Best Location for Gym Enthusiast in London

Clustering London Location with Most Number of Gym in Surrounding Area

Husen Wahyu Adi
June 14, 2021

1. INTRODUCTION

1.1 BACKGROUND

1.1.1 Body Building

Body Building is the process of putting on muscle through working out and dedicated yourself to a steady routine. It also involves constructing one's diet to help the muscle create mass and instigate muscle growth. A type of person who decides to begin bodybuilding must be very disciplined and commit to an exhausting workout regiment, and to change their eating habits consistently each day. A person who is just beginning to commit to becoming a bodybuilder should always seek out professional advice. Some will consult with a nutritionist for the food side of the bodybuilding program, and then consult a personal trainer for the physical side of the process. Both will help you keep safe during the bodybuilding process while you increase muscle mass and increase the potential for success. To properly build your body and increase muscle size, professional bodybuilders have used three main strategies to maximize their muscles: Strength Training, Specialized Nutrition, and Adequate Rest.

There are two main factors of successful bodybuilding: workout consistently and consume good nutrition. As we can see, people who want to begin doing body building have to do workout routine to get the body they want. As well as people who has been on body building for a long time have to maintain their workout routine to stay in shape.

1.2 PROBLEM

To maintain their routine, instead of doing workout in their own home, it is likely to be more effective for them to do their work out in the gym / fitness center. If they go to gym/fitness center, not only they can get the best equipment they can use, but they can also meet a personal trainer or other bodybuilder to help their bodybuilding progress. Unfortunately, they can be reluctant to go to the gym regularly (especially for beginner) because of:

a) The gym is too far away, and
b) There are very little options of another gym, so if they do not like some gyms near from their home they does not have other options as alternative.

1.3 IDEAS

We are going to help the body building enthusiast, the beginners, and the experts, to choose the best place for them to live in a city, in this case is London Area. We will find the areas in London where have many options of gym around its small surrounding area. The methodology is as follow:

a. Retrieve nearby gym venues on each London Borough to be analyzed.
b. Cluster gym venues using Density-Based Spatial Clustering of Applications with Noise (DBSCAN).
c. Analyze the clusters whether it is already show the most dense gym venues in small area
d. If clusters is too large, find the subcluster to obtain the most suitable area which surrounded by many gym venues

2. DATA ACQUISITION AND DATA CLEANING

2.1 DATA TO BE USED

1. List of London Borough
   This data can be retrieved from Wikipedia page of London Boroughs, in here. Data can be collected using BeautifulSoup Python Library. From this data, data of borough name will be used for searching the coordinate of borough, and then finding the nearby gym venues.
2. Coordinate of Borough
   Coordinate data (longitude and latitude) of each borough can be retrieved with Geopy Python library. This Data will be used for searching nearby gym venues around the borough area.
3. Gym Venues
   This data can be collected from Foursquare API. Data of Gym Venues is the most important data to cluster the gym venues based on their location.

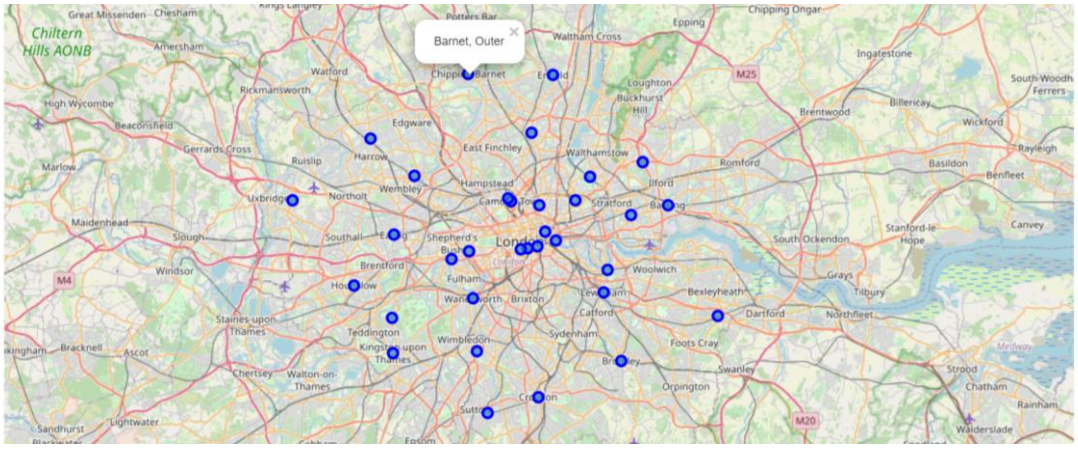
2.2 DATA CLEANING

1. List of London Borough

List of London Borough downloaded or scraped from Wikipedia page of London Borough. This data can be retrieved by BeautifulSoup Python Library. After the table is downloaded, there are some minor problems of this dataset. First, some borough name has additional string which is correspond to Wikipedia notes of information about the specific borough. This problem can be handled easily by deleting the specific additional string on some borough name data. The second problem is city of London area did not listed on the Wikipedia table. So, the city of London area is added to the dataframe manually. We have 32 names of borough plus one city of London area.

