

Comparison between Tokyo and Osaka with Data Science

COURSERA CAPSTONE PROJECT

TOMONAGA NOGAMI

Index

1. Introduction
2. Business Problem
3. Data Description
4. Methodology
5. Results and Discussion
6. Conclusion

1. Introduction

Tokyo and Osaka are the two largest cities in Japan. You can find a lot of information from internet or books which introduce landmarks, restaurants, shopping streets, parks and so on. But I decided to explore the cities with different approach by using data science. What are the characteristics of Japan's two major cities? Are Tokyo and Osaka similar or not? Could data science find interesting discovery?



2. Business Problem

Data science has a potential to tell us what people won't be able to notice and what is not written in books or magazines. The aim of this project is to provide new insights to tourists who are interested in Japan or to travel agencies who publish books or provide information on the internet.



3. Data Description

Special Wards of Tokyo is an inner city of Tokyo Prefecture and has 23 districts, whereas Osaka city is a center of Osaka Prefecture and has 24 districts. I will use clustering analysis methodology for each district and it's convenient that two cities have similar number of districts. In addition, below is basic comparison of Tokyo and Osaka, which could complement the results of data analysis (although I won't use these info in Python programming).

	Special Wards of Tokyo	Osaka City
Population	9,659,769	2,750,812
Area	627.57km ²	225.21km ²
Population Density	15,392/km ²	12,214/km ²
Districts	23	24

3.1 Data Source

For data source, names of all districts in Wards of Tokyo and Osaka city are necessary. Below Wikipedia pages provide the data as a table structure. There are so many information and data on these pages but the process for scraping is simplified thanks to Python BeautifulSoup library.

Special Wards of Tokyo

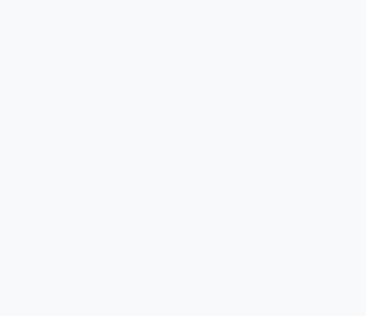
<https://en.wikipedia.org/wiki/Tokyo>

Special Wards of Tokyo			
	Place Name		Map of the Special Wards
	Rōmaji	Kanji	
1	Adachi	足立区	Red
2	Arakawa	荒川区	Green
3	Bunkyo	文京区	Yellow
4	Chiyoda	千代田区	Orange
5	Chūō	中央区	Green
6	Edogawa	江戸川区	Green
7	Itabashi	板橋区	Yellow
8	Katsushika	葛飾区	Yellow
9	Kita	北区	Orange
10	Kōtō	江東区	Yellow
11	Meguro	目黒区	Orange
12	Minato	港区	Yellow



Osaka city

<https://en.wikipedia.org/wiki/Osaka>

	Name	Kanji	Population	Land area in km ²	Pop. density per km ²	Map of Osaka
1	Abeno-ku	阿倍野区	107,000	5.99	18,440	
2	Asahi-ku	旭区	90,854	6.32	14,376	
3	Chūō-ku	中央区	100,998	8.87	11,386	
4	Fukushima-ku	福島区	78,348	4.67	16,777	
5	Higashinari-ku	東成区	83,684	4.54	18,433	
6	Higashisumiyoshi-ku	東住吉区	126,704	9.75	12,995	
7	Higashiyodogawa-ku	東淀川区	176,943	13.27	13,334	

3.2 Data Processing Libraries

- Beautifulsoup
BeautifulSoup is a Python library for pulling data out of HTML and XML files. In this project, this library is used for scraping district table from Wikipedia website.
- Geocoding
Geocoding is the process of converting addresses into geographic coordinates of latitude and longitude. In this project, this library is used for converting each district name into latitude and longitude.
- Foursquare
Foursquare is a technology company that build a massive dataset of accurate location data. Specific type of venues or stores around a given location can be searched by Foursquare API.

4. Methodology

This chapter consists of 2 parts. In the first part Tokyo and Osaka are clustered into 3 segments individually based on venues information for each districts and see what kinds of characteristics we could get. In the last part, clustering is done with mixed data of Tokyo and Osaka and see how the result is different with the first part.

