CSE 519 FINAL PROJECT REPORT
Education Statistics/Rankings

January 5, 2024

1 Introduction

The quantification of research impact has been a topic of intense investigation in the evolving landscape of academia. Citation metrics, a key component of this endeavor, traditionally assess the impact of academic papers and articles by the frequency with which they are cited in other scholarly works[9]. This seemingly simple metric, however, encounters difficulties, such as the lack of oversight, the size of the research community, and the overall coverage of a paper.

Efforts to establish a universal citation index have seen historical milestones, most notably the introduction of Eugene Garfield's Science Citation Index (SCI) and Social Science Citation Index, ensuring quality regardless of their contributions. Modern research recognizes the limitations of traditional metrics and seeks novel approaches to quantifying the cumulative impact and ranking the relevance of scientific research.

In this context, our research project sets out to improve understanding of university rankings by delving into the complex web of variables that contribute to disparities between different ranking systems[1]. Our global analysis includes information from over 800 universities and is based on data from reputable sources such as WalletHub, Forbes, Rate My Professors, and Steven Skiena's dataset.

Our primary goal is to conduct a comparative ranking analysis of publicly available university rankings, revealing significant differences due to diverse target audiences, priorities, data sources, and methodologies. Notably, certain rankings' emphasis on research quality over student career achievements adds layers of complexity to the evaluation process[5].

To navigate this complication, we have carefully chosen indicators that are aligned with the multifaceted nature of university performance. Essential characteristics are captured by core indicators such as Student Selectivity Rank, Faculty Resources Rank, Education Outcome Rank, and Career Outcome Rank. Complementary measures, such as the novel Galloway Score, Completion Rate, and Sports Score, offer a more nuanced picture of university performance.

The Location Score, an innovative addition to our methodology, considers the subtle influences of a university's location on its appeal. Recognizing the increased attractiveness of universities located in major cities, this novel inclusion adds to our research.

By integrating datasets from various sources and refining our ranking techniques, we hope to contribute to the ongoing debate on university rankings. This effort is critical for all stakeholders, from prospective students looking for information to legislators shaping education policies. Our project aims to capture the full significance of a university's role within the academic community by providing insights beyond traditional measurements.

As we move forward, we will investigate the complexities inherent in the world of university rankings, providing a thorough analysis of our findings and methodologies. Through

this endeavor, we hope to add to the existing body of knowledge and provide valuable insights that go beyond traditional evaluations of academic institutions.

2 Datasets

Our research project relies on a meticulous selection of datasets sourced from reputable platforms to gain a nuanced understanding of university rankings. Each dataset brings a unique perspective to the table, allowing us to conduct a thorough and comparative analysis of university rankings. Our methodology entails combining various datasets to capture a wide range of variables that influence a university's performance and reputation.

WalletHub Collection[6] In our investigation of institutional rankings, WalletHub emerges as a critical source of data. This platform provides a thorough examination of various variables, such as financial measurements, indicators of academic success, and student happiness polls. The richness of WalletHub's dataset provides us with valuable insights into the multifaceted dimensions that contribute to academic institution evaluation.

Forbes Collection[2] The dependable Forbes dataset is critical to our investigation of university rankings. Forbes provides detailed statistics on alumni accomplishments, reputation scores, and financial metrics. We gain a deeper understanding of the various factors that shape a university's standing by leveraging this dataset, allowing us to refine our ranking algorithm and provide a more holistic view of academic performance.

Skiena Dataset[4] The Skiena dataset is yet another valuable resource that has been incorporated into our research. This dataset adds to our understanding of university rankings by providing a wealth of data across multiple dimensions. Skiena's dataset adds to our analysis, allowing us to investigate new aspects that contribute to the complex landscape of academic institutions.

Rate My Professor[3] Including data from Rate My Professor, a popular platform where students rate their professors and courses, adds a dynamic layer to our analysis. Student satisfaction and professor performance are important considerations when assessing academic institutions. Rate My Professor data improves our ability to evaluate the qualitative aspects of universities by providing a nuanced perspective that goes beyond traditional metrics.

