**Description of the Problem and Discussion of the Background**

**(Introduction Section, Week 1)**

**Prospects of a Lunch Restaurant, Close to Office Areas in Tokyo, Japan.**

Tokyo, where I am currently staying, is the most populous metropolitan area in the world. Currently ranked 3rd in the global economic power index, Tokyo is definitely one of the best places to start up a new business.

During the daytime, specially in the morning and lunch hours, office areas provide huge opportunities for restaurants. Reasonably priced (one lunch meal) shops are usually always full during the lunch hours (11 am -- 2 pm) and, given this scenario, we will go through the benefits and pitfalls of opening a breakfast cum lunch restaurants in highly densed office places. Usually the profit margin for a decent restaurant lie within range but, it can even go high enough to , as discussed [here](https://www.ichefpos.com/en-sg/blog/japanese-restaurants-profits). The core of Tokyo is made of 23 wards (municipalities) but, I will later concentrate on 5 most busiest business wards of Tokyo - Chiyoda (千代田区), Chuo (中央区), Shinjuku (新宿区), Shibuya (渋谷区) and Shinagawa (品川区), to target daily office workers.



We will go through each step of this project and address them separately.

<font color=orange>**Target Audience**</font>

1. 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. Freelancer who loves to have their own restaurant as a side business. This analysis will give an idea, how  
   benificial 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 wants to implement some of the most used Exploratory Data Analysis techniques to  
   obtain necessary data, analyze it and, finally be able to tell a story out of it.

**2. Initial Data Preparation:**

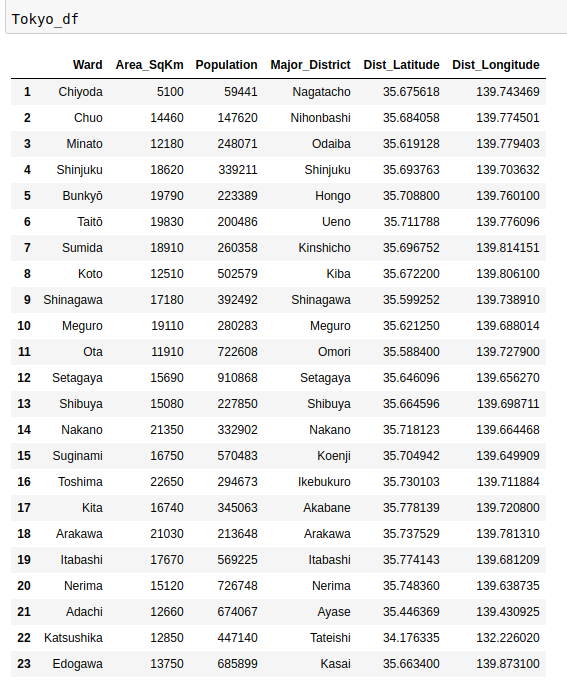
**2.1. Web-Scrapping and Cleaning (Week 1)**

**2.1.1. Get The Names of Wards, Major Districts and Population from Wikipedia**

The wikipedia page of [Tokyo Wards](https://en.wikipedia.org/wiki/Special_wards_of_Tokyo) contains the table of 23 wards of Tokyo, area, population and major districts. I have used [Beautifulsoup4](https://pypi.org/project/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 inital preparation, I moved on to the next step to obtain coordinates using [Geopy](https://geopy.readthedocs.io/en/stable/) library.

**2.1.2. 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

.

