Locations for Chinese Restaurant Startup in London

Applied Data Science Capstone by IBM/Coursera

Tan Kok Peng

14 SEP 2020

1. Introduction: Business Problem

The purpose of this project is to find optimal locations for a Chinese restaurant in London. This report is targeted at stakeholders, primarily but not limited to Chinese residing in London, British Chinese as well as new Chinese migrants who are interest in starting and operating a Chinese restaurant business in London.

1.1 Target Audience & Interest

According to Wikipedia (https://en.wikipedia.org/wiki/British_Chinese), Chinese form 1.4% of the population of London (107,100 / 7,456,000). The Chinese population is extremely dispersed, according to Rob Lewis, a senior demographer at the Greater London Authority: ***"The reason for their thin spread all over London, is because of the idea that you want to set up a Chinese restaurant that's a little way away from the next one."***

The statement made by Rob Lewis reinforces the applicability of the analysis discussed in this paper. We will be using geolocation data to identify low density Chinese restaurant areas in London suitable for startup which is a prime consideration for our stakeholders.

1.2 The Problem

Since there are lots of restaurants in London, the following will outline the selection criteria:

1. Locations that are not already crowded with restaurants
2. Areas with no Chinese restaurant in the vicinity
3. Near to city center of London (assuming the first two conditions are met)

This paper will use data science approach with geolocation data to identify low density Chinese restaurant areas within the boroughs of London and generate a list of suitable locations based on the criteria above. The advantages of each area will then be clearly expressed so that the best possible final location can be considered and chosen by stakeholders.

2. Data

Based on definition of the problem, factors that will influence the stakeholders' decision are:

- number of existing restaurants in the neighborhood (any type of restaurant)
- number of and distance to Chinese restaurants in the neighborhood, if any
- distance of neighborhood from city center of London

We decided to use regularly spaced grid of locations, centered around city center of London, to define our neighborhoods.

Following data/API sources will be needed to extract/generate the required information:

- centers of candidate areas will be generated algorithmically and approximate addresses of centers of those areas will be obtained through reverse geocoding using **Google Maps Geocoding API**
- number of restaurants and their type and location in every neighborhood will be obtained using **Foursquare API**
- coordinate of city center of London will be obtained using **Google Maps Geocoding API**
- London Bouroughs Boundaries available at https://skgrange.github.io/www/data/london_boroughs.jsonhttps://skgrange.github.io/www/data/london_boroughs.json

Since we will be dealing with addresses and geographic coordinates in this project, Google Maps Geocoding API is an effective tool to convert between the two.

2.1 Geolocation Data of Neighbourhood Candidates

We first began by getting the latitude & longitude of city center of London using Google Maps Geocoding API. We then created a grid of cells covering our area of interest which is approximately 12x12 kilometers centered around city center of London.

Each cell is centered around city center and within ~6km from city center of London. Our neighbourhoods is defined as circular areas with a radius of 300 meters, so our neighbourhood centers is 600 meters apart.

To accurately calculate the distances, we need to create our grid of locations in Cartesian 2D coordinate system which allows us to calculate distances in meters (not in latitude/longitude degrees). These coordinates are projected back to latitude/longitude degrees to be shown on Folium map using a transform function. The functions to convert between WGS84 spherical coordinate system (latitude/longitude degrees) and UTM Cartesian coordinate system (X/Y coordinates in meters) are defined and created.

In order to create a hexagonal grid of cells, we offset every other row, and adjust vertical row spacing so that every cell center is equally distant from it's neighbours. Figure 1 shows the Folium plot of city center of London location and candidate neighborhood centers (hexagonal grid of cells).

