IBM Data Science Capstone Project
Hanoi Population density and Wards clustering

Kiennt (at) iscale.vn

A. Introduction

1. Description of the problem and Discussion of the background

Hanoi, the capital of Vietnam, is my hometown. The city was founded in 1010 by monarch Ly Thai To. Since founded, it has remained the most important political and cultural center of Vietnam. Today, Hanoi covers an area of 3,328.9 square kilometers with an estimated population of 8.1 million as of 2019. The city is divided into 12 urban districts, 1 district-level town and 17 suburban districts (source: Wiki). Nearly half of the city population is living in the urban districts which are further divided into 168 wards.

Most of the economic and social activities are taking place in the urban area of the city. In the recent years, more and more restaurants and store chains have chosen Hanoi as their next destination. Famous chains include Starbucks, McDonald's, KFC, Lotte, Jollibee, Highlands coffee and CGV … Usually when investors come, there are two important questions that they want to seek answers to:

- Where to open the first restaurant/shop?
- Where will be the next locations after the first successful one?

Population density by district might be a good answer to the first question while location similarity in terms of popular venues can help answer the second one. In this capstone project, I will try to collect, analyze and visualize public data found on the internet using maps, charts and utilizing Foursquare API as well as clustering technique to help the investors make right decision.

2. Description of the data

The following data will be used in this project:

- GADM website (https://gadm.org/download_country_v3.html) provides the 3 levels of administrative map of Vietnam that we can use to get coordinates of all the districts and wards in Hanoi. The data is free to be used for non-commercial purpose.

  The shapefiles was downloaded and converted to geojson files (Hanoi_geo_2.json and gadm36_VNM_3.json) using https://mapshaper.org/

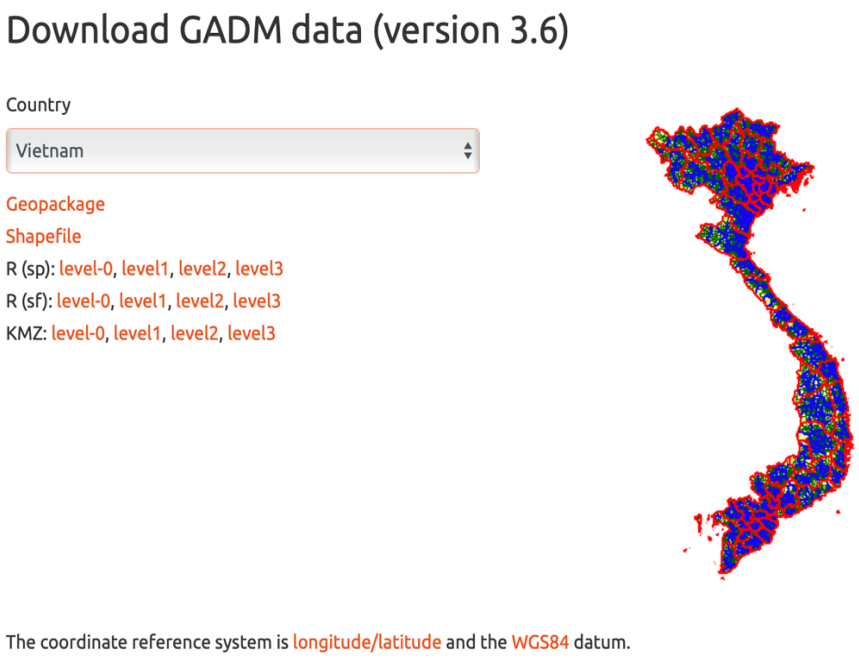


Figure 1. Website for downloading Vietnam 2nd and 3rd level shapefiles

- This site provides list of all districts and wards within districts in Vietnam from that we can extract the sub-list for Hanoi: https://www.gso.gov.vn/dmhc2015/

