Capstone Project - The Battle of

Neighborhoods

**Report**

**1. Introduction**

**1.1 Background**

The average American moves about eleven times in their lifetime. This brings us to the

question: Do people move until they find a place to settle down where they truly feel happy,

or do our wants and needs change over time, prompting us to eventually leave a town we

once called home for a new area that will bring us satisfaction? Or, do we too often move to

a new area without knowing exactly what we’re getting into, forcing us to turn tail and run at

the first sign of discomfort?

To minimize the chances of this happening, we should always do proper research when

planning our next move in life. Consider the following factors when picking a new place to

live so you don’t end up wasting your valuable time and money making a move you’ll end

up regretting. 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.

**1.2 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.

**1.3 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 and Cleaning**

**2.1 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:

● **lsoa\_code** : code for Lower Super Output Area in Greater London.

● **borough** : Common name for London borough.

● **major\_category** : High level categorization of crime

● **minor\_category** : Low level categorization of crime within major category.

● **value** : monthly reported count of categorical crime in given borough

● **year** : Year of reported counts, 2008-2016

● **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:

● **Borough** : The names of the 33 London boroughs.

● **Inner** : Categorizing the borough as an Inner London borough or an Outer London

Borough.

● **Status** : Categorizing the borough as Royal, City or other borough.

● **Local authority** : The local authority assigned to the borough.

● **Political control** : The political party that control the borough.

● **Headquarters:** Headquarters of the Boroughs.

● **Area (sq mi)** : Area of the borough in square miles.

● **Population (2013 est)[1]** : The population in the borough recorded during the year

2013.

● **Co-ordinates** : The latitude and longitude of the boroughs.

● **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:

**● Neighborhood:** Name of the neighborhood in the Borough.

**● Borough:** Name of the Borough.

**● Latitude:** Latitude of the Borough.

**● Longitude:** Longitude of the Borough.

**2.2 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. After the web scraping, string manipulation is required to get the names of the

boroughs in the correct form (see *fig 2.2* ). This is important because we will be merging the

two datasets together using the Borough names.

The two datasets are merged on the Borough names to form a new dataset that combines

the necessary information in one 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 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 left blank

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.

**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* ).

*Fig 3.1.1 Statistical description of the London crimes*

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 .

**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

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. Hence

we will consider the next borough with the lowest crime rate as the safest borough in

London which is Kingston upon Thames.

**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 .

*Fig 3.1.4 Neighborhoods in Kingston upon Thames*

**3.2 Modelling**

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. This returns a json file containing all

the venues in each neighborhood which is converted to a pandas dataframe. This data

frame contains all the venues along with their coordinates and category .

One hot encoding is done on the venues data. (One hot encoding is a process by which

categorical variables are converted into a form that could be provided to ML algorithms to

do a better job in prediction). 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.

**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

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

The second cluster has one neighborhood which consists of Venues such as Restaurants,

Golf courses, and wine shops.

The third cluster has one neighborhood which consists of Venues such as Train stations,

Restaurants, and Furniture shops.

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 cluster

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.

Visualising the clustered neighborhoods on a map using the folium library (see fig 4.6).

*Fig 4.6 Clustered neighborhoods in the Borough of Kingston upon Thames*

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.

**5. 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.

**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 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.