1. Introduction

"L.A. is a constellation of microclimates and microcosms, a library with dozens of special collections," wrote Meghan Daum. " Lying on the *Southern California Basin*, Los Angeles is the biggest and one of the most popular cities on the west coast of the United States. It is the entertainment capital of the world, a cultural mecca having numerous world-class museums and a paradise of sunshine. From tourist spots like the Universal Studio to those job opportunities presented in various industry, Los Angeles is the place to be.

LA is a rich city and the food culture here puts all of it on display. Thanks to the unique natural environment, the golden sunshine, and multicultural demographic, LA has become one of the best places for you to eat in the country. No matter it's middle east hummus, Mexican birria tacos, or Chinese dim sum, LA had anything and everything for you. The culinary possibility here is only beyond your imagination.

With the flourish of the food industry, it might be a good choice to open up a new restaurant in the LA area. If you have thought about it and are interested in it, this project is the right thing for you. This project is the final delivery of the IBM Data Science Professional Certificate, and its objective is to find out the best potential neighborhood to open up a new restaurant in LA. In this article, I will go through the introduction, methodology, analysis, results, and conclusion section by section, and the detailed code and report can be found at the end of this article.

2. Data Preparation

The following data will be necessary for this project and the reason for each one will be explained later:

- List of all neighbourhoods in LA and their coordinates — https://usc.data.socrata.com/Los-Angeles/Los-Angeles-Neighborhoods/xegr-9bnh
- Venue information(restaurant and their category, coordinates within each neighborhood) — Foursquare APIs
- Population, population density, population growth, median rent, median household income — https://www.niche.com/
- LA crime data — https://www.areavibes.com/los+angeles-ca/crime/

3. Methodology

3.1 Data pre-processing

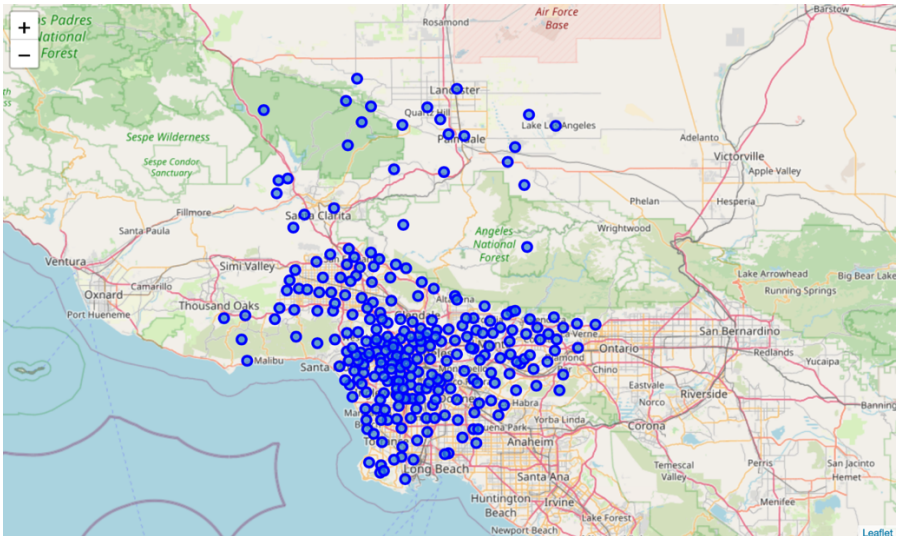
A dataframe of all the LA neighborhoods and their coodidates is obtained by using Pandas libiary. The data source is listed

	set	slug	the_geom	kind	external_i	name	display_na	sqmi	type	name_1	slug_1
0	L.A. County Neighborhoods (Current)	acton	MULTIPOLYGON (((-118.20261747920541 34.5389897...	L.A. County Neighborhood (Current)	acton	Acton	Acton L.A. County Neighborhood (Current)	39.33910895	unincorporated-area	NaN	NaN
1	L.A. County Neighborhoods (Current)	adams-normandie	MULTIPOLYGON (((-118.30900800000012 34.0374109...	L.A. County Neighborhood (Current)	adams-normandie	Adams-Normandie	Adams-Normandie L.A. County Neighborhood (Curr...	0.805350188	segment-of-a-city	NaN	NaN
2	L.A. County Neighborhoods (Current)	agoura-hills	MULTIPOLYGON (((-118.76192500000009 34.1682029...	L.A. County Neighborhood (Current)	agoura-hills	Agoura Hills	Agoura Hills L.A. County Neighborhood (Current)	8.146760298	standalone-city	NaN	NaN
3	L.A. County Neighborhoods (Current)	agua-dulce	MULTIPOLYGON (((-118.2546773959221 34.55830403...	L.A. County Neighborhood (Current)	agua-dulce	Agua Dulce	Agua Dulce L.A. County Neighborhood (Current)	31.46263195	unincorporated-area	NaN	NaN
4	L.A. County Neighborhoods (Current)	alhambra	MULTIPOLYGON (((-118.12174700000014 34.1050399...	L.A. County Neighborhood (Current)	alhambra	Alhambra	Alhambra L.A. County Neighborhood (Current)	7.623814306	standalone-city	NaN	NaN

