2020

Similarity Analysis among New York, Toronto and Hong Kong

Vinci Poon

Coursera Capstone Project

2020/2/1

Table of Contents

1. Introduction

1.1 Business Background

New York, Toronto and Hong Kong and famous cities in the world. It is worthy to have a similarity analysis among these cities to understand the different cultures of different nationalities. According to previous study and based on that, we would like to use Foursqaure data to provide the most popular type of venues in different neighborhoods for analysis. Where New York and Toronto are in North America but different countries, and Hong Kong is in Asia, we expect Toronto is more similar to New York than Hong Kong.

1.2 Problem

Is New York City more similar to Toronto rather than Hong Kong?

2. Data acquisition and cleaning

2.1 Analytic Approach

Given that we have the geolocation data of New York and Toronto, we also prepare the data for Hong Kong referring to their post offices' addresses. This data is provided by https://geodata.gov.hk/.
Make use of the geolocations, it represents the neighborhoods of an area and we can use Foursqaure API to get the list of venues surrounding. In order to compare the cities, we will sum up the counts of common places of the city to prepare the top 10 common places to compare.
Down to the neighborhoods, we will also apply K-MEAN clustering to label the different neighborhoods to see if the classes are evenly distribute or not.

2.2 Data sources

1. List of Neighborhoods and Boroughs of New York with geolocations
2. List of Neighborhoods and Boroughs of Toronto with geolocations
3. List of Neighborhoods and Boroughs of Hong Kong with geolocations
4. List of places surround Neighborhoods in New York City by Foursquare
5. List of places surround Neighborhoods in Toronto City by Foursquare
6. List of places surround Neighborhoods in Hong Kong City by Foursquare

2.3 Data processing

Detail steps refer to Jupyter Notebook section 3:
https://github.com/pwpvinci/Coursera_Capstone/blob/master/coursera_capstone.ipynb

3. Methodology

3.1 Analysis on City level

Base on the Lists of places surround Neighborhoods in these cities, we can sum up all the places count of a city and have their top-ranking places category.
Consider more demanding of the place types, the more places to be built. If the more top-ranking places are a-like, the more similar among the cities. Therefore, we have made matrix to show the counts of common top-ranking places:

Common top 10 ranking places count:

	City	NYC	Toronto	HK
0	NYC	100.00%	50.00%	20.00%
1	Toronto	50.00%	100.00%	40.00%
2	HK	20.00%	40.00%	100.00%

Common top 20 ranking places count:

	City	NYC	Toronto	HK
0	NYC	100.00%	50.00%	20.00%
1	Toronto	50.00%	100.00%	40.00%
2	HK	20.00%	40.00%	100.00%

Common top 50 ranking places count:

	City	NYC	Toronto	HK
0	NYC	100.00%	58.00%	52.00%
1	Toronto	58.00%	100.00%	50.00%
2	HK	52.00%	50.00%	100.00%

Common top 100 ranking places count:

	City	NYC	Toronto	HK
0	NYC	100.00%	66.00%	63.00%
1	Toronto	66.00%	100.00%	63.00%
2	HK	63.00%	63.00%	100.00%

3.2 Analysis on Neighborhoods level

In this analysis, we normalize the Lists of places surround Neighborhoods in these cities. Each place counts as a contributor to identify the nature of neighborhoods. Take top 10 places of the city as Features then use to cluster the neighborhoods.

