Applied Data Science - Capstone Project

Using location based and clustering to build recommendation system

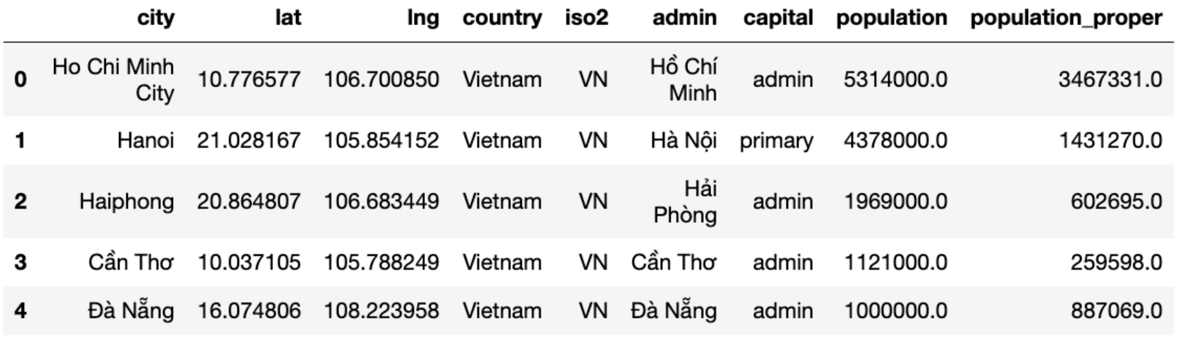
I. Introduction Business Problem.

Ho Chi Minh city is largest city of Vietnam which have nearly ten million in terms of population and have a lot of non-residents such as business travelers or tourists. The statistics showed that there was more than 6 million international tourist visited Ho Chi Minh city in the first nine months of 2019. This is amazing number for any related business and this is the reason why a group of young investors would like to find a good location to start their business by setting up a restaurant or coffee shop in this crowded and dynamic city. The investor would like to leverage the data analyse advise them where is good location to open their business.

From this stand point, there are several ways of approach such as identify where are the most attractives of people in the city or where are the business centres and so on. One of the approach is using available location based data to analyse it and make the recommendation.

II. Description of Data

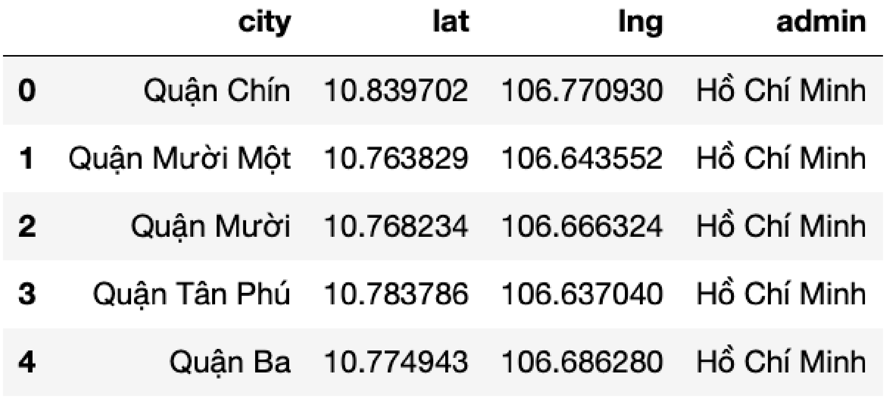
In solving this problem, the location data comes from a csv file which define the latitude, longitude and other information of all the cities in Vietnam as well as its neighborhoods. This is the sample data of city and its neighborhoods:



	city	lat	lng	country	iso2	admin	capital	population	population_proper
0	Ho Chi Minh City	10.776577	106.700850	Vietnam	VN	Hồ Chí Minh	admin	5314000.0	3467331.0
1	Hanoi	21.028167	105.854152	Vietnam	VN	Hà Nội	primary	4378000.0	1431270.0
2	Haiphong	20.864807	106.683449	Vietnam	VN	Hải Phòng	admin	1969000.0	602695.0
3	Cần Thơ	10.037105	105.788249	Vietnam	VN	Cần Thơ	admin	1121000.0	259598.0
4	Đà Nẵng	16.074806	108.223958	Vietnam	VN	Đà Nẵng	admin	1000000.0	887069.0

Source file: https://raw.githubusercontent.com/dodtoan/Coursera_Capstone/master/vn.csv

