

# The Battle of Neighborhoods

## 1. Introduction

**Business problem:** London is the capital of England, UK, with a population of almost 9 million. It's a global financial and commercial hub, so it's a real challenge for entrepreneurs from all over the world to invest in any field of industry.

A very well-known Asian healthy Food and Supplements Stores-chain, wants to expand its businesses in Europe by opening, firstly, a new store in London. Despite their online-shops, company's success is based, mainly, on small stores providing healthy food-snacks and supplements to people who have adopted or want to adopt a healthy lifestyle. Board of directors of the company has already decided to try to apply the same recipe of success in London and they need us to investigate some good locations for that purpose.

We are going to collect all the appropriate data, first, and then we will analyze them with the help of K-Means algorithm in order to finally give our suggestions to directors.

## 2. Data

In order to build our recommender system, we need lots of data and, especially, specific data defining what venues exist in each neighborhood and where they are. In this way, we can examine which places can attract more of our providing products, for example venues like gyms and fitness centers. So, first, we need to find how many neighborhoods London has and what their names are. This data will be acquired from wikipedia. Secondly, we need to find their exact locations (Latitude, Longitude) from Google Maps Geocoding API. We need this info, because by using these coordinates in Foursquare, we can find nearest venues for each neighborhood. The result will be a table containing info of most common venues in each neighborhood and, thus, our data is ready for further process.

## 3. Methodology

All steps we followed analytically, are:

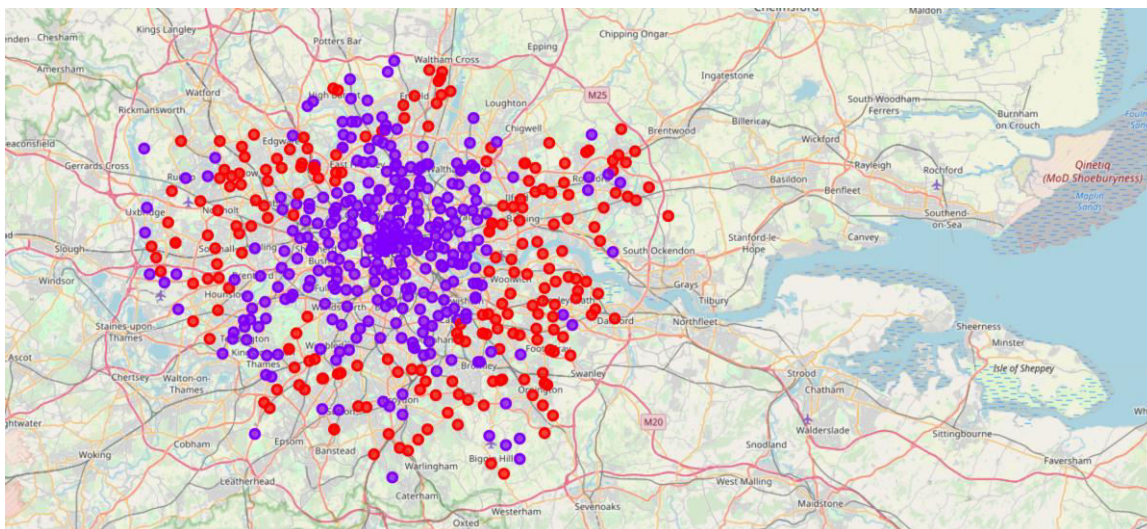
- gathering data from wikipedia. We used BeautifulSoup package to collect the list of areas of London from wikipedia. We then, stored them to a dataframe.
- gathering exact location (latitude, longitude) of each area/neighborhood of London, by using Google Maps Geocoding API and add them to the dataframe. We mapped all them in folium map.
- gathering data from Foursquare. Specifically, we got nearest

venues(name, category) of each neighborhood.

- pre-processing data. We applied one-hot encoding and created a more convenient dataframe containing neighborhoods and their nearest venues as percentages of occurrence. We also, found 10 most common venues per neighborhood.
- next step was to cluster all those neighborhoods based on their characteristics(different categories of venues). So, we applied, first silhouette analysis, as evaluation metric, to find out what K is optimum(K=2). After, we deployed K-Means algorithm and mapped the results in a folium map.
- we, then, examined each cluster and found which neighborhoods of which cluster have as 1st common venue Gym / Fitness Center, which was our target from the beginning.

