Venues Data Analysis of CanGio, HoChiMinh city
Khuất Thùy Phương
March, 04, 2019

1. Introduction/Business Problem

- HoChiMinh city is the largest and most populous city in Vietnam. Currently, according to statistics in 2018, the city has about 14 million people and it has a population density of 4097 people per square kilometer, an area of about 2095 kilometer. Born and raised in this city, I decided to choose HoChiMinh city in my project. The city is divided into 24 districts. However, the population is unevenly distributed, with many districts overloaded, there are also sparsely populated districts.
- With the population density and area mentioned above, we can see that this is a city with a high population density. When considering the population density of each county, it was found that some districts had very high population densities, small areas, high land price such as district 11: 46507 people / square kilometer (total area of 5 square kilometers, 19 millions vnd/ square metre) a, district 4: 45815 people / square kilometer (total area of 4km square, 17 millions vnd / square metre). Some districts have low population densities and large areas such as Can Gio: about 100 people per square kilometer (total area of 704 square kilometers, only 762000vnd /square metre), Cu Chi: 817 people per square kilometer (435 square kilometers, 762000vnd / square metre)... This is a disparity in population distribution. Businesses, companies, and shops often focus on populated areas, leading to employees focusing on these places. This makes these places expensive to consume, expensive houses, polluted environments ...
- What the city leaders want is to be able to balance the population situation, promote the development of low-density districts by encouraging investors to enter these districts. Choosing to analyze this issue, we hope to provide suggestions for businesses planning to open or expand their businesses, which will select districts with low population density, large area, and only low real estate fees, low cost of living. This will help city residents to choose the appropriate new residence, reducing the load for densely populated districts.

- When considering the above issues, we will create charts of districts, clustered by population density, based on the location of each county. In order to be able to support decision making for everyone.

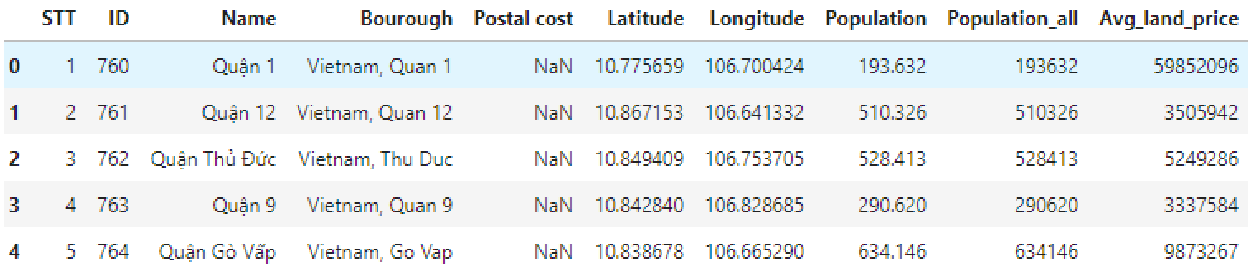
2. Data Description

With the above problem, we determine the need for relevant data as follows:

- Data of all districts in the city include: district name, area, population, longitude, latitude. This data is not available from a source that must be gathered from multiple sources. This is raw data, then we clean up the information as listed. Using this information can create related maps using choropleth map.
- City population data, including population of each district.
- City land price, including land price of each district.
- Use the Forsquare API to filter the venues of Borough of HoChiMinh City.
- From there select one Borough (specific districts like Can Gio) to analyze venues; use additional data obtained from Google Map and cluster data of the Borough.
- Because data is not available, collecting and cleaning data takes a lot of time.

