Predicting hubs of student population in London for targeted marketing

Introduction/Business Problem

Background

Students are a sizable demographic in London since the city's sprawling metropolis is speckled with world class colleges and universities. They also make a very lucrative target audience for companies, especially those looking to introduce new products in the market, since they are generally more open to exploration and experimentation. College and university students are also geographically stagnant for an average period of 4 years which gives companies the opportunity to develop marketing strategies and product packages for this group and see what works. This is the reason why bigwigs such as Apple and Spotify have student discounts and packages and websites such as UNiDAYS exist. Apart from this, high school students are a big target market for universities as possible future students. Having established the need for targeted marketing students, the next step is to do so effectively. It would be advantageous to narrow down the areas that have a high concentration of students so that companies can cut down costs by only targeting those areas.

Problem

London is a huge city covering a whopping area of 607 sq. miles and the purpose of this analysis is to break it down into chunks and identify those neighborhoods where the concentration of students is expected to be high. These neighborhoods, and the accompanying student population, will then be broken down into further clusters for more effective target marketing.

Possible Stakeholders

This analysis may be useful for companies and educational institutes specially looking to target students for marketing.

Data Acquisition and Preprocessing

For the purpose of this analysis, data regarding the neighborhoods and educational institutes of London was required. Data regarding neighborhoods was required so that the city might be broken down into neighborhoods and the target neighborhoods could be identified. Data regarding the location of educational institutes such as colleges and universities was imperative because the target audience i.e. students would most probably be found near these locations.

Data Sources

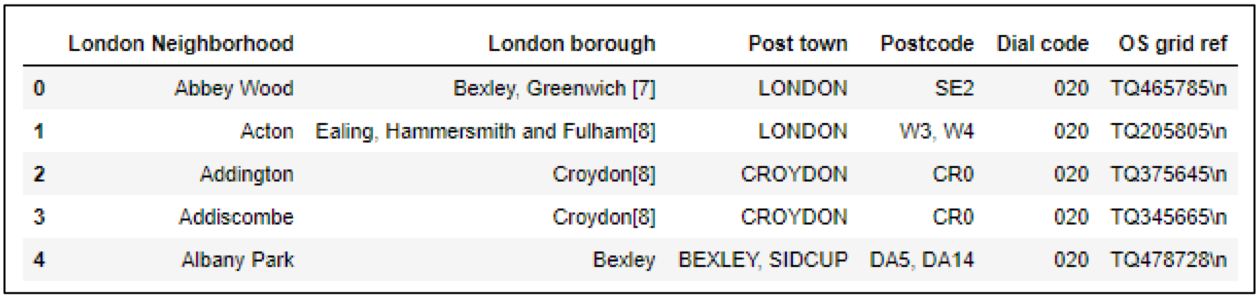
In this project, data for neighborhoods was acquired from Wikipedia (found here), and the data regarding educational institutes was obtained from Foursquare location data. The table available obtained from Wikipedia was scrapped using Beautiful Soup.

Data Usage

The neighborhood data consisting of a breakdown of London into its boroughs and neighborhoods was used as a foundation on which the project was built. The data set consisted of the following information:

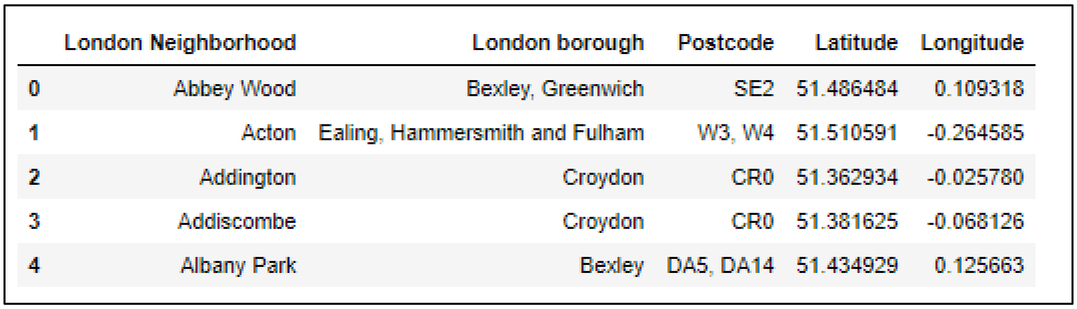
- Location (Neighborhood)
- Borough
- Post Town
- Post Code District
- Dial Code
- OS Grid reference

