# A Novel Analysis of Collegiate Ranking Data

## Identifying University Attributes which Correlate with Ranking

Ian Johnson
Southern Methodist University

September 19, 2016

# Contents

# Executive Summary

This report explores collegiate ranking data from a number of independent ranking organizations and attempts to identify trends among ranking attributes for each university. A number of university attributes are identified as key factors in university ranking, and other attributes are identified as relatively unimportant with respect to ranks. Southern Methodist University is examined in detail, and SMU's weaknesses and strengths are analyzed to identify what factors would be most valuable for the university to optimize in an effort to increase ranking. Finally, the distribution of top-schools throughout the USA is analyzed and visualized to identify top-performing states and regions.

# 1 Business Understanding

## 1.1 Purpose

The purpose of exploring collegiate ranking data is to identify trends which may provide insights to universities, governments, employers, and students which may help inform their decisions. The following are impetuses for each of those groups:

- **Universities** would like to learn how they can increase their rankings. Discovering trends in ranking data may help administrations discern what factors are most important in optimizing ranking.
- **Governments** on a local and national level would benefit from understanding what draw students to universities, as college students contribute significantly to an economy, and it is in the best interest of any government to have a well-educated constituency.
- **Employers** may take interest in identifying schools, regions, or countries which are likely to have top-tier students so that they can efficiently recruit top talent.
- **Students**, especially those in high school, as well as their parents, take great interest in the rankings of the schools to which they apply. A well-informed understanding of those rankings could help a student decide what colleges are of interest.

## 1.2 Potential Results

Two general sets of results may be of interest to the groups listed above. The first is a somewhat novel result: a clear understanding of the relative rankings of universities. Aggregating the data may help identify which universities are truly top-tier. The second, more difficult to achieve result, is an understanding of correlations between certain university attributes and their rankings.

The latter set of results may be of interest to **universities**, who perpetually seek to increase their own rankings, and **governments**, who take interest in the rankings of their constituent universities, which represent a source of significant positive economic influence. **Students** and **employers**, on the other hand, are more likely to take interest in aggregated rank data, so that they can identify what schools are the most likely to help them succeed, or help them find top talent.

## 1.3 Measure of Success

For each of the two identified goals of the forthcoming analyses, a metric must be defined to evaluate the significance of the results. For the novel goal of aggregating rankings, a successful analysis will provide a clear comparison between any two schools. With respect to the goal of identifying correlations between ranking and other metrics, a successful analysis will be one which describes specific school metrics and how they correlate with overall rank. Additionally, a successful analysis will allow a university to identify what to focus on in an effort to increase ranking.

# 2 Data Understanding

The remainder of this report will refer to a number of datasets, all of which are referenced below. Data analysis on these datasets was done using the R programming language, and a number of 3rd party R packages.

## 2.1 Attribute Information

A number of distinct university ranking datasets will be used. Each of the three main datasets includes many attributes about each university. Two additional datasets will be used which provide information on education expenditure and attainment by country.

### 2.1.1 Times Higher Education Data [1]

The THE dataset contains collegiate ranking data spanning from 2011-2016, and contains the following attributes:

- **world_rank** *ordinal*: the world-wide rank for the university (can be an individual number or a range)
- **university_name** *nominal*: the name of the university
- **country** *nominal*: the country where the university is located
- **teaching** *ratio*: the THE score for teaching
- **international** *ratio*: the THE score for international outlook
- **research** *ratio*: the THE score for research, based on volume, income, and reputation
- **citations** *ratio*: the THE score for citations and research influence
- **income** *ratio*: the THE score for industry income
- **total_score** *ratio*: the THE total score, used for ranking
- **num_students** *ratio*: the number of students attending the university
- **student_staff_ratio** *ratio*: the number of students per staff member
- **international_students** *ratio*: the percentage of students who are international
- **female_male_ratio** *ratio*: the number of female students per male student
- **year** *interval*: the year that this ranking occurred

### 2.1.2 Shanghai Data [2]

The Shanghai Ranking dataset contains collegiate ranking data from 2005-2015, and contains the following attributes:

- **world_rank** *ordinal*: the world-wide rank for the university (can be an individual number or a range)
- **university_name** *nominal*: the name of the university
- **total_score** *ratio*: the Shanghai Ranking total score, used for ranking
- **alumni** *ratio*: alumni score based on the number of alumni winning nobel prizes and fields medals
- **award** *ratio*: metric for the number of staff winning nobel prizes and fields medals
- **hici** *ratio*: metric for the number of highly-cited researchers at the university
- **ns** *ratio*: metric for the number of papers published in *Nature and Science*

- **pub** *ratio*: metric for the number of papers indexed in *Science Citation Index-Expanded* and *Social Science Citation Index*
- **pcp** *ratio*: weighted scores of above five indicators, divided by number of full time academic staff
- **year** *interval*: the year that this ranking occurred

### 2.1.3 CWUR Data [3]

The CWUR Ranking dataset contains collegiate ranking data from 2012-2015, and contains the following attributes:

- **world_rank** *ordinal*: the world-wide rank for the university
- **university_name** *nominal*: the name of the university
- **country** *nominal*: the country where the university is located
- **national_rank** *ordinal*: the nation-wide rank for the university
- **quality_of_education** *interval*: CWUR rank for quality of education
- **alumni_employment** *interval*: CWUR rank for alumni employment
- **quality_of_faculty** *interval*: CWUR rank for quality of faculty
- **publications** *interval*: CWUR rank for publications
- **influence** *interval*: CWUR rank for influence
- **citations** *interval*: CWUR rank for citations
- **broad_impact** *interval*: CWUR rank for broad impact (2014/2015 only)
- **patents** *interval*: CWUR rank for patents
- **score** *interval*: CWUR total score, used for world rank
- **year** *interval*: the year that this ranking occurred

Note: Although rankings are ordinal, not interval, some visualizations will treat rankings as interval data. These operate under the simple assumption that, on average, the difference between any two universities ranked $n$ and $n-1$ is the same as the difference between any two other universities ranked $i$ and $i-1$.

### 2.1.4 Supplimentary Educational Attainment and Expenditure Data [4][5]

The following supplementary datasets will be used for analyses:

- **Barro-Lee Dataset**: The average years of schooling among age and gender groups in 144 countries (1985-2015 every 5 years)
- **NCES Dataset**: The amount of public direct expenditure on education by country (1995-2010 every 5 years)

Because these datasets are not simple table data, they are described above based on contents, rather than based on table schema.

## 2.2 Data Quality

### 2.2.1 Times Higher Education Data

The THE data includes a number of data quality issues to deal with:

- Rank data includes ranges (200-250, for example), and some ranks include equals signs (=85). These data problems are dealt with by removing equals signs, and replacing ranges with the lower end of the range.
- Ratio data is given as x:y instead of as a quotient. This is converted to a quotient in pre-processing
- Percentage data is given in string form (including % sign). The % sign is removed.
- There is missing data for a number of attributes. Predominantly for the *income* column. Missing data was imputed using the per-country 5%-trimmed-mean by attribute.

Data processing for this dataset was performed using the CRAN package 'Zoo' [6]

#### 2.2.2 Shanghai Data

The Shanghai data is much simpler to work with, but it still has a few issues:

- Rank data includes ranges (200-250, for example). This is solved by replacing ranges with the lower end of the range.
- The *total_score* attribute is NA for all rows where the rank is in a range. Therefore, the *total_score* attribute is ignored. The *world_rank* attribute is used in its place, as it essentially represents the same thing.

#### 2.2.3 CWUR Data

The CWUR data is by far the cleanest dataset being used in this report. There are 200 missing values for *broad_impact*, which are imputed using the per-country mean for that attribute.

#### 2.2.4 Supplimentary Educational Attainment and Expenditure Data

The supplementary educational attainment data contains numerous rows of data which represent various statistics about the educational status of a country. The data is very highly dimensional. There are dozens of statistics for each individual country, and each statistic is provided for many years. In order to reduce the dimensional of the data, the average of the educational statistics was taken, to reduce the dataset to a simple 1-1 mapping of a country name to an overall education score. Many of the rows of the dataset were population data which were not included in the computation of the means.

The Expenditure dataset has a number of missing-data related issues:

- Private educational spending data is only included for one year of the study. Because this report does not focus on private education expenditures, this data is omitted.
- There is considerable missing data for the public expenditures of countries. However, for each country, there is at least 3-years worth of data. For that reason, the data is reduced to a two-column set where the first column is the name of the country and the second column is the average expenditure on university education by that country over the 5 years that the data was collected.

