# CLUSTERING OF POTENTIAL AREAS TO INVEST

# ON SERVICES FOR COLLEGE STUDENTS

**Christian Vilca**

**May 26, 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.

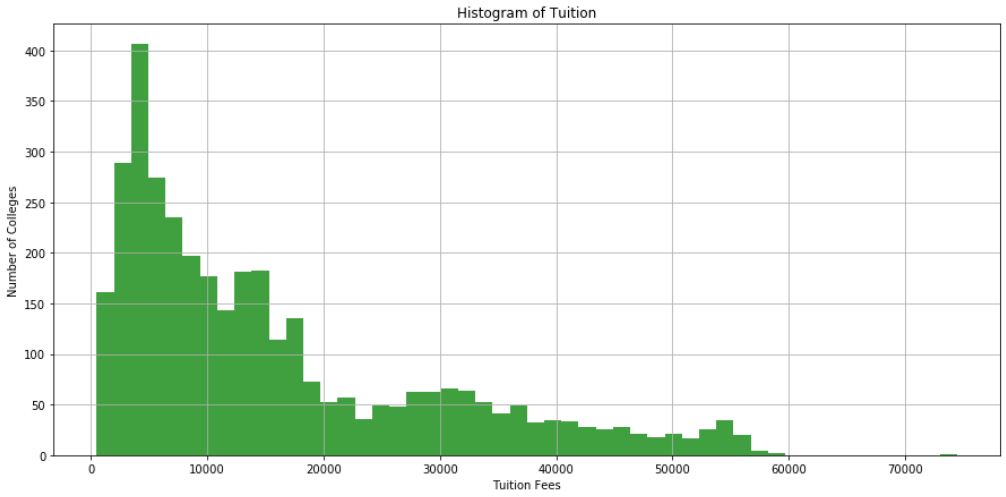



Figure 1. Histogram of Tuition Fees

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).

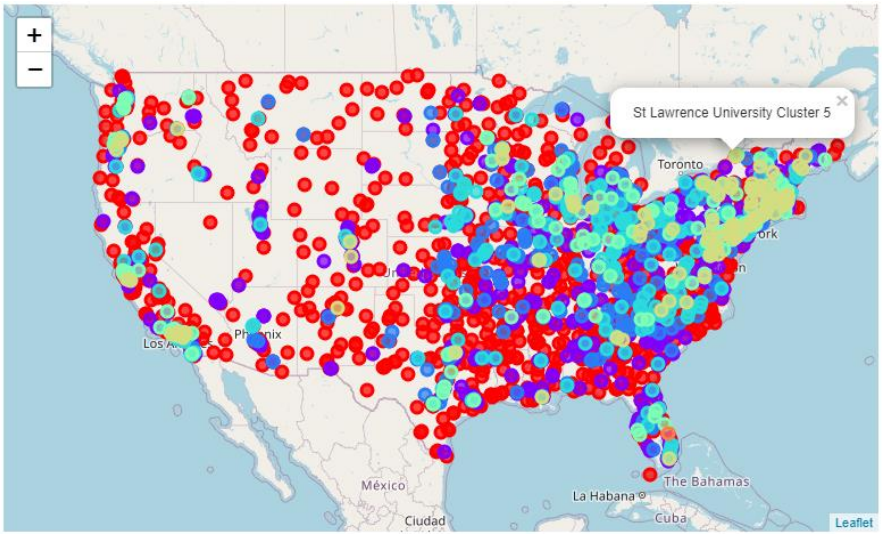



Figure 2. Demographic map of Tuition Fees Cluster

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

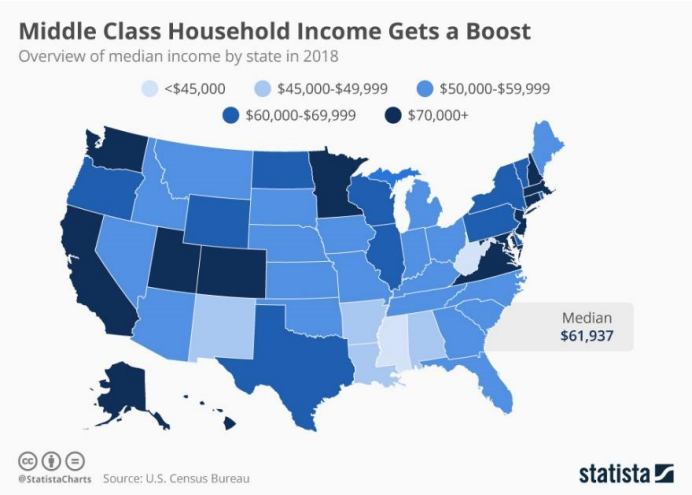



Figure 3. Middle Class Household Income

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.