The dataframe is cleaned as below:

	Neighborhood	Latitude	Longitude
0	Acton	34.497355	-118.1698102
1	Adams-Normandie	34.031461	-118.300208
2	Agoura Hills	34.146736	-118.7598845
3	Agua Dulce	34.504927	-118.3171037
4	Alhambra	34.085539	-118.136512

Then a folium map is created with superimposed makers on it. Each maker represents each neighborhood of Los Angeles. It seems like there are a lot of things we get.



3.2 Exploratary Data Analysis

3.2.1 EDA of Ktown

Before we move forward, let's take some time to explore the Koreatown which is the neighborhood where I currently live in. All the venues information includes the name, category, latitude and longitude is obtained by using the Foursquare API in this case.

First, let's take a look at what's really inside Ktown.

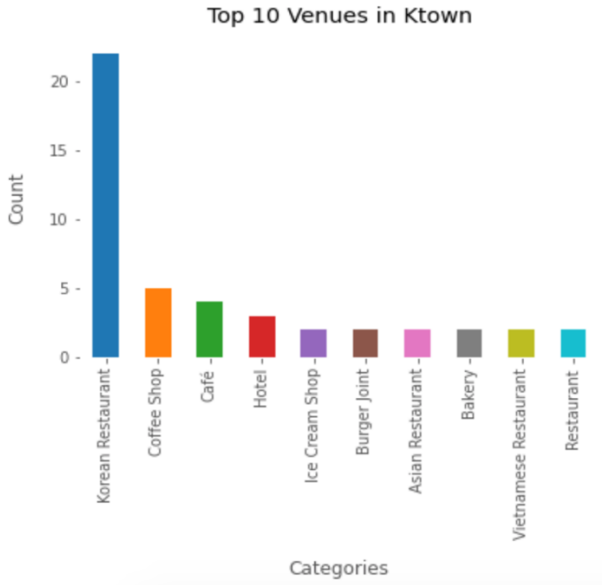
	name	categories	lat	lng
0	Ahgassi Gopchang	Korean Restaurant	34.063397	-118.303863
1	Kyochon Chicken	Korean Restaurant	34.063830	-118.306490
2	Hae Jang Chon Korean BBQ Restaurant	Korean Restaurant	34.063888	-118.306075
3	Sushi One	Sushi Restaurant	34.063571	-118.308160
4	Byul Yang Gopchang	BBQ Joint	34.063668	-118.305999

Then, let's find out how many venues are within 500 meters radius of Ktown.

```
print('{} venues were returned by Foursquare.'.format(nearby_venues.shape[0]))
```

```
80 venues were returned by Foursquare.
```

Last, let's find out what kind of venue is the most popular one in Ktown.



Unquestionably, korean restaurant is the most popular venue in Ktown.

3.2.1 EDA of each neighborhood

As we are searching for the best neighborhood to open up a new restaruant, exploring all the neighborhoods a little bit would be a good choice for us to have a glimpse in mind. Let's take a look at each neighborhood and all the venues within that specific neighborhood.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Acton	34.497355	-118.169810	Epik Engineering	34.498718	-118.168046	Construction & Landscaping
1	Acton	34.497355	-118.169810	Alma Gardening Co.	34.494762	-118.172550	Construction & Landscaping
2	Adams-Normandie	34.031461	-118.300208	Orange Door Sushi	34.032485	-118.299368	Sushi Restaurant
3	Adams-Normandie	34.031461	-118.300208	Shell	34.033095	-118.300025	Gas Station
4	Adams-Normandie	34.031461	-118.300208	Little Xian	34.032292	-118.299465	Sushi Restaurant

Next, let's check out how many venues are returned by each neighborhood.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Acton	2	2	2	2	2	2
Adams-Normandie	9	9	9	9	9	9
Agoura Hills	27	27	27	27	27	27
Agua Dulce	1	1	1	1	1	1
Alhambra	14	14	14	14	14	14

Then, let's use one-hot encoding technique to unfold the venue category for each neighborhood, group them by neighborhood and take the mean of each category's frequency of occurrence.

