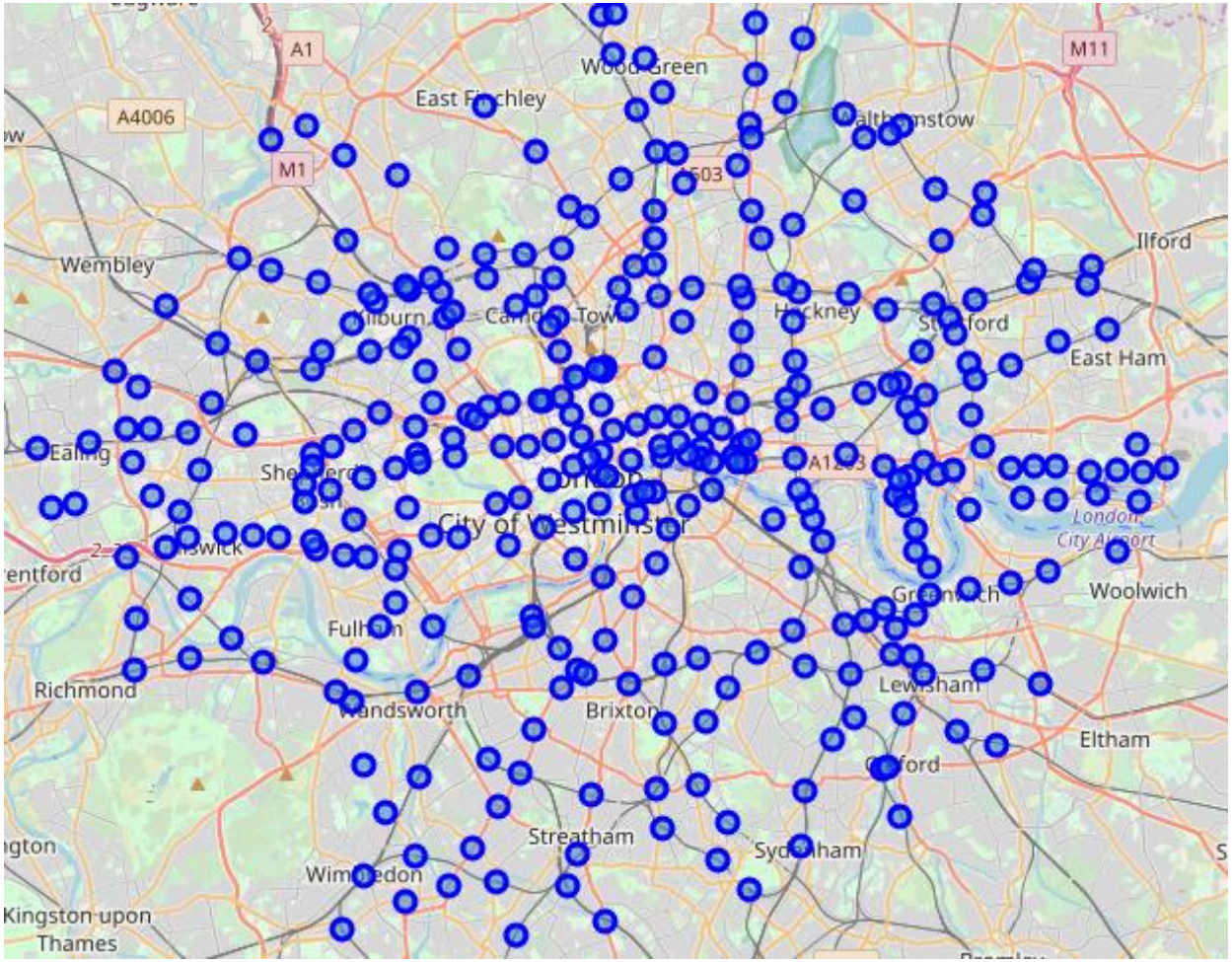
London Clusters

WALTER D'ANTINO

Business Case

Oppor-tunity	Transport, travel and tourism companies are interested in recommending destinations to their users tailored on their preferences.
Idea	Using **k-means** to cluster geographical areas (destinations) by their amenities could help such companies deliver relevant recommendations.
Business Case	Applied business case to London, UK, where **destinations** are underground and over-ground stations and **amenities** are Foursquare venues

Destinations - Stations



About 340 stations selected for this business case.

Stations cover zones 1, 2, and 3 in London UK.

Stations in London are evenly distributed around the town.

Collecting the Foursquare venues information around each station provides a good representation of what goes on around the city

Foursquare data preparation

The Foursquare venues information is worked so to have a final feature set (to be used in the k-mean algorithm) which uses 1) **Macro categories** and is 2) weighted by the number of **Likes** of the venues.

