# Battle of Neighbourhoods

**Problem description and discussion of the background**

# Introduction

## Prospects of lunch Restaurants close to office are in Tokyo, Japan

Tokyo is the most populated metropolitan area in the world. It is one of the best places to start up new business. During the day time, especially in the morning and lunch hours, office area provides huge opportunities for restaurants. Reasonably priced (one lunch meal 8$) shops are usually always full during the lunch hours (11 am — 2 pm) and, given this scenario, we will go through the pros and cons of opening a breakfast cum lunch restaurant in highly dense office places. Usually the profit margin for a decent restaurant lie within 15−20% range but, it can even go high enough to 35%. The core of Tokyo is made of 23 wards (municipalities) but, I will concentrate on 5 busiest business wards of Tokyo — Chiyoda, Chuo, Shinjuku, Shibuya and Shinagawa to target daily office workers.

We will go through each step of this project and address them separately. I first outline the initial data preparation and describe future steps to start the battle of neighbourhood's in Tokyo.

## Target Audience

What type of clients or a group of people would be interested in this project?

1. Business personnel who want to invest or open a restaurant. This analysis will be a comprehensive guide to start or expand restaurants targeting the large pool of office workers in Tokyo during lunch hours.

2. Freelancer who loves to have their own restaurant as a side business. This analysis will give an idea, how beneficial it is to open a restaurant and what are the pros and cons of this business.

3. New graduates, to find reasonable lunch/breakfast place close to office.

4. Budding Data Scientists, who want to implement some of the most used Exploratory Data Analysis techniques to obtain necessary data, analyse it, and, finally be able to tell a story out of it.

## Data Section

For this week describing the initial data preparation and future steps to start the battle of neighbourhoods in Tokyo.

1. Obtain Data
    a. Name the 23 wards, area and population from web scrapping.
    b. Obtain information about best business districts.
    c. Use foursquare data to obtain info about restaurants.
2. Data Visualization and some simple statistical analysis.
3. Analysis using clustering, specially K-Means Clustering
    a. Maximize number of clusters.
    b. Visualize using Chloropleth Map.
4. Compare the Neighbourhood to find the best place for starting up a restaurant.
5. Inference from the results and related conclusions.

# Data Preparation

## 1. Web-Scrapping and Cleaning

- ### Get the Names of Wards, Major Districts and Population from Wikipedia

The Wikipedia page of Tokyo Wards contains the table of 23 wards of Tokyo, area, population and major districts. I have used Beautifulsoup4 and pandas library to create the initial data-frame. For a clean and understandable data-frame some of the wards are renamed for example 'Chiyoda, Tokyo' to 'Chiyoda'. Here I have taken the first entry of the major districts column in the Wikipedia table to concentrate on. Even though not complete but it gives us quite a detailed picture of the corresponding ward, as later on I have considered top most venues within 1 kilometer radius of the major district. After this initial preparation, I moved on to the next step to obtain coordinates using Geopy library.

## 2. Get the Coordinates of the Major Districts

Some of the coo-ordinates of the major districts returned by Geopy are wrong and I have figured out the reason for this is the name of the major districts in the data-frame are

different from their actual names, for example Hongō to Hongo. In these cases (4 of them), I had to Google search and replace using pandas library. After little manipulation the obtained data-frame looks as below

```
Tokyo_df
```

```
Old latitude list:  [35.675618, 35.684058, 35.61912805, 35.6937632, 32.5093796, 35.7117587, 35
38, 35.599252, 35.62125, -38.9047057, 35.646096, 35.6645956, 35.718123, 35.7049419, 35.7301957
37529, 35.774143, 35.74836, 35.4463689, 34.7337515, -5.3498001]
Old Longitude list:  [139.7434685, 139.774501377979, 139.779403349221, 139.7036319, -116.29700
39.8150431, -80.8328748, 139.73891, 139.688014, 175.7552111, 139.65627, 139.6987107, 139.66446
9.7111534, 139.7207999, 139.78131, 139.681209, 139.638735, 139.4309254, 135.3315861, 21.424098
[35.675618, 35.684058, 35.61912805, 35.6937632, 35.7088, 35.7117587, 35.6963122, 35.6722, 35.5
5.5884, 35.646096, 35.6645956, 35.718123, 35.7049419, 35.7301957, 35.7781394, 35.737529, 35.77
5.4463689, 34.7337515, 35.6634]
[139.7434685, 139.774501377979, 139.779403349221, 139.7036319, 139.7601, 139.7776445, 139.8150
9.73891, 139.688014, 139.7279, 139.65627, 139.6987107, 139.664468, 139.649909, 139.7111534, 13
31, 139.681209, 139.638735, 139.4309254, 135.3315861, 139.8731]
```

