The Living Standard Analysis of the top 50 Cities in GDP

Ruixuan Han
March 10, 2021

1. Introduction

1.1 Background

With China's economic development, more and more people choose overseas investment and immigration. In recent years, due to the rapid increase in the number of Chinese immigration applicants, some popular countries have raised immigration thresholds and tightened immigration policies. As more and more countries launch projects to attract immigrants, the destinations of overseas Chinese immigrants will become more diversified, including some South American and African countries, which will gradually enter people's field of vision. The experience of living in many new countries and cities is very scarce, and the cost may also very high. Therefore, it is particularly important to obtain some suggestions based on data science.

1.2 Problem

Cities with larger GDP will generally have greater economic opportunities and higher standards of living, which is the first important factor for immigration that should be considered. In this project, the living standard of analysis of the top 50 cities in GDP was chosen to compare and cluster. Through the collection of living environment data in the entertainment, nature, sports, historical attractions, and culture, the world's most beautiful and historical cities will be selected. Through the GDP and population data analyze, the cities with high living standard were selected. The result can provide references for those planning to emigrate.

1.3 Interest

The main audience of this project should be the individuals planning to immigrate or relocate abroad and immigration agencies. The results may also interest the individuals and companies planning to invest, cooperate or travel internationally.

2. Data acquisition and cleaning

2.1 Data sources

For this project, the following data were used:

- World Cities Database (Basic update in 2020). Data source:World-Cities

	city	city_ascii	lat	lng	country	iso2	iso3	admin_name	capital	population	id
0	Tokyo	Tokyo	35.6897	139.6922	Japan	JP	JPN	Tōkyō	primary	37977000.0	1392685764
1	Jakarta	Jakarta	-6.2146	106.8451	Indonesia	ID	IDN	Jakarta	primary	34540000.0	1360771077
2	Delhi	Delhi	28.6600	77.2300	India	IN	IND	Delhi	admin	29617000.0	1356872604
3	Mumbai	Mumbai	18.9667	72.8333	India	IN	IND	Mahārāshtra	admin	23355000.0	1356226629
4	Manila	Manila	14.5958	120.9772	Philippines	PH	PHL	Manila	primary	23088000.0	1608618140

Description: The World Cities Database contains cities' population, latitude, and longitude.

- 150 richest cities in the world by GDP in 2020. Data source: 150 Richest Cities

	Rank	City	Country	GDP in 2020 in US$bn	Est annual growth 2005-2020
1	1	Tokyo	Japan	1602	2.0%
2	2	New York	USA	1561	2.2%
3	3	Los Angeles	USA	886	2.2%
4	4	London	UK	708	3.0%
5	5	Chicago	USA	645	2.3%

Description: The GDP top 50 city list with GDP was obtained from this source. The coordinator of each city was obtained from the world cities database which was used to explore the city.

- Data source: Fousquare API. Description: By using this API we will get 100 venues in each city.

2.2 Data cleaning

The data of World Cities Database was downloaded from the website and then upload to the GitHub repository. The data of the 150 richest cities in the world by GDP in 2020 was scraped from the website. I decided to only use the data of the top 50 cities in GDP, which have more attraction for the main audience. The data types of GDP and Growth rate for the 50 cities were changed from string to float.

The latitude and longitude of those 50 cities were obtained by combine the data with the World Cities Database. There are several problems with the dataset.

- First, some city names in the data of richest 50 cities do not match with that in the World Cities Database. The city names in the richest city dataframe were modified manually to match.

- Second, the cities were identified by their names. However, there were different cities with the same names, which cause get several different coordinators for one city. Since the World cities database was sorted by the population, and top GDP always come from the big cities. The first coordinator of the city was considered as the correct one. Therefore, cities with duplicate names just keep the first data and remove the rest duplicated.

After combine and cleaning, my master data has the main components *City, Country, Latitude, Longitude, GDP, Population and Estimate annual growth rate* information's of the top 50 cities.

