Capstone Project – The Best of Tokyo

Problem and Background

Tokyo, the capital of japan, is one of the most desirable places people would love to visit. A trip without visiting this legendary capital city is incomplete. The cherry blossom festival and the fall foliage attract people around all around the world.

Besides gazing out from Tokyo sky tree, and wrestling with sumo culture in Ryōgoku, Tokyo is also a "battleground" for good food. Ramen is a must-have in Tokyo, though you may have tried tons of ramen, Tokyo represents the cutting edge of ramen. Edomae-zushi with your personalized toppings, monjayaki - both delicious and healthy, and unaju - fresh eel served over juicy rice.

In this capstone project, we are going to explore Tokyo, and have an insight of districts and restaurants across Tokyo cities, which will help visitors quickly locate and choose the place to visit and enjoy a nice meal.

Dataset

We will use Foursquare location information to retrieve the raw data, and perform analysis by using clustering methods like one-hot encoding and k-means.

The dataset is from https://en.wikipedia.org/wiki/Special_wards_of_Tokyo

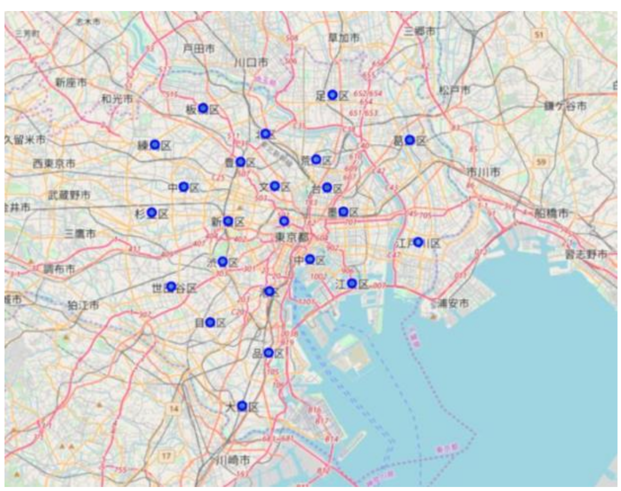
We will use the list of special wards table from the Wikipedia. It contains the basic information of the district name, population, density, area size, which would be the base line of our analysis.

We also need the restaurants information, we used the one from Foursquare APIs. We can use the location information to filter the restaurants information for our task.

	No.	Name	Population	Population(as of October 2016	Density	Area	Major districts
0	01	Chiyoda	千代田区	59441	5100	11.66	Nagatachō, Kasumigaseki, Ōtemachi, Marunouchi,...
1	02	Chūō	中央区	147620	14460	10.21	Nihonbashi, Kayabachō, Ginza, Tsukiji, Hatchōb...
2	03	Minato	港区	248071	12180	20.37	Odaiba, Shinbashi, Hamamatsuchō, Mita, Roppong...
3	04	Shinjuku	新宿区	339211	18620	18.22	Shinjuku, Takadanobaba, Ōkubo, Kagurazaka, Ich...
4	05	Bunkyō	文京区	223389	19790	11.29	Hongō, Yayoi, Hakusan
5	06	Taitō	台東区	200486	19830	10.11	Ueno, Asakusa
6	07	Sumida	墨田区	260358	18910	13.77	Kinshichō, Morishita, Ryōgoku
7	08	Kōtō	江東区	502579	12510	40.16	Kiba, Ariake, Kameido, Tōyōchō, Monzennakachō,...
8	09	Shinagawa	品川区	392492	17180	22.84	Shinagawa, Gotanda, Ōsaki, Hatanodai, Ōimachi,...
9	10	Meguro	目黒区	280283	19110	14.67	Meguro, Nakameguro, Jiyugaoka, Komaba, Aobadai
10	11	Ōta	大田区	722608	11910	60.66	Ōmori, Kamata, Haneda, Den-en-chōfu
11	12	Setagaya	世田谷区	910868	15690	58.05	Shimokitazawa, Kinuta, Karasuyama, Tamagawa
12	13	Shibuya	渋谷区	227850	15080	15.11	Shibuya, Ebisu, Harajuku, Daikanyama, Hiroo
13	14	Nakano	中野区	332902	21350	15.59	Nakano
14	15	Suginami	杉並区	570483	16750	34.06	Kōenji, Asagaya, Ogikubo
15	16	Toshima	豊島区	294673	22650	13.01	Ikebukuro, Komagome, Senkawa, Sugamo
16	17	Kita	北区	345063	16740	20.61	Akabane, Ōji, Tabata
17	18	Arakawa	荒川区	213648	21030	10.16	Arakawa, Machiya, Nippori, Minamisenju
18	19	Itabashi	板橋区	569225	17670	32.22	Itabashi, Takashimadaira
19	20	Nerima	練馬区	726748	15120	48.08	Nerima, Ōizumi, Hikarigaoka
20	21	Adachi	足立区	674067	12660	53.25	Ayase, Kitasenju, Takenotsuka
21	22	Katsushika	葛飾区	447140	12850	34.80	Tateishi, Aoto, Kameari, Shibamata

