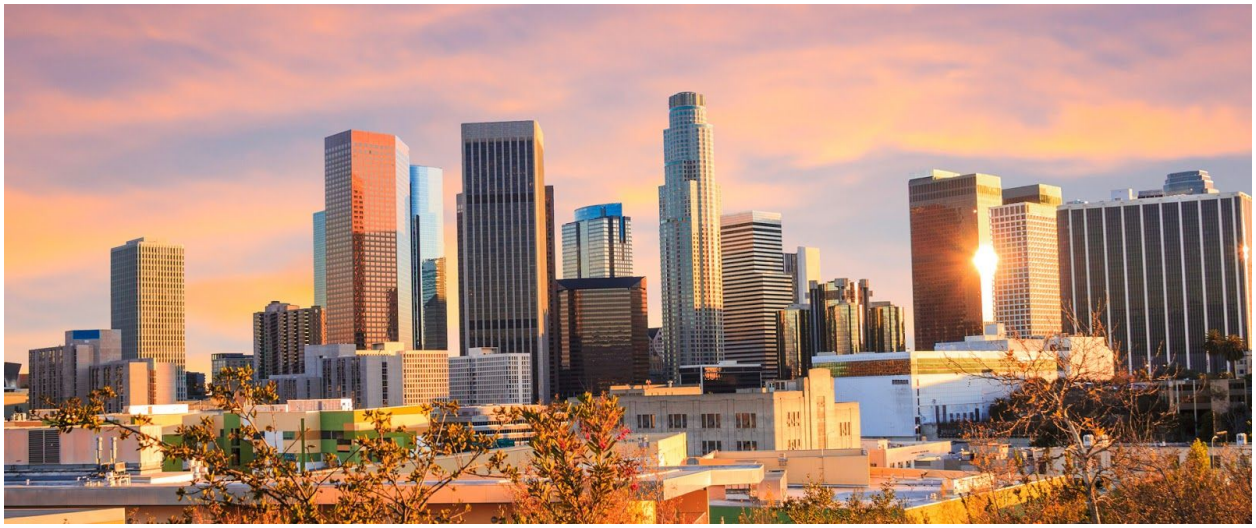
Coursera Capstone

IBM Applied Data Science Capstone

Opening More Fitness Centers in Los Angeles, California

By: Christian Barreto

November 2020



Introduction

The age of big data has just begun and there is already an ocean of information on consumer interest, habits, and spending for the past decade. Now more than ever, businesses can leverage modern data science practices where they can truly measure what are ideal business ventures to pursue and where are the prime locations to target. That being said it doesn't take a data scientist to determine that the people of Los Angeles commute large distances often and spend a good portion of their day sitting in their vehicles for the inevitable traffic jams that occur on their way to and from work. It also doesn't take a data scientist to determine that the american workforce is becoming more sedentary with new collar jobs on the rise. In other words, sitting, lots of sitting, is happening throughout the daily lives of angelenos and that trend needs to be offset for the sake of maintaining a healthy population. For property developers searching for a lucrative opportunity, this sedentary population dilema begs the question: is it time to build more gyms and fitness centers throughout the city of Los Angeles? More specifically, which neighborhoods in Los Angeles have yet to be tapped in regards to opening a community fitness center?

Business Problem

There is an unmet demand for the people of Los Angeles to conveniently access a local gym or fitness center without having to add another long commute to their daily routine. The goal is to determine optimal locations for a new gym or fitness centers within Los Angeles, California. Specifically, the end result is to compile a list of neighborhoods that can serve as target consumer areas for stakeholders interested in opening a gym or fitness center in Los Angeles, California. The neighborhoods of interest will be those with no gyms or fitness centers within 3000 meters from their center points.

Target Audience

The data within this report will be aimed at commercial property developers and real-estate investors that are in the market to open gyms or fitness centers in Los Angeles, California. Los Angeles City health officials can also benefit from having data on which neighborhoods lack sufficient recreational fitness resources as the city is still struggling with reducing the percentage of overweight individuals. According to data gathered by the Centers for Disease Control and Prevention, between the years 2011-2018, 35-36% of adults in California aged 18 years and older have consistently been classified as overweight. In addition, of that same group and time frame, 24-26% have obesity.