A snippet of the initial scraped data set is as follows:



	London Neighborhood	London borough	Post town	Postcode	Dial code	OS grid ref
0	Abbey Wood	Bexley, Greenwich [7]	LONDON	SE2	020	TQ465785\n
1	Acton	Ealing, Hammersmith and Fulham[8]	LONDON	W3, W4	020	TQ205805\n
2	Addington	Croydon[8]	CROYDON	CR0	020	TQ375645\n
3	Addiscombe	Croydon[8]	CROYDON	CR0	020	TQ345665\n
4	Albany Park	Bexley	BEXLEY, SIDCUP	DA5, DA14	020	TQ478728\n

Since, information regarding Dial Code and Post Town was nor relevant to this project, these columns were dropped. The OS Grid reference was used to find the latitude and longitude coordinates of the neighborhoods. Some OS Grid references were found missing and were manually fed in the data set. The OSGridConverter package was used to convert these references to location coordinates. A portion of the final data set used for further analysis is as follows:



	London Neighborhood	London borough	Postcode	Latitude	Longitude
0	Abbey Wood	Bexley, Greenwich	SE2	51.486484	0.109318
1	Acton	Ealing, Hammersmith and Fulham	W3, W4	51.510591	-0.264585
2	Addington	Croydon	CR0	51.362934	-0.025780
3	Addiscombe	Croydon	CR0	51.381625	-0.068126
4	Albany Park	Bexley	DA5, DA14	51.434929	0.125663

These coordinates were then fed into the Foursquare API along with a ‘search’ query for educational institutes to identify possible student hubs in the neighborhoods. For this purpose:

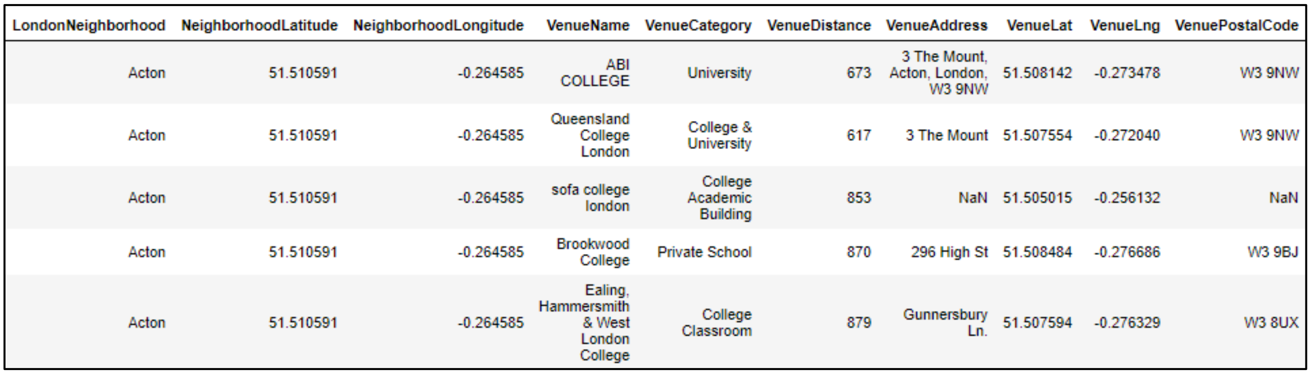
- Different search queries were tried and the query which yielded the most relevant results was ‘college university’.

- The search radius was limited to 1000m.
- The search results were limited to 20 to avoid overlap between neighboring areas.

