**IBM Professional Data Scientist Specialization**

**The Battle of Neighborhoods**

**Finding the best area for a student in London**

**1. Introduction**

**1.1 Background**

London is a popular destination for higher education where diverse students from all around the world gather to study.

According to the data published by the Higher Education Statistical Agency (HESA), in the academic year 2016-17, there were 2317880 students at UK higher education institutions. As of 2016/17, according to official international enrollment statistics, 442,750 international students were attending university in the UK. Of which 112205 students were in London which make up 24 percent of immigrant students at higher education institutions. This means that at least 110,000 students are looking for a new home in London every year, even if domestic students from outside of London are not considered.

**1.2 Problem**

Student halls are the most reliable means of housing for students, especially if one is completely new to the city and is not familiar with how rental contracts work. However, as they are in high demand, it is not easy to secure a place in one. Therefore this project aims to explore different neighborhoods of London and find the best area to build a new student hall for international students to solve this persistent problem as well as to find a new business opportunity.

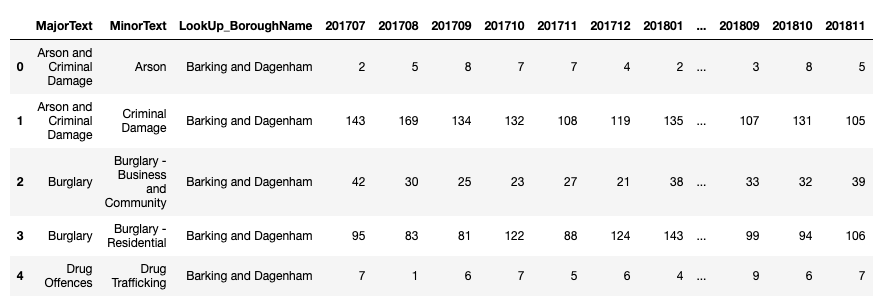
This research is expected to benefit real-estate investors looking for a profitable location or international students looking for a place to live in London.

From the student perspective, a lot of factors come into play when finding the best accommodation, including location and rent. In this project, however, the study will only focus on the safety and the general atmosphere of the neighborhood for simplification. Distances to universities are also an important factor in choosing a student hall, but as student halls accept students from different universities, it will be disregarded in this project.

**2. Data Acquisition and Preprocessing**

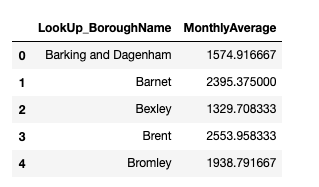
In this project, three different datasets will be used to solve the problem - London Recorded Crime, List of London Boroughs, and Foursquare API. After acquiring them from original and reliable sources, they will be wrangled and cleansed into more useful forms for our further analysis.

**2.1 London Recorded Crime**



Shown above is London crime records classified by boroughs and crime type in the recent 24 months. It consists of 1584 Observations and 27 columns. It was acquired directly from London Data Store.

For further analysis, the numbers of crimes were calculated into monthly averages, and crime categories were not considered in this research for simplification. This process turned the above dataset to a simple one as below

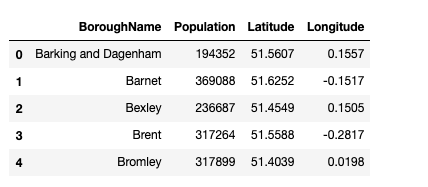
. 

**2.2 List of London Boroughs**

The second dataset used was information on boroughs in London, scrapped from Wikipedia.

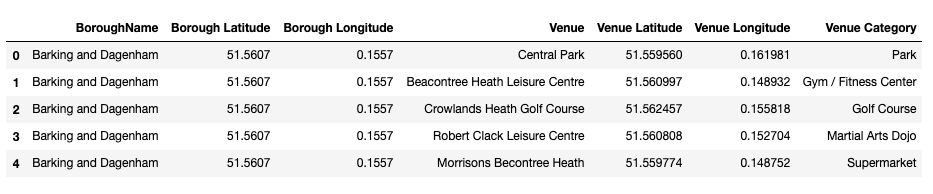


From the original data, we will only use on population and coordinates. Population can be used to calculate the ratio of reported crime to population for better comparison, and coordinates can be used to get neighborhood data from Foursquare. So the simplified data for our analysis looks as following.



