University Ranking Analysis:

Comparing and Identifying Biases Across Different University Ranking Systems

| Bingjie Chen  University of Washington  [cbjmandy@uw.edu](mailto:cbjmandy@uw.edu) | **Nelson D’souza**  University of Washington  [nelsonds@uw.edu](mailto:nelsonds@uw.edu) | Maria George  University of Washington  [gmaria@uw.edu](mailto:gmaria@uw.edu) | Cory Shyu  University of Washington  [coryshyu@uw.edu](mailto:coryshyu@uw.edu) |
| --- | --- | --- | --- |

# ABSTRACT

## Ranking is a key decision-making factor for students in their university selection process. Ranking methodology and criteria can vary significantly across different ranking systems and it can be difficult to understand and compare various ranking systems. Our study focuses on comparing university rankings by three reputable ranking systems and identifying potential biases among them. There are three major findings in our study. Firstly, volatility in the Times ranking system is consistently much higher than others, which may indicate systematic biases. Secondly, difference in ranking between certain ranking systems is highly statistically significant, which may be driven by biases in certain ranking criteria. Thirdly, we implemented our clustering model to visualize university clusters based on difference in university ranking and found evidence for potential biases in ranking with respect to certain countries across all three ranking systems. We also found that top 40 universities tend to have the least bias in terms of magnitude as well as the number of inconsistencies, and universities beyond top 40 are subject to much larger biases.

## Author Keywords

University ranking, volatility in university ranks, factor bias, K-Means clustering

# INTRODUCTION

University ranking has always been an important issue for schools, students, industry, and academia. This can be observed from the increasing number of annual rankings published (Aguillo et al, 2010) as well as the number of research on the topic.  Rankings are used by the administrations to receive funding, which can be used to improve their research performance, and further attract talents. University ranking also plays a key role in students’ application and enrollment decision-making process (Claassen, 2015).

While there is little dispute on the importance of university ranking, there is no consensus on the ranking criteria and methodology across different ranking systems. Thus, it is difficult to analyze and compare university rankings across different ranking systems.

Through this research study, we intend to analyze potential biases in the rankings published by the following different ranking systems: Times Higher Education World ranking system (“Times”), Academic Ranking of World Universities, also known as the Shanghai ranking (“Shanghai”), and the Center for World University Rankings (“CWUR”). We performed volatility analysis to identify high variation in ranking within each ranking system between 2012 and 2015. Additionally, we compared similar ranking criteria between different ranking systems to identify potential biases in the ranking input factors. Lastly, we implemented similarity measures and performed clustering analysis to identify and visualize favoritism toward certain universities as well as universities in certain countries.

# Relevant works

The study by Huang (2011) compared the university rankings by three different ranking systems. The research method identified the common universities among all the ranking lists and took into account the normalized difference in rankings. The researchers then devised a new mathematical approach, similarity measure, for comparison. The study by Aguillo, Ilan, Levene and Ortega (2010), compares the annual rankings of five different systems using three complementary measures, the size of the overlap, the Speraman’s footrule, and the M measure.

In the recent study by Claassen (2015), Bayesian hierarchical latent trait model was used to measure university quality. By measuring the accuracy of each ranking systems as well as degree of bias towards universities in particular countries, the model developed in the study computes a common ranking system. Furthermore, the study by Soo & Dill (2005) pointed out that it is necessary to have an international census for the cross-national university ranking, which can minimize the cultural and political bias and help students as well as policy makers make decisions wisely.

The similarity measure computed in this research study is inspired from the study by Huang (2011) and Aguillo et al (2010). However, we have combined the idea of difference in rankings and number of inconsistencies in the impartiality score. Additionally, we have performed volatility analysis and examined differences in similar ranking criteria between different systems.

# Data

The university ranking dataset was obtained from Kaggle[[1]](#footnote-1). There are three main datasets:

1. Times: The dataset contains world university rankings by Times Higher Education. Founded in the UK in 2010, it is considered to be one of the most reputable ranking institution. It has the top 200 university rankings from 2011, the top 400 rankings from 2012 to 2015 as well as the top 800 rankings from 2016.
2. Shanghai: This dataset contains world university rankings by ARWU, also known as the Shanghai Rankings. Founded in China in 2003, it has the top 500 rankings from 2005 to 2015.
3. CWUR: The CWUR ranking comes from Saudi Arabia, founded in 2012 and it contains the top 100 rankings from 2012 and 2013 as well as stop 1000 rankings from 2014 and 2015.

