Opening a new Fitness Center in Mumbai,IN

IBM Applied Data Science Capstone

AUGUST 2020

Introduction

In today's stressful daily lives, a fitness center serves as the perfect place to improve physical as well as mental well-being. It serves as an escape from the world where your entire focus is on yourself and how you can improve. It is very often a place where people can let go of their stress and focus on pushing themselves to the limit, to achieve their goals, whatever they may be. Simply signing up to a fitness center is considered a sign of commitment, and it teaches a lot of values, not least of which being that hard work always pays off.

With a large young population, Mumbai is one of the hotspots for trending activities- a fitness center is one of those at the front of that list. The demographic of Mumbai falls right into those of the desired prospective clients- the young and working population who are very conscious about how they look and are willing to go the extra mile in order to improve their physique.

Shortlisting a nice location for the fitness center can be a daunting task, since Mumbai is a large city and the location determines the per capita wealth for that neighborhood. As such, picking a good location, which is convenient to reach and appropriately priced, becomes a task of paramount importance.

Business Problem

The objective of this project is to identify the best location in Mumbai to open a new fitness center. We will be applying a lot of data science methodologies and techniques such as clustering to find the most suitable neighborhoods in the city of Mumbai where a new fitness center can be opened up.

Target Audience

The target audience for this project is developers and investors or even fitness enthusiasts who are looking to open up their own fitness centers but don't know which location should be selected. With a lot of fitness centers going out of business during the current Coronavirus pandemic, it serves an opportunity to take over considerable market share by choosing the right locality.

Data

For this project, we will use the following data components:

- List of neighborhoods in Mumbai- This list will be extracted from WIkipedia and will contain a list of all the neighborhoods and their respective regions in the city of Mumbai
- Latitude and Longitude coordinates for these neighborhoods will be available from the above table as well as from the Geocoder library

- Venue and location data in and around points of interest- To identify regions of interest and shortlist clusters which can be potentially key to determining the new location

Data Sources

https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai contains a list of all the neighborhoods in Mumbai along with their regions and latitude and longitude data. Using pandas library's "read_html" module, we will extract our desired table for further analysis. We will then use the geocoder package to verify and update the latitudes and longitudes.

Once that is done, we will use the Foursquare API to determine the data for venues within the neighborhoods. Foursquare also provides the category types for these venues, allowing for further analysis of the data. Once the data is cleaned and munging is done, we will begin exploring the data and how we can cluster locations of interest(particularly fitness centers) to develop insights for a new location for the fitness center.

Methodology

The first step in the preparation is to get the list of neighbourhoods in Mumbai. This page on Wikipedia contains a list of all the neighborhoods in Mumbai.
We will perform web scraping using the pandas library in Python to extract the list of neighbourhoods data. We will use the Geocoder package that will allow us to convert the address from the above data into geographical coordinates in the form of latitude and longitude, which are required to use the Foursquare API.

