An analysis of the world university rankings and their methodologies.

Nazia Ferdause Sodial

Abstract — **University rankings have become an effective element for judging the quality of higher education institutions as a result of globalization and the expectation for public entities to be accountable, efficient, and transparent. As a result, it's critical to find out exactly what these rankings are assessing. The Academic Ranking of World Universities (Shanghai ranking) and the Times Higher Education Ranking are explored in this study. Following a description of the ranking methodology, it is established that university rankings are stable over time but vary amongst the two rankings [1]. Additionally, it explores that citations overpower Shanghai ranking systems. Finally, a predictive model is created using the most correlated performance indicators with the rankings of THE ranking to understand if this approach can be adapted by universities to improve their rankings.**

I. Introduction

The education system has been dominating the world for ages and it will never cease to evolve or grow. Many people consider it as an investment and being part of a good university has always been a major quest. As a result of this desire, world university rankings are well-established tools that students, university managers, researchers, and policymakers read and use [2].

The Times Higher Education World University Ranking and The Academic Ranking of World Universities, also known as the Shanghai Ranking are the two most renowned ranking systems. Each ranking has its own methodologies and indicators. The Shanghai Ranking introduces publication and citation data, variables related to the number of Nobel Prizes or full-time equivalent staff. The Times Higher Education World Universities Ranking includes staff-student ratios and a reputational survey within its variables [3]. Globally, universities can take all these methodologies into consideration to improve their rankings and enhance their position in the education domain. The universities not only earn reputation from it but these rankings help them draw students across the globe which in turn boosts the income. These ranking systems influence the education industry majorly and if the dataset can be explored to identify certain characteristics and patterns that can help the universities improve their rankings, then it can be highly rewarding to the whole industry. In this paper, the effort is to explore the top countries, universities, and methodologies that are dominating both the ranking system and then use one of the ranking system's datasets that outperforms the other to predict the rankings.

II. Data and Analytical Questions

A. Data

The data for this study is collected from Kaggle and it includes two CSV files. One of the files belongs to The Times Higher Education World University Ranking (THE) and the other belongs to The Academic Ranking of World Universities, also known as the Shanghai Ranking.

As shared by THE World University ranking system, the below-calibrated performance indicators were taken into consideration to rank a university.

- Teaching (the learning environment)
- Research (volume, income, and reputation)
- Citations (research influence)
- International outlook (staff, students, and research)
- Industry income (knowledge transfer).

As per the Shanghai ranking site, the below methodologies were adapted to rank a university.

- Quality of Education - Alumni of an institution winning Nobel Prizes and Fields Medals
- Quality of Education - Alumni of an institution winning Nobel Prizes and Fields Medals; Staff of an institution winning Nobel Prizes and Fields Medals
- Quality of Faculty - Staff of an institution winning Nobel Prizes and Fields Medals; Highly Cited Researchers
- Research Output - Papers published in Nature and Science; Papers indexed in Science Citation Index-Expanded and Social Science Citation
- Per Capita Performance - Per capita academic performance of an institution.

B. Analytical Questions

The main aim of this study is to find the analytical answers to the below questions:

- Which countries and universities have been dominating these ranking systems?
- Are there any certain components majorly influencing the ranking systems?
- Which ranking system can be argued to have adapted fair methodology?
- Can universities use these indicators to predict their rankings?

Educational institutions can use these analytical questions to estimate and predict their rankings in the near future, based on historical trends. And these models and analysis can be utilized to develop decision-making support [4].

III. Data characteristics

THE ranking dataset has temporal data with ranks in the range of 1 to 800 and the performance indicators are in the form of scores from 0 to 100. Most of the columns have continuous metric and ordinal data. It includes the names of the universities and their respective countries. The dataset doesn’t provide many details about the breakdown of each indicator.

The Shanghai ranking dataset has similar data as THE ranking dataset. The columns include university ranks, names, years, and scores of performance indicators.

The performance indicators are major columns for this study as they provide key insights into the methodologies adopted to rank universities.

IV. Analysis.

A. *Data Preparation*

The below steps were followed to prepare the data for analysis and prediction:

- Cleansing the data – All the unnecessary values, columns were removed including the range values in the rank column.
- Merging the datasets – Both the datasets were merged to visualize how all the methodologies were adopted by individual universities.
- Imputing missing values with median – Upon comparison graphically, median proved to be a better imputation method as it helped to retain the existing variation in the data.
- Handling the outliers – As per figures 1 & 2, the required columns seem to have few outliers. However, these are not experimental errors. It is possible for a university to perform exceptionally and earn a very high score. And it can be unfair to disregard these scores as outliers.
- Standardization – A percentile rank-based method is used by THE system: A cumulative probability function is generated for all indicators except the Academic Reputation Survey, and it is evaluated where a given institution's indication lies within that function using a variation of Z-scoring [5].