**2.3 Foursquare API**

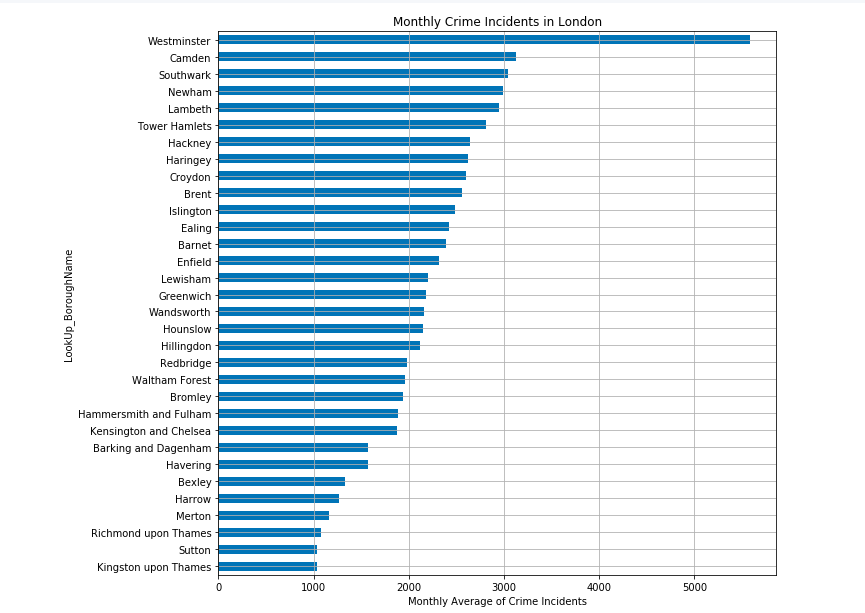
Finally, Foursquare API was used to call the top 50 popular venues in each neighborhood. This was done using the ‘explore’ function of requesting URL. We were able to acquire data looking like this.



**3. Methodology**

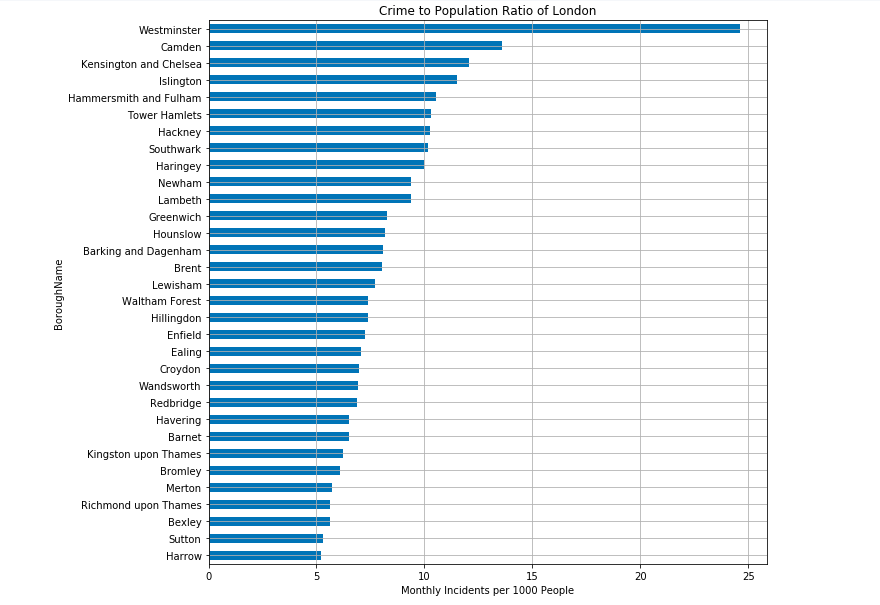
**3.1 Exploratory Analysis**

After cleansing datasets to more useful forms, we created some visualizations to interpret the data we have better.



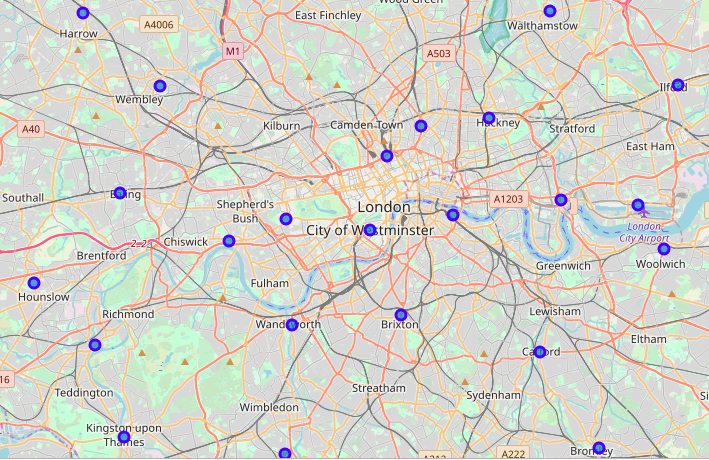
This is a bar chart displaying boroughs in descending order of monthly crime incidents. Westminster has the biggest number of reported crime, followed by Camden, Southwark and Newham.

However, as different boroughs have different sizes of population, it is not wise to directly compare the absolute number of incidents. Instead, we should consider the ratio of crime incidents to people. Thus, I have used the population to calculate the number of recorded crimes per 1000 people in each borough.



It is noticeable that Westminster and Camden still remains the top two most dangerous places in terms of recorded crime ratio to population. However, from the rank has been changed from the third borough.

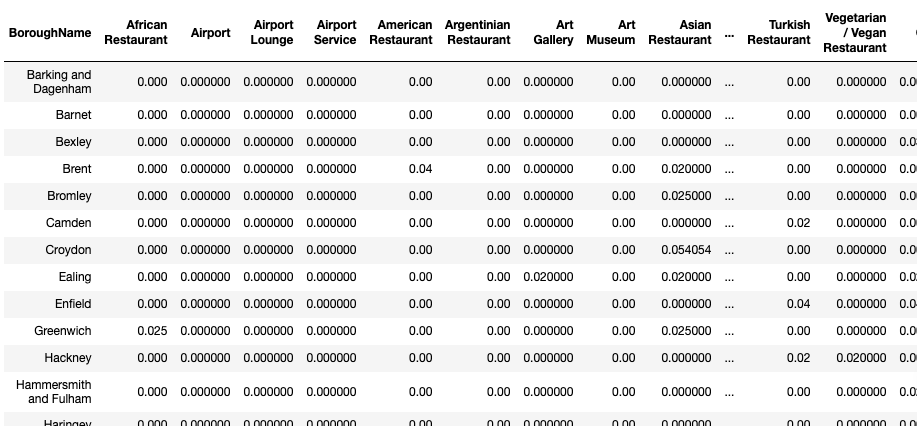
And before commencing further with the analysis, I have observed the locations of each borough to get an idea of the Greater London area.



**3.2 Cluster Analysis**

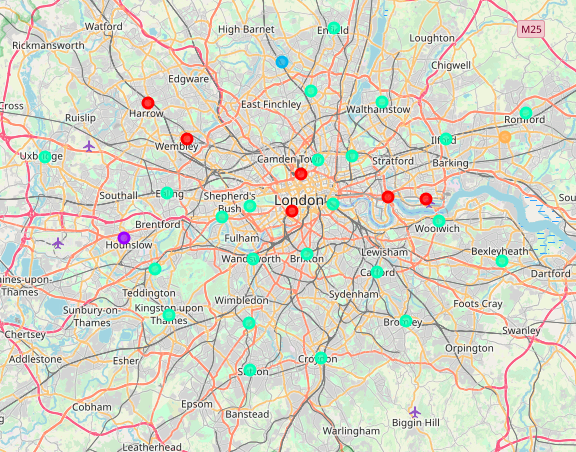
Afterwards, K-means clustering was conducted in order to group the boroughs according to what venues they have using Foursquare data, in order to feel the atmosphere of each borough.

As the first step of cluster analysis, one hot encoding was conducted to give binary values to each venue categories.



Then, the data was grouped by borough names to find out how many venues of each category exist in the boroughs within the top 50 venues. However, as some boroughs display less than 50 venues due to lack of Foursquare data, the category counts were altered to frequency of how often the category appears among others. Based on the frequency, we could attain a list of most common venue categories in each borough as follows. 

Based on the venue categories, K-means clustering was conducted to group the boroughs into 5 different clusters based on their similarity. The color dots below represent different clusters.

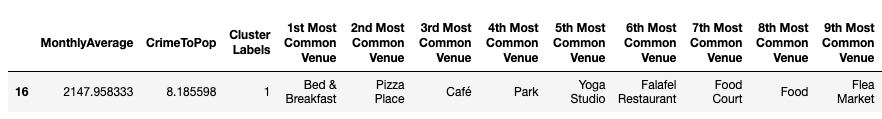


After observing each clusters and the characteristics they possess, we have given names for each clusters that best depicts their characteristics.

**Cluster 0** : Traveler area (B&B, hotel, airport...)



**Cluster 1** : Lively area (B&B, Cafe, ...)



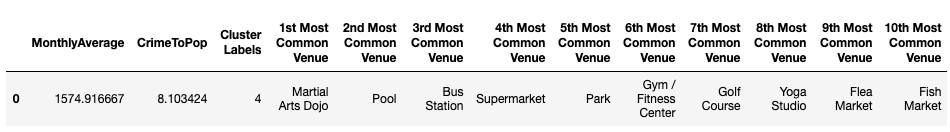
**Cluster 2** : Quiet area (Bus stop, yoga studio..)



**Cluster 3** :  Busy area (Coffee shop, clothing store...)



**Cluster 4** : Healthy area (gym, park, pool, ...)



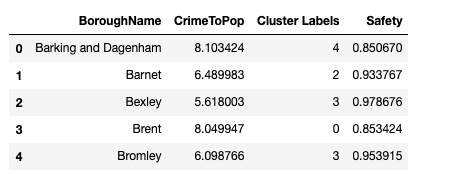
**4. Results**

Upon different analysis, we were able to discover the best neighborhoods based on our criteria of safety and atmosphere. Now we will review all the analysis made in this project before we make a conclusion on which area to live as an international student or invest as a student accommodation builder.

Like mentioned in the beginning, our key criteria of location decision will be based on safety and atmosphere.

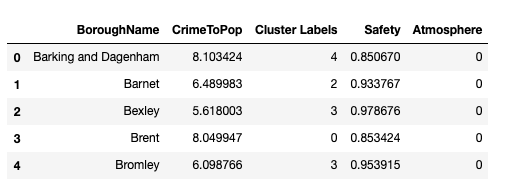
**4.1 Safety**

For safety, we normalized crime to population ratio and reversed the score so that 1 represents the neighborhood with least crime per person.

****

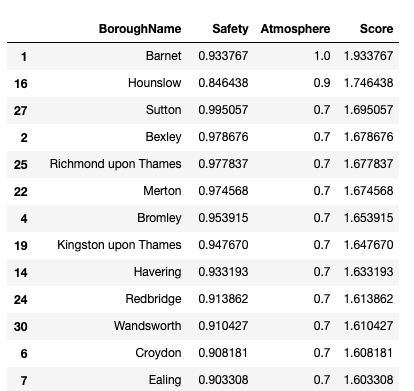
**4.2 Atmosphere**

For atmosphere, we gave an arbitrary score to each cluster based on personal preference, as preference is not easy to quantify without subjectivity. Highest score was given to Busy area (Cluster 3), which I prefer, and lowest score was given to Traveler area (Cluster 0).



**4.3 Final score**

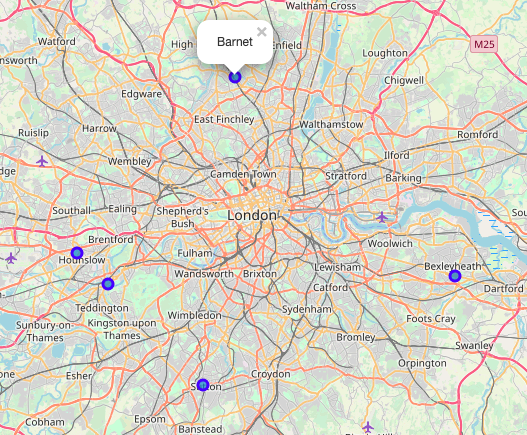
By adding the two scores, the best neighborhood scoring highest 1.93 points, Barnet.



**5. Conclusion**

**5.1 Final result of analysis**

From this analysis, we have found that the five boroughs below are the best places to build a student hall, based on safety and atmosphere of the neighborhood. The top five boroughs all belong to the Busy Area cluster, with many coffee shops and clothing stores. Therefore, what differentiates them is the safety score, which was calculated from monthly-recorded crimes per 1000 people.



**5.2 Limitations and recommendation for future study**

However, when we map the top five neighborhoods to live in, it is easily noticeable that they are all located in far out suburbs. This is due to many limitations this research holds.

Among numerous factors that determine a good neighborhood, we only took into consideration what type of venues are popular and how many crime incidents are recorded for the sake of simplification. This means that serious crimes like homicide was treated the same as a comparatively petty crime like shoplifting. Moreover, the number of stores in the neighborhood may be as important as what type of stores there are.

To overcome the limitations of this study, we will need further data such as distance to city center, housing prices or ratio of the number of stores to population. Also, taking crime categories into factor and weighting them differently may be helpful.

Despite some limitations, this research was still enjoyable in that we were able to explore the neighborhoods in depth.

**References**

* “London Recorded Crime: Geographic Breakdown”, London Datastore
* “List of London Boroughs”, Wikipedia
* Foursquare API
* “The Economic Impact of London’s International Studnets”, London & Partners (2018)
* Lecture notes from IBM Professional Data Science Specialization, Coursera