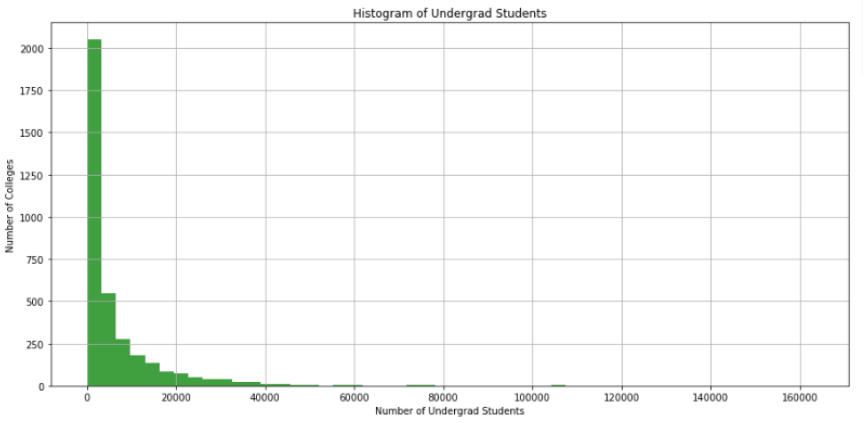



Figure 4. Histogram of Undergrad Students

**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

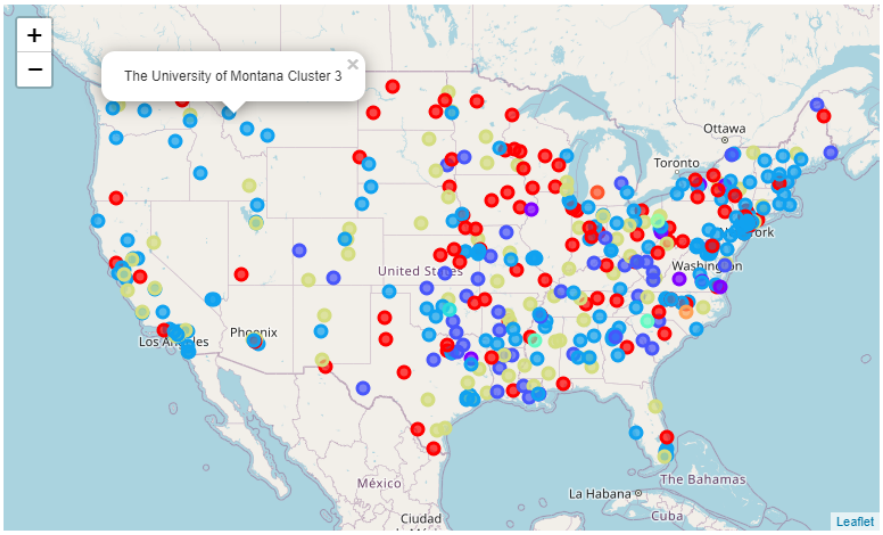
| Tuition Fee USD | # Colleges | # Samples | Clus_hist | Clus_ kmeans | # Colleges | Colleges clustered |
|---|---|---|---|---|---|---|
| **480 – 10,000** | 1641 | 500 | 0 | 0 | 118 | 494 |
| | | | | 1 | 9 | |
| | | | | 2 | 61 | |
| | | | | 3 | 208 | |
| | | | | 4 | 1 | |
| | | | | 5 | 2 | |
| | | | | 6 | 1 | |
| | | | | 7 | 92 | |
| | | | | 8 | 1 | |
| | | | | 9 | 1 | |
| **10,000 – 20,000** | 939 | 500 | 1 | 0 | 13 | 497 |
| | | | | 1 | 1 | |
| | | | | 2 | 85 | |
| | | | | 3 | 50 | |
| | | | | 4 | 1 | |
| | | | | 5 | 172 | |
| | | | | 6 | 159 | |
| | | | | 7 | 1 | |
| | | | | 8 | 14 | |
| | | | | 9 | 1 | |
| **20,000 – 30,000** | 348 | 348 | 2 | 0 | 28 | 346 |
| | | | | 1 | 1 | |
| | | | | 2 | 14 | |
| | | | | 3 | 1 | |
| | | | | 4 | 1 | |
| | | | | 5 | 24 | |
| | | | | 6 | 1 | |
| | | | | 7 | 1 | |
| | | | | 8 | 136 | |
| | | | | 9 | 139 | |
| **30,000 – 40,000** | 337 | 337 | 3 | 0 | 2 | 336 |
| | | | | 1 | 4 | |
| | | | | 2 | 27 | |
| | | | | 3 | 88 | |
| | | | | 4 | 9 | |
| | | | | 5 | 21 | |
| | | | | 6 | 1 | |
| | | | | 7 | 181 | |
| | | | | 8 | 2 | |
| | | | | 9 | 1 | |
| **40,000 – 50,000** | 171 | 171 | 4 | 0 | 13 | 171 |
| | | | | 1 | 119 | |
| | | | | 2 | 1 | |
| | | | | 3 | 1 | |
| | | | | 4 | 1 | |
| | | | | 5 | 31 | |
| | | | | 6 | 1 | |
| | | | | 7 | 2 | |
| | | | | 8 | 1 | |
| | | | | 9 | 1 | |
| **50,000 – MAX** | 115 | 115 | 5 | 0 | 1 | 114 |
| | | | | 1 | 105 | |
| | | | | 2 | 1 | |
| | | | | 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.

