

Hanoi Population density and Wards clustering

[Kiennt \(at\) iscale.vn](mailto:Kiennt (at) iscale.vn)

Introduction

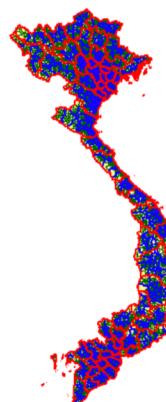
- Problem description
 - Study in Hanoi area
 - Answer questions:
 - Where to open the first restaurant/shop? => population density
 - Where will be the next locations after the first successful one? => ward clustering
- Data description
 - 2nd and 3rd level Geo shapefiles for Hanoi city
 - Area, population and population density on Hanoi committee website
 - Foursquare API to get most common venues to be used for clustering

Data collection and preparation

Download GADM data (version 3.6)

Country
Vietnam

Geopackage
Shapefile
R (sp): level-0, level1, level2, level3
R (sf): level-0, level1, level2, level3
KMZ: level-0, level1, level2, level3



The coordinate reference system is [longitude/latitude](#) and the [WGS84](#) datum.

	ENGTYPE_2	geometry	District	Urban	Area	Avg_Population	Population_density	Wards
1	Urban District	POLYGON ((105.72945 21.05656, 105.72884 21.056...)	Bac Tu Liem	1	45.32	333.7	7363	13
2	Urban District	POLYGON ((105.80135 21.03023, 105.80577 21.035...)	Ba Dinh	1	9.21	243.2	26406	14
3	District	MULTIPOLYGON (((105.46503 21.07101, 105.46525 ...))	Ba Vi	0	423.00	284.1	672	30
4	Urban District	POLYGON ((105.81349 21.00882, 105.81209 21.007...)	Cau Giay	1	12.32	280.5	22768	8
5	District	POLYGON ((105.57925 20.91554, 105.58212 20.912...))	Chuong My	0	237.38	332.8	1402	30
6	Urban District	POLYGON ((105.81349 21.00882, 105.80473 21.015...))	Dong Da	1	9.95	422.1	42422	21
7	District	POLYGON ((105.71733 21.07171, 105.70901 21.073...))	Dan Phuong	0	78.00	164.2	2015	15
8	District	POLYGON ((105.84167 21.08116, 105.83591 21.085...))	Dong Anh	0	185.62	384.7	2073	23
9	District	POLYGON ((105.96449 20.96341, 105.96065 20.963...))	Gia Lam	0	116.71	277.2	2375	20

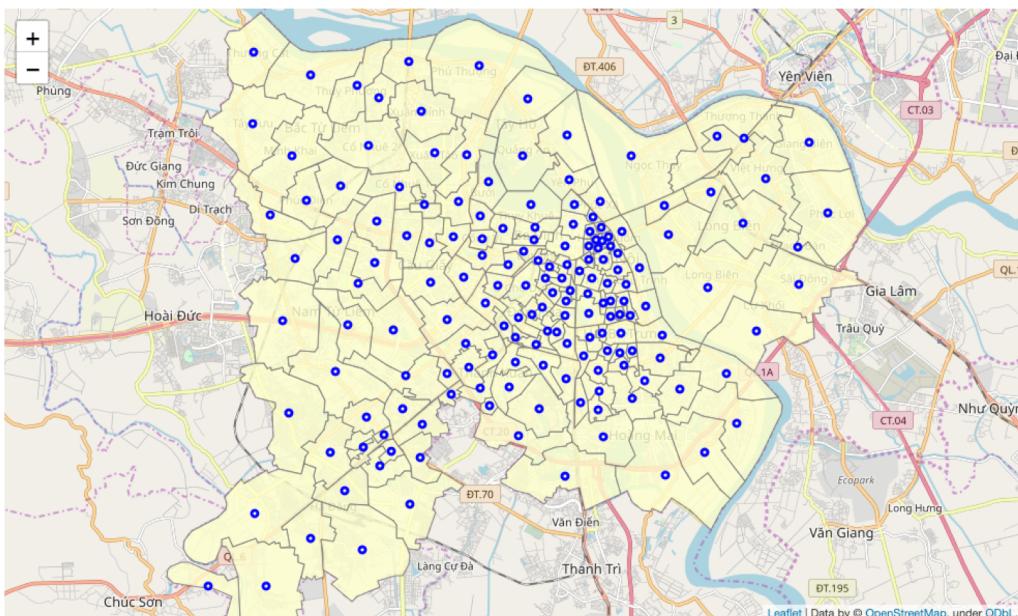
16 Diện tích, dân số, mật độ dân số và đơn vị hành chính năm 2018 phân theo đơn vị hành chính

