Chris O’Sullivan
Data Science Professional Certification
27 June 2020

Employee Secondment: Neighbourhood Comparisons and Recommendation

Background Introduction

A large multi national firm offers secondment opportunities to different locations around the world for employees; this can enhance both the employee's and the businesses performance by offering a variety of experience and exposure to alternative markets.

The HR team have asked a data scientist to produce an analysis that can make pertinent suggestions of where to base living arrangements for employees who are about to be seconded.

This is important for the firm to provide a high standard of job satisfaction, looking after their employees needs in what can be a difficult transition abroad for many people. Ideally this analysis would recommend regions of cities that are most similar to the employee's home city in terms of nearby amenities.

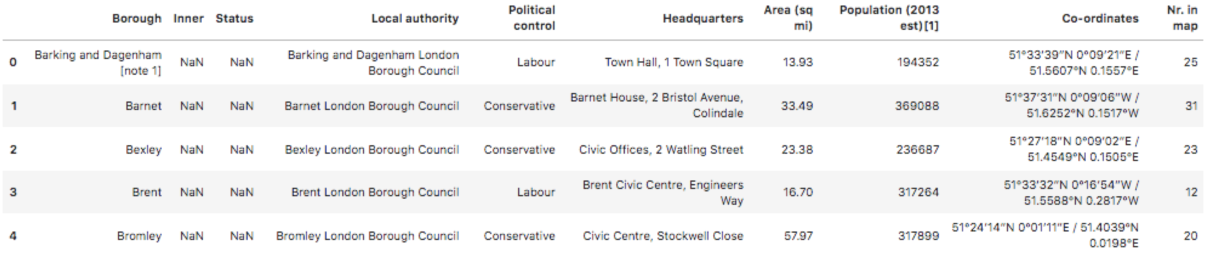
Problem

- Given the current living location of an employee, can the analyst find the most similar neighbourhood in alternative cities given a list of common amenities?
- In our example we will find the most suitable neighbourhood within New York for an employee who is currently living in Camden, London. We will compare NY to London and determine which is the most suitable neighbourhood for our employee's secondment, considering only local amenities in this case.
- This will be done by performing a similarity analysis in a K Means Classifier on the list of common venues to determine the locations with the most similarity to Camden Town.
- A general visual comparison of the two cities can also be performed from the results of the analysis.

Data Sources and Usage

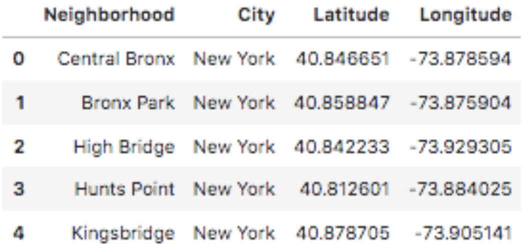
The following data sources will be used to demonstrate our solution to this location clustering problem:

- City Neighbourhoods for London and New York:
    - To obtain the list of neighbourhoods, we will perform web scraping. For example shown below are the results of the London web scraping from https://en.wikipedia.org/wiki/List_of_London_boroughs. This data already contains latitude and longitudinal coordinates but must be cleaned to split them.



	Borough	Inner	Status	Local authority	Political control	Headquarters	Area (sq mi)	Population (2013 est) [1]	Co-ordinates	Nr. in map
0	Barking and Dagenham [note 1]	NaN	NaN	Barking and Dagenham London Borough Council	Labour	Town Hall, 1 Town Square	13.93	194352	51°33′39″N 0°09′21″E / 51.5607°N 0.1557°E	25
1	Barnet	NaN	NaN	Barnet London Borough Council	Conservative	Barnet House, 2 Bristol Avenue, Colindale	33.49	369088	51°37′31″N 0°09′06″W / 51.6252°N 0.1517°W	31
2	Bexley	NaN	NaN	Bexley London Borough Council	Conservative	Civic Offices, 2 Watling Street	23.38	236687	51°27′18″N 0°09′02″E / 51.4549°N 0.1505°E	23
3	Brent	NaN	NaN	Brent London Borough Council	Labour	Brent Civic Centre, Engineers Way	16.70	317264	51°33′32″N 0°16′54″W / 51.5588°N 0.2817°W	12
4	Bromley	NaN	NaN	Bromley London Borough Council	Conservative	Civic Centre, Stockwell Close	57.97	317899	51°24′14″N 0°01′11″E / 51.4039°N 0.0198°E	20

