Suggesting the moving options for people from NYC to Toronto

1. Introduction

1.1 Background

Several people who are living in NYC right now plan move to Toronto. All of them are very satisfied of the location they are in NYC. Help them to find the similar locations in Toronto.

1.2 Problem description

The person right now lives in different places NYC. They trying to find similar convenient place in Toronto. Based on the similar venues in the area, suggest the options for them to move.

2. Data acquisition and cleaning

2.1 Data sources

Neighborhoods of NYC data is from the json file from course, which contains 5 boroughs and 306 neighborhoods (it's called: yu_2451_34572-geojson.json), it serves as a reference for comparison with the desired future location in Toronto.
Neighborhoods of Toronto data is from the WikiPedia (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M).

2.2 Data cleaning

Data of NYC is arranged by json to dataframe, and only pick 'Borough', 'Neighborhood', 'Latitude', 'Longitude' from it.

	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

Data of Toronto is obtained by BeautifulSoup, and used postal code to match the neighborhood with their latitude and longitude.

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

2.3 Data from Foursquare

From Foursquare, I explored the neighborhood of NYC and Toronto.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
2	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
3	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	Sun Life	43.754760	-79.332783	Construction & Landscaping
2	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
3	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
4	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop

3. Exploratory Data Analysis

3.1 Count the venues in neighborhood

Count the venues in NYC

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Allerton	29	29	29	29	29	29
Annadale	12	12	12	12	12	12
Arden Heights	4	4	4	4	4	4
Arlington	4	4	4	4	4	4
Arrochar	20	20	20	20	20	20
Arverne	18	18	18	18	18	18
Astoria	92	92	92	92	92	92
Astoria Heights	12	12	12	12	12	12
Auburndale	18	18	18	18	18	18
Bath Beach	47	47	47	47	47	47
Battery Park City	62	62	62	62	62	62

Taking the mean of the frequency of occurrence of each category

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Animal Shelter	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum
0	Allerton	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1	Annadale	0.000000	0.000000	0.000000	0.000000	0.083333	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
2	Arden Heights	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
3	Arlington	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
4	Arrochar	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
5	Arverne	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
6	Astoria	0.000000	0.000000	0.000000	0.000000	0.010870	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
7	Astoria Heights	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
8	Auburndale	0.000000	0.000000	0.000000	0.000000	0.055556	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

Then figure out the most common venues in the neighborhood

	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	Allerton	Pizza Place	Deli / Bodega	Supermarket	Department Store	Fried Chicken Joint	Spanish Restaurant	Breakfast Spot	Gas Station	Fast Food Restaurant	Grocery Store
1	Annadale	Pizza Place	Train Station	Bakery	Dance Studio	American Restaurant	Sports Bar	Diner	Restaurant	Cosmetics Shop	Pharmacy
2	Arden Heights	Pharmacy	Coffee Shop	Bus Stop	Pizza Place	Financial or Legal Service	Exhibit	Eye Doctor	Factory	Falafel Restaurant	Farm
3	Arlington	Bus Stop	Deli / Bodega	Coffee Shop	Yoga Studio	Fishing Spot	Factory	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant
4	Arrochar	Bus Stop	Italian Restaurant	Deli / Bodega	Supermarket	Middle Eastern Restaurant	Food Truck	Liquor Store	Outdoors & Recreation	Bagel Shop	Sandwich Place

Do the same process for Toronto.

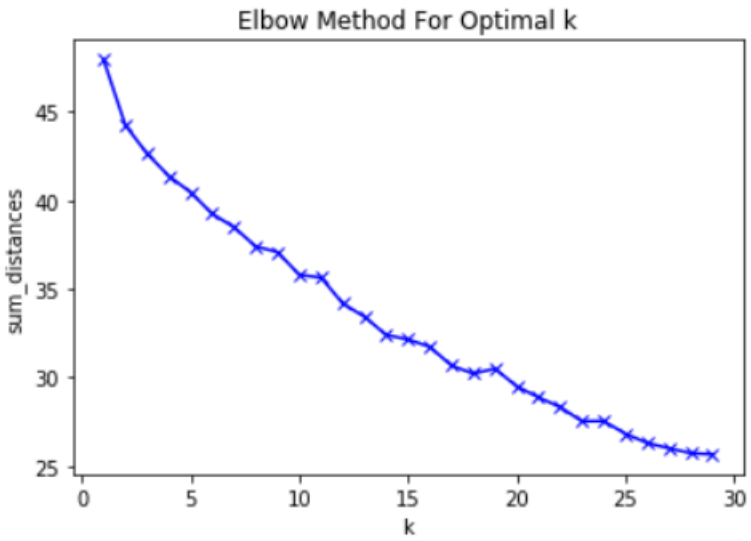
3.2 Cluster Neighborhoods

Because I was comparing to cities, so I needed to combine the data of these 2 cities together.
And I added the city column for later separate the neighborhoods in Toronto.

