FIT5147 Data Exploration Project

QS World University Rankings

Rahul Bharadwaj Mysore Venkatesh
31322239
Fahimeh Sadat Saleh, Ying Yang

Introduction

University ranking is a measurable outcome of multiple factors that are considered to evaluate the standard of education, faculty, resources, and infrastructure. Every year, Universities are ranked by different organizations around the world like CWUR, Times Higher Education, Quacquarelli Symmonds (QS) and many others. We are using a dataset that contains rankings of the world universities as maintained by QS.

Quacquarelli Symmonds (QS) is a British think-tank company specializing in the analysis of higher education institutions throughout the world. It uses 6 factors for their ranking framework wiz. Academic Reputation, Employer Reputation, Faculty to Student Ratio, Number of Citations per Faculty, International Faculty, and International Students. Another feature included in this data was Classification (which is not used for ranking) which included the institution's size, subject range, research intensity, age, and status.

This Data Exploration Project is an effort to answer some questions around the analysis of higher education institutions such as the following –

1. What factors other than rank is more desirable when deciding the quality of a University? In other words, how do Universities compare in terms of the 6 factors in QS factor classification?

2. Which Universities top in each of the specific factors?

3. Is there a correlation between different classification factors like Country, Age of the University, Reputation of the University, Size and International Student Numbers?

Data Wrangling – Cleaning and Transformations

- Kaggle link for dataset - https://www.kaggle.com/divyansh22/qs-world-university-rankings.

- It is a Tabular Data: 1K rows x 22 columns. It has simple text in the form of ".csv"

- For the purpose of our analysis, we make use of only the top 100 Universities of the world as it seems to have rich data and people are usually more likely to compare among the top 100.

- The Tools and Technologies used for this project are Microsoft Excel, R & RStudio, Tableau 2020.2

visdat: Visualizing Whole Data Frames in R - by Nicholas Tierney

- visdat() is an R package that has a variety of functions to visualize datasets and give a visual overview of the dataset of interest. We read the data into R environment after filtering top 100 observations in Microsoft Excel and use the clean data.

```
```{r ReadData}
Uni2020 <- read.csv(here::here("Data/Clean/2020-QS-World-University-Rankings.csv"))
```
```

Fig 1: Reading Data into R Environment

- The raw data has all columns in character format, and we convert it into appropriate types using the following procedure in RStudio.

```
```{r Transformation}
TopUni2020 <- Uni2020 %>% select(-Rank.in.2019) %>% mutate(
                    SIZE = as.factor(SIZE),
                    FOCUS = as.factor(FOCUS),
                    RESEARCH.INTENSITY = as.factor(RESEARCH.INTENSITY),
                    AGE = as.factor(AGE),
                    STATUS = as.factor(STATUS),
                    AcademicSCORE = as.double(AcademicSCORE),
                    EmpSCORE = as.double(EmpSCORE),
                    RatioSCORE = as.double(RatioSCORE),
                    CiteSCORE = as.double(CiteSCORE),
                    IntFacSCORE = as.double(IntFacSCORE),
                    IntStuSCORE = as.double(IntStuSCORE),
                    AcademicRANK = as.integer(AcademicRANK),
                    EmpRANK = as.integer(EmpRANK),
                    RatioRANK = as.integer(RatioRANK),
                    CiteRANK = as.integer(CiteRANK),
                    IntFacRANK = as.integer(IntFacRANK),
                    IntStuRANK = as.integer(IntStuRANK),
                    Overall.Score = as.double(Overall.Score))
```
```

Fig 2: Code for Data Transformation using mutate() in R

- We deselect 'Rank.in.2019' column, mutate classification factors to factor type, Score fields to double type and Rank fields to integer type which is suitable for mathematical operations. Operations cannot be performed on character fields and hence, this data transformation in necessary for mathematical analysis of fields.

- Next, we visualize the data using the following commands on R.

```
```{r DataCheck}
#checking the data
vis_dat(Uni2020)
```
```

