14 Feb 2020
Shudh Datta

BANGKOK
MACHINE LEARNING A CITY

By Shudh Datta

INTRODUCTION

Every city has its own distinct character which is expressed through its neighbourhoods. Some neighbourhoods are somnolent whereas others are bustling. The nature of neighbourhoods are fluid in time where the features of one neighbourhood spills over another.

Thailand’s capital Bangkok is one such city that hosted more than 22 million international overnight travelers [*Mastercard’s Global Destination Cities Index*]. Paris and London followed in second and third with just over 19 million each. To know a city neighbourhood with such a magnitude poses a huge challenge to the travellers

Traditionally travelers relied on the wisdom of travel guidebooks or portals to familiarize them with a city but with the explosion of global cities with new venues and attractions it is a constant challenge to keep the guidance literature updated.

Another way to know about a city is to search the internet for data - sifting through masses of information, pulling together everything relevant, and making it all accessible and useful. But this is a Herculean task.

In this paper we will use a type of unsupervised machine learning model to learn about a city . Rather than the traditional approach in travel guides which first divides a city by defining neighbourhood groups and then looks at the data viz. The features of the area, this model will reverse this approach.

This model will feed features of neighbourhoods to an unsupervised machine learning algorithm and allow it to find and analyze Bangkok based on the features that have formed organically.

DATA

Wikipedia - This will be our first source of geographical data on Bangkok. We will dynamically scrape and know about the districts of Bangkok.

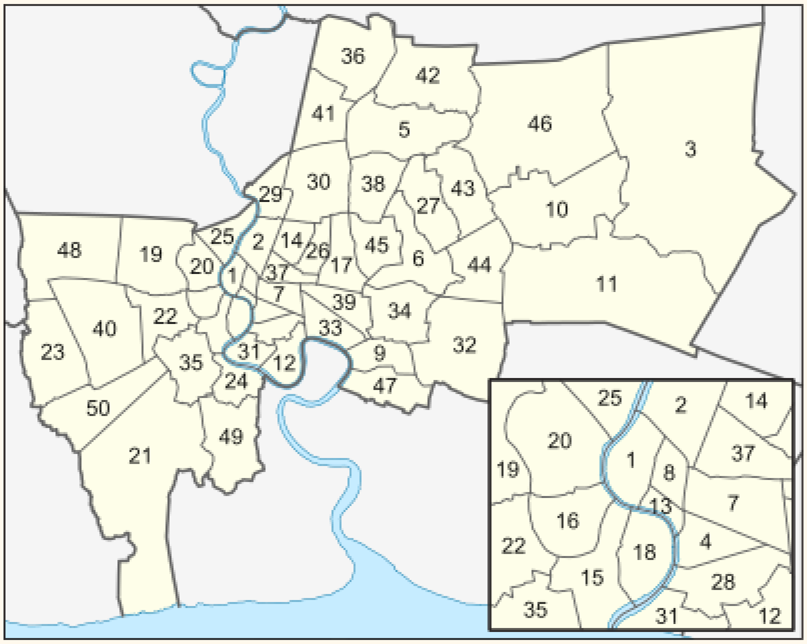
Nominatim for Geocoding Data - Nominatim (from the Latin, 'by name') is a tool to search OSM data by name and address (geocoding) and to generate synthetic addresses of OSM points (reverse geocoding). It can be found at nominatim.openstreetmap.org.

Foursquare Data - We will programmatically access Foursquare dataset in Bangkok. Foursquare is a treasure trove of location information for Bangkok as it remains one of the most popular cities of Foursquare with around 417132 venues with 103000000 check-ins [*4sqstat.com*]

