27 MAY, 2016

5147 VISULIZATION REPORT
TIMES HIGHER EDUCATION WORLD UNIVERSTIY RANKING ANALYSIS

JIALU LIN
22533206

Table of Contents

Problem Description & Motivation

Times Higher Education (THE) World University Rankings is an annual publication of university rankings by Times Higher Education magazine. It is considered as one of the most widely observed university measures. The reason I chosen THE dataset is due to the following reasons.

1. I want to see the performance of Asian universities against other universities in the world for the period of 2015-2016.
2. I want to see how good is the best university in each country compare to other universities in this country?
3. I want to see the performance of the university from 2011-2016.

I will mainly be looking at the THE's performance indicators, which they used to calculate the ranking scorings to rank the university.

Ranking score formula:

Overall = Teaching (30%) + Research (30%) + Citations (30%) + International (7.5%) + Industry Income (2.5%)

Due to the important underlying data changes in 2015-2016 data, THE strongly against direct comparisons with previous data, so my data analysis will be mainly focus on 2015-2016.

Data Wrangling Process

1. Wrong country name in the THE dataset e.g. "Unted Kingdom", "Unisted States of America".
   a. Used gsub function to replace the wrong country name with correct country names
2. Missing values in data columns "teaching", "international", "research", "citations", "income"
   a. Calculate the mean value of each column based on country and year, then replacing the missing value with its corresponding mean value.
3. Missing scores in "Overall" column for the university rankings below 200.
   a. Use the ranking score formula to calculate the overall scores.
   b. Use the cor.test to test the correlation between the new scores for the top 200 and the old scores of top 200 universities. The result shows they are highly correlated.
   c. Update the missing scores with new calculated scores.
4. Add new column name "rank_range" to divide top 200 universities into different ranking range for visualization purpose.
5. Use geocode to find the longitude and latitude for each country.

Data Exploration

1. In the world top 200 universities, USA has the most universities on the list. Only few universities from Asian make to the list.

Countries by number of universities in THE
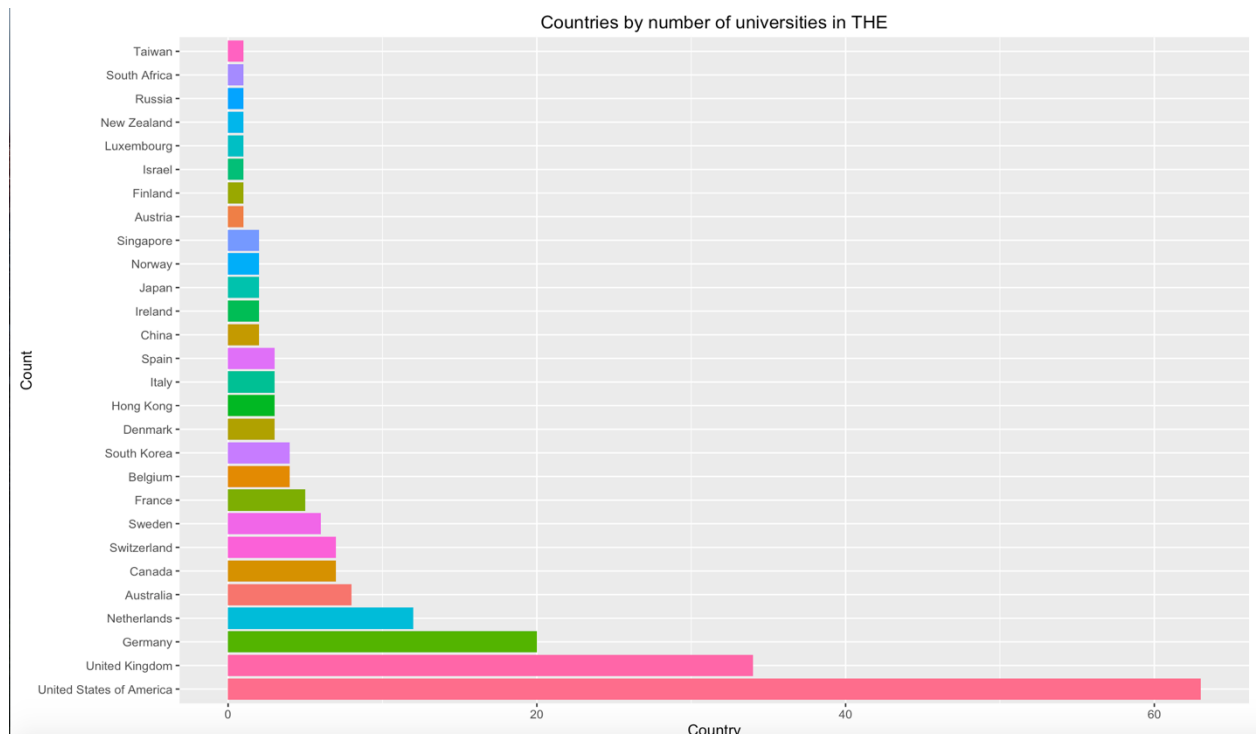
2. Compare three different university ranking system (THE, Shanghai Ranking and CWUR) to find the variance between rankings for the top 100 universities. I combined the universities rankings based on university name, I can only find 34 common universities amongst them.

