Institution of Choice in London

Utt Assoratgoon

1. Introduction

1.1. Background

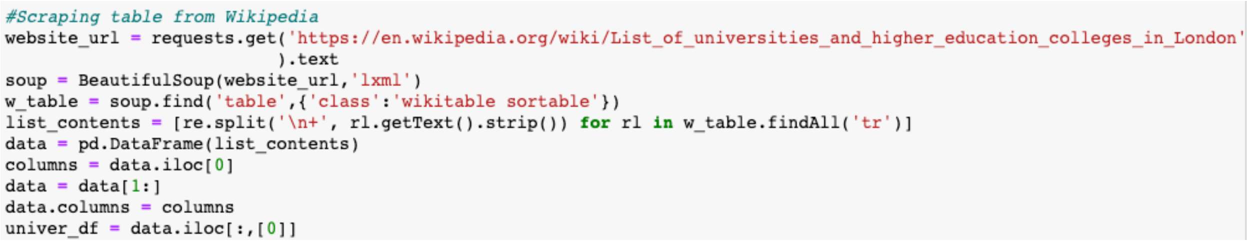
The UK has become one of the top destinations to study around the world. Not only because degrees and qualifications from UK higher education institutions are known around the world as high quality and world-class but studying in the UK also take less time to complete than other counties. Because of 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.

2. Data preparation

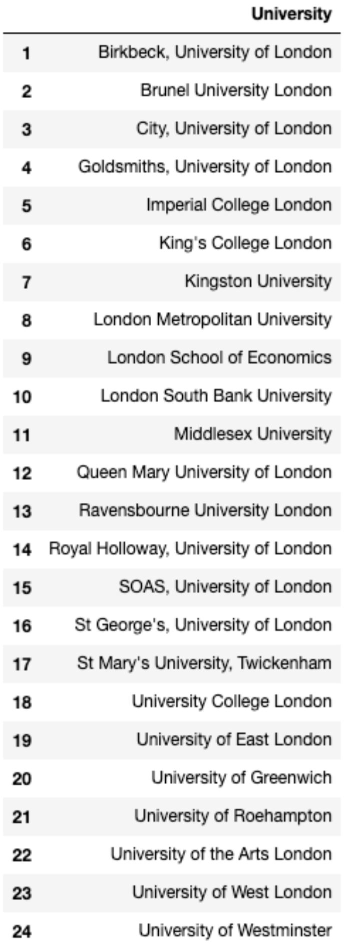
2.1. Web-Scrapping

2.1.1. 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, 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.



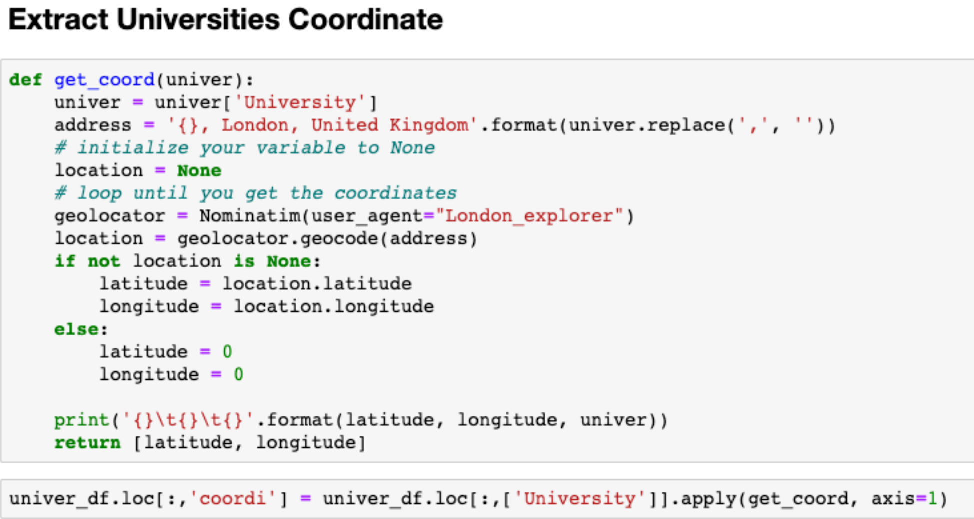
```
#Scraping table from Wikipedia
website_url = requests.get('https://en.wikipedia.org/wiki/List_of_universities_and_higher_education_colleges_in_London'
                          ).text
soup = BeautifulSoup(website_url,'lxml')
w_table = soup.find('table',{'class':'wikitable sortable'})
list_contents = [re.split('\n+', rl.getText().strip()) for rl in w_table.findAll('tr')]
data = pd.DataFrame(list_contents)
columns = data.iloc[0]
data = data[1:]
data.columns = columns
univer_df = data.iloc[:,[0]]
```



