IBM Capstone Project: The Battle of Neighborhoods

Alfredo Montero Fernández

17 de enero de 2021

Introduction to the problem

Tokyo is one of my favourites cities across all the world. It has so many things to offer: nice people, beautiful sights, a lot of entertainments and really really good food. Being as big as it is you can find a lot of different zones in the city that it will make you think you are travelling between cities.

The number of overseas visitors is getting incremented year by year, getting a peak of nearly 32 million in 2019! Sadly 2020 has been a really hard time for tourism and hostelry, but we expect that in the near future everything will return to the good days for visting this amazing countries.

With the vast amount of restaurants distributed across the 23 wards of the city, it can be really hard to decide which one to choose. It even has 226 restaurants with a star on the Michelin guide! So with the help of Foursquare and Machine Learning algorithms, we will be able to get a good recommendation for you depending on what you want, recommending the best district for you! This way if you are a visitor to the rising sun country, continue reading, this is for you!

Data preparation

- First we need the data of the different Tokyo's wards. We can obtain it from: https://en.wikipedia.org/wiki/Special_wards_of_Tokyo#List_of_special_wards

- Secondly we will clean it and convert it to a data frame with the geographic location of every ward.

After this step we obtain a data frame with its 5 first rows shown in the Figure 1

	Neighborhood	Kanji	Population	Density	Area	Latitude	Longitude
0	Chiyoda	千代田区	59441	5100	11.66	35.693810	139.753216
1	Chūō	中央区	147620	14460	10.21	35.666255	139.775565
2	Minato	港区	248071	12180	20.37	35.643227	139.740055
3	Shinjuku	新宿区	339211	18620	18.22	35.693763	139.703632
4	Bunkyō	文京区	223389	19790	11.29	35.718810	139.744732

Figura 1: Data obtained from Wikipedia

- Lastly, with the help of Foursquare we will retrieve data for the restaurants on every ward and we will use it to compare each one of them and divide them into different klusters using K-means methods, so deppending on your tastes you will be able to know which wards are going to reward you with your favourite food. We will also obtain data from leaflet with the library folium so we can display a nice visualization map.

Methodology

After wrangling the necessary data, we use the python library *folium* to display a map of Tokyo and its 23 wards using the longitude and latitude values obtained. The map can be seen in the Figure 2

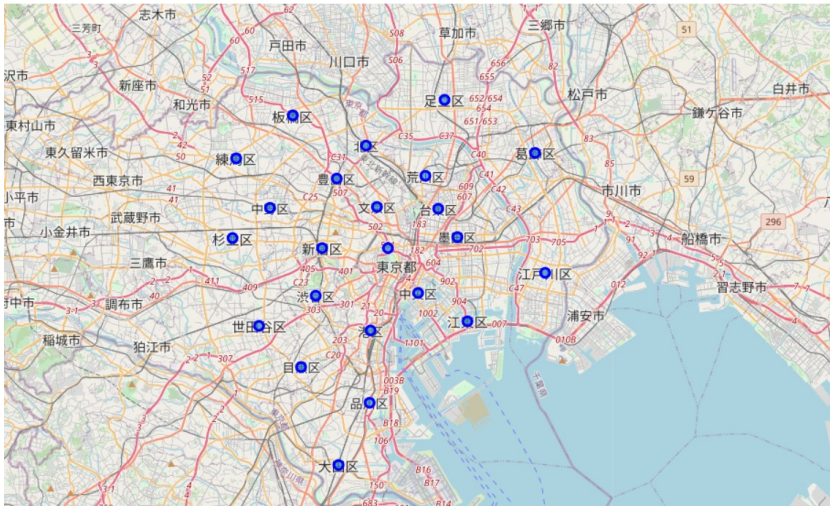

Figura 2: Map of Tokyo and its wards

Using the Foursquare API we will retrieve data which will help the future travellers reading this to get knowledge on the kind of food mostly served on each of the wards by grouping them into clusters for those wards which are alike.

To show how the data is displayed, first we just focus on Minato, one of my favourite Tokyo's wards. Retrieving the data we get 48 distinct restaurant types in the zone, where the top 10 in frequency of this zone are shown in the Figure 3.

Japanese Restaurant	13
Chinese Restaurant	9
Italian Restaurant	6
Coffee Shop	5
Soba Restaurant	5
Indian Restaurant	4
Sake Bar	3
BBQ Joint	3
Ramen Restaurant	3
Yakitori Restaurant	2

Figura 3: Top 10 most frequent type of restaurants in Minato

Then we do the same for all the 23 wards and retrieve a barplot showing the top 10 more frequent types of restaurants. This barplot is shown in the Figure 4.

