Applied Data Science Capstone Project

The Battle of Neighbourhoods:
using an Unsupervised Machine Learning Algorithm: KMean clustering
Cindy Yu

Introduction

Toronto and New York City are the financial, entertainment and cultural centres of North America. They are less than 90-mintue flight apart. People sometimes refer Toronto to the “New York North”. However, with location-based data being more accessible to public nowadays, it would be interested in knowing the following questions:

- Whether it’s possible to quantify how similar (or dissimilar) these cities are by utilizing this data?
- Are we able to build a cluster model that captures the city dynamic and characterizes the urban neighbourhood?

This report is a preliminary effort to analyze the dynamic in both cities and compare their similarity. It can be a reading material to people who are planning a move from one city to another but not sure about the uncertainty in changing of environment, or to people who are simply curious about these two famous North American cities. It can also be beneficial for readers who want to have some tourism guidances or would like to gain a sense on neighbourhood planning strategies in these two famous cities. By the end of the report, we will be able to disclose whether Toronto is the “New York North” or not.

Data

For our analysis, three main data sources will be used: location data for New York and Toronto. Venue data for both locations.

1. New York City Neighbourhood Dataset:

This dataset that contains the 5 boroughs in NYC and the neighbourhoods that exist in each borough as well as the the latitude and longitude coordinates of each neighbourhood.

Link: https://geo.nyu.edu/catalog/nyu_2451_34572

Out[3]:

	Borough	Neighborhood	Latitude	Longitude	City
0	Bronx	Wakefield	40.894705	-73.847201	New York
1	Bronx	Co-op City	40.874294	-73.829939	New York

2. Toronto Neighbourhood Dataset:

Similar to the New York City Dataset, this dataset also contains boroughs and the neighbourhoods that exist in each borough in Toronto, as well as the the latitude and longitude coordinates.

Link: *https://en.wikipedia.org/w/index.php title=List_of_postal_codes_of_Canada:_M&oldid=862527922*

Out[2]:

	Borough	Neighborhood	Latitude	Longitude	City
0	North York	Parkwoods	43.753259	-79.329656	Toronto
1	North York	Victoria Village	43.725882	-79.315572	Toronto

3. Foursquare API:

Foursquare is a location technology platform dedicated to collect trusted location and venue data. In this result, it is used to get the top 100 venues within a radius of 500 meters of a neighbourhood. Data was retrieved using API calls.

Link: https://api.foursquare.com/v2/venues

Out[33]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Parkwoods	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy

Methodology

In order to answer the business question, an unsupervised machine learning clustering algorithm (K-Mean) was applied in this project. K-means algorithm is an iterative algorithm famous for partitioning the dataset into Kpre-defined distinct groups (clusters). It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible.

An essential step to implement K-means is to determine the optimal number of clusters (i.e. k). The Elbow Method is one of the standard metrics to specify k. It calculates the sum of squared distances of samples to their closest cluster center for different values of 'k'. The optimal number of clusters is the value after which there is no significant decrease in the sum of squared distances.

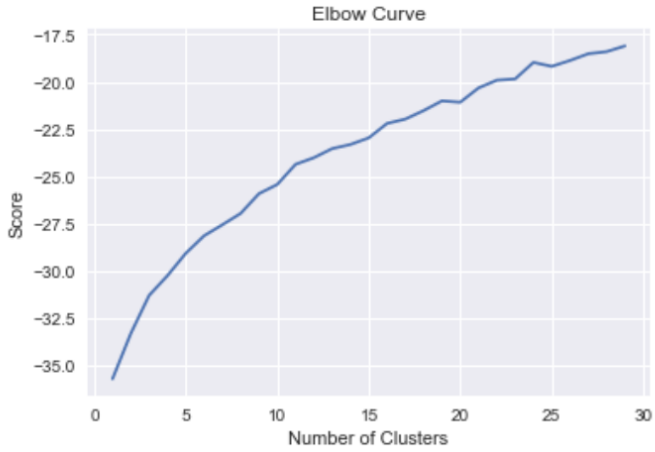

Fig. 1 Elbow Curve metrics to determine the best k in K-Mean algorithm. In this case, the elbow point is roughly around 3-5.

