Capstone Project - The Battle of Neighborhoods

Geolocation analysis for opening shopping mall in Bangkok

Phiphat Prapapanpong

January 12, 2020

1. Introduction

Bangkok is the capital and most populous city of Thailand. Over 8 million people lived within Bangkok and this city is the most visited city in the world with approximately 23 million tourists per year. The shopping malls are one of the places that Thai people and tourists usually visit and spend a lot of money. Therefore, if we are able to help decision making on finding the best location to build a shopping mall. It can reduce the risk for investors and also able to attract more tourists to visit and spend. Businesses can make more informed decisions that can improve both efficiency and effectiveness.

1.1 Business Problem

In this project, I will try to understand the preferences of each district by leveraging venue data from Foursquare's 'Places API' and 'k-means clustering' machine learning algorithm.

2. Data Acquisition and Cleansing

2.1 Data sources

Based on definition of our problem the following sources will be needed for the analysis

- List of districts in Bangkok. This defines the scope of this project which is confined to the city of Bangkok from https://en.wikipedia.org/wiki/List_of_districts_of_Bangkok, In this data, we can see district code, district name, population, latitude and longitude
- The list of Latitude and longitude coordinates of these districts is required in order to plot the map and also to get the nearby venue data from Foursquare
- Venue data from Foursquare API, particularly data related to shopping malls. I will use this data to perform clustering on the districts.

2.2 Data extraction and data cleansing

Web scraping technique “pandas.io.html.read_html” was used to extract the data from the Wikipedia page “https://en.wikipedia.org/wiki/List_of_districts_of_Bangkok”,

```
from pandas.io.html import read_html
url='https://en.wikipedia.org/wiki/List_of_districts_of_Bangkok'
wikitable = read_html(url,  attrs={"class":"wikitable"},header =0)
wiki_df = pd.DataFrame(wikitable[0])
wiki_df= wiki_df.rename(columns={'District(Khet)':'District'})
wiki_df= wiki_df.rename(columns={'No. ofSubdistricts(Khwaeng)':'No_of_Subdistricts'})
wiki_df= wiki_df.rename(columns={'Thai':'DistrictThai'})
wiki_df['Latitude'].fillna(0, inplace=True)
wiki_df['Longitude'].fillna(0, inplace=True)
wiki_df.head(11)
```

This data provides information of Bangkok subdivided into 50 districts including district name, district code, latitude, longitude and also provide population of each districts.

	District	Code	DistrictThai	Population	No_of_Subdistricts	Latitude	Longitude
0	Bang Bon	50	บางบอน	105161	4	0.000000	0.000000
1	Bang Kapi	6	บางกะปิ	148465	2	13.765833	100.647778
2	Bang Khae	40	บางแค	191781	4	13.696111	100.409444
3	Bang Khen	5	บางเขน	189539	2	13.873889	100.596389
4	Bang Kho Laem	31	บางคอแหลม	94956	3	13.693333	100.502500

However, some records do not provide data as show in the figure below;

	District	Code	DistrictThai	Population	No_of_Subdistricts	Latitude	Longitude
0	Bang Bon	50	บางบอน	105161	4	0.0	0.0
19	Khan Na Yao	43	คันนายาว	88678	2	0.0	0.0
44	Thawi Watthana	48	ทวีวัฒนา	76351	2	0.0	0.0
46	Thung Khru	49	ทุ่งครุ	116473	2	0.0	0.0
47	Wang Thonglang	45	วังทองหลาง	114768	4	0.0	0.0

Python Geocoder package used to extract the latitude and longitude coordinates of these records After that, we will use Foursquare API to get the venue data for those districts. Foursquare API will provide many categories of the venue data, we are particularly interested in the Shopping Mall category in order to help us to solve the business problem. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium).

