Applied Data Science Capstone Project

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

Cindy Yu
April 2020

Introduction

Toronto and New York City are the financial, entertainment and cultural centres of North America.

Research 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?

Data

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.

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.

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.

Methodology

1.An unsupervised machine learning clustering algorithm (K-Mean)

K-means: known for making the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible

2. To determine the optimal number of clusters

A gradual decrease with k greater than 5. As a consequence, k=5 is choose to segment the data for this project.

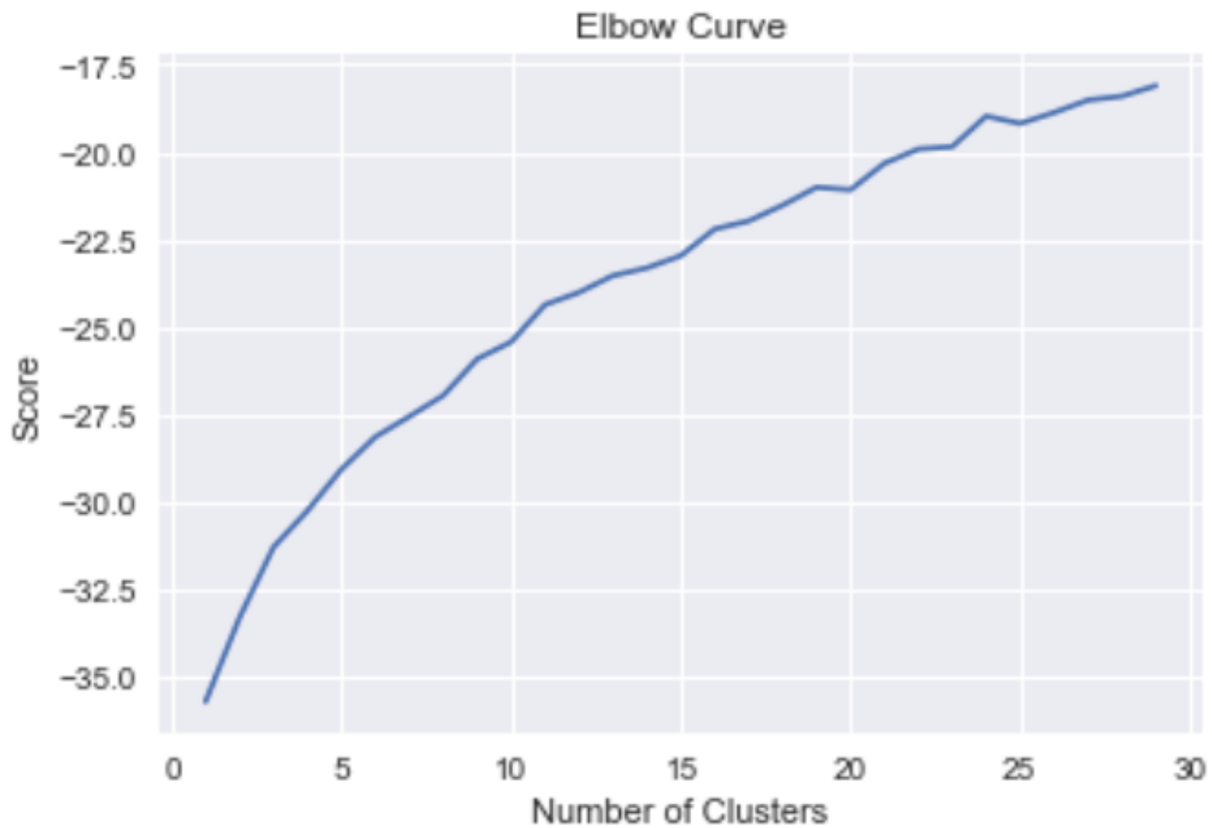

Fig. 1 Elbow Curve metrics to determine the best k in K-Mean algorithm. In this case, the elbow point is around 3-5.

Results & Discussion

The model categorizes the neighbourhoods of Toronto and New York into 5 different types.

Common characteristic: Type 0 and Type 3 are the most common types of neighbourhoods in both cities.

Given the high similarity in the distribution pattern, it's reasonable to conclude that neighbourhoods in Toronto and New York share quite a lot in common.

