Using location data to find profitable locations to setup Healthy Food Centers in New York city

Gourab Ghosh

16 August, 2019

1. INTRODUCTION

1.1 Background

Technology has taken over literally everything in every aspect of our life. It has made our lives easier, better and more enjoyable. People are no longer required to put on gruesome efforts to get things done. Do you remember the last time when you sweat profusely while doing some household or professional activity? For most of us the answer would be a big NO. And with all these leisure and comfort, technology has also brought numerous health problems due to lack of physical activity. Another important aspect to be mentioned here is the damage that unhealthy food habits have inflicted upon us. Both these factors have led to many health problems.

In recent years, people have been able to understand the ill-effects of comfort which technology has provided and they have got conscious about their health and are determined to stay fit and healthy. This is why we have seen a tremendous increase in the number of people enrolling for fitness activities like Gym or Yoga.

All these has led to a steady increase in the demand for healthy foods. Off late, there has been a realisation across the globe, with respect to health and wellness. Most people these days make it a point to invest their time and energy into staying fit and eating right. The global wellness industry grew 12.8 percent from 2015-2017, from a 3.7 trillion dollar to a 4.2 trillion dollar market. To put that in the economic context, from 2015-2017, the wellness economy grew 6.4 percent annually, nearly twice as fast as global economic growth (3.6 percent). (source: https://globalwellnessinstitute.org). There is a huge scope and untapped market still to be explored in healthy food sector.

1.2 Problem

Business Ventures which decide to open their chain of healthy food centers across a geographical area has to keep a lot of things in mind and as most of us know, among the 4 P's of business marketing, Place holds a vital role. A perfect location helps the business to grow manifolds, but finding that location is really a challenging task. The location should be easily accessible to the targeted consumer base, otherwise it might not be profitable to run the business. This project aims to predict profitable locations where *healthy food centers* can be opened in New York city.

1.3 Interest

The Business Ventures which are interested in opening healthy food centers or chain of healthy food centers across the city of New York are the prime audience in this case. Additionally, people interested in reality sector, angel investing and data science may also be a suitable audience for this report.

2. DATA

The target customer base for a healthy food center are the fitness conscious people. They are the primary and returning customers and contribute to the majority of the revenue. This is why the store location has to be in the vicinity of these people.

So where can we find these people at a large scale? The answer lies in the fact that these people frequent gym and yoga venues.

2.1 Data Sources

Foursquare is one of the leading location data providers in the world. It not only provides precise location data but also gives related details along with it. We shall be using the Places API provided by Foursquare and fetch venues as per our search terms. For this, one needs to open an account in Foursquare developer space (https://developer.foursquare.com). We have used the free plan for this account.

2.2 Data cleaning and Feature selection

The venue search API of Foursquare returns a list of places around a particular place within a certain radius. We have taken the center around New York city and made the searches. For this search, we have considered the radius of New York city as 20 miles (32187 meters).

In the free plan of Foursquare, a maximum of 50 venues are returned per search. Therefore, we have made two search hits - one for 'Gym' and another for 'Yoga'. Next we have refined each result set to transform the search data into a more readable and analysable format.

As a next step, both the result sets are appended together to form the master location dataset. This dataset still contains a lot of redundant/unnecessary data about the venues which are not required for our study. So, we filter and remove those columns and keep only the name, address and the latitude and longitude columns. The dataset at this stage looks like the following:

	name	**address**	**lat**	**lng**
0	New York by Gehry Gym	8 Spruce St.	40.710655	-74.005709
1	Gym @ Barclay Tower	10 Barclay St	40.712360	-74.009429

3. METHODOLOGY

As stated earlier, the concentration and footfall of health conscious people is expected to be more around the areas where there are fitness centers, gym or yoga studios. There the prospect of a healthy food center is better than other places. People tend to buy food items for consumption after or before their workouts.

3.1 Calculation of target variable

The target variable is the latitude and longitude of the gym or yoga studio venues. We need this sole location data to build our model.

3.2 Clustering of venues

Next, we move on to cluster the venues based on their latitude and longitude. This will be providing a congregation of similar venues which are geographically located near to each other. This ensures maximum density of health conscious people around the area.

k-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function know as squared error function given by:

where, '||xi - vj||' is the Euclidean distance between xi and vj. 'ci' is the number of data points in ith cluster.
'c' is the number of cluster centers.

