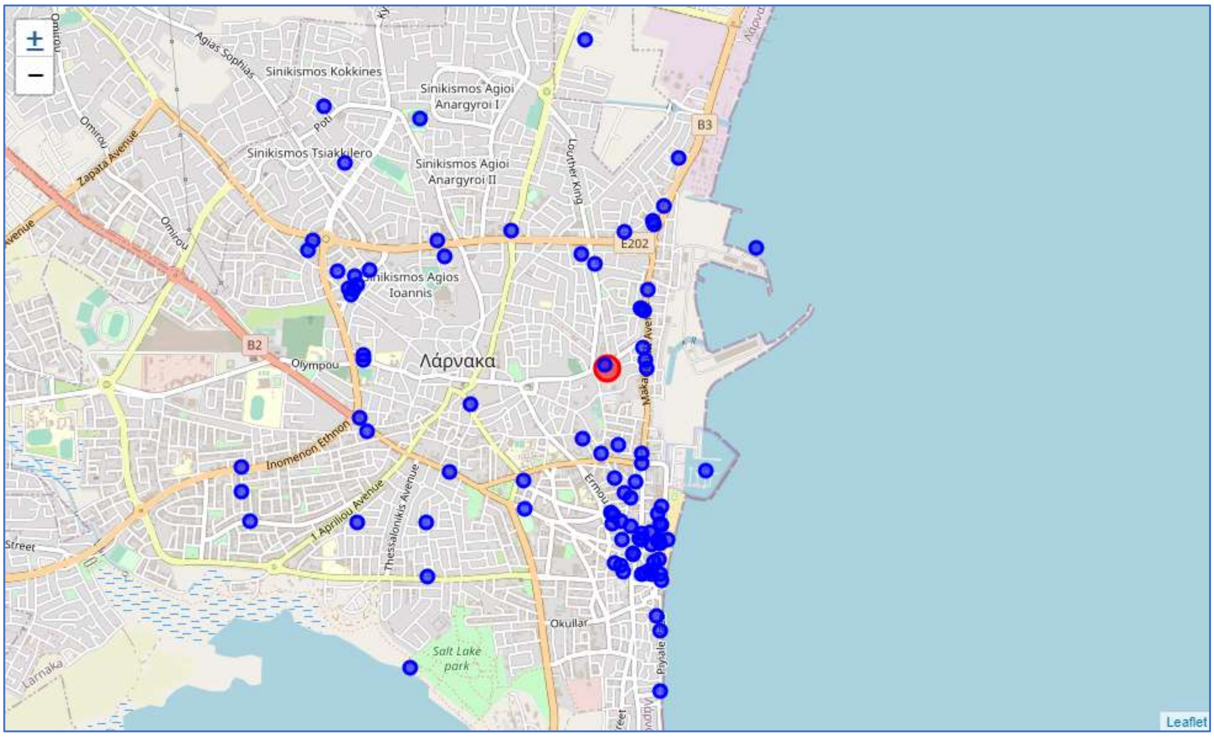
The Battle of Neighbourhoods – Larnaca



Applied Data Science – Capstone Project

IBM Data Science Processional Certificate

Thomas Buchan

1. Introduction

Background

People must relocate to new cities for a variety of work and family reasons and this process can be a stressful and challenging time. Having additional information about areas of a new city/country that your relocating too can aid in the process making it smoother, less stressful and leading to an overall better outcome.

This project would be interesting to any user who is having to relocate to a different city/country. They could utilise this project template to compare between different cities. It could also prove useful to companies' human resources departments to aid in the relocation process of their staff to new locations.

Business Problem

This project aims to utilise location data to analyse and compare similarities between different locations and neighbourhoods within cities and allow any user interested in relocating to determine the neighbourhood in a new city that best fits them.

Specifically, this project will use a relocation from Cyprus to London as an example however the methodology and notebooks would be such that the process could be repeated from the users current location to another city.

2. Data Section

This project will use data from London and Cyprus. The aim being to determine common elements of the area in Cyprus where I currently reside and then attempt to discover similar services within a borough in London. This is to be conducted to allow a user looking to relocate to be able to find an area in another country/city with similar traits to their home.

Neighborhoods

In order to carry out this project location data will be sourced for Cyprus and London. The Cyprus information will consist of a single location point of the coordinates of my current city. The list of London Boroughs was pulled from Wikipedia https://en.wikipedia.org/wiki/List_of_London_borough and Beautiful soup was used to construct a data frame. This table contained Borough Names and coordinates so only a dingle data source had to be web scraped to get this information.

Venue Data

Foursquare location data for each of the locations in the data frames, the 32 London boroughs and Larnaca were obtained through foursquare API.
https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}

Further Data

Additional data sets of information such as real estate prices, crim statistics, area demographics etc can be added as additional sorting information but were not included within this report.

