423 Class Project - Milestone 10 Final Report

Project Title: Graduation Rates

Team Field Museum Members:

Anmol Raibhandare
Breanne Fahey
Elizabeth Unzueta
Lesley Bosniack
Sarita Patel

1. Introduction

American universities and colleges can base their admission on many factors like high school performance as an example and strive to achieve high graduation rates, and students base their choice of school on a number of criteria including graduation rates. In this study, we will explore what makes a good applicant for a particular school. Institutions desire applicants who will perform well and graduate within the timeline for the chosen degree. They must keep their entry criteria somewhat stringent to ensure that the educational quality does not dip to accommodate the less prepared students. They must also admit enough students who are likely to accept their admittance offer so that enrollment and incoming tuition does not fall below the level of financial solvency. On the other side, the students have the challenge of deciding to which colleges they will apply, balancing program choice, cost and quality. The school profiles are determined by student performance in examinations such as average SAT and ACT scores and other data points such as their ethnicity mix, percent of students receiving financial aid, tuition and fees, etc.

The specific problems considered are how universities select students for admission from a number of applicants and how future students choose the right university. Decisions by these two sets of parties contribute to the complexion of each university and in turn may explain graduation rates. Graduation rate is the response variable in this study. Admission Rate (Admitted / Applicants), one variable considered in explaining graduation rate, may indicate the selectiveness of a school with less selective schools yielding lower graduation rates and the more selective yielding higher rates. Standardized ACT/SAT tests are designed to predict a student's readiness for higher level learning and so may be a good predictor of graduation rate. Ethnicity mix may be representative of the cultural behaviors and expectations that might have a relationship with graduation rate. Average Tuition and Average Price may contribute to the actual or perceived quality of the education received, which in turn contributes to higher graduation rates. A higher percentage of full time students would also contribute to more students graduating within the four year timeframe. Intuition suggests that full-time students demonstrate a more concentrated and uninterrupted approach to their education and are following the timeframe set by the institution thus contributing to higher graduation rates for schools with higher percentages of full-time students. These are just a handful of variables considered to explain the varying levels of graduation rates. Given the significance of factoring Graduation Rate into the decision-making of universities and students, this study focuses on understanding the influence of various factors on the graduation rate of an institution, and based on this understanding, predicting graduation rates.

The National Center for Educational Statistics (NCES) is the primary federal entity for collecting and analyzing data related to education in the United States and other nations. The original source of the data for this study comes from the NCES. Due to the size of the NCES dataset and the tools available for analysis, an abbreviated version of this data was selected, the "American University Data" IPEDS dataset from Kaggle. This data can be found at https://www.kaggle.com/sumithbhongale/american-university-data-ipeds-dataset.

The abbreviated dataset contains details about the universities and colleges in the United States for the year 2013. The dataset contains 145 variables with 1534 instances of data with each instance representing an individual university. Four variables are unique to each academic institution and are not variables considered in the analysis with the exception of acting as identifiers: Name, ID, Longitude, Latitude, Address. In total, the dataset contains 114 numerical, 14 binary, 12 categorical including one ordinal, and 4 unique university identifiers, one of which is the primary key. With certainty, the unique identifiers, U.S. State, Degrees Awarded, County Name, Zip Code, Year, and Tribal College variables will be excluded from the regression analysis due to being instance identifiers and not being useful to the regression, a duplicated variable, a variable with many categorical values or a variable having identical values for all instances.

In this study, graduation rate was selected as the response variable. Two key factors contributing to the selection of this response variable were:

- Graduation rate provides a solid performance metric for a university. This information is useful to a university in improving upon the selection criteria for their admission process.
- Graduation rate is an important criterion for applicants to understand the quality of the university in terms of the course offerings, teaching, and satisfaction.

There are three variables for graduation rate in the dataset - Graduation rate - Bachelor's degree within 4 years, Graduation rate - Bachelor degree within 5 years and Graduation rate - Bachelor degree within 6 years. We selected the Graduation rate - Bachelor's degree within 4 years ("Graduation Rate") as our response variable for this study.

The explanatory variables are comprised mostly of numerical variables but also binary and categorical. As itemized above, variables have been categorized into SAT and ACT Percentile Scores and Percent Submitting Scores (10 numerical variables), Estimated Enrollment (12 numerical), Actual Enrollment (12 numerical), Tuition and Fees (4 numerical), Cost to

Attend (2 numerical), Profile of Enrollees (41 numerical), Profile of University (21 categorical, 1 binary), Percent of Freshmen receiving financial support (10 numerical) and Endowments (2 numerical). Note that a limited number of variables will be selected as explanatory variables whether used in the analysis as is or through some kind of feature extraction. Variable details of these categories is provided in Appendix - A.

2. Data Preparation

Our review of the dataset identified variables that are unnecessary, duplicate or potentially closely describe information contained elsewhere in the dataset, categorical variables of greater granularity than is necessary for our analysis, instances missing from a few to numerous entries, metrics that once transformed provide more value, among other findings. Combining or transforming variables, removing variables, and removing select instances are amongst the efforts taken to clean the dataset.

The dataset was provided as a single data frame. The scope was limited to undergraduate only leading to deletion of all variables with data for graduate level programs and students. This included variables labeled as graduate or as total,ie - Graduate enrollment.

We identified initial data preprocessing needs in addition to the above variable elimination with the most challenging aspect of working with this dataset is feature reduction given the level of detail in which the explanatory variables are provided and similarities of the variables. Some explanatory variables are provided across a series of columns with the variables supplied as a percentage of the number of students which raises the concern of dependencies across explanatory variables. Although the data quality of the dataset is high, there are fields with null values. In particular for the Endowment field, whether a null means the information was not provided vs endowments equal zero is important to understand. Given the importance of this variable, understanding the coding of the Endowment field was important.

Year, Level of Institution, and Tribal College contained identical values for every instance, and for SAT Writing 25th and 75th Percentile Score variables, approximately 45% of instances were lacking entries. As a result, these variables were also removed from the dataset. The level of detail provided in Zip Code, County Name, Longitude, Latitude, and the twelve binary variables that profile the Degrees Offered by a university are beyond the detail necessary for this study, while the Applicants Total, Admissions Total and Enrolled Total are either provided in greater and more meaningful granularly further into the dataset or the dataset contains relationships between the three fields (Percent Admitted and Admissions Yield variables) that are more meaningful than the counts themselves. Enrolled Total is also very similar to a variable further in the dataset. These variables were also removed.

The eleven “Degrees Awarded” and the seven “Number of Students Receiving Degrees” variables are response variables as is Graduation Rate. The response variable selection of Graduation Rate was chosen due to its focus on undergraduate students and its normalized nature in being a percent of counts and not counts themselves. As a result, Graduate Rate was viewed as the stronger choice as the response variable, eliminating the need for the other eighteen possible response variables.

State Abbreviation and Sector of Institution duplicate other variables and were eliminated for this reason. In addition, the twelve Estimated Enrollment variables are identical or very close to actual enrollment values contained in the dataset except for a few instances. Estimated Enrollment and Tuition years other than 2013-2014 variables provide little additional value and were eliminated from the dataset.The ten variables in "Financial Aid" contained somewhat identical values for every instance with few missing values. Only Percent of freshmen receiving institutional grant aid was included in the dataset due to high values among all others, that will eventually help in the analysis. Thus, other variables in Financial aid are excluded from the dataset.

In addition to maintaining the race/ethnicity percentages variables, we combined the ethnic groups into a Diversity Index. The Ethnic Diversity Index is intended to measure how much "diversity" or "variety" a school or district has among the ethnic groups in its student population reflect how evenly distributed these students are among the race/ethnicity categories. The formula was obtained from Ed-Data. Because enrollments relative to other variables were expected to be more meaningful than actual counts, Undergraduate Actual Enrollment as a Percent of Total Actual Enrollment and Full-time Actual Undergraduate Enrollment as a Percent of Total Undergraduate Enrollment were also calculated from existing variables. Because endowments appear to be reported from either GASB or FASB tax accounting standards but not both, the two Endowments Assets Per FTE variables were combined. The average SAT score for all topics combined was merged into the dataset from another source. This improved the number of instances available for SAT variable.

Lastly, there are a number of categorical variables providing a profile of each university. Examples of these are Highest Degree Offered and Control of Institution (Public or Private). To consider these categoricals as explanatory variables, conversion to dummy variables is necessary, increasing the size of an already large dataset. Further data analysis and investigation strengthened the ability to draw conclusions about the quality and applicability of each potential explanatory variable.

Since the study's focus is on undergraduate performance, the ten instances that have 'Implied No' or 'No' under "Offers Bachelor's degree" were removed. Those are Claremont Graduate University Row 95, Fielding Graduate University Row 99, Rensselaer Hartford Graduate Center Inc Row 170, Maryland University of Integrative Health Row 519, Teachers College at Columbia University Row 893, Pennsylvania State University-Penn State Great Valley Row 1121, SIT Graduate Institute Row 1353, Antioch University-New England Row 1477, Mayo Graduate School Row 1483, Union Graduate College Row 1521. We also removed two instances with incomplete data across a number of variables - U of North Georgia Row 259 and Texas A&M Galveston row 1310. Review of outliers and other dataset modifications were considered during formal data diagnostics and regression analysis.

With the above steps, the dataset was prepared for formal data diagnostics and future regression and consisted of 56 variables and 1,522 instances which was believed to provide many explanatory variables to choose from and a robust number of instances with which to work.

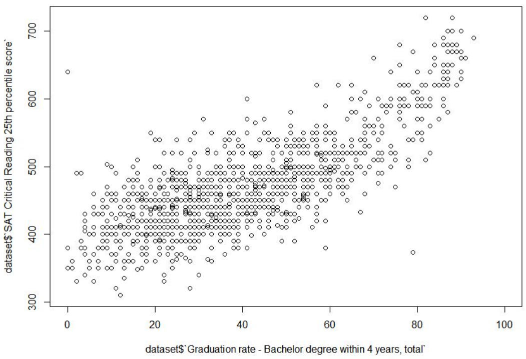
3. Data Analysis

Different techniques were used to analyze the explanatory variables. For numerical variables, five number summary, histograms, box plots and scatter plots were used. Correlations were also reviewed to understand the relationship between Graduation Rate and each explanatory variable. For categorical variables, bar charts and pivot tables were used along with linear regression exercises.

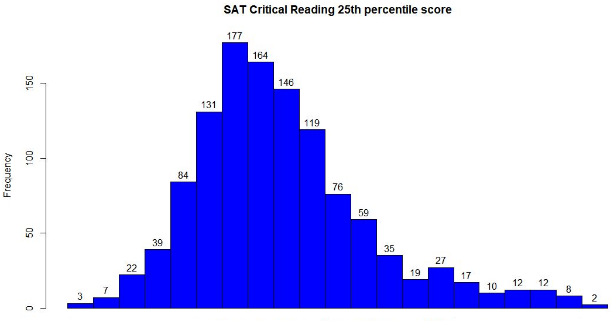
The **SAT Critical Reading 25th percentile score, SAT Critical Reading 75th percentile score, SAT Math 25th percentile score, SAT Math 75th percentile score, Average SAT Equivalent Score of Students Admitted, ACT Composite 25th percentile score and ACT Composite 75th percentile score** variables show a positive linear association pattern. To exemplify this, the relationship analysis between 'SAT Critical Reading 25th percentile score' variable and Graduation rate - Bachelor degree within 4 years, total is analyzed below.

The scatterplot below shows a strong, positive, linear association with a possible slight concavity between the SAT Critical Reading 25th percentile score and Graduation Rate. There are a few outliers in the data. Data points are scattered closely.

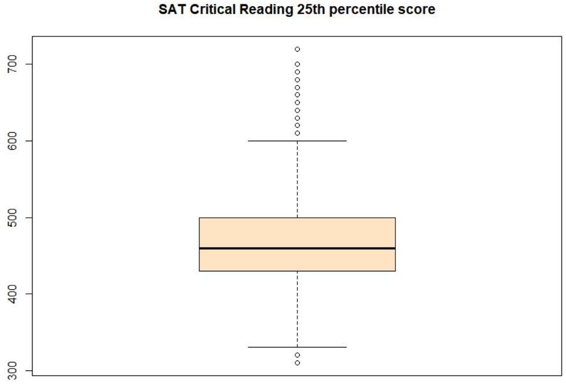
Correlation = 0.78 with range for other ACT/SAT of 0.74 - 0.80



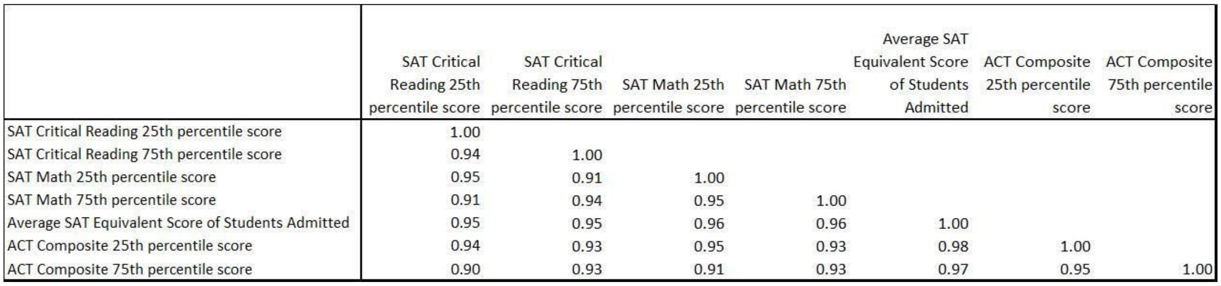
In the histogram below, the distribution of SAT Critical Reading 25th percentile score is skewed right. Scores are centered between 420 and 440. There is a high concentration of data between scores of 400 and 500 and a long tail to the right. .



