

World Rankings for Universities

James Tsai – Spring 2016, MSDS6372 – Experimental Statistics II

Introduction

From sports to education, there is a fascination with rankings. There is a desire for one's favorite team to be the best, and for every parent to want their child to attend the very best college or university. While sports teams compete on the playing field to determine which team is better, it is not so clear-cut when it comes to ranking universities. The size and purpose of institutions and the methodologies used to rank them is a hotly debated issue. Unfortunately, despite the subjectivity of these methodologies, the appeal of university ranking publications has become more influential across the globe as students have become ever more mobile and are looking far beyond their borders. For universities, having a good reputation translates to money in the form of more applications, more tuition dollars, and greater levels of alumni giving.

Problem Statement

The purpose of the study is to take one of the most influential and observed university measures, The Times Higher Education World University Ranking for 2016, and determine whether any of the nine available variables has the greatest influence on the total score. Understanding these variables may also give us some insight into the nature of the ranking methodology and whether specific criticisms levied against the ranking system are justified.

Constraints and Limitations

Since the scope of this observational study is for The Times Higher Education World University Ranking for 2016, no casual inference can be made between the explanatory and response variables. It is possible that there are more important measures, which better explain the overall ranking of the universities, which were not published.

Furthermore, we must bear in mind that the methodologies and weightings for each criterion used by The Times Higher Education World Ranking maybe adjusted from year-to-year. We will weight the explanatory variables equally as we are investigating the influence of all variables equally. Finally, to get another perspective of this study, we note that seven schools have consistently taken one of top seven spots under this ranking methodology from 2011-2016. They are: California Institute of Technology, University of Oxford, Stanford University, University of Cambridge, Massachusetts Institute of Technology, Harvard University, and Princeton University. Such a small group of 'winner take all' in the ranking results may lead one to question the ranking methodology for the top universities as it may not have practical significance to the student applying for admission.

Data Set Description

The analysis is based on publicly available dataset, available from the Kaggle website and also from The Times Higher Education World University Rankings website. Please refer to:

<https://www.kaggle.com/mylesoneill/world-university-rankings>

<https://www.timeshighereducation.com/world-university-rankings/2016/world-ranking#!/page/0/length/25>

In Figure 1, the data for the top 200 universities was sorted by the response variable total score, which determines the overall ranking of the university.

Variable	Usage	Description	Type	Range
University	Identifier	University name	String	N/A
Total Score	Response	Total score, used to determine rank	Decimal	0 to 100
Teaching	Explanatory	University score for teaching (the learning environment)	Decimal	0 to 100
International	Explanatory	University score for international outlook (staff, students, research); ability to attract undergrads, postgraduates, and faculty from all over the planet is key to it's success on the world stage	Decimal	0 to 100
Research	Explanatory	University score for research (volume, income, reputation)	Decimal	0 to 100
Citations	Explanatory	University score for citations (research influence); indicator at the role in spreading new knowledge and ideas	Decimal	0 to 100
Income	Explanatory	University score for industry income (knowledge transfer); this category seeks to capture the ability to help industry with innovations, inventions, and consultancy	Decimal	0 to 100
# Of Students	Explanatory	Total Number of Students; the number of full time equivalent students at the University	Integer	>0
Student/Staff	Explanatory	# of Students to Staff; ratio of full time equivalent students to the number of academic staff, those involved in teaching or research	Decimal	>0
International	Explanatory	Percentage of International Students; students originating from outside of the country of the University	Percent	0 to 100
Female/Male	Explanatory	# of Females to # of Males	Decimal	>0

Figure 1. List of all variables.

The next five variables, Teaching, International, Research, Citations, and Income were used by The Times Higher Education World University Rankings to determine total score. However, after some research on the methodology, we note that the non-equal weighting used in 2016 by The Times Higher Education World University Rankings to determine total score as shown in Figure 2.

Category	Weighting
Teaching	30% of overall score
International	7.5% of overall score
Research	30% of overall score
Citations	30% of overall score
Income	2.5% of overall score

Figure 2. Weighting for each category.