Area, population, population density and administrative units 2018 by district

	Diện tích Area (Km ²)	Dân số trung bình (1000 ng) Average Population (Thous. Pers.)	Mật độ dân số (Người/km ²) Population density (Person/km ²)	Đơn vị hành chính (Administrative units)	
				Phường/Xã Precincts, communes	Thị trấn Town under district
TỔNG SỐ - TOTAL	3358,59	7852,6	2338	563	21
Ba Đình	9,21	243,2	26406	14	-
Hoàn Kiếm	5,29	153,0	28922	18	-
Tây Hồ	24,39	166,8	6839	8	-
Long Biên	59,82	294,5	4923	14	-
Cầu Giấy	12,32	280,5	22768	8	-
Đống Đa	9,95	422,1	42422	21	-
Hai Bà Trưng	10,26	311,8	30390	20	-
Hoàng Mai	40,32	443,6	11002	14	-
Thanh Xuân	9,09	286,7	31540	11	-
Sóc Sơn	304,76	341,1	1119	25	1
Đông Anh	185,62	384,7	2073	23	1
Gia Lâm	116,71	277,2	2375	20	2
Nam Từ Liêm	32,19	240,9	7484	10	-
Thanh Trì	63,49	266,5	4198	15	1
Bắc Từ Liêm	45,32	333,7	7363	13	-
Mê Linh	142,46	228,5	1604	16	2
Hà Đông	49,64	353,2	7115	17	-
Sơn Tây	117,43	151,3	1288	15	-
Ba Vì	423,00	284,1	672	30	1

