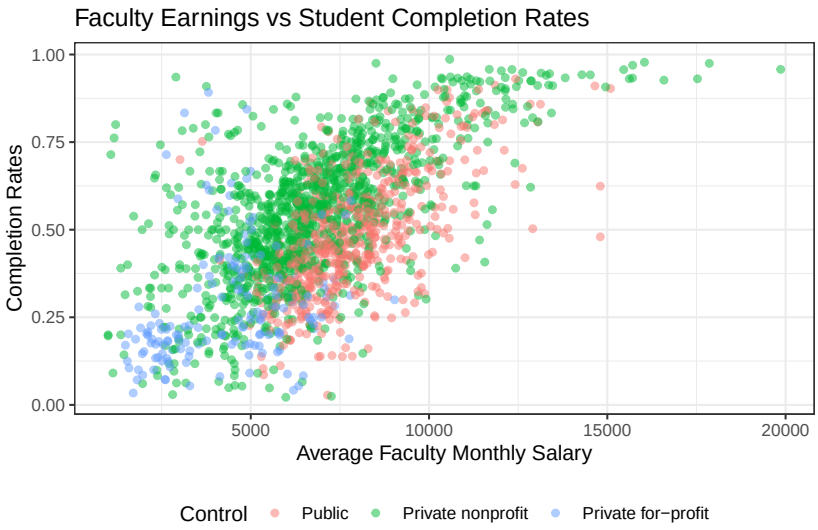


Figure 3: This box plot shows the distribution of average cost by control. There is a stark difference between public and private institutions. This makes sense as public schools are positioned to be the more accessible and affordable higher education option, thus should cost less than private schools on average.



This plot examines the relationship between average faculty monthly salary and completion rate. As observed, there is a generally positive relationship between faculty salary and completion rate. We also can observe a bit of clustering for private for-profit schools that are associated with lower faculty pay and lower completion rates. Though faculty pay is not indicative or associated with better teaching, this can lead to further conversations of mechanisms driving better quality of instruction which can lead to better student learning experiences and completion rates.

Student Demographic Related Variables

This section looks at more student-level variables in relationship to completion rates. Though much of literature points to student demographics as a motivating factor behind completion rate, it is important to keep in mind that this often points to a deficit mindset when it comes to student achievement. For instance, if there are more students who hold x identity, which is associated with lower completion rate, this does not necessarily mean there is something "wrong" with x students. This can be a good starting point to thinking about other other structural barriers at play and institutional level resources that can be provided to meet the needs of students.

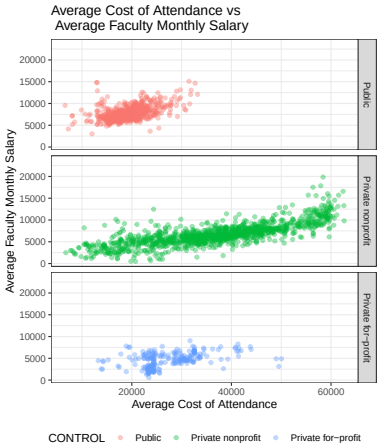
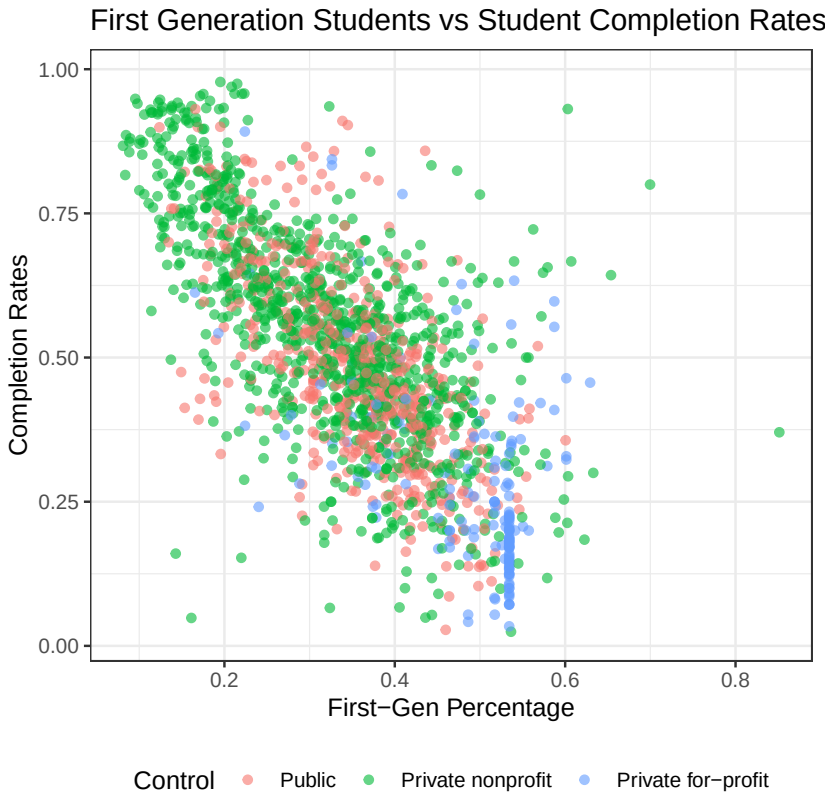
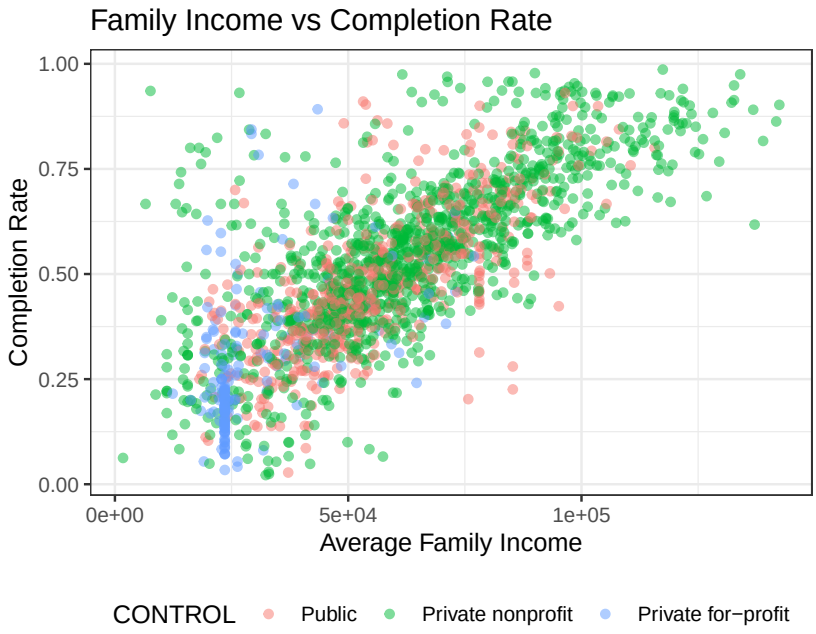


