

# **Capstone Project - The Battle of Neighborhoods**

**Best Neighborhoods in London Borough**

Author: Soumya Narayanan

Date: January 2020

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
1.1	Background . . . . .	3
1.2	Problem . . . . .	3
1.3	Interest . . . . .	3
<b>2</b>	<b>Data acquisition and cleaning</b>	<b>3</b>
2.1	Data sources . . . . .	3
2.2	Data Cleaning . . . . .	4
<b>3</b>	<b>Methodology</b>	<b>7</b>
3.1	Exploratory Data Analysis . . . . .	7
3.1.1	Analysis of crime data . . . . .	7
3.1.2	Analysis of Housing price . . . . .	9
3.1.3	Analysis of hapiness index . . . . .	10
3.2	Modelling . . . . .	11
<b>4</b>	<b>Results</b>	<b>13</b>
<b>5</b>	<b>Discussion</b>	<b>15</b>
<b>6</b>	<b>Conclusion</b>	<b>15</b>

# 1 Introduction

## 1.1 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 to set 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.

## 1.2 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.

## 1.3 Interest

Obviously, expats who are considering to relocate 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.

# 2 Data acquisition and cleaning

## 2.1 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](#) .
- 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](#) .
- 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 Boroughs revealed by ITV](#) . 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](#)

## 2.2 Data Cleaning

To find the London borough names, we use BeautifulSoup to scrap the wikipage. 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 through out the project, we will be using the Borough names as the index for the data frames.

**Out[4]:**

	London_Borough
0	Barking and Dagenham
1	Barnet
2	Bexley
3	Brent
4	Bromley
5	Camden

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.

	Isao_code	borough	major_category	minor_category	value	year	month
0	E01004177	Sutton	Theft and Handling	Theft/Taking of Pedal Cycle	1	2016	8
1	E01000733	Bromley	Criminal Damage	Criminal Damage To Motor Vehicle	1	2016	4
2	E01003989	Southwark	Theft and Handling	Theft From Shops	4	2016	8
3	E01002276	Havering	Burglary	Burglary in a Dwelling	1	2016	8
4	E01003674	Redbridge	Drugs	Possession Of Drugs	2	2016	11

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](#) 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 the wiki page](#) 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.

	local authority code	local authority name	year	price per m2
0	E06000001	Hartlepool	2016	987
1	E06000002	Middlesbrough	2016	1120
2	E06000003	Redcar and Cleveland	2016	1182
3	E06000004	Stockton-on-Tees	2016	1254
4	E06000005	Darlington	2016	1260
5	E06000006	Halton	2016	1339

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](#) is scrapped and the district names are obtained. Borough name will be Sutton. The geographical coordinated are obtained using **geopy client**.

	<b>District</b>	<b>Borough</b>	<b>Latitude</b>	<b>Longitude</b>
<b>0</b>	Bandon Hill	Sutton	51.364777	-0.134833
<b>1</b>	Beddington	Sutton	51.371988	-0.132393
<b>2</b>	Beddington Corner	Sutton	51.386942	-0.149532
<b>3</b>	Belmont	Sutton	51.343785	-0.201152
<b>4</b>	Benhilton	Sutton	51.371642	-0.191571
<b>5</b>	Carshalton	Sutton	51.365788	-0.161086

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 is 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

	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bandon Hill	51.364777	-0.134833	The Plough	51.367633	-0.132089	Pub
1	Bandon Hill	51.364777	-0.134833	Demesne Rd Allotments	51.362752	-0.140418	Garden
2	Beddington	51.371988	-0.132393	Carew Manor	51.370983	-0.136604	Park
3	Beddington	51.371988	-0.132393	Wickes	51.375476	-0.131014	Hardware Store
4	Beddington	51.371988	-0.132393	Asif Balti House	51.367795	-0.132356	Indian Restaurant
5	Beddington	51.371988	-0.132393	The Plough	51.367633	-0.132089	Pub

Figure 5: Top 100 Venues in each Neighborhoods of Sutton

## 3 Methodology

The methodology in this project consists of two parts:

### 3.1 Exploratory Data Analysis

In this section, we will analysis 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 boroughs of London and extract the 15 values with the least values.
- Analyse the happiness index and get the one with top 15 values.