Approach to Holistic Analysis: By combining these disparate datasets, our project takes a comprehensive approach to university ranking analysis. This approach enables us to investigate the complex interplay of factors that influence a university's position in the academic landscape. Our research seeks to capture the full significance of a university's role within the academic community, going beyond traditional evaluations.

Data Description

Student Selectivity Rank: A ranking based on the selectivity of undergraduate students, reflecting the competitiveness of the admissions process.

Faculty Resources Rank: A ranking based on the resources allocated to faculty, providing insights into the university's commitment to supporting its teaching staff.

Education Outcome Ranks: A ranking indicating the educational outcomes, reflecting the effectiveness of the university's academic programs.

Career Outcome Ranks: A ranking reflecting the career outcomes of graduates, gauging the success of students in their post-graduation professional pursuits.

Galloway Score: A measure of future financial security proposed by Galloway, reflecting the economic stability of the university.

Completion Rate: The percentage of students who graduate within six years of matriculation, reflecting the university's success in supporting student success.

Sports Score: For NCAA Division I schools, a score representing the competitive level of sports conferences provides insights into the university's athletic prominence.

Location Score: A score indicating the desirability of the university's location, considering factors such as climate, amenities, and local attractions.

Rate my Professor Score: An aggregated score based on student ratings of professors, providing insights into the teaching quality at the university.

Final Score: An overall score calculated as the sum of various university-ranking factors, offering a comprehensive evaluation.

Final Score out of 100: The final score was scaled to a range of 1 to 100, providing a standardized measure for comparison.

Final Rank: The rank of universities based on the scaled final score, with 1 being the highest.

3 Methods

3.1 Data Collection

Our research project hinges on a diverse array of data variables meticulously sourced from esteemed platforms such as WalletHub, Forbes, Rate My Professors, and Steven Skiena's dataset. This robust dataset is designed to encapsulate a multifaceted view of university performance, encompassing an extensive range of performance indicators. Among the pivotal metrics considered are the traditional measures of Student Selectivity Rank, Faculty Resources Rank, Education Outcome Rank, and Career Outcome Rank. To deepen our analysis, we have integrated innovative factors such as the Galloway Score, Completion Rate, and Sports Score, offering a more nuanced perspective on university performance[7]. Going beyond conventional metrics, we delve into the subtleties of a university's appeal by incorporating Location Scores and Faculty Scores from Rate My Professor. This comprehensive approach ensures that our research explores the intricacies of university rankings, providing stakeholders with a thorough and well-rounded understanding of the diverse factors that contribute to academic excellence and institutional appeal.

3.2 Data Matching and Cleaning

In our pursuit of a cohesive and accurate analysis, we leverage the powerful capabilities of the fuzzywuzzy library in Python to seamlessly integrate datasets while accommodating variations in university names[8]. This sophisticated library enables the computation of string similarity scores, predominantly employing the Levenshtein distance algorithm. The foundation of our approach lies in the definition of a matching function, which harnesses the fuzz.ratio method to quantitatively measure the similarity ratio between different university names. This method ensures a nuanced assessment that goes beyond exact matches.

The subsequent creation of a mapping dictionary serves as the linchpin in establishing robust connections between university names across disparate datasets. To refine this process, a judiciously chosen similarity threshold filters out marginal matches, enhancing the precision of our mappings. The iterative nature of this methodology is underscored by a crucial manual review and update phase, addressing any discrepancies and fortifying the accuracy of the final mapping. The pivotal outcome of this meticulous process is the augmentation of datasets with standardized university names. This enhancement, in turn, facilitates a seamless and effective merging of datasets, setting the stage for a comprehensive and cohesive analytical framework that transcends challenges posed by variations in university naming conventions.

3.3 Data Imputation

In our project, we strategically employ regression imputation as a robust statistical technique to address missing values within our dataset. This method proves particularly effective when dealing with variables Y that contain incomplete data. The core of regression imputation lies in constructing a predictive model, represented by:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$

where:

- $X_1, X_2, \ldots, X_n$ are observed variables.

- $beta_0$ is the intercept.

