**Optimum Location for a Student hostel**

Chinmay K.R

May 2020

1. **Introduction**
   1. **Background**

The USA is home to some of the best universities and colleges. It also boasts of a large number of higher education institutions and colleges. With such a huge and well-formed educational infrastructure, the country is bound to host many local and foreign students aspiring to graduate from these prestigious universities. But not all parts of the country are equally populated with colleges. It is therefore advantageous to know regions with high number of colleges and regions where the infrastructure needs some development. A quick and simple search on the internet would show that the state of New York alone is home to more than 70 colleges. This will cause a huge influx of student population. For instance, this huge student population will definitely have accommodation needs which can be identified and solved with the right data.

* 1. **Problem**

Most of the students pursuing a degree are on a budget, and house rents can be daunting. Low budget student hostels or dormitories can be a solution to this problem. A hostel which is situated in the right place and offering services at the right price can benefit a lot of students. This project aims to predict a suitable location for the construction of student hostels.

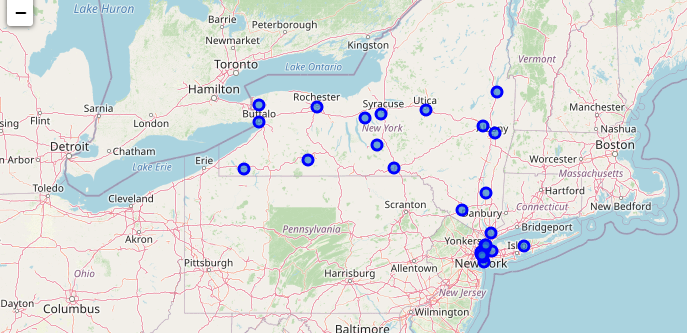
* 1. **Interest**

The stake holders interested in this project would be a hypothetical construction company, willing to enter into hostels and hospitality sector, trying to identify the right locations in the state of New York for establishing hostels so that it can be profitable as well as budget friendly. Also, the students enrolled in colleges and universities in New York.

1. **Data acquisition and Cleaning**
   1. **Data sources**

For the sake of simplicity, we will only be studying the community colleges in NY. The model can be scaled with details from all the colleges if and when needed. The primary source of community college data was web scraped from the site [community college review](https://www.communitycollegereview.com/college-size-stats/new-york) .

We will be using the ‘Location’ feature from this data to obtain **geo coordinates** of each college from geopy library. Below, we can see the exact location of each of these colleges.



We will then use **Foursquare API** to get hotspots around a college where students are most likely to hang out. For example : bookstores, libraries, bars and pubs, Cafes etc.

* 1. **Data cleaning**

The data was aggregated from multiple sources. The primary data being web scraped from an online site. The website contained information about the name of the college, The number of students enrolled in the college, and the location of the college in New York.

This website however, only possessed addresses of colleges in text format, which had to be converted into location coordinates using ‘geopy’ library. The latitude and longitude coordinates are then merged with the original college information data.

Based on the location coordinates, I then used Foursquare API’s explore end point to obtain the trending locations within 200 meters of each college. This fetched a result set including all sorts of venue categories. (Like kids’ stores, office spaces etc.) I then cleaned and filtered the trending locations to only include hotspots where students are most likely to hangout. Hotspots such as pubs and bars, pizza places, eateries, libraries, supermarkets and bookstores were retained whereas offices, grocery stores, kids stores where filtered out.

This data was then one hot encoded for different venue categories and grouped by the college name to take the mean value.

Final dataframe after data preparation looks like below :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| COLLEGE | # Students enrolled | LATITUDE | LONGITUDE | One hot encoded columns |

1. **Methodology** 
   1. **Exploratory Data Analysis**

**Feature engineering and Data munging**

The [website](https://www.communitycollegereview.com/college-size-stats/new-york) lists all 35 colleges in the state of New York. This data is pulled into the dataframe “cc\_df”. Some of the college addresses were not in a correct format to extract the location coordinates, for instance, the location feature needed modification for ‘Nassau Community College’ in order to extract the address.

The location data extracted from the geopy library had a few NA values. This could’ve been obtained via other sources like google earth, but for the simplicity of the model, these colleges were removed.

After obtaining the trending locations, it is filtered based on what could possibly be student hangout hotspots(Cafes, Pizza places etc.) This data is merged with the original college dataand the venue categories are one hot encoded. The data is then grouped based on each college and the mean value is taken. This will be the final data that will be used for modelling.

