05/29/2020

BEST GYM LOCATION IN BUENOS AIRES USING I.A.

IBM DATA SCIENCE CAPSTONE

MATHIAS A. HOCHKOFLER

TABLE OF CONTENTS

A. Introduction

A1. Description & Discussion of the Background

I have been acquiring skills related to data science by taking the IBM Data Science Professional Course on Coursera. The last course contains a capstone project. This project is about applying data science toolset and obtained skills to analyze a problem and creating value. My project's theme concerns a topic that I have been really interested in: Gym and health industry. My analysis was performed in Python. The details are pushed to Github.

A2. Business problem

In recent years, there is a great boom in the healthy living industry. She is interested to opening a new unit, which will focus on offering her clients a personalized routine according to their weight, age, expectations and time. Considering the financial plan in which the gym will operate, the intention is to find an optimal location in an area of Buenos Aires. The following criteria should be considered:

- Nearby competitors
- Metropolitan area

The assumption behind the analysis is that we can use unsupervised machine learning to create district groups that will provide us with a list of areas for potential gym locations. The purpose is that the gym is located near one of the most populated areas with less competition and easy access.

A3. Data requirements

To perform this analysis, we will need the following data:

- List of the districts Buenos Aires, Argentina Geo-coordinates of the districts in Buenos Aires Top venues of districts List of districts will be obtained from Wikipedia. (http://download.geonames.org/export/zip/AR.zip)
- Geo-coordinates of districts will be obtained with the help of the geocoder tool in the notebook.
- Top venues data will be obtained from Foursquare through an API.

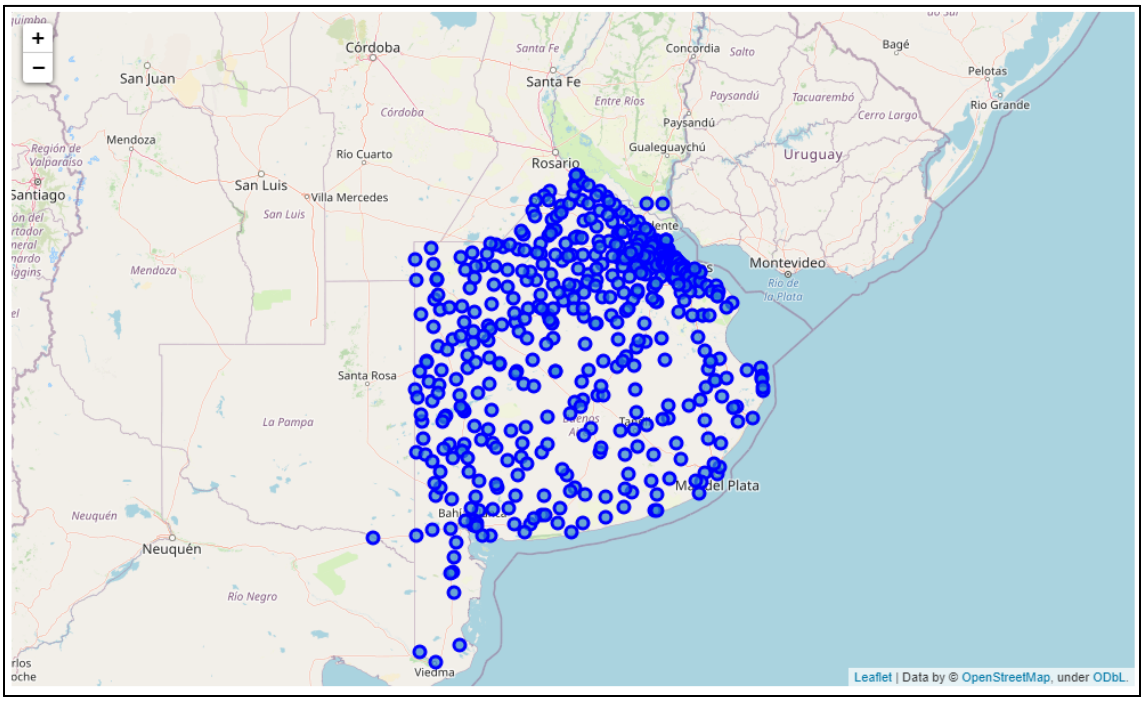
B. Solution

B1. Cluster Neighborhoods by venues frequency

Get Location Data of Buenos Aires

Geonames.org provided us a data set that contains all the postal codes / zip-codes of Argentina, categorized by state and breaking down the coordinates of each point. As shown in the next table and map:

	Zip-code	Latitude	Longitude	Neighborhood	Borough
0	1601	-34.5167	-58.5389	ISLA MARTIN GARCIA	Buenos Aires
1	1602	-34.5167	-58.5000	FLORIDA,PUENTE SAAVEDRA,JUAN B. JUSTO (ESTACIO...	Buenos Aires
2	1605	-34.5333	-58.5500	MUNRO,MUNRO ESTAFETA No.2,CARAPACHAY	Buenos Aires
3	1607	-34.5167	-58.5389	JOSE MARTI,BARRIO OBRERO FERROVIARIO,BARRIO ARCA	Buenos Aires
4	1609	-34.5000	-58.5667	BOULOGNE,BOULOGNE SUR MER,BOULOGNE ESTAFETA No...	Buenos Aires



Note: Buenos Aires – Argentina Map

Explore venues in Buenos Aires

In this way, with a simple filter we got the coordinates to use them as input parameters into Foursquere API and get a response a set of the most relevant venues around each zip-code.

Foursquere provided us 1901 venues related to zip-codes in Buenos Aires which we drop 362 (Foursquare could not find venues related to their respective zip-code) of 548 zip-codes. So, we have 186 zip-code (neiborhoods) with 1539 related venues to analyze.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	1601	-34.5167	-58.5389	Sandwiches Fredy	-34.514734	-58.542054	Sandwich Place
1	1601	-34.5167	-58.5389	Acacia Pastelería	-34.514233	-58.541880	Deli / Bodega
2	1601	-34.5167	-58.5389	La Colón	-34.519342	-58.544130	Bakery
3	1601	-34.5167	-58.5389	La Reja	-34.518106	-58.543510	Ice Cream Shop
4	1601	-34.5167	-58.5389	Retaceria Mary	-34.518633	-58.540269	Department Store
5	1601	-34.5167	-58.5389	Vicente Lopez Futbol	-34.516129	-58.528349	Soccer Field
6	1601	-34.5167	-58.5389	El Retorno	-34.511655	-58.540818	Argentinian Restaurant
7	1601	-34.5167	-58.5389	City Bar	-34.510774	-58.532785	Rock Club
8	1601	-34.5167	-58.5389	Colonial Helados & Cafe	-34.512004	-58.541162	Ice Cream Shop
9	1601	-34.5167	-58.5389	Plaza Almirante Brown	-34.518788	-58.547269	Pedestrian Plaza

Prepare data model

The Sklearn python library (artificial intelligence toolkit) allows us to use clustering models to group the data based on the input we put into the model. These data significantly affect the result of the analysis, so these should be cleaned and normalized property.

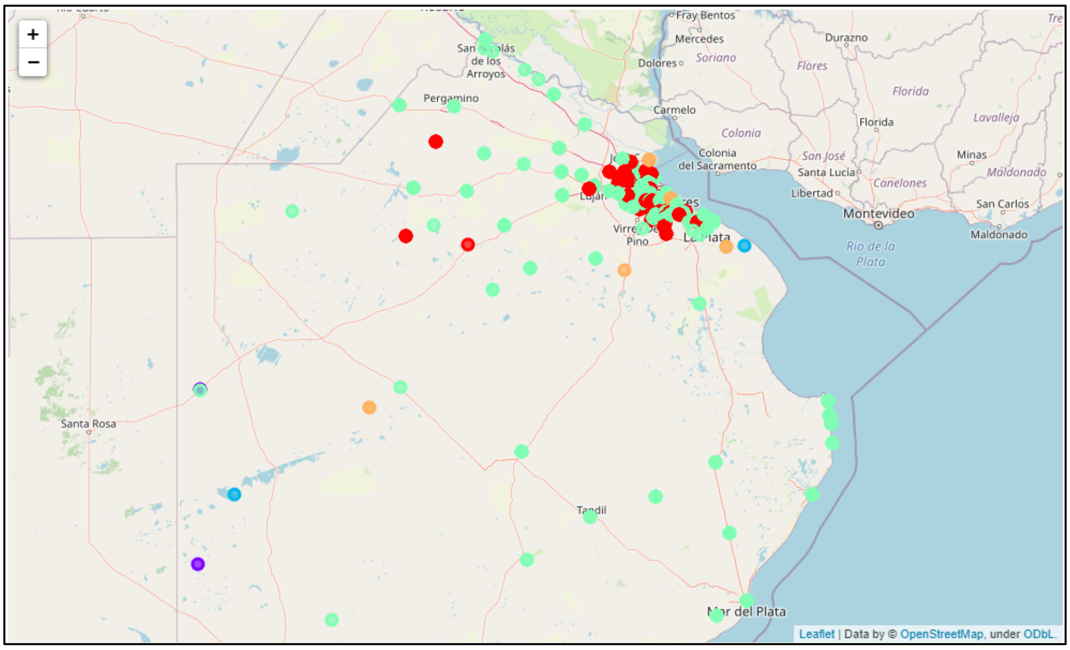
In the case of study, the input data for the first clustering model are the frequencies of each venue category related to each of the zip-codes.

