Best City in Los Angeles Country to Open a Fitness Store

Elias Lampietti

July 17, 2019

1. Introduction

1.1 Background

Los Angeles is home to Hollywood, one of the largest centers in the entertainment industry. This attracts upcoming actors and actresses looking for a way to break into the industry and launch their career. One quality that casting companies often look for is an actor or actress that looks healthy and takes care of themselves. This results in many aspiring actors and actresses joining a gym and maintaining a healthy lifestyle. This goal of maintaining a healthy physical appearance creates a high demand for gyms and fitness stores that carry products used to supplement training such as protein powder or pre workout.

1.2 Problem

Fitness centers in Los Angeles County are necessary and highly demanded in this county, which raises the question - where is the best city to open one?

1.3 Interest

The project would mainly be targeting aspiring small business owners looking to open up their own fitness store. This project will provide guidance towards the best and worst places to open up the store. These clients will be targeting actors and actresses, however the scope can expand to generally any active person looking for a store to help supplement their active and healthy lifestyle.

2. Data

2.1 Data Acquisition

The data source for this project is CSV file containing data for all of the zip codes in the US from https://simplemaps.com/data/us-zips. This contains the longitude, latitude, county name, and city for every zip code in the United States.

2.2 Data Cleaning

Since this file is too large to upload in its entirety to Watson Studio, I isolated only the rows with "Los Angeles" in the "county_name" column and created a new CSV file with these values. I then uploaded this in my IBM Cloud Object Storage as an asset for the notebook.

Since the data is indexed by zip code instead of cities, there are duplicate rows with the same value for "city" with only slight differences in their longitude and latitude values. Since I am only interested in cities as a whole, I grouped the rows by city and used the average of their latitudes, longitudes, populations, and densities as the new aggregate rows (Table 1).

Table 1. The head of a cleaned data frame of all Los Angeles cities.

	city	lat	lng	population	density
0	Acton	34.46511	-118.214160	7993.0	42.9
1	Agoura Hills	34.12274	-118.757270	25488.0	288.6
2	Alhambra	34.08285	-118.136885	41528.5	4128.4
3	Altadena	34.19544	-118.137960	36126.0	1689.5
4	Arcadia	34.13232	-118.037450	32905.0	2153.1