3. Methodology

This section details the methodology and libraries utilised to arrive at the results. A Python Jupyter notebook was used throughout the project the libraries and methods utilised throughout are described below;

- Pandas and Numpy – Common python libraries for structured data
- Matplotlib – A plotting library
- Plotly – Another visualisation package
- Scikit Learn – Machine Learning for the K means clustering
- Requests – for HTTP requests
- Geopy – To obtain location coordinates
- Folium – visualise data on a Leaflet map

The main steps involved in the process were Web Scraping, Data Cleaning, Data Exploration, Machine Learning, Results and visualisations.

Web Scraping

The different boroughs in London were obtained through web scraping the Wikipedia page and obtaining the borough names and coordinate information.

Additionally the Foursquare API requests was sent to obtain data on the 32 London Boroughs and the Larnaca point.

Data Cleaning

The data was cleaned and set into data frame format to be easily manipulated through the machine learning algorithms and visualised. The location data was cleaned into a single data frame containing Borough name, Latitude, Longitude and could then be plotted as a quality check, this is shown in the figure below.

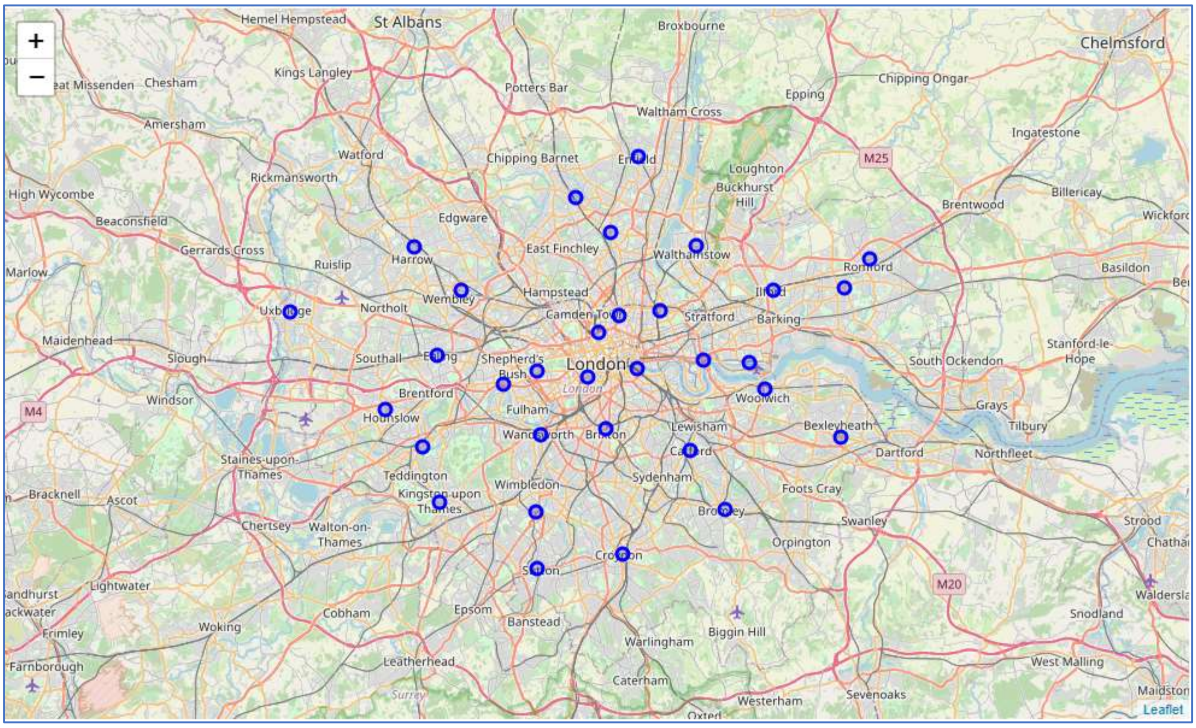
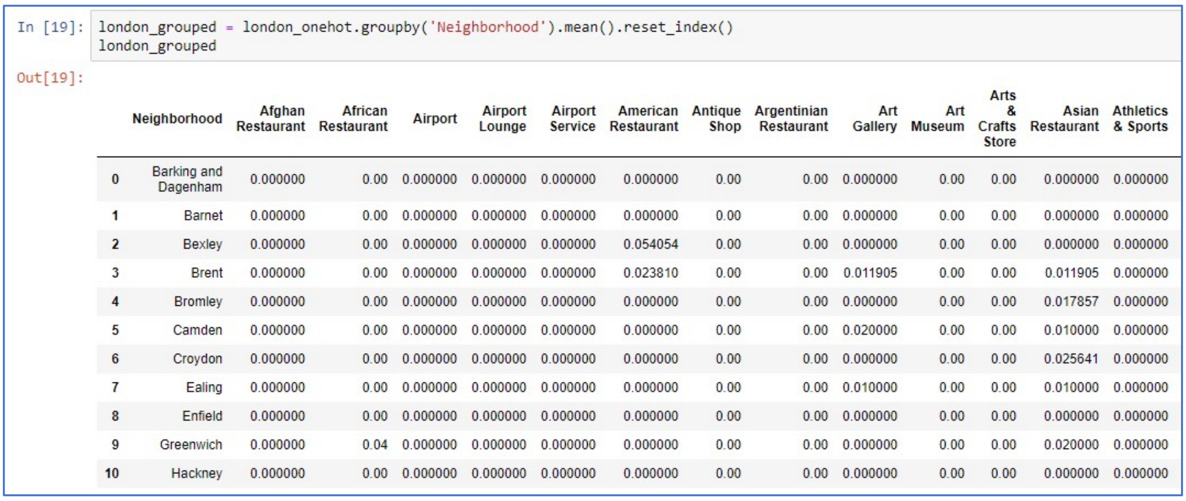


