# 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.

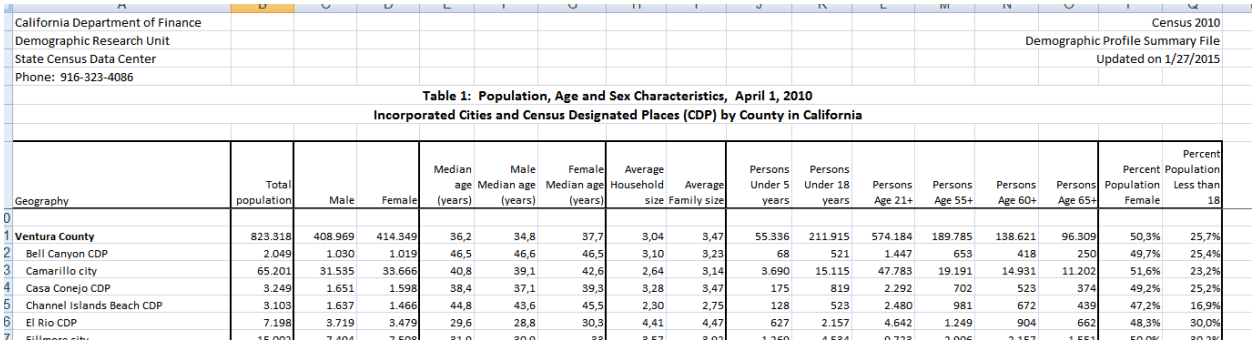
This list of requirements, when applied to a state the size of California, leaves plenty of options to choose from, and data analysis will be needed to filter potential locations and find the ones with the best chances of success. Given budget limitations, all data will need to come from free sources.

# 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 — Census 2010
Demographic Research Unit — Demographic Profile Summary File
State Census Data Center — Updated on 1/27/2015
Phone: 916-323-4086

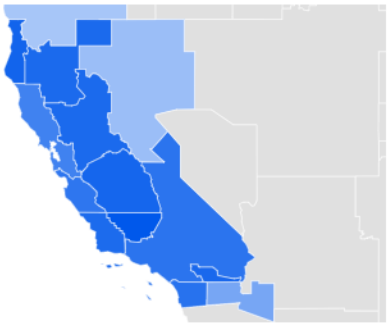
**Table 1: Population, Age and Sex Characteristics, April 1, 2010**
**Incorporated Cities and Census Designated Places (CDP) by County in California**

| | 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 | 30,9 | 33 | 3,57 | 3,92 | 1.269 | 4.584 | 9.723 | 2.006 | 2.157 | 1.551 | 50,0% | 30,1% |

## 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.

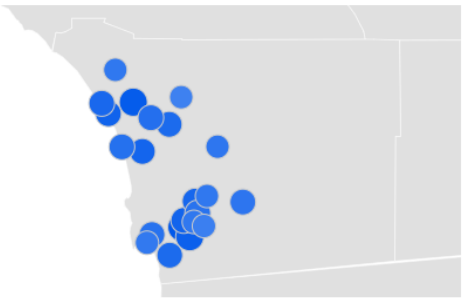
The challenge is that the page uses a tree-structure that makes it impossible to view detailed data for more than one geographical area at a time. For this reason, it is possible to compare the intensity of searches between metropolitan areas of the same state, and between cities of the same metropolitan 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.



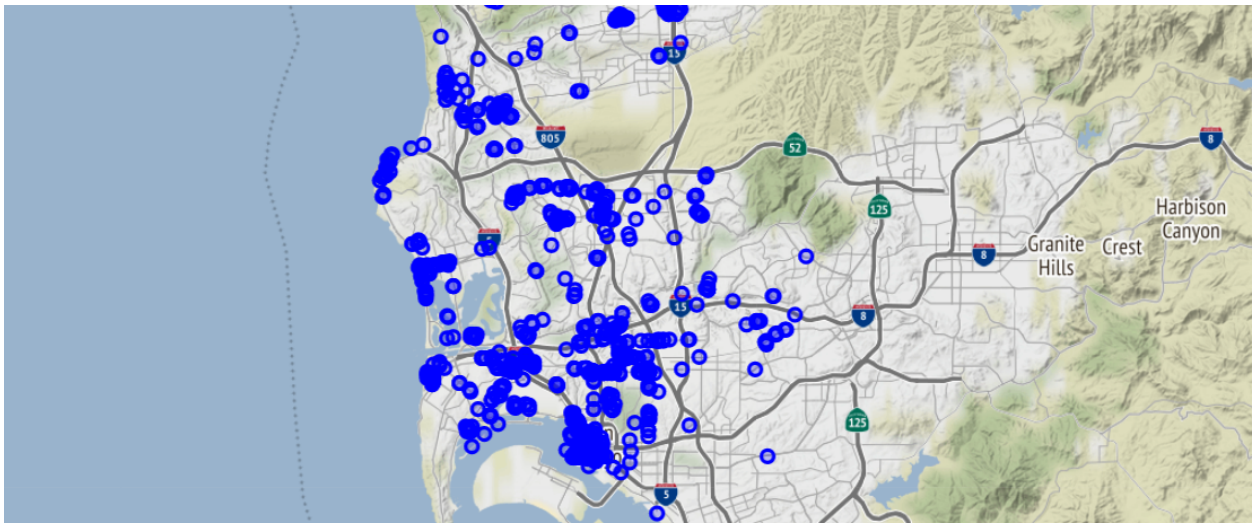
| | | |
|---|---|---|
| 1 | Lemon Grove | 100 |
| 2 | Vista | 97 |
| 3 | La Presa | 93 |
| 4 | Carlsbad | 91 |
| 5 | La Mesa | 90 |

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.

# Methodology

This project has been set up as a multi-stage filter that will help a family-owned business identify the best place to relocate. It consists of four stages that bring down the number of potential location from 1585 to just 3 analyzing the whole of California. This methodology section has been structured in the same way.

## Step 1: Potential locations in California

The first step in the project was a fairly straightforward one. The client had made it clear to us that they wanted to relocate to California and that they were only interested in cities with at least 200 thousand inhabitants.

A full list of cities and populations was sourced using the California census page that allowed data to be downloaded as xlsx files. The cleaning of this dataset proved to be the most work-intensive part and an early stage was done directly in Excel with the removal of all unnecessary tabs and columns.
The file was imported in Python and stored as a Pandas Dataframe.

