

2020

# A comparative study between Canadian cities and United- States cities

Olfa BELTAIEF



---

## Summary

1	Introduction.....	3
2	Business Problem .....	3
3	Data acquisition and cleaning .....	3
3.1	Data sources .....	3
3.2	Data cleaning.....	4
3.2.1	First data: Toronto Data Set: .....	4
3.2.2	Second data: New York Data Set .....	5
4	Exploratory Data Analysis.....	6
4.1	Comparing the most popular Venues categories by neighborhood .....	6
4.2	Comparing the top 10 popular Venues categories in New York and Toronto .....	6
4.3	Comparing the 10 least Venues categories in New York and Toronto.....	7
4.4	Comparison between the Venues Categories Distribution of New York and Toronto .....	7
5	Clustering.....	8
6	Conclusions.....	10

## 1 Introduction

Tourism is one of the pillar industries in Toronto and New York; they have accumulated a set of more perfect and mature management system in travel agency and operation mechanism. The comparative study of Toronto and New York will promote our tourism industry management process. The study focuses on the difference of the Toronto and New York's venues categories. Then it puts forward corresponding suggestions according to the present situation of the development of travel industry.

## 2 Business Problem

We suppose that a person who wants to choose the best place to travel during his vacation, so the chosen place should represent his preferences as much as possible. In other words, if the person prefers, for example, Vietnamese Restaurants, then the place chosen should reflect their choice and it should contain the maximum of Vietnamese restaurants. If he hates a specific category, then the chosen place must contain this category as a minimum. This study will help people to choose the right place for their trip. In this context, we chose as a first study, to compare a Canadian city such as Toronto and a United States city such as New York.

It is very important to note that Toronto is one of the largest and the most principal cities in Canada which is situated in the southern part of the province of Ontario. It is a modern beautiful city overlooking Lake Ontario, built on the northwestern shores of the lake.

Besides, New York is the most populous city in the United States. With an estimated 2018 population of 8,398,748 distributed over about 302.6 square miles (784 km<sup>2</sup>), New York is also the most densely populated major city in the United States. Located at the southern tip of the U.S. state of New York, the city is the center of the New York metropolitan area, the largest metropolitan area in the world by urban landmass.

Concretely, the two chosen cities are among the big cities, which explicate our choice.

In this study, we will compare their neighborhoods, the most popular Venues Categories, the least popular Venues categories. Finally we will use the clustering in order to identify the different cluster in each city.

## 3 Data acquisition and cleaning

### 3.1 Data sources

The first source is a list of postal codes in Canada where the first letter is M. Postal codes beginning with M are located within the city of Toronto in the province of Ontario. Only the first three characters are listed, corresponding to the Forward Sortation Area.

Canada Post provides a free postal code look-up tool on its website, via its applications for such smartphones as the iPhone and BlackBerry, and sells hard-copy directories and CD-ROMs. Many vendors also sell validation tools, which allow customers to properly match addresses and postal codes. Hard-copy directories can also be consulted in all post offices, and some libraries. We can find this data set in the following links:

- [https://en.wikipedia.org/wiki/List\\_of\\_postal\\_codes\\_of\\_Canada:\\_M](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)
- [https://github.com/beltaief/Coursera\\_Capstone/blob/master/Geospatial\\_Coordinates.csv](https://github.com/beltaief/Coursera_Capstone/blob/master/Geospatial_Coordinates.csv)

The second source represents the New York City Neighborhood Names point file which was created as a guide to New York City's neighborhoods that appear on the web resource, "New York: A City of Neighborhoods". We can find this data set in the following link:

- [https://geo.nyu.edu/catalog/nyu\\_2451\\_34572](https://geo.nyu.edu/catalog/nyu_2451_34572).

## 3.2 Data cleaning

### 3.2.1 First data: Toronto Data Set:

The Toronto data set was scraped from the web. Once scrapped, we convert the data into a data frame. The following figure represents the first version of our data.

