Exploring Restaurants in Los Angeles

1. Introduction

Restaurant business is one of the most sought-after businesses in Los Angeles (LA). LA, in particular, is an amazing place to dine out owing to its wide variety of international cuisines, the quality of food and the services offered. Given the high demand for quality, a restaurant needs to get several factors right in order to survive here. In this brief report, I tried to investigate some of these factors by exploring the restaurants based in the various neighborhoods of LA. I have tried to keep the discussion general when it comes to the kinds of restaurant and their features and rather focus on the global features like density of restaurants in a region, their frequency and so on.

The *business* direction that I have in mind is actually two-fold. One is directed towards individuals looking for places to eat or even people looking for places to rent based on restaurant types or frequency. The other direction is for corporations/individuals looking for an optimal location to open a new restaurant.

1.1 Specific Plan

I plan to use 2 ideas for segregating/clustering the neighborhoods based on their restaurant venues. One idea is to find the density of restaurants in each neighborhood. This will help anyone trying to open a new restaurant by either avoiding overcrowded areas or alternately could help finding popular venues to avoid competition.
The next idea to explore is the kind of restaurant. Here, one can again use clustering but now based on the frequency of occurrence of each restaurant in a Neighborhood just like the one done in the New York data set. Once we find the relevant cluster here, we can then look for its intersection with the clusters found above to fine tune the relevant neighborhood where one wants to open his/her restaurant.

2. Data collection, cleaning and wrangling

I acquired the location data from the website – https://usc.data.socrata.com/dataset/Los-Angeles-Neighborhood-Map/r8qd-yxsr. Also, there is an API endpoint link for the json file in this website which could be directly loaded into the python notebook.

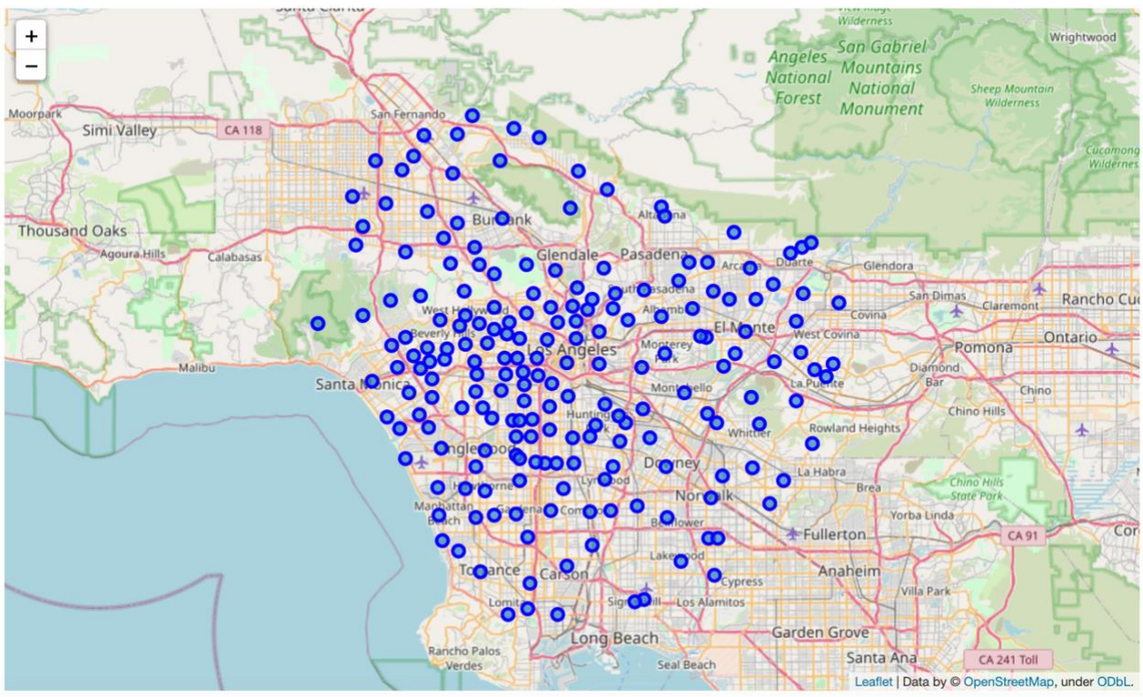
2.1 Data cleaning

Data cleaning was pretty straightforward. Several redundant columns were dropped. Some columns like the 'type' or the geometry of the boundaries ('the_geom') of a neighborhood were also removed. One major problem which I came to realize after plotting the folium map is that latitude and longitude values are swapped with each other. Once that is fixed along with renaming of some columns, the final table looked as shown below:

Table 1:

	Neighbourhood	sqmi	Longitude	Latitude
0	Acton	39.3391089485	-118.16981019229348	34.497355239240846
1	Adams-Normandie	0.805350187789	-118.30020800000011	34.031461499124156
2	Agoura Hills	8.14676029818	-118.75988450000015	34.146736499122795
3	Agua Dulce	31.4626319451	-118.3171036690717	34.504926999796837
4	Alhambra	7.62381430605	-118.13651200000021	34.085538999123571

The data consisted of 272 unique neighborhoods. To make the analysis slightly easier, I used a distance function from the LA central coordinates to reduce the number of rows to 199. This is just so that I can reduce the number of calls to the foursquare location app. The final map with all the neighborhoods is shown below –



The centroid coordinates of LA was found by using the geocoder library in the geopy module of python. Then to explore venues in a given neighborhood, we use Foursquare which was introduced earlier in the course.

2.2 Data Wrangling

Now, we need to transform the data to bring it into a useful form for analysis. Nearby venues for each of the neighborhoods were collected using Foursquare and only the restaurant venues were stored for further analysis.

Table 2:

:

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Adams-Normandie	34.031461	-118.300208	Orange Door Sushi	34.032485	-118.299368	Sushi Restaurant
1	Adams-Normandie	34.031461	-118.300208	Little Xian	34.032292	-118.299465	Sushi Restaurant
2	Adams-Normandie	34.031461	-118.300208	Sushi Delight	34.032501	-118.299454	Sushi Restaurant
3	Alhambra	34.085539	-118.136512	Manny's Tacos	34.087148	-118.135275	Mexican Restaurant
4	Alhambra	34.085539	-118.136512	Wendy's	34.087705	-118.135010	Fast Food Restaurant

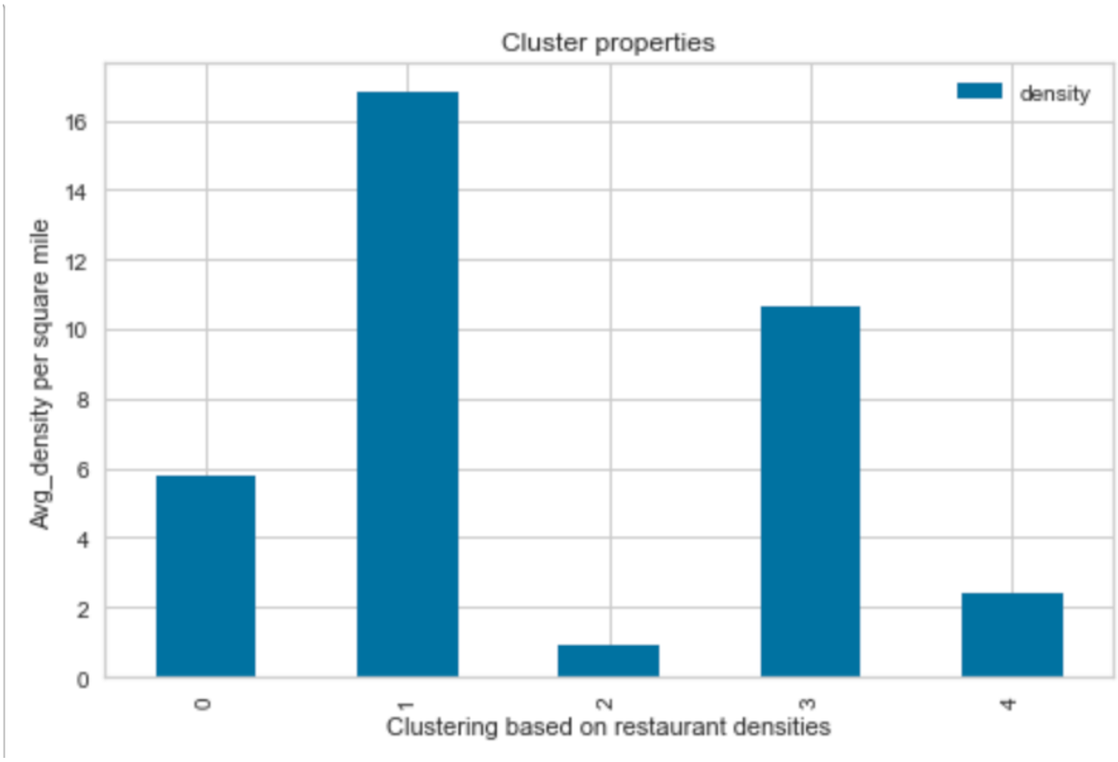
As mentioned earlier, we will be analyzing restaurant location to predict areas for new restaurant set ups. First part is basically exploring the restaurants and computing the densities for each neighborhood. Then we can form clusters with high, low and medium densities. This is a very good starting point for stakeholders as it gives them an idea of which regions are already saturated with restaurants and which are still open to welcoming new restaurants.

