****

**DATA DOMINATORS; COMPARITIVE STUDY OF TOP GLOBAL UNIVERSITIES IN DATA ANALYSTICS**

**NM2023TMIB06062 - PROJECT**

***Submitted by***

HEMA.V (Reg.No.912620104002)

KALAIVANI .S (Reg.No.912620104004)

NANDHINI.P (Reg.No.912620104008)

SUBASHINI.M (Reg.No.912620104019)

*in partial fulfillment for the award of the degree*

*of*

BACHELOR OF ENGINEERING

***in***

COMPUTER SCIENCE AND ENGINEERING

SRI BHARATHI ENGINEERING COLLEGE FOR WOMEN

KAIKKURICHI, PUDUKKOTTAI-622 303

ANNA UNIVERSITY: CHENNAI 600 025

OCTOBER 2023

# 

# ACKNOWLEDGEMENT

This project is dedicated to Almighty God for blessing us with inspiration parent, teacher and good friends.

First and for most we would like to express our thanks to our Chairman **Mr.G.DHANASEKARAN.,M.Com.,M.A.,M.Phil**., for his cooperation in providing the required facilities and support to complete our mini project.

We extend our heartfelt gratitude and pleasure to our Secretary **Mr.L.DAWOODGANI.,B.Sc.,**andCorrespondent **Mr**.**N.KANAGARAJAN.,B.E.,** and our Institution Administrative Officer **Mr.A.KRISHNAMOORTHY.,B.Sc.,** and all our trust members for their moral support.

We express our gratitude to our Principal **Dr.S.THILAGAVATHI.,M.E.,Ph.D.,** for her cooperation in providing the required facilities and support to complete our mini project.

### We express our heartfelt thanks to Mrs.G.SUGAPRIYA.,M.E.,M.S., HOD of CSE department for her generous support, who has been consistently encouraging us with has valuable suggestion and providing us the facilities to complete our mini project at the stipulate time.

We convey our gratitude to our Project Coordinator **Mrs.S.LAVANYA PRABHA.,M.E.,** and for their excellent guidance and for their prudential, suggestive and eminent help.

We convey our heartly thanks to our teaching staff, non-teaching staff, parents, family members, friends and well-wishers for their valuable suggestion and co- operation for the successful completion of our mini project.

# 

# Problem

John, just completed his schooling and eagerly waiting to pursue higher education.Now, comes the biggest question!!

**Of all the universities in the world, which is the best?**

Ranking universities is a difficult and tedious practice. There are hundreds of different national and international university ranking systems which can be used to compare different universities.

**Dataset chosen**

The Times Higher Education World University Ranking is widely regarded as one of the most influential and widely observed university measures. In this project, we have used the Times Higher Education world university ranking data (times.csv) to develop various visualisations.

The data has several different variables or attributes. Name and country is nominal and then there are several quantitative scores. We have divided the dataset by year just to make the coding easy.Below are the different attributes available for the data and their description -

* world\_rank - world rank for the university.
* university\_name - name of university.
* country - country of each university.
* teaching - university score for teaching (the learning environment).
* international - university score international outlook (staff, students, research).
* research - university score for research (volume, income and reputation).
* citations - university score for citations (research influence).
* income - university score for industry income (knowledge transfer).
* total\_score - total score for university, used to determine rank.
* num\_students - number of students at the university.
* student\_staff\_ratio - Number of students divided by number of staff.
* international\_students - Percentage of students who are international.
* female\_male\_ratio - Female student to Male student ratio.
* year - year of the ranking (2011 to 2016 included).

**Aim**

The aim of this project is to present the Times Higher Education World University Rankings for a creative and interactive way.

**Setup**

Python server and Flask as the web framework.

D3.js to render all the visualizations.

**Different Visualizations**

1. **Bubble Chart**

A bubble chart plots circles with area of circle corresponds to its value.We are planning to plot a bubble chart which forms manybubbles based on the countries. The size of each bubbles corresponds to the number of universities in that country.  On clicking any particular bubble, we would provide many other visualizations to compare universities within that country.

1. **Bar Chart**

We would like to provide the following two types of bar charts -

1. In the first bar chart, frequency is calculated as the total number of students in all the universities in that  country and each bin represents a country. Bar chart will plot this frequency against the corresponding country.
2. In the second bar chart, we will form the bins based on the total number of students. Frequency is the number of universities with their corresponding number of students falling into that particular bin. We are planning to fix the total number of bins as 20 and vary the bin width as (total number of students / total number of bins).

On mouse-over, we would display the value(frequency) of the bar on top of the bar.

**3.Pie Chart**

We would like to provide the following two types of pie charts -

1. The first pie chart has the number of sectors equal to the number of countries in the data and the size of each sector is calculated as the total number of students in all the universities in that country.
2. The second pie chart has the number of sectors equal to the number of bins formed in the corresponding bar chart.  The size of the sector is the number of universities falling into that sector.

 On mouse-over, we would display the value of each sector.

**4.Parallel Coordinates**

Parallel coordinates chart is an excellent visualization for showing many discrete entity values across many different attributes. We would like to provide the following two types of parallel coordinates charts -

1. For the first one, we decided to show the 5 quantitative scores, i.e., teaching, income, citations, research and international and to not show student staff ratio, number of students and sex ratio because the 5 scores seem to be at same conceptual level from a point of view of comparing the universities. Another reason for choosing these were because they had same scale and hence comparable across attributes. Apart from the quantitative scores, we also decided to include name and country, which are nominal variables. Each of the 7 attributes from the data set will be represented in a different axis in the parallel coordinates chart.
2. The second one is country wise. So, this would have only the 6 features i.e., Name of the university, teaching, income, citations, research and international. Each of these features will be represented along different parallel axes.
3. **Radar chart**

A radar chart visualizes multivariate data in a 2D chart of three or more quantitative variables represented on axes. We are planning to provide the following radar charts –

* 1. Each of the 5 quantitative scores, i.e., teaching, income, citations, research and international represent 5 different axes. For each country, the corresponding aggregate value of each feature is plotted and a polygon corresponding to that country is formed.
  2. Within each country, we consider the top 5 universities. For each university, the corresponding value of the feature is plotted and a polygon corresponding to that university is formed.

**Analysis to be done**

**Across-country analysis:**

* Find number of universities by region or country.
* Find the countries with highest and least number of students enrolled. It can visualized using bar-chart or pie-chart.
* Find countries with greater industry income.
* Find countries with maximum International student population.
* Among top Universities, which attribute among teaching, research, income, citations are most related and contributes the most to the University. This can be visualized using the parallel coordinates plot.
* Use radar chart to find commonality and striking outlier attributes for each country.

**Country-wise Analysis:**

* Find the number of Universities within a given range of students. This will help us evaluate student population across Universities.
* Within a country which attribute contributes the most to the University.

**Attribute Analysis:**

* Find the correlation between different attributes/dimensions such as country, teaching, research, income and citations. It will help John to understand which attributes are most related and which attributes contribute the most to the University Ranking.

