Finding the location of new gym in L.A.

June 7, 2020

Yuan-Ming Hsu

1. Introduction

- *Background*

L.A. is the city which has multiple cultures, huge diversity and population composition. It's the largest city in California, which has estimated 4 million people living in this area. The city of Los Angeles and its neighborhood are the home of eleven top level professional sports teams from different leagues like NBA, NHL, MLS and WNBA. Sport industry is popular in this area so it can be seen how competitive this industry is here. Gym is the common place that everyone can use and can be found everywhere in cities. To invest new gym in a place, we should consider the location first. Good place decides the basis of cash flow or income of enterprise.

- *Problem*

The project aims to find the potential location of gym which can be invested in the populated area.

- *Interest*

People who want to run a gym business in L.A area and figure out what location would have low competition and enough population.

2. Data acquisition

- *Data sources*

Data mainly come from two sources which are made by L.A. times and foursquare:

Neighborhood geospatial data and population data can be found in its website.

The first dataset contains the latitude and longitude of neighborhood area in L.A. but it provides the boundary location of neighborhood instead of center location. It will need some data pre-processing step to extract the information I need as shown in Table 1.

Table 1. Geospatial data in L.A.

	Neighborhood	Latitude	Longitude
0	Acton	34.49548	-118.172
1	Adams-Normandie	34.03226	-118.302
2	Agoura Hills	34.14012	-118.744
3	Agua Dulce	34.49395	-118.309
4	Alhambra	34.08308	-118.136

The second dataset contains the average population in every neighborhood area so it gives us a basic information to evaluate if the place has enough population as shown in Table 2

Table 2. Average population in L.A

	Rank	Neighborhood	Population per Sqmi
0	1	Koreatown	42611
1	2	Westlake	38214
2	3	East Hollywood	31095
3	4	Pico-Union	25352
4	5	Maywood	23638

The last data is from foursquare website which has various of venues data. But it need to sign up an account to retrieve the data from the website.

Table 3. Venue data in L.A.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Acton	34.49548	-118.172	Alma Gardening Co.	34.49476	-118.173	Construction & Landscaping
1	Adams-Normandie	34.03226	-118.302	Orange Door Sushi	34.03249	-118.299	Sushi Restaurant
2	Adams-Normandie	34.03226	-118.302	Shell	34.0331	-118.3	Gas Station
3	Adams-Normandie	34.03226	-118.302	Loren Miller Recreational Park	34.03134	-118.304	Playground
4	Adams-Normandie	34.03226	-118.302	Tacos La Estrella	34.03223	-118.301	Taco Place

3. Exploratory analysis

To give a basic idea of these data look like, I plot the center location of every neighborhood on the map. L.A geospatial data shows that there are 265 neighborhoods L.A area as shown in Figure 1.

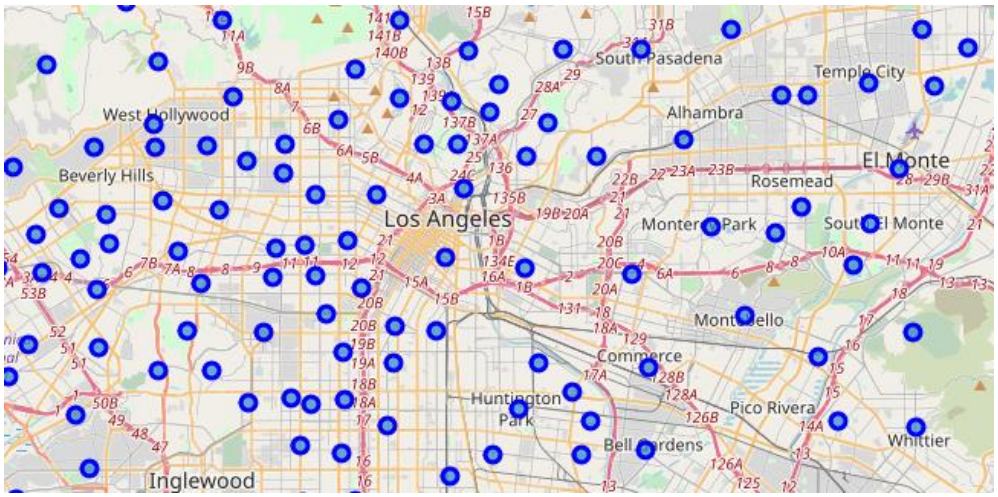

Figure 1. Center location of neighborhood in L.A.

The population data shows that the highest population is in Koreatown, the second highest is Westlake and the third highest is East Hollywood which are only three area higher than 30000 person per square miles. But if I look at the venue data in these three areas, it's not the target I have. To filter out the area

that fit our target, I set the threshold of population should be higher than 15000 person per square miles and there is at least one gym and the gym is not popular in this area. Based on this rule, I can sort out the candidate area - Hollywood.