Fig 3: Dataset Visualization Code

- We can see a visual representation of the dataset as follows –

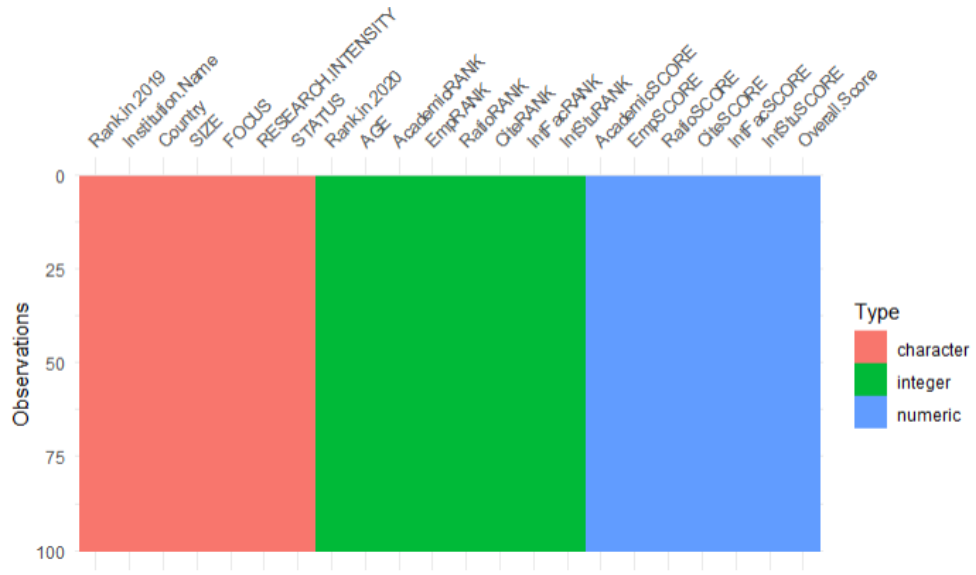


Fig 4: Initial Dataset Visualization with Datatypes

- The above visualization shows column names on x-axis and number of observations on y-axis. The 'Type' legend displays the data type for the observations for corresponding columns.

- We can see that there are only 3 datatypes in the data. They are character, integer, and numeric. Some of the classification factors are converted into factor using as.factor() function in R.

- After checking the data for its types and missing values, we can move on to further data exploration and analysis. The following visualization shows the converted type after transformation in R.

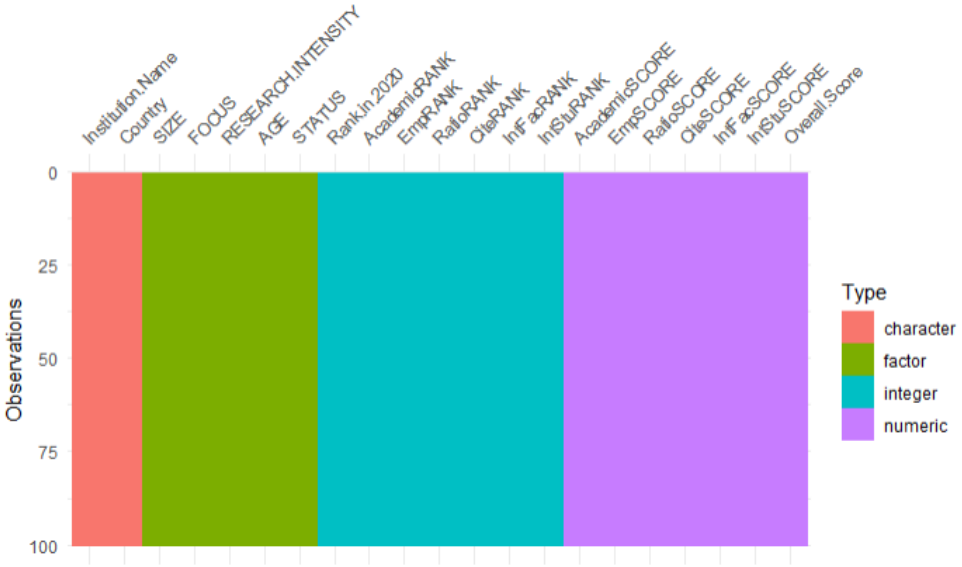


Fig 5: Transformed Data with Factor type

- The above figure shows that factor type is added for classification variables and 2019 rank column is deleted. This data is clean without any errors and is ready for exploration and analysis.

Data Checking – Errors in Data

- Data checking is the process of scrutinizing the data before using it for analysis. Data can be checked both simply in the text form like in MS Excel or using advanced visual methods in R.
- Some packages in R are very useful to check the overall data and its types in each row and column as described in the previous section. Now, let's check the simple text inconsistencies.
- Some of the manual cleaning done on MS Excel before transformations in R is as follows -