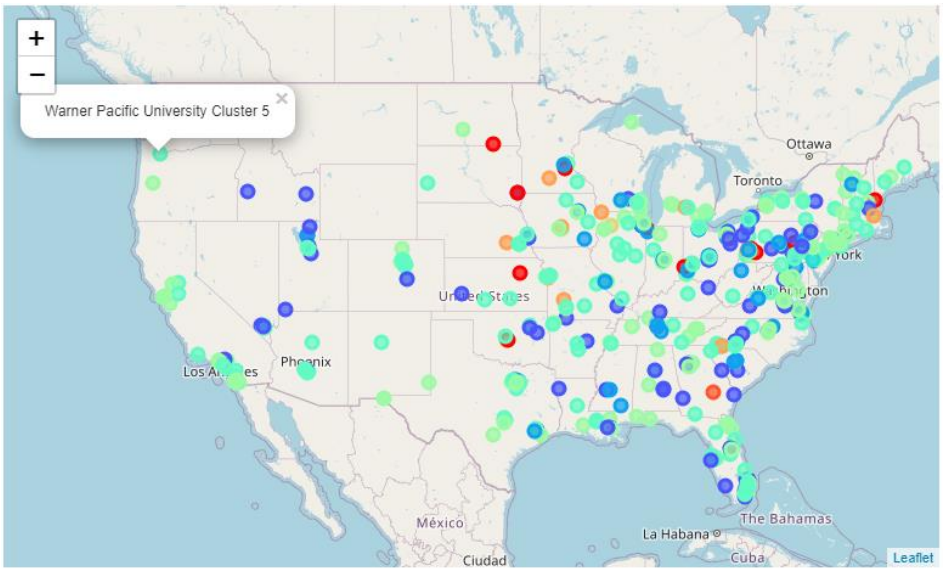
### 1. Colleges whose tuition fees are between USD 480 – USD 10,000




