Exploratory data analysis of QS world university rankings

Sandeep Parmar, Lakshika, Rishiraj Sutar, Ringan Majumdar

Under the guidance of Dr. Dootika Vats

Introduction

From the very childhood we have often heard many of our acquaintances who are going USA to pursue his PhD or Germany to pursue masters, so it came naturally to us, "Does India not have any good universities? Why do these people need to go abroad for pursuing higher studies? Isn't our country sufficient?". Moreover, most of the Nobel laureates belonged to these universities like Harvard, MIT or Cambridge. Hence, there grew a curiosity for these Universities. So, we decided to scrape the data of the best universities in the world and analyse what factors drive them to be the best and where are we lacking behind?

For this we choose QS World University Ranking.

QS University Rankings

QS World Ranking is an annual publications of university rankings by *Quacquarelli Symonds*. The QS ranking receive approval from the International Ranking Expert Group (IREG). According to Alexa internet, it is the most widely viewed university ranking worldwide.

Scraping of the data

We have scraped the QS university ranking data from its official website https://www.topuniversities.com/. This website also has a source page, we were unable to scrape data from this using web scraping. After some research, we found that this problem can be solved using API scraping.

What is API?

API stands for **Application Programming Interface**. APIs are mechanisms that enable two software components to communicate with each other using a set of definitions and protocols. Basically, we request the server (which provides data to the website) to provide the data we want.

For our project following libraries are required to extract data.

1) Httr
2) jsonlite
3) dplyr
4) stringr

Code of scraping

In the first line, we request the server using get() function in httr library. The link shown in GET function is basically not the main website link. This specific link contains data in text format. It can easily be found in Chrome developer tools for the main website. We extract data in JSON format and convert it into data frame which we require. Here is some highlight of main code of scraping.

```
library(httr)
library(jsonlite)
library(dplyr)

link <- "https://www.topuniversities.com/sites/default/files/qs-rankings-data/en/914824.txt?rk1fb0"

response <- GET(link)  # Request server
data_json <- content(response,encoding="UTF-8")
data <- jsonlite::fromJSON(data_json) # Data in JSON format
df <- data.frame(data) # Final scraped data frame
```

Cleaning of data

Now comes data cleaning part. In this we create new data frame and insert our 12 desired variable in it from JSON data. After we mainly clean the string of "HTML code".

For example, look the following string.

" Massachusetts Institute of Technology (MIT) "

In this string our main target is to extract name of universities.For this we use **stringr** library.

CSV file data

Besides this, we also downloaded a CSV file of each country's GDP per capita income for the years 2020,2021 and 2022 from the World Bank's official website. And merge this data with the latitude-longitude data of the country.

Description of our different data-sets

Our project is based on the QS World University Rankings, which are updated yearly. This data mainly consists of the academic reputation, employer reputation, faculty per student ratio, citations per faculty, and the number of international students, which, combined, result in the university's overall score. Based on the score given, the rank of the university is formulated. Also, the city, country and continent where the university is located are given to get an idea of the regional disparity. For our project, we have scraped the data for 4 years: 2020, 2021, 2022, and 2023. There are around 1000 observations for each year with 12 variables. Talking about the variables used in the dataset :

1. **ACADEMIC PEER REVIEW:** This variable is the academic reputation indicator. This is calculated based on a survey in which active academicians are asked to vote for a maximum of 30 universities, excluding their own. However, this is the most controversial part of the methodology. In the overall score, it consists of about 40% weightage.

2. **FACULTY STUDENT RATIO:** To measure the teaching commitment.This consists of 20% weightage.

3. **CITATIONS PER FACULTY:** This represents the citations of published research per faculty. Using citations in world university ranks is still questioned because arts and humanities generate comparatively few sources. Also, the sites used for citations have a language biases, biased towards English research papers. Weightage in the overall score is 20%.

4. **EMPLOYER REPUTATION**: This is based on a review of the employer. Similar to academic review, this is based on a survey of recruiters who hire graduates on a global level. Weightage is 10%.

5. **INTERNATIONAL STUDENT RATIO:** Measure of the diversity of the student community. The weightage of this is 5%.

6. **INTERNATIONAL FACULTY RATIO:** Measure the diversity of the academic staff. The weightage of this factor is 5%. These play a significant role in calculating the overall score and ranking. Hence, using all these variables as the basis for the rankings, QS calculates each university's overall score using a formula that uses all these factors. The dataset we are scraping here is essential in that we should be aware of where we stand in the global world right now, how it has changed over the years, and what we can do to improve?

Some Interesting Questions

After analyzing our clean data-sets we have posed a few questions which we have thought can be interesting. The questions are as follows,

1. What is the proportion of the universities for the different continents? Has this proportions changed over the 4 years? Has it truly changed or we are missing something?

2. Does the overall score and its indicators change over for years for a specific university?

3. How have the ranks of the top 10 universities changed over the course of 4 years?

4. How is the overall score affected by its indicators? Do the indicators have any biases?

5. Are the higher ranks clustered around specific locations? If, yes, can we guess a factor responsible for that?

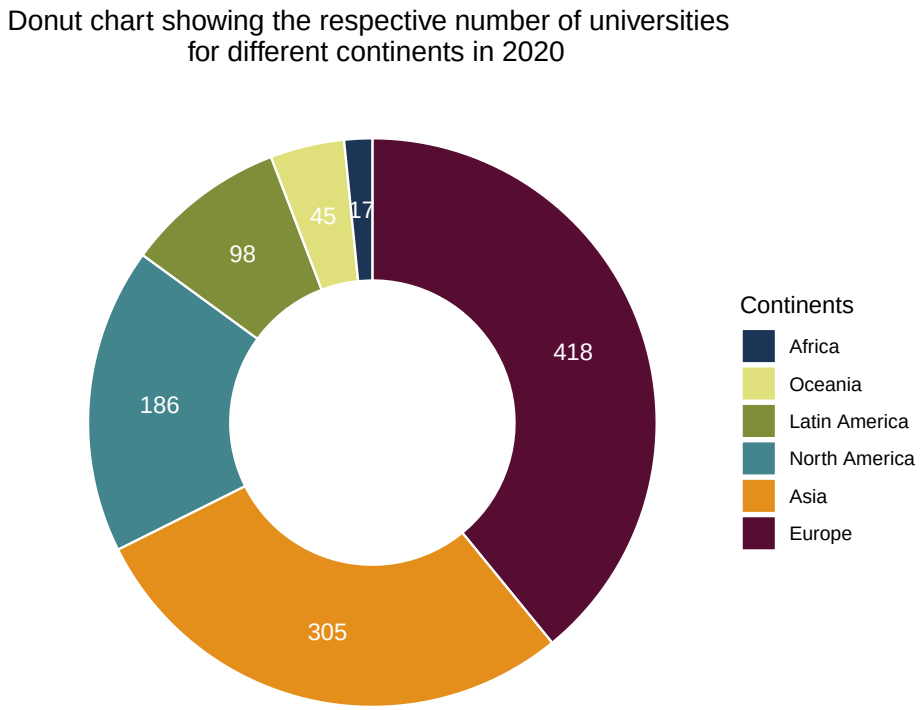
6. Do the richer countries have higher number of QS ranked universities? Why?

Data visualizations

We have posed a bunch of questions after scrutinizing and analyzing our data-sets. Now, we will try to answer as many as questions we can. The best thing we can do now is to perform different types of data visualizations and then analyze our plots. The next part will talk about the different visualizations.

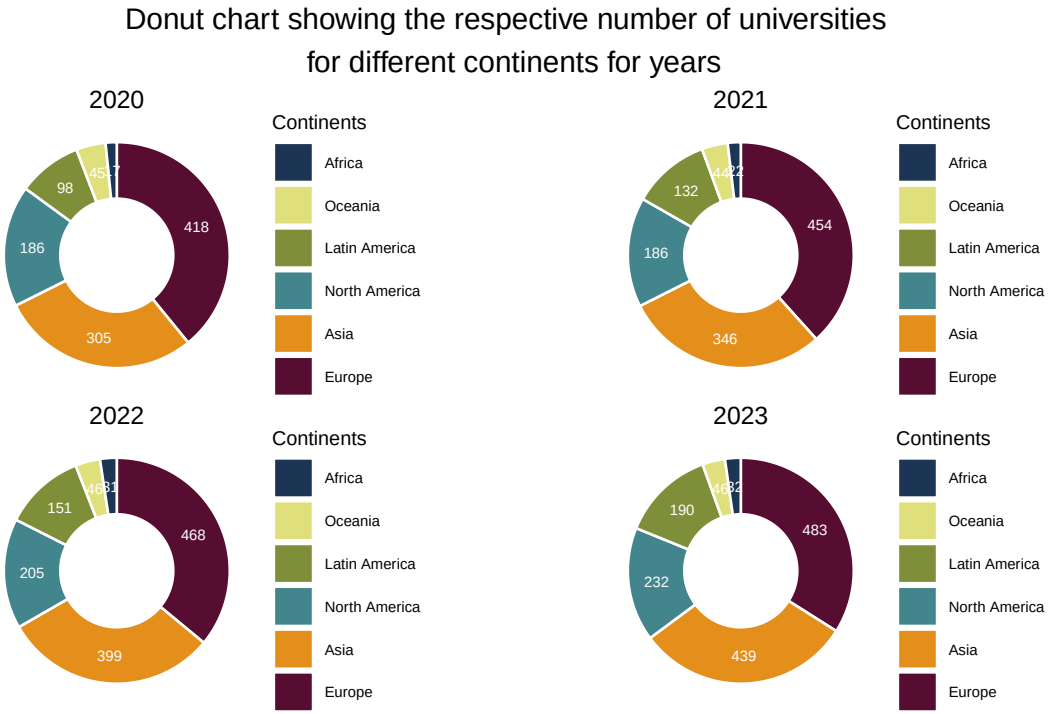
Donut Charts

The very first kind of the data visualization which we have done are donut charts. Let us first plot the frequency distribution of the different universities for different continents of year 2020.



