Capstone Final Assignment

The Battle of the Neighborhoods

1. Introduction

* **Background:** 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:** 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-means clustering.
* **Interest:** Expats who are considering to relocate to London will be interested to identify the safest borough in London and explore it's neighborhoods and common venues around each neighborhood.

2. Data Acquisition and Cleaning

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 neighborhoods in the Royal Borough of Kingston upon Thames as found on the Wikipedia page.

* **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 year of 2016 were 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 BeautifulSoup library in python. Using this library we can extract the data in the tabular format as 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 neighborhoods and the name of the borough with the latitude and longitude obtained using Geopy geocoding.
  - The new data set 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.

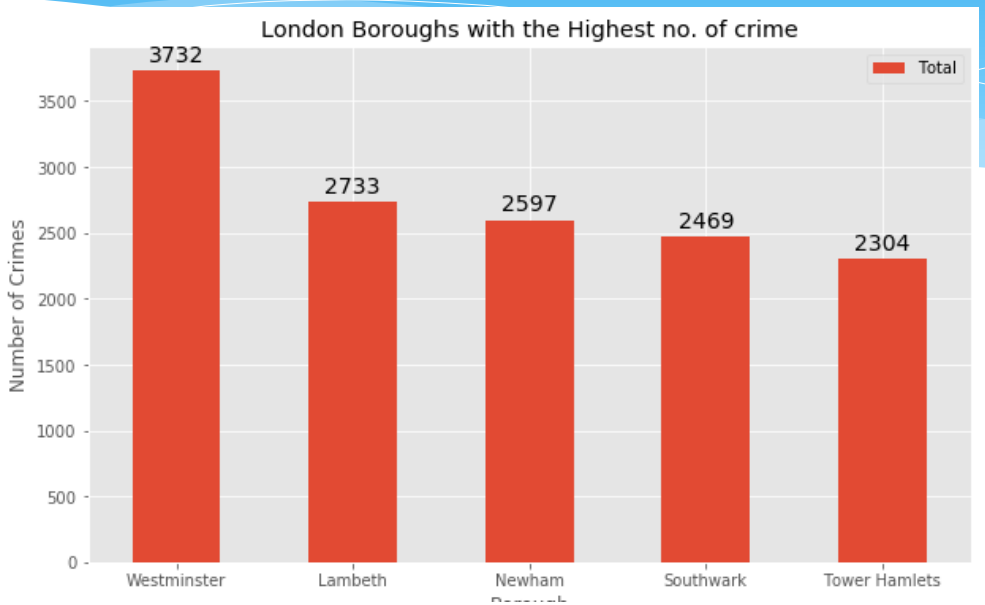
3. Methodology

	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	157.787879	150.787879	96.181818	35.424242	50.242424	693.303030	551.878788	1735.606061
std	56.646247	49.085230	55.441103	15.960949	31.826709	356.684514	203.826942	700.219427
min	0.000000	0.000000	1.000000	0.000000	2.000000	5.000000	1.000000	9.000000
25%	112.000000	108.000000	62.000000	23.000000	24.000000	459.000000	460.000000	1321.000000
50%	161.000000	153.000000	90.000000	34.000000	48.000000	659.000000	583.000000	1770.000000
75%	197.000000	180.000000	140.000000	47.000000	68.000000	839.000000	714.000000	2200.000000
max	264.000000	245.000000	255.000000	75.000000	123.000000	2054.000000	893.000000	3732.000000

