World University Ranking Analysis

Executive Summary

Our team did much research to learn more about the things that affect university rankings, which is an important and interesting topic in higher education. Understanding how the rankings work is essential for students who are making educational choices. We learned about the factors that affect a university's ranking in international lists from these. Our main research question was to find out how these rankings are affected by things like the number of international students, the quality of the research, the number of male and female students, and the school's location.

Understanding university rankings is important for students choosing where to go to school and schools that want to become more well-known around the world. These rankings often affect how well a university is known, how easy it is to get money, and how well it can attract top students. So, it is important to know what factors affect these rankings, both from a practical and an intellectual point of view.

As part of our plan, we used analytical methods. We mostly used Jupyter Notebook and Python to move around and understand the datasets. We looked at a lot of numbers when we analyzed the data, such as the total number of students, the staff-to-student ratio, and different scoring metrics for things like teaching, research, and global outlook. We also looked at categorical variables, like the names and locations of universities. We used several statistical methods to figure out how these variables were related to each other. ANOVA, the Chi-Square Test, Principal Component Analysis (PCA), K-means clustering, and Linear and Decision Tree Regression were used to find out how important the connections were between categorical variables and university rankings. These five statistical methods were used to ensure that the university rankings data was carefully analyzed.

During our research we found some interesting statistics. For example, universities in wealthy areas tend to rank higher. This could be because they get more money and resources. There was also a clear link between university rankings and gender balance. Schools with more gender balance tended to rank higher. This means that a lively, friendly atmosphere could help a university's reputation and output. There was also a strong link discovered between a university's citation score and the amount of research it produced, showing how important high-quality research is for academic prestige. The number of international students at a university seemed to have a positive effect on its international outlook score, which measures how engaged it is with the world.

Although our research provides us with a lot of useful information, we also have to be considerate about the flaws that it comes with. Correlation does not always imply causation, and it is very much possible that we may have left some factors unnoticed, and they could have influenced the results. In diverse cultural and geographical settings, if we think practically, the data may not be as insightful as we found it to be in our analysis. Further studies could investigate more complex topics, like how certain school policies or digital learning platforms affect the rankings of universities.

In the end, our study makes it clear what makes a university different in international rankings. It shows how important things like location, gender equality, high-quality research, and participation from international students are. These new facts may help students make smart choices about their academic goals and help schools plan their way to more international recognition.

Abstract

Higher education is an ever-evolving field that depends on numerous essential elements to operate as it does. Using Kaggle's "World University Ratings" dataset, this study is significant in finding the primary factors influencing university rankings, such as gender ratios, global outlook ratings, and a variety of measures. The findings are essential for universities looking to create more effective strategic plans, boost performance ratings, and drive long-term student planning. The emphasis on top global institutions is crucial for identifying commonalities in an increasingly diverse and more equitable society. The research focuses on economics, gender, academic prowess, and geographic location, all of which have a significant impact on educational policy. In conclusion, this research is an important resource for institutions navigating the challenges of modern university assessments and planning, to create a more inclusive and equitable academic sphere for future generations of students pursuing higher education.

Introduction

In an era where the global landscape of higher education is constantly changing, there is a greater need than ever for insightful methods to assess and comprehend the standing of universities worldwide. This program, which is based on rankings of universities throughout the world, is noteworthy as an essential option to meet the growing demand for comprehensive evaluation. When we delve into the complex world of academia, our goal is to not only understand the rankings themselves but also the complex network of elements that determine a university's place internationally. Because of Python's flexibility and analytical strength, our inquiry takes a dynamic turn, exposing patterns, trends, and correlations that impact the educational environment. People are interested in the aspects that could affect a university's ranking, like location, gender balance, research ratings, and the percentage of international students. The objective is to pinpoint the main factors influencing these rankings in order to offer guidance for bettering educational policies and helping potential students. According to early views, schools in wealthy or historically prominent school districts might rank better. Additionally, it is anticipated that universities with a diverse student body would rank higher, partly due to effective resource management. Additionally, it is anticipated that universities with a diverse student body would rank higher, partly due to effective resource management.

This paper, and the following code is based on the Kaggle dataset “World University Rankings 2023”, and “GDPPrayer”, from IMF, also known as the International Monetary Fund. Utilizing these datasets facilitates an understanding of how the aforementioned factors contribute to the global ranking of a university.