2]:

| | Ward | Area_SqKm | Population | Major_District | Dist_Latitude | Dist_Longitude |
|---|---|---|---|---|---|---|
| 1 | Chiyoda | 5100 | 59441 | Nagatacho | 35.675618 | 139.743469 |
| 2 | Chuo | 14460 | 147620 | Nihonbashi | 35.684058 | 139.774501 |
| 3 | Minato | 12180 | 248071 | Odaiba | 35.619128 | 139.779403 |

| | Ward | Area_SqKm | Population | Major_District | Dist_Latitude | Dist_Longitude |
|---|---|---|---|---|---|---|
| 1 | Chiyoda | 5100 | 59441 | Nagatacho | 35.675618 | 139.743469 |
| 2 | Chuo | 14460 | 147620 | Nihonbashi | 35.684058 | 139.774501 |
| 3 | Minato | 12180 | 248071 | Odaiba | 35.619128 | 139.779403 |
| 4 | Shinjuku | 18620 | 339211 | Shinjuku | 35.693763 | 139.703632 |
| 5 | Bunkyō | 19790 | 223389 | Hongo | 35.708800 | 139.760100 |
| 6 | Taitō | 19830 | 200486 | Ueno | 35.711759 | 139.777645 |
| 7 | Sumida | 18910 | 260358 | Kinshicho | 35.696312 | 139.815043 |
| 8 | Koto | 12510 | 502579 | Kiba | 35.672200 | 139.806100 |
| 9 | Shinagawa | 17180 | 392492 | Shinagawa | 35.599252 | 139.738910 |

## 3. Obtain the Average Land Price Data from Web-Scrapping

The average land-price data for each ward of Tokyo was obtained from Tokyo land market value page. Even though this data wasn't used for clustering but it definitely helps us to compare different districts of Tokyo for potentially opening a restaurant.

## 4. Foursquare Data

Use Foursquare API to obtain the 100 most common venues within 1 kilometer of each major district.

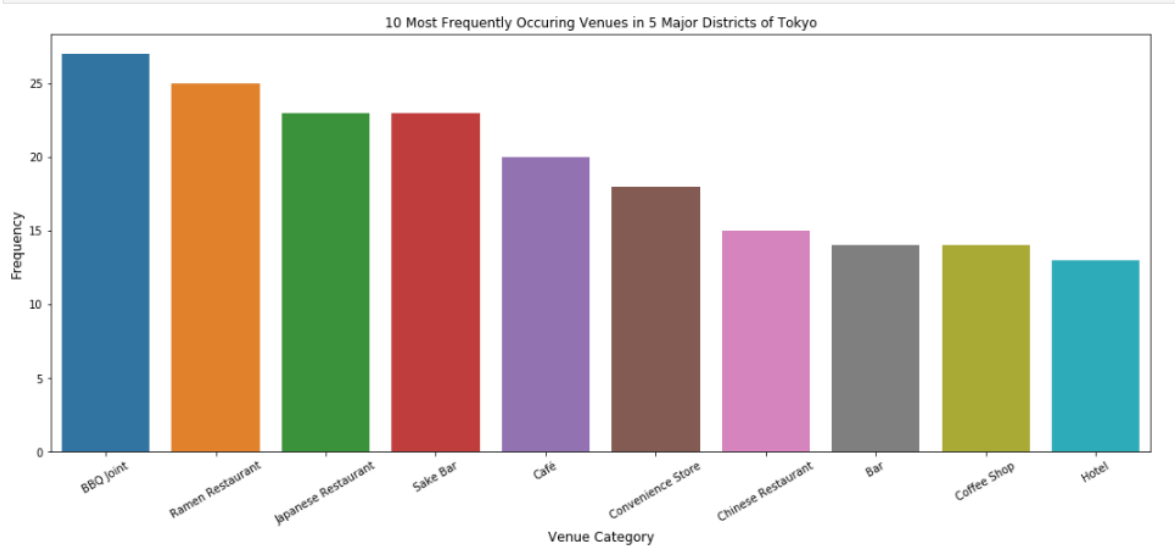
# Exploring the Data and Major Districts of Tokyo