	Postalcode	Borough	Neighborhood
3	M3A	North York	Parkwoods
4	M4A	North York	Victoria Village
5	M5A	Downtown Toronto	Regent Park , Harbourfront
6	M6A	North York	Lawrence Manor , Lawrence Heights
7	M7A	Downtown Toronto	Queen's Park , Ontario Provincial Government
...	...	...	...
161	M8X	Etobicoke	The Kingsway , Montgomery Road , Old Mill North
166	M4Y	Downtown Toronto	Church and Wellesley
169	M7Y	East Toronto	Business reply mail Processing Centre
170	M8Y	Etobicoke	Old Mill South , King's Mill Park , Sunnylea ,...
179	M8Z	Etobicoke	Mimico NW , The Queensway West , South of Bloo...

In the second step, we use another data set to complete the longitude and the Latitude of each Borough. You can find this data set [here](#) . After that we merge the two data sets.

The following figure represents our data after merging (the first five rows):

	Postalcode	Borough	Neighborhood	Latitude	Longitude
3	M3A	North York	Parkwoods	43.7533	-79.3297
4	M4A	North York	Victoria Village	43.7259	-79.3156
5	M5A	Downtown Toronto	Regent Park , Harbourfront	43.6543	-79.3606
6	M6A	North York	Lawrence Manor , Lawrence Heights	43.7185	-79.4648
7	M7A	Downtown Toronto	Queen's Park , Ontario Provincial Government	43.6623	-79.3895
...	...	...	...	...	...
161	M8X	Etobicoke	The Kingsway , Montgomery Road , Old Mill North	43.6537	-79.5069
166	M4Y	Downtown Toronto	Church and Wellesley	43.6659	-79.3832
169	M7Y	East Toronto	Business reply mail Processing Centre	43.6627	-79.3216
170	M8Y	Etobicoke	Old Mill South , King's Mill Park , Sunnylea ,...	43.6363	-79.4985
179	M8Z	Etobicoke	Mimico NW , The Queensway West , South of Bloo...	43.6288	-79.521

In the third step, we have filtered our data set. We only keep the Toronto Borough. The following figure represents the result after filtering.

	Postalcode	Borough	Neighborhood	Latitude	Longitude
5	M5A	Downtown Toronto	Regent Park , Harbourfront	43.6543	-79.3606
7	M7A	Downtown Toronto	Queen's Park , Ontario Provincial Government	43.6623	-79.3895
14	M5B	Downtown Toronto	Garden District , Ryerson	43.6572	-79.3789
23	M5C	Downtown Toronto	St. James Town	43.6515	-79.3754
31	M4E	East Toronto	The Beaches	43.6764	-79.293

Finally, we use Foursquare location data to complete our data with the different venues categories of each neighborhood. The following figure represents our data after cleaning (the first five rows):

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Regent Park , Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
1	Regent Park , Harbourfront	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
2	Regent Park , Harbourfront	43.65426	-79.360636	Morning Glory Cafe	43.653947	-79.361149	Breakfast Spot
3	Regent Park , Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center
4	Regent Park , Harbourfront	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa

### 3.2.2 Second data: New York Data Set

The first data is a json file. The data was downloaded and we convert it to data frame. The following figure represents the first version of our data (the first five rows):

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

In the second step, we use Foursquare location data to complete our data with the different venues categories of each neighborhood. The following figure represents our data after cleaning (the first five rows):

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
1	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	Coffee Shop
4	Marble Hill	40.876551	-73.91066	Dunkin'	40.877136	-73.906666	Donut Shop

