Beyond the Brochure: A Predictive Model for Determining Cost-Effective Education

Zach Quinn

Winter 2020-21

Portfolio Link: https://zachlq.github.io/ZachQuinnDSCPortfolio/

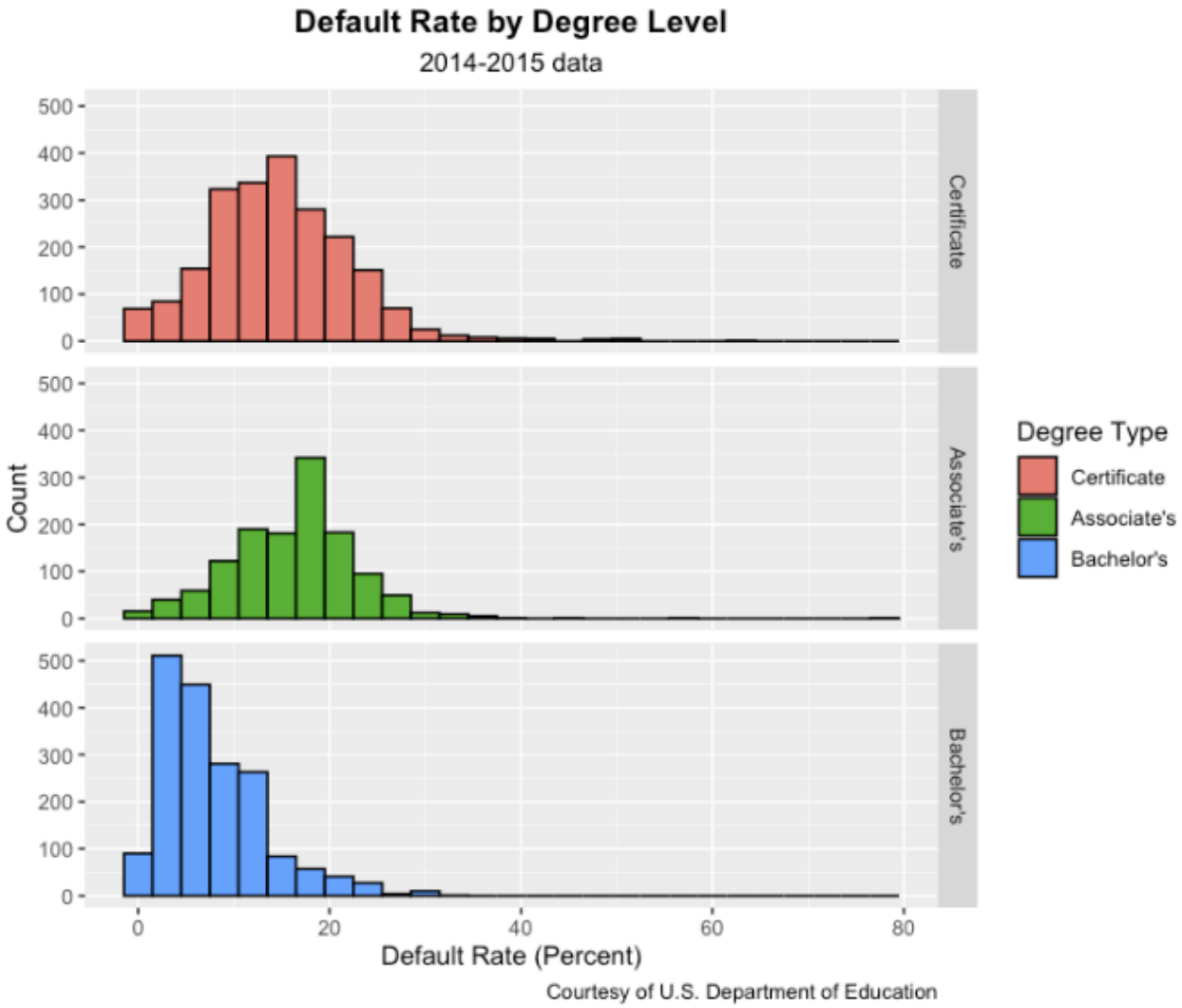
Bellevue University

Introduction

For decades, choosing an institute of higher learning has ranked among the most significant of life choices simultaneously representing a milestone of academic, biological and professional maturity. However, in recent years, the ideology of the perfect campus experience has been replaced by a preference for value over prestige. Nearly one-fifth (18.6 percent) of students admitted to their first-choice college for the 2016 – 2017 declined admission for financial reasons (Seltzer, 2017). Instead, these students sought a more cost-friendly option such as an in-state school or a junior college that afforded them the ability to transfer to a four-year institution. The percentage of students concerned about the cost of a degree was nearly double that of those concerned about other issues such as campus environment, financial aid and proximity to their homes. The Royall & Co. study is not an isolated incident. Incidentally, a USA Today reader poll determined that 45% of over 40,000 participants opted out of attending their first-choice college because of cost (Market, 2019). It must be noted that both the Royall & Co. and USA Today studies reflect pre-COVID ideals. If anything, the pandemic has underscored the importance of cost in choosing a college. For the 2020-2021 classes (and perhaps beyond), since on-campus activities have been limited, students will be judging universities based almost entirely upon the strength of their online instruction and overall educational merit. Consequently, 93% of college students surveyed during the 2020 school year believe that tuition should be reduced for an online-only experience (Hess, 2020). Even given this increased attention on cost (and the reluctance of students to pay for a watered-down educational experience), colleges do not prioritize data related to affordability in their marketing campaigns. Instead, institutions tout academic offerings, comfortable dormitory spaces and the value of building community. They avoid the financial reality of incurring debt, facing delinquency and defaulting. Therefore, the goal of this project will be to develop a linear regression and CART model that predicts whether students will default on outstanding debts. Visualizations will help to frame the financial problems inherent in the college selection process.

