Affordability vs Achievement

Peyton Sharpe, Keira McArthur, Joe Conrad, Stephan Cipcigan, Andy Huarcaya

Department of the School of Data Science, University of North Carolina at Charlotte

DTSC 1302: Data and Society

Professor Marco Scipioni and Ilieva Ageenko

Date 12/08/24

Examining the Impact of College Costs on Graduation Rates in Private vs. Public Institutions

Higher education affordability is a crucial factor that influences not only student success within an institution but the institution's effectiveness as well. The research question that our group selected places interest on affordability metrics like room and board, out-of-state tuition, book expenses, and personal spending and how they relate to graduation rates at private and public colleges nationwide. Examining these selected variables, and exploring the relationships between each factor, the impact of financial factors on education outcomes may become increasingly more evident. The research conducted was extracted from data cleaning college universities across the United States to identify trends and correlations between the factors discussed previously. Using predictive modeling, the data can be transformed into quantitative affordability metrics linked to graduation rates. This report, aimed at informing policymakers and institutions about higher education, will address this central question. How do affordability metrics impact graduation rates across private and public colleges in the United States?

Context and Implications

Affordability in higher education has always been a topic of interest among students, policymakers, educators, and even researchers as a whole. Having a very crucial role in accessibility and student success, it is a topic that continues to raise questions today. Throughout our research, we have seen the financial barriers proposed, such as tuition costs, living expenses, books, etc., that effectively impact low-income students in turn contribute to lower graduation rates as well as widening these ever-so-evident social disparities (Andreae, 2024; Bell, 2023). Throughout not only our research but also other work in the industry, a clear correlation between affordability and education has been discovered.

The stakeholders in this research include people like students, educational institutions, and policymakers. Students, in particular, are directly impacted as affordability influences their ability to

persist and succeed in higher education (Bell, 2023). These institutions greatly benefit from increased graduation rates making a higher demand evident through an enhanced reputation (Bareham, 2023). Policymakers who are tasked with creating equitable funding policies while addressing these budget constraints from operational costs are also greatly impacted. Of course, there may be a conflict of interest, as the measures of "improving" affordability may require reallocating institutional resources or even increasing public funding (Hood, 1988).

Ethical considerations regarding the potential biases evident in the dataset must be addressed, especially in a large study. The data studied may not fully capture diverse student populations or experiences. For example, part-time students or students with non-traditional educational backgrounds may not have the same student experience as a "regular" student (Data Home: College Scorecard, 2024). Unforeseen student populations that may have childcare costs or even emergency costs that do not relate to the main student population may uncover potential biases. Recognizing these limitations is extremely important to make sure that the analysis provides an accurate representation of the data, as well as the relationship between graduation rates and affordability (Bauer et al., 2022).

Measurement

We aimed to examine the effect of affordability metrics on graduation rates at different kinds of colleges, public vs private universities. In doing this, we analyzed large datasets and identified key factors to explore the relationship between college costs and graduation rates. In particular, we looked at how costs like tuition, fees, room and board, and textbooks affect students' ability to graduate at different types of schools. So our goal was to provide trends and conclusions that may aid policymakers and colleges in understanding the impact of financial burdens on student success. This research shows that higher college costs are linked to lower graduation rates. In conclusion, our study helps talk about ways to make college more affordable so more students can graduate.

To make sure our research is clear and correct, we focus on a few important numbers. We use correlation to see how graduation rates are connected to college costs (like tuition and housing). R-squared (R^2) is key for showing how well our model explains graduation rates based on college expenses. This explains how changes in college costs can affect graduation rates.

Mean Squared Error (MSE) is a number that shows how close our predictions are to the real results, helping us see if our model is good. The starting point of our model tells us what we think the graduation rate would be if college didn't cost anything.

Another key variable is coefficients. In the model, we show how a change in college expenses can impact graduation rates, which ultimately helps us to understand how likely students are to graduate. By looking at these important factors, we hope to show how college costs influence graduation at different schools.

