Analysis of Student Loan Repayment Rates

In Partial Fulfillment of

the Requirements for Completion of

the Microsoft Professional Program

for Data Science

by

Hiram S. Foster

July 2017

Twitter: @hiramf05 LinkedIn: hiramf Email: hiramfoster.co@gmail.com

Executive Summary

This capstone project is the culmination of all the skills imparted throughout the Microsoft Professional Program for Data Science. For the July 2017 Capstone (DAT102X), the goal was to "predict the repayment rate for the student loans given to students at United States institutions of higher education."[1] Each Data Science student individually explored and analyzed the data, then built a machine learning model predicting the repayment rate for higher-education institutions. Results from a test-set were posted a course leaderboard, ranking models according to their Root Mean Square Error (RMSE). There were 750 participants and the model discussed here scored 4th overall with an RMSE of 6.32 (as of July 21, 2017).

The first task was to explore and analyze the data according to principles learned in earlier courses such as statistical inference, analytic visualization, and data cleaning. According to the website:

> *There are 443 variables in this dataset. Each row in the dataset represents a United States institution of higher education and a specific year.... We don't provide a unique identifier for an individual institution, just a `row_id` for each row.*[1]

The 443 features were separated into eight categories: Academics, Admissions, Aid, Completion, Cost, Report Year, School and Student, and were a mixture between continuous numerical data, categorical text data, and integer denoted categories. After analyzing and visualizing the data, strong relationships were identified and new features were engineered and added to the dataset. Finally, a boosted tree regression model was created to predict student loan repayment rates.

Many of the features influenced or had strong linear correlation with student loan repayment rates. However, the most significant features were found using Microsoft Azure's Permutation Feature Importance Module and by plotting the feature importance attributes from Python modules (scikit-learn and xgboost). The following patterns emerged as significant predictors for repayment rates:

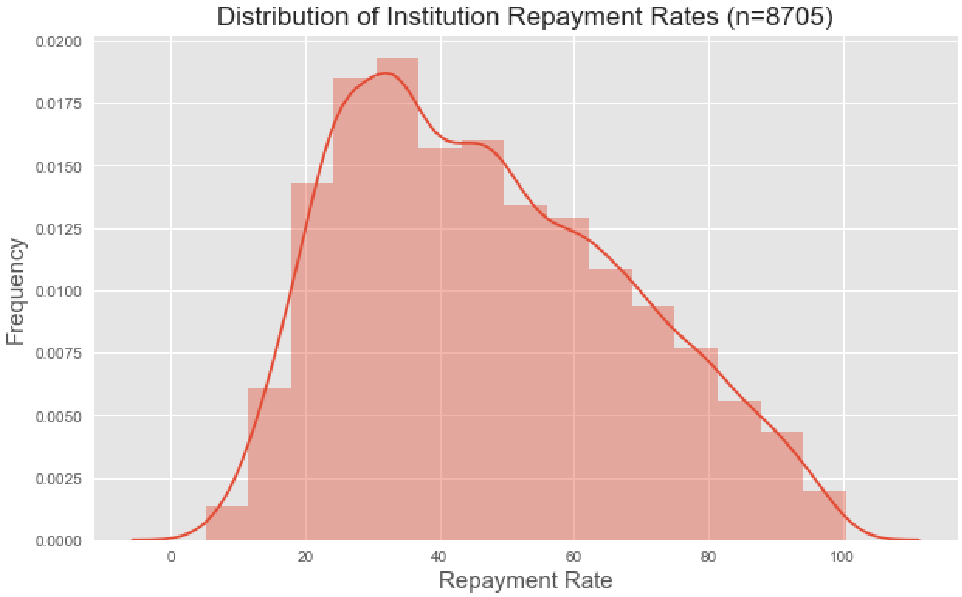
- **Student Family Income**: Institutions with a higher average family income (r = 0.83)* have higher repayment rates. Institutions with a higher percentage of low income families are strongly correlated with low repayment rates (r = -0.84)
- **Demographics**: Institutions with more Black students have lower repayment rates (r = -0.50). Those with a higher percentage of Asian students tend to have higher repayment rates (r = 0.31)*. Additionally, age of entry (r = -0.56)* and percentage of students that are married (r = -0.42)* are correlated with lower repayment rates.
- **Student Financial Aid**: Pell grants are negatively correlated with lower repayment rates, the more students that receive Pell grants at a school, the lower the institution repayment rate (r = -0.79) . However, there is a positive correlation between loan principal amount and repayment rate (r = 0.43).
- **Institution Attributes**: Certain types of institutions have lower repayment rates—four-year, selective schools have higher repayment rates. The location of the institution also affects repayment rates. For example, institutions in New England have higher repayment rates.

* After logarithmic transformation

[1] Microsoft – DAT102X: Predict Student Loan Repayment. Data Science Capstone Challenge. https://www.datasciencecapstone.org/competitions/1/student-loans. Accessed July 17, 2017.

Data Exploration

The target variable for this model was student loan repayment rate, which was measured on a scale of 0 to 100, representing the percentage of students repaying their loans. The distribution is slightly right-skewed, demonstrating that most institutions have a low repayment rate. A large standard deviation indicates considerable variation in repayment rates. Nearly every feature contained many missing values, however, the machine learning algorithm used (xgboost) is able to handle null values. Therefore, imputing missing values would often result in adding noise and false information into the data. For features undergoing logarithmic transformation, missing or zero values were transformed into log(1) because log(0) is undefined. For some categorical features, the presence of a non-missing value was imputed as one and all other rows imputed as zero. Finally, a new feature was created indicating the total number of missing categorical values for each institution.



