Clustering the Neighborhoods in Los Angeles

Tunan Jia August 31, 2020

Introduction

Los Angeles is a great city with different district, each district is filled with venues classified by categories, such as university, shops, food and event. Those general categories for venues are the first insights for new arrivers who are considering to either purchase houses or live for working, studying, or simply for visiting.

We will use this projection data and foursquare API for following analysis:

1. Classifying neighborhoods are highly developed, downtown and underdeveloped.
2. Classifying neighborhoods by the venue's frequencies, understanding the venues to distinguish each neighborhood.
3. Recommend places for people with searching purposes.

Data

We need data from reliable sources for analysis.

1.districts in Los Angeles, from

https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Los_Angeles

2. Latitude and Longitude of each neighborhood is retrieved using Geocode from Geopy library of python

3. Foursquare Developers Access to venue data: http://foursqaure.com/

Methodology

- Data scraping, exploration data analysis.
- K-means clustering algorithm to segment neighborhoods.
- Elbow and Silhouette methods to select the appropriate k numbers.
- Graphs to visualize the results.

Data retrieval

Firstly, get all districts in Los Angeles, from

https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Los_Angeles

Then, using Geocoder to obtain all latitudes and longitudes:

	Neighborhood	Latitude	Longitude
0	Central Los Angeles	34.053691	-118.242767
1	Eastside Los Angeles	34.030625	-118.246639
3	Northwest Los Angeles	-11.275489	-67.382600
4	San Fernando Valley	34.214885	-118.499820
5	South Los Angeles	33.928291	-118.278813

As there are too many venues for each neighborhood, and we are only interest to an overall development and focus of each district, so I used foursquare to explore the types of category, for which there are 10 in total. As a result, this limit our exploration down to a narrower filler.

```
Arts & Entertainment (4d4b7104d754a06370d81259)
College & University (4d4b7105d754a06372d81259)
Event (4d4b7105d754a06373d81259)
Food (4d4b7105d754a06374d81259)
Nightlife Spot (4d4b7105d754a06376d81259)
Outdoors & Recreation (4d4b7105d754a06377d81259)
Professional & Other Places (4d4b7105d754a06375d81259)
Residence (4e67e38e036454776db1fb3a)
Shop & Service (4d4b7105d754a06378d81259)
Travel & Transport (4d4b7105d754a06379d81259)
'4d4b7104d754a06370d81259'
```

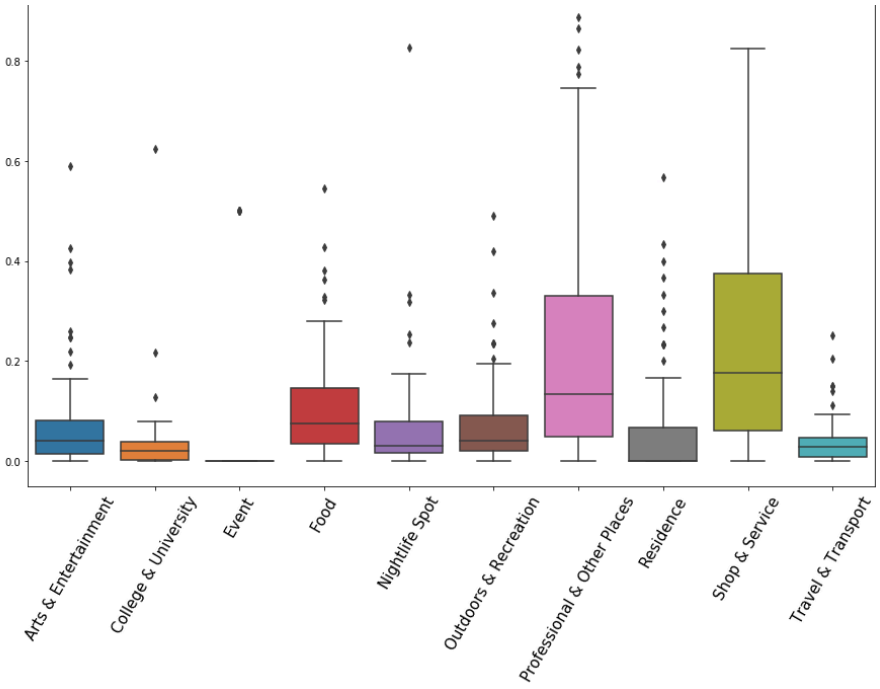
A resulting data frame:

	Neighborhood	Latitude	Longitude	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	Central Los Angeles	34.053691	-118.242767	73	8	2	171	63	98	141	30	144	107
1	Eastside Los Angeles	34.030625	-118.246639	10	1	2	56	4	7	77	1	128	6
3	Northwest Los Angeles	-11.275489	-67.382600	5	3	0	16	7	7	35	0	47	10
4	San Fernando Valley	34.214885	-118.499820	2	2	0	8	1	5	11	0	19	4
5	South Los Angeles	33.928291	-118.278813	3	63	1	47	5	48	100	17	69	7

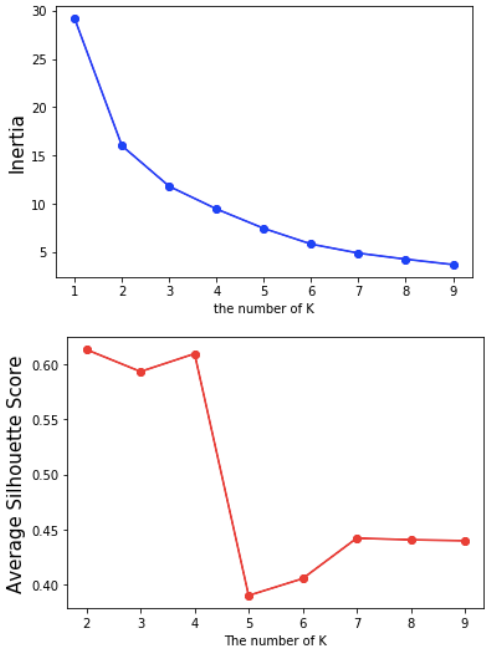
Clustering

Before doing the clustering analysis, some scaled preprocessing needs to be done. Since I am interested in the frequency, so I uniformed each category into range from 0 to 1. Here is the screenshot for the scaled dataframe and a boxplot to visualize their distribution:

	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	1.000000	0.079208	1.0	1.000000	1.000000	1.000000	1.000000	1.000000	0.791209	1.000000
1	0.136986	0.009901	1.0	0.327485	0.063492	0.071429	0.546099	0.033333	0.703297	0.056075
2	0.068493	0.029703	0.0	0.093567	0.111111	0.071429	0.248227	0.000000	0.258242	0.093458
3	0.027397	0.019802	0.0	0.046784	0.015873	0.051020	0.078014	0.000000	0.104396	0.037383
4	0.041096	0.623762	0.5	0.274854	0.079365	0.489796	0.709220	0.566667	0.379121	0.065421

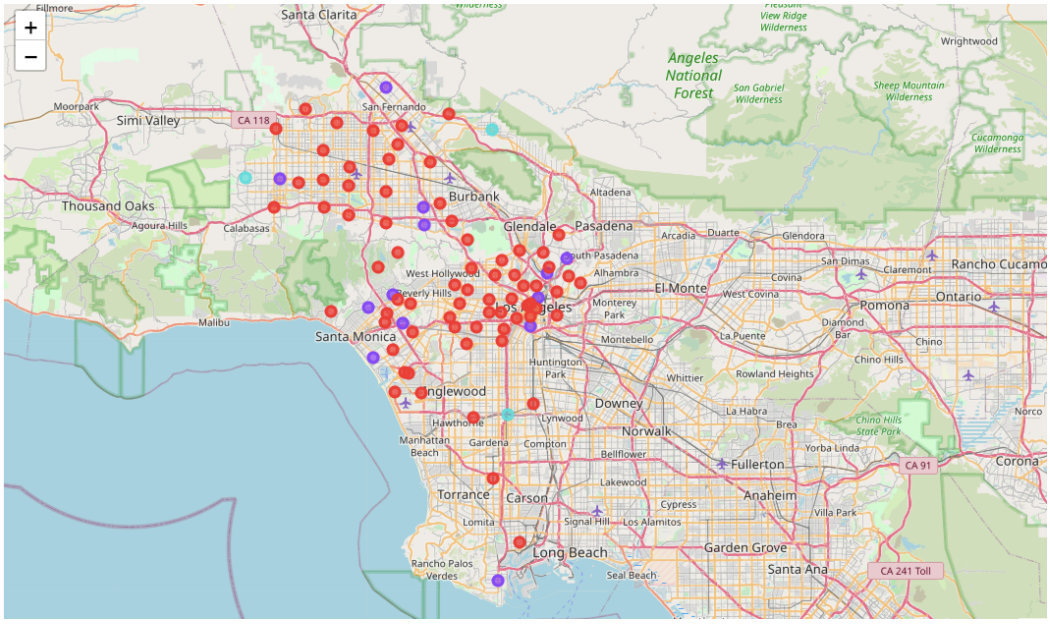


To choose the ‘most-right’ number to segment, I used both elbow and silhouette method



