Top cities mapping and clustering

Qiaozhou Xiong

1. Introduction

The world is experiencing the biggest urbanization in history. Over 65 million people move to city every year, equivalent to adding 3 new Beijing. And now it's the first time in history the number of people living in urban area larger than that in rural. Cities make the modern society more productive and concentrated. Cities are playing the most important role in modern life.

Since cities are still growing, it's important for stakeholders to know the current top city status and predict the developing city's future and finding business opportunity there. This project will benefit the commercial activities in three aspects. First, similarity and dissimilarity of the top cities would be explored for different countries. Second, this study provides the top components of the top cities. Top cities "today" would be the "tomorrow" of emerging but small cities. These most common facilities or services would be in need for the city development, where business opportunity lies. Last, these cities would be clustered into 5 types based on their GDP, People, most common venues, etc. These similarities could be meaningful to the market exploration and business management or operation.

In this project, I grabbed the basic information of the top cities in the world which contributes to the GDP most or carrying the most people from the website. Meanwhile, the venues information of each city and explore the top 10 venues in each city were from Foursquare. The two data source willed merged together after data cleaning.

2. Data cleaning

Data source:

Since the top GDP cities was found in Wikipedia while the population and area could be found in another Wikipedia webpage. Before we enquire the top venues in each city through FourSquare API, we need to know the coordinate of each city, which was acquired through Google API.

Data cleaning

Originally, I just merged the information provided by two webpages, which provide the GDP and population data, respectively. But I found there were only 87 cities have the population information. I employed another data source (population density webpage), and merge them together. Finally, I merged the city population and area information with the top GDP city table, only considering the cities has complete information. There we have around 43 cities in our list.

There were several problems with the data collection. Since the GDP data were not recorded in the same year, the period of the data recording was from 2007 to 2017. This might bring some errors to our calculation. Another, the population counting might be ambiguous. Some cites only count the people in the downtown, while some figures consider the whole population in the whole city area. And this population data should correspond to the GDP. While, most data just neglect this point. Hence, to ensure the authenticity, we used the data from Wikipedia, though the data there is limited.

3. Exploration data analysis

3.1 Top GPD cities exploration

After grabbing the data from Wikipedia, we can see the top five super cities which contributes the most GDP as shown in Table 1. Especially, the USA has two super cities, New York and Los Angeles. While, Tokyo contributed 1893B ranking the first. We can see the top 5 cities all belong to the most developed countries.

Table 1. Top 5 GDP cities

	city	country	gdp	gdp_ranking
0	Tokyo	Japan	1893.000	1
1	New York	United States	1717.712	2
2	Los Angeles	United States	1043.735	3
3	Seoul	South Korea	738.600	4
4	Paris	France	724.000	5

Grouping the top 253 cities by the country name and sorting them according to the GDP sum, we can see the top 10 countries has the largest urban GDP as shown in Table 2. The city economics make United States the greatest country in the world. Among the top 10 country, there are two countries belongs to the developing country, China and Brazil.

Table 2. Sorting the top cities according to its country

	country	cities	gdp_sum
44	United States	New York, Los Angeles, Chicago, Dallas-Fort Wo...	14113.838
7	China	Shanghai, Beijing, Guangzhou, Shenzhen, Chongq...	6268.800
23	Japan	Tokyo, Osaka–Kobe, Nagoya, Fukuoka–Kitakyushu	3127.000
4	Brazil	São Paulo, Rio de Janeiro, Brasília, Curitiba,...	1536.188
14	Germany	Rhine-Ruhr, Berlin, Munich, Stuttgart, Nürnber...	1366.600
43	United Kingdom	London, Manchester, Birmingham, Glasgow, Brist...	1274.800
12	France	Paris, Marseille, Lyon, Lille, Toulouse, Borde...	1201.300
22	Italy	Milan, Rome, Turin, Naples, Venice, Bologna, F...	1123.420
0	Australia	Sydney, Melbourne, Brisbane, Perth, Adelaide	823.800
38	South Korea	Seoul	738.600

We visualize the difference as shown in Figure 1, there we can see the USA cities' GDP sum might be equal to that of the 9 remaining counties.

