APPLIED DATA SCIENCE CAPSTONE "THE BATTLE OF NEIGHBORHOODS"

By: Aditya Rajneesh Singh

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-mean clustering.

- **Interest:**

  Expats who are considering to relocate to London will be interested to identify the safest borough in London and explore its neighborhoods and common venues a round 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.
- The third data source is the list of Neighborhoods in the Royal Borough of Kingston upon Thames as found on a Wikipedia page.

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.
- 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. The two datasets are merged on the Borough names to form a new dataset that combines the necessary information in one dataset.
- 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.
- 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.

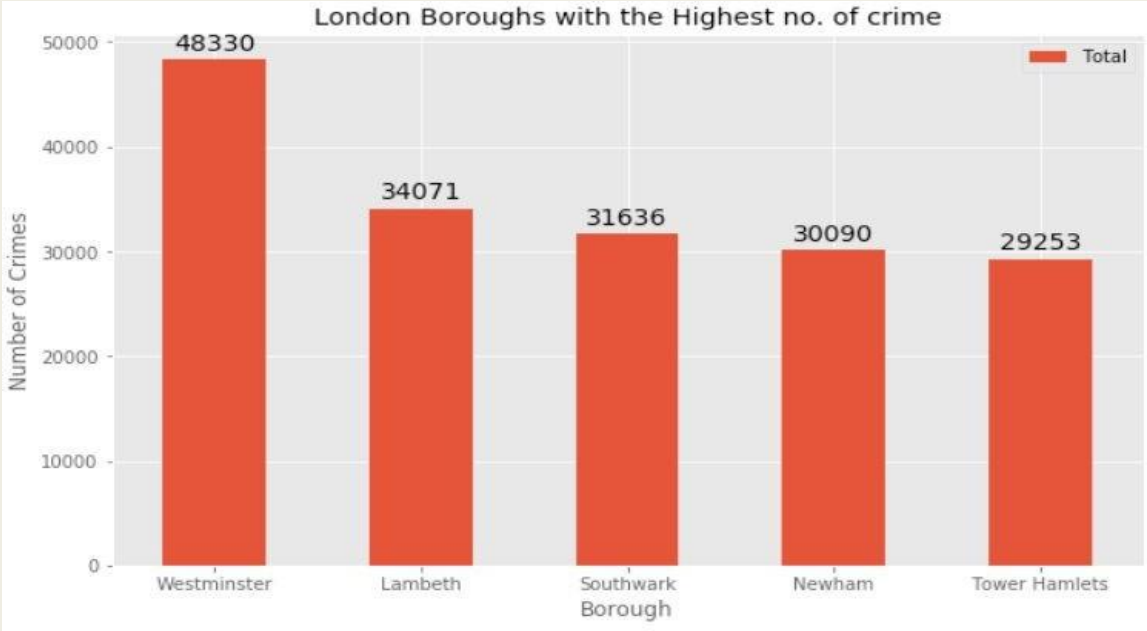
3. Methodology

Exploratory Data Analysis:

	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

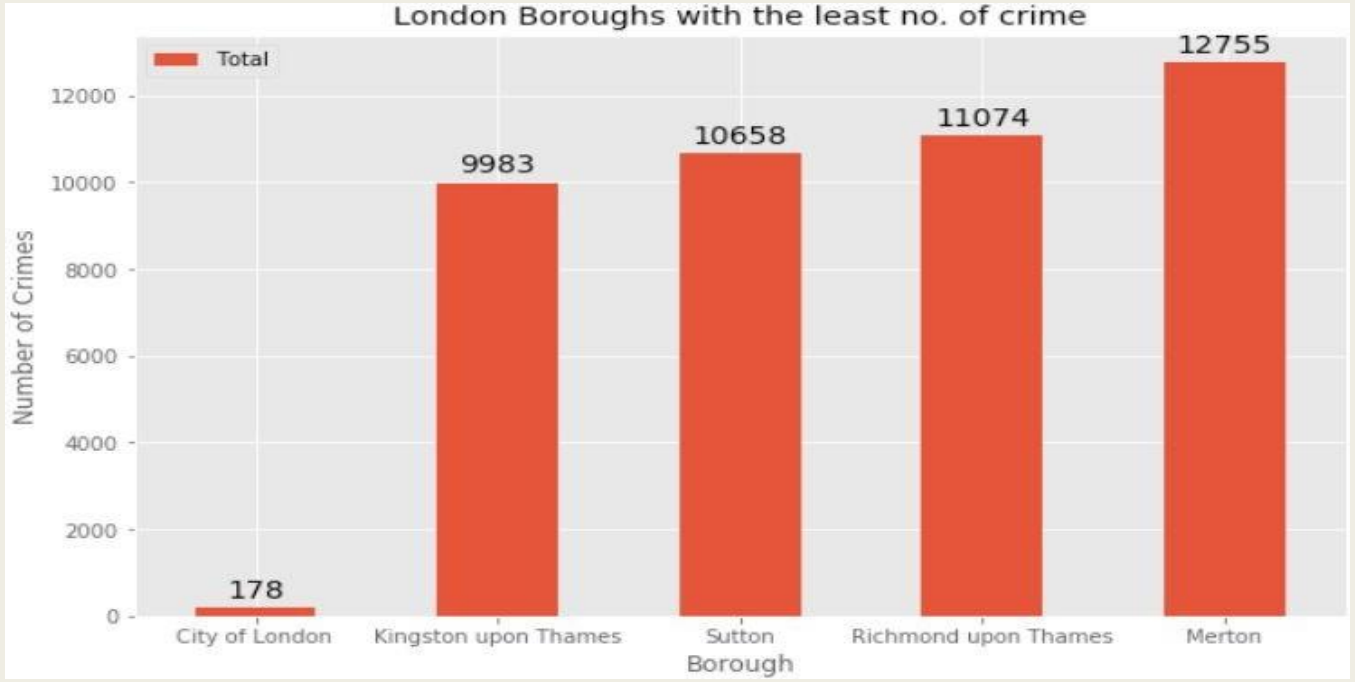
The count for each of the major categories of crime returns the value 33 which is the number of London boroughs. ‘Theft and Handling’ is the highest reported crime during the year 2016 followed by ‘Violence against the person’, ‘Criminal damage’. The lowest recorded crimes are ’Drugs’, ‘Robbery’ and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

Boroughs with highest crime rates:



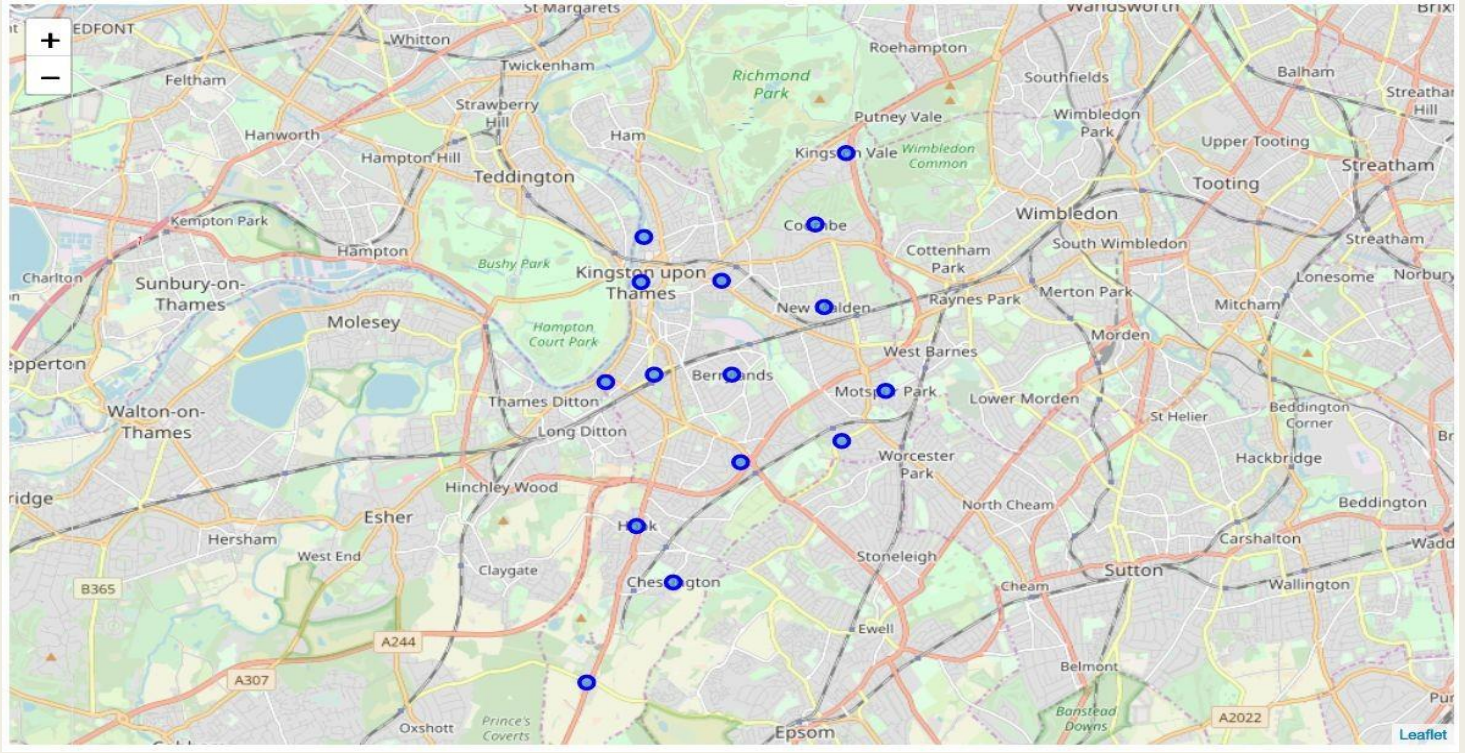
Comparing five boroughs with the highest crime rate during the year 2016 it is evident that Westminster has the highest crimes recorded followed by Lambeth, Southwark, Newham and Tower Hamlets. Westminster has a significantly higher crime rate than other 4 boroughs.

Boroughs with 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.

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:

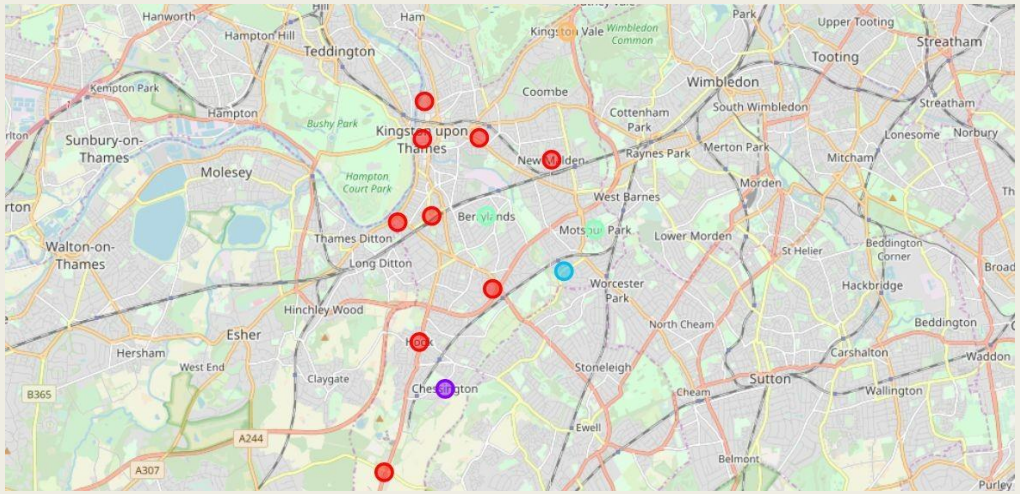
- Using the final dataset containing the neighborhoods in Kingston upon Thames along with the latitude and longitude, we can find all the venues within a 500-meter radius of each neighborhood by connecting to the Foursquare API.

	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. (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).
- We will use a cluster size of 5 for this project that will cluster the 15 neighborhoods into 5 clusters.

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. Looking into the neighborhoods in the first cluster.



- 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.

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 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.

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 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.