- $beta_1, beta_2, ..., beta_n$ are regression coefficients.

- ϵ signifies the error term.

Once the model is established and coefficients are estimated from available data, it becomes a powerful tool for predicting missing values

$hatY$. The imputation process involves substituting observed values of $\beta_1, \beta_2, ..., \beta_n$ into the regression equation:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_n X_n$$

Our regression imputation procedure is systematically applied to each variable with missing values, resulting in a comprehensive imputed dataset. Notably, our approach tailors the regression imputation formula for each variable, employing separate regression models based on relevance and correlation. This customization ensures a nuanced and accurate imputation process that preserves intricate relationships between variables. This meticulous methodology enhances the overall reliability and integrity of our dataset, contributing to the robustness of our research findings.

3.4 Rate My Professor Score Calculation

In our pursuit of a comprehensive university assessment, we've formulated a scoring mechanism to distill the essence of various factors, including the number of professors, individual professor ratings, and the average quality and difficulty scores. This University Score US formula is designed to be adaptable, allowing you to tailor the weights assigned to each component based on their relative importance. The formula is expressed as follows:

$$US = \left(\frac{Total\ Rating\ Points}{Total\ number\ of\ Professors} \right) X\ Professor\ weight$$
$$+ \left(\frac{Avg\ Quality\ Score\ +\ Avg\ Difficulty\ Score}{2} \right)$$

Here's a concise explanation of the variables in the formula:

- Total Rating Points: The aggregate of ratings for all professors.

- Total Number of Professors: The overall count of professors.

- Professor Weight: A user-defined weight for the professor rating component (ranging from 0 to 1).

- Average Quality Score: The average quality score of all professors.

- Average Difficulty Score: The average difficulty score of all professors.

- Average Score Weight: A user-assigned weight for the average quality and difficulty component (ranging from 0 to 1).

```
def calculate_university_score(total_rating_points, total_professors, professor_weight,
                               average_quality_score, average_difficulty_score, average_score_weight):
    professor_component = (total_rating_points / total_professors) * professor_weight
    average_score_component = ((average_quality_score + average_difficulty_score) / 2) * average_score_weight
    university_score = professor_component + average_score_component
    return university_score
```

3.5 Location Score Calculation

In our thorough evaluation of states for university selection, we have devised a comprehensive scoring system that integrates key factors, placing emphasis on employment opportunities, the presence of international students, and student body diversity. This scoring approach aims to offer a nuanced and balanced assessment, catering to the diverse needs and aspirations of potential university attendees.

For each state, we assigned a weight of 0.5 to the employment rate, which was normalized between 0 and 1, with 1 indicating the highest employment opportunities. Simultaneously, the percentages of international students and diversity factors collectively received a weight of 0.5. This reflects our acknowledgment of the importance of cultivating a globally inclusive and diverse academic environment. Both these factors were normalized between 0 and 1, where 1 represents the highest percentage of international students and the greatest diversity among the student body. The composite score C for each state is calculated using the formula:

$$C = 0.5 \times E + 0.25 \times I + 0.25 \times D$$

where:

- E denotes the normalized employment rate.

- I represents the percentage of international students.

- D symbolizes the diversity factor.

To map the composite score C to a user-friendly 5-point scale, the following formula was employed:

$$Score\ out\ of\ 5 = 1 + 4 \times C$$

This approach ensures that the final scores fall within a standardized range of 1 to 5, where 5 indicates the most favorable state for prospective students. This holistic evaluation takes into account employment opportunities, international student presence, and student body diversity. The scoring system provides a comprehensive perspective aligned with the diverse priorities and considerations of individuals navigating the complex landscape of university selection.

```python
def calculate_composite_score(employment_rate, international_students, diversity_factor):
    normalized_employment_rate = employment_rate / max_employment_rate
    normalized_international_students = international_students / max_international_students
    normalized_diversity_factor = diversity_factor / max_diversity_factor

    composite_score = (
        0.5 * normalized_employment_rate +
        0.25 * normalized_international_students +
        0.25 * normalized_diversity_factor
    )

    return composite_score

def map_to_5_point_scale(composite_score):
    score_out_of_5 = 1 + 4 * composite_score
    score_out_of_5 = min(5, max(1, score_out_of_5))

    return score_out_of_5
```

