

New York City – Toronto Neighborhoods Clustering


HASSAN FARAHANI

IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

JAN. 2021



Agenda

- Introduction & Business Problem
 - Data Acquisition and Preparation
 - Exploratory Data Analysis
 - Clustering of Neighbourhoods
 - Discussion
 - Conclusion
- 

Introduction & Business Problem

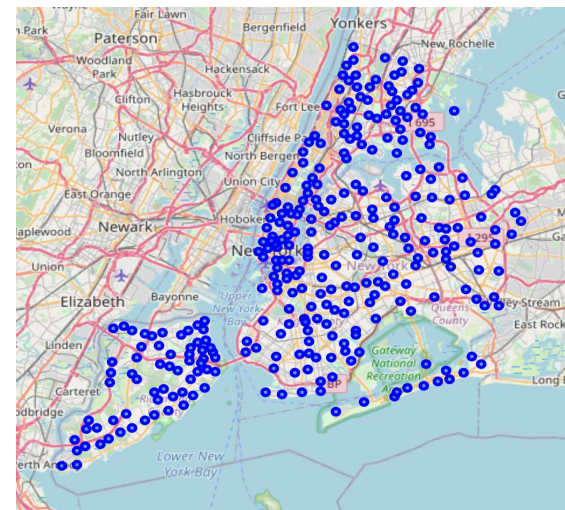
- New York City (USA) and the city of Toronto (Canada) is going to clustered based on the similarities in the venue categories of their neighbourhoods
- Machine Learning method: K-Means algorithm (unsupervised)
- Goal:
 - types of businesses in different neighbourhoods of both cities,
 - the geographic distribution of most common business in both cities,
 - the common businesses in both cities
- Output:
 - how similar or dissimilar these two cities are
 - Suitable business types in both cities
 - Undesirable business type in both cities

Data Acquisition and Preparation - NYC

- **Neighbourhood data**
 - a JSON file provided by the course coordinator, which converted into a dataframe

	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 New York neighbourhood dataframe (306 rows)



New York City map and its neighbourhoods (radius=500m)

Data Acquisition and Preparation - NYC

- **Venues data**

- Foursquare API to get venues data for each neighbourhood
- Foursquare API requirements:
 - Credentials (by registering a Foursquare developer account)
 - Latitude and longitude of each neighbourhood
 - The radius used by API to search for venues
 - Number of returned venues
- URL example to make request to the Foursquare API:

`https://api.foursquare.com/v2/venues/search?&client_id=9823&client_secret=8565&v=20180605&ll=40.87655077879964,-73.91065965862981&radius=500&limit=100`

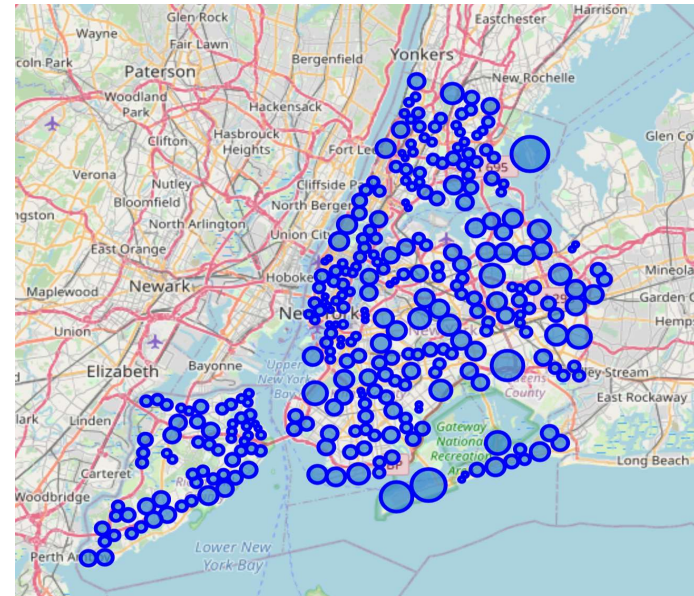
- As the area cover by neighborhoods overlaps, it is better to calculate different radius for each neighborhood using the following formula (latitude, ϕ , and longitude, λ):

$$d = 2r \arcsin \left(\sqrt{\sin^2 \left(\frac{\phi_2 - \phi_1}{2} \right) + \cos(\phi_1) \cos(\phi_2) \sin^2 \left(\frac{\lambda_2 - \lambda_1}{2} \right)} \right)$$

Data Acquisition and Preparation – NYC (Contd.)