Business Problem

Since most of the research regarding college preference with regards to finance focuses on graduating high school students, this project will focus on an all-too-often forgotten college demographic: The transfer student. Specifically, this project will focus on the population of transfer students. Unlike an 18 – 22 college student, these individuals are typically working adults who earn low to median incomes (approximately 30,000 – 48,000 dollars per year, based on College ScoreCard data definition for low-income) and will have a far tougher time searching for a college to finish their degrees while balancing full-time jobs and families. Additionally, since company benefits were either reduced or cut entirely, these individuals are suddenly in the position to choose to pay for the remainder of their degree, or seek out more cost-effective options. Since education benefits were eliminated, these individuals will also need to seek out financing channels such as grants, loans and scholarships in order to be able to afford a four-year undergraduate degree without accumulating an insurmountable amount of debt.



Hypothesis

The hypothesis of this project is: Can we leverage available federal education data to fuel a predictive model that determines whether transfer students would be at risk of defaulting on debt?

Methodology

The focus of this project is developing an easy-to-navigate visual college finance guide for transfer students who have suddenly lost their company-sponsored tuition assistance. The primary source data, the College Score Card was initially obtained from the U.S. Department of Education's website and cleaned using Python's pandas library. Over 2,000 variables were reduced to 38 numeric variables, representing pertinent financing information such as the number of pell grants awarded in a given academic year. A Pearson's correlation test was conducted on the numeric variables to select those that might display correlations that would be the most useful for a student seeking to understand a school's financial costs. A supplemental dataset was downloaded from a Kaggle repository gleaned from TuitionTracker.org as well as

the National Center for Education Statistics. Four datasets, comprising historical tuition data, school demographic data, salary potential and tuition income were combined using R's tidyr library and visualized in ggplot2. Both the Score Card data and Kaggle data were used to make accessible visualizations in Tableau.

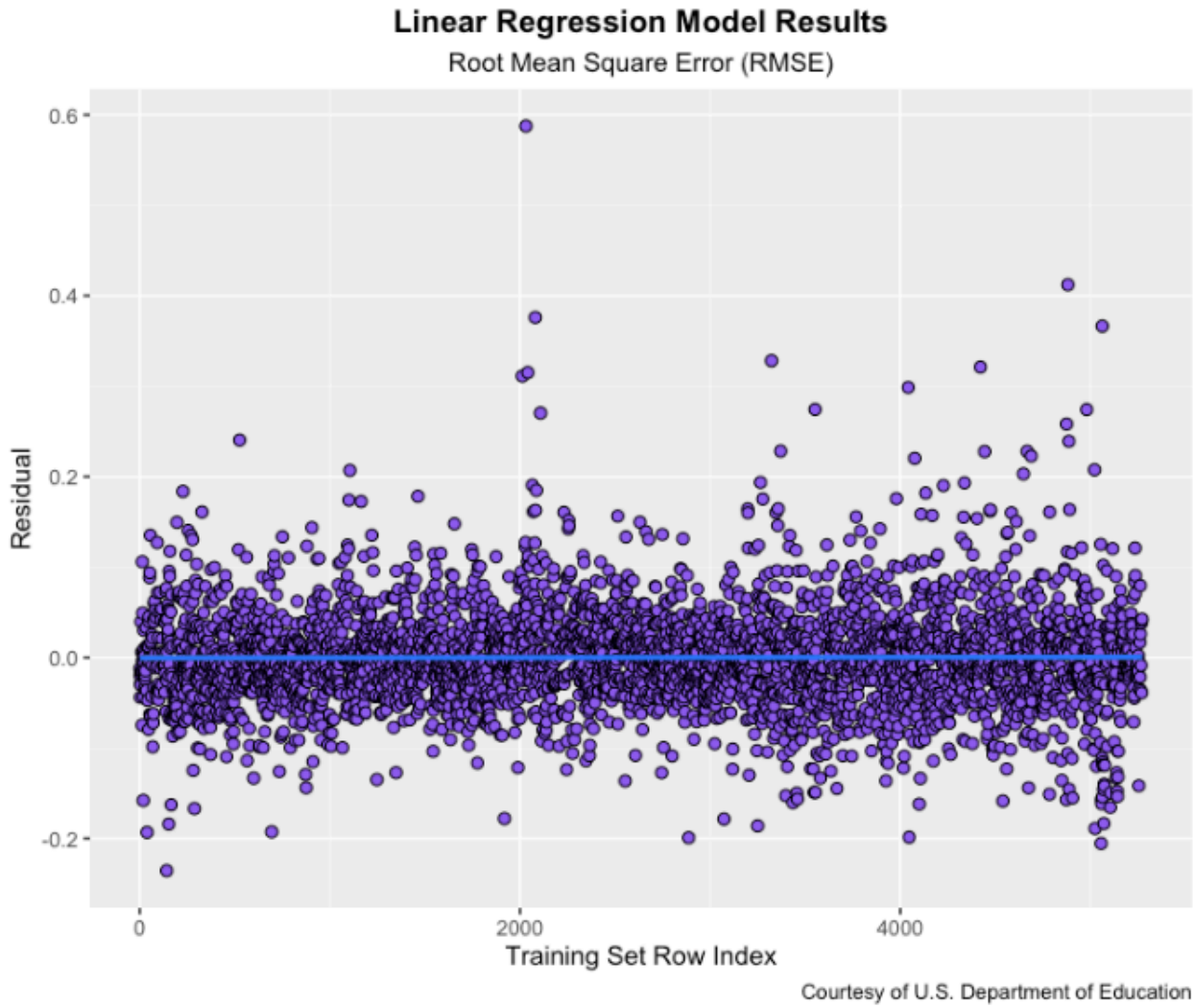
The project utilized two predictive modeling algorithms: Linear regression and Classification and Regression Trees (CART). While linear regression allows for the manual selection of multiple variables to create a linear model, CART automates the selection of variables, choosing only the features that have the most significant impact on a chosen target. Root mean square error was used as the predominant performance metric, along with residuals. 19 features fueled the linear regression algorithm, including (but not limited to): Percentage of students awarded a Pell Grant, debt at graduation, degree type and first generation student. Conversely, the CART chose average family income (dependents), average family income (independents), net tuition revenue per full-time student, and median debt for students who have completed an academic program as the most significant features. Ggplot2's library was utilized to visualize results for the linear regression model, while the rpart library helped to visualize the resultant CART model. Ultimately, the project will seek to offer a solution that aligns with the student's academic goals as well as an option that minimizes their debt-to-income (DTI) ratio upon graduation.

