The Battle of Neighborhoods between Great Lakes Cities: Cleveland and Toronto

Introduction of the problem and background

Toronto, ON and Cleveland, OH are two of the major cities in the Great Lakes area. The distance is less than 300 miles, which will take about a 4-hour driving. We can find a lot of similarities between the two cities: lake-shore cities, English as the predominant language, nearby the US-Canada border, etc. On the other hand, the two cities are also very different, for example, Toronto has larger area and population, Cleveland has warmer weather.

Since Toronto and Cleveland are so close by each other, people may have such questions: is the life style in these two cities same? What are the similarities and what are the major difference?

If we look at these in a business view point, some interesting questions might arise. A successful Cleveland restaurant owner want to know if she can repeat her business in Toronto, where should the new restaurant be located. A real estate agent would recommend a house to his client who plan to move from Toronto to Cleveland but rather keep the similar living environment. Questions like these can related to anyone or any business area.

The main targets of this project are:

1. to figure out the which neighborhoods are similar to each other,
2. to find whether there are neighborhoods distribution difference between of the two cities.

Data selection and sources

For this project, the following data is needed:

1. Neighborhoods names and coordinates of Toronto and Cleveland.
2. Venues data

Data sources:

Location Data of Toronto

- Toronto: We can use Toronto's information on Wikipedia as resource:

  https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- BeautifulSoup is a Python library used for pulling data out of HTML. We will use it to parse the Wikipedia page
- Geolocation data: http://cocl.us/Geospatial_data

Location Data of Cleveland

- We use the Cleveland neighborhoods data downloaded from Keggle.com

https://www.kaggle.com/jkortis2121/cleveland-neighborhoods-clean

- This data set contains the latitudes and longitudes of each neighborhood.

Venues Data (Foursquare API)

- Foursquare API provides information about venues and geolocation. We will use Foursquare API to get the venue data for all neighborhoods. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data such as name, location, hours, rating, prices, etc.

Approach:

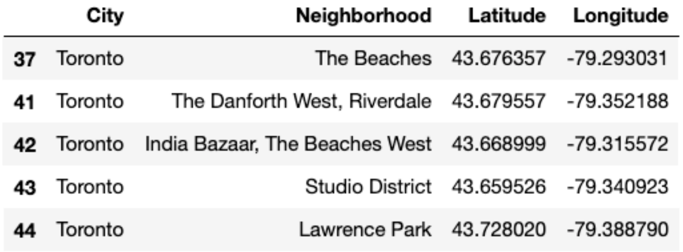
1. Collect required neighborhoods data from sources listed above, construct a data set for both city
2. Using Foursquare API to get all venues for each neighborhood.
3. Find the top 5 most popular venues in each neighborhood.
4. Apply K-Means approach to cluster all neighborhoods
5. Visualize clusters on city maps
6. Analyze the most important cluster of each city to find the differences/similarity

Methodology

1. Data preparing

Toronto:

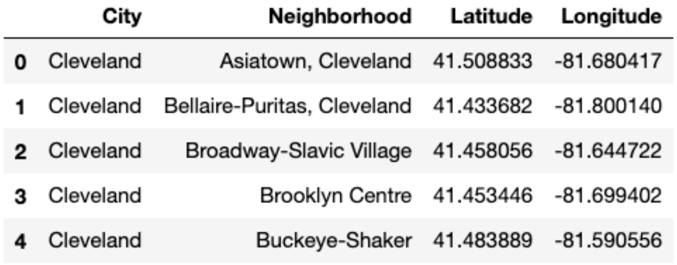
- We use BeautifulSoup to scraping table from the Wikipedia page that contains the neighborhoods name, borough, postal code of Toronto. Next, remove the instances that the “Borough” is not assigned or contain any null field.
- Combine the table we obtained with the geolocation data, such that in the data set each neighborhood has its latitude and longitude information added.
- Since there are 103 neighborhoods in Toronto, to refine our research, we will focus only on boroughs that contain the word “Toronto” (total 39 neighborhoods). Also, in order to do further analysis, we add city name into the table. After processing, the neighborhood dataset of city Toronto looks like the table below:



	City	Neighborhood	Latitude	Longitude
37	Toronto	The Beaches	43.676357	-79.293031
41	Toronto	The Danforth West, Riverdale	43.679557	-79.352188
42	Toronto	India Bazaar, The Beaches West	43.668999	-79.315572
43	Toronto	Studio District	43.659526	-79.340923
44	Toronto	Lawrence Park	43.728020	-79.388790

