**Capstone Project - The Battle of Neighborhoods**

## Best Neighborhoods in London Borough

Author: Jeetendra Swami Date: May 2020

# Contents

1. [**Introduction**](#_heading=h.30j0zll) **3**
   1. [Background](#_heading=h.3znysh7) 3
   2. [Problem](#_heading=h.tyjcwt) 3
   3. [Interest](#_heading=h.1t3h5sf) 3
2. [**Data acquisition and cleaning**](#_heading=h.2s8eyo1) **3**
   1. [Data sources](#_heading=h.3rdcrjn) 3
   2. [Data Cleaning](#_heading=h.lnxbz9) 4
3. [**Methodology**](#_heading=h.1ksv4uv) **7**
   1. [Exploratory Data Analysis](#_heading=h.2jxsxqh) 7
      1. [Analysis of crime data](#_heading=h.3j2qqm3) 7
      2. [Analysis of Housing price](#_heading=h.4i7ojhp) 9
      3. [Analysis of hapiness index](#_heading=h.1ci93xb) 10
   2. [Modelling](#_heading=h.2bn6wsx) 11
4. [**Results**](#_heading=h.3as4poj) **13**
5. [**Discussion**](#_heading=h.49x2ik5) **15**
6. [**Conclusion**](#_heading=h.147n2zr) **15**

# Introduction

## Background

Moving when you have a family can be daunting, especially when you’re moving to a city as diverse as London is. So as part of a career change, when I was considering setting up my new home in London, even though it was exciting, it was certainly not an easy task. Currently settled in a peaceful locality in the suburbs of Michigan, the bustling life and current political scenario of London was daunting to say the least. Upon some research, I could find that the slump in London’s housing market may be about to end. And this is the right time to invest in a property.

However, before making an investment, I decided to do a little research on the best neighborhood.

## Problem

Data that might contribute to determining the best neighborhood include how safe the neighborhood is, affordability, accessibility and facilities like parks, restaurants etc. This project aims to predict the best neighborhood based on the above factors.

## Interest

Obviously, expats who are considering relocating to London as part of career change could utilise this report for finding a safe borough and district in London for buying/renting a house. People who are currently in London, but has not invested in a house yet due to the current political/economical scenario, could also use this report.

# Data acquisition and cleaning

## Data sources

The first step in the analysis was to find all the boroughs in London. This data could be obtained from a wiki page, [List of London Boroughs](https://en.wikipedia.org/wiki/List_of_London_boroughs) .

The latest crime data records could not be obtained for the boroughs. However, the data pertaining to year 2016, is there as a kaggle dataset, [Kaggle dataset for London Crime data](https://www.kaggle.com/jboysen/london-crime) .

Next, we can check the affordability of the houses using a dataset from Office for National Statistics. ONS Dataset for Houseprice.

During the research, I also came across a survey by itv, which records the happiness index of the people in uk. The link for the report is [Happiest](https://www.itv.com/news/london/2016-08-16/londons-happiest-boroughs-revealed-as-richmond-hits-the-top-spot/) [Boroughs revealed by ITV](https://www.itv.com/news/london/2016-08-16/londons-happiest-boroughs-revealed-as-richmond-hits-the-top-spot/) . The data from this web page can be scrapped and the best boroughs can be found out.

Finally to find the districts in Sutton, which is the best borough for buying a house, we scarp the wiki page: [Districts in London Borough of Sutton](https://en.wikipedia.org/wiki/London_Borough_of_Sutton#Districts)

## Data Cleaning

**To find the London borough names**, we use BeautifulSoup to scrape the wiki page. After that, string manipulation is done using regular expression and the exact name of the boroughs are extracted. Extracting the exact names is very important, as throughout the project, we will be using the Borough names as the index for the data frames.

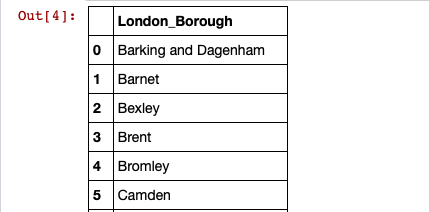


Figure 1: List of London Borough

**For the crime rate analysis**, the data set from Kaggle is downloaded. The dataset consist of:

* + - 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 csv file is read using pandas. Only the data from the year 2016 is extracted. If the ’value’ column is 0, then is filtered out. Finally the dataset can be grouped together on borough name to get the count/crime rate.

