IBM Data Science Specialization – Capstone Project

Looking for best place to open a Lebanese restaurant in Los Angeles, US.

By Abbass Hammoud

Introduction

Problem description and background

Fadi, a Lebanese-American, has been living in San Diego for 12 years, where he runs a family owned restaurant. Since he has been very successful, he is looking to expand his business and opening a new Lebanese restaurant in Los Angeles. LA is a very big city which is home to around 4 million inhabitants. It is considered the heart of the state of California and is a hot spot for tourists and visitors. Since LA is a very diverse city, having an authentic Lebanese restaurant offering fine oriental cuisine would be very appealing to the local population, as well as for visitors.

Since the city is big and has many neighborhoods, Fadi doesn't know which neighborhood he should pick for the location of his restaurant, in order to maximize his success and eventually his potential revenues. So he calls our data consulting firm to help him figure out the best neighborhoods, and avoid the less attractive ones.

Data description

To look more deeply into the problem, we start by collecting data about the city's neighborhoods. We can find the list of neighborhoods in the city of Los Angeles in the Wikipedia page: https://en.wikipedia.org/wiki/List_of_districts_and_neighborhoods_of_Los_Angeles

We process the page and extract the list using the BeautifulSoup library. Then we use the library geopy to find the geographical coordinates of each of neighborhood.

After this step, we collect information about the population. There are two main groups of people who would be interested in the restaurant and can affect where to place it: the local population, and the visitors and tourists. Places with high number of inhabitants and visitors are more attractive to open a restaurant, since they represent a higher potential of revenues.

For the local population, we can find the number of inhabitants in each neighborhood in this page from the university of UCLA: http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_LA_Neighborhoods_Data

Since neighborhoods can be small or large in area, the number of inhabitants alone is not decisive to have a fair comparison. Therefore, we find the surface area of each neighborhood and then calculate the number of people per mile square, which is more indicative.

For the category of visitors and tourists, since it's not possible to find the exact number of people who visit each of the neighborhoods, we look at the points of attractions instead. We use the foursquare api to obtain data about the categories of the venues in each neighborhood. Overall, we collect information about 6 categories for each neighborhood:

- Arts & Entertainment
- College & University
- Events (although events are not permanent, they can give an indication of how much certain places are frequented)
- Outdoors & Recreation
- Professional & Other Places
- Travel & Transport

Then we sum the total number of attractions in a new column.

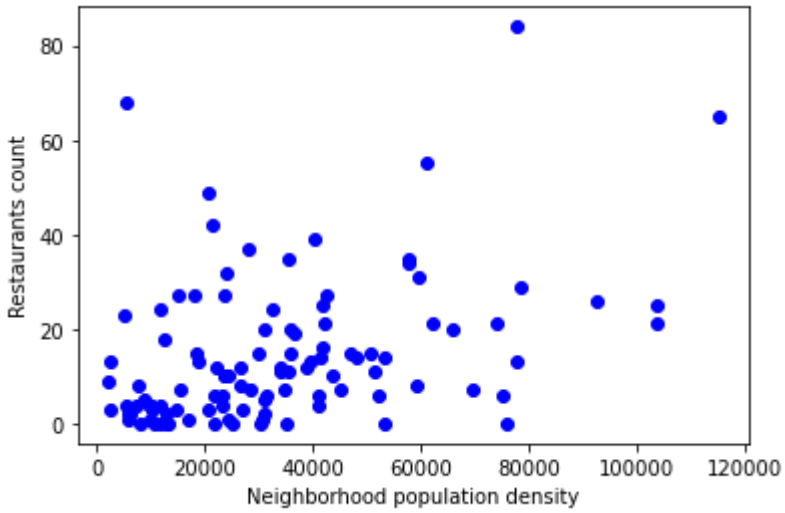
In addition to that, we collect data about restaurants in the city. Since it's more likely to have success in places where there are less restaurants to serve people, we store the number of all restaurants within a neighborhood, and also the number of Lebanese ones among them.

At this point, we store all the collected information in one dataframe, ready to be processed.