Elbow analysis in Fig.1 suggests that the turning point is not too obvious but we can still see a gradual decrease with k greater than 5. As a consequence, k=5 is choose to segment the data for this project.

Results

With k defined, the k-mean algorithm returns to 5 label types. Please note that Python indexing starting from 0 and thus our analysis will present the segment groups as Type 0,1,2,3 and 4.

Fig.2 shows the number of neighbourhoods in each category. The result clearly indicate that New York City has more neighbourhoods than Toronto. In both cities, almost 90% of the neighbourhoods fall into the category of Type 0 and Type 3. Type 2 neighbourhood is the least common in both cities.

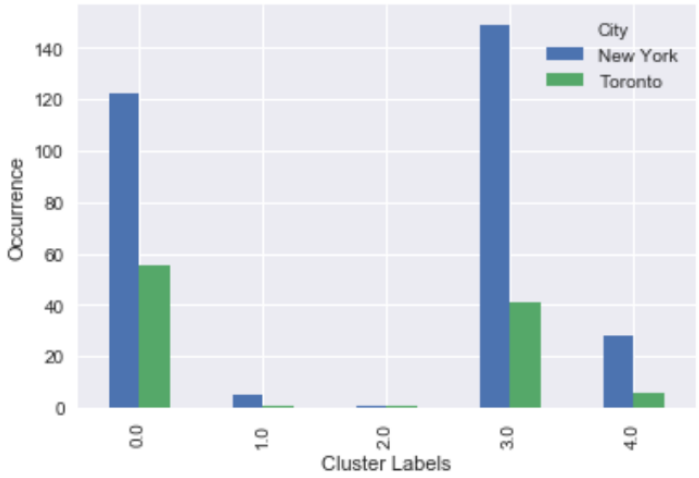

Fig. 2 The number of different types of neighbourhoods in New York City and Toronto.

As shown in Fig. 3, New York's neighbourhood allocation is very clear cut. The Manhattan island is occupied almost exclusively by Type 3 neighbourhoods while the Type 0 neighbourhoods are mostly located in the Brooklyn, Queens, and the Bronx area. Type 4 neighbourhoods are scattered among the Type 0; and we can see that Type 4 is much more common in Staten Island than the rest of New York, which is in line with our expectation.

By comparison, the allocation in Toronto neighbourhood types are more intertwined with one-another without clear lines to separate different areas (Figure 4). Similar as to New York, Type 0 and Type 3 neighbourhoods are the most popular types in the city. Unlike Manhattan, centre of New York City, which is dominated by only one type of neighbourhood (Type 3), Downtown Toronto allows two types of neighbourhoods. In the discussion session, we will explore each neighbourhood type to better understand the implication from the model results.