From the Foursquare data, we could see that there are 134 unique categories, but for this analysis.

I mostly later on concentrated in Restaurant category. As the focus is on 5 major business districts (Nagatacho, Nihombashi, Shibuya. Shinjuku, and Shinagawa), we found that there are 193 restaurants (searching for keyword Restaurant) among the 500 top venues in these 5 districts. I have used Folium library to plot a leaflet map of only these restaurants in these 5 major districts of Tokyo which is as shown below, where the colors representations are the follwing-- Nihombashi- Green, Nagatacho- Red, Shibuya- Orange, Shinjuku- Magenta, Shinagawa- Blue.


Here we have found out that

- Ramen restaurants top the charts of most common venues in the 5 districts, followed by Japanese restaurants and BBQ joints.
- A plot of the ten most frequent venues in these 5 districts are as below




## Next step was to obtain information about the top 5 venues of each district. And to do that, I proceed as follows

- Create a data-frame with pandas one hot encoding for the venue categories.
- Use pandas groupby on District column and obtain the mean.
- Transpose the data-frame at step 2 and arrange in descending order.

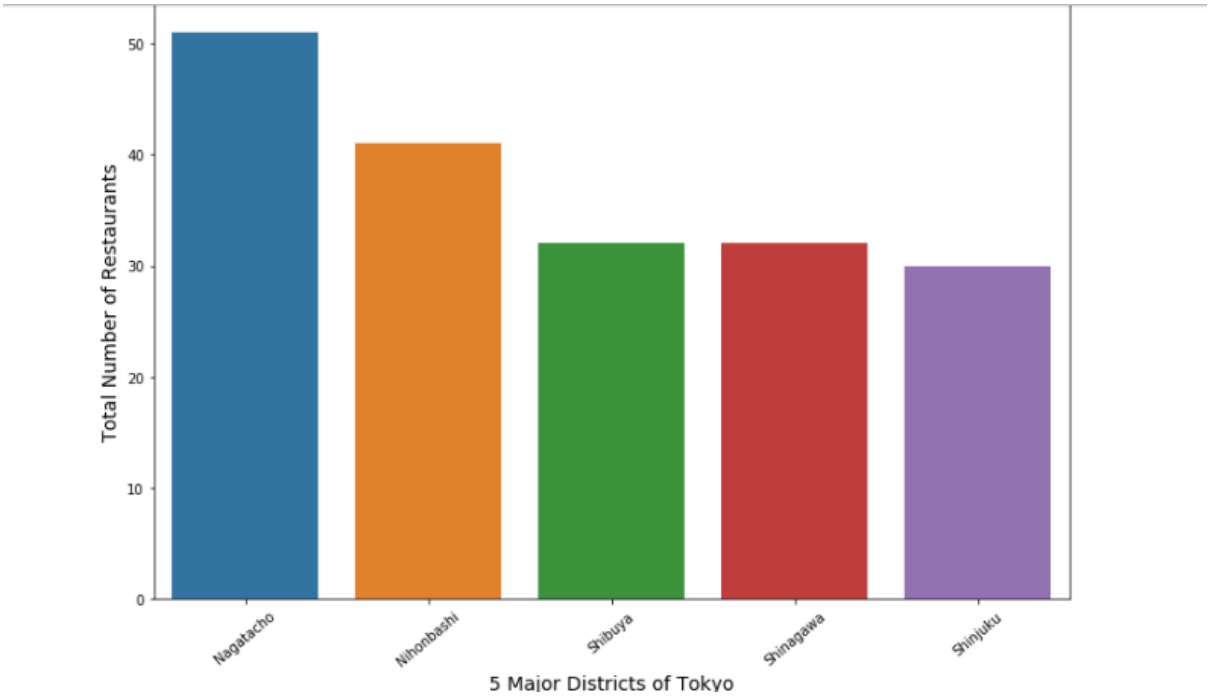
Implementing them in Pandas outputs the following--

```
    print(temp.sort_values('Freq', ascending=False).reset_index
    print('\n')
```

```
%%%%%%%%%%Nagatacho%%%%%%%%%
                 Venue  Freq
0  Japanese Restaurant  0.10
1            BBQ Joint  0.07
2          Coffee Shop  0.06
3     Ramen Restaurant  0.05
4                Hotel  0.05


%%%%%%%%%%Nihonbashi%%%%%%%%%
                 Venue  Freq
0  Japanese Restaurant  0.07
1            BBQ Joint  0.05
2      Soba Restaurant  0.04
3                Café  0.04
4           Hobby Shop  0.04


%%%%%%%%%%Shibuya%%%%%%%%%
               Venue  Freq
0               Café  0.10
1        Record Shop  0.07
2        Coffee Shop  0.05
3   Sushi Restaurant  0.04
4  French Restaurant  0.03
```