2. Coordinate of Borough

Second data to be retrieved is borough location data, which is coordinate of each borough. This borough coordinate data can be downloaded using GeoPy Python Library. The GeoPy Library assign the borough longitude and latitude data based on the right borough name. From the process of retrieving, there are no problems on the process of retrieving. After the borough location data is added to the dataframe, we can check the accuracy of location data by plotting it using folium map.



3. Gym Venues

The location data of each borough has been added to the dataframe. The next data preparation process is retrieving gym venues from each borough. This data is downloaded using Foursquare API. The Foursquare API calls are specified only on category id parent of gym / finess center, which is '4bf58dd8d48988d175941735'. After the downloading process from Foursquare API, we can collect 1032 venues.

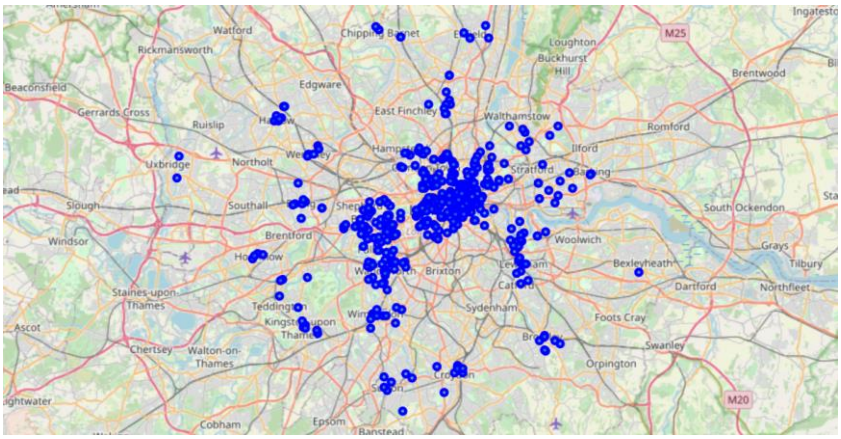
There are some problems that have to be handled from this gym venue data. First, the venue data still contain some category Id of venues that are categorized as gym / fitness center, but not for bodybuilding purposes, such as gymnastic gym, climbing gym, boxing gym, etc. We only select the category of gym and fitness center and drop the row of unmatching category. After this process, the data of gym venues that still on the data frame is 828 rows.

Second process of data cleaning of this gym venue data is deleting duplicate data. Because of the retrieving venues from foursquare using parameter of each borough location, the venues can be assigned on two or more borough. Duplicates can be found by searching the rows which have the same venue name, venue latitude, and venue latitude. After this process, the remaining rows that remains on the dataset are 541 rows.

The problem after deleting duplicates is the probability remaining borough did not assign the nearest borough. Because of the process of deletion is keeping the first unique data from dataset, it only keeps the venues on the data frame of the first letter alphabet borough name. Such as, the gym venues between Barking and Newham firstly retrieved on those two boroughs. But because the duplicate deletion process, the same venue data on Newham borough is deleted. But the deleted venue is the gym venue that is nearer by Newham borough, so it should be assign to the Newham borough, instead of Barking. So, to overcome this problem is reassigning all remaining gym venues to its nearest borough. We measure the distances of each venue to each borough, using coordinate data. After the nearest borough is collected, the old borough data is deleted and replaced by the new nearest borough data. After this process, the gym venues data also visualized on map to comprehend the spread of gym locations in London.

	Venue	Venue Latitude	Venue Longitude	Venue Category	Borough
0	PureGym	51.539250	-0.143077	Gym / Fitness Center	Camden
1	Barry's Bootcamp	51.527075	-0.131056	Gym / Fitness Center	Camden
2	PureGym	51.554052	-0.144984	Gym / Fitness Center	Havering
3	Urban Kings	51.531300	-0.121950	Gym / Fitness Center	Camden
4	Somers Town Community Sports Centre	51.532768	-0.133157	Gym / Fitness Center	Camden
...	...	...	...	...	...
536	Blitz CrossFit	51.448928	-0.332068	Gym	Richmond upon Thames
537	Go-Gym	51.360355	-0.195039	Gym	Sutton
538	Leyton Leisure Centre	51.573975	-0.010304	Gym	Waltham Forest
539	Shoreditch House Gym	51.523687	-0.076177	Gym	City of London
540	Montcalm Gym	51.520826	-0.091617	Gym	City of London

541 rows × 5 columns



3. EXPLORATORY DATA ANALYSIS (METHODOLOGY & RESULTS)

3.1 METHODOLOGY

In order to find the area in London which has most number of gym in their surroundings, we could use Density Based Clustering Algorithm, in this case Density-Based Spatial Clustering of Applications with Noise (DBSCAN) Method. Density-Based Clustering refers to unsupervised learning methods that identify distinctive groups/clusters in the data, based on the idea that a cluster in data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a base algorithm for density-based clustering. It can discover clusters of different shapes and sizes from a large amount of data, which is containing noise and outliers. The DBSCAN algorithm to cluster the dataset are as follows.

