IBM Data Science Professional Certificate Capstone Project – Battle of Neighbourhoods

Table of Contents

1. Introduction/Business Problem

1.1 Problem Statement

In the country of United Kingdom, London is the one of the hotspots for IT related jobs. In this project, one of the main problems that the people who started their career in IT can face will be targeted. From the beginning of the process, there are various problems. Even though there is a plenty of demand for skilled people out there, landing a good job has been always hard. However, a problem that can be faced after landing a good job will be treated in this project. Supposedly the person has already landed their job, and now it is time to find a place to live. Finding a good neighbourhood is always a big concern to deal with. Especially in big cities like London, the dilemma of price and quality can be challengingly sharp. While the price range can easily widen, the person may also end up living an undesirable neighbourhood. High average of rental prices may be a sign for more enjoyable or safer neighbourhood.

1.2 Approach for Problem

A reasonable approach to this problem is comparing crime numbers and rental prices of places for whole city of London and then analysing neighbourhood according to their amenities for our IT person. To start with, one assumption is made. Wage of a person who makes their living as an IT specialist will probably be more than average wage and this person will probably want to live somewhere close to central London, to benefit London's beauties. For this reason, the first step is filtering London boroughs by their inner/outer designation. Inner London boroughs will be the matter of this project. Then London boroughs will be compared by using the rental price and crime number aspects. London boroughs will be ordered by the total number of crimes and most dangerous boroughs will be removed from candidate list. Then the remaining boroughs will be ordered by rental price. There is a tricky point which should be taken into consideration here. By knowing the supply and demand relationship from Economy, low price can be an indicator of low demand which can be caused by insufficient social amenities of neighbourhood. Therefore, a medium price band will be focussed, instead of focusing on the lowest price band. After finding our target borough, neighbourhoods of that borough will be analysed by their venues. As a young, relatively comfortable in financial person may have different expectations from where they live. Those neighbourhoods will be clustered by their venues to come up with potential best places to live for our IT person.

2 Data

2.1 Data Description

For this project, the following data will be used:

- List of London Boroughs with their designations: https://en.wikipedia.org/wiki/London_boroughs
- Number of Crimes by London Boroughs: https://drive.google.com/file/d/1TTOhuQit4gkenrUizsUn9wloHalRffhE/view

- London Rental Prices by Boroughs: https://www.ons.gov.uk/peoplepopulationandcommunity/housing/adhocs/11100privaterentalmarketinlondonjanuary2019todecember2019

- Foursquare API

However, these sources give the data for multiple purposes. This is why they are needed to be cleaned and wrangled accordingly. Some of the columns and row may be needed to be removed, some others may be combined, some nonsense entity may be handled according to need.
First source is a Wikipedia page. Therefore, it is needed to be scrapped by using python library. The second is a csv file which can easily be turned to Pandas dataframe, however it does not have total number of crimes. It has data for different kind of crimes separately from March 2018 to February 2020. So there will be some calculations to generate useable data for our purpose. The third source is an xls file with some of text before the needed table. Also the table is formed in a different structure. It has data for different types of properties separately, instead of an average rental price for each borough. This source has also rental price data recorded between January 2019 to December 2019. So there will be calculation to find average rental prices for each boroughs as well. Final source is Foursquare's API that we used before, during the course. Course's instructions will be followed to obtain venue data for neighbourhoods of selected borough.

2.1.1 Crime Data Explanation:

Major Text: A general classification that express the type of crime committed.
Minor Text: A little more detailed classification for the crime.
Borough Name: The name of the corresponding London borough.
201xxx: The number of corresponding crime in the year and the month.
Number of Crime: Total number of crime committed in a borough.

2.1.2 London Rental Price Data:

Borough: Name of the London Borough.
Bedroom Category: Type of the property according to number or type of bedroom.
Count of Rents: Number of rented property in a class of corresponding borough and bedroom category.
Mean: The arithmetic average of rental prices in GBP for corresponding property type.
Avg_Rent: Average rental price for a property in a borough regardless of property type.