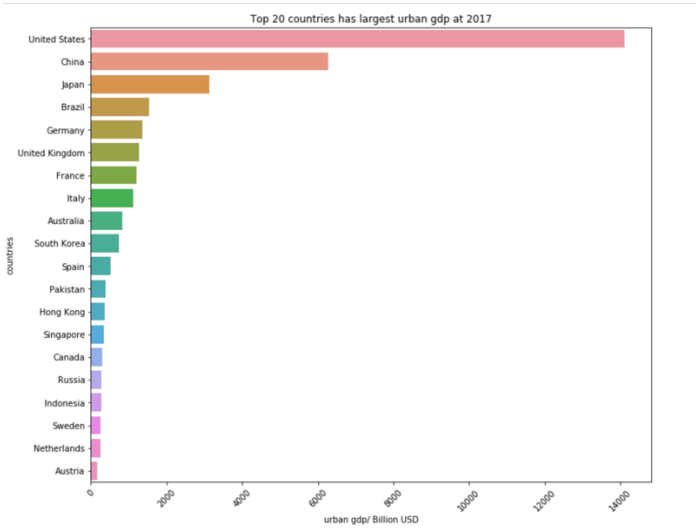

Figure 1. Top GDP cities grouping by countries

3.2 Top population city exploration

We could find the top 5 cities with the most populations. However, most of them belongs to the developing countries.

Table 3. Top population cities

	city	pop	area	popdst	pop_ranking
0	Chongqing	30165500.0	82403.0	366.072837	1
1	Shanghai	24183300.0	6340.5	3814.099834	2
2	Beijing	21707000.0	16411.0	1322.710377	3
3	Istanbul	15029231.0	5196.0	2892.461701	4
4	Karachi	14910352.0	3780.0	3944.537566	5

Another, if we divide the total GDP of each city by its population, we can know how much the people in each created in 2017 which is shown in *Figure 2*. The top 10 cities belongs to the most developed countries.

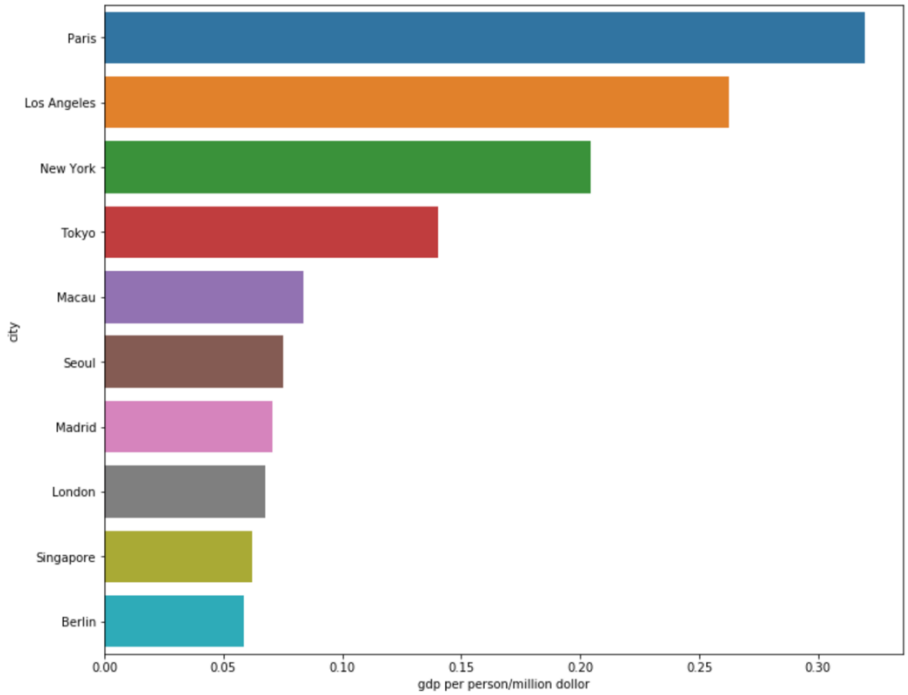

Figure 2. Top 10 cities with most GDP per person

In sum, we can conclude much more people lives in the top cities in the developing countries. While high GDP per person can be an indicator to differentiate a developed country. In other

words, the emerging city has the advantage of population, while the developed city has the advantage of higher efficiency.