## 4 Exploratory Data Analysis

### 4.1 Comparing the most popular Venues categories by neighborhood

The following table displays the top 10 venues for each neighborhood in Toronto. Each neighborhood has a different venues categories distribution. For example, the first common venue category in Berczy Park is Coffee Shop. In the other hand, the first common category in Business reply mail is the Yoga Studio.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Berczy Park	Coffee Shop	Cocktail Bar	Restaurant	Seafood Restaurant	Bakery	Beer Bar	Cheese Shop	Café	Gourmet Shop	Pub
1	Brockton, Parkdale Village, Exhibition Place	Café	Coffee Shop	Breakfast Spot	Bar	Intersection	Nightclub	Grocery Store	Gym	Convenience Store	Furniture / Home Store
2	Business reply mail Processing Centre	Yoga Studio	Skate Park	Fast Food Restaurant	Light Rail Station	Auto Workshop	Spa	Farmers Market	Recording Studio	Pizza Place	Burrito Place
3	CN Tower, King and Spadina, Railway Lands, Har...	Airport Lounge	Airport Service	Airport Terminal	Coffee Shop	Harbor / Marina	Sculpture Garden	Boat or Ferry	Rental Car Location	Plane	Boutique
4	Central Bay Street	Coffee Shop	Italian Restaurant	Café	Sandwich Place	Burger Joint	Bar	Japanese Restaurant	Salad Place	Thai Restaurant	Bubble Tea Shop

The following table displays the top 10 venues for each neighborhood in New York. Each neighborhood has a different venues categories distribution. For example, the first common venue category in Battery Park is Park. In the other hand, the first common category in Central Hariem is the African Restaurant.

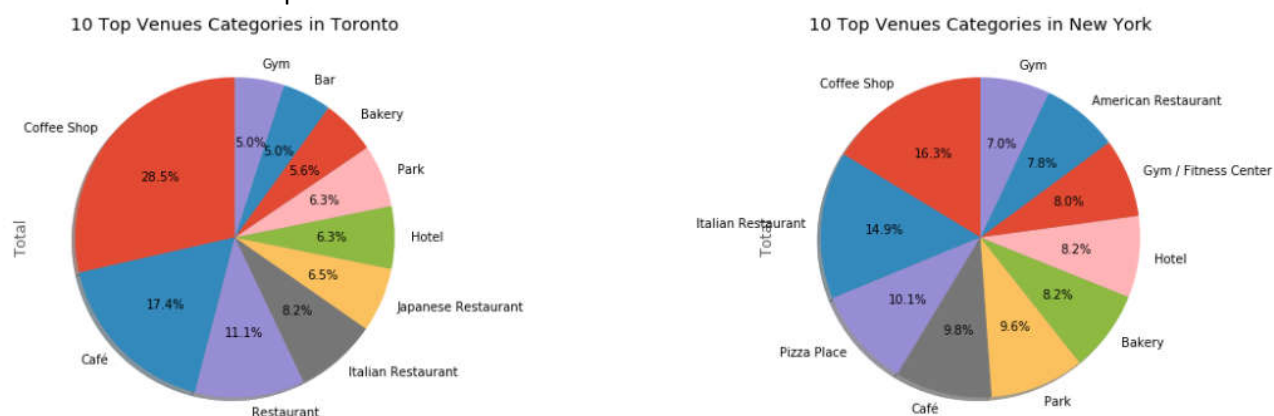
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Battery Park City	Park	Hotel	Memorial Site	Gym	Coffee Shop	Food Court	Beer Garden	Shopping Mall	Gourmet Shop	Boat or Ferry
1	Carnegie Hill	Coffee Shop	Grocery Store	Pizza Place	Japanese Restaurant	Gym / Fitness Center	Gym	Yoga Studio	Café	Bookstore	Bar
2	Central Harlem	African Restaurant	Cosmetics Shop	Chinese Restaurant	French Restaurant	American Restaurant	Bar	Seafood Restaurant	Public Art	Bookstore	Library
3	Chelsea	Art Gallery	Coffee Shop	Italian Restaurant	Ice Cream Shop	Boutique	Cycle Studio	Park	Café	Theater	Market
4	Chinatown	Chinese Restaurant	Bakery	Cocktail Bar	Coffee Shop	American Restaurant	Spa	Optical Shop	Salon / Barbershop	Noodle House	Hotel

### 4.2 Comparing the top 10 popular Venues categories in New York and Toronto

The following figure represents the top popular venues categories en New York and Toronto. In fact,