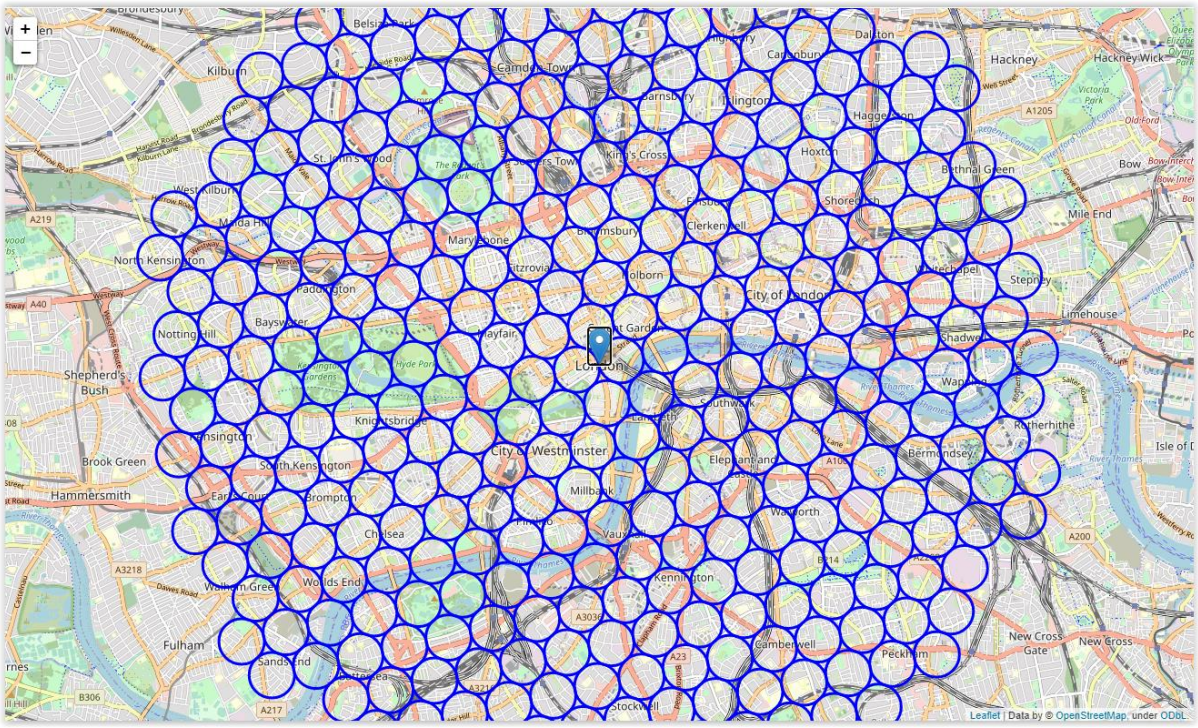

Figure 1. City Center of London and Candidate Neighbourhood Centers

We have created 364 candidate neighbourhoods around the center of London and obtained the latitude, longitude, distance from center, Cartesian coordinates X Y of each of these neighbourhood. Using Google Maps Geocoding API, we converted the latitude and longitude data of each of these candidate neighbourhoods into a known address and store them into a dataframe shown in Table 1.

	Address	Latitude	Longitude	X	Y	Distance from center
0	Warley House, 91 Elms Cres, Clapham Common, Lo...	51.454434	-0.135968	-548819.423256	5.809846e+06	5992.495307
1	4 Prague Pl, Brixton, London SW2 5ED, UK	51.455536	-0.127632	-548219.423256	5.809846e+06	5840.376700
2	56 Brixton Hill, Brixton, London SW2 1QS, UK	51.456637	-0.119296	-547619.423256	5.809846e+06	5747.173218
3	57 Rattray Rd, Coldharbour, London SW2 1BB, UK	51.457737	-0.110959	-547019.423256	5.809846e+06	5715.767665
4	18, Bessemer Park, 250 Milkwood Rd, Herne Hill...	51.458837	-0.102622	-546419.423256	5.809846e+06	5747.173218
5	Denmark Hill, Herne Hill, London SE5 8DX, UK	51.459936	-0.094284	-545819.423256	5.809846e+06	5840.376700
6	67 Green Dale, London SE5 8JZ, UK	51.461035	-0.085946	-545219.423256	5.809846e+06	5992.495307
7	Mount Pond, 8A Windmill Dr, London SW4 9DE, UK	51.457290	-0.149999	-549719.423256	5.810366e+06	5855.766389
8	Brook House, 47-48 Clapham Common South Side, ...	51.458393	-0.141664	-549119.423256	5.810366e+06	5604.462508
9	137 Clapham Park Rd, Ferndale, London SW4 7BH, UK	51.459495	-0.133327	-548519.423256	5.810366e+06	5408.326913

Table 1. Dataframe of Candidates Neighbourhoods around London

2.2 Foursquare API

Foursquare API is used to get info on restaurants in each candidate neighborhood which we have mapped out earlier.

Since we are interested in venues of the 'food' category, also only those that are proper restaurants, we have created keywords filter to include only venues that have 'restaurant' and other keywords in the category name, so we could capture them in the list. Coffee shops, pizza places, bakeries etc. are

not direct competitors and will be excluded. We have also included all the subcategories of 'Chinese restaurant' available through the Foursquare API in the neighbourhood.

Based on latitude and longitude of each candidate neighbourhood, we obtained latitude, longitude, food category of all restaurants, Chinese restaurants with a radius of 350 meters to have a good overlap between neighbourhoods. We stored the queries using dictionaries to remove any duplicates resulting from area overlaps.

The Foursquare API returned venue dataset and we run through Numpy to obtain the following statistics:

- Total number of restaurants: 2946
- Total number of Chinese restaurants: 152
- Percentage of Chinese restaurants: 5.16%
- Average number of restaurants in neighborhood: 7.3791208791208796

Figure 2 shows the Folium plot of the restaurants we have obtained through the Foursquare API query.

