Capstone Project: Entertainment Facility Recommendations in London

Remon Do

March 12, 2020

1 Introduction

1.1 Description of the problem

Entertainment place recommendation is a challenging task because of the variance in what an individual perceives as comfortable versus another. For example, children prefer play grounds or theme parks, young people like adventure and adult like some uproarious locations like bars or pubs. Additionally, the preference of entertainment varies in different time intervals or period. For example, people tend to visit natural lanscapes in summer rather than go to museums, or in winter much more custumers usually go to BBQ restaurants instead of the hot weather like summer. Entertainment preference also varies from culture to culture. For example, visiting religious locations like churches or temples would be prefered over stadiums in the weekend. A recommendation system would be highly effective if it can deal with as many of such challenges as possible.

1.2 Background

As we know, London is one of the most attractive locations in the world. Attracting 27 million visitors every year, London becomes the most visited city in Europe. It's no surprise that London is top of so many people's travel plans. Today, London is one of the most diverse cities in the world, with a rich history and some of the most outstanding culture on the planet. Of course, London has many entertainment facilities that being the most interests tourists. However, among a thousands of selections of which places that are convenient, high-rated, and affordable should be given priority. To this end, an efficient recommendation system should be considered to suggest and help tourists and travellers making a good decision. For aforementioned justifications, this project will investigate some common places that are highly recommended and visited in London. Then it can give reasonable suggestions for people to find the right place to enjoy their vacations or holidays.

2 Methodology

2.1 Data

Many of the world's largest tech companies rely on Foursquare data to add location into their apps and services. As the knowledge being acquired in this course, this project will take benefits from the FoursquareAPI application to access available data for the investigation. Foursquare lets users search for restaurants, nightlife spots, shops and other places of interest in their surrounding area. It is also possible to search other areas by entering the name of a remote location. The app displays personalized recommendations based on the time of day, displaying breakfast places in the morning, dinner places in the evening etc. Recommendations are personalized based on factors that include a user's check-in history, their "Tastes" and their venue ratings. For this reason, this project can get information of London such as Postal codes and theirs Borough, the geographical locations including Latitude and Longitude, the common places that are offen visited and recommended by custumers like Hotels, Bars, Restaurants. This approach is easy because of the open-source of data base location given by Wikipedia. In this case, data about locations of London can be easily fetched from this URL link: https://en.wikipedia.org/wiki/List_of_areas_of_London. After collecting necessary data, it should be preprocessed in the following steps.

2.2 Handling data

Data obtained will be wrangled in this step. First, BeautifulSoup library in Python is used to read the HTML content of data, then obtained data will be fed into dataframe to extract the most informative attributes. Some rich information such as Postal codes, Borough, and Latitude & Longitude should be retained because most of places are recommended based on these information. After that, matching physical locations with geographical locations will be done by applying FoursquareAPI. This step is performed by providing Client ID, Secret ID of the user and Version of FoursquareAPI application. Then, based on data analysis, the most common places will be recommended. Finally, a map of these locations should be displayed to facilitate the visualization for custumers. Details of these processed will be given in the following sections.

3 Data Acquisition and Processing

3.1 Data Acquisition

Data collected from London postal district database includes several important attributes such as:

- **Location:** Contains areas in London such as Abbey Wood, Acton, or Albany Park
- **London borough:** Contains 32 local authority districts that make up Greater London; each is governed by a London borough council.
- **Post town:** A post town is a required part of all postal addresses in the United Kingdom and Ireland, and a basic unit of the postal delivery system. Including the correct post town in the address increases the chance of a letter or parcel being delivered on time.

- **Postcode district:** Postcode district is made up of a postcode area and one or two numbers that can define and split the area into smaller geographical areas. On average each area is split 21 districts, but this varies from area to area. A postcode district can be 2-4 digits in total.
- **Dial code:** Indicates telephone number prefixes for reaching telephone subscribers in the networks of the member countries.
- **OS grid ref:** OS grid ref stands for Ordnance Survey National Grid reference system is a system of geographic grid references used in Great Britain, distinct from latitude and longitude.

The summary of these attributes on first 5 locations in London is shown in Figure 1.

	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

Figure 1: Data fetched from Wikipedia database.

3.2 Feature selection

Aiming at selecting the most informative and useful features (attributes) that have the strong effect to the recommendation system's performance, feature selection plays a pivotal role in improving the performance of the proposed system. As investigations from Wikipedia [1], and knowledge from the previous course. The most important attributes will be retained as the processes in Figure 2 and Figure 3.

