SAFETY (OR LACK OF IT) IN LONDON

Applied Data Science Capstone Project

IBM Data Science Professional Certificate

Synthesis

The analysis of the criminality in London boroughs might be useful if one is thinking about looking for a new home or office location

Mariana Martinho

July 2020

Index

Introduction 2
Introduction and problem definition 2
Objective and scope 2
Data sources and description 2
Methodology 4
Results 5
Exploratory Data Analysis 5
Segmenting neighbourhoods of the safest borough in London 9
Clustering Neighbourhoods of safest borough in London 10
Discussion 11
Conclusions 13
References 13

Introduction

Introduction and problem definition

Nowadays we are living in an exponentially globalized world where people are increasingly moving from one city to another, between different countries (or even continents) chasing their dreams or just to find a better place to live and call home. One important factor is safety that sometimes is underappreciated in highly developed countries (mainly because criminality rate is decreasing) but it is essential to choose the place where we would live and mostly when we are moving to a new/ unknown place.

London is the capital and largest city of England and United Kingdom (UK), as well as one of most important cities in western part of Europe, with a total population of around 9 million people, distributed by their 33 boroughs (including the City of London). Although it is well known as a crucial metropolis at many different levels, London is also a city with a vast criminal record and therefore this variable plays an important role, or being even decisive, at the time of choosing a place when we decide to move and to settle a new life there.

Objective and scope

Based on this premise, the project main goal can be posed by the following question: "what is the most suitable/ safest area to live within London city?" Putting the hands on the data will help to discover the answer to make a recommendation for the target audience. In fact, this project is very interesting to everybody who is coming to London from abroad (or even within London area) and want to rent/ buy a new house or to anyone who might be planning moving an office for a new location.

Data sources and description

In order to answer the question, and to achieve the objective described above, available free data is used and analysed throughout the project, according to the London criminal records:

- number and type of crimes,
- evolution of criminality over time
- which boroughs are the most dangerous and safest, etc.

Main dataset: London Crime Data between 2008-2016. It is composed by 3419099 rows and 7 columns and can be found on the following link (https://www.kaggle.com/jboysen/london-crime)

- **Variables of main dataset:**
  - lsoa_code (code for Lower Super Output Area in Greater London (lsoa)).
  - borough (Name of London boroughs).
  - major_category (Major categorization of crimes).
  - minor_category (Minor categorization of crimes according to major category).
  - value (Number of crimes monthly reported in given borough).
  - year (Year of reported crimes, Jan/2008 - Dez/2016).
  - month (Month of reported crimes, Jan - Dez (1-12)).

Additional dataset: List of London Boroughs. It is composed by 33 rows and 10 columns and can be found on the following link (https://en.wikipedia.org/wiki/List_of_London_boroughs).

- **Variables of additional dataset:**
  - Borough
  - Inner
  - Status Local Authority
  - Political Control
  - Headquarters
  - Area
  - Population
  - Coordinates

For the final part of this work, a list of neighbourhoods was used, regarding the selected borough, and obtained from wikipedia website:

(https://en.wikipedia.org/wiki/List_of_districts_in_the_Royal_Borough_of_Kingston_upon_Thames).

Methodology

The methodology for the development of this project was three-folded:

1) **the exploratory data analysis**

This was an important section of this work because this type of analysis is very useful for a more statistical comprehensive interpretation through the combination of several variables of London criminal records. In this section, this interpretation considered the number and type of crimes committed in the boroughs of London, the evolution of criminality over time, the comparison of crimes between boroughs, etc. Different visualization tools were used to understand and to determine the safest borough.

2) **the segmentation of neighbourhoods**

After choosing the most suitable/ safest borough in London, a new dataframe was created with the list of 15 neighbourhoods. At this point, the geopy library was used to get the values of geographical coordinates (latitude and longitude) associated to each neighbourhood. Then, location data from Foursquare API was used to explore neighbourhoods within the selected borough, retrieving the most common venue categories in each neighbourhood. This part of the project is the so-called Segmentation of Neighbourhoods, where the 10 most common venues with a radius of 750 meters, around each neighbourhood coordinates centre, were retrieved. Finally, the folium library was also used to visualize the location of neighbourhoods in the maps of London's borough.

