Applied Data Science Capstone

Segmenting and Clustering areas in Pune for new Gyms

Yuvraj Thakare

IBM Data Science Professional Certificate

**Introduction:**

Pune is very populated city and it has a lot of educational institutions.Therefore it has a large student population which comes to pune for educational purpose. It also has large IT sectors, attracting large amount of employees all around the India.

Considering all these factors there is more of youth population in Pune. In today’s generation, youth is giving importance to “Healthy Living” and more into Gym’s and workouts stuff. One more reason behind this is Social Media, youth is constantly active on social media posting pictures, making videos, etc. Hence, to look good fitness is really important. There is a lot of demand for gyms in pune but accordingly proper professional gyms are really less as compared to people.

That’s why through this project I’m trying to segment the neighborhoods of pune according to the availability of gyms to help the new gyms to open in locations where gyms are very less and have great potential for business.

**Data :**

• List of Neighborhoods in Pune .This basically sets a scope of our

project .

• Area Markers i.e Longitudes and Latitude of the Scope and also

the neighborhoods targeted for analysis.

• Venue data , for calculation of customer occurrences and

availability of specific venue in an area.

**Sources for Data Availability:**

• First and foremost Source for list of neighborhood in Pune can be

made available through Wikipedia . A csv file is created containing

the list which is further merged with area markers.

• The Coordinates of Neighborhood can be extracted from Python

Geocoder Package.• Venue data can be extracted from Foursquare.com site , which is

prestigious site four venue details , ratings etc. A Developer

account needs to be created so as to avail API calls.

**Methodology :**

Project Starts from gathering the data , initial list of neighborhoods is obtained

by creating a delimited file manually the list of neighborhoods was obtained

from Wikipedia (https://en.wikipedia.org/wiki/List\_of\_neighbourhoods\_in\_Pune) .

All the Areas (incl.Pimpri-chinchwad )were included in this file , the next step

was to gather the coordinates of these neighbourhoods along with coordinates

of Pune , this was swiftly done by using python geocoder package which easily

helped us to convert addresses to respective latitudes and longitudes , which

inturn where fed to Foursquare API , which primarily depends upon coordinates

to return the venues and details specifically residing in those coordinate ,for this

a developer account was created beforehand with limited API calls .

Next, we obtain top 100 venues allocated around (2km) radius , API calls were

used by feeding in the coordinates using a python loop which returned the

venue details in JSON format and key elements like venue name , venue

category ,venue latitude ,venue longitude were extracted . After extraction starts

the Analyzing part ,where each neighborhood was analyzed by calculating the

mean of frequency of occurrence of each venue category. Also after analyzing

we filter the data according to Gym venues as the project is limited to

this scope.

After all the data preparation K-means clustering is applied on prepared data

frame. *k*-means clustering tends to find clusters of comparable spatial extent,

while the expectation-maximization mechanism allows clusters to have

different shapes. Value of k is set to three (k=2) , i.e 2 clusters of varied

frequency Gym occurrences in different neighborhood . Result will allow

us to identify which neighborhood have higher or lower concentrations of Gyms ,it will help us answer the question as to which neighborhoods are

suitable to open new Gym.

**Results :**

The results after clustering (k=2) clearly exhibits that we can categorize the

neighborhoods into 2 clusters based on the frequency of occurrence for “Gym ”:

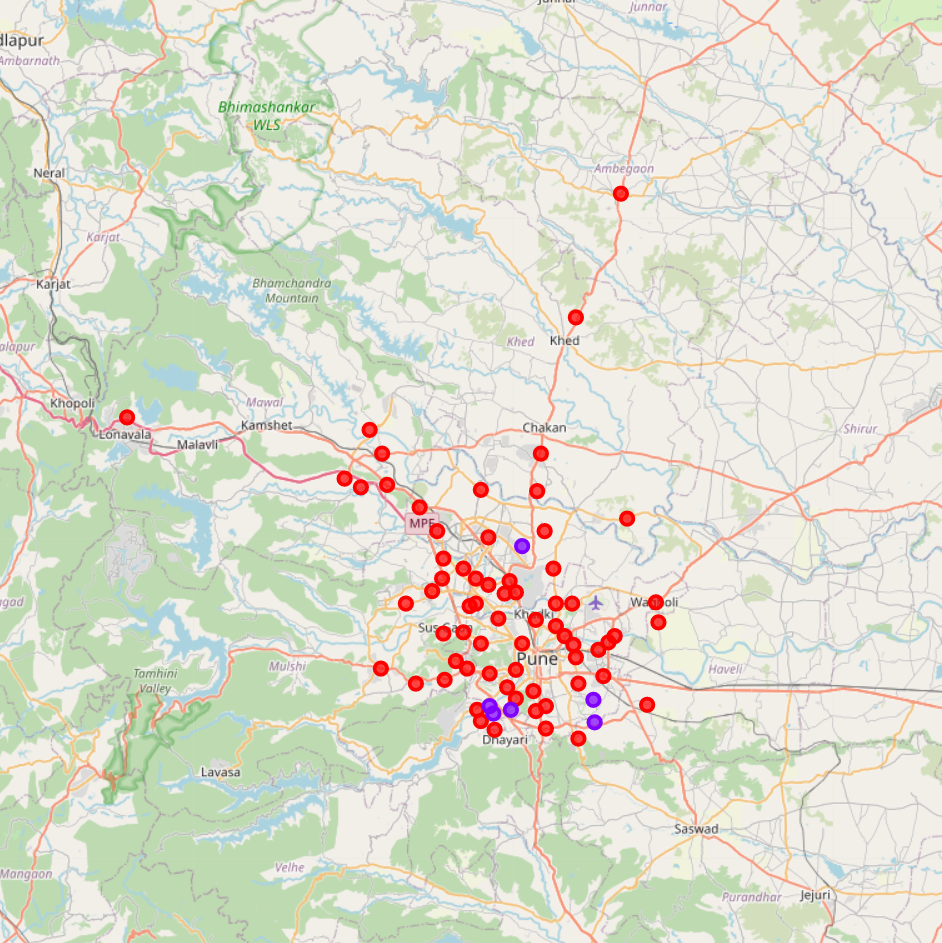


Fig.All the clusters are visualized with markers.

**Discussion :**

As inferred from the visualization of clusters it is clear that the franchise

targeting the areas in Cluster1 will be facing a lot of competition from its

peers as the number of gyms in this areas is highly concentrated

whereas Cluster0 serves to be the new market for cafes as these areas

gravely lack in those amenities , also Cluster0 can be a fair choice but

moderation in number wouldn’t hype up the sales but can surely attract

moderate amount of customers with fair amount of competition.

Best Areas to be targeted : Areas in Cluster 0

**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 2 clusters based on

their similarities, and lastly providing recommendations to the relevant

stakeholders i.e Gym franchise eager to invest in Pune region by opening a

new outlet. To answer the business question that was raised in the

introduction section, the answer proposed by this project is: The

neighborhoods in cluster 0 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 Gym.

**References:**

• Category: Neighborhoods in Pune , India .Retrieved from

( https://en.wikipedia.org/wiki/List\_of\_neighbourhoods\_in\_Pune)

• Foursquare Developers Documentation. Foursquare. Retrieved from

https://developer.foursquare.com/docs.

Thank

You.