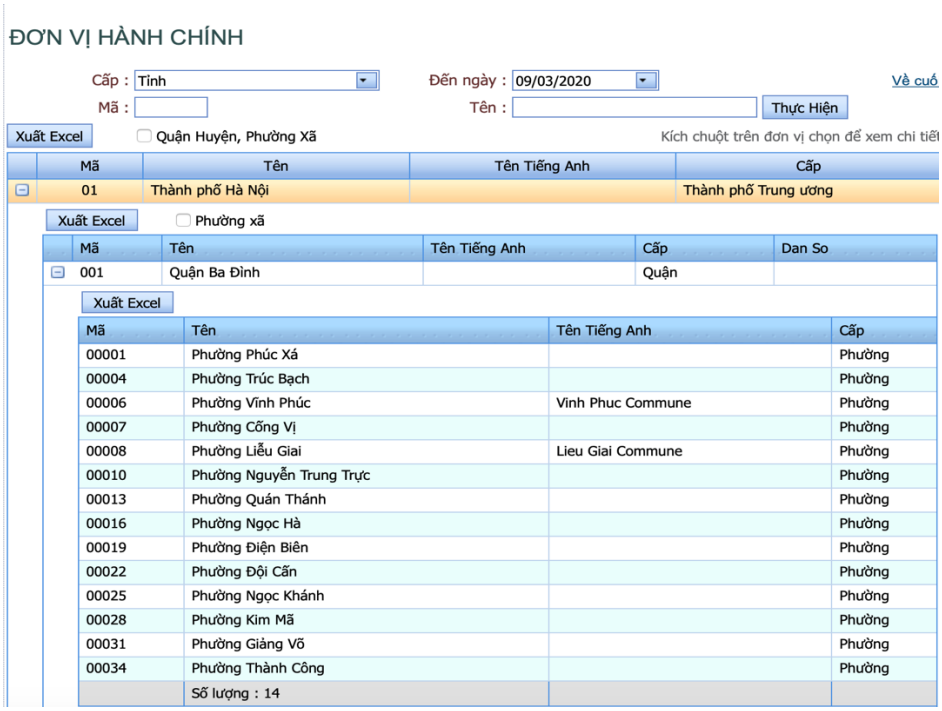
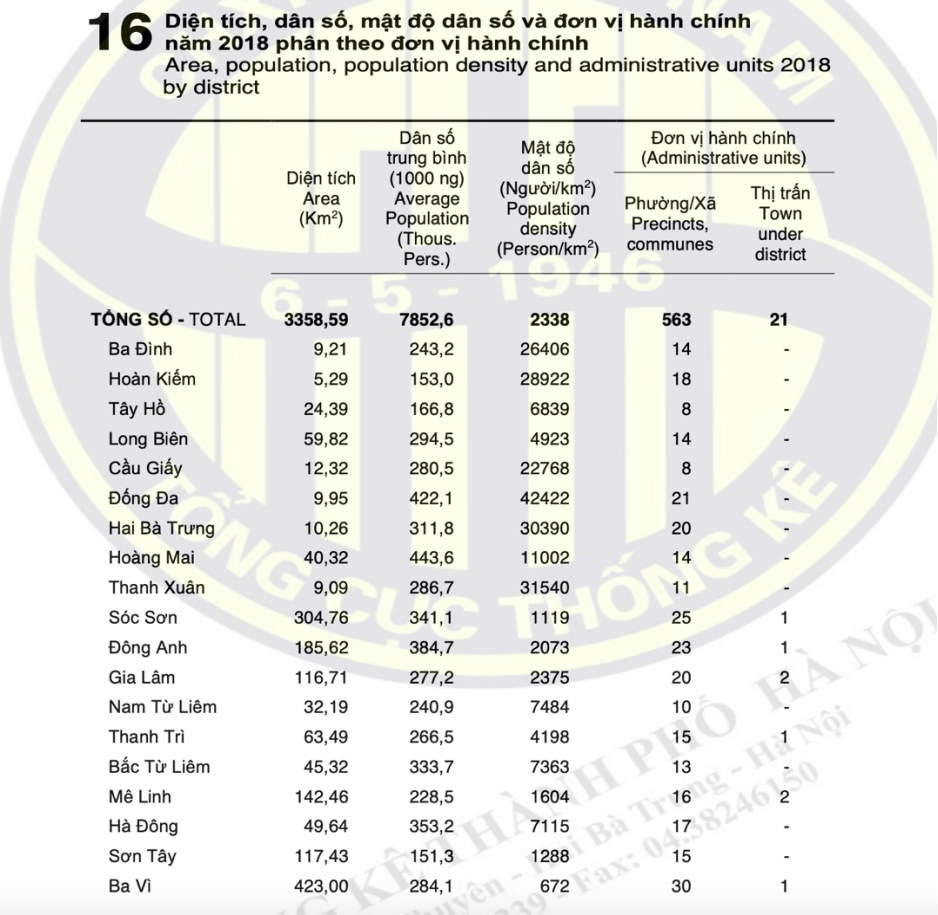


Figure 2. Official website from that list of districts and wards can be extracted

Alternatively, list of Hanoi districts and wards can also be extracted from the geojson file and that was the method used in this project.

- Area and population of each district of Hanoi was collected from this site and saved to "Hanoi_Area_population.csv" file: http://thongkehanoi.gov.vn/a/nien-giam-thong-ke-2018-1579246334/



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

Figure 3. Hanoi – 2018 Area, population, population density and administrative units by district

- Last but not least, Foursquare API was used to get most common venues for clustering the wards in the urban area of Hanoi.

B. Methodology

Population density

Population density by district was collected from the official site of Hanoi city committee [Figure 3]. The data has latest updates in 2018. This data was read and merged with geo data [Figure 1] to create the geopandas dataframe shown in [Figure 4].

	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

Figure 4. Geopandas dataframe including geometry, area and population density by district

I then used the python Folium library to display Hanoi map with districts colored based on the population density. Detail result is shown in the Results section.

