Tokyo analysis – Where to set a new business?

Capstone final project – Applied Data Science

Author: Joao Nunes

Date: 20/06/2021

Introduction

1. Background and problem description

Tokyo, or officially known as Tokyo Metropolis, is the capital of Japan and the most populated prefecture in the entire country. The city holds around 13,960,236 people across 23 special wards. Such a high amount of people aggregated in one city causes its density to reach around 6,363 people per square kilometre.

Densely populated areas tend to lead to a highly diversified market demand for food and other catering services. This can easily turn into a double-edged sword. On one hand, successful businesses can thrive at a faster pace and expand, however, this also means that businesses have added pressure to keep up with world trends and to cater to new customer needs in order to out-compete their massive competition. Furthermore, new businesses have an even harder time to enter this already established ecosystem.

Location, location, location. **Where should one start?**

- From a shop owner perspective, a place that is located in a highly dense area, with "hopefully" lower land costs and even more "hopefully" less direct competition would be a good start.
- From an investor perspective, the same information could be quite insightful to understand a business potential longevity and challenges (competition wise) in the short to mid-term.

This project aims at providing a solution that relies heavily on the "easy visualization" data staple. So that both new shop owners and investors can quickly gather insight on viable new opportunities.

2. Data description

I. Initially, a potential ward of interest will be shortlisted based on population density and land price factors.
II. From this, its respective boroughs will be analysed in terms of common venues - for potential indirect/direct competitors and synergies with other businesses.
III. Lastly, the land price values for each borough will be overlapped on a world map with the common venues clustering information to further help reduce the potential areas to setup a new business

The required information will be extracted from publicly available resources:

- Wards Density information [1]
- Tokyo Land market value list [2]
- ZIP codes within Tokyo [3]
- Location coordinates using Geocoder Python Package [4]
- **Foursquare location** to extract respective borough information. [5]

Methodology (new bakery)

A. Data scrapping and pre-processing

To gather information for our potential case study – the establishment of a new bakery within Tokyo – the first few steps involved web-scapping Tokyo's wards data (e.g. name, density, postal code and average price per land) and compiling it into a data frame. Any unwanted extra information was removed, and some pre-processing was made to guarantee that the data frame had the correct type of data. As an example, "Average Price (JPY/sq.m)" column had to be cleaned up since originally this column had string type data and for the purpose of the analysis, integers were needed.

B. Selecting the most promising ward

As mentioned in the Introduction, the first two assumptions made were that a new owner would want to establish a business in a location with maximum exposure to clients (high density area) while avoiding expensive areas that would require a bigger initial investment to either rent/buy land space.

For this purpose, the density versus average land prices were plotted against each other to understand the overall distribution within Tokyo's wards (Figure 1). The initial analysis revealed that Chiyoda prefecture had the least density but was the most expensive ward in terms of average land price. On the other end of the spectrum, Toshima had the highest density and relatively average land price.

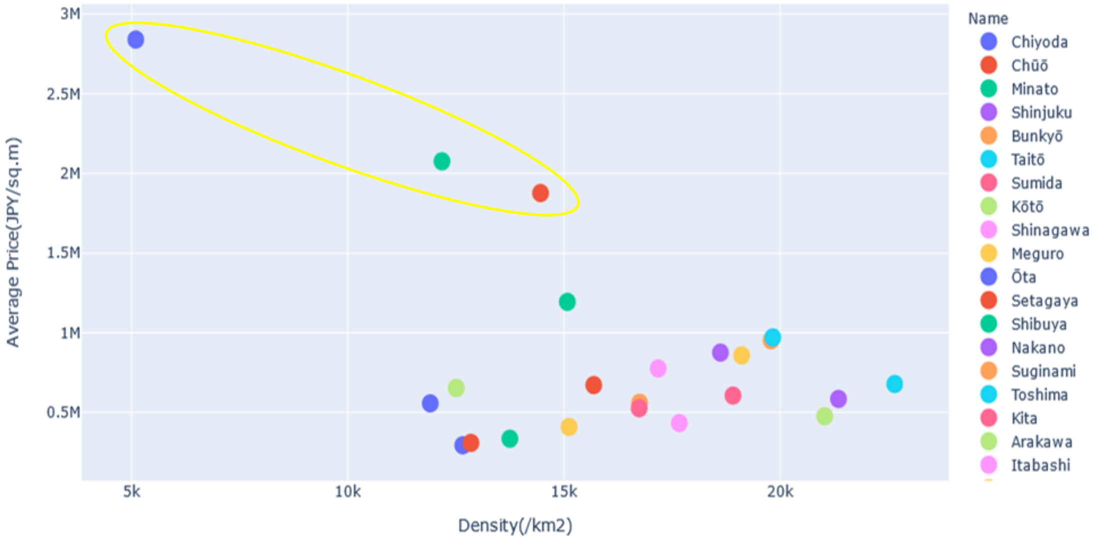

Figure 1 - Tokyo ward analysis. All wards within Tokyo were plotted regarding their density versus average price. There was a clear distinction between Chiyoda, Shibuya and Setagaya ward compared to all the others in terms of land value.

