Employee Secondment: Neighbourhood Comparison and Recommendation

1. Business Problem
2. Data Collection
3. Methodology
4. Results + Discussion
5. Conclusions

Chris O’Sullivan

Data Science Capstone Final Presentation

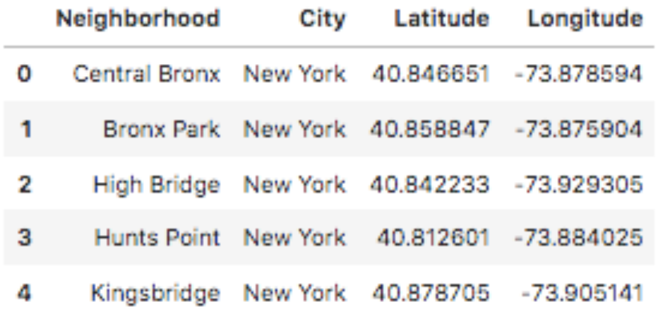
27 June 2020

Business Problem

- Employee Secondments are a common feature in development for both businesses and their employees.
- In this study we take the role of a HR department, tasked with using a data science and machine learning methodology to compare an employees current (London) and intended city (New York)
- Can we leverage the Foursquare API to produce a neighbourhood comparison and suggest a neighbourhood with a similar lifestyle to that of an employees home town of Camden, London?

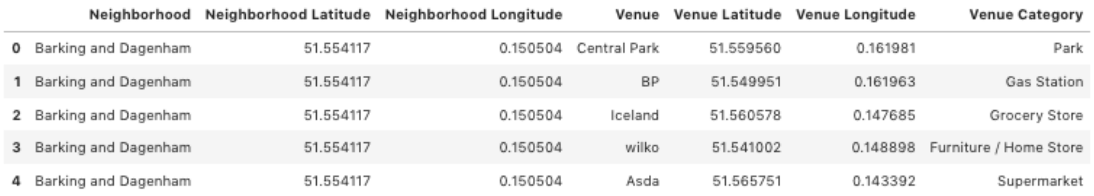
Data

- Listings for 32 Boroughs in London and 54 Neighbourhoods in New York were used alongside the Nomatim API call to obtain Latitude and Longitude Information.



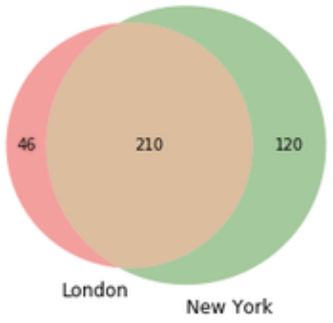
	Neighborhood	City	Latitude	Longitude
0	Central Bronx	New York	40.846651	-73.878594
1	Bronx Park	New York	40.858847	-73.875904
2	High Bridge	New York	40.842233	-73.929305
3	Hunts Point	New York	40.812601	-73.884025
4	Kingsbridge	New York	40.878705	-73.905141

- The Foursquare API call was used with this data to return the 100 nearest venues within 1 mile (1600m) of our neighbourhood coordinates. We decided to analyse a dataset of common venue types, of which there were 210.



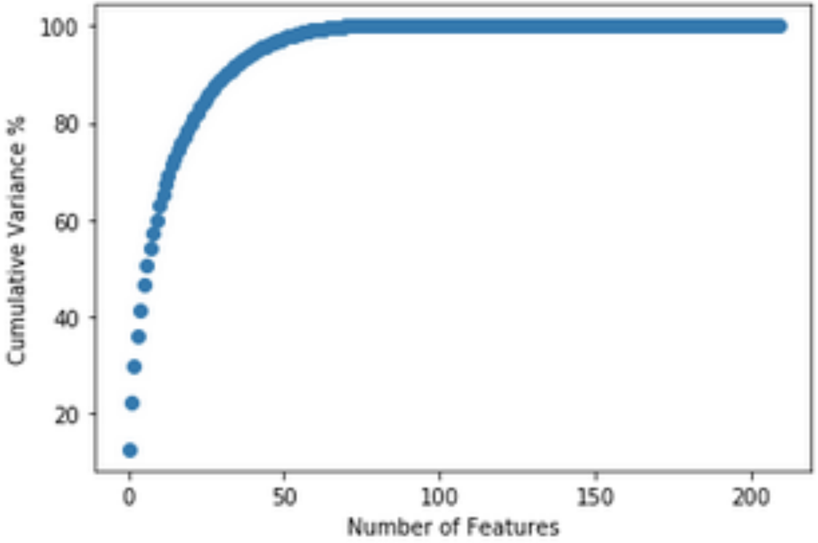
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Barking and Dagenham	51.554117	0.150504	Central Park	51.559560	0.161981	Park
1	Barking and Dagenham	51.554117	0.150504	BP	51.549951	0.161963	Gas Station
2	Barking and Dagenham	51.554117	0.150504	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

