Capstone Project - The Battle of Neighborhoods

Coursera

William Ho Swee Chiong

July 30 2021

Introduction:

Los Angeles, one of the greatest cities in the United States of America and even in the world. LA is known for Hollywood and also the sunny weather. LA is also a diverse city where the resident in Los Angeles is made up of every races and ethnics, which also make this city shine. However, LA is a huge cities which has a 500 mi^2 of total land area. As comparison, Manhattan Island only has 33 mi^2 of land area. Therefore, every different part of LA will show different characteristics.

Our client MR. Wang is a new immigrant to the United States. Here's is some background of our client: He's a member of middle class income group in China. He owned a Chinese canteen in China before he and his family migrate to the US. He wish to start a business in Los Angeles city. However, he doesn't know much about this cities. Therefore, he seek for business advice to find the best location for him to start the business (open a restaurant) .

Due to there is some constraint due to his budge, he will not be afford a overbudget property listing. Therefore, he will be looking a house that is both affordable and also be convenience.

Our goal

- Using the dataset and webpage available online to build the model to predict the area
- Using foursquare API to cluster the neighborhood that consist of more Asian restaurant and Asian grocery store.

Dataset Chosen

- https://data.lacounty.gov/GIS-Data/ZIP-Codes-and-Postal-Cities/c3xr-3jw2/data – This website consist of a list of LA neighborhood ,it's postal code and also coordinate
- A set of data from Zillow which show every single US post code with it respective average housing price
- FourSquare API – Using this API also us to access the venue around the location

Using this LA neighborhood dataset to help us to find the solution for our client MR. Wang to found a neighborhood that is most suitable for him to settle down and start his business to ensure him to have a better life. We will perform use some data science skill to process the data such as Data cleaning, data wrangling and also map visualization by using Folium. Next, we will use K-means clustering to perform the machine learning computation.

Methodology

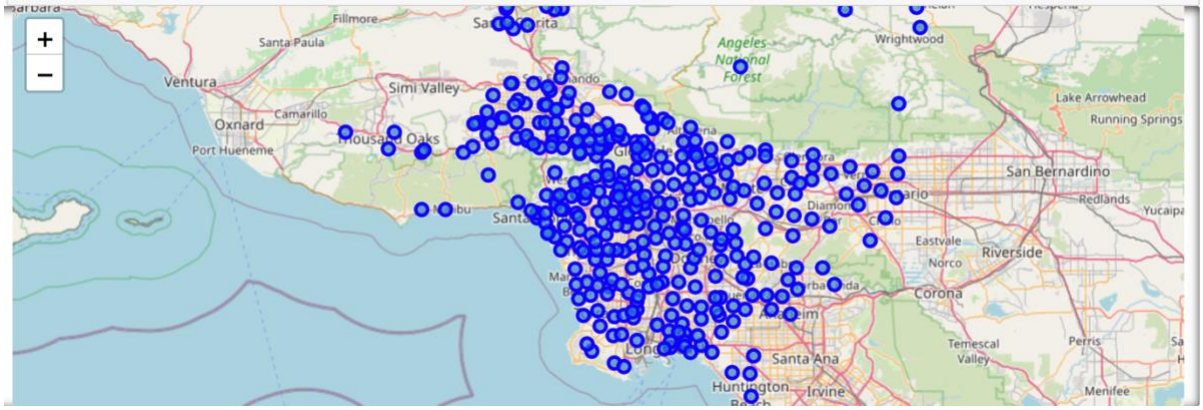
- Folium is used to visualize the map and set the point for every neighborhood.
- SK learn will be used to create the model to cluster the neighborhood by the nearest venue.
- Pandas is used to visualize the dataframe.
- Numpy is used to do some calculation for the array.

- Nominatim from Geopy.geocoder is used to find out the latitude and longitude for los angeles'

Result:
The dataframe that after the data processing:
The dataset included ZIP Code to identify the cities cause there is some of the city's name that is duplicated.