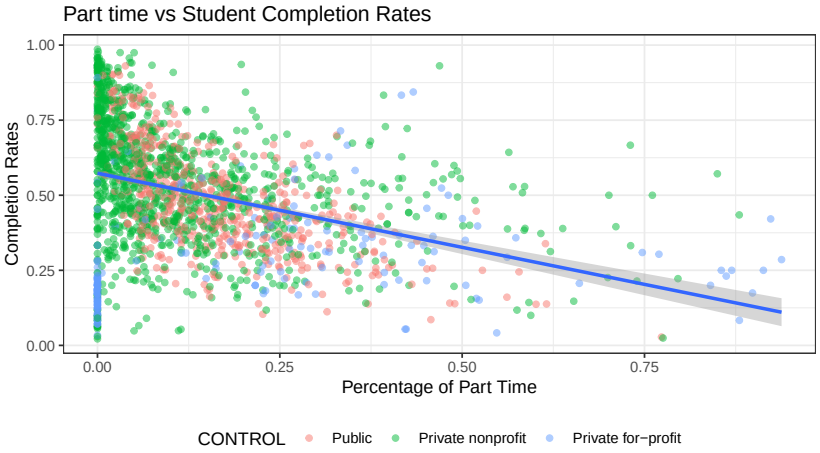
Figure 4: This plot shows the relationship between cost of attendance and average montly faculty pay motivated from observations seen in the last two plots. Cost and pay are positively associated with compltion rate, are cost and pay associated with each other? To some degree, yes. Across all three controls there is a positive association between cost and pay, with the most evident in private nonprofit schools.



First-generation college students are students who neither parent completed college. First generation college students face unique challenges compared to their non-first generation counter parts such as navigating the college system without the guidance from their parents. This plot shows the relationship between the percentage of first generation students and completion rates. As observed there is a negative relationship between these two variables. We also notice most schools have a first generation minority, as indicated by the majority points to the left of the 50% mark. This association can point to first generation students facing challenges with completion as they may not have the resources some of their peers have.



Another metric that can possibly impact completion rate is wealth. Ths scatterplot demonstrates the relationship between average family income and completion rate. We notice that there is positive relationship between family income and completion rates. By using family income of an institution as a proxy for the general wealth of the student body, this plot shows us that schools with wealthier families tend to have higher completion rates.



Part-time students are students who enroll at an institution on a reduced course load. Due to their reduced course load, it is probably that it takes more time for these students to complete their degree. This plot shows the relationship between percentage of part time students and completion rates. Many institutions appear to have very little part time students. There is some negative association between

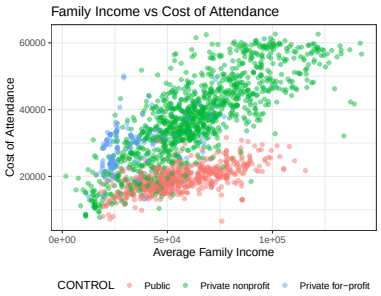


Figure 5: Though as seen in a previous plot between cost and completion, perhaps there is also some association between the types of schools wealthier families attend. This plot shows a positive association between family income and cost. This suggests that more costly schools may attract wealthier families, and is particularly more pronounced at private non profit schools compared to public schools.

part time students, but is fairly weak. Though on individual level, part-time students tend to take longer to complete their program than their full-time counterparts, having more full-time students may mitigate this effect.

Methods

Based on the exploratory data analysis in the previous section and my previous domain knowledge, I hypothesize that higher completion rates are associated with schools that are not for-profit privates, pay their faculty more, have higher retention rates, have a lower percentage of first generation students and have higher family income. To test this hypothesize and to further discover what factors influence completion rate, I used a multiple regression model.

For our purposes performing the regression, there cannot be any missing values in our data set. Any identifying information such as school name and city was removed since that is not be necessary for the model building. The state variable was also removed since in some cases there was only 1-2 schools in a given state which would have been challenging to build our model. Some of the variables that were included for the purposes of the exploratory data analysis but have a substantial amount of missing data, such as SAT scores, income, debt, etc. were removed from the data set. In addition, the average in-state and tuition variables were removed because it is collinear with in-state for non-public schools.

The data set used to perform this analysis consists of 1772 observations representative of all 3 controls (1005 Private nonprofit, 549 Public, 168 Private nonprofit) and representative of 9 out of the 10 original regions[1].

[1] Since US Service schools originally only consisted of 4 schools, there was not a large sample size to begin with. These four schools were missing data in variables necessary for the analysis

The data was first split in half. The first half of the data was used to perform model selection. To determine the best model, each model was evaluated based on its R-Squared value and 10-fold Cross Validation Test Error. The models that were tested with their corresponding R-Squared value and 10-fold Cross Validation Test Error are shown in the table below.