3) **the clustering neighbourhoods.**

The final section of the methodology describes the clustering of neighbourhoods, consisting in grouping the neighbourhoods into clusters. For this, firstly an one hot encoder (source: https://hackernoon.com/what-is-one-hot-encoding-why-and-when-do-you-have-to-use-it-e3c6186d008f) was applied to perform the "binarization" of each category, i.e. the top 10 most common venues associated to each neighbourhood. Then, and before applying the clustering algorithm, the several rows were grouped by neighbourhood and the mean of the frequency of venues was taken. The K-means clustering algorithm was then applied to generate the several clusters of neighbourhoods, using lastly the folium library to visualize the neighbourhoods in London and their emerging clusters. Before the decision/ recommendation making of most suitable neighbourhood, the analysis of each cluster and respective most common venues was performed.

Before the first method being applied, the data needed to be scrapped and manipulated to obtain the structured format and generate the dataframes. Regarding the main dataset, the rows with variable "value" equal to zero needed to be removed from the dataframe, i.e., the crimes were reported in the given boroughs, however they were classified as zero since they probably did not truly happened and therefore these rows were discarded from the original dataset. Then, the merge of both datasets was done, grouping the rows by boroughs.

Results

Exploratory Data Analysis

Figure 1 illustrate the bidistribution, traduced by scatter plot between the total number of crimes reported in 2012 and the population number in the same year for all the boroughs of Great London.

A good positive correlation can be achieved between the data. It is also possible to interpret that the higher the population number for a given borough, the higher is the number os crimes reported.

The "Westminster" and "City of London" boroughs are the outliers of the data, being also the dangerous and safest boroughs in London, respectively.

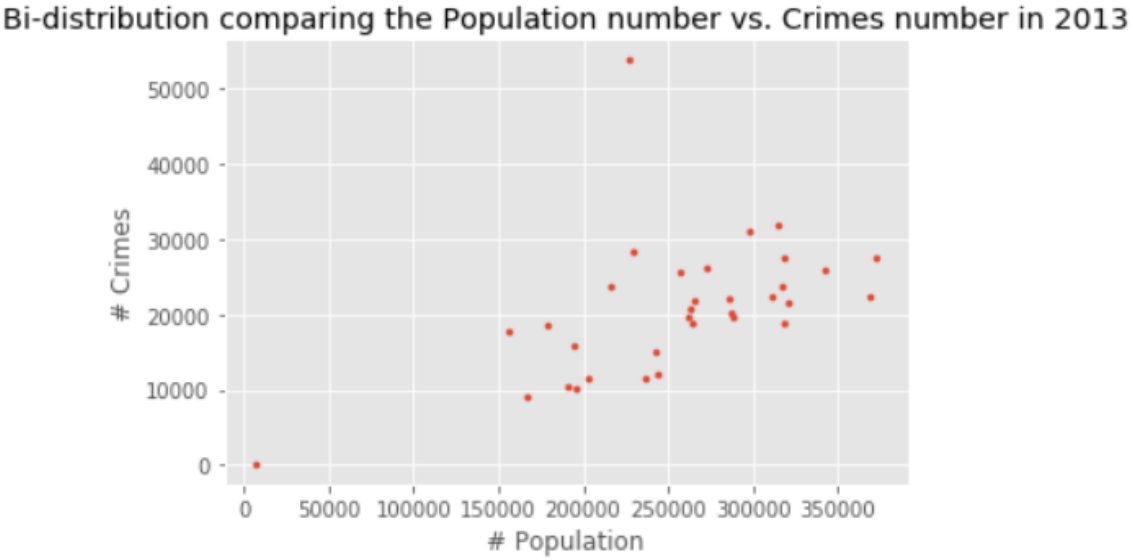

Figure 1. Scatterplot between population and crimes in London in 2013.

Figure 2 and **Figure 3** are line plots thar illustrate the evolution of Top 2 types of crimes in London over the past years and months, respectively. Starting from Figure 2, the general behaviour is quite different between both crime types over the years. The highest number of reports happened in 2012 for "Theft and Handling", in spite of "Violence Against the Person" crime type which is the second most common in London, has a stable behaviour between 2008 and 2013, significantly increasing afterword's.