#### 3.1.1 Analysis of crime data

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

	isoa_code	borough	major_category	minor_category	value	year	month
count	392042	392042	392042	392042	392042.000000	392042.0	392042.000000
unique	4835	33	7	28	NaN	NaN	NaN
top	E01033583	Lambeth	Theft and Handling	Harassment	NaN	NaN	NaN
freq	256	17605	129159	36213	NaN	NaN	NaN
mean	NaN	NaN	NaN	NaN	1.877659	2016.0	6.543077
std	NaN	NaN	NaN	NaN	2.650033	0.0	3.423461
min	NaN	NaN	NaN	NaN	1.000000	2016.0	1.000000
25%	NaN	NaN	NaN	NaN	1.000000	2016.0	4.000000
50%	NaN	NaN	NaN	NaN	1.000000	2016.0	7.000000
75%	NaN	NaN	NaN	NaN	2.000000	2016.0	10.000000
max	NaN	NaN	NaN	NaN	149.000000	2016.0	12.000000

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 were reported in the year 2016, Theft and Handling were most of them.

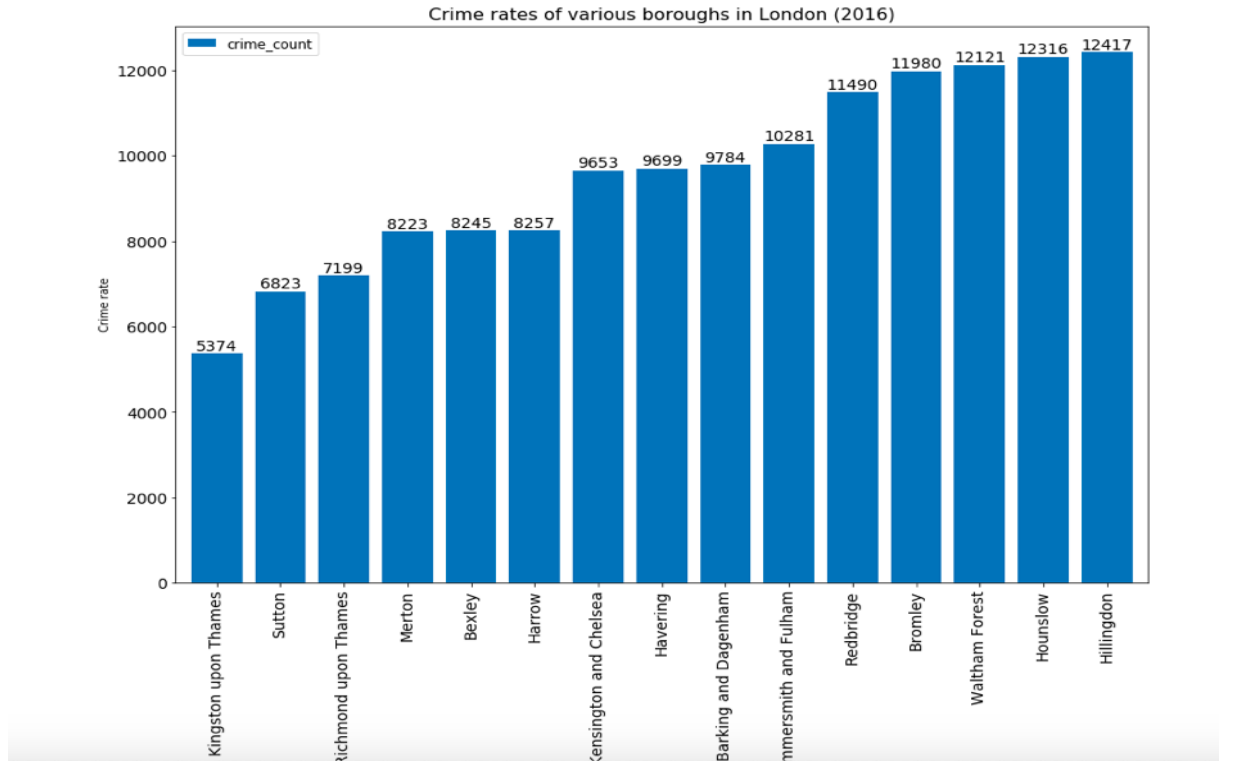



Figure 7: Boroughs with lowest crime rates



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

### 3.1.2 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.



	<b>price per m2</b>
<b>count</b>	32.000000
<b>mean</b>	7473.187500
<b>std</b>	3423.874446
<b>min</b>	3994.000000
<b>25%</b>	5262.500000
<b>50%</b>	6510.000000
<b>75%</b>	8549.750000
<b>max</b>	19439.000000

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:

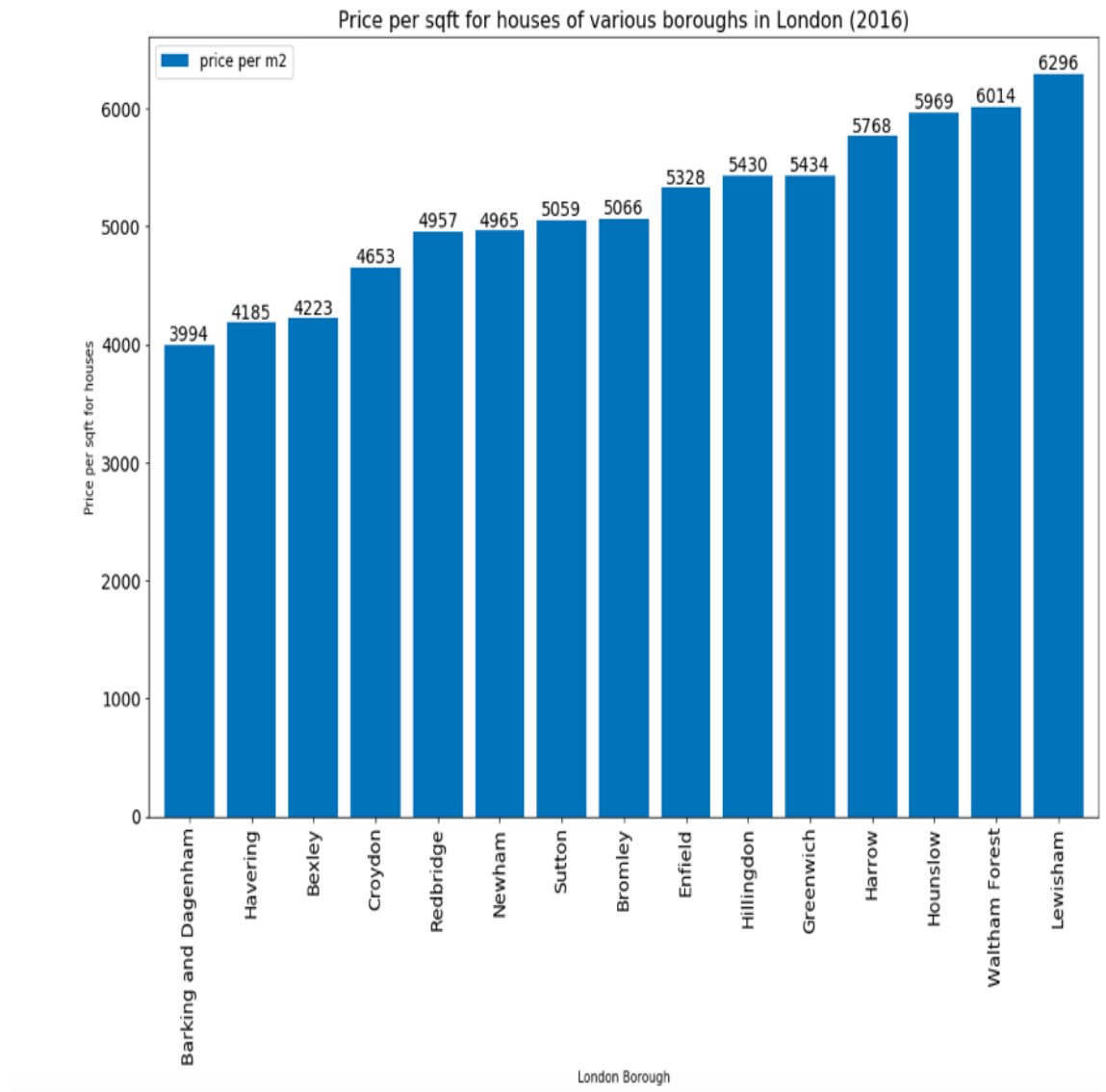


Figure 9: Boroughs with lowest Housing price

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

### 3.1.3 Analysis of happiness index

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

	<b>London_Borough</b>
<b>0</b>	Richmond Upon Thames
<b>1</b>	Kingston upon Thames
<b>2</b>	Bromley
<b>3</b>	Sutton
<b>4</b>	Wandsworth
<b>5</b>	Camden
<b>6</b>	Barnet
<b>7</b>	Ealing
<b>8</b>	Greenwich
<b>9</b>	Havering
<b>10</b>	Hackney
<b>11</b>	Waltham Forest
<b>12</b>	Merton
<b>13</b>	Kensington & Chelsea
<b>14</b>	Hammersmith & Fulham

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.

### 3.2 Modelling

Now, using the data set containing the geographical coordinates of Sutton Districts 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 understandable 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.

