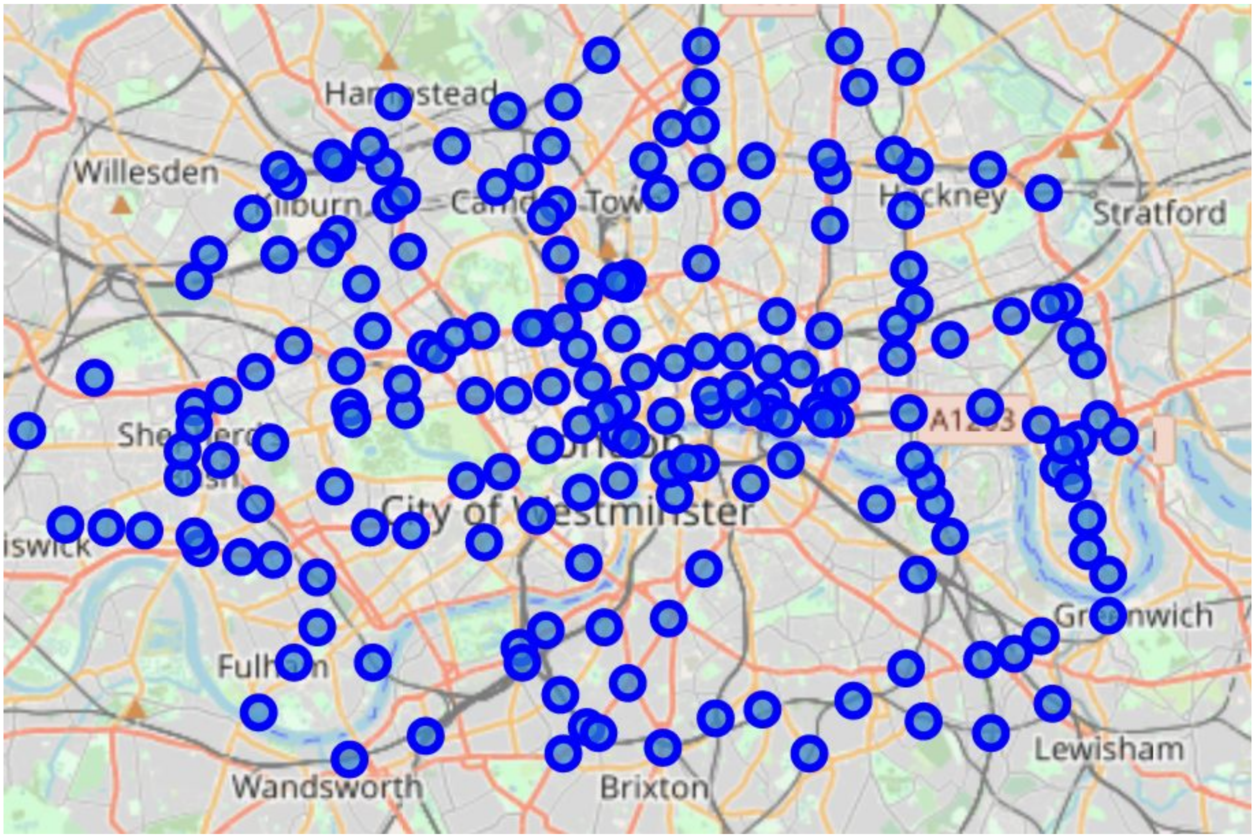
London Station Clustering

1. Introduction

1.1 Background

London is a big and diverse city, with 194 different train and tube stations serving just zones one to three:



We would like to understand what each of these different stations is like compared to the other stations, so you could know what types of businesses and attractions are likely to be around before getting off there.

1.2 Problem

Unsupervised clustering and classification is difficult to get right, the problem is that it doesn’t know which problem to solve, so we have to be very careful with the training data to ensure that it is classifying based on things we care about.

1.3 Interest

Anyone who wants to understand the different train stations of London better, for example a travel guide could use this to describe different tube stops characteristics. Another use could be if you owned a business and wanted to expand to another area, and wanted to choose somewhere similar to where you already are.

