London Clusters – Report

Contents

Introduction - Business Case

Companies operating in the tourism, travel, transport sector are interested in suggesting itineraries and destinations tailored to their audience's interests.

When visiting a city, for example, some travellers may be more interested in visiting museums and art places. Some others may be more interested in a shopping kind of tourism. Some others may be interested in gastronomy, and so on. It is also realistic to think that tourists, when visiting a city, would want to do a mix of things, for example may want to visit art places in the morning, do some shopping in the afternoon, and go out in the evening to some cool night spots area.

Tourism agencies may therefore want to make use of machine learning, and specifically clustering, to provide customers with relevant suggestions as to what areas to go to.

Data

As a test case, I will try to leverage Foursquare's data to cluster neighbourhoods in London, United Kingdom. The idea here is to be able to classify each neighbourhood at an overall level in terms of the main attractions it has to offer, and possibly categorise each neighbourhood in main classes, e.g. predominantly 'shopping area', or 'nightlife area', and so on. (Obviously, some areas have more to offer than only one type of attraction, e.g. there are usually many restaurants around art places. Nevertheless, the clustering should be able to pick on these elements and return clusters which are 'mixed'. Tourists only interested in food, for example, may still be interested in visiting a 'mixed' area which scores very high on 'food'.) The analysis will be run by employing the use of the k-means algorithm.

In order to create the cluster, I will build the underlying dataset so that:

• Each venue category is recoded to its own macro category as per Foursquare category tree (see https://developer.foursquare.com/docs/resources/categories). For example, an Italian restaurant and a Chinese restaurant will be both recoded to their own Macro category 'Food'. This will allow for the k-means algorithm to work with 9 aggregated macro variable (Food, Travel & Transport, Shop & Service, Arts & Entertainment, Nightlife Spot, Professional & Other Places, Outdoors & Recreation, College & University, Residence) rather than a hundreds of variables (i.e. the venue specific category)

• Each venue will have its own 'weight' in terms of 'likes'. For each venue, I will retrieve the count of 'likes' (i.e. how many people liked the particular venue). This will allow distinguishing between a major and a minor venue (e.g. a relatively unknown music venue 'The Blue Studios' has 6 likes, whereas an important music venue 'The O2 Arena' has 3154 likes) rather treating all venues as equal.

As for the geographical units to take into consideration, I have opted for the rail stations (overground and underground) in London zones 1, 2 and 3 (the central zones) as they are evenly spread across the city, whereas the actual administrative units (London boroughs) vary quite dramatically in terms of size amongst themselves.

Methodology¶

In this section I will carry out the data processing and analysis.

As briefly explained above I will look at the target geographical areas (underground and overground stations in London zones 1, 2, and 3) and extract their 'essence' as a whole by

grouping together the venue types under few macro categories, and weight the dataset by the number of venue 'likes'.

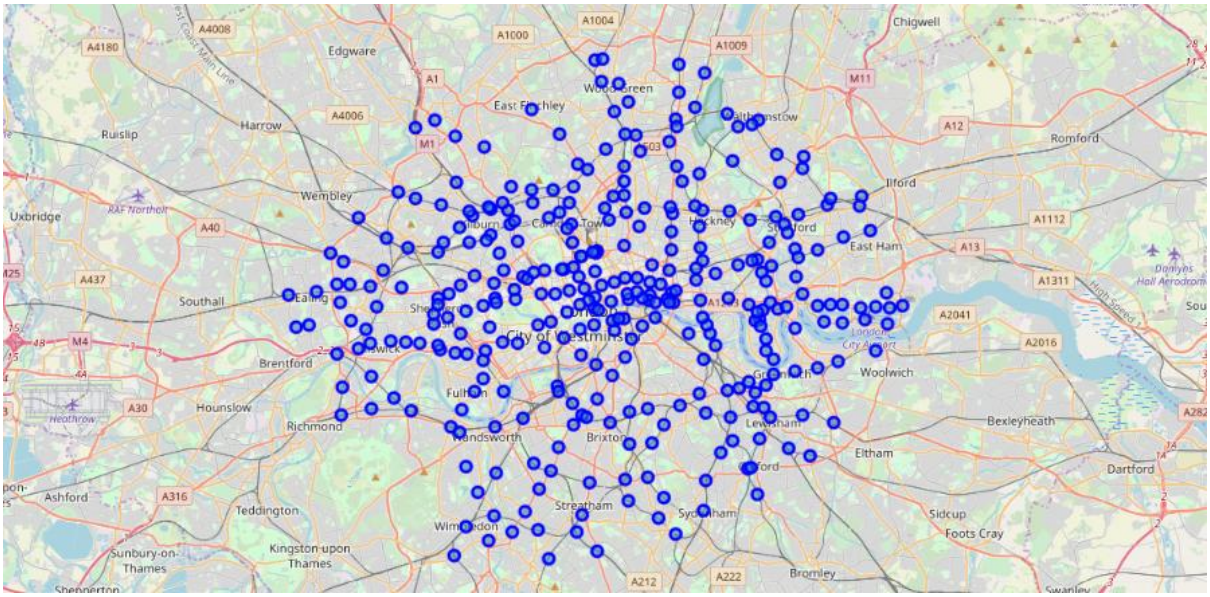
I will then explore the resulting dataset by looking at the correlation between macro categories and get some extra insight by creating a dendrogram (although I will not be running a hierarchical clustering).