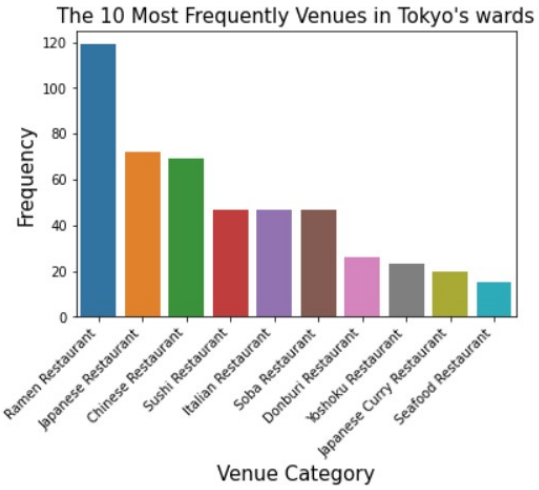

Figura 4: Top 10 most frequent type of restaurants in Tokyo

As you can see, Ramen wins by far. This is somehow normal there. Ramen restaurants are fast and delicious food that a lot of workers visit between the morning and the afternoon turns of work. But this did not happen in Minato, so we would like to know on which wards are ramen restaurants that common that they get the most common of Tokyo.

In the Barplot 5 we can observe that Minato is the second ward with more proportion of restaurants, but nothing compared with *Chūō*, which is the number one! This is normal since Ginza and Nihonbashi, two of the most popular upscale shopping areas of Tokyo, are districts of this warp.

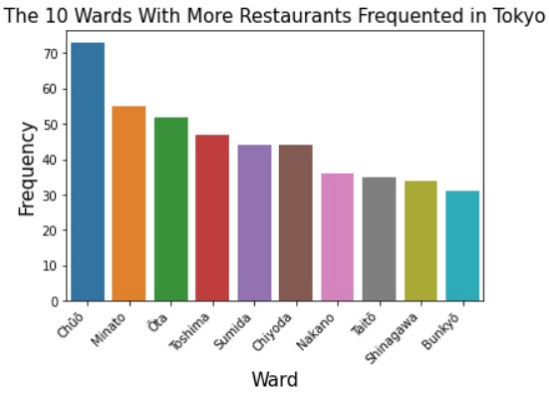

Figura 5: Top 10 wards with more proportion of restaurants

To answer the question of how the different wards compare to each other we first create a one-hot enconding using *pandas*. This way we can calculate the mean of the frequency of ocurrence of each restaurant category. Obtaining this way a data frame with its first 6 rows showed in the Figure 6

	Neighborhood	African Restaurant	American Restaurant	Asian Restaurant	Bangladeshi Restaurant	Belgian Restaurant	Brazilian Restaurant	Cantonese Restaurant	Chinese Restaurant	Dim Sum Restaurant	Donburi Restaurant	Dongbei Restaurant
0	Adachi	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.130435	0.000000	0.086957	0.000000
1	Arakawa	0.000000	0.000000	0.000000	0.04	0.000000	0.000000	0.000000	0.200000	0.000000	0.120000	0.000000
2	Bunkyō	0.000000	0.032258	0.000000	0.00	0.000000	0.000000	0.000000	0.096774	0.000000	0.096774	0.000000
3	Chiyoda	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.113636	0.000000	0.022727	0.000000
4	Chūō	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.013699	0.000000	0.027397	0.000000
5	Edogawa	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.047619	0.000000	0.095238	0.000000

Figura 6: Proportion of all the restaurants on its kind and ward

And then we create a data frame with the top 10 most frequent restaurant type for each ward. We show the first 5 rows in the Figure 7.