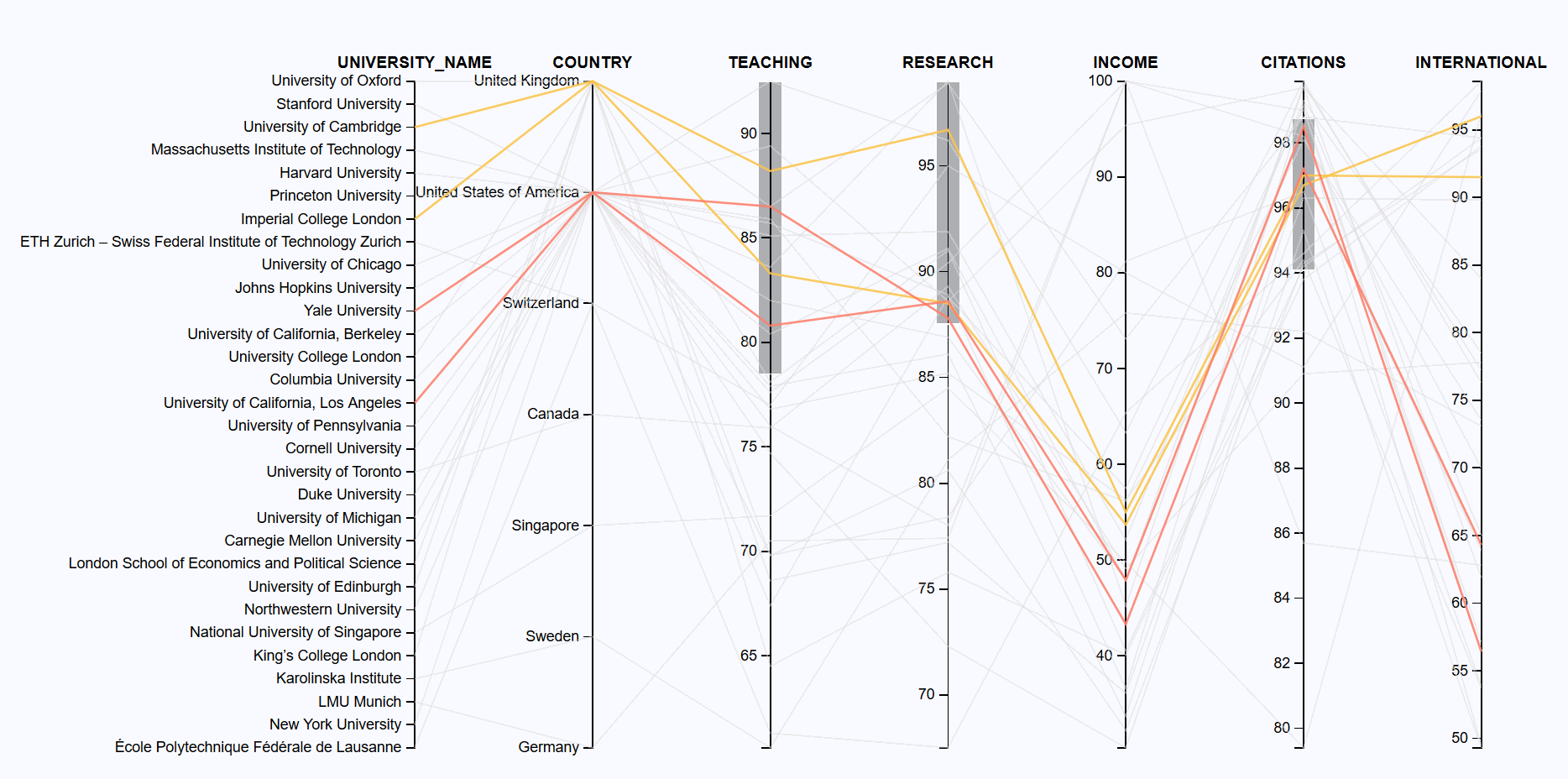
**Work done until Mid-Review**

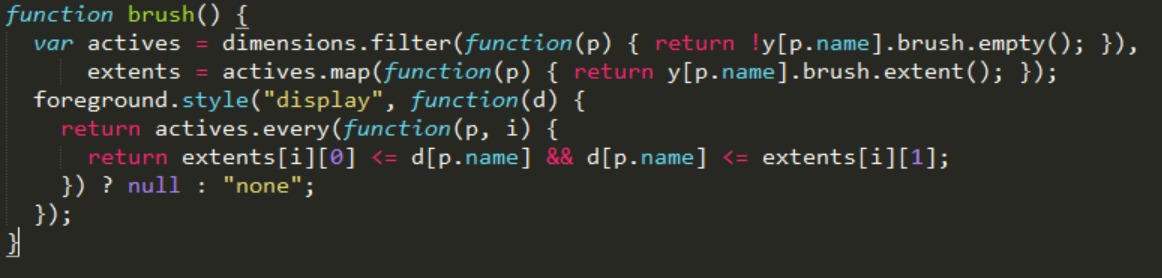
The following are the different visualization we have done over the data and the analysis which we made using the visualizations –

1. **Parallel Coordinates**

We have used color-coding for lines representing universities in different countries. This is useful to compare the general performance of different countries. We have used disparate colors for all the countries so that the difference is marked when we brush for some score in the quantitative scale. Brushing means selecting a subset of the data items with an input device (mouse). This is usually done to highlight this subset if the user wants to focus only on the subset. In our implementation of parallel coordinates, the user can select a subset on each axis so as to see only the universities falling into that subset.

The highlighted attribute filters for teaching, research and citation show the brushing implementation in the parallel coordinates.



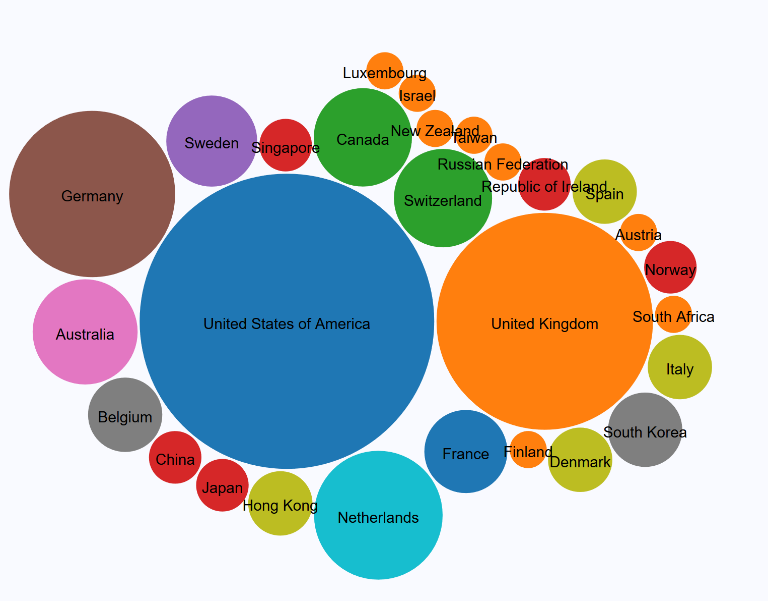
**Code Snippet -**

**Analysis:**

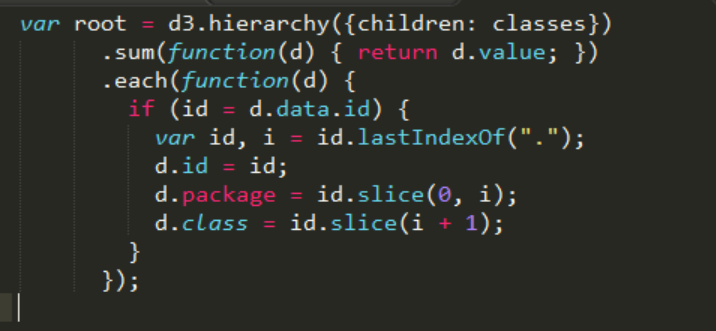
As we implemented brushing on each axis, the user can set a range on any of the attributes and find if there is any university which satisfies his interests. For example, if he wants a university with good research score and teaching score, he can set the corresponding brushes on the axis of parallel-coordinates chart.

1. **Bubble Chart**

As part of data exploration, it helps a user understand the university density in each country.Bubble charts encode data in the area of circles. Larger the area of circle, larger the value represented by the circle. With respect to timesData, bubble chart was plotted such that each bubble represents a country and the size of a bubble represents the number of universities in a country.

****

**Code Snippet –**



**Analysis**

Since we are analyzing only the top 200 universities, the user observes that most of the top universities are in United States of America and United Kingdom. So, he/she has more options to choose from these countries.