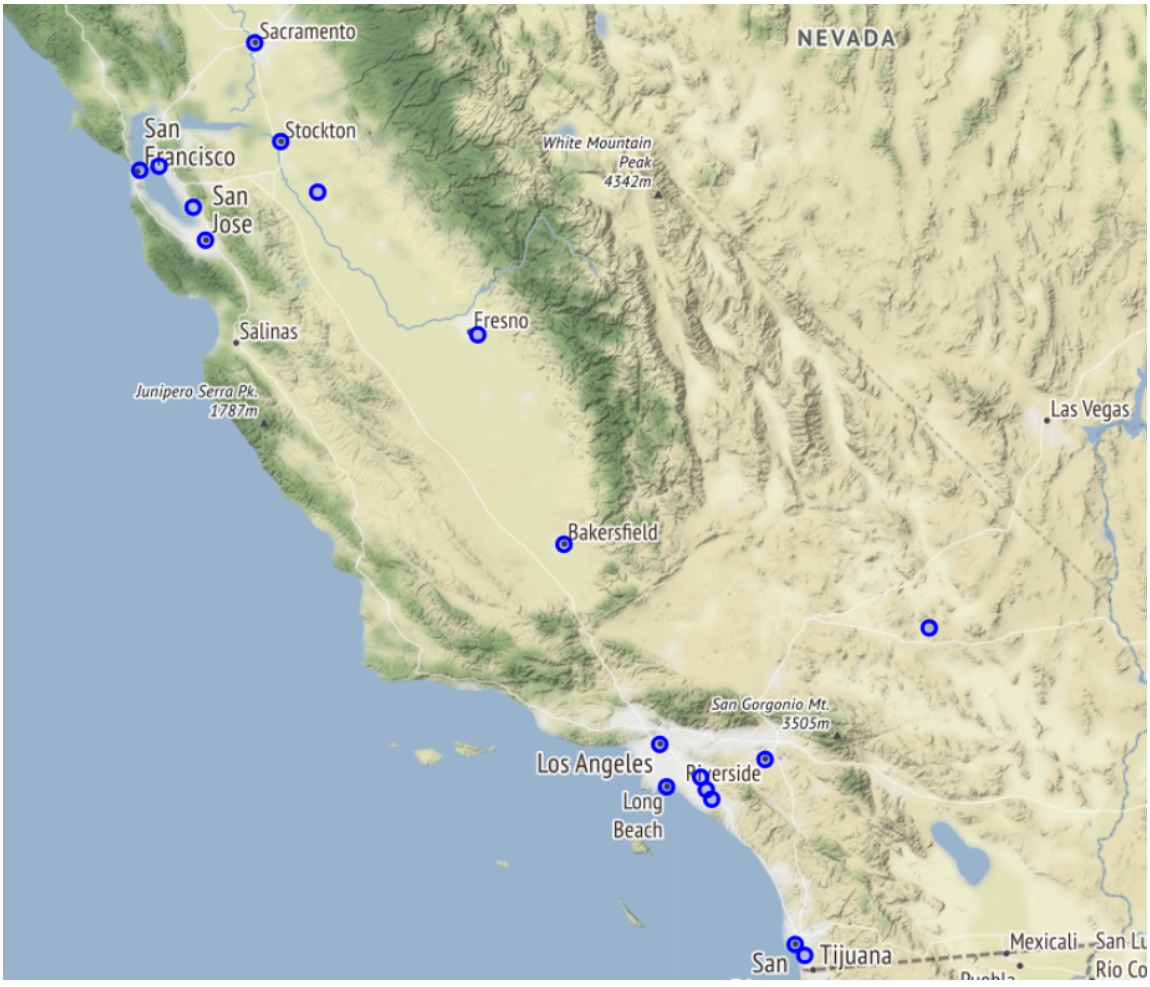
## Data preparation

One important step in the cleaning process was the removal of all suffixes from the location names. In the original document, each place carried the suffix “city”, “town” or “CDP” depending on the legal status. The removal was necessary to allow names to be more easily matched with other data sources that were used later. Another element that had to be removed were the totals for counties and regions that could have led to problems.

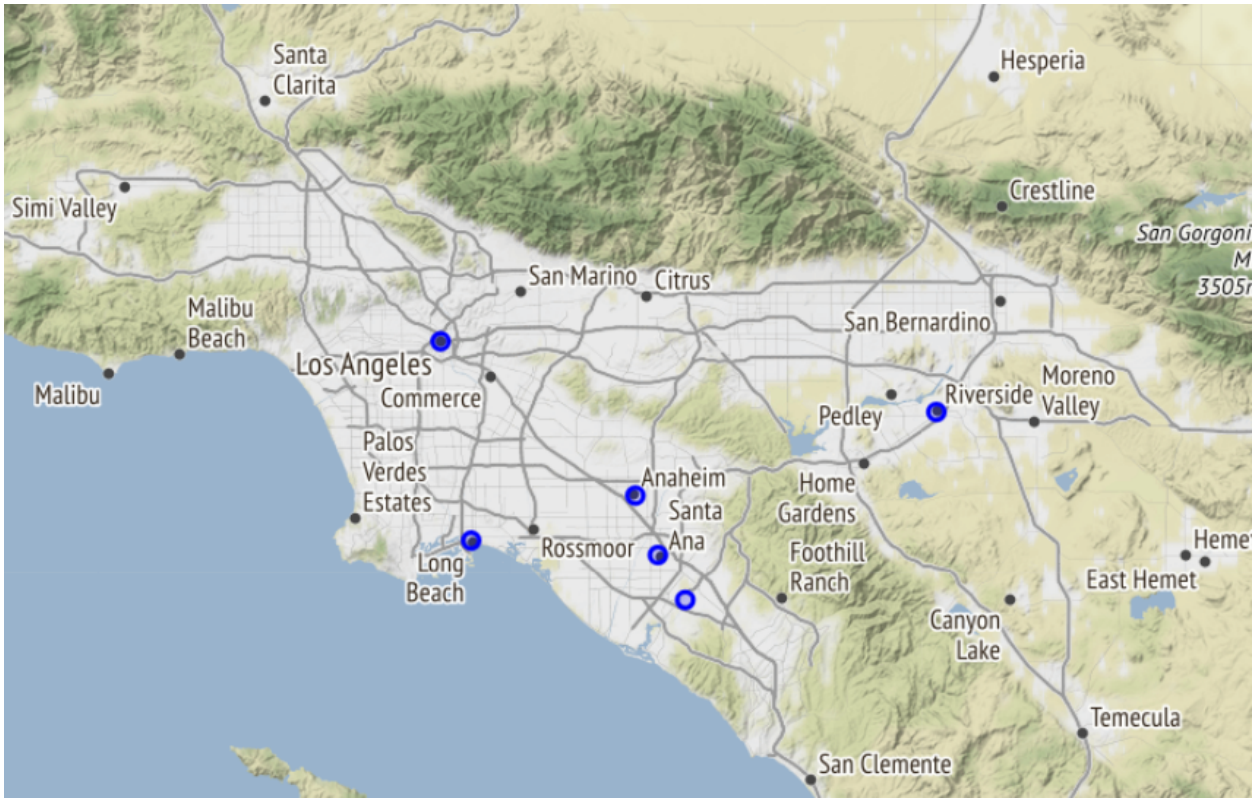
As an initial filter, all cities of less than 200 thousand people were also removed at this stage, bringing the number of potential locations down to just 18.