	ZIP Code	Postal City 1	Latitude	Longitude
0	90713	Lakewood	33.84871142900005	-118.11357922799999
1	91306	Winnetka	34.208404020000046	-118.57593995299999
2	90002	Los Angeles	33.94895070600006	-118.24697958699994
3	90506	Torrance	33.88535286100006	-118.32659746799999
4	90069	West Hollywood	34.08940300900008	-118.37978902499998
...	...	...	...	...
365	90011	Los Angeles	34.007903741000064	-118.259036977
366	90247	Gardena	33.89189209400007	-118.29849831699994
367	90601	Whittier	33.995472680000034	-118.04046581999995
368	90630	Cypress	33.82014301800007	-118.03980972399995
369	91759	Mt Baldy	34.237141886000074	-117.65801843599996

The following picture show that the map of los angeles with the marker that pointed out the neighbor in my dataset:



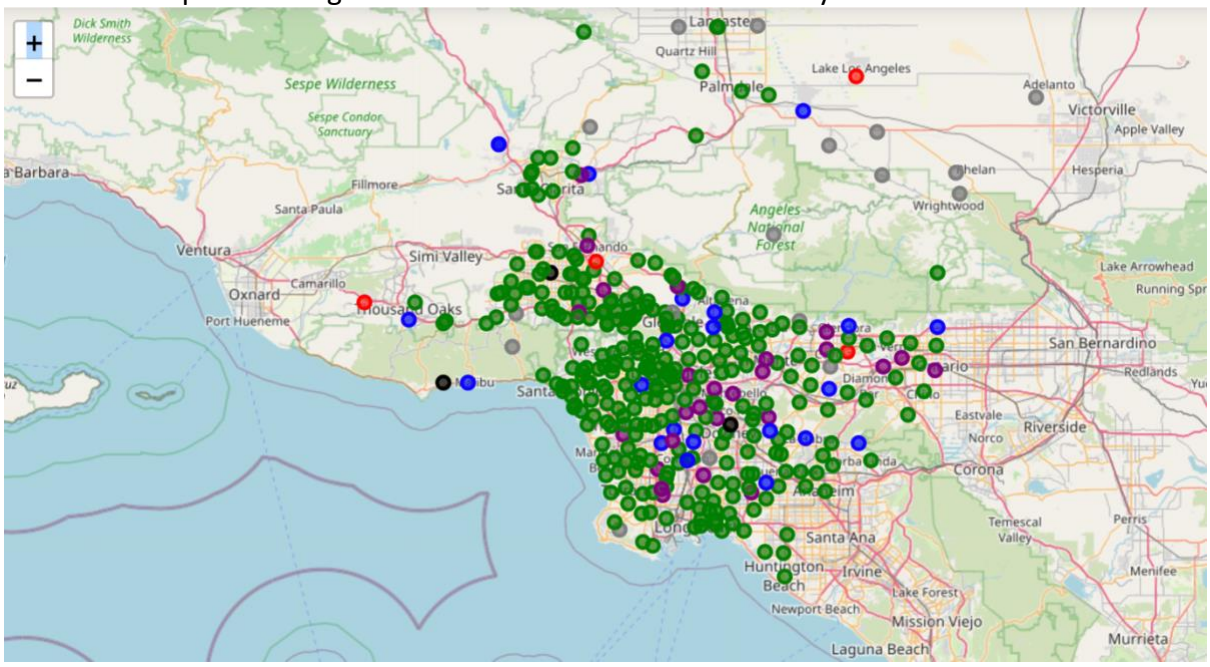
The following table show the top 10 most common venue in each neighborhood.