## 2.3 First Look at Attributes

### 2.3.1 Times Higher Education Data

To take a first look at the THE data, the data is aggregated by column per year and the mean of each column-year is calculated:

| | year | teaching | international | research | citations | income | total_score | num_students | student_staff_ratio |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2011 | 54.75650 | 54.38921 | 55.45750 | 71.58950 | 50.98029 | 60.42950 | 24155.24 | 15.96545 |
| 2 | 2012 | 37.83806 | 51.27114 | 35.88458 | 57.28706 | 47.00281 | 57.73552 | 23819.15 | 17.93707 |
| 3 | 2013 | 41.68300 | 52.36650 | 40.77750 | 65.26800 | 49.97788 | 59.46234 | 23805.48 | 18.32376 |
| 4 | 2014 | 37.27000 | 54.30200 | 35.56275 | 66.53675 | 50.65175 | 57.52633 | 23507.69 | 18.47540 |
| 5 | 2015 | 38.37082 | 56.03292 | 37.20274 | 68.48379 | 51.02604 | 58.22617 | 23637.81 | 18.67683 |
| 6 | 2016 | 31.63748 | 48.38465 | 28.19245 | 51.40528 | 46.80333 | 58.74096 | 24128.69 | 19.10854 |

*Figure 2.1 - Mean scores per year for THE data*

Figure 2.1 shows is that, in general, THE scores have gone down over the course of the last 5 years. At this point, it's not possible to identify if this is caused by decreasing qualities of universities or by increasing standards from the Times Higher Education scorers.

Figure 2.1 also shows an increasing average student-to-staff ratio over the last 5 years among sampled universities. However, the average number of students is not decreasing significantly. This suggests that the size of the faculty of ranked universities may be decreasing. One possible explanation for this would be the increased prevalence of adjunct faculty members in the united states. The AAUP (American Association of University Professors) recently claimed that over half of US University professors are part time [7]. This seems to suggest that the increasing number of adjunct faculty is responsible for the rise in student-to-staff ratio.

An additional possible reason is that in 2016, nearly 800 universities were included in the dataset, while in 2011 only 200 were included. Figure 2.2 shows the number of samples for the student-to-staff ratio year-by-year:

| | year | student_staff_ratio |
|---|---|---|
| 1 | 2011 | 200 |
| 2 | 2012 | 402 |
| 3 | 2013 | 400 |
| 4 | 2014 | 400 |
| 5 | 2015 | 401 |
| 6 | 2016 | 795 |

*Figure 2.2 - Sample size by year for THE data*

Because so many additional schools were sampled, it's possible that the additional, lower-ranked schools considerably increased the average student-to-staff ratio. This will be explored more in later sections.

To quickly identify if any of the measured attributes are heavily skewed, the following tables show the median of each attribute, for comparison against the above means.

| | year | teaching | international | research | citations | income | total_score |
|---|---|---|---|---|---|---|---|
| 1 | 2011 | 51.45 | 54.54503 | 51.05 | 71.35 | 47.23696 | 56.95000 |
| 2 | 2012 | 33.40 | 49.20000 | 30.45 | 55.60 | 40.35000 | 56.69219 |
| 3 | 2013 | 37.50 | 51.35000 | 35.55 | 63.85 | 43.65000 | 58.55997 |
| 4 | 2014 | 33.75 | 54.20000 | 30.30 | 66.20 | 44.70000 | 56.03333 |
| 5 | 2015 | 34.50 | 54.70000 | 32.50 | 68.50 | 44.50000 | 56.58438 |
| 6 | 2016 | 27.00 | 45.50000 | 22.20 | 50.40 | 39.10000 | 58.55997 |

| | num_students | student_staff_ratio |
|---|---|---|
| 1 | 22712.5 | 14.50000 |
| 2 | 21535.5 | 16.02622 |
| 3 | 21426.0 | 16.00000 |
| 4 | 21306.5 | 16.00000 |
| 5 | 21379.0 | 16.00000 |
| 6 | 20174.0 | 16.60000 |

*Figure 2.3 - Median scores per year for THE data*

A novel look at the medians in figure 2.3 show that there is no significantly skewed data. A table of the ratio of the median and modes of the data show no significant differences for any data attribute. The table is omitted from this report for the sake of brevity.

### 2.3.2 Shanghai Data

To take a cursory look at the Shanghai dataset, the various statistics from the dataset are aggregated by year, and their means are computed:

| | year | alumni | award | hici | ns | pub | pcp |
|---|---|---|---|---|---|---|---|
| 1 | 2005 | 9.263655 | 6.677309 | 15.14116 | 15.72831 | 36.70663 | 19.80602 |
| 2 | 2006 | 9.116466 | 6.604016 | 15.34538 | 15.40462 | 37.16145 | 21.36687 |
| 3 | 2007 | 8.907480 | 6.620472 | 15.19173 | 15.24587 | 36.32815 | 20.75197 |
| 4 | 2008 | 8.587226 | 6.836926 | 15.52834 | 15.11617 | 37.54930 | 21.48263 |
| 5 | 2009 | 8.594188 | 6.912625 | 15.63908 | 14.93126 | 37.31884 | 21.31042 |
| 6 | 2010 | 8.554418 | 7.010241 | 15.64418 | 15.20060 | 38.12189 | 20.23835 |
| 7 | 2011 | 8.634809 | 7.250905 | 15.90382 | 15.65875 | 37.86942 | 19.89879 |
| 8 | 2012 | 12.512367 | 12.185512 | 22.52650 | 21.24947 | 44.11696 | 23.09894 |
| 9 | 2013 | 24.013265 | 28.237755 | 36.25102 | 33.17245 | 52.96122 | 30.26531 |
| 10 | 2014 | 8.038431 | 7.219920 | 15.21831 | 15.85453 | 38.94648 | 21.42354 |
| 11 | 2015 | 7.960442 | 7.434739 | 15.24839 | 15.28755 | 38.85402 | 21.79357 |

*Figure 2.4 - Mean scores per year for Shanghai data*

The first insight from figure 2.4 is that, in general, aggregate scores (*pcp*) have not changed significantly over the course of the years sampled. However, individual statistics have changed somewhat. The *alumni* score, for example, has steadily decreased over the years, while the *pub* score has steadily increased. Interestingly, citation averages for U.S universities by-year have been decreasing since 2001 [8]. One possible explanation of the increasing citation scores is that the scores are cumulative citation scores, as opposed to year-by-year scores. The result of such a measurement system would be that scores have a tendency to increase over time. The principle issue with such a system is that it would heavily favor universities that were elite in the past, and lose focus on which universities are producing the best research on a year-to-year basis. The Shanghai dataset provides no documentation on the meaning of this attribute to discern which of these two measurement strategies is being used [2].

The second major insight that figure 2.4 indicates is that scores were very high in 2012 and 2013. These seem well outside the norm. To examine why, the following table shows the number of universities sampled, year-by-year.

| | year | pcp |
|---|---|---|
| 1 | 2005 | 498 |
| 2 | 2006 | 498 |
| 3 | 2007 | 508 |
| 4 | 2008 | 501 |
| 5 | 2009 | 499 |
| 6 | 2010 | 498 |
| 7 | 2011 | 497 |
| 8 | 2012 | 283 |
| 9 | 2013 | 98 |
| 10 | 2014 | 497 |
| 11 | 2015 | 498 |

*Figure 2.5 - Sample size by year for the Shanghai data*

Figure 2.5 shows that the Shanghai data has the opposite problem of the THE data. The years 2012 and 2013 have far fewer sampled universities, so in those years only a select few elite schools were ranked. This is what caused the significant mean score inflation for those two years.

To check for skewed attributes, the figure 2.6 shows the median of each attribute by year. If any of these medians significantly differ from their respective means, then the data in that column is skewed.

| | year | alumni | award | hici | ns | pub | pcp |
|---|---|---|---|---|---|---|---|
| 1 | 2005 | 0.0 | 0.0 | 11.1 | 12.45 | 33.90 | 17.25 |
| 2 | 2006 | 0.0 | 0.0 | 10.9 | 12.30 | 34.45 | 18.80 |
| 3 | 2007 | 0.0 | 0.0 | 12.8 | 11.90 | 33.90 | 18.25 |
| 4 | 2008 | 0.0 | 0.0 | 12.6 | 11.90 | 35.20 | 19.00 |
| 5 | 2009 | 0.0 | 0.0 | 12.6 | 11.90 | 35.20 | 19.00 |
| 6 | 2010 | 0.0 | 0.0 | 12.5 | 12.30 | 35.85 | 18.40 |
| 7 | 2011 | 0.0 | 0.0 | 12.5 | 12.70 | 35.50 | 18.10 |
| 8 | 2012 | 12.1 | 0.0 | 19.1 | 17.90 | 41.20 | 20.80 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 9 | 2013 | 19.3 | 23.7 | 32.4 | 29.65 | 51.45 | 27.10 |
| 10 | 2014 | 0.0 | 0.0 | 12.2 | 12.30 | 36.90 | 19.70 |
| 11 | 2015 | 0.0 | 0.0 | 12.3 | 12.10 | 36.75 | 19.90 |

*Figure 2.6 - Median scores per year for Shanghai data*

A cursory glance at the medians in figure 2.6 show no skewed data attributes. Once again, a matrix of median-to-mean ratios was computed to compare the medians and the means, and no significant differences were found. Also note that some items have a median of zero. This is due to missing data that was not imputed, so some rows had scores of zero for some attributes. They are imputed or ignored in upcoming visualizations.