Figure 1 - London Borough Location Data

The data cleaning process for the Foursquare API request consisted of cleaning up the information received from the request so that the data frame contained usable information on Venue, Venue category, Coordinates and Borough. This was then used for the data exploration.

Data Exploration

The London borough Foursquare data collection shows that there were 251 categories of venue with 1953 total venues returned. It was noted that 11 of the 32 boroughs maxed out with the request size of 100 venues. One hot encoding was then used to manipulate the venue categories to numerical formats as shown below;



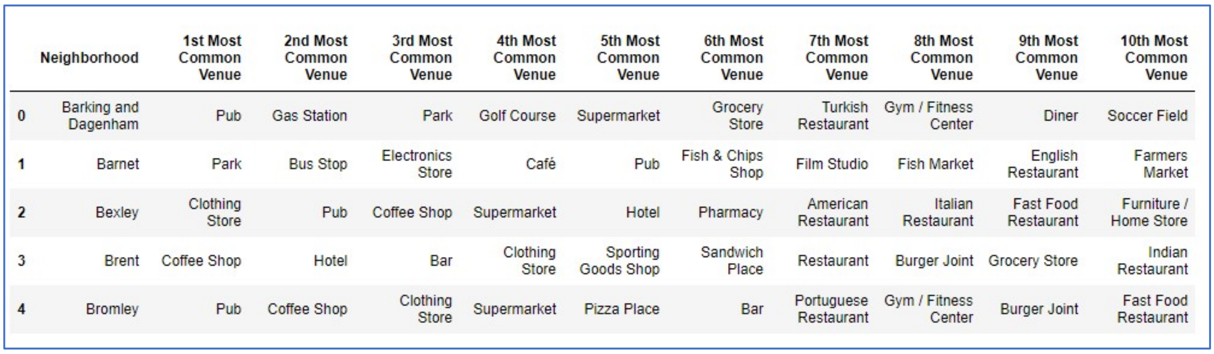
```
In [19]: london_grouped = london_onehot.groupby('Neighborhood').mean().reset_index()
         london_grouped
```

Out[19]:

	Neighborhood	Afghan Restaurant	African Restaurant	Airport	Airport Lounge	Airport Service	American Restaurant	Antique Shop	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports
0	Barking and Dagenham	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.00	0.000000	0.00	0.00	0.000000	0.000000
1	Barnet	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.00	0.000000	0.00	0.00	0.000000	0.000000
2	Bexley	0.000000	0.00	0.000000	0.000000	0.000000	0.054054	0.00	0.00	0.000000	0.00	0.00	0.000000	0.000000
3	Brent	0.000000	0.00	0.000000	0.000000	0.000000	0.023810	0.00	0.00	0.011905	0.00	0.00	0.011905	0.000000
4	Bromley	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.00	0.000000	0.00	0.00	0.017857	0.000000
5	Camden	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.00	0.020000	0.00	0.00	0.010000	0.000000
6	Croydon	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.00	0.000000	0.00	0.00	0.025641	0.000000
7	Ealing	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.00	0.010000	0.00	0.00	0.010000	0.000000
8	Enfield	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.00	0.000000	0.00	0.00	0.000000	0.000000
9	Greenwich	0.000000	0.04	0.000000	0.000000	0.000000	0.000000	0.00	0.00	0.000000	0.00	0.00	0.020000	0.000000
10	Hackney	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.00	0.000000	0.00	0.00	0.000000	0.000000

Figure 2 - Foursquare data manipulated through one hot encoding

The ten most common categories of venues at each location were then grouped from this information ready to further manipulate.



	Neighborhood	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	9th Most Common Venue	10th Most Common Venue
0	Barking and Dagenham	Pub	Gas Station	Park	Golf Course	Supermarket	Grocery Store	Turkish Restaurant	Gym / Fitness Center	Diner	Soccer Field
1	Barnet	Park	Bus Stop	Electronics Store	Café	Pub	Fish & Chips Shop	Film Studio	Fish Market	English Restaurant	Farmers Market
2	Bexley	Clothing Store	Pub	Coffee Shop	Supermarket	Hotel	Pharmacy	American Restaurant	Italian Restaurant	Fast Food Restaurant	Furniture / Home Store
3	Brent	Coffee Shop	Hotel	Bar	Clothing Store	Sporting Goods Shop	Sandwich Place	Restaurant	Burger Joint	Grocery Store	Indian Restaurant
4	Bromley	Pub	Coffee Shop	Clothing Store	Supermarket	Pizza Place	Bar	Portuguese Restaurant	Gym / Fitness Center	Burger Joint	Fast Food Restaurant

Figure 3 - Neighborhoods_Venues_Sorted

The table shows the top ten most common venue types for each of the boroughs and will be used to cluster the areas.