A glimpse of the Raw Data is as follows –

Rank in 2020	Rank in 2019	Institution Name	Country	Classification					Academic Reputation		Employer Reputation		Faculty Student		Citations per Faculty		International Faculty		International Students		Overall Score
				SIZE	FOCUS	RESEARCH	AGE	STATUS	SCORE	RANK	SCORE	RANK	SCORE	RANK	SCORE	RANK	SCORE	RANK	SCORE	RANK	
1	1	Massachusetts Ir	United S	M	CO	VH	5	B	100	5	100	4	100	15	99.8	7	100	43	94.1	87	100
2	2	Stanford Universi	United S	L	FC	VH	5	B	100	4	100	5	100	12	98.6	13	99.8	62	67.7	196	98.4
3	3	Harvard Universit	United S	L	FC	VH	5	B	100	1	100	1	98.7	40	99.6	8	86.3	186	62.2	221	97.4
4	5	University of Oxfo	United K	L	FC	VH	5	A	100	3	100	3	100	8	84.7	45	99.7	64	98.5	50	97.2
5	4	California Institu	United S	S	CO	VH	5	B	97.8	23	81.2	74	100	4	100	4	99.4	72	87.3	121	96.9
6	7	ETH Zurich (Swiss	Switzerli	L	CO	VH	5	A	98.4	22	96.7	25	85	113	98.4	15	100	19	98	53	95.9
7	6	University of Cam	United K	L	FC	VH	5	A	100	2	100	2	100	10	74.2	74	100	58	97.6	58	95
8	10	UCL (University Cc	United K	XL	FC	VH	5	A	99.3	14	98.7	18	98.1	43	76.7	68	99.1	79	100	16	94.8
9	8	Imperial College	United K	L	FC	VH	5	A	98.6	21	99.9	9	99.8	30	72.1	85	100	46	100	14	94.1
10	9	University of Chic	United S	L	FC	VH	5	B	99.5	13	93.7	37	96.5	54	78.5	60	70.2	249	81	143	92
11=	12	Nanyang Technol	Singapo	L	FC	VH	3	A	92.1	37	93.8	35	93.9	66	88.8	36	100	21	74.2	169	91.8
11=	11	National Univers	Singapo	XL	FC	VH	5	A	99.8	11	99.2	14	88.3	89	75.7	71	100	23	76.4	158	91.8
13	13	Princeton Univers	United S	M	CO	VH	5	B	99.9	10	98.6	19	70.7	192	100	3	64.9	272	67.6	197	90.9
14	14	Cornell University	United S	L	FC	VH	5	B	98.7	19	90.4	49	65.7	215	95.1	21	93.8	142	72.3	174	89.3
15	19	University of Penr	United S	L	FC	VH	5	B	95.8	32	93.7	38	100	20	67.6	105	84	199	63.6	216	88.9
16	17	Tsinghua Univers	China	XL	FC	VH	5	A	97.4	26	99.1	15	92.4	72	80.4	53	68	259	30.1	429	88.6
17	15	Yale University	United S	L	FC	VH	5	B	99.9	9	99.9	7	100	5	53.5	190	78.1	224	57.4	249	87.7

Fig 6: Raw Data in Excel

- On the first look, we can observe some main columns and sub columns for the 6 factors. Also, some records have inconsistent values in the "Rank in 2020" column like the 11th and 12th rank in the above example.

- Such small inconsistencies are first manually cleaned on MS Excel for easy processing in RStudio for further Analysis. After making manual changes, the clean data looks something like this –

Rank in 2020	Rank in 2019	Institution Name	Country	SIZE	FOCUS	RESEARCH INTENSITY	AGE	STATUS	AcademicSCORE	AcademicRANK	EmpSCORE	EmpRANK	RatioSCORE	RatioRANK	CiteSCORE	CiteRANK	IntFacSCORE	IntFacRANK	IntStuSCORE	IntStuRANK	Overall Score
1	1	Massachusetts Instit	United States	M	CO	VH	5	B	100	5	100	4	100	15	99.8	7	100	43	94.1	87	100
2	2	Stanford University	United States	L	FC	VH	5	B	100	4	100	5	100	12	98.6	13	99.8	62	67.7	196	98.4
3	3	Harvard University	United States	L	FC	VH	5	B	100	1	100	1	98.7	40	99.6	8	86.3	186	62.2	221	97.4
4	5	University of Oxford	United Kingdom	L	FC	VH	5	A	100	3	100	3	100	8	84.7	45	99.7	64	98.5	50	97.2
5	4	California Institute o	United States	S	CO	VH	5	B	97.8	23	81.2	74	100	4	100	4	99.4	72	87.3	121	96.9
6	7	ETH Zurich (Swiss Fed	Switzerland	L	CO	VH	5	A	98.4	22	96.7	25	85	113	98.4	15	100	19	98	53	95.9
7	6	University of Cambri	United Kingdom	L	FC	VH	5	A	100	2	100	2	100	10	74.2	74	100	58	97.6	58	95
8	10	UCL (University Colle	United Kingdom	XL	FC	VH	5	A	99.3	14	98.7	18	98.1	43	76.7	68	99.1	79	100	16	94.8
9	8	Imperial College Lon	United Kingdom	L	FC	VH	5	A	98.6	21	99.9	9	99.8	30	72.1	85	100	46	100	14	94.1
10	9	University of Chicago	United States	L	FC	VH	5	B	99.5	13	93.7	37	96.5	54	78.5	60	70.2	249	81	143	92
11	12	Nanyang Technologic	Singapore	L	FC	VH	3	A	92.1	37	93.8	35	93.9	66	88.8	36	100	21	74.2	169	91.8
12	11	National University c	Singapore	XL	FC	VH	5	A	99.8	11	99.2	14	88.3	89	75.7	71	100	23	76.4	158	91.8
13	13	Princeton University	United States	M	CO	VH	5	B	99.9	10	98.6	19	70.7	192	100	3	64.9	272	67.6	197	90.9
14	14	Cornell University	United States	L	FC	VH	5	B	98.7	19	90.4	49	65.7	215	95.1	21	93.8	142	72.3	174	89.3
15	19	University of Pennsyl	United States	L	FC	VH	5	B	95.8	32	93.7	38	100	20	67.6	105	84	199	63.6	216	88.9

Fig 7: Cleaned Data in Excel

- We can observe that the sub column structure is made into a single column name structure and the inconsistent "Rank in 2020" column in corrected.

