

Determining the Optimal Neighborhood for a Bookstore in Santa Barbara, CA Using K-Means Clustering

Aymiee Lee

April 18, 2021

1. Introduction

1.1 Description and Discussion

The location of a bookstore will determine its success. In a wrong location, one will not attract the amount of customers needed to stay in business. A location near a busy restaurant, movie theatre or any place that will increase traffic in the vicinity of your store is ideal. Good location decisions can significantly boost a company's long-term performance. Since bookstores typically have low margins, other ancillary sales such as stationary, tee shirts, gifts, coffee, snacks are important to grow sales.

I've lived in Santa Barbara for over sixteen years. One of my favorite past times was taking my daughter to the Border's book store downtown and spending a couple of hours browsing for books, watching my daughter partake in children events, perusing gifts, and enjoying a drink and snack or meet up with friends for coffee. Border closed its door back in 2011 and since then, we've frequented other book stores, but not with the same intensity. We don't linger for hours like we used to ...however, we still buy a book.

Santa Barbara is a small city 90 miles north of Los Angeles. It's roughly 42 square miles compared to 468.7 sq. miles of Los Angeles. Santa Barbara has roughly 30 neighborhoods. I put together a list of the neighborhood in Santa Barbara and used Google maps manually to capture the coordinates. Some may or may not have their own unique list of venues. I will use the K-Means Clustering algorithm to cluster the neighborhoods into similar groups. Next I will determine the ideal neighborhood by comparing the venues in each neighborhood. Our ideal location will be next to busy venues to drive traffic to the book store.

The objective of this capstone project is to locate the optimal neighborhood for an independent book store using Foursquare API.

1.2 Benefits

Santa Barbara, often regarded as the American Riviera, is home to a beautiful landscape and great climate for living, playing and working. Santa Barbara is well known for its strong sense of community, agricultural land, myriad of wineries, and attractive cultural and tourism opportunities. Lots of businesses open and sadly close (abruptly) in Santa Barbara. I hope that this analysis shows a new business owner where to look for real estate opportunities, saving time and allowing one to open new or more units faster.

2. Data Description

2.1 Data Sources

Data for this project will come from 2 sources:

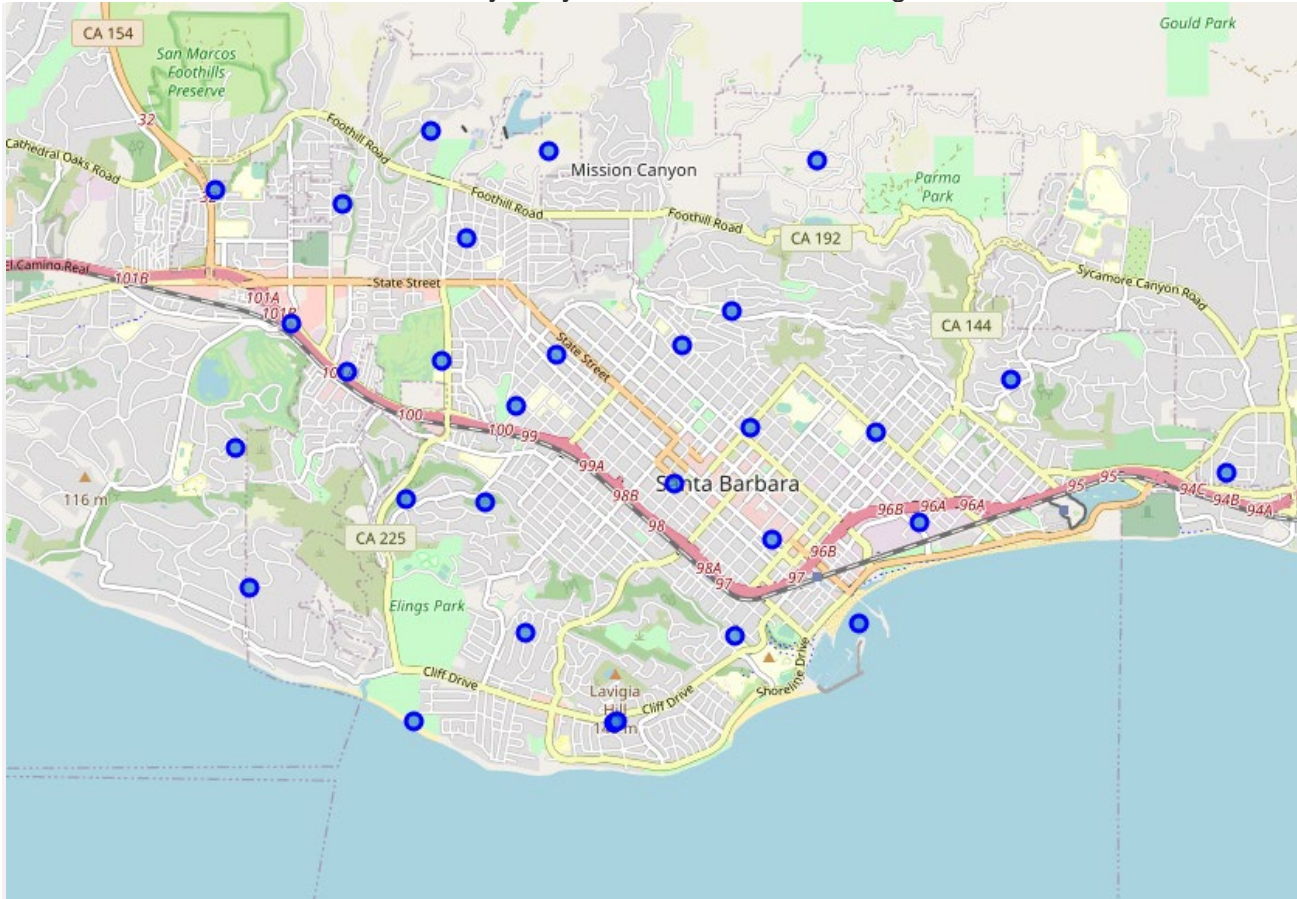
- Google Maps — The list of neighborhoods with their latitudes and longitudes were taken from the URL in Google maps. I saved them into a csv file and uploaded the file to GitHub. This data will help visualize the neighborhoods in Folium (a Python library used for visualizing geospatial data). The coordinates will be passed to the Foursquare API to gather venue data for each neighborhood
- Foursquare API will be used to obtain venue data for each neighborhood. Once again, finding a location with busy venues to drive traffic to the book store will be ideal.

2.2 Neighborhood Candidate

There was not a one stop site to obtain all of the neighborhoods of Santa Barbara and their coordinates. So, I did the next best thing which was entering the neighborhoods into Google Maps and manually storing the coordinates in a csv file on GitHub. The file was read into a python dataframe: Neighborhoods, Latitude and Longitude. There were thirty neighbors that I identified from my knowledge of living in Santa Barbara and from Google Maps. Not all 30 neighborhoods will have venues.

	Neighborhood	Latitude	Longitude
0	Alta Mesa	34.408464	-119.725023
1	Bel Air	34.420665	-119.738266
2	Campanil	34.412598	-119.755684
3	Cielito	34.451463	-119.692851
4	Coast Village	34.423107	-119.647475
5	Downtown	34.422050	-119.708637
6	East Beach	34.418564	-119.681380
7	East Mesa	34.400292	-119.715212
8	East San Roque	34.444448	-119.731579
9	East side	34.400331	-119.714883
10	Eucalyptus Hill	34.431490	-119.671408
11	Foothill	34.454180	-119.735598
12	Hidden Valley	34.425332	-119.757201
13	Hitchcock	34.432211	-119.744874
14	Hope	34.448782	-119.759308
15	Laguna	34.427191	-119.700211
16	Lower Riviera	34.434713	-119.707776
17	Lower State	34.417027	-119.697798
18	Milpas	34.426705	-119.686192
19	Mission Canyon	34.452332	-119.722557
20	Oak Park	34.429134	-119.726149
21	Riviera	34.437763	-119.702240
22	Samarkand	34.433200	-119.734321
23	San Roque	34.447537	-119.745265
24	Upper East	34.433816	-119.721574
25	Upper State	34.436661	-119.750920
26	Waterfront	34.409293	-119.688224
27	West Beach	34.408144	-119.701799
28	West Mesa	34.400390	-119.737399
29	Westside	34.420422	-119.729518

I used folium, data visualization library in Python, to visualize the neighborhoods in Santa Barbara.