For further detail in ranking criteria and methodology, please refer to Appendix I.

These particular data sets required significant effort in data wrangling. There were two major challenges: matching university names across the three ranking systems and handling missing ranking information. In many cases, the name of the same university was recorded in three different ways across different universities. For instance, Texas A&M University was shown as Texas A&M University, College Station in one ranking system, Texas A&M university in another system and Texas A & M university (note, space between A & M) in another system. Moreover, the same university was recorded differently within a ranking system as well because of the prevalence of special characters (e.g. University of Michigan, Ann Arbor, University of Wisconsin-Madison). To ensure that no duplicate university names were used, we assigned a unique identification to each unique university for all three ranking systems. In order to compare rankings across three systems, we imputed ranks when an exact rank was missing. We used the midpoint when given a range for both Shanghai and Times. For universities that were present in one ranking system but absent in another, for the purpose of distance calculation we assigned a ranking of 700. Please refer to the discussion in the distance clustering section.

# volatility analysis

Volatility in ranking can be an indicator of potential bias. Therefore, we analyzed the volatility in each ranking system for all top 100 universities between 2012 and 2015. The variance in ranking for a ranking system was computed as the average variance of all the top 100 universities in the ranking system. This required the assumption that rank of one university was independent from all other universities. While a change in ranking for one university could necessarily impact ranking of some other universities, the criteria used to compute the ranking were independent from one institution to another. Ultimately it is the criteria that drive changes in rankings. Therefore, we used the independent assumption and ignore the interaction terms between universities. Table 1 summarizes the volatility for each ranking system.

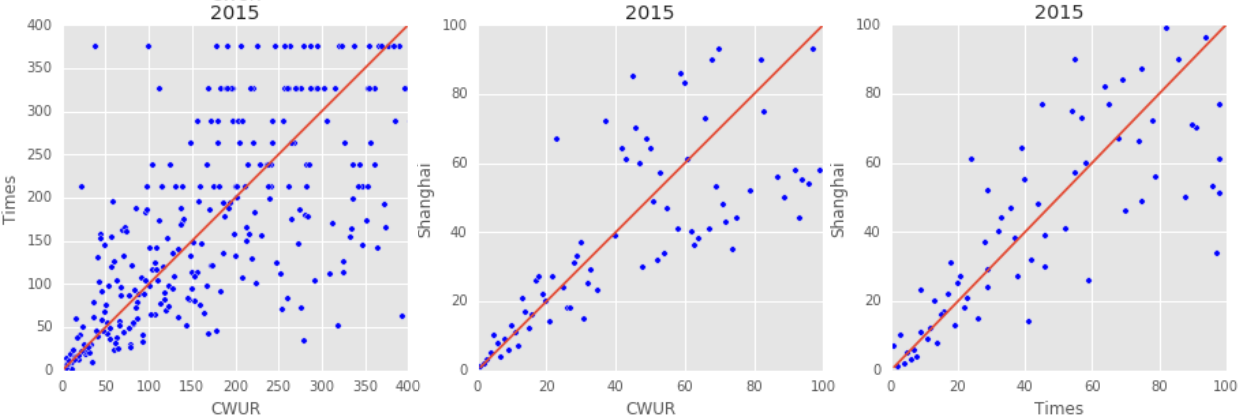
**Table 1. Volatility in Ranks**

| **Ranking System** | **Volatility** |
| --- | --- |
| Times | 20.178 |
| CWUR | 11.900 |
| Shanghai | 6.926 |

University ranks in Shanghai are clearly more stable over time relative to the other two, whereas Times shows high volatility in ranks among the three and CWUR is somewhere in the middle. Volatility is defined as the standard deviation in ranking. Therefore, we expect to see more than 30% of the top 100 universities change by more than 20 ranks from one year to another in the Times ranking system, versus 7 in Shanghai, and 12 in CWUR.