- Some columns related to 6 factor Ranks contained data such as 601+ which cannot be processed for operations in R. We find and replace all instances '601+' with '601'. This data is clean enough to be further processed in RStudio and Tableau for further analysis.

Data Exploration

- Data Exploration is the process of making sense of the data through visualizations. It not only helps us answer the questions defined for the analysis, but also helps us discover new insights we did not intend to find that can be useful.

- It sometimes helps us find insights about new things and formulate new questions that can further enhance the details about the primary answers to the questions defined beforehand.

1. What factors other than rank is more desirable when deciding the quality of a University? In other words, how do Universities compare in terms of the 6 factors in QS?

   - We make use of the geom_smooth() function in R to generate a correlation between the Rank and the 6 factors as shows in the 6 tiles below.

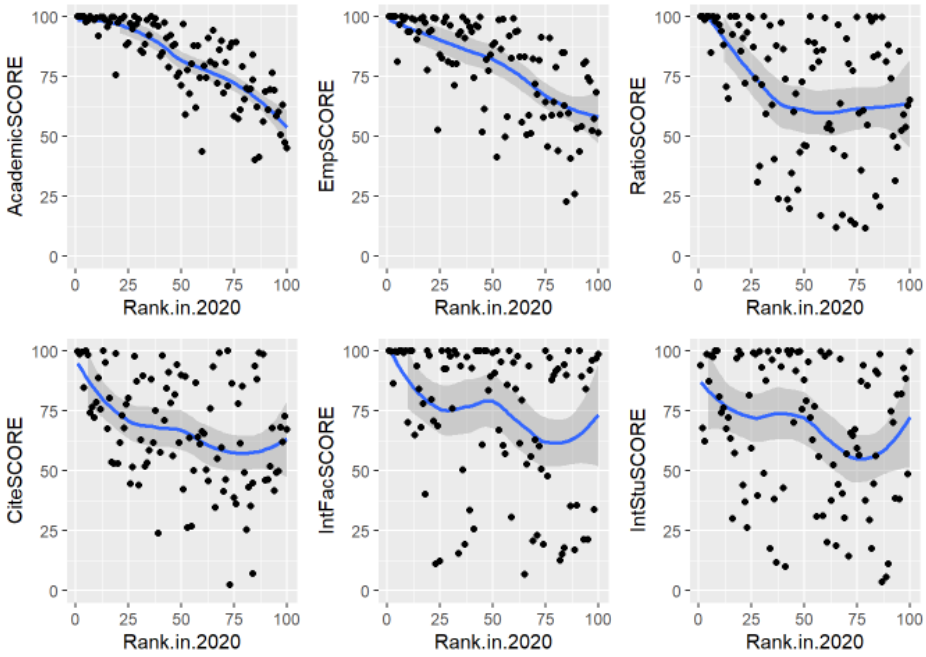


Fig 8: Correlation between Rank of the University and 6 factors

   - From the figure above, Academic Reputation and Employer Reputation Scores have a negative correlation as the Rank value increases from 1-100. This means that better the University rank, better is the Academic Quality and Employers seek graduates from top universities in general.

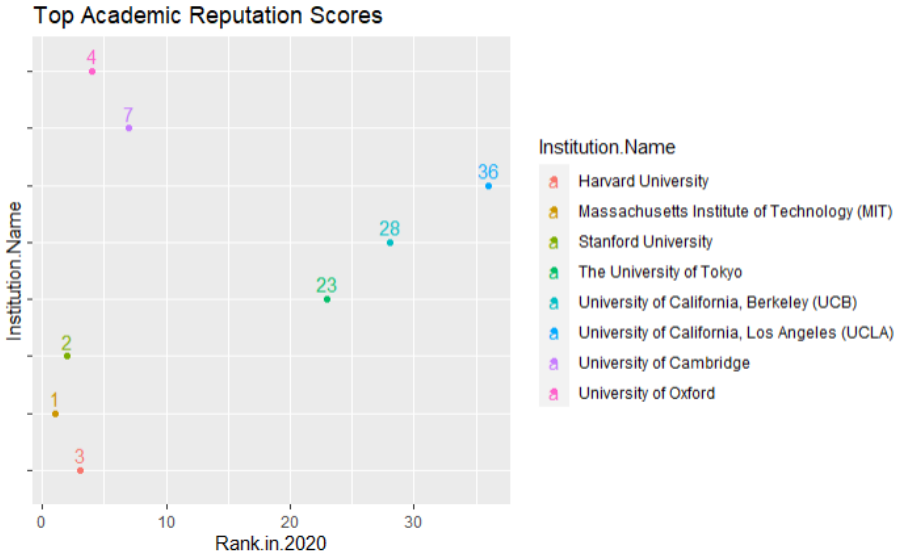
   - Faculty to Student Ratio and Citations per Faculty are somewhat similar after the 35th Rank and shows lesser correlation than the previous two factors.

