Coursera Capstone – The Battles of Neighborhoods

Open Japanese Restaurant in Hochiminh City

Table of contents

1. Introduction: Business Problem

Hochiminh City is the most populous city in Vietnam with a population of 8.4 million (13 million in the metropolitan area) as of 2017. As a major gateway to Vietnam, the city received over 8.6 million international visitors in 2019. Therefore, it would have a very number of potential customer if a restaurant is open. So if someone ask to open a Japanese Restaurant in Hochiminh City, where it should be opened so it's profit the best?

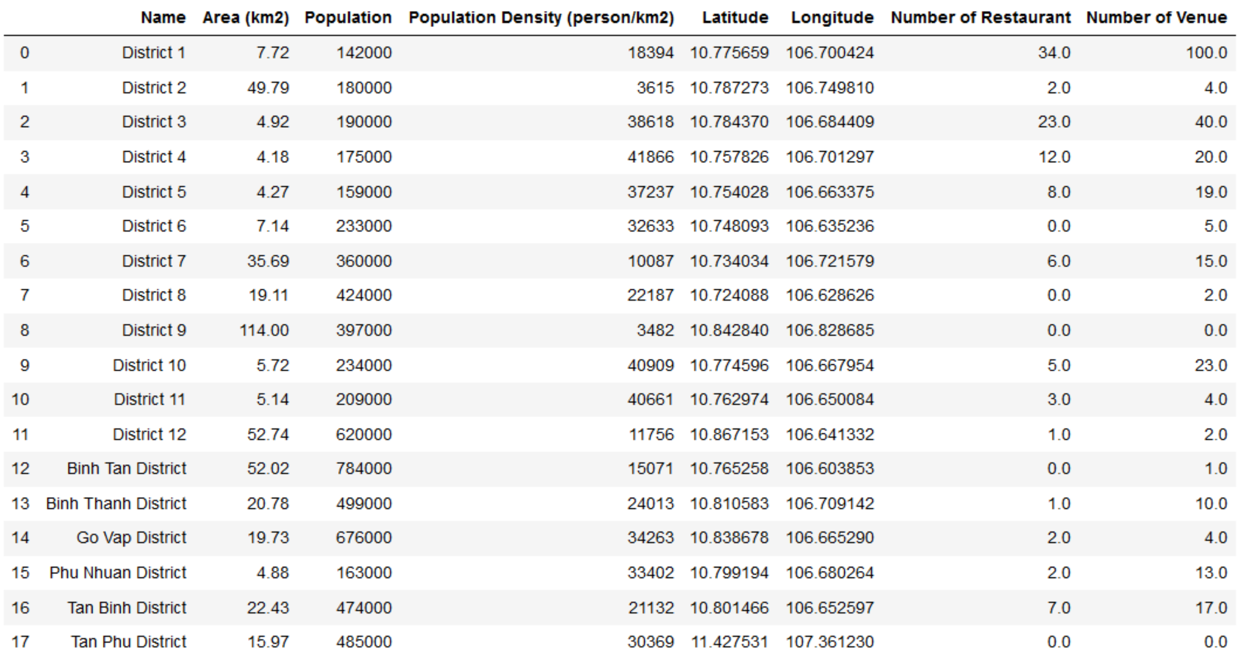
2. Data

In order to solve the above question, the data of Hochiminh City like district, population, area, population density Geolocation data collected from FourSquare wil also be used to collect number of restaurants and their type and location in every neighborhood as well as other venues.

3. Methodology ¶

In this project, we will invest the data of Hochiminh City like district, population, area, population density finds out if there is any correlation between them and the number of restaurant opened by using FourSquare API. If there is, we can narrow some neighborhoods to open the restaurants.

