**Introduction**

The trend of people students going abroad for their higher studies has been rising tremendously, thanks to globalization and easier mode of transports.

Parents are often worried about kids going to a different country and living there all by themselves so naturally they would want the place to be safe and thus safety is a top concern when moving to a new area.

**THE PROBLEM**

The crime statistics dataset of London found on Kaggle has crimes in each Boroughs of London from 2008 to 2016. The year 2016 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. So as to decide which is the safest neighbourhood to live in.

**INTREST**

Students 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

**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:

1. lsoa\_code​: code for Lower Super Output Area in Greater London.
2. borough​: Common name for London borough.
3. major\_category​: High level categorization of crime
4. minor\_category​: Low level categorization of crime within major category.
5. value​: monthly reported count of categorical crime in given borough
6. year​: Year of reported counts, 2008-2016
7. 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:
8. Borough​: The names of the 33 London boroughs.
9. Inner​: Categorizing the borough as an Inner London borough or an Outer London Borough. Status​: Categorizing the borough as Royal, City or other borough.
10. Local authority​: The local authority assigned to the borough.
11. Political control​: The political party that control the borough.
12. Headquarters: ​Headquarters of the Boroughs.
13. Area (sq mi)​: Area of the borough in square miles.
14. Population (2013 est)[1]​: The population in the borough recorded during the year 2013.
15. Co-ordinates​: The latitude and longitude of the boroughs.
16. 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:

1. Neighborhood: ​Name of the neighborhood in the Borough.
2. Borough: ​Name of the Borough.
3. Latitude: ​Latitude of the Borough.
4. longitude:​ Longitude of the Borough.

1. **Data Cleaning**

**STEP 1**

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 borough as per the category.

**STEP 2**

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.

**STEP 3**

The two datasets are merged on the Borough names to form a new dataset that combines the necessary information in one dataset (see ​fig 2.3 ​ ). 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.

**STEP 4**

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 are left blank.

**STEP 5**

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

**And finally**

The new dataset is used to generate the venues for each neighborhood using the Foursquare API

**3. Methodology**

**3.1 Exploratory Data Analysis**

**3.1.1 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

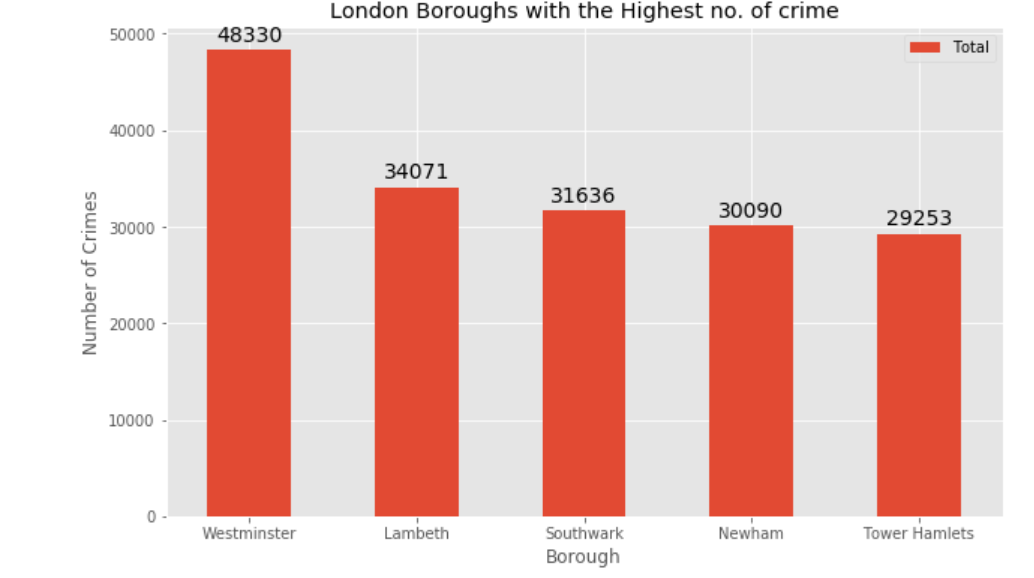


(See fig 3.1.1)

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’.