There are less variances for Yale University, University of Cambridge and Stanford University. Very large variances for Kyoto University, THE ranked this university at 88 for 2015, however CWUR ranked it at 17.

From this result, I can see each ranking system employed very different methodology to rank the university. There is a big discrepancy in university names amongst three ranking systems. Based on this two fact, I think to conduct comparisons amongst the ranking system require a lot text mining to make sure all the university names are consistency across three ranking systems, so my main focus of the data analysis will be on THE ranking.

Variance in University Rankings
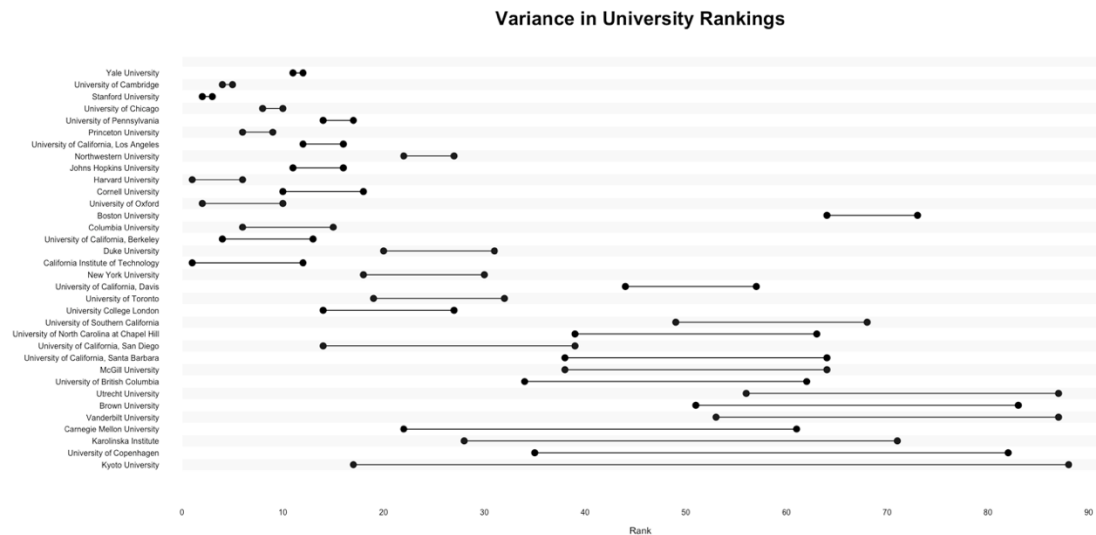
3. An overview of Research, Teaching, Citations, International, Income and Overall score in the world.

   Research (Volume, income and reputation) – United Kingdom and Sweden universities are the best in research.

   USA and Canada Universities are also doing pretty well with their research. Australia universities comes after them in research area. It looks like the rest of the world didn't really put much efforts in researches.

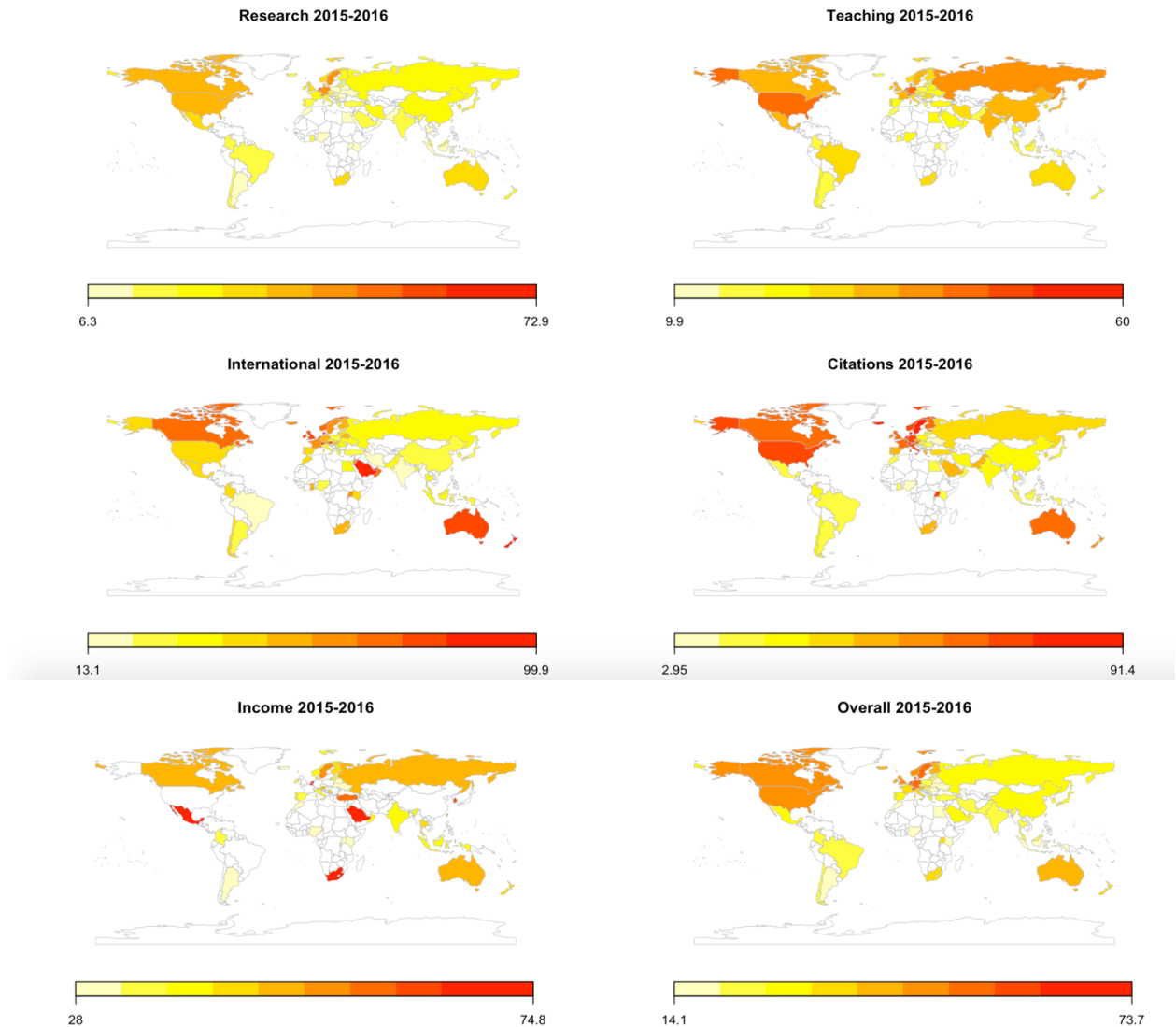   **Teaching** (Learning Environment) – The quality of teaching is very high in USA. Russian and Canada comes next. The interesting finding from this graph is China and Mexico universities have better teaching quality than Australia.

