11/18/2019

Final Project Reports

The Battle of the Neighbourhoods

By Mulaedza Mathoho

1. Introduction

1.1. Background

According to the United Nations, in 2017 there were 258 million international migrants worldwide. These people represent a very small proportion of the world's population: approximately 3.4%. Their numbers are rising, however: from 2010 to 2015, the total number of international immigrants rose from 220 million to 248 million, corresponding to an average increase of 2.4% per year.
Of the 258 million international migrants in 2017, 106 million were born in Asia. Europe is the birth region that accounts for the second largest number (61 million), followed by Latin America and the Caribbean (38 million) and Africa (36 million).
Refugees, who account for only 10% of international migrants, were estimated at 25.9 million in 2016. Most refugees (82.5%) live in developing countries.

1.2. Problem

There are certain factors that people look into when they decide to migrate to a different location. It might be because they are not happy where they are currently or maybe they are used to moving from place to place. To minimize the chances of this issue, we should always do proper research when planning on migrating.

The crime statistics dataset of London found on Kaggle has Crime in major metropolitan areas, such as London, occurs in distinct patterns. This data covers the number of criminal reports by month, LSOA borough, and major/minor category from Jan 2008-Dec 2016. The year 2016 being the latest we will be considering the data of that year which is actually old information as of now. The crime rates in each borough may have changed over time.

1.3. Interest

This project helps in selecting the safest borough in London based on the total crimes, explore the neighbourhoods of that borough to find the 10 most common venues in each neighbourhood and finally cluster the neighbourhoods using k-mean clustering. This report will be targeted to people who are looking to relocate to London. Inorder to finalise a neighbourhood to hunt for an apartment, safety is considered as a top concern when moving to a new place. If you don't feel safe in your own home, you're not going to be able to enjoy living there. The crime statistics will provide an insight into this issue.

2. Data Acquisition and Cleaning

2.1. Data Sources

Based on definition of our problem, factors that will influence our decision are:

- The total number of crimes commited in each of the borough during the last year.
- The most common venues in each of the neighbourhood in the safest borough selected

In this project a number of datasets were used. This is the list of the 3 datasets that were used for the project:

1. London crime data:
   - lsoa_code: code for Lower Super Output Area in Greater London.
   - borough: Common name for London borough.
   - major_category: High level categorization of crime
   - minor_category: Low level categorization of crime within major category.
   - value: monthly reported count of categorical crime in given borough
   - year: Year of reported counts, 2008-2016
   - month: Month of reported counts, 1-12

2. list of London boroughs:
   - Borough: The names of the 33 London boroughs.
   - Inner: Categorizing the borough as an Inner London borough or an Outer London Borough.
   - Status: Categorizing the borough as Royal, City or other borough.
   - Local authority: The local authority assigned to the borough.
   - Political control: The political party that control the borough.
   - Headquarters: Headquarters of the Boroughs.
   - Area (sq mi): Area of the borough in square miles.
   - Population (2013 est)[1]: The population in the borough recorded during the year 2013.
   - Co-ordinates: The latitude and longitude of the boroughs.
   - Nr. in map: The number assigned to each borough to represent visually on a map

3. list of Neighbourhoods in the Royal Borough of Kingston upon Thames:
   - Neighbourhood: Name of the neighbourhood in the Borough.
   - Borough: Name of the Borough.
   - Latitude: Latitude of the Borough.
   - Longitude: Longitude of the Borough.

2.2. Data Cleaning

Data preparation for each of the 3 datasets (London crime data, list of London boroughs, list of Neighbourhoods in the Royal Borough of Kingston upon Thames) was done separately. Considering the London crime data, only crimes during the most recent year were selected (2016). Fig 2.1 Shows the data before Pre-processing, Fig 2.2 shows the data after Pre-processing.

Fig 2.1:

	lsoa_code	borough	major_category	minor_category	value	year	month
0	E01001116	Croydon	Burglary	Burglary in Other Buildings	0	2016	11
1	E01001646	Greenwich	Violence Against the Person	Other violence	0	2016	11
2	E01000677	Bromley	Violence Against the Person	Other violence	0	2015	5
3	E01003774	Redbridge	Burglary	Burglary in Other Buildings	0	2016	3
4	E01004563	Wandsworth	Robbery	Personal Property	0	2008	6

Fig 2.2:

	Borough	Burglary	Criminal Damage	Drugs	Other Notifiable Offences	Robbery	Theft and Handling	Violence Against the Person	Total
0	Barking and Dagenham	1287	1949	919	378	534	5607	6067	16741
1	Barnet	3402	2183	906	499	464	9731	7499	24684
2	Bexley	1123	1673	646	294	209	4392	4503	12840
3	Brent	2631	2280	2096	536	919	9026	9205	26693
4	Bromley	2214	2202	728	417	369	7584	6650	20164

Using the Beautiful Soup library, the second data (list of London boroughs) was scraped from Wikipedia. Using this library we can extract the data in the tabular format as shown in the website.

```
# getting data from internet
wikipedia_link='https://en.wikipedia.org/wiki/List_of_London_boroughs'
raw_wikipedia_page= requests.get(wikipedia_link).text

soup = BeautifulSoup(raw_wikipedia_page,'xml')
print(soup.prettify())
```

String manipulation was also done to get the names of the boroughs in the correct format. Fig 2.3 below shows the format of the data after merging the two datasets using the Borough names.

Fig 2.3:

	Borough	Inner	Status	Local authority	Political control	Headquarters	Area (sq mi)	Population (2013 est) [1]	Co-ordinates	Nr. in ma
28	Tower Hamlets	NaN	NaN	Tower Hamlets London Borough Council	Labour	Town Hall, Mulberry Place, 5 Clove Crescent	7.63	272890	51°30′36″N 0°00′21″W / 51.5099°N 0.0059°W	8
29	Waltham Forest	NaN	NaN	Waltham Forest London Borough Council	Labour	Waltham Forest Town Hall, Forest Road	14.99	265797	51°35′27″N 0°00′48″W / 51.5908°N 0.0134°W	28
30	Wandsworth	NaN	NaN	Wandsworth London Borough Council	Conservative	The Town Hall, Wandsworth High Street	13.23	310516	51°27′24″N 0°11′28″W / 51.4567°N 0.1910°W	5
31	Westminster	NaN	City	Westminster City Council	Conservative	Westminster City Hall, 64 Victoria Street	8.29	226841	51°29′50″N 0°08′14″W / 51.4973°N 0.1372°W	2
32	City of London	([note 5]	Sui generis;City;Ceremonial county	Corporation of London;Inner Temple;Middle Temple	?	Guildhall	1.12	7000	51°30′56″N 0°05′32″W / 51.5155°N 0.0922°W	1

After visualizing the crime in each borough we can find the borough with the lowest crime rate and hence tag that borough as the safest borough. The third source of data is acquired from the list of neighbourhoods in the safest borough on wikipedia. This dataset is created from scratch, the pandas data frame is created with the names of the neighbourhoods and the name of the borough with the latitude and longitude left blank (see fig 2.4).

Fig 2.4:

	Neighborhood	Borough	Latitude	Longitude
0	Berrylands	Kingston upon Thames	51.393781	-0.284802
1	Canbury	Kingston upon Thames	51.417499	-0.305553
2	Chessington	Kingston upon Thames	51.358336	-0.298622
3	Coombe	Kingston upon Thames	51.419450	-0.265398
4	Hook	Kingston upon Thames	51.367898	-0.307145

The new dataset is used to generate the 10 most common venues for each neighbourhood using the Foursquare API, finally using k means clustering algorithm to cluster similar neighbourhoods together.

3.Methodology

3.1. Exploratory Data Analysis

3.1.1. Crime Statistical summary

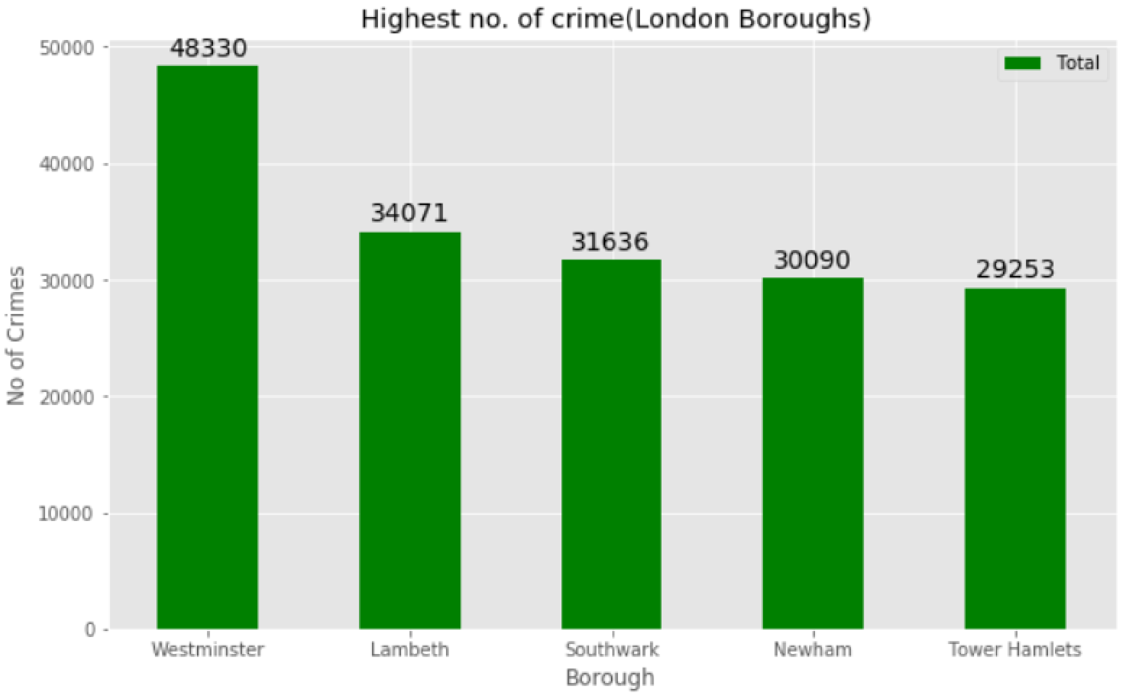
London Crime statistical summary showing the Count,Mean,Standard deviation,minimum,nmaximum,1st Quartile, 2nd Quartile and 3rd Quartile. Fig 3.1.

	Burglary	Criminal Damage	Drugs	Other Notifiable Offences	Robbery	Theft and Handling	Violence Against the Person	Total
count	33.000000	33.000000	33.000000	33.000000	33.000000	33.000000	33.000000	33.000000
mean	2069.242424	1941.545455	1179.212121	479.060606	682.666667	8913.121212	7041.848485	22306.696970
std	737.448644	625.207070	586.406416	223.298698	441.425366	4620.565054	2513.601551	8828.228749
min	2.000000	2.000000	10.000000	6.000000	4.000000	129.000000	25.000000	178.000000
25%	1531.000000	1650.000000	743.000000	378.000000	377.000000	5919.000000	5936.000000	16903.000000
50%	2071.000000	1989.000000	1063.000000	490.000000	599.000000	8925.000000	7409.000000	22730.000000
75%	2631.000000	2351.000000	1617.000000	551.000000	936.000000	10789.000000	8832.000000	27174.000000
max	3402.000000	3219.000000	2738.000000	1305.000000	1822.000000	27520.000000	10834.000000	48330.000000

3.1.2. Boroughs with the highest crime rate

The Bar graph below shows the top 5 boroughs with the highest crime rate. We will stay away from this place.

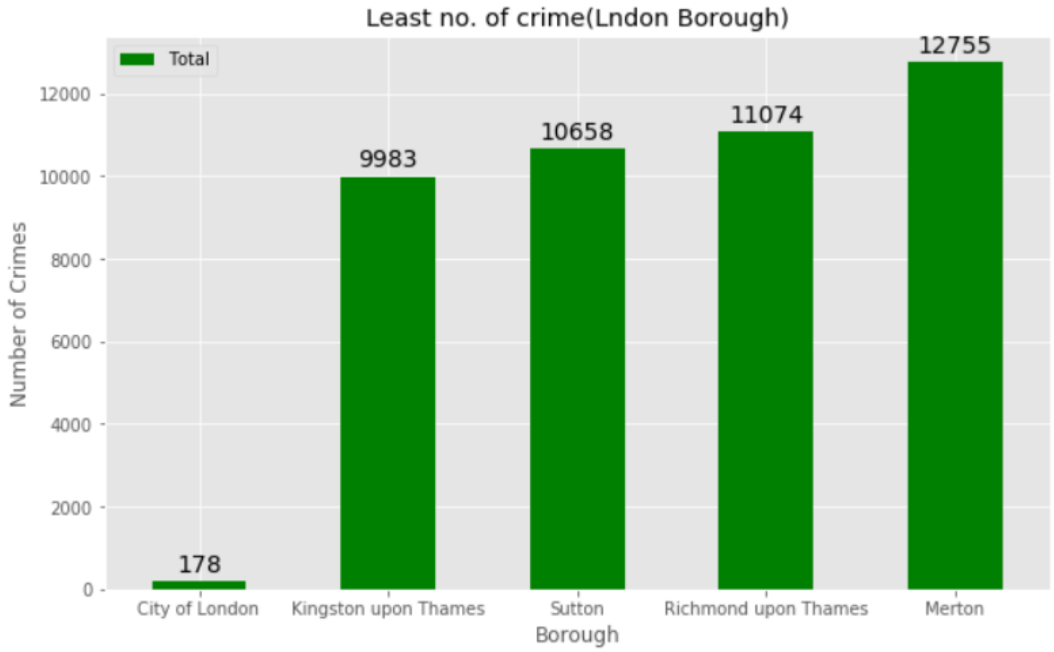
Fig 3.1.2:



3.1.3. Boroughs with the lowest crime rate

The Bar gragh below shows the top 5 boroughs with the lowest crime rate

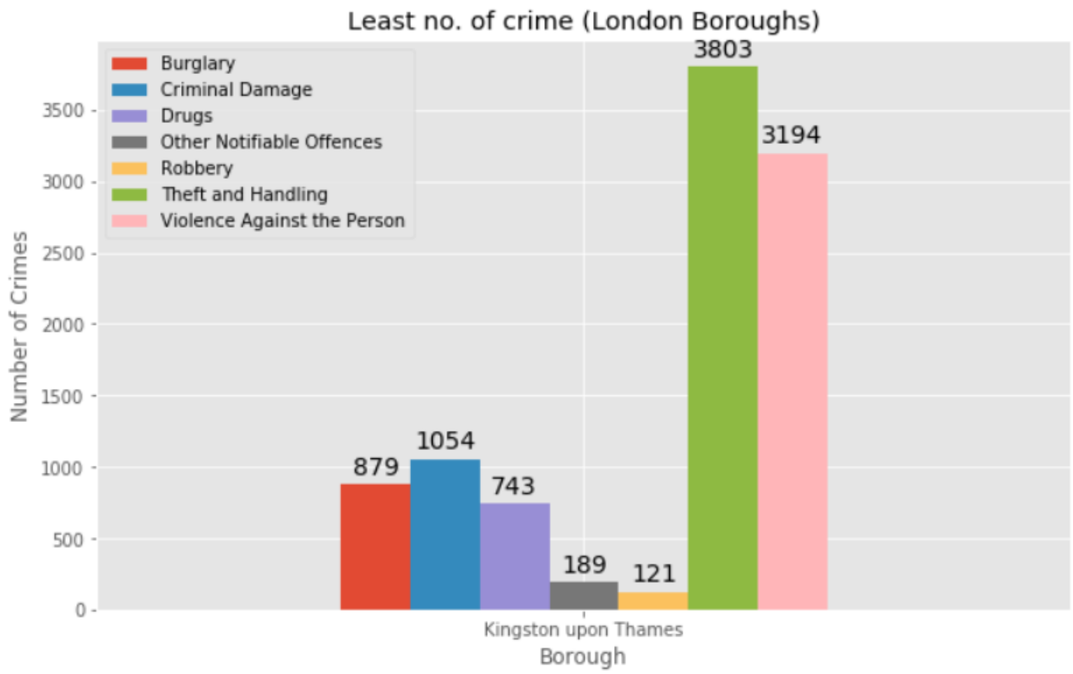
Fig 3.1.2a:



As per the wikipedia page, The City of London is the 33rd principal division of Greater London but it is not a London borough. Hence we will focus on the next borough with the least crime i.e. Kingston upon Thames.

Fig 3.1.2b below shows the category of crimes in Kingston upon Thames Borough

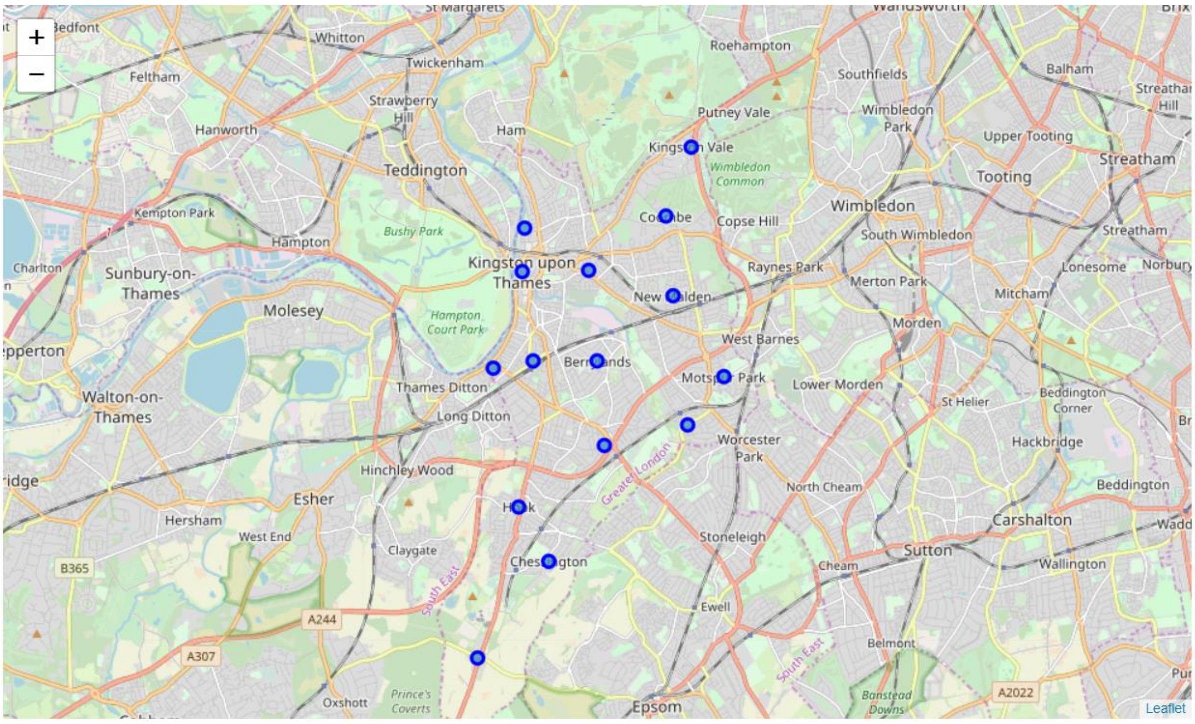
Fig 3.1.2b



3.1.4. Neighbourhoods in Kingston Thames

The map below shows the 15 neighbourhoods of Kingston Upon Thames.

Fig 3.1.4:



3.2. Modelling

Using the final dataset containing the neighbourhoods in Kingston upon Thames, this were the goals achieved:

- Finding all the venues within a 500-meter radius of each neighbourhood.
- Perform one hot ecoding on the venues data.
- Grouping the venues by the neighbourhood and calculating their mean.
- Performing a K-means clustering (Defining K = 5)

The data frame below contains all the venues along with their coordinates and category.

Fig 3.2.1

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Berrylands	51.393781	-0.284802	Surbiton Racket & Fitness Club	51.392676	-0.290224	Gym / Fitness Center
1	Berrylands	51.393781	-0.284802	24Hrs-Berrylands Minicabs-0208540444-Mini Cabs	51.393757	-0.285130	Taxi Stand
2	Berrylands	51.393781	-0.284802	K2 Bus Stop	51.392302	-0.281534	Bus Stop
3	Berrylands	51.393781	-0.284802	La Monaliza	51.389936	-0.283165	Colombian Restaurant
4	Canbury	51.417499	-0.305553	Canbury Gardens	51.417409	-0.305300	Park

4. Results

After running the K-means clustering we can access each cluster created to see which neighbourhoods were assigned to each of the five clusters.

Fig 4.1 Shows the first Cluster generated.

Fig 4.1:

	Neighborhood	Borough	Latitude	Longitude	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	8tl Co
1	Canbury	Kingston upon Thames	51.417499	-0.305553	0	Pub	Park	Fish & Chips Shop	Indian Restaurant	Supermarket	Spa	Hotel	Sh Se
7	Malden Rushett	Kingston upon Thames	51.341052	-0.319076	0	Garden Center	Convenience Store	Pub	Restaurant	Discount Store	Construction & Landscaping	Cosmetics Shop	Del Bo
11	Old Malden	Kingston upon Thames	51.382484	-0.259090	0	Train Station	Food	Pub	Child Care Service	Construction & Landscaping	French Restaurant	Fried Chicken Joint	Fis Chi Sh

The cluster one is the second biggest cluster with 3 of the 15 neighbourhoods in the borough Kingston upon Thames. Upon closely examining these neighbourhoods we can see that the most common venues in these neighbourhoods are Restaurants, Pubs, Cafe, Supermarkets, and stores.

Fig 4.2 Shows the second Cluster generated.

Fig 4.2:

	Neighborhood	Borough	Latitude	Longitude	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
2	Chessington	Kingston upon Thames	51.358336	-0.298622	1	Grocery Store	Wine Shop	Farmers Market	Convenience Store	Cosmetics Shop	Deli / Bodega	Department Store	Discount Store

The second cluster has one neighbourhood which consists of Venues such as store, wine shops and markets.

Fig 4.3 Shows the third Custer generated.

Fig 4.3:

	Neighborhood	Borough	Latitude	Longitude	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
4	Hook	Kingston upon Thames	51.367898	-0.307145	2	Indian Restaurant	Bakery	Supermarket	Fish & Chips Shop	Wine Shop	Convenience Store	Cosmetics Shop
5	Kingston upon Thames	Kingston upon Thames	51.409627	-0.306262	2	Coffee Shop	Café	Pub	Burger Joint	Sushi Restaurant	Department Store	German Restaurant
9	New Malden	Kingston upon Thames	51.405335	-0.263407	2	Gastropub	Korean Restaurant	Chinese Restaurant	Sushi Restaurant	Supermarket	Bar	Gym
10	Norbiton	Kingston upon Thames	51.409999	-0.287396	2	Indian Restaurant	Pub	Food	Italian Restaurant	Wine Shop	Fried Chicken Joint	Grocery Store
12	Seething Wells	Kingston upon Thames	51.392642	-0.314366	2	Indian Restaurant	Coffee Shop	Italian Restaurant	Café	Pub	Fish & Chips Shop	Fast Food Restaurant
13	Surbiton	Kingston upon Thames	51.393756	-0.303310	2	Coffee Shop	Pub	Italian Restaurant	Grocery Store	Breakfast Spot	Deli / Bodega	Gym / Fitness Center

The third cluster has Six neighbourhoods, making it the largest cluster, it consists of Venues such as Pubs, Restaurants, Stores and Wine shops.

Fig 4.4 Shows the fourth Cluster generated.

Fig 4.4

	Neighborhood	Borough	Latitude	Longitude	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 Com V
6	Kingston Vale	Kingston upon Thames	51.431850	-0.258138	3	Grocery Store	Bar	Soccer Field	Italian Restaurant	Wine Shop	Farmers Market	Cosmetics Shop	Deli Bode
8	Motspur Park	Kingston upon Thames	51.390985	-0.248898	3	Construction & Landscaping	Park	Gym	Soccer Field	Electronics Store	Convenience Store	Cosmetics Shop	Deli Bode
14	Tolworth	Kingston upon Thames	51.378876	-0.282860	3	Grocery Store	Pharmacy	Sandwich Place	Train Station	Discount Store	Hotel	Coffee Shop	Pizz Plac

The fourth cluster has three neighbourhoods in it, these neighbourhoods have common venues such as Parks, Soccer fields, Train stations,Stores,etc.

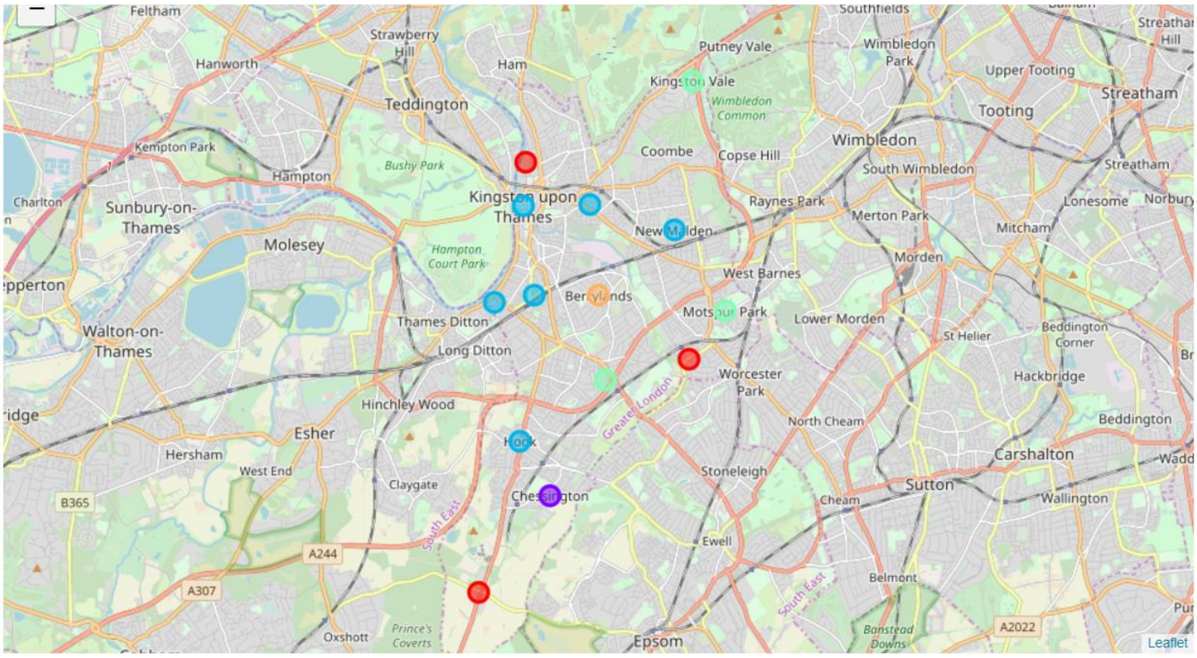
Fig 4.5 Shows the fifth Cluster generated

Fig 4.5:

	Neighborhood	Borough	Latitude	Longitude	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
0	Berrylands	Kingston upon Thames	51.393781	-0.284802	4	Colombian Restaurant	Park	Gym / Fitness Center	Bus Stop	Electronics Store	Cosmetics Shop	Deli / Bodega	Department Store

The fifth cluster has one neighbourhood which consists of Venues such as Park, Gym/Fitness Center,Electonic store,etc.

Visualization of the Neighbourhoods



Each cluster is color coded for the ease of presentation, we can see that majority of the neighborhood falls in the red cluster which is the first cluster. Three neighborhoods have their own cluster (Blue, Purple and Yellow), these are clusters two three and five. The green cluster consists of two neighborhoods which is the 4th cluster.

4. Discussion

The aim of this project is to help people who want to relocate to the safest borough in London, expats can choose the neighbourhoods to which they want to relocate based on the most common venues in it. For example if a person is looking for a neighbourhood with good connectivity and public transportation we can see that Clusters 3 and 4 have Train stations and Bus stops as the most common venues. If a person is looking for a neighbourhood with stores and restaurants in a close proximity then the neighbourhoods in the first cluster is suitable. For a family I feel that the neighbourhoods in Cluster 4 are more suitable dues to the common venues in that cluster, these neighbourhoods have common venues such as Parks, Gym/Fitness centers, Bus Stops, Restaurants, Electronics Stores and Soccer fields which is ideal for a family.

5. Conclusion

This project helps a person get a better understanding of the neighbourhoods with respect to the most common venues in that neighbourhood. It is always helpful to make use of technology to stay one step ahead i.e. finding out more about places before moving into a neighbourhood. We have just taken safety as a primary concern to shortlist the borough of London. The future of this project includes taking other factors such as cost of living in the areas into consideration to shortlist the borough based on safety and a predefined budget.