Python II Project Report:

What factors can have a significant influence in the worldwide rankings and scoring of a university? Can the publications and citations attributed to a university impact its ranking and scoring?

By:

Abdiel Lugo (800776400)

Brandon Gresham (800656861)

Cabel McCandless (800782628)

CS 154/454– Python II

Instructor: Poom Pianpak

Spring 2023

New Mexico State University (NMSU)

5/12/2023

Introduction:

University rankings are a familiar phenomenon in higher education all over the world. Based on a literature overview, the earliest worldwide university rankings were found to be in the year 2004 and ever since it has been a yearly publication showing a university's increasing, decreasing, or stabilized improvement and reputation. Worldwide rankings differ significantly from association to association depending on the methodology applied in their evaluation. In most methodologies, the ranking is directly influenced by the overall scoring attributed to a university. This poses a key aspect in a university's worldwide ranking; hence it needs to be considered in any evaluation of these types of evaluations. Two of the most well-known associations that provide their worldwide university rankings and scoring based on different evaluation methodologies and factors are the Times Higher Education University Rankings and the Shanghai Academic Ranking of World Universities. Even though both ranking and scoring methodologies consider different factors for evaluation, are there some semi-common factors between them that can influence a university's world ranking/scoring? Can factors such as the publications and citations attributed to a university be influential in their ranking and scoring?

The focus on both publications and citations as a possible indicator of good ranking and scoring in a university in this work is mostly attributed to the fact that successful publications of research projects can bring essential attention to scholars and their institutions. Several factors can affect the rate and quality of publications in a university, with some of the most thoroughly documented ones being personal, environmental, and situational factors (Wahid et al., 2021). In addition, any minimum changes in the three zones of global, local and university policies can seriously affect universities short-term and even long-term publication achievements (Moghdam et al., 2015). The goal for this project is to show what factors can significantly influence a university's world ranking along with its overall scoring. Factors such as the number of alumni, awards, citation indicators, and others are to be considered to answer the proposed question. In addition, identification of how different factors can directly impact an institution's overall score will also be explored. It's important to note that two worldwide university ranking datasets from both the Times and Shanghai rankings ranging from the year 2012-2015 are going to be investigated to assess the proposed rankings and assess how publications and citations indicators, among other indicators, accredited to a university can have a correlation between its ranking and overall scoring.

Project Proposal Question:

What factors can have a significant influence in the worldwide rankings and scoring of a university? Can the publications and citations attributed to a university impact its ranking and scoring?

Dataset Overview:

Link for the datasets utilized in this study: https://www.kaggle.com/datasets/mylesoneill/world-university-rankings

The two datasets of focus were the Times Higher Education University Rankings (THEUR) and the Shanghai Academic Ranking of World Universities (SARWU). The annual THEUR started in 2004 and aims at the effective provision of the definitive list of the best universities. The current methodology for the evaluation of the different universities was developed in 2010 and it evaluates across five key areas of Teaching (the learning environment), Research (volume, income, and reputation), Citations (research influence), International Outlook (staff, students, and research), and Industry Income (knowledge transfer). The THEUR data is trusted by governments and universities and is a vital resource for students, helping them choose where to study. The SARWU dataset considers every university that has any Nobel Laureates (alumni), Fields Medalists (award), Highly Cited Researchers (hici), or papers published in Nature or Science (ns). In addition, universities with a significant number of papers indexed by Science Citation Index-Expanded and Social Science Citation Index are also included (pub). In total, more than 2,500 universities are ranked, and the best 1000 are published. Dataset information was obtained through the of panda's data frame .info() function and are shown below in Figure 1:

```
Times University Rankings Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1603 entries, 200 to 1802
Data columns (total 14 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   world_rank            1603 non-null   int64
 1   university_name       1603 non-null   string
 2   country               1603 non-null   object
 3   teaching              1603 non-null   float64
 4   international         1603 non-null   float64
 5   research              1603 non-null   float64
 6   citations             1603 non-null   float64
 7   income                1603 non-null   float64
 8   total_score           1603 non-null   float64
 9   num_students          1557 non-null   object
 10  student_staff_ratio   1557 non-null   float64
 11  international_students  1553 non-null  object
 12  female_male_ratio     1453 non-null   object
 13  year                  1603 non-null   int64
dtypes: float64(7), int64(2), object(4), string(1)
memory usage: 175.5+ KB
```

```
Shanghai University Rankings Information:
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1383 entries, 3514 to 4896
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   world_rank       1383 non-null   int64
 1   university_name  1382 non-null   string
 2   national_rank    1382 non-null   object
 3   total_score      398 non-null    float64
 4   alumni           1382 non-null   float64
 5   award            1381 non-null   float64
 6   hici             1381 non-null   float64
 7   ns               1376 non-null   float64
 8   pub              1381 non-null   float64
 9   pcp              1381 non-null   float64
 10  year             1383 non-null   int64
dtypes: float64(7), int64(2), object(1), string(1)
memory usage: 129.7+ KB
```

