Clustering Neighborhoods in New York City and Toronto

Applied Data Science Capstone by IBM on Coursera

1. Introduction

New York and Toronto are both recognized as being between the most multicultural and cosmopolitan cities in the world. They are also the most populous cities and the financial capitals of their respective countries. As such, there's a large influx of people and businesses moving into both cities every year.

Moving into a new city involves a lot of thinking and planning. One important aspect a person or a business could consider when thinking about moving is which area in the new city is similar to the region they are currently at, or to a region they are familiar with.

In this project, we explore neighborhoods in New York and Toronto and cluster them into groups of similar neighborhoods. In doing so, we aim to aid households and business looking into moving between both cities to better choose which area in the new city they might be most interested in moving to.

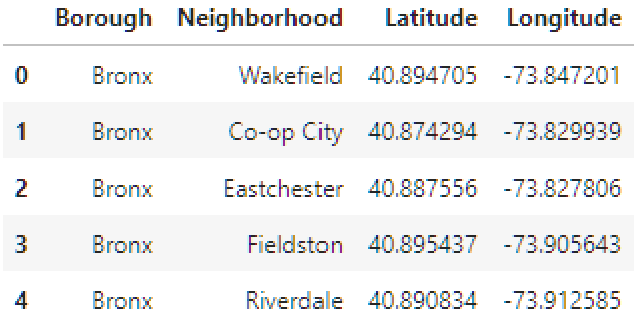
2. Data

To compare the regions of both cities, we grouped neighborhoods according to most common types of venues in each of them. First, we needed a list of neighborhoods in each of the cities and their coordinates, which we then used to create dataframes using the Pandas library.

NYU Spatial Data Repository's dataset was used to create the dataframe of neighborhoods in New York; the original data is available at https://geo.nyu.edu/catalog/nyu_2451_34572was.

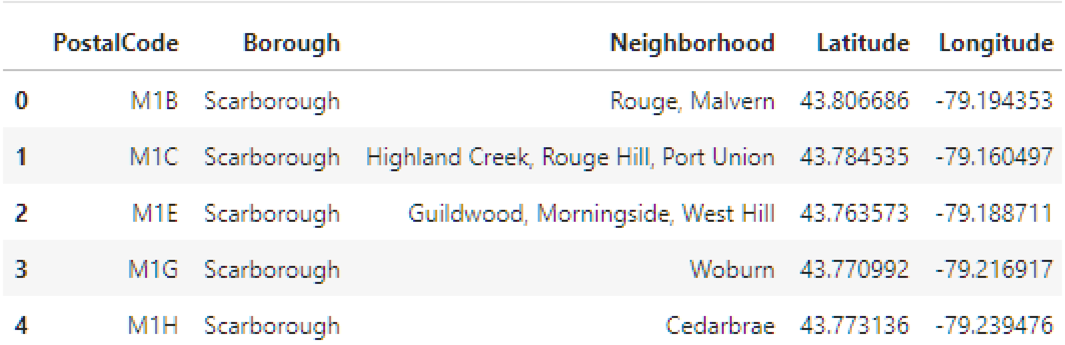
Data for neighborhoods in Toronto was not as readily available, so for Toronto's dataframe we scraped Wikipedia's page on Toronto postal codes in order to obtain the names of the neighborhoods and boroughs, and then used Python's Geocoder package to obtain the coordinates of each postal code. Given this difference in data sources, notice that we will be working with neighborhoods in New York and with postal codes in Toronto (which may include more than on neighborhood); for simplicity, in the rest of the project we will refer to both as simply "neighborhoods".

