Massachusetts Cities Clustering

Raul Alejandro Gonzalez Gallo
School of Engineering
Universidad Anahuac
Queretaro, Mexico
raulalejandro273@gmail.com

April, 7th, 2020

I. INTRODUCTION

Choosing which college or university to attend is one of the biggest decisions that you will make in terms of your personal and professional development, and it can be overwhelming if you don't know where to start. College is important for many reasons, including long-term financial gain, job stability, career satisfaction and success outside of the workplace. In essence, college teaches us how to learn and grow in every aspect of our lives. As Ralph Waldo Emerson says, "The things taught in schools and colleges are not an education, but the means to an education." Of course, the quality of the education offered is a very important factor in deciding which college to attend, as well as the overall on-campus experience. However, another important factor is the off-campus experience offered, including the historical, cultural and social background of the college's city since at the end attending college is preparing us to become active members of our societies.

Most of the time prospective college students can't visit the city of the college they are planning to attend because of time or economic limitations and are sometimes insecure if they are going to like the environment offered by that city. Also, many students choose to attend a certain college just because they really like the city in which it is located. It seems that knowing in advance the different types of environments offered by the cities would facilitate prospective college students the important task of choosing which college or university to attend. Maybe a student would prefer to attend a college in a city with a lot of Italian restaurants in the surroundings or maybe a city with many museums and parks to visit during weekends.

With its incredibly significant, undeniably unique place in American history and culture, from the earliest historical period of colonial America onward, Massachusetts continues to play a primary contributing roll to American high-culture and fine-arts. Massachusetts is home to countless world-class museums and national historical sites and has produced some of Americans most famously creative academics, artists, writers, and musicians. Massachusetts' role in American education is also without equal. Massachusetts is home to the United States' oldest high school, the first public library, oldest boarding school, oldest college, and the first women's college. Additionally, top level universities such as Harvard and MIT, which consistently rank among the world's best universities year after year, are located in this state. Massachusetts has 12% of the top research universities and 15% of the top 40 liberal arts colleges. Several of the world's best medical and technology facilities are located here as well as numerous multinational corporations. In summary, the state of Massachusetts seems like a great place to decide to accomplish college studies. This project aims to cluster of find the different environments offered by the cities of Massachusetts that houses the state´s colleges and universities, seeking to facilitate prospective college students, wanting to study there, the important task of choosing the right college.

II. DATA

The data that will be used for this project will be a List of colleges and universities in Massachusetts published in Wikipedia in the following link: https://en.wikipedia.org/wiki/List_of_colleges_and_universities_in_Massachusetts. First, I will extract all the data from Wikipedia using pandas. I will use the first table which comprises all the active colleges and universities in Massachusetts. Figure 1 shows the data without any preprocessing and cleaning.

	School	Location[note 1]	Control[1]	Type[1]	Enrollment[16]	Founded	Accreditation[16]
0	American International College	Springfield	Private not-for-profit	Master's university	2,177[17]	1885[17]	AOTA, APTA, CCNE, NEASC[17]
1	Amherst College	Amherst	Private not-for-profit	Baccalaureate college	1,817[18]	1821[18]	NEASC[18]
2	Anna Maria College	Paxton	Private not-for-profit	Master's university	1,455[19]	1946[19]	NASM, NEASC, NLNAC[19]
3	Assumption College	Worcester	Private not-for-profit	Master's university	2,813[20]	1904[20]	NEASC[20]
4	Babson College	Wellesley	Private not-for-profit	Special-focus institution	3,250[21]	1919[21]	NEASC[21]
5	Bard College at Simon's Rock	Great Barrington	Private not-for-profit	Baccalaureate/associate's college	354[22]	1964[22]	NEASC[22]
6	Bay Path University	Longmeadow	Private not-for-profit	Baccalaureate college	2,370[23]	1897[23]	AOTA, NEASC[23]
7	Bay State College	Boston	For-profit	Associate's college	1,721[24]	1946[24]	ABHES, APTA, NEASC, NLNAC[24]
8	Becker College	Worcester	Private not-for-profit	Baccalaureate college	1,826[25]	1784[25]	APTA, NEASC, NLNAC[25]
9	Benjamin Franklin Institute of Technology	Boston	Private not-for-profit	Special-focus institution	475[26]	1908[26]	NEASC[26]

Figure 1. List of active colleges and universities in Massachusetts

Starting with the data cleaning processes, I will first drop the Accreditation column, as well as remove all the values between brackets since they don't provide any useful information. Figure 2 shows the resulting table.