* 1. **Insights**

Some of the insights obtained in the process have been highlighted below :

* Total number of community colleges : 29
* Number of colleges with trending venues around them : 26

Top 10 hotspots by frequency of appearance :

1. **Coffee Shop 13**
2. **Pizza Place 7**
3. **Sandwich Place 5**
4. **Chinese Restaurant 4**
5. **Gym 4**
6. **Fast Food Restaurant 4**
7. **Bookstore 3**
8. **Burger Joint 3**
9. **Bar 3**
10. **Italian Restaurant 3**

* The most popular hotspots are Coffee shops and Pizza places.
* Gyms are the only non-eatery place that have made to the top 5 hotspots

Top 5 colleges with most number of hotspots around them :

1. **ASA College 12**
2. **CUNY LaGuardia Community College 10**
3. **Fashion Institute of Technology 9**
4. **CUNY Hostos Community College 8**
5. **Monroe College 7**

* These colleges' neighbourhoods could potentially make for a great place for a student hostel since it has all the amenities around and is also less than 200 meters from the college.
* But we will also have to take into account the number of students enrolled in each college.

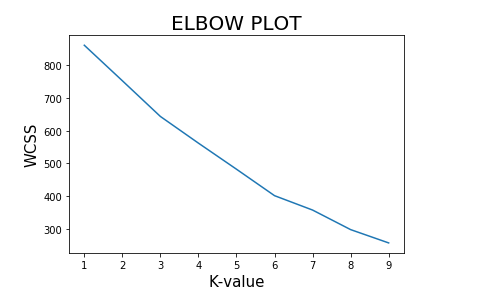
1. **Predictive Modelling**

Since we’re dealing with geospatial data, clustering algorithm has been used. This is an unsupervised learning method, since we do not have any historical labelled data to study from. A few different variations have been tried out with different feature sets. The algorithms used are K-means and DBSCAN.

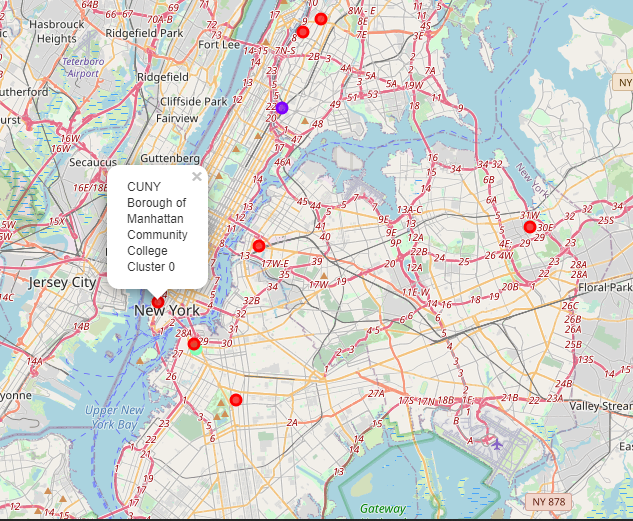
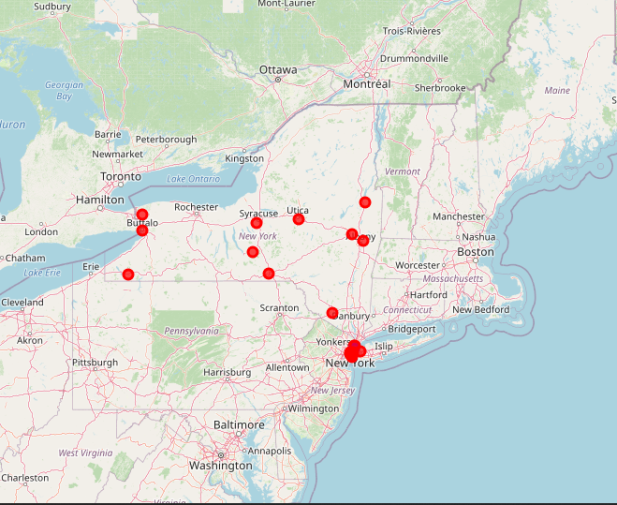
* 1. **KMeans clusters**

KMeans is an iterative algorithm which has a random centroid initialized at first. The algorithm converges by updating the centroids after each iteration. The mean distance (Euclidean in our case) is calculated from the cluster centre to all the points and points are sorted into different clusters based on their vicinity to the centroid. In the next iteration, the centroid is updated by calculating the mean of the centre’s cluster and this process repeats until the algorithm converges.

In our scenario, KMeans was tested for multiple values of k and an elbow plot was plotted to determine the reduction in the WCSS ( Within Cluster Sum of Squares).



The elbow plot for our data is ambiguous since there is no discernible elbow joint. However, we can see a faint elbow at k = 3 clusters. Hence, we will build a model with 3 clusters.



