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

1.1. 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.

1.2. 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

2.1. Data Source

- District Code, District Name, Population, Acreage can be found in Wikipedia. Ward Code, Ward Name can be found in official Ho Chi Minh City government website. 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.

2.2. 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:

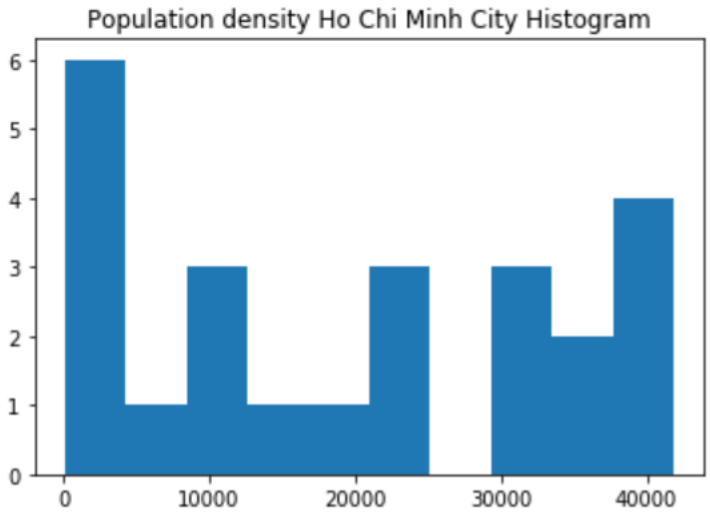
$$Radius = \sqrt{average\frac{acreage}{\pi}}$$

- After we got longitude and latitude. We get venue location information using the foursquare API. We will get all venues with 3 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.