Obviously, this data is raw data and need to be cleaned before actually use. The cleaned data can be used as "source" data to explore further venues in the every single neighborhood using FourSquare API. There are some unnecessary fields should be removed cause Foursquare just need the latitude and longitude of the cities only and the purpose of the analyse just focuses on the Ho Chi Minh city so other city information would be removed too. The data after cleansing would be like:



	city	lat	lng	admin
0	Quận Chín	10.839702	106.770930	Hồ Chí Minh
1	Quận Mười Một	10.763829	106.643552	Hồ Chí Minh
2	Quận Mười	10.768234	106.666324	Hồ Chí Minh
3	Quận Tân Phú	10.783786	106.637040	Hồ Chí Minh
4	Quận Ba	10.774943	106.686280	Hồ Chí Minh

By exploring the venue data from Foursquare, clustering algorithm would be applied to categorize the neighborhoods in to several clusters which they have the similar properties and from that view, the good location to start cafe/restaurant business can be suggested.

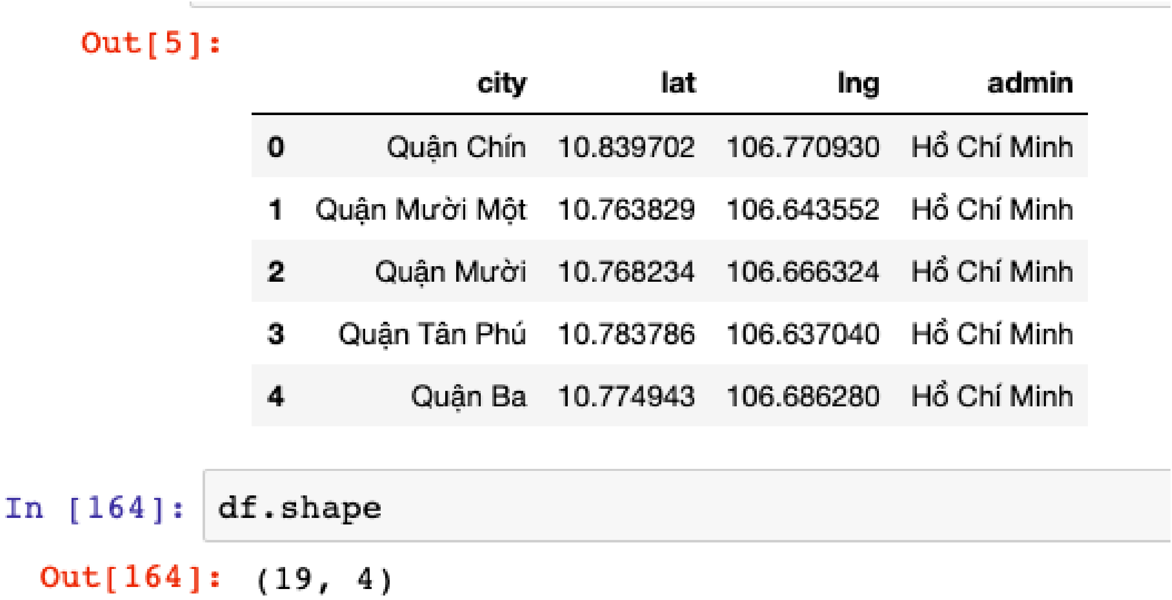
To make a suggestion, some properties of data from Foursquare would be leverage to analyse to find the pattern and relation between the venues. They are:

1. Name of venue;
2. Categories;
3. Latitude;
4. Longitude;

III. Methodology

Stated in the business problem, the expectation outcome is a recommendation where is a suitable location to settle a restaurant. This kind of question would be good use case to utilize ***unsupervised*** machine learning and more precise, it is ***K-mean*** clustering algorithm and integrate the outcome with ***Foursquare API*** and ***folium*** library to visualize the result.

Let take a detail look into the data. The source data of location based for Ho Chi Minh city and its neighborhood has the size of (19,4), it means there are 19 neighborhoods in the investigated area.