The next four variables called 'Key Statistics' are: # of Students, Student/Staff, International, and Female/Male, and are not part of the criterion used in the calculation of the total score. The

Exploratory Data Analysis and Screening

Examining the summary statistics, we notice that the ranges of values for explanatory variables `num_students` and `female_male_ratio` differ by several factors, and thus we must standardize the data before conducting the PCA analysis to determine which variables have the highest influence.

Finally, we examine the Pearson correlation matrix to examine if the variables have high correlation with each other. From a high-level, PCA is about identifying variables that describe a similar construct; we can investigate the correlation matrix to get a glimpse of which explanatory variables maybe related. Judging from Figure 6, we find that there is a high correlation between income/num_students and international/student_staff_ratio.

Summary Statistics						Incomplete Data
Variable	N	Mean	Std Dev	Minimum	Maximum	
total_score	200	62.5200000	12.0405554	48.8000000	95.2000000	Columbia University
teaching	200	50.2500000	16.2886402	25.0000000	95.6000000	University of Bonn
international	200	66.7890000	19.6010968	26.1000000	99.8000000	Purdue University
research	200	53.9055000	19.7651543	18.1000000	99.0000000	University of Florida
citations	200	82.9150000	12.7983500	8.6000000	100.0000000	Yeshiva University
income	195	56.1205128	23.1508141	28.0000000	100.0000000	
num_students	198	23828.84	13155.51	462.0000000	83236.00	
student_staff_ratio	198	17.2075758	11.5178076	3.6000000	85.8000000	
international_students	198	19.9242424	10.2684435	1.0000000	54.0000000	
female_male_ratio	181	1.0926377	0.3459290	0.3513514	2.8000000	

Figure 3. Summary statistics for response and explanatory variables.

Figure 4. Universities with missing 'income' variables.

Scatter Plot Matrix

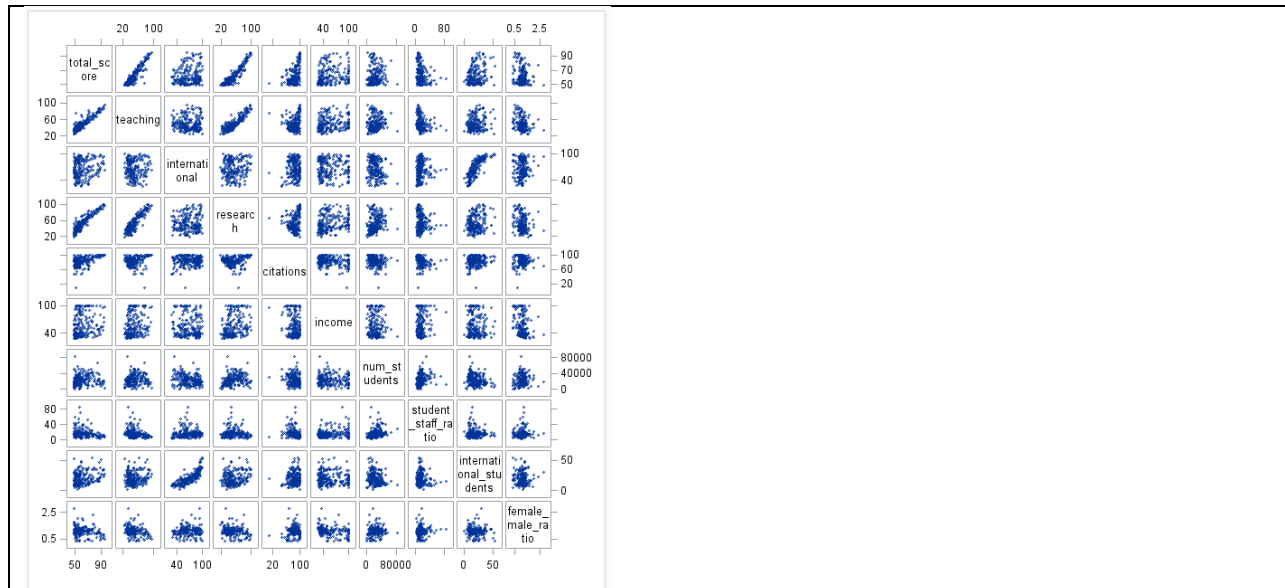


Figure 5. Scatterplot matrix of response and explanatory variables. Visual confirmation shows some multicollinearity but no nonlinear relationships.

