A Novel Analysis of Collegiate Ranking Data

Identifying University Attributes which Correlate with Ranking

Ian Johnson
Southern Methodist University

September 15, 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

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** *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 [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** *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 [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 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

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

What this table 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.

This table 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 explaination 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. The following table 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
```

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
```

A novel look at the medians 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 ommited 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
```

The first insight from this matrix 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 Shanhai 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 this matrix 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
```

The Shanghai data, it seems, 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 following table 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
```

A cursory glance at the medians 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.

2.3.3 CWUR Data

For the CWUR data, the final simple table dataset for this report, the same strategy will be used to gain some cursory insight into the data. The following matrix represents the year-by-year averages for every attribute in the CWUR dataset.

```
  year alumni_employment quality_of_faculty publications influence citations
1 2012            75.390             56.930       55.020    54.890    54.420
2 2013            75.910             56.060       54.670    56.280    53.930
3 2014           363.991            188.002      500.411   500.163   447.349
```

```
4 2015           406.536            194.253      500.419   500.275   451.334
  broad_impact patents    score
1     371.3195  63.650 54.94090
2     377.2275  63.550 55.27120
3     496.7350 448.968 47.27141
4     496.6640 491.674 46.86385
```

From a quick glance at the matrix, it looks like 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, the following matrix 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
```

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 supplimentary attainment and expenditure data was reduced to single-dimensional data with an artificial (unitless) 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 as follows:

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

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 as follows:

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

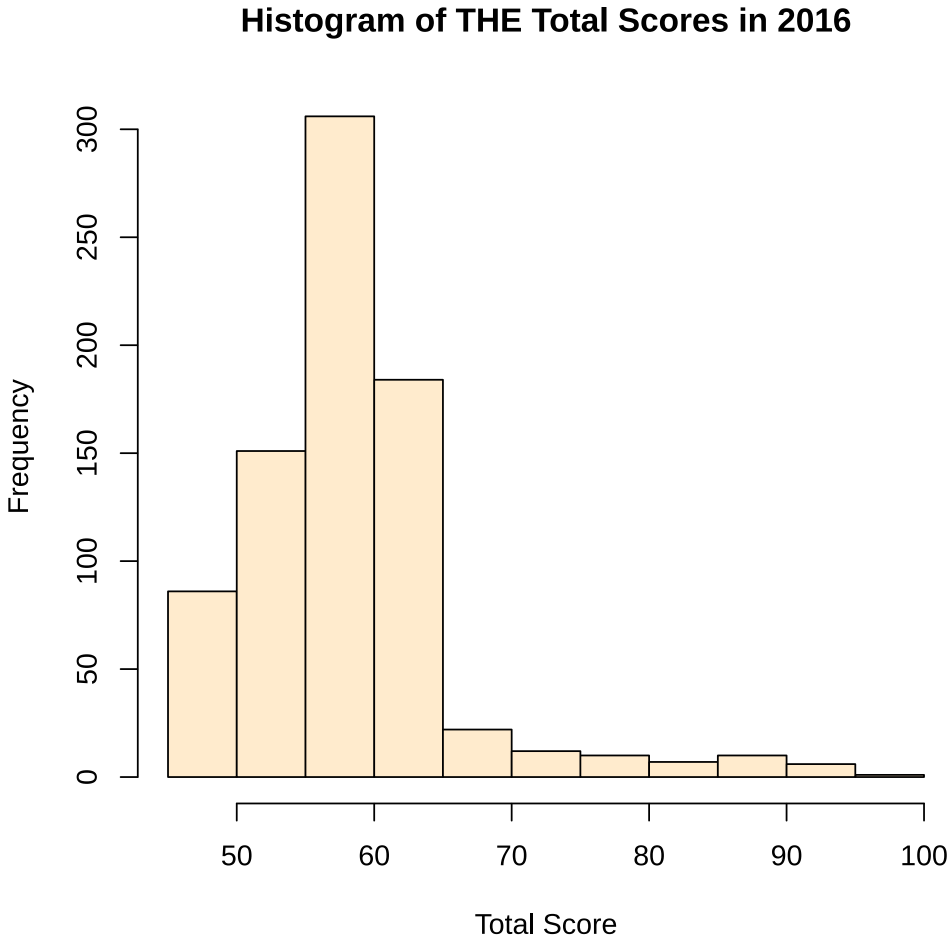
This data, like the attainment dataset, is a two column table, where the first column is the name of a country and the second column is an aggregate public university expenditure score, computed from the multi-year educational expenditure dataset [5].