3.3 Top common venues exploration

Exploring the nearby venues within 10 km of the city center with Four Square API, we can count the frequency of each category (part of the results is shown as Table 4).

Table 4. Top 10 most common venues in each city

	city	The 1st common venue	The 2nd common venue	The 3rd common venue	The 4th common venue	The 5th common venue	The 6th common venue	The 7th common venue	The 8th common venue	The 9th common venue	The 10th common venue
0	Bandung	Hotel	Bakery	Coffee Shop	Café	Snack Place	Sundanese Restaurant	Indonesian Restaurant	Multiplex	Sushi Restaurant	Shopping Mall
1	Beijing	Historic Site	Hotel	Park	Peking Duck Restaurant	Café	Dumpling Restaurant	Chinese Restaurant	Coffee Shop	Yunnan Restaurant	Beijing Restaurant
2	Berlin	Coffee Shop	Park	Bookstore	Concert Hall	Sandwich Place	Ice Cream Shop	Vegetarian / Vegan Restaurant	Wine Bar	Science Museum	Bakery
3	Changsha	Coffee Shop	Hotel	Shopping Mall	Park	Chinese Restaurant	Fast Food Restaurant	Historic Site	Multiplex	Café	Outdoor Sculpture
4	Chengdu	Hotel	Shopping Mall	Coffee Shop	Fast Food Restaurant	Hostel	Furniture / Home Store	Café	Noodle House	Bar	History Museum

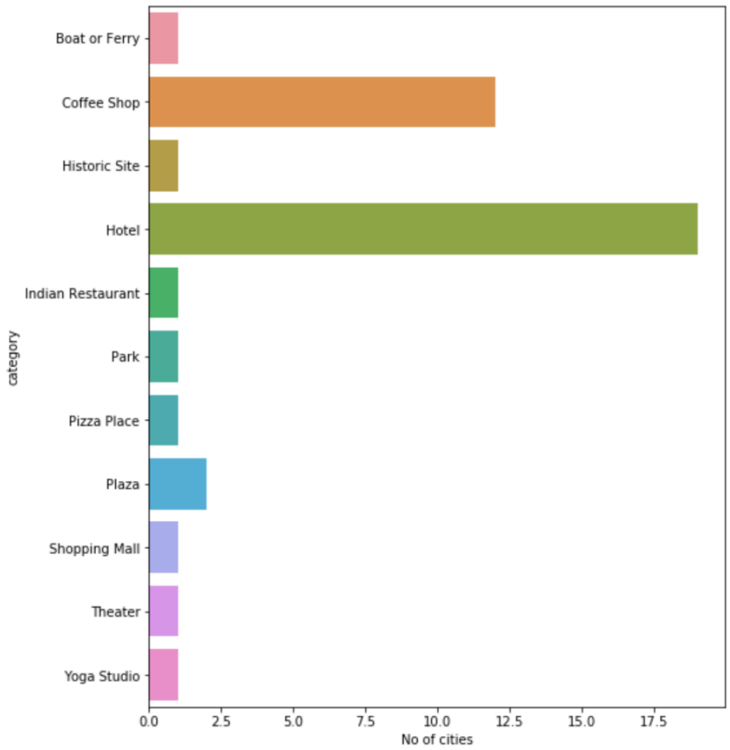

Figure 3. the top venues in all the cities.

First, let's count the frequency for the most common venues in all the cities, and the results are shown in **Figure 3**. Hotel was the most common venues in all the cities, followed by coffee shop. Further investigation reveals that those cities with hotel as the most common venue are all Asian cities except for London, which might indicate these cities has quite high population density.

