Project Report

Exploring Factors Influencing College Rankings

Kelvin Wu | Luke Miller | May 8, 2025

INTRODUCTION

In a society like ours where college education is increasingly deemed as the pathway to personal and professional success, college rankings have become an essential tool for students, parents, educational institutions, and policymakers. The rankings influence decision-makings on all levels – from student applications to government funding, and from institutional marketing strategies to educational strategies. However, different rankings systems use a wide array of methodologies and weightings, which can lead to varying results for the same college. This inconsistency raises an important question: **What underlying characteristics do high-ranking colleges tend to share, regardless of the ranking source?** In other words, we are seeking to uncover the consistent parameters – across different rankings sources – that point to institutional prestige.

In this project, we aim to explore the relationship between specific institutional features and college rankings by integrating two well-known college rankings sources: the QS College Rankings[1] and College Raptor[2]'s rankings. By merging these datasets and conducting exploratory analysis, statistical testing, and machine learning modeling, we seek to uncover patterns and correlations that define high-ranking colleges. In particular, we are interested in assessing the impact of academic reputation, internationalization (e.g., the international faculty and students), sustainability efforts, employer reputation, and other institutional metrics.

We further enhance our investigation by implementing both supervised and unsupervised machine learning methods – linear regression and K-means clustering – to quantify and visualize patterns in the data. Ultimately, this project demonstrates the value of data wrangling, integration, and analysis in understanding complex and societally relevant phenomena such as college rankings.

---

[1] This is a downloadable dataset on Kaggle, found under the following URL:
https://www.kaggle.com/datasets/darrylljk/worlds-best-universities-qs-rankings-2025
Because our project focuses on the colleges in the United States due to our research parameter, we have filtered out the foreign universities – which undeniably also glow in their prestige.

[2] This is a scrapable dataset found under the following URL:
https://www.collegeraptor.com/college-rankings/details/SchoolRanking

DATA

Our project makes use of two comprehensive real-world datasets: the QS College Rankings and College Raptor's institutional rankings. These datasets differ in structure, scope and the metrics they prioritize.

QS U.S. College Rankings

The QS dataset contains a variety of performance indicators. Our research has focused on the following metrics found in this dataset:

Features	Description
Institution Name	Name of the university
Size	Size category of the institution
Academic Reputation	Score of the institution's academic standing
Employer Reputation	Score reflecting how the employers perceive the university's graduates
Faculty Student	Ratio: academic staff to students, serving as a proxy to teaching quality
Citations per Faculty	Measure of research impact and quality, indicating the number of citations received per faculty member
International Faculty	Percentage of international faculty members, reflecting diversity in academic staff
International Students	Percentage of international students enrolled at the institution, indicating diversity in the student body
International Research Network	Score reflecting the extent of international research collaborations
Employment Outcomes	Score indicating the employability and career prospects of graduates
Sustainability	Score evaluating the institution's commitment to sustainability practices and research, including its impact on global

The variables represent both quantitative and qualitative assessments of institutional quality. Due to the prepared nature of the original dataset, we found little need to carry out substantial data cleaning procedures. Notably, we have removed several features from the dataset that we deemed not directly relevant to our evaluation of university excellence – such as location, rankings from prior years, etc. – as well as QS Overall Score in particular: as this "overall score" is in the same dataset as the foregoing

features, it is likely highly correlated with them. We extract overall assessments of university quality from our second, scraped dataset from College Raptor.

College Raptor Rankings

The College Raptor dataset provided the following parameters:

Features	Description
School Name	The name of the university
Rank	Current ranking from 2025
Prior Rank	Ranking from the prior year
Scaled Score	Scaled score reflecting the overall assessment of the university in the current year of 2025
Prior Scaled Score	Scaled score reflecting the overall assessment of the university in the prior year

Although simpler in structure, this dataset required data cleaning due to formatting issues such as newline characters (e.g., "\n(1.2%)") embedded within those score fields. We extracted the numerical portion of the Scaled Score using string manipulation and converted it to float for analysis. In addition, we dropped the rows where missing values exist in any of the listed features.