**2.1.3. Obtain the Average Land Price Data from Web-Scrapping**

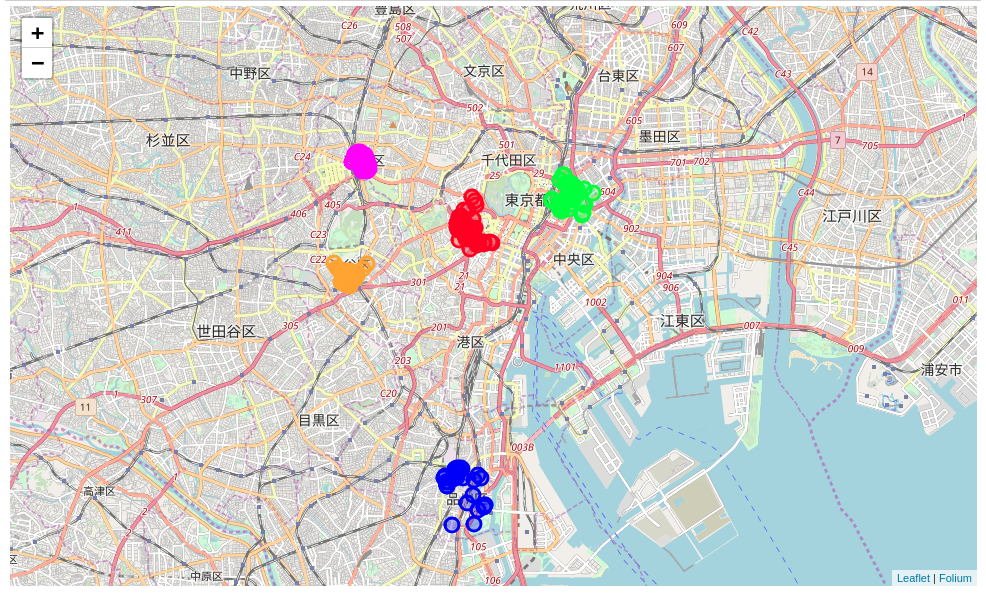
The average land-price data for each ward of Tokyo was obtained from [Tokyo land market value](https://utinokati.com/en/details/land-market-value/area/Tokyo/) page. Even though this data wasn't used for clustering but it definitely helps us to compare different districts of Tokyo for potentially opening a restaurant.

**2.2. Foursquare Data (Week 2)**

Finally, I make use of [Foursquare API](https://developer.foursquare.com/) to obtain the 100 most common venues within 1 kilometer of each major district.

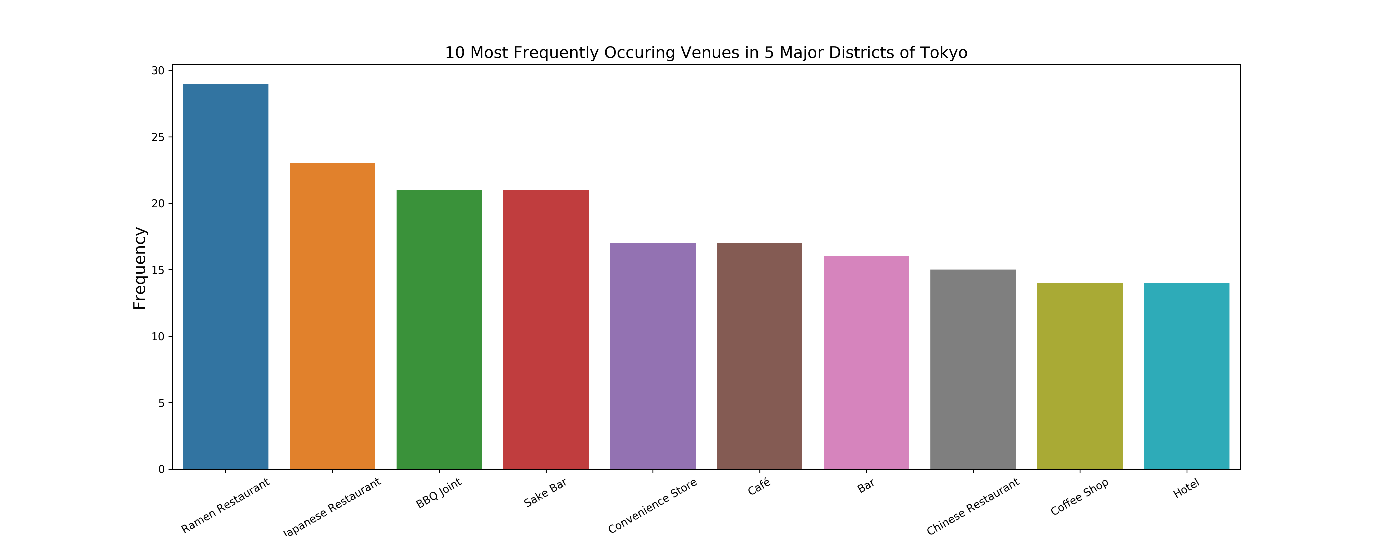
**3.1. Exploring the Data and Major Districts of Tokyo**

From the Foursquare data, we could see that there are  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](https://python-visualization.github.io/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 following-- 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



