Setting up a Café in Ho Chi Minh City

A final report for the course "Applied Data Science Capstone" given by IBM on Coursera

Anh-Thi DINH

May 17, 2019

Contents

1 Problem's description

If you have a chance to visit Ho Chi Minh City (HCMC), you will see that it's one of the most active cities in the world. People here love to talk, to chat everywhere. The catering services in this city, therefore, very thriving.

We need to clarify the differences between a *coffee shop* and a *café*. Coffee shop has no similar connotations. From personal experience in the United States, a café serves meals, while a coffee shop usually just sells snacks (muffins, scones, shortbread). This is not strictly the case, and both usually serve coffee. In this project, we suppose to work only on the café.

Although there are already a lot of cafés in HCMC, their density between districts is not uniform. There are some districts containing too many café while there are less in some others. If we have some knowledge about the *population density*, the *housing price* in each district

coupling with an overview of *the number of café*, we can have a better idea to set up a new business there.

If we think of it by an investor, we expect to choose a place where the population density is high but less competitors. If the housing price in that place is low, it's more attractive to us.

By using Data Science and some geometric factors about the relation between districts in HCMC, we can give good answers of following questions to the investors so that they can have a better vision about not only the café but also about other venues in HCMC.

1. *How many venues in each district?* Answering this question gives us a better understanding about *the dynamic level* of a district.
2. *How many categories in each district?* Answering this question helps us know about *the diversity in business* of a district.
3. *How many venues in each category?* This question shows the magnitude of a category in a district.
4. *What are the most popular categories in each district?* If investors change their mind to focus on other commercial fields instead of opening a café.
5. *How many clusters we can use to categorize the districts based on the popularity of cafés?*
6. *In which districts, the average housing price is low and the number of cafés is low also?*
7. *Where there are many people but less cafés?*
8. Visualize all information on the map so that we can have a better look on what we want to find the answers!

2 Data presentation

In order to explor the questions, we need to use following data in the research.

1. List of Ho Chi Minh City administrative units from Wikipedia. It gives us a list of all urban districts of HCMC with their area (in Km^2), population (in 2015) and the density of each district (people/Km^2). The list is given in `http://bit.ly/30r0yU8`.
2. List of the coordinates (latitude, longitude) of all urban districts in HCMC. This list can be generated based on the name of each district and package *geopy.geocoders.Nominatim*.
3. List of average housing prices per m^2 in HCMC. The list is frequently updated in `https://mogi.vn/gia-nha-dat`.
4. A *.json* file contains all coordinates where we use it to create a choropleth map of Housing Sales Price Index of HCMC. I create this file by myself using `https://nominatim.openstreetmap.org`.

3 Methodology

1. First, we need to collect the data by scraping the table of HCMC units on the wikipedia page and the average housing price (AHP) on a website. The *BeautifulSoup* package is very useful in this case.

2. The column *Density* is calculated later based on columns *Population* and *Area* of each district.

3. Throughout the project, we use *numpy* and *pandas* packages to manipulate dataframes.

4. We use *geopy.geocoders.Nominatim* to get the coordinates of districts and add them to the main data frame.

5. We use *folium* package to visualize the HCMC map with its districts. The central coordinate of each district will be represented as a small circle on top of the city map.

6. We use *Foursquare API* to explore the venues in each district and segment the districts based on them.

7. For clustering the "Café" venues between districts, we use *K-Means Clustering* method and the package *scikit-learn* will help us implement the algorithm on our data. In order to indicate how many K for the method, we try with 10 different values of K from 1 to 10 and use the "elbow" method to choose the most appropriate one.

8. In order to visualize the charts, we use pakage *matplotlib*.

9. We use again the package *folium* to visualize the clusters on the main map and the choropleth map of AHP.

4 Results

We will answer all questions in the Section 1.

4.1 The main data frame

After scraping all information from the internet, we have a table like in Figure 1.

4.2 Venues per District

We plot a chart in order to compare visually the different of number of venues between districts. This chart is shown in Figure 2.