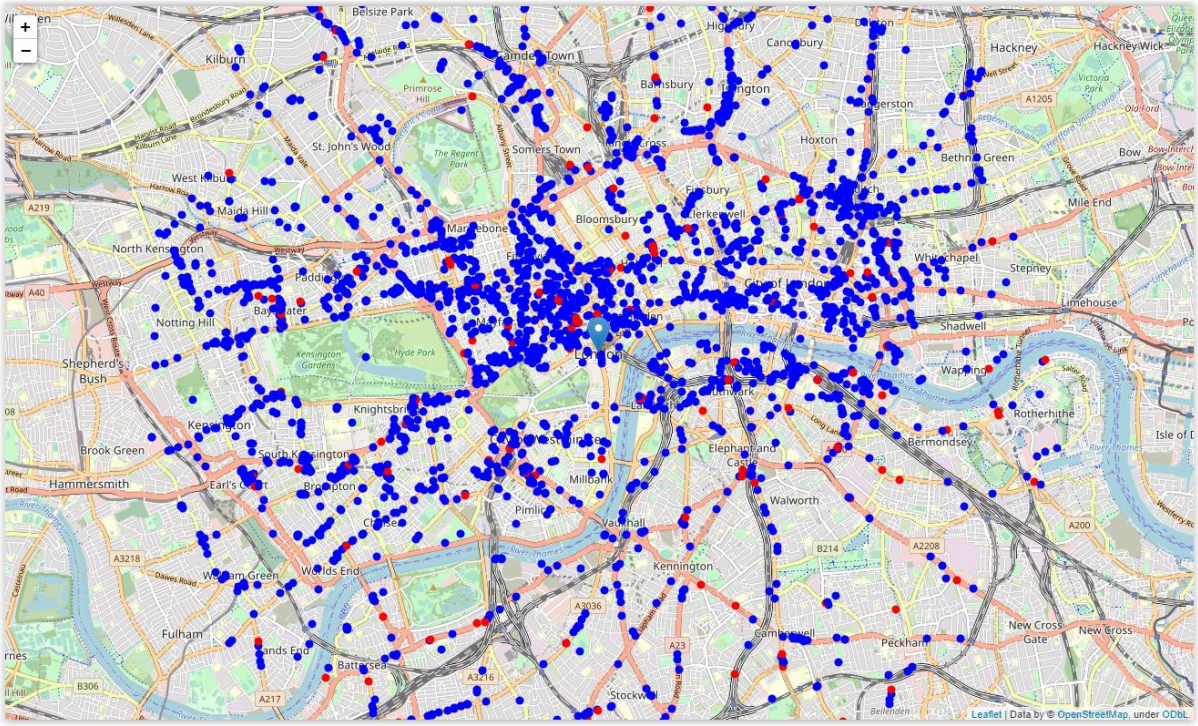

Figure 2. Restaurants Locations around the Candidate Neighbourhoods of London

The plot shows restaurant in blue and Chinese restaurant in red in candidate neighbourhoods within few kilometers from center of London. This concluded the data gathering phase.

3. Methodology

We now direct our efforts on detecting areas around center of London that have low restaurant density, particularly those with low number of Chinese restaurants. We will limit our analysis to area ~6km around city center.

Prior to this, we have already collected the required data: **location and type (category) of every restaurant within 6km from center of London**. We have also **identified those that are Chinese restaurants** (according to Foursquare API categorization).

Now we will apply the information we have into our analysis; using them for calculation and exploration of 'restaurant density' across different areas of London. We will use heatmaps overlay plot in Folium to identify a few suitable areas close to center with low number of restaurants and no Chinese restaurants in vicinity and focus our attention on those areas.

Finally, we will focus on these suitable areas identified. Using K-Means unsupervised clustering algorithm to create clusters of locations that meet the following criteria within those areas:

1. Not more than 2 restaurants in a radius of 250 meters
2. No Chinese restaurant in a radius of 400 meters for location

The result along with the Folium plots to identify general zones, neighbourhoods and suitable location addresses will be used as a starting point in discussion with the stakeholders. This will be the basis for any follow-up or final 'street level' exploration and search for suitable venue undertake by the stakeholders.

4. Analysis

We performed some basic explanatory data analysis and derived some additional information from the data we have collected earlier. We counted the number of restaurants in every candidate neighbourhood as well as the distance to nearest Chinese restaurant from neighbourhood center. These are stored into the dataframe shown in Table 2.

	Address	Latitude	Longitude	X	Y	Distance from center	Restaurants in area	Distance to Chinese restaurant
0	Warley House, 91 Elms Cres, Clapham Common, Lo...	51.454434	-0.135968	-548819.423256	5.809846e+06	5992.495307	4	758.153624
1	4 Prague Pl, Brixton, London SW2 5ED, UK	51.455536	-0.127632	-548219.423256	5.809846e+06	5840.376700	0	868.522156
2	56 Brixton Hill, Brixton, London SW2 1QS, UK	51.456637	-0.119296	-547619.423256	5.809846e+06	5747.173218	2	763.749781
3	57 Rattray Rd, Coldharbour, London SW2 1BB, UK	51.457737	-0.110959	-547019.423256	5.809846e+06	5715.767665	0	386.255432
4	18, Bessemer Park, 250 Milkwood Rd, Herne Hill...	51.458837	-0.102622	-546419.423256	5.809846e+06	5747.173218	0	659.600478
5	Denmark Hill, Herne Hill, London SE5 8DX, UK	51.459936	-0.094284	-545819.423256	5.809846e+06	5840.376700	0	1023.679291
6	67 Green Dale, London SE5 8JZ, UK	51.461035	-0.085946	-545219.423256	5.809846e+06	5992.495307	0	477.670768
7	Mount Pond, 8A Windmill Dr, London SW4 9DE, UK	51.457290	-0.149999	-549719.423256	5.810366e+06	5855.766389	1	1073.897656
8	Brook House, 47-48 Clapham Common South Side, ...	51.458393	-0.141664	-549119.423256	5.810366e+06	5604.462508	1	502.782254
9	137 Clapham Park Rd, Ferndale, London SW4 7BH, UK	51.459495	-0.133327	-548519.423256	5.810366e+06	5408.326913	4	268.930870

Table 2. Dataframe of Candidate Neighbourhoods with Restaurant Count and Distance to Chinese Restaurant

Average distance to closest Chinese restaurant from each area center: 514.0762671817615

The analysis shows on average Chinese restaurant can be found within 500 meters approximately from every centers of candidate neighbourhoods. This is fairly close, thus we have to filter our areas carefully!

We then created a map showing heatmap / density of restaurants (Figure 3) and try to extract some meaningful information from that. The borders of London boroughs is also plotted on our map and a few circles indicating distance of 1km, 2km and 3km from the city center.