Methodology

We plan to perform statistical methods like one-hot encoding and k-means on the foursquare API, which can provide information about the popular restaurants in each distributes of Tokyo. We will further plot the information on the Folium leaflet map as a good visualization.



Use Chuo (a Tokyo district) as an example, we first retrieve its location information from Nominatim.

In [216]:

```
neighborhood_latitude = df.loc[1, 'Latitude'] # neighborhood latitude value
neighborhood_longitude = df.loc[1, 'Longitude'] # neighborhood longitude value

neighborhood_name = df.loc[1, 'Name'] # neighborhood name

print('Latitude and longitude values of {} are {}, {}.'.format(neighborhood_name,
                                                               neighborhood_latitude,
                                                               neighborhood_longitude))
```

```
Latitude and longitude values of Chūō are 35.666255, 139.775565.
```

In [217]:

```
LIMIT = 100

radius = 500

url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}
    CLIENT_ID,
    CLIENT_SECRET,
    VERSION,
    neighborhood_latitude,
    neighborhood_longitude,
    radius,
    LIMIT)
url # display URL
```

After clustering, we can see clearly the categories of the places of interest and restaurants and also the categories.

	name	categories	lat	lng
0	築地 すし大 別館	Sushi Restaurant	35.664831	139.774704
1	Tsukiji JISAKU (つきじ治作)	Kaiseki Restaurant	35.665236	139.777358
2	Karamenya En (からめんや えん)	Noodle House	35.664916	139.774722
3	Tsukiji Tutto Bene! (築地トゥットベーネ)	Italian Restaurant	35.666144	139.774815
4	Sushi Keita (鮨 桂太)	Sushi Restaurant	35.665428	139.773654

In [221]:

```
print (nearby_venues['categories'].value_counts()[0:10])
```

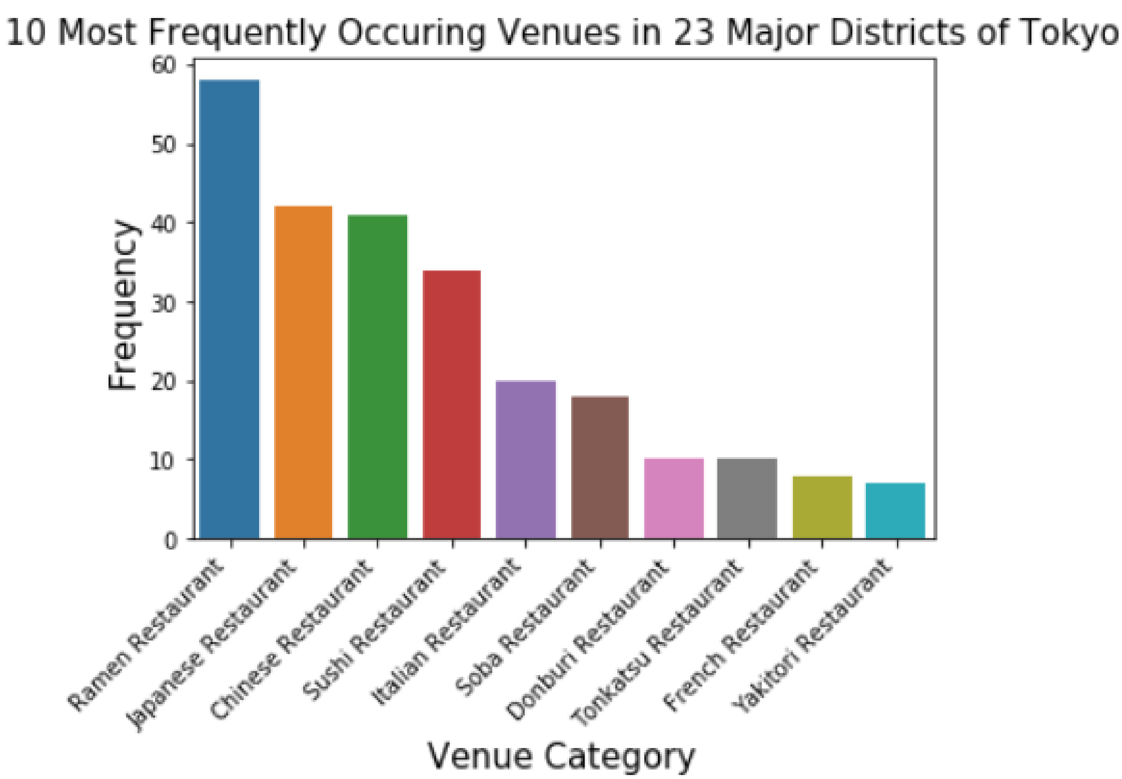
```
Sushi Restaurant       20
Japanese Restaurant    11
Park                    4
Convenience Store       4
Coffee Shop             4
Soba Restaurant         3
Italian Restaurant      3
Bakery                  3
Chinese Restaurant      2
Ramen Restaurant        2
Name: categories, dtype: int64
```