From this chart, we see that the districts **1, 10, 3, 5, Phu Nhuan** are the most dynamic ones. For the districts 1, 3 or 5, they are three center districts of HCMC, thus the high number of venues in these districts are not so strange. We pay attention to **Phu Nhuan** which is not a center district. We also notice on the **District 4** which has more venues than the others although in reality, who live in HCMC will think that is strange.

	District	Subdistrict	Area (km2)	Population 2015	Density (pop/m2)	Average Housing Price (1M VND)	Latitude	Longitude
0	1	10 wards	7.73	193632	25049.418	384	10.774540	106.699184
1	2	11 wards	49.74	147168	2958.745	58.8	10.791116	106.736729
2	3	14 wards	4.92	196333	39905.081	236	10.783529	106.687098
3	4	15 wards	4.18	186727	44671.531	70.3	10.759243	106.704890
4	5	15 wards	4.27	178615	41830.211	241	10.756129	106.670376
5	6	14 wards	7.19	258945	36014.604	95.5	10.746928	106.634495
6	7	10 wards	35.69	310178	8690.894	74.9	10.736573	106.722432
7	8	16 wards	19.18	431969	22521.846	56	10.740400	106.665843
8	9	13 wards	114	290620	2549.298	41.2	10.824543	106.818015
9	10	15 wards	5.72	238558	41705.944	203	10.773198	106.667833
10	11	16 wards	5.14	230596	44863.035	154	10.764208	106.643282
11	12	11 wards	52.78	510326	9668.928	39.7	10.867233	106.653930
12	Go Vap	16 wards	19.74	634146	32124.924	95	10.840150	106.671083
13	Tan Binh	15 wards	22.38	459029	20510.679	136	10.797979	106.653805
14	Tan Phu	11 wards	16.06	464493	28922.354	97.3	10.791640	106.627302
15	Binh Thanh	20 wards	20.76	487985	23506.021	136	10.804659	106.707848
16	Phu Nhuan	15 wards	4.88	182477	37392.828	168	10.800118	106.677042
17	Thu Duc	12 wards	49.76	528413	10619.232	49.9	10.852588	106.755838
18	Binh Tan	10 wards	51.89	686474	13229.408	57.2	10.749809	106.605664

Figure 1: The main data frame.

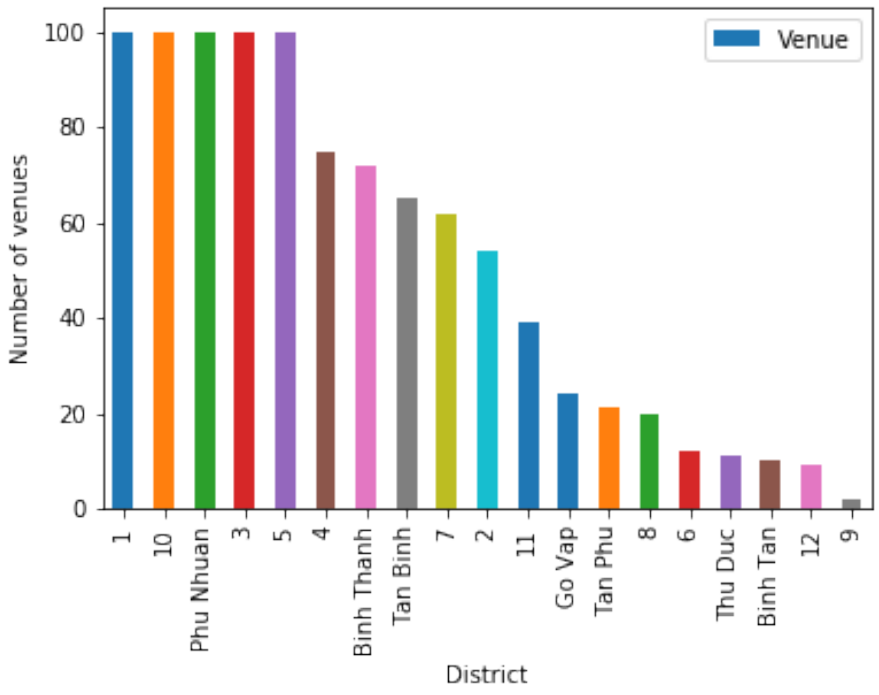

Figure 2: The number venues in each district.