| | Cluster 0 | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 | Cluster 7 | Cluster 8 | Cluster 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Avg. Tuition Fee (USD)** | 8247.0 | 8682.3 | 8017.6 | 8137.3 | 6570.0 | 7054.0 | 7000.0 | 8277.9 | 8328.0 | 7300.0 |
| **Avg. Underg. Students** | 7453 | 2572 | 4826 | 7484 | 43 | 1756 | 89 | 13159 | 45 | 534 |
| **Colleges Clustered** | **118** | **9** | **61** | **208** | 1 | 2 | 1 | **92** | 1 | 1 |
| **1st Common Venue** | 0.37 Pizza Place | 0.67 Americ. Rest. | 0.61 Fast Food Rest. | 0.07 Fast Food Rest. | Construc. & Landscap. | Discoun. Store | Bar | 0.24 Coffee Shop | Stables | Athletics & Sports |
| **2nd Common Venue** | 0.16 Bar | 0.11 Diner | 0.05 Discoun. Store | 0.07 Hotel | | | | 0.22 Sandwich Place | | |
| **3rd Common Venue** | 0.06 Fast Food Rest. | 0.11 Light Rail Station | 0.05 Pharmacy | 0.05 Coffee Shop | | | | 0.08 Fast Food Rest. | | |
| **4th Common Venue** | 0.04 Park | 0.11 Park | 0.03 Pizza Store | 0.05 Park | | | | 0.05 Conven. Store | | |
| **5th Common Venue** | 0.03 Hotel | | 0.03 Mexican Rest. | 0.03 Bar | | | | 0.04 Bar | | |

The colleges with tuition fees less than USD 10,000 can be categorized mainly in 4 clusters, and the most important categories per cluster are: Pizza Places, American Restaurants, Fast Food Restaurants and Coffee Shops. Also, the clusters 0 and 7 gather the highest number of students and colleges clustered. In cluster 0, there are 118 colleges surrounded mainly by Pizza Places and Bar, while in cluster 7 are 92 colleges surrounded mainly by Coffee Shops and Sandwich Places.

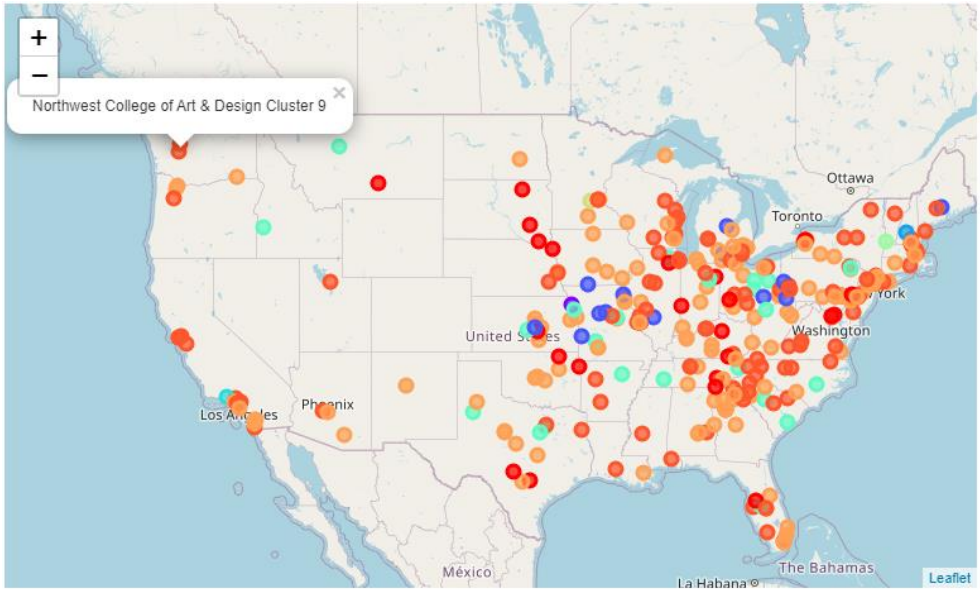
### 2. Colleges whose tuition fees are between USD 10,000 – USD 20,000




| | Cluster 0 | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 | Cluster 7 | Cluster 8 | Cluster 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Avg. Tuition Fee (USD)** | 16,454.5 | 17,665.0 | 16,755.5 | 16,259.5 | 14,468.0 | 16,364.8 | 16,427.8 | 16,100.0 | 17,556.1 | 18,240.0 |
| **Avg. Underg. Students** | 2857 | 157 | 1180 | 1714 | 1587 | 4024 | 2906 | 113 | 1905 | 985 |
| **Colleges Clustered** | **13** | 1 | **85** | **50** | 1 | **172** | 159 | 1 | **14** | 1 |
| **1st Common Venue** | 0.54 Bar | Intersecti on | 0.53 Fast Food Rest. | 0.70 Hotel | Church | 0.19 Pizza Place | 0.08 Coffee Shop | Trail | 0.36 Park | Chinese Rest. |
| **2nd Common Venue** | 0.15 Café | | 0.06 Pizza Place | 0.14 Convenie. Store | | 0.12 Clothing Store | 0.08 Coffee Shop | | 0.14 Construct. & Landscap. | |
| **3rd Common Venue** | 0.08 General Enterta. | | 0.05 Convenie. Store | 0.02 Fast Food Rest. | | 0.11 Sandwich Place | 0.07 Italian Rest. | | 0.14 Baseball Field | |
| **4th Common Venue** | 0.08 Disc Golf | | 0.04 Chinese Store | 0.02 American Rest. | | 0.06 American Rest. | 0.06 Bar | | 0.07 Athletics & Sport | |
| **5th Common Venue** | 0.08 Discoun. Store | | 0.04 Discount Store | 0.02 Golf Club | | 0.04 Coffee Shop | 0.06 Mexican Rest. | | 0.07 Skate Park | |

