Capstone Project – The Battle of Neighborhoods

Location analysis for launching new hotel in Bangkok

1. Business Problem

Bangkok is the capital of Thailand. Before the pandemic began, Bangkok was one of the world's top tourist destination cities. Each year approximately 22.7 million international visitors arrive in Bangkok. Also, MasterCard ranked Bangkok as the top destination for global travelers. Till now, there are millions of people have been vaccinated around the world. The travelling will be return soon. To serve the travelers around the world are looking forward to visiting Bangkok, the lodging is dedicated to perfecting the travel experience.

Therefore, if we can help decision makers on finding the potential location to launch new hotel. It can reduce the risk of investors.

This project aims to provide the stakeholder with necessary information for example, name and location of Bangkok districts and the number of hotels in each district. Also, generate hotel maps in Bangkok

Also, this project tries 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

Based on definition of our problem and the aim of this project the following sources will be used for analysis

2.1. Data sources

- The list of districts get from this https://en.wikipedia.org/wiki/List_of_districts_of_Bangkok link
- Venue data from Foursquare API, particular data related to hotel. I will use this data to perform clustering on the districts

2.2. Data extraction

pandas.io.html is web scraping technique which use to extract the data from Wikipedia page

https://en.wikipedia.org/wiki/List_of_districts_of_Bangkok to get Bangkok districts

```
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)':'DistrictEng'})
wiki_df= wiki_df.rename(columns={'MapNr':'code'})
wiki_df= wiki_df.rename(columns={'Post-code':'Postcode'})
wiki_df= wiki_df.rename(columns={'Thai':'District'})
wiki_df= wiki_df.rename(columns={'Popu-lation':'Population'})
wiki_df= wiki_df.rename(columns={'No. ofSubdis-trictsKhwaeng':'No_of_Subdistricts'})
wiki_df.head()
```

Below is the sample of Bangkok districts data from Wikipedia which contain district name in both English and local language, district code, postcode, population, latitude, and longitude

	DistrictEng	code	Postcode	District	Population	No_of_Subdistricts	Latitude	Longitude
0	Bang Bon	50	10150	บางบอน	105161	4	13.659200	100.399100
1	Bang Kapi	6	10240	บางกะปิ	148465	2	13.765833	100.647778
2	Bang Khae	40	10160	บางแค	191781	4	13.696111	100.409444
3	Bang Khen	5	10220	บางเขน	189539	2	13.873889	100.596389
4	Bang Kho Laem	31	10120	บางคอแหลม	94956	3	13.693333	100.502500

3. Exploratory data analysis

3.1. Explore dataset

In this section, we explore the data with Pandas, Matplotlib and map visualization with Folium library. After extract data from Wikipedia website, we populate data into Pandas data frame then visualize data.

According to Wikipedia, there are 50 districts in Bangkok and the total population is approximately 5.6 million people.

```
#checking the number of districts in Bangkok
wiki_df.shape
```

```
(50, 8)
```