2.1.3 Foursquare Data Explanation:

Neighbourhood: Name of the neighbourhood in the selected London borough.
Latitude: Latitude value of the location for corresponding neighbourhood
Longitude: Longitude value of the location for corresponding neighbourhood.
Venue: Name of the venue
Venue Latitude: Latitude value of the venue location
Venue Longitude: Longitude value of the venue location
Venue Category: Category of the venue

2.2 Data Wrangling/Handling/Cleaning

2.2.1 Crime Data

The data tells about crime was a csv file. Data handling process for this part started with Pandas read_csv function. This function transformed csv file to dataframe.

```
df_crime = pd.read_csv('Downloads/London_crime.csv')
```

```
df_crime.head()
```

	MajorText	MinorText	LookUp_BoroughName	201803	201804	201805	201806	201807	201808	201809	...	201905	201906	201907	201908	201909	201
0	Arson and Criminal Damage	Arson	Barking and Dagenham	6	3	4	12	6	5	3	...	11	3	5	3	6	
1	Arson and Criminal Damage	Criminal Damage	Barking and Dagenham	115	122	126	123	127	101	107	...	138	113	134	118	109	
2	Burglary	Burglary - Business and Community	Barking and Dagenham	38	36	24	33	30	18	33	...	22	27	31	35	37	
3	Burglary	Burglary - Residential	Barking and Dagenham	122	75	93	77	94	84	99	...	114	96	71	67	80	
4	Drug Offences	Drug Trafficking	Barking and Dagenham	7	3	8	6	9	7	10	...	8	6	8	6	6	

5 rows × 27 columns

As it can be seen clearly, this data set has crime numbers individually for each month between March 2018 to February 2020. Numbers from each month will be added to obtain a total number of crime for each type of crime.

```
cols = df_crime.columns[3:]
df_crime['2018-2020']= df_crime[cols].sum(axis=1)
df_crime=df_crime.drop(df_crime.columns[3:-1],axis=1)
df_crime.head(10)
```

	MajorText	MinorText	LookUp_BoroughName	2018-2020
0	Arson and Criminal Damage	Arson	Barking and Dagenham	129
1	Arson and Criminal Damage	Criminal Damage	Barking and Dagenham	2775
2	Burglary	Burglary - Business and Community	Barking and Dagenham	726
3	Burglary	Burglary - Residential	Barking and Dagenham	2430
4	Drug Offences	Drug Trafficking	Barking and Dagenham	155
5	Drug Offences	Possession of Drugs	Barking and Dagenham	1987
6	Miscellaneous Crimes Against Society	Bail Offences	Barking and Dagenham	1
7	Miscellaneous Crimes Against Society	Bigamy	Barking and Dagenham	2
8	Miscellaneous Crimes Against Society	Dangerous Driving	Barking and Dagenham	26
9	Miscellaneous Crimes Against Society	Disclosure, Obstruction, False or Misleading S...	Barking and Dagenham	1

Then groupby method is used to have total number of crime for each borough.

```
df_crime=df_crime.groupby('Borough',as_index=False)[['2018-2020']].sum()
df_crime.head(10)
```

	Borough	2018-2020
0	Barking and Dagenham	38796
1	Barnet	60534
2	Bexley	34099
3	Brent	61154
4	Bromley	48810
5	Camden	76081
6	Croydon	65738
7	Ealing	60811
8	Enfield	59263
9	Greenwich	55330

2.2.2 Rental Price Data

The format of this data is xls file. The original file is formed by multiple sheets and has some text before the table that is needed for the analysis. This is because second sheet is used and the first 13 lines are removed. Then it is transformed to a pandas dataframe.

```
df_rent = pd.read_excel('Downloads/londonrentalstatisticsq42019.xls',sheet_name='Table 1.2',header=11,usecols="B:H",nrc
df_rent=df_rent.drop(columns=df_rent.columns[4:])
df_rent.head(10)
```

	Borough	Bedroom Category	Count of rents	Mean
0	Barking and Dagenham	Room	20	482
1	Barking and Dagenham	Studio	20	742
2	Barking and Dagenham	One Bedroom	220	984
3	Barking and Dagenham	Two Bedrooms	380	1190
4	Barking and Dagenham	Three Bedrooms	260	1428
5	Barking and Dagenham	Four or More Bedrooms	40	1653
6	Barnet	Room	20	544
7	Barnet	Studio	110	864
8	Barnet	One Bedroom	400	1149
9	Barnet	Two Bedrooms	770	1388

After a brief look, it was found that there are some missing values which is in for of non-integer data. These rows are dropped. Then in order to obtain an average data for each borough regardless of bedroom type, average rental prices for each bedroom category multiplied by count of rents. Obtained values are summed then divided to total number of count of rents. After that, rows are grouped according to their borough values by using groupby method.

```
df_rent.drop(index=[36,40,41,156],inplace=True)
df_rent['Tot_Mean'] = df_rent['Mean'] * df_rent['Count of rents']
df_rent=df_rent.drop(['Bedroom Category','Mean'],axis=1)
df_rent=df_rent.groupby(['Borough'],as_index=False).sum()
df_rent['Avg_rent'] = df_rent['Tot_Mean'] / df_rent['Count of rents']
df_rent=df_rent.drop(['Count of rents','Tot_Mean'],axis=1)
```

```
df_rent=df_rent.sort_values(by='Avg_rent').reset_index(drop=True)
df_rent
```