To make sure that the chosen ward had the best ratio between density and land price, the top 10 wards with the highest ratio were plotted (Figure 2). This led to the selection of Arakawa ward as the target city to be further explored (upon further reading, this ward revealed to be a residential area).

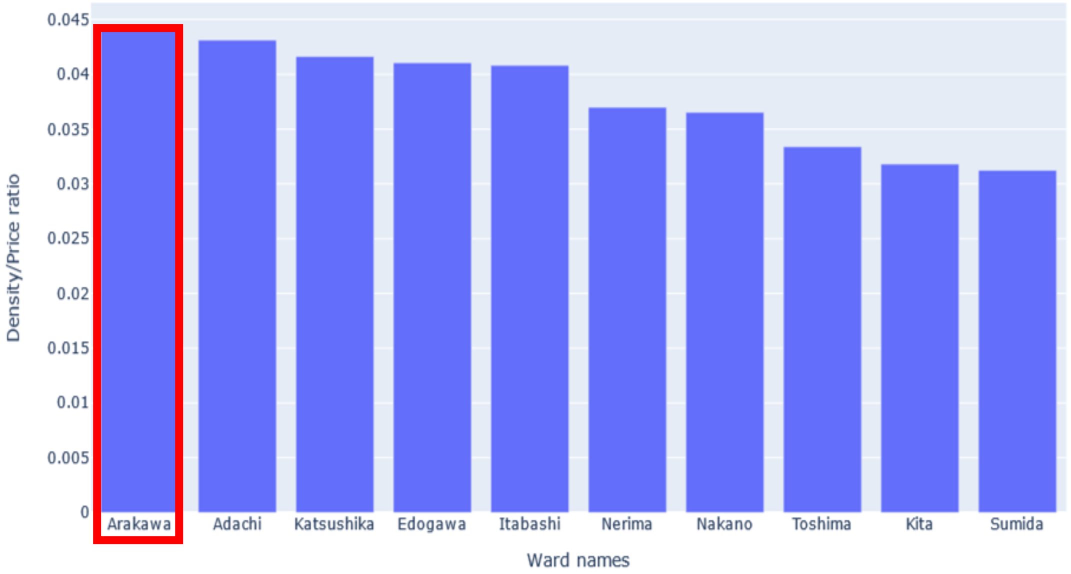

Figure 2 - Tokyo ward analysis. Top 10 wards with the highest density/average price ratio. Depending on the target business, Adachi could have also been chosen.

C. Extract information from the selected ward

With a ward in mind, the next step involved selecting data regarding each neighborhoods' name, postal code, coordinates as well as land price to combine it into a data frame. For this section, web-scrapping did not work as intended and some data had to be manually compiled into a text file. Reasons for this involved (1) html native webpage structure being difficult to extract desired data (2) neighborhoods' names being written in Japanese and had to be translated to English (3) *geolocator* library being stuck within the *while* loop never retrieving any coordinates. Google maps had to be used as an alternative to obtain the coordinates for each neighborhood.

D. Neighborhoods' venues analysis

Using *Foursquare* with a limit of 200 venues and a radius of 500, venues' names, coordinates and category types were extracted. Further analysis on the unique venue categories revealed that there could be 3 potential direct competitors – bakeries, sandwich places and pastry shops. In order to understand each neighborhood structure, *K-means* clustering algorithm was chosen to cluster neighborhoods based on their venues.

K-means inertia was used for the "elbow method" when choosing the optimal number of clusters (Figure 3). However, this result was slightly ambiguous with potentially cluster number 3 and 4 working. Between the former two, clustering with *k = 3* was chosen for division as results can be seen on Figure 4.

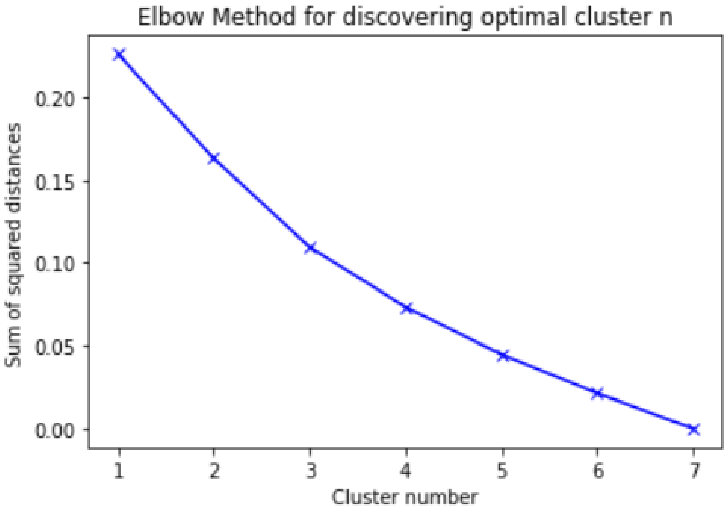

Figure 3 – Inertia (also named as the sum of squared distances) was obtained to identify the optimal clustering number for this analysis.

