DATA SCIENCE CAPSTONE PROJECT

SIMILARITY AND DISSIMILARITY BETWEEN DOWNTOWN NEW YORK AND DOWNTOWN TORONTO

05 November 2020

1. Introduction

1.1 Background

Manhattan is the most densely populated of New York City's 5 boroughs. It's mostly made up of Manhattan Island, bounded by the Hudson, East and Harlem rivers. Among the world's major commercial, financial and cultural centers, it's the heart of "the Big Apple." Its iconic sites include skyscrapers such as the Empire State Building, neon-lit Times Square and the theaters of Broadway. Downtown Toronto is a buzzing area filled with skyscrapers, restaurants, nightlife, and an eclectic mix of neighbourhoods. It's also home to iconic attractions like the CN Tower, St. Lawrence Market, and the Royal Ontario Museum, with exhibits on natural history. Bloor Street is an upscale shopping area, and the Eaton Centre is a huge, multistory mall. On the lake, the Harbourfront area has parks and cultural venues.

Lower Manhattan, also known as Downtown New York or Downtown New York, is the southernmost part of Manhattan, the central borough for business, culture, and government in New York City. What can be the similarity and the dissimilarity between the neighborhoods of the two boroughs: Manhattan and Downtown Toronto? Is Downtown Toronto more like Downtown New York?

1.2 Interest

Since this report will include information about foursquare location on Toronto and New York cities, the study will interest the mayors of both cities to help them evaluate the downtowns' organization. Indeed it will allow them to open objective discussion and to take informed decisions for the cities infrastructures. The study will also serve as an accessible reference for the entrepreneurs of both cities to develop some areas of business which are showing dissimilarity.

2. Data description

New York dataset exists and is accessible at the link: https://geo.nyu.edu/catalog/nyu_2451_34572. The dataset has four columns which are: 'Borough', 'Neighborhood', 'Latitude', and 'Longitude'. New York Neighborhood has a total of 5 boroughs and 306 neighborhoods. In order to segment the

neighborhoods and explore them specially Manhattan borough, I will extract data on Manhattan Borough. To retrieve Toronto dataset I will scrape the Wikipedia page https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M and parse HTML code using the Python package pandas, and convert data into a Pandas dataframe. This first part of the data contains 180 records with 2 columns: 'Postal Code' and 'Neighbourhood'. I will merge it with coordinate data containing longitude and latitude at the link:'https://cocl.us/Geospatial_data' using 'Postal Code' as the key to join both dataframe into one. Not assigned Borough will be removed and not assigned Neighborhoods will be replaced by the assigned borough. In order to segment the neighborhoods and explore them specially Downtown Toronto borough, I will extract data on that specific Borough.

I will use the Foursquare API to explore neighborhoods in both cities. I will use the explore function to get the most common venue categories in each neighborhood, and then use this feature to group the neighborhoods into clusters. I will use the k-means clustering algorithm to complete the comparison between Manhattan and Downtown Toronto. Finally, I will use the Folium library to visualize the neighborhoods in both cities and their emerging clusters.

2.1 Required Tools

- API FourSquare to explore neighborhoods
- Folium for mapping
- Geocoders library convert an address into latitude and longitude values
- Requests library to handle API request URL

3. Analytic approach

Since the business problem has been clearly stated, the analytic approach step entails expressing the problem in the context of statistical and machine-learning techniques, so that the entity or stakeholders with the problem can identify the most suitable techniques for the desired outcome.

In this case study according to the goal, it is a clustering problem. The suitable model for the geographical data will be k-means.
Criteria for data requirements: data selected should be New York and Toronto geographical data.

3.1 How data will be used to solve the problem?

- Data will help to analyse and build the model. We need extensive data of different neighborhoods.
- The machine learning model should be able to predict the clusters similarity
- To build a good model, the dataset should be rich and contain many observations (rows) and various neighborhoods.

Here are the steps that we will follow:

- Data wrangling: to identify and handle missing value, to standardize and normalize the data.
- Data exploratory by analyzing neighborhoods using visualization, descriptive statistical analysis
- Model development: k-means will be developed to predict the clusters. A Model will help to understand the exact relationship both cities.

4. Data wrangling

4.1 Data wrangling: New York

New York dataset was downloaded as a json file with many features in the following structure:


```
{'type': 'Feature',
 'id': 'nyu_2451_34572.1',
 'geometry': {'type': 'Point',
  'coordinates': [-73.84720052054902, 40.89470517661]},
 'geometry_name': 'geom',
 'properties': {'name': 'Wakefield',
  'stacked': 1,
  'annoline1': 'Wakefield',
  'annoline2': None,
  'annoline3': None,
  'annoangle': 0.0,
  'borough': 'Bronx',
  'bbox': [-73.84720052054902,
   40.89470517661,
   -73.84720052054902,
   40.89470517661]}}
```


For the purpose of our study, I extracted 4 features from each Feature type of the json list and renamed them as followed:

```
Borough = data['properties']['borough']
Neighborhood = data['properties']['name']
Latitude = data['geometry']['coordinates'][1]
Longitude = data['geometry']['coordinates'][0]
```

The extracted features were set into a dataframe and after the transformation the dataset has all 5 boroughs and 306 neighborhoods.

4.2 Data wrangling: Toronto

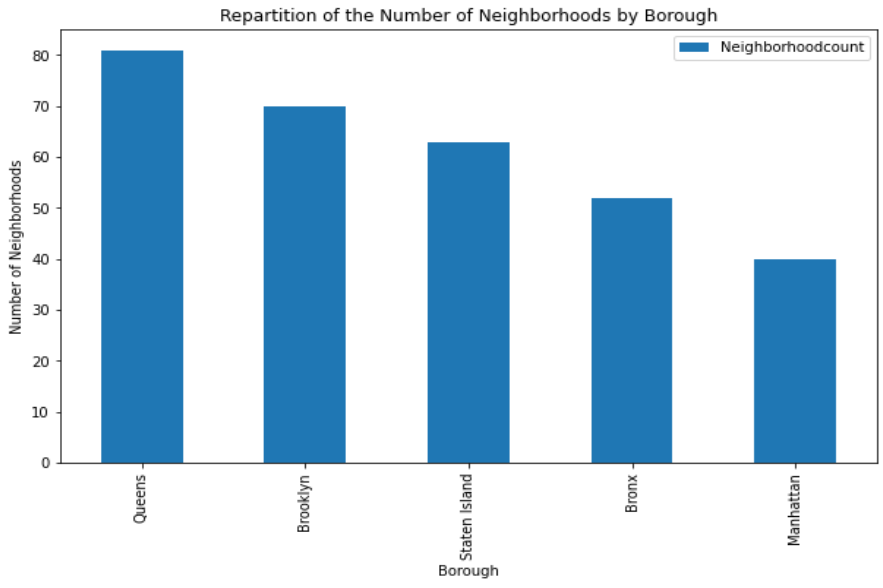
By using the Pandas package the Toronto dataset was scraped from the Wikipedia page: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. The extracted table from that site was set into a dataframe shaped of 180 records of 3 fields: 'Postal Code', 'Borough', 'Neighbourhood'. Two fields were renamed as followed: 'Postal Code': 'PostalCode' and 'Neighbourhood': 'Neighborhood'.