Additionally, to the London data information. The foursquare request for Larnaca was obtained and plotted onto the chart using Folium.

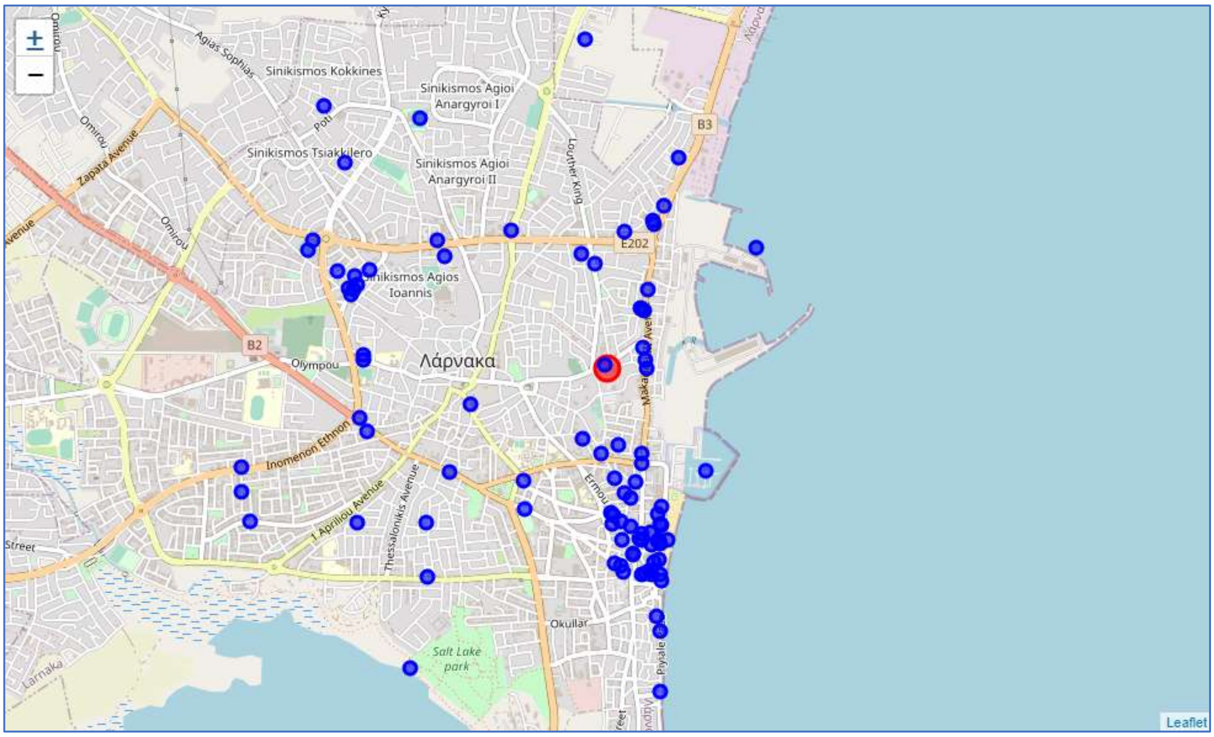


Figure 4 - Larnaca Venue Location Map

Machine Learning

The different boroughs in London were then clustered, first determining number of clusters required through the elbow method, according to the top venues available in its area. The algorithm K-Means as part of the scikit learn package was used to group the boroughs by the most common types of venues within each borough. This machine learning algorithm groups the data into the desired number of different clusters, k, so that boroughs with similar venue characteristics will be grouped together.

The steps carried out were to determine the number of clusters required. The optimum number of clusters was determined through using the elbow method. The name deriving from the point of inflection of the line determining the best fit for k, in this case 4 clusters were required as shown in the figure below.

