# APPLIED DATA SCIENCE CAPSTONE

"THE BATTLE OF NEIGHBORHOODS"

By: Mylavarabhatla Pratyush

# 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 are 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 – means clustering.

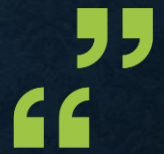
**Interest:**

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 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.

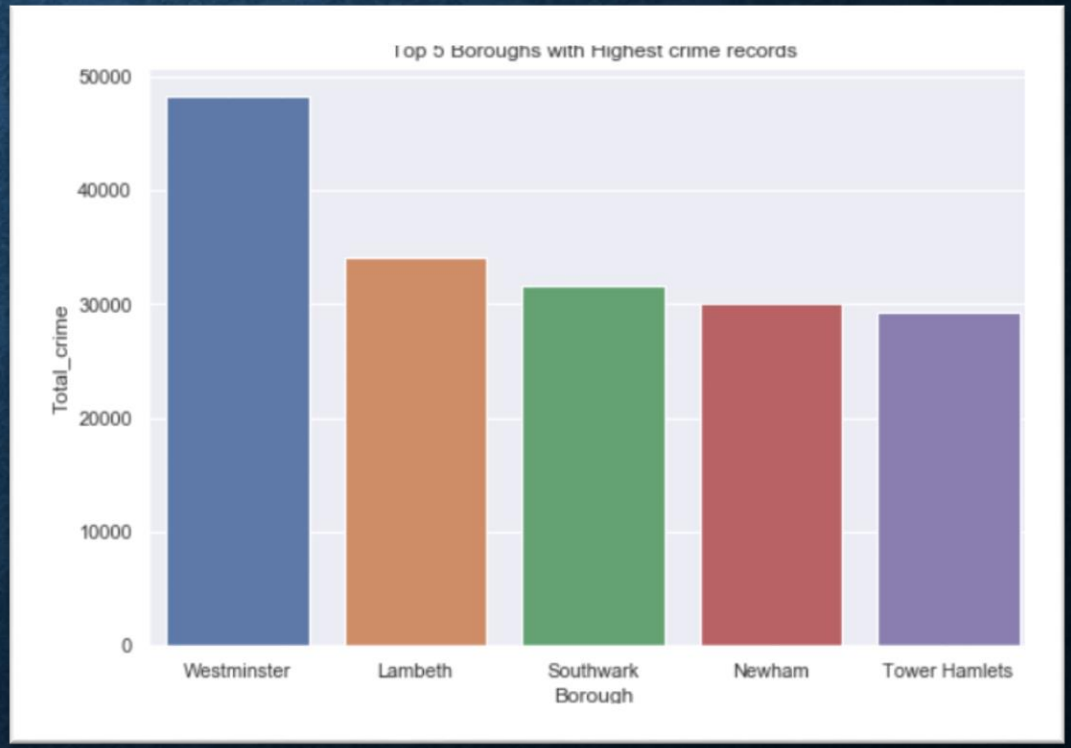
The third data source is the list of Neighborhoods in the Royal Borough of Kingston upon Thames as found on a Wikipedia page.

# DATA CLEANING

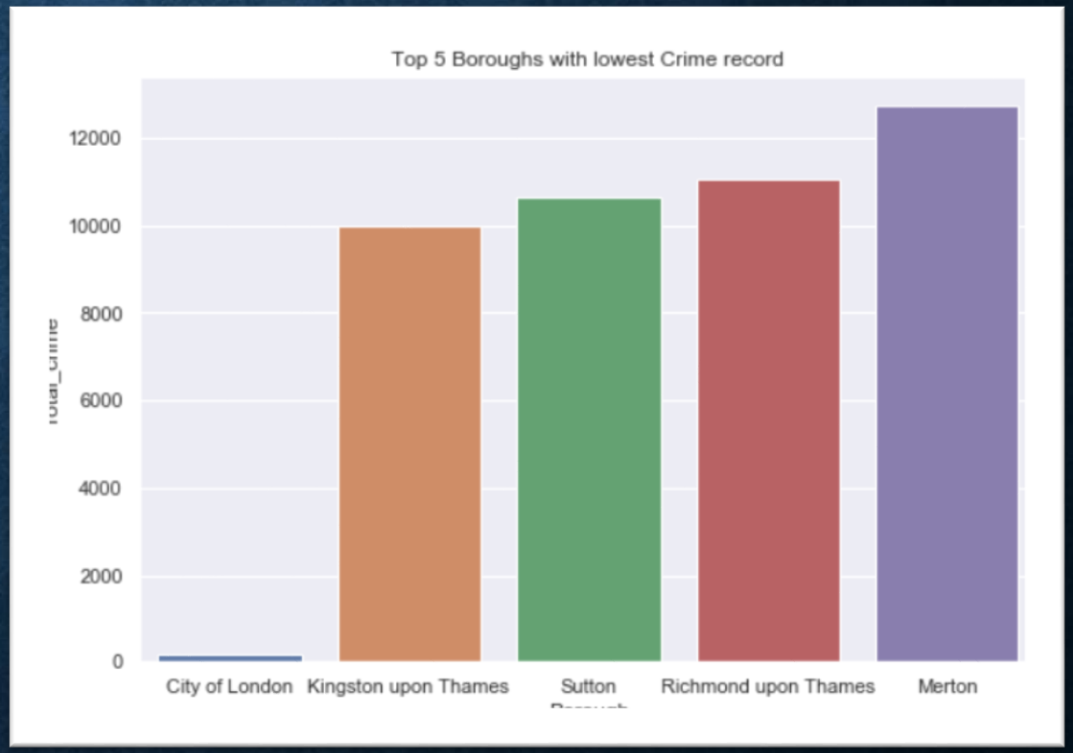
- The data preparation of each of the three sources of data is done separately:
- From the London crime data, the crimes during the most recent year 2016 are only selected. The major categories of crime are pivoted to get the total crimes.
- The second data is scraped from a Wikipedia page using the **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 cam find the borough with the lowest crime record 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.