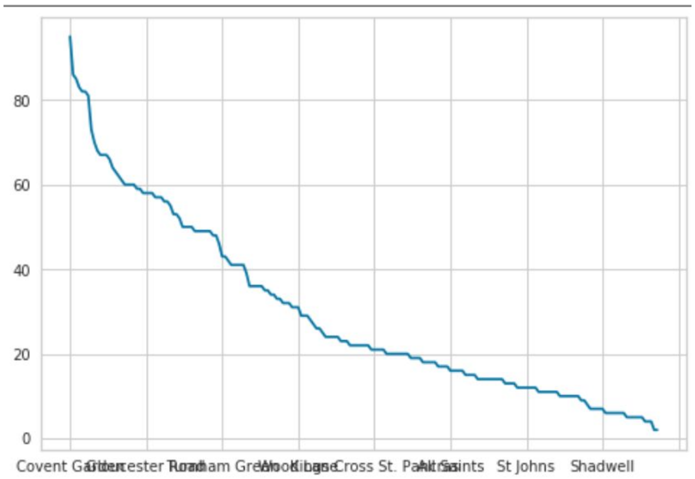
2. Data

2.1 Data Sources

Doogal - 'https://www.doogal.co.uk/london_stations.php' will be used for a table of London stations and their coordinates.

Foursquare - https://developer.foursquare.com/docs/api-reference/venues/explore/. API will be used to get venues around the different stations, along with the categories API for help converting venue types. ttps://developer.foursquare.com/docs/api-reference/venues/categories/

2.2. Data cleaning



Data from Doogal was cross referenced against google maps and showed the same data, so was good to use as was once filtered down to just zone 3.
When collecting venues from foursquare, care was taken to reduce the radius to ensure that less than 100 venues were being returned per search which is the foursquare limit:

3. Methodology

3.1 Feature Selection

We mapped all of the 338 venue categories returned by FourSquare to the below label types, this was done using foursquares categories to map these back up to a higher level, as well as with a human overlay where the foursquare category tree did not give an appropriate aggregation:

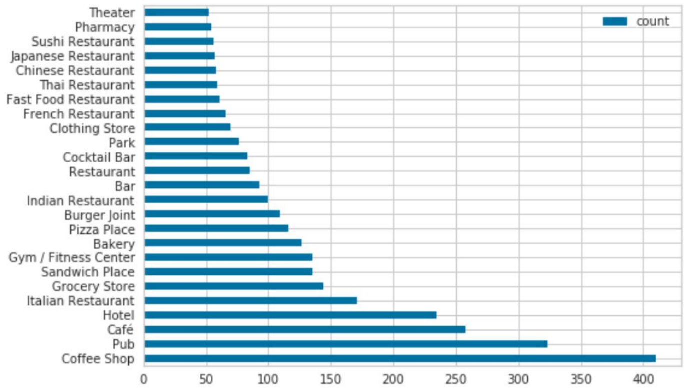
Accomodation - hotels, hostels, bed and breakfast etc.
Arts - Art Galleries, Exhibitions, etc.
Bar - Anywhere that calls itself a bar
Pub - Traditional English drinking establishment
Coffee Shop/Café
Entertainment - Cinemas, etc.

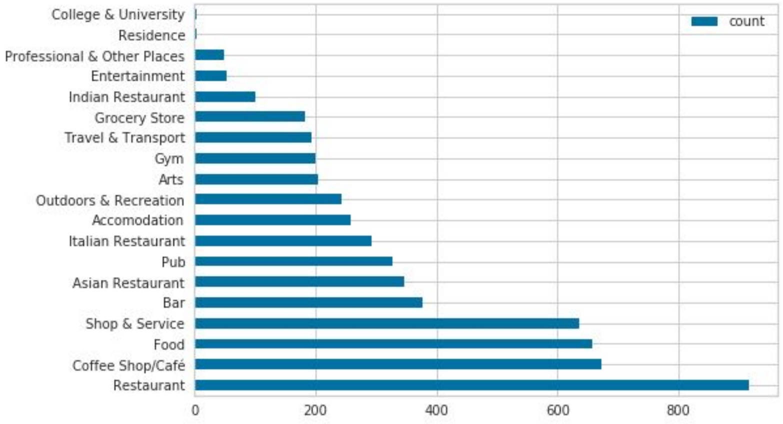
Gym - Includes studios and speciality gyms
Outdoors & Recreation - Parks etc.