Now, when we observe this chart, we can notice that the maximum number of universities is situated in Europe, followed by Asia and North America. On the other hand the least number of the universities is situated in Africa followed by Oceania and Latin America.

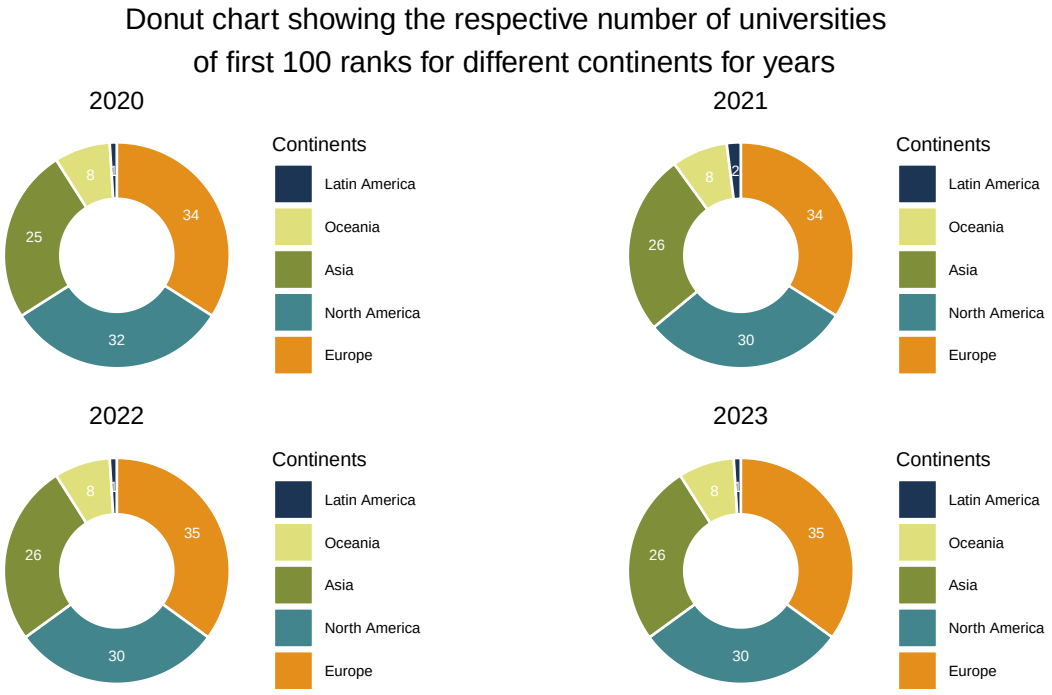
Next, let us plot the donut charts for all 4 years.

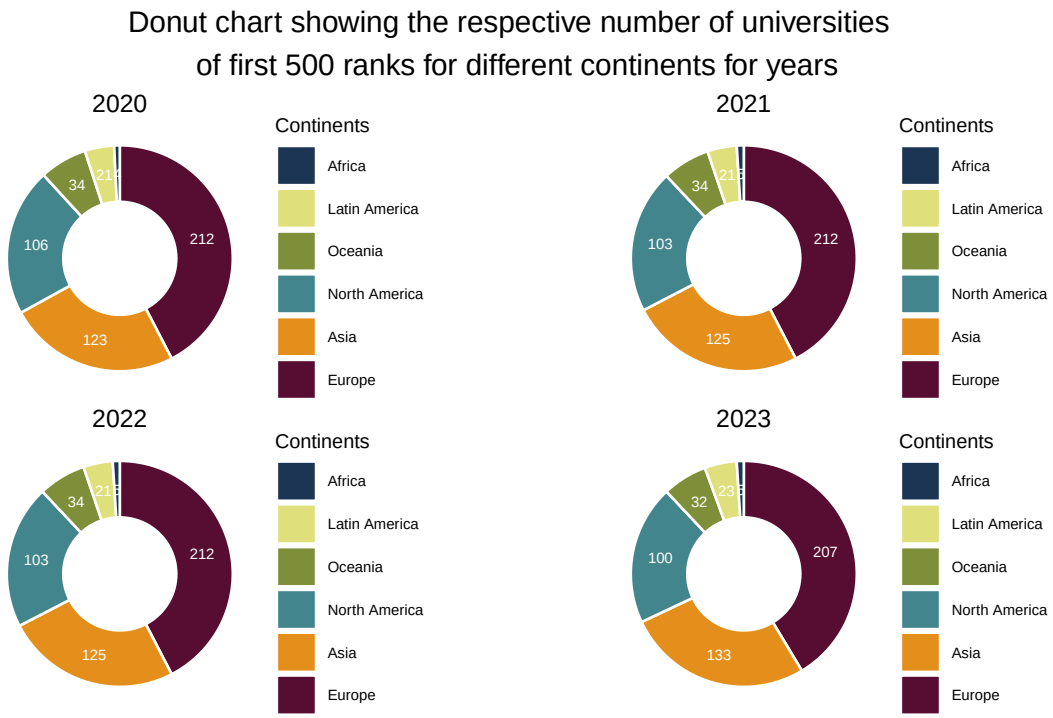


This plots shows that the number of the universities have increased over time for all the continents, be it Africa or Europe.

But this, plot does not make the ranking distribution of the universities clear. In the other words we don't know whether the top ranks are coming from Africa or Europe.

For understanding this let us plot the donuts for the first 100 and the 500 hundred ranks for the 4 years.



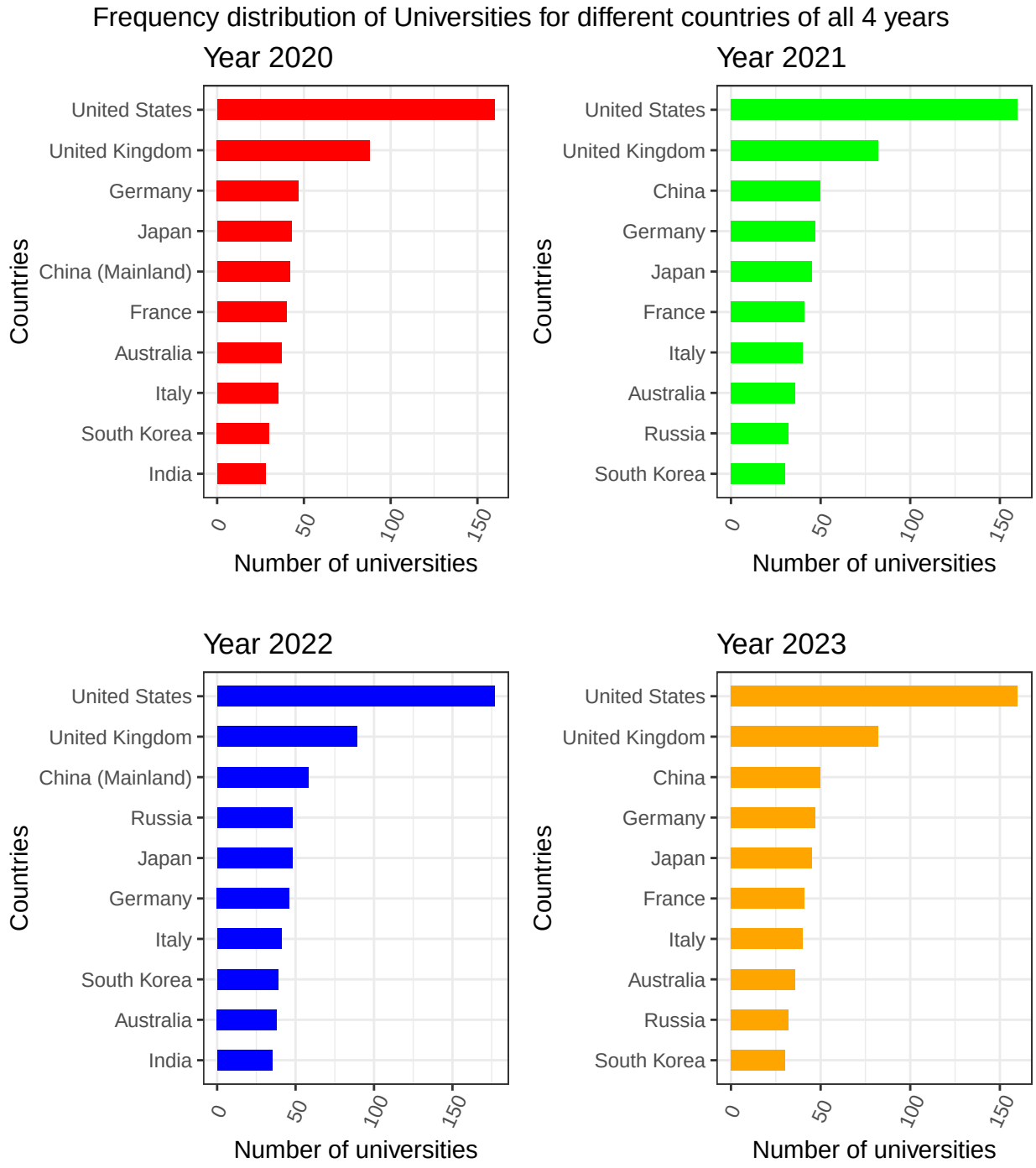


Observe that, in case of first 100 ranks, no a single university is coming from Africa and very few from latin America or Oceania. Most of the world's best universities are situated in Europe, followed by North America and Asia. The same pattern is noticed in the plots for first 500 ranks too.