# EXPLORATORY DATA ANALYSIS

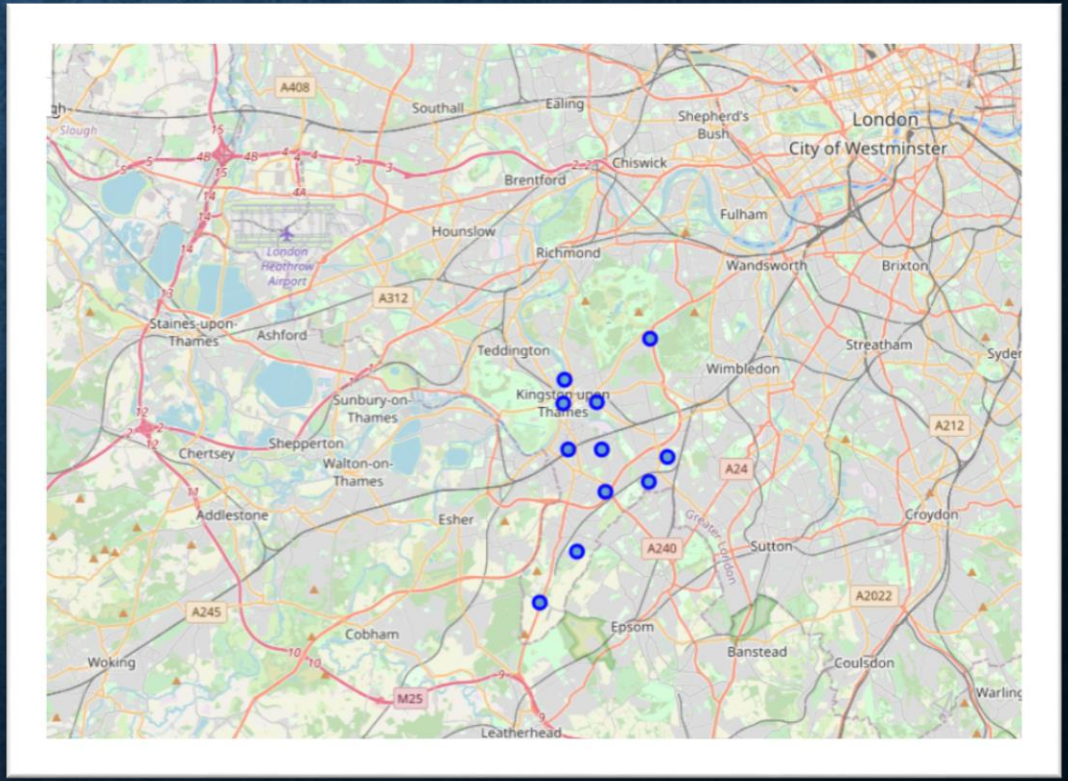
- **Boroughs with highest crime rates:**
- Comparing the five Boroughs with the highest crime record, Westminster has the highest crimes recorded followed by Lambeth, Southwark, Newham and Tower Hamlets. Westminster has a significantly higher crime rate than other 4 boroughs.



- **Boroughs with lowest crime rates:**
- Comparing five Boroughs with the lowest crime record, City of London has the lowest recorded crimes followed by Kingston upon Thames, Sutton, Richmond upon Thames and Merton.



