Capstone Project - The Battle of Neighbourhoods
Report

Atharva Kulkarni

May, 2020

1. Introduction

1.1 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.

1.2 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 neighbourhoods of that borough to find the 10 most common venues in each neighbourhood and finally cluster the neighbourhoods using k-mean clustering.

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 and 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 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

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 following are the 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 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.

2.2 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

	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

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.

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.

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

	Borough	Inner	Status	Local authority	Political control	Headquarters	Area (sq mi)	Population (2013 est)[1]	Co-ordinates	Nr. in map
0	Barking and Dagenham []	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	Barnet House, 2 Bristol Avenue, Colindale	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 coordinates of the neighbourhoods is be obtained using Google Maps API geocoding to get the final dataset. 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 Statistical summary of crimes

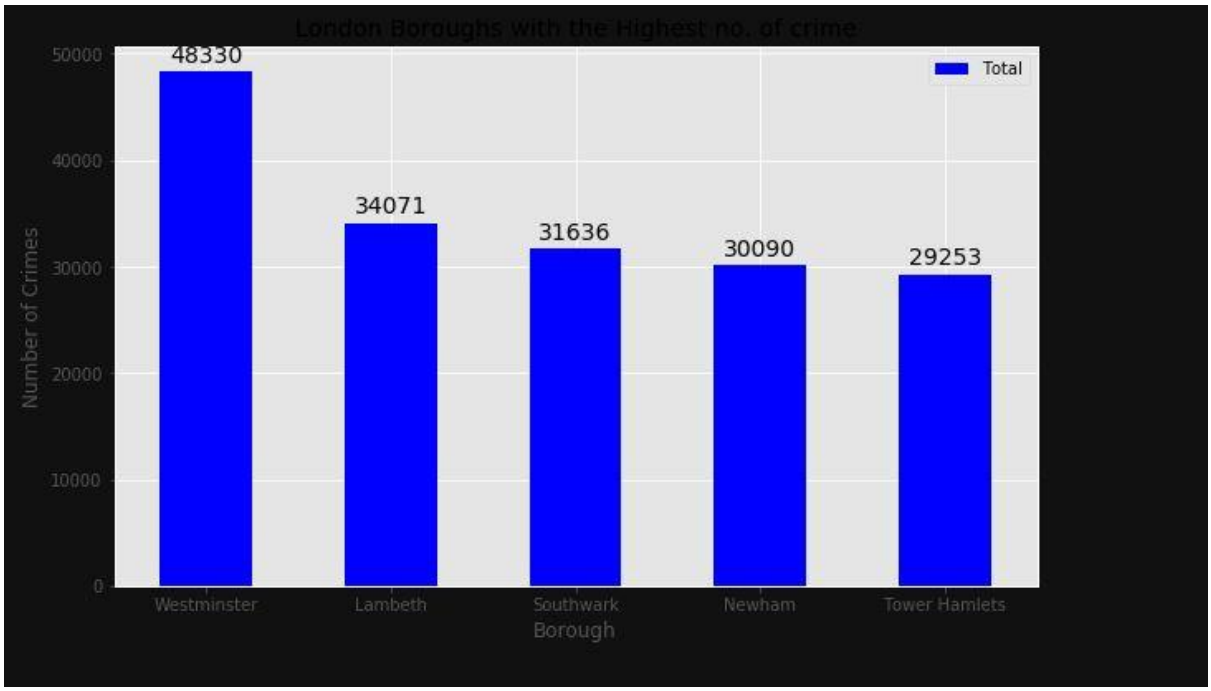
The describe function in python is used to get statistics of the London crime data, this returns the mean, standard deviation, minimum, maximum, 1st quartile (25%), 2nd quartile (50%), and the 3rd quartile (75%) for each of the major categories of crime

	No_of_CrimesBurglary	No_of_CrimesCriminal Damage	No_of_CrimesDrugs	No_of_CrimesOther Notifiable Offences	No_of_CrimesRobbery	No_of_CrimesTheft and Handling	No_of_CrimesViolence 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’.