Correlation Matrix

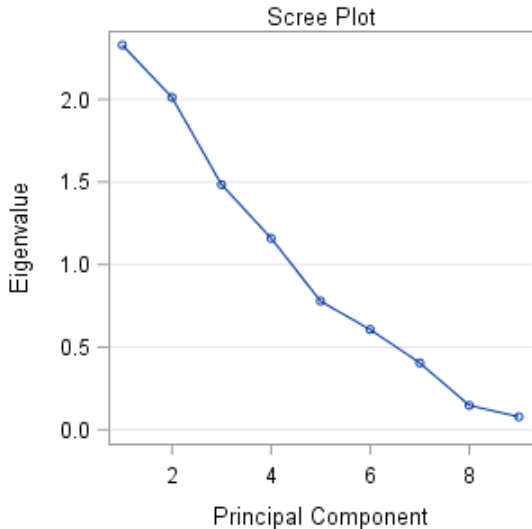
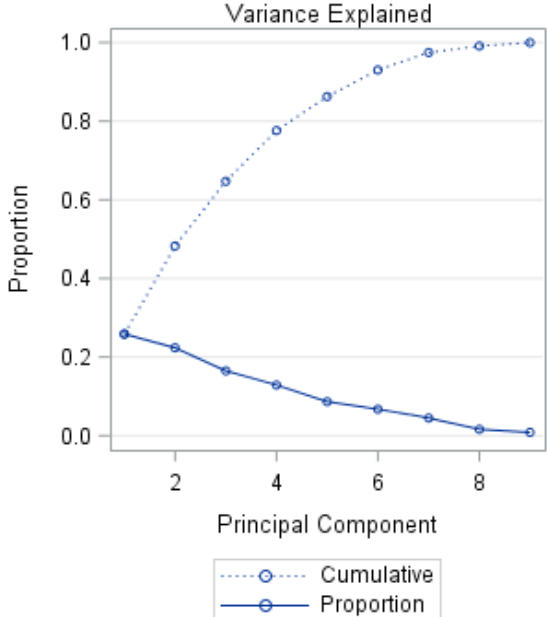
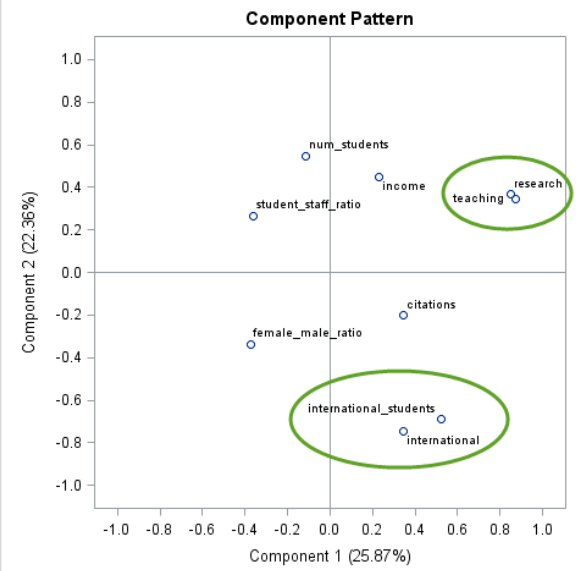
Pearson Correlation Coefficients Prob > r under H0: Rho=0 Number of Observations										
	total_score	teaching	international	research	citations	income	num_students	student_staff_ratio	international_students	female_male_ratio
total_score	1.00000 200	0.90992 <.0001 200	0.15111 0.0327 200	0.92324 <.0001 200	0.46658 <.0001 200	0.17943 0.0121 195	0.07175 0.3151 198	-0.20918 0.0031 198	0.28266 <.0001 198	-0.20086 0.0067 181
teaching	0.90992 <.0001 200	1.00000 200	-0.04339 0.5418 200	0.89483 <.0001 200	0.18638 0.0082 200	0.20379 0.0043 195	0.12591 0.0771 198	-0.26084 0.0002 198	0.12493 0.0795 198	-0.28042 0.0001 181
international	0.15111 0.0327 200	-0.04339 0.5418 200	1.00000 200	0.05788 0.4156 200	0.06386 0.3026 200	-0.07419 0.3026 195	-0.24234 0.0006 198	0.02264 0.7516 198	0.81230 <.0001 198	0.12751 0.0872 181
research	0.92324 <.0001 200	0.89483 <.0001 200	0.05788 0.4156 200	1.00000 200	0.15074 0.0331 200	0.26071 0.0002 195	0.22284 0.0016 198	-0.08539 0.2317 198	0.17946 0.0114 198	-0.23549 0.0014 181