In our process for operationalizing our key variables, we needed to carefully choose and analyze the important factors. This started with making simple graphs like scatter plots and histograms to visualize the data. These methods essentially helped evaluate the distribution and connections between the variables, which gave a basic understanding necessary for further study. We next used regression analysis to determine coefficients for several kinds of cost measures, including things like room and board, tuition, and book expenses. This step was especially important to measure the direct impact of changes in these expenses on graduation rates. By using these approaches, we hope to provide a complete and thorough understanding of how indicators affect students and inform policy choices to support educational institutions in developing plans to improve student results.

Data

The data manipulation process for our research question involved several key steps that integrated materials and concepts introduced and learned in class. This methodology will ensure that the dataset is well prepared for analysis and interpretation, furthering our conclusions on

what the data is trying to tell us. The dataset is taken from a public repository on GitHub and then imported into Python using pandas. This was a fundamental tool we used to collaborate, introduced during class time, and explored on our own. First, exploratory data analysis was performed to get a feel for the dataset. This included viewing the first few rows using "via.head()", and descriptive statistics "via.describe()", and checking for missing data or errors by using methods such as "isnull()". The base techniques listed here reflected foundational EDA practices covered in class and provided a structured path forward in understanding the data and being able to see what we were working with.

Numeric columns were isolated and drawn out using the select_dtypes() function, focusing on affordability-related variables regarding graduation rates, including 'Room.Board', 'Personal', 'Outstate', and 'Books'. We used this to target specific variable types for statistical analysis and visualization. These variables were chosen because affordability metrics and academic outcomes were central to the research questions of the project. One of the most important aspects of this that was an addition to our research was an analysis to compare and contrast public and private colleges.

By categorical filtering, the dataset was divided into two sets, using the Private column of the data. The colleges were categorized as either private or public with yes or no in the dataset. This step implemented skills learned in class providing a means to directly compare metrics of affordability and their impacts on Grad. Rate. Each was analyzed separately, showing differences between these factors and graduation rates by institution type. We used further skills learned in class to combine subsets into a single data frame with a new categorical variable called Type that categorized colleges as public or private.

This allowed for a comparison of visual and statistical differences between the groups. Box plots and pair plots were made for comparisons among distributions and relationships of the variables, demonstrating the effectiveness of visual analytics in conveying such information. The above methodologies are implemented using matplotlib and seaborn, which prove to be a very good way of representing data effectively. To prepare the dataset for predictive modeling, the data was split into training and testing sets using the train_test_split() function from the scikit-learn library. The features/predictors (X) included affordability metrics ('Room.Board', 'Personal', 'Outstate', and 'Books'), and the target variable/outcome (y) was 'Grad.Rate'. These selections reflected features relevant to the research question and ensured they provided meaningful input to our predictive model.

Further, correlation analyses were run using the .corr() method to determine the significant relations among these variables. This was one way we could directly see the correlation without looking at previous charts and graphs such as heatmaps and scatter plots. This step was informed by class discussions on using correlation matrices to guide analysis and focus on impactful relationships.

We used correlation heatmaps to visualize the relationships between the affordability metrics and graduation rates. The heatmaps helped us understand how room and board, personal spending, and other factors were related to graduation rates with notable differences between private and public colleges. For our regression analysis, we built a linear regression model to predict graduation rates based on room and board, personal spending, out-of-state tuition, and book costs. From the model coefficients, increases in room and board costs were associated with slight increases in graduation rates, but higher out-of-state tuition and book costs appeared to

lower graduation rates, suggesting that financial barriers might negatively impact student graduation.