From the Foursquare results obtained, apart from the venue name and category, relevant location data was retained for further analysis and for possible use by companies and institutes targeting students. This included:

- Distance of venue from neighborhood coordinates
- Venue Address
- Venue coordinates
- Venue postal code

Information returned by Foursquare deemed irrelevant was dropped. A sample of this data is as follows:



LondonNeighborhood	NeighborhoodLatitude	NeighborhoodLongitude	VenueName	VenueCategory	VenueDistance	VenueAddress	VenueLat	VenueLng	VenuePostalCode
Acton	51.510591	-0.264585	ABI COLLEGE	University	673	3 The Mount, Acton, London, W3 9NW	51.508142	-0.273478	W3 9NW
Acton	51.510591	-0.264585	Queensland College London	College & University	617	3 The Mount	51.507554	-0.272040	W3 9NW
Acton	51.510591	-0.264585	sofa college london	College Academic Building	853	NaN	51.505015	-0.256132	NaN
Acton	51.510591	-0.264585	Brookwood College	Private School	870	296 High St	51.508484	-0.276686	W3 9BJ
Acton	51.510591	-0.264585	Ealing, Hammersmith & West London College	College Classroom	879	Gunnersbury Ln.	51.507594	-0.276329	W3 8UX

Methodology

Exploratory Data Analysis

The first part of our analysis was examining the venue categories returned by Foursquare and shortlisting and/or modifying them to make our target data more relevant. A total of 127 categories were obtained and their details are as follows:

Adult Education Center	College Quad	High School	Performing Arts Venue
Art Gallery	College Rec Center	Historic Site	Pharmacy
Athletics & Sports	College Residence Hall	Hookah Bar	Physical Therapist
Bank	College Science Building	Hospital	Plaza
Bar	College Soccer Field	Hospital Ward	Pool
Basketball Court	College Stadium	Lake	Preschool
Bike Rental / Bike Share	College Technology Building	Language School	Private School
Bookstore	College Tennis Court	Laundry Service	Professional & Other Places

Building	College Theater	Law School	Pub
Bus Line	Community College	Library	Rental Car Location
Bus Stop	Convenience Store	Medical Center	Research Station
Business Service	Convention Center	Medical Lab	Residential Building (Apartment / Condo)
Café	Coworking Space	Medical School	Restaurant
Cafeteria	Dance Studio	Meeting Room	Road
Candy Store	Daycare	Middle School	Rock Club
Church	Dentist's Office	Miscellaneous Shop	Salon / Barbershop
Coffee Shop	Doctor's Office	Modern European Restaurant	Sandwich Place
College & University	Elementary School	Monument / Landmark	School
College Academic Building	Emergency Room	Mosque	Sculpture Garden
College Administrative Building	Event Space	Movie Theater	Snack Place
College Arts Building	Field	Museum	Soccer Field
College Auditorium	Food Court	Music School	Social Club
College Bookstore	Food Truck	Music Venue	Spiritual Center
College Cafeteria	Fraternity House	None	Student Center
College Classroom	Garden	Non-Profit	Tech Startup
College Communications Building	General College & University	Nursery School	Tennis Court
College Engineering Building	General Entertainment	Office	Theater
College Gym	Government Building	Other Great Outdoors	Trade School
College History Building	Grocery Store	Paper / Office Supplies Store	University
College Lab	Gym	Park	Urgent Care Center
College Library	Harbor / Marina	Parking	Yoga Studio
College Math Building	Health & Beauty Service	Pedestrian Plaza	

In the table above, it may be seen that some of the categories such as Bus Stop and Church are irrelevant. These categories along with others deemed unconnected to the problem at hand were removed.

For analyzing schools, only schools containing high school and above were included in this analysis since they are part of our targeted market. Students belonging to middle school and younger were deemed to young and were excluded from this analysis

