Clustering Top US Colleges Based on Average Tuition and Nearby Attractions

Problem:

Over 1 million students enroll in college each year. For many, tuition is the deciding factor, and for others, location and social life is the highest priority – choosing between colleges is always a struggle. Many online sources aggregate data pertaining to tuition, campus life, location, rating, etc. to help students make this decision, but the insight provided is often at a broad level. In this project, we will use Python to explore the top 50 U.S. colleges and cluster them based on specific venues nearby.

Data:

We will extract names and average tuitions of the top 50 U.S. colleges from Business Insider. The Foursquare API will then provide us with data for nearby venues and their categories, which we will use for the clustering analysis.

	College	Average Tuition	Coordinates
0	Harvard University	16205	(42.36782045, -71.1266665287448)
1	Massachusetts Institute of Technology	21576	(42.3583961, -71.0956778766393)
2	Yale University	18319	(41.25713055, -72.9896696015223)
3	Columbia University	22973	(40.8071772, -73.9625279772072)
4	California Institute of Technology	26839	(34.13710185, -118.125274866116)
5	Stanford University	16695	(37.43131385, -122.169365354983)
6	Brown University	25264	(41.82687235, -71.4012277069681)
7	Duke University	19950	(36.0001557, -78.9442297219588)
8	Princeton University	17732	(40.34829285, -74.66308325)
9	University of Pennsylvania	22944	(39.9492344, -75.191989851901)
10	Cornell University	30014	(42.4505507, -76.4783512955428)
11	Dartmouth College	21177	(43.7047927, -72.2925909)
12	Northwestern University	29326	(42.0551164, -87.6758111348217)
13	University of Chicago	31068	(41.78468745, -87.6007493265106)
14	Rice University	22061	(29.71679145, -95.4047811339379)
15	Carnegie Mellon University	35250	(37.4102193, -122.059654865858)
16	University of Southern California	32932	(34.0224149, -118.286344073446)
17	Washington University in St Louis	28824	(38.64724015, -90.3084017323959)
18	Vanderbilt University	23150	(36.1442594, -86.8027428817193)
19	Emory University	24804	(33.7915703, -84.3183726165067)
20	Johns Hopkins University	27352	(39.2964392, -76.592394032674)
21	Amherst College	19055	(42.37289, -72.518814)
22	Williams College	18167	(42.7130236, -73.2030082)
23	Pomona College	18140	(34.0947694, -117.7146921)
24	University of California, Los Angeles	14236	(34.07088865, -118.446731966638)
25	University of Notre Dame	26683	(41.70456775, -86.2382202601727)
26	New York University	35147	(40.72925325, -73.9962539360963)
27	University of Michigan — Ann Arbor	16107	(42.2942142, -83.710038935096)
28	Wellesley College	20013	(42.29182055, -71.3033260683231)
29	Georgetown University	26625	(38.90893925, -77.0745796206083)
30	Swarthmore College	19641	(39.9035501, -75.354092055757)
31	Tufts University	28076	(42.40629165, -71.1197504981564)
32	University of California, Berkeley	17160	(37.87094645, -122.266398722925)
33	Claremont McKenna College	30527	(34.1023497, -117.7067162)
34	Carleton College	28587	(44.47183535, -93.1414580590152)
35	Boston University	31539	(42.35050035, -71.1025599049017)
36	Middlebury College	21437	(44.0090777, -73.1767946)
37	University of North Carolina at Chapel Hill	10077	(35.90503535, -79.0477532652511)
38	Case Western Reserve University	33124	(41.50138695, -81.6007021615902)
39	Haverford College	21144	(40.0071506, -75.3069423257631)
40	University of California, Davis	16039	(38.52247515, -121.751392674913)
41	Smith College	24258	(42.3148532, -72.6401382)
42	Purdue University West Lafayette	11693	(40.4275052, -86.9122769)
43	Bowdoin College	24888	(43.9075035, -69.9617742423256)
44	University of California, San Diego	14770	(32.87935255, -117.231100493855)
45	Wesleyan University	20490	(41.5551478, -72.6569115610163)
46	University of Miami	37424	(25.7172788, -80.2786915764625)
47	University of Illinois at Urbana-Champaign	16683	(40.101976, -88.2314378)
48	Lehigh University	27478	(40.6068028, -75.3782488)
49	Bryn Mawr College	31900	(40.02813555, -75.3159205851816)

After calling the Foursquare API to search for nearby venues, a new table is created consisting of the top 100 venues in a 2000m radius from the college and the venue's broad and specific category. Below is a sample for Harvard University.