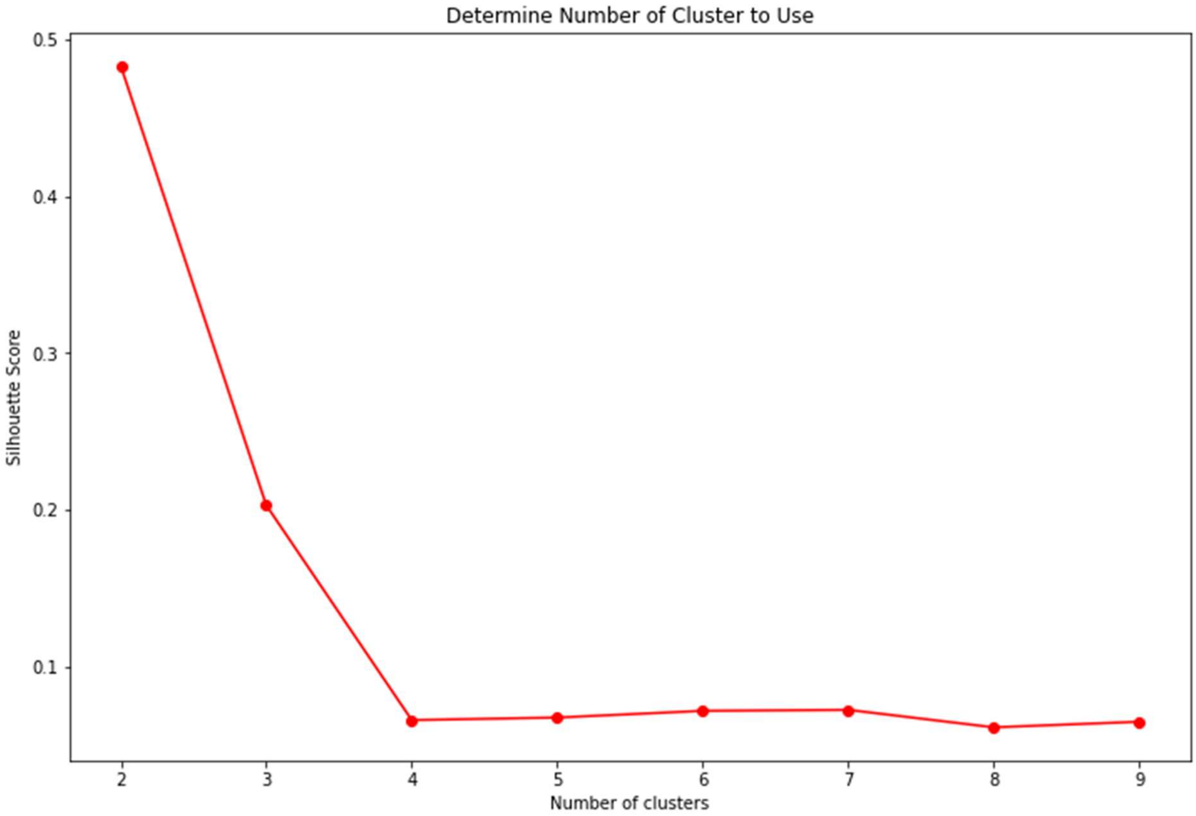
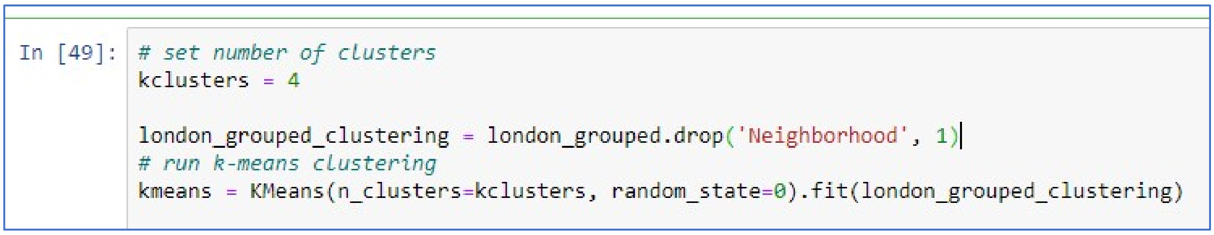


Figure 5 - Number of Clusters Calculation

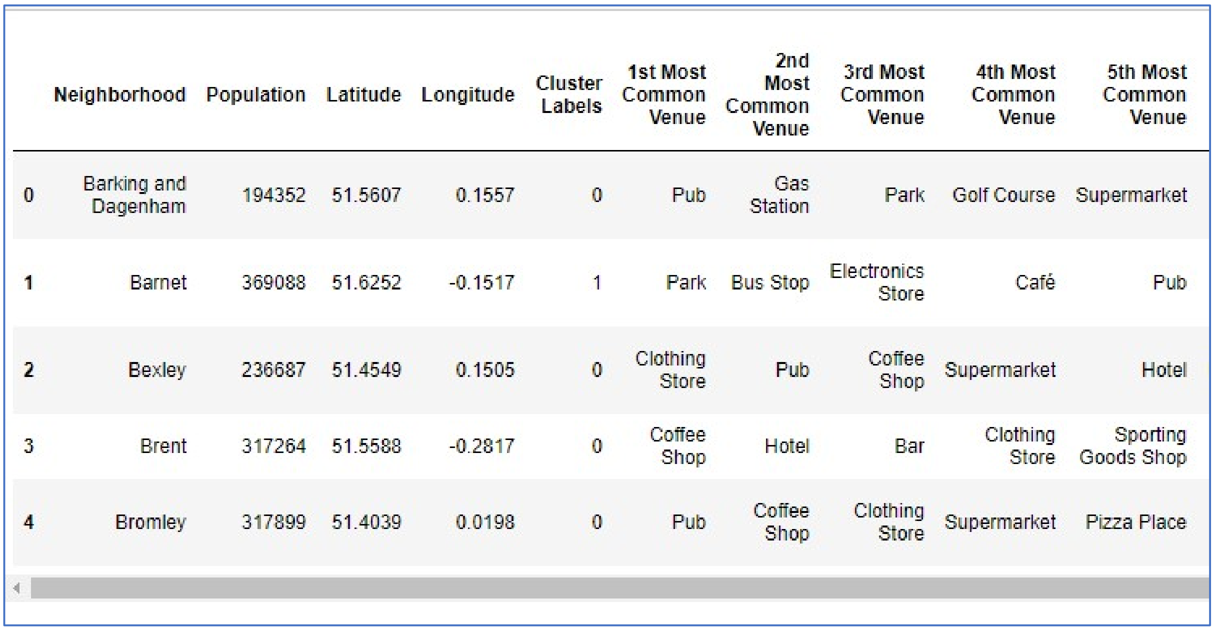
Then the clusters were determined then the groupings of the clusters plotted using Folium with a different colour representing each cluster.



```
In [49]: # set number of clusters
         kclusters = 4

         london_grouped_clustering = london_grouped.drop('Neighborhood', 1)
         # run k-means clustering
         kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(london_grouped_clustering)
```