3.6 Baseline Model for University Ranking

Total Score Calculation

We will calculate a baseline total score for each university by combining key metrics relevant to university performance. The formula for the total score (Total Score$_{\text{Baseline}}$) is as follows:

$$Total\ Score_{Baseline} = w_1 * Metric_1 + w_2 * Metric_2 \\ + ... + w_n * Metric_n$$

where:

- $Metric_1, Metric_2, ..., Metric_n$ are different performance metrics

- $w_1, w_2, \ldots, w_n$ are weights assigned to each metric, reflecting their relative importance.

Metric Weights

Assign weights to each metric based on its perceived importance. For example:

$$Selectivity = 0.2 \quad w_{\text{Selectivity}} = 0.2$$
$$Resources = 0.15 \quad w_{\text{Resources}} = 0.15$$
$$Education = 0.25 \quad w_{\text{Education}} = 0.25$$
$$Career = 0.2 \quad w_{\text{Career}} = 0.2$$
$$Additional\ Metrics = 0.2 \quad w_{\text{Addit Metrics}} = 0.2$$

### Advantages and Considerations

Advantages:

- Combines multiple metrics to provide a comprehensive ranking.

- Allows flexibility in adjusting weights based on the perceived significance of each metric.

Considerations:

- The weights are subjective and may require fine-tuning based on feedback and analysis.

- This baseline model assumes a linear combination of metrics, which may be refined in more advanced models.

This baseline model provides a starting point for ranking universities and can be adjusted and expanded upon as needed for your project.

4 Results
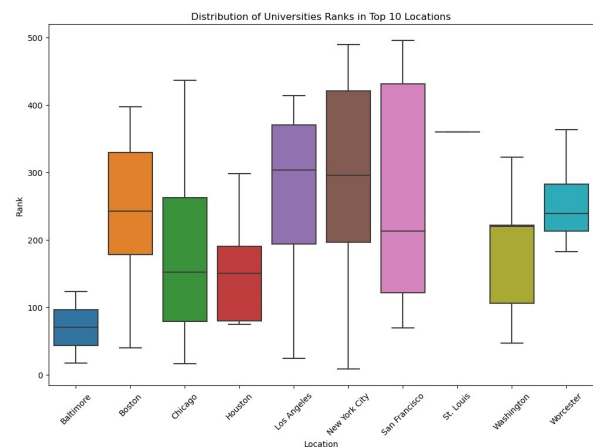
The box plot suggests varying degrees of rank dispersion among universities in the top 10 locations, with some cities like New York City and Los Angeles showing a wide

range of university rankings, while others like Baltimore exhibit a more concentrated rank distribution, indicating a closer performance level among their universities.
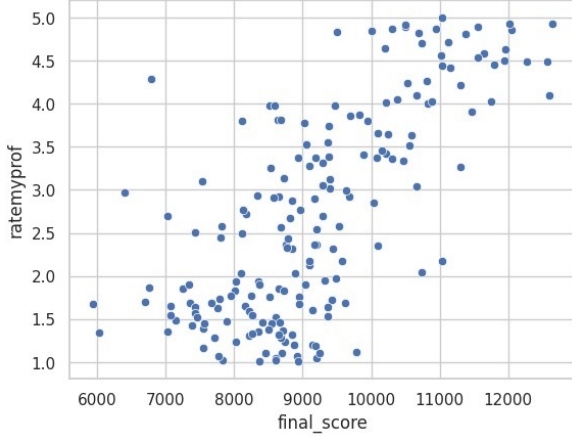
The scatter plot shows a distribution of data points representing the relationship between 'final_score' on the x-axis and 'ratemyprof' on the y-axis. There appears to be a mild positive correlation, indicating that higher final scores might be associated with higher professor ratings.
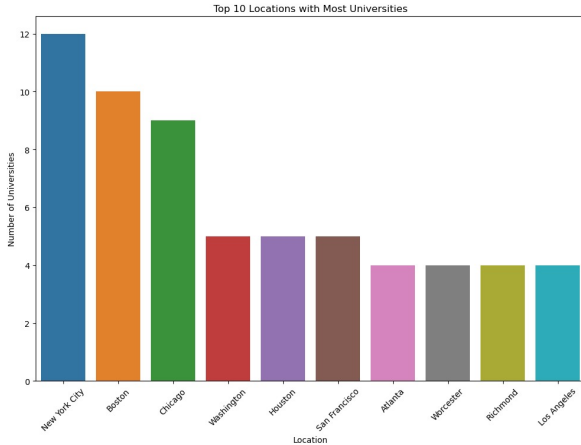
The bar graph provides insight into the concentration of educational institutions across various locations, with New York City having the highest number of universities compared to the other listed cities. This suggests a higher density of educational opportunities or a larger student population in New York City which aligns with the proposed ranking system

