Capstone Project – The Battle of Neighbourhoods: Which Borough in London is for Me?

Introduction & The Problem

The capital of the UK, London, experience a growth of 200,000 domestic residents every year. The city attracts all types of people from around the UK and around the world as it offers many opportunities and experiences. The city is growing in population. However, with every large city comes the abundant amount of criminal offenses. The London Datastore has a table consisting of the crime rate for each borough as well as the number of offences.

When people move to a new location, it is in their interest to identify the area they would be most comfortable in. This depends on what type of person but for the project's sake, we will be looking to find the ideal location in London. The project will attempt to identify the safest borough in London based on the rates and the number of incidents, there will then be exploration of the neighbourhoods to find common venues and the neighbourhoods will also be clustered using k-mean methods.

The project will attempt to identify the safest borough in London based on the rates and the number of incidents. There will then be exploration of the neighbourhoods within the safest borough to find common venues and the neighbourhoods will also be clustered using k-mean methods. The project should showcase the features of specific neighbourhoods in the chosen borough. The information showcased should be enough for an individual who is moving to London to find an optimum neighbourhood for themselves.

Data

The London Datastore has a dataset that has the crime rate for each borough from year 1999-00, to 2016-17. The most relevant data will be the data in 2016-17. This has been extracted from the 'data.london.gov' website found with this link: https://data.london.gov.uk/download/recorded_crime_rates/c051c7ec-c3ad-4534-bbfe-6bdfee2ef6bb/crime%20rates.csv

From the wiki page: https://en.wikipedia.org/wiki/List_of_London_boroughs, data about the boroughs will be scraped and processed. This will be used along with Fouresquare API to locate venues nearby and other relevant information on the neighbourhood.

With exploration of the project, I was also required to acquire some data from another wiki page, this was to find a list of neighborhoods in the borough determined safest. The link to this is found here: https://en.wikipedia.org/wiki/List_of_districts_in_the_Royal_Borough_of_Kingston_upon_Thames

The data to be used:

- From the wiki page:
    - **Borough**: the name of the borough
    - **Inner**: Categorising the borough as an inner London or outer London Borough
    - **Status:** Royal, City or other
    - **Local Authority:** The local authority assigned to the borough
    - **Political Control**: The political party controlling the borough
    - **Headquarters:** HQ of the Borough
    - **Area (sq mi):** Area of the Borough in Sq miles
    - **Coordinates:** The latitude and longitude of the borough
- From the London Datastore dataset
    - **Code**: Code of records
    - **Borough**: Name of borough
    - **Year:** year of reporting
    - **Offences:** type of offences
    - **Rate:** Crime rate
    - **Number of offences:** number of offences
- The data that will be requested using the Foresquare API:
    - Venue names
    - Venue location (longitude and Latitude)
    - Venue category

Methodology

Data control

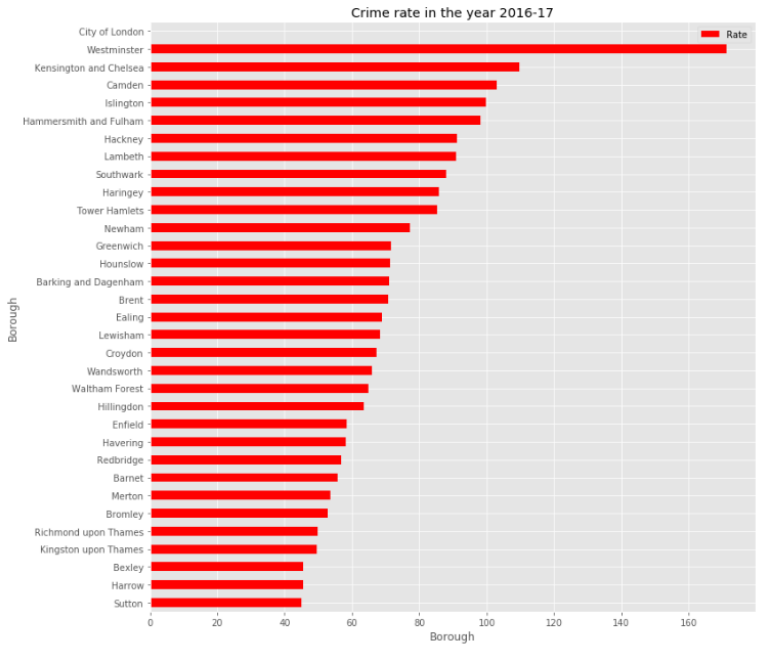
London Dataset: When the data is acquired from the data sources, not all of the data is used. Some data needed to cleaned. From the data from the London Datastore, there was information from the years 1999 all the way to the year 2017. This was too much data and the crime rate from time that early is not relevant to a modern project. '2016-17' being the most recent data, it was used. All rows that weren't associated with that year were dropped. There was also some data that were not indicative of any boroughs. These were labels such as: 'England and Wales', 'Inner London', 'Outer London' and 'Met Police Area'. The finished data frame gave the total recorded offence per borough as well as the total offences of a specific category in each borough.