As we looking into the whole Tokyo city, the distribution of restaurants would be like

```python
# Create a Data-Frame out of it to Concentrate Only on Restaurants

Tokyo_Venues_only_restaurant = Tokyo_venues[Tokyo_venues['Venue Category'].str.contains('Restaurant')].reset_index(
Tokyo_Venues_only_restaurant.index = np.arange(1, len(Tokyo_Venues_only_restaurant )+1)
print (Tokyo_Venues_only_restaurant['Venue Category'].value_counts())
```

```
Ramen Restaurant               58
Japanese Restaurant            42
Chinese Restaurant             41
Sushi Restaurant               34
Italian Restaurant             20
Soba Restaurant                18
Donburi Restaurant             10
Tonkatsu Restaurant            10
French Restaurant               8
Yakitori Restaurant             7
Indian Restaurant               7
Yoshoku Restaurant              6
Unagi Restaurant                6
Dumpling Restaurant             5
Seafood Restaurant              5
Japanese Family Restaurant      5
Thai Restaurant                 5
Teishoku Restaurant             4
Japanese Curry Restaurant       4
Restaurant                      3
Tempura Restaurant              3
Kushikatsu Restaurant           3
Udon Restaurant                 3
South Indian Restaurant         2
Nabe Restaurant                 2
Vietnamese Restaurant           2
Kaiseki Restaurant              2
Korean Restaurant               2
Szechuan Restaurant             2
```

The visualization of the top 10 restaurants would be like



And top restaurants list would be

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
1	Chiyoda	35.69381	139.753216	Jimbocho Kurosu (神保町 黒須)	35.695539	139.754851	Ramen Restaurant
2	Chiyoda	35.69381	139.753216	Kanda Tendonya (神田天丼家)	35.695765	139.754682	Tempura Restaurant
3	Chiyoda	35.69381	139.753216	Bondy (欧風カレー ボンディ)	35.695544	139.757356	Japanese Curry Restaurant
4	Chiyoda	35.69381	139.753216	Sushi Masa (九段下 寿司政)	35.695234	139.752227	Sushi Restaurant
5	Chiyoda	35.69381	139.753216	咸亨酒店	35.696010	139.756730	Chinese Restaurant
6	Chiyoda	35.69381	139.753216	Genrai Shuka (源来酒家)	35.695671	139.754409	Chinese Restaurant
7	Chiyoda	35.69381	139.753216	Yojinbo (用心棒)	35.696409	139.756696	Ramen Restaurant
8	Chiyoda	35.69381	139.753216	Fukumen Tomo (覆麺 智)	35.696403	139.757070	Ramen Restaurant
9	Chiyoda	35.69381	139.753216	Mueang Thai Nabe (ムアン・タイ・なべ)	35.695344	139.757728	Thai Restaurant
10	Chiyoda	35.69381	139.753216	たいよう軒	35.696454	139.754809	Chinese Restaurant

Results

We want to have more insights about each restaurant and the popularity within different districts, so we started with one-hot encoding.

one-hot encoding

we use neighborhood column to count the mean, and also use the frequency information:

	Neighborhood	African Restaurant	Asian Restaurant	Brazilian Restaurant	Chinese Restaurant	Donburi Restaurant	Dongbei Restaurant	Dumpling Restaurant	Fast Food Restaurant	French Restaurant	German Restaurant	Hotpot Restaurant	Ind Restaur
0	Adachi	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
1	Arakawa	0.000000	0.142857	0.000000	0.285714	0.142857	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
2	Bunkyō	0.000000	0.000000	0.000000	0.250000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
3	Chiyoda	0.000000	0.000000	0.000000	0.189189	0.000000	0.00	0.000000	0.000000	0.081081	0.000000	0.000000	0.081
4	Chūō	0.000000	0.000000	0.000000	0.035088	0.035088	0.00	0.000000	0.000000	0.017544	0.017544	0.000000	0.017
5	Edogawa	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.250
6	Itabashi	0.000000	0.000000	0.000000	0.500000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
7	Katsushika	0.000000	0.000000	0.000000	0.200000	0.400000	0.00	0.200000	0.000000	0.000000	0.000000	0.000000	0.000
8	Kita	0.000000	0.000000	0.000000	0.071429	0.071429	0.00	0.071429	0.000000	0.000000	0.000000	0.000000	0.000
9	Kōtō	0.000000	0.000000	0.000000	0.666667	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.333
10	Meguro	0.000000	0.000000	0.000000	0.333333	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
11	Minato	0.000000	0.000000	0.000000	0.181818	0.000000	0.00	0.000000	0.000000	0.090909	0.000000	0.000000	0.090
12	Nakano	0.000000	0.000000	0.000000	0.111111	0.111111	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
13	Nerima	0.000000	0.000000	0.000000	1.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
14	Setagaya	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.142857	0.000000	0.000000	0.000000	0.000
15	Shibuya	0.047619	0.000000	0.047619	0.142857	0.047619	0.00	0.000000	0.000000	0.095238	0.000000	0.000000	0.000
16	Shinagawa	0.000000	0.000000	0.000000	0.000000	0.166667	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
17	Shinjuku	0.000000	0.000000	0.030303	0.030303	0.000000	0.00	0.000000	0.000000	0.030303	0.000000	0.030303	0.000
18	Suginami	0.000000	0.000000	0.000000	0.090909	0.000000	0.00	0.090909	0.000000	0.000000	0.000000	0.000000	0.000
19	Sumida	0.000000	0.000000	0.000000	0.375000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
20	Taitō	0.000000	0.000000	0.000000	0.120000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
21	Toshima	0.000000	0.000000	0.000000	0.100000	0.050000	0.05	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
22	Ōta	0.000000	0.000000	0.000000	0.111111	0.000000	0.00	0.044444	0.000000	0.000000	0.000000	0.000000	0.000

Each district with its most popular venues:

```
Adachi
                           venue  freq
0            Japanese Restaurant  0.67
1     Japanese Family Restaurant  0.33
2             African Restaurant  0.00
3                Nabe Restaurant  0.00
4                     Restaurant  0.00


Arakawa
                   venue  freq
0     Chinese Restaurant  0.29
1       Ramen Restaurant  0.29
2     Italian Restaurant  0.14
3       Asian Restaurant  0.14
4     Donburi Restaurant  0.14
```