10 Research Questions

1. Which category of schools, on the whole are more affordable (private, public or for-profit?)
2. Which degree programs are more financially viable (two-year of four-year)?
3. What percentage of four-year transfer students qualify for Pell grants/federal aid?
4. What are historical averages for tuition rates?
5. What correlation (if any) exists between tuition paid and recent graduate earnings?
6. What correlation (if any) exists between tuition paid and mid-career graduate earnings?
7. What is the average amount of debt a four-year college student accumulates?
8. Which institutions are more affordable: In-state or out-of-state colleges?
9. What correlation (if any) exists between tuition paid and faculty salary?
10. What tuition threshold can a low-income (defined as 30,000 to 48,000 by U.S. Department of Education) student reasonably afford to pay?

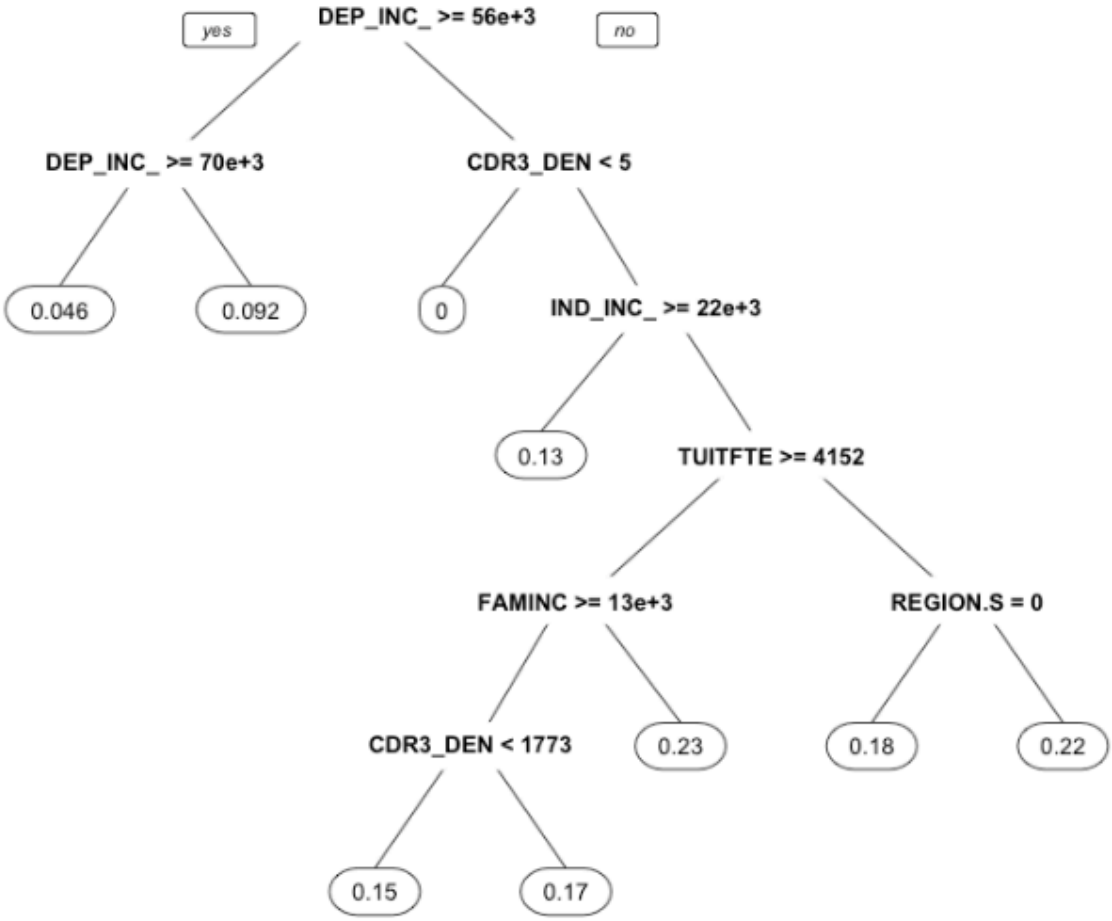
Results: Linear Regression

Linear regression was, by a slim margin, the more precise model of the chosen methodologies. After creating a linear model of student default rate (CDR3) as a function of the 19 features, the model was fairly accurate in its predictions. The resultant summary indicated an RMSE of 0.052. This is a statistically significant result because an ideal RMSE is a value closest to zero. A visualization of the model's performance on a scatter plot is included below.