1. New York City:

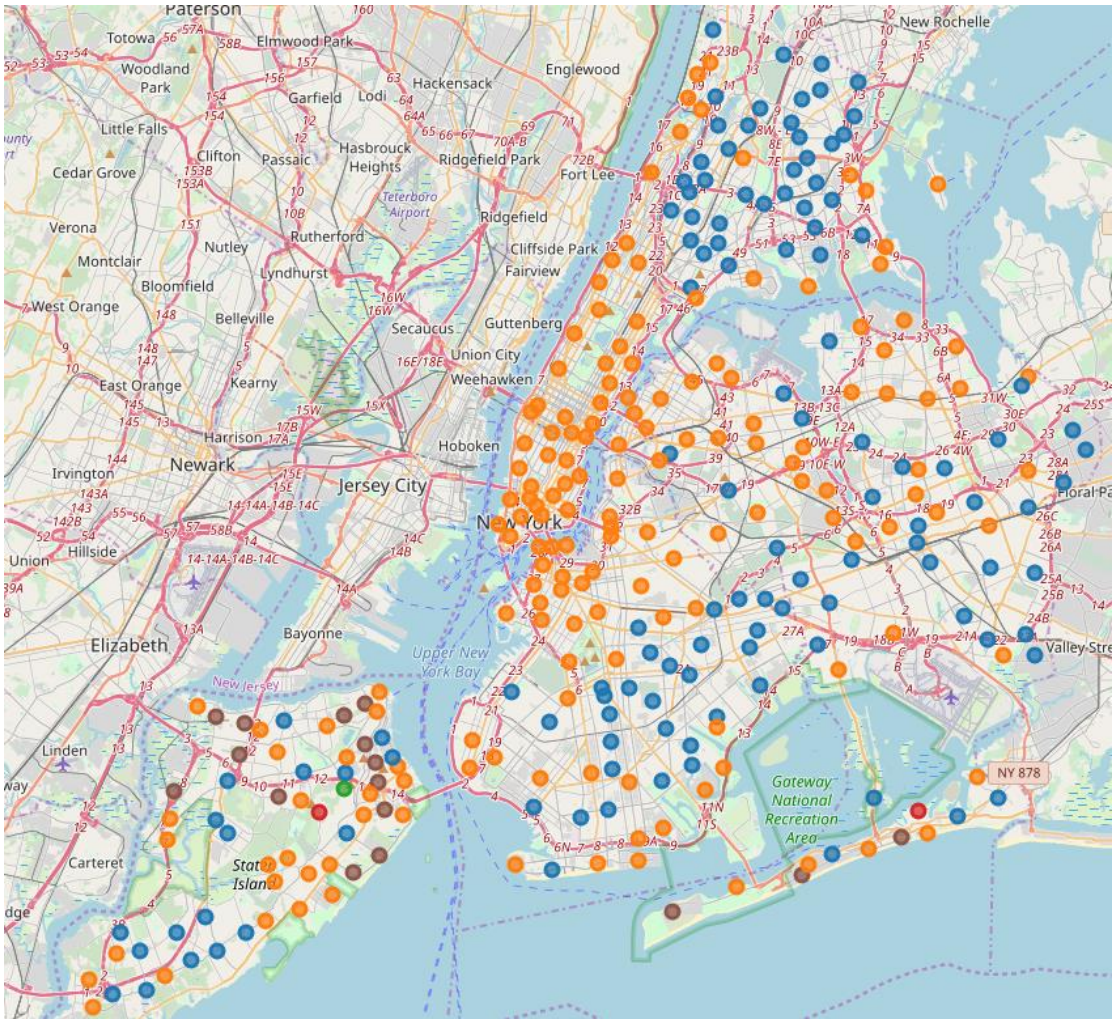

Figure 3.2.1.1 Neighborhoods clustering in New York City

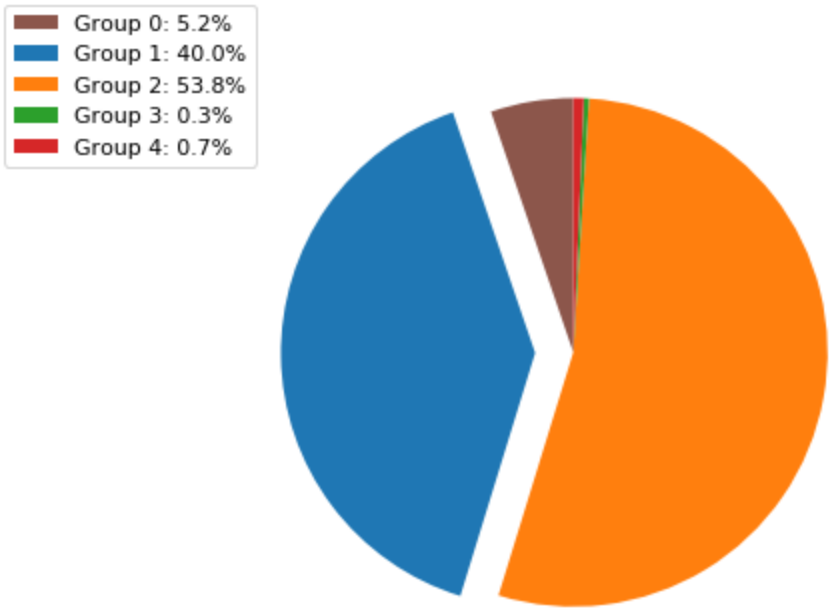

Figure 3.2.1.2 Clustering in New York City

2. Toronto:

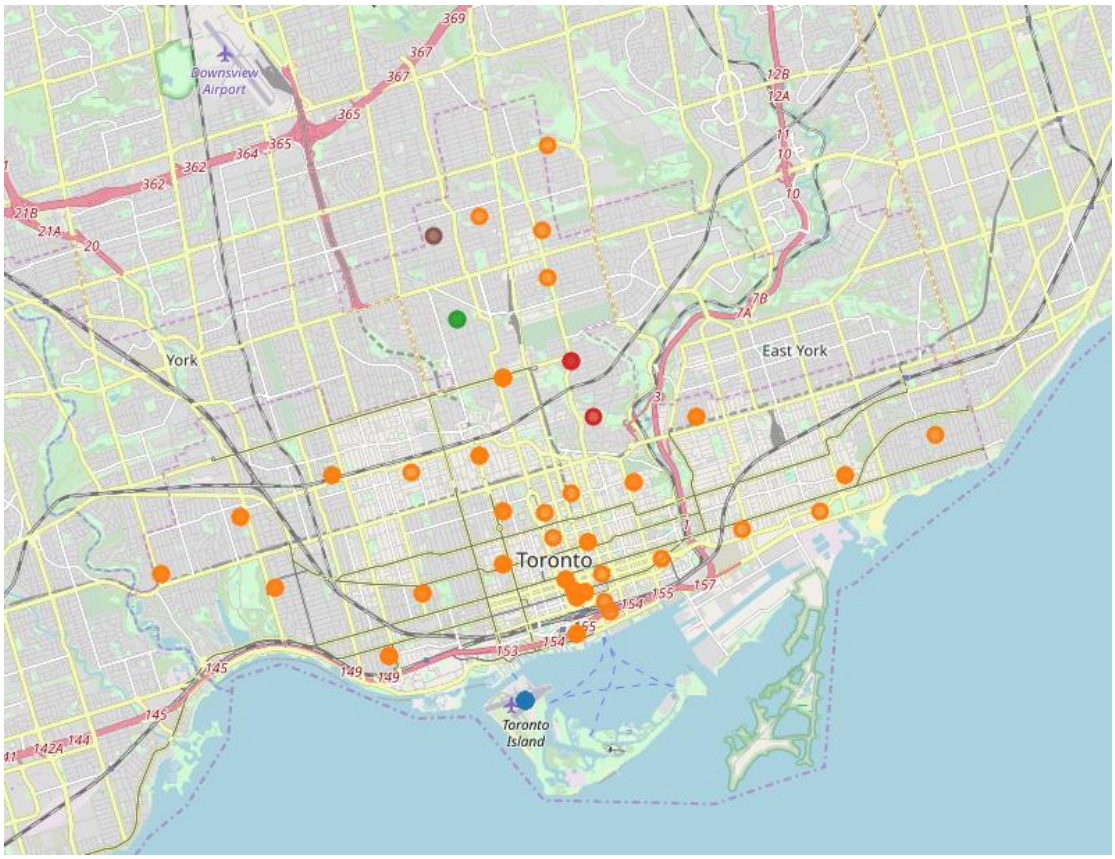

Figure 3.2.2.1 Neighborhoods clustering in Toronto

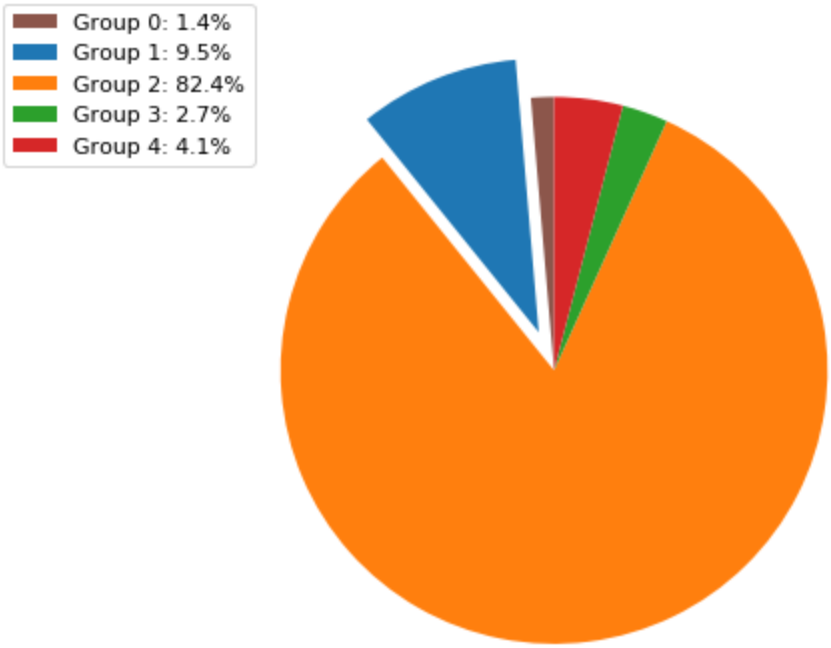

Figure 3.2.2.2 Clustering in Toronto

3. Hong Kong:

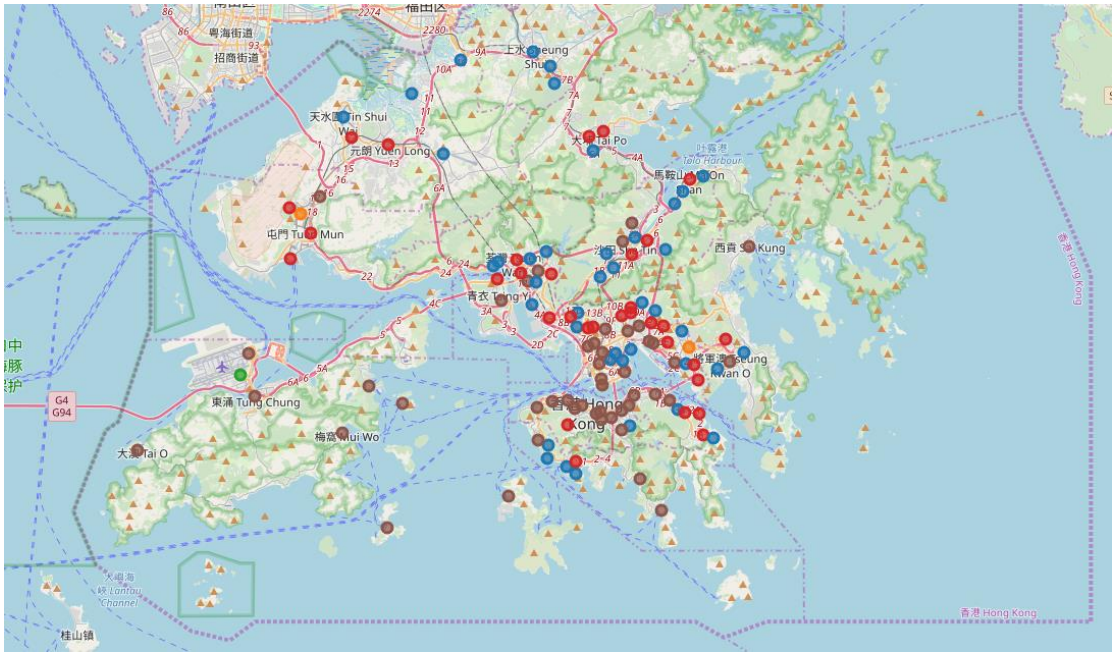

Figure 3.2.3.1 Neighborhoods clustering in Hong Kong

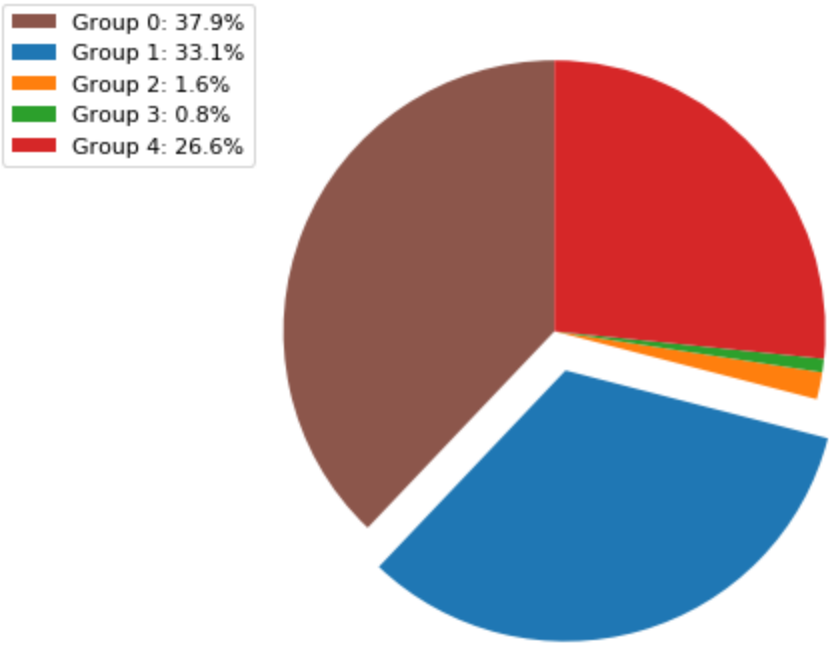

Figure 3.2.3.2 Clustering in Hong Kong

4. Discussion

4.1 Results

Let's recap our problem: Is New York City more similar to Toronto rather than Hong Kong?