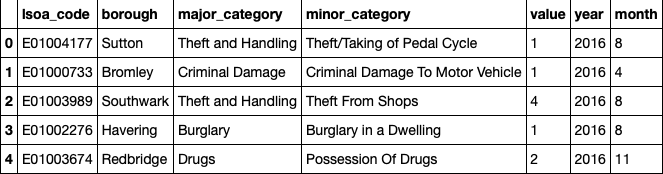


Figure 2: London Crime Data for the year 2016

For the analysis of price per sq.m of houses in the boroughs, the dataset from [Office of National Statistics](https://www.ons.gov.uk/) is extracted. This contains:

* + - local authority code
    - local authority name
    - year
    - price per m2

Only local authority name which are [there in the Borough list extracted from](https://en.wikipedia.org/wiki/List_of_London_boroughs) [the wiki page](https://en.wikipedia.org/wiki/List_of_London_boroughs) is filtered and taken into a dataset. Next, the local authority code and year can be dropped. This will give the dataset with the borough name and price/m2. The dataset when sorted in the ascending order will give the most affordable areas.

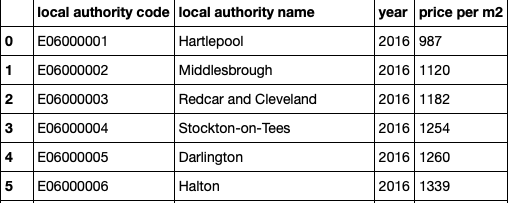


Figure 3: Comparison of House price per m2 in London(2016)

From the analysis of the above three data source, we conclude the best borough

for real estate in London to be Sutton. Next we need to build a dataset with the neighborhoods in Sutton.

For that, the wiki page [London Borough of Sutton #Districts](https://en.wikipedia.org/wiki/London_Borough_of_Sutton#Districts) are scrapped and the district names are obtained. Borough name will be Sutton. The geographical coordinates are obtained using a geopy **client**.

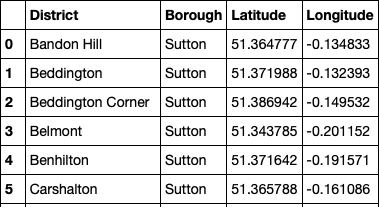


Figure 4: Geographical Coordinates of Sutton Disticts

Finally Using **Foursquare Location Data**, the 100 most popular venues in a radius of 500m for each district Sutton is obtained. The data obtained is a JSON file, and we need to turn that into a data-frame. This final dataset will contain:

* + - District
    - District Latitude
    - District Longitude
    - Venue
    - Venue Latitude
    - Venue Longitude
    - Venue Category



Figure 5: Top 100 Venues in each Neighborhoods of Sutton

# Methodology

The methodology in this project consists of two parts:

## Exploratory Data Analysis

In this section, we will analyze the data frames built in the above section. And come up with the analysis on the best neighborhood in Sutton. It consists of the below steps:

Visualise the crime rates in the London boroughs to identify the safest borough and extract the neighborhoods in that borough to find the 15 most common venues in each neighborhood.

Visualise the price per m2 in each borough of London and extract the 15 values with the least values.

* + - Analyse the happiness index and get the one with top 15 values.

### Analysis of crime data

Using the describe function, we could get the statistical values of the columns in the dataset.

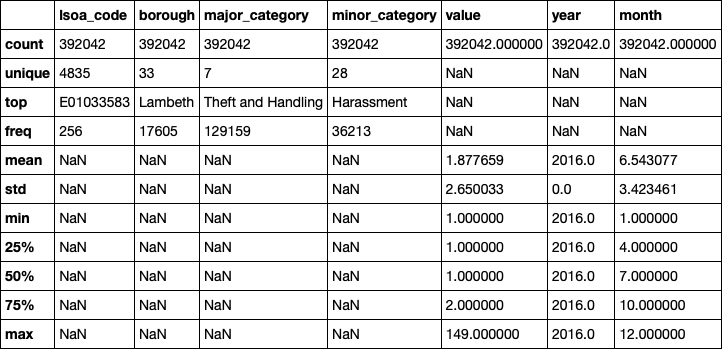


Figure 6: Crime data Analysis

From the above analysis, we could find that among the 33 boroughs of London, Lambeth has the highest crime rate. Out of the 392042 crimes reported in the year 2016, Theft and Handling were most of them.