I will finally proceed with the clustering of the London stations via the k-means algorithm by applying the optimal number of cluster as suggested by the analysis of the sum of square distances ('elbow method').

Stations data

1.1 retrieve the stations list and their coordinates using beautiful soup (source: https://www.doogal.co.uk/london_stations.php). Filtered the results to obtain 340 stations in zones 1, 2 and 3

1.2 chart the stations :



Foursquare data

2.1 Build a Foursquare category look-up table. With this table I will be able to say what micro-category translate into what macro category (e.g. a pizza place and a burrito restaurant will both fall in the 'Food' macro category). I now have my look-up table: for each venue category (column 1) I know its macro category. E.g. venue category id

56aa371be4b08b9a8d5734db ('Amphitheater') is now associated with its macro category id 4d4b7104d754a06370d81259 ('Arts & Entertainment'). Visit https://developer.foursquare.com/docs/resources/categories for a better understanding.

	0	1	2
0	4d4b7104d754a06370d81259	56aa371be4b08b9a8d5734db	Arts & Entertainment
1	4d4b7104d754a06370d81259	4fceea171983d5d06c3e9823	Arts & Entertainment
2	4d4b7104d754a06370d81259	4bf58dd8d48988d1e1931735	Arts & Entertainment
3	4d4b7104d754a06370d81259	4bf58dd8d48988d1e2931735	Arts & Entertainment
4	4d4b7104d754a06370d81259	4bf58dd8d48988d1e4931735	Arts & Entertainment

2.2 for each station, gather the venues in a radius of 1000 meters. In this loop, alongside with collecting the venue info (e.g. name, category,...), I will also calculate the distance from the venue and the station, together with retrieving the venue macro category via the look-up table above.

2.3 for each venue, retrieve the number of 'likes'. In here I will loop through all the unique venues collected (10,000+) and collect their number of 'likes' given by Foursquare users. Here below the resulting dataframe head:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue Name	Venue ID	Venue Latitude	Venue Longitude	Venue Distance from Neigh	Venue Category	Venue Category ID	Macro Category Name	Likes
0	Gloucester Road	51.494500	-0.183529	Byron	4a5f9446f964a520e0bf1fe3	51.494793	-0.182468	0.001101	Burger Joint	4bf58dd8d48988d16c941735	Food	193
1	Piccadilly Circus	51.509697	-0.133697	Haymarket Hotel	4abcec53f964a520b98720e3	51.508267	-0.131315	0.002779	Hotel	4bf58dd8d48988d1fa931735	Travel & Transport	47
2	Tottenham Court Road	51.516211	-0.131096	Patara	4abe5714f964a520d68c20e3	51.514136	-0.130855	0.002089	Thai Restaurant	4bf58dd8d48988d149941735	Food	151
3	Knightsbridge	51.501355	-0.160650	Harrods	4abf8c03f964a520079120e3	51.499572	-0.162698	0.002715	Department Store	4bf58dd8d48988d1f6941735	Shop & Service	10475
4	Hyde Park Corner	51.502584	-0.152460	The Athenaeum Hotel	4ac51183f964a52045a020e3	51.504589	-0.147353	0.005486	Hotel	4bf58dd8d48988d1fa931735	Travel & Transport	115