Ward clustering

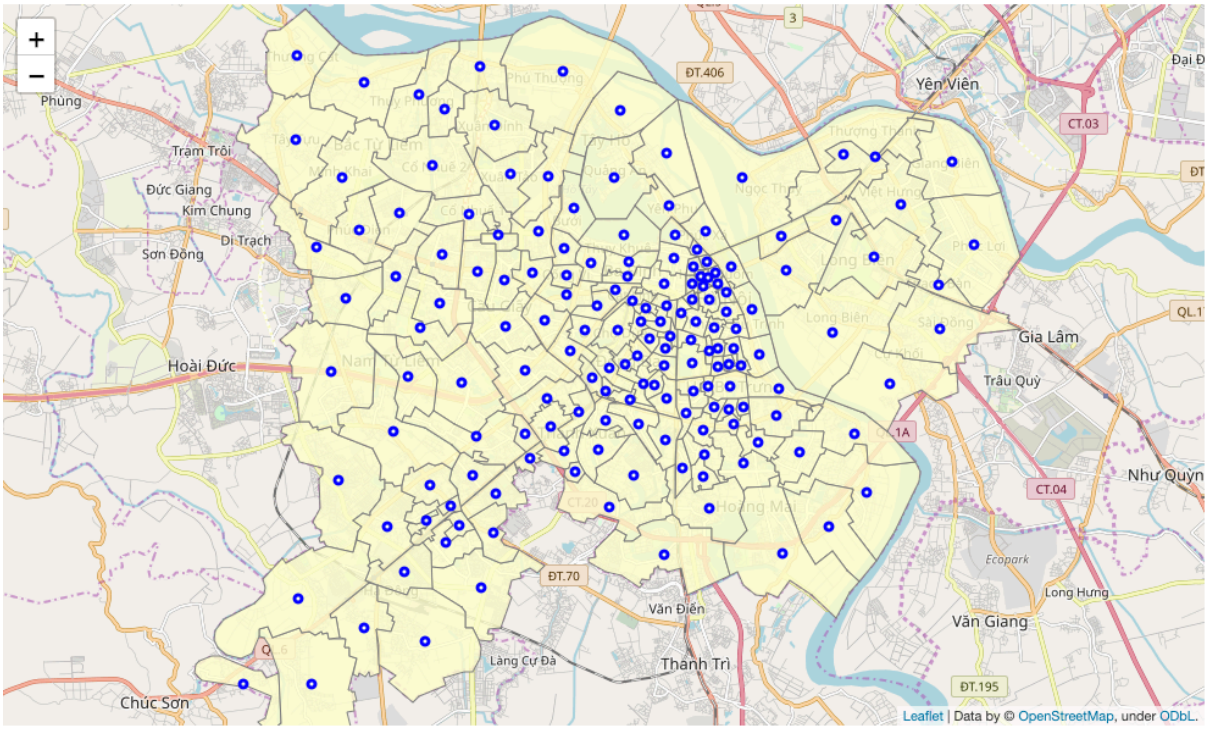


Figure 5. Map showing all the wards in the urban area and their centers (centroids)

For choosing good location for a business, besides the population density, it is very useful if we can have good insights of the most common venues around the locations as well as the similarities among them. To do this, I used Foursquare API to search for nearby venues of each ward center then applying K-Means clustering algorithm to cluster the wards. The wards' coordinates were collected from the 3rd level shape files by GADM. Center of each ward was set to its centroid by using its polygon geometry.centroid.x (longitude) and geometry.centroid.y (latitude) in python. [Figure 5] displays a map of all the wards in the urban area with their corresponding centers (geo centroids).

To find common venues near a location, the Foursquare API requires a searching radius. However, using a fixed radius for all the wards might not be a good idea. I have confirmed this by testing getting all the common venues near the ward centers with searching radius set to 500m. [Figure 6] shows the result with 16 wards colored in grey – the ones that have no popular venues in 500m from the centers of the wards. For these wards, the searching radius should be set to at least 2,000m from the centroids.

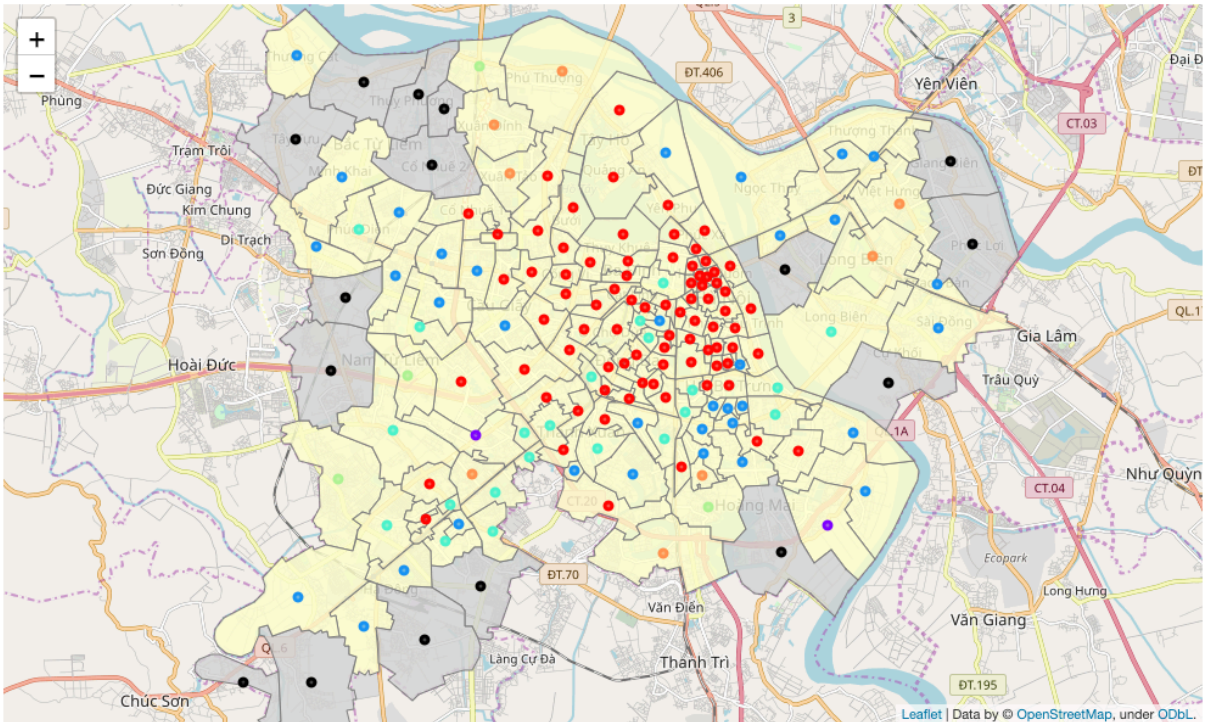


Figure 6. The grey colored wards are the ones that have no common venues returned by Foursquare if the radius is set to 500m

The largest ward is Phu Thuong with the area of ~7.47 sq. km while the smallest ones are Hang Bac and Hang Dao each is only 0.06 sq. km in area.