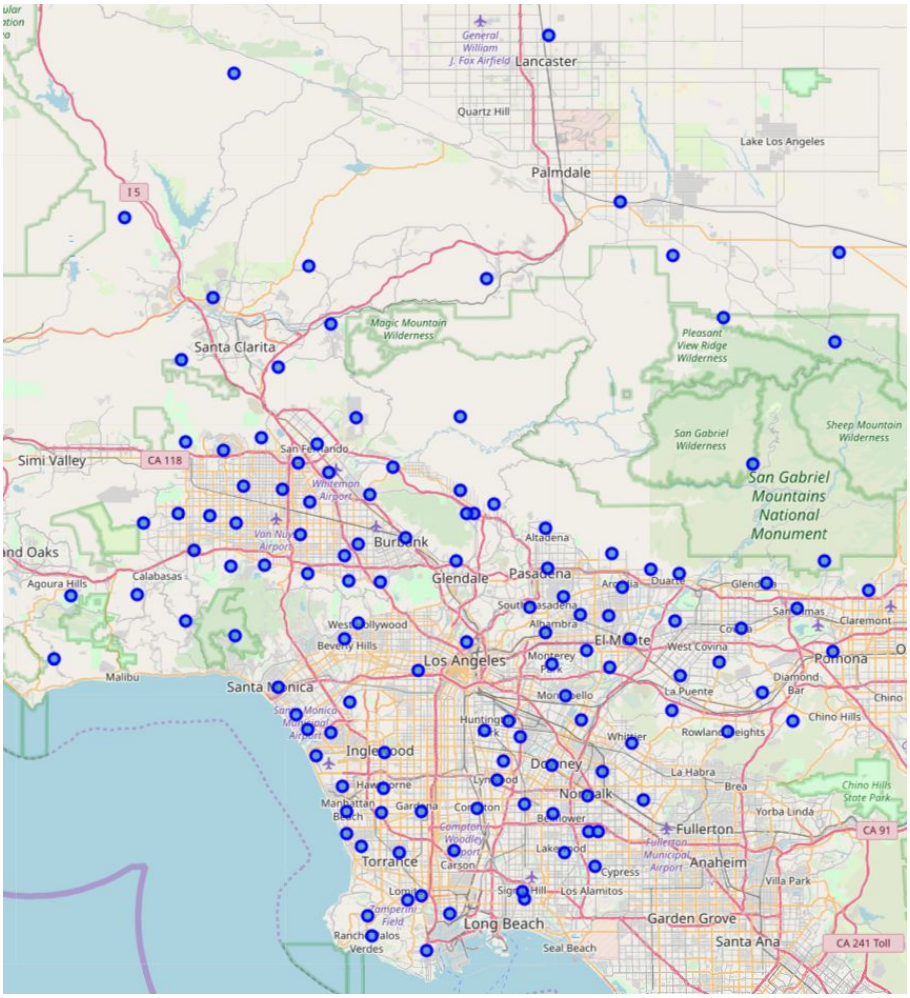
3. Methodology

3.1 Exploratory Data Analysis

In order to better visualize the geographical layout of Los Angeles County and the cities I will be ranking, I create a map using Folium centered on Los Angeles County. This map utilizes the latitude and longitude of each city from my data frame in order to pinpoint and mark each city on a map (Figure 1).

This map reveals that there are a few outliers amongst the marked cities that are far from urban centers and most likely not prime locations to open up a new fitness store due to a lack of traffic. I will keep this in mind while analyzing my results to see if any of these outliers were reported as good locations due to interfering factors.

Figure 1. Folium map with markers for each city in LA County.



The next step was to explore the venues located within the bounds of a certain city and familiarize myself with the quantity and category of the top venues in the city. In order to retrieve this information, I utilized the Foursquare API with the latitude and longitude of Los Angeles City to get the top 100 venues within a 500 meter radius of the city center (Table 2).

Table 2. Head of a data frame with the venues in Los Angeles City

	name	categories	lat	lng
0	Blu Elefant Café	Café	34.039827	-118.303951
1	Skinny B*tch Pizza	Pizza Place	34.039816	-118.298767
2	Regular Guys Pizza	Pizza Place	34.039885	-118.298287
3	Dollar Tree	Discount Store	34.044013	-118.301424
4	Burger Factory	Burger Joint	34.039826	-118.305305

3.2 Classification

The next step is to retrieve all the venues for each city in Los Angeles County and then sort the cities with those having the most gyms and parks at the top of the list and those with none or close to none at the bottom.

In order to accomplish this I repeated the process I used to retrieve all the venues in Los Angeles City, but this time for every city in Los Angeles County. This required iterating through the original dataframe, and sending the latitude and longitude coordinates for each city through to the Foursquare API in order to retrieve all the venues for every city (Table 3).

Since I am not interested in the distinct venues themselves, but rather the category they belong to, I needed to manipulate the data frame so that the columns were only venue categories indexed by the city name. These columns would be one hot encoded for each venue category and each row would represent a venue (Table 4).

This resulted in a lot of rows having the same value in the city column as each row represents a venue, however I am only interested in the city as a whole. Therefore, I

grouped the dataframe by city and used the mean value for each category as the new aggregate (Table 5).

Now I have a data frame that is populated with many venue categories that have no relevance in this project, so I searched the list of unique venue categories and combined all categories related to gyms or fitness together and all venues related to parks or trails together. The "Gyms" category consists of "Gym / Fitness Center", "Gym", "Gymnastics Gym", and "Boxing Gym" while the "Parks" category consists of "Park" and "Trail". I then trimmed the data frame so that I now have only 3 columns, "City", "Gyms", and "Parks" sorted first by prevalence of gyms and then by parks (Table 6).

I put more weight on gyms when comparing cities because parks contain people of all ages, including elderly people who go there solely to relax and are generally not interested in a fitness store. This is different from people who go to gyms that pay for a membership and therefore are more likely to be willing to purchase fitness supplements from a fitness store.

Table 3. Head of a data frame containing the venues for each city in Los Angeles County

	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Agoura Hills	34.12274	-118.75727	Final Construction Clean Up	34.123121	-118.757486	Construction & Landscaping
1	Agoura Hills	34.12274	-118.75727	RG Electric Services - Agoura Hills Electrical...	34.123963	-118.757025	Electronics Store
2	Agoura Hills	34.12274	-118.75727	Bwana Trail	34.121424	-118.753927	Trail
3	Agoura Hills	34.12274	-118.75727	Dead Man Overlook	34.123557	-118.761030	Scenic Lookout
4	Agoura Hills	34.12274	-118.75727	Medicine Woman Trail	34.118563	-118.757534	Trail

Table 4. Head of a data frame containing the one hot encoded venue categories for each venue

	City	ATM	Adult Boutique	American Restaurant	Antique Shop	Arcade	Argentinian Restaurant	Art Gallery	Arts & Crafts Store	Arts & Entertainment	...
0	Agoura Hills	0	0	0	0	0	0	0	0	0	...
1	Agoura Hills	0	0	0	0	0	0	0	0	0	...
2	Agoura Hills	0	0	0	0	0	0	0	0	0	...
3	Agoura Hills	0	0	0	0	0	0	0	0	0	...
4	Agoura Hills	0	0	0	0	0	0	0	0	0	...