The box plot is comparatively short with several outliers beyond the upper quartile in the dataset.



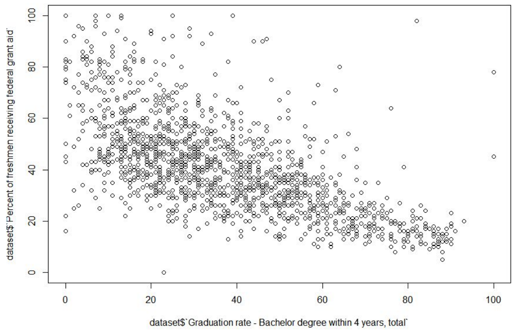
Note that correlations between most of the SAT and ACT variables (“SAT/ACT”) as exhibited below are very high which may suggest redundancy in their contribution to Graduation Rate. Ultimately, this is addressed further in the study.



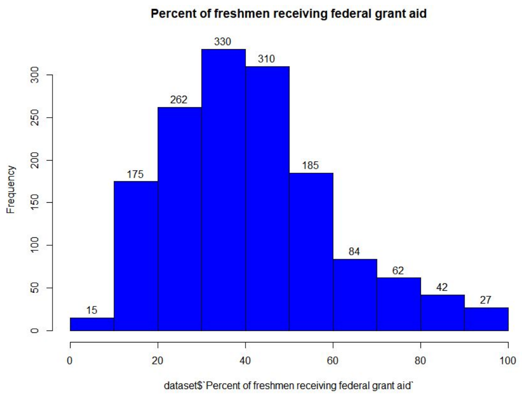
	SAT Critical Reading 25th percentile score	SAT Critical Reading 75th percentile score	SAT Math 25th percentile score	SAT Math 75th percentile score	Average SAT Equivalent Score of Students Admitted	ACT Composite 25th percentile score	ACT Composite 75th percentile score
SAT Critical Reading 25th percentile score	1.00						
SAT Critical Reading 75th percentile score	0.94	1.00					
SAT Math 25th percentile score	0.95	0.91	1.00				
SAT Math 75th percentile score	0.91	0.94	0.95	1.00			
Average SAT Equivalent Score of Students Admitted	0.95	0.95	0.96	0.96	1.00		
ACT Composite 25th percentile score	0.94	0.93	0.95	0.93	0.98	1.00	
ACT Composite 75th percentile score	0.90	0.93	0.91	0.93	0.97	0.95	1.00

We believe that the SAT/ACT scores are a major factor in predicting the Graduation Rate. Success in standardized tests provides an indication of a student’s baseline level of knowledge and the potential success in higher level education thus contributing to Graduation Rate.

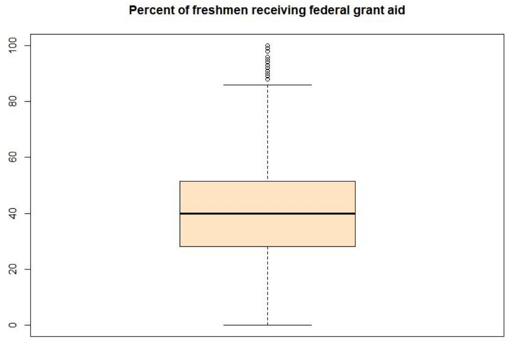
The relationship between **The Percent of Freshmen Receiving Federal Grant Aid and Percent of Freshmen Receiving Pell Grants** variables show a negative linear association pattern. The scatterplot shows a strong, negative, linear association with mild concavity between the Percent of Freshmen Receiving Federal Grant Aid and Graduation Rate. In the scatter plot, the data points are scattered far away from the plotted linear regression line, with few outliers.



Correlation = -0.65



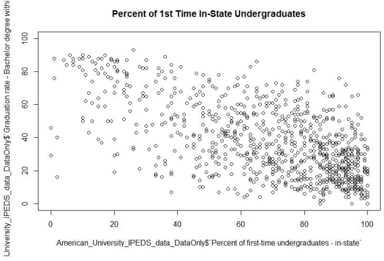
The distribution of Federal Grant Aid variable is skewed right with a concentration of data between 20-50%. Roughly half of the graduation rates are between 30-40% of the total graduation. There are possible outliers in the data, but the histogram cannot accurately depict this.

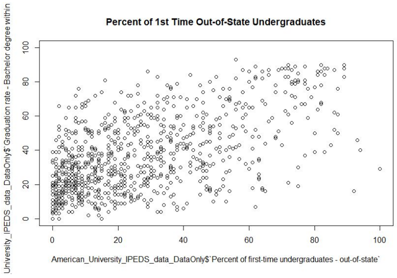


The box plot is comparatively short. This suggests that overall Federal Grant Aid variable has a high level of correlation with each other. The upper quartile is at 50%, and the lower quartile is at 30%. The median is at 40% with several outliers beyond the upper quartile in the dataset.

'Percent of Freshmen Receiving Pell Grants variable' and Graduation Rate show identical results. Other financial aid variables delivered less significant relationships to Graduation Rate. Thus far, the combination of the metrics delivers a favorable linear regression result. Linear regression indicates that the financial aid variables explain approximately 51% of the variability in Graduation Rate. Comparable to the Financial Aid variables, the **Percent of Undergraduate Enrollment that are of a particular ethnicity** provides metrics that suggest that these variables are not predictive of Graduation Rates. However, linear regression results of the group of these variables suggest otherwise. Linear regression indicates that ethnicity variables explain approximately 23% of the variability in Graduation Rates. Intuitively, ethnicity may be representative of cultural behaviors and standards of living that would suggest a relationship with Graduation Rate.

Percent of First Time Undergraduates - In-State variable shows a negative linear association pattern with a possible mild downward concavity, while **Percent of First Time Undergraduates- Out-Of-State** variable shows a positive linear association pattern again with a possible mild downward concavity. The scatterplots shows a fairly strong, negative, linear association between the In-State variable and Graduation Rate. No outliers were observed in the data. In the scatter plot, data points are widespread over the graph and appear to have a mild downward arc. Observations are similar for the Out-of-State variable showing few outliers but a fairly strong, positive association.

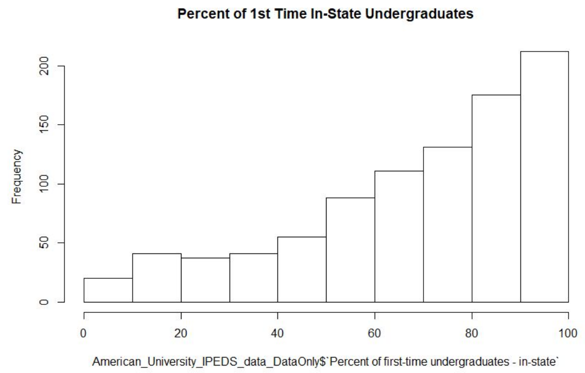


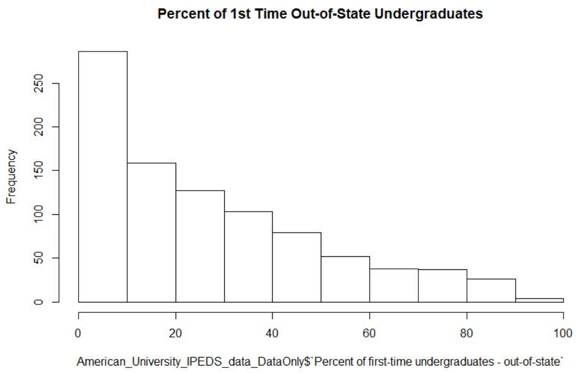


Correlation = -0.60

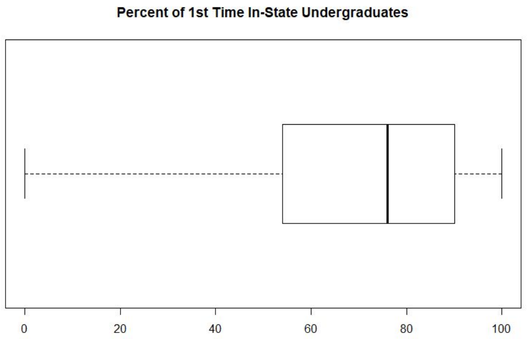
Correlation = 0.57

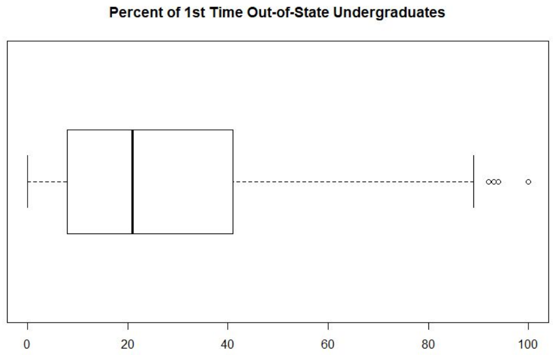
The distribution of the In-State variable is skewed to the left while Out-of-State is skewed to the right. The growth in the In-State Undergraduates is gradual across the histogram.



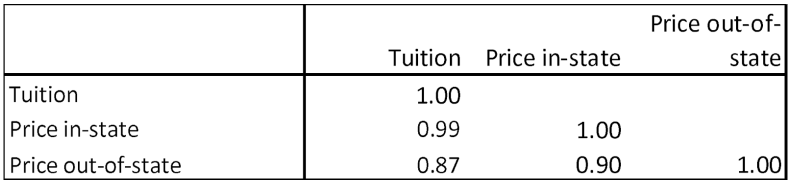


Box plots suggests a high concentration on higher percentages for In-State and on lower percentages for Out-of-State. It is not clear to us why these metrics produce the observed relationships to Graduation Rate particularly with respect to In-State.





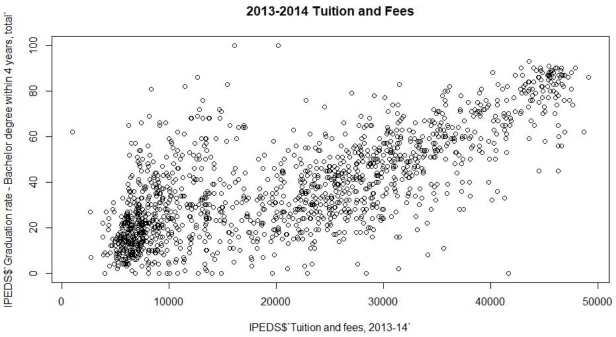
Tuition and Fees 2013-14, Total Price for In-State Students Living on Campus 2013-14, and Total Price for Out-of-State Students Living on Campus 2013-14 variables all show positive linear association patterns. The Tuition and Fees variable and possibly Price variables exhibit a concave up tail on the right side of the scatterplot. Note that correlations between Tuition and Price are very high which may suggest redundancy in their contribution to Graduation Rate.

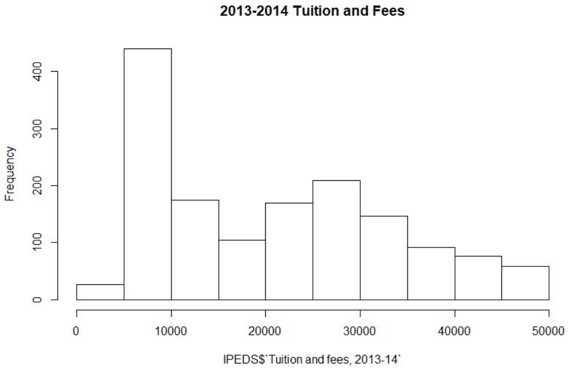


	Tuition	Price in-state	Price out-of-state
Tuition	1.00		
Price in-state	0.99	1.00	
Price out-of-state	0.87	0.90	1.00

The scatterplot shows a strong, positive, linear association between Tuition and Graduation Rate with some concave up curvature at lower tuitions. No outliers are observed. The scatter plot is scattered fairly tightly beyond $15,000 but widespread below $15,000.

Correlation = 0.73

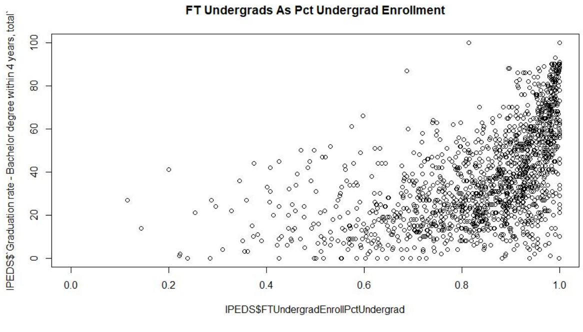




The distribution of Tuition is skewed right. with concentration of points in the $5,000-10,000 tuition amounts.