[1]https://en.wikipedia.org/wiki/London_postal_district

	Location	Borough	Postcode	Post-town
0	Abbey Wood	Bexley, Greenwich	SE2	LONDON
1	Acton	Ealing, Hammersmith and Fulham	W3	LONDON
2	Acton	Ealing, Hammersmith and Fulham	W4	LONDON
3	Angel	Islington	EC1	LONDON
4	Angel	Islington	N1	LONDON

Figure 2: Most informative attributes retained.

	Location	Borough	Postcode
0	Abbey Wood	Bexley, Greenwich	SE2
1	Acton	Ealing, Hammersmith and Fulham	W3
2	Acton	Ealing, Hammersmith and Fulham	W4
3	Angel	Islington	EC1
4	Angel	Islington	N1

Figure 3: Final attributes being used.

3.3 Geographical Location Acquisition

After cleaning data and putting them into data frame, the next important step is to map the physical locations into geographical locations. Thanks to the advantage of the location acquisition application named FoursquareAPI, this problem can be easily done. Firstly, geocoder library that is available in Python packages can be used to the latitude and longitude locations of physical places in London (Figure 4). Then Client ID, Secret ID of the user and Version of FoursquareAPI application should be provided to access the Foursquare database by the getting function named ***getNearbyVenues()*** that is modelled in the Python source code. Finally, nearby venues in London can be displayed by calling function ***getNearbyVenues()*** above, results are shown in Figure 5 and Figure 6.

	Location	Borough	Postcode	Latitude	Longitude
0	Abbey Wood	Bexley, Greenwich	SE2	51.49245	0.12127
1	Acton	Ealing, Hammersmith and Fulham	W3	51.51324	-0.26746
2	Acton	Ealing, Hammersmith and Fulham	W4	51.48944	-0.26194
3	Angel	Islington	EC1	51.52361	-0.09877
4	Angel	Islington	N1	51.53792	-0.09983

Figure 4: Geographical locations in London.

```
In [14]: london_venues = getNearbyVenues(names=df_london['Borough'],
                                          latitudes=df_london['Latitude'],
                                          longitudes=df_london['Longitude']
                                         )

Bexley,  Greenwich
Ealing, Hammersmith and Fulham
Ealing, Hammersmith and Fulham
Islington
Islington
Brent
Barnet
Lambeth, Wandsworth
Islington
Barnet, Enfield
Barnet
Havering
Merton
Barnet
Barnet
Bexley
Bromley
Bromley
Croydon
Kingston upon Thames
Croydon
Westminster
Hillingdon
Hounslow
Havering
```


Figure 5: Display of all venues in London.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bexley, Greenwich	51.49245	0.12127	Sainsbury's	51.492824	0.120724	Supermarket
1	Bexley, Greenwich	51.49245	0.12127	Lesnes Abbey	51.489526	0.125839	Historic Site
2	Bexley, Greenwich	51.49245	0.12127	Lidl	51.496152	0.118417	Supermarket
3	Bexley, Greenwich	51.49245	0.12127	Abbey Wood Railway Station (ABW)	51.490825	0.123432	Train Station
4	Bexley, Greenwich	51.49245	0.12127	The Abbey Arms	51.490693	0.121182	Pub

Figure 6: Geographical location mapping.

One-hot encoding representation will be used to normalize locations *Venue Category*, then they are grouped by the average value by ***mean()*** function. This could result in the comprehensive table of common venues that are highly rated by the proposed recommendation system, as Figure 7. The final step is using a machine learning technique to cluster the obtain venues. In this project, k-means clustering with $k = 5$ will be used to perform the clustering task.

In [19]:
```
df_cluster = df_onehot.groupby('Neighborhood').mean().reset_index()
df_cluster.head()
```

Out[19]:

	Neighborhood	Accessories Store	Afghan Restaurant	African Restaurant	American Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Arts & Entertainment	...	Turkish Restaurant	Veg Res
0	Barking and Dagenham	0.0	0.0	0.009950	0.000000	0.0	0.029851	0.019900	0.000000	0.0	...	0.000000	0.0
1	Barnet	0.0	0.0	0.000000	0.010989	0.0	0.000000	0.000000	0.010989	0.0	...	0.032967	0.0
2	Barnet, Brent, Camden	0.0	0.0	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.0	...	0.000000	0.0
3	Barnet, Enfield	0.0	0.0	0.010204	0.000000	0.0	0.030612	0.020408	0.000000	0.0	...	0.000000	0.0
4	Bexley	0.0	0.0	0.009967	0.000000	0.0	0.029900	0.019934	0.000000	0.0	...	0.000000	0.0

5 rows × 194 columns

Figure 7: Average score representation of common venues in London.