	Borough	Avg_rent
0	Bexley	1119.492958
1	Croydon	1126.730924
2	Sutton	1158.394231
3	Havering	1159.202020
4	Barking and Dagenham	1202.723404
5	Hillingdon	1218.068966
6	Lewisham	1299.909091
7	Waltham Forest	1310.406780
8	Enfield	1313.475610
9	Redbridge	1319.348571

2.2.3 London Borough Data

The list of London Boroughs is found on Wikipedia. Therefore, it is needed to be scrapped. However, the original table had some unnecessary data as well. They are excluded during the web scrapping procedure. Finally, a pandas dataframe which includes list of London boroughs and their designation about if they are inner London or outer London is obtained.

```
df_Borough_designation = pd.DataFrame(list(zip(rows_data_col1,rows_data_col2)),columns=col_names_table1)
```

```
df_Borough_designation
```

	Borough	Designation
0	Camden	Inner
1	Greenwich	Inner
2	Hackney	Inner
3	Hammersmith[notes 2]	Inner
4	Islington	Inner
5	Kensington and Chelsea	Inner
6	Lambeth	Inner
7	Lewisham	Inner
8	Southwark	Inner
9	Tower Hamlets	Inner
10	Wandsworth	Inner
11	Westminster	Inner
12	Barking[notes 3]	Outer
13	Barnet	Outer

2.2.4 Foursquare Data

Foursqare data is obtained by following the instructions from lab exercises which were done before through the course. To access the venues around the neighbourhood, radius is set to 500 meters and limit is set to 30.

2.2.5 Extra Data

In the process of the project, there are some data handling steps between analysis methodology stages as well. Besides the main data sources, a list of neighbourhood of selected borough is used too. This list is from Wikipedia. The list is transformed to python list manually. Then thanks to the libraries, location data is obtained for these neighbourhoods. Names of neighbourhoods, longitudes and latitudes formed a pandas dataframe.

```
df_Neigh_Wandsworth = pd.DataFrame(list(zip(neigh_Wandsworth,Latitude,Longitude)),columns=['Neighbourhood','Latitude','
df_Neigh_Wandsworth
```

	Neighbourhood	Latitude	Longitude
0	Balham	51.445645	-0.150364
1	Battersea	51.470793	-0.172214
2	Earlsfield	51.446448	-0.189394
3	Furzedown	51.424389	-0.153702
4	Nine Elms	51.478743	-0.136263
5	Putney	51.462552	-0.216746
6	Putney Heath	51.442842	-0.232207
7	Putney Vale	51.438019	-0.245970
8	Roehampton	51.449877	-0.241267
9	Southfields	51.445775	-0.206614
10	Streatham Park	51.424939	-0.145060
11	Tooting	51.426659	-0.169077
12	Tooting Bec/Upper Tooting	51.435609	-0.159655
13	Wandsworth	51.457027	-0.193261

3 Methodology

Methodology used for this project has two steps.

3.1 Exploratory Data Analysis

The intuition which is necessary for the purpose of project is gained from the datasets mentioned above. Most dangerous boroughs of London are excluded immediately. Then percentiles of rental prices data are checked and the price band of 25% to 75% is used for the further analysis. The remaining set of borough is sorted and most appropriate borough is selected. Then neighbourhoods of selected borough are listed. Finally, 10 most common venues of each neighbourhood is obtained by using Foursquare API

