Project Report

World University Rankings & Post-Grad Salaries

Mallory Drake, Paige Turner, & Luke Weeklund

1. Introduction

In today's competitive global job market, higher-education institutions play a crucial role in shaping graduates' career prospects and earning potential. Prospective students often consider factors such as academic reputation, global ranking, and potential future earnings when making their college choices. However, navigating through vast amounts of data to make informed decisions can be overwhelming, especially as many students move away from home. This project aims to analyze relationships between university rankings and the factors that contribute to these rankings such as graduates' salaries, percentage of STEM majors at the university, and academic reputation by integrating two real-world data sources.

Through our analysis we hope to help prospective university students understand the reasons behind institutional rankings and how it will influence their post-graduation financial compensation. Not only is this information helpful for students, but it's also valuable for hiring managers as they review resumes and search for strong candidates.

2. Data

This project uses two primary sources of data: QS World University Rankings[1] of different institution ranking details, and PayScale's 2024[2] salary report that details average salary amounts from university graduates.

2.1 World University Rankings

The first data source we will use is the QS World University Rankings 2024 from Kaggle. It features 29 columns and 1,497 rows. This dataset contains many features that support the rank each school received and includes the size of the institution, faculty scores, academic reputation

[1] https://www.kaggle.com/datasets/joebeachcapital/qs-world-university-rankings-2024/data

[2] https://www.payscale.com/research-and-insights/college-impact/

scores, and employment outcome scores. Some *2024 Rank* values were expressed as ranges (e.g., 1201-1400). To address this, we remapped the values to match the index plus 1.

This dataset had numerous columns, some of which are not important to our intended analysis, so we decided to drop *2023 Rankings, Country Code*, *Focus, Citations per Faculty Score, Citations per Faculty Rank, International Faculty Score, International Faculty Rank, International Students Score, International Students Rank, International Research Network Score, International Research Network Rank, RES, and Status.* columns. This left the dataset with 16 remaining columns and the same 1,497 rows of data.

2.2 Average Graduate Salaries

We scraped our second data source from PayScale's 2024 College Salary Report, gathering data from 61 pages on the site. The resulting dataset includes 1,503 rows and 5 columns (excluding *% High Meaning).* It provides average salary information across two stages for graduates from various universities.

2.3 Combining Rankings & Salaries

To successfully merge these datasets, we renamed the *Institution Name* column from our first data source to *School Name* to match the second dataset. We also began cleaning the data by stripping any spacing on the *School Name* data values and removing any abbreviations in parentheses from the Kaggle dataset. This approach maximized the number of matching school names when performing our join. After correcting data types and merging our two datasets we were left with a data frame consisting of 125 rows and 21 columns.

We recognized there were still many insignificant columns, so we dropped *School Type*, *Academic Reputation Rank*, *Employer Reputation Rank*, *Faculty Student Rank*, *Employment Outcomes Rank*, *Sustainability Rank*, *Sustainability Score*.

Originally, we merged the two datasets using an inner join on '*School Name*.' We manually mapped additional schools that had similar names but different formatting to get to a final count of 182 rows and 14 columns. We will use this merged data frame to perform our analysis.

Table 1 Data Dictionary

Field	**Type**	**Description**
2024 Rank	Numeric	Overall rank of the university in 2024
School Name	Text	Name of the university
Country	Text	Country where the university is located
Size	Text	Size of the university (S = Small, M = Medium, L = Large, XL = Extra Large)
Age	Numeric	Age category of the university
Academic Reputation Score	Numeric	Score for Academic Reputation
Employer Reputation Score	Numeric	Score for Employer Reputation
Faculty Student Score	Numeric	Score for Faculty-Student Ratio
Employment Outcomes Score	Numeric	Score for Employment Outcomes
Overall Score	Numeric	Final score determining University ranking
Postgrad Salary Rank	Numeric	Rank for post-grad salary of alumni
Early Career Pay	Numeric	Median salary of alumni with 0-5 years of experience

Mid-Career Pay	Numeric	Median salary of alumni with 10+ years of experience
Percent of STEM Degrees	Numeric	% degrees awarded in science, technology, engineering and/or mathematics

3. Analysis

3.1 Overall University Rank and Mid-Career Pay

We wanted to explore the relationship between *2024 Rank* and *Mid-Career Pay*. First, constructed a Scatter Plot to visualize the relationship between *2024 Rank* and *Early-Career Pay*. We expect a moderate negative relationship.