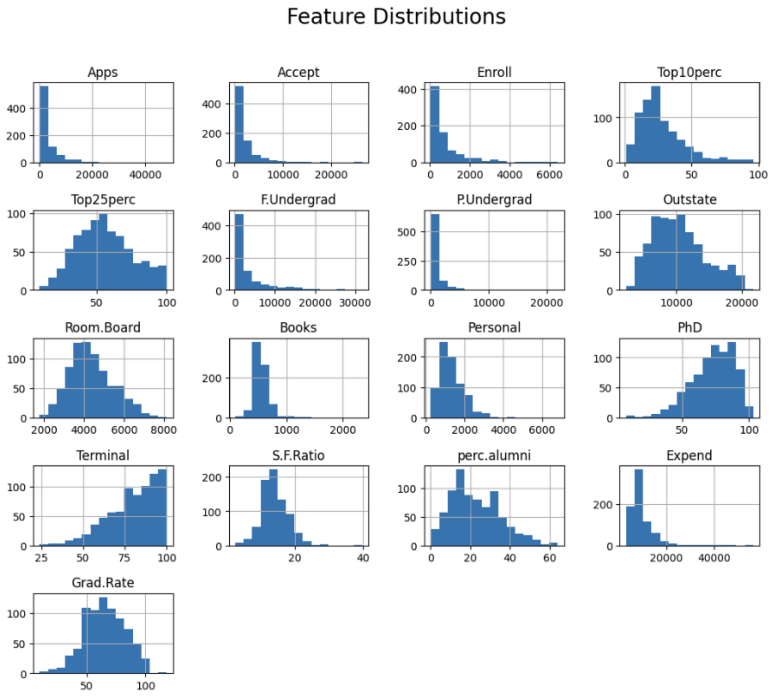
Looking at the model's performance, we see that the MSE was 138.08, which means our model's predictions were fairly off from the actual graduation rates. The R-squared value came out to 0.413, which reflected that it could explain about 41.3% of the variance in the graduation rates. While this suggests that the model captures some of the relationship between affordability and graduation rates, it also highlights that other factors, such as student support services or institutional quality, could be influencing graduation outcomes. For example, when we input sample values for room and board, tuition, personal spending, and books, the predicted graduation rate was -3.27%, which showed that our model may be missing some important factors or not taking into account the complexity of graduation outcomes. We also got an intercept of 71 which translates to the baseline graduation rate when all the affordability metrics (room and board, tuition, personal spending, and books) are zero. This doesn't have any real-world implications but it is important to know when understanding your model.

To compare the affordability metrics between private and public colleges, we used boxplots to visualize the differences in room and board, books, out-of-state tuition, and personal spending. These visualizations helped in understanding how the affordability varies between both types of colleges. We also created histograms to show the distribution of various features and pair plots with regression lines as a way to explore relationships between the affordability metrics and graduation rates. It was overall a valuable analysis, but the low R^2 value and the negative graduation rate prediction appeared unrealistic, so the model could be improved by considering additional factors or refining the current ones. These are all ways that we used to visualize a summary of the outcomes we were given through our project.

Throughout our project, we kept coming back to our original visuals. They very clearly show why our target variables are meaningful. We started our project by creating visuals to gather an idea of which variables would be the most rewarding to study. The first visualization we implemented was a feature distribution. This type of visualization is a key tool used commonly in exploratory data analysis. That provides insight into how the values of individual features are spread across the dataset. A feature distribution is crucial to understanding the nature of each variable and how it might affect the analysis or modeling. We looked for the shape of the distribution, central tendency, and skewness.

Figure 1

Feature distribution for all variables (columns) - Histograms



Note. Created by Peyton Sharpe using Python's Pandas and Matplotlib libraries.

Using the figure above we can easily see the shapes of distributions, central tendencies, and skewness. 'Top10perc' (the percentage of students who are in the top 10% of grades),

‘Top25perc’ (the percentage of students who are in the top 25% of grades), and ‘Grad.Rate’ (percentage of those who finish their degree promptly upon enrolling) have a normal distribution. This indicates that the data in these features are symmetrical and fall around a central point, with most values clustering around the middle. Essentially, this means that most data points fall close to the mean, while fewer fall away from it.