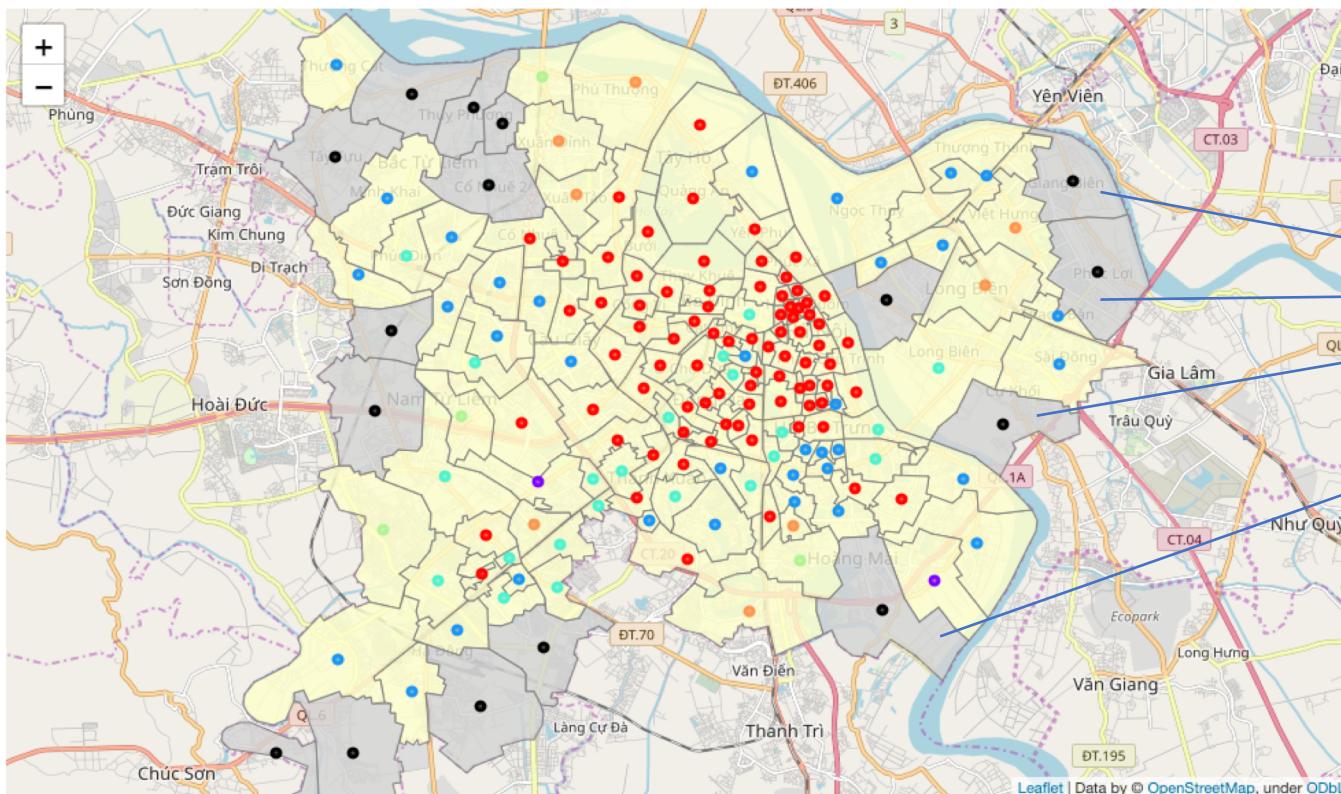
Data collection and preparation

```
# To use Foursquare API, we need a latitude and longitude for each ward  
# the centroid of each ward polygon is a good start  
uhanoi_gdf['Latitude'] = uhanoi_gdf['geometry'].centroid.y  
uhanoi_gdf['Longitude'] = uhanoi_gdf['geometry'].centroid.x
```



168 urban wards with corresponding centroids calculated by python geometry.centroid.y and geometry.centroid.x