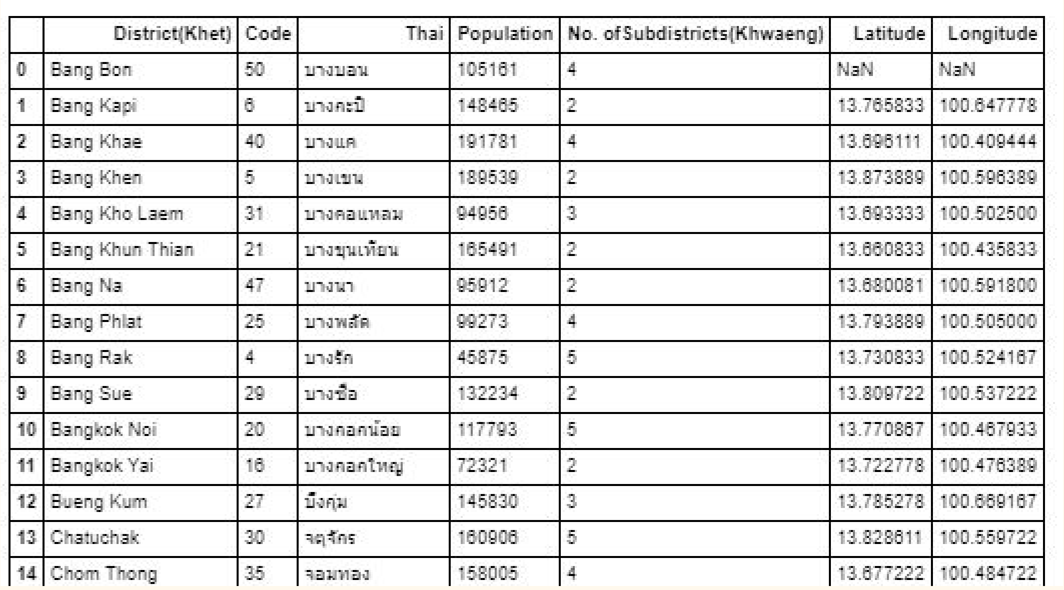
METHODOLOGY

Bangkok is subdivided into 50 districts (khet, เขต, pronounced [kʰèːt], also sometimes wrongly called amphoe as in the other provinces, derived from Pali khetta, cognate to Sanskrit kṣetra), which are further subdivided into 180 subdistricts (khwaeng, แขวง, pronounced [kʰwɛ̌ːŋ]), roughly equivalent to tambon in the other provinces.

For the Bangkok neighborhood data, a Wikipedia page exists - https://en.wikipedia.org/wiki/List_of_districts_of_Bangkok



The following data was extracted from this page

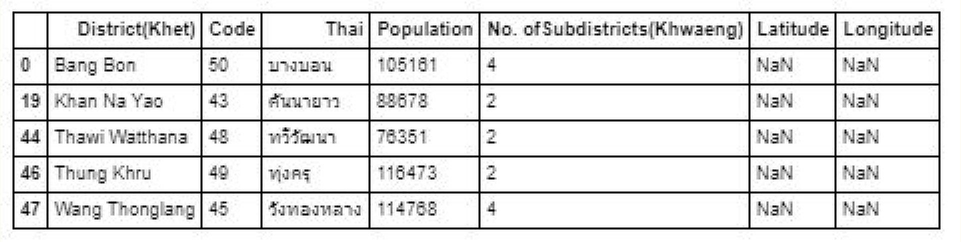


	District(Khet)	Code	Thai	Population	No. of Subdistricts(Khwaeng)	Latitude	Longitude
0	Bang Bon	50	บางบอน	105161	4	NaN	NaN
1	Bang Kapi	6	บางกะปิ	148465	2	13.765833	100.647778
2	Bang Khae	40	บางแค	191781	4	13.696111	100.409444
3	Bang Khen	5	บางเขน	189539	2	13.873889	100.596389
4	Bang Kho Laem	31	บางคอแหลม	94956	3	13.693333	100.502500
5	Bang Khun Thian	21	บางขุนเทียน	165491	2	13.660833	100.435833
6	Bang Na	47	บางนา	95912	2	13.680081	100.591800
7	Bang Phlat	25	บางพลัด	99273	4	13.793889	100.505000
8	Bang Rak	4	บางรัก	45875	5	13.730833	100.524167
9	Bang Sue	29	บางซื่อ	132234	2	13.809722	100.537222
10	Bangkok Noi	20	บางกอกน้อย	117793	5	13.770867	100.467933
11	Bangkok Yai	16	บางกอกใหญ่	72321	2	13.722778	100.476389
12	Bueng Kum	27	บึงกุ่ม	145830	3	13.785278	100.669167
13	Chatuchak	30	จตุจักร	160906	5	13.828611	100.559722
14	Chom Thong	35	จอมทอง	158005	4	13.677222	100.484722

[Partial list of district extracted from Wikipedia]

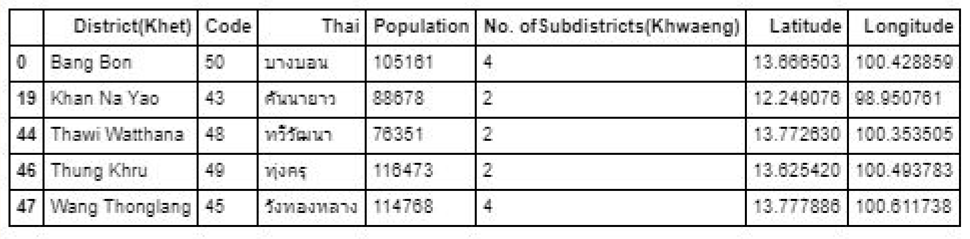
This has the information of districts but the data is not complete as some latitude and longitude values appear as blank.

Next stage we will connect to openstreetmap (nominatim) to geocode the undefined latlong values shown in the table below.