The 3 clusters can be summarised as :

1. The colleges in **cluster 1** (RED cluster) are surrounded by a wide variety of hotspots. These college neighbourhoods have a lot of amenities around them and it would be a safe bet to open a student hostel around any of these colleges.

The neighbourhood around CUNY Borough of Manhattan Community College would be the best bet since it alone has more than 26000 students enrolled and also the neighbourhood boasts of a wide variety of hotspots. The vicinities have different types of gyms, grounds, bookstores, multi-cuisine restaurants, supermarkets, bars etc.

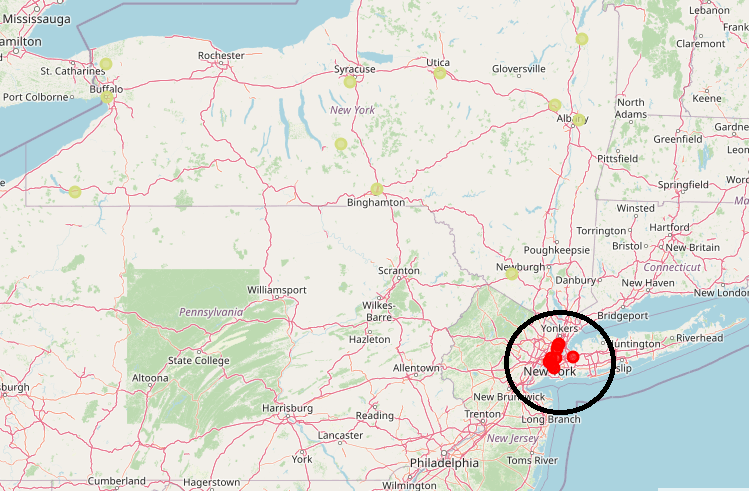
1. The colleges in **cluster 2** (BLUE cluster) are surrounded by hotspots like Bars and fast food restaurants. These neighbourhoods are considerably less fancy compared to the college neighbourhoods in cluster 1. The students might not find as many hotspots or amenities around this neighbourhood.
2. The colleges in **cluster 3** are surrounded by hotspots like International cuisine restaurants.

The interesting point in this cluster is that it has only 1 college and it is surrounded by as many as 13 differet types of hotspots.

* 1. **DBSCAN**

The DBSCAN algorithm views clusters as areas of high density separated by areas of low density.The central component to the DBSCAN is the concept of core samples, which are samples that are in areas of high density. A cluster is therefore a set of core samples, each close to each other (measured by some distance measure) and a set of non-core samples that are close to a core sample (but are not themselves core samples). There are two parameters to the algorithm, min\_samples and eps, which define formally what we mean when we say dense.

In our data, we can use the DBSCAN to get a completely different perspective or approach to the problem. Since Dbscan is basically used to identify dense regions, we can use this to identify regions of the NY state which are densly populated with colleges. Based on this approach, we can use the location of the colleges to cluster.



From the image, it is apparent that we have one high density region,(Red cluster) i.e the downtown NYC and the remaining colleges in the state (Golden points) are considered as outliers as they are not dense enough to form a cluster (min\_samples = 4).

Based on DBSCAN :

* If the builder were to establish just one hostel in the state of New York, then it is clear that he would make the most profit by situating the hostel in the dense region, i.e. Manhattan/Brooklyn/Bronx region.
* A closer look at the cluster on the map would reveal that situating the hostel at the center of this dense region would be the best approach. Based on this, it would be best to establish the hostel near Central park, Manhattan.

1. **Results**

To summarise the different models, we can state that,

* If the plan is to build multiple hostels in feasible areas, we should refer to the KMeans output. In KMeans, Colleges in cluster 1 have the best facilities in the neighbourhood and the number of students enrolled in these colleges also is on the higher side.
* If the builder is planning on establishing just 1 hostel in the state, then we will have to go with the DBSCAN model. The centre of the dense cluster would be a good location which happens to be near Central park, Manhattan.

1. **Future directions**

* The future scope would be to obtain a survey from a statistically large enough sample of students about their preferences in hostels and hotspots. More data would definitely help in obtaining further insights and different types of models could be developed.
* We could assign different weights to the hotspots based on the survey to obtain a more realistic model.
* In the current project, DBSCAN has been used with default distance function. We can try out haversine distance which could be a bit more accurate considering the geo spatial data.
* We could also try dimensionality reduction methods before applying clustering. But this would be more or less the same as obtaining insights from the students about the preferred neighbourhood attractions, and choosing only the right features to model.