4 Results

Figure 8 shows detailed information about every location and its common venues as well as the cluster label.

	Location	Borough	Postcode	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	7th Most Common Venue	8th I Com Ve
0	Abbey Wood	Bexley, Greenwich	SE2	51.49245	0.12127	2	Hotel	Monument / Landmark	Theater	Wine Bar	Plaza	Cocktail Bar	Garden	Pub
1	Acton	Ealing, Hammersmith and Fulham	W3	51.51324	-0.26746	2	Hotel	Monument / Landmark	Theater	Wine Bar	Plaza	Cocktail Bar	Garden	Pub
2	Acton	Ealing, Hammersmith and Fulham	W4	51.48944	-0.26194	2	Hotel	Monument / Landmark	Theater	Wine Bar	Plaza	Cocktail Bar	Garden	Pub
3	Angel	Islington	EC1	51.52361	-0.09877	2	Hotel	Monument / Landmark	Theater	Wine Bar	Plaza	Cocktail Bar	Garden	Pub
4	Angel	Islington	N1	51.53792	-0.09983	2	Hotel	Monument / Landmark	Theater	Wine Bar	Plaza	Cocktail Bar	Garden	Pub

Figure 8: Common venues by cluster labels.

Figure 9 visualizes the clustered neighborhoods in London on a map using the Folium library.

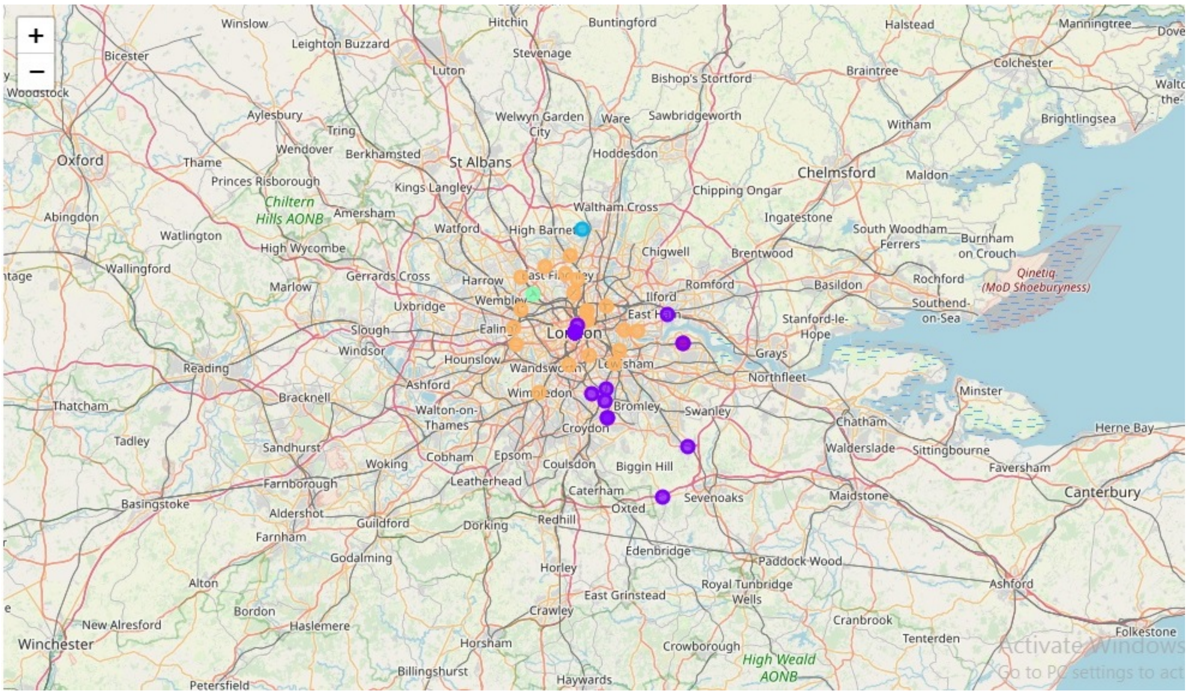

Figure 9: Folium map of common venues in London.

It can be easily seen that most of the location investigated are clustered in cluster 2 and cluster 3. Details about clustered locations will be given in Figure 10 - Figure 14.