citations	0.46658 <.0001 200	0.18638 0.0082 200	0.06386 0.3690 200	0.15074 0.0331 200	1.00000 200	-0.22393 0.0016 195	-0.18379 0.0095 198	-0.22688 0.0013 198	0.14474 0.0419 198	0.08143 0.2758 181
income	0.17943 0.0121 195	0.20379 0.0043 195	-0.07419 0.3026 195	0.26071 0.0002 195	-0.22393 0.0016 195	1.00000 195	-0.00147 0.9838 193	0.22457 0.0017 193	-0.06576 0.3636 193	-0.30110 <.0001 177
num_students	0.07175 0.3151 198	0.12591 0.0771 198	-0.24234 0.0006 198	0.22284 0.0016 198	-0.18379 0.0095 198	-0.00147 0.9838 193	1.00000 198	0.25334 0.0003 198	-0.26082 0.0002 198	0.07960 0.2868 181
student_staff_ratio	-0.20918 0.0031 198	-0.26084 0.0002 198	0.02264 0.7516 198	-0.08539 0.2317 198	-0.22688 0.0013 198	0.22457 0.0017 193	0.25334 0.0003 198	1.00000 198	-0.11655 0.1020 198	0.07881 0.2916 181
international_students	0.28266 <.0001 198	0.12493 0.0795 198	0.81230 <.0001 198	0.17946 0.0114 198	0.14474 0.0419 198	-0.06576 0.3636 193	-0.26082 0.0002 198	-0.11655 0.1020 198	1.00000 198	-0.04090 0.5846 181
female_male_ratio	-0.20086 0.0067 181	-0.28042 0.0001 181	0.12751 0.0872 181	-0.23549 0.0014 181	0.08143 0.2758 181	-0.30110 <.0001 177	0.07960 0.2868 181	0.07881 0.2916 181	-0.04090 0.5846 181	1.00000 181

