16:958:589:01 ADV PROGRAMMING REPORT

Policy Recommendations for Colleges Using

Data, Algorithms, and Statistics

Professor: Ji Zheng

Course: Advanced Programming 16:958:589:01

May 5th, 2024

Group 1: Kalaivani Sundharam Pillai, Sahitee Ramya Yerramilli Venkata Naga, Mazin Rafi

TABLE OF CONTENTS

1. Introduction

In this project, we explore the evolving landscape of higher education quality in the United States, focusing specifically on predicting shifts in university rankings and their potential impacts on various stakeholders, including students, faculty, and policymakers. By utilizing a range of advanced algorithms, statistical methods, and unsupervised learning techniques—such as linear/lasso regression analysis, learning algorithms like K-means clustering, PCA, and Random Forest—and ranking aggregation techniques, we aim to provide actionable insights and recommendations to assist universities in enhancing their global standing and educational effectiveness.

Our investigation will address key questions:

- How does one assess higher education quality using PCA, K-means clustering, and regression methods?
- Using these results, which of these factors are influencing academic rankings and outcomes?
- How could these shifts affect university policies, student enrollment patterns, and international collaborations?
- How do we develop a robust set of international university rankings?
- How to create and implement a set of policy recommendations for universities based on our findings?

By integrating data-driven predictions with policy analysis, our goal is to bridge the gap between technical analysis and practical application, delivering significant value to the educational community.

To achieve these objectives, we will analyze datasets from respected sources such as the 2024 Times Higher Education, the 2023 Forbes Rankings, and the 2024 QS World University Rankings. These datasets will be instrumental in building our predictive models and conducting thorough scenario analysis to explore the potential impacts of changes in university rankings. Our report is structured to not only highlight the technical processes but also to emphasize practical strategies that can be implemented to maintain or improve the academic standings of higher education institutions.

2. Methodology

2.1 Dataset Selection and Description

Our study leverages multiple datasets to critically evaluate the evolving dynamics of higher education quality and rankings in the United States. The datasets chosen for this analysis include:

1. **2024 Times Higher Education (THE)**: This comprehensive dataset includes detailed information on university rankings with metrics such as the number of students per staff, international student ratio, and domain-specific scores for teaching, research, and industry income.
2. **2023 Forbes Rankings:** This dataset assesses universities based on factors including alumni salaries, debt levels, and overall financial health, which are critical for understanding economic impacts on educational quality.
3. **2024 QS World University Rankings**: Provides a global perspective on U.S. universities, with data on academic and employer reputation among other key performance indicators. The fields are Institution Name, Location, Academic Reputation, Employer Reputation, Faculty-Student Ratio, Citations per Faculty, International Faculty, International Students, International Research Network, Employment Outcomes, and Overall Score.
4. **Machine Learning Approaches for Rankings:**
    - Identify Feature Importance: Apply different models to identify the important features.
    - Prediction: Ranking Analysis from different methods and models including Linear/Lasso Regression, PCA, Random Forest, and K-means clustering.
    - Ranking Aggregation from QS: Borda rank aggregation method to derive new ranks for the QS University Ranking Dataset.
5. **Metrics & Visualization:** Prompt different flavors of questions in Chat GPT and provide multiple visualizations based on key features & various analyses.
6. **Results**: The visualization output is interpreted at a high level for precise understanding.

2.2 Data Collection Methods

The data collection strategy includes:

1. **Direct Import**: Primary sources like THE and QS rankings were directly imported from their respective official websites to ensure data accuracy and reliability.
2. **Web Scraping**: For datasets not available in a structured format, such as certain aspects of the Forbes rankings, we employed Python scripts to scrape and subsequently clean and format the data for analysis.

2.3 Process Flow

- We analyzed datasets together where relevant to create a comprehensive overview, ensuring that data from different sources could be compared effectively.
- We filtered the data to focus solely on U.S. universities, aligning our analysis with the project's scope. For the ranking aggregation, we have extended the QS dataset to include all international universities as well.
- Using Python and ‘scikit-learn’, we implemented solutions to gain insights into trends and patterns, focusing on how various factors influence university rankings and the quality of education.
- Visualizations were created to effectively communicate our findings, emphasizing clarity and interpretability.

3. Data Analysis: Findings and Interpretations

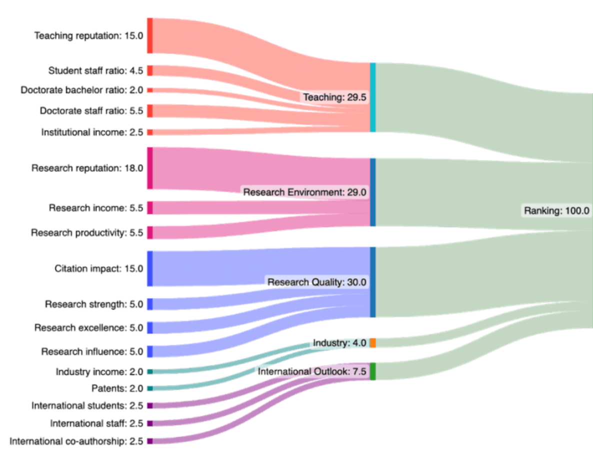
Our analysis emphasizes the complexity of defining and measuring the quality of higher education institutions. While certain factors like academic reputation and faculty-to-student ratios consistently influence rankings, other metrics such as financial aid effectiveness and social mobility rankings offer a more nuanced view of what constitutes quality in higher education. Our findings advocate for stakeholders to consider a broad array of indicators when assessing the performance and quality of educational institutions.

3.1 Visualizations and Their Relevance

Each dataset analysis is accompanied by tailored visualizations, such as cluster groups, scores representation, heat map, PCA explained maximum/minimum variance, elbow line, pie/bar charts, pair plots, and Borda Ranking summary, which highlight key findings and enhance the interpretability of the data. These visualizations are detailed in subsequent sections.

3.2 Times Higher Education

The 2024 Times Higher Education Ranking Dataset, a detailed evaluation of U.S. universities, reveals several key insights based on performance indicators across five areas: Teaching, Research Environment, Research Quality, International Outlook, and Industry (income and patents). A detailed breakdown of the methodology can be found here. Our comprehensive analysis of these rankings has yielded significant insights into the factors that influence university standings.



3.2.1 Regression Analysis of the Times Dataset

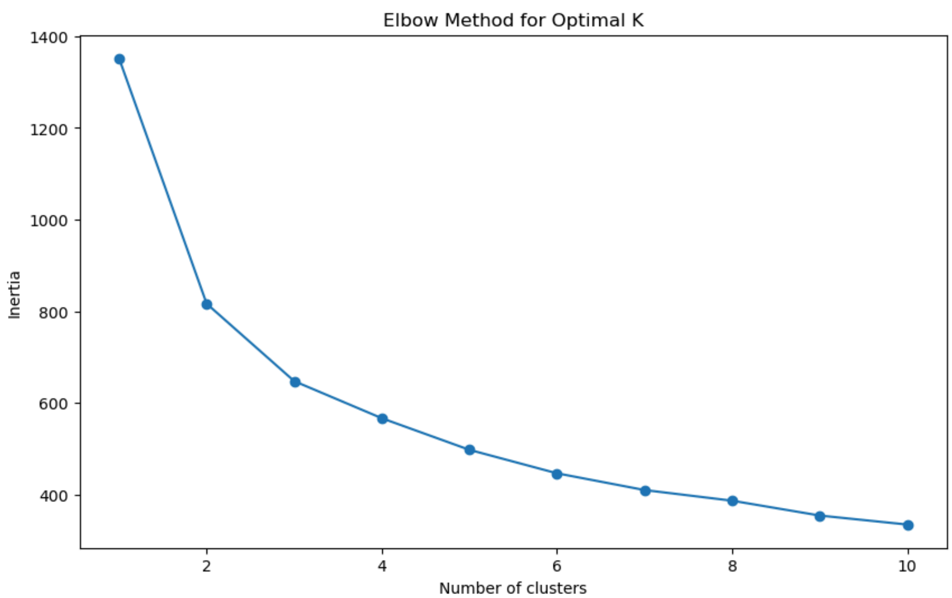
Our regression analysis, utilizing the 2024 Times Higher Education Ranking Dataset, quantitatively assessed the impact of key performance indicators on U.S. university rankings. These indicators are Teaching, Research Environment, Research Quality, Industry, and International Outlook.

