Clustering Neighborhoods in New York City and Toronto

Applied Data Science
Capstone by IBM
on Coursera

1. Introduction

- New York and Toronto both big, multicultural cities and the financial capitals of their respective countries, so several people and businesses move into them every year.
- One aspect a person / business could consider when 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.

2. Data

First, we needed a list of neighborhoods in each of the cities and their coordinates.

Neighborhoods in New York

We used NYU Spatial Data Repository's dataset.

	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

Neighborhoods in Toronto

For Toronto's dataframe we scraped Wikipedia's page on Toronto postal codes to obtain the names of the neighborhoods and boroughs, and then used Python's Geocoder package to obtain the coordinates of each postal code.

	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

Getting venue information

We then used Foursquare API to get venue information for the neighborhoods of both cities. More specifically, we extracted the venues' categories, which we used to cluster the neighborhoods into groups with similar venue categories.

	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

3. Methodology and Analysis

- We used **k-means clustering** algorithm to group the neighborhoods according to venue category
- To choose the best value for **k** we ran the algorithm for different number of clusters and choose the one for which the model had best **silhouette score.**

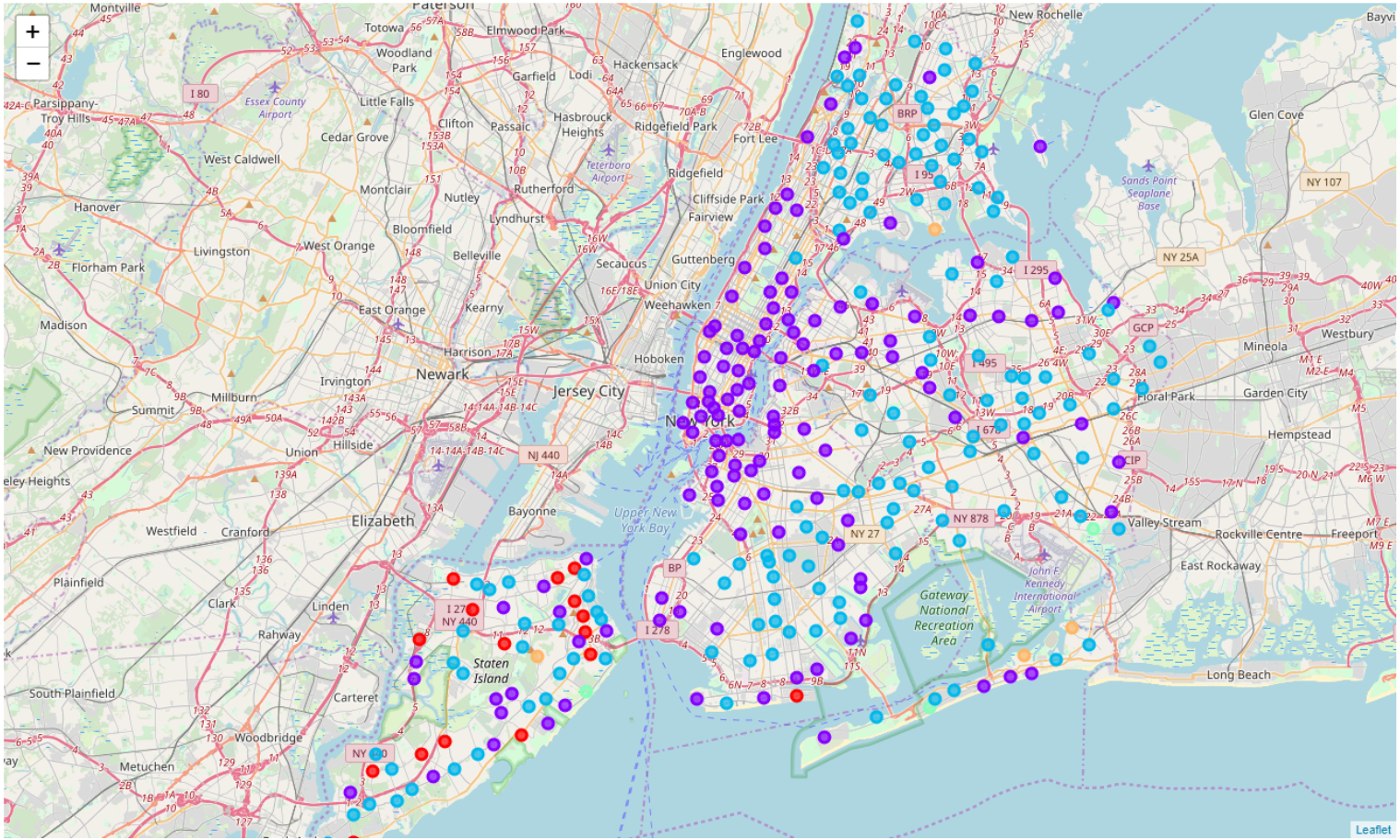
- The best number of clusters (in the range of 4 to 9) was 5.