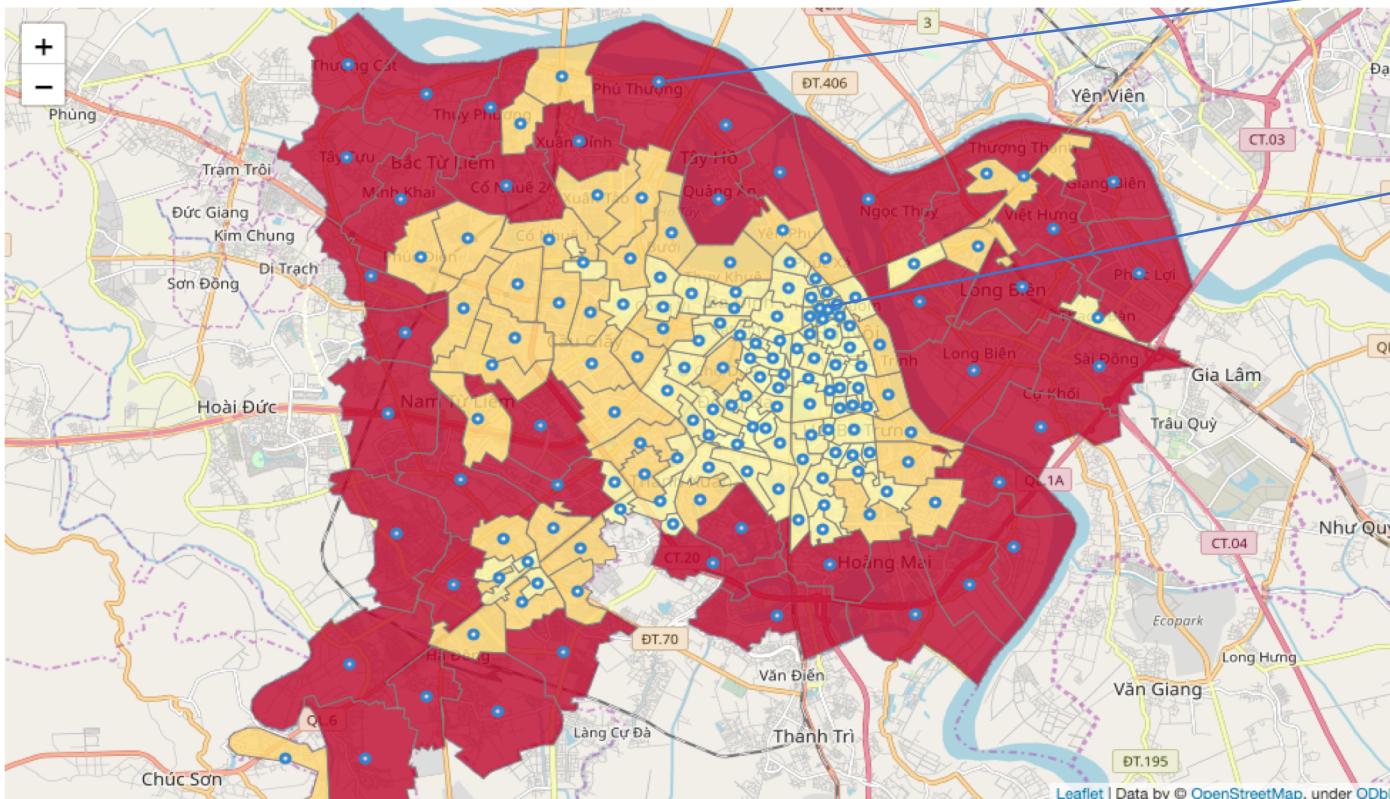
Data collection and preparation



16 “grey” wards with no nearby venues returned by Foursquare when searching radius set to 500m

Using Foursquare API with fixed searching radius is not good for all the wards with very different areas

Data collection and preparation



Largest ward
Phu Thuong: 7.47 sq. km

Smallest wards
Hang Bac, Hang Dao: 0.06 sq. km

=> Use different searching radius for each ward center

```
def getRadius(quant, area):
    if area <= quant.index[0]:
        return 300
    if area <= quant.index[1]:
        return 500
    if area <= quant.index[2]:
        return 800
    if area <= quant.index[3]:
        return 1500
    else: return 2000
```