Then we again use the same clustering technique to cluster neighborhoods based on their frequency in the neighborhood. This is similar to what was done in the New York dataset. Finally, one can use this cluster information to find particular kind of restaurants in a region and also determine if they belong dense or sparse neighborhood.

A further exploration could also be looking at the distance of a particular restaurant from neighborhood center and cluster regions based on their proximity to a particular kind of restaurant. We don't explore this here. But one can easily do a follow up in this direction using the analysis presented here.

3. Analysis

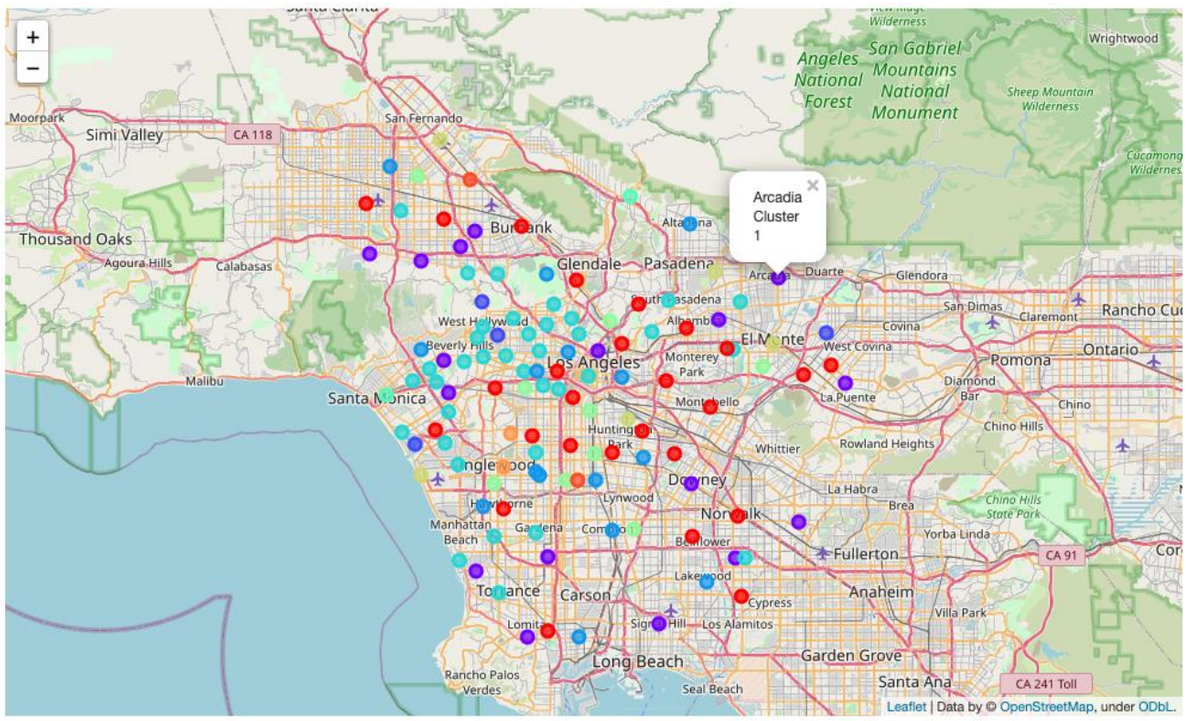
Similar to the New York dataset, we create one hot vectors for each restaurant venue. For the densities part, we group by neighborhood and find the number of restaurants in each neighborhood by using the sqmi column in Table 1 above. Then we evoke the KMeans() algorithm. First we use the elbow technique to find the optimal k for the KMeans(). This was found to be 5. Hence, we use n_clusters = 5 for our analysis. The cluster distribution was obtained as follows :



We can clearly see the segregation into high, medium and low densities. To, simplify our analysis, we rename the labels for the clusters as follows :

{1 & 3 : high, 0 & 4 : medium, 2 : low}

Next, we find the most common restaurant venues for each neighborhood by computing their frequency of occurrence. This would give us an idea about the most sought-after cuisine in each neighborhood. Then we use this to run the KMeans() algorithm again. The elbow technique was not significantly helpful as it was throwing different values everytime we ran it. So we just chose an average of those different values which turned out to be close to 10. The segmentation resulted in the following map :



The clusters are color coded above. Just as an example, I showed a popup in the above figure for a neighborhood in cluster 1 called Arcadia marked in purple. All the purple marked neighborhoods are in cluster 1.

