

Applied Data Science

Capstone Final Assignment

The Battle of Neighbourhoods

April 2020

Submitted by

Sayan Mondal

In partial fulfilment for

IBM Data Science Professional Certificate

on



1. Introduction

1.1 Background:

- ❖ Safety is a top concern when moving to a new area. If we don't feel safe in our own home, we're not going to be able to enjoy living there.

1.2 Problem:

- ❖ This project aims to select the safest borough in London based on the total crimes, explore the neighbourhoods of that borough to find the ten most common venues in each neighbourhood and finally cluster the neighbourhoods using k-means.

1.3 Interest:

- ❖ Expats who are considering to relocate to London will be interested to identify the safest borough in London and explore its neighbourhoods and common venues around each neighbourhood.

2. Data Acquisition & Cleaning

2.1 Data Acquisition:

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

- ❖ The first data source of the project uses a London crime data that shows the crime per borough in London
- ❖ The second source of data is scraped from a Wikipedia page that contains the list of London boroughs. This page contains additional information about the boroughs.
- ❖ The third data source is the list of neighbourhoods in the Royal Borough of Kingston upon Thames as found on the Wikipedia page.

2.2 Data Cleaning:

The data cleaning process for each of the three sources of data are done separately.

- ❖ From the London crime data, the crimes during the most recent year (2016) are only selected. The major categories of crime are pivoted to get the total crimes per the boroughs for each major category.
- ❖ The second data is scraped from a Wikipedia page using the Beautiful Soup library in Python. Using this library we can extract the data in the tabular format a shown in the website.
- ❖ The two data sets are merged on the borough names to form a new data set. The purpose of this data set is to visualize the crime rates in each borough and identify the borough with the least crimes recorded during the year 2016.
- ❖ After visualizing the crime in each borough we can find the borough with the lowest crime rate. The third data set is created, with the names of the neighbourhoods and the name of the borough with the latitude and longitude obtained using Google Maps API geocoding.
- ❖ The new data set is used to generate the ten most common venues for each neighbourhood using the Foursquare API, finally using k-means to cluster similar neighbourhoods together.