Regression Coefficients	Value
Teaching Score	0.296726
Research Environment Score	0.288351
Research Quality Score	0.299535
Industry Score	0.039826
International Outlook Score	0.075175

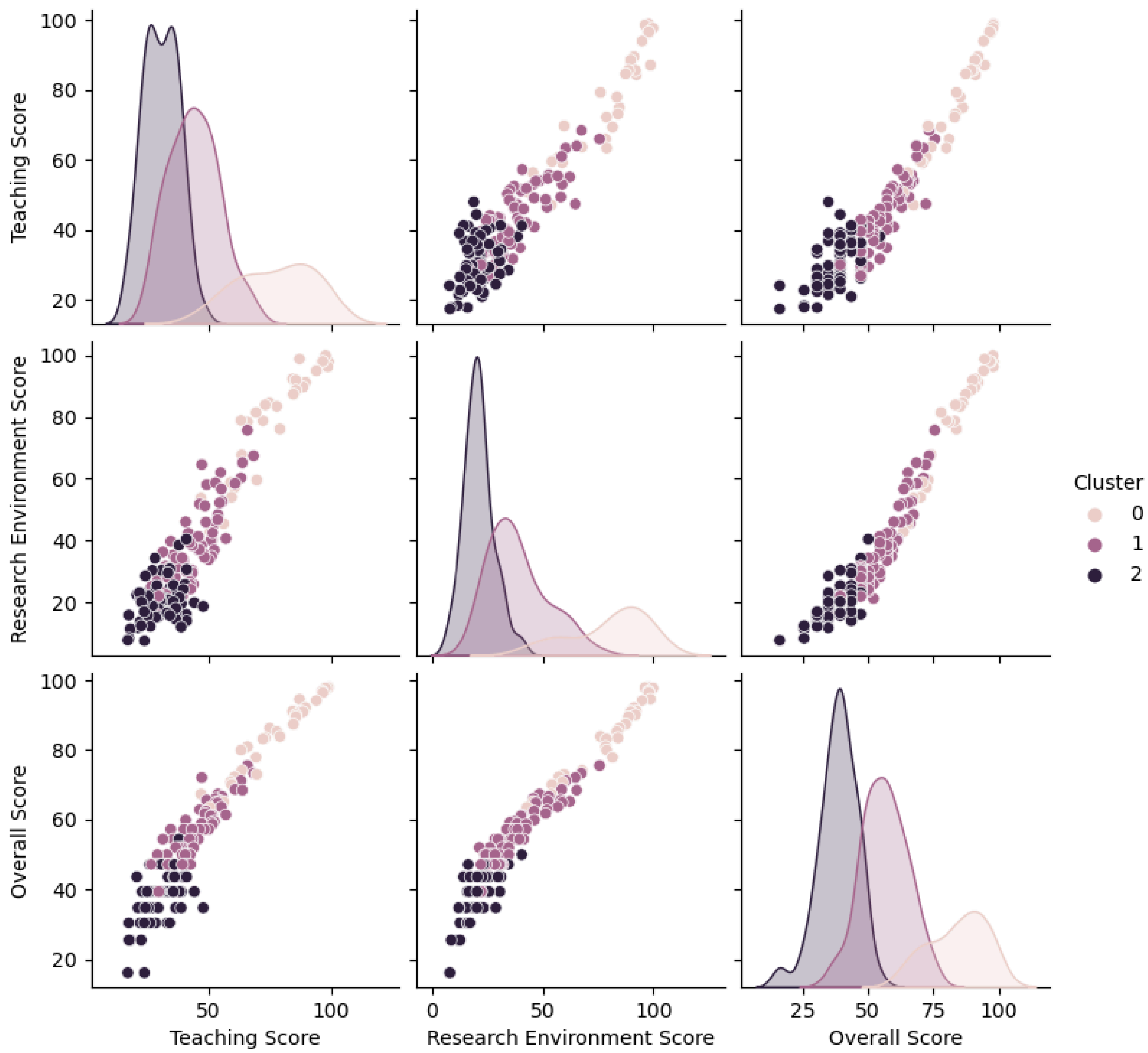
The results confirmed the proposed methodology, with the following regression coefficients demonstrating the importance of each factor and matching closely with the proposed scores.

3.2.2 Clustering Analysis of the Times Dataset

We employed the K-means clustering algorithm to categorize universities into homogeneous groups based on the Times Higher Education Rankings dataset. The optimal number of clusters was determined using the elbow method, where we analyzed inertia to strike a balance between detailed data partitioning and broad interpretability. We selected three clusters (k=3) as this provided a clear and meaningful differentiation among university groups.



We then note an application of the K-means clustering algorithm by comparing the differently assigned clusters below.

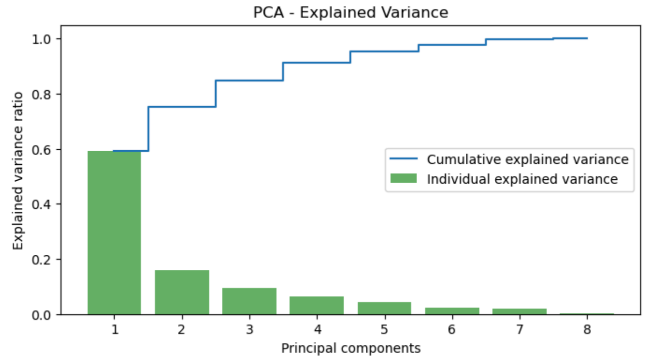


We consider some common features and interpretations of the clusters below:

- Cluster 0 (High Achievers): This group of universities stands out with high teaching, research, industry connections, and international outlook scores, contributing to their leading rankings. Examples of universities within this cluster include MIT, Stanford, and top Ivy Leagues.
- Cluster 1 (Middle Contenders): The middle-tier cluster is characterized by moderate overall scores and large student populations, indicating potential areas for enhancement in teaching and research quality to climb the ranking ladder. State schools fell within this category.
- Cluster 2 (Developing Institutions): The final cluster encompasses universities with the most significant potential for growth, given their lower performance metrics and high student-to-faculty ratios. Focused investments could elevate these institutions in future rankings.

3.2.3 Principal Component Analysis (PCA)

PCA revealed that a few principal components captured the majority of the variance within the dataset. The first principal component is strongly influenced by academic and research-related scores, while the second component seems to represent the student-to-faculty ratio. The PCA has allowed us to distill the dataset into its most critical elements, simplifying the complexity of the data while retaining its most informative attributes.



3.2.4 Times Ranking Policy Recommendations

Our analysis of the Times dataset leads to these cluster-specific policy recommendations:

- For High-ranking Universities (Cluster 0):
  - Excellence Maintenance: Continue to prioritize and invest in high-quality teaching and cutting-edge research. Implement faculty retention programs and recognize academic achievements to sustain their competitive advantage.
  - Strategic Internationalization: Develop global learning communities by expanding study abroad programs, international research partnerships, and dual degree programs. This will bolster the international outlook and diversity, enhancing the educational experience.
- For Middle-ranking Institutions (Cluster 1):
  - Teaching Enhancement: Adopt innovative teaching methods, such as blended learning and flipped classrooms, to improve engagement and learning outcomes. Invest in professional development for faculty to foster pedagogical excellence.
  - Research Amplification: Increase research funding opportunities, streamline grant application processes, and establish interdisciplinary research centers to catalyze innovation and collaboration.
- For Developing Institutions (Cluster 2):
  - Infrastructure Investment: Address the high student-to-faculty ratios by building new facilities, hiring additional faculty, and enhancing student services. This can improve the overall student experience and academic support.
  - Performance Metrics Improvement: Set clear targets for improvement across key performance indicators and monitor progress with a dashboard that can guide strategic decisions. Engage in benchmarking against institutions that have recently improved their rankings to identify best practices.

Each recommendation is tailored to the unique attributes and challenges identified in the respective clusters. They are intended to be implemented in a phased and evaluative manner, ensuring that the impacts can be measured, learned from, and adjusted over time. By aligning policy initiatives with data-driven insights, universities can systematically advance their missions and elevate their standings in global education.

