Applied Data Science Capstone

by IBM

REPORT

The battle of the neighborhoods:

Relocation between Toronto and New York City

by Ricardo F Reategui

May 2020

Introduction

Thousands of skilled workers, businessmen, science and technology professionals, entrepreneurs, move from one city to another anywhere in the world every year seeking better opportunities for themselves and their families.

Imagine a relocation company which offers services to these clients in a unique way. This particular company could leverage the Foursquare location data with the needs and expectations of its clients in terms of housing, recreation, diversity, transportation, and other characteristics in order to tailor the needs of their clients who want to relocate to a neighborhood as similar as possible to their original place.

For this project I will use as a relocation model a company with a client who wants to move from Toronto to New York City or vice versa.

Data

- Datasets of Toronto and New York
- Foursquare API
- Folium library
- k-means algorithm

The dataset for New York City is available on the web (https://geo.nyu.edu/catalog/nyu_2451_34572). It consisted of a nested Python dictionaries that can be retrieved as a JSON file. An inspection of the data revealed that the relevant data is in the "Features" key; therefore, these data is filled into an empty pandas dataframe that displays borough, neighborhood, latitude, and longitude of each New Yprk neighborhood. A portion of this dataframe is displayed Table 1.

Table 1. First 5 rows of a pandas dataframe with the coordinates of New York City neighborhoods

	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

The dataset for Toronto was obtained by reading a Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) directly into a pandas dataframe. The Wikipedia page contained the postal codes and the boroughs of the Toronto neighborhoods.

The data had to be cleaned by ignoring not assigned boroughs to a neighborhood. This process generated a new dataframe that listed each postal code with its respective neighborhood.

A Geospatial data of the Toronto postal codes and their coordinates were already available as a CSV file (http://cocl.us/Geospatial_data). This file was read into a pandas dataframe.

Joining the cleaned Toronto dataframe with the Geospatial data of Toronto generated the final Toronto dataframe that will be used for this project. Table 2 shows the first 5 rows of this final dataframe.

Information about the number of boroughs and neighborhoods can be obtained from this dataframe.

Table 2. First 5 rows of a pandas dataframe with the coordinates of Toronto neighborhoods

	Postal code	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

The Foursquare API will be used with Toronto and New York dataframes to explore their respective neighborhoods and segment them. In order to use Foursquare, credentials such as Client ID, Client Secret, and Version of Foursquare will be defined.

A new dataframe will be generated for each city, Toronto and New York. Each dataframe will contain the venues that Foursquare encounters for each neighborhood. Venues will be grouped by neighborhood.

Clustering of the venues will be subsequently performed using the *k*-means clustering algorithm.

The Folium library will be used to visualize the clusters in the maps of New York City and Toronto.

Each cluster of New York City and Toronto will be analyzed in order to find similarities in venues. The objective will be to select the clusters that are similar the most.

Methodology

For this project, a code was written in a Jupyter notebook (IBM Watson Studio) in which several programs were imported:

1. Pandas, a Python library for data manipulation and analysis.
2. Numpy, a library of Python which handles data in multidimensional arrays.
3. Json, a library that handles JavaScript Object Notation (JSON) files that store data, for example, New York City.
4. Geopy, a Python client that allows location of coordinates of neighborhoods, postal codes, cities, countries, landmarks, across the world.
5. Requests, a library that makes HTTP requests to retrieve data from a specified source.
6. Matplotlib, a library that allows the creation of visualization in Python.
7. K-means, a clustering algorithm to explore our data to identify subgroups with similar characteristics based on distinctive defined features.
8. Folium, a Python library especially useful when spatial data visualization is needed and generate insights from that generated clusters.

A Foursquare developer account was created for this project. The account provided a Client ID and a Client Secret to interrogate the database. For this purpose, a getNearbyVenues function was defined with a request URL:

```
url =
'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&
v={}&ll={},{}&radius={}'.format(
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            radius)
```

This URL was used to make a GET request to Foursquare to retrieve selective data from the neighborhoods that were filled into dataframes with new venues for each Toronto and New York City, specifically the venues with their latitude, longitude and category.