	Borough	Neighborhood	Latitude	Longitude	Radius
0	Bronx	Wakefield	40.894705	-73.847201	565.191216
1	Bronx	Co-op City	40.874294	-73.829939	481.235842
2	Bronx	Eastchester	40.887556	-73.827806	742.734963
3	Bronx	Fieldston	40.895437	-73.905643	388.116468
4	Bronx	Riverdale	40.890834	-73.912585	388.116468

New York neighbourhood dataframe along with radius (first five rows)



New York City map and its neighbourhoods (different radius for each neighbourhood)

Data Acquisition and Preparation – NYC (Contd.)

- **Venues data**

- After sending request to the API using the new neighbourhood dataframe, a JSON file containing all the venues details for a specific neighbourhood will be returned.
- Here is the venues dataframe obtained from transformation of the JSON file:
- Number of rows: 9915 and number of unique venue categories: 422
- Some venue categories like Metro Station, Bus Stop, Bus Line, Train Station, Gas Station, Neighborhood, Field, Bridge, Office, Train, Platform, River and Church have been removed from the dataframe

	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	Walgreens	40.896528	-73.844700	Pharmacy
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

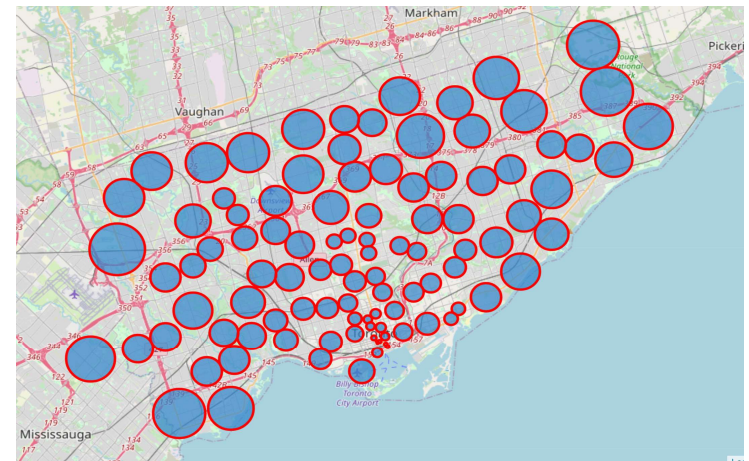
Data Acquisition and Preparation - Toronto

- **Neighborhood data**

- Combining data from two different sources:
 - https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M contains neighbourhood and Borough data
 - .csv file containing coordinates of each neighbourhood

	Postal Code	Borough	Neighborhood	Latitude	Longitude	Radius
0	M3A	North York	Parkwoods	43.753259	-79.329656	992.961518
1	M4A	North York	Victoria Village	43.725882	-79.315572	1018.563373
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	614.195007
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	934.471639
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	256.276246

Toronto neighbourhood dataframe (103 rows)



*Toronto map and its neighbourhoods
(different radius for each neighbourhood)*

Data Acquisition and Preparation - Toronto

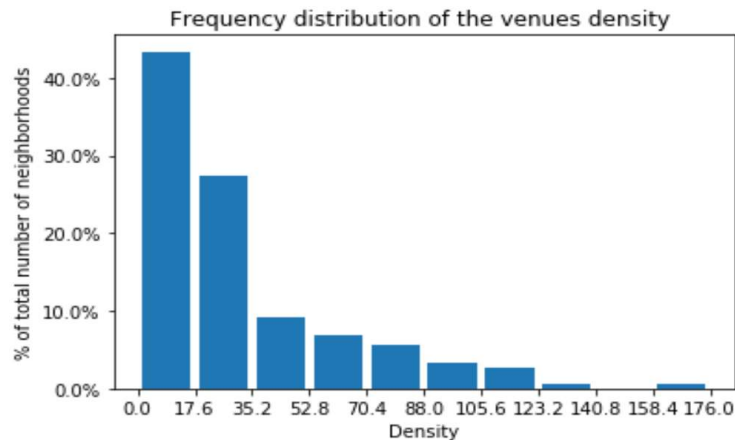
- **Venues data**

- Foursquare API to get venues data for each neighbourhood
- Number of rows: 3207 and number of unique venue categories: 314

