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.