By preprocessing the data we have, we will have the details of information of each District like below:



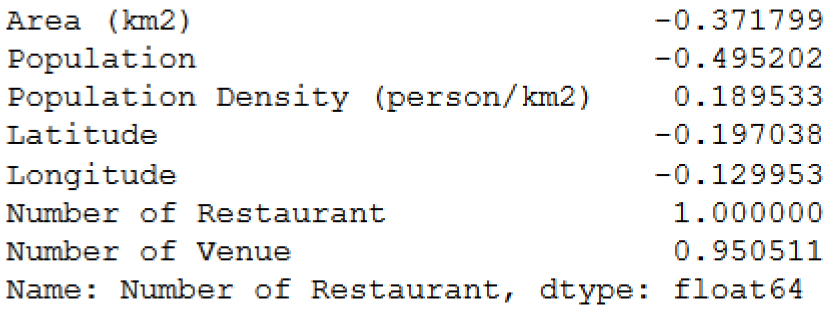
	Name	Area (km2)	Population	Population Density (person/km2)	Latitude	Longitude	Number of Restaurant	Number of Venue
0	District 1	7.72	142000	18394	10.775659	106.700424	34.0	100.0
1	District 2	49.79	180000	3615	10.787273	106.749810	2.0	4.0
2	District 3	4.92	190000	38618	10.784370	106.684409	23.0	40.0
3	District 4	4.18	175000	41866	10.757826	106.701297	12.0	20.0
4	District 5	4.27	159000	37237	10.754028	106.663375	8.0	19.0
5	District 6	7.14	233000	32633	10.748093	106.635236	0.0	5.0
6	District 7	35.69	360000	10087	10.734034	106.721579	6.0	15.0
7	District 8	19.11	424000	22187	10.724088	106.628626	0.0	2.0
8	District 9	114.00	397000	3482	10.842840	106.828685	0.0	0.0
9	District 10	5.72	234000	40909	10.774596	106.667954	5.0	23.0
10	District 11	5.14	209000	40661	10.762974	106.650084	3.0	4.0
11	District 12	52.74	620000	11756	10.867153	106.641332	1.0	2.0
12	Binh Tan District	52.02	784000	15071	10.765258	106.603853	0.0	1.0
13	Binh Thanh District	20.78	499000	24013	10.810583	106.709142	1.0	10.0
14	Go Vap District	19.73	676000	34263	10.838678	106.665290	2.0	4.0
15	Phu Nhuan District	4.88	163000	33402	10.799194	106.680264	2.0	13.0
16	Tan Binh District	22.43	474000	21132	10.801466	106.652597	7.0	17.0
17	Tan Phu District	15.97	485000	30369	11.427531	107.361230	0.0	0.0

After that, we will use k-means clustering to segment and cluster the neighborhoods in the city of New York and see what kind of restaurant is opened in district so we can avoid it and choose a better place.

4. Analysis

4.1 Correlation:

The correlation between the number of Restaurant with the rest information:

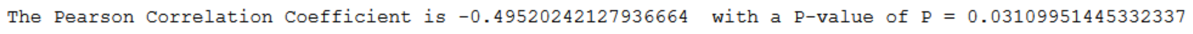


```
Area (km2)                         -0.371799
Population                         -0.495202
Population Density (person/km2)     0.189533
Latitude                           -0.197038
Longitude                          -0.129953
Number of Restaurant                1.000000
Number of Venue                     0.950511
Name: Number of Restaurant, dtype: float64
```

The correlation coefficient between Number of Restaurant and Population, and between Number of Restaurant and Number of Venue are noticeable: -0.495202, 0.950511.

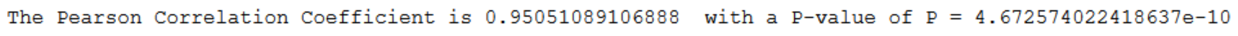
Detail Pearson coefficient and P_value of these 2 groups:

- Number of Restaurant and Population:



```
The Pearson Correlation Coefficient is -0.4952024212793664  with a P-value of P = 0.03109951445332337
```

