Big City Similarity Clustering

Sirui Wang

31 August 2019

1. Introduction

1.1Background

A New York Luxury Brand has built its business in several cities in the United States, including Los Angeles, New York, and Chicago. Due to its success and growing popularity in these cities, the CEO and his team wants to expand their business to other cities in the United States and also explore their market in big cities in other countries, such as China and UK. Before setting up the business in these cities, the company needs to do extensive research to be more familiar with these cities and consider different business modes to operate in different places.

1.2 Problem

Now the CEO has hired a data scientist and assigned her a task to find out the similarity between different big cities in the world and group the cities into various clusters, so that the Board of Directors can make a better decision of which business mode to operate in new cities. (For Example: If London has the been grouped into the same cluster with New York, the business mode operated in New York market will be considered for London). The similarity test should be based on various factors, including but not limiting to geolocations, economic development, cultures, population components and so on. In order to carry out the task, the data scientist should make a full use of FourSquare API and collect a dataset for at least 15 cities, including those in the United States and those in other countries.

2. Data Acquisition and Cleaning

2.1. Data Source

We chose 27 most popular cities in the world and clustered them based on three factors, venues distribution, GDP indicator, and climate types. The location information of these cities, including latitudes and longitudes, are obtained by using geolocator package on python. The venues information is retrieved from FourSquare API and at most 500 venues were selected for each city, while the GDP information and climate type information are scraped from online Wikipedia pages. The GDP data is released by Brookings Institution

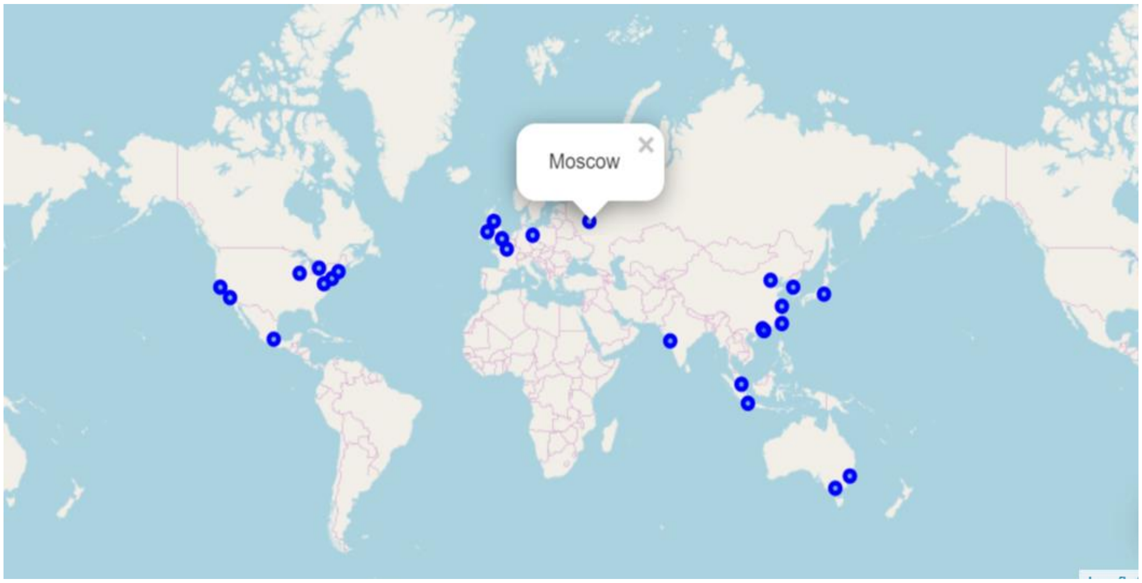

Figure 1. World map with location points

2.2. Data Cleaning

Data scraped from online sources contain extensive information that we might not need for analysis. Thus, we dropped out irrelevant data and only select those we need – venues category, annual GDP, and average annual temperature. Since venue categories are of the type string and need to be quantified for modeling, we apply one-hot coding to the venue category. The resulting data frame is as follows:

	City	Accessories Store	Adult Boutique	African Restaurant	American Restaurant	Amphitheater	Aquarium	Arcade	Ar Restau
0	New York	0	0	0	0	0	0	0	
1	New York	0	0	0	0	0	0	0	
2	New York	0	0	0	0	0	0	0	
3	New York	0	0	0	0	0	0	0	
4	New York	0	0	0	0	0	0	0	

Table 1. City Venue Category (one-hot coding)

Temperature and GDP data frame are as follows. The entry values are converted from strings to float numbers, and they are also normalized for modeling