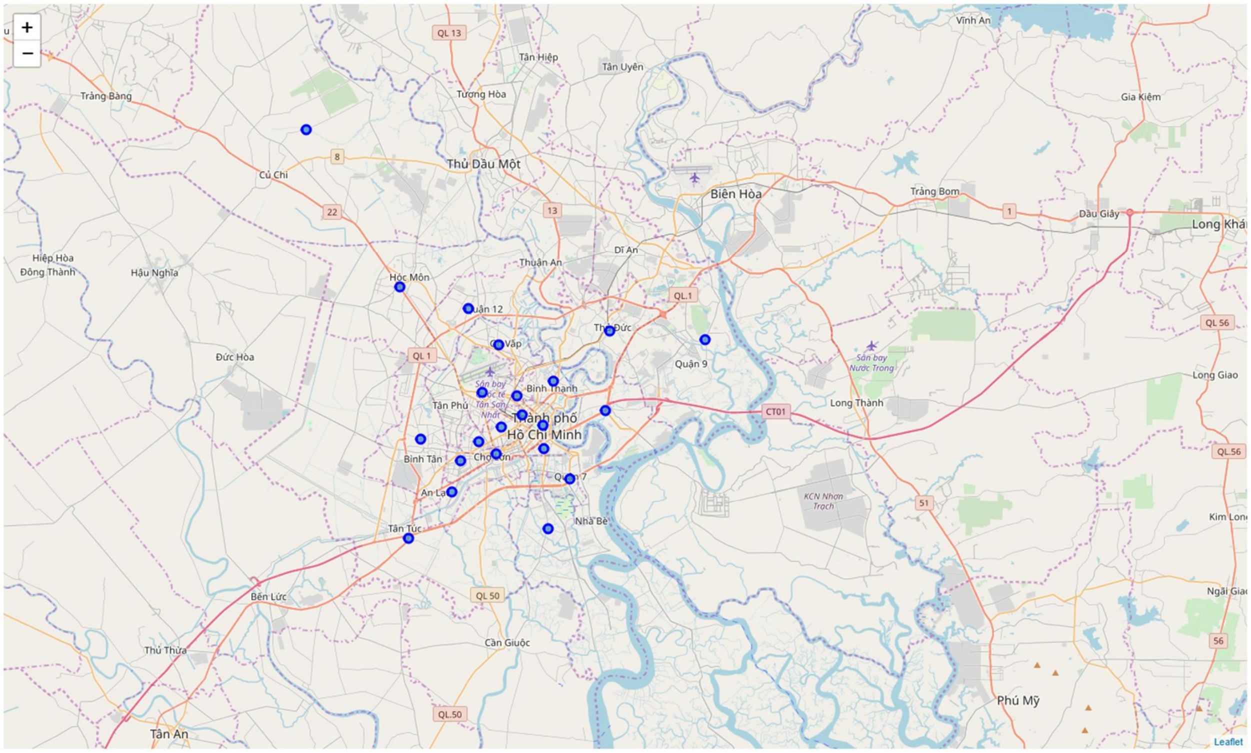
3. Methodology

- As a database, I used GitHub repository in my study. My master data which has the main components Name (of each district), Bourough, Latitude, Longitude, Population_all, Avg_land_price informations of the city.

