Coursera Capstone

IBM Applied Data Science Capstone

Opening a new Fitness center in Bay Area , California

Introduction

In recent years, the number of fitness and health services have increased, expanding the interest among the population. Fitness centers now provide both physical exercises with gym equipment and also halls for yoga to help with mental health. People enjoy going to a fitness center that provides an all round experience, from work out spaces to sauna facilities and in some cases even a cafe that caters to pre and post workout menus.

Business problem

The objective of this project is to analyze and select the best locations in the Bay Area to open a new fitness center. Using Data Analysis and machine learning algorithms like clustering, this project will aim to provide an answer to : If a property developer was looking to open a new fitness center, where would you recommend he do so?

Target audience

This project will be particularly interesting to property developers looking to open or invest in fitness centers. With the current scenario of mental health and obesity in the USA, these centers will aid people in the journey to a better lifestyle and put them in touch with equipment and professionals in this field. The growing need to be healthy is rising rapidly and thus establishments like fitness centers will surely be a valuable investment to property developers.

Data

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

- List of neighborhoods in the Bay Area. This defines the scope of this project.
- Latitude and longitude coordinate of those neighborhoods. This will be needed to get venue data and plot on the map
- Venue data, particularly regarding fitness centers. This will be used to perform clustering.

Sources of data and methods that will be applied

This wikipedia page (https://en.wikipedia.org/wiki/Category:Counties_in_the_San_Francisco_Bay_Area) contains the list of neighborhoods in the San Francisco Bay Area,with a total of 9 neighborhoods. We will use web scraping techniques to extract data from Wikipedia page,with the help of Python requests and BeautifulSoup packages.Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude of the neighborhoods.
After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare API has one of the largest databases of over 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data.However, we are particularly interested in the fitness related categories(gyms,spas, fitness centers and yoga studios) that will help solve our business problem.
This project will make use of many data science skills from web scraping , working with Foursquare API, data cleaning, data wrangling, to machine learning (k-means clustering) and map visualisation using Folium.

Methodology

Firstly, we need to get a list of the neighborhoods in the Bay Area, California. This list is available in the wikipedia page (https://en.wikipedia.org/wiki/Category:Counties_in_the_San_Francisco_Bay_Area).We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighbourhoods 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 neighbourhoods 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 Bay Area limits.

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 analysing the fitness related data, we will filter the gyms and yoga studios as venue category for the neighbourhoods.

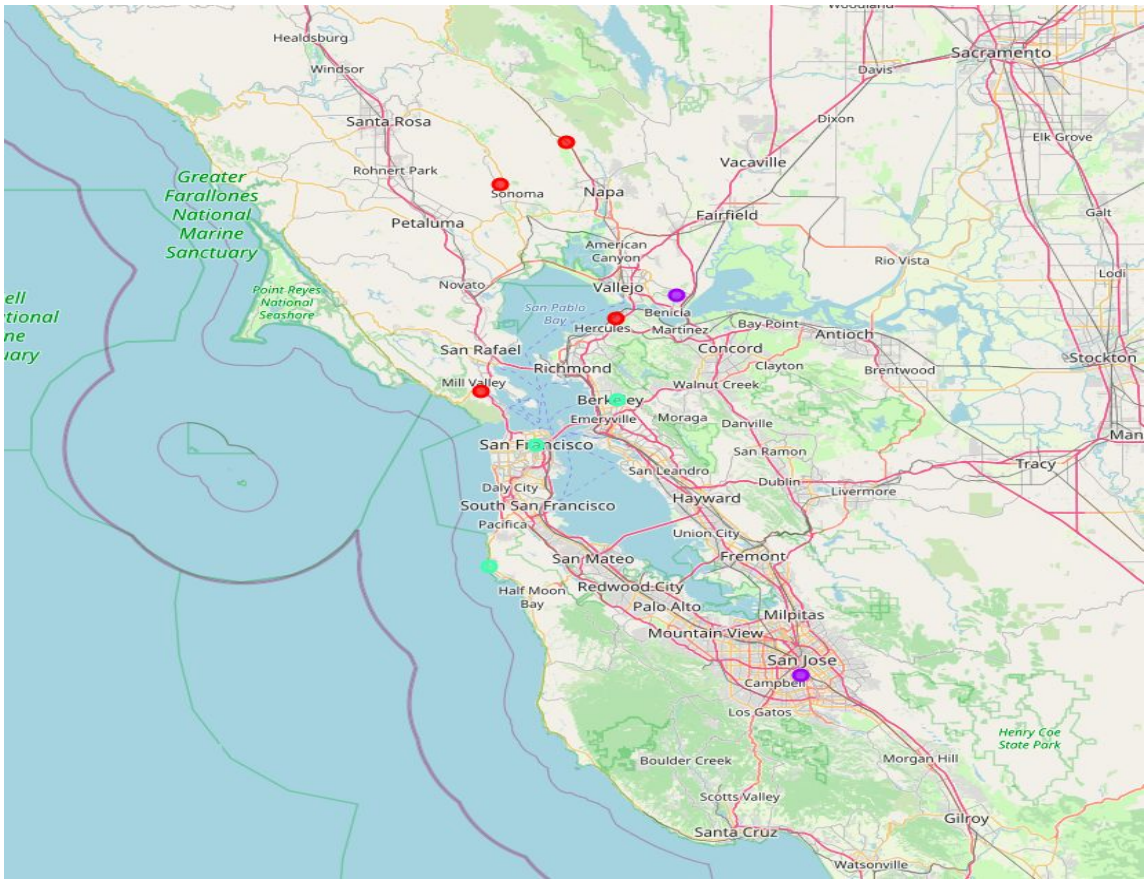
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine 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 all the venue centers listed above. The results will allow us to identify which neighbourhoods have higher concentration of fitness centers while which neighbourhoods have fewer number of fitness centers. Based on the occurrence of gyms, yoga studios and fitness centers in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new fitness center.

Results

The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for fitness center/gyms/yoga studios:

- **Cluster 0:** Neighbourhoods with low number to no existence of fitness centers
- **Cluster 1:** Neighbourhoods with high concentration of fitness centers
- **Cluster 2:** Neighbourhoods with moderate concentration of fitness centers

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.



Discussion

As observations noted from the map in the Results section, most of the Fitness Centers are concentrated in Cluster 1 and moderate number in Cluster 2. On the other hand, cluster 0 has very low number to totally no Fitness Centers in the neighborhoods. This represents a great opportunity and high potential areas to open new Fitness Centers as there is very little to no competition from existing gyms/Yoga Studios. Meanwhile, Fitness Centers in cluster 1 are likely suffering from intense competition due to oversupply and high concentration of multiple centers.

Therefore, this project recommends property developers to capitalize on these findings to open new Fitness centers/Gyms/Yoga Studios in neighborhoods in cluster 0 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new centers in neighborhoods in cluster 2 with moderate competition. Lastly, property developers are advised to avoid neighborhoods in cluster 1 which already have high concentration of Fitness Centers 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,fitness centers and yoga studios, there are other factors such as population and income of residents that could influence the location decision of a new fitness center. However, to the best knowledge of this researcher such data are not available to the neighbourhood 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 shopping mall. In addition, this project made use of the free 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 fitness center. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 0 are the most preferred locations to open a new fitness center. The findings of this project will help the relevant stakeholders to capitalize on the opportunities in high potential locations while avoiding overcrowded areas in their decisions to open a new fitness center.