Figure 1: THEUR and SARWU dataset information and datatypes

Approach and Assessment:

The successful integration of multiple datasets was originally a key primary step into providing an exploratory data analysis. Yet, upon further evaluation, a complete integration of both datasets was not possible due to the differing methodologies and indicators involved in both datasets. While a successful integration of the different datasets can help to potentially generate possible trends that may indicate as favorable to boosting a university's world ranking, it will not provide accurate answers since the metrics are different in both datasets. Therefore, in this work the two datasets of focus, THEUR and SARWU, were evaluated separately with the purpose of identifying commonalities between them in terms of their differing metrics to influence the ranking and scoring of a university. An initial step for the evaluation of both datasets consisted in transforming the datatypes in both from object to floats and strings. After the conversion of the data types, a correlation matrix was done using Seaborn heatmap plot to provide the correlation of the different features in both datasets to the targets (Ranking and Total Scoring). These are shown in Figure 2 and Figure 3. The ranking and total scoring columns and rows are highlighted in red. It's important to note that the lower the ranking the better the university, hence negative correlations are expected to impact positively for the ranking of a university. For the THEUR dataset (Figure 2), it can be observed that the features that influence both the ranking and total scoring are the research, citations, and teaching features. For the SARWU dataset (Figure 3), it can be observed that the features that influence both the ranking and total scoring are the papers published in Nature or Science (ns), highly cited researchers (hici), and staff awards (award). Therefore, a focus of comparison between universities is going to be centered around the influence of these features in both ranking and the score of a university.

