CLUSTERING OF POTENTIAL AREAS TO INVEST

ON SERVICES FOR COLLEGE STUDENTS

Christian Vilca

May 27, 2020

1. INTRODUCTION

1.1 BACKGROUND

When an investor wants to open a new store or offer a new service, he selects a location according to several characteristics that would guarantee the success of his investment. In the specific case of an investor whose target customers are college students, there are three main characteristics for a location:

1. The quantity of students that surround the area, as potential customers
2. An approximation of the economic status of students to check if they would pay premium services or products.
3. The kind of venues that are most common around each university.

1.2 PROBLEM

The data that might contribute to cluster the socio-economic status and habits of college students in USA might be the location of all universities in USA, their tuition fee, the quantity of undergrad students enrolled per year, and the venues that surround each university. This project aims to cluster different locations to give each investor a better approach if a certain location has the appropriate rate of potential customers (college students around the location), if the customers are tempted to pay premium services or products that will be offered (approximate economic status) and if the customers will accept the kind of services that the investor will offer (most common venues).

1.3 INTEREST

Investors whose target customers are college students would be interested in determine the best location to open a new store. Also, students that are planning to make an exchange would find useful to know the universities that not only share the amount of tuition fee but also similar venues to adapt easily.

2. DATA

2.1 DATA SOURCES

To determine the places that are most crowded by students I would have to check the geolocation of each one, but this approach will be require a lot of resources. So, I decided to choose the only place where students always get together which is colleges. The

information and statistics about colleges can be found in the College Scorecard of the U.S. Department of Education, whose most recent dataset is from March 30, 2020. This dataset will be used to get information about location, tuition fees, amount of undergrad students and name of universities. The data set is in: https://ed-public-download.app.cloud.gov/downloads/Most-Recent-Cohorts-All-Data-Elements.csv

In order to get more detail of each column within the College Scorecard, I used the data dictionary to get the meaning of each column label and the type of data within them. The data dictionary is:
https://collegescorecard.ed.gov/assets/CollegeScorecardDataDictionary.xlsx

To get the most common venues surrounding each university I have used the the FOURSQUARE API to explore each location.

2.2 DATA CLEANING

The dataset from the College Scoreboard has 6806 rows and 1982 columns. So, the first step was reading the "CollegeScorecardDataDictionary.xlsx" to select the most relevant attributes for the analysis.

To get information about the university I selected the column: ['INSTNM']. To get information about the location I selected the columns: ['CITY'], ['LATITUDE'], ['LONGITUDE']. To get information about the tuition I selected ['TUITIONFEE_IN'], ['TUITIONFEE_OUT'], ['TUITIONFEE_PROG']. To get the number of undergrad students I selected the columns ['UG'], ['UG12MN'] and ['G12MN']

After making a description for all columns, I preprocessed the data of each column:

Columns of university ['INSTNM']: Despite there were some repeated college's names, I chose to analyze the rest of columns to eliminate the rows that had missing values or NaN and this way eliminate the repeated rows that were empty.

Columns of Location ['CITY'], ['LATITUDE'], ['LONGITUDE']: I didn't make any change because the data inside the columns was completed.

Columns of Tuition ['TUITIONFEE_IN'], ['TUITIONFEE_OUT'], ['TUITIONFEE_PROG']: I decided to delete the column ['TUITIONFEE_PROG'] because most of its values were NaN. Then, I eliminated the rows that were NaN from the columns ['TUITIONFEE_IN'] and ['TUITIONFEE_OUT'].

Columns of Undergrad Students ['UG'], ['UG12MN'], ['G12MN']: I decided to delete the column ['UG'] and ['G12MN'] since most of its values were NaN. Then, I eliminated the rows that were NaN from the column ['UG12MN'].

To summarize I deleted the columns whose values NaN were more than a half of each column

2.3 FEATURE SELECTION

After data cleaning, there were 3551 rows and 7 columns. Upon examining the meaning of each feature, I realize that the columns ['TUITIONFEE_IN'] and ['TUITIONFEE_OUT'] share a similarity in values, however while some colleges share the same value for both, other increase the value for ['TUITIONFEE_OUT'] which is the out-of-state tuition fee. This increase is part of the politics of each university, and does not represent a variable to trust, because varies according to each university. So, I deleted the column ['TUITIONFEE_OUT'].

As a result, the columns selected from the College Scoreboard are:

Table 1. Attributes selected of the College Scoreboard

COLUMN	DATATYPE	DESCRIPTION
INSTNM	STRING	Institution Name
CITY	STRING	City
LATITUDE	FLOAT	Latitude
LONGITUDE	FLOAT	Longitude
TUITIONFEE_IN	INTEGER	In-state tuition and fees
UG12MN	INTEGER	Count of Undergraduate students enrolled during a 12 month period

And the columns selected from the FOURSQUARE API are:

Table 2. Attributes selected of the Venues Information

COLUMN	DATATYPE	DESCRIPTION
['venue']['name']	STRING	Venue Name
['venue']['location']['lat']	FLOAT	Venue Latitude
['venue']['name']['lng']	FLOAT	Venue Longitude
['venue']['categories'][0]['name']	STRING	Venue Category

3. METHODOLOGY

3.1 EXPLORATORY DATA ANALYSIS

The socioeconomic status of the students was not a feature in the dataset, so I used the attribute ['TUITIONFEE_IN'] to approximate the socioeconomic status of students attending to each college.

