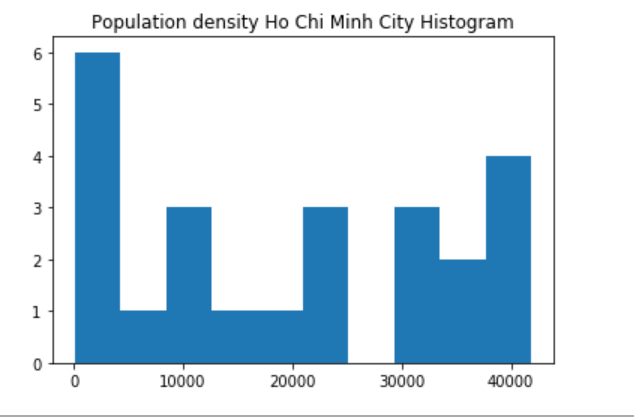
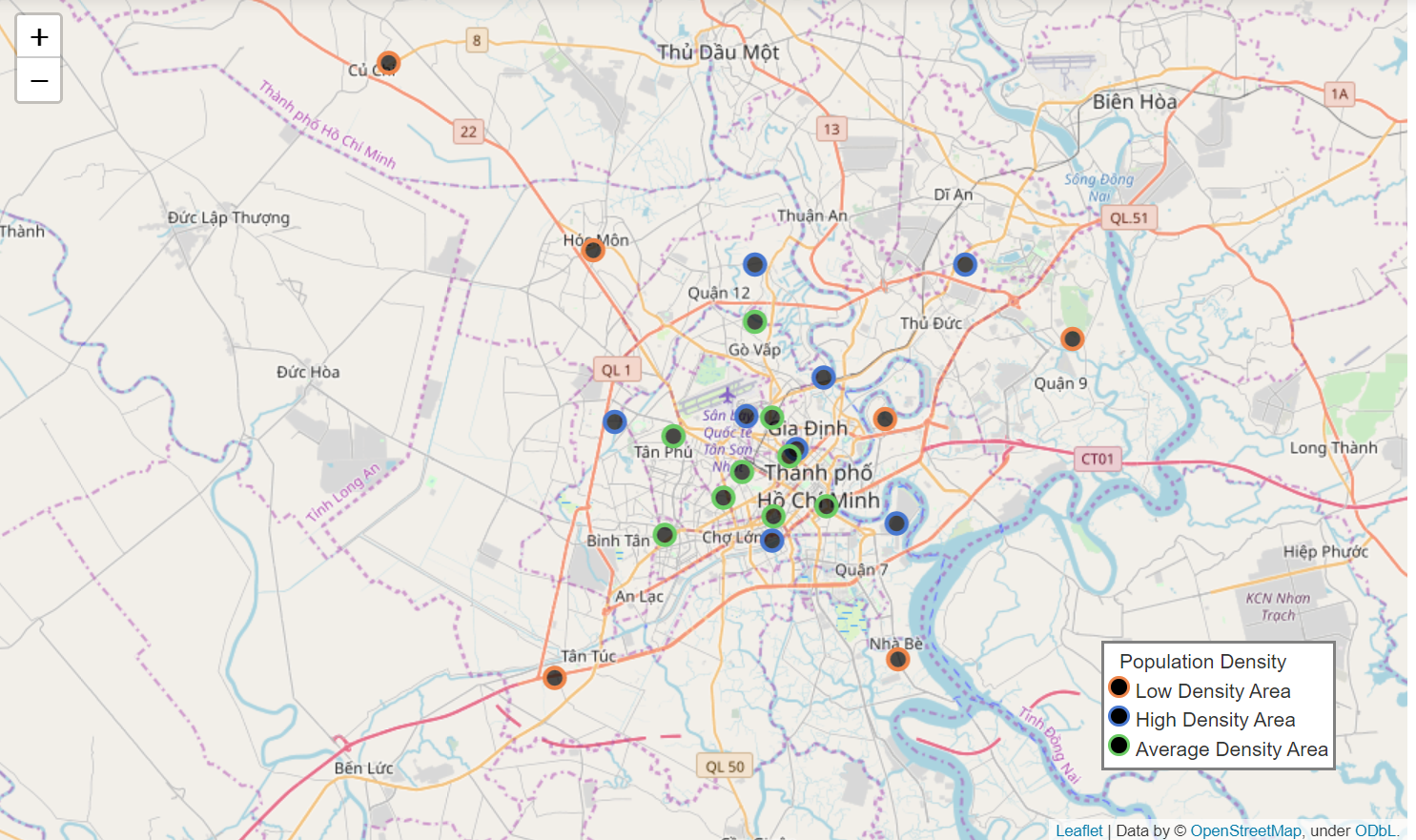