4. Visualization of Clusters

Finally, let's see how this cluster information can be combined to evoke some interesting results. First, we combine both the cluster labels into one table by using the join function using the 'Neighborhood' column as the key.
The following shows a portion of the resulting table –

Table 3 :

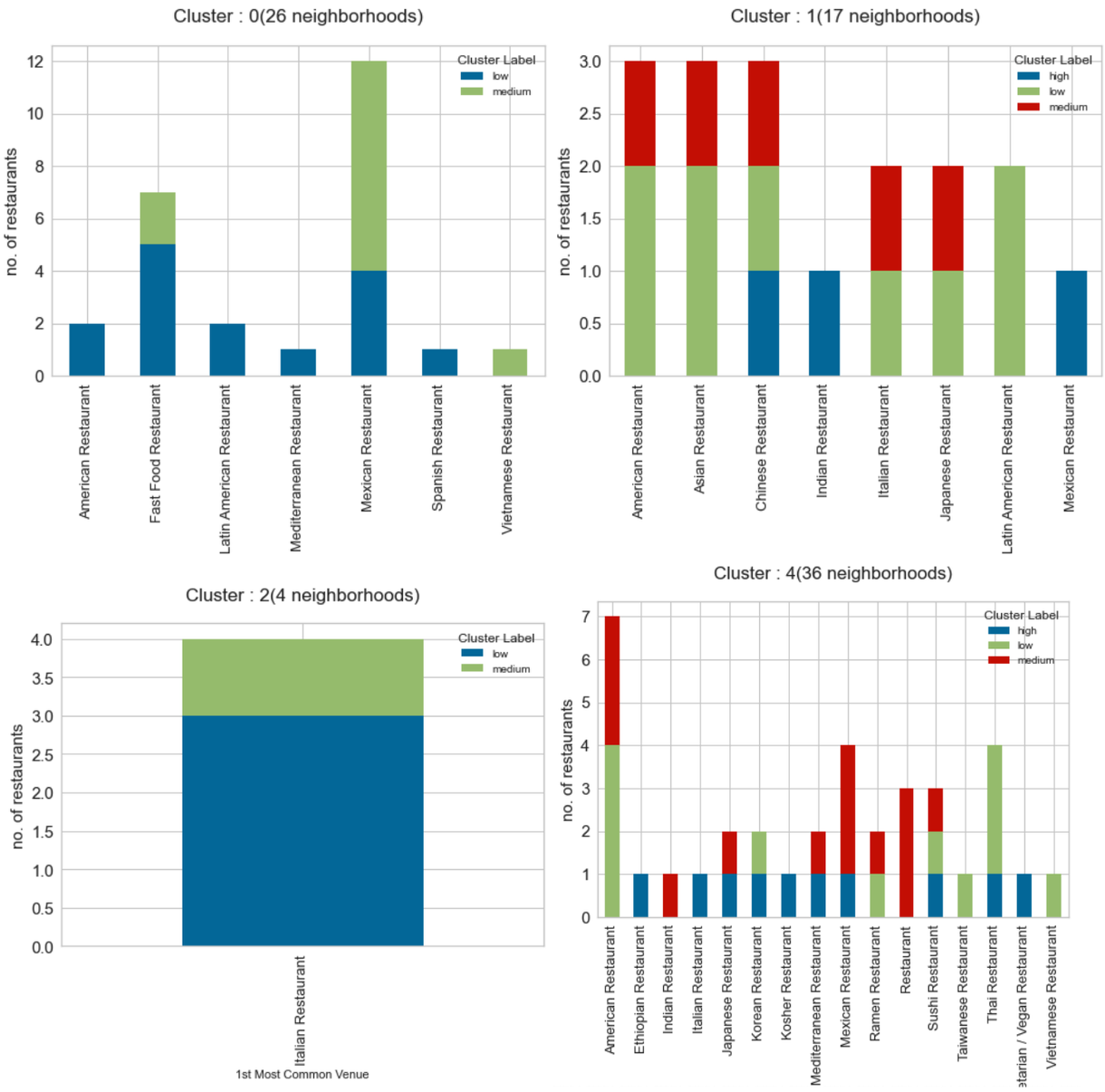
Cluster Label	Latitude	Longitude	Cluster Labels 2	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	34.031461	-118.300208	4	Sushi Restaurant	Vietnamese Restaurant	Doner Restaurant	Hawaiian Restaurant	Greek Restaurant
2	34.085539	-118.136512	0	Fast Food Restaurant	Mexican Restaurant	Vietnamese Restaurant	Dongbei Restaurant	Hawaiian Restaurant
4	34.133230	-118.030419	1	Italian Restaurant	Thai Restaurant	Japanese Restaurant	Fast Food Restaurant	Mexican Restaurant
3	34.044910	-118.323408	4	Vegetarian / Vegan Restaurant	Korean Restaurant	Latin American Restaurant	Mexican Restaurant	Sushi Restaurant
3	33.866896	-118.080101	1	Indian Restaurant	Chinese Restaurant	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Thai Restaurant