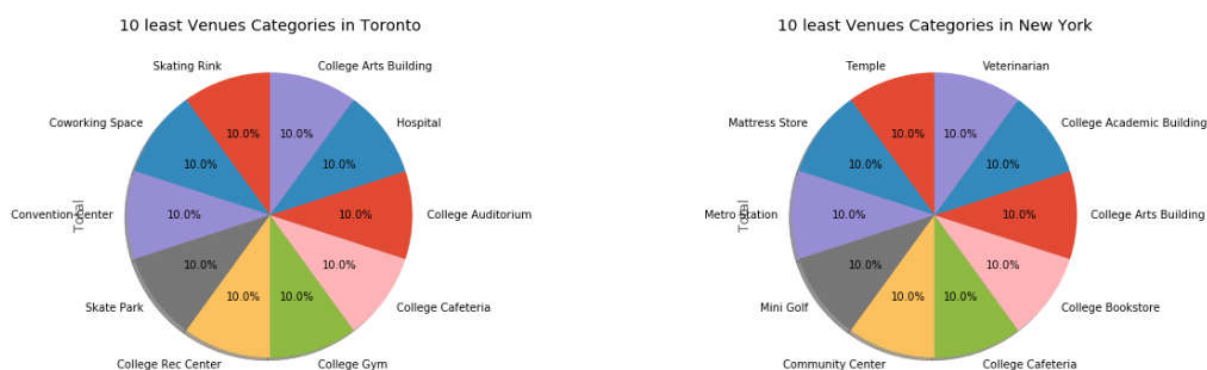


The percentage of coffee shops represents the largest percentage in Toronto and New York. Note that New York has more parks than Toronto.



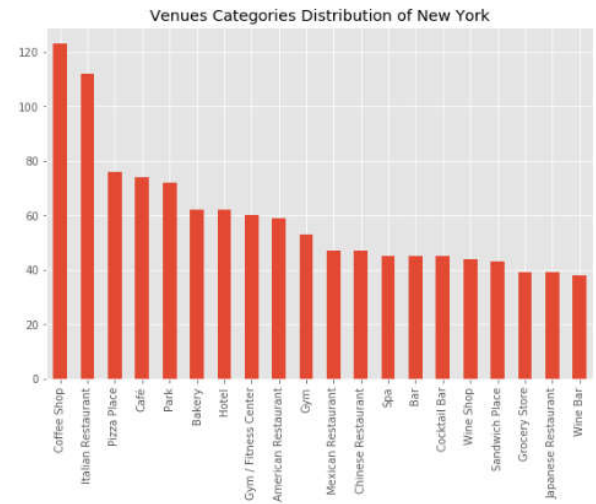
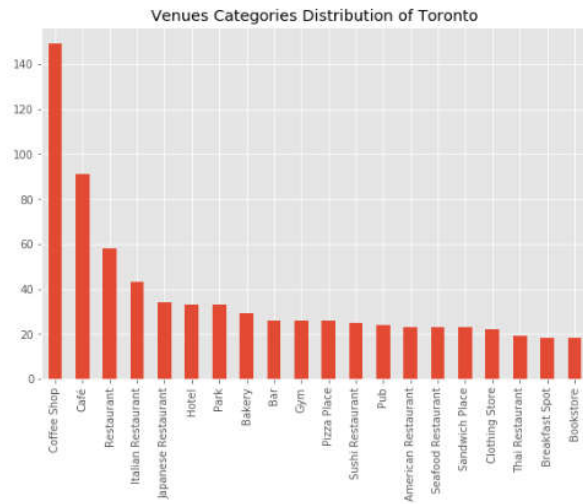
### 4.3 Comparing the 10 least Venues categories in New York and Toronto

The following figure represents the 10 least venues categories in New York and Toronto. We note that the number of hospitals is among the lowest in Toronto. Note also, the metro stations services, the mini golf services are among the least existing services in New York.