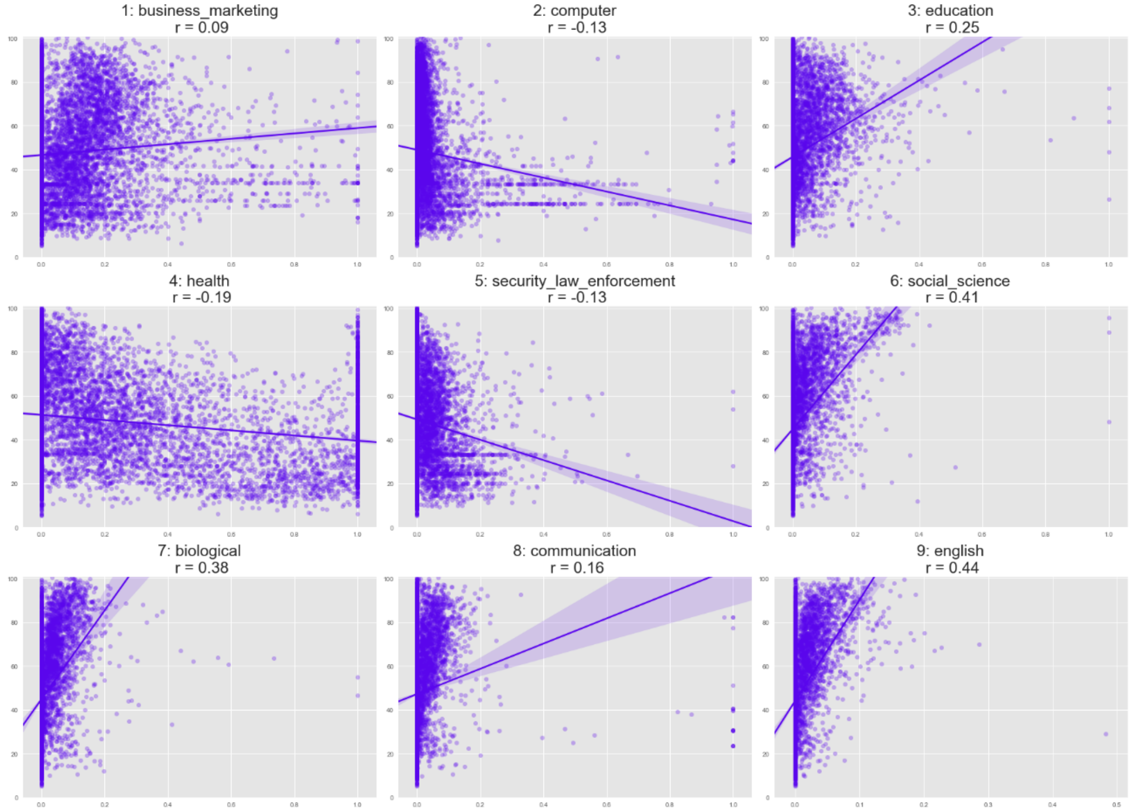
Mean	47.37
Std	20.99
Min	5.16
25%	30.23
50%	44.86
75%	62.62
Max	100.47

Feature Categories

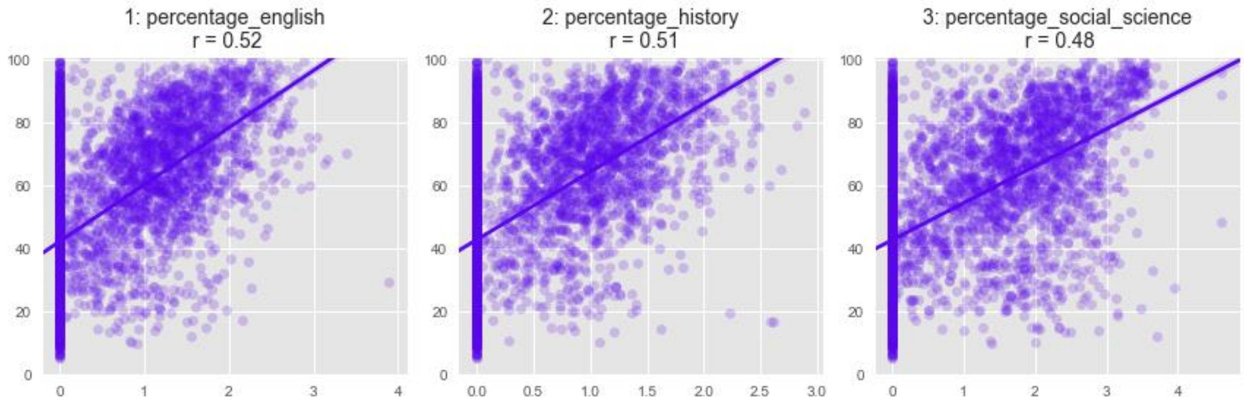
There were 230 numerical features and 213 categorical features from 7 different categories: Academics, Admissions, Aid, Completion, Cost, School, and Student. By writing code to create a grid of scatterplots, it was easy to quickly identify features with linear relationships to repayment rate. Because of the considerable number of numerical features (n_features = 230) it would be impractical to create a scatterplot matrix to check for collinearity between all features. For feature selection, it would also be largely unnecessary since the algorithm used is robust and able to perform feature selection on its own. In this case, checking for collinearity after significant predictors are identified might provide some explanatory power. Categorical features were explored using boxplots and violin plots. For each category of features, notable or characteristic examples from are shown with some notes.

Academics

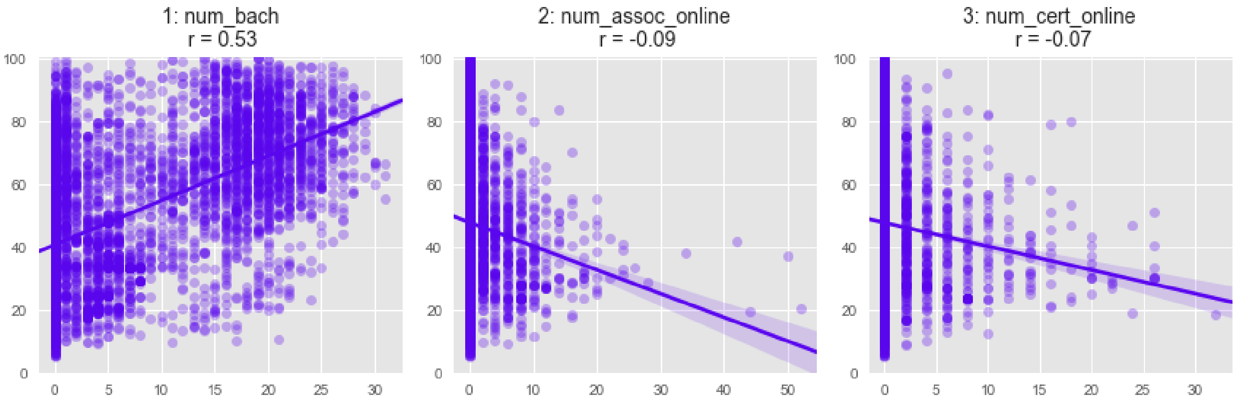
The numerical features in the academics category referred to the percentage of degrees awarded for each program at an institution. Some programs had a clear linear or exponential relationship with repayment rate, which others had no relationship at all. There was no information about what type of degrees these were, but one could guess that some programs correspond to a liberal arts education versus a technical school. For example, it seems that Computer Schools (figure 2) correspond to somewhat lower repayment rates. More information would be needed to make strong conclusions about schools with many health degrees: it would be interesting to explore further into what types of institutions offer higher percentages of health degrees.