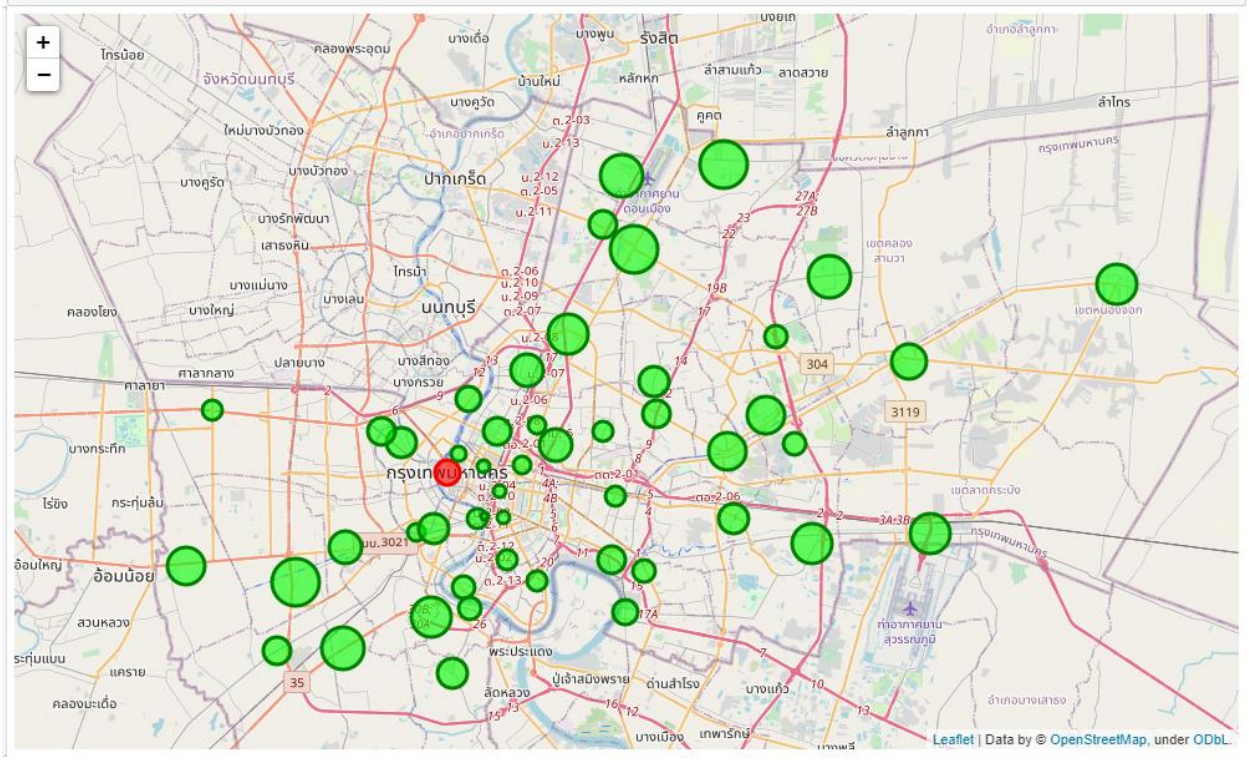
```
#Checking the total popluation
wiki_df['Population'].sum()
```

```
5671070
```

3.2. Visualization

To extract the location of each district, Python Geocoder package is the technique use to plot into the map by using latitude and longitude from Wikipedia.

Red circle represents the central area of Bangkok. Size of green circle represent population in each district.

