Coursera - IBM Data Science Capstone Project

Clustering Similar Areas in New York and London

Matthew Yeung - 4 January 2019

Introduction

Hustle and bustle New York City is famous for crowded and fully packed tall buildings. Shops, restaurants, malls, business centres and cafes are all around each corners. It is one of the busiest cities in the world with no doubt.

London is a city where people are all diversified. People come from different countries and cities to work and live there. With its multicultural diversities, the city are also popular to live at.

One may think that living in both the cities will have a similar livelihood and you can easily adapt to the environment if you moved from New York to London. Others may argue that as the difference in cultural diversities, it is hard to get used to the new environment.

A large number of expatriates migrated to London every single day. As a business and financial centre, some of them move there to work. Others may move there to study their master degrees and some may just move there to live. However, London is not a small city. There are 119 areas to choose for. In order to adapt to the environment easily, people may want to choose an areas that are similar to the place that they used to live at. This is a tough work as an expat as they are not usually familiar with the new place and it will be hard to choose a place to start for.

Machine Learning can solve the above problem. First of all, we assume that an employee is relocated from New York City to London. He would like to find a similar place to live in London that can provide him with same level of enjoyment of living, i.e. He can find the restaurants as easy as he does in New York City.



Description of the data

In order to compare the areas of the 2 different cities, we need to have both the location information of both places. We are going to use a location provider called *Foursquare*. The location provider can provide the location information including the shops, restaurants, cafes, leisure centres and a lot of different kinds of shops that around the areas. It will be a comprehensive comparison for the areas in the researching areas.

We need the location code including the latitude and longitude for the location provider to function. So we can easily find the relevant information online.

By extracting the information below:

New York: https://geo.nyu.edu/catalog/nyu_2451_34572

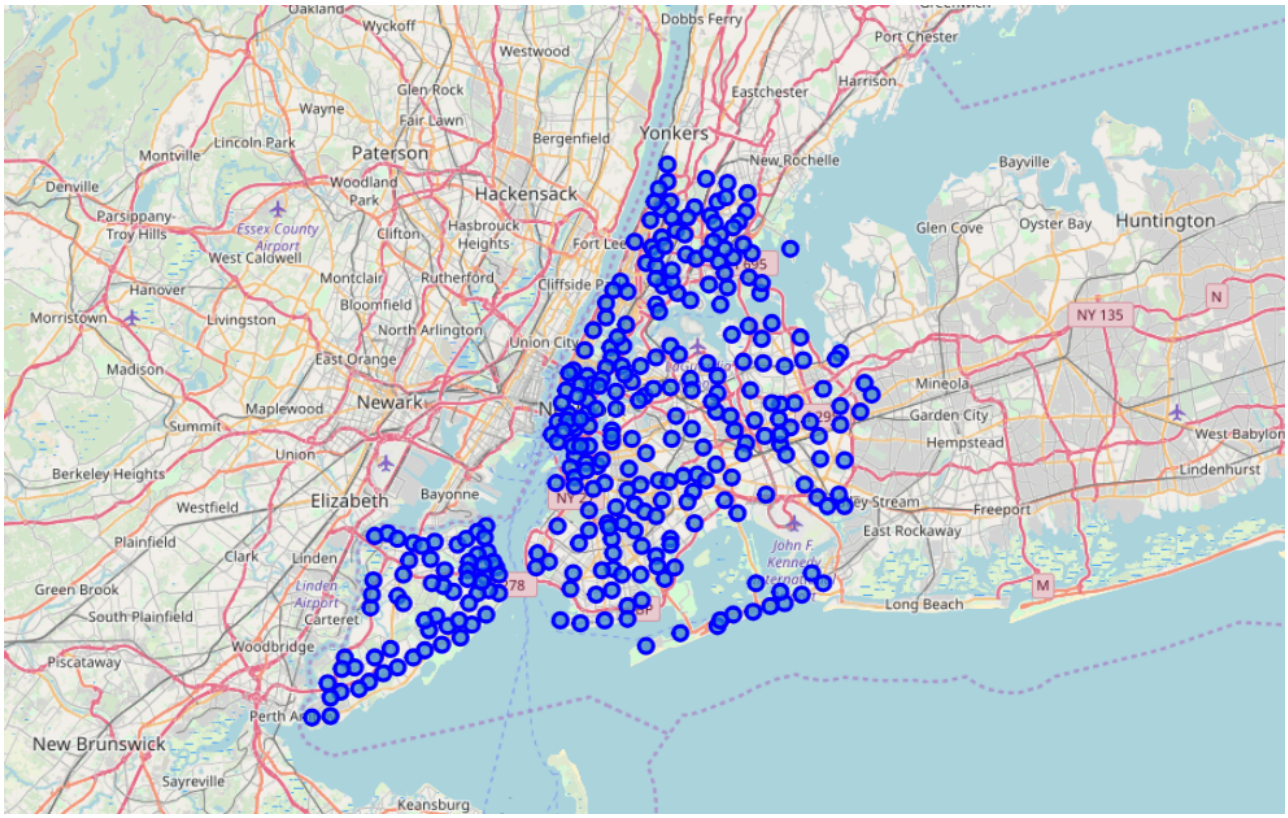
London: https://www.maps.thehunthouse.com/Streets/London_Postal_District_and_Area_Name_finding_aid.htm