The colleges with tuition fees between USD 10,000 to USD 20,000 can be categorized mainly in 5 clusters, and the most important venues per cluster are: Bar, Fast Food Restaurants, Hotels, Pizza Places, and Parks. The clusters 2 and 5 gather the majority of students and the venues are related Fast Food Restaurants and Pizza Places. While the clusters 0, 3 and 8 gather venues related to social places like Bar, Café, Hotels, Parks and so on.

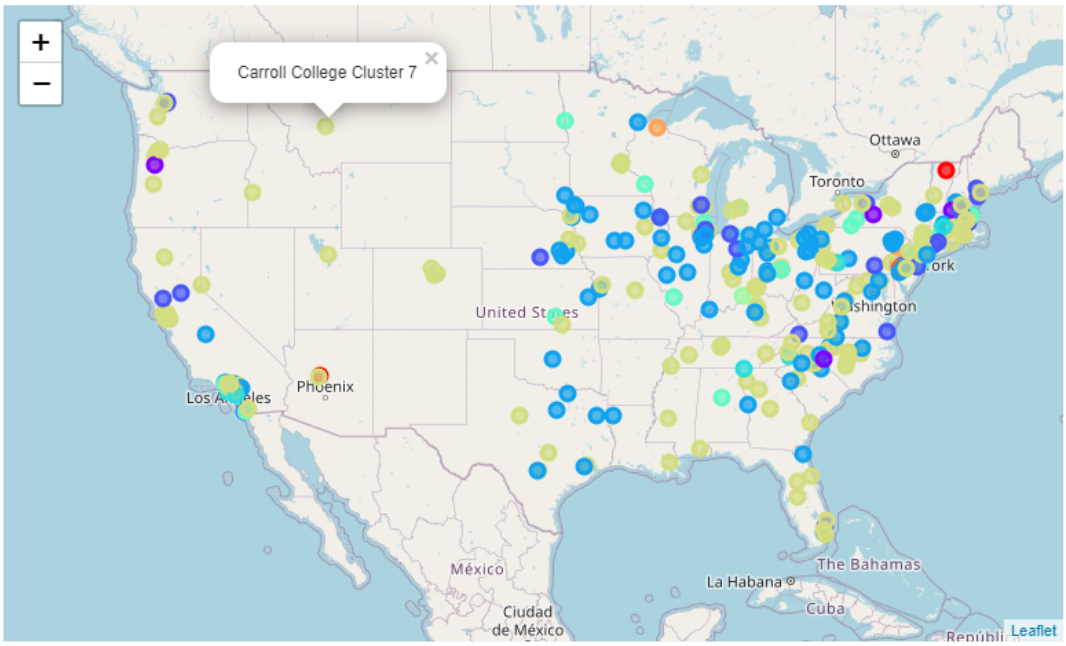
### 3. Colleges whose tuition fees are between USD 20,000 – USD 30,000