	Neighborhood	City	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Anim Shel
0	Allerton	NYC	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.
1	Annadale	NYC	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.083333	0.
2	Arden Heights	NYC	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.
3	Arlington	NYC	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.
4	Arrochar	NYC	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.
5	Arverne	NYC	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.
6	Astoria	NYC	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.010870	0.
7	Astoria Heights	NYC	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.
8	Auburndale	NYC	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.055556	0.

3.3 Determine the optimal k

Try different k values to find out which k works best for clustering.



And I found out the best k is 25.

3.4 Cluster the neighborhood

Run k-means to cluster the neighborhood into 25 clusters.

Cluster 1

```
neighborhoods_venues_sorted.loc[neighborhoods_venues_sorted['Cluster Labels'] == 0, neighborhoods_venues_sorted.columns[[1] + li
```

	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	City
198	Oakwood	Lawyer	Bar	Bus Stop	Yoga Studio	Factory	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant	Field	NYC
290	Williamsbridge	Caribbean Restaurant	Nightclub	Bar	Soup Place	Eye Doctor	Factory	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant	NYC
56	Northwest	Rental Car Location	Bar	Drugstore	Yoga Studio	Doner Restaurant	Diner	Discount Store	Distribution Center	Dog Run	Donut Shop	Toronto
57	Northwood Park, York University	Caribbean Restaurant	Massage Studio	Coffee Shop	Bar	Yoga Studio	Donut Shop	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Toronto
67	Rouge Hill, Port Union, Highland Creek	Moving Target	Bar	Yoga Studio	Donut Shop	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Drugstore	Dim Sum Restaurant	Toronto

Cluster 2

```
es_sorted['Cluster Labels'] == 1, neighborhoods_venues_sorted.columns[[1] + list(range(2, neighborhoods_venues_sorted.shape[1]))]
```

	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	City
4	Arrochar	Bus Stop	Italian Restaurant	Deli / Bodega	Supermarket	Middle Eastern Restaurant	Food Truck	Liquor Store	Outdoors & Recreation	Bagel Shop	Sandwich Place	NYC
20	Belle Harbor	Beach	Pub	Spa	Deli / Bodega	Mexican Restaurant	Chinese Restaurant	Pharmacy	Donut Shop	Construction & Landscaping	Bakery	NYC
22	Belmont	Italian Restaurant	Pizza Place	Deli / Bodega	Bakery	Bank	Donut Shop	Grocery Store	Dessert Shop	Mexican Restaurant	Food & Drink Shop	NYC
25	Blissville	Deli / Bodega	Donut Shop	Rental Service	Hotel	Sporting Goods Shop	Movie Theater	Skating Rink	Clothing Store	Mattress Store	Bar	NYC
30	Briarwood	Deli / Bodega	Mexican Restaurant	Arts & Crafts Store	Fast Food Restaurant	Indian Restaurant	Playground	Plaza	Coffee Shop	Bus Station	Farm	NYC
32	Broad Channel	Pizza Place	Deli / Bodega	Clothing Store	Dive Bar	Sporting Goods Shop	Other Nightlife	Yoga Studio	Factory	Falafel Restaurant	Farm	NYC
86	Edgewater Park	Italian Restaurant	Pizza Place	Deli / Bodega	Japanese Restaurant	Asian Restaurant	Chinese Restaurant	Park	Coffee Shop	Bar	Liquor Store	NYC
88	Elm Park	Italian Restaurant	Cosmetics Shop	Chinese Restaurant	Toll Plaza	Pizza Place	Café	Ice Cream Shop	Laundry Service	Bus Stop	American Restaurant	NYC
159	Madison	Deli / Bodega	Bagel Shop	Italian Restaurant	Restaurant	Pilates Studio	Candy Store	Pizza Place	Insurance Office	Dessert Shop	Falafel Restaurant	NYC
165	Manor Heights	Deli / Bodega	Donut Shop	Bagel Shop	Pharmacy	Chinese Restaurant	American Restaurant	Dog Run	Campground	Farmers Market	Fast Food Restaurant	NYC
168	Mariner's	Italian	Deli /	Pizza Place	Home	Yoga Studio	Fish & Chips	Eye Doctor	Factory	Falafel	Farm	NYC

3.5 Find the options for people to move

Provide the options in the same cluster and in Toronto for the people to move.
Person 1 is from Williamsbridge, person 2 is from Chinatown. Find the cluster of them and find the corresponding neighborhood in the same clustering and in Toronto.