3. Methodology

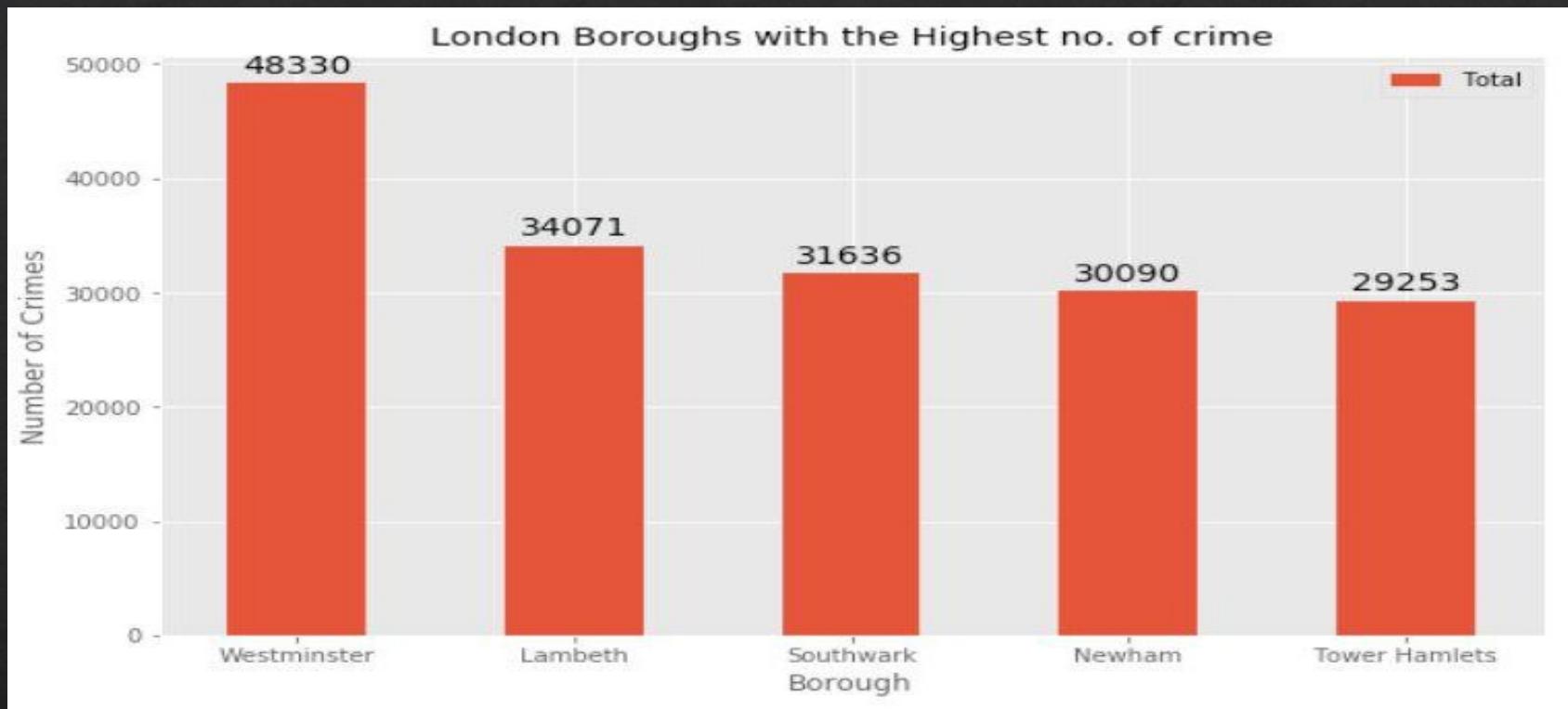
3.1 Exploratory Data Analysis

3.1.1 Statistical summary of crimes

	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

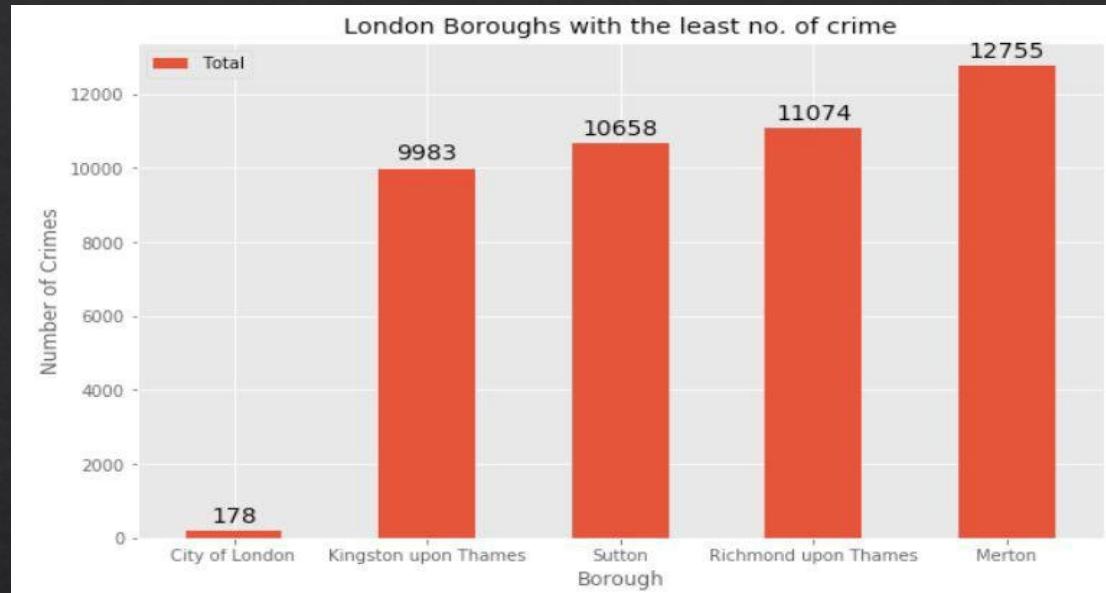
The count for each of the major categories of crime returns the value 33 which is the number of London boroughs. ‘Theft and Handling’ is the highest reported crime during the year 2016 followed by ‘Violence against the person’, ‘Criminal damage’. The lowest recorded crimes are ‘Drugs’, ‘Robbery’ and ‘Other Notifiable Offences’.

