Business Problem

In the United Kingdom there are many places to live, many of them are full of diversity and history. As a result, it is a popular country to live in and properties are highly sought after by many people, both nationally and internationally. One of the most popular - if not the most popular - locations is within the county of Greater London. Greater London is densely populated, therefore starting any business there with an expected high foot traffic would be beneficial to that business.

Greater London is made up of 32 boroughs and 33 districts, with one of the districts being the City of London that does not identify as being a borough. Each borough has its own set of postcodes with designations as to the general location they are found in, such as E for East, NW for North West. Note that there are no S for South or NE for North East, they fold into the other postcodes.

However, not all areas of Greater London are safe to live in. Some boroughs are safer than others, so choosing to live at that location could be more hazardous. We'll use Data Science tools to look at the clustering of crime across all the Greater London boroughs and locate which borough is safest from robberies. With the safest boroughs from robberies selected, we can then work out where it would be best to start a Restaurant within that vicinity where the main venues are not restaurants.

Data

Data will come from three locations. They are:

- Crime Data (https://data.police.uk/data/)
- Borough postcode data (e.g. for Eastern Postcodes https://en.wikipedia.org/wiki/E_postcode_area)

Crime Data

Data will be for crimes reported by the Metropolitan Police and City of London Police as they together cover the Greater London area. The data we'll look into would be the 2019 set of data as it is the most recent complete year.

The data files will be csv files with the following properties:

- Crime ID
- Month
- Reported by
- Falls within

- Longitude
- Latitude
- Location
- LSOA code
- LSOA name
- Crime type
- Last outcome category
- Context
- The key columns are the Longitude and Latitude to use mark the locations.

Borough Postcode Data

Once the safest borough has been identified, the postcodes for the borough would be retrieved from Wikipedia. Postcodes would be run through geocoder to retrieve the latitudes and longitudes before passing them through the Foursquare Venue Explore API to retrieve local venues.

Methodology

The approach is in two parts, locating the safest borough from robberies and then using K-means clustering to segment customers and figure out which areas see the most foot-traffic.

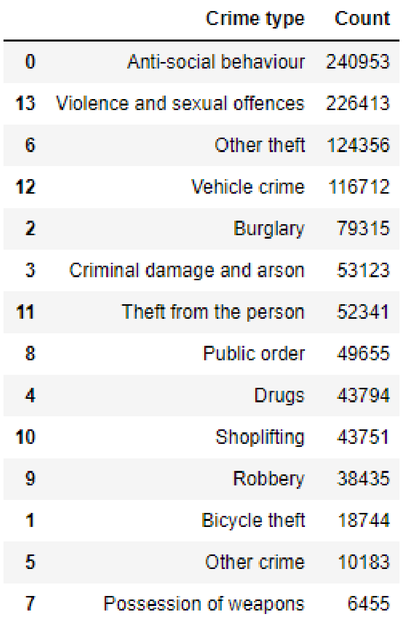
Data Group 1

Stage A – Crime Data

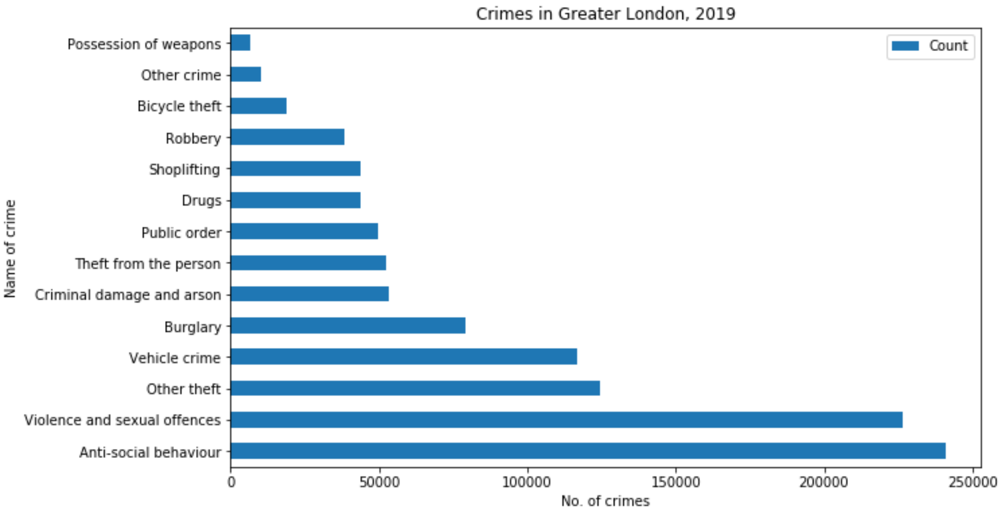
1. Retrieve crime data in csv format from https://data.police.uk/data/
2. Data to be retrieved is from the City of London Police and Metropolitan Police data, which covers most - if not all - of Greater London
3. Files are split into different directories by month. Merge the files into one large file for ease of loading into Pandas dataframe

Stage B – Review and filter

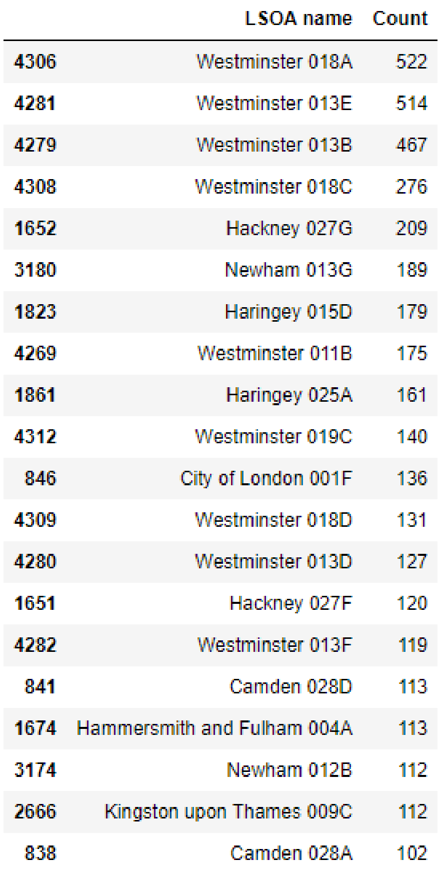
1. Get the shape of the dataframe to review size
2. Review of all crime data just to see how they are distributed



	Crime type	Count
0	Anti-social behaviour	240953
13	Violence and sexual offences	226413
6	Other theft	124356
12	Vehicle crime	116712
2	Burglary	79315
3	Criminal damage and arson	53123
11	Theft from the person	52341
8	Public order	49655
4	Drugs	43794
10	Shoplifting	43751
9	Robbery	38435
1	Bicycle theft	18744
5	Other crime	10183
7	Possession of weapons	6455



3. Filter crimes on Robbery only, order and review where most of them occur from



	LSOA name	Count
4306	Westminster 018A	522
4281	Westminster 013E	514
4279	Westminster 013B	467
4308	Westminster 018C	276
1652	Hackney 027G	209
3180	Newham 013G	189
1823	Haringey 015D	179
4269	Westminster 011B	175
1861	Haringey 025A	161
4312	Westminster 019C	140
846	City of London 001F	136
4309	Westminster 018D	131
4280	Westminster 013D	127
1651	Hackney 027F	120
4282	Westminster 013F	119
841	Camden 028D	113
1674	Hammersmith and Fulham 004A	113
3174	Newham 012B	112
2666	Kingston upon Thames 009C	112
838	Camden 028A	102

Stage C – Review distribution

1. Plot the robbery crimes on a map of Greater London using Folium
2. Review and choose an area with least about of robberies within Greater London
3. That borough to be used for next step.

Data Group 2

Stage A – Retrieve Postal Codes

1. With the winning borough selected, the postal codes for it will be web scraped from Wikipedia
2. Only postal codes associated to the borough would be kept as some postal codes overlap between boroughs
3. Postal codes stored in Pandas dataframe

Stage B – Obtain co-ordinates

1. With the postal codes, submit along with the borough to retrieve longitude and latitude co-ordinates using GeoPy
2. Store details within a new dataframe with Postal Code
3. Merge original postal code dataframe and co-ordinates dataframe into a new one for next step

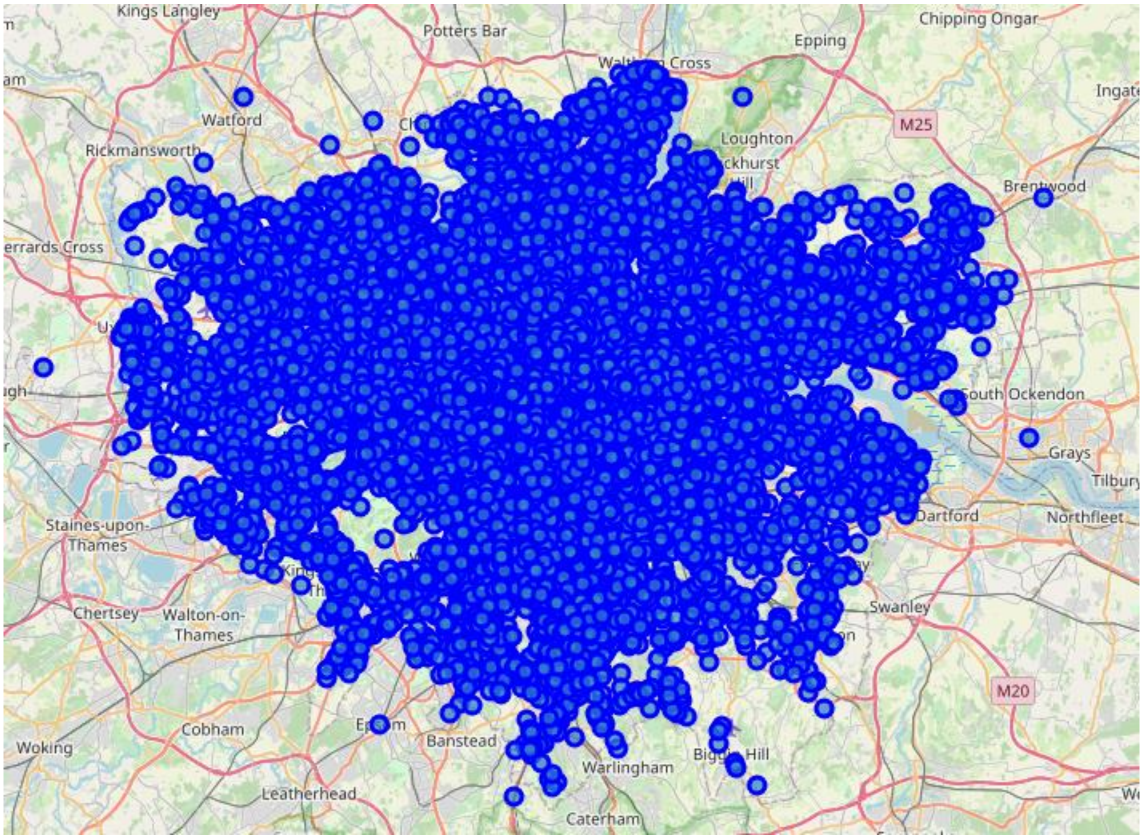
Data Group 3

Stage A – Foursquare Venues

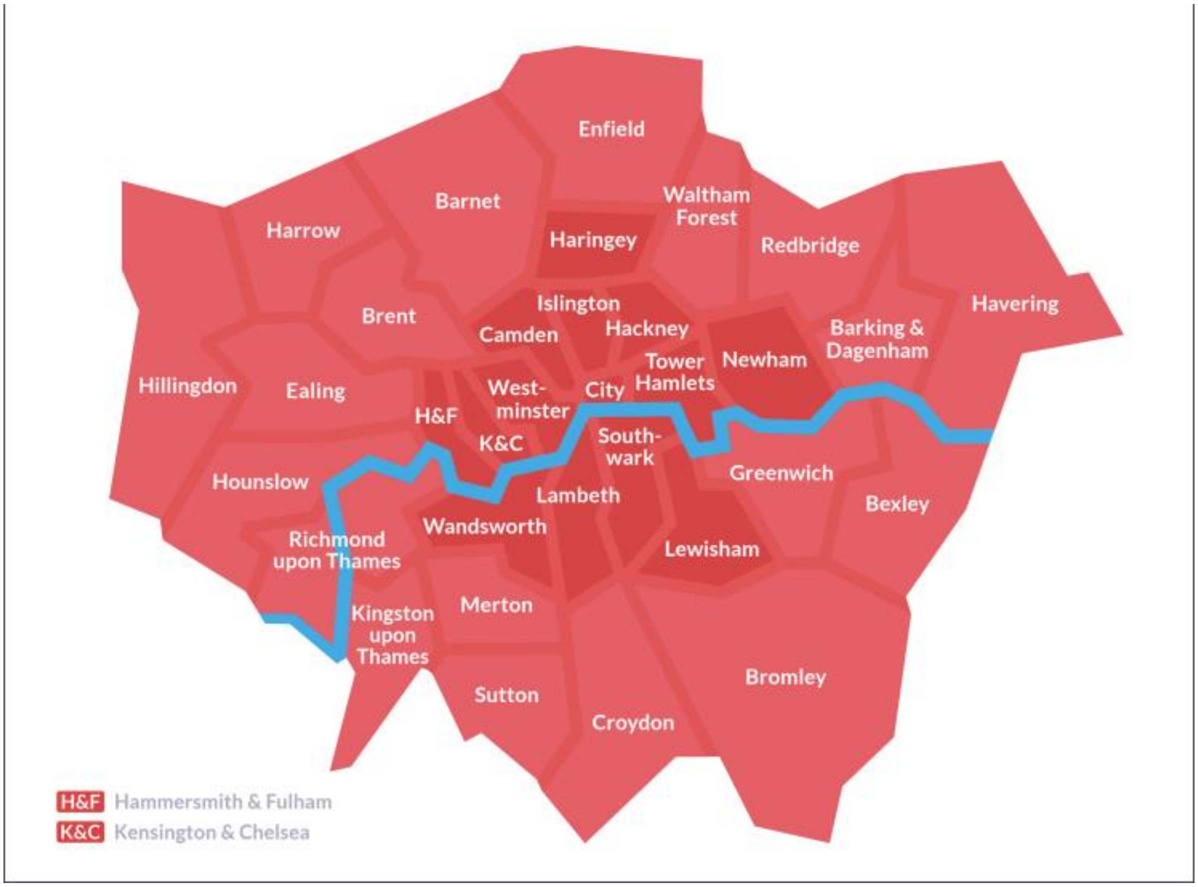
1. Using geographical co-ordinates for each postal code for selected borough, make calls to Foursquare API to retrieve the top 100 venues within 500 metres
2. Only postal codes associated to the borough would be kept as some postal codes overlap between boroughs
3. Perform the k-means clustering algorithm steps on dataset and display in Folium to visualize the postal codes in the borough and their emerging clusters

Results

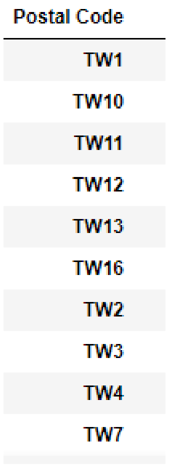
Initial findings after performing distributed crime hotspots for robbery show that a section of Greater London was least affect in 2019 in the south west:



Overlaid with a map that displays the boroughs, it is clear it is Richmond upon Thames:

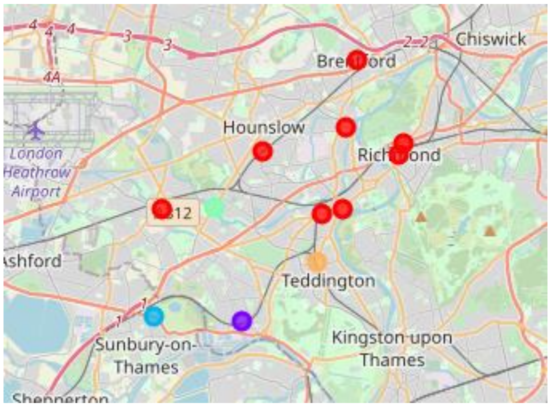


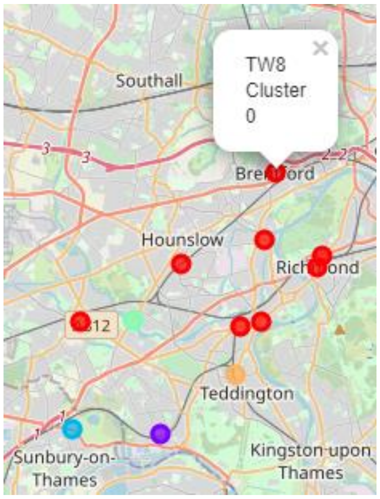
The borough chosen is clearly Richmond upon Thames. With this we perform web scraping to return the postal codes of Richmond upon Thames. Filtering only the Richmond upon Thames postal codes we are left with a small list:



Postal Code
TW1
TW10
TW11
TW12
TW13
TW16
TW2
TW3
TW4
TW7

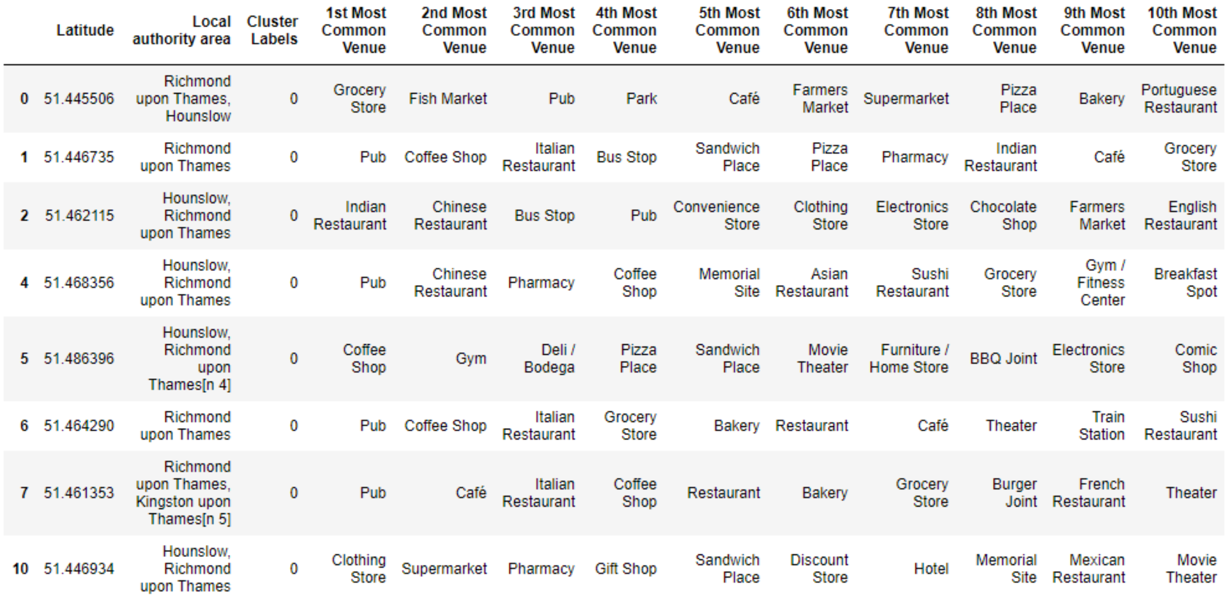
With the postal code data ready, we can the final step of the investigation by returned the top 100 venues them and then perform k-means to cluster the postal codes into five clusters. This is to allow for a good spread of clusters to see how the venues are clustered together per area.





The largest cluster appear to be red and far outweigh the others in size. The label for it is "Cluster 0".

Review of the cluster data shows that the venues for it are numerous:



	Latitude	Local authority area	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	51.445506	Richmond upon Thames, Hounslow	0	Grocery Store	Fish Market	Pub	Park	Café	Farmers Market	Supermarket	Pizza Place	Bakery	Portuguese Restaurant
1	51.446735	Richmond upon Thames	0	Pub	Coffee Shop	Italian Restaurant	Bus Stop	Sandwich Place	Pizza Place	Pharmacy	Indian Restaurant	Café	Grocery Store
2	51.462115	Hounslow, Richmond upon Thames	0	Indian Restaurant	Chinese Restaurant	Bus Stop	Pub	Convenience Store	Clothing Store	Electronics Store	Chocolate Shop	Farmers Market	English Restaurant
4	51.468356	Hounslow, Richmond upon Thames	0	Pub	Chinese Restaurant	Pharmacy	Coffee Shop	Memorial Site	Asian Restaurant	Sushi Restaurant	Grocery Store	Gym / Fitness Center	Breakfast Spot
5	51.486396	Hounslow, Richmond upon Thames[n 4]	0	Coffee Shop	Gym	Deli / Bodega	Pizza Place	Sandwich Place	Movie Theater	Furniture / Home Store	BBQ Joint	Electronics Store	Comic Shop
6	51.464290	Richmond upon Thames	0	Pub	Coffee Shop	Italian Restaurant	Grocery Store	Bakery	Restaurant	Café	Theater	Train Station	Sushi Restaurant
7	51.461353	Richmond upon Thames, Kingston upon Thames[n 5]	0	Pub	Café	Italian Restaurant	Coffee Shop	Restaurant	Bakery	Grocery Store	Burger Joint	French Restaurant	Theater
10	51.446934	Hounslow, Richmond upon Thames	0	Clothing Store	Supermarket	Pharmacy	Gift Shop	Sandwich Place	Discount Store	Hotel	Memorial Site	Mexican Restaurant	Movie Theater

Discussion

From the results presented, the following observations and recommendations can be made.

- In terms of safest area from robberies, it is Richmond upon Thames. For the entire crime data of 2019, there is a huge gap over Richmond upon Thames where the crime did not occur. The visualisation presented makes this very clear upon first glance.

- K-means clustering with venues for Richmond upon Thames postal codes has most of the nodes clustered around the Hounslow, Brentford and Richmond areas (cluster 0)
- It may seem counter-intuitive, but the restaurant should be opened around the area for cluster 0. The many nodes clustered together means higher foot traffic so potentially more customers
- Restaurant just needs to serve a cuisine not around in the top 10 frequent venues for best results. E.g. Vietnamese restaurant

Conclusion

In conclusion, we can say that the safest place from robbery is in Richmond upon Thames, and that a restaurant of a cuisine not common there would be best started around Hounslow, Brentford and Richmond areas, as the higher foot traffic in those areas to improve the chances of customers entering. Although the scope of the analysis is somewhat limited in that it does not address the cost of the locations, if money if no object, the recommendation stands.