Table 5. Head of a data frame grouped by city with mean values for each venue category

	City	ATM	Adult Boutique	American Restaurant	Antique Shop	Arcade	Argentinian Restaurant	Art Gallery	Arts & Crafts Store	Arts & Entertainment	...
0	Agoura Hills	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
1	Alhambra	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
2	Altadena	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
3	Arcadia	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
4	Artesia	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...

Table 6. Head of a data frame showing the prevalence of gyms and parks in each city

	City	Gyms	Parks
45	Lynwood	0.333333	0.0
91	Sunland	0.250000	0.0
93	Tarzana	0.250000	0.0
17	Culver City	0.200000	0.0
24	Encino	0.200000	0.0

3.3 Mapping

The final step is to split the data frame up into 4 sections representing “Great Location”, “Good Location”, “Average Location”, “Bad Location”, and map these accordingly using markers and Folium.

Folium requires latitude and longitude coordinates in order to place a marker, so the first step was to merge the current dataframe with the original data frame’s latitude and longitude columns with the shared index of “City” (Table 7).

Next I decided to split the data frame into the 2 main sections dependent on whether a city contained at least one non-zero value for either the “Gyms” or “Parks” venue category. I then split the section that contained cities with non-zero values for “Gyms” and “Parks” and split that section into 3 sections. The cities with the highest values for the “Gyms” and then “Parks” would be in the first section representing “Great Locations” while the following 2 would represent “Good Locations” and “Average Locations”. Finally the second main section consisted of only cities that had zero values for both “Gyms” and “Parks”, this would make up the “Bad Locations”.

Now that I have my sections of the data frame sorted accordingly, I can plot them on a folium map with each location ranging from “great” to “bad” marked with a different color.

Table 7. Head of the merged data frame with longitude and latitude values

	City	Gyms	Parks	lat	lng
46	Lynwood	0.250000	0.0	33.923650	-118.20053
93	Tarzana	0.250000	0.0	34.155080	-118.54751
17	Culver City	0.222222	0.0	34.008005	-118.39321
23	El Segundo	0.200000	0.0	33.916950	-118.40206
24	Encino	0.200000	0.0	34.155630	-118.50449

4. Results

4.1 Maps

This map is marked with red markers representing “Bad Locations”, yellow markers representing “Average Locations”, blue markers representing “Good Locations”, and green markers representing “Great Locations”. The first generated map displays all of these markers in Los Angeles Country (Figure 2).

Next I mapped each set of locations separately, including “Great Locations” (Figure 3), “Good Locations” (Figure 4), “Average Locations” (Figure 5), and “Bad Locations (Figure 6).

