APPLIED DATA SCIENCE

CAPSTONE PROJECT

REPORT

Introduction

Background

It is always a hassle when someone needs to relocate to a new region. It involves a lot of research and investment. So it is necessary for one to be absolutely content with their selection of residence. This selection involves a lot of research and figuring out what the enmities and facilities are around in a particular neighborhood, does it fulfill and desired satisfaction levels and many more. Safety is one of the factors of foremost concern.

Problem

What is the safest area based on total crimes in a city say London, and what are the most popular venues in the said area?

Data

In this project, the data is from a combination of 3 sources.

1. London Crime Data from Kaggle which contains all the crimes committed per borough from the years ranging from 2008 to 2016. For this case, we will use only the data from 2016 as it is the most current. It consists of the following features:
    - lsoa_code – code for Lower Super Output Area in Greater London.
    - borough – common name for London borough.
    - major_category – High level categorization of the committed crime.
    - minor_category – Low level categorization of the committed crime within major category.
    - value – monthly reported count of categorical crime in the given borough.
    - year – year of reported counts.
    - month – month of reported counts.
2. Second source is scraped from a Wikipedia page containing the list of all the London Boroughs.
    - Borough – Names of all 33 boroughs of London.
    - Inner - categorizing the borough as either Inner London or Outer London borough.
    - Status – categorizing the borough as either Royal, City or other.
    - 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: 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. Third data source is the list of Neighborhoods in the Royal Borough of Richmond upon Thames as found on a Wikipedia page. This dataset is created from scratch using the list of neighborhood available on the site, the following are columns:

    - Neighborhood: Name of the neighborhood in the Borough.
    - Borough: Name of the Borough.
    - Latitude: Latitude of the Borough.
    - Longitude: Longitude of the Borough.

Data Cleaning

The data preparation for each of the three sources of data is 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.

	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

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 as shown in the website. After the web scraping, string manipulation is required to get the names of the boroughs in the correct form. This is important because we will be merging the two datasets together using the Borough names.

	Borough	Inner	Status	Local authority	Political control	Headquarters	Area (sq mi)	Population (2013 est)[1]	Co-ordinates	Nr. in map
0	Barking and Dagenham [note 1]	NaN	NaN	Barking and Dagenham London Borough Council	Labour	Town Hall, 1 Town Square	13.93	194352	51°33′39″N 0°09′21″E / 51.5607°N 0.1557°E	25
1	Barnet	NaN	NaN	Barnet London Borough Council	Conservative	North London Business Park, Oakleigh Road South	33.49	369088	51°37′31″N 0°09′06″W / 51.6252°N 0.1517°W	31
2	Bexley	NaN	NaN	Bexley London Borough Council	Conservative	Civic Offices, 2 Watling Street	23.38	236687	51°27′18″N 0°09′02″E / 51.4549°N 0.1505°E	23
3	Brent	NaN	NaN	Brent London Borough Council	Labour	Brent Civic Centre, Engineers Way	16.70	317264	51°33′32″N 0°16′54″W / 51.5588°N 0.2817°W	12
4	Bromley	NaN	NaN	Bromley London Borough Council	Conservative	Civic Centre, Stockwell Close	57.97	317899	51°24′14″N 0°01′11″E / 51.4039°N 0.0198°E	20

The two datasets are merged on the Borough names to form a new dataset that combines the necessary information in one dataset. The purpose of this dataset is to visualize the crime rates in each borough and identify the borough with the least crimes recorded during the year 2016.