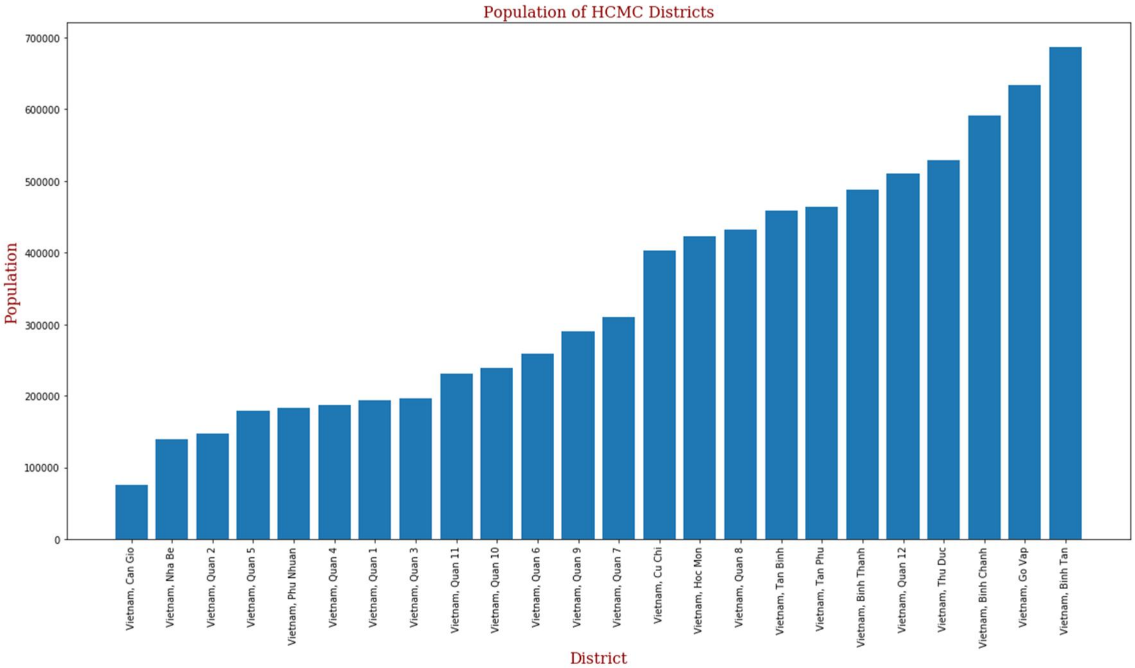


	STT	ID	Name	Bourough	Postal cost	Latitude	Longitude	Population	Population_all	Avg_land_price
0	1	760	Quận 1	Vietnam, Quan 1	NaN	10.775659	106.700424	193.632	193632	59852096
1	2	761	Quận 12	Vietnam, Quan 12	NaN	10.867153	106.641332	510.326	510326	3505942
2	3	762	Quận Thủ Đức	Vietnam, Thu Duc	NaN	10.849409	106.753705	528.413	528413	5249286
3	4	763	Quận 9	Vietnam, Quan 9	NaN	10.842840	106.828685	290.620	290620	3337584
4	5	764	Quận Gò Vấp	Vietnam, Go Vap	NaN	10.838678	106.665290	634.146	634146	9873267