3.1.2 Boroughs with the highest crime rates



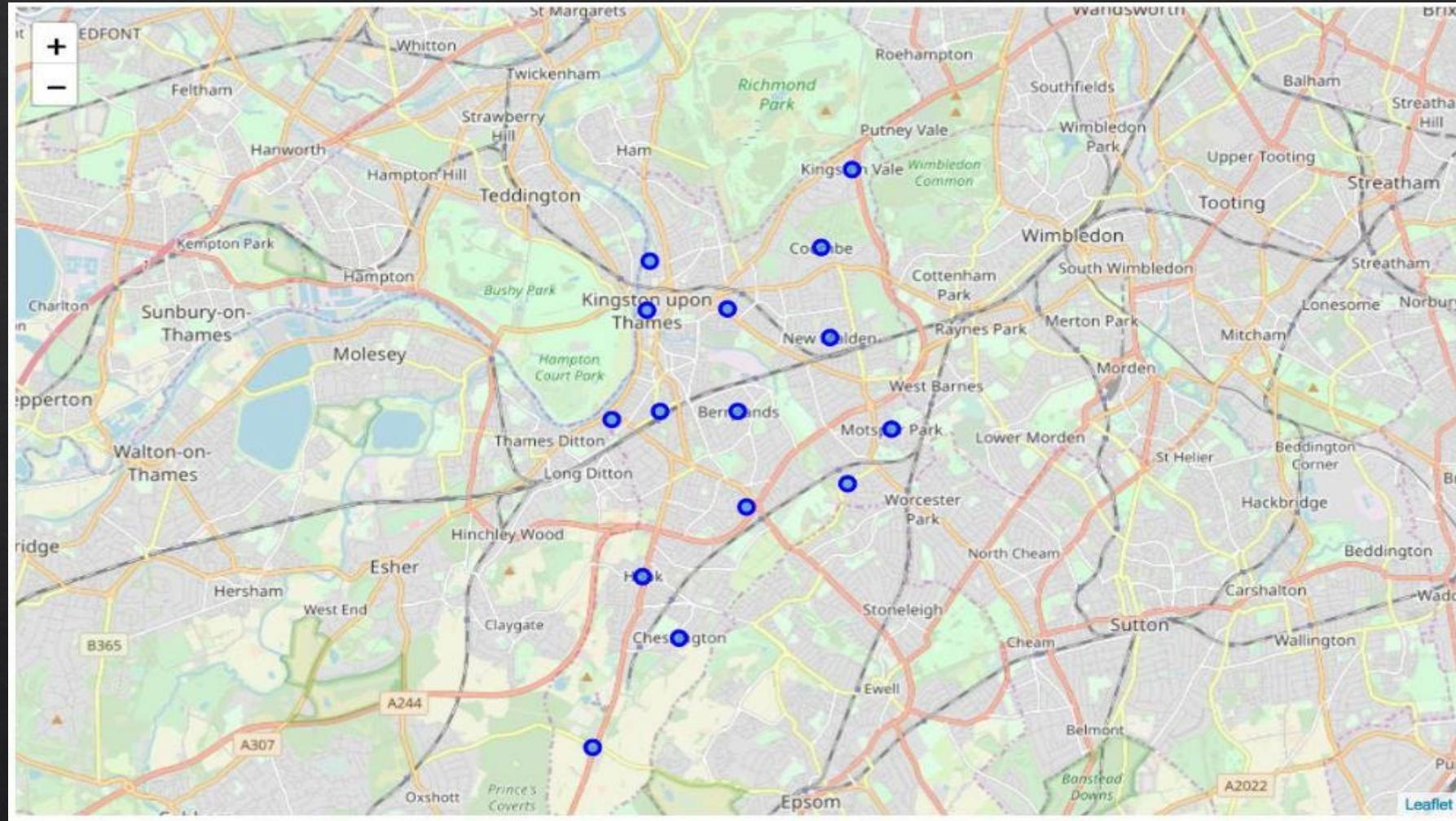
Comparing five boroughs with the highest crime rate during the year 2016, it is evident that Westminster has the highest crimes recorded followed by Lambeth, Southwark, Newham and Tower Hamlets. Westminster has a significantly higher crime rate than the other 4 boroughs.

3.1.3 Boroughs with the lowest crime rates



- ❖ Comparing five boroughs with the lowest crime rates during the year 2016, City of London has the lowest recorded crimes followed by Kingston upon Thames, Sutton, Richmond upon Thames and Merton.
- ❖ City of London has a significantly lower crime rate because it is the 33rd principal division of Greater London but is not a London borough. It has an area of 1.12 square miles and a population of 7000 as of 2013 which suggests that it is a small area.
- ❖ We consider the next borough with the lowest crime rate as the safest borough in London which is Kingston upon Thames.

3.1.4 Neighbourhoods in Kingston upon Thames



There are 15 neighbourhoods in the royal borough of Kingston upon Thames, they are visualized on a map using Folium on Python.