One possible explanation for difference is that Shanghai assigns more weight to criteria that are more stable over time. For instance, 30% weight is assigned to number of alumni and staff winning Nobel Prizes and Fields Medals. In contrast, CWUR includes “quality of education” which represents faculty members who have won major international awards, prizes, and medals. The range of relevant awards is broader compared to Shanghai’s criteria. Additionally, “quality of education” for CWUR is adjusted to the university's size, which allows the accomplishment of outstanding faculties in small-sized schools to have a significant influence on their university ranking.

Moreover, the “research” criterion focuses on natural and social science (Saisana, d’Hombres, Saltelli, 2010). Research in these fields can take much longer to produce meaningful publications. Therefore, we expect low throughput rate in scientific research, which leads to less ranking volatility for Shanghai. However, both CWUR and Times give more weight to an institution’s influence in industry, such as number of alumni assuming executive positions in top companies. These factors change relatively more frequently.



**Figure 1: Pairwise Ranking**

# pairwise ranking comparison

Figure 1 shows the pairwise ranking scatterplots for all 3 pairs of ranking systems in 2015. The pattern is fairly similar across all years between 2012 and 2015. For universities that were ranked within the top 40, variations in ranking tend to be much smaller. As ranking increases, the variation in rankings widens.

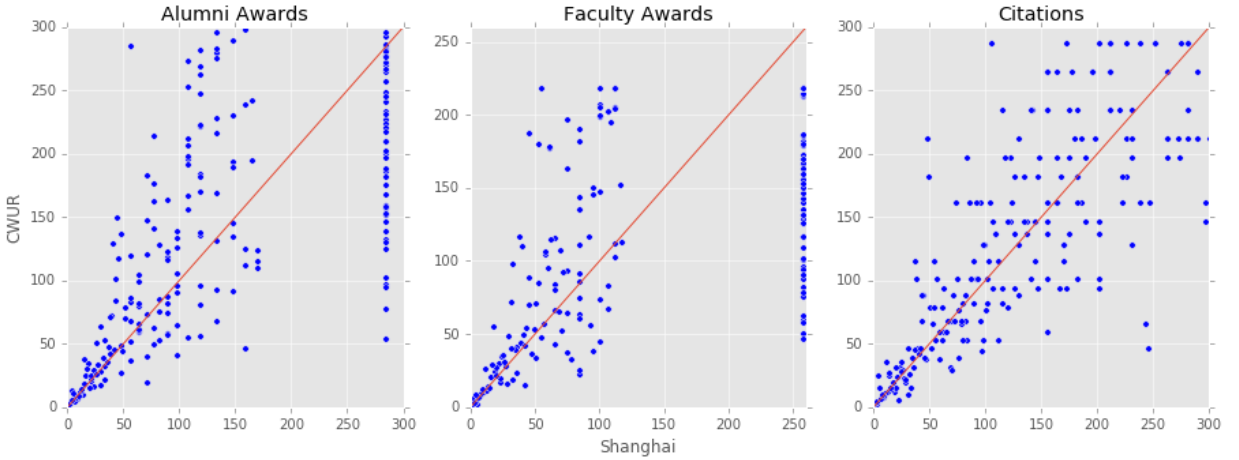
Table 2: P-values of pairwise T-Test for university rankings beyond top 40

|  | CWUR/  Times | CWUR/  Shanghai | Times/  CWUR |
| --- | --- | --- | --- |
| 2013 | 0.0237 | 0.4515 | 0.4702 |
| 2014 | 0.0038 | 0.1663 | 0.0000 |
| 2015 | 0.0000 | 0.0035 | 0.1768 |
| 2016 | 0.0000 | 0.0056 | 0.2407 |

Table 2 shows the p-values associated with the T-test for difference in ranking beyond top 40 between each pair. None

of the pairs has statistically significant difference for the top 40 universities. The difference in ranks between CWUR and Time has mostly been statistically significant while the difference in ranking between Shanghai and Times has not. However, given that some of the ranks were imputed using the midpoint the T-test is likely to underestimate the statistical significance of the difference.