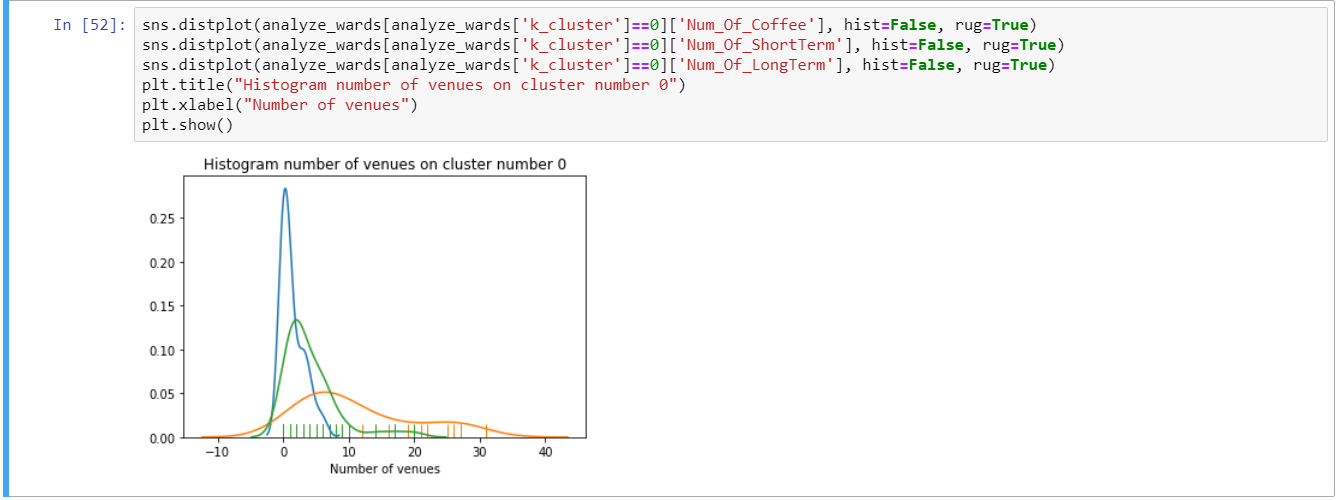
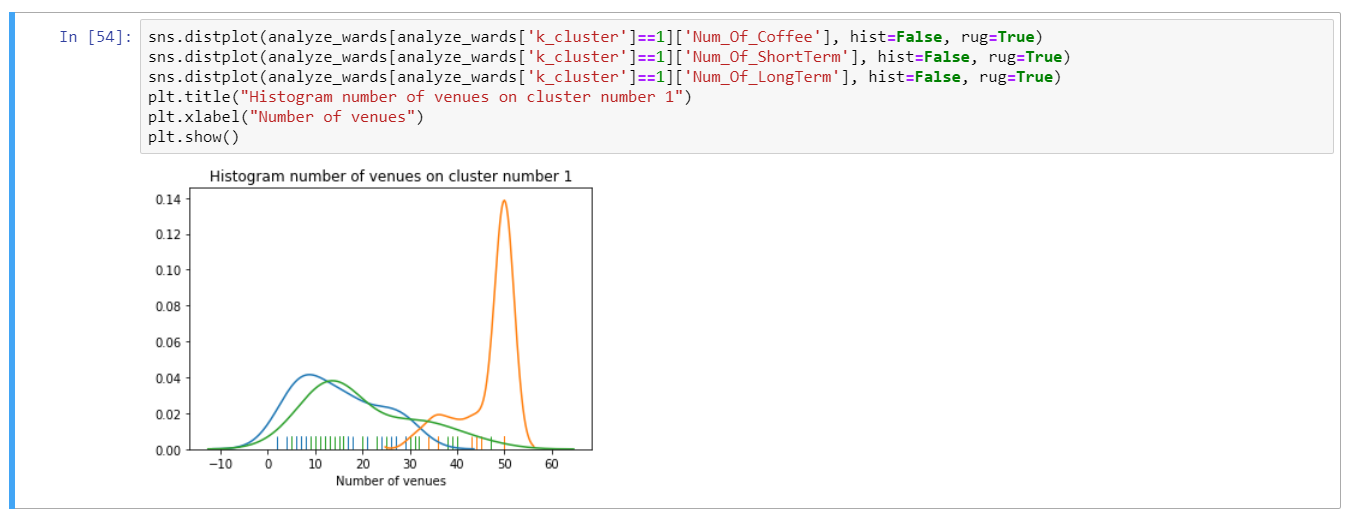