Results: Classification and Regression Tree (CART)

While the CART performed slightly worse than the linear regression model, it was successful in choosing variables that would serve as the best predictors for the outcome of defaulting on debt. The four variables that CART chose included: Average family income (dependents), average family income (independents), net tuition revenue per full-time student, and median debt for students who have completed an academic program. At 0.053, CART's RMSE is marginally larger than that of the linear regression model. This means that the linear regression model is more precise in the scope and accuracy of its predictions whether a student will default on debt. The regression tree's performance is excerpted below.


yes
DEP_INC_ >= 56e+3
no
DEP_INC_ >= 70e+3
CDR3_DEN < 5
0.046
0.092
0
IND_INC_ >= 22e+3
0.13
TUITFTE >= 4152
FAMINC >= 13e+3
REGION.S = 0
CDR3_DEN < 1773
0.23
0.18
0.22
0.15
0.17

Conclusion

While linear regression and CART as methodologies are not incredibly complex or as powerful as deeper learning neural nets like CNNs, their ability to combine multiple features is helpful in not only making a prediction, but determining which demographic or financial qualities puts a student at a higher risk of default. To be clear, the goal of this project is not to recommend any kind of overhaul to the current ecosystem of student debt creditors and borrowers. Instead, a prospective transfer student should closely examine and consider the data included in this paper and accompanying presentation as they decide whether to finish at their current institution, transfer or abstain from higher education in favor of more cost-effective options. Even though the data used in this presentation is from a few years ago, it should still serve as a stark reminder of how even those with seemingly valuable degrees, including bachelor's degree students, can become delinquent or default on their debt in a short period of time. Additionally, as the visualizations demonstrated, those with smaller amounts of debt, such as transfer students, who only have a portion of a program to complete, are at a higher risk for becoming delinquent or defaulting on their student debt. Hopefully, any prospective transfer student reviewing this information will choose an institution that fits their academic needs while maintaining an equitable debt-to-income ratio for future professional and financial wellbeing.

References

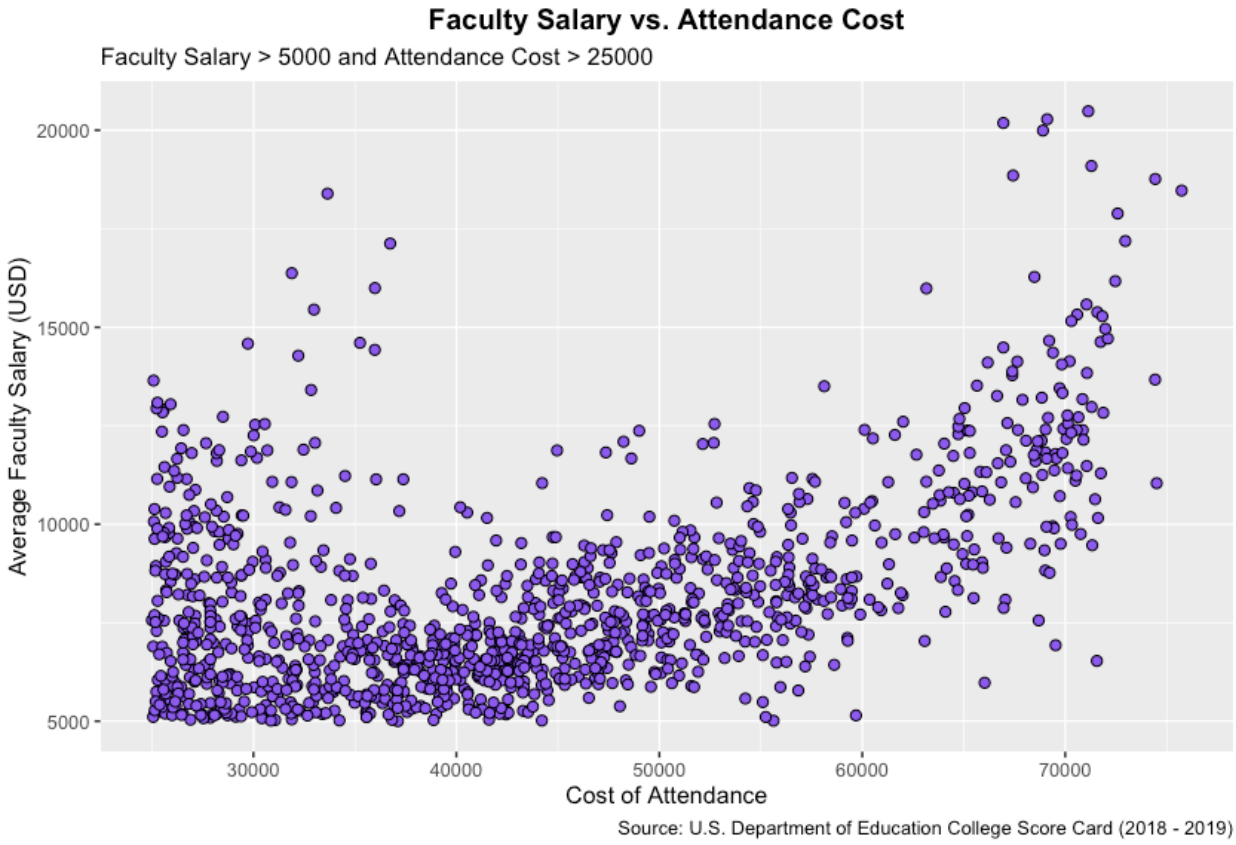
Hess, A. (2020). More than 93% of U.S. College Students Say Tuition Should Be Lowered if Classes are Online. CNBC. https://www.cnbc.com/2020/07/27/93percent-of-college-students-say-tuition-should-be-cut-for-online-classes.html

Kagan, J. (2020). 28/36 Rule. *Investopedia*. https://www.investopedia.com/terms/t/twenty-eight-thirty-six-rule.asp