Many features, such as: ‘Apps’, ‘Accept’, ‘Enroll’, ‘Outstate’, ‘F.Undergrad’, and ‘P.Undergrad’ have a right-skewed distribution (long tail to the right). This indicates that most schools have lower values for these features. However, we did observe that there were a couple of features with a left-skewed distribution (long tail to the left), such as: ‘PhD’ and ‘Terminal’. This indicates that most schools have higher values for these features. ‘Books’ and ‘Room.Board’ had very compact distributions, indicating less variability in these features across institutions.

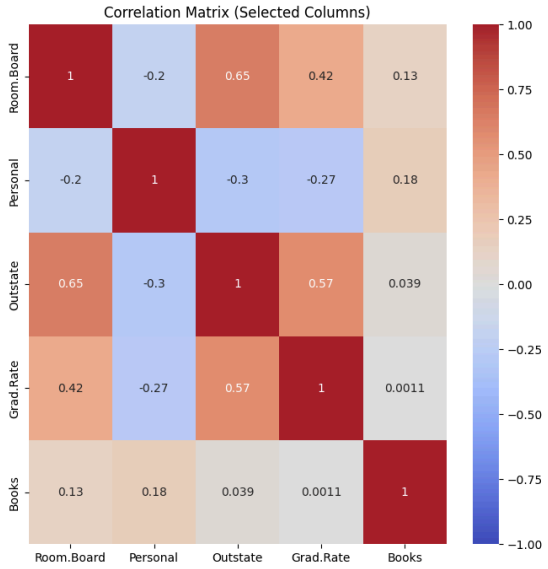
Looking at the feature distribution was reassuring to us as we refined our research question. We knew from the start that we were interested in examining the graduation rate as it is a huge marker for success for both the students and the institutions. Visualizing the other features helped us easily grasp where the averages for each distribution lay. We then grew curious as to why some of the features were skewed. In particular, we were intrigued by the variables for out-of-state tuition, room and board, personal spending, and books. Since these are the variables we are most familiar with throughout our college experience.

Moving from the feature distribution we needed a way to see the correlation of many different variables all at once. The correlation heatmap is a fantastic tool to highlight both positive and negative correlations. Reading the heatmap can be very straightforward. Each cell represents the correlation coefficient between two variables. Values range from -1, a perfect

negative correlation to 1, a perfect positive correlation. The color saturation indicates the strength and direction of the correlation. Red represents positive correlations and blue indicates a negative correlation.

Figure 2

Correlation matrix of selected columns - Heatmap



Note. Created by Joe Conrad using Python's Seaborn and Matplotlib libraries.

As seen in the figure above, the variables 'Room.Board' (representing the cost of room and board across universities) and 'Outstate' (representing the cost of out-of-state tuition across universities) have a moderately strong positive correlation of 0.65. This means that as the cost of out-of-state tuition rises, so does the cost of room and board and vice versa. The variable 'Personal' (the average amount of personal spending for students, including items like: transportation, clothing, and other miscellaneous costs) has a weak negative or near zero correlation with most other variables. We can interpret this as either the amount of personal spending will not affect the other variables. In some cases, as personal spending increases other variables will slightly decrease. The variable 'Books' (representing the average cost of books or