3.2 Modelling

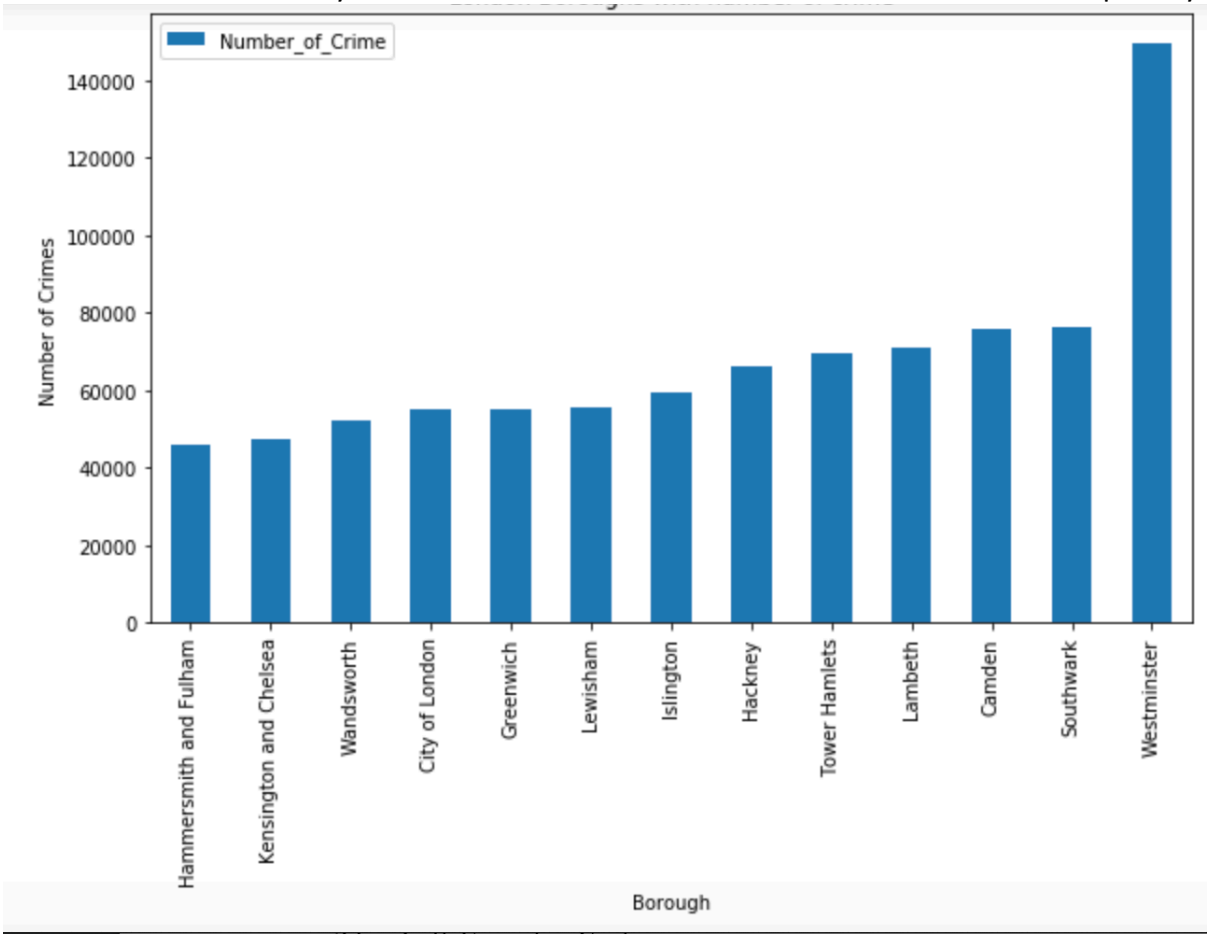
In order to come up with ideal neighbourhoods to live in a safe and moderately pricy borough, K-means clustering method is used. K-mean clustering method clustered neighbourhoods with similar venues. The number of cluster is predefined and it is 5, since the number of neighbourhood is low.