	Location	Borough	Postcode	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	7th Most Common Venue	8th Most Common Venue
5	Church End	Brent	NW10	51.53916	-0.25123	0	Hotel	Monument / Landmark	Theater	Wine Bar	Plaza	Cocktail Bar	Garden	Pub

Figure 10: Common venues in cluster 0.

```
In [23]: london_join[london_join['Cluster Labels']==1]
```

Out[23]:

	Location	Borough	Postcode	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	7th Most Common Venue	8 C
9	Cockfosters	Barnet, Enfield	EN4	51.506420	-0.127210	1	Hotel	Monument / Landmark	Theater	Wine Bar	Plaza	Cocktail Bar	Garden	P
11	Collier Row	Havering	RM5	51.506420	-0.127210	1	Hotel	Monument / Landmark	Theater	Wine Bar	Plaza	Cocktail Bar	Garden	P
15	Colyers	Bexley	DA8	51.506420	-0.127210	1	Hotel	Monument / Landmark	Theater	Wine Bar	Plaza	Cocktail Bar	Garden	P
16	Aperfield	Bromley	TN16	51.271728	0.074711	1	Hotel	Monument / Landmark	Theater	Wine Bar	Plaza	Cocktail Bar	Garden	P
17	Coney Hall	Bromley	BR4	51.506420	-0.127210	1	Hotel	Monument / Landmark	Theater	Wine Bar	Plaza	Cocktail Bar	Garden	P

Figure 11: Common venues in cluster 1.

In [24]: `london_join[london_join['Cluster Labels']==2]`

Out[24]:

	Location	Borough	Postcode	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	7th Most Common Venue	8 C
0	Abbey Wood	Bexley, Greenwich	SE2	51.49245	0.12127	2	Hotel	Monument / Landmark	Theater	Wine Bar	Plaza	Cocktail Bar	Garden	P
1	Acton	Ealing, Hammersmith and Fulham	W3	51.51324	-0.26746	2	Hotel	Monument / Landmark	Theater	Wine Bar	Plaza	Cocktail Bar	Garden	P
2	Acton	Ealing, Hammersmith and Fulham	W4	51.48944	-0.26194	2	Hotel	Monument / Landmark	Theater	Wine Bar	Plaza	Cocktail Bar	Garden	P
3	Angel	Islington	EC1	51.52361	-0.09877	2	Hotel	Monument / Landmark	Theater	Wine Bar	Plaza	Cocktail Bar	Garden	P
4	Angel	Islington	N1	51.53792	-0.09983	2	Hotel	Monument / Landmark	Theater	Wine Bar	Plaza	Cocktail Bar	Garden	P

Figure 12: Common venues in cluster 2.

In [25]: `london_join[london_join['Cluster Labels']==3]`

Out[25]:

	Location	Borough	Postcode	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	7th Most Common Venue	8th M Comı Ve
29	Cricklewood	Barnet, Brent, Camden	NW2	51.56237	-0.22131	3	Hotel	Monument / Landmark	Theater	Wine Bar	Plaza	Cocktail Bar	Garden	Pub

Figure 13: Common venues in cluster 3.

In [26]: `london_join[london_join['Cluster Labels']==4]`

Out[26]:

	Location	Borough	Postcode	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	7th Most Common Venue	8th Mos Commo Venu
28	Crews Hill	Enfield	EN2	51.65366	-0.110614	4	Hotel	Monument / Landmark	Theater	Wine Bar	Plaza	Cocktail Bar	Garden	Pub

Figure 14: Common venues in cluster 4.

5 Discussion

Through results and analysis in the previous section, it can be seen that London offers one of the planet's greatest concentrations of cultural attractions. From accommodation palaces such as Hotel to the people's entertainment needs, from Theater and Plaza to Monument/ Landmark for breathtaking views. Moreover, tourists could spend endless days exploring London's night life facilities like Wine/Cocktail Bar or Art without ever running out of unique things to see and do.

6 Conclusions

The key idea was to use user data from Foursquare and recommend individuals to venues. The primary conclusion I reached to after trying several recommendation models was: The concept of geographic distance is as important as the user's taste in all venue categories. In simpler terms, this means that the likelihood of a person visiting a restaurant depends a lot on the geographic location of the restaurant and not just the users taste alone! So how could this conclusion help a data engineer come up with an answer to the venue recommendation problem? This project will help users to use data and data models to come up with an answer for queries like: Given a venue (e.g. an Indian Restaurant), select individuals who are likely to visit it. These kinds of queries come under the broad purview of "venue recommendation".