International Home Searching Using Web Scraping and Foursquare

Alejandro Aylwin

May 23, 2020

Introduction

Buying a house is among the biggest and most important decisions people make during their life.

In an ideal situation, a future home owner will visit all potential properties multiple times and evaluate the benefits and shortcomings of each one before making a decision.

But how can this be done from thousands of miles away?

The following project will attempt to provide a solution for people living far away from their future homes that do not have the chance to evaluate their options in person.

We will consider our first case consisting of a young couple from South America moving to Santa Barbara, California for business.

Data

The first set of data used will be the characteristics of the available homes.

- Price
- Address
- Type (Condo or Townhouse)
- Number of bedrooms
- Number of bathrooms
- Area (Square feet)

The second set of data will describe the main venues in proximity to the available homes that the future owners can benefit from.

This second set of data will be obtained using the Foursquare API to retrieve the necessary information from the database.

Methodology

1. Conduct Interview with Clients

The first stage in our methodology consists of interviewing the clients to determine their budget, requirements and priorities. From the interview we gather the following information:

Budet: Maximum $1.000.000

Requirements: The couple are pet owners so their main requirement is to have either a park or a dog run within the top venues closeby.

The order of priorities after their initial requirement are: number of bedrooms, number of bathrooms and then price.

2. Web Scraping

We can now scrape the real estate website zillow.com for properties within the budget. To do this we use ParseHub .We create and save a csv file with all the information we need. The csv file contains the following table with 34 available homes.

Type	Price $	Address	Latitude	Longitude	Bedrooms	Bathrooms	Area sqr_ft
Townhouse for sale	995,000	282 N La Cumbre Rd, Santa Barbara, CA 93110	34.4439	-119.75038	3	2	1,472
Condo for sale	990,000	105 W De La Guerra St UNIT D, Santa Barbara, CA 93101	34.41782	-119.70016	1	2	1,116
Townhouse for sale	959,000	262 Calle Esperanza, Santa Barbara, CA 93105	34.43687	-119.74597	3	3	1,646
Townhouse for sale	949,000	105 W De La Guerra St UNIT E, Santa Barbara, CA 93101	34.41782	-119.70016	1	2	1,066

Extract From the File Created

3. Pandas Dataframe & Mapping Available Homes

We then open the csv file in our Jupyter Notebook using Pandas and map the available homes using Folium.


Goleta
Isla Vista
Santa Barbara
Montecito
Summerland
CA 192
CA 144
CA 225
Campus Point State Marine Conservation Area

4. Obtaining & Formatting Second Dataset

We use our Foursquare credentials and version to request the nearby venues for every address on the available homes dataframe.

From our request we make sure to gather only data that will be useful to us which include the names of the names, locations (latitude and longitude) and categories of the venues.

Once we have all the requested information we now must format the data. To do this we use one hot encoding and group by Address.

After this we sort the rows to show the top 5 nearby venues for each address.

Address	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
282 N La Cumbre Rd, Santa Barbara, CA 93110	Fast Food Restaurant	Hotel	Bakery	Pizza Place	Coffee Shop
105 W De La Guerra St UNIT D, Santa Barbara, CA 93101	Sushi Restaurant	Pizza Place	Bar	Ice Cream Shop	Burger Joint
262 Calle Esperanza, Santa Barbara, CA 93105	Pizza Place	Bank	Accessories Store	Mobile Phone Shop	Shipping Store
105 W De La Guerra St UNIT E, Santa Barbara, CA 93101	Sushi Restaurant	Pizza Place	Bar	Ice Cream Shop	Burger Joint

