THE BATTLE OF NEIGHBORHOODS
TORONTO & NEW YORK CITY

Wan Chwee Ying

19th June 2021

1. Introduction

Toronto is the capital city of the Canadian province of Ontario. With a recorded population of 2,731,571 in 2016, it is the most populous city in Canada and the fourth most populous city in North America.

New York City is the most populous city in the United States. With an estimated 2019 population of 8,336,817 distributed over about 302.6 square miles, New York City is also the most densely populated major city in the United States.

Toronto and New York City are the popular vacation and migration destinations for people all around the world. They are diverse and multicultural and offer a wide variety of experiences that is widely sought after. We try to group the neighborhoods of Toronto and New York City respectively and draw insights to what they look like now.

2. Business Problem

The aim is to help people choose their destinations depending on the experiences that the neighborhoods have to offer and what they would want to have. This also helps people make decisions if they are thinking about migrating to Toronto or New York City or even if they want to relocate neighborhoods within the city. Our findings will help stakeholders make informed decisions and address any concerns they have including the different kinds of cuisines, provision stores and what the city has to offer.

3. Data Description

We require geolocation data for both Toronto and New York City. Postal codes in each city serve as a starting point. Using Postal codes, we use can find out the neighborhoods, boroughs, venues, and their most popular venue categories.

3.1 Toronto

To derive our solution, we scrape our data from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

This Wikipedia page has information about all the neighborhoods, we limit it Toronto.

Also, it lacks information about the geographical locations. To solve this problem, we use the Geocoder Python package.

3.2 New York City

To derive our solution, we leverage JSON data available at https://geo.nyu.edu/catalog/nyu_2451_34572

3.3 Foursquare API Data

We will need data about different venues in different neighborhoods of that specific borough. In order to gain that information, we will use "Foursquare" locational information. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos. As such, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API.

After finding the list of neighborhoods, we then connect to the Foursquare API to gather information about venues inside each and every neighborhood. For each neighborhood, we have chosen the radius to be 500 meters.

The data retrieved from Foursquare contained information of venues within a specified distance of the longitude and latitude of the postcodes. The information obtained per venue as follows:

- Neighborhood: Name of the Neighborhood
- Neighborhood Latitude: Latitude of the Neighborhood
- Neighborhood Longitude: Longitude of the Neighborhood
- Venue: Name of the Venue

- Venue Latitude: Latitude of Venue
- Venue Longitude: Longitude of Venue
- Venue Category: Category of Venue

Based on all the information collected for both Toronto and New York City, we have sufficient data to build our model. We cluster the neighborhoods together based on similar venue categories. We then present our observations and findings. Using this data, our stakeholders can take the necessary decision.

4. Methodology

We will be creating our model with the help of Python so we start off by importing all the required packages.

```
import pandas as pd
import numpy as np
import requests
import json
from bs4 import BeautifulSoup

import matplotlib.cm as cm
import matplotlib.colors as colors

from geopy.geocoders import Nominatim
!conda install -c conda-forge folium=0.5.0
import folium

from sklearn.cluster import KMeans
```

Package breakdown:

- *pandas*: To collect and manipulate data in JSON and HTML and then data analysis.
- *requests*: Handle http requests.
- *matplotlib*: Detailing the generated maps.
- *folium*: Generating maps of Toronto and New York City.
- *sklearn*: To import KMeans which is the machine learning model that we are using.

The approach taken here is to explore each of the cities individually, plot the map to show the neighborhoods being considered and then build our model by clustering all of the similar neighborhoods together and finally plot the new map with the clustered neighborhoods. We draw insights and then compare and discuss our findings.

4.1 Data Collection

To get the neighborhoods in Toronto, we start by scraping the list of postal codes of Canada Wikipedia page using the following code:

```
url = "https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M"
html_data  = requests.get(url).text
soup = BeautifulSoup(html_data,"html.parser")
table_contents = pd.DataFrame(columns=['Postal Code', 'Borough', 'Neighborhood'])

table_contents=[]
table=soup.find('table')
for row in table.findAll('td'):
    cell = {}
    if row.span.text=='Not assigned':  #To skip if the borough name is 'Not Assigned'
        pass
    else:
        cell['PostalCode'] = row.p.text[:3]
        cell['Borough'] = (row.span.text).split('(')[0]
        cell['Neighborhood'] = (((((row.span.text).split('(')[1]).strip(')')).replace(' /',',')).replace(')',' ')).strip(' ')
        table_contents.append(cell)
df=pd.DataFrame(table_contents)
```

The data looks like this:

	PostalCode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Queen's Park	Ontario Provincial Government
...	...	...	...
98	M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North
99	M4Y	Downtown Toronto	Church and Wellesley
100	M7Y	East TorontoBusiness reply mail Processing Cen...	Enclave of M4L
101	M8Y	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu...
102	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...