   Citations (research influence – universities' role in spreading new knowledge and ideas)- USA and United Kingdom universities are doing very well in this area. Australia universities also have very high citations.

   International outlook (staff, students, research) - The graph shows that Australia and Canada have high ability to attract undergraduates, postgraduates and faculty from all over the planet.

   Industry Income (Knowledge transfer) – Canada, Russian Federation and Australia all have high ability to help industry with innovations, inventions and consultancy.

   Overall Scores – Universities in United Kingdom and some European countries have the best overall scores. Then comes with universities in USA, Canada and Australia.

5

Research 2015-2016

6.3   72.9

Teaching 2015-2016

9.9   60

International 2015-2016

13.1   99.9

Citations 2015-2016

2.95   91.4

Income 2015-2016

28   74.8

Overall 2015-2016

14.1   73.7

4. Based on the analysis above, it will be interesting to see where are all the students go? We can see most of the students are distributed in Russian Federation, Europe, USA, Canada and Brazil. China also have lots of students, maybe its because it's high population.
I think the result of this graph is having very high similarity to the Teaching graph above. It seems like when students are deciding which university to choose, teaching quality as its number one considerations.

29   791
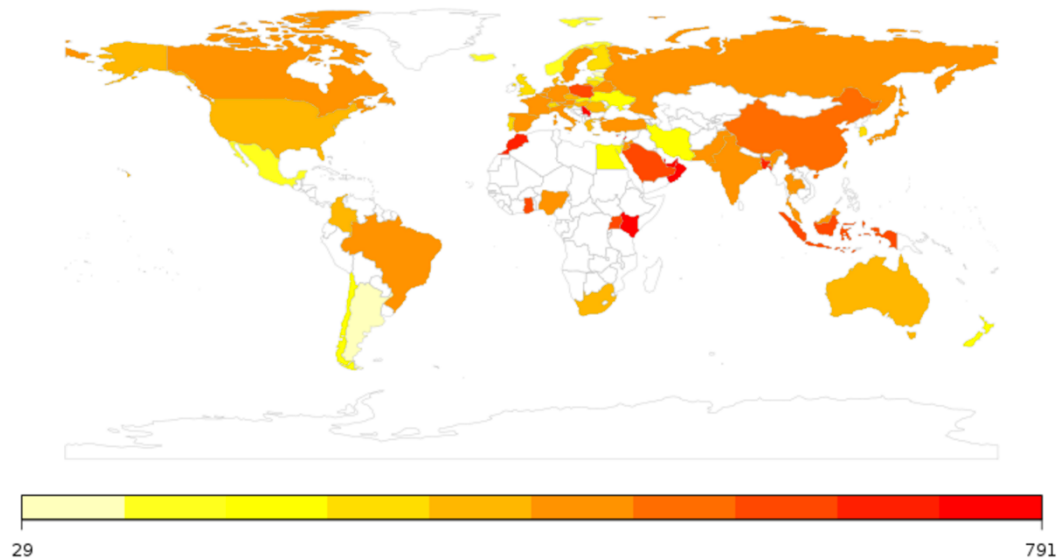
5. Discovering the correlations between attributes.
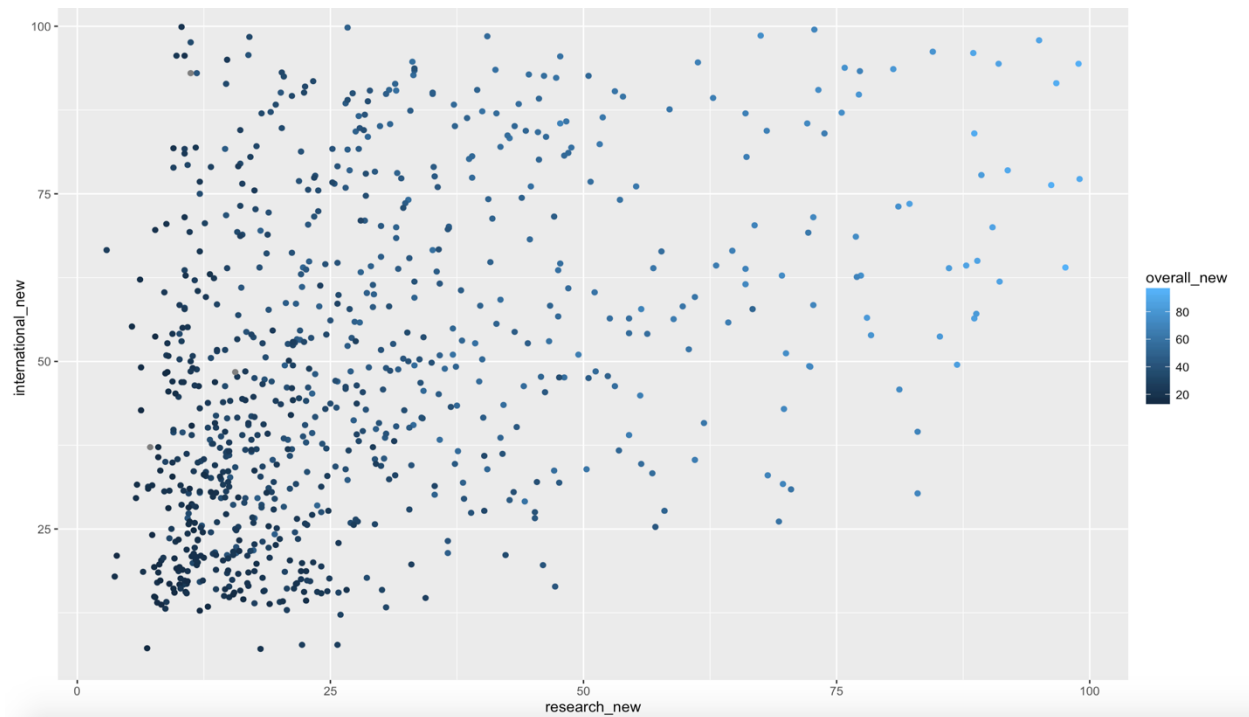   a. A strong positive relationship between Research and Teaching
      This diagram shows the best university in the world are very good with their research and teaching quality.
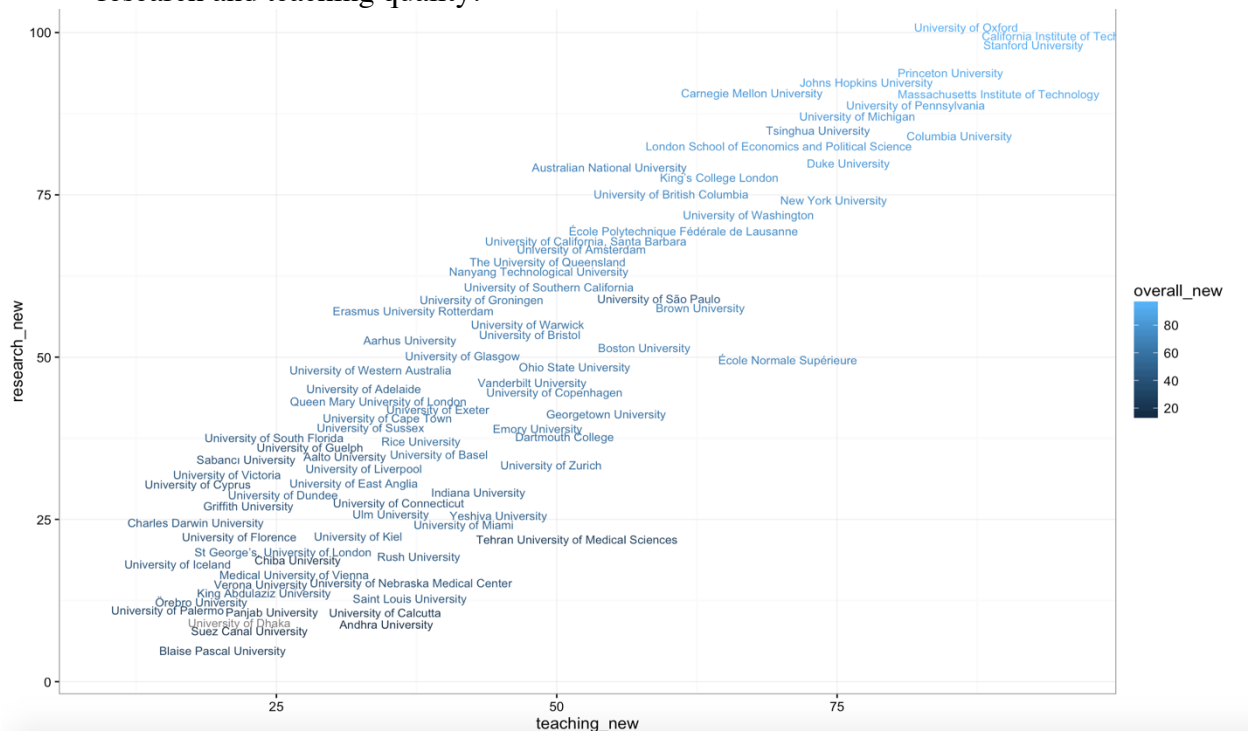

