London city tours by venue themes

Danuphan Suwanwong
June 5, 2018

Introduction

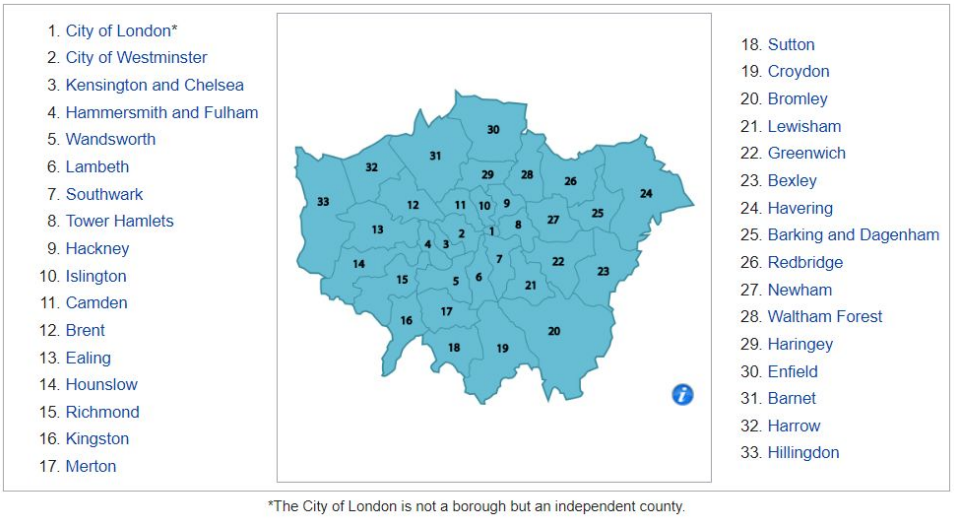
As a London tour agency, we are planning to promote new city tours in Greater London. We would like to know which boroughs are similar or different from the others by venues in each area. This will help us make a decision which tourist groups would be targeted in our tours e.g. "Food Lover" or "Shopping Lover".

On the other side, tourists can benefit from this project. They can look at results and plan for their trips by themselves whether they want to visit similar or different boroughs in Greater London.

Data Acquisition

London Boroughs

The data is from Wikipedia (https://en.wikipedia.org/wiki/Greater_London). There are 33 boroughs we will take into account. GPS locations of these boroughs will be retrieved from Foursquare API when querying venues since our tours will base on nearby venues in each borough, not a specific GPS location, where Foursquare API supports this search parameter.



*The City of London is not a borough but an independent county.

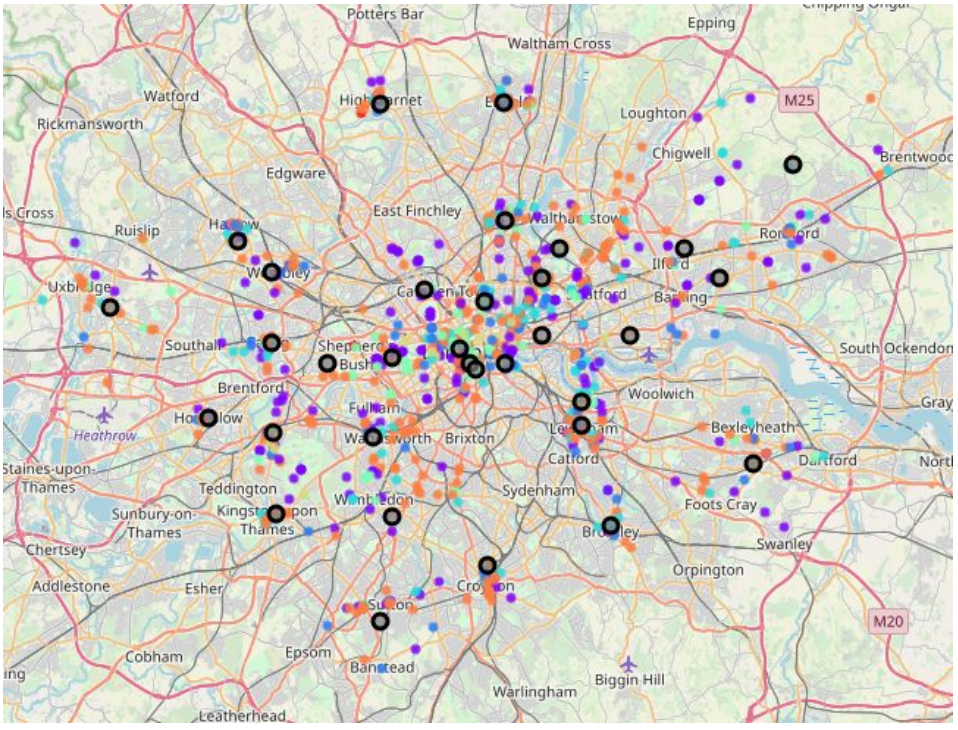
Venues

The venue list is from Foursquare API (https://developer.foursquare.com/) by the "Get Venue Recommendations" endpoint (https://developer.foursquare.com/docs/api/venues/explore). However, categories returned by API are too fine-grained and sparse so we hardly see the common pattern among boroughs. To handle this issue, we get a category tree from another Foursquare endpoint "Get Venue Categories" (https://developer.foursquare.com/docs/api/venues/categories) and then replace the original categories with their root categories. For example,

Original categories	Root categories
Bistro	Food
Dim Sum Restaurant	Food
Movie Theater	Arts & Entertainment

Exploratory Data Analysis

After we completed data acquisition, we then plot on a map using Folium python package. In the figure below, we can see that some parts have many “Food” venues (orange) grouping together and somes have “Outdoors & Recreation” venues (Purple).



