

Comparing neighborhoods between New York City and Toronto

Airton Carlos Nunes Raimundo

March 21, 2020

1. Introduction

1.1 Background

In the context of globalization, migratory currents are increasingly strong and people are increasingly experiencing the cultural diversity that the world provides.

However many people have habits of consumption that are part of their lives and they would like to maintain them even during a change of country.

1.2 Problem

The question we will try to answer in this project is: "What is the best way to choose a neighborhood that provides the services that I am used to consuming?"

1.3 Interest

The results presented here aim to reach people in temporary or permanent change with little taste for changes in their essential services. Although we are restricted to Toronto and New York only, the intention of this project is to present a generic way of comparing two major cities and making the best decision.

2. Data acquisition and cleaning

2.1 Data sources

Initially we will need a dataset containing all the neighborhoods in both cities, these datasets can be downloaded [here](#) and [here](#). The second part will be collecting the locations of New York and Toronto using the geopy library to help us visualize our data. To top it off, let's get information from nearby vendors using the Foursquare API.

2.2 Data preparation

The data comes from different sources, so it was necessary to adjust column names and eliminate unnecessary data.

The biggest difficulty encountered was with the Foursquare API because the categories are very detailed and we needed to group them into more comprehensive categories before proceeding to our analysis.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	CFB Toronto	43.737473	-79.464763	Toronto Downsview Airport (YZD)	43.738883	-79.470111	Airport
1	CFB Toronto	43.737473	-79.464763	Ancaster Park	43.734706	-79.464777	Park
2	The Beaches	43.676357	-79.293031	Glen Manor Ravine	43.676821	-79.293942	Trail
3	The Beaches	43.676357	-79.293031	The Big Carrot Natural Food Market	43.678879	-79.297734	Health Food Store
4	The Beaches	43.676357	-79.293031	Grover Pub and Grub	43.679181	-79.297215	Pub

Fig 1: Toronto's neighborhoods and venues

3. Methodology

First, we will explore the data to determine what are the characteristics that make each neighborhood unique, thus being able to differentiate and group them. After that, we will use the k-means algorithm to group similar neighborhoods into clusters based on the categories previously determined to visualize the similarities between the two cities.

4. Exploratory data analysis

4.1 Category selection

For each neighborhood of the two cities we search for all commercial establishments within a radius of 500 meters. After that we group the categories into more generalist groups and select the 22 most representative categories.

Restaurant	321
Coffee Shop	162
Bar	126
Fast Food	96
Park	76
Grocery Store	50
Bakery	45
General Store	45
Gym	29
Clothing Store	25
Supermarket	23
Pharmacy	18
Bank	18
Spa	14
Theater	13
Shopping Mall	13
Bus Line	12
Hotel	12
Tea Room	12
Pet Store	10
Art Gallery	8
Bookstore	8

Fig 2: Selected categories by frequency

4.2 Performing K-means algorithm

To perform k-means clustering we first count the number of occurrences of each category in each neighborhood and then apply the algorithm using $K = [10, 15, 20]$.

The choice of these K's was based on the number of neighborhoods in Toronto that is just over 100.