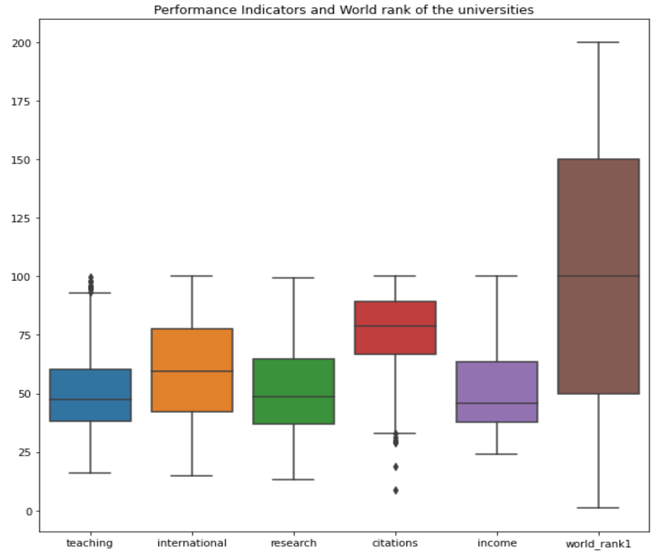

Figure 1. Box plot of THE university ranking’s key columns

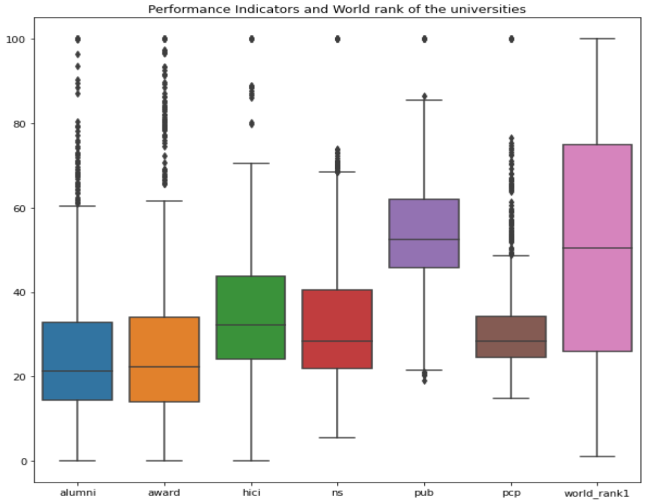

Figure 2. Box plot of Shanghai university ranking’s key columns

B. *Data Derivation*

A word cloud, as shown in Figure 3, was used to find the answer to the first analytical question, revealing that countries such as the United States of America, the United Kingdom, and Australia dominate the ranking system. To get a better insight count plots were used to check the dominance of the countries in the top 100 and 10 as well. From the above steps, it was observed that the USA has been exceptionally dominating the ranking system.

In agreement with Li, Shankar, and Tang, the United States' supremacy is attributable to its enormous population and economic heft, which is bolstered by its high R&D spending (2.7 percent of GDP vs. 0.89 percent in the sample) and English as its primary language [6]. Also, given that the rankings predominantly consider English-language media, that seems to be an easy yes at first glance.

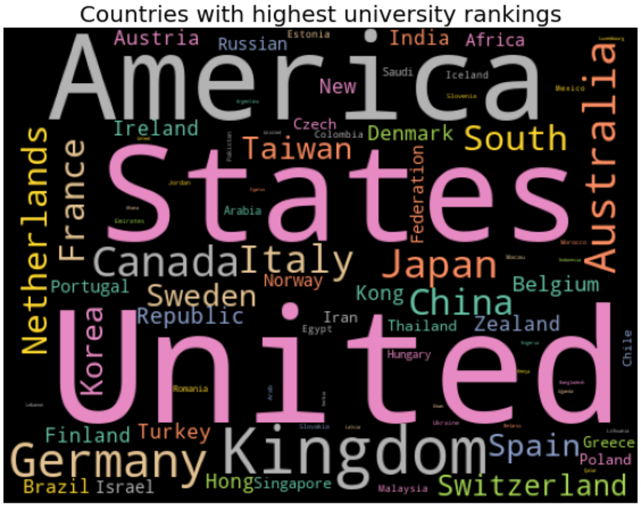

Figure 3. Word cloud of all the countries included in THE university ranking system

From figures 4 and 5, it is deduced that some universities have been consistently dominating the top five ranks in both the ranking system. However, Shanghai rankings have been quite consistent and this curiosity has arisen the next analytical question of taking a look into the methodologies of both the ranking systems.