	Adjusted R-Squared	10-fold Cross Validation
Full Model	0.7681	0.0972127
Full Model Forward	0.7675	0.0950073
Full Model Backward	0.7686	0.0952015
Full Model w/ Interaction	0.7953	0.0934800

	Adjusted R-Squared	10-fold Cross Validation
Full Model w/ Interaction Forward	0.7748	0.0938188
Full Model w/ Interaction Backward	0.7940	0.0904593

As seen by the table, the backward step wise model with the interaction terms resulted in the highest adjust R squared value of 0.794 and lowest CV error of 0.0904593.

The interaction terms that was included in the 4th, 5th, and 6th model were *Percentage of First Generation Students x Percentage of Pell Grant students*, *Percentage Part-Time and Percentage of Students Over the 25 y/o*, and *Cost and Control* The first interaction term was based off of some domain knowledge that many first generation students are low-income, thus qualify for the Pell Grant.[2] In addition, the terms 'first-generation' and 'low-income' have often been grouped together to reflect common overlapping experiences these student may have. The second interaction term was included based on an article (Marcus) that that mentions older students, many of whom are parents, are also working. As a result, they enroll on a part time basis. The third interaction term was included based off the notion that public schools are positioned as the most cost effective option compared to private schools, thus there might be some effect control has on cost.

[2] Though these terms are not indicative of each other. A student can be first generation and not low-income. Similar a student can be low-income and not first generation.

Results

Variable	Estimate	Std. Error	p-value	Sig Level
(Intercept)	-0.0481	0.074256	0.517313	
Control - Private nonprofit	0.253688	0.033562	1.07E-13	***
Control - Private for-profit	-0.01511	0.068248	0.824825	
Undergraduate Population	1.76E-06	5.82E-07	0.002558	**
% Black Undergrad	-0.05384	0.032558	0.098566	
% Hispanic Undergrad	-0.06652	0.025112	0.008231	**
% Asian Undergrad	0.045338	0.077509	0.558748	
% AIAN Undergrad	-0.15127	0.095759	0.114545	
% NHPI Undergrad	0.343397	0.096852	0.000414	***
% Part-time Undergrad	-0.09675	0.062356	0.121132	
Average Faculty Salary	7.1E-06	2.73E-06	0.009347	**
% Pell Recipient	-0.35097	0.085722	4.65E-05	***
Retention Rate	0.553243	0.038594	7.92E-42	***
% Undergrad > 25y.o.	-0.07653	0.031049	0.013907	*
Net Tuition Revenue	-1.5E-06	1.08E-06	0.160908	
Tuition In-state	-1.3E-06	1.82E-06	0.47523	
Cost of Attendance	1.33E-05	2.11E-06	5.09E-10	***
% First Gen	-0.57164	0.098644	9.68E-09	***
Average Family Income	8.54E-07	4.14E-07	0.039635	*
% Undergrad Women	0.067965	0.025618	0.00813	**
HBCU1	-0.02549	0.027071	0.346603	
Net Price	-1.8E-06	1.14E-06	0.108272	
% Pell Recipient:% First Gen	0.837149	0.173574	1.68E-06	***
% Part-time Undergrad:% Undergrad > 25 y.o.	0.118484	0.092196	0.199101	
Control - Private nonprofit:Cost of Attendance	-1E-05	1.63E-06	1.37E-09	***
Control - Private for-profit:Cost of Attendance	-3E-07	2.57E-06	0.907101	
Adjusted R-squared	0.7958			

The backward step wise model with interaction terms resulted in an adjusted R-squared of 0.7958 which means the variables included in this models accounts for 79.58% of the variation in the completion rate variable. From the proposed hypothesis, we notice variables such retention rate, faculty salary, and family income are statistically significant meaning the variables have an effect on completion rate.

In the case of retention rate, if we were to keep everything constant, a one percentage point increase in retention rate would correspond to a .55 percentage point increase in completion rate. Similarly, in the case for faculty salary, an increase in salary of $10,000 would correspond to a .07 percentage point increase.

Our hypothesis also looks at the effects of first generation students and completion rates. This model incorporates an interaction term

between first generation students and Pell grants. The effect of more first generation students is mitigated by more Pell Grant recipients. For instance, holding all other variables constant, consider an institution that has 50% first generation students (which is on the upper-bound of first gen students). $Complete = -.35Pell + .84(Pell \times .5) = -.35 + .42(Pell) = .7Pell$. Even though the first generation coefficient has a value of -.57, some of its effects can be mitigated by the Pell Grant variable. So a 1 percentage point increase in Pell Grant recipients corresponds with an .7 percentage point increase in completion rate. Thus the percentage first generation variable does have an effect on completion rate but may be partially mitigated by a higher percentage of Pell Grant recipients.

In the exploration of the completion rate variable we noticed some association between cost and completion rate. We can formalize this by looking at the coefficients. Since this model includes the interaction term between control and cost, we can evaluate the effect of cost by control on completion rate. In the case for public schools, an increase in cost of $10,000 corresponds to a .132 percentage point increase in completion rate. Similarly, since the coefficient for interaction between private for-profit and cost is nearly zero (-3e-7), cost has a similar effect on completion rate in the public and private for-profit case, corresponding to a ~.13 percentage point increase. On the other hand, the coefficient on the interaction term for private nonprofit is -1e-5. This results in the overall effect of a $10,000 increase in cost to be an increase in a .03 percentage point. Thus the effect of cost on public and private for-profit school completion rate is much greater than the effect on cost on private nonprofit completion rate.

In addition, variables such as percentage of Native Hawaiian/Pacific Islander and percentage of women are also statistically significant. However, this can point to the other underlying models and factors driving them to be important variables in the model. For instance, generally schools in the continental US have low percentage of NHPI students, however these students may make up the majority at institutions in Hawaii or Guam. As a result the effect of NHPI may only be applicable to certain institutions, and is not necessarily generalizable for the average US institution. Further exploration of institutions with high number of NHPI students can be helpful in understanding the true effects of this variable.