## Coordinates and visualization

Coordinates for each location were then added to the dataframe using Geolocator and Nominatim and Folium was used to create a map.



A visual analysis of the map revealed that the 18 locations were spread throughout California and that some clustering was present in the Los Angeles area. Looking at the area in more detail it became evident that some of these cities were actually suburbs or regions of Los Angeles.




These locations were not merged into a single one because despite being part of a single metropolitan area their size and distance from each other justified an individualized analysis.

# Step 2: Understanding the interest in Pizza

Google Trends' data was downloaded from the site in a series of 13 CSV files. 12 represented metropolitan areas* and one the state of California. This data source was discussed in detail in the Data section.

For the purposes of this section, it is important to keep in mind that the data is provided in the form of a score (from 0 to 100) that compare the intensity of searches for the keyword "pizza" in the regions contained in each CSV file.

(* NOTE:: the term "metropolitan area" is used by Google in a loose sense and usually represents just a subsection of a state)

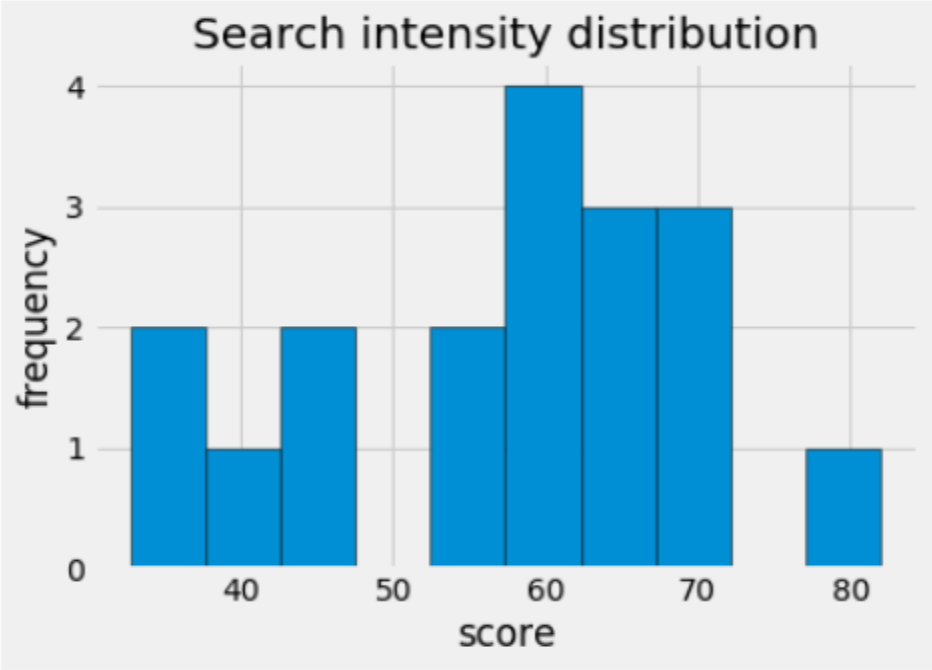
## Importing and merging

The first metropolitan region CSV file was imported individually to test the cleaning process and the weighting of the search values with input from the state-level sheet. Once this was successfully completed a function was created to automate the process.

The weighting of city scores was performed by multiplying the score by that for its municipal area and dividing the result by 100. The resulting DataFrames were then merged into a single one using the append method.