You can refer to full Python codes from below link.

https://github.com/TomonagaNogami/Coursera_Capstone/blob/master/Capstone_Project_Final_Week_2.ipynb.ipynb

4.1 Generate initial DataFrame

DataFrame has 3 columns (District, Latitude and Longitude), Tokyo has 23 rows and Osaka has 24 rows.

Below table shows first 5 lines of Tokyo and Osaka DataFrame.

	District	Latitude	Longitude
0	Adachi	35.774811	139.804537
1	Arakawa	35.736093	139.783403
2	Bunkyo	35.707595	139.752210
3	Chiyoda	35.693930	139.753711
4	Chūō	35.670572	139.771988

Tokyo

	District	Latitude	Longitude
0	Abeno-ku	34.638732	135.518467
1	Asahi-ku	34.721168	135.544269
2	Chūō-ku	34.681144	135.509884
3	Fukushima-ku	34.692308	135.472220
4	Higashinari-ku	34.669951	135.541270

Osaka

4.2 Plot DataFrame to map

With DataFrame defined in 4.1, we can plot districts to the map. Same scale is applied to both maps and area in Tokyo is about 3 times larger than Osaka.



Tokyo



Osaka

4.3 Get venues information

Explore and get venues around districts. Search for venues within a radius of 500 meters from district origin and maximum number of venues for one district is limited to 100. Below is the first 5 lines of Tokyo and Osaka venues.

Tokyo

	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Adachi	35.774811	139.804537	MEGA Don Quijote (MEGAドン・キホーテ 環七梅島店)	35.778288	139.804967	Discount Store
1	Adachi	35.774811	139.804537	Ikinari Steak (いきなり!ステーキ)	35.777730	139.802890	Steakhouse
2	Adachi	35.774811	139.804537	Nitori (ニトリ)	35.778200	139.802592	Furniture / Home Store
3	Adachi	35.774811	139.804537	Yoshinoya (吉野家)	35.773290	139.803560	Donburi Restaurant
4	Adachi	35.774811	139.804537	7-Eleven (セブンイレブン 足立梅島1丁目店)	35.771822	139.803176	Convenience Store

Osaka

	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Abeno-ku	34.638732	135.518467	FamilyMart (ファミリーマート 松崎町店)	34.639983	135.517041	Convenience Store
1	Abeno-ku	34.638732	135.518467	Royal Host (ロイヤルホスト 文の里店)	34.637175	135.517652	Restaurant
2	Abeno-ku	34.638732	135.518467	甘辛や	34.638189	135.522473	Okonomiyaki Restaurant
3	Abeno-ku	34.638732	135.518467	FamilyMart (ファミリーマート 文の里二丁目店)	34.636626	135.517734	Convenience Store
4	Abeno-ku	34.638732	135.518467	Gusto (ガスト 文の里店)	34.639614	135.518109	Restaurant

4.4 Check how many venues we got

The population density ratio between Tokyo and Osaka is 4 : 3 (15,392/km² : 12,214/km²) whereas the number of venues ratio is 2 : 1 (1523 : 774).

Moreover, the number of venue is reached the maximum of 100 in some districts in Tokyo, which means data shows how Tokyo's venues are dense.

District	
Adachi	21
Arakawa	15
Bunkyo	100
Chiyoda	100
Chūō	100
Edogawa	33
Itabashi	41
Katsushika	29
Kita	35
Kōtō	72
Meguro	90
Minato	90
Nakano	100
Nerima	56
Setagaya	39
Shibuya	100
Shinagawa	56
Shinjuku	100
Suginami	31
Sumida	74
Taitō	76
Toshima	93
Ōta	72

Tokyo (1523)

District	
Abeno-ku	22
Asahi-ku	15
Chūō-ku	64
Fukushima-ku	50
Higashinari-ku	30
Higashisumiyoshi-ku	16
Higashiyodogawa-ku	46
Hirano-ku	20
Ikuno-ku	11
Jōtō-ku	42
Kita-ku (administrative center)	73
Konohana-ku	11
Minato-ku	27
Miyakojima-ku	26
Naniwa-ku	54
Nishi-ku	46
Nishinari-ku	36
Nishiyodogawa-ku	22
Suminoe-ku	17
Sumiyoshi-ku	17
Taishō-ku	23
Tennōji-ku	26
Tsurumi-ku	19
Yodogawa-ku	61

Osaka (774)

4.5 Convert DataFrame

Venue category is a good feature for applying clustering algorithm but machine learning can't handle text information. Strings needs to be converted to numerical number.