To verify the distribution of universities according to their tuition fees, I plotted a histogram (Figure 1). This suggest that there is a range of tuition fees between USD 480 to USD 75,000 and there are more universities whose tuition is less than USD 10,000.00.

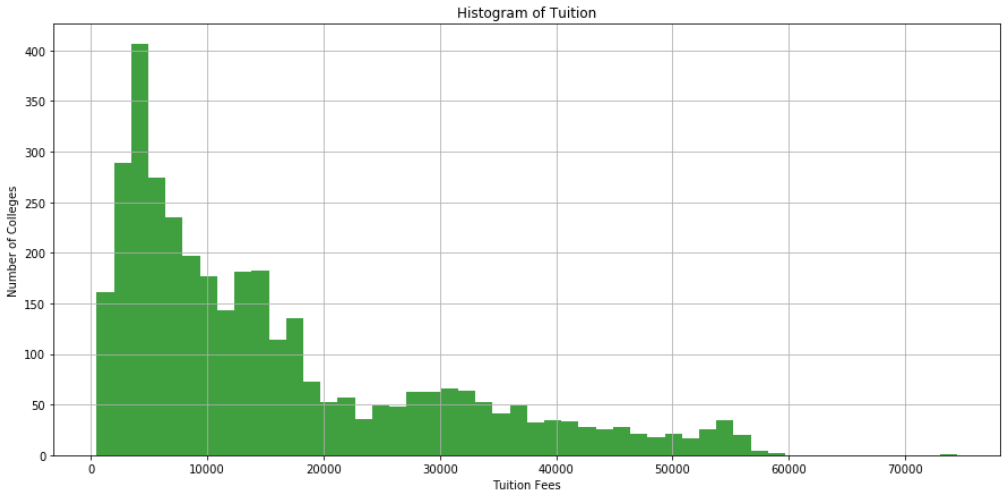


Figure 1. Histogram of Tuition Fees

With the histogram, I decided to set ranges to cluster each university according to the tuition fees in ranges of USD 10,000. So, I got 6 clusters from 0 to 5 and the first cluster was between USD 480 to USD 10,000, and it arranged 1641 colleges (46% of the total colleges filtered) and the last cluster arranged 115 colleges whose tuition was between USD 50,000 to MAX

RELATIONSHIP BETWEEN TUITION FEES OF COLLEGES AND MIDDLE CLASS HOUSEHOLD INCOME

To validate the assumption that the college income could be used to determine the socioeconomic status of students from each university, I plotted a demographic map pointing each university with its corresponding cluster of tuition fees (Figure 2).

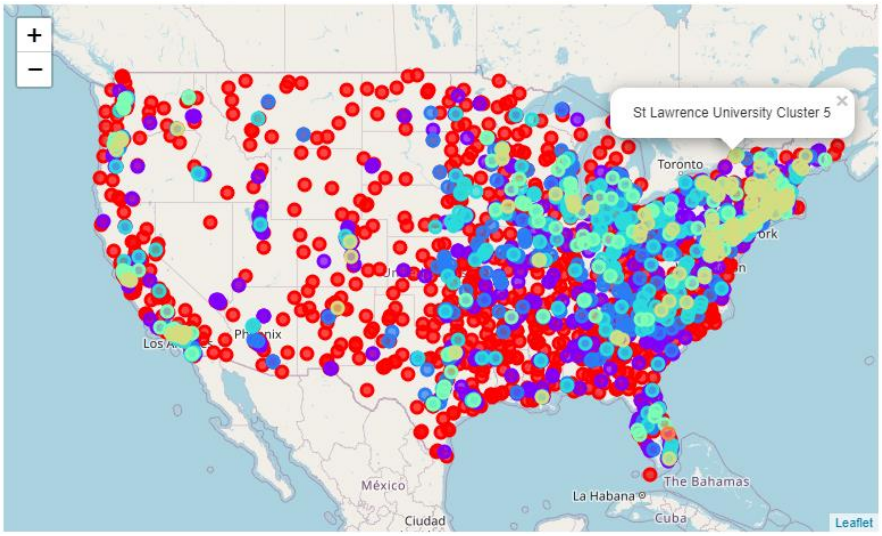


Figure 2. Demographic map of Tuition Fees Cluster

The demographic map previously plotted was compared with a Middle Class Household Income from the Statista Page (Figure 3)

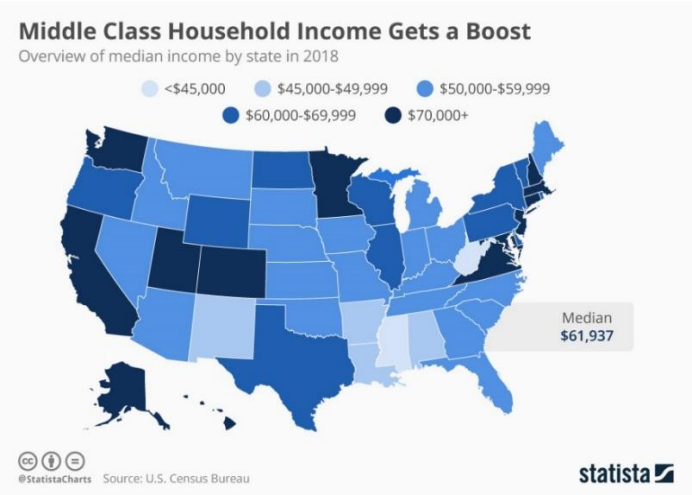


Figure 3. Middle Class Household Income

Source: Statista Charts by Sarah Feldman

After the comparison we can assure that the colleges with higher tuition fees are located in states that have higher household income, for example: Boston, Philadelphia, New York, Los Angeles, San Francisco and so on. To summarize, the attribute [‘TUITIONFEE_IN’] will be used to categorize the socioeconomic level of students.