Cleveland:

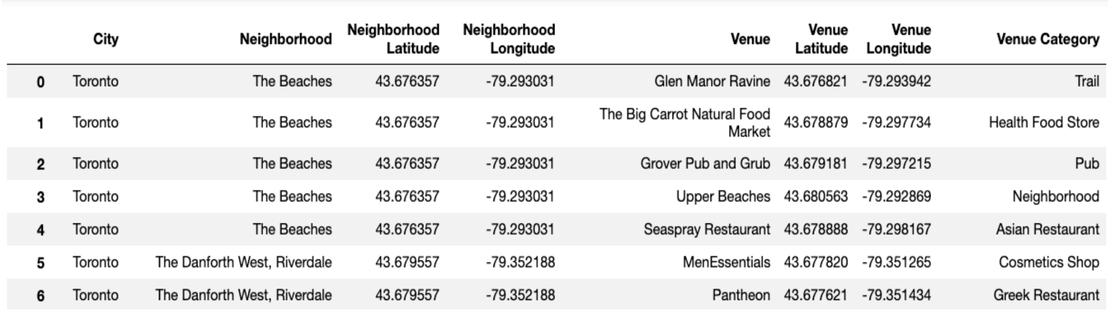
- The data we downloaded from keggle.com already contained all the information we need, including neighborhood name, latitude and longitude.
- There are 33 neighborhoods in the Cleveland dataset, similar to that of Toronto.



	City	Neighborhood	Latitude	Longitude
0	Cleveland	Asiatown, Cleveland	41.508833	-81.680417
1	Cleveland	Bellaire-Puritas, Cleveland	41.433682	-81.800140
2	Cleveland	Broadway-Slavic Village	41.458056	-81.644722
3	Cleveland	Brooklyn Centre	41.453446	-81.699402
4	Cleveland	Buckeye-Shaker	41.483889	-81.590556

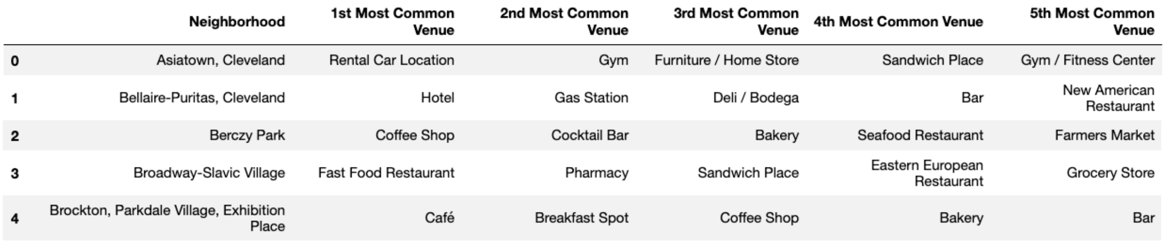
2. Obtain venue data

Once we combine the neighborhood data of two cities, we can use Foursquare API to obtain the venues of each neighborhood. return a JSON file and we need to turn that into a data-frame. In this project I've chosen 100 popular spots for each neighborhood with a radius of 500 meters from their given latitude and longitude information. Here is the head of a returned list with venue name, venue category, latitude and longitude from Foursquare API.



	City	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Toronto	The Beaches	43.676357	-79.293031	Glen Manor Ravine	43.676821	-79.293942	Trail
1	Toronto	The Beaches	43.676357	-79.293031	The Big Carrot Natural Food Market	43.678879	-79.297734	Health Food Store
2	Toronto	The Beaches	43.676357	-79.293031	Grover Pub and Grub	43.679181	-79.297215	Pub
3	Toronto	The Beaches	43.676357	-79.293031	Upper Beaches	43.680563	-79.292869	Neighborhood
4	Toronto	The Beaches	43.676357	-79.293031	Seaspray Restaurant	43.678888	-79.298167	Asian Restaurant
5	Toronto	The Danforth West, Riverdale	43.679557	-79.352188	MenEssentials	43.677820	-79.351265	Cosmetics Shop
6	Toronto	The Danforth West, Riverdale	43.679557	-79.352188	Pantheon	43.677621	-79.351434	Greek Restaurant

Furthermore, we can sort the venues by popularity and find the top 5 most popular venues for each neighborhood. Here is a head of sorted venues list:

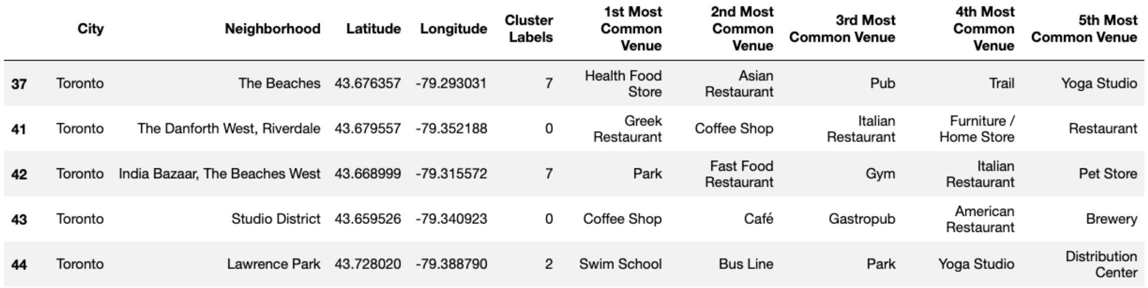


	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Asiatown, Cleveland	Rental Car Location	Gym	Furniture / Home Store	Sandwich Place	Gym / Fitness Center
1	Bellaire-Puritas, Cleveland	Hotel	Gas Station	Deli / Bodega	Bar	New American Restaurant
2	Berczy Park	Coffee Shop	Cocktail Bar	Bakery	Seafood Restaurant	Farmers Market
3	Broadway-Slavic Village	Fast Food Restaurant	Pharmacy	Sandwich Place	Eastern European Restaurant	Grocery Store
4	Brockton, Parkdale Village, Exhibition Place	Café	Breakfast Spot	Coffee Shop	Bakery	Bar

3. K-Means clustering

Now we are ready to do some exploratory analysis with the venue data. I choose K-Means as clustering method here. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and it is highly suited for this project as well.

I make K=10 to cluster the all neighborhoods into 10 clusters. Here is the merged table with cluster labels added for each neighborhood.



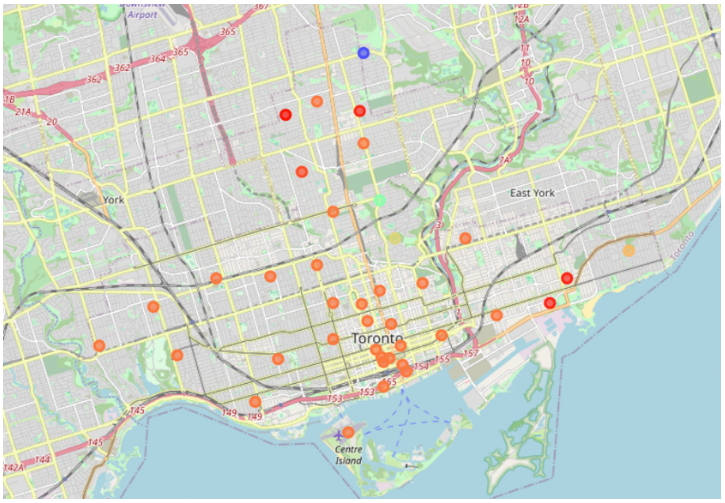
	City	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
37	Toronto	The Beaches	43.676357	-79.293031	7	Health Food Store	Asian Restaurant	Pub	Trail	Yoga Studio
41	Toronto	The Danforth West, Riverdale	43.679557	-79.352188	0	Greek Restaurant	Coffee Shop	Italian Restaurant	Furniture / Home Store	Restaurant
42	Toronto	India Bazaar, The Beaches West	43.668999	-79.315572	7	Park	Fast Food Restaurant	Gym	Italian Restaurant	Pet Store
43	Toronto	Studio District	43.659526	-79.340923	0	Coffee Shop	Café	Gastropub	American Restaurant	Brewery
44	Toronto	Lawrence Park	43.728020	-79.388790	2	Swim School	Bus Line	Park	Yoga Studio	Distribution Center