| | Cluster 0 | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 | Cluster 7 | Cluster 8 | Cluster 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Avg. Tuition Fee (USD)** | 25,215.6 | 29530.0 | 25,876.9 | 24,000.0 | 25,000.0 | 25,661.9 | 25,511.0 | 26,200.0 | 25,212.9 | 25,180.3 |
| **Avg. Underg. Students** | 1963 | 2193 | 1540 | 71 | 376 | 2377 | 404 | 1178 | 1899 | 2088 |
| **Colleges Clustered** | **28** | 1 | **14** | **1** | 1 | **24** | 1 | 1 | **136** | **139** |
| **1st Common Venue** | 0.29 Park | Park | 0.57 Pizza Place | Lounge | Trail | 0.54 Fast Food Rest. | Art Gallery | Crown College | 0.20 Pizza Place | 0.13 Coffee Shop |
| **2nd Common Venue** | 0.11 Americ. Rest. | | 0.14 Business Service | | | 0.08 Sandwich Place | | | 0.10 Fast Food Rest. | 0.12 Hotel |
| **3rd Common Venue** | 0.11 Construc. & Landscap. | | 0.14 Gym | | | 0.04 Mexican Rest. | | | 0.07 Bar | 0.05 Italian Rest. |
| **4th Common Venue** | 0.04 Soccer Field | | 0.07 Antique Shop | | | 0.04 Cafeteria | | | 0.05 Sandwich Place | 0.04 American Rest. |
| **5th Common Venue** | 0.04 College Quad | | 0.07 Smoke Shop | | | 0.04 Cosmetic Shop | | | 0.04 Mexican Rest. | 0.04 Bar |

The colleges with tuition fees between USD 20,000 to USD 30,000 can be categorized mainly in 5 clusters, and the most important venues per cluster are: Park, Pizza Place, Fast Food Restaurant and Coffee Shop. The universities from clusters 2, 5 and 8 gather half of the total students and the venues from this clusters are related to Fast Food Restaurants and Pizza Places. However, clusters 0 and 9 gather social places like Parks, Landscapes, Coffee Shops, Hotels and so on.

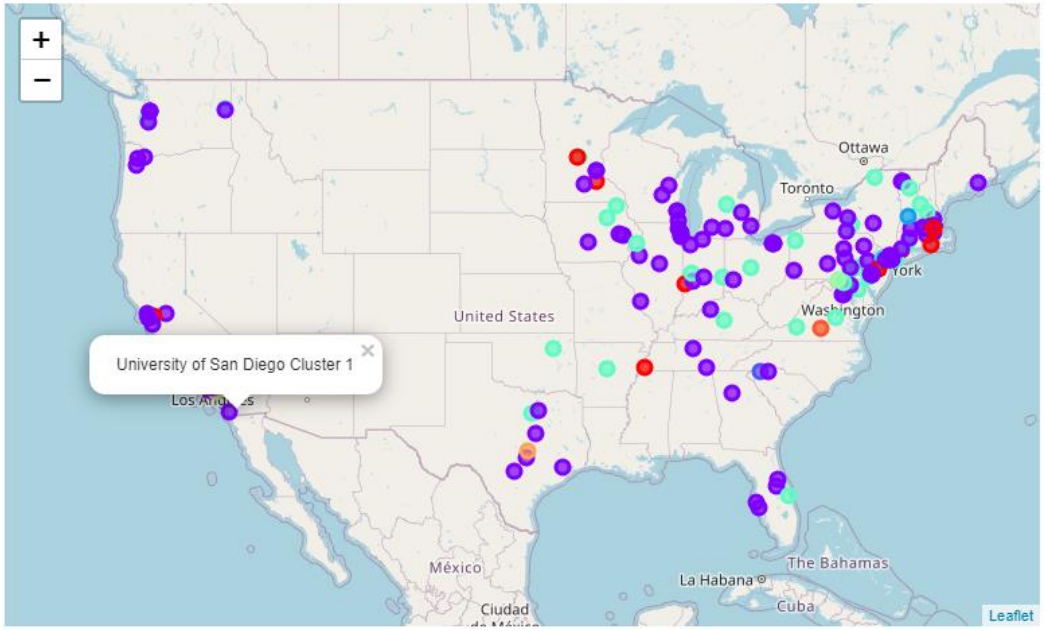
### 4. Colleges whose tuition fees are between USD 30,000 – USD 40,000