Merging Datasets

Merging these datasets required resolving inconsistencies in institutional naming conventions across the two different platforms. For example, "University of California, Berkeley (UCB)" in QS and "University of California-Berkeley" in Raptor refer to the same institution but use different styles. Indeed, people call the same university many different things colloquially – e.g., University of Iowa, U Iowa, UI, Iowa, etc.… To address the issue of naming inconsistency, we proceeded with the following steps:

1. Lowercasing all names.
2. Removing punctuation and extra descriptors (e.g., text in parentheses).
3. Applying fuzzy string matching (with a similarity threshold of 90%).

After these steps, we successfully merged 145 colleges – a significant improvement from the 93 that initially matched using only simple lowercasing.

ANALYSIS

Correlation Matrix Overview

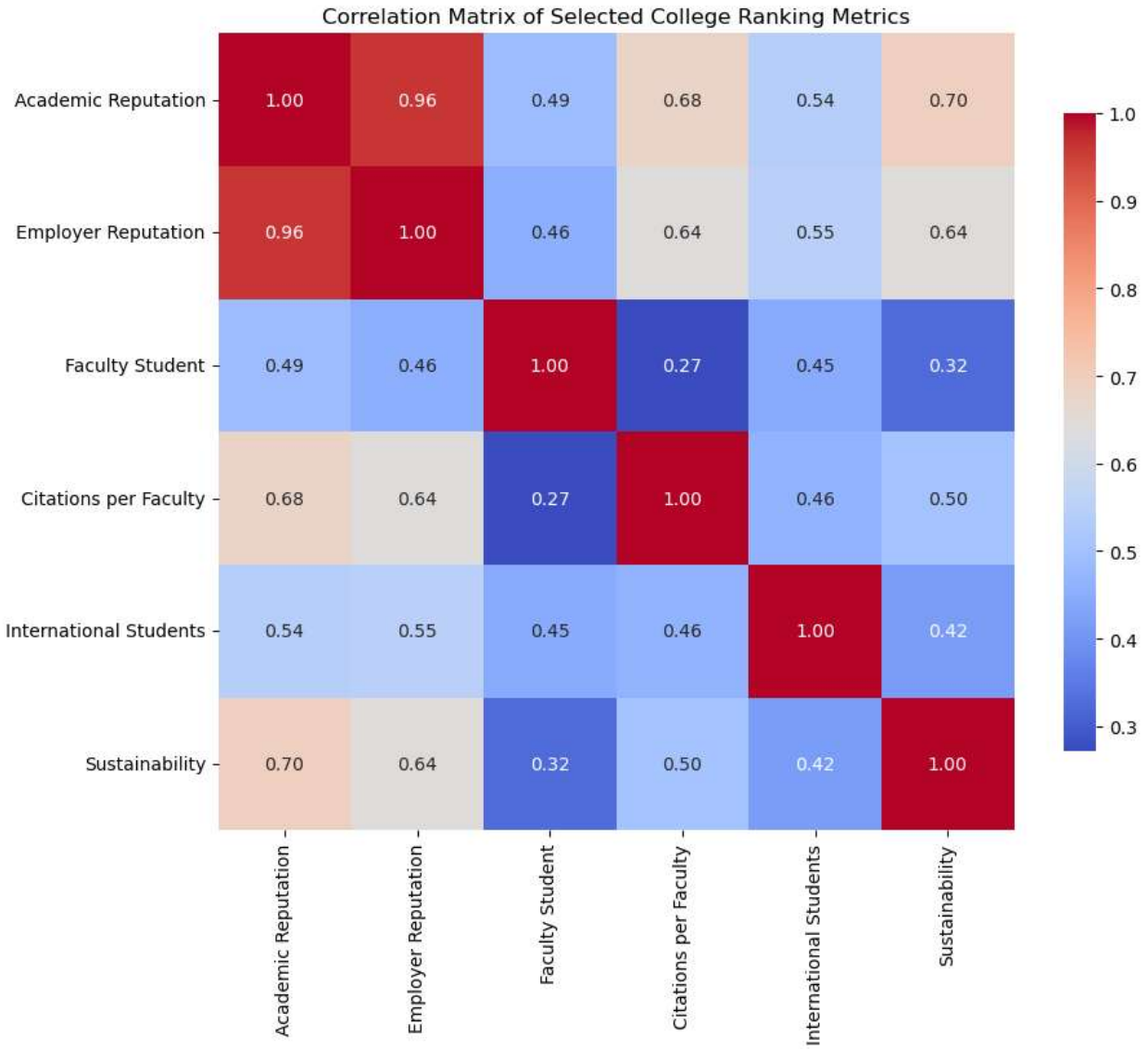

Figure 1: Correlation matrix of selected college ranking metrics from the QS dataset. Values closer to 1 indicate stronger positive relationships between features.

To provide a comprehensive view of how key institutional features relate to one another, we created a correlation matrix of selected QS metrics (Figure X). The matrix confirms the strongest relationships are between Academic Reputation and Employer Reputation (r = 0.96), and between Academic Reputation and Sustainability (r = 0.70). These high correlations suggest that institutions perceived as academically strong are also seen as more sustainable and well-regarded by employers.

Other notable insights include a moderate relationship between Citations per Faculty and Sustainability (r = 0.50), and between International Students and Faculty-Student Ratio (r = 0.45). These relationships offer additional context for interpreting our bivariate and regression results.

Topic 1: How Does Academic Reputation Relate to Scaled Score?