The attribute that shows the number of undergrad students enrolled per year [‘UG12MN’] can’t be used to cluster the colleges because the number of students per university varies according to the capacity of the university, the reputation of the university or the tuition fee. I plotted a histogram that shows that most universities have less than 20000 undergrad students and it can’t be used to cluster institutions. (Figure 4). However, the attribute [‘UG12MN’] will be useful in an advance analysis when clustering by venues the location of each university.

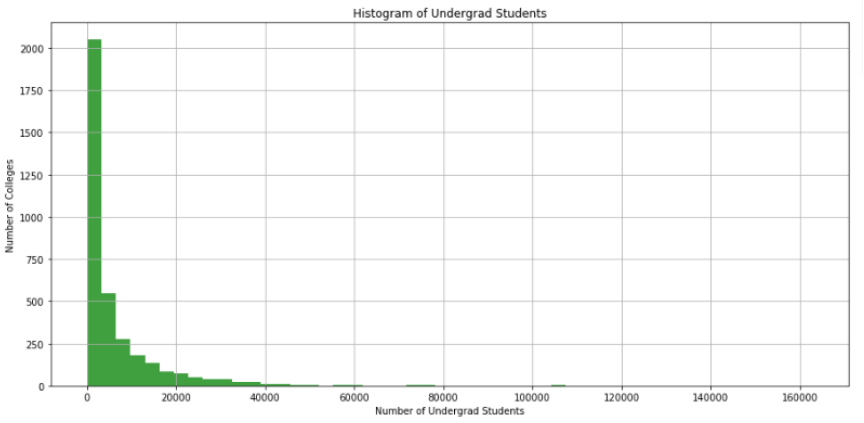


Figure 4. Histogram of Undergrad Students

VENUES SURROUNDING EACH UNIVERSITY LOCATION USIGN FOURSQUARE API

To get more information about the services, stores and public places surrounding each location and due to the restrictions of Foursquare API requests per hour, I have selected a sample of rows from each cluster of tuition fee:

Table 3. Number of samples to cluster with k-means

# Rows	Cluster	# Samples
1641	0	500
939	1	500
348	2	348
337	3	337
171	4	171
114	5	114

Then I passed the latitude and longitude of each college to the Foursquare API to get information about the 100 venues surrounding each college in a radius of 1000 meters. Then, I analyzed and segmented the information that returned from the API according to the category of each venue, so the 10 most common venues per university will appear in a new data frame. For example, for cluster 0 whose tuition fees are from USD 480 to USD 10,000, I have selected the first 500 colleges to get the surrounding venues and identifying the 10 most common. (Figure 5)

	Univ. Name	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	Arizona College-Las Vegas	Sandwich Place	Pool	Casino	Mexican Restaurant	Seafood Restaurant	Pizza Place	Pub	Bridal Shop	Buffet	Café
1	Arizona College-Mesa	Convenience Store	Mexican Restaurant	Grocery Store	Gas Station	Donut Shop	Bar	Coffee Shop	Sandwich Place	Electronics Store	Asian Restaurant
2	Avila University	College Gym	Sandwich Place	Park	Baseball Field	Farm	Falafel Restaurant	Eye Doctor	Event Space	Doctor's Office	Electronics Store
3	California Jazz Conservatory	Coffee Shop	Theater	Yoga Studio	Sushi Restaurant	Brazilian Restaurant	Asian Restaurant	Music Venue	American Restaurant	Café	Sandwich Place
4	Calumet College of Saint Joseph	Bar	Mexican Restaurant	Dive Bar	Pizza Place	Bus Station	Eye Doctor	Dry Cleaner	Eastern European Restaurant	Electronics Store	Event Space

Figure 5. Most common venues per university

3.2 K-MEANS

To find the relationship among the colleges, I used an unsupervised model because the datasets were unlabeled and Foursquare API return the top10 venues that surrounded each university, so I selected the clustering technique **k-means**. This partitioning clustering divides the data into k=10 non-overlapping subsets whose objects are very similar. In this case, colleges within each cluster have similar venues surrounding them. I repeated the k-means algorithm to each range of tuition fee. (Table 4)

