Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Gym in St. Louis, Missouri

By: Chule Hou

April 2020

Introduction

For many sport fans, going to the gym is a great way to relax and enjoy themselves during weekends and holidays. They can take exercise in the gym, release stress, or get fit for other fitness exercises. Gyms are especially important in every area. The biggest benefit of fitness is that aerobic exercise works the heart, strengthens the lungs, improves circulatory system function, burns fat, increases lung capacity, lowers blood pressure, prevents diabetes, and reduces the incidence of cardiovascular disease. To achieve a long and healthy life by exercising for a healthier body. Fitness is essential for some people every day, they spend some time, can be in a very good fitness environment inside the most effective exercise, can make the body in a very healthy state. For every neighborhood, providing a gym for local residents is very important, there are many gyms in St. Louis, and more are being built, opening a gym can bring convenience to the surrounding residents while also making profit for investors. Of course, as with any business decision, opening a new gym requires serious consideration and is much more complicated than it seems. In particular, the location of the gym is one of the most important factors in determining whether or not a gym can open successfully.

Business Problem

The objective of this capstone project is to analyze and select the best locations in the city of St. Louis, Missouri to open a new gym. Using data

science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city of St. Louis, Missouri, if a property developer is looking to open a new gym, where would you recommend that they open it?

Target Audience of this project

The project is particularly useful for real estate developers and investors who want to open or invest in stadiums in St. Louis, Missouri. The project is timely because the city is currently suffering from shortages in the gym. Especially during the new coronavirus. More and more residents hope to have suitable gymnasiums around their communities. The audience of this project is a large number of real estate developers and investors, and perhaps there can also be investors in the community.

Data

To solve the problem, I will need the following data:

• List of neighborhoods in St. Louis. This defines the scope of this project which is confined to the city of St. Louis, the biggest city of Missouri in the middle of America.

• Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.

• Venue data, particularly data related to shopping malls. I will use this data to perform clustering on the neighborhoods.

Sources of data and methods to extract them

This Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighborhoods_in_St._Louis) contains a list of neighborhoods in St. Louis, with a total of 88 neighborhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests 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 of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Shopping Mall category in order 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, I need to get the list of neighborhoods in the city of St. Louis. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighborhoods_in_St._Louis). I 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. I need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, I will use the wonderful Geocoder package that will allow me to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, I will populate the data into a pandas DataFrame and then visualize the neighborhoods in a map using Folium package. This allows me to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Kuala Lumpur.

Next, I will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. I need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. 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 I will extract the venue name, venue category, venue latitude and longitude. With the data, I can check how many venues were returned for

each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, I will analysis each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, I also preparing the data for use in clustering. Since I am analyzing the “Gym” data, I will filter the “Gym” as venue category for the neighborhoods.

Lastly, I will perform clustering on the data by using k-means 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. I will cluster the neighborhoods into 3 clusters based on their frequency of occurrence for “Gym”. The results will allow me 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 me to answer the question as to which neighborhoods are most suitable to open new gyms.

