Similarity the capital city in the world

Introduction

There are about 244 countries in the world. The capital city of those countries is all the same as they are the heart of the country. However, each city has its characteristics. The question is can we do clustering for those capital cities? Can we group those capital cities?

If we can group similar cities to each other, a business model which is a success in a certain city might also be success if we implement it in the similar city in the other country
Characteristic of the city should be able to be identified by using similarity of the venues located in the center of the city. With this assumption, if we use the data containing the category of venues nearby the center of the city that we can get it from Foursquare location data, we should be able to do clustering for the cities. Also, can we use population data to make data visualization for some information display?

Venues which are located nearby the center of the capital city might reveal the character of the city. After that comparing individual character of many cities and clustering them can be done.

Not only nearby venues information, capital city location, and population, there are also more information related to each country that we can utilize to find our similarity of the cities. Combining all the information we might be able to make interesting visualization to present the finding.

Clustering the capital cities we might see group of country which are similar to each other inside the cluster. Again, any business or technology that is currently doing well in any particular country might also be doing well in other cites which is similar to the based location.

Data section

To solve the problem mentioned in the introduction session, we need to have the data set which includes the data of country name, capital name, population, coordinates of the capital cities. Luckily, there are available data on the internet but not all the required data are available in a single data source. Therefore, we need to pull the data from a couple of data sources and process it to be ready for solving our business problem.

<u>Country name, capital name, and population from wikipedia.org</u>

https://en.wikipedia.org/wiki/List_of_national_capitals_by_population

List [edit]

Rank	Country/Territory	Capital	Population	Year	% of country's population
-	*Hong Kong*	Hong Kong	7,482,500[14]	2018	100%
1	China PR	Beijing	21,542,000[1]	2010	1.5%
2	Japan	Tokyo	13,929,286[2]	2017	11.03%
3	Russia	Moscow	12,506,468[3]	2011	8.52%
4	DR Congo	Kinshasa	11,855,000[4]	2012	12.9%
5	Indonesia	Jakarta	10,075,310[5]	2011	3.76%
6	South Korea	Seoul	9,838,892[6]	2015	19.03%
7	Egypt	Cairo	9,500,000	2012	9.54%
8	Mexico	Mexico City	8,918,653[7]	2015	7.05%
9	Bangladesh	Dhaka	8,906,039 [8]	2011	5.52%
10	United Kingdom *England*	London	8,908,081[9]	2015	13.19%
11	Peru	Lima	8,852,000[10]	2012	26.74%
12	Iran	Tehran	8,693,706	2014	10.53%
13	Thailand	Bangkok	8,305,218[11]	2010	11.96%
14	Vietnam	Hanoi	7,781,631[12]	2014	8.14%

<u>Country name, capital name and its coordinates are available in provided webpage.</u>

https://www.jasom.net/list-of-capital-cities-with-latitude-and-longitude/

```
Country,Capital,Latitude,Longitude
Abkhazia,Sukhumi,43.001525,41.023415
Afghanistan,Kabul,34.575503,69.240073
Aland Islands,Mariehamn,60.1,19.933333
Albania,Tirana,41.327546,19.818698
Algeria,Algiers,36.752887,3.042048
American Samoa,Pago Pago,-14.275632,-170.702036
Andorra,Andorra la Vella,42.506317,1.521835
Angola,Luanda,-8.839988,13.289437
Anguilla,The Valley,18.214813,-63.057441
Antarctica,South Pole,-90,0
Antigua and Barbuda,St. John's,17.12741,-61.846772
Argentina,Buenos Aires,-34.603684,-58.381559
Armenia,Yerevan,40.179186,44.499103
Aruba,Oranjestad,12.509204,-70.008631
Australia,Canberra,-35.282,149.128684
```

Also, not only country name, capital city population, its coordinates, there are other data that can be utilized.