Data

The following data resources will be used for this project:

1. The official list of neighborhoods in Los Angeles.
    - Description: Names of each Los Angeles neighborhood are listed. The scope of this project does not extend to the greater Los Angeles area and is confined to the official neighborhood boundaries under Los Angeles.
    - Data Source: geohub.lacity.org
2. Coordinates for each Los Angeles neighborhood.
    - Description: Latitude and longitude coordinates for the center of each Los Angeles neighborhood. These coordinates will be used to plot circle markers over a map of Los Angeles and to explore each of them.
    - Source: geopy.geocoder package
3. Venue Data in proximity of each Los Angeles neighborhood.
    - Description: Using Foursquare API a search query will be done against the coordinates for each Los Angeles neighborhood and venue data that is within a radius of 3000 meters of each neighborhood will be generated. This venue data will then be filtered to account for only data related to gyms and fitness centers.
    - Data source: Foursquare API

Methodology

Data Gathering

The list of Los Angeles neighborhoods needs to be gathered first. This information can be conveniently downloaded from a csv file found in the city's platform for location-based open data, Geohub (geohub.lacity.org). For additional convenience this CSV file with the list of Los Angeles neighborhoods was also uploaded to the public Github repository that will be used to store all files related to this project (c2barreto/Coursera_Capstone).