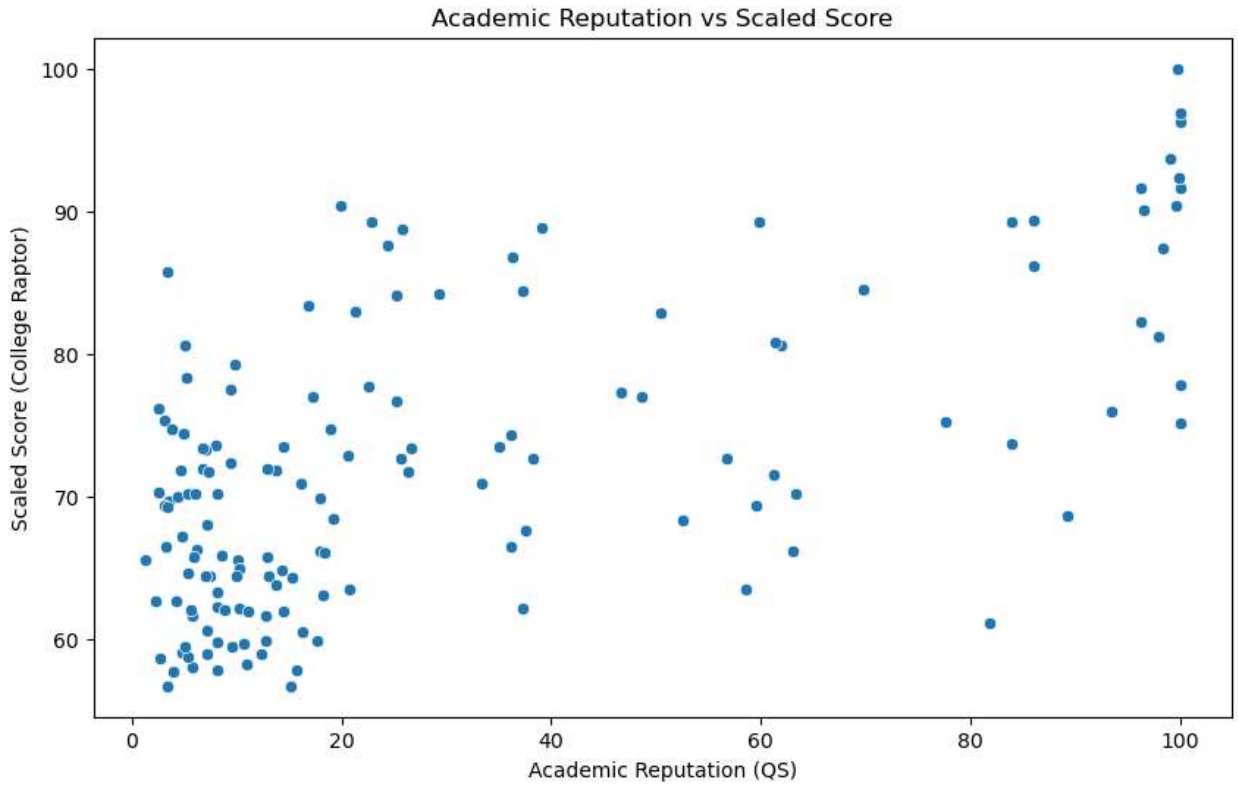

Figure 2: Scatterplot Reflecting Correlation Between Academic Reputation (QS) and Scaled Score (College Raptor)

We explored the relationship between Academic Reputation (QS) and Scaled Score (Raptor) using correlation analysis and scatterplots. We felt that these two variables are likely correlated, as academic excellence is intuitively a strong indicator of a university's prestige.

In our analysis, a strong positive correlation was observed (r = 0.62), indicating that colleges with higher academic reputation tend to receive higher composite performance scores. This relationship was visually confirmed through a clear upward trend in the scatterplot, suggesting a consistent association across the 145 merged institutions. These findings support the interpretation that perceived academic prestige, often tied to research output and faculty credentials, plays a critical role in how institutions are evaluated in ranking systems.

Topic 2: Does Sustainability Relate to Higher Rank?