3.2 Modeling

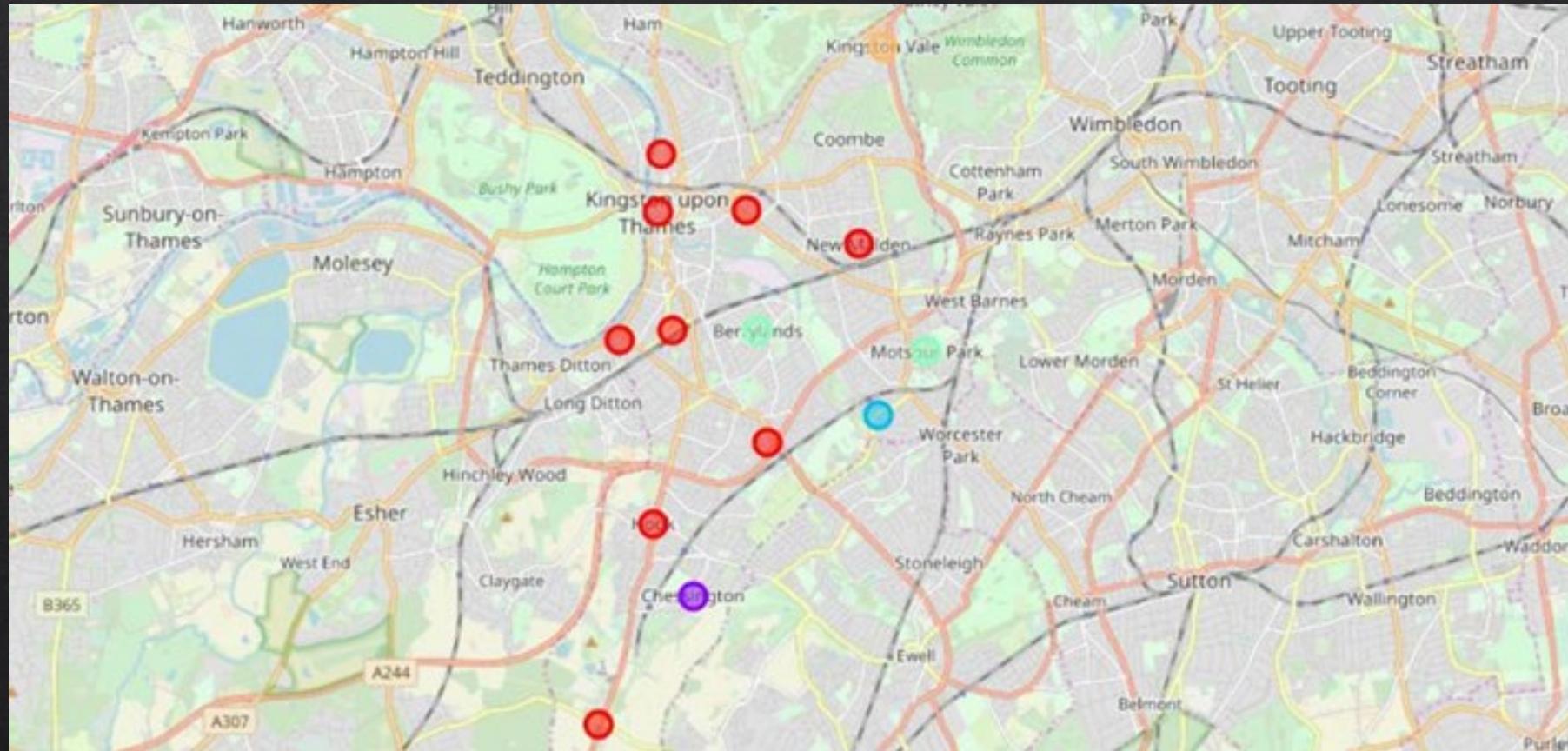
- Using the final dataset containing the neighbourhoods in Kingston upon Thames along with the latitude and longitude, we can find all the venues within a 500 meter radius of each neighbourhood by connecting to the Foursquare API.

	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	Berrylands	51.393781	-0.284802	Cafe Rosa	51.390175	-0.282490	Café
4	Canbury	51.417499	-0.305553	The Boater's Inn	51.418546	-0.305915	Pub

- One hot encoding is done on the Venues data. The Venues data is then grouped by the Neighbourhood and the mean of the venues are calculated, finally the 10 common venues are calculated for each of the neighbourhoods.
- To help people find similar neighbourhoods in the safest borough we cluster similar neighbourhoods using k-means, a form of unsupervised machine learning algorithm that clusters data based on predefined cluster size.
- We use a cluster size of 5 for this project that will cluster the 15 neighbourhoods into 5 clusters. The reason to conduct a k-means clustering is to cluster neighbourhoods with similar venues together so that people can shortlist the area of their interests based on the venues/amenities around each neighbourhood.