	Zip Code	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
0	90001	Grocery Store	Shoe Store	Burger Joint	Pizza Place	Sandwich Place	Fast Food Restaurant	Mexican Restaurant	Donut Shop	Pharmacy	ATM
1	90002	Park	Pharmacy	ATM	Noodle House	Outdoor Sculpture	Outdoor Gym	Other Repair Shop	Other Great Outdoors	Organic Grocery	Optical Shop
2	90003	Bakery	Southern / Soul Food Restaurant	Pizza Place	Fast Food Restaurant	Restaurant	Taco Place	ATM	Noodle House	Outdoor Gym	Other Repair Shop
3	90004	Pizza Place	Cocktail Bar	Mexican Restaurant	Sandwich Place	Sushi Restaurant	Convenience Store	Rental Car Location	Bank	Spa	New American Restaurant
4	90005	Korean Restaurant	Restaurant	Hotel	Ice Cream Shop	Clothing Store	Golf Course	BBQ Joint	Mexican Restaurant	Steakhouse	North Indian Restaurant

The following table show the top 10 most common venue in every neighborhood with the cluster label that calculated by the KMean function

	ZIP Code	Postal City 1	Latitude	Longitude	Cluster Labels	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 M Com Ve
0	90713	Lakewood	33.848711	-118.113579	2.0	Fast Food Restaurant	Smoke Shop	Japanese Restaurant	Video Store	Convenience Store	Discount Store	Coffee Shop	Music Store	Superma
1	91306	Winnetka	34.208404	-118.575940	2.0	Fried Chicken Joint	Bar	Grocery Store	Ice Cream Shop	South American Restaurant	Convenience Store	Latin American Restaurant	Mexican Restaurant	Fil Restau
2	90002	Los Angeles	33.948951	-118.246980	1.0	Park	Pharmacy	ATM	Noodle House	Outdoor Sculpture	Outdoor Gym	Other Repair Shop	Other Great Outdoors	Org Gro
3	90506	Torrance	33.885353	-118.326597	2.0	Fast Food Restaurant	Pizza Place	Mediterranean Restaurant	Mexican Restaurant	Restaurant	Concert Hall	Hookah Bar	Paper / Office Supplies Store	Coffee S
4	90069	West Hollywood	34.089403	-118.379789	2.0	Sushi Restaurant	New American Restaurant	Coffee Shop	Hotel	Salad Place	Gym	Gay Bar	Burger Joint	

This is the map for los angeles and the cluster is differentiate by the color of the marker



Examine the cluster

Cluster 0

	ZIP Code	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
8	91702	Mexican Restaurant	Taco Place	BBQ Joint	ATM	Outdoor Supply Store	Outdoor Sculpture	Outdoor Gym	Other Repair Shop	Other Great Outdoors	Organic Grocery
26	91733	Sporting Goods Shop	Furniture / Home Store	Convenience Store	Construction & Landscaping	Food	Mexican Restaurant	ATM	North Indian Restaurant	Outdoor Supply Store	Outdoor Sculpture
30	90602	Indie Theater	Plaza	Convenience Store	Hotel	Mexican Restaurant	North Indian Restaurant	Outdoors & Recreation	Outdoor Supply Store	Outdoor Sculpture	Outdoor Gym
35	90033	Mexican Restaurant	Fast Food Restaurant	Intersection	Pharmacy	Burger Joint	Taco Place	Thai Restaurant	Bank	Bakery	Seafood Restaurant
41	90304	Mexican Restaurant	Taco Place	Convenience Store	Rental Car Location	Park	Seafood Restaurant	Food	Pizza Place	Latin American Restaurant	Organic Grocery

scroll output; double click to hide

Cluster 1

	ZIP Code	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
2	90002	Park	Pharmacy	ATM	Noodle House	Outdoor Sculpture	Outdoor Gym	Other Repair Shop	Other Great Outdoors	Organic Grocery	Optical Shop
5	91361	Park	Pool	Boat or Ferry	ATM	Noodle House	Outdoor Supply Store	Outdoor Sculpture	Outdoor Gym	Other Repair Shop	Other Great Outdoors
15	90018	Park	Market	Skate Park	ATM	North Indian Restaurant	Outdoor Sculpture	Outdoor Gym	Other Repair Shop	Other Great Outdoors	Organic Grocery
51	91208	Park	Cosmetics Shop	Plaza	Middle Eastern Restaurant	North Indian Restaurant	Outdoors & Recreation	Outdoor Supply Store	Outdoor Sculpture	Outdoor Gym	Other Repair Shop
57	90061	Park	Board Shop	Business Service	ATM	Noodle House	Outdoor Supply Store	Outdoor Sculpture	Outdoor Gym	Other Repair Shop	Other Great Outdoors