3.3 Forbes Rankings: Universities Financial Analysis

The Forbes College Rankings evaluate institutions on criteria like degree variety, research output, and the specialized focuses that can steer students towards specific career paths. Key financial factors are also assessed, including alums' median salary, grant aid, and the average debt carried by graduates, providing a realistic picture of a college education's investment and potential returns. The dataset can be viewed here for further information: https://www.forbes.com/top-colleges/

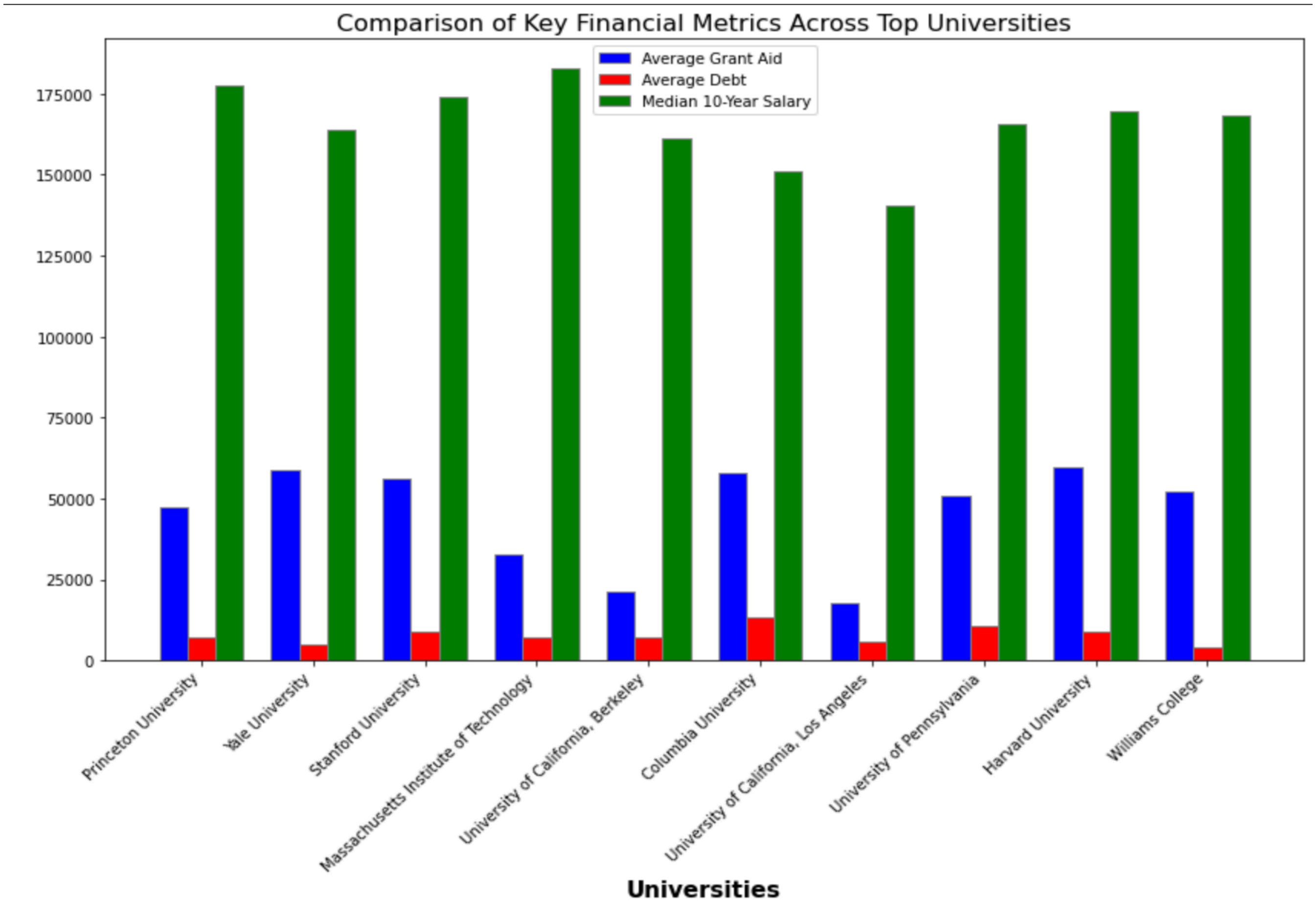
3.3.1 Forbes Rankings: Lasso Regression Analysis on University Rankings

The analysis aimed to identify key financial factors influencing university rankings. Lasso regression was employed due to its efficacy in feature selection, helping to pinpoint the most impactful variables by penalizing less important ones. Key Financial Metrics Identified: Average Debt (AV. DEBT): Higher debt levels were associated with poorer rankings, indicating that universities where students graduate with less debt tend to be viewed more favorably.

Median 10-Year Salary: This metric showed a strong inverse relationship with rankings; higher median salaries are correlated with better university rankings, suggesting a strong post-graduation earning potential as a significant factor.

Average Grant Aid: Although a less impactful factor, more substantial grant aid was slightly positively correlated with better rankings, suggesting that financial support enhances a university's attractiveness and ranking.

Interpretation of Coefficients: The positive coefficient for AV. DEBT implies that as debt increases, rankings worsen. The negative coefficient for MEDIAN 10-YEAR SALARY indicates that higher salaries lead to better (lower) numerical rankings. A positive, albeit minor, influence of AV. GRANT AID on rankings suggests that more aid slightly improves a university's standing. Practical Implications: These insights underline the importance of financial outcomes and support for students when assessing university performance. Prospective students and parents can use these findings to evaluate universities not only on academic quality but also on financial impacts like student debt and earning potential. Recommendations: Universities should focus on strategies to reduce student debt and enhance financial aid to improve their rankings and attractiveness. Higher education policymakers might consider these factors in funding models and educational assessments.