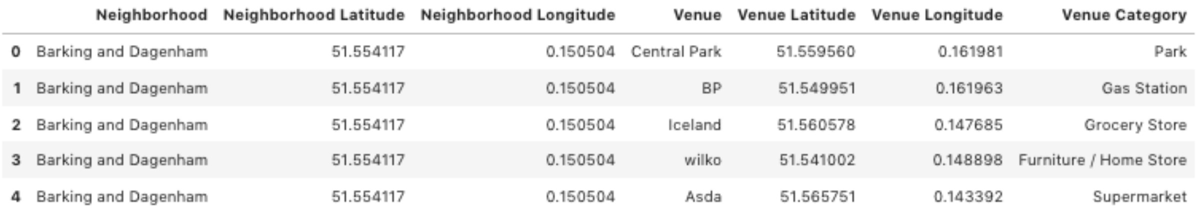
- To obtain the latitude and longitude for New York, we will use the geopy geocode python module to extract the data from the borough and city combination, amending the results to our data frame. An example of the final data for New York is shown below.



	Neighborhood	City	Latitude	Longitude
0	Central Bronx	New York	40.846651	-73.878594
1	Bronx Park	New York	40.858847	-73.875904
2	High Bridge	New York	40.842233	-73.929305
3	Hunts Point	New York	40.812601	-73.884025
4	Kingsbridge	New York	40.878705	-73.905141

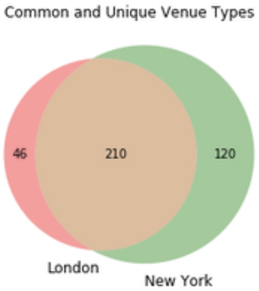
- In total there are 32 London neighbourhoods and 54 New York neighbourhoods.

- Foursquare venue data for all neighbourhoods to produce a data frame of amenities.
    - We will select the 100 nearest amenities in a 1600m radius to the coordinates we have been given. An example of this call is shown below for London:



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Barking and Dagenham	51.554117	0.150504	Central Park	51.559560	0.161981	Park
1	Barking and Dagenham	51.554117	0.150504	BP	51.549951	0.161963	Gas Station
2	Barking and Dagenham	51.554117	0.150504	Iceland	51.560578	0.147685	Grocery Store
3	Barking and Dagenham	51.554117	0.150504	wilko	51.541002	0.148898	Furniture / Home Store
4	Barking and Dagenham	51.554117	0.150504	Asda	51.565751	0.143392	Supermarket

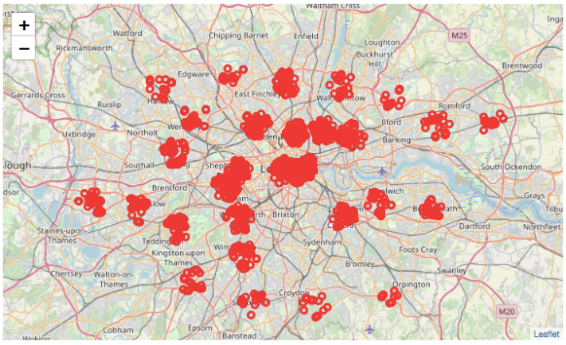
    - In total we found 256 venue types in London and 330 Types in New York.
    - To perform representative analysis, we will reduce the data frame to contain only common venue types across the three cities. This is a field of 210 venue types, a reasonable number for classifying similarity across neighbourhoods, including items such as train stations, restaurants, gyms and bars.