	University
1	Birkbeck, University of London
2	Brunel University London
3	City, University of London
4	Goldsmiths, University of London
5	Imperial College London
6	King's College London
7	Kingston University
8	London Metropolitan University
9	London School of Economics
10	London South Bank University
11	Middlesex University
12	Queen Mary University of London
13	Ravensbourne University London
14	Royal Holloway, University of London
15	SOAS, University of London
16	St George's, University of London
17	St Mary's University, Twickenham
18	University College London
19	University of East London
20	University of Greenwich
21	University of Roehampton
22	University of the Arts London
23	University of West London
24	University of Westminster

2.1.2. Get coordinates of the institution

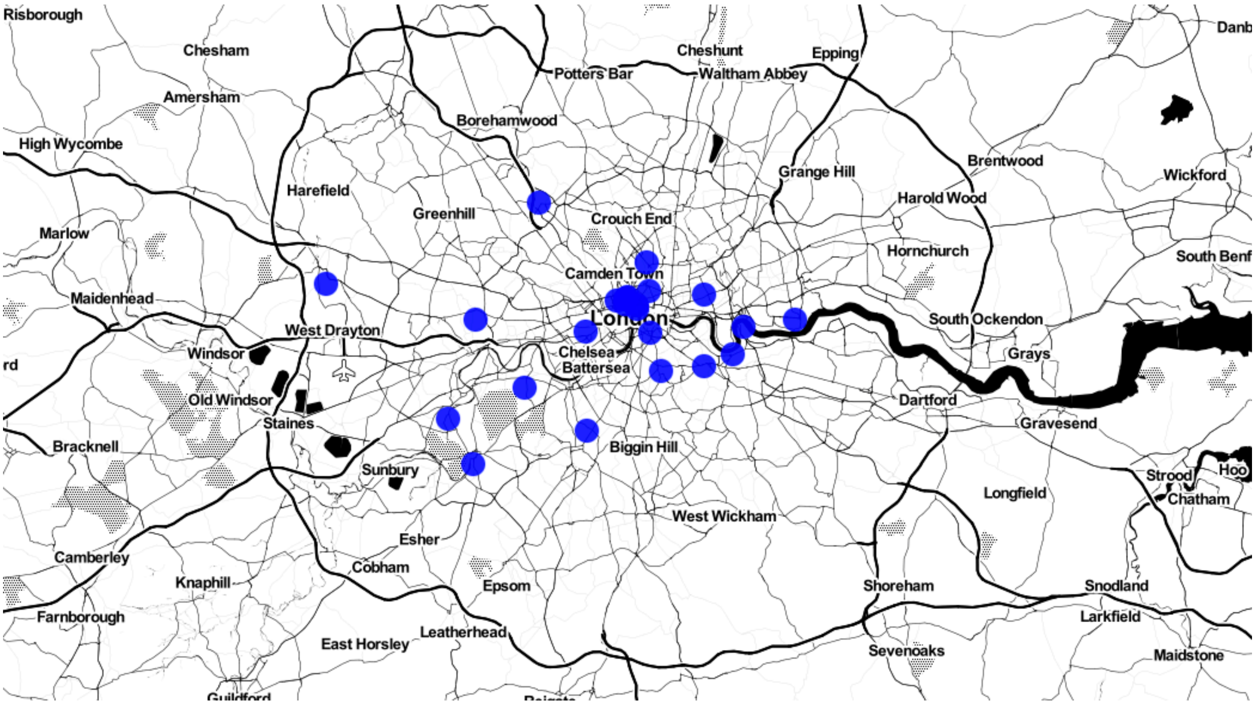
The Next step is to get coordinates of these 25 institutions by using the Geopy library. But there are two institutions coordinates that Geopy cannot found the correct location, so I had to manually search in Google Map and replace the wrong coordinates with the right one using Pandas library.



Extract Universities Coordinate

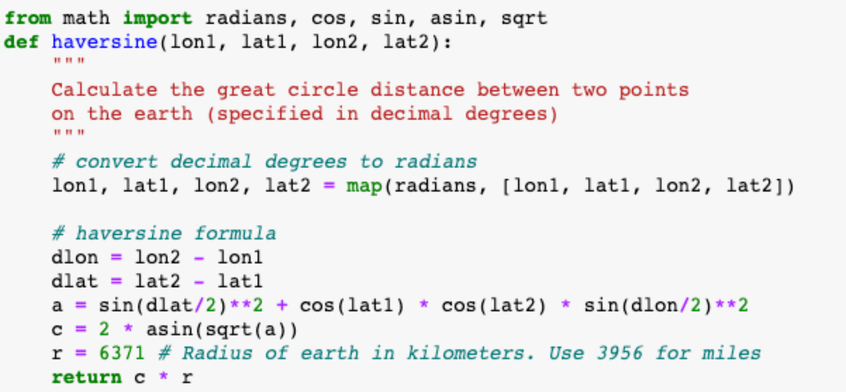
```
def get_coord(univer):
    univer = univer['University']
    address = '{}, London, United Kingdom'.format(univer.replace(',', ''))
    # initialize your variable to None
    location = None
    # loop until you get the coordinates
    geolocator = Nominatim(user_agent="London_explorer")
    location = geolocator.geocode(address)
    if not location is None:
        latitude = location.latitude
        longitude = location.longitude
    else:
        latitude = 0
        longitude = 0

    print('{}\t{}\t{}'.format(latitude, longitude, univer))
    return [latitude, longitude]
```

```
univer_df.loc[:,'coordi'] = univer_df.loc[:,['University']].apply(get_coord, axis=1)
```