Macro Category	Each venue id is recoded to its macro venue id (e.g. an Aquarium is goes with its Macro Category 'Arts'). This allows k-means to work with fewer, more meaningful variables.
Likes	Each venue will be repeated as many times as its number of Likes. This is give a relative weight / importance to each venue.

Arts & Entertainment
4d4b7104d754a06370d81259

Amphitheater
56aa371be4b08b9a8d5734db

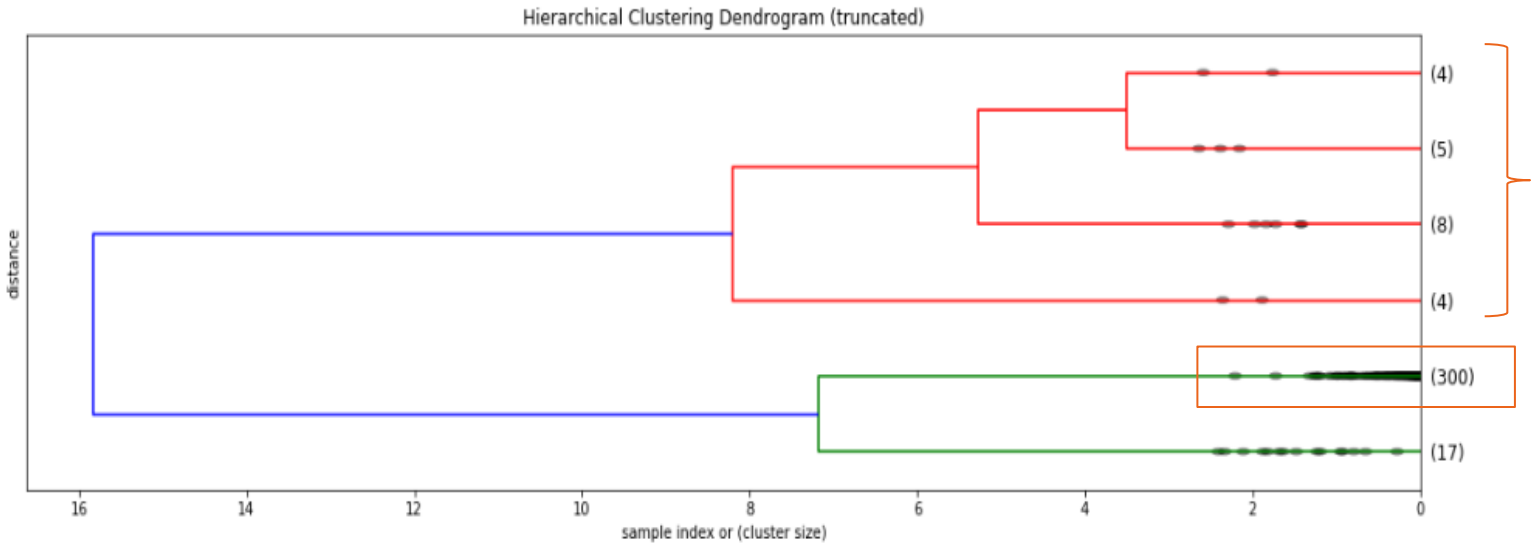
Aquarium
4fceea171983d5d06c3e9823

Feature set – correlation analysis

	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	-0.00668414	-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	-0.00668414	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

The correlation matrix shows an overall low correlation between the Macro Category variables –i.e. 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. 'Food' is common to many areas and may not be a good area distinguisher.

Feature set – data hierarchy

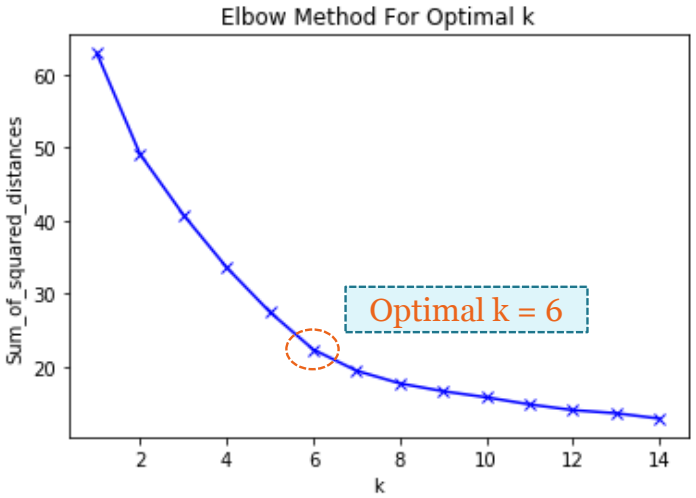


Central London Stations – major point of interest

London Suburbs Stations – low differentiation

The dendrogram above highlights the difference in 2 main groups of stations.
The red group contains most vibrant areas in central London (e.g. Oxford Circus, Westminster, Leicester Square, etc.). The distance from the 'red group' from 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'.
The green group contains 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.