To explain the possible reasons for the above observation, we compared similar ranking criteria (alumni award, faculty award, and citation) between Shanghai and CWUR in 2015 as shown in Figure 2. While these criteria are based on measures such as the number of faculty and alumni winning major awards, and the number of citations of important publications, there is a surprisingly significant divergence between the two systems. The seemingly objective measures also show similar patterns as the ranking variable. The T-test for each of the factor has a corresponding p-value of close to 0, which that differences between the two systems on similar ranking criteria are statistically significant. Biases in ranking criteria may be the main driver for differences in ranking we have seen earlier.

**Figure 2: Input Factor Comparison Between Shanghai and CWUR**

# Clustering Analysis

We have seen in the previously section that there is statistically significant difference between some of the ranking systems. The next step is to develop a metric to visualize the distance between each ranking. We chose the Manhattan distance as the distance measure, which is the absolute value of the distance. While there are multiple options to measure distance such as Euclidean, L-Norm (multi-dimensional), and Chebyshev, we chose the Manhattan measure because it is more intuitive in the context of ranking.

There is a total of 142 universities in the top 100 list of any of the 3 ranking systems. 20 universities within the top 100 list were not common to at least 2 of the ranking systems. For these universities, we imputed a dummy rank of 700. This would ensure that the ones which were not common to the ranking systems would form a cluster. Therefore, the distance between the universities in the two ranking systems would have to be at least twice as large as the largest difference of 316. The distribution of distance between each ranking system pair over 2014 and 2015 is shown in Appendix II.

All of the pairwise difference in ranking for 2014 and 2015 are almost similarly distributed. As shown in the histograms, the majority of the difference is between 0 and 100, while there is fat tail on the right indicating extreme differences.

Given the above distribution, we hypothesized that 50 would be a good threshold for classification of bias. In other words,

the universities which had a rank difference of greater than

50 would be considered to be biased (either towards or against) by a ranking system.

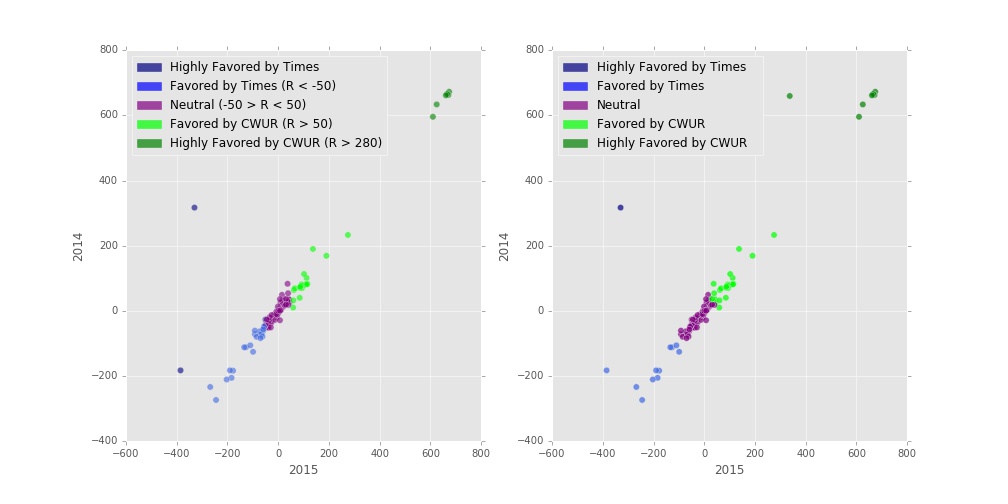
We classified universities into four groups based on the degree of inconsistency in ranks among the three ranking systems, as described below:

1. **Consistent** - This category represents universities which have consistent ranking across all the 3 ranking systems and none of the pairs shows significant difference.
2. **Somewhat consistent** - This category represents universities which have consistent rankings across 2 pairs of rankings but not for the third one.
3. **Somewhat inconsistent** - This category represents universities which have inconsistent rankings across 2 pairs of rankings, but not for the third pair.
4. **Inconsistent** - This category represents universities which have inconsistent ranking across all the 3 ranking system pairs.

