Battle of Neighborhoods

Data Science Project

Determining the Optimum Venue to Open in Los Angeles County

Yonghoon Andrew Kim

10th June 2020

Table of Contents

Introduction

Background / Interest

I have recently moved to Los Angeles (LA), California after living near Chicago, Illinois for four years. As a fairly new resident, I had my hopes up to explore one of the most vibrant cities in the U.S. However, the recent covid-19 outbreak had rendered this impossible, and I've been stuck at home ever since.

Since I've had more spare time, I've decided to take the data science course on Coursera. After learning about the ins and outs of data science and machine learning, I was eager to put all this knowledge to use. To my surprise, the course had a capstone, this felt like the perfect opportunity to merge my newly acquired knowledge and my desire for exploration.

Figure 1 below depicts the high quantity of distinct locations in LA. This led me to believe that there will be enough locational data and enough Foursquared data which I could explore into.

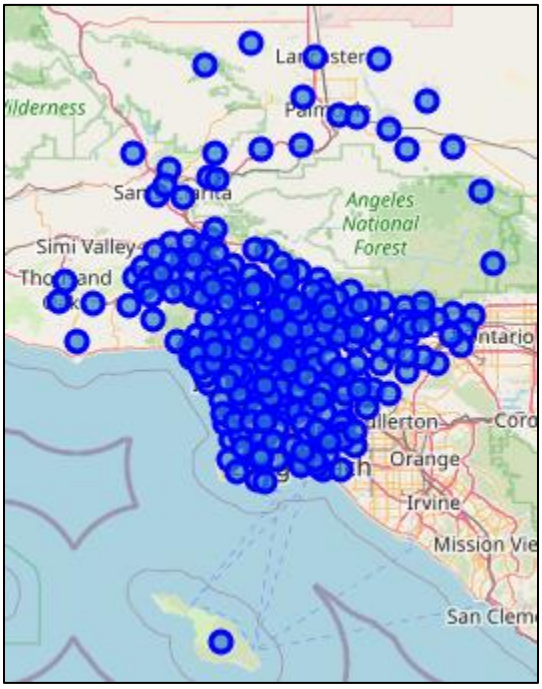

Figure 1. Cities of Los Angeles County

Problem

I have decided to go with a slightly different prompt from the given one. My idea was a slight altercation of the given prompt: "In a city of your choice, if someone is looking to open a restaurant, where would you recommend that they open it? Similarly, if a contractor is trying to start their own business, where would you recommend that they setup their office?"

My exploration highlights will take this a step further in explore the different venues grouped by median incomes of the cities. This will then be able to tackle the question of which area / city is most suitable to open what type of business.

Furthermore, the income data could potentially allow some rough predictions of how the business will perform in that specified area. All in all, the final problem prompt may be summarized as: “What kind of venues are popular for different income classes in LA?”

Data

Data Sources

The first dataset to be explored was the summary of city name with its respective zip codes. This data was very commonplace and was acquired through opendatasoft.com. The validity and reliability was checked by looking at a few cities and googling its zip code.

The second dataset described the median income of LA county cities per zip code. This was acquired through census.gov. The validity and reliability of this data is rather high as the dataset is government issued.

The third and final dataset will be acquired through the required FourSquared.com developer portal. As discussed in the lessons, this dataset is created by users and has practical value, therefore it is considered useful and valid. However, the reliability is a separate issue, but this will be overlooked as there are no comparators.

Data Cleaning

The data cleaning process was rather simple for the first two datasets as they only had to be joined by the zip code. In order to ensure that there were no empty cells, a .dropna() was used.

Feature Selection

The most important feature that allowed the joining of the income and coordinate data was the zip code section. This was especially important as integers are more reliable in joining tables compared to strings.

Therefore for the first dataset (city name and coordinates), the zip code, city name, latitude, and longitude was kept. Then this dataframe was joined with the income dataset by joining by zip codes.

Methodology

Merging

As discussed in the section above, income data and locational data were merged through the zip code variable. Tables 1 and 2 below are the first two datasets of locational data and income data respectively. A lot of the extra variables that are of no help were discarded as discussed in the feature selection section. Both tables 1 and 2 show the first 5 rows of the dataset.