Here is the head of each of the dataframes we created:



	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

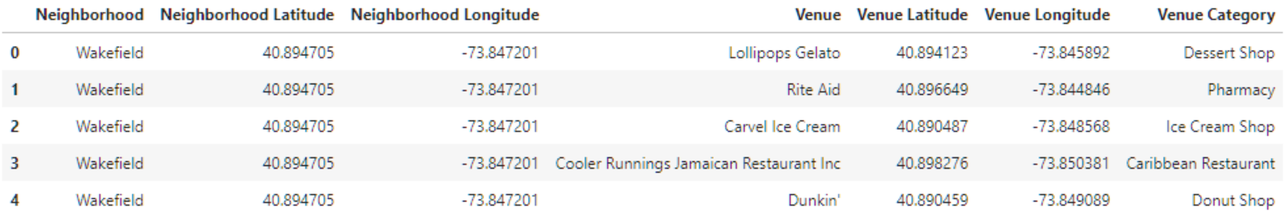
Dataframe 1: dataframe of New York City's neighborhoods and their coordinates.



	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

Dataframe 2: dataframe of Toronto's neighborhoods and their coordinates.

Next we combined both dataframes and used Foursquare API to get venue information for the neighborhoods (or postal codes) of both cities, and used the venues' categories to cluster the neighborhoods (or postal codes) into groups with similar venue categories. Here’s the head of the dataframe containing information on venues:



	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	Rite Aid	40.896649	-73.844846	Pharmacy
2	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
3	Wakefield	40.894705	-73.847201	Cooler Runnings Jamaican Restaurant Inc	40.898276	-73.850381	Caribbean Restaurant
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

Dataframe 3: dataframe of venues, containing their coordinates and categories.

3. Methodology

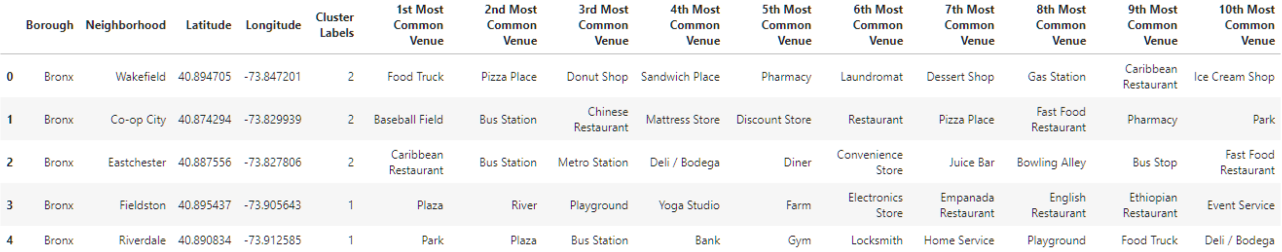
Having created the list of nearby venues, we used the **k**-means clustering algorithm to group the neighborhoods and identify the ones with similar venue categories. **k**-means clustering was chosen because our data is unlabeled, so we had to select an unsupervised algorithm, and because it is relatively efficient.

Another advantage of **k**-means clustering for our problem is that we can choose the value **k**, the number of clusters. This is important for us because we didn’t want the number of clusters to be too small (such as 2 or 3), since that would mean most clusters would contain too many neighborhoods and, thus, not provide much information to compare them.

Correspondingly, we also didn’t want **k** to be too big, since in that case several clusters could contain too few neighborhoods or only contain neighborhoods in one of the cities. Of course, it is not an issue if some clusters are small or are in only one city: that can be useful to indicate that some few neighborhoods share some specific characteristics, or that some regions in New York City have no counterpart in Toronto, and vice-versa. However, if too many clusters are small, that could limit the options of similar neighborhoods to choose from.

Therefore, we decided the number of clusters should be between 4 and 9 (not including 9). To select **k**, we ran the algorithm for values of **k** in the selected range, and then evaluated the model’s accuracy using it's silhouette score. We then chose the number of clusters as the one which returned best model accuracy, which was **k** = 5.