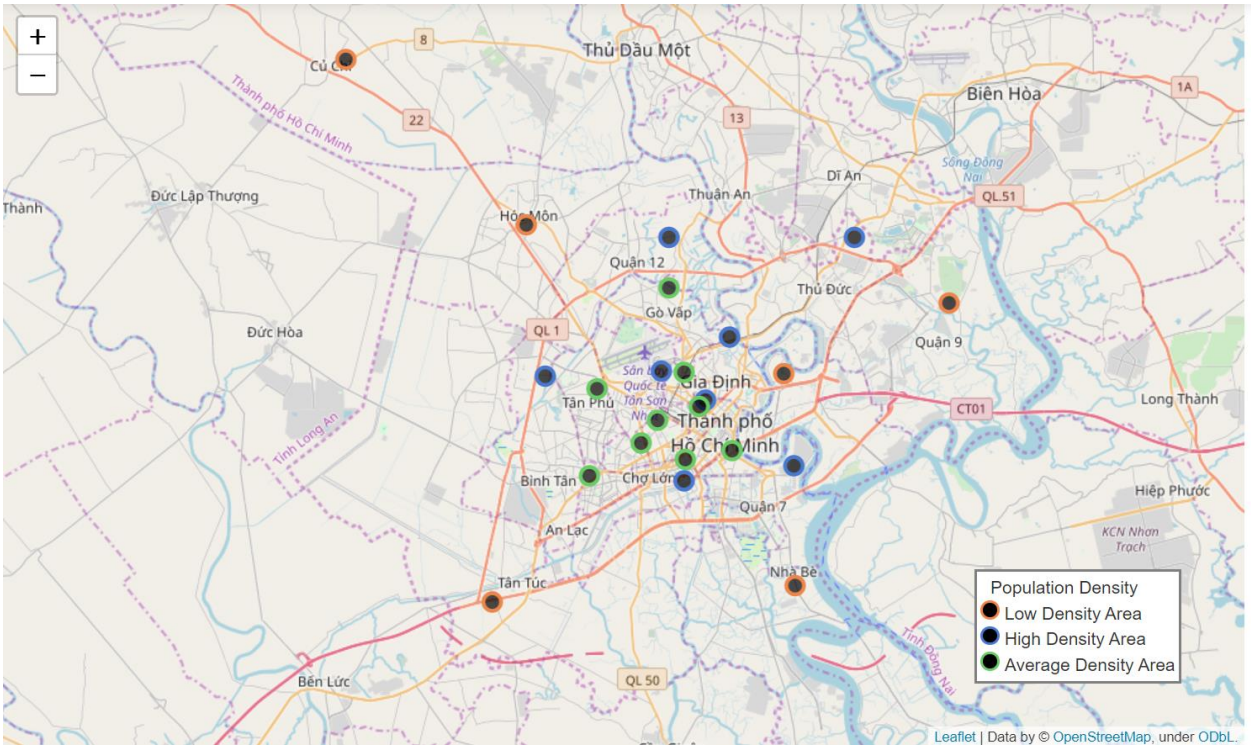
3. EXPLORATORY DATA ANALYSIS

3.1. 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 (km^2). 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 (km^2). 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.

3.2. Cluster districts using KMeans.

- Cluster districts into 4 segment:

Out[73]:

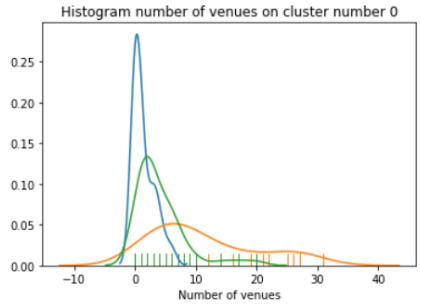
	District Code	District Name	Acreage/Coffee	Acreage/ShortTerm	Acreage/LongTerm	Population/Coffee	Population/ShortTerm	Population/LongTerm	k_Cluster
0	760	Quận 1	17.1556	21.2672	18.2506	315.556	391.185	335.697	0
1	761	Quận 12	224.426	108.967	196.059	2638.3	1280.99	2304.83	3
2	762	Quận Thủ Đức	206.926	95.6	185.271	2562.77	1184	2294.57	3
3	763	Quận 9	525.346	253.898	326.648	1829.49	884.187	1137.54	0
4	764	Quận Gò Vấp	40.5133	25.7908	45.6713	1388.09	883.66	1564.81	0
5	765	Quận Bình Thạnh	24.1067	22.4406	24.3041	578.886	538.877	583.626	0
6	766	Quận Tân Bình	31.2396	36.7705	31.3706	660.167	777.049	662.937	0
7	767	Quận Tân Phú	40.7398	29.1956	42.7005	1237.24	886.654	1296.79	0
8	768	Quận Phú Nhuận	7.58942	6.50667	7.79553	253.499	217.333	260.383	0
9	769	Quận 2	212.778	127.34	146.441	769.231	460.358	529.412	0
10	770	Quận 3	7.43202	7.03863	8.25503	287.009	271.817	318.792	0
11	771	Quận 10	9.0938	7.62667	8.62745	372.019	312	352.941	0
12	772	Quận 11	20.9796	6.9932	17.1333	853.061	284.354	696.667	0
13	773	Quận 4	17.7119	6.211	12.8615	741.525	260.03	538.462	0
14	774	Quận 5	12.0621	5.69333	8.10247	449.153	212	301.708	0
15	775	Quận 6	63.75	12.3958	44.625	2080.36	404.514	1456.25	0
16	776	Quận 8	65	28.9985	43.73	1442.18	643.399	970.252	0
17	777	Quận Bình Tân	230.177	116.637	199.31	3469.03	1757.85	3003.83	3
18	778	Quận 7	97.7808	91.2788	85.7933	986.301	920.716	865.385	0
19	783	Huyện Củ Chi	7246.17	1274.99	3423.39	7700	1354.84	3637.8	2
20	784	Huyện Hóc Môn	565.648	295.854	791.087	2808.29	1468.83	3927.54	3
21	785	Huyện Bình Chánh	2745.22	825.359	1350.59	7663.04	2303.92	3770.05	2
22	786	Huyện Nhà Bè	1304.29	554.862	707.254	2675.32	1138.12	1450.7	3
23	787	Huyện Cần Giờ	78272.2	7915.17	14676	7888.89	797.753	1479.17	1

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.