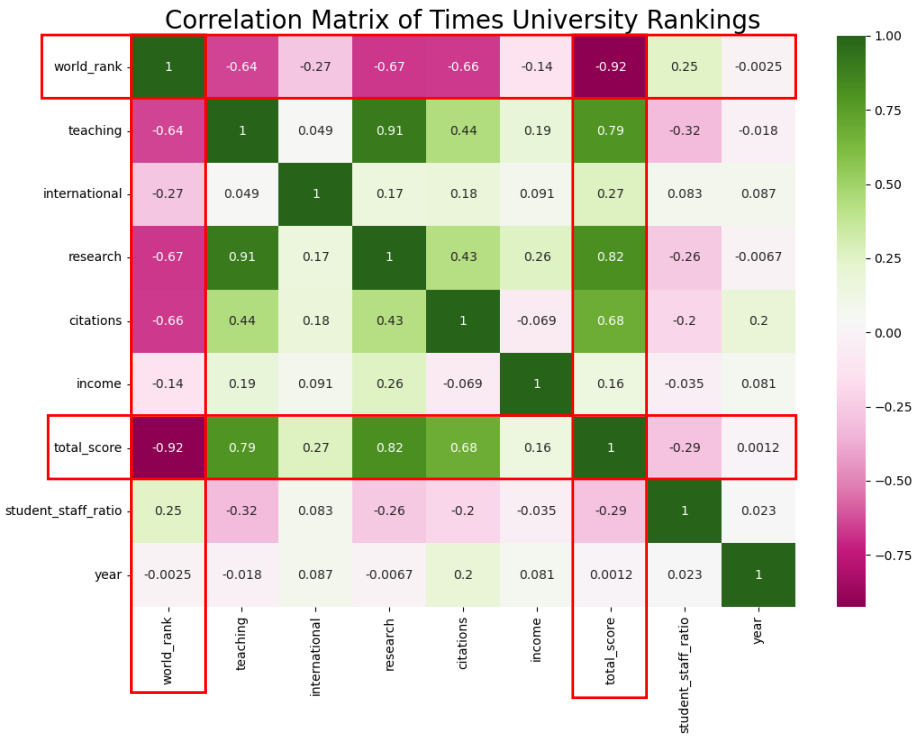

Figure 2: Seaborn heatmap representing the correlation of features in THEUR dataset.

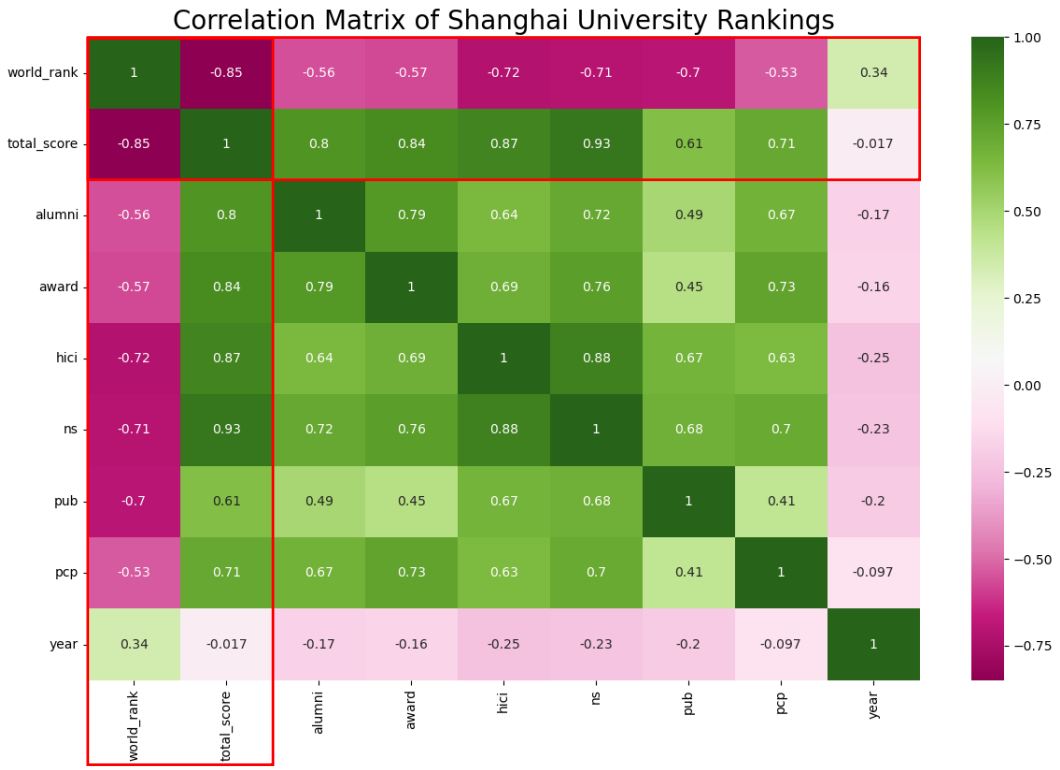

Figure 3: Seaborn heatmap representing the correlation of features in SARWU dataset.

To avoid large amounts of data processing, only the top 25 universities in both datasets were evaluated. It's important to note that the ranking on both these datasets differ quite a bit, therefore the goal is to evaluate universities they both have in common inside their top 25 rankings. In this case, the number of universities that both datasets have in common inside the top 25 were 15 universities:

1. California Institute of Technology
2. Columbia University
3. Cornell University
4. Harvard University
5. Princeton University
6. Stanford University
7. University College London
8. University of California, Berkeley
9. University of California, Los Angeles
10. University of Cambridge
11. University of Chicago
12. University of Oxford
13. University of Pennsylvania
14. University of Washington
15. Yale University

With these universities being the center of focus in this evaluation, the developed Python program gives the user the options of selecting two universities to compare them in either their ranking or scoring.