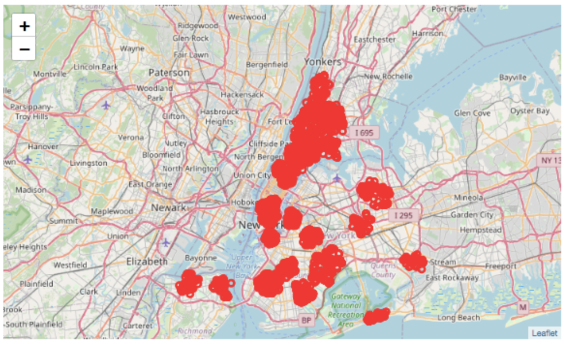


    - We will train our classifier on this data. This data will be categorical so to train a K means classifier, this data will be converted to numeric data via one hot encoding for all venue types.
    - A similarity analysis will allow us to characterise the locations in New York that are most similar to the borough of Camden Town and to perform a general visual comparison of the cities and clusters.

Methodology

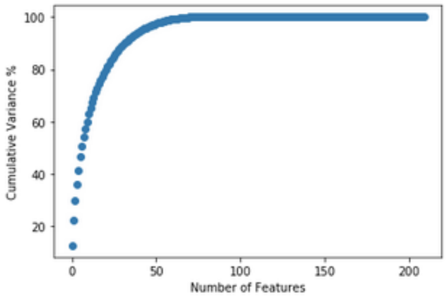
Here we set out to obtaining common venue types through the foursquare API, deleting duplicates where boroughs were found in close proximity. We ended up with a range of venue densities and a good amount of coverage of the map in both cities as shown below. We do however have gaps in the flushing meadows area of New York and the Brixton Area of London, potentially missing out on capturing some key cultural venues. This is partly due to the borough coordinates being sparsely distributed through the cities in the Nomatim API calls.





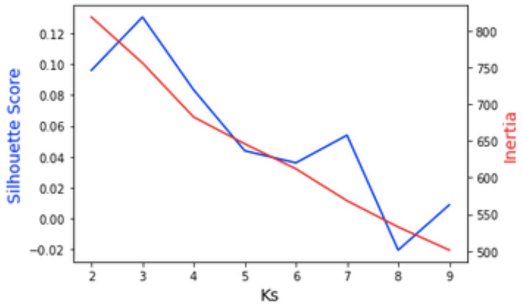
After this we created a set of one hot encoded categorical variables ending up with a dataset of 86 Neighbourhoods and 210 Features, representing 3846 venues in total.

This means significantly more features than samples and so singular value decomposition was performed on the feature covariance matrix to reduce the size of the feature space, using the features that describe a sufficient amount of the variance between samples. Below shows a plot of cumulative variance against feature number. The most important 30 features described >85% of the variance in our dataset and so the final dataset for clustering was reduced to 86x30 using the 30 eigenvectors of the covariance matrix.



The alternative to this is not to perform this reduction and potentially allow noise to dominate, with no clear distinction between clusters. Prior to this SVD decomposition we used a robust scalar on our feature set to remove the influence of outliers on largely skewed data profiles.

