A Novel Analysis of Collegiate Ranking Data

Identifying University Attributes which Correlate with Ranking

Ian Johnson

Southern Methodist University

September 11, 2016

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

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 which will be used 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 idenfity what to focus on in an effort to increase ranking.

2 Data Understanding

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

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

- **world_rank** *interval*: 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

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

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

- **world_rank** *interval*: 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** *interval*: 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

2.1.4 Supplimentary Educational Attainment and Expenditure Data

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.

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 will be 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 supplimentary 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 dimensionality 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 ommitted.
- 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 will 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

firstlook

2.4 Attribute Visualizations

vis

2.5 SMU: A Case Study

SMU

2.6 Attribute Relationships

relats

2.7 Geographic Relationships

geography

References

[1] Michel Goossens, Frank Mittelbach, and Alexander Samarin. *The LaTeX Companion*. Addison-Wesley, Reading, Massachusetts, 1993.