	PostalCode	Borough	Neighborhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
5	M6A	North York	Lawrence Manor, Lawrence Heights
6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government
7	M8A	Not assigned	Not assigned
8	M9A	Etobicoke	Islington Avenue, Humber Valley Village
9	M1B	Scarborough	Malvern, Rouge
10	M2B	Not assigned	Not assigned
11	M3B	North York	Don Mills

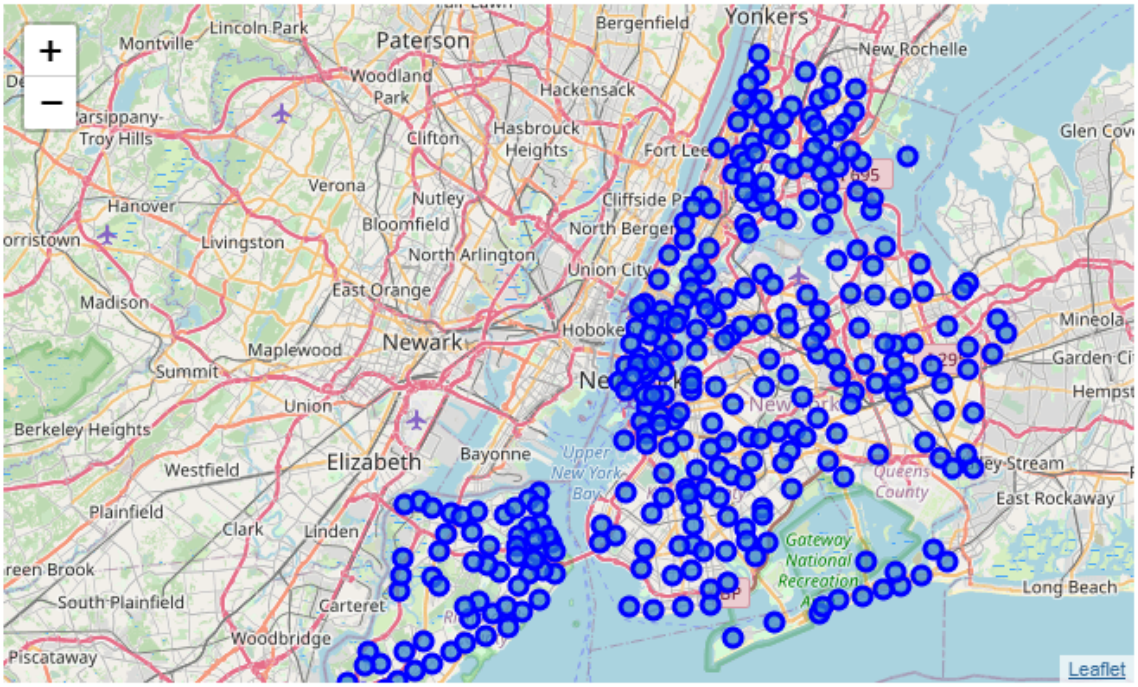
For reliable analysis, I removed all “Not assigned” Borough and then replaced all “Not assigned” Neighborhood by the assigned Borough. After all these changes the remain size of the dataframe was 103 records (42,78% of missing information). I retrieved the latitude and longitude coordinates by using the geocoder.google function.

5. Data exploratory of dataset

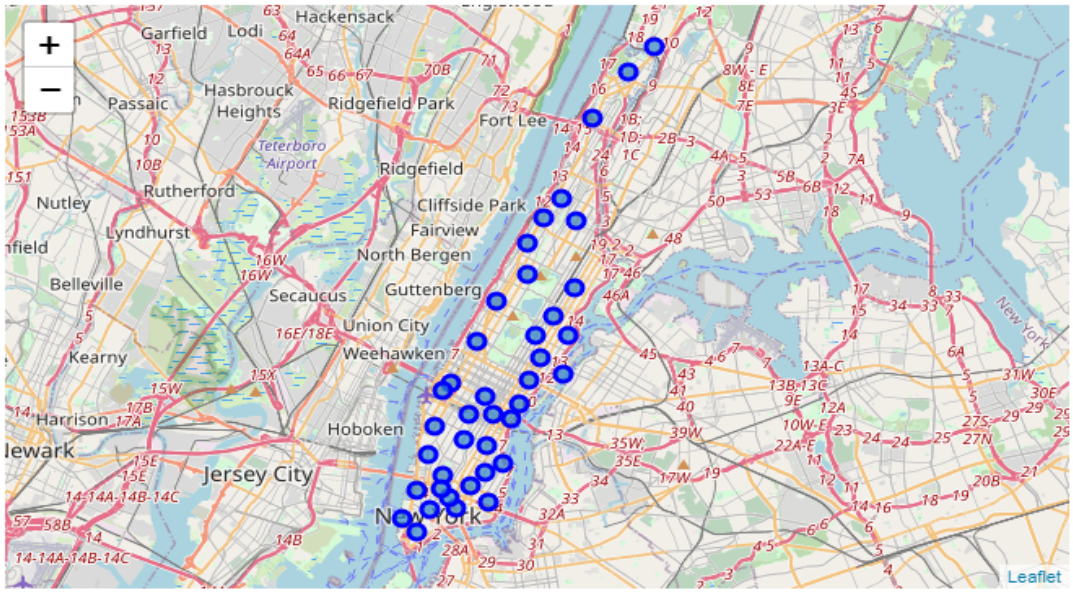
5.1 Data exploratory of Manhattan



Per the above graphic, showing the number of the neighborhood per borough, we can see that has 5 boroughs and Manhattan has the smallest number of neighborhoods (40). Queens is the Borough which has the highest number of neighborhoods (81). A map visualisation of all the 306 neighborhood that count New York City shows the following graphic



