Data collection, preprocessing (Week 4)

1. The data sets

- Main data
  The information collected by Foursquare on the most common locations in all 23 special wards of Tokyo, which is used to determine the most popular categories of venues in each district, building the foundation for future business.
- Supplementary data
  Data on the geographical, demographic, social, and economic conditions of Tokyo ’s 23 special wards to understand statistic information such as local income levels, consumption levels, and consumption preferences etc. , which are helping to define and refine the business target population in Tokyo.

2. Data pre-processing

2.1 Retrieve data from below original data source

① Wikipedia webpage for introduction of Tokyo Special Wards
Data source : https://en.wikipedia.org/wiki/Special_wards_of_Tokyo
Here you can find basic information about Tokyo's 23 special wards (population, population density, area, main neighborhoods in each area, etc. This will help us get a rough idea about basic information of each special wards in Tokyo.

② Demographic data about population in Tokyo Special Wards
Data source : https://www.toukei.metro.tokyo.lg.jp/jsuikei/2020/js203a0000.xls
The demographic information of Wikipedia above is 2016, which is relatively outdated, so in order to obtain the most accurate information, we queried the website from the Tokyo Metropolitan Bureau of Statistics and found the actual demographic data of 2020 in it. The latest data for 2020 will also be used in further processing.

③ Average monthly rental market prices in Tokyo Special Wards
Data source : https://www.daiwahouse.co.jp/chintai/tokyo/souba/
Here you can find the average monthly rent price information of each district in Tokyo. It should be noted that the price used here is the average market price of all room types, that is, other factors such as room size and specific location are not considered.

④ Education level (graduation rate of university) in Tokyo Special Wards
Data source : http://wildhog.hatenablog.com/entry/2018/05/01/170000
This website provides the university graduation rate of each district in Tokyo in 2018 (the information for 2019 and later is temporarily not found), from which we can know the

total number of college students in each district, the number of university graduates, the university graduation rate and so on.

⑤ Average annual income in Tokyo Special Wards
Data source :
https://www.nenshuu.net/prefecture/shotoku/shotoku_pre.php?prefecture=%E6%9D%B1%E4%BA%AC%E9%83%BD
Here you can find information on the annual income per capita of each district in Tokyo. The data is very simple, without considering the specific industry, type of work, gender and other factors.

⑥ Geographical data
GeoPy : https://geopy.readthedocs.io/en/stable/
GeoPy is a Python module used for locating the coordinates of addresses, cities, countries, and landmarks worldwide. This project is used to obtain the geographical coordinates of each special wards in Tokyo.

⑦ Foursquare
Foursquare : https://de.foursquare.com/
Foursquare is a location-based recommendation service for restaurants and other places. It is used to obtain information on the most popular venues (restaurants, cafes, sightseeing spots, etc.) in various districts of Tokyo. It is also the main source of information for this project.

2.2 **Convert to pandas data frames and clean the data**

After collecting the original data from the above data sources, imported all of them into the project notebook, and then import the required tools and libraries. From now, we can start the data pre-processing.

- Demonstration of processed datasets
  (Note: here only shows the head of each data frame i.e. first 5 rows)

① Wikipedia webpage for introduction + ② Updated demographic data

	Name	Kanji	Major_districts	Population(2020)	Area(km²)	Density(/km²)
0	Chiyoda	千代田区	Nagatachō, Kasumigaseki, Ōtemachi, Marunouchi,...	66080.0	11.66	5667.0
1	Chūō	中央区	Nihonbashi, Kayabachō, Ginza, Tsukiji, Hatchōb...	168553.0	10.21	16509.0
2	Minato	港区	Odaiba, Shinbashi, Hamamatsuchō, Mita, Roppong...	260535.0	20.37	12790.0
3	Shinjuku	新宿区	Shinjuku, Takadanobaba, Ōkubo, Kagurazaka, Ich...	349101.0	18.22	19160.0
4	Bunkyō	文京区	Hongō, Yayoi, Hakusan	236043.0	11.29	20907.0

③ Average monthly rental market prices in Tokyo Special Wards

	Kanji	RMP/(MM_yen)
0	千代田区	22.2
1	中央区	13.9
2	港区	19.4
3	新宿区	14.5
4	文京区	12.9

④ Education level (graduation rate of university) in Tokyo Special Wards

	Kanji	Total	Graduates	Uni_Grad	Uni_ratio
0	千代田区	41978	38922	14290	36.7%
1	杉並区	432766	396403	143649	36.2%
2	中央区	109813	104866	37566	35.8%
3	文京区	182238	163476	58188	35.6%
4	港区	179914	170019	53193	31.3%

⑤ Average annual income in Tokyo Special Wards

	Kanji	AvAn/(MM_yen)
0	港区	1217
1	千代田区	1081
2	渋谷区	872
3	中央区	690
4	目黒区	637

⑥ Geographical data

	Name	Kanji	Major_districts	Population(2020)	Area(km²)	Density(/km²)	MD_Latitude	MD_Longitude
0	Chiyoda	千代田区	Nagatachō, Kasumigaseki, Ōtemachi, Marunouchi,...	66080.0	11.66	5667.0	35.693810	139.753216
1	Chūō	中央区	Nihonbashi, Kayabachō, Ginza, Tsukiji, Hatchōb...	168553.0	10.21	16509.0	35.666255	139.775565
2	Minato	港区	Odaiba, Shinbashi, Hamamatsuchō, Mita, Roppong...	260535.0	20.37	12790.0	35.643227	139.740055
3	Shinjuku	新宿区	Shinjuku, Takadanobaba, Ōkubo, Kagurazaka, Ich...	349101.0	18.22	19160.0	35.693763	139.703632
4	Bunkyō	文京区	Hongō, Yayoi, Hakusan	236043.0	11.29	20907.0	35.718810	139.744732

⑦ Foursquare (Here can only demonstrate a very small part of the whole data frame due to its huge size)

	Major_districts	MD_Latitude	MD_Longitude	Venue	VN_Latitude	VN_Longitude	Category
0	Nagatachō, Kasumigaseki, Ōtemachi, Marunouchi,...	35.69381	139.753216	Bondy (欧風カレー ボンディ)	35.695544	139.757356	Japanese Curry Restaurant
1	Nagatachō, Kasumigaseki, Ōtemachi, Marunouchi,...	35.69381	139.753216	Nippon Budokan (日本武道館)	35.693356	139.749865	Stadium
2	Nagatachō, Kasumigaseki, Ōtemachi, Marunouchi,...	35.69381	139.753216	National Museum of Modern Art (東京国立近代美術館)	35.690541	139.754694	Art Museum
3	Nagatachō, Kasumigaseki, Ōtemachi, Marunouchi,...	35.69381	139.753216	Kitanomaru Park (北の丸公園)	35.691653	139.751201	Park
4	Nagatachō, Kasumigaseki, Ōtemachi, Marunouchi,...	35.69381	139.753216	Kanda Coffee	35.697455	139.754686	Café

(Codes will be displayed in Week 5 part 2)

2.3 **Saving the data for further processing and analyses**

(To be continue... to part 2)