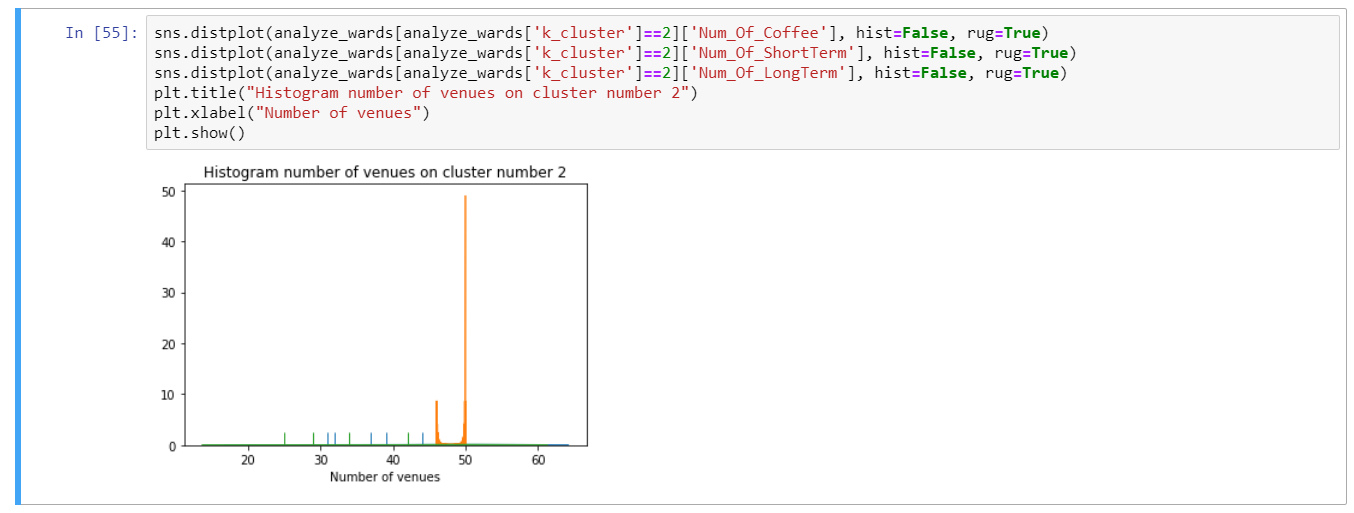
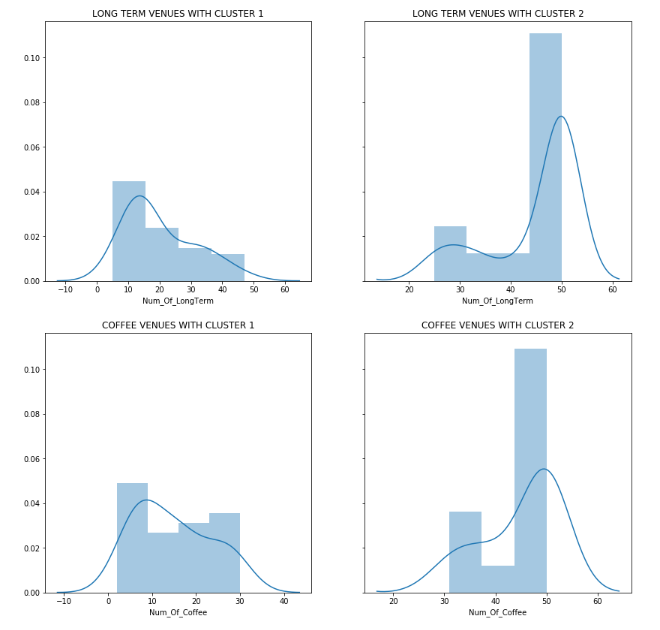