Borough	Local authority	Political control	Headquarters	Area (sq mi)	Population (2013 est) [1]	Co-ordinates	Burglary	Criminal Damage	Drugs	Other Notifiable Offences	Robbery	Theft and Handling	Violence Against the Person	Total
Barking and Dagenham	Barking and Dagenham London Borough Council	Labour	Town Hall, 1 Town Square	13.93	194352	51°33′39″N 0°09′21″E / 51.5607°N 0.1557°E	1287	1949	919	378	534	5607	6067	16741
Barnet	Barnet London Borough Council	Conservative	North London Business Park, Oakleigh Road South	33.49	369088	51°37′31″N 0°09′06″W / 51.6252°N 0.1517°W	3402	2183	906	499	464	9731	7499	24684
Bexley	Bexley London Borough Council	Conservative	Civic Offices, 2 Watling Street	23.38	236687	51°27′18″N 0°09′02″E / 51.4549°N 0.1505°E	1123	1673	646	294	209	4392	4503	12840
Brent	Brent London Borough Council	Labour	Brent Civic Centre, Engineers Way	16.70	317264	51°33′32″N 0°16′54″W / 51.5588°N 0.2817°W	2631	2280	2096	536	919	9026	9205	26693
Bromley	Bromley London Borough Council	Conservative	Civic Centre, Stockwell Close	57.97	317899	51°24′14″N 0°01′11″E / 51.4039°N 0.0198°E	2214	2202	728	417	369	7584	6650	20164

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 neighborhoods in the safest borough on Wikipedia. This dataset is created from scratch, the Pandas data frame is created with the names of the neighborhoods and the name of the borough with the latitude and longitude left blank.

	Neighborhood	Borough	Latitude	Longitude
0	Barnes	Richmond upon Thames		
1	Castelnau	Richmond upon Thames		
2	East Sheen	Richmond upon Thames		
3	East Twickenham	Richmond upon Thames		
4	Fulwell	Richmond upon Thames		
5	Ham	Richmond upon Thames		
6	Hampton	Richmond upon Thames		
7	Hampton Hill	Richmond upon Thames		
8	Hampton Wick	Richmond upon Thames		
9	Kew	Richmond upon Thames		
10	Mortlake	Richmond upon Thames		
11	North Sheen	Richmond upon Thames		
12	Petersham	Richmond upon Thames		
13	Richmond	Richmond upon Thames		
14	St Margarets	Richmond upon Thames		
15	Strawberry Hill	Richmond upon Thames		
16	Teddington	Richmond upon Thames		
17	Twickenham	Richmond upon Thames		
18	Whitton	Richmond upon Thames		

The coordinates of the neighborhoods are being obtained using **Google Maps API geocoding** to get the final dataset.

	Neighborhood	Borough	Latitude	Longitude
0	Barnes	Richmond upon Thames	51.471896	-0.238744
1	Castelnau	Richmond upon Thames	51.482695	-0.237688
2	East Sheen	Richmond upon Thames	51.462371	-0.267094
3	East Twickenham	Richmond upon Thames	51.446744	-0.328189
4	Fulwell	Richmond upon Thames	51.433748	-0.349685
5	Ham	Richmond upon Thames	51.434764	-0.309299
6	Hampton	Richmond upon Thames	51.415027	-0.369141
7	Hampton Hill	Richmond upon Thames	51.430996	-0.360211
8	Hampton Wick	Richmond upon Thames	51.414452	-0.312674
9	Kew	Richmond upon Thames	51.480663	-0.291929
10	Mortlake	Richmond upon Thames	51.469887	-0.268523
11	North Sheen	Richmond upon Thames	51.468886	-0.282283
12	Petersham	Richmond upon Thames	51.443691	-0.305293
13	Richmond	Richmond upon Thames	51.461353	-0.303277
14	St Margarets	Richmond upon Thames	51.456709	-0.322412
15	Strawberry Hill	Richmond upon Thames	51.439168	-0.339301
16	Teddington	Richmond upon Thames	51.427784	-0.333653
17	Twickenham	Richmond upon Thames	51.446744	-0.328189
18	Whitton	Richmond upon Thames	51.451169	-0.357976

