Capstone Project – The Battle of Neighbourhood – Presentation

By Navid Farha

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 around each neighborhood.

2.Data Aquisation 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 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.

•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 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 Google Maps API 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

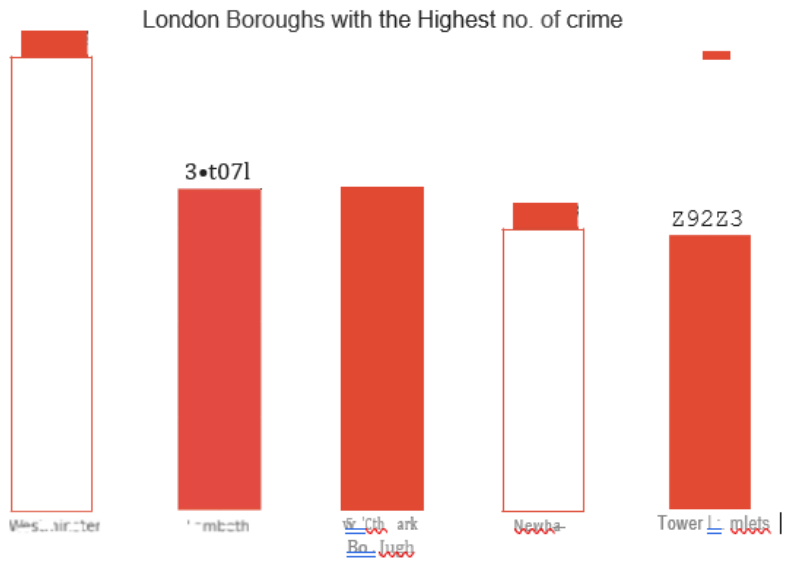
Exploratory Data Analysis

Statistical summary of crimes

	Burglary	Criminal Damage	Drugs	Other Notifiable Offences	Robbery	Theft and Handling	Violence Against the Person	1bt•l
mean	2068.24Z424	1941.54545S	117B.212121	47B.060B	682.666007	8B13.121212	7041.848480 22SA.B96970	
std	737.448644	825.207070	586.406418	229.288898	441.425300	4620.565054	2519.601SS1 8828 748	
(zdn	2.0000€0	2.00000 I	10.0€I0000	6.000000	4.000000	128.£ IOOIXI	25.000000 178.000000	
06X 1S91.0000€€t		18 DQAXXEI	743.£XAI0IXI	480.00tXXAI	SOB.IAI000€I	5818.IAI00IXI	S098.00IXXAI 18B3Z. XXAI00	
5096 2071.0000€0		1880.00000 I	1£B8.0€I0000	551.00IXAI0	038.£XXI00€I	8925.000000	740B.00IXXAI 22730.1 I000	
y696 2831.0000€Q		ZZS1.QAXAXI	1617.£I£I00IXI	1305.00tXXXI	1822.IXI000€I	10780./AI000LI	8832.000tAXI 27174. XI000	
max	3402.000000	3219.000000	2738.000000	378.000000	377.000000	27520.000000	10834.00IXAXI 48330.TOO	

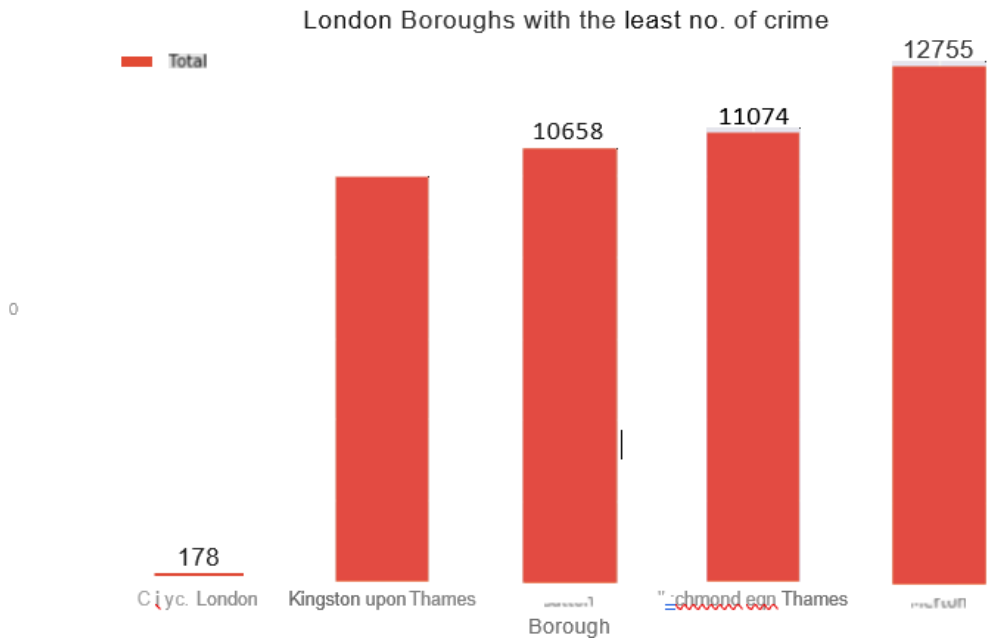
The count for each of the major categories of crime returns the value 33 which is the number of London boroughs. ‘Theft and Handling’ is the highest reported crime during the year 2016 followed by ’Violence against the person', ’Criminal damage’. The lowest recorded crimes are ’Drugs', ’Robbery’ and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

Boroughs with 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.