The difference in areas between the largest and smallest wards in the urban area of Hanoi is very big - 124 times. As mentioned above, using fixed radius for Foursquare API is not a good idea since if the radius is too small, Foursquare might find very few or even not find any venue at all near a given coordinate in large wards; while if the radius is too big, the nearby venues of a location will overlap with those of its neighbors, especially in the old quarter of Hanoi.

Therefore, I decided to set radius to be used with Foursquare API based on the quantiles of the ward areas as follows:

- Area = 0.06 sq. km - Radius = 300m
- 0.06 < Area <= 0.3475 sq. km - Radius = 500m
- 0.3475 < Area <= 0.86 sq. km - Radius = 800m
- 0.86 < Area <= 2.215 sq. km - Radius = 1500m

- 2.215 < Area <= 7.47 km - Radius = 2000m

Area of each ward was approximately calculated by python geometry.area.

The map bellow shows the difference in areas of the wards:

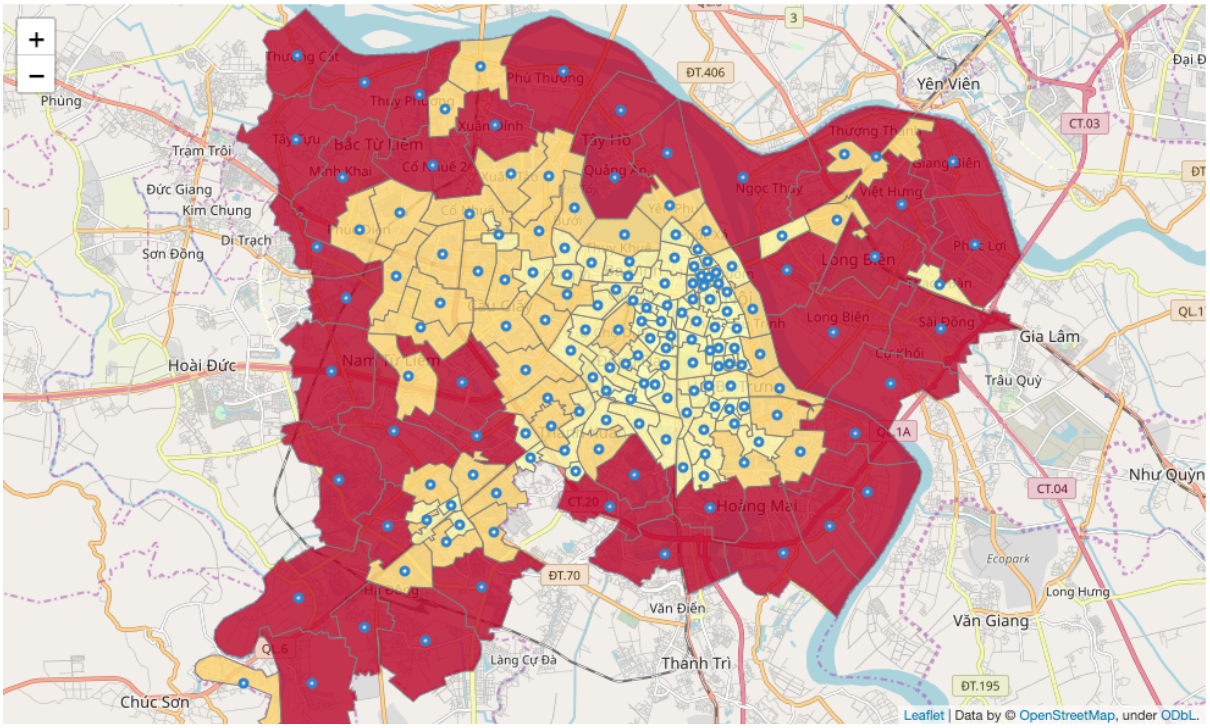


Figure 7. Hanoi urban wards colored by area scale

The final ward geopandas dataframe to be used to get nearby venues and to plot cluster maps is as follows:

	District	Ward	geometry	Latitude	Longitude	Area	radius
0	Bac Tu Liem	Co Nhue 1	POLYGON ((105.79462 21.04619, 105.79076 21.046...	21.051265	105.782960	1.97	1500
1	Bac Tu Liem	Co Nhue 2	POLYGON ((105.78443 21.06429, 105.78713 21.058...	21.064118	105.772726	3.76	2000
2	Bac Tu Liem	Duc Thang	POLYGON ((105.78256 21.08275, 105.78159 21.082...	21.078739	105.776018	1.09	1500
3	Bac Tu Liem	Dong Ngac	POLYGON ((105.78256 21.08275, 105.77682 21.086...	21.090057	105.786137	2.05	1500
4	Bac Tu Liem	Lien Mac	POLYGON ((105.77170 21.09449, 105.77061 21.089...	21.085823	105.753425	4.23	2000
5	Bac Tu Liem	Minh Khai	POLYGON ((105.73260 21.05441, 105.73428 21.056...	21.060784	105.747278	4.37	2000
6	Bac Tu Liem	Phu Dien	POLYGON ((105.77396 21.04789, 105.77150 21.045...	21.051546	105.763309	2.20	1500
7	Bac Tu Liem	Phuc Dien	POLYGON ((105.76146 21.04165, 105.76067 21.040...	21.046910	105.752069	1.92	1500
8	Bac Tu Liem	Tay Tuu	POLYGON ((105.72945 21.05656, 105.72884 21.056...	21.070666	105.734077	4.79	2000
9	Bac Tu Liem	Thuy	POLYGON ((105.77170 21.09449, 105.77672 21.096...	21.082668	105.768932	2.36	2000

Figure 8. Ward geopandas dataframe

There are 168 wards each with corresponding coordinates, area and searching radius. The results returned by Foursquare API (with limit set to 100 venues), includes 4,507 venues in 210 unique categories in total.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Co Nhue 1	21.051265	105.78296	Bánh Giò Hồ Tây	21.043694	105.790049	Food Truck
1	Co Nhue 1	21.051265	105.78296	Chợ Nghĩa Tân	21.044103	105.793873	Market
2	Co Nhue 1	21.051265	105.78296	Somerset Hoa Binh	21.046280	105.795193	Hotel
3	Co Nhue 1	21.051265	105.78296	Metro Cash & Carry	21.054612	105.780599	Supermarket
4	Co Nhue 1	21.051265	105.78296	Highlands Coffee	21.046310	105.795244	Café
5	Co Nhue 1	21.051265	105.78296	Bánh Mỳ Chảo Cột Điện	21.042417	105.793345	Sandwich Place
6	Co Nhue 1	21.051265	105.78296	Hat & Snack Rounds	21.040736	105.789069	Snack Place
7	Co Nhue 1	21.051265	105.78296	Siêu Thị Điện Máy HC	21.051528	105.781686	Electronics Store
8	Co Nhue 1	21.051265	105.78296	CGV Vincom BẮC TỪ LIÊM	21.053304	105.780262	Multiplex
9	Co Nhue 1	21.051265	105.78296	Sumo BBQ	21.046234	105.791102	Japanese Restaurant

Figure 9. List of all nearby common venues

The result was then transformed by one hot encoding, grouped by wards and calculating the mean of the frequency of occurrence of each venue category.
Then a dataframe including 10 most common venues in each ward was created as bellow:

	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

Figure 10. 10 most common venues in each ward

This most-common-venue dataframe was then used as input to the K-Means clustering algorithm to cluster the wards into different groups. After some experiments with different number of clusters, I decided to set the number of clusters to 3 since the result clusters would show the best the differences among them.

Cluster labeling

K-Means put 97 wards to the first cluster, 67 wards to the second cluster and ony 3 wards to the third one.

To give good labels to the clusters, I created bar charts to compare the numbers of the most common venues in each cluster [Figure 11][Figure 12][Figure 13].

