CSE 564 Project Proposal

Swaminathan Sivaraman (110951180)
Sriram Sundar (110921718)

Demo Video: https://youtu.be/IGrcJqyNeWo

Background & Problem Description

There are numerous universities around the globe, and figuring out which university is best and ranking them based on things like education quality, research and so many other attributes is an ever-present problem. Which university is the best for research? Which university has the best faculty? Which one has the best labs? Which one will help to land the best jobs? Which one is affordable? These questions constantly need to be answered and every year, the data keeps changing and we also find newer metrics to analyze. More and more datasets about more and more universities also keep popping up, waiting to be explored. Our project aims to analyze some of this data and use interactive visualizations to try to learn something about the current state of universities across the planet.

Multiple ranking systems for universities exist. 3 of them are the most popular. The Times Higher Education World University Ranking ranks universities based on industry income, international diversity, teaching, research and citations. However, it has been criticised for undermining non-English instructing universities. The Academic Ranking of World Universities (Shanghai Ranking) ranks universities based on quality of education, quality of faculty, research output and per capita performance. But, it is said to value research more, undermining quality of instructions sometimes. The Center for World University Ranking ranks universities based on quality of education, alumni employment, quality of faculty, number of publications, citations, scientific impact and number of patents. As seen, there are multiple ranking systems and it is not clear which one is better or which attribute from which one should be preferred. Also, these rankings don't take into account the expenditure involved in the universities, one of the main factors that influences someone's decision to enroll in an university.

We aim to combine these rankings and also factor in the expenditure data for each university. Using visualizations, we want to show how the universities compare across different attributes. We will explore two avenues - one is to collate the rankings and try to visualize the university data using data from all rankings, the other is to compare the different ranking systems themselves and see which one is better for which attribute. This involves combining data from different datasets and analyzing them. We hope to find interesting information through our investigation.

Dataset Description

Dataset URL: https://www.kaggle.com/mylesoneill/world-university-rankings

The dataset consists of five primary CSV files which contain metrics used for evaluating the universities by the respective ranking institutions. The metrics used are also mentioned,

1. Times Higher Education World University Ranking - timesData.csv
    - world_rank, university_name, country, teaching (score for teaching & learning environment), international (international footprint - staff, students & research), research score (inclusive of reputation, volume & income), citations (research influence), income (industrial), total_score, num_students, sudent_staff_ratio, international_students, female_male_ratio, year (year of ranking)
    - Site: *https://www.timeshighereducation.com/world-university-rankings*
2. Academic Ranking of World Universities - shanghaiData.csv
    - World_rank, university_name, national_rank (rank of university within its country), total_score, alumni (alumni score, based on the number of alumni of an institution winning nobel prizes and fields medals), award (Award Score, based on the number of staff of an institution winning Nobel Prizes in Physics, Chemistry, Medicine, and Economics and Fields Medals in Mathematics), hici (HiCi Score, based on the number of Highly Cited Researchers selected by Thomson Reuters), ns (N&S Score, based on the number of papers published in Nature and Science), pub (PUB Score, based on total number of papers indexed in the Science Citation Index), pcp (PCP Score, the weighted scores of the above five indicators divided by the number of full time academic staff), year
    - Site: *http://www.shanghairanking.com/*
3. Center for World University Rankings - cwur.csv
    - world_rank, university_name, country, national_rank (rank of university within its country), quality_of_education (rank for quality of education), alumni_employment (rank for alumni employment), quality_of_faculty (rank for quality of faculty), publications (rank for publications), influence (rank for influence), citations (rank for citations), broad_impact (rank for broad impact), patents (rank for patents), score (total score), year
    - Site: *http://cwur.org/*
4. Education Expenditure data - education_expenditure_supplementary_data.csv
    - country, institute_type, direct_expenidture_type, 1995, 2000, 2005, 2009, 2010, 2011 (expenditure in these years)
5. University to country mapping - school_and_country_table.csv
    - school_name, country

Objectives

1. The data is split across multiple csv files - particularly, the expenditure data is in a separate csv file. We plan to combine with the expenditure data into the main rankings data. We will also identify if there are any attribute values missing for any data points and use some metrics (eg. mean of other values) to fill them up.
2. Since different ranking institutions use different metrics to evaluate the universities, the next step would be to preprocess and normalize the data to ensure consistency while visualizing.
3. Identify similarity amongst universities using clustering algorithms. We plan to initially use the k-means clustering algorithm and find the best value of k using an elbow plot. We also plan to use the parallel coordinate map to identify clusters.
4. Figure out which evaluation metric contributes the most to the ranking of a university using techniques like PCA. Eg. figure out how expenditure is likely to influence the ranking of the university.
5. Analyze how these 3 rankings compare with each other for different combinations of evaluation metrics.
6. Visualize the analyzed results in the form of pie-charts, bar-charts, line-charts, plots, maps, etc.
7. Try out novel visualization methods like Tree Map, Data Context Map etc.
8. Use these visualizations to show which university would suit which needs of a particular student.

Methods chosen to achieve project objectives

1. Cleaning up data
    - We will do this by some basic pre-processing of the csv files using Python, find missing values and replace them with appropriate values. We may also need to do sampling if the number of data points are too big to produce an understandable visualization.
2. Preprocessing and Scaling data
    - Standardization of datasets is quite essential in data processing and visualization. We plan to use Python's sklearn library to preprocess and normalize the data.
3. Clustering of similar universities
    - Clusters of similar universities could be identified using a parallel coordinate map. This would enable us to visualize which combination of metrics are most predominant across top-tier universities, mid-tier universities and the rest. We will use d3.js to do the visualization of the parallel coordinate map. We will also do k-means clustering using Python and find the best value of k.

4. Top evaluation metrics
    - We will identify the top evaluation metrics using principal component analysis and figure out the attributes having maximum sum of squared loadings. This would aid people to shortlist universities easily and enable them to make an appropriate

decision based on its most influential metrics rather than going through all of its metrics. We will use Python's sklearn library to do PCA and also produce scree plots.

5. Analysis of the three rankings
    - We will then compare the three ranking systems themselves. We will identify common attributes that are shared by all three systems and use those for comparison. We will then normalize these values and compare how the attributes influence the ranks across the three ranking systems.
6. Visualizations
    - We plan to majorly use bar charts, pie charts and scatter plots for visualization, as these are the plots that can best represent our data. We will make them interactive and modifiable. Once we have used these to identify the most interesting attributes, we plan to create more novel visualizations like Tree Map etc. for those attributes. We will also try to fuse the data and attributes together into one map (like a Data Context Map) to allow the user to visually see which universities are clumped together and which are far apart. We will use d3.js for all plot drawing and also some mapping libraries like Anymap.js.
7. System design
    - We will use Python to process and analyze data. We will also use a Python-based flask server as the backend to serve data. The front-end will primarily use d3.js to produce all plots.
8. Identifying and reporting results
    - Once we have done our analysis and produced all visualizations, we will first report the most important attributes. We will then use our visualizations as a guide to try and say which ranking is best for which attribute. We will also look for and report any new notable observations, if we find any.

Preliminary Report and Results

Swaminathan Sivaraman (110951180)
Sriram Sundar (110921718)

Current Progress

1. **Clean-up, Pre-processing and Scaling of data**
    a. The dataset had a few values missing for some attributes, for instance, a few universities in the **timesData.csv** file had some values missing for attributes like income, num_students and student_staff_ratio. We read the data into a pandas dataframe and interpolated these missing values from the averages obtained for these respective attributes. Further, we also removed the entries of a few universities which had more than 3 attribute values missing.
    b. We pre-processed and normalized the data using sklearn library. Though the range of most attributes was in [0,100], there were a few attributes like num_students which had quite vast ranges. Thus, we normalized the data so that we could do dimensionality reduction more accurately.
    c. Further, our dataset consists of ranking from three different ranking institutions. To analyse and visualize the relative comparison of a university with respect to all these three ranking systems, we merged and grouped the data (based on the university names) from individual data files into a single pandas dataframe.

2. **Identifying top evaluation metrics**
    a. We observed that all the ranking systems have multiple evaluation metrics to rank a particular university. This makes the lives of students difficult as they have to skim over all of these individual metrics to make a decision. Therefore we did a few steps to make this process easier.
    b. We first found out the intrinsic dimensionality of the data using PCA. We considered the axes (principal components) with eigenvalues greater than 1. The intrinsic dimensionality obtained for our data was 4. We also visualized the same using the scree plot as shown in **Fig.1**.