Figure 2. A folium map displaying all of the locations from bad to great

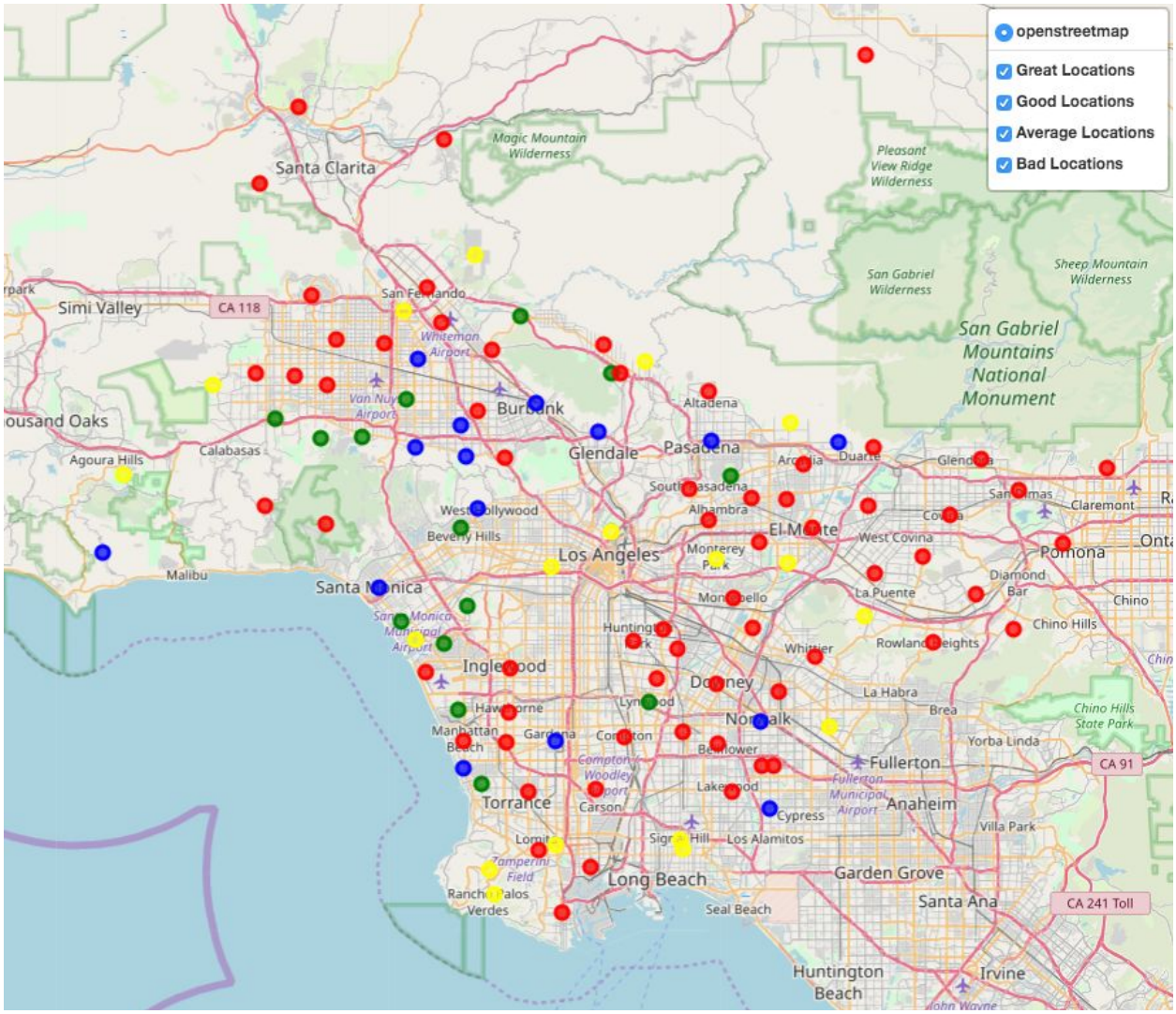