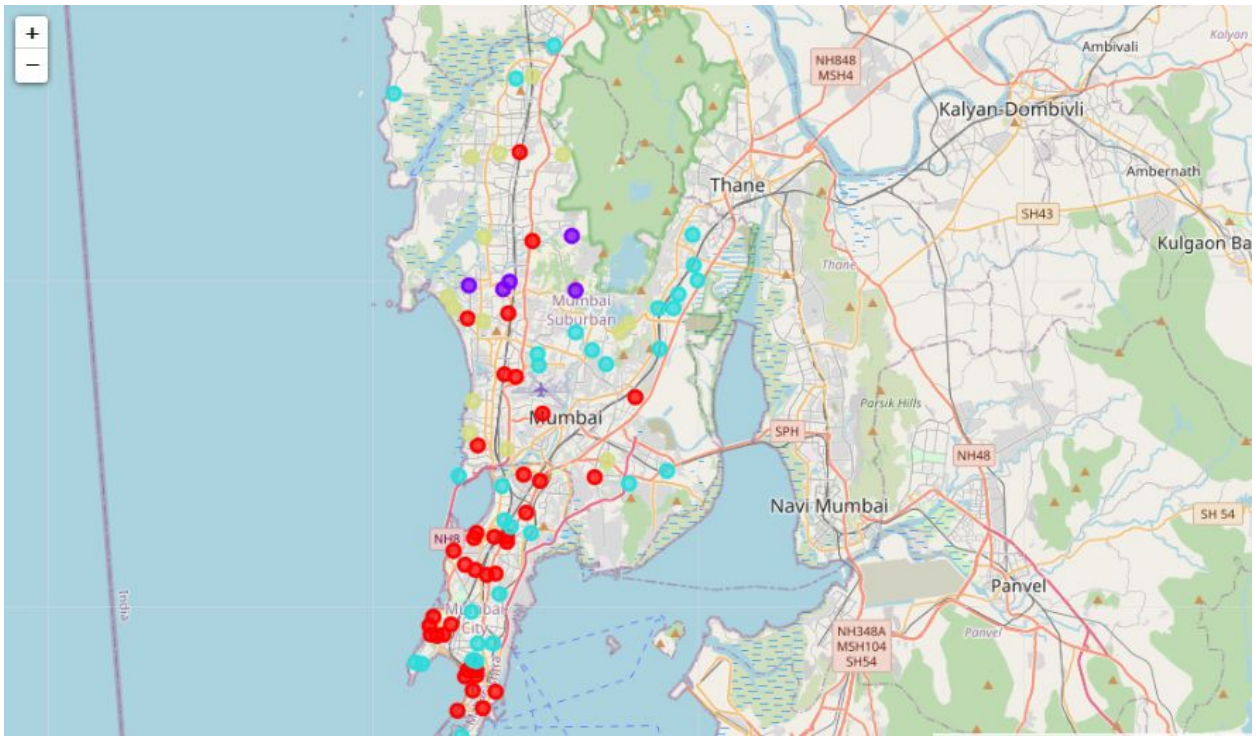
Once the data is gathered, we will populate the data into a pandas DataFrame and then visualize the neighbourhoods in a map using Folium package, to verify the neighborhoods.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 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 neighbourhoods 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 neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood 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 looking for fitness centers, we will filter the "Fitness Center" as venue category for the neighbourhoods.
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 neighbourhoods into 3 clusters based on their frequency of occurrence for "Fitness Center". The results will allow us to identify the concentration of fitness centers in different neighbourhoods. Based on the concentration of fitness centers in different neighbourhoods, we can decide which neighbourhoods are most suitable to open new fitness centers.

Results

From the result below, we can see that there are 4 main clusters as results from the k-means clustering based on the frequency of occurrence for "Fitness Center":
- Cluster 0(Red): Neighbourhoods with moderate number of fitness centers
- Cluster 1(Purple): Neighbourhoods with highest number of fitness centers
- Cluster 2(Blue): Neighbourhoods with lowest concentration of fitness centers
- Cluster 3(Mint Green): Neighbourhoods with moderate number of fitness centers

The results of the clustering are visualized in the map below.



Discussion

As observations noted from the map in the Results section, most of the fitness centers can be found in the clusters 0, 1 and 3 in Mumbai, with the highest number in cluster 1. Cluster 2 has no fitness centers in the neighborhoods. This represents a great opportunity and high potential areas to open new fitness centers. Meanwhile, cluster 1 is likely suffering from intense competition due to a large number of fitness centers. Therefore, this project recommends capitalizing on these findings to open new fitness centers in neighborhoods of cluster 2, and with unique selling propositions, can even succeed in cluster 0 with moderate competition. Lastly, potential investors are advised to avoid neighborhoods in cluster 1 which already have high concentration of fitness centers and suffering from intense competition.

The results also show that the oversupply of fitness centers mostly happening to the north of Suburban Mumbai. Therefore, this project recommends property developers to

capitalize on these findings to open new fitness centers in the neighbourhoods of cluster 2 with little to no competition, below the suburban Mumbai region.

Conclusion and Suggestions for Future Research

In this project, we used the frequency of occurrence to influence the location decision of a new fitness center. Future research could devise a methodology to determine the preferred locations to open a new fitness center using additional factors such as income.

Based on the data, this project recommends property developers to capitalize on these findings to open new fitness centers in the neighbourhoods of cluster 2 with little to no competition, below the suburban Mumbai region.