	Neighborhood	ATM	Accessories Store	Airport	Airport Terminal	American Restaurant	Andhra Restaurant	Antique Shop	Arcade	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Arts & Entertainment	A Restau
0	Acton	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1	Adams-Normandie	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
2	Agoura Hills	0.0	0.0	0.0	0.0	0.037037	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
3	Agua Dulce	0.0	0.0	1.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
4	Alhambra	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

Last, output each neighborhood along with top 5 most popular venues.

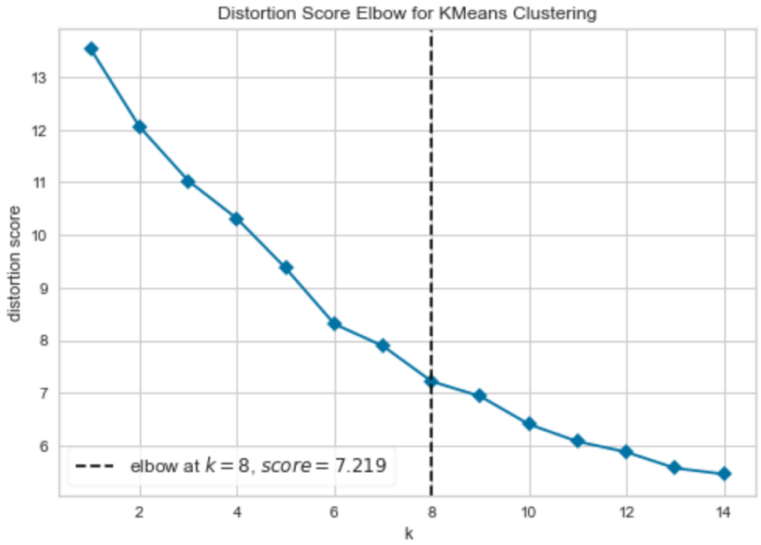
```
----Adams-Normandie----
                venue  freq
0  Sushi Restaurant  0.33
1       Taco Place  0.11
2      Gas Station  0.11
3     Grocery Store  0.11
4      Home Service  0.11


----Agoura Hills----
                  venue  freq
0  Fast Food Restaurant  0.15
1    Chinese Restaurant  0.07
2        Breakfast Spot  0.07
3          Burger Joint  0.04
4    Mexican Restaurant  0.04
```

3.3 Clustering Analysis

Before we can determine which neighborhood is the one we are looking for, we can use clustering analysis to narrow down our search by clustering them into groups. In our case, we use K-Means clustering technique to do that, and we cluster each neighborhood based on their venue similarities.

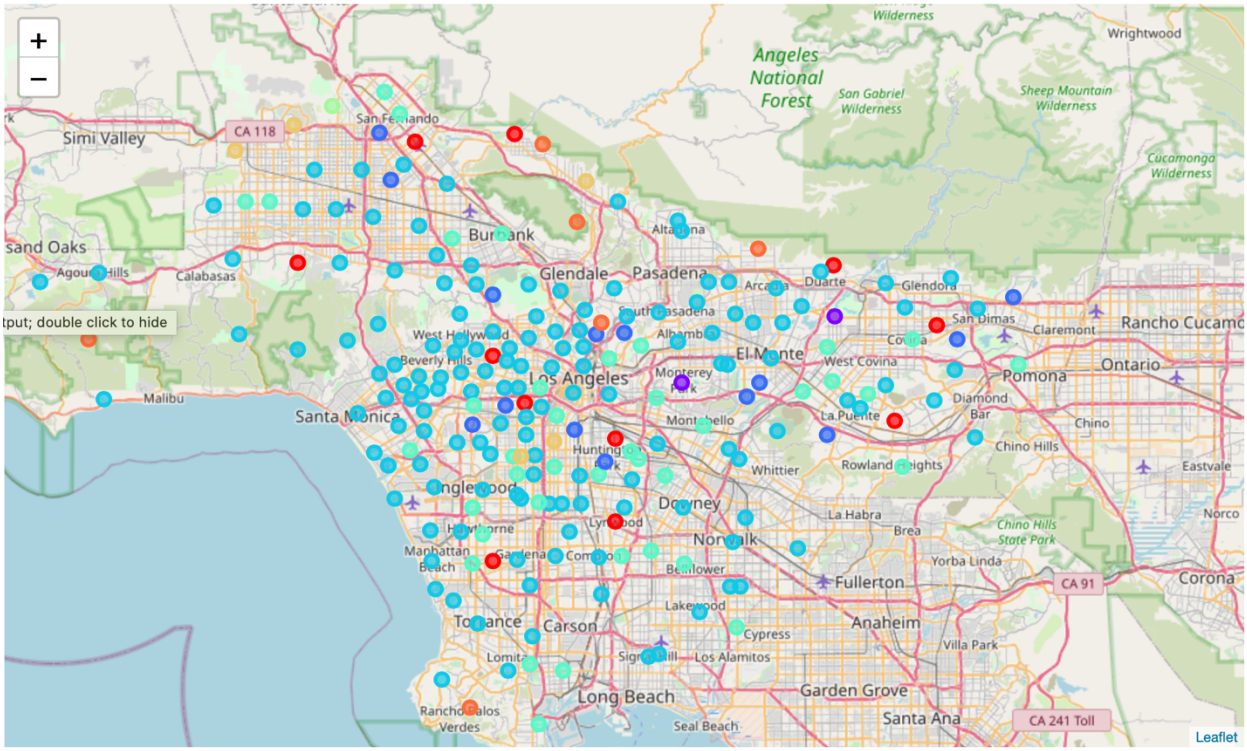
The first step in clustering the neighborhood is to determine the optimal value of K for the dataset. This is carried out by using the Elbow method. The elbow figure below show us that the distortion score is lowest when we set K equal to 8. So, as a result, we should group all the neighborhoods into 8 clusters.