List of Boroughs from Wiki: Here most of the data acquired from the webpage was relevant. There was some irrelevant data attached to some of the values. The data needed to be looped and checked through to amend these redundancies. Some of the boroughs needed to be renamed too so that they matched with the aforementioned data frame, this was done so when merging the two data frames, there was no issues.

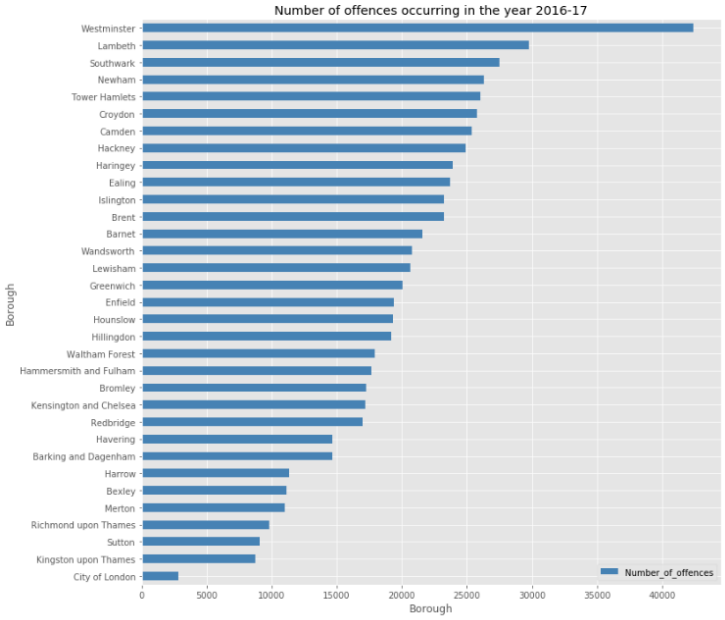
List of Neighbourhoods from Wiki: This was only used to see the name of the streets. They were manually inputted into an array that was used to make the data frame.

Exploratory Data Analytics

The main analysis was ordering the data and determining the safest and least safe boroughs in London. This was done with the rate:



And with the number of offences:



The graph which used the ‘number of offences’ column was more relevant than the graph created using the ‘rate’ column. This was because the rate was not explained from the data source and was unclear what it indicated.

Evidently, the plot shows that the boroughs with the most crime occurrences in the year 2016-17 is Westminster, followed by Lambeth, Southwark, Newham and Tower Hamlets. Westminster had a much larger crime rate than the others.

The boroughs with the least occurring crime were City of London followed by Kingston upon Thames, Sutton, Richmond upon Thames and Merton. The City of London is not actually a borough of London and cannot be considered a viable option when determining an optimal borough. This is probably an explanation for the much lower crime rate recorded. The borough with the lowest crime rate is therefore Kingston upon Thames.

Modelling

The 15 neighbourhood's coordinates located in Kingston upon Thames as well as the Fouresquare API were used to find venues in the vicinity of each neighbourhood. This gave us a JSON file that was converted into a pandas data frame. The data frame had all the venues along with their category, coordinates and which neighbourhood they were close to.

One hot encoding was used on the data – this helped with the machine learning grouping process. The data on the venues was grouped by the Neighbourhood and the mean of the venues was calculated. The 10 most common venues are calculated for each of the venues. The clustering process involved clustering similar neighbourhoods using K – Means clustering that clustered the data based on a size of 5. The clustering resulted in 5 clusters that held neighbourhoods similar to each other based on the venues.

Results

Cluster 1: The first cluster contains one neighbourhood from Kingston upon Thames, we can see that the most common venue is Grocery Store, followed by more recreational venues, perhaps the neighbourhood would be good for a family.

	Neighborhood	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
7	Malden Rushett	51.341052	-0.319076	3.0	Grocery Store	Pub	Garden Center	Restaurant	Fish & Chips Shop	Department Store	Dry Cleaner	Electronics Store	Farmers Market	Fast Food Restaurant

Cluster 2: This cluster contains a large number of the neighbourhoods in the borough, these venues that are popular in these neighbourhoods are the food venues and recreational venues.

