Coursera Capstone Project
IBM Applied Data Science Capstone

Explore Chinese Restaurants in London

Author: Tzu-Hsuan Chang
2021/07

1. Introduction

   The idea in my mind is to help clients to find an exotic restaruant in a certain city based on my orientation of hometown. Because as I am also in a role of client, take my feet into other people's shoes is also the same thought as the project's target.

   London is already an international city, with a lot of population and immigrants grown for years. However, although Asian food has already served in several locations here. It is sometimes really struggling and find it difficult to search for a good place to enjoy the cuisines and delicacy because there are too many options to select, also not so good and common in certain neighborhoods and locations.

   The other reason for this is that as I chose to study master in United Kingdom for the next 2 years, the homesickness of food and taste of cuisines always draw my attraction while walking on the street of cities. Every spices and food smell in such an atmosphere would be a real happiness for me only to enjoy just one or two meals for three days a week.

   So I would like to research the suitable district to find a cosy environment for tourists like me to get the best location of **Asian restaurant** in **London, United Kingdom** with several combinations from especially East Asian and North-east Asian neighbourhoods (because it is kind of too difficult to find the pure Taiwanese flavour food) to help the people who have the same background like me who wish to find a good place to dine in.

1-1.   Business Problem

 The brief question of this would be :
   1. Which neighborhoods should be the best decision of opening an Asian restaurant in **London, United Kingdom**?
   2. How to choose the most suitable places?
   3. And for explorers(or tourists), where would be the best neighbourhoods to choose to have cuisine?

1-2.   How to decide?

 My idea is to decide from these steps:

   1. Locate and find the data of neighborhoods in **City of London or suburban area of Greater London**, but better in city center of London
   2. Discover and construct the datas of similar restaurants in London, like Japanese, Korean, Chinese and other Asian restaurants.
   3. Foursquare API to find the venues
   4. **KMeans and clustering** to find the statistics which would be in use for the cluster on maps
   5. Analyse the results and give some suggestions of locations which suit for the purpose of opening an Asian restaurant in London
   6. Conclusion

1-3.   Target Audience

As I mentioned, London is an international and multicultural city for long. So our target would be **every tourists , immigrants and long-living foreigners** from East Asian countries who seek for the missing flavour or to open up a new business related to restaurants industry of Asian related flavor, especially from east Asia.

2. Data

2-1. How does the dataset come from?

To explore, we need to import and build neighborhoods data of London from the most common used data resources, **Wikipedia for the datasets from London, a brief summary of Locations, Borough, and later on we would add coordinates(latitude, longitude) on the dataset for it to be easier in use mapping and Foursquare which is used for finding the venues especially the categories of restaurants**. Then these two datasets will combined into one set for the newly data frame used for clustering.

2-2. How to solve the problem with datasets?

The most straightforward solution is to find the venues with the neighbourhoods of London Boroughs. Foursquare API would be considered of the assistance tools. The neighborhoods to transport connection and other amenities would be correlated, at the same time, accessibility and ease of supplies are also the priorty selection.

2-3. Data process

For datasets, we need to have a basic cleaning and combination.

