The problem

A small, family-run pizza restaurant is looking to relocate to California and needs to find the best possible location. The owners don't know much about the region and don't have much money to spend to identify the ideal location.

They do, however, have a small list of requirements that have been derived from their long experience in the business. Their ideal location is a city with at least 200k residents, that already has a "restaurant hotspot" where locals go to when they are looking to eat out.

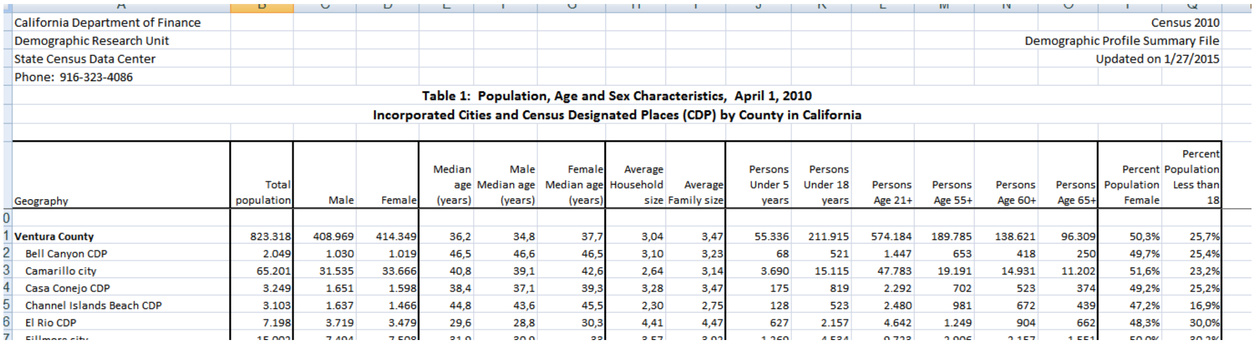
This list of requirements, when applied to a state the size of California, leaves plenty of options to choose from, and data analysis will be needed to filter potential locations and find the ones with the best chances of success. Given budget limitations, all data will need to come from free sources.

The data

Data for this project will come from four sources: the California census, Google Trends, Open Maps and Foursquare in this order.

California census

The census will be used to create an initial list of all cities in California that have at least 200 thousand residents. This data can be freely downloaded from the census page in an xlsx format. Unfortunately, the format is optimized for viewing on Excel and will require significant cleaning to eliminate unnecessary spaces, totals and other unwanted information.



California Department of Finance
Demographic Research Unit
State Census Data Center
Phone: 916-323-4086

Census 2010
Demographic Profile Summary File
Updated on 1/27/2015

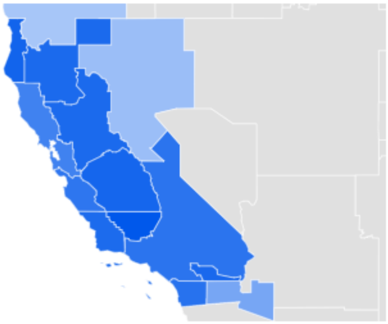
Table 1: Population, Age and Sex Characteristics, April 1, 2010
Incorporated Cities and Census Designated Places (CDP) by County in California

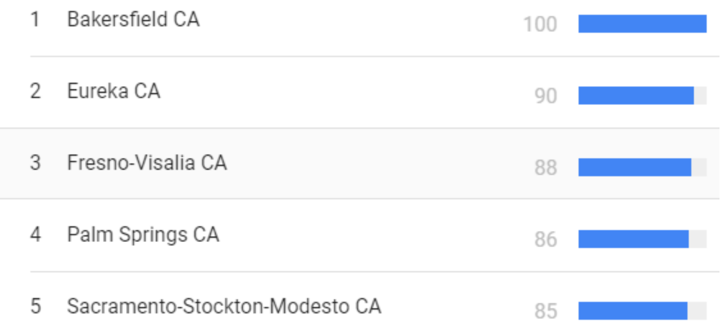
	Geography	Total population	Male	Female	Median age (years)	Male Median age (years)	Female Median age (years)	Average Household size	Average Family size	Persons Under 5 years	Persons Under 18 years	Persons Age 21+	Persons Age 55+	Persons Age 60+	Persons Age 65+	Percent Population Female	Percent Population Less than 18
0																	
1	**Ventura County**	823.318	408.969	414.349	36,2	34,8	37,7	3,04	3,47	55.336	211.915	574.184	189.785	138.621	96.309	50,3%	25,7%
2	Bell Canyon CDP	2.049	1.030	1.019	46,5	46,6	46,5	3,10	3,23	68	521	1.447	653	418	250	49,7%	25,4%
3	Camarillo city	65.201	31.535	33.666	40,8	39,1	42,6	2,64	3,14	3.690	15.115	47.783	19.191	14.931	11.202	51,6%	23,2%
4	Casa Conejo CDP	3.249	1.651	1.598	38,4	37,1	39,3	3,28	3,47	175	819	2.292	702	523	374	49,2%	25,2%
5	Channel Islands Beach CDP	3.103	1.637	1.466	44,8	43,6	45,5	2,30	2,75	128	523	2.480	981	672	439	47,2%	16,9%
6	El Rio CDP	7.198	3.719	3.479	29,6	28,8	30,3	4,41	4,47	627	2.157	4.642	1.249	904	662	48,3%	30,0%
7	Fillmore city	15.002	7.494	7.508	31,0	30,0	33	3,57	3,92	1.260	4.534	9.723	2.006	2.157	1.551	50,0%	30,2%

Search data

The filtering of census data will provide a good list of potential cities. To start filtering them it will be important to understand how much interest/demand for a pizza place there is in each. Under normal circumstances, this data could be obtained using surveys and commercially available data. In this case, however, we are limited to freely available sources and have selected Google Trends as the best source of data.