We can make a hypothesis that some boroughs can be grouped together by their common venues e.g. “Food” Lover and “Outdoors & Recreation” Lover. We will continue on further analysis in the next section.

Clustering by K-Mean

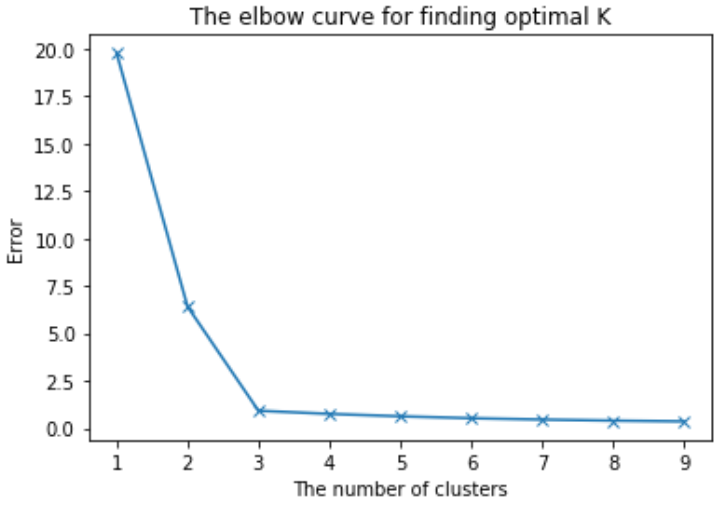
Our main goal is to group boroughs by their venues for a city tour. Obviously, we can use a clustering method i.e. K-Mean.
At the first step, we pre-processed our data using one-hot encoding to encode categorical data into numeric.

	Borough	Arts & Entertainment	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Shop & Service	Travel & Transport
15	City of London	0	0	0	1	0	0	0
538	Sutton	0	1	0	0	0	0	0
84	Kensington and Chelsea	0	0	0	1	0	0	0
209	Southwark	0	1	0	0	0	0	0
692	Havering	0	0	0	1	0	0	0

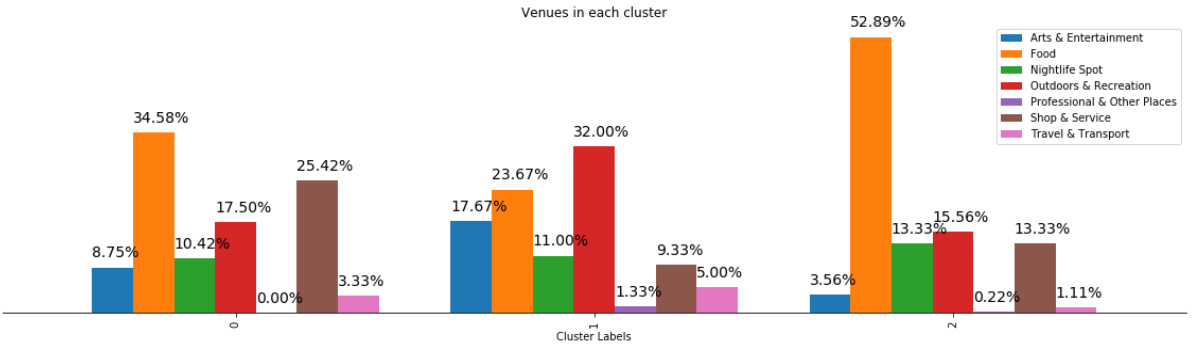
We then grouped them by boroughs to see the venue distribution on average in each area.

	Borough	Arts & Entertainment	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Shop & Service	Travel & Transport
22	Lambeth	0.200000	0.333333	0.133333	0.233333	0.033333	0.000000	0.066667
6	City of London	0.033333	0.066667	0.100000	0.633333	0.000000	0.133333	0.033333
5	Camden	0.200000	0.200000	0.100000	0.333333	0.000000	0.100000	0.066667
18	Hounslow	0.000000	0.466667	0.033333	0.066667	0.000000	0.366667	0.066667
15	Harrow	0.000000	0.533333	0.200000	0.100000	0.000000	0.166667	0.000000

Now we applied the K-Mean method. The optimal number of clusters is 3 which can be found by Elbow curve.



We applied K-Mean with 3 clusters again and see how boroughs were clustered.