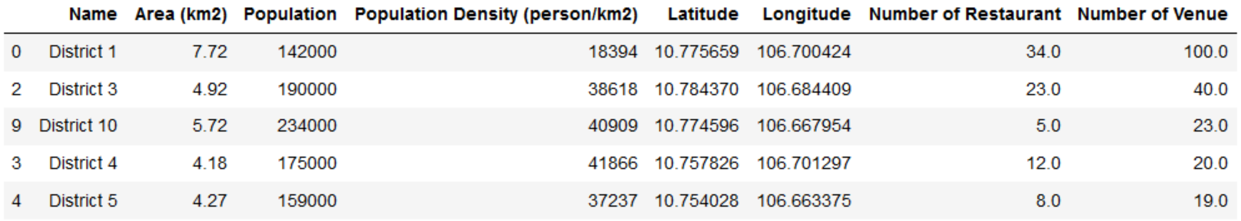
- Number of Restaurant and Number of Venue:



```
The Pearson Correlation Coefficient is 0.95051089106888  with a P-value of P = 4.672574022418637e-10
```

So the number of Venue for each district will be used as main critical to decide the economy of the district as well as the district to open a restaurant

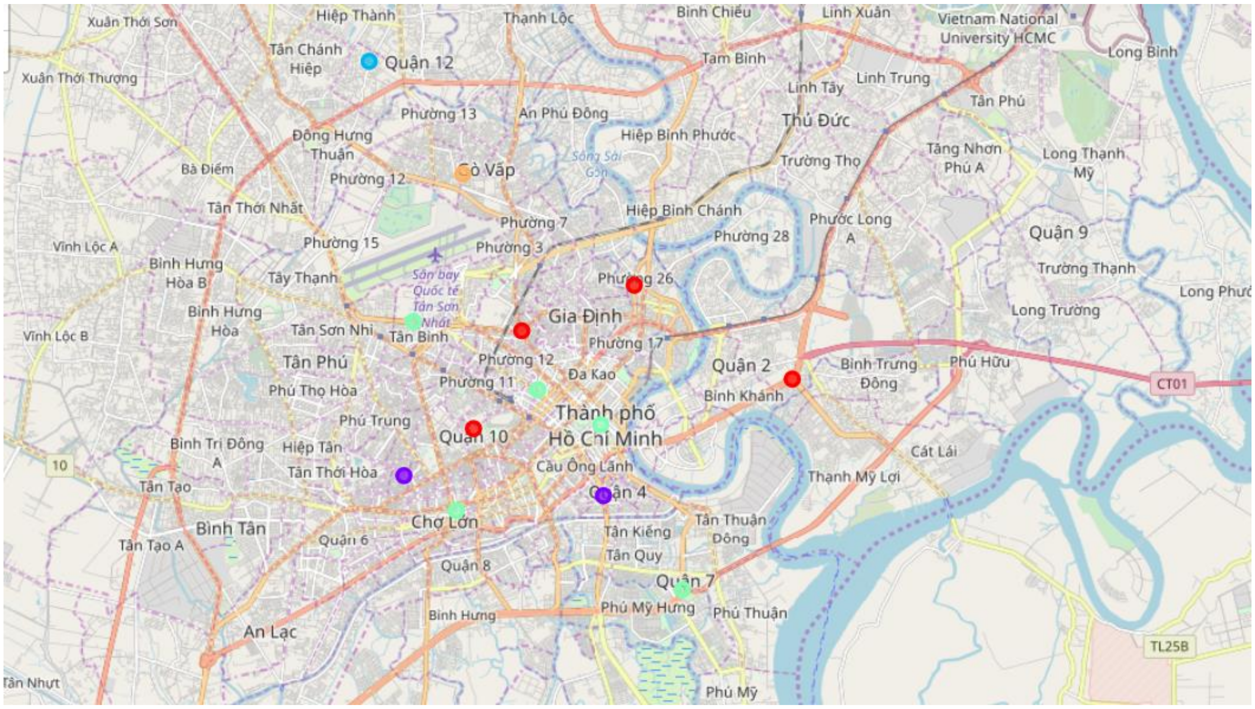
Therefore, the top five district have highest number of Venue:



	Name	Area (km2)	Population	Population Density (person/km2)	Latitude	Longitude	Number of Restaurant	Number of Venue
0	District 1	7.72	142000	18394	10.775659	106.700424	34.0	100.0
2	District 3	4.92	190000	38618	10.784370	106.684409	23.0	40.0
9	District 10	5.72	234000	40909	10.774596	106.667954	5.0	23.0
3	District 4	4.18	175000	41866	10.757826	106.701297	12.0	20.0
4	District 5	4.27	159000	37237	10.754028	106.663375	8.0	19.0

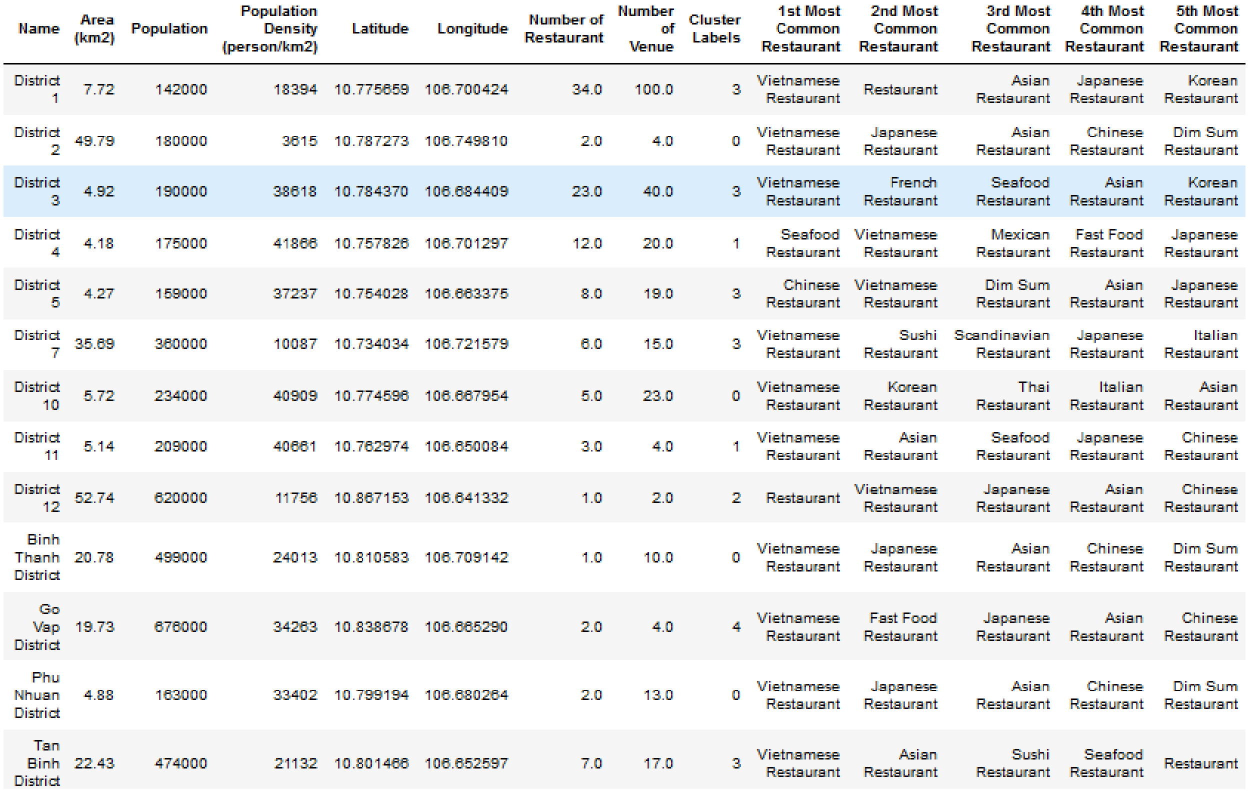
4.2 K-means

- With K-means clustering technique, the top 5 clusters of similar neighborhoods have been apparent in the result below:



This information will be necessary so we can target on the cluster that offer the largest business expansion.

