Applied Data Science Capstone

Which Boston College/University to start my career as a Data Scientist?

Tyler Lareau

Introduction

With the new skills I have developed as a data scientist, I want to apply to for a new role in an educational community to begin my career in data science. I picture myself working for a college/university. and I really want to work in my hometown of Boston. My problem is which college or university in the Greater Boston do I want to work for? To assist in my selection, I am going to leverage the power of public datasets and location data to select a school which I think will best fit my new career as a data scientist. With over 50 colleges/universities in the city of Boston, the criteria I am going to use to select a school to apply to work for are

-School located in the Greater Boston Area

-School must have over 5000 students. As a data scientist, I am going to want to intake as much data as possible and believe that 5000 is a solid minimum to retrieve valuable and diverse data.

-Sports Complexes/Arenas must be popular venues within the 500M radius of campus. Sports is one of the industries with high demands for data scientists and want to be within walking distance to all sports events on campus.

-School must have access to public transportation in the immediate radius (500M)

The target audience for this project would be for any prospective student or employee interested in applying to and learning more about the population size and surrounding neighborhood for schools located in Boston with over 5000 students.

Data Acquisition and Cleaning

For this project I will be leveraging public datasets from Analyze Boston (https://data.boston.gov/) to retrieve the names, coordinates and number of students enrolled in all college/universities. There are over 159 datasets available and I will be utilizing the College and Universities CSV.

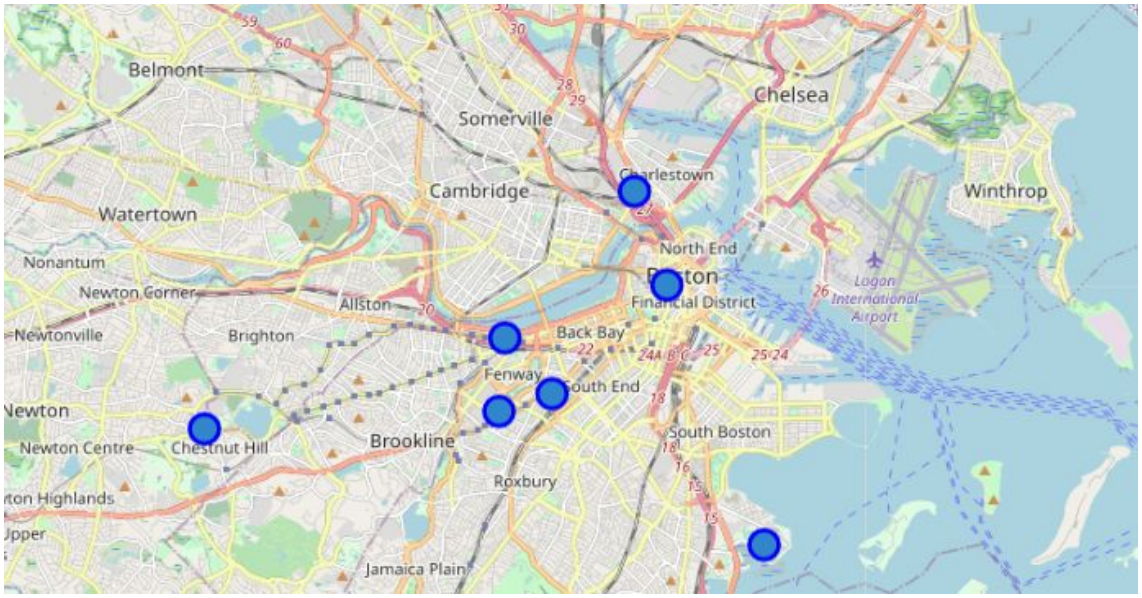
This dataset includes 60 College/Universities in the Boston City Limits and has over 30 columns including data points such as the school name, address, school Id, location coordinates, number of students (years 2012 and 2013), website and even the year the college was built. To better visualize the list I created a pandas dataframe from the CSV and only kept the School Name, Latitude, Longitude and Number of Students in 2013 as that was the most current year available. Leveraging the Foursquare API, the location coordinates will be used to retrieve frequently checked-in venues in each college's surrounding neighborhood.

Methodology

The first step in making the decision on which school in Boston to work for was to 53manipulate the pandas dataframe created from the Boston Colleges and Universities CSV to only show schools that have a student body over 5000. This brought the number of schools down from 60 to 7, an 88% decrease of schools listed. With the shorter list it will be much easier to compare the neighborhoods around each school.

	Name	Latitude	Longitude	NumStudents13
0	Boston University	42.349560	-71.099709	32411
1	University of Massachusetts-Boston	42.313809	-71.039202	16277
2	Boston College	42.333833	-71.169719	14309
3	Bunker Hill Community College	42.375117	-71.069572	14023
4	Suffolk University	42.358905	-71.061948	8675
5	Northeastern University	42.340048	-71.088892	8479
6	MCPHS University	42.336880	-71.101120	6548

Next, using the coordinates I visualized the locations of all the schools by marking them on a map of Boston.