Methodologies adopted by the Shanghai university ranking system are very inclined towards research-related components. As per its official site, 90% of the weightage is given to awards, papers, and citations related to research. Due to which some universities have consistently secured the top positions leaving no room for improvement in other major educational components.

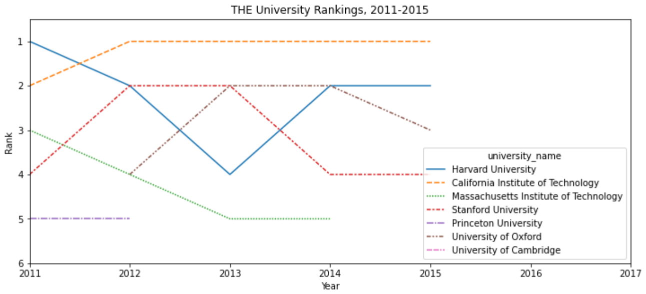

Figure 4. Top 5 rankings of THE University Ranking

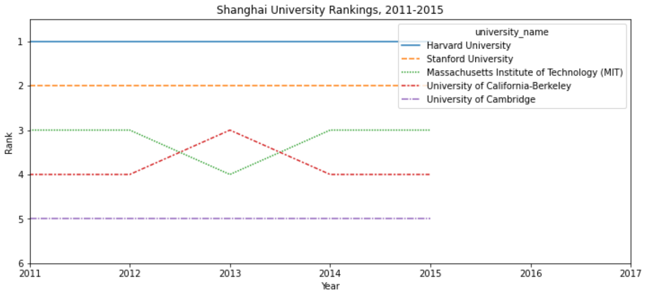

Figure 5. Top 5 rankings of Shanghai Ranking

As per the correlation matrix and pairwise plot, all the components have a positive relationship with the ranks. Although, the sign is negative and shows a negative correlation it is an actually positive relationship as rank 1 is considered as the highest value. Out of these, components like highly cited researchers and papers published in Natural Science are highly correlated with each other and with rankings. In THE University Ranking, teaching and research have a very high positive relationship with the ranks followed by citations. This analysis answers the second analytical question emphasizing the two key components – research and citations in both the ranking systems.

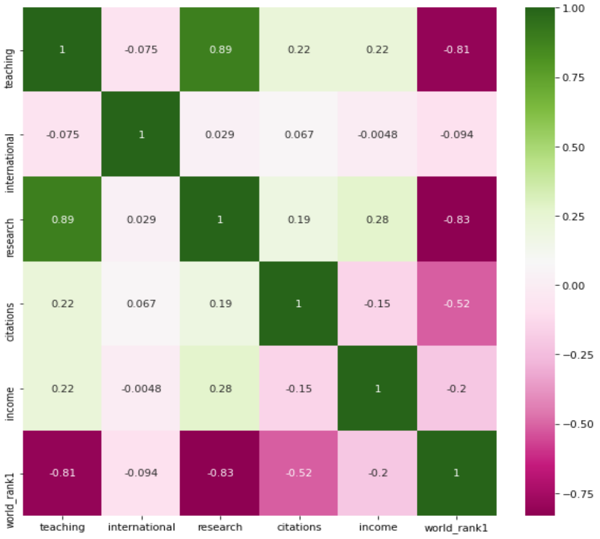

Figure 6. The correlation heatmap of all the performance indicators of THE University Ranking

C. *Construction of predictive model*

Since THE ranking system is much more diverse than the Shanghai ranking in terms of methodology, this dataset was used to create the predictive model. With the help of multiple linear regression, the predictive model was constructed. To build this model the below assumptions were taken into consideration.

- Independence – The first choice of predictors to create the model was teaching and research and citations after considering the correlation matrix as per figure 6. But as per figure 7, teaching and research are highly correlated. Due to multicollinearity, teaching was dropped as it had a comparatively lower correlation coefficient with ranks.