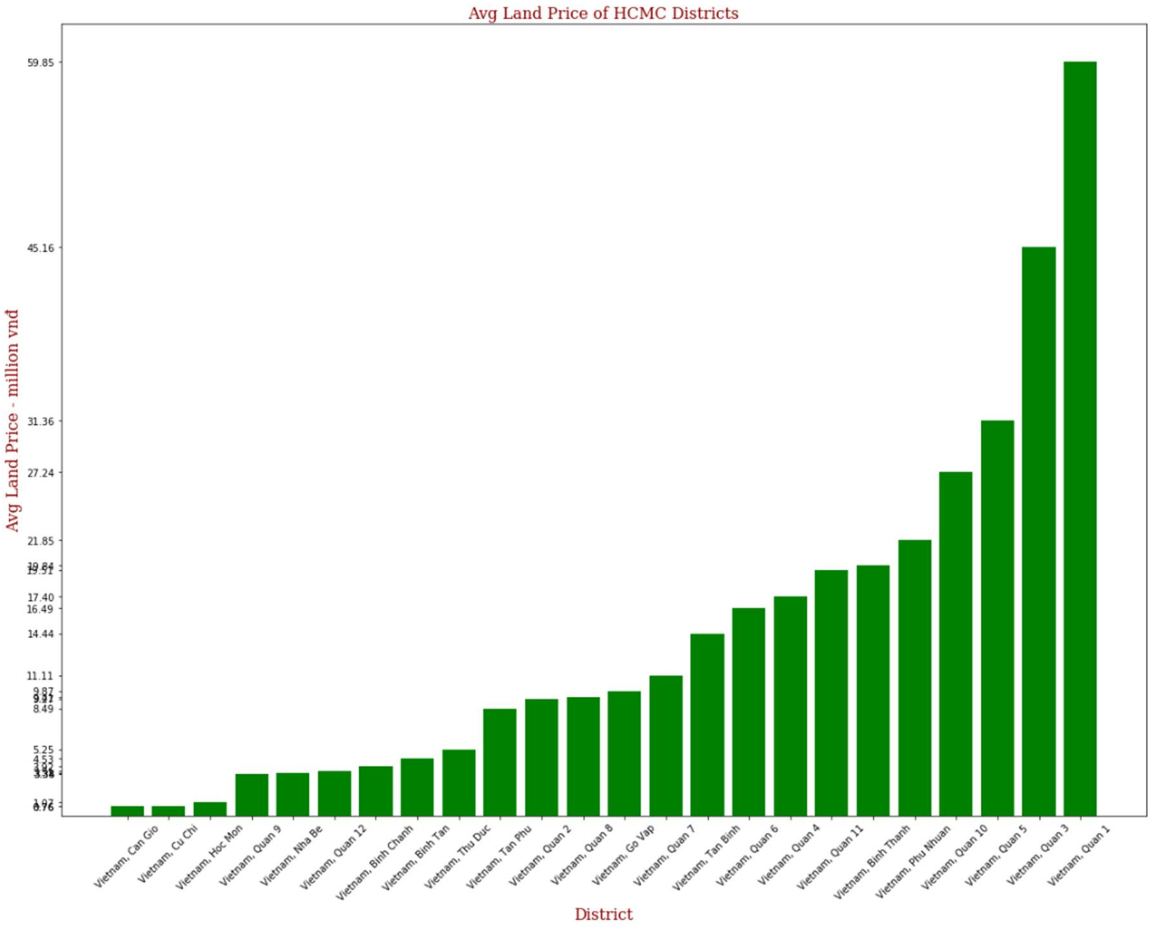
- I used Folium to to visualize geographic details of HoChiMinh city with all the districts. I used latitude and longitude values to get the visual as below:



- I used Matplotlib to show the population of each district in HoChiMinh city. I used Borough and Population_all values to get the visual as below:



- I used Matplotlib to show land price of each district in HoChiMinh city. I used Borough and Avg_land_price values to get the visual as below:

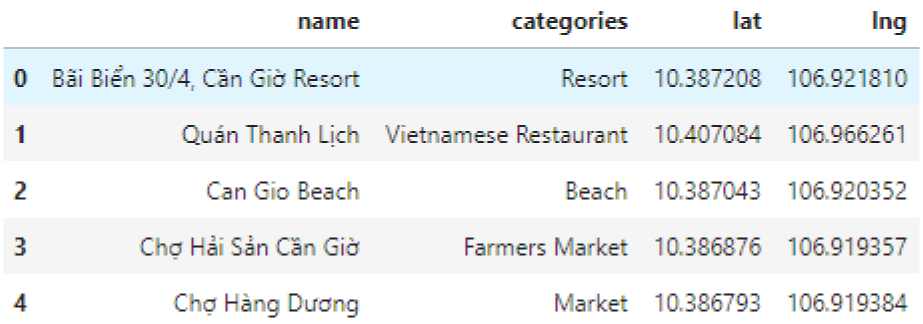


- Look at two charts above, we see that Can Gio has lowest population and also has lowest average land price.

- Now we have to find out the strengths of Can Gio so that we can recommend to businesses, organizations and individuals to come and invest. From there, it is possible to shift population, increase land value.