Population of the whole countries

The data is quite up to date, we get the population of each country in 2019

https://en.wikipedia.org/wiki/List_of_countries_by_population_(United_Nations)#cite_ref-16

Countries and areas ranked by population in 2019

	Country or area	UN continental region[4]	UN statistical region[4]	Population (1 July 2018)	Population (1 July 2019)	Change
1	China[a]	Asia	Eastern Asia	1,427,647,786	1,433,783,686	+0.43%
2	India	Asia	Southern Asia	1,352,642,280	1,366,417,754	+1.02%
3	United States	Americas	Northern America	327,096,265	329,064,917	+0.60%
4	Indonesia	Asia	South-eastern Asia	267,670,543	270,625,568	+1.10%
5	Pakistan	Asia	Southern Asia	212,228,286	216,565,318	+2.04%
6	Brazil	Americas	South America	209,469,323	211,049,527	+0.75%
7	Nigeria	Africa	Western Africa	195,874,683	200,963,599	+2.60%
8	Bangladesh	Asia	Southern Asia	161,376,708	163,046,161	+1.03%
9	Russia	Europe	Eastern Europe	145,734,038	145,872,256	+0.09%
10	Mexico	Americas	Central America	126,190,788	127,575,529	+1.10%
11	Japan	Asia	Eastern Asia	127,202,192	126,860,301	−0.27%
12	Ethiopia	Africa	Eastern Africa	109,224,414	112,078,730	+2.61%
13	Philippines	Asia	South-eastern Asia	106,651,394	108,116,615	+1.37%
14	Egypt	Africa	Northern Africa	98,423,598	100,388,073	+2.00%
15	Vietnam	Asia	South-eastern Asia	95,545,962	96,462,106	+0.96%

GDP per capita of each country is also up to date

https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)_per_capita

List of per capita nominal GDP for countries and dependencies [edit]

International Monetary Fund (2019)[3]			World Bank (2018)[4]			United Nations (2017)[5]		
Rank	Country/Territory	US$	Rank	Country/Territory	US$	Rank	Country/Territory	US$
1	Luxembourg	113,196	1	Luxembourg	114,341	1	Liechtenstein	166,022
2	Switzerland	83,716	—	*Macau*	86,365	2	Monaco	165,421
—	*Macau*	81,151	2	Switzerland	82,839	3	Luxembourg	106,806
3	Norway	77,975	3	Norway	81,807	—	*Bermuda*	102,192
4	Ireland	77,771	4	Ireland	77,450	—	*Macau*	80,893
5	Qatar	69,687	5	Iceland	73,191	4	Switzerland	80,101
6	Iceland	67,037	6	Qatar	69,027	5	Norway	75,295
7	United States	65,111	7	Singapore	64,582	6	Iceland	73,060
8	Singapore	63,987	8	United States	62,641	7	Ireland	69,604
9	Denmark	59,795	9	Denmark	60,596	—	*Cayman Islands*	65,472
10	Australia	53,825	10	Australia	57,305	8	Qatar	63,506
11	Netherlands	52,367	11	Sweden	54,112	9	United States	60,055
12	Sweden	51,241	12	Netherlands	52,978	10	Australia	57,613
13	Austria	50,022	13	Austria	51,513	11	Denmark	57,533
—	*Hong Kong*	49,334	14	Finland	49,960	12	Singapore	56,737
14	Finland	48,868	—	*Hong Kong*	48,717	13	Sweden	54,043
15	San Marino	47,279	15	Germany	48,196	14	San Marino	50,588

With the accessible data sources, we can process the data and get the answer to the business problem. After all data are processed, we can make data visualization to report the project.

Clustering the cities

First, we begin clustering a similar city by scraping data from the website. Start with the population of each capital city. There are 244 countries in the database.

```
Number of country and its capital city from the database = 244
```

	Rank	Country/Territory	Capital	Population	Year	% ofcountry'spopulation
0	1	China	Beijing	21542000	2010	1.5%
1	2	Japan	Tokyo	13929286	2017	11.03%
2	3	Russia	Moscow	12506468	2011	8.52%
3	4	DR Congo	Kinshasa	11855000	2012	12.9%
4	5	Indonesia	Jakarta	10075310	2011	3.76%

Top 5 rows

Next, grabbing the capital cities' population data and cleaning the dataset, we start scraping capital cities' coordinates.

```
Number of country, its capital and its coordinate from the database = 252
```

	Country	Capital	Latitude	Longitude
0	Abkhazia	Sukhumi	43.001525	41.023415
1	Afghanistan	Kabul	34.575503	69.240073
2	Aland Islands	Mariehamn	60.1	19.933333
3	Albania	Tirana	41.327546	19.818698
4	Algeria	Algiers	36.752887	3.042048

Top 5 rows

Since the number of the coordinate data is larger than the population dataset, we merge the data based on the population dataset. After merging the two datasets we found that there are some data errors and missing.