4.3 Categories per District

The chart in Figure 3 gives us an overview of the number of categories in each district.

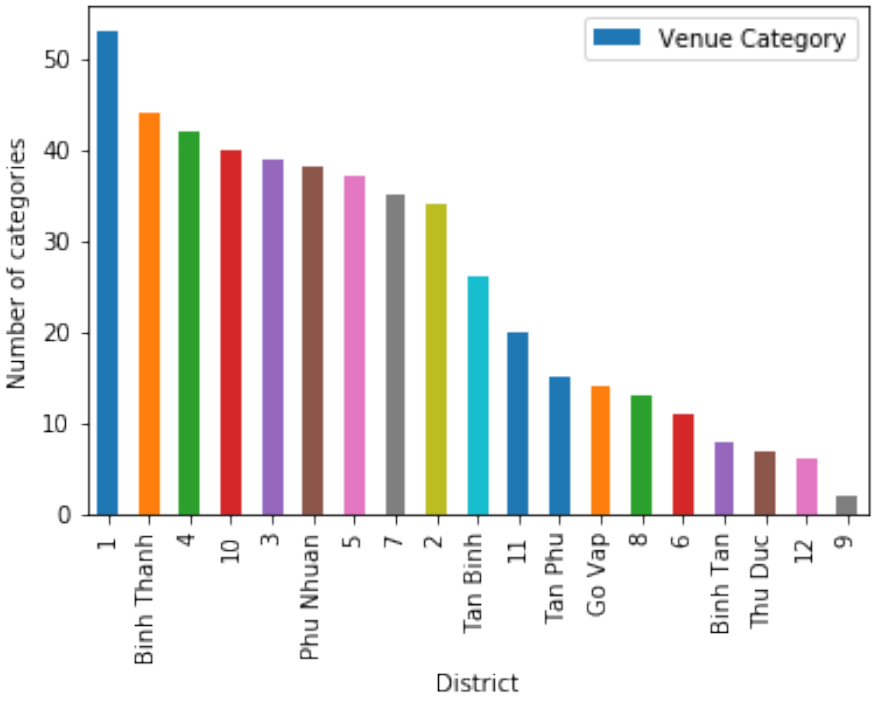

Figure 3: The number categories in each district.

Again, the district 1 wins the top. However in this time, the **Binh Thanh** runs the second instead of district 3 or 5 like in Figure 2. The **District 4** is still very diversity. The reason for that there are many venues but les categories in some districts is maybe there are some principle categories in these districts. Those principle categories play the major role in the commercial activities of these districts.

4.4 Venues per Category

Look at the most 5 categories, we have *Vietnamese Restaurant* (133), *Café* (127), *Coffee Shop* (72), *Seafood Restaurant* (33), *Asian Restaurant* (29). **The café** is the main category in the drinks business with 127 different venues!

4.5 Top 10 venue categories in each district

Figure 4 shows us the most 10 categories in each district. For less copetition, we can choose districts whose fitst most common venue is not café. For examples, districts 1, 10, 2, 3, 4, 5.

4.6 How many clusters?

We consider the data relating to category "café" only. We want to cluster them into several groups. First, we need to determine the number of groups (or K for the K-means method). Using the elbow method with different values of K, Figure 5 shows that **3** is the best choice.