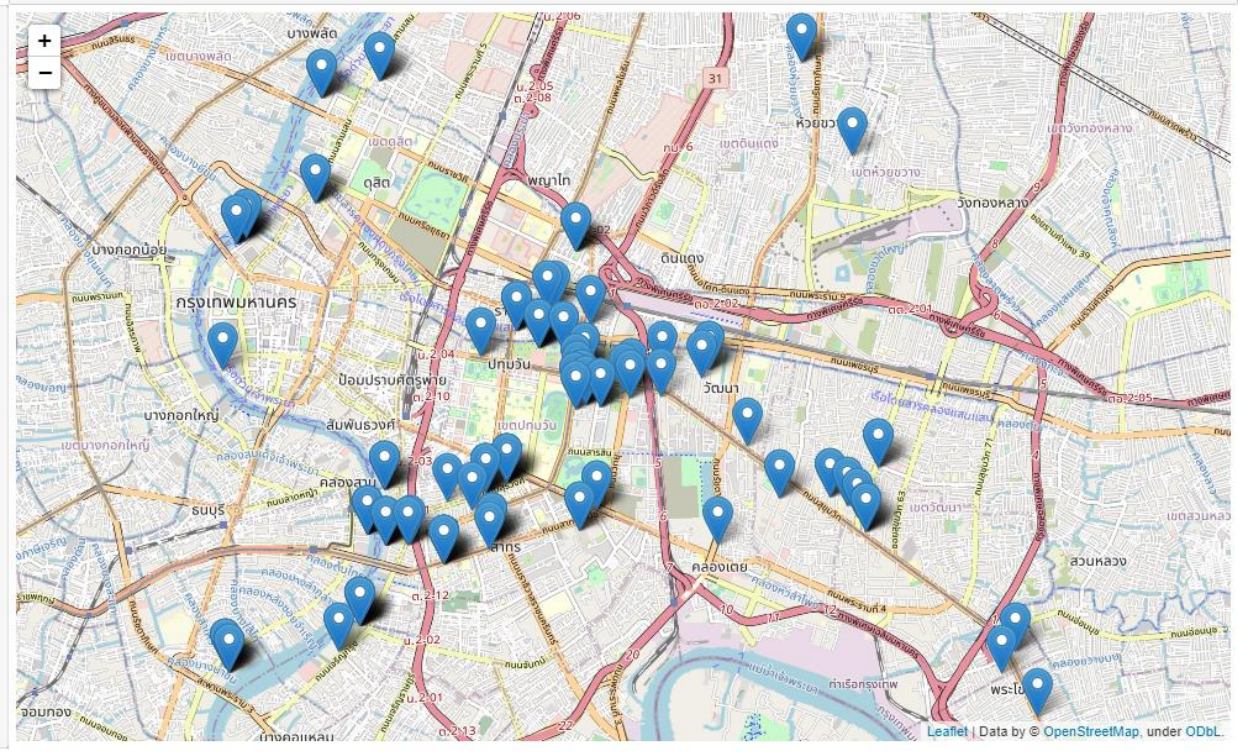


For gathering more information, Foursquare API is used to get the top 100 venues that are within a radius of 5,000 meters. Foursquare will return the venue data in JSON format and will extract the venue name, venue category, venue latitude and longitude.

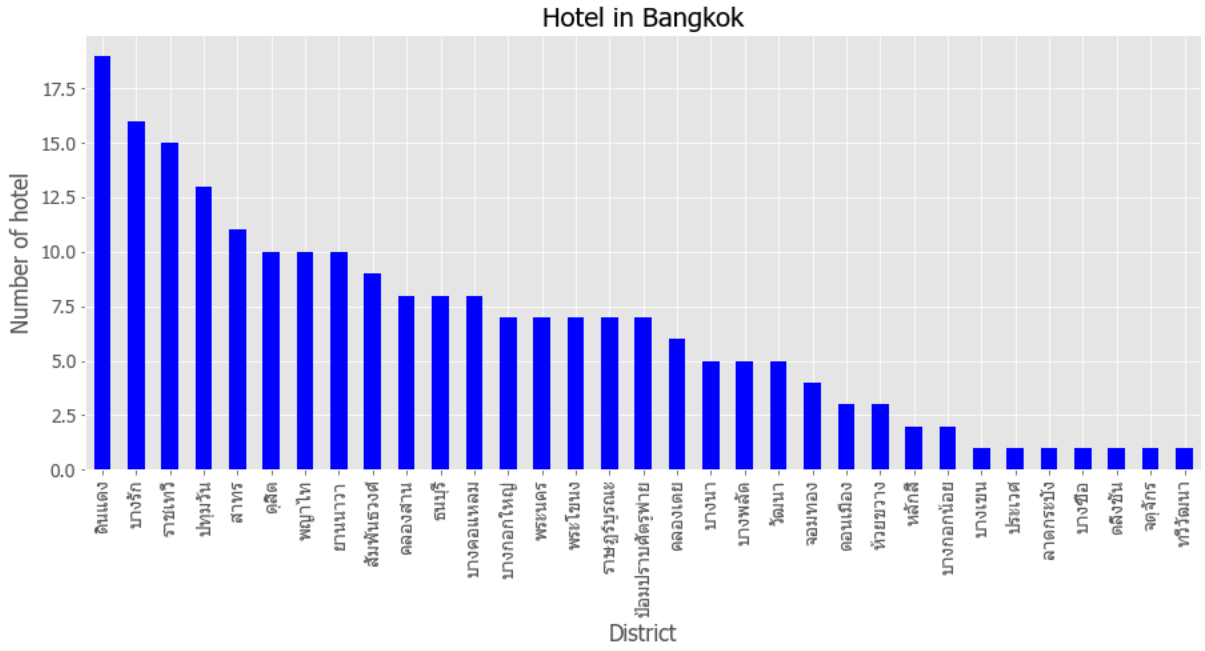
	District	Population	Code	Latitude	Longitude	VenueName	VenueId	VenueLatitude	VenueLongitude	VenueCategory
0	บางบอน	50	105161	13.659200	100.399100	ขาหมูบางหว้า	4e880a81f790e992e01d7284	13.657136	100.395230	Thai Restaurant
1	บางบอน	50	105161	13.659200	100.399100	ข้าวต้มม้วน	4ec7d62f0e6158bafef41324	13.666550	100.412108	Asian Restaurant
2	บางบอน	50	105161	13.659200	100.399100	ร้านต้นไม้ ริมถนนกาญจนาภิเษก	4bf8e392508c0f4796f13e31	13.654098	100.405054	Garden Center
3	บางบอน	50	105161	13.659200	100.399100	KFC	570f0a1bcd109301f16f70fa	13.670449	100.405502	Fast Food Restaurant
