SI 601: Regional Analysis of Relations between GDP and University Ranking

Introduction

It is universally known that the United States has the best universities in the world. It is to a great extent due to the substantial amount of money invested on the academia by the U.S. government. I am then curious to ask how the economics of a country or region impacts the resources in universities. How and to what degree has it affected the university ranking? Do regions with high GDP overlap with regions with top colleges? For this project, I am planning to use data to exam the relations between GDP of countries or regions and college ranking.

Data Sources

Summary:

There are three datasets I used for this project: the Times Higher Education World University Ranking dataset, the country-continent dataset, and the global GDP dataset.

Details:

I obtained the Times Higher Education World University Ranking dataset from Kaggle. This dataset is in csv file format. It includes university name, ranking, country of university, scores (teaching, international outlook, research, citation, industry income, total), number of students, percentage of international students, student-staff ratio, female-male ratio, and year(2011-2016). I kept columns university name, ranking, scores, number of students, percentage of international students, and records of top 200 universities in year 2016.

Another dataset is global GDP dataset, which I download as excel file from the website www.worldeconomics.com. The excel file has 5 sheets: an overview sheet that defines concepts and links to the other sheets, three sheets that contain GDP data respectively at Purchasing Power Parity, at current prices, and at constant prices, and the last sheet that contains population data. In each of the last four sheets, the columns include country, continent, currency, database year, and years from 1870 to 2015. I used the sheet of GDP-PPP data because the data incorporates the factor of purchasing power and thus is comparable across different countries. I kept the columns country and year 2015, and all records. I chose the year 2015, because I think it could well reflect the metric that has effect on the university ranking in 2016.

The Country-Continent dataset is downloaded from datahub.io as a csv file. The dataset has 195 rows, and contains variables including country and its corresponding continent. The reason why I use the country-continent mapping in this dataset instead of that in the global GDP dataset is because this dataset has more coverage on countries and has more categories in classification of continents.

Data Manipulation

Summary:

My data manipulation procedure mainly consists of processing inconsistent and missing data, converting data type, merging three datasets and aggregating statistics.

Details:

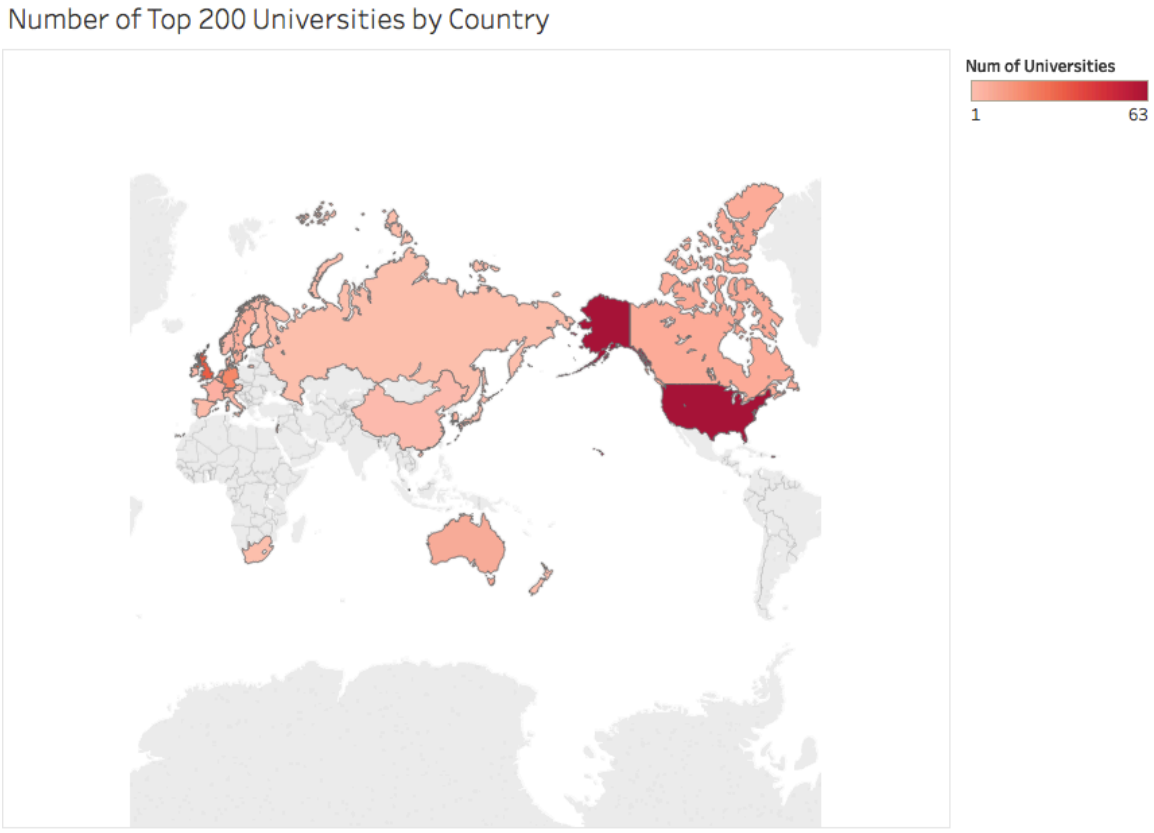
For the Times Higher Education World University Ranking dataset, I changed the country value "Macau" to "China" because Macau is part of China. I deleted rows with country value equal to "Taiwan" because it is a debatable land. I also changed the country value "South Korea" to "Korea, South", and "Republic of Ireland" to "Ireland" in order to prepare for the next step because of different naming in these datasets. In addition, I also converted the column percentage of international students from string to numeric value by stripping off the "%" sign and dividing the number by 100. For the Country-Continent dataset, I changed the column names to lowercase for the convenience of merging dataset. For the global GDP dataset, I changed the country value "Russia" to "Russian Federation" because "Russian Federation" is the one used in the other two dataset.

