

Capstone Project

Clustering and Exploring the Neighborhoods of Toronto and NYC


by Zhaocheng Li

1. Introduction

1-1. Interests

Toronto and New York city are ones of most modern and international cities in the world, and their population and scales grow rapidly. They are all the economical capitals of their respective countries(CANADA/USA). However, the distance between them are surprisingly closer than most people think. It takes only an hour flight to arrive from one to another.

But, it is not common to have two world-class metropolises that are close to each other like Toronto and NYC, in terms of close geographical distance. Which makes me wonder, is it possible for Toronto and NYC to have more similarities? The similarities/dissimilarities can be discussed in fields of population distribution, personal wealth distribution, shops distribution, and race distribution, etc..



1-2. Business Problems

Many business problems will rise after this problem. Once we understand the similarities/dissimilarities of neighborhoods of Toronto & of NYC, we can reveal some very important characteristics such as multiculture(races/languages), population distribution and income per capita differential. The stakeholders can decide the optimized location for business based on these factors. For example, if we want to open a chinese restaurant, we want to choose a location not just in a business district, but most importantly, a location with a high asian population, and maybe with big asian markets.



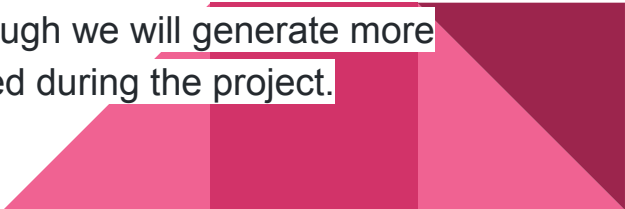
2. Data acquisition and cleaning

2-1. Data acquisition

To accomplish this project, we acquire datasets from many sources:

1. [Foursquares API](#), where we scrape accurate location data of venues for exploration.
2. [List of postal codes of Canada: M from Wikipedia](#), where we obtain the info about boroughs and neighborhoods of Toronto. This dataset will be converted into pandas dataframe and used for exploring neighborhoods around Toronto.
3. [NYC geo location data from IBM](#). Similarly, we will use this JSON file to load NYC boroughs/neighborhoods dataset and explore it.

These are all start-up datasets we collect online. As the project goes through we will generate more datasets used for research. They will be mentioned and explicitly explained during the project.



2-2. Data cleaning

After acquiring the datasets, we clean and format the datasets into the *pandas* dataframes.

`df_toronto`, the source dataset for the location data of Toronto, including columns:

- `Borough`, borough name
- `Neighbourhood`, the neighbourhood name
- `Latitude`, the latitude of the the neighbourhood location
- `Longitude`, the longitude of the neighbourhood location

`df_nyc`, with the similar structure, is the source dataset for the location data of New York city, including columns:



df_toronto dataframe example

(210, 4)

	Borough	Neighbourhood	Latitude	Longitude
0	North York	Parkwoods	43.753259	-79.329656
1	North York	Victoria Village	43.725882	-79.315572
2	Downtown Toronto	Harbourfront	43.654260	-79.360636
3	North York	Lawrence Heights	43.718518	-79.464763
4	North York	Lawrence Manor	43.718518	-79.464763

dft & dfn, the source dataset population details and economy information of Toronto & NYC

(175, 13)

	Name	FM	Census Tracts	Population	Land area (km2)	Density (people/km2)	% Change in Population since 2001	Average Income	Transit Commuting %	% Renters	Second most common language (after English) by name	
0	Toronto CMA Average	NaN	All	5113149	5903.63	866	9.0	40704	10.6	11.4	NaN	
1	Agincourt	S	0377.01, 0377.02, 0377.03, 0377.04, 0378.02, 0...	44577	12.45	3580	4.6	25750	11.1	5.9	Cantonese (19.3%)	
2	Aldenwood	E	0211.00, 0212.00	11656	4.94	2360	-4.0	35239	8.8	8.5	Polish (6.2%)	
3	Alexandra Park	OCOT	0039.00	4355	0.32	13609	0.0	19687	13.8	28.0	Cantonese (17.9%)	
4	Allenby	OCOT	0140.00	2513	0.58	4333	-1.0	245592	5.2	3.4	Russian (1.4%)	