	Country	Capital	Population	Latitude	Longitude
0	China	Beijing	21542000	39.904211	116.407395
1	Japan	Tokyo	13929286	35.709026	139.731992
2	Russia	Moscow	12506468	55.755826	37.6173
3	DR Congo	Kinshasa	11855000	NaN	NaN
4	Indonesia	Jakarta	10075310	-6.208763	106.845599

Top 5 rows

Some data errors

Therefore, data cleaning and correction process is needed, and after the cleaning is done we can get the cleaned dataset.

```
Number of country after data cleaing = 244 Countries
Remaining Dataset with missing data/value =  0 Case
```

Top 5 rows

	Country	Capital	Population	Latitude	Longitude
0	China	Beijing	21542000	39.904211	116.407395
1	Japan	Tokyo	13929286	35.709026	139.731992
2	Russia	Moscow	12506468	55.755826	37.6173
3	DR Congo	Kinshasa	11855000	-4.441931	15.266293
4	Indonesia	Jakarta	10075310	-6.208763	106.845599

Finally, we use the foursquare function to get information about venues which are located nearby the center of each capital city in the radius of 1,500 meters (limit 100 places per city) and the model shows that there are over 11,000 venues located nearby the cities with 510 unique venue categories.

```
There are 510 uniques categories.
Number of total venues nearby all of the capital cities =  11346 Places
```

Top 5 rows

	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Beijing	39.904211	116.407395	Grand Hyatt Beijing (北京东方君悦大酒店)	39.908060	116.408349	Hotel
1	Beijing	39.904211	116.407395	New World Beijing Hotel (北京新世界酒店)	39.897296	116.405923	Hotel
2	Beijing	39.904211	116.407395	WF Central (王府中环)	39.910960	116.404725	Shopping Mall
3	Beijing	39.904211	116.407395	吴裕泰 Wu Yu Tai Tea Shop	39.910257	116.405064	Tea Room
4	Beijing	39.904211	116.407395	Din Tai Fung (鼎泰丰)	39.913630	116.405766	Dumpling Restaurant

Next step, we create the new dataset and display the top 10 venues for each capital city and at this step, we found that 232 cities out of 244 cities have Foursquare data.

```
Number of city with have Foursqare data =  232 City
```