	City	Normalized GDP	GDP
0	Tokyo	0.509116	1617.0
1	New York	0.441738	1403.0
2	Los Angeles	0.270930	860.5
3	Seoul	0.266333	845.9
4	London	0.263122	835.7

Table 2. City with GDP table

	City	Normalized Temperature	Temperature
0	Mumbai	0.307823	27.1
1	Singapore	0.306687	27.0
2	Jakarta	0.303279	26.7
3	Hong Kong	0.264659	23.3
4	Taipei	0.261252	23.0

Table 3. City with Temperature table

3. Clustering Modeling

3.1 Feature Summary

We visualized the data to have a look at all the features. Here is the table with the top 10 venue categories for all cities.

	City	1th Most Common Venue	2th Most Common Venue	3th 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	Beijing	Historic Site	Hotel	Park	Chinese Restaurant	Yunnan Restaurant	Café	Coffee Shop	Hostel	Temple	Peking Duck Restaurant
1	Berlin	Coffee Shop	Bookstore	Park	Ice Cream Shop	Concert Hall	Sandwich Place	Garden	Bakery	Hotel	Wine Bar
2	Boston	Park	Bakery	Hotel	Seafood Restaurant	Theater	Mexican Restaurant	Historic Site	Gym	Pizza Place	Gastropub
3	Chicago	Hotel	Park	Theater	Italian Restaurant	Coffee Shop	New American Restaurant	Boat or Ferry	Mediterranean Restaurant	Restaurant	Burger Joint
4	Dublin	Coffee Shop	Café	Pub	Restaurant	Park	Cocktail Bar	Burger Joint	Hotel	Italian Restaurant	Indie Movie Theater
5	Edinburgh	Café	Coffee Shop	Bar	Hotel	French Restaurant	Park	Cocktail Bar	Museum	Pub	Whisky Bar
6	Guangzhou	Hotel	Coffee Shop	Park	Shopping Mall	Chinese Restaurant	Turkish Restaurant	Restaurant	Cantonese Restaurant	Electronics Store	Cocktail Bar
7	Hong Kong	Hotel	Bar	Japanese Restaurant	Italian Restaurant	Gym / Fitness Center	Scenic Lookout	Cocktail Bar	Lounge	Café	Steakhouse
8	Jakarta	Hotel	Coffee Shop	Restaurant	Shopping Mall	Indonesian Restaurant	Dessert Shop	Sushi Restaurant	Food Truck	Asian Restaurant	Buffet
9	London	Hotel	Cocktail Bar	Theater	Art Museum	Park	Bookstore	Department Store	Coffee Shop	Hotel Bar	Clothing Store
10	Los Angeles	Coffee Shop	Brewery	Theater	Taco Place	Ice Cream Shop	Hotel	American Restaurant	Bookstore	Sushi Restaurant	Art Gallery
11	Melbourne	Coffee Shop	Café	Cocktail Bar	Park	Wine Bar	Ice Cream Shop	Italian Restaurant	Asian Restaurant	Theater	Music Venue
12	Mexico City	Ice Cream Shop	Mexican Restaurant	Art Museum	Hotel	Park	Coffee Shop	Public Art	Seafood Restaurant	Spa	Bakery
13	Moscow	Yoga Studio	Hotel	Theater	Park	Pizza Place	Garden	Road	Jewelry Store	Coffee Shop	Art Gallery
14	Mumbai	Indian Restaurant	Café	Hotel	Ice Cream Shop	Dessert Shop	Pizza Place	Fast Food Restaurant	Coffee Shop	Italian Restaurant	Bakery
15	New York	Park	Cycle Studio	Ice Cream Shop	Bookstore	Scenic Lookout	Italian Restaurant	Gym	Theater	Wine Shop	Yoga Studio
16	Paris	Plaza	Cocktail Bar	Hotel	Italian Restaurant	Seafood Restaurant	Bookstore	Wine Bar	Bistro	Ice Cream Shop	Art Museum
17	San Francisco	Park	Coffee Shop	Yoga Studio	Grocery Store	Ice Cream Shop	Marijuana Dispensary	Gym	Dance Studio	Wine Bar	Bakery
18	Seoul	Coffee Shop	Korean Restaurant	Park	Multiplex	Market	BBQ Joint	Bakery	Golf Course	Soccer Stadium	Café
19	Shanghai	Hotel	Hotel Bar	Lounge	Dumpling Restaurant	French Restaurant	Italian Restaurant	Café	Chinese Restaurant	Scenic Lookout	Shopping Mall
20	Shenzhen	Hotel	Coffee Shop	Shopping Mall	Café	Chinese Restaurant	Park	Hotpot Restaurant	Japanese Restaurant	Bookstore	Electronics Store
21	Singapore	Hotel	Park	Ice Cream Shop	Bakery	Supermarket	Event Space	Performing Arts Venue	Steakhouse	Trail	Buffet
22	Sydney	Café	Park	Theater	Coffee Shop	Scenic Lookout	Hotel	Bakery	Cocktail Bar	Thai Restaurant	Italian Restaurant
23	Taipei	Hotel	Café	Dessert Shop	Japanese Restaurant	Bakery	Dumpling Restaurant	Noodle House	Bookstore	Park	Mountain
24	Tokyo	Hotel	Sake Bar	Tonkatsu Restaurant	Kaiseki Restaurant	BBQ Joint	Chinese Restaurant	Wagashi Place	Coffee Shop	Japanese Curry Restaurant	Garden
25	Toronto	Coffee Shop	Café	Restaurant	Hotel	Park	Japanese Restaurant	Pizza Place	Yoga Studio	Sandwich Place	Italian Restaurant
26	Washington	Monument / Landmark	Hotel	Art Museum	Coffee Shop	History Museum	Theater	Salon / Barbershop	American Restaurant	Science Museum	Mediterranean Restaurant