Logarithmic transformation showed a clearer correlation between some programs and institutional repayment rates. These patterns seem to defy intuition; institutions granting more degrees for soft sciences (English, History, Social Science) seem to have a better outcome than professional degrees like business and computer (as seen above in figures 1 and 2).

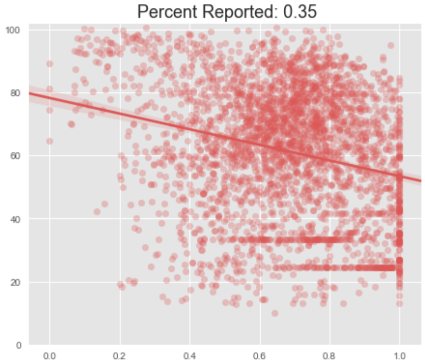


The Academics category also provided categorical features (n_features = 190) for whether an institution offered a program at a particular level (Certificate, Associate, Bachelor's) and if it was offered only online. To get a better understanding of these data, it was helpful to zoom out a little and create features for the total number of Certificates, Associate, and Bachelor's degrees offered in-person at the institution. There was a strong correlation between the number of Bachelor's degrees offered in-person and repayment rate, and a weak negative correlation between the number of Associate and Certificate degrees offered only online. (There was no correlation with the number of Associate and Certificates offered in-person).

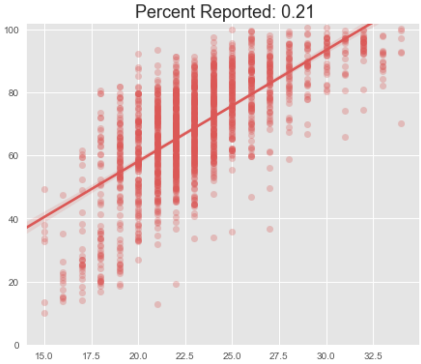


Admissions

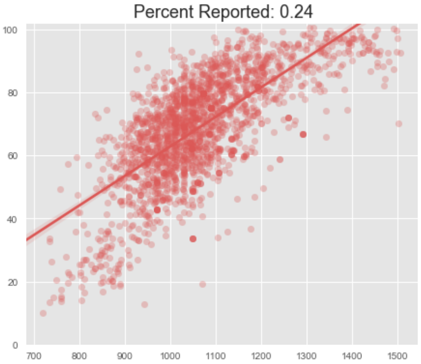
Higher admissions rates correspond to lower repayment rates. This makes sense given that institution selectivity was found to be a strong predictor. SAT and ACT scores had a positive linear correlation with student loan repayment rates. However, not all schools provided these data, as noted by 'Percent Reported' in each figure.



1: admissions__admission_rate_overall
r = -0.14
Percent Reported: 0.35



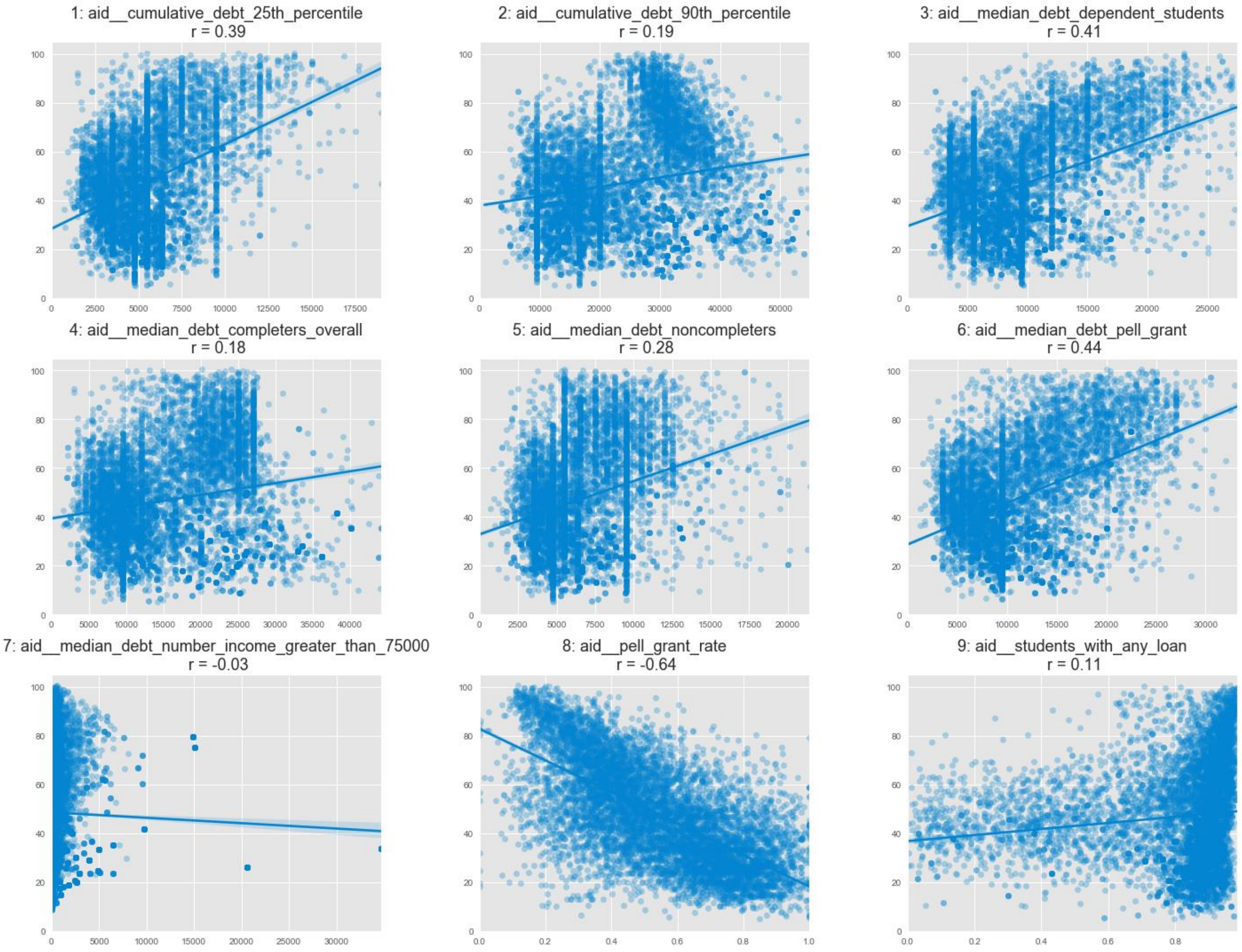
2: admissions__act_scores_midpoint_cumulative
r = 0.26
Percent Reported: 0.21