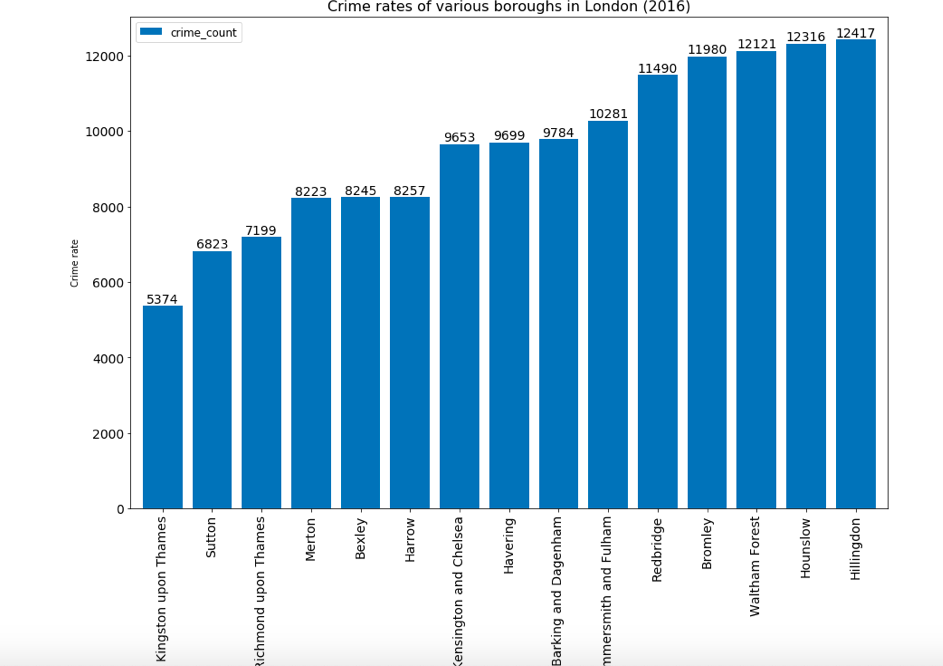


Figure 7: Boroughs with lowest crime rates

From the plot, we could see that Kingston Upon Thames has the lowest crime rate, followed by Sutton, Richmond upon Thmaes and Merton.

### Analysis of Housing price

We need to analyse the boroughs which are affordable by comparing the general trend in the price of the houses in the neighborhood. For that we have built a dataframe with housing price published by ONS.

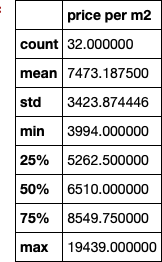


Figure 8: Analysis of Housing price

From the above analysis, the mean value of the housing price per m2 in London is 7473.2, with a minimum value of 3994 and a maximum value of 19439.

Now we can check the boroughs which has the lowest rates:

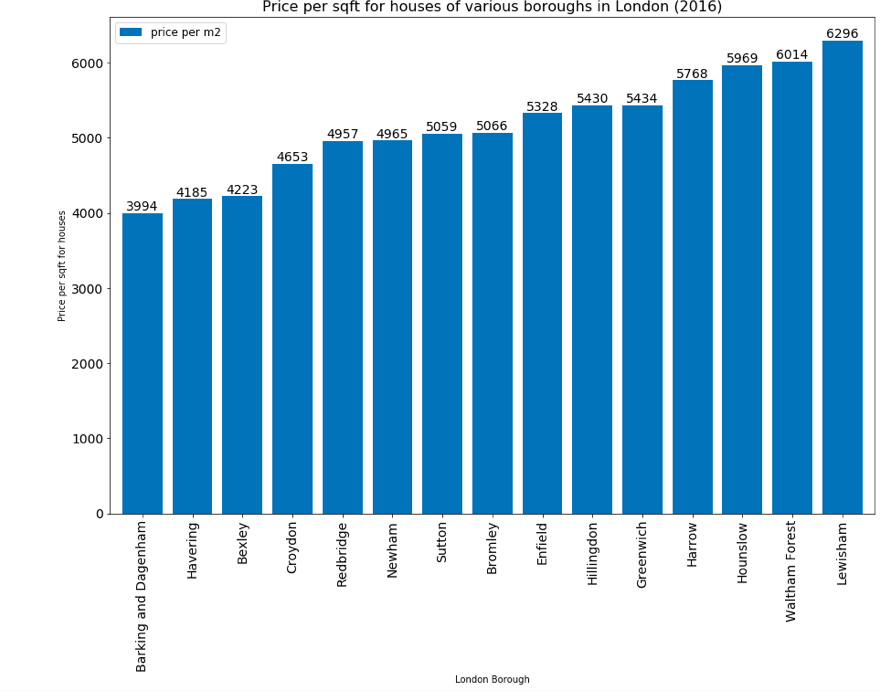


Figure 9: Boroughs with lowest Housing price

From the plot above, some of the most affordable regions in London are in Barking and Dagenham, Havering, Bexley and Croydon.

### Analysis of happiness index

The boroughs with the best happiness index can be shown as:



Figure 10: Boroughs with best Happiness index

Considering the crime rates, affordability in terms of price per sq feet and the happiness index, I have decided that these boroughs in London are the best ones for buying a house: **Sutton, Bromley, Havering, Waltham Forest**, with **Sutton being the best choice.** It is placed second in terms of least crime rate, seventh in affordability, which is much less than the mean value of houses in London(7473.1875) and fourth in happiness index.

## Modelling

Now, using the data set containing the geographical coordinates of Sutton Dis- tricts and FourSquare API, we can find the 100 top most venues in each district. There are 60 unique categories of venues in Sutton. One hot encoding is then performed on these venues to convert the categorical values to a form under- standable by the ML algorithm called binarization.