Cluster 2

	ZIP Code	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
0	90713	Fast Food Restaurant	Smoke Shop	Japanese Restaurant	Video Store	Convenience Store	Discount Store	Coffee Shop	Music Store	Supermarket	Outdoor Gym
1	91306	Fried Chicken Joint	Bar	Grocery Store	Ice Cream Shop	South American Restaurant	Convenience Store	Latin American Restaurant	Mexican Restaurant	Filipino Restaurant	Office
3	90506	Fast Food Restaurant	Pizza Place	Mediterranean Restaurant	Mexican Restaurant	Restaurant	Concert Hall	Hookah Bar	Paper / Office Supplies Store	Coffee Shop	Sushi Restaurant
4	90069	Sushi Restaurant	New American Restaurant	Coffee Shop	Hotel	Salad Place	Gym	Gay Bar	Burger Joint	Café	Sandwich Place
6	90064	Japanese Restaurant	Comic Shop	Pizza Place	Chinese Restaurant	Bar	Salon / Barbershop	Mexican Restaurant	Pet Store	Miscellaneous Shop	Mobile Phone Shop

Cluster 3

	ZIP Code	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
23	91724	Construction & Landscaping	ATM	Noodle House	Outdoors & Recreation	Outdoor Supply Store	Outdoor Sculpture	Outdoor Gym	Other Repair Shop	Other Great Outdoors	Organic Grocery
20	91320	Construction & Landscaping	ATM	Noodle House	Outdoors & Recreation	Outdoor Supply Store	Outdoor Sculpture	Outdoor Gym	Other Repair Shop	Other Great Outdoors	Organic Grocery
20	91331	Shop & Service	Construction & Landscaping	ATM	Noodle House	Outdoor Supply Store	Outdoor Sculpture	Outdoor Gym	Other Repair Shop	Other Great Outdoors	Organic Grocery
57	93591	Construction & Landscaping	Garden Center	ATM	Noodle House	Outdoor Supply Store	Outdoor Sculpture	Outdoor Gym	Other Repair Shop	Other Great Outdoors	Organic Grocery

Cluster 4

	ZIP Code	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
76	91325	Home Service	Flower Shop	Non-Profit	Outdoor Supply Store	Outdoor Sculpture	Outdoor Gym	Other Repair Shop	Other Great Outdoors	Organic Grocery	Optical Shop
174	90240	Home Service	ATM	Non-Profit	Outdoor Supply Store	Outdoor Sculpture	Outdoor Gym	Other Repair Shop	Other Great Outdoors	Organic Grocery	Optical Shop
276	90265	Home Service	ATM	Non-Profit	Outdoor Supply Store	Outdoor Sculpture	Outdoor Gym	Other Repair Shop	Other Great Outdoors	Organic Grocery	Optical Shop

Cluster 5 is return an empty cluster

Los Angeles House Price Data

This is the dataframe for the average housing price in every single zip code in United States

	RegionName	State	City	2021-06-30
0	10025	NY	New York	1209625.0
1	60657	IL	Chicago	426430.0
2	10023	NY	New York	1752665.0
3	77494	TX	Katy	302597.0
4	60614	IL	Chicago	480745.0
...	...	...	...	...
27556	4109	ME	Portland	534643.0
27557	21405	MD	Annapolis	839754.0
27558	10118	NY	New York	2725221.0
27559	86343	AZ	Crown King	180944.0
27560	89155	NV	Las Vegas	229348.0