- **Neighborhood in Kingston upon Thames:**
- There are 15 neighborhoods in the Royal Borough of Kingston upon Thames, they are visualized on a map using the Folium library.



# DATA MODELLING

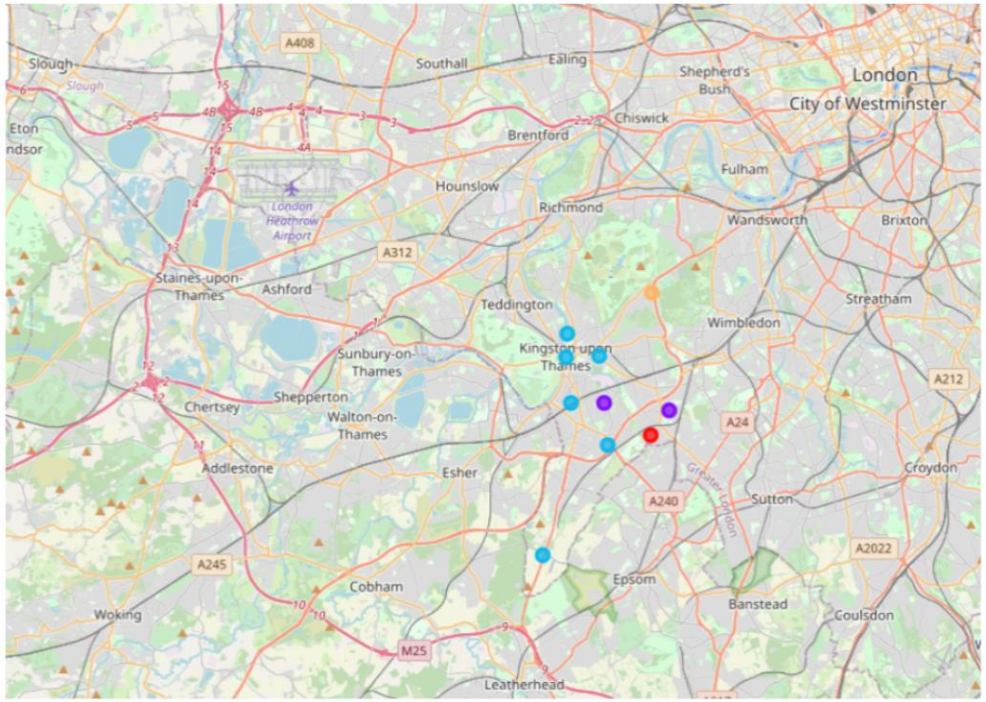
- Using the final dataset containing the neighborhoods in Kingston upon Thames along with the latitude and longitude, we can find all venues within a 500-meter radius of each neighborhood by connecting to the Foursquare API.
- One hot encoding is done on the venues data.
- We will use cluster size of 5 for this project that will divide the neighborhood into 5 clusters.

| | 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 | SK Superstores | 51.389901 | -0.283278 | Convenience Store |
| 4 | Canbury | 51.417499 | -0.305553 | Canbury Gardens | 51.417409 | -0.305300 | Park |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 224 | Tolworth | 51.378876 | -0.282860 | Manttex UK | 51.377009 | -0.285152 | Furniture / Home Store |
| 225 | Tolworth | 51.378876 | -0.282860 | LloydsPharmacy | 51.381273 | -0.283112 | Pharmacy |
| 226 | Tolworth | 51.378876 | -0.282860 | Tolworth Bus Stop B | 51.377780 | -0.279041 | Bus Stop |
| 227 | Tolworth | 51.378876 | -0.282860 | Viana Restaurant | 51.382067 | -0.284743 | Restaurant |
| 228 | Tolworth | 51.378876 | -0.282860 | Tolworth Railway Station (TOL) | 51.377385 | -0.279454 | Train Station |

# 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.

- Each cluster id color coded for the ease of presentation; we can see that majority of the neighborhood falls in the blue cluster which is the third cluster. Two neighborhoods have their own cluster (Red, Orange), these are clusters 4 and 5. The purple cluster consists of two neighborhoods which is the second 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, people who are looking for a neighborhood with good daily commute options then clusters 1 and 2 would be recommended.
- If people are looking for neighborhoods with good eating joints, then clusters 4 and 5 would be advisable.
- For people who are looking to shift with family, cluster 3 would be recommended because it has the most amenities/facilities in close vicinity.

# 6. CONCLUSIONS

- This project helps a person get a better understanding of the neighborhoods with respect to the most common venues in that neighborhoods. The neighborhoods selected in this project were done based on crime records in various Boroughs.
- Improvements can be made to this project by taking into consideration more factors such as Cost of living, Budget etc. Projects like these can really help people out and prevent them from moving out later due to neighborhood issues.