2.4 build a dataframe in which each venue row is repeated a 'Likes' number of times (e.g. if venue X has 3 likes, venue X will be repeated 3 times). This dataframe is then passed to the one-hot encoding method. The repetitions are to account for the number of likes for the clustering.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue Name	Venue ID	Venue Latitude	Venue Longitude	Venue Distance from Neigh	Venue Category	Venue Category ID	Macro Category Name	Likes
0	Gloucester Road	51.4945	-0.183529	Byron	4a5f9446f964a520e0bf1fe3	51.494793	-0.182468	0.001101	Burger Joint	4bf58dd8d48988d16c941735	Food	193
1	Gloucester Road	51.4945	-0.183529	Byron	4a5f9446f964a520e0bf1fe3	51.494793	-0.182468	0.001101	Burger Joint	4bf58dd8d48988d16c941735	Food	193
2	Gloucester Road	51.4945	-0.183529	Byron	4a5f9446f964a520e0bf1fe3	51.494793	-0.182468	0.001101	Burger Joint	4bf58dd8d48988d16c941735	Food	193
3	Gloucester Road	51.4945	-0.183529	Byron	4a5f9446f964a520e0bf1fe3	51.494793	-0.182468	0.001101	Burger Joint	4bf58dd8d48988d16c941735	Food	193
4	Gloucester Road	51.4945	-0.183529	Byron	4a5f9446f964a520e0bf1fe3	51.494793	-0.182468	0.001101	Burger Joint	4bf58dd8d48988d16c941735	Food	193

	Neighborhood	Arts & Entertainment	College & University	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	Gloucester Road	0	0	1	0	0	0	0	0	0
1	Gloucester Road	0	0	1	0	0	0	0	0	0
2	Gloucester Road	0	0	1	0	0	0	0	0	0
3	Gloucester Road	0	0	1	0	0	0	0	0	0
4	Gloucester Road	0	0	1	0	0	0	0	0	0

Preliminary analysis

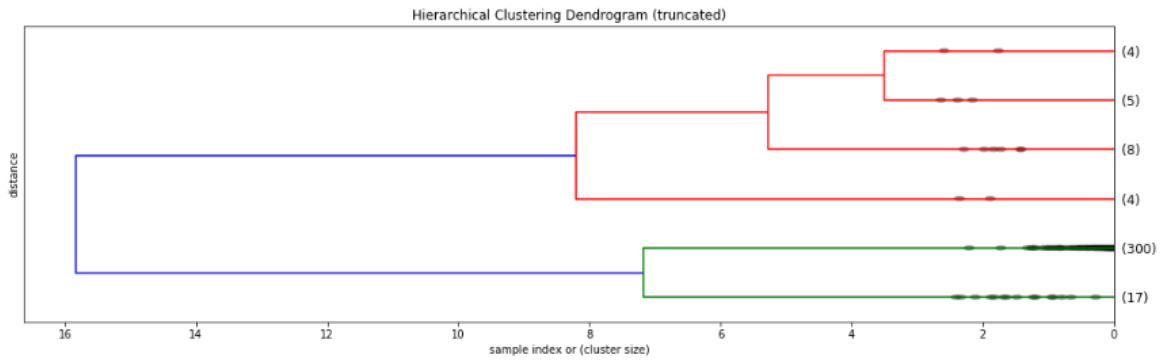
3.1 overall low correlation between variables. I could interpret this as every variable pulling in its own direction (which is good news for the purpose of clustering). However, also note that Food correlates mildly strongly with Nightlife and Shop & Services. This makes sense intuitively.

	Arts & Entertainment	College & University	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
Arts & Entertainment	1	-0.00235511	0.375284	0.237124	0.18775	0.0821877	-0.0147167	0.126498	0.103806
College & University	-0.00235511	1	0.0326766	0.0415591	-0.0128941	-0.00125825	-0.00068414	-0.00870741	0.012155
Food	0.375284	0.0326766	1	0.63891	0.380545	0.0740216	0.046342	0.735733	0.487547
Nightlife Spot	0.237124	0.0415591	0.63891	1	0.29441	0.110444	0.0324091	0.385691	0.298087
Outdoors & Recreation	0.18775	-0.0128941	0.380545	0.29441	1	0.094505	-0.0112021	0.285808	0.235801
Professional & Other Places	0.0821877	-0.00125825	0.0740216	0.110444	0.094505	1	-0.00581219	0.0375124	0.0958777
Residence	-0.0147167	-0.00068414	0.046342	0.0324091	-0.0112021	-0.00581219	1	-0.00196117	0.123419
Shop & Service	0.126498	-0.00870741	0.735733	0.385691	0.285808	0.0375124	-0.00196117	1	0.29212
Travel & Transport	0.103806	0.012155	0.487547	0.298087	0.235801	0.0958777	0.123419	0.29212	1

3.2 looking at venue groups via dendrograms: the chart below highlights the difference in 2 main groups of stations / neighbourhoods (these would be my clusters in hierarchical clustering). Looking at the labels we can see that in the red group there are the most vibrant areas in central London (i.e. where the world famous names are: Oxford Circus, Westminster, Leicester Square, etc.).