Conclusion

Ethical Implications

The College Scorecard is the first step in higher education accountability. The Actionable Intelligence for Social Policy classifies data initiatives on a scale from low to high risk and low to high benefit for society. The College Score can possibly falls into a moderate-risk higher-benefit category. Since this level is aggregate data from all universities, it is reflective of the undergraduate population, which does not disclose much about individuals identities. There is relatively low risk in having this data set publicly available, as there have been practices to anonymize sensitive data. However, one potential risk that has been pointed out of the College Scorecard should be used to observe larger macro trends rather than influencing individual college-going decisions. Much of this data is collected on the administrative level such as asking a student their race, gender, or family income, and may not be suited to make larger claims about students of a certain background. Thus, in this analysis it is important to talk about associations between variables rather than implying causality which can often point to deficit models in education achievement.

Two important factors within the data life cycle include data collection and data access.

Data collection refers to the process of gathering information. In the context of the higher education, it is also important to consider data that is not collected. Often many efforts, including those of my own, fail to recognize the nuances of race in their data collection practices. In an article by Pilar Diaz in honor of Asian Pacific American Heritage Month, Diaz discusses the difficulties of collecting data for Native Hawaiian/Pacific Islander folks. Often times they are left out of data collection process, data analysis[3], and ultimately policy decisions because of their small population size. However, this raises the even greater need to push for accurate data collection for these populations. Other data collection initiatives within higher education have looked at data disaggregation for Asian American communities.

The data included in this data set touches on many areas in college college access literature ranging from topics such as affirmative action, need-based aid, and test optional admissions. Higher education has been branded as this equalizing force, but from initial observations have shown to be differences among various subgroups. Though there have been many skeptics that continue to think higher education is accessible, having access to macro level college data allows for us to use data driven evidence that points to larger issues in higher education. Data access allows us quantify structural barriers within higher education systems that make it more difficult for students to enter, succeed, and complete college.

[3] When I first did my analysis, I considered not including Indigenous and Pacific Islander undergraduates from my analysis since they only made up a small percentage of the population, however recognize by practicing this "statistically insignificant" mentality I was actively contributing to the neglect of these identities in other parts of society (government representation, access to resources, media representation, etc)

Discussion & Limitations

Results from this analysis were able to support my initial hypothesis that factors such as control, family income, first-generation students, high retention rate, and high faculty salaries drive completion rate. The model selected was able to account for 79.5% of the variance in the completion rate variable. By utilizing observational data, no causal statements can be made. However we were able to see some hidden effects variables such as Pell Grant had on First Generation. This discovered relationships can be the beginning of further exploration, research, and intervention.

The results in this analysis are different than the results in a similar analysis done by Lu and Uzzi. The focus of their paper is similar, to evaluate the factors associated with graduation rates. Their approach was also to utilize a multiple regression model utilizing the College Scorecard but wanted to examine the effects of three specific variables: family income, Pell Grant, and SAT scores. Though my model uses the first two variables in my model, it does not include the SAT score. SAT score is an interesting variable that has been found to discretely encode information on student's identities and backgrounds. SAT scores have been shown to effectively predict college success, which may have partially been covered by variables such as retention rate in my model. Though their model cannot imply causation either, this can point to underlying mechanisms that may make students more persistent throughout college.

A limitation of this data was only being able to think about completion rate from a larger macro level. Possible next steps to rethinking completion rates are asking questions like "How do about specific institutional resources and quality of student experiences affect completion rate?" This question touches on more causal perspectives as the intervention of a program or individual (like a professor) can affect a student's trajectory. Since students cannot change whether they are first generation or low income, this analysis may not be of use to make large policy decisions but rather understanding the college landscape and the fact that various factors and influencing completion rates that are out of one's control. The next step is to think about what action and interventions can mitigate some of these effects, and how can we quantify the effectiveness of these interventions on completion rates.

References

- Actionable Intelligence for Social Policy. "A Toolkit for Centering Racial Equity Throughout Data Integration", `https://aisp.`

upenn.edu/centering-equity/
- Beach et al. “Pathways to Economic Mobility: Key Indicators”, https://www.pewtrusts.org/~/media/legacy/uploadedfiles/wwwpewtrustsorg/reports/economic_mobility/pewempchartbook12pdf.pdf
- Diaz, Pilar. “Beyond Just Data, Connecting with Native Hawaiian Pacific Islanders”, https://www.first5la.org/article/beyond-just-data-connecting-with-native-hawaiian-pacific-islanders/
- Hanover Research. “Retention and Graduation Rate Analysis”, https://www.clarion.edu/about-clarion/offices-and-administration/university-support-and-business/office-of-institutional-research/retention-and-graduation-rate-analysis-clarion-university.pdf
- Li, Lu and Mary Jane Uzzi. “An Evaluation of the Factors Influencing College Graduation Rate”, https://www.causeweb.org/usproc/sites/default/files/usclap/2020-1/An%20Evaluation%20of%20the%20Factors%20Influencing%20College%20Graduation%20Rate.pdf
- March, Jon. “Universities that are recruiting older students often leave them floundering”, https://hechingerreport.org/universities-that-are-recruiting-older-students-often-leave-them-floundering/
- The Hundred-Seven. “HBCU Listing”, http://www.thehundred-seven.org/hbculist.html
- The Institute for College Access and Success. “Takeaways from New Program-Level Data on the College Scorecard”, https://ticas.org/accountability/data-evidence-and-information/takeaways-from-new-program-level-data-on-the-college-scorecard/
- U.S. Department of Education. The College Scorecard, https://collegescorecard.ed.gov/data/
- Will, Madeline. “Higher Pay Leads to Smarter Teachers, Global Study Says”, https://www.edweek.org/teaching-learning/higher-pay-leads-to-smarter-teachers-global-study-says/2019/02