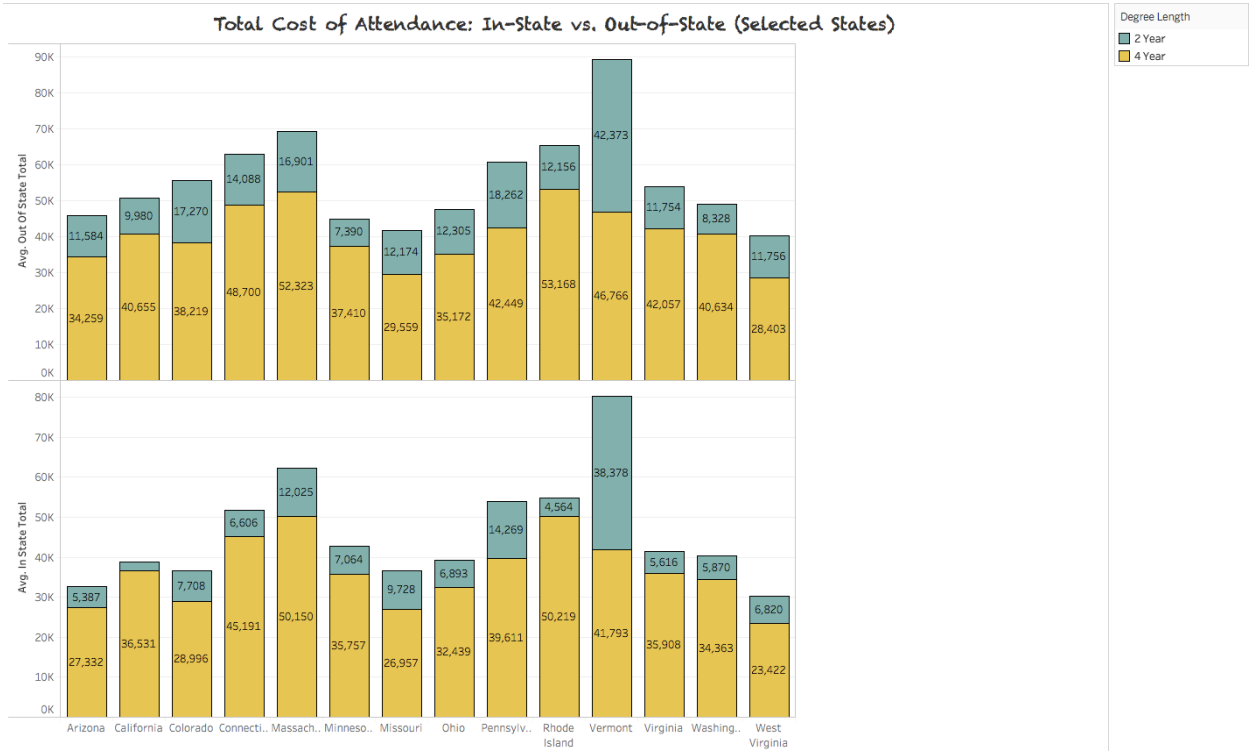
Market, J. (2019). Twitter Poll: Cost is top factor when choosing a college. https://www.usatoday.com/story/sponsor-story/college-ave-student-loans/2019/05/30/twitter-poll-cost-top-factor-when-choosing-college/1290536001/

Seltzer, R. (2017). Turning Down Top Choices. Inside Higher Ed. https://www.insidehighered.com/news/2017/03/23/study-shows-how-price-sensitive-students-are-selecting-colleges

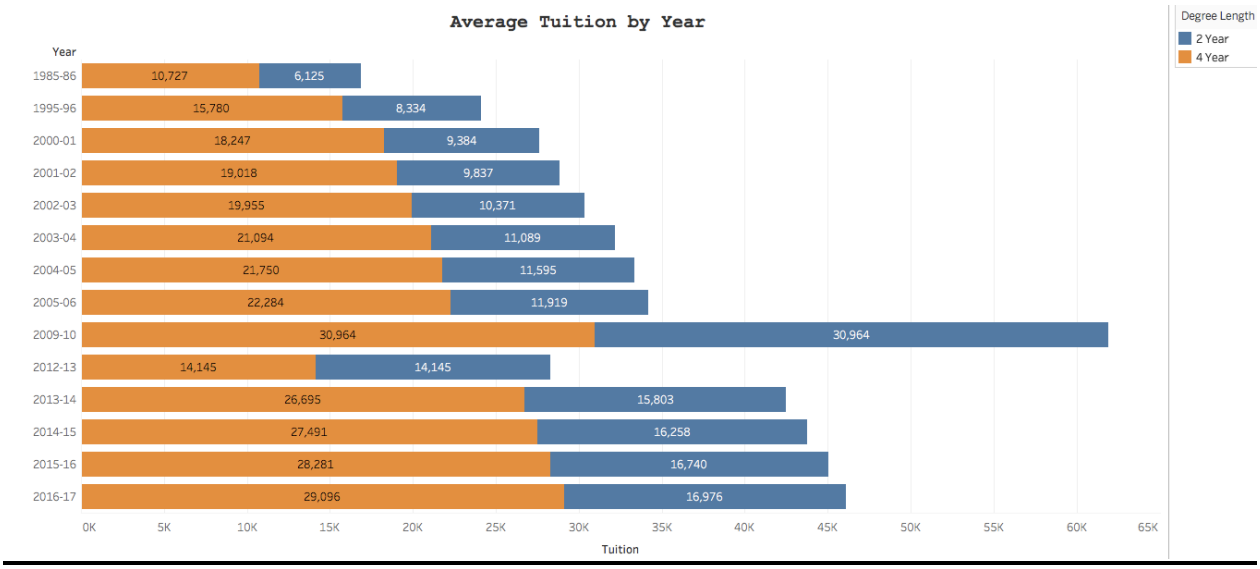
Appendix A: Faculty Salary vs. Cost of Attendance