	City	Country	lat	lng	GDP	population	Est annual growth 2005-2020
1	Tokyo	Japan	35.6897	139.6922	1602	37977000.0	2.0%
2	New York	USA	40.6943	-73.9249	1561	18713220.0	2.2%
3	Los Angeles	USA	34.1139	-118.4068	886	12750807.0	2.2%
4	London	UK	51.5072	-0.1275	708	10979000.0	3.0%
5	Chicago	USA	41.8373	-87.6862	645	8604203.0	2.3%

Python **folium** library was used to visualize geographic distribution of the 50 cities in the world. The latitude and longitude values were used to get the visual as below:

The Foursquare API were utilized to explore the cities and segment them. I designed the limit as 100 venue and the radius 50km for each city from their given latitude and longitude information. Here is a head of the merged table of cities and venues. Venues name, category, latitude and longitude information from Foursquare API.

	city	lat	lng	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Tokyo	35.6897	139.6922	Park Hyatt Tokyo (パークハイアット東京)	35.685575	139.690639	Hotel
1	Tokyo	35.6897	139.6922	BERG	35.691945	139.701082	Pub
2	Tokyo	35.6897	139.6922	Ohitsuzen Tanbo (おひつ膳 田んぼ)	35.682386	139.699894	Japanese Restaurant
3	Tokyo	35.6897	139.6922	Shinjuku Gyoen (新宿御苑)	35.685268	139.709528	Garden
4	Tokyo	35.6897	139.6922	Anshin Oyado (安心お宿 新宿駅前店)	35.689432	139.702708	Bed & Breakfast

In summary of this data, 4991 venues for all the 50 cities were returned by Foursquare. The top 49 cities reached the 100 limit of venues. While 91 venues were returned for Kolkata. Since 91 is

very close to 100, may not influence the analyze results. Therefore, we keep all the 50 cities for the further analyze.

2.3 Feature selection

After data cleaning, there were 4991 venues and 405 unique categories in the data. We only choose 85 categories and 1318 venues which relative with the living environment and grouped them in to 6 group.

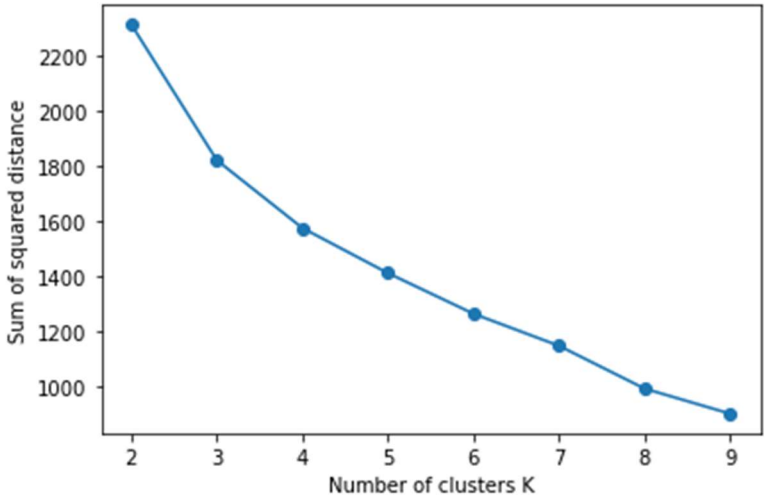
Group	Categories number	Categories List
Art	13	'Jazz Club', 'Art Museum', 'Art Gallery', 'Performing Arts Venue', 'Music Venue', 'Sculpture Garden', 'Street Art', 'Public Art', 'Outdoor Sculpture', 'Concert Hall', 'Comedy Club', 'Dance Studio', 'Country Dance Club',
Entertainment	11	'Indie Theater', 'Theater', 'Movie Theater', 'General Entertainment', 'Playground', 'Indie Movie Theater', 'Opera House', 'Theme Park Ride / Attraction', 'Theme Park', 'Resort', 'Pedestrian Plaza',
Nature	26	'Trail', 'National Park', 'Forest', 'Reservoir', 'Boat or Ferry', 'Hot Spring', 'Bay', 'Garden', 'Park', 'Scenic Lookout', 'Cricket Ground', 'Beach', 'Mountain', 'Island', 'Fountain', 'Zoo Exhibit', 'Water Park', 'Zoo', 'State / Provincial Park', 'Pier', 'Nature Preserve', 'Lake', 'Botanical Garden', 'Waterfront', 'Other Great Outdoors', 'Outdoor Event Space'
Historic	8	'Historic Site', 'Monument / Landmark', 'History Museum', 'Palace', 'Bridge', 'Shrine', 'Mosque', 'Pelmeni House'
Sport	14	'Racetrack', 'Rock Climbing Spot', 'Soccer Field', 'Baseball Stadium', 'Squash Court', 'Athletics & Sports', 'Basketball Court', 'Pool', 'Sports Club', 'Soccer Stadium', 'Climbing Gym', 'Tennis Court', 'Golf Course', 'Dive Bar'
Cultural	13	'Bookstore', 'Planetarium', 'Museum', 'Science Museum', 'Aquarium', 'Church', 'Cultural Center', 'Spiritual Center', 'Beer Garden', 'Temple', 'Buddhist Temple', 'Library', 'Martial Arts School'