The challenge is that the page uses a tree-structure that makes it impossible to view detailed data for more than one geographical area at a time. For this reason, it is possible to compare the intensity of searches between metropolitan areas of the same state, and between cities of the same metropolitan area, but it is not possible to directly compare cities located in different metropolitan areas.

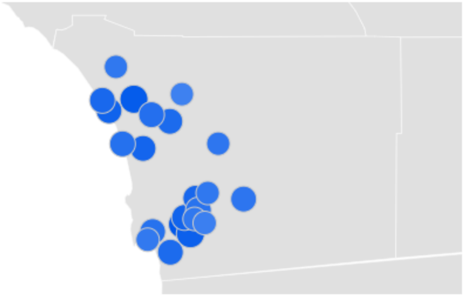


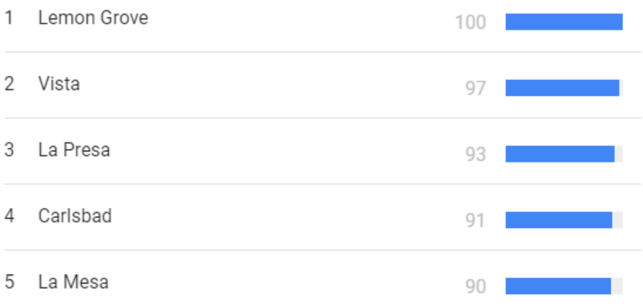


1	Bakersfield CA	100
2	Eureka CA	90
3	Fresno-Visalia CA	88
4	Palm Springs CA	86
5	Sacramento-Stockton-Modesto CA	85

The page does offer a city-level view but this view only shows a limited number of cities. In the case of California, a state-level search would only return 50.

To get around this problem it was decided to first obtain the score of each metropolitan area of California and then those for every city within each metropolitan area. The scores of each city will then be weighted using the score of their metropolitan area to get data that is directly comparable. This method delivered data for 370 cities.





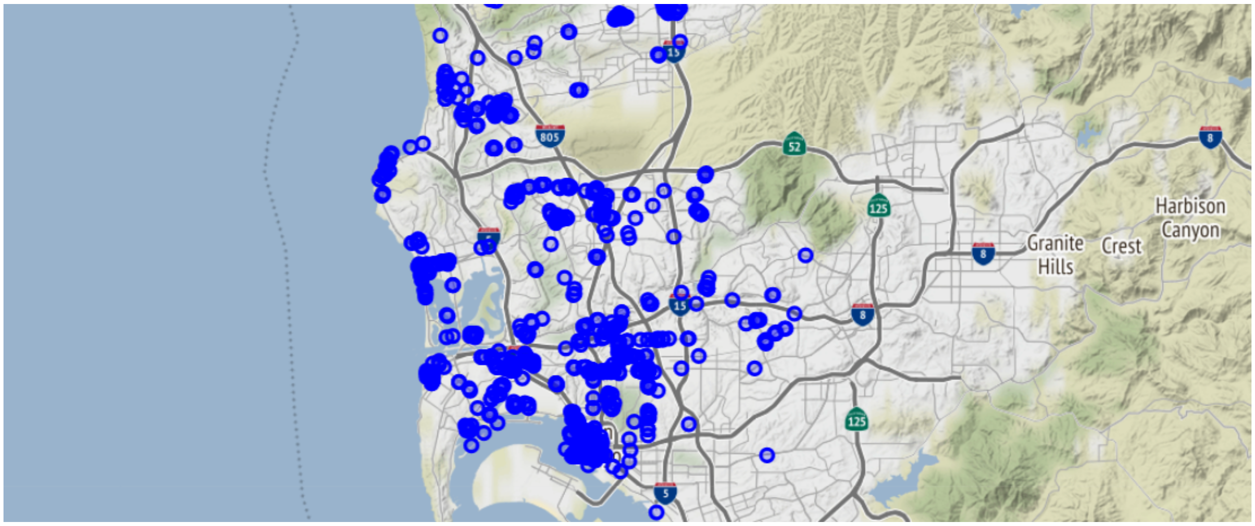
1	Lemon Grove	100
2	Vista	97
3	La Presa	93
4	Carlsbad	91
5	La Mesa	90

NOTE: Google also displays the variation of data over time but in this particular case the search volume was stable over time and the average value for the past year was used.

Open street map data

Open Street Map data was selected as the best source to identify the clusters of restaurants that were described in the problem section. This data source was chosen because it allows searches over large regions without limitations in the number of results that are returned.

The data obtained is not as detailed as the one provided by other sources but it will be sufficient to identify clusters. To query the database we'll use the OverPass API.



NOTE: the Foursquare API was also considered for this step but its limit of 50 results per search made it unviable for this stage.

Foursquare data

Once the most promising clusters have been selected we'll use Foursquare data to better understand what the competition and environment are like in each potential cluster. The data is available through a partially-free API that can return a maximum of 50 results per search.

This limitation prevented the usage of this service for the previous levels of data gathering presents less of a problem in this stage. Nevertheless, the data will need to be obtained in stages to ensure searches don't hit the limit of results and return a complete dataset.

For this reason, the first set of queries to the API will focus only on pizza places in a radius of 1.5 km from the center of each cluster. This will allow initial filtering of areas that are already oversaturated with pizza-options.

The second set of queries will focus on obtaining a list of all eating options in each cluster. In this case, since a single query would likely hit the limit and return an incomplete dataset it will be necessary to perform a series of smaller-range queries using a pattern of coordinates that will be calculated with a custom function. The sum of all the smaller queries should return the results of an area of at least 1 km around the original cluster center.

The image on the right shows the pattern of small searches around the original cluster center. Each search had a 500 m radius and its center was offset 500 m from the original one.