Now by grouping these venues on the basis of neighborhood, we calculate the mean of the venues. Based on the mean values, we can find out the top 10 venues in each venues.



Figure 11: Top 10 venue categories the neighborhood

Finally we can utilize K-Means clustering algorithm, to group the similar neighborhoods. The neighborhoods are classified into 6 categories using elbow method.

# Results

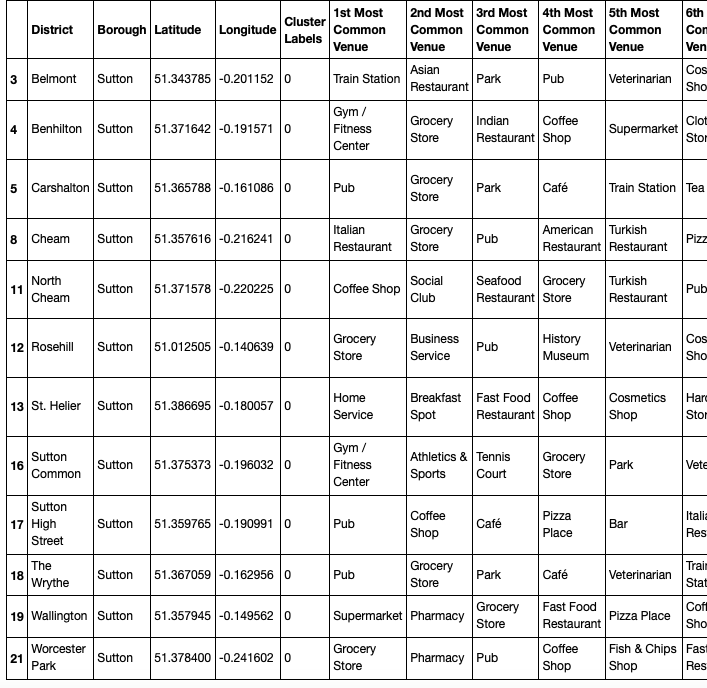


Figure 12: First cluster

The cluster one is biggest with 12 out of 21 neighborhoods; which contains Train station, restaurants, gym, pubs, pharmacy etc. This seems like a happening place of Sutton, with a lot of floating crowd.

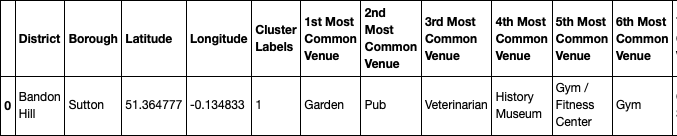


Figure 13: Second cluster

Second cluster consist of Bandon Hill. It has garden, pubs, gym, museum

etc



Figure 14: Third cluster

Third cluster contains three districts which are well connected, with lot of train stations, parks, river, bakery, grocery stores etc. This could be a tourist desti- nation.



Figure 15: Fourth cluster

Fourth cluster is Beddington Corner which has a racetrack, business services and shops. It was not grouped with any other clusters due to the unique venue category.



Figure 16: Fifth cluster

Fifth cluster consist of Bandon Hill with Gardens, hardware stores, pubs,museums etc.



Figure 17: Sixth cluster

Sixth cluster consisting of Little Woodcote & Woodcote Green has lots of garden, parks, coffee shop, gym and grocery stores. This seems like an idyllic place with lots of green space.

# Discussion

After clustering these neighborhoods into 6 clusters, we could see that the first cluster contains almost 12 neighborhood. If people prefer to live in a place where all the amenities like pubs, gym, pharmacy etc are at one hand distance, we could choose the first cluster. These are the areas which attracts a lot of floating crowd. Second cluster consist of only one area which has a museum, garden, pub etc. This could be where the tourists prefer(may be people who wanted to set up an AirBnB can check this area :p) Third cluster is a well connected cluster with train stations being the most common venues. People who travel to work can use this area. Fourth cluster is Beddington Corner which has a racetrack, business services and shops. It was not grouped with any other clusters due to the unique venue category. Fifth cluster consist of Bandon Hill with Gardens, hardware stores, pubs,museums etc. Sixth cluster consisting of Little Woodcote & Woodcote Green has lots of garden, parks, coffee shop, gym and grocery stores. This seems like an idyllic place with lots of green space. For a family, I think **sixth cluster has the best neighborhoods**.

# Conclusion

This project helps me get a better understanding of the neighborhoods with respect to the most common venues in that neighborhood. We have just taken

safety, affordability and accessibity as the primary concerns to shortlist the borough of London. We also have issues in getting the latest dataset for analysis. The future of this project includes taking factors such as cost of living in the areas, employment rate, transportation etc into consideration to shortlist the borough.