Figure 6. Pearson Correlation Matrix. The red boxes highlight the variables with high multicollinearity.

PCA Results

As we have fulfilled the suitability for PCA analysis, we can continue the analysis using the correlation matrix. We examine the Scree Plot as shown on Figure 7. Unfortunately, we do not see a steep drop, so

there is not clear-cut answer at this point as to the number of principal components to retain. However, we do note that the first four principal components account for approximately 80% of the cumulative variance as shown in Figure 8 and 9. We note the close grouping of research/teaching in Principal Component 1 and international_students/international in Principal Component 2.

Scree Plot	Cumulative Variance																																																							
																																																								
Figure 7. Scree Plot showing no sharp drop off.	Figure 8. Cumulative variance as explained by each principal component.																																																							
Eigenvalues	Principal Components 1 & 2																																																							
<table><tr><th colspan="5">Eigenvalues of the Correlation Matrix</th></tr><tr><th></th><th>Eigenvalue</th><th>Difference</th><th>Proportion</th><th>Cumulative</th></tr><tr><td>1</td><td>2.32861463</td><td>0.31650258</td><td>0.2587</td><td>0.2587</td></tr><tr><td>2</td><td>2.01211205</td><td>0.52777917</td><td>0.2236</td><td>0.4823</td></tr><tr><td>3</td><td>1.48433288</td><td>0.32460116</td><td>0.1649</td><td>0.6472</td></tr><tr><td>4</td><td>1.15973172</td><td>0.38173404</td><td>0.1289</td><td>0.7761</td></tr><tr><td>5</td><td>0.77799768</td><td>0.17174396</td><td>0.0864</td><td>0.8625</td></tr><tr><td>6</td><td>0.60625372</td><td>0.20119961</td><td>0.0674</td><td>0.9299</td></tr><tr><td>7</td><td>0.40505411</td><td>0.25778041</td><td>0.0450</td><td>0.9749</td></tr><tr><td>8</td><td>0.14727370</td><td>0.06864422</td><td>0.0164</td><td>0.9913</td></tr><tr><td>9</td><td>0.07862949</td><td></td><td>0.0087</td><td>1.0000</td></tr></table>	Eigenvalues of the Correlation Matrix						Eigenvalue	Difference	Proportion	Cumulative	1	2.32861463	0.31650258	0.2587	0.2587	2	2.01211205	0.52777917	0.2236	0.4823	3	1.48433288	0.32460116	0.1649	0.6472	4	1.15973172	0.38173404	0.1289	0.7761	5	0.77799768	0.17174396	0.0864	0.8625	6	0.60625372	0.20119961	0.0674	0.9299	7	0.40505411	0.25778041	0.0450	0.9749	8	0.14727370	0.06864422	0.0164	0.9913	9	0.07862949		0.0087	1.0000	
Eigenvalues of the Correlation Matrix																																																								
	Eigenvalue	Difference	Proportion	Cumulative																																																				
1	2.32861463	0.31650258	0.2587	0.2587																																																				
2	2.01211205	0.52777917	0.2236	0.4823																																																				
3	1.48433288	0.32460116	0.1649	0.6472																																																				
4	1.15973172	0.38173404	0.1289	0.7761																																																				
5	0.77799768	0.17174396	0.0864	0.8625																																																				
6	0.60625372	0.20119961	0.0674	0.9299																																																				
7	0.40505411	0.25778041	0.0450	0.9749																																																				
8	0.14727370	0.06864422	0.0164	0.9913																																																				
9	0.07862949		0.0087	1.0000																																																				
Figure 9. Eigenvalues of the Correlation Matrix.	Figure 10. First two principal components.																																																							

Eigenvectors									
	Eigenvectors								
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9
teaching	0.571112	0.243576	-0.114009	0.120196	-0.041789	-0.170104	-0.261697	0.027080	0.697966
international	0.225329	-0.525751	0.379151	0.177798	-0.036377	0.000202	0.087801	-0.693289	0.088104
research	0.557107	0.259914	0.018835	0.246847	0.040146	-0.098551	-0.244558	-0.064088	-0.696813
citations	0.224952	-0.139703	-0.507302	0.050629	0.638333	0.425581	0.272559	-0.081657	0.020402
income	0.151420	0.315754	0.471548	-0.289380	0.417998	-0.338576	0.528316	0.042342	0.031723
num_students	-0.076040	0.385141	0.032214	0.624552	-0.320290	0.262048	0.525885	-0.046720	0.069198
student_staff_ratio	-0.235364	0.186882	0.508690	0.266692	0.441285	0.398190	-0.457937	0.061814	0.113901
international_students	0.342633	-0.487575	0.286518	0.112295	-0.112826	0.189892	0.155290	0.690744	-0.011794
female_male_ratio	-0.246137	-0.237689	-0.145889	0.576978	0.321058	-0.634284	-0.014655	0.151383	0.014430

Figure 11. Eigenvectors representing the loadings for the nine principal components.

We are now at the point where we examine the Eigenvectors in Figure 11 to see if we can associate real-world meaning to the principal components.

The first principal component associates the traditional measurements of teaching and research. That is, both increase in teaching and research have a positive association with the overall score, which isn't too surprising. The second principal component groups international outlook and international students, having a negative association with the total score. Somewhat surprising at first glance, but remember the weighting for international is only 7.5%, and the negative association in this principal component confirms that teaching, research, and citations are the main driving factors. The third principal component shows almost perfect contrast between citations and student to staff ratio. That is, citations decrease as the number of students to staff increases. The fourth principal component groups number of students and female to male ratio. That is the tendency to have a greater female to male population as the number of students increase.

Based on these rough interpretations, it would seem reasonable to proceed with fitting a regression model, which according to Figure 9, should account for 77.6% of the variation in the data.

Model Selection and Regression Analysis

By definition, principal components are orthogonal, thus we can represent the regression equation as such:

$$Total\ Score = \beta_0 + \beta_1 P_1 + \beta_2 P_2 + \beta_3 P_3 + \beta_4 P_4$$

Where P_1, P_2, P_3 , and P_4 are principal components.