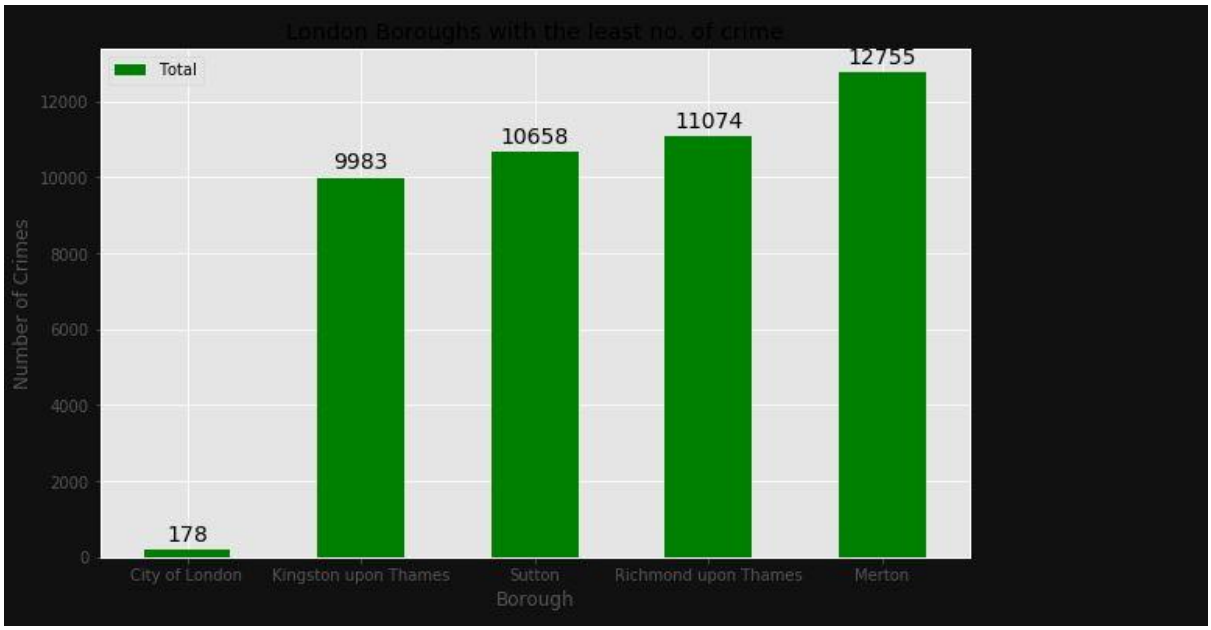
3.1.2 Boroughs with the highest crime rates

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

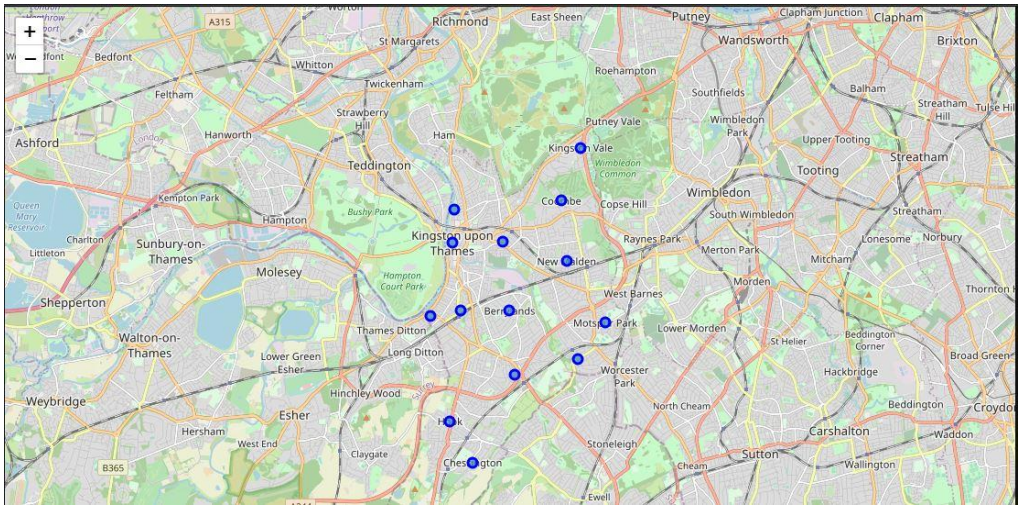
Comparing five boroughs with the lowest crime rate during the year 2016, City of London has the lowest recorded crimes followed by Kingston upon Thames, Sutton, Richmond upon Thames and Merton (see fig 3.1.3).



