Coursera Capstone Project

Similarity Analysis among New York, Toronto and Hong Kong

Vinci Poon
23rd Feb, 2020

Agenda

- Business Backgrounds
- Data acquisition and cleaning
- Analysis
- Discussion
- Conclusion

Business Backgrounds

- 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.
- **So, Is New York City more similar to Toronto rather than Hong Kong?**

Data acquisition and cleaning

- 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 distributing or not.

Data acquisition and cleaning

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

Analysis on city level

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%

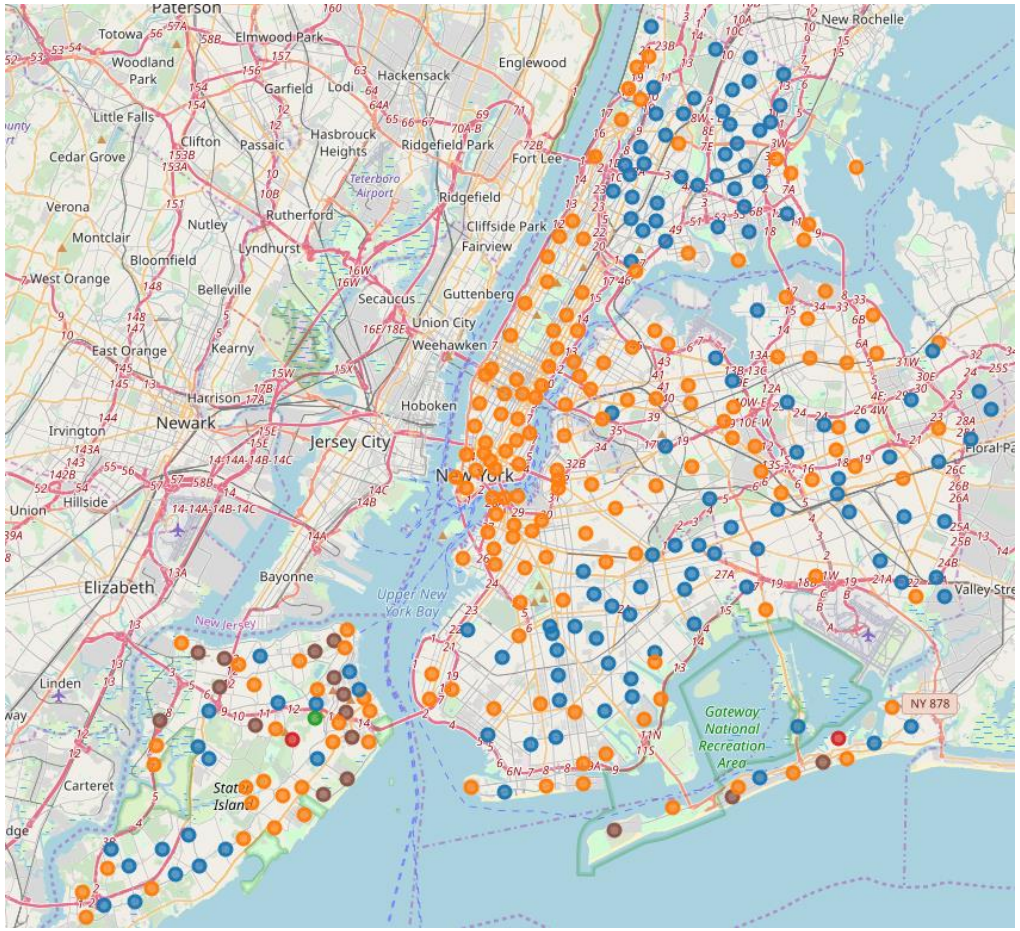
Common top 50 ranking places count:

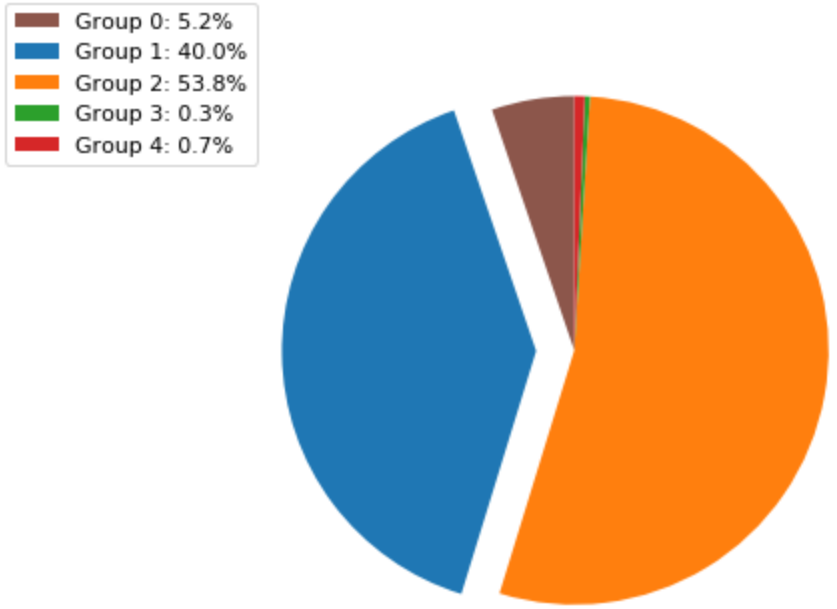
	City	NYC	Toronto	HK
0	NYC	100.00%	58.00%	52.00%
1	Toronto	58.00%	100.00%	50.00%
2	HK	52.00%	50.00%	100.00%

Common top 100 ranking places count:

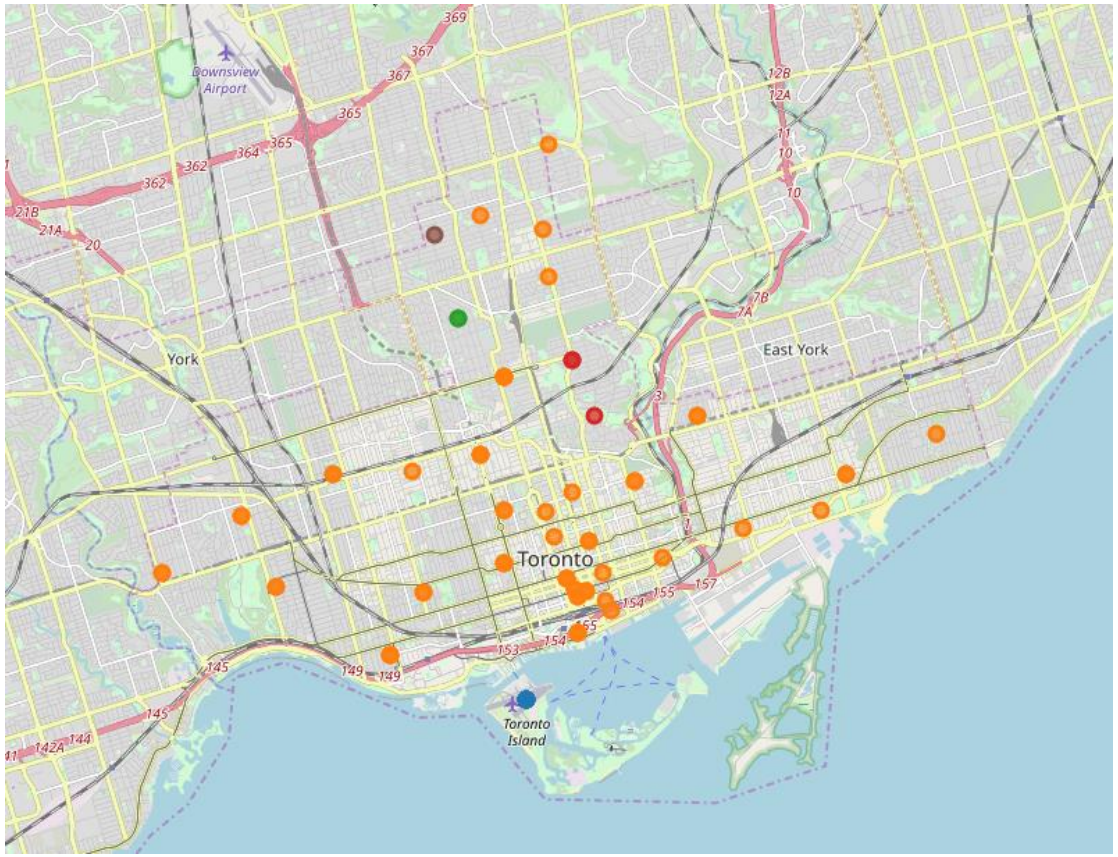
	City	NYC	Toronto	HK
0	NYC	100.00%	66.00%	63.00%
1	Toronto	66.00%	100.00%	63.00%
2	HK	63.00%	63.00%	100.00%