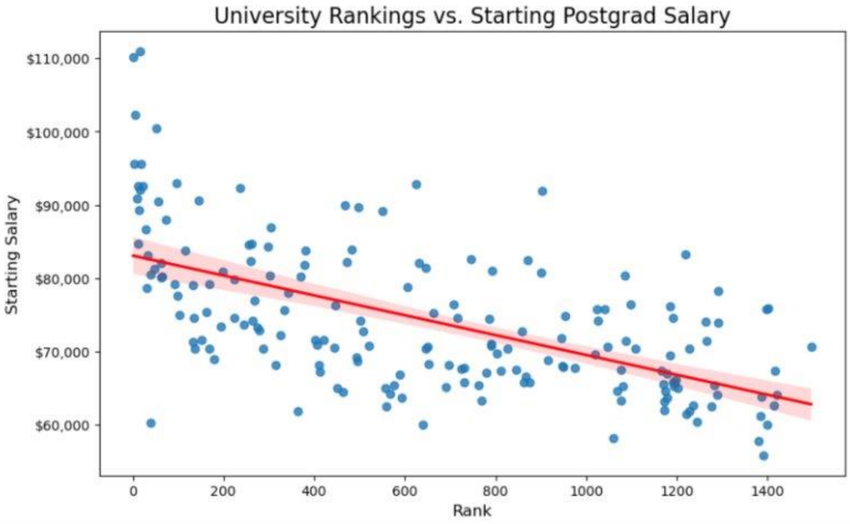

Figure 1 Scatter Plot of Rank and Starting Salary

We found a negative correlation in Figure 1, which means the worse the rank, the lower the starting salary. We wanted to verify that Early Career Pay was the better predictor compared to Mid-Career Pay, so we created a Jitter Plot to see the distribution comparisons between early and mid-career salary.

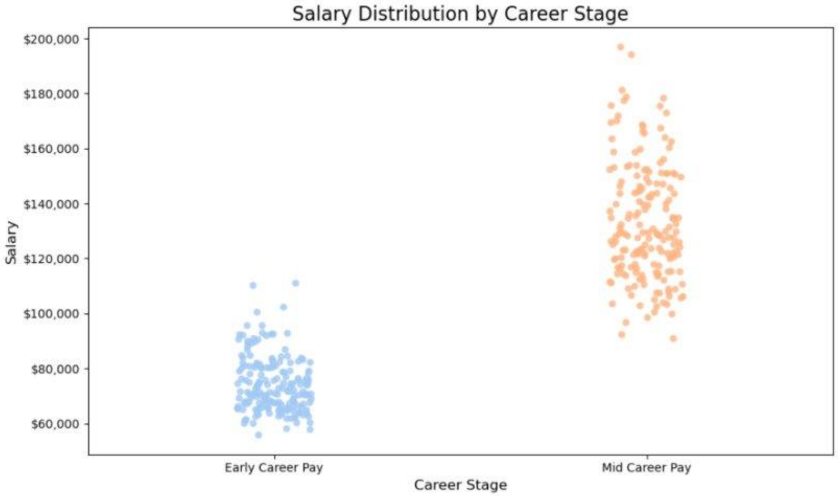

Figure 2 Jitter Plot of Salary Distribution

From the distribution in Figure 2, we found that mid-career salary has a larger distribution than early-career, indicating that other factors play a significant role in shaping later-stage earnings. This finding led us to change our target variable to mid-career salary after finding that higher-ranked universities tend to have a higher mid-career salary.

We then built a linear regression model with *2024 Rank* as the independent variable and *Mid-Career Pay* as the dependent variable. We used training and testing data to fit the model, finding an intercept of 150,988 and a coefficient of –26.60 for *2024 Rank*. Next, we found that the model had an accuracy score of 28.89%, showing *2024 Rank* is a poor predictor of *Mid-Career Pay*.