Finally, we added the cluster labels to the neighborhoods dataframe, and also listed the most common venue category in each neighborhood:



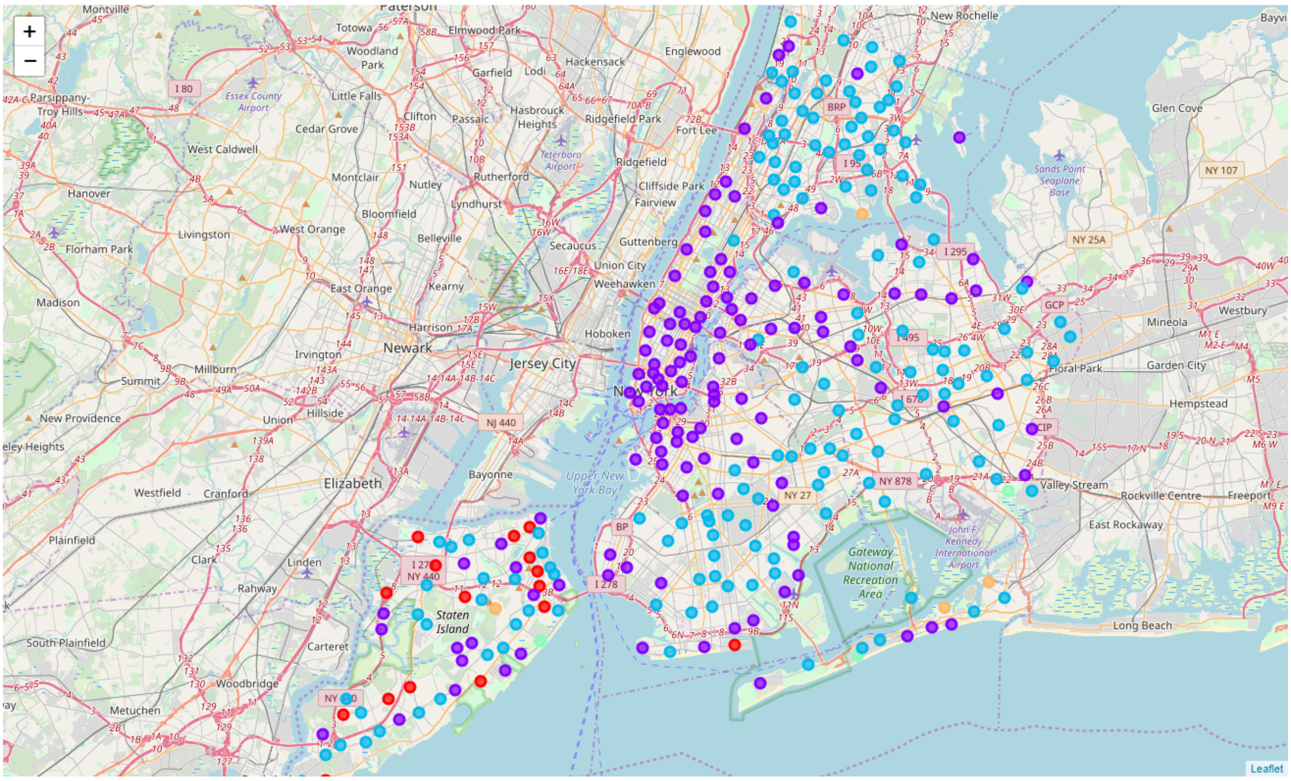
	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bronx	Wakefield	40.894705	-73.847201	2	Food Truck	Pizza Place	Donut Shop	Sandwich Place	Pharmacy	Laundromat	Dessert Shop	Gas Station	Caribbean Restaurant	Ice Cream Shop
1	Bronx	Co-op City	40.874294	-73.829939	2	Baseball Field	Bus Station	Chinese Restaurant	Mattress Store	Discount Store	Restaurant	Pizza Place	Fast Food Restaurant	Pharmacy	Park
2	Bronx	Eastchester	40.887556	-73.827806	2	Caribbean Restaurant	Bus Station	Metro Station	Deli / Bodega	Diner	Convenience Store	Juice Bar	Bowling Alley	Bus Stop	Fast Food Restaurant
3	Bronx	Fieldston	40.895437	-73.905643	1	Plaza	River	Playground	Yoga Studio	Farm	Electronics Store	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Event Service
4	Bronx	Riverdale	40.890834	-73.912585	1	Park	Plaza	Bus Station	Bank	Gym	Locksmith	Home Service	Playground	Food Truck	Deli / Bodega

Dataframe 4: dataframe listing each neighborhood’s cluster label and its top venues.