4	บางบอน	50	105161	13.659200	100.399100	Burger King (เบอร์เกอร์ คิง)	5884172803e29a6e0f77197e	13.670830	100.405089	Fast Food Restaurant
...	...	...	...	...	...	...	...	...	...	...
4866	ยานนาวา	12	81521	13.696944	100.543056	Pullman Bangkok Hotel G (โรงแรมพูลแมน กรุงเทพฯ...	4b9b1d27f964a520b8f335e3	13.726129	100.525795	Hotel
4867	ยานนาวา	12	81521	13.696944	100.543056	Rengaya (เรนกายา)	4d96ae48af3d236acf1311c7	13.729172	100.533386	BBQ Joint
4868	ยานนาวา	12	81521	13.696944	100.543056	25 Degrees (ทเวนตี้ไฟว์ ดีกรีส์)	4fb34c70e4b0d8fe963a2826	13.725711	100.525800	Burger Joint
4869	ยานนาวา	12	81521	13.696944	100.543056	สวนหรรษา@สวนลุมพินี	50a7247de4b027fd39cdc095	13.730093	100.539104	Park
4870	ยานนาวา	12	81521	13.696944	100.543056	Tealicious Café (แซ่บ เวอร์)	52a578c0498ef669db82cdfb	13.722545	100.516866	Thai Restaurant