### 2.3.3 CWUR Data

For the CWUR data, the final simple table dataset for this report, the same strategy is used to gain some cursory insight into the data. Figure 2.7 indicates the year-by-year averages for every attribute in the CWUR dataset.

| | year | alumni_employment | quality_of_faculty | publications | influence | citations | broad_impact | patents | score |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2012 | 75.390 | 56.930 | 55.020 | 54.890 | 54.420 | 371.3195 | 63.650 | 54.94090 |
| 2 | 2013 | 75.910 | 56.060 | 54.670 | 56.280 | 53.930 | 377.2275 | 63.550 | 55.27120 |
| 3 | 2014 | 363.991 | 188.002 | 500.411 | 500.163 | 447.349 | 496.7350 | 448.968 | 47.27141 |
| 4 | 2015 | 406.536 | 194.253 | 500.419 | 500.275 | 451.334 | 496.6640 | 491.674 | 46.86385 |

*Figure 2.7 - Mean scores per year for CWUR data*

In figure 2.7, it seems that there are two sets of year pairs during which the rankings were very similar. The scores from 2012 and 2013 are nearly identical, and the same is true for 2014 and 2015. From this, two things are apparent. First, it seems that the CWUR data is the most constant over time. Second, it appears that one of two things occurred between 2013 and 2014: either the scoring system changed, or the number of universities sampled dramatically increased. To check which of those is true, figure 2.8 shows the number of universities scored each year over the course of the 4 years.

| | year | score |
|---|---|---|
| 1 | 2012 | 100 |
| 2 | 2013 | 100 |
| 3 | 2014 | 1000 |
| 4 | 2015 | 1000 |

*Figure 2.8 - Sample size by year for the CWUR data*

It appears that the latter of the possibilities occurred: in 2014, the number of universities sampled increased ten-fold.

### 2.3.4 Supplimentary Educational Attainment and Expenditure Data

Because the supplementary attainment and expenditure data was reduced to single-dimensional data with an artificial (unit-less) scale, the only meaningful summary of those datasets is a simple summary with means and quantiles of the new scale of data.

For the educational attainment dataset, the distribution of scores is shown in figure 2.9.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  9.801  12.530  14.560  14.320  16.260  18.780
```

*Figure 2.9 - Distribution of educational attainment scores*

Recall this data is two column table data, where the first column is the name of a country and the second column is an aggregate educational attainment score for that country, computed from the multi-dimensional data in the educational attainment dataset [4].

For the educational expenditure data, the distribution of scores is shown in figure 2.10

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1000  0.8500  0.9917  1.0150  1.1810  1.7830
```

*Figure 2.10 - Distribution of educational expenditure scores*

This data, like the attainment dataset, is a two column table, where the first column is the name of a country and the second column is an aggregate public university expenditure score, computed from the multi-year educational expenditure dataset [5].

## 2.4 Attribute Visualizations

### 2.4.1 Times Higher Education Data

***THE Data Total Score Distribution***

The first important metric to visualize for the THE dataset is the distribution of total scores for the schools surveyed. Figure 3.1 is a histogram of the total scores of all schools included in the 2016 THE collegiate ranking report.

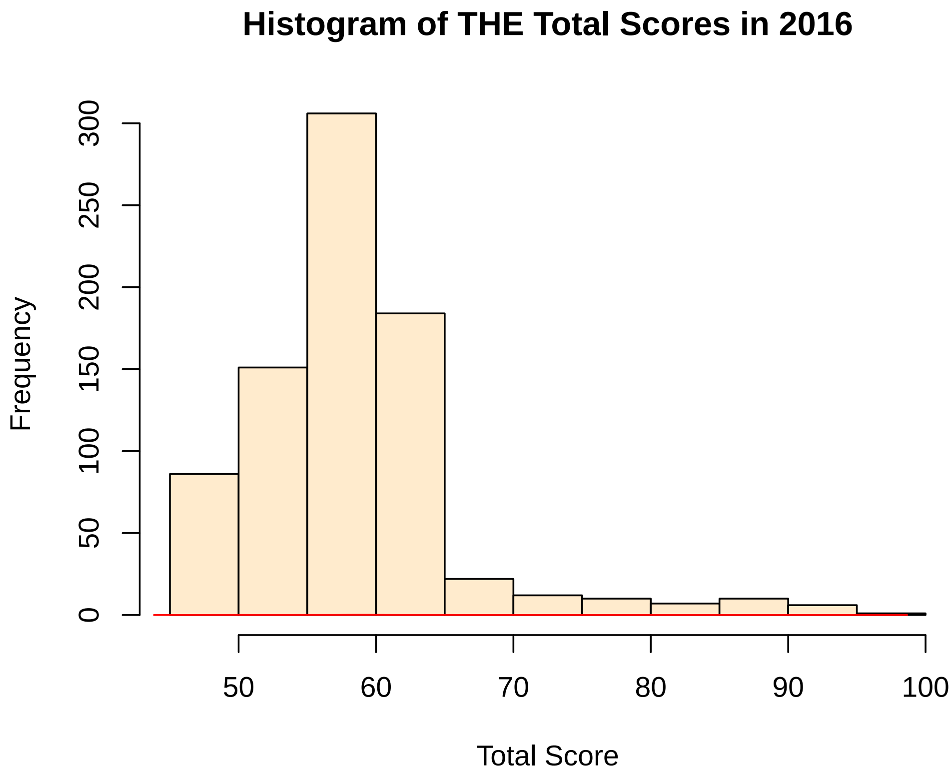


*Figure 3.1 - Distribution of THE total scores*

Figure 3.1 indicates that the total score of a university, as measured by THE, is unimodally distributed and heavily left skewed. There is a very noticeable drop-off in frequency when total score goes from the 60-65 range to the 65-70 range. What this indicates is that the THE dataset has a clear set of elite universities. It also provides a simple cutoff to define a class of 'top-tier' universities. Such a cutoff seems to exist at the 65 total score mark.

#### ***THE Data Individual Score Distributions***

To identify the distributions of the remaining sub-scores of the THE data, a violin plot is used to visualize the teaching, international, research, and citation attributes of each university. The violin plot is built using the CRAN package 'vioplot' [9].

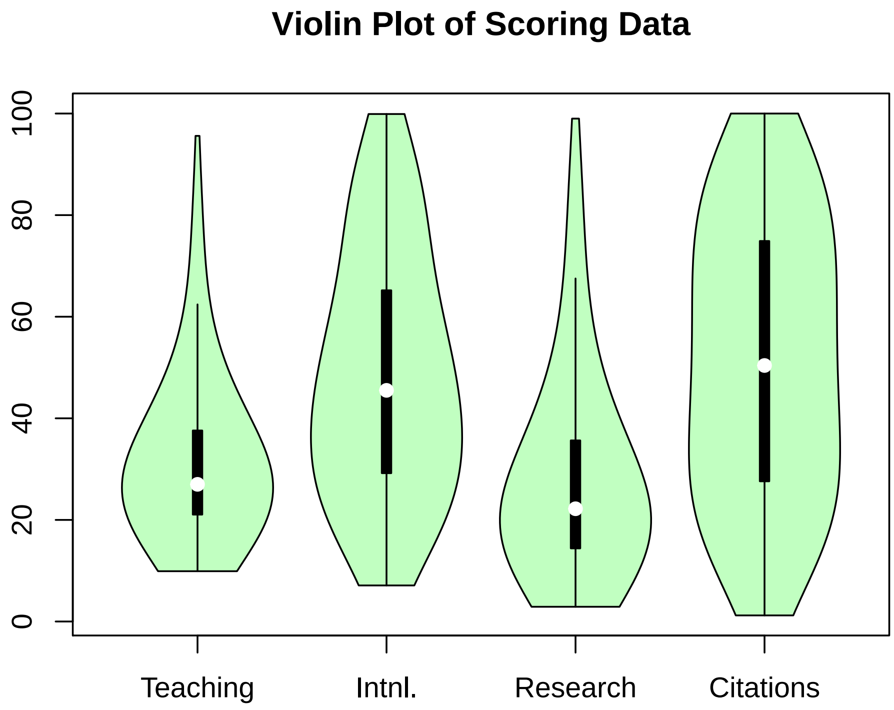


*Figure 3.2 - Distributions of individual THE scores*

Figure 3.2 shows, interestingly, that teaching and research seem to follow somewhat normal distributions which have been cut off by a lower bound, while international and citations scores follow a nearly-uniform distribution. The teaching and research attributes seem to fit the working understanding that there are very few universities in a 'top-tier,' which is reflected by the narrowness of those two violins. The international and citations violins' broadness at the upper end of the scoring spectrum may indicate that international and citation scores are a poor predictor of overall score and overall rank, because there is no elite group based on international and citation scores.

### 2.4.2 Shanghai Data

***Shanghai Data Total Score Distribution***

Much like for the THE data, the first way to assess the Shanghai data is to draw a histogram of the total scores. If the Shanghai data conforms to the trend seen in the THE data – that there is a very small set of elite universities with considerably higher scores than the rest of the universities, then figure 3.3 should once again show a heavily left-skewed distribution.