3: admissions__sat_scores_average_by_ope_id
r = 0.28
Percent Reported: 0.24

Aid

Many features in the aid categories showed linear relationships. Some features (2 and 4) showed interesting clustering, which was later used to create a new feature for cluster assignments that gave some further predictive value. The features for the *number* of students in a particular debt category showed virtually no correlation as seen in figure 7. Pell Grants are interesting: the greater the share of students that receive Pell Grants, the lower the institution repayment rate, but a larger median debt for Pell students corresponded to a better repayment rate. Finally, most students received a federal loan (figure 9), but that did not seem to affect repayment rate. The general trend is that the greater the median debt for an institution, the higher the repayment rate.



1: aid__cumulative_debt_25th_percentile
r = 0.39

2: aid__cumulative_debt_90th_percentile
r = 0.19

3: aid__median_debt_dependent_students
r = 0.41

4: aid__median_debt_completers_overall
r = 0.18

5: aid__median_debt_noncompleters
r = 0.28

6: aid__median_debt_pell_grant
r = 0.44

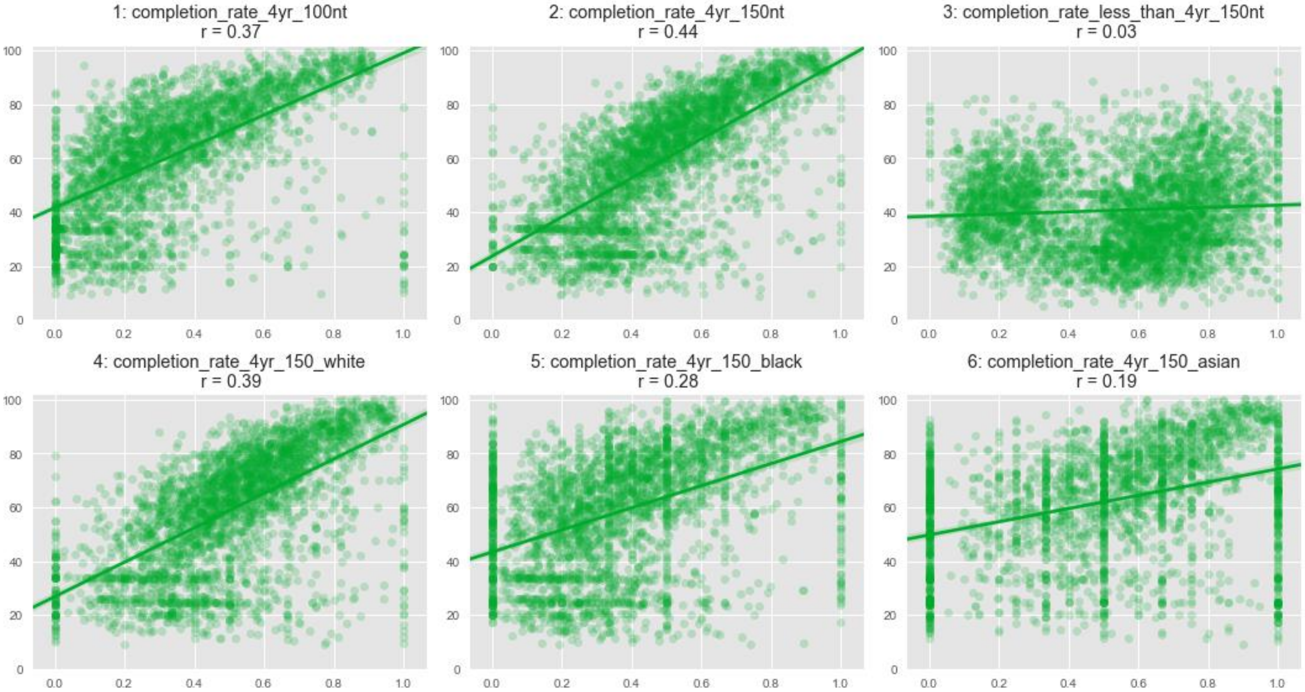
7: aid__median_debt_number_income_greater_than_75000
r = -0.03