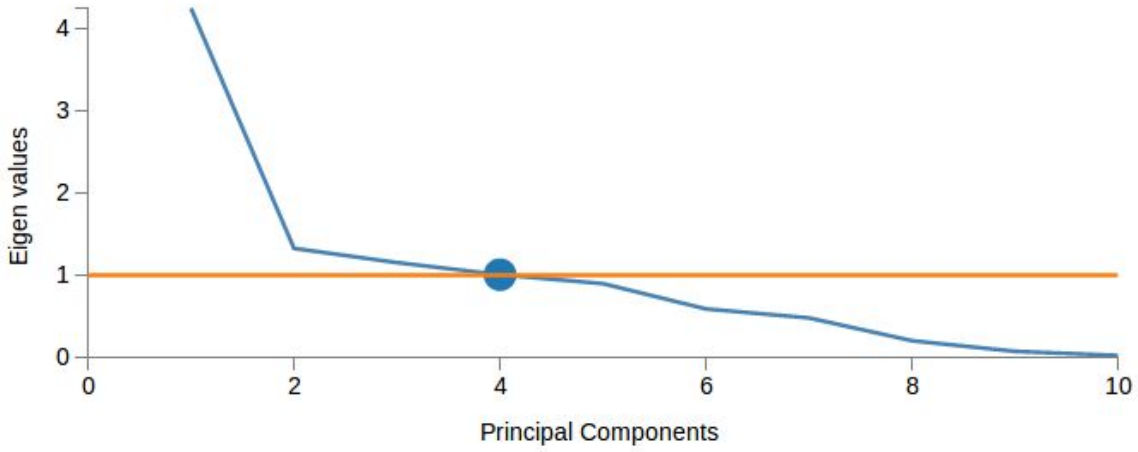

Fig.1 PCA Scree Plot

c. The next step was to identify the most important attributes that influence the ranking a particular university. For this, we obtained the top 4 attributes with highest PCA loadings. The top 4 attributes were **student_staff_ratio, num_students, total_score, income**. This also makes sense as we could categorize a university as lower ranked or higher ranked based on these 4 metrics. For example, top ranked universities would have **a low value of student_staff_ratio** (students can have more 1-1 interactions with the staff), **a low size of num_students** (top universities are more selective and admit a very few only), **a high total_score** (evaluation of infrastructural facilities like lab, etc.) and **a moderate income** (we observed that income is high only when there are a lot of students. Since they have a lesser number of students, the top ranked universities comparatively have a moderate income that's generally obtained from research grants, fundings, etc). All of these analysis and findings can also be verified from the parallel coordinate map shown in the forthcoming sections. We plotted the scatter plot matrix for the top 4 attributes as shown in **Fig.2**.

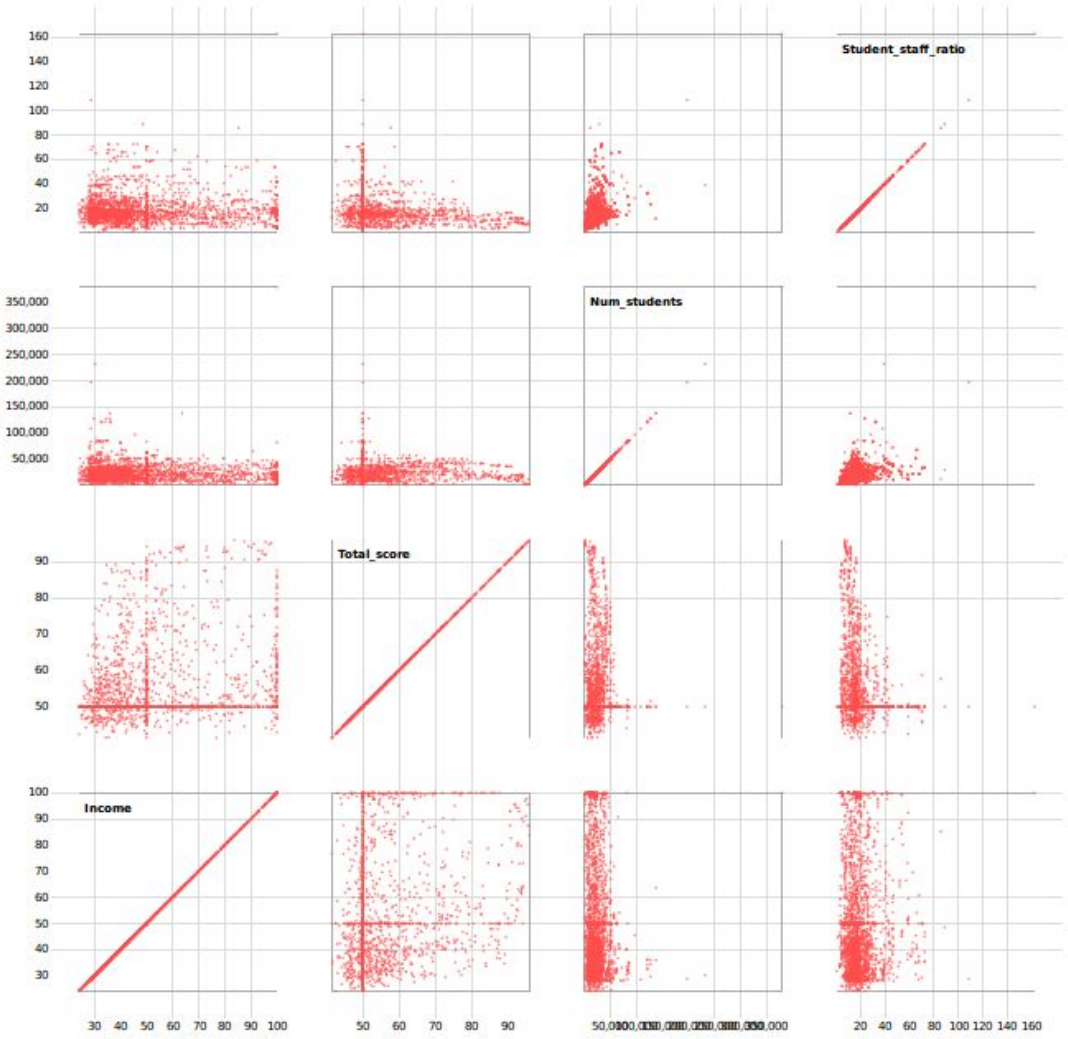

Fig.2 Scatterplot matrix for top 4 attributes

3. Analysis of the three rankings

- We used the merged data to analyze how a particular university has been ranked by the 3 different ranking institutions over the past 4 years. This would help us to evaluate the extent of similarity between the different ranking systems. We visualized it using a **time series plot - year VS ranking** as shown in **Fig.3**. The plot shown below is for University of Cambridge.

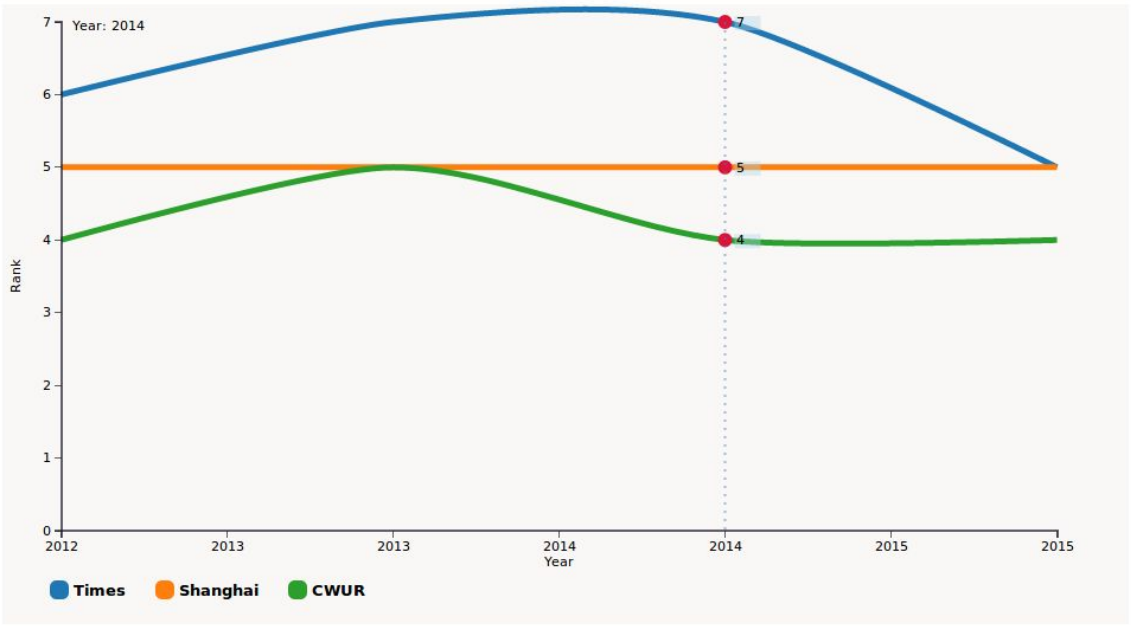

Fig.3 - University of Cambridge was ranked 4,5,7 by the three institutions in 2014

4. Parallel Coordinate - Analysis and findings

We used a parallel coordinate map to identify which combination of metrics are most predominant across top-tier universities, mid-tier universities and the rest. Further we also implemented a brush for filtering purposes. For instance in **Fig.4a**, we use a brush on the first coordinate (world_rank). We filter out the lower ranked universities using the brush and analyze the corresponding metric values in the other parallel coordinates. We observed that the lower ranked universities have **a low teaching score, less research activities, lesser citations, low student_staff_ratio**, **a moderate size of num_students**, **a low total_score**, **varied income.**