Table 4. Cities with the top 10 popular venues

Temperature and GDP data are sorted and plotted out as bar charts. We can view the ranking from low to top

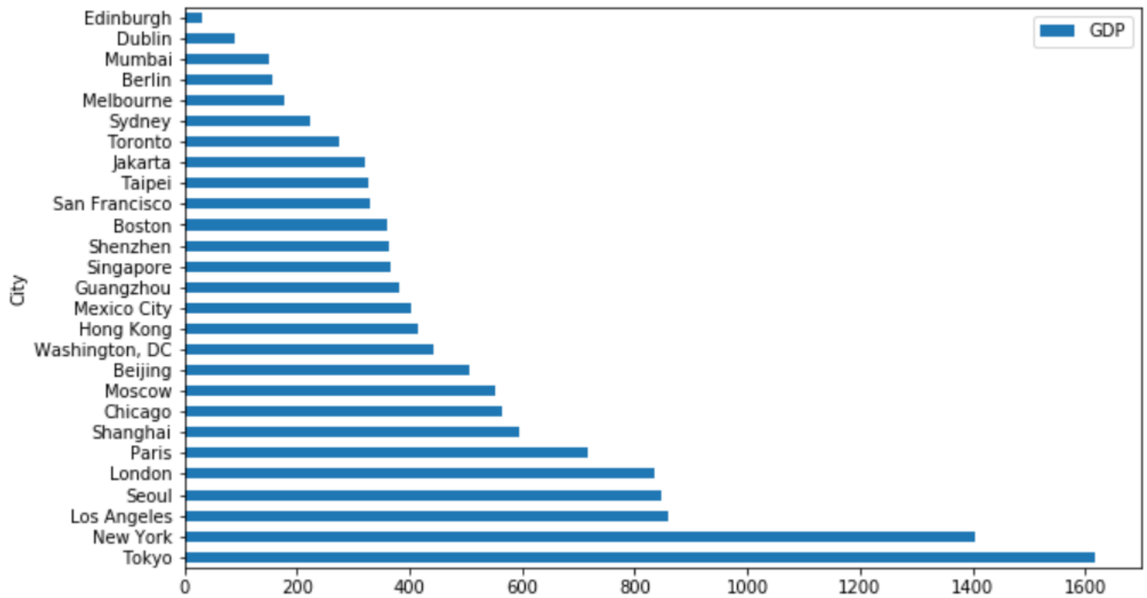

Figure 2. City rankings in GDP

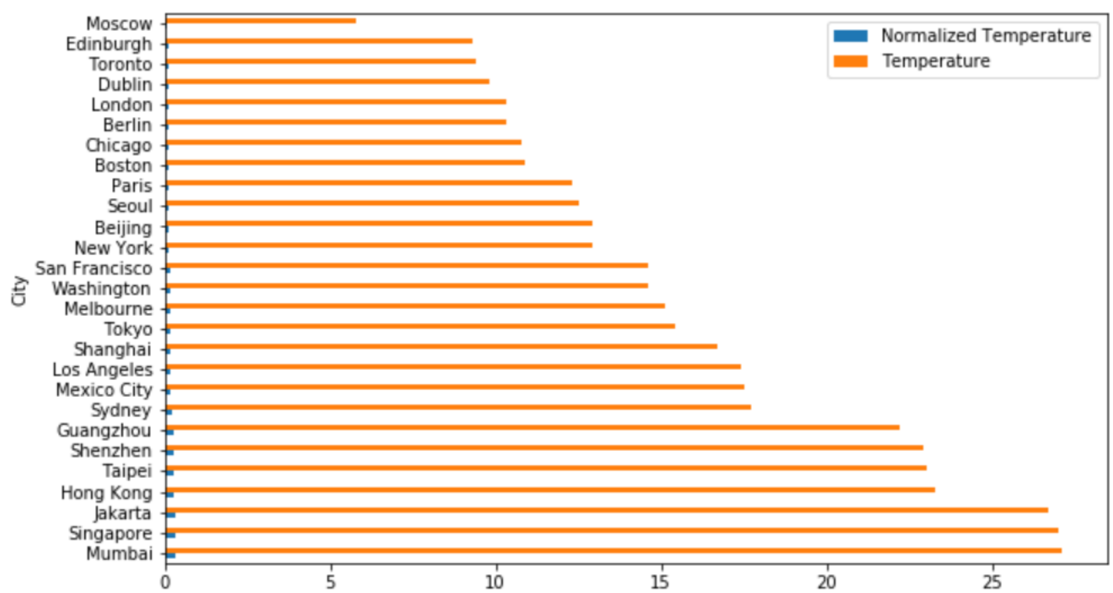

Figure 3. City rankings in Average Annual Temperature

3.2 K-means Clustering

We chose k-means clustering model as our data are all continuous and numeric (after cleaning). Besides, we randomly and manually chose the city sample from across the world, and k-means could be an appropriate model to categorize the cities into clusters with similar traits.

We calculate the percentage of different venue categories for each city, and normalized the temperature and GDP values. The final dataset for modeling is as follows:

	City	Accessories Store	Adult Boutique	African Restaurant	American Restaurant	Amphitheater	Aquarium	Arcade	Arepa Restaurant	Argentinian Restaurant	...	Women's Store	Xinjiang Restaurant	Yakitori Restaurant	Yoga Studio
0	Beijing	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	...	0.00	0.01	0.00	0.00
1	Berlin	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	...	0.00	0.00	0.00	0.01
2	Boston	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	...	0.00	0.00	0.00	0.02
3	Chicago	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	...	0.00	0.00	0.00	0.02
4	Dublin	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	...	0.00	0.00	0.00	0.00
5	Edinburgh	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	...	0.00	0.00	0.00	0.00
6	Guangzhou	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...	0.00	0.00	0.00	0.00
7	Hong Kong	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...	0.00	0.00	0.00	0.03
