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.