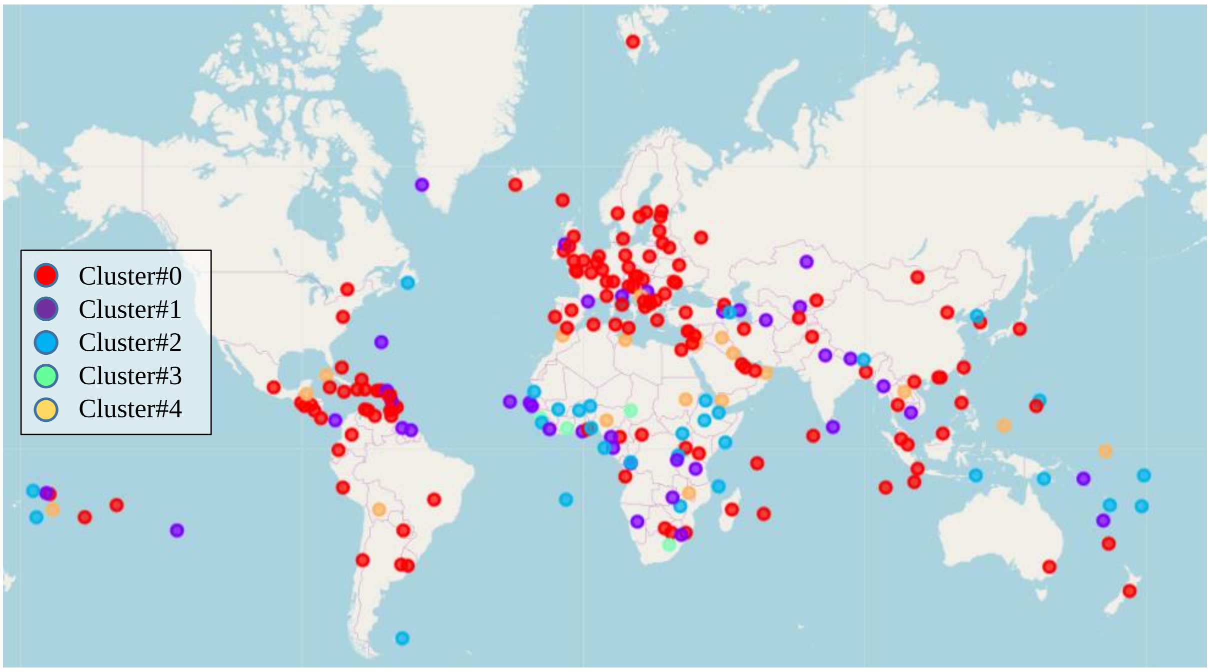
Top 5 rows

	Capital	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	Abu Dhabi	Currency Exchange	Tourist Information Center	Asian Restaurant	Music Venue	Fish Taverna	Fish Market	Fish & Chips Shop	Filipino Restaurant	Festival	Zoo Exhibit
1	Abuja	Flea Market	Lounge	African Restaurant	Café	Festival	Fish Taverna	Fish Market	Fish & Chips Shop	Filipino Restaurant	Zoo Exhibit
2	Accra	Hotel	Restaurant	African Restaurant	Hotel Pool	Chinese Restaurant	Nightclub	Pizza Place	Café	Soccer Field	Middle Eastern Restaurant
3	Adamstown	Nature Preserve	Zoo Exhibit	Flower Shop	Event Service	Event Space	Exhibit	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant
4	Addis Ababa	Hotel	Restaurant	Fried Chicken Joint	Greek Restaurant	Nightclub	Friterie	Frozen Yogurt Shop	Ethiopian Restaurant	Event Service	Event Space

We cluster the cities into 5 clusters by using Kmeans method.

Cluster Labels	Country	Capital	Population	Latitude	Longitude
0	142	142	142	142	142
1	38	38	38	38	38
2	31	31	31	31	31
3	4	4	4	4	4
4	19	19	19	19	19

With the clustering, we found that the majority number of the capital cities in the world are clustered into cluster#0.



By examining the cluster#0 we found that most of the big country, developed and developing countries are clustered into cluster#0

	Country	Capital	Cluster Labels
0	China	Beijing	0
1	Japan	Tokyo	0
2	Russia	Moscow	0
4	Indonesia	Jakarta	0
5	South Korea	Seoul	0
6	Egypt	Cairo	0
7	Mexico	Mexico City	0
8	United Kingdom England	London	0
9	Bangladesh	Dhaka	0
10	Peru	Lima	0
11	Iran	Tehran	0
12	Thailand	Bangkok	0
15	Hong Kong	Hong Kong	0
16	Colombia	Bogotá	0
18	Chile	Santiago	0
19	Singapore	Singapore	0
20	Turkey	Ankara	0
21	Germany	Berlin	0
23	Algeria	Algiers	0
24	Spain	Madrid	0

By looking into the 10 most common venues we can see that the venues located near the center of the cities of this cluster are the most likely tourist attraction, restaurants, and hotels.

Cluster Labels	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	Hotel	Coffee Shop	Chinese Restaurant	Peking Duck Restaurant	Clothing Store	Historic Site	Shopping Mall	Beijing Restaurant	Pizza Place	Hotel Bar
0	Ramen Restaurant	Italian Restaurant	Sake Bar	Japanese Restaurant	Bakery	BBQ Joint	French Restaurant	Café	Chinese Restaurant	Dessert Shop
0	Plaza	Hotel	Concert Hall	Coffee Shop	Theater	Boutique	History Museum	Café	Clothing Store	Asian Restaurant
0	Coffee Shop	Indonesian Restaurant	Hotel	Food Truck	Pizza Place	Bakery	Asian Restaurant	Café	Multiplex	Park
0	Hotel	Korean Restaurant	Café	Chinese Restaurant	Coffee Shop	Bookstore	Lounge	Japanese Restaurant	Plaza	Historic Site

Top 5 rows

Next, we scope down and re-do the clustering process for the cities which are clustered in the cluster#0 only. There are 142 cities to be clustered.

```
Number of city to be clustered =  142 City
```

	Country	Capital	Population	Latitude	Longitude
0	China	Beijing	21542000	39.904211	116.407395
1	Japan	Tokyo	13929286	35.709026	139.731992
2	Russia	Moscow	12506468	55.755826	37.6173
4	Indonesia	Jakarta	10075310	-6.208763	106.845599
5	South Korea	Seoul	9838892	37.566535	126.977969

Top 5 rows

With the Foursquare data about nearby venues, even the number of cities reduce from 232 cities to 142 cities (reduce 38 %), the number of nearby venues reduces only by 15% to be 9,623 places.

```
There are 490 uniques categories.
Number of total venues nearby all of the capital cities =  9623 Places
```