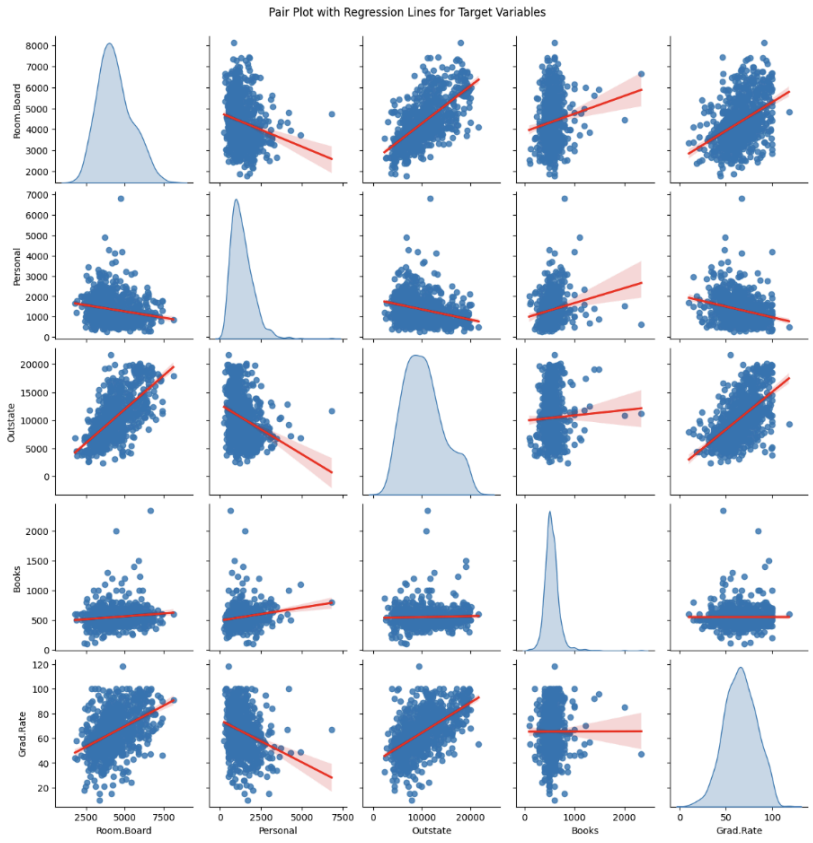
course materials for students during a year) exhibits a very weak correlation with the other variables. This indicates that the cost of books is not statistically significant in predicting other factors.

Considering our research question, we are most interested in evaluating the 'Grad.Rate' variable. We can interpret from the figure that 'Grad.Rate' has a positive correlation of 0.42 with 'Room.Board'. Meaning as the cost of room and board increases, so does the percentage of graduation rates. 'Grad.Rate' also has a moderate positive correlation of 0.57 with 'Outstate'. Meaning as the cost of out-of-state tuition increases, so does the percentage of graduation rates. 'Grad.Rate' also has a slight negative correlation of -0.27 with personal spending. This can be interpreted to show that as personal spending increases, graduation rates decrease.

After analyzing our heatmap, we were excited to see what a pair plot with regression lines would look like for our dataset. A scatterplot will show how one variable relates to another, and the distribution of the data in that variable (which will display any outliers). Adding in regression lines will reveal whether a relationship between the variables is linear, nonlinear, or non-existent. A positive slope indicates a positive relationship, a negative slope indicates a negative relationship and a flat line will indicate no linear relationship. The steepness of the slope will indicate the strength of the positive or negative correlation.

Figure 3

Pair Plot with Regression Lines for Target Variables



Note. Created by Peyton Sharpe using Python's Seaborn and Matplotlib libraries.

We can see in the figure above, 'Books' and 'Personal" have distributions centered around lower values, suggesting limited variability. 'Room.Board' and 'Outstate' have a strong positive linear relationship. This suggests that higher room and board costs often correlate with higher out-of-state tuition and vice versa. 'Grad.Rate' has a weaker relationship with most variables, but shows a slight positive correlation with 'Room.Board' and 'Outstate'. This implies again, that when out-of-state tuition goes up, so does the cost of room and board. I was also interested in how 'Grad.Rate' has a slightly negative correlation with 'Personal'. This implies that the more personal spending a student does, the lower their chances of graduation.

Conclusion

Our project strived to uncover how affordability metrics shape graduation rates at public and private colleges across the United States. Through a combination of exploratory data analysis and predictive modeling, we identified several key findings by turning our data into quantifiable data. Most notably, we observed that increases in room and board costs were associated with a slight increase in graduation rates, while higher out-of-state tuition and book costs were linked to lower graduation rates. Initially, we couldn't confidently conclude that affordability affects graduation rates, as the observed trends could have been due to chance. So to ensure the accuracy of our findings, we conducted further statistical tests and refined our models to account for potential confounding variables.