1. MinPts and eps are determined.
    a. minPts: The minimum number of points (a threshold) clustered together for a region to be considered dense.
    b. eps (ε): A distance measure that will be used to locate the points in the neighborhood of any point.
2. A starting point is selected at random at it's neighborhood area is determined using radius eps. If there are at least minPts number of points in the neighborhood, the point is marked as core

point and a cluster formation starts. If not, the point is marked as noise. Once a cluster formation starts (let us say cluster A), all the points within the neighborhood of initial point become a part of cluster A. If these new points are also core points, the points that are in the neighborhood of them are also added to cluster A.

3. Next step is to randomly choose another point among the points that have not been visited in the previous steps. Then same procedure applies.
4. This process is finished when all points are visited.

By applying these steps, DBSCAN algorithm is able to find high density regions and separate them from low density regions. In the case of finding the area with most number of gym in their surroundings, DBSCAN could cluster the area in London based on its density of gym venues in those area.

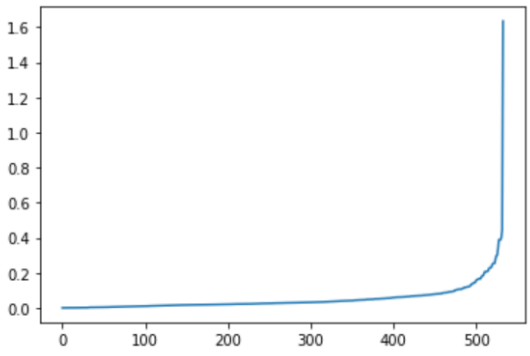
3.2 CLUSTERING GYM DATA USING DBSCAN

3.2.1 Finding Gym Cluster Using DBSCAN

a. Preparation

To cluster the gym venues, we select the features to be processed. Because the goal of clustering is to get the most densely populated area of gym venues based on its location, we select the location data of venues. The selected features are Venue Latitude and Venue Longitude. Another important step of preparation to cluster the data is Data Normalization. We use the StandardScaler from scikit learn preprocessing library to normalize the data.

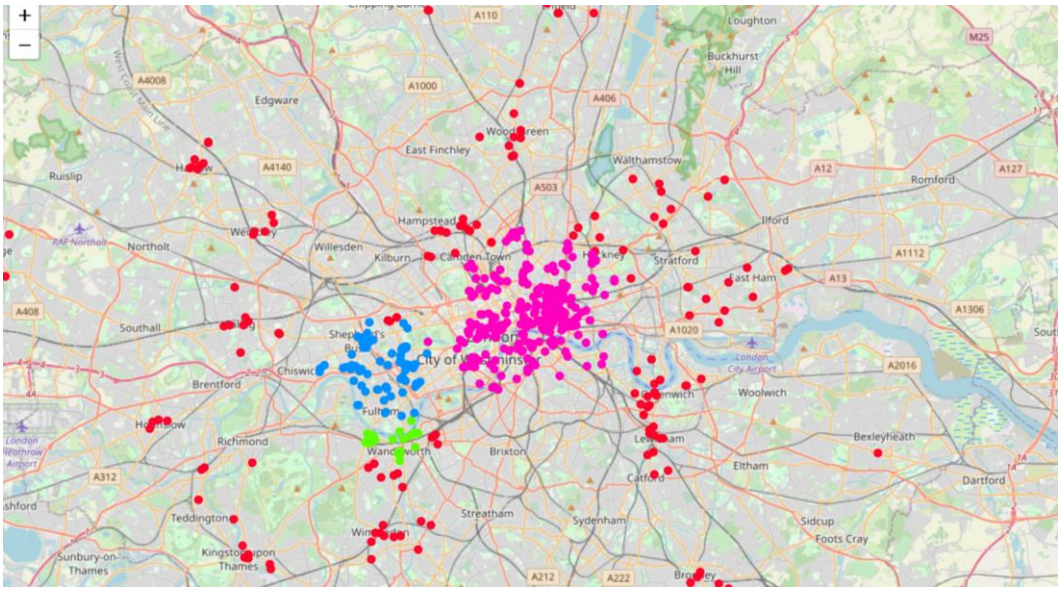
As we mention above in Methodology section, two main parameters on clustering using DBSCAN are minPts and eps. The minPts value that we select for this clustering is 20. It means, that the minimum number of gym venues neighbors on a given point in order to be classified as a core point are 20 gym venues. To determine the optimum eps value, we could calculate the distance to the nearest n points for each point, sorting and plotting the results. Then we look to see where the change is most pronounced (think of the angle between your arm and forearm) and select that as epsilon. By using this method, we determine that the optimum value of epsilon is 0.25.