## 4. Results and Discussions

The result of the above recommender system is that it produces a list of most common venues in every neighborhood, which, in its turn, belongs to a cluster comprised of similar venues. The produced clusters here in London are two. By observing the map, we can see that first cluster is a circle around the center of London and a bigger second circle around first one. This segmentation of neighborhoods of London, based on their neighborhoods, reveals that almost all neighborhoods have the same categories of venues. Venues are located almost equally to different areas of London and that is why London is one of the most modern cities all around the world in our days.



We found that neighborhoods of Colindale, The Hyde and the Upper Ruxley in first cluster and closer to the city centre have most common venue Gym / Fitness Center, so they are very good and promising locations to open a healthy food and supplements store, as the people of gyms and generally of fitness are more prone to buy our products. Similar neighborhoods in second

cluster, are Burroughs, Canary and Old Oak Common. So, these six locations are our answer to board of directors as the very good and promising locations to open the first store in Europe.

|     | Location       | Dial code | OS grid ref | Latitude  | Longitude | Cluster Labels | 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 |
|-----|----------------|-----------|-------------|-----------|-----------|----------------|-----------------------|---------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|------------------------|
| 73  | Burroughs, The | 020       | TQ227891    | 51.587840 | -0.229035 | 1              | Gym / Fitness Center  | Pub                       | Coffee Shop           | Park                  | Café                  | Department Store      | History Museum        | Grocery Store         | Pizza Place           | Golf Course            |
| 77  | Canary Wharf   | 020       | TQ375802    | 51.505431 | -0.023533 | 1              | Gym / Fitness Center  | Coffee Shop               | Plaza                 | Park                  | Bakery                | Burger Joint          | Hotel                 | Ramen Restaurant      | Pizza Place           | Bar                    |
| 105 | Colindale      | 020       | TQ213897    | 51.589691 | -0.249103 | 0              | Gym / Fitness Center  | Supermarket               | Coffee Shop           | Pub                   | Hotel                 | Park                  | Fast Food Restaurant  | Grocery Store         | History Museum        | Chinese Restaurant     |
| 253 | The Hyde       | 020       | TQ215888    | 51.586400 | -0.249812 | 0              | Gym / Fitness Center  | Pub                       | Park                  | Supermarket           | Coffee Shop           | Pet Store             | Hotel                 | History Museum        | Chinese Restaurant    | Korean Restaurant      |
| 351 | Old Oak Common | 020       | TQ216823    | 51.520470 | -0.250717 | 1              | Gym / Fitness Center  | Middle Eastern Restaurant | Pub                   | Park                  | Gastropub             | Thai Restaurant       | Coffee Shop           | Office                | Fast Food Restaurant  | Grocery Store          |
| 479 | Upper Rusley   | 020,01689 | TQ4970      | 51.412110 | 0.151023  | 0              | Gym / Fitness Center  | Golf Course               | Italian Restaurant    | Restaurant            | Convenience Store     | Event Service         | Coffee Shop           | Fast Food Restaurant  | Park                  | Paintball Field        |

Of course, our model can be improved somehow by adding more data or constraints to it. For example, in our case, we didn't have any constraint about the budget we can use or to look for safer area or even richer neighborhood. So, by finding first data concerning commercial-rental prices or data with crime rates per neighborhood or data with average income per neighborhood, we could target safer and cheaper in rents neighborhoods (in conjunction with gyms and fitness centers as common venues) or target middle-class people (by using average income). And, therefore, our model could be more accurate and , maybe, effective. At any case, the most important thing is **DATA** and how well we can manipulate them.

## 5. Conclusion

Our goal was to find some very good locations for a well-known Asian healthy Food and Supplements Stores-chain, which wants to expand its businesses in Europe by opening a new store in London. Those locations were places where most common venues were Gyms and Fitness centers, as people going there are more prone to buy our providing products. We started by gathering data concerning neighborhoods of London and venues belonging to them. After a pre-processing session, we used K-Means algorithm to cluster neighborhoods by their most common venues. We, finally, found which locations are ideal, always based on neighborhood's venues occurance. In discussion segment, some ways to improve the existing model were analyzed, concluding that the most significant Tool in our days is DATA and by managing it very well, we can achieve great things at any king of industry.