Once all data for analysis were obtained, they had to be prepared for segmentation, since categorical data cannot be used in clustering algorithms such as K-means. These algorithms require all input variables and output variables to be numeric. For this purpose, a one-hot encoding was applied to the venues dataframes of Toronto and New York City.

The resulting one-hot encoded dataframes venues for each city were then grouped by neighborhoods and the mean of occurrence for each category was calculated. These new grouped dataframes contained numerical values that were required for the clustering algorithm.

The new grouped dataframes were used to obtain the 10 most common venues for each neighborhood in Toronto and New York City through the *return_most_common_venues* function. Additionally, 10 cluster labels for each Toronto and New York City were obtained from these dataframes by using the clustering algorithm K-means.

Finally, in order to have all the most common venues for each neighborhood with their coordinates, the cluster labels were added to each of the original pandas dataframes of New York City (Table 1) and Toronto (Table 2) that were merged with the corresponding dataframes containing the most common venues.

The merged dataframes were then used with the Folium library to visualize the clusters in a map of either Toronto or New York City.

Results

The map of New York City and Toronto with the corresponding clusters are displayed in Figure 1 and Figure 2, respectively.

Figure 1. The 10 clusters generated by K-means algorithm superimposed on the map of New York City.

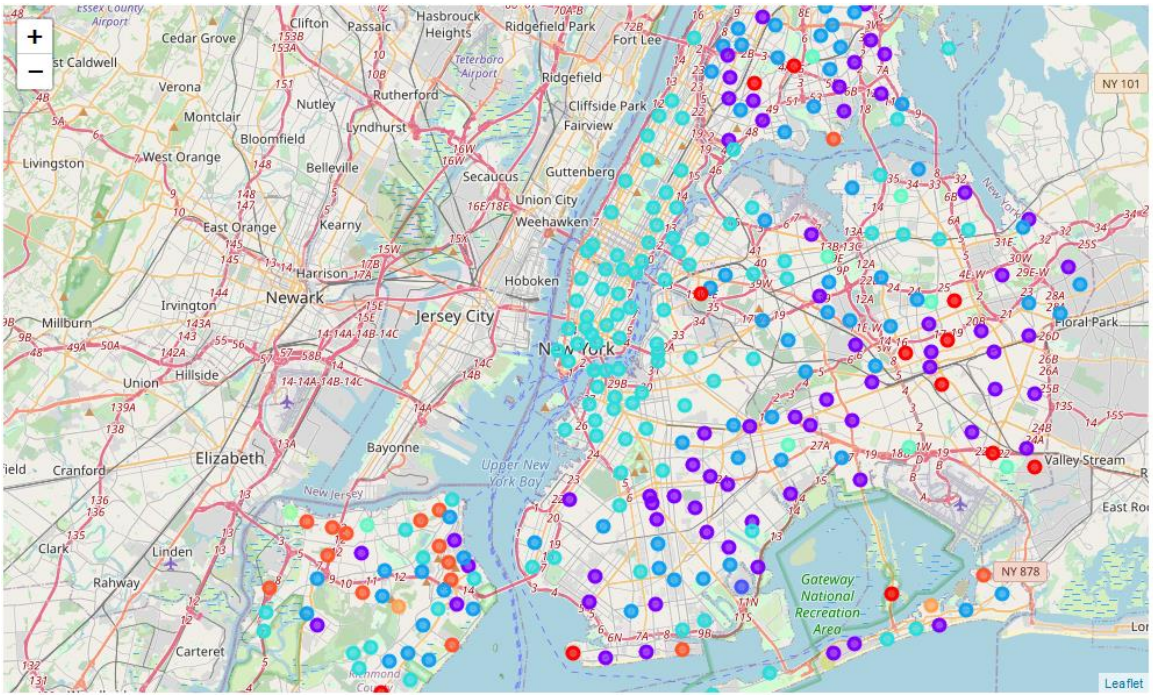
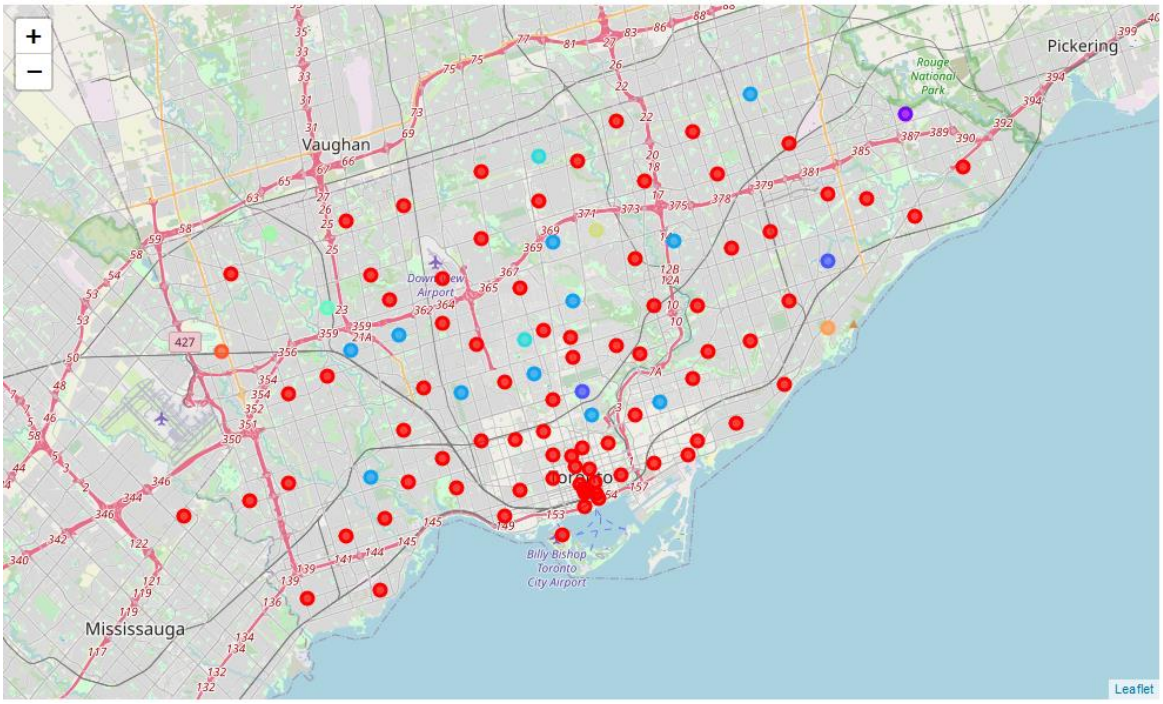


Figure 2. The 10 clusters generated by K-means algorithm superimposed on the map of Toronto.



Toronto, with 10 boroughs and 96 neighborhoods was segmented into 10 clusters. Likewise, New York City, with 5 boroughs and 300 neighborhoods was also segmented into 10 clusters for comparison.

A list of each individual cluster can be found in the Jupyter notebook (https://github.com/Ricardo-Reategui/Coursera_Capstone/blob/master/Capstone%20Project.ipynb). The clusters are not of equal size, some of them have only one neighborhood, others have more than 70. The results reveal similarities in some clusters from New York and Toronto. For example, Cluster 4 (Table 3) from Toronto compares very well with Cluster 9 (Table 4) from New York. Both have parks as the most common venue.