As the website cannot directly give out the required information in a table format, we need to further process the data by merging the postcode and latitude and longitude. In this process, Panda DataFrame will be used. The cleared tables will be used for finding the location information by *Foursquare*. I am going to use 'Explore given location' query in the Foursquare API to explore the each areas and find out the habitat of the areas. The output from it will be a JSON file including the Name of locations, ID of the locations, location info and category of the locations. In each areas, I am going to choose the top 100 findings for comparison. Next, the location data will be put in to the K-Mean Algorithm for grouping. K-Mean Algorithm is an unsupervised machine learning algorithm to group the similar feature data. All the data from New York and London will be put into the same

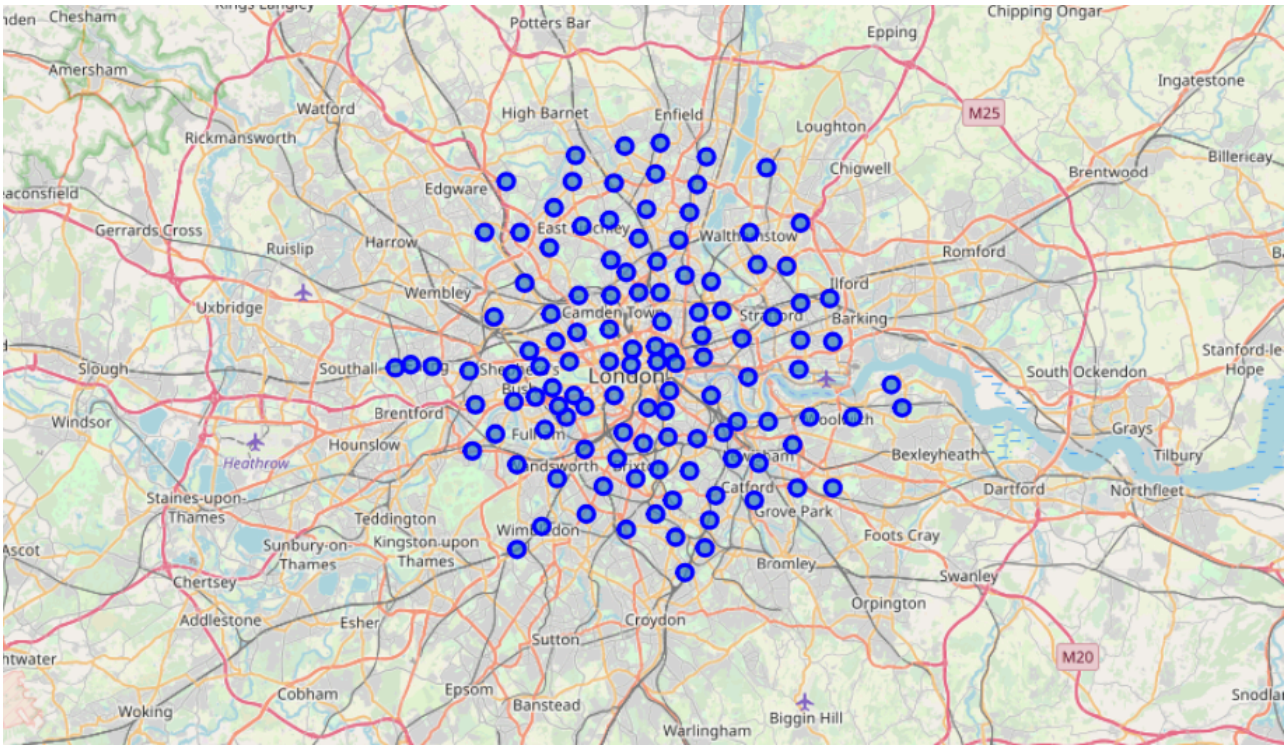
algorithm and the output will be the areas which are grouped by their features. After all, the result will indicate the similarity of New York city and London. For the ease of observation, *Folium* which is a map view visualisation Python library will be used.