4 Analysis

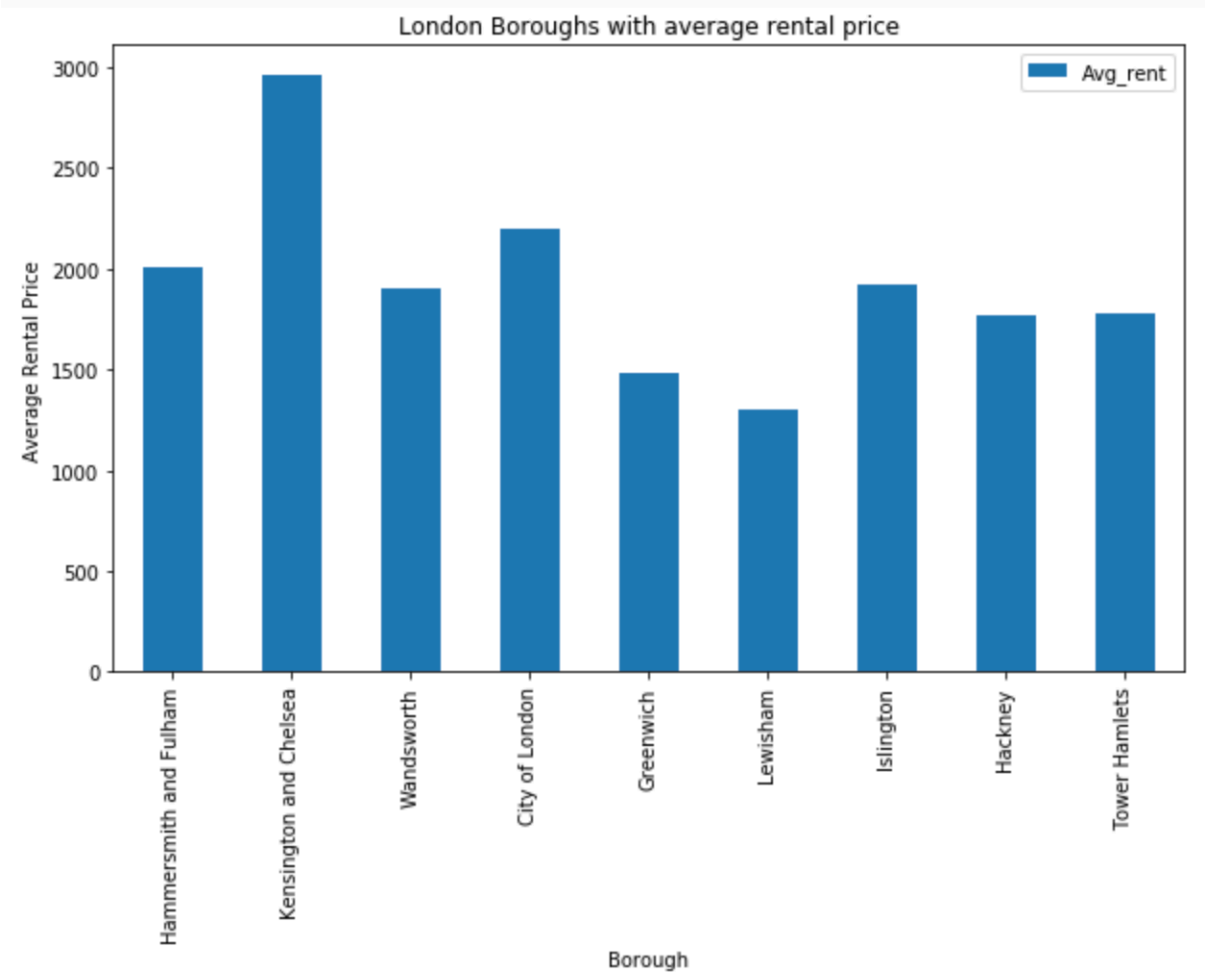
4.1 Exploratory Analysis

Analysis started with excluding outer London Boroughs. Since our targeted stakeholder is a position to demand more from being a Londoner, far boroughs of London would not even be option for us.

After that, list of London boroughs are sorted according to their total number of crimes, because safety is the first priority.



While there is not massive difference between each borough, some boroughs like Westminster has enormous number of crime. Therefore, the most dangerous 25% of the all London boroughs are removed from the list, before even checking the average rental prices. Next step is checking rental prices for each of remaining London boroughs.



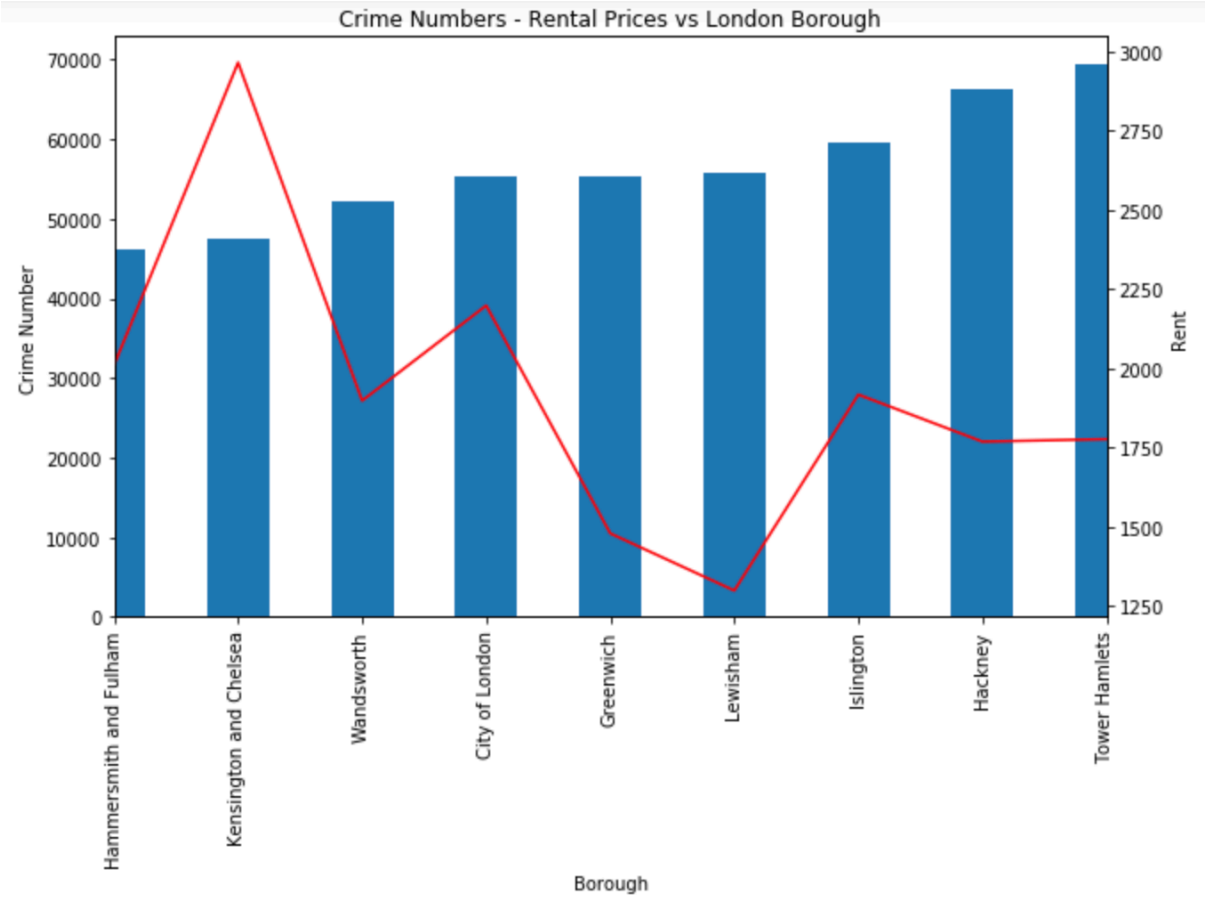
Besides the first look, fundamental statistical information about rental prices are also considered.

```
df['Avg_rent'].describe()
```

```
count       9.000000
mean     1924.193351
std       474.128656
min      1299.909091
25%      1769.586207
50%      1898.714829
75%      2010.779070
max      2964.680180
```