One interesting observation in this case can be that the proportion of the universities situated in the different continents does not change much over time.

Bar Charts

The next type of visualizations we have done are bar charts. We have plotted the frequency distribution of the universities situated in different countries for all 4 years. We did this because in the donut charts we could see the dominant continents but, her we can see the dominant countries for different continents. One thing thing to note, in this plots only top 10 bars are shown.



Looking at this plot we can observe, the country with the highest number of universities is USA followed by Great Britain. The other positions more or less varies over the 4 years. India came to 10th position in the year 2020 and 2022.

Two interesting observations can be made in this case.

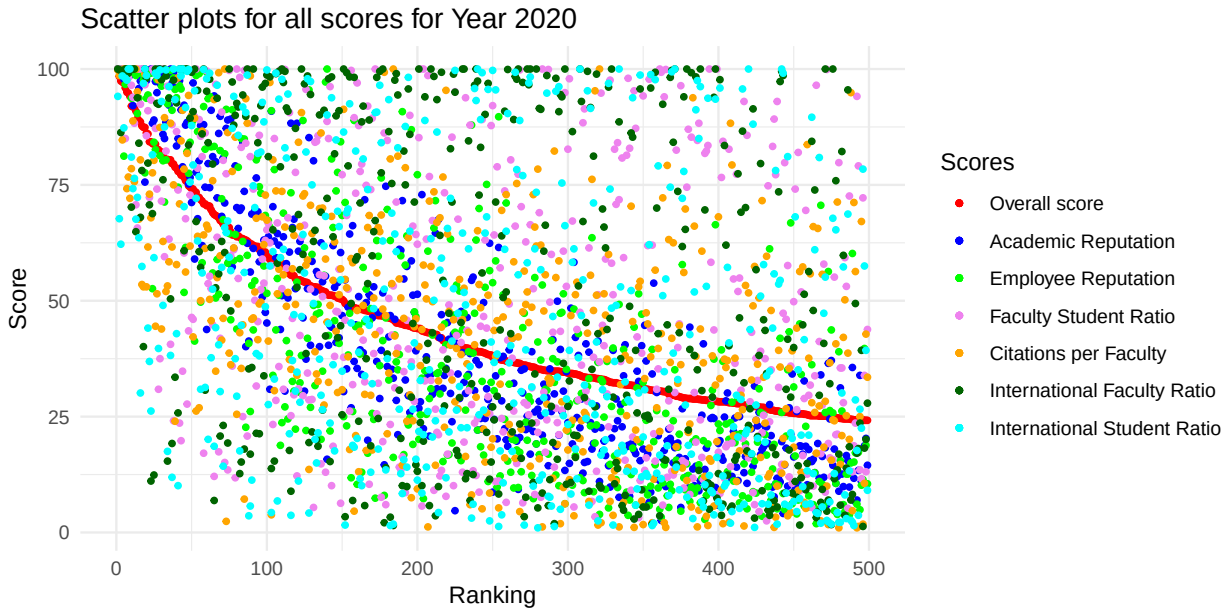
- The pattern of the frequency distribution does not change much over the course of 4 years.
- The countries with high number QS ranked universities tend to be rich countries.

Scatter Plots

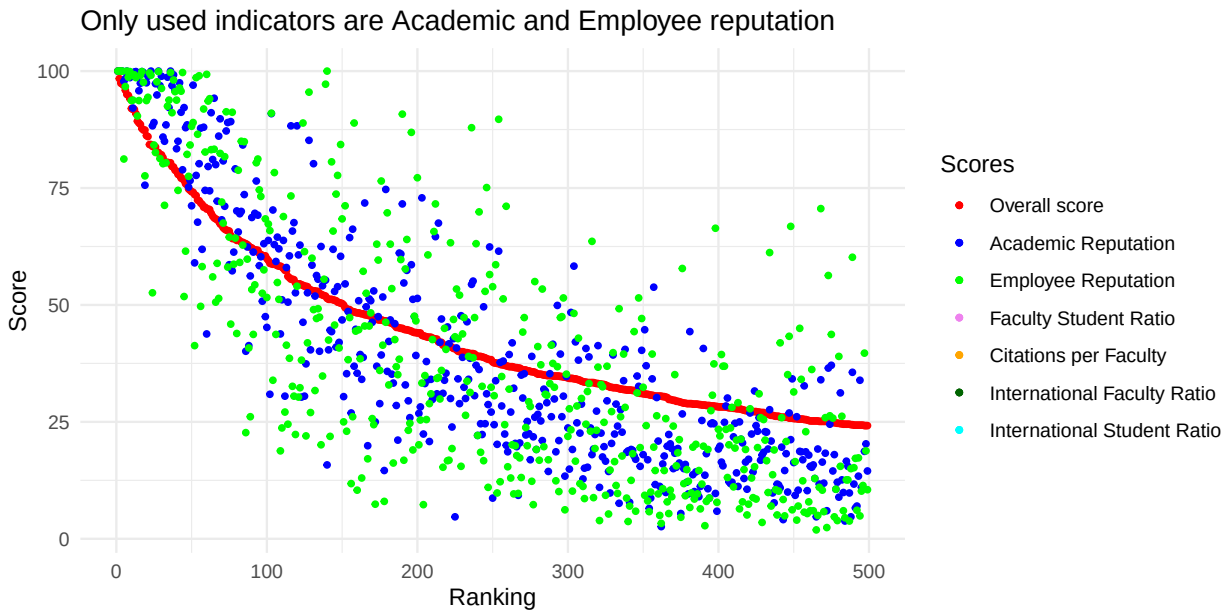
We have data for the overall score of the universities and its different indicators. So, the first kind of visualization that came to our mind is scatter plot. By plotting this we can see whether the scores follow a pattern? So, let us first try to plot the scatter plot for the overall score and all its indicators all overlaid on the same plot.

We faced a problem we tried to plot scores and its indicators for all the universities. The universities which have ranks after 500 didn't have all the indicator values and this occurs for the data of all 4 years. So, we used only first 500 ranks for the scatter plots

This is the scatter plot for the overall score and all its indicators all overlaid on the same plot for year 2020.

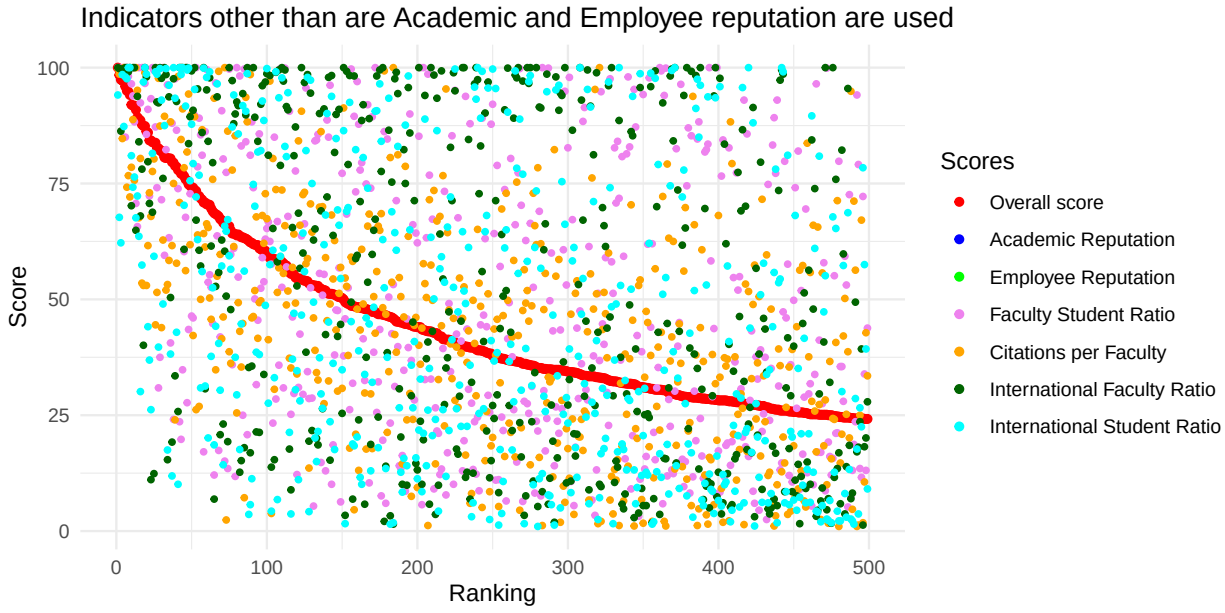


The red points denote the overall score. One can see how the overall scores increases gradually with the ranks. Now, for the indicators it seems completely random at a glimpse. But, when you observe the plot clearly you can see the indicators like Academic Reputation and Employee Reputation are scattered very closely to the overall scores. So let us plot the overall score with Academic Reputation and Employee Reputation for the first 500 ranks.