4. Results

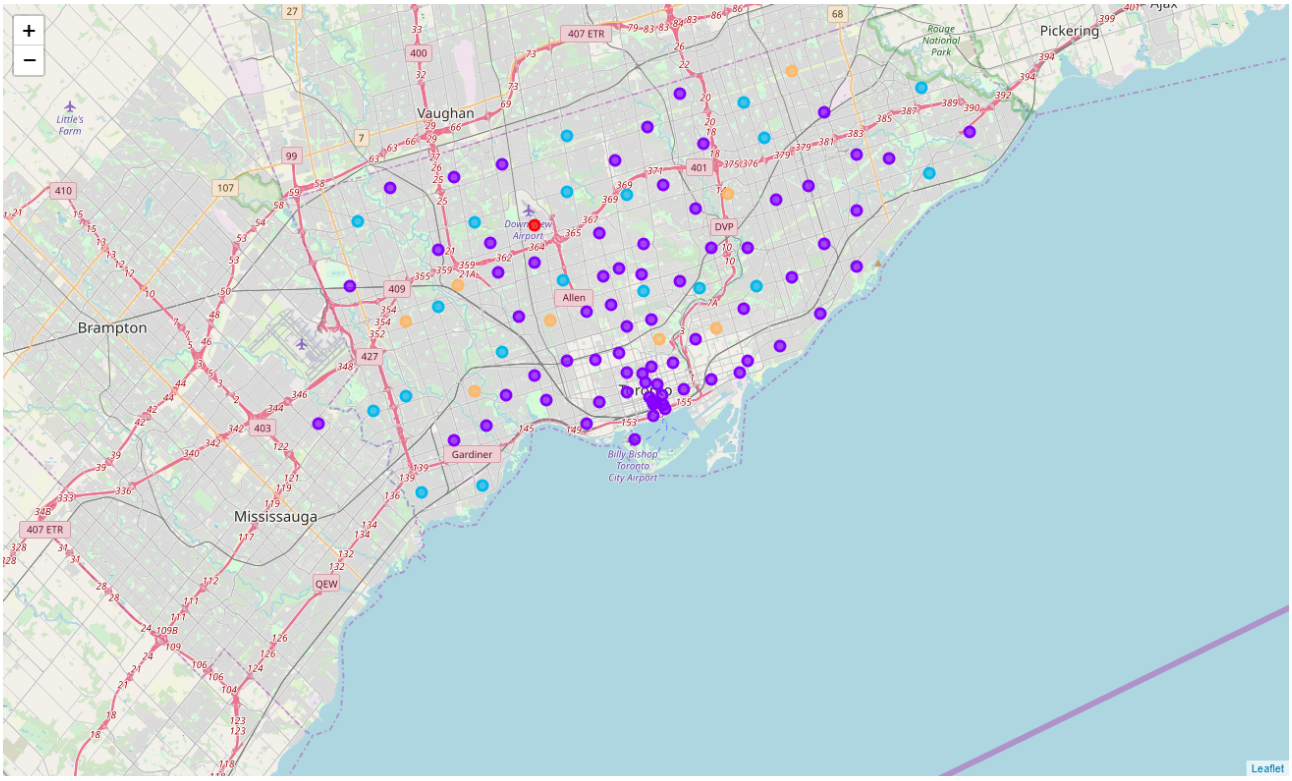
To visualize the clusters we obtained, we used the Folium library to create maps of New York City and Toronto and labeled their neighborhoods using their clusters.

Here’s the map of New York City:



Map 1: map of New York City with neighborhoods labeled according to their cluster.

And the map of Toronto:



Map 2: map of Toronto with neighborhoods labeled according to their cluster.