2.3 Data Cleansing

The Foursquare API was used to obtain venues for each neighborhoods. Since Santa Barbara is a small area, venues may overlap. I limited the venues to 50 within a 500-meter radius from each neighborhood. I also removed all of the neighborhoods that had less than 10 venues as my goal is to choose neighborhood with the most venues to drive traffic to the bookstore.

Figure 1 - 30 Neighborhoods in Santa Barbara

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Alta Mesa	32	32	32	32	32	32
Bel Air	6	6	6	6	6	6
Campanil	4	4	4	4	4	4
Cielito	4	4	4	4	4	4
Coast Village	50	50	50	50	50	50
Downtown	50	50	50	50	50	50
East Beach	50	50	50	50	50	50
East Mesa	39	39	39	39	39	39
East San Roque	50	50	50	50	50	50
East side	41	41	41	41	41	41
Eucalyptus Hill	4	4	4	4	4	4
Foothill	6	6	6	6	6	6
Hidden Valley	3	3	3	3	3	3
Hitchcock	49	49	49	49	49	49
Hope	4	4	4	4	4	4
Laguna	50	50	50	50	50	50
Lower Riviera	23	23	23	23	23	23
Lower State	50	50	50	50	50	50
Milpas	45	45	45	45	45	45
Mission Canyon	3	3	3	3	3	3
Oak Park	33	33	33	33	33	33
Riviera	16	16	16	16	16	16
Samarkand	50	50	50	50	50	50
San Roque	48	48	48	48	48	48
Upper East	50	50	50	50	50	50
Upper State	50	50	50	50	50	50
Waterfront	50	50	50	50	50	50
West Beach	50	50	50	50	50	50
West Mesa	8	8	8	8	8	8
Westside	8	8	8	8	8	8

Figure 2- Neighborhoods filtered by Venue >= 20

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Alta Mesa	32	32	32	32	32	32
Coast Village	50	50	50	50	50	50
Downtown	50	50	50	50	50	50
East Beach	50	50	50	50	50	50
East Mesa	39	39	39	39	39	39
East San Roque	50	50	50	50	50	50
East side	41	41	41	41	41	41
Hitchcock	49	49	49	49	49	49
Laguna	50	50	50	50	50	50
Lower Riviera	23	23	23	23	23	23
Lower State	50	50	50	50	50	50
Milpas	45	45	45	45	45	45
Oak Park	33	33	33	33	33	33
Riviera	16	16	16	16	16	16
Samarkand	50	50	50	50	50	50
San Roque	48	48	48	48	48	48
Upper East	50	50	50	50	50	50
Upper State	50	50	50	50	50	50
Waterfront	50	50	50	50	50	50
West Beach	50	50	50	50	50	50

The data was sorted by mean frequency to show the 10 most popular venues.

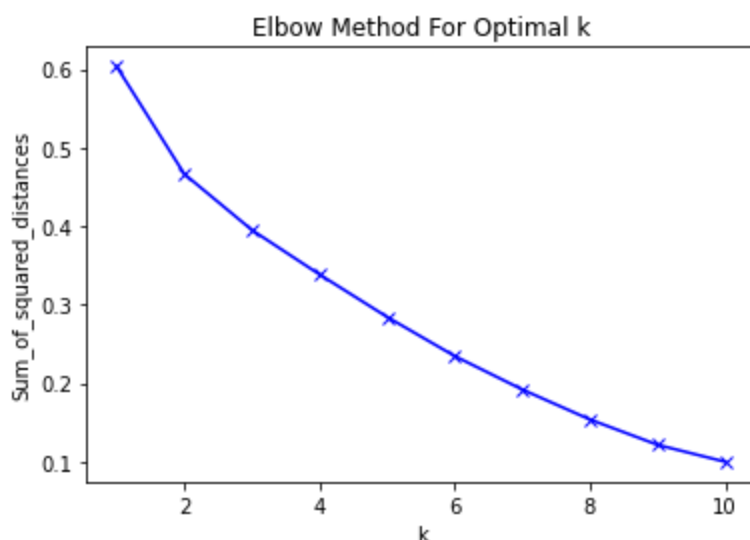
Figure 3- Sample to show sorted by mean frequency