   b. Moderate positive relationship between International outlook and Research
      It shows good universities with high research resources are more attractive to international students/stuff/researchers

c. Strong positive relationship between Teaching and Citations
    It shows good universities with high quality of teaching environment will have
    more influence on the world.
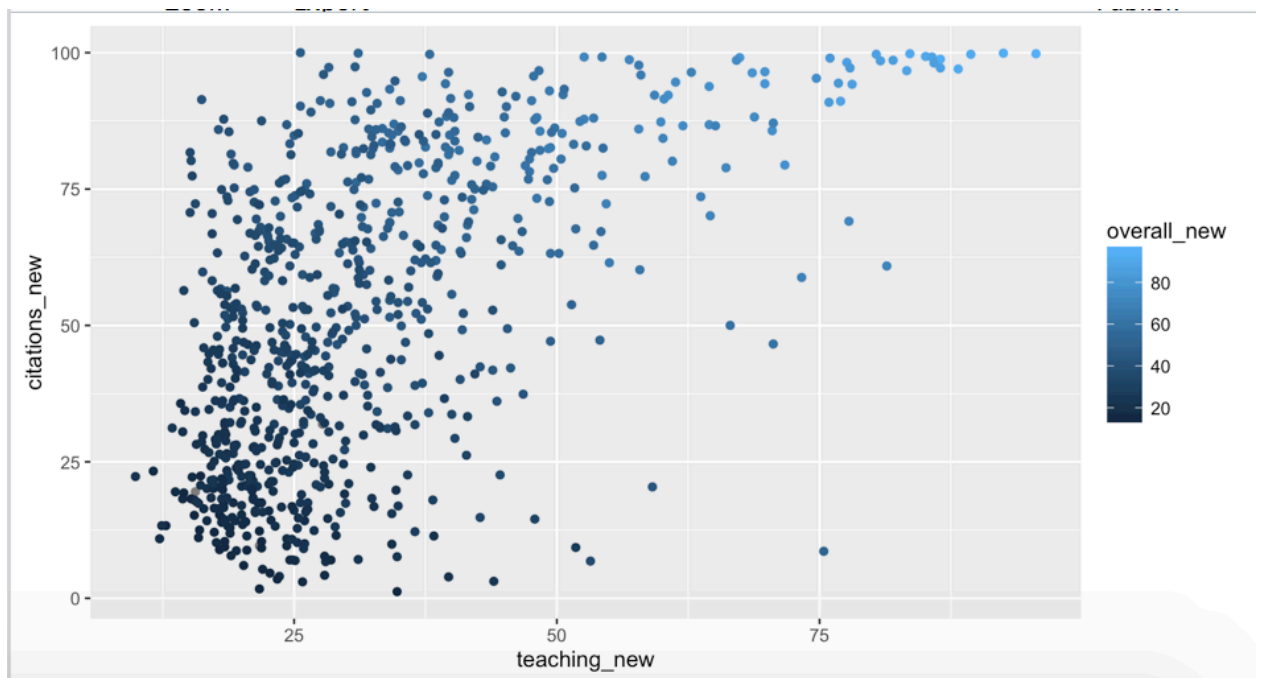
Intended Audience & Message to Delivery

The intended audience is for this visualization project is general public.

My first visualization is a map view of No 1. Universities of each country around the world. A bar chart shows the comparison between the best university and the others universities which makes onto the THE ranking list per country.

The message I want to delivery - how balance is an education system in the country by comparing its best university to its other good universities. Through the analysis of the data, I find in most of the country there is a huge difference between the best university and the other universities. I think one reason is due to lots of universities in USA and European are private universities, so they will get a lot more funding in its researches, which will help them with their ranking.

My second visualization uses scatterplot, parallel co-ordinate plot, bar chart and line chart to find out how universities in Asia performed compare to the rest of world universities performances. The message I want to delivery - the best universities in the country such as Singapore, Hong Kong, Japan and China, they are actually doing pretty well in term of Teaching and Research compare to Australia, but this only happens to one or two of the best universities in the country. The majority of the universities in Asia are fall into the rank between 500-800 on the THE list. The country such as USA, United Kingdom and Australia have a lot more universities on the THE list and lots of the universities from USA and UK are occupied in top 200 rankings.