Algorithmic steps for k-means clustering

Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers.

1) Randomly select 'c' cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4) Recalculate the new cluster center using:

$$\boldsymbol{v}_i = (1 / c_i) \sum_{j=1}^{c_i} \boldsymbol{x}_i$$

where, 'ci' represents the number of data points in ith cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

Here we have taken the number of clusters to be 7 considering the size of the available data points. The number of clusters to be generated can vary depending upon the size of location venues and the business need.

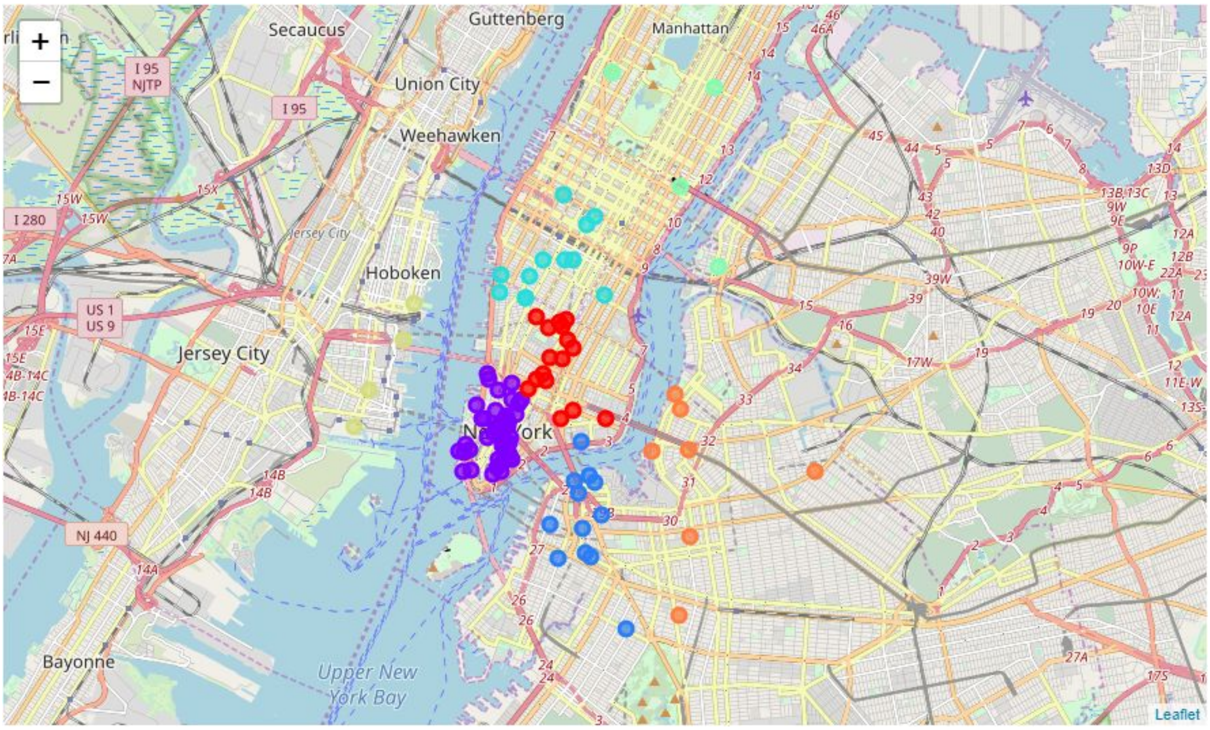
3.3 Plotting on map

We generate a map of New York and on that map, we put all the location dataset members (the fitness venues of gym or yoga) using colored circles. Each cluster points are colored in a unique color to separate it out visually from the other clusters. We can see 7 distinct colored points on the map.

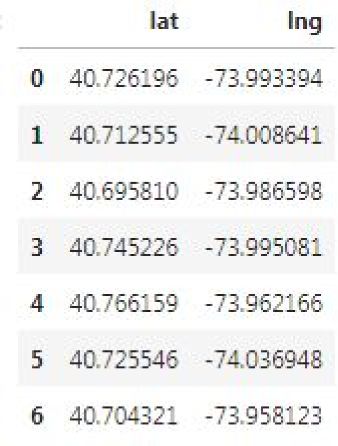
Finally, we get the centroids of each cluster and plot them on the map in black colored circles.