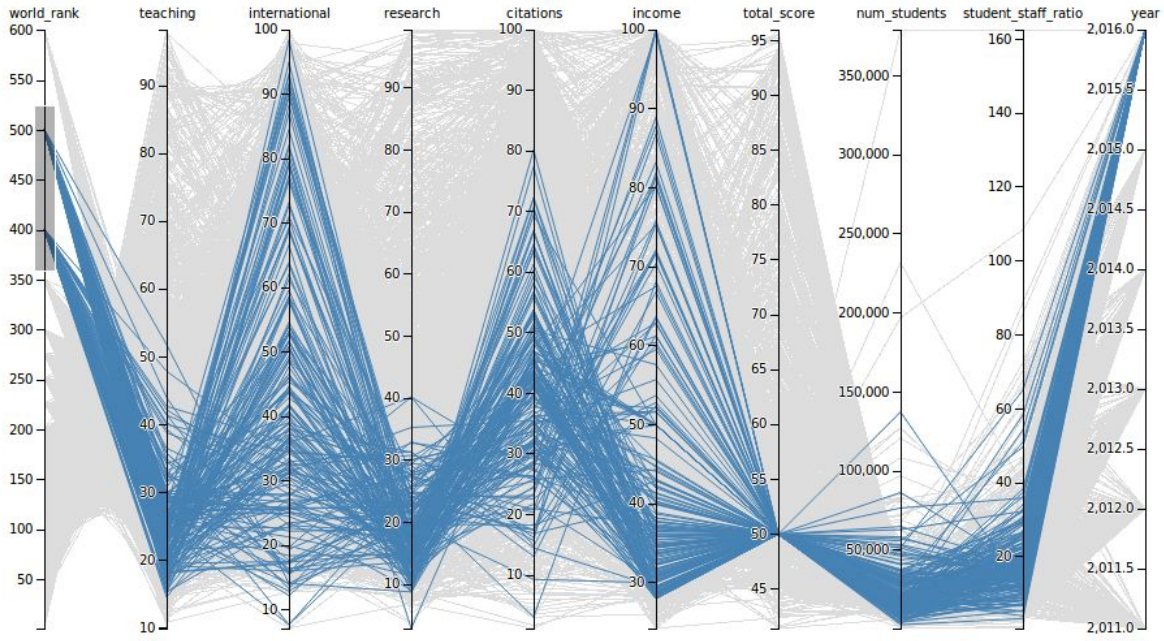

Fig.4a - Parallel coordinate plot with brushes applied to lower ranked universities

Similarly, as shown in **Fig.4b** we observed that the top ranked universities have **a high teaching score, a low student_staff_ratio, a high research score, higher number of citations, a very low num_students, a high total_score and varied income.**

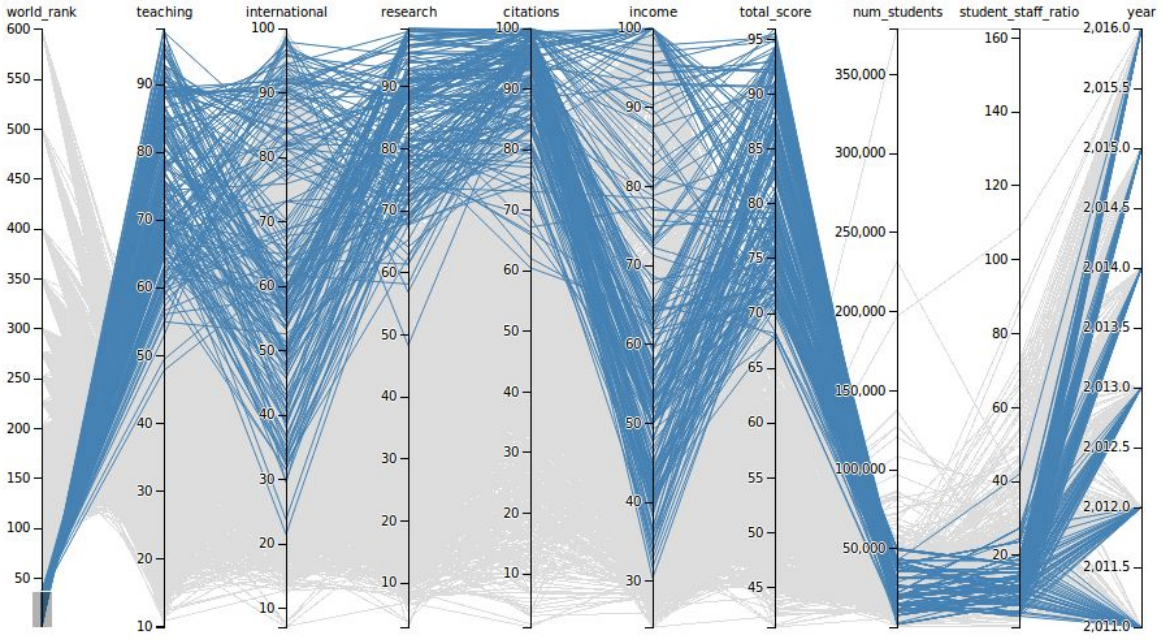

Fig.4b - Parallel coordinate plot with brushes applied to higher ranked universities

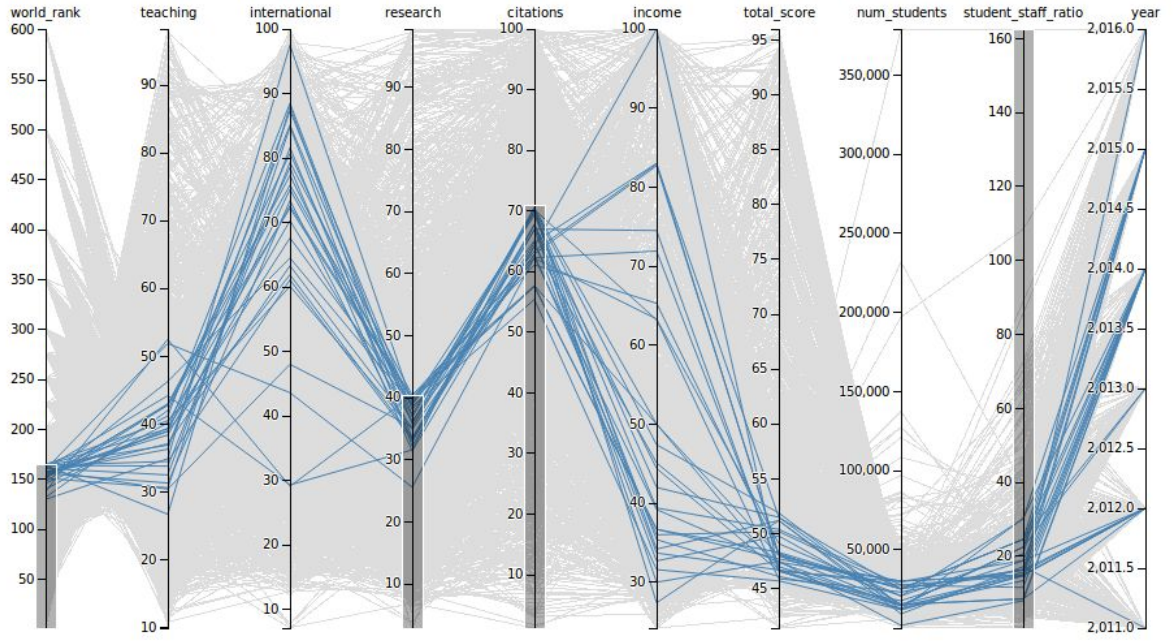

Fig.5a - Parallel Coordinate Plot with brushes applied to top-ranked and low-research universities

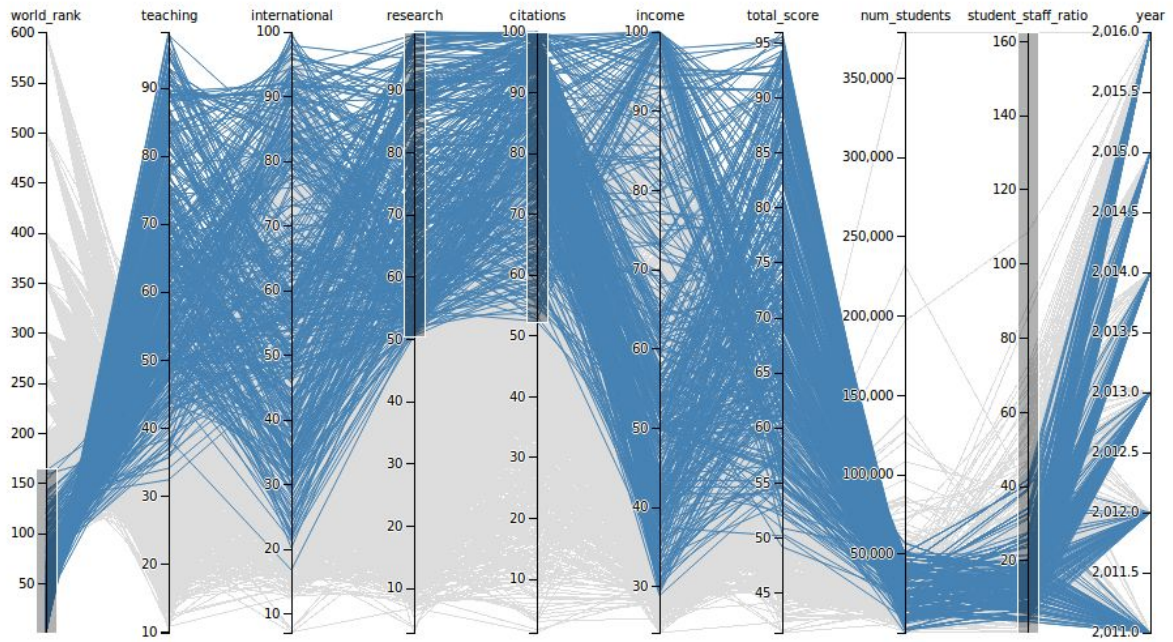

Fig.5b - Parallel Coordinate Plot with brushes applied to top-ranked and low-research universities

We took the parallel coordinate plot and used a brush and selected only the 150 top-ranked universities. Then we applied more brushes to select universities with low research content and citations. The final result is shown in **Fig 5a**. We can clearly see that

there are very few universities that fit this criteria - thereby showing very few top-ranked universities have low research or citations. A curious observation is that all such universities happen to have higher number of international intake.