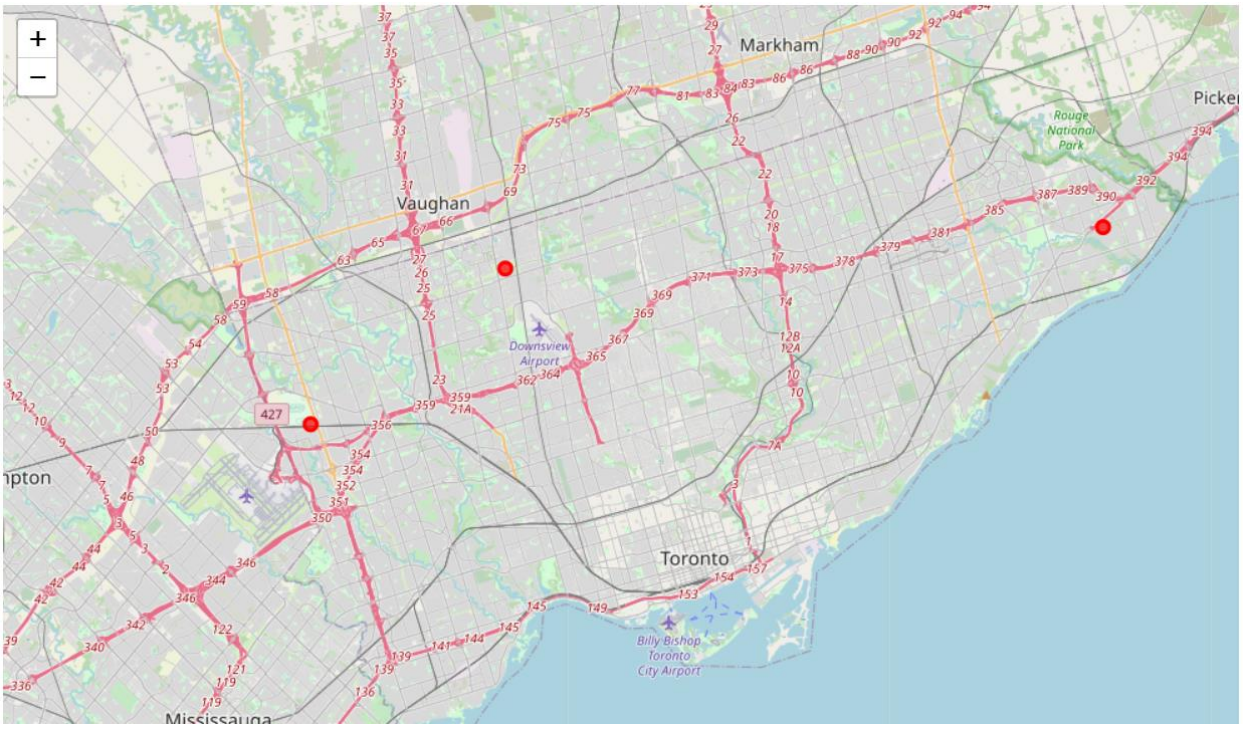
4. Result

4.1 Person 1

The options are: 'Northwest',

'Northwood Park, York University',

'Rouge Hill, Port Union, Highland Creek'