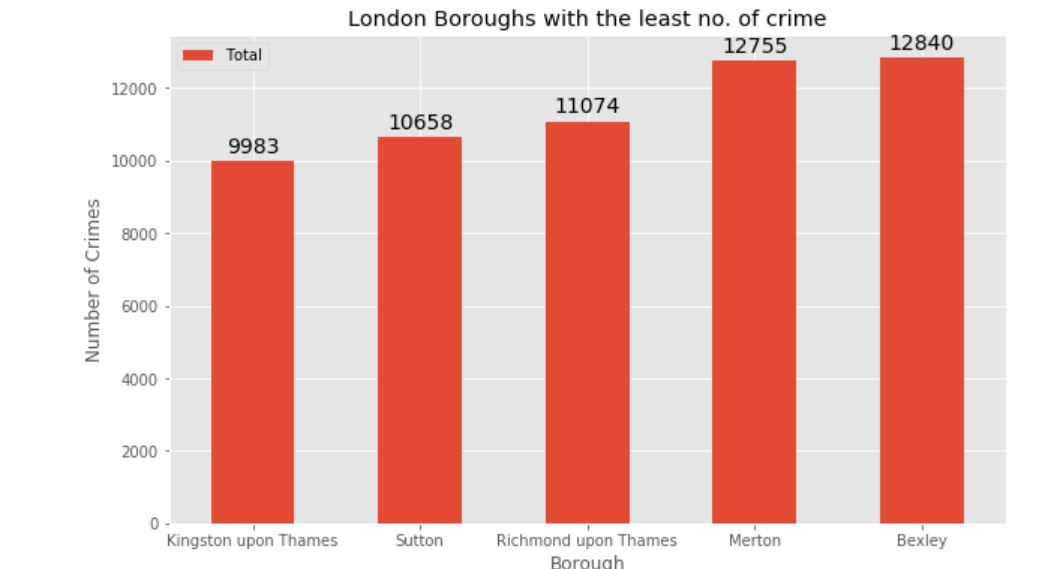
**3.1.2 Boroughs with the highest crime rates**

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 (see fig 3.1.2).



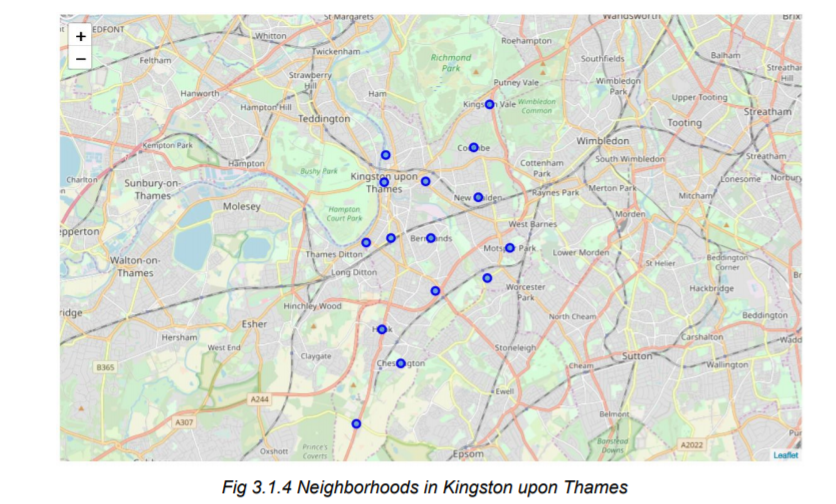
**3.1.3 Boroughs with the lowest crime rates**

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 (see fig 3.1.3).



3.1.4 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 (see fig 3.1.4).



**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. Looking into the neighborhoods in the first 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. Looking into the neighborhoods in the second, third 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

Cluster 2 The second cluster has one neighborhood which consists of Venues such as Restaurants, Golf courses, and wine shops.

Cluster 3 The third cluster has one neighborhood which consists of Venues such as Train stations, Restaurants, and Furniture shops.

Cluster 4 The fourth cluster has two neighborhoods in it, these neighborhoods have common venues such as Parks, Gym/Fitness centers, Bus Stops, Restaurants, Electronics Stores and Soccer fields etc.

Cluster 5 The fifth cluster has one neighborhood which consists of Venues such as Grocery shops, Bars, Restaurants, Furniture shops, and Department stores. We will look into the neighbourhoods in the fourth .

**6.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 to a new place to start a new chapter of your life which in this case is a higher education . 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.