COURSERA CAPSTONE-FINAL REPORT

April, 2021

Outline

1 Introduction

In this project, we will use data science to evaluate how similar or dissimilar two boroughs/ cities are in terms of venues. We will work on Manhattan and Toronto boroughs to illustrate the concept.
The system will be helpful to solve several practical issues. For example:

- Business case 1: Company K which is already in Toronto wants to extend his business in Mahanttan, and want to know how similar or dissimilar Toronto and Mahnattan are in term of venues. This information is critical to company K, since it can decide to replicate the same distribution system or to change it.

- Business case2: Mr. Dupont is living now in the “Marble Hill” neighborhood in Manhattan. Mr. Dupont loves this neighborhood very much. He is leaving for Toronto and wants to find something quite similar to “Marble Hill” in Toronto.

Aside from those two business cases illustrated here, there are many other real-world problems that the tool developed here could solve.

2 Data

We will use the data :

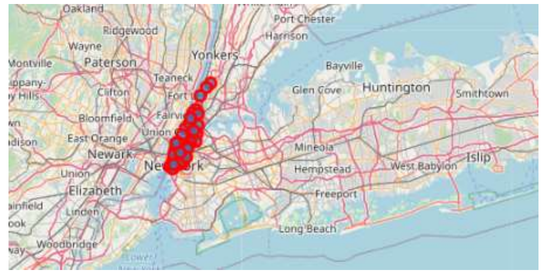
- For the Toronto borough
  - Web scrapping (url= https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) to get the list of postal code, Borough and Neighborhood in Toronto (
  - Geospatial data for Toronto: ‘Geospatial_Coordinates.csv’
  - The data is saved in the file Toronto_Data.csv and is available in the same repository on Github.
- For Manhattan borough:
  - Download and process list of boroughs and neighborhoods in NewYork https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json
  - Save the data specific to Manhattan in a file called ‘Manhattan_Data.csv’ and available on Github

3 Methodology

3.1 Preparing the data

As mentioned in the previous section, we first acquired Toronto borough data through webscrapping, then the data on Manhattan neighborhood. The two data sets are merged in a single dataframe. We use Folium library to visualize the different neighborhoods on the map.

(a) Toronto neighborhoods



(b) Manhattan neighborhoods

Figure1: Toronto's and Manhattan's neighborhoods

3.2 Getting all venues using Foursquare API

Foursquare API is used to get venues in radius of 500 meters of each neighborhood. The maximum number of venues is limited to 100 (Fig.2).

	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue
4806	Manhattan	Hudson Yards	40.756658	-74.000111	StarDust
4807	Manhattan	Hudson Yards	40.756658	-74.000111	George's
4808	Manhattan	Hudson Yards	40.756658	-74.000111	Jake's
4809	Manhattan	Hudson Yards	40.756658	-74.000111	Gray Line New York Sightseeing Cruises - Pier 78
4810	Manhattan	Hudson Yards	40.756658	-74.000111	Pier Cafe

(a) Sample of venues in Hudson Yards neighborhood (Manhattan)

	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue
0	Toronto	Regent Park, Harbourfront	43.65426	-79.360636	Tandem Coffee
1	Toronto	Regent Park, Harbourfront	43.65426	-79.360636	Roselle Desserts
2	Toronto	Regent Park, Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA
3	Toronto	Regent Park, Harbourfront	43.65426	-79.360636	Impact Kitchen
4	Toronto	Regent Park, Harbourfront	43.65426	-79.360636	Body Blitz Spa East

(b) Sample of venues in Regent Park neighborhood (Toronto)

Figure2: Sample of venues obtained using Foursquare API

3.3 Clustering using KNN

After applying the panda's "get_dummies" method to the dataframe, we use KNN algorithm to cluster the combined to both Manhattan's and Toronto's neighborhoods (Fig.3).

Borough	Neighborhood	Latitude	Longitude	Cluster Labels
Toronto	Regent Park, Harbourfront	43.654260	-79.360636	0
Toronto	Garden District, Ryerson	43.657162	-79.378937	0
Toronto	St. James Town	43.651494	-79.375418	0
Toronto	The Beaches	43.676357	-79.293031	0
Toronto	Berczy Park	43.644771	-79.373306	0

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels
74	Manhattan	Turtle Bay	40.752042	-73.967708	0
75	Manhattan	Tudor City	40.746917	-73.971219	0
76	Manhattan	Stuyvesant Town	40.731000	-73.974052	0
77	Manhattan	Flatiron	40.739673	-73.990947	0
78	Manhattan	Hudson Yards	40.756658	-74.000111	0