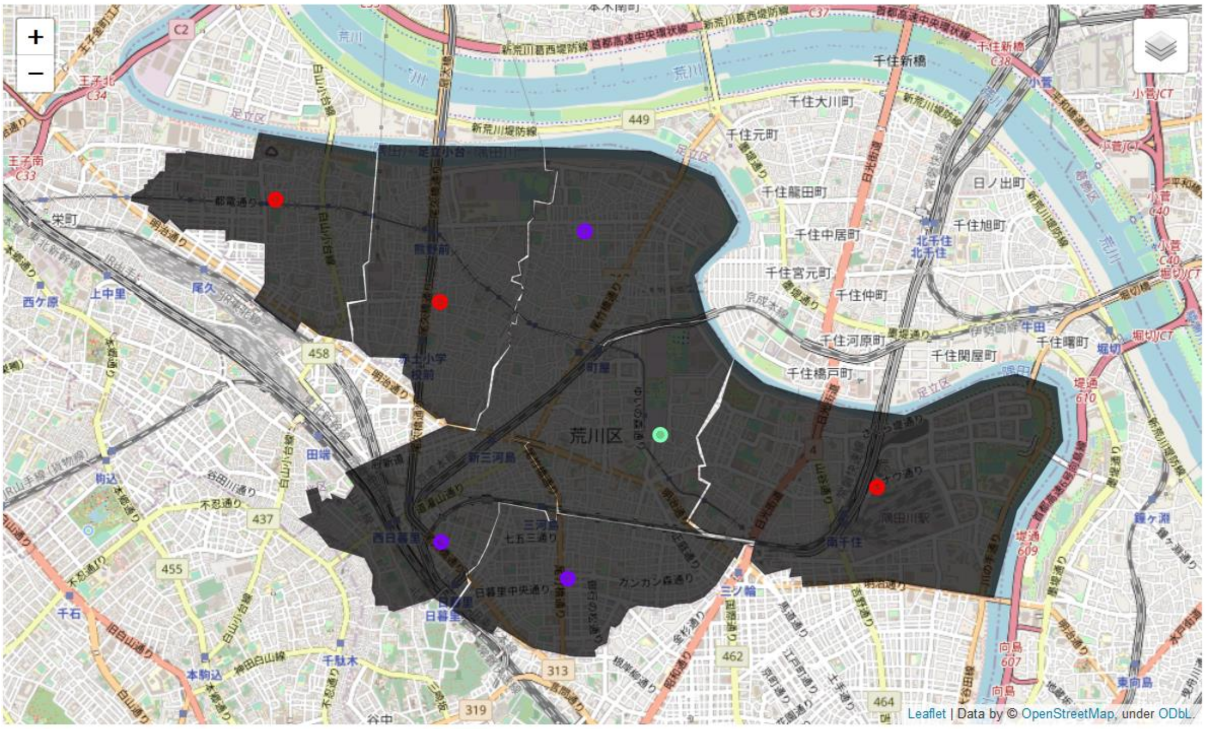

Figure 4 – Choropleth map of Arakawa ward with shaded areas showing the vicinity for each neighborhood and coloured markers locating the centre of each respective neighborhood. Colours are based on the clustering results with *k = 3* as follows: Cluster 1 (red) - Higashiogu, Minamisenju and Nishiogu; Cluster 2 (purple) - Higashinippori, Machiya and Nishinippori; Cluster 3 (green) - Arakawa.

Further analysis on this cluster division revealed that the issue was due to venues similarity across all neighborhoods. Considering all other venues as similar, cluster 1 neighborhoods had a higher amount of train/tram stations compared to the other two clusters. Cluster 2 had a higher number of supermarkets while cluster 3 revealed a large number of venues being parks. For this reason, venue clustering did not provide significant results for choosing a potential neighborhood since there was no major difference across neighborhoods, which explains the results obtained for Figure 3.

E. Neighborhood selection

To further narrow down the list of potential neighborhoods, data obtained regarding direct competitors as well as average land price was used (Figure 5).