3. Exploratory Data Analysis

3.1 Explore dataset

In this section, we explore the data with Pandas, Matplotlib and Map visualization work library, namely Folium. After extract the data from website, we populate the data into a pandas DataFrame and then visualize the data in a map using Folium package

According data from Wiki, there are 50 districts in Bangkok and total population of Bangkok is approximately 5,671,070 people.

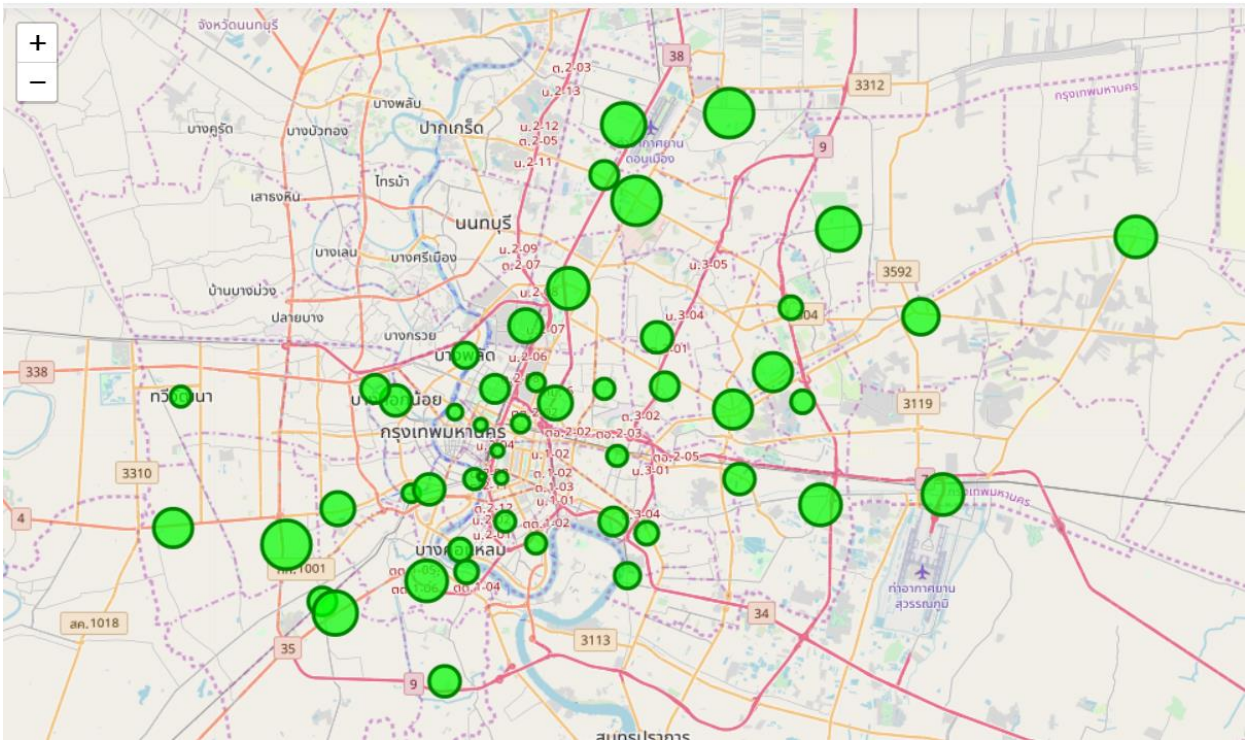
```
1  wiki_df.shape
```

```
(50, 7)
```