	District(Khet)	Code	Thai	Population	No. of Subdistricts(Khwaeng)	Latitude	Longitude
0	Bang Bon	50	บางบอน	105161	4	NaN	NaN
19	Khan Na Yao	43	คันนายาว	88678	2	NaN	NaN
44	Thawi Watthana	48	ทวีวัฒนา	76351	2	NaN	NaN
46	Thung Khru	49	ทุ่งครุ	116473	2	NaN	NaN
47	Wang Thonglang	45	วังทองหลาง	114768	4	NaN	NaN

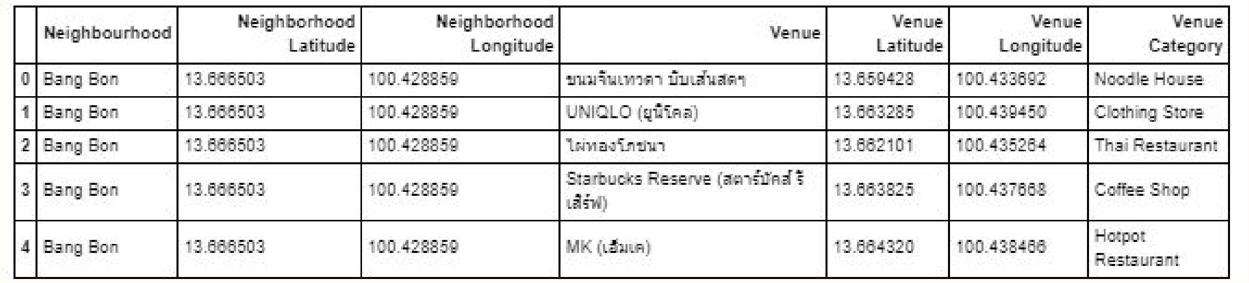
The data returned from Nominatim api for the undefined latlong values are shown below.



	District(Khet)	Code	Thai	Population	No. of Subdistricts(Khwaeng)	Latitude	Longitude
0	Bang Bon	50	บางบอน	105161	4	13.666503	100.428859
19	Khan Na Yao	43	คันนายาว	88678	2	12.249076	98.950761
44	Thawi Watthana	48	ทวีวัฒนา	76351	2	13.772630	100.353505
46	Thung Khru	49	ทุ่งครุ	116473	2	13.625420	100.493783
47	Wang Thonglang	45	วังทองหลาง	114768	4	13.777886	100.611738

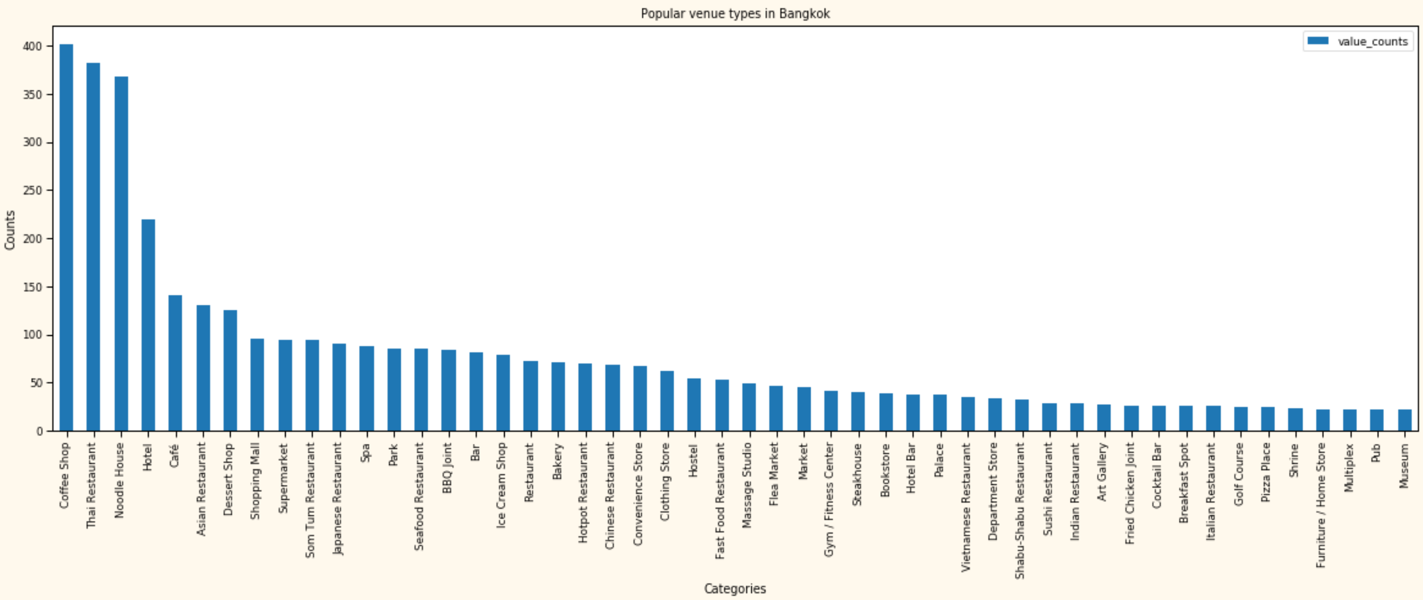
Finally, the data is merged and we have the complete districts of Bangkok data with latitude and longitude.