103 rows × 3 columns

To collect data for New York City, we load the JSON data using the following codes:

```
!wget -q -O 'newyork_data.json' https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json
with open('newyork_data.json') as json_data:
    newyork_data = json.load(json_data)
```


```
{'type': 'FeatureCollection',
 'totalFeatures': 306,
 'features': [{'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]}},
```

4.2 Data Pre-processing

For Toronto, we replace some of the *borough* using:

```
df['Borough']=df['Borough'].replace({'Downtown TorontoStn A PO Boxes25 The Esplanade':'Downtown Toronto Stn A',
                                     'East TorontoBusiness reply mail Processing Centre969 Eastern':'East Toronto Business',
                                     'EtobicokeNorthwest':'Etobicoke Northwest','East YorkEast Toronto':'East York/East Toronto',
                                     'MississaugaCanada Post Gateway Processing Centre':'Mississauga'})
```

The data looks like:

	PostalCode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Queen's Park	Ontario Provincial Government
...	...	...	...
98	M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North
99	M4Y	Downtown Toronto	Church and Wellesley
100	M7Y	East Toronto Business	Enclave of M4L
101	M8Y	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu...
102	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...

103 rows × 3 columns

4.3 Feature Selection

For both of our datasets, we need only the borough, neighborhood, postal codes and geolocations (latitude and longitude). So, we select the columns that we need by:

```
# define the dataframe columns
column_names = ['Borough', 'Neighborhood', 'Latitude', 'Longitude']

# instantiate the dataframe
neighborhoods = pd.DataFrame(columns=column_names)
```

4.4 Feature Engineering

Looking over our Toronto dataset, we can see that we do not have the geolocation data. We need to extrapolate the missing data for our neighborhoods.

```
data = pd.read_csv("https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs_v1/Geospatial_Coordinates.csv")
combined_data = df.join(data.set_index('Postal Code'), on='PostalCode', how='inner')
combined_data
```

	PostalCode	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	Queen's Park	Ontario Provincial Government	43.662301	-79.389494
...	...	...	...	...	...
98	M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North	43.653654	-79.506944
99	M4Y	Downtown Toronto	Church and Wellesley	43.665860	-79.383160
100	M7Y	East Toronto Business	Enclave of M4L	43.662744	-79.321558
101	M8Y	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu...	43.636258	-79.498509
102	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999

103 rows × 5 columns

As for our New York City dataset, we do not need to get the geo coordinates using an external data source since we already have it stored in the data.

We just need to transform the data into a pandas dataframe:

```
# instantiate the dataframe
neighborhoods = pd.DataFrame(columns=column_names)

for data in neighborhoods_data:
    borough = neighborhood_name = data['properties']['borough']
    neighborhood_name = data['properties']['name']

    neighborhood_latlon = data['geometry']['coordinates']
    neighborhood_lat = neighborhood_latlon[1]
    neighborhood_lon = neighborhood_latlon[0]

    neighborhoods = neighborhoods.append({'Borough': borough,
                                          'Neighborhood': neighborhood_name,
                                          'Latitude': neighborhood_lat,
                                          'Longitude': neighborhood_lon}, ignore_index=True)

neighborhoods.head()
```

The first 5 rows of the data look like:

	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

For illustration purposes, let us simplify and cluster only the neighborhoods in Manhattan. So, let us slice the original dataframe and create a new dataframe of the Manhattan data.