Food - food to take out
Indian Restaurant - Includes continent of India
Italian Restaurant - Includes pizza places
Asian Restaurant - Include full subset of Asian food
Restaurant - Catch all for everything not covered above, there is opportunity to split further

Shop & Service - any shops and services not a supermarket or grocery store.
Grocery Store - includes supermarkets
Travel & Transport - anything related to travel and transport

for this list, the frequency of the top 100 can be seen below:





We used the default category returned by Foursquare, of which there are 336 unique categories, the top 100 of which can be seen in the table on the left to the categories on the right, College & University and Residence were removed due to low incidence.

3.2 Feature Normalisation

Multiple methods for feature weighting were considering for this project

1. We could normalise so that all 17 categories above have the same weight in the clustering. This was not chosen as these categories are pretty arbitrary, and would give too much prominence to otherwise lowly weighted categories.
2. We could normalise across a location, so that each horizontal equals 1 split between the categories. As this would remove the number of venues from mattering, it would lose a large magnitude of the detail
3. Mapping a weighting onto the different venues depending on how important the venue is, this would have been a preferred method, however requires premium calls to the FourSquare API to gain insights into popularity and size of venue. In this way a museum would not be worth the same as a coffee shop.

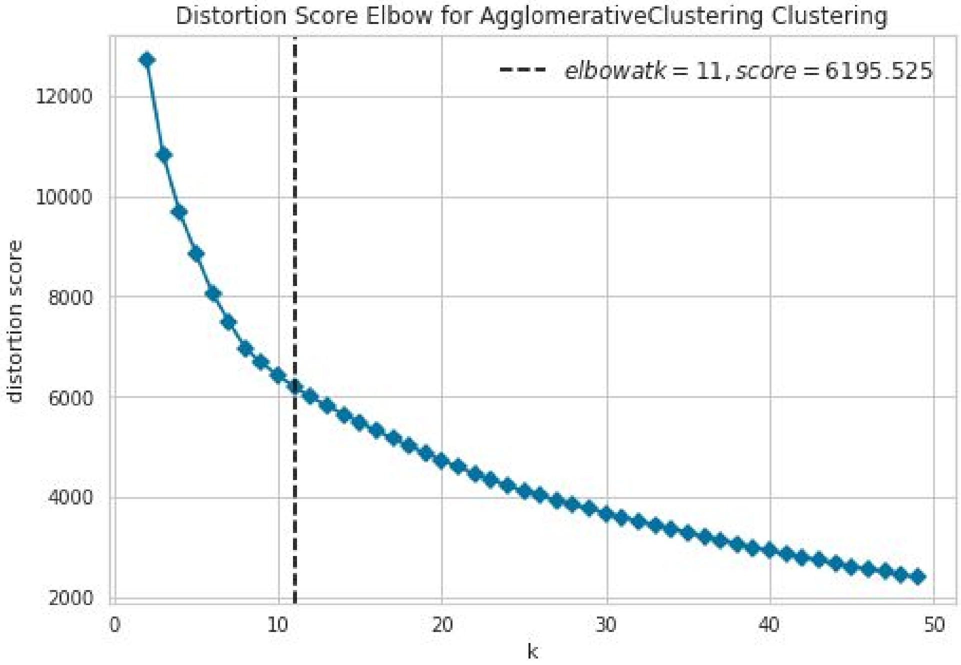
For this reason, we kept the size of different categories equal in the model, although option 3 would have been preferred.