Clustering with k-means

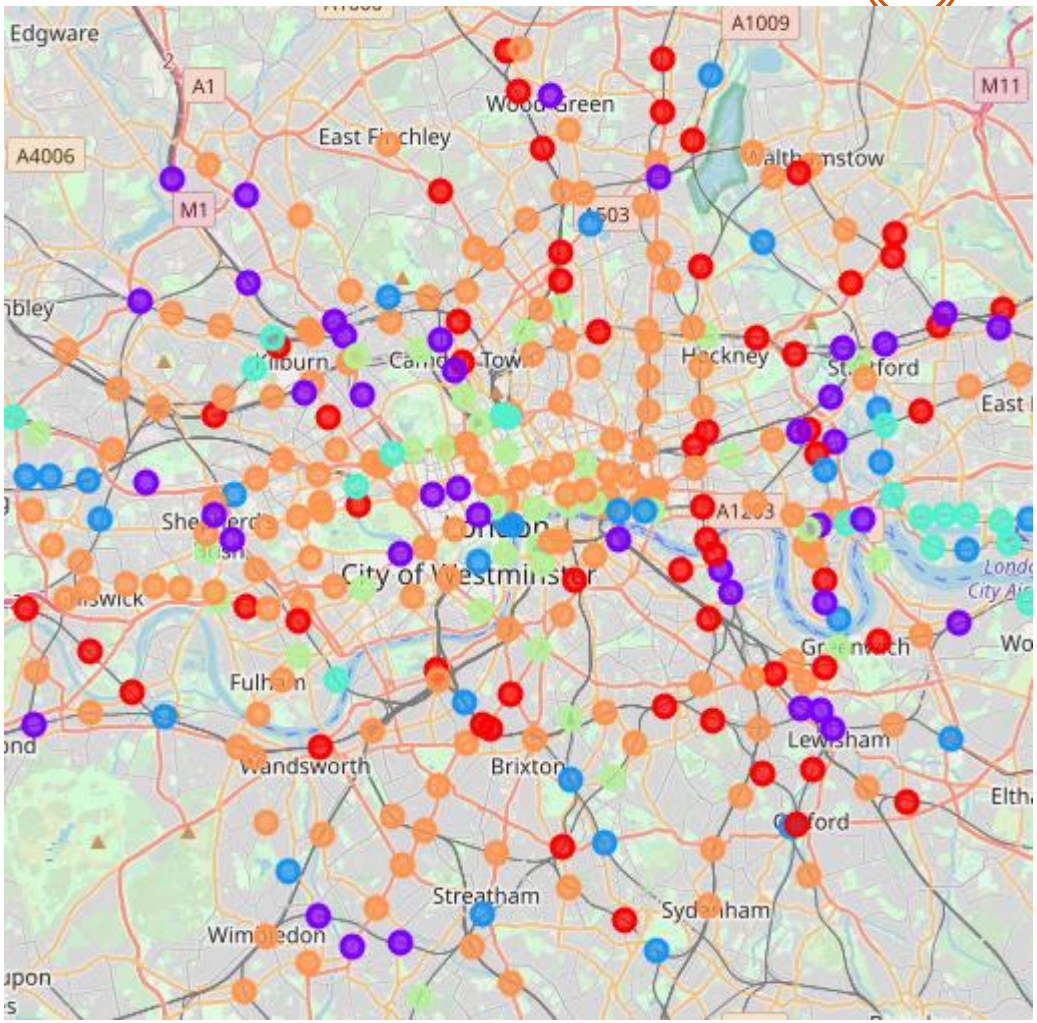


By applying the 'elbow method' in finding the optimal number of k, the following clusters emerge:

0 - **Fun** (Uni + Nightlife + Food)
1 - **Shopping** (Shops + Food)
2 - **Outdoor** (Parks)
3 - **Accommodation** (Hotels + major transport hubs)
4 - **Major attractions** (Museums, theatres, monuments)
5 - **Food** (Food and 'Other')

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

Conclusion



Model Pros	A feature set based on macro categories and weights allows to capture more closely the essence of an area.
Model Cons	Cluster 5 ‘Food (+ Other)’ is indistinct and accounts for a disproportionate amount of stations. This cluster needs further refining.
Conclusion	This model (k-means clustering + specially configured feature set) could potentially be used by tourism company as a base for a destination recommendation system.