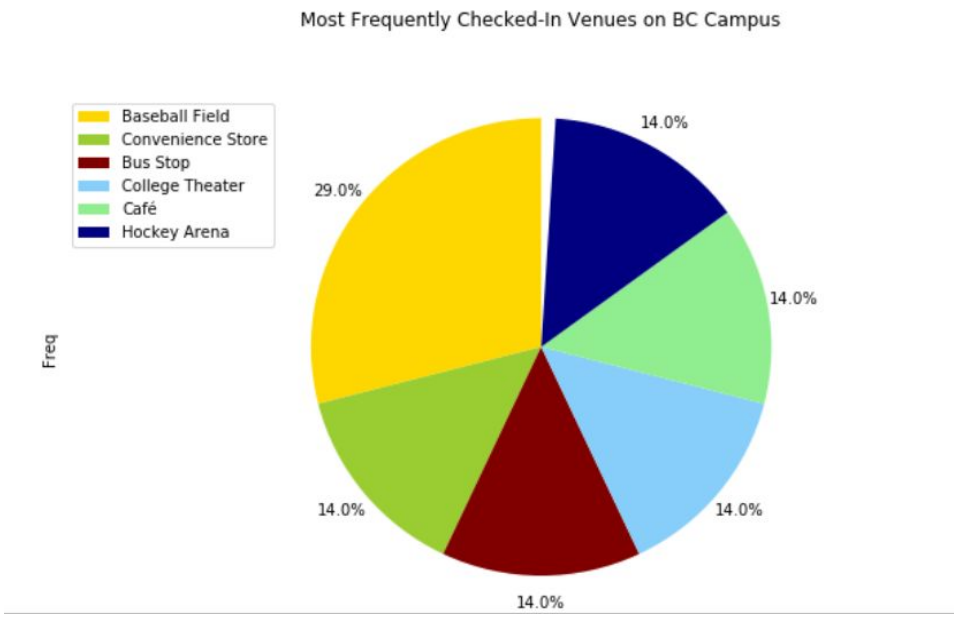


To learn more about each school's surrounding neighborhood I needed to leverage the Foursquare API and location coordinates. I created a function using the latitude, longitude coordinates for each of the seven remaining schools to retrieve the top 10 most frequent visited venues in a 500 meter radius from the Foursquare API. This would give me insight of what schools have surrounding neighborhoods which meet my criteria to want to live and work there. This required one-hot encoding to clean the JSON results that were retrieved from the Foursquare API as well as an additional function to add the venues and schools into a pandas dataframe. As these were frequently visited venues near schools I am interested in applying to so I listed them as "Popular Venues".

	School	1st Popular Venue	2nd Popular Venue	3rd Popular Venue	4th Popular Venue	5th Popular Venue	6th Popular Venue	7th Popular Venue	8th Popular Venue	9th Popular Venue	10th Popular Venue
0	Boston College	Baseball Field	College Theater	Convenience Store	Hockey Arena	Café	Bus Stop	Food Court	Cosmetics Shop	Deli / Bodega	Department Store
1	Boston University	American Restaurant	Lounge	Sports Bar	Coffee Shop	Hotel	Sushi Restaurant	Café	Pub	Gym / Fitness Center	Brewery
2	Bunker Hill Community College	Coffee Shop	Yoga Studio	Shopping Mall	Bank	Convenience Store	Donut Shop	Gastropub	Grocery Store	Light Rail Station	Liquor Store
3	MCPHS University	Coffee Shop	Sushi Restaurant	Sandwich Place	Convenience Store	Pizza Place	Pub	Donut Shop	Café	Falafel Restaurant	Gastropub
4	Northeastern University	Sandwich Place	Pizza Place	Grocery Store	Arts & Crafts Store	Middle Eastern Restaurant	Concert Hall	Café	Caribbean Restaurant	Burrito Place	Restaurant
5	Suffolk University	Coffee Shop	Historic Site	Seafood Restaurant	Mediterranean Restaurant	New American Restaurant	Hotel	Restaurant	American Restaurant	Market	Italian Restaurant
6	University of Massachusetts-Boston	Museum	Coffee Shop	Donut Shop	Fast Food Restaurant	Food Court	Concert Hall	Convenience Store	Cosmetics Shop	Deli / Bodega	Department Store

Results

Only one college has more than one venue that meets the listed criteria of having public transit as well as popular sports complexes on campus, Boston College. Boston College's most popular venue is their Baseball Field with 29% of check-ins (graphic below) and also haS Hockey Arena and Bus stop within the top 6 most checked in venues. Two schools have access to public transport: Bunker Hill Community College(light rail station) and MCPHS(Bus Station). The other 4 schools do not have any top 10 frequently visited venues that meet the criteria.



Discussion

As the only college to have more than 1 venue that meets the listed criteria for selection, Boston College will be my choice school to start as a Data Scientist. With frequently visited sports complexes (hockey arena/baseball field) there should be many opportunities to leverage data in athletics as well as become a fan. It was also interesting to see that there are over 50 schools in Boston with less than 5000 students which may be useful for projects in the future where smaller sample sizes are needed.

Conclusion

In this study, I analyzed the surrounding neighborhoods of seven colleges located in Boston, MA. I used the new skills I have learned over the past few months to solve a problem leveraging data. I used the location coordinates to learn what venues are frequently visited in effort to select an area to want to live and work. This type of analysis could be duplicated in any major city or town to make a location data based decision.