Neighbourhood Type	Type 3	Type 1	Type 4	Type 2	Type 1
New York City	49%	40%	9.8%	3,87%	0.33%
Toronto	37%	50%	11.1%	0.95%	0.95%

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

Results & Discussion

New York City

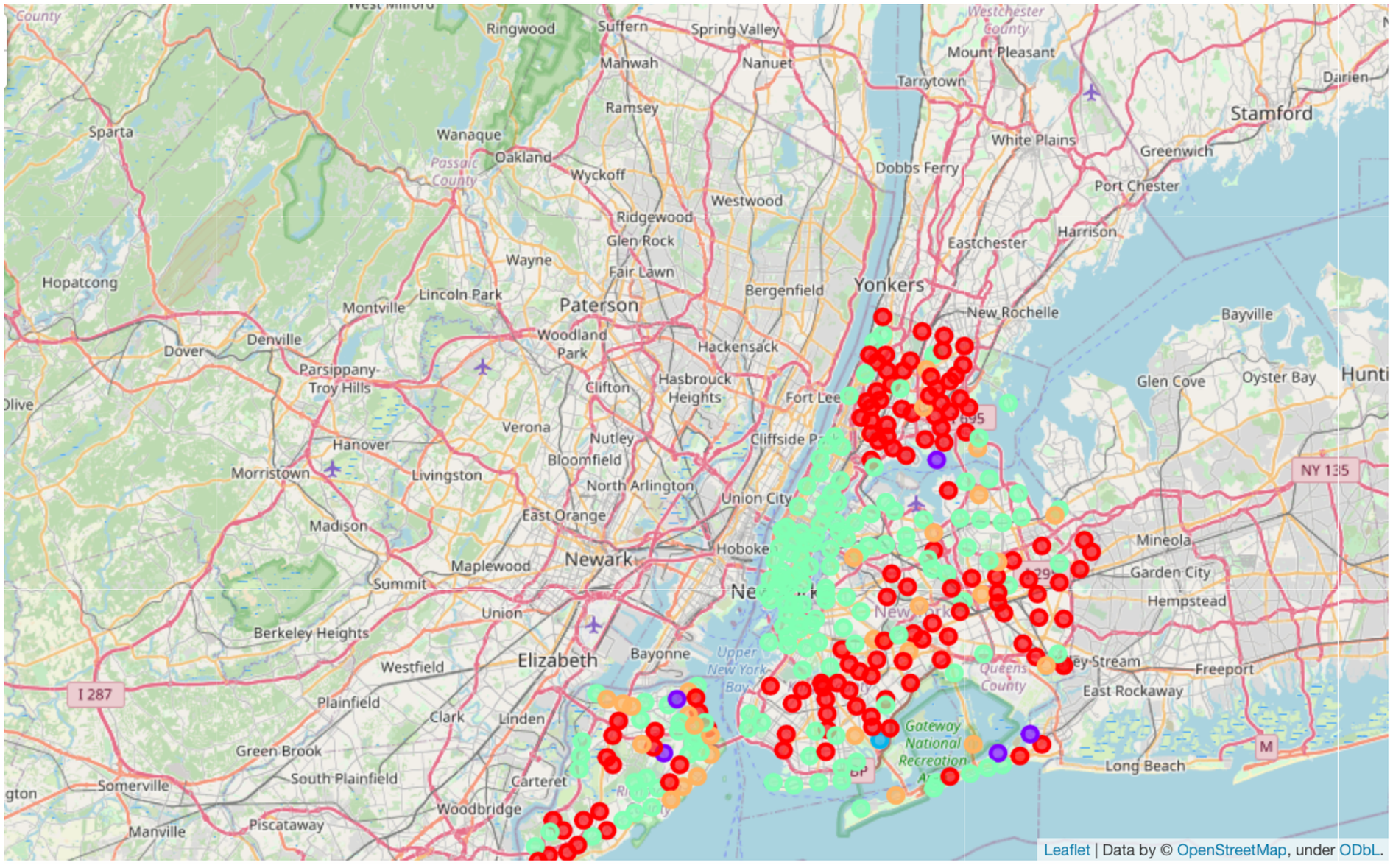

Fig. 3 Cluster results from K-Mean analysis in New York City. Red: cluster label =0; Purple: cluster label =1; Blue: cluster label =2; Green: cluster label=3; Orange: cluster label = 4.