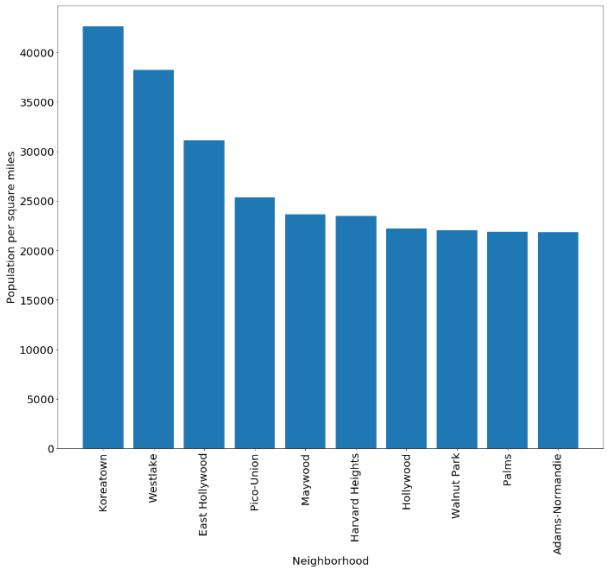

Figure 2. Rank of average population of neighborhood in L.A.

As seen in Table 4, the top 10 venue in Hoolywood doesn't include gym but there is at least one gym and the population in this area also higher than 15000 person per square miles.

Table 4. Selected area in L.A.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Population per Sqmi
1	Hollywood	Coffee Shop	Museum	Thai Restaurant	Hotel	American Restaurant	Movie Theater	Juice Bar	Fast Food Restaurant	Cosmetics Shop	Chocolate Shop	22193

Here's one ***thing to note that Hollywood has a yogurt place which isn***'t categorized as gym***. But I see it as a gym as shown in Table 5.***

Table 5. Venue list in Hollywood

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Hollywood	34.098904	-118.339189	L'Antica Pizzeria da Michele	34.099063	-118.337444	Italian Restaurant
Hollywood	34.098904	-118.339189	Qwench Juice Bar	34.097824	-118.340460	Juice Bar
Hollywood	34.098904	-118.339189	In-N-Out Burger	34.098273	-118.341667	Fast Food Restaurant
Hollywood	34.098904	-118.339189	El Capitan Theatre	34.101102	-118.339794	Movie Theater
Hollywood	34.098904	-118.339189	Drnk Coffee & Tea	34.097859	-118.340388	Coffee Shop
Hollywood	34.098904	-118.339189	vegan house thai bistro	34.097178	-118.338794	Thai Restaurant
Hollywood	34.098904	-118.339189	Chick-fil-A	34.097757	-118.338254	Fast Food Restaurant
Hollywood	34.098904	-118.339189	Catalina Bar and Grill	34.098222	-118.337471	Jazz Club
Hollywood	34.098904	-118.339189	Javista Organic Coffee Bar	34.098218	-118.336544	Coffee Shop
Hollywood	34.098904	-118.339189	The Hollywood Museum	34.101087	-118.338613	Museum
Hollywood	34.098904	-118.339189	Ghirardelli Soda Fountain & Chocolate Shop	34.101697	-118.339765	Chocolate Shop
Hollywood	34.098904	-118.339189	Wanderlust Hollywood	34.095953	-118.339080	Yoga Studio

4. Modeling

Here only k-means method would be used to decide the boundary of the place because K-means method is a distance-based metric to cluster the data point into several regions. Our target is to find the region that doesn't ***include the existing gym. This kind of area would be a potential area to invest the gym.***

The blue dot refers to the all venues in Hollywood while the red dot gives the place of gym and yoga studio. If the number of cluster is 3, then it can get 3 clustered area like Figure 3.

Figure 3 shows there's only one place doesn't contain the gym – Selma Avenue. But it can be observed the distribution of place is not even. These venues locate in the upper and bottom side of Selma Avenue and there are more venues near the northwestern side of Selma Avenue. So it comes out the idea that we should select the location which is closer to the northwestern boundary of Selma Avenue instead of center location of circle.

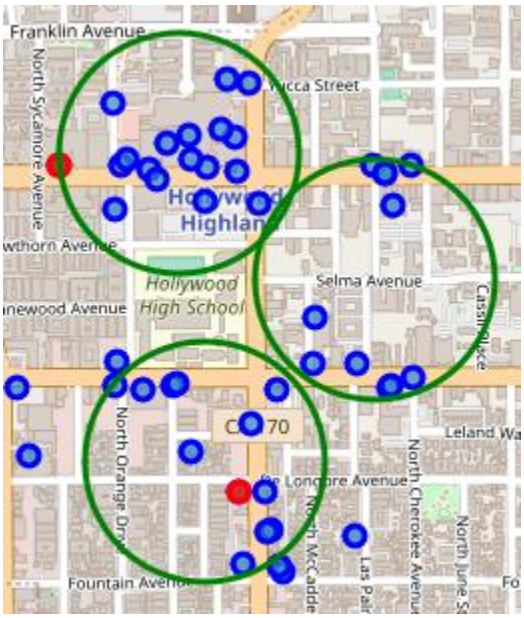

Figure 3. Clustered result for potential location in Hollywood

5. Conclusion

In this project, I analyzed the venue/population/neighborhood data from L.A area to figure out if there's a place would be a potential location for gym investment. After applying some filtering criteria and kmean method on the clean dataset, it gives the Selma Ave would be a place to invest and based on the observation, it could be closer to northwestern side of this area.

6. Future direction

To improve the model performance, we can consider more data together and put into the model like

1. ***Average wage in every area***
2. ***Average age in every area***