An Analysis of National Student Loan Data from 2009-2015

Madelyn Rubenstein, Surya Menon, Nanditha Sundararajan, Camellia Debnath

Summary: When deciding where to attend college, choices between public versus private and local versus out-of-state can mean the difference between acquiring an immense amount of debt due to student loans, or going to a more affordable but less prestigious school. Paying off student loans can take quite some time, upwards of 10 plus years, and these debts continue to increase when considering intentions of pursuing a graduate degree[1]. The debt that can follow a student after college may take a while to pay off—whether or not they actually complete a degree—and can impact other factors such as their credit score, the ability to buy a house, or taking out loans for other life investments[3,4].

While the costs for higher education continues to grow, recently companies such as IBM, Google, and Apple have stopped requiring college degrees to apply for jobs[2]. Students looking to apply to college may think twice about their decision, considering how much time and money it would take to complete their degree if companies are hiring them regardless of whether or not they attended college. With this in mind, we wanted to look at factors contributing to students' ability to repay their college loans, and draw conclusions on important parameters that students should consider when making decisions about their college investment.

We were able to explore some of the potential factors related to student loans through the College Scorecard dataset, which is a data source generated and maintained by the U.S. Department of Education. The dataset was created under the Obama Administration in 2013, and is designed to increase transparency in higher education by allowing the public to assess schools on how well they are serving their students[5]. We chose to use this dataset because it contains detailed information on every undergraduate institution in the country in the following categories: Academics, Admissions, Student Body, Cost, Aid, Completion, Repayment, and Earnings[6]. This dataset gives a comprehensive look at national universities and spans more than 20 years, from 1996-2017.

We looked at national student loan data from 2009-2015 across all 4-year public and private universities to see if we could predict the percentage of students at a school who could pay off their loans within three years. We first subsetted the dataset to examine variables we thought could potentially relate to our response variable, the 3-year repayment rate, such as student demographics, location, cumulative debt, cost of tuition, family earnings, and visualized these variables on a national and local level. Through extensive exploratory data analysis (EDA) and observing trends within these variables, we were able to further narrow down the variables considered as potential predictors for modeling. We then examined the relationships of these variables with the response variable and from these results fit a linear regression model to predict the 3-year repayment rate.

From our analysis and modeling we found that variables such as the debt-to-earnings ratio, the percentage of students receiving Pell grants, the median earnings of federally aided students after graduation, the cost of attending the institution, and median family income are highly important in determining the 3-year repayment rate at an institution.

Methods:

Tidying and Variable Selection: The College Scorecard dataset contains 2,146 variables, so one of our first steps in data tidying was choosing the variables for which we could visualize trends through EDA, along with those we thought would be good predictors for modeling. Through an iterative process over five rounds, we narrowed down the variables that we wanted to further analyze. We first attempted to eliminate variables that would not contribute to any sort of analysis, such as the institution ID and URL link. Additionally, we eliminated variables for which we consistently found a lot of null or irrelevant data across various years. For example, although there was student demographic information available in the dataset, it was specifically related to enrollment rather than in regards to financial aid, so we determined that these variables would not be helpful for our analysis. Furthermore, there were a lot of highly specific variables related to student completion and transfer rates, which we also decided to exclude, because they were either unrelated to student loans or too specific for the more general analysis we were interested in conducting.

Through the next couple of rounds of variable selection, we attempted to choose variables that intuitively made sense in terms of financial aid and student loans. After these iterations, we still had around 180 variables left to consider, so we further narrowed down the variables by focusing on measures that were clearly defined and would be easily understood by students. In particular, a lot of the financial metrics we considered were highly specific but poorly defined due to vague descriptions in the data dictionary, such as 3-year repayment rate for non-first-generation students suppressed for n=30, were redundant, so we ended up excluding those variables from analysis as well. Through this entire process, we were able to reduce our dataset to 81 variables of interest to look at for visualization and modeling.

The datasets are split into separate files by year, so to prepare for visualization, we joined together the years we were interested in analyzing. We decided to limit our focus to five years of data for this project to keep the data to a manageable size that could be analyzed effectively across all group members' computers. We also wanted to choose a group of years that had the least amount of null data for the variables we had chosen, and to focus on years with more recent data as a preference. Through an initial examination of the datasets, we determined that the years 2009-10, 2011-12, 2012-13, 2013-14, 2014-15 generally had the least amount of null data for the variables of interest. We then filtered our data to only include public and private 4-year institutions, and subsetted to the 81 variables we had chosen.

EDA: Through some of our initial EDA, we discovered it was a little difficult to visualize and interpret trends on a national level for each year, so we decided to focus some of our visualizations on local data to facilitate our understanding of some of the variables, and to assuage our personal interest in Massachusetts data (refer to Figure B in the Appendix). To look at regional trends, we subsetted the data to specifically look at schools in New England states, which include Connecticut, Massachusetts, Maine, Rhode Island, New Hampshire, and Vermont. To look at metrics over Massachusetts schools, we chose the top five public and private universities in the state as reported by the U.S. News and World Report 'Best 2019 College Rankings'[7]. In this report, the schools are evaluated across 16 different measures of academic quality using various types of data collected[8]. Our list included Harvard University, Massachusetts Institute of Technology (MIT), Tufts University, Brandeis University, and Boston College for private universities, and University of Massachusetts (UMass) Amherst, UMass Lowell, UMass Boston, UMass Dartmouth, and Massachusetts College of Liberal Arts (MCLA) for public institutes.

From our EDA we were further able to narrow down the variables we would consider as potential predictors for modeling. In particular, we decided to examine variables for which we saw interesting trends, such as median family income and the percentage of students receiving federal grants. Additionally, for some of these metrics, we most consistently saw null values when visualizing data between 2010-2011, so for modeling we decided to exclude this year from analysis to generate a more robust and consistent model.

Modeling: We decided to fit a linear regression model to see if we could predict the 3-year repayment rate at a university. Specifically, our response variable was the 3-year repayment rate for completers, which is defined as the fraction of borrowers at an institution that completed their degree and were able to pay off their student loans within three years[(6)]. We found it was common in other models to use a 3-year period as a measurement window when looking at repayment rates, as opposed to 1, 5, or 7 years[(9)].

We split our combined dataset from the years 2009-15 into training (60%), validation (20%), and test (20%) sets; we decided to split the dataset this way for ease of modeling, since conducting separate year-wise splits of the data and merging these partitions would result in similar results to the general randomized split we performed. We conducted visualization on the training set with the variables we determined as highly likely to be related to the response variable. Through EDA, we were able to eliminate variables with which we did not see a relationship with 3-year repayment rate, such as the number of students in the median debt cohort and family education level. We evaluated the correlation between each of the predictors and the response variable with scatter plots, and selected the predictor variables that showed strong correlations, and additionally had little or no correlation with one another to avoid possible over-parameterization in the model. Also, we found that some variables, such as the median earnings of students 6, 8, and 10 years after graduation, exhibited analogous relationships with the response variable across all years and were tightly correlated with each other, so we decided to include only one of them in the model due to their similarities. From these procedures we determined around 20 variables to examine as potential predictors for our model.