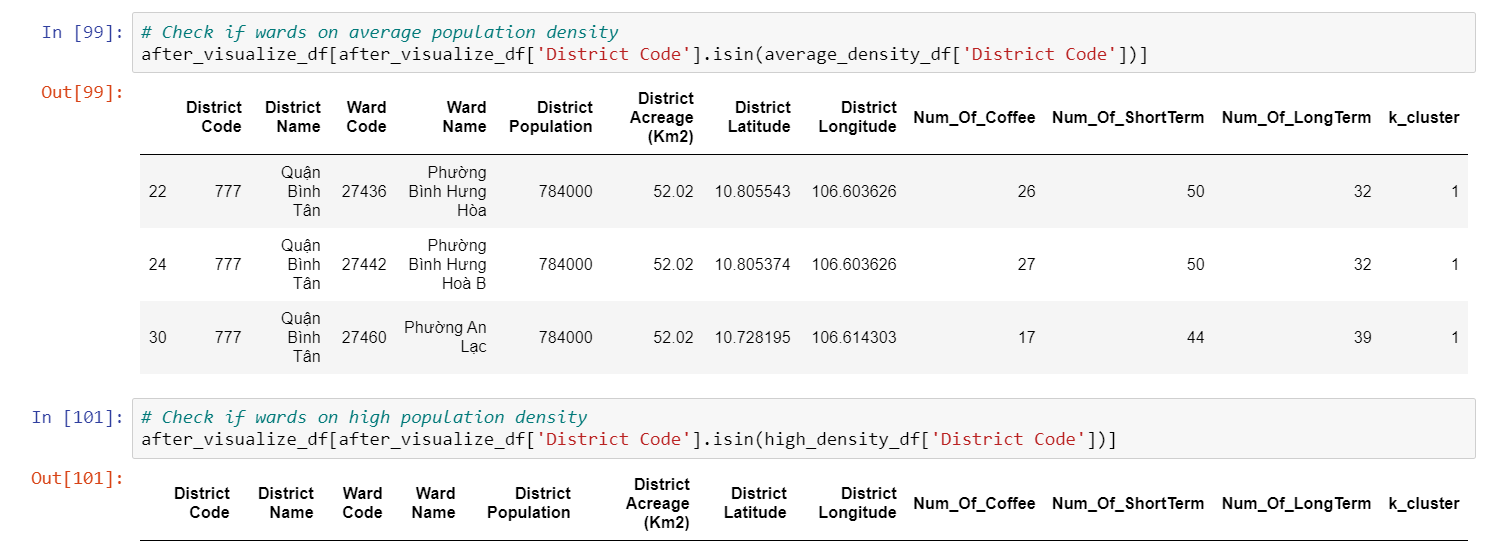
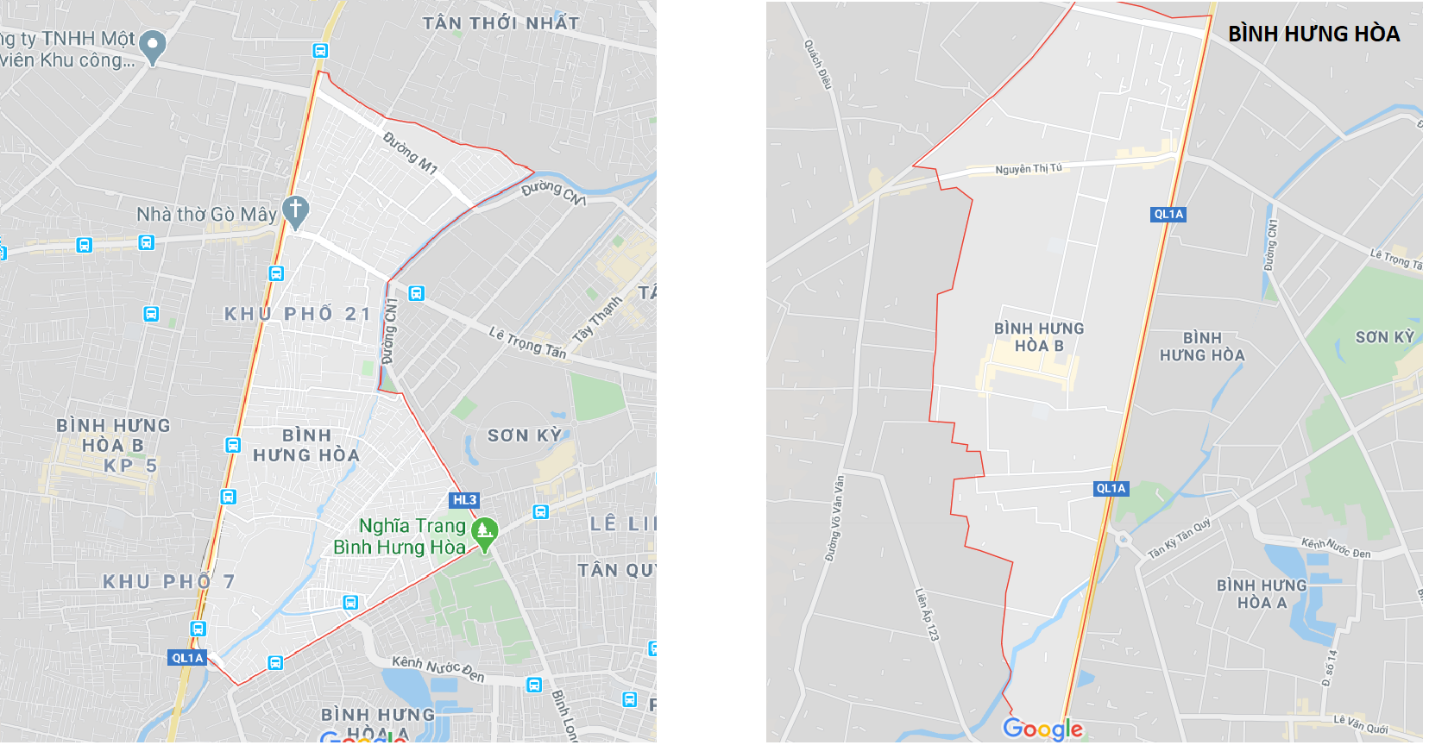
