London vs Manhattan:

A social venues based comparative analysis

A. Introduction

A.1. Business Scenario

The owner of a successful restaurant chain in London is looking to expand in New York and wants to identify neighbourhoods similar to their current locations in London as they have been very successful in running their business there.

The goal is to identify similar neighbourhoods in Manhattan, New York to launch their next two restuarants. The similarity comparision of the neighbourhoods whould be in terms of the other business/social establishments that exist in the proposed neighbourhoods.

The chain has two categories of restuarants - the first one is in Covent Garden, City of London, London and , it is a fine dining restuarant and offers Indian cuisine, the second one is in Southall, in the Ealing borough , London and offers Indian fast/street food. Both cater to very different clientele and therefore it is important for the business to identify the right kind of neighbourhoods in Manhattan for each restaurant.

So the problem statement is as follows:

Is it possible to segment London and Manhattan into neighbourhood clusters and then identify clusters that extend across both cities thereby implying similarity in the neighbourhoods in London and Manhattan?

Do the neighbourhoods of the two restaurants in London fall in different clusters? If yes, for each cluster are we able to identify similar neighbourhoods in Manhattan to expand the restaurant business?

A.2. Data Description

The data that will be used for this analysis will include four square data to analyze the different neighbourhoods in New York and London and combine them into one single dataset of neighbourhoods , then segment this dataset into different clusters and identify which clusters Southall and Covent Garden fall into respectively and are there neighbourhoods in New York which fall in the same clusters. If there are, then these could potentially be good locations for the launch of the new restuarants.

We will start with identifying the names of all the boroughs of London by webscraping the following page : https://en.wikipedia.org/wiki/List_of_London_boroughs and the neighbourhoods in Manhattan from a pre-existing dataset used earlier in this course. After that we use this list to gather geospatial data like latitude and longitude for each borough/neighbourhood in each city which is used for gathering social venues information from Four Square through an API request.

The main data to be used for this analysis is Four Square data which typically includes the following fields for each location in a given Neighbourhood: Name ,Neighborhood Latitude, Neighborhood Longitude, Venue name, Venue Latitude, Venue Longitude, Venue Category, for top 100 venues within a 500 metre radius.

After identifying the social venues data we will use the K Means algorithm to cluster the data into similar groups and see if we can define the characteristics if each group of neighbourhoods , and then make final recommendations on the locations of the restaurants to be launched in Manhattan.

B. Methodology

As mentioned above the names all the boroughs of London were identified by webscraping the following page : https://en.wikipedia.org/wiki/List_of_London_boroughs and the neighbourhoods in Manhattan from a pre-existing dataset used earlier in this course.

After geospatial data like latitude and longitude for each borough/neighbourhood in each city was gathered using geocoder and the result was as follows:

	Borough	Neighborhood	Latitude	Longitude
6	Manhattan	Marble Hill	40.876551	-73.910660
100	Manhattan	Chinatown	40.715618	-73.994279
101	Manhattan	Washington Heights	40.851903	-73.936900
102	Manhattan	Inwood	40.867684	-73.921210
103	Manhattan	Hamilton Heights	40.823604	-73.949688
104	Manhattan	Manhattanville	40.816934	-73.957385
105	Manhattan	Central Harlem	40.815976	-73.943211
106	Manhattan	East Harlem	40.792249	-73.944182
107	Manhattan	Upper East Side	40.775639	-73.960508
108	Manhattan	Yorkville	40.775930	-73.947118
109	Manhattan	Lenox Hill	40.768113	-73.958860
110	Manhattan	Roosevelt Island	40.762160	-73.949168
111	Manhattan	Upper West Side	40.787658	-73.977059
112	Manhattan	Lincoln Square	40.773529	-73.985338
113	Manhattan	Clinton	40.759101	-73.996119
114	Manhattan	Midtown	40.754691	-73.981669
115	Manhattan	Murray Hill	40.748303	-73.978332
116	Manhattan	Chelsea	40.744035	-74.003116
117	Manhattan	Greenwich Village	40.726933	-73.999914
118	Manhattan	East Village	40.727847	-73.982226
119	Manhattan	Lower East Side	40.717807	-73.980890
120	Manhattan	Tribeca	40.721522	-74.010683
121	Manhattan	Little Italy	40.719324	-73.997305
122	Manhattan	Soho	40.722184	-74.000657
123	Manhattan	West Village	40.734434	-74.006180
124	Manhattan	Manhattan Valley	40.797307	-73.964286
125	Manhattan	Morningside Heights	40.808000	-73.963896
126	Manhattan	Gramercy	40.737210	-73.981376
127	Manhattan	Battery Park City	40.711932	-74.016869
128	Manhattan	Financial District	40.707107	-74.010665
247	Manhattan	Carnegie Hill	40.782683	-73.953256
248	Manhattan	Noho	40.723259	-73.988434
249	Manhattan	Civic Center	40.715229	-74.005415
250	Manhattan	Midtown South	40.748510	-73.988713
271	Manhattan	Sutton Place	40.760280	-73.963556
273	Manhattan	Turtle Bay	40.752042	-73.967708
274	Manhattan	Tudor City	40.746917	-73.971219
275	Manhattan	Stuyvesant Town	40.731000	-73.974052
276	Manhattan	Flatiron	40.739673	-73.990947
301	Manhattan	Hudson Yards	40.756658	-74.000111
0	London	Barking and Dagenham	51.560228	0.171961
1	London	Barnet	51.627300	-0.253760
2	London	Bexley	51.452078	0.069931
3	London	Brent	51.609783	-0.194672
4	London	Bromley	51.601511	-0.066365
5	London	Camden	51.591180	-0.165040
6	London	Croydon	51.593470	-0.083380
7	London	Ealing	51.508383	-0.305200
8	London	Enfield	51.540024	-0.077502
9	London	Greenwich	51.477890	-0.013340
10	London	Hackney	51.531820	-0.061780
11	London	Hammersmith and Fulham	51.482600	-0.212880
12	London	Haringey	51.589270	-0.106405
13	London	Harrow	51.513180	-0.106980
14	London	Havering	51.544610	-0.144260
15	London	Hillingdon	51.484230	-0.096477
16	London	Hounslow	51.467701	-0.361718
17	London	Islington	51.534380	-0.108940
18	London	Kensington and Chelsea	51.522660	-0.207930
19	London	Kingston upon Thames	51.409008	-0.303598
20	London	Lambeth	51.494471	-0.120066
21	London	Lewisham	51.465280	-0.013210
22	London	Merton	51.544520	-0.166860
23	London	Newham	51.519937	0.055882
24	London	Redbridge	51.475773	-0.080698
25	London	Richmond upon Thames	51.480270	-0.237540
26	London	Southwark	51.505734	-0.100002
27	London	Sutton	51.512243	-0.053659
28	London	Tower Hamlets	51.499990	-0.010450
29	London	Waltham Forest	51.581765	-0.276968
30	London	Wandsworth	51.467826	-0.144992
31	London	Westminster	51.628249	0.012986
32	London	City of London	51.508530	-0.125740

London has 32 boroughs and the area of the City of London. Manhattan is itself a borough of New York and has 40 Neighbourhoods. To avoid confusion we will refer to both London boroughs and Manhattan Neighbourhoods as Neighbourhoods going forward.

Also even though the average area of Manhattan neighbourhoods is much smaller than London , the concentration of venues in Manhattan neighbourhoods is high, therefore in terms of number of social venues within a neighbourhood, London neighbourhoods and Manhattan neighbourhoods can be considered comparable.

Here is a spatial representation of the identified neighbourhoods using a folium map of London and Manhattan:

LONDON:

MANHATTAN:

In all 73 Neighbourhoods were analyzed across London and Manhattan and the number of venues identified across these 73 neighbourhoods was 4392. Here is a snapshot of the data returned from the Four Square API:

```
  Neighborhood  Neighborhood Latitude  Neighborhood Longitude         Venue  \
0  Marble Hill              40.876551               -73.91066   Bikram Yoga
1  Marble Hill              40.876551               -73.91066      Arturo's
2  Marble Hill              40.876551               -73.91066  Tibbett Diner
3  Marble Hill              40.876551               -73.91066       Dunkin'
4  Marble Hill              40.876551               -73.91066      Starbucks

   Venue Latitude  Venue Longitude Venue Category
0       40.876844       -73.906204   Yoga Studio
1       40.874412       -73.910271   Pizza Place
2       40.880404       -73.908937         Diner
3       40.877136       -73.906666    Donut Shop
4       40.877531       -73.905582   Coffee Shop
There are 371 uniques categories.
```

As you can see a total of 371 unique venue categories were identified and the neighbourhoods where number of venues identified was less than 10 were excluded from the analysis. Here is a snapshot of the data grouped by neighbourhood to find out which neighbourhoods had less than 10 venues retrieved for them.