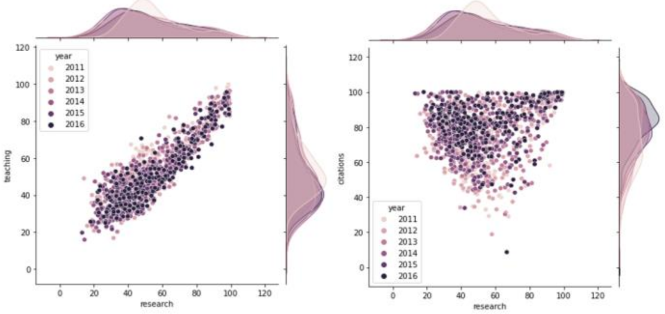

Figure 7. Scatter plot displaying the multicollinearity of the independent variables amongst each other

- Normality – The independent and dependent variables must have normal distribution but with the help of histogram and QQ plots, it was verified that world ranks were not normally distributed. But as confirmed by THE University Ranking that the values are already normalized using z score, no change was observed after log transformation.

- Linearity – As per Figures 6 and 8, the predictors - research and citation have a linear relationship. However, the nonlinearity was also checked using Spearman's coefficient but the linearity score was higher.

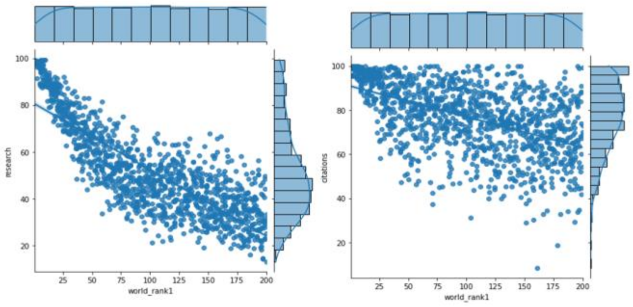

Figure 8. Scatter plot displaying the linearity of the independent and dependent variables.

- Homoscedasticity- While constructing the model, it was assumed that the variance between the residuals is constant.

The formula of multiple linear regression is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

where Y (dependent variable) is world rank, X_1 (independent variable) is research, and X_2 (independent variable) is citations.

And the null hypothesis (H_0) and alternative hypothesis (H_1) were considered while constructing the model.

$$H_0: \beta_1,\ \beta_2 = 0$$

$$H_1: \beta_1, \beta_2 \neq 0$$

D. *Validation of results*

The R^2 and adjusted R^2 values were the same with a value of 0.81 which seemed to be a good fit. The intercept obtained from the model is 322.0315 and the coefficient of the predictors are $\beta_1 = -2.2252$ and $\beta_2 = -1.3756$. The total F value of the model and the corresponding p-value were examined to see if there is a statistically significant association between research and citations with university ranks. Since this p-value is less than $\alpha = 0.05$, the null hypothesis was rejected. Figure 9, was used to visualize the multiple linear regression model.

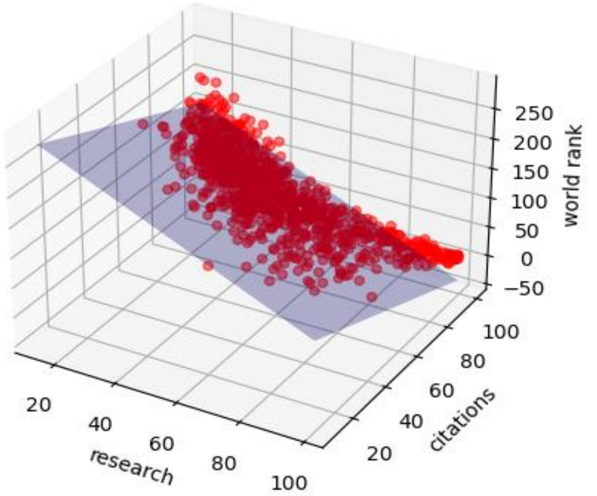

Figure 9. 3D view of the predictive model

V. Finding, reflections, and further work

The below points summarize the findings and reflections of this study.

- To answer the first analytical question, as per Figures 3 and 11, THE World University Ranking system is dominated by a single country – the USA. Also, with further analysis and per Figures 4 and 5, universities from the USA have been securing the top 5 ranks consistently namely Harvard University, Stanford University, MIT, University of California Berkley, and Princeton University from the USA in both the ranking systems. Followed by the University of Cambridge and the University of Oxford from the UK. Since university rankings in Shanghai ranking are determined by the statistical indicators underpinning the ranking score, such as the number of journal publications and the number of Nobel laureates, the 'reason' that the US is well ahead of other countries is because its universities have published many more journal articles, recruited many more Nobel laureates and so forth [6]. However, in agreement with the research paper [6], since the methodologies majorly rely on research, the position of the USA may be challenged if its economic power weakens.