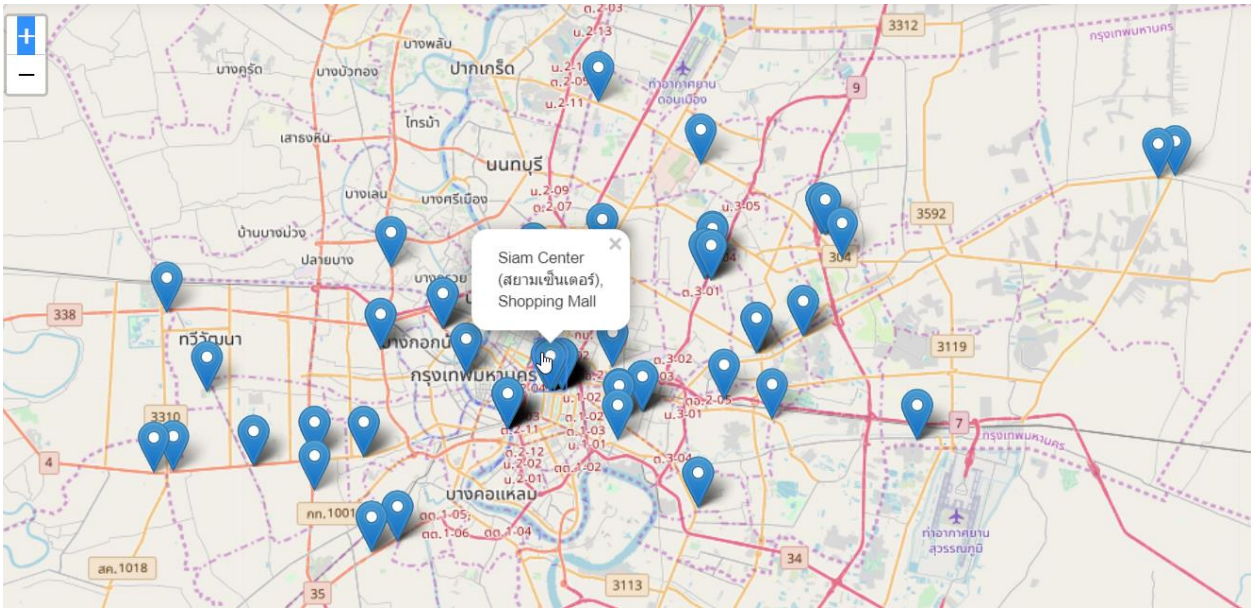
```
1  wiki_df['Population'].sum()
```

```
5671070
```

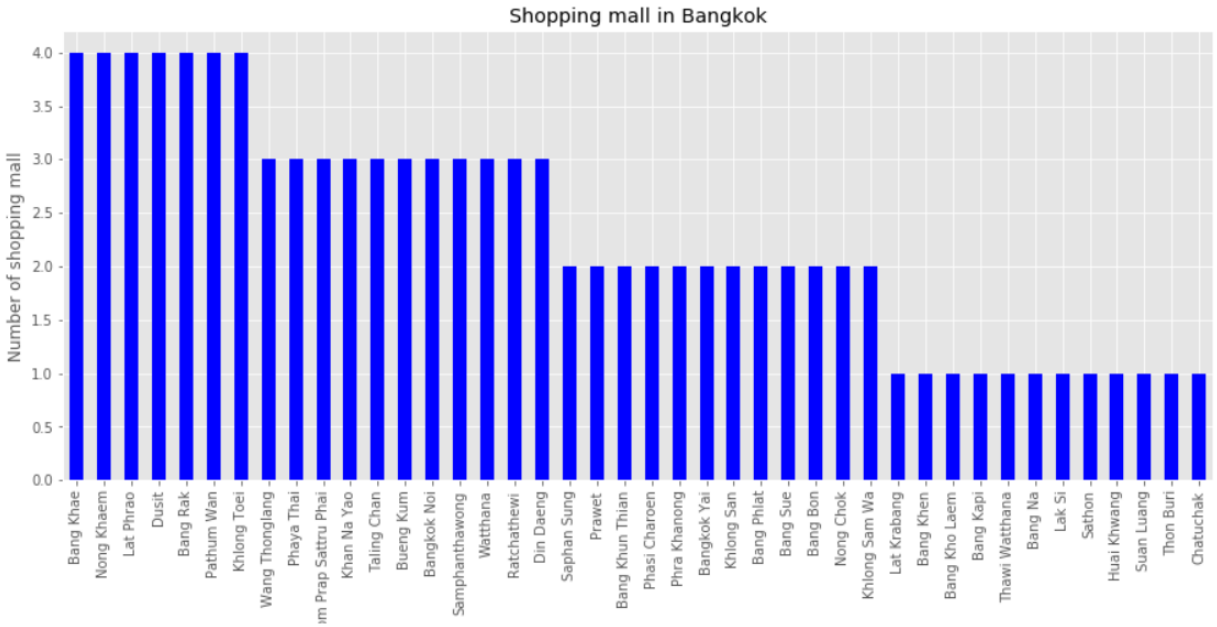
From map visualization, we found that people in Bangkok lived in suburb area as show in the figure below. Size of circle represent population in each district.