Fitting a linear regression model for the 3-year repayment rate included several iterations of verifying how including or excluding specific predictors affected the root-mean-square error (RMSE), and checking with the validation set to avoid overfitting, by checking the R-squared, adjusted R-squared and predicted R-squared values. We ultimately determined five variables that gave us a reasonably low RMSE for our model: the debt-to-earnings ratio (refer to Snippet 4 in the Appendix), the median earnings of federally aided students after graduation, the cost of attending the institution, the proportion of undergraduate students who received Pell grants, and the median family income. With the response variable and these explanatory variables, we used the *lm* function from the R *stats* package to build a linear regression model on our training dataset. We plotted the predicted values with the actual values from the dataset to visually judge whether our model appeared to be a good fit. We also examined the residuals by plotting them against each of our predictor variables and examining the distribution; the random noise pattern we observed in these plots confirmed that our model was not violating any linearity assumptions (refer to Figure H in the Appendix). Finally, we calculated the RMSE on the test data to assess the predictive ability of our model.

Results: The following figures depict the EDA and modeling results for the predictor variables we ultimately ended up using in our model with 3-year repayment rate as our response variable. Figure 1 depicts the average debt-to-earnings ratios for the top five public and private schools in Massachusetts over a 5-year period. We observe the trend that private universities appear to overall have a lower ratio compared to the public universities, with Harvard consistently having the lowest ratio. A lower debt-to-earnings ratio indicates a good balance in debt and income and can give an individual a better chance at acquiring a mortgage or loan; a higher ratio, however, may indicate that an individual has too much debt for the amount of money they are earning[16]. We observe that students at private universities in Massachusetts may be able to more successfully manage their debt based on their earnings, compared to those at public schools.

Figure 2 shows the median earnings for students eight years after graduating from the top five public and private universities in Massachusetts over a 5-year period. Students graduating from MIT, Harvard, and Tufts have the highest median earnings compared to the rest of the public and private universities. This finding is not necessarily surprising, but it is interesting to see not only that all the private institutions' earnings outnumber the public universities each year, but also the actual amounts by which these universities tend to differ; in particular, we see that MIT students consistently have the highest median earnings, more than double the amount compared to students at MCLA.

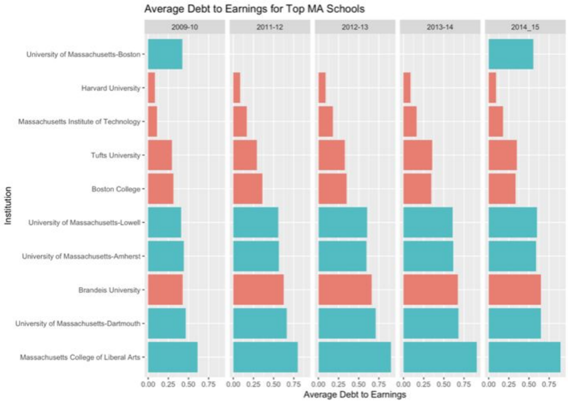


Figure 1: Average debt-to-earnings ratio for top Massachusetts schools, 2009-2015

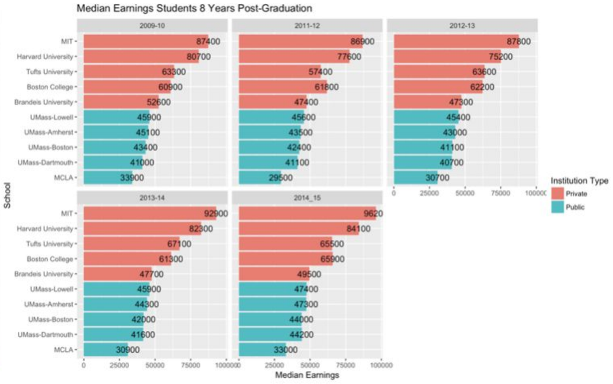


Figure 2: Median earnings of students post graduation, for top Massachusetts schools, 2009-2015

Figure 3 shows the average cost of attendance by state for the most recent year of data available, 2016-17. The District of Columbia (DC) has the highest cost of attendance at $47,597 followed by Massachusetts at $46,977 and Rhode Island at $45,472, with Vermont, Connecticut, and California following; in contrast we see that states such as Wyoming and Utah have some of the lowest costs of attendance, at $19,169 and $20,715, respectively. This shows that it is generally pretty expensive to attend school in many of the New England states, and this could be a good factor for students to examine when initially determining a school or region in which to study.

Figure 4 depicts the average percentage of Pell grant recipients for New England states over a 7-year time period. A Pell grant is a subsidy from the U.S. federal government, and is limited to students with financial need, who have not earned their first bachelor's degree or who are enrolled in certain post-baccalaureate programs, through participating institutions[15]. A Pell grant is generally considered a foundation of a student's financial aid package, to which other forms of aid are added, and therefore is an important metric when examining students who receive financial aid. Across all the years we examined, the percentage of Pell grant recipients is well under 40%, showing that there is a pretty small amount of students in each state receiving these grants in New England. Notably, public universities in Maine consistently have the highest percentage of Pell grant recipients from 2011-2015, followed closely by both public and private universities in Vermont.

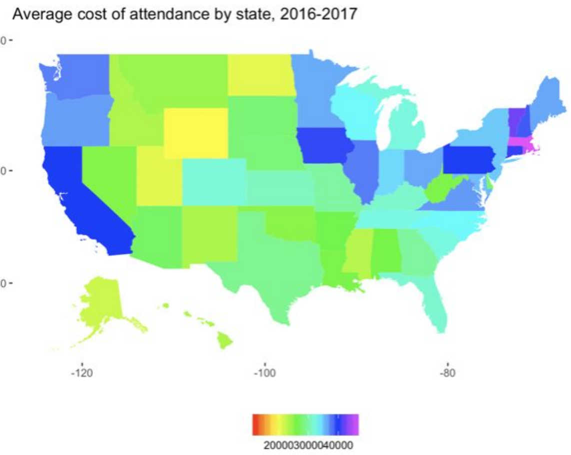


Figure 3: Average cost of attendance, 2016-2017

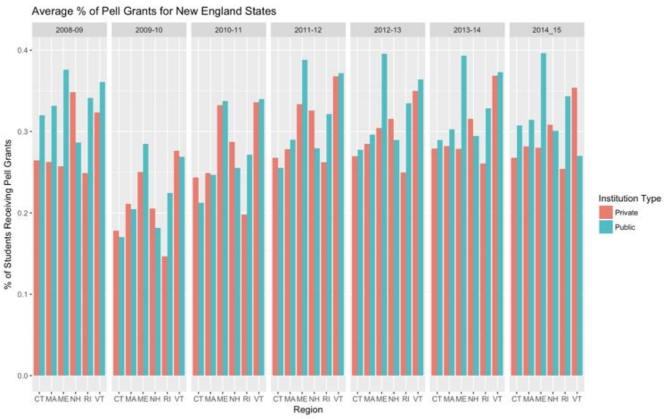


Figure 4: Average percentage of Pell grant recipients, New England states, 2008-2015

Lastly, Figure 5 depicts the 3-year repayment rate for completers and non-completers at a university. Completers are defined as students who were able to finish their degrees in four years, while non-completers are defined as students who withdrew from their university during that same time period[6]. Despite this difference, the repayment rate for both groups is well over 0.5 for both public and private universities, indicating that students were able to pay back their loans during this time period. This an interesting finding, since some prior research has observed that completers at an institution are more likely to pay off their loan principals than those who do not complete their degrees[17].