Our findings reveal that affordability metrics do have some influence on graduation rates (See Figures 1 & 2). One reason for the positive correlation between room and board costs and graduation rates can be explained by the notion that higher costs may correlate with greater student resources, which in turn could foster better student outcomes, such as higher graduation rates displayed in our findings. On the other hand, the negative correlation between out-of-state tuition and graduation rates suggests that affordability plays a crucial role in determining whether students can complete their education.

However, while these correlations exist, it is important to consider the relatively low R^2 value of our predictive models. With an R^2 value of 0.413, only 41.3% of the variance in graduation rate can be explained by affordability. This indicates that other factors, beyond affordability, may be contributing to graduation outcomes. To increase graduation rates, addressing financial barriers alone may not be sufficient. We need to consider broader strategies, such as improving accessibility to financial aid, which are also necessary to ensure students can succeed.

Future research could look into other factors that affect graduation rates, starting with how effective student support services are. Measuring their impact might reveal insights we hadn't considered before. It would also be helpful to include data from a broader range of students, as this could show us how different backgrounds and experiences play a role in graduation rates. By including more factors in the analysis, we can get a better understanding of what truly influences graduation rates.

References

Andreae, G. (2024, May 3). *Navigating higher Ed's rising costs: Strategies for affordability*. American Council of Trustees and Alumni. https://www.goacta.org/2024/04/navigating-higher-eds-rising-costs-strategies-for-affordability/

Bareham, H. (2023, July 13). *College graduation statistics*. Bankrate. https://www.bankrate.com/loans/student-loans/college-graduation-statistics/

Bauer, L., Kelchen, R., Mary Helen Immordino-Yang, D. R. K., & Jon Valant, K. M. (2022, March 9). *To boost college graduation rates, look for the successful "positive deviants."* Brookings. https://www.brookings.edu/articles/to-boost-college-graduation-rates-look-for-the-successful-positive-deviants

Bell, L. (2023, August 16). *College affordability still out of reach for students with lowest incomes, students of color*. IHEP. https://www.ihep.org/college-affordability-still-out-of-reach-for-students-with-lowest-incomes-students-of-color/

College affordability and transparency. U.S. Department of Education. (2024). https://www.ed.gov/higher-education/paying-college/college-affordability-and-transparency

Conrad, J. (2024). CollegeData12: A collection of code and data for our research. GitHub. https://github.com/JoeConrad11/CollegeData12

Data Home: College Scorecard. Data Home | College Scorecard. (2024). https://collegescorecard.ed.gov/data/

Fitzgerald, S. R., Turner, C., & Graham, A. (n.d.). *The Faculty Role in College Affordability: Syllabus Creation and Resource Affordability*. College & Research Libraries. https://crl.acrl.org/index.php/crl/article/view/26415

Hood, J. (1988). *Why college costs are rising*. Foundation for Economic Education. https://fee.org/articles/why-college-costs-are-rising/?gad_source=1&gclid=CjwKCAiA3ZC6BhBaEiwAeqfvynBguYkEtskUDQgiag6KLaQaI92BofkWYxMsmmI-CFHjRdkY_Z03MhoCPXAQAvD_BwE

Mary Helen Immordino-Yang, D. R. K., Jon Valant, K. M., & David F. Damore, C. J. S. (2023, November 8). *How much should college cost students?*. Brookings. https://www.brookings.edu/articles/how-much-should-college-cost-students/

Taylor, Z., & Alsmadi, I. (2020). *College affordability and U.S. News & World Report Rankings: : Analyzing national and regional differences*. Journal of Interdisciplinary Studies in Education. https://www.ojed.org/index.php/jise/article/view/1360

Trends in higher education outcomes. (2024). https://www.ncsl.org/education/trends-in-higher-education-understanding-policy-and-outcomes

Yahoo! (2023). *What college graduation statistics tell us about higher education*. Yahoo! Finance. https://finance.yahoo.com/news/college-graduation-statistics-035013402.html