Literature Review

Several studies have been carried out in an effort to gain insight into university rankings and the factors that influence them. With this research project, we intend to take an in-depth examination of the variables that affect university rankings. Our study entails a thorough examination, delving into topics like learning the key determinants of ranking universities, how gender ratios, international outlook score, and other similar factors have a significant impact. We aim to go beyond simple numerical measurements to understand the practical consequences of these rankings for academic institutions as well as students. Our goal is to provide insightful analysis that advances our understanding of the university ranking system and provides useful advice to universities aiming for high standards in higher education.

Investigating the influence of funding on university rankings, Benito et al. (2019) conducted a quantitative evaluation. By using statistical methods and regression analysis on Principal Components, they examined the funding of the institutions in the top 300 rankings. The well-funded top 100 universities had twice as many resources as the schools in positions 101 to 200 and three times as many as the schools in positions 201 to 300. The results of the study show that funding has the biggest influence on institution rankings. Knowledge of funding dynamics and how they impact rankings is crucial to our research on university ranking analysis because it offers useful information for universities looking to strategically improve their academic standing.

In a different approach, Mikryukov et al. (1970) examined the relationship between university performance indicators and ranking functionality using correlation regression and factor analysis techniques. The primary indicators showed a significant link with ranking functionality in the correlation regression analysis. Latent factors impacting baseline indicators were found by factor analysis, with a focus on the importance of some additional entities. The report's recommendations for improving university performance metrics were based on the latent components, offering practical and useful information for our own analysis of university rankings. By implementing their report's suggestions, we can customize our own approaches to enhance university performance metrics.

Sharif et al. (2015) used quantitative methods to analyze factors influencing Times Higher Education (THE) rankings. They used Pairwise correlations and regression models to analyze factors like research, citations, and the international outlook. Pairwise correlation analysis resulted in strong relationships between the independent variables and THE component scores. We position our study to take advantage of a comparable quantitative framework by referencing the findings of Sharif et al. (2015), which expands the range and practicality of our own analysis of university rankings. Authors provide an insightful foundation, providing a conceptual and methodological framework that supports our objectives in carrying out an extensive and practically relevant university ranking analysis.

The three studies that we have cited show numerous ways in which we may use programming for conducting research, collection of data and preprocessing, data analysis, testing types of analysis methods and lastly evaluate the model (if used). The authors of each of these journal studies have used statistical analysis methods, factor analysis, correlation analysis, various regression techniques for analyzing the factors that account for university rankings worldwide and also go ahead to suggest some recommendations on if a specific factor stands out in university rankings, enhance reliability and ensure that the data analysis and models are valid. Additionally, the studies glance at the data to examine the effect of each key metric on the overall ranking of universities separately and offer a deeper understanding on ranking information with analysis methodologies. The studies are of great help to the students and universities irrespective of whether they are big or small, to provide some insights on what factors affect the university rankings. Their analytical techniques seem reliable as they've used reputed data sources for data collection, and appropriate and transparent statistical methods.

Our intention is to incorporate the research findings from the referenced studies into finding any trends that explain why some universities have a higher ranking than others, identifying specific elements and variables that have a significant impact on a university's ranking, and exploring the characteristics of universities that have a high ranking. We also want to investigate the different factors that influence or affect the university rankings. This way we can uncover the influential factors that are directly related to university rankings.

Our main objective is to learn/determine what components and traits of university rankings contribute to a student's decision-making process while applying to college, this will enable in-depth research using descriptive statistics and ranking analysis. Our initial hypothesis tries to correlate gender balance with a university's rank positively. We predict that universities with more evenly distributed gender ratios are probably better in teaching and research, which helps them score higher in those fields. We also anticipate that location would impact a university's rating as we predict that cities with solid facilities for teaching, research will likely perform better and receive higher rankings. Since the exact weightage of each component in calculating the overall rating may vary, we think that each aspect contributes uniquely to the overall position.

This research would assist students to discover potential rank-influencing elements and make them take necessary steps while applying to the universities. Students are enabled to take appropriate measures in their decision-making process while choosing a college or university by verifying which are the actual influential factors, and hence our project research holds immense value.

Methods

To address significant challenges that include the impact of location, gender parity, research scores, and the proportion of foreign students on university rankings, our research aims to determine what influences university rankings and how important it is to take into account the significant impact of factors like international outlook scores and gender ratios. This understanding is essential for developing performance evaluations, guiding strategic planning for prospective students, and influencing educational policy.