Methodology

1. Download and extracting data:
In my research, I acquired both New York City Neighbourhood and London Neighbourhood data from the web which are open resources and easily assessable. New York data if from the website of University of New York. The site provide the JSON file containing all Boroughs and Neighbourhoods and their corresponding geographical coordinations. Further extract the data from the JSON file was needed. For London Data, as I cannot find a file that containing both the Neighbourhood name and their geographical coordinations, I need to find them separately and combine the tables. One of the website containing the postcode and Neighbourhood name and another containing the geographical coordinations and postcode. After merging both the tables, the final table was used in the research.



Blue spots show the Neighbourhoods of New York City.



Blue spots show the Neighbourhoods of London

2. Handle the duplicated Neighbourhood problem:
By looking at the final table that combining the New York City and London, there were some Neighbourhoods share the same names. In fact, they are not the same place as some of the Neighbourhood in New York have the same name as that in London. This would cause some problem in the later stage of the research as the Neighbourhood will be used as index. So, in order to determine them, I replace the name of the Neighbourhood column with a name that including the borough or postcode which is the first column. So the Neighbourhood Name would be something like 'BronxWakefield' which 'Bronx' is the Borough name and 'Wakefield' is the Neighbourhood name.

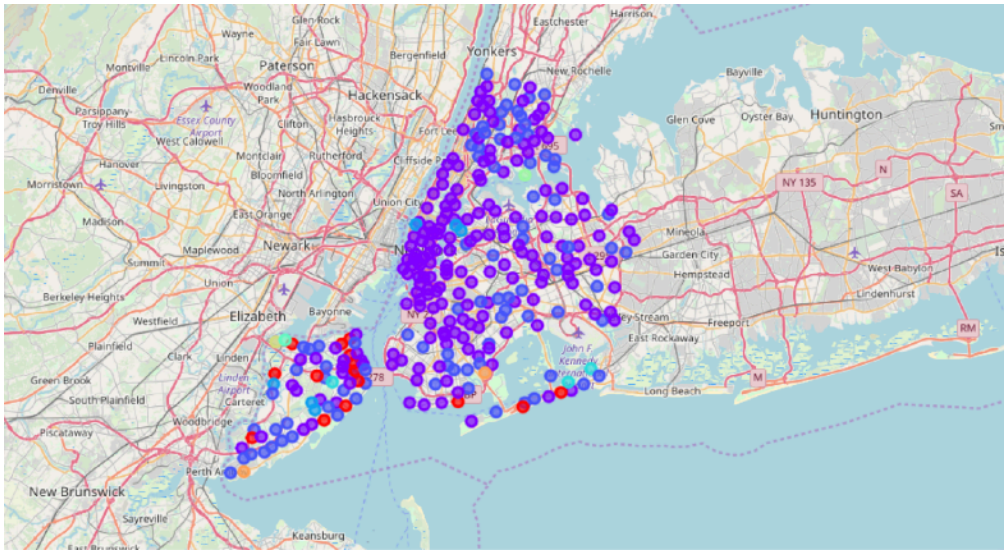
	borough_postcode	Neighbourhood	Latitude	Longitude
0	Bronx	BronxWakefield	40.894705	-73.847201
1	Bronx	BronxCo-op City	40.874294	-73.829939
2	Bronx	BronxEastchester	40.887556	-73.827806
3	Bronx	BronxFieldston	40.895437	-73.905643
4	Bronx	BronxRiverdale	40.890834	-73.912585