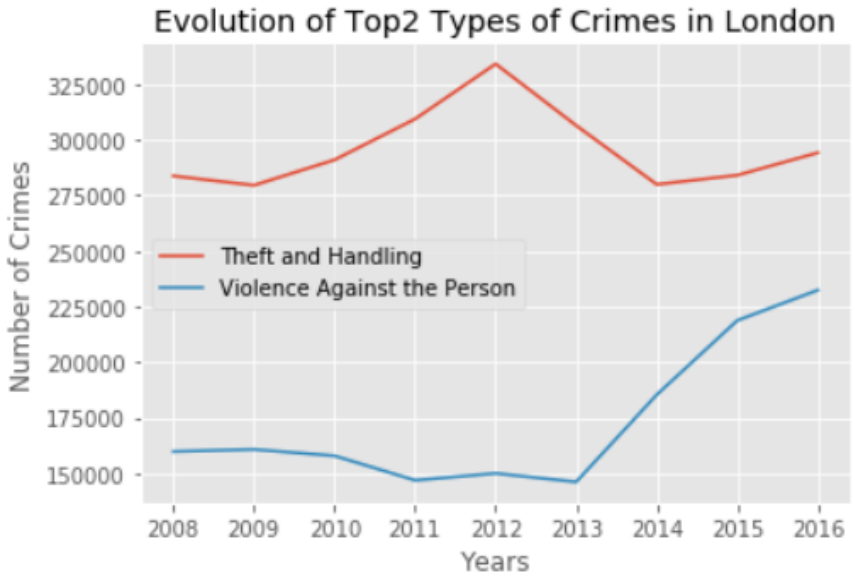

Figure 2. Line plot of the evolution of Top 2 types of crimes in London over the years.

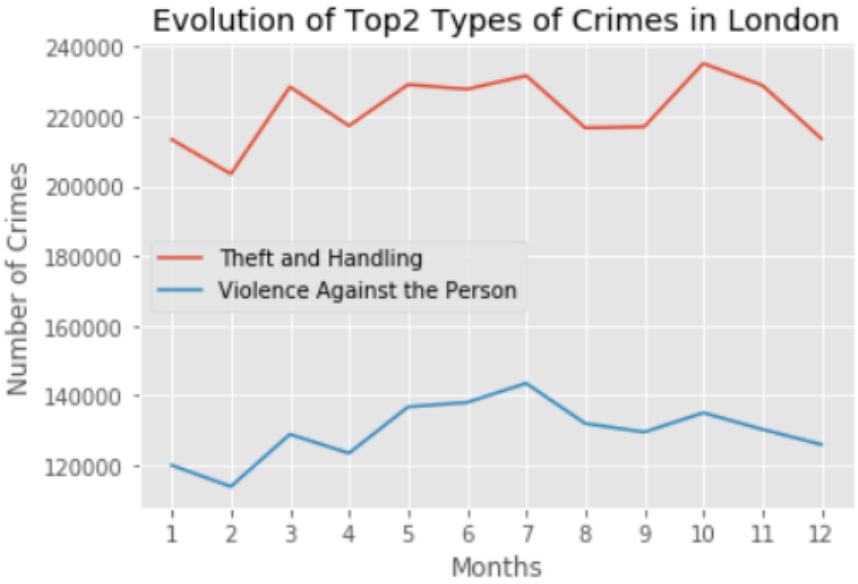

Figure 3. Line plot of the evolution of Top 2 types of crimes in London over the months.

Figure 4 and **Figure 5** are histograms of Top 5 boroughs with highest and lowest number of crimes, respectively, in Great London. Clearly between 2008-2016 "Westminster" borough presents the higher number of reported crimes (Figure 4), therefore it's the most dangerous (almost half a million times); City of London is the safest borough in London with 780 reported crimes.

Figure 4. Histogram of Top 5 (higher values) of total crimes in London boroughs.