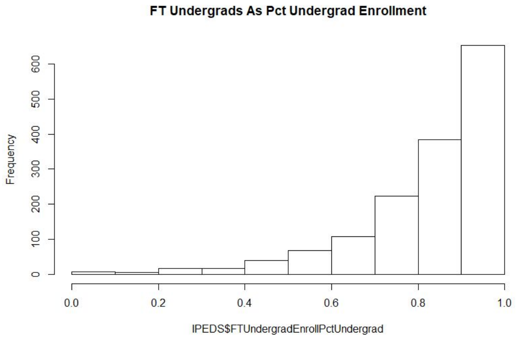
The In-State Price variable has a similar distribution but shifted to the right by approximately $15,000. In addition, the right skewed tendency is not as significant as Tuition. The Out-of-State distribution has much more of an appearance of a normal distribution than the other two variables and a peak at a greater price level than In-State as expected. We believe that Tuition and Price may contribute to the actual or perceived quality of the education received, which in turn contributes to higher graduation rates.

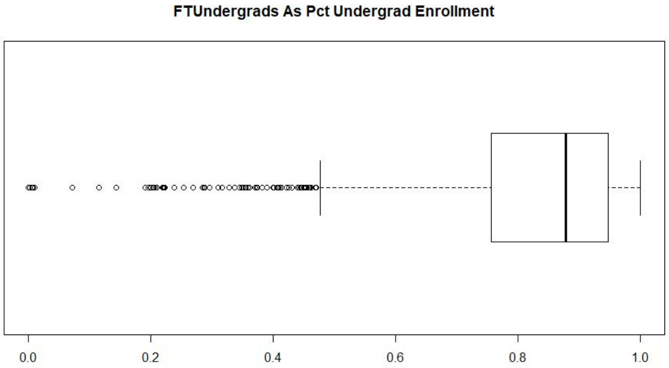
The analysis between **'FT Undergrads As Pct Undergrad Enrollment'** variable has a fairly strong, positive, concave up relationship to Graduation Rate. This variable was tested for inclusion in the regression model as a second order term.



Correlation = 0.54

The distribution of FT Undergrads As Pct of Undergrad Enrollment is skewed to the left with a long thin tail and a high concentration above 80%.





The box plot is comparatively short further demonstrating the high concentration.

There are approximately 40 outliers in the data for which no notable cause was found for their levels. Intuition suggests that full-time students demonstrate a more concentrated and uninterrupted approach to their education which improves their academic success thus contributing to higher graduation rates for schools with higher percentages of full-time students.

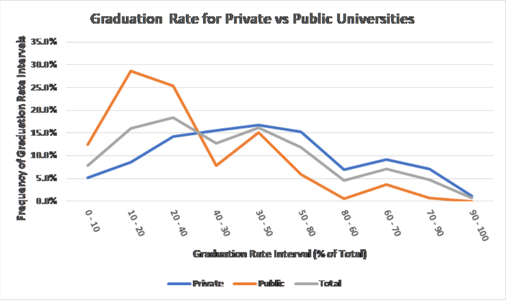
Categorical Variables

Carnegie Classification
FIPS State Code
Historically Black College or University
Control of Institution
Geographic Region
Degree of Urbanization
Highest Degree Offered
Religious Affiliation

Bar charts of the distributions of categorical variables were analyzed to understand the number of values, shape of the distribution, outlying values and also whether combining values makes sense. Each variable was pivot tabled against the Graduation Rate to identify distributional similarities and differences across categorical values for Graduation Rate intervals. Due to the challenges of analyzing categorical variables, linear regression was performed on each. Of the variables considered above, Carnegie Classification and Control of Institution provided interesting results and/or provided the most significant results of all categoricals in the pivot table and linear regression analyses.

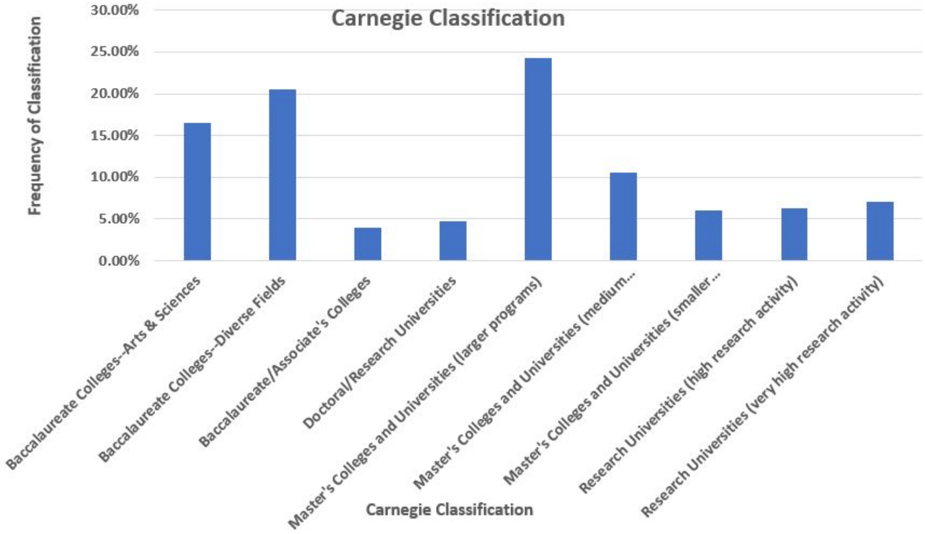
Control of Institution results suggest a difference in the distribution of Graduation Rates of Public vs Private schools.

Graduation Rate Interval	Private	Public	Total
0 - 10	5.2%	12.4%	7.9%
10 - 20	8.5%	28.6%	16.0%
20 - 40	14.2%	25.3%	18.4%
40 - 30	15.5%	7.8%	12.7%
30 - 50	16.7%	15.1%	16.1%
50 - 80	15.3%	5.8%	11.8%
80 - 60	7.0%	0.5%	4.6%
60 - 70	9.2%	3.6%	7.1%
70 - 90	7.1%	0.7%	4.7%
90 - 100	1.2%	0.0%	0.7%
Total	100%	100%	100%

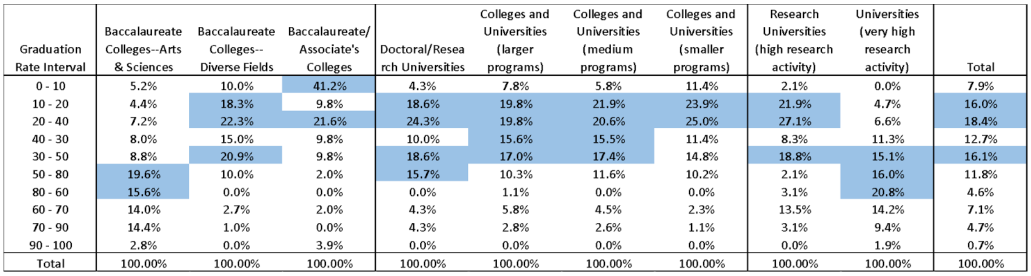


Univariate linear regression indicates that Control of Institution explains approximately 16% of the variability in Graduation Rate. We believe higher graduation rates for private schools relative to public may interact with quality of education and tuition levels. We believe these interactions are real and will be explored in the regression analysis.

The spread of **Carnegie Classifications** is across nine categories with distributions spanning from 4% to 24%. Three classes represent the majority of the distribution, but all classes are fairly well represented.



The distributional results of Graduation Rate by Carnegie Classification in the chart below indicate differences in distributions across many of the classes.



Graduation Rate Interval	Baccalaureate Colleges--Arts & Sciences	Baccalaureate Colleges--Diverse Fields	Baccalaureate/ Associate's Colleges	Doctoral/Research Universities	Colleges and Universities (larger programs)	Colleges and Universities (medium programs)	Colleges and Universities (smaller programs)	Research Universities (high research activity)	Universities (very high research activity)	Total
0 - 10	5.2%	10.0%	41.2%	4.3%	7.8%	5.8%	11.4%	2.1%	0.0%	7.9%
10 - 20	4.4%	18.3%	9.8%	18.6%	19.8%	21.9%	23.9%	21.9%	4.7%	16.0%
20 - 40	7.2%	22.3%	21.6%	24.3%	19.8%	20.6%	25.0%	27.1%	6.6%	18.4%
40 - 30	8.0%	15.0%	9.8%	10.0%	15.6%	15.5%	11.4%	8.3%	11.3%	12.7%
30 - 50	8.8%	20.9%	9.8%	18.6%	17.0%	17.4%	14.8%	18.8%	15.1%	16.1%
50 - 80	19.6%	10.0%	2.0%	15.7%	10.3%	11.6%	10.2%	2.1%	16.0%	11.8%
80 - 60	15.6%	0.0%	0.0%	0.0%	1.1%	0.0%	0.0%	3.1%	20.8%	4.6%
60 - 70	14.0%	2.7%	2.0%	4.3%	5.8%	4.5%	2.3%	13.5%	14.2%	7.1%
70 - 90	14.4%	1.0%	0.0%	4.3%	2.8%	2.6%	1.1%	3.1%	9.4%	4.7%
90 - 100	2.8%	0.0%	3.9%	0.0%	0.0%	0.0%	0.0%	0.0%	1.9%	0.7%
Total	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

Univariate linear regression indicates that Carnegie Classification explains approximately 23% of the variability in Graduation Rate. We expect that academic institutions with higher levels of education than baccalaureate may create positive learning opportunities delivering greater success to undergraduate students, but for schools with graduate programs, instructors may be more focused on research, leaving students to rely on teaching assistants or alternative resources to learn. This may be one of multiple factors contributing to the unexpected observations in the chart above.

4. Model Building

In this analysis, we use multivariate linear regression modeling to predict Graduation Rates in American universities. Specifically we will identify what variables are most significant to explaining Graduation Rates. Because a number of explanatory variables that have been identified are related, we consider the relationships between explanatory variables in the regression analysis. Further, we dissect the analysis into public and private schools to identify the differences, if any, in the variables that contribute to the levels of Graduation Rates. Because the number of instances in the dataset was reduced by approximately 15% due to missing ACT/SAT scores, we analyze the impact of the instance reduction. To do this, we measure the impact on other explanatory variables' contributions to predicting Graduation Rate.

Our goal was to analyze the dataset and evaluate a large number of institutional variables that may help predict Graduation Rate. Using the IPEDS data, we aimed to answer the following questions:

- What are the key criteria that drive Graduation Rate?
- Which criteria are most significant?
- Do the drivers vary by public vs private institution and how?
- And lastly, how does the removal of instances affect the prediction of Graduation Rate for non-ACT/SAT variables?

In order to accomplish this, we perform a comprehensive analysis considering all explanatory variables and all instances in the dataset and partition the study as follows.

Comprehensive Analysis

In this project, the team built various regression models to predict the Graduation Rate, after preprocessing the entire dataset and selecting most significant explanatory variable (using correlation and Model Selection Methods), using Multi-variable Linear Regression Model. With a clean, full dataset and most important (correlated) explanatory variables, we build the regression model and further refine the explanatory variables using the Model Selection Methods – Backward, Forward, and Stepwise. The previous step will give the regression model with the best explanatory variables to use in the final model. Now, we use these explanatory variables from the previous step to build our model and make predictions for the Graduation Rate. The final step will be to explain the model and describe the accuracy of the model in predicting the Graduation Rate.

Public vs Private Institutions

Because we believe that the factors contributing to Graduation Rates for public and private may vary, we segment the dataset into public vs private schools and perform separate regression analysis on each. Public schools comprise about one third of instances in the dataset. These schools were tested to see if there is a correlation between public schools and Graduation Rate and also all other variables within the public school subset. Due to the large number of private institutions, the rows for private universities are split in half by Carnegie Classification to aid in the comparison to public institutions. By analyzing separately, we hoped to identify the factors that contribute to predicting graduation rates of private institutions. The dataset was split into two segments based on Carnegie Classifications: Baccalaureate or Doctoral/Master/Research institutions. Analysis of each subset included Model Selection Methods – Backward, Forward, and Stepwise, on their subsets. N-Folds validation was also considered.

Impact of Missing ACT/SAT Scores

Because we believe that the number of instances being removed from the full dataset for missing ACT/SAT variables is not insignificant and may have an effect on the relationship of non-ACT/SAT variables with Graduation Rate, we analyze the impact of the instances removed. Because the point of the analysis is to measure impacts of non-ACT/SAT variables on Graduation Rate, ACT/SAT variables are excluded from the analysis. The dataset is cleaned for all variables except ACT/SAT. Due to a number of instances missing ACT/SAT values, this increased the number of instances in the dataset over the complete dataset by approximately 15%. Multivariate linear regression models using stepwise regression is used to narrow down the explanatory variables and inform the building of the model. Due to the collinearity of some variables (Distributions of Ethnicity and Financial Aid), Variance Inflation Factor analysis ("VIF") analysis informs the choice of variables for modeling. N-Folds Cross Validation assists in model validation. Residuals are reviewed for randomness as well. Findings of this analysis are compared to other dataset analyses to identify significant similarities and differences in predictors of Graduation Rate.

