

PROCESS BOOK

DATA SCIENCE – GRAD SCHOOL FINDER

Pranavnath Dommata

u1068846

u1068846@utah.edu

Neha Pathapati

u1069211

u1069211@utah.edu

Project Repository: <https://github.com/u1068846/VisProject.git>

BACKGROUND AND MOTIVATION:

The United States of America is a home to a large number of schools that are well acclaimed for their Data Science programs. Hence, for an individual intending to pursue a graduate degree in Data Science in the US, it may be overwhelming to zero down upon a list of schools that best suit his requirements, which he can apply to. Hence, we propose to design an interactive visualization tool that assists an individual in easing up this process.

PROJECT OBJECTIVES:

With the gigantic increase in the volume of data, there seems to be a huge demand in the industry for professionals who are trained to create value out of this data. Generally, such professionals are designated as – Data Scientists, Data Analysts, Data Engineers, and so on.

To cater to the rising need in the industry for data professionals who can mine enormous datasets to extract insights, several universities have begun to introduce Data Science programs into their curricula at the graduate level. Ultimately, they intend to produce graduates who can contribute to stemming the shortage of data scientists in the industry.

For an individual wanting to pursue a Data Science degree in the United States, there are undoubtedly a broad spectrum of options. The “Data Science – Grad School Finder” is a tool that allows one to shortlist a set of schools offering Data Science degrees based on his/her set preferences for parameters such as – type of the program (Masters’, Doctoral, or Certificate), delivery mode of the program, world rank of the program, pre-requisites needed for the program, and the state of the United States where the school is located. Additionally, the user can explore the shortlisted schools in more detail, draw up two schools to compare them closely, and sort the shortlisted schools based on his/her preferred parameter.

To reiterate, our objective is to build a comprehensive and an interactive visualization tool that assists potential data-science grads to shortlist, explore, and compare and contrast schools offering Data Science degrees in the US of A.

PROBLEM STATEMENT:

The “Data Science – Grad School Finder” tool finds an optimized set of schools in the US offering Data Science graduate degrees based on the user’s preferences set for – the level of the program (Masters’/Doctoral), the delivery mode, the state of US where the school is located, and pre-requisites

for the program. From the list of schools, the user can explore details like – the location of the school (state, city), the several Data Science programs offered by the school, the departments within the school offering these programs, the world rank of the program, a link to the program’s website, and so on.

Additionally, from the obtained list of shortlisted schools, a user can choose any two schools to compare them more closely. Lastly, the tool also enables the user to sort the shortlisted schools based on his preferred parameter, for instance, the world rank, faculty-student ratio, research citation score, and so on.

DATASET:

An existing data set of 951 entries will be employed for this project. The source of the dataset is stated in [1].

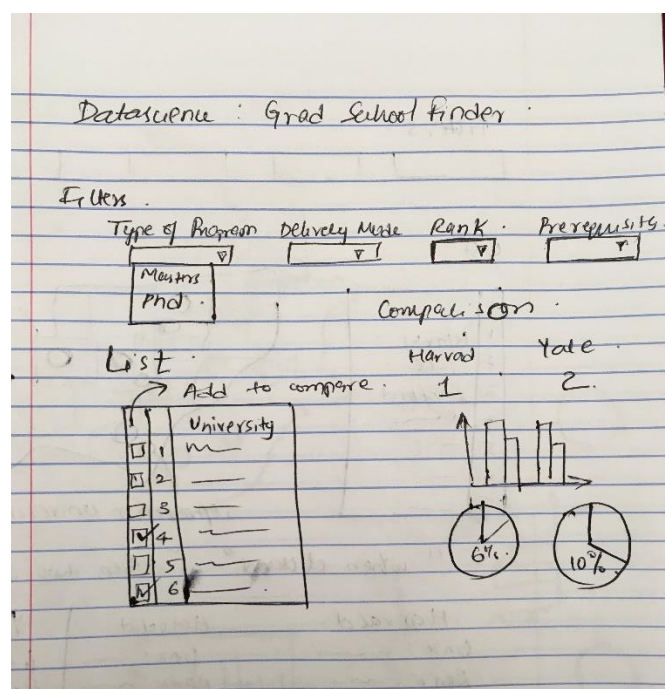
The dataset comprises of a comprehensive set of attributes that are sufficient for all components of our project. The attributes of the dataset include – School, State, City, Program, Type, Department, Delivery, Duration, Prerequisites, Link to the program website, Longitude and Latitude of the school, World Rank, Teaching Score, International Score, Research Score, Total Score for the school, Student Population Size, Student Faculty Ratio, International Student Population, and Male to Female Ratio.