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

Methodology

Exploratory Data Analysis

Statistical summary of crimes

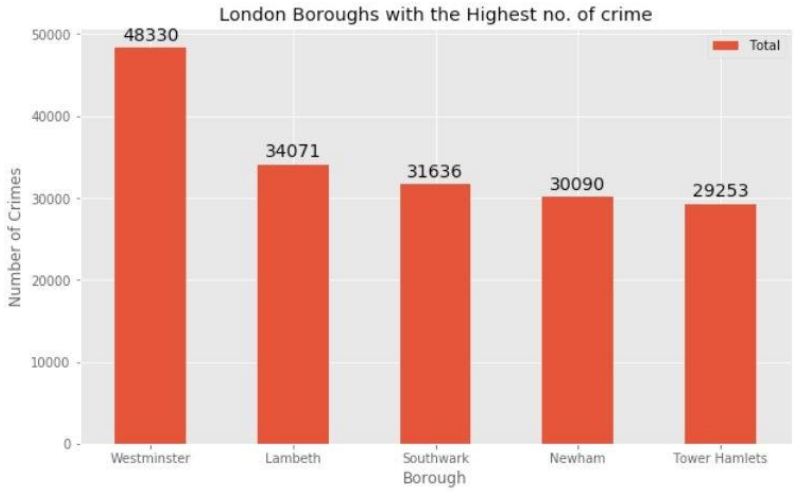
The describe function in python is used to get statistics of the London crime data, this returns the mean, standard deviation, minimum, maximum, 1st quartile (25%), 2nd quartile (50%), and the 3rd quartile (75%) for each of the major categories of crime.

	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 offenses'.

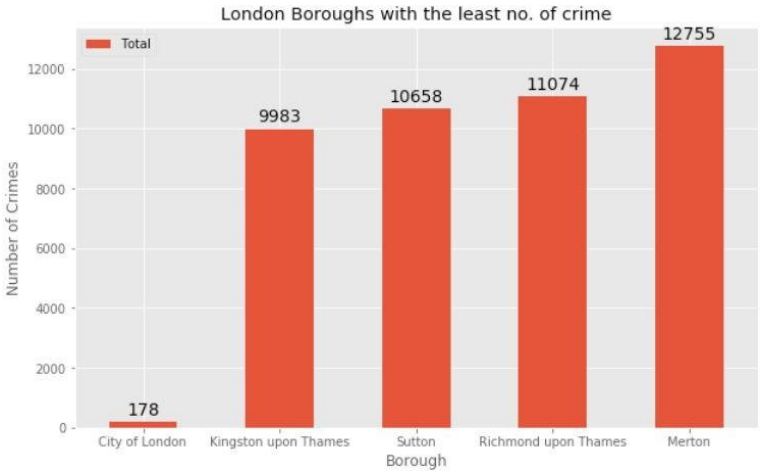
Boroughs with the highest crime rates

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.



Boroughs with the lowest crime rates

Comparing five boroughs with the lowest crime rate during the year 2016, City of London has the lowest recorded crimes followed by Kingston upon Thames, Sutton, Richmond upon Thames and Merton.