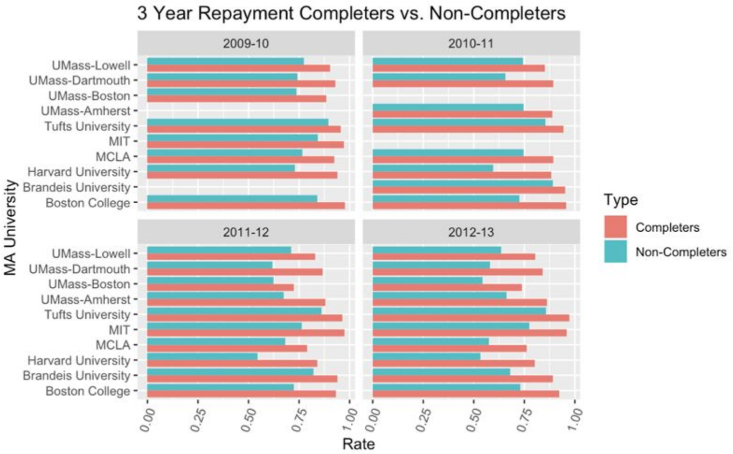


Figure 5: 3-year repayment rate for completers versus non-completers, top Massachusetts schools, 2009-2013

In terms of modeling, we observed in Figures 6 and 7 that the median family income and the median earnings of students after graduation have positive relationships with our response variable, with correlations of 0.7743 and 0.6578, respectively. We further saw that the median family income has a nonlinear relationship with the response variable, so we transformed this predictor in our final linear model. Other variables that we determined had linear relationships with the 3-year repayment rate from their correlation plots were the debt-to-earnings ratio, the cost of attending the institution, and the proportion of undergraduate students who received Pell grants (refer to Figure A in the Appendix). As a result, we included these variables as predictors in our linear model.

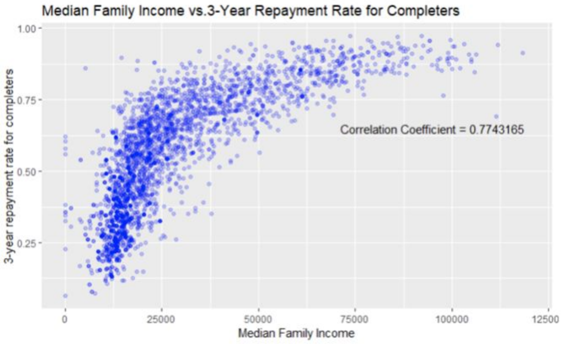


Figure 6: Correlation between median family income and 3-year repayment rate. Reference year: 2013-14

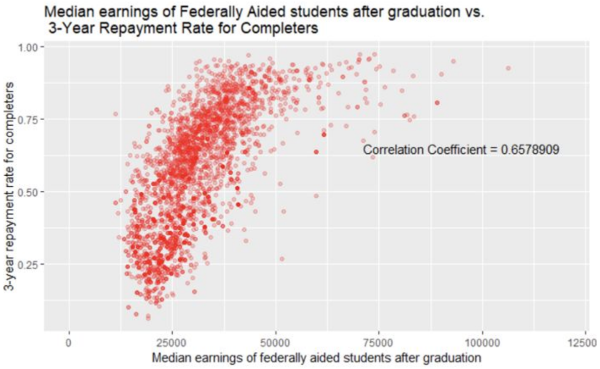


Figure 7: Correlation between median earnings 8 years after graduation and 3-year repayment rate. Reference year: 2013-14

From our linear regression model, we found that the training RMSE was 0.0922 and the test RMSE was 0.0939. We see that the test error is slightly higher than the training error, which indicates we are not overfitting our model. Overall, the low value of our test RMSE suggests that our model is able to accurately predict the response variable. In Figure 8, we plot the predicted and actual values together and confirm that the observed values in the dataset are generally close to the predicted values in the model, so we are also visually able to confirm that our model is a good fit for predicting the 3-year repayment rate.

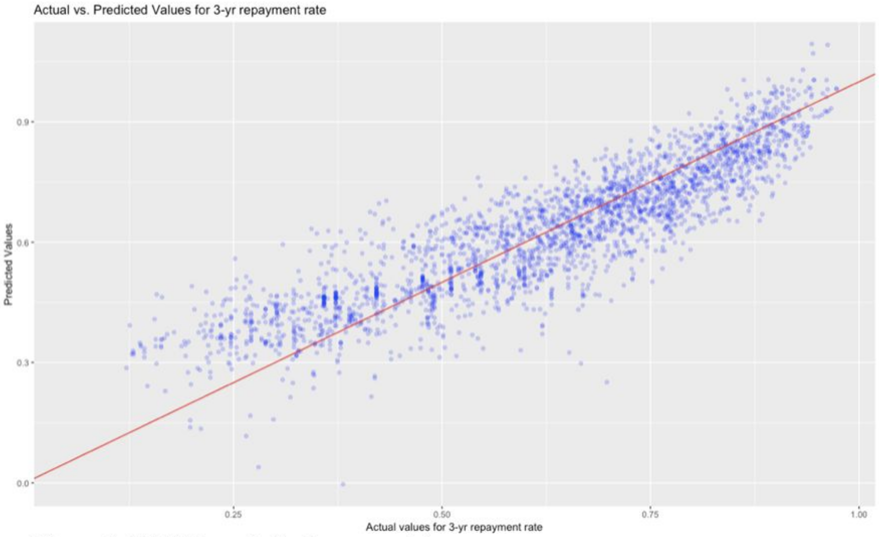


Figure 8: RMSE graph for linear model

Discussion: The College Scorecard dataset is a valuable resource for students to use to make informed decisions about where to attend college. However, with the extensive amount of variables available, it may seem like a daunting task for some to sort through all the data. Through this project, we attempted to create a model to predict repayment rates of student loans within three years, and from this process, we identified important and coherent variables that students can evaluate when deciding where to attend college and determining how feasible it would be to pay off their student loans in a designated time period. With college education only getting more expensive and federal grant opportunities decreasing, it is becoming increasingly important for students to evaluate how worthwhile a college education can be, and how making this investment may affect their future[10, 11].

Overall, we generally determined comprehensible and sensible predictor variables when creating our model for the 3-year repayment rate. We attempted to choose variables that would intuitively make sense when thinking about loan repayments, and that students themselves could look up and understand. For example, using the percentage of undergraduate students who received Pell grants as an explanatory variable is a good indicator because students receiving grants are not burdened with as much student debt as those who have taken out federal loans, since grants do not have to be paid back. Additionally, debt-to-earnings is another important metric to look at when discussing loan repayment, since it indicates an individual's ability to pay off loans based on their income. When looking at the correlation between median earnings of federally aided students after graduation and the repayment rate, a positive relationship is expected, since a higher salary would indicate a better ability for a student to pay off their loan faster. The cost of attending an institute could positively correlate to repayment rate—as we observed in our visualization—perhaps because students who complete degrees at more expensive institutions are possibly obtaining higher paying jobs after graduation that will help them pay off their loans sooner. Lastly, median family income is a reasonable predictor because it partly determines how much federal loan money a student will receive, and could potentially dictate how long it would take the student to pay off the loan.