Table 4. Clustering with k-means

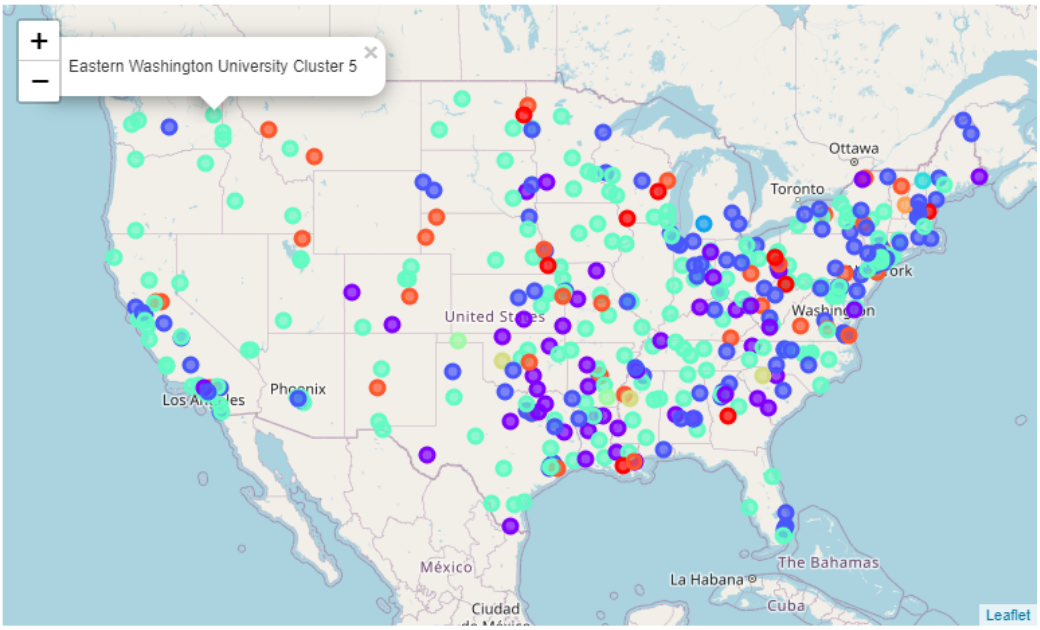
Tuition Fee USD	# Colleges	# Samples	Clus_hist	Clus_ kmeans	# Colleges	Colleges clustered
480 – 10,000	1641	500	0	0	11	494
				1	62	
				2	158	
				3	1	
				4	1	
				5	214	
				6	2	
				7	3	
				8	1	
				9	41	
10,000 – 20,000	939	500	1	0	10	497
				1	7	
				2	154	
				3	1	
				4	1	
				5	10	
				6	41	
				7	1	
				8	4	
				9	268	
20,000 – 30,000	348	348	2	0	1	345
				1	5	
				2	52	
				3	240	
				4	1	
				5	3	
				6	1	
				7	1	
				8	2	
				9	39	
30,000 – 40,000	337	337	3	0	23	336
				1	38	
				2	8	
				3	78	
				4	1	
				5	6	
				6	1	
				7	1	
				8	179	
				9	1	
40,000 – 50,000	171	171	4	0	1	170
				1	2	
				2	45	
				3	9	
				4	1	
				5	1	
				6	94	
				7	1	
				8	15	
				9	1	
50,000 – MAX	115	115	5	0	71	115
				1	9	
				2	28	
				3	1	
				4	1	
				5	1	
				6	1	
				7	1	
				8	1	
				9	1	

The difference between the column "# Samples" and "Colleges clustered" is due to Foursquare API, because some locations have no data registered and the rows are deleted.

4. RESULTS

I plotted a demographic map pointing the location of each university that was clustered by k-means and a description of the top5 venues category that have each cluster.

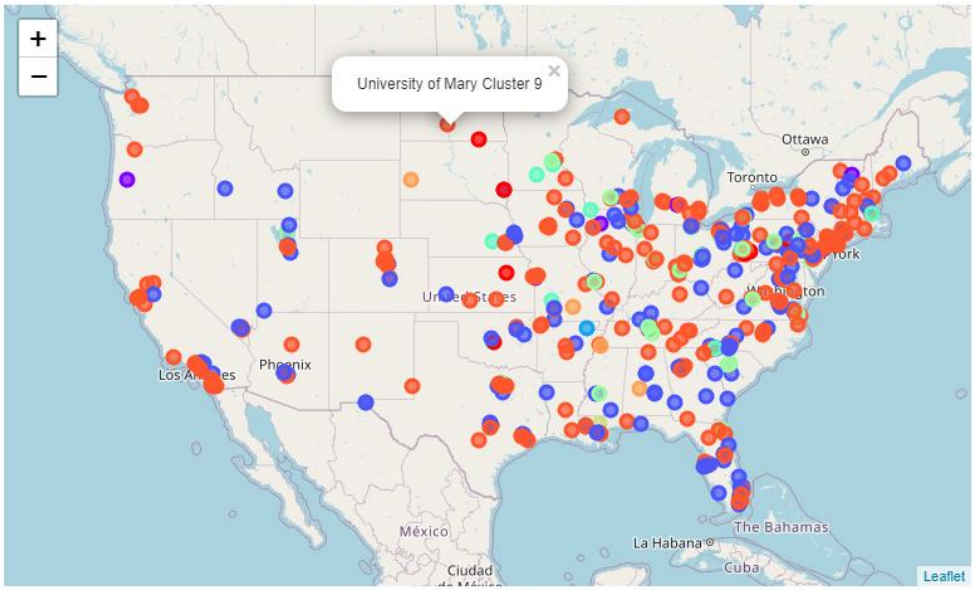
1. Colleges whose tuition fees are between USD 480 – USD 10,000



