Location Recommendations in London

Priyesh Gupta

June 9th, 2020

1. Introduction

1.1 Background

London, the capital city of England and United Kingdom, is one of the most happening places in the world. Business, education, travel and what not. It is a destination in every person's bucket list. Now, every person encounters one major question, LOCATION!. In case of a business where to set it up, let's say restaurant in London. Which place would be the best to set up a restaurant or which place would be good for a service providing business. Similarly if you are a student who has moved in for education, where should you rent a space for living. Whether you want peaceful locations or you prefer to be around happening spots. I have created a very generalised model to answer these questions which can give recommendations for places in any part of the world. In this project, I will run the model specifically for London.

1.2 Problem

An important determinant while determining a location fit for any particular purpose is the kind of neighborhood it is in. What are the various venues in the neighborhood. Can these be of benefit to the stakeholder. A student like me would love to be in a neighborhood with some good cafes and party places for weekends. While a service provider would like to start his business in a

neighborhood with high density of offices or homes depending on the type of service to be provided, a restaurant could be opened in an area with less no. of restaurants to have lesser competition.

1.3 Interest

Obviously, every person would be interested in getting location recommendations. Infact, a similar kind of approach is used by several planners for tourist recommendations as well.Thus, it has a great scope for everyone and anyone.

2. Data Acquisition and Cleaning

2.1 Data Sources

The first requirement is the set of Neighborhoods and Boroughs in London. This data was obtained from wikipedia. I scrapped this page with the help of Panda. Alternatively, this task can also be done with the help of BeautifulSoup package. This helped in getting a good dataframe with all Boroughs and neighborhoods. Some formatting had to be done so as to make data more handy for algorithms.

The latitudes and longitudes needed for plotting on map was obtained with help of geopy.gecoders package.

Now, the most important data, on the basis of which we will cluster neighborhoods, the location based data was obtained by the help of Foursquare API.

2.2 Data Cleaning

Initial data after scrapping from wikipedia:

	Location	London borough	Post town	Postcode district	Dial code	OS grid ref
0	Abbey Wood	Bexley, Greenwich [7]	LONDON	SE2	020	TQ465785
1	Acton	Ealing, Hammersmith and Fulham[8]	LONDON	W3, W4	020	TQ205805
2	Addington	Croydon[8]	CROYDON	CR0	020	TQ375645
3	Addiscombe	Croydon[8]	CROYDON	CR0	020	TQ345665
4	Albany Park	Bexley	BEXLEY, SIDCUP	DA5, DA14	020	TQ478728
5	Aldborough Hatch	Redbridge[9]	ILFORD	IG2	020	TQ455895
6	Aldgate	City[10]	LONDON	EC3	020	TQ334813

All the irrelevant columns like Postcode district, Dial code, OS grid reference data were dropped right in the first step.

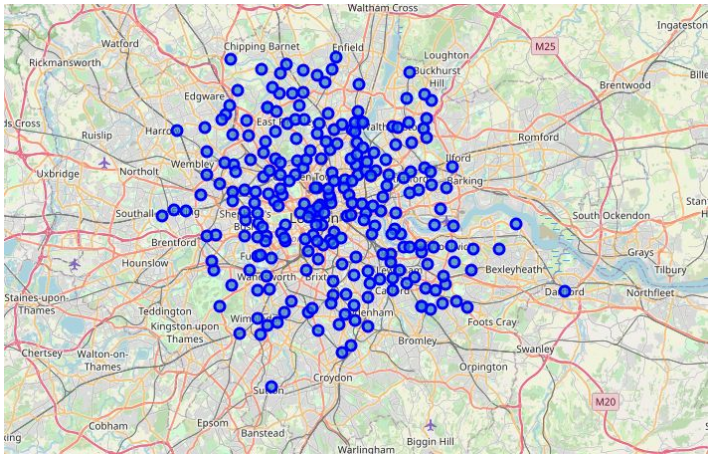
A lot of useless data came along while scrapping which had to be removed. With the names of Boroughs and neighborhoods, several superscript texts also came in which had to be removed.