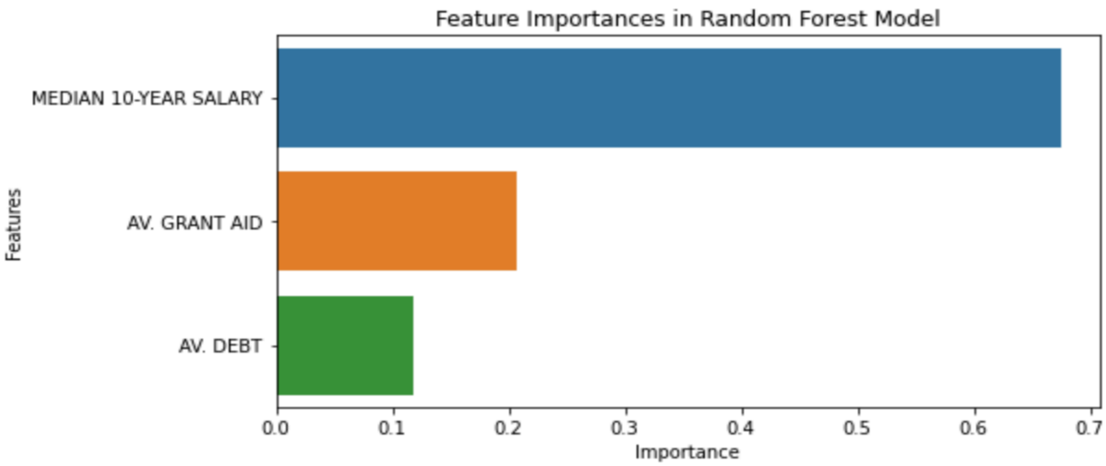


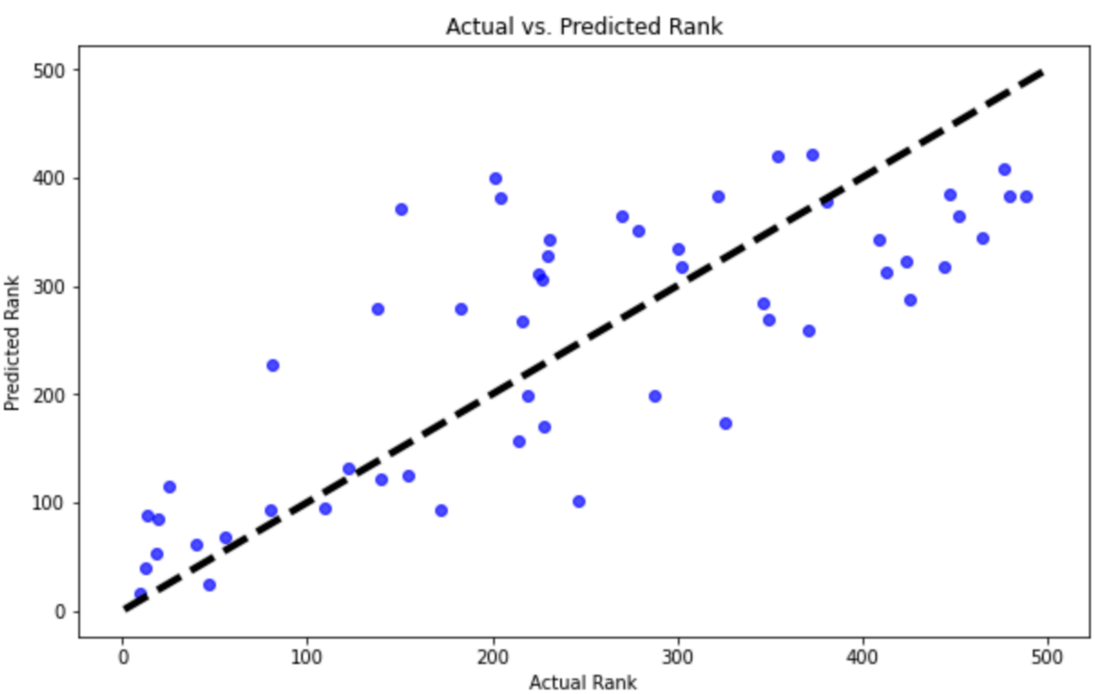
Index	NAME	Score
3	Massachusetts Institute of Technology	1181.313844
70	Harvey Mudd College	1162.342816
0	Princeton University	1161.872637
9	Williams College	1137.202092
2	Stanford University	1136.344354
8	Harvard University	1114.596675
1	Yale University	1110.197642
46	California Institute of Technology	1103.405818
116	Babson College	1102.162246
15	Dartmouth College	1090.396766

3.3.2 Forbes Rankings: Random Forest

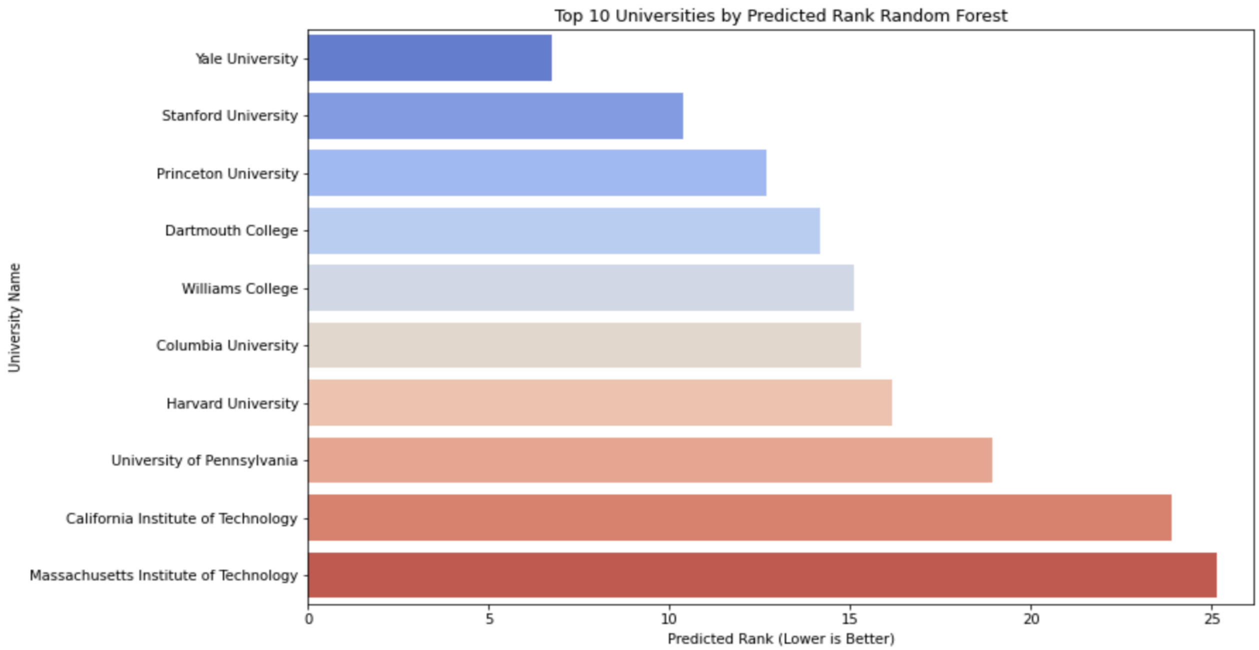
Random Forest presents an analysis using the Random Forest method to assess the impact of financial factors on university rankings. The analysis reveals that the Median 10-Year Salary is the dominant factor, contributing 66.88% to the feature importance in predicting university rankings. This highlights the perception that salary outcomes ten years post-graduation are critical indicators of a university's performance and reputation.

Significant secondary factors include Average Grant Aid, accounting for 20.88% of the importance, suggesting that financial aid enhances a university's ranking by making education more accessible. Average Debt also plays a role, with 12.24% importance, indicating that higher debt levels are typically associated with lower rankings, reflecting the financial burden on students.





The predictive model's performance is depicted in the accompanying plots. The bar chart of feature importance clearly shows the relative influence of each factor, with the Median 10-Year Salary being the most impactful. Another plot shows the top 10 universities by predicted rank, although without annotations, which suggests varying prediction accuracy. Some predictions align closely with actual rankings, demonstrating the model's effectiveness in certain scenarios, while discrepancies in others suggest potential areas for improvement through the integration of additional features or further model refinement. This nuanced analysis provides valuable insights into the factors influencing university rankings and highlights areas for potential improvement in predictive modeling.



3.3.3 Forbes Rankings: PCA Analysis

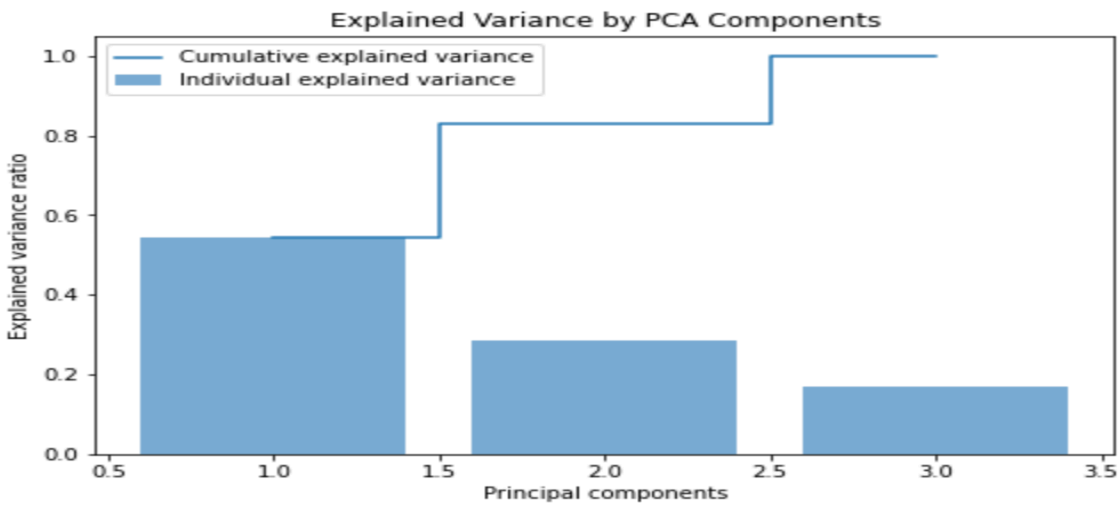
PCA Analysis presents a Principal Component Analysis (PCA) focused on discerning the financial factors that significantly impact university rankings. This analysis simplifies the complex data set, pinpointing the most influential variables in shaping financial health and thereby university standings. The PCA reveals that the Median 10-Year Salary is the most substantial factor, with a feature importance score of 0.686519, highlighting its critical role in influencing university rankings. This suggests that outcomes ten years post-graduation, particularly in terms of salary, are seen as vital indicators of a university's performance and reputation. The analysis also identifies Average Grant Aid as a significant secondary factor with a score of 0.201074, indicating that financial aid positively affects university rankings by making education more accessible. Average Debt, though less impactful with a score of 0.112408, still reflects its notable influence, where higher debt levels are typically associated with lower-ranked universities, underscoring the financial burdens on students.