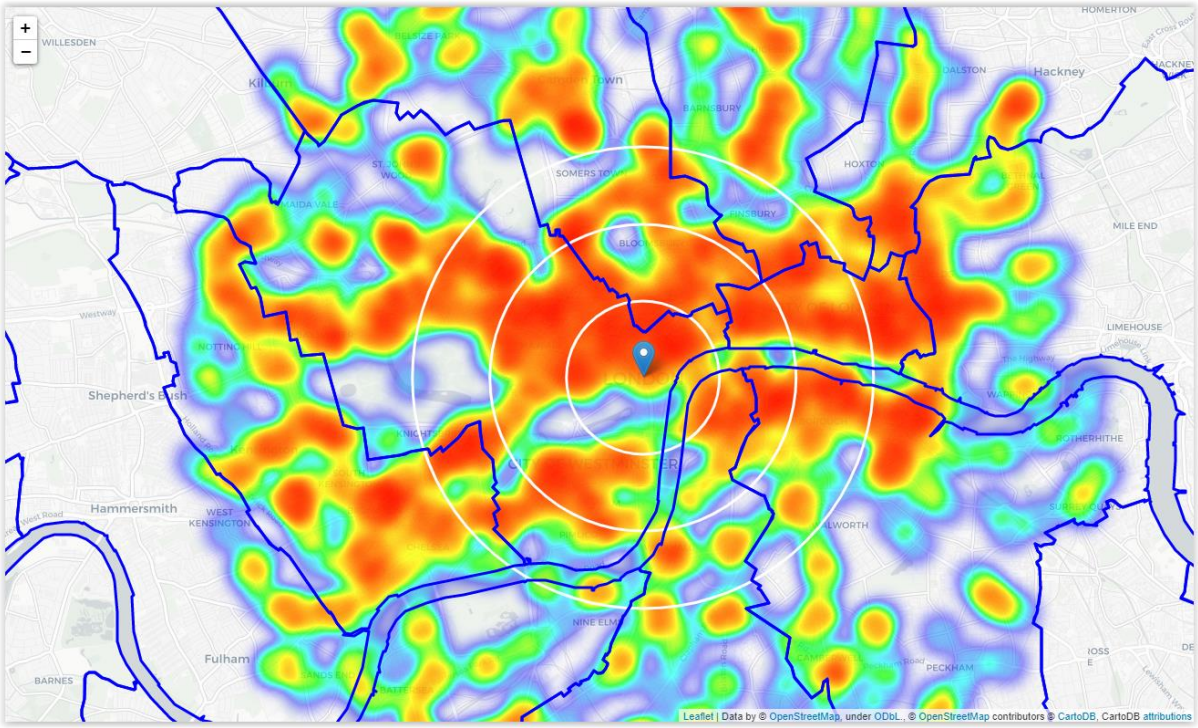

Figure 3. Heatmap Plot of Restaurants around London

From heatmap plot of the restaurants, there seems to a few pockets of low restaurant density closest to city center around the **south-west, south, south-east and east from city center** from 2-3km radius.

This is verified using another heatmap map plot showing only **heatmap/density of Chinese restaurants**, shown in Figure 4.