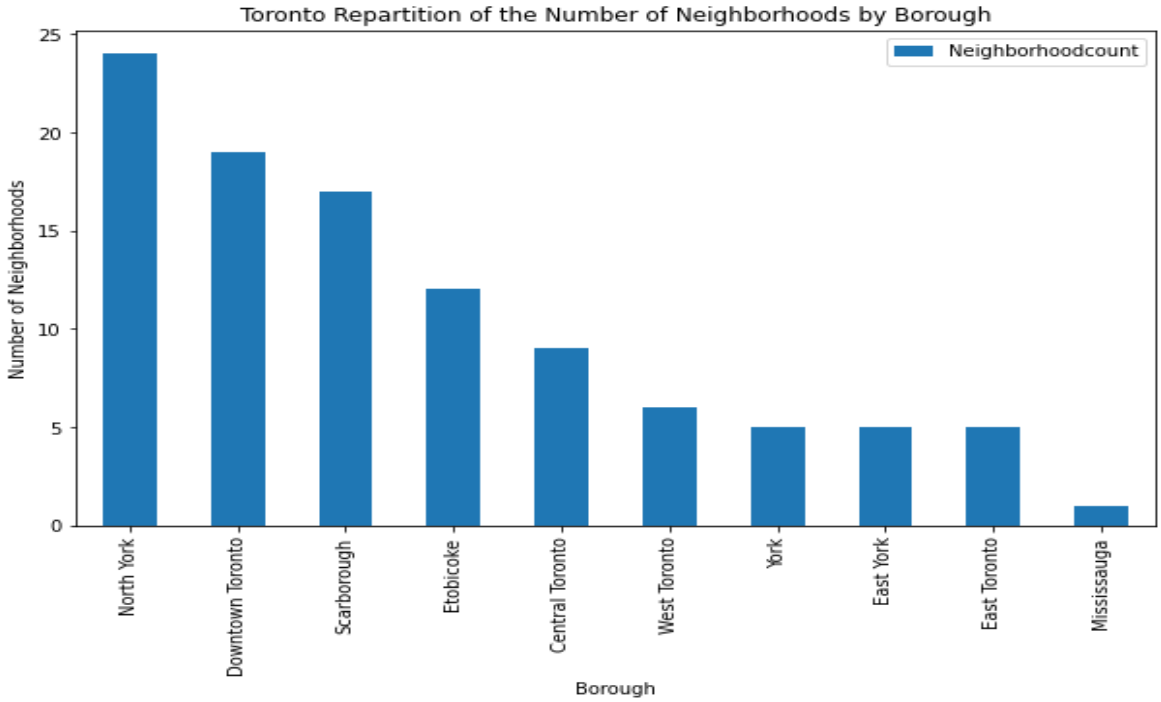
Let's now extract only Downtown New York neighborhood and display the map representation of Manhattan with its 40 Neighborhoods:



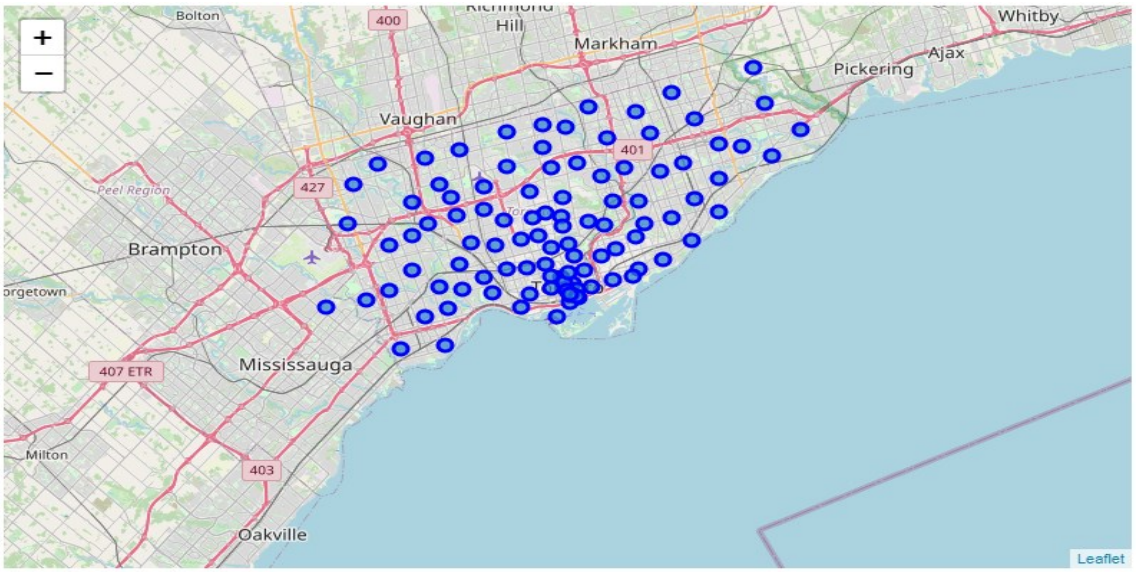
It appears that in Manhattan Borough some neighborhoods are well dispersed while others are more concentrated.

5.2 Data exploratory of Downtown Toronto

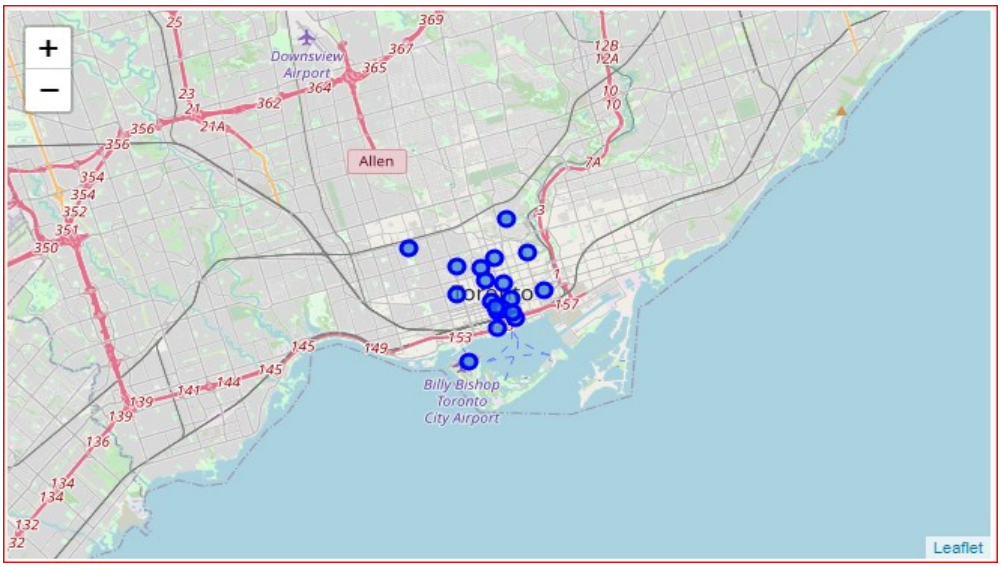
Toronto has 10 Boroughs with 103 neighborhoods and Downtown Toronto is the second most populated in number of neighborhood with 19 neighborhoods. 6 Boroughs have less than 10 neighborhoods and Mississauga has only 1 neighborhood.