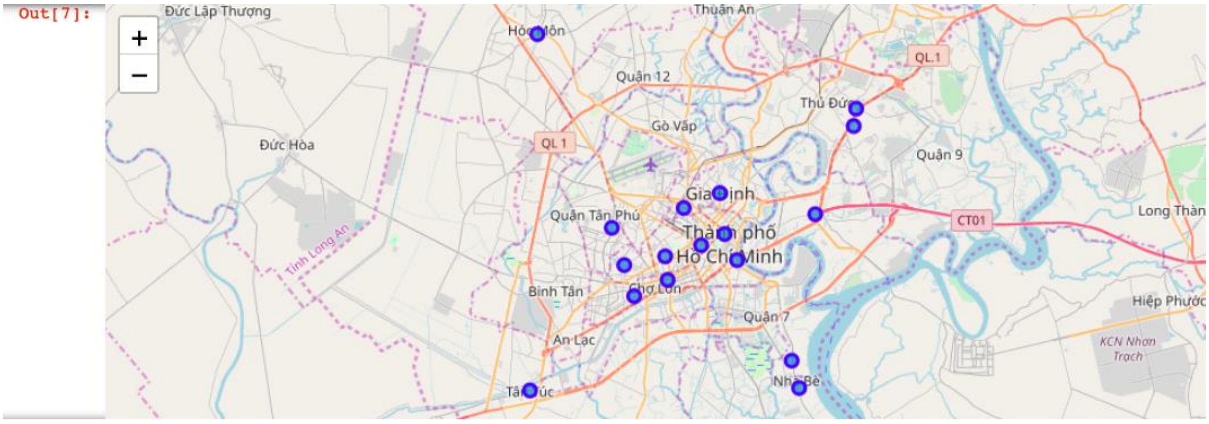
Out[5]:

	city	lat	lng	admin
0	Quận Chín	10.839702	106.770930	Hồ Chí Minh
1	Quận Mười Một	10.763829	106.643552	Hồ Chí Minh
2	Quận Mười	10.768234	106.666324	Hồ Chí Minh
3	Quận Tân Phú	10.783786	106.637040	Hồ Chí Minh
4	Quận Ba	10.774943	106.686280	Hồ Chí Minh

```
In [164]: df.shape
```

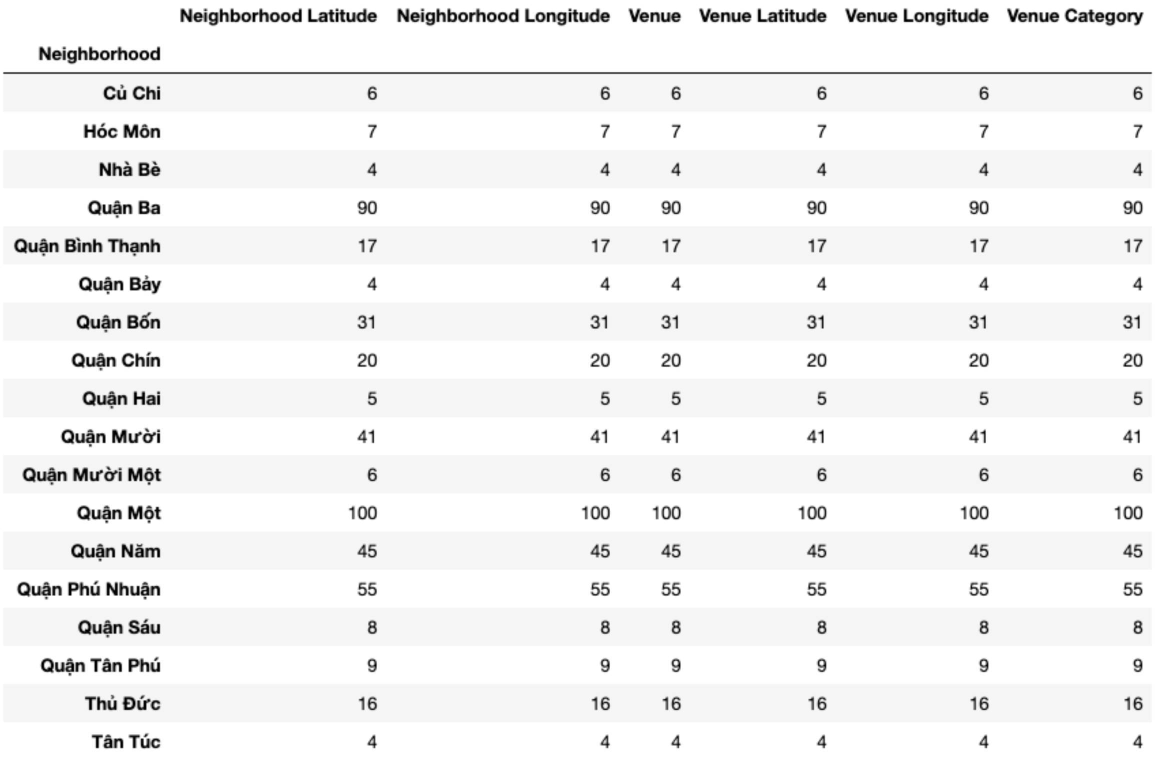
```
Out[164]: (19, 4)
```

With the ***folio*** library, the data set can be represented on the map as below:

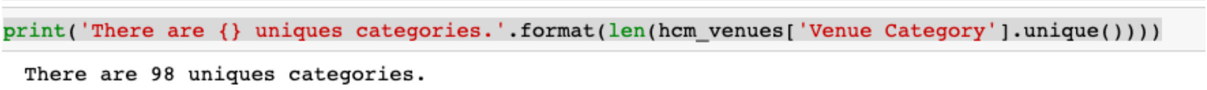


Out[7]:

From this point, the next step is using Foursquare to explore the venues in the neighborhoods. Cause the limited of the subscription, there is maximum of 100 venues in results and the radius for the exploration was set for 700m. This make a result as below:



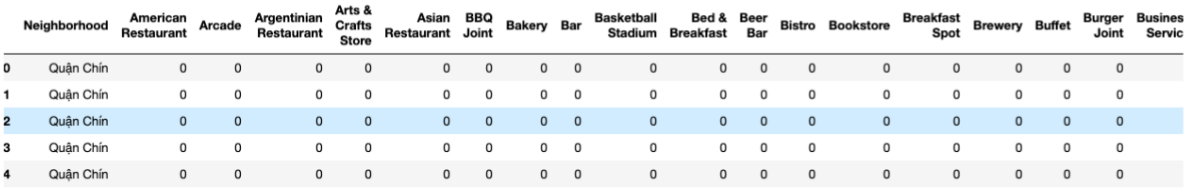
Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Củ Chi	6	6	6	6	6	6
Hóc Môn	7	7	7	7	7	7
Nhà Bè	4	4	4	4	4	4
Quận Ba	90	90	90	90	90	90
Quận Bình Thạnh	17	17	17	17	17	17
Quận Bảy	4	4	4	4	4	4
Quận Bốn	31	31	31	31	31	31
Quận Chín	20	20	20	20	20	20
Quận Hai	5	5	5	5	5	5
Quận Mười	41	41	41	41	41	41
Quận Mười Một	6	6	6	6	6	6
Quận Một	100	100	100	100	100	100
Quận Năm	45	45	45	45	45	45
Quận Phú Nhuận	55	55	55	55	55	55
Quận Sáu	8	8	8	8	8	8
Quận Tân Phú	9	9	9	9	9	9
Thủ Đức	16	16	16	16	16	16
Tân Túc	4	4	4	4	4	4



```
print('There are {} uniques categories.'.format(len(hcm_venues['Venue Category'].unique())))
```

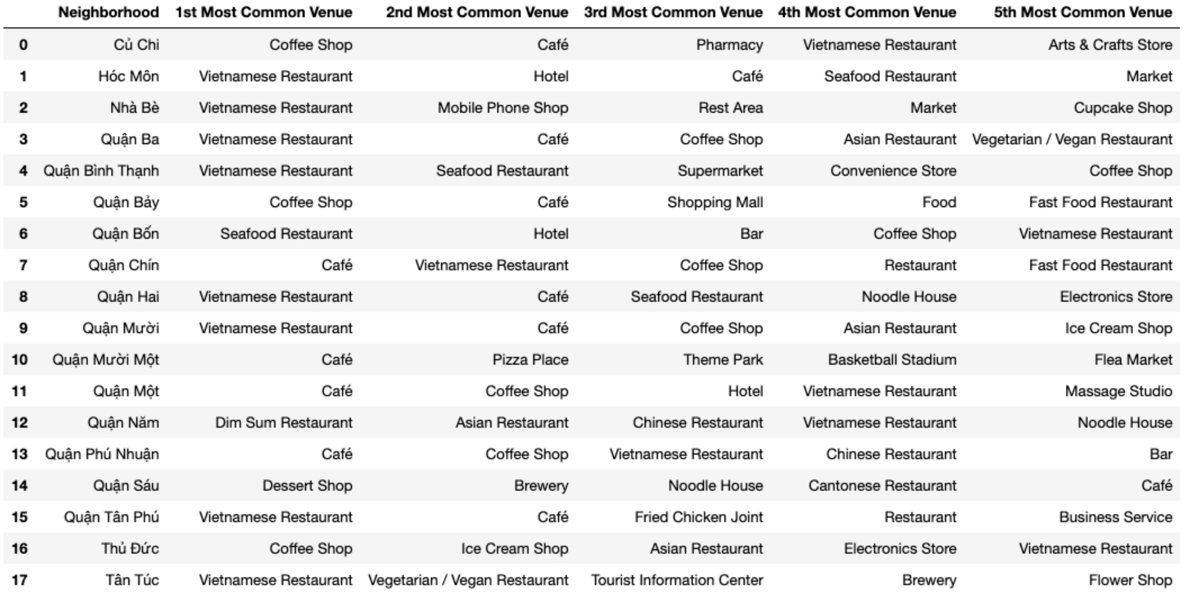
```
There are 98 uniques categories.
```