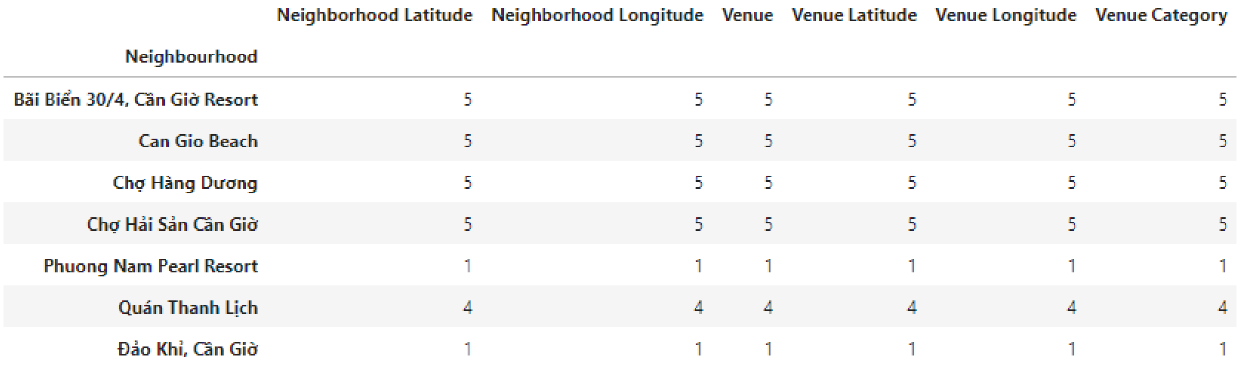
Analysis CanGio

- With CanGio, used the Foursquare API to explore the boroughs and segment them: limit = **20 venue** and radius = **10000 meter** (CanGio's area is very large: 704 square kilomet). Here is a head of the list Venues name, category, latitude and longitude informations from Forsquare API.



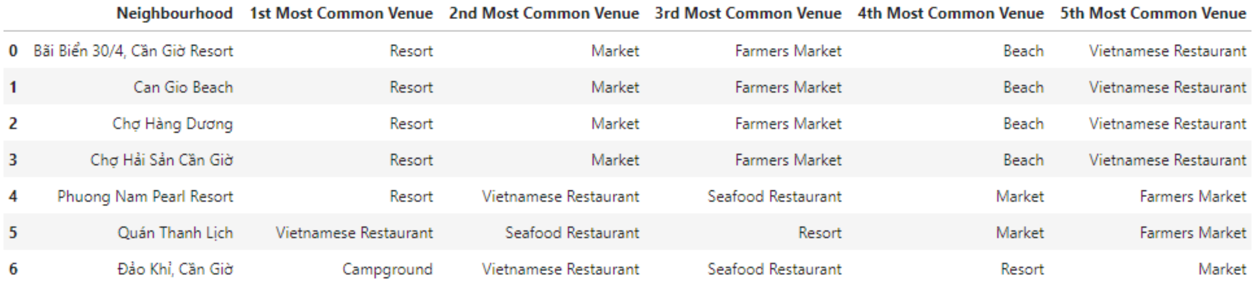
	name	categories	lat	lng
0	Bãi Biển 30/4, Cần Giờ Resort	Resort	10.387208	106.921810
1	Quán Thanh Lịch	Vietnamese Restaurant	10.407084	106.966261
2	Can Gio Beach	Beach	10.387043	106.920352
3	Chợ Hải Sản Cần Giờ	Farmers Market	10.386876	106.919357
4	Chợ Hàng Dương	Market	10.386793	106.919384

- In summary of this data 7 venues were returned by Foursquare
- Then, I explore the neighborhood in CanGio:



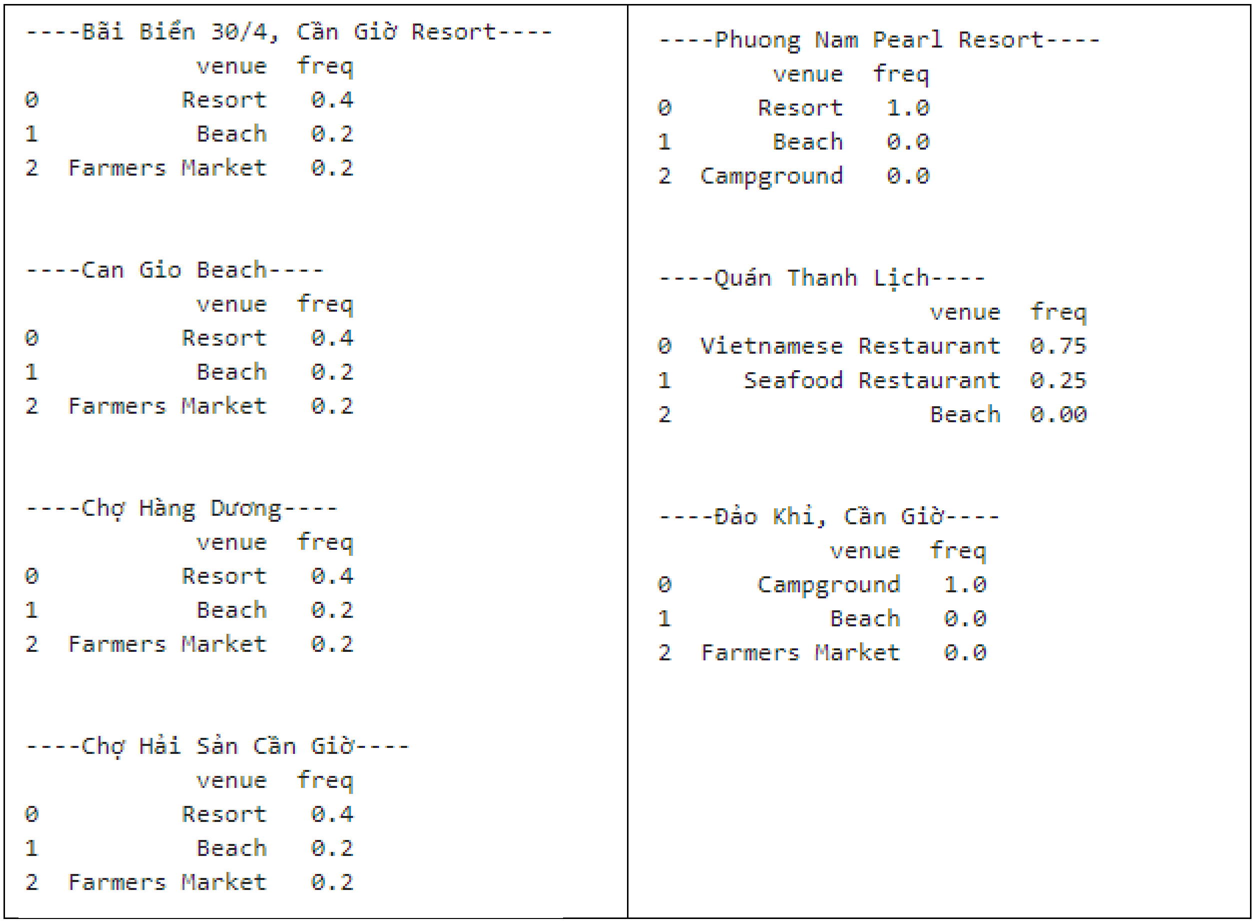
	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighbourhood						
Bãi Biển 30/4, Cần Giờ Resort	5	5	5	5	5	5
Can Gio Beach	5	5	5	5	5	5
Chợ Hàng Dương	5	5	5	5	5	5
Chợ Hải Sản Cần Giờ	5	5	5	5	5	5
Phuong Nam Pearl Resort	1	1	1	1	1	1
Quán Thanh Lịch	4	4	4	4	4	4
Đảo Khỉ, Cần Giờ	1	1	1	1	1	1

- There are list of 5 most common nenus:



	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Bãi Biển 30/4, Cần Giờ Resort	Resort	Market	Farmers Market	Beach	Vietnamese Restaurant
1	Can Gio Beach	Resort	Market	Farmers Market	Beach	Vietnamese Restaurant
2	Chợ Hàng Dương	Resort	Market	Farmers Market	Beach	Vietnamese Restaurant
3	Chợ Hải Sản Cần Giờ	Resort	Market	Farmers Market	Beach	Vietnamese Restaurant
4	Phuong Nam Pearl Resort	Resort	Vietnamese Restaurant	Seafood Restaurant	Market	Farmers Market
5	Quán Thanh Lịch	Vietnamese Restaurant	Seafood Restaurant	Resort	Market	Farmers Market
6	Đảo Khỉ, Cần Giờ	Campground	Vietnamese Restaurant	Seafood Restaurant	Resort	Market

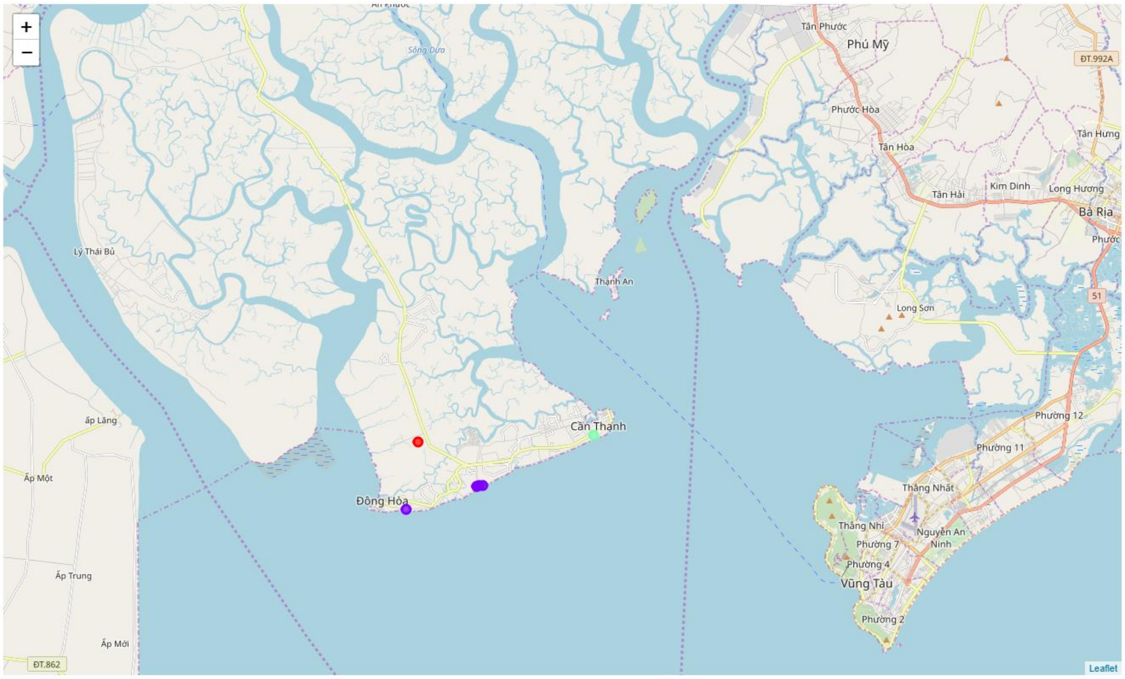
- Each neighborhood along with the top 3 most common venues:



```
----Bãi Biển 30/4, Cần Giờ Resort----
            venue  freq
0          Resort   0.4
1           Beach   0.2
2  Farmers Market   0.2


----Can Gio Beach----
            venue  freq
0          Resort   0.4
1           Beach   0.2
2  Farmers Market   0.2


----Chợ Hàng Dương----
            venue  freq
0          Resort   0.4
1           Beach   0.2
2  Farmers Market   0.2


----Chợ Hải Sản Cần Giờ----
            venue  freq
0          Resort   0.4
1           Beach   0.2
2  Farmers Market   0.2
```