Some case studies were selected in this evaluation:

1. Harvard University vs Princeton University

This first case study shows a comparison between Harvard University and Princeton University, two highly rated Universities that are consistently in the top 10 universities according to Shanghai rankings. From both ranking (Figure 4) and scoring (Figure 5) shows that Harvard University outranks (and outscored) Princeton University relative to the dataset they originate from.

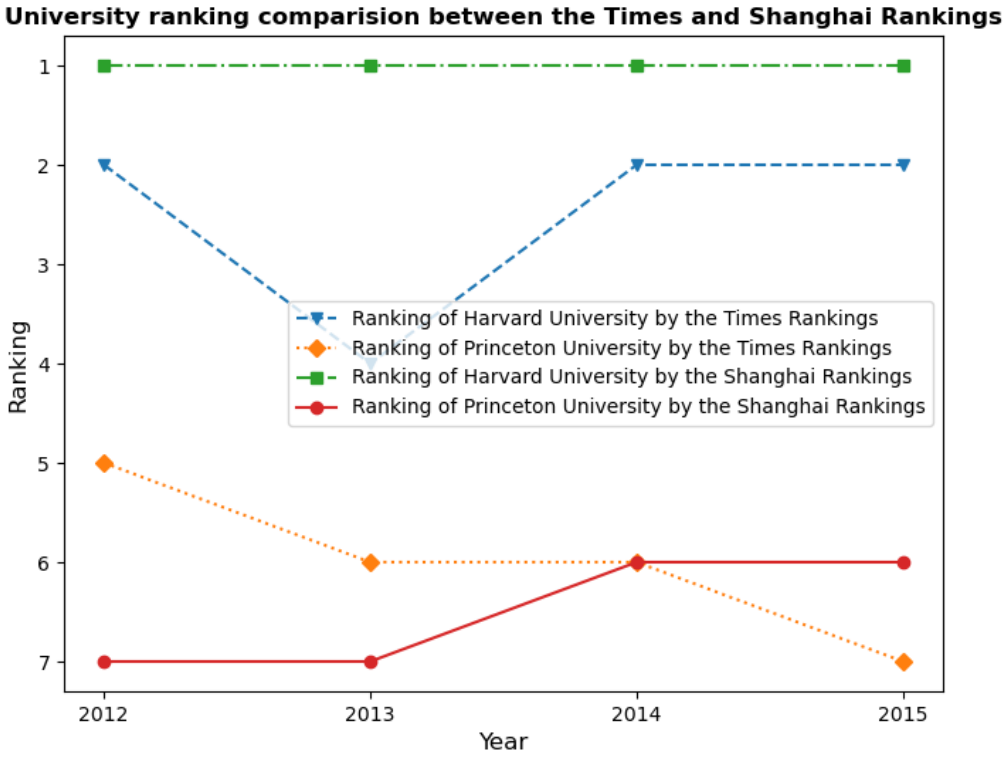

Figure 4: University Ranking Comparison Between the Times and Shanghai Rankings of Harvard University and Princeton University