Before categorizing the universities using a clustering algorithm, we needed a validation cluster to test the algorithm and verify our hypothesis for the difference threshold. We performed validation on the Times-CWUR ranking system pair. In order to create validation clusters, we considered the difference in ranking between each ranking system pair. For each pair there were 5 possible clusters. The first cluster having universities for which the absolute difference in ranking was below the threshold of 50. The second and third clusters contained universities for which the difference in the rankings exceeded or fell below the threshold of +50 and +316 or -50 and -316. Ideally we would only have these 3 clusters. However the impact of extreme distance due to absence of university in a ranking system influenced the clusters significantly. Therefore, we introduced two more clusters for universities missing in either ranking system pair by considering rank difference of greater than 316 or less than -316. All the validation clusters except the one which had universities in the range -50 to +50 were considered as biased.

Our logical model is derived using a semi-supervised learning method because we know what our clusters should be but we do not know the optimal threshold for each of the clusters. To cluster universities based on an optimal threshold, we applied the K-Means algorithm for the Times-CWUR ranking system pair by keeping K=5. We performed the analysis on both 2014 and 2015 data in order to control for volatility in a University ranking over time. As shown in Figure 3, our logical cluster is fairly similar to the K-Means algorithm. In order to check the accuracy of our logical model, we computed a confusion matrix:

We found via the confusion matrix that the K-Means algorithm had **82.23%** accuracy. The data in both cases were clustered into 5 groups. Firstly, we had a neutral cluster having universities which did not exhibit a high difference in ranking for both systems. We had two other clusters which had universities favored by either Times or CWUR and finally two more clusters which had universities missing in either of the ranking system pairs.

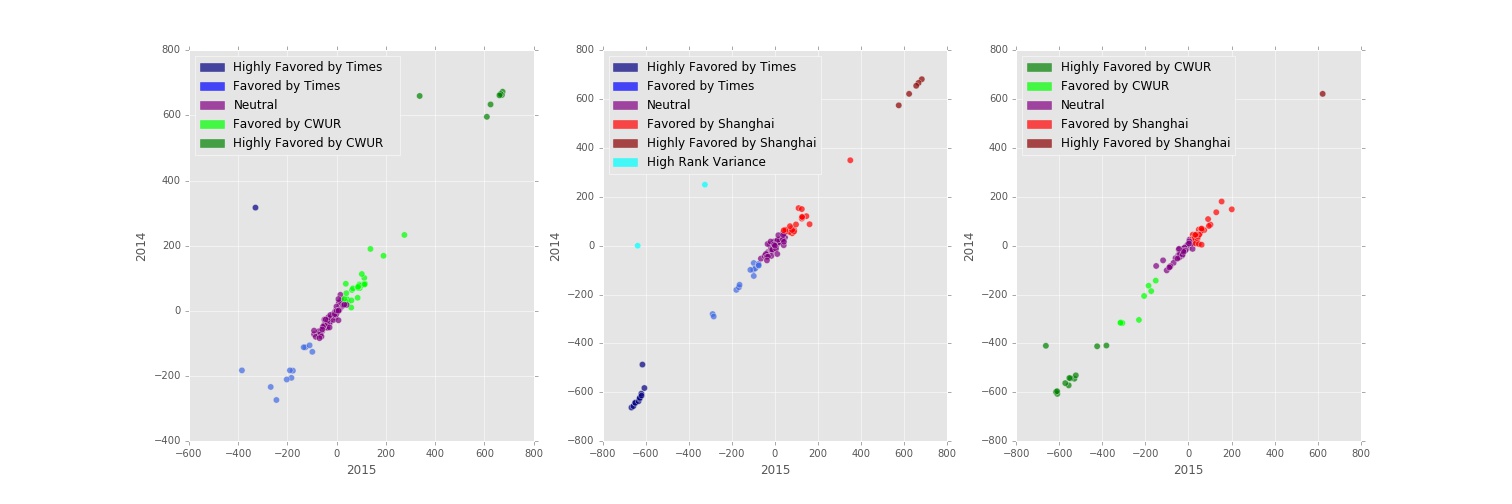


**Figure 3: Validation Model and K-Means Clustering Algorithm Comparison**

Having validated our clustering model using the K-Means clustering algorithm, we implemented the methodology for all 3 pairs, as shown in Figure 4. These plots help visualize the count of universities as well as the magnitude of inconsistency between any two ranking systems. Universities at the diagonally far end are highly inconsistent with respect to the other ranking system, whereas the ones nearer to the center are fairly consistent.