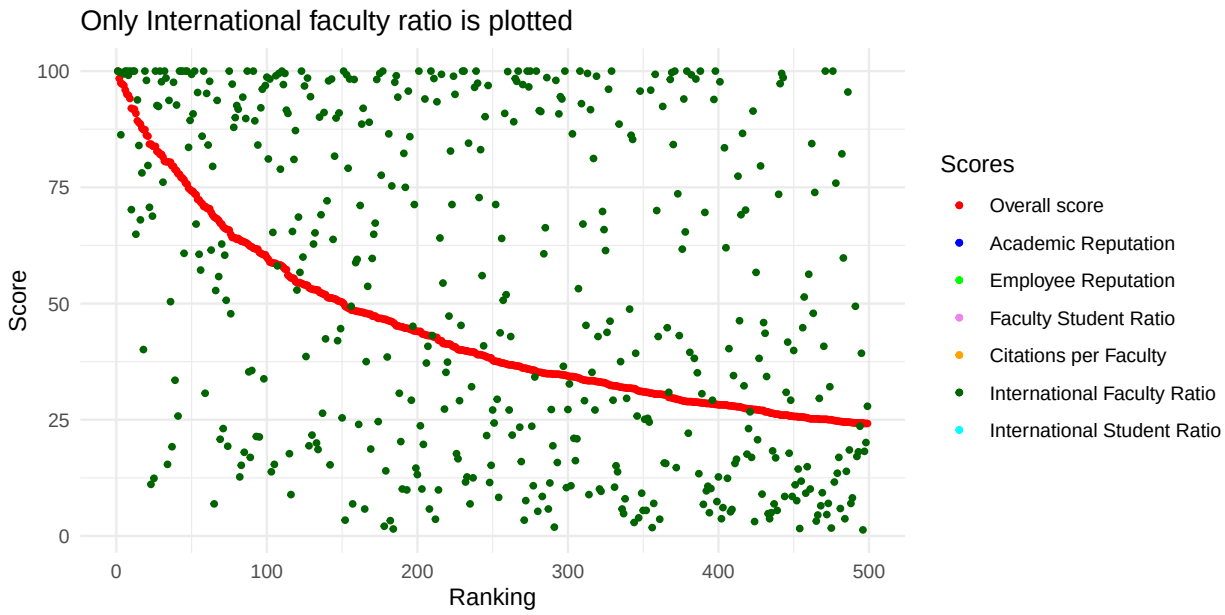


And, yes our guess was right.

Now, let us plot all the remaining indicators with overall score of first 500 ranks for year 2020.



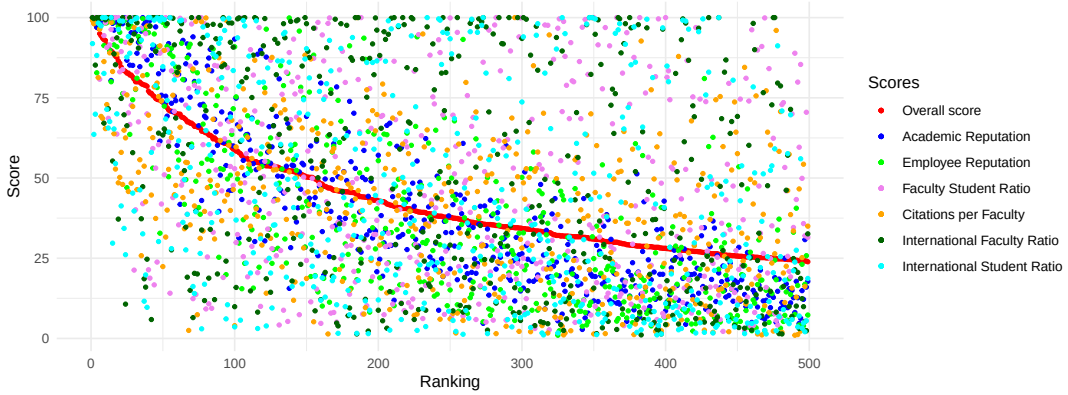
When you look at this plot you can see the indicator International Faculty Ratio is highly clustered around at the top or at the bottom. To see this clearly lets plot overall score with only indicator International Faculty Ratio



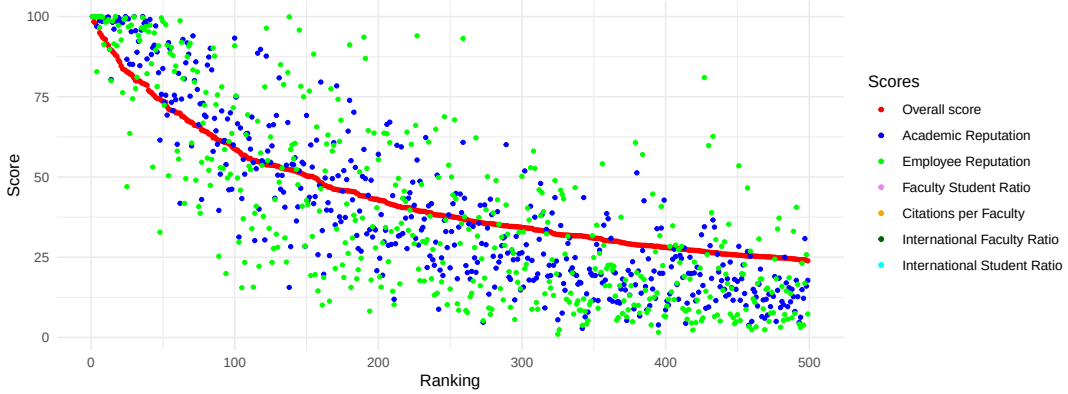
Now, one may think what can be the reason behind this. Why do some universities have so high number of international faculties and yet some of the universities have so low? We were not able answer this only the basis of our data.

Now, the same 4 plots are done for the remaining 3 years

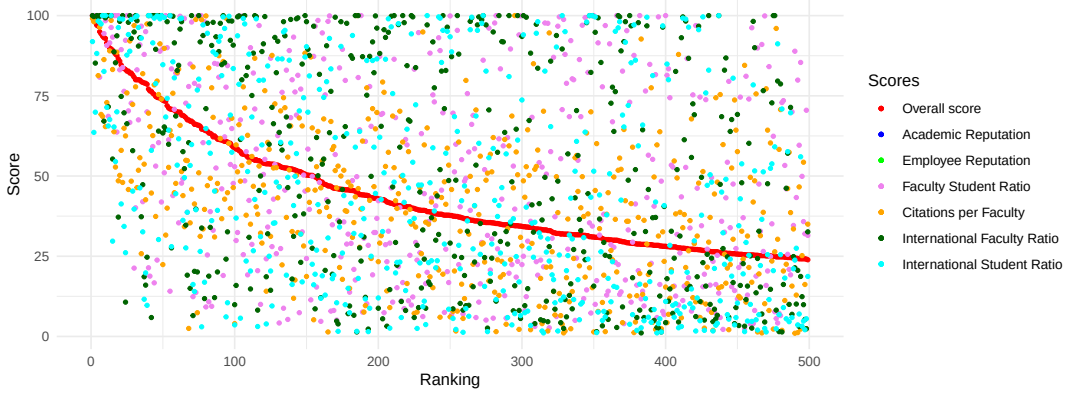
Scatter plots for all scores for Year 2023



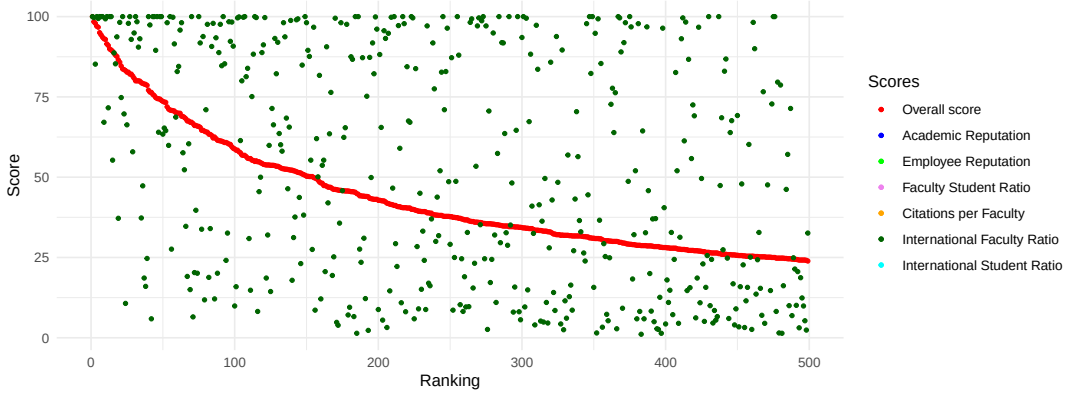
Only used indicators are Academic and Employee reputation