	School	Location	Control	Type	Enrollment	Founded
0	American International College	Springfield	Private not-for-profit	Master's university	2177	1885
1	Amherst College	Amherst	Private not-for-profit	Baccalaureate college	1817	1821
2	Anna Maria College	Paxton	Private not-for-profit	Master's university	1455	1946
3	Assumption College	Worcester	Private not-for-profit	Master's university	2813	1904
4	Babson College	Wellesley	Private not-for-profit	Special-focus institution	3250	1919
5	Bard College at Simon's Rock	Great Barrington	Private not-for-profit	Baccalaureate/associate's college	354	1964
6	Bay Path University	Longmeadow	Private not-for-profit	Baccalaureate college	2370	1897
7	Bay State College	Boston	For-profit	Associate's college	1721	1946
8	Becker College	Worcester	Private not-for-profit	Baccalaureate college	1826	1784
9	Benjamin Franklin Institute of Technology	Boston	Private not-for-profit	Special-focus institution	475	1908

Figure 2. Table with the useful data

The data contains four categorical features and two numeric features. First, lets explore and gain some insight of the numeric features. Using the correct methods presented in the Jupyter Notebook we obtained the following insights:

- The oldest college in Massachusetts is Harvard University founded in 1636
- The newest college in Massachusetts is Olin College founded in 1997
- The college with the highest enrollment is Boston University with 32603 students
- The college with the lowest enrollment is Conway School of Landscape Design with 18 students
- The average enrollment for a college in Massachusetts is approximately 4780 students
- The total number of students enrolled in Massachusetts is 497170 students
- 1 college was founded in the 17th century
- 2 colleges were founded in the 18th century
- 40 colleges were founded in the 19th century
- 61 colleges were founded in the 20th century

Likewise, exploring the categorical features we obtained the following insights:

- We can see that there are 104 distinct colleges in the dataset, in a total of 53 different locations
- The location with most colleges is Boston with a total of 24 colleges
- The most common type of college are Special-focus institutions
- Most colleges are in Boston, Worcester and Cambridge
- 72 colleges are Private not-for-profit, 30 are Public and 2 are For-profit
- 28 colleges are Special-focus institutions, 22 are Baccalaureate colleges, 20 are Master's universities, 20 are Associate's colleges and 14 are Research universities

Next, I will prepare the data in a convenient way by finding the number of colleges for each city, as well as grouping the data by cities and separating the schools in each city with commas. Finally, I will just keep the columns of cities, number of colleges and colleges' names. Figure 3 shows the resulting table.

	City	Number of Colleges	Colleges
0	Amherst	3	Amherst College,Hampshire College,University o...
1	Andover	1	Massachusetts School of Law
2	Beverly	2	Endicott College,Montserrat College of Art
3	Boston	24	Bay State College,Benjamin Franklin Institute ...
4	Bourne	1	Massachusetts Maritime Academy
5	Bridgewater	1	Bridgewater State University
6	Brighton	1	Saint John's Seminary
7	Brockton	1	Massasoit Community College
8	Brookline	2	Boston Graduate School of Psychoanalysis,Helle...
9	Cambridge	6	Cambridge College,Harvard University,Hult Inte...

Figure 3. Table with the needed features for the implementation

III. METHODOLOGY

Now that the data preparation is done, I will start with the implementation. For the implementation, I will start by finding the location of each city using Geopy and start using the Foursquare API to get the closest venues for each city, as well as the 10 most common venues in each city. Finally, I will use k-means to cluster the different environments offered by the cities of Massachusetts that houses the state´s colleges and universities, seeking to facilitate prospective college students, wanting to study there, the important task of choosing the right college. Now I will break this in parts.

First, I will use the Nominatim function from Geopy to find the latitude and longitude of each city and combine it with the last table obtained. Figure 4 shows the resulting table.