After clustering, we assigned binary values to each of the 5 clusters present in each pairwise university ranking. A value of 1 denoted a university cluster which was consistent while a value of 0 denoted otherwise.

We summed the binary values for each university to calculate the Impartiality Score as given in Table 3. These scores correspond to the categories we had defined above wherein 3 signifies score of group of universities with consistent ranking and 0 signifies score of group of universities with inconsistent ranking. For visualization of university clusters, please refer to Appendix III.



**Figure 4: Pairwise Clustering**

**Table 3. Impartiality Scores**

| **Impartiality Score** | **Number of Universities** |
| --- | --- |
| 3 | 46 |
| 2 | 45 |
| 1 | 31 |
| 0 | 20 |

# Country-specific biases

To explore other sources of biases, we also analyzed for systematic biases against or towards universities by country using the same threshold of 50 as the threshold for bias. The results are shown in Appendix IV.

Times exhibited high bias towards universities from the UK compared to the other two ranking systems. This was not surprising given that Times is based in the UK. Surprisingly for Shanghai, a ranking system from China exhibited high bias against universities from China as well as South Korea, a neighboring country. CWUR, a ranking system from Saudi Arabia exhibited very high bias against universities from Europe, Australia and North America, but none from Asian countries.

# Conclusion

Based on our exploratory analyses, we concluded that there were significant biases among the three ranking systems. In the volatility analysis, we found that university ranks in Times were almost 3 times as volatile as that of Shanghai and ranks in CWUR were almost 2 times as volatile. While we attempted to explain the difference based on each system’s ranking criteria, the volatility should have been theoretically similar, which was not the case.

We also examined differences in similar ranking criteria between Shanghai and CWUR and found that the differences are statistically significant. This signifies that variations in ranking criteria may be a driver for difference in ranks.

Based on theManhattan distance measure, we derived a clustering model for each university to visualize the similarity and differences in ranking across the three ranking systems. We also used the Impartiality Score to visualize the number of instances in which the 3 ranking system pairs are consistent. Through the visualization of university clusters, we reaffirmed that universities within the top 40 were generally more consistent, and as the ranking increased, the number as well as the magnitude of inconsistency increased.

Finally, we explored how each ranking system either favored or disapproved of universities from a particular country. We found that all three ranking systems exhibited strong bias with respect to different set of countries.

# Limitations and Future Work

We clustered top 100 universities from the 3 ranking systems, however, we could not fully utilize these clusters due to absence of common features across the 3 rankings. Country was the only common feature available and because of this, the scope of the bias analysis was rather limited.

The next steps for this study are increasing the number of common features and also incorporating a greater number of ranking systems. From Wikipedia, we could scrape a host of additional features such as university type (public/private), number of undergraduate students, founding year of university, etc. We can incorporate multiple dimensions in the distance calculation and moreover, identify which critical features drive the differences in ranking across different system pairs. By considering a greater number of ranking systems, we can validate or adjust our model accordingly. Finally, with a larger number of features, we can conduct linear regression analysis to identity the key features driving difference in ranks across different ranking systems.

We have made our findings available on our website[[2]](#footnote-2).  The website also contains interactive visualizations to explore our results in further detail. We hope people find it useful, especially those in the critical phase of school search process.

