Project Report

1. Team info

- Group no.: 11
- Group members: 6
- Course enrolled: Data Science Specialization
- Location: Pune
- Role of each member in the project: Have done the project individually as I was travelling most of time in and out of Pune.

2. Domain of the project: Education

3. Topic: World University Rankings

4. Introduction

With millions of people opting for higher education every year globally, sometimes it becomes extremely difficult and challenging to choose from all the renowned and prestigious universities spread around the world. That's the reason many prestigious organizations come up with ranking of universities worldwide following different methodologies. Prominent among them are Times Higher Education, Center for World University Rankings(CWUR), SanghaiRanking.

World University Rankings project is about analyzing one such dataset from which the rankings are derived. The objective of this project is to understand the dataset, clean it up, analyze it and work on different findings either through numbers or plot. Some of those are as follows:

- Top 25 universities/ countries by World Rank
- Top 25 countries by International Students
- Top 25 countries by Female students
- Top 10 Countries with Most Universities Ranked each year 2011-2016
- Countries by number of universities
- Correlation between different features

5. Dataset description (describe the data files used in the project)

Among the three datasets available, I decided to go with timesData. The primary reason for that its self-explanatory columns. Along with that there were many challenges to work on this dataset i.e. lots of missing values, special characters in data, data format not in a desired way.
Below are the steps performed to clean up the data:

1. Upload Dataset.
2. Convert factor variables to numeric.
3. Convert female_male_ratio variable to numeric from its present form(33:67:00).
4. Impute Inf with NA
5. Impute NA with mean of the corresponding column.
6. Calculate total score for all the observations and replace it where currently no data available.
7. Outlier detection and removal:
    - There seems to be some outliers in num_students, female_male_ratio and student_staff_ratio.

- Outliers were removed by putting specific filters for each of these features.

6. Business questions identified

Question 1:

1. Top 25 universities/ countries by World Rank
2. Top 25 countries by International Students
3. Top 25 countries by Female students