For the data processing part, I removed the other states and othe keep the row with the state of California

	RegionName	State	City	2021-06-30
2945	90001	CA	Florence-Graham	465037.0
3465	90002	CA	Los Angeles	460436.0
1319	90003	CA	Los Angeles	483458.0
253	90004	CA	Los Angeles	935004.0
1306	90005	CA	Los Angeles	730415.0
...	...	...	...	...
11966	96145	CA	Tahoe City	736177.0
17836	96146	CA	Tahoe City	748273.0
22643	96148	CA	Tahoe Vista	611146.0
5811	96150	CA	South Lake Tahoe	465142.0
6996	96161	CA	Truckee	664957.0

After the calculation of the data and the comparison of the zip code in LA neighborhood and the housing price data, I found out with a data that listed the cluster with its average housing price:

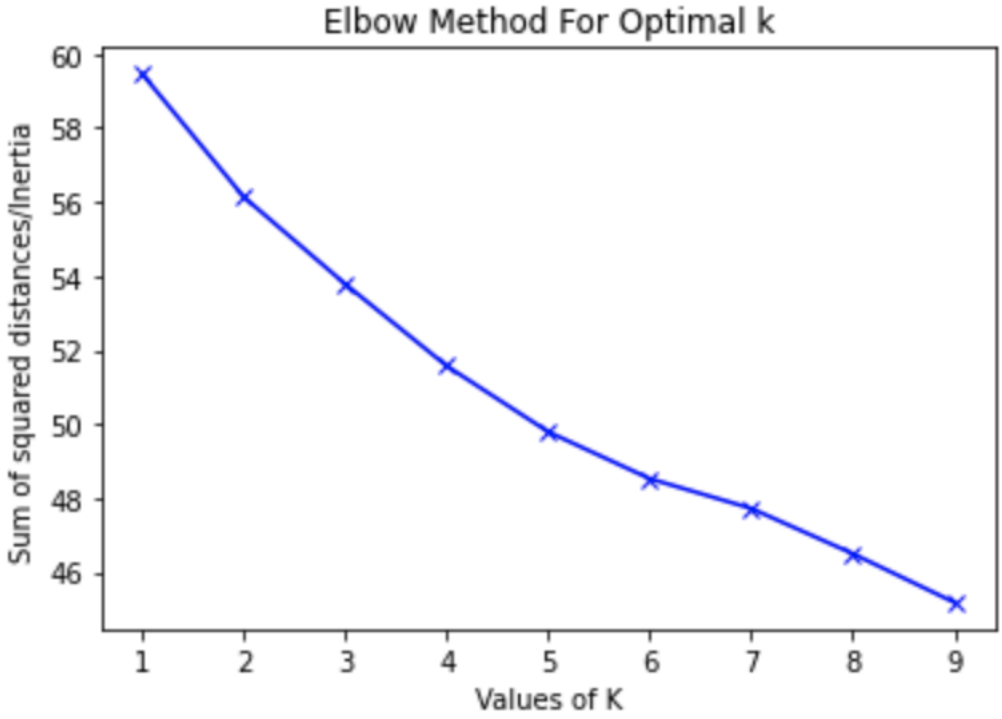
	Cluster	Housing Mean Price
0	0	524636.761905
1	1	637202.368421
2	2	741686.772358
3	3	400524.750000
4	4	918945.000000