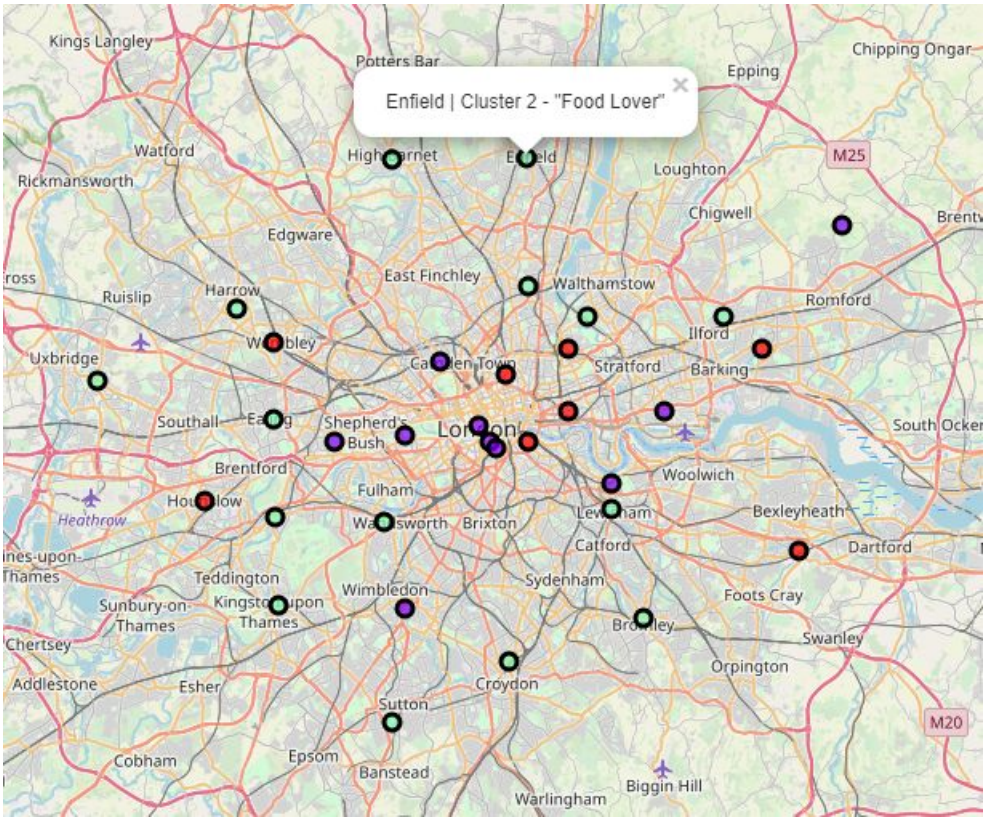
As the figure above, we can clearly assign themes to each cluster:

- Cluster 0: “Food and Shopping Lover”
- Cluster 1: “Outdoors and Entertainments Lover”
- Cluster 2: “Food Lover”

Results

After we assigned cluster labels to each borough, we then plotted on the map again with assigned cluster labels.

Clusters	Color	Boroughs
0: “Food and Shopping Lover”	Red dots	'Southwark', 'Tower Hamlets', 'Hackney', 'Islington', 'Brent', 'Hounslow', 'Bexley', 'Barking and Dagenham'
1: “Outdoors and Entertainments Lover”	Blue dots	'City of London', 'City of Westminster', 'Kensington and Chelsea', 'Hammersmith and Fulham', 'Lambeth', 'Camden', 'Merton', 'Greenwich', 'Havering', 'Newham'
2: “Food Lover”	Green dots	'Wandsworth', 'Ealing', 'Richmond', 'Kingston', 'Sutton', 'Croydon', 'Bromley', 'Lewisham', 'Redbridge', 'Waltham Forest', 'Haringey', 'Enfield', 'Barnet', 'Harrow', 'Hillingdon'



Finally, the tour agency can use this information to promote city tours by themes according to clusters we had found so far.

Discussion

In this project, we applied K-Mean clustering to venues in each borough and used Elbow curve to find the optimal number of clusters. The boroughs were well-clustered and meaningful for each cluster.

Not surprisingly, the central of London has a good balance of venues comparing to the outer boroughs which mainly have a popularity of food. So we would not recommend tourists to go to boroughs assigned to cluster 2 due to the tourists surely not coming here for only food.

Regarding data acquisition, we can see that venues listed by borough names from Foursquare API sometimes go beyond borough boundary. Furthermore, in our experiment, we queried only the top 30 venues recommended by Foursquare which could be biased to the central area where there could be more venues to be explored in the area. If the Foursquare API could precisely give venues within a given borough, it would improve our cluster accuracy since we can precisely assign venues to their boroughs.

In future studies, we could apply similar ideas to other cities, cluster them and find different tour themes in different cities, or even try clustering many cities together to find similarities (or dissimilarities) at a higher viewpoint.

Conclusion

In this study, we applied the K-Mean clustering algorithm to find venue themes, according to recommended venues by Foursquare, in London boroughs to help the tour agency target groups of tourists. The result supports our hypothesis at the beginning and recommended to not doing group tours in outer boroughs. In summary, the tour agency should target “Food and Shopping Lover” and “Outdoors and Entertainments Lover” customers.