A. Introduction:

1. Background

Ho Chi Minh City (Vietnamese: Thành phố Hồ Chí Minh;] , abbreviated as HCMC, also known by its former name Saigon, is the most populous city in Vietnam with a population of 9 million (13 million in the metropolitan area) as of 2019. Located in southeastern Vietnam, the metropolis surrounds the Saigon River and covers about 2,061 square kilometres (796 square miles). From 1955 to 1975, Saigon was the capital of the Republic of Vietnam, commonly known as South Vietnam.

Ho Chi Minh City is the financial centre of Vietnam and is classified as a Beta+ World City by Globalization and World Cities Research Network. It is home to the Ho Chi Minh City Stock Exchange, the largest stock exchange by total market capitalization in Vietnam and the headquarters of many national and international banks and companies.

As a major gateway to Vietnam, the city received over 8.6 million international visitors in 2019, of which 4.1 million were overnight visitors. The main passenger airport serving the metropolitan area is Tan Son Nhat International Airport, the busiest airport in Vietnam handling over 40 million passengers in 2019

Hanoi is the capital of Vietnam. It covers an area of 3,328.9 square kilometres (1,285 sq mi). With an estimated population of 8.1 million as of 2019, it is the second largest city in Vietnam. The metropolitan area, encompassing nine additional neighbouring provinces, has an estimated population of 16 million. Located in the central area of the Red River Delta, Hanoi is the commercial, cultural, and educational centre of Northern Vietnam. Having an estimated nominal GDP of US$32.8 billion, it is the second most productive economic centre of Vietnam, following Ho Chi Minh City.

Originally a small settlement along the banks of the Red River, the city was founded as Thăng Long, the capital of Imperial Vietnam, in 1010 by monarch Lý Thái Tổ. Thăng Long would remain the most important political and cultural centre of Vietnam until 1802, when the Nguyễn dynasty, the last imperial dynasty of Vietnam, moved the capital to Huế. Thăng Long renamed to its current name Hanoi in 1831. In 1873, Hanoi was conquered by the French, and from 1883 to 1945, the city was the administrative centre of French Indochina. The French colonisation left a lasting impact on the city's architecture that is visible today, as showcased through the juxtaposition of French-styled avenues, bridges, buildings, and traditional Vietnamese architecture.

2. Problem Description:

The best pho in Hanoi is arguably found at Pho Thin, a nondescript eatery on Lo Duc in the city's French Quarter. When it opened in 1979, Vietnam was still recovering from the war and food was rationed. To add more flavour to the standard raw beef, broth and herb combination, owner Nguyen Trong Thin made the controversial call to stir-fry the beef in garlic before adding it to the soup.

In the years since, more Pho Thin outposts have opened across Asia, including wider Vietnam and Japan. Last week, Australia's first Pho Thin opened softly (it opens officially September 19) in Hardware Lane.

Now the owner wants to open a restaurant in HCM. Wouldn't it be great if you are able to determine neighborhoods on the other side of the country that are the same as your current neighborhood. Phothin has a branch in Ba Dinh District, so I will combine features of Ba Dinh and districts of HCMC then clusters, looking for which one is closest similar to Ba Dinh district.

3. Objective

The aim of this report is to study and analyze the combined neighborhoods of Ho Chi Minh city and Ba Dinh (where the original district PhoThin located) then group them into similar clusters and, to analyze those clusters to gather meaningful information. That information can be used to find out neighborhoods that are the same as your current neighborhood or at least similar.

B. Data Description:

To consider the objective stated above, we can list the below data sources used for the analysis.

1. Ho Chi Minh Neighborhood Data: The following Wikipedia page was scraped to pull out the necessary information: https://en.wikipedia.org/wiki/Ho_Chi_Minh_City

	District	Subdistrict	Area (km2)	Population 2015	Density (pop/Km2)
0	1	10 wards	7.73	193632	25049.418
1	2	11 wards	49.74	147168	2958.745
2	3	14 wards	4.92	196333	39905.081
3	4	15 wards	4.18	186727	44671.531
4	5	15 wards	4.27	178615	41830.211
5	6	14 wards	7.19	258945	36014.604
6	7	10 wards	35.69	310178	8690.894
7	8	16 wards	19.18	431969	22521.846
8	9	13 wards	114.00	290620	2549.298
9	10	15 wards	5.72	238558	41705.944
10	11	16 wards	5.14	230596	44863.035
11	12	11 wards	52.78	510326	9668.928
12	Go Vap	16 wards	19.74	634146	32124.924
13	Tan Binh	15 wards	22.38	459029	20510.679
14	Tan Phu	11 wards	16.06	464493	28922.354
15	Binh Thanh	20 wards	20.76	487985	23506.021
16	Phu Nhuan	15 wards	4.88	182477	37392.828
17	Thu Duc	12 wards	49.76	528413	10619.232
18	Binh Tan	10 wards	51.89	686474	13229.408