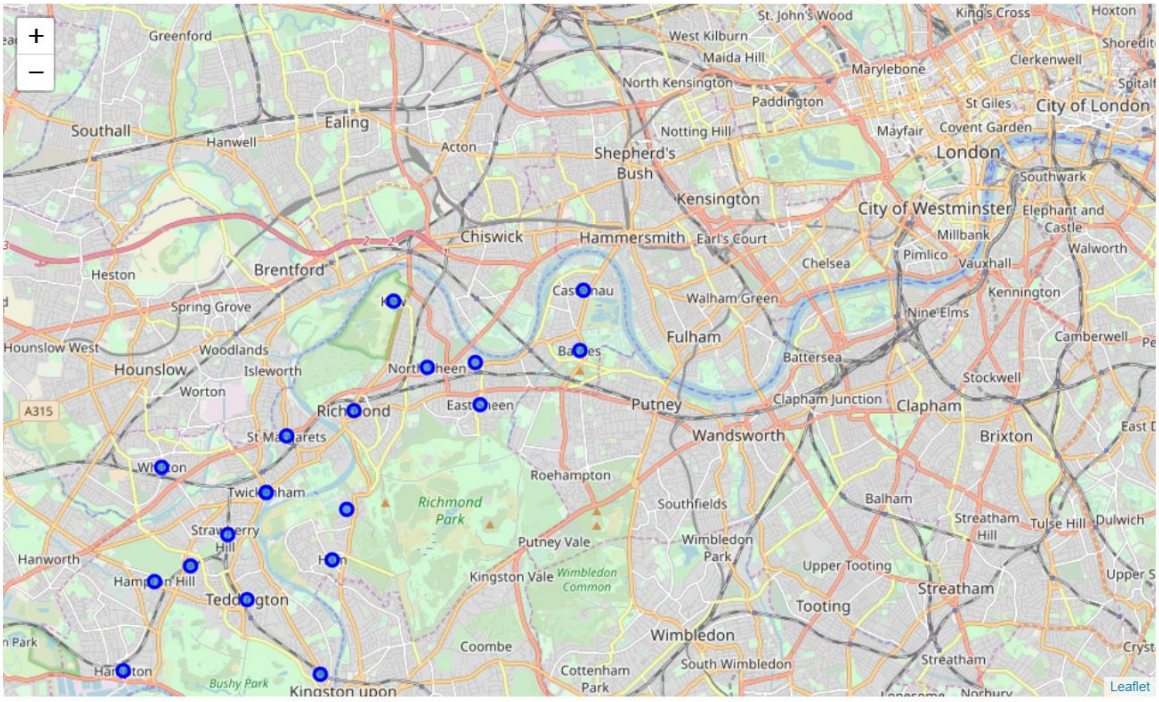


City of London has a significantly lower crime rate because it is the 33rd principal division of Greater London but it is not a London borough. It has an area of 1.12 square miles and a population of 7000 as of 2013 which suggests that it is a small area. Hence we will consider another borough, in this case that is the London Borough of **Richmond upon Thames.**

	Borough	Total	Area (sq mi)	Population (2013 est)[1]
6	City of London	178	1.12	7000

Neighborhoods in Richmond upon Thames

There are 19 neighborhoods in the royal borough of Richmond upon Thames, they are visualized on a map using folium on python.



Modelling

Using the final dataset containing the neighborhoods in Richmond 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.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Barnes	51.471896	-0.238744	Olympic Studios Cafe + Dining Room	51.475158	-0.240333	Indie Movie Theater
1	Barnes	51.471896	-0.238744	ArteChef	51.474705	-0.241282	Pizza Place
2	Barnes	51.471896	-0.238744	Awesome Thai Cuisine	51.474905	-0.240909	Thai Restaurant
3	Barnes	51.471896	-0.238744	Alma Café	51.474880	-0.239207	Breakfast Spot
4	Barnes	51.471896	-0.238744	Olympic Cinema	51.475171	-0.240435	Movie Theater

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

After running the K-means clustering we can access each cluster created to see which neighborhoods were assigned to each of the five clusters. 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	Castelnau	Richmond upon Thames	51.482695	-0.237688	0	Café	French Restaurant	Indian Restaurant	Grocery Store	Italian Restaurant	Recreation Center	Lake	Cosmetics Shop	R
7	Hampton Hill	Richmond upon Thames	51.430996	-0.360211	0	Café	Burger Joint	Pub	Grocery Store	Park	Food & Drink Shop	Cycle Studio	Deli / Bodega	
9	Kew	Richmond upon Thames	51.480663	-0.291929	0	Botanical Garden	Café	Garden	Playground	Wine Shop	Restaurant	Coffee Shop	Historic Site	
15	Strawberry Hill	Richmond upon Thames	51.439168	-0.339301	0	Wine Shop	Train Station	History Museum	Pub	Thai Restaurant	Café	Convenience Store	Fish Market	

The cluster one consists of 4 neighborhoods out of 19 in the borough Richmond upon Thames. Upon closely examining these neighborhoods we can see that the most common venues in these neighborhoods are Cafes, Restaurants and Gardens.