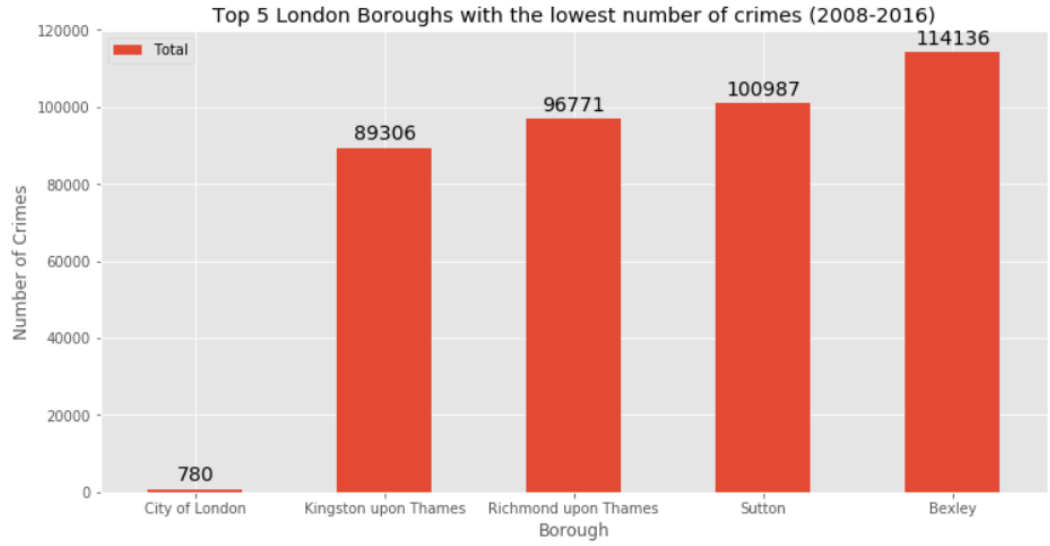

Figure 5. Histogram of Top 5 (lower values) of total crimes in London boroughs.

According to Wikipedia, "City of London" does not consist to a borough of Great London, although it was considered for the previous exploratory data analysis. For this reason, the results illustrated in **Figures 6**, **7** and **8** encompass only the interpretation between "Kingston Upon Thames" and "Richmond upon Thames" to select the safest Borough in London.

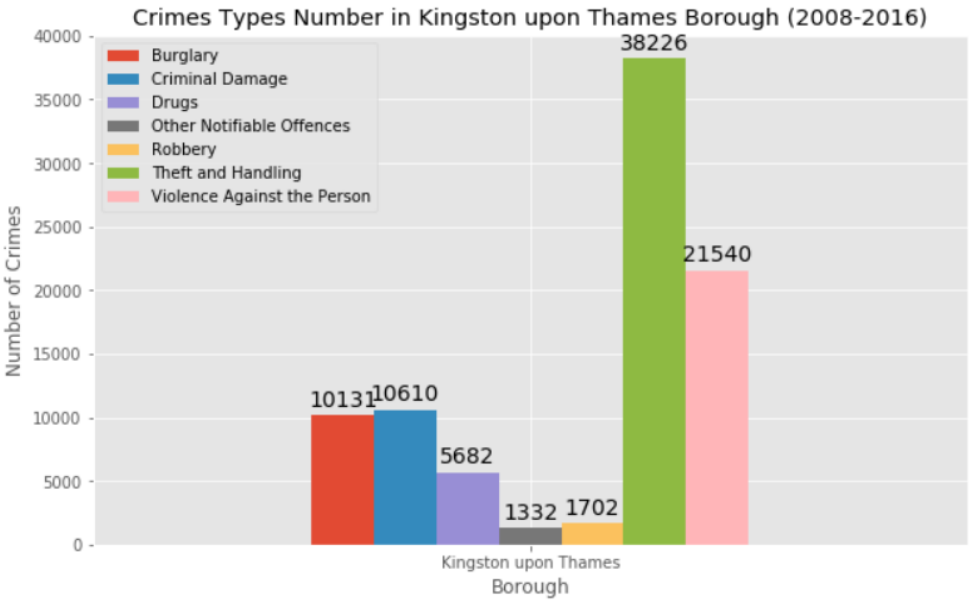

Figure 6. Histogram of major crime types in Kingston Upon Thames borough.

Figure 7 and **Figure 8** show a good similarity in the general behaviour of crime types, within Major Category, between both analysed boroughs (although frequency number is different). Clearly the crime types "Theft and Handling" and "Violence Against the Person" happen more often. Similarly, between both Boroughs, crime types classified as "Robbery" and "Other Notifiable Offences" are the less reported.