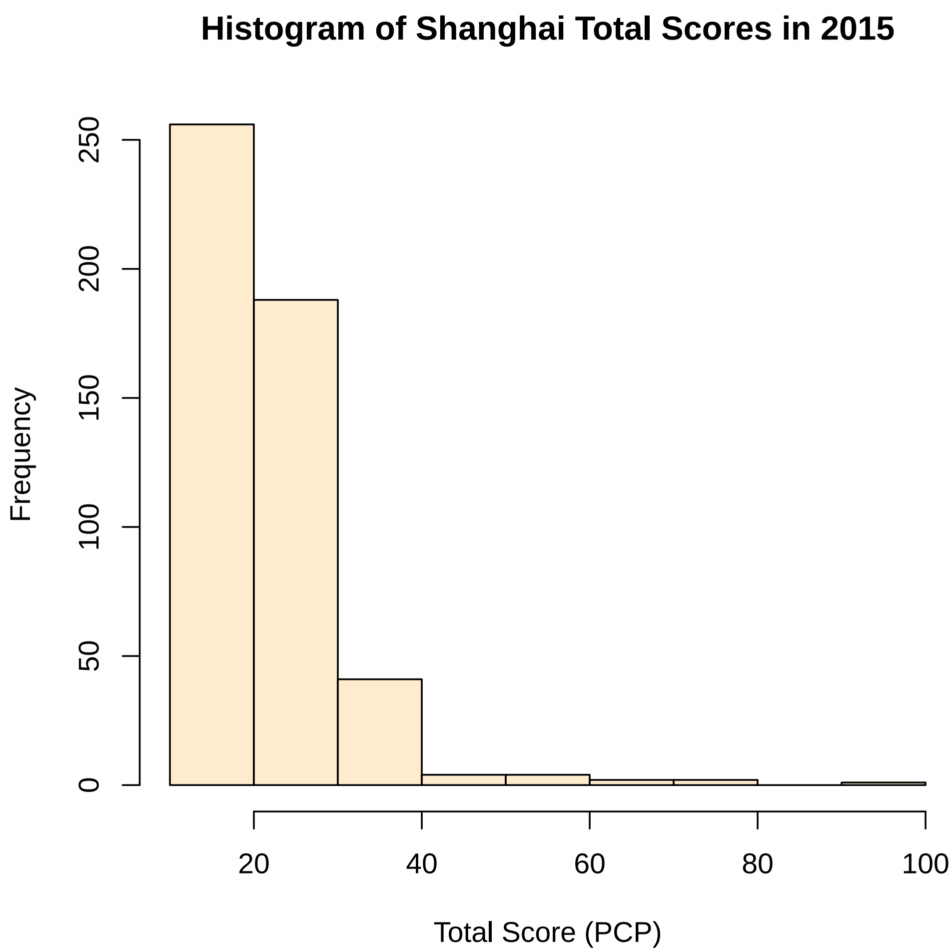


*Figure 3.3 - Distribution of Shanghai total scores*

Figure 3.3, indeed, shows that the Shanghai total score data follows a left-skewed distribution which conforms to the 'top-tier' model indicated by the THE data. In fact, the Shanghai data seems to have an even more heavily skewed distribution of total scores than the THE data. Very few schools achieved a total score of more than 40.

Interestingly, the top-15 schools by THE score and the top 15 schools have 8 schools in common. This means that there are 8 schools which fall into the top-tier by both rankings, and 14 schools that fall into the top-tier by one ranking, but not the other.

### 2.4.3 CWUR Data

***CWUR Data Total Score Distribution***

For the final ranking set, the same histogram visualization is used to identify if, in fact, all 3 of the ranking systems follow a left-skewed score distribution with an elite few top-tier schools.

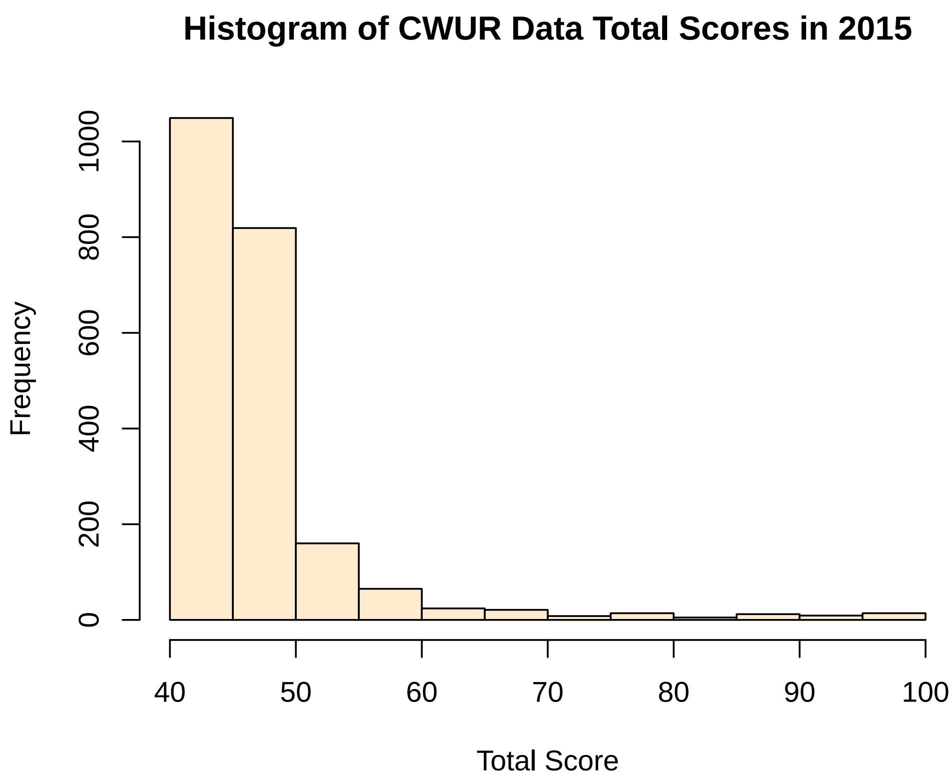


*Figure 3.4 - Distribution of CWUR total scores*

Figure 3.4 shows that the CWUR data, much like the Shanghai and THE data, follows a left-skewed distribution which clearly isolates the top-tier universities from the others. To assess if the three rankings identify the same universities as being among the elite, the top 50 universities from each ranking is combined, and the overlap of the 3 top-50 sets is assessed.

***Comparison of Top-50 Schools from Each Ranking***

When the top-50 universities from each ranking are selected, 14 of the universities present are top-50 in all 3 rankings, 14 of the universities are present in 2 of the 3 top-50 rankings, and 80 of the 150 universities are in only one of the 3 rankings. What this demonstrates is that the ranking systems are clearly not consistent relative to each-other. This suggests that the ranking process is highly subjective, because any objective ranking system would have a higher incidence of overlap in 3 independent top-50 sets.

### 2.4.4 Supplimentary Educational Attainment and Expenditure Data

***Distribution of Attainment Scores by Country***

To establish an idea for the meaning of the attainment score data (which was artificially constructed from the existing data, and therefore has no underlying unit), a histogram with density lines is used. Because the data does not come from any known distribution, and is in fact an aggregation of other distributions, it is expected to follow a normal distribution due to the central limit theorem.

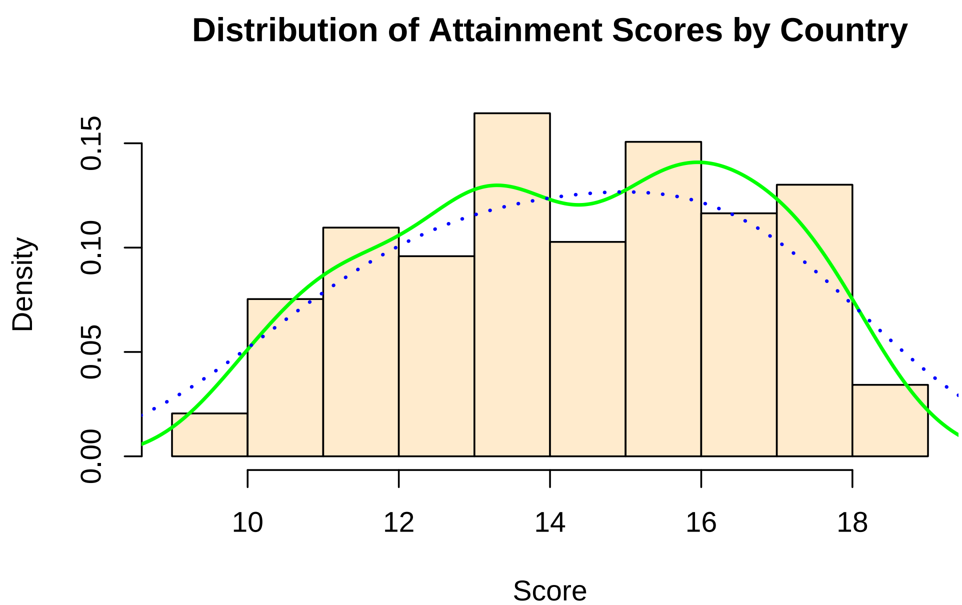


*Figure 3.5 - Distribution of educational attainment scores*

As expected, figure 3.5 shows that the data seems to follow a somewhat-normal distribution. However, it's not a perfectly normal distribution. A simple density line shows a bi-modal distribution, where a density line with a higher bandwidth shows a uni-modal distribution. As a result, educational attainment data alone can not be considered a meaningful metric for a country's educational success on its own, because the distribution of the data follows no clear pattern.