Figure 3. A folium map displaying all of the great locations

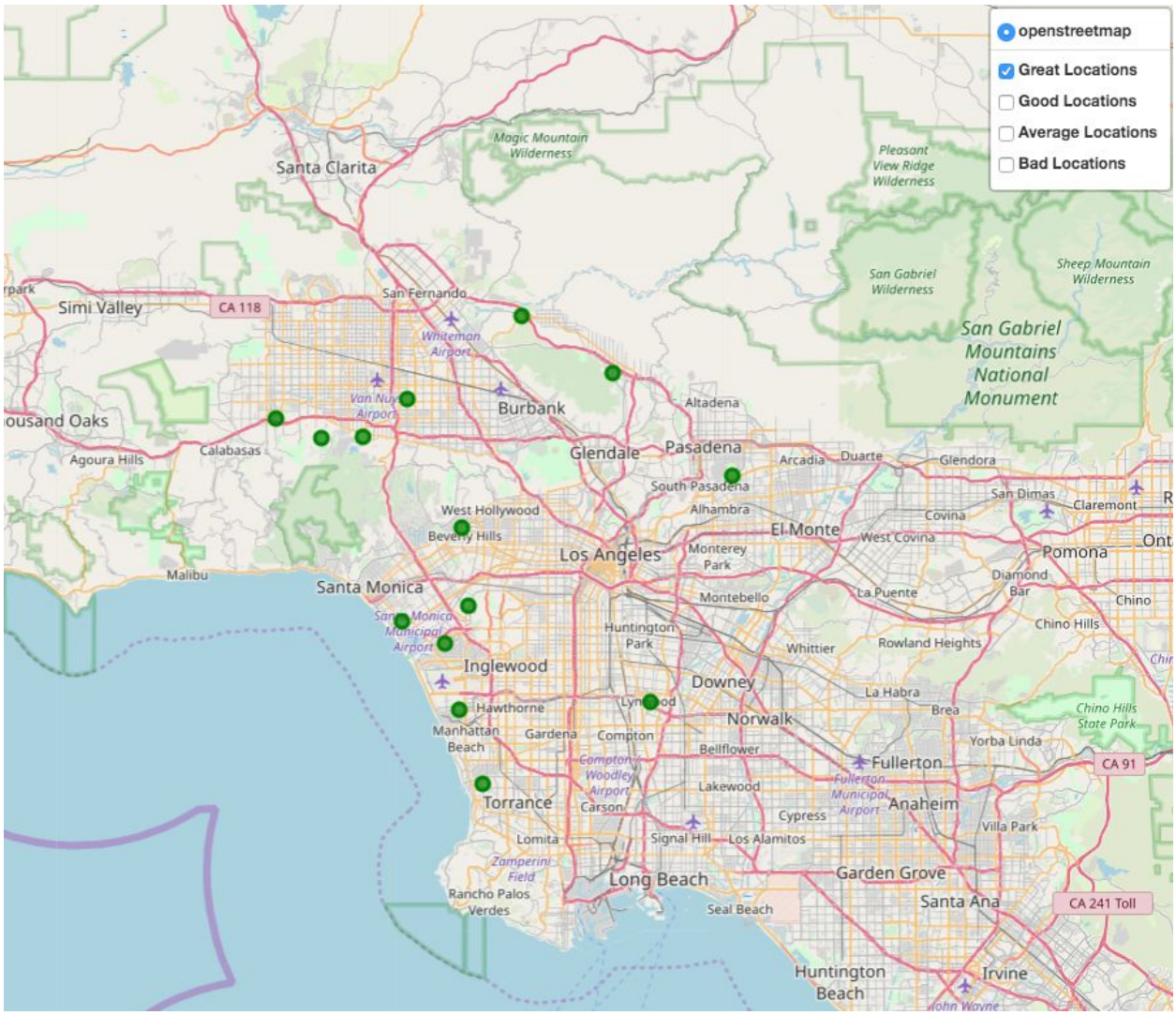