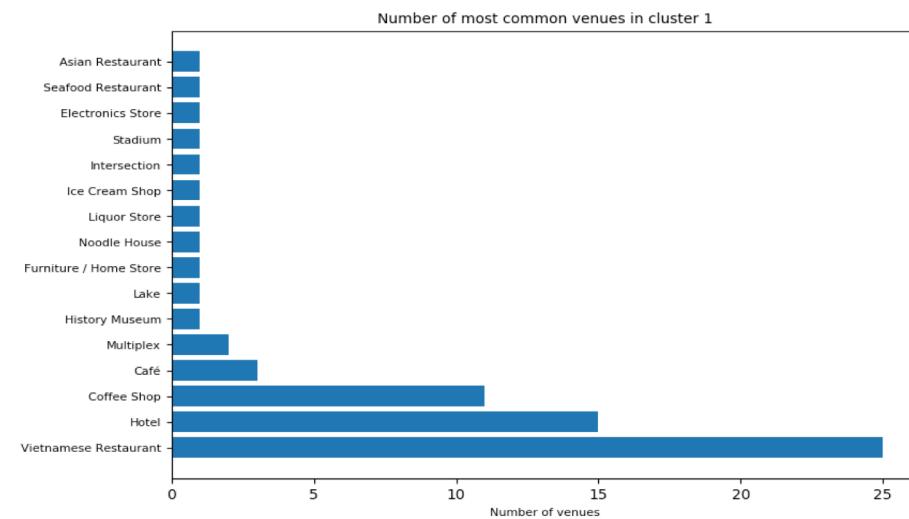
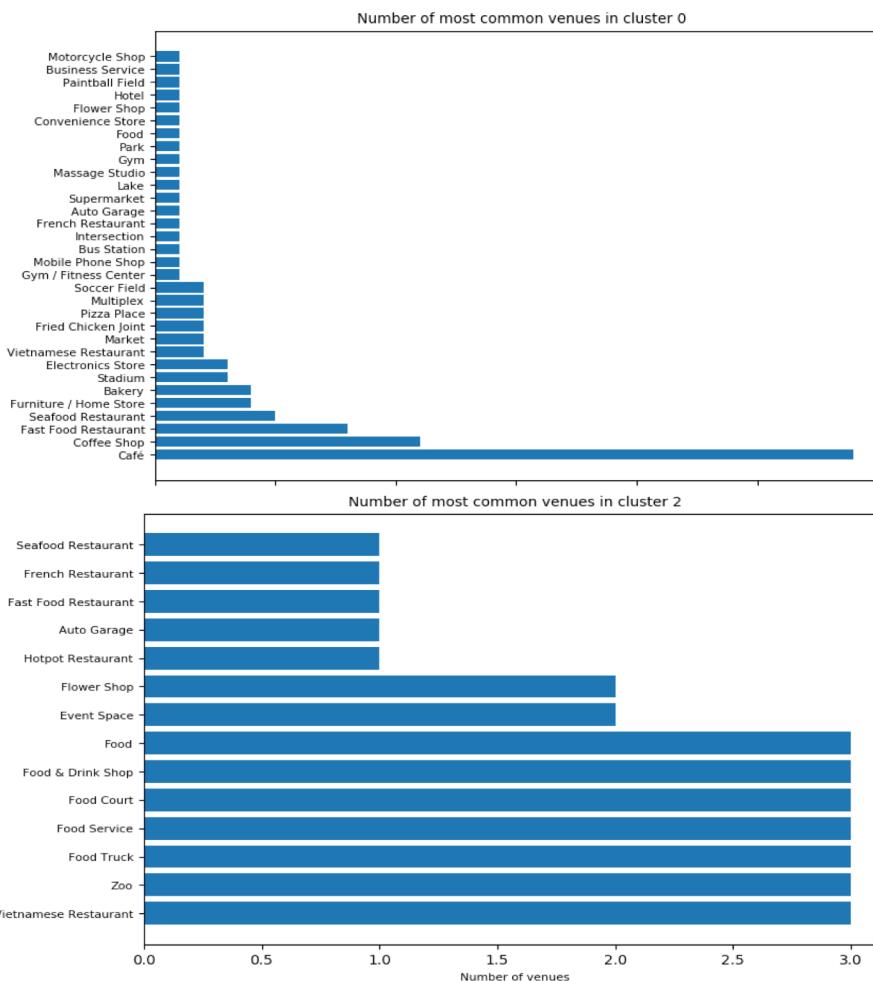
Utilizing Foursquare API

	Ward	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	Bach Dang	Vietnamese Restaurant	Coffee Shop	Café	Noodle House	Japanese Restaurant	Hotel	Sushi Restaurant	Restaurant	Tea Room	Comic Shop
1	Bach Khoa	Coffee Shop	Café	Vietnamese Restaurant	Bus Station	BBQ Joint	Fast Food Restaurant	Halal Restaurant	Bakery	Sushi Restaurant	Bubble Tea Shop
2	Bach Mai	Fast Food Restaurant	Halal Restaurant	Dessert Shop	Asian Restaurant	Vietnamese Restaurant	Café	Coffee Shop	Food	Fishing Spot	Flea Market
3	Bien Giang	Hotpot Restaurant	Vietnamese Restaurant	Zoo	Event Space	Food Truck	Food Service	Food Court	Food & Drink Shop	Food	Flower Shop
4	Bo De	Vietnamese Restaurant	Hotel	Coffee Shop	Café	Ice Cream Shop	Lounge	Vegetarian / Vegan Restaurant	Cocktail Bar	Fast Food Restaurant	Gastropub