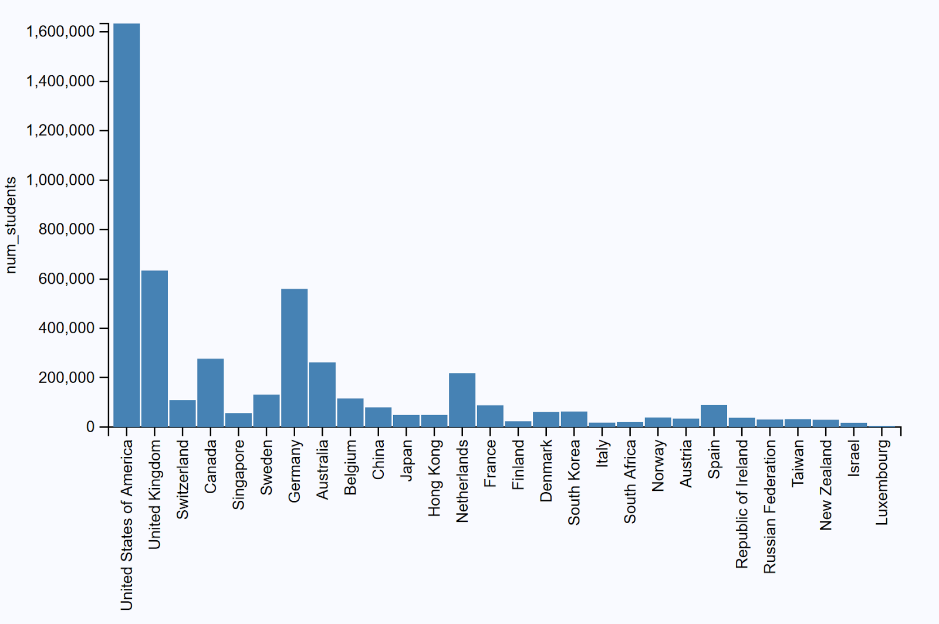
1. **Bar Chart**

We have implemented a bar chart which can be used to see the number of universities that come under a particular bin and we have provided a drop down to select the binning attribute.We also have provided another drop down to select the year to visualize that particular year’s data. On mouse-over a particular bar, it displays the number of universities in that bin.

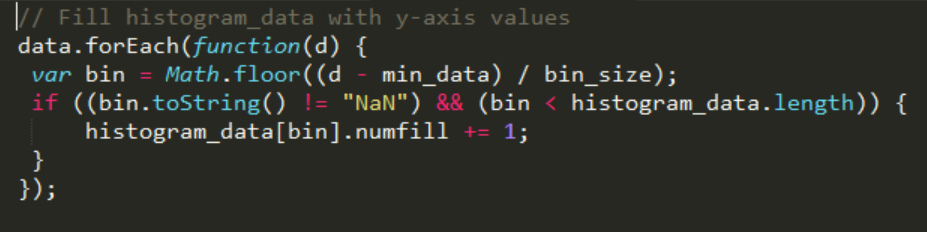
***Analysis***

With these bar charts, the user would get an idea of how many universities he could consider based on the value of each attribute he wishes for. He gets an estimate of the number of universities he could consider as an option.

**Bar chart across country:** Frequency is calculated as the total number of students in all the universities in that country and each bin represents a country.

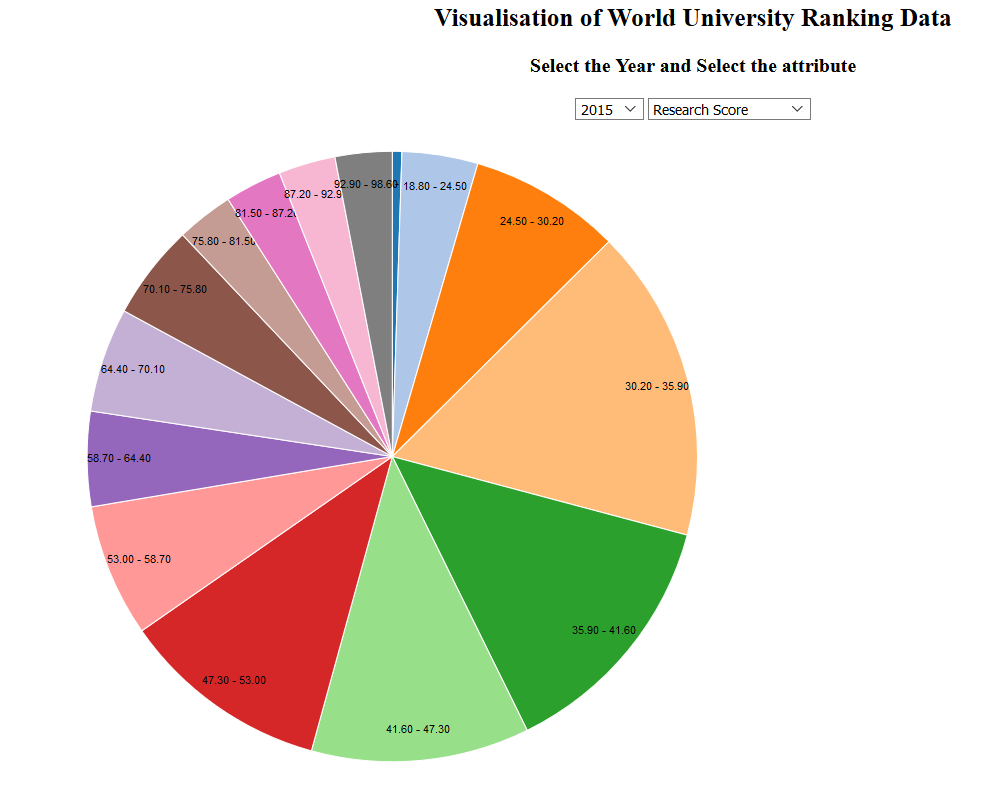
****

**Code Snippet**



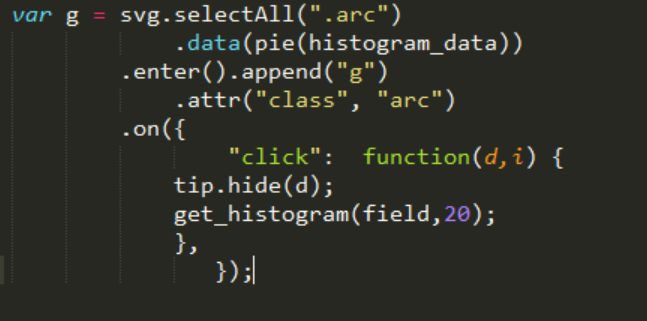
1. **Pie Chart**

On clicking the bar chart, we implemented the visualization in the form of a pie chart. This pie chart would display the exact range of values the sector would hold.



***Analysis***With these pie charts, the user would be able to see the size share of each sector clearly and could see the actual range of values that the sector hold. On mouse-over, we display the number of universities that fall into that sector.

**Code Snippet**



**Proposing additional analysis:**

**1.Evolution of criteria of ranking:**

* We observed that in case of our times-data, the attributes such as teaching, international, research, citations and income are all scores calculated in a scale of 100.
* So, for each year we can find the averaged scores of each attribute to find the contribution of each attribute towards the total score which is used for ranking.
* Here we do it for all the Universities in all countries in each year as we want to understand how each attribute's contribution to University's ranking developed gradually over the years.

**Visualization proposed in order to do the above analysis - Stacked bar chart :**

In a stacked bar chart, it is easier to see the total sizes of all attributes-combined in each year. This visualization will help us understand how each attribute’s contribution to the total ranking evolved over the years.

**2.Evolution of a University ranking across years.**

We want to see how the rank of a university fluctuates over the years. We observed that very few universities tend to maintain its rank over the years. Based on the popular observation quoted above, we can do a hypothesis testing here.

NULL HYPOTHESIS: College rank remains same across the years.

Based on the data let us see if we can disprove the null hypothesis during our final project presentation.

**Visualization proposed in order to do the above analysis – Line Plot:**

* For each University do a line plot with x-axis as year and y-axis as the rank of the University.
* Plot multiple line plots of each University across all the years in single chart.
* Each university will be represented using a unique color. So, the corresponding color will be used for its line plot.

**Data Analysis:**

Times Higher Education Data ranking (timesData) is in very nice format. Teaching, International students, Citations, Research, Income are given a percentage score out of 10. Apart from these, there are other interesting criteria such as number of students, student-staff ratio etc.