Extract of the table

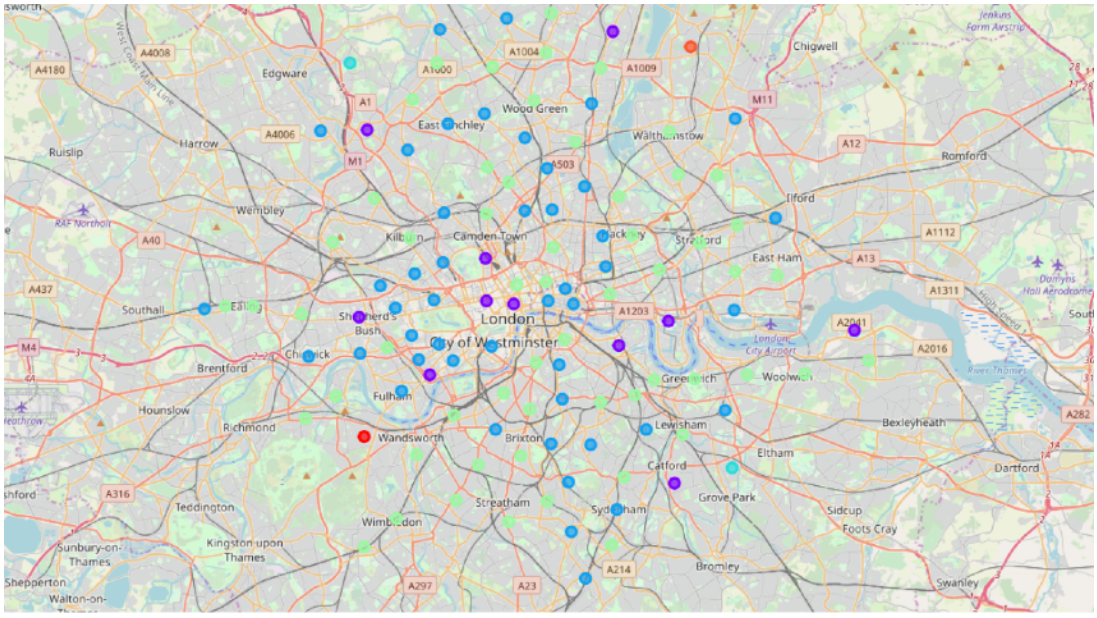
3. Explore Neighbourhoods
After handled the duplicated problem, the table are pass to the Foursquare API to explore the details about the Neighbourhoods. First, a function is created to repeatedly finding the top 100 venues to all the neighbourhoods in the table. The 'explore' function of the Foursquare API is used. The function require both the latitude and longitude of the places. 100 venues is a suitable amount as the parameter since some of the venues do not have too many information in the API. The result shows there are 460 uniques categories found. The result is grouped by a table showing the all the categories of the findings.

4. Sort out the Most Common Venues
Some of the neighbourhoods were found more venues but some of them found less. To have a better comparison, the top ten most common venues of each neighbourhood are sorted out. By taking the mean on the frequency of occurrence of each category, the top ten most common venues can be found. A table of neighbourhoods showing the top ten most common venues was created and these data were plugged in to the algorithm later on.