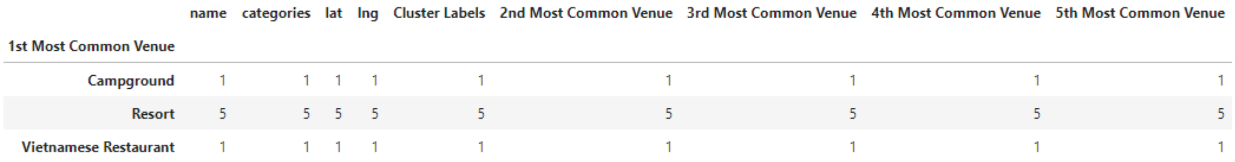
```
----Phuong Nam Pearl Resort----
        venue  freq
0      Resort   1.0
1       Beach   0.0
2  Campground   0.0


----Quán Thanh Lịch----
                   venue  freq
0  Vietnamese Restaurant  0.75
1     Seafood Restaurant  0.25
2                  Beach  0.00


----Đảo Khỉ, Cần Giờ----
            venue  freq
0      Campground   1.0
1           Beach   0.0
2  Farmers Market   0.0
```

- At analyzing venuses of CanGio, I saw that I need to use K-Means clustering to cluster venus of CanGio to three clusters:

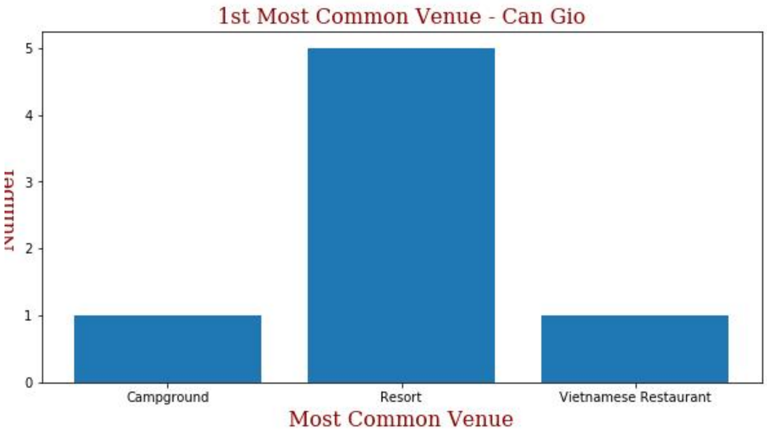


4. Results

- This is the result after clustering: we have 3 clusters: *Resort, Campground and Vietnamese Restautant*:



	name	categories	lat	lng	Cluster Labels	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1st Most Common Venue									
Campground	1	1	1	1	1	1	1	1	1
Resort	5	5	5	5	5	5	5	5	5
Vietnamese Restaurant	1	1	1	1	1	1	1	1	1



- According to the Google map, we get through Folium, Can Gio is a land with a long coastline. According to the analysis results, the strong points of CanGio that we can investment and development are fishing and aquaculture, tourism business, restaurant, hotel.

5. Discussion

- As I have mentioned in the introduction, HCM city is a big city with 24 districts but the population density and land prices are uneven. Therefore, many consequences are not good.
- The goal is to find districts with low population density, low land prices and find the strengths of these places, suggesting investors and individuals to come and develop.
- I have found CanGio and applied K-Means to cluster main investment groups as stated in the results section.
- Can Gio is only a part of the job. In the future, it is necessary to analyze other districts with low population density and low land prices to suggest to investors and individuals.

6. Conclusion

- With this result, it can help the organizations, businesses and individuals plan to invest in districts which have small population, cheap land but have potential for development with a clearer view.

- However, in order to do this well, it is necessary to provide information channels such as websites, electronic portals so that people can look up the analyzed information.