	College	Average Tuition	College Coordinates	Venue	Venue Coordinates	Broad Category	Specific Category
0	Harvard University	16205	(42.36782045, -71.1266665287448)	Harvard Stadium	(42.366997, -71.12680128)	Arts & Entertainment	College Stadium
1	Harvard University	16205	(42.36782045, -71.1266665287448)	Trader Joe's	(42.3633439875157, -71.12994385534071)	Shops	Grocery Store
2	Harvard University	16205	(42.36782045, -71.1266665287448)	John F. Kennedy Memorial Park	(42.37080162572463, -71.12280545734018)	Parks & Outdoors	Park
3	Harvard University	16205	(42.36782045, -71.1266665287448)	Flour Bakery + Cafe	(42.3731171074856, -71.12234866100246)	Food	Bakery
4	Harvard University	16205	(42.36782045, -71.1266665287448)	Our Fathers Deli	(42.36352042227399, -71.12945843588452)	Nightlife	Bar
5	Harvard University	16205	(42.36782045, -71.1266665287448)	Bright Hockey Center	(42.36807488929957, -71.12699337955992)	Arts & Entertainment	College Hockey Rink
6	Harvard University	16205	(42.36782045, -71.1266665287448)	Orinoco	(42.37193332225955, -71.12061225278022)	Food	Arepa Restaurant

Exploratory Data Analysis:

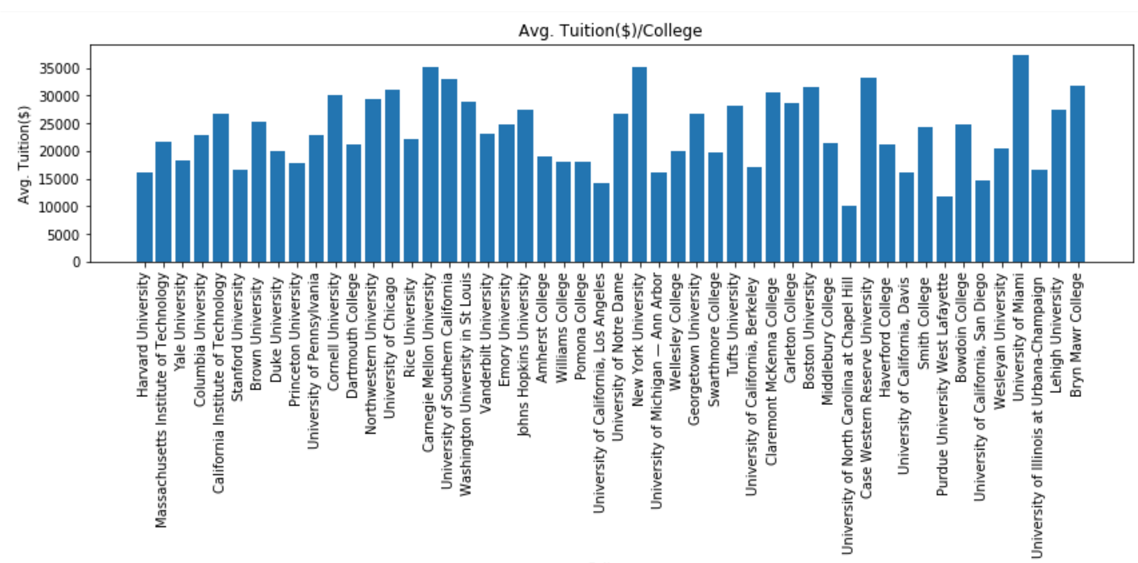
Presented to the right is a table displaying the descriptive statistics on the number of venues per college. We see that the minimum is 4 venues, belonging to Carleton College, which could be a potential outlier. The majority of colleges have 100 venues in their search radius.

	# of Venues
count	50.000000
mean	87.740000
std	22.552261
min	4.000000
25%	82.500000
50%	100.000000
75%	100.000000
max	100.000000

As shown on the table above, a broad and specific category is specified for each venue. We will explore both types of categories to obtain the most comprehensive analysis.

It is also apparent that there are 8 unique broad venue categories: Arts & Entertain ment, Shops, Parks & Outdoors, Food, Nightlife, Travel, Building, and Education.

Here is a bar chart displaying each college's average tuition. We see that Wesleyan U niversity, Carnegie Mellon University, and New York University have the highest tuit ion.



The pie charts below display the distribution of broad venue categories for each college. Refer to the Jupyter Notebook to see the distribution of specific categories.

The most common venue across all colleges appears to be Starbucks, followed by Dunkin' Donuts and Subway.