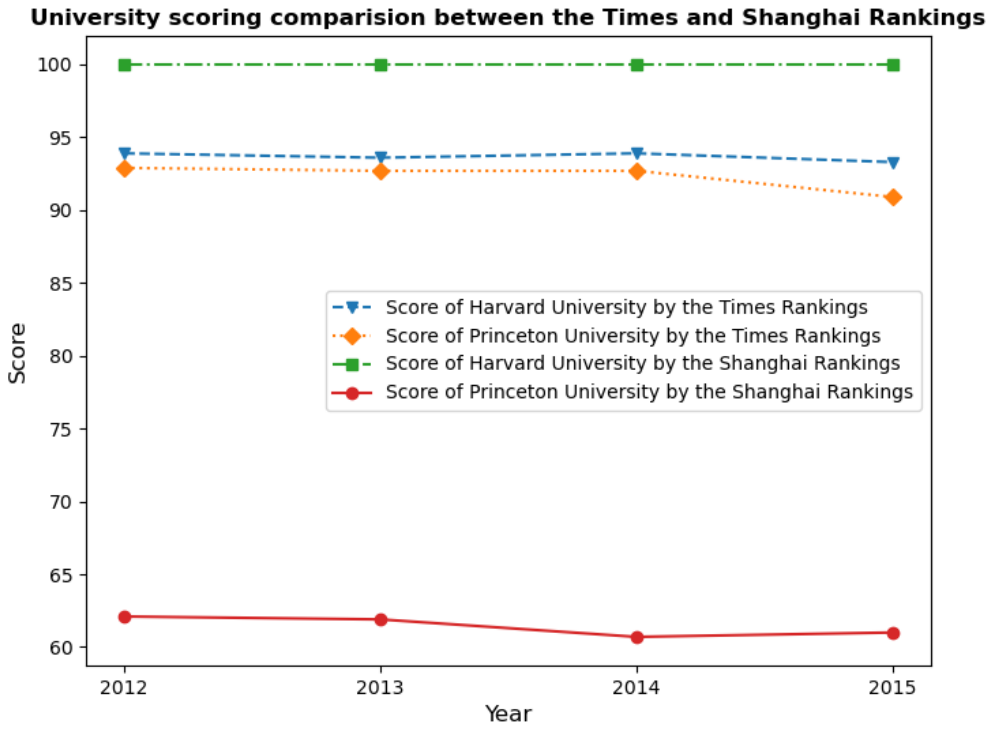

Figure 5: University Scoring Comparison Between the Times and Shanghai Scoring of Harvard University and Princeton University

In Figure 6 there is a general correlation between the parameter of the teaching rankings and the overall rank according to the Times Dataset. When observing the teaching rating of Princeton University, the rankings are grouped around a teaching score of 84 to 91. Likewise, when observing the teaching rankings for Harvard University that rankings are grouped around 93 to 95. The data from the citations and research sections shows that both universities perform exceptionally well in these fields. Both universities received a score above 98 for citations and above 97 for research between the years 2012-2015. This data shows that even though they are relatively matched in both citations and research, because of their differences in teaching it causes Harvard University to have a higher ranking in the Times dataset.

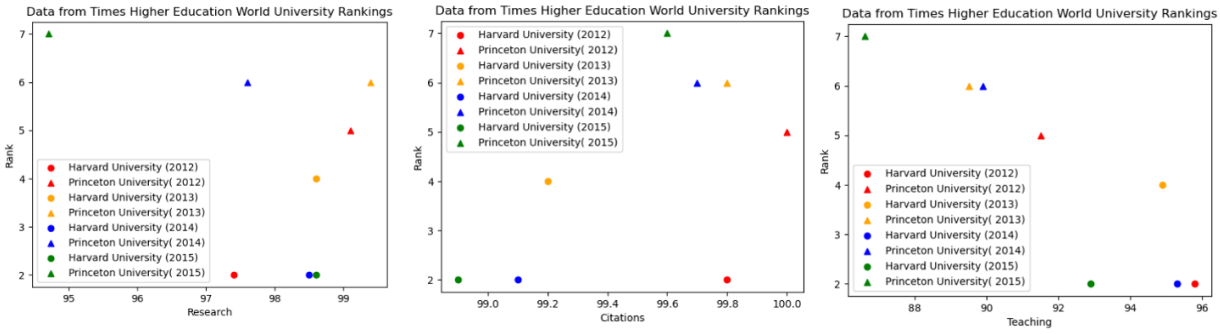

Figure 6: Comparison of the Times Dataset showing Citations, Research, and Teaching Ranking compared to the ranking of the university.