A glimpse of the midway cleaned data:

	Borough	Neighborhood	Post town	latitude	longitude
0	Barnet	Barnet Gate	LONDON, BARNET	51.641827	-0.242985
1	Greenwich	Blackheath Royal Standard	LONDON	51.477735	0.020490
2	Barnet	Brent Cross	LONDON	51.576760	-0.218380
3	Brent	Brent Park	LONDON	51.563826	-0.275760
4	Lewisham	Catford	LONDON	51.445321	-0.019753

Latitude and Longitude data was apparently not available for all neighborhoods. In this case we can try using some other service provider like Google. I simply ignored those neighborhoods as their number was very low (5 out of 263) and plotted on map to see if everything works fine.

This is how the map looks with neighborhoods plotted:



Then, the foursquare API was used to get venues in all the locations of neighborhoods and borough so obtained. The json files received from API calls were put in loops with appropriate Keys of nested dictionaries and indices of lists so as to just extract relevant informations which included Venue Name, Venue Latitude, Venue Longitude, Venue Category.

The Venue category is the most important thing for clustering while latitude and longitude data for visualization of clusters on map.

The final cleaned data looked like the following:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Barnet Gate	51.641827	-0.242985	The Gate	51.641994	-0.242582	Pub
1	Barnet Gate	51.641827	-0.242985	Hadley FC	51.642660	-0.243032	Soccer Field
2	Barnet Gate	51.641827	-0.242985	Barnet Gate Wood	51.639062	-0.242836	Forest
3	Blackheath Royal Standard	51.477735	0.020490	Mara Interiors & Café	51.477672	0.019368	Furniture / Home Store
4	Blackheath Royal Standard	51.477735	0.020490	M&S Simply Food	51.476772	0.020189	Grocery Store

3. Methodology

3.1 Clustering :

Cluster analysis or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense) to each other than to those in other groups (clusters). It is basically a type of *unsupervised learning method* . An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labelled responses. Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples.

3.2 Why Clustering?

Clustering is very much important as it determines the intrinsic grouping among the unlabeled data present. There are no criteria for a good clustering. It depends on the user, what is the criteria they may use which satisfy their need. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding “natural clusters” and describe their unknown properties (“natural” data types), in finding useful and suitable groupings (“useful” data classes) or in finding unusual data objects (outlier detection). This algorithm must make some assumptions which constitute the similarity of points and each assumption make different and equally valid clusters.

3.3 K-means Clustering:

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. In my project, I will be using the sklearn library’s inbuilt k- means Clustering algorithm to cluster neighborhoods of London on the basis of the similarity of the venues in them. This will result in neighborhoods having similar type of venues or precisely, similar types of most visited venues will be clustered together. This means that they will be given same cluster label and accordingly a fixed color for visualization on the map.

4. Result

After a lot of trials for different cluster values, I decided to use 14 clusters, this resulted in getting 7 major clusters while the others can be treated as outliers. These clusters could be identified with their specific characteristics differentiating each of them from the others. The major clusters and their characteristics are as follows:

Cluster 11

Shopping centres, hotels, entertainment, parks, good places for tourists

Cluster 3

Party place, high density of pubs. And also some other shopping centres and stores.