2. The information obtained i.e. the table of postal codes was transformed into a pandas data frame for further analysis. Coordinate data for each Neighborhood in HCM city: The following csv file gave us the geographical coordinates of each postal code:

	District	Subdistrict	Area (km2)	Population 2015	Density (pop/Km2)	Average Housing Price (1M VND)/m2	Latitude	Longitude
0	1	10 wards	7.73	193632	25049.418	451	10.774540	106.699184
1	2	11 wards	49.74	147168	2958.745	88.5	10.791116	106.736729
2	3	14 wards	4.92	196333	39905.081	282	10.783529	106.687098
3	4	15 wards	4.18	186727	44671.531	93.6	10.759243	106.704890
4	5	15 wards	4.27	178615	41830.211	260	10.756129	106.670376
5	6	14 wards	7.19	258945	36014.604	115	10.746928	106.634495
6	7	10 wards	35.69	310178	8690.894	88.1	10.736573	106.722432
7	8	16 wards	19.18	431969	22521.846	69.4	10.740400	106.665843
8	9	13 wards	114.00	290620	2549.298	48.9	10.824543	106.818015
9	10	15 wards	5.72	238558	41705.944	217	10.773198	106.667833
10	11	16 wards	5.14	230596	44863.035	167	10.764208	106.643282
11	12	11 wards	52.78	510326	9668.928	46.3	10.867233	106.653930
12	Go Vap	16 wards	19.74	634146	32124.924	99.4	10.840150	106.671083
13	Tan Binh	15 wards	22.38	459029	20510.679	153	10.797979	106.653805
14	Tan Phu	11 wards	16.06	464493	28922.354	106	10.791640	106.627302
15	Binh Thanh	20 wards	20.76	487985	23506.021	137	10.804659	106.707848
16	Phu Nhuan	15 wards	4.88	182477	37392.828	190	10.800118	106.677042
17	Thu Duc	12 wards	49.76	528413	10619.232	66.1	10.852588	106.755838
18	Binh Tan	10 wards	51.89	686474	13229.408	66.6	10.749809	106.605664

3. List of average housing prices per m2 in HCMC and Ba Dinh. The list is frequently updated in https://mogi.vn/gia-nha-dat.

	District	Average Housing Price (1M VND)/m2
0	1	451
1	2	88.5
2	3	282
3	4	93.6
4	5	260
5	6	115
6	7	88.1
7	8	69.4
8	9	48.9
9	10	217
10	11	167

C. Methodology:

1._Scrape the Wikipedia page and gather data into a Pandas dataframeTo start with our analysis, we used the BeautifulSoup package to transform the data in the table on the Wikipedia page into the below pandas dataframe.

2. Generating a map of Ho Chi Minh city and plotting the Neighborhood data on it We use folium package to visualize the HCMC map with its districts. The central coor-dinate of each district will be represented as a small circle on top of the city map.

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

4. Analyze each neighborhood. We use One Hot Encoding, use the neighborhood to group data, and find out the top ten venues present in each neighborhood.For clustering the Vietnamese Restaurant" 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.

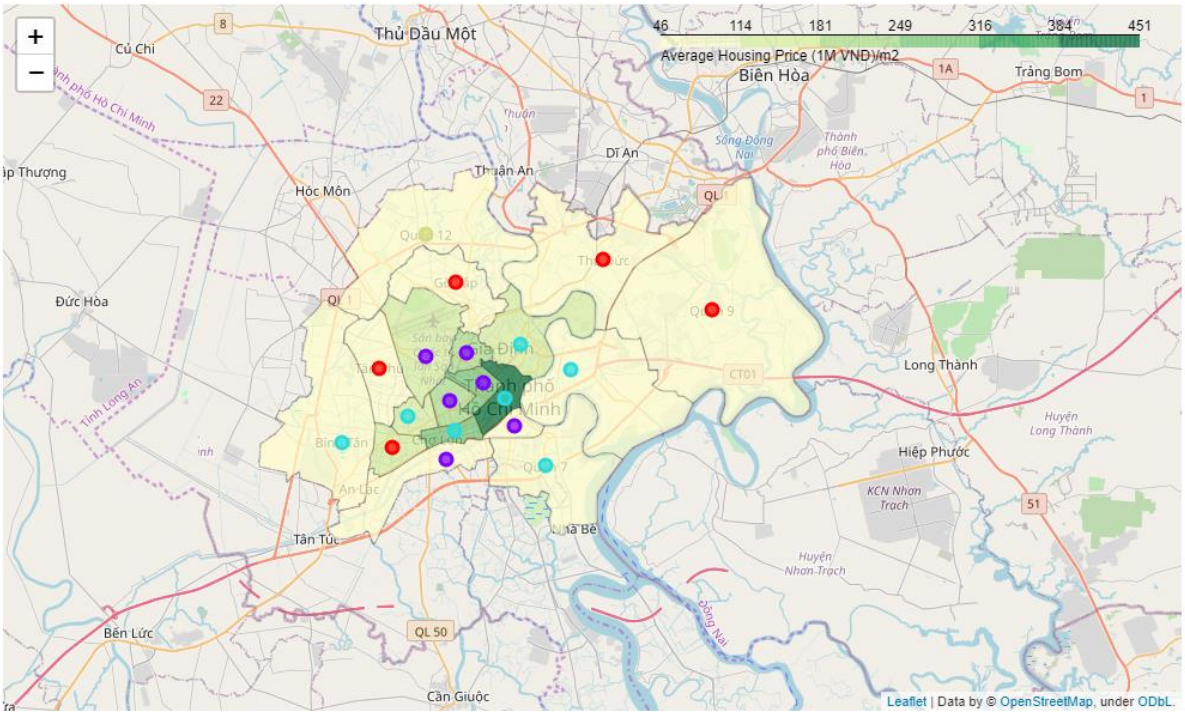
D. Results:

1. Adding the Cluster Labels to the Venue Data The below table depicts the clustered data along with the top 10 most common venues in that cluster.

	District	Vietnamese Restaurant	Cluster Labels	Population 2015	Density (pop/Km2)	Average Housing Price (1M VND)/m2	Latitude	Longitude
19	Thu Duc	0.000000	0	528413	10619.23200	66.1	10.852588	106.755838
15	Go Vap	0.032258	0	634146	32124.92400	99.4	10.840150	106.671083
8	6	0.047619	0	258945	36014.60400	115.0	10.746928	106.634495
18	Tan Phu	0.000000	0	464493	28922.35400	106.0	10.791640	106.627302
11	9	0.000000	0	290620	2549.29800	48.9	10.824543	106.818015
1	10	0.190000	1	238558	41705.94400	217.0	10.773198	106.667833
17	Tan Binh	0.229730	1	459029	20510.67900	153.0	10.797979	106.653805
16	Phu Nhuan	0.150000	1	182477	37392.82800	190.0	10.800118	106.677042
5	3	0.150000	1	196333	39905.08100	282.0	10.783529	106.687098
6	4	0.159420	1	186727	44671.53100	93.6	10.759243	106.704890
10	8	0.166667	1	431969	22521.84600	69.4	10.740400	106.665843
14	Binh Thanh	0.071429	2	487985	23506.02100	137.0	10.804659	106.707848
13	Binh Tan	0.100000	2	686474	13229.40800	66.6	10.749809	106.605664
0	1	0.130000	2	193632	25049.41800	451.0	10.774540	106.699184
7	5	0.130000	2	178615	41830.21100	260.0	10.756129	106.670376
4	2	0.103448	2	147168	2958.74500	88.5	10.791116	106.736729
2	11	0.128205	2	230596	44863.03500	167.0	10.764208	106.643282
12	Ba Dinh	0.070000	2	247100	26829.53312	167.0	21.035800	105.821700
9	7	0.083333	2	310178	8690.89400	88.1	10.736573	106.722432
3	12	0.333333	3	510326	9668.92800	46.3	10.867233	106.653930

2. Visualizing the resulting Clusters

We use the matplotlib and folium packages to visualize the clusters on a map of Ho Chi Minh city



The intent with which analysis was carried out was to find out similar neighborhoods for a person relocating within the city.

As we analyze the results section, we can analyze the clusters and see similar neighborhoods in different parts of the city. For example, if we compare the different neighborhoods clustered in cluster 2.

14	Binh Thanh	0.071429	2	487985	23506.02100	137.0	10.804659	106.707848	Medium
13	Binh Tan	0.100000	2	686474	13229.40800	66.6	10.749809	106.605664	Low
0	1	0.130000	2	193632	25049.41800	451.0	10.774540	106.699184	Very High
7	5	0.130000	2	178615	41830.21100	260.0	10.756129	106.670376	High
4	2	0.103448	2	147168	2958.74500	88.5	10.791116	106.736729	Low
2	11	0.128205	2	230596	44863.03500	167.0	10.764208	106.643282	Medium
12	Ba Dinh	0.070000	2	247100	26829.53312	167.0	21.035800	105.821700	Medium
9	7	0.083333	2	310178	8690.89400	88.1	10.736573	106.722432	Low

F. Conclusion:

From all above results, we conclude that, the best place for us to set up a new restaurant is in district Binh Tan because there are a lot of people living there (high density), there are alot of shopping malls, Multiplexies, there are not many already-working Vietnamese Restaurant and the average housing price is low.