2.4 Attribute Visualizations

2.4.1 Times Higher Education Data

THE Data Total Score Distribution

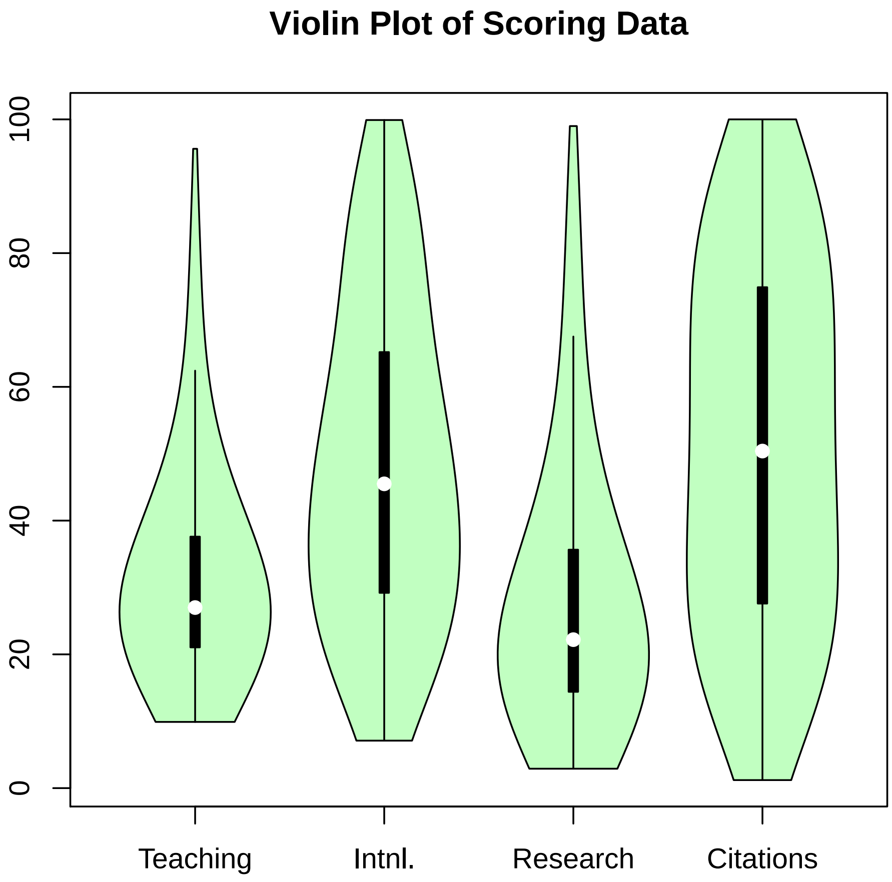
The first important metric to visualize for the THE dataset is the distribution of total scores for the schools surveyed. The following chart is a histogram of the total scores of all schools included in the 2016 THE collegiate ranking report.



What the above histogram indicates is 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 will be used to visualize the teaching, international, research, and citation attributes of each university. The violin plot is built using the CRAN package 'vioplot' [9].



The violin plots of the score distributions 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

2.4.3 CWUR Data

2.4.4 Supplimentary Educational Attainment and Expenditure Data

2.5 SMU: A Case Study

SMU

2.6 Attribute Relationships

relats

2.7 Geographic Relationships

geography

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