Are neighbourhoods of New york and Toronto similar?

Business problem:

We know that the cities New York and Toronto are very diverse and are the financial capitals of their respective countries. One interesting idea would be to compare the neighborhoods of the two cities and determine how similar or dissimilar they are. Is New York City more like Toronto? Can we identify how similar are the neighbourhoods of New York to the neighbourhoods of toronto?

Background and Importance:

We need to explore the neighbourhoods of the two cities using Foursquare API and apply the k-means clustering algorithm for grouping the neighbourhoods of the two cities. By forming the clusters, we can understand how many neighbourhoods of New York are similar to Toronto.

This problem would be helpful to the people who are having trouble in taking a decision of choosing a neighbourhood of the city to take residence or choosing a neighbourhood of the city to have breakfast,lunch or dinner or choosing a neighbourhood of the city to travel as some people choose to travel to cities which are different from the cities they have already visited while some people choose to travel to cities which are similar to the cities they have already visited.

Data Understanding:

As the neighbourhoods in New York are very large in number, I have chosen only the neighbourhoods in the borough Manhattan for representing New York. We can collect the location data of the neighbourhoods of New York and Toronto and explore the top 100 venues around each neighbourhood using the Foursquare API. Moreover, we are exploring the top 100 venues around each neighbourhood for both cities because we need some means of comparison between the neighbourhoods for grouping. By using these venues for the neighbourhoods we can apply one hot encoding and build a metric of comparison based on the category types of each venue.

Step 1:

Create a dataframe which shows the location coordinates of the neighbourhoods of the two cities.

	Borough	Neighborhood	Latitude	Longitude
0	Manhattan	Marble Hill	40.876551	-73.910660
1	Manhattan	Chinatown	40.715618	-73.994279
2	Manhattan	Washington Heights	40.851903	-73.936900
3	Manhattan	Inwood	40.867684	-73.921210
4	Manhattan	Hamilton Heights	40.823604	-73.949688

Step 2:

Get the top 100 venues around each neighbourhood location into a new dataframe for both cities.It looks as shown below.

	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

Step 3:

Perform one hot encoding on the dataframe obtained from step 2 for each neighbourhood based on the category type for both cities into a new dataframe. It consists of all category types as the columns.

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	...	Vietnamese Restaurant	Volleyball Court
0	Marble Hill	0	0	0	0	0	0	0	0	0	...	0	0
1	Marble Hill	0	0	0	0	0	0	0	0	0	...	0	0
2	Marble Hill	0	0	0	0	0	0	0	0	0	...	0	0
3	Marble Hill	0	0	0	0	0	0	0	0	0	...	0	0
4	Marble Hill	0	0	0	0	0	0	0	0	0	...	0	0

Step 4:

Perform the group by neighbourhood operation on the dataframe in Step 3 and perform the mean operation on each group which will give the mean occurence/frequency of each category in the respective neighbourhood.Perform it for both cities.

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	...	Vietnamese Restaurant	Volleyball Court	Waterfr
0	Battery Park City	0.000000	0.00	0.00	0.000000	0.010526	0.00	0.00	0.000000	0.000000	...	0.000000	0.00	0.000
1	Carnegie Hill	0.000000	0.00	0.00	0.000000	0.010101	0.00	0.00	0.000000	0.010101	...	0.020202	0.00	0.000
2	Central Harlem	0.000000	0.00	0.00	0.068182	0.045455	0.00	0.00	0.000000	0.000000	...	0.000000	0.00	0.000
3	Chelsea	0.000000	0.00	0.00	0.000000	0.030000	0.00	0.00	0.000000	0.000000	...	0.000000	0.00	0.000
4	Chinatown	0.000000	0.00	0.00	0.000000	0.040000	0.00	0.00	0.000000	0.000000	...	0.030000	0.00	0.000
5	Civic Center	0.000000	0.00	0.00	0.000000	0.030000	0.01	0.00	0.000000	0.000000	...	0.000000	0.00	0.000
6	Clinton	0.000000	0.00	0.00	0.000000	0.040000	0.00	0.00	0.000000	0.000000	...	0.000000	0.00	0.000
7	East Harlem	0.000000	0.00	0.00	0.000000	0.000000	0.00	0.00	0.000000	0.000000	...	0.000000	0.00	0.000
8	East Village	0.000000	0.00	0.00	0.000000	0.020000	0.01	0.00	0.020000	0.010000	...	0.020000	0.00	0.000
9	Financial District	0.010000	0.00	0.00	0.000000	0.040000	0.00	0.00	0.000000	0.000000	...	0.000000	0.00	0.000
10	Flatiron	0.000000	0.00	0.00	0.000000	0.040000	0.00	0.00	0.000000	0.000000	...	0.000000	0.00	0.000
11	Gramercy	0.000000	0.00	0.00	0.000000	0.030000	0.00	0.01	0.000000	0.000000	...	0.020000	0.00	0.000
12	Greenwich Village	0.000000	0.00	0.00	0.000000	0.020000	0.00	0.00	0.000000	0.000000	...	0.020000	0.00	0.000
13	Hamilton Heights	0.000000	0.00	0.00	0.000000	0.000000	0.00	0.00	0.000000	0.000000	...	0.000000	0.00	0.000
14	Hudson Yards	0.000000	0.00	0.00	0.000000	0.060241	0.00	0.00	0.000000	0.000000	...	0.000000	0.00	0.000
15	Inwood	0.000000	0.00	0.00	0.000000	0.037037	0.00	0.00	0.000000	0.000000	...	0.000000	0.00	0.000
16	Lenox Hill	0.000000	0.00	0.01	0.000000	0.000000	0.00	0.00	0.000000	0.000000	...	0.000000	0.00	0.000
17	Lincoln Square	0.000000	0.00	0.00	0.000000	0.030000	0.00	0.00	0.000000	0.000000	...	0.000000	0.00	0.000
18	Little Italy	0.000000	0.00	0.00	0.000000	0.000000	0.00	0.00	0.000000	0.000000	...	0.010000	0.00	0.000
19	Lower East Side	0.000000	0.00	0.00	0.000000	0.000000	0.00	0.00	0.000000	0.019608	...	0.019608	0.00	0.000
20	Manhattan	0.000000	0.00	0.00	0.000000	0.000000	0.00	0.00	0.000000	0.000000		0.020833	0.00	0.000