Appendix

Appendix 1: Data Dictionary

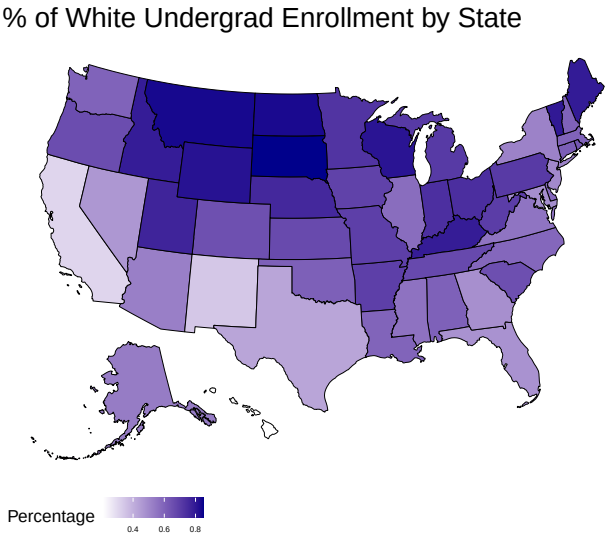
Variable Name	Definition		
INSTNM	Institution name	NPT4_PUB	Average net price for Title IV institutions (public institutions)
CITY	City	NPT4_PRIV	Average net price for Title IV institutions (private for-profit and nonprofit institutions)
STABBR	State postcode	AVGFACSAL	Average faculty salary
PREDDEG	Predominant undergraduate degree awarded	PCTPELL	Percentage of undergraduates who receive a Pell Grant
HIGHDEG	Highest degree awarded	C150_4	Completion rate for first-time, full-time students at four-year institutions (150% of expected time to completion)
CONTROL	Control of institution	RET_FT4	First-time, full-time student retention rate at four-year institutions
REGION	Region (IPEDS)	PCTFLOAN	Percent of all undergraduate students receiving a federal student loan
ADM_RATE	Admission rate	UG25ABV	Percentage of undergraduates aged 25 and above
ACTCMMID	Midpoint of the ACT cumulative score	GRAD_DEBT_MDN	The median debt for students who have completed
SAT_AVG	Average SAT equivalent score of students admitted	MD_EARN_WNE_P10	Median earnings of students working and not enrolled 10 years after entry
UGDS	Enrollment of undergraduate certificate/degree-seeking students	GT_25K_P10	Share of students earning over $25,000/year (threshold earnings) 10 years after entry
UGDS_WHITE	Total share of enrollment of undergraduate degree-seeking students who are white	TUITFTE	Net tuition revenue per full-time equivalent student
UGDS_BLACK	Total share of enrollment of undergraduate degree-seeking students who are black	TUITIONFEE_IN	In-state tuition and fees
UGDS_HISP	Total share of enrollment of undergraduate degree-seeking students who are Hispanic	TUITIONFEE_OUT	Out-of-state tuition and fees
UGDS_ASIAN	Total share of enrollment of undergraduate degree-seeking students who are Asian	COSTT4_A	Average cost of attendance (academic year institutions)
UGDS_AIAN	degree-seeking students who are American Indian/Alaska Native	FIRST_GEN	Share of first-generation students
UGDS_NHPI	degree-seeking students who are Native Hawaiian/Pacific Islander	FAMINC	Average family income in real 2015 dollars
UGDS_2MOR	degree-seeking students who are two or more races	UGDS_WOMEN	Total share of enrollment of undergraduate degree-seeking students who are women
UGDS_NRA	degree-seeking students who are non-resident aliens	ENDOWBEGIN	Value of school's endowment at the beginning of the fiscal year
UGDS_UNKN	Total share of enrollment of undergraduate degree-seeking students whose race is unknown	DISTANCEONLY	Flag for distance-education-only education
PPTUG_EF	Share of undergraduate, degree-/certificate-seeking students who are part-time	HBCU	Flag for HBCU

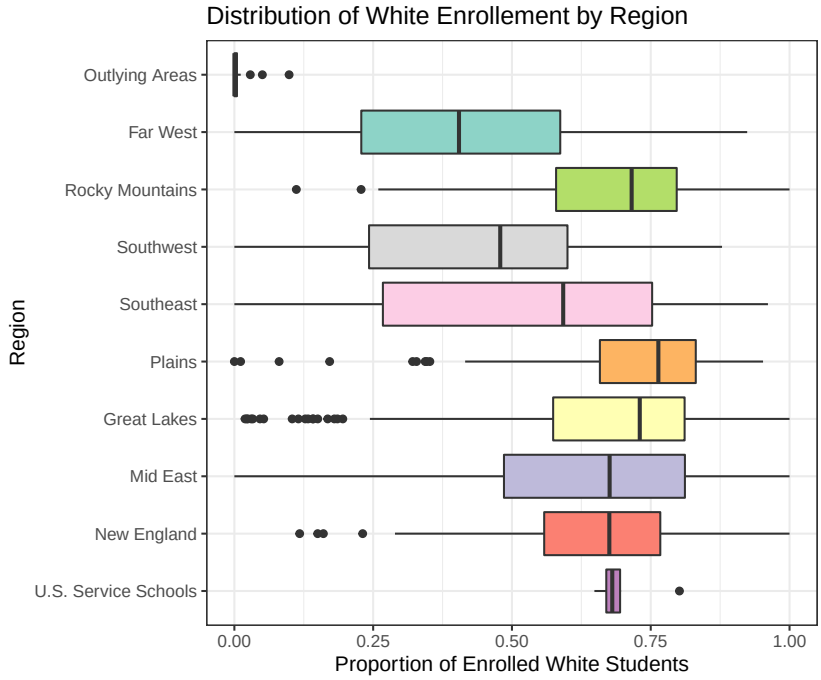
This data dictionary provides a definition for the variables in the data set used in this analysis.