Although we were able to construct a strong linear model, there were some significant weaknesses with this dataset. There was a lack of consistency in the data over time, with some variables disappearing for certain years. As a result, we had to choose variables that were available and consistent across different years. Additionally, we encountered a lot of null data when working with this dataset across many variables of interest for particular years; this meant we ultimately had to limit the span of years of data we used in our analyses. Furthermore, although the dataset provided an extensive data dictionary, the documentation for many variables—specifically for a lot of the financial aid and loan repayment information—was often times vague or difficult to interpret, which created some difficulties when conducting our extensive variable selection process. Lastly, the median earnings information that is provided in the dataset only accounts for students receiving federal financial aid, which makes it difficult to apply these financial metrics to those who do not receive loans and to generalize our findings to college students as a whole, since a significant number of students attending college do not fit into this categorization(12). Therefore, our analyses are limited to the scope of only students that received loans and grants from the federal government.

One of the biggest challenges of the project was variable selection. With over 2,000 variables to initially consider, the process of determining relevant and interesting variables for visualization and modeling was a painstaking process that took considerable time. While we were finally able to narrow down our variables to a small but strong set of predictors for our model, there are many other possible combinations of variables that could have also resulted in a predictive and insightful model. However, to avoid overfitting, we tried to keep the number of predictor variables to a minimum, with understandable meanings in relation to our response variable.

Although we selected the 3-year repayment rate as the response variable in our model, the dataset also provided 1, 5, and 7-year repayment rate measurements. A potential next step could be constructing models for different repayments rates to see how they compare; a research question for example, would be do different repayment rates use the same predictors? Additionally, we chose to work with repayment rates because they are generally considered to be more sensitive than default rates, which measure the worst-case scenario for repayment outcomes and can be manipulated through the use of allowable non-repayment options like deferments and forbearances(6). It would be interesting to look at the differences in models using default rates versus repayment rates as well. Also, perhaps having access to national student-level data in the future could also be an interesting avenue of research to further examine how repayment rates may be influenced by student-specific demographics, and to potentially create a predictive model that is more generalizable to all college students across the country.

The College Scorecard dataset only reports median earnings for students receiving federal financial aid, so another interesting follow-up to this study could involve looking at general median earnings for all college students. Some websites such as PayScale gather self-reported earnings data from college alumni into a database, but this data has a sample bias, so it may be problematic to use(13). However, interestingly, we discovered that the Trump Administration has plans to further expand the College Scorecard data by providing more measures regarding student outcomes after graduation. Specifically, there are plans to not only report earnings data that is more representative of the whole study body, but to provide earnings and student debt information by major field of study(14). Having this information could allow us to conduct a more well-rounded analysis and possibly create a more robust model. Alternatively, an expanded dataset could also provide opportunities to create models looking at predicting student earnings in the future. The College Scorecard is a nascent but powerful source of information regarding college attendance and finances. As it continues to expand with more accurate and relevant information, there will be more opportunities to truly understand and evaluate the value of a college education and predict certain outcomes.

Statement of contributions:

1. Madelyn Rubenstein - Identified dataset used for project, wrote up project proposal, participated in variable selection and conducted EDA, assisted in modeling, prepared slides for project presentation, and wrote project report.
2. Surya Menon - Wrote up project proposal, participated in variable selection and conducted EDA, assisted in modeling, prepared slides for project presentation, and wrote project report.
3. Nanditha Sundararajan - Wrote up project proposal, participated in variable selection, assisted in modeling, prepared slides for project presentation, and assisted in writing project report.
4. Camellia Debnath - Wrote up project proposal, participated in variable selection, performed modeling, prepared slides for project presentation, created GitHub repository with project code, and assisted in writing project report.

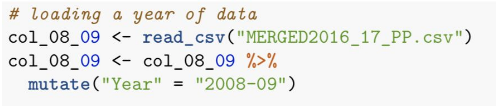
References:

1. Hess, Abigail. "This Is the Age Most Americans Pay off Their Student Loans." *CNBC*, CNBC, 3 July 2017. (https://www.cnbc.com/2017/07/03/this-is-the-age-most-americans-pay-off-their-student-loans.html)
2. Connley, Courtney. "Google, Apple and 12 Other Companies That No Longer Require a College Degree." *CNBC*, CNBC, 8 Oct. 2018. (https://www.cnbc.com/2018/08/16/15-companies-that-no-longer-require-employees-to-have-a-college-degree.html)
3. "Life Delayed: New Study Shows Student Debt Impacts Financial Security of Borrowers Across All Institution Types, Credentials". BusinessWire, 17 December 2015. https://www.businesswire.com/news/home/20151217006048/en/Life-Delayed-New-Study-Shows-Student-Debt
4. Trull, Jeffrey. "How Do Student Loans Affect Your Credit Score?" *Student Loan Hero*, Student Loan Hero, 3 Dec. 2014. (https://studentloanhero.com/student-loans/student-loan-repayment/how-do-student-loans-affect-your-credit-score/)
5. US Department of Education. "College Scorecard Data." *College Scorecard*, US Department of Education, 28 Sept. 2018. (https://collegescorecard.ed.gov/data/)
6. US Department of Education. "Data Documentation for College Scorecard ." *College Scorecard*, US Department of Education, 28 Sept. 2018. (https://collegescorecard.ed.gov/assets/FullDataDocumentation.pdf)
7. U.S. News & World Report. "2019 Best Colleges in Massachusetts | US News Rankings." *U.S. News & World Report*, U.S. News & World Report, 2018. (https://www.usnews.com/best-colleges/ma)
8. Morse, Robert, et al. "How U.S. News Calculated the 2019 Best Colleges Rankings." *U.S. News & World Report*, U.S. News & World Report, 9 Sept. 2018. (www.usnews.com/education/best-colleges/articles/how-us-news-calculated-the-rankings)
9. Miller, Ben. "Getting Repayment Rates Right." *Center for American Progress*, Center for American Progress, 10 July 2018. (www.americanprogress.org/issues/education-postsecondary/reports/2018/07/10/453199/getting-repayment-rates-right/)
10. Martin, Emmie. "Here's How Much More Expensive It Is for You to Go to College than It Was for Your Parents." *CNBC*, CNBC, 29 Nov. 2017. (www.cnbc.com/2017/11/29/how-much-college-tuition-has-increased-from-1988-to-2018.html)
11. deHahn, Patrick. "Bad News for Low-Income College Students in Trump 2017 Budget." *USA Today*, Gannett Satellite Information Network, 16 Mar. 2017. (www.usatoday.com/story/college/2017/03/16/bad-news-for-low-income-college-students-in-trump-2017-budget/37429281/)
12. Chingos, Matthew M., and Grover J. "Russ" Whitehurst. "Deconstructing and Reconstructing the College Scorecard." *The Brookings Institution*, The Brookings Institution, 15 Oct. 2015. (www.brookings.edu/research/deconstructing-and-reconstructing-the-college-scorecard/)