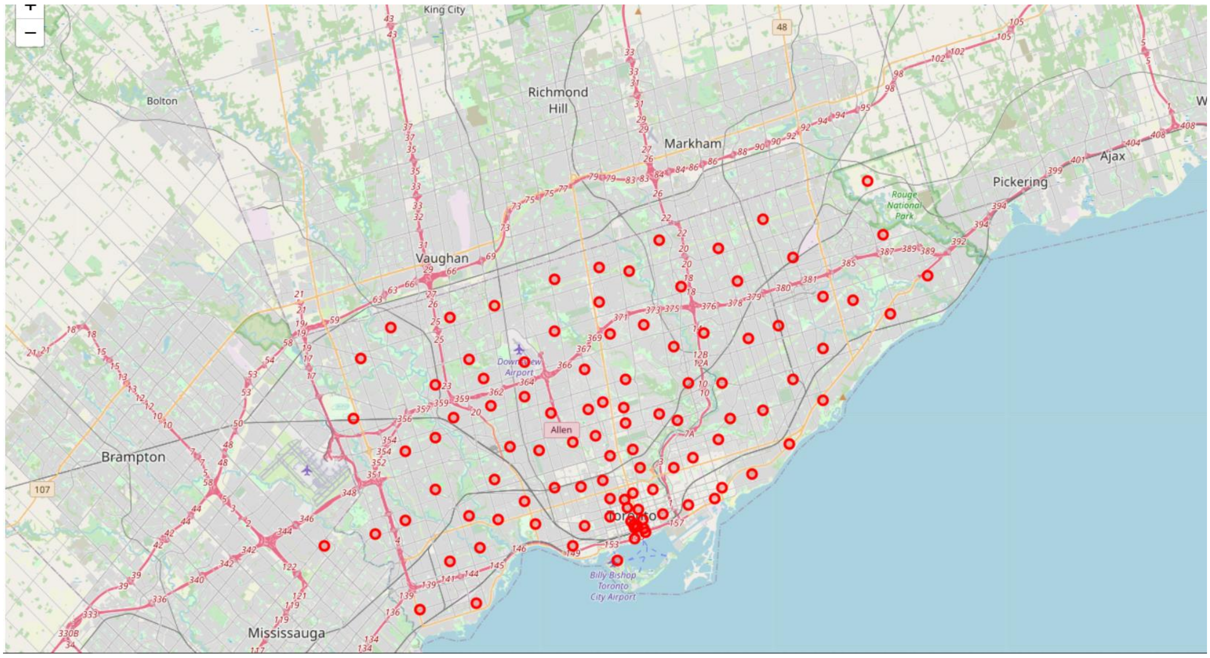
```
manhattan_data = neighborhoods[neighborhoods['Borough'] == 'Manhattan'].reset_index(drop=True)
manhattan_data.head()
```

	Borough	Neighborhood	Latitude	Longitude
0	Manhattan	Marble Hill	40.876551	-73.910660
1	Manhattan	Chinatown	40.715618	-73.994279
2	Manhattan	Washington Heights	40.851903	-73.936900
3	Manhattan	Inwood	40.867684	-73.921210
4	Manhattan	Hamilton Heights	40.823604	-73.949688

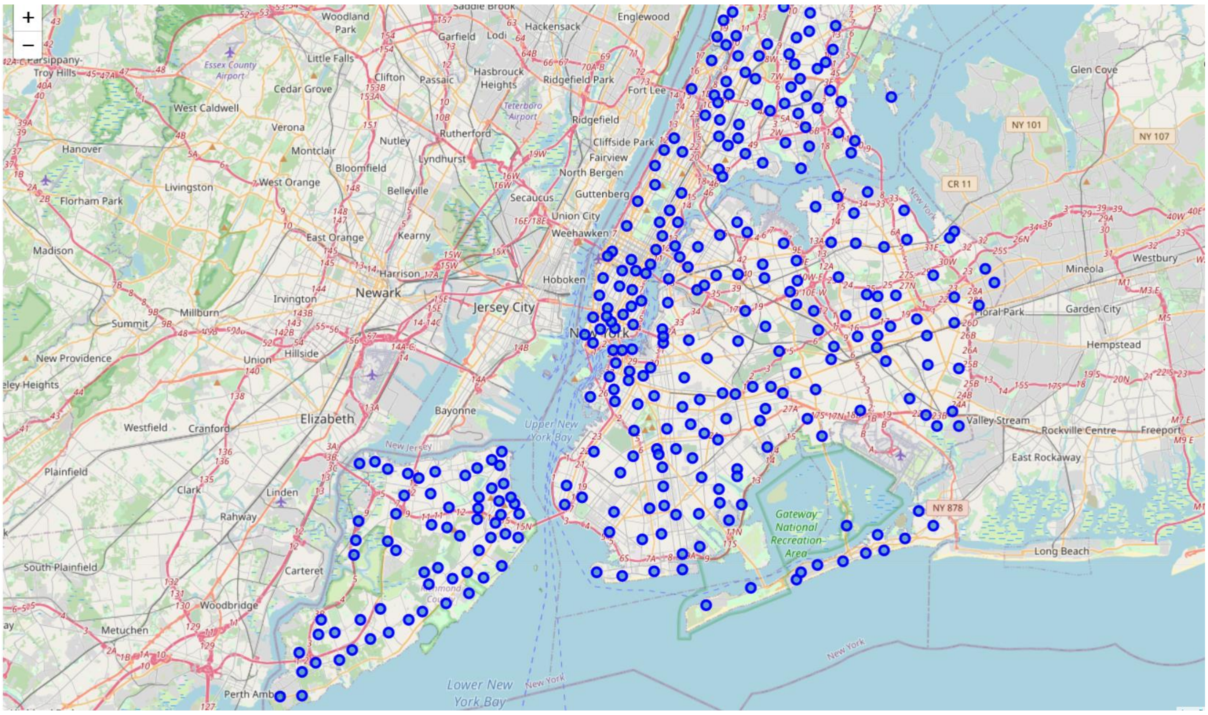
4.5 Visualizing the Neighborhoods of Toronto and New York City

With our datasets ready, using the Folium package, we can visualize the maps of Toronto, New York City and Manhattan with the neighborhoods that we collected.