Common and Unique Venue Types



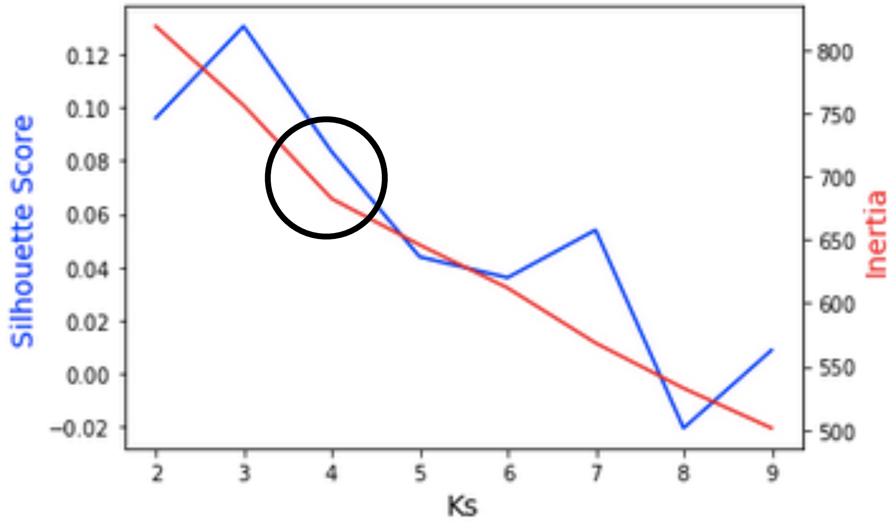
Methodology - Singular Value Decomposition

- To perform a K Means Clustering on our data to find like groups of neighbourhoods, we had to reduce our dataset using singular value decomposition due to the fact that we have a significant amount more features (210) than samples (86).
- SVD uses the most significant eigenvectors of the covariance matrix of features (size 210 by 210) to reduce the size of the dataset, whilst leaving a dataset that still contains a significant amount of the original variance (>85%). This reduces the effect of noise on the fitting of the clusters