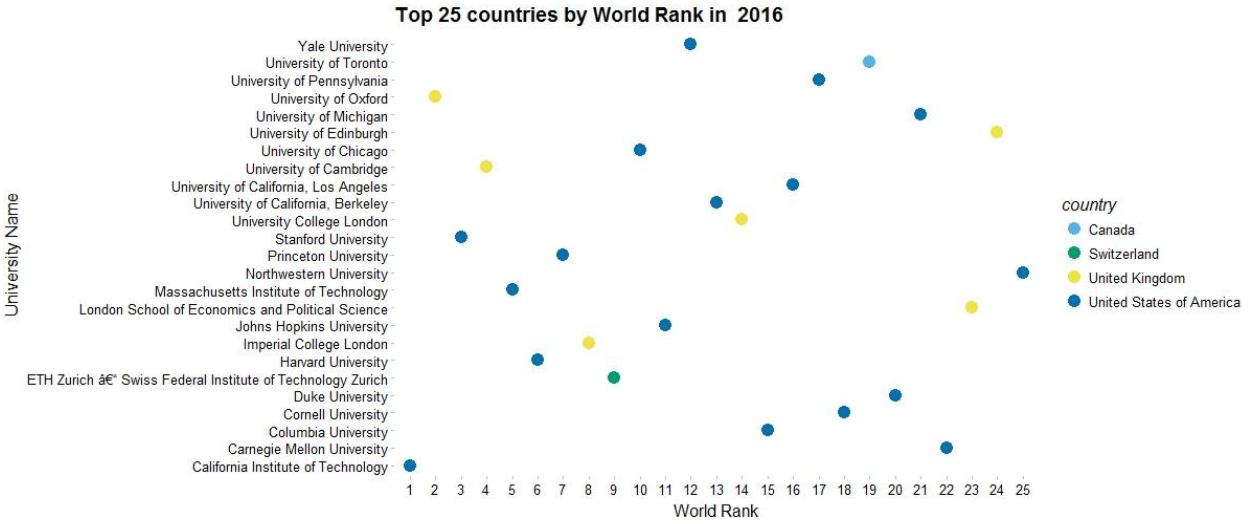
Approach

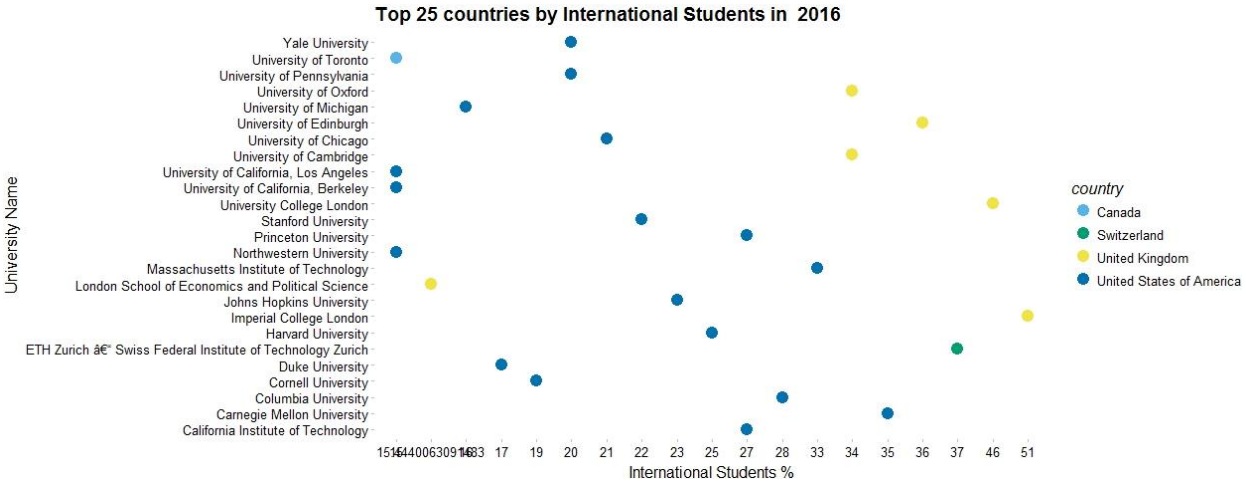
For all the above three questions, the data is already prepared at the end of outlier step. So, in my opinion a visualization should be appropriate to answer them.

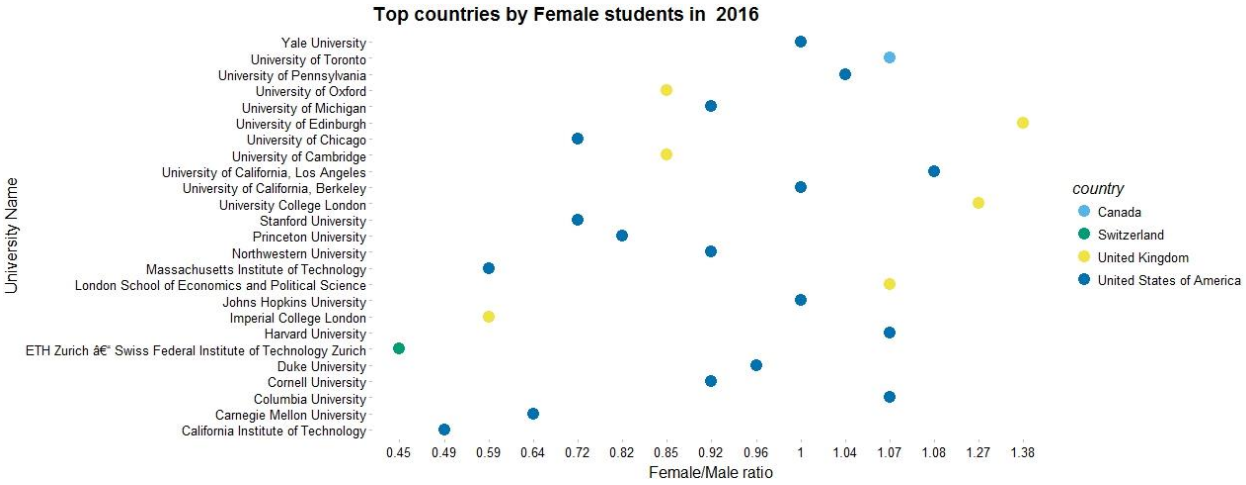
Code Analysis

As there are 6 year's rankings from 2011 till 2016, I have plotted the visualizations in a loop iterating over the years of ranking. So, each question will be answered in 6 visualizations one for each year.

Findings and visualizations







Question 2:

1. Top 10 Countries with Most Universities Ranked each year 2011-2016

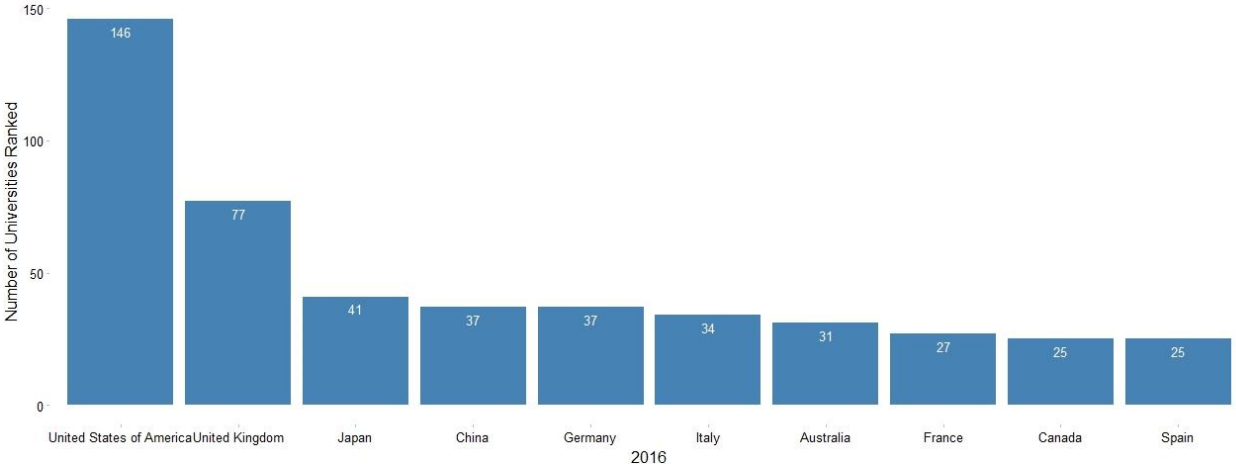
Approach

For this, I have created a bar chart which will visualize the answer to the above question.

Code Analysis

The bar chart is plotted after grouping the data by country and year and then summarizing it by count. There are 6 visualization as usual for each year.

Findings and visualizations



Question 3:

- Countries by number of universities

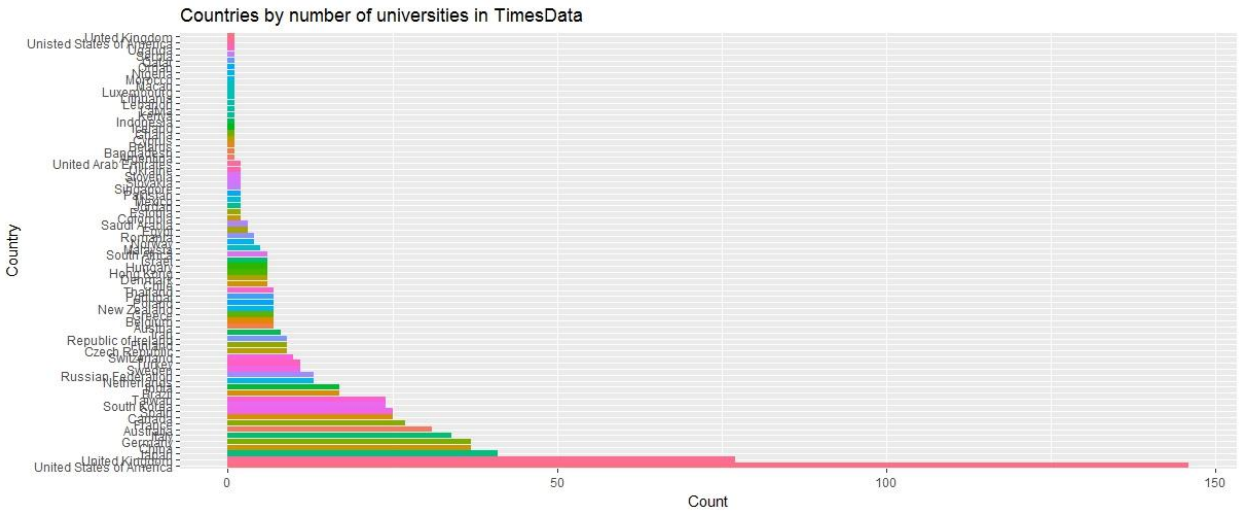
Approach

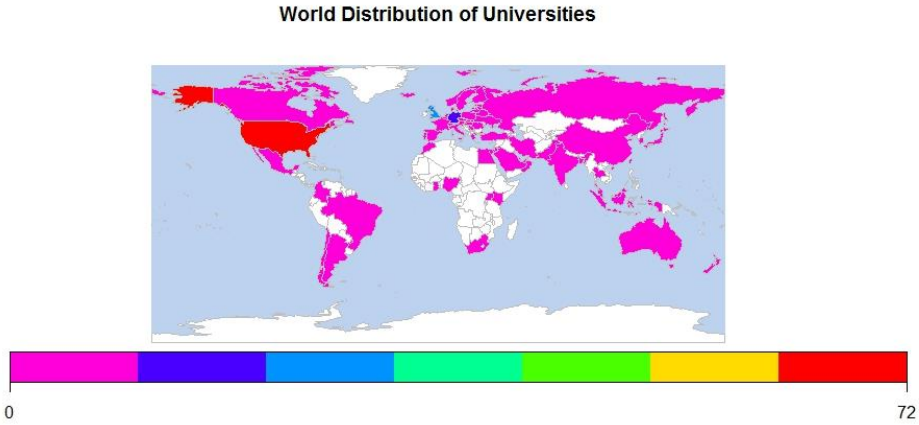
For this, I have created a bar chart which will visualize the answer to the above question.