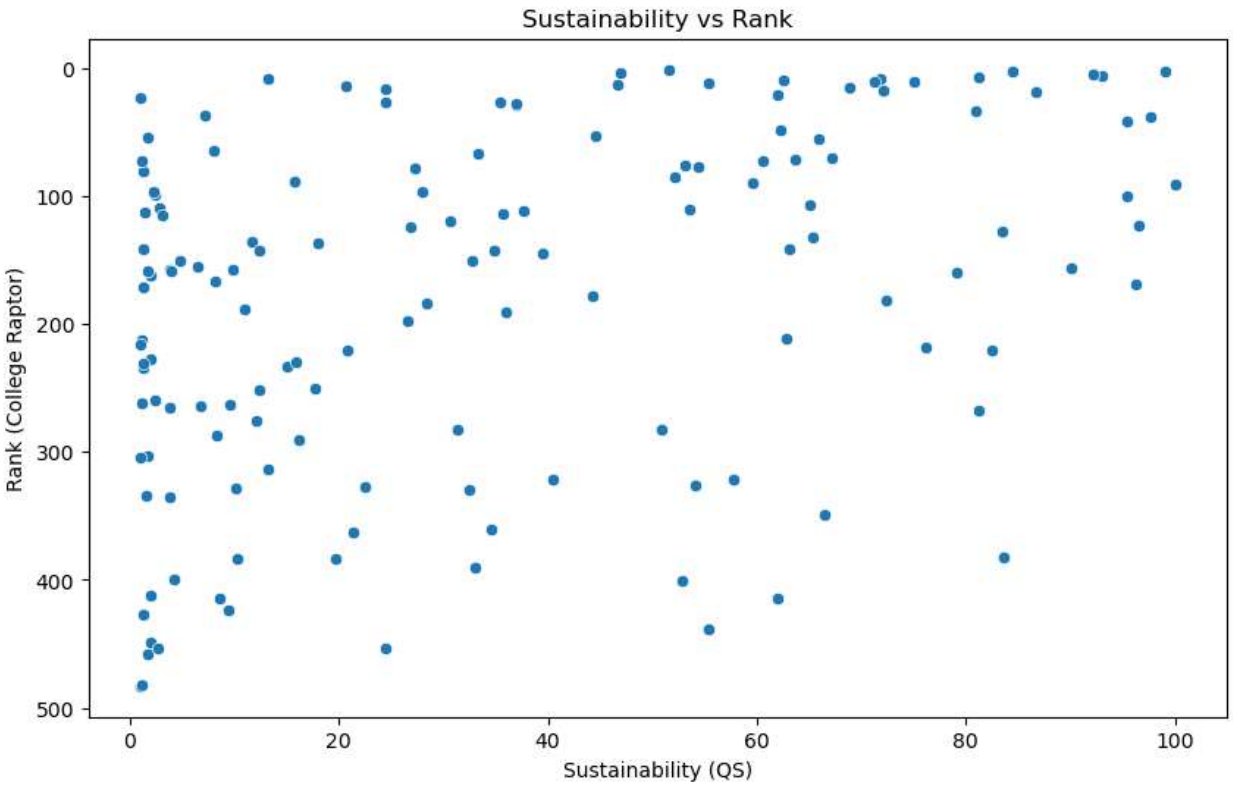

Figure 3: Scatterplot Reflecting Correlation Between Sustainability (QS) and Rank (College Raptor)

To assess whether sustainability efforts impact rankings, we compared the Sustainability Score (QS) with the Rank (Raptor). We suspected likely correlation between these two variables, as a college's investments in sustainability reflects its level of concern for the environment and the world community, a strong indicator of its humanitarian spirit.

A moderate negative correlation was found (r = -0.39), meaning that colleges with stronger sustainability performance