	District	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	1	Vietnamese Restaurant	Hotel	Café	Coffee Shop	French Restaurant	Massage Studio	Bar	Clothing Store	Thai Restaurant	Middle Eastern Restaurant
1	10	Vietnamese Restaurant	Café	Coffee Shop	Dessert Shop	Seafood Restaurant	Spa	Ice Cream Shop	Music Venue	Market	Bookstore
2	11	Café	Vietnamese Restaurant	Chinese Restaurant	Coffee Shop	Seafood Restaurant	Dumpling Restaurant	Cantonese Restaurant	American Restaurant	Gym / Fitness Center	French Restaurant
3	12	Café	Vietnamese Restaurant	Department Store	Diner	Seafood Restaurant	Asian Restaurant	Flower Shop	French Restaurant	Food Truck	Food Court
4	2	Vietnamese Restaurant	Café	BBQ Joint	Restaurant	Shopping Mall	Multiplex	Asian Restaurant	Coffee Shop	Thai Restaurant	Bistro
5	3	Vietnamese Restaurant	Coffee Shop	Asian Restaurant	Vegetarian / Vegan Restaurant	Café	Hotel	Japanese Restaurant	Breakfast Spot	Yoga Studio	Restaurant
6	4	Vietnamese Restaurant	Seafood Restaurant	Coffee Shop	Snack Place	Café	Food	Hotel	Flea Market	Burger Joint	Bar
7	5	Vietnamese Restaurant	Chinese Restaurant	Coffee Shop	Café	Dim Sum Restaurant	Dessert Shop	Noodle House	Asian Restaurant	Seafood Restaurant	Vegetarian / Vegan Restaurant
8	6	Café	Department Store	Movie Theater	Fast Food Restaurant	Pizza Place	Coffee Shop	Food Court	Boutique	Dessert Shop	Asian Restaurant
9	7	Café	Vietnamese Restaurant	Japanese Restaurant	Gym / Fitness Center	Sushi Restaurant	Flea Market	Seafood Restaurant	Steakhouse	Restaurant	Spa
10	8	Vietnamese Restaurant	Dim Sum Restaurant	Dessert Shop	Coffee Shop	Chinese Restaurant	Grocery Store	Plaza	Fast Food Restaurant	Café	Food Truck
11	9	Seafood Restaurant	Racetrack	Yoga Studio	Flea Market	French Restaurant	Food Truck	Food Court	Food	Flower Shop	Farmers Market
12	Binh Tan	Coffee Shop	Shopping Mall	Multiplex	Café	Food Court	Pizza Place	Bubble Tea Shop	Fast Food Restaurant	Flea Market	Fried Chicken Joint
13	Binh Thanh	Café	Coffee Shop	Vietnamese Restaurant	Seafood Restaurant	Soup Place	Bakery	Multiplex	Fast Food Restaurant	Pizza Place	Diner
14	Go Vap	Café	Multiplex	Market	Coffee Shop	Shopping Mall	Department Store	Warehouse Store	Vietnamese Restaurant	Farmers Market	Asian Restaurant
15	Phu Nhuan	Café	Coffee Shop	Vietnamese Restaurant	Hotel	Vegetarian / Vegan Restaurant	Japanese Restaurant	Diner	Seafood Restaurant	Chinese Restaurant	Spa
16	Tan Binh	Vietnamese Restaurant	Café	Coffee Shop	Noodle House	Pizza Place	Seafood Restaurant	Flea Market	Multiplex	Asian Restaurant	Hotel
17	Tan Phu	Café	Japanese Restaurant	Diner	Supermarket	Restaurant	Coffee Shop	Flea Market	Shopping Mall	Shopping Plaza	Cafeteria
18	Thu Duc	Café	Jewelry Store	Shopping Mall	Multiplex	Tennis Court	Pizza Place	Diner	Farmers Market	Electronics Store	Fast Food Restaurant

Figure 4: Top 10 venue categories for each district.