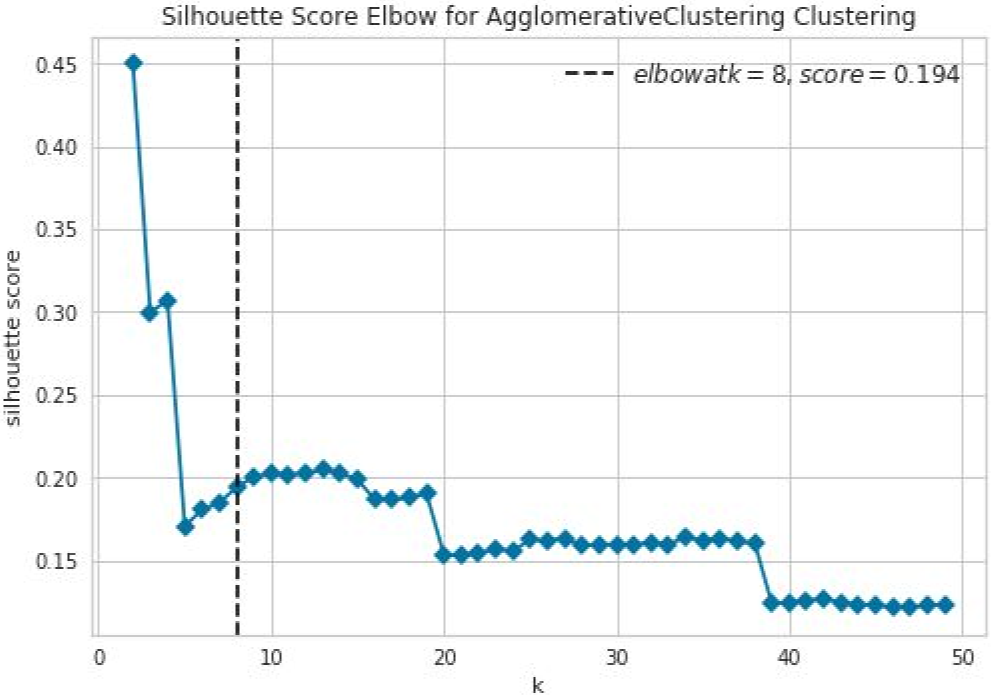
3.3 Cluster Modelling

For Cluster Modelling, I chose a K-Cluster based model approach as it's a good starting point for these types of problems, I chose k using the elbow method using the distortion score, whilst also comparing to the Silhouette Score.