Then, a new dataframe which include the corresponding cluster label and the top 10 venues of each neighborhood is created as below.

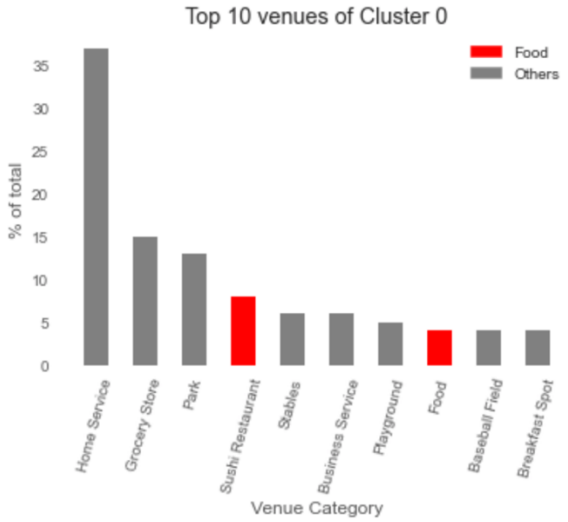
	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Acton	34.497355	-118.169810	5	Construction & Landscaping	Yoga Studio	Fast Food Restaurant	English Restaurant	Escape Room	Ethiopian Restaurant	Fabric Shop	Falafel Restaurant	Farm	Farmers Market
1	Adams-Normandie	34.031461	-118.300208	0	Sushi Restaurant	Park	Home Service	Gas Station	Playground	Grocery Store	Taco Place	Falafel Restaurant	Empanada Restaurant	English Restaurant
2	Agoura Hills	34.146736	-118.759884	3	Fast Food Restaurant	Chinese Restaurant	Breakfast Spot	Restaurant	Thai Restaurant	Liquor Store	Lounge	Sushi Restaurant	Bakery	Shipping Store
3	Agua Dulce	34.504927	-118.317104	3	Airport	Yoga Studio	Filipino Restaurant	Escape Room	Ethiopian Restaurant	Fabric Shop	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant
4	Alhambra	34.085539	-118.136512	3	Convenience Store	Bagel Shop	Sporting Goods Shop	Fast Food Restaurant	Pet Store	Hardware Store	Video Store	Mexican Restaurant	Business Service	Pizza Place

Meanwhile, we use another folium map with differently colored makers on it to visulize those neighborhood. Every color in this case represent every cluster. Same color means same cluster, different color means different cluster.



Then, several bar charts are generated to show the top 10 venues in each neighborhood, and with food venue highlighted.

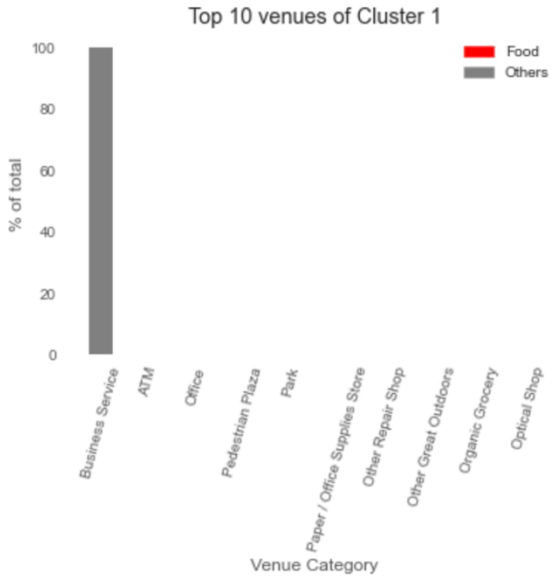
Cluster 0:



As we can see, there are only 2 food venues within the top 10 of cluster 0, and only 1 of them is about “Restaurant” which is “Sushi Restaurant”. Meanwhile, the presence of venues that see a

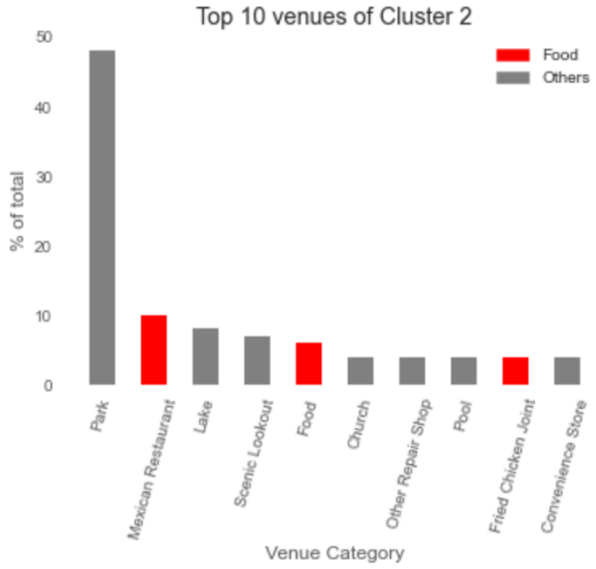
high footfall such as home service and grocery store in the list may further indicates that the population density in these neighborhoods are fairly high. All the obervations point to the direction that cluster 0 being nomiated as the cluster to explore further.

Cluster 1:



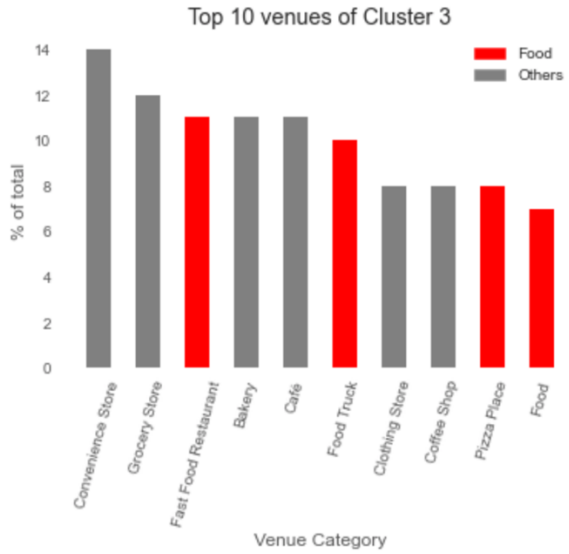
We only have 2 neighborhoods in cluster 1 which are “Irwindale” and “Monterey Park”. From the venue information Foursquare API returned, both of them are not appropriate to open up a restaurant. So, we simply drop cluster 1 right now.

Cluster 2:



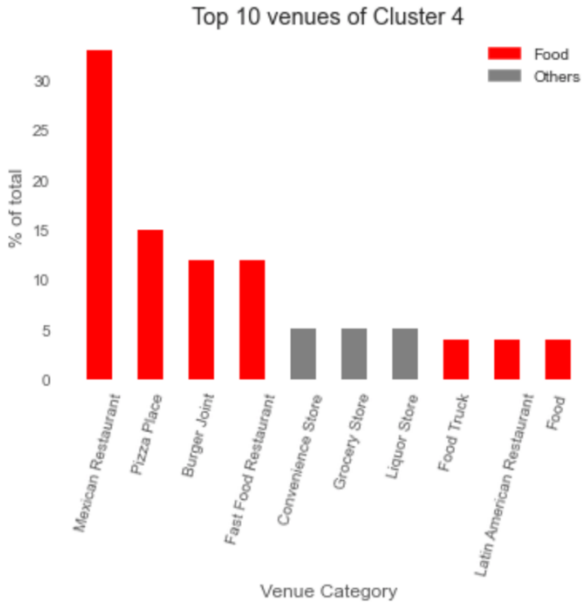
As we can see, there are 3 venues which is food-related in cluster 2. They are “Mexican Restaurant”, “Food” and “Food Chicken joint”. But all of them take less than 10% of total which might suggests a not that furious environment in these area. And the presence of “Lake”, “Park” and “Scenic Lookout” on the list also indicates that those neighborhoods are the destinations that people will choose for fun or for vacation and opening up a restaurant in those area may be a good choice. So, we nominate cluster 2 as the cluster to explore further.