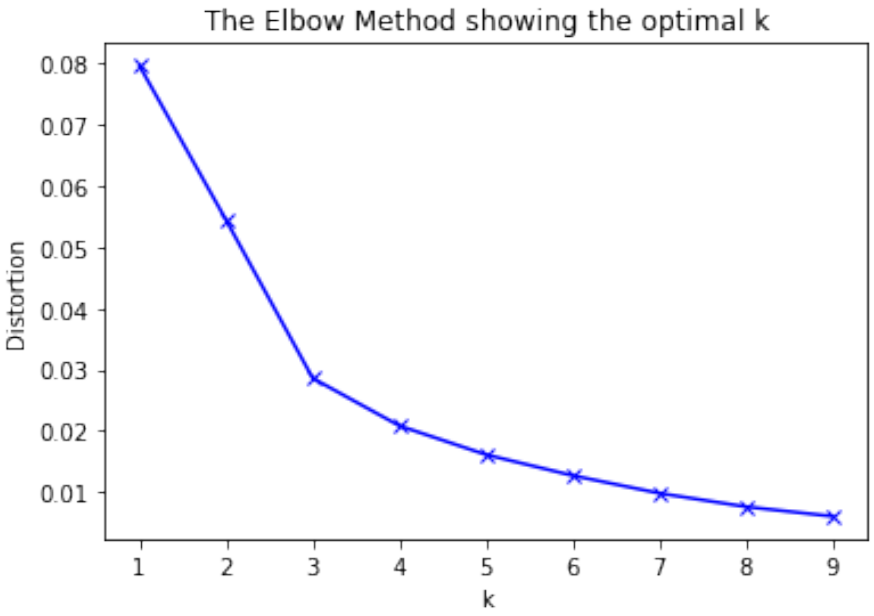

Figure 5: The optimal number of groups/clusters.

We can name the clusters like these,

- **Cluster 0** : There are not many café shops in these districts.
- **Cluster 1** : There are a lot of café shops in these districts.
- **Cluster 2** : The number of café shops in these districts is medium.

Figure 6 illustrates the clusters of all urban districts in HCMC. With this map, we can easily distinguish the clusters between districts.

4.7 AHP vs the number of café

Look back to the average housing price table (AVH), we categorize them into 4 groups (unit: million VND). Figure 7 indicates that the low price housing take the majority. We need to focus on the **Low** and **Medium** housing price to set up our business.

- **Low** : $30 < AHP \leq 100$.
- **Medium** : $100 < AHP \leq 200$.
- **High** : $200 < AHP \leq 300$.
- **Very high** : $300 \leq AHP$.

Look at Figure 8, we focus on:

- **Low** AHP & **not many café** (cluster 0) : district 2, **district 4**, district 8, district 9 and Binh Tan.
- **Low** AHP & medium number of café : district 12, Go Vap, Thu Duc.