Table 1: Locational dataset of cities within LA County

	Zip	City	State	Latitude	Longitude	Timezone	Daylight savings time flag	geopoint
0	92232	Calexico	CA	33.026203	-115.284581	-8	1	33.026203,-115.284581
1	93227	Goshen	CA	36.357151	-119.425371	-8	1	36.357151,-119.425371
2	93234	Huron	CA	36.209815	-120.084700	-8	1	36.209815,-120.0847
3	93529	June Lake	CA	37.765218	-119.077690	-8	1	37.765218,-119.07769
4	93761	Fresno	CA	36.746375	-119.639658	-8	1	36.746375,-119.639658

Table 2: Dataset of median income per zip code within LA county

	Zip	Community	Estimated Median Income
0	90001	Los Angeles (South Los Angeles), Florence-Graham	35660
1	90002	Los Angeles (Southeast Los Angeles, Watts)	34000
2	90003	Los Angeles (South Los Angeles, Southeast Los ...	34397
3	90004	Los Angeles (Hancock Park, Rampart Village, Vi...	46581
4	90005	Los Angeles (Hancock Park, Koreatown, Wilshire...	32461

Table 3 on the next page shows a portion of the final dataset that's been merged and cleaned from tables 1 and 2.

Table 3: Merged table

	Zip	City	Latitude	Longitude	Income
7	90038	Los Angeles	34.089459	-118.32850	36996.0
8	90063	Los Angeles	34.045161	-118.18650	44121.0
23	90301	Inglewood	33.955913	-118.35868	42100.0
60	90220	Compton	33.890566	-118.23666	54014.0
63	91302	Calabasas	34.133513	-118.66464	122967.0

K Means Clustering

The aim was to cluster these areas using three major variables: latitude, longitude, and income. Therefore K means clustering was performed on a new temporary dataset that only had these three variables as seen in table 4 below. Parameters of the k means clustering process were chosen as follows: 10 clusters and 100 iterations.

Table 4: Temporary dataset created for K means clustering

	Latitude	Longitude	Income
7	34.089459	-118.32850	36996.0
8	34.045161	-118.18650	44121.0
23	33.955913	-118.35868	42100.0
60	33.890566	-118.23666	54014.0
63	34.133513	-118.66464	122967.0

There was one problem that arose with this produced dataset of 10 cluster center points. These had to be related back into a city name data to ensure that the foursquared call was able to process its name to return the venues at each city. In order to convert the produced latitudes and longitudes of the center points into address data, reverse geocoding had to be used.

Reverse Geocoding

The principle behind reverse geocoding is quite simple, entering a valid address will spit out the coordinates, and vice versa. Therefore the two columns of latitudinal and longitudinal data were entered into the reverse geocode function to return an address for that exact point in the map. Results are outputted as an array, which was divided and entered into a new dataframe seen in table 5 below.

Table 5: Center point addresses

	0	1	2	3	4	5	6	7	8
0	Solano Avenue	Elysian Park	Los Angeles	Los Angeles County	California	90012	United States of America	None	None
1	1545	Palisades Circle	Pacific Palisades	Los Angeles	Los Angeles County	California	90272	United States of America	None
2	I-10 Metro ExpressLanes	Brooklyn Heights	Boyle Heights	Los Angeles	Los Angeles County	California	90033	United States of America	None
3	267	Avenue 33	Lincoln Heights	Los Angeles	Los Angeles County	California	90031	United States of America	None
4	4343	Don Diablo Drive	Baldwin Hills/Crenshaw	Los Angeles	Los Angeles County	California	90008	United States of America	None
5	841	Solano Avenue	Elysian Park	Los Angeles	Los Angeles County	California	90012	United States of America	None
6	Los Angeles Academy Middle School	644	East 56th Street	South Park	Los Angeles	Los Angeles County	California	90011	United States of America
7	2842	Eva Terrace	Lincoln Heights	Los Angeles	Los Angeles County	California	90031	United States of America	None
8	7572	Mulholland Drive	Hollywood Hills West	Los Angeles	Los Angeles County	California	90046	United States of America	None
9	960	North Kenter Avenue	Westgate Heights	Brentwood	Los Angeles	Los Angeles County	California	90049	United States of America

As can be seen above, the dataframe has a slight problem of points not having the equal strength in specificity, therefore some columns are empty. However, to my luck, I only needed the city name data, which can be seen between columns 2 and 4.

These data cells needed to be organized in a way such that I could access the city names as a single column. Therefore, another function was created and applied which yielded table 6 seen below.