4. RESULTS

After the successful execution of the k-means clustering, we get 7 distinct clusters. The following map shows the same:

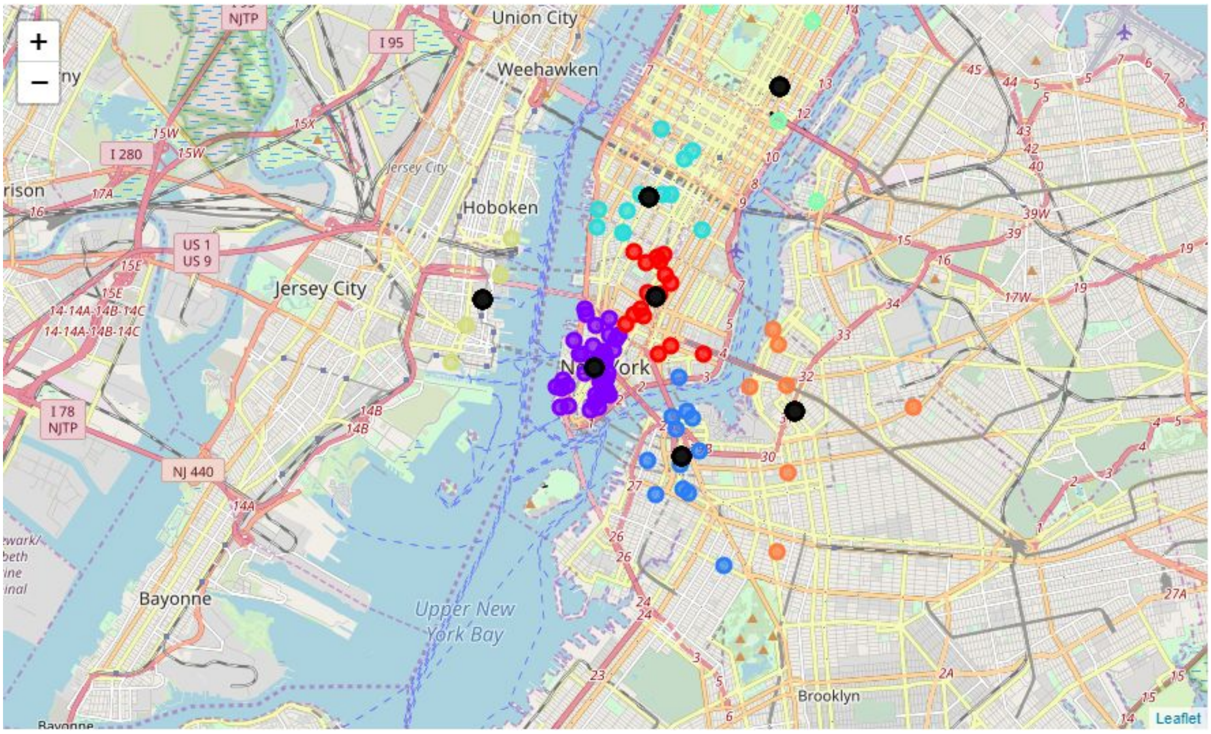


From the k-means clustering, we also get the centroids for these 7 clusters. The location of the centroids are as follows:



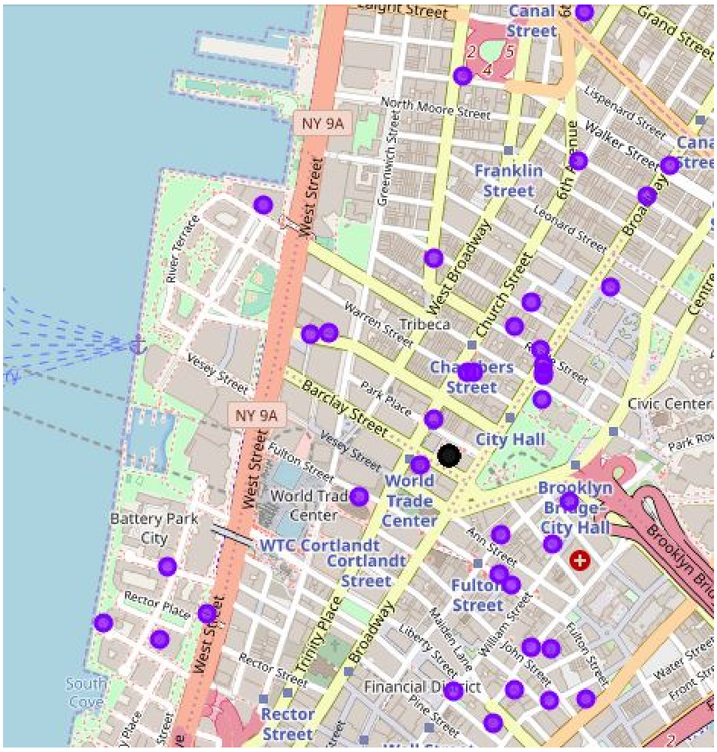
	lat	lng
0	40.726196	-73.993394
1	40.712555	-74.008641
2	40.695810	-73.986598
3	40.745226	-73.995081
4	40.766159	-73.962166
5	40.725546	-74.036948
6	40.704321	-73.958123

Once we have these centroids, we plot them on top of the existing map containing clusters data. The centroids are colored in black circles. The following map shows the same:

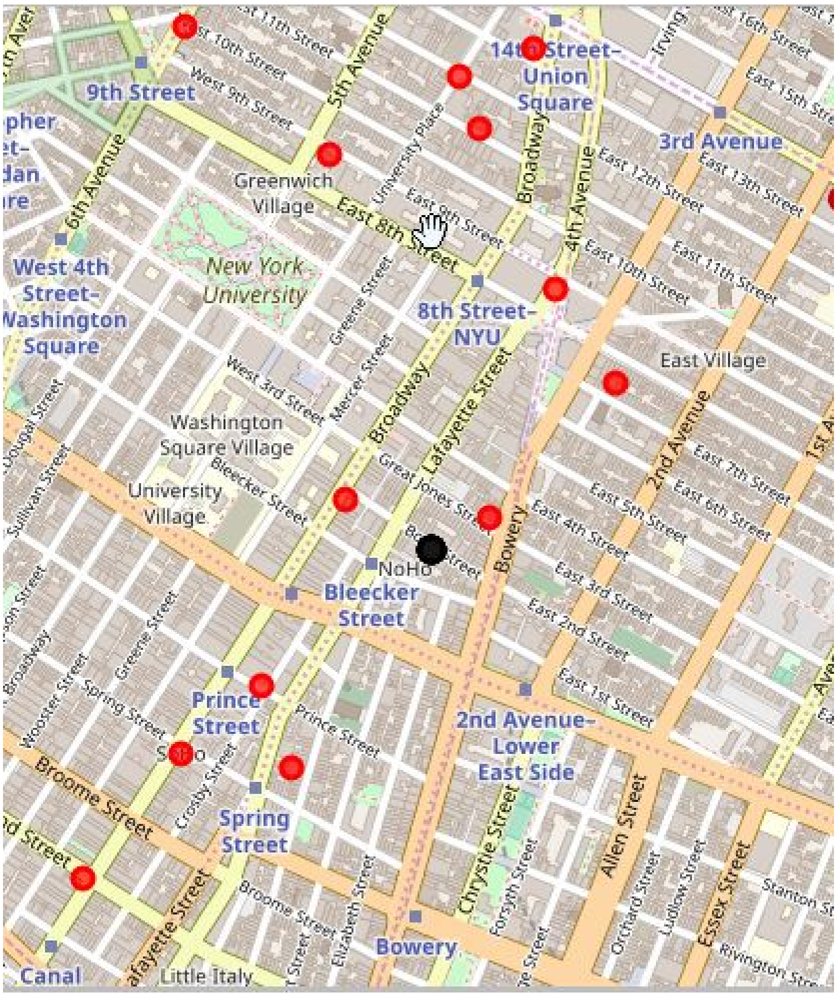


5. DISCUSSION

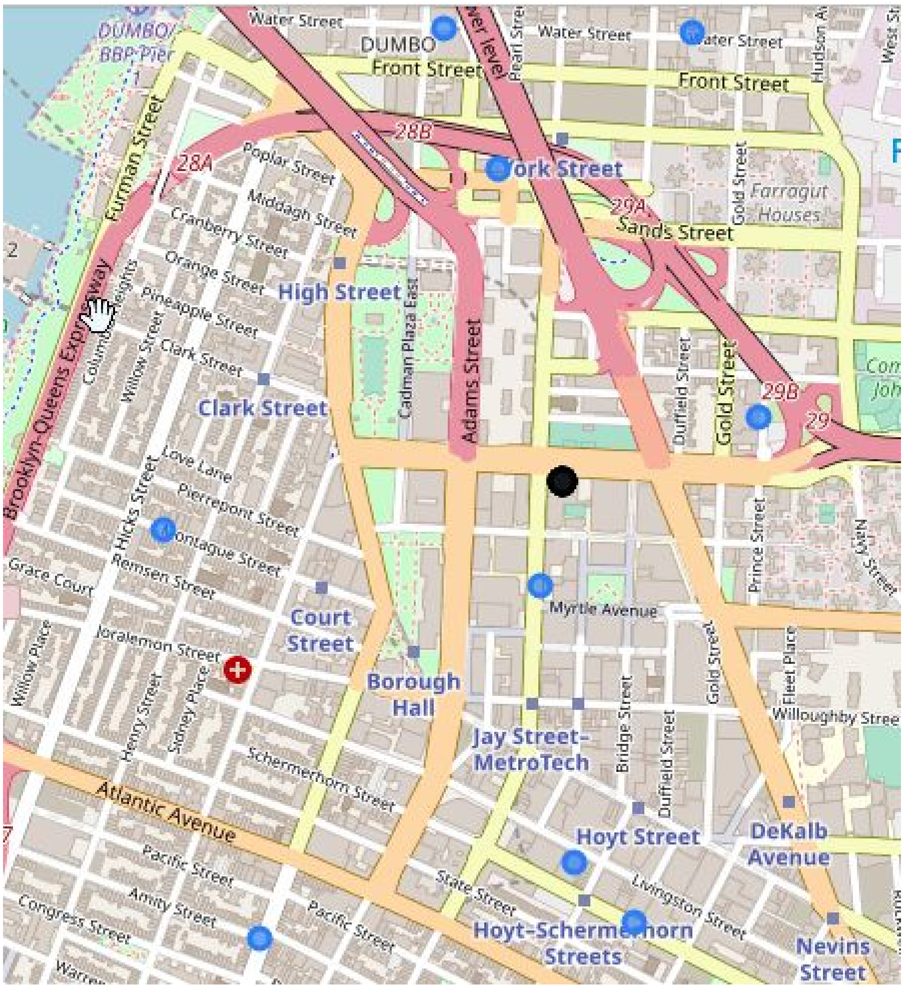
From the map shown above, we can conclude that our clustering mechanism shows lower Manhattan bay area as having the most dense occurrence of gym and yoga centers. The area is colored in violet. The centroid location centered there with co-ordinates as **(40.71255502453713,-74.00864113580623)** near Barclay Street can be the most profitable venue to open a healthy food center. In real life, the venue of the food center can vary based on budget and other constraints to a certain permissible extent.



The second most profitable location is the area around Bond Street with co-ordinates as **(40.72619606949153,-73.99339449287002)** marked by deep orange color in this map. It has density lesser than the first one.



The third most profitable location is the area around Tillary Street with coordinates as **(40.6958098909231,-73.98659838836905)** marked by deep blue color in this map. It has a density less than the first and the second one.



This way, we can see other points also. But the thing is that density falls drastically in the other clusters and the venues are located far away from each other. So, it is doubtful whether opening up a healthy food center will really be helpful in those places or not.

6. CONCLUSION

We can conclude that location data along with data clustering techniques can be a helpful tool in predicting the profitable locations for a healthy food chain stores in a city.

Since the city of New York is very big in size and there are numerous gym and yoga centers, we can improve the efficiency of our method by running the study based on specific boroughs of the city. For example, we can take Manhattan as the base location and by taking its radius, we perform a search for location datasets. This can significantly improve the prediction.

Also, if some business owner is interested in opening up the store in a preferred borough of his/her, we can take the center of our study around that location and make searches.