13. “College Salary Report Methodology.” *PayScale*, PayScale, 2018. (www.payscale.com/college-salary-report/methodology)
14. Vedder, Richard. “A Good Idea: More Job Earnings Data On The College Scorecard.” *Forbes*, Forbes Magazine, 20 Aug. 2018. (www.forbes.com/sites/richardvedder/2018/08/20/a-good-idea-more-job-earnings-data-on-the-college-scorecard/#17cfe4fd6881)
15. Staff, Investopedia. “Pell Grant.” *Investopedia*, Investopedia, 27 Sept. 2010. (www.investopedia.com/terms/p/pell-grant.asp)
16. Josephson, Amelia. “Debt-to-Income Ratio.” *SmartAsset*, SmartAsset, 22 May 2018. (https://smartasset.com/credit-cards/what-is-a-good-debt-to-income-ratio)
17. Kreighbaum, Andrew. “The Link Between Completion and Loan Repayment.” *Inside Higher Ed*, Inside Higher Ed, 8 Aug. 2018. (www.insidehighered.com/news/2018/08/08/link-between-college-completion-and-student-loan-repayment)

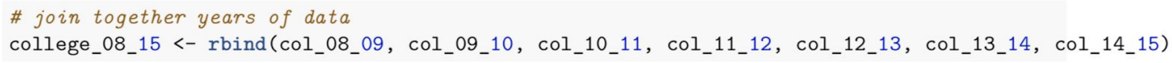
Appendix:

I. **Relevant Code -** For reference, we have included some snippets of the code from the project. The complete code for the project can be found at: https://github.com/debnath-c/DS5110**.**



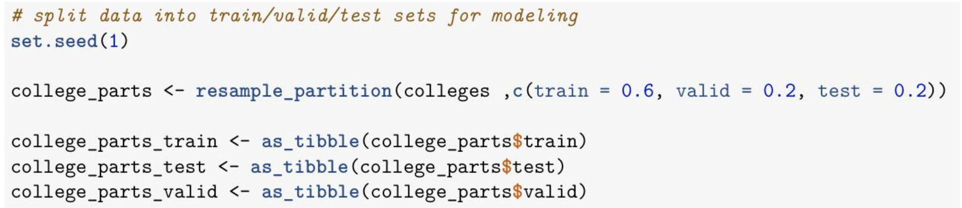
```
# loading a year of data
col_08_09 <- read_csv("MERGED2016_17_PP.csv")
col_08_09 <- col_08_09 %>%
  mutate("Year" = "2008-09")
```

Snippet 1: Loading datasets.



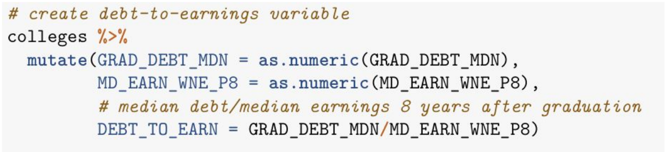
```
# join together years of data
college_08_15 <- rbind(col_08_09, col_09_10, col_10_11, col_11_12, col_12_13, col_13_14, col_14_15)
```

Snippet 2: Joining years of datasets together.



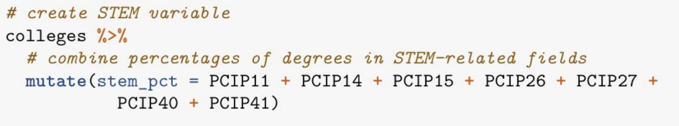
```
# split data into train/valid/test sets for modeling
set.seed(1)

college_parts <- resample_partition(colleges ,c(train = 0.6, valid = 0.2, test = 0.2))

college_parts_train <- as_tibble(college_parts$train)
college_parts_test <- as_tibble(college_parts$test)
college_parts_valid <- as_tibble(college_parts$valid)
```

Snippet 3: Partitioning the dataset.



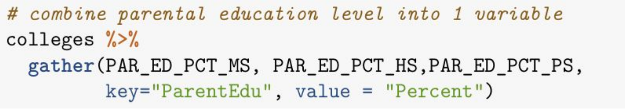
```
# create debt-to-earnings variable
colleges %>%
  mutate(GRAD_DEBT_MDN = as.numeric(GRAD_DEBT_MDN),
         MD_EARN_WNE_P8 = as.numeric(MD_EARN_WNE_P8),
         # median debt/median earnings 8 years after graduation
         DEBT_TO_EARN = GRAD_DEBT_MDN/MD_EARN_WNE_P8)
```

Snippet 4: Creating debt-to-earnings ratio variable, which was ultimately used in the linear model. Debt-to-earnings is a measure of financial security, and specifically is the ratio between debt and income; higher ratios indicate stronger likelihoods to default on mortgages and other debt repayments(16).



```r
# create STEM variable
colleges %>%
  # combine percentages of degrees in STEM-related fields
  mutate(stem_pct = PCIP11 + PCIP14 + PCIP15 + PCIP26 + PCIP27 +
           PCIP40 + PCIP41)
```

Snippet 5: Creating a variable to measure percentage of STEM degrees at an institution.



```r
# combine parental education level into 1 variable
colleges %>%
  gather(PAR_ED_PCT_MS, PAR_ED_PCT_HS,PAR_ED_PCT_PS,
         key="ParentEdu", value = "Percent")
```

Snippet 6: Creating a parental education level variable by gathering variables (highest level of education being middle school, high school, or post-secondary) together.