5. Cluster Neighbourhoods
K-Mean is an unsupervised machine learning algorithm which can cluster a large number of object into small clusters by their features automatically and K-Mean algorithm is from the 'sklearn.cluster ' library. In this research, K-Mean clustering was used as the research topic relating to find the similarities between the areas which the areas features can be used as the data in the algorithm. Before putting the data in the algorithm, we need to drop the 'Neighbourhood' columns as the algorithm need to be consisting only numbers in the data. I set 10 clusters as the output. The output is visualised by map views.



New York Neighbourhoods clustered with different colours spot

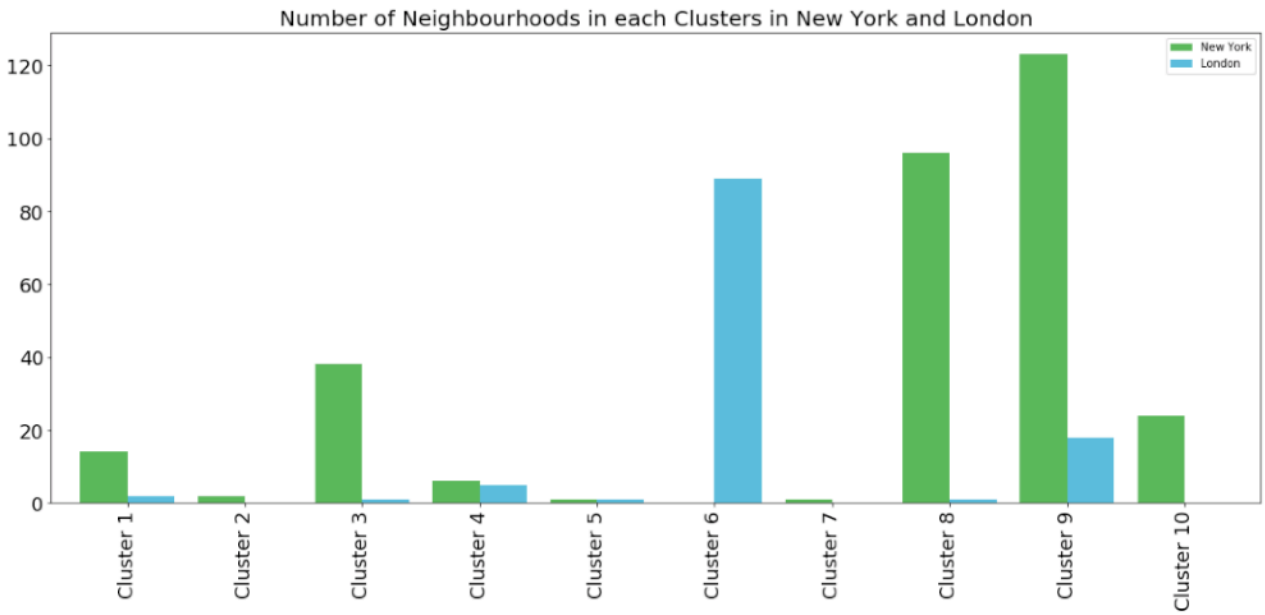


New York Neighbourhoods clustered with different colours spot

6. Statistics of the clustering

There are 10 clusters produced in the output. The Neighbourhoods in each clusters were displayed. Furthermore, a table showing the number of neighbourhoods in each clusters was produced. Finally, a bar chart showing the corresponding data was created.