The PCA plot shows the cumulative and individual explained variances by the principal components, with the first component predominantly influenced by Median 10-Year Salary, affirming its dominance. The second principal component, reflecting contributions from Average

Grant Aid and Average Debt, further illustrates the secondary but meaningful impact these factors have on university rankings.

Principal components

	Feature	Importance/Score
0	MEDIAN 10-YEAR SALARY	0.686519
1	AV. GRANT AID	0.201074
2	AV. DEBT	0.112408



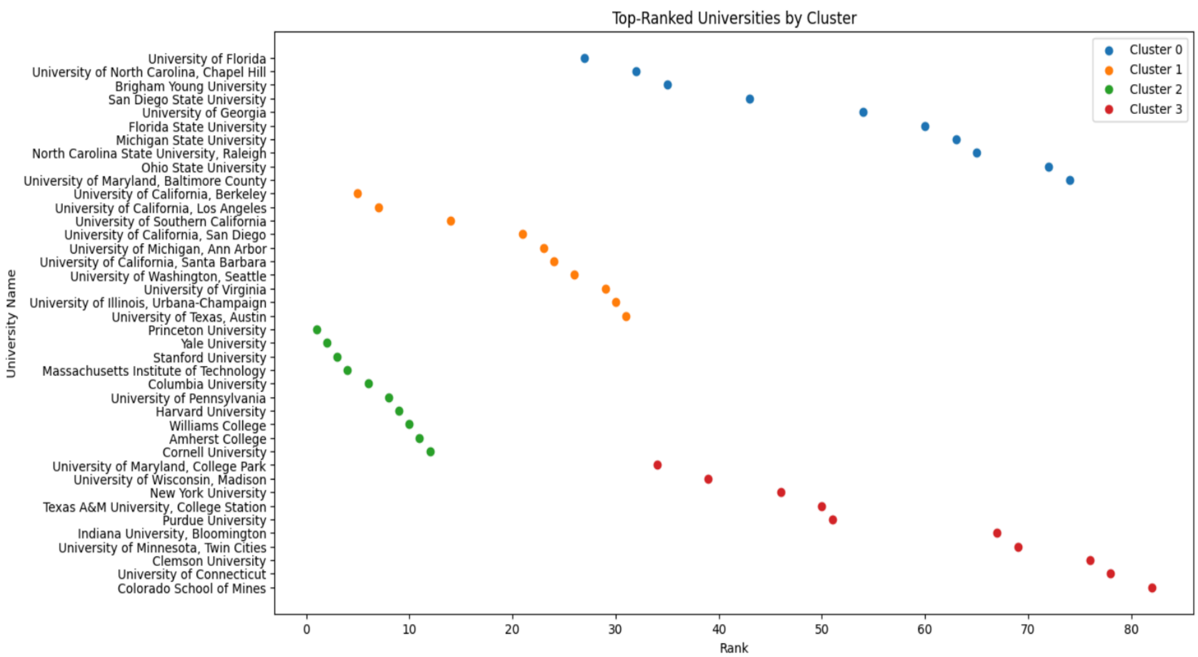
In summary, the PCA streamlined the analysis by focusing on essential elements—salaries, grant aid, and debt—that determine university rankings. This quantified approach aids higher education institutions in strategic financial planning to enhance their rankings and attractiveness through effective financial structuring and support mechanisms.

3.3.4 Forbes Rankings: PCA Analysis Cluster Insights

PCA Analysis Cluster Insights delves into the clustering of universities based on Principal Component Analysis (PCA), which groups institutions into four distinct clusters based on their financial and academic attributes.

- Cluster 0 includes prominent public universities such as the University of Florida and the University of Georgia. These institutions are well-ranked and are noted for potentially offering favorable financial terms, which may contribute to their high standings.

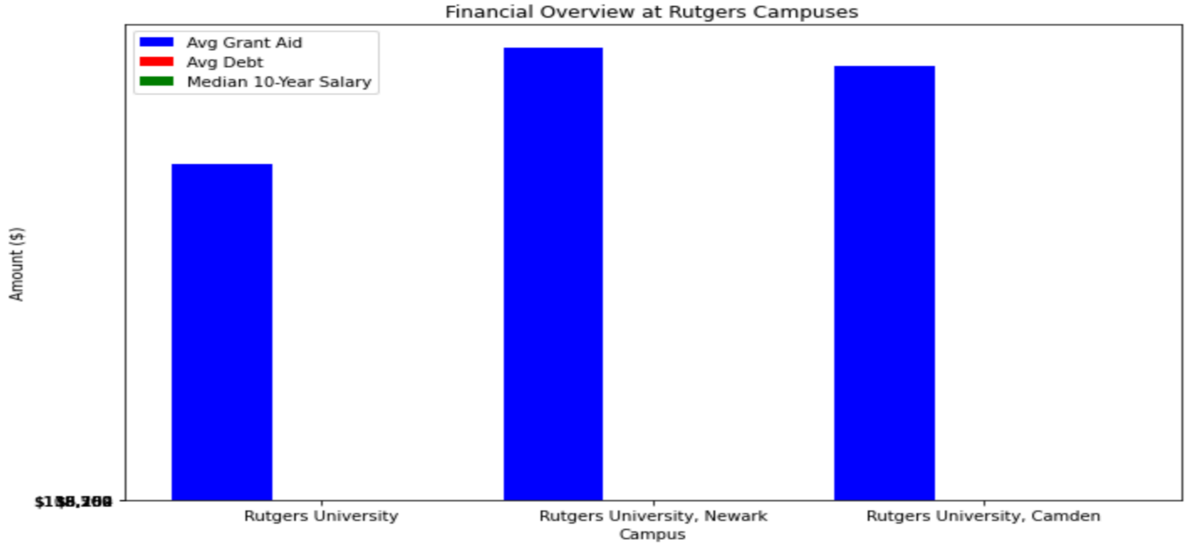
- Cluster 1 comprises a mix of strong public and private institutions, like the University of Maryland and New York University, which are recognized for their solid academic standing and robust financial offerings.
- Cluster 2 features elite public universities from the University of California system and other top public institutions, indicating a high quality of education and favorable financial conditions. This cluster reflects the high standards of education combined with strong financial backing.
- Cluster 3 is dominated by Ivy League and other top private institutions such as Stanford and MIT. These universities are known for their extreme academic prestige and strong financial resources, positioning them uniquely in terms of attractiveness and performance metrics.



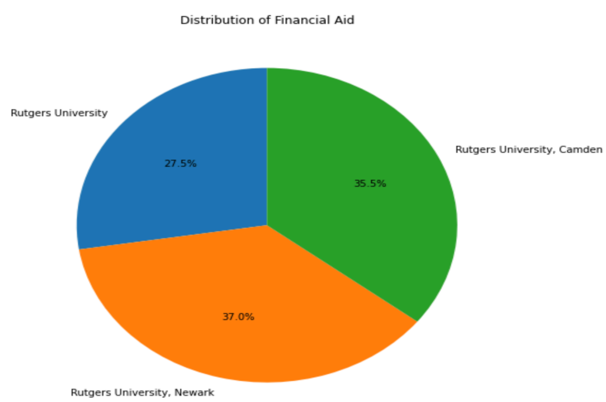
The scatter plot visually categorizes these universities into the defined clusters, plotted against their rankings. Each cluster is color-coded, showing a clear distribution and ranking trend across the clusters. The plot illustrates the positioning of universities within each cluster, with rankings improving from Cluster 0 through Cluster 3, demonstrating how financial and academic attributes assessed by the PCA correlate with higher rankings and presumably better overall performance. This analysis provides insightful categorizations that aid in understanding the strategic positioning and competitive landscapes of major universities.