We then tried out the reverse. For the same 150 top-ranked universities, we created brushes to select ones with high research content and citations **(shown in Fig. 5.b)**. Immediately, the plot becomes dense, proving again that top-ranked universities do a lot of research work. However, the industrial income that can be obtained after graduating from these universities is still about average for a lot of such universities. Good research work doesn't seem to be rewarded with heavy income, apparently.

5. Bar graph and Pie chart distribution

We took the four most important attributes in the data based on our PCA analysis - student_staff_ratio, num_students, total_score and income. We also picked 3 other attributes that we thought were interesting - research, citations and teaching. We created option choosers for the attribute and year. The user will choose a particular attribute and a particular year. We will then filter the data based on that attribute and year and then bin it and create a fixed-width histogram. The y-axis of the histogram will contain the number of attributes while the x-axis will be the chosen attribute. Hovering over a bar will expand a bar and highlight it. Now, the user might be also interested in how the data is distributed within a bin. So, we created a pie chart that distributes the number of universities based on number of countries. The visualization **(Fig.6)** will plot the pie chart for the top 5 countries, since many countries have only one university and it would be useless to plot that. When we hover over a particular bar, the pie chart transitions to represent the distribution of universities across countries only in that bar. The visualization in action is shown below. The total_score attribute for the year 2016 has been binned and plotted. When the user hovers over the bar with size 49, the pie chart transitions to show from which countries (only top 5) those 49 universities come from.

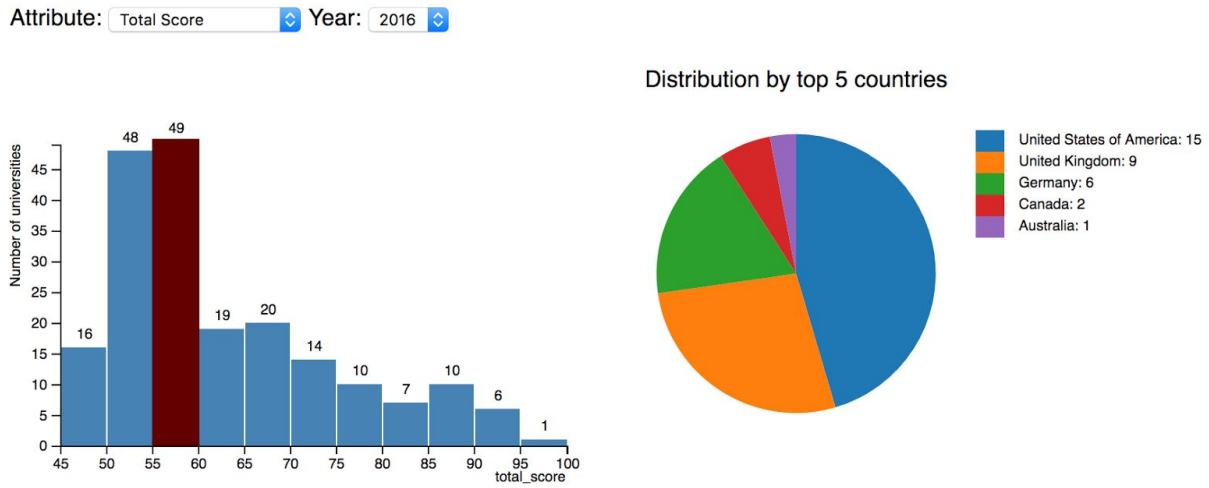

Fig.6 Visualization of distribution of the Total score attribute for the year 2016

Remaining work

While working with the data, we observed that users exploring the data would be most interested in three things - the rankings comparison, how universities are spread out for a particular attribute, and most importantly the statistics related to one particular university. So, we have decided to include a map that will initially be an overview of the entire global map. All university points will be marked on the map and country regions as well will be marked. When the user hasn't clicked on anything, he/she will be seeing visualizations for all universities across all countries. This would involve the parallel coordinates visualization, the scatter plot matrices of the most important attributes determined using PCA, and the bar and pie charts that show a distribution of one particular selected attribute. Next, the user can select a country region on the map and that will act as a filter, and only visualizations for universities for that university will be shown. Finally, the user can click on any one particular university on the map, and that will load up visualizations pertaining to that university alone (like the time-series visualization shown above). So, the map will act as a driver of the visualization and will allow users to interact with the data easier since the world map is something everyone is familiar with.

We have created many individual visualizations and already shown them here. But we will integrate them into a dashboard whereby a user can see all visualizations in one go. We will also add brush effects such that brushing done on one visualization can create effects in other visualizations in the dashboard as well. We will essentially have two dashboards, placed under two tabs. One dashboard will contain all university visualizations while the other dashboard will contain per-university visualizations.

Final Report (contd.)

There are numerous universities around the globe, and figuring out which university is best and ranking them based on things like education quality, research and so many other attributes is an ever-present problem. Which university is the best for research? Which university has the best faculty? Which one has the best labs? Which one will help to land the best jobs? Which one is affordable? These questions constantly need to be answered and every year, the data keeps changing and we also find newer metrics to analyze. More and more datasets about more and more universities also keep popping up, waiting to be explored. When a student tries to make a decision on choosing a university, he/she would have to take a lot of parameters into consideration. In a nutshell, our project analyses the ranking dataset and uses an interactive dashboard to learn about the current state of universities across the planet.

We observed that users exploring the data would be most interested in three things - the rankings comparison, how universities are spread out for a particular attribute, and most importantly the statistics related to one particular university. So, we have included a **zoomable heatmap (Fig.1)** that will initially be an overview of the entire global map representing the concentration of the universities across the globe.

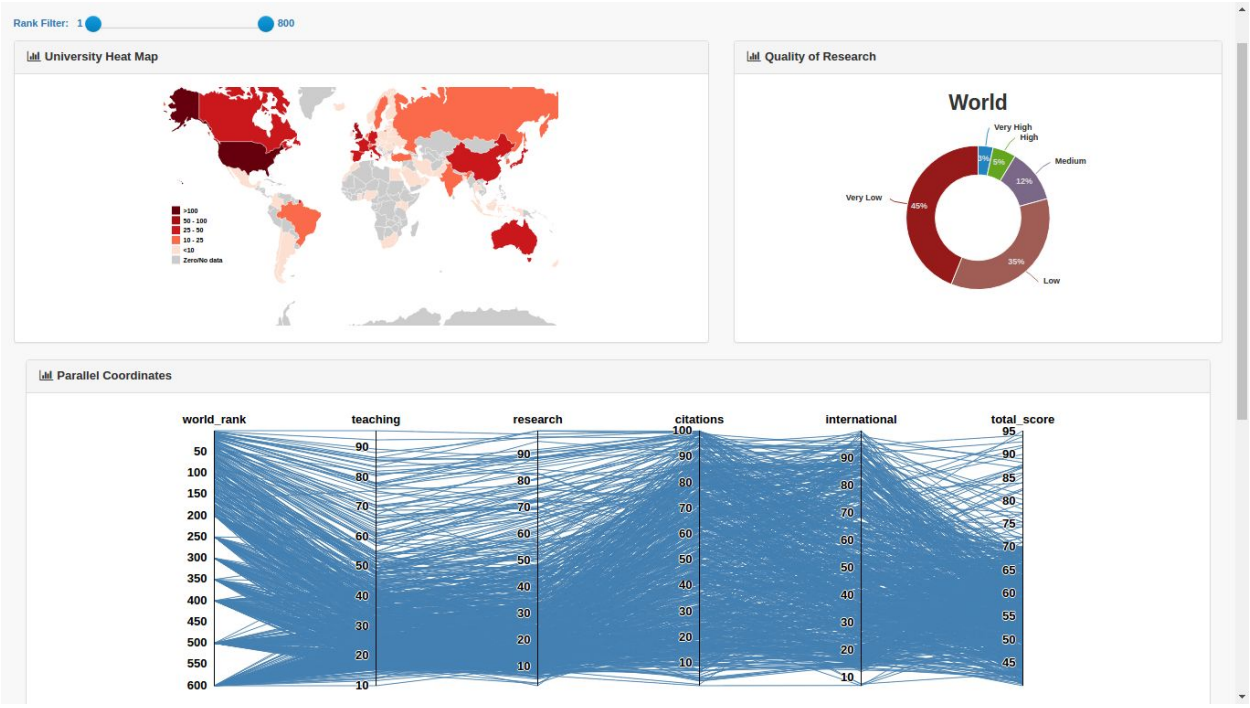

Fig.1 Dashboard with heatmap, donut chart and parallel coordinates

In Fig.1, we can observe the heatmap of all university concentrations across the globe. Since research quality is an important parameter while determining rank of a university, we can see a donut chart on the right representing the quality of research across all universities in the world. **We observe that ~80% of the universities in the world have low/very low research quality.** This states that out of the 800 universities that we have in our dataset, around 600 of them focus less on research and the remaining 200 have active research going. This helps us to get an general idea of how active research is done across global universities.