Figure 4. A folium map displaying all of the good locations

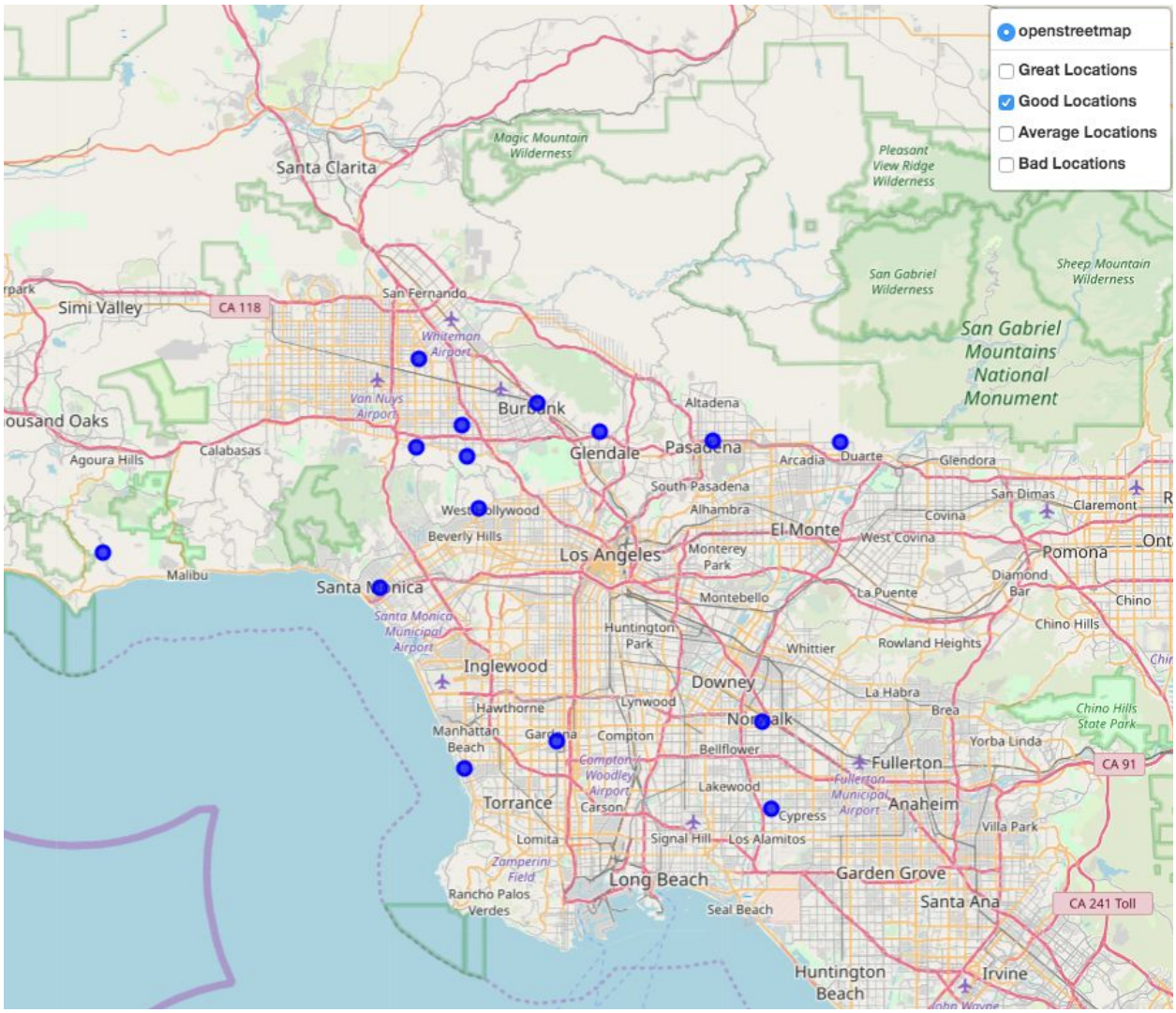


Figure 5. A folium map displaying