3. **Methodology**

3.1 Living environment data

The unsupervised learning K-means algorithm was used to cluster the venue numbers in the above 6 Category groups of the cities.

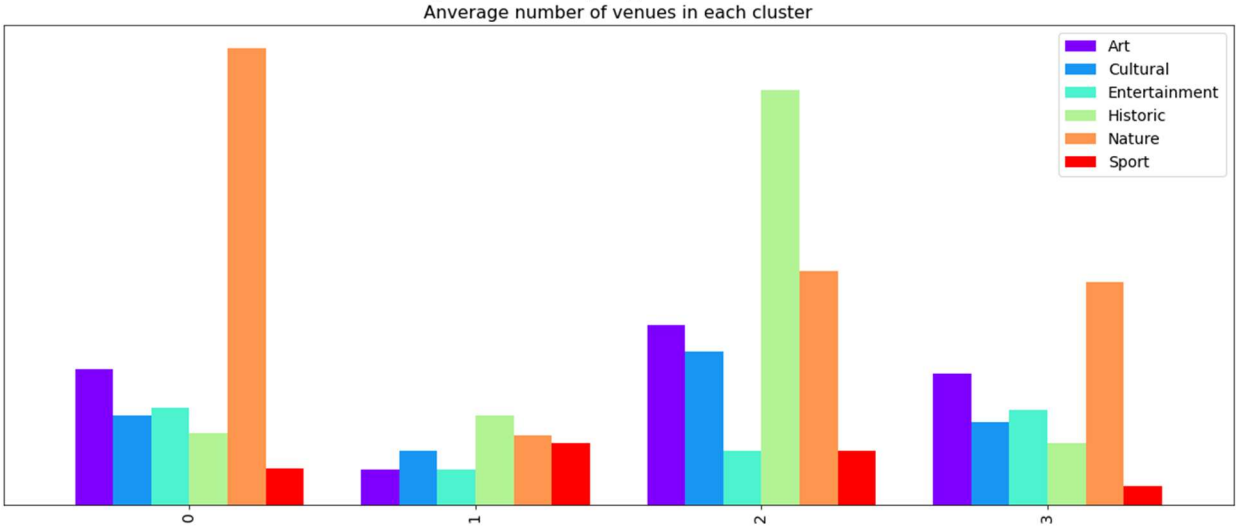
First, the elbow method was used to find the optimum the k of the K-Means. The results were showed in below figure. From the figure, we can find the elbow point is in 4 clusters.