DATA PROCESSING:

The dataset obtained from [1] was in CSV format. Hence, we are using our data in the raw form and did not find the need to subject it to any form of processing.

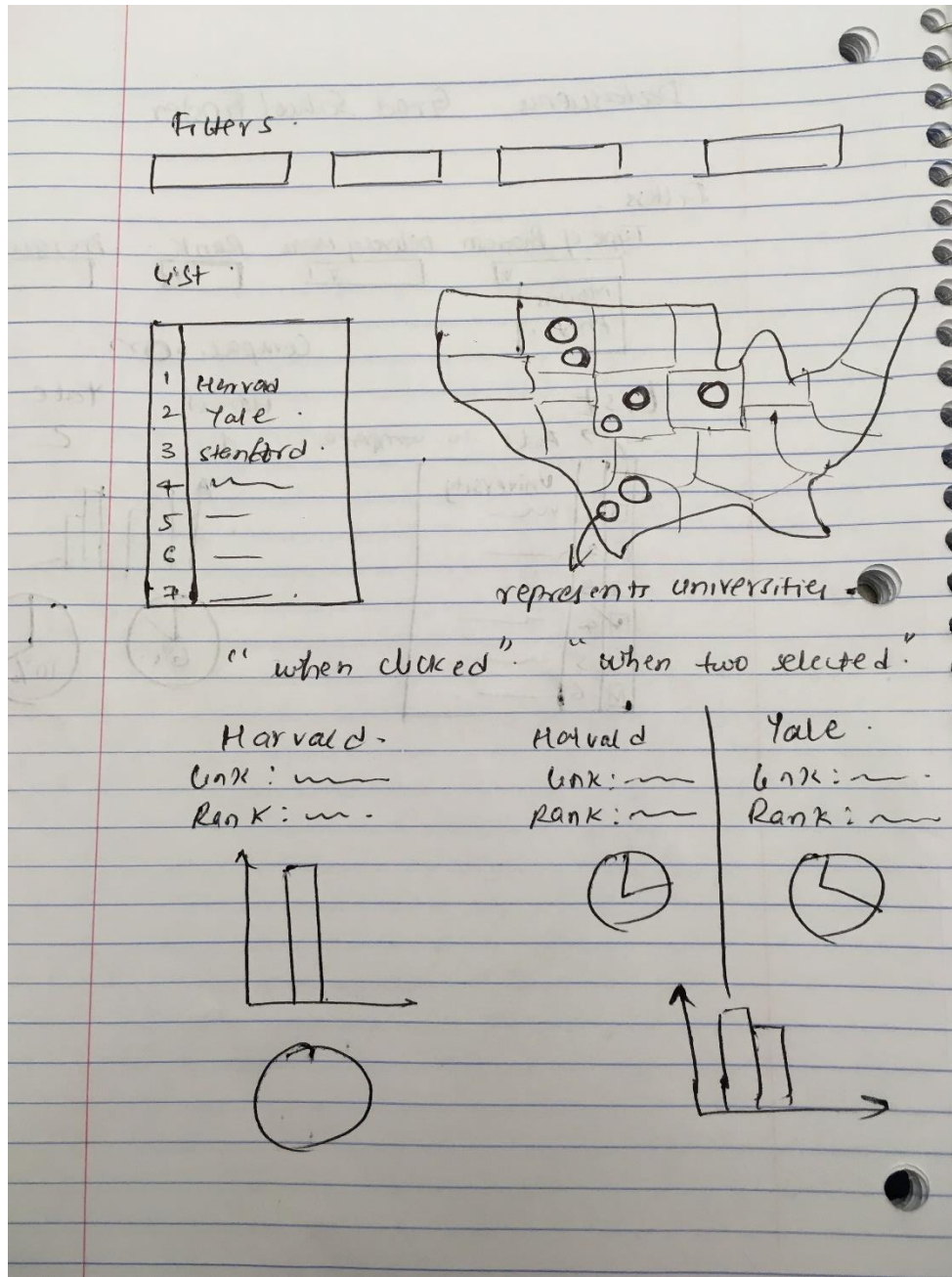
VISUALIZATION DESIGN:

DESIGN #1:



This is our first design based on our original ideas. Here, our visualization had three components – First, filters to allow the user to set his preferences for the type of Data Science program he wishes to pursue, the list of schools derived from the dataset based on the filter values, and the comparison of two schools selected by the user in the list using bar charts, donut charts, tiles, scatterplots, and so on.

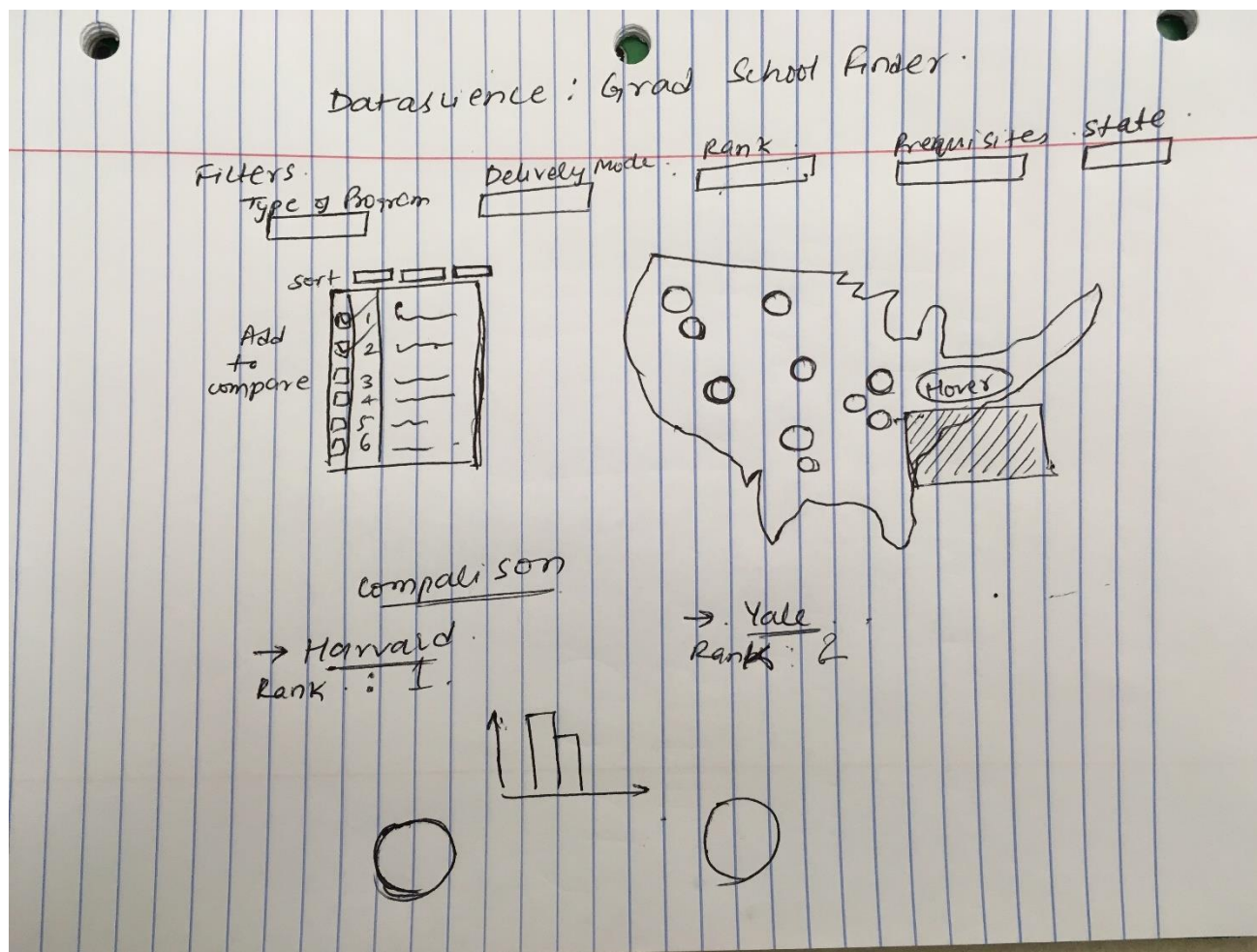
DESIGN #2:



In our second design, we made the following amendments: First, we decided to add a map of the US and highlight the location of the schools shortlisted by the tool to give the user an idea of the location of the school. This decision was primarily influenced by the presence of the latitude and longitude attributes for the schools in our dataset. Second, when the user clicks upon a highlighted school on the map, additional details of the school such as – the location, the degree names, the departments within the school offering these degrees, the world rank, a link to the program's website,

etc. are displayed below the map. Third, when the user selects two schools on the map, a comparison of the two schools is provided using bar charts, tiles, scatterplots, donut charts, and so on.

DESIGN #3:



This is our best design till date. Here, we added functionalities such as – allowing user to hover over a school on the map and get details of it in a tooltip, and sort the schools shortlisted by the tool based on his preferred parameter like for instance, the world rank, the student-faculty ratio, and so on.

MUST HAVE FEATURES:

1. Filters provided for the user to define more precise dataset he/she wants to look into.
2. Display a map of the US with interactive features like hovering over the university (represented with circles) to explore more about the school and clicking adds the university to compare with the university the user clicks next.
3. Selection of the criteria upon which the user wants to compare the two selected universities (multiple selections possible).
4. Visualizing the parameters using different designs like using grouped bar charts, pie-charts for representing percentages, donut charts, etc. to encode the data in a visibly effective manner.
5. The user is given an option to sort the list of universities displayed based on the parameter he chooses.

OPTIONAL FEATURES:

1. Enabling the comparison for more than two university selections.
2. Being able to add to compare from the list displayed instead of selecting from the map.
3. Zoom feature in the map which makes it more fun to interact with.

PROJECT SCHEDULE:

<u>Week</u>	<u>Date</u>	<u>Deadline</u>	
		Neha	Pranav
1	Nov 4	Filters	Comparison
2	Nov 11	Map	University List
3	Nov 18	Map Interaction	Sort
4	Nov 28	Zoom	Comparison designs

OVERVIEW OF PROJECT DELIVERABLES:

An interactive visualization that has the following functionalities:

1. Allows the user to set his preferences for – the graduate degree level (Master's/Doctoral), mode of delivery, the location of the school, and the prerequisites for the program.
2. Obtains a set of schools based on the user's preferences and highlights the schools location on the map of the US of A.
3. Upon hovering over a school on the map, allows the user to explore the school to a greater detail by displaying - the location of the school (state, city), the world rank of each of the programs, the student population size, the student faculty ratio, the research citation score, and so on.
4. Select two schools on the map and compare them closely.
5. Sort the schools shortlisted by the tool based on the user's preferred parameter like – the world rank, research citation score, and so on.

RELATED WORK:

Our work is inspired by the reference provided by our advisor – Yogesh Mishra [2]. This reference is a website that allows the user to explore and compare the schools of New York City.

PEER FEEDBACK:

A peer feedback session was conducted on November 3rd. We received valuable feedback from another team during this session. This session also facilitated us in improving the design of our project and motivated us to include more effective visualizations. We were critiqued by Sunipa Dev and group. Their critique was fair and helpful.