To better understand the clusters of neighborhoods, we created a table listing the most common venue for each cluster:

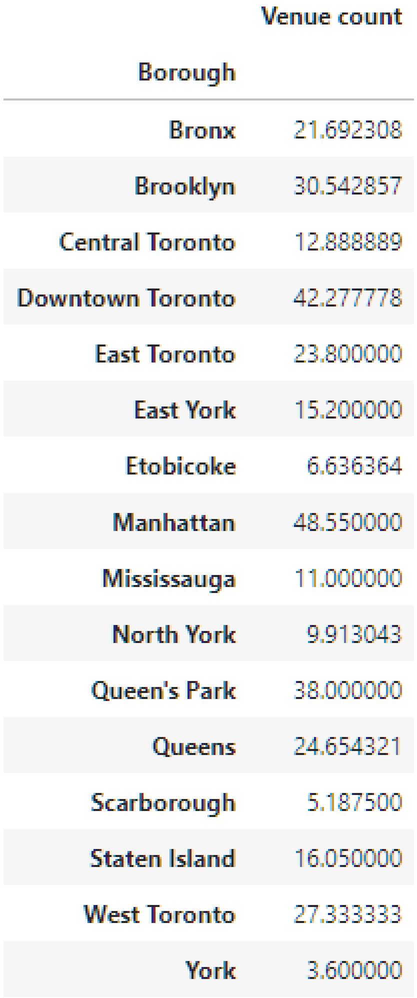
Cluster label	Marker color on map	Top 5 venue categories (in descending order)
0	**Red**	Bus Stop, Pizza Place, Park, Bagel Shop, Coffee Shop
1	**Purple**	Coffee Shop, Café, Italian Restaurant, Bar, Park
2	**Light blue**	Pizza Place, Deli / Bodega, Pharmacy, Bank, Donut Shop
3	**Light green**	Deli / Bodega, Beach, Pier, Beach Bar, Athletics & Sports
4	**Orange**	Park, Playground, Fast Food Restaurant, Pizza Place, River

5. Discussion

The data provided by the maps is easily understood, so recommendations based on which regions of both cities are most similar to one another are straightforwardly provided by them. Households and businesses looking into moving between both cities can then use this information to better choose a neighborhood in the new city (and, as we've repeatedly pointed out, the similarity between areas is only one of several aspects to examine).

It is worth highlighting that we used one specific metric to compare neighborhoods: namely, the categories of venues in the area. There are, of course, various other features that could be used to group the neighborhoods, although some might be harder to find extensive and quantitative data for. As such, approaches other than machine learning algorithms may be more suited for exploring the similarity of neighborhoods using other traits.

Notice, however, that even the approach used here has limitations due to data availability: using Foursquare API some neighborhoods returned significantly more venues than others (and that is one of the reasons we capped the number of returned venues per neighborhood to 50), therefore it's possible there wasn't enough information on venue category for some regions to meaningfully compare them. The table below illustrates this imbalance:



	Venue count
Borough	
Bronx	21.692308
Brooklyn	30.542857
Central Toronto	12.888889
Downtown Toronto	42.277778
East Toronto	23.800000
East York	15.200000
Etobicoke	6.636364
Manhattan	48.550000
Mississauga	11.000000
North York	9.913043
Queen's Park	38.000000
Queens	24.654321
Scarborough	5.187500
Staten Island	16.050000
West Toronto	27.333333
York	3.600000

Dataframe 5: dataframe showing the average number of venues returned for neighborhood by borough.

On the other hand, that suggests that having our analysis in Toronto be based on postal codes instead of neighborhoods may have helped reduce this imbalance, since each postal code can contain information on venues of several neighborhoods.

Despite the aforementioned limitations on data, altogether we believe our analysis can still provide helpful insights to those considering to move between both cities.

6. Conclusion

The purpose of this project was to cluster neighborhoods in New York and Toronto into groups with similar characteristics. To do so, we used Foursquare API to get information on venues and then used the **k**-means clustering algorithm to group neighborhoods according to venue category, choosing the most adequate value of **k** in a certain range. In the end, we visualized the distribution of neighborhoods, labeled by their clusters, in a map of each city.

Although there are several factors to consider when moving between New York and Toronto, the analysis we conducted here can aid that decision by showing which areas of both cities are most similar to one another.