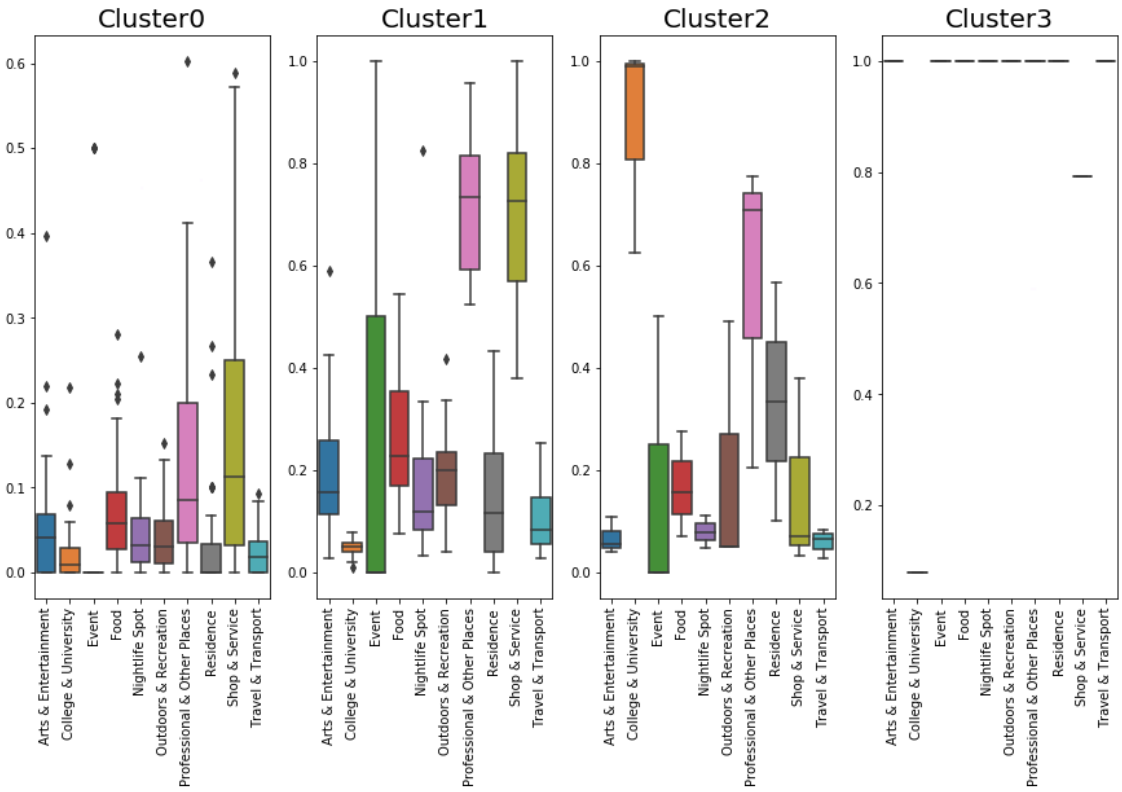
As shown, the elbow appear at k equals 3 or 4, but the silhouette suggests a higher score at k equals 4, so I’ll use 4 as the number to segment clusters.

After filling the data frame with cluster labels, I used folium to show the geographic information about the resulted clusters.



From the graph, we can see most neighborhoods are labeled as cluster 0 (red), second most are the cluster 1 (purple), only few are labeled as cluster 2 (blue). Note the neighborhood labeled with 3 are not shown in this screenshot. Actually, its latitude and longitude are wrongly detected by the Geopy, the potential problem could be the district name, I'll leave those group not analyzed.

To examine the clusters, I created a plot to compare them:



Conclusion and Recommendation

Cluster 1 and 2 can been seen more developed than cluster 0, but in different ways.

'Academic atmosphere' is an important element to segment different neighborhoods, if people are looking forward to living near the university and college, he or she should move to those labeled cluster'2';

If people are working for placed to develop career and related to more professional events/services, areas labeled 1 can be the first choice.

Neighborhoods labels in 0 can be seen relatively undeveloped, with lower frequency of all categories.