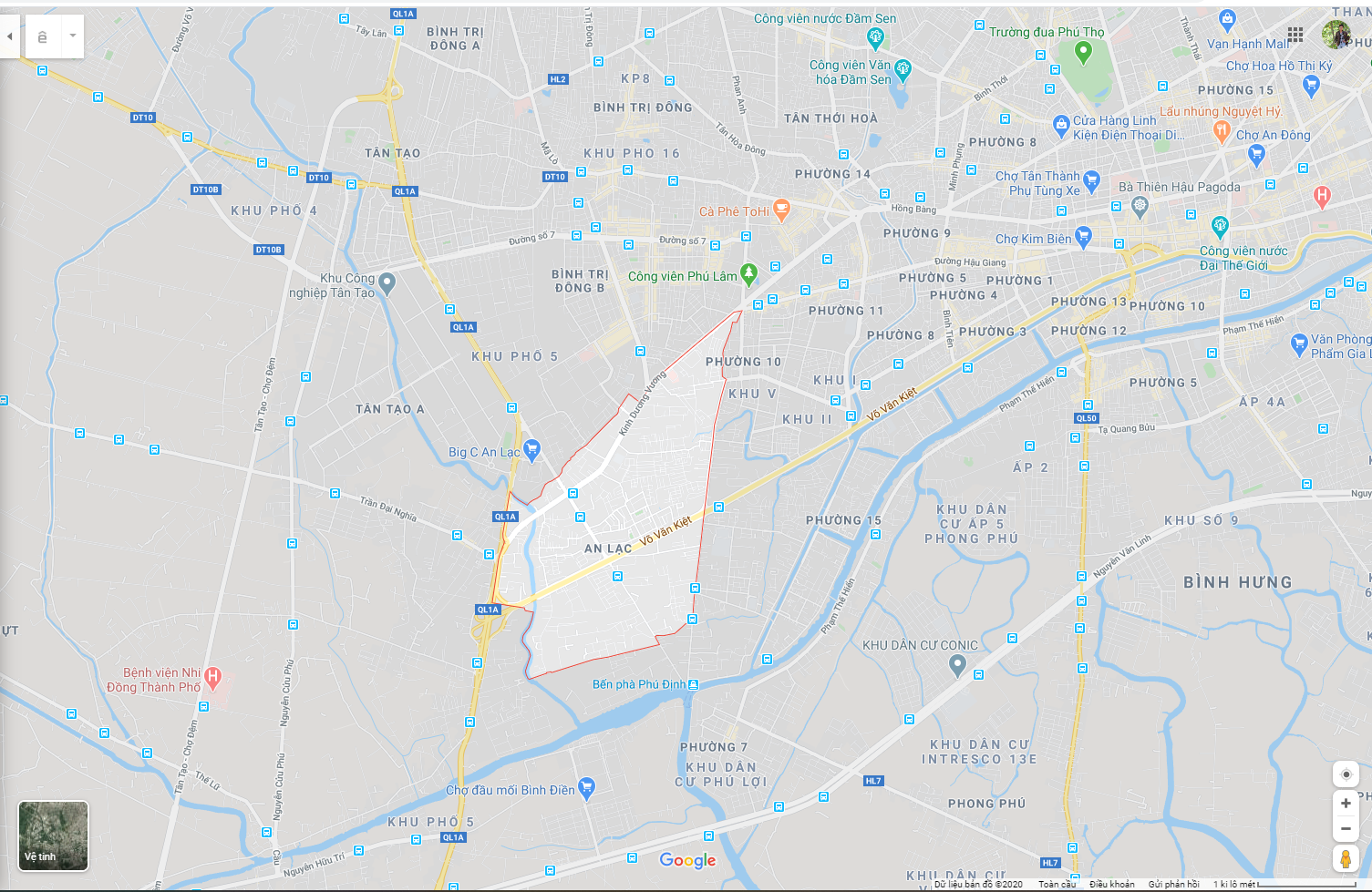
**BATTLECAPSTONE PROJECT: THE BATTLE OF THE NEIGHBORHOODS**