	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
Avg. Tuition Fee (USD)	7309.9	7975.7	8250.6	7300.0	7970.0	8271.5	7696.0	7147.7	7298.0	8043.9
Avg. Underg. Students	4005	7604	4092	534	1527	12070	2599	2936	1659	5017
Colleges Clustered	**11**	**62**	**158**	**1**	1	214	2	**3**	1	41
1st Common Venue	0.64 Bar	0.64 Fast Food Rest.	0.22 Pizza Place	Athletics & Sports	Restaur.	0.18 Coffee Shop	Football Stadium	Discount Store	New American Rest.	0.31 Park
2nd Common Venue	0.09 IT Service	0.06 Bakery	0.16 Fast Food Rest.			0.11 Sandwich Place		Convenie. Store		0.09 Trail
3rd Common Venue	0.09 Diner	0.03 Mexican Rest.	0.04 Grocery Store			0.10 Bar				0.07 Golf Course
4th Common Venue	0.09 Café	0.03 Discount Store	0.04 Hotel			0.10 Pizza Place				0.07 Café
5th Common Venue	0.09 Pizza Place	0.03 Pizza Place	0.04 Discount Store			0.05 Hotel				0.05 Grocery Store

The colleges with tuition fees less than USD 10,000 can be categorized mainly in 5 clusters, and the most important categories per cluster are: Bar, Fast Food Restaurant, Pizza Place, Coffee Shop and Park. Also, the clusters 0,1,2,5 and 9 gather the highest number of students and colleges clustered. In cluster 1 and 2, there are 220 colleges surrounded mainly by Pizza Places and Fast Food Restaurants, while in cluster 0,5,9 are 234 colleges surrounded mainly by Coffee Shops, bar and Parks.

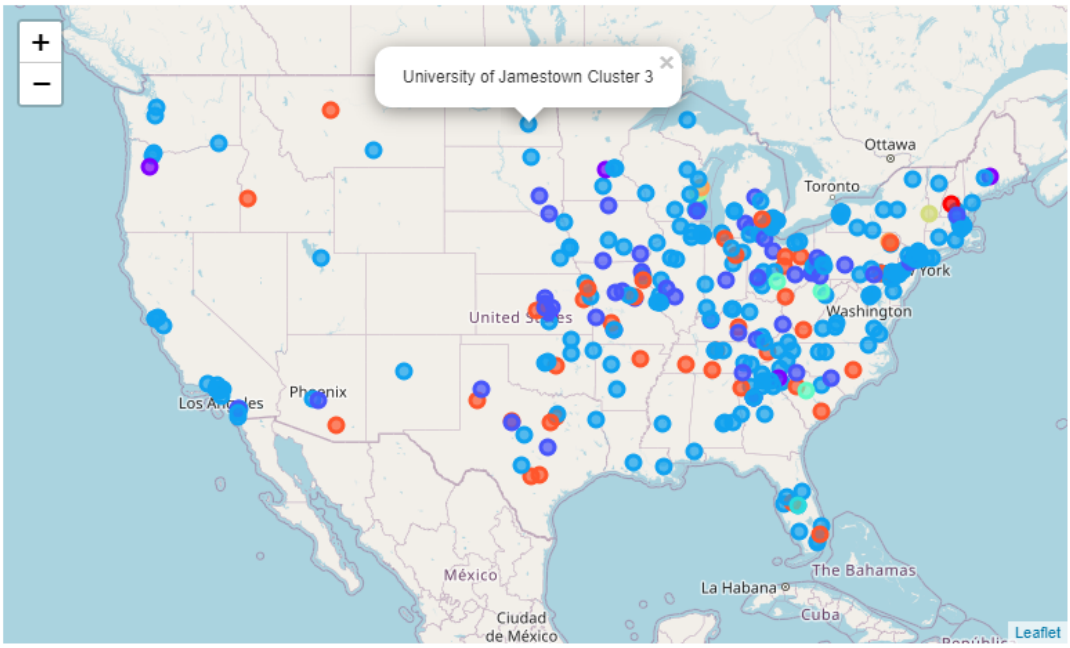
2. Colleges whose tuition fees are between USD 10,000 – USD 20,000



	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
Avg. Tuition Fee (USD)	16,942.5	17,375.7	16,602.8	17,320.0	17,665.0	16,906.6	16,173.1	18,230.0	16,710.8	16,391.4
Avg. Underg. Students	486	504	2069	539	157	2357	1957	149	2628	3595
Colleges Clustered	**10**	7	**154**	**1**	1	**10**	41	1	**4**	268
1st Common Venue	0.7 Bar	0.29 Park	0.29 Fast Food Rest.	Baseball Field	Intersect.	0.2 Park	0.83 Hotel	Bookstore	0.25 Jewlry Store	0.09 Pizza Place
2nd Common Venue	0.1 General Entertai.	0.14 Construc. & Landsc.	0.10 Conveni. Store			0.2 Construc. & Landsc.	0.05 Fast Food Rest.		0.25 American Rest.	0.09 Clothing Store
3rd Common Venue	0.1 Grocery Store	0.14 Trail	0.10 Pizza Place			0.2 History Museum	0.02 College Cafeteria		0.25 Pizza Place	0.09 Coffee Shop
4th Common Venue	0.1 Italian Rest.	0.14 Gym	0.04 Grocery Store			0.1 Scenic Lookout	0.02 American Rest.		0.25 Airport Terminal	0.07 Sandwich Place
5th Common Venue		0.14 Café	0.04 Pharmacy			0.1 Golf Course	0.02 Café			0.06 Hotel

