Coursera Capstone
IBM Applied Data Science Capstone

Opening More Fitness Centers in Los Angeles, California

By Christian Barreto

Introduction: Business Problem

The Problem: There is an unmet demand for the people of Los Angeles to conveniently access a local gym or fitness center.

The Goal: Determine optimal locations for a new gym or fitness centers within Los Angeles, California.

Final Deliverable For Target Audience: A list of neighborhoods that can serve as key consumer areas for property developers, real-estate investors, and Los Angeles City Officials.

Data

The following data resources will be used for this project:

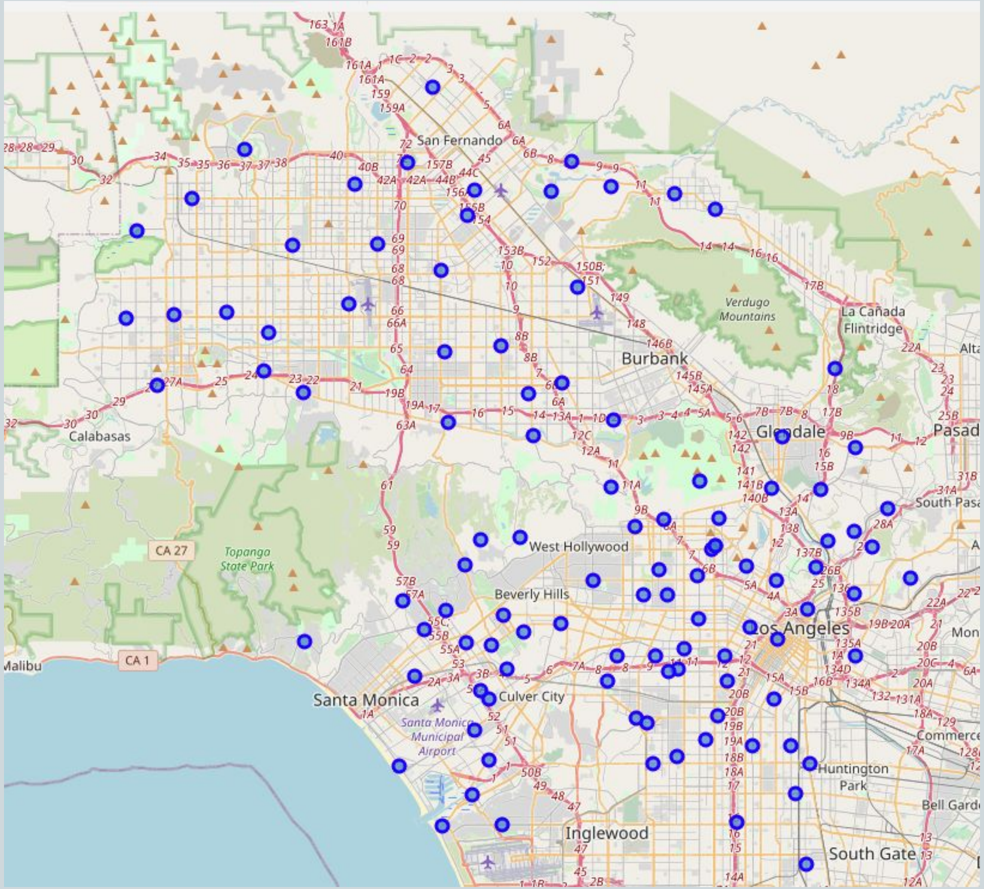
- **The official list of neighborhoods in Los Angeles.**
  - **Source:** geohub.lacity.org
- **The latitude and longitude coordinates of each neighborhood.**
  - **Source:** geopy.geocoder package
- **Fitness center venue data in proximity of each neighborhood.**
  - **Source:** Foursquare API

Methodology: Data Gathering & Geocoding

- Get CSV file found on Geohub that contains LA neighborhood lists.
- Upload CSV files to github repository (c2barreto/Coursera_Capstone)
- Query each neighborhood name with Geocoder API to obtain one latitude and longitude coordinate that is respective to their center point.
- Important detail: only one set of coordinates was required for each neighborhood since there is a limit to how many queries can be done using the free version of Foursquare API.

