I. Introduction/Business Problem

This project is a way of comparing colleges. The user can perform comparison and clustering based on details about the universities themselves (including tuition/fees, acceptance rate, student body size, etc.) as well as attractions in the cities in which they are located. This tool is designed to help students make choices about which university to attend.

I specifically worked with the eight Ivy League schools (Brown University, Columbia University, Cornell University, Dartmouth College, Harvard University, Princeton University, University of Pennsylvania, and Yale University), however the code could be easily modified to compare other universities/cities.

The analysis that follows is based on the wants/needs of a hypothetical student, and it is meant to be modified to meet the needs of any student using the tool. **The hypothetical student in this particular analysis wants to attend an elite university and enjoys going to museums, attending live performances, and exploring nature.**

II. Data

University Data

For this project, I obtained data on the universities themselves data from these two URLs:
http://blog.collegetuitioncompare.com/2015/05/ivy-league-2015-2016-estimated-tuition.html
https://www.collegetuitioncompare.com/best-schools/ivy-league/
From these two sources, I used webscraping to obtain information about 2020 tuition/fees and on-campus housing costs, as well as the acceptance rates, SAT scores, graduation rate, and student/faculty ratio. All of these pieces of information are useful to a prospective student trying to choose a university.

Foursquare Data

I used Foursquare location data to explore the attractions/amenities nearby each of the universities. This is helpful for students looking to narrow their choices based on the type of city they'd like to live in: Is it important to them to have lots of restaurants nearby? Or museums? Or parks? Or music venues? To start, simply found the latitude/longitude location of each university and put them on a map. Later, I used these locations to find nearby venues.

Here is all of the data I used to compare and cluster the universities and search for nearby venues:

Name	Acceptance Rate	SAT Score	Graduation Rate	Population	Student-Faculty Ratio	City	Undergraduate Tuition & Fees	Graduate Tuition & Fees	Cost of Living (On-Campus)	Latitude	Longitude
Brown University	0.08	1485	0.95	10257	0.1667	Providence, RI	58404	58180	17454	41.826148	-71.404674
Columbia University in the City of New York	0.06	1505	0.96	31077	0.1667	New York, NY	61788	49968	16670	40.807619	-73.962367
Cornell University	0.11	1465	0.95	23600	0.1111	Ithaca, NY	57222	29585	18066	42.447505	-76.483603
Dartmouth College	0.09	1490	0.95	6572	0.1429	Hanover, NH	57638	55947	18414	43.703747	-72.289352
Harvard University	0.05	1520	0.98	31566	0.1429	Cambridge, MA	51925	49214	20875	42.373711	-71.117264
Princeton University	0.05	1505	0.96	8374	0.2	Princeton, NJ	52800	53770	20300	40.348158	-74.659310
University of Pennsylvania	0.08	1485	0.95	25860	0.1667	Philadelphia, PA	57770	40182	18136	39.952056	-75.195065
Yale University	0.06	1515	0.97	13433	0.1667	New Haven, CT	55500	43300	20095	41.311007	-72.926121

III. Methodology

Data Collection and Cleaning

In order to obtain the data above, I did substantial formatting/cleaning of the data I scraped from the above URLs. I converted all data (except the city names) to floats, and got rid of extra characters such as commas, dollars signs, and percent signs. I obtained the latitude and longitude values for each university by querying Foursquare for venues matching each university name and using those locations.

Initial Cluster Analysis

The next step was to narrow the student's choices by performing an initial cluster analysis of the attributes listed in the above section. I used k-cluster analysis and looked for four clusters. Two or three clusters would also have been fairly effective to narrow the selection down into larger groups, but I chose to make four clusters so the clusters would each be small.

The universities were clustered based in similarities in acceptance rate, SAT scores, graduation rate, student/faculty ratio, population size, and cost. Based on these clusters, the student could choose which type of university they're most interested in attending (easier to get into or more elite, smaller or larger student body, lower tuition, lower cost to live on-campus, etc.). Next, the student can create a "shortlist" consisting of as many schools as they would like to examine more closely.

Within-Cluster Analysis

Once the hypothetical student has selected the cluster(s) they're interested in investigating further, they can begin to explore the surrounding area to determine if a particular location is more or less appealing than the others. I kept the exploration radius to 1 mile (reasonable walking-distance from campus).