Table 3. Cluster 4 of the segmentation of Toronto

Borough	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
North York	Parkwoods	Bus Stop	Park	Food & Drink Shop	Women's Store	Drugstore	Donut Shop	Dog Run	Distribution Center	Discount Store	Diner
York	Caledonia-Fairbanks	Park	Women's Store	Pool	Curling Ice	Drugstore	Donut Shop	Dog Run	Distribution Center	Discount Store	Diner
East York	East Toronto	Park	Coffee Shop	Convenience Store	Women's Store	Dance Studio	Eastern European Restaurant	Drugstore	Donut Shop	Dog Run	Distribution Center
North York	North Park, Maple Leaf Park, Upwood Park	Park	Construction & Landscaping	Bakery	Women's Store	Deli / Bodega	Eastern European Restaurant	Drugstore	Donut Shop	Dog Run	Distribution Center
Central Toronto	Lawrence Park	Park	Swim School	Bus Line	Dance Studio	Drugstore	Donut Shop	Dog Run	Distribution Center	Discount Store	Diner
York	Weston	Park	Women's Store	Dance Studio	Eastern European Restaurant	Drugstore	Donut Shop	Dog Run	Distribution Center	Discount Store	Diner
North York	York Mills West	Park	Bank	Convenience Store	Women's Store	Deli / Bodega	Eastern European Restaurant	Drugstore	Donut Shop	Dog Run	Distribution Center
Central Toronto	Forest Hill North & West	Park	Jewelry Store	Trail	Sushi Restaurant	Curling Ice	Drugstore	Donut Shop	Dog Run	Distribution Center	Discount Store
Scarborough	Milliken, Agincourt North, Steeles East, L'Amo...	Park	Playground	Bakery	Curling Ice	Drugstore	Donut Shop	Dog Run	Distribution Center	Discount Store	Diner
Downtown Toronto	Rosedale	Park	Playground	Trail	Cuban Restaurant	Drugstore	Donut Shop	Dog Run	Distribution Center	Discount Store	Diner
Etobicoke	The Kingsway, Montgomery Road, Old Mill North	River	Park	Women's Store	Drugstore	Donut Shop	Dog Run	Distribution Center	Discount Store	Diner	Dim Sum Restaurant

Table 4. Cluster 9 of the segmentation of New York City

Borough	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
Queens	Somerville	Park	Women's Store	Irish Pub	Entertainment Service	Ethiopian Restaurant	Event Service	Event Space	Exhibit	Factory	Falafel Restaurant

Borough	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
Staten Island	Todt Hill	Park	Women's Store	Irish Pub	Entertainment Service	Ethiopian Restaurant	Event Service	Event Space	Exhibit	Factory	Falafel Restaurant

Discussion

The most common venues in Table 4 are identical, revealing that Somerville in Queens and Todt Hill in Staten Island share similar characteristics. Similar venues in Toronto align with Rosedale, York Mills West, and Weston.

Similar analysis can be performed by comparing other clusters between Toronto and New York City.

From the analysis of the clusters and the maps rendered by Folium it is interesting to point out that the clusters are dispersed across both cities Toronto and New York, each borough in general seems to have venues that can be present in other boroughs. This is one characteristic that makes New York and Toronto similar.

If a company is offering relocation services to individuals who wanted to move, for example, from New York City to Toronto, it could take advantage of the information retrieved from Foursquare to recommend its client that neighborhoods in Toronto such as Rosedale, York Mills West, or Weston may contain similar amenities as Somerville or Todt Hill in New York, easing the transition that usually occurs when and individual or family move from one city to another.

Following the methodology described above, all this information can be obtained by interrogating Foursquare in less than a day, allowing the relocation company to offer its clients with a comprehensive package linking the analysis described with real estate agents and other services.

Conclusion

The Foursquare database was interrogated to obtain information about most common venues clusterd and grouped by neighborhoods in Toronto and New York City.

To obtain this information it is necessary to have a Foursquare account, data with the neighborhood coordinates of both Toronto and New York City and a Jupyter notebook to run the code.

The methodology was explained with an example of a company offering relocation services to individuals who wanted to move, for example, from New York City to Toronto. Analysis of the clusters obtained from Foursquare and the clustering algorithm K-means would allow the company to recommend its client that neighborhoods in Toronto such as Rosedale, York Mills West, or Weston are somehow similar as Somerville or Todt Hill in New York.

The relocation company could apply this methodology to other cities across the globe providing rapid and targeted information to its clients who wants to make informed decisions when they move to one place to another.