A map representation of Toronto city shows the following graph:



Let's now extract only Downtown Toronto neighborhood and display its 19 Neighborhoods in a map representation:



It clearly appears that the neighborhoods at Downtown Toronto are mainly concentrated.
New York has fewer boroughs (5) than Toronto (10). However, Toronto has fewer neighborhoods (103) than New York (306). Downtown Toronto has 19 neighborhoods which are mainly concentrated while Manhattan has more than twice more Neighborhoods 40 and they look more dispersed.

Let's analyse deeper the similarity and dissimilarity between both cities with venues.

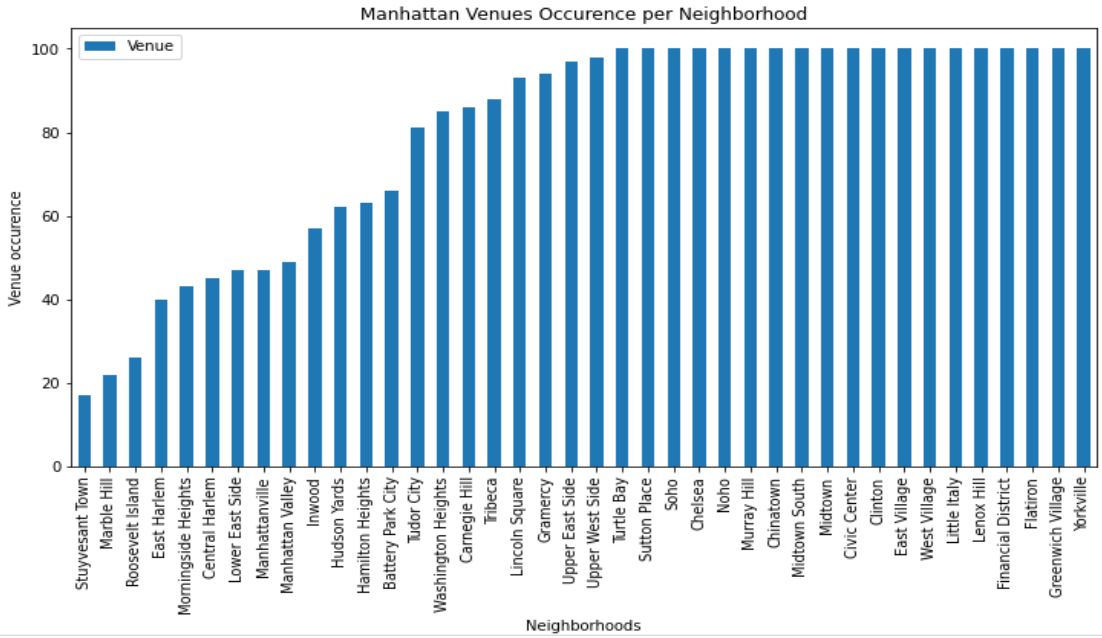
6. Explore Neighborhoods in Downtown Toronto vs Manhattan

To explore more deeply the neighborhoods in Downtown Toronto and Manhattan, I used Foursquare API in both cases. I started by created a Foursquare API account to use my credentials for the API requests.

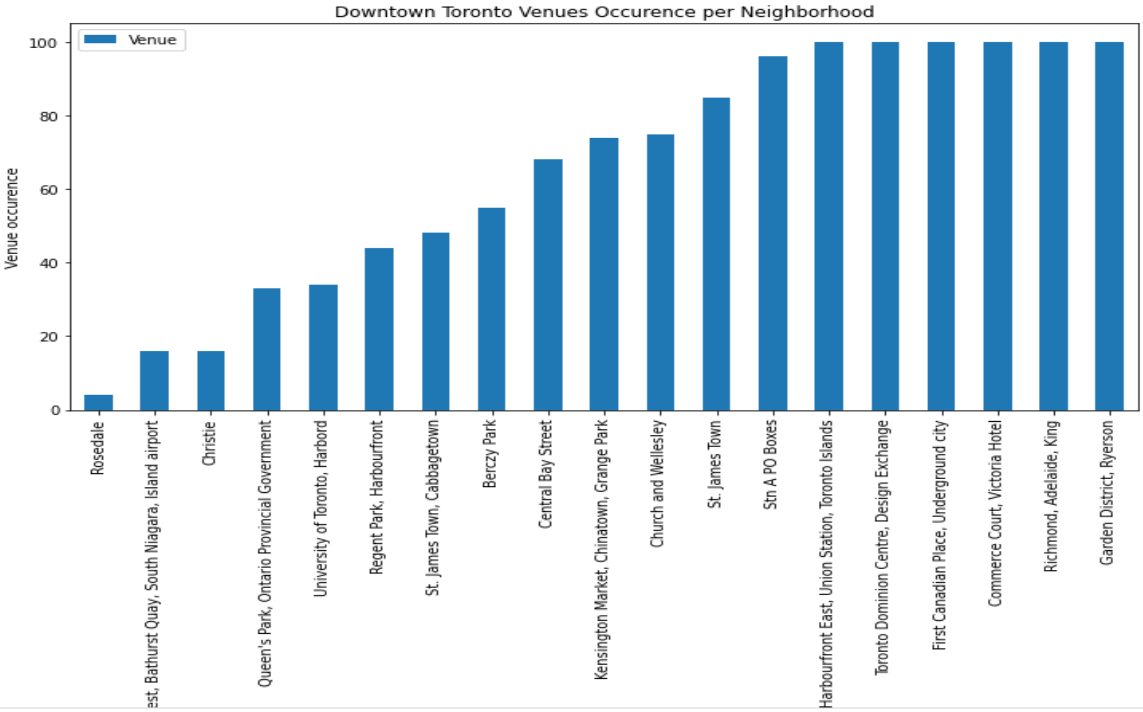
6.1 Explore Neighborhoods in Manhattan

I created a function called getNearbyVenues() to retrieve all the venues nearby Manhattan in their 40 neighborhoods. Here is the representation of the each neighborhood and the number of venues inside. The most populated neighborhoods have 100 venues categories and among them are: Yorkville, West Village, Turtle Bay, Sutton Place, Noho, Murray Hill, Midtown South, Midtown, Little Italy, Lenox Hill, Greenwich Village, Flatiron, Financial District, East Village, Clinton, Civic Center, Chinatown and Chelsea. Stuyvesant Town is the neighborhood less popular in terms of venues with only 17 venues categories, after come Marble Hill with 22 venues categories and Roosevelt Island with 26 venues categories.