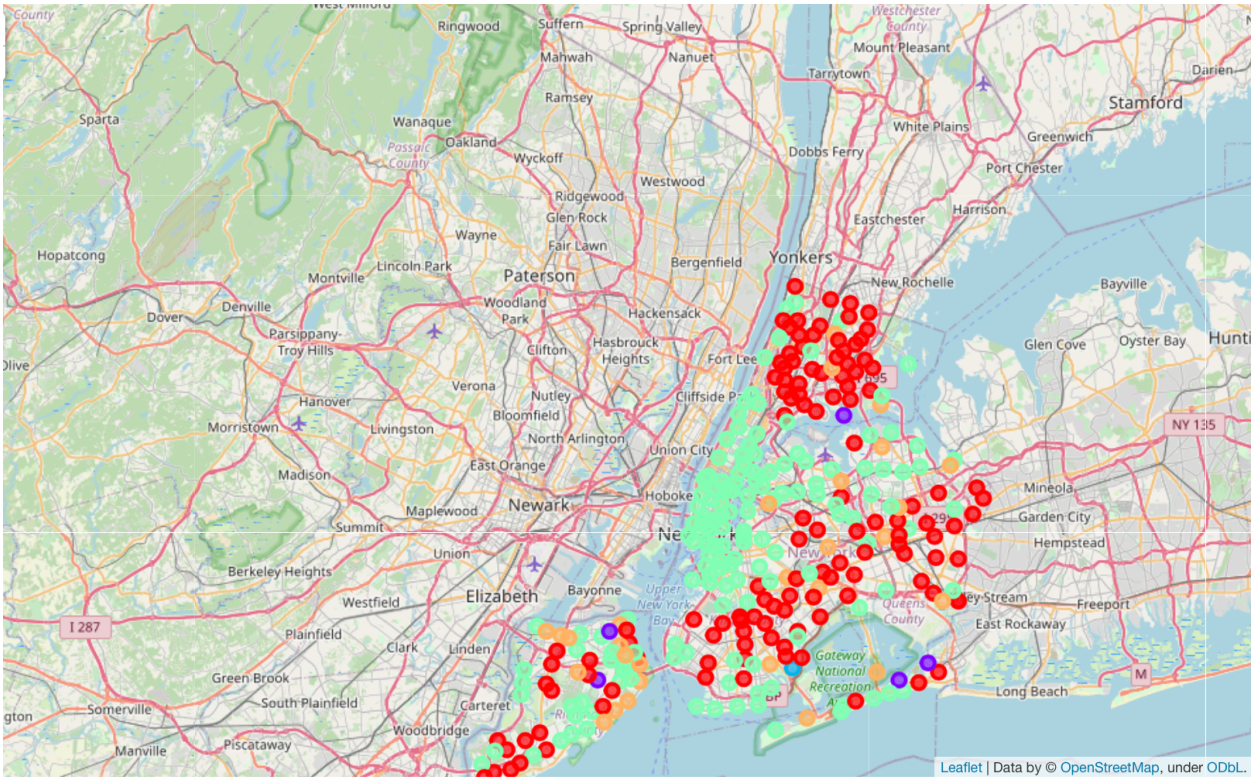

Fig. 3 Cluster results from K-Mean analysis in New York City. Red: cluster label =0; Purple: cluster label =1; Blue: cluster label =2; Green: cluster label=3; Orange: cluster label = 4.