Accessories Store	American Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Asian Restaurant	Athletics & Sports	Auto Garage	Auto Workshop	...	Theme Park	Theme Park Ride / Attraction	Thrift / Vintage Store	Toy / Game Store
0.0	0.0	0.066667	0.0	0.0	0.0	0.066667	0.0	0.0	...	0.0	0.0	0.0	0.0
0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.0	0.0
0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.0	0.0
0.0	0.0	0.066667	0.0	0.0	0.0	0.066667	0.0	0.0	...	0.0	0.0	0.0	0.0
0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.0	0.0

Result of the model

The model processes the data input and responds by classifying each zip-code according to the frequencies sent. The result can be seen in the 'Cluster Label' column of the following table and on its respective map below it.

Cluster Labels	zip number	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
0	1601	Soccer Field	Ice Cream Shop	Department Store	Rock Club	Grocery Store	Gym / Fitness Center	Bakery	Hardware Store	Pharmacy	Athletics & Sports
0	1602	Sports Club	Bakery	Gym	Butcher	Gym / Fitness Center	Tennis Court	Gymnastics Gym	Fruit & Vegetable Store	Deli / Bodega	Train Station
0	1605	Plaza	Dessert Shop	Pharmacy	Bakery	Women's Store	Event Space	Food & Drink Shop	Flower Shop	Fish Market	Financial or Legal Service
0	1607	Soccer Field	Ice Cream Shop	Department Store	Rock Club	Grocery Store	Gym / Fitness Center	Bakery	Hardware Store	Pharmacy	Athletics & Sports
0	1609	Fast Food Restaurant	Grocery Store	Women's Store	Event Space	Food Court	Food & Drink Shop	Flower Shop	Fish Market	Financial or Legal Service	Farmers Market



Note: Cluster0: red ; Cluster1:purple ; Cluster2:light blue ; Cluster3:light green ; Cluster4:orange.

The previous model is not enough to decide, that is why a second clustering model will be carried out in which the input data will be related in the competition with other gyms.

B2. Cluster by characteristics of nearby gyms

Get Gyms Location Data in Buenos Aires

The data necessary to carry out the model were obtained through the Foursquare API, which we set the central coordinates (latitude and longitude) of Buenos Aires and a base radius (10.000 meters) as input, we get every gym's location inside defined radius. the data found is shown in the following table:

	Venue Id	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue Rating
0	50203111e4b0ab4947270817	Vuelta al Hipódromo de San Isidro	-34.480489	-58.519612	Track	8.6
1	4bd4c2f629eb9c74115392e1	Club Nordelta	-34.404626	-58.655813	Gym	8.1
2	4df52479ae609e69dd9f7334	Corredor Aeróbico Muñiz	-34.559512	-58.693454	Track	6.6
3	59aad25b646e382e654d1f98	SportClub	-34.405893	-58.620206	Gym / Fitness Center	5.0
4	506d7a9fe4b0377aa9d05e47	San Fernando Centro	-34.441638	-58.555633	Gym / Fitness Center	5.0