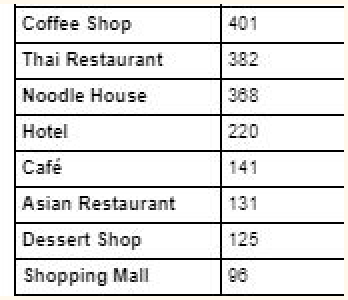
At this stage we are ready to send the geocoded data to Foursquare API to explore the venues in Bangkok. After formatting the JSON returned from the API the data for the venues looks like this.



	Neighbourhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bang Bon	13.666503	100.428859	ขนมจีนเทวดา บิ๊บเส้นสดๆ	13.659428	100.433692	Noodle House
1	Bang Bon	13.666503	100.428859	UNIQLO (ยูนิโคล่)	13.663285	100.439450	Clothing Store
2	Bang Bon	13.666503	100.428859	ไผ่ทองโภชนา	13.662101	100.435264	Thai Restaurant
3	Bang Bon	13.666503	100.428859	Starbucks Reserve (สตาร์บัคส์ รีเสิร์ฟ)	13.663825	100.437668	Coffee Shop
4	Bang Bon	13.666503	100.428859	MK (เอ็มเค)	13.664320	100.438466	Hotpot Restaurant

Though the venue names are in Thai the venue categories are in English for our analysis. We will limit our search result to 100 most relevant results in a district and find out numbers of unique categories & their counts.