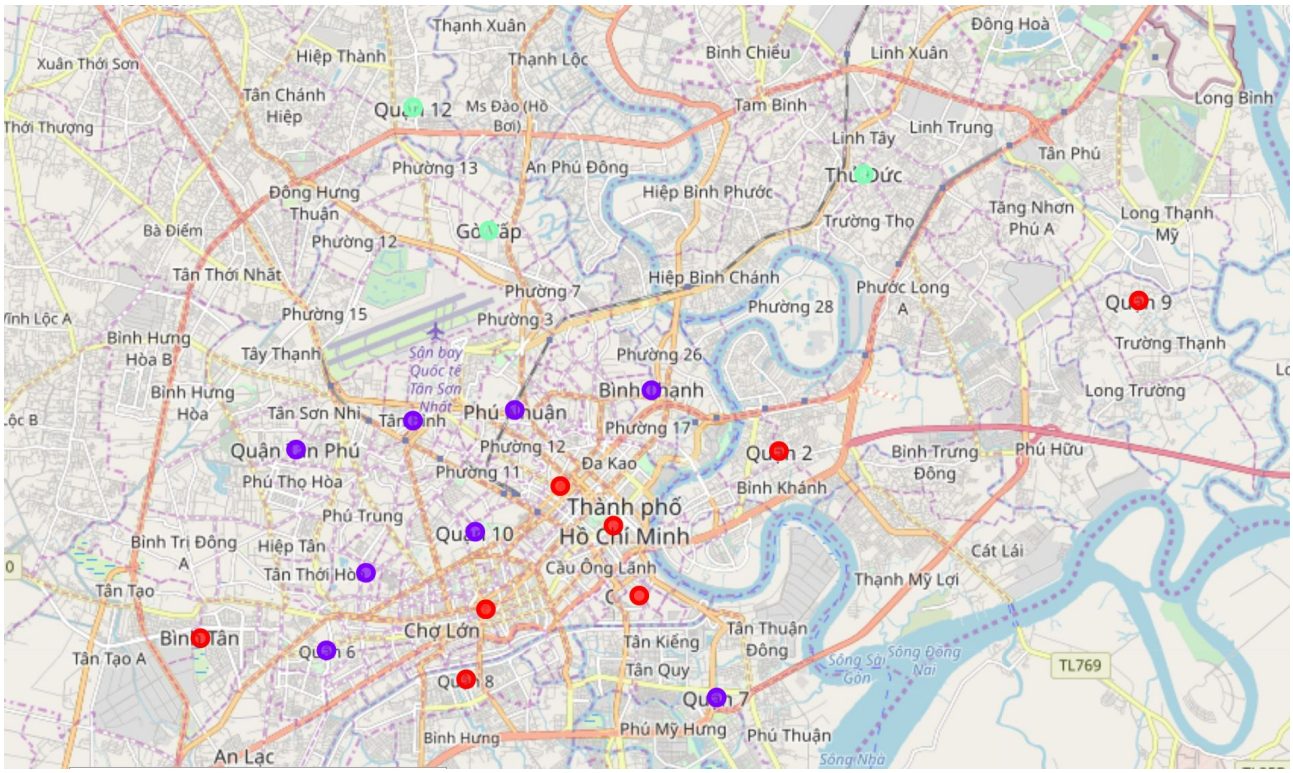

Figure 6: The maps of clusters. Cluster 0 (Red), Cluster 1 (Violet), Cluster 2 (Cyan).

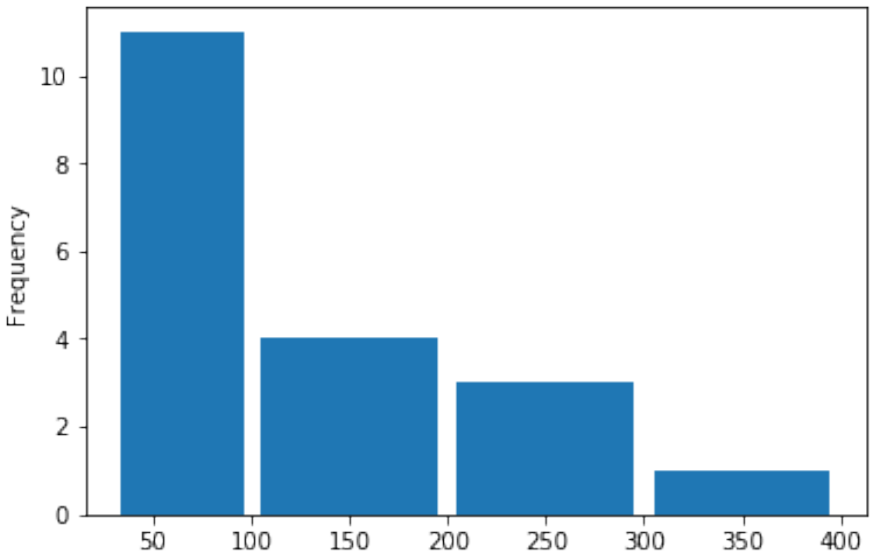

Figure 7: The distribution of AHP.