3.3. Cluster wards and find the best ward which could open a new coffee shop.

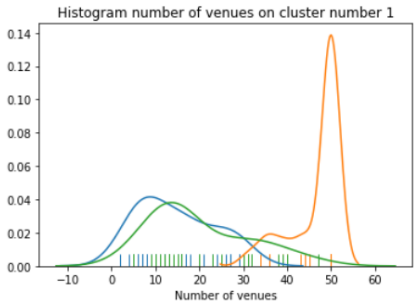
- Cluster wards in remain district into 3 cluster: analyze each cluster.

```
In [52]: sns.distplot(analyze_wards[analyze_wards['k_cluster']==0]['Num_Of_Coffee'], hist=False, rug=True)
         sns.distplot(analyze_wards[analyze_wards['k_cluster']==0]['Num_Of_ShortTerm'], hist=False, rug=True)
         sns.distplot(analyze_wards[analyze_wards['k_cluster']==0]['Num_Of_LongTerm'], hist=False, rug=True)
         plt.title("Histogram number of venues on cluster number 0")
         plt.xlabel("Number of venues")
         plt.show()
```

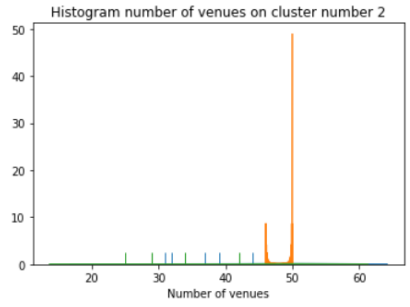


- With cluster number 0: There are less venues location on this cluster (less than 10 venues for each wards). This means they are rural areas with limit of entertainment, building and other services. We don't want to open a new coffee shop in this area. We will removed this cluster from our data frame.

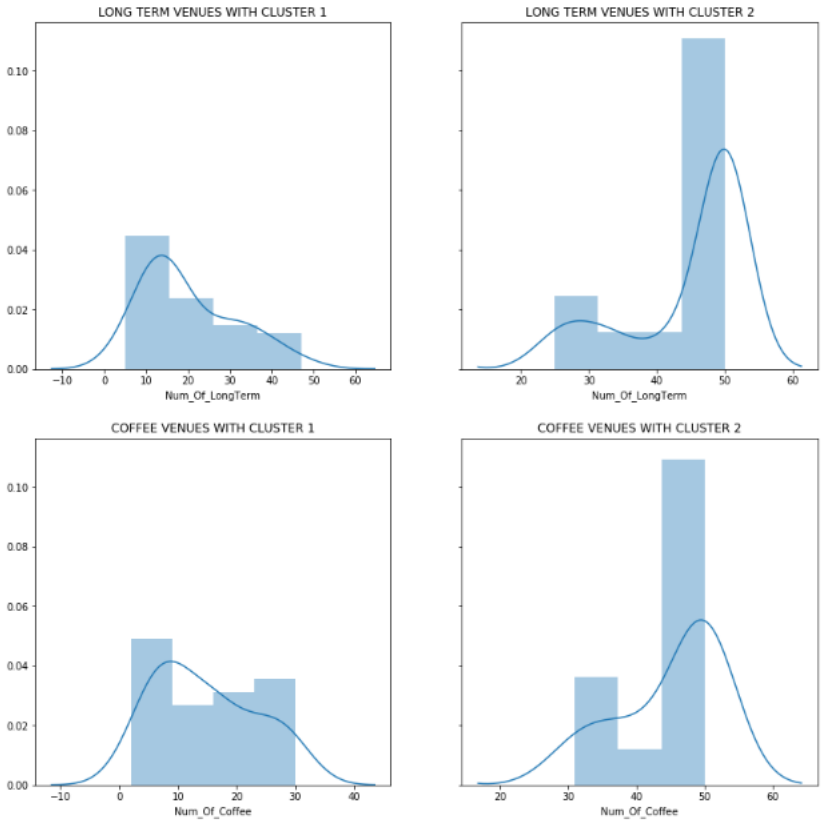
```
In [54]: sns.distplot(analyze_wards[analyze_wards['k_cluster']==1]['Num_Of_Coffee'], hist=False, rug=True)
         sns.distplot(analyze_wards[analyze_wards['k_cluster']==1]['Num_Of_ShortTerm'], hist=False, rug=True)
         sns.distplot(analyze_wards[analyze_wards['k_cluster']==1]['Num_Of_LongTerm'], hist=False, rug=True)
         plt.title("Histogram number of venues on cluster number 1")
         plt.xlabel("Number of venues")
         plt.show()
```




```
In [55]: sns.distplot(analyze_wards[analyze_wards['k_cluster']==2]['Num_Of_Coffee'], hist=False, rug=True)
         sns.distplot(analyze_wards[analyze_wards['k_cluster']==2]['Num_Of_ShortTerm'], hist=False, rug=True)
         sns.distplot(analyze_wards[analyze_wards['k_cluster']==2]['Num_Of_LongTerm'], hist=False, rug=True)
         plt.title("Histogram number of venues on cluster number 2")
         plt.xlabel("Number of venues")
         plt.show()
```