Cluster 8

High density of restaurants, eateries and cafeterias

Cluster 1

Good bus connectivity, gymnasiums and sports, good density of grocery shops and supermarkets, low density of restaurants

Cluster 5

Gorcery stores, parks, decent connectivity and entertainment sources. Ideal living places for families

Cluster 4

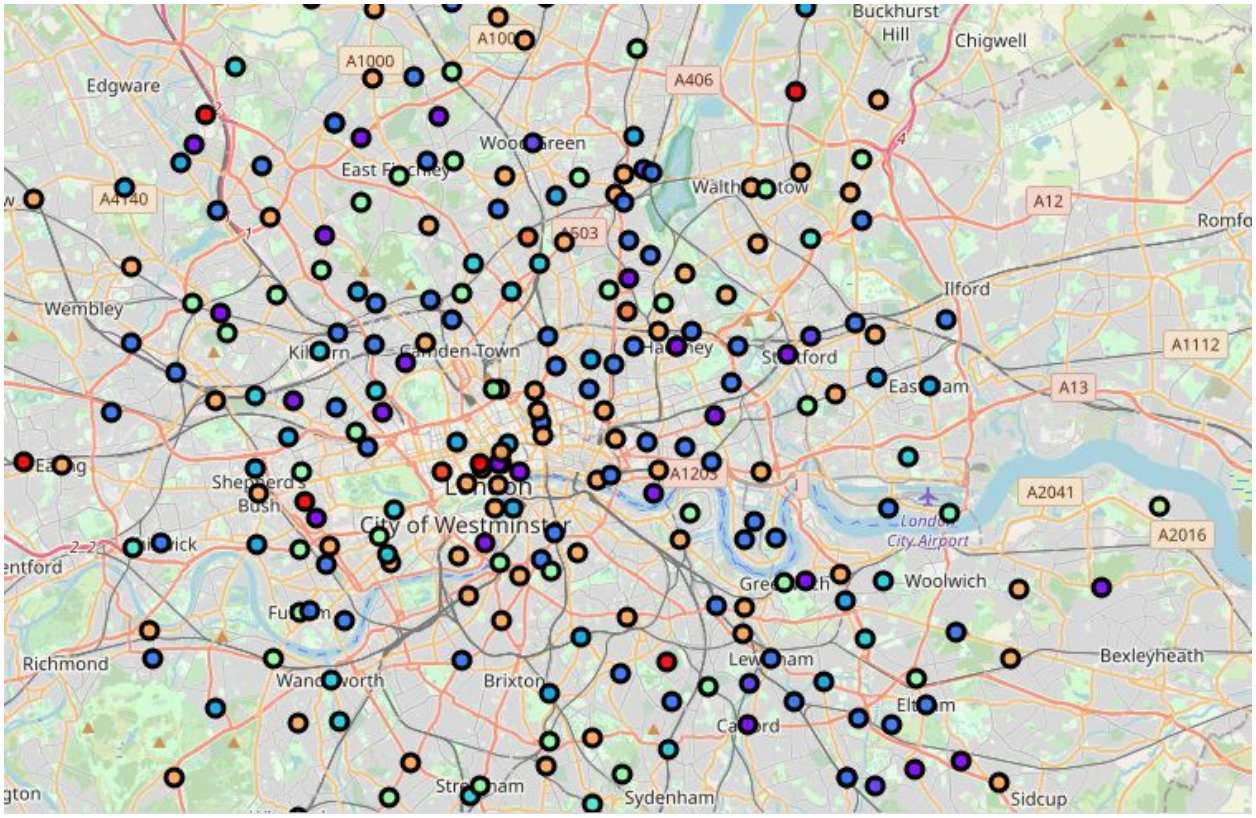
Complete **Party place**, very high density of Pubs. Also moderate no. of restaurants and minimal other services

Cluster 0

Nearby platforms and a mix of all shops, stores, restaurants, cafes

Cluster 2

Parks, farms, pet stores and zoos. Very low density of restaurants and event spaces.



5. Conclusion

On the basis of the distinct characteristics of neighborhoods, identified by clustering, we can recommend locations to Stakeholders. For example, a person like me would like to be around neighborhoods of cluster 3 or 4, with a lot of happening places and party spots. A person who wants to live closer to nature and in peace could choose to be in neighborhoods of cluster 2. Those who want to stay close to good public transportation means could choose Cluster 0 or 1 according to their needs. Similarly suggestions can be provided for almost any purpose as per demands and wishes of people.