Neighborhood	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 Com V
Canbury	51.417499	-0.305553	1.0	Pub	Shop & Service	Supermarket	Plaza	Café	Indian Restaurant	Hotel	Spa	Park	F (
Hook	51.367898	-0.307145	1.0	Bakery	Indian Restaurant	Supermarket	Fish & Chips Shop	Yoga Studio	Food	Department Store	Dry Cleaner	Electronics Store	Far M
Kingston upon Thames	51.409627	-0.306262	1.0	Café	Coffee Shop	Sushi Restaurant	Burger Joint	Pub	Asian Restaurant	Department Store	Market	Mexican Restaurant	Electr :
New Malden	51.405335	-0.263407	1.0	Indian Restaurant	Gastropub	Chinese Restaurant	Bar	Sushi Restaurant	Supermarket	Korean Restaurant	Fish & Chips Shop	Dry Cleaner	Electr :
Norbiton	51.409999	-0.287396	1.0	Food	Indian Restaurant	Italian Restaurant	Pub	Pizza Place	Rental Car Location	Wine Shop	Dry Cleaner	Hardware Store	
Old Malden	51.382484	-0.259090	1.0	Food	Indian Restaurant	Pub	Construction & Landscaping	Train Station	Gastropub	Garden Center	Furniture / Home Store	Fried Chicken Joint	Fr Resta
Seething Wells	51.392642	-0.314366	1.0	Indian Restaurant	Pub	Coffee Shop	Yoga Studio	Café	Chinese Restaurant	Fast Food Restaurant	Fish & Chips Shop	Golf Course	(Fi C
Surbiton	51.393756	-0.303310	1.0	Coffee Shop	Pub	Italian Restaurant	Pharmacy	Grocery Store	Gastropub	Bistro	Pizza Place	Hotel	Far M
Tolworth	51.378876	-0.282860	1.0	Grocery Store	Bowling Alley	Bus Stop	Furniture / Home Store	Train Station	Coffee Shop	Thai Restaurant	Pharmacy	Pizza Place	S

Cluster 3: This cluster only contains one of neighbourhoods. This cluster's most popular venue is Tea room, then Yoga Studio then food. This is clearly quite different from the neighbourhoods seen in cluster 2.

	Neighborhood	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
3	Coombe	51.41945	-0.265398	2.0	Tea Room	Yoga Studio	Food	Department Store	Dry Cleaner	Electronics Store	Farmers Market	Fast Food Restaurant	Fish & Chips Shop	French Restaurant

Cluster 4: Another cluster only containing one of the neighbourhoods from the borough. This neighbourhood consists of venues such as grocery stores, pubs, garden centres and restaurants.

	Neighborhood	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
7	Malden Rushett	51.341052	-0.319076	3.0	Grocery Store	Pub	Garden Center	Restaurant	Fish & Chips Shop	Department Store	Dry Cleaner	Electronics Store	Farmers Market	Fast Food Restaurant

Cluster 5: This cluster has 2 neighbourhoods. This cluster showcases neighbourhoods that are similar in terms of fitness venues. Gyms and fitness centres along with parks are the most popular venues followed by other venues.

	Neighborhood	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	51.393781	-0.284802	4.0	Gym / Fitness Center	Park	Bus Stop	Yoga Studio	Fish & Chips Shop	Dry Cleaner	Electronics Store	Farmers Market	Fast Food Restaurant	Food
8	Motspur Park	51.390985	-0.248898	4.0	Gym	Soccer Field	Park	Bus Stop	Department Store	Dry Cleaner	Electronics Store	Farmers Market	Fast Food Restaurant	Fish & Chips Shop

Discussion

The objective here was to aid individuals in the choice of living given the idea that the people are crime averse. The project should show to the people moving to London that the safest borough in London is Kingston upon Thames (according to the London Dataset data). The project then furthers creates clusters around the neighbourhoods existing in Kingston upon Thames, this would allow for an easier understanding of the neighbourhood and therefore a better decision on where their ideal or optimal neighbourhood would be.

Conclusion

A person is better equipped to make decisions on neighbourhoods in relevance to the common surrounding venues. This is an example of utilising machine learning to solve a real-world problem/task. The potential futures of this project could be taking on other factors so that the person gets an even better understanding of the neighbourhoods as well as being better equipped to make decisions.