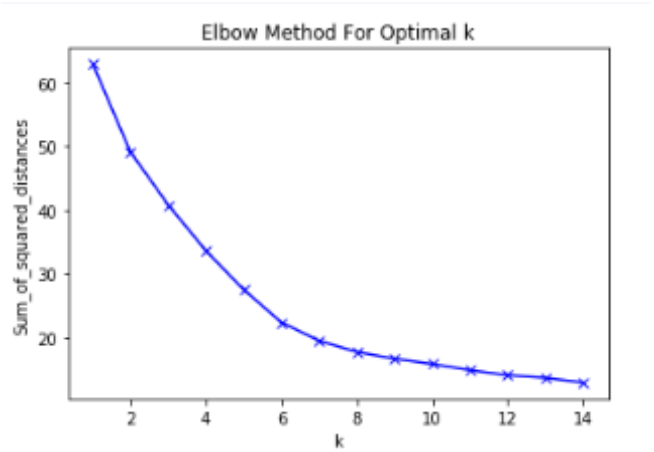
The distance from the latter group to the rest is quite considerable. This is because the number of 'Likes' is much higher for the major venues (e.g. Big Ben, Harrods, Tower of London, to mention few) compared to other venues. In other terms, the data structure is very much stretched by the number of 'Likes'.

Finally, the dendrograms also show in the green group a middle band of many stations with very little difference (close to 0) amongst them. These are stations associated with very few and / or not very much 'liked' venues. These are mainly stations in the suburbs (zones 2 and 3). As a result of the few likes these stations appear as an indistinct group (in the below dendrogram, value 300).



Clustering with k-means:

4.1 look for the optimal k (via elbow method). The elbow method suggests the optimal number of clusters I should be using is 6. I will apply this to the k-mean clustering.



4.2 clustering with k-means algorithm and merging the stations and venues datasets. Final dataframe grouped by mean – i.e. the k-mean centroids.

	Neighborhood	Zone	Postcode	Latitude	Longitude	Easting	Northing	Cluster Labels	Arts & Entertainment	College & University	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	Abbey Road	3.0	E15 3NB	51.531952	0.003738	539081	183352	2	0.166667	0.0	0.000000	0.000000	0.777778	0.0	0.0	0.000000	0.055556
2	Acton Central	2.0	W3 6BH	51.508758	-0.263416	520613	180299	2	0.024561	0.0	0.115789	0.329825	0.449123	0.0	0.0	0.042105	0.038596
3	Acton Main Line	3.0	W3 9EH	51.516887	-0.267676	520296	181196	2	0.000000	0.0	0.000000	0.000000	0.428571	0.0	0.0	0.285714	0.285714
4	Acton Town	3.0	W3 8HN	51.503071	-0.280288	519457	179639	5	0.282828	0.0	0.606061	0.000000	0.000000	0.0	0.0	0.000000	0.111111
8	Aldgate	1.0	EC3N 1AH	51.514342	-0.075613	533629	181246	5	0.000000	0.0	0.748711	0.167526	0.015034	0.0	0.0	0.003866	0.064863

Cluster Labels	Arts & Entertainment	College & University	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	0.0536797	0.000159012	0.24046	0.496934	0.0639649	0.0116497	0	0.0770291	0.0561241
1	0.0566454	0	0.244842	0.070776	0.0589283	0.0141956	0	0.497122	0.0574907
2	0.024453	0	0.151738	0.0724089	0.568078	0.00551639	0	0.0986991	0.0791072
3	0.028984	0	0.217085	0.0994214	0.0708042	0	0	0.0268797	0.556825
4	0.435881	7.36864e-06	0.261052	0.121039	0.0622205	0.0264479	0	0.0515838	0.0417674
5	0.0366682	6.93962e-05	0.58769	0.17004	0.0639494	0.00562778	7.67016e-05	0.087067	0.0488116

4.3 cluster interpretation. In the table above we can see the 6 clusters created and their strengths. Matching macro category labels and clusters is now quite intuitive:

0 'Fun' cluster (predominance of Nightlife and trending with College & University)

1 'Shopping' cluster (predominance of Shopping)