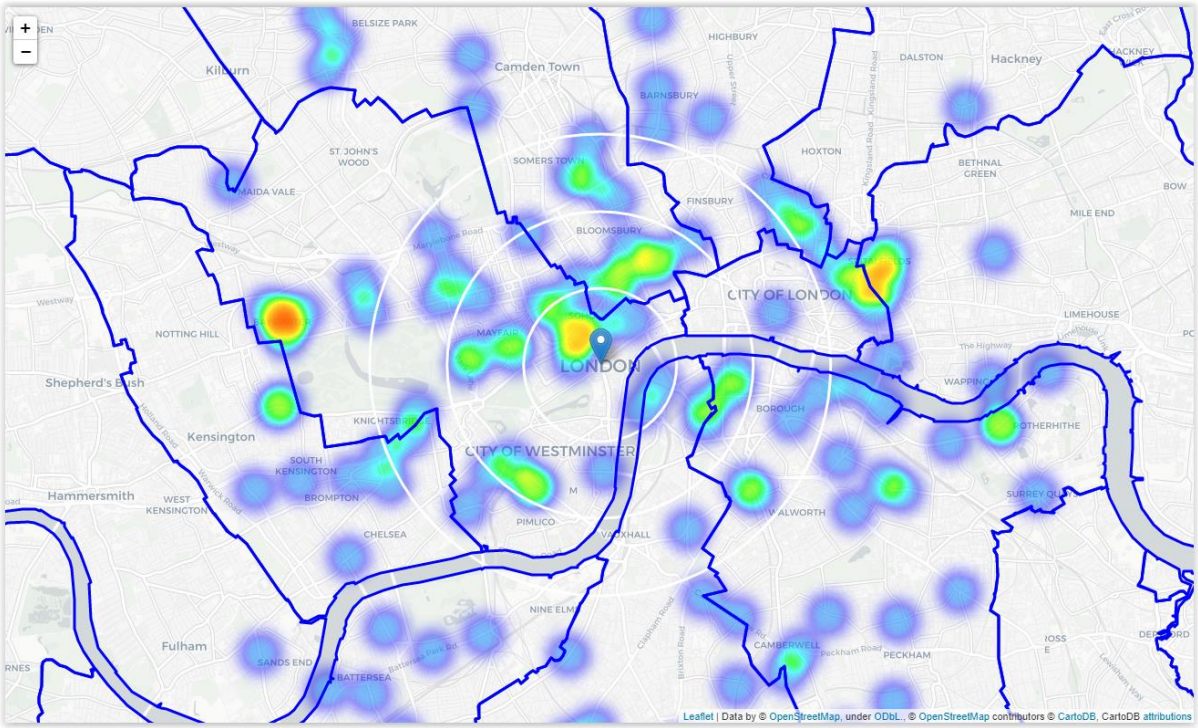

Figure 4. Heatmap Plot of Chinese Restaurants around London

This map is not so 'hot' as can be seen from the plot. We have calculated earlier that Chinese restaurants represent a subset of ~5% of all restaurants in London. The plot also indicates higher density of existing Chinese restaurants directly north and west of London within a kilometer radius. The plot identifies closest pockets of **low Chinese restaurant density positioned south-west, south, south-east and east from city center**.

Based on this we will now focus our analysis on areas *south-west, south, south-east and east from city center* - we will move the center of our area of interest and reduce its size to have a radius of **2.5km**. This places our location candidates mostly in boroughs **Westminster, City of London, Southwark, Lambeth and Wandsworth**.

4.1 Exploring the London Boroughs

Based on the discovery above, we now defined a narrow region of interest, which will include low-restaurant-count parts of regions of boroughs in London. This region of interest is 2.5km offset from the city center of London and plotted using Folium in Figure 5.

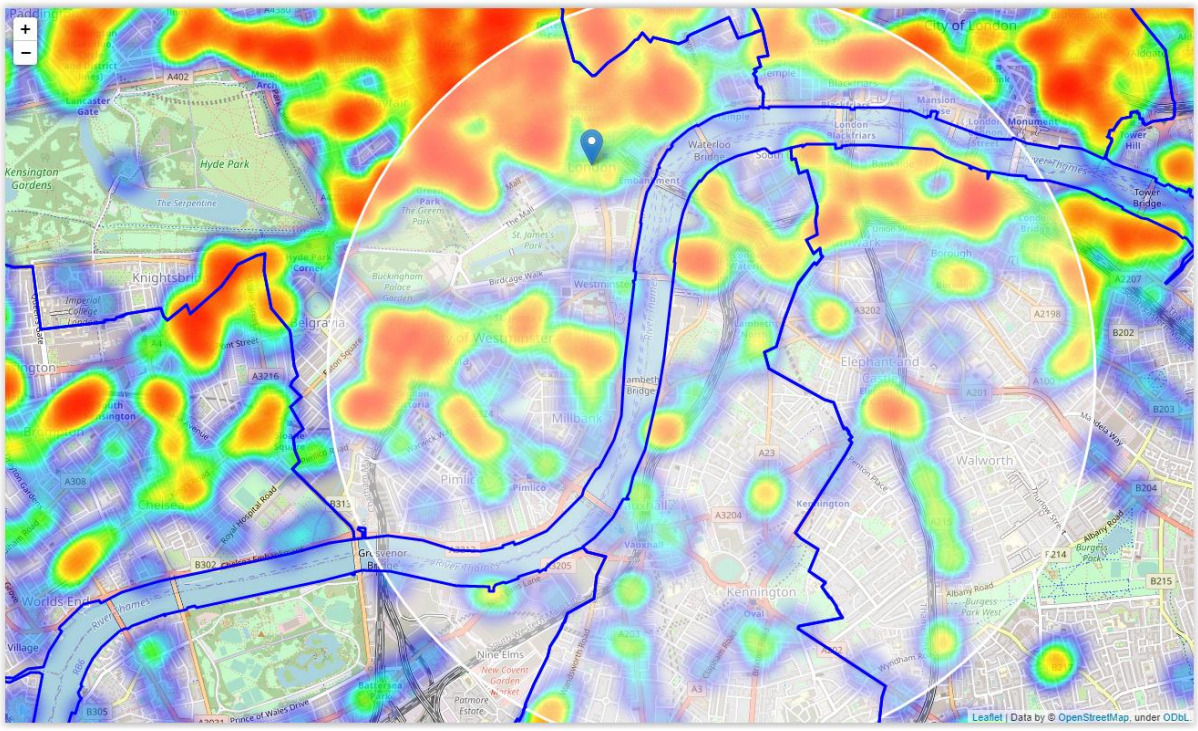

Figure 5. Region of Interest

The plot nicely covers all the pockets of low restaurant density in Westminster, City of London, Southwark, Lambeth and Wandsworth closest to city center.

Now we that have a new region of interest we have created new candidate neighbourhoods, more densely packed between neighbourhoods making them 100m apart. This created 2261 candidate neighbourhoods around the center of the region of interest and obtained the latitude, longitude, Cartesian coordinates X Y of each of these neighbourhood, nearby restaurant count and distance to Chinese restaurant shown in Table 3.