8: aid__pell_grant_rate
r = -0.64

9: aid__students_with_any_loan
r = 0.11

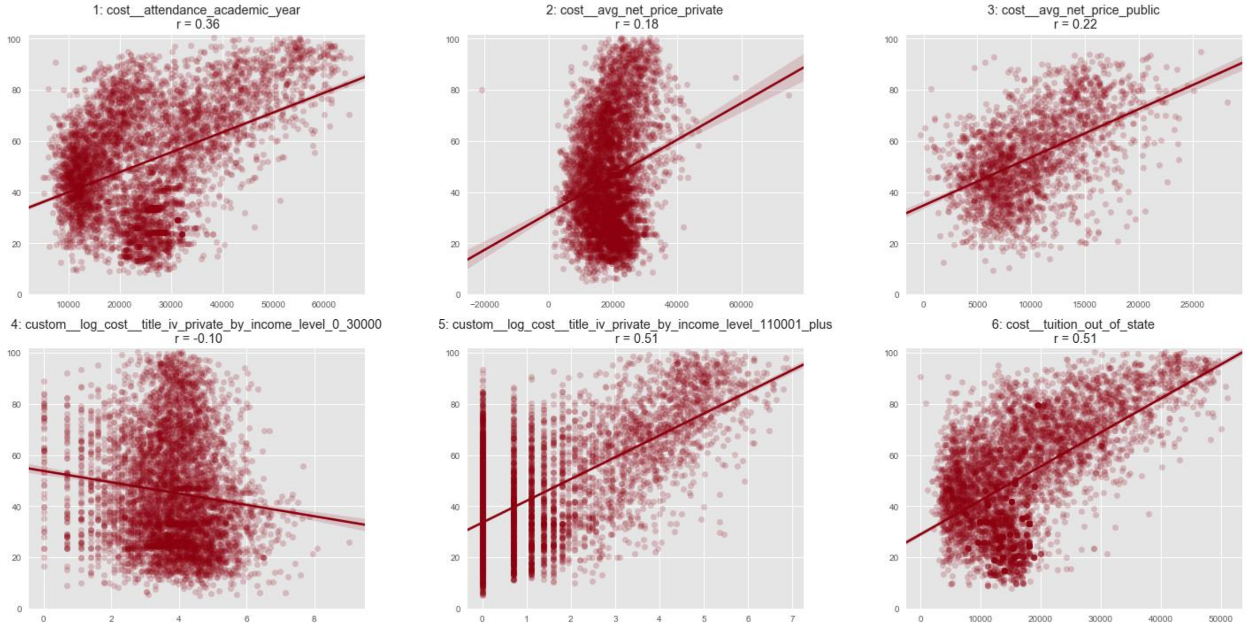
Completion

The data showed a clear linear correlation between completion rates and repayment rates. It was easy to compare the effect of race on repayment rates here: completion rates for White students were more strongly correlated with repayment rates than for Black students, and even less so for Asian students. However, this correlation was only present in four-year institutions; institutions that were less than Four-Years showed virtually no relationship between completion rate and repayment rate (figure 3). The data also showed that students who spent more time at their four-year institution had higher repayment rates (figures 1 and 2).

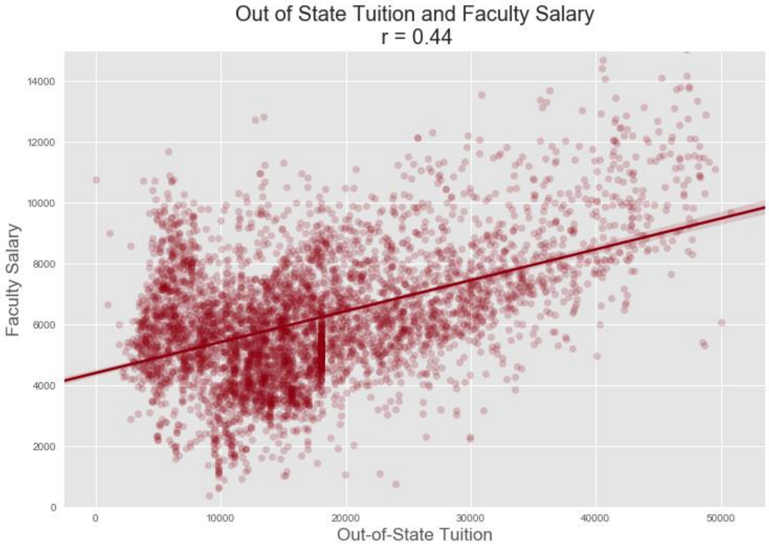


Cost

Overall, only a few categories showed a liner relationship between cost and repayment rate. These linear relationships implied that more expensive institutions correspond with better repayment rates (figures 1 and 6). Interestingly, net price for students at private institutions had virtually no correlation with repayment rate (figure 2), while net price for public institutions showed weak correlation (figure 3). Separating by income level demonstrated that private institutions benefited from having more students from wealthier families (figure 5). Institution repayment rates did not improve as the number of low-income students increased. The data gives credence to the idea that institutions with students from higher socioeconomic levels experience better repayment rates.

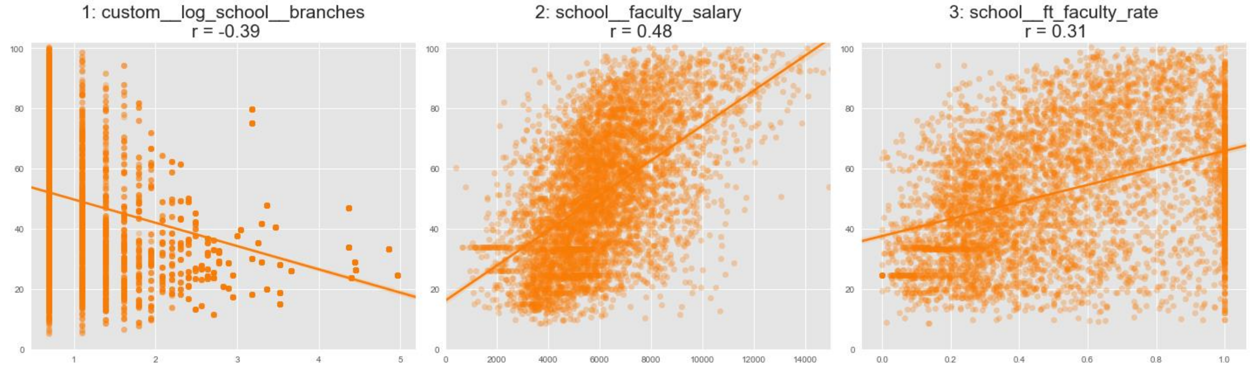


The data shows a correlation between tuition and faculty salary. Since faculty salary and tuition both exhibit positive correlation with repayment rate, one could hypothesize that faculty commanding higher salaries improve student outcomes. It would be interesting to compare more expensive institutions with less expensive ones and where they spend their budget (academics, facilities, athletics).