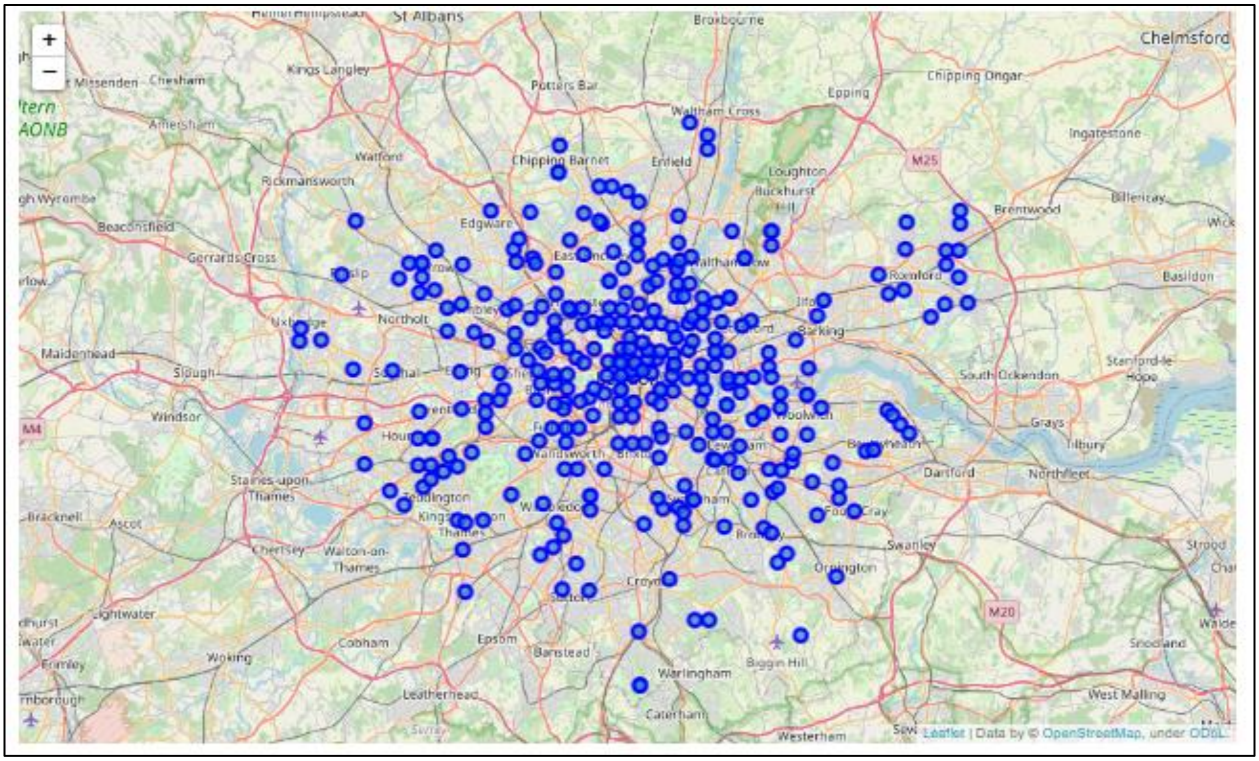
It was also found that some of the categories are overlapping; for instance, categories such as College Arts Building, College Administrative Building are superfluous since they are parts of venues that are already covered in other categories from a broader perspective.

Further inspection of the data revealed that the category 'College Academic Building' included colleges and universities. This category along with other categories containing colleges and universities were grouped under one umbrella category called 'General College & University' to make further analysis more meaningful since they all contained similar venues.

The final venue categories which were selected for further inclusion and analysis are as follows:

- Adult Education Center
- College Residence Hall
- Community College
- Coworking Space
- General College & University
- Government Building
- High School
- Language School
- Law School
- Medical School
- Music School
- Private School
- Residential Building (Apartment / Condo)
- School
- Trade School

The neighborhoods containing these categories were then mapped to see their distribution. The following distribution was obtained:

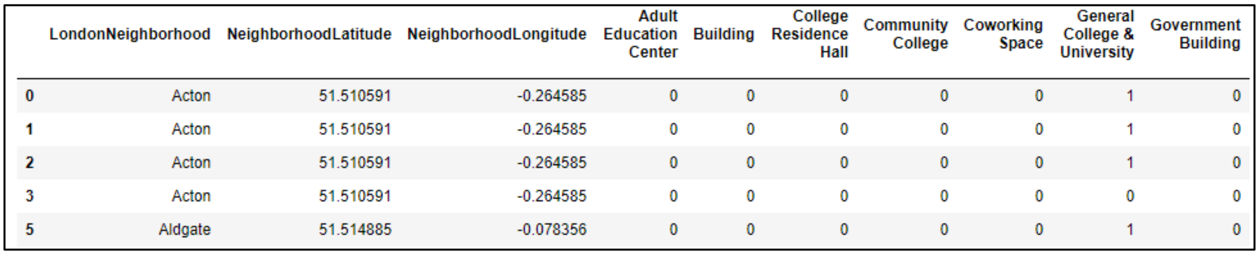


In the map above, it may be seen that most of our neighborhoods of interest are concentrated towards the center of London.

The next part of our analysis will be to further breakdown these neighborhoods into clusters depending upon the extracted information.

K-means clustering

To start off clustering, one-hot encoding was used to quantify which categories are present in our shortlisted neighborhoods: a binary system was used in which a present category was denoted by 1 whereas an absent category was denoted by 0. A sample of the encoded data is as follows:



	LondonNeighborhood	NeighborhoodLatitude	NeighborhoodLongitude	Adult Education Center	Building	College Residence Hall	Community College	Coworking Space	General College & University	Government Building
0	Acton	51.510591	-0.264585	0	0	0	0	0	1	0
1	Acton	51.510591	-0.264585	0	0	0	0	0	1	0
2	Acton	51.510591	-0.264585	0	0	0	0	0	1	0
3	Acton	51.510591	-0.264585	0	0	0	0	0	0	0
5	Aldgate	51.514885	-0.078356	0	0	0	0	0	1	0

After this, the data was grouped by neighborhood and the mean of the grouped data was obtained to determine the frequency of occurrence of each venue category for each neighborhood. This data was then used to form clusters by using k-means clustering.

Different values of number of clusters were tried and the most relevant and insightful clusters were obtained with a value of 4.

The breakdown of the number of neighborhoods in each cluster is as follows:

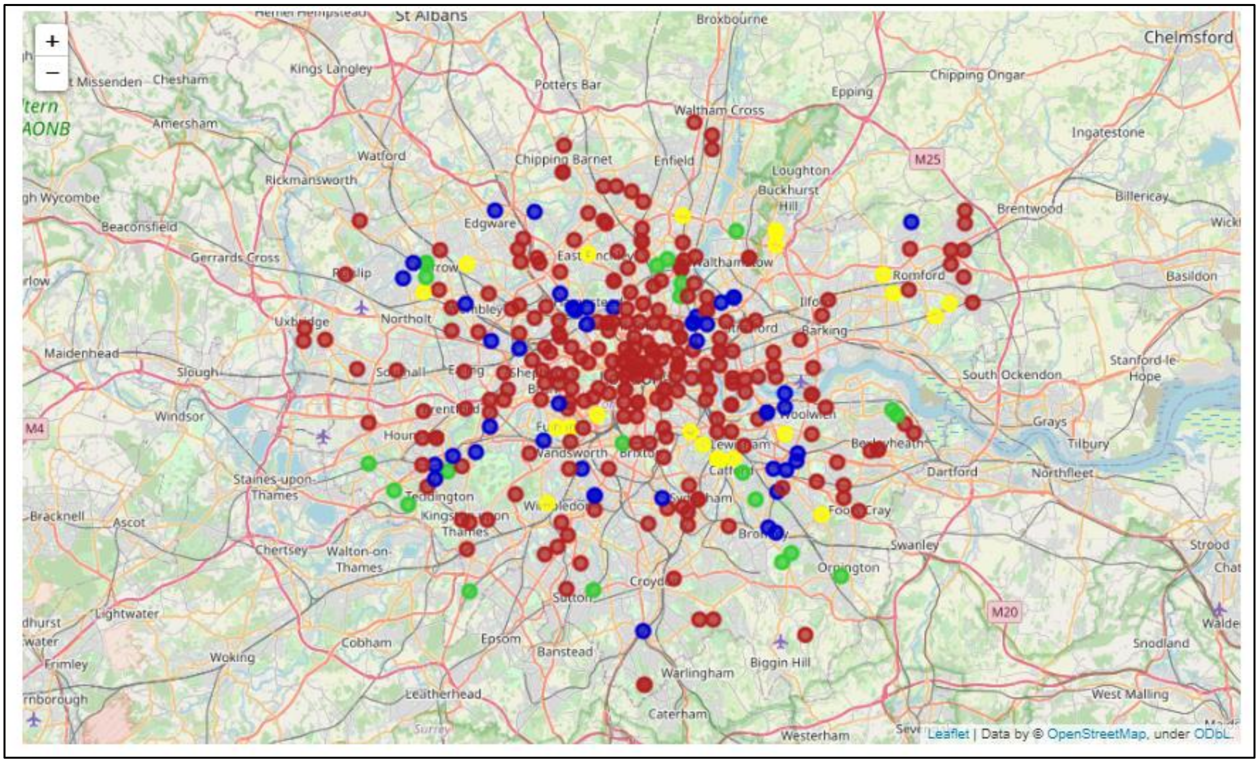
Cluster No.	No. of Neighborhoods in Cluster
0	248
1	45
2	22
3	22