It appears that Manhattan is a Borough with a great variety of venues around because more often each neighborhood has more than 60 category venues. Indeed there is 3211 venues at Downtown New York (Manhattan) with 321 unique categories among them are: Yoga Studio, Pizza place, Diner, Coffee Shop, Donut Shop, etc.



6.2 Explore Neighborhoods in Downtown Toronto



By using the same function getNearbyVenues() I sent the Foursquare API request to retrieve all the venues at Downtown Toronto. It results that Downtown Toronto has 1248 venues with 213 unique venues categories. As shown above in the graph, Rosedale has the smallest count of venues categories with 4 while the following neighborhoods have the highest(100) venues categories: Commerce Court,

Victoria Hotel, First Canadian Place, Ryerson, Harbourfront East, Union Station, Toronto Islands, Richmond, Adelaide, King, Toronto Dominion Centre, Design Exchange and Underground city, Garden District.

In terms of the number of venues, Manhattan (3211) has 2,5 times more venues than Downtown Toronto. In Manhattan and Downtown Toronto the maximum number of venue category per neighborhood is 100.

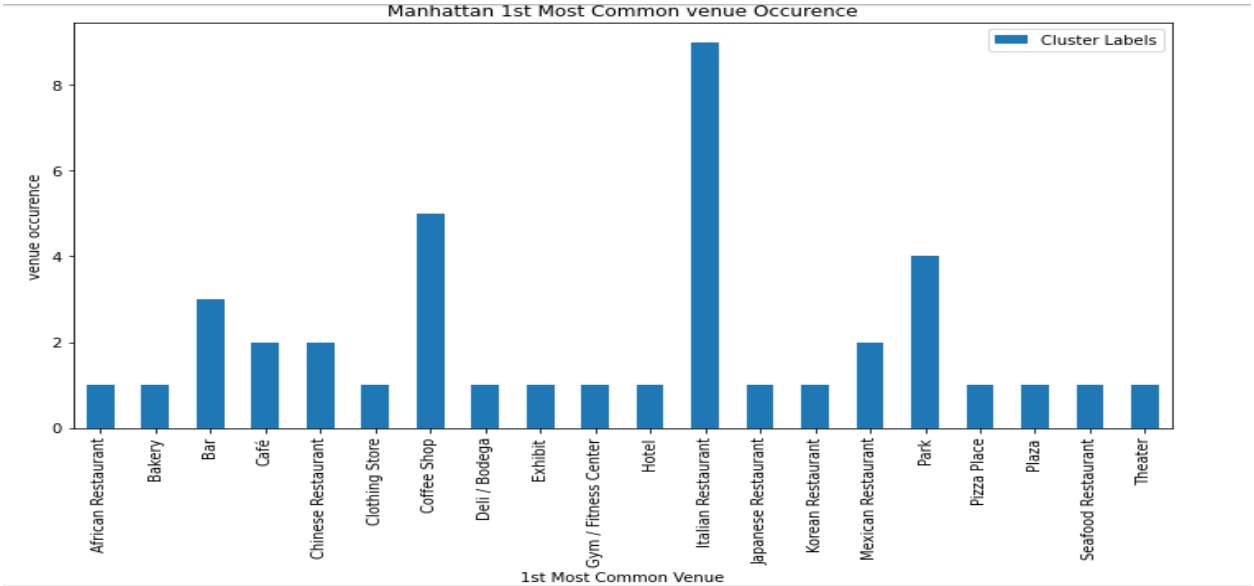
7. Analyse of Each Neighborhood in Manhattan vs Downtown Toronto

The analysis of neighborhood in Manhattan vs Downtown Toronto begin with one hot encoding of the venue category in order to get the frequency of occurrence of each category and classify the most common venue per neighborhood.

7.1 Analyse of Neighborhoods in Manhattan

	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 Com V
0	Battery Park City	Park	Hotel	Gym	Coffee Shop	Memorial Site	Shopping Mall	Plaza	Burger Joint	Gou
1	Carnegie Hill	Coffee Shop	Café	Bookstore	Italian Restaurant	Gym / Fitness Center	Gym	French Restaurant	Yoga Studio	Wine
2	Central Harlem	African Restaurant	Chinese Restaurant	Bar	Seafood Restaurant	American Restaurant	French Restaurant	Cosmetics Shop	Fried Chicken Joint	Carib Resta
3	Chelsea	Coffee Shop	Art Gallery	American Restaurant	Bakery	Italian Restaurant	Ice Cream Shop	Japanese Restaurant	Park	S
4	Chinatown	Chinese Restaurant	Bakery	Cocktail Bar	American Restaurant	Dessert Shop	Optical Shop	Noodle House	Hotpot Restaurant	
5	Civic Center	Coffee Shop	Spa	Gym / Fitness Center	Yoga Studio	French Restaurant	Hotel	Cocktail Bar	Sandwich Place	Resta
6	Clinton	Theater	Gym / Fitness Center	American Restaurant	Sandwich Place	Coffee Shop	Gym	Spa	Hotel	Resta
7	East Harlem	Mexican Restaurant	Bakery	Thai Restaurant	Deli / Bodega	Latin American Restaurant	Sandwich Place	Spa	Liquor Store	Taco
8	East Village	Bar	Mexican Restaurant	Pizza Place	Ice Cream Shop	Wine Bar	Coffee Shop	Speakeasy	Korean Restaurant	Veget / V Resta
9	Financial District	Coffee Shop	Pizza Place	Cocktail Bar	American Restaurant	Bar	Gym	Juice Bar	Park	Steakh
10	Flatiron	Italian Restaurant	Japanese Restaurant	New American Restaurant	Mediterranean Restaurant	Cycle Studio	Gym	Gym / Fitness Center	Furniture / Home Store	Spo G

It results from the analysis that Restaurants, coffee Shop, Café, Bar, Park, Hotel, Pizza place, Plaza, Bakery, Gym/Fitness Center, Exhibit, Deli/Bodega and Theater are the first most common venues at Manhattan. The graph below shows the 1st most common venues repartition in Manhattan city.



As we can clearly see Italian Restaurant venue comes more often than all others venues. After it is Coffee Shop, in third position Park and fourth position with Bar. In one sentence I could said that Manhattan city is a place to mainly eat Italian food, drink Coffee and Beer and take a rest in the Park.

7.2 Analyse Neighborhoods in Downtown Toronto

In Downtown Toronto the first most common venues are: Coffee Shop, Airport Lounge, Grocery Store, Mexican Restaurant, Park and Café. Here are the venues classified according to their frequencies.