often held better ranks (lower numbers). This finding suggests that, while sustainability may not dominate ranking criteria, it contributes to the overall perception of institutional quality. This is important as universities continue to adopt sustainability initiatives to improve campus operations, reduce environmental impact, and engage students in global challenges.

Topic 3: Is Employer Reputation Linked to Rank Improvements?

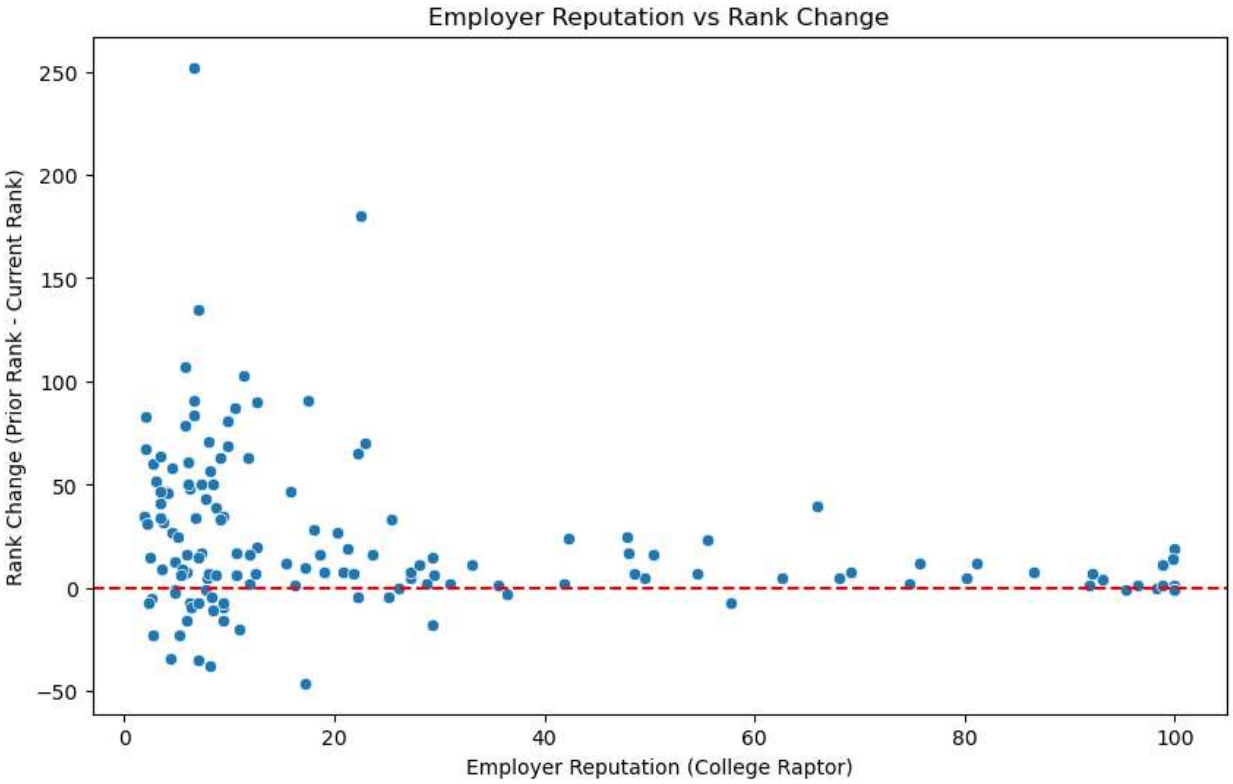

Figure 4: Scatterplot Reflecting Correlation Between Employer Reputation (College Raptor) and Rank Change (Prior Rank - Current Rank)