| | Cluster 0 | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 | Cluster 7 | Cluster 8 | Cluster 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Avg. Tuition Fee (USD)** | 36,977.0 | 33,054.5 | 33,781.4 | 33,763.4 | 34,786.3 | 33,965.9 | 37,670.0 | 34,788.4 | 36,791.5 | 31,920.0 |
| **Avg. Underg. Students** | 1502 | 1414 | 2301 | 2413 | 1298 | 1977 | 1113 | 2979 | 1767 | 1587 |
| **Colleges Clustered** | 2 | 4 | **27** | **88** | **9** | **21** | 1 | **181** | 2 | 1 |
| **1st Common Venue** | Bookstore | Café | 0.26 Park | 0.25 Pizza Place | 0.44 Trail | 0.29 Hotel | Bus Station | 0.14 Coffee Shop | Perform. Arts Venue | Golf Course |
| **2nd Common Venue** | Zoo | Beach | 0.15 Golf Course | 0.18 Fast Food Rest. | 0.11 Harbor / Marina | 0.09 Beach | | 0.08 Bar | Gym / Fitness Center | |
| **3rd Common Venue** | | Lake | 0.07 Baseball Field | 0.08 Mexican Rest. | 0.11 Pool | 0.09 Theater | | 0.07 American Rest. | | |
| **4th Common Venue** | | | 0.07 Intersect. | 0.08 Sandwich Place | 0.11 Ice Cream Shop | 0.05 Diner | | 0.07 Pizza Place | | |
| **5th Common Venue** | | | 0.07 Athletics & Sports | 0.05 Discount. Store | 0.11 Park | 0.05 Snack Place | | 0.05 Hotel | | |

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: Park, Pizza Place, Trail, Hotel and Coffee Shop. Most of colleges and students are gather in clusters 2, 4, 5, and 7 whose social places are Parks, Golf Clubs, Trails, Pools, Hotels and in a greater quantity Coffee Shops and Bar. While the cluster 3 gather venues related to Pizza Place and Fast Food Restaurants.

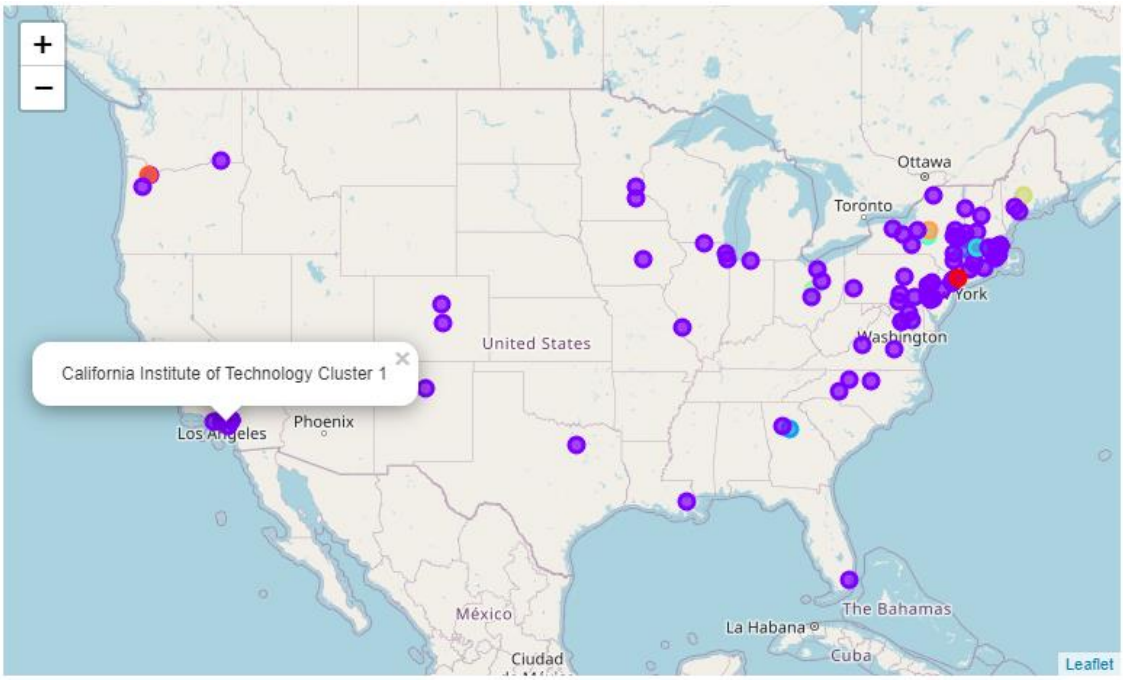
### 5. Colleges whose tuition fees are between USD 40,000 – USD 50,000