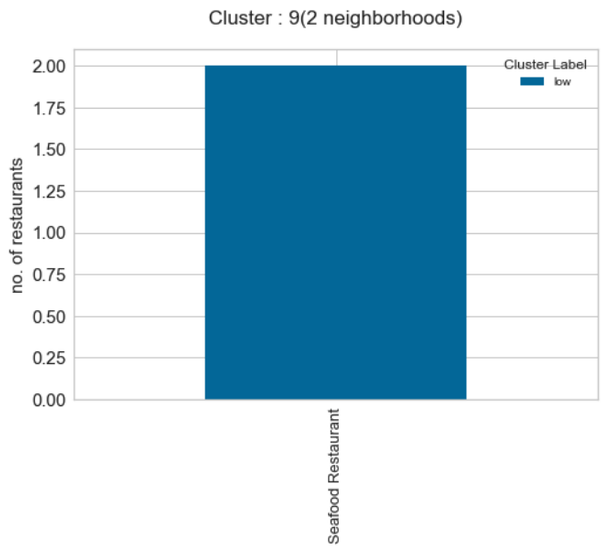
Next, we defined a function that takes all neighborhoods in one cluster, and finds the number of restaurants in the '1st most common' category' belonging to each restaurant type. To this, we also stack the information about whether these restaurants belong to the high, medium or low density cluster category found earlier. So, overall this gives you information about the most common venues in a cluster and whether they belong to a crowded or sparse neighborhood characterized by the density of the cluster they belong to.

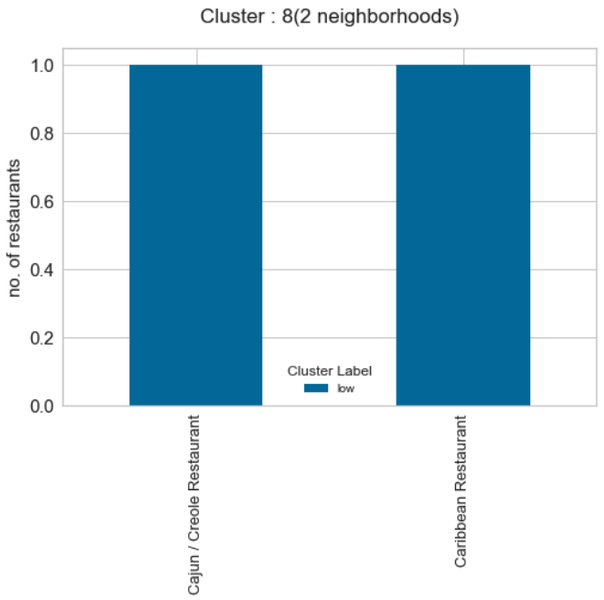
This idea could be refined further by quantifying the density of the clusters rather than using just 3 labels. But here we use just 3 labels to make the visualization a bit straightforward. Some of these figures are shown below. We can derive some obvious conclusions by looking at these plots and some a bit hand wavy. Cluster 2 is dominantly Italian while 0 is leaning a bit towards Mexican. Similarly, anyone's interested in seafood would likely visit a neighborhood in cluster 9 or cluster 8 if craving some Cajun or Caribbean dish.

Now for stakeholders,let's say someone's planning to open an Italian restaurant. Firstly, one could certainly avoid Cluster 2. One could argue that

the restaurants in cluster 2 are actually in low density regions, but the other factor to consider is popularity. A low-density region could also be synonymous with it being less popular. Cluster 1 in that regard could be a good option. It has a very diverse option of cuisines (could correspond to a diverse audience in the vicinity) and the Italian restaurants are actually located in medium and low-density regions which provides a good tradeoff between competition and popularity. There are of course many other factors to look for, but this certainly gives a good starting point.







5. Conclusion

Similar conclusions as above can be derived for other restaurants as well. One can also look at the second most common venues as well to gain some additional information about the clusters. Further exploration could also be based on things like distance from public transport, availability of parking lots, population, etc. So, one can naturally carry forward this analysis further based on the information collected in this report and make even stronger predictions.