Standard derivation is about 25% of mean value. Maximum average rent is almost £3000, while the minimum average rent is still £1300. In my personal intuition, the lowest point of

economic indicators for an area generally shows lower living quality. On the other hand, the most expensive boroughs are hard to be afforded.



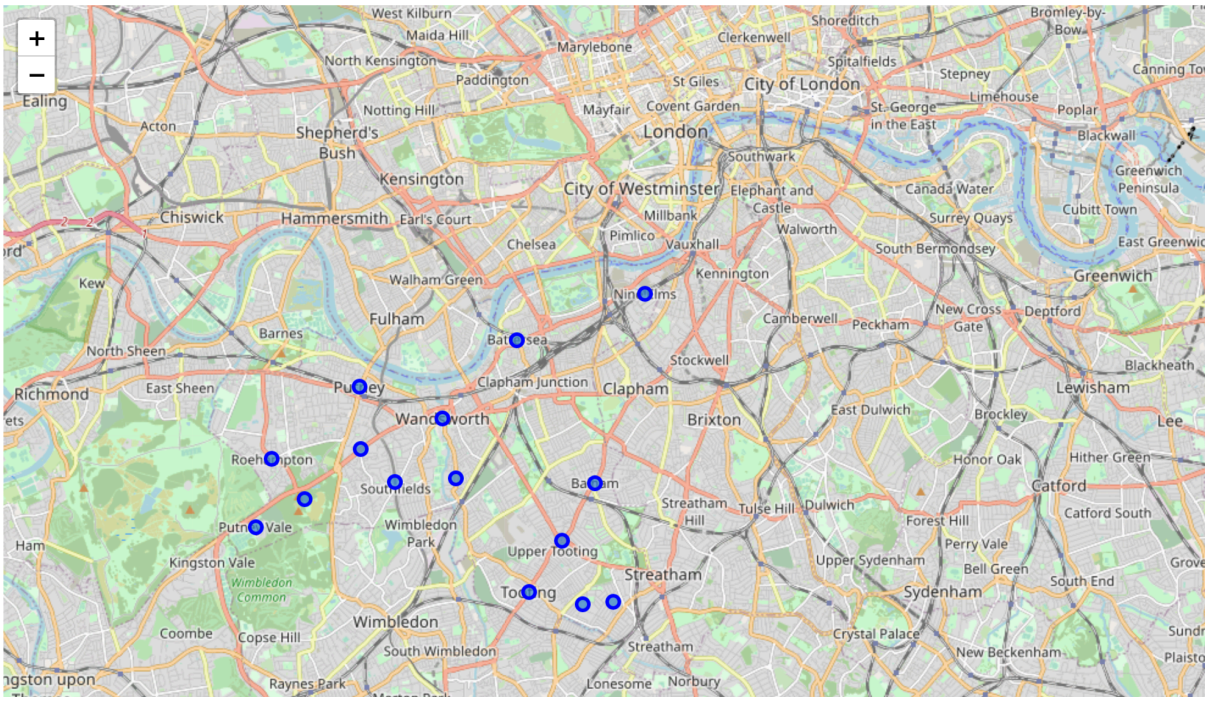
Therefore, the range from 25% to 75% according to average rental price is taken as for the purpose, by considering the fact std/mean ratio is 25%. The final list obtained in this stage is:

	Borough	Avg_rent	Number_of_Crime
2	Wandsworth	1898.714829	52270
6	Islington	1917.857143	59540
7	Hackney	1769.586207	66324
8	Tower Hamlets	1777.375000	69452

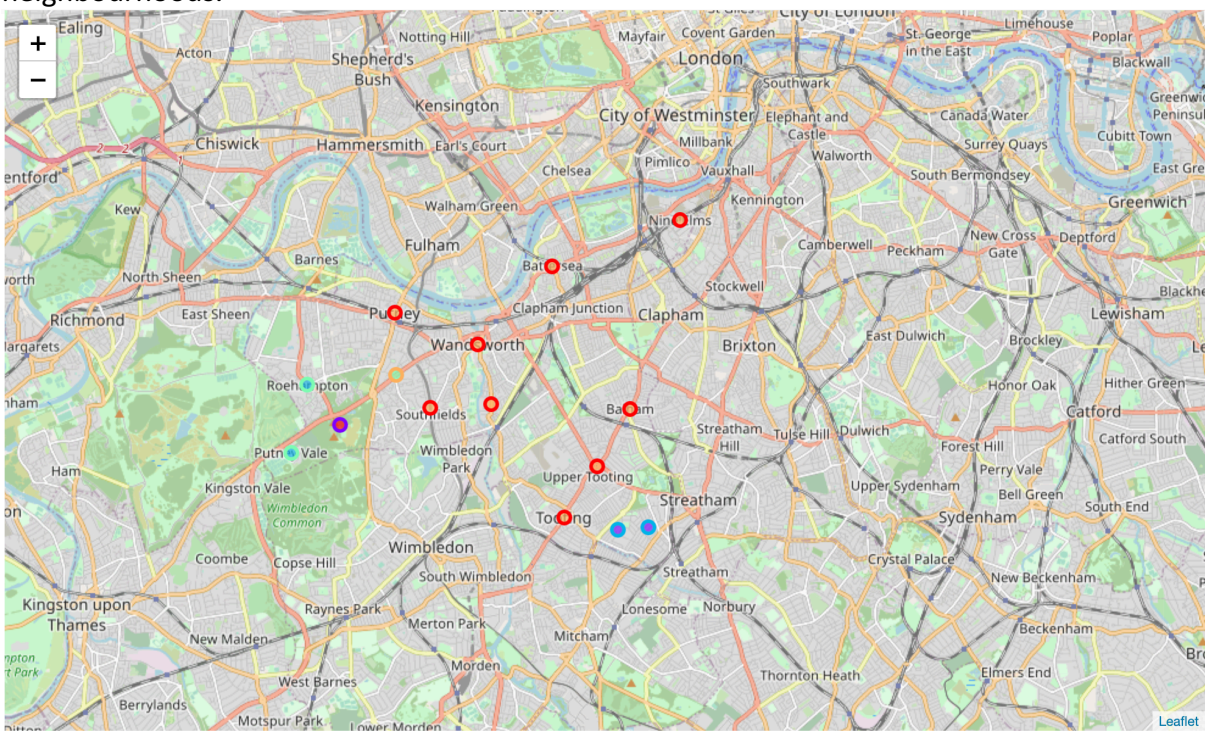
Thus, Wandsworth has the lowest number of crime within reasonable and desirable price range and this situation makes it a potentially best place to move.

4.2 Modelling

After deciding on Wandsworth borough, neighbourhoods of Wandsworth are listed. Then location data, longitude and latitude, is collected for each neighbourhood.



Then, all the venues are collected within the radius of 500 meters for each neighbourhood. This data for venues includes Venue name, Venue longitude, Venue latitude and Venue category. There are 86 unique venue category. One hot encoding is applied for venue category, because K means clustering is done according to venue category. Encoded data is grouped by neighbourhood names and final outcome is matrix of each neighbourhood and their venues according to venue categories. After K means clustering, 5 clusters are obtained. First cluster has most of the neighbourhoods, while the other clusters have one or two neighbourhoods.