Step 5:

Create a new dataframe which shows neighbourhood along with top 10 venues in the respective neighbourhood. Perform it for both cities.

	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	Coffee Shop	Hotel	Wine Shop	Women's Store	Clothing Store	Gym	Memorial Site	Pizza Place	Grocery Store
1	Carnegie Hill	Coffee Shop	Pizza Place	Cosmetics Shop	Yoga Studio	Bakery	Gym	Bookstore	Café	Japanese Restaurant	Wine Shop
2	Central Harlem	African Restaurant	Chinese Restaurant	American Restaurant	Bar	Cosmetics Shop	Seafood Restaurant	French Restaurant	Fried Chicken Joint	Bookstore	Caribbean Restaurant
3	Chelsea	Coffee Shop	Bakery	Italian Restaurant	Ice Cream Shop	Nightclub	Theater	American Restaurant	Hotel	Tapas Restaurant	Cocktail Bar
4	Chinatown	Chinese Restaurant	Cocktail Bar	American Restaurant	Salon / Barbershop	Bakery	Spa	Optical Shop	Vietnamese Restaurant	Hotpot Restaurant	Sandwich Place

Step 6:

Concatenate the two dataframes of two cities obtained from step 4. It simply combines the neighbourhoods of New York and Toronto. Each row shows a category metric of each neighbourhood based on which we perform k-means clustering.

	Neighborhood	Yoga Studio	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	...	Turkish Restaurant	Udon Restaurant	U Bookst
0	Adelaide,King,Richmond	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.02	...	0.0	0.0	
1	Berczy Park	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	...	0.0	0.0	
2	Brockton,Exhibition Place,Parkdale Village	0.083333	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	...	0.0	0.0	
3	Business Reply Mail Processing Centre 969 Eastern	0.055556	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	...	0.0	0.0	
4	CN Tower,Bathurst Quay,Island airport,Harbourf...	0.000000	0.0	0.058824	0.058824	0.058824	0.117647	0.176471	0.117647	0.00	...	0.0	0.0	

Step 7:

Add a City column to the two dataframes obtained from Step 5. Concatenate the two dataframes of two cities obtained from step 5. It simply combines the neighbourhoods of new york and toronto. Each row shows the top 10 venues of the respective neighbourhood.

	Neighborhood	City	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,King,Richmond	Toronto	Coffee Shop	Café	Bar	Hotel	Restaurant	Steakhouse	Burger Joint	Cosmetics Shop	Asian Restaurant	Thai Restaurant
1	Berczy Park	Toronto	Coffee Shop	Cocktail Bar	Cheese Shop	Seafood Restaurant	Farmers Market	Beer Bar	Bakery	Steakhouse	Café	Gourmet Shop
2	Brockton,Exhibition Place,Parkdale Village	Toronto	Café	Yoga Studio	Breakfast Spot	Coffee Shop	Gym	Pet Store	Performing Arts Venue	Italian Restaurant	Intersection	Gym / Fitness Center
3	Business Reply Mail Processing Centre 969 Eastern	Toronto	Light Rail Station	Yoga Studio	Garden	Butcher	Fast Food Restaurant	Auto Workshop	Farmers Market	Burrito Place	Spa	Pizza Place
4	CN Tower,Bathurst Quay,Island airport,Harbourf...	Toronto	Airport Service	Airport Terminal	Airport Lounge	Sculpture Garden	Plane	Coffee Shop	Boat or Ferry	Boutique	Harbor / Marina	Airport Gate

The concatenated dataframe in step 6 can be directly used for k-means clustering algorithm to group the neighbourhoods of New York and Toronto. Based on the results, we can conclude what neighbourhoods in New York are more similar to Toronto.