Venue Limit = 100

Total: 4,507 venues in 210 unique categories

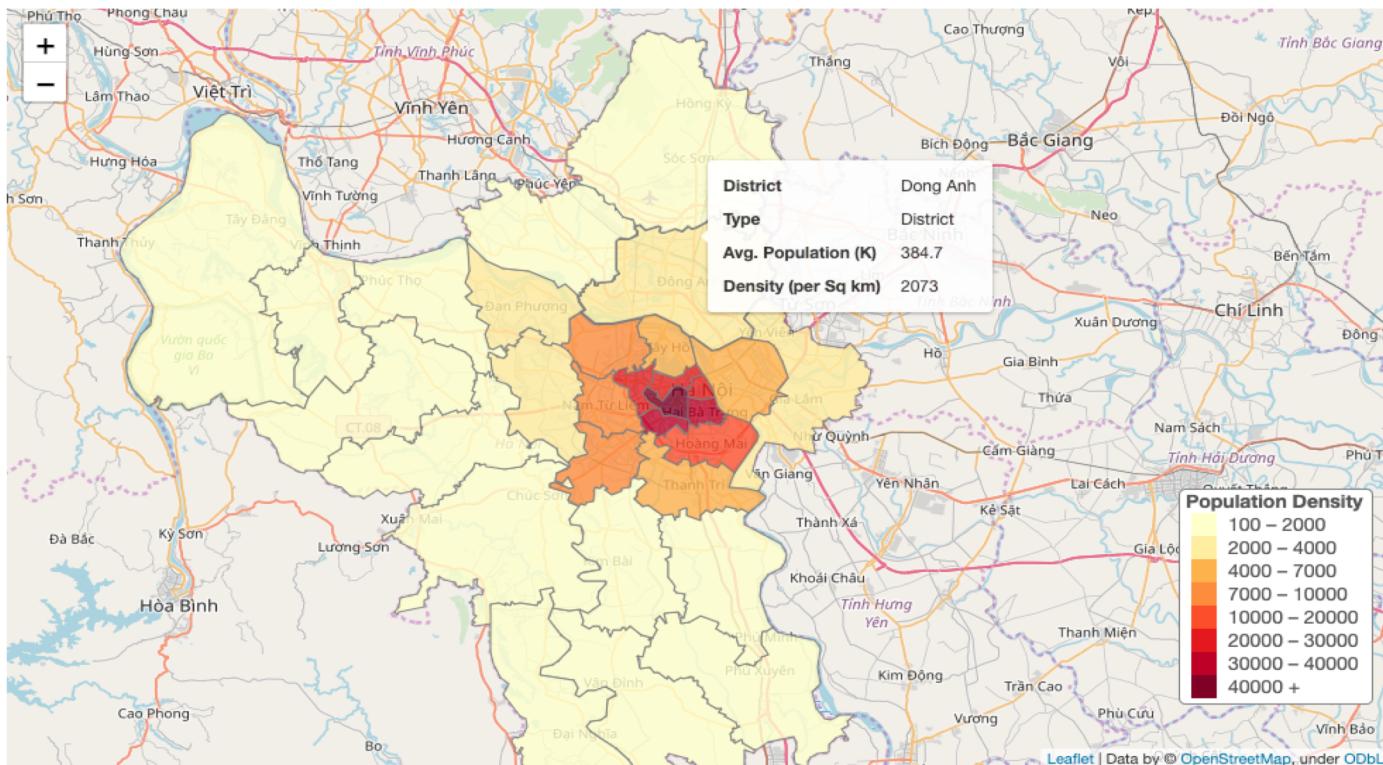
Applying Kmeans algorithm



3 clusters:

- Cluster 0: Café, Coffee Shops, Fast Food, Stores
- Cluster 1: Restaurant, Hotel, Coffee Shops
- Cluster 2: Outing Places