all of the average locations

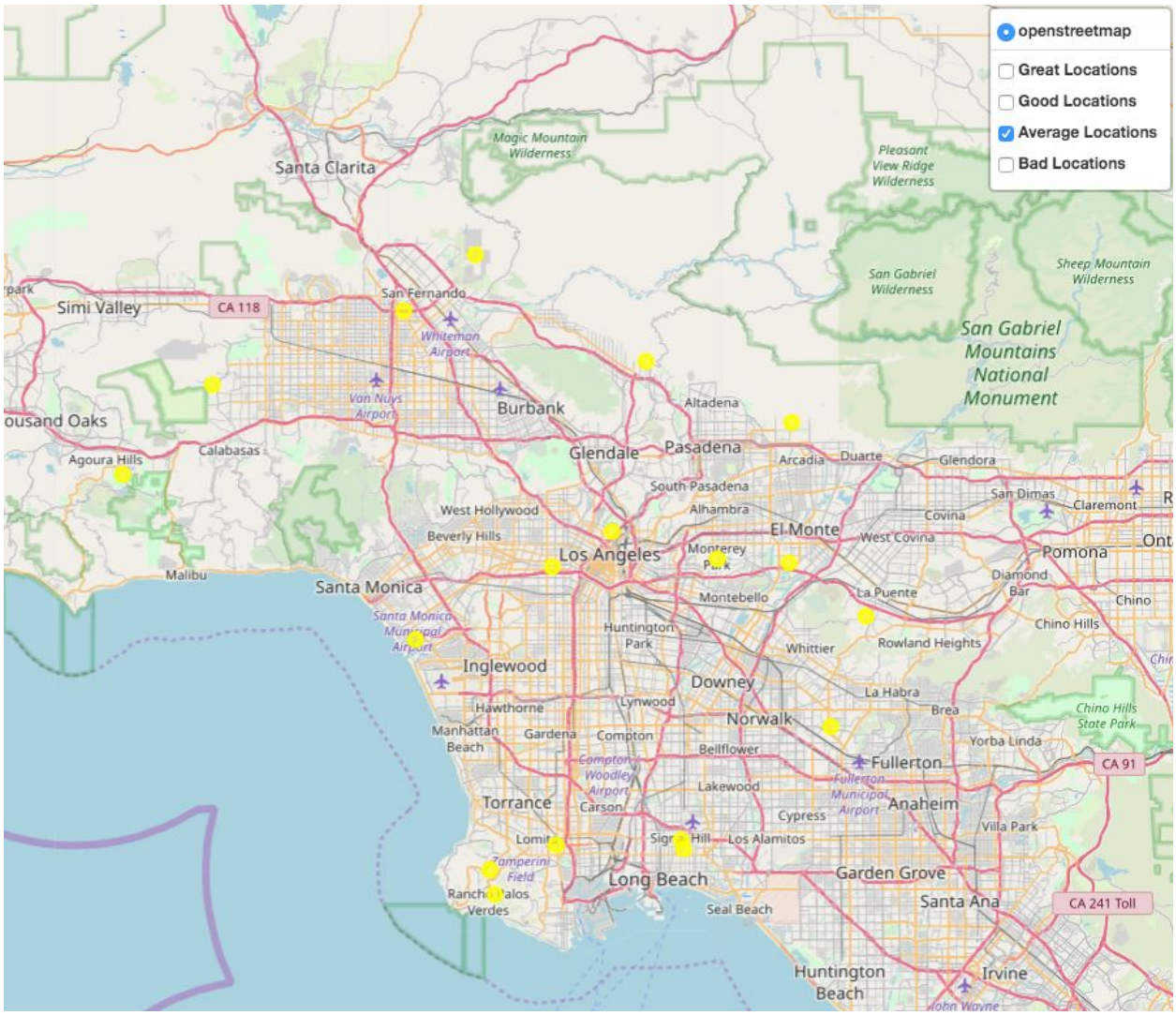
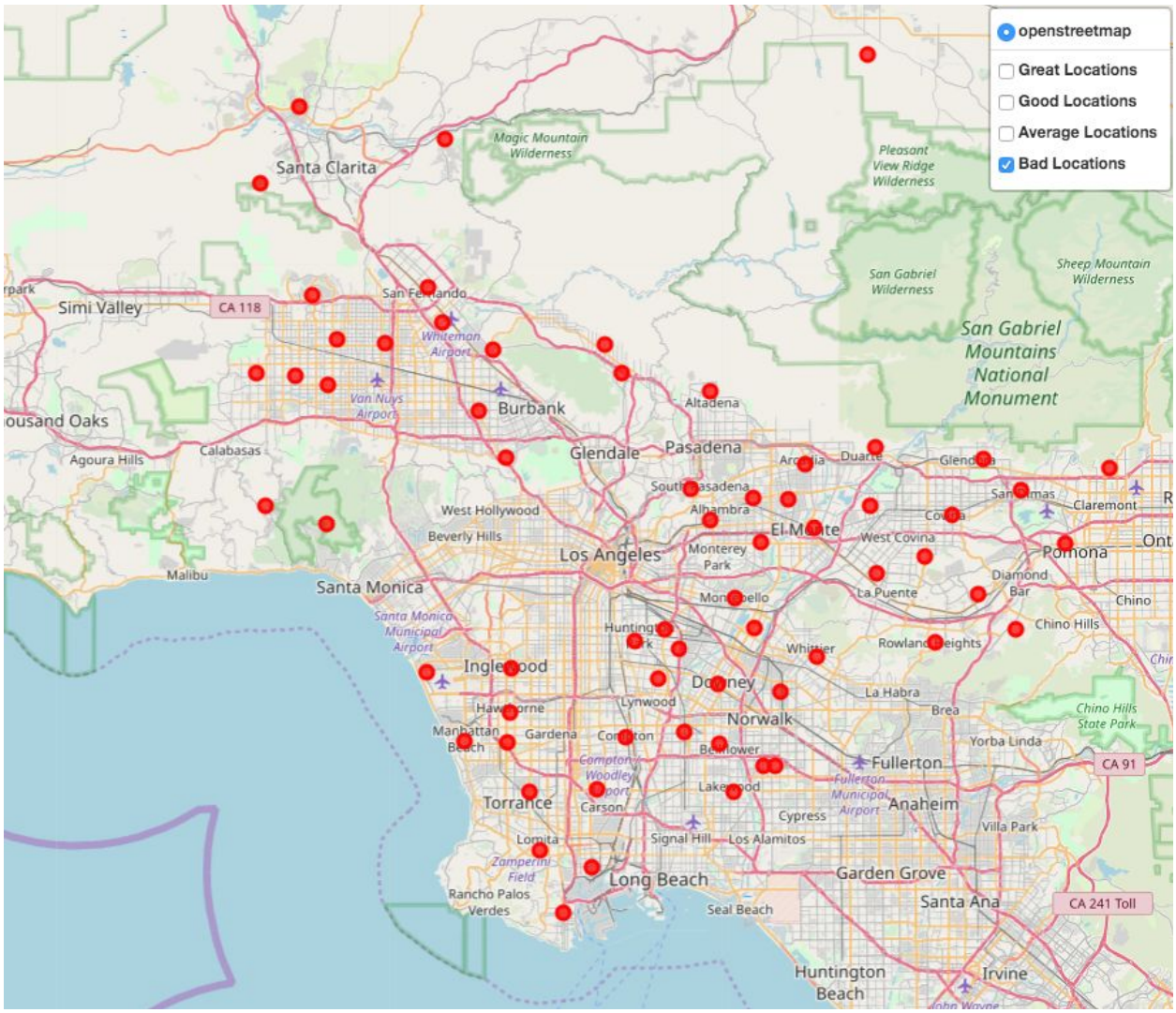