We adopt the places found by Foursqaure API to observe the features of culture of these 3 cities. First of all(in section 3.1), we study the common top ranked places in these city among top 10, 20, 50 and 100. We assume that the more popular places, the more in demand. In each comparison matrix, we still get the same result that New York City has more common popular places with Toronto compare to Hong Kong. It is suggested that people in the city of New York City and Toronto willing to have the same demanding trend on the places.

Secondly, we use k-Means to classify the neighborhoods in these 3 cities and observe their distributions. If we use k-Means clustering analysis on the places in different neighborhoods to classify, we have the result as the mapping and pie charts. According to the figures, Toronto and New York City's neighborhoods are not evenly distributed. They are dominated by the top 2 groups. In other case, the top 3 groups

of neighborhoods in Hong Kong dominated the classifications.

4.2 Limitation and Improvement

There are some limitations on this analysis:

1. too few cities for analysis
2. variety of data for analysis
3. data transformation

Due to insufficient data set, we only made analysis on places' categories of these 3 cities. In fact, to make the analysis more fruitful, we should have more cities' information to prepare the analysis.

In order to compare the cities, we should not only focus on the categories of places. We can add the cities' population, GDP, vacation and education data, then we can compare the cities in different angles.

Lastly, we observed that the data retrieved from Foursequare API are too detailed in categories, e.g. Chinese Restaurant, Japanese Restaurant and Pizza Store should be group into restaurant and different from bookstore. Therefore, there should be some cases the same type of category are splitted into small group and their significances are minimized.

4.3 Further Study

As mentioned in section 5.2, we can add more dataset for further study. Here are 2 suggestions:

1. Similarity among cities in the same continents
2. Similarity of cities on Economy, Education and Population

5. Conclusion

In conclusion, the analysis result shows New York City is more similar to Toronto rather than Hong Kong. For further study, we should include more dataset of cities' information to have a more fruitful result.