Top 5 rows

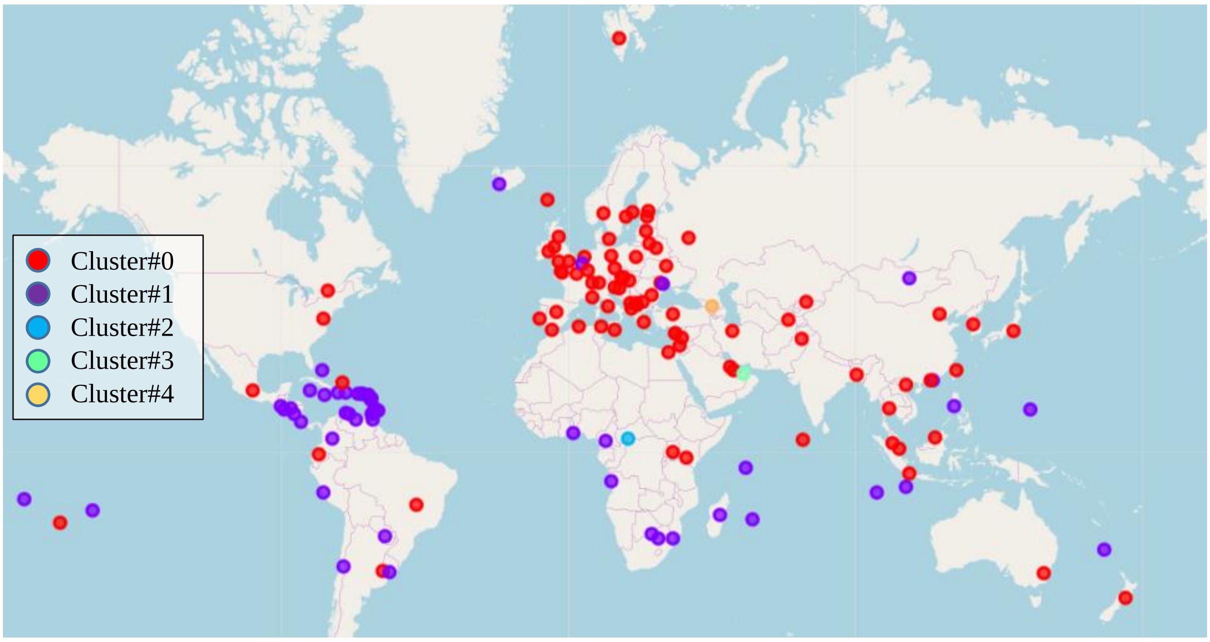
	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Beijing	39.904211	116.407395	Grand Hyatt Beijing (北京东方君悦大酒店)	39.908060	116.408349	Hotel
1	Beijing	39.904211	116.407395	New World Beijing Hotel (北京新世界酒店)	39.897296	116.405923	Hotel
2	Beijing	39.904211	116.407395	WF Central (王府中环)	39.910960	116.404725	Shopping Mall
3	Beijing	39.904211	116.407395	吴裕泰 Wu Yu Tai Tea Shop	39.910257	116.405064	Tea Room
4	Beijing	39.904211	116.407395	Din Tai Fung (鼎泰丰)	39.913630	116.405766	Dumpling Restaurant

Again, we cluster the cities into 5 clusters by using Kmeans method.

Cluster Labels	Country	Capital	Population	Latitude	Longitude
0	86	86	86	86	86
1	53	53	53	53	53
2	1	1	1	1	1
3	1	1	1	1	1
4	1	1	1	1	1

This time, clustering result shows that only cluster#0 and cluster#1 contains most of the cities

The result of the 2nd clustering.



By examining the clusters reveal that most of the developing and the developed countries are clustered into cluster#0.

By looking into the 10 most common venues of both cluster#0 and cluster#1, we can see the difference between the two clusters. It seems like nearby the center of the cities of cluster#1 are more like residential area. There are a lot of parks, restaurants, and public places; on the other hand, for cluster#0, common venues are tourist attractions, restaurants, business areas, and hotels.