Appendix 2: Geographic Distribution by Race

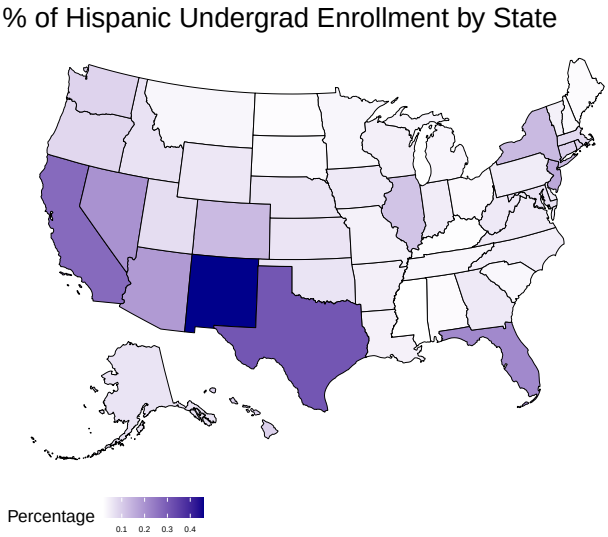
These box plot shows the distribution of student demographic by region for white, Black, Hispanic, and Asian, Native American/Alaskan Native, and Native Hawaiian/Pacific Islander undergraduates Though the US is a diverse country, there are higher concentrations of racial groups in various areas of the US whether this be due to the history of higher education, immigration patterns, or urban sprawl.[4]. The map shows the proportion of undergrads in a designated racial category to the total undergraduate population for a certain state. The box plot shows the distribution of the percentage of undergraduates in a racial category for every institution represented in the data set.

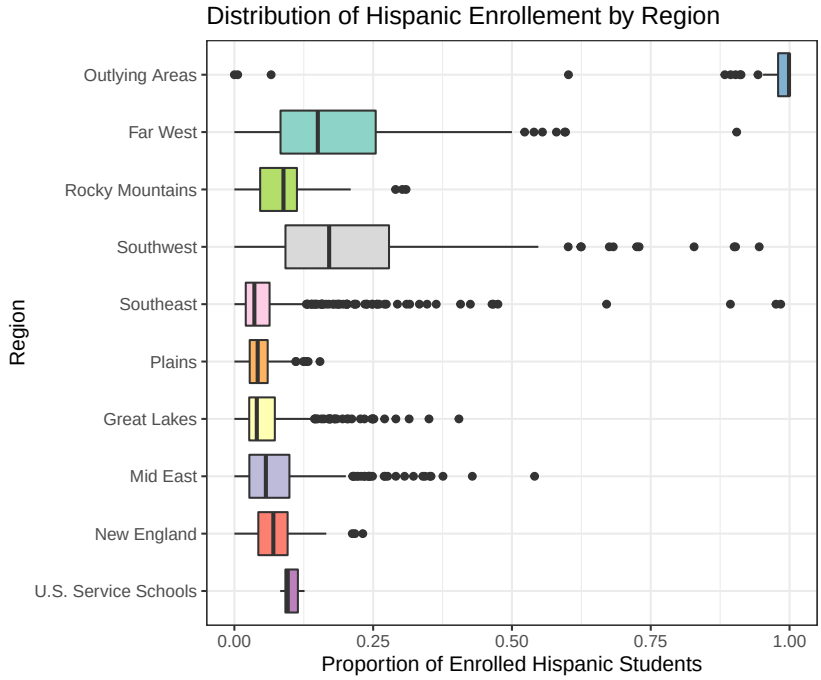
[4] The population has changed since 2013. But as a reference point, from the 2020 Census, the racial demographics were 58% white, 19% Hispanic, 12% Black, 6% Asian, 1% American Indian/Alaska Native, and 0.2% Native Hawaiian/Pacific Islander



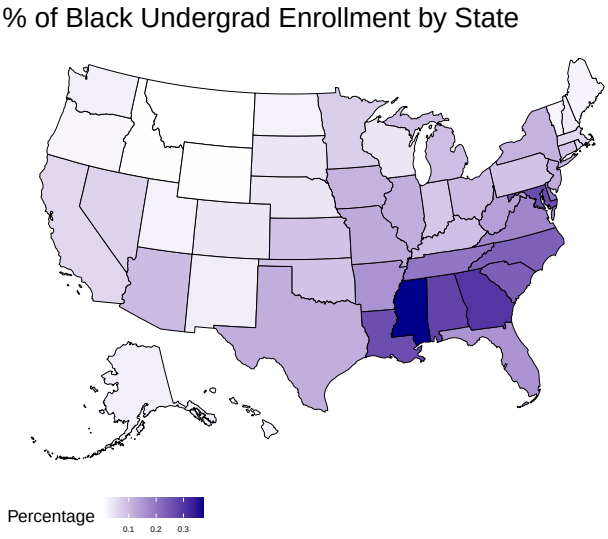


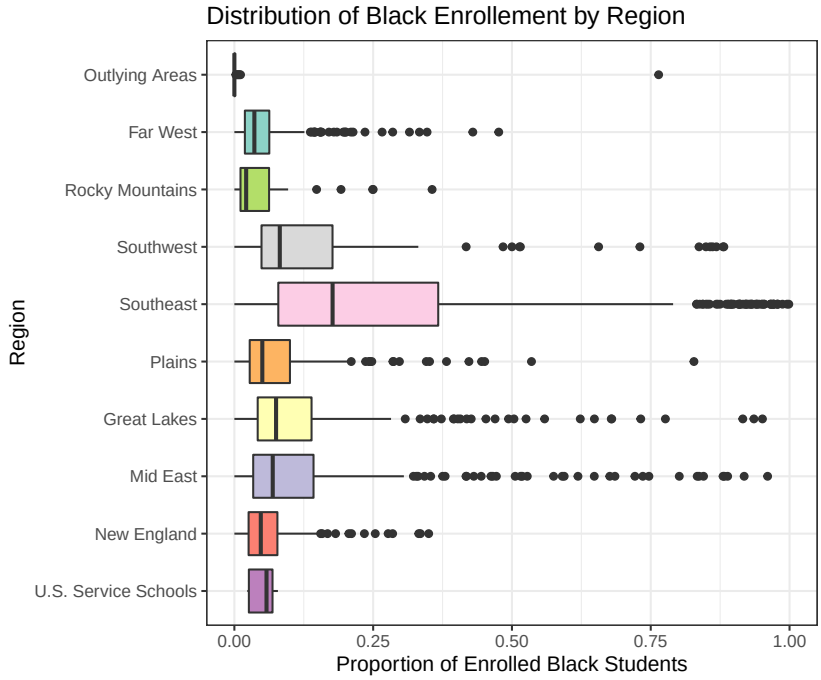
We observe there is a higher proportion of white students across most regions. White undergrad population are most notably the minority in Hawaii and states near the Mexican boarder such as California, Nevada, Arizona, New Mexico, and Texas. From the box plot, we white undergraduates typically make up the majority of undergraduate populations across the institutions represented in the data set.