After expanding the list of the common venues into top 10, we can find the 34 from 43 cities has hotel as the top 10 common venues. Still, hotel and coffee shop have the top 2 occurrences in these cities. We could also conclude that these ten venues are the basic elements for a top city. For fast developing city, these categories or facility construction would bring a lot of business opportunity.

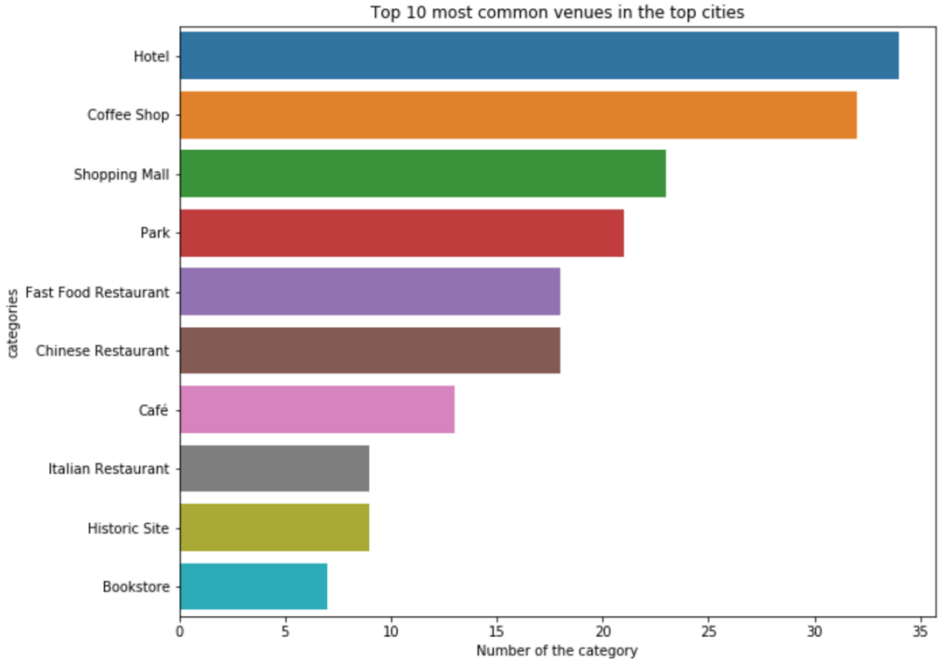

Figure 4. Top 10 common venues in these cities

4. Cities clustering

After integrating the data from Wikipedia and Four-Square API, the data is shown as Table 5. We can get the GDP data, area data, as well as their ranking.

After adding the top venues of each cities into consideration, we now already grab the basic information and characteristic of each city. Clustering will be performed based on their GDP, population, area, and top venues.

Table 5. The final information for the top cities

	city	country	gdp	gdp_ranking	pop	area	popdst	pop_ranking	gdp_pp	gdp_pa
0	Tokyo	Japan	1893.000	1.0	13515271.0	626.99	21555.799933	7	0.140064	3.019187
1	New York	United States	1717.712	2.0	8398748.0	786.30	10681.353173	30	0.204520	2.184550
2	Los Angeles	United States	1043.735	3.0	3976322.0	1213.85	3275.793549	68	0.262488	0.859855
3	Seoul	South Korea	738.600	4.0	9806000.0	605.25	16201.569599	22	0.075321	1.220322
4	Paris	France	724.000	5.0	2265886.0	105.40	21497.969639	95	0.319522	6.869070

The clustering results are shown in Figure 5. We set the cluster number as 5 and standardized each data to ensure no bias on each data. We can find We can find New York, Beijing, Chongqing are quite special cities. And Los Angeles and most cities in western are more similar than New York. I guess population and area make Chongqing special in this clustering, which was encoded in red color on the map.