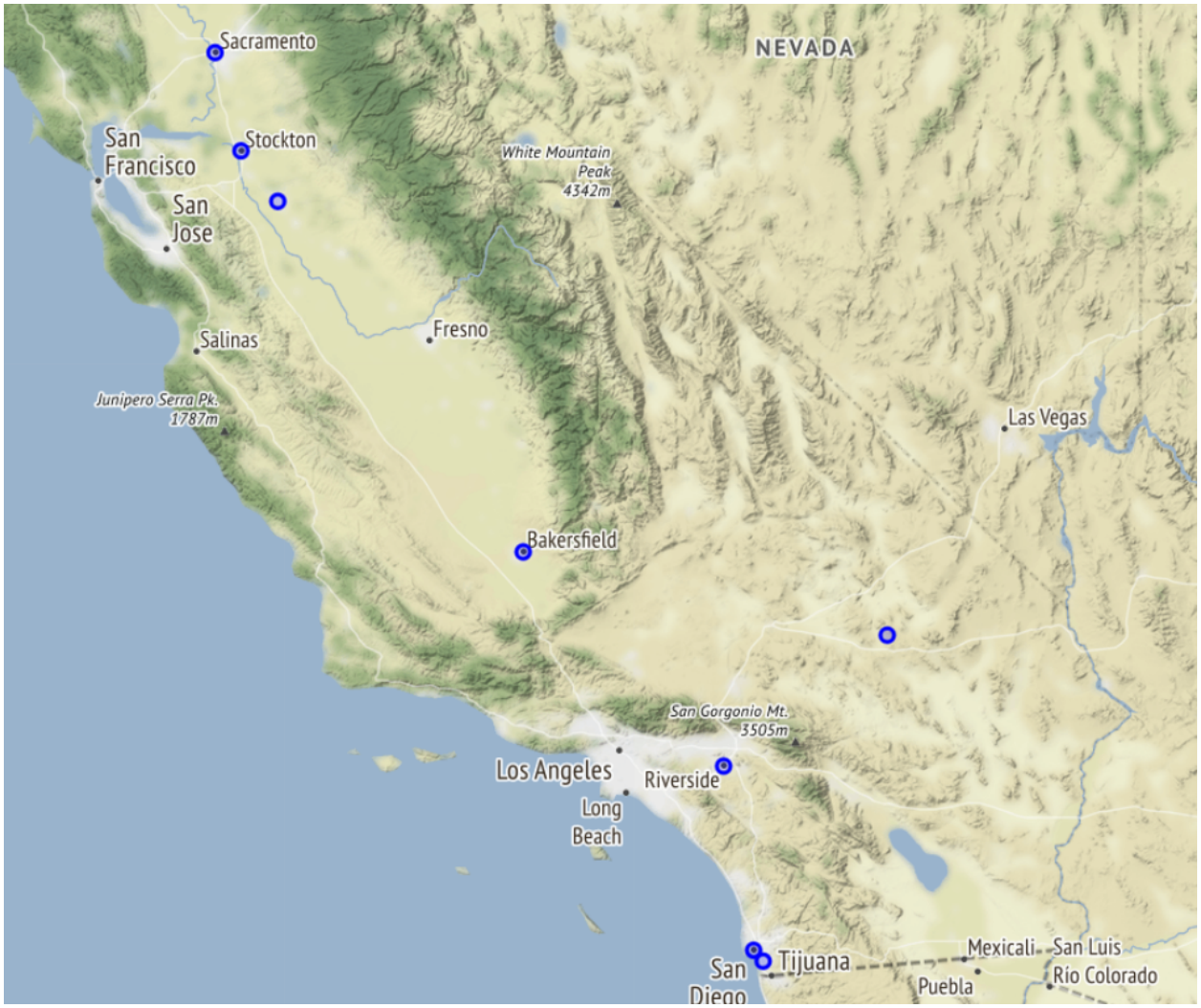
## Combining with census data and filtering

At this stage, the dataframe with the 18 cities selected in the previous stage was matched with the search score dataframe to add search score values to each. The resulting dataframe was then plotted to better understand the distribution.




Statistical analysis of the data revealed that the median search score was 60.6 and the data point was used to further reduce the number of candidate locations. By selecting only those with scores above the median the set was reduced to 8 cities.

Interestingly despite the reduction in number the geographical distribution did not vary significantly in terms of how spread out they were in the state but there was a marked decline in the metropolitan areas of San Francisco and Los Angeles




# Step 3: Identifying restaurant clusters

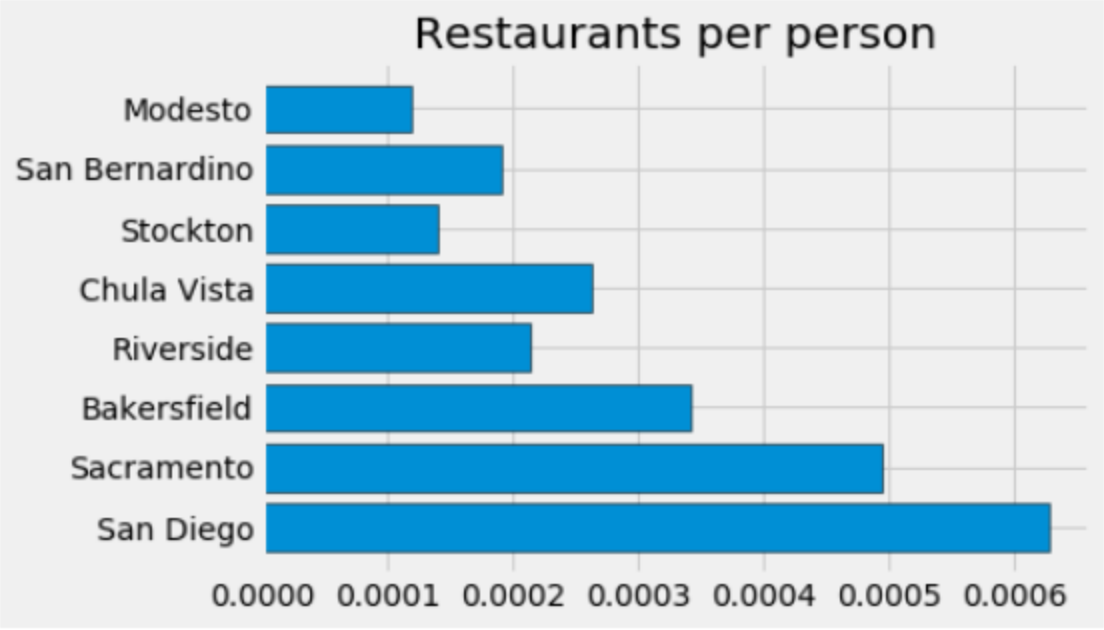
Having created a shortlist of suitable cities it is now time to go into more detail and look for potential clusters of restaurants in each city. The client informed us that clusters are important because, as a general rule, restaurants that are located in these areas tend to do better since the cluster attracts potential customers and creates a better atmosphere.

The goal in this stage is to create shortlist of clusters. These clusters will be analyzed individually as local hotspots and may or may not be part of the same city. The starting point to identify the clusters will be the shortlist of 8 cities.

The ideal cluster will be well-defined, contain enough restaurants to be viable and ideally only a small % of pizza places. To find it data from Open Street Maps, obtained via the OverPass API will be used.

## Searching for restaurants

As a first step, the API was queried to all restaurants in the shortlisted areas and the results were outputted to a dataframe. Then results were grouped by the city to understand how the number of restaurants compared with its population.