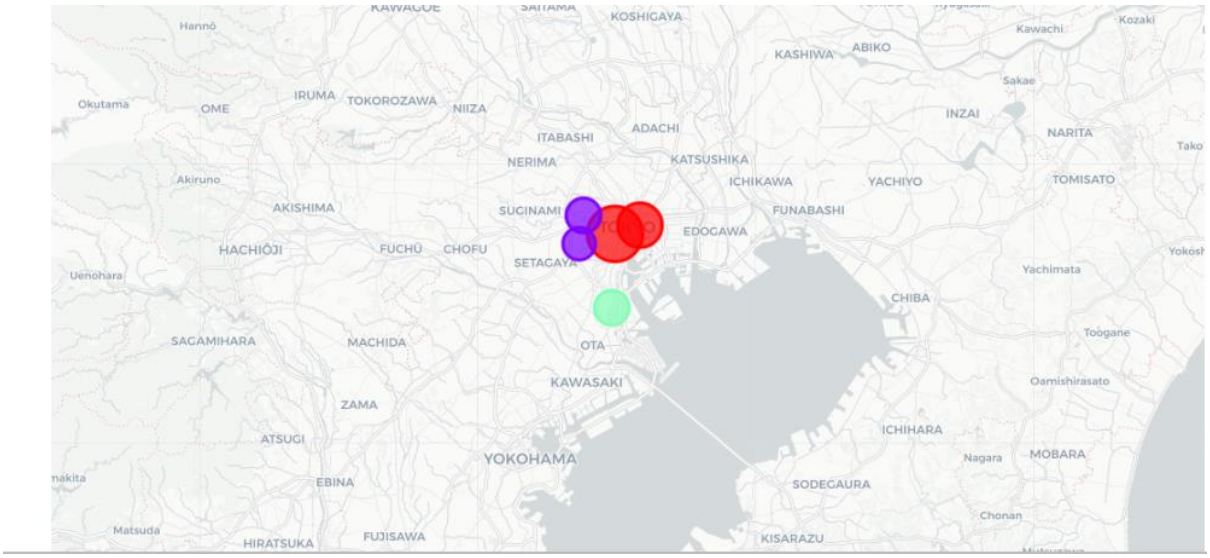
```
%%%%%%%%%%Shinagawa%%%%%%%%%
                 Venue  Freq
0  Convenience Store  0.17
1   Ramen Restaurant  0.09
2           Sake Bar  0.08
3          BBQ Joint  0.06
4               Park  0.03


%%%%%%%%%%Shinjuku%%%%%%%%%
                  Venue  Freq
0             Sake Bar  0.09
1                  Bar  0.09
2     Ramen Restaurant  0.06
3            BBQ Joint  0.06
4  Japanese Restaurant  0.04
```