Parameter Estimates	Analysis of Variance
---------------------	----------------------

Parameter Estimates						Analysis of Variance					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Intercept	1	62.29435	0.19071	326.64	<.0001	Model	4	24344	6085.98087	945.36	<.0001
Prin1	1	7.28619	0.12533	58.14	<.0001	Error	172	1107.29087	6.43774		
Prin2	1	1.60179	0.13483	11.88	<.0001	Corrected Total	176	25451			
Prin3	1	-1.50000	0.15698	-9.56	<.0001	Root MSE		2.53727	R-Square	0.9565	
Prin4	1	2.31079	0.17760	13.01	<.0001	Dependent Mean	62.29435	Adj R-Sq	0.9555		
						Coeff Var	4.07303				

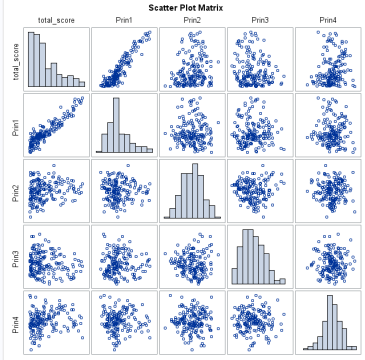
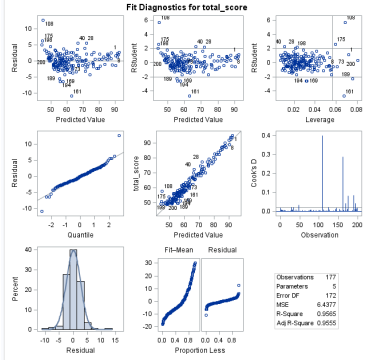
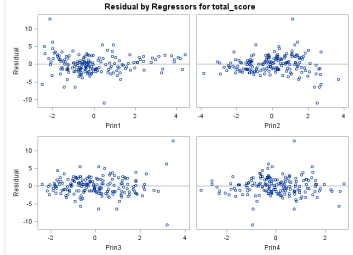
Figure 12. Parameter estimates comprising of the first four principal components.

Figure 13. Results of fitting the first four principal components.

After fitting the first four principal components, we have a good overall fit ($F=945.36$, $p<0.0001$), indicating the first four principal components are significant as shown in Figure 13. The R^2 value of 95.65% indicates about 4.35% of the variation in the data is not explained by the data. As show in Figure 12, there are significant effects for all four principal components ($t=58.14$, $P<0.0001$, $t=11.88$, $P<0.0001$, $t=-9.56$, $P<0.0001$, $t=13.01$, $P<0.0001$). As expected, the scatterplot matrix in Figure 14 shows no evidence of a nonlinear relationship to the response variable.

Now we move on to check the assumptions. Overall, the residual errors look good, as show in the diagnostic plots of Figure 15. There is no evidence to suggest interdependence for the predicted value plots. There is no evidence of non-normal distribution in the quantile and residual histogram plots. Finally, the residual plots for the principal components in Figure 16 show no evidence of non-constant variance.

Next, we examine the Outliers and Leverage points and note that we have four observations that are high Cook's D value and Outliers and Leverage points, they are 108, 161, 175, and 189. These correspond to the Universities: University of Mannheim, Lomonosov Moscow State University, University of Konstanz, and Arizona State University. There is no reason to exclude these observations, as we have no reason to expect that these four universities are not part of the model.

Scatter Plot Matrix	Fit Diagnostics	Residual by Regressors
		
Figure 14. Scatterplot matrix of the response variable and first four principal components.	Figure 15. Diagnostic plots for the fitted model using all four principal components.	Figure 16. Residuals for each principal component.

We have no good reason to simplify the model any further, and thus our regression equation for the final model remains with all four principal components intact:

$$Total\ Score = 62.3 + 7.3P_1 + 1.6P_2 - 1.5P_3 + 2.3P_4$$

Conclusions

In the quest to find the variables that have the greatest influence of university rankings using principal component analysis, the findings must be framed in an appropriate manner. While a prediction model is not really useful, the discovery of how variables were grouped in the principal components yielded some interesting results.