Methodology: Mapping Los Angeles

- Geographical coordinates that cover the entire area of Los Angeles were also queried using the geocoder package in order to create a Los Angeles Map using the Folium package.
- With Folium, each neighborhood coordinate was then superimposed on top of the Los Angeles map.



Methodology: Foursquare API (Part 1)

- Utilized Foursquare API in order to get the top venues that are within 2000 meters of each neighborhood coordinate.
- The Los Angeles Venue dataframe was structured to contain a row for each unique venue found along with columns that distinguish what neighborhood that venue belongs to, the venue coordinates, and what the venue category is.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Adams-Normandie	34.07809	-118.3012	Noshi Sushi	34.076159	-118.305374	Sushi Restaurant
1	Adams-Normandie	34.07809	-118.3012	Jaraguá	34.076364	-118.306646	Cocktail Bar
2	Adams-Normandie	34.07809	-118.3012	Kim Sun Young Hair Beauty Salon (Kim Sun Young...	34.076453	-118.308921	Salon / Barbershop
3	Adams-Normandie	34.07809	-118.3012	Guatemalteca Bakery	34.076303	-118.297168	Restaurant
4	Adams-Normandie	34.07809	-118.3012	Cactus Mexican Food	34.076194	-118.304147	Mexican Restaurant

Methodology: Foursquare API (Part 2)

- The Los Angeles Venue data frame was then grouped according to neighborhood
- A tally of how many venues were returned for each neighborhood so that they can be examined.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Adams-Normandie	100	100	100	100	100	100
Arleta	59	59	59	59	59	59
Arlington Heights	100	100	100	100	100	100
Atwater Village	100	100	100	100	100	100
Baldwin Hills/Crenshaw	78	78	78	78	78	78
Bel-Air	65	65	65	65	65	65
Beverly Crest	45	45	45	45	45	45
Beverly Grove	28	28	28	28	28	28
Beverlywood	100	100	100	100	100	100

Methodology: Foursquare API (Part 3)

- A one hot encoding function was then defined in order to create a matrix of all the unique venue categories that are present in the Los Angeles Venue data frame
- Then the mean of the frequency of occurrence of each venue category per neighborhood was applied.