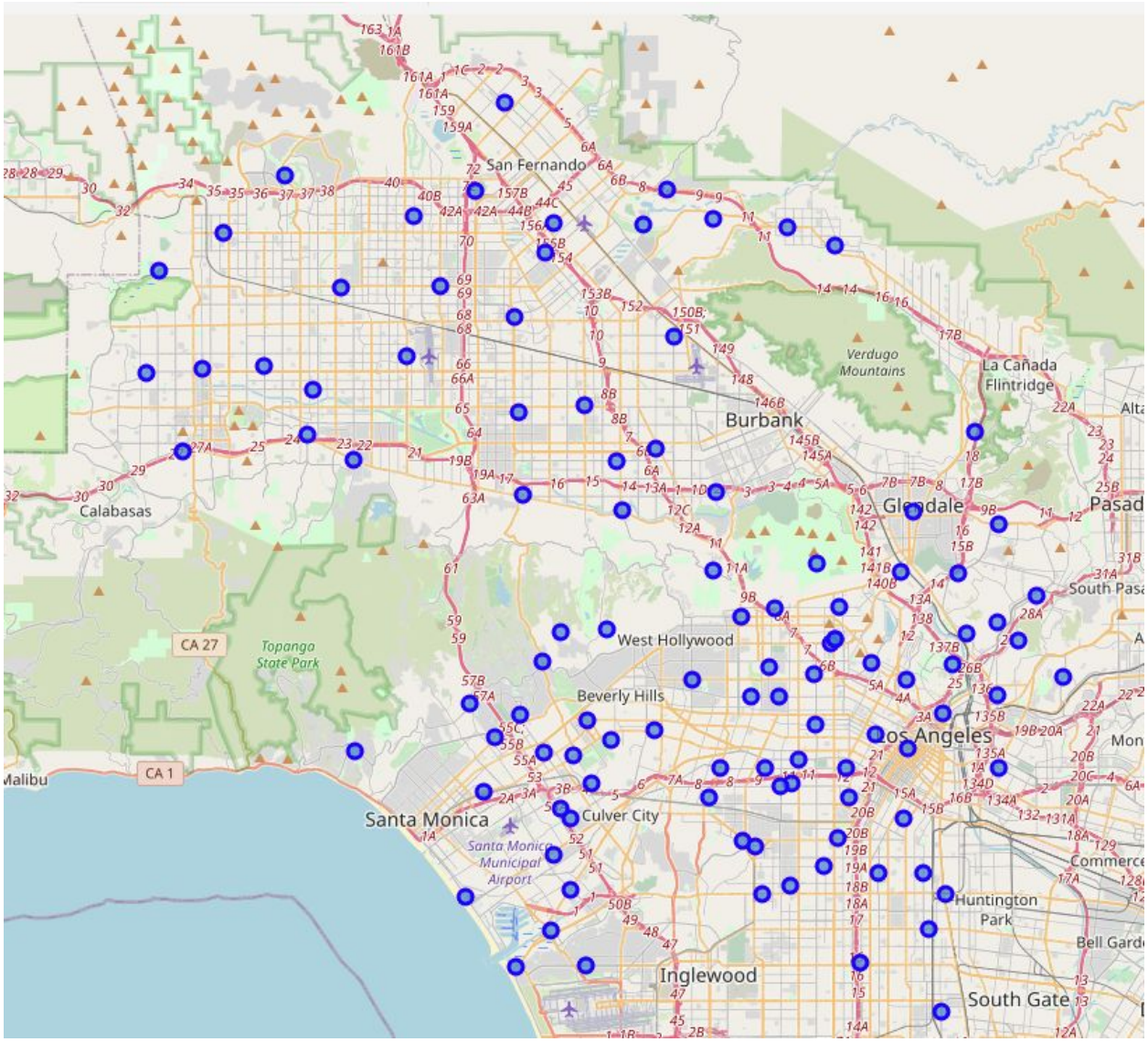
Geocoding

With just the list of their unique names, the geographical coordinates of each Los Angeles neighborhood can be obtained utilizing the Geocoder package in Python. More specifically, each neighborhood name was queried with the Geocoder API to obtain one latitude and longitude coordinate that is respective to their center point. From there on those coordinates were merged onto a pandas data frame that holds each neighborhood name and their latitude and longitude coordinates. These neighborhood

coordinates are needed later to query for venue data that is in the vicinity of those neighborhoods. One important detail to note is that for this project, 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.

Data Visualization: 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.



Exploring Neighborhoods With Foursquare API

Next was to utilize Foursquare API in order to get the top venues that are within 2000 meters of each neighborhood coordinate. Since the free version of Foursquare API was used, there is a limit of up to 100 venues that can be generated for each neighborhood coordinate. A python function was defined in order to loop through multiple API calls to Foursquare until every neighborhood coordinate was queried. The output from Foursquare is one large JSON file but all the relevant data that is desired for the project is then passed onto a dataframe. 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

The Los Angeles Venue data frame was then grouped according to neighborhood so that a tally of how many venues were returned for each neighborhood 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

Furthermore, 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, which totaled to about 423 categories out of the 9771 venues collected. 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

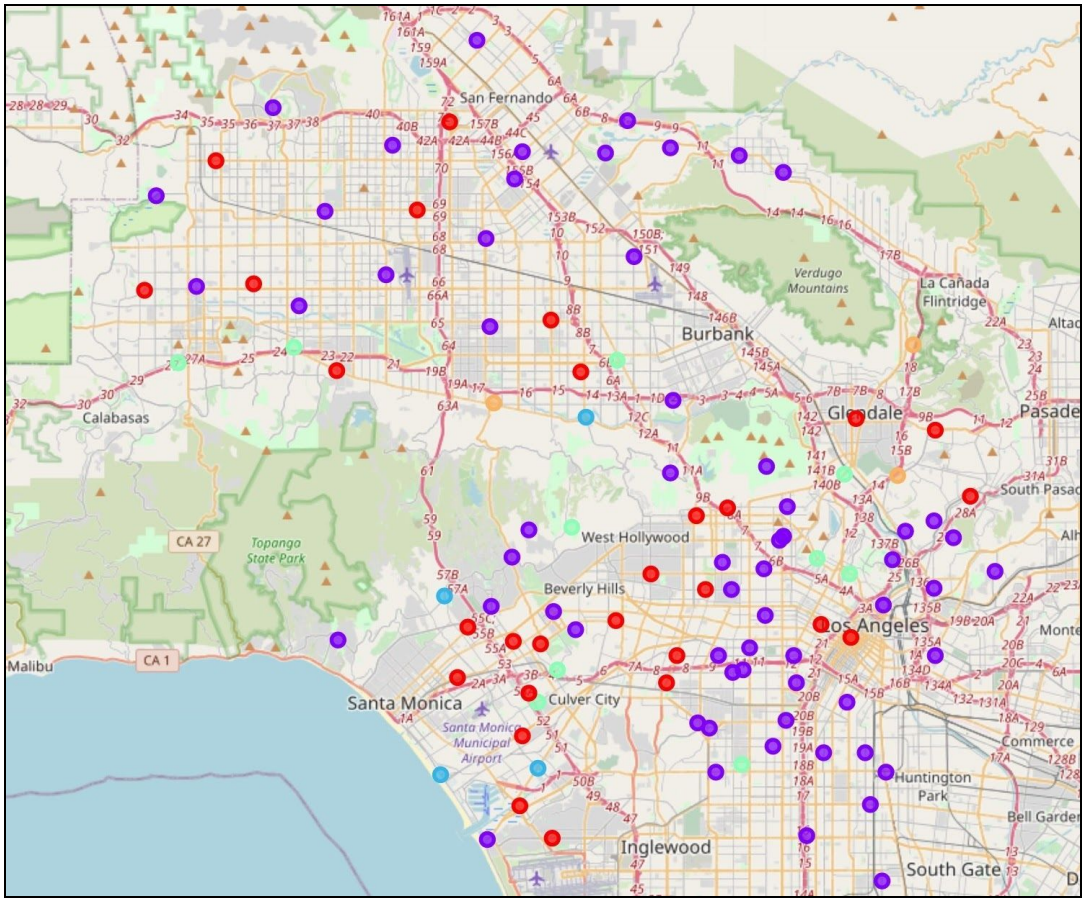
Clustering Neighborhoods by Gym & Fitness Center Data