	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
0	Berczy Park	Coffee Shop	Restaurant	Seafood Restaurant	Cheese Shop	Farmers Market	Cocktail Bar	Beer Bar	Bakery	Shopping Mall
1	CN Tower, King and Spadina, Railway Lands, Har...	Airport Lounge	Airport Service	Boutique	Boat or Ferry	Plane	Rental Car Location	Coffee Shop	Bar	Harbor / Marina
2	Central Bay Street	Coffee Shop	Café	Sandwich Place	Italian Restaurant	Thai Restaurant	Salad Place	Japanese Restaurant	Bubble Tea Shop	Burger Joint
3	Christie	Grocery Store	Café	Park	Candy Store	Restaurant	Italian Restaurant	Baby Store	Athletics & Sports	Coffee Shop
4	Church and Wellesley	Coffee Shop	Japanese Restaurant	Gay Bar	Sushi Restaurant	Restaurant	Yoga Studio	Men's Store	Mediterranean Restaurant	Hotel
5	Commerce Court, Victoria Hotel	Coffee Shop	Restaurant	Café	Hotel	Gym	American Restaurant	Seafood Restaurant	Japanese Restaurant	Deli / Bodega
6	First Canadian Place, Underground city	Coffee Shop	Café	Gym	Restaurant	Hotel	Japanese Restaurant	Salad Place	Seafood Restaurant	Asian Restaurant
7	Garden District, Ryerson	Coffee Shop	Clothing Store	Café	Bubble Tea Shop	Cosmetics Shop	Japanese Restaurant	Ramen Restaurant	Furniture / Home Store	Diner
8	Harbourfront East, Union Station, Toronto Islands	Coffee Shop	Aquarium	Hotel	Café	Fried Chicken Joint	Restaurant	Brewery	Scenic Lookout	Music Venue
9	Kensington Market, Chinatown, Grange Park	Mexican Restaurant	Coffee Shop	Vegetarian / Vegan Restaurant	Bar	Café	Vietnamese Restaurant	Dessert Shop	Bakery	Park

It result that In Downtown Toronto Coffee Shop comes more often than all others venues. It seems that people are more willing to drink coffee there.

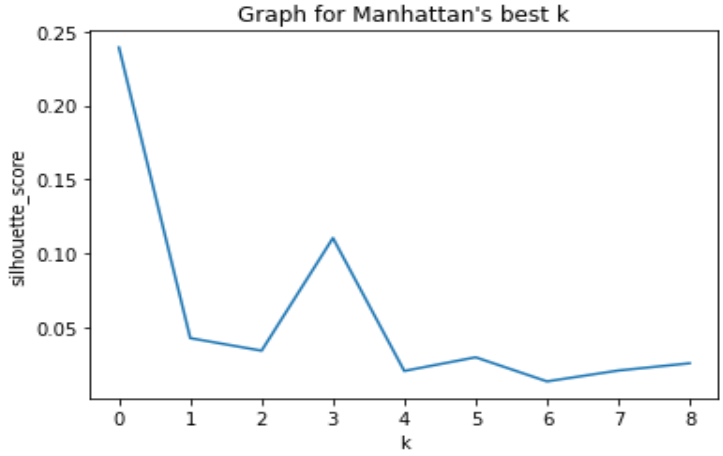
At this point of analysis, It results that at Manhattan, Italian Restaurant is the most frequent venue while Coffee Shop at Downtown Toronto. At the same time Coffee Shop is the second most frequent venue in Manhattan.

8. Cluster Neighborhoods Manhattan vs Downtown Toronto

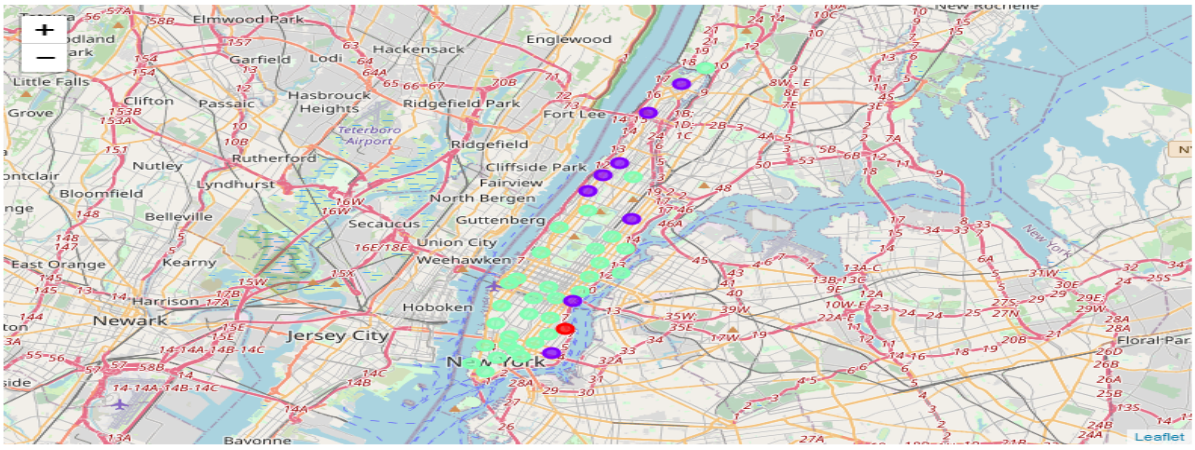
Neighborhood dataset in both cities are not labeled data of category venue in boroughs Manhattan and Downtown Toronto. For that reason I used unsupervised learning K-means algorithm to cluster the boroughs. K-Means algorithm is one of the most common cluster methods of unsupervised learning.

8.1 Cluster Neighborhoods Manhattan

The Sihouette_Score graph for multiple k applied on Downtown New York (Manhattan) shows that there is a clear peak at k = 3. Hence, it is optimal. Manhattan data can then be optimally clustered into 3 clusters as shown below.



Here is a map representation of Manhattan clusters.

