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. The search results for each neighborhood were limited to 10 to avoid overlap between neighborhoods and the radius was kept at 1000m. The results obtained through the Foursquare query were then analyzed to identify relevant venue categories and narrow neighborhoods which have these venue categories. These neighborhoods were identified as the target neighborhoods. Other relevant location data regarding the obtained venues was kept on record for further analysis and for possible use by companies and institutes targeting students. Information returned by Foursquare deemed irrelevant was dropped.