Gyms in Boston:
To Do or Not To Do

IBM Data Science
Capstone Project

Reha Patel

1. Introduction

1.1 Background

In a country where almost a fifth of all people go to the gym, there is certainly some money to be made in the fitness space. In fact, in 2017 it was reported that over 60 million Americans had some sort of fitness membership, according to Statista.com. What is interesting to note is that this number does not include the total number of Americans that perform some sort of physical activity. Surely there is a group of people that perform physical activity, but choose to not go to a gym due to factors such as vicinity, cost, etc.

1.2 Problem

This report will focus on the vicinity aspect and specifically in Boston, Massachusetts, USA. In a city of almost 700,000 people with many more commuting in and out every day, this report will examine the different neighborhoods to determine if there is a neighborhood that holds potential for development of a gym.

1.3 Interest

As the Greater Boston Area continues to expand and develop into its surrounding neighborhoods such as Dorchester, Roxbury, etc., this information will become relevant and can even be manipulated to target the development of movie theaters, housing complexes, etc. instead of gyms. Therefore, the target audience of this report will be developers and investors that are looking to profit on the production of such development. According to Norada Real Estate Investments, the average housing prices in Boston are increasing by about 5.7% per year. This might not seem like much until you realize that the most expensive neighborhood of Beacon Hill has an median housing price of over $2 million dollars. Having such information, but for all neighbors can let investors know if it's really worth placing a gym in an area where housing is so expensive.

2. Data Acquisition and Cleaning

2.1 Data Sources

The primary dataset we're going to be using in this project is that from the Foursquare API, specifically the venue data that lists the gyms in the Greater Boston Area. In regards to the area we will be surveying, I will refer to the 22 neighborhoods listed in this Wikipedia article. In addition to this, the Foursquare API will be an immensely valuable data source to this report. Given the neighborhood longitude and latitude information, I will be able to find a limit of 100 venues within each neighborhood.

2.2 Data Cleaning

In terms of cleaning data, there was a minimal amount to be done given what had been web scraped. Two of the Boston neighborhoods included in parentheses additional information such as "X-area is included." This additional information was ultimately removed because of the potential for error when attempting to gather longitude and latitude data for the neighborhoods. After that, because it was a relatively small data set there were no other modifications.

3. Data Analysis

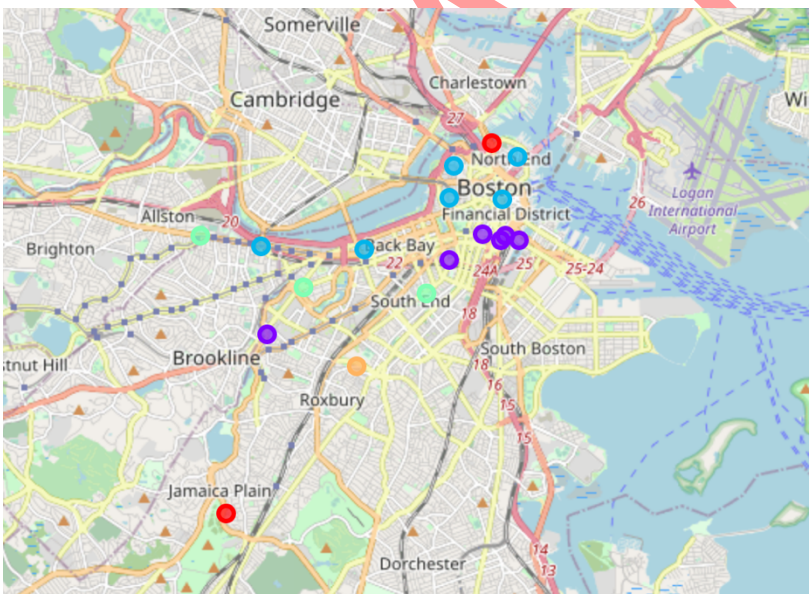
3.1 Creation of Data Frames

Throughout the process of gathering and manipulating the data, a number of data frames were created. The initial data frame began as just listing the different neighborhoods that were scraped from the Wikipedia article. From there, it was cleaned and was looped through a method that called *geocoder* in order to get the coordinates of the neighborhood. The next step in this was to create a data frame after connecting the Foursquare API and included additional information such as Venue Name and Venue Category. Because the call to the API referenced the existing neighborhoods, a new data frame had been created. Lastly, a data frame that specifically referenced the proportion of gyms in the neighborhood was created. The final data frame included the cluster name as well and was ultimately a culmination of all previous data frames.

3.2 Clusters

The 22 neighborhoods were clustered into 5 groups. The clusters as well as the final data frame went as follows:

	Neighborhood	Gym	Cluster Labels	Latitude	Longitude
5	Charlestown	0.000000	0	42.367771	-71.059016
13	Mattapan	0.000000	0	42.278222	-71.096083
11	Hyde Park	0.000000	0	42.274773	-71.119898
12	Jamaica Plain	0.000000	0	42.305849	-71.119092
18	South Boston	0.030000	1	42.352250	-71.055690
2	Bay Village	0.030000	1	42.348165	-71.068470
6	Chinatown/Leather District	0.030000	1	42.352510	-71.060900
7	Dorchester	0.030000	1	42.351355	-71.052848
14	Mission Hill	0.030000	1	42.335710	-71.109800
9	East Boston	0.030000	1	42.351418	-71.056714
0	Allston	0.010000	2	42.350531	-71.111091
16	Roslindale	0.012987	2	42.281820	-71.137104
15	North End	0.010000	2	42.365490	-71.052970
21	West Roxbury	0.013333	2	42.282201	-71.146000
8	Downtown	0.010000	2	42.358290	-71.056630
3	Beacon Hill	0.010000	2	42.358420	-71.068600
1	Back Bay	0.010000	2	42.349990	-71.087650
20	West End	0.010000	2	42.363940	-71.067390
4	Brighton	0.020000	3	42.352134	-71.124925
19	South End	0.020000	3	42.342560	-71.073580
10	Fenway Kenmore	0.020000	3	42.343550	-71.101570
17	Roxbury	0.040000	4	42.330304	-71.089469



We can see that neighborhoods in Cluster 0 have gym = 0. In Cluster 1, gym = 0.03, and so on. When examining a map of the clusters, we also see that the neighborhoods in Cluster 0 are primarily found in the outskirts of the Greater Boston Area, whereas the neighborhoods with a greater proportion of gyms are found closer to Downtown, near more populous areas.

4. Conclusions

Based on the analysis conducted above, we can come to a few conclusions. In general, we see that there are fewer gyms as you move away from the literal city of Boston and into its surrounding neighborhoods. Cluster 0 which does not have any gyms within a mile radius of the neighborhoods coordinates includes neighborhoods which are almost entirely not within the city. On the other hand we see the cluster with an average number of gyms to be closer to Downtown, where there are a greater number of businesses and vicinity to the MBTA train system. It is possible this number actually appears to be lower on average because in Clusters 1 and 2, there are a greater number of businesses in the area.

In regards to where an investor should open a gym, I would suggest opening one in Cluster 0. Cluster 0 is a generally residential area and would potentially find benefits in development. Because of this fact, it is actually possible that residents in Cluster 0 are traveling to gyms in other neighborhoods. On the other hand, I would deter investors from opening any gyms in Clusters 1 and 2. Because these areas are densely populated with venues ranging from bars to restaurants, there are also many gyms and therefore a new gym would face severe competition. One issue the gym would face in this area is convincing gym-goers to switch from their home gym to the new gym and it is ultimately not worth the risk.