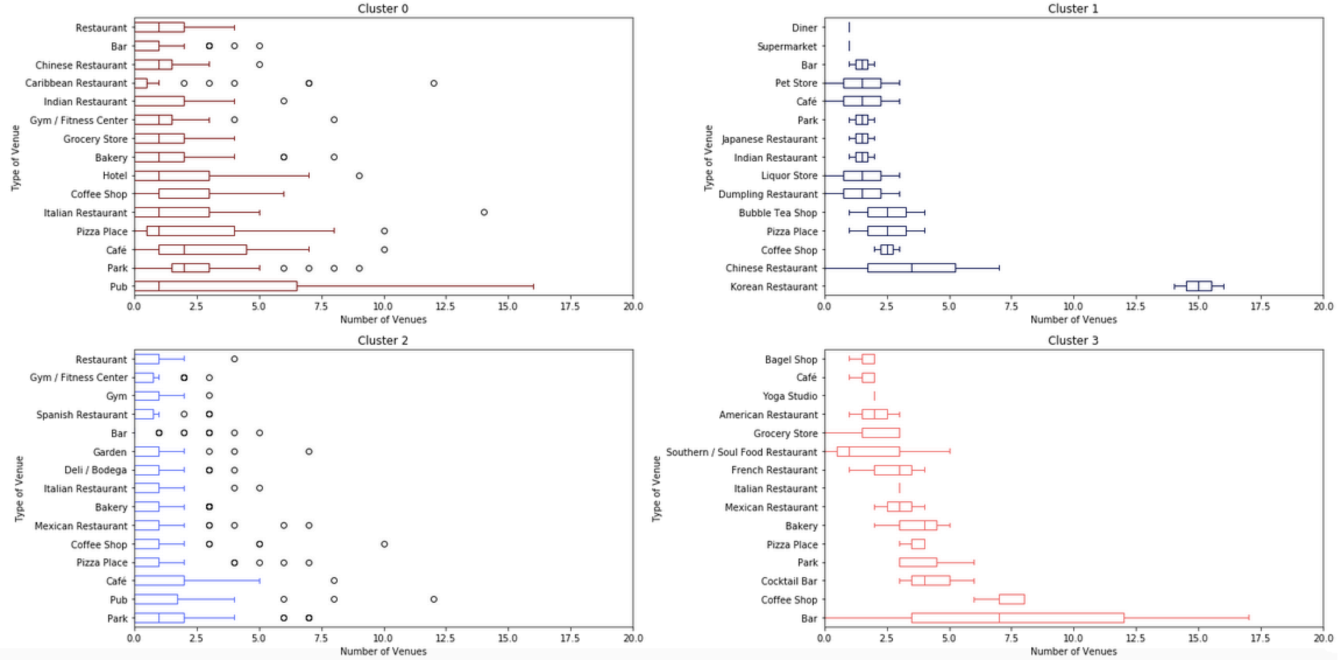


Methodology - K Means Clustering

- With the final dataset of size 86 (samples) x 30 (eigenvector decomposed features) we perform a sweep of K Means Clustering fits to determine the best amount of clusters.
- We find using the Elbow method of Inertia and the Silhouette Score, that the most suitable number of good clusters is 4 in this case.