	City	Number of Colleges	Colleges	Latitude	Longitude
0	Amherst	3	Amherst College,Hampshire College,University o...	42.368566	-72.505714
1	Andover	1	Massachusetts School of Law	42.657170	-71.140878
2	Beverly	2	Endicott College,Montserrat College of Art	42.558428	-70.880049
3	Boston	24	Bay State College,Benjamin Franklin Institute ...	42.360253	-71.058291
4	Bourne	1	Massachusetts Maritime Academy	41.741217	-70.598920
5	Bridgewater	1	Bridgewater State University	41.990379	-70.975043
6	Brighton	1	Saint John's Seminary	42.350097	-71.156442
7	Brockton	1	Massasoit Community College	42.083433	-71.018379
8	Brookline	2	Boston Graduate School of Psychoanalysis,Helle...	42.331764	-71.121163
9	Cambridge	6	Cambridge College,Harvard University,Hult Inte...	42.375100	-71.105616

Figure 4. Table with the locations of each city

Then, I will use the Folium library to map the state of Massachusetts and its cities with colleges, as shown in figure 5.

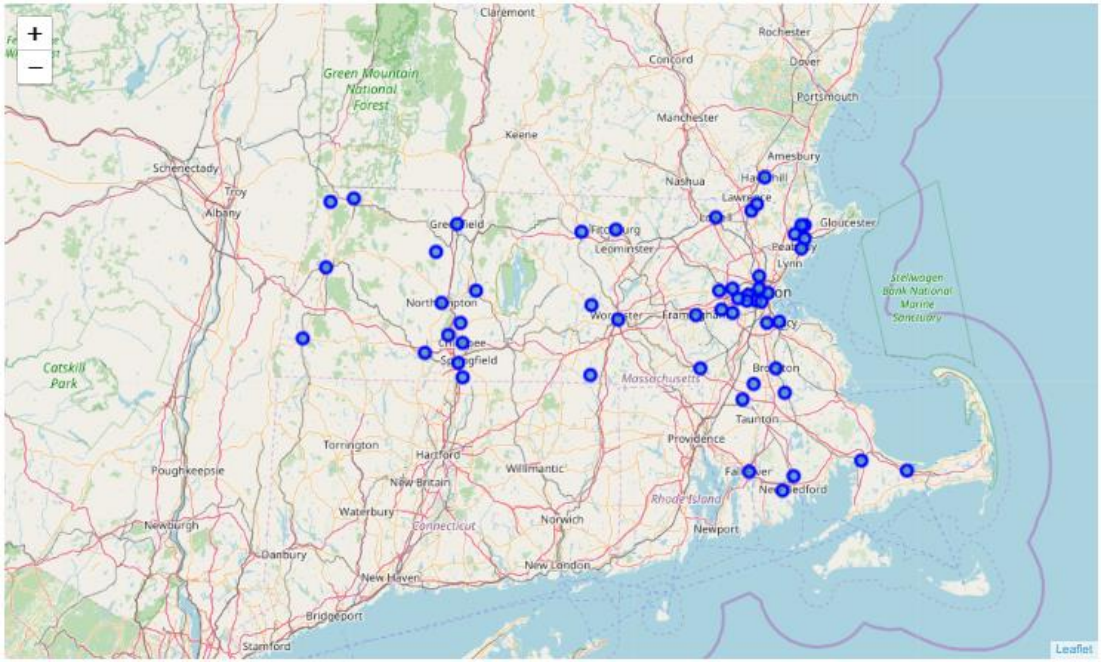

Figure 5. Massachusetts' map with the cities that have colleges and universities

Having the locations of each city I can start using the Foursquare API to get the closest venues for each city using a radius of 2km and a limit of 100 venues per city. I will then analyze every city by taking the mean of the frequency of occurrence of each category and keeping the top 10 venues for each city. Figure 6 shows the result of this process.