	District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
0	Bandon Hill	Garden	Pub	Veterinarian	History Museum	Gym / Fitness Center	Gym	Grocery Store	Gastropub	Garden Center
1	Beddington	Indian Restaurant	Park	Pub	Hardware Store	Veterinarian	Cosmetics Shop	Gym	Grocery Store	Gastropub
2	Beddington Corner	Bridal Shop	Business Service	Racetrack	Veterinarian	Cosmetics Shop	Gym / Fitness Center	Gym	Grocery Store	Gastropub
3	Belmont	Train Station	Asian Restaurant	Park	Pub	Veterinarian	Cosmetics Shop	Gym	Grocery Store	Gastropub
4	Benhlilton	Gym / Fitness Center	Grocery Store	Indian Restaurant	Coffee Shop	Supermarket	Clothing Store	Pizza Place	Park	Fish & Chips Shop

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.

## 4 Results

	District	Borough	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
3	Belmont	Sutton	51.343785	-0.201152	0	Train Station	Asian Restaurant	Park	Pub	Veterinarian	Cos Sho
4	Benhillton	Sutton	51.371642	-0.191571	0	Gym / Fitness Center	Grocery Store	Indian Restaurant	Coffee Shop	Supermarket	Clot Stor
5	Carshalton	Sutton	51.365788	-0.161086	0	Pub	Grocery Store	Park	Café	Train Station	Tea
8	Cheam	Sutton	51.357616	-0.216241	0	Italian Restaurant	Grocery Store	Pub	American Restaurant	Turkish Restaurant	Pizz
11	North Cheam	Sutton	51.371578	-0.220225	0	Coffee Shop	Social Club	Seafood Restaurant	Grocery Store	Turkish Restaurant	Pub
12	Rosehill	Sutton	51.012505	-0.140639	0	Grocery Store	Business Service	Pub	History Museum	Veterinarian	Cos Sho
13	St. Helier	Sutton	51.386695	-0.180057	0	Home Service	Breakfast Spot	Fast Food Restaurant	Coffee Shop	Cosmetics Shop	Har Stor
16	Sutton Common	Sutton	51.375373	-0.196032	0	Gym / Fitness Center	Athletics & Sports	Tennis Court	Grocery Store	Park	Vete
17	Sutton High Street	Sutton	51.359765	-0.190991	0	Pub	Coffee Shop	Café	Pizza Place	Bar	Itali Res
18	The Wrythe	Sutton	51.367059	-0.162956	0	Pub	Grocery Store	Park	Café	Veterinarian	Trail Stat
19	Wallington	Sutton	51.357945	-0.149562	0	Supermarket	Pharmacy	Grocery Store	Fast Food Restaurant	Pizza Place	Coff Sho
21	Worcester Park	Sutton	51.378400	-0.241602	0	Grocery Store	Pharmacy	Pub	Coffee Shop	Fish & Chips Shop	Fast Res

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.

	District	Borough	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
0	Bandon Hill	Sutton	51.364777	-0.134833	1	Garden	Pub	Veterinarian	History Museum	Gym / Fitness Center	Gym

Figure 13: Second cluster

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

	District	Borough	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
6	Carshalton Beeches	Sutton	51.357196	-0.169351	2	Train Station	Grocery Store	Bakery	Italian Restaurant	Veterinarian
9	Hackbridge	Sutton	51.379613	-0.156754	2	Park	Train Station	River	Supermarket	Veterinarian
15	Sutton	Sutton	51.357511	-0.173640	2	Train Station	Grocery Store	Bakery	Italian Restaurant	Park

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 destination.

	District	Borough	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
2	Beddington Corner	Sutton	51.386942	-0.149532	3	Bridal Shop	Business Service	Racetrack	Veterinarian	Cosmetics Shop	Gym / Fitness Center

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.

	District	Borough	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Beddington	Sutton	51.371988	-0.132393	4	Indian Restaurant	Park	Pub	Hardware Store	Veterinarian
14	South Beddington	Sutton	51.371988	-0.132393	4	Indian Restaurant	Park	Pub	Hardware Store	Veterinarian

Figure 16: Fifth cluster

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

	District	Borough	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
10	Little Woodcote	Sutton	51.346076	-0.145932	5	Garden Center	Park	Sports Club	Coffee Shop	Veterinarian
20	Woodcote Green	Sutton	51.347991	-0.146830	5	Park	Garden Center	Café	Coffee Shop	Construction & Landscaping

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.

## 5 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**.

## 6 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 accessibility 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.