**Anna Van Roekel Final Project Report**

**College ROI Analysis**

**1. Introduction**

Over the past two decades, the cost of college has increased significantly across the United States. According to the Education Data Initiative[[1]](#footnote-1), the average tuition and fees at public four-year colleges have risen from approximately $4,870 in 2000 to over $10,940 in 2023. As tuition rises, many students and families are asking whether paying more for college actually leads to better financial outcomes.

This project investigates whether there is a measurable return on investment (ROI) for college tuition by exploring the relationship between institutional costs and graduate salaries. In particular, I am interested in whether students who attend higher-cost public universities tend to earn more over time, and whether other factors—such as graduation rate or geographic location—also help explain salary outcomes.

To conduct this analysis, I collected and combined data from two separate sources: a scraped dataset from CollegeTuitionCompare.com containing tuition and institutional characteristics, and an earnings dataset retrieved from the U.S. Department of Education’s College Scorecard API. After cleaning and merging the datasets using fuzzy string matching on school names, I created a final dataset containing over 100 matched public four-year colleges.

The following three research questions guide my analysis:

1. Does higher tuition correlate with higher graduate salaries?
2. Does graduation rate predict salary outcomes?
3. Are there regional differences in earnings among public four-year colleges?

This project uses descriptive statistics, visualizations, hypothesis testing, and machine learning models to explore how well tuition and institutional characteristics can predict long-term salary outcomes for college graduates.

**2. Data**

*2.1 Tuition Data*

I collected tuition data from CollegeTuitionCompare.com[[2]](#footnote-2), a public website that provides structured cost data for U.S. colleges and universities. Using Python and Selenium, I developed a web scraping script to automate the collection process across hundreds of school pages. The scraped data included fields such as in-state and out-of-state tuition, acceptance rate, standardized test scores (SAT/ACT), graduation rate, and total student enrollment. In total, the script retrieved data for 461 public four-year colleges from across the United States, current as of the 2023–2024 academic year.

The web scraper saved the results into a CSV file that could be later read into Python and cleaned for merging. I removed extraneous symbols such as dollar signs, commas, and percent signs, and converted applicable fields to numeric format.

*2.2 Earnings Data*

The salary data was collected from the U.S. Department of Education’s College Scorecard API[[3]](#footnote-3). Using Python’s requests library and an API key from api.data.gov, I extracted data on median earnings 10 years after entry for graduates of U.S. colleges. This metric is based on tax record data and reflects long-term earning potential rather than immediate post-graduation salaries. Other relevant fields, such as school name and state, were also collected to assist with merging.

After removing null values and converting monetary columns to numeric types, I saved the API response data into a CSV file. This provided a structured view of graduate income outcomes across a large number of institutions.

*2.3 Data Cleaning and Integration*

Once the two datasets were cleaned, I merged them into a single data frame. Because school names often appeared differently across the two sources (e.g., “University of Iowa” vs. “The University of Iowa”), I used fuzzy string matching to identify and link records between the datasets. A similarity threshold was applied to ensure reliable matches.

Despite collecting tuition data on 461 schools, only just over 100 colleges successfully matched with corresponding salary records and contained complete information on tuition, graduation rate, and earnings. In cases where overlapping variables existed (e.g., tuition), I retained the values from the CollegeTuitionCompare.com dataset for consistency.

This integration produced a clean, horizontally merged dataset suitable for descriptive and predictive analysis. Although SAT and ACT scores were included in the final dataset, they were missing for many institutions and were not used in the analysis. Table 1 provides descriptions of the variables included in the final merged dataset.

