Which university is ‘the best’?

Gan Ruyi 3035637905

Structure

Problem Definition

- Data science questions to answer
- Why the questions are important
- Difficulties on answering the questions
- Data available for answering the questions
- Existing works

Results

- Methods used
- Findings
- Key solutions and highlights

Conclusion & future works

- Answers to the questions
- Possible future works and directions

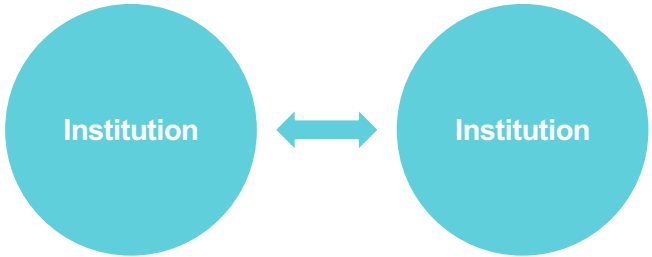
Acknowledgements and References

Data science questions to answer

Factor Analysis

1. Why are the rankings varied so much according to different **ranking factors** and which one is better?

Probablistic Modelling



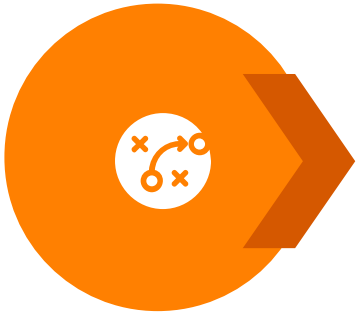
2. What are the **possible strengths** of each institution according to various attributes?

Why the questions are important?

Topic:
Popular

The question of which university is the best has always been an endlessly debated and controversial topic.

Problem:
Confusing



Although many ranking systems have been developed and posted for students to choose their desired university, their ranking criteria are not clear enough, and their widely varied rankings can sometimes become more confusing.

Aspiration:
Effiencient

Therefore, I'm coming up with the idea to compare the different ranking systems and clearly state some critical thoughts on their criteria, advantages, and why they receive potential criticisms.

Difficulties on answering the questions

Data

Existing datasets are too old

01

02

Web scraping

Challenging

Information Alignment

Troublesome with too many details

03

04

Modeling

Limited by my regression & machine learning knowledge

Existing works

Data available for answering the questions

1. Here is the dataset containing world university rankings from 2012 to 2016 from Kaggle, which includes 6 different tables: THE, CWUR, ARWU rankings, and supplementary data on educational attainment and expenditure.
https://www.kaggle.com/datasets/mylesoneill/world-university-rankings?datasetId=27&sortBy=voteCount&language=R&select=cwurData.csv
2. Here is the dataset of THE from Kaggle:
https://www.kaggle.com/datasets/r1chardson/the-world-university-rankings-2011-2023
3. Here is the dataset containing qs rankings from 2017 to 2022 from Kaggle,
https://www.kaggle.com/datasets/padhmam/qs-world-university-rankings-2017-2022.
4. However, these datasets are incomplete, and more information should be grasped from the official websites:
https://www.timeshighereducation.com/world-university-rankings
https://www.topuniversities.com/university-rankings

Existing Works

1. There are some existing works on world university ranking data visualization, comparison, data analysis, and modeling on Kaggle.
2. More detailed references can be seen from the 'Acknowledgements and References' part.