Below is converted DataFrame of Tokyo, each column shows unique venue category name and value is appearance rate of the venue category.

	District	ATM	Accessories Store	African Restaurant	American Restaurant	Aquarium	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Automot s
0	Adachi	0.00	0.0	0.0	0.00	0.0	0.00	0.00	0.00	0.00	0.000000	0.00	
1	Arakawa	0.00	0.0	0.0	0.00	0.0	0.00	0.00	0.00	0.00	0.066667	0.00	
2	Bunkyo	0.00	0.0	0.0	0.01	0.0	0.01	0.01	0.00	0.00	0.000000	0.01	
3	Chiyoda	0.01	0.0	0.0	0.00	0.0	0.00	0.00	0.02	0.00	0.000000	0.00	
4	Chūō	0.00	0.0	0.0	0.00	0.0	0.00	0.00	0.00	0.01	0.000000	0.00	

5 rows × 214 columns

4.6 Frequency of venue appearance

Let's see top 5 venue categories for some districts.

In the next page, I'll summarize data and see if there are any features and differences.

```
----Shibuya----
      venue  freq
0      Café  0.10
1  Record Shop  0.07
2   Nightclub  0.05
3    Sake Bar  0.04
4   Rock Club  0.04
```

```
----Shinagawa----
      venue  freq
0  Convenience Store  0.11
1 Japanese Restaurant  0.09
2  Donburi Restaurant  0.07
3   Ramen Restaurant  0.05
4           Theater  0.04
```

```
----Shinjuku----
      venue  freq
0    Sake Bar  0.09
1   BBQ Joint  0.07
2  Ramen Restaurant  0.07
3         Bar  0.05
4   Rock Club  0.03
```

Tokyo

```
----Ikuno-ku----
      venue  freq
0  Grocery Store  0.18
1  Shopping Mall  0.09
2    Bath House  0.09
3  Convenience Store  0.09
4   Udon Restaurant  0.09
```

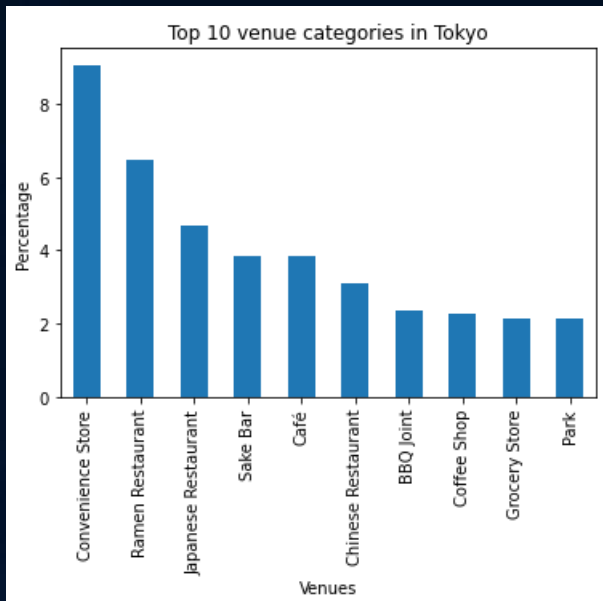
```
----Jōtō-ku----
      venue  freq
0  Convenience Store  0.17
1  Ramen Restaurant  0.10
2    Restaurant  0.07
3    BBQ Joint  0.07
4   Coffee Shop  0.07
```

```
----Kita-ku (administrative center)----
      venue  freq
0   Coffee Shop  0.12
1  Convenience Store  0.11
2   Train Station  0.08
3        Hotel  0.07
4        Café  0.04
```