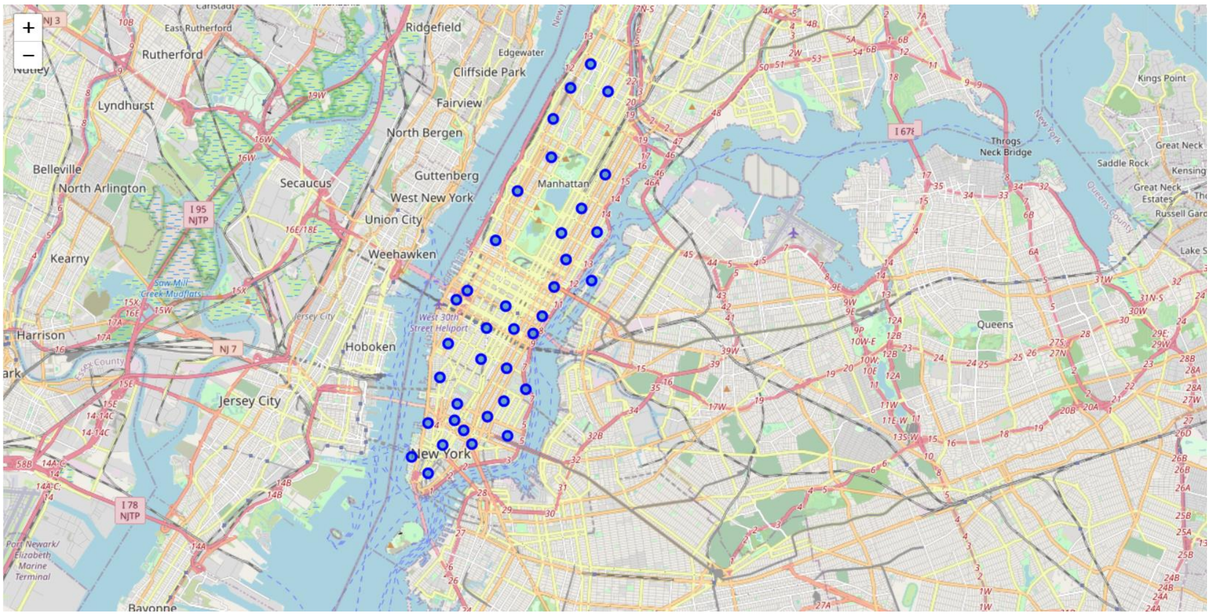
- Neighborhood map of Toronto



- Neighborhood map of New York City



- Neighborhood map of Manhattan



Now that we have visualized the neighborhoods, we need to find out what each neighborhood is like and what are the common venue and venue categories within a 500m radius.

This is where Foursquare plays its part. With the help of Foursquare we define a function which collects information pertaining to each neighborhood including that of the name of the neighborhood, geo-coordinates, venue and venue categories.

```
def getNearbyVenues(names, latitudes, longitudes, radius=500):

    venues_list=[]
    for name, lat, lng in zip(names, latitudes, longitudes):
        print(name)

        # create the API request URL
        url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            radius,
            LIMIT)

        # make the GET request
        results = requests.get(url).json()["response"]['groups'][0]['items']

        # return only relevant information for each nearby venue
        venues_list.append([(
            name,
            lat,
            lng,
            v['venue']['name'],
            v['venue']['location']['lat'],
            v['venue']['location']['lng'],
            v['venue']['categories'][0]['name']) for v in results])

    nearby_venues = pd.DataFrame([item for venue_list in venues_list for item in venue_list])
    nearby_venues.columns = ['Neighborhood',
                  'Neighborhood Latitude',
                  'Neighborhood Longitude',
                  'Venue',
                  'Venue Latitude',
                  'Venue Longitude',
                  'Venue Category']

    return(nearby_venues)
```

Resulting data of Toronto looks like:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	KFC	43.754387	-79.333021	Fast Food Restaurant
1	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
2	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
3	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
4	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant

Resulting data of Manhattan looks like:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
1	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	Rite Aid	40.875467	-73.908906	Pharmacy
4	Marble Hill	40.876551	-73.91066	Subway	40.874667	-73.909586	Sandwich Place

4.6 One Hot Encoding

Since we are trying to find out what are the different kinds of venue categories present in each neighborhood and then calculate the top 10 common venues to base our similarity on, we use the One Hot Encoding to work with our categorical datatype of the venue categories. This helps to convert the categorical data into numeric data.