*Table 1 Data Dictionary*

|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Type** | **Source** | **Description** |
| School Name | Text | Both | Official name of the college or university |
| State | Text | College Scorecard | U.S. state where the institution is located |
| In-State Tuition | Numeric | College Tuition Compare | Annual tuition for in-state students (USD) |
| Out-of-State Tuition | Numeric | College Tuition Compare | Annual tuition for out-of-state students (USD) |
| Median Earnings (10 yrs) | Numeric | College Scorecard | Median salary 10 years after entry (USD) |
| Acceptance Rate | Numeric | College Tuition Compare | Percentage of applicants admitted (stored as whole number, e.g., 74.9) |
| Student Population | Numeric | College Tuition Compare | Total number of enrolled students |
| Graduation Rate | Numeric | College Tuition Compare | Percentage of students who graduate (stored as whole number, e.g., 47.0) |
| SAT Score | Numeric | College Tuition Compare | Median SAT score (if reported) |
| ACT Score | Numeric | College Tuition Compare | Median ACT score (if reported) |

**3. Analysis**

*3.1 Does Higher Tuition Correlate with Higher Graduate Salaries?*

To investigate whether higher tuition leads to higher post-graduation earnings, I analyzed the relationship between tuition costs and median earnings 10 years after college entry. Specifically, I looked at both in-state and out-of-state tuition compared to reported graduate salaries.

I began with descriptive statistics and a correlation analysis to measure the strength of the relationship between tuition and earnings. The correlation coefficient between in-state tuition and median earnings was approximately 0.71, suggesting a moderate positive correlation. This indicates that, on average, students who attend higher-cost institutions tend to earn higher salaries a decade later. A similar correlation was observed for out-of-state tuition.

To test for statistical significance, I performed a two-sample t-test comparing the median earnings of colleges with above-average tuition to those with below-average tuition. The results yielded a p-value < 0.01, indicating that the difference in earnings between high-tuition and low-tuition schools is statistically significant.

I also fit a simple linear regression model, using in-state tuition as the independent variable and graduate salary as the dependent variable. The model returned an R² value of approximately 0.50, meaning that about 50% of the variation in graduate earnings can be explained by differences in tuition alone.

These results suggest that while tuition is not a perfect predictor of post-college salary, it is moderately associated with higher earnings. However, tuition likely serves as a proxy for other factors, such as institutional quality or available resources.

A graph of a line with blue dots

AI-generated content may be incorrect.

Figure 1: Scatterplot of In-State Tuition vs. Median Earnings

*A moderate positive correlation is visible between in-state tuition and graduate salaries 10 years after entry.*

*3.2 Does Graduation Rate Predict Salary Outcomes?*

Another factor that may influence graduate earnings is graduation rate. A higher graduation rate could signal better institutional support, student preparedness, or program effectiveness, all of which could contribute to post-college financial success.

To evaluate this relationship, I began by calculating the correlation coefficient between graduation rate and median earnings. The result showed a strong positive correlation of approximately 0.78, suggesting that institutions with higher graduation rates tend to have graduates with higher earnings.

I also performed a two-sample t-test, comparing graduate earnings between schools with graduation rates above and below the median. The test returned a p-value < 0.001, confirming a statistically significant difference in earnings between high-graduation and low-graduation institutions.

Instead of analyzing the raw graduation rate as a continuous variable, I grouped schools into four graduation rate ranges: 0–25%, 26–50%, 51–75%, and 76–100%. I then calculated the average median earnings within each group to better visualize how salary outcomes vary by institutional completion rates. Each bar represents the average salary of graduates from institutions falling within a specific graduation rate range.

These findings indicate that graduation rate is not only strongly correlated with earnings but may be a more reliable signal of long-term financial outcomes than tuition price.

A graph of increasing columns

AI-generated content may be incorrect.

Figure 2: Average Median Earnings by Graduation Rate Group

*Schools with higher graduation rates tend to produce graduates with higher average salaries.*

*3.3 Are There Regional Differences in Graduate Earnings?*

Geographic location is a major consideration for students choosing a college, and it may also influence post-graduation earnings. This analysis explores whether median graduate salaries vary meaningfully by state across public four-year institutions.