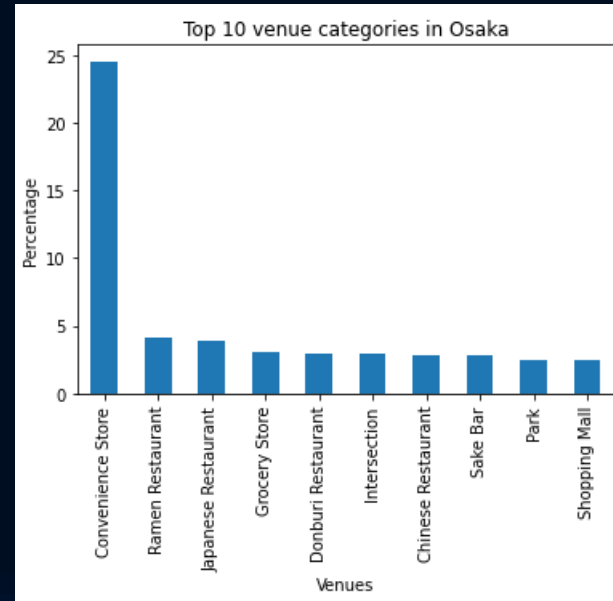
Osaka

4.7 Japanese culture?

It looks like Japanese cities are occupied by a convenience stores, especially in Osaka city (about 25% is convenience stores!). Another interesting characteristic is 2nd place of Ramen restaurant for both Tokyo and Osaka. Believe me, Japanese Ramen is amazing.



Tokyo



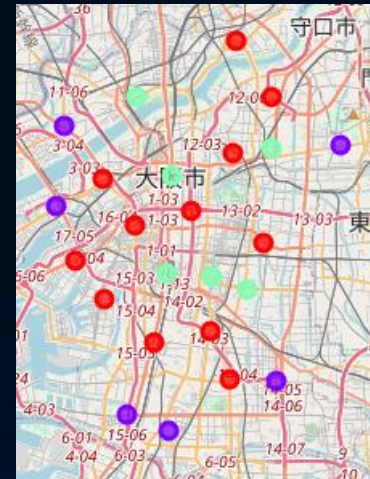
Osaka

4.8 Clustering and mapping

Run k-means clustering with DataFrame made in chapter 4.5 and plot to the map. There's a trend that purple mark shows outer districts, but we are not sure the relation between 2 cities as individual data set is used for clustering.



Tokyo



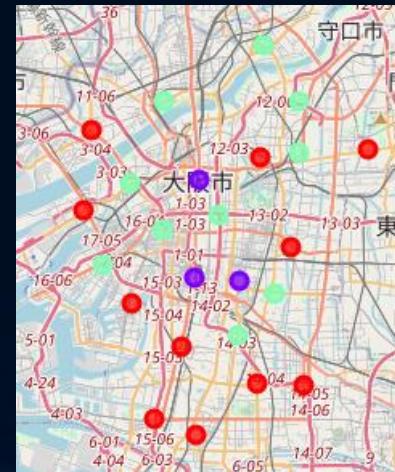
Osaka

4.9 Clustering with mixed DataFrame

This time k-means clustering will be applied to new DataFrame with mixed data of Tokyo and Osaka. Interesting point is that 2 major districts in Osaka (Kita and Naniwa, purple mark) are clustered as same as most part of Tokyo.



Tokyo



Osaka

5. Results and Discussion

- Venues in Tokyo is much dense than Osaka. Venue density ratio of Tokyo and Osaka is much higher than population density ratio. (4.4)
- Convenience store is everywhere, especially in Osaka. (4.7)
- In the result of individual data clustering, k-means clearly separates inner and outer districts. (4.8)
- In the result of mixed data clustering, two representative districts of Osaka are classified in the same cluster as the most of districts in Tokyo. (4.9)

6. Conclusion

By utilizing data science and machine learning, we were able to extract features that are difficult to notice from a human perspective.

This time the experiment is done on the subject of Japanese cities but data science could be applied for every field. I hope that more convenient and advanced services will be provided by combining human experience and sensibility with data science.