Visualization Design Process

I used leaflet to build an interactive map for user to click on the map and obtain required information on the map.
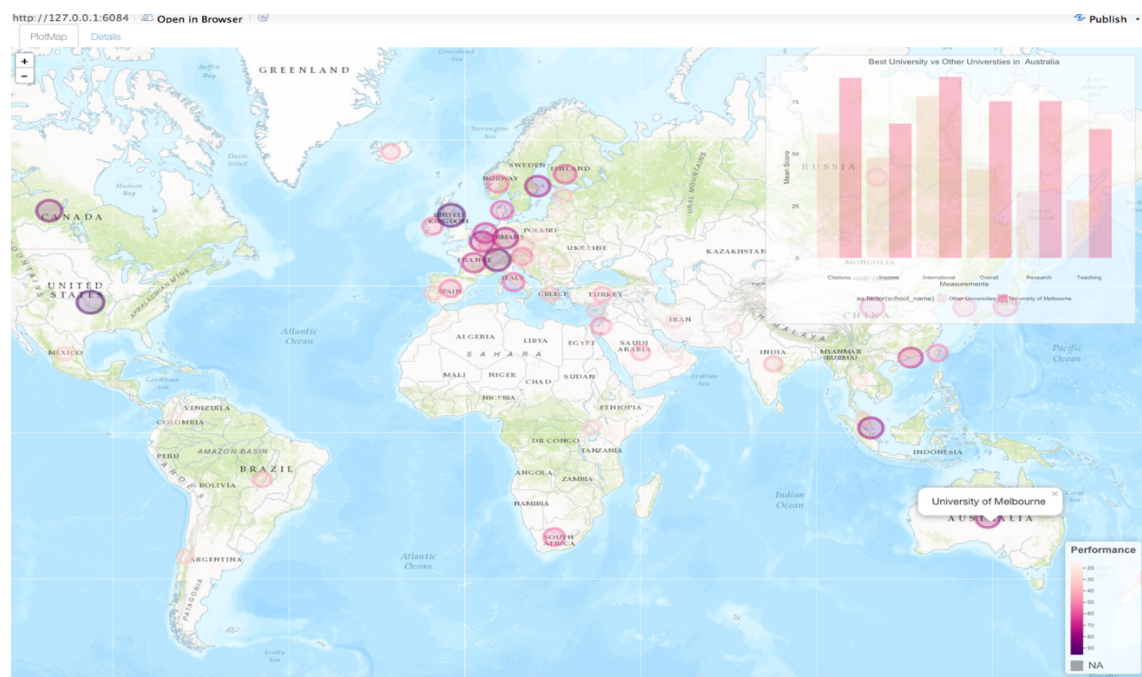
The circles on the map represents the No.1 university for each country (Tooltip shows the name of this university). The color of the circle indicates the overall score of the university, the darker the color is the higher the score. The size of of the circle also calculated by the overall score, a higher score will have the bigger cycle. The score is used to rank the university. The higher the score is the higher the rank for this university on THE.

When user click on the circle, a popup will display on the screen and it will show the name of the best university of this country. The bar chart on the side panel will be updated to show the performance of the best university and the other universities of this country when user click on different circle.

The bar chart uses five performance indicators (Teaching, Citations, International Outlook, Research and Industry Income) to compare between the best university in this country to the rest universities which makes onto the THE rank. The performance scores are calculated by finding the mean scores amongst the universities in this country. For example, in Australian, The University of Melbourne have the highest ranking on the THE ranking system, so the No.1 university is University of Melbourne. The other universities of Australia such as Australian National University, University of Sydney, Monash University are also on the THE list. I summed up all these universities performance scores and calculated mean score to represent all of them.

The initial design is to use text to describe the statistic values.

The reason I changed from static text representation to use bar chart is because the bar charts is more easy for user to understand the statistics and easier for them to compare the values. I also changed the markers on the map to the circle markers. Its easy for user to compare amongst the best universities to see which university have higher scores. The higher score also means the higher ranking of the university.

My second design acts like a dashboard. It contains 3 user controls. "University in Asia" combo box allows users to choose a university country in Asia which includes China, Japan, Taiwan, Turkey, Israel, Hong Kong, India, Singapore, South Korea, Macau, Thailand, Iran and Saudi Arabia. "Universities in the rest of the world" combo box let user to select the rest country. "Performance Indictor" combo box let user to choose different performance measurements.

| Universities in Asia	Universities in the rest of the World	Performance Indicators
China ▾	Australia ▾	Teaching ▾

Scatter Plot Design
The scatter plots are excellent for showing correlations between two sets of values. By using fixed overall scores on the x-axis and changeable values on y-axis allows user to explore the relationship between overall scores and each performance indicators. E.g. The following diagram reveals that the university have higher scores/ranking will have better teaching quality.
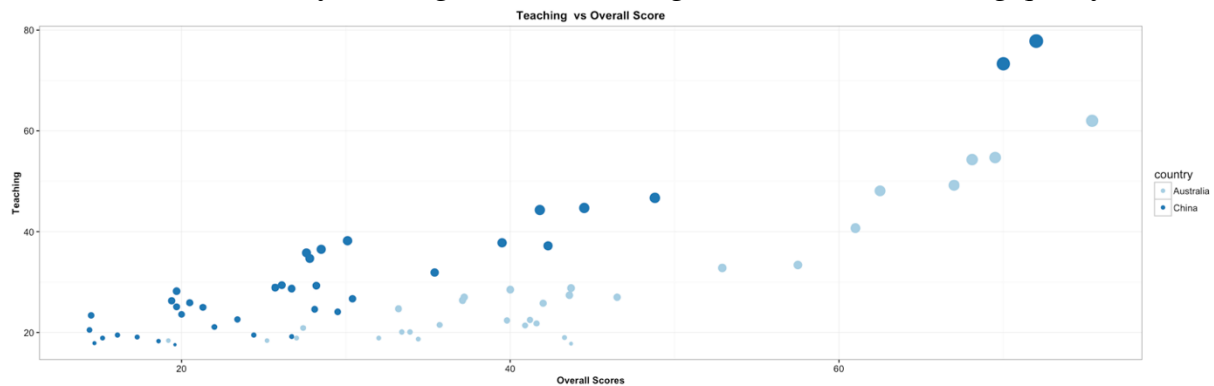
- x-axis represent the overall scores for the universities
- y-axis represents the different performances indicators which user has chosen
- the color is defined by the country.
- A legend on the side of the graph indicates that Australia is light blue and China is dark blue in this case.
- the size represents by the performance score. The better the university perform the bigger the dots on the plot

Based on the graph, it clearly shows that universities in Australia are having better teaching quality compare to China in general, but for one or two of the best universities in China, it seems like that the teaching quality is catching up with the best university in Australia according to THE ranking.
Most of the universities in China are below 50 for its overall scores and teaching quality. However, quite few Australia universities are scored above 60 for both measurements.
The reason I use scatter plot is to allow the user to explore relationships between variables, but also to let users have good ideas about how universities performed in this context.

