Battle of Neighborhoods

1. Problem Statement and Background
Prospects of a Restaurant in Tokyo, Japan.

Tokyo is one of the most populous metropolitan area and one of the best places to start up a new business in the world. Currently it is number 3 in the global economic power index.
Office areas in Tokyo provide huge opportunities for restaurants. Average priced shops are mostly full during the peak hours. Considering this we will check Go/No-Go decision of opening a breakfast cum lunch restaurants near CBD area of Tokyo.

The core of Tokyo is made of 23 wards (municipalities) but, I will later 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.
Target Audience

1. 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.
2. Budding Data Scientists, who wants 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.
3. New graduates, to find reasonable lunch/breakfast place close to office.
4. Business personnel who wants 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. Data Preparation:
2.1. Web-Scrapping and Cleaning (Week 1)
2.1.1. Getting 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 topmost venues within 1 kilometre radius of the major district. After this initial preparation, I moved on to the next step to obtain coordinates using Geopy library.

2.1.2. Getting 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
```

	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	35.711788	139.776096
7	Sumida	18910	260358	Kinshicho	35.696752	139.814151
8	Koto	12510	502579	Kiba	35.672200	139.806100
9	Shinagawa	17180	392492	Shinagawa	35.599252	139.738910
10	Meguro	19110	280283	Meguro	35.621250	139.688014