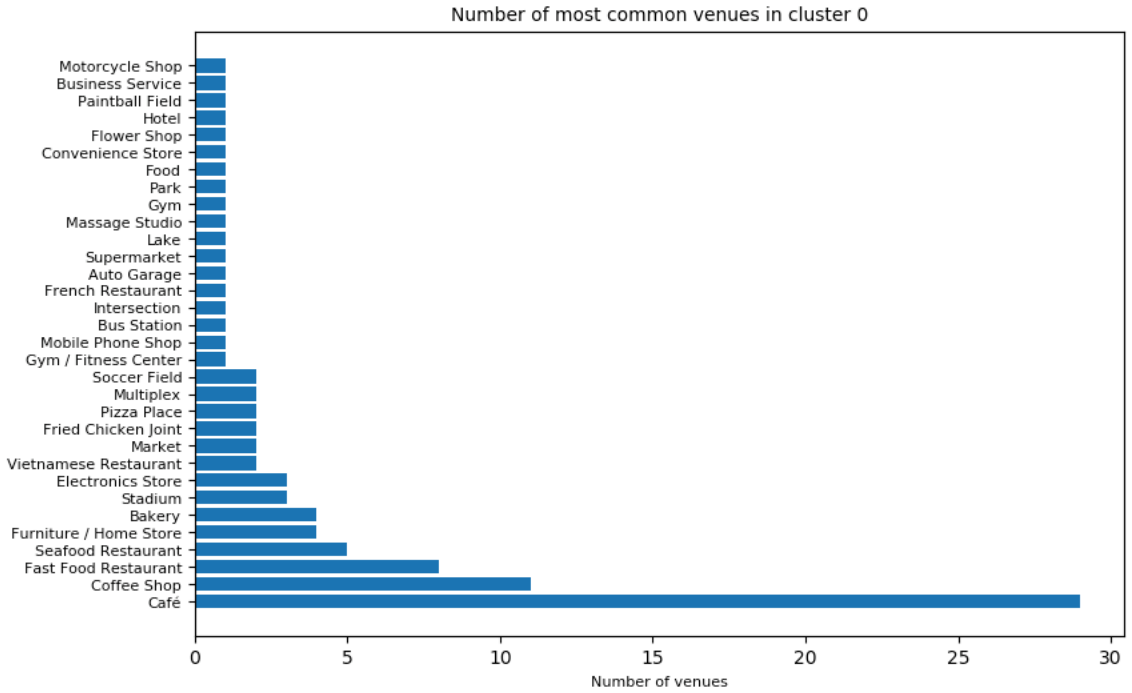


Figure 11. Comparison of numbers of most common venues in Cluster 0

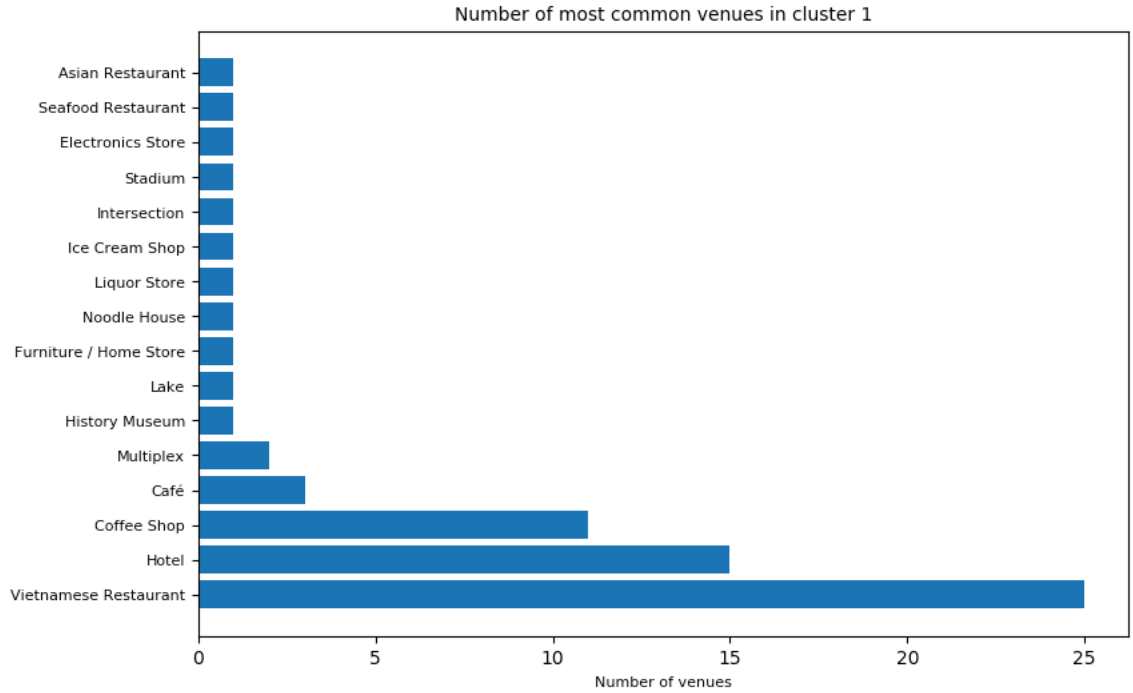


Figure 12. Comparison of numbers of most common venues in Cluster 1

There are only 3 wards in the third cluster and the 1st most common venues in the cluster do not tell much. So, to have better insights of the venues in the cluster, I plotted a chart using all 10 most common venues of the cluster.

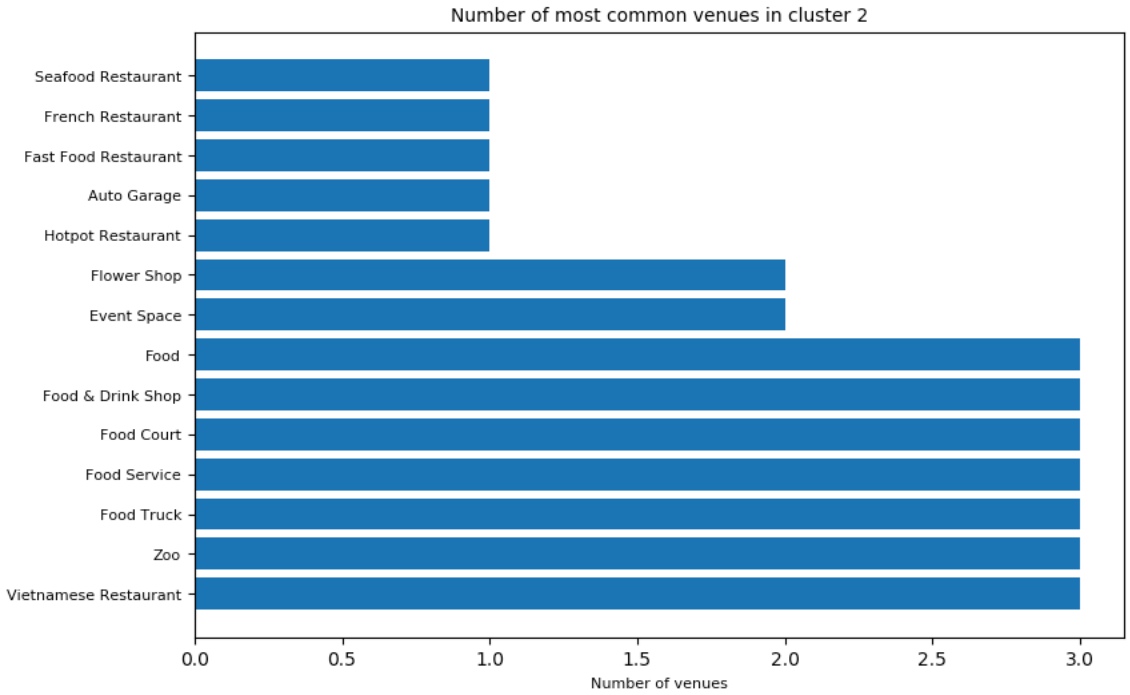


Figure 13. Comparison of numbers of most common venues in Cluster 2

The charts above show that the first cluster (cluster 0) includes more Café, Coffee Shops, Fast Food restaurants and Stores while Restaurants, Hotels and Coffee Shops are dominant in the second cluster (cluster 1). This also suggests that the venues in the second cluster are more high-end compared to those in the first cluster.