The Optimum Epsilon is around 0.25

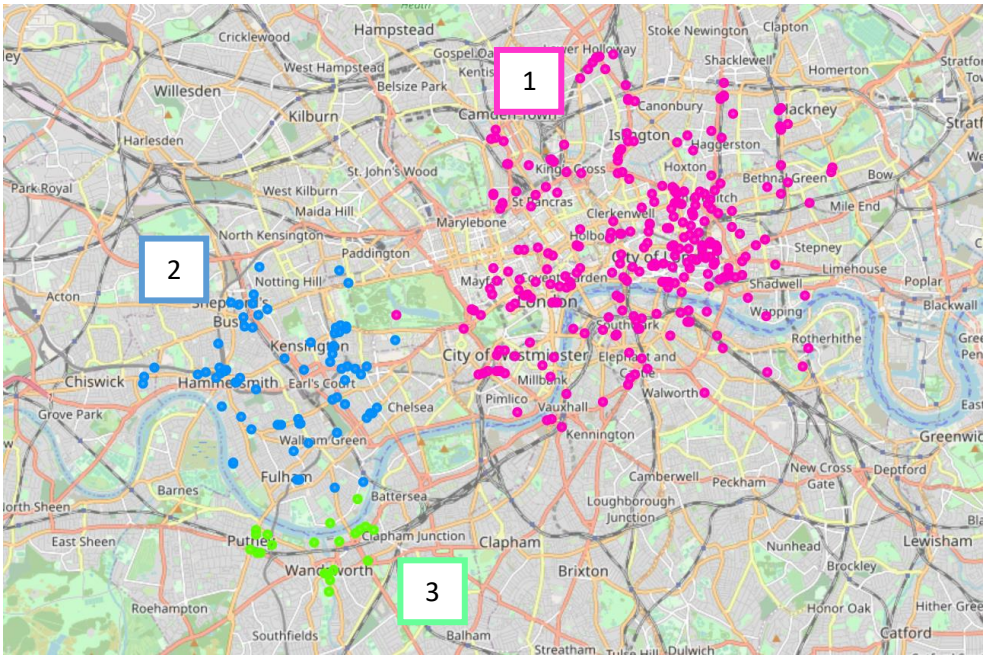
b. Clustering Result

After the preparation and determining the parameter for classification, the DBSCAN clustering method can be deployed. The resulting clusters from this method are 3 clusters. Those clusters are spreading on City of London, Camden, Hammersmith, Kensington, Wandsworth and surrounding area. The area on outer London seems does not meet the requirement to have minimum gym venues on determined radius on surrounding area.



Cluster Recap Table

Cluster	Cluster 1	Cluster 2	Cluster 3	No Cluster (Outlier)
Color	Pink	Blue	Green	Red
Number of Venues	273	73	21	174
Most Area Covered	City of London, Westminster, Tower Hamlets, Islington	Hammersmith and Kensington and Chelsea	Wandsworth	Outer



c. Analysis

From the clustering method using DBSCAN, we already found some areas in London which has higher density of gym venues from others. Gym venues that are not assigned to a cluster did not meet the requirement of clustering, which are number of neighbor gym venues in specified epsilon value. In other word, the area of gym venues that are not assigned to a cluster did not dense enough to be clustered and recommended to a body builder. From the result above, all cluster are located in inner London Area, so we can conclude that the inner area of London has higher density of gym venues than outer area of London.

The clustering result seems successfully separate groups of gym locations from others. But it is still didn't answer our problem to find the most densely area of gym venues in London in narrow area. The 3 clusters are still wide enough to be recommended.

In other way, we already know which area on greater London, which has higher density of gym venues from others. But the resulting dataset could not specify the little neighborhood area of the most densely populated area by gym venues in London. So, we should deploy next clustering method to find the subcluster in the clusters to identify the dense narrow area of gym venues.

3.2.2 Finding Gym Sub Cluster Using DBSCAN (Gym Clustering Level 2)

The outcome of first clustering has successfully produce some clusters of gym venues, but the clusters output didn't satisfy our requirement. To solve this problem, we should deploy DBSCAN clustering level 2, which is clustering the gym venues using the same method of first clustering (DBSCAN) from the clusters dataset. So, instead of using the whole gym venues location we've already retrieved from foursquare, we only find cluster from gym venues which are assigned to clusters in first clustering. This process hopefully could obtain the expected result, which is the narrow area in London which has the most density of gym venues on its surrounding area.