* **Exploring Data Analysis**

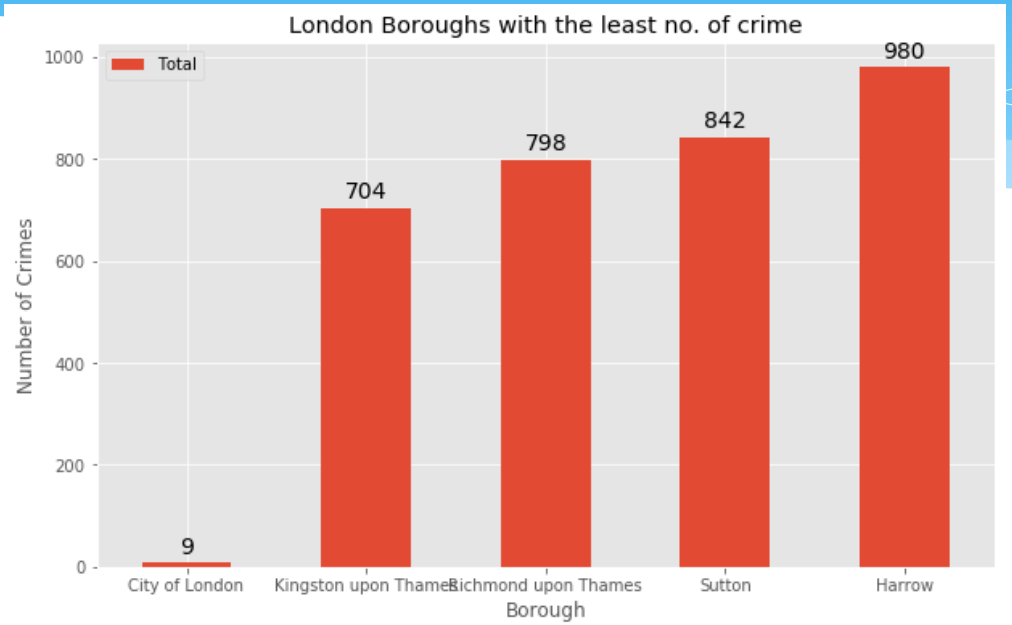
Statistical summary of crimes: The count for each of the major categories of crime returns the value 33 is the number of London boroughs. ‘Theft and Handling’ is the highest reported crime during the year 2016 followed by ‘Violence against the person’ and ‘Criminal damage’. The lowest recorded crimes 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 is evident that Westminster has the highest crimes recorded followed by Lambeth, Newham, Southwark and Tower Hamlets. Westminster has significantly higher crime rate than the other 4 boroughs.

Boroughs with the lowest crime rates

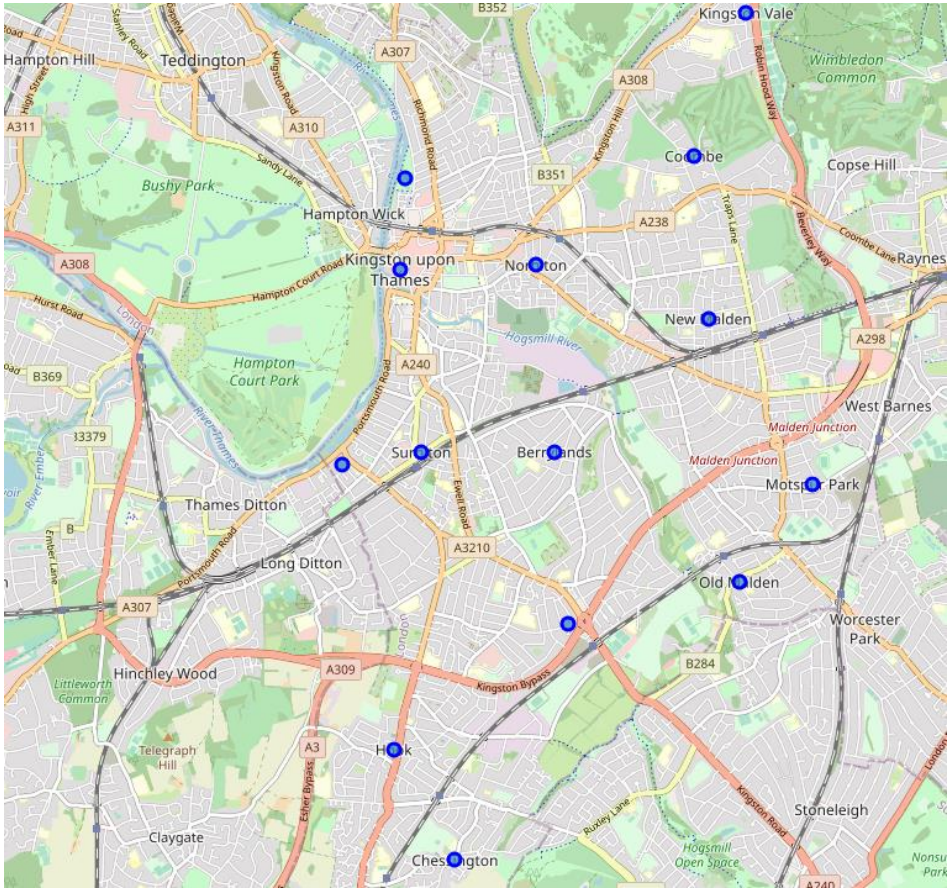


Comparing five boroughs with the lowest crime rate during the year 2016, City of London has the lowest recorded crimes followed by Kingston, Richmond, Sutton and Harrow.