Bar chart Design
The bar chart shows which ranking range the universities are belonging to based on the countries.

Rankings Comparision
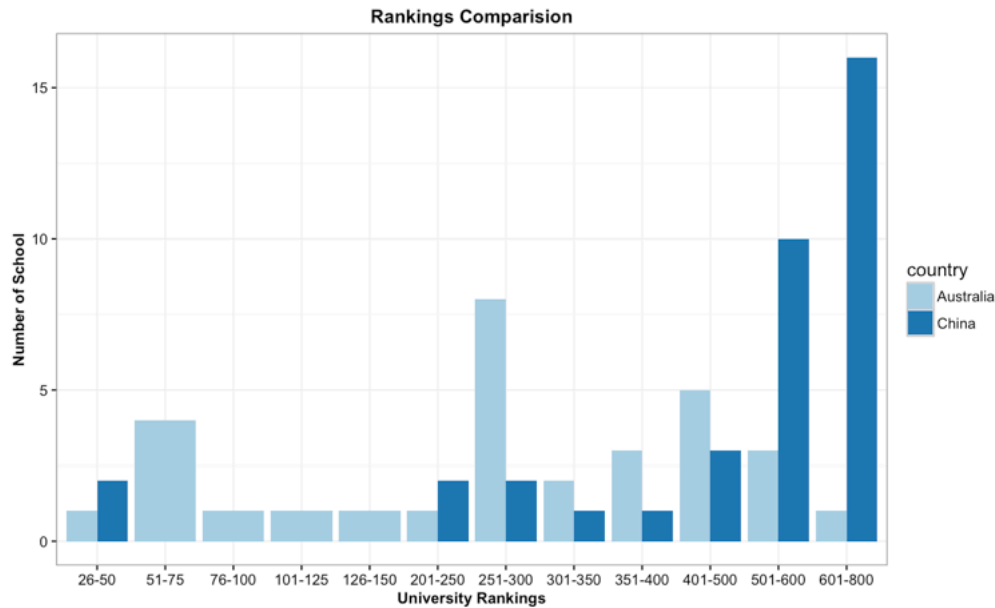
- x-axis represent the university rankings
- y-axis represents the number of universities in each different ranking.
- the color is defined by the country.
- A legend on the side of the graph indicates that Australia is light blue and China is dark blue in this case.

From the graph we can see there are quite a few Australia universities are ranked in top 200. The universities in China are mostly in 500-800.

The reason I choose bar chart is because its easy to compare the numeric values between two categories.

Parallel Co-ordinates Plot Design

I used Parallel co-ordinates plot as my multivariate data design choice. This diagram shows how universities compare to each other based on its Teaching, International, Research, Citations, Income and Overall performances.
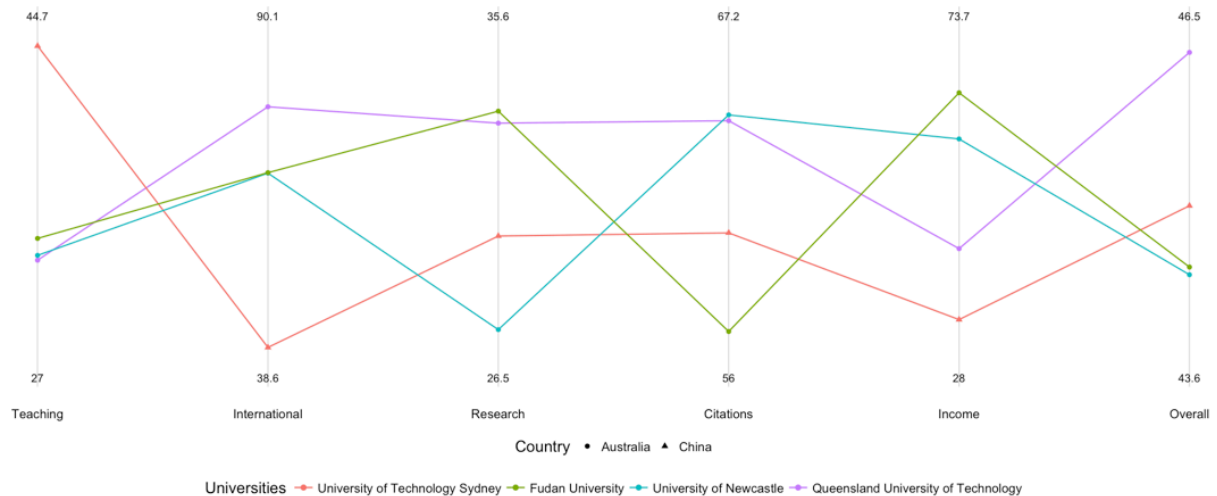
Six vertical lines represents 6 different dimensions. Each line has a name label to show which dimension it represents. It also has a max and min range specified for it based on the values in the dataset. The color of the line represents the school name and the shape is defined by the country.