Out[22]:

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Barking and Dagenham	1	1	1	1	1	1
Barnet	1	1	1	1	1	1
Battery Park City	81	81	81	81	81	81
Bexley	4	4	4	4	4	4
Brent	4	4	4	4	4	4
Bromley	10	10	10	10	10	10
Camden	25	25	25	25	25	25
Carnegie Hill	92	92	92	92	92	92
Central Harlem	43	43	43	43	43	43
Chelsea	100	100	100	100	100	100

Certain venue categories had to be modified to accommodate differences in the way certain establishments are named in UK vs USA. Eg a bar in USA is typically referred to as a pub in UK, a coffee shop in USA is likely to be called a café in UK etc.Also venues like beer bar, wine bar, cocktail bar etc were all clubbed into one single category named bar. This exercise reduced the number of unique categories to 364. There were other similar anomalies in the data eg Gym, Gym/Fitness Centre/Gym Pool/ Gymanstics Gym – but since the frequency of these venues was low, so these were not corrected.

```
In [20]: #Cleaning the Data : Venue Categories data - replacing the values as a pub in London is the same as a bar in Manhattan, and a Cafe is the same as a Co

lon_ny_venues.loc[lon_ny_venues['Venue Category'] == 'Pub', 'Venue Category'] = 'Bar'
lon_ny_venues.loc[lon_ny_venues['Venue Category'] == 'Café', 'Venue Category'] = 'Coffee Shop'
lon_ny_venues.loc[lon_ny_venues['Venue Category'] == 'Cocktail Bar', 'Venue Category'] = 'Bar'
lon_ny_venues.loc[lon_ny_venues['Venue Category'] == 'Whisky Bar', 'Venue Category'] = 'Bar'
lon_ny_venues.loc[lon_ny_venues['Venue Category'] == 'Wine Bar', 'Venue Category'] = 'Bar'
lon_ny_venues.loc[lon_ny_venues['Venue Category'] == 'Beer Bar', 'Venue Category'] = 'Bar'
lon_ny_venues.loc[lon_ny_venues['Venue Category'] == 'Cafeteria', 'Venue Category'] = 'Coffee Shop'
```

```
In [21]: print('There are {} uniques categories.'.format(len(lon_ny_venues['Venue Category'].unique())))

There are 364 uniques categories.
```

The features used for K means clustering were the venue categories.To create the dataset for K means the one hot encoding technique was applied on the dataset of venues as follows:

Out[25]:

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auditorium	Australian Restaurant	Aus Restai
0	Marble Hill	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	Marble Hill	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	Marble Hill	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	Marble Hill	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	Marble Hill	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

And then the data was grouped by neighbourhoods and the mean values for each category calculated.

Out[27]:

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auditorium	Australian Restaurant	Re:
0	Battery Park City	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.00000	0.000000	0.000000	0.0	0.0	0.012346	0.012346	0.0	
1	Bromley	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.00000	0.000000	0.000000	0.0	0.0	0.000000	0.000000	0.0	
2	Camden	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.00000	0.000000	0.000000	0.0	0.0	0.000000	0.000000	0.0	
3	Carnegie Hill	0.0	0.0	0.0	0.000000	0.010870	0.0	0.0	0.01087	0.000000	0.021739	0.0	0.0	0.000000	0.000000	0.0	
4	Central Harlem	0.0	0.0	0.0	0.046512	0.046512	0.0	0.0	0.00000	0.023256	0.000000	0.0	0.0	0.000000	0.000000	0.0	

Each neighbourhood was sorted in terms of the most frequently occurring venue categories for us to be able to visually define the characteristics of the clusters later.