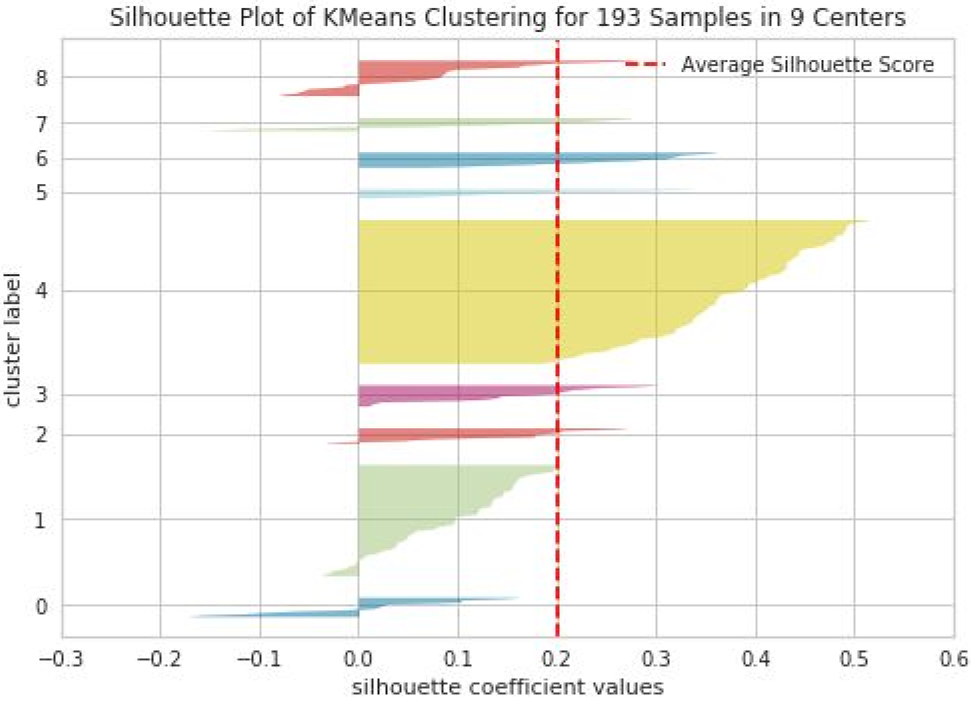
4. Results

4.1 Choosing K and Scoring the model



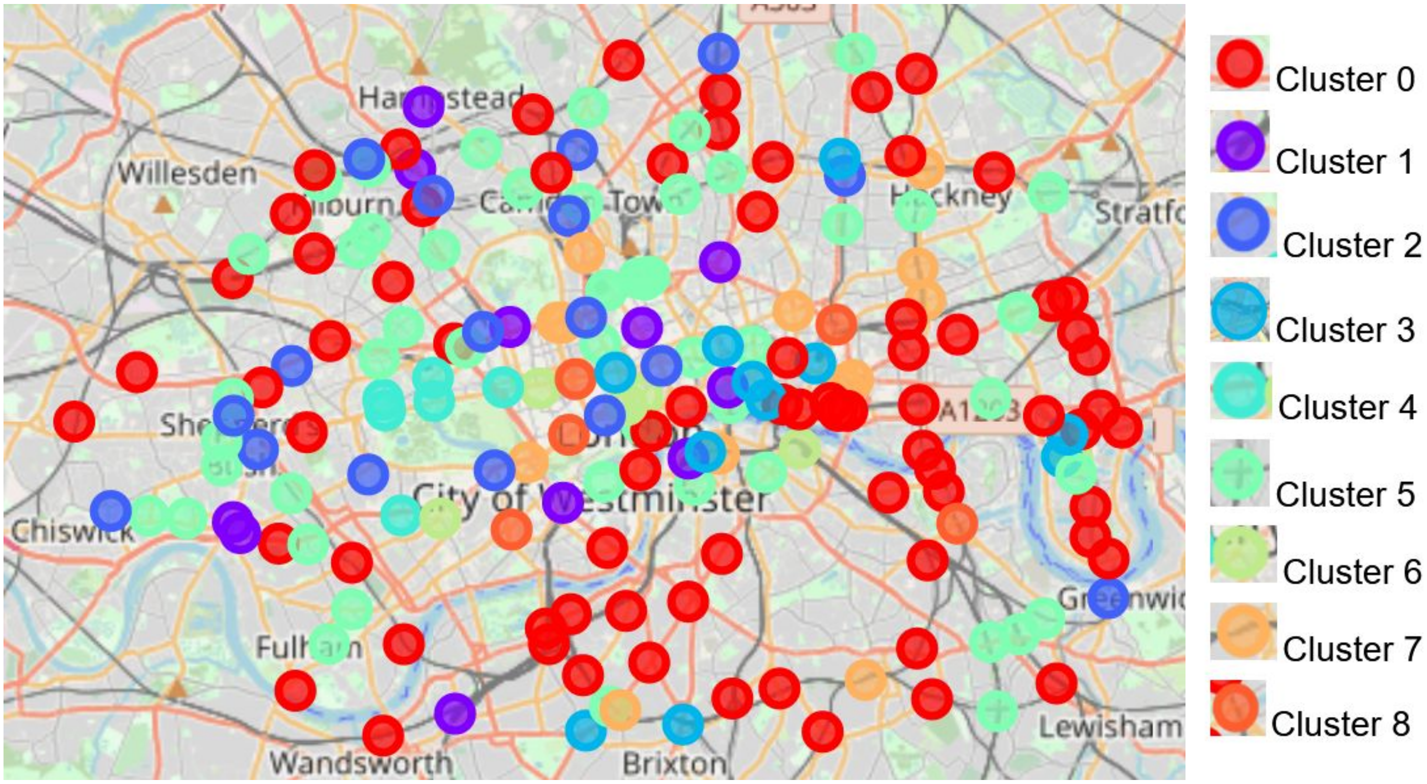


In this instance, I chose 9 clusters as a good compromise between the two scores.



The results at this stage compare very favourably to the default method provided in the course without mapping of venues. A comparison can be seen in the discussion section.