**Applied Data Science Capstone by IBM/Coursera**

**Open a coffee shop in Ho Chi Minh City**

***NGUYEN TAN HUY***

***March 17, 2020***

1. **INTRODUCTION**
   1. **Background**
      * Coffee is the most popular drinking in the world. The coffee industry is highly profitable. In Viet Nam, Coffee is the regular drinking. People drink coffee every morning. They are working, dating, playing game in the coffee shop. My company wants to open a new coffee shop in Ho Chi Minh City which is millions of coffee shops. We want to find locations that are not already crowded with the coffee shop in Ho Chi Minh City and many other services as possible.
   2. **Problems**
      * In Ho Chi Minh City, we have 24 districts with 324 wards. Some districts with high population density, other low population density. And we have millions of coffee shops in these districts, find out where can we open a new coffee shop is a difficult task. With a little help of data, we can have an overview of coffee shop market in Ho Chi Minh City. Find out the best wards to open a new coffee shop.
2. **DATA ACQUISITION AND DATA CLEANING**
   1. **Data Source**
      * District Code, District Name, Population, Acreage can be found in [Wikipedia](https://vi.wikipedia.org/wiki/Th%C3%A0nh_ph%E1%BB%91_H%E1%BB%93_Ch%C3%AD_Minh). Ward Code, Ward Name can be found in [official Ho Chi Minh City government website](http://www.pso.hochiminhcity.gov.vn/web/guest/home;jsessionid=110E7DB2FD8D09EF0F14A33EB055CC9E). Data on Wikipedia about population is also getting from official Ho Chi Minh City government website in 2018. New population data is not public yet.
   2. **Data collection and data cleaning**
      * Data download and scraped from multiple sources. Data from Wikipedia contain population and acreage without ward code and district code, data from official Ho Chi Minh City government website contain ward code and district code. We need to define a list of district code and combine two data frames into one data frame.
      * Getting longitude and latitude from google API. Because of the diverse shapes of the ward, and coordinates getting from API is not in the middle of the ward. I trying to get the exact central point of each ward hand by hand using google map.
      * Getting data from foursquare using radius in foursquare API. Radius will be found be getting average acreage each ward. Here is radius formula:
      * After we got longitude and latitude. We get venue location information using the foursquare API. We will get all venues with 3 [categories](https://developer.foursquare.com/docs/resources/categories) : coffee shop venues categories which are all kind of coffee shop in Ho Chi Minh City includes Cafeteria, Pet Café, Corporate Coffee Shop, Gaming Café… Long term venues includes: Residence, Home, Residential Building, Office building, government building which are all kind of venues provide the regular customer. Short term venues include all kind of shops, entertainment building… which are all kind of venues provide traveler, the non-regular customer.   
        We will analyze our data base on 2 features: average acreage per venues and average population per venues each district.
3. **EXPLORATORY DATA ANALYSIS**
   1. **Calculation of population density**
      * We don’t want our coffee shop open in rural area which is low population density. We will choose districts which are have high or average number of population density.  
          