In Figure 7 the data presented shows a clear correlation between the university ranking of the Shanghai Dataset and the parameters of Staff Awards, Highly Cited Research, and Publications in Nature and Science Journals. Harvard University received the highest ranking of all universities on the dataset. Meanwhile, Princeton University had an overall ranking around 6-7 for this dataset. Harvard university scored a perfect score of 100 for Staff Awards, Highly Cited Research, and Publications in Nature and Science Journals. While Princeton University scored 88-94 for staff awards, lower than 50 for Highly Cited Research, and from 60 to 65 in publications in Nature and Science Journals. Overall, we see that these factors have significant influence on the rankings of these universities in the Shanghai dataset.

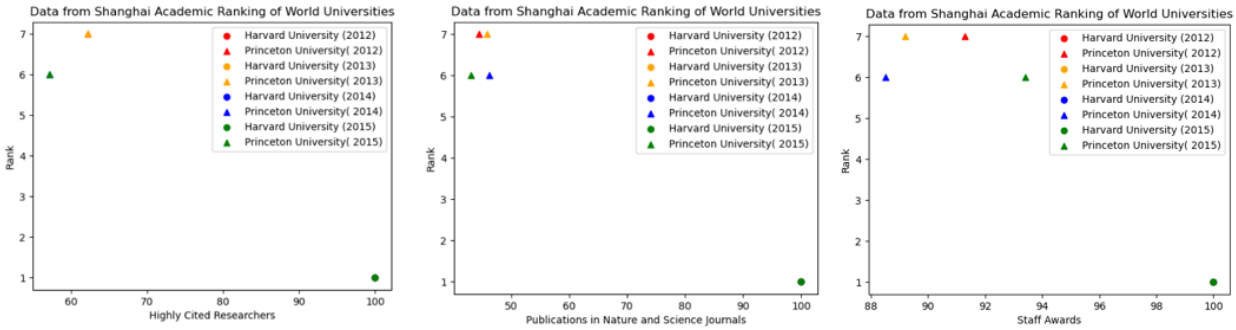

Figure 7: Comparison of the Shanghai Dataset showing Staff Awards, Highly Cited Research, and Publications in Nature and Science Journals compared to the ranking of the university.

2. Cornell University vs University of Pennsylvania

This second case study presents a comparison of Cornell University and University of Pennsylvania. The ranking Figure 8 and scoring Figure 9 shows that the University of Pennsylvania outranks (and outscores) Cornell University in the times dataset. In contrast, Cornell University outranks (and outscores) the University of Pennsylvania in the Shanghai dataset. This can be used to explore the factors that affect the world-wide ranking of universities of both datasets.

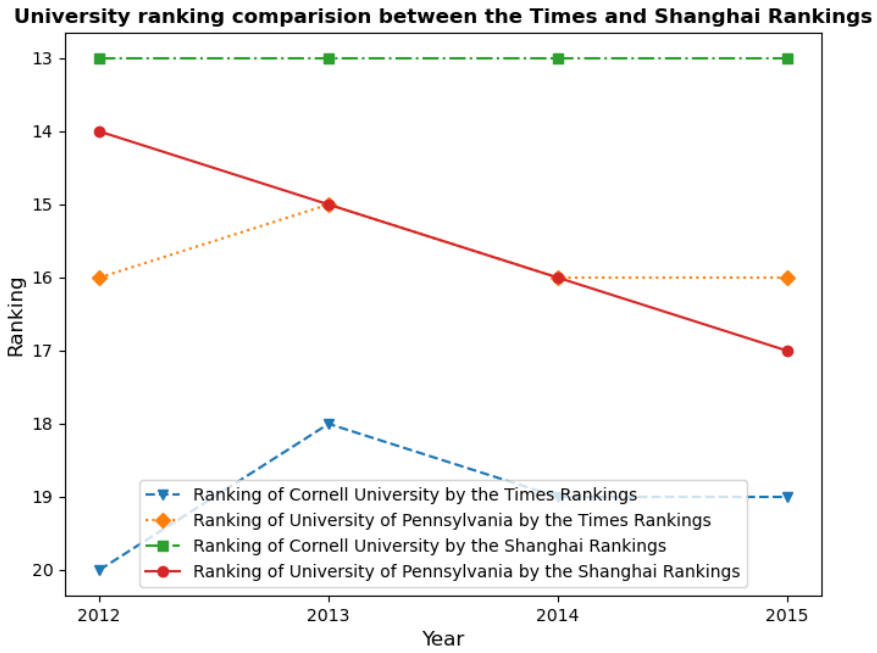

Figure 8: University Ranking Comparison Between the Times and Shanghai Rankings of Cornell University and University of Pennsylvania