Our initial hypothesis proposes that universities in nations with strong economies or well-established educational infrastructures will do better overall. Furthermore, we hypothesize that universities with a balanced gender ratio will rank higher, suggesting a varied and inclusive atmosphere, while institutions with a reasonable student population will score higher due to resource balance.

We attempted to use multiple statistical methods, such as Linear Multiple Regression, Principal Component Analysis, Decision Tree Regression, K-Means Clustering, and Correlation Analysis, as well as hypothesis testing methods such as ANOVA for comparing means across groups and the Chi-square test for exploring associations in categorical data. For Correlation analysis, we analyzed the strength and direction of relationships between variables such as those between research scores and citation frequency. Further, Linear regression was significantly useful in assessing linear relationships, providing insights into the impact of factors like geographical location and gender balance on overall university scores. K-Means Clustering helped to identify patterns and group universities that share similar characteristics, potentially including geographical location or the proportion of international students. The decision tree regression method discovered nonlinear correlations and elaborate patterns, reflecting intricacies beyond linear models. Principal Component Analysis (PCA) made it easier to reduce dimensionality, which in turn made it easier to understand intricate correlations between several variables. By classifying nations according to their economic standing or geographic location, for example, or comparing means across different groups, ANOVA allowed us to evaluate how different university rankings were depending on different factors. The Chi-square test examined correlations between categorical variables, such as analyzing the relationship between gender balance and rankings and discovered significant associations within categorical data.

We utilized Python 3.9.6 as the primary programming language in the Jupyter Notebook web application platform version 6.4.3 for statistical data analysis and data preprocessing. We used several important libraries and packages to thoroughly analyze university rankings and the factors that influence them, including NumPy for numerical operations and the Pandas package for the effective handling of structured data, Scikit-learn for machine learning functionalities such as linear regression and decision tree classification. We used Matplotlib and Seaborn for data visualization to create informative plots and charts. Using detailed data and tools, we aim to clarify what makes a university notable. Although our approach is thorough, it's vital to remember our research's limits.

Discussion and Results

In the Linear Regression Analysis of this study, the model we implemented was designed to evaluate the relationship between university rankings and various factors, potentially including research excellence, gender balance, or ratios of international students. The results showed a relatively low Mean Squared Error (MSE), which suggests that the model's predicted rankings closely aligned with the actual ranking. This really captured the underlying relationships between the variables and the overall ranking. Additionally, the high R-squared value signifies that the model accounts for a significant proportion of the variability in university rankings.

Our next statistical method was K-Means Clustering. This statistical method categorized universities into clusters based on key metrics such as the Teaching Score, Research Score, and International Outlook Score. The identified clusters ranged from those with lower mean scores across all dimensions, representing universities with developing academic profiles, to clusters with moderate and high scores, representing middle-tier and top-tier universities respectively. This clustering method was crucial in grouping universities that shared similar characteristics, including potential geographical location or international student proportions.

Our third statistical method was Principal Component Analysis (PCA). This method was employed to reduce the dataset to two principal components, which together accounted for about 96% of the variance. The first principal component, a weighted combination of teaching, research, and citation scores, predominantly reflects academic prowess. In comparison, the second component captured a significant variance portion, which aligns more closely with the international outlook, highlighting aspects beyond academic scores that contribute to a university's global reputation.

Our fourth statistical method was the Decision Tree Regressor. The analysis revealed a high predictive accuracy, with an R-squared value of 0.961. This model successfully identified complex patterns and interactions between variables influencing university rankings, such as the synergistic effect of high research and teaching scores on improving rankings. However, the variability in the model's predictions, especially for mid-ranked universities, was notable.

The fifth statistical method grouped in with the Chi-Square Test was the ANOVA Test. With an F-statistic of 15.52 and a near-zero p-value, this indicated significant differences in university scores across various categories, such as geographic location or institution type. This finding indicated the importance of these categories in determining a university's score.

Lastly, the Chi-Square Test had a statistic of 838.92 and a p-value close to zero. This showed a strong association between two categorical variables, such as university ranking categories and university types. This result suggests a non-independence between these variables, indicating, for example, that public and private universities might be unequally distributed across different ranking categories.

Across these analyses, several common themes emerged. To start with, the linear regression and decision tree regression both highlight the significant impact of academic scores on university rankings. Also, K-Means clustering and PCA reveal how universities can be grouped based on their scores, indicating common profiles or tiers. Finally, how both ANOVA and Chi-Square tests reinforce the importance of categorical variables such as geographic location and university type in influencing rankings.