The second cluster has one neighborhood which consists of Venues such as Pubs, Stores, Bakeries, and Cafes.

	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 Mo Comm Ven
5	Ham	Richmond upon Thames	51.434764	-0.309299	1	Pub	Convenience Store	Bakery	Park	Café	Fish Market	Creperie	Cycle Studio	De Bode
12	Petersham	Richmond upon Thames	51.443691	-0.305293	1	Pub	Bus Station	Garden Center	Café	Sports Club	Playground	Fish Market	Creperie	Cy Stu

The third cluster is the biggest one, consisting of 11 out of 19 neighborhoods. The most common places are Parks, Cafes, Pubs and Coffee Shops.

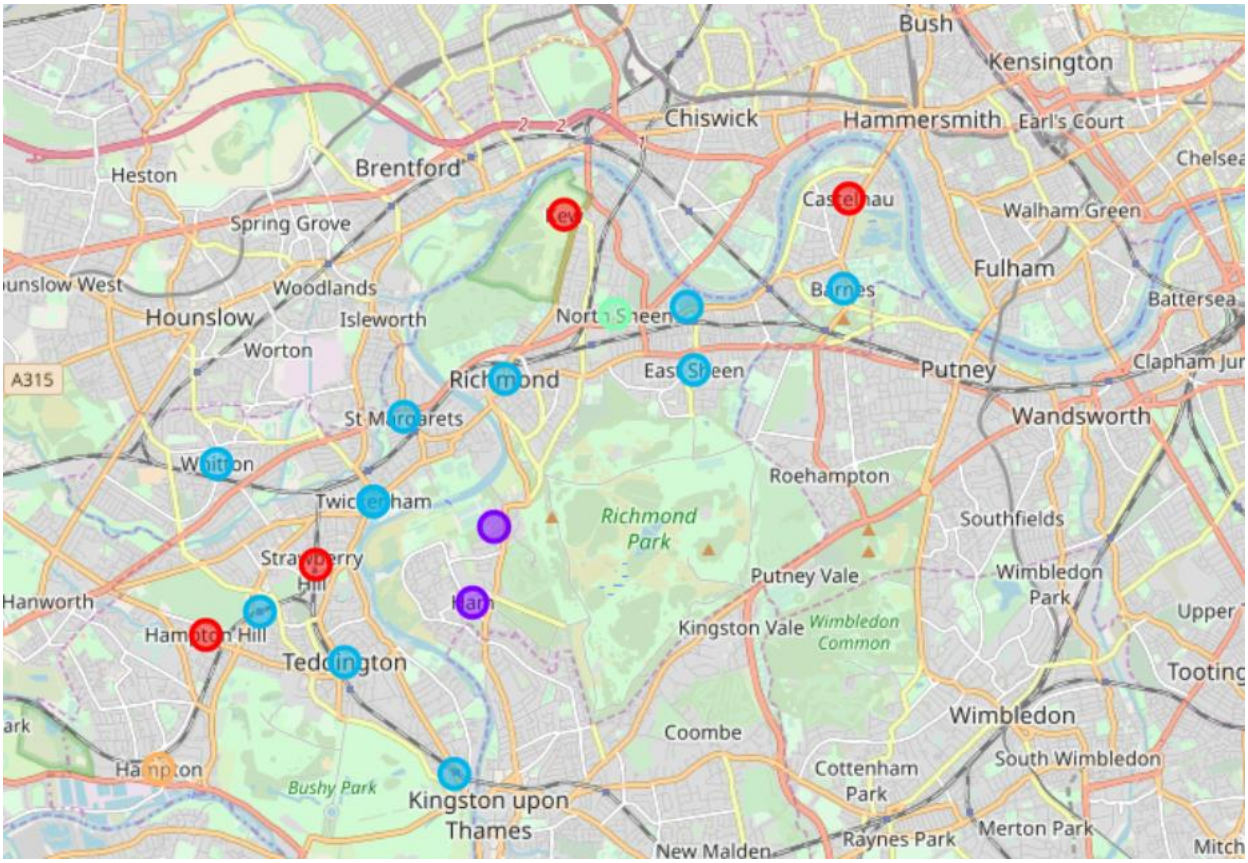
	Neighborhood	Borough	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
0	Barnes	Richmond upon Thames	51.471896	-0.238744	2	Park	Bookstore	Farmers Market	Pub	Café	Food & Drink Shop	Italian Restaurant	Breakfast Spot
2	East Sheen	Richmond upon Thames	51.462371	-0.267094	2	Coffee Shop	Pub	Pizza Place	Plaza	Stationery Store	Italian Restaurant	Middle Eastern Restaurant	Creperie
3	East Twickenham	Richmond upon Thames	51.446744	-0.328189	2	Pub	Coffee Shop	Italian Restaurant	Indian Restaurant	Café	Garden	Japanese Restaurant	Fish Market
4	Fulwell	Richmond upon Thames	51.433748	-0.349685	2	Pizza Place	Garden Center	Pub	Diner	Seafood Restaurant	Chinese Restaurant	Café	Liquor Store
8	Hampton Wick	Richmond upon Thames	51.414452	-0.312674	2	Hotel	Pub	Plaza	Train Station	Coffee Shop	Park	Indian Restaurant	Italian Restaurant
10	Mortlake	Richmond upon Thames	51.469887	-0.268523	2	American Restaurant	Platform	Grocery Store	Golf Course	Pub	Coffee Shop	Tapas Restaurant	Park
13	Richmond	Richmond upon Thames	51.461353	-0.303277	2	Coffee Shop	Italian Restaurant	Pub	Restaurant	Gift Shop	Gastropub	Waterfront	Juice Bar
14	St Margarets	Richmond upon Thames	51.456709	-0.322412	2	Pub	French Restaurant	Comedy Club	Grocery Store	Deli / Bodega	Italian Restaurant	Coffee Shop	Park