II. Relevant Figures -

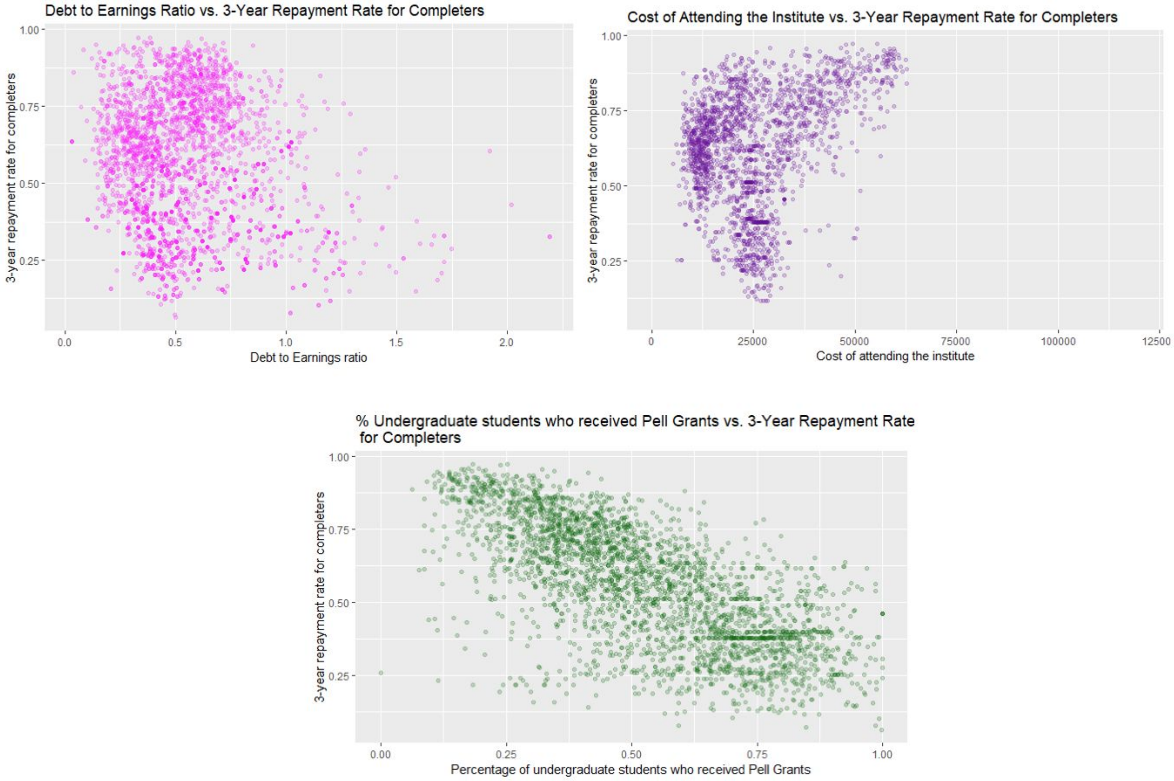


Fig. A: Correlations between 3-year repayment rate and debt-to-earnings, percentage of undergraduates receiving Pell grants, and the average cost of attendance. These variables were included in the linear model, with appropriate transformations; similar trends observed across all years looked at for analysis. Reference year: 2013-14.

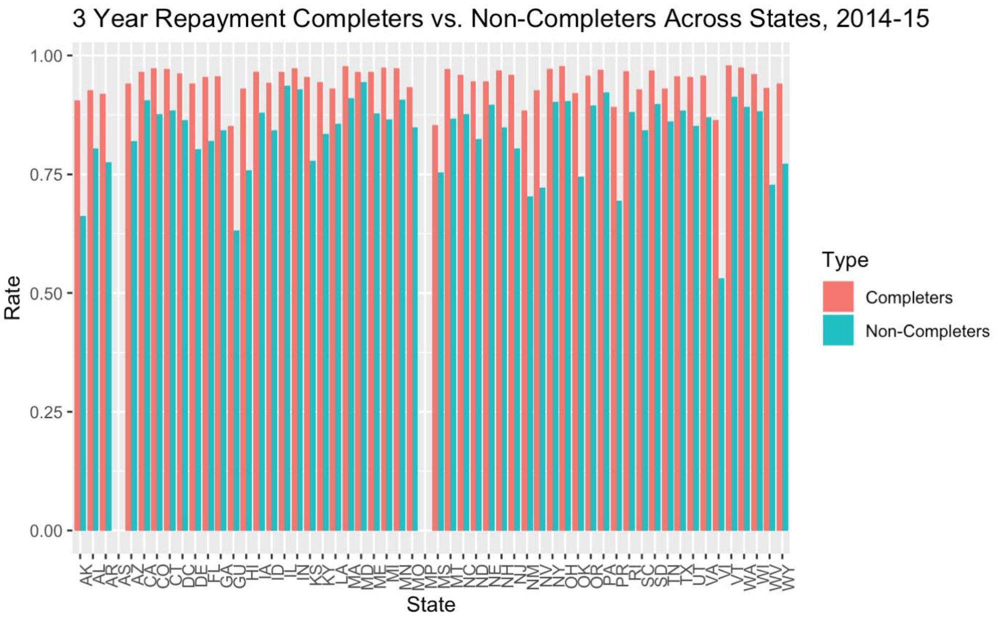


Fig. B: The 3-year repayment rates for completers and non-completers by U.S. state for 2014-15. This data was difficult to visualize nationally, especially when additionally faceted by year, therefore we decided to conduct much of the exploratory data analysis on a local level, as seen in the report.

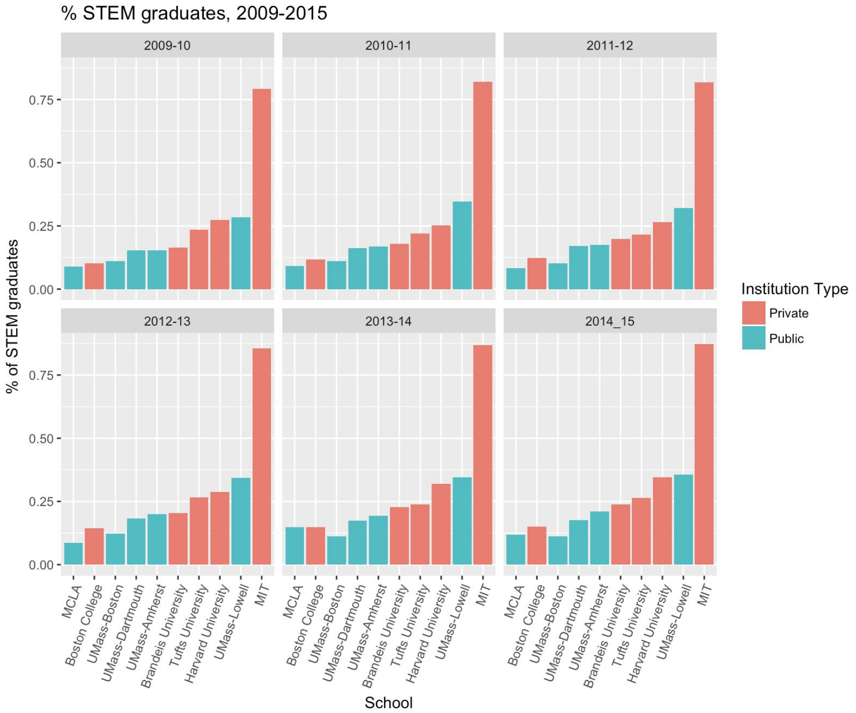


Fig. C: The percentage of graduates in a STEM field at the top 5 public and private universities in Massachusetts from 2009-2015. We can see that MIT overwhelmingly has the highest percentage of these graduates.

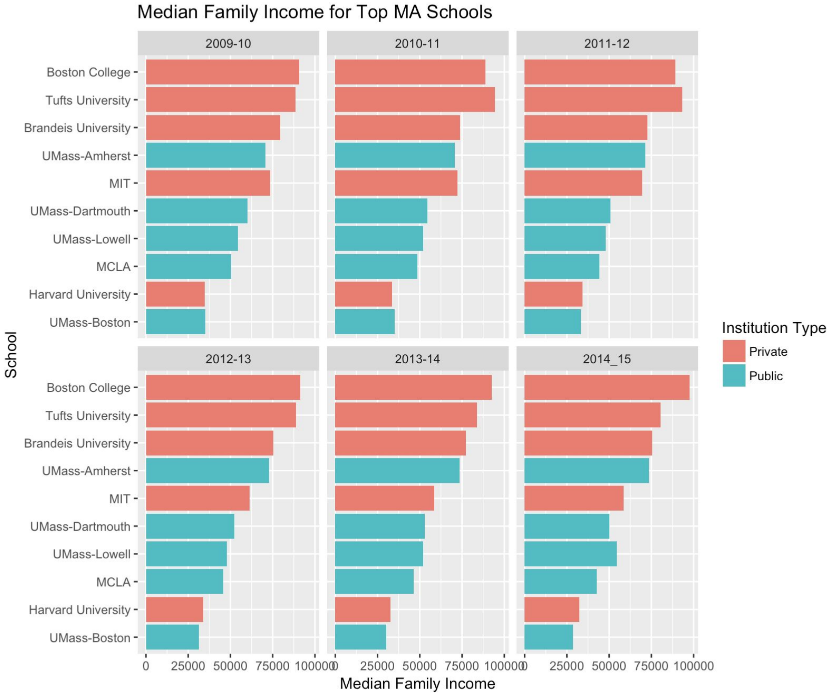


Fig. D: The median family income at the top 5 public and private universities in Massachusetts from 2009-2015. We observe that Boston College and Tufts University students consistently have higher median family incomes. Interestingly, Harvard University has the second lowest median family income reported across all years.