Methods

Platform

R Studio

Version

R 4.2.2

Libraries: Data management

Tidyverse, dplyr, tidyr…

Libraries: Visualization

Ggplot2, plotly, ggrepel, gplots, corrplot…

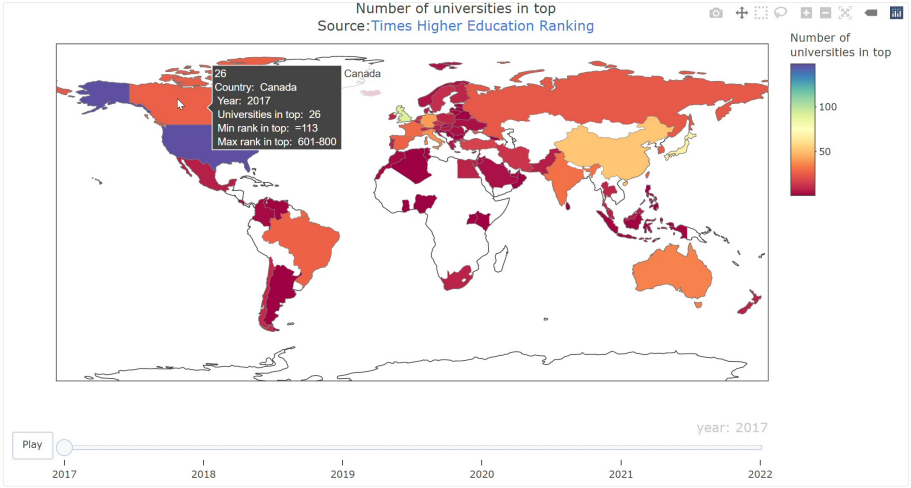
Libraries: Web Scraping

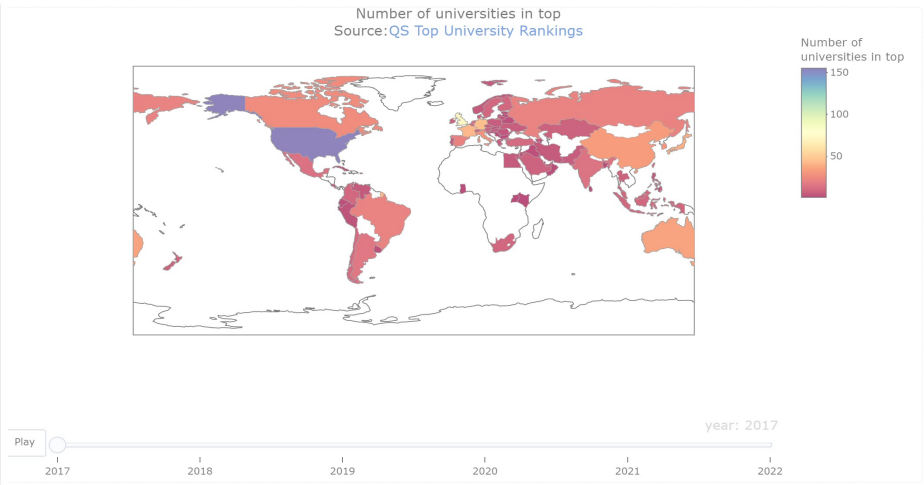
Rvest…

Libraries: Factor Analysis

Stats…

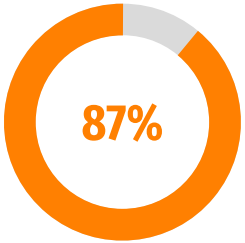
Findings: Comparison & Visualization



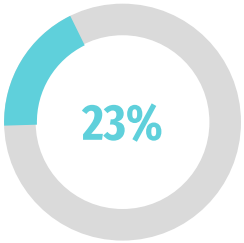


Increase Rate (China)

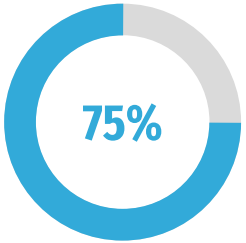
Increase Rate (UK)



Increase Rate (mainland China)

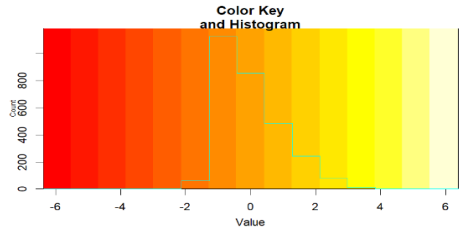


Increase Rate (UK)



Combination:

HeatMap & dendrogram



Y: Universities

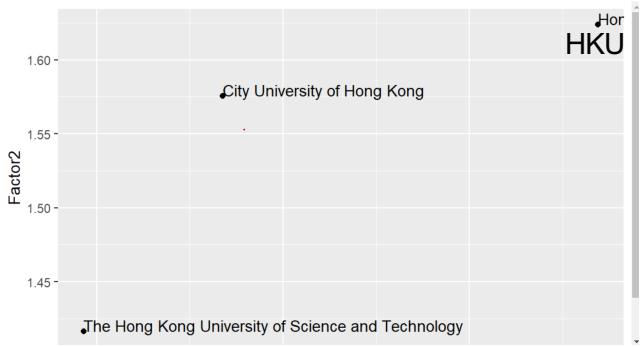
X: Factors on Different Ranking Systems

Highlights: Factor Analysis

Table

	factor1	factor2	factor3
ti_scores_overall_rank	**0.963186386382816**	-0.244903527148565	0.0857824266469249
ti_scores_teaching_rank	**0.659491983886174**	0.0770423070711403	**0.616024696686295**
ti_scores_research_rank	**0.752318017592238**	-0.123886503535311	0.38507901844214
ti_scores_citations_rank	**0.832310862543658**	-0.249494064990921	-0.274817673929056
ti_scores_industry_income_rank	0.151322671628744	0.175377207641551	**0.428837933454003**
ti_scores_international_outlook	-0.179911380578139	**0.977589238924892**	0.0834702453171591
ti_scores_international_outlook_rank	0.207689755063185	**-0.956619603884458**	-0.0758441086085277
ti_stats_student_staff_ratio	-0.113371732081103	0.189830117931534	0.257687363933498
qs_score	**-0.536930715349169**	0.271758679819003	**-0.57897733108039**
qs_student_faculty_ratio	0.00680548751092073	-0.0149688012306455	**0.490914237923292**

Plots



01 PCA Analysis

02 Factanal()

03 fa_load
v.s. fa_rotate_load

04 Factor Analysis
& visualisation

Conclusions

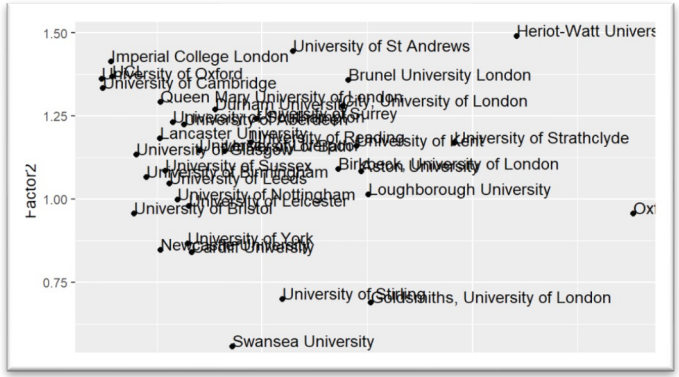
Insight 1

factanal(univ_num,factors=3,rotation="none")

```
                Factor1 Factor2 Factor3
SS loadings       3.515   1.917   1.194
Proportion Var    0.351   0.192   0.119
Cumulative Var    0.351   0.543   0.663

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 242.6 on 18 degrees of freedom.
The p-value is 2.6e-41
```

We can use factor analysis to judge the performance of different universities according to specific criteria.



Compared with QS rankings, THE teaching, research, citations may take an important role in line with university ranking values.

Insight 2

Possible Future Works

Bayesian Plackett-Luce Rankings Model

Use Machine Learning Models to propose a new ranking system according to analyzed important factors.

Cluster of Institutions

Now the factor analysis is quite on the surface and can be further developed.
Based on components analyzed today, observe and find different institutions in the top clusters.

Acknowledgements and References:

1. POZDNIAKOV, 2016. https://www.kaggle.com/code/pozdniakov/which-universities-do-good-science
2. GABRIEL PREDA, 2019. https://www.kaggle.com/code/gpreda/world-university-rankings-advanced-analysis?scriptVersionId=7833309
3. JEREMY LEIPZIG, 2016. https://www.kaggle.com/code/leipzig/factor-analysis-of-times-and-cwur-sets
4. APOLLO_STAR, 2017. https://www.kaggle.com/code/apollostar/bayesian-plackett-luce-rankings-model
5. Datasets used in the presentation are listed before in the existing work part.