Therefore, the K-Means was run to cluster the cities into 4 clusters.
Here is top 5 line of my merged table with cluster labels for each city.

	index	Cluster Labels	city	Art	Cultural	Entertainment	Historic	Nature	Sport
0	25	3	Milan	5	5	1	6	8	0
1	20	3	Madrid	8	2	5	5	3	0
2	47	3	Tokyo	9	3	1	3	5	0
3	43	3	Shanghai	4	0	3	0	7	0
4	40	3	Sao Paulo	8	6	3	1	3	3

A bar chart was also created to help us to find proper labels for each cluster.



When we examine above graph, we can label each cluster as follows:

- Cluster 0: “Nature & Entertainment City”
- Cluster 1: “Ordinary City with less venues”
- Cluster 2: “Historic, Art & Cultural City”
- Cluster 3: “Ordinary City with some Art and Nature Venues”

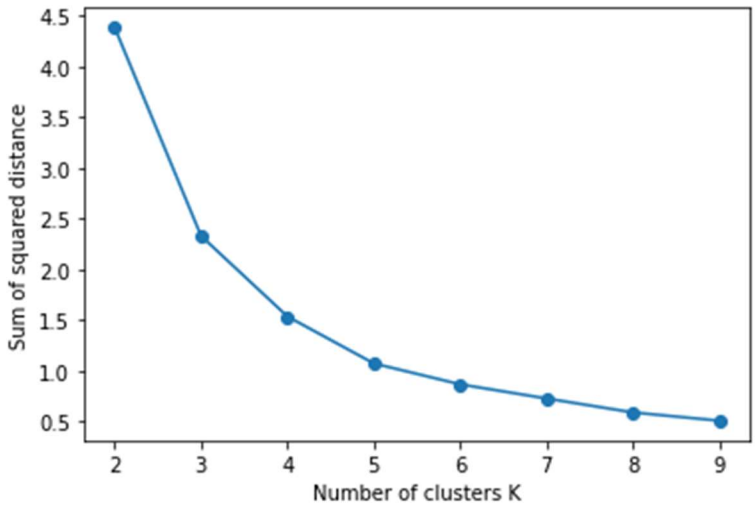
3.2 GDP, Growth rate and Population data

The GDP, Growth rate and population data of the 50 cities were normalized by the Min-Max Normalization. Here is top 5 line of the data table after normalized.

	GDP	population	Est annual growth 2005-2020%
0	1.000000	1.000000	0.075472
1	0.971429	0.474033	0.113208
2	0.501045	0.311238	0.113208
3	0.377003	0.262862	0.264151
4	0.333101	0.198022	0.132075

The unsupervised learning K-means algorithm was also used to cluster GDP, Growth rate and Population data.

First, the elbow method was used to find the optimum the k of the K-Means. The results were showed in below figure. From the figure, we can find the elbow point is in 5 clusters.