	Neighborhood	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	10th Most Common Venue
0	Adachi	Ramen Restaurant	Japanese Family Restaurant	Chinese Restaurant	Fast Food Restaurant	Donburi Restaurant	Indian Restaurant	Yoshoku Restaurant	Japanese Restaurant	Restaurant	Italian Restaurant
1	Arakawa	Chinese Restaurant	Italian Restaurant	Donburi Restaurant	Ramen Restaurant	Sushi Restaurant	Fast Food Restaurant	Japanese Restaurant	Japanese Curry Restaurant	Tonkatsu Restaurant	Japanese Family Restaurant
2	Bunkyō	Italian Restaurant	Donburi Restaurant	Chinese Restaurant	Japanese Curry Restaurant	Ramen Restaurant	Japanese Restaurant	Dumpling Restaurant	Soba Restaurant	Japanese Family Restaurant	Vietnamese Restaurant
3	Chiyoda	Japanese Curry Restaurant	Ramen Restaurant	Chinese Restaurant	Tempura Restaurant	Soba Restaurant	Yoshoku Restaurant	Sushi Restaurant	Okonomiyaki Restaurant	Udon Restaurant	Russian Restaurant
4	Chūō	Sushi Restaurant	Japanese Restaurant	Monjayaki Restaurant	Soba Restaurant	Italian Restaurant	Kaiseki Restaurant	Seafood Restaurant	Tonkatsu Restaurant	Japanese Curry Restaurant	Tempura Restaurant

Figura 7: Top 10 most frequent restaurant types sorted for each ward

Next, for the process of clustering the wards, we will use the machine learning technique of K-means (you can consult https://en.wikipedia.org/wiki/K-means_clustering for more information about it) taking on account the most common type of restaurant for each ward as a variable. This technique of unsupervised learning works really well on this kind of problems, so it returns at least a nice visualization result. We will use a value of $k = 5$ so it does not divide our wards too much since there are only 23. After aplying the algorithm, we display again the map of Tokyo and its wards, but this time we show with a circle of 5 different colours the wards, where 2 wards with the same color belong to the same cluster, and the radius of the circle is bigger as more proportion of restaurants it has.

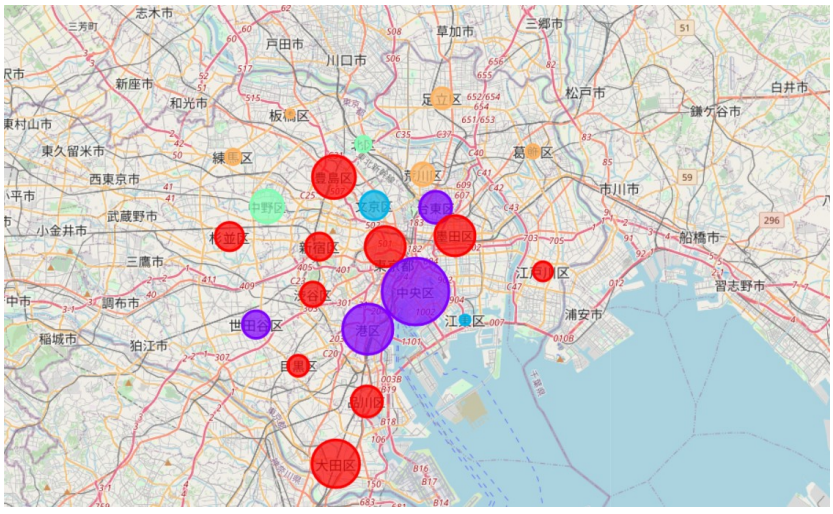

Figura 8: Map of Tokyo and its wards clustered

There seems to be some kind of relationship between the number of restaurants and the most frequent ones as we can see in the Figure 8.

Results and Discussion

We found some nice results. For instance we observated that ramen restaurants are the most frequent ones across Tokyo. This is mainly beacause of what we commented before. Ramen restaurants are, generally, cheap, fast, but also delicious. On the other hand, the two wards with more restaurants were *Chūō* and Minato. In *Chūō* the most frequent kind of restaurant is no more than sushi restaurants! This could be explained because as ramen restaurants are cheap, fast and delicious, sushi restaurants are also really fast and delicious, but they are more expensive, so only in the richest parts of the city they can be used day by day by local workers, and as we commented before this is what happens in Ginza (a district of *Chūō*), one of the most expensive districts of all Japan and where the extremely high-quality products are located.

On thing we did not take into account is that we just gor clusters in function of the most common types of restaurants, but we did not think about the price of those, or the distance from the center of the city, the quantity of workers and other variables.

Conclusion

Data science can be used in a lot of different scenarios. This project is just a very simple example that could be extended by taking into account more variables as well as retrieving more precise data. In this case we wanted to ask the question: ïf I am a traveller in Tokyo and want to eat a specific type of food, in which ward should I search so Iit will be easier to find what I exactly want?üsing some data wragling, using some machine learning techniques and plotting results so it is easier for the reader to understand. And about the question made, for example if you want to eat sushi (and probably really really good sushi), and you have the money for it, just go to Ginza!