8	Jakarta	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...	0.00	0.00	0.00	0.00
9	London	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...	0.00	0.00	0.00	0.00
10	Los Angeles	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	...	0.00	0.00	0.00	0.00

n nt	Amphitheater	Aquarium	Arcade	Arepa Restaurant	Argentinian Restaurant	...	Women's Store	Xinjiang Restaurant	Yakitori Restaurant	Yoga Studio	Yoshoku Restaurant	Yunnan Restaurant	Zhejiang Restaurant	Zoo	Normalized GDP	Normalized Temperature
1	0.00	0.00	0.00	0.00	0.00	...	0.00	0.01	0.00	0.00	0.00	0.04	0.01	0.00	0.239020	0.219792
0	0.00	0.00	0.00	0.00	0.01	...	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.074478	0.175493
1	0.00	0.01	0.00	0.00	0.00	...	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.170067	0.185716
1	0.01	0.00	0.00	0.00	0.00	...	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.265987	0.184012
0	0.00	0.00	0.01	0.00	0.00	...	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.042552	0.166974
0	0.00	0.00	0.00	0.00	0.00	...	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.015349	0.158455
0	0.00	0.00	0.00	0.00	0.00	...	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.179607	0.378247
0	0.00	0.00	0.00	0.00	0.00	...	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.01	0.196468	0.396989
0	0.00	0.00	0.00	0.00	0.00	...	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.151743	0.454919
0	0.00	0.00	0.00	0.00	0.00	...	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.394683	0.175493
3	0.00	0.00	0.00	0.00	0.00	...	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.406395	0.29646
0	0.00	0.00	0.00	0.00	0.01	...	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.084254	0.257276