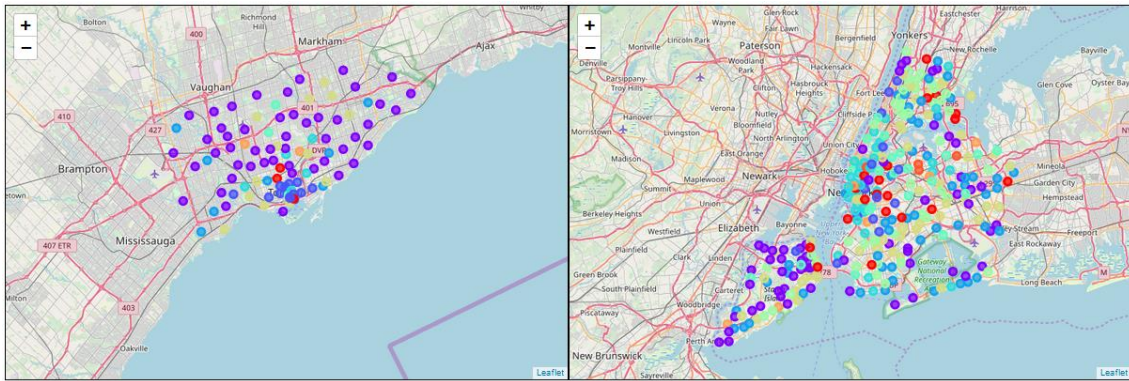


Fig 3: Toronto on the left and New York City on the right for $K = 10$

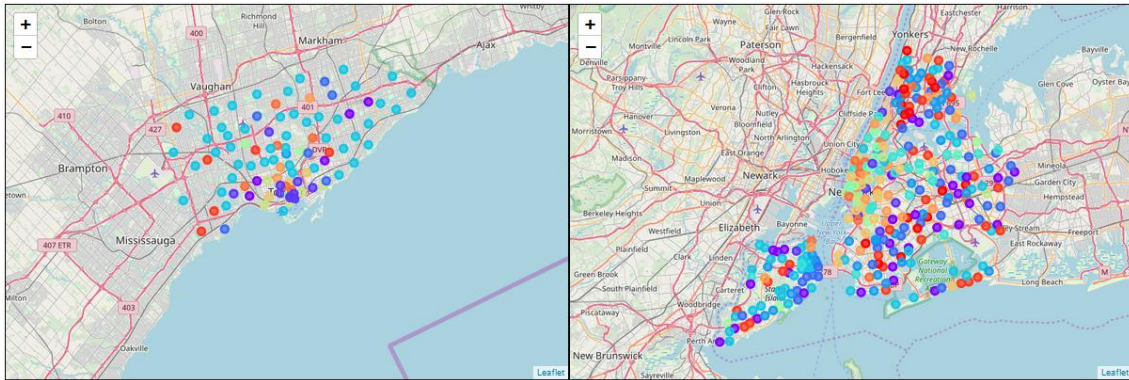


Fig 4: Toronto on the left and New York City on the right for $K = 15$

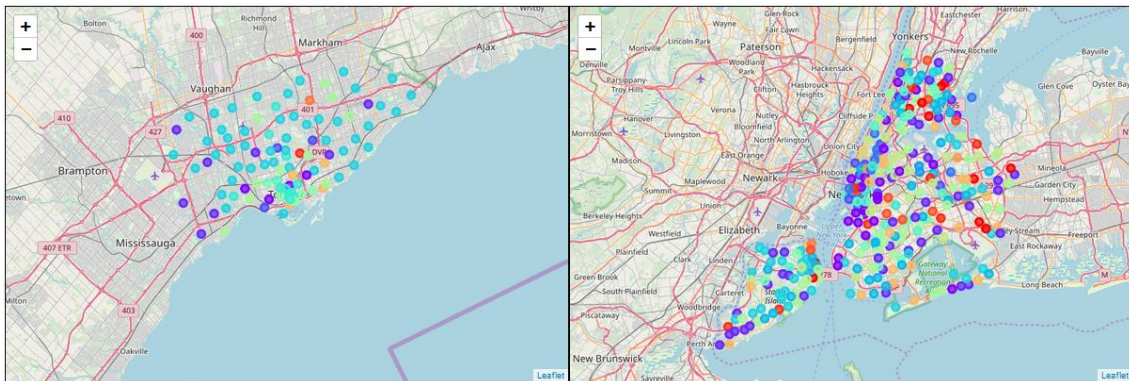


Fig 5: Toronto on the left and New York City on the right for $K = 20$

5. Results and discussion

The results show us that the city of Toronto is very homogeneous in terms of the main services offered by the city and depends very little on the chosen K , on the other hand, New York City has much more variety.

Our graphical presentation of the results allows us to directly conclude which are the most similar neighborhoods between the two cities.

6. Conclusion

This work satisfactorily demonstrated a procedure to compare the neighborhoods of two cities using public databases and machine learning and data visualization tools with the practical objective of choosing the most similar neighborhood.