Figure 6 - K-means Clustering



	Neighborhood	Population	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Barking and Dagenham	194352	51.5607	0.1557	0	Pub	Gas Station	Park	Golf Course	Supermarket
1	Barnet	369088	51.6252	-0.1517	1	Park	Bus Stop	Electronics Store	Café	Pub
2	Bexley	236687	51.4549	0.1505	0	Clothing Store	Pub	Coffee Shop	Supermarket	Hotel
3	Brent	317264	51.5588	-0.2817	0	Coffee Shop	Hotel	Bar	Clothing Store	Sporting Goods Shop
4	Bromley	317899	51.4039	0.0198	0	Pub	Coffee Shop	Clothing Store	Supermarket	Pizza Place

Figure 7 - Borough Clusters, Location and Venue Data Frame

The above data frame with the results of the k means clustering, the Borough coordinates and the most common venue types was then used as the basis for the plotting and visualisations.

Results / Visualizations

The clusters were then be looked at to determine the closest fit to the Cyprus location information and this used as a basis for the areas to look to relocate too.

This process could then be iterated down to a reduced geographical area to further refine the process. Also, additional data sources such as crime figures, real estate prices, transport links and school information could be compared to find the most appropriate location. The final results will be described in the next section.

4. Results and Discussion

There were a number of results taken from this data with more questions being discovered than answers found with more data analysis required to obtain the desired results. This section details the current results and discusses them.

Results

The initial results for the process steps described in the previous sections consisted of a folium generated map of London with the cluster selections plotted on the map, this is shown in the figure below.