2.2. London crime records from Kaggle

Download London crime records dataset 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.



```
from math import radians, cos, sin, asin, sqrt
def haversine(lon1, lat1, lon2, lat2):
    """
    Calculate the great circle distance between two points
    on the earth (specified in decimal degrees)
    """
    # convert decimal degrees to radians
    lon1, lat1, lon2, lat2 = map(radians, [lon1, lat1, lon2, lat2])

    # haversine formula
    dlon = lon2 - lon1
    dlat = lat2 - lat1
    a = sin(dlat/2)**2 + cos(lat1) * cos(lat2) * sin(dlon/2)**2
    c = 2 * asin(sqrt(a))
    r = 6371 # Radius of earth in kilometers. Use 3956 for miles
    return c * r
```

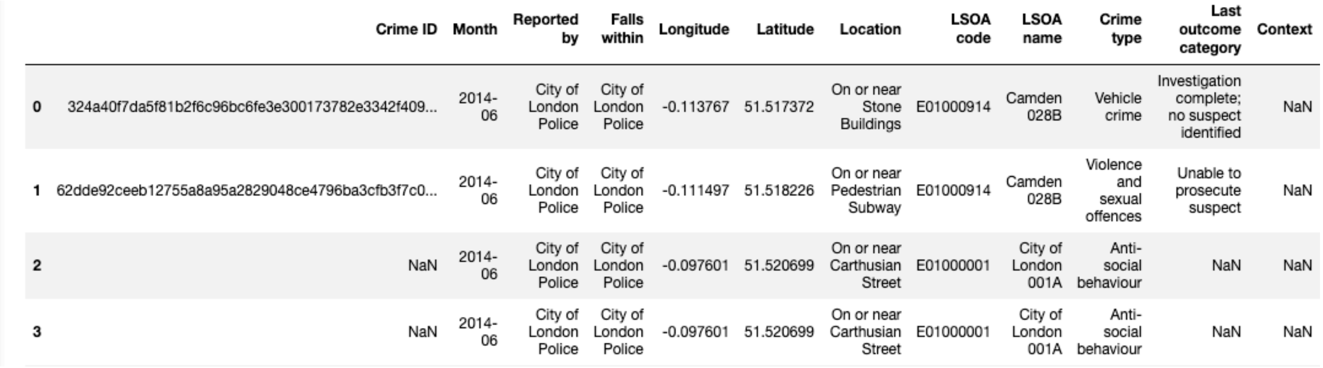
2.3. Foursquare API data

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

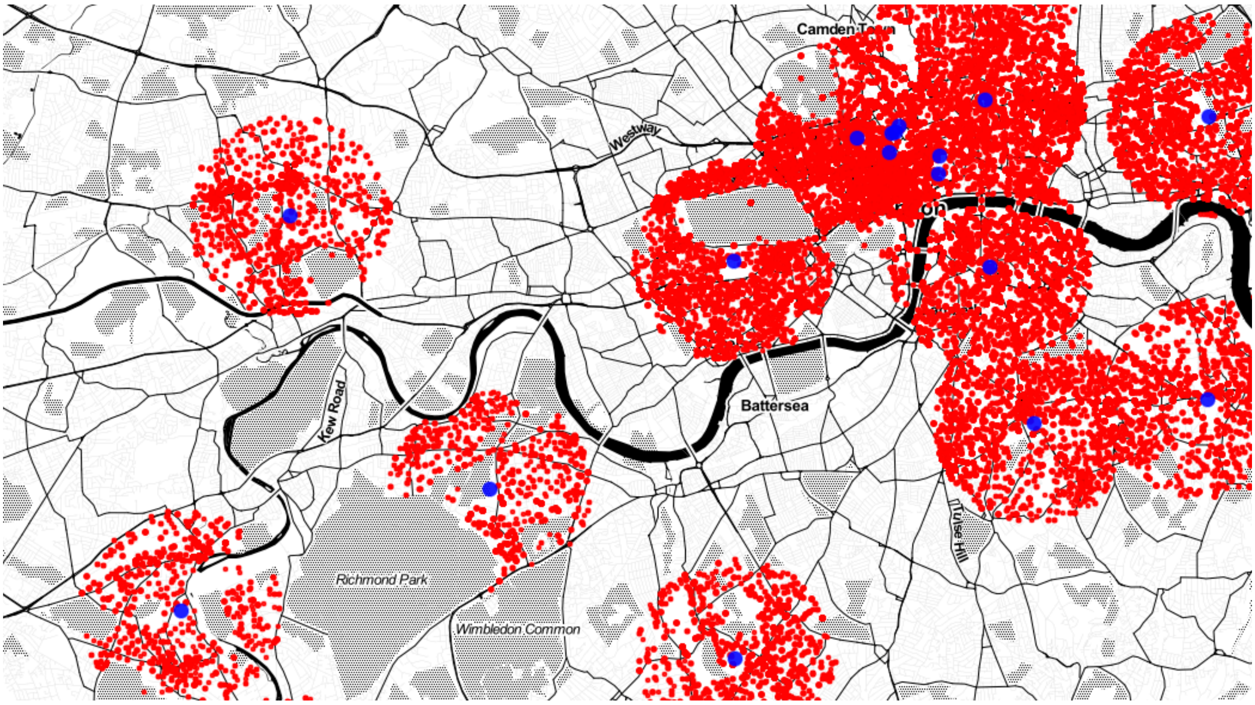
3. Data Exploration and Visualization

3.1. Crime records dataset exploration

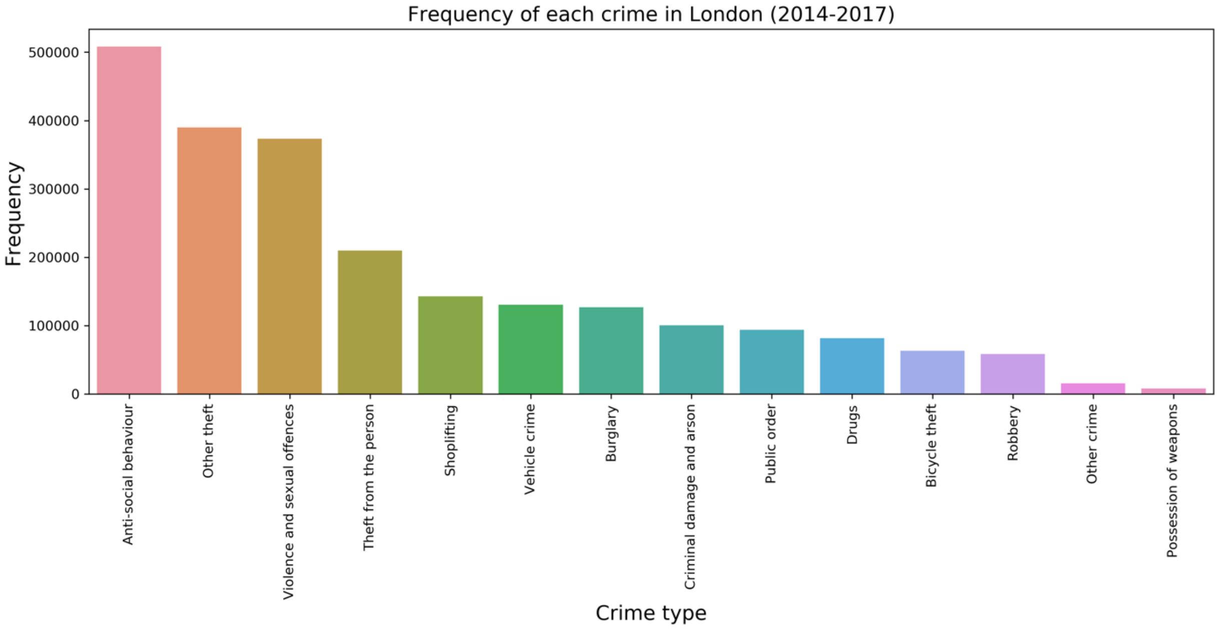