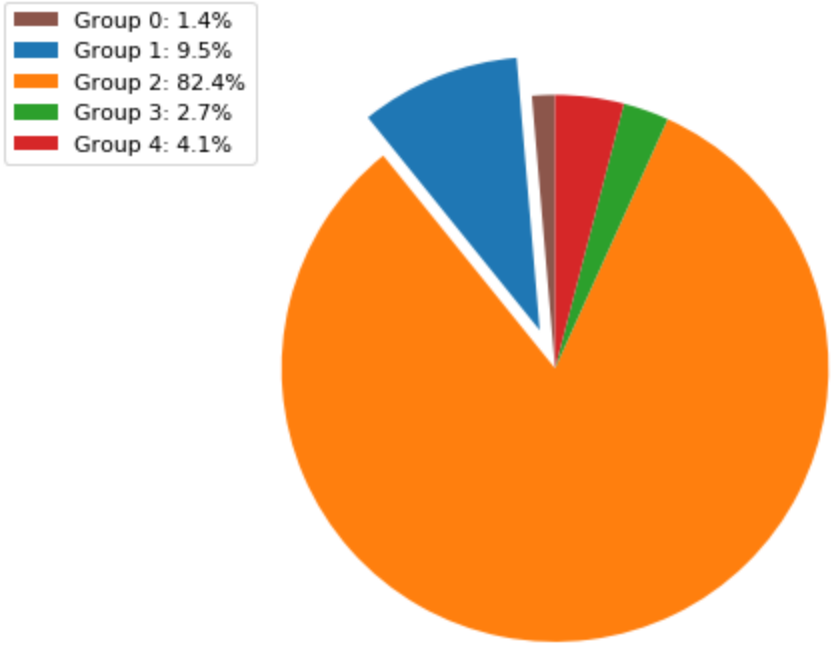
Analysis on Neighborhoods level – New York City



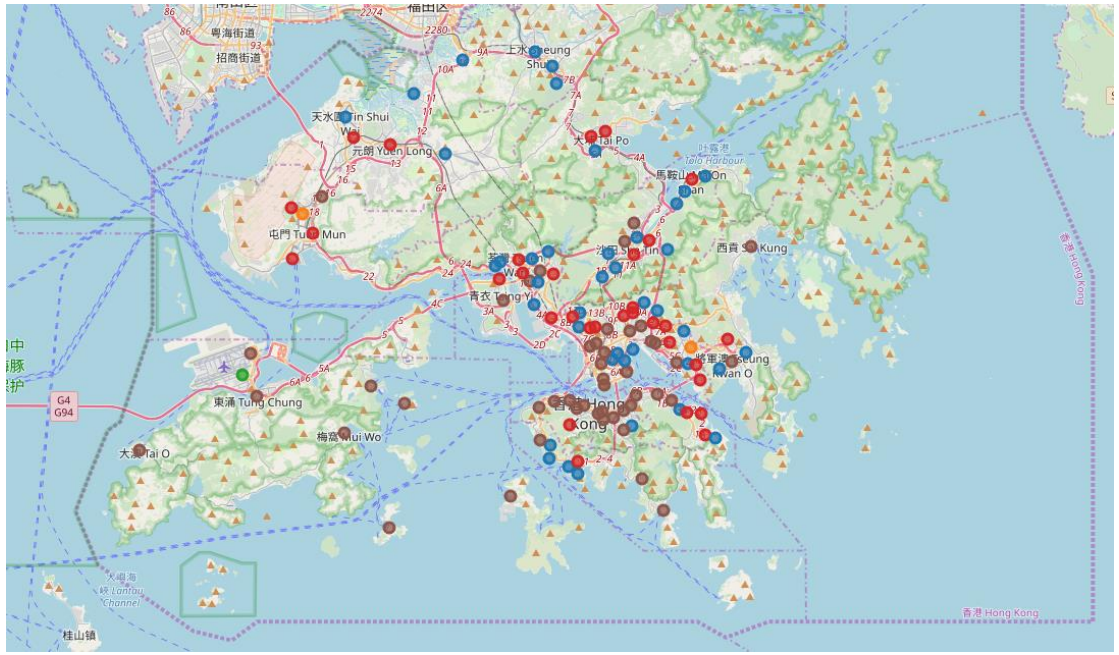


Analysis on Neighborhoods level – Toronto





Analysis on Neighborhoods level – Hong Kong



Discussion - results

- 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.

Discussion

Limitation and Improvement

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

Further Study

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

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.