Capstone Project-The Battle of Neighborhoods

Introduction

Background :

Safety is a top concern when to a new area. If you don't feel safe in your own home, you are not going to be able to enjoy living there.

Problem :

This Project aim to select the safest borough in London based on the total crimes, exploring the neighborhoods of the borough to find the top 5 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 find the safest borough in London and explore its neighborhoods and common venue around each neighborhood.

Data Acquisition & Cleaning

Data Acquisition:

The data acquired for this project is a combination of data from three sources.

1) The first data source of the project uses a London crime data that shows the crime per borough in London.

2) 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.

3) 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 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 datasets are merged on the Borough names to form a new 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 2016.

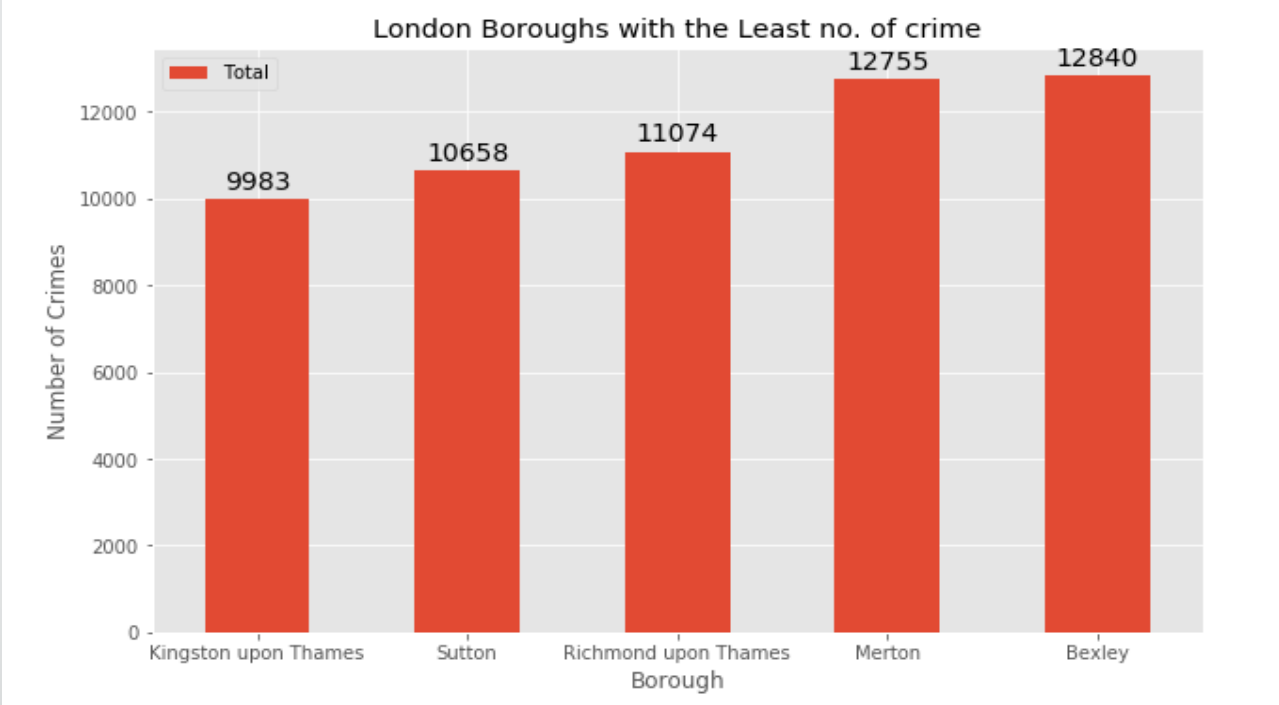
- After visualizing the crime in each borough we can find the borough with the lowest crime rate. The coordinates of the neighborhoods is be obtained using **Google Maps API geocoding** to get the final dataset

- 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.