I have used the Folium library to plot a map of institutions and crimes, in which blue represents the location of institutions and red represents the location where crimes occurred. But because there are too many records in the dataset. I had to select only crimes that occur in 2017 for a clearer appearance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



	Crime ID	Month	Reported by	Falls within	Longitude	Latitude	Location	LSOA code	LSOA name	Crime type	Last outcome category	Context
0	324a40f7da5f81b2f6c96bc6fe3e300173782e3342f409...	2014-06	City of London Police	City of London Police	-0.113767	51.517372	On or near Stone Buildings	E01000914	Camden 028B	Vehicle crime	Investigation complete; no suspect identified	NaN
1	62dde92ceeb12755a8a95a2829048ce4796ba3cfb3f7c0...	2014-06	City of London Police	City of London Police	-0.111497	51.518226	On or near Pedestrian Subway	E01000914	Camden 028B	Violence and sexual offences	Unable to prosecute suspect	NaN
2	NaN	2014-06	City of London Police	City of London Police	-0.097601	51.520699	On or near Carthusian Street	E01000001	City of London 001A	Anti-social behaviour	NaN	NaN
3	NaN	2014-06	City of London Police	City of London Police	-0.097601	51.520699	On or near Carthusian Street	E01000001	City of London 001A	Anti-social behaviour	NaN	NaN