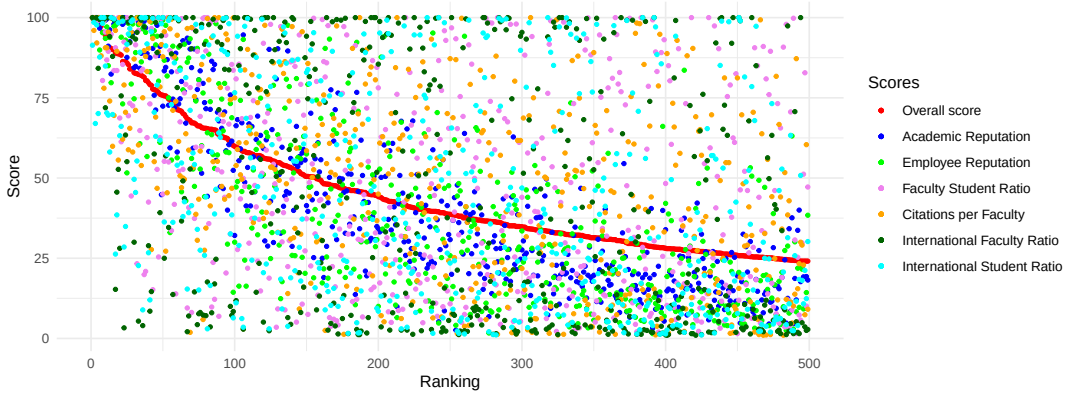
Indicators other than are Academic and Employee reputation are used



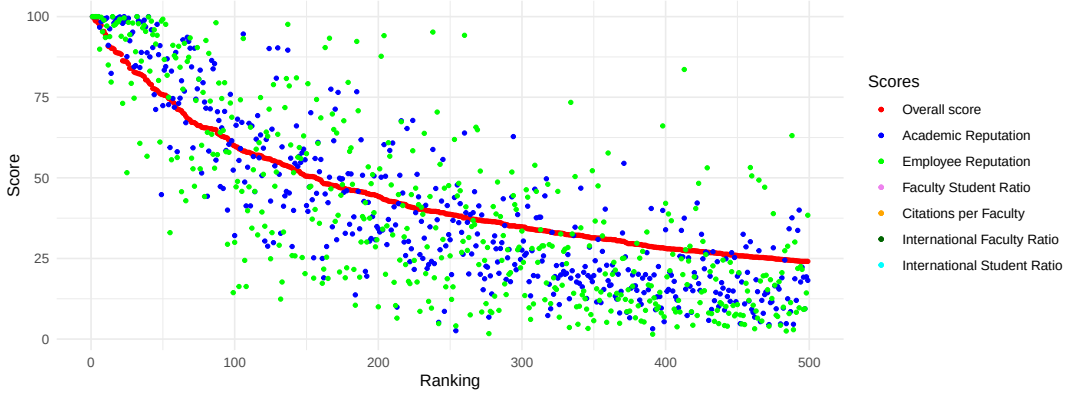
Only International faculty ratio is plotted



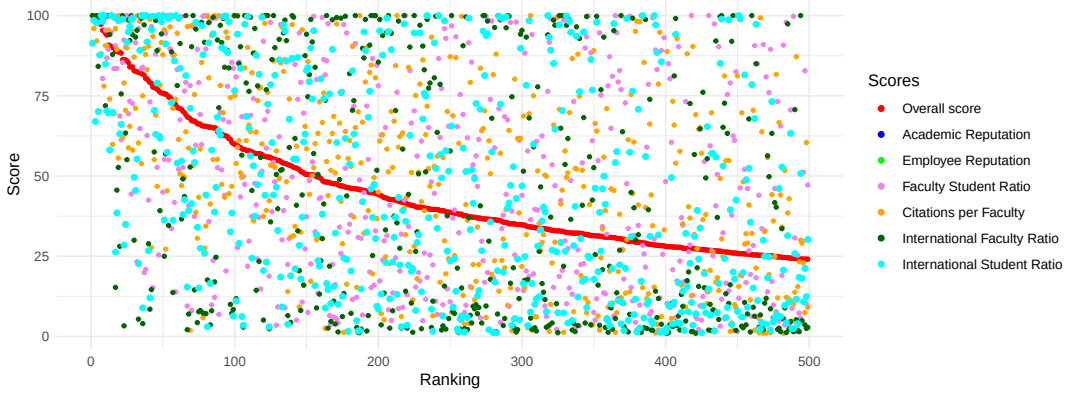
Scatter plots for all scores for Year 2022



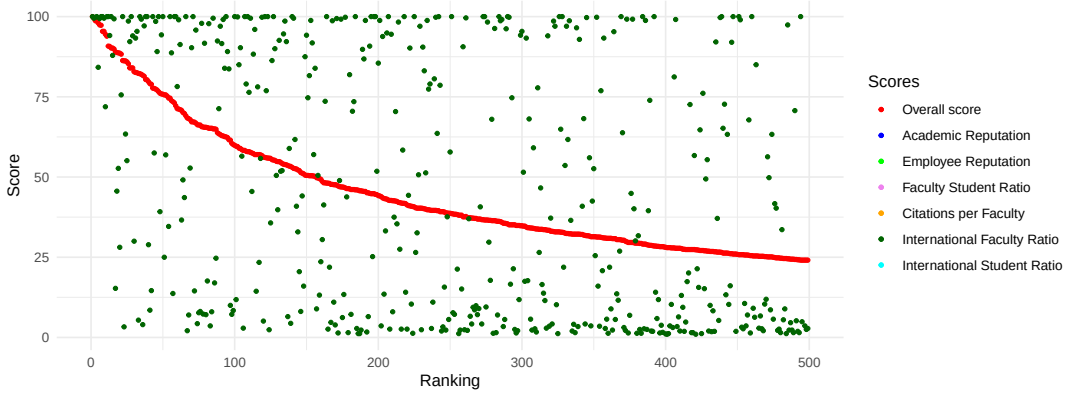
Only used indicators are Academic and Employee reputation



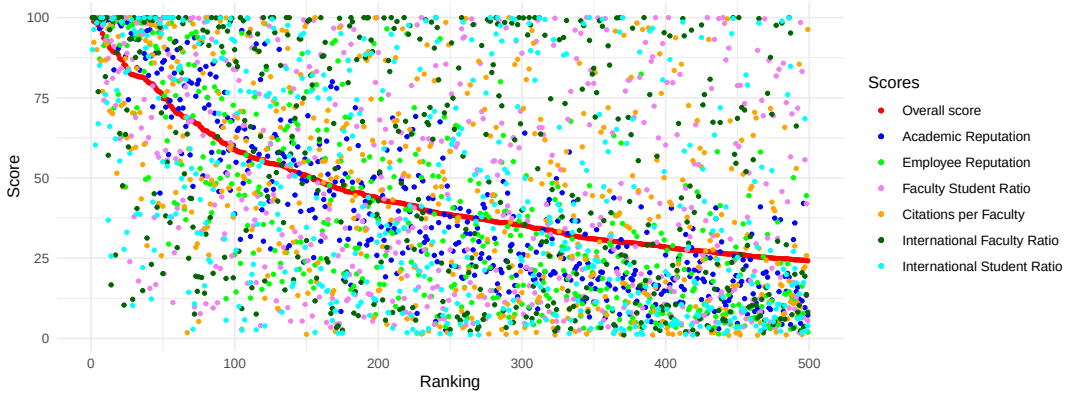
Indicators other than are Academic and Employee reputation are used



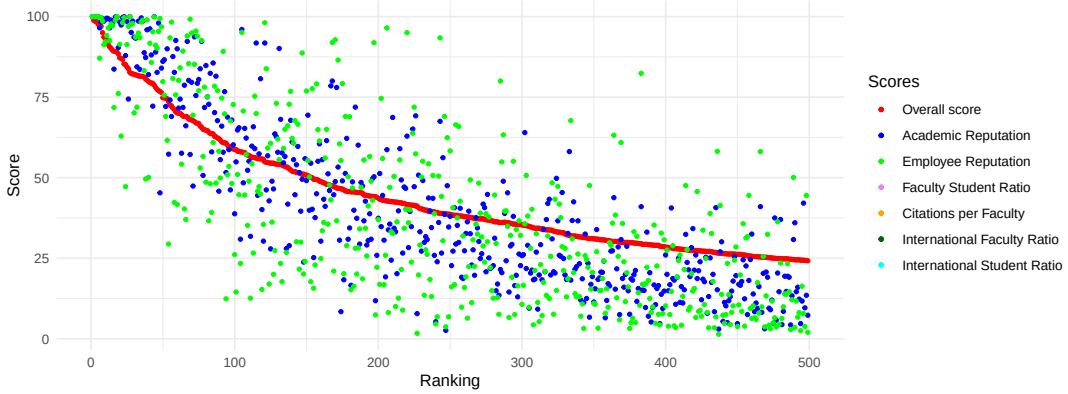
Only International faculty ratio is plotted