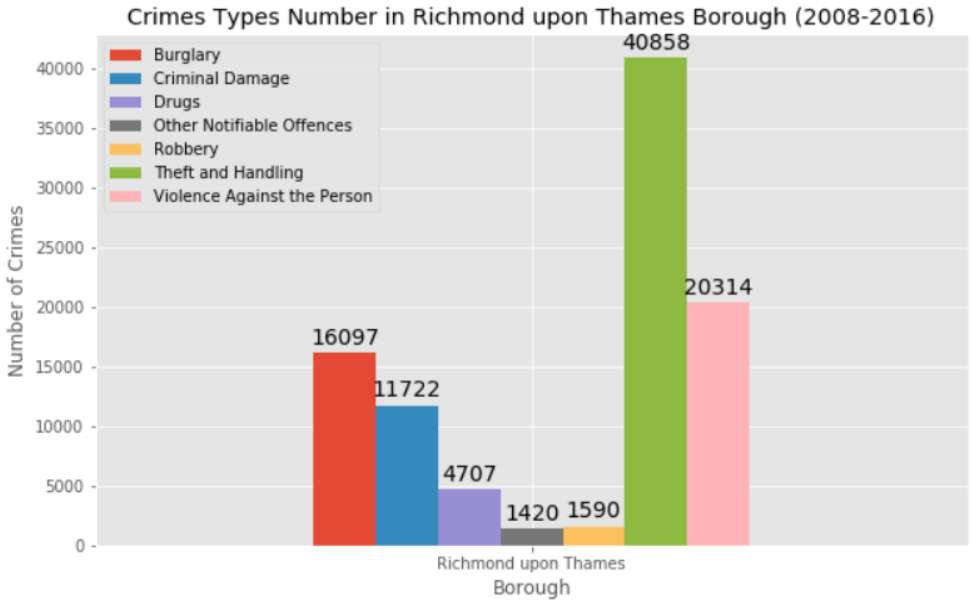

Figure 7. Histogram of major crime types in Richmond Upon Thames borough.

In **Figure 8,** between time intervals of 2008-2009 and 2015-2016, the evolution behaviour is similar. Critical years of criminality maximums were 2008, 2012 and 2016, for "Richmond upon Thames", and mostly 2008 for "Kingston upon Thames". "Richmond upon Thames" registered a peak in crimes between 2010 and 2013. "Kingston upon Thames" shows a significant decrease in crimes between 2012 and 2013.

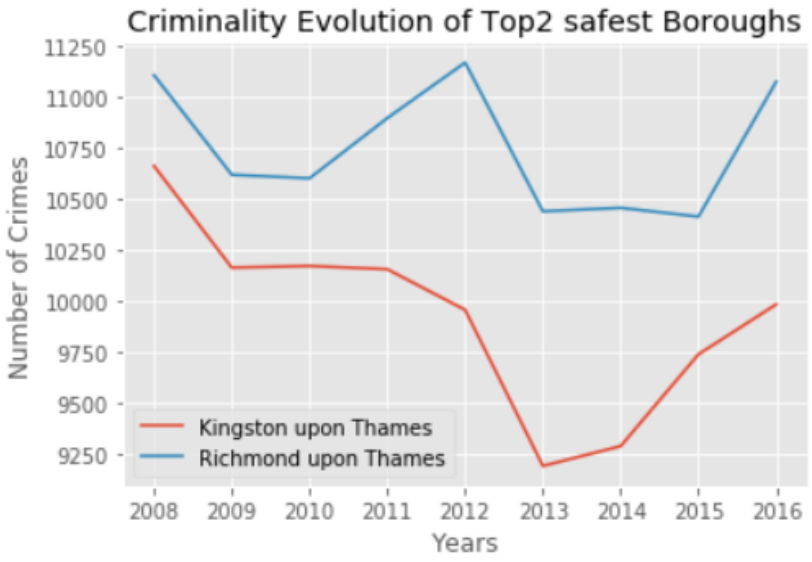

Figure 8. Line plot of the evolution of total crimes over the years in Kingston upon Thames and Richmond upon Thames.

Based on this final exploratory data analysis between the two safest boroughs, "Kingston upon Thames" is the selected borough to carry this project on.