5 Results

According to results of clustering, each cluster has slightly different characteristics. Even though each cluster seems like they have good amenities, there are some obvious distinctions as well. First cluster has more pubs and cafes. It also has variety of restaurants from different cuisines. It also has gym and fitness facilities. Most of the neighbourhoods in this cluster also have bus stops. Second cluster is more rural area. It has outdoor activity facilities and fastfood options. However, there is no information about public transportation. Third cluster does not have fitness centres or gyms while it has most of the other facilities. The other clusters are more or less same. They have grocery stores and supermarkets which may be needed daily. However, they do not have any restaurant options.

Venues categories for cluster 1:

	Neighbourhood	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	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Balham	0	Pub	Coffee Shop	Italian Restaurant	Bakery	Bar	Pizza Place	Beer Store	Café	English Restaurant	Farmers Market
1	Battersea	0	Pub	Bakery	Hotel	Italian Restaurant	Grocery Store	Plaza	Pet Store	Japanese Restaurant	Heliport	Harbor / Marina
2	Earlsfield	0	Grocery Store	Thai Restaurant	Pub	Italian Restaurant	Café	Indoor Play Area	Gym	Lounge	Gastropub	Music Venue
4	Nine Elms	0	Grocery Store	Bar	Coffee Shop	Gym / Fitness Center	Indian Restaurant	Pizza Place	Fish & Chips Shop	Restaurant	Café	Bus Stop
5	Putney	0	Japanese Restaurant	Café	Burger Joint	Pub	Coffee Shop	Gastropub	Pizza Place	Pie Shop	Ice Cream Shop	Gym / Fitness Center
9	Southfields	0	Coffee Shop	Grocery Store	Pub	Italian Restaurant	Bus Stop	Gym	Lebanese Restaurant	Flea Market	Park	Pharmacy
11	Tooting	0	Pub	Indian Restaurant	Bar	Market	Coffee Shop	Asian Restaurant	Café	Juice Bar	Lebanese Restaurant	Middle Eastern Restaurant
12	Tooting Bec/Upper Tooting	0	Coffee Shop	Indian Restaurant	Bakery	Pub	Italian Restaurant	Restaurant	Bar	Burger Joint	Convenience Store	Gas Station
13	Wandsworth	0	Pub	Asian Restaurant	Supermarket	Coffee Shop	Clothing Store	Restaurant	Portuguese Restaurant	Chocolate Shop	Gym / Fitness Center	Pharmacy

Cluster 2:

	Neighbourhood	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	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
6	Putney Heath	1	Park	Baseball Field	Yoga Studio	Gastropub	Fast Food Restaurant	Fish & Chips Shop	Flea Market	Food Truck	Fried Chicken Joint	Garden Center

Cluster 3:

	Neighbourhood	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	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
3	Furzedown	2	Italian Restaurant	Convenience Store	Park	Chinese Restaurant	Café	Pizza Place	Gas Station	Fish & Chips Shop	Flea Market	Food Truck
10	Streatham Park	2	Convenience Store	Pub	Park	Grocery Store	Café	Yoga Studio	Gas Station	Fish & Chips Shop	Flea Market	Food Truck

Cluster 4:

	Neighbourhood	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	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
7	Putney Vale	3	Bus Stop	Gas Station	Grocery Store	Supermarket	Outdoors & Recreation	Café	Yoga Studio	Fish & Chips Shop	Flea Market	Food Truck
8	Roehampton	3	Café	Bus Stop	Bakery	Bar	Wine Shop	Paper / Office Supplies Store	Supermarket	Pub	Grocery Store	Bus Station

Cluster 5:

	Neighbourhood	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	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
14	West Hill	4	Gym	Park	Bus Stop	Wine Shop	Pub	Scenic Lookout	Bus Station	Music Venue	Gas Station	Fish & Chips Shop

6 Discussion

Even though it may change from individual to individual, as a young IT professional, I would like to live a neighbourhood with social life and fancy facilities and where safety is not a big concern. In the first cluster, we can see that there are a lot different options for social life. It also has enough option for eating your dinner out. On the other hand, you can still find grocery store, if you want to do your shopping and cook for yourself. Having a gym or a fitness centre close enough to your house is also a plus for healthy and enjoyable life.
My personal choice would be Nine Elms, because it has all necessary amenities, also public transportation and it is the nearest neighbourhood to heart of London.

7 Conclusion and Further Studies

This project enables us to analyse London in different perspectives. Each person has their own needs, concerns and priorities. Definition of a best place to call ‘home’ can massively change person to person. A safe and affordable neighbourhood with satisfactory amenities for a young IT professional is what we needed to find out for this project. However, the main focus of this project is problems can be sorted by using the data. In order to produce better outcomes, input should be enhanced. Some other metrics can be added. For example, happiness indexes, or public transportation statistics depending on person’s workplace. While simplicity makes decision making process easier, comprehensive analysis on extensive data will make understanding clear.