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.

Hence, we will consider the next borough with the lowest crime rate as the safest borough in London which is Kingston upon Thames.

Neighborhoods in Kingston upon Thames



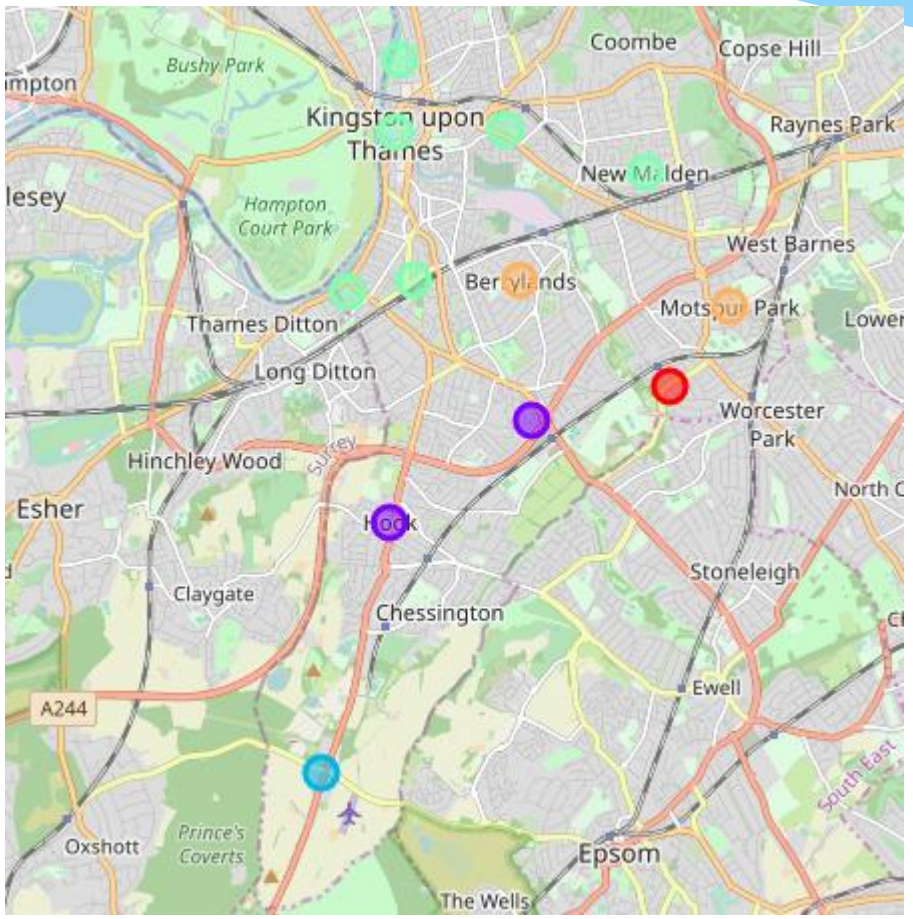
* There are 15 neighborhoods in the royal borough of Kingston upon Thames, they are visualized on a map using Folium on Python.

Modelling

	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	Alexandra Park	51.394230	-0.281206	Park
3	Berrylands	51.393781	-0.284802	K2 Bus Stop	51.392302	-0.281534	Bus Stop
4	Canbury	51.417499	-0.305553	Canbury Gardens	51.417409	-0.305300	Park

* Using the final data set containing neighborhoods in Kingston upon Thames along with the latitude and longitude (screenshot above) we can find all the venues within a 500 meter radius of each neighborhood by connecting to the **Foursquare API**.
* One hot encoding is done on the venues data. 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 cluster 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.

4. 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. Visualizing the clustered neighborhoods on a map using the folium library.

5. Discussion

The aim of this project is to help people who want to relocate to the safest borough in London, expats can chose 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.

6. 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 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.