Segmenting neighbourhoods of the safest borough in London

In this section, the neighbourhoods segmentation' results from "Kingston upon Thames" borough are presented. **Figure 9** shows a table with the results after using the geolocator (with a user_agent = "London agent") to find the latitude and longitude of each neighbourhood.

	Neighborhood	Borough	Latitude	Longitude
0	Berrylands	Kingston upon Thames	51.393781	-0.284802
1	Canbury	Kingston upon Thames	51.417499	-0.305553
2	Chessington	Kingston upon Thames	51.358336	-0.298622
3	Coombe	Kingston upon Thames	51.419450	-0.265398
4	Hook	Kingston upon Thames	51.367898	-0.307145
5	Kingston upon Thames	Kingston upon Thames	51.409627	-0.306262
6	Kingston Vale	Kingston upon Thames	51.431850	-0.258138
7	Malden Rushett	Kingston upon Thames	51.341052	-0.319076
8	Motspur Park	Kingston upon Thames	51.390985	-0.248898
9	New Malden	Kingston upon Thames	51.405335	-0.263407
10	Norbiton	Kingston upon Thames	51.409999	-0.287396
11	Old Malden	Kingston upon Thames	51.382484	-0.259090
12	Seething Wells	Kingston upon Thames	51.392642	-0.314366
13	Surbiton	Kingston upon Thames	51.393756	-0.303310
14	Tolworth	Kingston upon Thames	51.378876	-0.282860

Figure 9. Neighbourhoods from Kingston upon Thames borough and respective coordinates retrieved using geopy llibrary.

The geolocator was also used to determine the central coordinates of the borough, in this case, the central point of "Kingston upon Thames", which is "Berrylands" with values of 51.3937811; -0,2848024. It is worth to mention that the coordinates of the center for the entire borough are very similar with those of "Berrylands" neighbourhood center (**Figure 9**).

Using the elements of **Figure 9**, the spatial distribution of neighbourhoods was plotted and shown in the map of **Figure 10**. Now that we have all the neighbourhoods' location let's move to the battle of neighbourhoods.

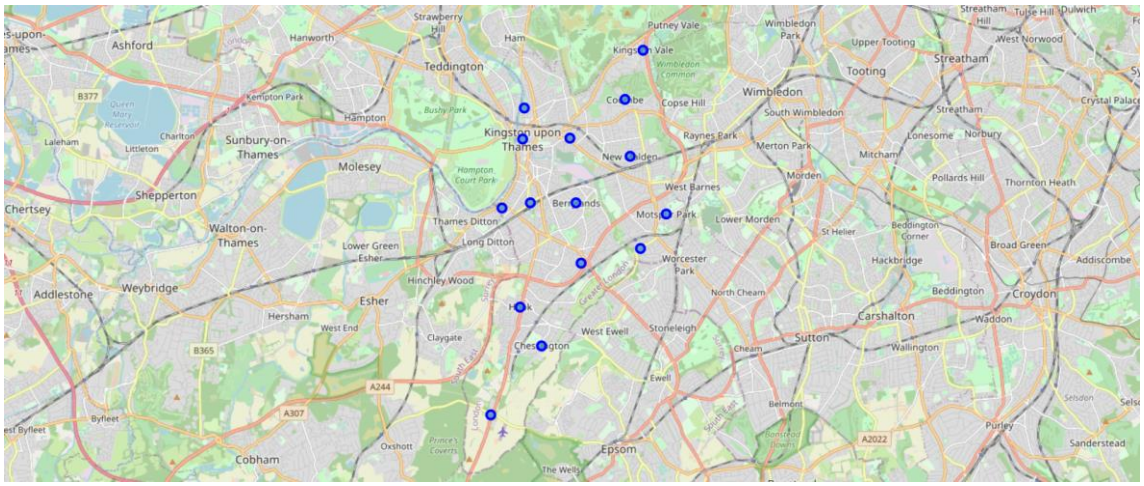

Figure 10. Map of the neighbourhoods (blue dots) from Kingston upon Thames borough.

Clustering Neighbourhoods of safest borough in London

The final results shown in this section are related with the modelling of the data using the Foursquare API to get the venues to every neighbourhood and applying the K-means clustering algorithm to find the similar groups of neighbourhoods based on similarity of ranked common venues.