1. **Web-Scrapping**: With BeautifulSoup, we could requests and get the url like: https://en.wikipedia.org/wiki/List_of_areas_of_London as the data resources. The dataset should be inspected first from the terminal to find the 'table':{'class' : 'wikitable sortable'}.
2. **Construction Coordinates**: Data Frame would be created for the new dataframe, however there are two columns, which is also the coordinates we mentioned above, 'Latitude' and 'Longitude', are required to add on to the data frame. So geocoder is required for the library, installation should be done and wait for a bit time to operate the process of coordinates since there are 500 more data
3. **Find the coordinates to finish the brief datasets**: Then the data frame is collected for the geographical coordinate, we could find the centroids of London via address and geolocator again to acquire the coordinates, which is easier for the further process.
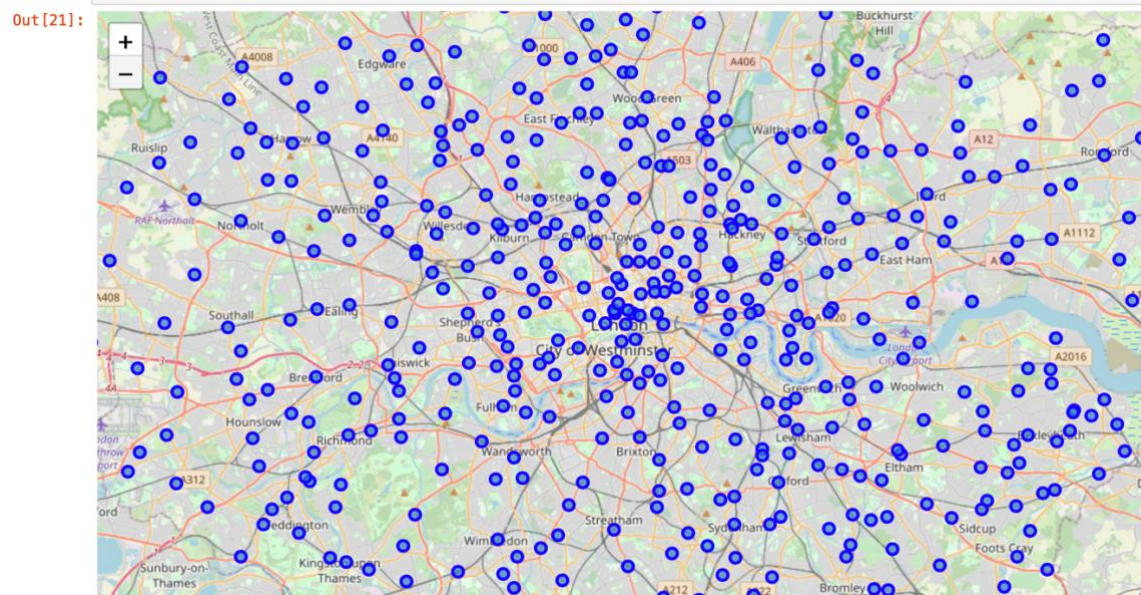
3. Methodology

3-1. Geocoding

For this step, we are going to work on geocoding with **FourSquare API**. The FourSquare API will be used to obtain the map and other venues data.

1. **Folium Mapping**:

    Just like Google api, this time we are going to work on Foursquare API to acquire the map of London first as it shown like this:

2. **Limit the range of data and repeat the same process**:

    We would like to limit the range of map with coordinates to find the longitude and latitude within a certain value to make it more concentrate to London Area. And reset the index and create again.

3. **Foursquare API:**

    Venues are required in the step, so we would use Foursquare API to find the venues of London City. Define a function with getNearbyVenues from latitude and longitude to acquire each neighborhoods' data and every business like stores, shops, restaurants, museums and other facilities. And especially, sort the "**restaurants**" out because the category is going to be our main target for research. There are 81 unique categories of restaurants in London as the output shown for several steps.

4. Results

4-1. Analysis

1. **One hot encoding:**

   From the project done before, one hot coding would be used again here in the dataset. This is an important step before clustering since we could define a new data frame easier as shown like Neighborhoods and other different styles of restaurants.

	Neighborhood	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Australian Restaurant	Austrian Restaurant	Brazilian Restaurant	Cantonese Restaurant	Caribbean Restaurant	Cau Res
11	Acton	0	0	0	0	0	0	0	0	0	0	0	
16	Acton	0	0	0	0	0	0	0	0	0	0	0	
40	Addiscombe	0	0	0	0	0	0	0	0	0	0	0	
42	Addiscombe	0	0	0	0	0	0	0	0	0	0	0	
50	Albany Park	0	0	0	0	0	0	0	0	0	0	0	

2. **Group up the data with means of neighborhoods :**

   Continue to work on the datasets, the means of neighborhoods could also be found for the coding with .mean( ) to get the rate or proportion in each neighborhood.