Discussion

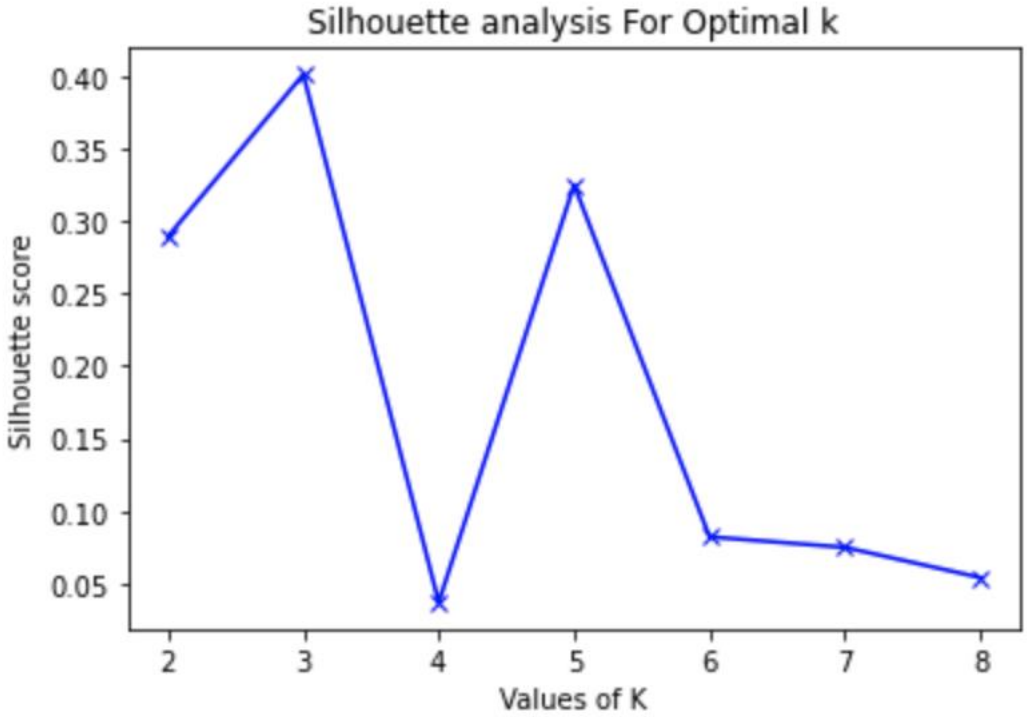
1) The data is processed by convert it to onehot data to allow the clustering method because it required a numeric properties to perform the calculation

	Zip Code	ATM	Accessories Store	Advertising Agency	Afghan Restaurant	African Restaurant	Airport	Airport Service	Airport Terminal	American Restaurant	...	Water Park	Weight Loss Center	Whisky Bar	Wine Bar	Wine Shop	Winery	Wing Joi
0	90713	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
1	90713	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
2	90713	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
3	90713	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
4	90713	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	

2) In order to find the best n cluster value, we need to use elbow method and also Silhouette method.
   - For Elbow method the best n-cluster value is 5 or 6



- For Silhouette method the best k value is 3 and 5



- Therefore the k value chosen here will be 5

Conclusion

For cluster 0, I believed that it's a Hispanic neighborhood because there are more Mexican restaurant and taco place. For cluster 1, I assumed that it is a cluster that consist of more Mediterranean and Indian ethnic. For cluster 2, I think is the most suitable cluster for the client to reside in. This is because it is a more diverse cluster, it consists of more Japanese restaurant, some Filipino restaurant and some entertainments are. For cluster 3, it consists of more construction company. Lastly, cluster 4 is a more high-end cluster that consist more outdoor activities shops and organic groceries.

When compared to the average housing price, cluster 2 will be the second most expensive cluster among all others.

Therefore, I assumed that cluster 1 and 2 will be nice for the client to start business and reside in. For cluster 2, it is very competitive for client to start a business. But on the other hand, it is also having the best market to start the business. For cluster 1, its highlight points are the housing price is more affordable and it is less competitive for client to start his business and ensure the business will be smoother.

In conclusion, I will recommend the client to choose cluster 1 for his first choice to start his business and more affordable housing price. However, if the client is able to afford the high housing price in neighborhoods that fall in cluster 2, we will also strongly recommend neighborhood in cluster 2 because living in cluster 2 will be more convenient and more accessible to some Asian immigrants.