| | Cluster 0 | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 | Cluster 7 | Cluster 8 | Cluster 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Avg. Tuition Fee (USD)** | 44,631.0 | 44,594.2 | 49,532.0 | 40,840.0 | 43,142.0 | 44,435.9 | 41,350.0 | 44,996.0 | 42,000.0 | 45,746.0 |
| **Avg. Underg. Students** | 2695 | 3404 | 2821 | 354 | 2886 | 2829 | 1961 | 3836 | 1430 | 1061 |
| **Colleges Clustered** | 13 | 119 | 1 | **1** | 1 | 31 | 1 | 2 | 1 | 1 |
| **1st Common Venue** | 0.31 Trail | 0.17 Coffee Shop | Food Service | Coffee Shop | Hotel | 0.55 Pizza Place | College Basketball Court | Playground | Perform. Arts Venue | Food Court |
| **2nd Common Venue** | 0.23 Park | 0.09 Pizza Place | | | | 0.06 Fast Food Rest. | | Park | | |
| **3rd Common Venue** | 0.08 Liquor Store | 0.08 Bar | | | | 0.06 Bar | | | | |
| **4th Common Venue** | 0.08 Breakfast Spot | 0.07 American Rest. | | | | 0.03 Sandwich Place | | | | |
| **5th Common Venue** | 0.08 Music Venue | 0.06 Italian Rest. | | | | 0.03 Theater | | | | |

The colleges with tuition fees between USD 40,000 to USD 50,000 can be categorized mainly in 3 clusters, and the most important venues per cluster are: Trail, Coffee Shops and Pizza Places. The cluster 1 gather most of the students and main venue category are Coffee Shops and Bars. Also the cluster 0 gather social places like trail and parks. On the other hand, cluster 5 gather Pizza places and Fast Food Restaurants.

### 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)** | 50,175.0 | 53,794.1 | 55,082.0 | 51,306.0 | 51,668.0 | 55,870.0 | 55,930.0 | 55,210.0 | 54,620.0 | 50,934.0 |
| **Avg. Underg. Students** | 98 | 5332 | 432 | 1002 | 1304 | 3034 | 1768 | 2081 | 2039 | 2240 |
| **Colleges Clustered** | 1 | 105 | 1 | **1** | 1 | 1 | 1 | 1 | 1 | 1 |
| **1st Common Venue** | Trail | 0.20 Coffee Shop | Soccer Field | Trail | Bus Stop | Pub | Hotel | Pub | Pub | Lawyer |
| **2nd Common Venue** | | 0.13 Pizza Place | | | | | | | | |
| **3rd Common Venue** | | 0.09 American Rest. | | | | | | | | |
| **4th Common Venue** | | 0.07 Bar | | | | | | | | |
| **5th Common Venue** | | 0.06 Italian Rest. | | | | | | | | |

The colleges with tuition from USD 50,000 to maximum are categorized mainly in 1 cluster, and the most important venues are: Coffee Shops, Pizza Places and Bar. 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 | 4,6,7,9 | 19% | 0,1,2, | 38% | 3,5,8 | 43% |
| 10,000 – 20,000 | 0,3,7,8 | 16% | 2,5,9 | 52% | 1,4,6 | 32% |
| 20,000 – 30,000 | 0,1,3,4,6,9 | 49% | 2,5,8 | 50% | 7 | 1% |
| 30,000 – 40,000 | 0,1,2,4,5,7 | 73% | 3 | 26% | 6,8,9 | 1% |
| 40,000 – 50,000 | 0,1,3,4,7, | 79% | 2,5,9 | 19% | 6,8 | 2% |
| 50,000 - Max | 0,1,2,3,5,6, 7,8 | 50% | 1 | 48% | 4,9 | 2% |

I have noticed that in colleges with tuition fees until USD 20,000, food places that surround universities have a bigger proportion than Social Places, it seems that is higher profitable to open a Fast food restaurant than bar or coffee shop. On the other hand, when the tuition fee is between USD 20,000 to 50,000, the more profitable business are bars, coffee shops and hotels. However, food places still keep a significant proportion. Surprisingly, when clustering colleges with tuition fees above USD 50,000 almost all colleges were clustered in 1 category that shares almost the same proportion for social places and food places.

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.