Table 6: Arranged address data for center points

	0	1	2	3	4	5	6	7	8
0	United States of America	90012	California	Los Angeles County	Los Angeles	Elysian Park	Solano Avenue	NaN	NaN
1	United States of America	90272	California	Los Angeles County	Los Angeles	Pacific Palisades	Palisades Circle	1545	NaN
2	United States of America	90033	California	Los Angeles County	Los Angeles	Boyle Heights	Brooklyn Heights	I-10 Metro ExpressLanes	NaN
3	United States of America	90031	California	Los Angeles County	Los Angeles	Lincoln Heights	Avenue 33	267	NaN
4	United States of America	90008	California	Los Angeles County	Los Angeles	Baldwin Hills/Crenshaw	Don Diablo Drive	4343	NaN

Columns 1 and 5 of table 6 was extracted and merged onto the first merged data according to the zip codes. This final dataset was then ready to be put through the foursquared call as seen in table 7 below.

Table 7: Final processed center points

	Zip	Latitude	Longitude	Income	Neighborhood
13	90011	34.007063	-118.25868	33824.0	South Park
14	90008	34.009754	-118.33705	36641.0	Baldwin Hills/Crenshaw
16	90033	34.050411	-118.21195	31683.0	Boyle Heights
36	90049	34.067409	-118.47528	121671.0	Brentwood
145	90012	34.061611	-118.23944	38786.0	Elysian Park
146	90012	34.061611	-118.23944	38786.0	Elysian Park
154	90046	34.098908	-118.36241	65990.0	Hollywood Hills West
181	90272	34.050505	-118.53374	180962.0	Pacific Palisades
235	90031	34.078710	-118.21610	41126.0	Lincoln Heights
236	90031	34.078710	-118.21610	41126.0	Lincoln Heights

FourSquared

The standard code seen in the lab section to retrieve and analyze venues were used as the produced table from reverse geocoding had the necessary parameters. This then returned 101 unique categories, which was onehot encoded, then a frequency table was produced.

Results

The final result, ordered by median income, is seen in table 8 below. This table is essentially the final product of the code as it is able to answer the primary question that I had created myself.

Table 8: Income and most popular venue

	Income	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
7	180962.0	Pacific Palisades	Gym	Deli / Bodega	Trail	Scenic Lookout	Theater	High School	State / Provincial Park	Wings Joint	Flea Market	Filipino Restaurant
3	121671.0	Brentwood	Home Service	Historic Site	Campground	Wings Joint	Frozen Yogurt Shop	Discount Store	Donut Shop	Fast Food Restaurant	Filipino Restaurant	Flea Market
6	65990.0	Hollywood Hills West	Coffee Shop	Pharmacy	Bank	Sushi Restaurant	Movie Theater	Gym / Fitness Center	Liquor Store	Fast Food Restaurant	Grocery Store	Nail Salon
8	41126.0	Lincoln Heights	Taco Place	Grocery Store	Mexican Restaurant	Pharmacy	Thrift / Vintage Store	Seafood Restaurant	Fast Food Restaurant	Bank	Frozen Yogurt Shop	Discount Store
4	38786.0	Elysian Park	Chinese Restaurant	Mexican Restaurant	Vietnamese Restaurant	Sandwich Place	Bakery	Plaza	Seafood Restaurant	Café	Tea Room	Bar
1	36641.0	Baldwin Hills/Crenshaw	Fast Food Restaurant	Sandwich Place	Mexican Restaurant	Lingerie Store	Department Store	Southern / Soul Food Restaurant	Wings Joint	Hardware Store	Paper / Office Supplies Store	New American Restaurant
0	33824.0	South Park	Fast Food Restaurant	Ice Cream Shop	Discount Store	Pizza Place	Mexican Restaurant	Fried Chicken Joint	Wings Joint	Frozen Yogurt Shop	Donut Shop	Filipino Restaurant
2	31683.0	Boyle Heights	Mexican Restaurant	Taco Place	Pharmacy	Sandwich Place	Seafood Restaurant	Coffee Shop	Burger Joint	Pizza Place	Thai Restaurant	Gym / Fitness Center

Conclusion

As can be seen in table 8, richer neighborhoods such as pacific palisades (which is located slightly north of long beach) had most popular venues which relates to activities. As the income level gets lower and lower, there are more restaurants, namely Chinese, Mexican, and fast food restaurants that become more and more popular. Therefore, I would recommend that a prospective business owner refer to this table to decide which venue to open at which location.