	Count
Starbucks	58
Dunkin'	22
SUBWAY	18
CVS pharmacy	17
sweetgreen	16
Chipotle Mexican Grill	15
Trader Joe's	14
Blaze Pizza	12
7-Eleven	12
Insomnia Cookies	11

<u>Cluster Modeling:</u>

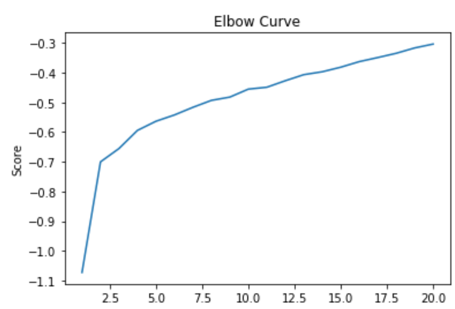
We will perform KMeans clustering on broad and specific venue categories.

Broad Venue Categories:

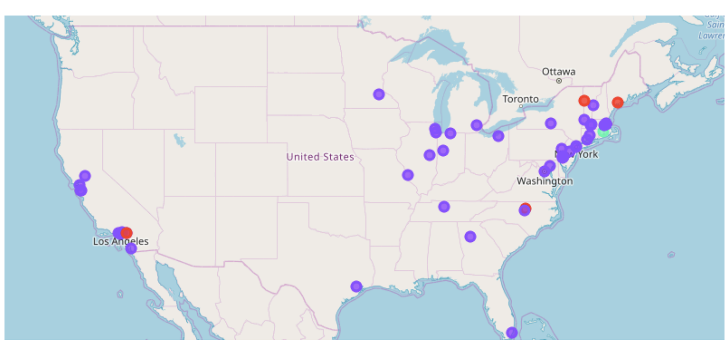
First, we must perform one-hot encoding on the table with colleges and venues and group by the mean frequency of venue categories for each college. The resulting table will look as follows:

	College	Arts & Entertainment	Building	Education	Food	Nightlife	Parks & Outdoors	Shops	Travel
0	Amherst College	0.051282	0.00000	0.012821	0.525641	0.025641	0.051282	0.307692	0.025641
1	Boston University	0.080000	0.06000	0.000000	0.470000	0.070000	0.090000	0.190000	0.040000
2	Bowdoin College	0.063830	0.00000	0.000000	0.659574	0.063830	0.021277	0.148936	0.042553
3	Brown University	0.060000	0.01000	0.010000	0.720000	0.080000	0.040000	0.070000	0.010000
4	Bryn Mawr College	0.036585	0.02439	0.012195	0.487805	0.060976	0.048780	0.317073	0.012195

Next, we will fit the model with the frequency data to produce an elbow curve that will help us determine the optimal number of clusters for the model. It appears that 3 is the optimal number of clusters.



Lastly, the map below is produced, displaying the clusters for easier visualization.

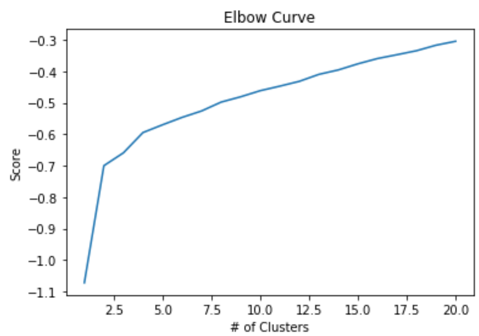


Specific Venue Categories:
The same process is repeated for specific venue categories.

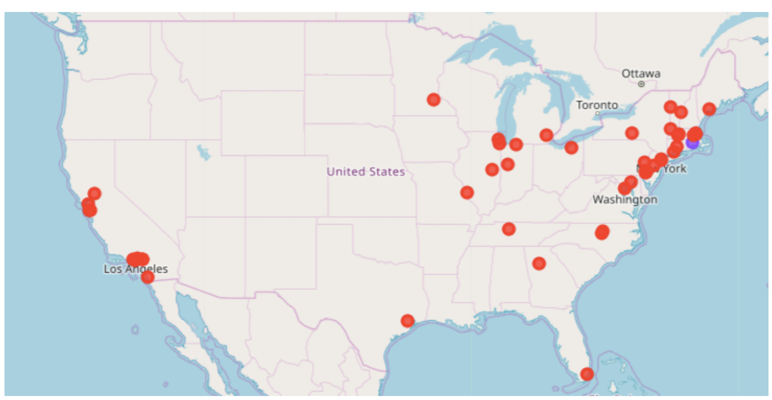
One-hot encoded table:

	College	ATM	Accessories Store	African Restaurant	Airport	Airport Food Court	American Restaurant	Amphitheater	Animal Shelter	Aquarium	...	Waterfall	Waterfront	Whisky Bar	Wine Bar	Wine Shop
0	Amherst College	0.0	0.0	0.0	0.0	0.0	0.051282	0.0	0.0	0.0	...	0.0	0.0	0.0	0.000000	0.00
1	Boston University	0.0	0.0	0.0	0.0	0.0	0.060000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.000000	0.01
2	Bowdoin College	0.0	0.0	0.0	0.0	0.0	0.042553	0.0	0.0	0.0	...	0.0	0.0	0.0	0.021277	0.00
3	Brown University	0.0	0.0	0.0	0.0	0.0	0.030000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.010000	0.00
4	Bryn Mawr College	0.0	0.0	0.0	0.0	0.0	0.024390	0.0	0.0	0.0	...	0.0	0.0	0.0	0.012195	0.00

Elbow curve suggesting 2 is the optimal number of clusters.