          
          
        Most of the districts have population density higher than 10000 persons per acreage (km2). We will segment our districts into 3 segments low density which is lower than 10000, average density which is between 10000 and 30000 and high density which is higher than 30000 persons per acreage (km2). We only choose districts with average and high population density**(Conclusion 1)**.
      *   
        (Population map on folium)
      * We are only select districts which average and high-density areas (green and blue circle). They are in the center of the City.
   2. **Cluster districts using KMeans.**
      * Cluster districts into 4 segment:  
          
          
          
        We can easily observe that cluster number 0 with low number of Acreage per venues and low number of Population per venues. This means we have many venues in this cluster. We will have many competitors in these districts if we decide to open a new coffee shop. So we will exclude this cluster from the data frame, we only keep districts in cluster number 1, number 2, number 3. We got the total of 96 wards in the remaining district.
   3. **Cluster wards and find the best ward which could open a new coffee shop.**
      * Cluster wards in remain district into 3 cluster: analyze each cluster.  
          
        
      * With cluster number 0: There are less venues location on this cluster (less than 10 venues for each wards). This means they are rural areas with limit of entertainment, building and other services. We don’t want to open a new coffee shop in this area. We will removed this cluster from our data frame.  
          
          
        
      * These clusters have a high number of venues for each ward. We will keep them and continue to analyze them. They also have a high number (around 50 venues each ward) of short term venues which is entertainment services. We don’t need to analyze number of short term venues on two remaining clusters, because they are the same. We are continuing to analyze the number of coffee and the number of long term venues.  
        
      * Cluster number 1 with low number of long term venues than cluster number 2. We also have low number of coffee shops in cluster number 1 than cluster number 2. So we will choose cluster number 1 with low competitors but only choose the wards with high number of long term venues in this cluster (higher than 30 venues).
   4. **Final data frame after analyze**  
      * We only choose wards in average and high population density area in [conclusion number 1](#Condition1).  
          
        We only got 3 wards in the average population density. This is final analyze. We will discuss and choose 1 ward to open new coffee shop
4. **DISCUSSION AND RESULT**
   1.   
      ( Image of BINH HUNG HOA AND BINH HUNG HOA B ward on google maps )  
        
      Two wards are in the same district and same area, separated by a road. But they are far from Ho Chi Minh downtown. And we have more coffee shop venues on these 2 wards (26 and 27 coffee shop venues). So we will not choose these 2 wards to open a new coffee shop.
   2.   
      ( Image of AN LAC WARD on google map )  
        
      This ward near Ho Chi Minh City downtown, border of District 10 which is high population and high density of administrative and career agencies. Võ Văn Kiệt Street and Kinh Duong Vuong Street are arterial roads in Ho Chi Minh City which is run through most of this wards area. We have lot of department, one industry park, 2 park in around this ward. There are less coffee shop venues than other wards (17 coffee shop venues).
5. **CONCLUSION**  
   We will open new coffee shop in AN LAC WARD with ward code 27460.