- With Foursquare API, we can also find out the top common restaurant on each District. This is critical we want to recommend a place with low competition as much as possible



Name	Area (km2)	Population	Population Density (person/km2)	Latitude	Longitude	Number of Restaurant	Number of Venue	Cluster Labels	1st Most Common Restaurant	2nd Most Common Restaurant	3rd Most Common Restaurant	4th Most Common Restaurant	5th Most Common Restaurant
District 1	7.72	142000	18394	10.775659	106.700424	34.0	100.0	3	Vietnamese Restaurant	Restaurant	Asian Restaurant	Japanese Restaurant	Korean Restaurant
District 2	49.79	180000	3615	10.787273	106.749810	2.0	4.0	0	Vietnamese Restaurant	Japanese Restaurant	Asian Restaurant	Chinese Restaurant	Dim Sum Restaurant
District 3	4.92	190000	38618	10.784370	106.684409	23.0	40.0	3	Vietnamese Restaurant	French Restaurant	Seafood Restaurant	Asian Restaurant	Korean Restaurant
District 4	4.18	175000	41866	10.757826	106.701297	12.0	20.0	1	Seafood Restaurant	Vietnamese Restaurant	Mexican Restaurant	Fast Food Restaurant	Japanese Restaurant
District 5	4.27	159000	37237	10.754028	106.663375	8.0	19.0	3	Chinese Restaurant	Vietnamese Restaurant	Dim Sum Restaurant	Asian Restaurant	Japanese Restaurant
District 7	35.69	360000	10087	10.734034	106.721579	6.0	15.0	3	Vietnamese Restaurant	Sushi Restaurant	Scandinavian Restaurant	Japanese Restaurant	Italian Restaurant
District 10	5.72	234000	40909	10.774596	106.667954	5.0	23.0	0	Vietnamese Restaurant	Korean Restaurant	Thai Restaurant	Italian Restaurant	Asian Restaurant
District 11	5.14	209000	40661	10.762974	106.650084	3.0	4.0	1	Vietnamese Restaurant	Asian Restaurant	Seafood Restaurant	Japanese Restaurant	Chinese Restaurant
District 12	52.74	620000	11756	10.867153	106.641332	1.0	2.0	2	Restaurant	Vietnamese Restaurant	Japanese Restaurant	Asian Restaurant	Chinese Restaurant
Binh Thanh District	20.78	499000	24013	10.810583	106.709142	1.0	10.0	0	Vietnamese Restaurant	Japanese Restaurant	Asian Restaurant	Chinese Restaurant	Dim Sum Restaurant
Go Vap District	19.73	676000	34263	10.838678	106.665290	2.0	4.0	4	Vietnamese Restaurant	Fast Food Restaurant	Japanese Restaurant	Asian Restaurant	Chinese Restaurant
Phu Nhuan District	4.88	163000	33402	10.799194	106.680264	2.0	13.0	0	Vietnamese Restaurant	Japanese Restaurant	Asian Restaurant	Chinese Restaurant	Dim Sum Restaurant
Tan Binh District	22.43	474000	21132	10.801466	106.652597	7.0	17.0	3	Vietnamese Restaurant	Asian Restaurant	Sushi Restaurant	Seafood Restaurant	Restaurant

5. Results and Discussion

Our analyst shows that there is a strong relation between Number of Restaurant and Number of Venue. So it seems that the restaurant should be opened where the economy is higher than other place. We can narrow down the top 5 district have highest economy: 1,3,10,4,5. Within these 5 district, the ratio between restaurant's number and Venue's number of district 1, 10 are quite low compare to other place so these 2 districts seem like the place we are looking for.

By using K-means, we found that on District 1, the Japanese restaurant is in the 4th most common Restaurant. So if we open a Japanese restaurant in District 1, there quite a number of rivalry. But in District 10, the Japanese restaurant isn't in top 5 common restaurant so District 10 is likely a place to open the Japanese restaurant.

6. Conclusion

Purpose of this project was to identify Hochiminh City area in order to aid stakeholders in narrowing down the search for optimal location for a new Japanese restaurant. By calculating restaurant density distribution from Foursquare data and the relation of between each characteristics of District, we have chosen District 10 as the starting point for final decision by stakeholders. The exact location will be made only after stakeholders considers on other characteristics of district like the attractiveness, real estate availability, estate's price, traffic, …