**Capstone Project – The Battle of Neighborhoods Report**

By Ghazali Akmal Rabbani

1. **INTRODUCTION**
   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. **Problem**

The crime statistics dataset of London found on Kaggle has crimes in each Boroughs of London from 2008 to 2019. The year 2019 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 neighborhoods of that borough to find the 10 most common venues in each neighborhood and finally cluster the neighborhoods using k-mean clustering.

* 1. **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 around each neighborhood.

1. **DATA ACQUISITION AND CLEANING**
   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 :

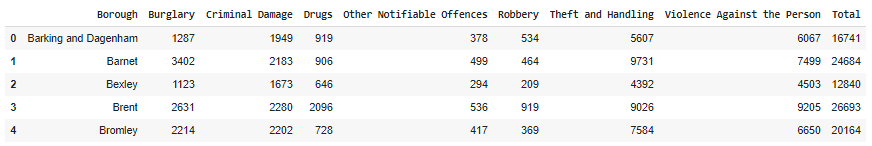
* **lsoa\_code:** code for Lower Super Output Area in Greater London
* **borough:** Common name for London borough
* **major\_category:** High level categorization of crime
* **minor\_category:** Low level categorization of crime with major category
* **value:** monthly reported count of categorical crime in given borough
* **year:** Year of reported counts, 2008-2019
* **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 Borough
* **Area (sq ml):** Area of the borough in square miles
* **Population (2019 est)[1]:** The population in the borough recorded during the year 2019
* **Co-ordinates**: The latitude and longitude of the boroughs
* **Nr. In map:** The number assigned to each borough to represent visually on a map.

The third data source is the list of Neighborhoods in the Royal Borough of Kingston 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
  1. **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 (2019) are only selected. The major categories of crime are pivoted to get the total crimes per the boroughs for each major category.

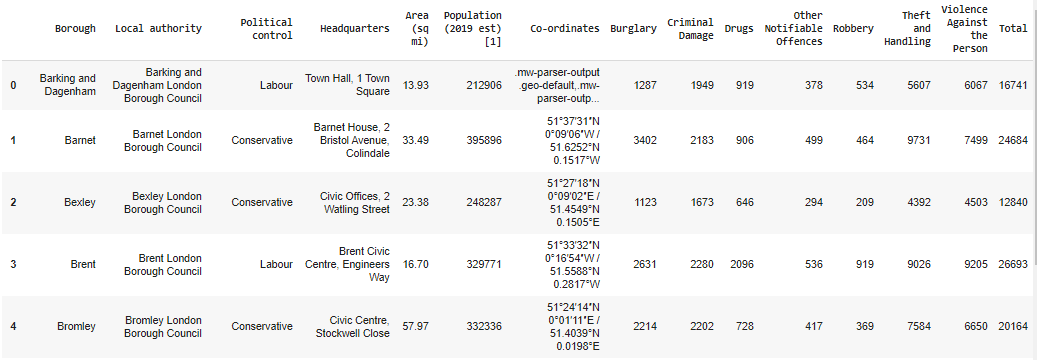
*Figure. London Crime Data After Preprocessing*

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.



*Figure. List of London Boroughs*

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 2019.



*Figure. London Borough Crime*

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.



*Figure. Neighborhoods of the safest borough*

The coordinates of the neighborhoods is be obtained using Google Maps API geocoding to get the final dataset.



*Figure.Neighborhoods of the safest borough*

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.

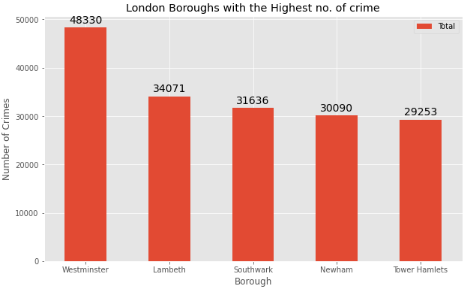
1. **METHODOLOGY**
   1. **Exploratory Data Analysis**
      1. **Statistical Summary of Crimes**

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

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 2019 followed by ‘Violence against the person’, ‘Criminal damage’. The lowest recorded crimes are ’Drugs’, ‘Robbery’ and ‘Other Notifiable offenses’.