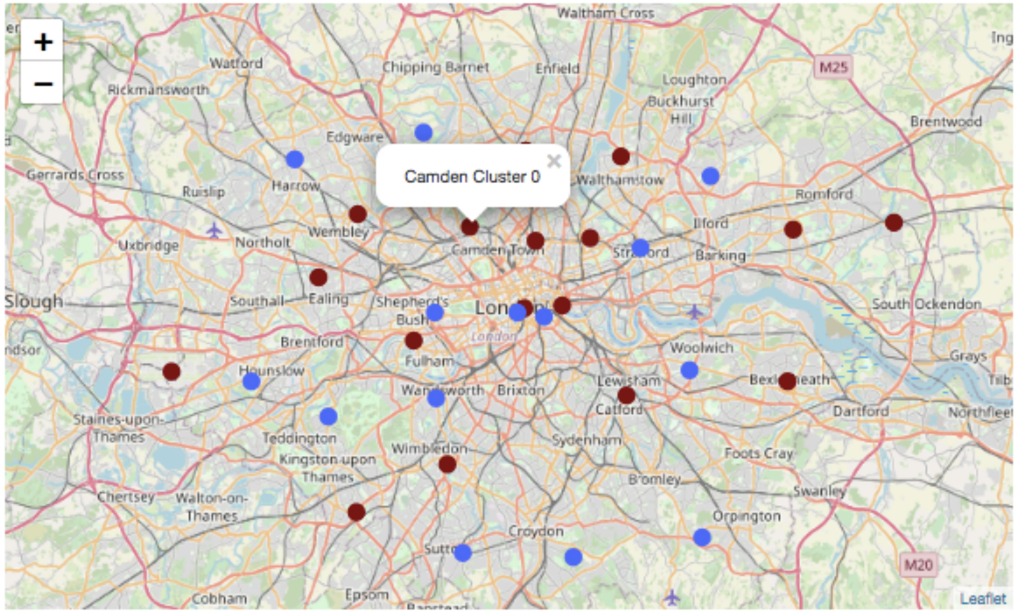
Results - Clustering

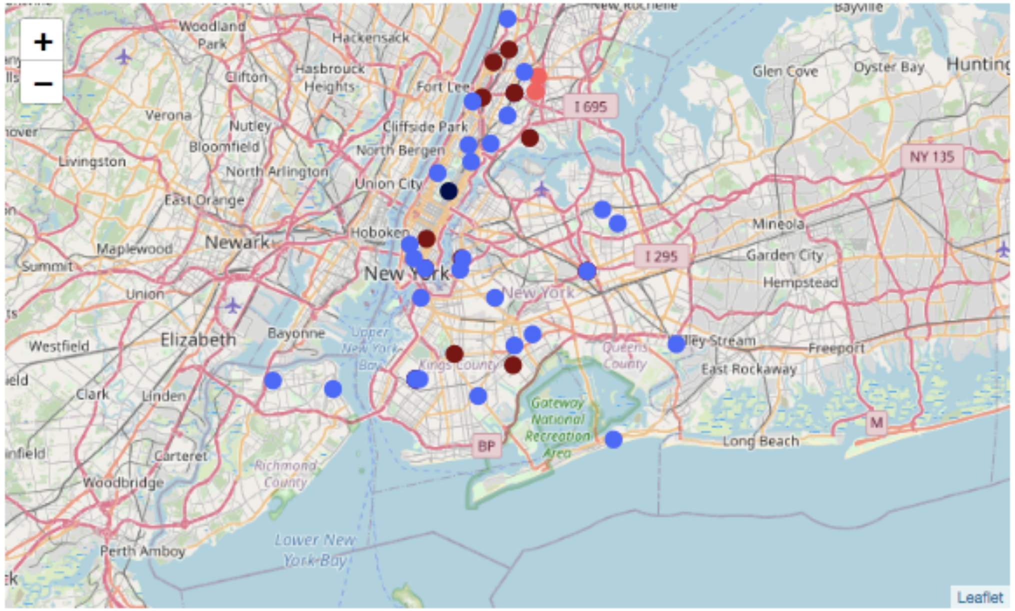


- Cluster 0 size is 35 out of 86
- Cluster 1 size is 2 out of 86
- Cluster 2 size is 46 out of 86
- Cluster 3 size is 3 out of 86

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.

Results - London vs New York

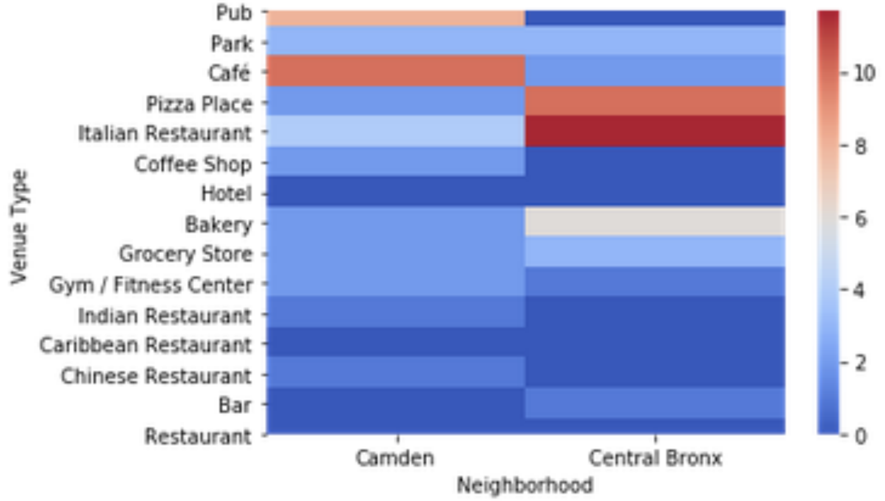




New York and London are both similarly made up of our higher density (cluster 0) and lower density (cluster 2) clusters but New York also has our outlying clusters suggesting more densely pocketed areas of ethnic variation. In general we can see that perhaps using Nomatim geocodes is not covering each cities area equally and a better method could be used.

Results - Secondment Recommendation

The closest similarity score to Camden in New York City Cluster 0 was Central Bronx and we can produce a heat map of the most common 15 venue types in this cluster.



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.