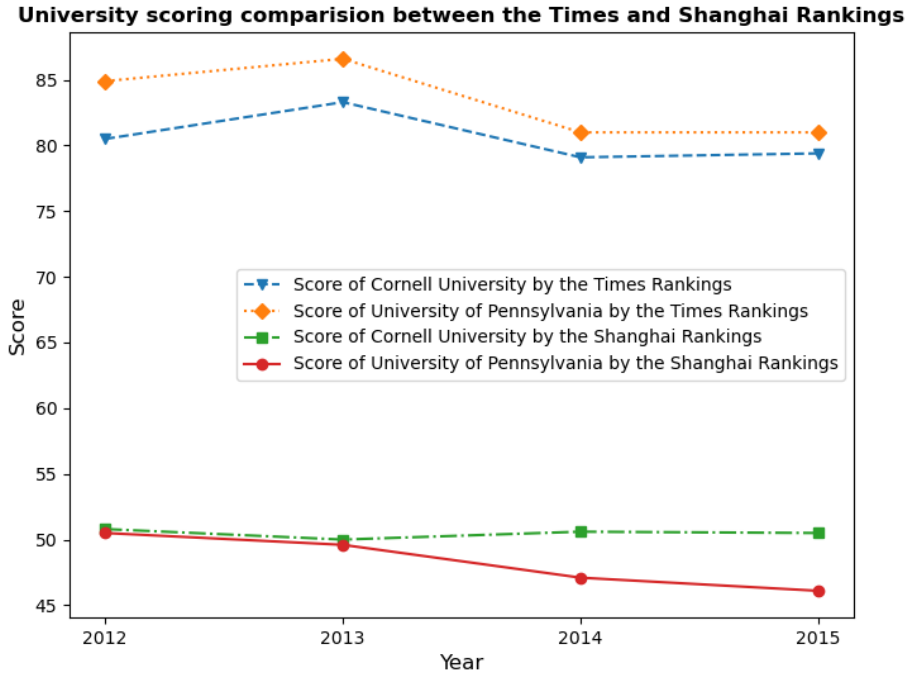

Figure 9: University Scoring Comparison Between the Times and Shanghai Rankings of Cornell University and University of Pennsylvania

It is established that the times dataset favored University of Pennsylvania over Cornell University in ranking and score. The University of Pennsylvania's score exceeds Cornell University's score in the factors of Citations and Teaching while not falling far behind in the Research Score seen in Figure 10. This establishes that Citations and Teaching are important to the ranking of university in the Times Dataset.

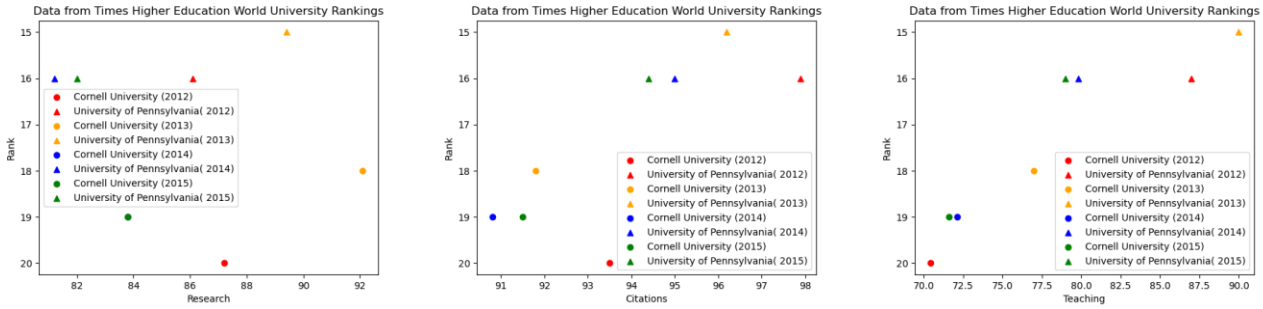

Figure 10: Comparison of the Times Dataset showing Citations, Research, and Teaching Ranking compared to the ranking of the university.