Since this reminded of delicious ramen so I need to take a small break

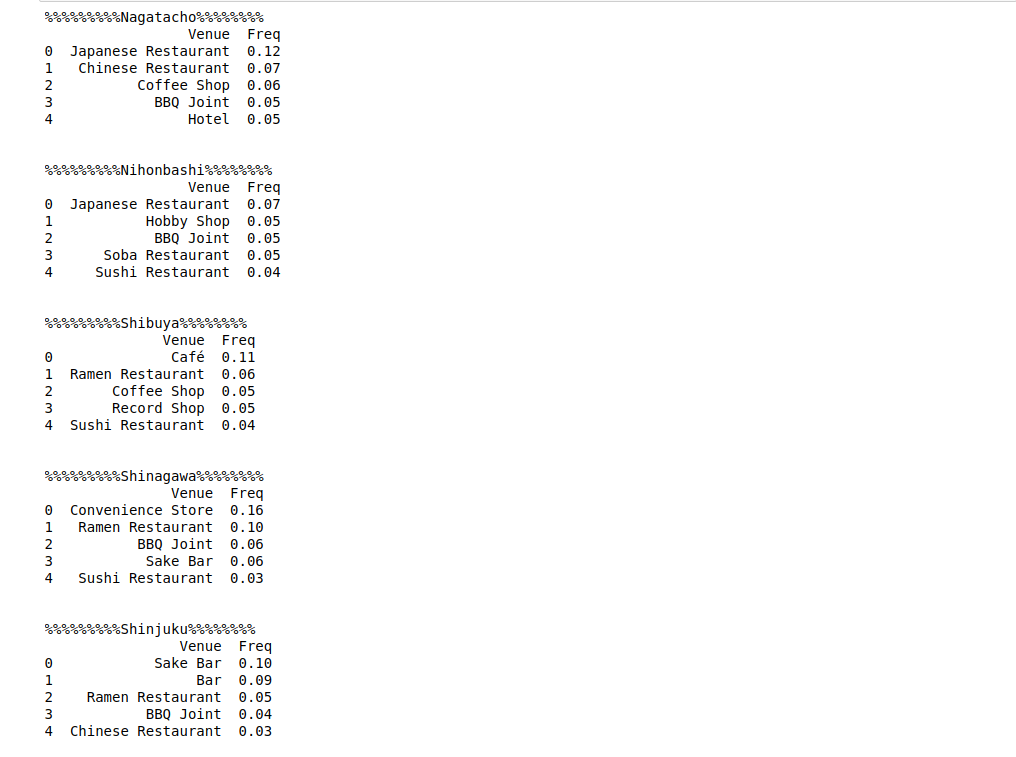


Okay, let's begin again.