Cluster 3:



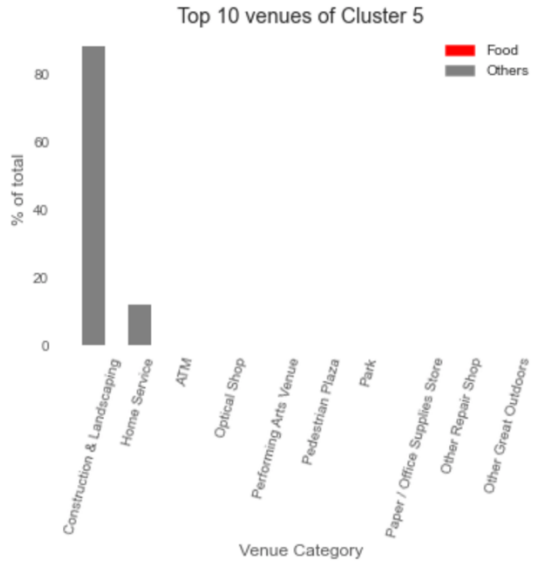
There are 4 food venues within the top 10 of Cluster 3 which indicates that cluster 3 may not be the proper one for setting up a new restaurant.

Cluster 4:



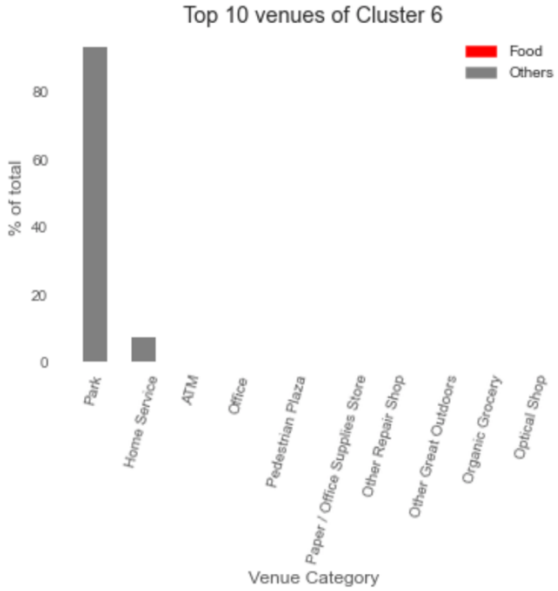
Cluster 4 is even further than cluster 3. There are 7 food venues in cluster 4 with Mexican Reastaurants making up more than 30% of all venues. Which suggests that cluster 4 may not be the one we want to open up a reastaurant in.

Cluster 5:



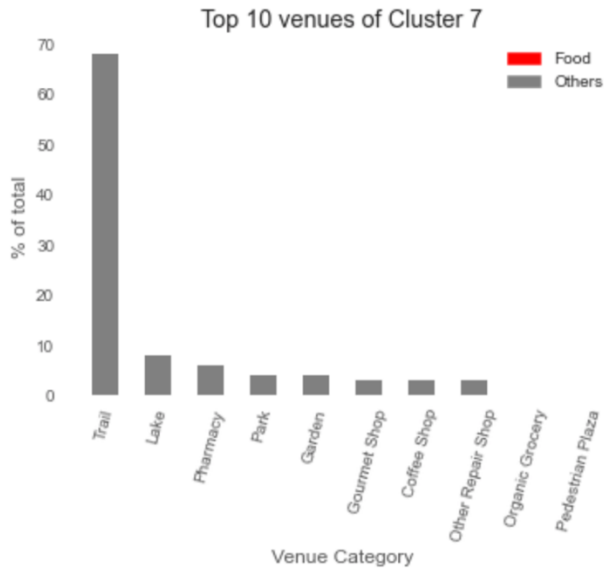
All the neighborhoods in cluster 5 are rural area of Los Angeles, and most of them only have "Construction & Landscaping" venues. Obviously, Cluster 5 is not the one we are looking for.

Cluster 6:



Similar to cluster 5, "Park" consisting of nearly all the venues in cluster 6 which indicates that those neighborhood don't have a high population density, thus not appropriate to open up a restaurant.

Cluster 7:



Cluster 7 also got a lot of natural landscapes on list which is similar to cluster 6. But the differences are the presences of “Pharmacy”, “Coffee Shop” and “Garden” in this case are telling us there is population actually living in those neighborhoods of cluster 7. And more importantly, there are no food venues on the top 10 venues list of cluster 7 which suggests a loose environment and even a bright future for opening up a new reastaurants in those neighborhoods. So, we nominate cluster 7 for us to explore further.