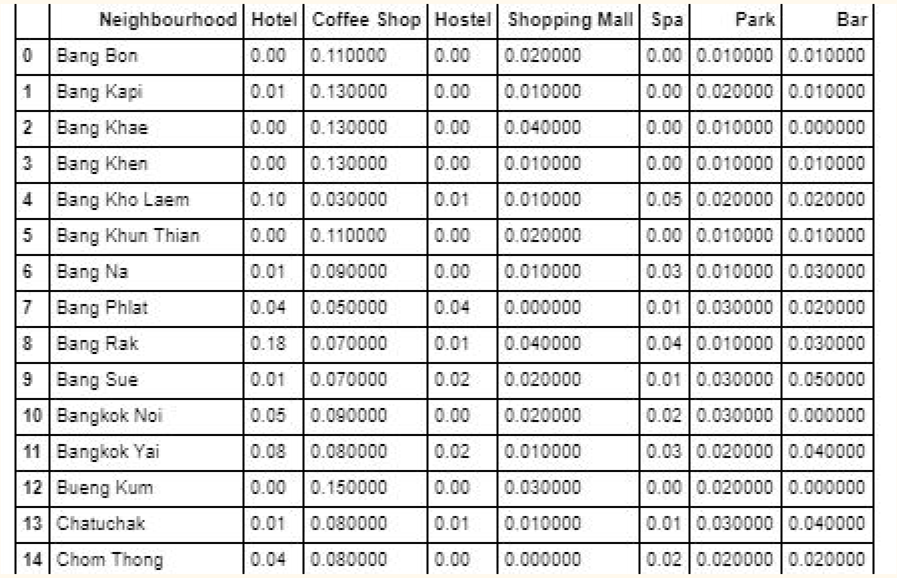




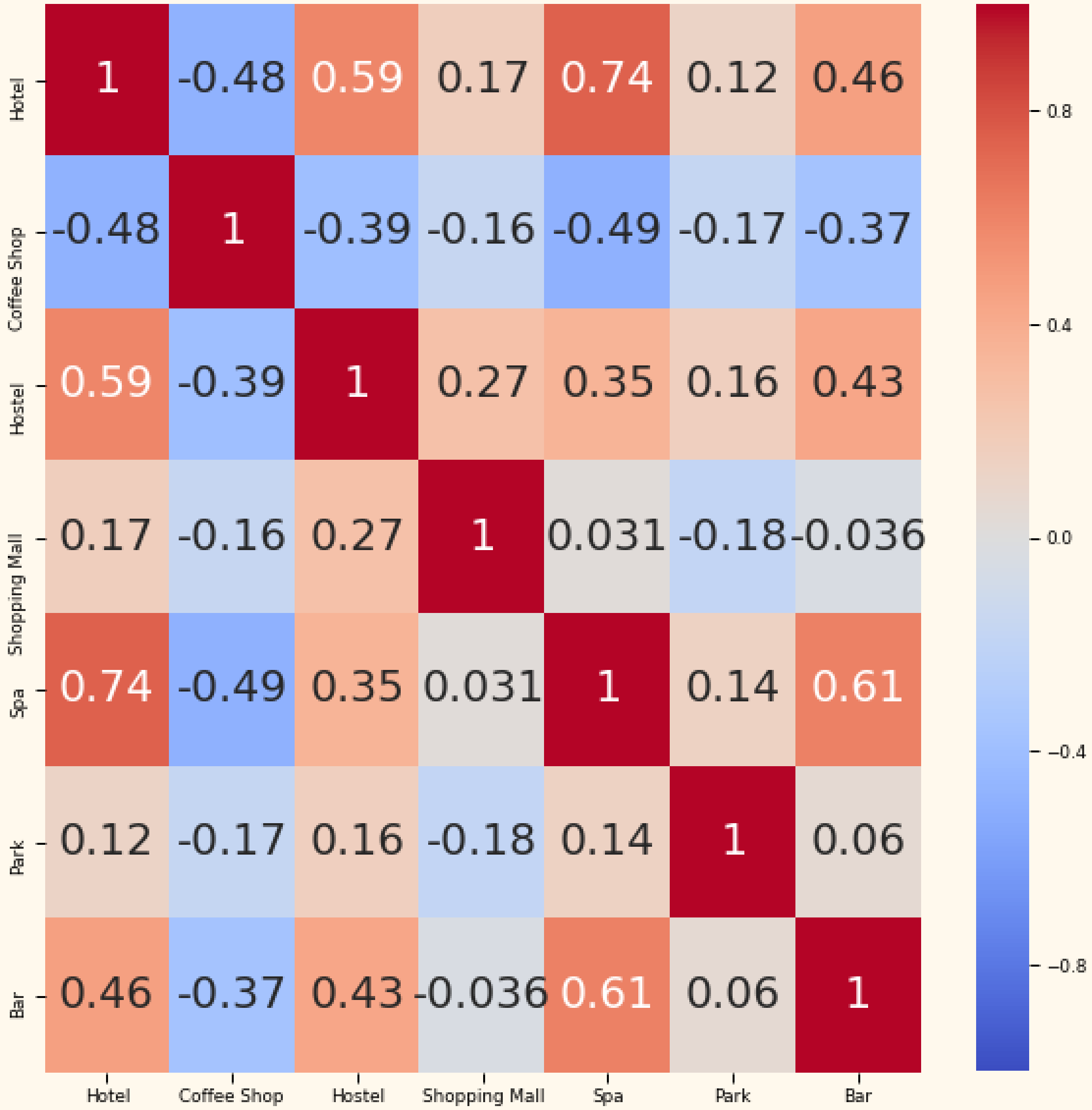
Coffee Shop	401
Thai Restaurant	382
Noodle House	368
Hotel	220
Café	141
Asian Restaurant	131
Dessert Shop	125
Shopping Mall	96

We can see that restaurants, hotels,shopping malls etc form the primary categories in Bangkok.

Now, we will identify some high frequency categories of Bangkok in each neighbourhood and infer if any relation between the categories could be found.



	Neighbourhood	Hotel	Coffee Shop	Hostel	Shopping Mall	Spa	Park	Bar
0	Bang Bon	0.00	0.110000	0.00	0.020000	0.00	0.010000	0.010000
1	Bang Kapi	0.01	0.130000	0.00	0.010000	0.00	0.020000	0.010000
2	Bang Khae	0.00	0.130000	0.00	0.040000	0.00	0.010000	0.000000
3	Bang Khen	0.00	0.130000	0.00	0.010000	0.00	0.010000	0.010000
4	Bang Kho Laem	0.10	0.030000	0.01	0.010000	0.05	0.020000	0.020000
5	Bang Khun Thian	0.00	0.110000	0.00	0.020000	0.00	0.010000	0.010000
6	Bang Na	0.01	0.090000	0.00	0.010000	0.03	0.010000	0.030000
7	Bang Phlat	0.04	0.050000	0.04	0.000000	0.01	0.030000	0.020000
8	Bang Rak	0.18	0.070000	0.01	0.040000	0.04	0.010000	0.030000
9	Bang Sue	0.01	0.070000	0.02	0.020000	0.01	0.030000	0.050000
10	Bangkok Noi	0.05	0.090000	0.00	0.020000	0.02	0.030000	0.000000
11	Bangkok Yai	0.08	0.080000	0.02	0.010000	0.03	0.020000	0.040000
12	Bueng Kum	0.00	0.150000	0.00	0.030000	0.00	0.020000	0.000000
13	Chatuchak	0.01	0.080000	0.01	0.010000	0.01	0.030000	0.040000
14	Chom Thong	0.04	0.080000	0.00	0.000000	0.02	0.020000	0.020000



	Hotel	Coffee Shop	Hostel	Shopping Mall	Spa	Park	Bar
Hotel	1	-0.48	0.59	0.17	0.74	0.12	0.46
Coffee Shop	-0.48	1	-0.39	-0.16	-0.49	-0.17	-0.37
Hostel	0.59	-0.39	1	0.27	0.35	0.16	0.43
Shopping Mall	0.17	-0.16	0.27	1	0.031	-0.18	-0.036
Spa	0.74	-0.49	0.35	0.031	1	0.14	0.61
Park	0.12	-0.17	0.16	-0.18	0.14	1	0.06
Bar	0.46	-0.37	0.43	-0.036	0.61	0.06	1

The correlation plot of the high frequency categories are shown above. From the above correlation heatmap we can see that Hotel and Spa are strongly correlated as the linear relationship is quite strong (~ 0.74). Also, the hotels and hostels could be found together in certain neighbourhoods. However, it does look like you might have to put some effort to get good coffee in those neighbourhoods (negatively correlated).Also, note that Spa and Bar are positively correlated though not as strong as the relationship between Hotels and Spas that we found earlier in certain neighbourhoods.