Students who decide to pursue their education in a specific country would want to see country specific statistics to determine which university and country would be apt for them. For instance, when we (the authors) decided to pursue higher studies, we first decided to pursue it in the USA. The next thing we did was to find out the top universities in that particular country. Hence, even though a global perspective is good, it always helps to have country specific details. **Thus, we have a couple of filters for the users to analyse the the universities within a specific rank range and in specific countries.** One filter is to filter the universities based on rank. So, when a person wants more information about universities in a specific rank range **(Fig.2),** he can apply this particular filter.

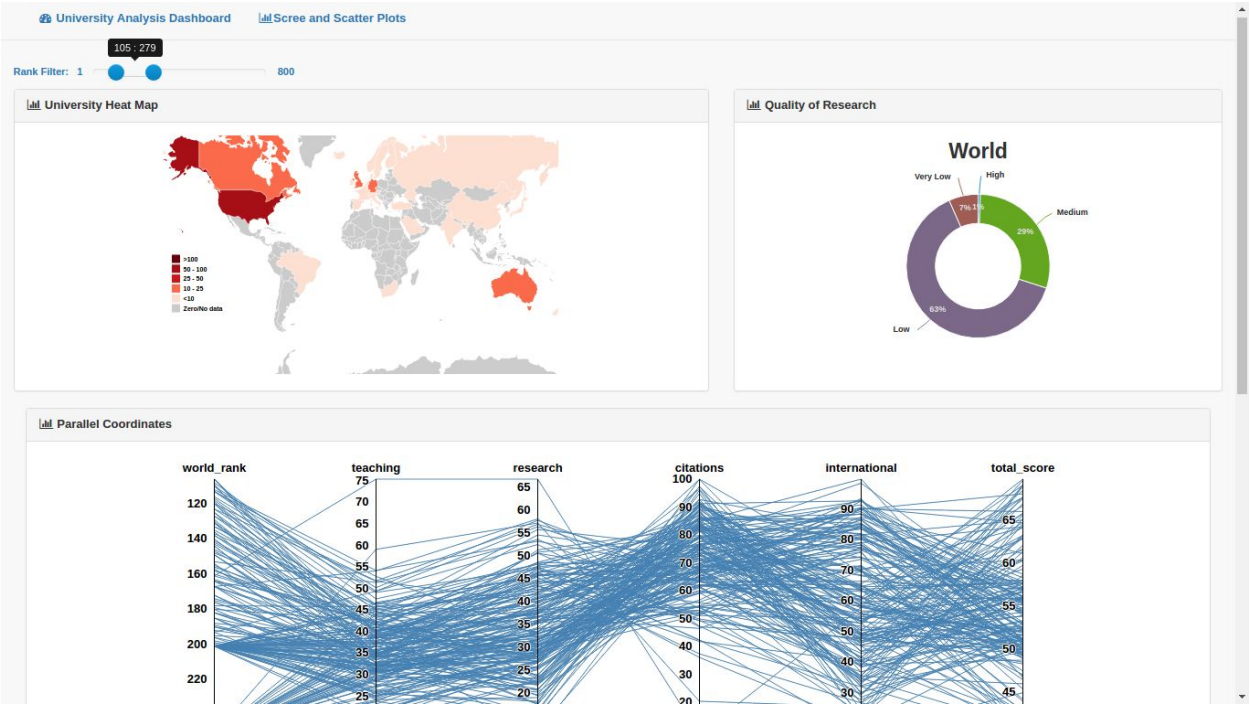

Fig.2 Rank filter applied - universities ranked between 105 and 279

In Fig.2 we can observe the rank filter applied. One important observation here is that **only 1% of the universities have high quality research.** This is quite indicative of the fact that the universities in this

particular range (105-279) do not stress much on research activities. We'll identify some patterns for top ranked universities in the forthcoming sections.

Having filtered based on rank, the next task is to filter based on countries **(Fig.3)**.

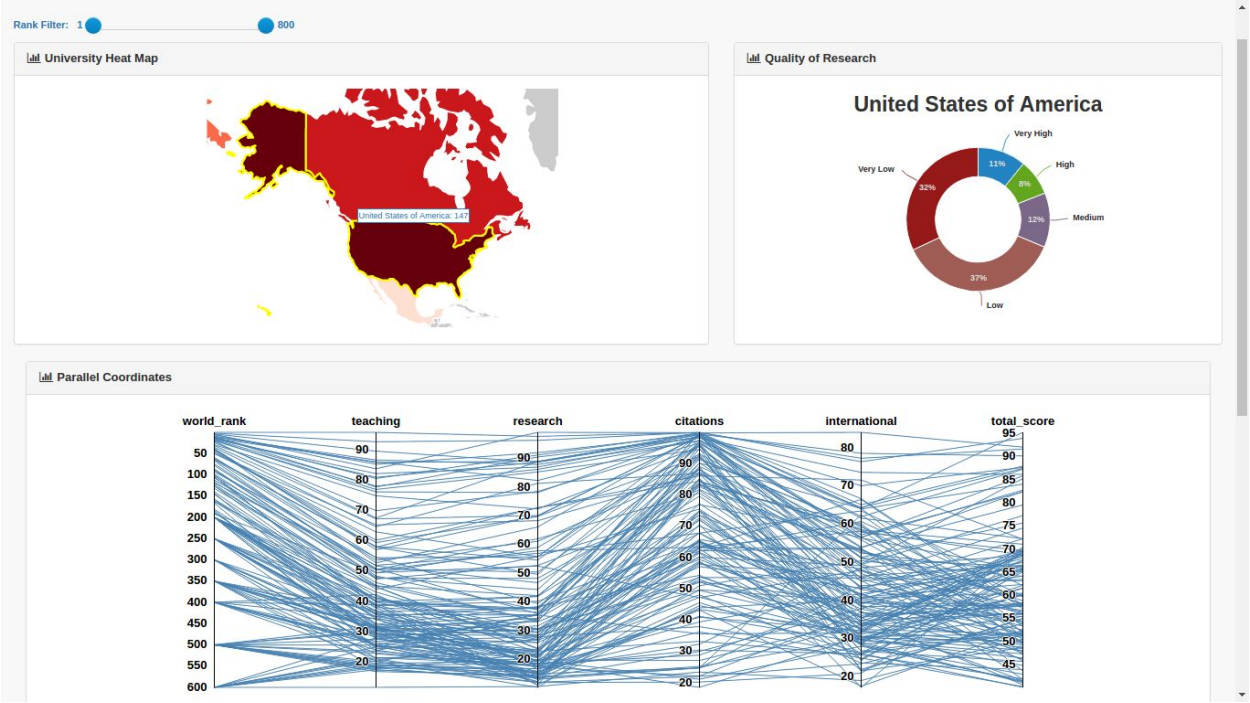

Fig.3 Universities in USA

In Fig.3, we have all universities in the USA. **In Fig.1, we had observed that only 8% (high + very high) of the universities had excellent research activities.** However, when we look at the USA, we observe that **19% (high + very high) of them have excellent research going on**. Comparative to other countries, **universities in the USA tend to focus more on research and the results are reflected on the parallel coordinate plot as well**. We can see universities with high research scores in the parallel coordinates.

Next, we apply further filters on the parallel coordinate plot using brushing techniques. We use this to find interesting insights on the features that contribute to ranking of the university. **In Fig.4 and FIg.5**, we identify the correlation amongst different ranking metrics used to rank a specific university using the parallel coordinate plot.