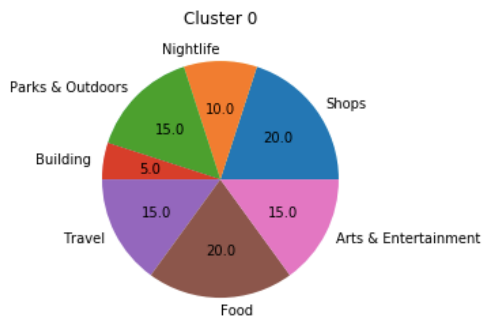
Map displaying clusters:

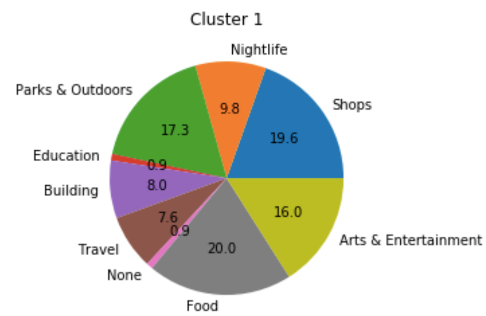


<u>Results:</u>
We will now interpret the clusters for each model (broad & specific).

Broad Venue Categories:

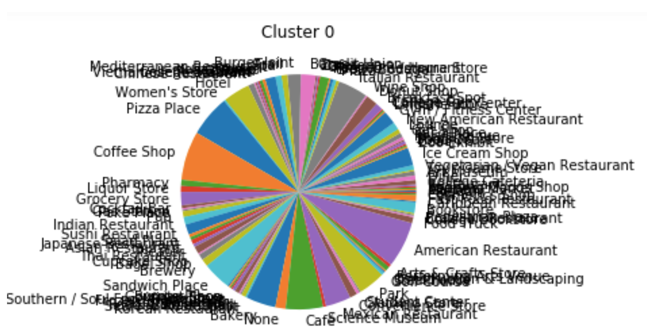
Each cluster appears to have Food and Shops as the top 2 venue categories, but they differentiate in other categories.

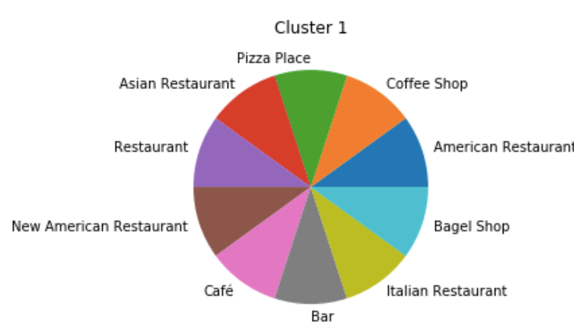




Cluster 0: contains majority of venues in Travel
Cluster 1: contains all venues in Education and majority in Building
Cluster 2: contains majority of venues in Nightlife, Arts & Entertainment, and Parks & Outdoors

Specific Venue Categories:





Interpreting the clusters above for specific venue categories is difficult and the results are inconclusive. It seems that clustering based on broad venue categories is most appropriate for this project.

<u>Discussion:</u>
In this project, we grouped colleges based on similar venue categories. Initially, we attempted clustering for both broad and specific venue categories but realized that clustering based on broad categories is most informative and appropriate. Python proved to be a versatile and powerful language for this project, due to its ample libraries and simplicity

in handling data and producing visuals. Utilizing Foursquare and making API calls was also a simple and quick task.

Errors/Modifications:

1. The Foursquare API provides a maximum of 100 venues for each college, so using other sources to gather additional venues would help create a more accurate model.
2. We utilized KMeans clustering, but another algorithm, such as Hierarchical Clustering or DBSCAN, could be more useful, considering the large number of categories.
3. Due to KMeans being an unsupervised algorithm, different clusters are produced for each execution of the algorithm. The local directory of this project contains 50 different models with different numbers of clusters for each broad and specific venue category. Aggregating this data and further analyzing it may help us better interpret the clusters.

The most difficult part of the project is selecting the optimal algorithm. I hope to better understand techniques to do this in future classes and projects.