5. Comprehensive Model Analysis:Sarita Patel

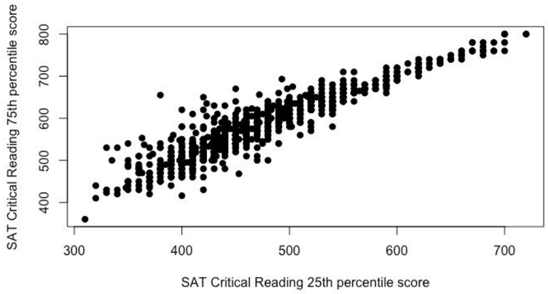
In the comprehensive model (included all variables and filtered rows containing null values) built by the team, we performed an exploratory data analysis of the variables filtered in the previous milestones to understand their impact on the response variable and their importance in the model. This process helped us to eliminate variables that were not having much impact or effect on the response variable. We used these refined variables to build models with first-order, second-order and interaction terms and evaluated them using the model evaluation metrics and scoring the models (train and test data). A final model was selected per the model evaluation metrics and model accuracy. A residual analysis of the final model was carried out to confirm that the model agrees to the assumption we made while building the model. Please find below the details of the steps performed during the Comprehensive Model building process.

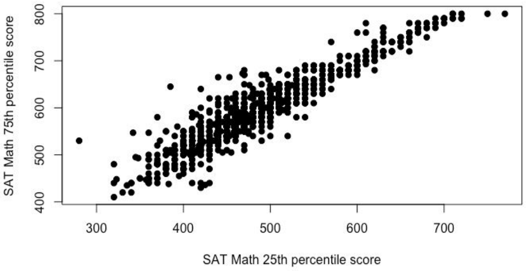
a. Feature Engineering

SAT/ACT Variables

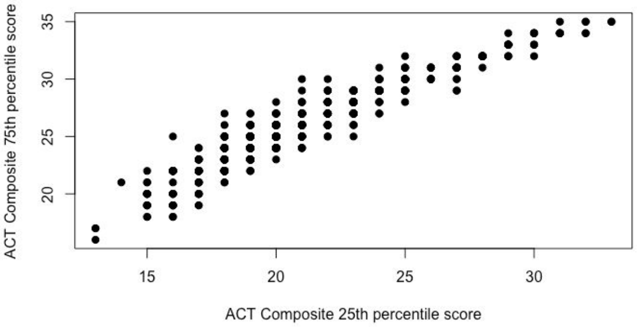
All of the SAT/ACT variables were highly correlated to each other and to the response variable.

SAT Critical Reading 25th percentile score and SAT Critical Reading 75th percentile score are very highly correlated. The variable "SAT Critical Reading 25th percentile score" appears to be the better candidate for our model as it could be close to the cut-off score for admission to a University. Therefore, we removed the "SAT Critical Reading 75th percentile score" variable from our model.





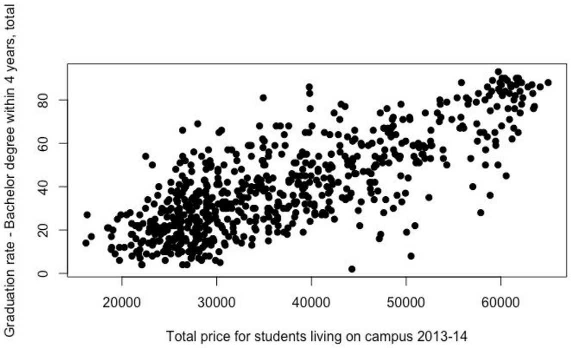
The SAT Math 25th percentile score and SAT Math 75th percentile score are highly correlated with each other as seen in figure with the value of correlation coefficient close to 0.95. Again, we used our domain knowledge and make the decision to retain the variable "SAT Math 25th percentile score" variable in our dataset.



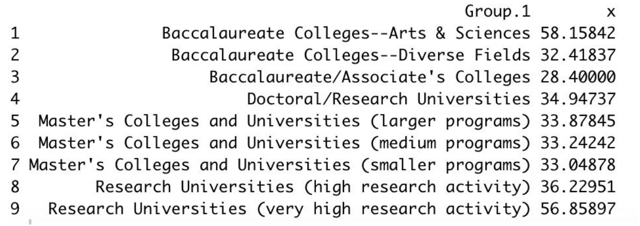
ACT Composite 25th percentile score and ACT Composite 75th percentile score are also highly correlated with each other. With our domain knowledge we made a decision to retain the variable "ACT Composite 25th percentile score" only as candidate of explanatory variable in the Comprehensive Model.

Total price for in-state students living on campus 2013-14 and Total price for out-of-state students living on campus 2013-14 variables are highly correlated to each other.With our domain knowledge, we created a new variable that was an average (in-state and out-of-state) of both these variables in our dataset. We called this variable as "Total price for students living on campus 2013-14".

The new variable had a strong linear relationship with the response variable with a correlation coefficient value of 0.77. We used this variable in further processed for the Comprehensive Model only.



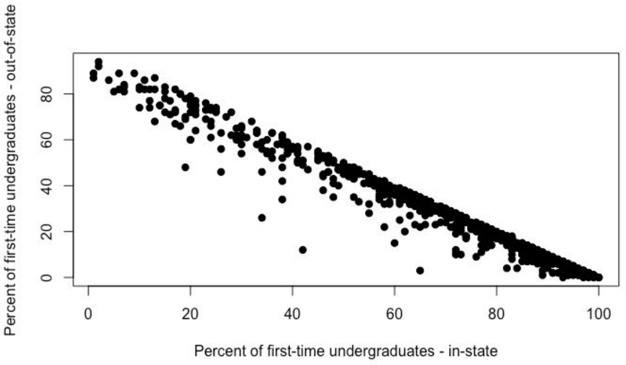
Carnegie Classification 2010: Basic



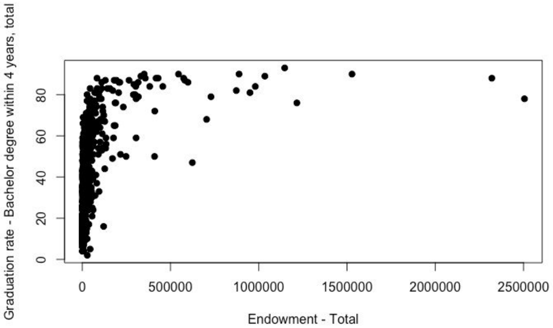
```
                                              Group.1        x
1              Baccalaureate Colleges--Arts & Sciences 58.15842
2               Baccalaureate Colleges--Diverse Fields 32.41837
3                   Baccalaureate/Associate's Colleges 28.40000
4                       Doctoral/Research Universities 34.94737
5  Master's Colleges and Universities (larger programs) 33.87845
6  Master's Colleges and Universities (medium programs) 33.24242
7 Master's Colleges and Universities (smaller programs) 33.04878
8         Research Universities (high research activity) 36.22951
9    Research Universities (very high research activity) 56.85897
```

In the figure above we see that this categorical variable has 9 values. The average value of the "Graduation rate - Bachelor degree within 4 years, total" is added against each of these values. While most of these categories has an average Graduation rate of approximately 30%, the categories "Baccalaureate Colleges--Arts & Sciences" and "Research Universities (very high research activity)" has an average graduation rate of more than 50%. Therefore, these two categories are important for the comprehensive model. We created two dummy variables "Baccalaureate Colleges--Arts & Sciences" and "Research Universities (very high research activity)" that we used it in the Comprehensive Model.

Percent of first-time undergraduates - in-state and Percent of first-time undergraduates - out-of-state variables have a strong negative linear relationship with each other with a correlation coefficient of -0.98. It appears that these two variables are complementary of each other. Therefore, we selected the in-state variables for the Comprehensive Model.



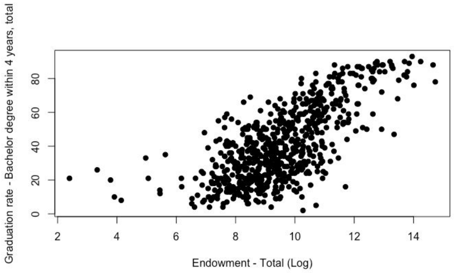
Endowment - Total



The correlation coefficient of this variable with the response variable is 0.4494183. So, it might have a linear relationship with the response variable.

Per the above figure, it seems that Graduation Rate increases exponentially with the increase in endowment. So, the "Endowment - Total" has an exponential relationship with the response variable. We applied a *log transformation (natural log)* on the "Endowment - Total" variable and checked the relationship again using R.

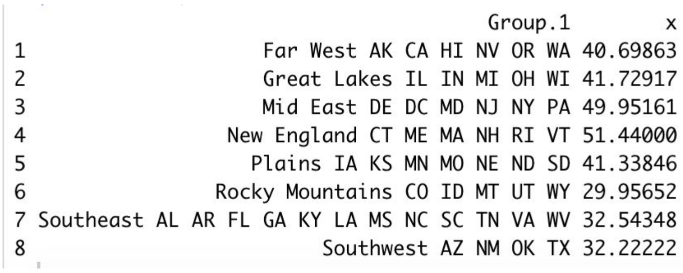
Next, we created a scatter plot of this variable with the response variable:



As seen in the figure above, the log transformed variable has a linear relationship with the response variable with a correlation coefficient of 0.70. Therefore, we considered this new transformed variable for building the Comprehensive Model.

Geographic region

This variable gives the geographic region per the classification given in the figure below. We looked at the average Graduation Rate by each these regions.



```
                                           Group.1        x
1                   Far West AK CA HI NV OR WA 40.69863
2                 Great Lakes IL IN MI OH WI 41.72917
3                   Mid East DE DC MD NJ NY PA 49.95161
4                 New England CT ME MA NH RI VT 51.44000
5                   Plains IA KS MN MO NE ND SD 41.33846
6                     Rocky Mountains CO ID MT UT WY 29.95652
7 Southeast AL AR FL GA KY LA MS NC SC TN VA WV 32.54348
8                          Southwest AZ NM OK TX 32.22222
```

Per the figure, geographic regions MidEast and NewEngland has graduation rate around 50%, geographic regions FarWest, GreatLakes, and Plain around 40% and other geographic regions has graduation rate of 30%. Therefore, this variable has an influence on the response variable. We created two dummy variables and include it in our Comprehensive Model using R.

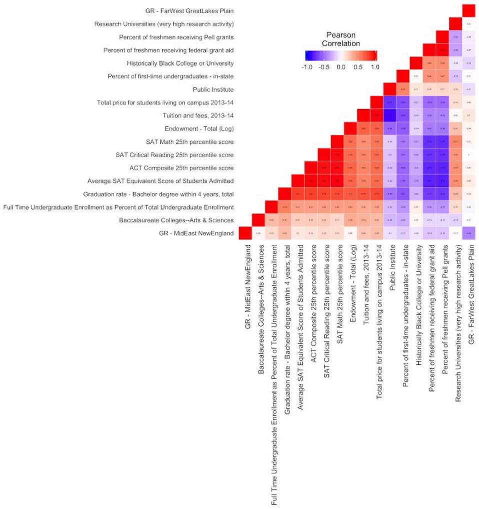
Variable Refinement

We were left with following variables that were a good candidate for the Comprehensive Model:

[1] "SAT Critical Reading 25th percentile score"
[2] "SAT Math 25th percentile score"
[3] "Average SAT Equivalent Score of Students Admitted"
[4] "ACT Composite 25th percentile score"
[5] "Tuition and fees, 2013-14"
[6] "Total price for students living on campus 2013-14"
[7] "Baccalaureate Colleges--Arts & Sciences"
[8] "Research Universities (very high research activity)"
[9] "Historically Black College or University"
[10] "GR - FarWest GreatLakes Plain"
[11] "GR - MidEast NewEngland"
[12] "Public Institute"
[13] "Full Time Undergraduate Enrollment as Percent of Total Undergraduate Enrollment"
[14] "Percent of first-time undergraduates - in-state"
[15] "Percent of freshmen receiving federal grant aid"
[16] "Percent of freshmen receiving Pell grants"
[17] "Endowment - Total (Log)"
[18] "Graduation rate - Bachelor degree within 4 years, total"

Multicollinearity Check

In our previous analysis, we did the multicollinearity checks for some of the variables in our dataset, but we performed another multicollinearity check for the variables listed above.