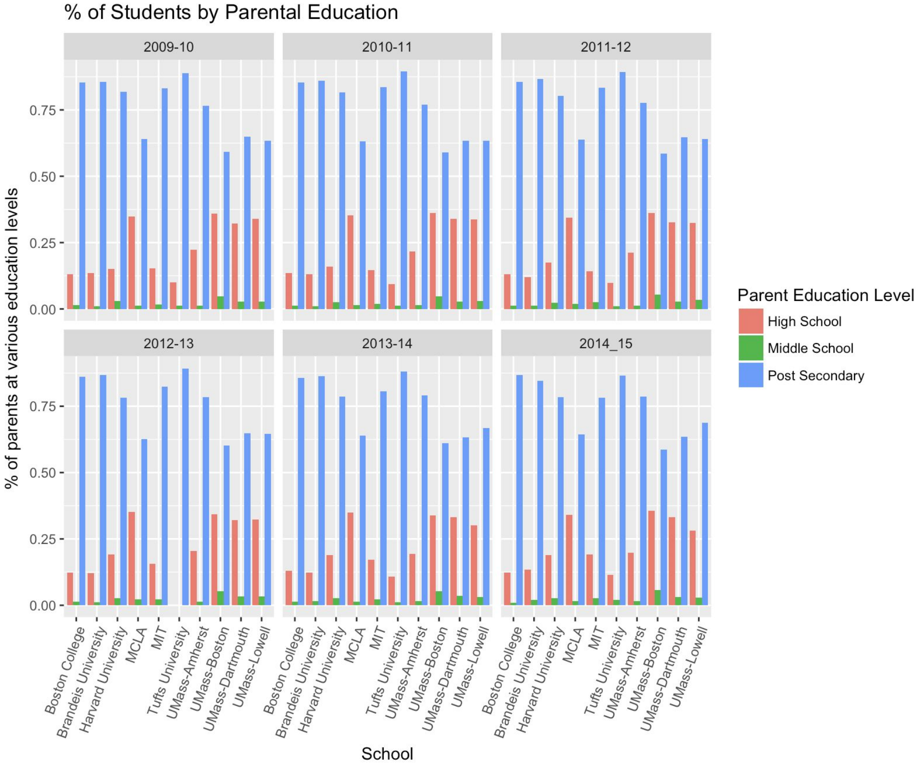


Fig. E: The percentage of students at the top 5 public and private universities in Massachusetts, broken down by parental education level, from 2009-2015. We see that overall, across all schools, most parents have completed some sort of post-secondary education. There is also a higher percentage of parents with a high school as their highest level education for public schools versus private schools.

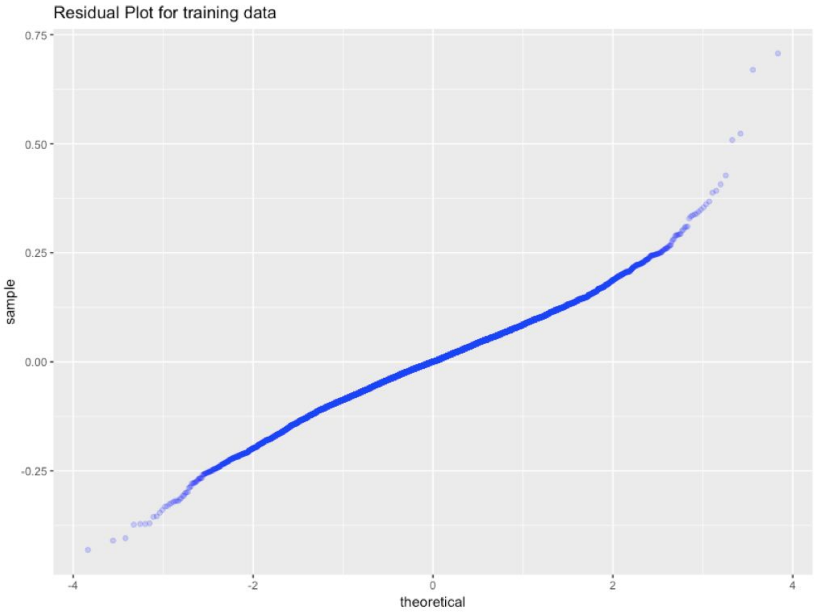


Fig. F: Quantile plot for training data. Checks normality assumptions for linear model.

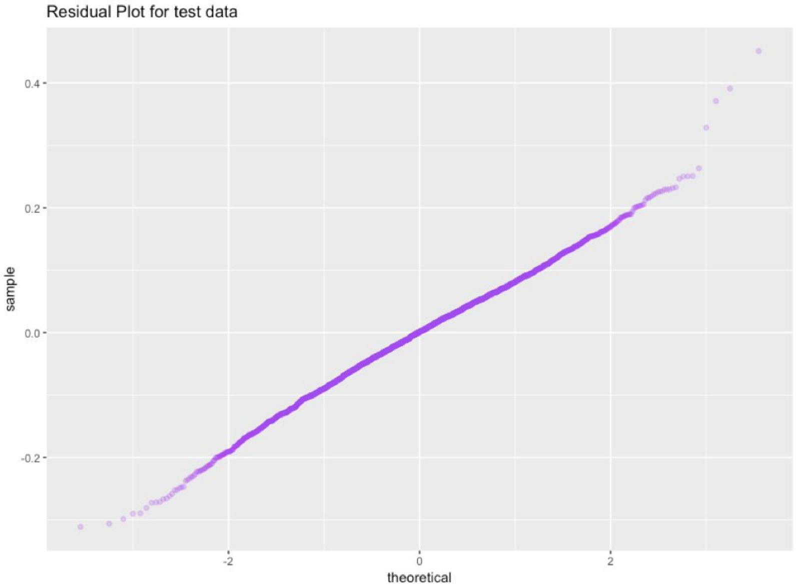


Fig. G: Quantile plot for test data. Checks normality assumptions for linear model.

