The Battle of Neighborhoods

1. Introduction

1.1 Background

California is a state in the Pacific region of the United States. The state has an area of about 163,696 square miles, which makes it the third largest state in the US in terms of land area. In terms of population, California is the biggest state. The capital city is Sacramento while Los Angeles is the most populated city in the state. The population size of Los Angeles is second only to that of New York City among the cities in the US. Another interesting fact about California is that it has the largest economy of any state. For these reasons and others, the state is known globally for several things ranging from entertainment and relaxation to business and economy. Therefore, it is advantageous to grow business in California.

1.2 Business Problem

A customer decided to expand his restaurant chain to California. But how do they decide where should open their restaurant? What type of restaurant he should open? Is it any competitors available? There could be lots of factors that could potentially affect how the business would go, so choosing the right neighborhood is critical!

The idea of this study is to help the customer planning to open a new restaurant in California to choose the right location by providing data about the population of each neighborhood as well as the competitors already present on the same regions.

1.3 Interest

The customers who interested to operate a restaurant in California.

2. Data section

2.1 Data Sources

Based on definition of our problem, factors that will influence our decision are:

1. Number of existing restaurants in the neighborhood (any type of restaurant)

2. Population of the neighborhood in California

Following data sources will be needed to extract/generate the required information:

1. List of cities and towns in California with population: https://en.wikipedia.org/wiki/List_of_California_locations_by_income

   To provide the stakeholders the necessary information the list of cities and towns in California will scrape from the Wikipedia page. It contains Population of each neighborhood in California.

2. Foursquare API to collect competitors on the same neighborhoods.

   The majority of the data used in this project will be taken from Foursquare API, which is a crowd sourced, comprehensive geographical data source. With Foursquare API, we will be able to get insight on the most popular venues in each neighborhood, their photos, ratings and customer comments on those venues.

3. Coordinate of neighborhoods in Los Angeles County

   We need to get the coordinate in order to utilize the Foursquare location data. The coordinates of neighborhoods in Los Angeles in available in this link. https://docs.google.com/spreadsheets/d/1hOuAeqMuu6lJwKrtTepyHdk0ljAouGr_BIWtCTHEvmM/export?format=csv&gid=2047060771

4. California's geojson

   To visualize population in California with a choropleth map, we need a geojson file of California. The geojson can be obtained from this link. https://gist.githubusercontent.com/threestory/ed0f322d7bb2e3be8ded/raw/2aa8df2f15817985c60b67d726e6d13197e8ffb6/cb_2014_us_county_5m.json

2.1 Data processing and cleaning

Firstly, I scrape the "List of cities and towns in California with population" from the Wikipedia page and wrangle the data. There are 2 tables in the web portal, one with City column and one without City Column. I extracts the first table and read it into a pandas dataframe (Table 1). I use this table to compare the populations of each county in California. There are some unclean data 0 or [7] and removed from the table.

	County	Population
0	Alameda	1559308
1	Alpine	1202
2	Amador	37159
3	Butte	221578
4	Calaveras	44921
5	Colusa	21424

Table 1

Then, I use the data from Table 1 to create a choropleth map to compare population of each county. From the map, I can choose a county with highest population and then extract the neighborhood from second table in the wikipedia page and remove unwanted data and column. The coordinates of each cities get from csv file mention in Data Source no.4 , Los Angeles cities table and csv table to form Table 2.

|	Neighborhood	County	Population	Latitude	Longitude
0	Acton	Los Angeles	6,956	34.467170	-118.197180
1	Agoura Hills	Los Angeles	20,630	34.147980	-118.781242
2	Agua Dulce	Los Angeles	2,898	34.496320	-118.326200
3	Alhambra	Los Angeles	84,400	34.095287	-118.127014
4	Alondra Park	Los Angeles	8,833	33.887980	-118.330580
5	Altadena	Los Angeles	44,622	34.185680	-118.131508
6	Arcadia	Los Angeles	57,251	34.127361	-118.045868
7	Artesia	Los Angeles	16,698	33.861370	-118.082161
8	Avalon	Los Angeles	3,768	33.342800	-118.327900
9	Avocado Heights	Los Angeles	15,432	34.036300	-117.991050
10	Azusa	Los Angeles	47,420	34.133480	-117.907539

Table2

After generated Table 2, I use Foursquare API to get the list of existing restaurants for each neighborhood. The list contain some NaN some, I remove the row with NaN value and then start analysis.