	Latitude	Longitude	X	Y	Restaurants nearby	Distance to Chinese restaurant
0	51.473456	-0.109742	-546569.423256	5.811562e+06	0	686.600667
1	51.473639	-0.108352	-546469.423256	5.811562e+06	0	626.334159
2	51.473199	-0.117641	-547119.423256	5.811648e+06	1	1030.317092
3	51.473382	-0.116251	-547019.423256	5.811648e+06	0	953.944447
4	51.473566	-0.114861	-546919.423256	5.811648e+06	1	882.307603
5	51.473749	-0.113471	-546819.423256	5.811648e+06	1	816.653784
6	51.473933	-0.112081	-546719.423256	5.811648e+06	1	758.538134
7	51.474116	-0.110691	-546619.423256	5.811648e+06	0	709.814623
8	51.474299	-0.109301	-546519.423256	5.811648e+06	0	653.141337
9	51.474483	-0.107911	-546419.423256	5.811648e+06	0	563.511506

Table 3. Dataframe of New Candidate Neighbourhoods

On the dataframe, we **filter** those locations: we're interested only in **locations with no more than two restaurants in radius of 250 meters**, and **no Chinese restaurants in radius of 400 meters**. These gives us all possible good locations for the restaurant startup and we stored them into a new dataframe containing all good locations. Below are the aggregated sum of the dataframe output:

- Locations with no more than two restaurants nearby: 942
- Locations with no Chinese restaurants within 400m: 1219
- Locations with both conditions met: 694

The locations are visualized using Folium plot shown in Figure 6.

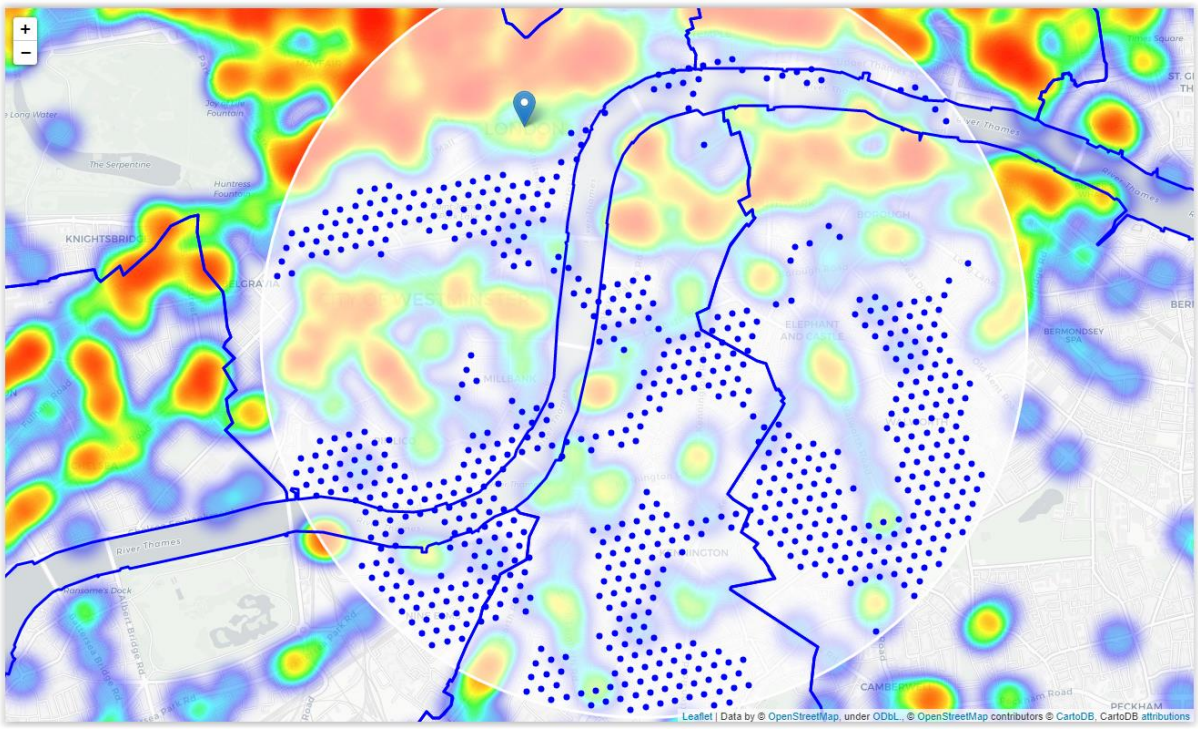

Figure 6. Good Locations in the Region of Interest

Based on the plot and the boroughs borders, the locations fall into mostly in Westminster, Lambeth and Southwark boroughs with clusters of locations from Wandsworth and City of London forming up at the south-west and north-east coners respectively. Each of these locations has no more than two restaurants in radius of 250m, and no Chinese restaurant nearer than 400m. Thus, any of these locations is a potential candidate for a new Chinese restaurant, **at least based on proximity of access and nearby competition**.

We plot the region of good location with heatmap overlay for better clarity in Figure 7.