***Distribution of Public University Expenditure Scores by Country***

The same methodology is used for the expenditure score data. Once again, the data is an aggregate of other datasets, and therefore is expected to be normally distributed.

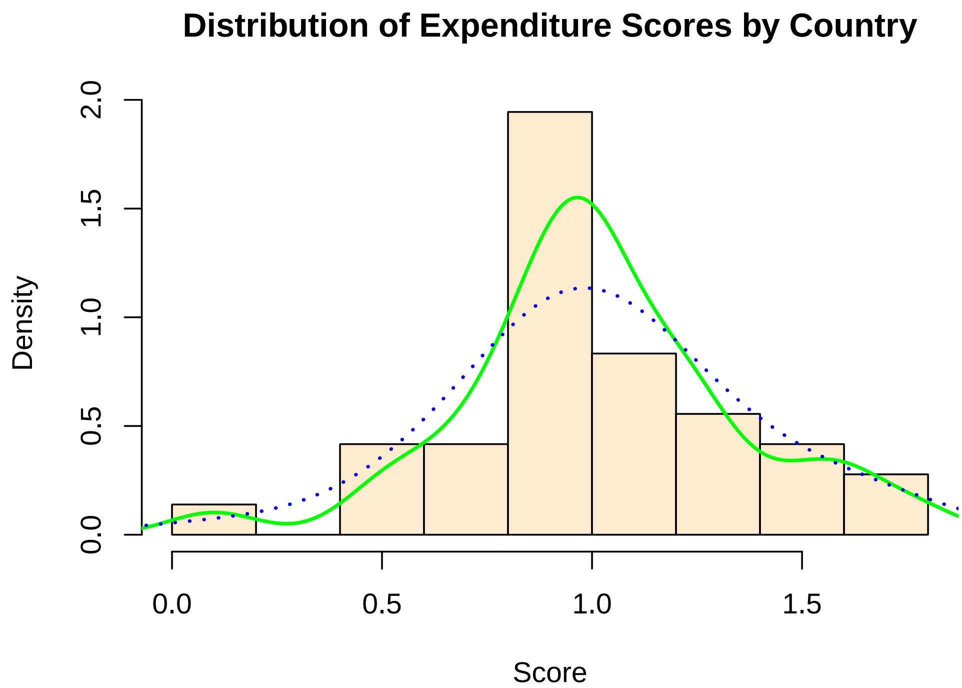


*Figure 3.6 - Distribution of educational expenditure scores*

The public university expenditure data appears to me more normalized than the attainment data, as shown by figure 3.6. However, there is some rightward skew in the data. Nonetheless, both density lines indicate a somewhat normal distribution, which indicates that the expenditure data, much like the attainment data, doesn't follow the top-tier skewed model. This is expected and reasonable, however, as the expenditure and attainment datasets refer to entire countries, not single universities.

## 2.5 SMU: A Case Study

The next objective is to take a close look at Southern Methodist University and how it measures up to other universities. There are three primary goals for assessing SMU as a university:

- Determine an aggregate ranking for SMU based on scores from all overall rankings
- Assess SMU's various strengths and weaknesses using its various individual attribute rankings
- Identify what factors would be best for SMU to improve upon in an effort to improve overall ranking

***SMU Aggregate Overall Rank***

In order to identify an overall rank for SMU, the SMU world ranks from the datasets are averaged for every year they were included in the rankings. Internal country rankings are ignored, as they essentially a direct result of world ranks.

SMU is ranked once in the THE dataset, and 5 times in the Shanghai dataset. The average of these six rankings is 393. This provides a lower-bound for the ranking that SMU truly has.

Since SMU is omitted in a number of rankings, it can be inferred that SMU ranks too low to appear on those rankings. Therefore, SMU's aggregate rank has to be at least 393. In the optimistic assumption case, where SMU is assumed to be ranked just below the cutoff to make each year's ranking, the new average ranking is 442. This average is calculated using the average of the existing rankings, and replacing missing rankings with the equivalent ranking of a university who fell one school short of being included on the list.

It is difficult to provide a pessimistic estimate for ranking, as an omitted data point could possibly range from the most ideal ranking to the worst ranking, where SMU is ranked as the last in the world. Because such an estimate would be difficult and wildly inaccurate, the optimistic rank of 442 is used.

### *SMU Aggregate Individual Scores*

For each of the non-ranking scores, SMU's aggregate score is calculated on a normalized scale of each attribute. Therefore, a score of 1 represents SMU being 1 standard deviation above average, a score of -1 indicates that SMU is 1 standard deviation below average, and so on. The attributes included in the table below are a combination of those found in the THE and Shanghai datasets, and in some cases are an aggregate of the two datasets.

| Attribute | Score |
|---|---|
| Num_Students | 0.0107 |
| Student_Staff_Ratio | −0.4291 |
| Teaching | −1.0922 |
| Citations | 0.2746 |
| Alumni | 0.3656 |

*Figure 4.1 - SMU aggregate scores*

Figure 4.1 shows that SMU has slightly above average student counts, citations scores, and alumni scores, and also has fewer students per faculty member, but that SMU's principal weakness is a very low teaching score.

### *How to Improve SMU's Overall Rank*

In order to identify what attributes are most inhibitive to SMU earning a higher ranking, the normalized mean scores from the attributes computed above for SMU is computed for the average of the top-20 schools in the THE and Shanghai datasets in figure 4.2.

| Attribute | Score |
|---|---|
| Num_Students | −0.1407 |
| Student_Staff_Ratio | −0.7697 |
| Teaching | 3.4486 |
| Citations | 1.7093 |
| Alumni | 2.9714 |

*Figure 4.2 - Top-20 aggregate scores*

In figure 4.2, the two factors that stand out as potential improvements for SMU are the teaching and publications scores. SMU does not vary significantly from the top 20 schools in terms of number of students, and it has an only slightly higher student-to-staff ratio, but it falls more than 4 standard deviations below the top-20 average in terms of teaching, and 1.5 standard deviations below the top-20 average in terms of citations. Therefore, it is in SMU's best interest to focus on improving teaching and research. This means that, for SMU to improve their rank, they need to turn their focus toward catering toward the interests of their existing faculty members and attracting strong new faculty members. This is highly opposed to SMU's current operational model, which has a strong focus on athletics and aesthetics.

## 2.6 Attribute Relationships

***Relationships in the THE Dataset***

The THE dataset can be broken down into two subsets: direct measurements and rankings. Each of the two sets of attributes is evaluated for possible correlations. Figure 5.1 shows a scatter matrix of the 4 ranking attributes in the THE dataset, while figure 5.2 shows a scatter matrix of the direct measurement attributes in the THE dataset.

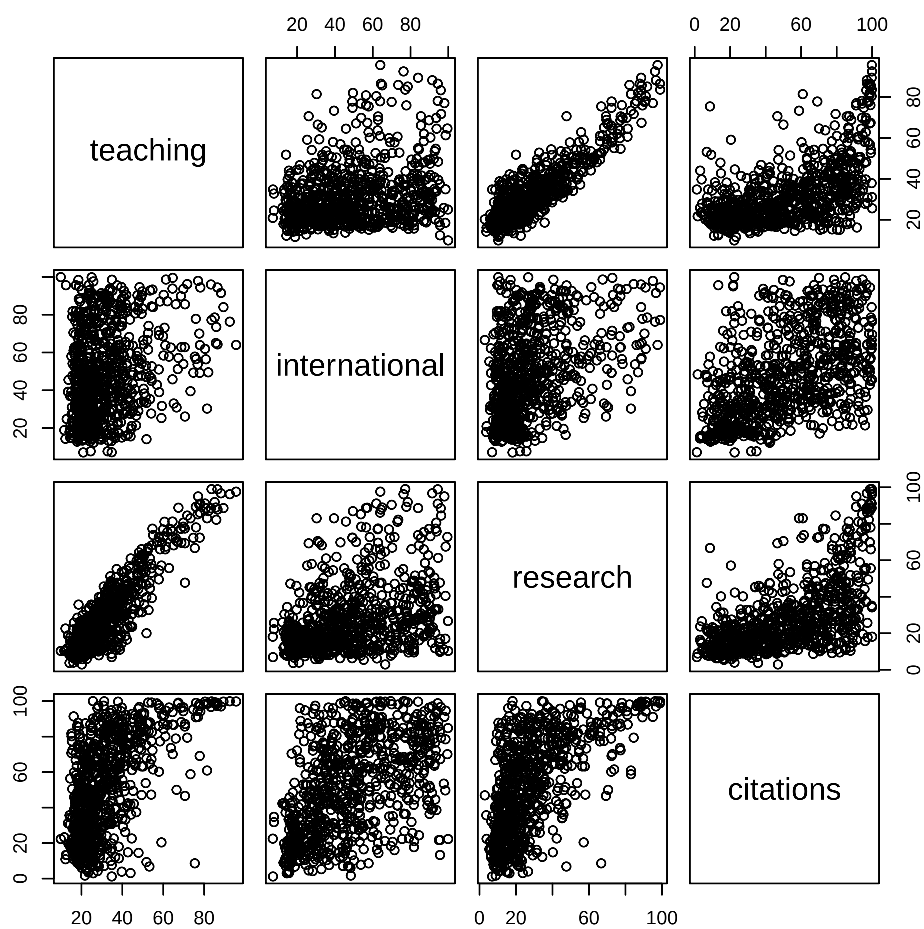


*Figure 5.1 - Scatter matrix of THE rank attributes in 2016*

Figure 5.1 shows that there is a strong positive correlation between teaching scores and research scores. This is likely due to the fact that teaching and research are generally done by the same group of people – the faculty. This solidifies the idea that SMU's best option for increasing rank is to cater to the needs of the faculty and to seek out outstanding new faculty members. Figure 5.1 also shows that there is very little correlation between international score and other scores. However, international students are very profitable for universities, which provides an impetus of its own for seeking out a large group of international students [10].