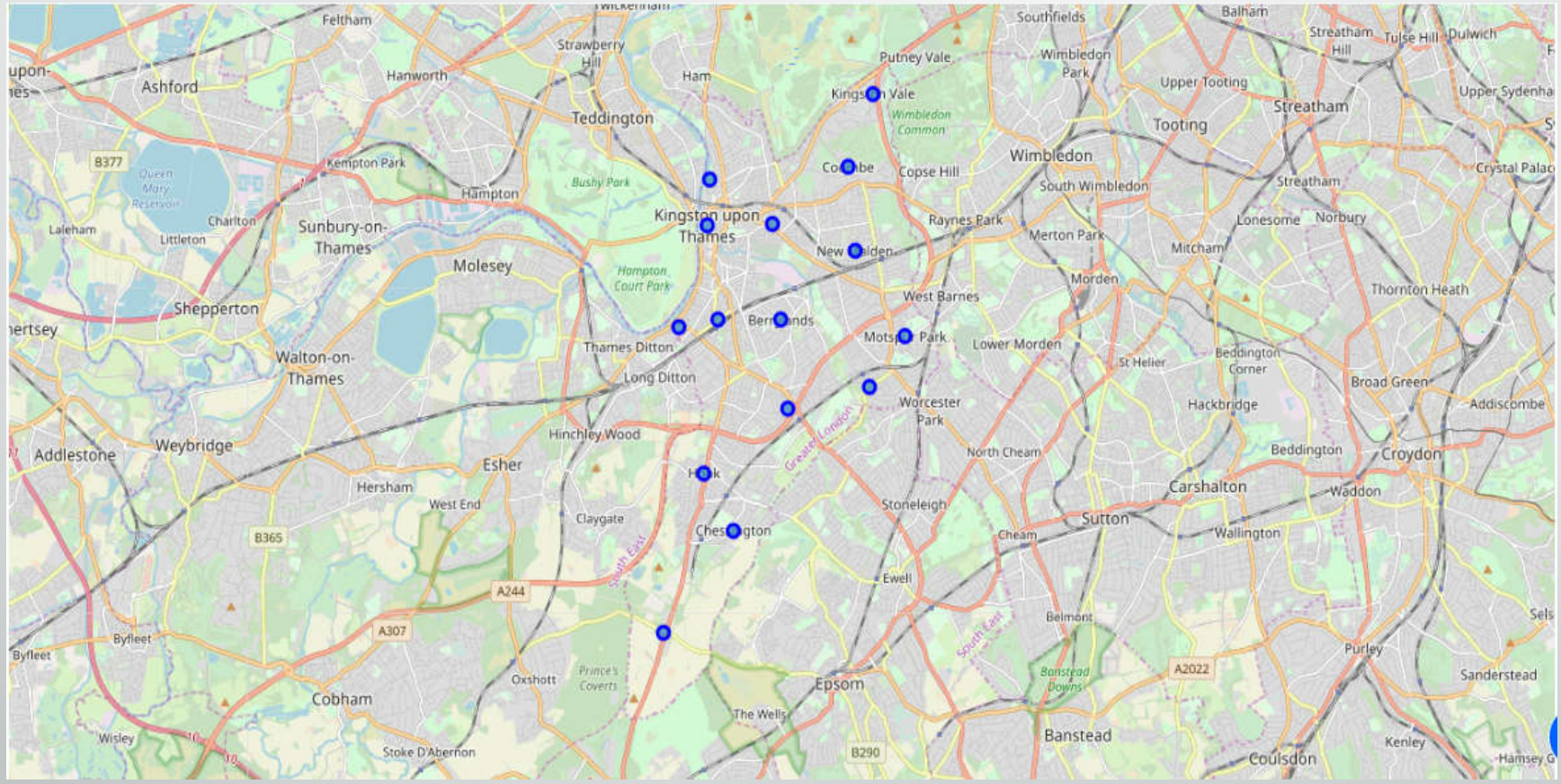
Methodology

Exploratory Data Analysis:

- 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’.



- 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.
- we will consider the next borough with the lowest crime rate as the safest borough in London which is 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

Modeling:

- 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.