a. Data Selection and Preparation

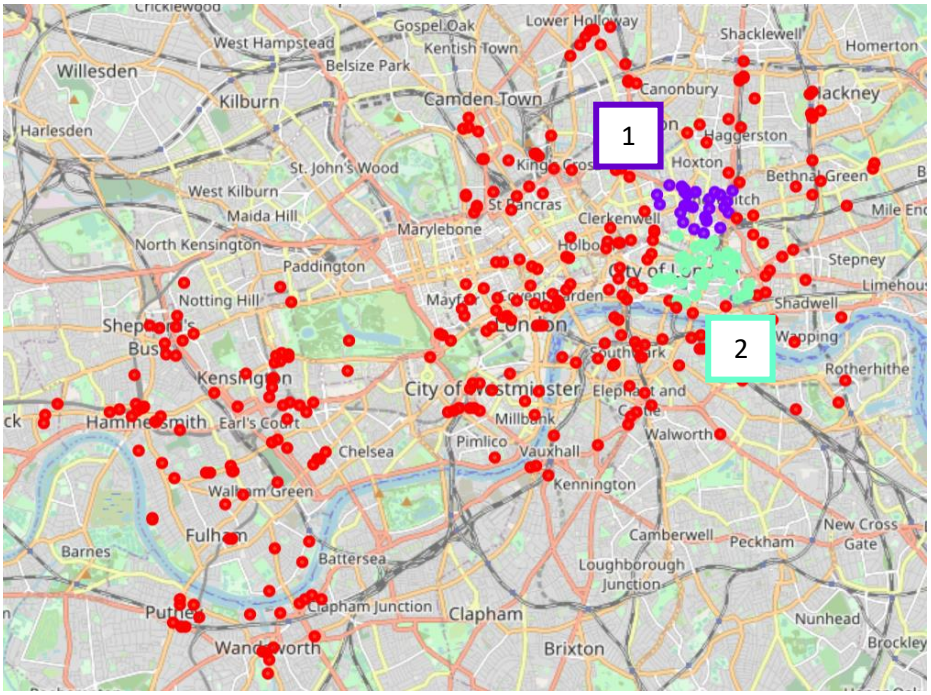
As we mention above, this clustering procedure only implemented on gym venues dataset which assigned to a cluster in first clustering process. The dataset to be processed are combined dataset form cluster 1, cluster 2, and cluster 3. Total gym venues to be clustered are 351 gym venues.

The next preparation method to find do the level 2 clustering is the same as the preparation of first level clustering. First, we select the location data of venues, which are Venue Latitude and Venue Longitude. After that we normalize the dataset using standardscaler. Then, the clustering parameter is determined. The minPts value that we select for this clustering is 15. The optimum epsilon value that is selected from the same method as first clustering is 0.15.

b. Clustering Result

The resulting clusters from this level 2 clustering method are 2 Sub clusters. Those clusters are spreading on City of London and Tower Hamlets area. All of this subcluster are located on cluster number 1. The gym venues on cluster 2 and cluster 3 didn't densely enough compared to the subclusters on cluster 1.

There are many gym venues in cluster 1 area outside the subcluster area which also didn't meet the requirement to be subclustered. So gym venues in cluster 1 outside subcluster and all gym venues on cluster 2 and cluster 3 area considered to be outlier in this level 2 clustering process.

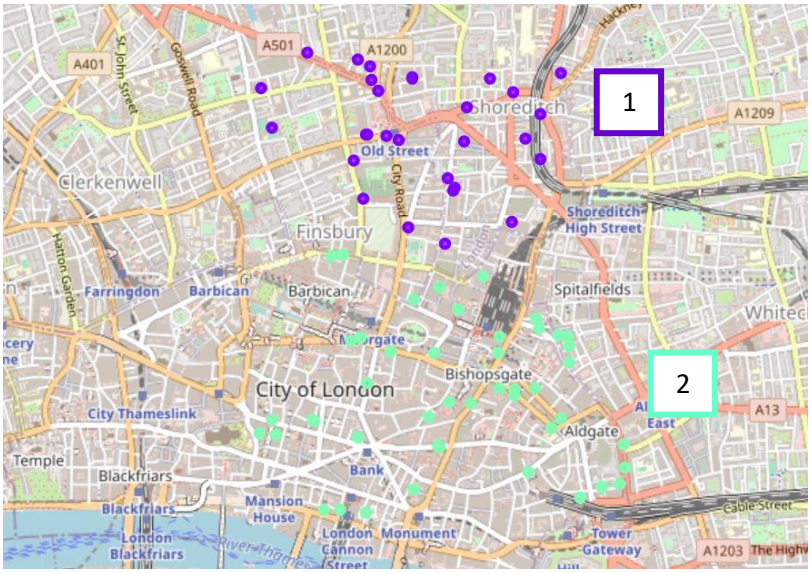


Sub Cluster Recap

Cluster	Cluster 1	Cluster 1	Cluster 1,2,3
Sub Cluster	Sub Cluster 1	Sub Cluster 2	No Sub Cluster
Color	Purple	Green	Red
Number of Venues	29	46	292
Most Area Covered	City of London	Tower Hamlets	Other

c. Analysis

This level 2 clustering method could obtain more narrow clusters compared to the first clustering. All of subclusters are located in the cluster 1 area, which means the gym venues in cluster 2 and cluster 3 didn't densely enough to be recommended. Subcluster 1 and Subcluster 2 are located in nearby area. Subcluster 1 are consisted of 29 gym venues and subcluster 2 are consisted of 46 gym venues.