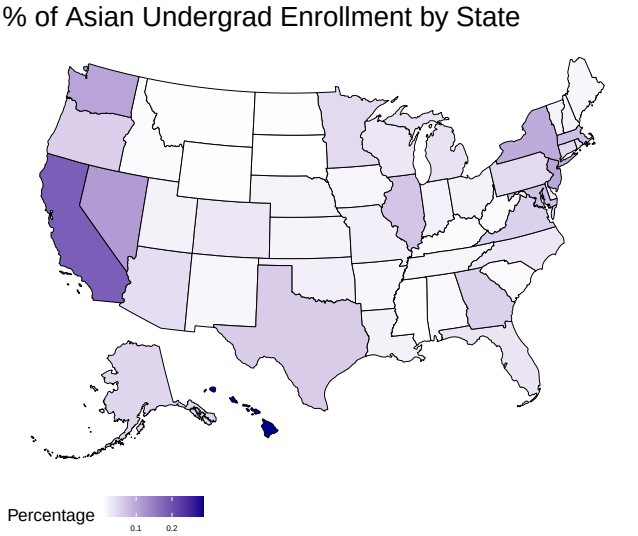
In contrast to the white undergraduate population, we see some of the states near the Mexican boarder to be associated with a higher Hispanic undergraduate population. From the box plot, we see these regions include the Southwest and Far West. This includes states such as California, Nevada, Arizona, New Mexico, Texas. In addition, Florida has a higher Hispanic population which may be attributed to its large Cuban population.

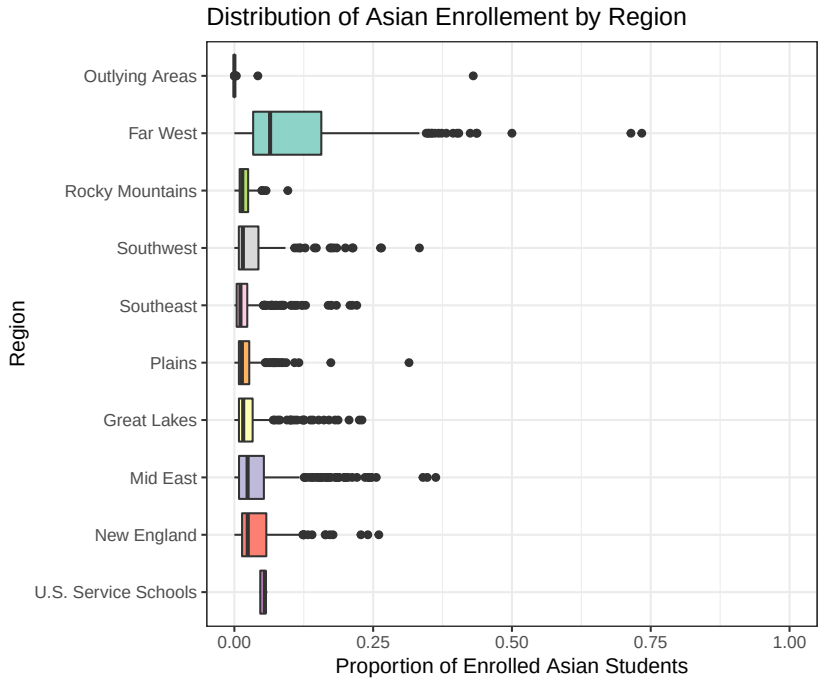




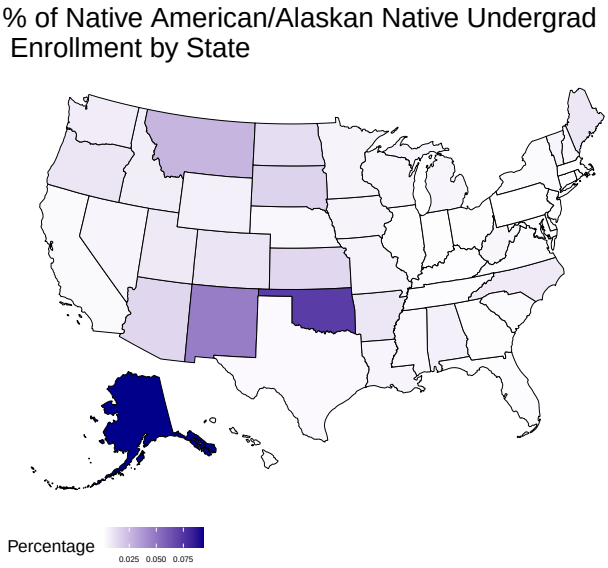
This map shows the average Black student enrollment percentage by state. Geographically, we see a higher concentration of Black undergrad enrollment in states located in the Southeast states. There is a larger population of Black people in the Southern States. In addition, many HBCUs are located in the Southeastern region which may contribute to a larger Black undergraduate population in these states. One observation from the box plot for enrolled black students is that the median is generally under 10% of the institution population, but there are many outliers as extreme as >75% for regions such as the

