Norberto Romero
6.1 - Sourcing Open Data

THE World University Rankings

Summary & Justification

The Times Higher Education World University Rankings (THEWUR) provide a ranking of world universities with an emphasis on research. It uses several performance indicators ranging from staff-student ratios to research citations and volume of publishings. The data contains rankings from 2011-2024 and spans the entire globe. The earliest available data contained 200 universities where the most recent year has over 2500 entries measured across similar metrics.

This data provides multiple years of analysis and allows users to compare global universities, across time, using the same metrics. The data was accessed through Kaggle, but is also available on the organization's website, along with the methodology for each year data was collected.

Kaggle & Times Higher Education | Methodology

Data Source

This is an external data source. THE is an organization that provides insights and analysis on higher education across the globe. They are gathering data about other institutions' performance and references to those institutions. The data is frequently referenced, and uses information that is also publicly available. The organization's reputable history and transparent practices lead us to trust this data source.

Data Collection

The data comes from several places and is collected in different ways. First, THE allows for universities to self-submit and report data directly to them. This data undergoes automated validation checks and comparison to open data sources to ensure accuracy. Data from the UK is provided by the government agency HESA/JISC and then confirmed by representatives from the institutions.

The research data used in the rankings is supplied by a company named Elsevier. This company analyzes bibliometric data from millions of published journals and provides results to
THE who then benchmark and compare the data.

Academic Reputation data comes from survey data that is regularly sent out by THE to scholars across the globe and then analyzed by the company. The survey addresses different categories and is weighted with reference data obtained.

Reference Data is obtained from several UN and governmental data sources.

Data Contents

The data provides a raw score and ranking for each of the 5 main categories used to calculate university rank along with an overall score and rank. Additionally, the data provides descriptive information and statistics about the school such as where it is located, number of students, gender breakdowns, subjects offered, accreditation, and other information.

Data Relevance

This data gives a comparative summary of the five categories Teaching, Research, Citations, Industry Income, and International Outlook when comparing world universities. These five categories provide several ways to analyze and compare countries and areas of the world. The inclusion of geographic data allows for spatial analysis. Multiple years of available data also allows for a time-analysis.

Data Profile

Variables and Data Types

'rank_order' ;	qualitative	time-variant	ordinal
'rank'	qualitative	time-variant	ordinal
'name'	qualitative	time-invariant	nominal
'scores_overall'	quantitative	time-variant	discrete
'scores_overall_rank'	quantitative	time-variant	discrete
'scores_teaching'	quantitative	time-variant	continuous
'scores_teaching_rank'	quantitative	time-variant	discrete
'scores_research'	quantitative	time-variant	continuous
'scores_research_rank'	quantitative	time-variant	discrete
'scores_citations'	quantitative	time-variant	continuous
'scores_citations_rank'	quantitative	time-variant	discrete
'scores_industry_income'	quantitative	time-variant	continuous
'scores_industry_income_rank'	quantitative	time-variant	discrete
'scores_international_outlook'	quantitative	time-variant	continuous
'scores_international_outlook_rank'	quantitative	time-variant	discrete
'location'	qualitative	time-invariant	nominal
'stats_number_students'	quantitative	time-variant	discrete
'stats_student_staff_ratio'	quantitative	time-variant	continuous
'stats_pc_intl_students'	quantitative	time-variant	continuous
'stats_female_male_ratio'	quantitative	time-variant	continuous
'stats_proportion_of_isr'	quantitative	time-variant	continuous
'aliases'	qualitative	time-invariant	nominal
'subjects_offered'	qualitative	time-variant	nominal
'closed'	qualitative	time-variant	binary
'unaccredited'	qualitative	time-variant	binary

Data Integrity Issues

- 1761 rows with missing scores (overall, citations, industry_income, international_outlook, research, teaching). These records are from 2022-2024 where some schools reported information but did not rank. They are labeled ‘Reporter’ in the ‘rank’ column
- Ranks above 200 are grouped into ranges for the ‘rank’ column
- Ranks tied for a position contain an equal sign that make inconsistent datatypes