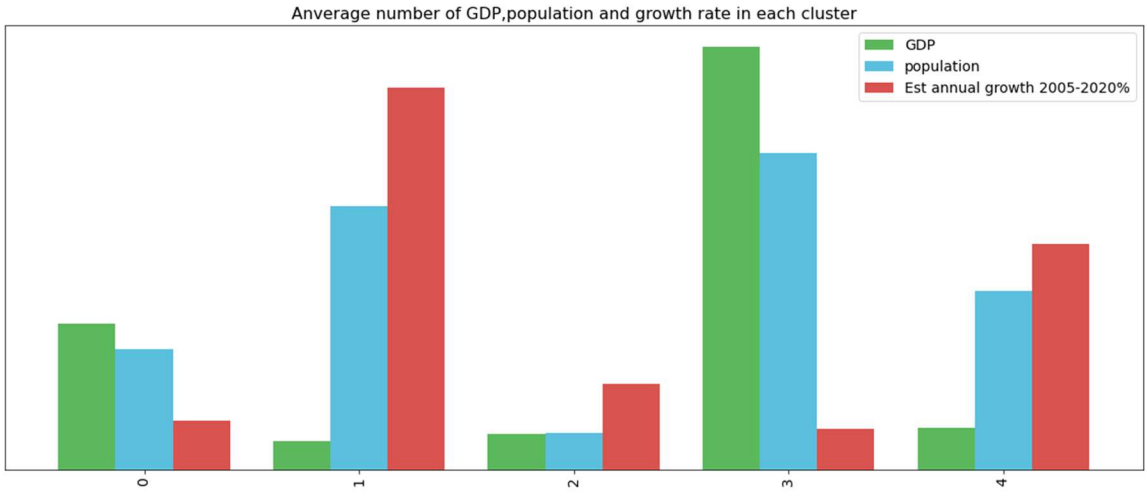


Therefore, the K-Means was run to cluster the cities into 5 clusters.

Here is top 5 line of the merged table with cluster labels for each city.

	index	GDP Cluster Labels	City	GDP	population	Est annual growth 2005-2020%
0	3	4	Los Angeles	886.0	12750807.0	2.2
1	4	4	London	708.0	10979000.0	3.0
2	5	4	Chicago	645.0	8604203.0	2.3
3	6	4	Paris	611.0	11020000.0	1.9
4	9	4	Osaka	430.0	14977000.0	1.6

A bar chart was also created to help us to find proper labels for each cluster.



When we examine above graph, we can label each cluster as follows:

- Cluster 0: “Middle GDP, middle population and low growth rate City”
- Cluster 1: “Low GDP, high population and high growth rate City”
- Cluster 2: “Low GDP, low population and middle growth rate City”
- Cluster 3: “High GDP, high population and low growth rate City”
- Cluster 4: “Low GDP, middle population and high growth rate City

4. **Result:**

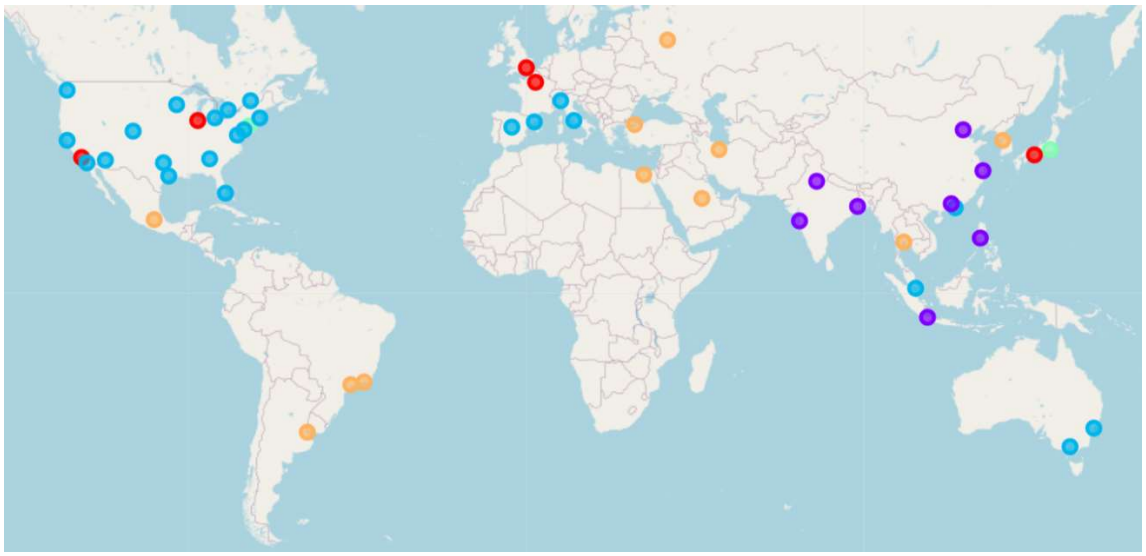
The two cluster labels were merge with the master tables and get the results table as below.