The third cluster (cluster 2) includes Restaurants, Zoos, Event Space, Food trucks and Food services. These are indications of outing places for outdoor events.

Now we can set appropriate labels to the clusters:

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

We will need also another label to assign to the location that does not belong to any cluster above (if the location does not have any nearby common venues in the given searching radius). Let's call it "Cluster 3: Not in any cluster".

Another interesting information that can be extracted from the result list of common venues is the average distance (the searching radius) from the wards' centers to the nearby venues in each cluster:

- Cluster 0: 1.314 km
- Cluster 1: 0.995 km
- Cluster 2: 1.666 km

C. Results

A population density map of Hanoi was created, including all urban and suburban districts. Users using the online version of the map can hover over the district polygons to show the district information.

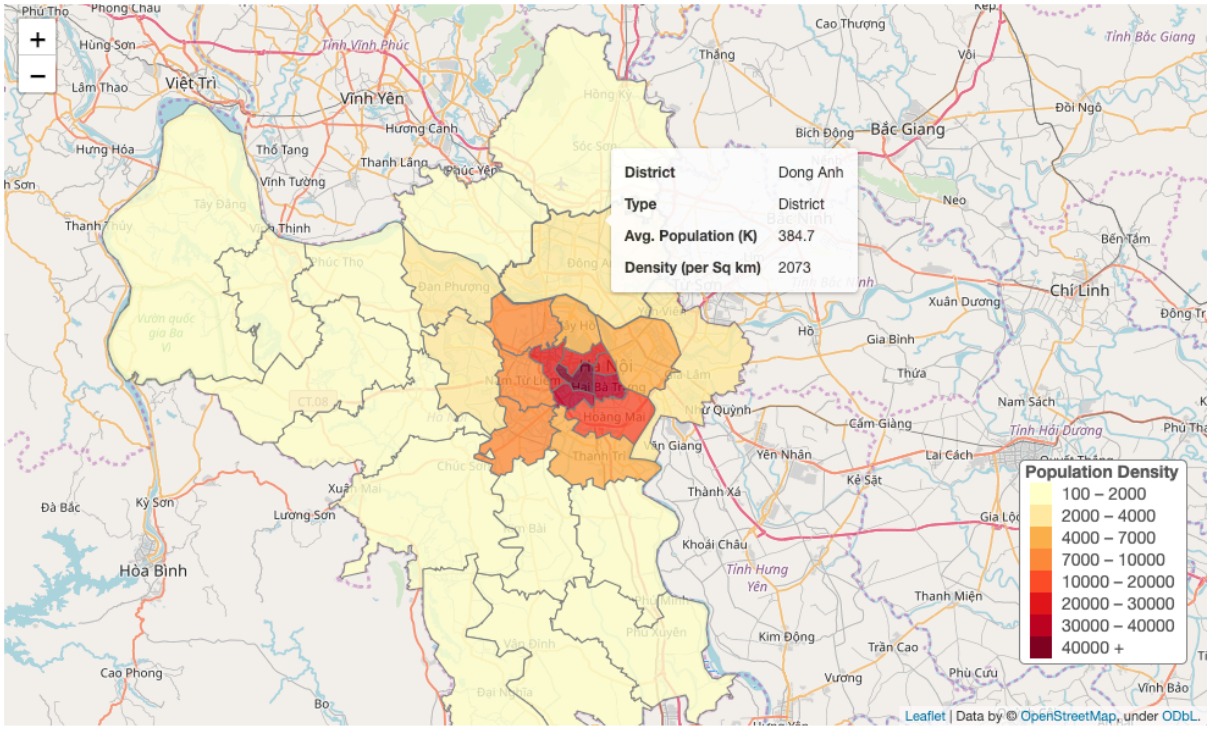


Figure 14. Hanoi Population density by districts (info updated 2018)

We can see from the map that more than half of the districts in Hanoi have fairly low population density of less than 2,000 people per square kilometer. There's only one suburban district - Thanh Tri - that has the density comparable to those of some urban districts.

The districts can be categorized into different groups based on population density as follows:

1. More than 40,000: Dong Da
2. 30,000 - 40,000: Hai Ba Trung, Thanh Xuan
3. 20,000 - 30,000: Ba Dinh, Hoan Kiem, Cau Giay
4. 10,000 - 20,000: Hoang Mai
5. 70,000 - 10,000: Bac Tu Liem, Nam Tu Liem, Ha Dong
6. 4,000 - 7,000: Tay Ho, Long Bien, Thanh Tri
7. Less than 4,000: the rest

When an investor comes, depending on the business, he can consider to choose the districts in each of the group above. For example, the most densely populated districts

might be good places to open restaurants/hotels/coffee shops but usually it comes with very high rental cost. If the business requires large land with fairly densed population, e.g. for hosting outdoor events, he can choose the districts in group 5-6 or 7, etc.