The colleges with tuition fees between USD 10,000 to USD 20,000 can be categorized mainly in 5 clusters, and the most important cluster is 9 with 268 Colleges and venues like Pizza Places and Coffee Shop. On the other hand, cluster 0,1,5 and 6 gather venues related to social places like Bar, Hotels, Parks and so on. Also, the clusters 2 gather venues related to Fast Food Restaurants and represent 154 colleges

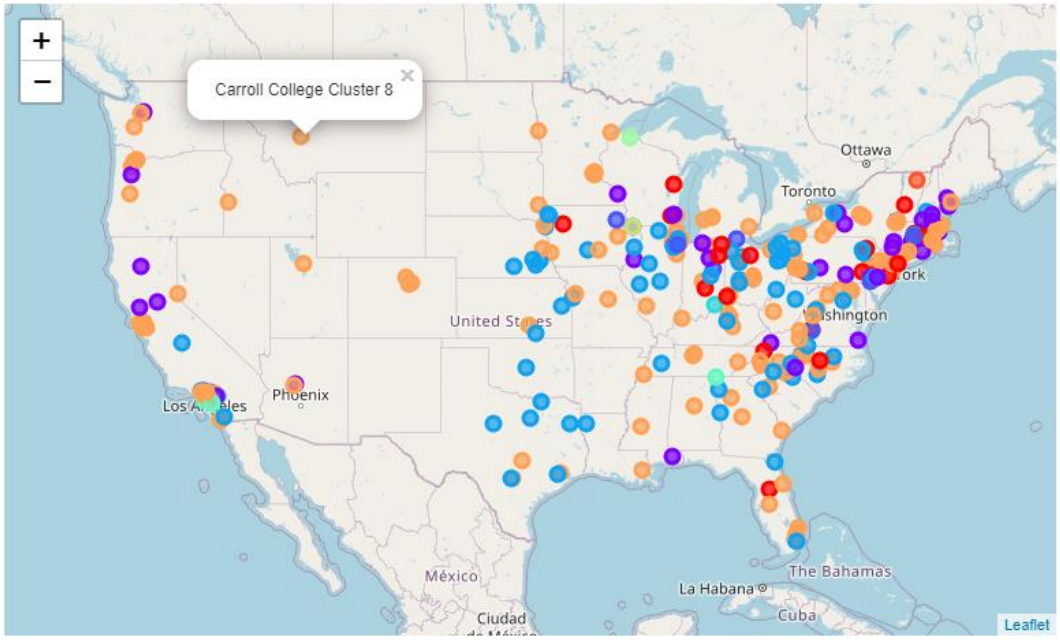
3. Colleges whose tuition fees are between USD 20,000 – USD 30,000



	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
Avg. Tuition Fee (USD)	24,000.0	26,284.4	25,662.0	25,097.0	21,694.0	24,135.3	28,870.0	25,511.0	27,555.0	25,765.4
Avg. Underg. Students	71	804	1663	1952	1123	2167	2900	404	1048	2913
Colleges Clustered	**1**	5	**52**	**240**	1	**3**	1	1	**2**	**39**
1st Common Venue	Lounge	0.6 Coffee Shop	0.33 Pizza Place	0.09 Coffee Shop	Lake	IT Services	Pub	Art Gallery	College Residence Hall	0.61 Fast Food Rest.
2nd Common Venue		0.2 Disc Golf	0.08 Grocery Store	0.08 Pizza Place		Football Stadium			Italian Rest.	0.05 Conveni. Store
3rd Common Venue		0.2 Brewery	0.08 American Rest.	0.07 Hotel		Convenie. Store				0.05 Baseball Field
4th Common Venue			0.06 Gym	0.06 Park						0.05 Discount Store
5th Common Venue			0.06 Mexican Rest.	0.05 Italian Rest.						0.03 Mexican Rest.

The colleges with tuition fees between USD 20,000 to USD 30,000 can be categorized mainly in 4 clusters, and the most important venues per cluster are: Coffee Shop, Pizza Place and Fast Food Restaurant. The universities from clusters 1 and 3 gather 70% of colleges are mainly surrounded by social places like Coffee Shops. On the other hand, clusters 2 and 9 gather venues related to Fast Food Restaurants and Pizza Places.