**REFERENCES**

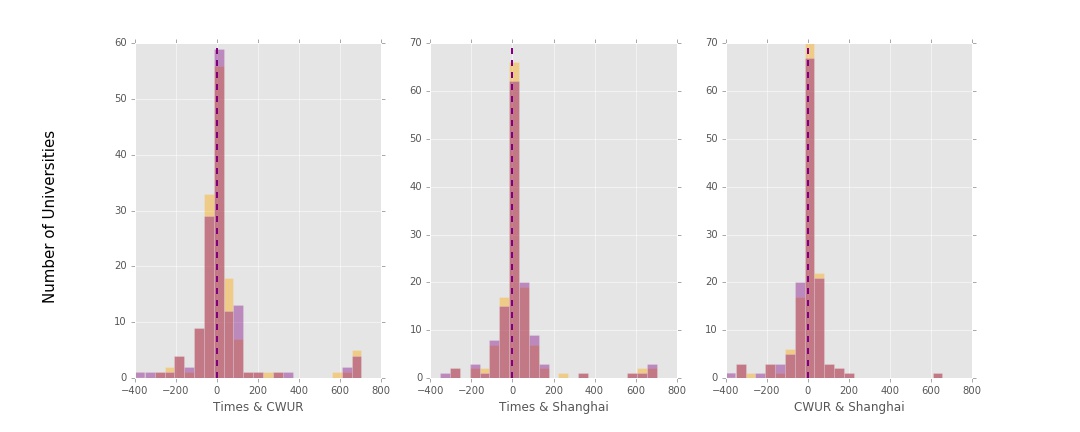
1. Aguillo, I. F., Bar-llan, J., Levene, M., & Ortega, J. L. (2009, July 17). Comparing University Rankings. <https://www.researchgate.net/publication/220365497_Comparing_university_rankings>
2. Claassen, C. (2015). Research Repository. <http://repository.essex.ac.uk/14726/>
3. Huang, M. (2011, June). A Comparison of Three Major Academic Rankings for World Universities: From a Research Evaluation Perspective. <http://jlis.lis.ntu.edu.tw/article/v9-1-1.pdf>
4. Soo, M. & Dill, D.D. (2005). Academic quality, league tables, and public policy: A cross-national analysis of university ranking systems. Higher Education. 49: 495- 533.
5. Saisana, M., d’Hombres, B., & Saltelli, A. (2010) Rickety numbers: Volatility of University rankings and policy. Research Policy. 40. 165 - 177. Elsevier B.V.

**Appendix**

1. **Criteria and methodologies**

|  |  |  |  |
| --- | --- | --- | --- |
| Ranking System | Ranking Criteria | Weight | Description |
| Times | Teaching | 30% | University score for teaching (the learning environment) |
|  | Research | 30% | University score for research (volume, income and reputation) |
|  | Citations  International | 30%  7.5% | University score for citations (research influence)  University score international outlook (staff, students, research) |
|  | Income | 2.5% | University score for industry income (knowledge transfer) |
| Shanghai | Award | 20% | Staff of an institution winning Nobel Prizes and Fields Medals |
|  | HiCi | 20% | Highly cited researchers in 21 broad subject categories |
|  | N&S | 20% | Papers published in Nature and Science |
|  | PUB  Alumni | 20%  10% | Papers indexed in Science Citation Index-expanded and Social Science  Citation Index  Alumni of an institution winning Nobel Prizes and Fields Medals |
|  | PCP | 10% | Per capita academic performance of an institution |
| CWUR | Quality of Education | 25% | The number of a university's alumni who have won major international awards, prizes, and medals relative to the university's size |
|  | Alumni Employment | 25% | The number of a university's alumni who have held CEO positions at the world's top companies relative to the university's size |
|  | Quality of Faculty | 25% | The number of academics who have won major international awards, prizes, and medals |
|  | Publications | 5% | The number of research papers appearing in reputable journals |
|  | Influence | 5% | The number of research papers appearing in highly-influential journals |
|  | Citations | 5% | The number of highly-cited research papers |
|  | Broad Impact | 5% | The university's h-index |
|  | Patents | 5% | The number of international patent filings |