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
Alta Mesa	Grocery Store	Pharmacy	Pizza Place	Fast Food Restaurant
Coast Village	Mexican Restaurant	Hotel	Italian Restaurant	French Restaurant
Downtown	American Restaurant	Bakery	Mexican Restaurant	Italian Restaurant
East Beach	Brewery	Winery	Mexican Restaurant	Grocery Store
East Mesa	Grocery Store	Beach	Park	Pizza Place

3. Methodology and Analysis

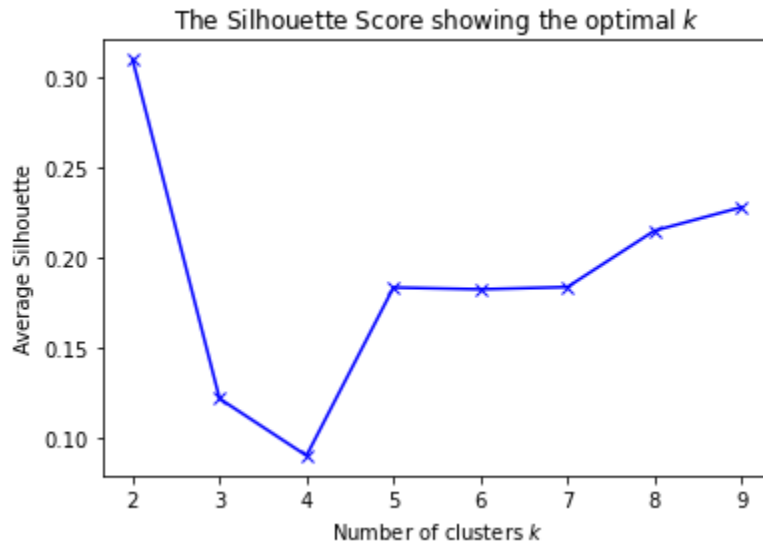
The K-Means clustering was applied to the venue dataframe (above)to put each neighborhood into similar clusters. This methodology requires one to initialize “K” — the number of clusters. Determining the number of clusters is user-defined and it can be difficult. I will use the Elbow Method to determine K — number of cluster. The method consists of plotting the variation as a function of the number of clusters. One would pick the elbow of the curve as the number of clusters to use. In the data below, there was not too much of an elbow.

Figure 4- Elbow Method



Since the Elbow Method was indeterminate, I used the Silhouette Score to determine the number of clusters. It is used to study the distances between resulting clusters. The score measure how close each point in one cluster is to points in the neighboring clusters.