   - International Faculty and Student Scores are not correlated with rank and show bimodal distribution with the increase in rank value from 1-100.

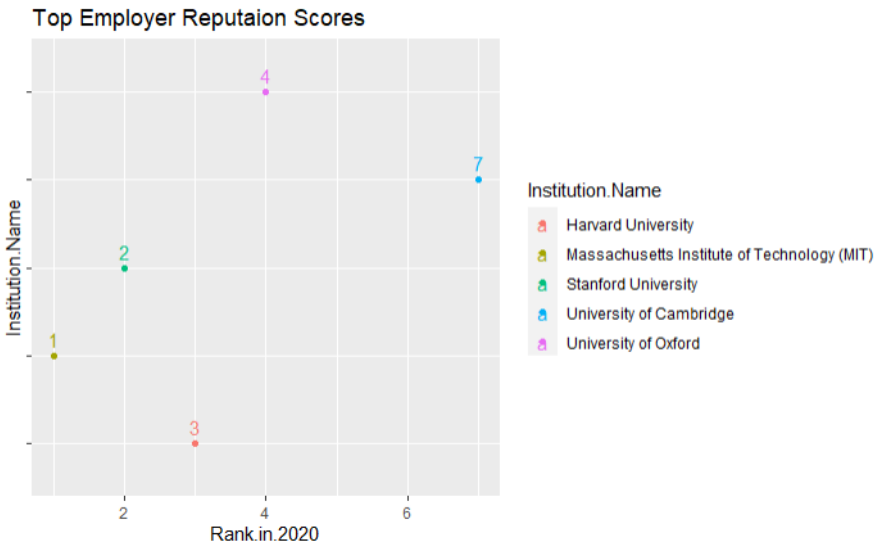
   - Thus, we can say that Academic Reputation and Employer Reputation are the factors other than rank that decides the quality of an institution in general.

2. Which Universities top in each of the specific factors?

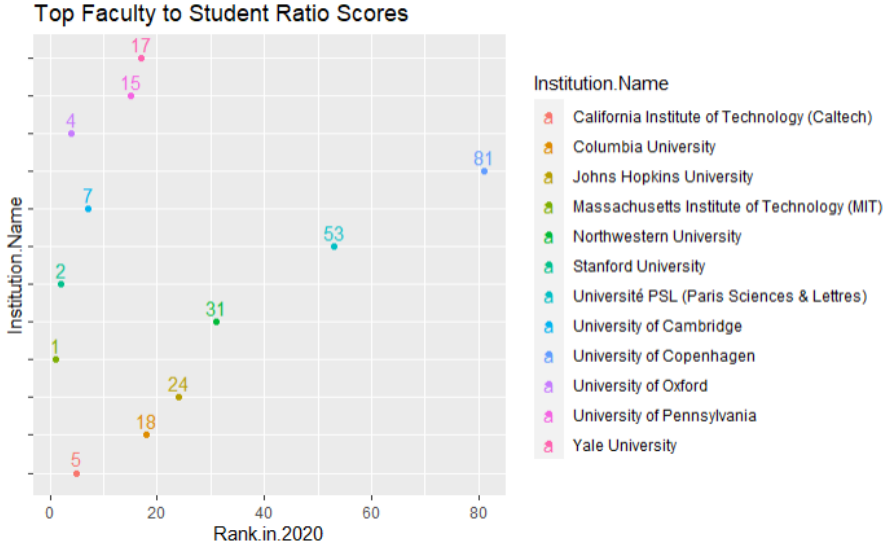
   - We consider the Universities that have a Score of 100/100 in each factor. The numbering in the plot represents the world university ranking in the QS list for that university.



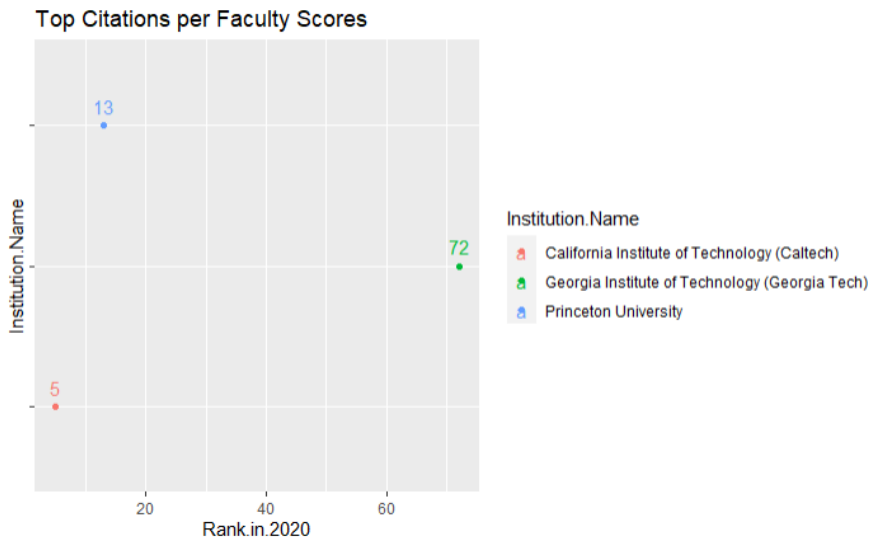
The figure on the left shows the Universities that are best at Academic Reputation in the QS World Rankings. We can observe that University of Tokyo, UCB, and UCLA which are not in the top 10 ranks have the best Academic Reputation. Harvard, MIT, Stanford, Cambridge, and Oxford which are among the top universities make it into this list which was expected.