### 4.4 Comparison between the Venues Categories Distribution of New York and Toronto

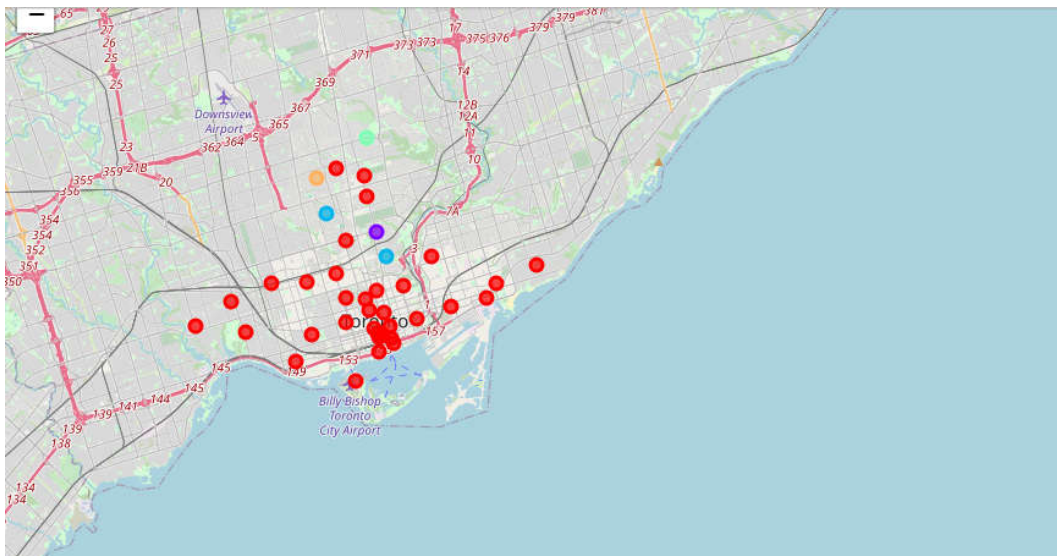
The following figure represents the venues categories distribution of New York and Toronto. We note that the number of coffee shops is almost similar whether in Toronto or New York. Note also that the number of "Italian restaurants" and "Pizza Place" in New York is significantly larger than in Toronto. Overall, New York has more services than Toronto.



## 5 Clustering

The following table represents the Toronto maps with the different cluster.

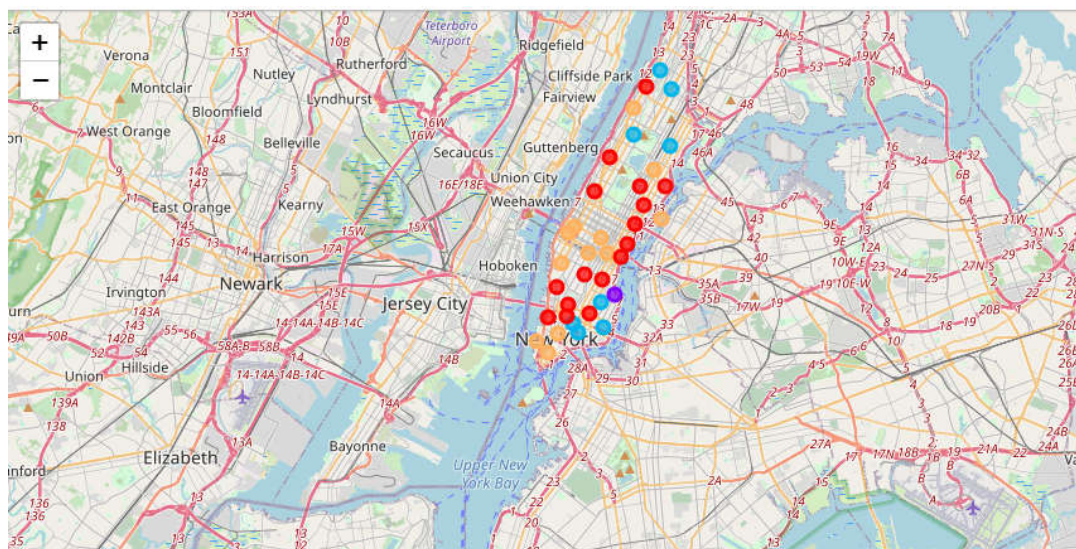
Note that the most popular cluster is the red cluster. In fact, the label of the red cluster is “0”. The most of neighborhoods of the red cluster are defined by the Coffee shop in the most common venue category (see the following table).