In our final review, we have plotted some interesting visualizations to determine how these parameters influence ranking across countries and within a country. We have also analyzed the trend in ranks and criteria over the years. In short, our tool will help any student to shortlist a university based on his preferred criteria.

**PART 1: ACROSS COUNTRY ANALYSIS**

Let us assume you are a student searching for a university which meets your requirement.

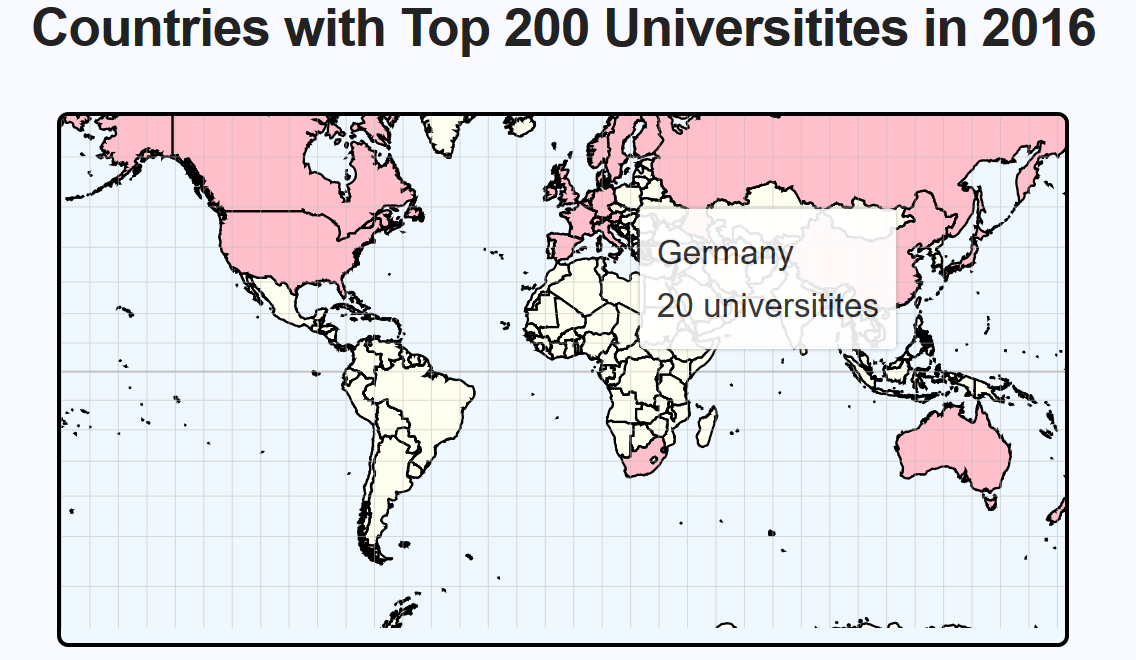
Note: By default, the data is for 2016.

* 1. **MAP**

As a student, you go to the tool to see where are top 200 universities When you hover your mouse on each country, you get the country name and the number of universities in that country.As shown in figure, Germany has 20 universities in top 200 in year 2016. This is the main chart of the tool.

OBSERVATIONS:

* Most of the top 200 universities lie in Europe and United States.
* None of the universities in South America are in Top 200 and only one university from Africa is in top 200.





* 1. **DISTRIBUTION OF UNIVERSITITES (HISTOGRAM, PIE CHARTS)**

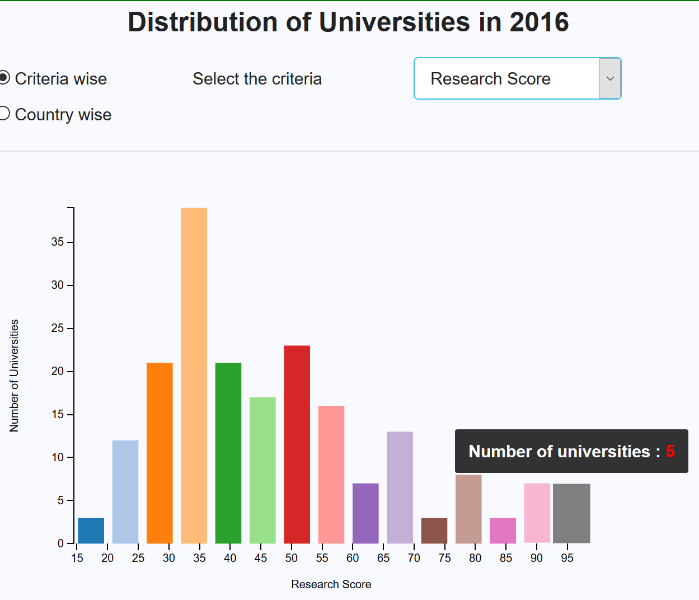
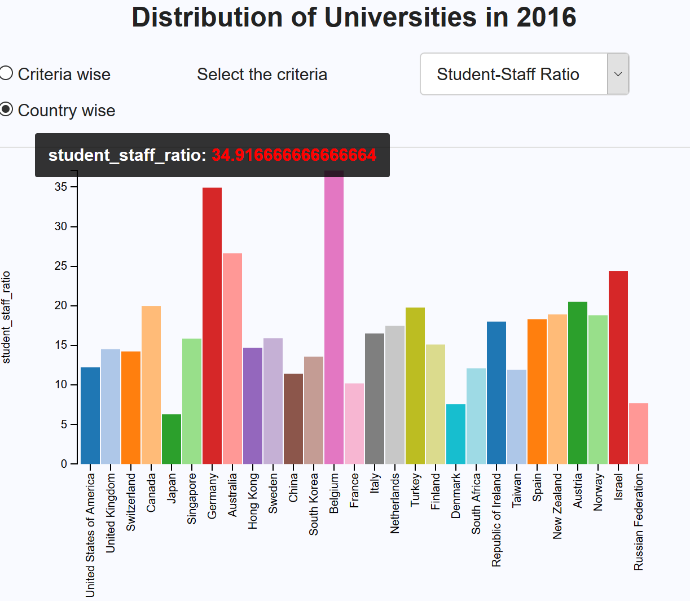
Now, you realize that top 200 universities are unevenly distributed. So, the dashboard provides you with histogram and pie chart visualizations. When you click on histogram, there is also a pie-chart visualization. There are two types of histograms.

First, to get a rough estimate how each criterion varies across country. For example, total number of students studying in Germany are 559304.

But you figure out that, restricting to a single country might limit your options to choose universities, so the second histogram provides you with a university count within a given range of criteria. For example, how many universities have industry income above 95%. Answer is 11 from histogram. Similarly, following are few interesting observations:

OBSERVATIONS:

* Only 5 of the total 200 universities have research score of 95% or above.
* There is only one university with more than 108,000 students. By checking the histogram plot across country, we figure out that it was the only university in South Africa, University of South Africa.After further research, figured out that this is a mega-university specialized on distance education.
* Germany and Belgium have very high student-staff ratio compared to other universities.

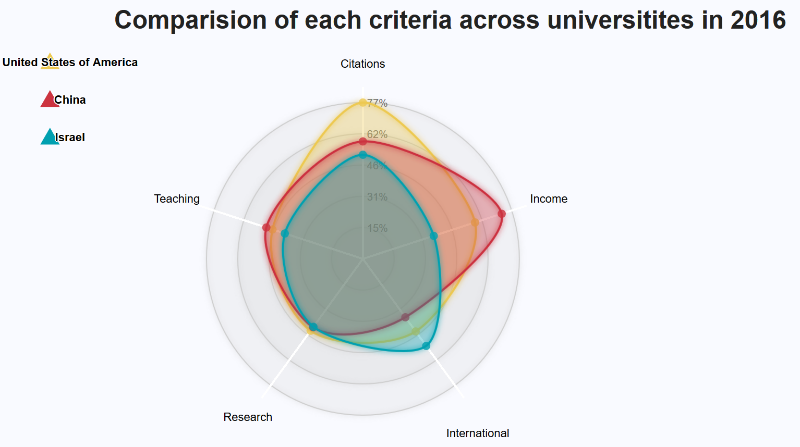


* 1. **COMPARISION OF EACH CRITERIA ACROSS COUNTRIES (RADAR CHART)**

Till now, you have visualized the criteria individually across countries. You have basic idea which countries you are interested in. So, here is a radar chart visualization, where for chosen countries you have holistic visualization how different criteria vary for each country.