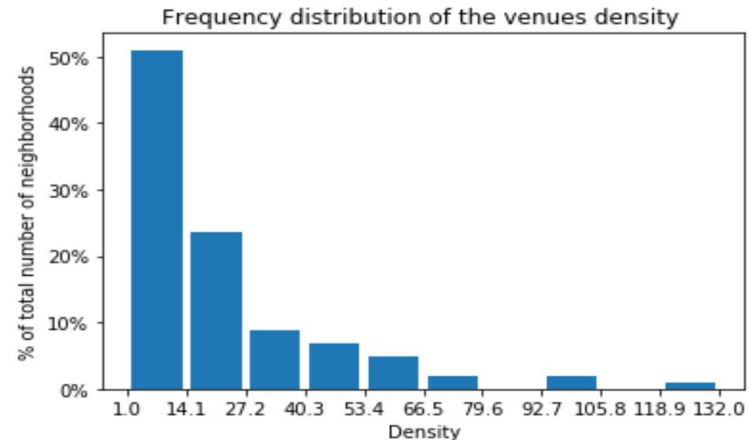
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Allwyn's Bakery	43.759840	-79.324719	Caribbean Restaurant
1	Parkwoods	43.753259	-79.329656	Tim Hortons	43.760668	-79.326368	Café
2	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
3	Parkwoods	43.753259	-79.329656	Bruno's valu-mart	43.746143	-79.324630	Grocery Store
4	Parkwoods	43.753259	-79.329656	High Street Fish & Chips	43.745260	-79.324949	Fish & Chips Shop

Exploratory Data Analysis

- **Density (number of venues per distance(km)) comparison**
 - Given that each neighbourhood has a different radius, it's better to represent the venues per neighbourhood in terms of density
 - Toronto has a low population density compared to NYC



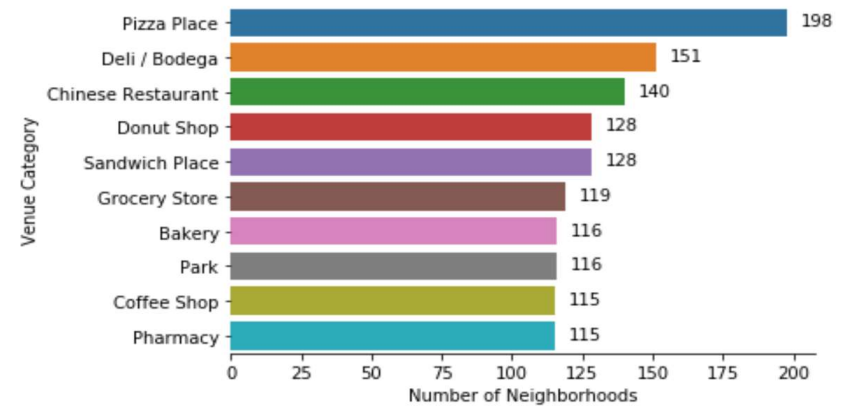
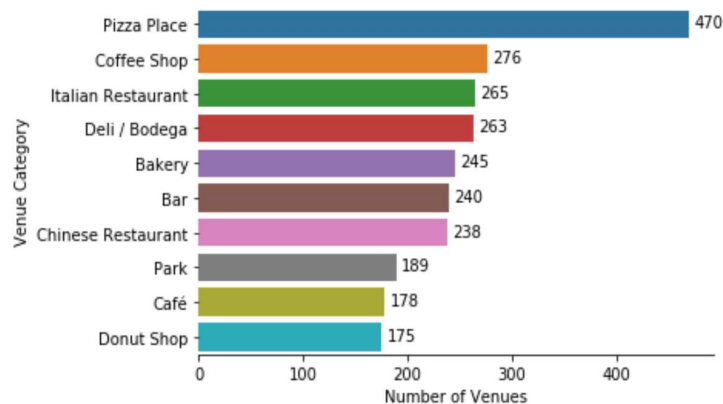
New York City