For gathering more information, Foursquare API was used to get the top 100 venues that are within a radius of 5000 meters. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. Then visualize shopping mall category of venues in map.



The total number of shopping mall in each district shows in the bar chart below;



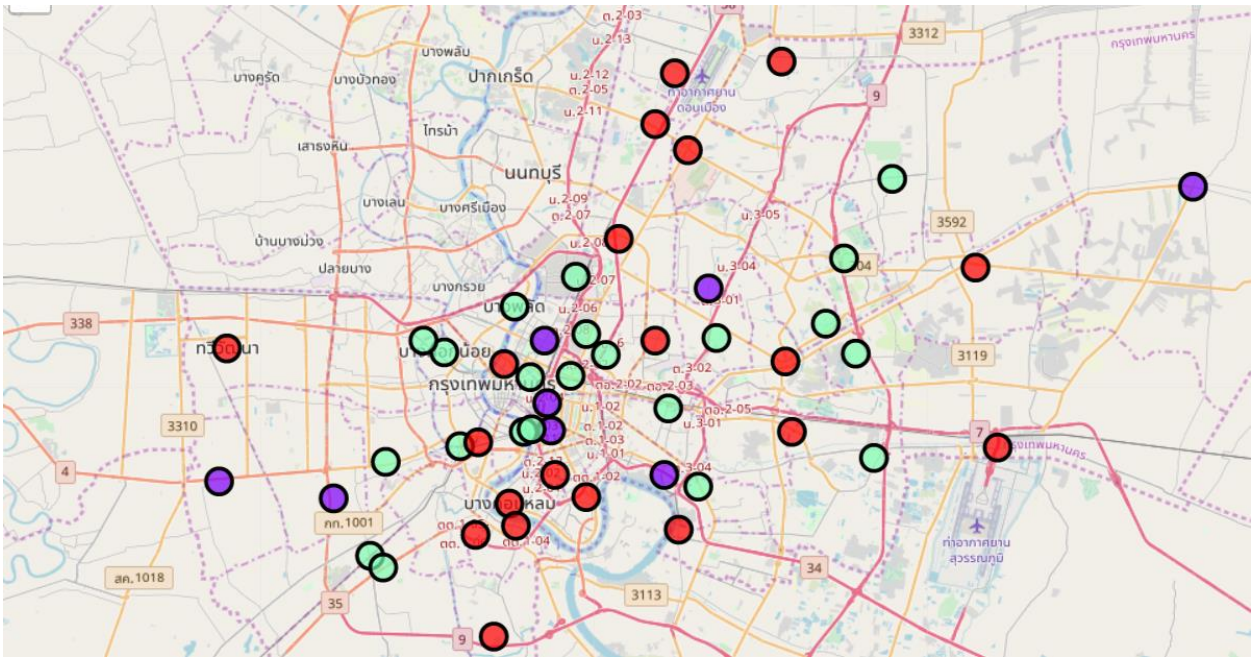
4. Machine Learning

We will cluster the district into 3 clusters based on their frequency of occurrence for “Shopping Mall”. Before using K-Mean, we have to transform data using one-hot encoder.