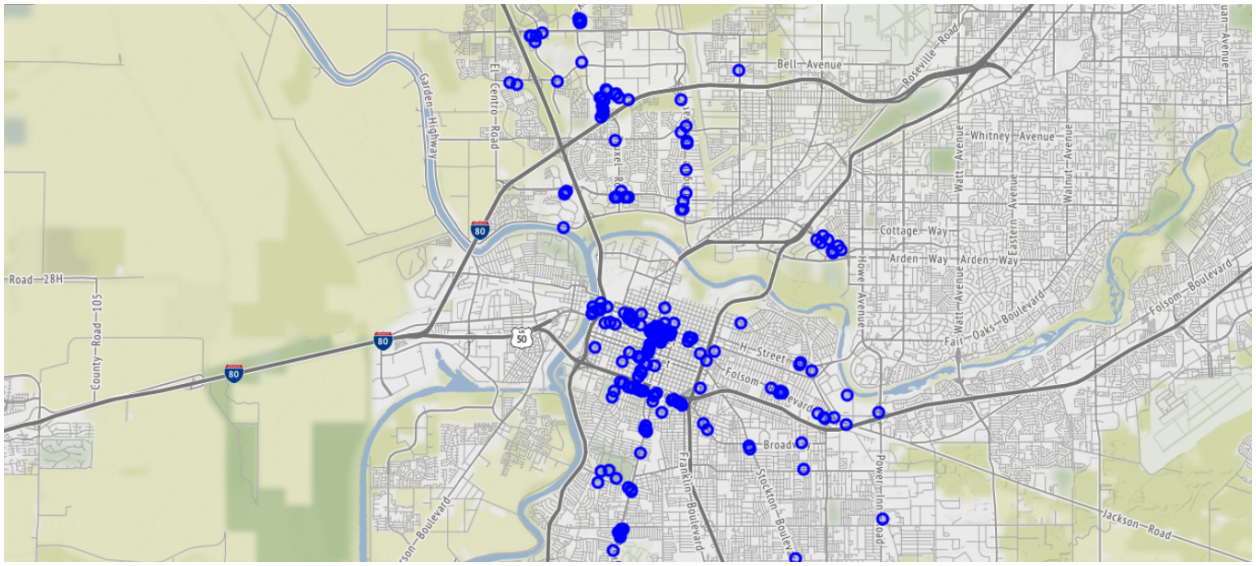



In the filtering process, a higher number of restaurants per person was considered as an advantage since it could indicate that there is a higher demand for the service.

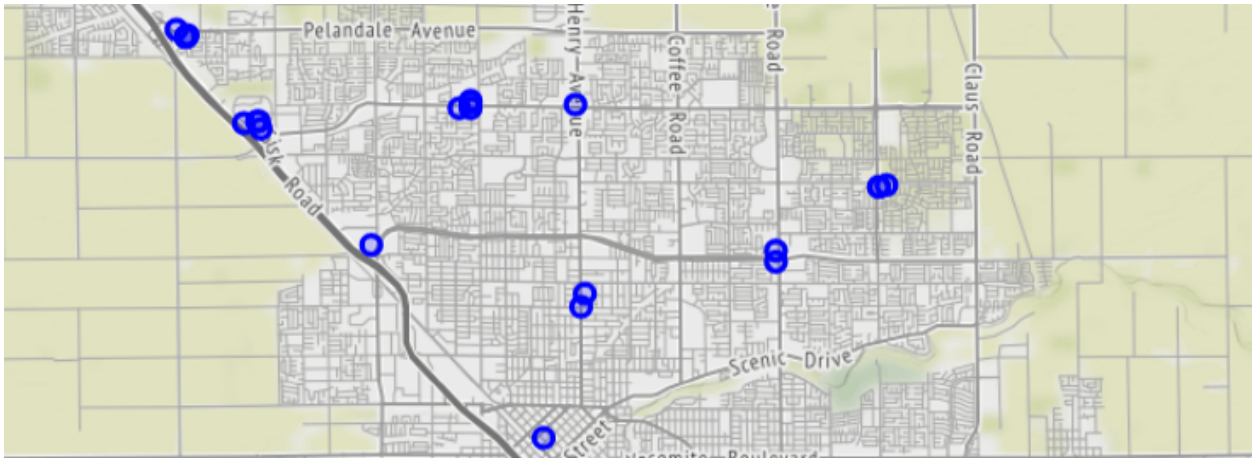
## Searching for clusters

The dataframe was then divided by city and individual maps were created to visualize the results. This quickly revealed that some cities had plenty of viable clusters while others didn't.

A good example is the city of Sacramento (see picture below) where a clear concentration of restaurants can be seen in the town center and some smaller clusters can be seen outside it.




This contrasts with cities like Modesto that tended to have more spread out patterns




## Shortlisting

As mentioned earlier for a restaurant to thrive it should ideally be located near a cluster of other restaurants that might attract clients. Looking at the maps above some cities like Sacramento, stand out for having clearly defined clusters. As the first step in this qualitative stage of the filtering process, we'll rule out the cities of Modesto, Riverside and Bakersfield since their restaurants appear to be very spread out in the city and do not have clear clusters. This leaves us with Sacramento, San Bernardino, Chula Vista, San Diego and Stockton.

Of these, the cities of San Bernardino and Stockton have the lowest number of restaurants per inhabitant. This could be interpreted both as an opportunity and as a sign of weak local demand. Looking at the map in more detail we can see that in Stockton the restaurants aren't very well

clustered and tend to line up with two main roads, which could suggest a less-than-lively local scene. The same is true for San Bernardino and Chula Vista

Ruling out these cities we are left with two contenders, San Diego and Sacramento. Both have a high concentration of restaurants and what appears to be multiple clusters of restaurants that could serve as potential locations for the new pizza place. To understand them better we'll first use machine learning to identify the centre of the clusters and then investigate them in more detail with the FourSquare API.

# Step 4: Detailed cluster analysis

Having isolated San Diego and Sacramento as the two contenders it is now time to dive deeper and identify which neighborhoods are hotspots for restaurants in each city and use them to determine the best place to open the pizza place.

However, since traditional neighborhoods borders are sometimes shaped by historical factors and not by the characteristics of each area we'll use a different system based on identifying clusters with machine learning and then using the center of each cluster and a radius as the definition of each area.