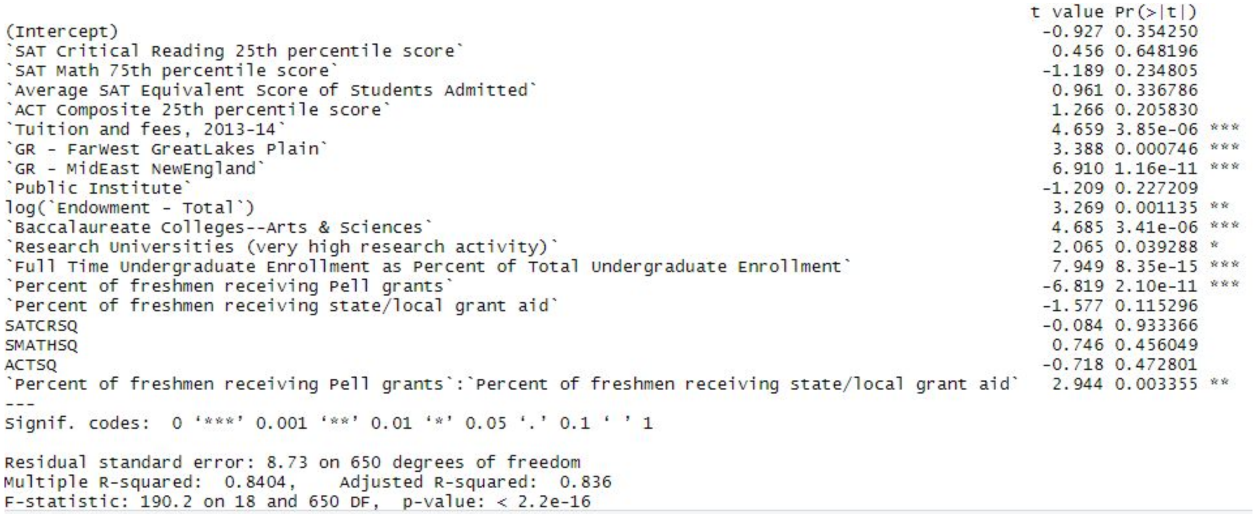


As seen in the figure above, the squares highlighted in dark red are strongly positively correlated and the squares highlighted in dark blue/purple are strongly negatively correlated. The “SAT Critical Reading 25th percentile score”, "SAT Math 25th percentile score" , "Average SAT Equivalent Score of Students Admitted", and "ACT Composite 25th percentile score" are strongly positively correlated with each other. So, we retained one or two variables out of these two variables for our model. Also, the variables "Percent of freshmen receiving Pell grants" and "Percent of freshmen receiving state/local grant aid" are strongly positively correlated with each other, so retained just one variable out of these two variables.

b. Model building and Evaluation

Using the variables refined after the Multicollinearity check, we built our model using various combinations of first-order, second-order and interaction terms. These models were evaluated using the MSE, Adj-R2, F-test of overall significance of the regression model, and t-test of the regression model coefficients. Furthermore, the model was scored using the training and test data to determine the accuracy of the model in predicting the Graduation Rate. We finally arrived at the following model that has a good performance, accuracy, and easy to understand.

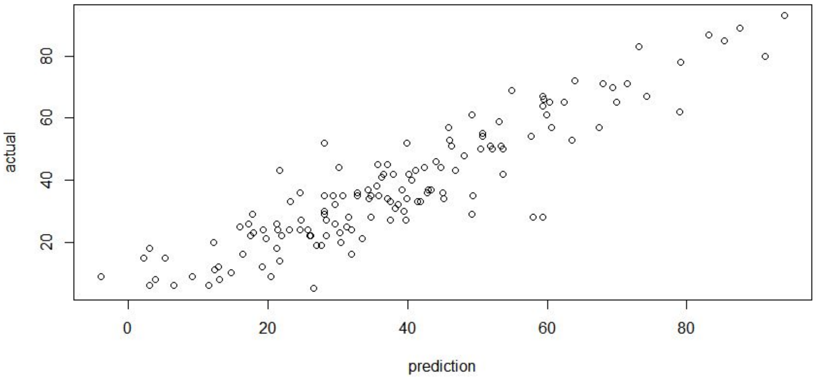
A residual analysis of this final regression model was performed to check and validate the assumptions made while building this model and the residuals: normality, linearity, homoscedasticity, and absence of multicollinearity.



```
                                                                                                  t value Pr(>|t|)
(Intercept)                                                                                        -0.927 0.354250
`SAT Critical Reading 25th percentile score`                                                        0.456 0.648196
`SAT Math 75th percentile score`                                                                   -1.189 0.234805
`Average SAT Equivalent Score of Students Admitted`                                                 0.961 0.336786
`ACT Composite 25th percentile score`                                                               1.266 0.205830
`Tuition and fees, 2013-14`                                                                         4.659 3.85e-06 ***
`GR - FarWest GreatLakes Plain`                                                                     3.388 0.000746 ***
`GR - MidEast NewEngland`                                                                           6.910 1.16e-11 ***
`Public Institute`                                                                                 -1.209 0.227209
log(`Endowment - Total`)                                                                            3.269 0.001135 **
`Baccalaureate Colleges--Arts & Sciences`                                                           4.685 3.41e-06 ***
`Research Universities (very high research activity)`                                               2.065 0.039288 *
`Full Time Undergraduate Enrollment as Percent of Total Undergraduate Enrollment`                   7.949 8.35e-15 ***
`Percent of freshmen receiving Pell grants`                                                        -6.819 2.10e-11 ***
`Percent of freshmen receiving state/local grant aid`                                              -1.577 0.115296
SATCRSQ                                                                                            -0.084 0.933366
SMATHSQ                                                                                             0.746 0.456049
ACTSQ                                                                                              -0.718 0.472801
`Percent of freshmen receiving Pell grants`:`Percent of freshmen receiving state/local grant aid`   2.944 0.003355 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.73 on 650 degrees of freedom
Multiple R-squared:  0.8404,    Adjusted R-squared:  0.836
F-statistic: 190.2 on 18 and 650 DF,  p-value: < 2.2e-16
```

The final model passed all these tests. The model also has a very good accuracy in predicting the value of Graduation Rate, close to 90%.



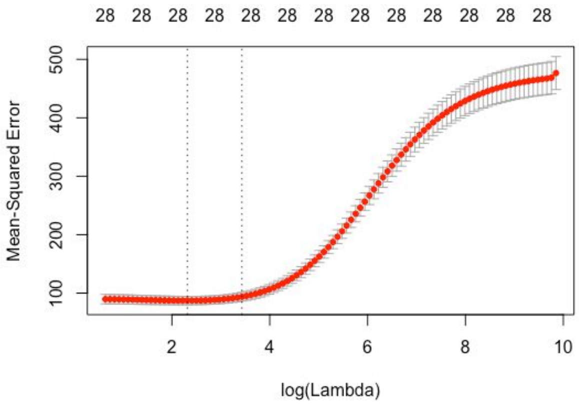
Correlation of 0.9072006

c. Key Observations and Findings

As discussed previously, the Comprehensive Model was a good stepping stone for the team in understanding the impact of most important variables on the Graduation Rate and therefore predicting the value of the Graduation Rate. The two most important variable in predicting the Graduation Rate for a 4 years Bachelor’s degree are “SAT Reading 25th percentile score” and “Tuition and fees, 2013-2014”. Their associated coefficient or slopes are also pretty high in the model shown above, therefore, universities should focus on these two pieces of information to increase their graduation rate. This finding was useful in other models built by team as the other models also had “ACT/SAT scores” and “Tuition and Fees” as leading contributor in their regression model.

7. Analysis of Private, Baccalaureate Institutions: Elizabeth Unzueta

The data set consisted of approximately 500 public institutions and 1,000 private institutions. Exactly 504 are baccalaureate institutions made up of 3 types: Baccalaureate Colleges of Art and Science, Baccalaureate Colleges- Diverse, and Baccalaureate/Associate Colleges. This is an important difference between other institutions and we think that these differences will have an important effect on the four-year graduation rates for students. Based on earlier stages, I began by finding the correlation between my variables. Multicollinearity seemed to play a larger role in private, baccalaureate institutions. A ridge regression was utilized due to very high correlations of about 85% and higher for many of the variables focused on test scores and tuition costs.The ridge plot may be seen on the left with a minimum lambda of around 10 being utilized. Suspicions of multicollinearity were confirmed with negative beta signs where there was a positive correlation found in earlier stages and after finding the VIF and dozens of variables had extremely high VIFs in the hundreds and higher.

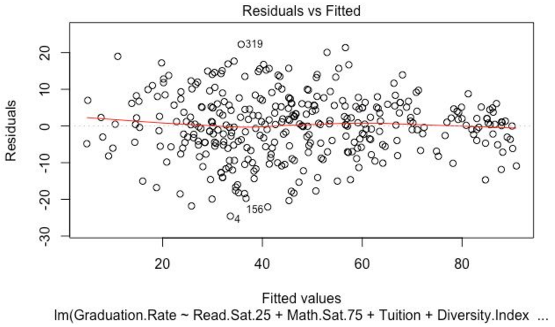


Continuing with the feature selection process, all of the categorical variables were eliminated. While a few were shown to have significance to include in the full dataset; for private, baccalaureate institutions, they did not pass the t-test. After eliminating variables that were highly correlated, backward and forward selection processes were used to cut down on the remaining independent variables. After investigating both and finding comparable models from each method, I choose a backward selection model. Both models contained all SAT/ACT variables, Tuition and fees, Price for in-state students, and the same financial aid variables with differences in the remaining variables. Both had similar adjusted R-squares (“ARS”), still in the 80% range and passing the f-test. After running the test for the VIF, many of the variables had a VIF>10. Many variables were removed until all fell below the 10 threshold: SAT Critical Reading 75th percentile score, SAT Math 25th percentile score, and ACT Composite 25th percentile score. Once these variables were eliminated, multicollinearity was so longer an issue for the model. The chosen backward selection model eliminated many variables including two that were shown to have predictive qualities for the entire dataset: Full Time Undergraduate Enrollment as Percent of Total Undergraduate Enrollment and Percent of freshmen receiving federal grant aid meaning that these variables had an effect on public schools but not for private, baccalaureate institutions.

An important factor before truly eliminating these variables was to try transforming variables. Our response variable, Graduation Rate, had a normal histogram and did not require any transformations. While transformations were tried for variables that had been eliminated up to this point, all were fruitless. None passed the t-test except for Part-time undergraduate enrollment. A logarithmic transformation was used for Part-time undergraduate enrollment. However, this variable was also eliminated as it did not add even half of one percent to our ARS. At this time, Endowment, even with a logarithmic transformation, and was also eliminated due to failing the t-test. Again, the logarithmic transformation of Endowment had an affect on the complete dataset but failed to show predictiveness for this subset.

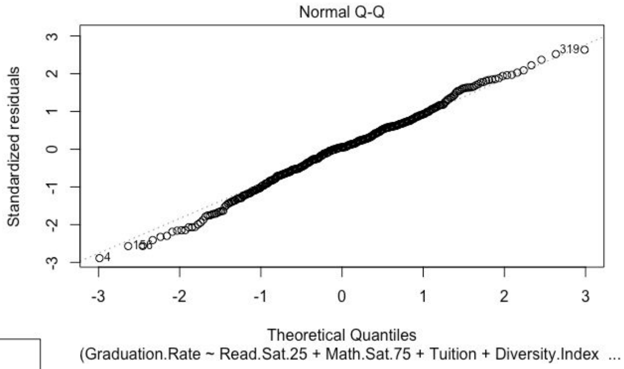
The next step was to run a residual analysis. The sum of the residuals for the final model was very close to zero at $3.19e^{-13}$. While the histogram of the residuals was fairly normal, there were a few outliers noticed and the following rows were eliminated: 16, 127, 189, 261, and 432. The Durbin-Watson test with a p-value of 0.018 showed that the residuals are independent. For private, baccalaureate institutions the following six variables were important for Graduation Rate: SAT Critical Reading 25th percentile score, SAT Math 75th percentile score, Tuition and fees for the 2013-2014 school year, Diversity Index, Percent of undergraduate enrollment that are women, and Percent of freshmen receiving Pell grants. Three of these: SAT Critical Reading 25th percentile score, Tuition and fees for the 2013-2014 school, and Percent of freshmen receiving Pell grants hold steady across Carnegie Class and for type of institution. For this subset, those six variables attributed to about 85% of the variability of Graduation Rate.

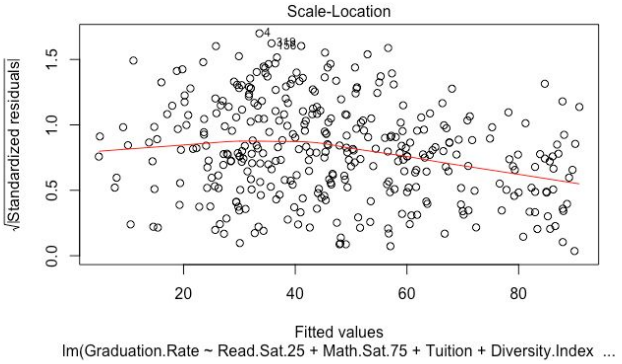
Similar to the whole dataset, no second-order or interaction terms are included in this subset. While interaction terms were considered, for example, Women and the Diversity Index, it did not add value or pass the necessary tests.Also, a second-order term was investigated for Percent of freshmen receiving Pell grants due to a curve in the plot against Graduation Rate. However, this also yielded nothing. Other curvatures were noticed for many of the SAT and ACT variables, and second-order terms were investigated for the remaining testing variables. None passed the t-test.



The Residual vs. Fitted plot shows a variance that is mostly homoscedastic throughout. While the top looks “fuller” than the bottom, it is not enough to say that it is heteroscedastic. Also, the plot isn’t perfect, but it is good enough. I think that if you were to put 2 or 3 points into the spaces on the bottom, it would be close to perfect and show the same variance throughout.

The Normal Q-Q plot shows that are dataset is normal as most of the dots line up on the diagonal. While a few more outliers are shown, it is safe to say that about 95% of the points fall on the line.





This plot shows the fitted values again but not against the square of the standardized residuals. Again,it shows homoscedasticity throughout.

In conclusion, it appears that the response variable, Graduation Rate, has many factors. The most important of which appears to be test scores and tuition price, Most of the remaining independent variables are ones that are easily accessible like tuition cost, test scores, and gender. A few like whether the student receives a Pell grant or the Diversity Index variables, while not at the same reliability level, are still required to be submitted from the schools for funding when gathered by the NCES.