Listed below are some of the concerns the peer feedback session instigated us to reconsider.

What are you doing that doesn't already exist on the Internet?

We are providing a tool that enables an individual to explore and compare schools in the US to pursue a Data Science degree based on his preferences. The preferences he can vary include - the level of the graduate degree he wants to pursue, the World Rank of the program, and the presence/absence of prerequisites for that program. We employ several visual encodings that further assist an individual to make an informed choice of the schools he wishes to gain admission into. For instance, (a). the parallel axes allows the individual to filter out schools based on the range of scores awarded to the school for – teaching, research, international, citation, and income, (b). the map of US of A encodes the location of the school and allows the user to easily select a school to explore and select greater than one school to compare closely, (c). the comparator chart employs several encoding mechanisms to compare the selected schools based on several parameters.

Most of the websites focus on providing statistics of a particular school, and rarely furnish a wholesome comparison of schools, specifically for the Data Science degrees. Hence, we provide a tool for an individual intending to pursue a Data Science degree in the US to customize his choices to a greater extent, which is rarity to find on the internet.

When there are a large number of schools generated by the filters in a state, how will you handle the clutter?

We intend to use a zoom technique, wherein a user can zoom into a cluttered region of the map for better visibility.

Are you allowing the user to select the prerequisites explicitly rather than just allowing him to choose programs with or without prerequisites?

During the latter part of the project, we do intend to incorporate the selection of specific prerequisites like – Mathematics, Statistics, and a Bachelor's degree in Computer Science and so on.

When the user selects two schools to compare, in which order are you displaying them in the Comparator chart?

We will sort them by world rank, and then compare them in an effective manner.

Would you consider the use of tooltips and popovers for the map?

Yes, the map will have tooltips and popovers for all the schools on it. Additionally, we also intend to provide a link to the official website of the school for the Data Science degree.

Will you be using animations while adding schools to the comparator?

Yes, we intend to implement a drag-and-drop animation to drag schools from the map and drop them into the comparator for comparison. As we are unsure of the feasibility of implementing this functionality, we would like to add this to our list of additional features.

When the filters result in a large number of schools, how do plan to display the list of schools?

We intend to display the schools in a well-defined box and use scrolling, if that is the case.

DATA COLLECTION AND PRE-PROCESSING:

The dataset we retrieved from [1] came in the form of a CSV file with 951 entries. The entries were organized primarily by the school name in alphabetical order.

Each entry comprised of the following data – school name, the state, the city, the name of the Data Science program, the type of the program (Masters’ or Certificate), the department of the school offering the program, the delivery mode of the program (on-campus or online), the duration, the prerequisites, a link to the official program website, the latitude and longitude on a map of the US, the world rank, several scores for teaching, research, international, citations, income, a cumulative score for the school, the number of students, the student-faculty ratio, ratio of international students, the female to male ratio, and the year. Additionally, some schools had multiple entries corresponding to different years. Each of such entries differed only with respect to the several scores assigned to the school.

Though this dataset seems wholesome, it required a considerable amount of cleaning up as some critical fields were missing for several school entries.

There were several schools that did not have entries for critical fields like world rank. Since, the world rank is one of the filtering parameter, we were forced to exclude schools that did not have a world rank. Hence, a large number of schools were eliminated from our dataset due to this reason.

DESIGN CHANGES:

1. Earlier, we had decided to use the state of the US within which the school is located as a filter. But, since several entries in the CSV dataset had to be excluded due to absence of world rank, further filtering the schools by state would drastically reduce the number of schools.
2. We decided to implement parallel axes to enable the user to explore schools having his specified range of teaching score, research score, international score, and citation score.
3. To prevent cluttering, we decided to place the map below the list and the parallel axes.
4. For increased flexibility, we decided that the user must be allowed to select schools to compare both from the map and the list of schools.

FILTER DESIGN:

The filters enable the user to narrow down what school they want to look at by having them choose their preferences. The filters we have chosen are parameters that someone would most commonly look at when trying to explore schools. These include:

1. The type of the graduate program (Master’s or Doctoral or Certificate).
2. The range of world ranks of schools the user wants to explore.
3. Whether the program has any prerequisites.