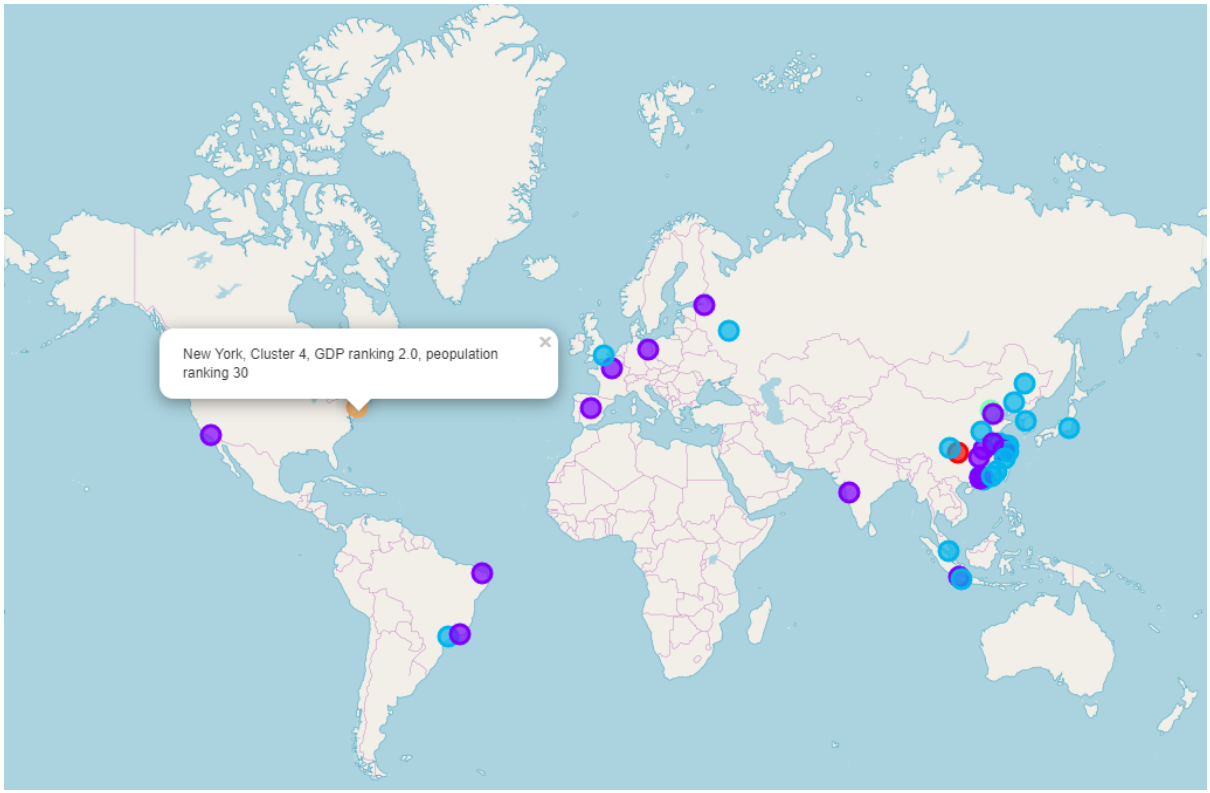

Figure 5. cities' clustering results on the map

5. Discussion and conclusion

In this project, I grabbed the list of the top cities in the world with the highest GDP and population, respectively. And analyzed them according to the countries and GDP per person. Also, I explored the top venues in each city within 10 near the city center. While we only have

43 cities has the complete information, due to the shortage of authentic source on the web. More features can be added into this study, which involves more about the underlying characteristic of the city.

We found following interesting phenomenon. Most cities with largest populations come from Asia emerging cities or other developing cities. While most top cities from developed countries has the very high GDP per person. Asian top cities tend to have more hotel than other cities. Hotel, coffee shop, shopping mall, fast food restaurant are the top facilities in top cities, which should be the investment opportunity in those developing cities. Last, the city clustering results might be helpful for the management decision making or business operation, while more in-depth investigations are yet to make.

6. Reference

[1] https://en.wikipedia.org/wiki/List_of_cities_by_GDP

[2] https://en.wikipedia.org/wiki/List_of_cities_proper_by_population

[3] https://en.wikipedia.org/wiki/List_of_cities_by_population_density