## Choosing an algorithm

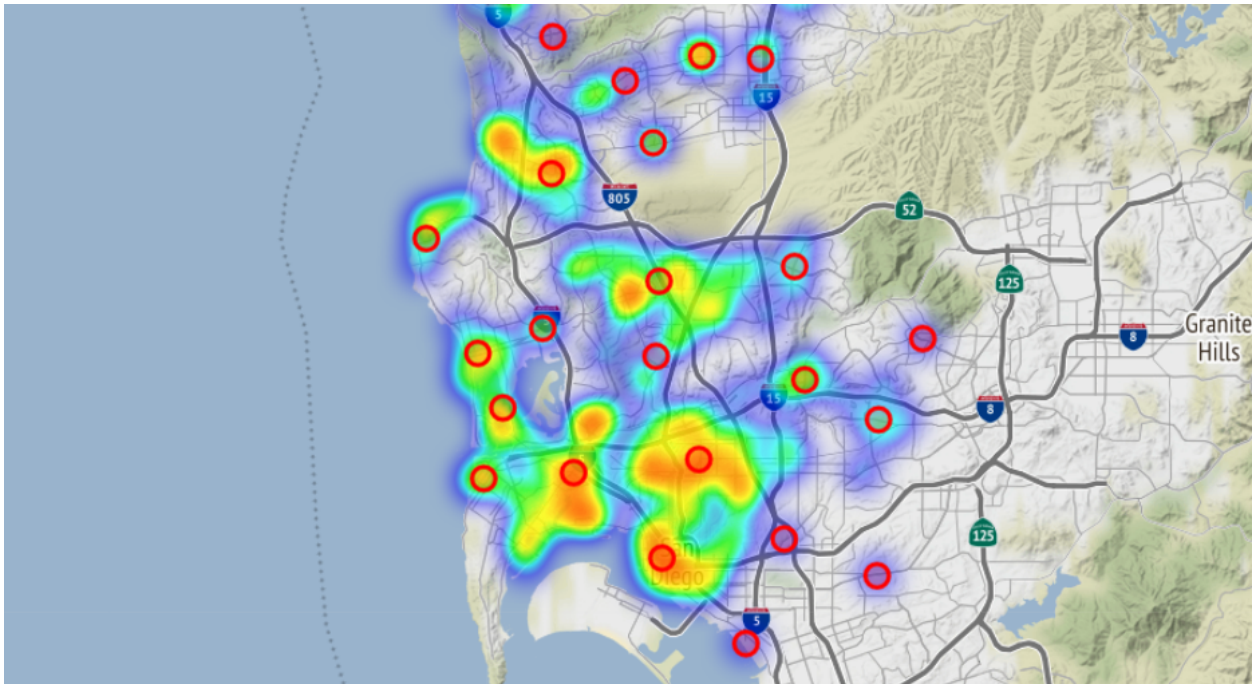
To be able to conduct detailed search using the Foursquare API we will need to define the cluster’s centres. In the previous step we have already been able to get an idea for their location but now it is time to apply machine learning to the problem and get the right coordinates.

The right algorithm for the job needs to be capable of finding high-density areas and ignore items that aren't part of them. This is because many restaurants are isolated from the rest and effectively act as noise in the data.

K-means is therefore not a suitable option since it is too heavily influenced by noise. Mean Shift, on the other hand, appears to be a better tool, especially if used in conjunction with a heat-map that will help validate the viability of each cluster.

## Cluster selection

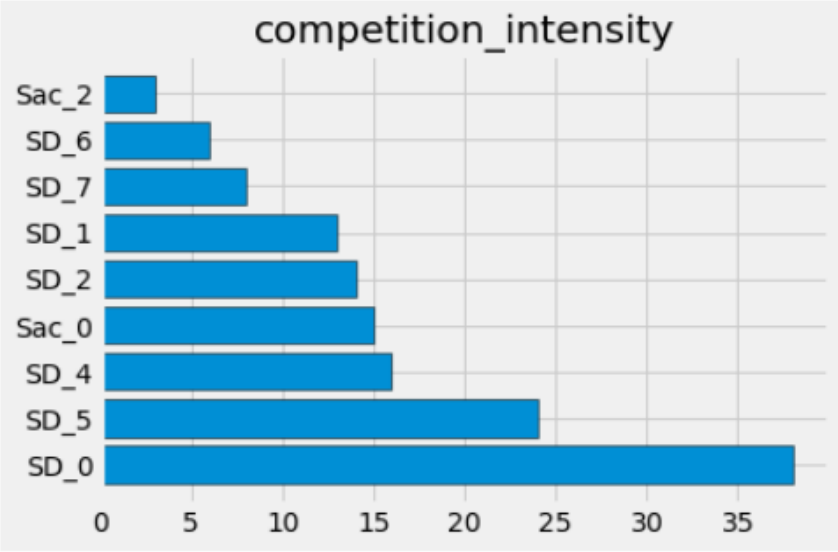
Applying Mean Shift to the dataset created a number of cluster centres. However, not all clusters were created equal as evidence by a heat map layer derived from the restaurants' dataset. (the example below is from the city of San Diego




From the maps of the two cities, the most promising clusters were shortlisted. Of these 7 belonged to the city of San Diego and 2 to Sacramento.

## First Foursquare search

To understand how many other pizza places were already present in each area Foursquare's API was used to conduct a search of potential competitors in a 1.5km radius from the center of the cluster. The results were then plotted:



Analysis of competition intensity led to the elimination of the two areas with the highest concentration and the reduction of the shortlist to 7.

## Bypassing Foursquare’s limitations

To get a clear idea of what can be found in each area it is necessary to obtain a full list of all the eating establishment that can be found in a range of 1 km from the centre of each area.

The challenge is that Foursquare's API can only return a maximum of 50 items per search. This means that if we were to query it for all the food places found within 1 km it would return a list of 50 items, but it would be an incomplete list.

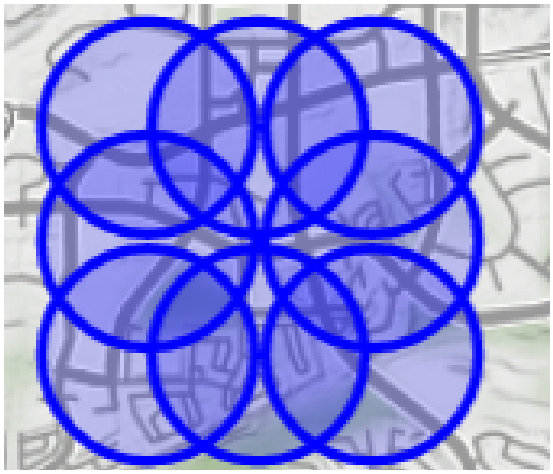
After some trials, it was determined that, in the densest areas, the maximum search range to get a complete list of results is 500 meters. If we convert this radius to an area we quickly discover that a circle of range 500 has only a quarter of the area of one of range 1000 and gives us a very limited search area.