11	Ota	11910	722608	Omori	35.588400	139.727900
12	Setagaya	15690	910868	Setagaya	35.646096	139.656270
13	Shibuya	15080	227850	Shibuya	35.664596	139.698711
14	Nakano	21350	332902	Nakano	35.718123	139.664468
15	Suginami	16750	570483	Koenji	35.704942	139.649909
16	Toshima	22650	294673	Ikebukuro	35.730103	139.711884
17	Kita	16740	345063	Akabane	35.778139	139.720800
18	Arakawa	21030	213648	Arakawa	35.737529	139.781310
19	Itabashi	17670	569225	Itabashi	35.774143	139.681209
20	Nerima	15120	726748	Nerima	35.748360	139.638735
21	Adachi	12660	674067	Ayase	35.446369	139.430925
22	Katsushika	12850	447140	Tateishi	34.176335	132.226020
23	Edogawa	13750	685899	Kasai	35.663400	139.873100

2.1.3. Obtaining the Average Land Price Data from Web-Scrapping

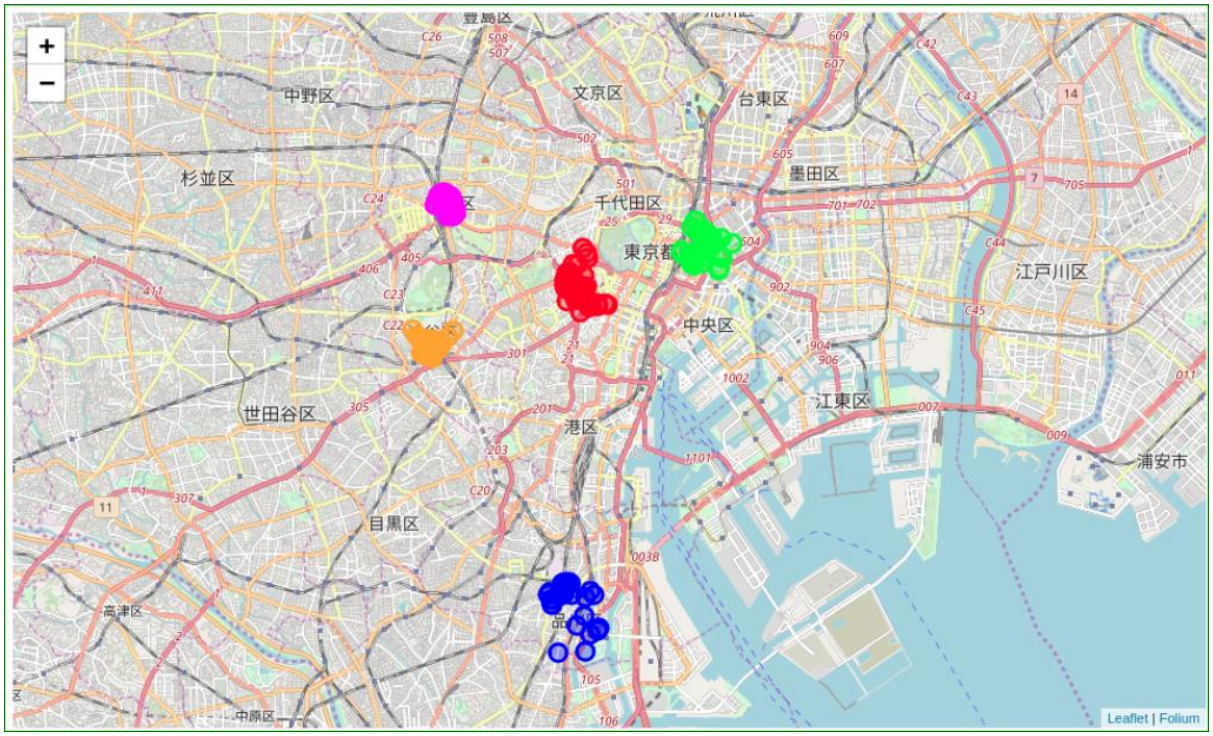