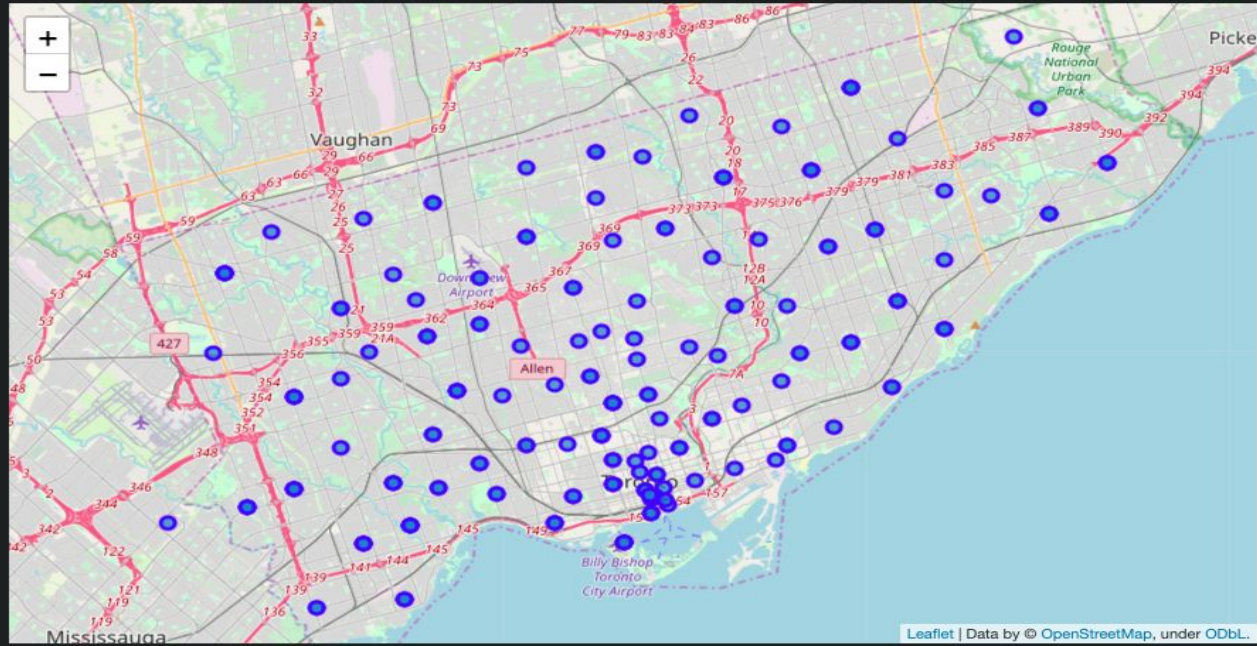
3. Methodology

Using the Foursquare API we can create maps of neighborhoods for both Toronto and NYC:



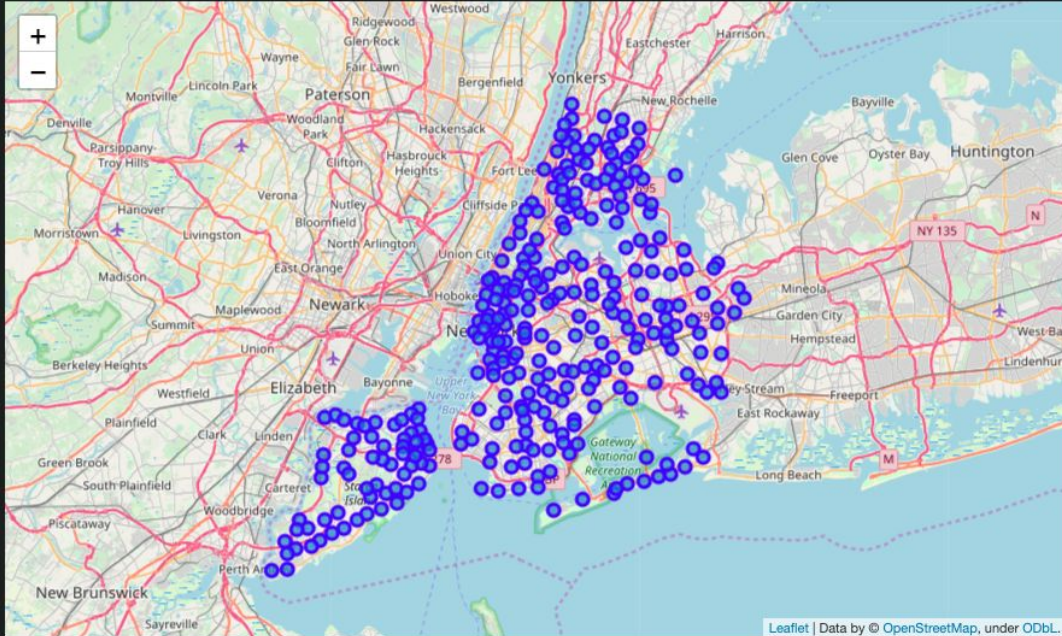
map_toronto

map_toronto



map_nyc

map_newyork

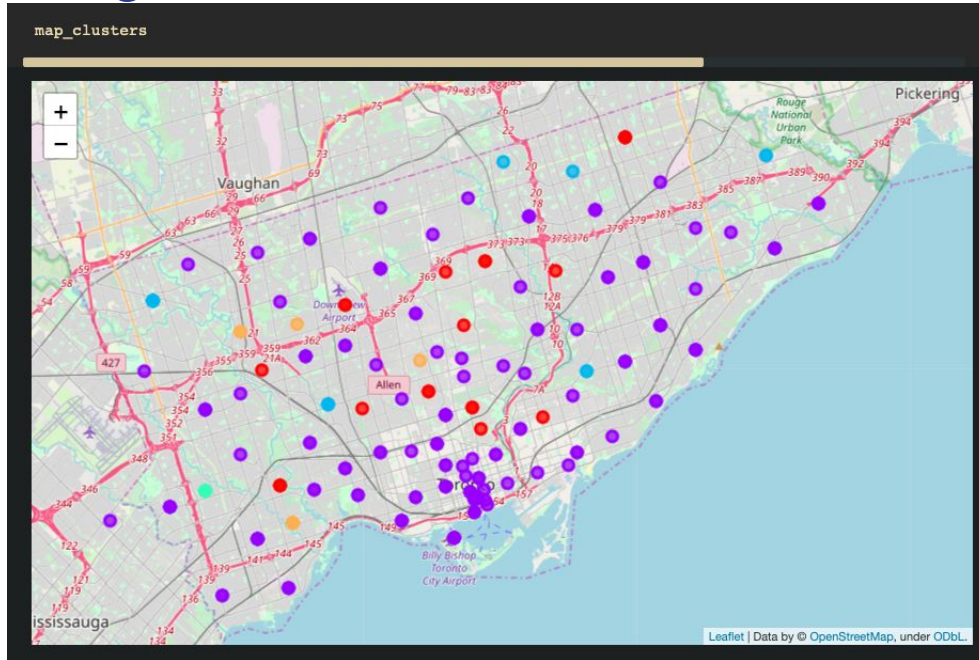


And we can also get the top ten venues for each neighborhood for both cities: for example,

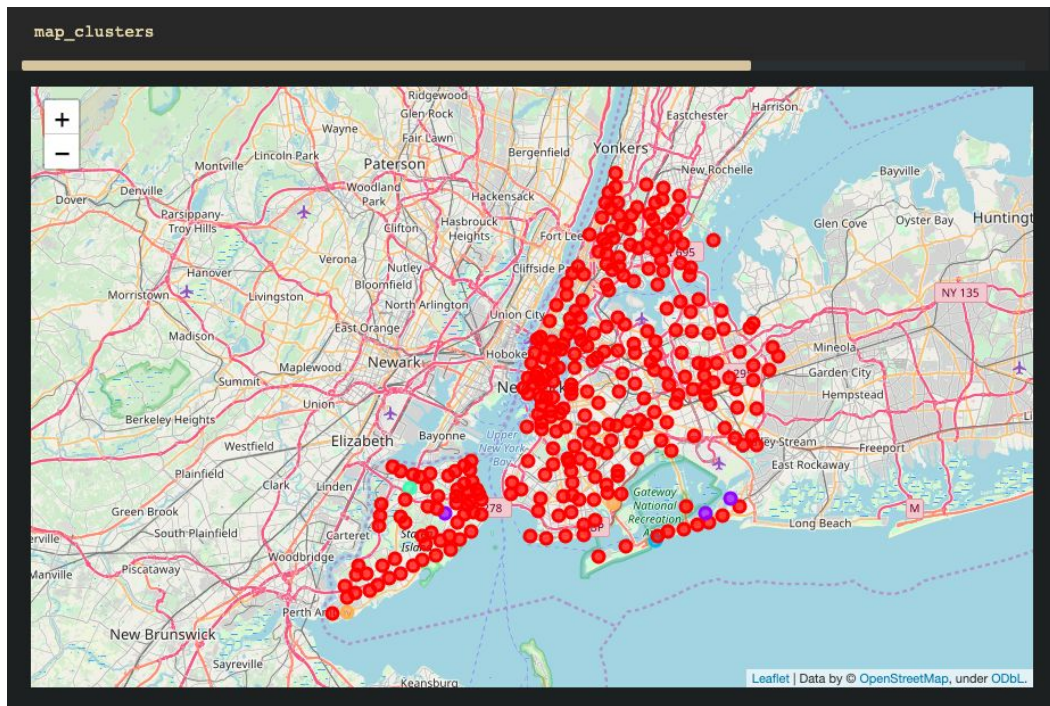
```
neighborhoods_venues_sorted.head()
```

