2020

A comparative study between Canadian cities and United-States cities

Olfa BELTAIEF

Capstone Project

08/05/2020

Summary

1 Introduction

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 km2), 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.

2 Data acquisition and cleaning

2.1 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://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.

2.2 Data cleaning

2.2.1 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

2.2.2 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