Now suppose that the investor has opened a first restaurant/shop in the urban area of Hanoi. It was very successful and he wants to find other locations to expand the business. With Foursquare API and utilizing K-Means clustering technique, I have created the map below that can help him to find similar locations to the current one in terms of popular venues around the location.

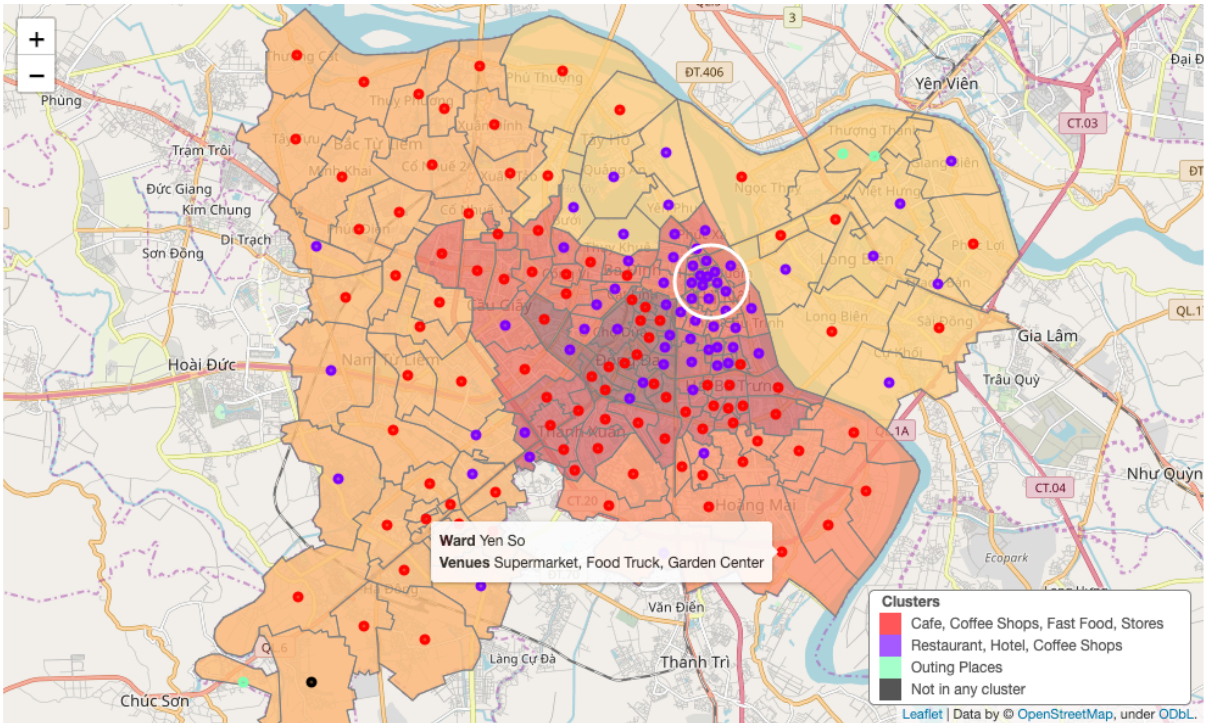


Figure 15. Visualization of clusters of common venues in the urban area with the white circle surrounding the old quarter (center) of Hanoi on the population density background map

The map clearly shows two dominant clusters - purple and red ones. The purples are places with high-end restaurants, hotels and coffee shops while the red ones have many Fast food and Café. Most of the purples are in small wards in the center of Hanoi (close to the old quarter - the white circle in the map) where the population is much denser, where many visitors come and stay and where there are many office buildings and shopping streets located. The red ones are located further to the west and south from the center, to the direction that the city has been expanded during the last few years.

The third cluster (light greens) includes wards at the edge of the urban area. Those wards are less densely populated, close to the rivers and suitable for outdoor events.

There is one ward shown at the bottom left of the map - Dong Mai - with the center colored in black as Foursquare returned no common venue for it, even with the searching radius

set to 2km. This shows that choosing appropriate value of searching radius is very important for the analysis.

This map will be very useful for investors to choose location to open their businesses in Hanoi and where to expand the chains after a successful one. The online version of this map was created in an interactive way such that hovering over a ward polygon will show the ward name, the district and the cluster that the ward belongs to; while hovering over a dot (ward center) will show the most common venues in the ward.

D. Discussion

The data that I have regarding district population was since in 2018. Hanoi is a rapid changing city with many apartments built further from the center. The maps will be useful if there's a way to update the population density yearly.

Foursquare API was used to search for nearby common venues of the wards' centers. Foursquare database is updated continuously, therefore the clusters and the maps shown above might change regularly. This project can be further updated with methods to automatically choose a good number of clusters and label them accordingly.

In the analysis above, I manually chose number of clusters to be 3. I actually did test with the Elbow method with number of clusters up to 100, however I did not see any clear elbow. This does not really mean clustering is not a suitable algorithm to solve the problem, but Elbow might not be a good method to be used for this dataset. Further study will be conducted with different method such as BIC and SSE.

E. Conclusion

In this project, I have demonstrated how to use different techniques, tools and algorithms to collect, explore, analyze and visualize public data on population density, area and nearby common venues in Hanoi. The result of the project can help investor to decide where to start their business in the city and what location will be good to expand in the future. Further studies and ways to make the project more useful was also discussed.