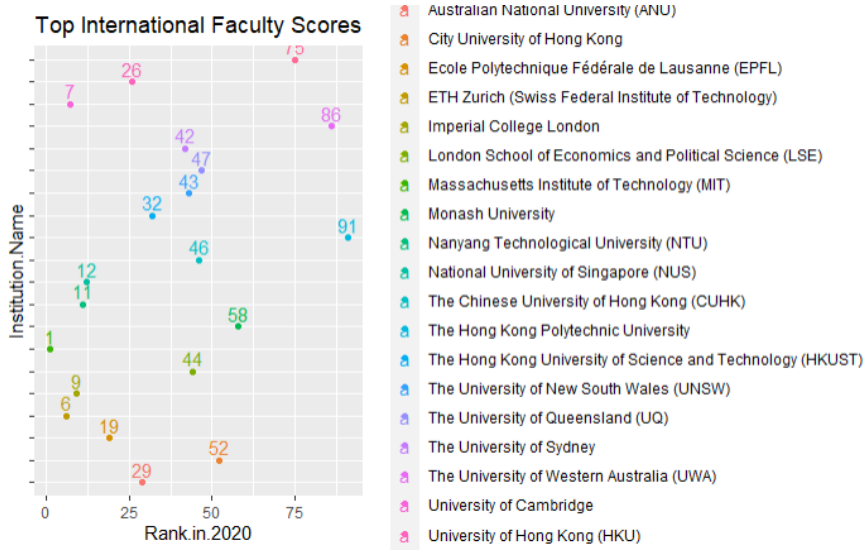
The figure on the left shows the Universities that are best at Employer Reputation in the QS World Rankings. We can observe that only five universities have a score of 100 in this factor. Harvard, MIT, Stanford, Cambridge, and Oxford, all of which are in the top ten have made it into this list. This shows that employers mostly seek students from the topmost universities.



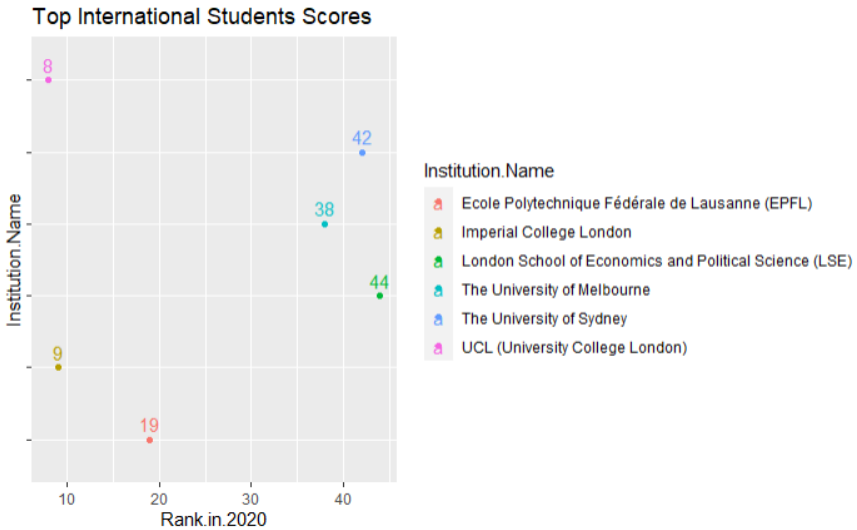
The figure on the left shows the Universities that are best at Faculty to Student Ratio in the QS World Rankings. We can see that Universities ranked at 31st, 53rd, and 81st have made it into this list. This means to say that the universities in this list have a greater number of faculty in comparison to the number of students in that university. We can observe that some of the universities ranked lower also have ample number of faculty per student.



The figure on the left shows the Universities that are best at Citations per Faculty in the QS World Rankings. We can observe that only 3 universities in the QS rankings have a score of 100 in this factor. Caltech, Georgia Tech, and Princeton universities have the most citations per faculty. This means to say that the faculty in these universities have the greatest number of research papers and citable work cited under their names.



The figure on the left shows the Universities that are best in the number of International Faculty in the QS World Rankings. There is a long list of universities that have a score of 100 under this factor. This goes to show that these universities are multicultural and have faculty from different countries from all around the globe.



The figure on the left shows the Universities that are best in the number of International Students in the QS World Rankings. Only 6 universities in the list have made it to this list with a score of 100 in the factor. This goes out to show that these universities are diverse and multicultural in their student population and accept students from a broad array of backgrounds and cultures.

