Applied Data Science Capstone Project

Using location based and clustering to build recommendation system

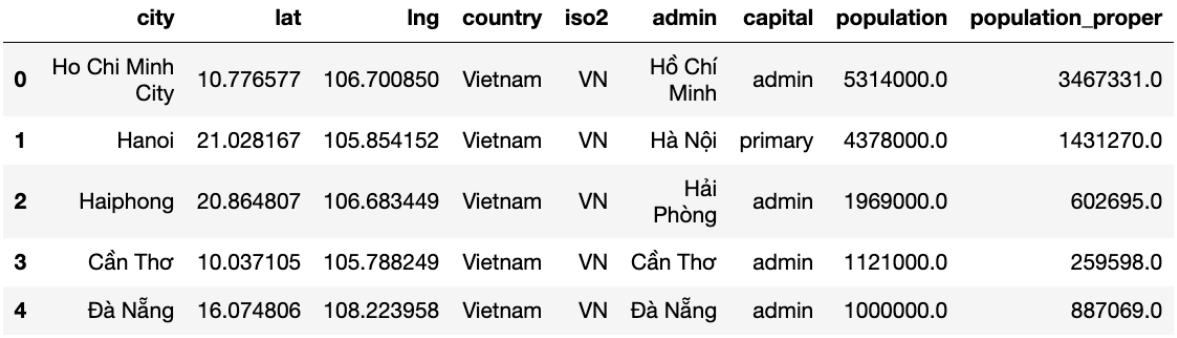
I. Introduction Business Problem.

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

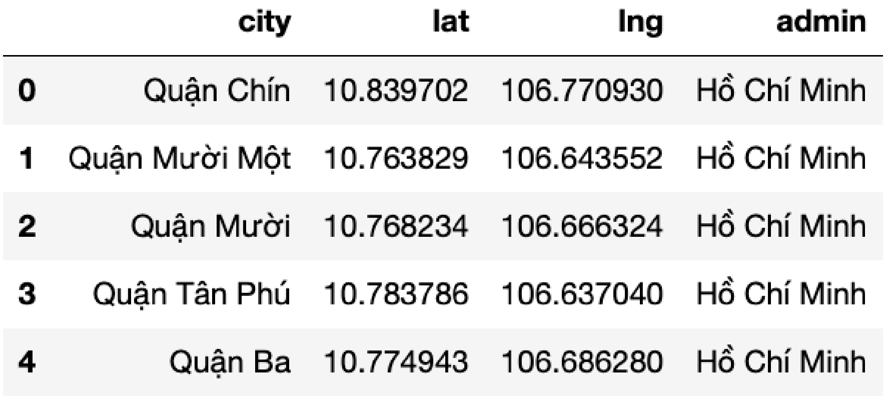
In solving this problem, the location data comes from a csv file which define the latitude, longitude and other information of all the cities in Vietnam as well as its neighborhoods. This is the sample data of city and its neighborhoods:



	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;