1. **Distribution of ranking difference in 2014 (Orange) and 2015 (Purple)**



1. **University Clusters**



1. **Biases against or for specific countries**



1. **Regression Results**

**CWUR**

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | **world\_rank** | **R-squared:** | **0.919** |
| **Model:** | **OLS** | **Adj. R-squared:** | **0.918** |
| **Method:** | **Least Squares** | **F-statistic:** | **3530.** |
| **Date:** | **Sat, 21 May 2016** | **Prob (F-statistic):** | **0.00** |
| **Time:** | **11:00:22** | **Log-Likelihood:** | **-12943.** |
| **No. Observations:** | **2200** | **AIC:** | **2.590e+04** |
| **Df Residuals:** | **2192** | **BIC:** | **2.595e+04** |
| **Df Model:** | **7** |  |  |
| **Covariance Type:** | **nonrobust** |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[95.0% Conf. Int.]** |
| **Intercept** | **-52.6402** | **5.670** | **-9.284** | **0.000** | **-63.759 -41.522** |
| **quality\_of\_education** | **0.0961** | **0.027** | **3.603** | **0.000** | **0.044 0.148** |
| **alumni\_employment** | **0.2630** | **0.013** | **19.503** | **0.000** | **0.237 0.289** |
| **quality\_of\_faculty** | **-0.1890** | **0.050** | **-3.746** | **0.000** | **-0.288 -0.090** |
| **publications** | **0.4223** | **0.014** | **29.545** | **0.000** | **0.394 0.450** |
| **influence** | **0.2857** | **0.014** | **19.858** | **0.000** | **0.257 0.314** |
| **citations** | **0.1494** | **0.014** | **10.365** | **0.000** | **0.121 0.178** |
| **patents** | **0.0886** | **0.010** | **9.283** | **0.000** | **0.070 0.107** |

**Times**

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | **world\_rank** | **R-squared:** | **0.841** |
| **Model:** | **OLS** | **Adj. R-squared:** | **0.840** |
| **Method:** | **Least Squares** | **F-statistic:** | **1134.** |
| **Date:** | **Mon, 23 May 2016** | **Prob (F-statistic):** | **0.00** |
| **Time:** | **19:13:31** | **Log-Likelihood:** | **-4900.6** |
| **No. Observations:** | **1078** | **AIC:** | **9813.** |
| **Df Residuals:** | **1072** | **BIC:** | **9843.** |
| **Df Model:** | **5** |  |  |
| **Covariance Type:** | **nonrobust** |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[95.0% Conf. Int.]** |
| **Intercept** | **349.2175** | **4.746** | **73.580** | **0.000** | **339.905 358.530** |
| **teaching** | **-0.8902** | **0.097** | **-9.211** | **0.000** | **-1.080 -0.701** |
| **international** | **-0.2192** | **0.034** | **-6.496** | **0.000** | **-0.285 -0.153** |
| **research** | **-1.4889** | **0.082** | **-18.187** | **0.000** | **-1.650 -1.328** |
| **citations** | **-1.3729** | **0.048** | **-28.699** | **0.000** | **-1.467 -1.279** |
| **income** | **-0.1291** | **0.034** | **-3.751** | **0.000** | **-0.197 -0.062** |

**Shanghai**

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | **world\_rank** | **R-squared:** | **0.752** |
| **Model:** | **OLS** | **Adj. R-squared:** | **0.748** |
| **Method:** | **Least Squares** | **F-statistic:** | **197.3** |
| **Date:** | **Sat, 21 May 2016** | **Prob (F-statistic):** | **5.48e-115** |
| **Time:** | **11:00:24** | **Log-Likelihood:** | **-1620.6** |
| **No. Observations:** | **398** | **AIC:** | **3255.** |
| **Df Residuals:** | **391** | **BIC:** | **3283.** |
| **Df Model:** | **6** |  |  |
| **Covariance Type:** | **nonrobust** |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[95.0% Conf. Int.]** |
| **Intercept** | **117.4401** | **4.146** | **28.329** | **0.000** | **109.290 125.590** |
| **alumni** | **0.0570** | **0.071** | **0.805** | **0.421** | **-0.082 0.196** |
| **award** | **-0.5346** | **0.062** | **-8.668** | **0.000** | **-0.656 -0.413** |
| **hici** | **-0.2979** | **0.101** | **-2.963** | **0.003** | **-0.496 -0.100** |
| **ns** | **-0.6262** | **0.124** | **-5.056** | **0.000** | **-0.870 -0.383** |
| **pub** | **-0.5290** | **0.091** | **-5.829** | **0.000** | **-0.707 -0.351** |
| **pcp** | **0.1750** | **0.091** | **1.927** | **0.055** | **-0.004 0.353** |

1. <https://www.kaggle.com/mylesoneill/world-university-rankings> [↑](#footnote-ref-1)
2. <http://univrankinganalysis.github.io/> [↑](#footnote-ref-2)