So, one can expect to discover the most number of spas and bars if they stay in the neighbourhoods that have the maximum frequency of hostels and hotels.

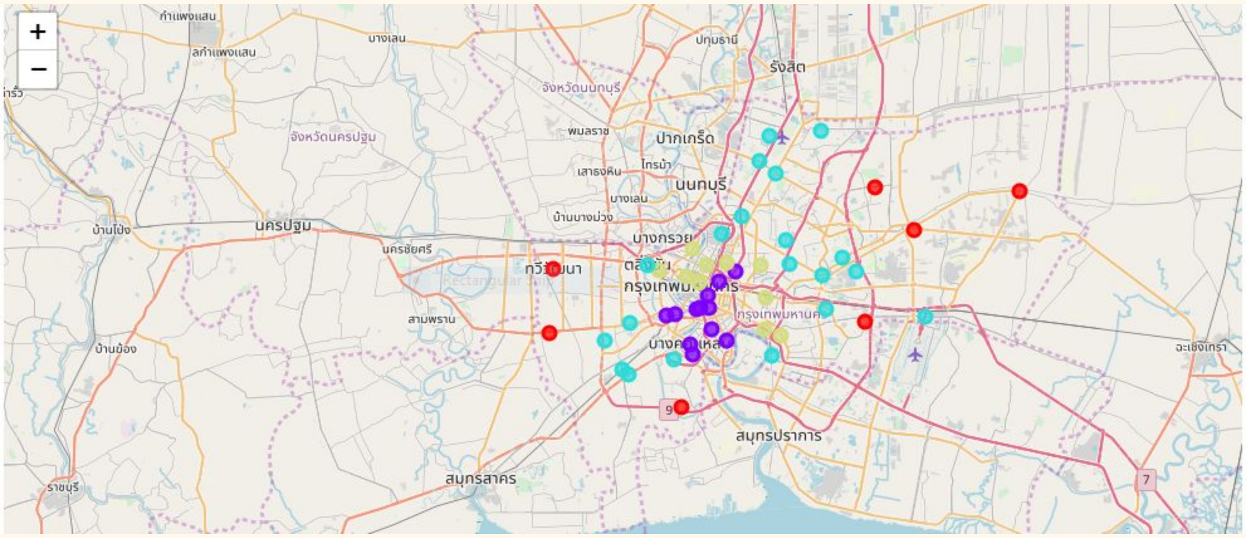
Unsupervised Machine Learning

As discussed earlier the selected algorithm should be able to identify subgroups (groups of neighbourhoods in Bangkok) in the data such that data points in the same subgroup (cluster) are very similar. So, the algorithm should be able to identify homogeneous neighbourhoods within the data such that data points in each neighbourhood are as similar as possible based on a similarity measure such as euclidean-based distance.

We selected K-means clustering algorithm for this purpose. K-means clustering is a type of unsupervised learning, which is used when on unlabeled data (i.e., data without defined categories or groups, in our case neighbourhoods). The goal of this algorithm is to find groups in the data.This is unsupervised learning because we don't have the labels or ground truth of neighbourhood clusters of Bangkok.

RESULT

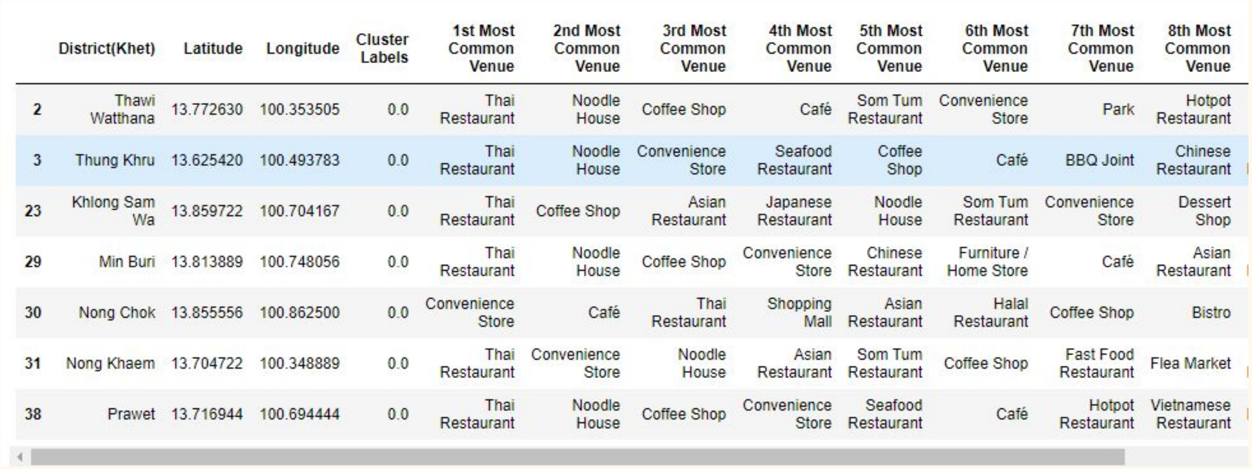
After running the K-means algorithm we plot the discovered neighbourhood clusters with the same colour dots on a folium map shown below.