4.2 Person 2

The options are: 'Agincourt',

'Bedford Park, Lawrence Manor East',

'Berczy Park',

'Birch Cliff, Cliffside West',

'Brockton, Parkdale Village, Exhibition Place',

'Business reply mail Processing Centre',

'CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara, Island airport',

'Canada Post Gateway Processing Centre',

'Central Bay Street',

'Christie',

'Church and Wellesley',

'Commerce Court, Victoria Hotel',

'Davisville',

'Davisville North',

'Del Ray, Mount Dennis, Keelsdale and Silverthorn',

'Don Mills',

'Eringate, Bloordale Gardens, Old Burnhamthorpe, Markland Wood',

'Fairview, Henry Farm, Oriole',

'First Canadian Place, Underground city',

'Garden District, Ryerson',

'Harbourfront East, Union Station, Toronto Islands',

'High Park, The Junction South',

'Hillcrest Village',

'Humewood-Cedarvale',

'India Bazaar, The Beaches West',

'Kennedy Park, Ionview, East Birchmount Park',

'Kensington Market, Chinatown, Grange Park',

'Lawrence Manor, Lawrence Heights',

'Leaside',

'Little Portugal, Trinity',

'North Toronto West',

'Parkdale, Roncesvalles',

"Queen's Park, Ontario Provincial Government",

'Regent Park, Harbourfront',

'Richmond, Adelaide, King',

'Runnymede, Swansea',

'St. James Town',

'St. James Town, Cabbagetown',

'Stn A PO Boxes',

'Studio District',

'Summerhill West, Rathnelly, South Hill, Forest Hill SE, Deer Park',

'The Annex, North Midtown, Yorkville',

'The Beaches',

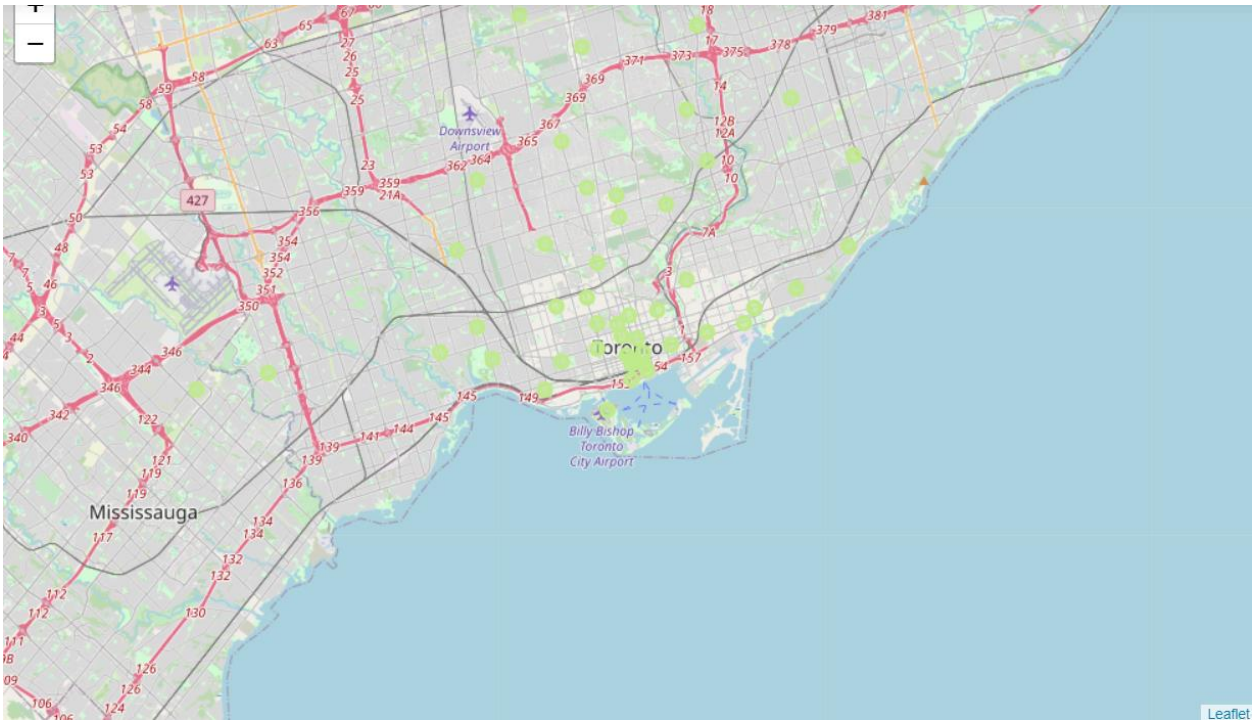
'The Danforth West, Riverdale',

'Toronto Dominion Centre, Design Exchange',

'University of Toronto, Harbord',

'Wexford, Maryvale',

'Willowdale'



5. Conclusion

Person 2 has more options than person 1. If the person is from a neighborhood which in a big clustering, it may be easier to find more options to move.

6. Discussion

If the person cannot find the options in clustering, the k should be lower down to make the clustering bigger to have the neighborhoods in Toronto.