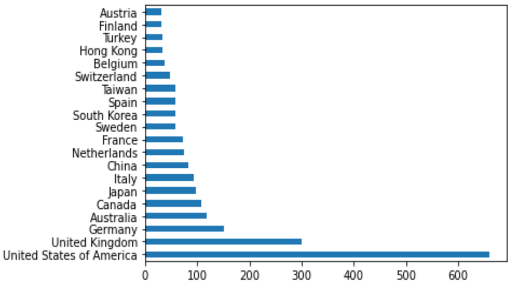

Figure 10. Bar chart of the number of universities included in THE ranking for each country.

- To answer the second, the two components – research and citation have a very strong positive relationship with both the ranks. Universities seeking to improve their rankings can instantly work on these two components.

- The Shanghai ranking system is quite consistent compared to THE rankings and this can be supported with the argument that the Shanghai ranking system is biased towards research and citation and leaves behind other major components like international engagement, the teacher-student ratio which in turn makes THE World University Ranking more reliable. This conclusion not only answers the third question but also leads to the selection of THE World University Ranking dataset for creating the predictive model.

- To answer the fourth question, the ranks of THE World University ranking can be predicted. Although, the model seemed a good fit satisfying almost all the assumptions including Figure 11 but upon checking the scatter plot between the predicted values and residuals, it was sensed that the homoscedasticity assumption was violated because of the presence of heteroscedasticity.

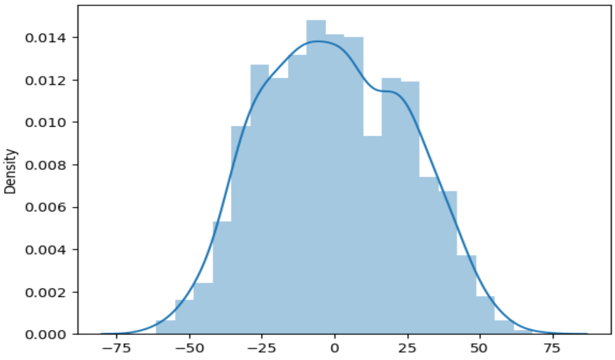

Figure 11. Histogram of the residuals.

To further confirm, Goldfeld Quandt test was conducted with the null hypothesis of the residuals to be homoscedastic and the alternate hypothesis as residuals are heteroscedastic. The p-value turned out to be higher than 0.05. Hence, it was confirmed that the residuals are heteroscedastic. Due to this, it is difficult to rely on the confidence interval and hypothesis testing results.

For future work, an extended analysis of this study can be conducted by embracing the below approaches:

- To take remedial measures for heteroscedasticity and improve the predictive model.
- The reliability of Shanghai ranking is less than THE World University Ranking due to the self-citation concept as well. Although, this investigation is beyond the scope of this study. However, it can be explored with access to data and evidence.
- As per the research paper [3], Principal Component Analysis can be used to support the conclusion of this study more strongly, if research and citations are two major components of these ranking systems.

WORD COUNT:

- ABSTRACT – 130
- INTRODUCTION – 257
- DATA AND ANALYTICAL QUESTIONS – 283
- DATA CHARACTERISTICS – 109
- ANALYSIS – 996
- FINDINGS, REFLECTION AND FUTURE WORK - 536

REFERENCES

[1] Selten, F. *et al.* (2020) ‘A longitudinal analysis of university rankings’, *Quantitative Science Studies*, 1(3), pp. 1109–1135. doi:10.1162/qss_a_00052.

[2] Hazelkorn, E. (2008) ‘Learning to Live with League Tables and Ranking: The Experience of Institutional Leaders’, *Higher Education Policy*, 21(2), pp. 193–215. doi:10.1057/hep.2008.1.

[3] Robinson-Garcia, N. *et al.* (2019) ‘Mining university rankings: Publication output and citation impact as their basis’, *Research Evaluation*, 28(3), pp. 232–240. doi:10.1093/reseval/rvz014.

[4] Althagfi, A. (2017) An Exploration Study of using the Universities Performance and Enrolments Features for Predicting the International Quality. Masters dissertation, Technological University Dublin, 2017. doi:10.21427/D7XK73

[5] Moed, H.F. (2017) ‘A critical comparative analysis of five world university rankings’, Scientometrics, 110(2), pp. 967–990. doi:10.1007/s11192-016-2212-y.

[6] Li, M., Shankar, S. and Tang, K.K. (no date) ‘Why does the US dominate university league tables?’, p. 38.