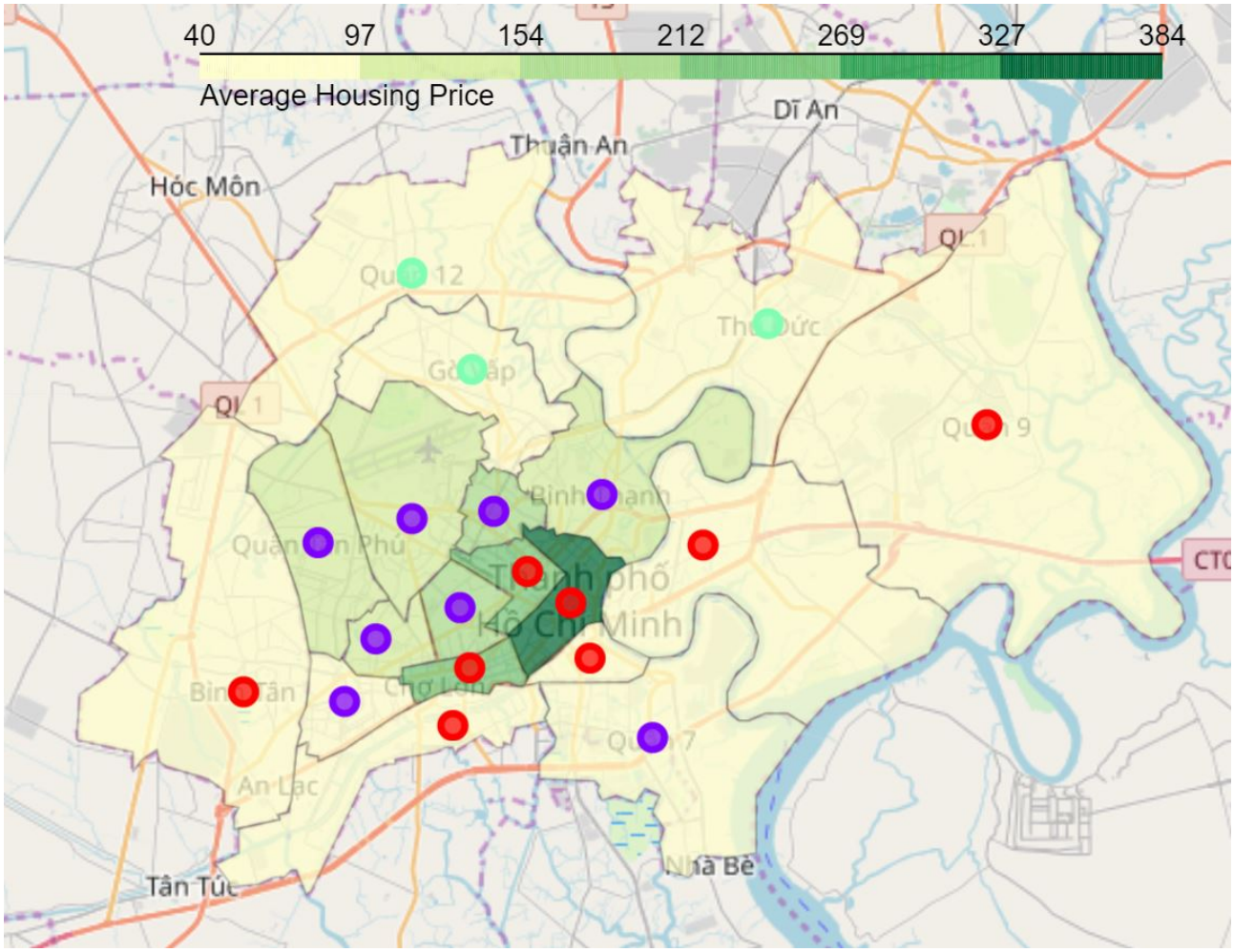

Figure 8: The couple maps of AHP and the clusters given in Section 4.6.

4.8 Population density vs the number of café

We should not rely only on the relationship between AHP and clusters. For example, district 9 has almost no café and it has also very low AHP but in reality, this district contains many industry zones and there are not many people living around here. That's why we need to consider also the density of each district. Just think that, if there are not enough people to come to our café, how can we make a profit?

Figure 9 gives us a full picture about the relation between population density and the clusters.