3. Methodology

Exploratory Data Analysis

I create a choropleth and boundary map of California to get a visual of population distribution, darker red means higher population and lighter yellow means lower population. From the map we can see Los Angeles has the highest population among the counties. So, I choose Los Angeles to analyze the existing restaurants.
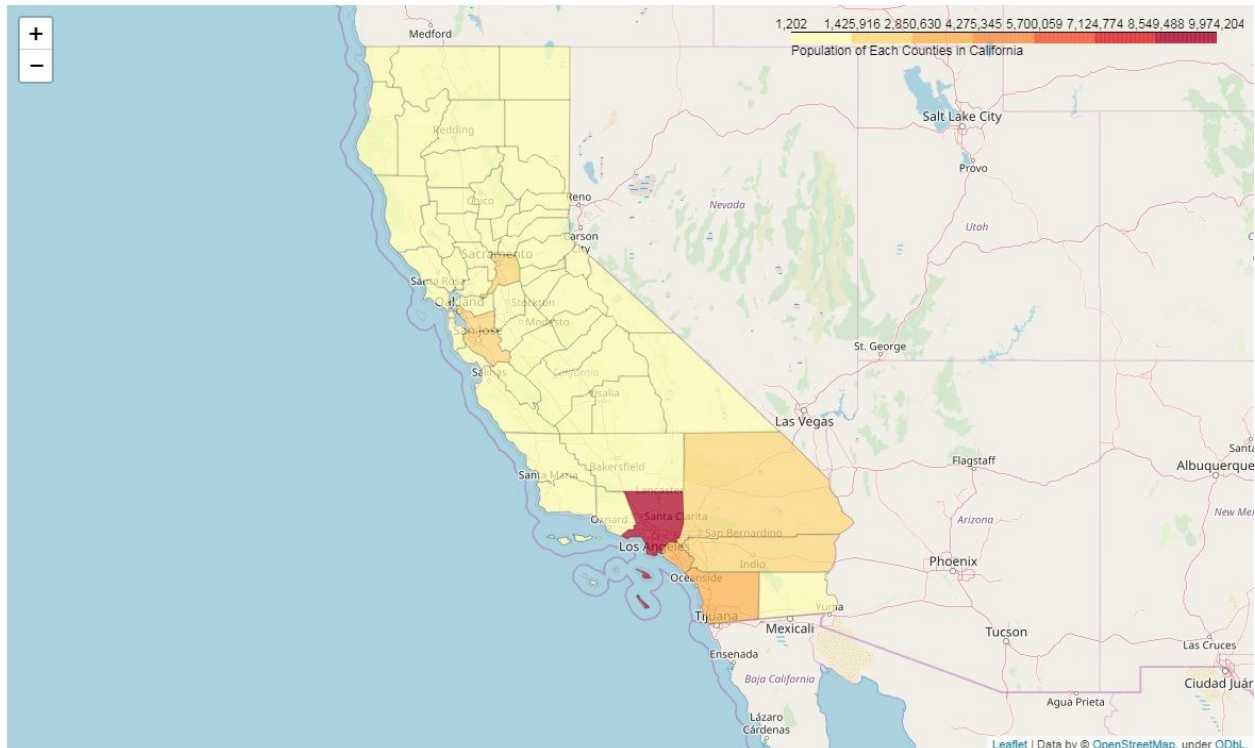
California's Population Choropleth Map

Analyzing neighborhoods in Los Angeles to get most common restaurant for each area using Foursquare API

Next, I downloaded restaurants data from Foursquare and analyze the venues and venue category. I created the table below to shows the restaurants with the most competitor.

Venue Category	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude
Mexican Restaurant	192	192	192	192	192	192
Pizza Place	126	126	126	126	126	126
Chinese Restaurant	116	116	116	116	116	116
American Restaurant	96	96	96	96	96	96
Sandwich Place	87	87	87	87	87	87
Fast Food Restaurant	86	86	86	86	86	86
Bakery	71	71	71	71	71	71
Burger Joint	68	68	68	68	68	68