A drop-down is used for all these selection parameters. Once the user has set his preferred values for these parameters, then he can hit the “Find Schools!” button to get the list of schools.

A screen shot of our filter implementation is shown below. Please note that this is not our final version, and we plan to improve it substantially.

DATA SCIENCE - GRAD SCHOOL FINDER

Find the best grad school to get your Data Science degree in!

Filters:

Grad Program Type: World Rank of Program: Prerequisites: [Find Schools!](#)

LIST DESIGN:

We implemented a simple list that displays the schools shortlisted by the tool based on the parameters set by the user for the filters.

The list displays the school name, the department within the school offering the Data Science degree, the duration of the program, the type of the program, and a checkbox. The checkbox allows the user to add schools to the comparator. Additionally, the shortlisted schools can also be sorted based on the school name, the department, and the duration of the program.

We intend to place all the schools listed in a well-defined box and enable scrolling within it. Also, we intend to add the glyph feature to the list.

A screen shot of our list implementation is shown below. Please note that this is not our final version, and we plan to improve it substantially.

University List

University Name	Department	Duration	Type	Check Box
Carnegie Mellon University	Heinz College	144 units	M	<input type="checkbox"/>
Cornell University	Department of Statistical Science	2 semesters	M	<input type="checkbox"/>
Duke University	Fuqua School of Business	10 Months	M	<input type="checkbox"/>
Harvard University	School of Public Health	80 credits/18 months	M	<input type="checkbox"/>
Massachusetts Institute of Technology	Sloan School of Management	1 Year	M	<input type="checkbox"/>
New York University	Computer Science Department	39 units	M	<input type="checkbox"/>
Stanford University	School of Medicine: Department of Biomedical Informatics	45 hours required	M	<input type="checkbox"/>
University of Illinois at Urbana-Champaign	Department of Statistics	30 credit hours	M	<input type="checkbox"/>
Washington University in St Louis	Olin Business School	39 Credits	M	<input type="checkbox"/>

PARALLEL AXES DESIGN:

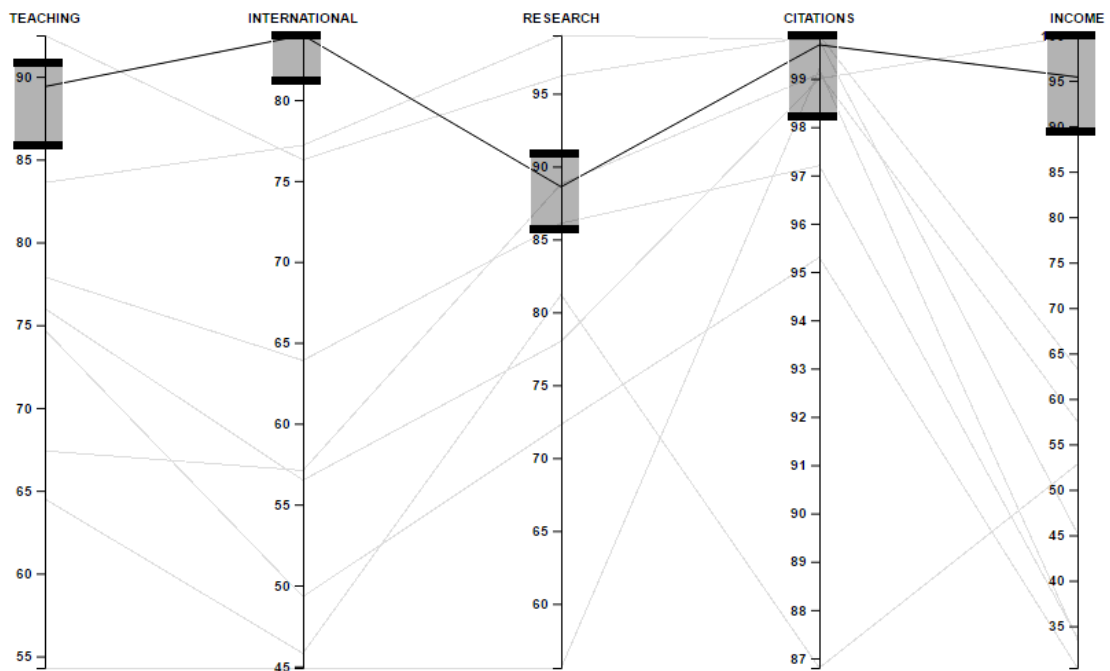
As suggested by our project advisor to provide the user more filters, we have implemented a parallel axes design which enables the user to filter on the universities further on some of the critical attributes by just brushing over the desirable values. We have implemented this design to give the user power to customize the filters effectively.