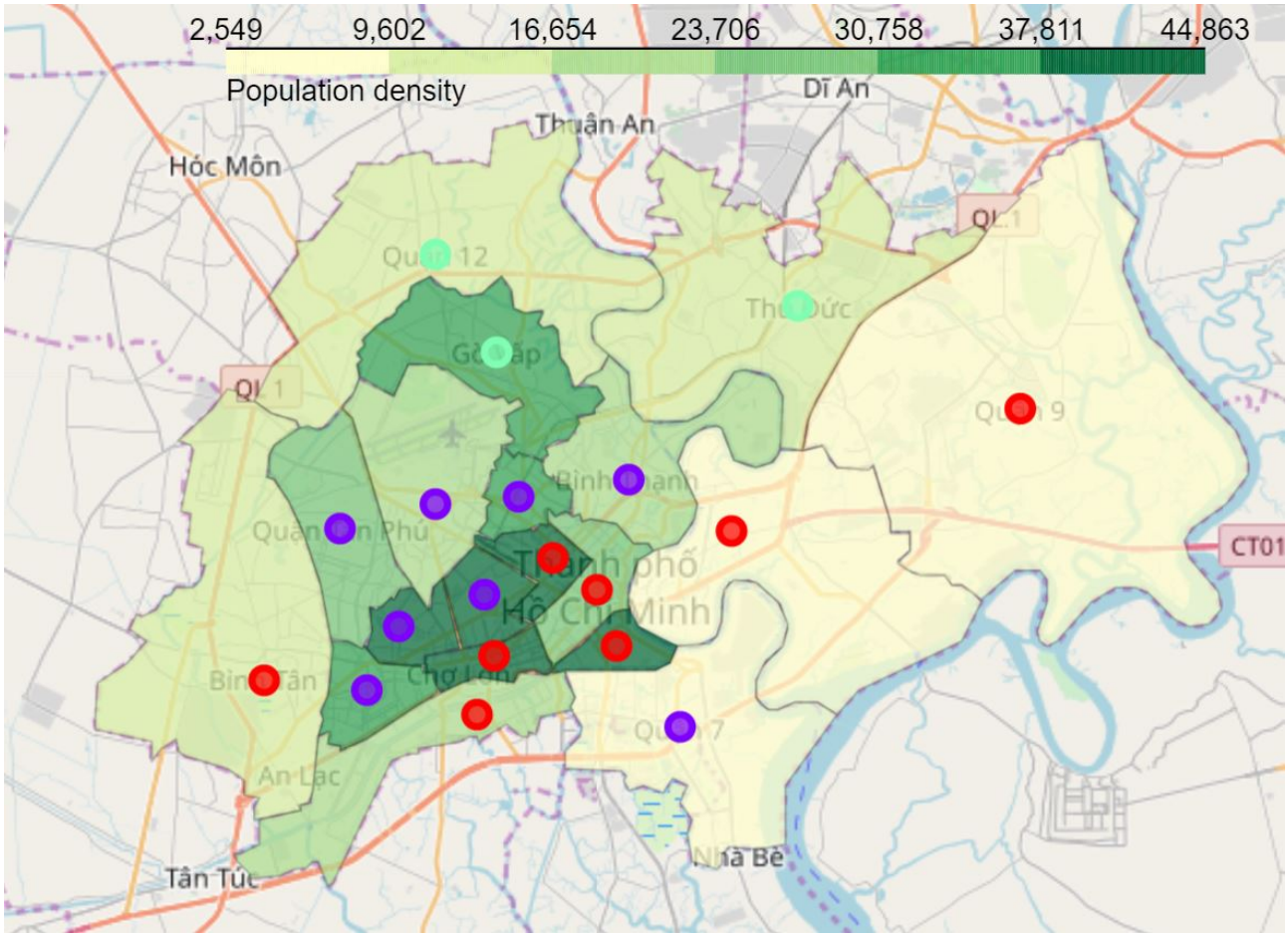

Figure 9: The couple maps of AHP and the population density of each district.

We focus on:

- **High density + not many café** : district 3, **district 4**, district 5.

5 Conclusion

From all above results, we conclude that, the best place for us to set up a new café is in **district 4** because there are a lot of people living there (high density), there are not many already-working café (cluster 0) and the average housing price is low.