2. Data acquisition and pre-processing

2.1 Data source

There were four types of data platforms adapted by the project. The borough and neighborhood information of New York and Toronto was obtained respectively from NYU Spatial Data Repository (https://geo.nyu.edu/catalog/nyu_2451_34572) and the Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M). Geographical coordinates datasets were downloaded through available sources by geopy library or link (http://cocl.us/Geospatial_data). As the project was supposed to analyze the biggest cluster of neighborhoods by its venues, for venues of interest, Foursquare API was also utilized.

2.2 Data pre-processing

2.2.1 Manhattan dataset

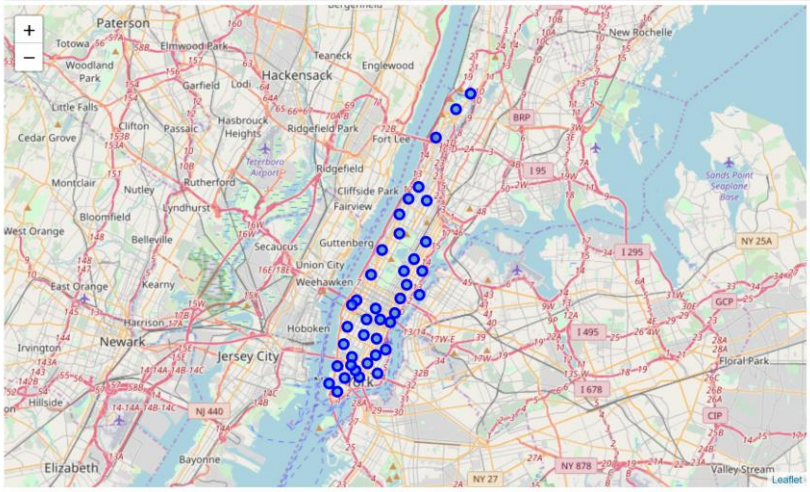
Manhattan dataset was extracted from the New York dataset. After downloading the New York json file, all the relevant data that in the features key was transformed into a pandas dataframe with the instantiation of column names. Geopy library was then used to get latitude and longitude values of New York City which would be put into the same dataframe. The completed version dataframe was created with attributes as following: borough, neighborhood, latitude and longitude (see Table 1).

Next, Manhattan data in the completed dataframe was selected and put into a new dataframe. Geographical coordinates were acquired as the New York dataset so that a specified map of Manhattan can be generated (see Fig 1).

Table 1 Manhattan dataset

	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

Figure 1 Manhattan map



2.2.2 Toronto area dataset

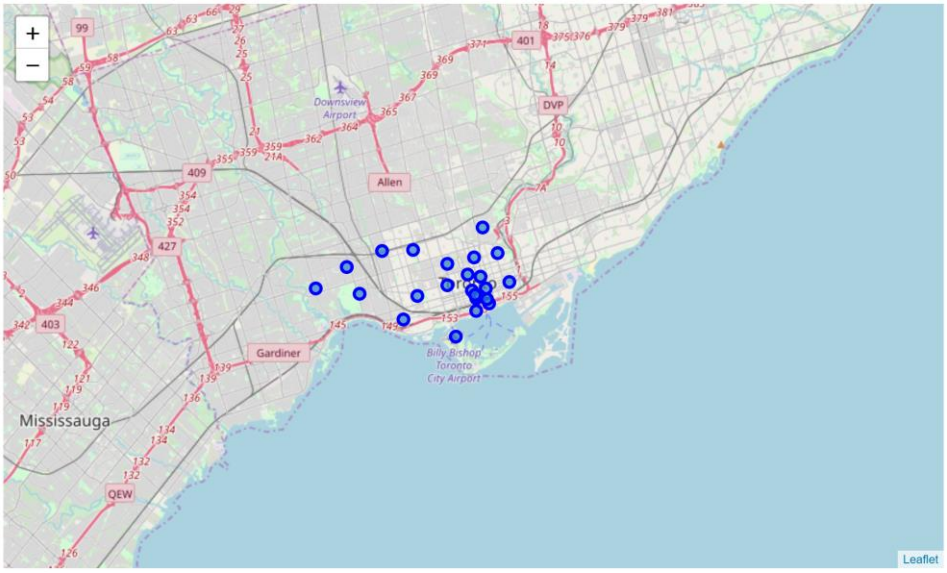
After scraping the table from Wikipedia page, all the relevant attributes including postal code, borough, and neighborhood were cleaned and put into a dataframe. Columns then were renamed by attributes name. It is obvious to see that the original dataset contains many missing values which are shown as “Not assigned”. Thus, the next step was to drop all the missing values and reset the dataframe index. Specially, if there was missing value in the neighborhood column, it would be replaced by its corresponding borough name.

Second, neighborhoods with the same postal code and borough were grouped together, remaining unique values of postal codes and boroughs. For each postal code, latitude and longitude could be found in the geographical coordinates file. It was then added into the dataframe through matching its postal code, which generated completed version dataframe. Finally, the dataset with the word ‘Toronto’ was extracted from its original dataset and put into a new dataframe (see Table 2). A map of Toronto area was plotted as following (see Fig 2).

Table 2 Toronto area dataset

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M4W	Downtown Toronto	Rosedale	43.679563	-79.377529
1	M4X	Downtown Toronto	Cabbagetown , St. James Town	43.667967	-79.367675
2	M4Y	Downtown Toronto	Church and Wellesley	43.665860	-79.383160
3	M5A	Downtown Toronto	Harbourfront , Regent Park	43.654260	-79.360636
4	M5B	Downtown Toronto	Ryerson , Garden District	43.657162	-79.378937

Figure 2 Toronto area map



2.2.3 Venue dataset

Before analyzing each neighborhood, there was a need to obtain the venue information of both target areas. First, a radius of 500 meters and the limited venue number of 100 were set. Second, after the

call to Foursquare API was made, a json file can be requested containing all the information of interest in the item key. Third, knowing that attributes to extract included venue name, venue categories, venue latitude and longitude, the json file was filtered and cleaned, leaving all the needed data for each venue which was then appended into a list. Finally, by corresponding to the neighborhood name and its coordinates, the venue information was combined with neighborhood information into a same dataframe (see Table 3 & 4). The size and venue categories of Manhattan dataset were (3331, 7) and 333 while Toronto area dataset had the size and venue categories of (1467, 7) and 222.

Table 3 Manhattan venue dataset

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
1	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	Coffee Shop
4	Marble Hill	40.876551	-73.91066	Dunkin'	40.877136	-73.906666	Donut Shop

Table 4 Toronto area venue dataset

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rosedale	43.679563	-79.377529	Rosedale Park	43.682328	-79.378934	Playground
1	Rosedale	43.679563	-79.377529	Whitney Park	43.682036	-79.373788	Park
2	Rosedale	43.679563	-79.377529	Alex Murray Parkette	43.678300	-79.382773	Park
3	Rosedale	43.679563	-79.377529	Milkman's Lane	43.676352	-79.373842	Trail
4	Cabbagetown , St. James Town	43.667967	-79.367675	Cranberries	43.667843	-79.369407	Diner