OBSERVATION:

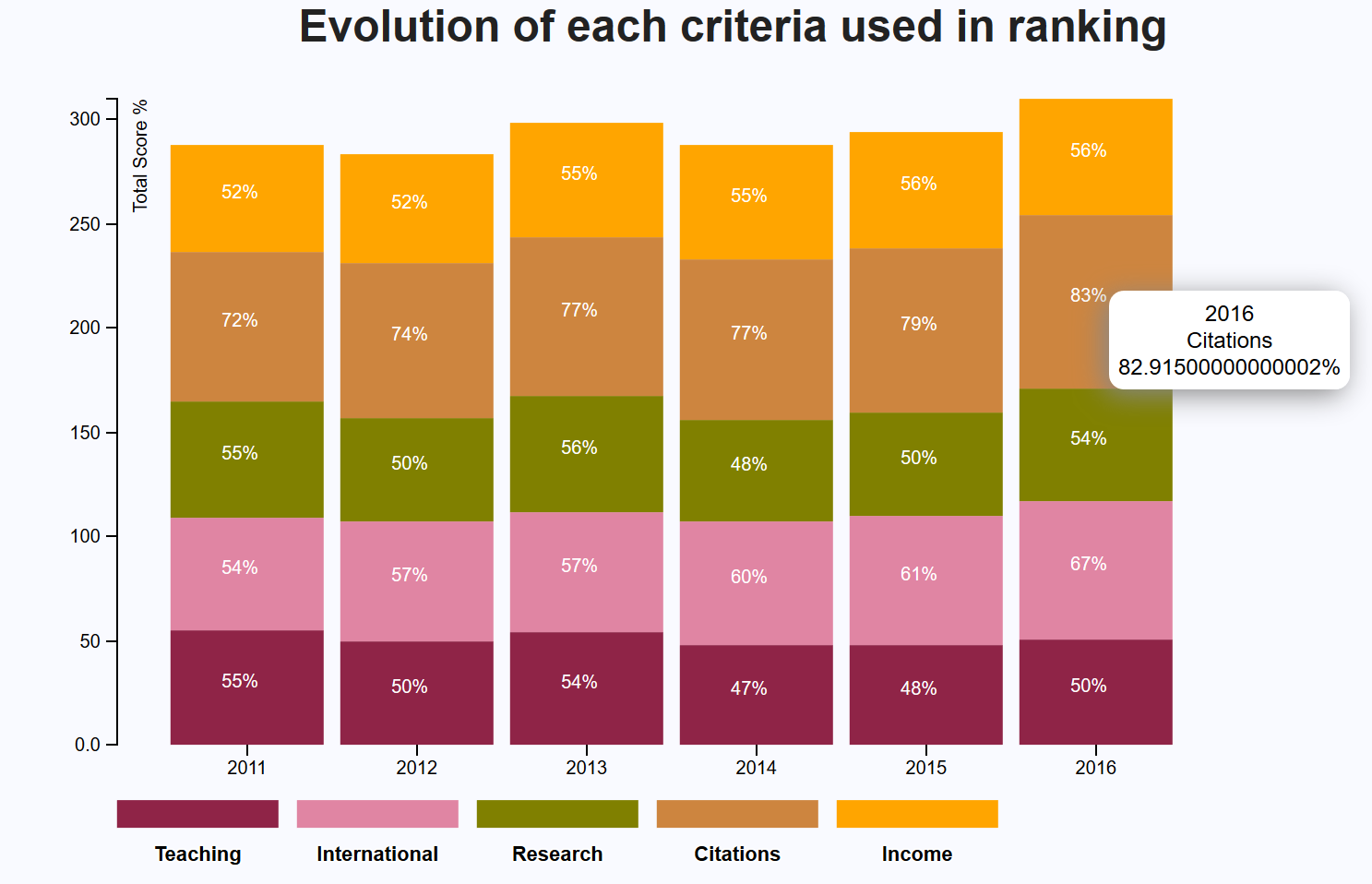
* From the below figure, it can be clearly seen that United States has very high citation score compared to China and Israel.





* 1. **EVOLUTION OF EACH CRITERIA (STACK BAR CHART)**

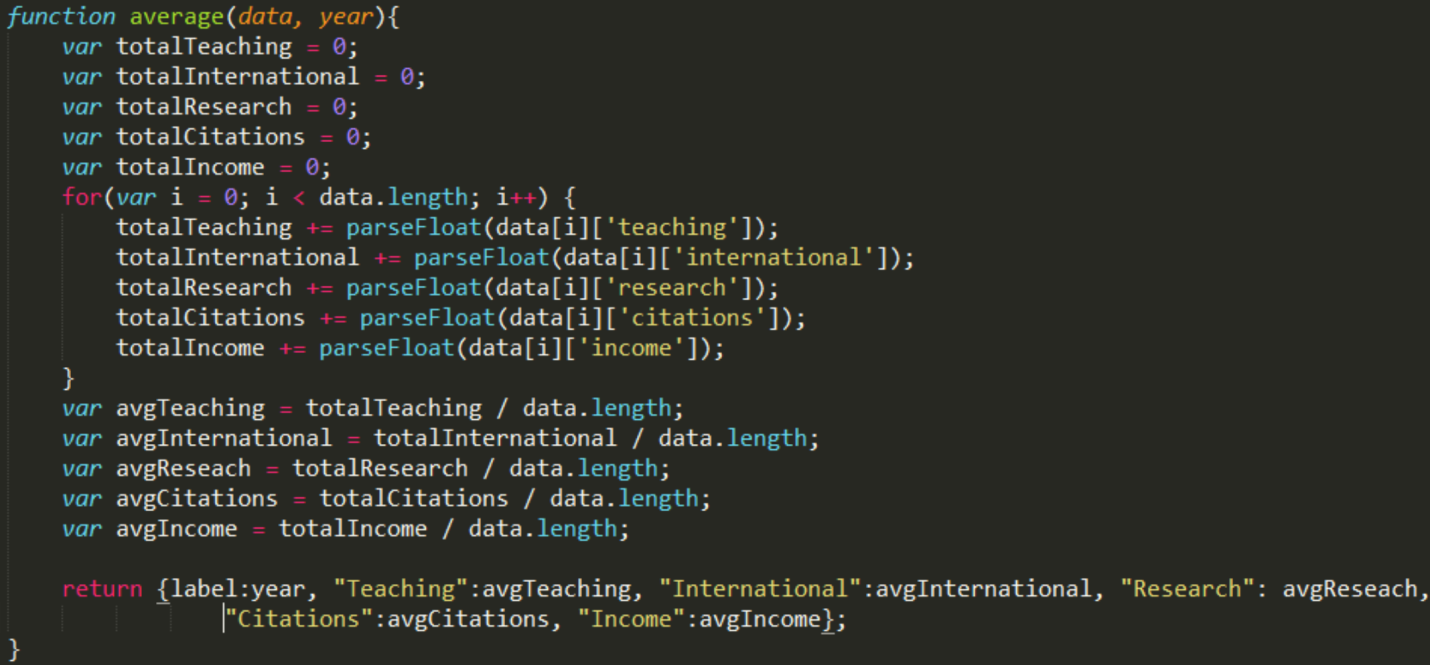
By now, you are almost sure, about the countries where your dream university lies in.Given that you have own criteria, like citations, teaching, research etc. to choose your university, you may want to check how is that criteria evolving over the years.