Figure 5- Silhouette Score

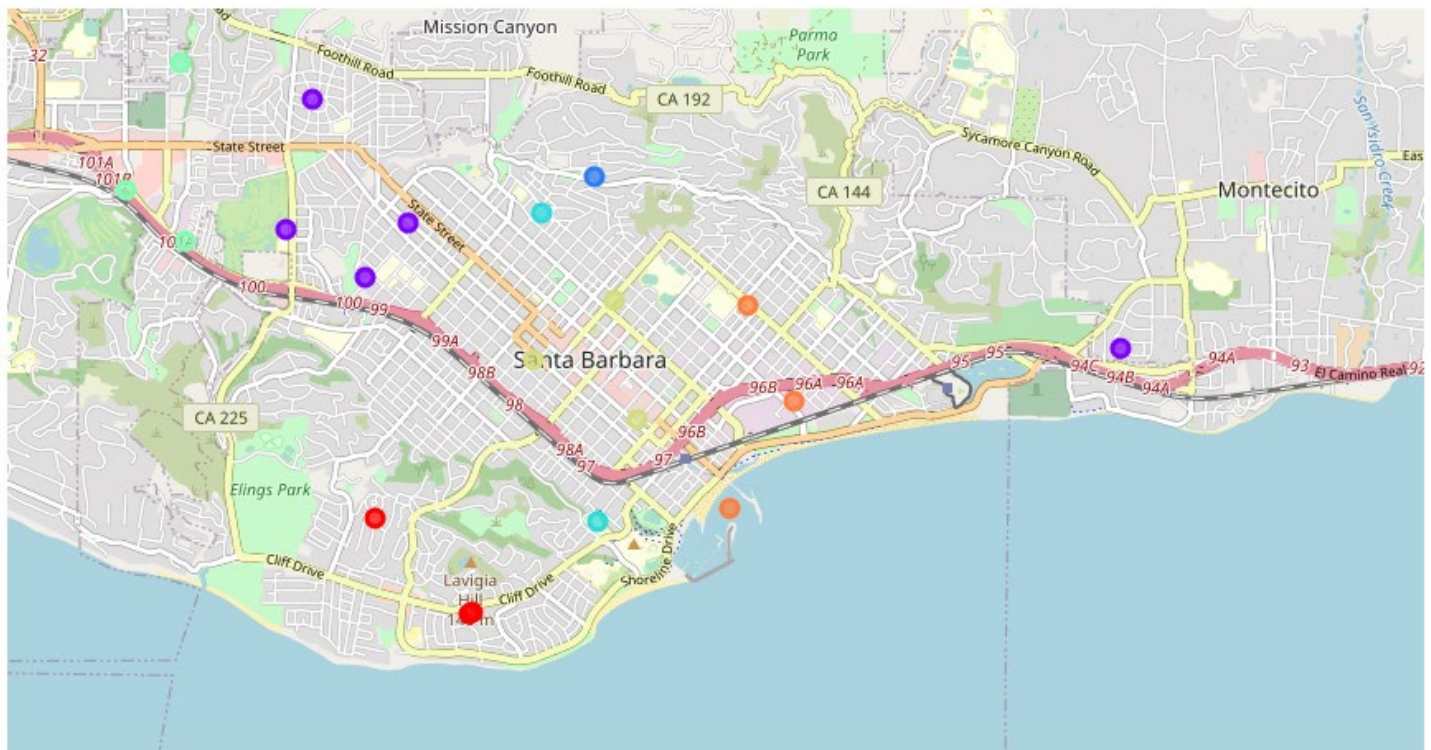


The appropriate number of clusters of neighborhoods = 7. Seven has a relatively high silhouette score and a relatively low inertia from the elbow method. After applying the K-Means Cluster algorithm, the labels were added back into the dataframe:

Figure 6 - Adding Cluster Labels

Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Alta Mesa	34.408464	-119.725023	0	Grocery Store	Pharmacy	Pizza Place
Coast Village	34.423107	-119.647475	1	Mexican Restaurant	Hotel	Italian Restaurant
Downtown	34.422050	-119.708637	5	American Restaurant	Bakery	Mexican Restaurant
East Beach	34.418564	-119.681380	6	Brewery	Winery	Mexican Restaurant
East Mesa	34.400292	-119.715212	0	Grocery Store	Beach	Park

The seven are depicted below using Folium.



Out of the 7 clusters, I identified 2 of the clusters that could bring enough foot traffic to sustain a book store.

- Purple Cluster – more of “Locals” neighborhood.

Neighborhood

Coast Village

East San Roque

Oak Park

Samarkand

Upper East

- Light Green Cluster - had venues such as Theaters and Farmer's Market. You're likely to have a lot of people interested in visiting a bookstore if they are coming from/to Theater and Farmer's market.

Neighborhood

Downtown

Laguna

Lower State

4. Results and Conclusion

According to SantaBarbara.com, The Santa Barbara South Coast welcomed an estimated 7.2 million total visitors between 2016 and 2017. The most common activities were dining, going to the beach, shopping and wine tasting. All of these activities are in walking distance of Light Green Cluster neighborhood.

In the neighbor of Light Green Cluster: Downtown, Laguna and Lower State, the choice would be between Lower State and Downtown. What Foursquare API doesn't show is another book store in Downtown called Book Den. We do not want to open another book store in the same neighborhood, so Lower State is the better option as it's also the most touristy (compared to Laguna).

According to Google Quick Facts: "The leafy streets of Lower State are lined with trendy eateries serving New American, Mexican, and pan-Asian fare, plus craft breweries and winery tasting rooms. National fashion chains cluster in the courtyards of the Paseo Nuevo shopping center, and buskers play amid stalls selling organic produce and flowers at the weekly Santa Barbara Farmers Market. The kid-friendly MOXI museum has interactive science exhibits."

Disclaimer: this was just an exercise in using K-Means Clustering and accessing Foursquare API data. There's other criteria that is also important such as rent. Rent downtown and on Lower State street is very high. You would get more traffic, but you'd be paying a lot in rent in Cluster 5 as oppose to Cluster 1. Which brings be back to the original statement that bookstores typically have low margins. Other ancillary sales such as stationary, tee shirts, gifts, coffee, snacks are important to grow sales.