Comparing Venue Categories

The student can do a search that returns all types of venues, but I found it more productive to search by venue type. I included searches on food venues, arts/entertainment venues, and outdoor/recreation venues. I found this method to allow for better comparisons between the cities; when searching through all venues, the vast majority of the top venues were restaurants, which made it difficult to learn more about the other types of venues.

Exploring Most Popular Venues

Within each venue type, the student can view the 'n' most popular types of venues within that category. I found 10 to be a good number, as numbers greater than that can become confusing. Additionally, if you set 'n' too high and a city doesn't have at least 'n' unique venue types in that category, the city may not actually have any of the nth most popular venue type.

Comparing Specific Venues Between Cities

If a student has a particular interest--such as the performing arts, physical fitness, outdoor activities, a specific cuisine, etc.--they can compare those venue types directly. I performed such a comparison on different kinds of museums, performing arts venues, and outdoor/nature venues. Because we are now comparing only a few venue types in a limited number of cities, the clearest way to view the results is with a bar chart comparing the relative quantities of each venue type in each city.

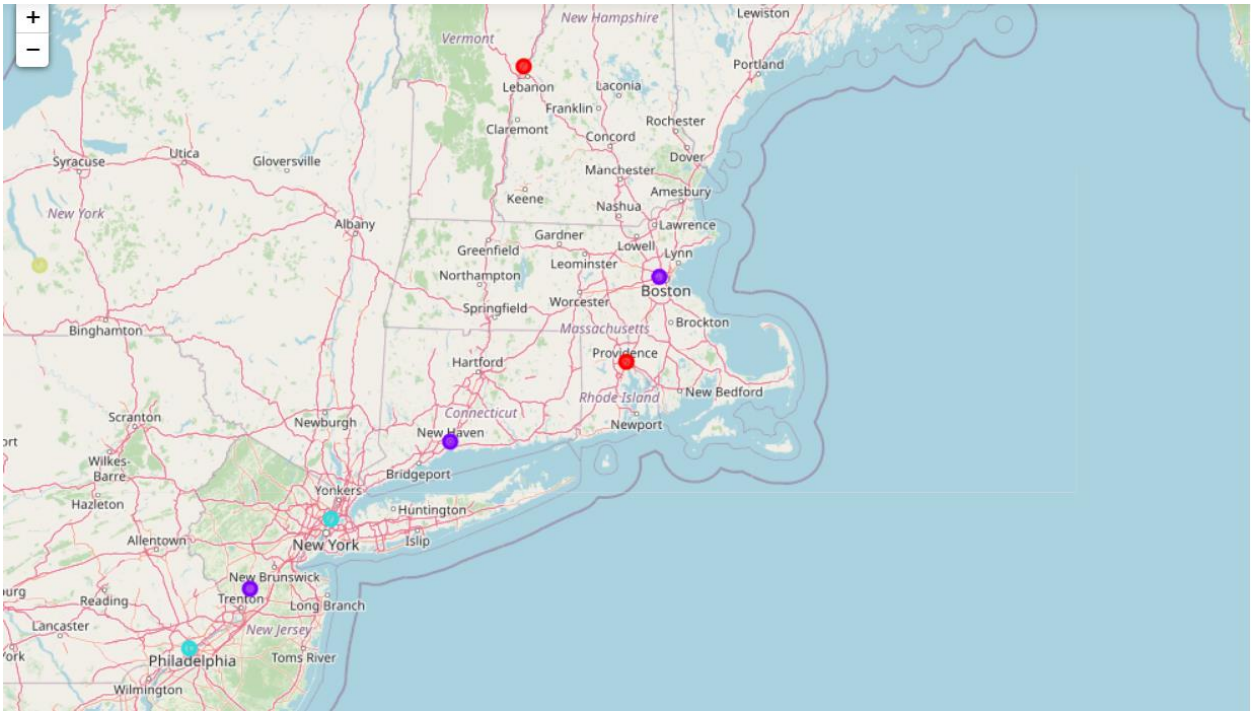
IV. Results

Initial Cluster Analysis

In this analysis, group 0 is Dartmouth and Brown, group 1 is Harvard, Yale, and Princeton, group 2 is Columbia and UPenn, and group 3 is Cornell. Here are the mean values for the four clusters:

Labels	Acceptance Rate	SAT Score	Graduation Rate	Population	Student-Faculty Ratio	Undergraduate Tuition & Fees	Graduate Tuition & Fees	Cost of Living (On-Campus)
0	0.085000	1487.500000	0.950	8414.5	0.154800	58021.000000	57063.500000	17934.000000
1	0.053333	1513.333333	0.970	17791.0	0.169867	53408.333333	48761.333333	20423.333333
2	0.070000	1495.000000	0.955	28468.5	0.166700	59779.000000	45075.000000	17403.000000
3	0.110000	1465.000000	0.950	23600.0	0.111100	57222.000000	29585.000000	18066.000000

Here is a map showing the four clusters:



Here is a summary of the noteworthy attributes of each cluster:

Harvard, Yale, and Princeton	Columbia and UPenn	Brown and Dartmouth	Cornell
lowest acceptance rate	second lowest acceptance rate	second highest acceptance rate	highest acceptance rate
highest SAT scores	second highest SAT scores	second lowest SAT scores	lowest SAT scores
mixed student population (Harvard is large, Yale and Princeton are small)	largest student population	smallest student population	large student population
lowest tuition for both undergraduate and graduate students	highest undergrad tuition, medium graduate tuition	high tuition for graduate and undergraduate students	lowest graduate tuition, fairly high undergraduate tuition
highest cost to live on-campus	low cost to live on-campus	low cost to live on-campus	fairly low cost to live on-campus worst faculty-student ratio
SUMMARY: Most elite and difficult to get accepted at	SUMMARY: Somewhat difficult to get accepted, good if you like large student bodies	SUMMARY: Easier to get into, good if you like very small student bodies	SUMMARY: Easiest to get accepted, good for students who need less personal attention from faculty, good financial option for graduate students

Within-Cluster Analysis

Based on the preferences of the hypothetical student, I assumed that the student is most interested in Harvard, Yale, and Princeton. The next step was to use the Foursquare data to explore the areas surrounding each school in the cities of Cambridge, New Haven, and Princeton to find out what types of venues are nearby.

Comparing Venue Categories

Beginning with the survey of different venue categories within a mile of campus, I found the following:

- Harvard has 100+ food venues, 57 arts/entertainment venues, and 100+ outdoor/recreation venues
- Princeton has 72 food venues, 20 arts/entertainment venues, and 33 outdoor/recreation venues

- Yale has 100+ food venues, 38 arts/entertainment venues, and 43 outdoor/recreation venues

Based on the initial analysis of these three categories, it appears that Princeton has the fewest venues overall, and Harvard has the most. You might also conclude from this data that Princeton is the smallest/least dense of the three cities, while Cambridge is the largest/most dense.

Exploring Most Popular Venues

I then took a closer look at the most common types of venues in each category in the three cities. Here are the 10 most common types of food venues near each university:

	University	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Harvard University	American Restaurant	Pizza Place	Bakery	Italian Restaurant	Café	Japanese Restaurant	Indian Restaurant	New American Restaurant	Burger Joint	Seafood Restaurant
1	Princeton University	Pizza Place	Mexican Restaurant	Sushi Restaurant	Sandwich Place	Chinese Restaurant	Italian Restaurant	Bakery	American Restaurant	New American Restaurant	Tapas Restaurant
2	Yale University	Pizza Place	American Restaurant	Italian Restaurant	Mexican Restaurant	Café	Deli / Bodega	Indian Restaurant	Thai Restaurant	Food Truck	Sandwich Place

Here are the 10 most common types of arts/entertainment venues near each university:

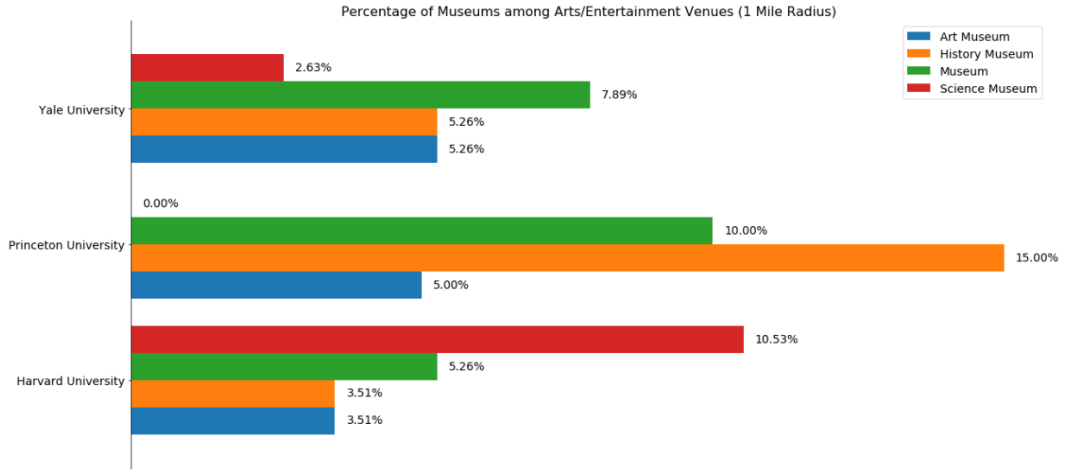
	University	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Harvard University	Theater	Music Venue	Dance Studio	Art Gallery	Science Museum	Outdoor Sculpture	Museum	Comedy Club	Jazz Club	History Museum
1	Princeton University	Theater	History Museum	Museum	Concert Hall	Dance Studio	Art Gallery	Movie Theater	Indie Movie Theater	Outdoor Sculpture	Art Museum
2	Yale University	Theater	Music Venue	Concert Hall	Art Gallery	Museum	Rock Club	Movie Theater	History Museum	Dance Studio	Art Museum

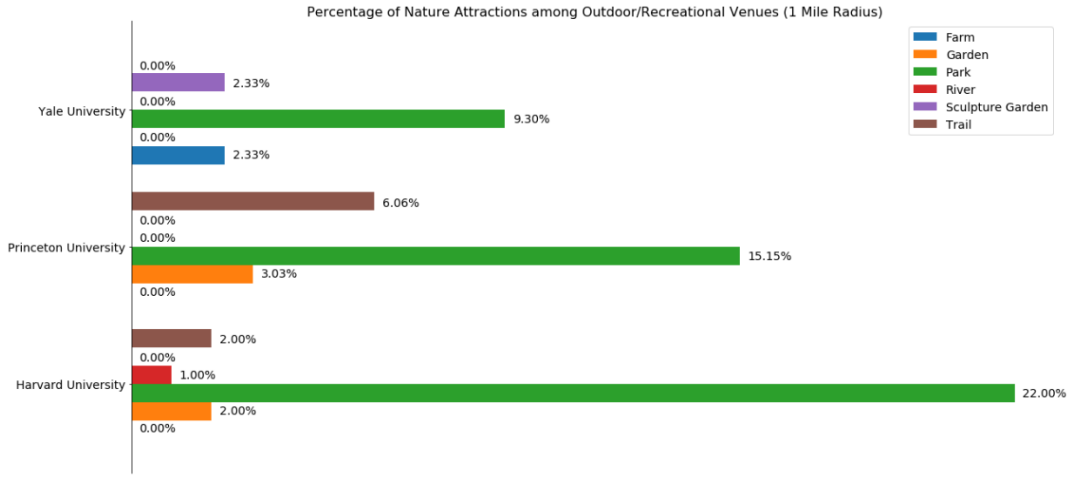
Here are the 10 most common types of outdoor/recreation venues near each university:

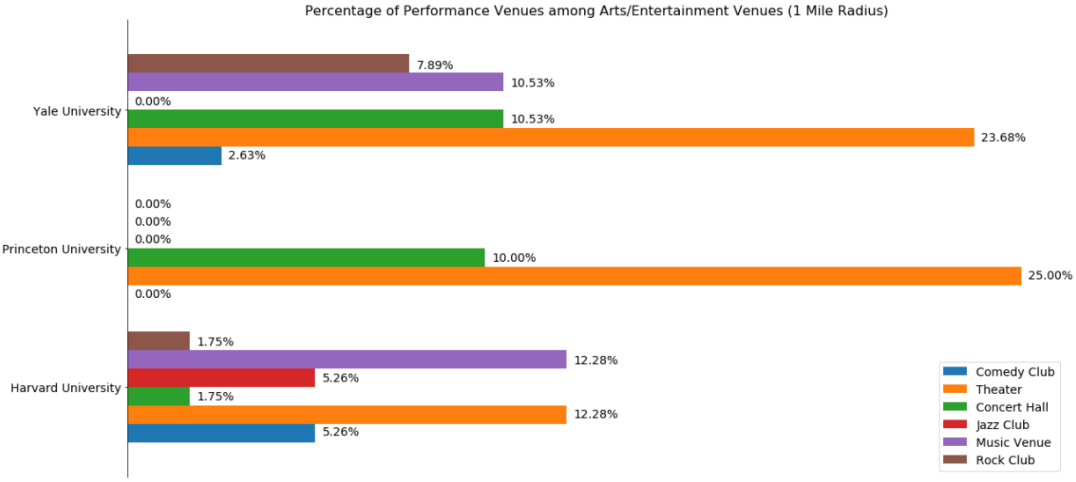
	University	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Harvard University	Park	Plaza	Gym	Yoga Studio	Gym / Fitness Center	Playground	Baseball Field	Athletics & Sports	Harbor / Marina	Skating Rink
1	Princeton University	Plaza	Yoga Studio	Park	Gym / Fitness Center	Athletics & Sports	Pool	Trail	Pedestrian Plaza	Baseball Field	Garden
2	Yale University	Gym	Plaza	Gym / Fitness Center	Yoga Studio	Park	Baseball Field	Squash Court	Cycle Studio	Sports Club	Sculpture Garden

Comparing Specific Venues Between Cities

Next, I looked specifically at the types of venues the hypothetical student is most interested in (museums, live performance venues, and outdoor/nature venues). Note that the percentages for the charts below reflect the number of venues of a particular type (e.g. an art museum) divided by the total number of venues in that category of venue (e.g. arts/entertainment venues) that were returned by our query within a 1 mile radius of each campus







Based on the above comparisons, a student could make more informed decisions about which area they might enjoy most. For example, while Princeton has the highest percentage of museums among its arts/entertainment venues, our query only returned 20 total arts/entertainment venues within a mile

radius of Princeton (compared with 38 arts/entertainment venues near Yale, and 57 near Harvard). Thus, if a student places high value on having lots of museums to explore during their free time, the best choice would be Harvard, where 22.81% of the 57 arts/entertainment venues are museums (Yale has a similarly high percentage--21.04%, but fewer overall venues).

Looking at live performance venues, 55.26% of Yale's 38 arts/entertainment venues are places where a student might see live music/entertainment. This seems much higher than Harvard's 38.58%, but taking into account the greater number of venues in Harvard the number of live performance venues is actually quite similar (22 in Harvard, 21 in Yale).

In terms of outdoor venues, Harvard once again returned the largest number of total outdoor/recreational venues (100 in total, as opposed to 43 near Yale and 33 near Princeton). Harvard also has the greatest number of unique types of outdoor venues (trails, rivers, parks, and gardens) and the highest percentage of nature attractions.

V. Discussion

All of the above analysis was based on the wants/needs of a hypothetical student. This student wants to attend an elite university and enjoys going to museums, attending live performances, and exploring nature. **According to this analysis, this hypothetical student should attend Harvard University**; out of the "elite" cluster (Harvard, Princeton, and Yale) Harvard has the lowest acceptance rate, highest SAT scores, and highest graduation rate (it the most elite) and it also has the most museums, performance venues, and outdoor/nature attractions.

For a student with different values, the analysis could be modified to meet their needs. For example, a student with slightly lower SAT scores might be more interested in looking at clusters C (Brown and Dartmouth) and D (Cornell). That student might also be less interested in the arts and more interested in living somewhere with diverse cuisine. Another student might be a grad student with children looking to pay lower tuition and live close to lots of parks/playgrounds. All of these needs can be accommodated by changing the specific parameters of the within-cluster analysis section.

VI. Conclusion

This is designed as a tool to help prospective students choose a university. It is currently set up only to compare the eight Ivy League universities, however the same data can just as easily be obtained on other groups of universities. This tool is set up to provide a lot of freedom to the user; at each stage, the prospective student can make decisions based on their personal preferences. This is obviously incomplete data and is not meant to represent every factor a student might take into account when making a decision to attend a college; it is only intended to provide comparison and visualization of various features of multiple schools in order to influence a decision.