Although this subclustering result are already show the more densely gym venues in narrower area, we could try to do the same clustering method of this subclusters, which means the level 3 clustering of gym venues to find whether there are some specific location that is more dense compared by other location in those two subclusters.

3.2.3 Finding More Specific Gym Sub Cluster (Gym Clustering Level 3)

The outcome of second clustering has succesfully produce two narrower clusters of gym venues, but we want to try the next level clustering in order to get the more specific gym cluster location. As well as the second level clustering, we do the clustering method of the gym venues using the DBSCAN from the subclusters dataset. Using the idea of second level clustering, in the third level clustering we only find cluster from gym venues which are assigned to subclusters in second clustering. This process hopefully could obtain the expected result, which is the narrowest area in london which has the most density of gym venues on its surrounding area.

a. Data Selection and Preparation

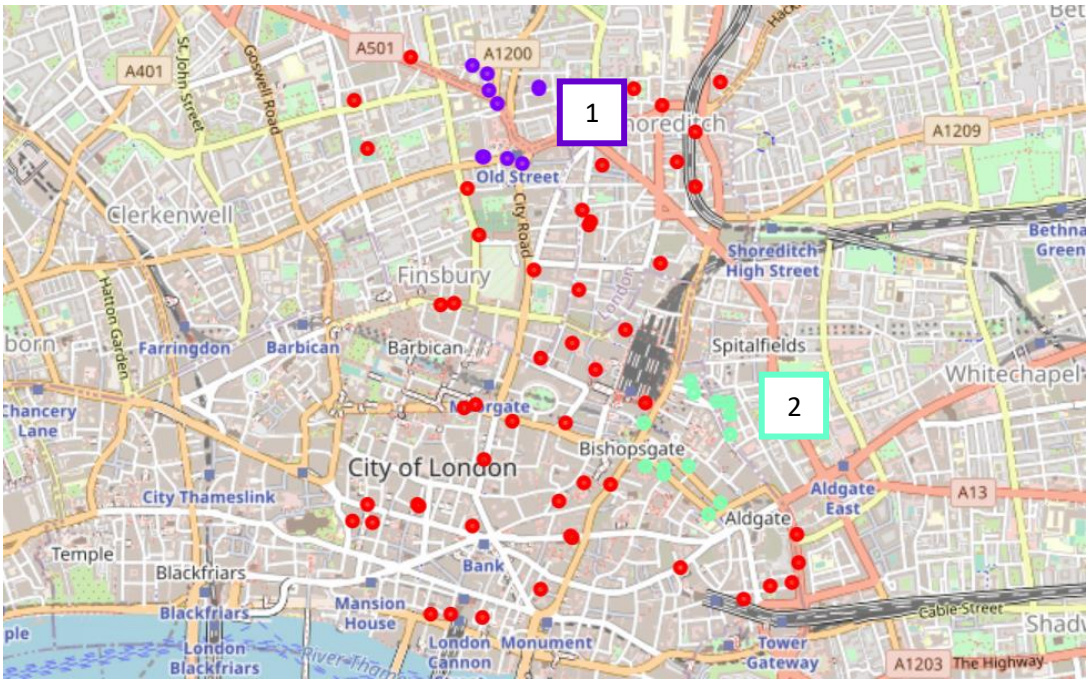
This clustering process only conducted on gym venues dataset which assigned to a subcluster in second clustering process. The dataset to be processed are combined dataset form Subcluster 1 and Subcluster 2, which are part of Cluster 1. Total gym venues to be clustered based on its location are 75 gym venues.

The next preparation method to find do the level 3 clustering is the same as the preparation of previous level clustering. First, we select the location data of venues, which are Venue Latitude and Venue Longitude. After that we normalize the dataset using standardscaler. Then, the clustering

parameter is determined. The minPts value that we select for this clustering is 10. The optimum epsilon value that is selected is 0.5.

b. Clustering Result

The resulting clusters from this level 3 clustering method are 3 Sub Sub clusters. Those clusters are spreading on City of London (Specifically Old Street Station) and Tower Hamlets (Specifically around Bishopgate) area. Each Sub Sub Cluster repesents the most dense area of each Sub Cluster. Which means, Sub Sub Cluster 1 is the most densely populated gym venues area of Sub Cluster 1 and Sub Sub Cluster 2 is the most densely populated gym venues area of Sub Cluster 2. The area on Subcluster 1 and Subcluster 2 that is not assigned to Sub Sub Cluster is considered as not dense enough compared to the clustered areas.