	City	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	Amherst	Coffee Shop	Grocery Store	Sandwich Place	Hotel	American Restaurant	Bakery	Pizza Place	Department Store	Liquor Store	Breakfast Spot
1	Andover	Coffee Shop	American Restaurant	Pizza Place	Italian Restaurant	Sandwich Place	Restaurant	Gym / Fitness Center	Fast Food Restaurant	Donut Shop	Burger Joint
2	Beverly	Coffee Shop	Italian Restaurant	Park	Pizza Place	Bakery	Ice Cream Shop	Pub	Sandwich Place	Indie Movie Theater	Brewery
3	Boston	Park	Bakery	Seafood Restaurant	Gym	Coffee Shop	Hotel	Historic Site	Pizza Place	Sandwich Place	New American Restaurant
4	Bourne	Seafood Restaurant	Donut Shop	Convenience Store	American Restaurant	Park	Sandwich Place	Restaurant	Beach	Gas Station	Breakfast Spot

Figure 6. Top 10 venues for each city in Massachusetts that has colleges and universities

Next, I can start using machine learning techniques to try and find relationships between cities. Since in this case we don't have labels, we need an unsupervised learning algorithm. I will run *k*-means to cluster the cities into 4 clusters using the mean of the frequency of occurrence of each category for the venues. Figure 7 shows the table with all the data used for the implementation as well as the cluster assigned and the top 10 venues.

	City	Number of Colleges	Colleges	Latitude	Longitude	Cluster Labels	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	Amherst	3	Amherst College,Hampshire College,University o...	42.368566	-72.505714	0	Coffee Shop	Grocery Store	Sandwich Place	Hotel	American Restaurant	Bakery	Pizza Place	Department Store	Liquor Store	Breakfast Spot
1	Andover	1	Massachusetts School of Law	42.657170	-71.140878	1	Coffee Shop	American Restaurant	Pizza Place	Italian Restaurant	Sandwich Place	Restaurant	Gym / Fitness Center	Fast Food Restaurant	Donut Shop	Burger Joint
2	Beverly	2	Endicott College,Montserrat College of Art	42.558428	-70.880049	0	Coffee Shop	Italian Restaurant	Park	Pizza Place	Bakery	Ice Cream Shop	Pub	Sandwich Place	Indie Movie Theater	Brewery
3	Boston	24	Bay State College,Benjamin Franklin Institute ...	42.360253	-71.058291	0	Park	Bakery	Seafood Restaurant	Gym	Coffee Shop	Hotel	Historic Site	Pizza Place	Sandwich Place	New American Restaurant
4	Bourne	1	Massachusetts Maritime Academy	41.741217	-70.598920	1	Seafood Restaurant	Donut Shop	Convenience Store	American Restaurant	Park	Sandwich Place	Restaurant	Beach	Gas Station	Breakfast Spot

Figure 7. Top 10 venues for each city in Massachusetts that has colleges and universities and corresponding cluster

Finally, now that the cities are grouped into clusters and the top venues are known for each one of them, we can visualize the different clusters and try to find similarities inside clusters and differences between different clusters in order to determine the different environments offered by the cities of Massachusetts that houses the state´s colleges and universities.

IV. RESULTS AND DISCUSSION

Figure 8 shows the different clusters in the cities of Massachusetts, were different colors represent different clusters.