Table 5. Clustering data for modeling

After modeling, 6 groups of cities are clustered out and the result is presented in the next section

4. Results and Discussion

4.1Clustering Results

We clustered out 6 labels and marked out the points on the world map as follows:

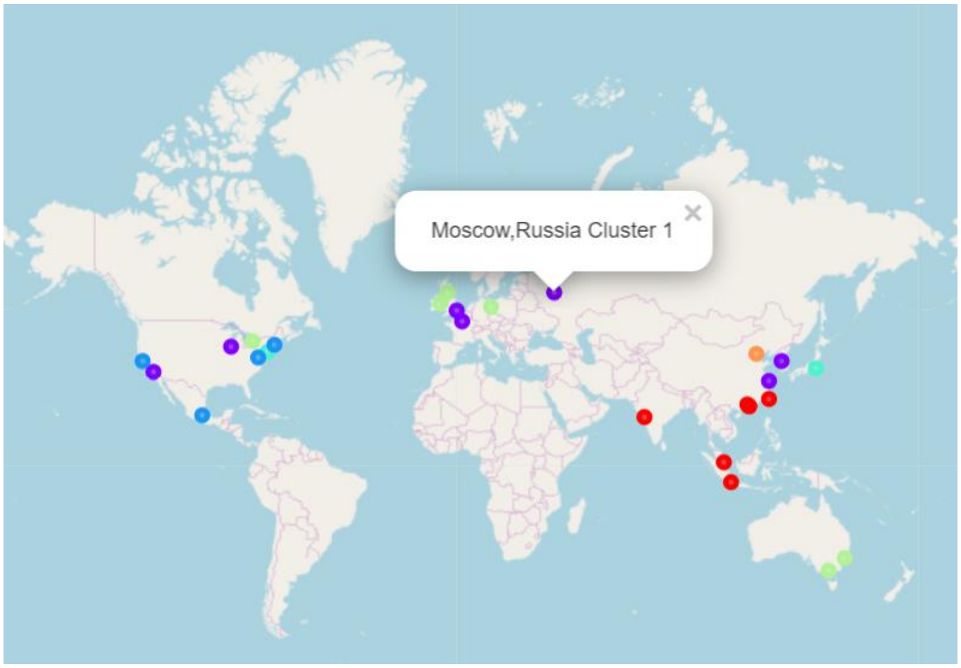

Figure 4. World Map with marked location points

4.2 Discussion of Results

The cities in the Cluster 0 are shown below:

	City	Country	Cluster Labels	1th Most Common Venue	2th Most Common Venue	3th Most Common Venue	4th Most Common Venue
5	Singapore	Singapore	0	Hotel	Park	Ice Cream Shop	Bakery
7	Hong Kong	China	0	Hotel	Bar	Japanese Restaurant	Italian Restaurant
16	Guangzhou	China	0	Hotel	Coffee Shop	Park	Shopping Mall
17	Shenzhen	China	0	Hotel	Coffee Shop	Shopping Mall	Café
18	Mumbai	India	0	Indian Restaurant	Café	Hotel	Ice Cream Shop
23	Taipei	China	0	Hotel	Café	Dessert Shop	Japanese Restaurant
25	Jakarta	Indonesia	0	Hotel	Coffee Shop	Restaurant	Shopping Mall

Table 6. Cluster 0 table

These 7 cities are all located in the southern Asian areas with similar climates and temperatures. Their GDP are close too and are lower than those of the cluster 1 cities. 6 of them have hotel as the most common venue and coffee shops/cafe are very popular too. This shows that tourism might be an essential source of income for these cities. The Board of

Director can consider a new business mode customized for cities in this cluster which might cater to the needs of tourists and the spending power of the residents.

The cities in the Cluster 1 are shown below:

	City	Country	Cluster Labels	1th Most Common Venue	2th Most Common Venue	3th Most Common Venue	4th Most Common Venue
1	London	UK	1	Hotel	Cocktail Bar	Theater	Art Museum
8	Los Angeles	US	1	Coffee Shop	Brewery	Theater	Taco Place
9	Chicago	US	1	Hotel	Park	Theater	Italian Restaurant
15	Shanghai	China	1	Hotel	Hotel Bar	Lounge	Dumpling Restaurant
21	Moscow	Russia	1	Yoga Studio	Hotel	Theater	Park
22	Paris	France	1	Plaza	Cocktail Bar	Hotel	Italian Restaurant

Table 7. Cluster 1 table

These 6 cities are from all across the world. One common feature among them is that theaters are pretty popular in these cities. London, LA, Chicago, and Moscow are four out of five cities among all with theaters in the top 3 most common venues, and they all have developed arts and entertainment industries. Besides, Hotels are popular in these cities and all 6 cities all have close GDP, which is nearly 2 times higher than that of cluster 0 cities. However, their climates are pretty different. Since Los Angeles and Chicago are categorized into this cluster, the Business mode for these two cities could be applied to other cities in this group.