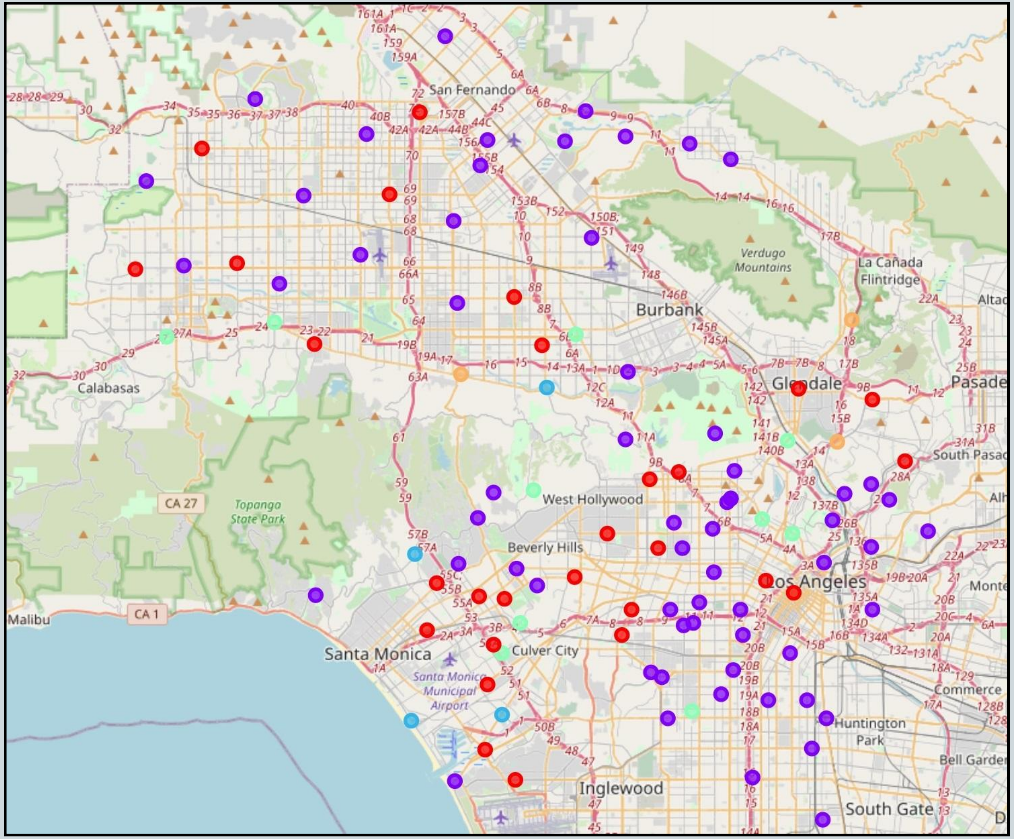
	Neighborhood	ATM	Accessories Store	Adult Boutique	Airport	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Amphitheater	Andhra Restaurant	Antique Shop
0	Adams-Normandie	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.000000	0.010000	0.000000	0.00	0.000000
1	Arleta	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.000000
2	Arlington Heights	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.000000	0.010000	0.000000	0.00	0.000000
3	Atwater Village	0.010000	0.00	0.00	0.000000	0.00	0.000000	0.000000	0.030000	0.000000	0.00	0.000000
4	Baldwin Hills/Crenshaw	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.000000	0.012821	0.000000	0.00	0.000000
5	Bel-Air	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.000000	0.030769	0.000000	0.00	0.000000
6	Beverly Crest	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.000000
7	Beverly Grove	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.000000
8	Beverlywood	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.000000	0.060000	0.000000	0.00	0.000000
9	Boyle Heights	0.010000	0.00	0.00	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.000000
10	Brentwood	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.000000	0.030928	0.000000	0.00	0.000000
11	Broadway-Manchester	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.000000	0.050000	0.000000	0.00	0.000000
12	Canoga Park	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.000000	0.030000	0.000000	0.00	0.000000
13	Carthay	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.000000
14	Central-Alameda	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.000000
15	Century City	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.000000	0.060000	0.000000	0.00	0.000000
16	Chatsworth	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.000000

Methodology: Clustering Neighborhoods

- The scope of this project only requires that gym and fitness center data be kept while everything else can be dropped.
- The final step was to perform clustering on the data related to fitness centers by using the k-means clustering algorithm.
- With K-means, 5 clusters were set and then the algorithm randomly assigned a neighborhood to a cluster based on the data of mean frequency of occurrence for gyms and fitness centers.

Results

- Cluster 0 (Red): 30 neighborhoods with a low concentration of fitness centers.
- Cluster 1 (Purple): 67 neighborhoods with no fitness centers in proximity.
- Cluster 2 (Blue): 4 neighborhoods with a mid-high concentration of fitness centers.
- Cluster 3 (Green): 10 neighborhoods with a low-mid concentration of fitness centers.
- Cluster 4 (Orange): 3 neighborhoods with a high concentration of fitness centers.



Discussion

- The majority of neighborhoods under Los Angeles county lack access to nearby fitness centers.
- Very few neighborhoods harbor a mid to high concentration of local fitness centers present. These are reflected in clusters 2, 3, and 4. Safe to recommend to avoid building new fitness centers in these areas.
- Notable geographic concentrations from cluster 1 that display a lack of fitness centers are South Los Angeles, East Los Angeles, Central Los Angeles, and the most northern regions of Los Angeles that hug San Fernando.
- These city regions from cluster 1 are also high population dense and low income areas so there is a high probability of demand for more convenient and cheap fitness centers in these areas.

Conclusion

- Despite there being some limitations as far as how much venue data can be queried from Foursquare, enough venue data was gathered where an accurate mean frequency of occurrence for gyms and fitness centers can be obtained for each Los Angeles neighborhood.
- From these mean frequency of occurrences for gyms and fitness centers, five unique clusters were able to be made and each held notable geographic insights.
- In Particular the most stand out cluster that can be referred to as the ideal list of neighborhoods to target for building new fitness centers is cluster 1.
- Property developers should aim to build a market of low-cost membership fitness centers in these regions.