We will not be using label encoding in this situation since label encoding might cause our machine learning model to have a bias or a sort of ranking which we are trying to avoid by using One Hot Encoding.

We perform One Hot Encoding and group the neighborhoods and calculate the mean of the frequency of occurrence of each category.

```
# one hot encoding
Toronto_onehot = pd.get_dummies(venues_in_toronto[['Venue Category']], prefix="", prefix_sep="")

# add neighborhood column back to dataframe
Toronto_onehot['Neighborhood'] = venues_in_toronto['Neighborhood']

# move neighborhood column to the first column
fixed_columns = [Toronto_onehot.columns[-1]] + list(Toronto_onehot.columns[:-1])
Toronto_onehot = Toronto_onehot[fixed_columns]

Toronto_grouped = Toronto_onehot.groupby('Neighborhood').mean().reset_index()
```

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Argentinian Restaurant	Art Gallery	Art Museum	...	Vietnamese Restaurant	Volleyball Court	Waterfront	Weight Loss Center	Whisky Bar
0	Battery Park City	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000
1	Carnegie Hill	0.000000	0.000000	0.00	0.000000	0.014085	0.000000	0.000000	0.000000	0.028169	...	0.000000	0.000000	0.000000	0.028169	0.000000
2	Central Harlem	0.000000	0.000000	0.00	0.081081	0.054054	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000
3	Chelsea	0.000000	0.000000	0.00	0.000000	0.030000	0.000000	0.000000	0.070000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000
4	Chinatown	0.000000	0.000000	0.00	0.000000	0.030000	0.000000	0.000000	0.010000	0.000000	...	0.010000	0.000000	0.000000	0.000000	0.000000
5	Civic Center	0.000000	0.000000	0.00	0.000000	0.011628	0.011628	0.000000	0.011628	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000

4.7 Top Venues in the Neighborhoods

In our next step, we need to rank and label the top venue categories in our neighborhood.

Let us define a function to get the top venue categories in the neighborhood.

```
def return_most_common_venues(row, num_top_venues):
    row_categories = row.iloc[1:]
    row_categories_sorted = row_categories.sort_values(ascending=False)

    return row_categories_sorted.index.values[0:num_top_venues]
```

There are many categories, we will consider top 10 categories to avoid data skew.

```
num_top_venues = 10

indicators = ['st', 'nd', 'rd']

# create columns according to number of top venues
columns = ['Neighborhood']
for ind in np.arange(num_top_venues):
    try:
        columns.append('{}{} Most Common Venue'.format(ind+1, indicators[ind]))
    except:
        columns.append('{}th Most Common Venue'.format(ind+1))

# create a new dataframe
neighborhoods_venues_sorted = pd.DataFrame(columns=columns)
neighborhoods_venues_sorted['Neighborhood'] = Toronto_grouped['Neighborhood']

for ind in np.arange(Toronto_grouped.shape[0]):
    neighborhoods_venues_sorted.iloc[ind, 1:] = return_most_common_venues(Toronto_grouped.iloc[ind, :], num_top_venues)

neighborhoods_venues_sorted.head()
```

Neighborhood of Toronto:

	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
0	Agincourt	Breakfast Spot	Lounge	Skating Rink	Clothing Store	Latin American Restaurant	Middle Eastern Restaurant	Molecular Gastronomy Restaurant	Modern European Restaurant	Mobile Phone Shop	Miscellaneous Shop
1	Alderwood, Long Branch	Pizza Place	Pub	Sandwich Place	Playground	Dance Studio	Coffee Shop	Skating Rink	Molecular Gastronomy Restaurant	Modern European Restaurant	Mobile Phone Shop
2	Bathurst Manor, Wilson Heights, Downsview North	Bank	Coffee Shop	Shopping Mall	Sushi Restaurant	Diner	Gas Station	Sandwich Place	Frozen Yogurt Shop	Ice Cream Shop	Bridal Shop
3	Bayview Village	Japanese Restaurant	Café	Bank	Chinese Restaurant	Museum	Motel	Moroccan Restaurant	Monument / Landmark	Molecular Gastronomy Restaurant	Modern European Restaurant
4	Bedford Park, Lawrence Manor East	Restaurant	Italian Restaurant	Coffee Shop	Sandwich Place	Indian Restaurant	Sushi Restaurant	Café	Liquor Store	Butcher	Thai Restaurant