8. Analysis Excluding ACT/SAT Variables-Lesley Bosniack

In the analysis performed on the dataset excluding ACT/SAT variables, additional data processing, multicollinearity and feature reduction, and variable transformation testing is performed before performing regression analysis. The regression analysis was multivariate stepwise similar to all other analysis starting with first order regression with no interaction terms. The analysis was refined to consider non-linear relationships to Graduation Rate, interactions between explanatory variables and removal of outliers. Cross validation and residual analysis supplemented the analysis similar to the analysis of other datasets.

Because the focus of this analysis is non-ACT/SAT variables, ACT/SAT variables are removed from the dataset. Additional data cleaning removed instances with missing values. With the additional processing, the dataset contains 47 variables and 1339 instances.

Collinearity in this dataset appears to exist due to repetition of similar variables, high correlations between variables, and variables being a function of other explanatory variables in the dataset. Correlations (above 0.9) and VIF near/above 10.0 were targeted to identify sets of variables with possible collinearity. Regression results for each set of were compared across models excluding each variable within a set individually. For models with similar ARS and correlations near and above 0.90, variables were considered for removal from regression modeling. Correlations and/or VIFs were significant for First-Time Undergraduate distributions, within Ethnicity distributions, within Enrollment percentages, within Enrollment counts, within Financial Aid variables, and for Tuition/Price variables, 6 different sets of variables possibly exhibiting collinearity. The different sets were prioritized according to VIF size with the largest addressed first. Values were missing for approximately one-third of the instances for Percent of First Time Undergraduate variables. This variable was removed, sacrificing the ARS by approximately four points keeping in mind that the ARS may improve during the modeling process with variable removals, transformation, higher-order terms.

Denominators are the same for the ethnicity distribution variables with the distributions across variables summing to 100%. Removal of Percent of Undergraduate Enrollment that are White is supported by comparing ARS values and minimal effect on F-tests. Correlations between Total Enrollment, Undergraduate Enrollment and Full Time Enrollment were between 0.96-0.98. Removal of each deliver similar ARS values. Total and Full Time Enrollment were chosen for removal based on t-test values for the remaining variables in each regression exercise. Financial aid variables overlapped in their definitions. This contributed to the high correlations between the variables for Federal Student Loans and for Student Loan Aid. In addition, Pell and Federal Grant Aid exhibited near perfect correlations. Student Loan Aid and Pell financial aid variable removals were judgmental and based on t-tests due to identical F-test and ARS. Correlations between Tuition and In(Out)-State Price was high. The Price variables were missing a number of entries. This contributed to their removal over Tuition. Region and State and also Percentage of Part time and Full-time enrollment were identified as alias coefficients (linearly dependent coefficients) by R-Studio leading to the removal Region and Percent of Part time Enrollment based on ARS comparisons. Intuitively the choices for inclusion vs removal made sense. The total effect on the ARS of removing the above variables was a decrease from 0.8475 to 0.8423 with no change to the model p-value. The reduced dataset contains 32 variables, a reduction of fifteen variables.

Per the Data Analysis finding, Tuition, Undergraduate Enrollment, Ethnicity, Financial Aid and Endowment variables were considered for log-transformations. Although t-tests improved for some variables, the ARS changed minimally, with a maximum change of 0.0048. Despite the small ARS change, forward regression analysis was pursued, delivered little/no value and resulting in no additional transformations in the non-ACT/SAT dataset.

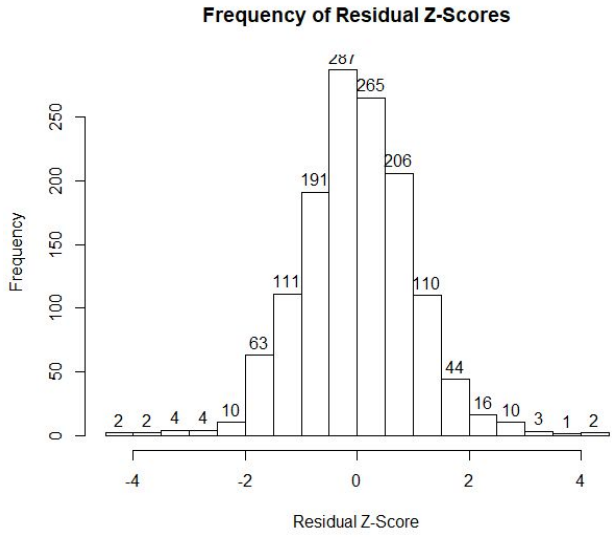
Initial forward and backward regression, with 66% of the data as the training set, were performed on the revised dataset resulting in identical results with 80.33% of the variability explained by the model. F-test, ARS, t-test results for Religious Affiliation suggest the removal of this variable. Its removal increased the p-value of Control of Institution to 0.5 from 0.005.

Further analysis led to the elimination of Control of Institution, Enrollment Yield, Historically Black College or University, and Religious Affiliation. The ARS lowered to 0.7935.

As with the other datasets, residual plots against explanatory variables confirmed lack of need for higher orders for any numeric variables. The Normal Q-Q plot confirmed normality with the exception of the far right and left quartiles. The mean of the residuals is near zero at -0.11, however the variance in the residuals does not appear consistent suggesting some heteroscedasticity. Due to the inconsistent variances, a binomial transformation of Graduation Rate was tested. The transformation very mildly improved the observed heteroscedasticity, the deciding factor in choosing to model Graduation Rate untransformed. Residual and Normal Q-Q plots are below.

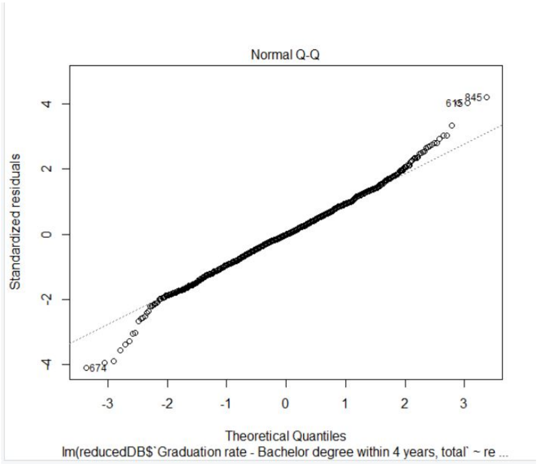
Six Number Summary – Residual Z-Score

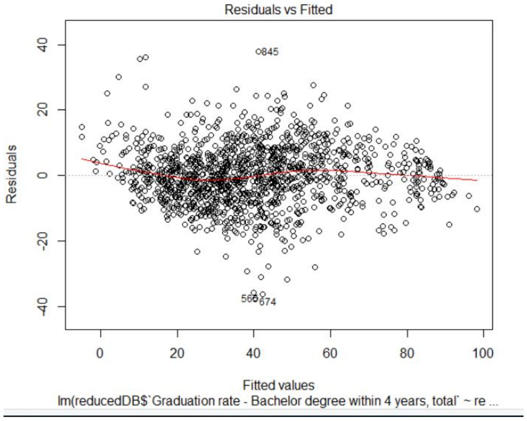
```
           summary(finalModel$residuals)
    Min.  1st Qu.   Median      Mean  3rd Qu.     Max.
-4.13870 -0.62346 -0.01216  0.00000  0.62030  4.186
```



Lastly, the Durbin-Watson test indicated the independence of residuals (p-value of 0.018). Metrics and graphs for residual testing is below.

The review of residuals did highlight outliers in the dataset for various reasons. Eight instances were removed from the analysis. Removing outliers increases ARS by nearly 2 points. Other instances were reviewed for outlier tendencies but did not identify unusual circumstances, and they were not removed from the dataset.