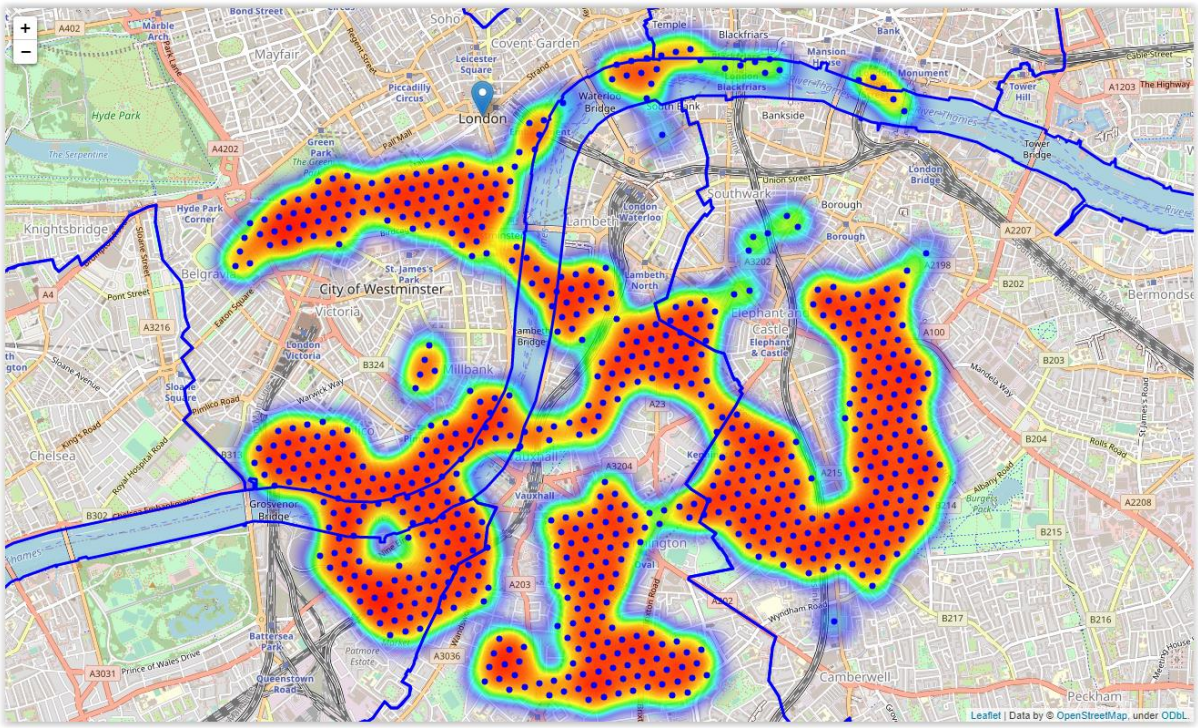

Figure 7. Heatmap of Good Locations

What we have now is a clear indication of region with low number of restaurants in vicinity, and ***no*** Chinese restaurants at all nearby.

4.2 K-Means Clustering

We applied K-Means clustering to **cluster** these good locations into **zones** and to create **centers of zones containing geographical coordinates**. The zones, their centers and finally the addresses that can be obtained using reverse geocoding through Google Maps Geocoding API will be the final result of our analysis.

We let **K = 20** so we can obtain a list of 20 addresses for stakeholders' review.

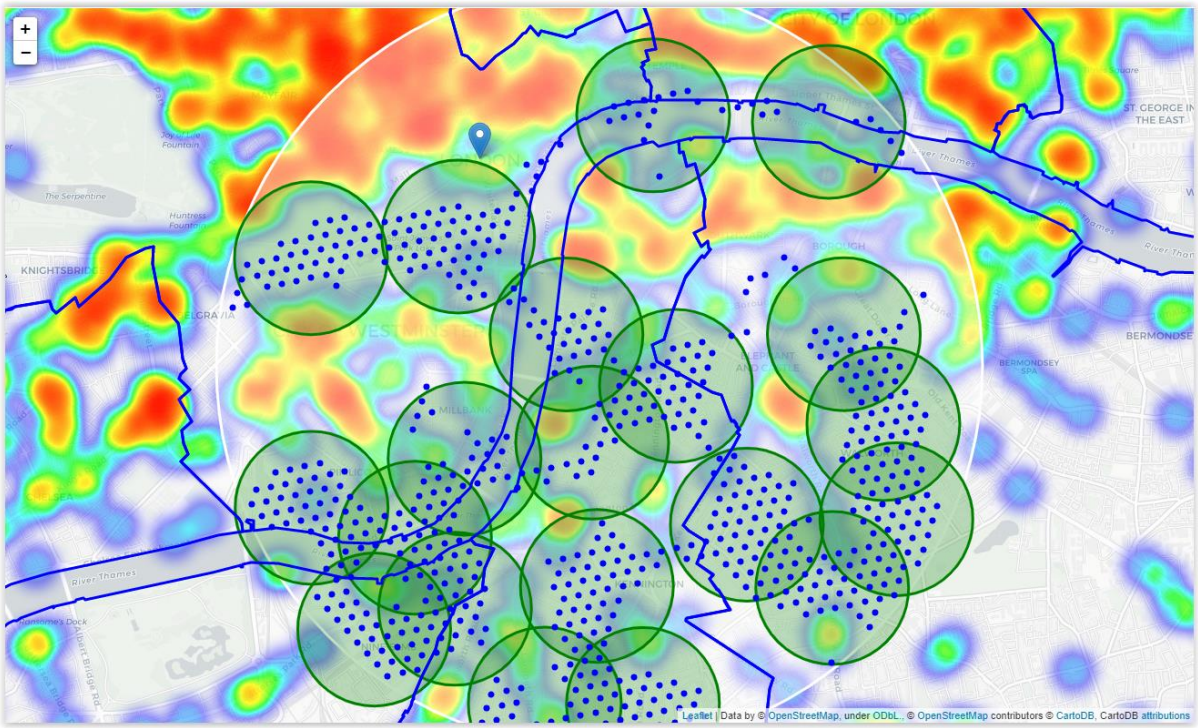

Figure 8. K-Means Clustering of Good Locations

Figure 8 shows the zones after applying K-Means clustering on the good locations. The clusters represented groupings of the candidate locations and cluster centers are placed nicely in the middle of the zones.

Addresses derived through reversed geocoding of these cluster centers will be a good starting point for exploring the neighborhoods to find the most suitable location for Chinese restaurant.