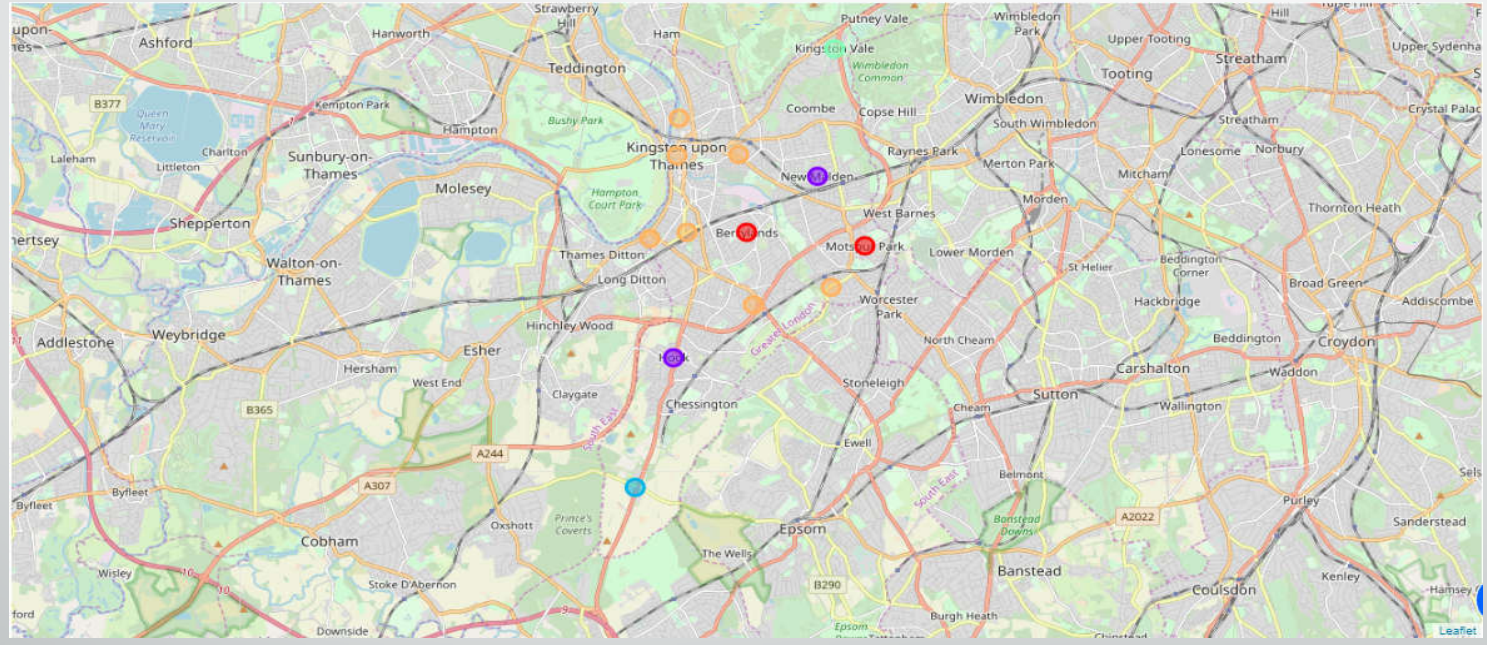
- One hot encoding is done on the venues data. The Venues data is then grouped by the Neighborhood and the mean of the venues are calculated, finally the 10 common venues are calculated for each of the neighborhoods.

- 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 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.

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.



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.

Cluster 1:

The cluster one is the biggest cluster with 9 of the 15 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.

	Neighborhood	Borough	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Canbury	Kingston upon Thames	51.417499	-0.305553	4	Pub	Café	Plaza	Fish & Chips Shop	Supermarket
5	Kingston upon Thames	Kingston upon Thames	51.409627	-0.306262	4	Coffee Shop	Café	Pub	Sushi Restaurant	Burger Joint
10	Norbiton	Kingston upon Thames	51.409999	-0.287396	4	Pub	Italian Restaurant	Indian Restaurant	Food	Wine Shop
11	Old Malden	Kingston upon Thames	51.382484	-0.259090	4	Train Station	Pub	Child Care Service	Food	Wine Shop
12	Seething Wells	Kingston upon Thames	51.392642	-0.314366	4	Indian Restaurant	Coffee Shop	Pub	Café	Fast Food Restaurant
13	Surbiton	Kingston upon Thames	51.393756	-0.303310	4	Coffee Shop	Pub	Grocery Store	Breakfast Spot	Gastropub
14	Tolworth	Kingston upon Thames	51.378876	-0.282860	4	Grocery Store	Pharmacy	Restaurant	Bowling Alley	Hotel

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.

- The choices of neighborhoods may vary from person to person.

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.