As the educational success of colleges is often measured by the job placements of its graduates, we intended to examine whether the employers' perceptions of a university's graduates somehow link to that university's rank change from the prior year.

We calculated the difference between a college's Prior Rank and Current Rank to derive Rank Change. The relationship between Employer Reputation and Rank Change was analyzed to assess whether stronger employer perceptions of the university's graduates correlate to improvements in rank. We found a weak but statistically significant negative correlation (r = -0.25), suggesting that colleges with higher employer reputation scores were more likely to maintain or improve their positions year-over-year. This could indicate that graduates of these institutions perform well in the job market, leading to favorable perceptions that feed back into institutional rankings.

Topic 4: Does the Presence of International Students Correlate to Scaled Scores?

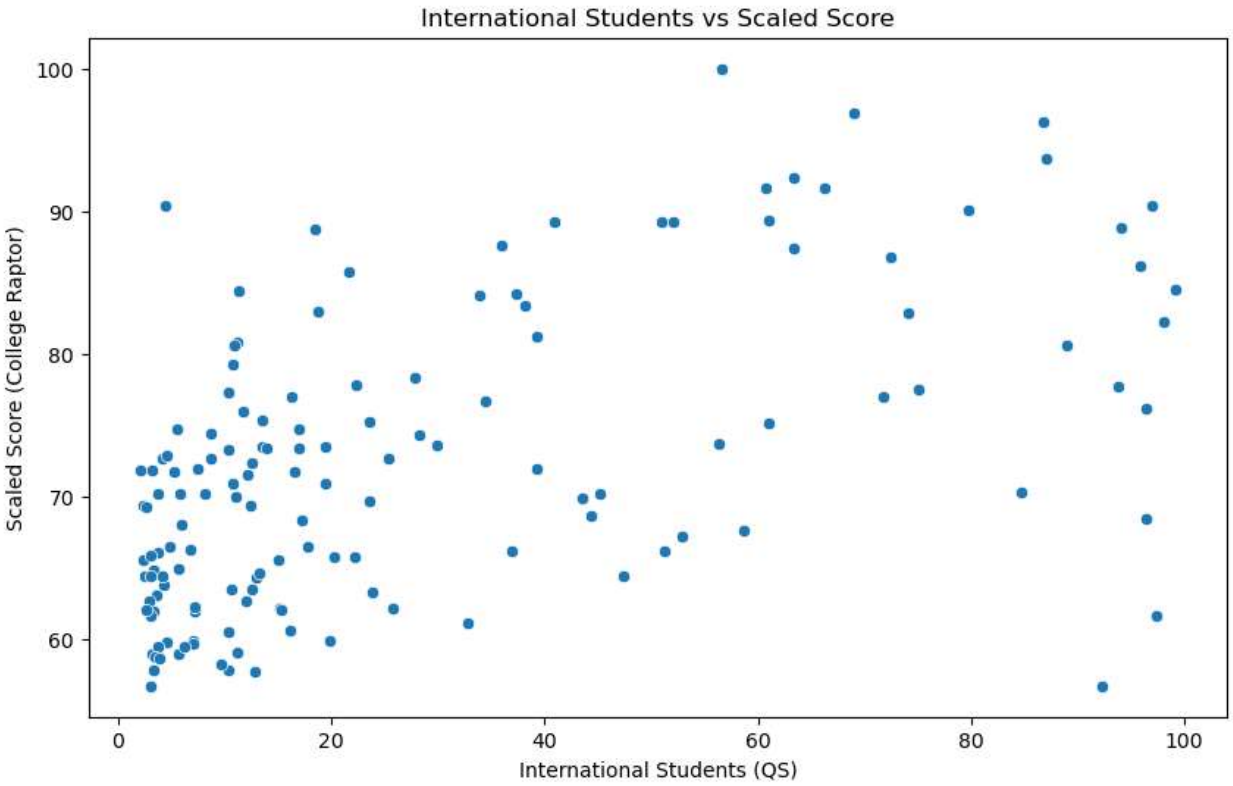

Figure 5: Scatterplot Reflecting Correlation Between International Students' Presence (QS) and Scaled Score (College Raptor)

As international students visit the United States for higher education, we found it reasonable to expect that these international students shall aim to study at the most prestigious universities. We also felt that the presence of international students contributes to a diversity of cultures and thought – recipes for richer and more dynamic intellectual environments on campus. We thus wondered if the presence of international students somehow correlates to a college's score.

We examined whether a college's percentage of international students correlates with its Scaled Score. The correlation was strong and positive ($r = 0.55$), suggesting that institutions with more globally diverse student populations tend to perform better in composite rankings. A higher international student ratio may reflect stronger global engagement, more inclusive learning environments, and the ability to attract talent from around the world – all of which align with prestige and global visibility.

Machine Learning Model 1: Predictive Modeling with Linear Regression

To identify the predictive power of selected features on institutional performance, we built a linear regression model to predict Scaled Score using Academic Reputation, Sustainability Score, and International Students %. The model achieved an R^2 of 0.56, meaning it explained 56% of the variance in Scaled Score. The mean squared error was 38.69, indicating moderate prediction error. Among predictors, Academic Reputation was the most influential (coefficient = 0.155), followed by International Students (coefficient = 0.092), with Sustainability having a minimal effect (coefficient = 0.007). This suggests that while sustainability is important, it does not yet heavily influence composite scores, unlike reputation and global outreach.