I created the table below which shows the top 10 most common restaurant of each neighborhood.
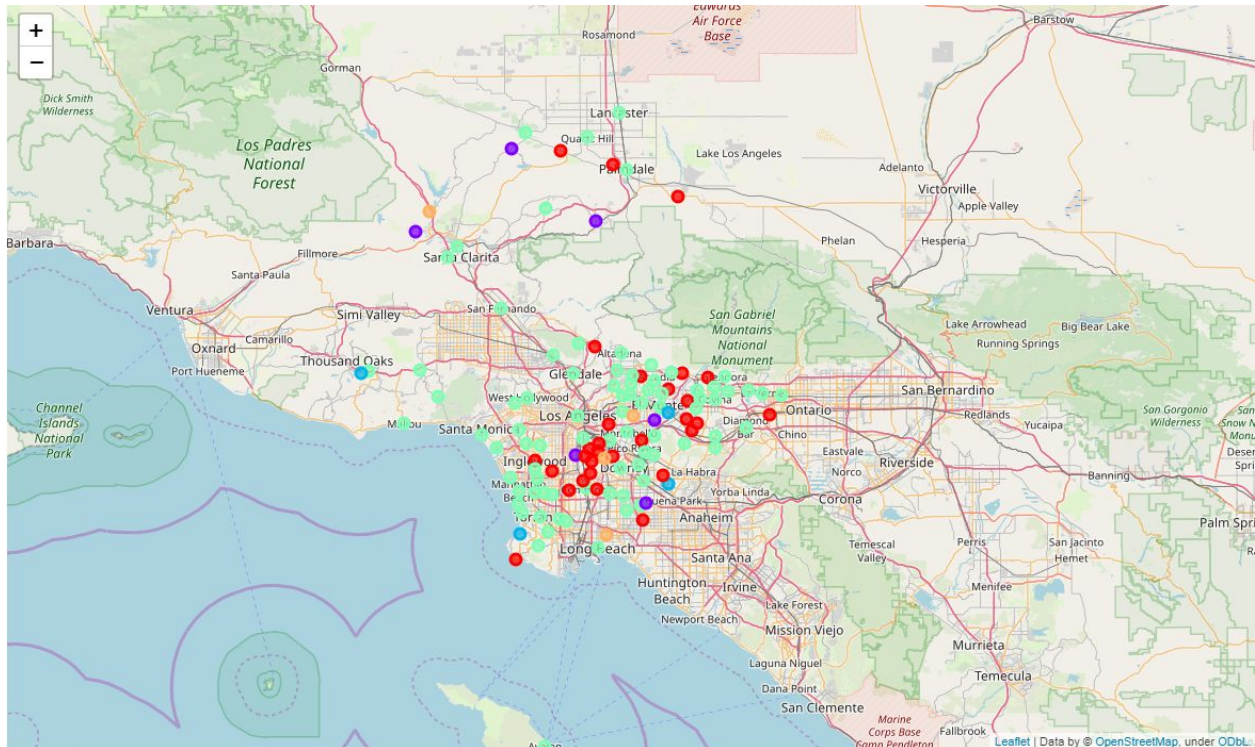
	Neighborhood	1st Most Common Restaurant	2nd Most Common Restaurant	3rd Most Common Restaurant	4th Most Common Restaurant	5th Most Common Restaurant	6th Most Common Restaurant	7th Most Common Restaurant	8th Most Common Restaurant	9th Most Common Restaurant	10th Most Common Restaurant
0	Acton	Café	Food	Wings Joint	Eastern European Restaurant	Deli / Bodega	Dim Sum Restaurant	Diner	Donburi Restaurant	Donut Shop	Dosa Place
1	Agoura Hills	Café	Ramen Restaurant	Sushi Restaurant	Italian Restaurant	Mediterranean Restaurant	Pizza Place	Gastropub	American Restaurant	Bagel Shop	Fast Food Restaurant
2	Agua Dulce	Restaurant	Pizza Place	Café	Bakery	Mexican Restaurant	Eastern European Restaurant	Dim Sum Restaurant	Diner	Donburi Restaurant	Donut Shop
3	Alhambra	Chinese Restaurant	Sushi Restaurant	Vietnamese Restaurant	Burger Joint	Diner	Seafood Restaurant	Korean Restaurant	Asian Restaurant	Bakery	Japanese Curry Restaurant
4	Alondra Park	Fast Food Restaurant	Breakfast Spot	Mediterranean Restaurant	Restaurant	Asian Restaurant	Sandwich Place	Donut Shop	Pizza Place	Burger Joint	Mexican Restaurant
5	Altadena	Pizza Place	Food	Bakery	Diner	Sandwich Place	Mexican Restaurant	Burger Joint	Breakfast Spot	Fast Food Restaurant	Food Truck

Clustering algorithm (K-mean)

I perform a K-Means clustering algorithm on the restaurant data and partition them into five different clusters.