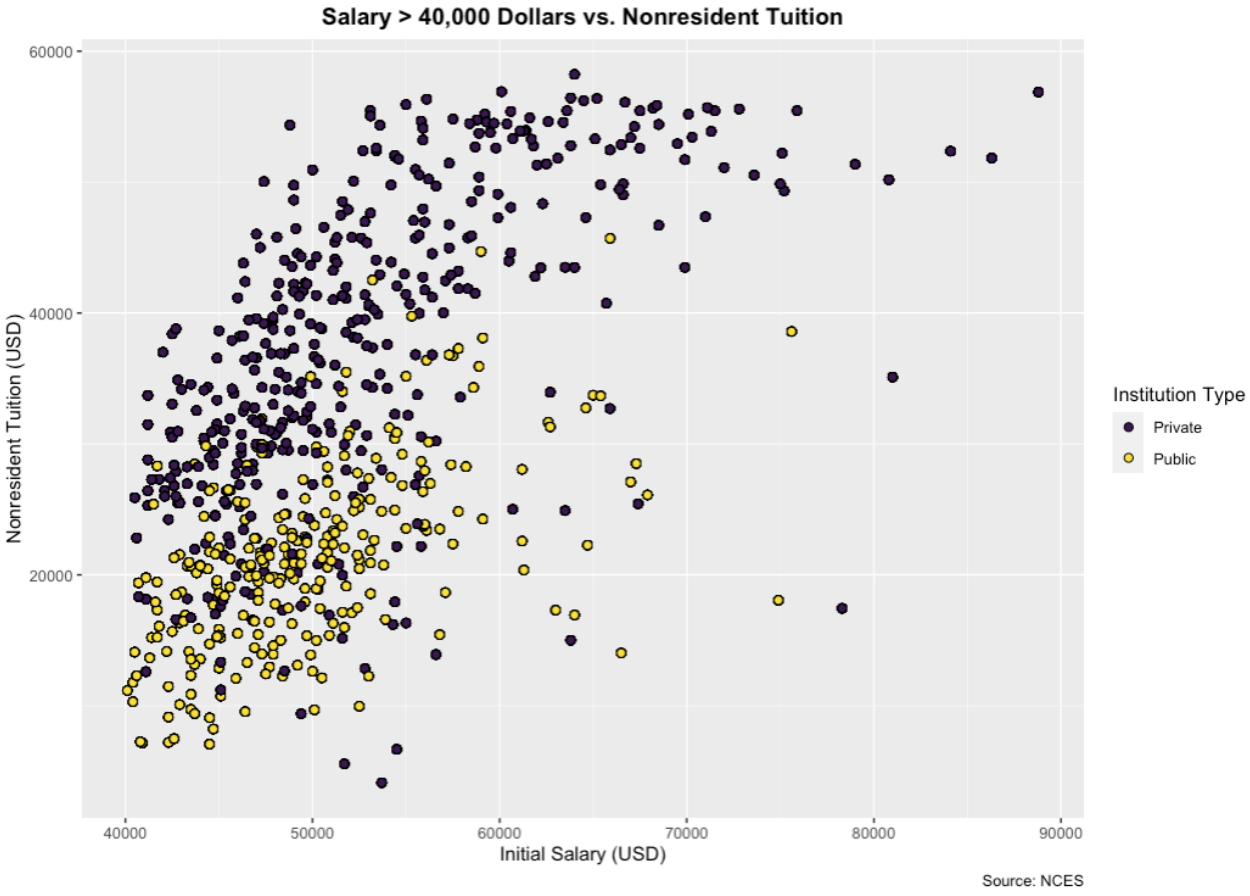
Appendix B: In-State vs. Out-of-State Attendance Costs