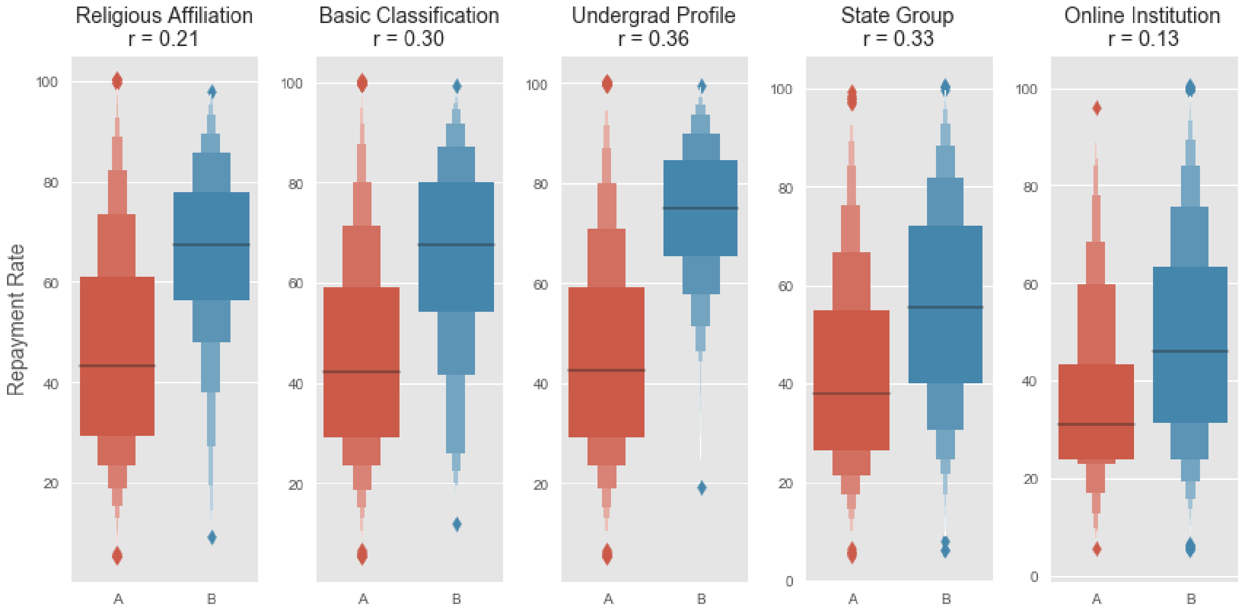


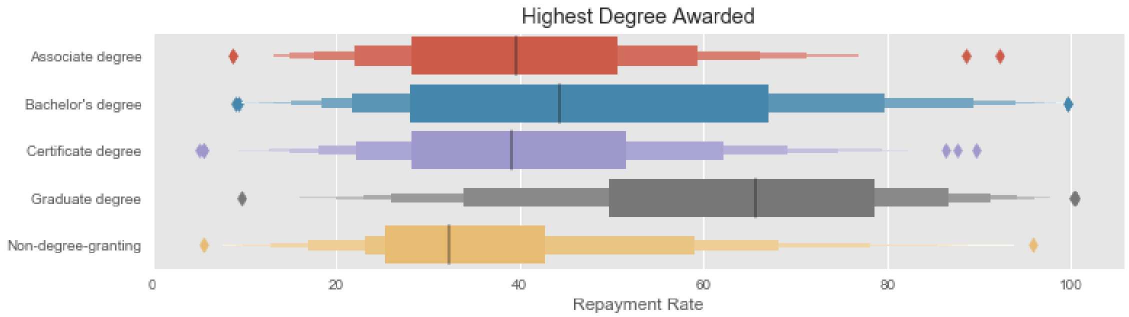
School

Only 3 features in this category (n_features = 27) were numerical and they all showed a positive linear relationship with repayment rate. Institutions experience better repayment rates with more full-time employees with higher salaries.

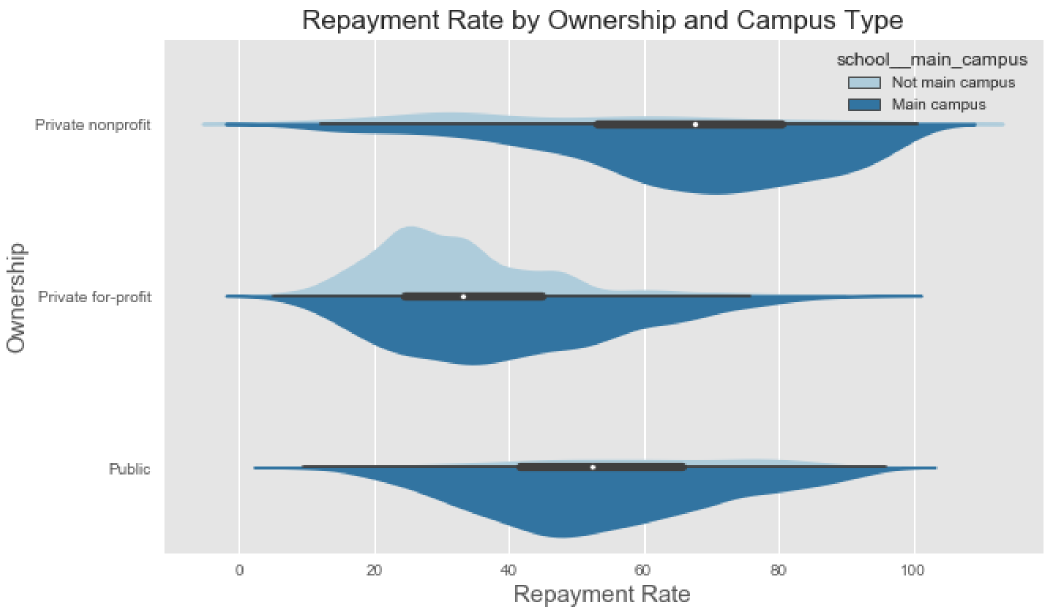


Nearly every one of the 25 categorical features in the School category had at least some impact on repayment rates. Some feature engineering was used to simplify the data and reduce the number of features: categories were grouped according to correlation with repayment rate and transformed into a binary feature indicating group membership.