I first grouped the top 200 universities in 2016 by country, aggregated the values of the columns I picked previously, and then sorted the rows. Besides using the count function on column "total score" to aggregate the number of universities in each country, I also used the mean function on every column to aggregate the average values for each country. I sorted the rows by the descending order of average total score within the descending order of number of universities, so that it can show the overall situation of higher education in different countries. I saved this data frame as python variable "a".

I used the left-join to merge the data frame "a" with Country-Continent dataset, based on the common column "country", and saved the resulting data frame as "dt". For the missing value in continent for Hong Kong, I filled in the blank with "Asia". After that I again used the left-join to merge data frame "dt" with the global GDP dataset, and saved the resulting data frame as "dt2". I changed the multi-index column names into single-index. I grouped the data frame "dt2" by "continent", aggregated the mean values of every existing column, and sorted them by the descending order of number of universities.

During this process, I came across some challenges, as I was not highly well versed with the Python data structures. For example, some columns had a multi-index hierarchical structure, so it caused me some trouble when I merged two data frames based on those columns. I eventually turned them into single-index columns.

Analysis and Visualization

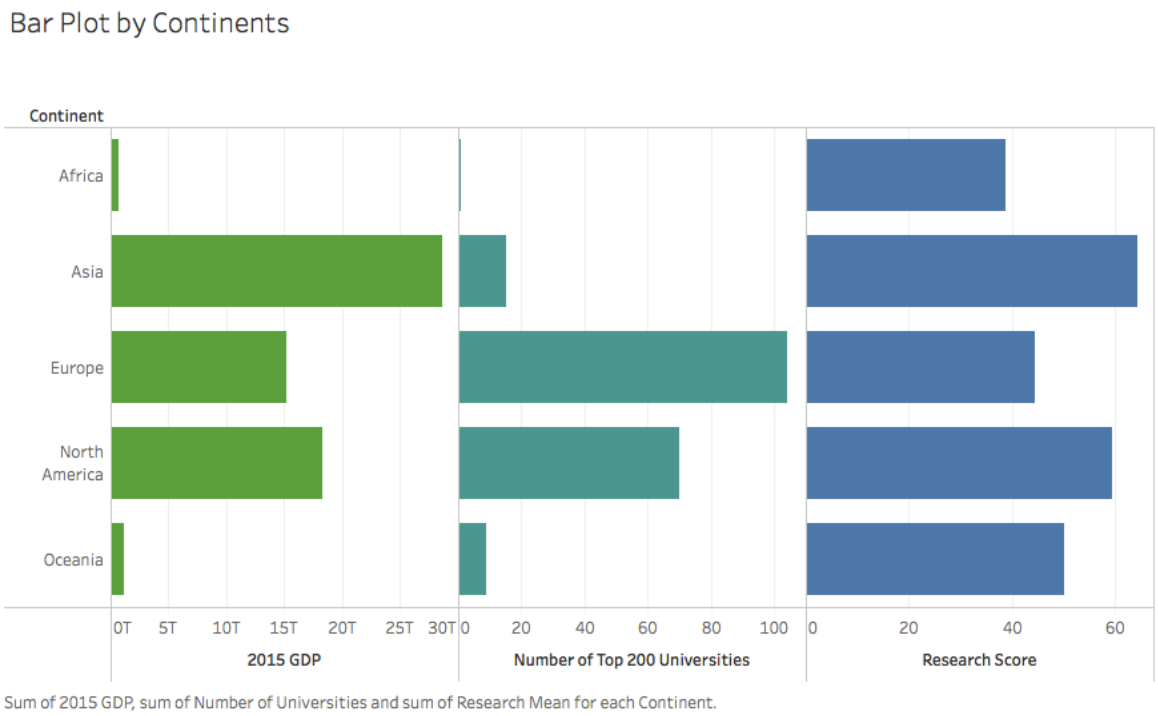


I am interested to know the distribution of top 200 universities in different countries, so I used Tableau to plot the data on the world map. The number of top 200 universities is shown in red, with lighter color representing lower number and deeper color representing higher number. We can see from the graph that United States and a few countries in Western Europe have far more top 200 universities than others.

Correlation	2015 GDP
international_students_mean	-0.455098358
total_score_mean	0.978240869
teaching_mean	0.72940887
total_score_num_universities	0.303873654

citations_mean	0.050304696
2015 GDP	1
research_mean	0.708538571
income_mean	-0.440650391
num_students_mean	0.597479078

I am curious to see a more general picture: how is higher education correlated with the college metrics across different continents. So I drew a table that lists the correlations between 2015 GDP and the 2016 continent-level aggregate metrics of higher education. One important finding is that GDP is extremely highly correlated (0.98) with the mean total score. However, the GDP surprisingly has a low correlation (0.30) with the number of top 200 universities. To find out why there seems to be a discrepancy between these two statistics, I further analyzed by making the following plot.



Among the 5 continents, Asia has the highest volume of GDP yet only the third most number of top 200 universities, whereas Europe has the third highest volume of GDP but the second most number of top 200 universities. This explanation for the question raised in the previous section is: the best resources of higher education concentrate on a small number of universities in Asia while spreading out in a larger number of universities in Europe. Moreover, the right most bar plot shows that Asia tops the list in terms of Research Score, followed in order by North America, Oceania, Europe and Africa. Combining these results, we can conclude that overall the higher GDP a country or region has, the higher colleges quality it has, but GDP is not relevant to the concentration of the top colleges.