	Sub Sub Cluster	Sub Cluster	Cluster	Venue	Venue Latitude	Venue Longitude	Venue Category	Borough
0	0	0	0	Gymbox	51.525592	-0.089320	Gym / Fitness Center	City of London
1	0	0	0	Ravercise Ltd	51.527277	-0.088667	Gym / Fitness Center	City of London
2	0	0	0	Outrivals	51.527736	-0.086560	Gym / Fitness Center	City of London
3	0	0	0	Beth Lavis Fitness	51.525530	-0.088180	Gym / Fitness Center	City of London
4	0	0	0	London Fight Factory	51.528507	-0.089985	Gym	City of London
5	0	0	0	Britannia Building Gym	51.528238	-0.089165	Gym	City of London
6	0	0	0	M by Montcalm Fitness	51.527703	-0.089121	Gym	City of London
7	0	0	0	OBSESSIVE GYM DISORDER	51.527809	-0.086541	Gym	City of London
8	0	0	0	Crossfit City Road	51.525554	-0.089410	Gym	City of London
9	0	0	0	Bezier Apartments Gym	51.525361	-0.087333	Gym	City of London

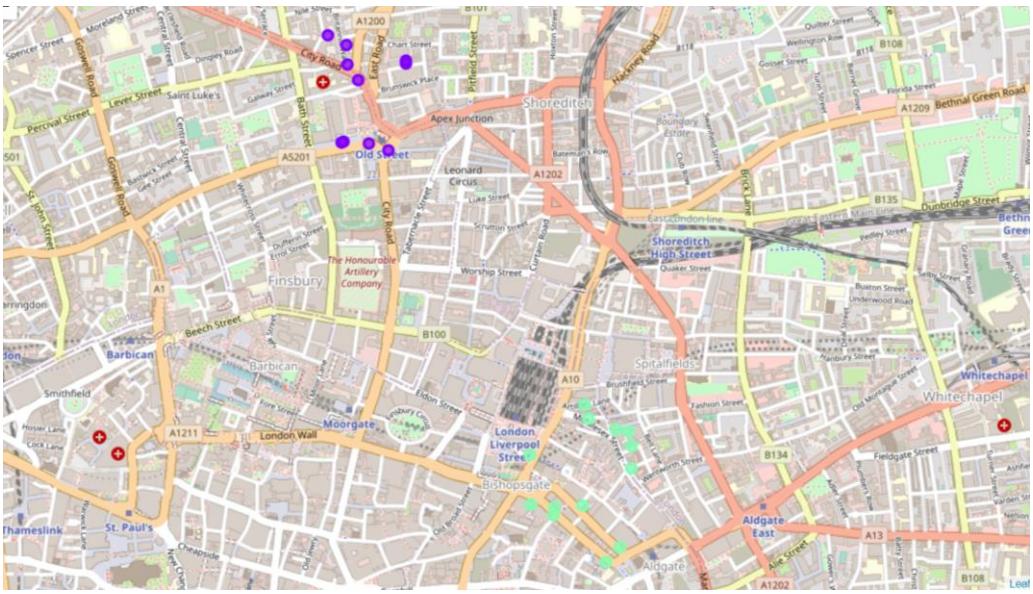
The Sub Sub Cluster 1 is located on City of London around Old Street Station. This Sub Sub Cluster consist of 10 gym venues around that area.

	Sub Sub Cluster	Sub Cluster	Cluster	Venue	Venue Latitude	Venue Longitude	Venue Category	Borough
0	1	1	0	1Rebel	51.515569	-0.080040	Gym / Fitness Center	City of London
1	1	1	0	PureGym	51.514475	-0.077173	Gym / Fitness Center	Tower Hamlets
2	1	1	0	Crossfit Aldgate	51.515630	-0.078790	Gym / Fitness Center	Tower Hamlets
3	1	1	0	F45 Liverpool Street	51.516653	-0.076648	Gym / Fitness Center	Tower Hamlets
4	1	1	0	DW Fitness First	51.517709	-0.077321	Gym / Fitness Center	Tower Hamlets
5	1	1	0	Roar Fitness	51.514090	-0.077761	Gym	Tower Hamlets
6	1	1	0	No1 Studio Training	51.515372	-0.080063	Gym	City of London
7	1	1	0	Equinox Bishopsgate	51.515655	-0.081053	Gym	City of London
8	1	1	0	IgnitePT	51.517270	-0.076664	Gym	Tower Hamlets
9	1	1	0	Sky Gym - Nido Spitalfields	51.517668	-0.076767	Gym	Tower Hamlets
10	1	1	0	Andaz Health Club	51.517024	-0.081103	Gym	City of London
11	1	1	0	Ten Fitness Studio	51.517999	-0.078537	Gym	City of London
12	1	1	0	Foundry:City	51.518371	-0.078728	Gym	City of London

Sub Sub Cluster 2 is located around Bishopgate area, which is between City of London and Tower Hamlets. This Sub Sub Cluster consist of 13 gym venues around the area.

c. Analysis

As we expected, this level 3 clustering method could obtain more narrow clusters compared to the second clustering. Each Sub Sub Cluster represents the most densely populated by gym venues area of each Sub Cluster. Sub sub cluster 1 is Consisted of 10 gym venues and Sub sub cluster 2 is consisted of slightly higher number of gym venues, which are 13.