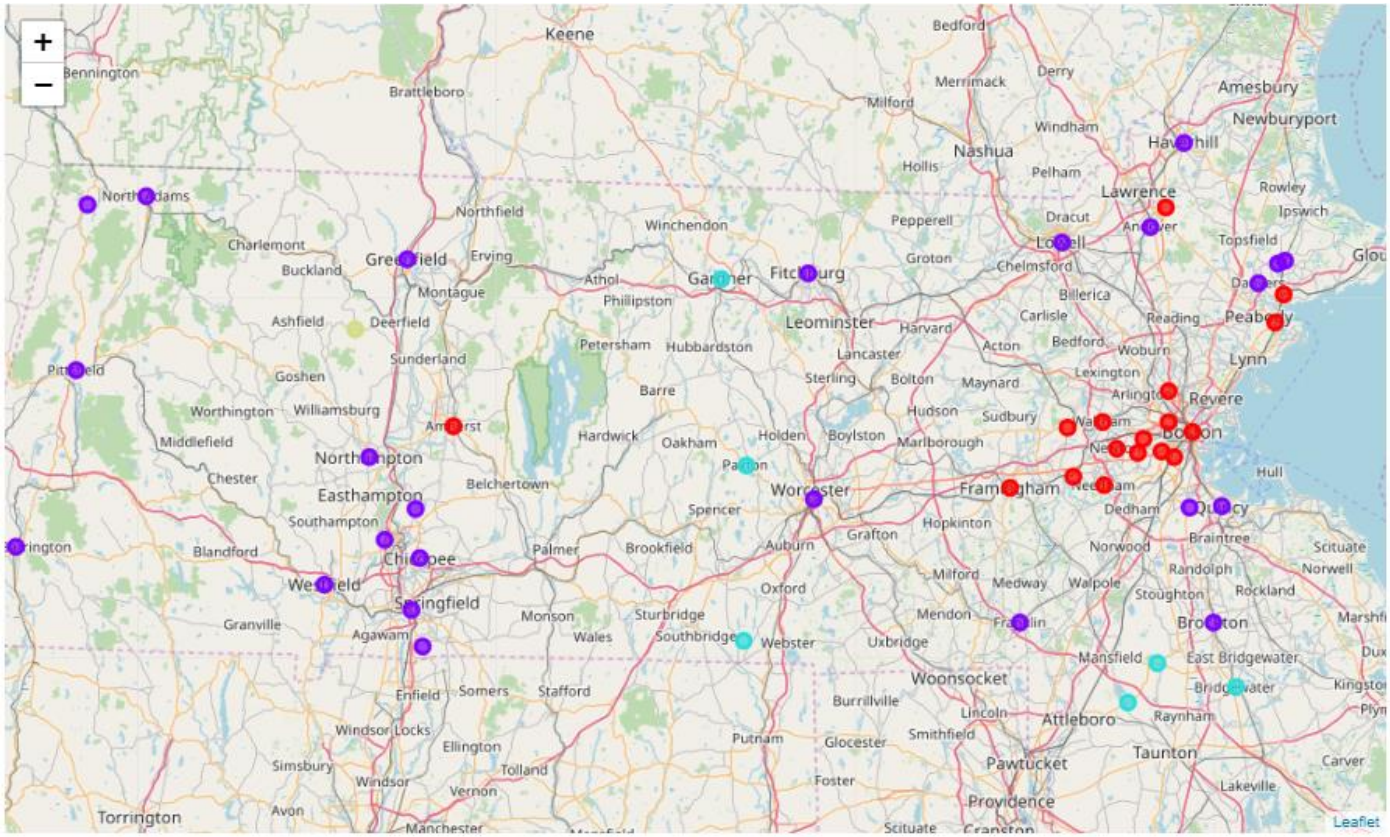

Figure 8. Massachusetts' map with the cities that have colleges and universities and their corresponding cluster

As we can observe, there are two big clusters (red and purple) and two small clusters, one of them containing only one city. In the Jupyter Notebook we can observe all the clusters and which cities belong to each one of them. Some cities in each cluster are the following:

- Cluster 0: Amherst, Boston, Brookline, Cambridge, Framingham, North Andover, amongst others
- Cluster 1: Andover, Danvers, Dartmouth, Greenfield, Northampton, Worcester, amongst others
- Cluster 2: Bridgewater, Dudley, Easton, Gardner, Norton and Paxton
- Cluster 3: Conway

It is very interesting that one of the clusters only has one city: Conway. If we analyze carefully the top common venues in Conway, we can see uncommon venues in other cities such as rivers, farms, flea markets, Construction and Landscaping.

The two clusters with most cities have in common venues of different types of restaurant, parks, coffee shops, historic sites, museums, hotels and others. As we can observe, this are mainly big cities near the capital of Massachusetts, Boston.

V. CONCLUSIONS

In this study, I analyzed the different environments offered by the cities of Massachusetts that houses the state´s colleges and universities based on the most common venues by cities by using one of the most common unsupervised learning machine learning algorithm, *k*-means. I used the pandas capabilities to scrape data from Wikipedia and manipulate the data in convenient ways. Also, I used the geopy and folium libraries to obtain locations and map them for useful visualization as well as the Foursquare API to explore common venues in each city. The clusters obtained help understand the relationships and characteristics between cities and what they have to offer to prospective college students, in order to help them choose the right college to attend.