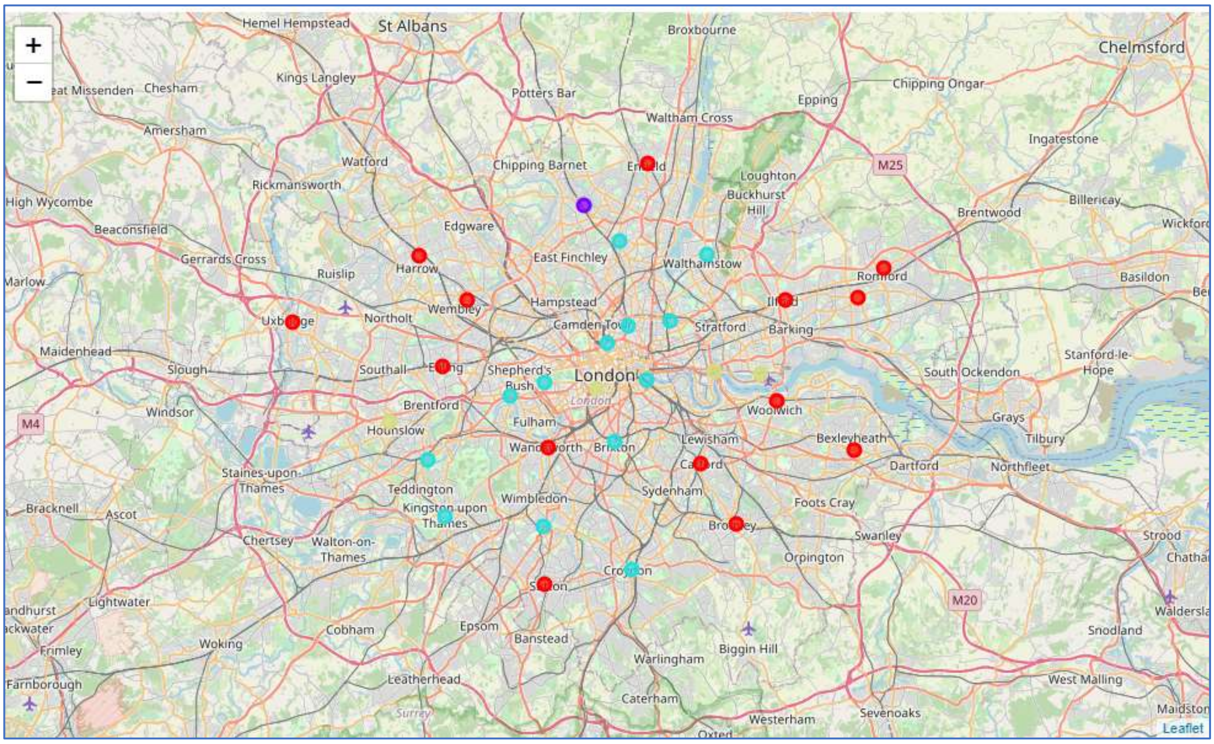


Figure 8 - London Boroughs Clustered by venue categories (key below)

Cluster 1	🟥
Cluster 2	🟪
Cluster 3	🟦
Cluster 4	🟨

As shown in the methodology four clusters were used for this London data set. The geographical relationships in the above clusters can be noted.

- Cluster 1 – This appears to be predominantly on the outskirts of the city
- Cluster 2 – Not much can be derived from this as it's a single Borough
- Cluster 3 – Seem to be closer to the centre than cluster 1 but almost dispersed amongst each other, I would expect to be seeing similarities with clusters 1 and 3.
- Cluster 4 – They appear to be in the city centre, along the river and near the airports.

The top five venue categories in each cluster were graphed and plotted below. This counts for each Borough this was present hence the Cluster 2 are all 1 due to only one Borough being present.

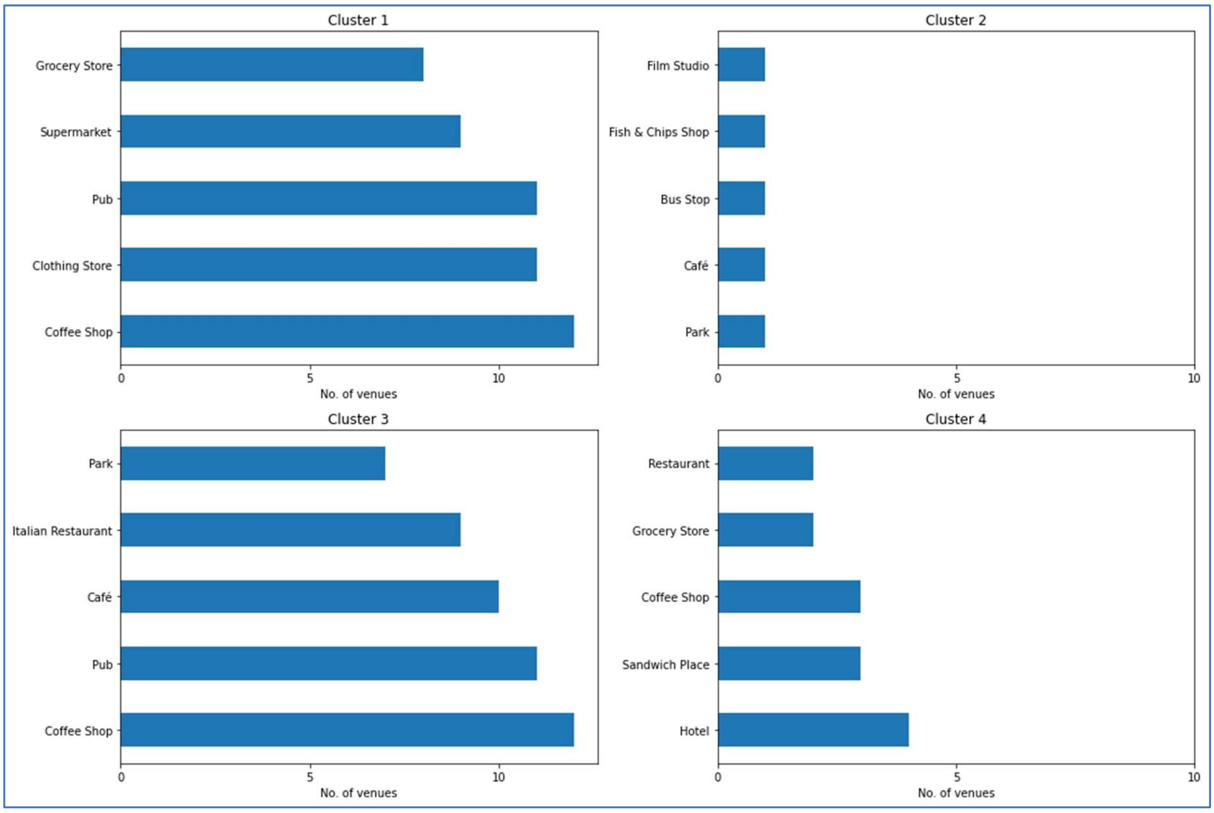


Figure 9 - The top 5 venue types for each London Cluster

For a comparison between the London Borough groupings and the current location of Larnaca the number of venues in the 5 most popular categories in Larnaca was plotted below.