The red dots in the above map belong to the cluster 0, which is identified by the algorithm. The districts of Bangkok that belongs to this cluster are

District(Khet)

1. Thawi Watthana
2. Thung Khru
3. Khlong Sam Wa
4. Min Buri
5. Nong Chok
6. Nong Khaem
7. Prawet



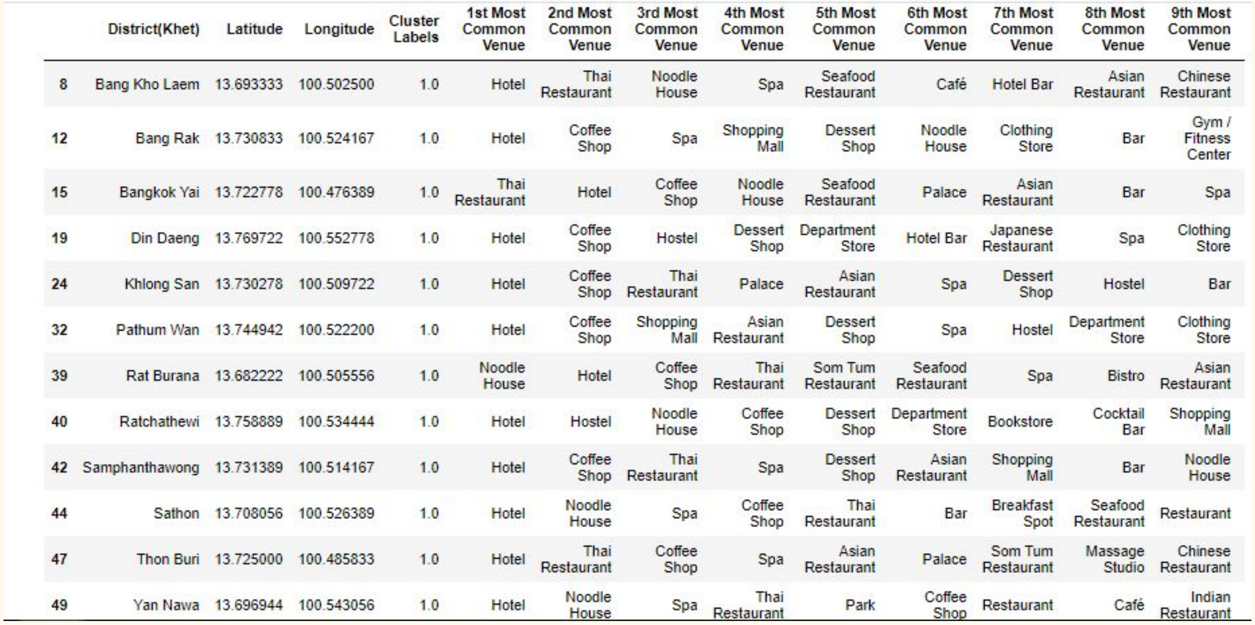
	District(Khet)	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
2	Thawi Watthana	13.772630	100.353505	0.0	Thai Restaurant	Noodle House	Coffee Shop	Café	Som Tum Restaurant	Convenience Store	Park	Hotpot Restaurant
3	Thung Khru	13.625420	100.493783	0.0	Thai Restaurant	Noodle House	Convenience Store	Seafood Restaurant	Coffee Shop	Café	BBQ Joint	Chinese Restaurant
23	Khlong Sam Wa	13.859722	100.704167	0.0	Thai Restaurant	Coffee Shop	Asian Restaurant	Japanese Restaurant	Noodle House	Som Tum Restaurant	Convenience Store	Dessert Shop
29	Min Buri	13.813889	100.748056	0.0	Thai Restaurant	Noodle House	Coffee Shop	Convenience Store	Chinese Restaurant	Furniture / Home Store	Café	Asian Restaurant
30	Nong Chok	13.855556	100.862500	0.0	Convenience Store	Café	Thai Restaurant	Shopping Mall	Asian Restaurant	Halal Restaurant	Coffee Shop	Bistro
31	Nong Khaem	13.704722	100.348889	0.0	Thai Restaurant	Convenience Store	Noodle House	Asian Restaurant	Som Tum Restaurant	Coffee Shop	Fast Food Restaurant	Flea Market
38	Prawet	13.716944	100.694444	0.0	Thai Restaurant	Noodle House	Coffee Shop	Convenience Store	Seafood Restaurant	Café	Hotpot Restaurant	Vietnamese Restaurant

The purple dots plotted on the map at the heart of Bangkok belongs to Cluster 1. Following districts were identified in this cluster

District(Khet)

1. Bang Kho Laem
2. Bang Rak
3. Bangkok Yai
4. Din Daeng
5. Khlong San
6. Pathum Wan
7. Rat Burana
8. Ratchathewi
9. Samphanthawong
10. Sathon
11. Thon Buri