- The top universities in each factor are listed above and this answers our question as to which is best in what factor. We can make a choice based on the kind of environment we're looking for in a university and the factor that matters most to us.

3. Is there a correlation between different classification factors like Country, Age of the University, Reputation of the University, Size and International Student Numbers?

➢ Distribution of Top 100 Universities in each country.

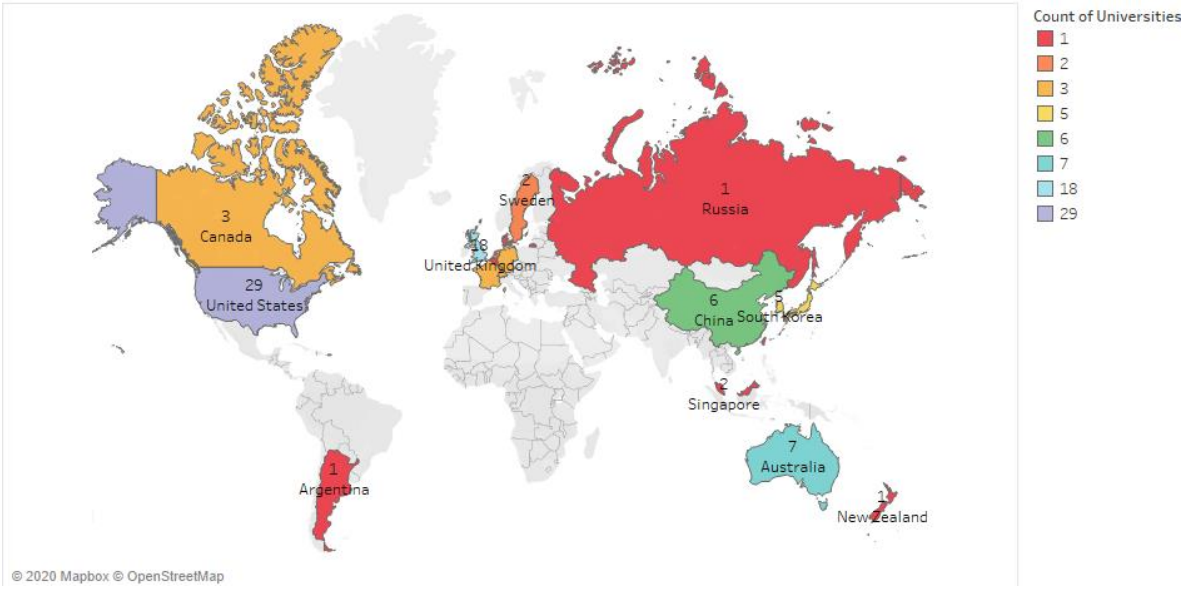
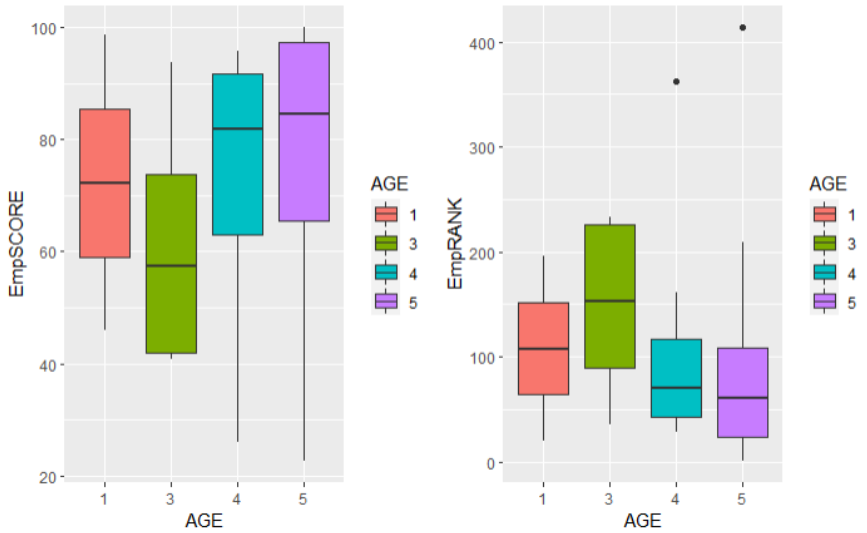


Fig 9: Country-wise Frequency Distribution of the Top 100 Universities in QS World Rankings.

- It is evident from the above figure that US/UK have the greatest number of universities in the Top 100 rankings worldwide with 29 and 18 universities respectively. Australia has the third highest number of Universities in the Top 100 with 7 and the rest of the countries in the list has 6 or lesser universities in the Top 100.

➢ Is there a relationship between age of the university and its employer reputation?