Boroughs with the lowest crime rates

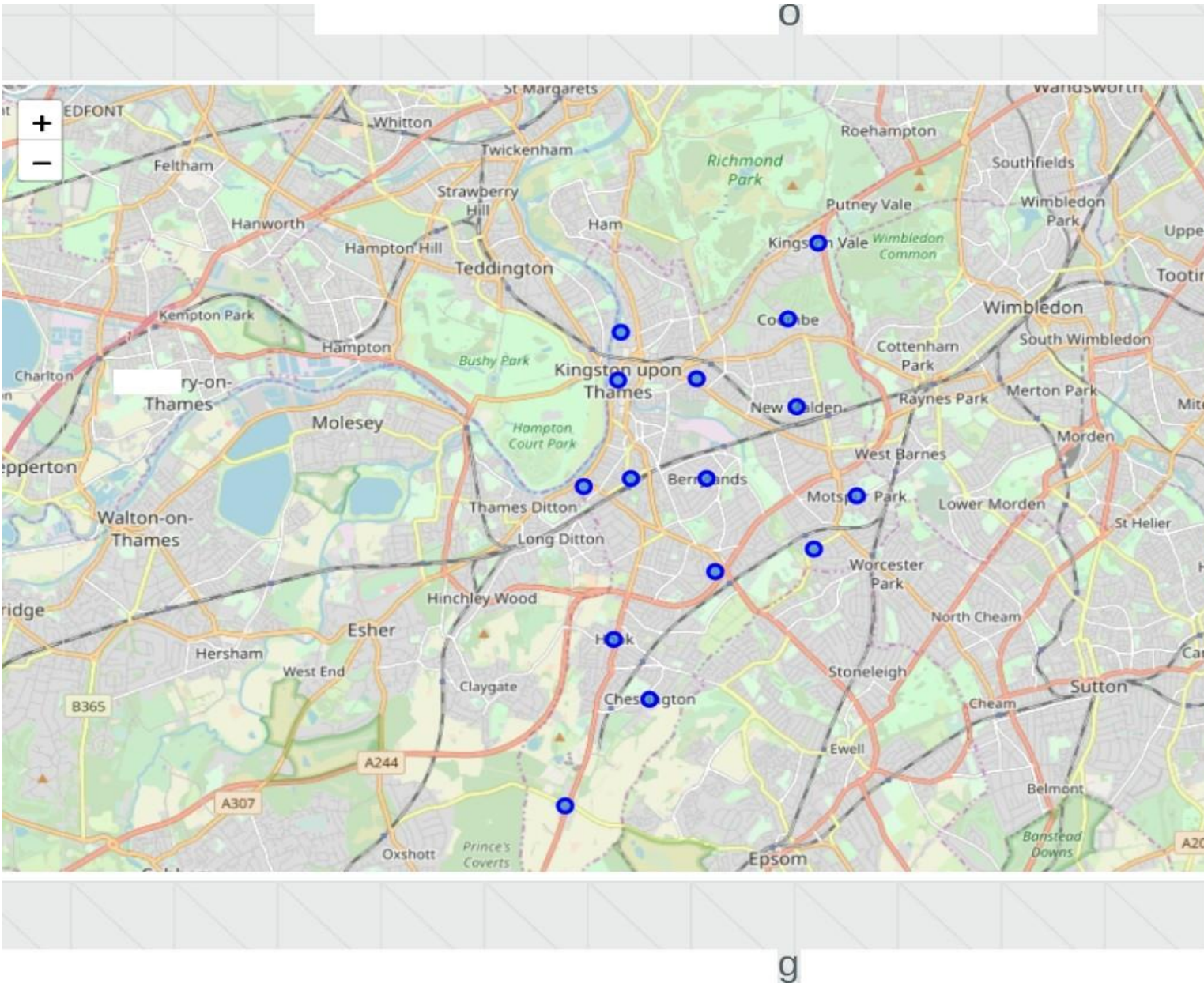


Comparing five boroughs with the lowest crime rate during the year 2016, City of London has the lowest recorded crimes followed by Kingston upon Thames, Sutton, Richmond upon Thames and Merton.

•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 2013 which suggests that it is a small area.

•We will consider the next borough with the lowest crime rate as the safest borough in London which is Kingston upon Thames.