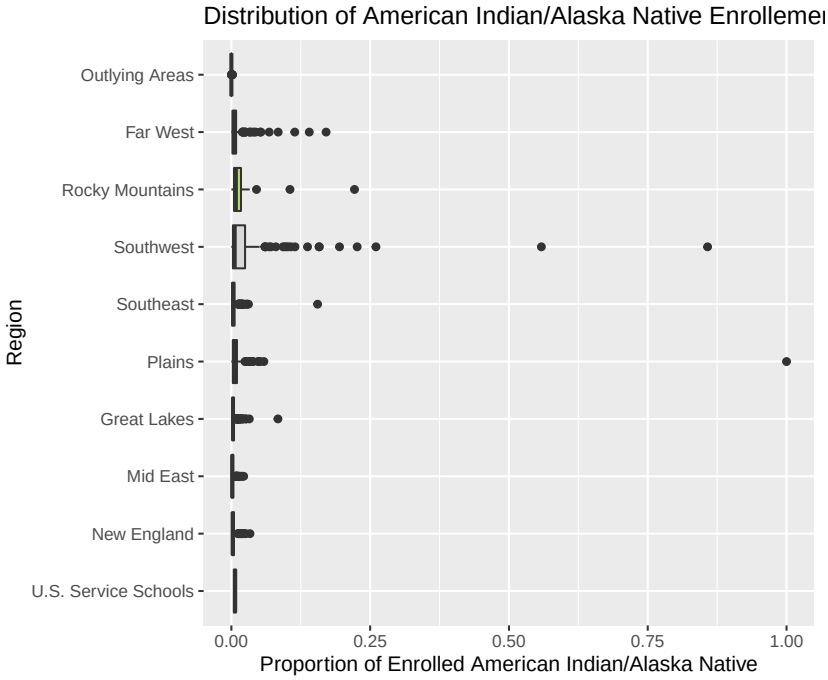
Southeast, Great Lakes, and the Mid East. One possible explanation is that many Historically Black College and Universities (HBCU) are located in these area, which can make up most of the outlying values.

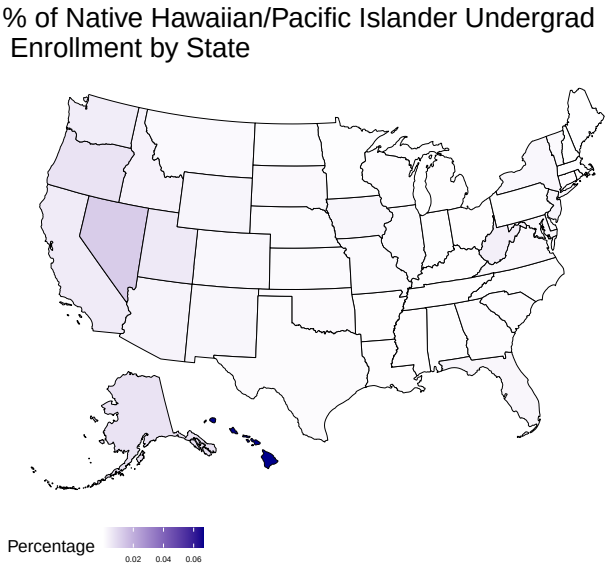


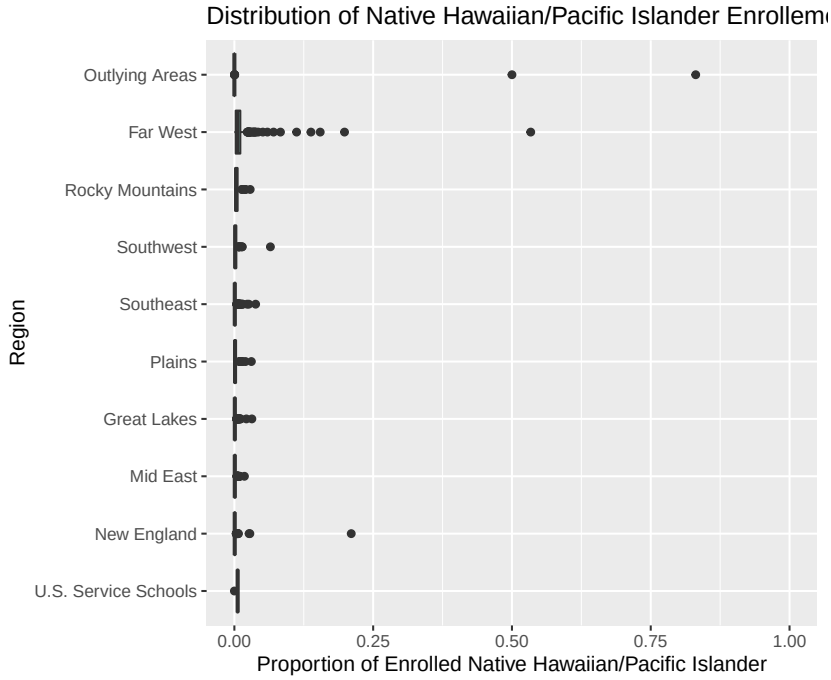


We notice there is a larger percentage of Asian undergraduates in states such as California, Nevada, Washington, New York, New Jersey, and Hawaii. These states are located on opposite coasts but share commonalities of being located near the coast.


% of Native American/Alaskan Native Undergrad Enrollment by State
Percentage
0.025
0.050
0.075


Distribution of American Indian/Alaska Native Enrollemei
Region
Outlying Areas
Far West
Rocky Mountains
Southwest
Southeast
Plains
Great Lakes
Mid East
New England
U.S. Service Schools
0.00
0.25
0.50
0.75
1.00
Proportion of Enrolled American Indian/Alaska Native





Though Indigenous and Pacific Islander students make ups a small percentage of the undergraduate population on most college campuses, it is important to acknowledge their presence in various states throughout the US. States. States such as Oklahoma, New Mexico and Alaska have the highest percentage of Indigenous Peoples. In addition, Hawaii has the largest percentage of Pacific Islander which can attribute its large undergraduate Pacific Islander/Native Hawaiian population.