3.2 Overall Rank and Percentage of STEM Majors

We explored the relationship between a university's overall rank and the proportion of its graduates earning STEM degrees. We expect no correlation or relationship between variables. To analyze the connection between these variables, we first created a scatter plot with a trend line.

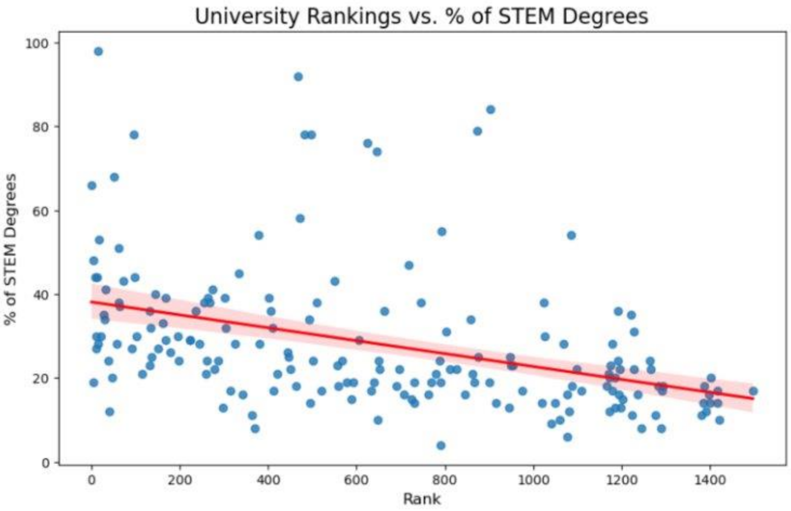

Figure 3 Scatter Plot of Rank and % STEM Degrees

As shown in Figure 3, there is a slight negative correlation. Higher-ranked universities (with lower values) tend to boast a higher percentage of STEM degrees. This correlation suggests that more prestigious institutions may attract or emphasize STEM-oriented studies. However, the trendline slope is modest, indicating that while rank may play a role, it is not a strong standalone predictor of STEM concentration. We ran a simple linear regression model between the two variables finding an R^2 score of only 16%.

To better understand the overall distribution of STEM degree prevalence, we also examined the frequency of STEM degree percentages across all universities.

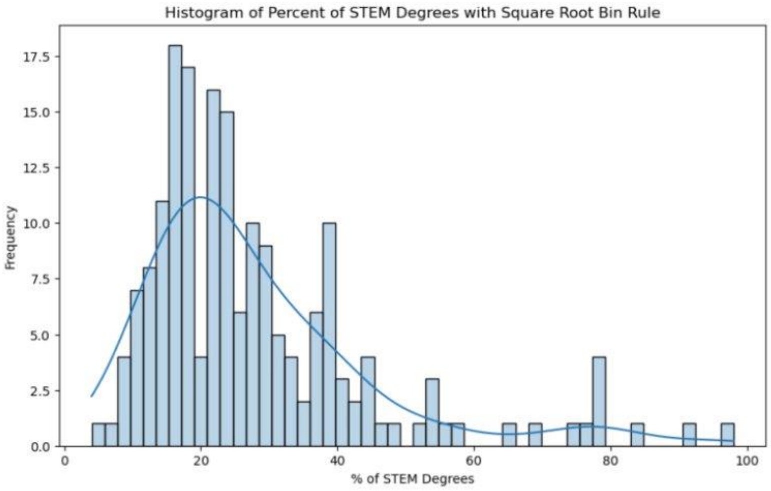

Figure 4 Frequency Histogram of % STEM Degrees

The histogram reveals that most universities award between 15% and 35% of their degrees in STEM fields, with the distribution skewed right. A small number of institutions individually award a high percentage of STEM degrees, in some cases 90–100%, likely reflecting specialized technical or engineering schools.

Together, these figures suggest that while top-ranked institutions may slightly favor STEM programs, the variation in STEM concentration is broad and influenced by institutional focus and mission. To further understand what factors influenced STEM concentration we ran three tests for two different machine learning models, linear regression and decision tree classification.