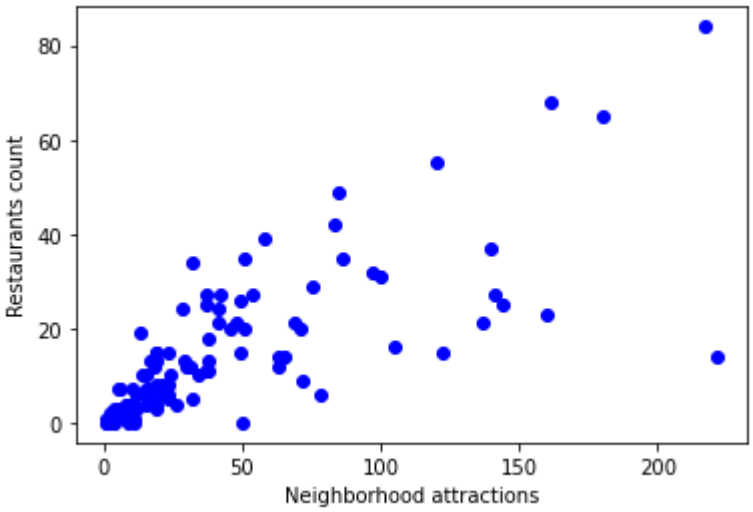
Methodology

Exploratory data analysis

We start first by exploring the data we have collected. To see the relation between the number of restaurants and the population density inside a neighborhood, we use a scatter plot:

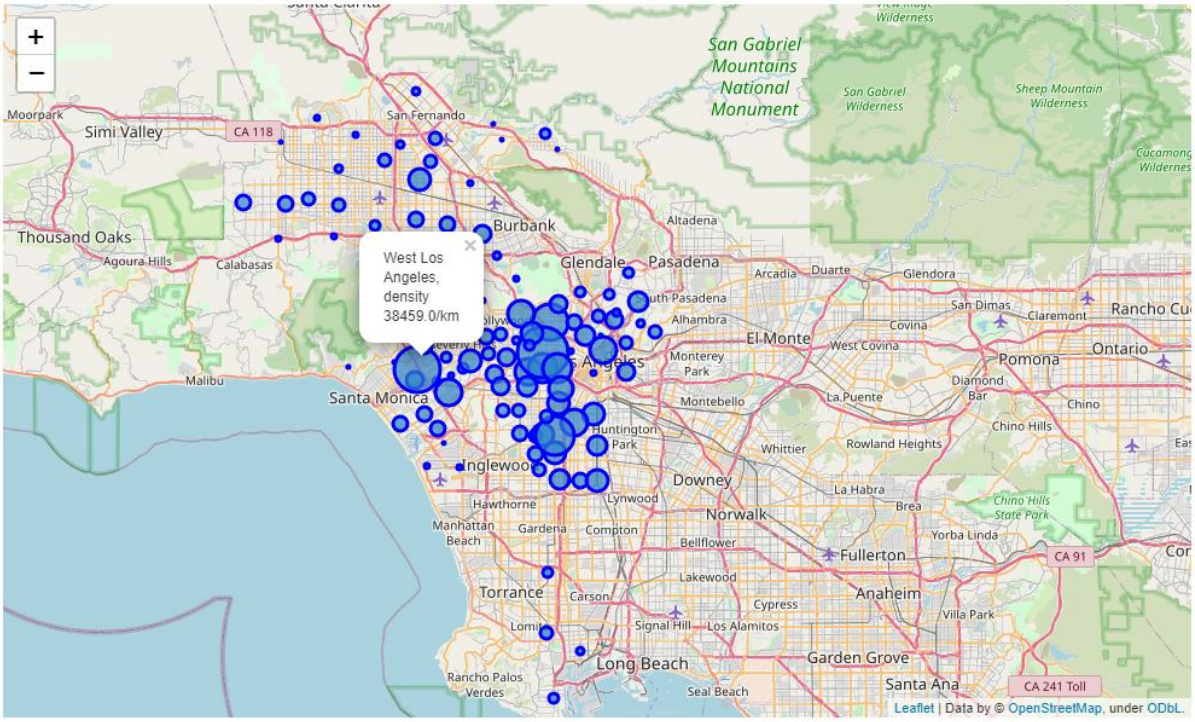


From the plot, we can see a general tendency, although not very clear, for the number of restaurants to increase when the population density increases. This is seen more clearly when we plot the number of restaurants versus the total number of attractions in a neighborhood:

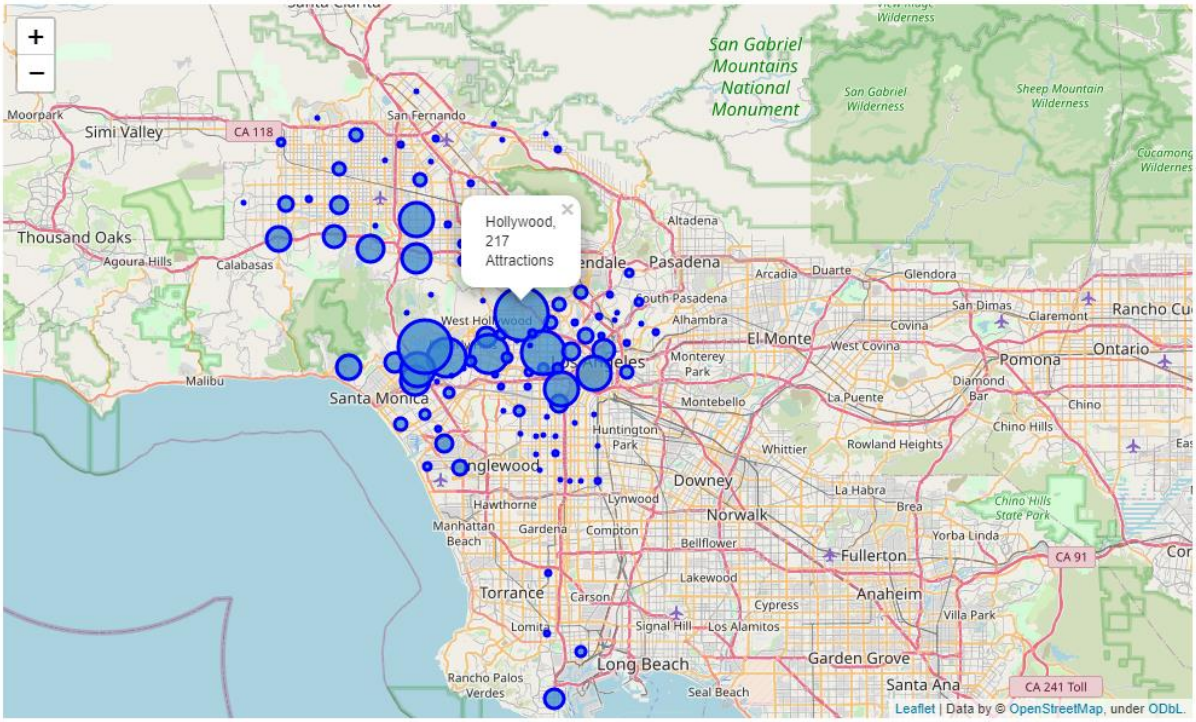


This shows that people logically tend to open restaurants in places with high population and high number of visitors/tourists who frequent nearby attractions.

To visualize the distribution of LA neighborhoods and their population density, we plot the map of the city. In the map, we can see that the neighborhoods around downtown, and to the west are the most dense, while other neighborhoods are less dense.



We do the same for the attractions count, and we can see also where the attractions are concentrated:



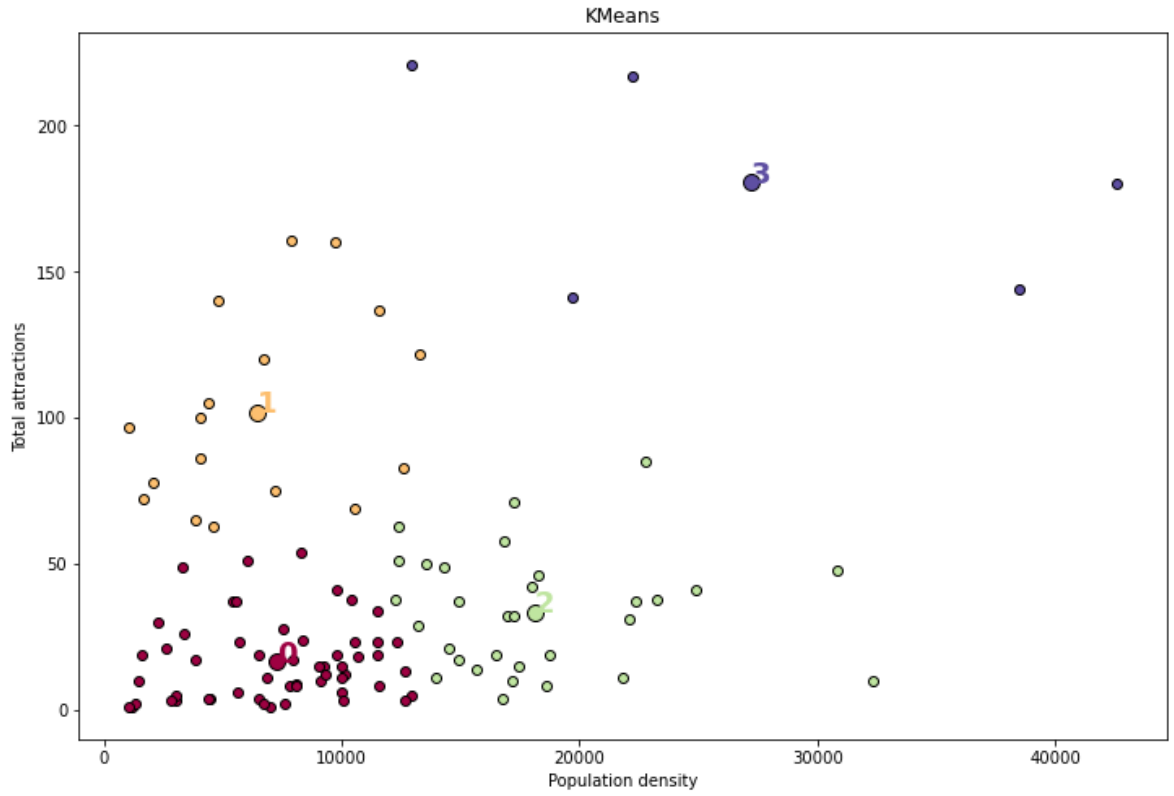
Machine learning methodology

The neighborhoods data is not labeled, so supervised learning techniques cannot be used, and we should use unsupervised learning instead. For unsupervised learning, there are several clustering algorithms that can be used. I have tested kmeans, hierarchical clustering, and DBscan. By comparing the three results, I have found that the mapping of neighborhoods to clusters is best using kmeans.