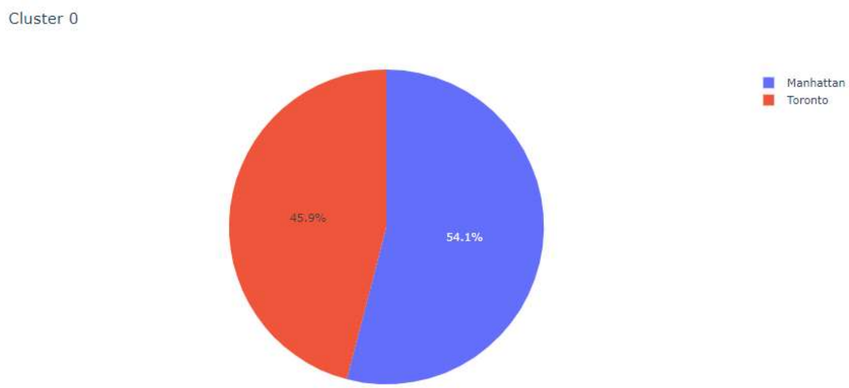
Figure 3: Results of the clustering

4 Business cases

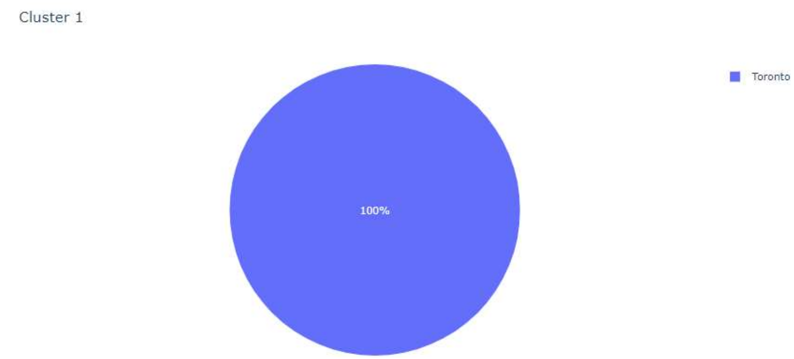
We will use the approach to solve two business cases.

4.1 Business case1: similarity between two boroughs

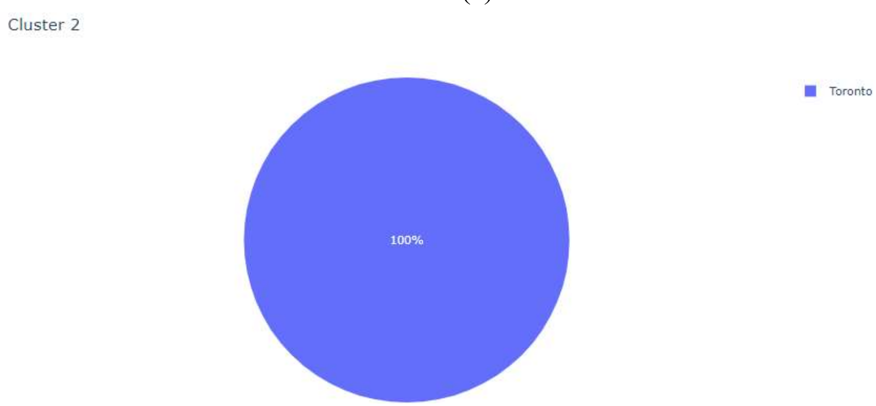
Using the example of Manhattan and Toronto boroughs, we want to find how far these two neighborhoods are similar. The figure 4 presents the reparation between Toronto's and Manhattan's boroughs in each cluster.