Figure 6. A folium map displaying all of the bad locations

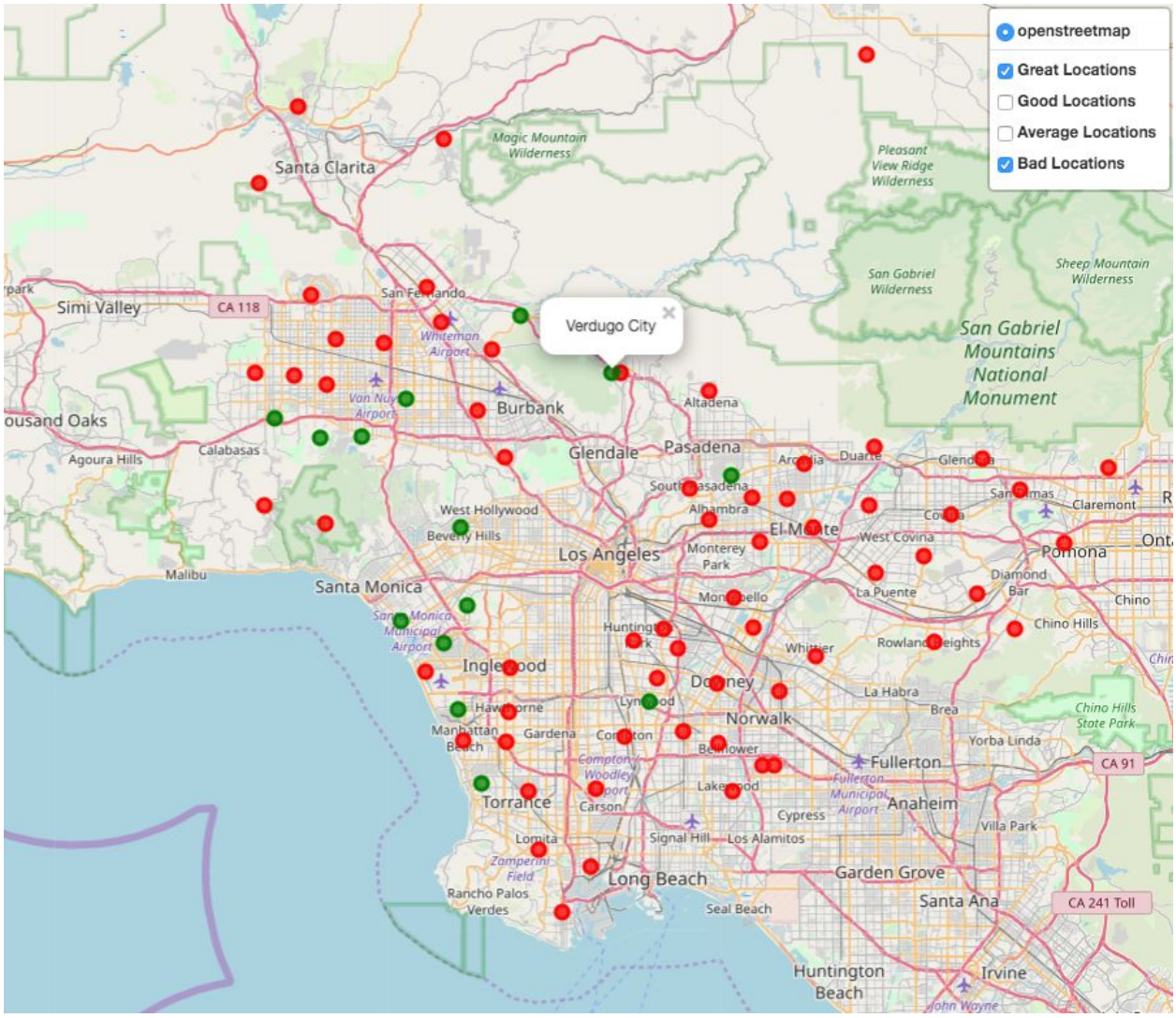


5. Discussion

5.1 Observatory Analysis

In order to see how accurate my results were, I plotted only the “Great Locations” and “Bad Locations” (Figure 7). Out of all the cities, there is only an overlap of great and bad markers on Verdugo City. This is a “Great Location” marker that should be disregarded as it is contradictory for it to have an overlapping “Bad Location”. Therefore, this faulty marker reduces the accuracy by 7 percent, giving the overall accuracy of great locations to open up a fitness store at 93%.

Figure 7. A folium map displaying only the great and bad locations



5.2 Recommendations

An observatory analysis of the map containing all of the great locations (Figure 3) reveals that any of the green cities along the coast or near a large public park are the best locations to open a fitness store. These cities include, but are not limited to, Venice, El Segundo, Redondo Beach, Playa Vista, and Woodland Hills.

6. Conclusion

This study's purpose was to locate the best cities in Los Angeles County to open up a fitness store based on the prevalence of gyms and parks. With the employment of the

Foursquare API, Folium maps, and classification techniques, I was able to produce an interactive map that shows not only the best locations to open up this fitness store, but also locations to avoid. A link to this map and the code used to create it is provided below.

https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/d955863f-01d9-4766-b8dc-ae2a5bbcc023/view?access_token=-