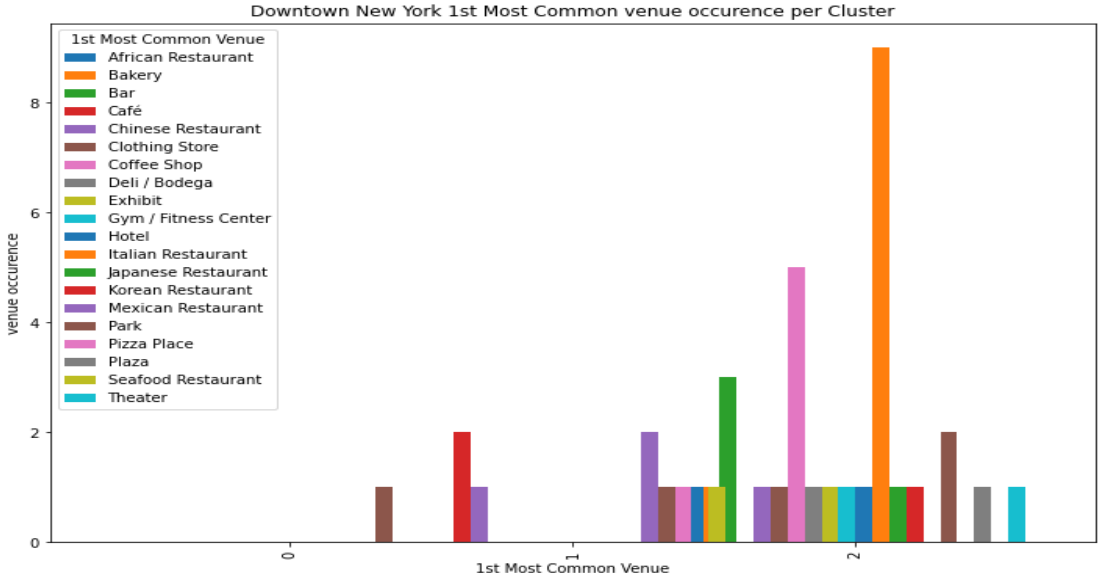


We can see that the physical distance between neighborhoods doesn't necessarily means that they are similar en term of venues. Let's now deeper analyse the clusters.

The "Cluster Labels" Columns in below table shows a clustering repartition of the venues.

	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
0	Manhattan	Marble Hill	40.876551	-73.910660	2	Coffee Shop	Sandwich Place	Discount Store	Gym	Supplement Shop	Donut Shop
1	Manhattan	Chinatown	40.715618	-73.994279	2	Chinese Restaurant	Bakery	Cocktail Bar	American Restaurant	Dessert Shop	Optical Shop
2	Manhattan	Washington Heights	40.851903	-73.936900	1	Café	Bakery	Grocery Store	Mobile Phone Shop	Bank	Sandwich Place
3	Manhattan	Inwood	40.867684	-73.921210	1	Mexican Restaurant	Café	Lounge	Restaurant	Park	Chinese Restaurant
4	Manhattan	Hamilton Heights	40.823604	-73.949688	1	Pizza Place	Coffee Shop	Café	Mexican Restaurant	Cocktail Bar	Indian Restaurant

Per the 1st Most Common Venues frequency graph below, Italian Restaurant is the most common venue in all Manhattan neighborhoods. The second most frequent venues are Coffee Shop and Park and the third most frequency is Bar. The fourth most frequencies venues are Café, Chinese Restaurant, Hotel and Mexican Restaurant. By displaying the 1st most common venues by cluster it results that following graph.



After executing the K-Means algorithm on venues in Manhattan neighborhoods it results the following:

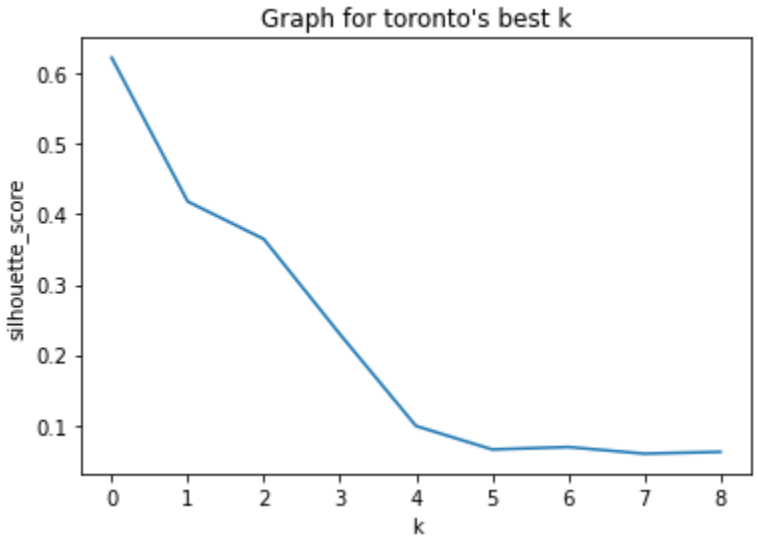
Cluster 0: "Park venues"

Cluster 1: "Restaurants (Mexican Restaurant, Seafood Restaurant and Chinese Restaurant), Café, Pizza Place and Park venues"

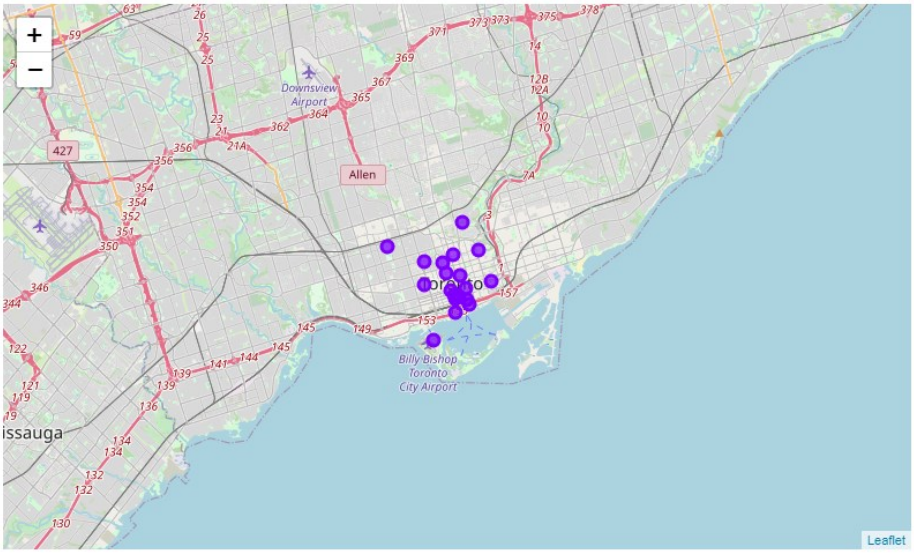
Cluster 2: "Restaurants (Chinese Restaurant, African Restaurant, Italian Restaurant, and Japanese Restaurant), Coffee Shop, Bar, Exhibit, Plaza, Hotel, Theater, Gym/Fitness, Park, Clothing Store, Bakery and Deli / Bodega venues"

8.2 Cluster Neighborhoods Downtown Toronto

The Sihouette_Score graph for multiple k applied on Downtown Toronto data clearly shows below that more Downtown Toronto's Neighborhoods venues data are clustered, more the silhouette_score decrease. The optimal k is then 1.