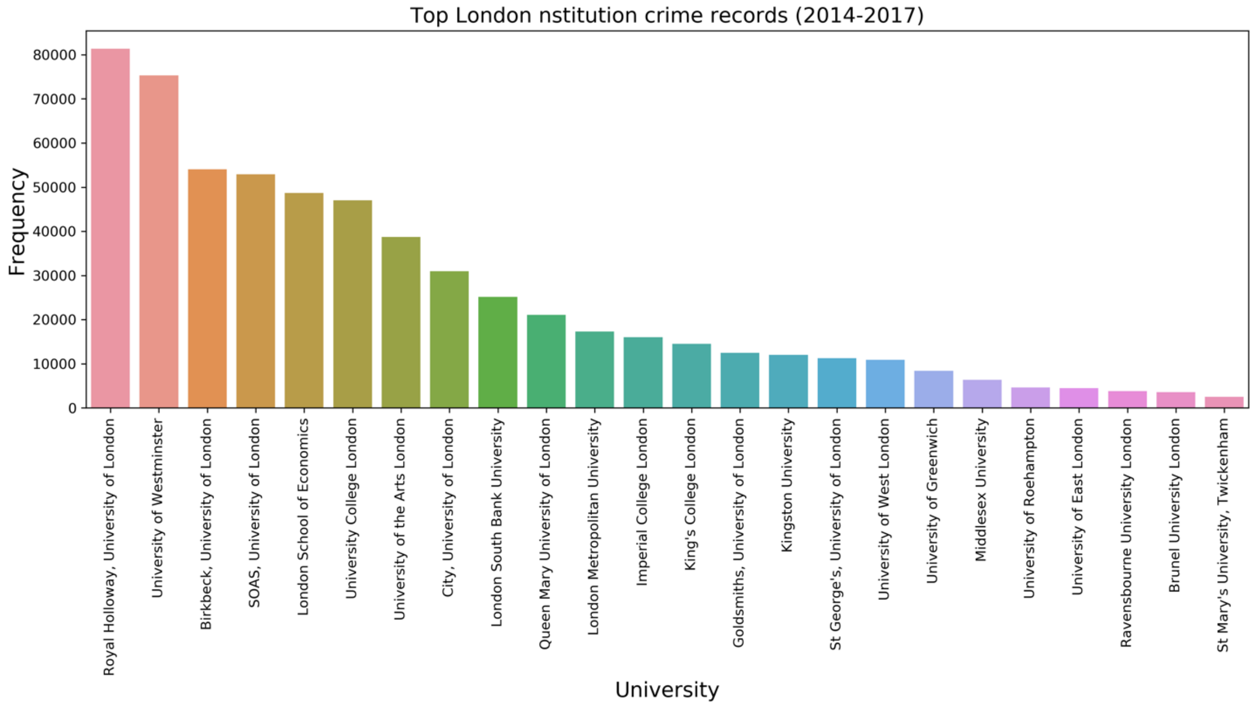


Next, building a bar chart of crime type.

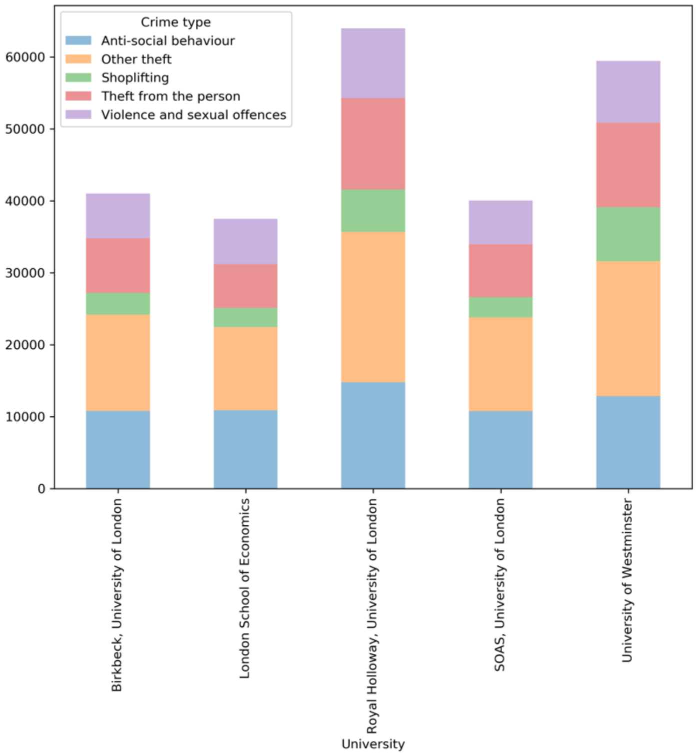


We found out that there is a large number of **Anti-social behavior** compare to other crime.