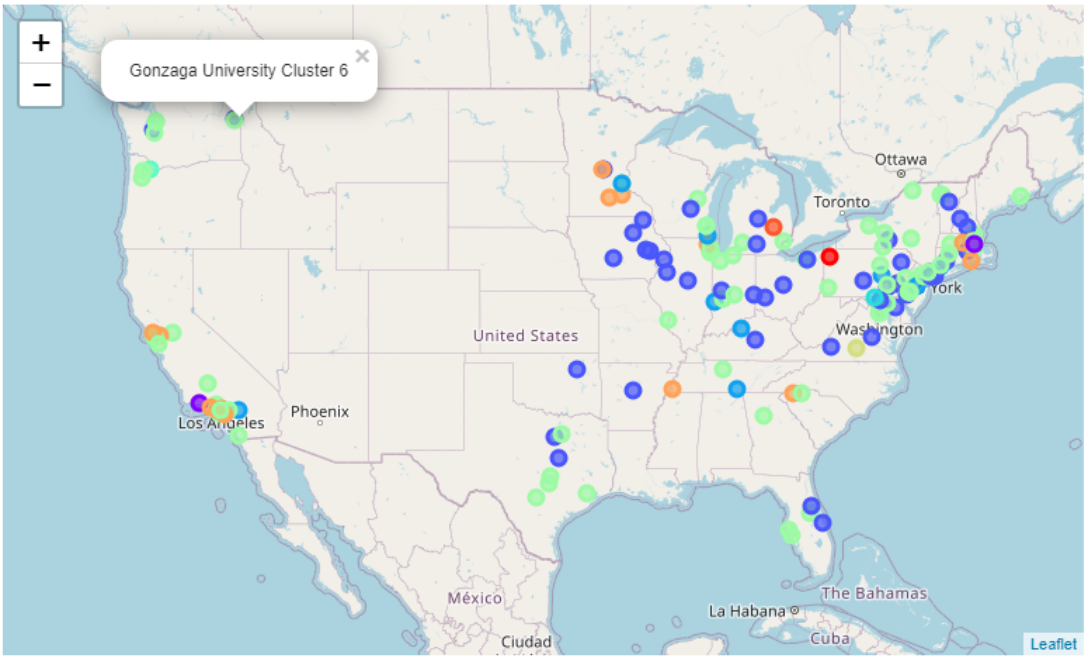
4. Colleges whose tuition fees are between USD 30,000 – USD 40,000



	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
Avg. Tuition Fee (USD)	34,274.3	34,220.2	34,167.4	33,525.1	37,670.0	34,518.3	36,183.0	33,350.0	34,782.7	38,300.0
Avg. Underg. Students	2090	4876	2642	2281	1113	1164	685	817	2454	234
Colleges Clustered	23	38	8	**78**	**1**	**6**	1	**1**	179	1
1st Common Venue	0.39 Bar	0.18 Golf Course	0.75 Pizza Place	0.22 Pizza Place	Bus Station	Trail	Gym Fitness	Dog Run	0.17 Coffee Shop	Bookstore
2nd Common Venue	0.09 Baseball Field	0.08 Ice Cream Shop	0.13 College Cafeteria	0.19 Fast Food Rest.					0.08 Hotel	
3rd Common Venue	0.09 American Rest.	0.08 Baseball Field	0.12 Bagel Shop	0.06 Sandwich Place					0.07 American Rest.	
4th Common Venue	0.04 Athletics & Sports	0.08 Café		0.06 Bank					0.05 Bar	
5th Common Venue	0.04 Diner	0.05 Pub		0.06 Discount Store					0.04 Pizza Place	

The colleges with tuition fees between USD 30,000 to USD 40,000 can be categorized mainly in 5 clusters, and the most important venues per cluster are: Bar, Golf Course, Pizza Place and Coffee Shop. Most of colleges and students are gather in clusters 0,1 and 8 whose social places are Bar, Golf Place and in a greater quantity Coffee Shops. While the cluster 2 and 3 gather venues related to Pizza Place and Fast Food Restaurants.

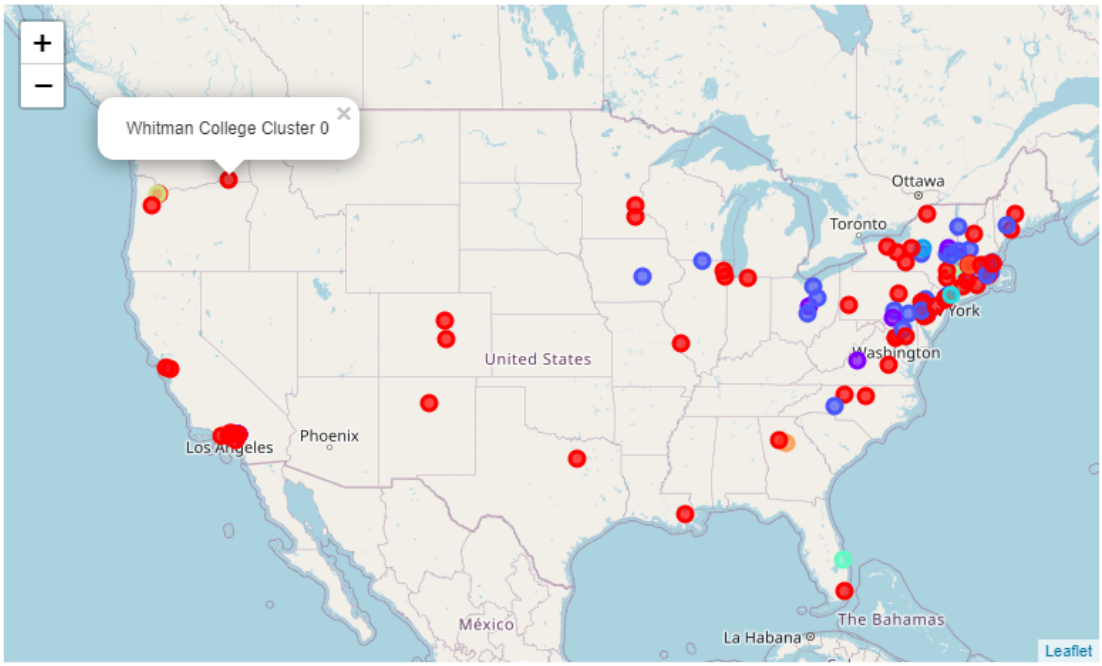
5. Colleges whose tuition fees are between USD 40,000 – USD 50,000