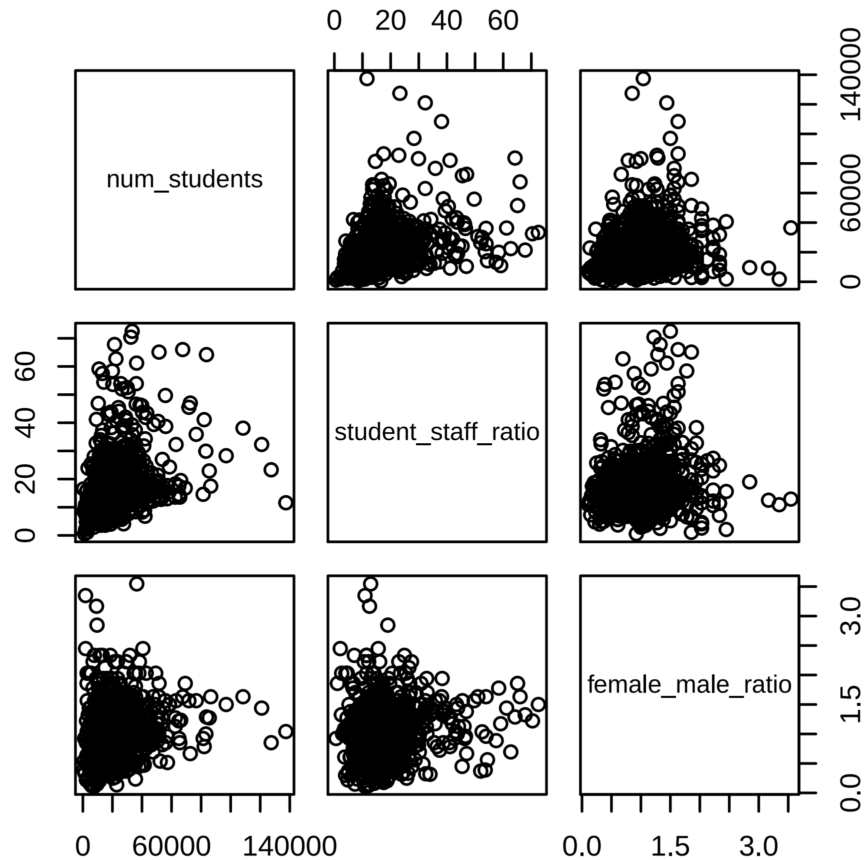


*Figure 5.2 - Scatter matrix of THE direct measurement attributes in 2016, excluding outliers*

Figure 5.2 shows that there is no strong correlation among the 3 directly measured attributes for the THE dataset in 2016, with outliers excluded.

***Relationships in the Shanghai Dataset***

The Shanghai dataset contains 6 scored attributes, none of which are direct measurements. All 6 of those attributes are compared in figure 5.3 using a correlation plot. The correlation plot is built using the CRAN package 'corrplot' [11].

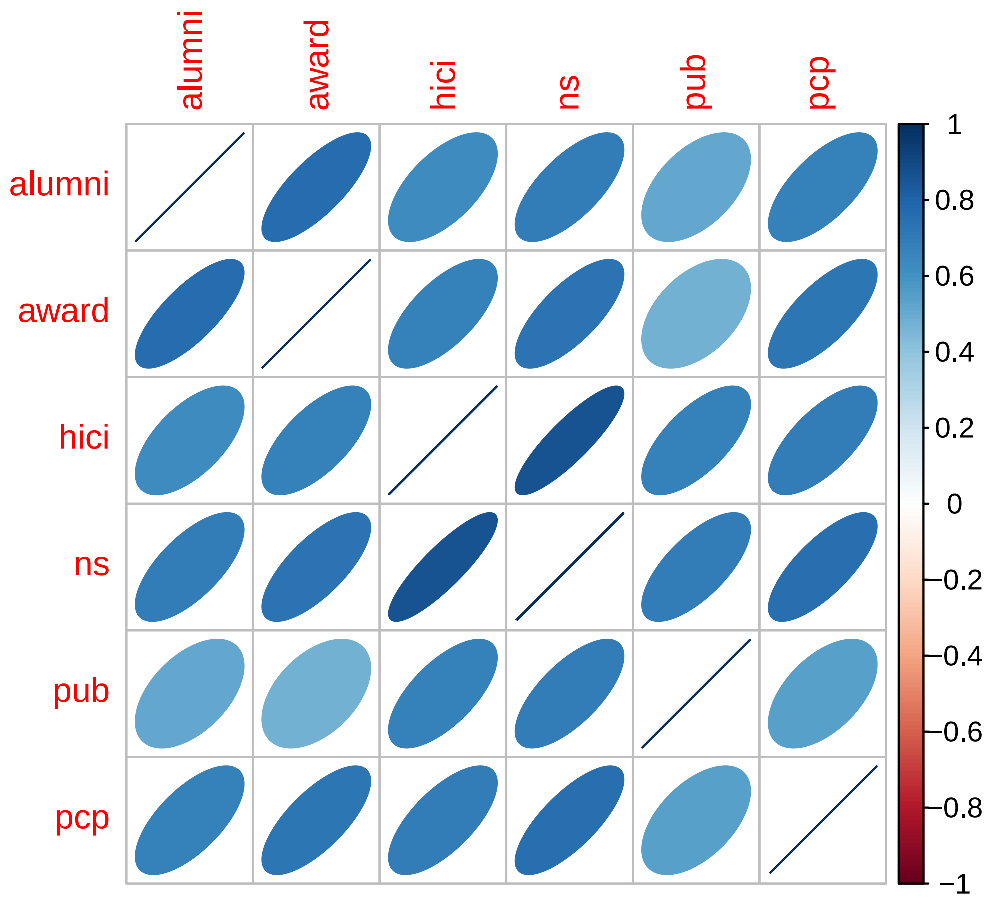


*Figure 5.3 - Correlation plot of Shanghai dataset scoring attributes*

Figure 5.3 shows that all of the Shanghai dataset's attributes are highly correlated with each other. With the exception of the publication ranking, all rankings are correlated ($r^2 > 0.5$) with all other rankings. Because of this, it may be possible to reduce the dimensionality of this dataset using PCA and still have a strong single-attribute predictor of overall rank.

### *Principal Component Analysis on the Shanghai Dataset*

Figure 5.4 shows that the first principle component of the Shanghai scores is considerably stronger than the subsequent 5. It appears, therefore, that PCA is powerful for reducing the dimensionality of the Shanghai dataset. Figure 5.5 shows the results of comparing the ranks of all universities in the Shanghai ranking with the first principal component of their 6 rankings.