Code Analysis

The bar chart is plotted after sub setting the data by country and year and then counting by countries. There are 6 visualization as usual for each year.

Findings and visualizations





Question 3:

- Correlation between different features

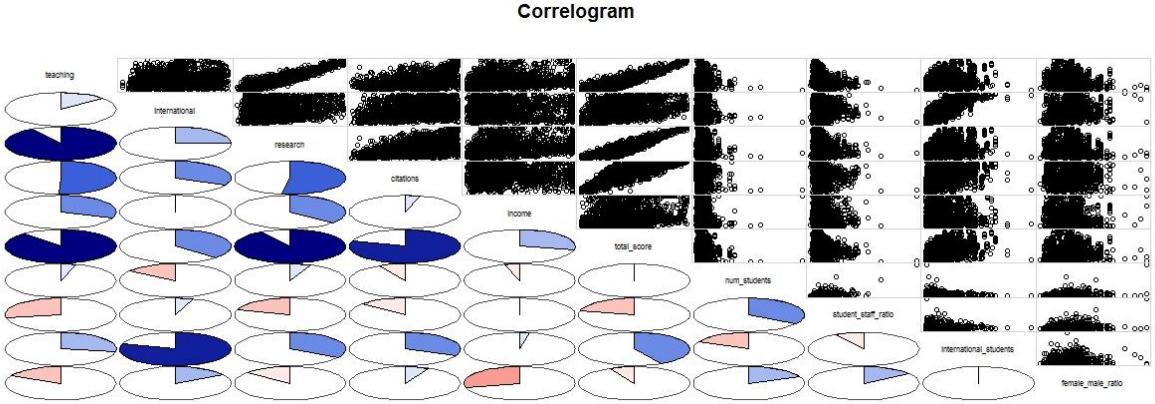
Approach

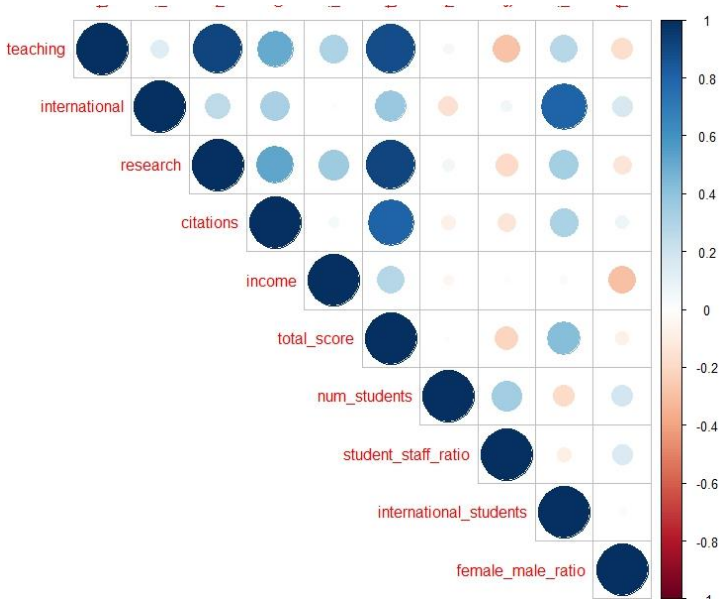
To find out correlation between different features in the dataset, correlation matrix and plotting has been used.

Code Analysis

Corrgram and corrplot packages has been used to create matrix and plot the correlation between different numeric features in the dataset.

Findings and visualizations





1. There seems to be a strong correlation between international_students and teaching (which can increase the level of university).
2. There seems to be a negative correlation between female_male_ratio and income. One can assume that women are given some discounts for their enrollment in some universities.
3. Some negative correlation exists between num_students and international_students. Presence of some universities which specifically attract international students rather than local students might explain this.
4. As expected, there is a positive correlation between num_students and student_staff_ratio.

7. Tools and technologies used

As I was more comfortable in R, I have attempted this project using R:

- R version 3.3.1
- RStudio

8. OS platform: Windows 10 Pro 64-bit

9. Conclusion

Some of the conclusions are:

- The more the number of international students, the higher the level of university. However, that is not enough for a good world rank.
- May be smaller universities admit more percentage of international students.