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 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 visualised on a map using folium on python



3.2 Modelling

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. This returns a json file containing all the venues in each neighbourhood which is converted to a pandas Dataframe. This data frame contains all the venues along with their coordinates and category

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 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 will be clustering similar neighbourhoods 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 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 clustering we can access each cluster created to see which neighbourhoods were assigned to each of the five clusters. Looking into the neighbourhoods

```
[139]: kut_merged[kut_merged['Cluster Labels'] == 0]
```

[139]:

	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 Most Common Venue
6	Kingston Vale	Kingston upon Thames	51.431850	-0.258138	0	Grocery Store	Sandwich Place	Bar	Soccer Field	Department Store	Discount Store	Dry Cleaner	Electronics Store	Farmers Market	Fast Food Restaurant
7	Malden Rushett	Kingston upon Thames	51.341052	-0.319076	0	Grocery Store	Garden Center	Pub	Restaurant	Farmers Market	Deli / Bodega	Department Store	Discount Store	Dry Cleaner	Electronics Store
8	Motspur Park	Kingston upon Thames	51.390985	-0.248898	0	Park	Gym	Soccer Field	Restaurant	Wine Shop	Farmers Market	Department Store	Discount Store	Dry Cleaner	Electronics Store
14	Tolworth	Kingston upon Thames	51.378876	-0.282860	0	Grocery Store	Restaurant	Train Station	Hotel	Indian Restaurant	Italian Restaurant	Discount Store	Coffee Shop	Furniture / Home Store	Pizza Place

The cluster one is the biggest cluster 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, Cafe, Supermarkets, and stores. Looking into the neighbourhoods in the second, third and fifth clusters, we can see these clusters have only one neighbourhood in each. This is because of the unique venues in each of the neighbourhoods, hence they couldn't be clustered into similar neighbourhoods

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

```
[140]: kut_merged[kut_merged['Cluster Labels'] == 1]
```

[140]:

	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 Most Common Venue
2	Chessington	Kingston upon Thames	51.358336	-0.298622	1	Construction & Landscaping	Fish & Chips Shop	Department Store	Discount Store	Dry Cleaner	Electronics Store	Farmers Market	Fast Food Restaurant	Food	Cosmetics Shop

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

```
[141]: kut_merged[kut_merged['Cluster Labels'] == 2]
```

[141]:

	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 Most Common Venue
11	Old Malden	Kingston upon Thames	51.382484	-0.25909	2	Construction & Landscaping	Pub	Food	Train Station	Bakery	Bar	German Restaurant	Gastropub	Garden Center	Furniture / Home Store

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

```
[142]: kut_merged[kut_merged['Cluster Labels'] == 3]
```

[142]:

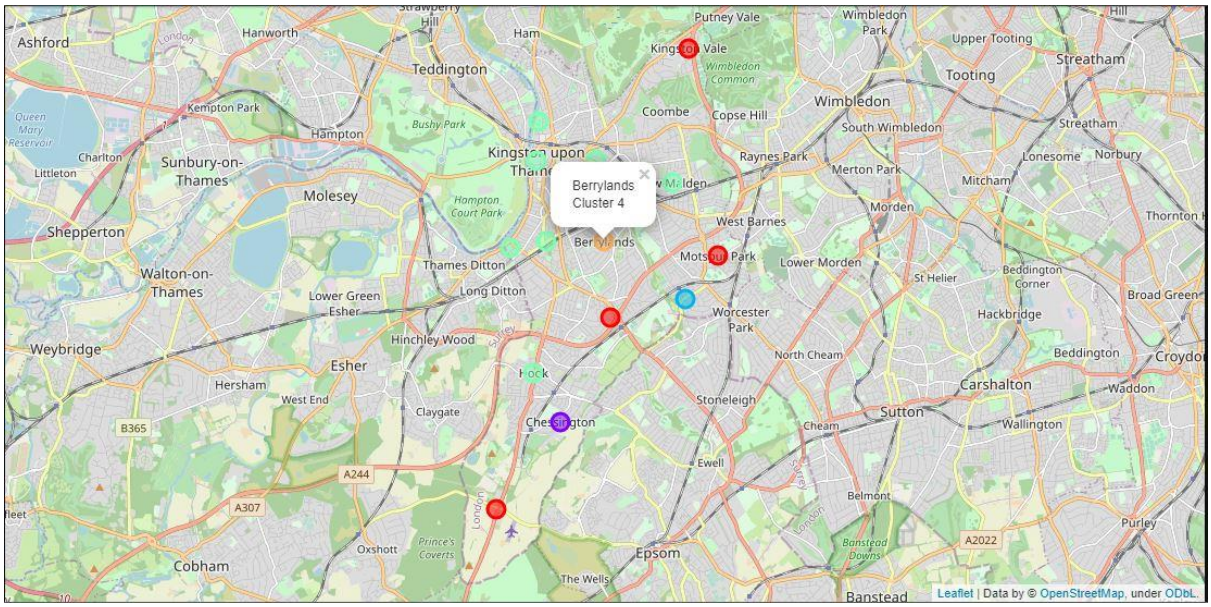
	Neighborhood	Borough	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	Canbury	Kingston upon Thames	51.417499	-0.305553	3	Pub	Park	Fish & Chips Shop	Supermarket	Spa	Gym / Fitness Center	Shop & Service	Plaza	Hotel	Indian Restaurant
4	Hook	Kingston upon Thames	51.367898	-0.307145	3	Indian Restaurant	Fish & Chips Shop	Bakery	Supermarket	Department Store	Discount Store	Dry Cleaner	Electronics Store	Farmers Market	Fast Food Restaurant
5	Kingston upon Thames	Kingston upon Thames	51.409627	-0.306262	3	Coffee Shop	Café	Sushi Restaurant	Burger Joint	Pub	Asian Restaurant	Portuguese Restaurant	French Restaurant	German Restaurant	Electronics Store
9	New Malden	Kingston upon Thames	51.405335	-0.263407	3	Office	Gastropub	Sushi Restaurant	Supermarket	Bar	Chinese Restaurant	Korean Restaurant	Indian Restaurant	Wine Shop	Electronics Store
10	Norbiton	Kingston upon Thames	51.409999	-0.287396	3	Indian Restaurant	Food	Pub	Italian Restaurant	Fried Chicken Joint	Japanese Restaurant	Hotel	Hardware Store	Wine Shop	Pizza Place
12	Seething Wells	Kingston upon Thames	51.392642	-0.314366	3	Indian Restaurant	Coffee Shop	Pub	Café	Italian Restaurant	Restaurant	Hotel	Fast Food Restaurant	Harbor / Marina	Gym / Fitness Center
13	Surbiton	Kingston upon Thames	51.393756	-0.303310	3	Coffee Shop	Pub	Grocery Store	Italian Restaurant	Pharmacy	Thai Restaurant	Tea Room	Gastropub	Train Station	Gym / Fitness Center

The fifth cluster has one neighbourhood which consists of Venues such as Grocery shops, Bars, Restaurants, Furniture shops, and Department stores. We will look into the neighbourhoods in the fourth cluster

```
kut_merged[kut_merged['Cluster Labels'] == 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	10th Most Common Venue
0	Berrylands	Kingston upon Thames	51.393781	-0.284802	4	Gym / Fitness Center	Park	Bus Stop	Wine Shop	Fast Food Restaurant	Discount Store	Dry Cleaner	Electronics Store	Farmers Market	Fish & Chips Shop

Visualising 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 and Yellow), these are clusters two three and five. The green cluster consists of two neighbourhoods which is the 4th cluster.



6.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 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.