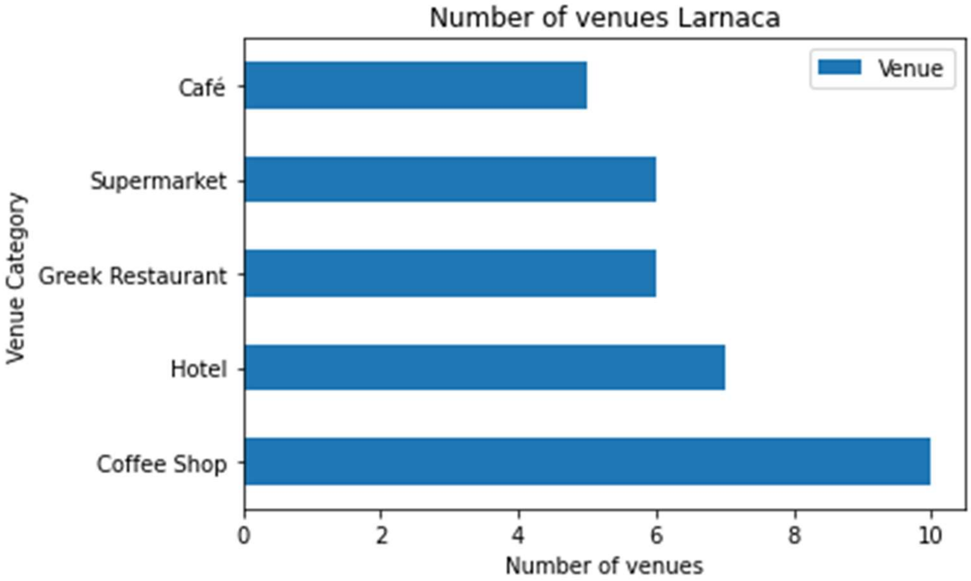


Figure 10 - The top 5 venue types for Larnaca

Discussion

There are several similar traits that can be taken from the clusters of venue categories. Cluster 2 only consisting of 1 borough hence why the top 5 venues all appear one time. It can be seen that the two most common clusters, Cluster 1 and Cluster 3, are quite similar in make up with coffee shops being the most popular and pubs also appearing in the top 5 categories. Cluster 3 has more cafes and less supermarkets/grocery stores appearing in the searches. Looking geographically on the map its interesting to see that generally cluster 1 is boroughs around the outskirts of London and cluster 3 are more central.

A feature of Cluster 4 is hotels as the top venue and these areas are in areas containing airports and the city centre. The other categories of Coffee shops, sandwich place and restaurants are also focused towards people travelling into the area whether on business or tourism.

Given this information some insights can be used to compare it to Larnaca however this information itself would not be enough to make a reasoned decision on location.

The previous chart shows the top venue types for Larnaca, the base location, with which to compare to the London clusters. From this is can be seen that as with both clusters 1 and 3 the top venue type is Coffee shop, with coffee shop also being the second highest category in Cluster 4. The types of restaurants and coffee shops were, with Larnaca being particularly heavy on the coffee shops. Also other groups may need to be carried out as there are coffee shops, cafes, coffee roasters etc. Venue categories that are different venue categories in name yet function the same as each other.

5. Conclusions

The area closest to central London or the airports is most likely closest to Larnaca based on this due to the hotels and cafes with Larnaca also being relatively tourist driven and next to the international airport. This will have effected the results and showed that this metric alone is not very useful but when refined further and coupled with additional metrics I would expect it to prove helpful at determining the most similar areas for a relocation.

- The size of a London borough is too large an area to make a comparison from the most frequently occurring within a maximum sample size of 100.
- Additional metrics would be required in addition to grouping of the most common venue types.

The purpose of the project was to aid in the decision making process involving relocation and although the concept of utilising location data for this is clear. More than the utilised data would be required both in terms of volume of venues and types of metrics in order to make a fully informed decision.

Future Work

Further work would include

- The size of a London Borough is too large to determine enough traits from 100 venues per borough to be used as the sole factor on a decision. The size of area could be reduced to each postal code region in London significantly increasing the number of areas to cluster.
- The Larnaca data point could be merged with the London data frame to see which cluster it falls in within the same data set.
- Each Borough venue results could individually be compared to the Larnaca results set to see if there is one specific one that compares.
- Additional metrics really would need to be included to aid with the decision making. These would allow primary and secondary groupings, could be done together or different groupings used for each set. The additional metrics that would be helpful to group would be;
    - Real Estate Rent/Prices
    - School Information
    - Transport Links
    - Population density
    - Population Demographic
    - Specific searches – for instance in case of Larnaca, Greek restaurants in London could be searched for.
    - Crime Rates

6. References

The information within these were obtained from

https://en.wikipedia.org/wiki/List_of_London_borough

https://api.foursquare.com

The Jupyter notebooks from previous exercises and courses were also utilised for guidance in the code.