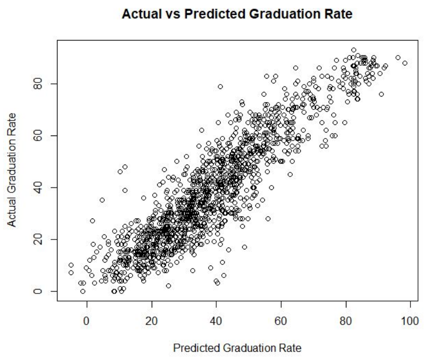
Below is the final model selection.

```
lm(formula = `Graduation rate - Bachelor degree within 4 years, total` ~
 `Admission Rate` + `Tuition and fees, 2013-14` + `FIPS state code` +
 `Degree of urbanization (Urban-centric locale)` +  `Carnegie Classification 2010: Basic` +
 `Full Time Undergraduate Enrollment as Percent of Total Undergraduate Enrollment`+
 `Percent of undergraduate enrollment that are Black or African American` +
 `Percent of undergraduate enrollment that are women` +
 `Percent of freshmen receiving federal, state, local or institutional grant aid` +
 `Percent of freshmen receiving institutional grant aid` +  `Percent of freshmen  receiving federal grant aid` +
`Percent of freshmen receiving federal student loans` + `Endowment - Total` +  `Diversity Index, undergraduate`)
```

```
Coefficients:
                                                              Estimate    Std. Error   t value    Pr(>|t|)
(Intercept)                                                    3.620e+01  4.947e+00   7.318 4.50e-13 ***
reducedDB$`Admission Rate`                                    -7.250e-02  1.818e-02  -3.988 7.05e-05 ***
reducedDB$`Tuition and fees, 2013-14`                         7.292e-04  3.863e-05  18.877  < 2e-16 ***
reducedDB$`FIPS state code`Arizona                            9.725e-01  4.394e+00   0.221 0.824893
reducedDB$`FIPS state code`Arkansas                           3.017e+00  3.137e+00   0.962 0.336418
reducedDB$`FIPS state code`California                         5.378e+00  2.478e+00   2.170 0.030173 *
reducedDB$`FIPS state code`Colorado                          -1.290e-01  3.315e+00  -0.039 0.968963
reducedDB$`FIPS state code`Connecticut                        4.670e+00  3.198e+00   1.461 0.144398
reducedDB$`FIPS state code`Delaware                           5.629e+00  5.823e+00   0.967 0.333912
reducedDB$`FIPS state code`District of Columbia               6.329e+00  4.200e+00   1.507 0.132064
reducedDB$`FIPS state code`Florida                            5.679e+00  2.602e+00   2.182 0.029275 *
reducedDB$`FIPS state code`Georgia                            6.572e+00  2.638e+00   2.491 0.012852 *
reducedDB$`FIPS state code`Hawaii                            -7.463e+00  4.425e+00  -1.687 0.091929 .
reducedDB$`FIPS state code`Idaho                              6.462e-01  4.419e+00   0.146 0.883774
reducedDB$`FIPS state code`Illinois                           8.215e+00  2.537e+00   3.238 0.001237 **
reducedDB$`FIPS state code`Indiana                            4.364e+00  2.631e+00   1.659 0.097446 .
reducedDB$`FIPS state code`Iowa                               5.189e+00  2.845e+00   1.824 0.068419 .
reducedDB$`FIPS state code`Kansas                             5.191e+00  3.000e+00   1.731 0.083779 .
reducedDB$`FIPS state code`Kentucky                           2.932e+00  2.895e+00   1.013 0.311343
reducedDB$`FIPS state code`Louisiana                         -1.397e+00  2.945e+00  -0.474 0.635300
reducedDB$`FIPS state code`Maine                              5.784e+00  3.300e+00   1.753 0.079897 .
reducedDB$`FIPS state code`Maryland                           8.256e+00  2.932e+00   2.816 0.004939 **
reducedDB$`FIPS state code`Massachusetts                      7.167e+00  2.538e+00   2.824 0.004825 **
reducedDB$`FIPS state code`Michigan                          -6.973e-01  2.636e+00  -0.265 0.791436
reducedDB$`FIPS state code`Minnesota                          9.327e+00  2.786e+00   3.348 0.000838 ***
reducedDB$`FIPS state code`Mississippi                        8.572e+00  3.216e+00   2.666 0.007780 **
reducedDB$`FIPS state code`Missouri                           5.837e+00  2.713e+00   2.152 0.031609 *
reducedDB$`FIPS state code`Montana                           -5.839e+00  4.199e+00  -1.390 0.164635
reducedDB$`FIPS state code`Nebraska                          -3.796e-01  3.432e+00  -0.111 0.911930
reducedDB$`FIPS state code`Nevada                            -2.822e+00  5.873e+00  -0.480 0.631030
reducedDB$`FIPS state code`New Hampshire                      6.289e+00  3.615e+00   1.740 0.082147 .
reducedDB$`FIPS state code`New Jersey                         7.186e+00  2.892e+00   2.485 0.013088 *
reducedDB$`FIPS state code`New Mexico                        -1.810e+00  4.791e+00  -0.378 0.705659
reducedDB$`FIPS state code`New York                           1.000e+01  2.361e+00   4.238 2.42e-05 ***
```

```
reducedDB$`FIPS state code`North Carolina                                          7.217e+00  2.493e+00   2.894 0.003864 **
reducedDB$`FIPS state code`North Dakota                                            -6.776e+00  4.184e+00  -1.619 0.105600
reducedDB$`FIPS state code`Ohio                                                    7.411e+00  2.584e+00   2.868 0.004197 **
reducedDB$`FIPS state code`Oklahoma                                                4.247e+00  3.029e+00   1.402 0.161183
reducedDB$`FIPS state code`Oregon                                                  4.345e+00  3.079e+00   1.411 0.158431
reducedDB$`FIPS state code`Pennsylvania                                             8.582e+00  2.387e+00   3.595 0.000337 ***
reducedDB$`FIPS state code`Rhode Island                                            9.748e+00  3.980e+00   2.450 0.014438 *
reducedDB$`FIPS state code`South Carolina                                          1.109e+01  2.755e+00   4.026 6.01e-05 ***
reducedDB$`FIPS state code`South Dakota                                            2.255e+00  3.780e+00   0.597 0.550941
reducedDB$`FIPS state code`Tennessee                                               4.622e+00  2.726e+00   1.696 0.090212 .
reducedDB$`FIPS state code`Texas                                                   4.205e+00  2.461e+00   1.709 0.087765 .
reducedDB$`FIPS state code`Utah                                                    -5.339e+00  4.715e+00  -1.132 0.257682
reducedDB$`FIPS state code`Vermont                                                 1.946e+00  3.513e+00   0.554 0.579725
reducedDB$`FIPS state code`Virginia                                                6.032e+00  2.575e+00   2.343 0.019310 *
reducedDB$`FIPS state code`Washington                                              1.043e+01  3.058e+00   3.411 0.000667 ***
reducedDB$`FIPS state code`West Virginia                                            -9.081e-03  3.234e+00  -0.003 0.997760
reducedDB$`FIPS state code`Wisconsin                                               -6.326e-01  2.759e+00  -0.229 0.818688
reducedDB$`FIPS state code`Wyoming                                                 -4.933e+00  9.725e+00  -0.507 0.612059
reducedDB$`Degree of urbanization (Urban-centric locale)`City: Midsize              3.436e+00  1.024e+00   3.356 0.000815 ***
reducedDB$`Degree of urbanization (Urban-centric locale)`City: Small               2.726e+00  9.921e-01   2.748 0.006079 **
reducedDB$`Degree of urbanization (Urban-centric locale)`Rural: Distant            2.238e+00  2.171e+00   1.031 0.302934
reducedDB$`Degree of urbanization (Urban-centric locale)`Rural: Fringe             2.497e+00  1.644e+00   1.519 0.128970
reducedDB$`Degree of urbanization (Urban-centric locale)`Rural: Remote              6.781e+00  2.979e+00   2.277 0.022971 *
reducedDB$`Degree of urbanization (Urban-centric locale)`Suburb: Large             1.364e+00  9.242e-01   1.475 0.140347
reducedDB$`Degree of urbanization (Urban-centric locale)`Suburb: Midsize           3.803e+00  1.611e+00   2.360 0.018411 *
reducedDB$`Degree of urbanization (Urban-centric locale)`Suburb: Small             3.728e+00  1.857e+00   2.008 0.044882 *
reducedDB$`Degree of urbanization (Urban-centric locale)`Town: Distant             5.323e+00  1.105e+00   4.818 1.63e-06 ***
reducedDB$`Degree of urbanization (Urban-centric locale)`Town: Fringe              2.481e+00  1.479e+00   1.678 0.093656 .
reducedDB$`Degree of urbanization (Urban-centric locale)`Town: Remote              3.727e+00  1.232e+00   3.025 0.002535 **
reducedDB$`Carnegie Classification 2010: Basic`Baccalaureate Colleges--Diverse Fields       -5.788e+00  9.751e-01  -5.936 3.78e-09 ***
reducedDB$`Carnegie Classification 2010: Basic`Baccalaureate/Associate's Colleges          -2.864e+00  2.288e+00  -1.252 0.210875
reducedDB$`Carnegie Classification 2010: Basic`Doctoral/Research Universities              -5.894e+00  1.400e+00  -4.210 2.74e-05 ***
reducedDB$`Carnegie Classification 2010: Basic`Master's Colleges and Universities (larger programs)  -4.832e+00  9.402e-01  -5.140 3.19e-07 ***
reducedDB$`Carnegie Classification 2010: Basic`Master's Colleges and Universities (medium programs) -5.522e+00  1.123e+00  -4.915 1.00e-06 ***
reducedDB$`Carnegie Classification 2010: Basic`Master's Colleges and Universities (smaller programs) -6.899e+00  1.335e+00  -5.168 2.75e-07 ***
reducedDB$`Carnegie Classification 2010: Basic`Research Universities (high research activity)       -2.740e+00  1.332e+00  -2.058 0.039830 *
reducedDB$`Carnegie Classification 2010: Basic`Research Universities (very high research activity)    5.450e+00  1.292e+00   4.220 2.62e-05 ***
reducedDB$`Full Time Undergraduate Enrollment as Percent of Total Undergraduate Enrollment`         2.120e-01  2.518e-02   8.420  < 2e-16 ***
reducedDB$`Percent of undergraduate enrollment that are Black or African American`         -6.700e-02  2.440e-02  -2.746 0.006116 **
reducedDB$`Percent of undergraduate enrollment that are women`                             1.849e-01  2.490e-02   7.427 2.05e-13 ***
reducedDB$`Percent of freshmen receiving federal, state, local or institutional grant aid`          -1.340e-01  3.411e-02  -3.928 9.03e-05 ***
reducedDB$`Percent of freshmen receiving institutional grant aid`                          6.801e-02  2.178e-02   3.123 0.001831 **
reducedDB$`Percent of freshmen  receiving federal grant aid`                                -3.891e-01  3.028e-02 -12.851  < 2e-16 ***
reducedDB$`Percent of freshmen receiving federal student loans`                            -7.277e-02  2.327e-02  -3.128 0.001802 **
reducedDB$`Endowment - Total`                                                               7.087e-06  2.148e-06   3.299 0.000997 ***
reducedDB$`Diversity Index, undergraduate`                                                  -2.552e-01  3.593e-02  -7.103 2.04e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

F-test p-value is 2.2e-16. ARS is 0.8086. The correlation between actual and predicted Graduation Rate is 0.8962. The relationship between actual and predicted values is linear, positive and strong.



The variance (and coefficient of variation) in residuals is small as well which indicated no concern of overfitting. In addition, many more than 95% of instances were within 2 standard deviations of the mean as represented in the Normal Q-Q plot above.

Variables contributing to the prediction of Graduation Rates are admission rates, tuition and endowments, state and urban location of the school, ethnicity/race and financial aid. Most findings thus far are intuitive: Graduation rates appear to be inversely related to admission rates. Schools with lower admission rates are perhaps the more selective schools, delivering higher graduation rates. Higher tuition and endowment schools, possible proxies for quality and support of future development of the school's capabilities, return higher Graduation Rates. Higher full-time enrollment percentages indicate higher Graduation Rates. Schools with higher distributions of Blacks/African Americans, unknown race/ethnicity and Nonresident Aliens and schools with higher diversity indices contribute less favorably to Graduations Rates. The percentage of students receiving most types of financial aid is inversely related to Graduation Rate with the exception of institutional grant aid. For categorical variables, the most significant states to the model are Illinois, Maryland, Minnesota, New York North Carolina and Washington with Illinois contributing the most favorably to graduation rate. Of the four levels of urbanization, cities and towns are the most significant to the model. Midsize schools contribute more favorably than small schools to Graduation Rate for all levels of urbanization. All but one Carnegie Classification is significant to the model, Baccalaureate/Associate Colleges. Research universities contribute most favorably to Graduation Rates while schools with Master's degrees having smaller programs and Baccalaureate schools of diverse fields contribute the less favorably.

9. Executive Summary

Given the significance of factoring Graduation Rate into the decision-making of universities and students, this study focuses on understanding the influence of various factors on the Graduation Rate of an institution, and based on this understanding, predicting Graduation Rates. We built our domain knowledge based on the review of correlations of the response variable to explanatory variables, correlations/multicollinearity analysis across explanatory variables, visualization of variable distributions and relationships between variables, and initial regression analysis. Through scaling, recategorization of categorical variables and other variable transformations, variance inflation analysis and further regression analysis, we narrowed down the explanatory variables considered for analysis. Multivariate linear regression was performed with data segregated into training and test sets. Higher-order terms and interaction terms along with further transformations of explanatory variables and of the response variable were tested to arrive at the models chosen. F-test, t-test, and adjusted R-square metrics along with cross-validation and residual analysis exercises, and also intuition, led to the selection of the chosen models that explains the variability in and predicts Graduation Rates. The chosen models are based not only on their statistical soundness and ability to explain and predict Graduation Rates but also because of their intuitive findings, reasonable level of complexity of modeling required, relative ease in programming the prediction of future Graduation Rates and manageable access to the data fields for future replication due to the reporting requirements that the NCES has set.

The steps listed below are important building blocks for the model creation: identify the explanatory variables that may contribute to explaining the variability in Graduation Rate variables, review summary distributions of all the variables, understand relationships between explanatory variables including multicollinearity,test and employ variable transformations, create an initial linear regression model to predict the Graduation Rate, introduce second order and interaction terms as valuable, develop model with different model selection techniques (ie. forward, backward), segment the dataset instances for model training/validation testing, assess the quality of the model results and identify outlier instances through cross validation and residuals analysis, refine dataset and model as necessary, create and finalize model for prediction of future data.

Modeling is performed with the entire dataset but also with subsets of the data. Because initial analysis indicated differences in Graduation Rates by Control of Institution (public vs private schools) and across Carnegie Classifications, the dataset was segmented into three subsets to test the similarities and differences in models for different Controls/Classification combinations. Additionally, because the number of instances in the dataset was reduced by approximately 15% due to missing SAT/ACT scores, a modeling exercise was performed excluding SAT/ACT variables to analyze the impact of the instance reduction for those datasets that include SAT/ACT explanatory variables.

Per the models built by the team members in their independent analyses, we observe that the SAT/ACT variables and Tuition are key variables contributing to Graduation Rate across all datasets with higher SAT/ACT scores and tuition levels delivering higher Graduation Rates. This finding is consistent with earlier analysis of correlations between explanatory variables and Graduation Rate. Specifically, “SAT Critical Reading 25th percentile score” and “SAT Math 75th percentile score” offer the greatest contributions with respect to standardized test scores. Isolating private schools that educate at the baccalaureate level, fewer and different additional variables contribute to predicting Graduation Rate than public schools and schools that offer learning above the baccalaureate level. In focusing on non-SAT/ACT variable modeling, many more variables were necessary to explain how Graduation Rate varies by school. All models chosen in the individual analysis of the team members, were able to explain at least 80% of the variability in Graduation Rate across schools. Relationships between the explanatory variables and Graduation Rate were almost exclusively linear with the exception of the Endowment in the comprehensive model, and interactions between variables that were not eliminated due to collinearity at early stages of modeling were few and far between.

The “SAT Critical Reading 25th percentile score” gives the 25th percentile score of a university for the Critical Reading section of the standardized SAT test. This variable has an associated coefficient (slope) of 4.083e-02 in regression analysis of the comprehensive dataset. A one-point increase in the “SAT Critical Reading 25th percentile score” translates into approximately 0.041% increase in the Graduation Rate. Intuitively, this makes sense as students with a higher critical reading score should be better able to perform in higher level education and have a greater chance of achieving an undergraduate degree.