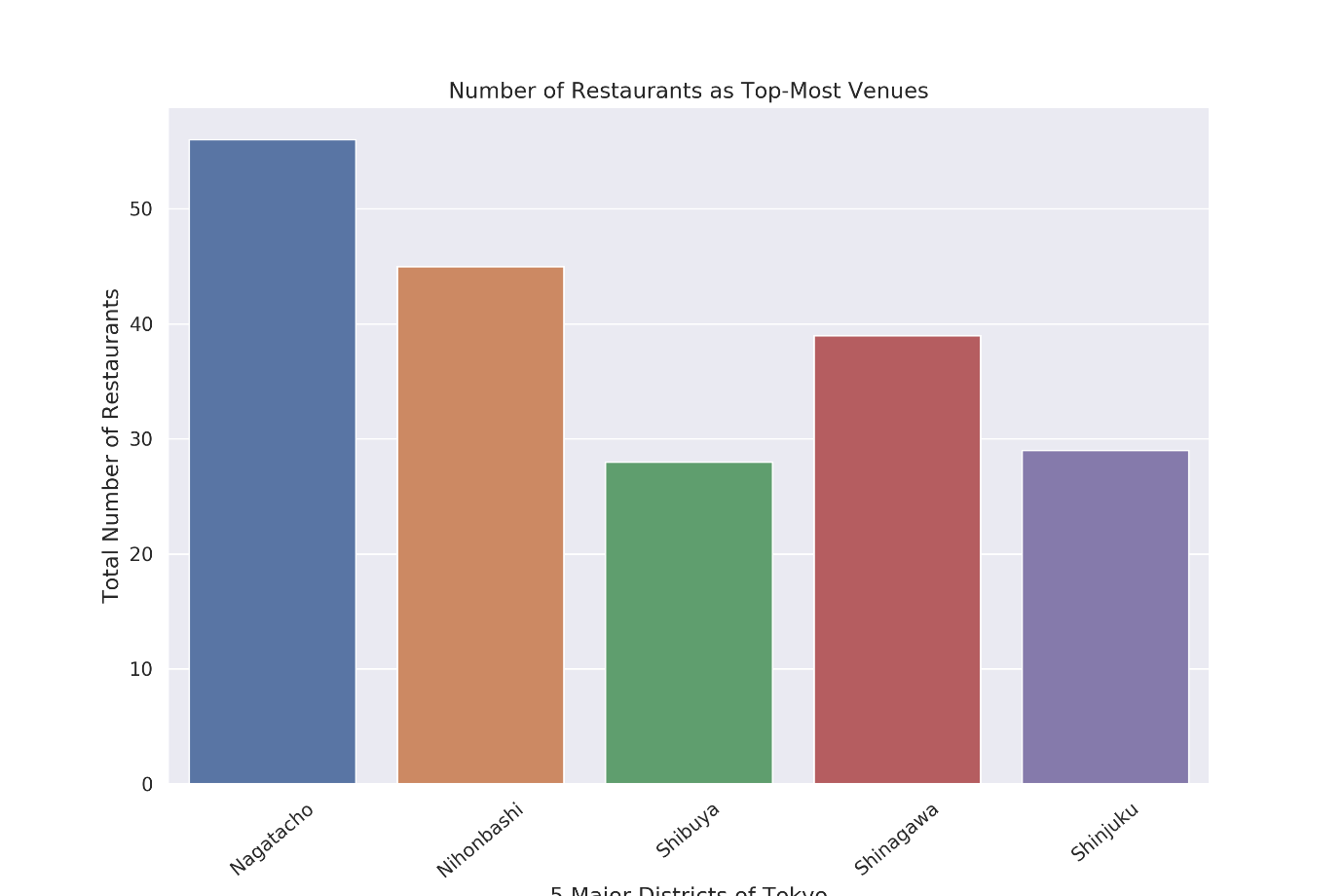
**Next step was to obtain information about the top 5 venues of each district. And to do that, I proceed as follows**

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

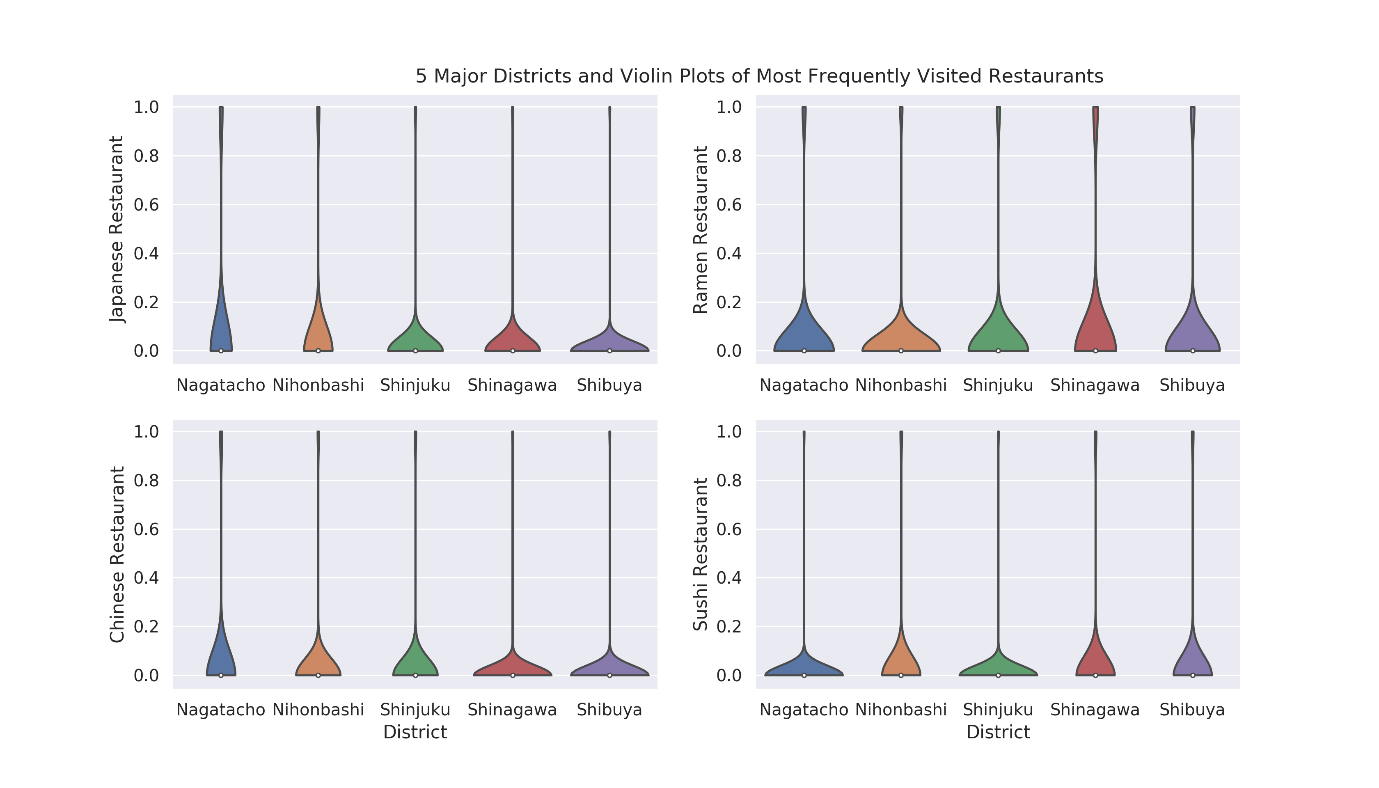
Implementing them in Pandas outputs the following--