<u>10 most common venues of cluster#0</u>

Top 5 rows

Cluster Labels	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	Hotel	Coffee Shop	Chinese Restaurant	Peking Duck Restaurant	Shopping Mall	Clothing Store	Historic Site	Pizza Place	Beijing Restaurant	Department Store
0	Ramen Restaurant	Japanese Restaurant	Italian Restaurant	Sake Bar	Bakery	French Restaurant	Café	BBQ Joint	Dessert Shop	Chinese Restaurant
0	Plaza	Hotel	Theater	Coffee Shop	Concert Hall	Boutique	History Museum	Clothing Store	Café	Russian Restaurant
0	Indonesian Restaurant	Coffee Shop	Hotel	Food Truck	Multiplex	Café	Bakery	Pizza Place	Park	Fast Food Restaurant
0	Hotel	Korean Restaurant	Café	Chinese Restaurant	Bookstore	Coffee Shop	Plaza	Lounge	Japanese Restaurant	Historic Site

<u>10 most common venues of cluster#1</u>

Top 5 rows

Cluster Labels	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
1	Bar	Restaurant	Plaza	Peruvian Restaurant	Sandwich Place	Bakery	Seafood Restaurant	Hotel Bar	Hotel	Beer Bar
1	Recreation Center	Mountain	Tourist Information Center	Park	Chinese Restaurant	Farm	Dim Sum Restaurant	Hotel	Bus Stop	Clothing Store
1	Pizza Place	BBQ Joint	Bakery	Sandwich Place	Coffee Shop	Restaurant	Steakhouse	Pub	Donut Shop	Candy Store
1	Coffee Shop	Pizza Place	Bar	Martial Arts Dojo	Pub	Hostel	Café	Restaurant	Arepa Restaurant	Asian Restaurant
1	Intersection	Fried Chicken Joint	Plaza	Basketball Stadium	Fast Food Restaurant	Falafel Restaurant	Festival	Farmers Market	Farm	Zoo

For the cluster#2 and #3, it seems like the cities have their own characteristic which is not similar to other cities.

Since this observation aims to find out cities' similarity and seek business opportunity so we decided to look more into the cities in cluster#0 of the 2nd clustering. So before going further, this is where the cities that we are considering, are located.

Before we try to do the clustering process for these cities again, we would like to add more useful information into the clustering model

First, we add the population of each country into the dataset so the clustering model will not only consider the population of the capital cities only but also the whole country population reflecting the bigger similarity model.

The country population data is grabbed from the available database (2019).

https://en.wikipedia.org/wiki/List_of_countries_by_population_(United_Nations)#cite_ref-16

	Country	Country Population
0	China	1433783686
1	India	1366417754
2	United States	329064917
3	Indonesia	270625568
4	Pakistan	216565318

Top 5 rows

We also add the information about GDP per capita into the considering dataset (2019).

	Rank	Country	GDP US$ per capita
0	1	Luxembourg	113196
1	2	Switzerland	83716
2	—	Macau	81151
3	3	Norway	77975
4	4	Ireland	77771

Top 5 rows

The GDP per capita information is also available on the Wikipedia webpage and the information was updated recently so we could get the GDP per capita of each country in 2019.

https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)_per_capita

International Monetary Fund (2019)[3]

Rank	Country/Territory	US$
1	Luxembourg	113,196
2	Switzerland	83,716
—	*Macau*	81,151
3	Norway	77,975
4	Ireland	77,771
5	Qatar	69,687
6	Iceland	67,037
7	United States	65,111
8	Singapore	63,987
9	Denmark	59,795
10	Australia	53,825

After we merge all the datasets into the single dataset and clean/correct error dataset, we can get the dataset for the final clustering process which contains of 69 countries to be clustered.

```
Number of city to be cosidered in the final clustering =  69 Cities
```

Top 5 rows

	Country	City	Population	Latitude	Longitude	Country Population	GDP US$ per capita
0	China	Beijing	21542000	39.904211	116.407395	1433783686	10098
1	Japan	Tokyo	13929286	35.709026	139.731992	126860301	40846
2	Russia	Moscow	12506468	55.755826	37.6173	145872256	11162
3	Indonesia	Jakarta	10075310	-6.208763	106.845599	270625568	4163
5	Egypt	Cairo	9500000	30.04442	31.235712	100388073	3046

For the final clustering step, we use all the information that we have which include nearby venues of each capital city, the population of each capital city, the population of each country, and GDP per capita of each country.