4871 rows × 10 columns

Then visualize hotel category of venue in map. Below is sample hotel in each district from Foursquare



The total number of hotels in each district shows as bar chart as below



4. Mythology

4.1. one-hot encoding

One-hot encoding categorize data and represents as numerical value of the dataset

	District	Airport	Airport Lounge	Airport Service	American Restaurant	Art Gallery	Art Museum	Asian Restaurant	Athletics & Sports	BBQ Joint	...	Water Park	Whisky Bar	Wine Bar	Wine Shop	Wings Joint	Women's Store	Yoga Studio	R
0	บางบอน	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
1	บางบอน	0	0	0	0	0	0	1	0	0	...	0	0	0	0	0	0	0	
2	บางบอน	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
3	บางบอน	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
4	บางบอน	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	

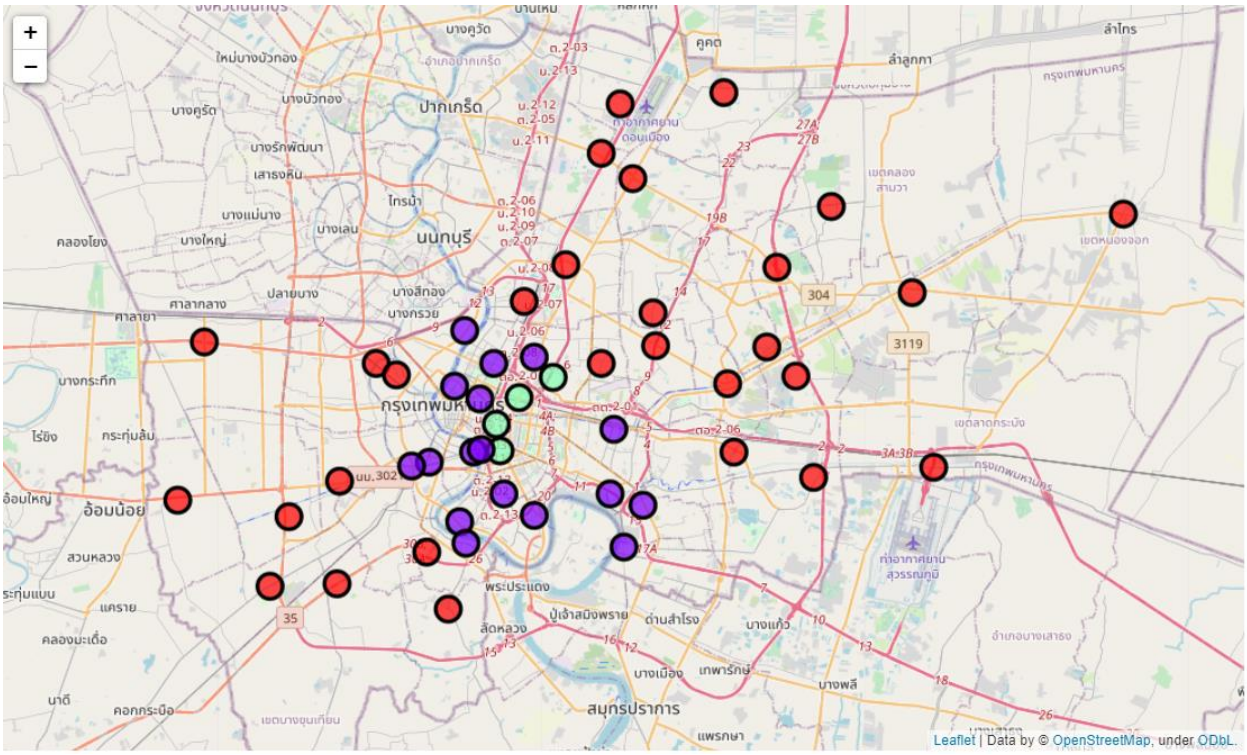
5 rows × 216 columns

4.2. K-Means

K-Means is the technique to use to cluster district into 3 clusters based on their frequency of occurrence for hotel.

	District	Hotel	Cluster	DistrictEng	code	Postcode	Population	No_of_Subdistricts	Latitude	Longitude
0	คลองสาน	0.08	1	Khlong San	18	10600	76446	4	13.730278	100.509722
1	คลองสามวา	0.00	0	Khlong Sam Wa	46	10510	169489	5	13.859722	100.704167
2	คลองเตย	0.06	1	Khlong Toei	33	10110	109041	3	13.708056	100.583889
3	คันนายาว	0.00	0	Khan Na Yao	43	10230	88678	2	13.827100	100.674300
4	จตุจักร	0.01	0	Chatuchak	30	10900	160906	5	13.828611	100.559722

The result allows us to identify which districts have a higher and lower concentration of hotel by using K-Means clustering. This result can help us to plan which district we should acquire land for developing new hotel.



5. Result

We found that hotels are clustered using number of hotels in each district so, it can be classified into high, medium, and low. However, there are other factors that can be used to cluster for example, the number of tourists or number of people traveled around these areas. This information can make clustering more accurate and will generate more information for decision making

	District	Hotel	Cluster	DistrictEng	code	Postcode	Population	No_of_Subdistricts	Latitude	Longitude
0	คลองสาน	0.080000	1	Khlong San	18	10600	76446	4	13.730278	100.509722
1	คลองสามวา	0.000000	0	Khlong Sam Wa	46	10510	169489	5	13.859722	100.704167
2	คลองเตย	0.060000	1	Khlong Toei	33	10110	109041	3	13.708056	100.583889
3	คันนายาว	0.000000	0	Khan Na Yao	43	10230	88678	2	13.827100	100.674300
4	จตุจักร	0.010000	0	Chatuchak	30	10900	160906	5	13.828611	100.559722
5	จอมทอง	0.040000	0	Chom Thong	35	10150	158005	4	13.677222	100.484722
6	ดอนเมือง	0.030000	0	Don Mueang	36	10210	166261	3	13.913611	100.589722
7	ดินแดง	0.190000	2	Din Daeng	26	10400	130220	2	13.769722	100.552778
8	ดุสิต	0.100000	1	Dusit	2	10300	107655	5	13.776944	100.520556
9	ตลิ่งชัน	0.010000	0	Taling Chan	19	10170	106604	6	13.776944	100.456667
10	ทวีวัฒนา	0.010000	0	Thawi Watthana	48	10170	76351	2	13.787800	100.363800
11	ทุ่งครุ	0.000000	0	Thung Khru	49	10140	116473	2	13.647200	100.495800
12	ธนบุรี	0.080000	1	Thon Buri	15	10600	119708	7	13.725000	100.485833
13	บางกอกน้อย	0.020000	0	Bangkok Noi	20	10700	117793	5	13.770867	100.467933
14	บางกอกใหญ่	0.070000	1	Bangkok Yai	16	10600	72321	2	13.722778	100.476389
15	บางกะปิ	0.000000	0	Bang Kapi	6	10240	148465	2	13.765833	100.647778
16	บางขุนเทียน	0.000000	0	Bang Khun Thian	21	10150	165491	2	13.660833	100.435833
17	บางคอแหลม	0.080000	1	Bang Kho Laem	31	10120	94956	3	13.693333	100.502500

6. Conclusion

According to objective of this project to support decision making on planning to launch new hotel in Bangkok by gone through the identifying process of business problem, data acquisition, data extraction and data preparation and clustering data with machine learning by clustering data into 3 groups

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

7. References

- List of districts in Bangkok: https://en.wikipedia.org/wiki/List_of_districts_of_Bangkok
- Foursquare Developers Document: https://developer.foursquare.com/docs