Result

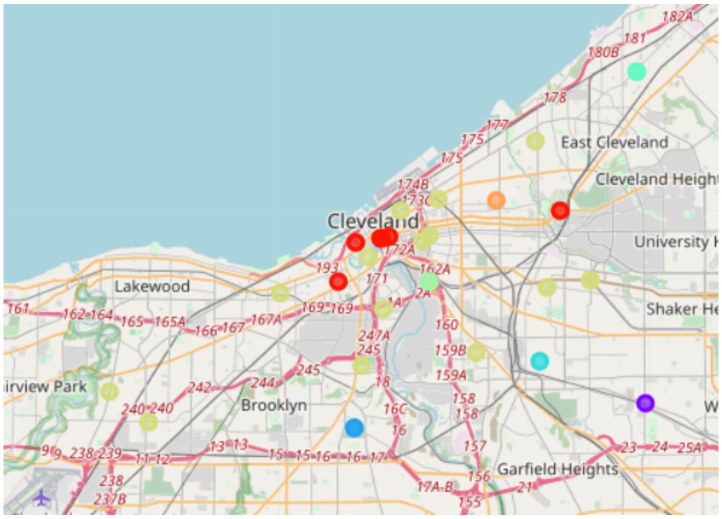
1. Cluster analysis

Are the two cities similar to each other?To address this question, we can compare the numbers of different clusters in each city to find which one is the most important for the city.

First, I used the Folium library to visualize the geographic details of each city. Clusters were marked as dot. From the maps, we can see clearly that in Toronto and Cleveland, the dominant clusters are different.

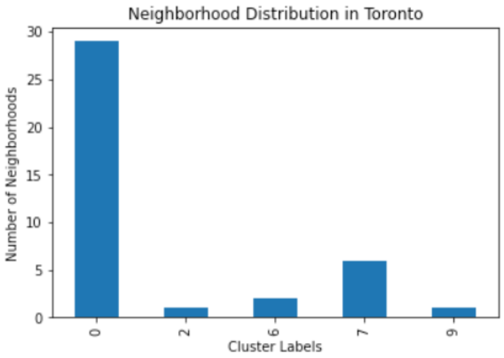


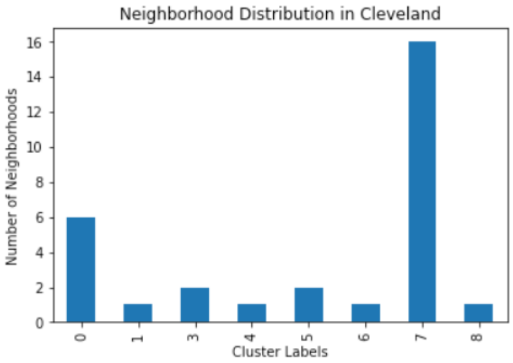
Cluster Map of Toronto



Cluster Map of Cleveland

Next, I grouped clusters in each city to find which clusters are the most important. Bar graphs work very well in showing the differences between numbers of neighborhood.

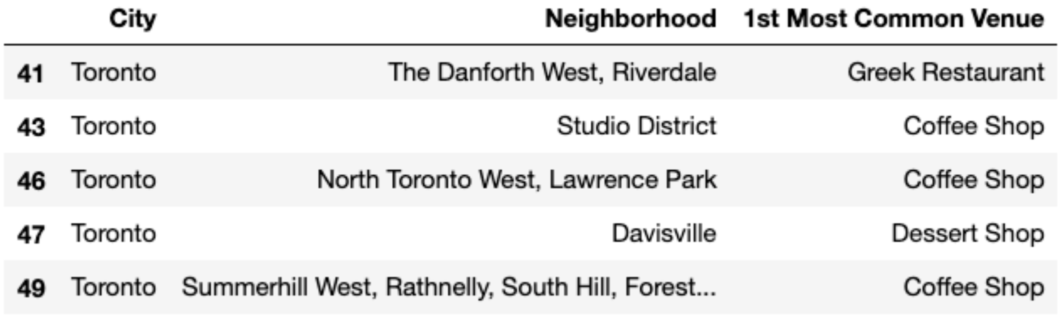




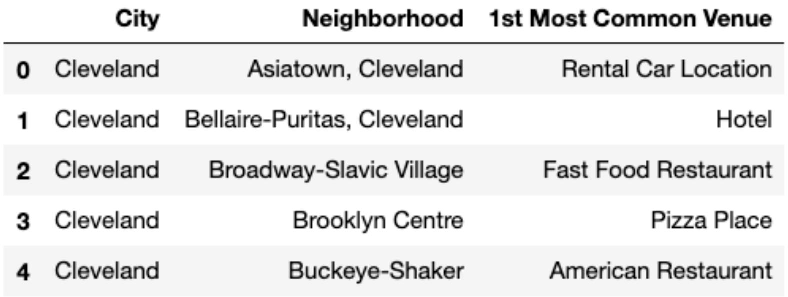
Result shows that cluster 0 is the dominant cluster in Toronto, while it is cluster 7 in Cleveland.