* 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



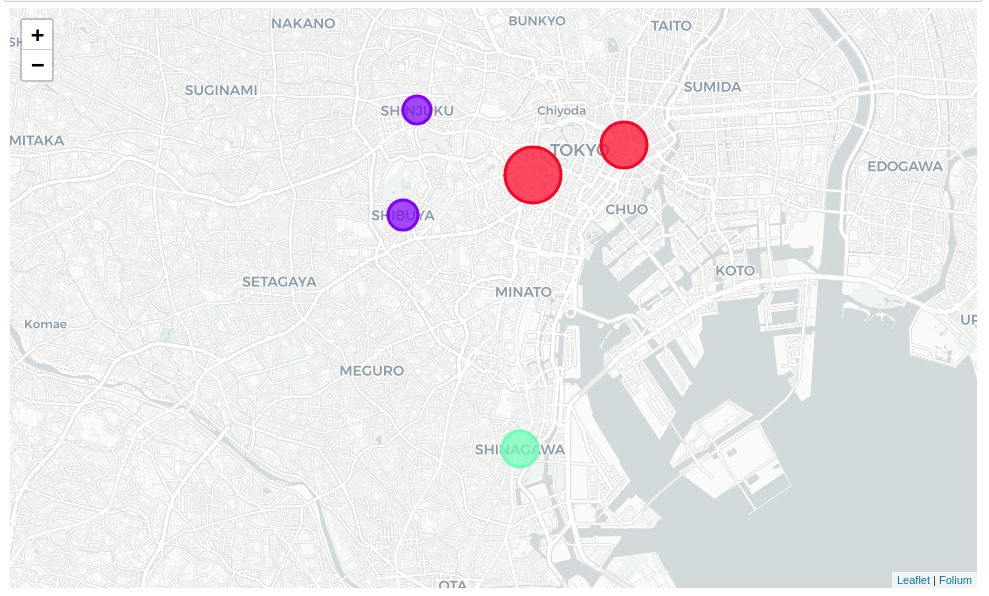
Since my focus is mostly on the competetions provided by restaurants on these districts, I also explored **how most common restaurant categories are distributed in each district** and the violin plot below of categorical variables give the answer.



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. Does that tell us something ?

**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 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).

**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, 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.

**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 librairies to scrap web-data, use Foursquare API to explore the major districts of Tokyo and saw the results of segmanetation of districts using Folium leaflet map. Potential for this kind of analysis in a real life business problem is dicussed 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 expecteafter staying 5 years in Tokyo.