	Neighborhood	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Australian Restaurant	Austrian Restaurant	Brazilian Restaurant	Cantonese Restaurant	Caribbean Restaurant
0	Acton	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1	Addiscombe	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
2	Albany Park	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
3	Aldgate	0.000000	0.000000	0.000000	0.000000	0.030303	0.060606	0.000000	0.000000	0.000000	0.030303	0.000000
4	Aldwych	0.000000	0.000000	0.038462	0.000000	0.000000	0.038462	0.000000	0.000000	0.000000	0.000000	0.000000
5	Alperton	0.000000	0.000000	0.000000	0.000000	0.000000	0.333333	0.000000	0.000000	0.000000	0.000000	0.000000
6	Anerley	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
7	Angel	0.050000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.050000	0.000000	0.000000	0.000000
8	Aperfield	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

3. List top 5 frequent numbers of restaurants in each neighborhoods:

   For example:

```
----Acton----
                venue  freq
0  Fast Food Restaurant   0.5
1       Chinese Restaurant   0.5
2        Paella Restaurant   0.0
3               Restaurant   0.0
4         Ramen Restaurant   0.0


----Addiscombe----
                venue  freq
0  Fast Food Restaurant   0.5
1       Chinese Restaurant   0.5
2        Paella Restaurant   0.0
3               Restaurant   0.0
4         Ramen Restaurant   0.0
```

4. Data frame created with top 10 most common venue:

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 Acton	Fast Food Restaurant	Chinese Restaurant	Yoshoku Restaurant	Ethiopian Restaurant	Filipino Restaurant	French Restaurant	German Restaurant	Gluten-free Restaurant	Greek Restaurant	Halal Restaurant
1 Addiscombe	Fast Food Restaurant	Chinese Restaurant	Yoshoku Restaurant	Ethiopian Restaurant	Filipino Restaurant	French Restaurant	German Restaurant	Gluten-free Restaurant	Greek Restaurant	Halal Restaurant
2 Albany Park	Indian Restaurant	Yoshoku Restaurant	Halal Restaurant	Falafel Restaurant	Fast Food Restaurant	Filipino Restaurant	French Restaurant	German Restaurant	Gluten-free Restaurant	Greek Restaurant
3 Aldgate	Indian Restaurant	Restaurant	Middle Eastern Restaurant	Thai Restaurant	Italian Restaurant	Mediterranean Restaurant	English Restaurant	Sushi Restaurant	Asian Restaurant	Vietnamese Restaurant
4 Aldwych	Restaurant	French Restaurant	Sushi Restaurant	Japanese Restaurant	Italian Restaurant	Peruvian Restaurant	Portuguese Restaurant	Ramen Restaurant	Mexican Restaurant	Lebanese Restaurant

4-2. Statistics and Clustering

1. Find K-Means and clustering model here:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
11	Acton	51.508140	-0.273261	Amigo's Peri Peri	51.508396	-0.274561	Fast Food Restaurant	1	Fast Food Restaurant	Chinese Restaurant	Yoshoku Restaurant	Ethiopian Restaurant	Fil... Resta...
16	Acton	51.508140	-0.273261	North China Restaurant	51.508251	-0.277435	Chinese Restaurant	1	Fast Food Restaurant	Chinese Restaurant	Yoshoku Restaurant	Ethiopian Restaurant	Fil... Resta...
40	Addiscombe	51.379692	-0.074282	KFC	51.380105	-0.073570	Fast Food Restaurant	1	Fast Food Restaurant	Chinese Restaurant	Yoshoku Restaurant	Ethiopian Restaurant	Fil... Resta...
42	Addiscombe	51.379692	-0.074282	Welcome Friends	51.378230	-0.073857	Chinese Restaurant	1	Fast Food Restaurant	Chinese Restaurant	Yoshoku Restaurant	Ethiopian Restaurant	Fil... Resta...
50	Albany Park	51.435384	0.125965	Ruchi	51.435973	0.126779	Indian Restaurant	4	Indian Restaurant	Yoshoku Restaurant	Halal Restaurant	Falafel Restaurant	Fast I... Resta...

