Coursera Applied Data Science Capstone

IBM Applied Data Science Specialization

Opening a New Gym in Dubai, United Arab Emirates

Introduction

For so many people, going to the gym is an amazing way to relax, build muscle, lose weight, and many other health benefits, but a lot of times it can be a bit annoying for people when there are no close gyms, and they have to travel good distance to reach there, hence it can be a good idea to start a gym/ Fitness center that is situated in a place that can cater to people in the surrounding area. But starting a gym is a big business decision, it's a lot more complicated than it seems, in particular, as mentioned the location of the gym is a very important decision that can determine if it will be a success or a failure

Business Problem

The objective of this project is to analyse and select the best locations in Dubai, United Arab Emirates, to open a new Gym. Using data science methodology and machine learning, this project will provide an answer to someone who wants to open a gym, the best recommended place to open.

Target Audience of this project

This project will be useful to property developers and investors looking to open a new gym in Dubai, United Arab Emirates. Which can be strategically placed that it can attract as much business as possible, and increase the property value.

Data

The following data will be used in this project:

- List of communities in Dubai.

- Latitude and longitude coordinates of those communities.

- Venue data, particularly data related to Gyms. Which will be used to perform clustering on the communities.

Data Sources and methods of extraction

The Wikipedia page ("https://en.wikipedia.org/wiki/Category:Communities_in_Dubai") contains a list of the communities in the city of Dubai, with 79 communities, with the help of beautifulsoup package, the data will be extracted, and then using Python Geocoder package, it will get the latitude and longitude coordinates of the communities.

After that, Foursquare API will be used to acquire the venue data for each of the communities, while it will provide a lot of categories, only the gym category will be the most useful, to solve the business problem.

Methodology

First step is to get a list of communities(places) in Dubai, United Arab Emirates. Fortunately, a list is available on Wikipedia (https://en.wikipedia.org/wiki/Category:Communities_in_Dubai), Then by doing web scraping using Python and beautifulsoup package, we will extract the data from Wikipedia. but this is just a list of names. We will need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. Using Geocoder package, it will allow us to convert addresses into geographical coordinates in the form of latitude and longitude, we will then populate the data into a pandas Data Frame and then visualize the communities in a map using Folium package. This will allow us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Dubai most suitable to open new gym.

After that, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We will make an API calls to Foursquare passing in the geographical coordinates of the communities 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 results we will analyse each community by grouping the rows by community 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 “Gym” data, we will filter the “Gym” as venue category for the community.

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 communities into 3 clusters based on their frequency of occurrence for “Gym”. The results will allow us to identify which communities have higher concentration of Gyms while which communities have fewer number of gyms. Based on the occurrence of gyms in different communities, it will help us to answer the question as to which communities are most suitable to open new gym.

Results

The results from the k-means clustering show that we can categorize the communities into 3 clusters based on the frequency of occurrence for “Gym”:

- Cluster 1: communities with low number to no existence of gyms.
- Cluster 2: communities with high concentration of gyms.
- Cluster 3: communities with moderate number of gyms.

The results of the clustering are visualized in the Figure 1 with cluster 1 in red colour, cluster 2 in purple colour, and cluster 3 in mint green colour.

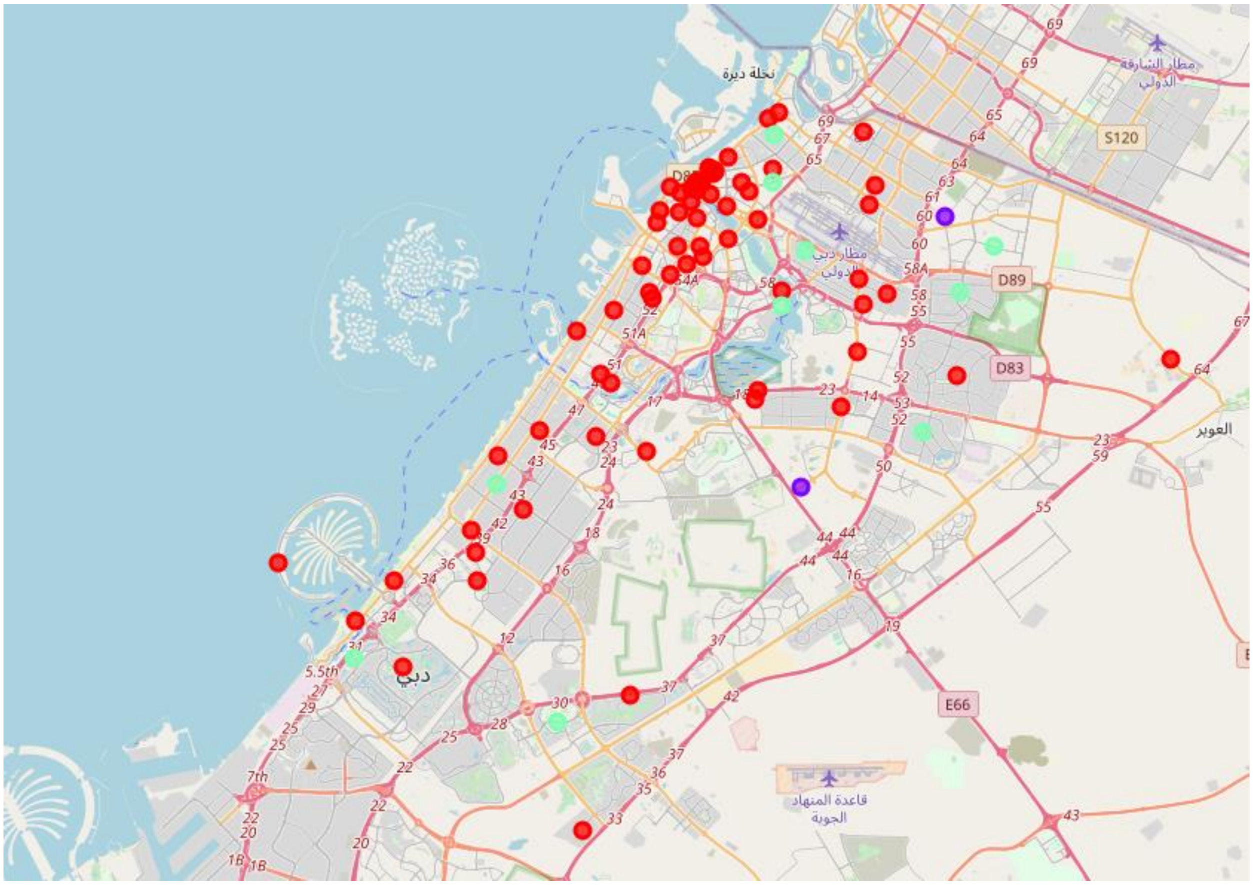

Figure 1: Map visulazing Clusters

Discussion

As observations noted from the map in the Results section, most of the gyms are concentrated in the newly built areas in Dubai, with the highest number in cluster 2 and moderate number in cluster 1. On the other hand, cluster 3 has very low number to no gyms. This represents a great opportunity and high potential areas to open new gyms 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. Therefore, this project recommends property developers to capitalize on these findings to open new gym in the communities 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 communities in cluster 3 with moderate competition. Lastly, property developers are advised to avoid communities in cluster 2 which already have high concentration of gyms and suffering from intense competition.

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 communities 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.