Neighborhood of 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 Most Common Venue	10th Most Common Venue
0	Battery Park City	Park	Hotel	Coffee Shop	Gym	Memorial Site	Food Court	Plaza	Boat or Ferry	Playground	Pizza Place
1	Carnegie Hill	Coffee Shop	Wine Shop	Pizza Place	Cosmetics Shop	Café	Gym	Yoga Studio	Pub	Gym / Fitness Center	Sushi Restaurant
2	Central Harlem	African Restaurant	Sandwich Place	Gym / Fitness Center	French Restaurant	Bar	Seafood Restaurant	Fried Chicken Joint	American Restaurant	Bank	Library
3	Chelsea	Art Gallery	Coffee Shop	Bakery	French Restaurant	American Restaurant	Wine Shop	Italian Restaurant	Ice Cream Shop	Tapas Restaurant	Hotel
4	Chinatown	Bakery	Chinese Restaurant	Cocktail Bar	American Restaurant	Spa	Dessert Shop	Salon / Barbershop	Ice Cream Shop	Optical Shop	Noodle House

4.8 Model Building – KMeans

We will use KMeans Clustering Machine Learning algorithm to cluster similar neighborhoods together. We will be going with the number of clusters as 5.

```
# set number of clusters
kclusters = 5

manhattan_grouped_clustering = manhattan_grouped.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(manhattan_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

Our model has labelled each of the neighborhoods, we add the label into our dataset. We then merge our datasets to add latitude and longitude for each of the neighborhood to prepare for visualization.

```
# add clustering labels
neighborhoods_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)

manhattan_merged = manhattan_data

# merge manhattan_grouped with manhattan_data to add latitude/longitude for each neighborhood
manhattan_merged = manhattan_merged.join(neighborhoods_venues_sorted.set_index('Neighborhood'), on='Neighborhood')