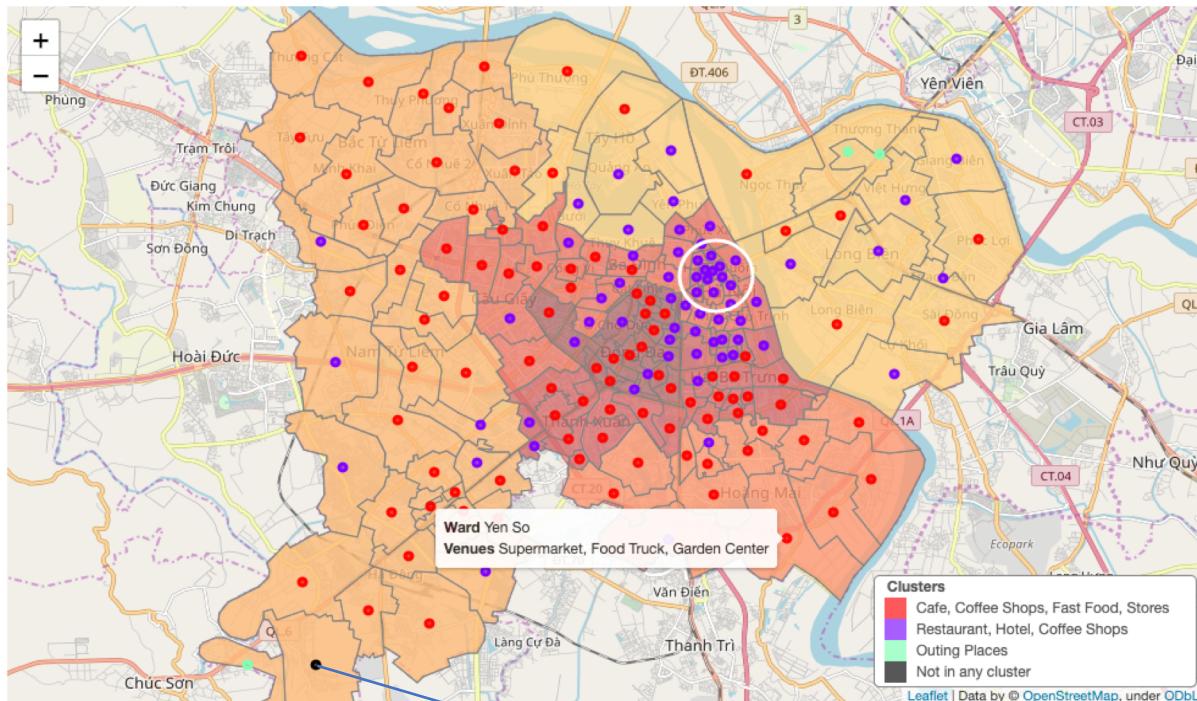
Result maps



Population density by district

- More than 40,000: Dong Da
- 30,000 - 40,000: Hai Ba Trung, Thanh Xuan
- 20,000 - 30,000: Ba Dinh, Hoan Kiem, Cau Giay
- 10,000 - 20,000: Hoang Mai
- 70,000 - 10,000: Bac Tu Liem, Nam Tu Liem, Ha Dong
- 4,000 - 7,000: Tay Ho, Long Bien, Thanh Tri
- Less than 4,000: the rest

Result maps



Not in any cluster
Foursquare return no data
even with radius = 2km

3 clusters:

- Cluster 0 (red): Café, Coffee Shops, Fast Food, Stores
- Cluster 1 (purple): Restaurant, Hotel, Coffee Shops
- Cluster 2 (green): Outing Places

Discussion

- Population statistics is not updated (since 2018) though population of districts in Hanoi change rapidly
- Foursquare updates database continuously, the clusters and maps will change regularly => need method to automatically choose number of cluster, assign appropriate labels and create maps
- Elbow was unclear with the dataset => further investigate BIC and SSE methods

Thank you!