4. Results

After running the k-means, we can access each cluster created to see which neighbourhoods were assigned to each of the five clusters. Visualizing the clustered neighbourhoods on a map using the Folium library:



Each cluster is colour coded for the ease of presentation. We can see that majority of the neighbourhood falls in the red cluster which is the first cluster. Three neighbourhoods have their own cluster (blue, purple & yellow), i.e. clusters 2, 3 & 5. The green cluster consists of two neighbourhoods which is the fourth cluster.

Cluster 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	8th Most Common Venue
1	Canbury	Kingston upon Thames	51.417499	-0.305553	0	Pub	Café	Plaza	Fish & Chips Shop	Supermarket	Spa	Shop & Service	Park
4	Hook	Kingston upon Thames	51.367898	-0.307145	0	Bakery	Convenience Store	Indian Restaurant	Fish & Chips Shop	Wine Shop	Food	Electronics Store	Farmers Market
5	Kingston upon Thames	Kingston upon Thames	51.409627	-0.306262	0	Coffee Shop	Café	Burger Joint	Sushi Restaurant	Pub	Record Shop	Cosmetics Shop	Market
7	Malden Rushett	Kingston upon Thames	51.341052	-0.319076	0	Convenience Store	Pub	Garden Center	Restaurant	Fast Food Restaurant	Discount Store	Dry Cleaner	Electronics Store
9	New Malden	Kingston upon Thames	51.405335	-0.263407	0	Gastropub	Gym	Sushi Restaurant	Supermarket	Korean Restaurant	Indian Restaurant	Fish & Chips Shop	Dry Cleaner
10	Norbiton	Kingston upon Thames	51.409999	-0.287396	0	Indian Restaurant	Pub	Food	Italian Restaurant	Platform	Grocery Store	Farmers Market	Dry Cleaner
12	Seething Wells	Kingston upon Thames	51.392642	-0.314366	0	Indian Restaurant	Coffee Shop	Italian Restaurant	Pub	Café	Wine Shop	Fast Food Restaurant	Chinese Restaurant
13	Surbiton	Kingston upon Thames	51.393756	-0.303310	0	Coffee Shop	Pub	Supermarket	Breakfast Spot	Grocery Store	Gastropub	French Restaurant	Train Station
14	Tolworth	Kingston upon Thames	51.378876	-0.282860	0	Grocery Store	Pharmacy	Furniture / Home Store	Train Station	Pizza Place	Discount Store	Coffee Shop	Bus Stop

The first cluster is the biggest one with 9 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, cafes, supermarkets and stores.

Cluster 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	9th Most Common Venue
2 Chessington	Kingston upon Thames	51.358336	-0.298622	1	Fast Food Restaurant	Wine Shop	Golf Course	German Restaurant	Gastropub	Garden Center	Furniture / Home Store	Fried Chicken Joint	French Restaurant

The second cluster has one neighbourhood which consists of venues such as restaurants, golf courses and wine shops.

Cluster 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	8th Most Common Venue	9th Most Common Venue
11 Old Malden	Kingston upon Thames	51.382484	-0.25909	2	Train Station	Pub	Food	Gastropub	Garden Center	Furniture / Home Store	Fried Chicken Joint	French Restaurant	Deli / Bodega

The third cluster has one neighbourhood which consists of venues such as train stations, restaurants and furniture shops.

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

The fourth cluster has two neighbourhoods in it, these neighbourhoods have common venues such as parks, gym/fitness centers, bus stops, restaurants, electronics stores and soccer fields, etc.

Cluster 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	9th Most Common Venue
6	Kingston Vale	Kingston upon Thames	51.43185	-0.258138	4	Grocery Store	Bar	Italian Restaurant	Soccer Field	Garden Center	Furniture / Home Store	Fried Chicken Joint	French Restaurant	Department Store

The fifth cluster has one neighbourhood which consists of venues such as grocery shops, bars, restaurants, furniture shops and department stores.

5. 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 & 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, the neighbourhoods in cluster 4 are more suitable due 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.
- ❖ The preference of venues may vary from person to person, as they select a neighbourhood based on priorities.

6. Conclusion

- ❖ This project helps us 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 safest 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, such as filtering areas based on a predefined budget.