Based on the table above, it may be seen that Cluster 0 contains the maximum number of neighborhoods and Cluster 2 and 3 have the minimum number. In order to further explore these clusters and their makeup, their bar charts were plotted.

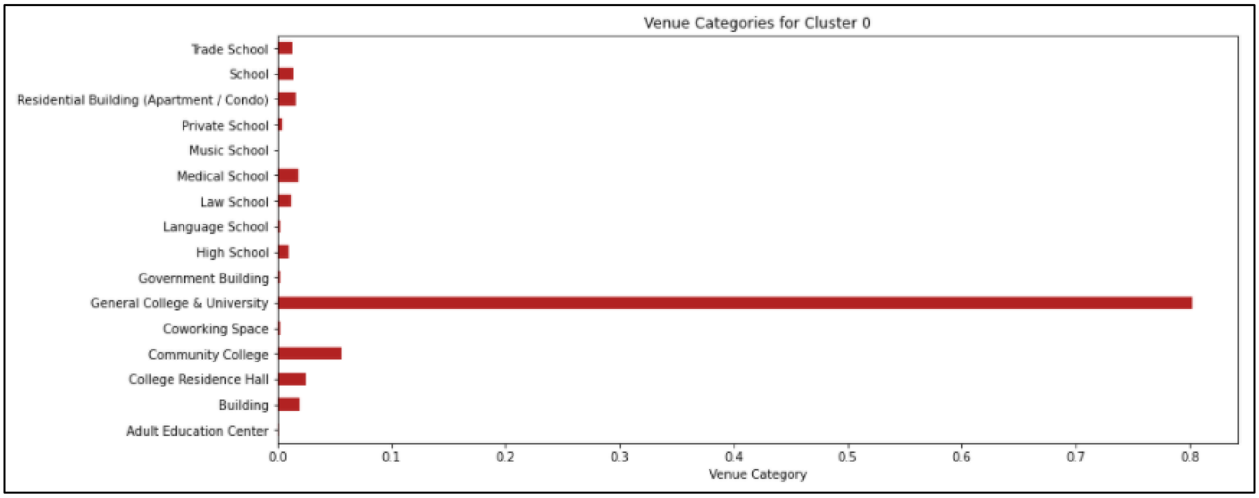
For plotting these charts, the data was grouped by clusters and the mean of the frequency of occurrence of each category in each cluster was obtained. Horizontal bar charts were then plotted with the categories on the Y-axis and the frequency on the X-Axis.