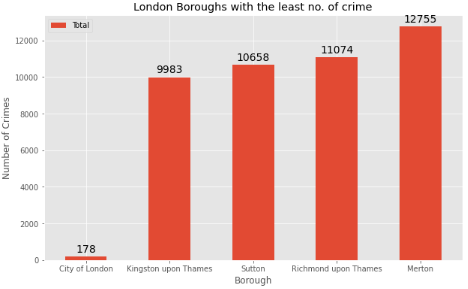
* + 1. **Boroughs with the highest crime rates**

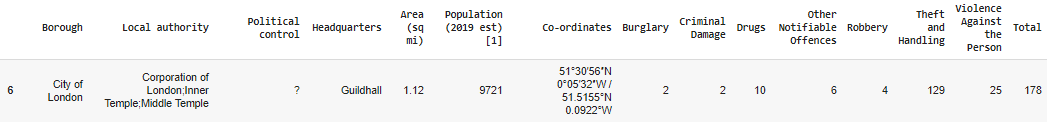
Comparing five boroughs with the highest crime rate during the year 2019 it is evident that Westminster has the highest crimes recorded followed by Lambeth, Southwark, Newham, and Tower Hamlets.



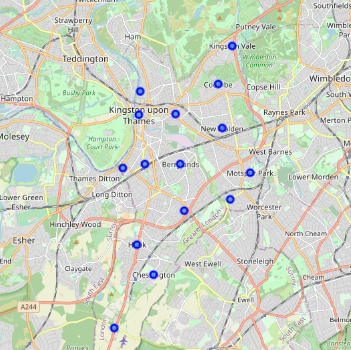
* + 1. **Boroughs with the lowest crimes rates**

Comparing five boroughs with the lowest crime rate during the year 2019, City of London has the lowest recorded crimes followed by Kingston upon Thames, Sutton, Richmond upon Thames and Merton.



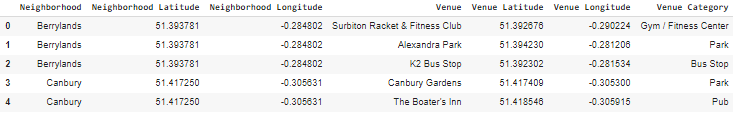
City of London has a significantly lower crime rate because it i is the 33rd principal division of Greater London but it is not a London borough. It has an area of 1.12 square miles and a population of 7000 as of 2019 which suggests that it is a small area. Hence we will consider the next borough with the lowest crime rate as the safest borough in London which is Kingston upon Thames.

* + 1. **Neighborhoods in Kingston upon Thames**



There are 15 neighborhoods in the royal borough of Kingston upon Thames, they are visualised on a map using folium on python

* 1. **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. This returns a json file containing all the venues in each neighborhood 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 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.

1. **RESULT**

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.



Figure. CLUSTER 1

The cluster one is the biggest cluster with 7 neighborhoods in the borough Kingston upon Thames. Upon closely examining these neighborhoods we can see that the most common venues in these neighborhoods are Restaurants, Pubs, Cafe, Supermarkets, and stores.

Looking into the neighborhoods in the fourth and fifth clusters, we can see these clusters have only one neighborhood in each. This is because of the unique venues in each of the neighborhoods, hence they couldn't be clustered into similar neighborhoods



Figure. CLUSTER 4

The forth cluster has 1 neighborhood which consists of Venues such as Store, Bar and Sandwich Place



Figure. CLUSTER 5

The fifth cluster has 1 neighborhood which consists of Venues such as Train Station, Coffee Shop, and Pub. And also There are the others two cluster in London Neighborhoods, which are Second Cluster and Third Cluster.

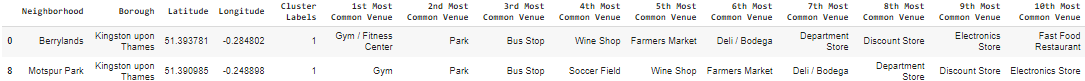


Figure. CLUSTER 2

The second cluster has 2 neighborhood which consists of Venues such as Gym, Store and Public places like Park, Bus Stop, and Soccer Field.

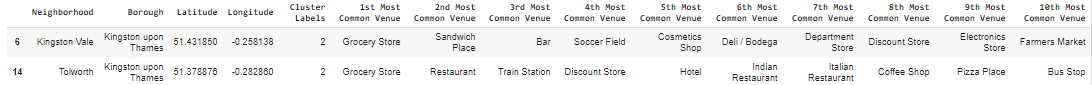


Figure. CLUSTER 3

And the last one is the third cluster which has 2 neighborhood which consists of Venues such as Grocery Store, Bar, Sandwich Place, and Restaurant.

1. **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 public spaces and transportations we can see that **Clusters 2** have Gym, Park and Bus Station 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 **Cluster 1, 3 and 4** is suitable.
* For a family I feel that the neighborhoods in **Cluster 5** are more suitable dues to the common venues in that cluster, these neighborhoods have common venues such as Gastropub, Gym, Supermarket and Store which is ideal for a family who want to have great life experience from daily life needed to entertainment

1. **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