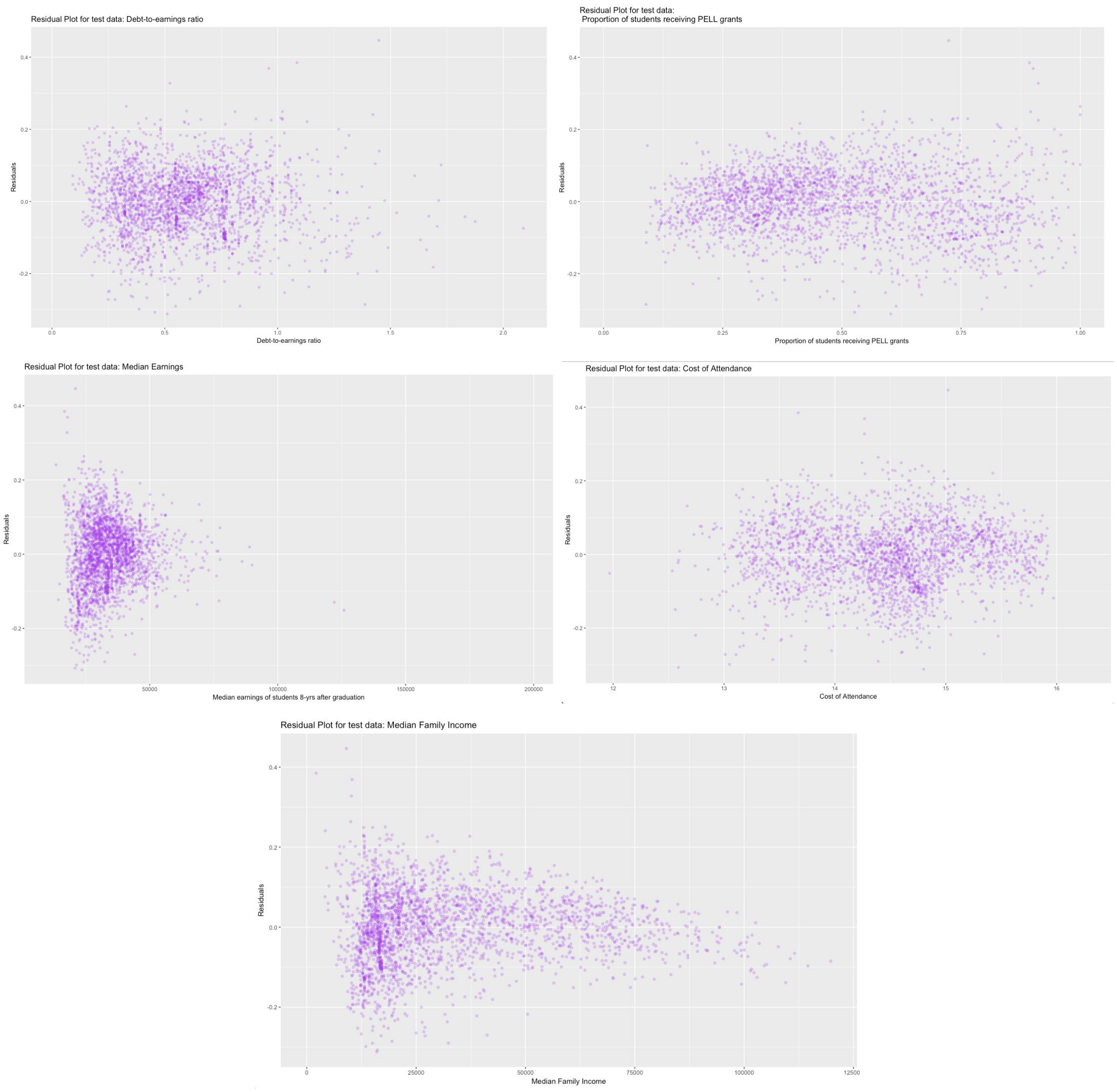


Figure H: Residuals plots of predictor variables included in the linear model.

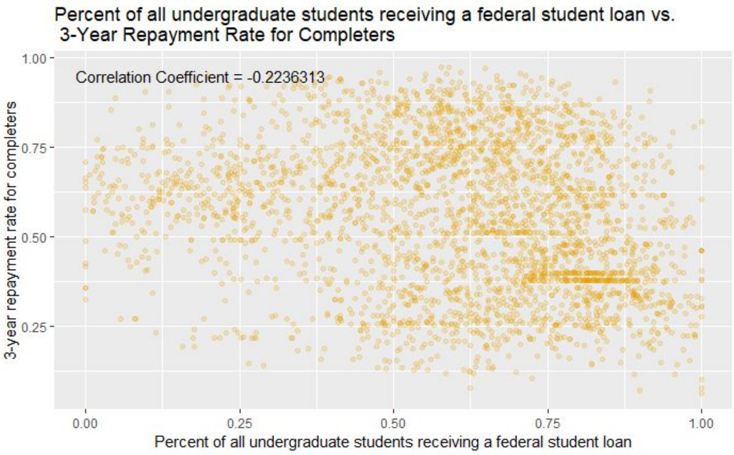
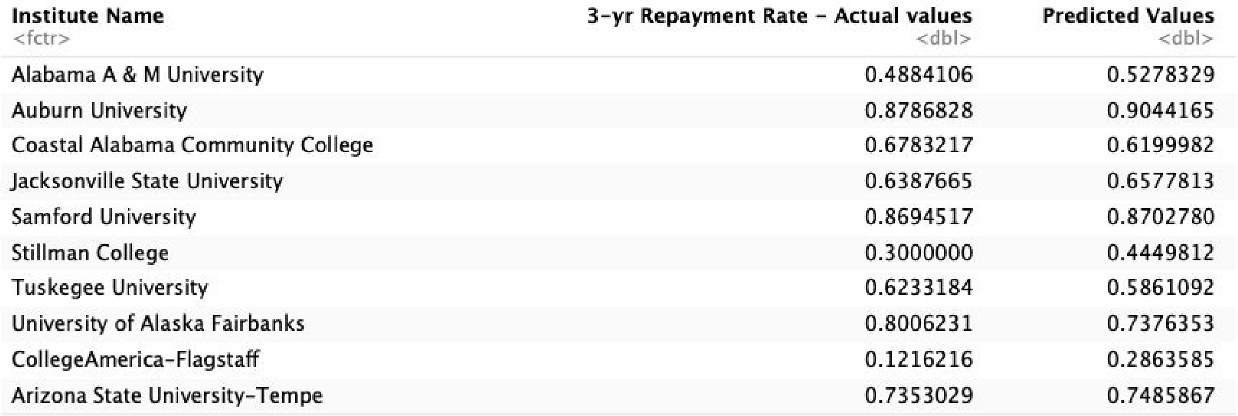


Figure I: Correlation between 3-year repayment rate and percentage of students receiving federal loans. We see no relationship between these variables, and therefore we did not add this variable to our linear model. This is an example of a variable we determined as not relevant for the model. Reference year: 2013-14.



Institute Name <fctr>	3-yr Repayment Rate - Actual values <dbl>	Predicted Values <dbl>
Alabama A & M University	0.4884106	0.5278329
Auburn University	0.8786828	0.9044165
Coastal Alabama Community College	0.6783217	0.6199982
Jacksonville State University	0.6387665	0.6577813
Samford University	0.8694517	0.8702780
Stillman College	0.3000000	0.4449812
Tuskegee University	0.6233184	0.5861092
University of Alaska Fairbanks	0.8006231	0.7376353
CollegeAmerica-Flagstaff	0.1216216	0.2863585
Arizona State University-Tempe	0.7353029	0.7485867

Figure J: Table: Actual versus predicted values for response variable, 3-year repayment rate.