Institution of Choice in London

Utt Assoratgoon

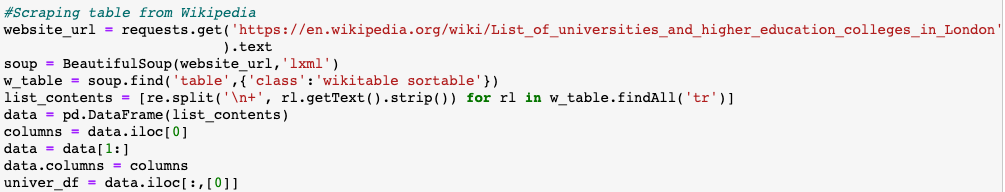
1. **Introduction**
   1. **Background**

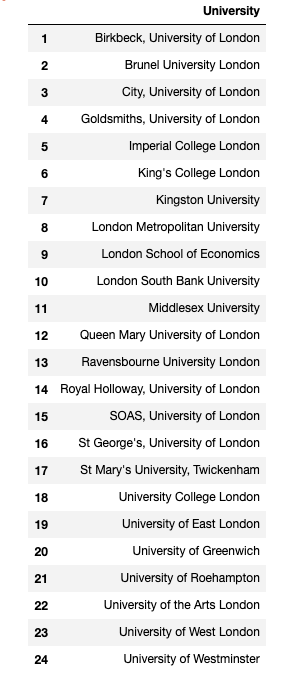
The UK has become one of the top destinations to study around the world. Not only because degrees and qualifications from UK higher education institutions are known around the world as high quality and world-class but studying in the UK also take less time to complete than other counties. Because of cultural diversity, students who study in the UK will meet new people and experience a variety of cultures along with gaining knowledge from the university. However, according to The Sun (A tabloid newspaper published in the United Kingdom), the total number of London murders, excluding victims of terrorism, has shot up by 38 percent since 2014. Therefore, safety should be one of the concerns for students who decide to join the university in London. In this project, we will walk you through an analyzing process that includes data exploration, data preparation, and analysis section to help you decide which institution is fit for your needs base on criminal records and venues around each institution in London.



1. **Data preparation**
   1. **Web-Scrapping**
      1. **Get the names of institution in London from Wikipedia**

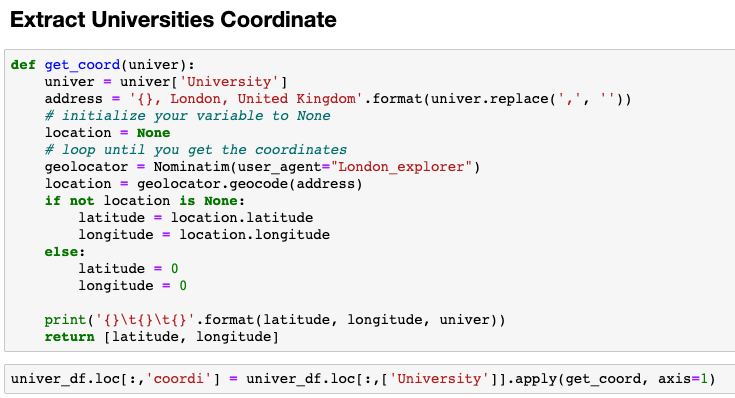
I have scrapped a sorted table from Wikipedia page, [List of universities and higher education colleges in London](https://en.wikipedia.org/wiki/List_of_universities_and_higher_education_colleges_in_London), which contains the names of 25 institutions in London, using BeautifulSoup and Pandas library to create an initial data frame. For the cleaning process, after extracting names from a scrapped table I have used .strp() method to trim each element in the data frame to make certain that there is not any whitespace on both ends of each name.

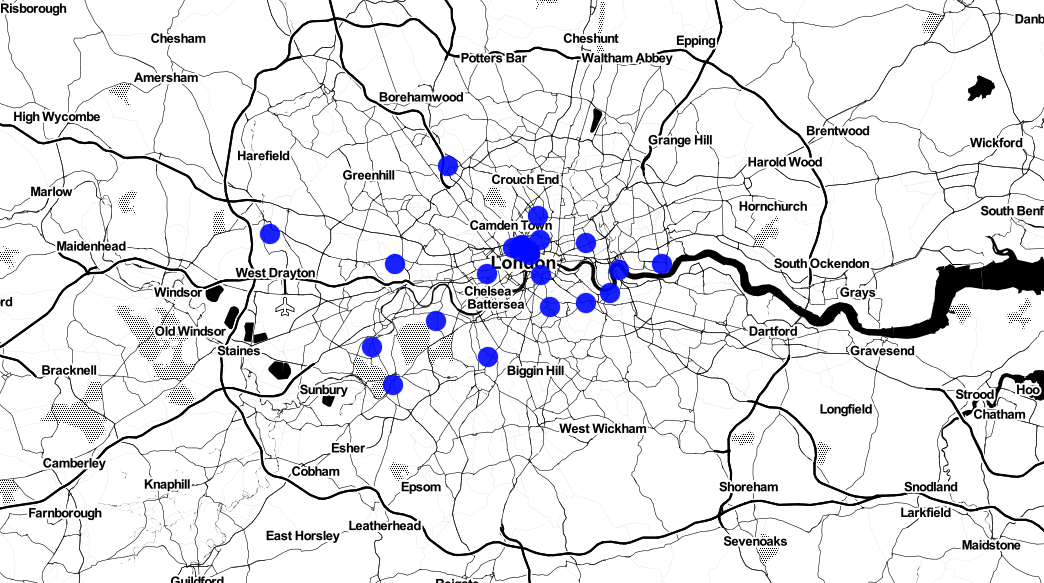




* + 1. **Get coordinates of the institution**

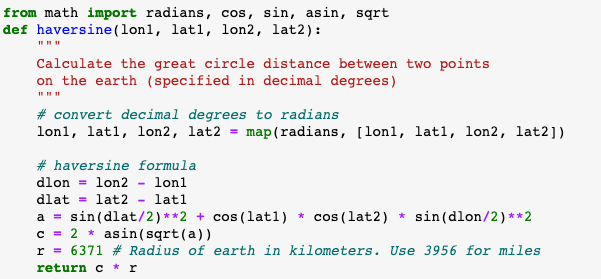
The Next step is to get coordinates of these 25 institutions by using the Geopy library. But there are two institutions coordinates that Geopy cannot found the correct location, so I had to manually search in Google Map and replace the wrong coordinates with the right one using Pandas library.





* 1. **London crime records from Kaggle**

Download [London crime records dataset](https://www.kaggle.com/sohier/london-police-records#london-street.csv) from Kaggle, an online community of data scientists and machine learners, owned by Google LLC. The dataset contains crime records from 2014 through mid 2017 of London. After, criminal records are loaded into the data frame. I have filtered just only crimes that occur within a radius of 1 kilometer around each institution. I used the Haversine function to calculate a distance between two points on earth which is shown below.

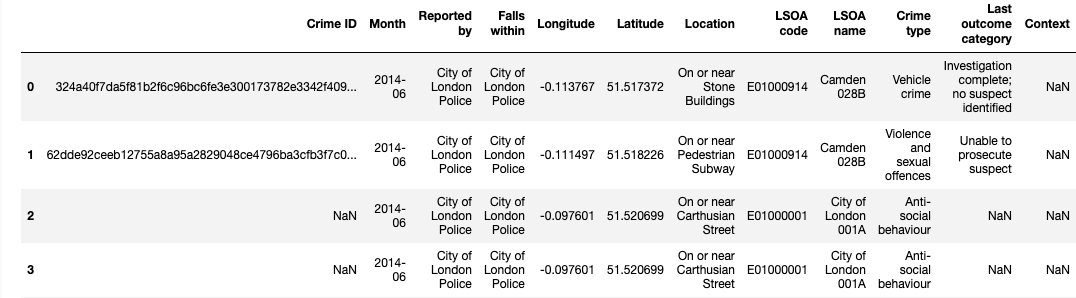


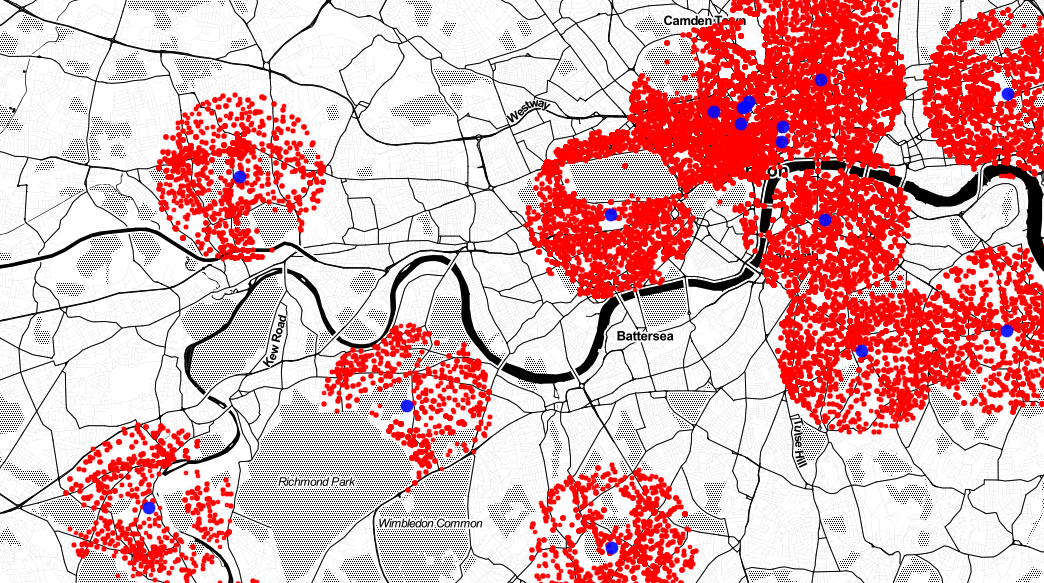
* 1. **Foursquare API data**

I have used Foursquare API to obtain list of venues within 1 kilometer around each institution.

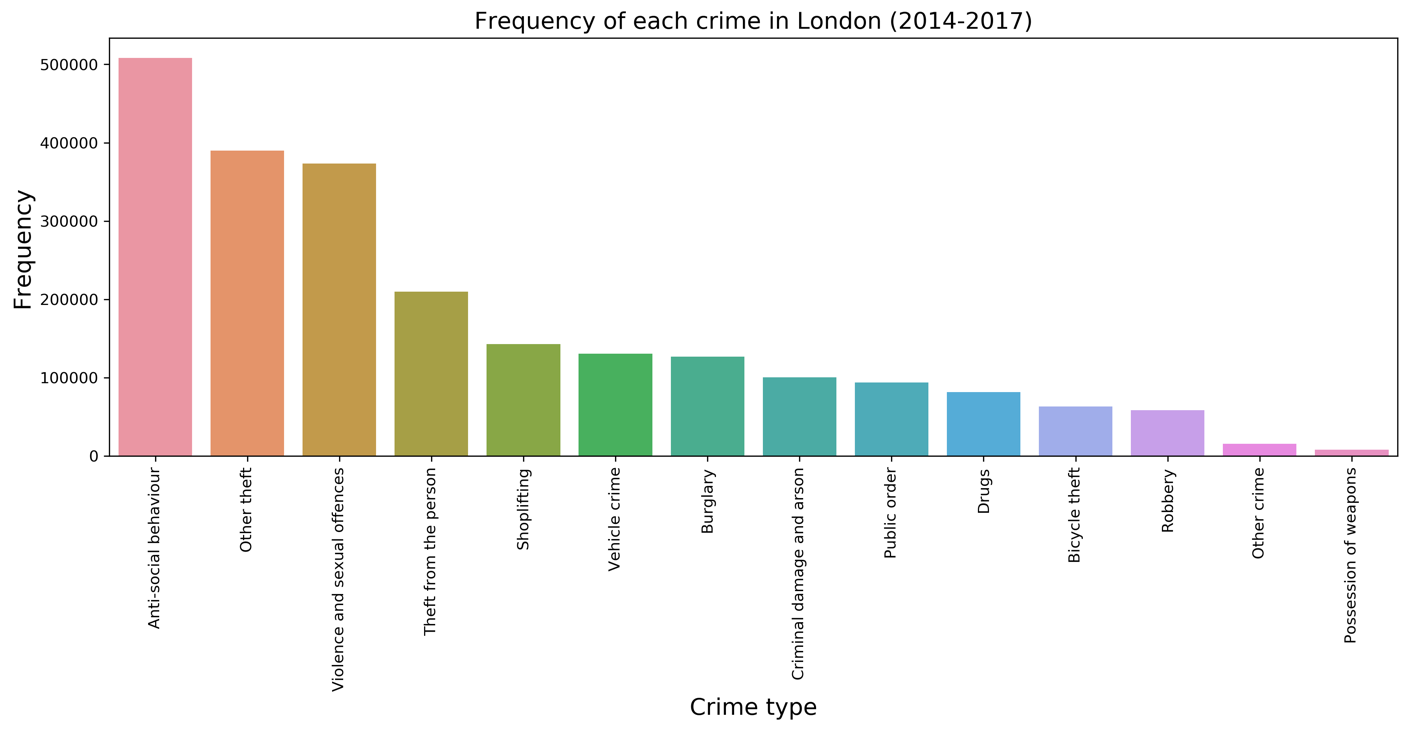
1. **Data Exploration and Visualization**
   1. **Crime records dataset exploration**

I have used the Folium library to plot a map of institutions and crimes, in which blue represents the location of institutions and red represents the location where crimes occurred. But because there are too many records in the dataset. I had to select only crimes that occur in 2017 for a clearer appearance when it appears on the map as shown below.



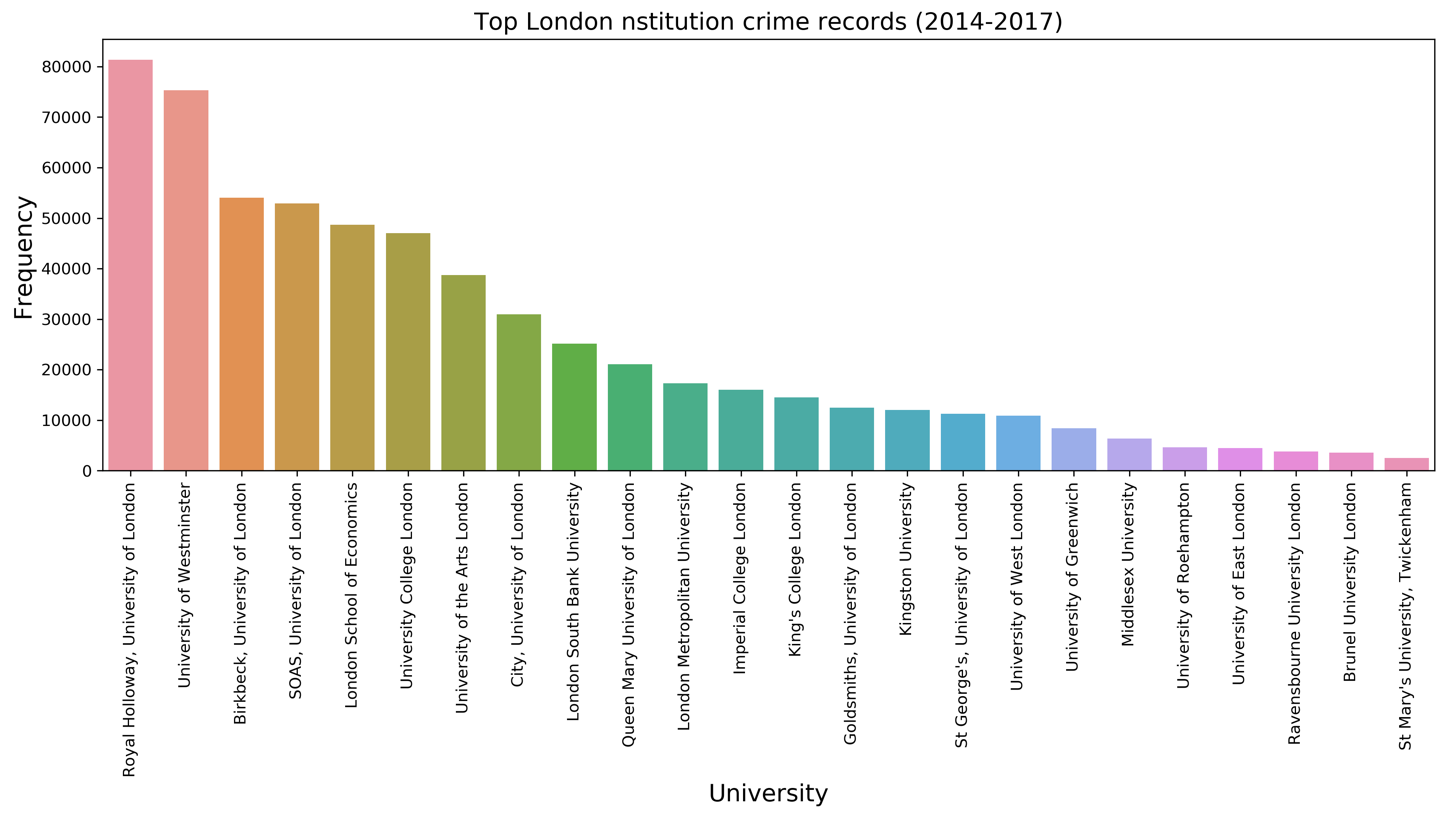


Next, building a bar chart of crime type.

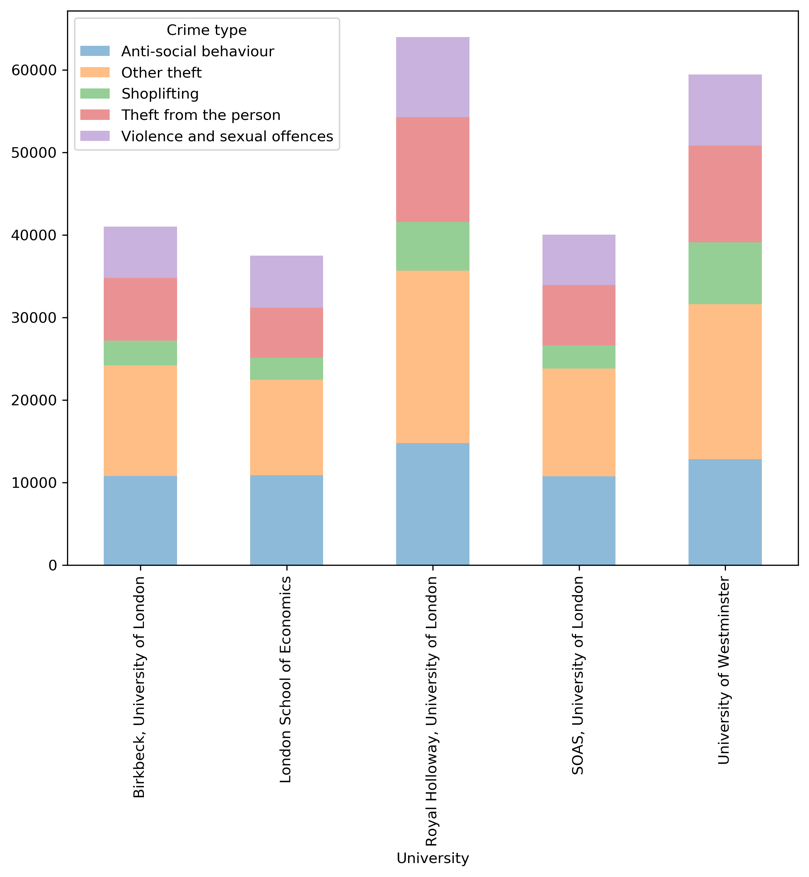


We found out that there is a large number of **Anti-social behavior** compare to other crime.

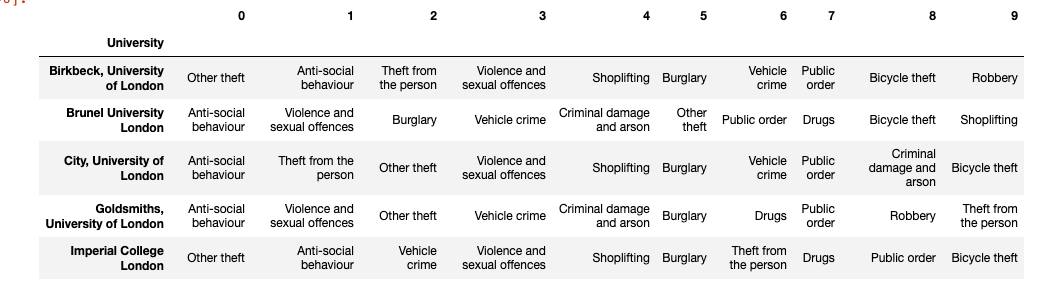
Next step, I also wanted to explore which institution has the highest number of crimes, so I build this bar chart to answer the question.



Now, let’s plot top 5 crimes against top 5 institutions.

****

For the next step, it was to gain more information about crime around each institution. First, using pandas one-hot encoding on the ‘Crime type’ column to convert from the categorical variable into a numeric variable. Second, groping each institution to obtain an average of each crime type using pandas Groupby on the ‘Crime type’ column. Finally, use the average that we obtain in the previous step to create a sorted data-frame of institutions and crimes.

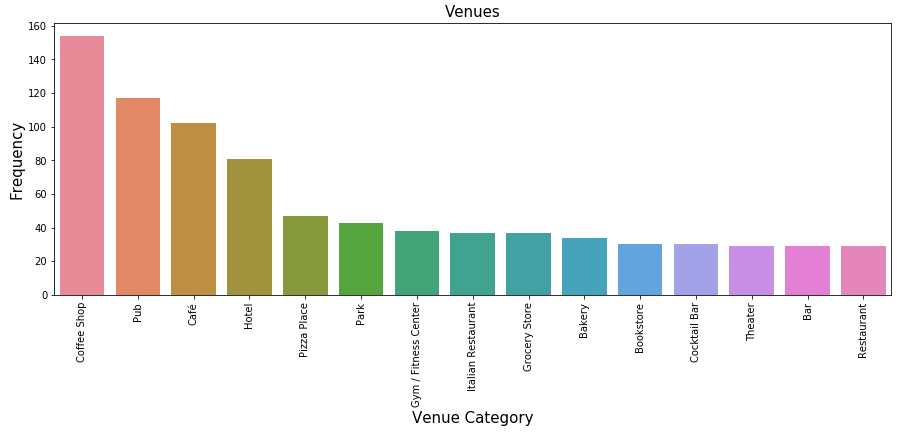
****

* 1. **Venues exploration**

It’s time to explore venues. First, I used Foursquare API to obtain information of popular venues around each institution.



Then let’s see, which category of venues mostly appear around institutions? We can find the answer from this plot below.



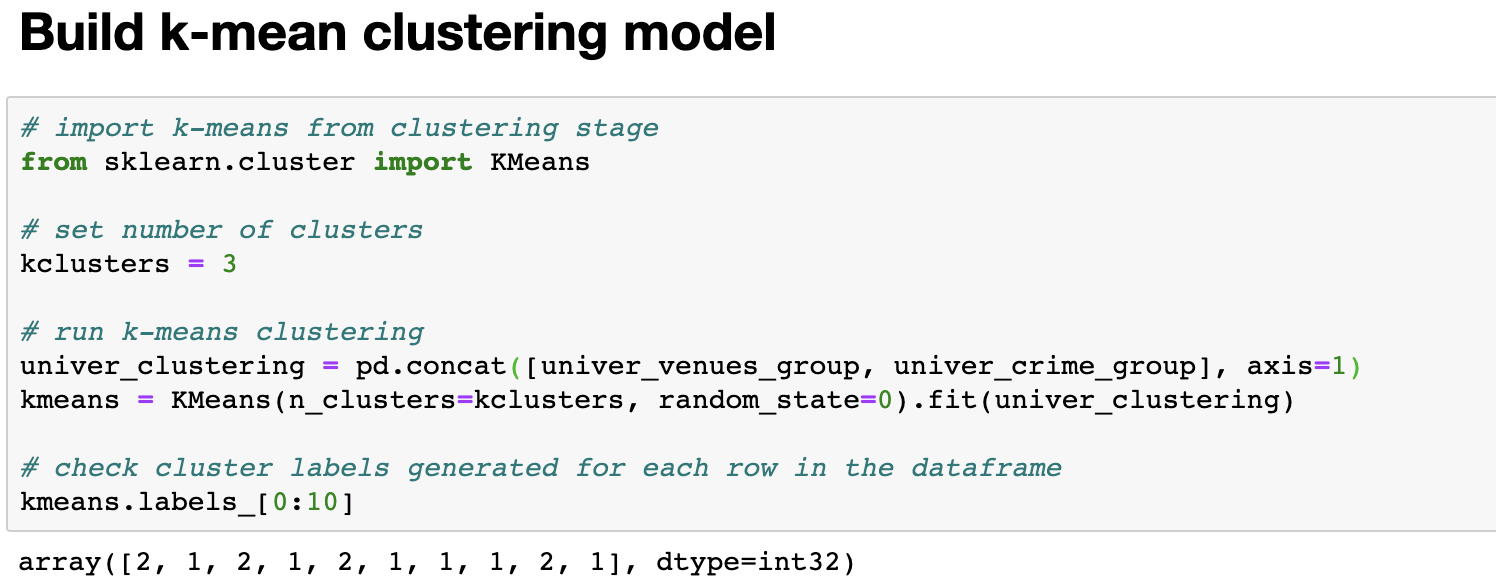
It seems that a **coffee shop** is very popular around an institution, perhaps students enjoy studying there more than their home.

Next, let’s do the same process that we did on criminal records to rank venues around each institution.

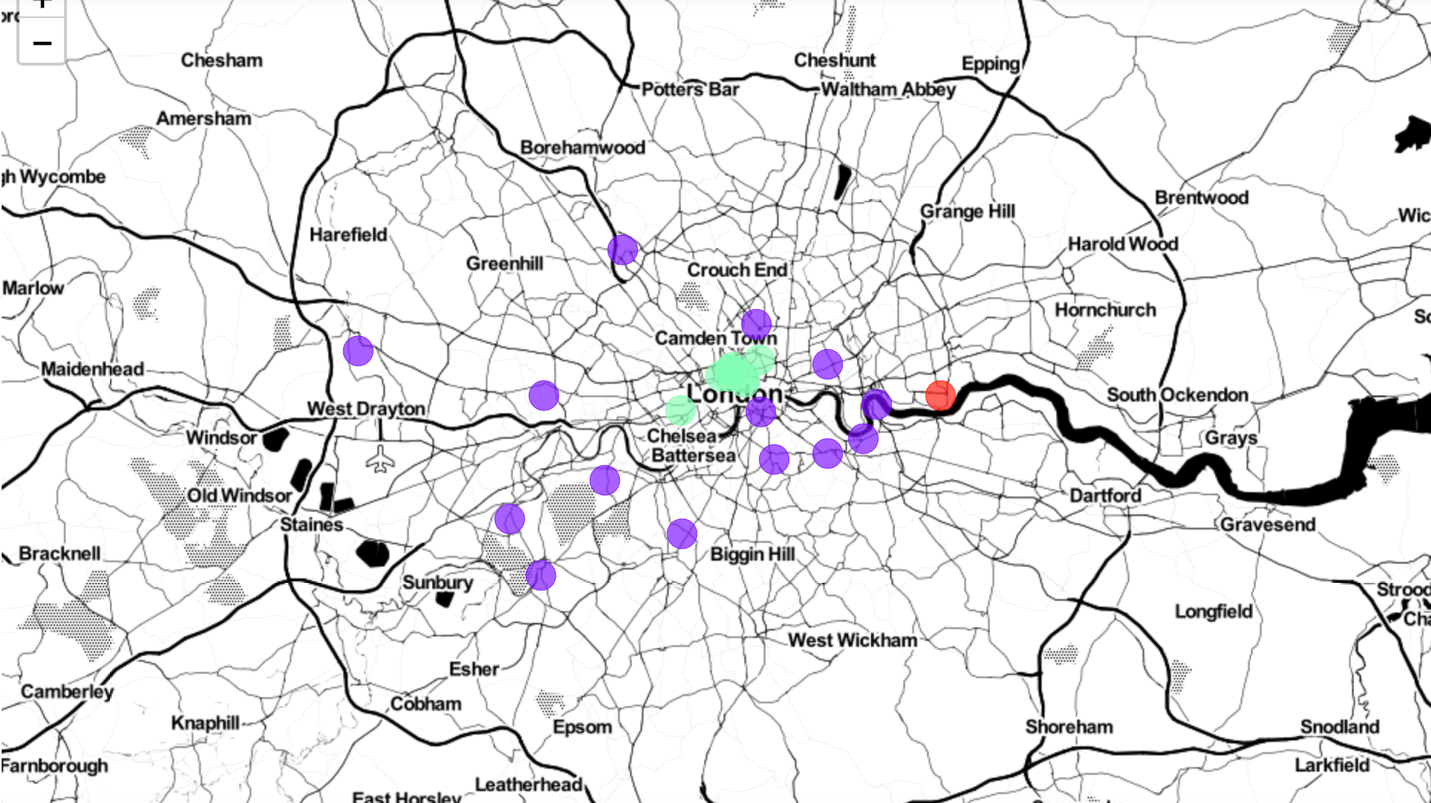


1. **Clustering**

Finally, I have used K-mean clustering to cluster these top institutions in London based on similarities of nearby venues and crime records around each institution. I have used code snippet below.



Then, visualize the result on London map using Folium library.



1. **Results**

After all the analysis processes that we did. Finally, we got the summarization of the result below.

* For the purple cluster which consist of

1. Brunel University London
2. Goldsmiths, University of London
3. King's College London
4. Kingston University
5. London Metropolitan University
6. London South Bank University
7. Middlesex University
8. Queen Mary University of London
9. Ravensbourne University London
10. St George's, University of London
11. St Mary's University, Twickenham
12. University of Greenwich
13. University of Roehampton
14. University of West London

Mostly they are in the suburb area with a moderate number of crimes. But most of the crimes that occurred around these institutions are **Anti-social behavior** and **Violence and sexual offenses** which may relate to a high number of the **pub** in this cluster.

* The green cluster consist of

1. Birkbeck, University of London
2. City, University of London
3. Imperial College London
4. London School of Economics
5. Royal Holloway, University of London
6. SOAS, University of London
7. University College London
8. University of Westminster
9. University of the Arts London

Institutions in this cluster are located in the heart of London’s downtown. Unsurprisingly, this cluster has a very high rate of crime compared to others. Besides, London is a famous destination for tourists and tourists are the main target of criminals. No doubt, the majority of crime that occurred near these institutions is **theft**. However, there are a lot of **cafes** in this area for students to study and relax.

* For the last red cluster, it consists of only one institution which is the University of East London. The University of East London has a very low rate of crime and instead of surrounding by coffee shops or pubs, it's surrounded by a **park**, **gym**, and **fitness center**.