Out[50]:

	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	Battery Park City	Park	Coffee Shop	Hotel	Clothing Store	Gym	Playground	Memorial Site	Boat or Ferry	Burger Joint	Pizza Place
1	Bromley	Bar	Hostel	Supermarket	Coffee Shop	Train Station	Mediterranean Restaurant	Soccer Stadium	Sporting Goods Shop	Outdoor Sculpture	Organic Grocery
2	Camden	Coffee Shop	Greek Restaurant	Pizza Place	Bus Stop	Grocery Store	Chinese Restaurant	French Restaurant	Health & Beauty Service	Shoe Store	Movie Theater
3	Carnegie Hill	Coffee Shop	Bar	Wine Shop	Yoga Studio	Gym	French Restaurant	Bookstore	Pizza Place	Gym / Fitness Center	Bakery
4	Central Harlem	Bar	Seafood Restaurant	African Restaurant	American Restaurant	Gym / Fitness Center	Chinese Restaurant	Coffee Shop	French Restaurant	Public Art	Bookstore

```
In [51]: neighborhoods_venues_sorted.shape

Out[51]: (64, 11)
```

K Means machine learning classification algorithm was then run on the day to classify it into 5 clusters which resulted in the following 5 clusters:

In [55]:
```
print(Cluster0.shape)
Cluster0
```

(13, 12)

Out[55]:

	Borough	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
118	Manhattan	East Village	Bar	Coffee Shop	Mexican Restaurant	Speakeasy	Vegetarian / Vegan Restaurant	Vietnamese Restaurant	Italian Restaurant	Pizza Place	Korean Restaurant	Bagel Shop
275	Manhattan	Stuyvesant Town	Bar	Boat or Ferry	Park	Coffee Shop	Harbor / Marina	Gas Station	Heliport	Fountain	Farmers Market	Baseball Field
4	London	Bromley	Bar	Hostel	Supermarket	Coffee Shop	Train Station	Mediterranean Restaurant	Soccer Stadium	Sporting Goods Shop	Outdoor Sculpture	Organic Grocery
7	London	Ealing	Bar	Coffee Shop	Italian Restaurant	Park	Clothing Store	Pizza Place	Bookstore	Performing Arts Venue	Hotel	Pharmacy
8	London	Enfield	Bar	Coffee Shop	Middle Eastern Restaurant	Restaurant	Lounge	Park	Bookstore	Burger Joint	Canal	Gay Bar
9	London	Greenwich	Bar	Coffee Shop	Gym / Fitness Center	Hotel	French Restaurant	Pizza Place	Burger Joint	Dance Studio	Thrift / Vintage Store	Theater

In [57]:
```
print(Cluster1.shape)
Cluster1
```

(36, 12)

Out[57]:

	Borough	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
6	Manhattan	Marble Hill	Gym	Sandwich Place	Yoga Studio	Tennis Stadium	Kids Store	Ice Cream Shop	Pharmacy	Pizza Place	Donut Shop	Diner
100	Manhattan	Chinatown	Chinese Restaurant	Bar	Bakery	American Restaurant	Hotpot Restaurant	Ice Cream Shop	Dessert Shop	Salon / Barbershop	Optical Shop	Spa
105	Manhattan	Central Harlem	Bar	Seafood Restaurant	African Restaurant	American Restaurant	Gym / Fitness Center	Chinese Restaurant	Coffee Shop	French Restaurant	Public Art	Bookstore
107	Manhattan	Upper East Side	Italian Restaurant	Exhibit	Gym / Fitness Center	Bakery	Coffee Shop	American Restaurant	Juice Bar	Hotel	Yoga Studio	Cosmetics Shop
108	Manhattan	Yorkville	Italian Restaurant	Bar	Coffee Shop	Gym	Deli / Bodega	Japanese Restaurant	Wine Shop	Sushi Restaurant	Mexican Restaurant	Diner
109	Manhattan	Lenox Hill	Coffee Shop	Italian Restaurant	Bar	Pizza Place	Sushi Restaurant	Gym / Fitness Center	Gym	Burger Joint	Taco Place	Cycle Studio

In [59]:
```
print(Cluster2.shape)
Cluster2
```

(1, 12)

Out[59]:

	Borough	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
27	London	Sutton	Grocery Store	Park	Pool	Fried Chicken Joint	Coffee Shop	Bar	Gym / Fitness Center	Indian Restaurant	Hotel	Street Art