Toronto

Exploratory Data Analysis

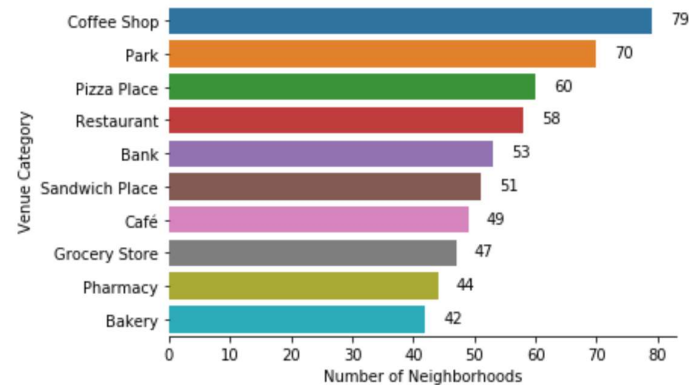
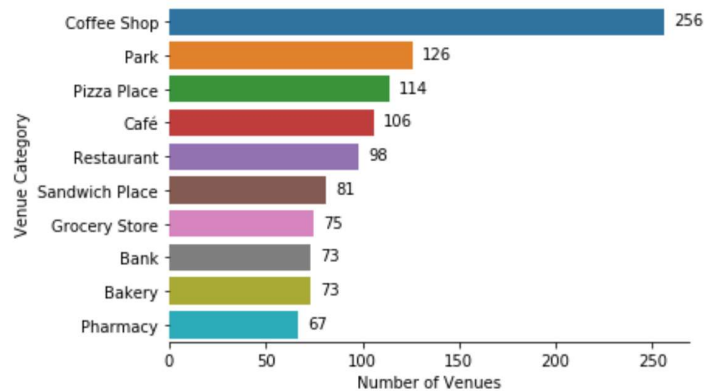
- **Most common & widespread venue categories for NYC**
 - which venue categories have more number of venues (more common)
 - which venue categories are existed in more number of neighbourhoods (more widespread)



New York City

Exploratory Data Analysis

- **Most common & widespread venue categories for Toronto**
 - which venue categories have more number of venues (more common)
 - which venue categories are existed in more number of neighbourhoods (more widespread)



Toronto

Clustering of Neighbourhoods

- ML method: K-Means algorithm (unsupervised)
- pre-processing steps:
 - Applying one-hot encoding on the 'Venue Category' feature of both dataframes.
 - Aggregating each dataframe
 - Combining both dataframes

Clustering of Neighbourhoods – one-hot encoding

- if any venue category (columns) has the value of one, it means that specific neighbourhood has a venue with that category.

	Neighborhood	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Lounge	Airport Service	American Restaurant	Animal Shelter	Antique Shop	Aquarium	Arcade	Arepa Restaurant	Argentir Restau
0	Wakefield	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Wakefield	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Wakefield	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Wakefield	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Wakefield	0	0	0	0	0	0	0	0	0	0	0	0	0	0

NYC

	Neighborhood	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Amphitheater	Animal Shelter	Antique Shop	G
0	Parkwoods	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Parkwoods	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Parkwoods	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Parkwoods	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Parkwoods	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Toronto

Clustering of Neighbourhoods – aggregation

	Neighborhood	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Lounge	Airport Service	American Restaurant	Animal Shelter	Antique Shop	Aquarium	Arcade	Arepa Restaurant	Argentir Restau
0	Allerton	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	
1	Annadale	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.083333	0.0	0.0	0.0	0.0	0.0	
2	Arden Heights	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	
3	Arlington	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.250000	0.0	0.0	0.0	0.0	0.0	
4	Arrochar	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	