To ensure reliability, I first filtered the dataset to include only states with data from at least three colleges. I then calculated the average median earnings 10 years after entry for each state and visualized the results using a choropleth map.

The map reveals noticeable regional differences in graduate earnings. States like California, Colorado, and Washington stand out with higher average salaries, while others—particularly in the South and parts of the Midwest—tend to show lower earnings. This may reflect the strength of local job markets, cost of living, or the presence of high-paying industries and top-ranked institutions in certain states.

While this analysis is correlational and not causal, it suggests that geography plays a meaningful role in shaping financial outcomes for graduates, even among public colleges.

A map of the united states

AI-generated content may be incorrect.

Figure 3: Average Graduate Earnings by State (3+ Colleges Only)

*States such as California, Colorado, and Washington show higher average earnings, suggesting regional differences in graduate outcomes.*

**4. Machine Learning**

To further explore the predictability of graduate earnings, I built a set of regression models using machine learning techniques. The goal was to evaluate how well a combination of institutional characteristics could predict median graduate salary 10 years after enrollment.

The features used in the model were:

* In-State Tuition
* Graduation Rate
* Student Population
* Acceptance Rate

These predictors were selected based on earlier correlation analysis and relevance to the concept of institutional “return on investment.”

The dataset was split into training and testing sets using an 80/20 ratio. I trained and evaluated three models:

* Linear Regression (baseline)
* Ridge Regression (L2 regularization)
* Lasso Regression (L1 regularization)

All three models produced comparable results, with R² scores around 0.78, indicating that approximately 78% of the variation in graduate earnings can be explained by the selected institutional features. The similarity in performance suggests that regularization had minimal effect on model accuracy, likely due to the small dataset size and low multicollinearity among predictors.

These results support the finding that factors such as tuition cost and graduation rate are meaningfully predictive of graduate salary outcomes.

A graph showing a comparison of a model

AI-generated content may be incorrect.Figure 4: R² Score Comparison for Regression Models

*All three regression models achieved similar predictive performance, explaining approximately 78% of the variation in graduate earnings.*

**5. Conclusion**

In this project, I analyzed three aspects of college ROI: tuition, graduation rate, and geographic location. Based on the analysis, I found the following results:

1. *Does higher tuition correlate with higher graduate salaries?*  
There is a moderate positive correlation between tuition and graduate earnings. Schools with above-average tuition tend to produce graduates with higher median salaries. However, tuition alone does not fully explain salary outcomes, suggesting other factors also play a role.

2. *Does graduation rate predict salary outcomes?*  
Yes. Graduation rate was more strongly correlated with salary than tuition. Schools with higher graduation rates showed consistently higher median earnings. This relationship was statistically significant and supported by both grouped and continuous analysis.

3. *Are there regional differences in earnings among public four-year colleges?*  
Yes. There were clear state-level differences in earnings, with states like California, Colorado, and Washington reporting the highest average graduate salaries. These variations may reflect differences in local labor markets and the presence of top-ranked public institutions.

4. *Can institutional factors be used to predict graduate earnings through machine learning?*  
Yes. A multiple regression model using tuition, graduation rate, student population, and acceptance rate explained about 78% of the variation in graduate salaries. Regularization methods (Ridge and Lasso) performed similarly, indicating strong signal from the selected features and minimal overfitting.

This project shows that publicly available college data can be used to evaluate the potential financial return of attending a particular institution. The findings suggest that graduation rate and location are especially meaningful indicators of post-graduation salary outcomes.

1. <https://educationdata.org/average-cost-of-college> [↑](#footnote-ref-1)
2. [https://www.collegetuitioncompare.com](https://www.collegetuitioncompare.com/college-search-by-tuition/?range=%7bprice%7d&type=public&level=4-year-colleges&page=%7bpage%7d) [↑](#footnote-ref-2)
3. [https://collegescorecard.ed.gov](https://collegescorecard.ed.gov/data) [↑](#footnote-ref-3)