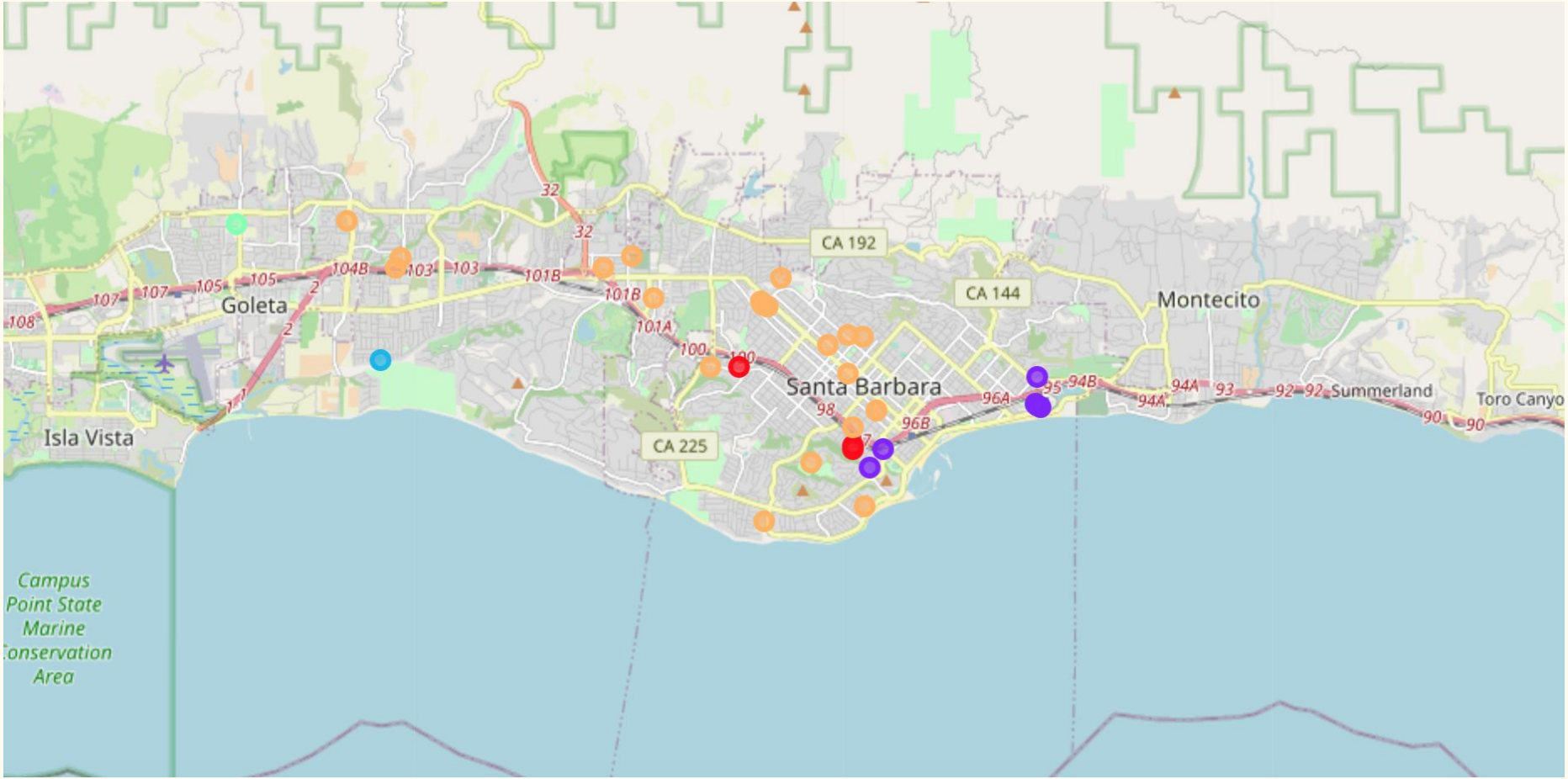
Extract From Second Dataset

5. Clustering & Labeling

We cluster and label our second dataset based on the nearby venue characteristics.

We use clustering because the highest priority of this case is to deliver a specific venue category requirement.

After we have assigned and labeled the addresses to their clusters we can now merge the first and the second datasets by address and map the available homes according to their clusters.


Goleta
Isla Vista
Santa Barbara
Montecito
Summerland
Toro Canyo
CA 192
CA 144
CA 225
Campus
Point State
Marine
Conservation
Area

Results

First Cluster (label 0)

- 621 W Ortega St APT A, Santa Barbara, CA 93101
- 633 W Ortega St, Santa Barbara, CA 93101

Second Cluster (label 1)