We began by running a linear regression model using only the 2024 Rank, with a 70/30 train-test split. This model yielded an R^2 of just 1%, indicating very limited predictive power. To improve performance, we ran another regression using five variables to predict the percentage of STEM degrees, which raised the R^2 to 27%. Finally, we created a model using nine variables (excluding *School Name, Country, Faculty Student Score, and Overall Score*) which resulted in an R^2 of 54%. This progression shows that increasing the number of relevant variables leads to stronger model performance.

Based on the skewed distribution found in Figure 4, we decided to group the STEM percentages into bins rather than individual datapoints. This grouping method allowed us to perform a decision tree classification model with the data. We continued testing with the same nine variables as the final linear regression model. Our initial grouping of bins was 0-20, 20-40, and 40-100, which produced an accuracy score of 61.8%. We reran this model using bins of 0-20, 20-60, and 60-100, which produced an accuracy score of 65.5%. We ran this model for a final time using bins of 0-33, 33-67, and 67-100, which produced an accuracy score of 74.5%.

By grouping STEM percentages into bins, we were able to apply decision tree classification and improve model performance. The final binning strategy yielded the highest accuracy at 74.5%, showing that thoughtful categorization can enhance predictive power.

3.3 School Size and Rank

To evaluate whether institutional size is associated with university rankings, we compared school sizes, categorized as Small, Medium, Large, and XLarge, with their *2024 Rank*. We expect, on average, larger institutions to have a better rankings. To explore each individual school size as it relates to rank, we created a collection of box plots to compare the distribution of rank within each size.

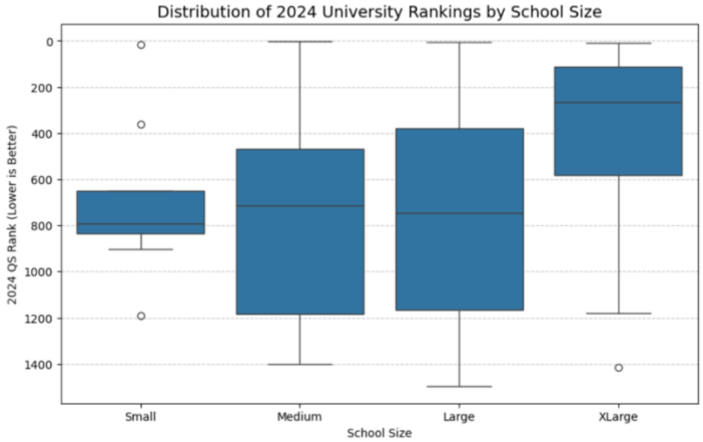

Figure 5 Distribution of 2024 University Rankings by School Size

In Figure 5, the box plots show how university rankings vary by school size, where lower values indicate better rankings. Small schools appear more tightly clustered around better ranks, while extra-large schools have a wider spread and a higher median rank. Medium and large schools show similar median rankings but also display considerable variability.

To further investigate whether size correlates with top performance, we created a bar chart that breaks down the percentage of schools in each size category that fall within the top 25% of rankings.