Overfitting is a common problem with larger datasets. High R-squared values, especially in regression, might suggest overfitting, particularly if the model is too complex. When we performed

K-Means clustering, it oversimplified the characteristics for universities. In PCA analysis, the interpretation of results could not cover the complete complexity of the variables. The dataset was very sensitive to features/variables when we found out the accuracy of our trained data in decision tree analysis. Another issue is that grouping universities based on K-Means clustering scores oversimplifies variation within each cluster, potentially neglecting the distinctive characteristics of particular colleges.

Given more in-depth research, when we think about completing the research differently, there are a few things that came to our mind. Firstly, we would have liked to have a diversifying range of predictive models to include more machine learning techniques, such as random forests or support vector machines, which would provide a broader understanding of the data. Ensemble methods could also be explored to improve prediction accuracy. There are a wide range of variables that should be considered in world rankings, so we may have looked at some more datasets that give us a context for geographical or cultural settings and perform subgroup analysis to get more insights on the rankings based on region of universities. We would have also thought about considering some additional predictor variables as the education industry is constantly evolving in its nature to keep our research more reasonable.

Conclusion

Our research provided new insights and a thorough understanding of the different factors influencing university rankings. As a group we have discovered that aspects like international student population, research output, and faculty quality play pivotal roles. These findings have substantial implications for universities striving for higher rankings and for students seeking the best educational opportunities. Our research contributes to the ongoing talks about the importance and implications of university rankings in the global educational landscape, offering a nuanced perspective on what drives these rankings and their broader impact.

References

Benito, M., Gil, P., & Romera, R. (2019, September 9). Funding, is it key for standing out in the university rankings? - scientometrics. SpringerLink. https://link.springer.com/article/10.1007/s11192-019-03202-z

Mikryukov, A., &amp; Mazurov, M. (1970, January 1). The task of improving the University Ranking based on the Statistical Analysis Methods. SpringerLink. https://link.springer.com/chapter/10.1007/978-3-030-67133-4_6

Sharif, F. S. (2015, July 1). Characteristics of highly ranked universities in the Times Higher Education (the) World University Rankings. VTechWorks Home. https://vtechworks.lib.vt.edu/handle/10919/79640

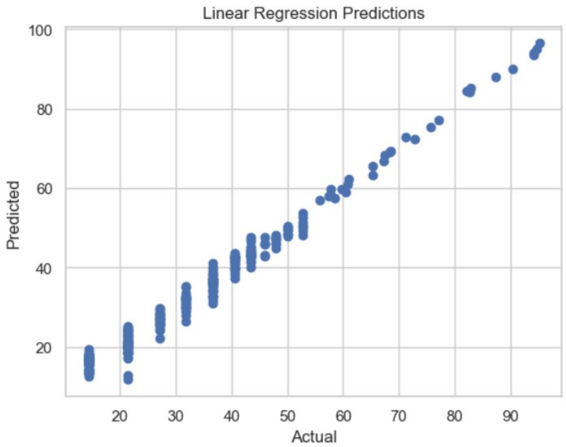
Table:

RESEARCH QUESTION	METHOD	OUTCOME
To what extent do teaching, research, and citation scores contribute to predicting and understanding the university's overall Score' in rankings?	Linear Regression Analysis	Low MSE for the model suggests a high level of accuracy in capturing the relationship between the variables and the overall university score
How do universities group based on teaching, research, and international outlook scores, and what insights do these clusters reveal about shared characteristics?	K- Means Cluster Analysis	The three clusters, characterized by different mean scores across teaching, research, and international outlook, potentially represent tiers of universities with varying academic profiles
What are the key factors contributing significantly to the variability in university rankings?	Principal Component Analysis	Variability in university rankings can be attributed to academic excellence and global presence/ reputation
How does the exploration of complex, non-linear	Decision Tree Regressor	Universities with a certain level of research score

relationships and interactions among influencing factors contribute to understanding university rankings?		significantly improve their ranking if coupled with high teaching scores, indicating a synergistic effect
Is there a significant association between university ranking categorical variables and university types?	Chi-Squared Test	There's a strong relationship between categorical variables, such as university ranking categories and university types (public or private)