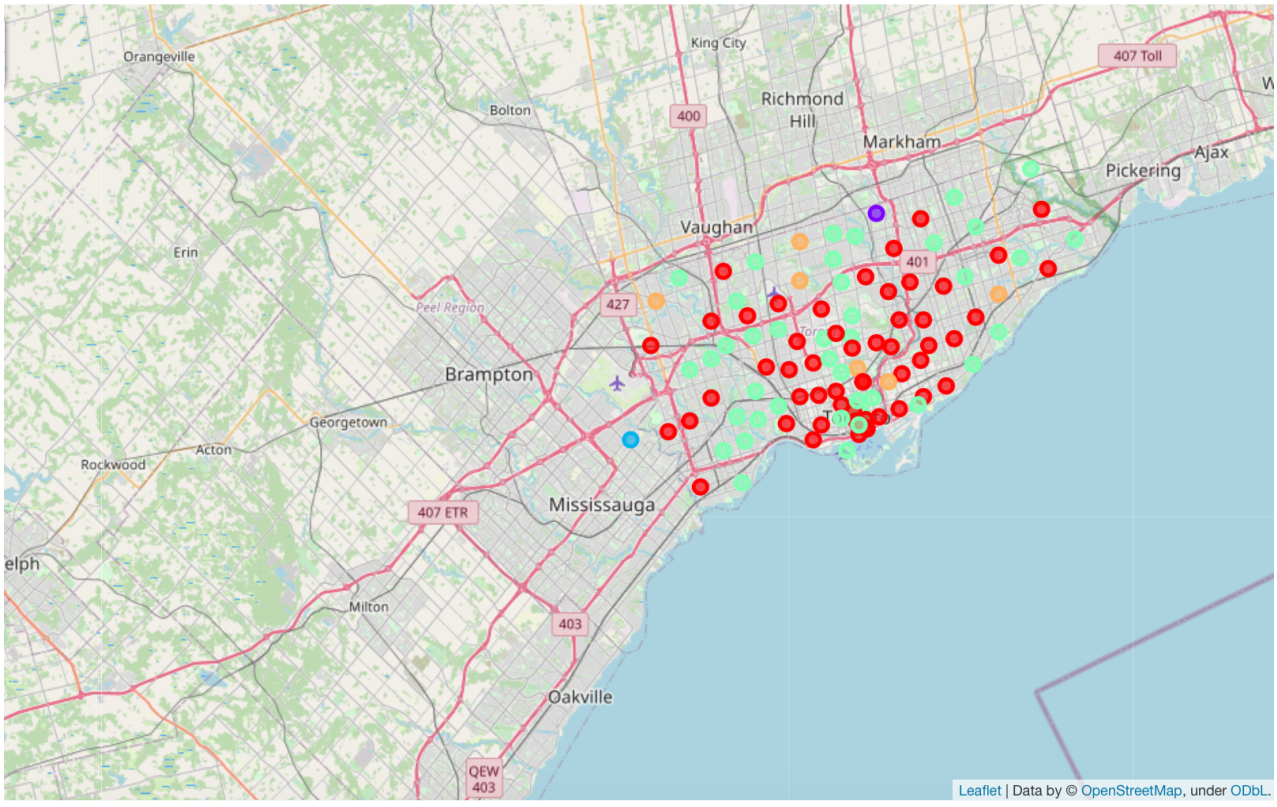

Fig. 4 Cluster results from K-Mean analysis in Toronto. Red: cluster label =0; Purple: cluster label =1; Blue: cluster label =2; Green: cluster label=3; Orange: cluster label = 4.

Discussion

1. Similarity between two cities

Fig.3 and Fig.4 answer the business question with regard to whether it's possible to quantify how similar (or dissimilar) these cities are by utilizing this data.

The model has categorized the neighbourhoods of Toronto and New York into 5 different Types. One common characteristic between the two cities is that Type 0 and Type 3 are the most common types of neighbourhoods in both cities. In New York City, Type 0 and Type 3 accounts for 40% and 43% of the total neighbourhoods respectively. In Toronto, the percentage of Type 0 and Type 3 is approximately 50% and 37%. Given the high similarity in those numbers, it's reasonable to conclude that neighbourhoods in Toronto and New York share quite a lot in common.

2. Implication from cluster model on city dynamic and urban neighbourhood characterization

Judging from cluster output, Type 0 represents neighbourhoods that are surrounded predominantly by restaurants, particularly casual food places such as pizza places, coffee shops, and sandwich places. These neighbourhoods also have a good mix of service amenities such as banks, gyms, and supermarkets. These neighbourhoods are likely residential areas based on these characteristics. The wide variety of street food/restaurant choice suggests New York City and Toronto are the best food cities.

Parks are the most common venues of Type 1 neighbourhoods. Presumably these neighbourhoods are located in the suburban areas of the cities

Type 2 neighbourhoods are considered the outlier group. It consists of only two neighbourhoods with the common characteristic of having "Pool" as the most common venue in the neighbourhood.

Type 3 neighbourhoods are well balanced with various type of amenities including, shops, restaurants, and entertainments such as bars and spas. These neighbourhoods are likely located in the downtown area of the cities. Infused with youth culture, this type of neighbourhood is more dynamic and will keep you entertained at all hours. It would be great to visit for tourism purpose.

Type 4 neighbourhoods, like Type 1, are heavily surrounded by restaurants, however with a focus on higher end restaurants rather than casual food places. It is also more common to see low-density places such as sports fields, art galleries, and beaches. Presumably these neighbourhoods are low-density residential areas.

Decades of immigration have created an invigorating melting pot of people, cultures and food, which is expected as both New York and Toronto are major metropolitans.

Conclusion

Applying the K-mean unsupervised Machine Learning Algorithm on neighbourhoods in Toronto and New York City segments the neighbourhoods to five clusters. The neighbourhood distribution is showing high similarity in pattern, which suggests that New York City and Toronto have similar convergence as a result of the cultural diversity. Meanwhile, they also have their unique characteristics judging by how neighbourhood groups structured.