Salin Dalabehera
IBM Applied Data Science Capstone
May 16, 2020

Opening of Gym in New York City

Coursera Capstone Project

Introduction:

Lifestyle behaviors are the most important factors in short-term and long-term well being. Developing healthy lifestyle habits will have a dramatic effect on overall wellness. Components of a healthy lifestyle include: physical activity, balanced and adequate nutrition, rest, recovery and stress management. Daily physical activity is one of the key components in developing and leading healthy lifestyle. Physical activity improve overall quality of life. It helps boost energy, assists with weight management and improves self-esteem.

"Health is wealth". This is the new mantra of everybody in metro cities now-a-days. Good health can enjoy the pleasure of life. People are more focus toward healthy lifestyle and for that they are willing to spend money for it. Instead of buying equipment for exercise in home, they normally prefer gym in nearby area.

The fitness industry has exploded in recents years. In the U.S., total revenues reached 3.1 billion U.S. dollars. There are 2111 health clubs in New York and increasing every year. In addition, many of the gym and fitness centers employ very little to no staff, which makes startup costs and barriers to entry low. With the proper mix of skills, training and commitment, starting a gym or fitness center can prove a successful business move.

Business Problem:

The objective of this project is to analyze the best location to open a gym/fitness center in New York city. New York has 5 boroughs i.e. Bronx, Queens, Manhattan, Brooklyn and Staten Island. By using the data science methodology and machine learning, we will find the best location to open a gym in any of the borough.

Audiences:

This project can be useful for the investors/new entrepreneur to open a gym in New York area. They can see the ares with high, moderate and low fitness centers and include in their analysis or decision.

Data:

For this project, we need following data to build our model:

* List of neighborhoods of New York city: List of neighborhoods of all the 5 boroughs
* Location coordinates of neighborhood: Latitude and Longitude of the neighborhoods
* Venues of the neighborhoods: Venues data near to the neighborhood, particularly Gym or fitness centers

Sources:

One of the wikipedia page (https://en.wikipedia.org/wiki/Neighborhoods_in_New_York_City) contains the neighborhoods of all 5 boroughs. We will use the python modules "requests" and "BeautifulSoup" to extract the data and save it to a pandas dataframe.

For latitude and longitude coordinates of each neighborhood, we will use python "geopy" package.

After that, we will use Foursquare developer api to extract venues of each neighborhood of New York city. Foursquare has the largest database with 150+millions places. Foursquare api provides many venues with category, which will help us to extract Gym from the result set.

In this project, we will use many tools and methods like web scrapping, Geo coordinates package, API(Foursquare), data cleaning, data wrangling, machine learning (Kmean clustering) and Folium to visualize the locations in map.

Methodology:

We will get the list of neighborhood of New York city from the wikipedia(https://en.wikipedia.org/wiki/Neighborhoods_in_New_York_City). We will use web scrapping method using requests and beautifulsoup package to extract the table with borough and neighborhood. However the neighborhoods are comma separated, so we split the column to multiple rows and boroughs are listed as community board. So we update the rows with common borough name.

Once we have the neighborhood table, we will find the geo coordinates. We will use geopy module to get the latitude and longitude of each neighborhood. As a part of data cleaning process, we will drop rows without latitude and longitude, if any.

Then using foursquare API, we will extract the 100 venues nearby within 500 meters radius. To use the foursquare API, we need register in their developer account and get the ID and secret key. Using this we pass latitude and longitude of each neighborhood and get the list of nearby venues in json format. We will pass the json object and create a table of each neighborhood. We will check the

uniqueness of the venues and calculate the means of the each venues, which will help us to apply the machine learning technique. As we are interested about the Gym only. We will filter the data with column "Gym" only.