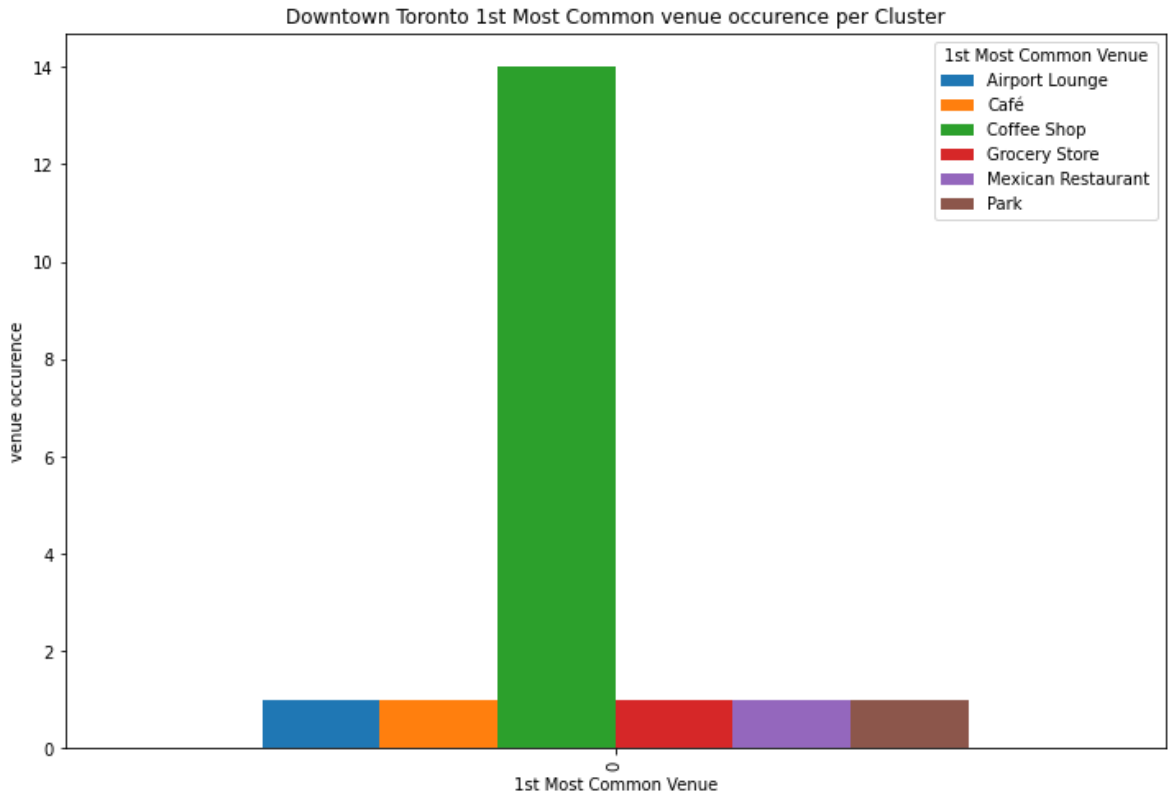
The entire Downtown Toronto neighborhoods data is conserved in one single cluster as shown in the map.



The “Cluster Labels” Columns in below table shows a clustering repartition of the venues.

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Mo Commo Venu
0	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	0	Coffee Shop	Bakery	Pub	Park	Breakfa Sp
1	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	0	Coffee Shop	Gym	Diner	Restaurant	Portugues Restaura
2	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937	0	Coffee Shop	Clothing Store	Café	Bubble Tea Shop	Cosmetic Shc
3	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418	0	Coffee Shop	Café	Cocktail Bar	Restaurant	Gastropu
4	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306	0	Coffee Shop	Restaurant	Seafood Restaurant	Cheese Shop	Farme Mark

Let's now display the graph of the Downtown Toronto 1st most common venues.



The Cluster: "Coffee Shop, café, Grocery Store, Mexican Restaurant & park venues"

Downtown Toronto only cluster is mainly similar to Manhattan Cluster 1: "Restaurants (Mexican Restaurant, Seafood Restaurant and Chinese Restaurant), Café, Pizza Place and Park venues"

9. Discussion

As I mentioned before, Manhattan is the most densely populated of New York City's 5 boroughs while it is the smallest populated in terms of neighborhoods (40). Manhattan is a Borough with a great variety of venues around and more often each neighborhood has more than 60 category venues. There are 3211 venues at Downtown New York (Manhattan) with 321 unique categories. Restaurants, coffee Shop, Café, Bar, Park, Hotel, Pizza place, Plaza, Bakery, Gym/Fitness Center, Exhibit, Deli/Bodega and Theater are the first most common venues at Manhattan. In order the most frequent venues are: Italian Restaurant, Coffee Shop, Park and Bar.

Downtown Toronto is the second most populated Borough among the 10 Toronto's Boroughs in terms of neighborhoods with 19 on the total of 103 neighborhoods. Downtown Toronto neighborhoods are mainly concentrated than Manhattan Neighborhoods which looks more dispersed. Downtown Toronto has 1248 venues with 213 unique venues categories. Each neighborhood in the borough often have more than 60 categories venues same as in Manhattan with a maximum of 100 venues categories in both cities. Coffee Shop comes is the 1st most common venue. The 1st most common venue is the Coffee Shop.

I used the Kmeans algorithm to cluster the neighborhoods in both cities. I used silhouette_score graph with differents k to find the best k in both cities. I set the optimum k value to 3 for Manhattan and 1 for Downtown Toronto. By analysing clusters information on the Manhattan and Downtown Toronto and comparing the results of both boroughs, Downtown Toronto is similar to one Manhattan's cluster.

For futures studies, those results can be carried out to direct entrepreneurs who could be interested to open a new business in those boroughs. The study can also help the managers of both cities to evaluate and well organize the cities.

10. Conclusion

Manhattan and Downtown Toronto are both great and interesting cities to study and compare. They look the same at many levels, mainly in the structure with Borough, Neighborhoods and venues etc. However they are different in their neighborhoods and venues sizes, the concentration of the neighborhoods on the map. They both have a large quantity and variety of unique category venue with. In one sentence, Downtown Toronto looks more homogenous than Manhattan. In Conclusion Manhattan city is a place to mainly eat Italian food, drink Coffee, Beer and take a rest in the Park. Downtown Toronto is a place with mainly Coffee Shop venue.

This study can benefit to the investors and cities managers.

11.References:

- [1] **Manhattan — Wikipedia**
- [2] **Toronto — Wikipedia**
- [3] **Forsquare API**