3.4 Candidate Neighborhood

Based on our discussion, the proper neighborhood for starting a new reastaurant would only come from Cluster 0, Cluster 2 and Cluster 7. So, a list of all the neighborhoods from those clusters are created. Those neighborhoods are our candidates that we want to explore further.

	Neighborhood
0	Adams-Normandie
1	Alondra Park
2	Green Valley
3	Bradbury
4	Charter Oak
5	Hancock Park
6	Lynwood
7	Pacoima
8	South San Jose Hills
9	Sunland
10	Tarzana
11	Vernon
12	Baldwin Hills/Crenshaw
13	Central-Alameda
14	Cypress Park
15	Hacienda Heights
16	Hollywood Hills
17	Huntington Park

18	Jefferson Park
19	La Verne
20	Mission Hills
21	Montecito Heights
22	Panorama City
23	South El Monte
24	Val Verde
25	West San Dimas
26	Whittier Narrows
27	Glendale
28	Mount Washington
29	Rancho Palos Verdes
30	Sierra Madre
31	Tujunga

3.5 Best Neighborhood

We then start to examine each of our candidate neighborhood one by one. We use U.S. Census Bureau data and data on Niche website to calculate the score for each neighborhood based on the demand for new restaurants and local business cost. Because these are the 2 most imprtant factors which could affect the success of newly opened restaurants.

To analyze the demand, we look at the *population*, *population density*, *population growth* and *median household income*. To access the local business cost, we look at the *crime rate* and *median rent price*.

So, a new dataframe with all the candidates and their corresponding score is generated as below.

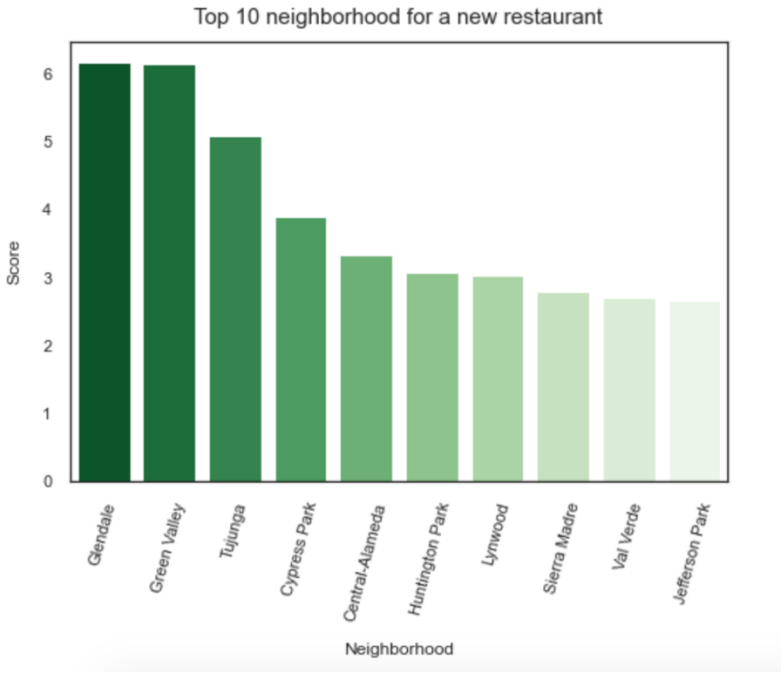
	Neighborhood	population	population_density	population_growth	Median_Household_Income	Median_Rent	Crime_rate
0	Adams-Normandie	18504	21948.0	-7.00	44190	1080.0	C
1	Alondra Park	8097	7518.4	8.82	63368	1144.0	D
2	Green Valley	1572	80.2	19.60	111884	1280.0	A
3	Bradbury	916	562.0	-7.66	154000	3501.0	B
4	Charter Oak	9760	10034.0	2.94	75371	1754.0	D

We clean the dataframe, transform the categorical data to numerical, normalize it and sort the neighborhood by its corresponding score. Then, we get the following dataframe:

	Neighborhood	population	population_density	population_growth	Median_Household_Income	Median_Rent	Crime_rate	score
27	Glendale	1.000000	0.266625	0.530448	0.195707	-0.202588	-0.25	6.160766
2	Green Valley	0.007400	0.002443	1.000000	0.629671	-0.101901	-0.00	6.150455
31	Tujunga	0.139244	1.000000	0.369458	0.226307	-0.213101	-0.25	5.087629
14	Cypress Park	0.238649	0.564903	0.369458	0.098016	-0.048524	-0.25	3.890009
13	Central-Alameda	0.216250	0.786650	0.369458	0.000000	-0.040032	-0.50	3.329303
17	Huntington Park	0.292607	0.803484	0.167278	0.003201	-0.000000	-0.50	3.066283
6	Lynwood	0.354161	0.605500	0.265591	0.082743	-0.053781	-0.50	3.016851
30	Sierra Madre	0.054503	0.152855	0.422230	0.495542	-0.178730	-0.25	2.785600
24	Val Verde	0.014510	0.039467	0.280998	0.395108	-0.052568	-0.00	2.710056
18	Jefferson Park	0.119432	0.683375	0.369458	0.054060	-0.059846	-0.50	2.665914