2. Final mapping with clustering model like this:
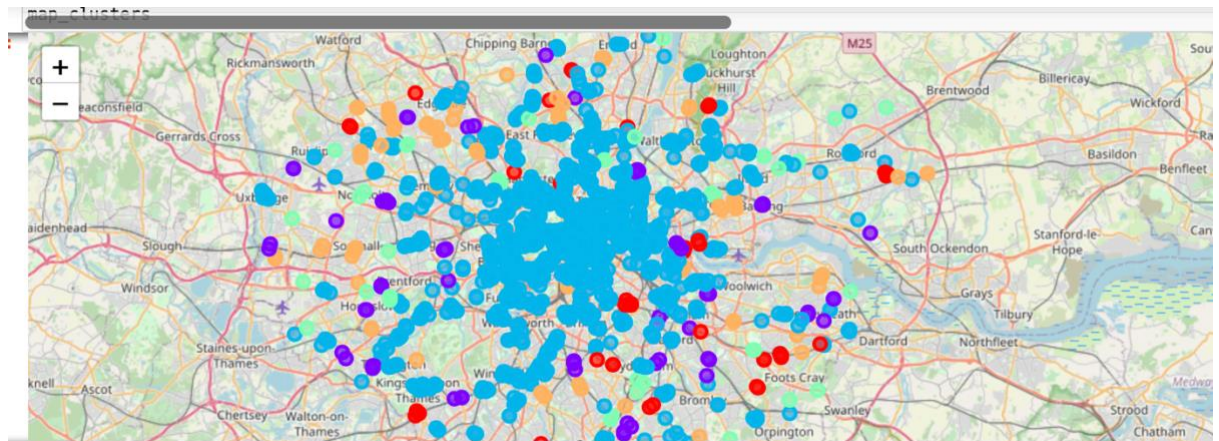
3. **Clustering**:
   Since kcluster = 5, so there would be five different cluster levels with different datasets and I conclude the results as it shown like this:

```
|Cluster| Most Common Restaurant | 2nd Common Restaurant | Least Common Restaurant|
|-------|------------------------|-----------------------|------------------------|
|   1   | Italian                | South European        | Gluten-free            |
|   2   | Fast Food              | Chinese               | Halal                  |
|   3   | Depends                | Depends               | Depends                |
|   4   | Fast Food              | Yoshoku               | Halal                  |
|   5   | India                  | Yoshoku               | Greek                  |
```

5. Discussion

What could we learn from the project?

First of all, if we choose to open up and explore a new restaurant. We would like to make two different thoughts.

1. For owner who would like to open up a new business, suggestion would be made for opening in least common restaurant since there would not be too much competitive. Like in Cluster 1, least common restaurant is Gluten-free related, so start a business with these kind of store would be a better choice and it could increase the popularity to open up a Chinese restaurant. But if location is not an important factor for the choice, then more common the style restaurnts in each cluster, the better to open the specific restaurant based on different kinds of restaurants.

2. For explorer(tourists), since we have certain flavour-like restaurants or cuisine. Most Common Restaurants in each cluster is the main factor to choose as the distinct thought, like Chinatown for Chinese food of course.

6. Conclusion

As an East-Asian oriented student, if I would like to open up an Asian-flavoured restaurant, **Cluster1** is definitely the best to choose. Because since the common restaurants here are mostly for European

cuisine, like Italian, Portuguese styles and so on. **This would not be the competitive area for a Chinese or East-Asian restaurant naive owner to open up a business.**

   In conclusion, this project should have better results since we do not have enough information and data collections to decide several factors like price, races, costs and other decisive reasons to decide and make sure if these would influence our selection of restaurants not only for opening a store but also for discovering a good and cozy place to have a meal. But for a general understanding, I think this would be enough to satisfy the problem we are facing now.