### Cluster offset function

To get around the problem a function was created to return a series of 6 points each offset 500 m in latitude and/or in longitude. By using the function it became possible to conduct separate searches of radius 500 on each of these points, combining them and discarding duplicate locations gives us a full picture of the 1000 m area.

Since converting meters to a change in coordinates is not a straightforward process the function used a while loop where the coordinates were increased by a small amount each time and the distance measured with GeoPy’s distance method until the 500m offset was achieved.

The resulting pattern of searches looked like this:



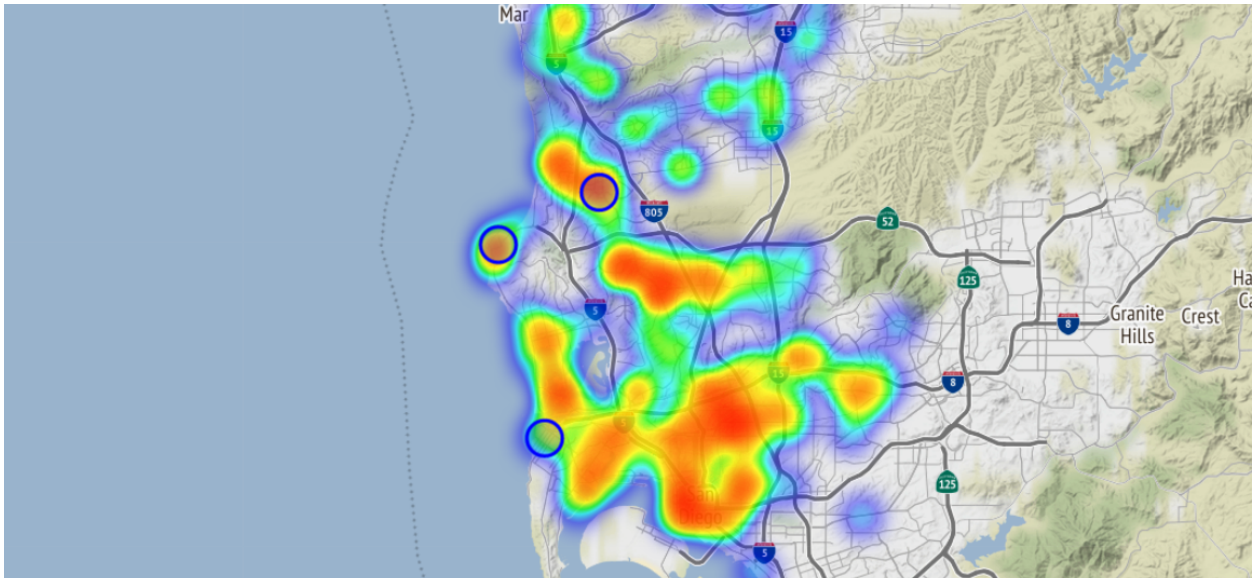
## Second Foursquare search

The new search patterns were then applied to Foursquare API queries that were made using the category code for “food”. This category was chosen since it was the only one that included all types of restaurants.  Unfortunately, the category also included places like bakeries or candy shops that had to be filtered out before the resulting dataframe could be used.

The data was then filtered to calculate the total number of restaurants in each area and the total number of pizza places.  A ratio of restaurants to pizza places was also calculated.

| | area_code | latitude | longitude | food_places | pizza_places | ratio |
|---|---|---|---|---|---|---|
| **0** | SD_4 | 32.867066 | -117.215628 | 112 | 6 | 18.67 |
| **1** | Sac_0 | 38.569378 | -121.486924 | 160 | 13 | 12.31 |
| **2** | SD_2 | 32.749573 | -117.205504 | 157 | 11 | 14.27 |
| **3** | SD_1 | 32.754667 | -117.147377 | 142 | 11 | 12.91 |
| **4** | SD_7 | 32.841568 | -117.274516 | 127 | 7 | 18.14 |
| **5** | SD_6 | 32.747568 | -117.247702 | 91 | 5 | 18.20 |
| **6** | Sac_2 | 38.535104 | -121.495582 | 32 | 3 | 10.67 |

From this table, the three areas with the highest ratio of restaurants per pizza place were selected as the finalists and plotted on a map. They turned out to be located not far from each other in the north side of the city of San Diego.



# Results

The data analysis process revealed three ideal candidates. These were initially identified only by their cluster codes of SD_4, SD_7 and SD_6 and their coordinates but are actually parts of the areas of Costa Verde (SD_4), La Jolla (SD_7) and Ocean Beach (SD_6).

(Ocean Beach, San Diego)

Looking at their numbers in more detail we could say that the area of Costa Verde (SD_4) is slightly more promising. The choice, however, lies with the owners of the pizza place since the other two areas are located by the sea and may attract a different public compared to Costa Verde.

| | area_code | latitude | longitude | food_places | pizza_places | ratio |
|---|---|---|---|---|---|---|
| **0** | SD_4 | 32.867066 | -117.215628 | 112 | 6 | 18.67 |
| **1** | SD_7 | 32.841568 | -117.274516 | 127 | 7 | 18.14 |
| **2** | SD_6 | 32.747568 | -117.247702 | 91 | 5 | 18.20 |

# Discussion

As with any data science project when looking at these results it is important to understand that they are the results of a set of constraints and that these constraints influence the project.