In Figure 8 Cornell University consistently ranks higher than the University of Pennsylvania in the Shanghai Dataset. In Figure 11 we see that Cornell University ranks higher in Staff Awards consistently and ranks higher more often in Publications in Nature and Science Journals and Highly Cited Research. These factors attribute to the overall higher ranking of the Shanghai Dataset.

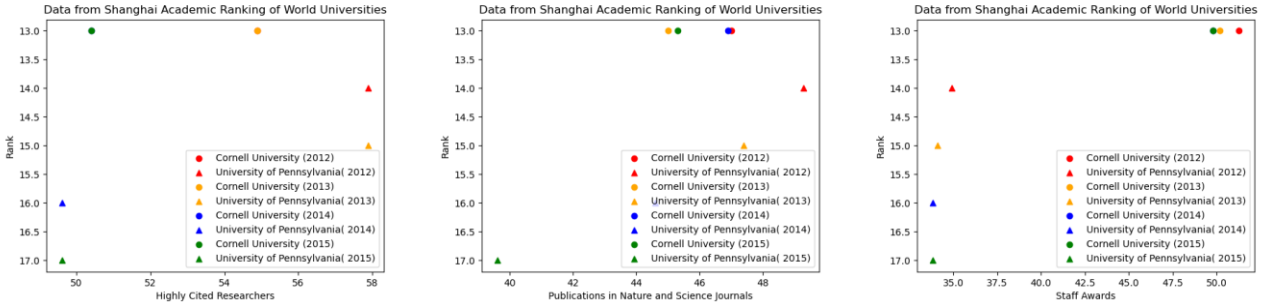

Figure 11: Comparison of the Shanghai Dataset showing Staff Awards, Highly Cited Research, and Publications in Nature and Science Journals compared to the ranking of the university.

Interpretation and Conclusion:

The SARWU and THEUR datasets compare the world's universities by rankings and creating a quantitative score attributed to the university. As an answer to the first question of this project, key factors were identified that have a good influence in both ranking and scoring of a dataset. The THEUR dataset tends to appropriate a higher weight in the ranking and scoring of a university on the features of accounting for Citations, Research, and Teaching Ranking. The SARWU dataset on the other hand focuses on Staff Awards, Highly Cited Research, and Publications in Nature and Science Journals. Both ranking datasets methodologies have shown citations and publications as

key factors that affect the ranking and scoring of the university. This in turn answers our second proposed question regarding the impact of the publications and citations in the ranking and scoring of a university. This study has limitations of only using two datasets as representatives of the worldwide university rankings and other factors such as lack of consideration of the financing of research in the respective universities, which was not present in either dataset, or could have had a correlation in the rankings. In addition to this, it's important to note that the methodologies in both datasets differ significantly in multiple areas and may have some bias involved in their rankings which makes them relatively unreliable in some capacity. Yet, this can't be confirmed and just be deemed as speculation.

References:

1. Wahid, Nazia & Warraich, Nosheen & Tahira, Muzammil. (2021). Factors influencing scholarly publication productivity: a systematic review. Information Discovery and Delivery. 10.1108/IDD-04-2020-0036].
2. Moghdam, Masoomeh & Salehi, Hadi & Ale Ebrahim, Nader & Mohammadjafari, Marjan & Gholizadeh, Hossein. (2015). Effective Factors for Increasing University Publication and Citation Rate. Asian Social Science. 11. 10.5539/ass.v11n16p338.