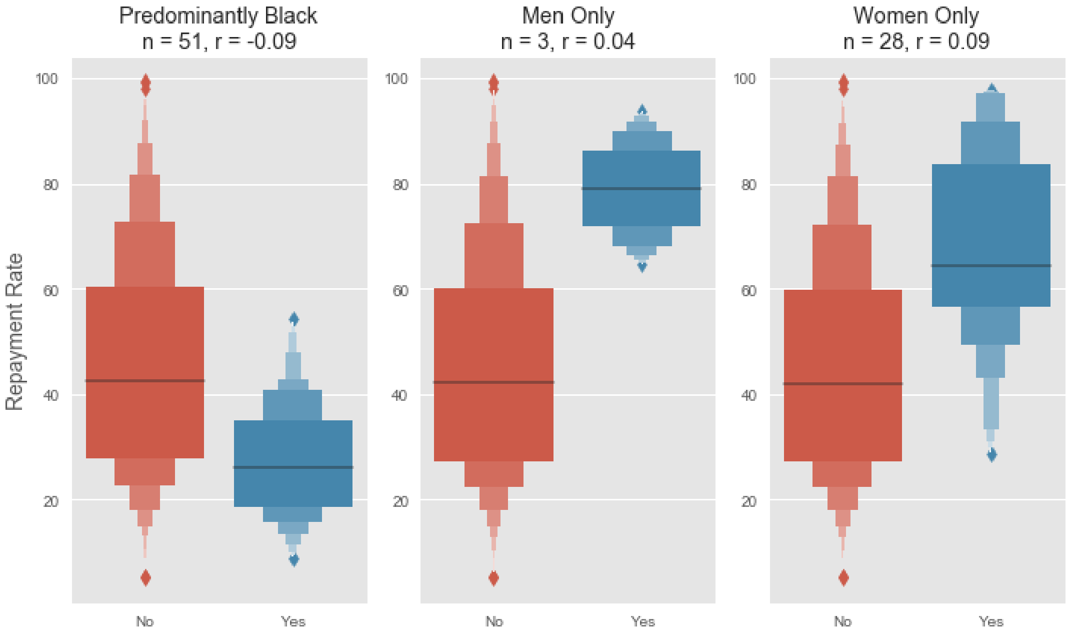




Our data did not come only from traditional universities, but also included for-profit institutions. Private, non-profit institutions have the highest repayment rates ($\bar{x}$ = 64.73) and private for-profit institutions have very low repayment rates ($\bar{x}$ = 35.75). Additionally, Private For-Profit institutions had many non-main campus locations which corresponded to very low repayment rates. Private and Public non-profit institutions had relatively few non-main campus locations; the repayment rates were high at the non-main campus locations for Public institutions and low at Private ones.

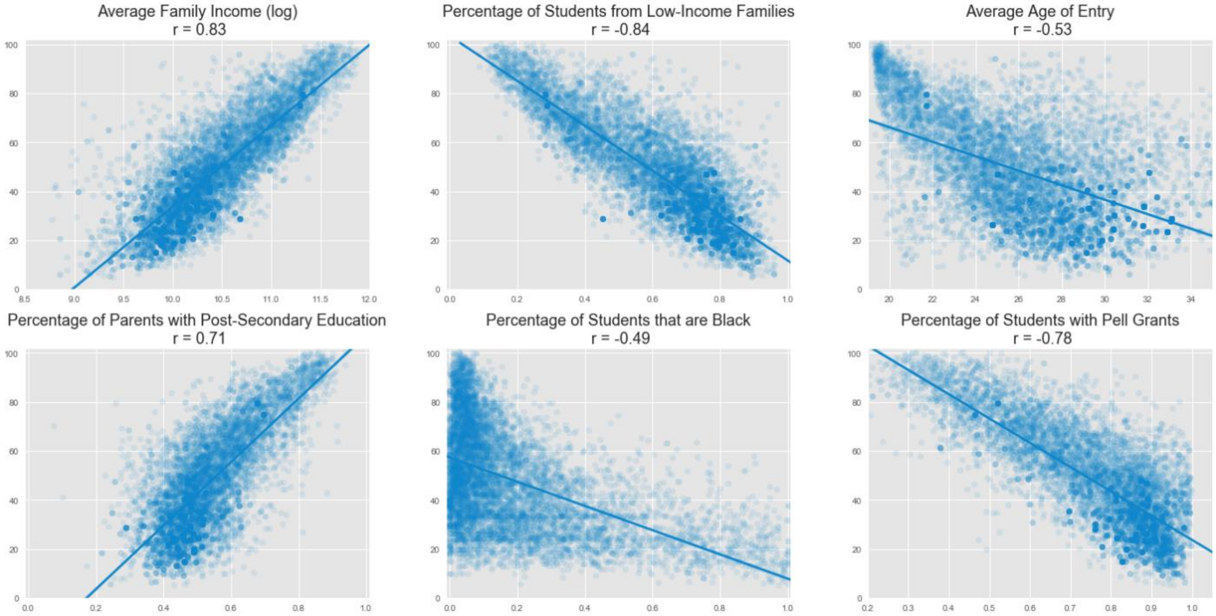


The demographic information provided in the School category was sparse, but provided some interesting information. Predominantly Black institutions suffered from low repayment rates and male/female institutions enjoyed higher repayment rates although they represent as very small sample size.



Student

All the features in the Student category were numerical. Income proved to be a significant factor in student loan repayment rates. Student Demographics were also found to be key features, amongst the most important were ethnicity, Pell Grants, and average age of entry, and parental education. Clearly, the socioeconomic composition of institutions is highly related to repayment rates.



Conclusion

The introduction to the project asked, “Can you help students understand which institutions are good investments, and which ones leave them in debt?” The question is somewhat misleading—most students leave their institutions with debt. A better question would be, “What *kind* of institutions provide an education that is worth going into debt for?”

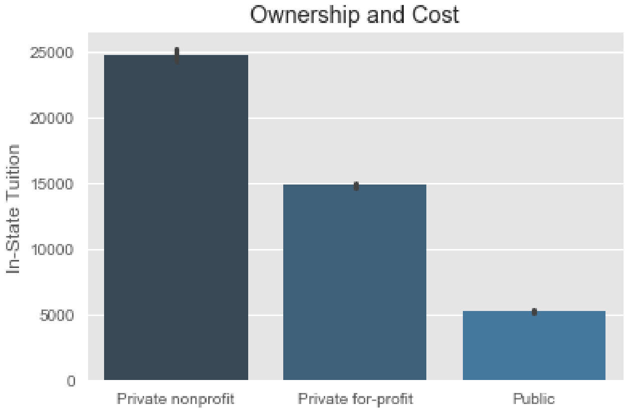
> *While outcomes for the typical student vary substantially across schools, there is a great deal of variation within schools in the outcomes as well. Students should know that while differences across schools may inform the question of relative school quality, these differences mask a great deal of heterogeneity in the outcomes of students.... The fact that there is so much variation in student outcomes within schools should not be taken as evidence that schools may not matter as much as other factors.*[2]

[2] Using Federal Data to Measure and Improve the Performance of U.S. Institutions of Higher Education. 2015 (revised 2017): 48-49. https://collegescorecard.ed.gov/assets/UsingFederalDataToMeasureAndImprovePerformance.pdf. Accessed July 20, 2017.