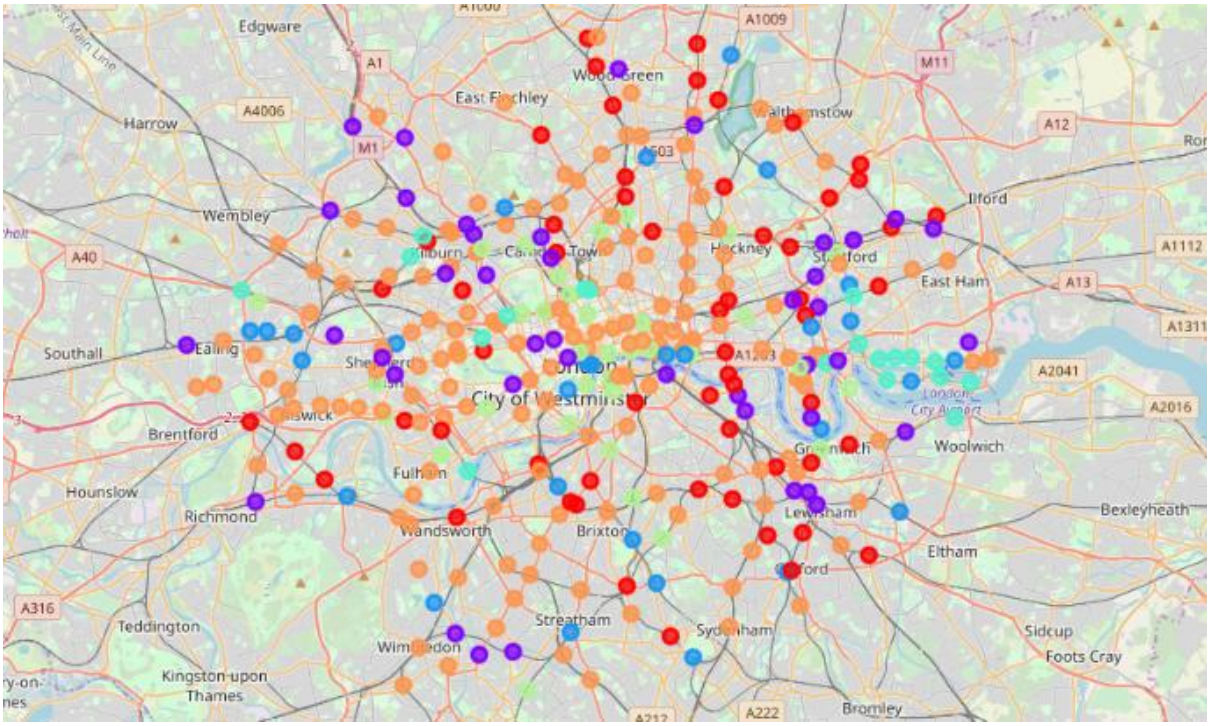
2 'Parks' cluster (predominance of Outdoors)

3 'Accommodation' cluster (note that under 'travel and transport' foursquare lists all hotels, hostels, and so on)

4 'Major attractions' cluster (predominance of Arts & Entertainment - museums, theatres, monuments, etc. are here)

5 'Food' cluster (predominance of Food - however this cluster is quite indistinct, 'Food' is common to many areas in any cities - so probably areas falling in this cluster could be interpreted as lacking any other particularly distinctive characteristic (e.g. cluster 5 could be called 'Other')

4.4 chart of clusters:



Results

The clustering of London station using a 'weighted' dataset (i.e. were venues had 'Likes' as weights) has added depth to the data analysis. The resulting 6 clusters do make intuitive sense, e.g. areas such as Oxford Circus, Piccadilly Circus, Knightsbridge, which are famous for shopping, fall in the 'shopping' cluster. Similarly, Westminster - Big Ben - South Kensington - where museums are concentrated - and so on fall in the cluster 'major attractions'. In a nutshell, the k-mean clustering seem to have captured the essence of an area.

Discussion

There are a few caveats to model presented in this report. Firstly, cluster 5 is far too big compared to the size (count) of the other cluster - this suggesting that such areas should be treated separately from the other clusters and further investigated. Secondly, the use of weights has had a skewing effect on the dataset (e.g. some venues had thousands of likes, many venues just a few). In third place, Foursquare users' specific demographics (slight skew towards young, males, and US as a country of origin - see: https://99firms.com/blog/foursquare-statistics/) may have had an effect in favouring certain venues / areas and therefore influencing the clusters.

Conclusion

The use of Foursquare data in conjunction with the k-mean algorithm, with the methodology as detailed in this report (use of macro categories and weighting data by number of likes) portrait quite a faithful picture of London areas (stations). The model can be improved with a less skewing use of weights and possibly by integrating other data sources aside Foursquare (e.g. Google Places to account for venues not covered by Foursquare users).

Company operating in the Transport, Travel and Tourism industry at large could use clustering to associate areas to their costumers' interest. This particular report focused on London as a test case – the model can be surely be replicated for other cities.