In [61]:
```
print(Cluster3.shape)
Cluster3
```

(2, 12)

Out[61]:

	Borough	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
24	London	Redbridge	Bus Stop	Coffee Shop	Grocery Store	Hotel	Park	Chinese Restaurant	Theater	Art Gallery	Asian Restaurant	Gay Bar
28	London	Tower Hamlets	Bus Stop	Sandwich Place	Grocery Store	Italian Restaurant	Park	Sushi Restaurant	Tennis Court	Bar	Hotel	Gym / Fitness Center

In [63]:
```
print(Cluster4.shape)
Cluster4
```

```
(12, 12)
```

Out[63]:

	Borough	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
101	Manhattan	Washington Heights	Coffee Shop	Bakery	Grocery Store	Mobile Phone Shop	Gym	Liquor Store	Pizza Place	Sandwich Place	Deli / Bodega	Bar
102	Manhattan	Inwood	Coffee Shop	Mexican Restaurant	Restaurant	Bar	Lounge	Bakery	Pizza Place	Deli / Bodega	Park	Chinese Restaurant
103	Manhattan	Hamilton Heights	Coffee Shop	Pizza Place	Bar	Deli / Bodega	Mexican Restaurant	Bakery	Indian Restaurant	Sandwich Place	Caribbean Restaurant	School
104	Manhattan	Manhattanville	Coffee Shop	Sushi Restaurant	Bar	Deli / Bodega	Mexican Restaurant	Italian Restaurant	Liquor Store	Chinese Restaurant	Climbing Gym	Seafood Restaurant

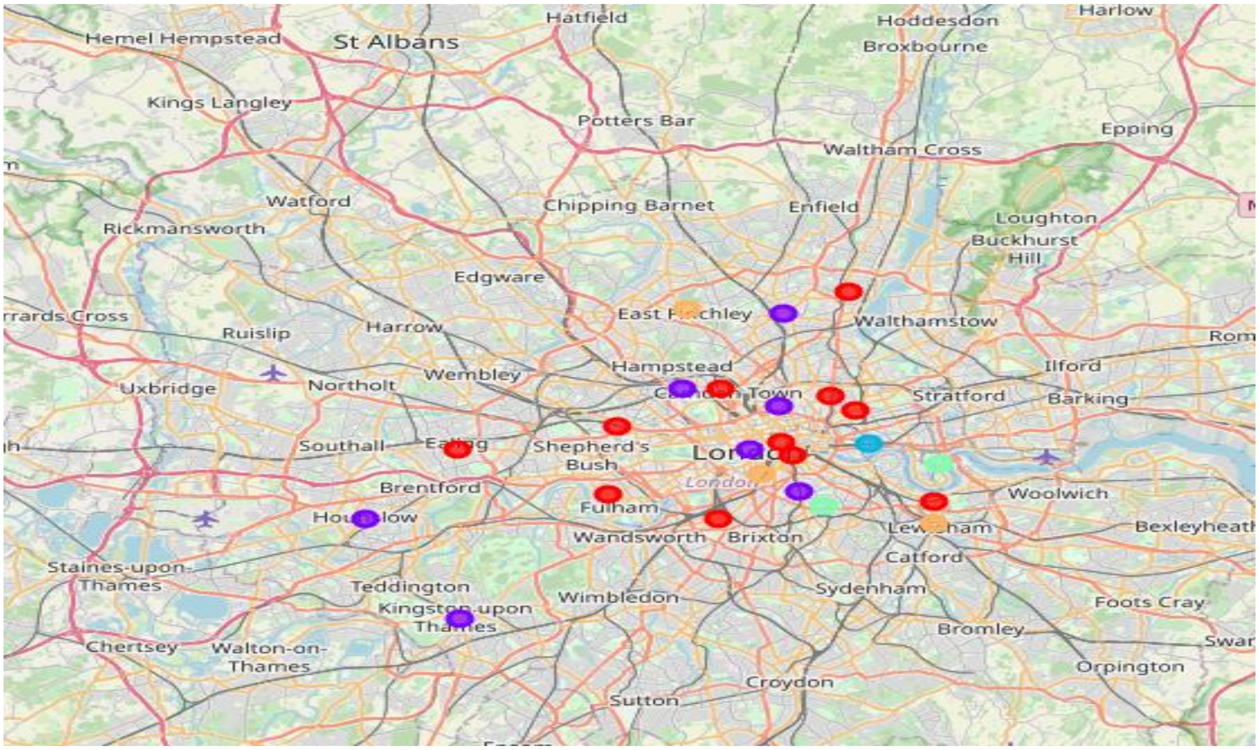
C. Results

The K Means Algorithm resulted in 5 clusters which are described as below:

Cluster Name	Details	Dimensions
Cluster 0	These neighbourhoods are dominated by Bars and Coffee Shops , with the top two most frequent venue categories being bars and coffee shops in 10 out of 13 neighbourhoods. The remaining top categories mostly included restaurants. These appear to be vibrant social venues.	13 Neighbourhoods : 2 Manhattan and 11 London
Cluster 1	These neighbourhoods are a mix of restaurants, coffee shops and bars alongwith grocery stores, gyms, parks as the other frequent categories so they appear to be areas that cater to amenities for local residents as well as socialising venues.	36 Neighbourhoods : 29 Manhattan and 7 London
Cluster 2	Just a single neighbourhood – so we will ignore this cluster	1 Neighbourhood : 0 Manhattan and 1 London
Cluster 3	Just two neighbourhoods, with the most frequent category being a bus stop – so we will ignore this cluster	2 Neighbourhoods : 0 Manhattan and 2 London
Cluster 4	In this cluster 10 out of 12 neighbourhoods had Coffee Shops as the most frequently occurring social venue	12 Neighbourhoods : 9 Manhattan and 3 London

As you can see the algorithm successfully classified the combined set of neighbourhoods across London and Manhattan into clusters of similar neighbourhoods. These are visually represented as below:

LONDON:



MANHATTAN:

D. Discussion

Based on the above results of the K means clustering we can now revisit the questions we had asked ourselves at the beginning of this project and see if we found any answers.

The first question was whether it is t possible to segment London and Manhattan into neighbourhood clusters and then identify clusters that extend across both cities thereby implying similarity in the neighbourhoods in London and Manhattan? The conclusion is that yes this possible. The algorithm was able to classify the combined dataset of London and Manhattan neighbourhoods into 5 Clusters. Two of the five clusters had very few datapoints so these clusters were disregarded. Three clusters extend across both London and Manhattan with a good mix of neighbourhoods from both cities.

The second question was that do the neighbourhoods of the two restaurants in London fall in different clusters? If yes, for each cluster are we able to identify similar neighbourhoods in Manhattan to expand the restaurant business? Now looking at our current restaurants and their locations in London, we find that Ealing and City of London were the two neighbourhoods that we were analyzing and Ealing falls in Cluster 0 and City of London falls in Cluster 1 and we are able to identify similar neighbourhoods in Manhattan. Neighbourhoods in Manhattan similar to Ealing where we could consider launching the Indian street food restaurant are : Stuyvesant Town and East Village and the neighbourhoods similar to City of London where we could consider launching the Indian fine dining restaurant are : Battery Park City, Financial District, Carnegie Hill, Noho, Civic Center , Midtown South, Sutton Place, Turtle Bay, Flatiron ,Hudson Yards

However one of the limitations of this approach due to time constraints and API request limits has been that we have only requested data for the top 100 venues in a 500 metre radius. Since London is less concentrated than Manhattan therefore this study needs to be redone with a larger radius for London neighbourhoods to get more accurate results

Another point to bear in mind is that we have taken the latitude and longitude coordinates of the neighbourhoods as the starting point to define the area around it. These coordinates may not always be the best points to identify the social venues around it as venues may not be distributed evenly around these coordinates .

Further analysis can be carried out by exploring the venues list differently based on the above considerations to fine tune this study.

Additionally multiple other factors like real estate cost, population density , presence of other Indian restaurants and their ratings can be used to further drill down into the identified neighbourhoods and provide a more precise list of possible options to consider.

F. Conclusion

This sort of inter city analysis can be extended to many more areas for example people looking to relocate to a different city and wanting to identify where to look for homes based on their current localities, people looking to compare different cities lets say for retirement purposes or businesses looking to launch into new geographies. A lot more data will need to be overlaid ob the Four Square data, eg population, real estate cost, cost of living etc to make these kind of studies more useful for the stakeholders.