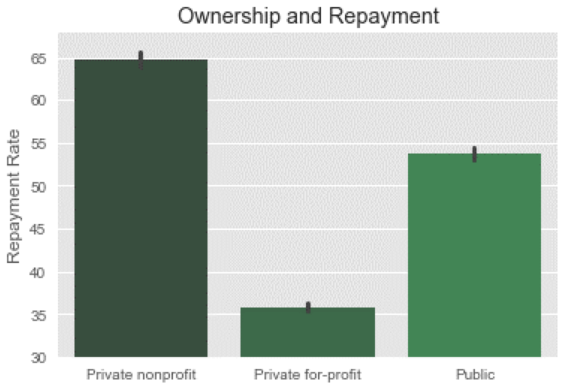
It is difficult to predict the outcome of individual students. Furthermore, the outcome for a student is not equivalent to their repayment status. Therefore, this report should not be used to establish causal relationships between prominent features and repayment rate. Rather, the patterns identified here can be used to inform various audiences about the efficacy of institutions to provide a return on investment (ROI).

Educational funding organizations can use these patterns to determine which institutions will benefit from funding and which ones which ones do not provide value to their students. Policy makers can use these data to determine what interventions or regulations may be necessary to curb predatory institutions from taking advantage of those seeking to obtain a useful education, as well as prevent discriminatory practices. Institutions themselves can establish best practices, create more conscientious budgets, and provide better services for students. Finally, students and parents can use this report to plan for a successful future.

Key Findings



- Private for-profit institutions do not provide value to their students. In fact, private for-profit institutions have been in the news recently for defrauding students and some have even been ordered to discharge student loans.[3] These institutions often serve a different, less-advantaged demographic than non-profit institutions. Research has shown “for-profit student [sic] have less institutional aid to cushion tuition costs and school fees. As a result, they face the highest borrowing rates of any sector for comparatively weaker job prospects post-graduation.”[4] Students and funders should avoid these institutions and policy makers should enact stricter regulations.

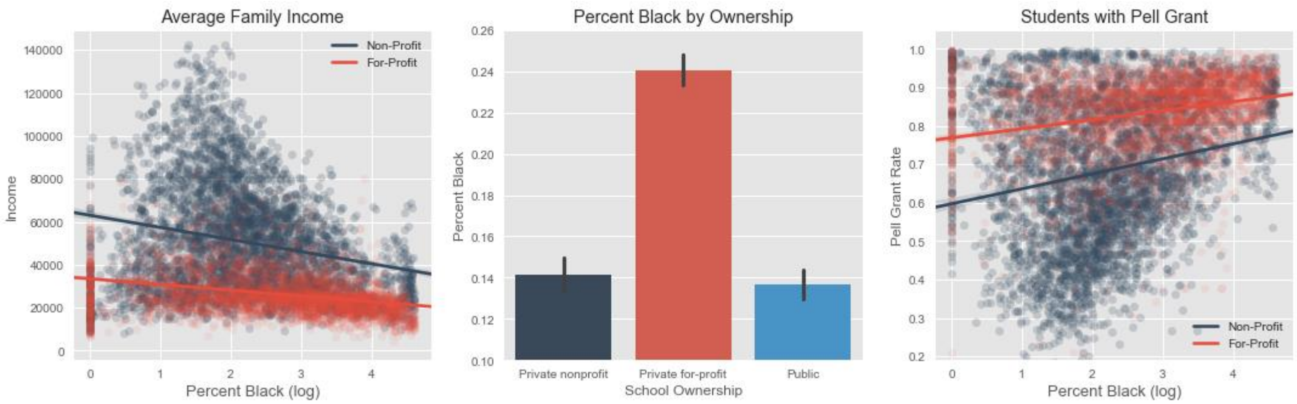


- There is a pattern of correlation between tuition/cost and repayment rates, particularly for public institutions. Although the cost of post-secondary education has outpaced economic growth and ROI for graduates, students should not avoid certain institutions based solely on cost. In fact, institutions where students leave with a greater amount of debt have higher repayment rates. More research should be conducted to discover patterns between types of expenditures, such as administrator salaries, instructor salaries, and spending on facilities, academics, and athletic programs. Private Non-Profit institutions have higher repayment rates, however, the cost of Private Non-Profit institutions does not affect repayment rates. Expensive public institutions may provide the best ‘bang for the buck.’

[3] Consumerist: For Profit College articles. https://consumerist.com/tag/for-profit-college/. Accessed July 21, 2017.

[4] Cellini, SG, Darolia, R. Different degrees of debt: Student borrowing in the for-profit, nonprofit, and public sectors. https://www.brookings.edu/research/different-degrees-of-debt-student-borrowing-in-the-for-profit-nonprofit-and-public-sectors/. Accessed July 21, 2017.

- Finally, socioeconomic composition may be the most significant factor in determining institutional repayment rates. Student bodies that are poorer, blacker, and older have lower repayment rates. Permutation Feature Importance scoring showed that the percentage of black students ranked third in importance, behind two features that measured income. For-Profit institutions enroll far more Black students and students from the poorest families, yet offer the lowest repayment rates. Regardless of ownership type, however, these demographics are strongly correlated with low repayment rates. Students from disadvantaged demographics should avoid institutions with compositions that are predominantly like their own. Institutions and policy makers should be aware of these patterns and work to prevent discriminatory practices (or predatory practices, in the case of For-Profit institutions).


Average Family Income
Non-Profit
For-Profit
Income
Percent Black (log)
Percent Black by Ownership
Percent Black
Private nonprofit
Private for-profit
Public
School Ownership
Students with Pell Grant
Pell Grant Rate
Non-Profit
For-Profit
Percent Black (log)