The result showed that there were 98 venue categories found. To analyze each neighborhood and how the relative with its venues, above data need to be standardized. After applying the *get_dummies()* method in Python and merging the result, the new dataset looks like:



	Neighborhood	American Restaurant	Arcade	Argentinian Restaurant	Arts & Crafts Store	Asian Restaurant	BBQ Joint	Bakery	Bar	Basketball Stadium	Bed & Breakfast	Beer Bar	Bistro	Bookstore	Breakfast Spot	Brewery	Buffet	Burger Joint	Busines Servic
0	Quận Chín	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	Quận Chín	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	Quận Chín	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	Quận Chín	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	Quận Chín	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

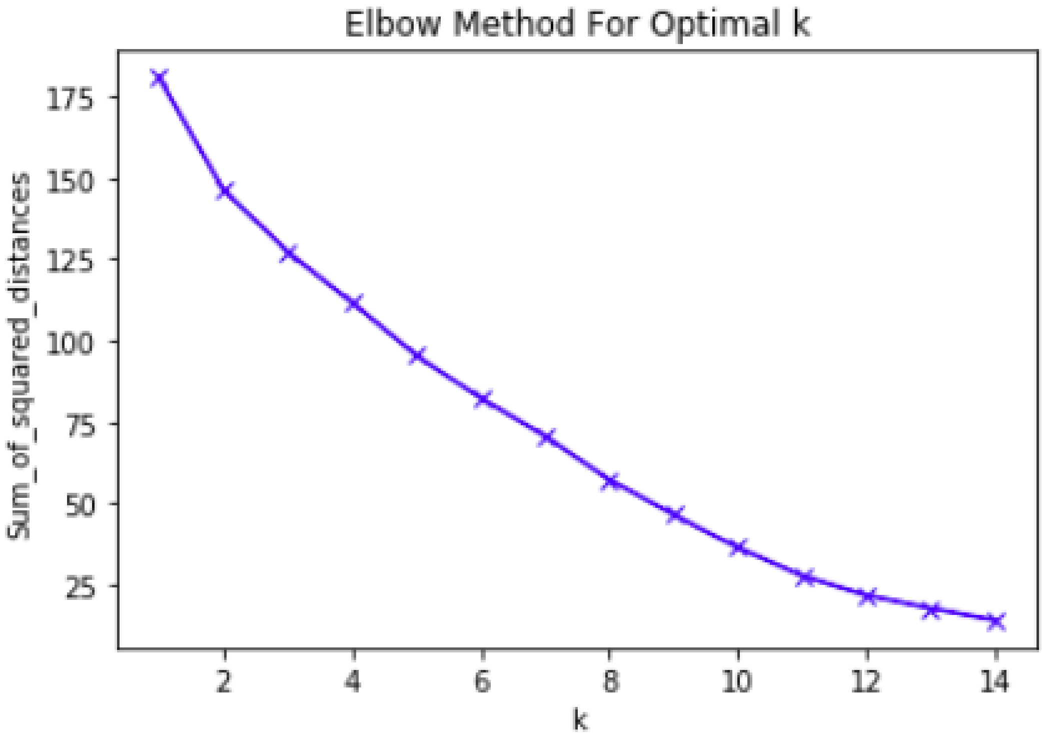
This dataset is still complicated to analyze cause there are 98 venues categories which most of them may not relevant to the features need to analyse then it would be transform to the new shape. The good idea is shorten the result into the top 5 common venues. After transforming the dataset, the new result would be:



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Củ Chi	Coffee Shop	Café	Pharmacy	Vietnamese Restaurant	Arts & Crafts Store
1	Hóc Môn	Vietnamese Restaurant	Hotel	Café	Seafood Restaurant	Market
2	Nhà Bè	Vietnamese Restaurant	Mobile Phone Shop	Rest Area	Market	Cupcake Shop
3	Quận Ba	Vietnamese Restaurant	Café	Coffee Shop	Asian Restaurant	Vegetarian / Vegan Restaurant
4	Quận Bình Thạnh	Vietnamese Restaurant	Seafood Restaurant	Supermarket	Convenience Store	Coffee Shop
5	Quận Bảy	Coffee Shop	Café	Shopping Mall	Food	Fast Food Restaurant
6	Quận Bốn	Seafood Restaurant	Hotel	Bar	Coffee Shop	Vietnamese Restaurant
7	Quận Chín	Café	Vietnamese Restaurant	Coffee Shop	Restaurant	Fast Food Restaurant
8	Quận Hai	Vietnamese Restaurant	Café	Seafood Restaurant	Noodle House	Electronics Store
9	Quận Mười	Vietnamese Restaurant	Café	Coffee Shop	Asian Restaurant	Ice Cream Shop
10	Quận Mười Một	Café	Pizza Place	Theme Park	Basketball Stadium	Flea Market
11	Quận Một	Café	Coffee Shop	Hotel	Vietnamese Restaurant	Massage Studio
12	Quận Năm	Dim Sum Restaurant	Asian Restaurant	Chinese Restaurant	Vietnamese Restaurant	Noodle House
13	Quận Phú Nhuận	Café	Coffee Shop	Vietnamese Restaurant	Chinese Restaurant	Bar
14	Quận Sáu	Dessert Shop	Brewery	Noodle House	Cantonese Restaurant	Café
15	Quận Tân Phú	Vietnamese Restaurant	Café	Fried Chicken Joint	Restaurant	Business Service
16	Thủ Đức	Coffee Shop	Ice Cream Shop	Asian Restaurant	Electronics Store	Vietnamese Restaurant
17	Tân Túc	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Tourist Information Center	Brewery	Flower Shop

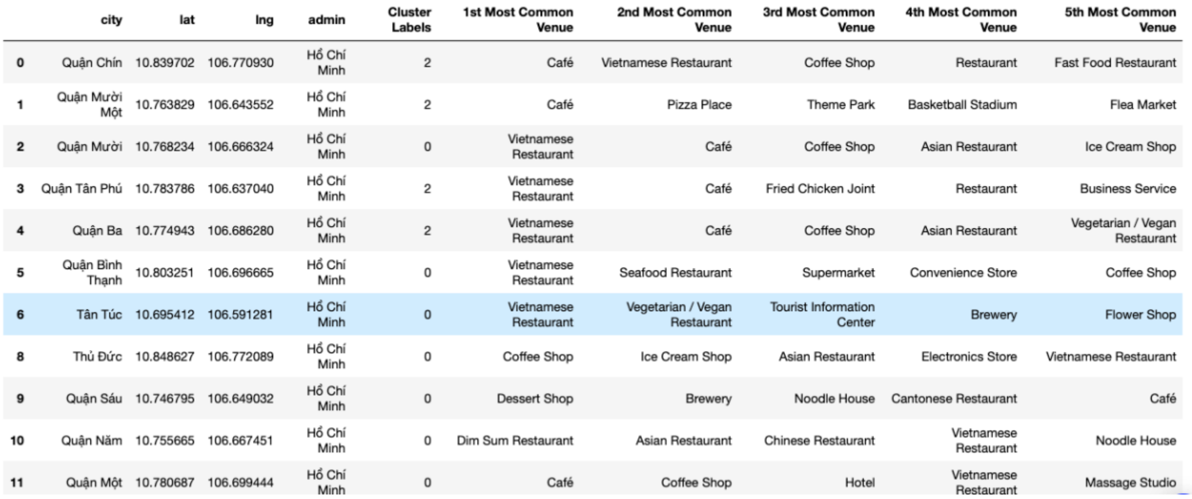
It is obviously found that now the dataset had only 18 cities in stead of 19 cities at the beginning. This missing will be discussed later but the new dataset is good enough to analyse.

It is time to apply the K-mean clustering algorithm. Before running the K-mean to the dataset, it would be necessary to find out what is best K. Using the Elbow method, the result showed that the K would be 3 or 4 but let's take 3.