- 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).

3.4. Final data frame after analyze

- We only choose wards in average and high population density area in conclusion number 1.

In [99]:
```
# Check if wards on average population density
after_visualize_df[after_visualize_df['District Code'].isin(average_density_df['District Code'])]
```

Out[99]:

	District Code	District Name	Ward Code	Ward Name	District Population	District Acreage (Km2)	District Latitude	District Longitude	Num_Of_Coffee	Num_Of_ShortTerm	Num_Of_LongTerm	k_cluster
22	777	Quận Bình Tân	27436	Phường Bình Hưng Hòa	784000	52.02	10.805543	106.603626	26	50	32	1
24	777	Quận Bình Tân	27442	Phường Bình Hưng Hoà B	784000	52.02	10.805374	106.603626	27	50	32	1
30	777	Quận Bình Tân	27460	Phường An Lạc	784000	52.02	10.728195	106.614303	17	44	39	1

In [101]:
```
# Check if wards on high population density
after_visualize_df[after_visualize_df['District Code'].isin(high_density_df['District Code'])]
```

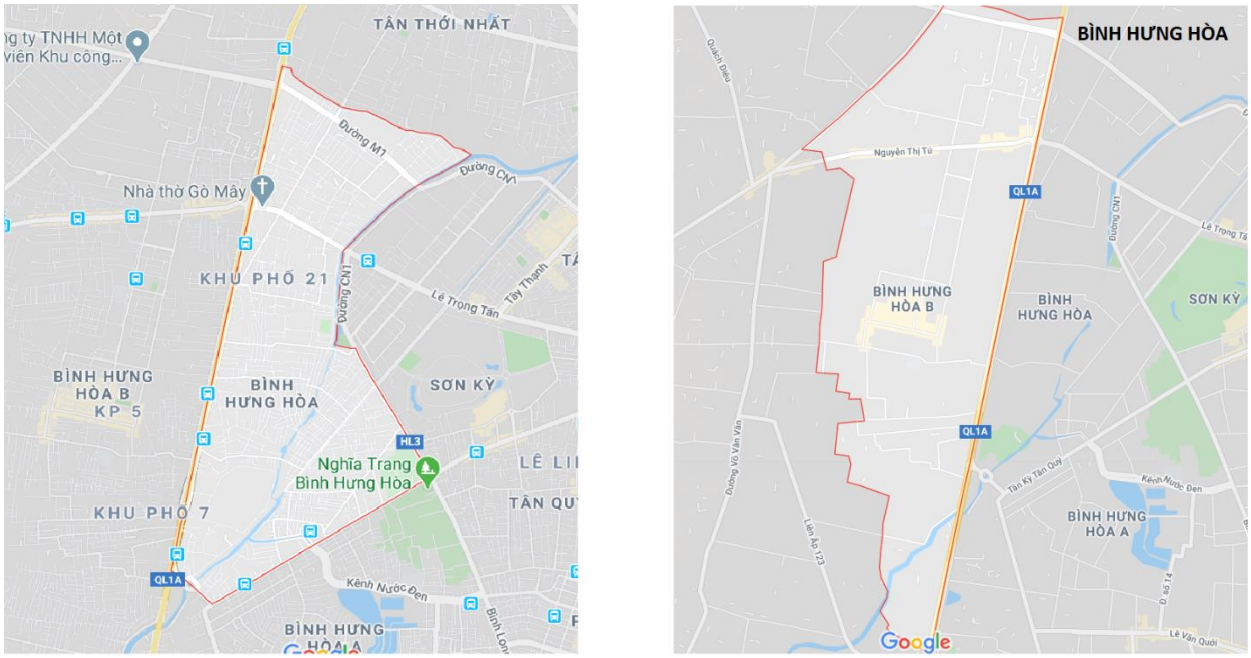
Out[101]:

	District Code	District Name	Ward Code	Ward Name	District Population	District Acreage (Km2)	District Latitude	District Longitude	Num_Of_Coffee	Num_Of_ShortTerm	Num_Of_LongTerm	k_cluster

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

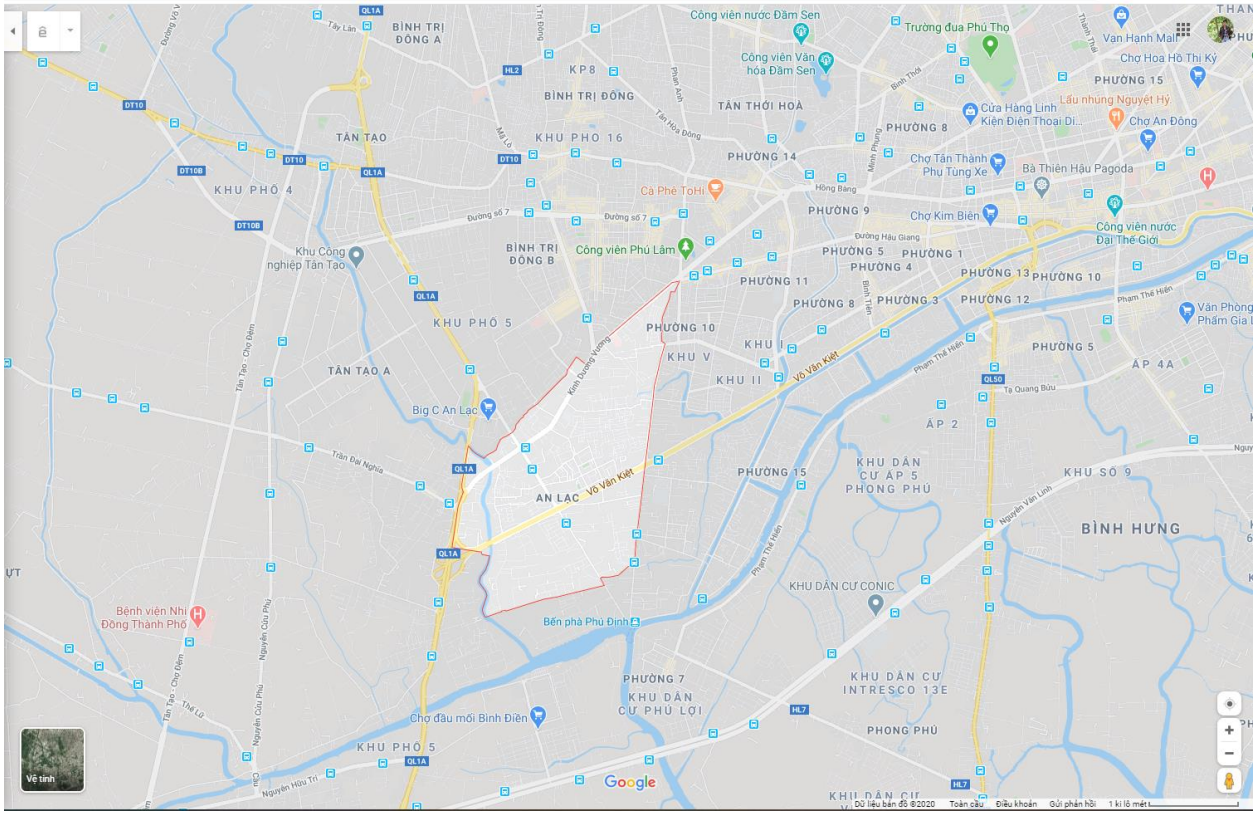
4.1.



(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.

4.2.



(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.