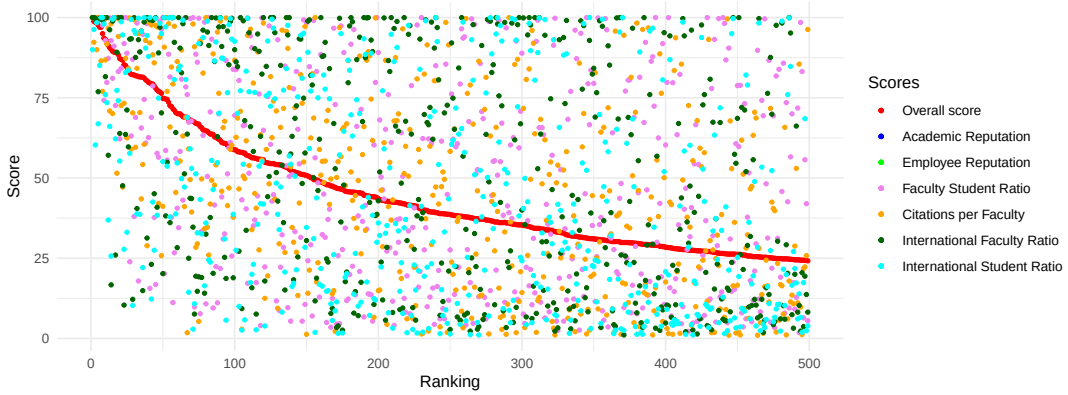
Scatter plots for all scores for Year 2023



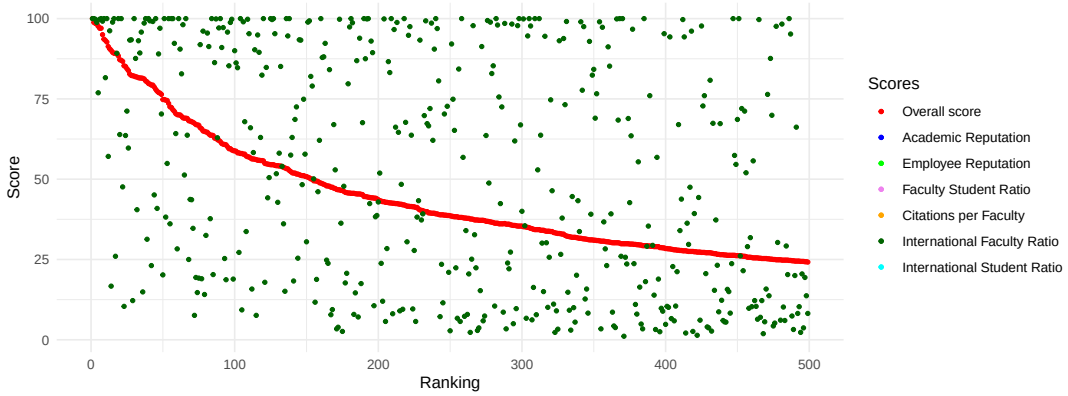
Only used indicators are Academic and Employee reputation



Indicators other than are Academic and Employee reputation are used



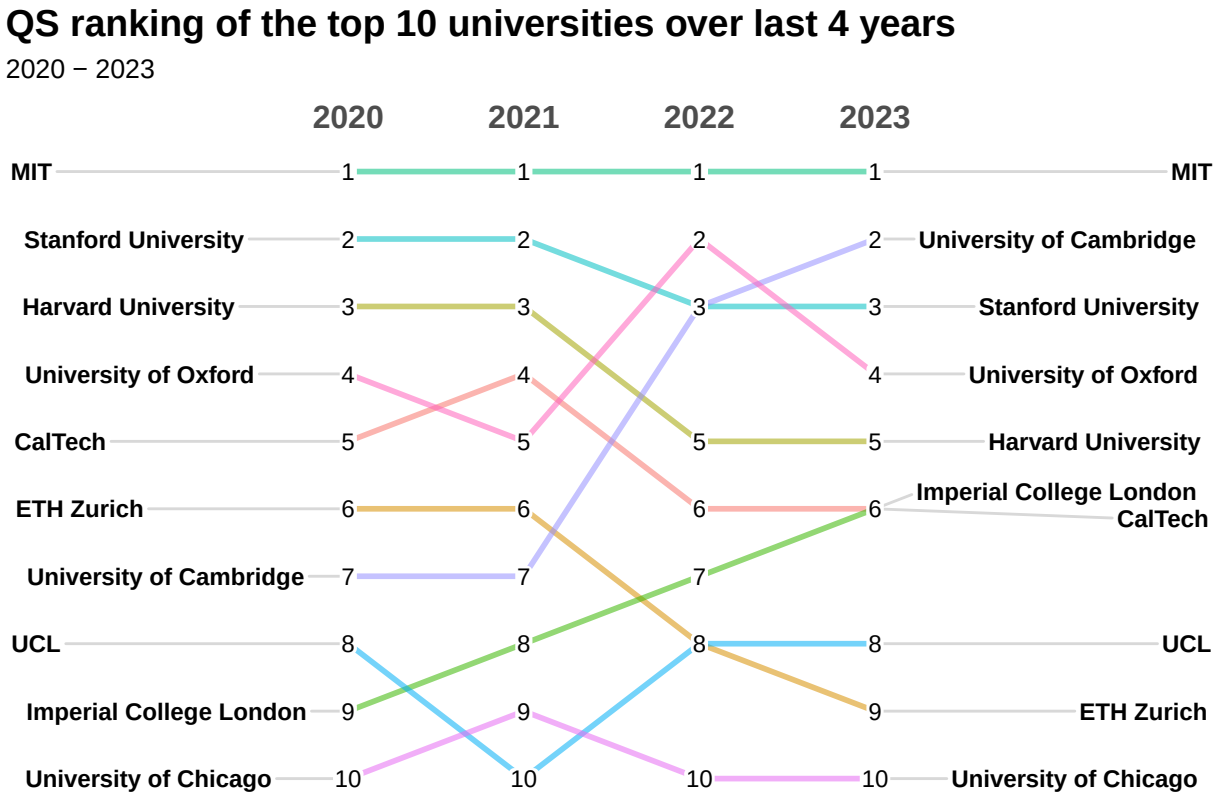
Only International faculty ratio is plotted



Interestingly, here again we see the same pattern for all 4 years.

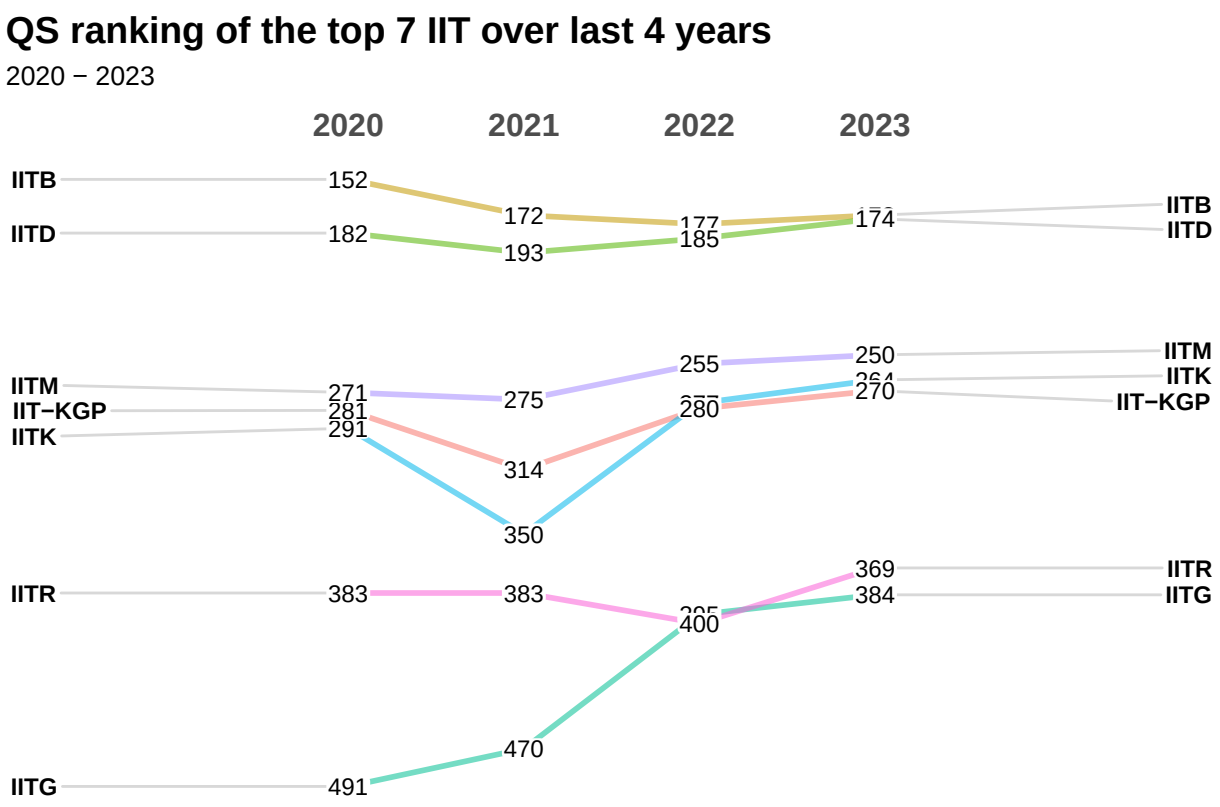
Ranking Comparisons

Next, we have done some a kind of visualization to compare the rankings of the universities. Let us try to visualize ranking of the top 10 universities have changed over the course of the 4 years.