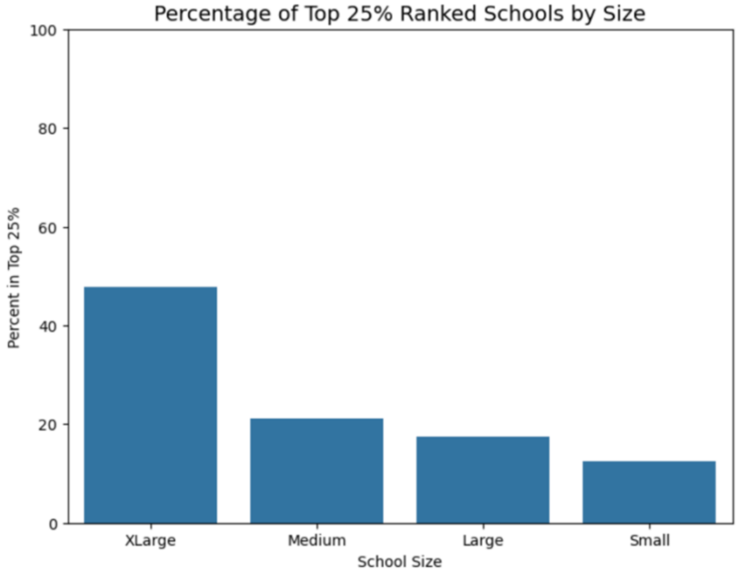

Figure 6 Percentage of Top 25% Ranked Schools by Size

Figure 6 highlights how XLarge schools dominate this metric, with nearly half of them ranking in the top quartile. In contrast, Small, Medium, and Large schools each contributed significantly smaller shares of top performers.

These findings suggest that while small schools may occasionally achieve high rankings, extra-large institutions are more consistently represented among the top-tier universities. This could be due to greater resources, research output, and global visibility often associated with larger institutions.

This led us to dive deeper into statistical hypothesis testing. We broke the *2024 Rank* data into percentiles and created a new binary variable identifying the top 25% as “Top Ranked Schools.” School sizes were categorized as Small = 1, Medium = 2, Large = 3, and XLarge = 4. We then counted how many top-ranked schools appeared in each size group.

Our first test was a two-proportion z-test comparing the proportion of top 25% ranked schools between Small and XLarge schools. The test yielded a z-statistic of -1.86 and a p-value of 0.0636. Since the p-value exceeded the 0.05 significance threshold, the result was not statistically significant. There is no evidence of a meaningful difference in the probability of a Small or XLarge school being in the top 25% of rankings.

Next, we ran the same test using a top 10% threshold while still comparing Small and XLarge schools. This test produced a z-statistic of -0.087 and a p-value of 0.931, which was also not statistically significant, now by a larger margin. Again, no significant difference was found between the groups.

We then compared Small and Large schools using a top 25% threshold. The z-statistic was -0.35 with a p-value of 0.724, indicating no significant difference in ranking performance between Small and Large schools.

Finally, we tested the difference between Small and Large schools using a broader threshold of the top 50%. The z-statistic was -1.016 and the p-value was 0.3098 – still not statistically significant.

Among tests, we found no strong evidence that school size affects the likelihood of a school being top ranked, though the first test came close.

4. Conclusion

In this project, we analyzed how different institutional factors impact post-graduate salary. In summary, from the analysis questions presented in our proposal, we found the following results.

1. *How does an institution's rank correlate with graduate salary outcomes? What implications does this relationship have for employers in terms of hiring decisions and for students when choosing educational institutions?*
   Rank is a weak predictor of mid-career pay, with an accuracy score of only 28.88%. The linear regression model used a 70/30 split and showed a coefficient of -26.65. Although higher-ranked universities often have higher salaries, the wide distribution of mid-career pay suggests rank alone isn't enough. More variables or complex models are needed for better predictions.

2. *How does the percentage of STEM majors at an institution relate to its overall ranking and what factors influence this correlation?*
   The percentage of STEM degrees across schools shows a right-skewed distribution, with most institutions falling between 15% and 25%, and a smaller peak near 80%. Predictive accuracy improved significantly when more variables were added, with the R^2 increasing from 16% to 54%. A machine learning model using a broader set of features achieved a 49.07% accuracy, highlighting the value of multivariable approaches in understanding STEM degree distribution.

3. *Do small schools have a higher probability of having a higher ranking than super large schools?*
   Two-proportion Z-tests across various ranking thresholds (10%, 25%, and 50%) found no statistically significant relationship between school size and the likelihood of being top-ranked. The closest result, comparing small and XL schools at the 25% threshold, produced a z-statistic of -1.86 and a p-value of 0.0636, which was not below the 0.05

significance level. Overall, school size does not appear to be a meaningful predictor of ranking performance.

This project has several limitations, including variable duplication, data scarcity issues, and a focus on American institutions rather than international. Despite these constraints, we found that institutional rank is a weak predictor of mid-career salary, school size does not significantly influence top rankings, and the percentage of STEM majors is more informative when combined with additional factors. Future work could involve expanding the dataset, incorporating international data, and applying advanced modeling techniques to better understand what drives higher post-graduate salaries.