	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

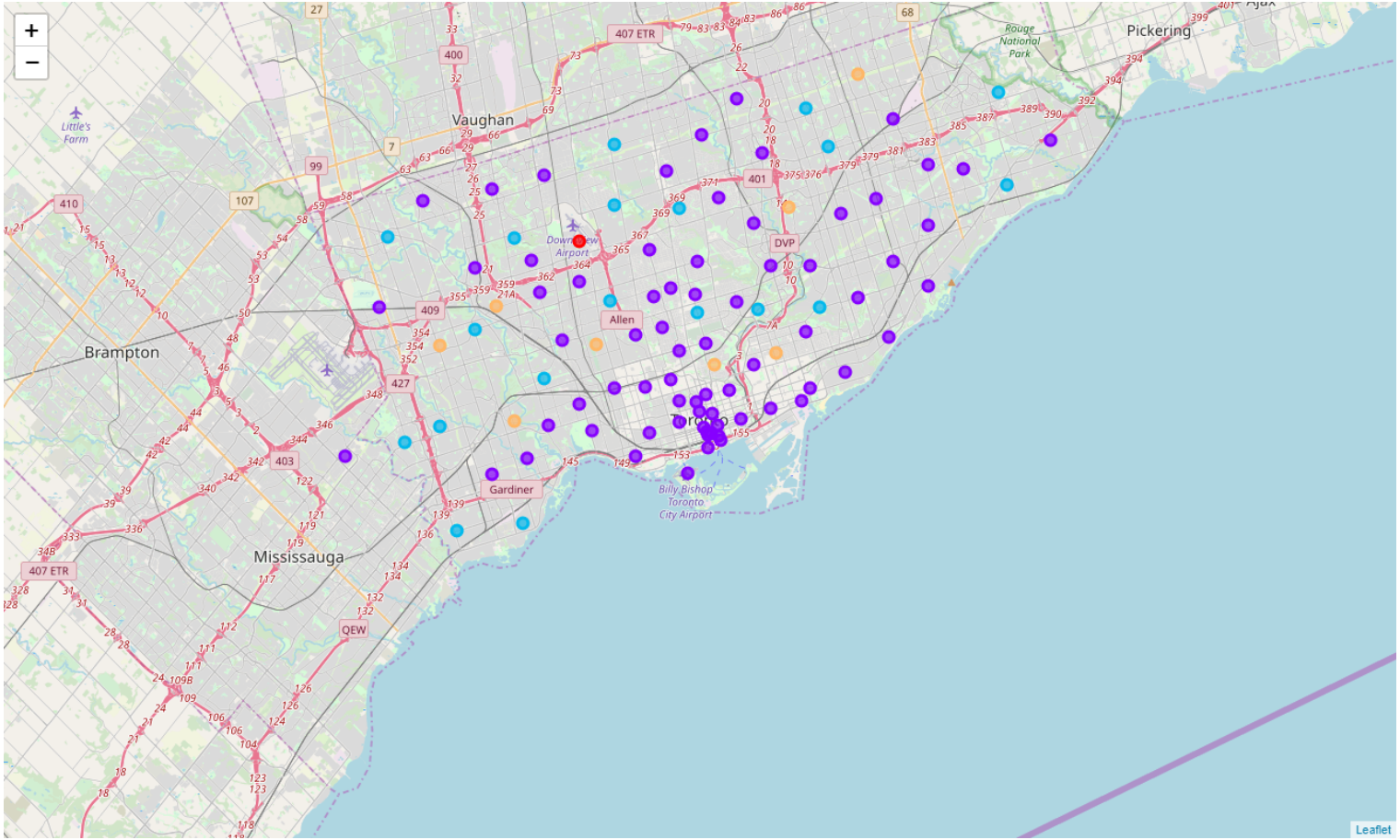
4. Results and Discussion

- We split the dataframe we obtained above in two, one for New York and one for Toronto.
- Each was used to create maps to visualize the clusters in a map, using the Folium library.
- We also explored which were the most common venue categories in each cluster.

Map of New York



Map of Toronto



Top 5 venue categories per 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. Conclusion

- Using k-means clustering algorithm, we grouped New York and Toronto Neighborhoods into clusters of similar venue categories.
- This can aid households and businesses looking into moving between both cities which areas are of both cities are most similar to one another.