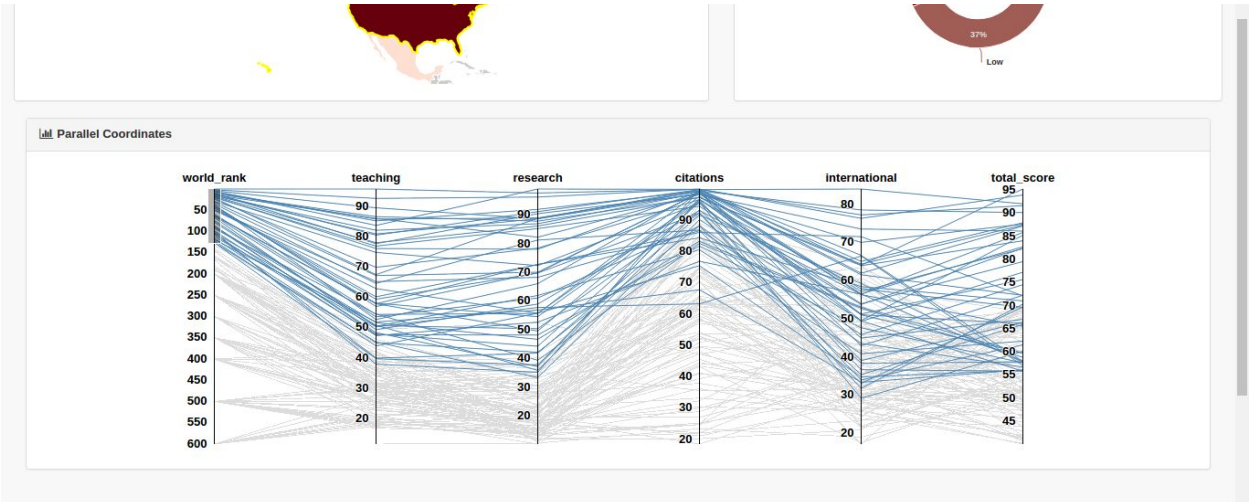

Fig.4 Parallel coordinate plot for top universities in the USA

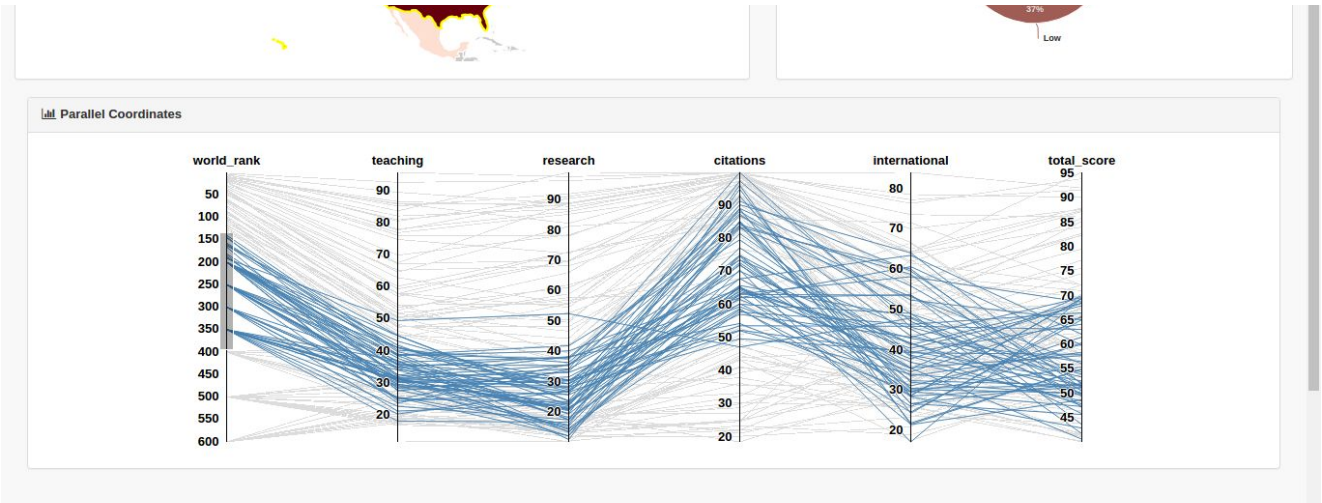

Fig.5 Parallel coordinate plot for lower ranked universities in the USA

We find that for all **top ranked universities in the USA**, the **research score and citations score is quite high**. This makes us to believe that there is a linear positive correlation between rank and research/citations. However, if we notice the plot for **lower ranked universities** (note the brush applied on the world_rank column), we observe that they **have less research scores but high citations score**. We thus conclude two key results from this,

1. world_rank and research have a positive linear correlation
2. world_rank and citations have no correlation

One possible explanation for a high citation scores despite a low research score is that quality research happens at lower ranked universities as well but not at the same scale compared to top universities. Though less research occurs at these universities, they are highly cited across the academic fraternity.

Further, if we look at the other attributes, **not all top ranked universities have high international impact/score**. However **for total_score, top universities have better total_scores (65 - 95) and lower ranked universities have a lower range of total_scores (40 - 70).** We thus conclude two key results from this,

1. world_rank and international impact have no correlation
2. world_rank and teaching score, world_rank and total_score have a positive linear correlation

We also added support for multiple brushes **(Fig.6)** to support multiple filters for students.

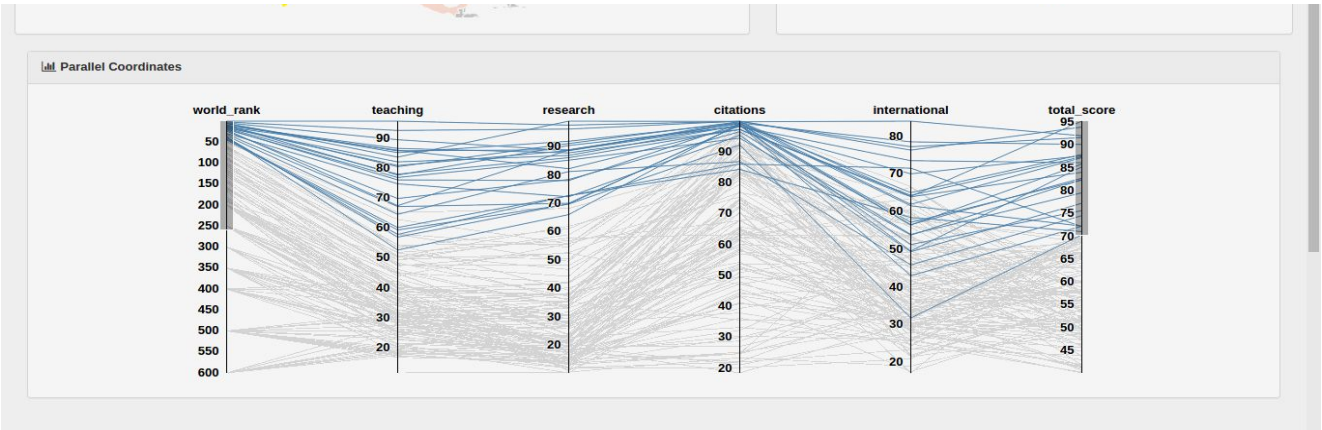

Fig.6 Parallel coordinate plot for top universities with high total_scores in the USA

We observe that **top universities with high total_score** are the universities which students dream of as they have high scores for all key ranking metrics like teaching, research, citations, etc.

However, a similar trend was not observed in UK. Most of the universities in the UK had a low teaching_score, a low research score and a low total_score as shown in **Fig.7**.

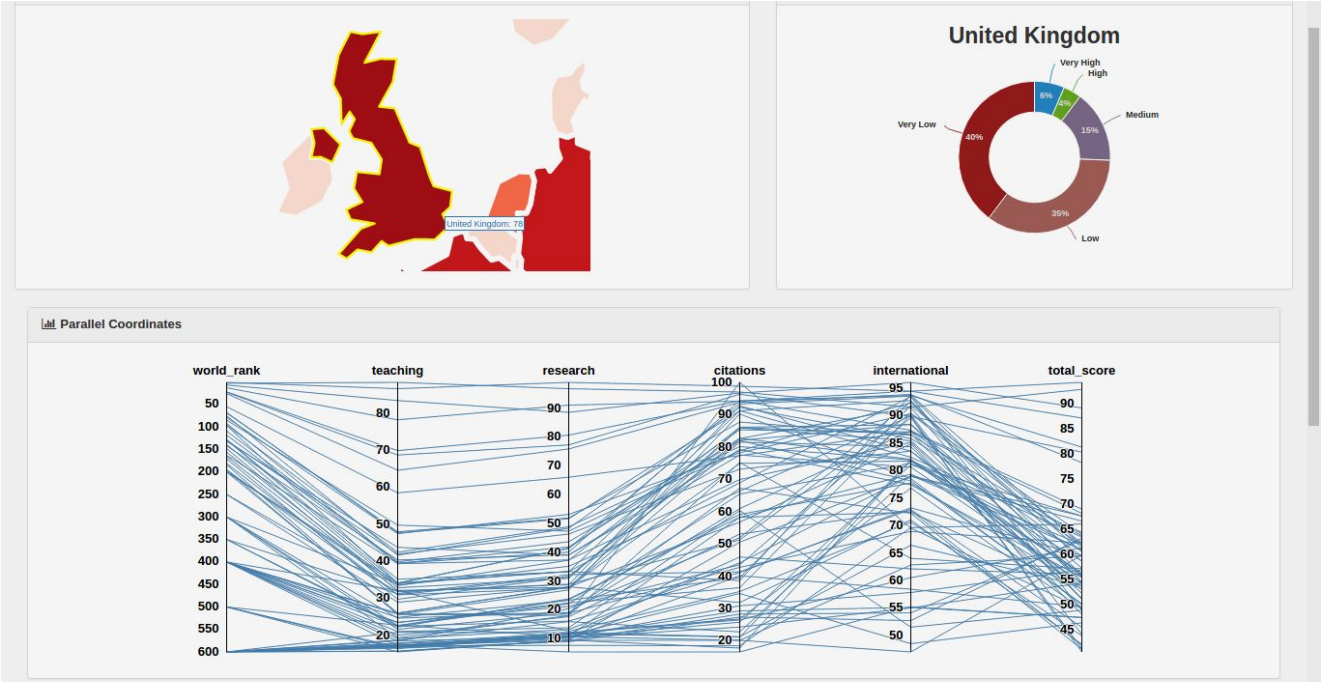

Fig.7 Parallel coordinate plot for universities in the UK

Next, we analyse **university specific details**. We have provided a dropdown to select a particular university. For the selected university, we show 3 visualizations.