	GDP Cluster Labels	City	GDP	population	Est annual growth 2005-2020%	Cluster Labels	Art	Cultural	Entertainment	Historic	Nature	Sport
1	3	Tokyo	1602.0	37977000.0	2.0	3	9	3	1	3	5	0
2	3	New York	1561.0	18713220.0	2.2	0	6	4	6	1	15	3
3	0	Los Angeles	886.0	12750807.0	2.2	0	5	0	3	0	20	2
4	0	London	708.0	10979000.0	3.0	0	9	3	4	3	16	0
5	0	Chicago	645.0	8604203.0	2.3	3	4	1	2	2	12	0

The geographic distribution of the 4 clusters of the living environment data was visualized as below:

- Cluster 0(Red): “Nature & Entertainment City”
- Cluster 1(Purple): “Ordinary City with less venues”
- Cluster 2(Blue): “Historic, Art & Cultural City”
- Cluster 3(Yellow): “Ordinary City with some Art and Nature Venues”

The geographic distribution of the 5 clusters of the GDP and population data was visualize as below:



- Cluster 0(Red): “Middle GDP, middle population and low growth rate City”
- Cluster 1(Purple): “Low GDP, high population and high growth rate City”
- Cluster 2(Blue): “Low GDP, low population and middle growth rate City”
- Cluster 3(green): “High GDP, high population and low growth rate City”
- Cluster 4(orange): “Low GDP, middle population and high growth rate City

5. **Discussion**

The top 50 cities in GDP are mainly distributed in the traditional developed countries in Europe and the North America, as well as several rapidly developing countries in Asia, South America, and Africa. By the Kmeans algorithm, the Living environment data was clustering to 4 group.

	Cluster Labels	city
0	0	20
1	1	7
2	2	3
3	3	20

- The Cluster 2"Historic, Art & Cultural City": Istanbul, Beijing, and the Rome were grouped into the Cluster 2. All these three cities have more than 2,000 years history, representing ancient Asian and European civilizations respectively, and are world-renowned tourist attractions cities.
- The Cluster 0: "Nature & Entertainment City" included 20 cities. From the distribution map, we can see most of these cities are coastal cities with abundant natural scenery. They are also recommended places to travel and live. It is worth mentioning that Brazil's Rio de Janeiro is also on this Cluster.
- The Cluster 3: "Ordinary City with some Art and Nature Venues" also included 20 cities. These cities may not be attractive as the Cluster 0 or 2 as tourist attractions cities but is still a good place to living.
- The Cluster 1(Purple): "Ordinary City with less venues" included 7 cities which have relative less venues in the 6 areas. You may feel boring if you are living here for a long time. So, it is not recommended to relocation to these cities.

The first 49 cities reached the 100 limits of venues in the data collection steps. The result does not mean that inquiry run all the possible results in the cities. It depends on given Latitude and Longitude information and in this project, we just run single Latitude and Longitude pair for each city. Therefore, the result has certain reference value, but the accuracy can still be improved.
The GDP and population data were clustering to 5 clusters.