3.3.5 PCA Analysis: Financial Performance Rutgers

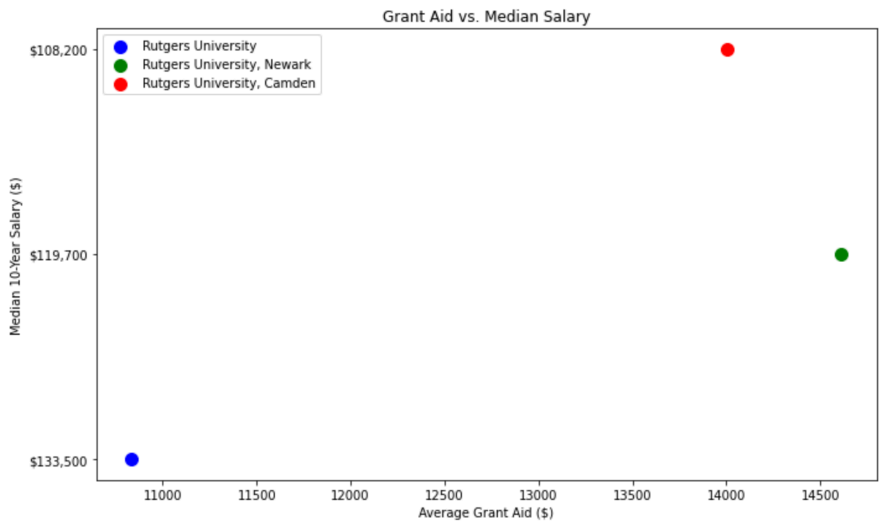
Financial Performance Rutgers provides a detailed analysis of Rutgers University's financial metrics across its main, Newark, and Camden campuses, positioning it within Cluster 2 of the PCA findings. This cluster includes universities with moderate to high financial aid, reasonable debt levels, and strong median salaries, reflecting overall solid financial health and academic outcomes.



Rutgers University's main campus is characterized by a median 10-year salary of $133,357, which surpasses the cluster average, indicating that graduates tend to have good earning potential. However, the campus offers an average grant aid of $14,831, which is below the cluster average, suggesting potential for improvement in financial aid offerings. The average debt amount is $7,767, slightly below the cluster average, depicting a comparatively favorable position in terms of student debt burden.



The pie chart shows the distribution of financial aid across the three Rutgers campuses, highlighting the proportionate share each campus contributes to the total financial aid. Additionally, a bar chart compares the financial metrics of Rutgers' main campus against averages for all universities, public universities, and private universities, specifically showcasing the average grant aid, average debt, and median 10-year salary. The plot titled "Grant Aid vs. Median Salary" graphically contrasts the grant aid and median salary figures for Rutgers' campuses, with the main campus notably higher in median salary but lower in grant aid compared to some peers.



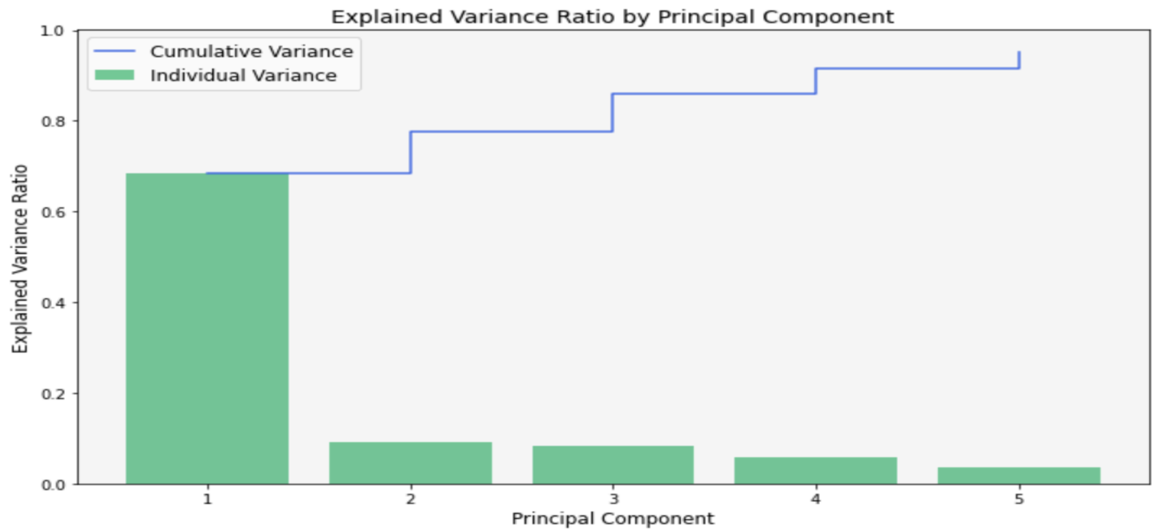
This comprehensive analysis paints a picture of Rutgers as a university with strong earnings outcomes for its graduates but with room to enhance its financial aid offerings to better support its students financially.

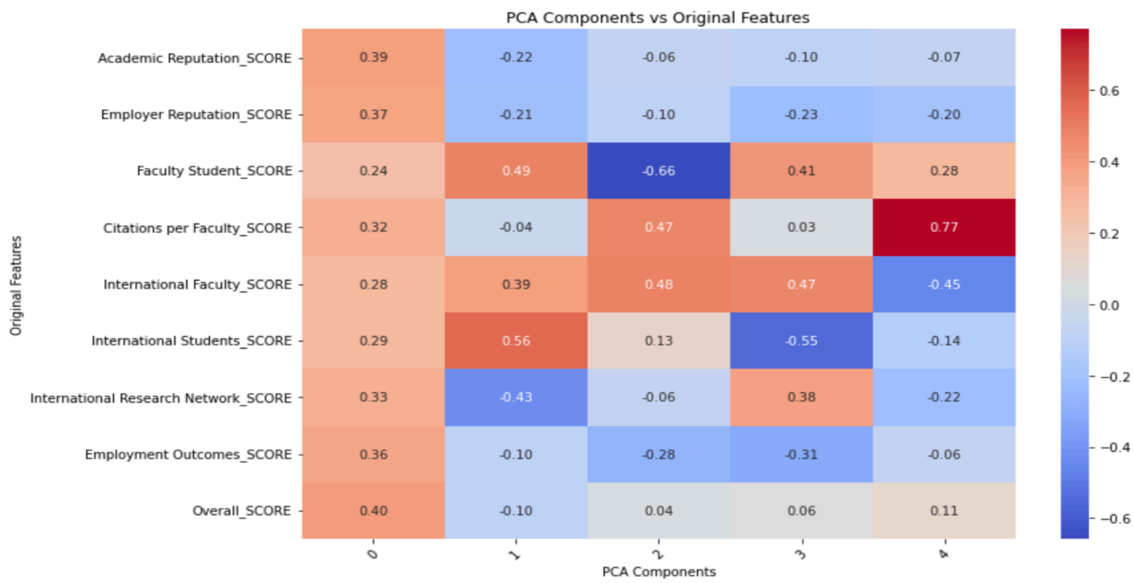
3.4 QS (Quacquarelli Symonds) World University Rankings

The QS University rankings were available globally. The PCA and K-means clustering analyses of the QS World University rankings specifically targeted the year 2024, focusing on universities within the United States. Additionally, rank aggregation encompassed all global universities.

3.4.1 PCA analysis

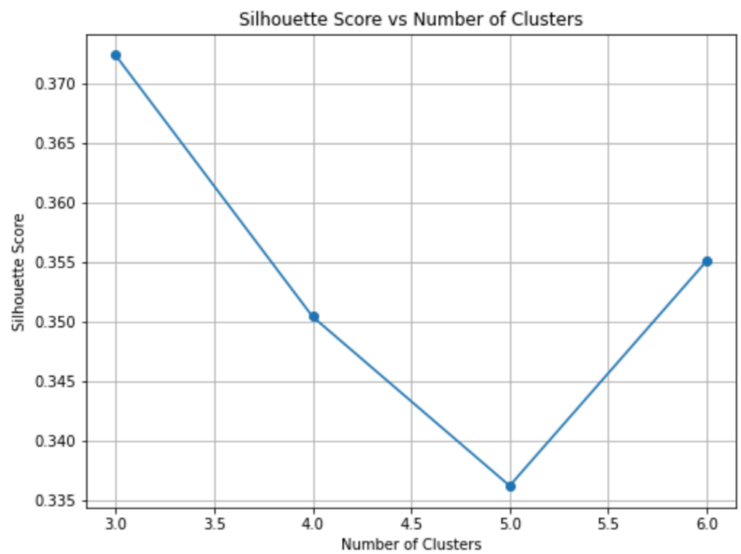
Principal Component Analysis (PCA) is a powerful dimensionality reduction technique widely used in exploratory data analysis and feature extraction. In this study, we applied PCA to the QS University Ranking Dataset to uncover underlying patterns and reduce the dimensionality of the dataset. By transforming the original set of university attributes into a new set of principal components, PCA enables us to capture the most significant sources of variance in the data while preserving as much information as possible. Through this analysis and interpretation of the principal components, we gained valuable insights into the key factors driving university rankings and the relationships among different evaluation criteria. Our findings shed light on the complex interplay of features. By leveraging PCA, we provide a comprehensive understanding of the QS dataset, enabling more informed decision-making in the area of higher education policy and practice.