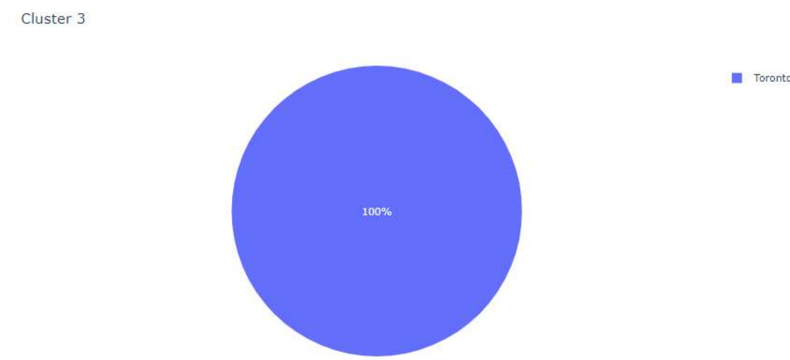
(a) Cluster 0



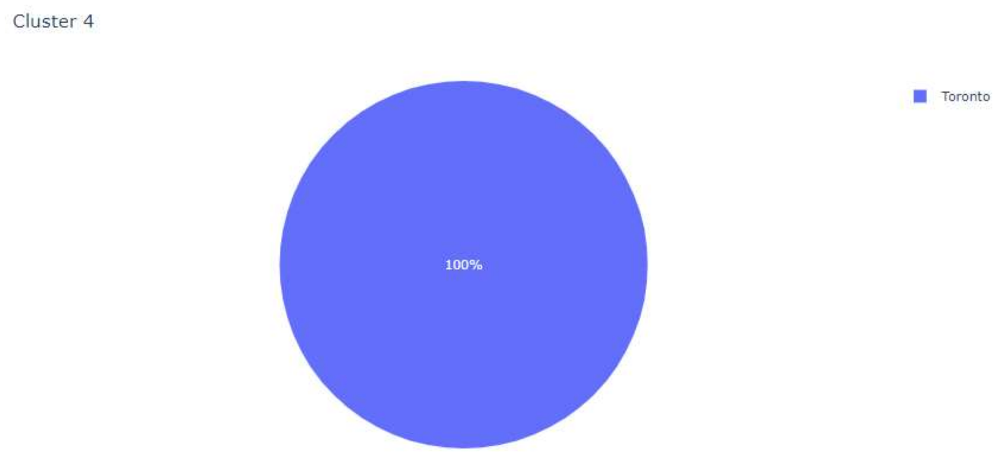
(b) Cluster 1



(c) Cluster 2



(c) Cluster 3



(d) Cluster 4

Figure 4: results of the clustering

We can see that:

-In cluster 0 there are 46% of Manhattan's neighborhood and 54% of Toronto's. This cluster seems to be balanced.

-In cluster 1, 2, 3, 4; there are only neighborhoods of Toronto and no neighborhood in Manhattan. In total there are 5 neighborhoods in those 4 clusters, all in Toronto.

In the next sections we will use Data Science tools to understand why those 5 neighborhoods are so unique according to our clustering. Let's investigate 2 different hypotheses:

-Hypothesis 1: The most common venues in those 5 neighborhoods are very different of those of the rest of the neighborhoods in Toronto.

-Hypothesis 2: The 5 neighborhoods are located on the outskirt of the city.

4.1.1 Testing Hypothesis 1: The most common venues are very different in those 5 neighborhoods

Figure 5 presents the most common venues in Cluster0, and Figure 6 presents the most common venues in cluster 1,2,3, and 4.

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
0	Toronto	Regent Park, Harbourfront	43.654260	-79.360636	0	Coffee Shop	Park	Pub	Bakery
1	Toronto	Garden District, Ryerson	43.657162	-79.378937	0	Clothing Store	Coffee Shop	Middle Eastern Restaurant	Italian Restaurant
2	Toronto	St. James Town	43.651494	-79.375418	0	Café	Coffee Shop	Gastropub	Cocktail Bar
3	Toronto	The Beaches	43.676357	-79.293031	0	Health Food Store	Pub	Pizza Place	Trail
4	Toronto	Berczy Park	43.644771	-79.373306	0	Coffee Shop	Cocktail Bar	Bakery	Farmers Market
5	Toronto	Central Bay Street	43.657952	-79.387383	0	Coffee Shop	Sandwich Place	Café	Italian Restaurant
6	Toronto	Christie	43.669542	-79.422564	0	Grocery Store	Café	Park	Baby Store
7	Toronto	Richmond, Adelaide, King	43.650571	-79.384568	0	Coffee Shop	Café	Restaurant	Hotel

Figure 5: Most common venues in cluster 0

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
9	Toronto	The Danforth East	43.685347	-79.338106	3	Park	Convenience Store	Metro Station	Nail Salon
18	Toronto	Lawrence Park	43.728020	-79.388790	1	Bus Line	Park	Swim School	Yoga Studio
19	Toronto	Roselawn	43.711695	-79.416936	2	Pool	Fast Food Restaurant	Garden	Music Venue
29	Toronto	Moore Park, Summerhill East	43.689574	-79.383160	4	Restaurant	Trail	Tennis Court	Lawyer
33	Toronto	Rosedale	43.679563	-79.377529	3	Park	Playground	Trail	Yoga Studio

Figure 6: Most common venues in cluster 1,2,3,4

Conclusion on hypothesis 1: We can see that in cluster 0, the most common venues are food-like venues (bar, coffee, restaurant, pizza place, ...), while in cluster 1, 2, 3, 4 the most common venues are sport-like venues such as: trail, park, pool, ...

4.1.2 Testing Hypothesis 2: The most common venues are located on the outskirt of the city

Using Folium, we plot in red the neighborhoods of cluster 1-4, and in blue the neighborhoods of cluster 0 (Fig.7).