NYC

	Neighborhood	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Amphitheater	Animal Shelter	Antique Shop	G
0	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	
1	Aldenwood, Long Branch	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	
2	Bathurst Manor, Wilson Heights, Downsview North	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.027027	0.0	0.0	0.0	
3	Bayview Village	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	
4	Bedford Park, Lawrence Manor East	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.021739	0.0	0.0	0.0	

Toronto

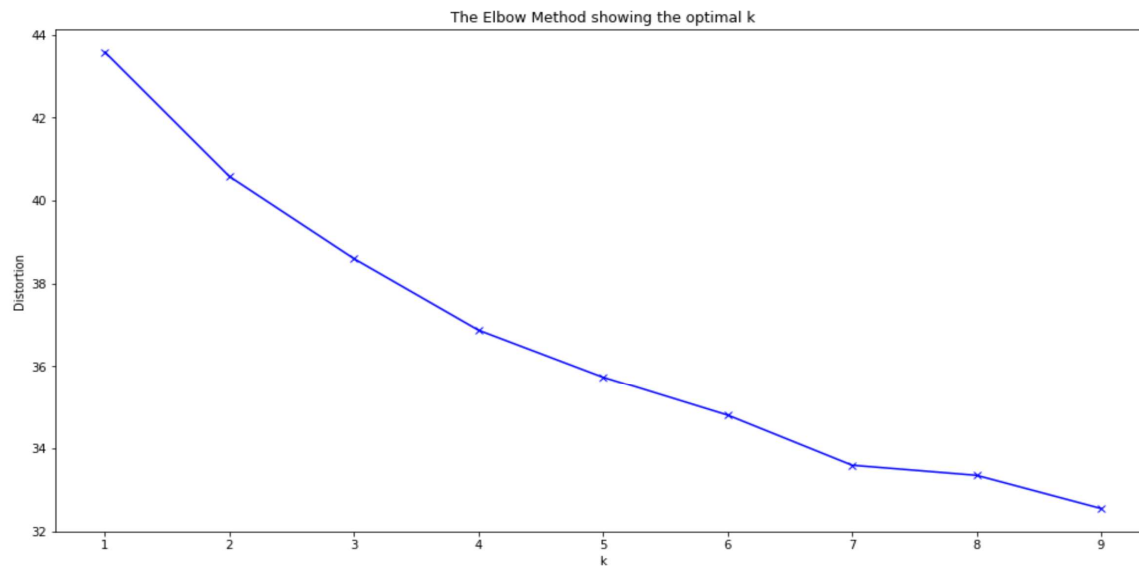
- each venue category compose how many percent of the returned venues for a specific neighbourhood

Clustering of Neighbourhoods – dataframes combination

	Neighborhood	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Lounge	Airport Service	American Restaurant	Animal Shelter	Antique Shop	Aquarium	Arcade	Arepa Restaurant	Argo Res
298	Woodlawn-nyc	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.022727	0.0	0.0	0.0	0.022727	0.000000	
299	Woodrow-nyc	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	
300	Woodside-nyc	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.044776	0.0	0.0	0.0	0.000000	0.014925	
301	Yorkville-nyc	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	
302	Agincourt-to	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	
303	Alderwood, Long Branch- to	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	

- Note that before combining two dataframes, we added '-nyc' and '-to' to the end of the neighbourhoods in NYC and Toronto respectively to distinguish between the neighbourhoods in both cities.

K-Means – Elbow method



- To apply K-Means algorithm, we need to predefine the number of clusters
- To find the optimum number of clusters, we used elbow method
- the optimal value of the number of clusters is defined as 7

K-Means (k = 7)

- output of K-Means method is seven clusters with the cluster labels of 0, 1, 2, 3, 4, 5, and 6

Cluster No.	Number of Neighbourhoods
0	297
1	14
2	4
3	1
4	85
5	1
6	2

- To obtain these numbers, we added the cluster labels as a new column (with the name of 'Cluster Labels') to the combined dataframe

Results

Venue Category % of Venues		
0	Coffee Shop	3.65207
1	Park	2.86153
2	Deli / Bodega	2.71033
3	Pizza Place	2.59208
4	Italian Restaurant	2.55749