The basic implementation of the parallel axes is done and shown in the figure below. We are still yet to decide on some the interactive features that we want to implement over this design.

Presently, the brushing over the axes is enabled and also the drag option is provided.

A screen shot of our parallel axes implementation is shown below.

Parallel Axes



MAP DESIGN:

The map provides a locational reference for the schools. This way the user can explore schools by their location, targeting schools that are closer to their location, or at locations that they wish to explore. This was done by placing the schools as circles on the map as shown below.

Several interactive features were also added to the map to facilitate better exploration and comparison. They include:

1. On hovering over the schools on the map, the name of the school will appear on the tooltip.
2. On clicking upon a school, detailed information will be displayed for the school in the Comparator chart below the map.
3. To select two schools to compare, the checkbox “Select two schools to compare” has to be checked, and then the two schools to be compared have to be selected on the map.
4. A reset button that clears all selections on the map.

We are aware of the fact that maps are to be implemented only when absolutely necessary. In this case, the graduate Data Science programs in the US attract a large number of international students. Keeping this in mind, we decided to include this feature of providing an insight of the location.

A screen shot of our list implementation is shown below. Please note that this is not our final version, and we plan to improve it substantially.

Map Locator

Click on a school to explore!

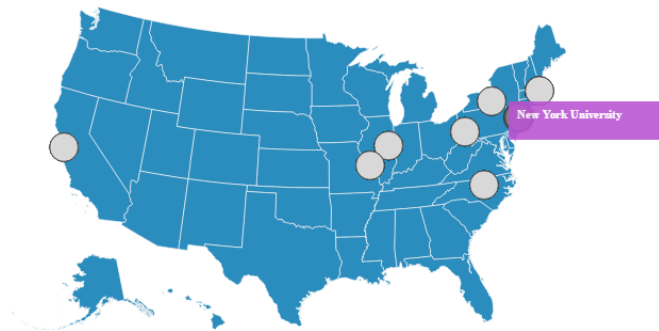
Select two schools to compare.

Reset

SCHOOL SELECTED TO EXPLORE

UNSELECTED SCHOOLS

SCHOOLS SELECTED TO COMPARE



COMPARATOR DESIGN:

After providing the user a wide range of filters to narrow down his choices by a significant level, we feel providing a one-one comparison between any two universities would be a great feature for the user to weigh them based on some key aspects. On the bigger picture, we intend to do a side by side comparison of the universities that the user selects.

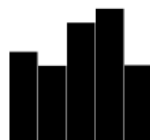
We intend to implement many interactive features and also make use of linked interactivity to throw light on the differences in an effective way and help users in finalizing their schools.

Below is the basic implementation of the comparator. We would like to put in majority of our efforts on making this section of the visualization as effective as possible.

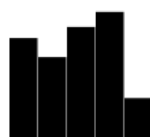
We haven't worked on the layout yet. Once it is done the below comparisons would be displayed side by side.

Comparator

- Carnegie Mellon University



- Cornell University



REFERENCE:

- [1] Srihari Rao. [Online]. Available: <https://www.kaggle.com/sriharirao/datascience-universities-across-us>
- [2] [Online]. Available: <http://itisaasta.com/nycs/nycs.html>