OBSERVATION:

* Teaching score seems to fluctuate but in general seems to decline over the years.
* International score has steadily increased over the years. This indicates that students are seeking for the best opportunities beyond their home country.
* Research seems to fluctuate the most i.e. Research is independent of time.
* Number of citations are increasing over time. This implies good quality research was performed which helped citations in future.
* Income score is almost constant over the years.

Hence, while choosing a university, there are high chances that its international score, citations might improve while income might remain the same the next year.





Here you realized which criteria fluctuates over the years and which remains the same. Finally let’s go to PART 2:

**PART 2: WITHIN COUNTRY ANALYSIS:**

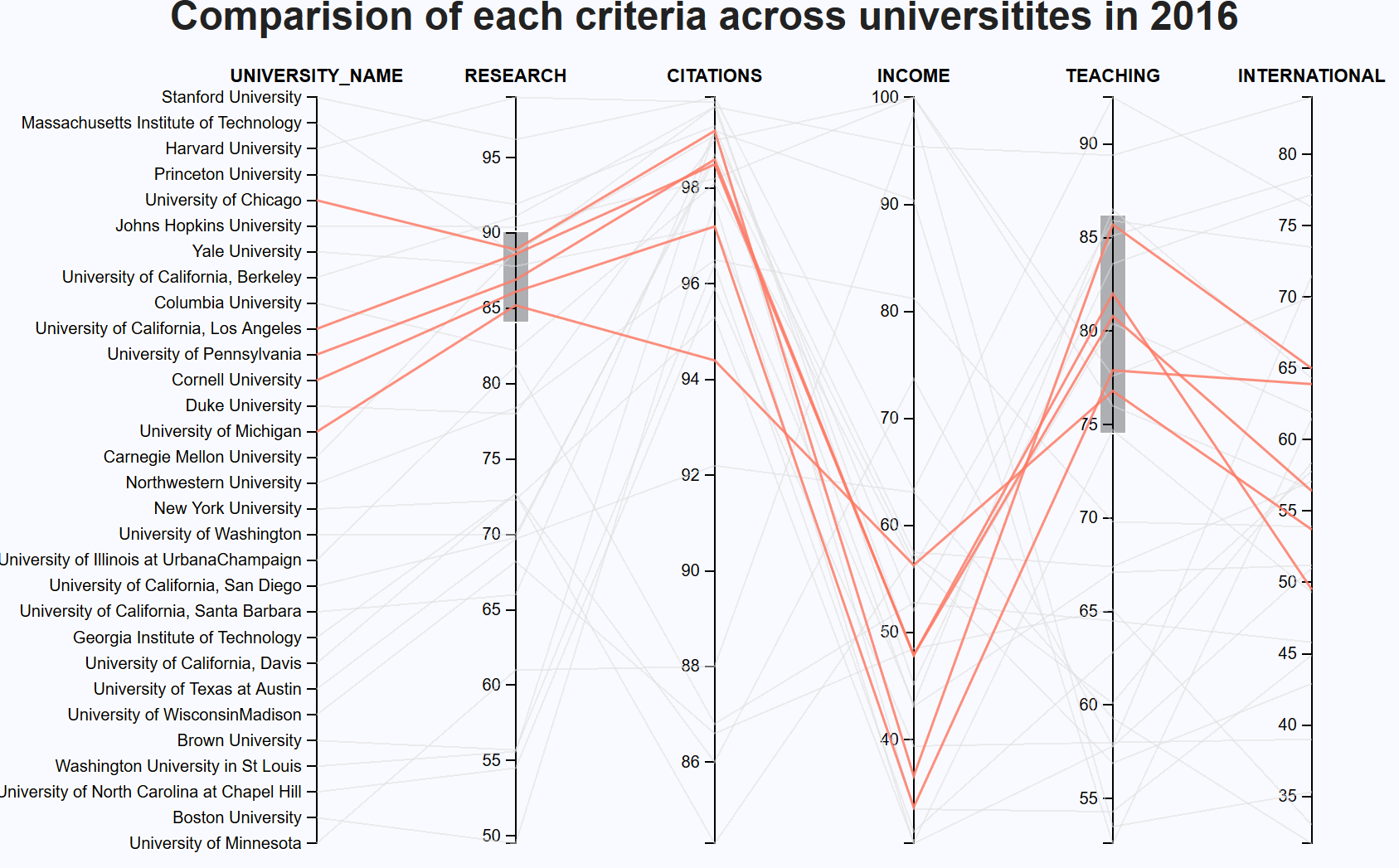
Now you click on your desired country in map. You get following two visualization to compare universities in that country.

* 1. **COMPARE CRITERIA OF EACH COUNTRY (PARALLEL COORDINATES)**

Now you chose United States to find your desired university. Here you wish to filter universities according to your preference. Suppose you want good research and moderate teaching then, accordingly you brush those axes to view your filtered universities.

OBSERVATIONS:

* United States distinctly had higher citations compared to other countries and other criteria.
* Germany had income on the higher end.
* United Kingdom has research and income are on the lower side.