Results from our comprehensive analysis of university rankings reveal substantial variations influenced by diverse methodologies, priorities, and data sources. The incorporation of innovative metrics such as the Galloway Score, Completion Rate, and Sports Score contributes to a nuanced understanding of university performance. The Location Score, considering the impact of a university's location, adds an insightful dimension to the evaluation process. By integrating datasets from reputable sources like WalletHub, Forbes, Rate My Professors, and Skiena's dataset, we provide a holistic perspective on over 800 universities globally. Our baseline model, combining key metrics with user-defined weights, offers a flexible approach for ranking universities. The methodological rigor applied to data matching, cleaning, and imputation enhances the reliability of our findings. Moving forward, our research aims to contribute valuable insights beyond traditional evaluations, shaping the discourse on university rankings.

5 Conclusion

In culmination, our research project has conducted an exhaustive exploration into university rankings, encompassing an array of metrics sourced from reputable outlets. By navigating the intricate landscape of varied methodologies, priorities, and target audiences, our endeavor aimed to unravel a nuanced comprehension of academic performance. The inclusion of carefully curated indicators, notably the groundbreaking Location Score, enriches our endeavor, providing a holistic panorama of university standing. The baseline model, harmonizing key metrics with customizable weights, not only serves as a versatile inception point for ranking universities but also underscores our commitment to adaptability in the pursuit of precision.

Acknowledging the inherent subjectivity of metric weights, our project aspires to transcend conventional measurements, furnishing invaluable insights tailored to the discerning needs of both prospective students and policymakers steering the course of educational policies. As we chart our trajectory forward, our dedication persists in refining methodologies, embracing advanced models, and propelling the dialogue on university rankings to new echelons of insight and significance.

References

[1] clarivate.com. `https://clarivate.com/wp-content/uploads/dlm_uploads/2021/07/Global-University-Ranking-Report.pdf`. [Accessed 08-12-2023].

[2] Forbes America's Top Colleges List 2023 — forbes.com. `https://www.forbes.com/top-colleges/`. [Accessed 08-12-2023].

[3] Rate My Professors — Find and rate your professor or school — ratemyprofessors.com. `https://www.ratemyprofessors.com/`. [Accessed 08-12-2023].

[4] Steven Skiena — www3.cs.stonybrook.edu. `https://www3.cs.stonybrook.edu/~skiena/519/`. [Accessed 08-12-2023].

[5] University Rankings: A Closer Look for Research Leaders — elsevier.com. `https://www.elsevier.com/academic-and-government/university-rankings-guide`. [Accessed 08-12-2023].

[6] WalletHub - Improve Your Credit, Compare Financial Products & Reach Top Wallet-Fitness — wallethub.com. `https://wallethub.com/`. [Accessed 08-12-2023].

[7] Ana Carmen Estrada-Real and Francisco J Cantu-Ortiz. A data analytics approach for university competitiveness: the qs world university rankings. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, 16(3):871–891, 2022.

[8] Erhard Rahm, Hong Hai Do, et al. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4):3–13, 2000.

[9] Marlo M Vernon, E Andrew Balas, and Shaher Momani. Are university rankings useful to improve research? a systematic review. *PloS one*, 13(3):e0193762, 2018.