29	Rancho Palos Verdes	0.210608	0.128908	0.421497	0.817860	-0.666397	-0.25	2.649906
19	La Verne	0.161113	0.158480	0.285326	0.400040	-0.173878	-0.25	2.324329
16	Hollywood Hills	0.112591	0.127666	0.604182	0.521666	-0.292762	-0.50	2.293373
7	Pacoima	0.391388	0.440302	0.281365	0.160456	-0.200970	-0.50	2.290162
28	Mount Washington	0.061993	0.287825	0.326486	0.399451	-0.256369	-0.25	2.277545
0	Adams-Normandie	0.091940	0.920487	0.024211	0.034434	-0.021027	-0.50	2.200181
22	Panorama City	0.340600	0.342233	0.369458	0.092327	-0.150829	-0.50	1.975158
25	West San Dimas	0.169211	0.089887	0.369458	0.391072	-0.282653	-0.25	1.947900
12	Baldwin Hills/Crenshaw	0.151262	0.437615	0.369458	0.043438	-0.079256	-0.50	1.690068
21	Montecito Heights	0.081919	0.352645	0.369458	0.220662	-0.135059	-0.50	1.558503
1	Alondra Park	0.039979	0.314710	0.604549	0.203067	-0.046907	-0.75	1.461593
9	Sunland	0.078424	0.042485	0.369458	0.345734	-0.272139	-0.25	1.255853
20	Mission Hills	0.095820	0.120437	0.284666	0.305489	-0.289527	-0.25	1.067538
26	Whittier Narrows	0.112511	0.243350	0.369458	0.078979	-0.344925	-0.25	0.837495
23	South El Monte	0.103040	0.306759	0.283654	0.076236	-0.092196	-0.50	0.709970
15	Hacienda Heights	0.274014	0.201948	0.336023	0.386411	-0.309341	-0.75	0.556222
10	Tarzana	0.181055	0.168598	0.292333	0.379834	-0.383744	-0.50	0.552301
5	Hancock Park	0.050998	0.270235	0.060895	0.674516	-0.432673	-0.50	0.495885
4	Charter Oak	0.048282	0.420319	0.388848	0.308610	-0.293571	-0.75	0.489955
11	Vernon	0.000000	0.000000	0.955979	0.261383	-0.222037	-1.00	-0.018699
8	South San Jose Hills	0.102371	0.571234	0.311519	0.211280	-0.207845	-1.00	-0.045765
3	Bradbury	0.004124	0.022670	0.000000	1.000000	-1.000000	-0.25	-0.892823

At this point, we can easily plot out the top 10 neighborhood for opening up a new restaurant based on their corresponding score.



4. Results & discussions

From the figure above we can see that 3 neighborhoods stand out among all others. They are Glendale, Green Valley, and Tujunga which is No.1, No.2, and No.3 on the list respectively. So, if I am planning to open up a new restaurant in the Los Angeles area, those 3 neighborhoods would be my first consideration.

The result of this analysis highlighted the potential neighborhoods where a newly opened restaurant might be favorable. But it solely from the geographical perspective, thus can only be served as a start point in the overall investigation. There are so many other factors and things that should be considered before opening up a new restaurant, such as your customer base, the availability of commercial space, the labor cost, ingredient cost, access to public transportation, and so on.

Besides, the methodology of this analysis could be improved in several ways. First, when we clustered the neighborhoods, we clustered them solely based on their venue information but completely ignored other factors such as the demography, traffic condition, and acreage, which could affect the veracity of the clustering result. Second, when we accessed the score of each neighborhood, we could have to get more industrial or professional data such as payroll costs and growth in labor which may be more useful for our analysis, but those data are kind hard to get on the neighborhood level.

5. Conclusion

As a data analyst, I have been always believing in the power of data and convincing myself that there are many real-life problems and scenarios where data can be used to find a solution to it. This project is technically the first project that I've ever finished on my own, and it is the final delivery of Coursera's Applied Data Science Capstone in pursuit of *the IBM Data Science Professional Certificate.* There are a lot of places that can be improved about this analysis, so if you have any ideas, suggestions, or comments, please let me know.