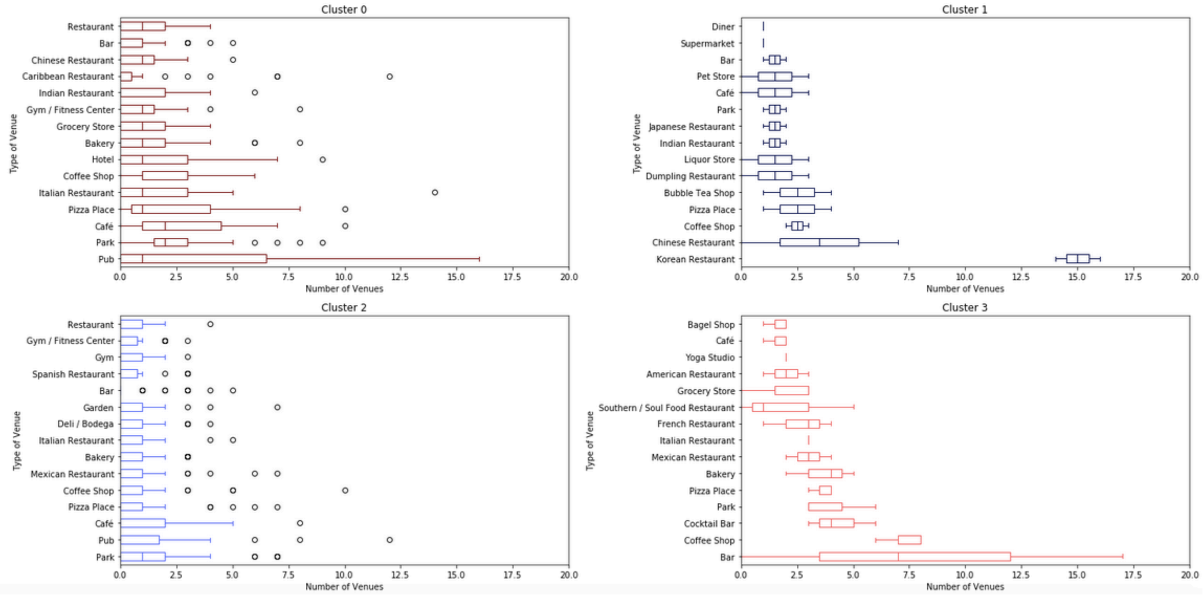
Following this SVD approach we performed our K Means clustering analysis, using the elbow method and silhouette score to select a good clustering number, which ended up being 4 with a clear elbow in inertia and still a higher silhouette score.



To visualise the clustered data we used Folium and box plots of the most significant 15 venue types in each cluster, ordered by mean values in each cluster. This gives us a good view of how each city and cluster is organised. When the cluster containing Camden is identified, we will use the closest cosine distance within that cluster for New York neighbourhoods to identify the most similar to Camden.

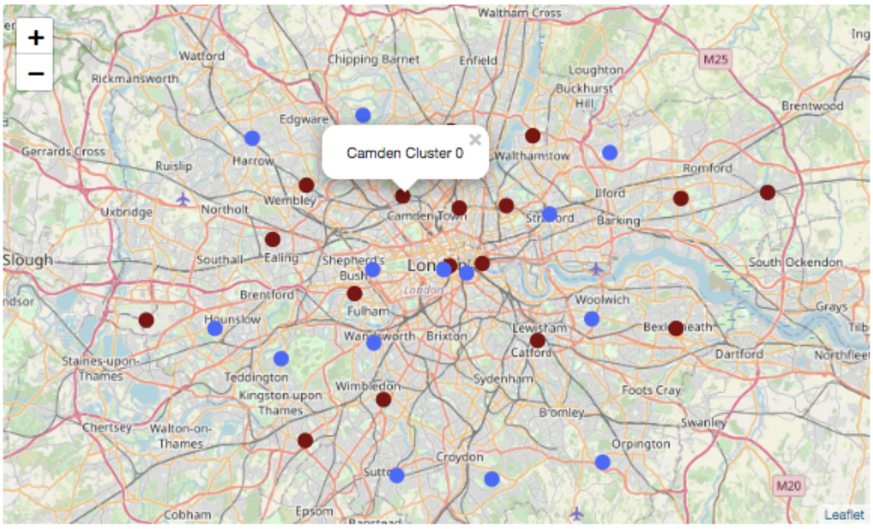
Results

Our main clustering results are shown in the box plot below. Here the most significant 15 venue types are plotted for each cluster.