	District	Airport	Airport Lounge	Airport Service	American Restaurant	Art Gallery	Art Museum	Asian Restaurant	Athletics & Sports	BBQ Joint	...	Veterinarian	Vietnamese Restaurant	Water Park	Whisky Bar	Wine Bar	Wine Shop
0	Bang Bon	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
1	Bang Bon	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
2	Bang Bon	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
3	Bang Bon	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
4	Bang Bon	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0

The results will allow us to identify which districts have a higher concentration of shopping malls while which districts have a fewer number of shopping malls by using k-means clustering. Based on the occurrence of shopping malls in a different district, it will help us to plan which district we should acquire land for developing new shopping mall.

	District	Shopping Mall	Cluster	Code	DistrictThai	Population	No_of_Subdistricts	Latitude	Longitude
0	Bang Bon	0.02	2	50	บางบอน	105161	4	13.666503	100.428859
1	Bang Kapi	0.01	0	6	บางกะปิ	148465	2	13.765833	100.647778
2	Bang Khae	0.04	1	40	บางแค	191781	4	13.696111	100.409444
3	Bang Khen	0.01	0	5	บางเขน	189539	2	13.873889	100.596389
4	Bang Kho Laem	0.01	0	31	บางคอแหลม	94956	3	13.693333	100.502500



5. Results

From the results, we found that the shopping malls were clustered using number of shopping malls in each district, so it can be classified to high, medium and low number fo shopping mall in each area. However, there are other factors which can be used to cluster such as number of tourists, income of people around that area, number of people traveled around these areas. These information can make clustering more accurate and will generate more information for decision making.

	District	Shopping Mall	Cluster	Code	DistrictThai	Population	No_of_Subdistricts	Latitude	Longitude
1	Bang Kapi	0.01	0	6	บางกะปิ	148465	2	13.765833	100.647778
3	Bang Khen	0.01	0	5	บางเขน	189539	2	13.873889	100.596389
4	Bang Kho Laem	0.01	0	31	บางคอแหลม	94956	3	13.693333	100.502500
6	Bang Na	0.01	0	47	บางนา	95912	2	13.680081	100.591800
13	Chatuchak	0.01	0	30	จตุจักร	160906	5	13.828611	100.559722

	District	Shopping Mall	Cluster	Code	DistrictThai	Population	No_of_Subdistricts	Latitude	Longitude
2	Bang Khae	0.04	1	40	บางแค	191781	4	13.696111	100.409444
8	Bang Rak	0.04	1	4	บางรัก	45875	5	13.730833	100.524167
17	Dusit	0.04	1	2	ดุสิต	107655	5	13.776944	100.520556
22	Khlong Toei	0.04	1	33	คลองเตย	109041	3	13.708056	100.583889
25	Lat Phrao	0.04	1	38	ลาดพร้าว	122182	2	13.803611	100.607500

	District	Shopping Mall	Cluster	Code	DistrictThai	Population	No_of_Subdistricts	Latitude	Longitude
0	Bang Bon	0.02	2	50	บางบอน	105161	4	13.666503	100.428859
5	Bang Khun Thian	0.02	2	21	บางขุนเทียน	165491	2	13.660833	100.435833
7	Bang Phlat	0.02	2	25	บางพลัด	99273	4	13.793889	100.505000
9	Bang Sue	0.02	2	29	บางซื่อ	132234	2	13.809722	100.537222
10	Bangkok Noi	0.03	2	20	บางกอกน้อย	117793	5	13.770867	100.467933

6. Conclusion

According to objective of this project we would like to support decision making on planning to develop new shopping mall in Bangkok, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new shopping mall.

The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new shopping mall.

7. References

- List of districts in Bangkok

  https://en.wikipedia.org/wiki/List_of_districts_of_Bangkok

- Foursquare Developers Documentation. Foursquare. Retrieved from

  https://developer.foursquare.com/docs