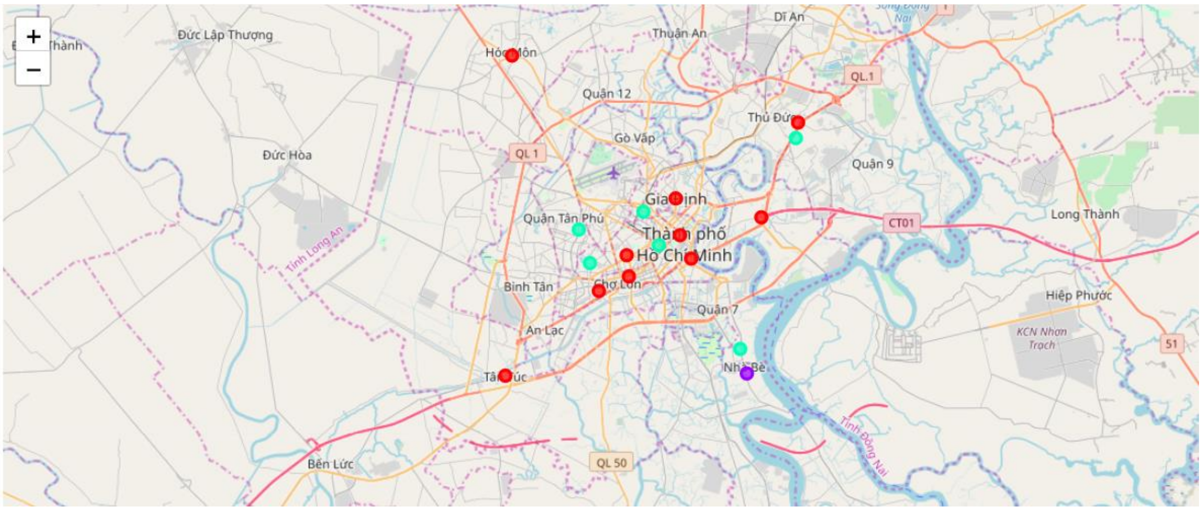
IV. Result:

Applying K-mean algorithm with K=3, merged with the original data, the result showed in the table below with 3 clusters:



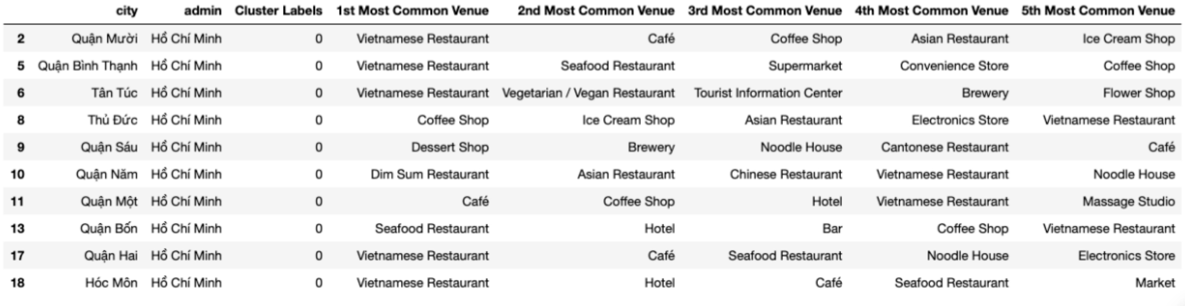
	city	lat	lng	admin	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Quận Chín	10.839702	106.770930	Hồ Chí Minh	2	Café	Vietnamese Restaurant	Coffee Shop	Restaurant	Fast Food Restaurant
1	Quận Mười Một	10.763829	106.643552	Hồ Chí Minh	2	Café	Pizza Place	Theme Park	Basketball Stadium	Flea Market
2	Quận Mười	10.768234	106.666324	Hồ Chí Minh	0	Vietnamese Restaurant	Café	Coffee Shop	Asian Restaurant	Ice Cream Shop
3	Quận Tân Phú	10.783786	106.637040	Hồ Chí Minh	2	Vietnamese Restaurant	Café	Fried Chicken Joint	Restaurant	Business Service
4	Quận Ba	10.774943	106.686280	Hồ Chí Minh	2	Vietnamese Restaurant	Café	Coffee Shop	Asian Restaurant	Vegetarian / Vegan Restaurant
5	Quận Bình Thạnh	10.803251	106.696665	Hồ Chí Minh	0	Vietnamese Restaurant	Seafood Restaurant	Supermarket	Convenience Store	Coffee Shop
6	Tân Túc	10.695412	106.591281	Hồ Chí Minh	0	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Tourist Information Center	Brewery	Flower Shop
8	Thủ Đức	10.848627	106.772089	Hồ Chí Minh	0	Coffee Shop	Ice Cream Shop	Asian Restaurant	Electronics Store	Vietnamese Restaurant
9	Quận Sáu	10.746795	106.649032	Hồ Chí Minh	0	Dessert Shop	Brewery	Noodle House	Cantonese Restaurant	Café
10	Quận Năm	10.755665	106.667451	Hồ Chí Minh	0	Dim Sum Restaurant	Asian Restaurant	Chinese Restaurant	Vietnamese Restaurant	Noodle House
11	Quận Một	10.780687	106.699444	Hồ Chí Minh	0	Café	Coffee Shop	Hotel	Vietnamese Restaurant	Massage Studio

Using folium library to visualize the result on the map, the clusters will be represented:



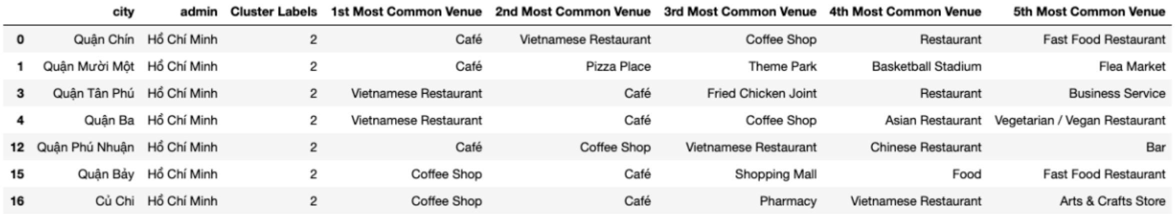
By exploring more detail on each cluster, the data showed that, it is recommended:

(i) to open the restaurant, the good location is in the *Cluster 0* which contains the neighborhoods as below:



	city	admin	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
2	Quận Mười	Hồ Chí Minh	0	Vietnamese Restaurant	Café	Coffee Shop	Asian Restaurant	Ice Cream Shop
5	Quận Bình Thạnh	Hồ Chí Minh	0	Vietnamese Restaurant	Seafood Restaurant	Supermarket	Convenience Store	Coffee Shop
6	Tân Túc	Hồ Chí Minh	0	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Tourist Information Center	Brewery	Flower Shop
8	Thủ Đức	Hồ Chí Minh	0	Coffee Shop	Ice Cream Shop	Asian Restaurant	Electronics Store	Vietnamese Restaurant
9	Quận Sáu	Hồ Chí Minh	0	Dessert Shop	Brewery	Noodle House	Cantonese Restaurant	Café
10	Quận Năm	Hồ Chí Minh	0	Dim Sum Restaurant	Asian Restaurant	Chinese Restaurant	Vietnamese Restaurant	Noodle House
11	Quận Một	Hồ Chí Minh	0	Café	Coffee Shop	Hotel	Vietnamese Restaurant	Massage Studio
13	Quận Bốn	Hồ Chí Minh	0	Seafood Restaurant	Hotel	Bar	Coffee Shop	Vietnamese Restaurant
17	Quận Hai	Hồ Chí Minh	0	Vietnamese Restaurant	Café	Seafood Restaurant	Noodle House	Electronics Store
18	Hóc Môn	Hồ Chí Minh	0	Vietnamese Restaurant	Hotel	Café	Seafood Restaurant	Market

(ii) to open the coffee shop, the good location is in the *Cluster 2* which contains the neighborhoods as below:



	city	admin	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Quận Chín	Hồ Chí Minh	2	Café	Vietnamese Restaurant	Coffee Shop	Restaurant	Fast Food Restaurant
1	Quận Mười Một	Hồ Chí Minh	2	Café	Pizza Place	Theme Park	Basketball Stadium	Flea Market
3	Quận Tân Phú	Hồ Chí Minh	2	Vietnamese Restaurant	Café	Fried Chicken Joint	Restaurant	Business Service
4	Quận Ba	Hồ Chí Minh	2	Vietnamese Restaurant	Café	Coffee Shop	Asian Restaurant	Vegetarian / Vegan Restaurant
12	Quận Phú Nhuận	Hồ Chí Minh	2	Café	Coffee Shop	Vietnamese Restaurant	Chinese Restaurant	Bar
15	Quận Bảy	Hồ Chí Minh	2	Coffee Shop	Café	Shopping Mall	Food	Fast Food Restaurant
16	Củ Chi	Hồ Chí Minh	2	Coffee Shop	Café	Pharmacy	Vietnamese Restaurant	Arts & Crafts Store

V. Discussion

Even though the algorithm generated the recommendation but actually there are several points need to be considered.

Firstly, from the data source point of view, it was not rich enough to analyse. It is both from the city data. To improve this barrier, the more detail data source would be collected, such as location of neighborhood at the ward level instead of district level as current situation.

Secondly, the business proposal used the simple features to analyse, that is venue categories. It would be suggested that the more features will be applied in the future version of the solution, such as *venue price, venue like, venue rate,..*

These above limitation can obviously found in the report when there is one missing city in the final result and the optimal K in the Elbow method looked not really good.

VI. Conclusion

Absolutely, machine learning could resolve many business problem nowadays but by this study, the important thing is the data for analyzing would be detail enough and also requires the analyst pay pretty much attention on exploring the data. Almost of the algorithms are integrated in the libraries and save a lot of effort in the data science project. In this example project, despite the are several aspects need to be improved but it definitely showed the result to audience in a pretty much visual way.