Also I explored which district has the highest number of restaurants as the most common venue and the plot below is the answer




## Clustering the Major Districts of Tokyo

Finally, we try to cluster these 5 districts based on the frequency of venue categories and, use K-Means clustering. So our expectation would be based on the similarities of venue categories, these districts will be clustered. Using K-Means algorithm from Scikit-learn library we obtain 3 clusters as shown below.




Here the radius of the circles represent the number of restaurants as most common venue for the corresponding district and, we have seen before that it is maximum for Nagatacho district (56) and minimum for Shibuya (26).

From the most common venues this clustering makes a complete sense as Shibuya, Shinjuku are dominated by pubs, bars and cafe falls under the purple cluster, whereas Nagatacho, Nihombashi dominated by Japanese and Chinese restaurants falls under red cluster and Shinagawa stands alone (green cluster).

## Results

The results of the exploratory data analysis and clustering are summarized below--

- Ramen restaurants top the charts of most common venues in the 5 districts.
- Nagatacho district in Chiyoda ward and Nihombashi in Chuo ward are dominated by Japanese and Chinese restaurants as the the most common venues.
- Shibuya and Shinjuku areas are dominated by bars, pubs, and cafe as most common venues.
- Nagatacho has maximum number of restaurants as the most common venue whereas has Shibuya area has the least. But, Cafe and BBQ joints are found to be among the most visited destinations in this area.
- ***Since the clustering was based only on the most common venues o each district, Shinjuku, Shibuya fall under the same cluster and, Nagatacho, Nihonbashi fall under another cluster. Shinagawa is separated from both of these clusters as, convenient stores stand out as the most common venue (with a very high frequency).***

## Discussion

According to this analysis, Shinagawa area will provide least competition for an upcoming lunch restaurant as convenience store is the most common venue in this area and the frequency of restaurants as common venue are very low compared to the remaining districts.

Also seen from the web-scrapped data, the average land price in and around Shinagawa is much cheaper compared to the districts close to central Tokyo. So, definitely this region could potentially be a target for starting quality restaurants.

Some drawbacks of this analysis are-- the clustering is completely based on the most common venues obtained from Foursquare data. Since land price, distance of the venues from closest stations, number of potential customers, benefits and drawbacks of Shinagawa being a port region, could all play a major role and thus, this analysis is definitely far from being conclusory. However, it definitely gives us some very important preliminary information on possibilities of opening restaurants around the major districts of Tokyo.

Also, another pitfall of this analysis could be consideration of only one major district of each ward of Tokyo, taking into account of all the areas under the 5 major wards would give us an even more realistic picture. Furthermore, this results also could potentially vary if we use some other clustering techniques like DBSCAN.

## Conclusion

Finally to conclude this project, We have got a small glimpse of how real life data-science projects look like. I have made use of some frequently used python libraries to scrap web-data, use Foursquare API to explore the major districts of Tokyo and saw the results of segmentation of districts using Folium leaflet map. Potential for this kind of analysis in a real life business problem is discussed in great detail. Also, some of the drawbacks and chance for improvements to represent even more realistic pictures are mentioned. Finally, since my analysis were mostly concentrated on the possibilities of opening a restaurants targeting the huge pool of office workers, some of the results obtained are surprisingly what I have expect after staying 5 years in Tokyo.