Opening a New Gym in Pune,India

Omkar Nagarkar

April 30,2020

1.Introduction

A gymnasium, also known as a gym, is a covered location for athletics. The word is derived from the ancient Greek gymnasium. They are commonly found in athletic and fitness centers, and as activity and learning spaces in educational institutions. "Gym" is also slang for "fitness centre ", which is often an area for indoor recreation. A gym may be open air as well. Gym apparatus such as barbells, jumping board, running path, tennis-balls, cricket field, and fencing area are used as exercises. In safe weather, outdoor locations are the most conducive to health. Their curricula included self-defense, gymnastic a medical, or physical therapy to help the sick and injured, and for physical fitness and sports, from boxing to dancing to skipping rope. Today, gym are commonplace in the around. The number of gyms in the India has more than doubled since the late 1980s.Gym has been popular among the youth. Property developers are also taking advantage of this trend to build more gyms to cater to the demand. As a result, there are many gyms in the city of Pune and many more are being built. Opening gym allows property developers to earn consistent rental income. Of course, as with any business decision, opening a new gym requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the gym is one of the most important decisions that will determine whether the mall will be a success or a failure.

2.Business Problem

The objective of this capstone project is to analyze and select the best locations in the city of Pune, India 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 Pune, India, if a property developer is looking to open a new gym, where would you recommend that they open it?

3.Target Audience of this project

This project is particularly useful to property developers and investors looking to open or invest in new gym in the Pune,India.

4.Data

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

- List of neighborhoods in Pune. This defines the scope of this project which is confined to the city of Pune, India.
- 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 Gyms. We 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/Template:Neighbourhoods_of_Pune) contains a list of neighborhoods in Pune, with a total of 41 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 database 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 Gym 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.

5.Methodology

Firstly, we need to get the list of neighborhoods in the city of Pune. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Template:Neighbourhoods_of_Pune). 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 in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address 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 Pune.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 5000 meters. We need to register a Foursquare Developer Account in order 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 curated 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 in clustering. Since we are analyzing the “Gym” data, we will filter the “Gym” as venue category for the neighborhoods. As we have venue category “Gym” and “Gym / Fitness Center” which are of same categories so we merge them together.

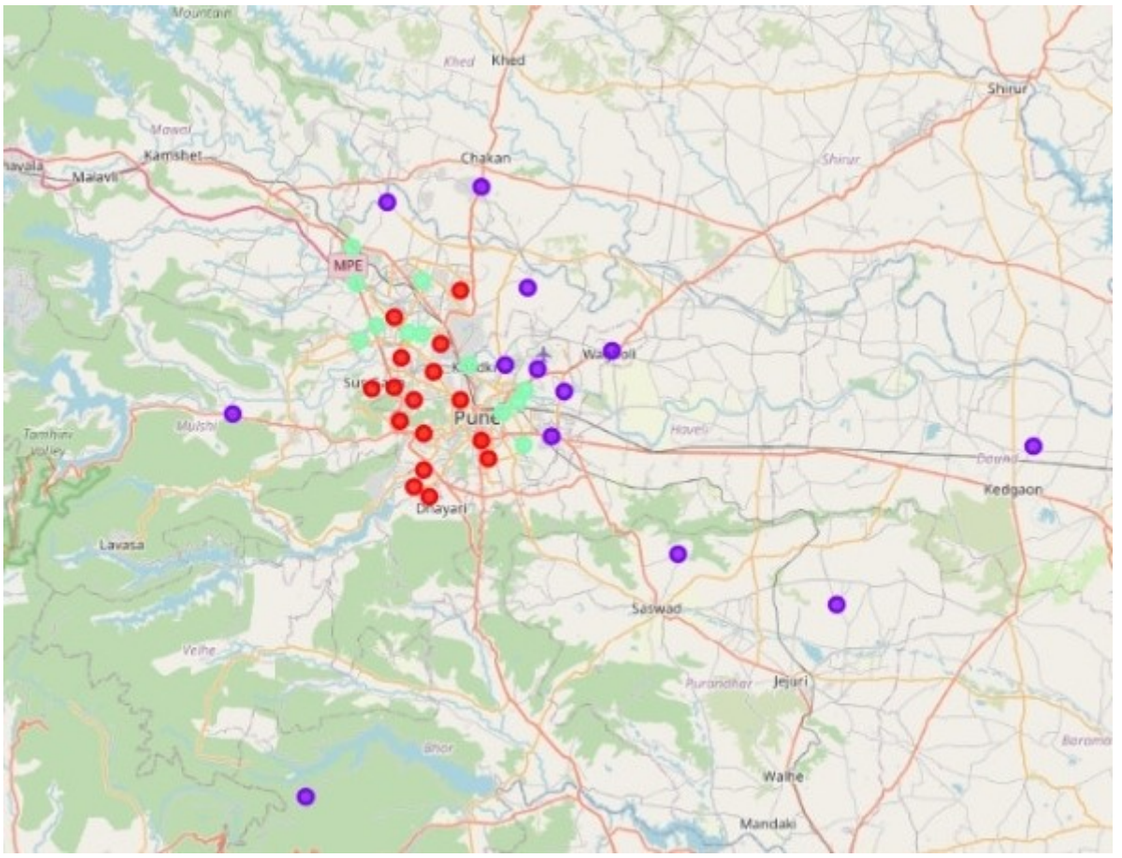
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids , and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 3 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 Gym.

6.Result

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 moderate number of Gym
- Cluster 1: Neighborhoods with low number to no existence of Gym
- Cluster 2: Neighborhoods with high concentration of Gym

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



7.Discussion

As observations noted from the map in the Results section, the highest number of gyms in cluster 2 and moderate number in cluster 0. On the other hand, cluster 1 has very low number to no Gym in the neighborhoods. This represents a great opportunity and high potential areas to open new gym as there is very little to no competition from existing gyms. Meanwhile, gyms in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of gyms. From another perspective, 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 1 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

0 with moderate competition. Lastly, property developers are advised to avoid neighborhoods in cluster 2 which already have high concentration of gyms and suffering from intense competition.

8.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 gyms. However, to the best knowledge of this researcher such data are not available to the neighborliness 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.

9.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 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 gym. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 1 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.

10.References

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

Template:Neighbourhoods of Pune,Wikipedia. Retreved from https://en.wikipedia.org/wiki/Template:Neighbourhoods_of_Pune