manhattan_merged.head()
```

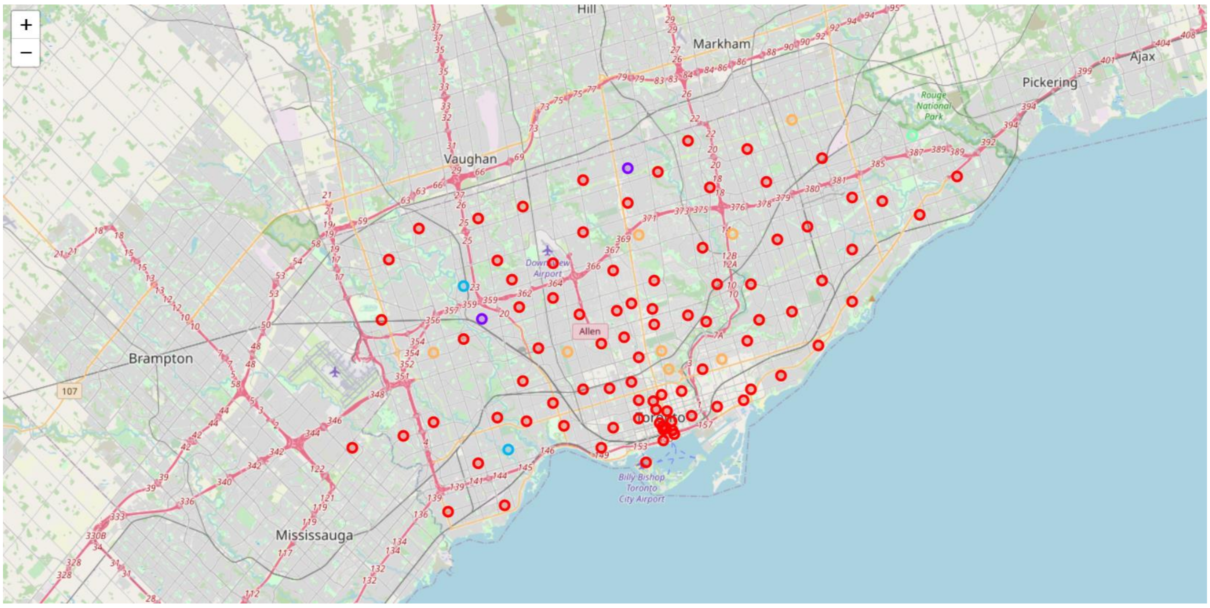
	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	M3A	North York	Parkwoods	43.753259	-79.329656	4.0	Fast Food Restaurant	Park	Food & Drink Shop	Martial Arts School	Malay Restaurant	Medical Center	Mediterranean Restaurant	Men's Store	Moroccan Restaurant	Metro Station
1	M4A	North York	Victoria Village	43.725882	-79.315572	0.0	Nail Salon	Grocery Store	Coffee Shop	Hockey Arena	Portuguese Restaurant	Yoga Studio	Miscellaneous Shop	Monument / Landmark	Molecular Gastronomy Restaurant	Modern European Restaurant
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	0.0	Coffee Shop	Park	Pub	Bakery	Café	Restaurant	Farmers Market	Sushi Restaurant	Beer Store	Performing Arts Venue
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	0.0	Clothing Store	Furniture / Home Store	Miscellaneous Shop	Accessories Store	Boutique	Vietnamese Restaurant	Coffee Shop	Moroccan Restaurant	Monument / Landmark	Molecular Gastronomy Restaurant
4	M7A	Queen's Park	Ontario Provincial Government	43.662301	-79.389494	0.0	Coffee Shop	Sushi Restaurant	Burrito Place	Yoga Studio	Salad Place	Mexican Restaurant	Burger Joint	Café	Sandwich Place	Restaurant

	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	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Manhattan	Marble Hill	40.876551	-73.910660	4	Sandwich Place	Department Store	Supplement Shop	Storage Facility	Steakhouse	Clothing Store	Coffee Shop	Seafood Restaurant	Deli / Bodega	Yoga Studio
1	Manhattan	Chinatown	40.715618	-73.994279	1	Bakery	Chinese Restaurant	Cocktail Bar	American Restaurant	Spa	Dessert Shop	Salon / Barbershop	Ice Cream Shop	Optical Shop	Noodle House
2	Manhattan	Washington Heights	40.851903	-73.936900	2	Café	Pizza Place	Bakery	Chinese Restaurant	Mobile Phone Shop	Grocery Store	Bank	Park	New American Restaurant	Supplement Shop
3	Manhattan	Inwood	40.867684	-73.921210	2	Mexican Restaurant	Café	Lounge	Bank	Pizza Place	Park	Deli / Bodega	Bakery	Wine Bar	Caribbean Restaurant
4	Manhattan	Hamilton Heights	40.823604	-73.949688	0	Pizza Place	Café	Sandwich Place	Mexican Restaurant	Bakery	Deli / Bodega	Coffee Shop	Cocktail Bar	Caribbean Restaurant	Indian Restaurant

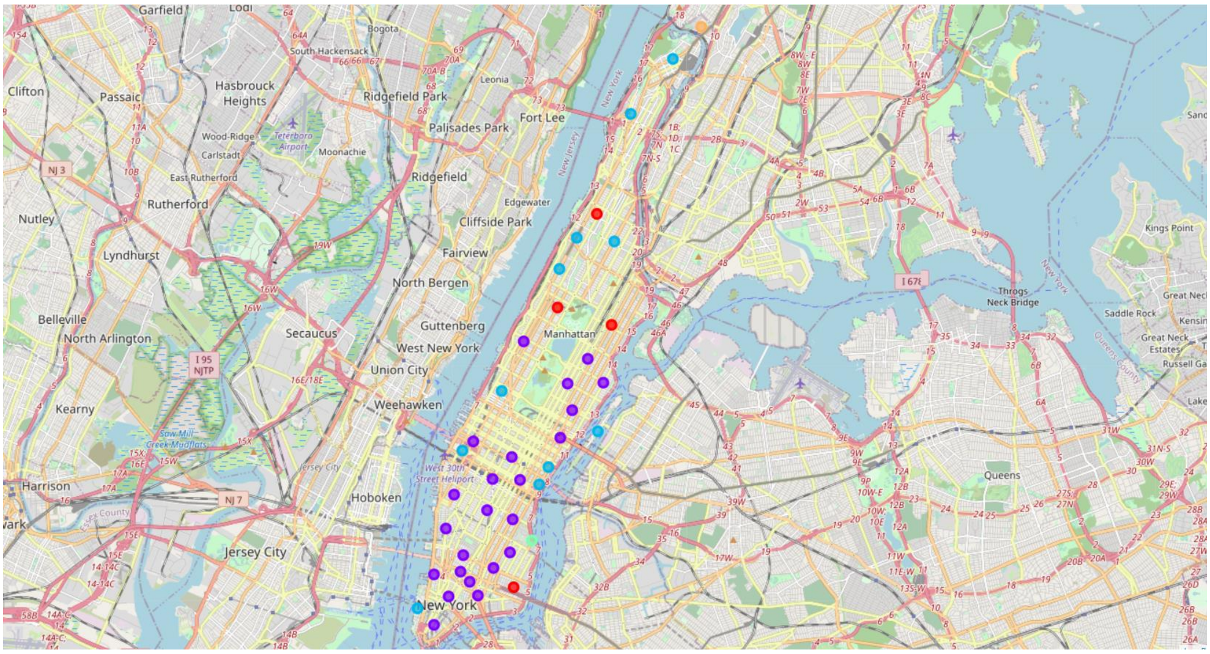
4.9 Visualizing the Clustered Neighborhoods

Our data is processed, missing data is collected and compiled. The Model is built and we can visualize the clustered neighborhoods on the map by using Folium package.

- Map of clustered neighborhoods of Toronto



- Map of clustered neighborhoods of Manhattan



4.10 Examining our Clusters

We could examine our clusters by expanding on our code using the Cluster Labels column:

Cluster 1