The cities in the Cluster 2 are shown below:

	City	Country	Cluster Labels	1th Most Common Venue	2th Most Common Venue	3th Most Common Venue	4th Most Common Venue
10	Boston	US	2	Park	Bakery	Hotel	Seafood Restaurant
11	San Francisco	US	2	Park	Coffee Shop	Yoga Studio	Grocery Store
13	Washington	US	2	Monument / Landmark	Hotel	Art Museum	Coffee Shop
26	Mexico City	Mexico	2	Ice Cream Shop	Mexican Restaurant	Art Museum	Hotel

Table 8. Cluster 2 table

In this cluster, 3 cities are from the United States and all 4 cities are located in the American continents. Their popular venues include parks and museums, which shows people in these cities are enjoying a rather slowly-paced life. These cities have a similar GDP too, which is slightly higher than that of cluster 0 but much lower than that of cluster 1. The Board of Directors can consider creating a rather relaxing and joyful environment for the customers when they are operating business in these cities, such as building a common space near the shopping stores

The cities in the Cluster 3 are shown below:

	City	Country	Cluster Labels	1th Most Common Venue	2th Most Common Venue	3th Most Common Venue	4th Most Common Venue	5th Mo Commc Venu
0	New York	US	3	Park	Cycle Studio	Ice Cream Shop	Bookstore	Scen Looko
19	Tokyo	Japan	3	Hotel	Sake Bar	Tonkatsu Restaurant	Kaiseki Restaurant	BB Joi

Table 9. Cluster 3 table

This cluster only has 2 cities, one in East Asia and the other in North America. These 2 cities have the highest GDP among all cities, and have similar climates too. However, their venue distribution are quite similar. Restaurants take up a large portion of venues in Tokyo, while venues in New York are pretty diverse. Since New York is in this cluster, the Business mode for NY can be considered for Tokyo as well.

The cities in the Cluster 4 are shown below:

	City	Country	Cluster Labels	1th Most Common Venue	2th Most Common Venue	3th Most Common Venue	4th Most Common Venue
2	Edinburgh	UK	4	Café	Coffee Shop	Bar	Hotel
3	Toronto	Canada	4	Coffee Shop	Café	Restaurant	Hotel
4	Sydney	Australia	4	Café	Park	Theater	Coffee Shop
6	Melbourne	Australia	4	Coffee Shop	Café	Cocktail Bar	Park
12	Dublin	Ireland	4	Coffee Shop	Café	Pub	Restaurant
24	Berlin	Germany	4	Coffee Shop	Bookstore	Park	Ice Cream Shop

Table 10. Cluster 4 table

This cluster has the lowest GDP among all clusters. Most cities are located in Europe and Australia. The similarity among these cities is that they are western countries and people in these cities generally enjoy a western lifestyle. They are unlike cities such as London or LA, which are more international and their popular venues are coffee shops, parks, and bars. The Board of Directors can consider a new Business mode that caters more to people with western lifestyles and relatively lower spending power.

The cities in the Cluster 5 are shown below:

	City	Country	Cluster Labels	1th Most Common Venue	2th Most Common Venue	3th Most Common Venue	4th Most Common Venue	5th Mo Comm Ven
14	Beijing	China	5	Historic Site	Hotel	Park	Chinese Restaurant	Yunn Restaur

Table 11. Cluster 5 table

Beijing is the only city in this cluster. Its GDP value and Average Annual Temperature value are in the middle. However, its venue distribution is rather unique compared to other cities. It is the only city with historic site as the most common venue, and also the only city with Yunnan and Pecking duck restaurants in the top 10 venues. This shows Beijing is a rather cultured and unique city differentiated with other cities. Thus, the Board of Directors might need to think more about how to customize and localize their business in Beijing.

5. Conclusion

In this assignment we have built up a clustering model to segment the major big cities into different groups. The result could be a valuable reference to the Board of Directors when they are making decisions on their business expansions into these cities. This results is straight-forward and takes different factors into account. However, there is also room for improvement, as there are a lot of features that influence the similarity between two cities and more variables could be included for higher accuracy of the clustering results.