Figure 7: In red neighborhoods of cluster 1-4/ in blue neighborhoods of cluster 0

Conclusion on hypothesis 2: This second hypothesis also seems to be verified. The neighborhoods of cluster 1,2,3,4 in red on the Folium map are not in central Toronto but instead on the outskirt of the city.

Conclusion on Business Case 1: The two hypotheses explained why the cluster 1,2,3, and 4 are so singular. It appears from this analysis that there are few sport-friendly neighborhoods in Manhattan. For somebody like me who is passionate about sport, I prefer Toronto and I will definitely choose one of the 5 neighborhoods mentioned previously.

4.2 Neighborhood recommendation system during a relocation

We will suppose that Mr.Dupont is living in Marble Hill in Manhattan. In two weeks, Mr. Dupont will leave Manhattan for Toronto, and he wants to find a list of neighborhoods that are similar to Marble Hill in Toronto. Even before thinking about the price of rents, rankings of schools... Let's find to Mr. Dupont a neighborhood which is similar in terms of venues to Marble Hill.

Figure 8 presents a sample of all neighborhoods that could satisfy the request.

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
0	Toronto	Regent Park, Harbourfront	43.654260	-79.360636	0	Coffee Shop	Park	Pub	Bakery
1	Toronto	Garden District, Ryerson	43.657162	-79.378937	0	Clothing Store	Coffee Shop	Middle Eastern Restaurant	Italian Restaurant
2	Toronto	St. James Town	43.651494	-79.375418	0	Café	Coffee Shop	Gastropub	Cocktail Bar
3	Toronto	The Beaches	43.676357	-79.293031	0	Health Food Store	Pub	Pizza Place	Trail
4	Toronto	Berczy Park	43.644771	-79.373306	0	Coffee Shop	Cocktail Bar	Bakery	Farmers Market
5	Toronto	Central Bay Street	43.657952	-79.387383	0	Coffee Shop	Sandwich Place	Café	Italian Restaurant
6	Toronto	Christie	43.669542	-79.422564	0	Grocery Store	Café	Park	Baby Store
7	Toronto	Richmond, Adelaide, King	43.650571	-79.384568	0	Coffee Shop	Café	Restaurant	Hotel
8	Toronto	Dufferin, Dovercourt Village	43.669005	-79.442259	0	Bakery	Pharmacy	Bank	Bar
10	Toronto	Harbourfront East, Union Station, Toronto Islands	43.640816	-79.381752	0	Coffee Shop	Aquarium	Café	Hotel
11	Toronto	Little Portugal, Trinity	43.647927	-79.419750	0	Bar	Café	Coffee Shop	Restaurant
12	Toronto	The Danforth West, Riverdale	43.679557	-79.352188	0	Greek Restaurant	Coffee Shop	Italian Restaurant	Furniture / Home Store
13	Toronto	Toronto Dominion Centre, Design Exchange	43.647177	-79.381576	0	Coffee Shop	Hotel	Café	Seafood Restaurant
14	Toronto	Brockton, Parkdale Village, Exhibition Place	43.636847	-79.428191	0	Café	Breakfast Spot	Coffee Shop	Intersection
15	Toronto	India Bazaar, The Beaches West	43.668999	-79.315572	0	Park	Brewery	Fish & Chips Shop	Liquor Store
16	Toronto	Commerce Court, Victoria Hotel	43.648198	-79.379817	0	Coffee Shop	Restaurant	Hotel	Café
17	Toronto	Studio District	43.659526	-79.340923	0	Coffee Shop	Café	Bakery	Brewery
20	Toronto	Davisville North	43.712751	-79.390197	0	Gym	Breakfast Spot	Hotel	Food & Drink Shop

Figure 8: Sample of all Toronto's neighborhoods that could satisfy Mr. Dupont request

Starting from the dataframe displayed in Figure 8, Mr.Dupont, using additional criteria such as school's rankings, could refine this list further.

5 Conclusion

In this project, we first collect data from Manhattan and Toronto neighborhoods. We merged the data in only one dataframe. We then used Foursquare API to have the list and category of venues nearby. We used a KNN algorithm to cluster the list of neighborhoods in Manhattan and Toronto. We analyzed the results through two business cases.

I think that the Foursquare API is a powerful tool that could help develop or even create a business. For example, it can be used to help a city to become more eco-friendly, tourist-friendly or business-friendly. I have many ideas in mind, and I will continue to work on them after this certification.

I want to thank IBM, Coursera, and Alex Aklson, who built this capstone project. This last course is simply excellent.

Thank you for reviewing my work.