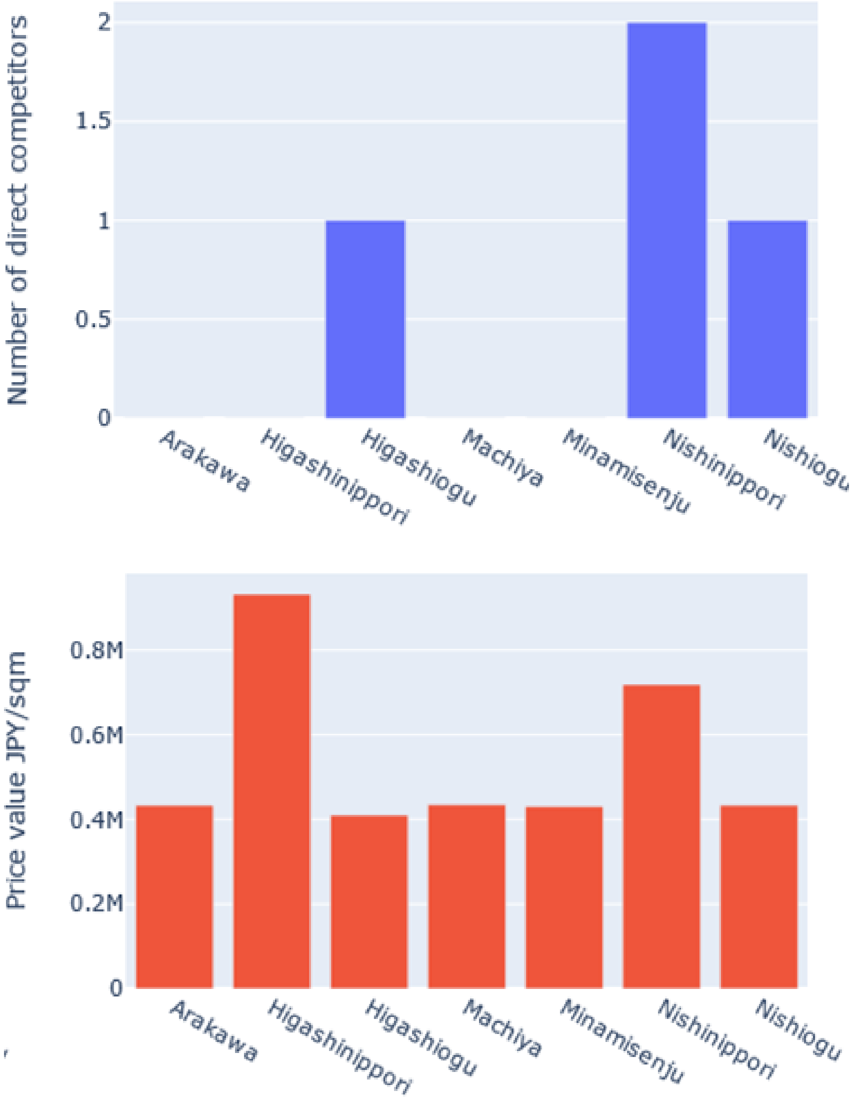

Figure 5 – Top bar plot represents the number of direct competitors within each neighborhood. Direct competitors were counted as any business within the following categories: bakery, pastry shop or sandwich place. Bottom bar plot exhibits the average land price value per neighborhood.

From the previous results, Higashiogu, Nishinippori and Nishiogu were excluded due to the presence of direct competitors. Higashinippori was also excluded due to the high average land price value. This left Arakawa, Machiya and Minamisenju as potential neighborhoods for the establishment of a new bakery business.

Discussion

As introduced before, Tokyo is a city with enormous potential for new businesses, however, it is a high risk, high reward kind of situation. When starting a new business in such a competitive established market, location and access to target customers are key for the longevity of the business.

Through this analysis, I have used public data to narrow the number of options down to the neighborhood area. I am aware that a few assumptions have been made about the chosen "bakery" study case that would need to be re-adapted for other businesses that could have different priorities.

Further research could involve data regarding the average age population within each neighborhood as well as commuting numbers in/out of each neighborhood. The main reason is because some services could want to establish themselves where there is a high number of people converging to during the working week. Restaurants, as an example, could wand to focus on workers that go around to find a meal during lunch time, therefore areas with the highest amount of companies/number of employees ratio could be attractive in this situation. As for the former type of information, some services have a target age group which was not accounted for in this analysis.

Conclusion

In such competitive ecosystems, it is difficult for a newcomer to know where to start a new business. It can also be challenging for investors to gauge how well a new business could perform in the short-midterm due to a multitude of variables. This analysis is not by far exhaustive, but hopefully can shed some insight on potential factors that could help in the decision or evaluation making process.

Reference:

Cover image obtained from - https://sensanalytics.com/executive-master-classes/introduction-data-science/