4.2 The Clusters



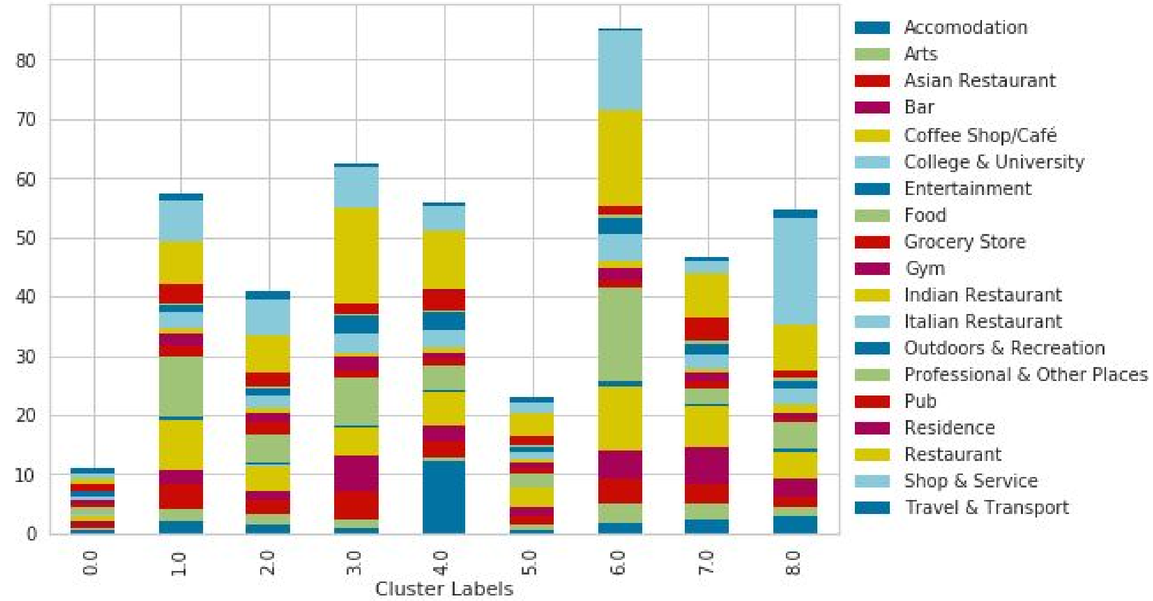


Chart: Average number of each venue present in each cluster

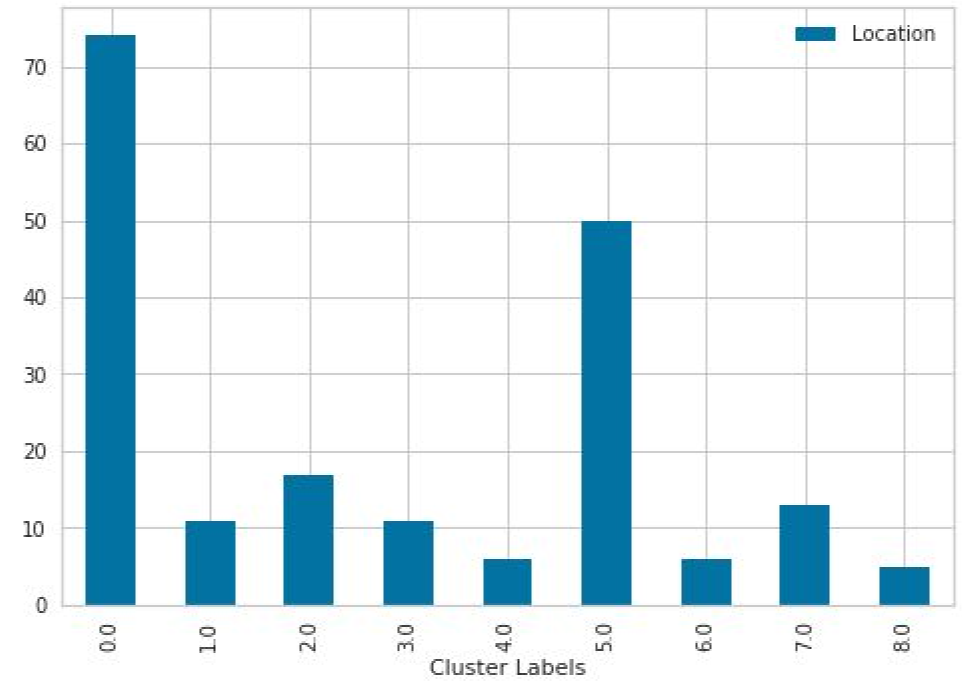


Chart: Number of locations in each Cluster

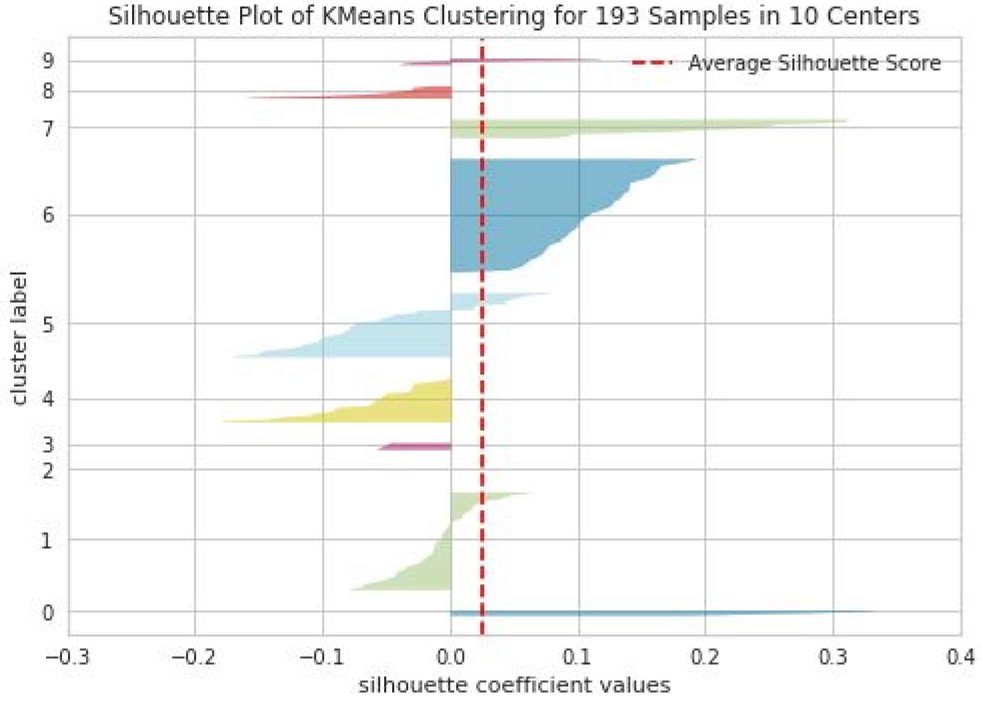
We can see some clear patterns in the 9 clusters:

Cluster 0 - This is the largest cluster focussed predominantly well outside the central zone, very low levels of business across the board, most interesting thing is that these were so well grouped together.
Cluster 1 - Lots of cafes and not so many bars, good place to get food to go but less restaurants
Cluster 2 - Arts and entertainment can be found here
Cluster 3 - Bars, Food, Gyms and Restaurants
Cluster 4 - This cluster has great accommodation, along with supporting services.
Cluster 5 - Is the second largest cluster, it's a bit of a mix of lots of different categories.
Cluster 6 - This comprises 6 of the biggest shopping and eating areas in London, Bond Street, Canary Wharf, Covent Garden, Leicester Square, London Bridge, and South Kensington with lots of food, coffee and restaurants supporting.
Cluster 7 - Bars, Cafes and Restaurants are common in these locations.
Cluster 8 - This compromises of areas good for shopping but less well known for the restaurants, Green Park, Oxford Circus, Shoreditch High Street, Sloane Square, Surrey Quays

Overall we can see that clustering has been quite successful, we are able to identify areas good for shopping, shopping and eating, accommodation, and arts which is really useful from a tourist perspective.

5. Discussion

Before looking to build an improved clustering system, I tested performance using the same method found in the IBM course, it was hard to find a good number for k, in particular we see very poor silhouette scores, across most clusters.



Given how much this improved this, it would be interesting to take categorisation further, you could:

- Scrape businesses details to better categorise what type they are.
- Build sub categories knowledge into cluster analysis, so categories can still be cluster on their headings as well as their sub type.

Further to this, we could find a better method for clustering, in particular i think anomaly detection is a problem were we have unusual locations which is to be expected, for this reason DBSCAN would be interesting to try. Agglomerative Clustering was tested for this work, but did not provide any meaningful uplift in results, although it also worked better on the categorized data.

As mentioned previously, I feel that some venues are under-indexed in that they are large venue that lots of people would attend, this could be solved relatively easily with the premium API for FourSquare and might improve our ability to detect areas of significance for Arts and entertainment which typically have larger venues.

6. Conclusion

Improving categorisation drastically improved the ability to build interesting clusters around parts of London to show what different stations are good for. Whilst this was my first exercise in clustering, the resultant clustering labels actually look very useful when investigated from a data standpoint and when comparing back my local knowledge.