Despite having gathered information for several different types of venues for each neighborhood, the scope of this project only requires that gym and fitness center data be kept while everything else can be dropped. At this point, all that remained was a column on neighborhoods and a column of the mean frequency of occurrence for gyms and fitness centers within 3000 meters of each neighborhood. 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

The results of the 5 clusters ultimately identify which neighborhoods have a low to high concentration of fitness centers and which neighborhoods have none.

- **Cluster 0 (Red)**: Cluster 0 groups together 30 neighborhoods with a low concentration of fitness centers in proximity. The venue data for these neighborhoods generated at least 1 occurrence of a fitness center.
- **Cluster 1 (Purple):** Cluster 1 groups together 67 neighborhoods with no existence of fitness centers in proximity. The venue data for these neighborhoods generated 0 occurrences of a fitness center.
- **Cluster 2 (Blue):** Cluster 2 groups together 4 neighborhoods with a mid-high concentration of fitness centers in proximity. The venue data for these neighborhoods generated 3 occurrences of a fitness center.
- **Cluster 3 (Green):** Cluster 3 groups together 10 neighborhoods with a low-mid concentration of fitness centers in proximity. The venue data for these neighborhoods generated at least 2 occurrences of a fitness center.
- **Cluster 4 (Orange):** Cluster 4 groups together 3 neighborhoods with a high concentration of fitness centers in proximity. The venue data for these neighborhoods generated at least 4 occurrences of a fitness center.



Discussion:

Based on the clustering data over the map of Los Angeles, 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. In fact, clusters 2, 3 and 4, which contain a mid to high concentration of fitness centers, can be combined and they would only total 17 neighborhoods out of the 114. Unsurprisingly, clusters 2, 3, and 4 contain suburban neighborhoods with a higher income population. Judging also by their higher concentration of fitness centers, it is safe to recommend to avoid building new fitness centers in these areas since they already have competition established. Even more so, the intent is to give more local access to neighborhoods that lack fitness centers.

Continuing on, the second largest cluster is cluster 0, which lists neighborhoods with at least one occurrence of a fitness center in proximity. Cluster 0 contained 30 neighborhoods out of the 114 in Los Angeles and geographically was concentrated towards Western Los Angeles and North West Los Angeles. Competition for memberships is probably not high in these neighborhoods so they can be considered as a secondary tier of neighborhoods for building more fitness centers around.

Every neighborhood that did not generate an occurence of a fitness center within a 2000 meter radius from their center point was assigned to Cluster 1. Cluster 1 had a total of 67 out of 114 neighborhoods assigned to its cluster. Geographically, the neighborhoods under cluster 1 that are located in the most southern side of Los Angeles have the least availability to access a fitness center in proximity.This discrepancy could be due to the fact high concentration of low income residents reside in the southern side of Los Angeles which in the past might have deterred property developers from establishing fitness centers there.

In Addition, there are also several other notable geographic concentrations from cluster 1 that display a lack of available fitness centers are 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 areas so there is a high probability of demand for more convenient 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. Cluster 1 revealed that there are at least three big and populous regions in Los Angeles that have zero fitness centers established and they are Southern Los Angeles, East los Angeles, and Central Los Angeles. While these regions have historically housed low-income residents, this factor should not limit property developers from pursuing to build fitness centers in these regions. In fact, property developers should aim to build a market of low-cost membership fitness centers in these regions. Better yet, the city of Los Angeles can serve the health needs of its low-income residents better by building community fitness recreational centers in these neighborhoods as well. Overall, there is a market and need to build more fitness centers for the neighborhoods listed under cluster 1.