An alternative representation of multivariate data is to use a radar/spider chart, but I prefer to use Parallel co-ordinates is because:
1. people are inclined to interpret the pentagon's size and compare area.
2. It's hard to compare lengths across the "spokes" because they are on uncommon axes.

Line chart Design
I used line chart to describe the universities performance over the period of time (2011-2016)
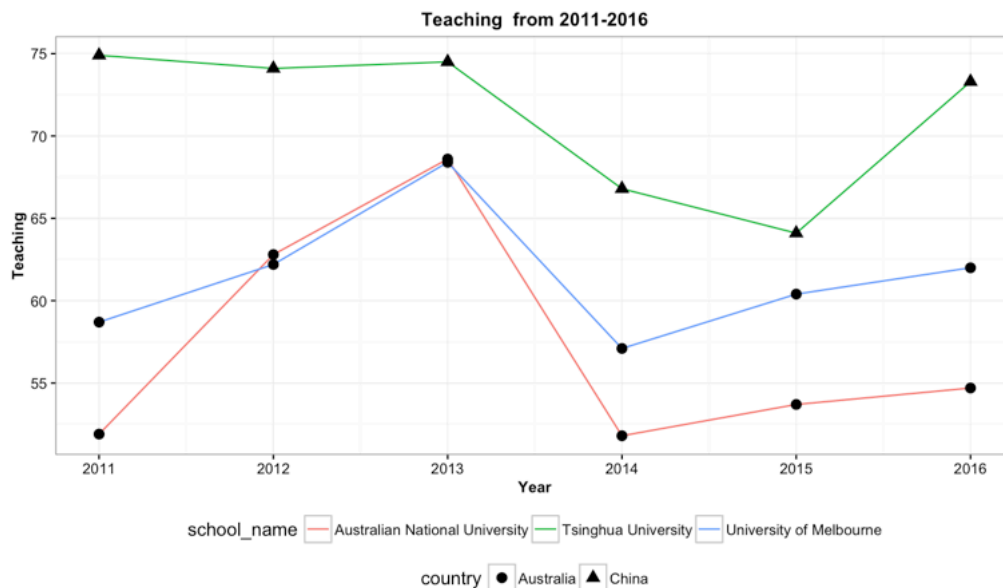
- x-axis shows the year
- y-axis shows the performance indicators
- The color of the line represents different universities

- the shapes on the line represent the country.

The reason I choose line is because its designed to display trends over a period of time. A small multiple design can also be used to show data at different time periods. However, the layout will be at lot harder to managing.

Visualization Implementations & Reasons

I used Shiny and R to implement this visualization project. Using interactive design can improve user experiences by allowing them to filter the input parameters and see the effects on the graph immediately.

The leaflet map allow user to click on the circle markers and see a bar chart to display the difference in university performance. Leaflet map also allow user to zoon in and zoom out the map.

Parallel co-ordinates plot and Line chart are controlled by the Scatter plot. By selecting different point or a range of points on the scatter plot, the parallel co-ordinate plot and line chart will show the performances of these points (universities). This allows users to choose the universities they want to get more information and compare the results.

The three combo boxes will allow user to choose different universities and performance measurements.

The combo box also gives the ability for user to filtering items.

I adapted tab panel for this project, the reason I'm using this design is because the messages I'm delivery are totally different. It makes easier for user to understand the message by separating the view.

I also encoded the longitude and latitude into the original data file, this will decrease the loading time when running the map.

Instructions for Viewing and Exploring

To run the project please make sure you've install the following packages into R environment
install.packages("shiny")
install.packages("MASS")
install.packages("plyr")
install.packages("dplyr")
install.packages("GGally")
install.packages("tidyr")
install.packages("leaflet")
install.packages("ggplot2")
install.packages("ggmap")
install.packages("ggalt")
install.packages("ggthemes")
install.packages("rgdal")
install.packages("gridExtra")
install.packages("stringr")

In "World University Map – (2015-2016)"
Please click on the circles to enquire relevant information.

In "Universities in Asian Country vs. The Rest of World Universities (2015-2016)" tab
To view one university's performance, please double click the points on scatter plot.
To view multiple universities performance, please draw a brush on the scatter plot to select the range of the data.
To changed the country for comparison please use the country selections on the top of the page.
To change the performance measurement please use the "Performance Indicators"

Conclusion

Achievements in this project:
1. Used R and Shiny create the interactive visualization for user to explore and learn the data
2. Used map, scatter plot, bar chart, parallel co-ordinate plot and line chart to effectively present the data to the user
3. Allowed user to to select the data through combo box selections; scatter plot double click and brush event; and map's market click events

Improvements for the project:
1. Use D3.js to enriched user experiences
2. Including the feature to allow users to have the abilities to change the color of the design and allow them to order the data in they way they prefer.
3. Using small multiples showing data at different time periods rather than just using line chart with multiple lines.
4. Engaging the real users in the design process. Having feedbacks from users will enhance the usability of the design, since the final products are used by the real human being.