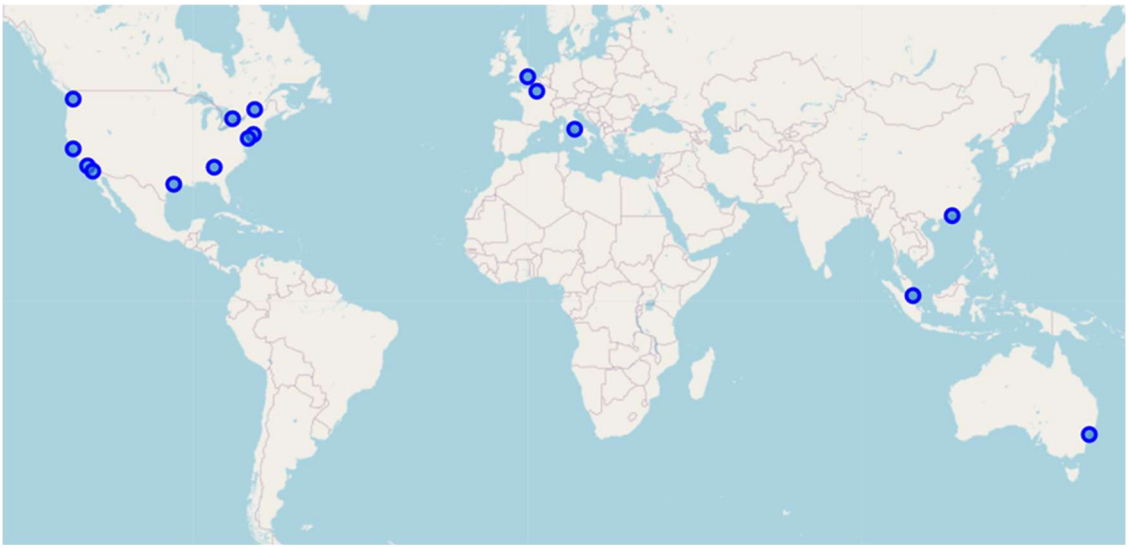
- The Cluster 3"High GDP, high population and low growth rate City" only have two cities, the Tokyo, and the New York. The two city have the top GDP. You may get a job with high salary in these two cities, but the GDP growth rate are relatively low which may not attractive for investment. The population in Tokyo are twice as that of New York. Based on the living environment data analyze result, the New York belong to Cluster 0(Red): "Nature & Entertainment City", the Tokyo belong to Cluster 3(Yellow): "Ordinary City with some Art and Nature Venues". So compared to Tokyo, you will feel much better for living at New York.
- Cluster 0(Red): "Middle GDP, middle population and low growth rate City" include 5 cities: Los Angeles, London, Chicago, Paris, Osaka. Angeles, London, Paris belongs to living environment Cluster 0(Red): "Nature & Entertainment City" which are

recommended to living. And the annual growth rate of London is 3.0% which is relatively high in the developed countries. So, the London is a good place to relocation.

- Cluster 2(Blue): “Low GDP, low population and middle growth rate City” include 24 cities. 11 cities in this group belong to the living Cluster 0(Red): “Nature & Entertainment City” Rome belongs to the living Cluster 2“Historic, Art & Cultural City”.
- Cluster 4(orange): “Low GDP, middle population and high growth rate City include 11 cities which is all located distributed in South America, Central Asia and other places. These group cities have a relative higher GDP per capita and a relative higher GDP growth rate. Four cities in this group belong to the living Cluster 0(Red): “Nature & Entertainment City” including the Rio de Janeiro, Tehran, Moscow, Seoul. Istanbul also belongs to the living Cluster 2“Historic, Art & Cultural City”:
- Cluster 1(Purple): “Low GDP, high population and high growth rate City” include 8 cities which is all located in Asian. The population of these cities are around or above 20 billion which may feel crowed than other types of cities. And the GDP per capita is much lower than other cities which is harder to maintain a high standard living. However, these cities have the highest GDP growth rate which will be attractive the investment.

	GDP Cluster Labels	GDP	population	Est annual growth 2005-2020%
0	0	5	5	5
1	1	8	8	8
2	2	24	24	24
3	3	2	2	2
4	4	11	11	11

In summary, 20 most beautiful cities (living environment cluster 0) and 3 historical cities(living environment cluster 2) were recommended to travel or relocation. 31 cities with high living standard(GDP cluster 0,3,2) were selected. The intersection between these two group cities are the 16 most recommended cities for immigration which were visualize as below.



6. Conclusion

In this project, the living standard of analysis of the top 50 cities in GDP was chosen to compare and cluster. Through the collection of living environment data in the entertainment, nature, sports, historical attractions, and culture, 20 most beautiful cities and 3 historical cities were recommended to travel or relocation. Through the GDP and population data analyze, 31 cities with high living standard were selected. Combine the two results, the 16 most recommended cities for immigration were visualize. The result can provide references for those planning to immigrate, invest or travel internationally.