Prepare data model

Comparing the data given (latitude, longitude and rating) with each zip-code, the following measurements were calculated:

- **Nearby gyms walking:** Gymnasiums less than 2,000 meters to the zip-code

- **Nearby gyms bike:** Gymnasiums at a distance between 2000 and 5000 meters to the zip-code

- **Nearby gyms car:** Gymnasiums at a distance between 5000 and 10000 meters to the zip-code

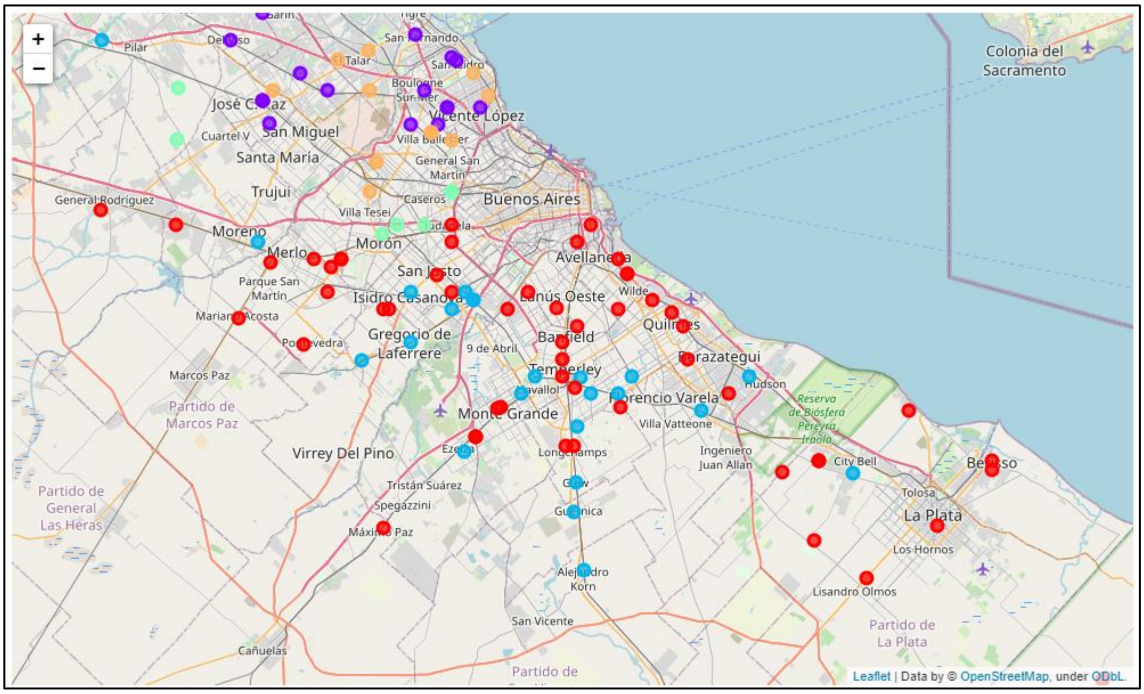
- **Rating:** social rating 1 to 10 of each gym.

setting the described measures and the result of the previous cluster model into a new cluster model calculate the next table:

	Cluster Labels	nearby gym walking	nearby gym bike	nearby gym car	min gym distance	mean gym distance	mean ratings nearby gyms
0	0	0	7	19	2172.0	6073.538462	5.380769
1	0	0	3	19	3826.0	6976.409091	5.450000
2	0	2	3	20	1275.0	6978.080000	5.396000
3	0	0	7	19	2172.0	6073.538462	5.380769
4	0	3	9	15	802.0	4906.407407	5.366667

Result of the model

	zip number	2nd Cluster Labels	nearby gym walking	nearby gym bike	nearby gym car	min gym distance	mean gym distance	mean ratings nearby gyms	1st Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	1601	1	0	7	19	2172.0	6073.54	5.38	0	Soccer Field	Ice Cream Shop	Department Store	Rock Club	Grocery Store
1	1602	1	0	3	19	3826.0	6976.41	5.45	0	Sports Club	Bakery	Gym	Butcher	Gym / Fitness Center
2	1605	1	2	3	20	1275.0	6978.08	5.40	0	Plaza	Dessert Shop	Pharmacy	Bakery	Women's Store
3	1607	1	0	7	19	2172.0	6073.54	5.38	0	Soccer Field	Ice Cream Shop	Department Store	Rock Club	Grocery Store
4	1609	1	3	9	15	802.0	4906.41	5.37	0	Fast Food Restaurant	Grocery Store	Women's Store	Event Space	Food Court



Note: Cluster0:red ; Cluster1:purple ; Cluster2: light blue; Cluster3:light green ; Cluster4:orange.