Clearly, MIT always tops the list, the undisputed leader. Sadly Stanford and Harvard university has gone down to ranks 3 and 5 from ranks 2 and 3 respectively. On the other hand, University of Cambridge has gone up to rank 2 from rank 7. Another interesting observation is that the top 10 remained top 10 for all 4 years.

We are IITians, so, we also to tried to compare the ranks of top 7 IITs.

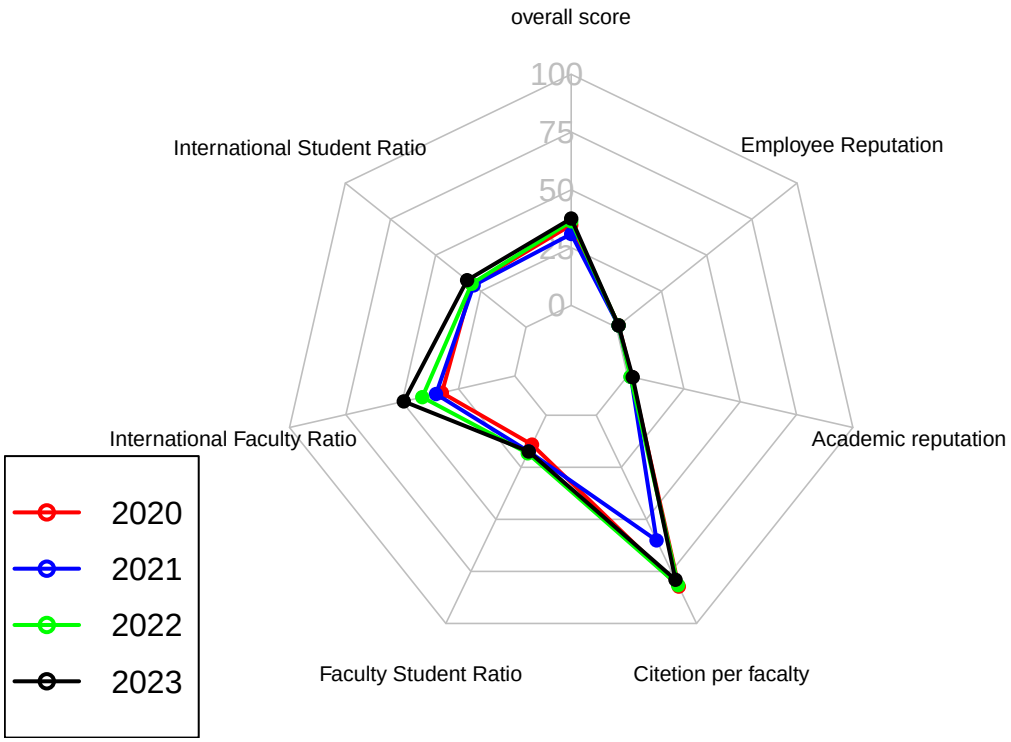


Sadly, not a single IIT ranks in top 100. IITB tops the list for all 4 years but, its rank has gone down a bit over the 4 years. IITK, on the other hand has gone up to rank 4 from rank 5. More or less every IIT has improved its rank with an only exception of IITB.

Radar Charts

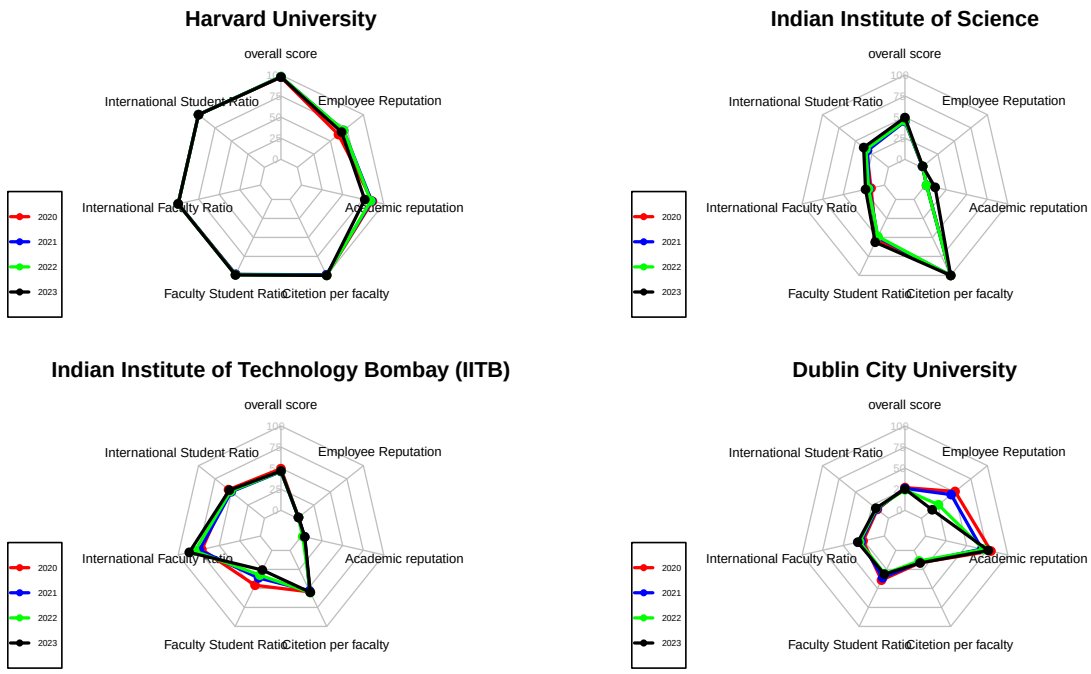
A radar chart is type of data visualization where multivariate data is plotted in a web like structure. Hence, it is also known as spider diagram. The axes of the web represent different variables. So, we have tried to do radar charts for different universities where the variables are score and its indicators. Moreover, we have used the data of all 4 years in the same radar charts.

Let us first try to have the radar chart of IIT Kanpur.



By looking at the radar chart of IIT Kanpur, we can say that the Overall Score and the International Students Ratio has increased over the course of four years. The Institute has always maintained a high Citation per Faculty. On the other hand, we can also observe that Employee Reputation and Academic Reputation has always been low throughout the four years.

Now, let us plot the Radar Chart of the 4 different Universities, namely, Harvard University, Indian Institute of Science, Indian Institute of Technology Bombay (IIT Bombay) and Dublin City University.



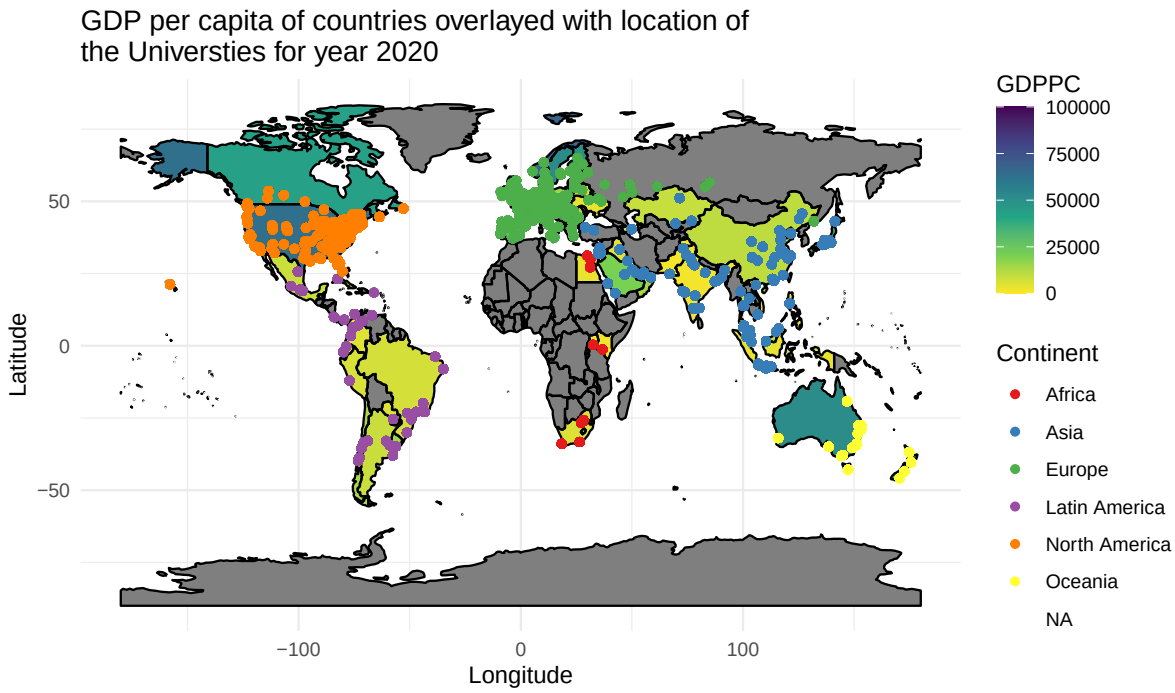
- By looking at the radar chart of Harvard University, we can easily say that it is one of the most prestigious Institute of the World since, throughout the four years it has maintained a high Overall Score, International Student Ratio, International Faculty Ratio , etc.
- Indian Institute of Science has an excellent Citation per Faculty whereas, the International Faculty Ratio has always been low. However, the Academic Reputation has shown improvement in 2023.
- The Indian Institute of Technology Bombay (IIT Bombay) has got a high International Student Ratio and increased even more in 2023.But Employee Reputation and Academic Reputation is quite low and it hasn't any improvement in the four years time.
- Dublin City University has a good Academic Reputation . From the radar chart, we can see that the Employee Reputation has fallen drastically from the year 2020 to 2023.

