COURSERA CAPSTONE PROJECT
IBM DATA SCIENCE PROFESSIONAL SPECIALIZATION
OPENING A NEW GYM IN Queens, NY
BY:- SHRIHARSH SHENDRE
JUNE 2019

1) INTRODUCTION

Gym is a great way to generate revenue. It helps the population to be fit and at the same time can generate huge revenues.

Nowadays, People strive to be healthy instead of spending 1000s of money at doctor. For that an average person choses gym over running in park or yoga. If a gym is located in a crowded and less competition environment one can generate huge revenue.

1.1) BUSINESS PROBLEM

The objective of this project is to find the area in Queens with low or no gym so owner can get first mover advantage and zero competition.

1.2) TARGET AUDIENCE

This project is particularly useful for the investors and developers who are looking to develop a Gym .The project will help in guiding the investors as well as the developers in identifying the locations with higher probability of generating high revenues.

2) Requirements

To solve the given problem the following data is required:
List of all the major Neighborhood/Borough in Manhattan.
Latitudes and longitudes of all those areas. This is required to plot the graph and also get the venue data.

Venue data, in this case the data about existing gyms. This data will be required to perform clustering.

2.1) Source of data

The data is available at public New York database at 'newyork_data.json' https://cocl.us/new_york_dataset. This dataset has lots of features for New York and covers majority of neighborhoods. On top of it , it possess latitude and longitude data of areas.

2.2)Features to use

city, latitude, longitude attributes will be used for the study.

3) METHODOLOGY

Firstly we need to get the location of all the Boroughs in Queens. Fortunately this data set can be found at https://cocl.us/new_york_dataset.

The dataset contains the list of all the Borough of the Queens and many unnecessary attributes or columns. Therefore the dataset need to be sorted and only the Neighborhoods present in Queens have to be selected.

Another requirement is the coordinates i.e. latitudes and longitudes of those Neighborhood.

Again the data set proved to be helpful because the dataset already contains the coordinates of the cities.

After acquiring the necessary details about the cities they were superimposed on the map of Queens which was created using Folium package.

After plotting the map, *Foursquare API* was used to acquire the venue details within 500 meter radius. They were inserted in a new data frame. Many values were returned for each city in the dataset. For example the neighborhood Astoria had 6 values in *Venue Category* whereas Hammelshad 15. Then it was checked that whether the *Venue Category* contained *GYMs*. Fortunately GYM entry was present along with *Hotel, diner, park, bakery, grocery store* etc. After that each city was analyzed and the results were grouped together.

K means clustering was used after the results were sorted. The numbers of clusters were selected as 3. Again Folium package was used to visualize the clusters.
Then finally the results were analyzed.

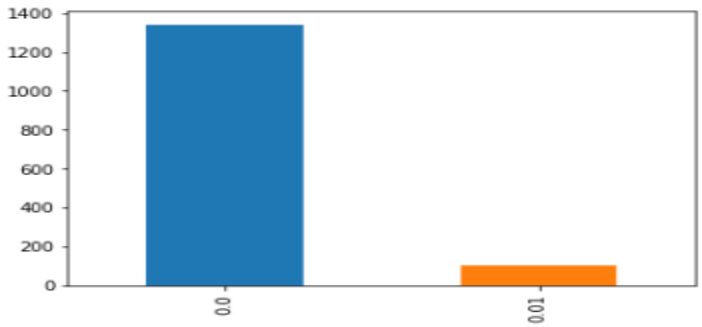
4) RESULTS

The cities were plotted using the Folium package and the following result was obtained:

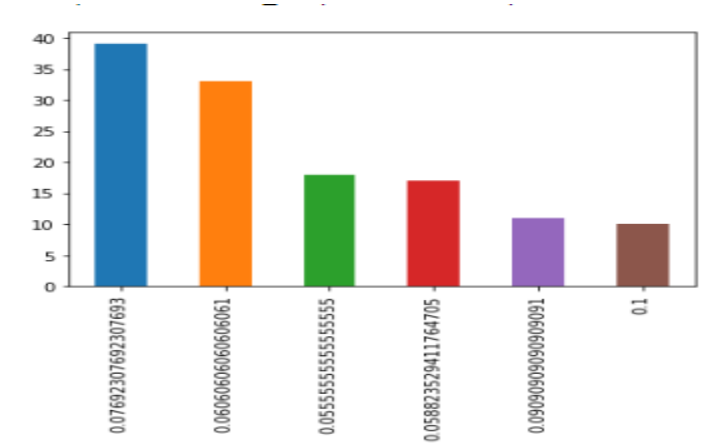
5) Observations

Cluster=0

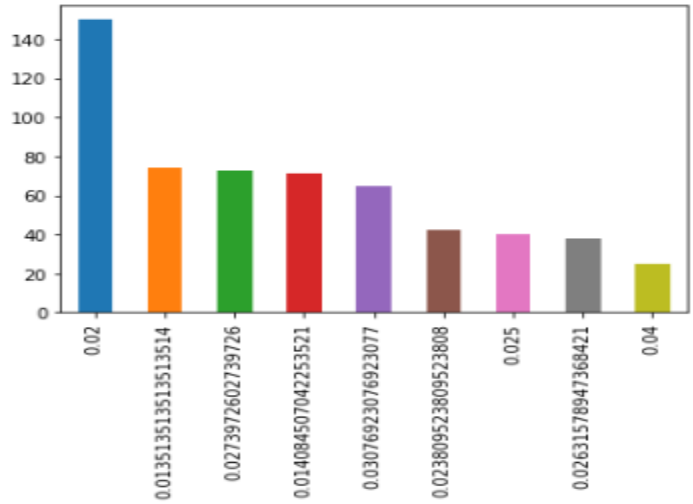
```
<matplotlib.axes._subplots.AxesSubplot at 0x1632b8999b
```



Cluster=1



Cluster=2



On the basis of above graph following observations are made

On the basis of clusters the following observations were made:

1. High number of GYMs exists in cluster 2.
2. Moderate number of Gyms exists in cluster 1.
3. Zero number of Gyms exists in cluster 0.

6) DISCUSSION

Based upon the results that were obtained it is safe to say that setting up a new gym at the location defined in cluster 2 will be very risky as very high concentration of gym are already present. Setting up a new gym there will result in competition. It is highly recommended that new gym should be setup at either location mentioned in cluster 1 or cluster 0, but preferably cluster 0.

7) CONCLUSION

In this project we have gone through the process identifying the business problem, specifying the data required, extracting the data, preparing the data, cleaning the data and performing machine learning by clustering the data based on similarities and lastly providing recommendations to the stakeholders i.e. developers and investors. It is recommended that setting up a new gym at locations mentioned in cluster 1 or cluster 0 can result in higher revenue generation