1. **Ranking comparison via Time Series plot**: Since we use the rankings of universities from three different ranking systems, we compare the ranking of a single university across the 3 ranking standards. In our dataset we had ranks of all universities from 2012 - 2016 for all 3 ranking systems. We use this plot to compare the ranking systems.
2. **Student-staff ratio via Pie Chart plot:** Another important factor in grading a specific university is to identify its student-staff ratio. The lower the student to staff ratio is, the better it is for a university as more students would gain individual attention. This is also an important attribute obtained from PCA analysis.
3. **Return on Investment (ROI) via custom fillable vector:** Students particularly look at the ROI prior to selecting a university. A student would only choose a university if and only if it would be beneficial for his/her career in the long term. Therefore, we have visualized this using a novel method. We obtained an svg vector online for the dollar image, and then used linear gradients to fill it up with color proportionate to the ROI value.

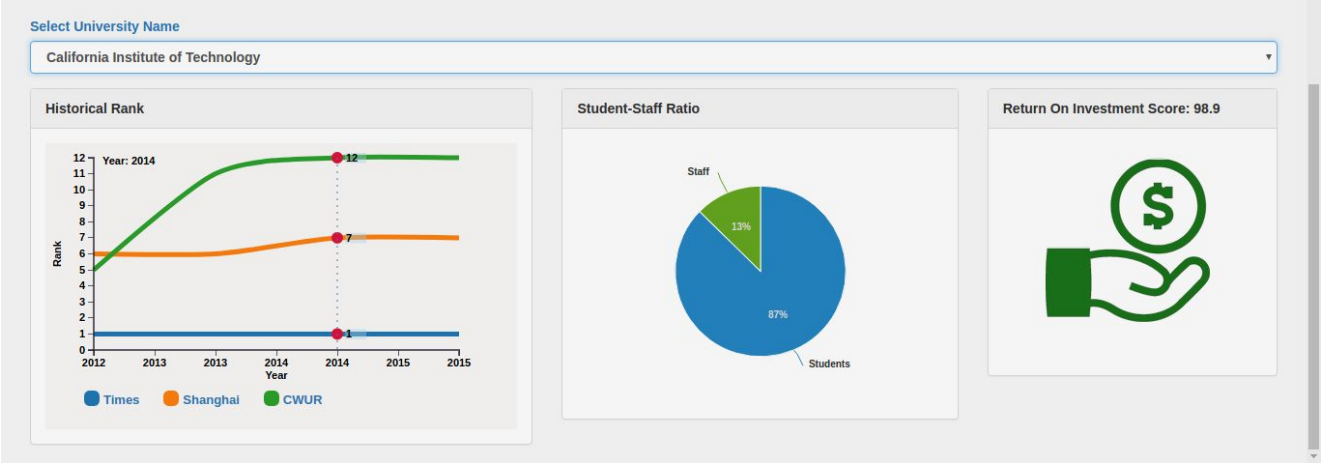

Fig.8 CalTech University details

We observe from **Fig.8** that the three ranking systems were quite varied in their evaluations. One key observations in this was that **even though top ranked universities had good student-staff ratio, they didn't have high ROI in several instances.** On the other hand, **some lower ranked universities had extremely good ROI.** One plausible reason for this is that a few top ranked universities are highly research-centric and do not have high ROI. Whereas other universities are located at prime locations and have a lot of industrial connection. This enables such universities to have high ROI. For example, in **Fig.9**, we observe that one of the best universities in the world (UCB) has a high rank but low ROI. On the other hand, in Fig. 10, we observe that Duke university has a lower rank but exceptional ROI (100%).

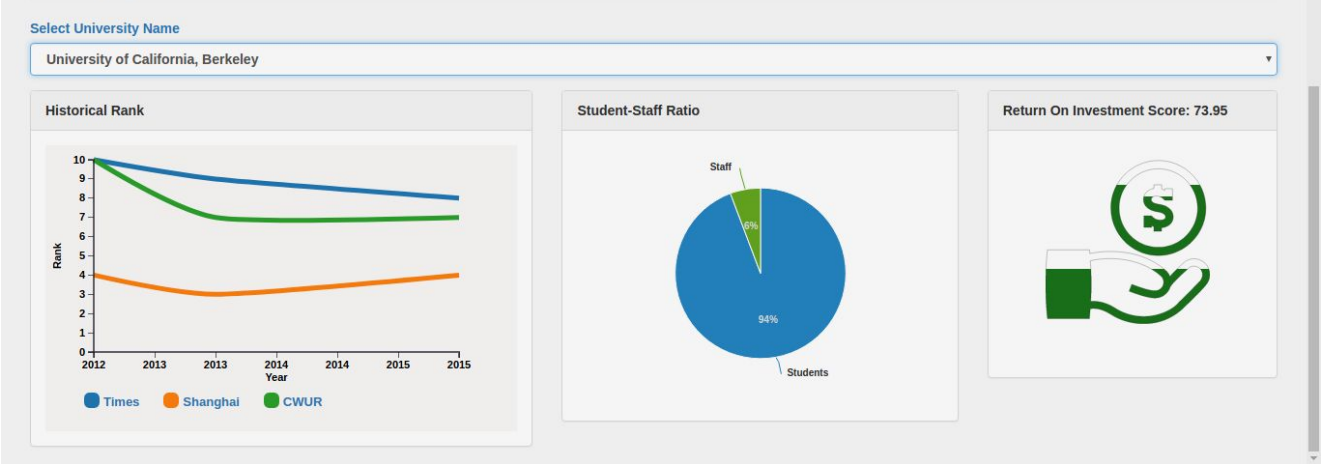

Fig.9 University of California, Berkeley details

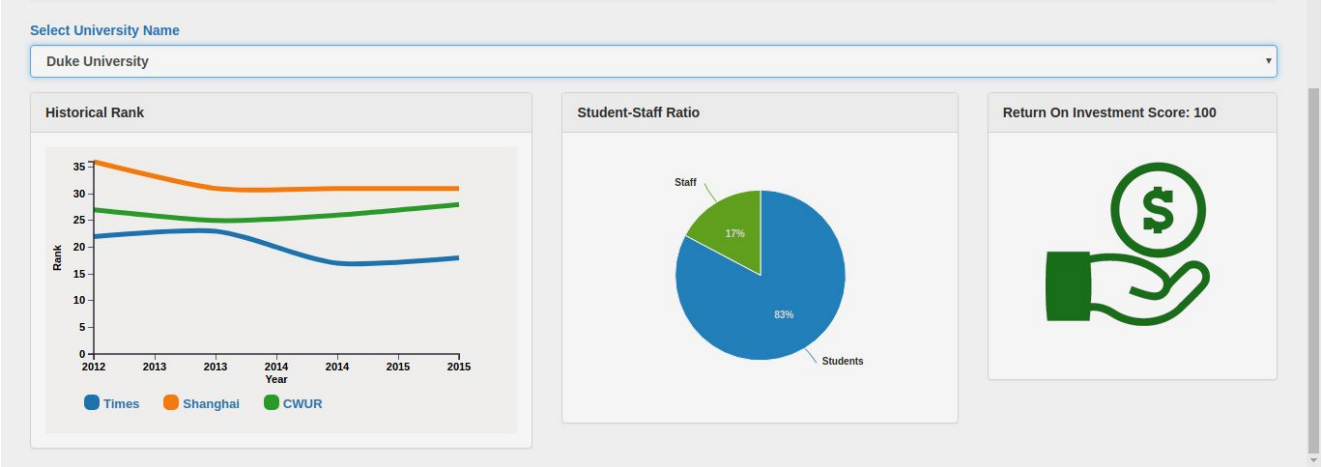

Fig.9 Duke University details

An important reason for Duke University's high ROI is its prime location. It is in North Carolina where the cost of living is quite reasonable and perhaps their tuition fee might not be as exorbitant as well. Thus it makes sense for its high ROI.

Apart from analysing the universities globally, country-wise and individually, we also did some data data analysis using PCA. We observed that all the ranking systems have multiple evaluation metrics to rank a particular university. This makes the lives of students difficult as they have to skim over all of these individual metrics to make a decision. Therefore we did a few steps to make this process easier.

We first found out the intrinsic dimensionality of the data using PCA. We considered the axes (principal components) with eigenvalues greater than 1. The intrinsic dimensionality obtained for our data was 4. We also visualized the same using the scree plot as shown in **Fig.10**.