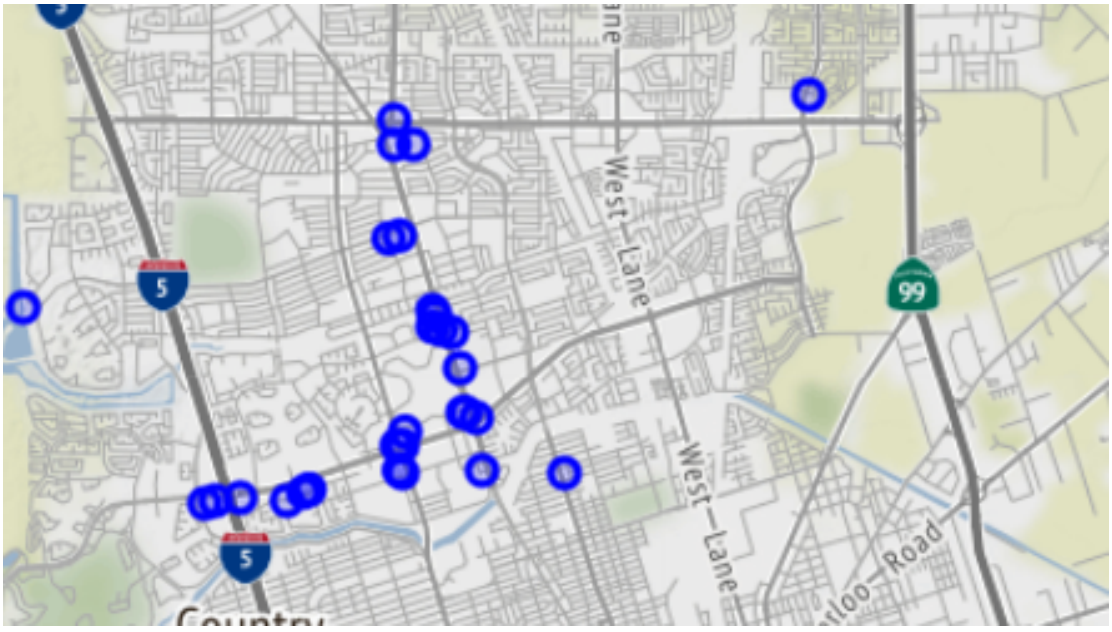
In this particular case, the main limitation was the need to use only free resources which limited the type and quality of data that could be used. For example, this project did not take into account real estate prices or any factors related to the cost of operating a pizza place in each area since the data required could not be obtained for free.

Having said that, the results do offer an interesting insight into how different nearby cities can be and especially in terms of the distribution of commercial activities.

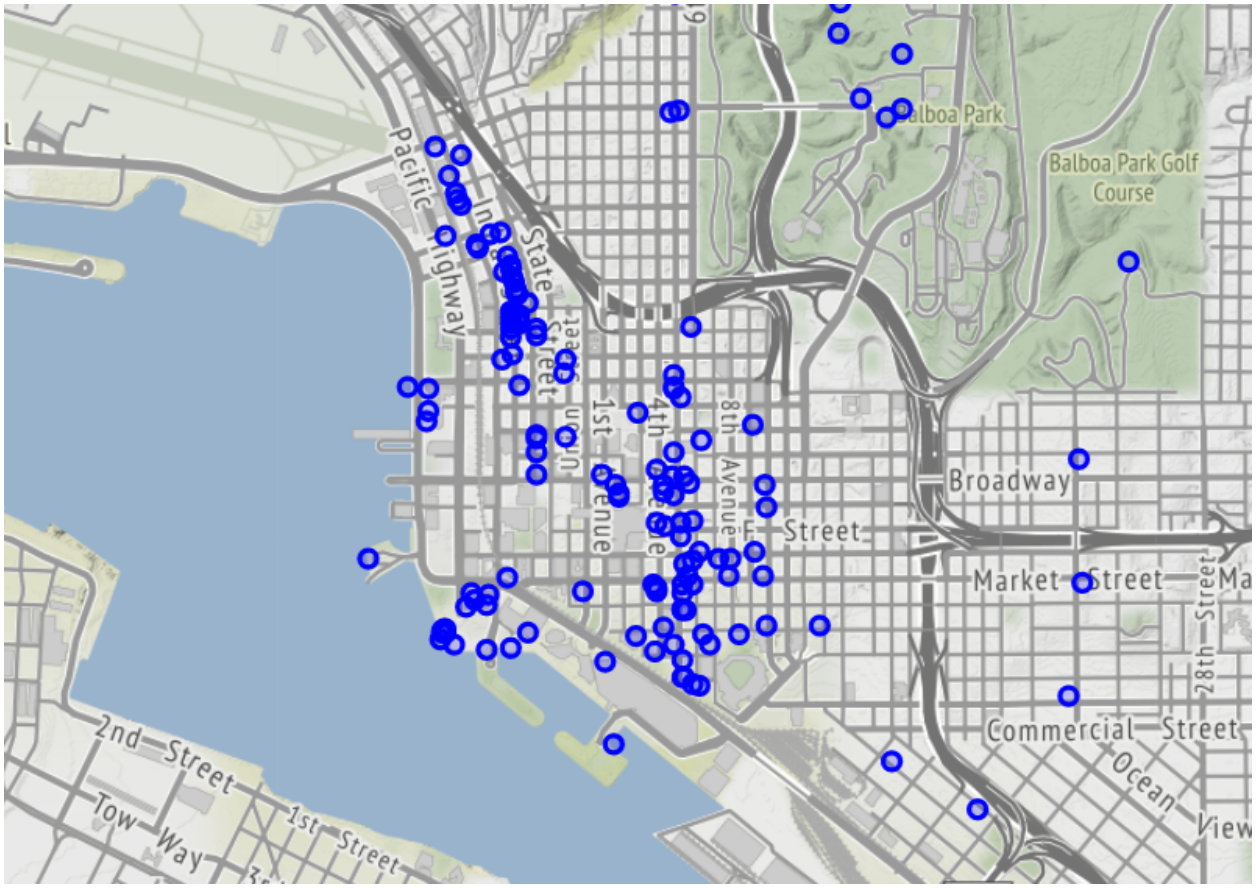
In some cases looking at the map, it became clear that the restaurants were arranged along a single street. This could be good for some businesses but it also suggests that most of the customers to the area arrive by car do not explore the area beyond the street they drive on.

For a small family restaurant, this could be a problem since their only option would be to open on that street (which will likely have inflated costs) and will require heavy investments in signage to become visible from cars.

A good example is the city of Stockton depicted below:



Other cities, by contrast, had what appeared to be lively centres where clusters were dense but restaurants were not limited to a single road or sets of streets.




# Conclusion

The final choice of location will ultimately depend on the type of pizza place, budget and style that the owners decide on. The code used is published on GitHub and can be adjusted to account for any additional input that the owners of the pizza place may wish to make.

Please note that this work was conducted as a capstone project for the IBM Data Science Professional Certificate and was not intended for commercial or real-world usage.