Cluster 1

Venue Category % of Venues		
0	Park	47.7976
1	Trail	7.61905
2	Coffee Shop	5
3	Beach	3.57143
4	Hotel	3.57143

Cluster 2

Venue Category % of Venues		
0	Deli / Bodega	70.8333
1	Market	12.5
2	Pharmacy	8.33333
3	Spanish Restaurant	8.33333
4	Plaza	0

Cluster 3

Venue Category % of Venues		
0	Construction & Landscaping	100
1	ATM	0
2	Pizza Place	0
3	Puerto Rican Restaurant	0
4	Public Art	0

Cluster 4

Venue Category % of Venues		
0	Pizza Place	13.3041
1	Pharmacy	4.07363
2	Bank	4.03491
3	Grocery Store	3.39211
4	Deli / Bodega	3.29925

Cluster 5

Venue Category % of Venues		
0	Boat or Ferry	100
1	ATM	0
2	Pizza Place	0
3	Puerto Rican Restaurant	0
4	Public Art	0

Cluster 6

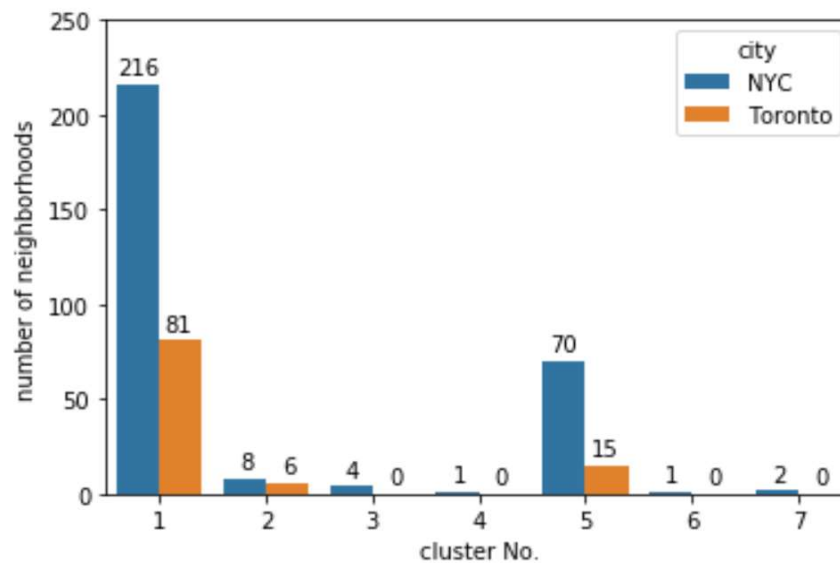
Venue Category % of Venues		
0	Beach	100
1	ATM	0
2	Playground	0
3	Puerto Rican Restaurant	0
4	Public Art	0

Cluster 7

Discussion

- In clusters 4, 6, and 7, there is only one venue category
- In both cluster 2 and 3, there is one venue category with a very high contribution compared to other venue categories in that cluster.
 - For example, 47.7 % of the venue categories of cluster two belongs to 'Park' category
- There is a uniform distribution in the most common venue categories of clusters 1 and 5
- The venue category of 'Deli / Bodega' is the only venue category that exist in three different clusters of (1, 3 and 5)

Discussion (contd.)



- There are four clusters (3, 4, 6, and 7) that Toronto has no contribution in them; but both cities have contribution in three clusters of 1, 2 and 5, which means NYC and Toronto are similar in that they share the same venue categories of 'Coffee Shop', 'Park' and 'Pizza Place'

Conclusion

- In this project, we clustered the neighbourhoods in NYC and Toronto based on the venue categories using K-Means algorithm
- Using Elbow method, the optimal number of clusters was defined as 7
- This clustering could help business individuals about:
 - business type that can thrive in both cities
 - The best locations for the business
 - Type of businesses which are suitable in each city
 - Type of business which are not desirable in both cities
- It could also help Tourists, travellers and new immigrants to find a place to visit, stay or hangout.