The average land-price data for each ward of Tokyo was obtained from Tokyo land market value page. Even though this data wasn't used for clustering, but it definitely helps us to compare different districts of Tokyo for potentially opening a restaurant.

2.2. Foursquare Data

Finally, I used Foursquare API to obtain the 100 most common venues within 1 kilometre of each major district.

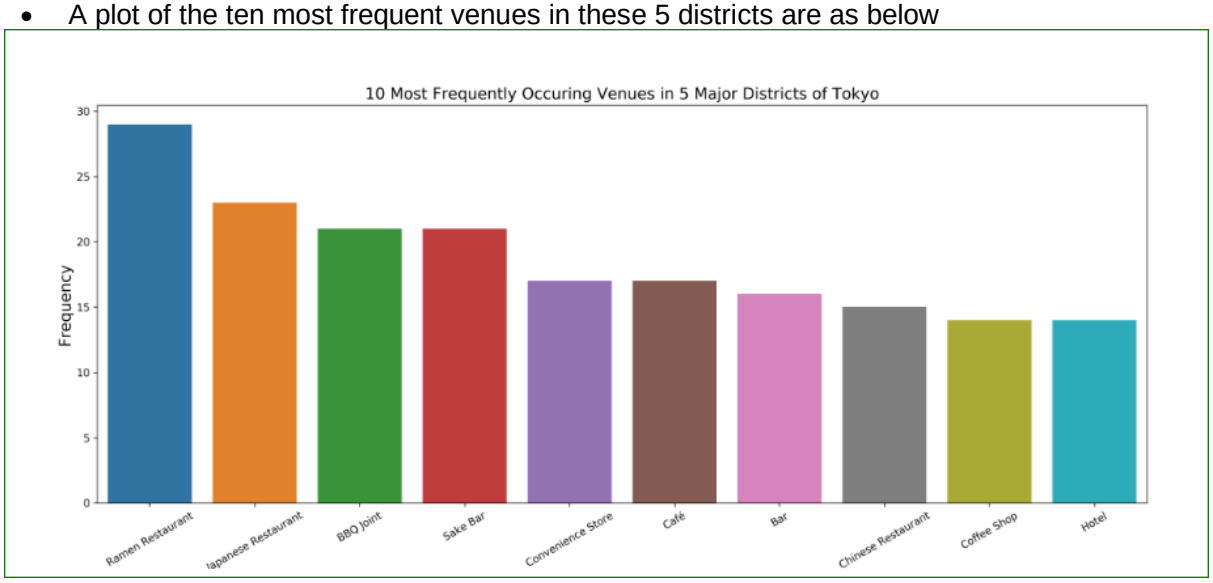
3.1. Exploring the Data and Major Districts of Tokyo

I had emphasized mostly on concentrated area of 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



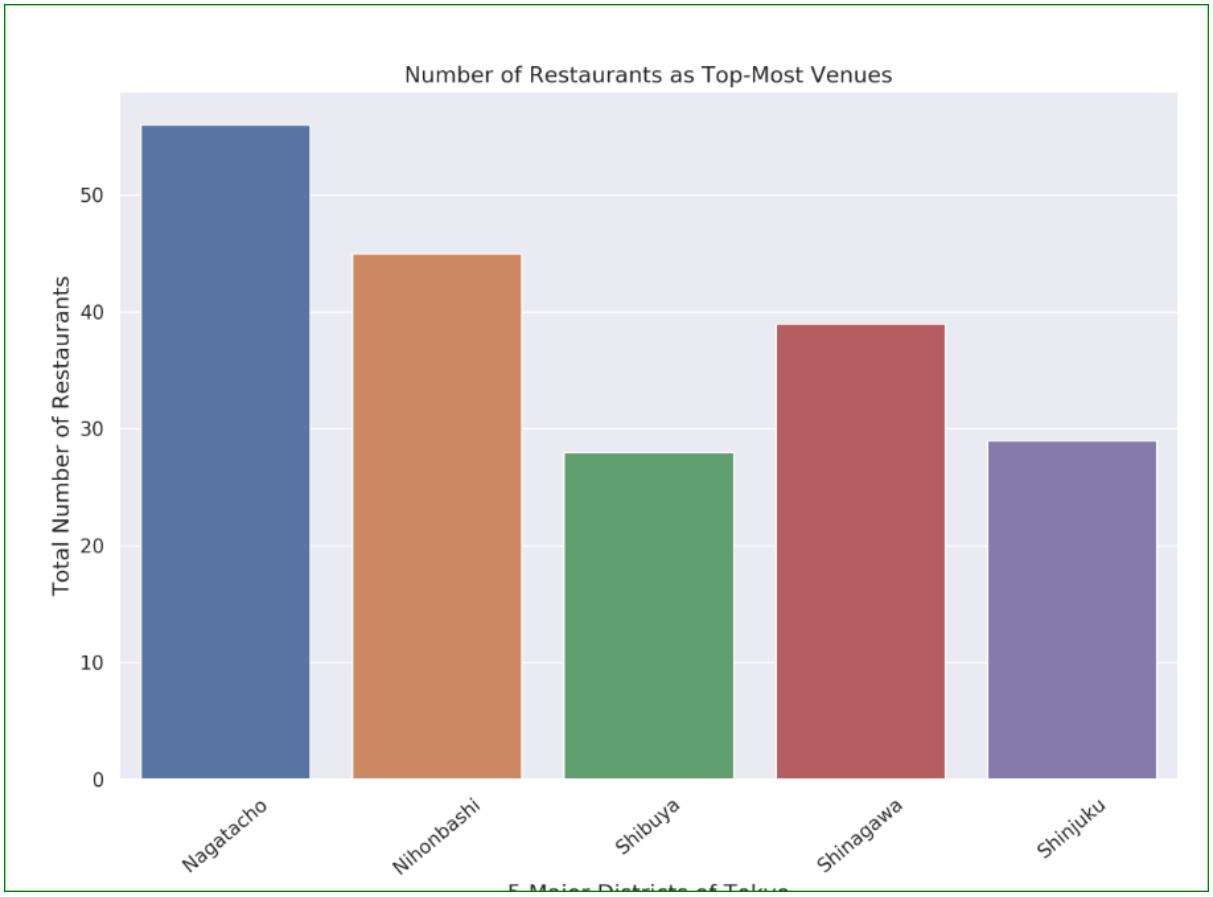
Next step was to obtain information about the top 5 venues of each district. And to do that, I proceed as follows

- 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.

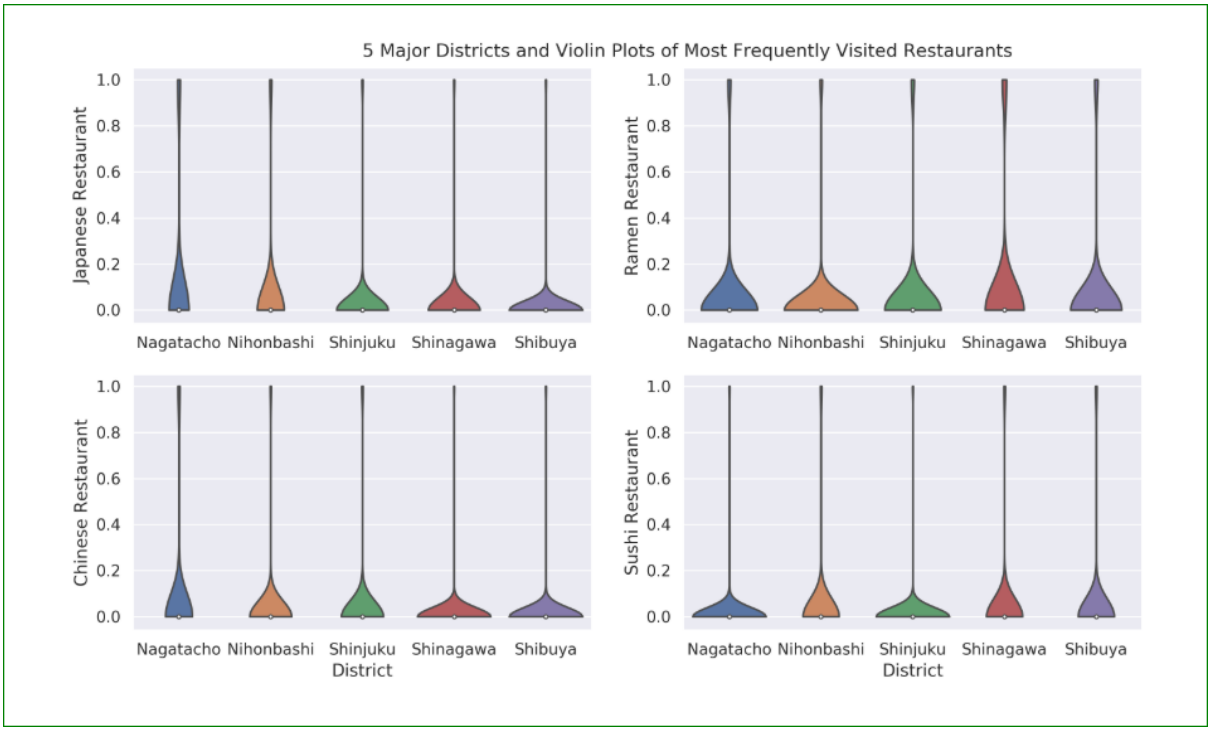
Implementing them in Pandas outputs the following--