12. Yan Nawa



	District(Khet)	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
8	Bang Kho Laem	13.693333	100.502500	1.0	Hotel	Thai Restaurant	Noodle House	Spa	Seafood Restaurant	Café	Hotel Bar	Asian Restaurant	Chinese Restaurant
12	Bang Rak	13.730833	100.524167	1.0	Hotel	Coffee Shop	Spa	Shopping Mall	Dessert Shop	Noodle House	Clothing Store	Bar	Gym / Fitness Center
15	Bangkok Yai	13.722778	100.476389	1.0	Thai Restaurant	Hotel	Coffee Shop	Noodle House	Seafood Restaurant	Palace	Asian Restaurant	Bar	Spa
19	Din Daeng	13.769722	100.552778	1.0	Hotel	Coffee Shop	Hostel	Dessert Shop	Department Store	Hotel Bar	Japanese Restaurant	Spa	Clothing Store
24	Khlong San	13.730278	100.509722	1.0	Hotel	Coffee Shop	Thai Restaurant	Palace	Asian Restaurant	Spa	Dessert Shop	Hostel	Bar
32	Pathum Wan	13.744942	100.522200	1.0	Hotel	Coffee Shop	Shopping Mall	Asian Restaurant	Dessert Shop	Spa	Hostel	Department Store	Clothing Store
39	Rat Burana	13.682222	100.505556	1.0	Noodle House	Hotel	Coffee Shop	Thai Restaurant	Som Tum Restaurant	Seafood Restaurant	Spa	Bistro	Asian Restaurant
40	Ratchathewi	13.758889	100.534444	1.0	Hotel	Hostel	Noodle House	Coffee Shop	Dessert Shop	Department Store	Bookstore	Cocktail Bar	Shopping Mall
42	Samphanthawong	13.731389	100.514167	1.0	Hotel	Coffee Shop	Thai Restaurant	Spa	Dessert Shop	Asian Restaurant	Shopping Mall	Bar	Noodle House
44	Sathon	13.708056	100.526389	1.0	Hotel	Noodle House	Spa	Coffee Shop	Thai Restaurant	Bar	Breakfast Spot	Seafood Restaurant	Restaurant
47	Thon Buri	13.725000	100.485833	1.0	Hotel	Thai Restaurant	Coffee Shop	Spa	Asian Restaurant	Palace	Som Tum Restaurant	Massage Studio	Chinese Restaurant
49	Yan Nawa	13.696944	100.543056	1.0	Hotel	Noodle House	Spa	Thai Restaurant	Park	Coffee Shop	Restaurant	Café	Indian Restaurant

The sky-blue colored dots on the cluster map belongs to the cluster 2. Following neighbourhoods belong to this cluster

District(Khet)

1. Bang Bon
2. Wang Thonglang
3. Bang Kapi
4. Bang Khae
5. Bang Khen
6. Bang Khun Thian
7. Bang Na
8. Bang Sue
9. Bueng Kum
10. Chatuchak
11. Chom Thong
12. Don Mueang
13. Lak Si
14. Lat Krabang
15. Lat Phrao
16. Phasi Charoen

17. Sai Mai
18. Saphan Sung
19. Suan Luang
20. Taling Chan

Finally, the olive green dots represent the cluster 3. Following districts belong to this cluster

District(Khet)

1. Bang Phlat
2. Bangkok Noi
3. Dusit
4. Huai Khwang
5. Khlong Toei
6. Phaya Thai
7. Phra Khanong
8. Phra Nakhon
9. Pom Prap Sattru Phai
10. Watthana

DISCUSSION

We observe the following from the results of our clustering analysis

Cluster 0 - The outskirts cluster

The identified cluster is far away from the city center. There are no hotels here. Thai restaurants and noodle houses cater to the local palate. There are many department stores in this cluster.

Cluster 1 - The Backpackers Cluster

This cluster is dotted with hostels and hotels also noodle houses, thai restaurants jostle for space with spas. It appears to be a cluster where one should visit atleast once even if they are not staying in the neighbourhood. Probably one of the central tourist destinations. Due to the presence of hostels this could be ideal for young backpackers.

Cluster 2 - The Resident Cluster

Lack of hotels/hostels hustle and bustle are the characteristics of this cluster. If you are an expat and plan to work in Bangkok you should consider renting a place in this cluster. Also, from the visualization map one can stay nearer to the airport and optimal distance away from the city center here. When you don't want to cook the friendly neighbourhood noodle houses will always welcome you.

Cluster 3 - The Affluent Traveler Cluster

This cluster is predominantly "The HOTEL" cluster which is the most common venue followed by coffee shops and shopping malls. Someone staying in this cluster could leisurely stroll into the air conditioned shopping malls . Shopping gaps could be filled by plush coffee or dessert.

CONCLUSION

We live in the ever increasing age of data and cheap computing power. It is now possible to generate various insights from the data. The model discussed in this paper is generic enough to be applied to the cities of your choice.