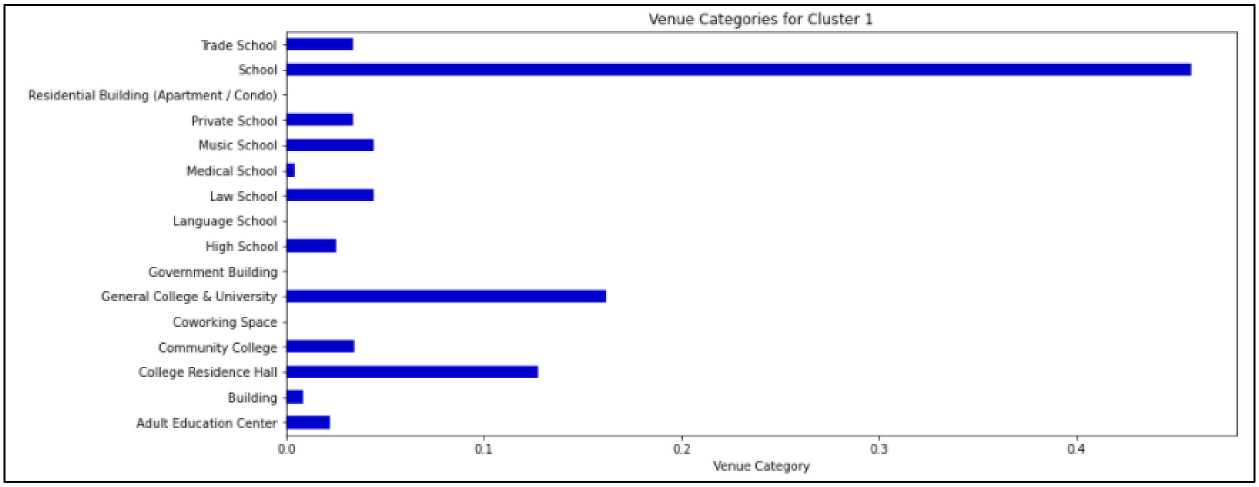
Results

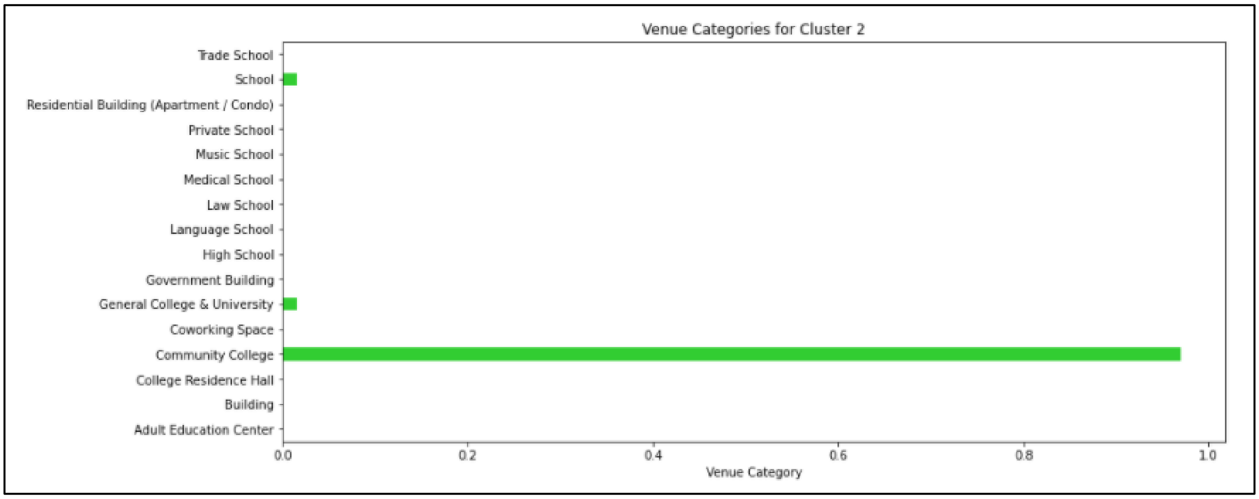
The clusters obtained were mapped and the following distribution was obtained:

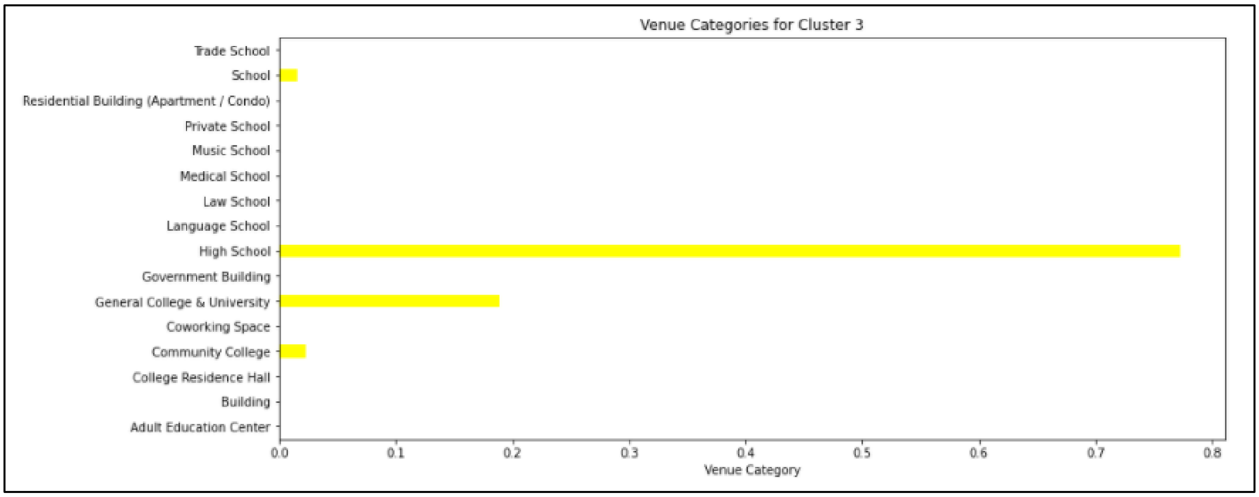


Corresponding to this distribution, the following bar charts were obtained for each cluster:









In the figures above, it may be seen that:

- Cluster 0 mostly consists of universities and colleges and is the largest cluster.

- Cluster 2 mostly almost purely consists of community colleges.
- Cluster 3 mostly consists of high schools.
- Cluster 1 is an amalgam of categories which are smaller in number than the categories already discussed but relevant, nevertheless. This includes student residential buildings, private schools, adult education centers, etc. Schools have the highest frequency of venue category in this cluster.

Discussion

Based on our analysis and results, we can derive the following insights:

- Most of our neighborhoods of interest are concentrated towards the center of London. The density of student hubs is diminishing the further we move away from the center. Hence, a broad suggestion for our stake holders would be to concentrate their marketing efforts towards the center of the city for targeting students belonging to all categories.

- If the stakeholders want to target college and university students, then they should focus on the neighborhoods contained in Cluster 0. These neighborhoods are mostly concentrated towards the center of the city.

- If the stakeholders want to target community colleges, then their areas of interest fall in Cluster 2. These areas can be mostly seen scattered at the peripherals of the city.

- Stakeholders, such as universities targeting future students, can target areas in Cluster 3 and some parts of Cluster 1 to reach their intended audience. These areas are smaller in number but are spread throughout the city. This noted smaller number of schools could be because of the search query used which does not include the word ‘school’. However, it was observed that including the word school in the search query was deviating the focus of the search away from our intended age group which is why it was dropped.

- Stakeholders interested in reaching and engaging a mixed audience instead of concentrating on specific vendor categories can focus on Cluster 1 for targeted marketing.

Conclusion

In this analysis, hubs of student population in London were defined for targeted marketing and. Based on the assumption that the presence of educational institutes is indicative of the presence of students, it was found out that most of our areas of interest in this regard are concentrated towards central London. This student population was then further broken down into clusters so that companies and educational institutes can target specific student populations. Three distinctive categories of students were obtained in this regard:

- Students belonging to colleges and universities who make up the largest number of the student population and are mostly concentrated towards the center of the city.
- Community colleges which are mostly located towards the peripherals of the city.
- School students who are spread throughout the city.

These findings effectively address the problem of locating student hubs in the city of London for effective target marketing, which was the intended purpose of this analysis