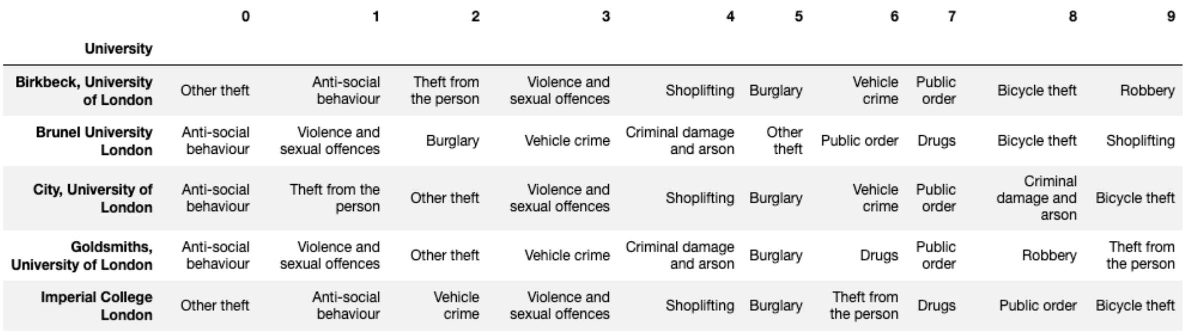
Next step, I also wanted to explore which institution has the highest number of crimes, so I build this bar chart to answer the question.



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



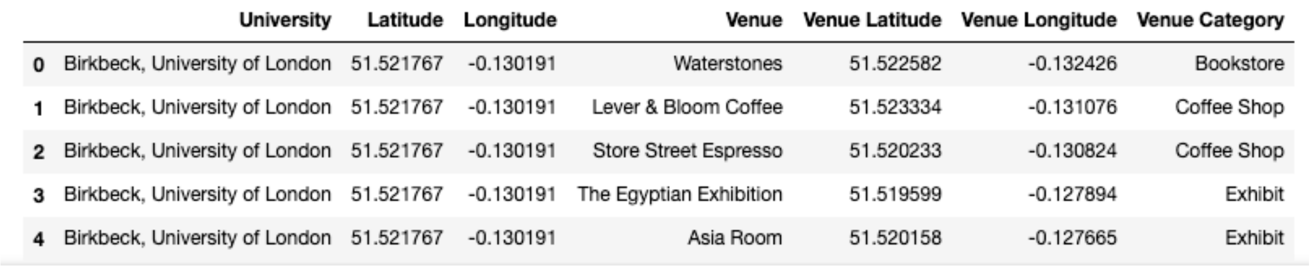
For the next step, it was to gain more information about crime around each institution. First, using pandas one-hot encoding on the ‘Crime type’ column to convert from the categorical variable into a numeric variable. Second, groping each institution to obtain an average of each crime type using pandas 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.



	0	1	2	3	4	5	6	7	8	9
University										
Birkbeck, University of London	Other theft	Anti-social behaviour	Theft from the person	Violence and sexual offences	Shoplifting	Burglary	Vehicle crime	Public order	Bicycle theft	Robbery
Brunel University London	Anti-social behaviour	Violence and sexual offences	Burglary	Vehicle crime	Criminal damage and arson	Other theft	Public order	Drugs	Bicycle theft	Shoplifting
City, University of London	Anti-social behaviour	Theft from the person	Other theft	Violence and sexual offences	Shoplifting	Burglary	Vehicle crime	Public order	Criminal damage and arson	Bicycle theft
Goldsmiths, University of London	Anti-social behaviour	Violence and sexual offences	Other theft	Vehicle crime	Criminal damage and arson	Burglary	Drugs	Public order	Robbery	Theft from the person
Imperial College London	Other theft	Anti-social behaviour	Vehicle crime	Violence and sexual offences	Shoplifting	Burglary	Theft from the person	Drugs	Public order	Bicycle theft

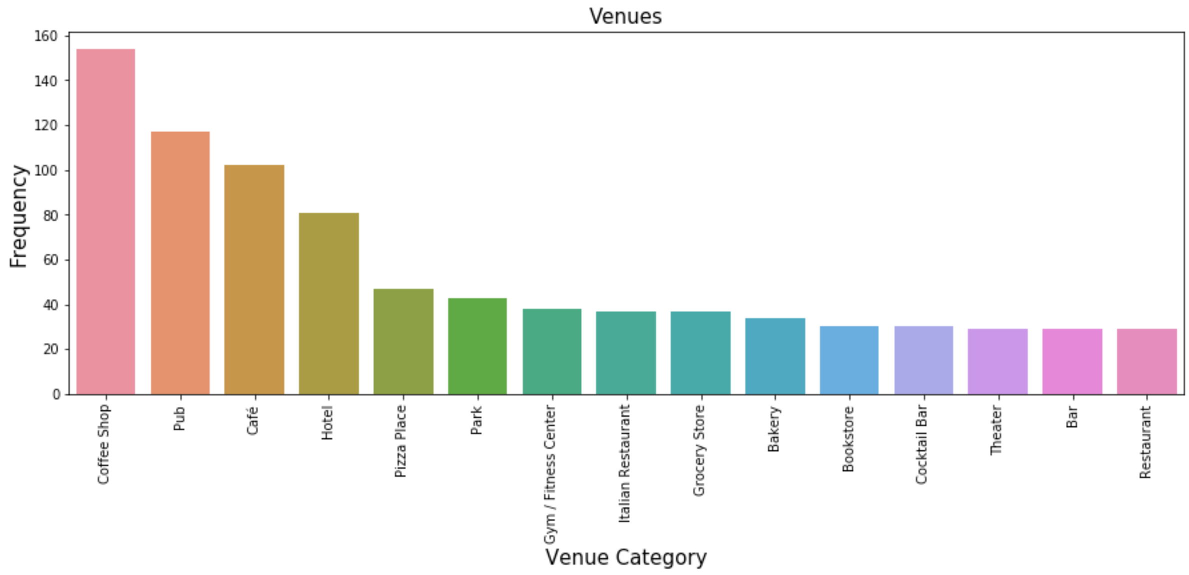
3.2. Venues exploration

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



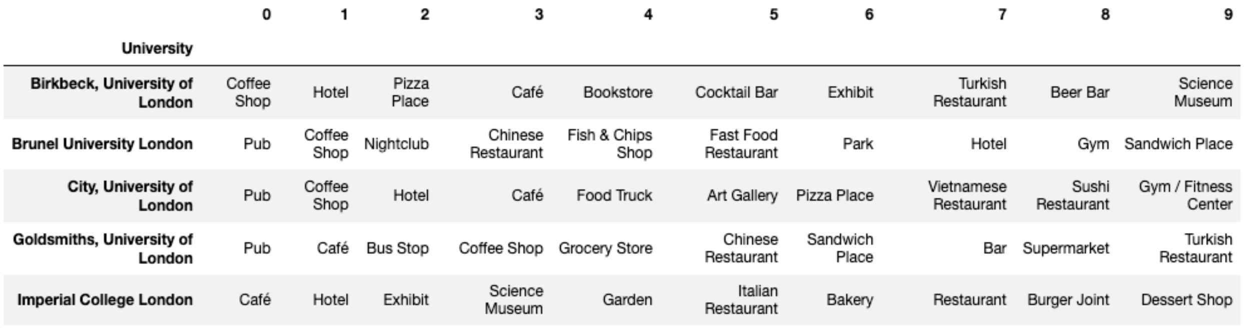
	University	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Birkbeck, University of London	51.521767	-0.130191	Waterstones	51.522582	-0.132426	Bookstore
1	Birkbeck, University of London	51.521767	-0.130191	Lever & Bloom Coffee	51.523334	-0.131076	Coffee Shop
2	Birkbeck, University of London	51.521767	-0.130191	Store Street Espresso	51.520233	-0.130824	Coffee Shop
3	Birkbeck, University of London	51.521767	-0.130191	The Egyptian Exhibition	51.519599	-0.127894	Exhibit
4	Birkbeck, University of London	51.521767	-0.130191	Asia Room	51.520158	-0.127665	Exhibit

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.

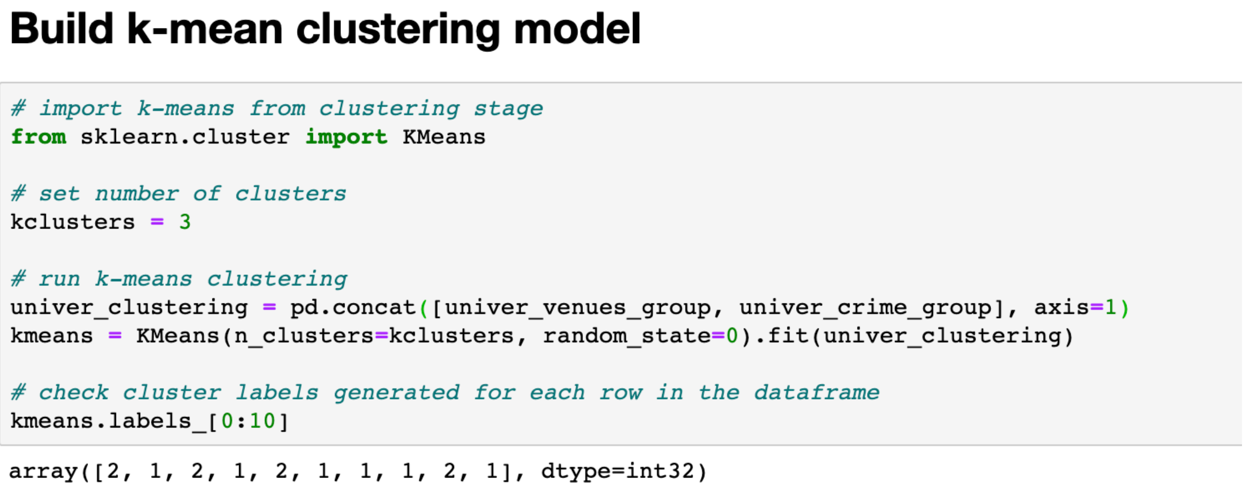
Next, let’s do the same process that we did on criminal records to rank venues around each institution.



University	0	1	2	3	4	5	6	7	8	9
Birkbeck, University of London	Coffee Shop	Hotel	Pizza Place	Café	Bookstore	Cocktail Bar	Exhibit	Turkish Restaurant	Beer Bar	Science Museum
Brunel University London	Pub	Coffee Shop	Nightclub	Chinese Restaurant	Fish & Chips Shop	Fast Food Restaurant	Park	Hotel	Gym	Sandwich Place
City, University of London	Pub	Coffee Shop	Hotel	Café	Food Truck	Art Gallery	Pizza Place	Vietnamese Restaurant	Sushi Restaurant	Gym / Fitness Center
Goldsmiths, University of London	Pub	Café	Bus Stop	Coffee Shop	Grocery Store	Chinese Restaurant	Sandwich Place	Bar	Supermarket	Turkish Restaurant
Imperial College London	Café	Hotel	Exhibit	Science Museum	Garden	Italian Restaurant	Bakery	Restaurant	Burger Joint	Dessert Shop

4. 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.

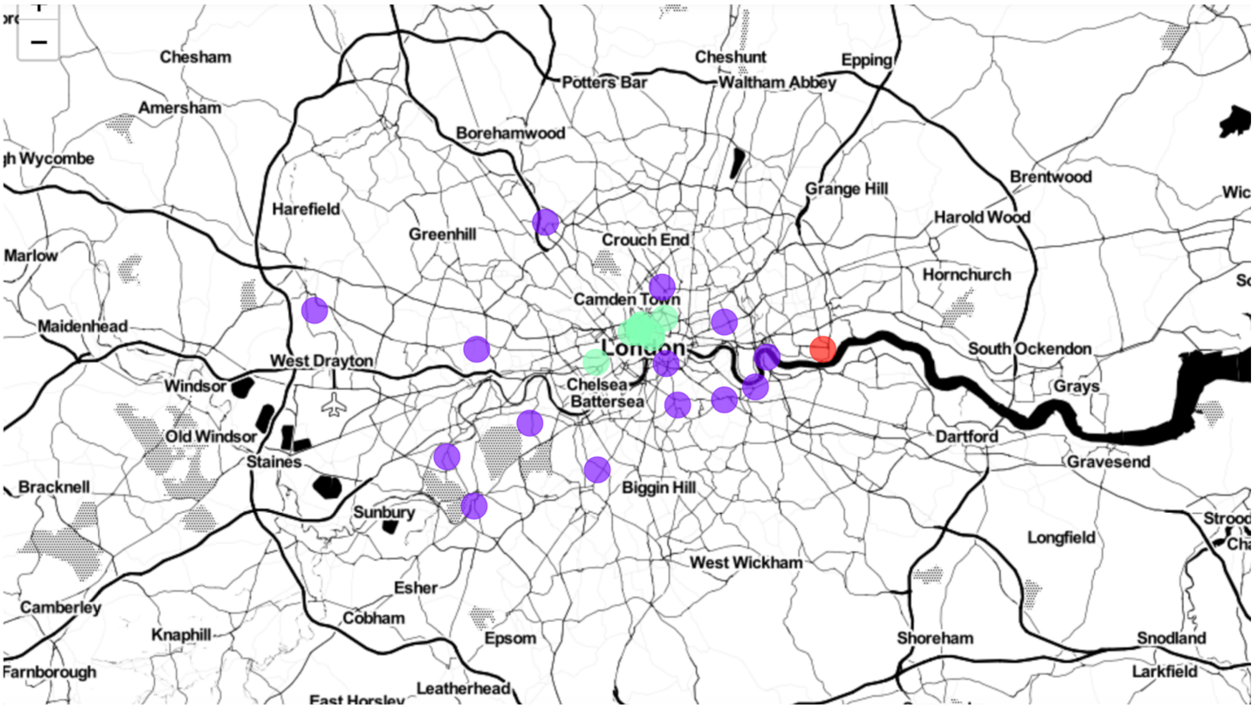


Build k-mean clustering model

```
# import k-means from clustering stage
from sklearn.cluster import KMeans

# set number of clusters
kclusters = 3

# run k-means clustering
univer_clustering = pd.concat([univer_venues_group, univer_crime_group], axis=1)
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(univer_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

```
array([2, 1, 2, 1, 2, 1, 1, 1, 2, 1], dtype=int32)
```

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



5. 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**.