```
Toronto_merged_nonan.loc[Toronto_merged_nonan['Cluster Labels'] == 0, Toronto_merged_nonan.columns[[1] + list(range(5, Toronto_merged_nonan.shape[1]))]]
```

Cluster 2

```
Toronto_merged_nonan.loc[Toronto_merged_nonan['Cluster Labels'] == 1, Toronto_merged_nonan.columns[[1] + list(range(5, Toronto_merged_nonan.shape[1]))]]
```

Cluster 3

```
Toronto_merged_nonan.loc[Toronto_merged_nonan['Cluster Labels'] == 2, Toronto_merged_nonan.columns[[1] + list(range(5, Toronto_merged_nonan.shape[1]))]]
```

Cluster 4

```
Toronto_merged_nonan.loc[Toronto_merged_nonan['Cluster Labels'] == 3, Toronto_merged_nonan.columns[[1] + list(range(5, Toronto_merged_nonan.shape[1]))]]
```

Cluster 5

```
Toronto_merged_nonan.loc[Toronto_merged_nonan['Cluster Labels'] == 4, Toronto_merged_nonan.columns[[1] + list(range(5, Toronto_merged_nonan.shape[1]))]]
```

For example, Cluster 1 of the neighborhood of 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 Most Common Venue	10th Most Common Venue
4	Hamilton Heights	Pizza Place	Café	Sandwich Place	Mexican Restaurant	Bakery	Deli / Bodega	Coffee Shop	Cocktail Bar	Caribbean Restaurant	Indian Restaurant
7	East Harlem	Mexican Restaurant	Bakery	Pizza Place	Steakhouse	Bank	Sandwich Place	Gas Station	Spanish Restaurant	Spa	New American Restaurant
20	Lower East Side	Café	Art Gallery	Bakery	Chinese Restaurant	Pizza Place	Sandwich Place	Diner	Pet Café	Tailor Shop	Filipino Restaurant
25	Manhattan Valley	Yoga Studio	Mexican Restaurant	Pizza Place	Coffee Shop	Indian Restaurant	Café	Hostel	Bubble Tea Shop	Playground	Liquor Store

5. Results and Discussion

The neighborhoods of Toronto are very multicultural. There are a lot of different cuisines including Middle Eastern, Moroccan, Latin American, Italian, Japanese, Thai and Chinese. Toronto seems to take a step further in this direction by having a lot of restaurants, pub, café, coffee shops, sandwich place, breakfast spots, ice cream shop, frozen yogurt shop etc. For leisure, the neighborhoods are set up to have gym, skating rink, lounge, yoga studio, dance studio, playground and monument/landmark. Overall, the city of Toronto offers a multicultural, diverse and certainly an entertaining experience.

New York City has a wide variety of cuisines and eateries including French, African, Japanese, Korean, Chinese etc. There are a lot of hangout spots including café, coffee shop, wine shop and cocktail bars. It has a lot of shopping options too with that of the shopping mall, cosmetics shop, bakery, optical shop etc. For leisure and sight-seeing, there are a lot of parks, playground, yoga studio, gym/fitness center, spa, memorial sites and art gallery. Overall, New York City seems like the relaxing vacation spot with a mix of parks, historic spots and a wide variety of cuisines to try out.

6. Conclusion

The purpose of this project was to explore the cities of Toronto and New York City and see how attractive it is to potential tourists and migrants. We explored both the cities based on their postal codes and then extrapolated the common venues present in each of the neighborhoods. Finally, we conclude with clustering similar neighborhoods together.

We could see that each of the neighborhoods in both the cities have a wide variety of experiences to offer which is unique in its own way. The cultural diversity is the evident which also gives the feeling of a sense of inclusion.

Both Toronto and New York City seem to offer a vacation stay or a romantic gateway with a lot of places to explore, beautiful landscapes and a wide variety of culture. In a nutshell, it depends on the stakeholders to decide which experience they would prefer more, and which would more to their liking.