Results

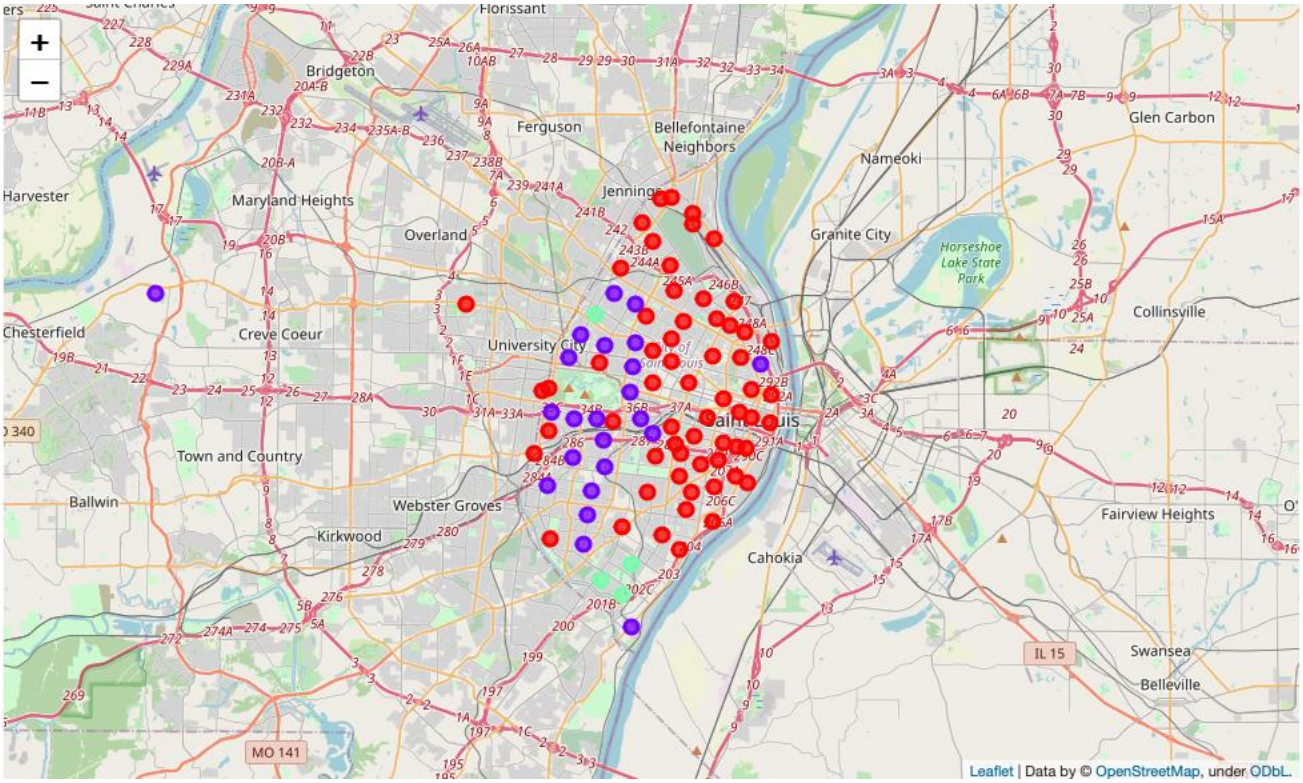
The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Gym”:

• Cluster 0: Neighborhoods with high concentration of gyms

• Cluster 1: Neighborhoods with moderate number of gyms

• Cluster 2: Neighborhoods with low number to no existence of gyms

The results of the clustering are visualized in the map below with cluster 0 in red color, cluster 1 in purple color, and cluster 2 in mint green color.



Discussion

As observations noted from the map in the Results section, most of the gyms are concentrated in the central area of St. Louis, with the highest number in cluster 0 and moderate number in cluster 1. On the other hand, cluster 2 has very low number to no gym in the neighborhoods. This represents a great opportunity and high potential areas to open new gyms as there is very little

to no competition from existing gyms. Meanwhile, gyms in cluster 0 are likely suffering from intense competition due to oversupply and high concentration of gyms. From another perspective, the results also show that the oversupply of gyms mostly happened in the central area of the city, with the suburb area still have very few gyms. Therefore, this project recommends property developers to capitalize on these findings to open new gyms in neighborhoods in cluster 2 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new gyms in neighborhoods in cluster 1 with moderate competition. Lastly, property developers are advised to avoid neighborhoods in cluster 0 which already have high concentration of gyms and suffering from intense competition.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of gyms, there are other factors such as population and income of residents that could influence the location decision of a new shopping mall. However, to the best knowledge of this researcher such data are not available to the neighborhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new gym. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research

could make use of paid account to bypass these limitations and obtain more results.

Conclusion

In this project, I have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new shopping mall. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 2 are the most preferred locations to open a new gym. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new gym.

Reference

Category: Neighborhoods in St. Louis. Wikipedia. Retrieved from https: //en.wikipedia.org/wiki/Category:Neighborhoods_in_St._Louis

Foursquare Developers Documentation. Foursquare. Retrieved from http s://developer.foursquare.com/docs