C. Results and Discussion

With the described variables and using artificial intelligence clustering models, the following clusters were found:

Cluster 0 and Cluster 2

The main characteristic is that they do not have gyms near at short, medium or long distance. That is why making an investment in these points is a great risk because there is no certainty that there is a potential demand in those areas.

	zip number	2nd Cluster Labels	nearby gym walking	nearby gym bike	nearby gym car	min gym distance	mean gym distance	mean ratings nearby gyms	1st Cluster Labels
13	1625	0	0	0	0	NaN	NaN	NaN	3
42	1702	0	0	0	0	NaN	NaN	NaN	3
46	1712	0	0	0	0	NaN	NaN	NaN	3
47	1713	0	0	0	0	NaN	NaN	NaN	3
48	1714	0	0	0	0	NaN	NaN	NaN	3

Note: Cluster 0

	zip number	2nd Cluster Labels	nearby gym walking	nearby gym bike	nearby gym car	min gym distance	mean gym distance	mean ratings nearby gyms	1st Cluster Labels
14	1633	2	0	0	0	NaN	NaN	NaN	0
53	1742	2	0	0	0	NaN	NaN	NaN	0
60	1757	2	0	0	0	NaN	NaN	NaN	0
61	1759	2	0	0	0	NaN	NaN	NaN	0
64	1765	2	0	0	0	NaN	NaN	NaN	0

Note: Cluster 2

Cluster 1

This group is characterized by being several gymnasiums at a medium and long distance, being more precise, the minimum distance to a gym is 1.800 meters and an average distance is 6,000 meters. This group represents a good market opportunity, although there is strong competition over long distances.

	zip number	2nd Cluster Labels	nearby gym walking	nearby gym bike	nearby gym car	min gym distance	mean gym distance	mean ratings nearby gyms	1st Cluster Labels
0	1601	1	0	7	19	2172.0	6073.54	5.38	0
1	1602	1	0	3	19	3826.0	6976.41	5.45	0
2	1605	1	2	3	20	1275.0	6978.08	5.40	0
3	1607	1	0	7	19	2172.0	6073.54	5.38	0
4	1609	1	3	9	15	802.0	4906.41	5.37	0
7	1613	1	0	3	17	2261.0	6312.10	5.08	0
8	1615	1	1	2	15	1486.0	6918.33	5.20	0
10	1619	1	0	2	3	2291.0	7107.00	6.02	0
18	1642	1	9	8	8	830.0	4155.48	5.30	0
19	1643	1	8	9	9	435.0	4223.88	5.28	0
21	1646	1	1	12	16	1466.0	5486.07	5.36	0
26	1655	1	1	3	22	1878.0	7138.23	5.38	0
29	1661	1	1	1	11	1923.0	6285.38	5.28	0

```
nearby gym walking            1.705882
nearby gym bike               4.529412
nearby gym car               13.235294
min gym distance           1820.000000
mean gym distance          6078.031765
mean ratings nearby gyms      5.351176
```

Note: mean of each column

Cluster 3

It is characterized by having no offer at short-medium distance and very low offer at long distance with an average of 3 gymnasiums, of which they have an average distance of 8,300 meters and a minimum distance of 7,700 meters. The recommended addresses in cluster 3 are detailed below:

	zip number	2nd Cluster Labels	nearby gym walking	nearby gym bike	nearby gym car	min gym distance	mean gym distance	mean ratings nearby gyms	1st Cluster Labels
11	1621	3	0	1	7	4955.0	6882.12	5.64	3
12	1623	3	0	0	5	8161.0	8737.60	6.02	0
15	1635	3	0	0	1	7534.0	7534.00	5.00	0
20	1644	3	0	0	3	9072.0	9464.00	5.00	4
22	1648	3	0	0	3	9072.0	9464.00	5.00	4
23	1649	3	0	0	3	9072.0	9464.00	5.00	4
35	1672	3	0	0	3	6603.0	7430.67	5.83	0
36	1674	3	0	0	3	6603.0	7430.67	5.83	0
37	1676	3	0	0	3	6603.0	7430.67	5.83	0
38	1678	3	0	0	3	6603.0	7430.67	5.83	0
43	1704	3	0	0	3	9320.0	9588.33	5.83	3
44	1706	3	0	0	2	7748.0	8569.50	6.25	3
45	1708	3	0	0	1	8125.0	8125.00	5.00	3
54	1744	3	0	0	1	8726.0	8726.00	5.00	0

```
nearby gym walking          0.000000
nearby gym bike             0.071429
nearby gym car              2.928571
min gym distance         7728.357143
mean gym distance        8305.516429
mean ratings nearby gyms    5.504286
```

Note: mean of each column

Cluster 4

It is characterized by being a small group with a high density of gyms, having an average of 5 gyms at short distance, 9 gyms at medium distance and 11.6 gyms at long distances. Given its high level of supply, investing in this group is risky due to the high level of competition.

	zip number	2nd Cluster Labels	nearby gym walking	nearby gym bike	nearby gym car	min gym distance	mean gym distance	mean ratings nearby gyms	1st Cluster Labels
5	1611	4	3	3	24	1278.0	7071.47	5.19	3
6	1612	4	0	5	17	3491.0	7331.27	5.14	3
9	1617	4	2	4	19	47.0	6517.24	5.19	3
16	1636	4	0	5	17	3877.0	6914.73	5.45	3
17	1640	4	4	8	10	1062.0	4694.64	5.45	3
24	1650	4	2	2	19	294.0	7298.96	5.43	3
25	1651	4	0	3	13	2419.0	7490.81	5.62	3
27	1657	4	2	2	19	294.0	7298.96	5.43	3
28	1659	4	2	2	19	294.0	7298.96	5.43	3
32	1665	4	0	3	10	2484.0	6105.31	5.28	3
39	1682	4	1	1	14	1342.0	7223.06	5.26	3
40	1686	4	0	1	10	3224.0	7811.82	5.37	3
41	1688	4	1	1	14	1342.0	7223.06	5.26	3

zip number	1650.307692
2nd Cluster Labels	4.000000
nearby gym walking	1.307692
nearby gym bike	3.076923
nearby gym car	15.769231
min gym distance	1649.846154
mean gym distance	6944.637692
mean ratings nearby gyms	5.346154

Note: mean of each column

Addresses of centers of areas recommended for further analysis:

In order to maximize opportunities with the least possible risk, the study indicates that the best option is in Cluster 3, with the following addresses:

- Club Newman, Benavídez, Partido de Tigre, Buenos Aires, 1621, Argentina
- Barrio Santa Isabel, Ingeniero Maschwitz, Partido de Escobar, Buenos Aires, B1623, Argentina
- Presidente Derqui, Partido del Pilar, Buenos Aires, 1635, Argentina
- Isla Nazar Anchorena, Primera Sección, Partido de Tigre, Buenos Aires, B1644BHH, Argentina
- 879, 411 - Beazley, Sáenz Peña, Partido de Tres de Febrero, Buenos Aires, B1674AVJ, Argentina
- Luis Antonio Beruti, Morón, Partido de Morón, Buenos Aires, B1708KCH, Argentina
- Boca Ratón Golf Club, Pilar Sur, Partido del Pilar, Buenos Aires, Argentina

D. Conclusion

Purpose of this project was to identify Buenos Aires areas close to center with low number of gyms in order to aid stakeholders in narrowing down the search for optimal location for a Gym and fitness center. By calculating Gym density distribution from Foursquare data, we have first identified general boroughs that justify further analysis, and then generated extensive collection of locations which satisfy some basic requirements regarding existing nearby gyms. Clustering of those locations was then performed in order to create major zones of interest (containing greatest number of potential locations).

Final decision on optimal gym 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, social and economic dynamics of every neighborhood etc.