Figure 11 shows the final map of "Kingston upon Thames" borough with the five clusters of neighbourhoods. The most important parameters used to obtain the final clustering were the definition of the radius to explore the venues of each neighbourhood (radius = 750 meters) and the pre-defined number of clusters for the K-means algorithm. Clearly **cluster 2** is in the western part of the borough and **cluster 3** regards the most eastern part of the borough. **Clusters 1** and **4** are clearly distinct in what common venues is concerned.

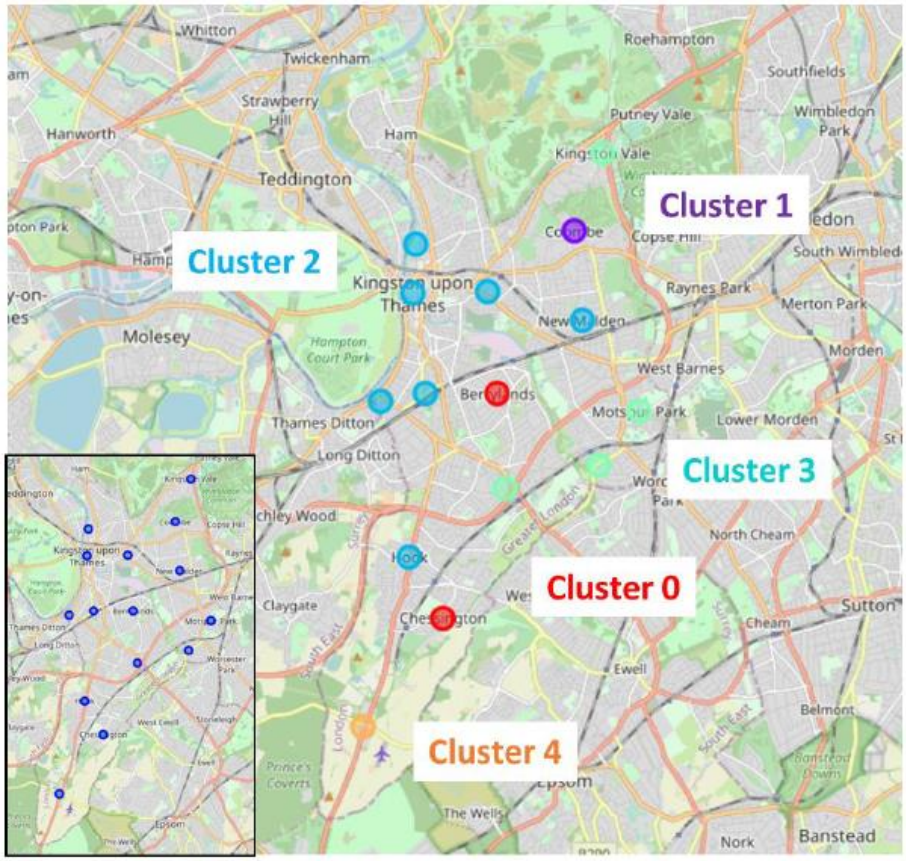

Figure 11. Map of clusters of neighbourhoods from Kingston upon Thames borough.

A more comprehensive analysis of each cluster individually and the comparison between each other is accomplished in the next section of discussion.

Discussion

The analysis of each cluster was performed, and the top 5 common venues were selected to facilitate a better understanding of the differences and the similarities of the neighbourhoods. It is important to notice that different k values were tested in the K-means algorithm, however there were always two distinct neighbourhoods when compared to the remaining ones.

Keeping the clustering k-values equal to 5 and the exploration radius of 750 meters, the following list of common venues, corresponding to each cluster, are the following:

Cluster 0	Pub, Train Station, Bus Stop, Coffee Shop, Golf Course
Cluster 1	Golf Course, Hotel, Garden, Spa, Food
Cluster 2	Café, Breakfast Spot, Coffee Shop, Supermarket, Restaurants
Cluster 3	Coffee Shop, Gym/Fitness Center, Grocery Store, Tennis Court, Football Field
Cluster 4	Theme Park Ride/Attraction, Pub, Restaurant, Garden Center, Hotel

After analysing the clusters, the neighbourhood recommendation would be Berrylands neighbourhood represented in **Figure 12.**

