IBM APPLIED DATA SCIENCE CAPSTONE

Opening a new Gym in Bangalore, Karnataka, India

By

Kishore Chandra Dash

INTRODUCTION

For Fitness enthusiast, Gym is a great place to visit. People in Bangalore especially young working class prefer to spend on an average 1.5 to 2 hours in Gym. As per the health experts regular exercise helps control weight when use with a balanced diet. Regular physical activity can help you prevent or manage a wide range of health problems and concerns, including stroke, metabolic syndrome, type 2 diabetes, depression, and certain types of cancer, arthritis and falls. Regular training helps reduce stress and improves moods. Hence the demand for fitness centers like Gym is very high in this city. For investors and property developer choosing the right place to build a Gym is very crucial. Opening a Gym in the right place will allow investors to gain a good profit as well as the satisfaction of keeping people health.

BUSINESS PROBLEM

The main objective of this project is to analyze and select the best location in the city of Bangalore, India to open a Gym (Fitness Center). 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 Bangalore, India, if property developer is looking to open a new Gym or Fitness center, where would you recommend them to open?

TARGET AUDIENCE

This project is particularly useful for people (Property Developer, Investor etc) looking to open or invest in a Gym in the city of Bangalore, India.

DATA

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

- List of neighborhoods in Bangalore, India. This defines the scope of the project which his confined to the city of Bangalore, 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 Gym. We will use this data to perform clustering on the neighborhoods.

Source of data and method to extract them:

The Wikipedia page https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Bangalore contains a list of neighborhoods in Bangalore, India, with a total of 129 neighborhoods. We will use web scrapping techniques to extract the data from the Wikipedia page with the help of Python requests and Beautifulsoup packages. Then we will get the coordinates of the neighborhoods using Python Geocode 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 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. This is a project that will make use of data science skills, from web scrapping, working with Foursquare API, data cleaning and data wrangling, to machine learning (K- mean clustering) and Map visualization (Folium). In the next section, we will present the Methodology where we will discuss the steps taken in the project, the data analysis we did and the machine learning technique that was used.

Methodology

First, we need to get the list of neighborhoods in the city of Bangalore, India. Fortunately, the list is available in the Wikipedia page https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Bangalore . We will do web scrapping 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 coordinate 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 coordinate's data returned by Geocoder are correctly plotted in the city of Bangalore, India.

Next, we will use Foursquare API to get the top 200 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare Developer Credentials. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in 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 curetted 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.

Lastly, we will perform clustering on the data by using k-mean clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest

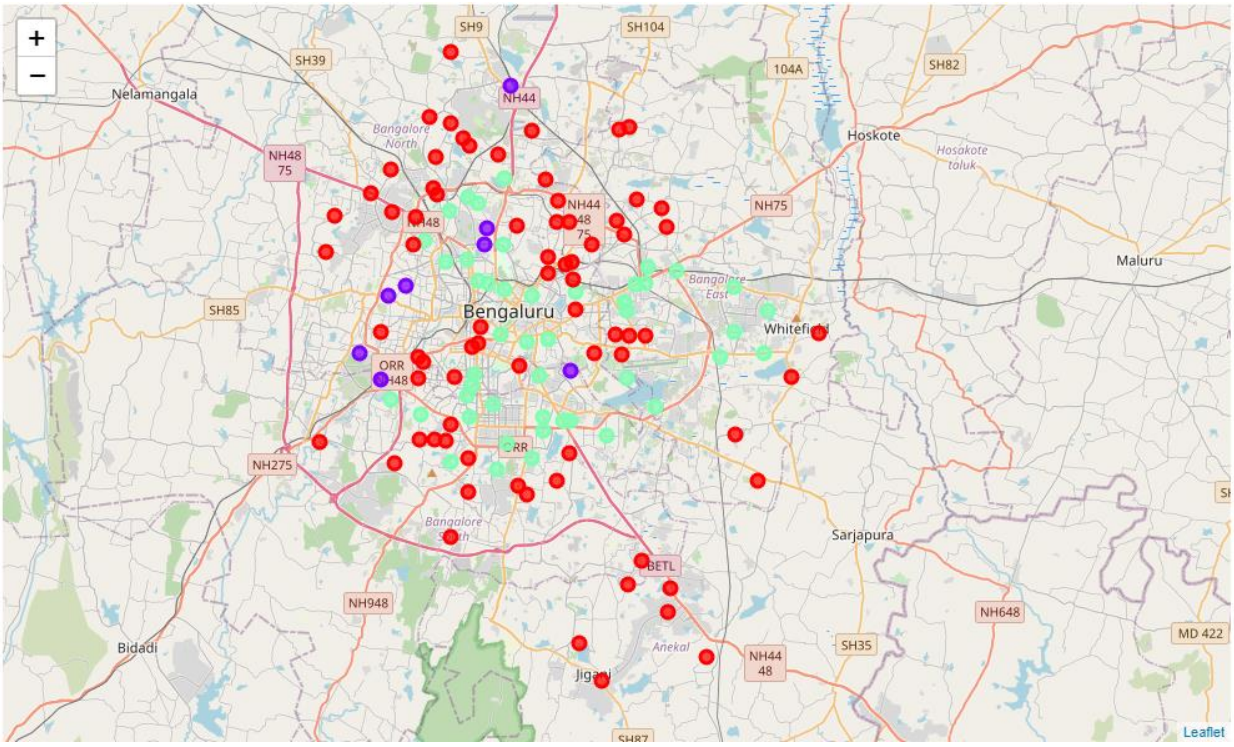
cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 3 clusters based on their frequency of occurrence for “Gym”. The results will allow us to identify which neighborhoods have less number of Gyms. Based on the occurrence of Gyms in different neighborhood, it will help us to answer the question as to which neighborhoods are most suitable to open new Gym.

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.

1. Cluster 0: Neighborhoods with high concentration of Gyms.
2. Cluster 1: Neighborhoods with moderate number of Gyms.
3. 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 purple color, cluster 1 in mini green color, and cluster 2 with red color.



Discussion

As observations noted from the map in the results section, most of the Gyms are concentrated in the central area of Bangalore city, 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 a new Gym 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 happen 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 neighborhood in cluster 0 which already have high concentration of Gyms and suffering from intense competition.

Limitations and Suggestions for Future Research

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 Gym. 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, 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 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.

References

- Category: Suburbs in Bangalore, Karnataka, India. Wikipedia Retrieved from: https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Bangalore.