16	Teddington	Richmond upon Thames	51.427784	-0.333653	2	Café	Coffee Shop	Pub	Hotel	Mediterranean Restaurant	Grocery Store	Italian Restaurant	Pharmacy
17	Twickenham	Richmond upon Thames	51.446744	-0.328189	2	Pub	Coffee Shop	Italian Restaurant	Indian Restaurant	Café	Garden	Japanese Restaurant	Fish Market
18	Whitton	Richmond upon Thames	51.451169	-0.357976	2	Coffee Shop	Grocery Store	Pizza Place	Bakery	Indian Restaurant	Pub	Café	Steakhouse

The fourth cluster has one neighborhoods, having common venues such as Parks, Supermarket, Coffee Shops etc.

	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 Mos Common Venue
11	North Sheen	Richmond upon Thames	51.468886	-0.282283	3	Park	Supermarket	Coffee Shop	Bus Stop	Food & Drink Shop	Cycle Studio	Deli / Bodega	Diner	Farmers Marke

The fifth cluster has one neighborhood which consists of Venues such as Grocery shops, Train Stations, Restaurants, Parks etc.

	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	C
6	Hampton	Richmond upon Thames	51.415027	-0.369141	4	Grocery Store	Train Station	Pizza Place	Park	Soccer Stadium	Wine Shop	Fish & Chips Shop	Creperie	Cycle Studio	



Each cluster is color coded for the ease of presentation, we can see that majority of the neighborhood falls in the blue cluster which is the third cluster. Three neighborhoods have their own cluster (Red, Purple and Yellow), these are clusters one, two and five. The green cluster consists of one neighborhood, which is the fourth cluster.

Discussion

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 Cluster 2 and 5 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 third 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, Super Markets, Bus Stops and Restaurants which is ideal for a family.

Conclusion

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 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.