Cluster	Cluster 1	Cluster 1
Sub Cluster	Sub Cluster 1	Sub Cluster 2
Sub Sub Cluster	Sub Sub Cluster 1	Sub Sub Cluster 2
Color	Purple	Green
Number of Venues	10	13
Most Area Covered	City of London around Old Street Station	Around Bishopgate, between City of London and Tower Hamlets

This level 3 clustering method has produced satisfying outcome to get the most densely populated area by gym venues in London. The retrieved cluster area are City of London around Old Street Station and Bishopgate between City of London and Tower Hamlets. But, when we are trying to specify one of the most densely populated area of gym venues, we would choose area of Bishopgate between City of London and Tower Hamlets. This consideration is based on the number of venues in Sub Sub Cluster 2 is higher than the number of gym venues in Sub Sub Cluster 1.

4. CONCLUSIONS

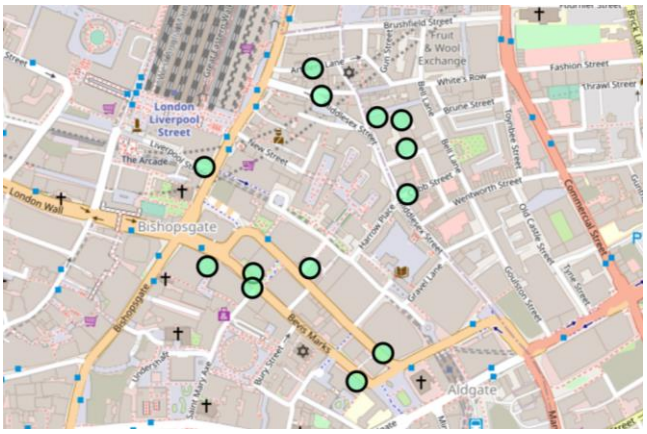
<table>
<tr><th>Clustering</th><th colspan="7">Output</th></tr>
<tr><td>Clustering Level 1</td><td colspan="4">Cluster 1</td><td>Cluster 2</td><td>Cluster 2</td><td>Outlier</td></tr>
<tr><td>Clustering Level 2</td><td colspan="2">Sub Cluster 1</td><td colspan="2">Sub Cluster 2</td><td>Outlier</td><td>Outlier</td><td>Outlier</td><td>-</td></tr>
<tr><td>Clustering Level 3</td><td>Sub Sub Cluster 1</td><td>Outlier</td><td>Sub Sub Cluster 2</td><td>Outlier</td><td>-</td><td>-</td><td>-</td><td>-</td></tr>
</table>

DBSCAN Clustering Method has been successfully extract some clusters of areas in London based on gym venues location data. In order to get the specific location on the most densely populated area by gym venues, we deploy 3 levels of DBSCAN Clustering method. The higher level of DBSCAN Clustering, is conducted to find the subclusters from retrieved clusters of previous clustering method.

From the 3 levels of DBSCAN Clustering results based on only location data of gym venues, the recommended area for body builders to live in London are as follows.

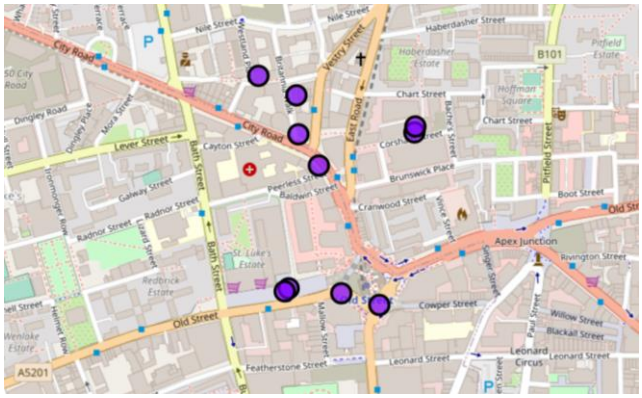
1. Sub Sub Cluster 2 (around Bishopgate, which is between City of London and Tower Hamlets)

   The most recommended area for body builders to live in London is this area. The area consists of 13 gym venues on its surrounding. Which means, that if the body builders got bored or not liking one of the nearby gym venues, there are still 12 other gym venues to try on.



2. Sub Sub Cluster 1 (City of London, around Old Street Station)

   This area consists of 10 gym venues on its surrounding. Which means, that if the body builders got bored or not liking one of the nearby gym venues, there are still 9 other gym venues as other alternative. This is the second recommended area for body builders to live in London in this area.



5. DISCUSSION AND RECOMMENDATION FOR FURTHER STUDY

The 3 levels of clustering method could obtain the specific location of the most densely populated area by gym venues in London. This result is satisfying enough as supporting data in order to recommend someone who has passion of body building finding the best location for him to live. However, this study is not including some of other component that could become consideration of finding place to live, such as housing price, public transportation access, etc. We hope that if someone want to conduct this kind of study in higher level which including more variables, those data (housing price, public transportation data, etc) could be included to the study in order to get more satisfying result.

References:

- IBM Professional Data Science Course
- https://bodyspartan.com/blogs/all-articles/what-is-bodybuilding
- https://en.wikipedia.org/wiki/London_boroughs
- https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html