	Neighborhood	County	Population	Latitude	Longitude	Cluster Labels	1st Most Common Restaurant	2nd Most Common Restaurant	3rd Most Common Restaurant	4th Most Common Restaurant	5th Most Common Restaurant	6th Most Common Restaurant	7th Most Common Restaurant	8th Most Common Restaurant	9th Most Common Restaurant
0	Acton	Los Angeles	6,956	34.467170	-118.197180	1.0	Café	Food	Wings Joint	Eastern European Restaurant	Deli / Bodega	Dim Sum Restaurant	Diner	Donburi Restaurant	Donut Shop
1	Agoura Hills	Los Angeles	20,630	34.147980	-118.781242	3.0	Café	Ramen Restaurant	Sushi Restaurant	Italian Restaurant	Mediterranean Restaurant	Pizza Place	Gastropub	American Restaurant	Bagel Shop
2	Agua Dulce	Los Angeles	2,898	34.496320	-118.326200	3.0	Restaurant	Pizza Place	Café	Bakery	Mexican Restaurant	Eastern European Restaurant	Dim Sum Restaurant	Diner	Donburi Restaurant
3	Alhambra	Los Angeles	84,400	34.095287	-118.127014	3.0	Chinese Restaurant	Sushi Restaurant	Vietnamese Restaurant	Burger Joint	Diner	Seafood Restaurant	Korean Restaurant	Asian Restaurant	Bakery
4	Alondra Park	Los Angeles	8,833	33.887980	-118.330580	3.0	Fast Food Restaurant	Breakfast Spot	Mediterranean Restaurant	Restaurant	Asian Restaurant	Sandwich Place	Donut Shop	Pizza Place	Burger Joint
5	Altadena	Los Angeles	44,622	34.185680	-118.131508	3.0	Pizza Place	Food	Bakery	Diner	Sandwich Place	Mexican Restaurant	Burger Joint	Breakfast Spot	Fast Food Restaurant
6	Arcadia	Los Angeles	57,251	34.127361	-118.045868	3.0	Restaurant	Himalayan Restaurant	Cuban Restaurant	Deli / Bodega	Dim Sum Restaurant	Diner	Donburi Restaurant	Donut Shop	Dosa Place
7	Artesia	Los Angeles	16,698	33.861370	-118.082161	3.0	Indian Restaurant	Chinese Restaurant	Vietnamese Restaurant	Bakery	Korean Restaurant	Seafood Restaurant	Sushi Restaurant	Japanese Restaurant	Pizza Place
8	Avalon	Los Angeles	3,768	33.342800	-118.327900	3.0	American Restaurant	Seafood Restaurant	Mexican Restaurant	Restaurant	Bakery	Deli / Bodega	Pizza Place	Breakfast Spot	Burger Joint

Use Folium map to visualize clustering result

4. Results

From the clustering result, I found that most of the neighborhood of Los Angeles has similar most common restaurants characteristics. 1st Most Common Restaurant of cluster 0 is Mexican Restaurant, cluster 1 is Food, cluster 2 is Cafe and cluster 4 is Taco Place, meanwhile cluster 3 has the most neighborhoods but there are variety of 1st Most Common Restaurant.

5. Discussion

The result of K-mean clustering is not 100% accurate, By examine the clustering results, I can see some neighborhoods have different most common restaurant compare with the neighborhood in the same cluster group.

I recommend the neighborhood for open restaurant based on number of population only. This recommendation may not fully advantageous for the customer. In future, I can add in the per capita income of each county for analysis, the result may different.

6. Conclusion

From the analysis results, I can conclude that:

1. Los Angeles is the best city in California to open a restaurant because it has highest population. High population will have high demand for food also.

2. Mexican restaurant, Pizza Place, Chinese Restaurant and American restaurant are the most common restaurants in Los Angeles. I will advise customer to avoid open this type restaurant since there are too many competitors.

3. I would recommend to open Korean restaurant since there are 19 existing competitors only. Besides that, I would recommend to open a Steamboat restaurant since there is no existing competitors.

Korean Restaurant	19	19		19	19	19	19