Once we have the neighborhoods table with its borough, we will apply KMeans clustering algorithm to each Borough to create 3 clusters. KMeans is the one of the most popular and simple clustering algorithm to use. We will cluster neighborhoods into 3 clusters based on their frequency of the Gym in its Borough. This results will allow us to identify neighbors with most number of Gyms, moderate numbers of the gym and few numbers of Gym. This model will help us to identified the neighborhood based on Borough to open a gym in New York city.

Results:

This model will show us 3 categories of the neighborhood.

* Cluster 0: Neighborhood with low number of Gyms
* Cluster 1: Neighborhood with moderate number of Gyms
* Cluster 2: Neighborhood with high number of Gyms

We can see results of each neighborhood based on borough.

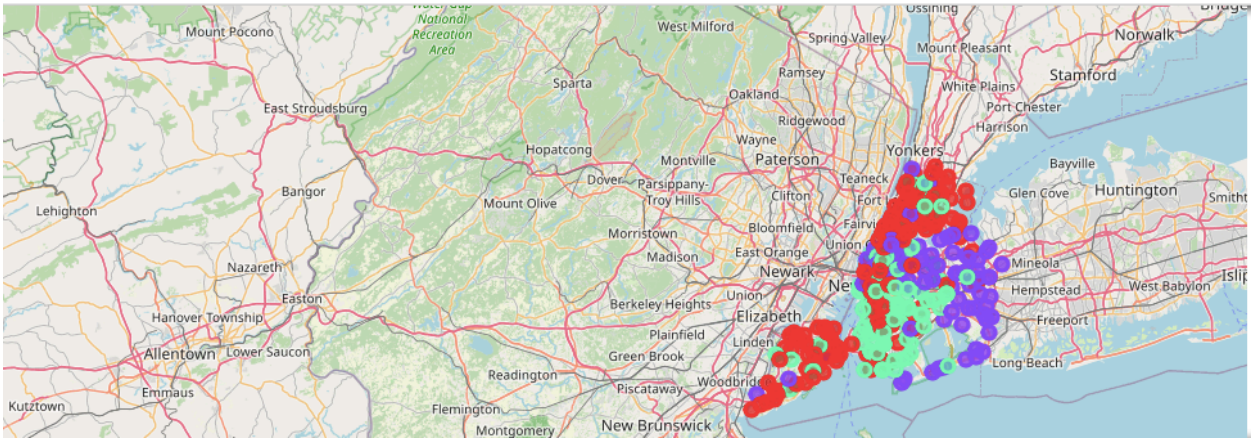
		Neighborhood	Gym	Latitude	Longitude
Borough	**cluster**				
Bronx	**0**	47	47	47	47
	1	3	3	3	3
	2	11	11	11	11
Brooklyn	**0**	32	32	32	32
	1	1	1	1	1
	2	57	57	57	57
Manhattan	**0**	31	31	31	31
	1	9	9	9	9
	2	12	12	12	12
Queens	**0**	6	6	6	6
	1	63	63	63	63
	2	16	16	16	16
Staten Island	**0**	42	42	42	42
	1	4	4	4	4
	2	5	5	5	5

We can visualize the location using folium.

* Red - cluster 0

* Blue - cluster 1

* Green - cluster 2



Discussion:

From the above results, we can see find the places based on the borough. Let's say we want to find places in Queens to open a gym. We can see 64 places with low number of gyms, 15 places with moderate and 7 places with high number of gym. We can select places with moderate number of gyms to consider opening of a gym as completion will be less compared to high concentrated place with cluster 2. We can consider places with low number of gym, to start with few peoples living nearby or without any completions.

One more thing we can consider in this model is the population of the neighborhood. As a new gym can attract in high populated places. This can be a future addition to the project.

Conclusion:

In this project we go through the business problem, data requirement, data extraction, data cleaning and machine learning to find our results. We segment neighborhoods in 3 clusters to identified places with high, moderate and low frequencies. This project may be useful for the entrepreneur/investor to open new gym in the neighborhood of the New York city and also be helpful to identified places with high/moderate/low number of gyms to join or live in their neighborhood.