Results & Discussion

Toronto

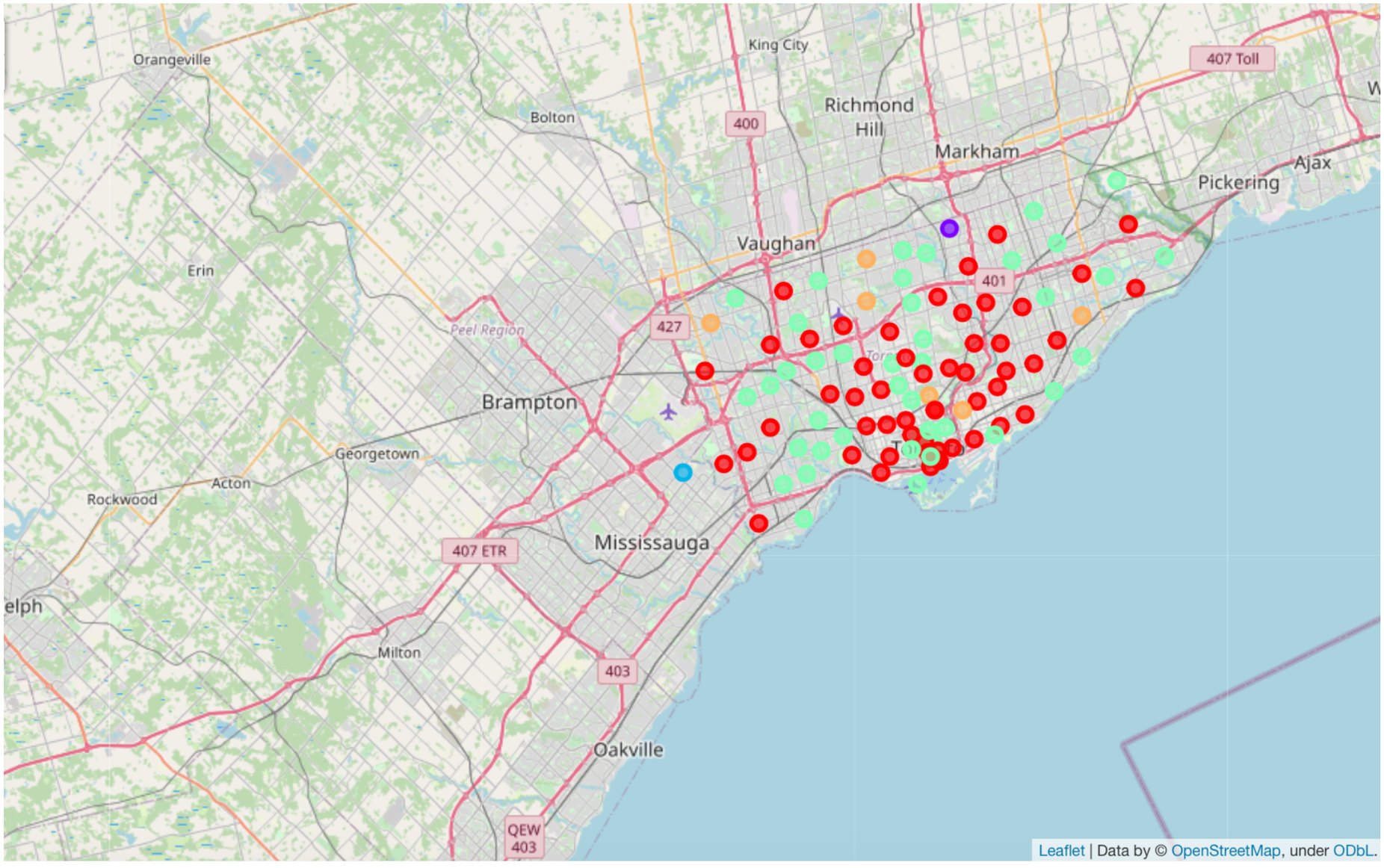

Fig. 4 Cluster results from K-Mean analysis in Toronto. Red: cluster label =0; Purple: cluster label =1; Blue: cluster label =2; Green: cluster label=3; Orange: cluster label = 4.

Results & Discussion

Neighbourhoods analysis -

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

Results & Discussion

Neighbourhoods analysis -

Type 1: 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.

Results & Discussion

Neighbourhoods analysis -

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.

Conclusion

- K-mean unsupervised Machine Learning Algorithm suggests that there are five types of neighbourhood segments in Toronto and New York City.

- 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: New York's neighbourhood allocation is very clear cut; By comparison, the allocation in Toronto neighbourhood types are more intertwined with one-another without clear lines to separate different areas.

Thank you!