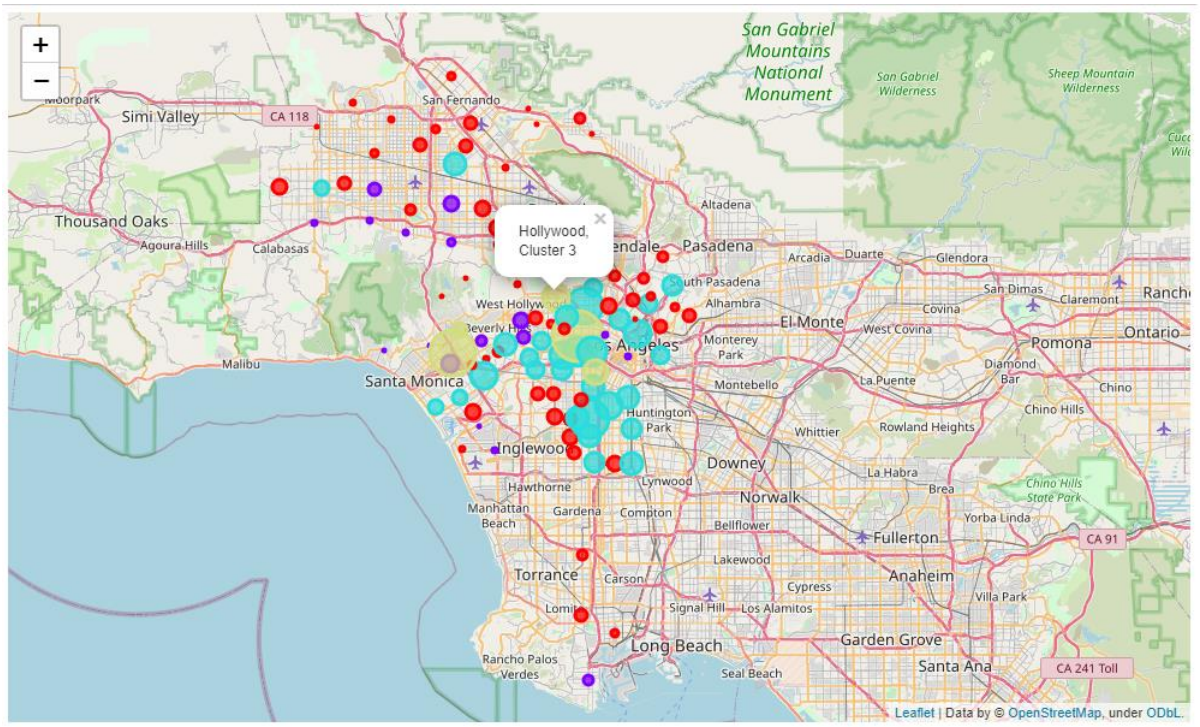
When using kmeans, one has to specify the number of clusters. So I tested 3, 4 and 5 and I found 4 clusters to give the most suitable clustering. Less than 3 or more than 5 is not useful to gain insights on the different classes of neighborhoods.

Results

We select the most important factors as features for learning: population density and total number of attractions. We normalize the data and then using kmeans with 4 clusters, we get this result:



Displaying the clusters on the map, we get this:



The neighborhoods details of each cluster are displayed on the notebook. In total, we can observe the following about the 4 clusters of neighborhoods:

- Cluster 0: contains the neighborhoods with low population density and low number of attractions
- Cluster 1: contains the neighborhoods with high number of attractions but relatively low population density
- Cluster 2: contains the neighborhoods with high population density but relatively low number of attractions
- Cluster 3: contains the neighborhoods with high population density and high number of attractions

Discussion

The problem was: where to place the new Lebanese restaurant in the city ?

Now looking at the clusters, we can say that in order to have the best potential to success and eventually higher revenues, Fadi should find a place in one of the neighborhoods of cluster 3: Hollywood, Koreatown, University Park, West Los Angeles, or Westwood

	Neighborhood	Cluster label	Population density	Total attractions	Restaurants count	Lebanese restaurants count
42	Hollywood	3	22234.0	217	84	1
47	Koreatown	3	42619.0	180	65	0
88	University Park	3	19663.0	141	27	0
102	West Los Angeles	3	38459.0	144	25	2
103	Westwood	3	12950.0	221	14	0

Among these neighborhoods, he might avoid Hollywood and West Los Angeles if he desires to avoid competition from same type restaurant. Moreover, if he's looking for a place with not much restaurants in general, he might go for University Park or Westwood, rather than Koreatown.

What if it is too hard to find a place for restaurant in these neighborhoods, or it's beyond the budget of Fadi? In this case, he can opt out for one of the neighborhoods in cluster 1 if he would like to target visitors/tourists, or one from cluster 2 if he would like to target the local population.

Finally the neighborhoods in cluster 0 will present the least potential, since they have lower population density and attractions to attract visitors and tourists. Therefore, he should better avoid this cluster.

Conclusion

In this report, we have studied the neighborhoods of Los Angeles, to help Fadi decide on the best place to open his new Lebanese restaurant. We started by presenting the problem, collecting data about the different neighborhoods from pages on the internet, then we used kmeans clustering technique to classify the neighborhoods. Finally, we presented the results and gave the recommendations on where is the best place to open the restaurant.