Maps

The last kind of visualization we have done are maps. We have used GDP per capita data of countries having QS ranked universities and then overlaid it with the location of the universities. We have again faced a problem here. The GDP data for year 2023 is not published yet so we have used GDP data for years 2020, 2021, 2022.

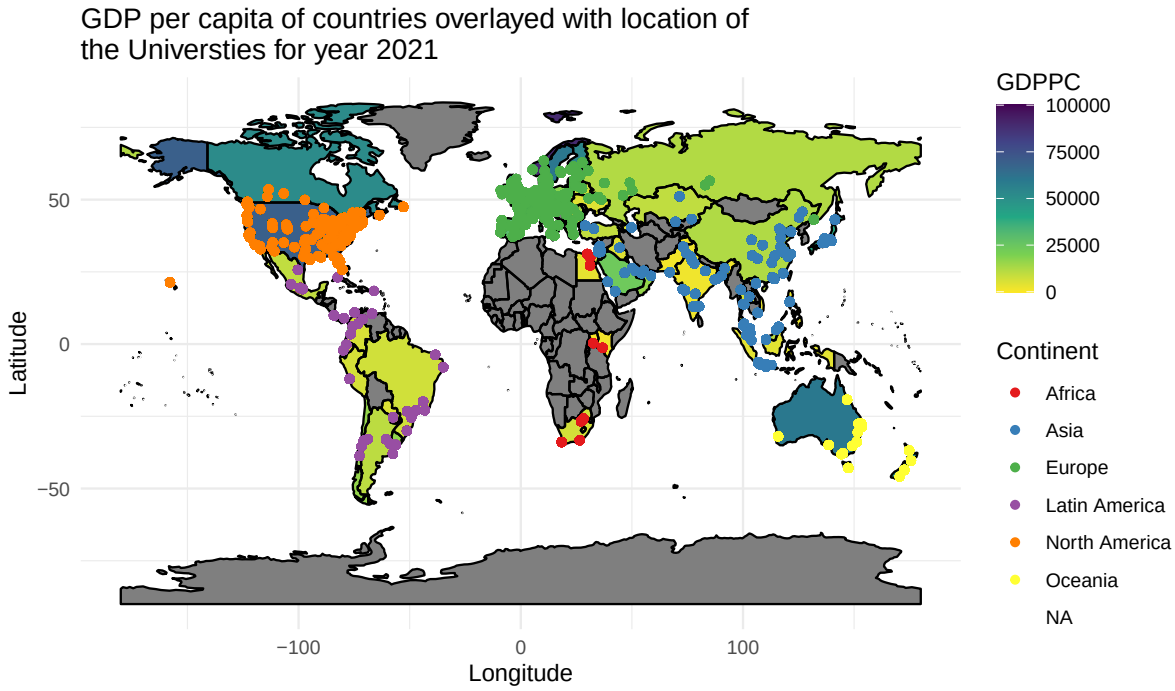
Now the question arises, "Why have we taken the GDP per capita data and not any other indicator?". One answer to this question is that in the previous plots we have noticed the richer countries tend to have a higher number of universities. Again, we know that Money drives everything. The more money a country has the more it can allocate to health or education. The more educated the people become, the more it will produce researchers, scientists or artists. And when a country has a huge number of quality researchers it will produce great professors. Hence, the country will very good universities. These universities will again produce good students who will then become entrepreneurs making the country richer again. This runs like a cycle. We tried to use the data where what percentage of GDP is allocated to a education for a specific country but, we were unable to extract to data for all the countries for a given year.

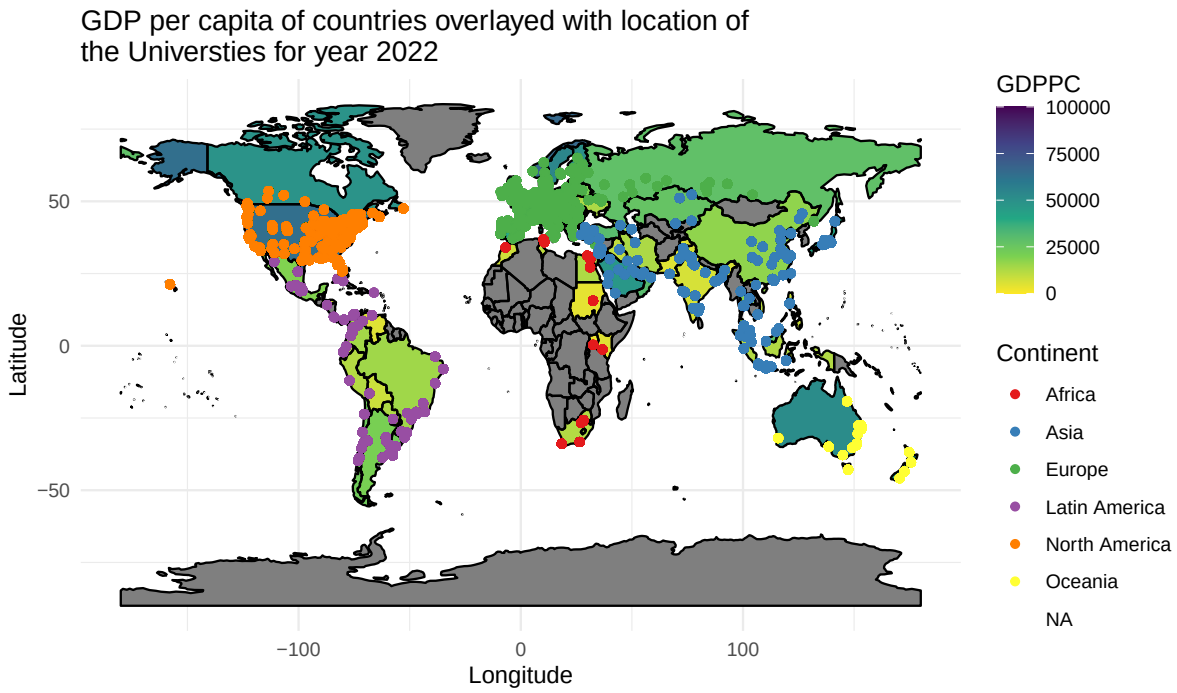
Let us first plot the GDP per capita data of countries having QS ranked universities and then overlaid it with the location of the universities for the year 2020.



It is very clear from the diagram that our guess was right. Points are clustered around dark green areas. Richer countries do have more number of universities. There are indeed high clusters on universities in the place like Europe and USA.

Now let us do the same for the years 2021 and 2022.





Notice, the positions of the clusters are not changing but the countries are getting richer as the shade of the green in getting darker as year changes.

A brief summary of our R shiny

Backend

The shiny app firstly loads all the datasets required and then many libraries like shiny, shinythemes, ggplot2, dplyr are used. It contains navbar which includes different tabs for different visualisations like donut, barplot, radarchart, globe etc. For every different tab, there is different sidebar panel and main panel. Reactive is used to switch the datasets according to the input given by the user and the output is changed accordingly. Then code for different plots is given. There is then the compilation of shiny app using ui and the server.

Front end

Shinythemes library is used in which sandstone theme is incorporated in our app. Every plot varies according to the options chosen. In most of the plots, year can be varies from 2020 to 2023. The radarchart is drawn for a specific university. In the scatter plot, different variables can be selected. The globe which depicts the gdp per capita and the location of universities, can be rotated sideways or upside-down by changing the respective sliders. There is another tab which has ranking comparisons for top 10 universities and the top 7 IITs. Lastly, there is the dataset tab showing the data used in the visualizations. Data will change according to the number of observations and year.

Conclusions

After plotting various graphs and studying the ranks of different universities of the world we conclude that:

- The most frequent pattern which we have observed more or less from all the visualizations is that the nature of the data has not change over the course of 4 years. Hence, we can conclude it takes to have a change.
- The continents like Europe and North America have a overwhelming number of high ranked universities.
- Massachusetts Institute of Technology has always been at the top of the ranking table while Stanford University's rank went down from 2nd to 3rd over the course of four years. Imperial College to London has shown improvements in their ranking as they went up from 9th in 2020 to 6th rank in 2022. But if we see as a whole, there are no major changes in ranking patterns over the four years namely 2020 , 2021, 2022 and 2023. In other words the ranks of universities have north changed drastically
- After plotting the universities in the world maps overlaid with GDP per capita and locations, we can easily come to a conclusion that major bulk of the universities are located in countries like USA, Japan or UK. They all happen to be the few of the richest countries. So, one may conclude GDP per capita is probably affecting the number of Universities from a specific countries.
- By studying the QS Ranking Dataset for the 4 years, we can easily observe that the top Universities of the World with high Overall Score, Academic Reputation, Employee Reputation, International Student's Ratio, etc, are mainly situated in the foreign countries like USA and Europe. This justifies our claim that most of the students and scholars migrate to different countries to pursue high education and research works.

References

- https://www.topuniversities.com/
- https://en.wikipedia.org/wiki/QS_World_University_Rankings
- https://data.worldbank.org/indicator/NY.GDP.MKTP.CD
- Google Developer tools