Coursera Capstone

IBM Applied Data Science

Opening a New Gym in Chicago, Illinois

By: Sahil Desai

June 2020

Introduction

Gyms are a great way for people to spend their leisure time and exercise. Events, group activities, and even some dining can be present at gyms and reap in a lot of profit. Gyms bring different types of consumers or target markets due to the diverse number of things people can do there. Some gym franchises have spas, pools, saunas, and daycare for children. Entrepreneurs and small business related to the health/fitness industry want to incorporate some product placement in these facilities. It is a great area for them advertise and launch new products, and a way to create a distribution channel. Property development companies want to take advantage of building gyms to cater the demand of health and fitness facilities and bring in profit. Chicago is listed as the third largest city by population in the United States. As a result, many gyms and other venues are already built to adhere to the large population. The goal of gym owners is to make enough money off their members to at least pay rental costs to contractors or developers. Like any business decision, they need to choose the location of where they build very carefully. Gyms on average have a rent for $6500, with startup costs being $10,000-$50,000 on average.

Business Problem

The objective of this capstone project is to find out where is the best location to build a gym in Chicago, Illinois. Throughout the teachings of the courses, data science methodology and machine learning techniques like clustering will be implemented to make data driven decisions to clients. Based off the findings, where would be the best area to build a new gym in Chicago as an aspiring gym owner?

Target Audience

The target audience of the project are gym owners, gym contractors, or a property developer looking to invest or open a gym in the Chicago suburban area. With the large and diverse population Chicago provides, Chicago is constantly building new things weather its new offices/buildings for their tech companies, or attractions to keep their city lively and citizens invested financially to the city’s economy. Building a gym there could be a good investment due to the already established and large population. It can be mutually beneficial to the citizens and the gym developers or owners.

Data

To figure out this problem, we will need the following data:

- List of neighborhoods in Chicago, Illinois. This defines the scope of this project, which is confined to this area, the capital city of the state of Illinois in midwestern region of the U.S
- Latitude and longitude coordinates of the neighborhoods. This is required for plotting of the maps and to get the venue data.
- Venue data, particularly data related to gyms. We will use this data to perform clustering on the neighborhoods.

Data Sources and method extraction

This Wikipedia page (https://simple.wikipedia.org/wiki/Category:Suburbs_of_Chicago,_Illinois) contains a list of neighborhoods in Chicago, Illinois. This consists of a total of 144 neighborhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of making API calls and beautiful soup packages. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods.

After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare has one of the largest databases and has venue data over 190 countries and is used by over 150,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the gym category to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

Methodology

Firstly, we need to get the list of neighborhoods in the city of Chicago. The list is available in the Wikipedia page (https://simple.wikipedia.org/wiki/Category:Suburbs_of_Chicago,_Illinois). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighborhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude to be able to use Foursquare API. To do so, we will use Geocoder package that converts addresses into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas dataframe and then visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Chicago.

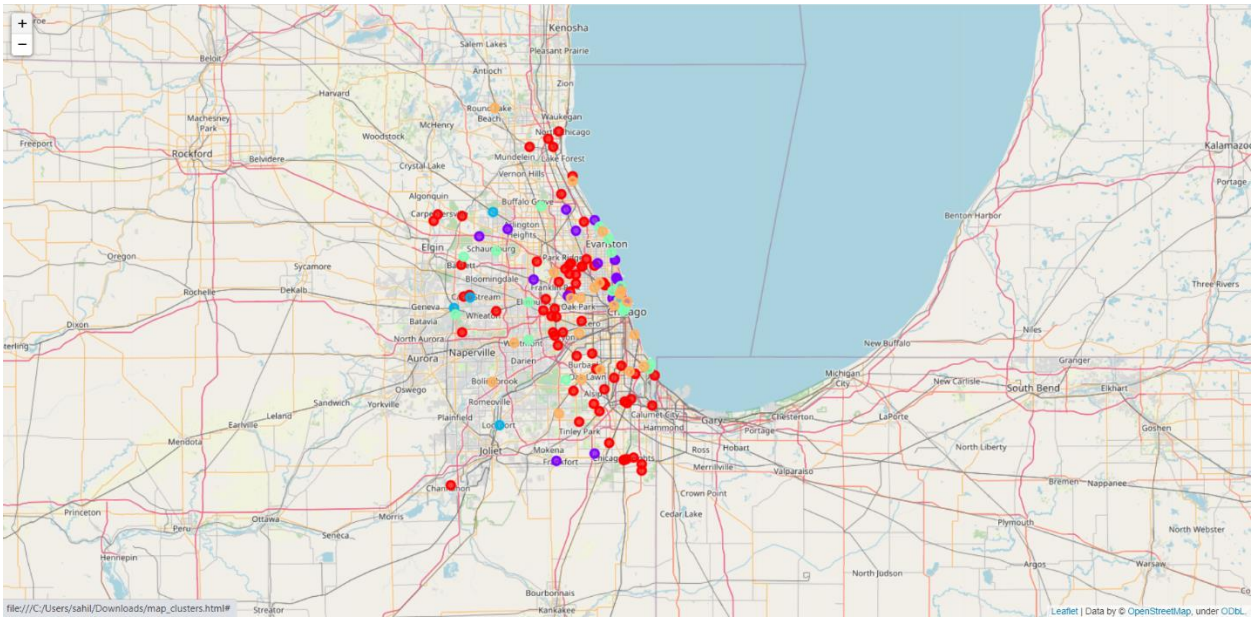
Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be considered from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use k-means clustering. Since we are analyzing “Gym” data, we will filter the “Gym” as venue category for the neighborhoods.

Lastly, we will perform an unsupervised machine learning method, clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 5 clusters based on their frequency of occurrence for “Gym”. The results will allow us to identify which neighborhoods have higher concentration of gyms while which neighborhoods have fewer number of gyms. Based on the occurrence of gyms in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new gyms.

Results

The results from the k-means clustering show that we can categorize the neighborhoods into 5 clusters based on the frequency of occurrence for “Gym”:

- Cluster 0: Neighborhood with no number of gyms (red)
- Cluster 1: Neighborhood with moderate concentration of gyms (blue)
- Cluster 2: Neighborhood with high concentration of gyms (light blue)
- Cluster 3 Neighborhood with moderate concentration of gyms (light green)
- Cluster 4 Neighborhood with low concentration of gyms (green)



Discussion

Cluster 0 takes a heavy portion of the neighborhoods and is widely scattered across near inner city and outer suburbs. It is by far the most diverse and has the lowest amount of gyms. I do believe Foursquare is missing a lot of data over gyms pertaining Cluster 0. However, I will still attempt to make an analysis based off this. We can come up with areas to avoid, such as Clusters 1 and 3 which have moderate competition and would remotely difficult to open a gym there. Cluster 2 is the smallest and most saturated with gyms out of any cluster, so it would be best to avoid this area all together. Cluster 4 has lower amount of gyms overall and could be a good opportunity to build. Overall, I do not think that Cluster 0 is accurate in the sense that it does not have any gyms in all those neighborhoods, especially near other clusters that have gyms. But I did get to see areas to avoid and a cluster that had low competition.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing unsupervised machine learning by clustering the data into 5 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders (gym contractors, gym owners, property developers etc.) regarding the best locations to open a new gym. I would have liked my API call to capture more information about Chicago venues that were categorized as gyms, because I had a large cluster unrealistically portray a big part of the city not having gyms. But based off my findings, we still found clusters that had high and low competition. As to answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 4 are the most preferred locations to open a new gym. The findings of this project will help those to make an informed decision on choosing high potential locations while avoiding over competitive areas in opening a new gym.