	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
5	Downtown Toronto	0	Coffee Shop	Bakery	Park	Pub	Café	Theater	Breakfast Spot	Restaurant	Electronics Store	Espresso Bar
7	Downtown Toronto	0	Coffee Shop	Sushi Restaurant	Beer Bar	Restaurant	Bank	Bar	Café	Diner	Discount Store	Yoga Studio
14	Downtown Toronto	0	Clothing Store	Coffee Shop	Café	Bubble Tea Shop	Cosmetics Shop	Restaurant	Japanese Restaurant	Italian Restaurant	Middle Eastern Restaurant	Tea Room
23	Downtown Toronto	0	Café	Coffee Shop	Gastropub	Cocktail Bar	American Restaurant	Clothing Store	Italian Restaurant	Seafood Restaurant	Beer Bar	Moroccan Restaurant
31	East Toronto	0	Trail	Health Food Store	Pub	Women's Store	Cupcake Shop	Eastern European Restaurant	Donut Shop	Doner Restaurant	Dog Run	Distributive Center
32	Downtown Toronto	0	Coffee Shop	Cocktail Bar	Restaurant	Seafood Restaurant	Bakery	Beer Bar	Cheese Shop	Café	Gourmet Shop	
41	Downtown Toronto	0	Coffee Shop	Italian Restaurant	Café	Sandwich Place	Burger Joint	Bar	Japanese Restaurant	Salad Place	Thai Restaurant	Bubble Shop
42	Downtown Toronto	0	Grocery Store	Café	Park	Candy Store	Italian Restaurant	Diner	Restaurant	Baby Store	Athletics & Sports	Nightlife

The following table represents the New York maps with the different cluster. Note that also the most popular cluster is the red cluster. In fact, the label of the red cluster is “0”. The most of neighborhoods of the red cluster are defined by the Italian Restaurant in the most common venue category (see the following table).



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
5	Manhattanville	Seafood Restaurant	Coffee Shop	Italian Restaurant	Mexican Restaurant	Fried Chicken Joint	Park	Deli / Bodega	Boutique	Supermarket	Spanish Restaurant
8	Upper East Side	Italian Restaurant	Juice Bar	Bakery	Exhibit	Gym / Fitness Center	Hotel	Yoga Studio	Wine Shop	Pizza Place	Cosmetics Shop
9	Yorkville	Coffee Shop	Italian Restaurant	Gym	Bar	Deli / Bodega	Japanese Restaurant	Mexican Restaurant	Sushi Restaurant	Diner	Wine Shop
10	Lenox Hill	Italian Restaurant	Coffee Shop	Pizza Place	Café	Sushi Restaurant	Cocktail Bar	Burger Joint	Gym / Fitness Center	Gym	Mexican Restaurant
12	Upper West Side	Italian Restaurant	Coffee Shop	Wine Bar	Bakery	Bookstore	Bar	Bagel Shop	Mediterranean Restaurant	Thai Restaurant	Seafood Restaurant
13	Lincoln Square	Plaza	Café	Concert Hall	Performing Arts Venue	Gym / Fitness Center	Theater	Italian Restaurant	American Restaurant	Wine Shop	Indie Movie Theater
18	Greenwich Village	Italian Restaurant	Gym	Sushi Restaurant	Wine Bar	Bakery	Coffee Shop	Pizza Place	Gourmet Shop	Dog Run	Dessert Shop
21	Tribeca	Italian Restaurant	Park	Wine Bar	Café	Spa	Art Gallery	Greek Restaurant	Skate Park	Scenic Lookout	Poke Place

## 6 Conclusions

This study can allow travelers to determine the right place to travel depending on their preferences. This study, in fact, shows us the most popular services by neighborhood. She also shows us the least existing venues categories in Toronto and in New York. In this way, the traveler can easily detect the place that best meets their needs. Finally, we showed the different clusters in a map for Toronto and also for New York. It is very important to note that we focused in this study only in the comparison of Toronto and New York .For future study, we can compare between more Canadian cities and more United-States cities.