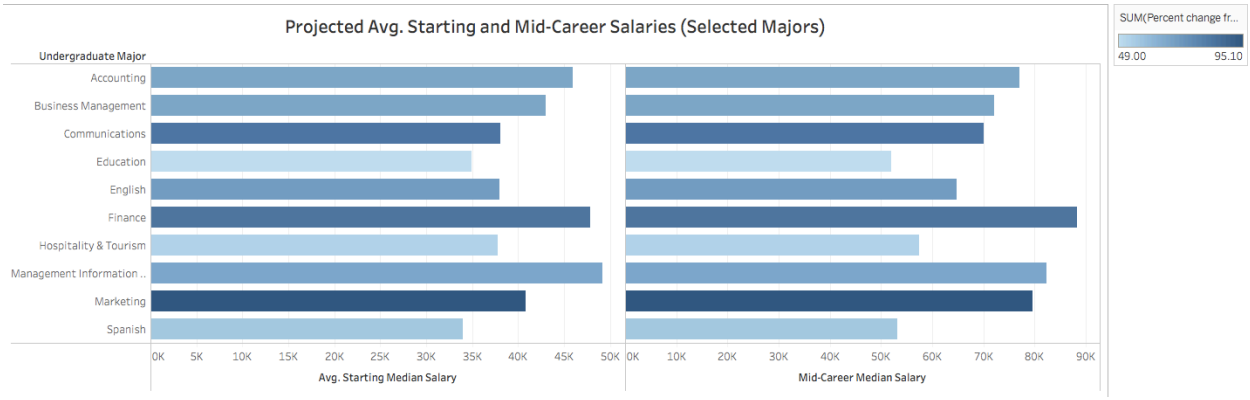
Appendix C: Tuition by Year



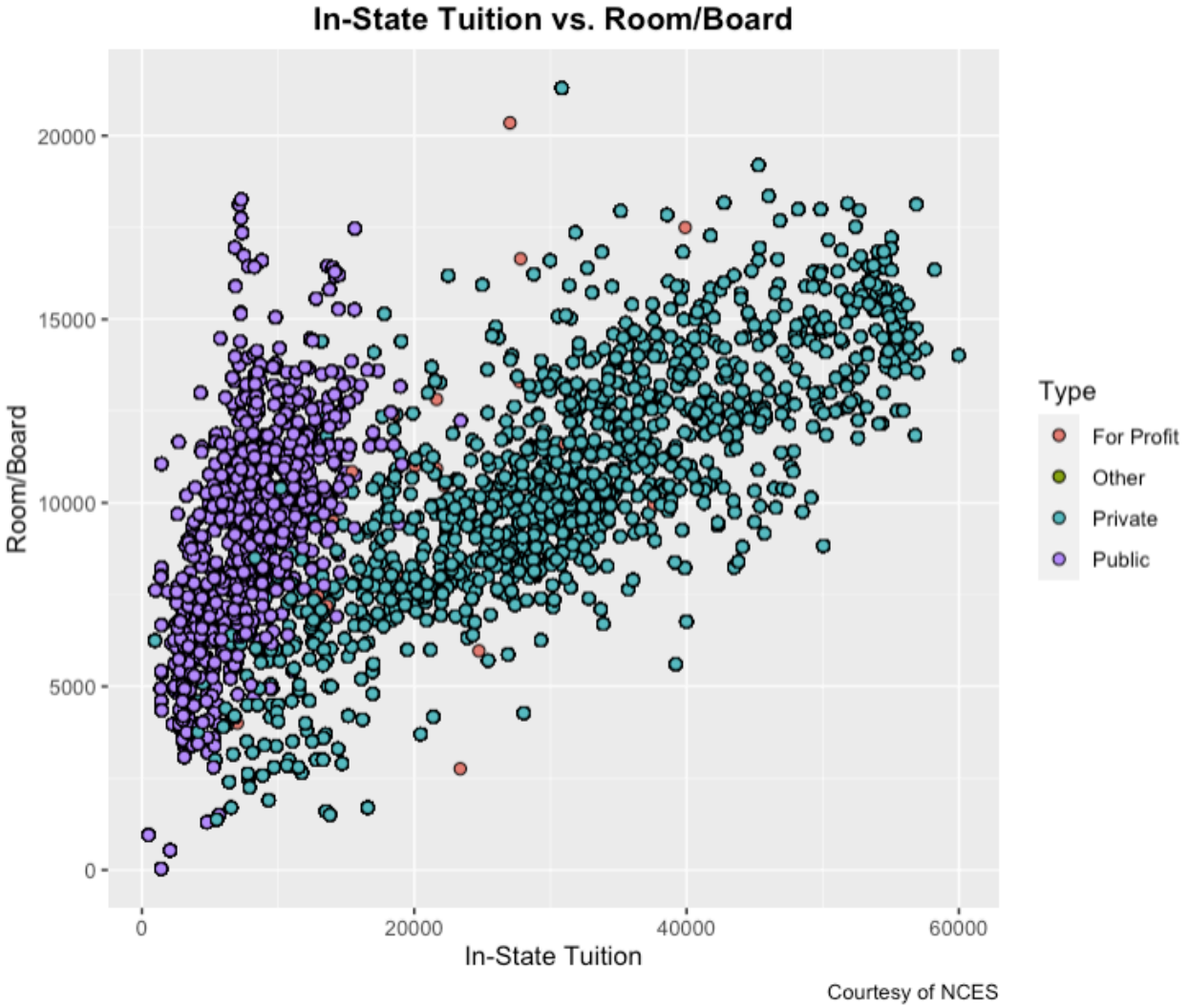
Appendix D: Salary Potential vs. Nonresident Tuition Cost



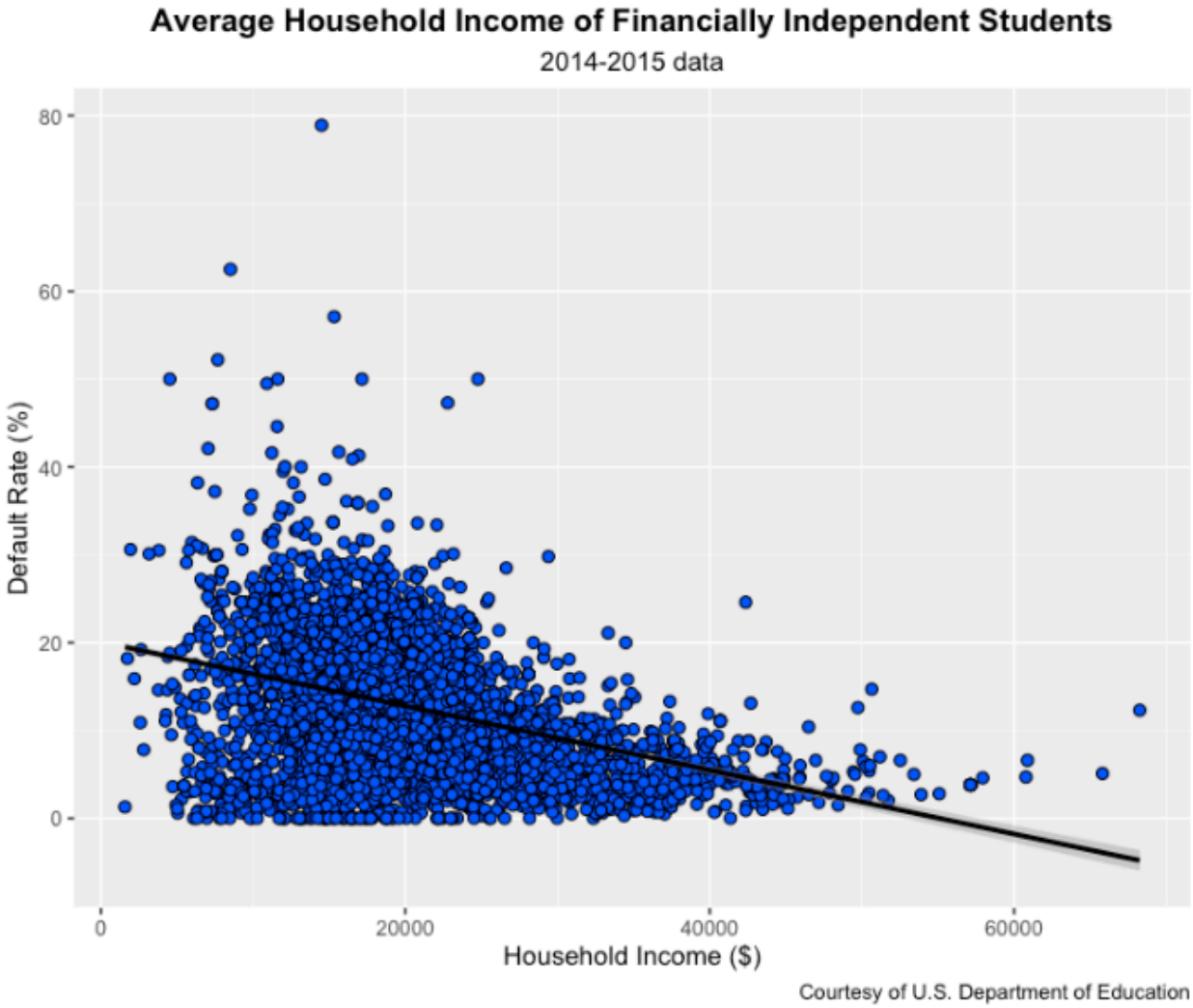
Appendix E: Projected Starting and Mid-Career Salary (Selected Majors)