Machine Learning Model 2: Clustering Colleges Using K-Means

We applied K-Means clustering (k=3) on three features: Academic Reputation, Sustainability, and International Students. After standardization, the clustering revealed the following groups:

- **Cluster 2**: Colleges with high reputation and high international student percentages — typically elite institutions
- **Cluster 1**: Institutions with moderate performance and mixed profiles
- **Cluster 0**: Schools with lower reputation and limited international diversity — potentially regionally focused colleges

These clusters reveal underlying structure in the dataset, suggesting that certain groups of colleges share similar strengths and institutional profiles. The clusters were visualized in a scatterplot and aligned closely with known categories of elite, mid-tier, and regional institutions. This clustering provides a useful perspective for understanding how similar institutions compare in performance metrics.

CONCLUSION

This project set out to uncover what institutional characteristics most strongly influence a college's performance across multiple ranking systems. Through our analysis, we identified several consistent themes. First and foremost, academic reputation emerged as a critical driver of institutional rank and scaled score, confirmed both through correlation analysis and as the most predictive feature in our linear regression model. Institutions that are widely perceived as academically rigorous or research-intensive reliably score higher.

Next, we found that international student representation plays a meaningful role in boosting a college's scaled score. This likely reflects how global engagement, diversity, and cultural exposure contribute to the perceived prestige and influence of an institution. Our clustering analysis also highlighted that elite colleges often exhibit high internationalization in tandem with academic strength.

While sustainability was moderately correlated with improved rankings, its influence was noticeably weaker in predictive modeling. This suggests that although sustainability is becoming more relevant in public discourse and institutional mission statements, it may not yet be weighted heavily in traditional scoring methodologies.

Together, these findings paint a clearer picture of how various dimensions — academic rigor, international visibility, sustainability, and employment outcomes — interact to shape a college's position in the competitive educational landscape.

Looking ahead, we envision expanding this project to include variables such as tuition rates, student loan default rates, alumni earnings, and faculty research output. Longitudinal tracking across multiple years would allow us to determine whether these relationships are stable or evolving. In doing so, we hope to support students, educators, and policymakers with a deeper and more nuanced understanding of the forces that drive institutional success.