2. Venue analysis

What makes cluster 0 and cluster 7 differ from each other?
Comparing the 1st most popular venue of these two clusters would give us some information to answer this question. Here are the head of list of cluster 0 in Toronto and cluster 7 in Cleveland.

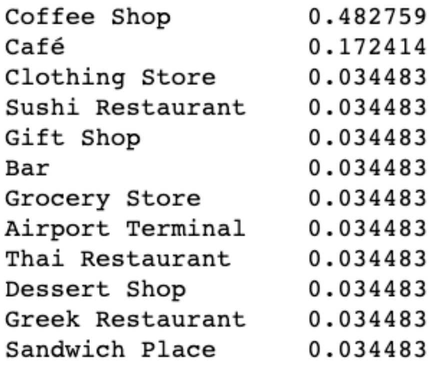


	City	Neighborhood	1st Most Common Venue
41	Toronto	The Danforth West, Riverdale	Greek Restaurant
43	Toronto	Studio District	Coffee Shop
46	Toronto	North Toronto West, Lawrence Park	Coffee Shop
47	Toronto	Davisville	Dessert Shop
49	Toronto	Summerhill West, Rathnelly, South Hill, Forest...	Coffee Shop

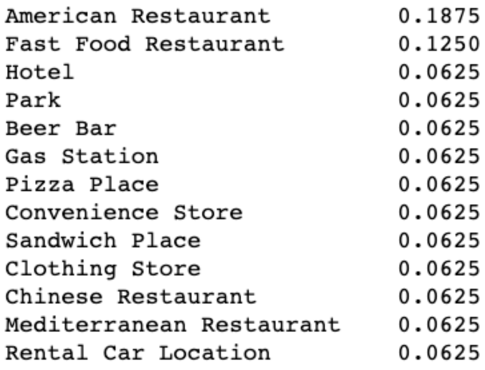


	City	Neighborhood	1st Most Common Venue
0	Cleveland	Asiatown, Cleveland	Rental Car Location
1	Cleveland	Bellaire-Puritas, Cleveland	Hotel
2	Cleveland	Broadway-Slavic Village	Fast Food Restaurant
3	Cleveland	Brooklyn Centre	Pizza Place
4	Cleveland	Buckeye-Shaker	American Restaurant

In addition, proportions of each category in the most common venue were calculated:



```
Coffee Shop          0.482759
Café                 0.172414
Clothing Store       0.034483
Sushi Restaurant     0.034483
Gift Shop            0.034483
Bar                  0.034483
Grocery Store        0.034483
Airport Terminal     0.034483
Thai Restaurant      0.034483
Dessert Shop         0.034483
Greek Restaurant     0.034483
Sandwich Place       0.034483
```



```
American Restaurant          0.1875
Fast Food Restaurant         0.1250
Hotel                        0.0625
Park                         0.0625
Beer Bar                     0.0625
Gas Station                  0.0625
Pizza Place                  0.0625
Convenience Store            0.0625
Sandwich Place               0.0625
Clothing Store               0.0625
Chinese Restaurant           0.0625
Mediterranean Restaurant     0.0625
Rental Car Location          0.0625
```

In Cluster 0, "Coffee Shop" takes 48% and the second popular category is "Café" which takes 17%. On the other hand, in Cluster 7, "American Restaurant" and "Fast Food Restaurant" combined take 31% of all venues.

Discussion

The cluster and venue analysis results show the dominant clusters of Toronto and Cleveland are quite different. In Toronto, the most popular venues are the coffee shops, while in Cleveland are American and fast food restaurants. Thus, if a fast food company want to open a franchise restaurant in Toronto, it might need avoid the neighborhoods that fall into Cluster 0. Similarly, if people who loves coffee want to move to Toronto, he will have a lot neighborhood in considering.

Of course, to compare cities and find their differences and similarities is not that simpe. This project just provides a single viewpoint of this question. Many other factors should be involved in further analysis. Demographics, transportation, weather condition may all affect analysis result.

Also, we should note in this project, we simplified the neighborhood data of Toronto for demonstration purpose. To gain an accurate result, more data points should be included in to the research.

Conclusion

To conclude, this project showed us how the data science knowledge and skills can be used to solve real-life issues. All aspects of our life could encounter some questions just like that aroused in the beginning section of this report. Not only the governments, business owners, but also every individual, could utilize data science to exam or improve the decision-making procedure.