Neighbourhoods in Kingston upon Thames



Modelling

- Using the final data set containing the neighbourhoods in Kingston upon Thames along with latitude and logtitude,we can find all the values within a 500 meter radius of each neighbourhood by connecting to the Foresquare API.

	Nelqhbarhood	Neighborhood Latitude	f4aighborhood Longitude	Vanue	Venue kaMtude	Ver	Longitude	Venue Category
0	Berrylande	S1.30g781	-0.284802	S+abltrxs Racket & Fknecs Club	51.392676		-0.290Z24	Gym / Ffizieea Center
S	Berrylanda	51.393781	•0.284B02	Aaxaridra Park	51.394230		-0.2B4108	Park
2	Berrylands	01.393781	-0.284802	K2 Bua Gtop	01.382302		-0.281534	Bue Gtop
S	Berrylande	51.393781	-J2&B02	Cafe Rasa	51.390175		-0.282480	Cafe
4	Canbury	51.417499	-0.305553	Tie Boater'e Inn	51.418548		-0.305B15	Pub

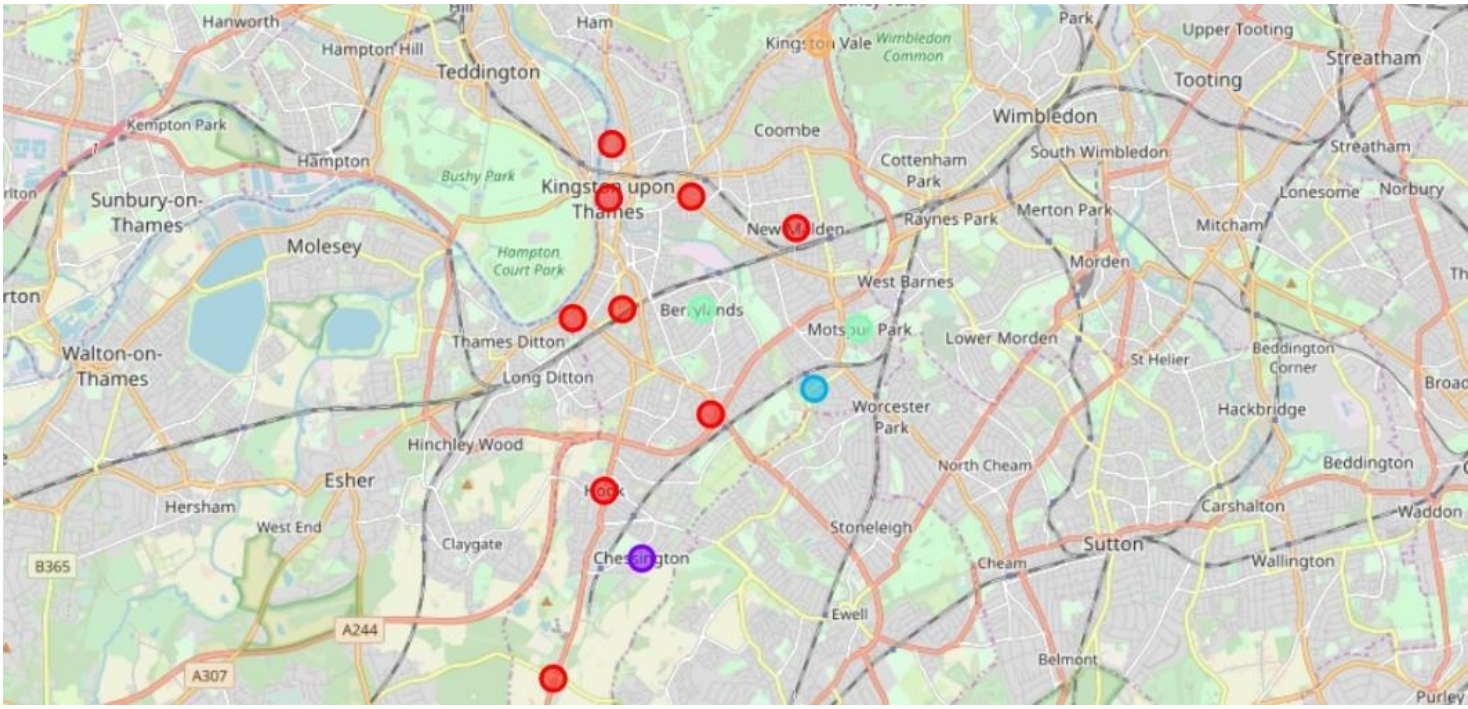
- One hot encoding is done on the Venues data.Venues data is then grouped by the Neighbourhood and the mean of the venues are calculated,finally the 10 common venues are calculated for each of the neighbourhoods.

- 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 cluter the 15 neighbourhoods into 5 clusters.The reason to conduct a K-means clustering is to cluster neighbouthoods with similar venues together so that people can shortlist area of their interest based on 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.Visualizing the clustered neighbourhoods on a map using the folium library.



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.