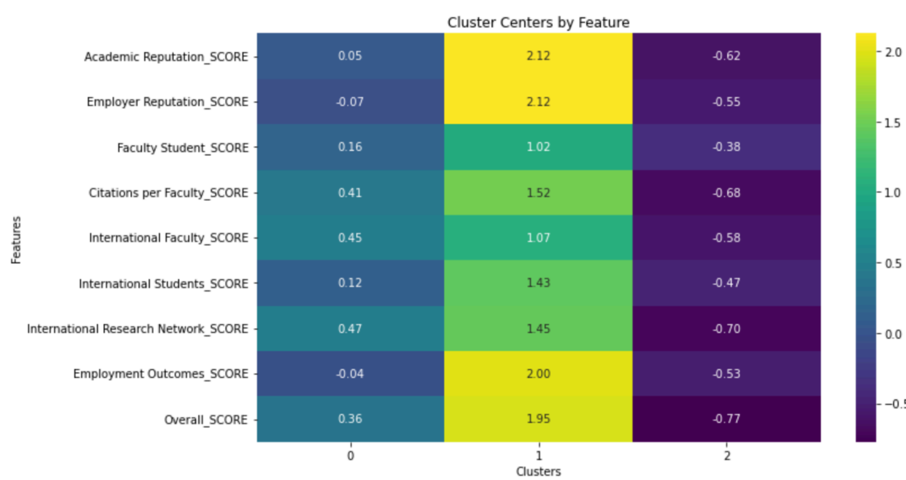
3.4.2 K-means Clustering

K-means clustering was applied to the QS US University rankings dataset to categorize universities into cohesive groups. The selection of the optimal number of clusters, set at 3, was guided by Silhouette analysis, a method that evaluates the similarity of objects within their clusters relative to others. The resulting heatmap visually illustrates the distribution of features across these clusters, enabling clear identification of distinctive patterns and disparities among them.



The three cluster centers suggest that:

- Cluster 0 represents universities with relatively low performance across most metrics.
- Cluster 1 represents universities with outstanding performance across all metrics.
- Cluster 2 represents universities with average to below-average performance, with some strength in Employment Outcomes.



3.4.3. Rank Aggregation :

Rank aggregation methods play a crucial role in synthesizing diverse opinions or evaluations into a unified ranking. In this study, we leveraged Borda ranking, a widely used rank aggregation technique, to consolidate multiple scores and derive new ranks for the QS University Ranking Dataset. By assigning scores to universities based on their positions in individual rankings and aggregating these scores, we construct a comprehensive and robust ranking that captures the collective preferences of various evaluators. The resulting Borda rank provides valuable insights into the relative standing of universities across different evaluation criteria.

Through empirical analysis and comparison with existing rankings, we demonstrate the effectiveness and applicability of the Borda ranking in the context of university rankings. Our findings underscore the importance of rank aggregation techniques in facilitating informed decision-making processes and providing stakeholders with reliable and actionable insights into the landscape of higher education institutions.

The Borda ranking has placed the University of Cambridge at Rank 1 while the QS University ranking placed MIT at Rank 1, which got placed at Rank 14 by Borda.

Institution	2024_Rank	Borda_Rank
University of Cambridge	2	1
University of Oxford	3	2
Imperial College London	6	3
Stanford University	5	4
Harvard University	4	5
UCL	9	6
University of Pennsylvania	12	7
Johns Hopkins University	28	8
Yale University	16	9
Columbia University	23	10
The University of Edinburgh	22	11
National University of Singapore (NUS)	8	12
University of Chicago	11	13
Massachusetts Institute of Technology (MIT)	1	14
The University of Manchester	32	15
King's College London	40	16
Cornell University	13	17
California Institute of Technology (Caltech)	15	18
Nanyang Technological University, Singapore (NTU)	26	19
University of California, Berkeley (UCB)	10	20

4. ChatGPT Findings

4.1 Integration of ChatGPT in the Research Process

Throughout our research, ChatGPT served as an invaluable tool, assisting in various capacities from data interpretation to hypothesis generation. We leveraged its capabilities to enhance our understanding of the datasets, explore analytical angles, and refine our research narratives.

4.2 Key Prompts and Insights

4.2.1 Exploratory Data Analysis Assistance:

- Prompt: What kind of Statistical analysis can we do with Forbes, Times, and QS datasets?
- Insight: Analyzing university rankings from datasets such as QS World University Rankings, Forbes, and Times Higher Education can provide valuable insights into various aspects of educational institutions. Here are some types of statistical analyses you can perform on these datasets:
    - Predictive Modeling:
        - Regression Analysis: Explore relationships between university rankings and other variables (such as tuition fees, faculty-student ratio, and research output).
        - Rank Prediction: Build predictive models to forecast future rankings based on historical data and other influencing factors.
    - Cluster Analysis:
        - Segmentation: Identify clusters of similar universities based on various attributes like location, size, or focus areas.
        - Anomaly Detection: Identify outliers in the data which could indicate errors or unique institutions.

4.2.2 Statistical Interpretation:

- Prompt: How to interpret the cluster centers results from the K-means clustering
- Insight:
    - Cluster 0:
        - This cluster has relatively low scores across most of the features compared to the other clusters. The scores are all below average. The scores for certain features are slightly below average but not as low as the other features.
    - Cluster 1:

        - This cluster has significantly high scores across all features compared to the other clusters. The Scores are all well above average. This cluster likely represents universities with outstanding performance across all evaluated metrics.
    - Cluster 2:
        - This cluster has scores that are generally below average, but not as low as Cluster 0. The Scores are all below average. However, the scores for certain features are higher than the other features, indicating some strength in this aspect compared to other clusters.

4.2.3 Rutgers Financial Performance Insights:

- Prompt: Can you Reveal additional insights for Rutgers Financial Health?
- Insight:
    - Rutgers, with its Newark and Camden campuses, mostly falls into Clusters 1 and 2. These clusters are characterized by universities offering moderate to high financial aid and strong median salary outcomes.
    - For Cluster 2, where Rutgers' main campus is categorized, the average grant aid is approximately $14,831, which is lower than the cluster average, suggesting that there is potential to increase financial aid. The average debt amount is about $7,767, which is slightly below the cluster average, indicating a relatively better situation in terms of student debt.
    - The median 10-year salary for graduates from Rutgers' main campus is $133,357, surpassing the cluster average and indicating that Rutgers graduates have good earning potential. This data underscores Rutgers' competitive position in terms of financial metrics when compared to other universities in the same cluster.

4.2.4 Ranking Aggregation Analysis:

- Prompt: What are some of the rank aggregation methods that can be used for QS University ranking?
- Insight:
    - Rank aggregation is a technique used to combine rankings from multiple sources into a single, aggregated ranking. This can be particularly useful when dealing with datasets like the QS World University Rankings, where universities are ranked by various criteria by different organizations.
    - There are several methods for aggregating rankings, including Borda count, Markov chain Monte Carlo (MCMC) methods, and various optimization algorithms. Each method has its advantages and disadvantages, so you'll need to choose one that best fits your dataset and requirements.