- 1 row with a missing location.
- 1803 records from 2011-2015 are missing stats
    - 1 row with a missing ‘stats_student_staff_ratio’ from a later year.
    - 620 rows have missing ‘stats_female_male_ratio’ from later years.
- 34 rows with missing subjects
- 191 rows have ‘-’ as the ‘scores_industry_income’ value.
- 11739 rows with missing ‘stats_proportion_of_isr’. All years before 2024 are missingLarge percentage of ‘scores_overall’ are score ranges(i.e.: 9.7-22.7) and make it inconsistent data types.
- Data type for ‘stats_number_students’ was mixed with str and NaN values

Changed/Fixed Records:

- Dropped ['scores_overall', 'aliases', 'closed','stats_proportion_of_isr'] columns.
- Removed ‘Reporter’ schools from the full database (1761 records)
- Split the ‘stats_female_male_ratio’ column into two, showing percentages of each gender respectively.
    - Dropped original ratio column
- Deleted the 'stats_proportion_of_isr' because of too many missing values, over 13k.

Summary:

- 14235 rows & 23 columns

Qualitative:

- subjects_offered
    - Values: 58110 records
    - Mode: Computer Science
- name
    - Values: 14235 records
    - Mode: Harvard University
- location
    - Values: 14235 records
    - Mode: United States

Quantitative:

- rank_order
    - min: 1
    - max: 19060
    - mean: 6257.217
    - median: 5650.0
    - mode: 301
- scores_overall_rank
    - min: 1
    - max: 19060
    - mean: 6258.3171
    - median: 5650.0
    - mode: 10
- scores_teaching
    - min: 8.2
    - max: 99.7
    - mean: 30.058349
    - median: 25.5
    - mode: 17.7
- scores_teaching_rank
    - min: 1
    - max: 1906
    - mean: 667.7183
    - median: 591.0
    - mode: 1
- scores_international_outlook
    - min: 0
    - max: 100
    - mean: 51.797
    - median: 51.849
    - mode: 0
- scores_industry_income
    - min: 0
    - max: 100
    - mean: 44.9552
    - median: 40.1
    - mode: 0
- scores_industry_income_rank
    - min: 1
    - max: 1906
    - mean: 667.7183
    - median: 591.0
    - mode: 105
- scores_research
    - min: 0.8
    - max: 100
    - mean: 26.040077
    - median: 19.8
    - mode: 9.2
- scores_research_rank
    - min: 1
    - max: 1906
    - mean: 667.7183
    - median: 591.0
    - mode: 2
- scores_citations
    - min: 0.7
    - max: 100
    - mean: 51.217724
    - median: 51.2
    - mode: 100
- scores_citations_rank
    - min: 1
    - max: 1906
    - mean: 667.7183
    - median: 591.0
    - mode: 8

- year
  - min: 2011
  - max: 2024
  - mean: 2019.7220
  - median: 2020
  - mode: 2024
- stats_number_students
  - min: 25
  - max: 1824383
  - mean: 23365.075
  - median: 17821.5
  - mode: 9928
- stats_student_staff_ratio
  - min: 0.3
  - max: 865.8
  - mean: 18.897708
  - median: 16.3
  - mode: 14.9
- stats_pc_intl_students
  - min: 0
  - max: 92
  - mean: 11.31093
  - median: 7.0
  - mode: 1.0
- stats_female_pct
  - min: 0
  - max: 100
  - mean: 50.117051
  - median: 53
  - mode: 55
- stats_male_pct
  - min: 0
  - max: 100
  - mean: 49.882949
  - median: 47
  - mode: 45

Data Limitations & Considerations

From 2011-16 the stats columns were not part of the analysis, and have missing values. This is not a concern because many schools have multiple years of data and trends can be inferred from the available data when time series analysis is conducted.

Additionally, the criteria for ranking schools has been altered throughout the years. It has always been research heavy and the THE organization does a good job maintaining records of its methodology and updates it has made. The criteria in this time period was objectively applied to all schools equally, thus making it fair to use for more detailed analysis.

The data was collected through volunteered information, public information, government data collection efforts, and a third party bibliographic data source. The data is aggregated at the university-level, so no PII is involved. This is a reliable dataset and should continue to be used.

Key Questions

1. Which countries have the most schools in the top 200 global universities?
2. Does gender distribution contribute to academic achievement?
3. Where are the countries where most works are cited from?
    a. Do one or two universities dominate that, or is it a competitive ranking?
4. Which countries produce the top earners in their respective industries?
5. Which schools have the best teachers?
    a. What are the staff student ratios?
    b. How many students attend these schools?
6. Which countries accept the most international students at their universities?