- 282 N La Cumbre Rd, Santa Barbara, CA 93110
- 105 W De La Guerra St UNIT D, Santa Barbara, CA 93101
- 262 Calle Esperanza, Santa Barbara, CA 93105
- 105 W De La Guerra St UNIT E, Santa Barbara, CA 93101
- 4004 Via Lucero APT 13, Santa Barbara, CA 93110
- 2605 Hacienda Ct, Santa Barbara, CA 93105
- 18 W Victoria St APT 107, Santa Barbara, CA 93101
- 5463 Tree Farm Ln, Santa Barbara, CA 93111
- 1426 Laguna St #A, Santa Barbara, CA 93101
- 211 Reef Ct, Santa Barbara, CA 93109
- 66 Barranca Ave APT 4, Santa Barbara, CA 93109
- 323 Ladera St #2, Santa Barbara, CA 93101
- 25 Ocean View Ave #C5, Santa Barbara, CA 93103
- 2627 State St APT 3, Santa Barbara, CA 93105
- 415 W Gutierrez St APT 6, Santa Barbara, CA 93101
- 316 Por La Mar Cir, Santa Barbara, CA 93103
- 2525 State St APT 25, Santa Barbara, CA 93105
- 2525 State St APT 11, Santa Barbara, CA 93105
- 1600 Garden St APT 35, Santa Barbara, CA 93101
- 1701 Anacapa St UNIT 3, Santa Barbara, CA 93101
- 2525 State St APT 31, Santa Barbara, CA 93105
- 432 Por La Mar Cir, Santa Barbara, CA 93103
- 5095 Rhoads Ave APT C, Santa Barbara, CA 93111
- 125 Por La Mar Cir, Santa Barbara, CA 93103
- 30 W Constance Ave UNIT 2, Santa Barbara, CA 93105
- 425 Transfer Ave APT A, Santa Barbara, CA 93101
- 2727 Miradero Dr APT 306, Santa Barbara, CA 93105

Third Cluster (label 2)

- 2330 Vista Madera, Santa Barbara, CA 93101

Fourth Cluster (label 3)

- 969 Miramonte Dr APT 5, Santa Barbara, CA 93109
- 27 N San Marcos Rd #B, Santa Barbara, CA 93111
- 5034 Birchwood Rd, Santa Barbara, CA 93111

Fifth Cluster (label 4)

- 5305 Traci Dr, Santa Barbara, CA 93111

Cluster Descriptions

The first and fourth clusters (labels 0 and 3) include venues consisting of outdoor activities.

The second cluster (label 1) groups addresses surrounded mainly by hotel and gastronomical services.

The third and fifth cluster (labels 2 and 4) group seemingly unrelated venue categories.

Candidates

The only properties with “Parks” and “Dog Runs” in their top nearby venues are:

- Option 1: 621 W Ortega St APT A, Santa Barbara, CA 93101 (Park as 1st Most Common Venue) from the first cluster.
- Option 2: 633 W Ortega St, Santa Barbara, CA 93101 (Park as 3rd Most Common Venue) from the first cluster.
- Option 3: 27 N San Marcos Rd #B, Santa Barbara, CA 93111 (Dog Run as 3rd Most Common Venue) from the fourth cluster.
- Option 4: 5034 Birchwood Rd, Santa Barbara, CA 93111 (Park as 1st Most Common Venue and Dog Run as 3rd Most Common Venue) from the fourth cluster.

Recommendation

Selection criteria Using Order of Priorities.

Option	Bedrooms	Bathrooms	Price
Option 1 ⇨⇨⇨	2 ⇨⇨⇨	3 ⇨⇨⇨	$780,000
Option 2 ⇨⇨⇨	2 ⇨⇨⇨	2	-
Option 3 ⇨⇨⇨	2 ⇨⇨⇨	1	-
Option 4 ⇨⇨⇨	2 ⇨⇨⇨	3 ⇨⇨⇨	$599,000

The best option is Option 4

- Address: 5034 Birchwood Rd, Santa Barbara, CA 93111
- Type: Townhouse
- Price: $599,000
- Bedrooms: 2
- Bathrooms: 3
- Area: 1,104 square feet
- 1st Most Common Venue: Park
- 2nd Most Common Venue: Pool
- 3rd Most Common Venue: Dog Run
- 4th Most Common Venue: Zoo Exhibit
- 5th Most Common Venue: Exhibit

Conclusion

In conclusion, we see how basic data can be transformed into useful and valuable information by following a structured methodology.

These series of steps are scalable, applicable and improvable to fit different requirements and cases.

This example as shown can help at least in some proportion to guide the decision making process.