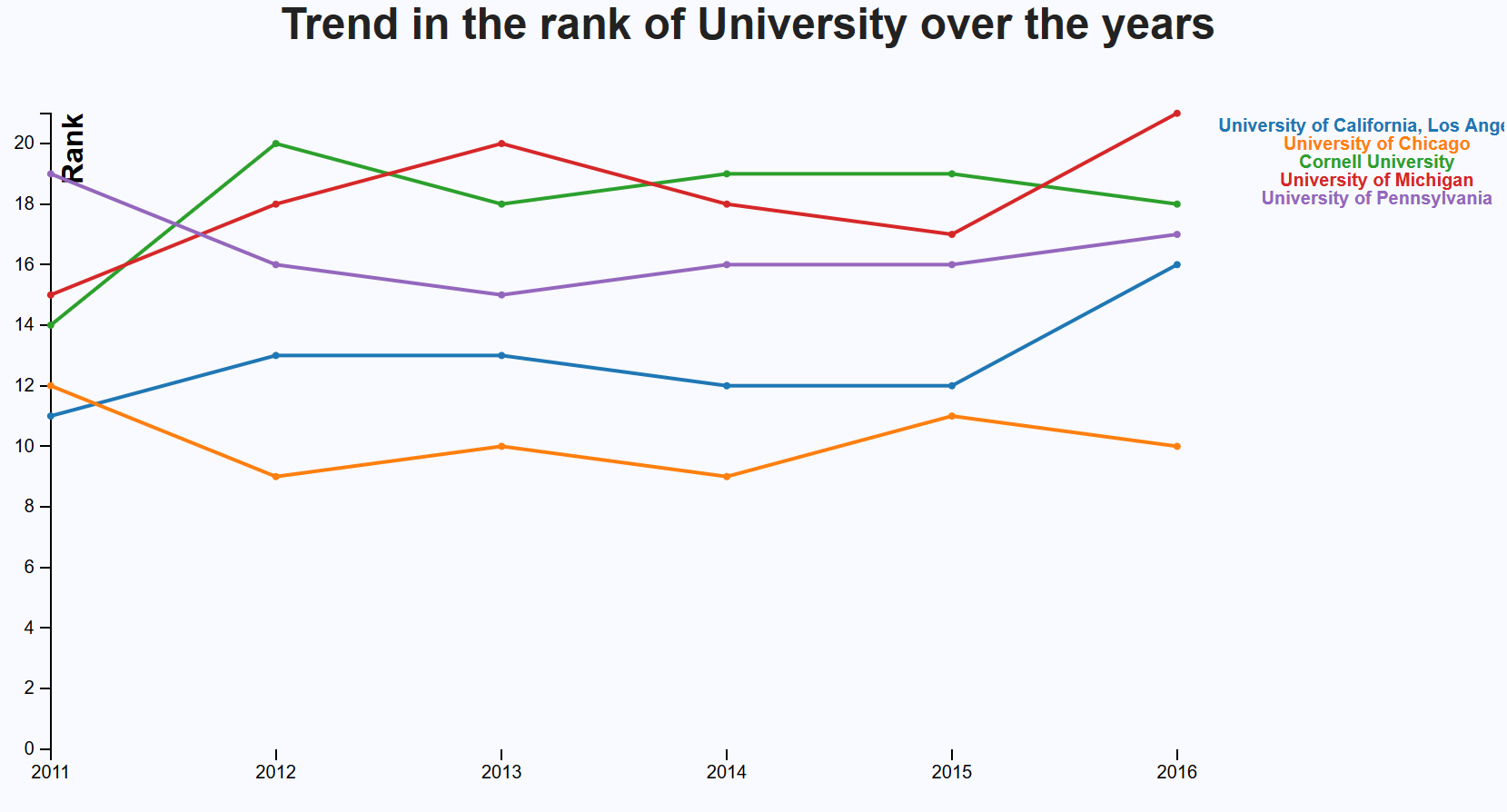
****

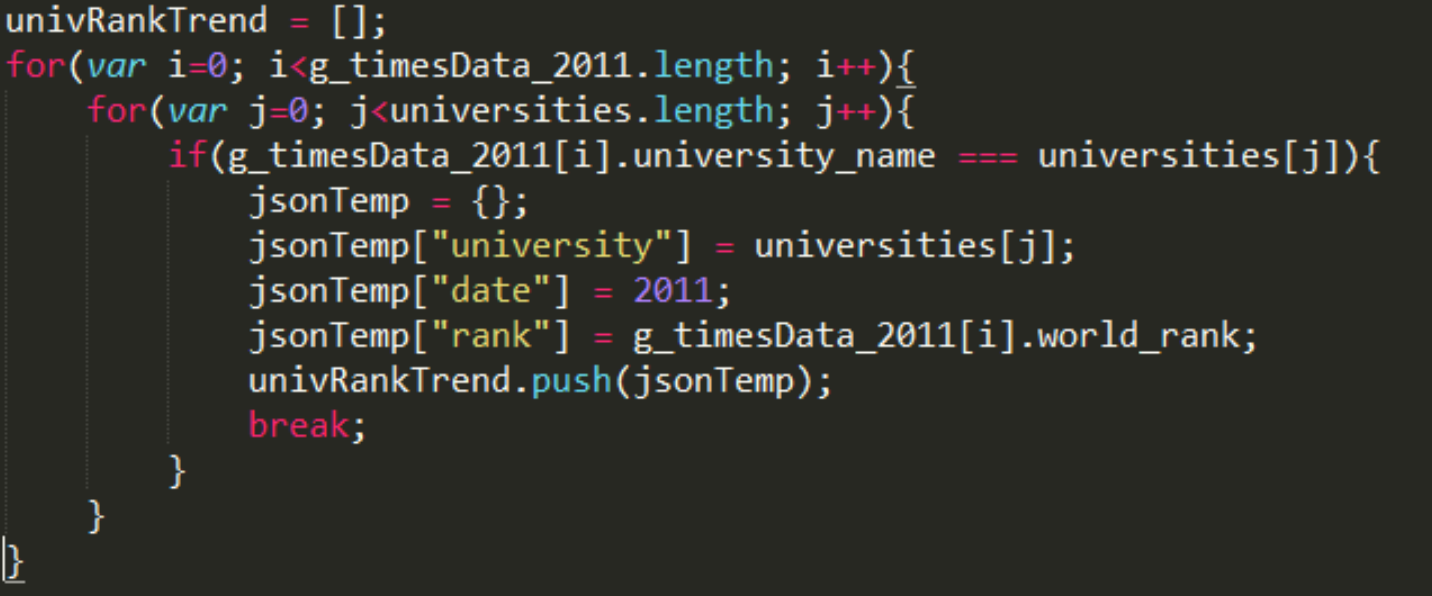
* 1. **TREND IN RANK OF UNIVERSITY (MULTI-LINE PLOT).**

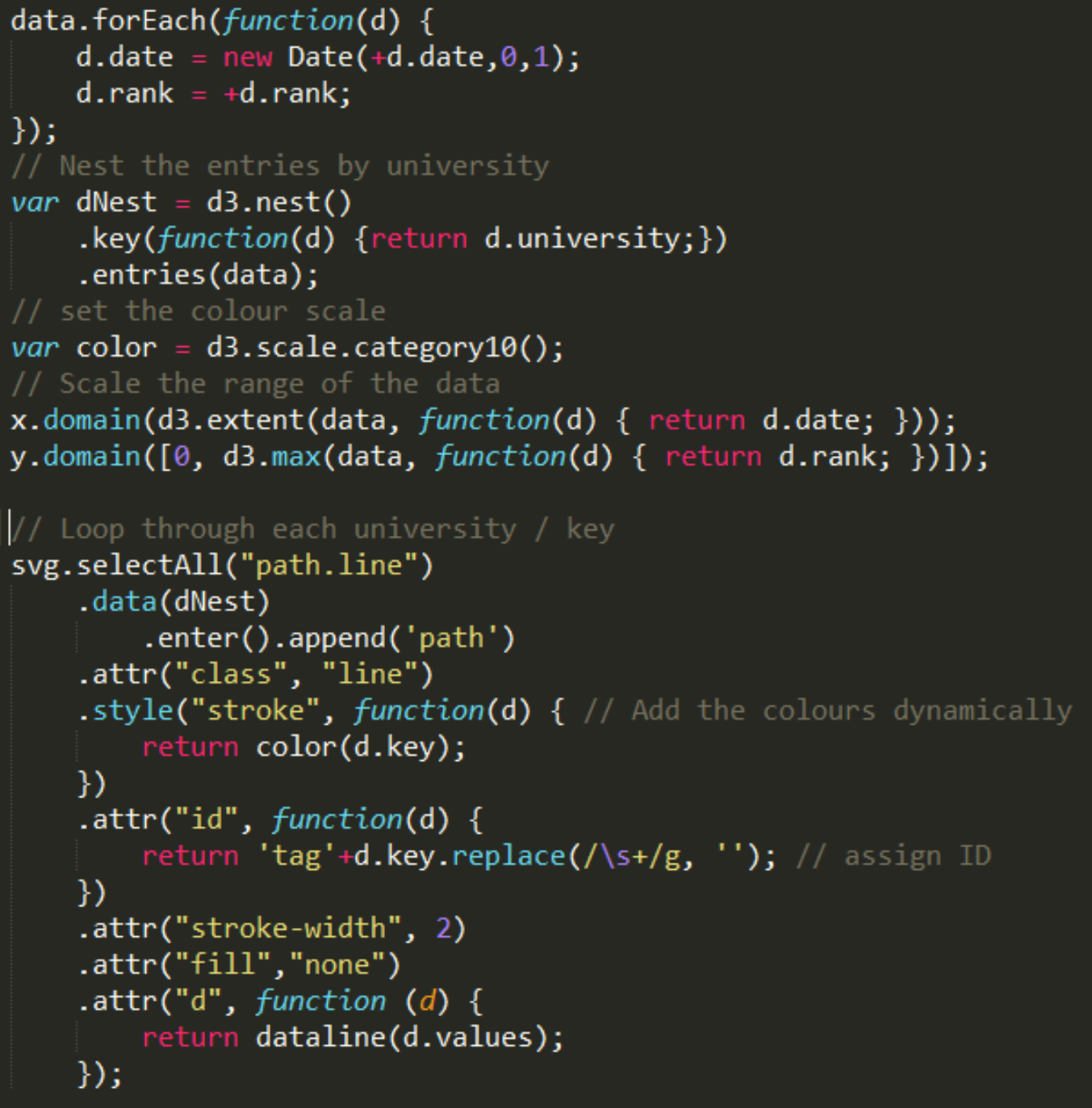
After you finished brushing, you can click on “Update Line Plots” to get updated multi-line plot showing the trend in rank of those filtered universities.

OBSERVATIONS:

* Most of the college rankings fluctuate a lot as their total scores are very similar. For example, 5 universities in United States have a total score as 87.
* Also, each college may bias their ranking depending on a criterion in abnormal proportion. For example, University of South Africa which we discussed was ranked high because of the huge number of students.

