**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.

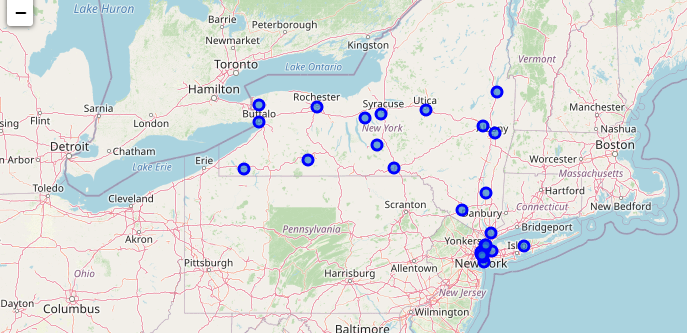
* 1. **Interest**

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 |