

World University Ranking 2022

台北大學統碩一 商巧昀

專案簡介

THE及QS分別是兩大世界大學排名機構，其排名依據各有不同特色，但皆參考各校方向？的學生人數作為排名依據。

本專案以學生人數為中心，分為兩個部分：

1. **差異比較**：觀察各校提供給兩大世界大學排名機構的學生人數是否有差異，又不同族群(排名區間、排名進步與否、地區)的學校的人數差是否會有很大的差異？
2. **資料轉換**：若兩個機構將資料做轉換，讓公正的第三方機構以不知道原始資料為何的情況下比較人數差異以判斷哪些學校的人數差距較不合理，如何轉換較好？

專案流程

差異比較

QS_his.R:
抓QS2021, 2020學校名稱&排名
(rank_table_2020.csv · rank_table_2021.csv)

重複: 共29筆重複。如果有學校的欄位不齊全, 會抓不到資料, 這筆資料會完全被前一筆全覆蓋, 也就是前一筆重複兩次。因欄位不齊而被覆蓋沒有抓到的學校以南美學校(16間)為主, 其次為歐洲(8間)。

QS.R:
1. 抓QS2022學校網址(href_table_2022.csv)
2. 進href_table_2022每個網站抓各校資料(qs_total_2022.csv)
repeat.R:
3. 刪除重複筆資料(qs_total_2022_不重複.csv)

THE.R:
1. 抓THE2022各校資料(the_total_2022.csv)
repeat.R:
2. 刪除重複筆資料(the_total_2022_不重複.csv)

重複: 共一筆重複, 為中國和美國的Northeastern University, 在THE的名稱一樣, (保留排名前面的美國Northeastern University)。

THE_his.R:
抓THE2021, 2020學校名稱&排名
(rank_table_2020.csv · rank_table_2021.csv)

rise_or_not.R:
1. 合併qs_total_2022_不重複.csv & rank_table_2020.csv & rank_table_2021.csv
2. 修改格式('%'改小數點, 刪千分位數...)
3. 計算2021相對2020上升或下降(代表變數名稱: qs_up)。

continent.R:
1. 讀入qs_total_2022_不重複.csv及the_total_2022_不重複.csv
2. 修改格式('%'改小數點, 刪千分位數...)
3. 合併QS及THE表成一個table(m)

region.R:
QS網站藉由篩選地區, 抓各洲的學校名稱(europe.csv, latin_america.csv, ... *6)

rise_or_not.R:
1. 合併the_total_2022_不重複.csv & rank_table_2020.csv & rank_table_2021.csv
2. 修改格式('%'改小數點, 刪千分位數...)
3. 計算2021相對2020上升或下降(代表變數名稱: the_up)。

rise_or_not.R:
1. 合併QS及THE表成一個table(m)
2. 計算diff及pdiff比較兩種機構提供的人數資訊
3. 以qs_up, the_up變數分組, 比較組別間的差異

group_rank.R:
1. 將THE的rank變數分為'1-200', '201-400', '401-600', ..., '1200+'共7組
2. 計算diff及pdiff比較兩種機構提供的人數資訊
3. 比較組別間的差異

continent.R:
1. 比對各洲的學校名稱與m的學校名稱欄, 並新增continent變數
2. 計算diff及pdiff比較兩種機構提供的人數資訊
3. 利用continent變數分組, 比較洲組別間的差異

專案流程

資料轉換

重複: 共29筆重複。如果有學校的欄位不齊全，會抓不到資料，這筆資料會完全被前一筆全覆蓋，也就是前一筆重複兩次。因欄位不齊而被覆蓋沒有抓到的學校以南美學校(16間)為主，其次為歐洲(8間)。

QS.R:

- 1.抓QS2022學校網址(href_table_2022.csv)
- 2.進href_table_2022每個網站抓各校資料(qs_total_2022 .csv)

repeat.R:

- 3.刪除重複筆資料(qs_total_2022_不重複.csv)

THE.R:

- 1.抓THE2022各校資料(the _total_2022.csv)

repeat.R:

- 2.刪除重複筆資料(the_total_2022_不重複.csv)

transform.R:

- 1.讀入qs_total_2022_不重複.csv及the_total_2022_不重複.csv
- 2.因為沒有要轉換其他資料，因此只取學生人數來修改格式(刪千分位)
- 3.資料轉換(ranking及z score)
合併QS及THE表成一個table(m)

重複: 共一筆重複，為中國和美國的Northeastern University，在THE的名稱一樣，(保留排名前面的美國Northeastern University)。