```
###########Nagatacho###########
                 Venue  Freq
0  Japanese Restaurant  0.11
1            BBQ Joint  0.07
2     Ramen Restaurant  0.06
3          Coffee Shop  0.06
4   Chinese Restaurant  0.06


###########Nihonbashi###########
                 Venue  Freq
0  Japanese Restaurant  0.09
1            BBQ Joint  0.05
2      Soba Restaurant  0.05
3                Café  0.05
4           Hobby Shop  0.04


###########Shibuya###########
               Venue  Freq
0               Café  0.10
1        Coffee Shop  0.05
2        Record Shop  0.05
3   Ramen Restaurant  0.04
4  French Restaurant  0.03


###########Shinagawa###########
               Venue  Freq
0  Convenience Store  0.14
1   Ramen Restaurant  0.09
2           Sake Bar  0.07
3      Grocery Store  0.06
4          BBQ Joint  0.05


###########Shinjuku###########
                 Venue  Freq
0             Sake Bar  0.09
1                  Bar  0.08
2     Ramen Restaurant  0.07
3  Japanese Restaurant  0.06
4                  Pub  0.04
```

Also I wanted to explore **which district has the highest number of restaurants as the most common venue** and the plot below is the answer



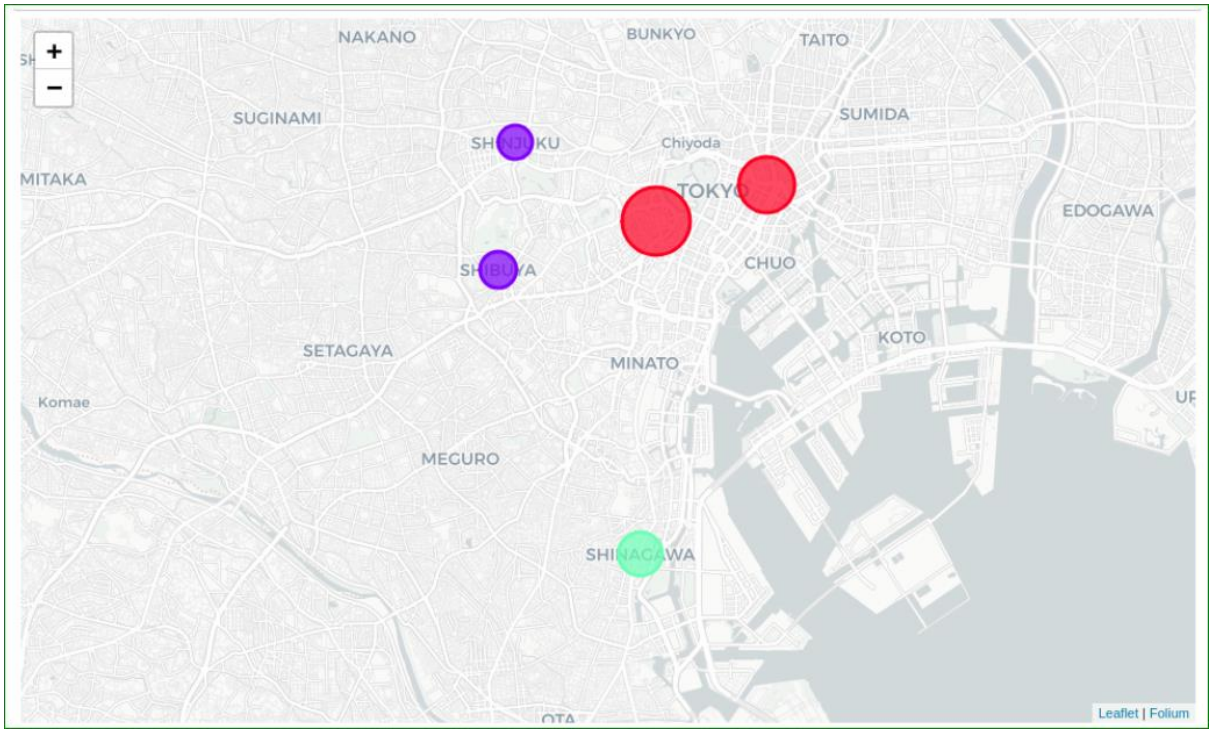
Since my focus is mostly on the competitions provided by restaurants on these districts, I also explored **how most common restaurant categories are distributed in each district** and the violin plot below of categorical variables give the answer.



From here we see that **lots of Japanese and Chinese restaurants are in Nagatacho, whereas Shinagawa has many Ramen restaurants.** Also, it is important to note here that most common venue in shinagawa are convenient stores.

3.2. 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 rom Scikit-learn library we obtain 3 clusters as shown below.



Here the **radius of the circles represents 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).

4. Results

The resutls 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).**

5. 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, 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 far from being conclusory. However, it 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, considering 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.

6. 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.