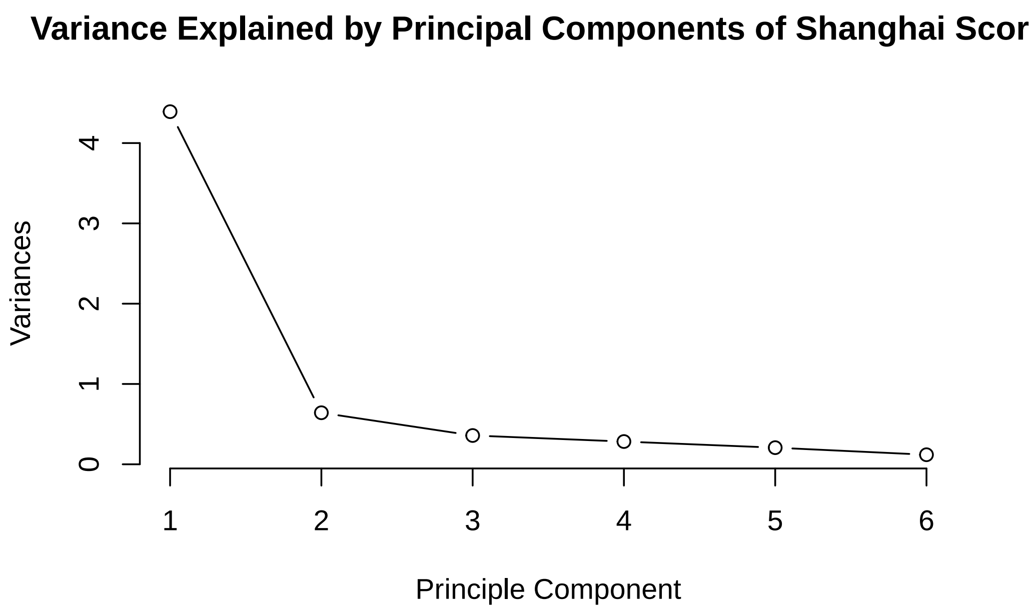


*Figure 5.4 - Variation explained by each of the principal components of the Shanghai dataset*

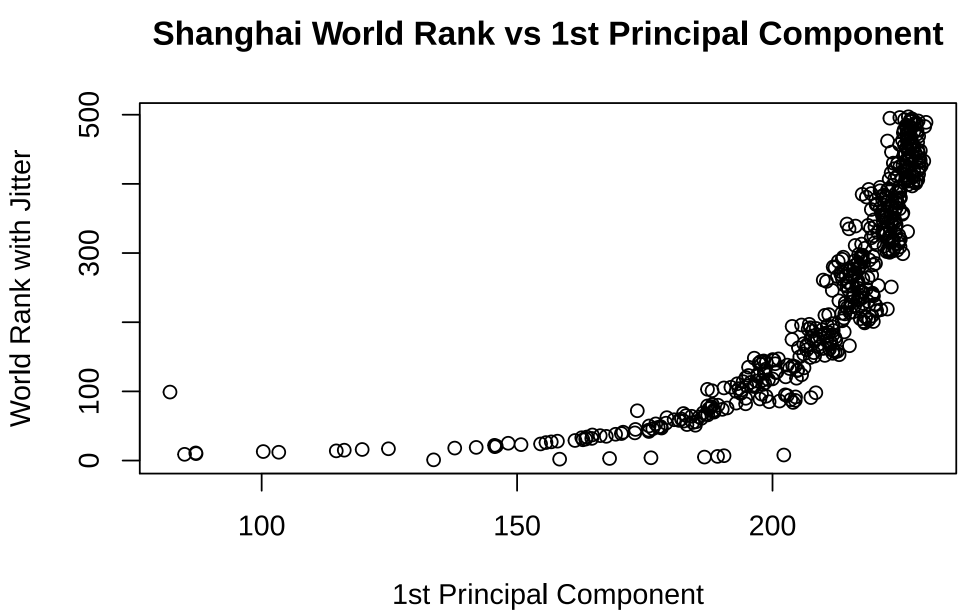


*Figure 5.5 - Shanghai rankings vs first principal component of Shanghai attributes*

The strong trend in figure 5.5 shows that PCA was quite successful on the Shanghai datasets. The figure shows a strong correlation between the world rank of a university and the first principal component of their 6 Shanghai scores. In fact, the trend is so strong that the weights of the principal components are relevant to assess the importance of each scoring in the overall ranking. Figure 5.6 shows the weights of each ranking in each principal component. The weights in the first principal components represent the importance of each score when predicting world rank.

The world rank (y-axis) includes some jitter in figure 5.5, to help visualize the trend despite the ranged rankings.

```
Standard deviations:
[1] 2.0955087 0.8010222 0.5986398 0.5320680 0.4548351 0.3447694

Rotation:
              PC1        PC2         PC3        PC4        PC5         PC6
alumni -0.3974386 -0.4059479  0.62903795 -0.1053974  0.5173036 -0.05337156
award  -0.4086869 -0.4600694  0.03625149 -0.1801179 -0.7631795 -0.07149584
hici   -0.4232831  0.2631814 -0.40096377 -0.4742163  0.2138976 -0.56582836
ns     -0.4446542  0.1575540 -0.23953040 -0.2294113  0.1100635  0.80953136
pub    -0.3594227  0.6988589  0.48049080  0.2756294 -0.2638092 -0.07728722
pcp    -0.4110315 -0.2026106 -0.39240149  0.7765093  0.1499754 -0.10278009
```

*Figure 5.6 - Shanghai rankings PCA weights*

Figure 5.6 shows that, interestingly, all 6 rankings seem to be weighted almost equally in the first principle component. This means that, on average, all individual rankings in the Shanghai dataset carry equal weight in deciding on the world rank.

#### *Relationships in the CWUR Dataset*

The CWUR dataset included exclusively rank data; no direct measurement data is present. Figure 5.7 shows a plot of all of the rank correlations of the ranks in the CWUR dataset.

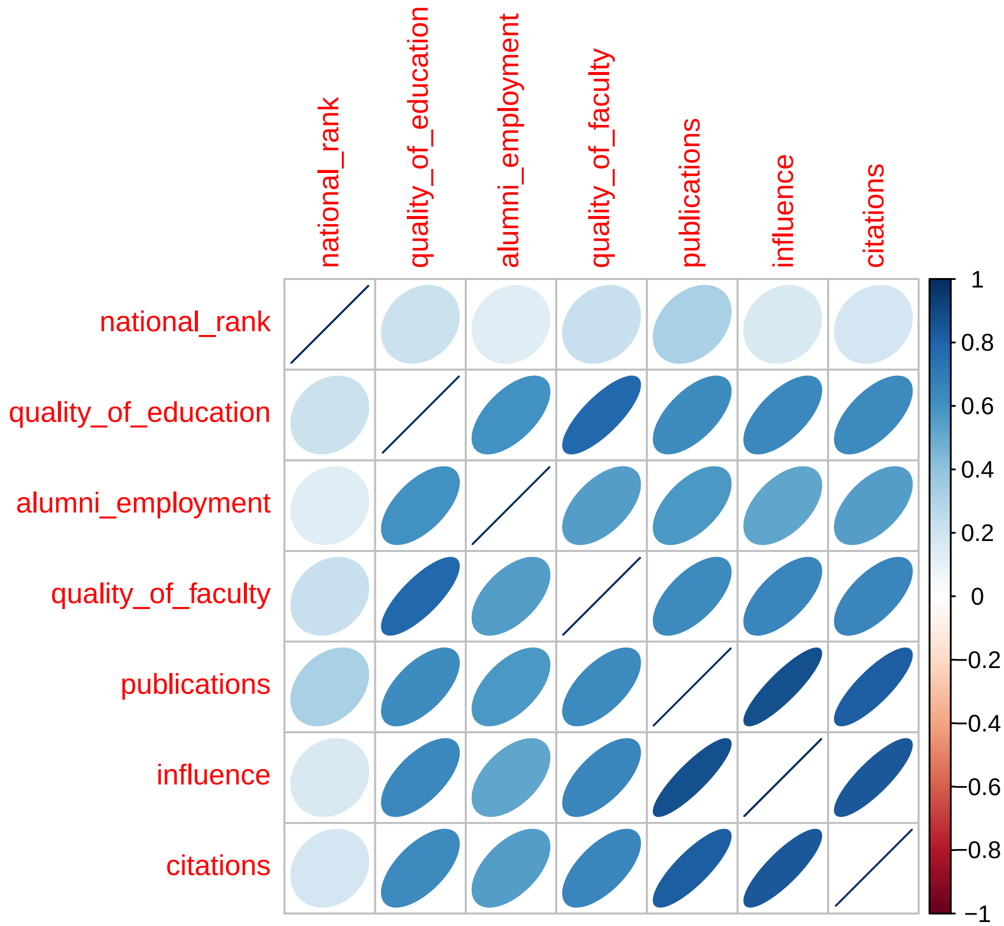


*Figure 5.7 - Rank Correlations of CWUR Individual Rankings*

Figure 5.7 shows that, much like the Shanghai dataset, individual attribute scores in the CWUR dataset are highly cross-correlated. Interestingly, the only attribute that isn't correlated with the others is the national rank. PCA can, once again, be used, to determine if the dimensionality of the data can be easily reduced.