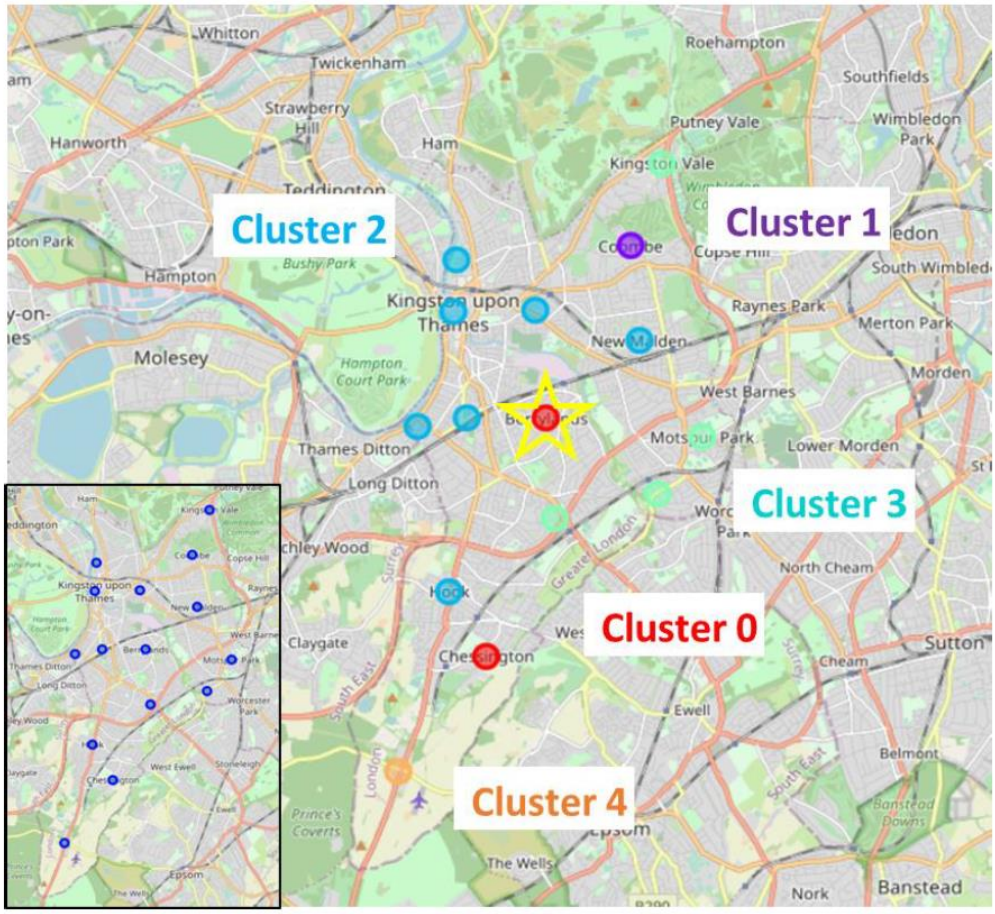

Figure 12. Map of the five clusters of neighbourhoods from Kingston upon Thames borough. The star represents the recommended neighbourhood.

Independently of the interested audience is just a citizen who wants to find an house or a someone who wants to move the office to another part of the city, neighbourhood Berrylands would be the most suitable area where in just a walk of few

minutes we can find social venues as pubs or coffee shops, or public transport points as train station and bus stops, or either if one is interested in taking a golf course which is also very close to the neighbourhood center.

Moving a bit further, 750 meters or more, food shops, restaurants, grocery, breakfast spots, supermarkets and restaurants can also be found as well as gyms, cafes and sport and leisure areas.

Conclusions

In this project, the exploratory analysis of London criminality data was done to better understand how criminality components evolved during the last years and to find the safest borough in the city (Kingston upon Thames).

Segmenting and clustering the neighbourhoods belonging to Kingston upon Thames borough was done using Foursquare API and K-means clustering algorithm, allowing to explore and retrieve the most common venues around each neighbourhood and cluster them based on venues similarity. After analysing the several clusters and respective venues, the final recommendation to everybody who is interested to find a new house, or a new office to work in, would be Berrylands neighbourhood.

As future work on this project, the extension of this method to other London boroughs would be valuable but also the incorporation of property/office costs to complement this work since topics of safety and common venues were considered.

References

All the references used for this project are shown in the code itself.