The first principal component associates the variables teaching and research. It is not surprising as these two fundamental areas reinforce each other in an institution of higher learning and contribute to a total higher score. We note that both teaching and research were assigned a weight of 30% in the original methodology, but surprisingly citations is not part of the first principal component as it was also assigned 30% weighting.

The second principal component groups international outlook and international students, having a negative association with the total score. At first glance, this may seem somewhat surprising given the international outlook one of the five published criteria as being essential for a university ("ability to attract undergrads, postgraduates, and faculty from all over the planet is key to it's success on the world stage"). However, taking a closer look at the weighting methodology, we see that international outlook is given only 7.5% weighting. We confirm this and note that many of the top ranked schools have relatively low international outlook scores. For example, California Institute of Technology is ranked at the top in 2016, but only has an international outlook score of 64 out of 100, below the average of all universities at 66.79. Interestingly, this principal component seems to provide some circumstantial evidence to support the criticism levied against The Times Higher Education World University Rankings that they are in fact undermining non-English speaking institutions. Please refer to:

https://en.wikipedia.org/wiki/Times_Higher_Education_World_University_Rankings

The highest non-English speaking ranked school at #9 is Swiss Federal Institute of Technology in Zurich, Switzerland with a score of 97.9 on international outlook. Given a higher weighting for international outlook would in fact put this school as a contender for the top seven spots.

The third principal component shows almost perfect contrast between citations and student to staff ratio. One interpretation of this principal component is that having fewer students per staff increases the quality of research and therefore leads to more citations. In fact, the top seven-ranked institutions have an average of 9.2 students to staff ratio and an average citations score of 99.19 out of 100. The student to staff ratio for all the universities is 17.2 with an average citations score of 82.915 (see Figure 2). We should also note here that having many citations might not have practical significance to the quality of undergraduate study.

The fourth principal component groups number of students and female to male ratio. As the number of students' increase, so does the female to male ratio. While this principal component does not contain any variables that were used to calculate the original total score, it does provide evidence that the ranking system is including schools that teach majors that not only include traditionally men dominated fields such as engineering.

While principal components one, three, and four were interesting, principal component two revealed the most surprising result. It's difficult to assess the how the weightings are determined, but there is definitely a bias towards teaching/research/citations. Finally, we should address some of the practical issues with these rankings as they are simply too broad when it comes to specific fields of study. A student that is interested in studying Music would have no reason to apply to California Institute of Technology, just as a student interested in studying Mathematics would not apply to Julliard. The best advice to an aspiring college student may well be to take these University rankings with a grain of salt and gain the perspective that these rankings are limited in their scope of analysis.

APPENDIX

```
/* Original Data Load */
PROC IMPORT OUT=WORK.PROJECT2
    DATAFILE="C:/Documents and Settings/james/My Documents/My SAS
Files/9.4/MSDS6372/Project2/timesData3.csv"
    DBMS=CSV REPLACE;
    GETNAMES=YES;
    DATAROW=2;
RUN;

PROC PRINT DATA=WORK.PROJECT2; RUN;

PROC MEANS DATA=WORK.PROJECT2 N MEAN MEDIAN STD MIN MAX;
    VAR total_score teaching international research citations income num_students
student_staff_ratio international_students female_male_ratio;
RUN;

PROC SGSCATTER DATA=WORK.PROJECT2;
    MATRIX total_score teaching international research citations income num_students
student_staff_ratio international_students female_male_ratio;
RUN;

PROC CORR DATA=WORK.PROJECT2 PLOTS(MAXPOINTS=NONE) = MATRIX(NVAR=ALL);
RUN;

PROC PRINCOMP PLOTS=ALL DATA=WORK.PROJECT2 OUT=PCA;
    VAR teaching international research citations income num_students student_staff_ratio
international_students female_male_ratio;
RUN;

PROC REG DATA=PCA outest=PCARESULT plots(label) = (rstudentbyleverage cooks);
    MODEL TOTAL_SCORE=PRIN1-PRIN4;
RUN;

PROC CORR DATA=PCA PLOTS=MATRIX(HISTOGRAM);
    VAR TOTAL_SCORE PRIN1-PRIN4;
RUN;
```