Importantly, before process the Kmeans method, we need to normalize the dataset first, due to a large number of population and GDP

```
Scaled number of population and GDP
```

	Capital	Population	Country Population	GDP US$ per capita
0	Beijing	1.000000	1.000000	0.082970
1	Tokyo	0.646505	0.088204	0.356466
2	Moscow	0.580437	0.101468	0.092434
3	Jakarta	0.467546	0.188504	0.030180
5	Cairo	0.440832	0.069735	0.020244

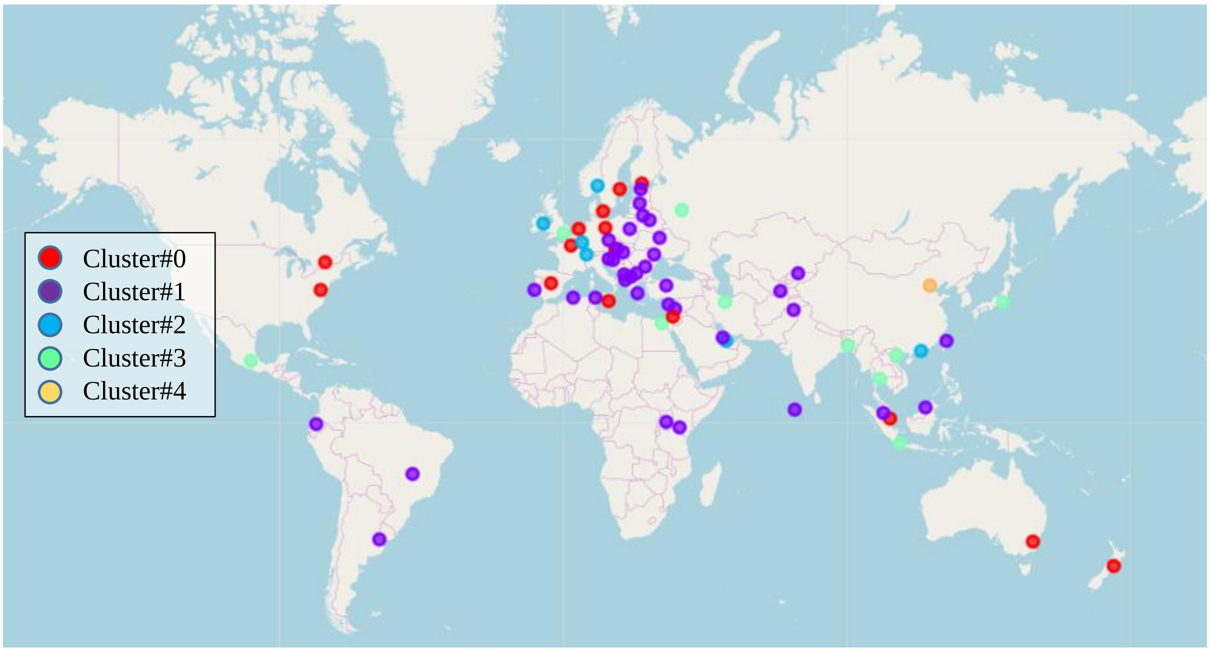
Top 5 rows

Finally, we get the final scaled dataset and we do the Kmeans method to cluster the cities into 5 clusters.

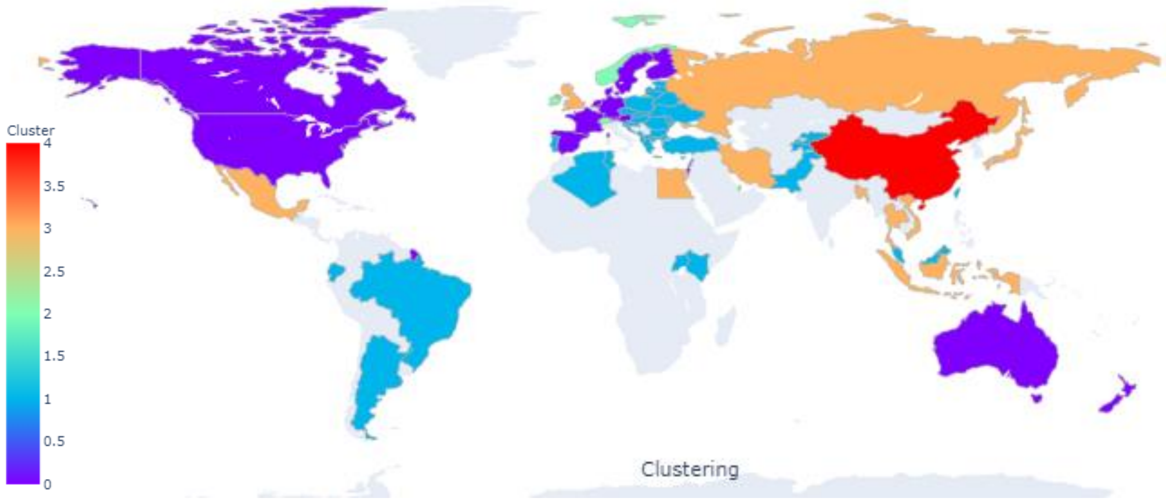
This is the result of the final clustering.