Figure 9 shows the zones on the city map without heatmap, using shaded areas to indicate our clusters.

Figure 9. K-Means Clustering of Good Locations without Heatmap

Finally, using Google Maps Geocoding API to **reverse geocode those candidate neighbourhood centers to get the addresses** which can be presented to stakeholders, shown in Table 4.

Addresses of centers of areas recommanded for further analysis	Distance from London center (km)
28 Walcot Square, Prince's, London SE11 4TZ, UK	2.0
A Rd, Nine Elms, London SW8 5AL, UK	3.2
6 Walnut Road, Westminster, Manchester M30 8LE, UK	1.3
51 Westcott Rd, London SE17 3QY, UK	3.0
21A Ponsonby Pl, Westminster, London SW1P 4PS, UK	2.0
78a Ashmole St, Oval, London SW8 1NE, UK	2.9
91 Rodney Rd, London SE17 1RF, UK	3.2
H.Q.S. Wellington, Victoria Embankment, London WC2R 2PN, UK	1.2
69 Horse Guards Rd, Westminster, London SW1A 2AG, UK	0.5
62 Wandsworth Rd, Oval, London SW8 2FA, UK	3.0
12 Villa St, London SE17 2EJ, UK	3.7
Wedgwood House, Pimlico, London SW1V 3BT, UK	2.6
Vintners Place, 68 Upper Thames St, Queenhithe, London EC4V 2AF, UK	2.3
Garlinge House, Gosling Way, Vassal, London SW9 6LU, UK	3.8
Falmouth Road Group Practice, 78 Falmouth Rd, London SE1 4JW, UK	2.7
A3036, South Bank, London SE1 7EH, UK	1.3
Sullivan House, Black Prince Rd, Vauxhall, London SE11 6JH, UK	2.0
7141 Albany Rd, London SE5 0DQ, UK	3.7
112 Portland Grove, Stockwell, London SW8 1JJ, UK	3.6
136 Grosvenor Rd, London SW1V 3JY, UK	2.5

Table 4. List of Addresses of the Good Locations Obtained through Google Maps Geocoding API

This concludes our analysis. The K-Means clustering algorithm created 20 addresses representing centers of the 20 zones. These are locations with low number of restaurants and no Chinese

restaurants nearby. All the zones are relatively near to city center of London, less than 4km approximately.

The addresses derived from the zone centers should be considered only as a starting point for exploring area in search for a suitable restaurant location. Most of the zones are located in boroughs discussed earlier. Since they are fairly close to city center, and well connected by public transport, they will not be short of tourists, office workers as well as local residents as patrons.

5. Results and Discussion

Our analysis shows that although there is a great number of restaurants in London (~3000 in our initial area of interest which was 12x12km around city center), there are pockets of low restaurant density fairly close to city center. Highest concentration of restaurants was detected north and west of London, so we focused our attention to areas south-west, south, south-east and east, corresponding to boroughs Westminster and Wandsworth, Lambeth, Southwark and City of London. They offer a combination of popularity among tourists and office workers, closeness to city center, strong socio-economic dynamics *and* a number of pockets of low restaurant density in the area.

We then look into these pockets of low restaurant density area in detail, covering approximately 5x5km south-east from city center London. We first created a dense grid of location candidates (spaced 100m apart); these locations are then filtered so that those with more than two restaurants in radius of 250m and those with a Chinese restaurant within 400m were removed. This provides a dataframe of good location candidates for a Chinese restaurant startup.

By applying K-Means clustering on the dataframe of good location candidates, we clustered the locations into zones of interest which contain greatest number of location candidates and determined the geographical coordinates (latitude, longtitude) of **center of each zone**. The addresses of centers of these zones are determined by parsing the latitude and longitude to Google Maps Geocoding API using reverse geocoding. The list of addresses is to be used as starting point for review and consideration with more detailed local/borough/neighbourhood analysis based on other factors.

The result is a list of addresses of the20 zones containing largest number of potential new restaurant locations based on number of and distance to existing venues - both restaurants in general and Chinese restaurants in particular. However, by using location and proximity data alone, does ***NOT*** imply these zones are optimal locations for a new Chinese restaurant! The purpose of this analysis approach is only the first step in providing information on areas close to London but not crowded with existing restaurants (particularly Chinese) and use it for further analysis. It is entirely possible that there is a very good reason for small number of restaurants in any of these zones identified; reasons which would make them unsuitable for a new restaurant regardless of lack of competition in the area. Recommended zones should therefore be considered only as a starting point for further and detailed analysis which could eventually result in a location which has not only has no nearby competition but also taking other factors into account and all other relevant criteria met e.g. rental, type of business approval, other legislation etc.

6. Conclusion

Purpose of this project is to identify areas close to center of London with low number of restaurants (particularly Chinese restaurants) in order to aid stakeholders in narrowing down the search for optimal location for a new Chinese restaurant. By calculating restaurant density distribution using Foursquare API data we have first identified general boroughs that justify further analysis (Westminster, Wandsworth, Lambeth, Southwark and City of London), and then generated extensive collection of locations which satisfy some basic requirements regarding existing nearby restaurants. Using K-Means clustering of these

locations is then performed in order to create major zones of interest (containing greatest number of potential locations) and addresses of these zone centers are determined using reverse geocoding. The list of addresses is used as starting points for final exploration by stakeholders.

Final discussion on optimal restaurant location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, legislation, social and economic dynamics of every neighborhood etc.