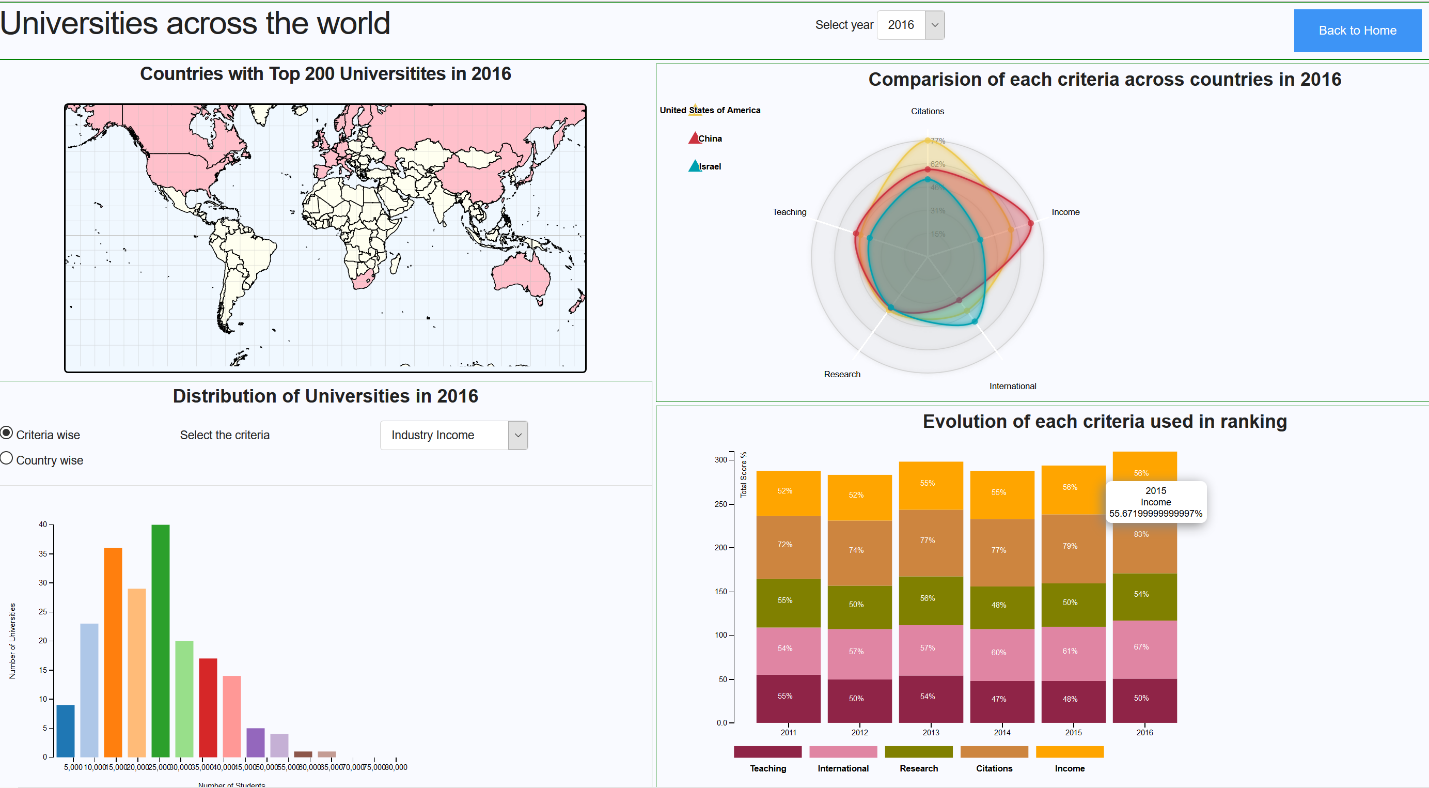


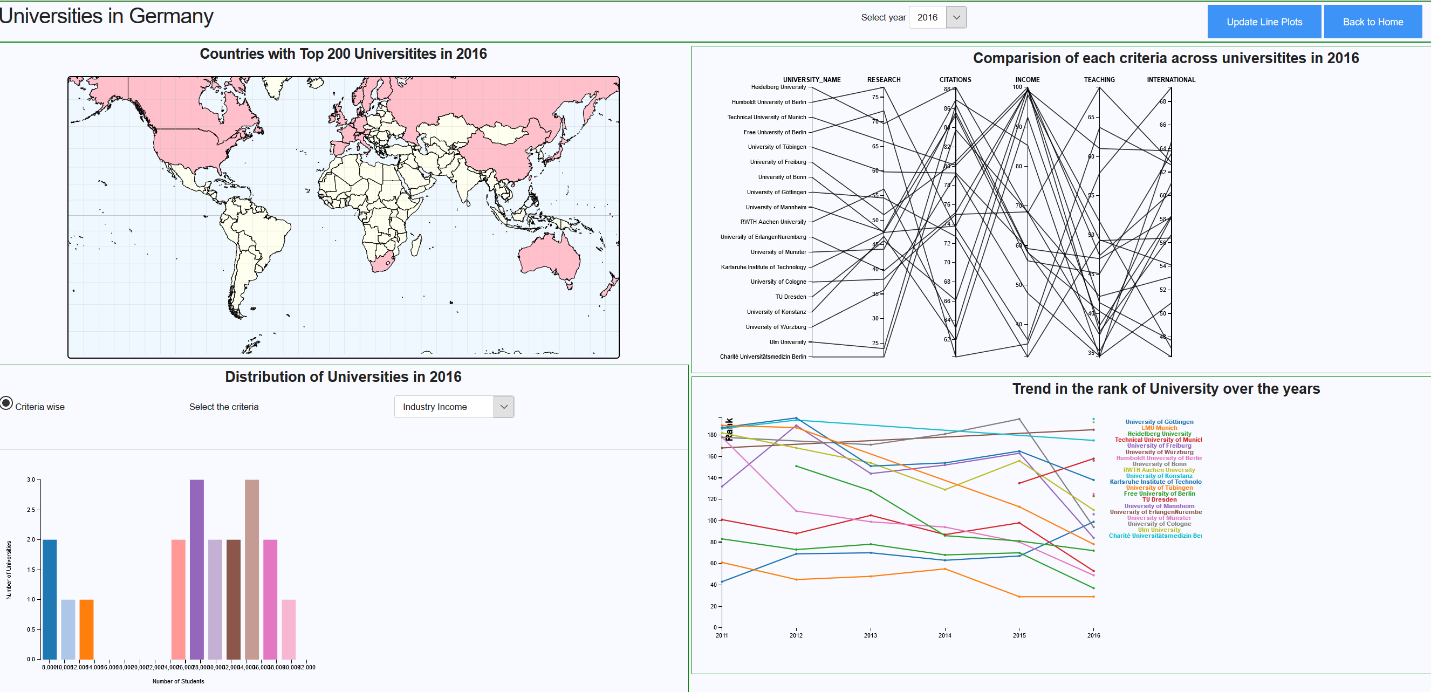




**PART 3: DASHBOARD**

* The data changes with year and criteria selected.
* Get country related data on choosing a country.
* Filter brushed data from parallel coordinates plot and plot multi-line plot.
* Seamless interaction to go back and forth country-wise data and world data.





**OBSERVATIONS AND CONCLUSION:**

1. This is a useful dashboard to shortlist universities based on different criteria like teaching, research, international score, income,and citations. The dashboard has two different components to compare universities across the world and universities within a country.
2. On an average only 5 of the total 200 universities have an average score of 95% and above.
3. University of South Africa is an outlier which is a mega university for distance education.
4. United States has a distinctly high citation score compared to other countries.Germany had income on the higher end. United Kingdom has research and income are on the lower side.
5. International and citation score tends to improve over the years while the income score remained steady with time.
6. Most of the college rankings fluctuate a lot as their total scores are very similar. For example, 5 universities in United States have a total score as 87.

**This project provides an interactive dashboard to find new insights about top universities across the world. The above dashboard is targeted towards students applying for universities across the world.**