Cluster Labels	Country	Capital	Population	Latitude	Longitude	Country Population	GDP US$ per capita
0	15	15	15	15	15	15	15
1	37	37	37	37	37	37	37
2	6	6	6	6	6	6	6
3	10	10	10	10	10	10	10
4	1	1	1	1	1	1	1

And this is how the result looks like in the world map.

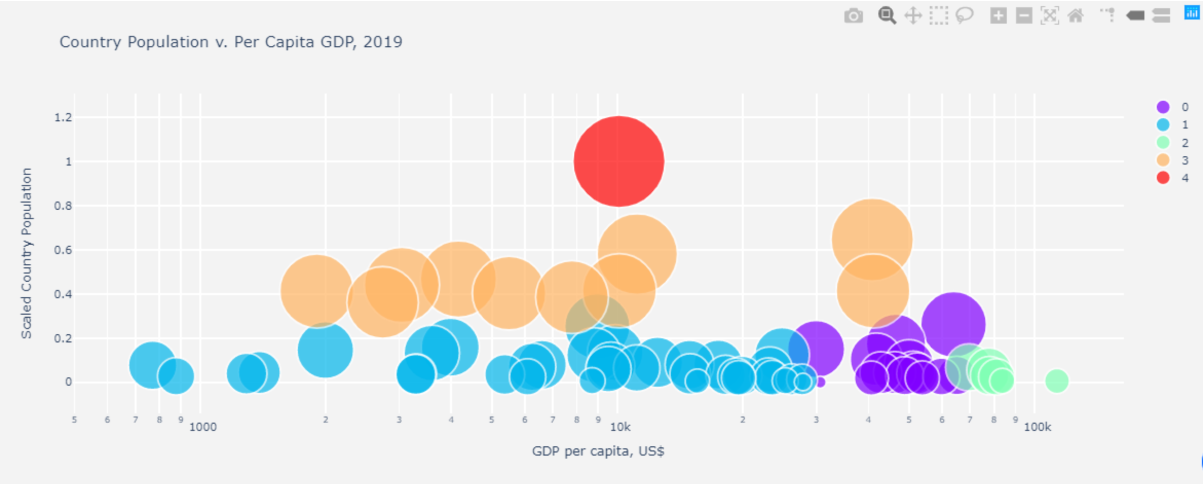


If we scale up the location of the considering city to its country, we can visualize the clustering result as below.



Conclusion

To examine the result of the clustering, we plot the result in a bubble chart with related attributes like the country population, the population of the capital cities, GDP per capita, and cluster labels.



Cluster#0: High GDP per capita with fewer number of country population while their capital city has the number of population in the range of low to medium.

- Singapore, Germany, Spain, France, Austria, Sweden, Canada, Israel, Netherlands, United States, Finland, Denmark, Australia, New Zealand, Malta

Cluster#1: Majority of the result, low to medium GDP per capita with less number of country population and their capital city also has the number of population in the range of low to medium.

- Turkey, Algeria, Kenya, Argentina, Ukraine, Taiwan, Brazil, Romania, Hungary, Poland, Belarus, Uganda, Lebanon, Ecuador, Malaysia, Bulgaria, Czech Republic, Pakistan, Kyrgyzstan, Croatia, Moldova, Tunisia, Albania, Latvia, Greece, Tajikistan, Portugal, Lithuania, North Macedonia, Estonia, Slovakia, Slovenia, Cyprus, Montenegro, Bahrain, Maldives, Brunei

Cluster#2: Highest GDP per capita with quite less number of country population and their capital city also has a small population.

- Qatar, Ireland, Norway, Macao, Switzerland, Luxembourg

Cluster#3: Medium to high GDP per capita with a high number of country population and their capital city also have a high number of population.

- Japan, Russia, Indonesia, Egypt, Mexico, United Kingdom, Bangladesh, Iran, Thailand, Vietnam

Cluster#4: This cluster contains only one country which is China. This city has its own characteristic, medium GDP per capita with a very high number of country populations and a very high population for the capital city which is Beijing.

- China