K-means clustering

We select number of clusters equal to 6

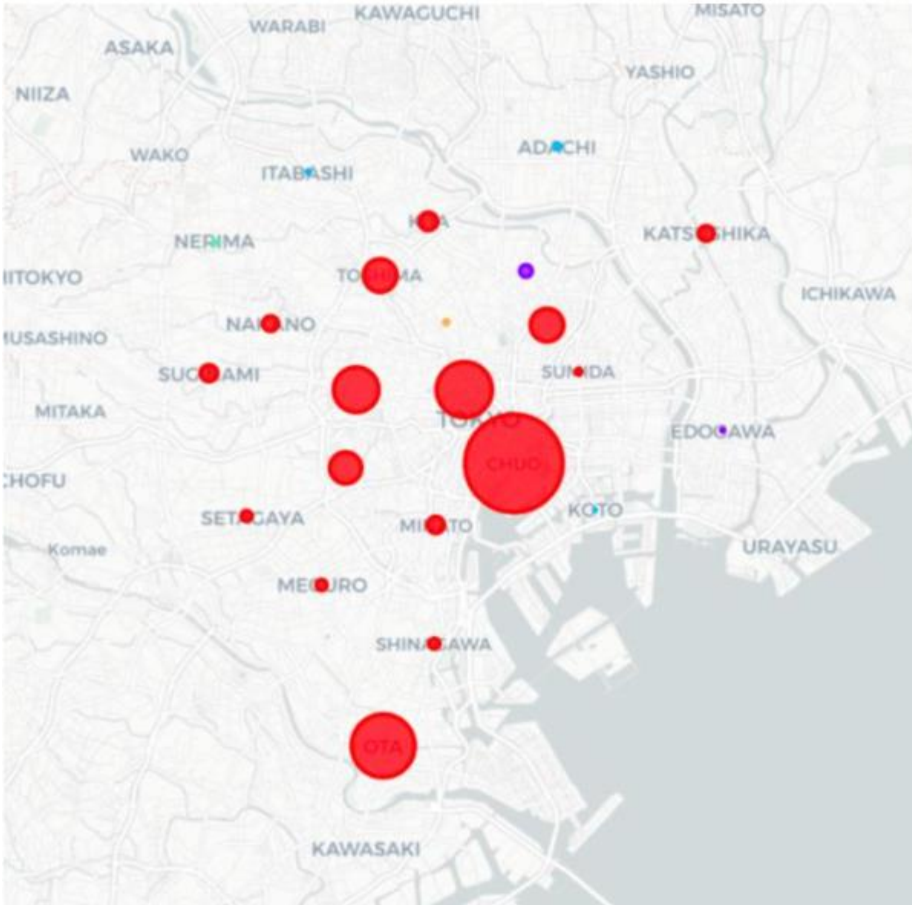
```
# set number of clusters to 6
kclusters = 6

Tokyo_grouped_clustering = Tokyo_grouped.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(Tokyo_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

```
array([2, 3, 4, 0, 0, 0, 5, 3, 3, 1])
```

```
# add clustering labels

neighborhoods_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)

tokyo_merged = df

tokyo_merged.rename(columns={'Name':'Neighborhood'}, inplace=True)

# merge toronto_grouped with toronto_data to add latitude/longitude for each neighborhood
tokyo_merged = tokyo_merged.join(neighborhoods_venues_sorted.set_index('Neighborhood'), on='Neighborhood')

tokyo_merged.head() # check the last columns!
```

	No.	Neighborhood	Population	Population (as of October 2016	Density	Area	Major districts	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
0	01	Chiyoda	千代田区	59441	5100	11.66	Nagatachō, Kasumigaseki, Ōtemachi, Marunouchi,...	35.693810	139.753216	0	Chinese Restaurant	Ramen Restaurant	French Restaurant	Japanese Curry Restaurant
1	02	Chūō	中央区	147620	14460	10.21	Nihonbashi, Kayabachō, Ginza, Tsukiji, Hatchōb...	35.666255	139.775565	0	Sushi Restaurant	Japanese Restaurant	Soba Restaurant	Italian Restaurant
2	03	Minato	港区	248071	12180	20.37	Odaiba, Shinbashi, Hamamatsuchō, Mita, Roppong...	35.643227	139.740055	0	Chinese Restaurant	Soba Restaurant	Indian Restaurant	French Restaurant
3	04	Shinjuku	新宿区	339211	18620	18.22	Shinjuku, Takadanobaba, Ōkubo, Kagurazaka, Ich...	35.693763	139.703632	0	Ramen Restaurant	Japanese Restaurant	Yakitori Restaurant	Tonkatsu Restaurant
4	05	Bunkyō	文京区	223389	19790	11.29	Hongō, Yayoi, Hakusan	35.718810	139.744732	4	Chinese Restaurant	Japanese Restaurant	Szechuan Restaurant	Italian Restaurant

Project the clusters on the map



Conclusions

Some conclusions can be drawn from the above analysis:

1. Within all the districts, Chuo has the largest number of restaurants
2. Ramen restaurants enjoy the highest popularities across all the categories, followed by Japanese restaurants
3. From the map visualization, we can see more restaurants locates in the central Tokyo, suburb areas have relatively fewer restaurants

Discussion

Overall, the foursquare API collaborates with statistical methods, like one-hot encoding and k-means clustering, enable us to have an insight about the popular restaurants in each distributes of Tokyo. The preliminary market along with Folium leaflet map facilitates good visualization, which display the location of the hotspots clearly on the map.

The future work could be done on the following aspects:

Combine the Folium leaflet display with google map, so visitors can get more comprehensive results from the searching.

Also, more machine learning methods can be applied on the datasets, from traditional classification and clustering methods, to the deep learning approaches, like CNN, and reinforcement learning.

To sum up, we explored the Tokyo neighborhood by using the machine learning methods and have an insight about the restaurants location and popularities, which can provide much information for the visitors the local residents.