	New York	London
Cluster 1	14.0	2.0
Cluster 2	2.0	0.0
Cluster 3	38.0	1.0
Cluster 4	6.0	5.0
Cluster 5	1.0	1.0
Cluster 6	0.0	89.0
Cluster 7	1.0	0.0
Cluster 8	96.0	1.0
Cluster 9	123.0	18.0
Cluster 10	24.0	0.0



Result

From the above table and bar chart, I can observe:

1. Cluster 1, 3, 8, 9 and 10 are dominating in New York (96.6%) . There are only a few neighbourhoods from Cluster 2, 4, 5, 6 and 7 (3.5%).
2. Cluster 6, 9 are dominating in London (91.5%) . There are only a few neighbourhoods from Cluster 1 - 5, 7 and 8 (8.5%).
3. Generally, when number of New York neighbourhoods is high in one cluster, number of neighbourhood will be low in London in that cluster. They tends to be negatively correlated. Cluster 4 and 5 are the exceptions. Looking at cluster 4 and 5, they have almost same number of neighbourhoods but both are less than 6 which is a very small proportion to the total neighbourhoods.
4. Cluster 1, 2, 3, 7, 8 and 10 consist solely or almost solely neighbourhoods from New York. But Cluster 6 consist solely neighbourhoods from London.
5. Looking at cluster 9, it is the highest neighbourhood numbers from New York (40%). For London, it is about 17% of the total neighbourhoods. For both Cities, this cluster consist of a substantial proportion of neighbourhoods.

Back to our research problem, if you are moving from New York City to London, the place best for you to live can be determined. By the above observation, we can make the following arrangement:

1. If you are originally living in Cluster 1-3, 7, 8 or 10 in New York, the above method to determine the best place to live is not statistically valid.
2. If you are from Cluster 4 and 5 in New York, by the analysis made, you can try to live in the corresponding cluster in London. However, as the neighbourhoods in the cluster is not enough, we do not confident that it is similar to the neighbourhood that you live in New York.
3. If you are from Cluster 9 in New York, you will probably find the same cluster in London to live as the analysis showed that it is statistically significant to have similar features to your old place in New York.

Discussion

In the analysis, I found that the number of neighbourhoods in New York and London in each cluster are negatively correlated. This will be a problem for the research as we are going to find the similar neighbourhoods in both cities. There are some reasons for that.

1. I chose only 10 venues in each Neighbourhoods for investigation. In fact, the output of the 10 venues were too diversified that the algorithm can not do the job to cluster them by the features since most of the features are different.

2. The venues name is also too diversified. By looking at the output table, there are too many different kinds of restaurants which may affect the efficiency of the algorithm. For example an ‘asian restaurant', ‘japanese restaurant’ and ‘Japanese cousins restaurant’ are in 3 different groups. In this way, the algorithm may treat the same type of feature by 3 different types. Also, the venues name may vary according to the countries that we explored.
3. K-Mean algorithm are random algorithm. The algorithm initiated a random seed to start the iteration of calculation. A different seed may result in a different output of the clusters. This is due to the different centroid of the clusters created by the algorithm. The consistency of the outcome is also a cons for the K-Mean algorithm.
4. The information of the venues given by the location provider (Foursquare API) based on the database of the function. The information may tend to have some bias in some case. This is understandable as some of the venues were paid to increase the search priority shown to the users. Then, the venues showed in the API is not solely consisting of the venues near the places that I explored.

Conclusion

In all, in the above research, the result is not concrete enough to be used. However, the research paved a way for further investigation to the similarities for two cities. There are different ways to improve the research accuracy. The improvements include:

1. Using a different location provider which can provide a fair comparison between the neighbourhoods.
2. Using more than 10 venues in each neighbourhoods.
3. Separating the data from central part of the city to that of the less busy part of the city. The central part of the city tends to be a lot more crowded and they have a lot more venues of outcome than those in the less busy areas. In order to have more accurate comparison, they should be investigated separately. For those neighbourhoods in central part, searching radius can kept to be 500 meters. However, in the less busy areas, we can raise the searching radius to 1km. This can increase the venues found in the less busy areas and so that more features will be assessable.
4. Using K-Mean algorithm several times and take the average. As K-Mean algorithm gives random output depending on the random seed created initially, taking average for the clusters is a way to alleviate the problem.
5. Considering to use other Machine Learning algorithm, for example Density-Based Spatial Clustering of Applications with Noise (DBSCAN Clustering).