5 = Greater than 100 years old (Historic Universities)

4 = 50-100 years old (Mature Universities)

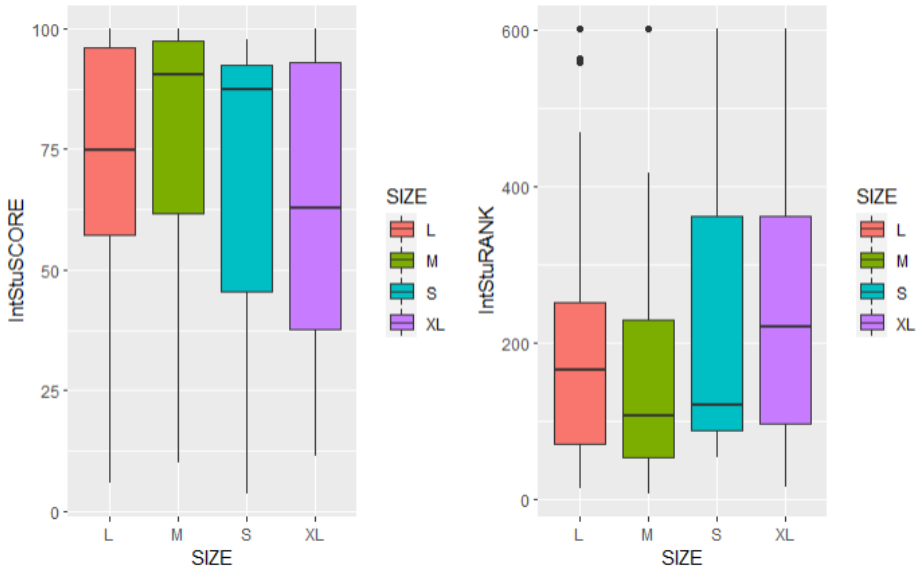
3 = 25-50 years old (Established Universities)

1 and 2 = Less than 25 years old (Young Universities)

Fig 10: Age versus Employer Reputation of Universities

- We can see that generally the historic and mature universities have better employer reputation. A higher SCORE and lower RANK is the desirable reading which is mostly found in the older universities. Thus, employer reputation generally increases with the age of the university.

- Is there a relationship between institution size and number of international students?



XL = Greater than 30,000
L = Greater than 12,000
M = Greater than 5,000
S = Lesser than 5,000

Fig 11: Size versus International Student Scores and Rank

- We can observe that the Small (S) and Medium (M) sized universities have the greatest diversity among students with a greater number of international students compared to Large (L) and Extra Large (XL) sized universities. They have a desirable reading of higher SCORE and lower RANK values. Thus, a higher number of students count does not mean that there will be more diversity in the student population.

Conclusion

From the Data Exploration and Visualization conducted in the previous section, we can conclude that –

- Academic Reputation and Employer Reputation are the factors other than rank that decides the quality of an institution in general. A high ranked University is generally good in Academic and Employer Reputation factors.

- Each University has its own factor of strength. It is not necessary for the Top Universities to have strength in all factors. There is a different list of Top Universities for each category of factors. Individuals need to assess what factors matter most to them and make a choice of institution.

- US, UK, and Australia have the greatest number of Universities in the Top 100 QS list.

- Employer reputation generally increases with the age of the university. Also, some of the young institutions have a better employer reputation than the established ones. This shows the quality of Young institutions is good in general.

- Size of the university does not indicate its diversity. Smaller and Medium sized Universities have more International Students in comparison with Large and Extra-Large Universities.

Reflection

From the above conducted Data Exploration Project, we learnt the importance of Academic Reputation, Employer Reputation, Faculty to Student Ratio, Number of Citations per Faculty, International Faculty, International Students as different factors that drive the quality of an institution and influence World University Rankings.

After this analysis, we were able to understand that face value of rank does not determine the quality of individual factor strengths. We need to assess each category of factors and make a choice based on what factors are most important to us. The choice is based on whether you desire employability, quality of education, diversity, or faculty to student ratio the most.

The topmost universities seem to be clustered in certain developed countries of the world. Usually, the older an educational institution, the better employability it provides. The size of an institution does not decide if it is diverse and multicultural. Universities with a greater number of students might still have lesser diversity among the student population. This analysis can be extended to similar questions to gain insights about all facets of university factors that will help us understand the science behind the university rankings.

Bibliography

Software Used –

- Microsoft Corporation. (2018). *Microsoft Excel*. Retrieved from https://office.microsoft.com/excel
- RStudio Team. (2015). *RStudio: Integrated Development Environment for R*. Boston, MA. Retrieved from http://www.rstudio.com/
- Tableau (Version 2020.2) [Windows]. Chabot, C., Stolte, C., Beers, A., & Hanrahan, P. (2020). Mountain View, California: Salesforce. Retrieved from https://www.tableau.com/

R Packages –

- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
- Kirill Müller (2017). here: A Simpler Way to Find Your Files. R package version 0.1. https://CRAN.R-project.org/package=here
- Tierney N (2017). “visdat: Visualising Whole Data Frames.” _JOSS_, *2*(16), 355. doi: 10.21105/joss.00355 (URL: https://doi.org/10.21105/joss.00355)
- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
- Alboukadel Kassambara (2020). ggpubr: 'ggplot2' Based Publication. Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr