IBM DATA SCIENCE

The Battle of the Neighborhoods for London Area

Abstract

To move or not to move....

Coursera Course

INTRODUCTION

To move or not to move….

BACKGROUND

The average American moves about eleven times in their lifetime. This brings us to the question: Do people move until they find a place to settle down where they truly feel happy, or do our wants and needs change over time, prompting us to eventually leave a town we once called home for a new area that will bring us satisfaction? Or, do we too often move to a new area without knowing exactly what we're getting into, forcing us to turn tail and run at the first sign of discomfort?

To minimize the chances of this happening, we should always do proper research when planning our next move in life. Consider the following factors when picking a new place to live so you don't end up wasting your valuable time and money making a move, you'll end up regretting. Safety is a top concern when moving to a new area. If you don't feel safe in your own home, you're not going to be able to enjoy living there.

PROBLEM

The crime statistics dataset of London found on Kaggle has crimes in each Boroughs of London from 2008 to 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.

This project aims to select the safest borough in London based on the total crimes, explore the neighborhoods of that borough to find the 10 most common venues in each neighborhood and finally cluster the neighborhoods using k-mean clustering.

OBJECTIVE

Expats who are considering relocating to London will be interested to identify the safest borough in London and explore its neighborhoods and common venues around each neighborhood.

DATA

To move or not to move….

DATA SOURCE

The data acquired for this project is a combination of data from three sources:

1. https://www.kaggle.com/jboysen/london-crime that shows the crime per borough in London. The dataset contains the following columns:
    - **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. https://en.wikipedia.org/wiki/List_of_London_boroughs that contains additional information about the boroughs. The dataset contains the following columns:
    - **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 another 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)** : 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. https://en.wikipedia.org/wiki/List_of_districts_in_the_Royal_Borough_of_Kingston_upon_Thames which is created from scratch using the list of neighborhoods available on the site. The dataset contains the following columns:
    - **Neighborhood**: Name of the neighborhood in the Borough
    - **Borough**: Name of the Borough
    - **Latitude**: Latitude of the Borough
    - **Longitude**: Longitude of the Borough

The coordinates of the neighborhoods are obtained using **Google Maps API geocoding** to get the final dataset. Then, the new dataset is used to generate the venues for each neighborhood using the Foursquare API.

METHODOLOGY

To move or not to move….

The methodology in this project consists of two parts:

- **Exploratory Data Analysis**: Visualize the crime rates in the London boroughs to identity the safest borough and extract the neighborhoods in that borough to find the 10 most common venues in each neighborhood.
- **Modelling**: To help people find similar neighborhoods in the safest borough we will be clustering similar neighborhoods using K - means clustering which is a form of unsupervised machine learning algorithm that clusters data based on predefined cluster size. We will use a cluster size of 5 for this project that will cluster the 15 neighborhoods into 5 clusters. The reason to conduct a K- means clustering is to cluster neighborhoods

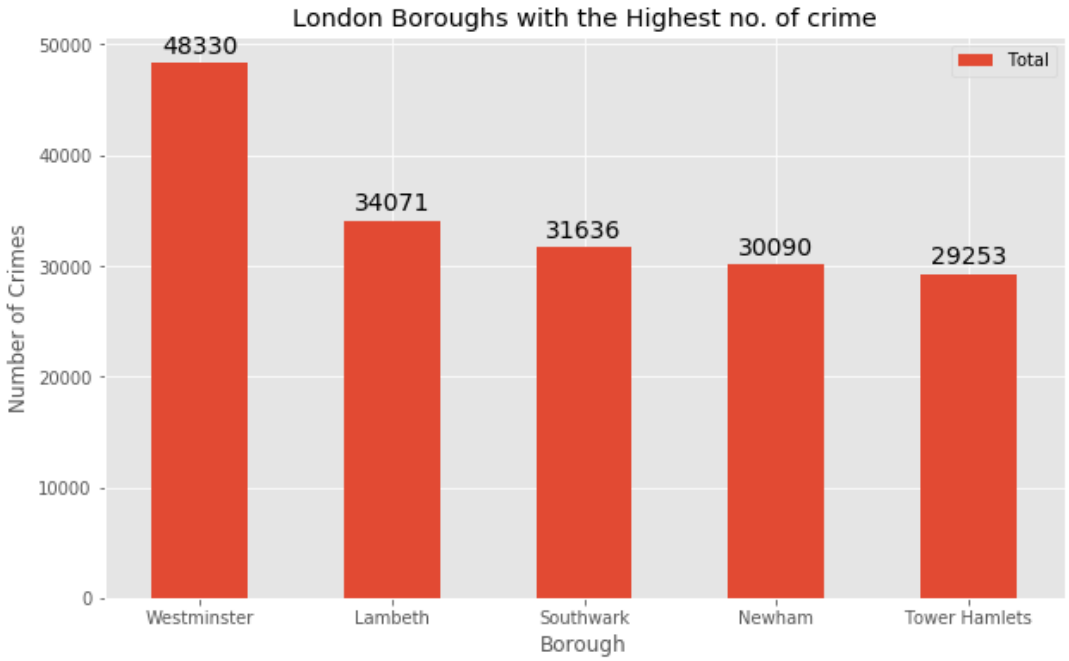
with similar venues together so that people can shortlist the area of their interests based on the venues/amenities around each neighborhood.

EXPLORATORY DATA ANALYSIS

Descriptive statistics of the data:

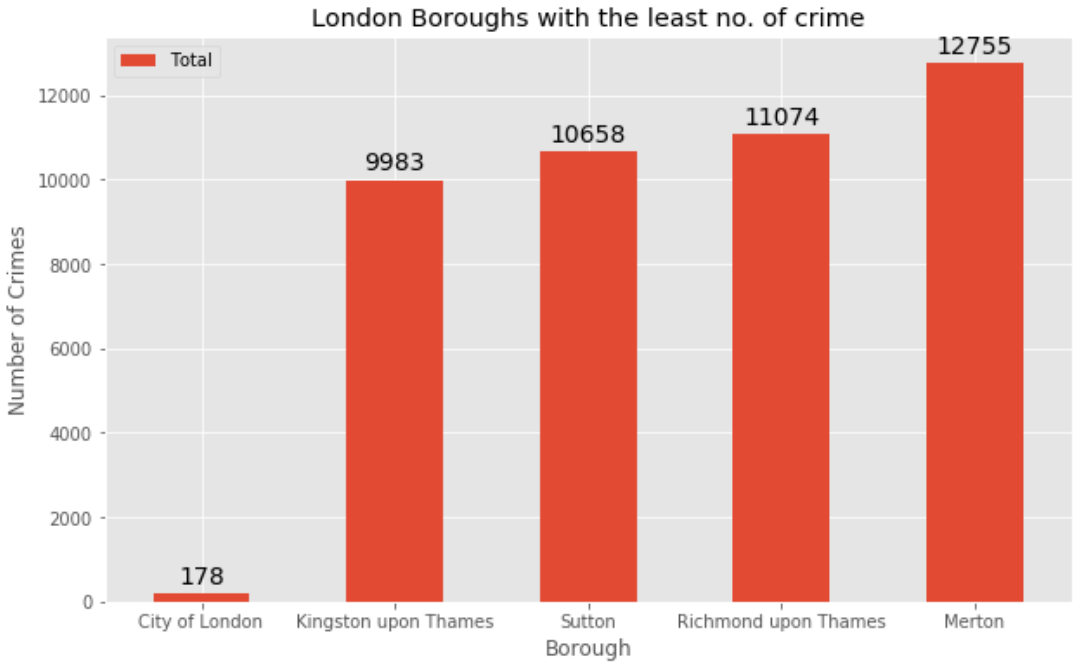
	Burglary	Criminal Damage	Drugs	Other Notifiable Offences	Robbery	Theft and Handling	Violence Against the Person	Total
count	33.000000	33.000000	33.000000	33.000000	33.000000	33.000000	33.000000	33.000000
mean	2069.242424	1941.545455	1179.212121	479.060606	682.666667	8913.121212	7041.848485	22306.696970
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

Top five boroughs with the highest number of crimes:



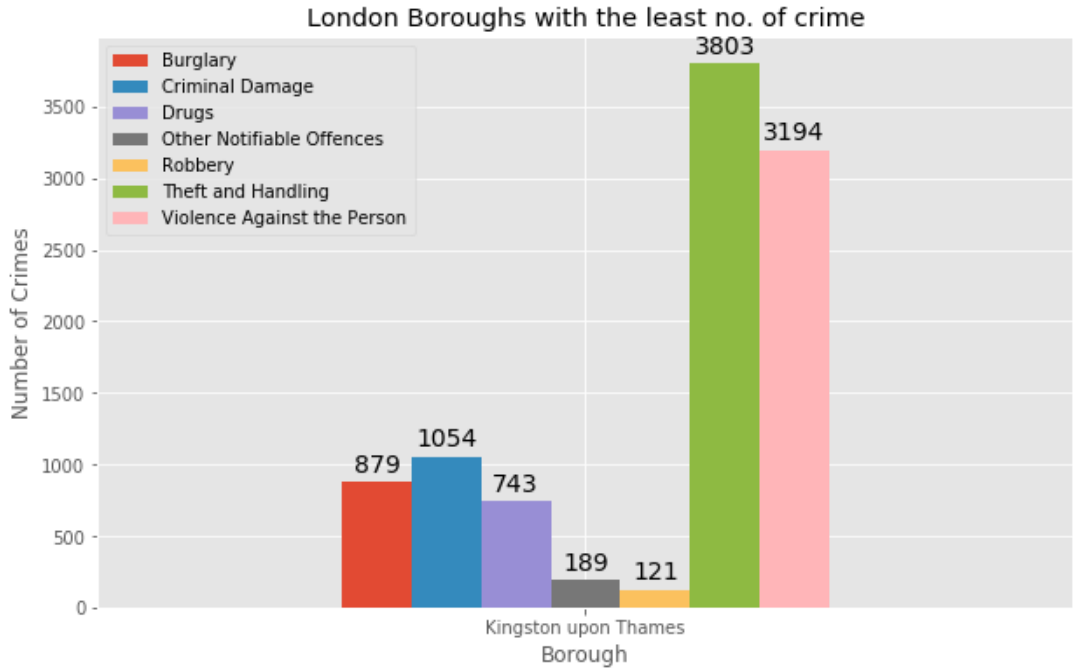
So, we will stay away from these places…

Top five boroughs with the lowest number of crimes:

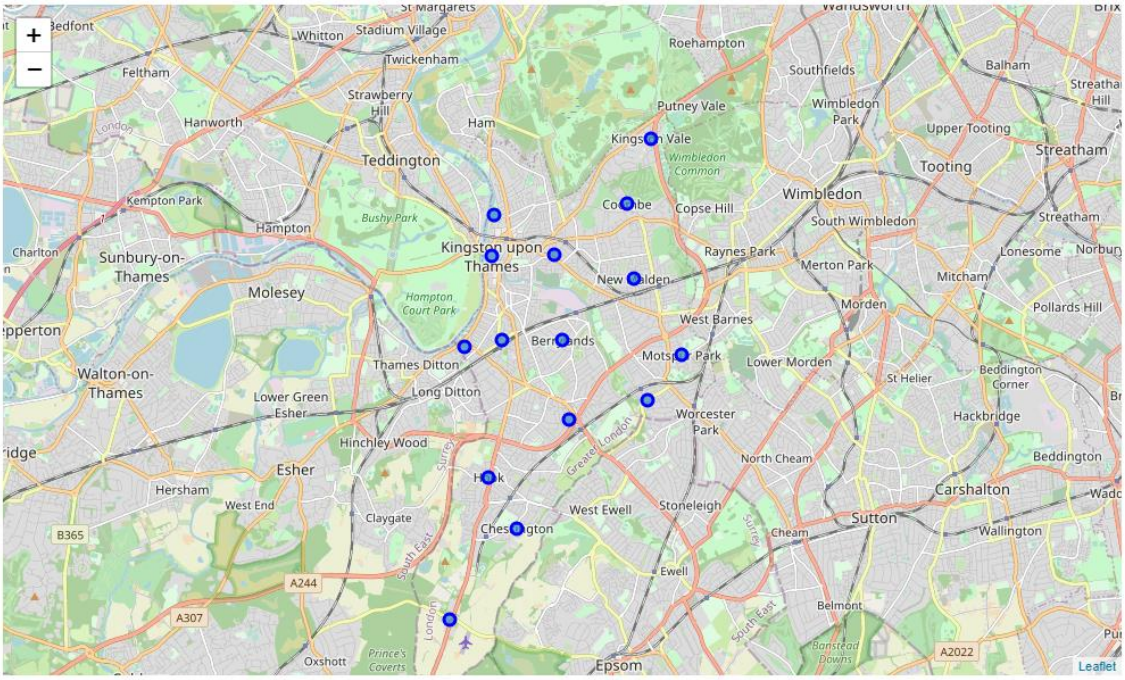


As per the Wikipedia (https://en.wikipedia.org/wiki/List_of_London_boroughs), The City of London is the 33rd principal division of Greater London but it is not a London borough, so we will focus on the next borough with the least crime i.e. Kingston upon Thames.

Different types of crimes in the borough 'Kingston upon Thames:



Visualize the Neighborhood of Kingston upon Thames Borough:



MODELLING

Steps involved in the modelling are:

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

Using the final dataset containing the neighborhoods in Kingston upon Thames along with the latitude and longitude, we can find all the venues within a 500-meter radius of each neighborhood by connecting to the Foursquare API. This returns a json file containing all the venues in each neighborhood which is converted to a pandas data frame. This data frame contains all the venues along with their coordinates and category as shown below.

	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	Alexandra Park	51.394230	-0.281206	Park
2	Berrylands	51.393781	-0.284802	K2 Bus Stop	51.392302	-0.281534	Bus Stop
3	Canbury	51.417499	-0.305553	Canbury Gardens	51.417409	-0.305300	Park
4	Canbury	51.417499	-0.305553	The Boater's Inn	51.418546	-0.305915	Pub

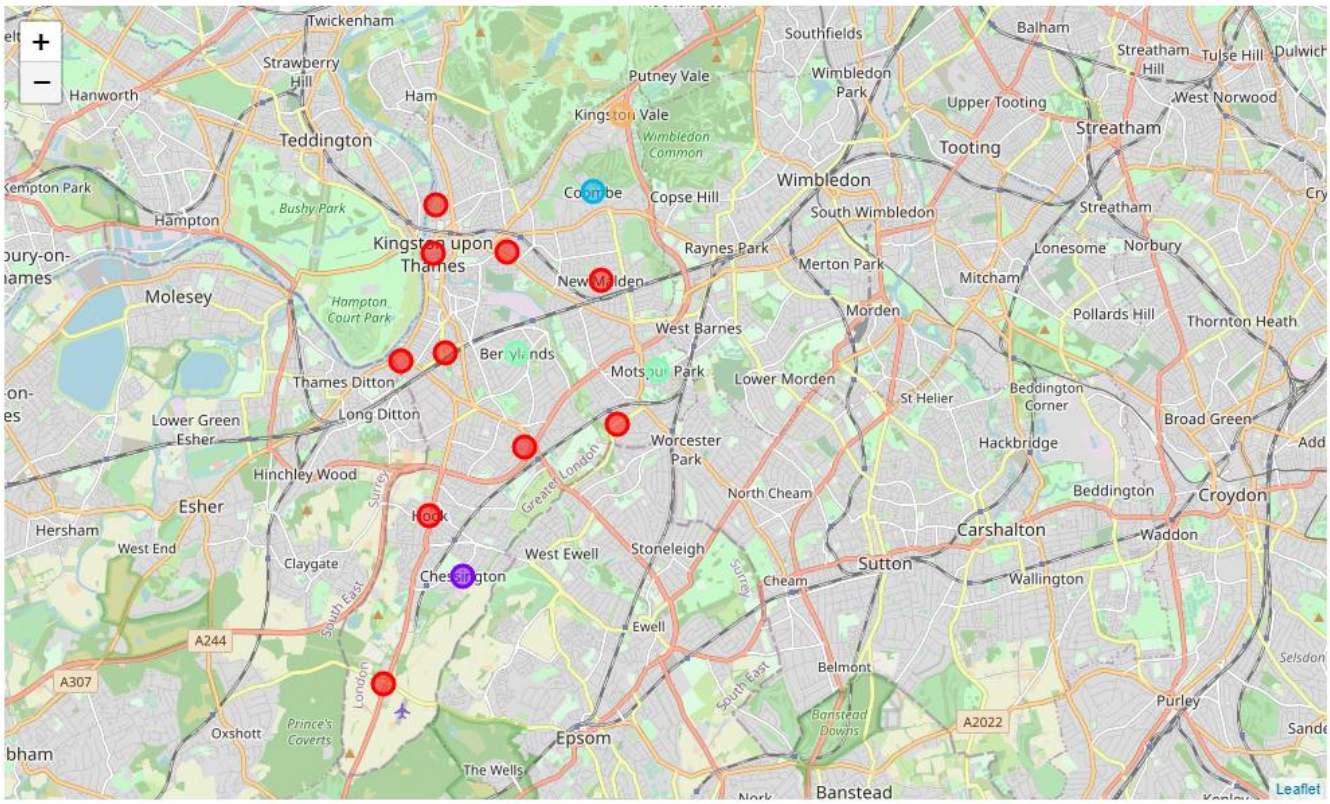
One hot encoding is done on the venues data. (One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction). The Venues data is then grouped by the Neighborhood and the mean of the venues are calculated, finally the 10 common venues are calculated for each of the neighborhoods.

To help people find similar neighborhoods in the safest borough we will be clustering similar neighborhoods using K - means clustering which is a form of unsupervised machine learning algorithm that clusters data based on predefined cluster size. We will use a cluster size of 5 for this project that will cluster the 15 neighborhoods into 5 clusters. The reason to conduct a K-means clustering is to cluster neighborhoods with similar venues together so that people can shortlist the area of their interests based on the venues/amenities around each neighborhood.

RESULTS

To move or not to move….

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



Looking into the neighborhoods in the first cluster:

	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
1	Canbury	Kingston upon Thames	51.417499	-0.305553	0	Pub	Park	Fish & Chips Shop	Supermarket	Spa	Gym / Fitness Center	Shop & Service	Plaza
4	Hook	Kingston upon Thames	51.367898	-0.307145	0	Bakery	Supermarket	Fish & Chips Shop	Indian Restaurant	Turkish Restaurant	Food	Discount Store	Dry Cleaner
5	Kingston upon Thames	Kingston upon Thames	51.409627	-0.306262	0	Café	Pub	Burger Joint	Sushi Restaurant	Coffee Shop	Turkish Restaurant	Electronics Store	Gift Shop
7	Malden Rushett	Kingston upon Thames	51.341052	-0.319076	0	Grocery Store	Garden Center	Pub	Restaurant	Fast Food Restaurant	Department Store	Discount Store	Dry Cleaner
9	New Malden	Kingston upon Thames	51.405335	-0.263407	0	Gastropub	Indian Restaurant	Gym	Sushi Restaurant	Supermarket	Bar	Chinese Restaurant	Korean Restaurant
10	Norbiton	Kingston upon Thames	51.409999	-0.287396	0	Pub	Indian Restaurant	Italian Restaurant	Food	Platform	Breakfast Spot	Japanese Restaurant	Hardware Store
11	Old Malden	Kingston upon Thames	51.382484	-0.259090	0	Train Station	Pub	Construction & Landscaping	Food	Gastropub	Garden Center	Furniture / Home Store	Fried Chicken Joint
12	Seething Wells	Kingston upon Thames	51.392642	-0.314366	0	Indian Restaurant	Pub	Café	Coffee Shop	Hotel	Fish & Chips Shop	Restaurant	Italian Restaurant
13	Surbiton	Kingston upon Thames	51.393756	-0.303310	0	Coffee Shop	Pub	Grocery Store	Italian Restaurant	Pharmacy	Gastropub	Thai Restaurant	Bakery
14	Tolworth	Kingston upon Thames	51.378876	-0.282860	0	Grocery Store	Pharmacy	Restaurant	Sandwich Place	Train Station	Hotel	Indian Restaurant	Discount Store

The cluster one is the biggest cluster with 9 of the 15 neighborhoods in the borough Kingston upon Thames. Upon closely examining these neighborhoods we can see that the most common venues in these neighborhoods are Restaurants, Pubs, Cafe, Supermarkets, and stores.

Looking the second cluster:

	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	9th Most Common Venue
2	Chessington	Kingston upon Thames	51.358336	-0.298622	1	Construction & Landscaping	Turkish Restaurant	Deli / Bodega	Discount Store	Dry Cleaner	Electronics Store	Farmers Market	Fast Food Restaurant	Fish & Chips Shop

The second cluster has one neighborhood which consists of Venues such as Restaurants, Golf courses, and wine shops.

Looking the third cluster:

	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	9th Most Common Venue	10th Co
3	Coombe	Kingston upon Thames	51.41945	-0.265398	2	Health & Beauty Service	Turkish Restaurant	Food	Discount Store	Dry Cleaner	Electronics Store	Farmers Market	Fast Food Restaurant	Fish & Chips Shop	Rest

The third cluster has one neighborhood which consists of Venues such as Train stations, Restaurants, and Furniture shops.

Looking at the fourth cluster:

	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	9th Most Common Venue
0	Berrylands	Kingston upon Thames	51.393781	-0.284802	3	Gym / Fitness Center	Park	Bus Stop	Turkish Restaurant	Fish & Chips Shop	Dry Cleaner	Electronics Store	Farmers Market	Fast Food Restaurant
8	Motspur Park	Kingston upon Thames	51.390985	-0.248898	3	Park	Gym	Soccer Field	Bus Stop	Restaurant	Turkish Restaurant	Fast Food Restaurant	Discount Store	Dry Cleaner

The fourth cluster has two neighborhoods in it, these neighborhoods have common venues such as Parks, Gym/Fitness centers, Bus Stops, Restaurants, Electronics Stores and Soccer fields etc.

Looking at the fifth cluster:

	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	9th Most Common Venue	1 M Comr Ve
6	Kingston Vale	Kingston upon Thames	51.43185	-0.258138	4	Grocery Store	Sandwich Place	Bar	Soccer Field	Discount Store	Dry Cleaner	Electronics Store	Farmers Market	Fast Food Restaurant	Fis C S

The fifth cluster has one neighborhood which consists of Venues such as Grocery shops, Bars, Restaurants, Furniture shops, and Department stores.

DISCUSSION

To move or not to move….

The aim of this project is to help people who want to relocate to the safest borough in London, expats can choose the neighborhoods to which they want to relocate based on the most common venues in it. For example, if a person is looking for a neighborhood 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 neighborhood with stores and restaurants in a close proximity, then the neighborhoods in the first cluster is suitable. For a family I feel that the neighborhoods in Cluster 4 are more suitable dues to the common venues in that cluster, these neighborhoods have common venues such as Parks, Gym/Fitness centers, Bus Stops, Restaurants, Electronics Stores and Soccer fields which is ideal for a family. The choices of neighborhoods may vary from person to person.

CONCLUSION

To move or not to move….

This project helps a person get a better understanding of the neighborhoods with respect to the most common venues in that neighborhood. 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 neighborhood. 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.