	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
Avg. Tuition Fee (USD)	47,540.0	42,687.0	44,136.5	45,559.4	41,350.0	45,904.0	44,732.6	45,746.0	44,382.3	43,490.0
Avg. Underg. Students	1859	2191	3446	2976	1961	4241	3307	1061	2483	2139
Colleges Clustered	1	2	45	**9**	1	1	94	1	15	1
1st Common Venue	Pharmacy	Breakfast Spot	0.53 Pizza Place	0.33 Coffee Shop	College Basketball Court	Cruise	0.16 Coffee Shop	Food Court	0.27 Park	Bar
2nd Common Venue		Baseball Field	0.11 Bar	0.33 Baseball Field			0.12 American Rest.		0.13 Trail	
3rd Common Venue			0.06 Fast Food Rest.	0.22 Park			0.07 Bar		0.07 Food Court	
4th Common Venue			0.04 American Rest.	0.11 College Theater			0.06 Café		0.07 Beach	
5th Common Venue			0.44 Sandwich Place				0.06 Italian Rest.		0.07 Pub	

The colleges with tuition fees between USD 40,000 to USD 50,000 can be categorized mainly in 4 clusters, and the most important venues per cluster are: Pizza Places, Coffee Shops and Parks. The cluster 3,6,8 and 9 gather most of the students whose social places are Coffee Shops, Parks and Bars. On the other hand, cluster 2 gather Pizza places.

6. Colleges whose tuition fees are between USD 50,000 – MAX



	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
Avg. Tuition Fee (USD)	53,755.8	54,760.0	53,781.8	54,620.0	50,175.0	74,514.0	55,082.0	50,934.0	51,306.0	51,668.0
Avg. Underg. Students	6495	1750	3210	2039	98	432	432	2240	1002	1304
Colleges Clustered	71	9	28	**1**	1	1	1	1	1	1
1st Common Venue	0.25 Coffee Shop	0.56 American Rest	0.39 Pizza Place	Pub	Trail	Airport	Soccer Field	ATM	Trail	Bus Stop
2nd Common Venue	0.08 Bar	0.11 Coffee Shop	0.11 Italian Rest.							
3rd Common Venue	0.07 Park	0.11 Hotel	0.07 Pub							
4th Common Venue	0.07 American Rest.	0.11 College Theater	0.07 American Rest.							
5th Common Venue	0.04 Café	0.11 Theater	0.07 Bar							

The colleges with tuition from USD 50,000 to maximum are categorized mainly in 3 cluster, and the most important venues are: Coffee Shops, Pizza Places and American Restaurants. The rest of colleges have different venues, so there is no similarity between them.

5. DISCUSSION

After the analysis, I have noticed that the clusters from each range of tuition fee can be categorized as food places and social places and their proportion vary among each range of tuition fee:

Table 5. Distribution of social and food places that are close to colleges

Range Tuition Fee	Social Places		Food Places		Mixed	
	# Cluster	Proportion	# Cluster	Proportion	# Cluster	Proportion
0 – 10,000	0,3,5,9	55 %	1,2,4,8	45 %	6,7	0 %
10,000 – 20,000	0,1,5,6	14 %	2	31 %	3,4,7,8,9	55 %
20,000 – 30,000	1,6,7,3	70 %	2,9	26 %	0,4,5,8	4 %
30,000 – 40,000	0,1,8	71 %	2,3	26 %	4,5,6,7,9	3 %
40,000 – 50,000	3,6,8,9	70 %	2,7	27 %	0,1,4,5	3 %
50,000 - Max	0,3	63 %	1,2	32 %	4,5,6,7,8,9	5 %

I have noticed that in colleges with tuition fees until USD 10,000, food places that surround universities have a similar proportion than Social Places. However, in the case of colleges with tuition fees between USD 10,000 to 20,000, most of the colleges were categorized in a clusters that doesn't show a predominance of social places or food places, but the rest colleges. Surprisingly, when the tuition fee is between USD 20,000 to 50,000, the more profitable business are bars, coffee shops and hotels, and food places have a smaller proportion, but that is still significant. Finally, when categorizing colleges with tuition fees above USD 50,000, the proportion of social places was also above food places, but not as much as previous ranges.

The most emblematic food places are pizza places and fast food restaurants, and in the case of social places, the most emblematic are coffee shops, parks and bar. Also, for the final cluster of colleges with the highest tuition fee, they seem to have a slight higher number of coffee shops than Pizza places.

6. CONCLUSION

In this study, I approximated the socio-economic status of students with the tuition fee of each university. To confirm this measure, I compared the tuition fees clustered in ranges against a choropleth map of the household income from the web Statista by Sarah Feldman. I requested information about the venues surrounding each university with Foursquare API. Then I used the partitioned-based clustering k-means, to find similarities among the venues that surround each university per category of tuition fee. The analysis can help an investor, whose target customer are college students, to identify the potentiality and opportunities

of any store that could be open close to a college. The analysis estimates the socio-economic status of each student to suggest the investor if a premium service will have potential customers and the amount of students that walks around each location. Furthermore, the analysis shows the most common venues that surround each university and how they vary according to the tuition fees of each university and the amount of students. This last approach will help an investor identify which service would have better chances within a specific location and know the competitors.