"Tuition and fees, 2013-2014" plays a critical role in predicting Graduation Rate. In the comprehensive model, the coefficient for the tuition and fees is 4.88e-04. Schools with higher tuition and fees have higher Graduation Rates at a rate of approximately 4.8 points per $100 of tuition and fees. One theory that may explain this finding is that the level of the tuition and fees is representative of the quality of the education received with higher priced schools delivering greater quality of education and contributing to student success and higher Graduation Rates. Another viable conclusion may be that students with the resources to attend more expensive universities have less impediments to graduation.

In isolating private, baccalaureate institutions, the key variables contributing to Graduation Rate were similar to those of the full dataset. Tuition and fees, SAT Critical Reading 25th percentile scores and SAT Math 75th percentile score played an important role. This means, that for private institutions offering only bachelor's degrees, the most important factors in picking students who will graduate within four years are these two testing variables and tuition costs. While other dataset segmentation found that Full Time Undergraduate Enrollment and Endowment amounts were valuable predictors, these variables did not have significant p-values for private, baccalaureate institutions. In the evaluation of the public schools, the Average SAT scores of admitted students was identified as a significant variable. Lastly, Endowment, found to be valuable in the modeling of the comprehensive dataset, did not pass the t-test when tested in the modeling of private, baccalaureate institutions regardless of whether untransformed or log-transformed.

Consistent with the modeling performed with other dataset segmentations and the comprehensive dataset, tuition is a key explanatory variable in the modeling when SAT/ACT variables are not considered as explanatory variables. Similar to the public and private, baccalaureate schools, the percent of women attending and diversity were found to predict Graduation Rate. The amount of Endowments untransformed, Carnegie Classification and the percent of full-time students were in common with the comprehensive analysis. A number of additional variables were found to be valuable to the explaining Graduation Rate: admission rate, state for largely populated states, some urbanization categories, Black or African/American ethnicities and a number of financial aid variables other than Pell. Excluding SAT/ACT variables from the analysis did not diminish the ability to explain and predict Graduation Rate. The exclusion of the SAT/ACT variables allowed for a larger number of instances to be included in the analysis, however many more variables and more of the categorical variables were necessary to the model. Why the composition of these alternative variables can deliver similar predictive ability is unclear. The findings of the analysis excluding SAT/ACT raises the question of whether these other additional variables act as a proxy for the SAT/ACT, if the strength of the SAT/ACT score to the model diminishes the value of some non-SAT/ACT variables, or if the instances with no SAT/ACT score in other datasets compromised the quality of other modeling.

10. Conclusions

In conclusion, several factors appear to contribute to the prediction for four-year undergraduate Graduation Rates. From the NCES data provided on both academic institutions and profiles of students attending these institutions, we found factors relating to SAT/ACT scores, tuition, Control of Institution, Carnegie Classification, geographical location, full time vs part time status, financial aid, amount of endowments, gender, admission rate and race/ethnicity contribute to explaining the variability in Graduation Rates in some/all models with SAT/ACT and Tuition being key variables and most relationships of explanatory variables to Graduation Rate being linear.

Appendix

A: DATA DICTIONARY

Attribute	**Data Type**	**Explanatory Category**	**Brief Description**
ID number	Unique Identifiers	Unique Identifier	Six-digit number assigned to each institution
Name	Unique Identifiers	Unique Identifier	Name of Institution
Highest degree offered	Categorical	School Descriptor	The highest degree offered by the institution
Religious affiliation	Categorical	School Descriptor	Religious affiliation (denomination) for private not-for-profit institutions
SAT Critical Reading 25th percentile score	Numerical	Exam Score	25th percentile score for incoming undergraduate students in reading
SAT Critical Reading 75th percentile score	Numerical	Exam Score	75th percentile score for incoming undergraduate students in reading
SAT Math 25th percentile score	Numerical	Exam Score	25th percentile score for incoming undergraduate students in math
SAT Math 75th percentile score	Numerical	Exam Score	75th percentile score for incoming undergraduate students in math
Average SAT Equivalent Score of Students Admitted	Numerical	Exam Score	Average score from College Scorecard
ACT Composite 25th percentile score	Numerical	Exam Score	25th percentile composite score for incoming undergraduate students
ACT Composite 75th percentile score	Numerical	Exam Score	75th percentile composite score for incoming undergraduate students

Admission Rate	Numerical	Actual	Admission - Total from the original dataset / Applicants - Total
Admissions yield - total	Numerical	Actual	Enrolled - Total / Admission -Total from the original dataset
Tuition and fees, 2013-14	Numerical	Tuition and Fees	Cost of tuition and fees for the 2013-2014 school year for first-time, full-time undergraduate students including those enrolled for associate’s or other vocational Degrees
Total price for in-state students living on campus 2013-14	Numerical	Price	Cost of enrollment including tuition and other school fees ie books for first-time undergraduate students from the same state as the institution
Total price for out-of-state students living on campus 2013-14	Numerical	Price	Cost of enrollment including tuition costs and other school fees ie room and board for students not from the same state as the institution
FIPS state code	Categorical	School Descriptor	US Postal Service State Abbreviation for location of school
Geographic region	Categorical	School Descriptor	Geographic region categories are: US Service schools, New England CT ME MA NH RI VT, Mid East DE DC MD NJ NY PA, Great Lakes IL IN MI OH WI, Plains IA KS MN MO NE ND SD, Southeast AL AR FL GA KY LA MS NC SC TN VA WV, Southwest AZ NM OK TX, Rocky Mountains CO ID MT UT WY, Far West AK CA HI NV OR WA, Outlying areas AS FM GU MH MP PR PW VI
Control of institution	Categorical	School Descriptor	Control categories are public, private not-for-profit, and private for-profit.
Historically Black College or University	Binary	School Descriptor	Institution that is one of the Historically Black College or University (HBCU) institutions.

Degree of urbanization (Urban-centric locale)	Categorical	School Descriptor	Identify the geographic status of the school on the range of urbanization (large city to rural) based on the schools physical address. **City: Large/Midsize/Small:** Territory inside an urbanized area and inside a principal city with population of >=250,000 / < less than 250,000 and >= 100,000 / <100,000. **Suburb: Large/Midsize/Small**: Territory outside a principal city and inside an urbanized area with population of >=250,000 / **<** 250,000 and >= 100,000 / < 100,000. **Town: Fringe/Distant/Remote**: Territory inside an urban cluster that is <= 10 miles / > 10 miles and <= 35 miles / > 35 miles of an urbanized area. **Rural: Fringe/Distant/Remote**: Census-defined rural territory that is <= 5 miles from an urbanized area & rural territory that is <= 2.5 miles from an urban cluster / > 5 miles and <= 25 miles & > 2.5 miles and <= 10 miles / > 25 miles & > 10 miles. Data through 2004 is used to categorize.
Carnegie Classification 2010: Basic	Categorical	School Descriptor	Six parallel classifications: Basic Classification (the traditional Carnegie Classification Framework), Undergraduate and Graduate Instructional Program classifications, Enrollment Profile and Undergraduate Profile classifications, and Size & Setting classification. These classifications are time-specific snapshots of institutional attributes and behavior based on data from 2008 to 2010. Complete description and technical details visit the Carnegie Foundation Website at www.carnegiefoundation.org/classifications.
Total enrollment	Numerical	Actual	Total men and women enrolled for credit in the fall of the 2013 academic year. (Keeping so we have the base number used for undergrad percentage)
Undergraduate enrollment	Numerical	Actual	Total men and women enrolled for credit in the 2013 academic year in a 4 or 5 year program
Full-time undergraduate enrollment	Numerical	Actual	Total men and women enrolled for full time credit in the 2013 academic year.
Part-time undergraduate enrollment	Numerical	Actual	Total men and women enrolled for full time credit in the 2013 academic year

Undergraduate Enrollment as a Percent of Total Enrollment	Numerical	Actual	Undergraduate enrollment / Total enrollment
Full Time Undergraduate Enrollment as a Percent of Total Undergraduate Enrollment	Numerical	Actual	Full-time undergraduate enrollment / Undergraduate enrollment
Percent of undergraduate enrollment that are American Indian or Alaska Native	Numerical	Student Descriptor	Percent of undergraduate students that are American Indian or Alaska Native in the 2013 academic year. American Indian or Alaska Native - A person having origins in any of the original peoples of North and South America (including Central America) who maintain cultural identification through tribal affiliation or community attachment. Collected in surveys. Ratios are converted to percentages by 100 and then are rounded to whole numbers.multiplying.
Percent of undergraduate enrollment that are Asian	Numerical	Student Descriptor	Percent of undergraduate enrollment that are Asian Asian - A person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian Subcontinent, including, for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam. Collected in surveys. Ratios are converted to percentages by multiplying by 100 and then are rounded to whole numbers.
Percent of undergraduate enrollment that are Black or African American	Numerical	Student Descriptor	Percent of undergraduate students that are Black or African American in the fall of the academic year. Black or African American (new definition) - A person having origins in any of the black racial groups of Africa. Collected in surveys. Ratios are converted to percentages by multiplying by 100 and then are rounded to whole numbers.

Percent of undergraduate enrollment that are Hispanic/Latino	Numerical	Student Descriptor	Percent of undergraduate students that are Hispanic/Latino in the fall of the academic year. Hispanic or Latino (new definition) - A person of Cuban, Mexican, Puerto Rican, South or Central American, or other Spanish culture or origin, regardless of race. Collected in surveys. Ratios are converted to percentages by multiplying by 100 and then are rounded to whole numbers.
Percent of undergraduate enrollment that are Native Hawaiian or Other Pacific Islander	Numerical	Student Descriptor	Percent of undergraduate enrollment that are Native Hawaiian or Other Pacific Islander Native Hawaiian or Other Pacific Islanders (new definition) - A person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands. Collected in surveys. Ratios are converted to percentages by multiplying by 100 and then are rounded to whole numbers.
Percent of undergraduate enrollment that are White	Numerical	Student Descriptor	Percent of undergraduate students that are White in the fall of the academic year. White (new definition) - A person having origins in any of the original peoples of Europe, the Middle East, or North Africa. Collected in surveys. Ratios are converted to percentages by multiplying by 100 and then are rounded to whole numbers.
Percent of undergraduate enrollment that are two or more races	Numerical	Student Descriptor	Percent of undergraduate students that are of two or more races in the fall of the academic year. Two or more races - Category used by institutions to report persons who selected more than one race. Collected in surveys. Ratios are converted to percentages by multiplying by 100 and then are rounded to whole numbers.
Percent of undergraduate enrollment that are Race/ethnicity unknown	Numerical	Student Descriptor	Percent of undergraduate students that are race/ethnicity unknown in the fall of the academic year. RACE/ETHNICITY UNKNOWN - This category is used ONLY if the student did not select a racial/ethnic designation, AND the postsecondary institution finds it impossible to place the student in one of the aforementioned racial/ethnic categories during established enrollment procedures or in any post-enrollment identification or verification process. Ratios are converted to percentages by multiplying by 100 and then are rounded to whole numbers.

Percent of undergraduate enrollment that are Nonresident Alien	Numerical	Student Descriptor	Percent of undergraduate students that are nonresident alien in the fall of the academic year. This variable is derived from the enrollment component. NONRESIDENT ALIEN - A person who is not a citizen or national of the United States and who is in this country on a visa or temporary basis and does not have the right to remain indefinitely. NOTE - Nonresident aliens are included here, rather than in any of the five racial/ethnic categories described below. Resident aliens and other eligible (for financial aid purposes) non-citizens who are not citizens or nationals of the United States and who have been admitted as legal immigrants for the purpose of obtaining permanent resident alien status (and who hold either an alien registration card (Form I-551 or I-151), a Temporary Resident Card (Form I-688), or an Arrival-Departure Record (Form I-94) with a notation that conveys legal immigrant status such as Section 207 Refugee, Section 208 Asylee, Conditional Entrant Parolee or Cuban-Haitian) are to be reported in the appropriate racial/ethnic categories along with United States citizens. Ratios are converted to percentages by multiplying by 100 and then are rounded to whole numbers.

Diversity Index	Numerical	Actual	FractionNoneReported=TotalNoneReported/TotalReported
			FractionAmIndian=(TotalAmIndian/(1-FractionNoneReported))/TotalReported
			FractionAsian=(TotalAsian/(1-FractionNoneReported))/TotalReported
			FractionAfricanAm=(TotalAfricanAm/(1-FractionNoneReported))/TotalReported
			FractionHispanic=(TotalHispanic/(1-FractionNoneReported))/TotalReported
			FractionPacIslander=(TotalPacIslander/(1-FractionNoneReported))/TotalReported
			FractionNonresident=(TotalNonresident/(1-FractionNoneReported))/TotalReported
			FractionWhite=(TotalWhite/(1-FractionNoneReported))/TotalReported
			Fraction2orMore=(Total2orMore/(1-FractionNoneReported))/TotalReported
			D(8) = SQRT(((FractionAfricanAm-(1/8))2) + ((FractionAmIndian-(1/8))2) + ((FractionAsian-(1/8))2) + ((FractionHispanic-(1/8))2)+ ((FractionFilipino-(1/8))2) + ((FractionPacIslander-(1/8))2) + ((FractionWhite-(1/8))2) + ((Fraction2orMore-(1/8))2))
			C1 = 100
			C2 = -100*SQRT(8*(8-1))/(8-1) = -100*SQRT(56)/7 = -106.90449
			EDI = C1 + (C2*D(8)) = 100 + (-106.90449*D(8))

Percent of undergraduate enrollment that are women	Numerical	Student Descriptor	Percent of undergraduate students who are women calculated by women undergraduate enrollment/ total enrollment for undergraduates *100 and rounded to the nearest whole number.
Percent of first-time undergraduates - in-state	Numerical	Student Descriptor	Percent of undergraduate students who live in the same state as the institution calculated by the In-State Enrollment/ Total-Enrollment of undergraduates * 100
Percent of first-time undergraduates - out-of-state	Numerical	Student Descriptor	Percent of students who do not live in the same state as an institution, a foreign country, or with unknown residence calculated by the Out-of-State enrollment/ Total-Enrollment of undergraduates *100
Percent of first-time undergraduates	Numerical	Student Descriptor	Percent of first- time undergraduates who live in a foreign country calculated by Number of students from foreign countries divided by the total enrollment of First-Time students * 100

- foreign countries			
Percent of first-time undergraduates - residence unknown	Numerical	Student Descriptor	Percent of Undergraduate students calculated as Percent of First-Time Undergraduates - residence unknown but living in the United States / Total Enrollment of First-Time Undergraduates * 100
Graduation rate - Bachelor degree within 4 years, total	Numerical	Target Variable	Response Variable: Number of Undergraduate Students completing a bachelor's or equivalent degree divided by adjusted bachelor cohort (students subtracted for circumstances ie death) in 4 years or less.
Percent of freshmen receiving institutional grant aid	Numerical	Financial Aid	Percent of freshmen receiving institutional grant aid calculated as Percent of freshmen receiving financial aid/ Total-enrollment of undergraduates *100
Endowment assets (year end) per FTE enrollment	Numerical	Endowment	Endowment Assets (gross investments of funds from endowment) divided by the 12-month full time enrollment for public institutions only, GASB and FASB versions of this variable are combined.