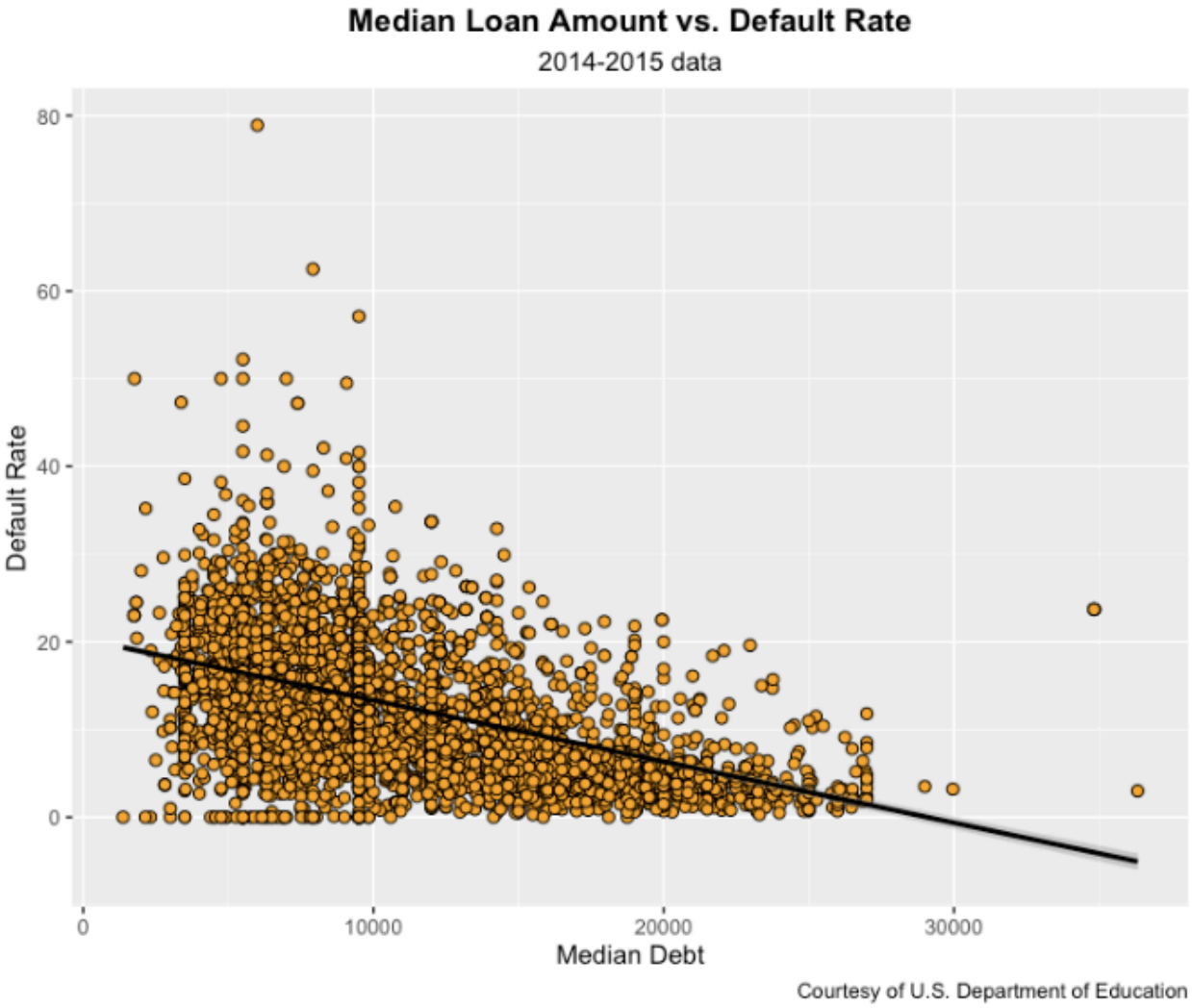
Appendix F: In-State Tuition vs. Room and Board



Appendix G: Average Household Income of Independent Students



Appendix H: Median Loan Amount vs. Default Rate



Appendix I: Data Links

College Score Card: https://collegescorecard.ed.gov/data/

Kaggle: https://www.kaggle.com/jessemostipak/college-tuition-diversity-and-pay?select=historical_tuition.csv