	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	Adelaide	Coffee Shop	Café	Bar	Steakhouse	Cosmetics Shop	American Restaurant	Restaurant	Bakery	Sushi Restaurant	Asian Restaurant
1	Agincourt	Lounge	Latin American Restaurant	Skating Rink	Breakfast Spot	Ethiopian Restaurant	Empanada Restaurant	Electronics Store	Event Space	Dim Sum Restaurant	Eastern European Restaurant
2	Agincourt North	Playground	Park	Coffee Shop	Women's Store	Donut Shop	Dim Sum Restaurant	Diner	Discount Store	Dog Run	Doner Restaurant
3	Albion Gardens	Pizza Place	Beer Store	Fried Chicken Joint	Japanese Restaurant	Fast Food Restaurant	Discount Store	Pharmacy	Sandwich Place	Grocery Store	Airport Terminal
4	Alderwood	Pizza Place	Pub	Gym	Pharmacy	Sandwich Place	Coffee Shop	Skating Rink	Department Store	Dessert Shop	Dim Sum Restaurant

And our main focus is by applying machine learning: clustering, and k-means cluster, we want to get the clusters of neighborhoods of two cities.



nyc_clusters



And another topic we interested is the information about
multiculture, income, population.

We can accomplish this by cleaning and formatting datasets

`dfn, dft`



Toronto data

```
In [9]: df_toronto.head(20)
```

	neighborhoods	population	income_per_capita	land_area(km2)	density(person/km2)	location
1	Agincourt	44577	25750	12.45	3580	toronto
2	Alderwood	11656	35239	4.94	2360	toronto
3	Alexandra Park	4355	19687	0.32	13609	toronto
4	Allenby	2513	245592	0.58	4333	toronto
5	Amesbury	17318	27546	3.51	4934	toronto
6	Armour Heights	4384	116651	2.29	1914	toronto
7	Banbury	6641	92319	2.72	2442	toronto
8	Bathurst Manor	14945	34169	4.69	3187	toronto
9	Bay Street Corridor	4787	40598	0.11	43518	toronto
10	Bayview Village	12280	46752	4.14	2966	toronto
11	Bayview Woods & Steeles	13298	41485	4.07	3267	toronto
12	Bedford Park	13749	80827	2.27	6057	toronto
13	Bendale	28945	29723	8.49	3409	toronto
14	Birch Cliff	12266	48965	3.48	3525	toronto
15	Bloor West Village	5175	55578	0.74	6993	toronto
16	Bracondale Hill	5343	41605	0.62	8618	toronto
17	Branson	8017	27156	1.25	6414	toronto
18	Bridle Path	1540	314107	3.46	445	toronto
19	Brockton	9039	27260	1.10	8217	toronto
20	Cabbagetown	11120	50398	1.40	7943	toronto

NYC data

```
In [10]: df_nyc
```

	borough	population	income_per_capita	land_area(km2)	density(person/km2)	location
0	The Bronx	1471160	19570	109.04	13231	nyc
1	Brooklyn	2648771	23900	183.42	14649	nyc
2	Manhattan	1664727	378250	59.13	27826	nyc
3	Queens	2358582	31310	281.09	8354	nyc
4	Staten Island	479458	23460	151.18	3132	nyc

4. Results and discussion

Based on all kinds of analysis plots, we can have various observations by comparing the performance of neighborhoods of Toronto and NYC

1. It can be noticed that Toronto has one big cluster and a smaller one, while others are insignificant compared to them. As for NYC, there are two big and one mid size clusters, while other two clusters are insignificant compared to them. Therefore, we see Toronto seems to have more uniform neighborhood type. New York has much more varieties. So, the two cities are different in terms of segmentation.



1. NYC and Toronto are similar in population, around 7 million, more or less, and the race ratios are also similar.
2. for the incomes, there are relatively big differential between boroughs for each city. In Toronto, the people working in central Toronto areas has 2 or 3 times more income than incomes of people from other boroughs. Meanwhile, in NYC, people in Manhattan earn dramatically more than people from other area. For example, people in Manhattan earn around 320,000 USD per capital, while people in Brooklyn earn 30,000 USD per capital.



5. Conclusion

Therefore, from my research, Toronto and NYC are different in segmentation, but they are similar in multicultural, population and income distributions.

This research still need more work to be done. Thank you very much for reviews and advices!!!!