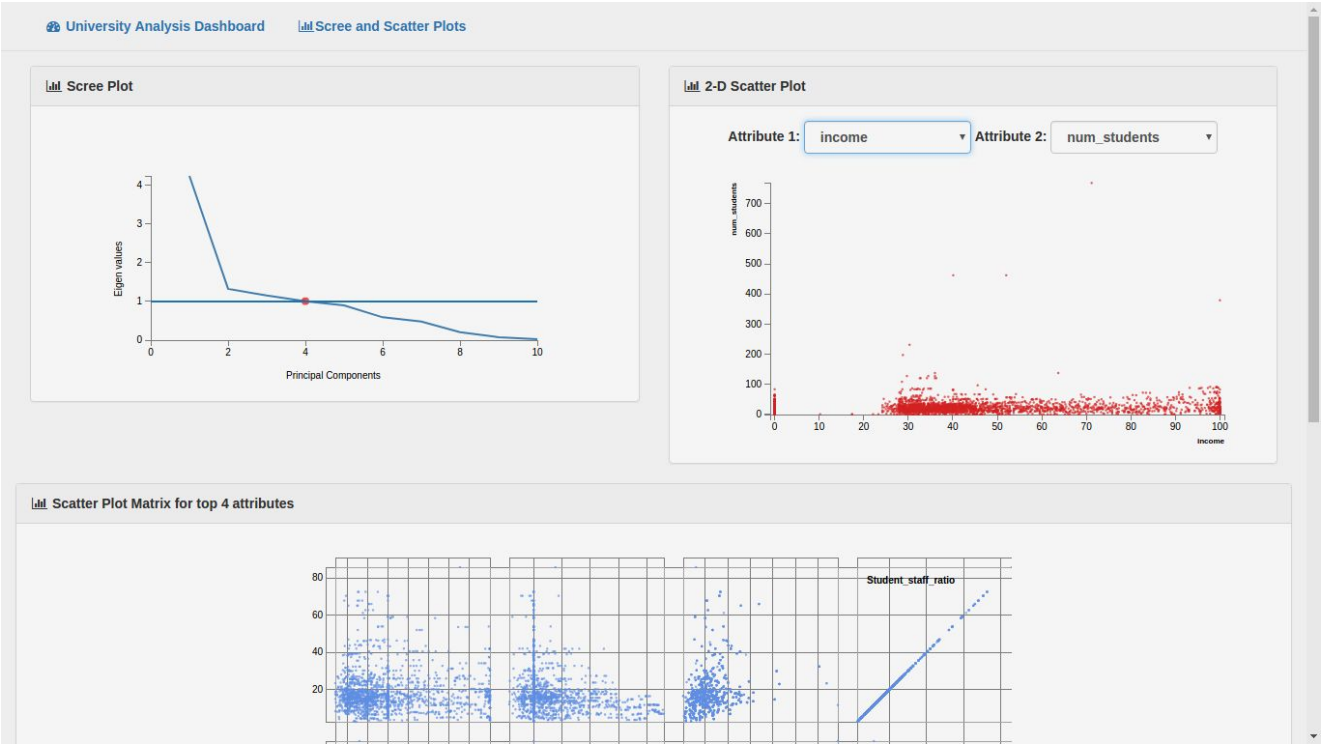

Fig.10 Scree Plot and 2D Scatter plot

The next step was to identify the most important attributes that influence the ranking a particular university. For this, we obtained the top 4 attributes with highest PCA loadings. The top 4 attributes were **student_staff_ratio, num_students, total_score, income**. This also makes sense as we could categorize a university as lower ranked or higher ranked based on these 4 metrics. For example, top ranked universities would have **a low value of student_staff_ratio** (students can have more 1-1 interactions with the staff), **a low size of num_students** (top universities are more selective and admit a very few only), **a high total_score** (evaluation of infrastructural facilities like lab, etc.) and **a moderate income** (we observed that income is high only when there are a lot of students. Since they have a lesser number of students, the top ranked universities comparatively have a moderate income that's generally obtained from research grants, fundings, etc). All of these analysis and findings can also be verified from the parallel coordinate map shown in the forthcoming sections. We plotted the scatter plot matrix for the top 4 attributes as shown in **Fig.11**.

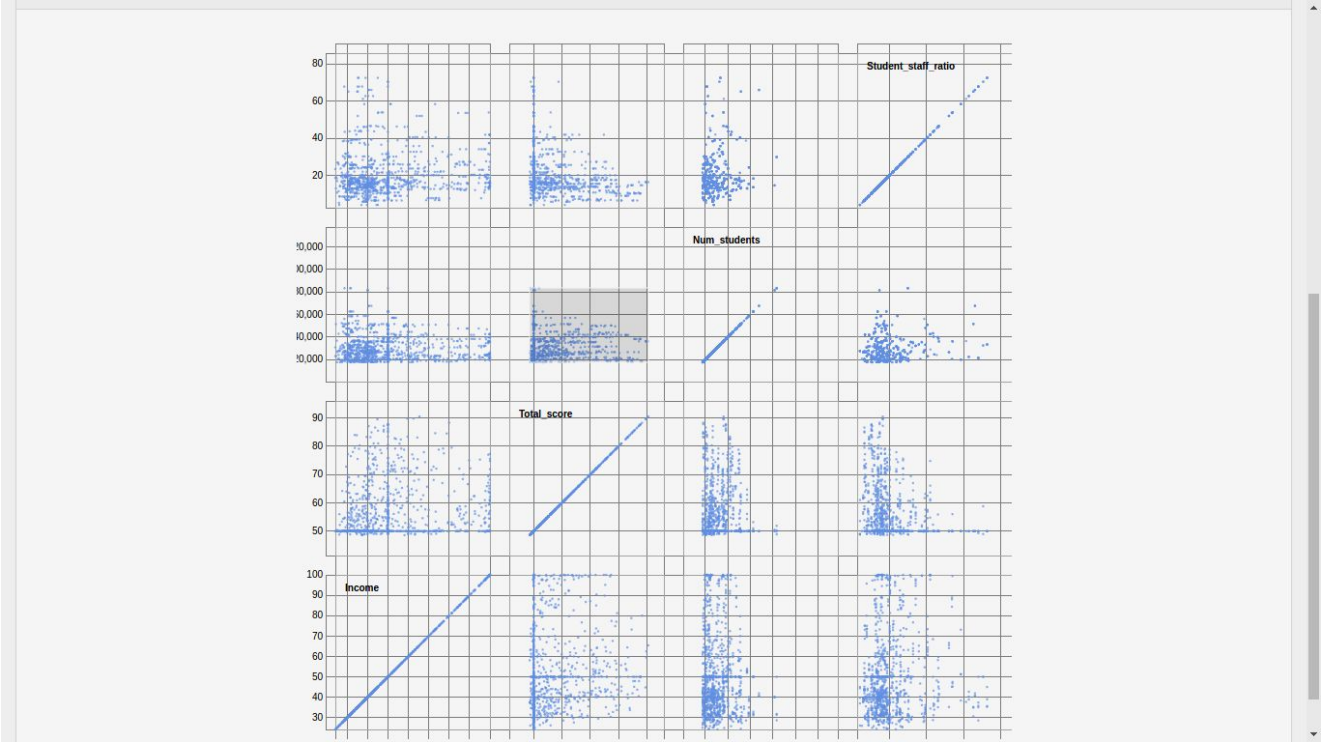

Fig.11 Scatterplot matrix

For the top 4 attributes, we identified their individual correlations using a 2D Scatterplot as shown in **Fig.12.**

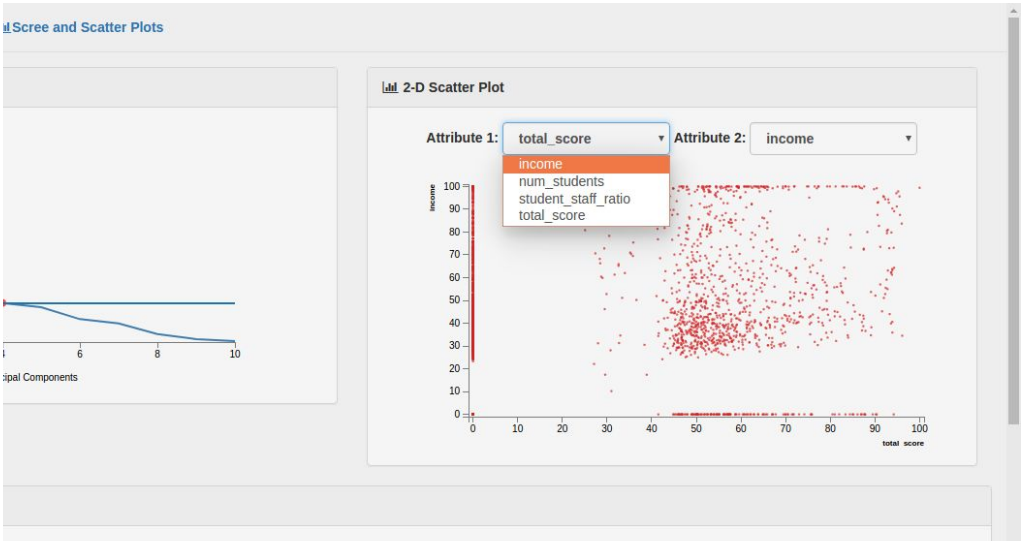

Fig.11 Individual 2D Scatter Plots of top 4 attributes

We saw a few interesting observations from the 2D scatterplot. **A large number (50% of universities) had a income score of around 30-50**. A majority of the **universities which had a total_score around 50 had less income score as well**.

Conclusion

Identifying the best university is often an arduous task. Ranking systems help us to an extent by evaluating all these universities quantitatively. Our project has extensively analysed these ranking systems and has given the user multiple visual depictions of the ranking system. It helps one to filter out the most appropriate university based on his/her priorities and see all the relevant comparison statistics. Through our visualizations, one can also understand the most important evaluation metrics used by these respective ranking systems. We have analysed the universities at a global level, at a national level and at an individual level. Our project would definitely provide a helping hand to any student deciding on what universities he/she should apply to.

Acknowledgements:

We thank all our reviewers for their valuable insights and feedback which has enabled us to develop a useful application.

Demo Video: https://youtu.be/lGrcJqyNeWo