4.2.5 Visualization Recommendations:

- Prompt: what kind of visualizations can I create for PCA, Random Forest, and Lasso Regression?
- Insight:
    - Principal Component Analysis (PCA):
        - Scree Plot: A bar chart showing the variance explained by each principal component. It helps in determining the number of components to retain for further analysis.
        - Loadings Plot: Visualize how strongly each variable influences a principal component.
        - Biplot: Combines a scatter plot of the scores of the first two principal components with vectors representing the loadings. This can help in understanding both the scores and the loadings simultaneously.
        - Cumulative Explained Variance Plot: Shows the cumulative proportion of the variance explained by the principal components, aiding in assessing how many components are needed to explain most of the variance.
    - RandomForest:
        - Feature Importance Plot: A bar chart showing the importance of each feature in the model. This helps in understanding which variables contribute most to the prediction.
        - Confusion Matrix: For classification problems, visualizing the confusion matrix can help assess the model's performance in terms of true positives, false positives, true negatives, and false negatives.
    - Lasso Regression:
        - Coefficient Path Plot: This shows the paths of the coefficients as the regularization penalty changes. It's useful for understanding how each feature's inclusion in the model changes with the penalty.
        - Regularization Plot: This can show the impact of different regularization values on model accuracy or other metrics, helping to choose an appropriate regularization penalty.

4.2.6 What didn't work?

- Prompt: There is a continuous poor R-squared value for the Forbes dataset. Why and what is the reason?
- Insight: There are several reasons why linear regression might not be a particularly effective model for predicting university rankings based on grant aid in the dataset you provided:
    - Weak Linear Relationship:
        - The low R-squared value indicates a weak linear relationship between grant aid and university ranking. Linear regression is most effective when

the relationship between the independent and dependent variables is strong and linear.

- Non-Linear Relationships:
    - The relationship between grant aid and rankings might be non-linear, which linear regression cannot adequately model. For example, there could be diminishing returns on grant aid; once a university provides a certain amount, additional aid might have less impact on improving or maintaining its ranking.
- Rank Reversal:
    - The ranking is a relative measure and inherently ordinal. The actual difference in quality or performance between adjacent ranks might not be uniform or linear. Linear regression assumes a constant interval scale, which is not the case with rankings.
- Distribution of Data:
    - If the distribution of grant aid or rankings is highly skewed, this could violate the assumptions of linear regression, affecting the model's accuracy and interpretation.

5. Policy Recommendations: FACE

Our refined strategic recommendations, encapsulated under the acronym FACE—Financial support, Alumni success, Continuous improvement, and Enhanced allocation—are directly informed by our comprehensive analysis of the Times, QS, and Forbes university rankings datasets. Each recommendation is backed by specific insights from these datasets, ensuring a data-driven approach to optimizing university performance and enhancing student outcomes.

Financial Support Structures (F): Increasing grant aid and reducing student debt burdens form the cornerstone of our financial support policies. Insights from the Forbes dataset, which shows a strong inverse relationship between student debt levels and university rankings, guide this recommendation. Additionally, the Times dataset highlights how financial aid availability correlates with higher academic outcomes and diversity, which in turn enhances university rankings.

Alumni Success Programs (A): We recommend enhancing career support services to ensure alumni achieve high median salaries, a significant factor as shown by the QS rankings, where alumni outcomes heavily influence university standings. The strong correlation between alumni salaries and university rankings, as highlighted in regression analyses from the Forbes dataset, underscores the value of sustained alumni support programs.

Continuous Analytical Improvement (C): Adopting sophisticated data analysis tools is essential for maintaining the accuracy of rankings and adapting to educational trends. This recommendation is based on the successful implementation of machine learning algorithms in the QS dataset analysis, which provided dynamic responses to changes in ranking factors, thereby ensuring that universities remain competitive.

Enhanced Resource Allocation (E): The final facet involves optimizing resource distribution using predictive models, a strategy derived from insights within the Times dataset, where PCA and clustering analyses identified key performance indicators that most effectively influence university rankings and student satisfaction. This strategic allocation of resources ensures that investments are directed toward initiatives most likely to enhance rankings and overall educational quality.

By leveraging data-driven insights from these prominent rankings and methodologies, the FACE recommendations provide universities with a comprehensive framework to improve their operational effectiveness and rankings. This approach not only aligns with the current educational data trends but also sets a foundation for fostering environments that enhance educational outcomes and elevate institutional prestige.

6. Future Analysis

Our future analysis section outlines a strategic approach designed to advance higher education analytics for a deeper understanding and enhancement of university rankings both regionally and globally. By integrating comprehensive, data-driven insights into strategic planning, we aim to adapt and excel in the evolving global educational landscape.

Firstly, we plan to extend the Borda ranking analysis to compare university rankings internationally using the Borda count method. This initiative will aggregate multiple global ranking lists to provide a comprehensive view of international educational standards, identifying regional strengths and areas of improvement in higher education. The outcome of this analysis is expected to be a thorough evaluation of performance that can guide strategic decisions.

Additionally, we will conduct a regional performance evaluation to assess how universities from various regions perform relative to each other. Utilizing clustering techniques, this analysis aims to visualize and analyze regional differences in educational quality and outcomes. The expected result is to highlight critical areas where regional policy adjustments and resource allocation can be optimized to enhance educational standards.

Furthermore, the impact of global trends on local rankings will be examined to understand how shifts in the global educational landscape influence local university rankings. Through scenario analysis, this method will offer insights into adapting local educational strategies to global changes and forecast potential future shifts that could impact local rankings.

Lastly, the strategic implementation of these findings will involve tailoring educational policies based on the insights gained from both regional and global analysis. We will use comparative graphs and heatmaps to represent performance across different regions and globally, enhancing strategic decision-making. This integration of local and global data insights is expected to significantly improve educational outcomes and institutional rankings.

Together, these strategies embody a forward-looking approach to leveraging analytical insights in higher education, aiming to provide a dynamic and responsive framework to navigate and thrive amid global educational challenges.

7. Impact of ChatGPT on the Research

The incorporation of ChatGPT into our research process significantly enhanced the efficiency and depth of our analysis of university rankings. By automating initial data explorations and providing rapid, accurate interpretations of complex statistical outputs, ChatGPT enabled us to identify key trends and anomalies quickly, allowing more time for strategic analysis. This not only sped up the research process but also improved the accuracy and relevance of our findings, especially in the application of advanced techniques like PCA and clustering. Additionally, ChatGPT's capabilities were instrumental in generating insight, leading to our policy recommendations, which both incorporated information from individual datasets such as Times and then applied a more holistic approach encompassing all datasets.

ChatGPT further supported our research by suggesting and helping to implement statistical methods appropriate for analyzing college rankings, including feature importance, the Borda rank aggregation method, the Silhouette analysis, and machine learning models. It also led to the creation of specific visualizations tailored to specific institutions such as Rutgers University, recognizing and differentiating between its multiple campuses. Furthermore, ChatGPT also assisted in code generation and correction, enabling us to resolve setup errors and refine coding approaches efficiently. Overall, ChatGPT demonstrates the transformative potential of integrating advanced AI tools in academic research, enhancing both the scope and impact of our analytical capabilities.

8. Conclusion

Throughout our analysis, we employed advanced analytics techniques to ascertain significant insights into the factors influencing university rankings. Insights from the Times dataset specifically showed how teaching quality, research environment, and international outlook all significantly influence rankings. By applying linear regression, K-means clustering, random forest, and lasso regression, we determined the weighted impacts of various metrics and quantified the influence of financial variables on rankings.

Our analyses highlighted the median 10-year salary as a pivotal determinant of university standings. Principal component analysis across different datasets emphasized the critical roles of median 10-year salary and grant aid in shaping a university's financial health and competitive ranking. The disparity in prediction accuracy between models like Random Forest and PCA illustrates the necessity for ongoing refinement and validation of analytical models. In our exploration of financial metrics, we confirmed that universities offering substantial grant aid and maintaining lower average debt levels significantly enhance their attractiveness and standings in global rankings.

The use of Borda rank aggregation allowed us to integrate various ranking lists, providing a comprehensive view that adjusts for anomalies present when different ranking methodologies are used independently. These findings emphasize the importance of continuous improvement in data-driven methodologies to ensure the reliability and accuracy of predictive analytics in higher education, guiding future policy recommendations and strategic implementations to enhance the educational landscape.

9. References Links

https://www.timeshighereducation.com/world-university-rankings

https://washingtonmonthly.com/2023-college-guide/national-1/

https://www.forbes.com/top-colleges/

https://www.topuniversities.com/world-university-rankings?tab=indicators&countries=us&sort_by=rank&order_by=asc

https://realpython.com/python-web-scraping-practical-introduction/

https://chat.openai.com/