Table 1: Summary of the research questions and their outcome

Figures

<table>
<tr>
<td>Figure 1: Linear Regression</td>
<td>


</td>
</tr>
</table>

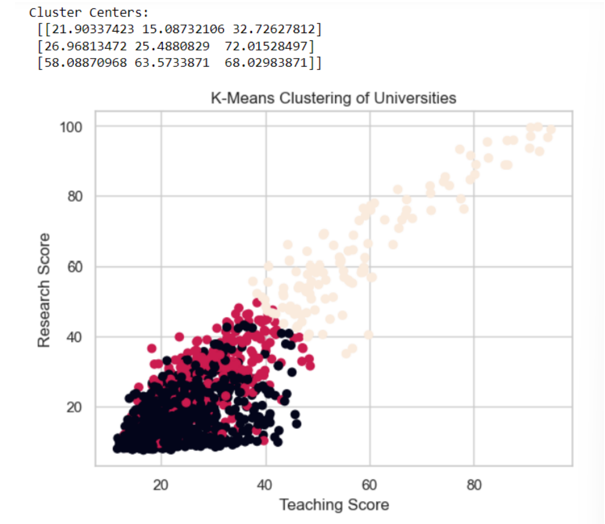
Figure 2: K-Mean Clusters	
	`Cluster Centers:`
	`[[21.90337423 15.08732106 32.72627812]`
	`[26.96813472 25.4880829  72.01528497]`
	`[58.08870968 63.5733871  68.02983871]]`

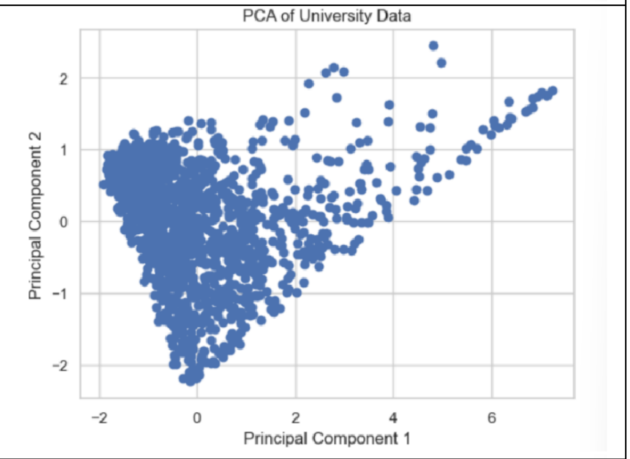
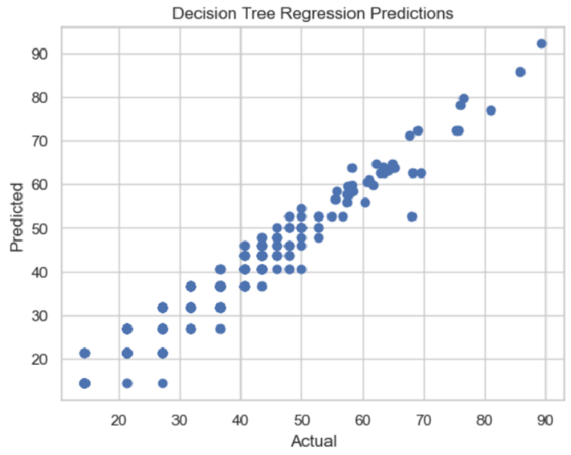
Figure 3: Principle Component Analysis	
Figure 4: Decision Tree Regressor	

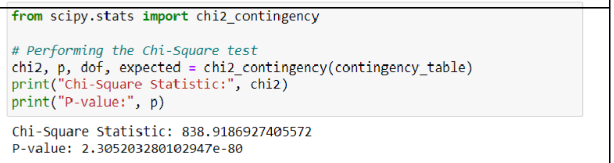
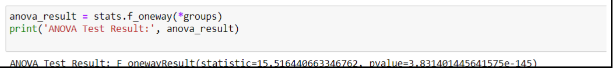
Figure 5: Chi-Squared Test	
	`from scipy.stats import chi2_contingency`
	
	`# Performing the Chi-Square test`
	`chi2, p, dof, expected = chi2_contingency(contingency_table)`
	`print("Chi-Square Statistic:", chi2)`
	`print("P-value:", p)`
	
	`Chi-Square Statistic: 838.9186927405572`
	`P-value: 2.305203280102947e-80`

Figure 6: ANOVA



```
anova_result = stats.f_oneway(*groups)
print('ANOVA Test Result:', anova_result)
```

```
ANOVA Test Result: F_onewayResult(statistic=15.516440663346762, pvalue=3.831401445641575e-145)
```