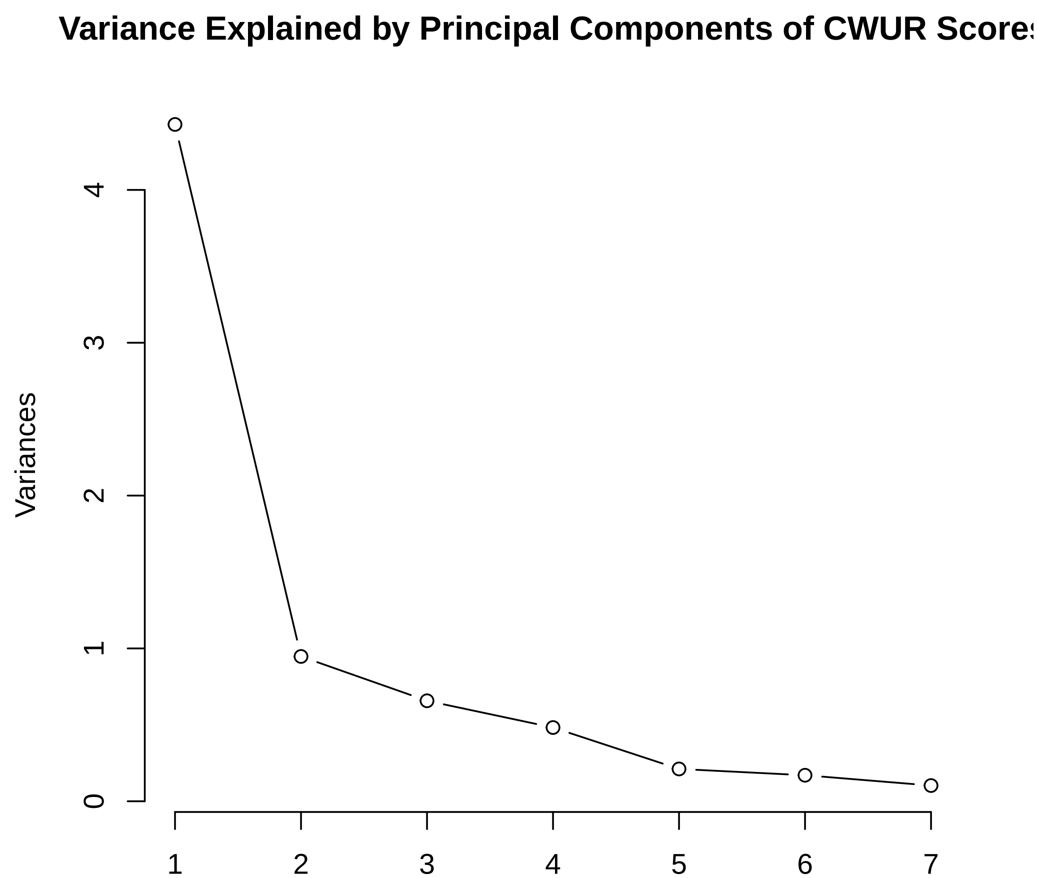


*Figure 5.8 - Variation explained by each of the principal components of the Shanghai dataset*

Figure 5.8 shows that, while the first principal component of the CWUR ranks does significantly outweigh the others, it doesn't do so to the same degree as the first principal component of the Shanghai dataset. For that reason, it is unlikely that the single-principal-component simplification of the CWUR dataset has the same predictive quality for overall rank.

## 2.7 Geographic Relationships

For each university, GPS data (or some sort of location data) allows for a visualization of the distribution of top universities around the country, or around the world. The THE dataset can be cross-referenced with a list of universities by state to provide such a visualization. This may provide insight into which states or regions provide the best education for students in universities. This information could help state and federal government organizations, and could also aid in the process of selecting a school for high school seniors.

Data for cross-referencing universities by name and location is gathered from the US Department of Education's Accreditation dataset [12].

Figure 6.1 shows a map of the United States where each state is color-coded based on the average THE score of universities in the THE rankings in that state. White states represent states where no top universities reside.

The following figures are generated with CRAN packages 'ggplot2' [13] and 'maps' [14]

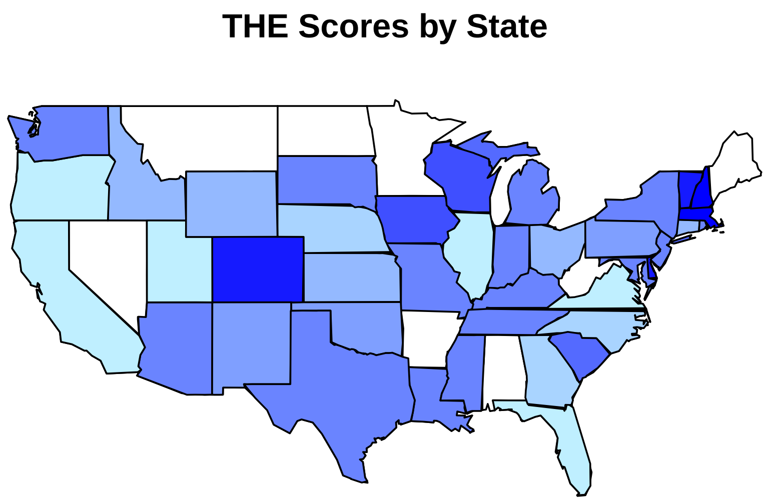


*Figure 6.1 - THE Scores of universities in each state*

Figure 6.1 shows no strong geographic trend, although the Northeast does appear to have a slight edge on the rest of the country when it comes to THE scores. It is also relevant to identify which states have the most top schools. The same visualization can be used, this time color-coded based on number of top schools in the THE ranking.

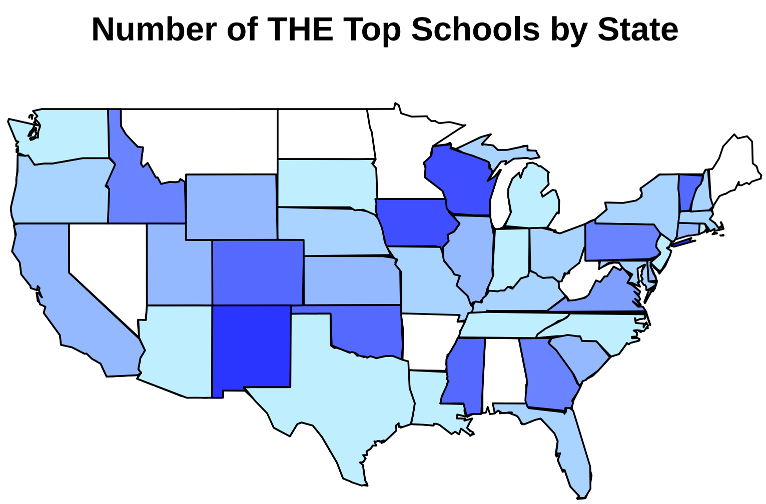


*Figure 6.2 - THE top-schools count in each state*

Figure 6.2 shows a similar result to figure 6.1. The two figures collectively identify which states have the best higher education systems. Some states, such as North Dakota, perform poorly in both figures, while some schools, like Colorado, perform exceptionally in both. This can inform decision-making for the federal government, as well as the state governments.

# 3 Conclusion

Careful analysis of the analyzed datasets yields a number of relevant insights. A number of correlations are identified between various rankings and the directly measured attributes of universities who received those rankings. Distributions of scores are visualized, and an elite-tier university model becomes apparent, where there exists a clearly separated top-tier of universities in many ranking systems. SMU's individual scores are analyzed, and specific ranking improvement methodologies are identified. The United States' universities are distributed geographically, and geographic trends in school ranking and top school density are identified.

Although there is no clear model for how rankings are generated, and no clear-cut path to generating top-ranked universities, many valuable inferences are made from the datasets provided.

# References


[1] THE Times Higher Education Rankings. *timeshighereducation.com*, THE World Rankings, 2016.

[2] Academic Ranking of World Universities. *shanghairanking.com*, Shanghai World Rankings, 2015.

[3] CWUR | Center for World University Rankings. *cwur.org*, Worlds Top Universities, Rankings by Country, 2015.

[4] Education Attainment Query. *datatopics.worldbank.org*, Barro-Lee Dataset, UNESCO Institute for Statistics, 2013.

[5] National Center for Education Statistics. *nces.edu.gov*, Digest of Education Expenditure Statistics, 2011.

[6] Achim Zeileis and Gabor Grothendieck (2005). zoo: S3 Infrastructure for Regular and Irregular Time Series. Journal of Statistical Software, 14(6), 1-27. URL http://www.jstatsoft.org/v14/i06/

[7] "Background Facts on Contingent Faculty." *AAUP*. American Association of University Professors, n.d. Web. 12 Sept. 2016.

[8] THE Citation Data, "Citation Averages, 2000-2010, by Fields and Years." THE. Times Higher Education, 22 May 2015. Web. 13 Sept. 2016.

[9] Daniel Adler (2005). *vioplot*: Violin plot. R package version 0.2. http://wsopuppenkiste.wiso.uni-goettingen.de/ dadler

[10] Stephens, Paul. "International Students: Separate but Profitable." Washington Monthly. Washington Monthly, 05 July 2016. Web. 18 Sept. 2016.

[11] Taiyun Wei and Viliam Simko (2016). corrplot: Visualization of a Correlation Matrix. R package version 0.77. https://CRAN.R-project.org/package=corrplot

[12] "U.S. Department of Education Database of Accredited Postsecondary Institutions and Programs." U.S. Department of Education Database of Accredited Postsecondary Institutions and Programs. DOE, n.d. Web. 18 Sept. 2016.

[13] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.

[14] Original S code by Richard A. Becker, Allan R. Wilks. R version by Ray Brownrigg. Enhancements by Thomas P Minka and Alex Deckmyn. (2015). maps: Draw Geographical Maps. R package version 3.0.1. https://CRAN.R-project.org/package=maps