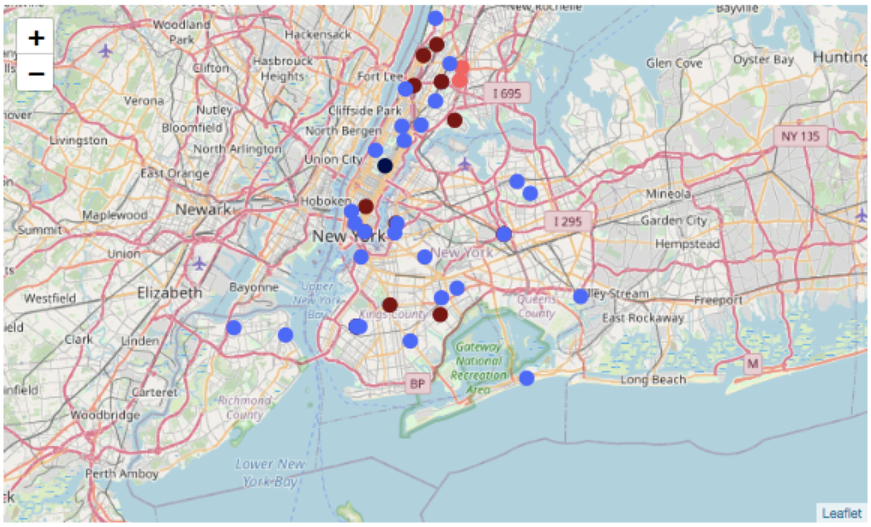


- Cluster 0 size is 35 out of 86 Neighbourhoods
- Cluster 1 size is 2 out of 86 Neighbourhoods
- Cluster 2 size is 46 out of 86 Neighbourhoods
- Cluster 3 size is 3 out of 86 Neighbourhoods

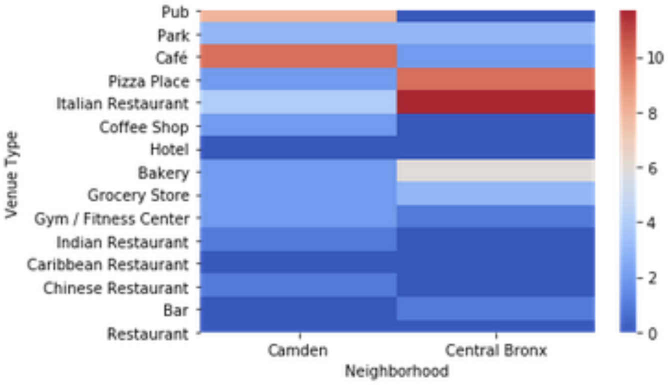
Our Folium Markers for each neighbourhood were then updated with the cluster ID to get a feel of how the two cities compare.
London: Contains only two clusters, 0 and 2. Camden is in cluster 0.



New York: Mainly clusters 0 and 2, but some outlying clusters also.



Our main enquiry for our employee was which neighbourhood in NY is closest to Camden and our analysis showed that to be Central Bronx. Below we present a heat map of the Clusters 15 significant venues to view a final comparison of both neighbourhoods.



Discussion

- *General:* Comparing London to New York, we have found two clusters with broadly similar characteristics but one with a lower density of venue counts indicating more residential and less business centric neighbourhoods. In New York we have also found some highly dense areas of Asian centric venues and other restaurant types.
- *Cluster Analysis*: Cluster 0 has Low average venue count < 2.5 mean but a wide range in density and dominated by pubs, cafes and parks. Cluster 1 is small cluster dominated by Asian cuisine. Cluster 2 is similar in nature to cluster 0 but with much lower density. Cluster 3 has Larger means > 2.5 and is a small cluster dominated by bars and coffee shops with large counts of different restaurant types.
- *Camden vs Central Bronx*: In General correlation is good in terms of the amount of venues in our key categories within this cluster, there is a noted difference in that Pubs and Cafes are more prevalent in Camden whereas Italian and Pizzerias are more prevalent in the Bronx. This is probably an expected difference given the nature of the UK and New York cultures. We can say that based on venues, the two neighbourhoods are broadly similar and would be a good starting location for our employee to consider basing themselves for their secondment.

Conclusion

It is possible to produce a comparison between neighbourhoods based on venues to produce a lifestyle recommendation for employee secondment but as we can see in our methodology section, the foursquare approach used here doesn’t capture the spread of venues across the city as we can see using Folium. As the number of features is large, we only needed the most significant 30 to describe much of the variance and to make the study better and benefit employees on different salary bands we could add the average rental costs of each neighbourhood as a feature engineering exercise to provide more defined clusters and a more suitable recommendations