Mapping Neighborhoods In Toronto & Manhattan : Profiling

Faisal Mazlan

August 18th, 2021

Introduction & Business Problem:

Problem Background:

Both New York and Toronto are major metropolitan areas and most populous in their respective countries. They are also considered major international centers for business, finance, arts and culture, having a very multicultural and cosmopolitan population.

There is plenty of opportunity for profitable ventures however the cost of doing business is also higher than other locations. The restaurant business catering to specific cuisines is addition to the multicultural cosmopolitan identity of these cities. An analysis of the popularity of these locations and the population distribution in the respective areas can help planners targeting the market.

For a potential traveller/immigrant/expatriate , knowing these hotspots for certain cuisines, type of shops, what these neighborhoods are known for would also help. It is much easier to settle in a neighborhood where there is a large proportion of people with similar backgrounds and with easy to access restaurants offering dishes from back home.

Problem Description

The restaurant business is important in both these cities. Torontonians dine out an average of 3.1 times per week while New Yorkers dine out roughly 3 times a week according to Zagat's 2012 survey. While the number of dining out patrons has declined tremendously due to the pandemic, the amount of delivery/takeouts have risen by 2x on average (Zagat Future of Dining study).

There is serious competition in both of these cities, however knowing what type of cuisine and the distribution of ethnicities in the city will help ABC Company LTD. in deciding what sort of cuisine and location to invest in.

Some of the factors that were investigated include

1. Population distribution between the two cities (New York : Manhattan Borough, Toronto : Toronto Boroughs)
2. Ethnicity (Asian, Latin American) population distribution
3. Distribution of restaurant and types among neighborhoods identified

While there are many more information that could further assist in this feasibility study such as the current price range, socioeconomic factors, transportation etc, the factors identified above were deemed sufficient for this sprint.

Target Audience:

While this study involves an initial feasibility check for prospective restaurateurs, the study might also help prospective immigrants or travellers looking to find a place to stay. For example, travellers might like to stay in areas where a certain cuisine or cafe is easier to find in either city.

Data

Two cities will be analysed in this project

New York : Manhattan Borough

Toronto : Neighborhoods containing Toronto

Data Sources & Processing

Data 1 (Location/Neighborhoods):

New York Neighborhood dataset has a total of 5 boroughs and 306 neighborhoods. This dataset includes the Boroughs, Neighborhood in them and the latitude and longitude coordinates of each neighborhood. We will initially focus only on the Manhattan borough.

link: https://geo.nyu.edu/catalog/nyu_2451_34572

The data was downloaded by JSON and converted to dataframe format. The dataset was filtered for only the Manhattan Borough. Further boroughs were planned to be added, however, there were certain errors when trying to combine with the next data set (venues). Decision was made to add the other boroughs in future sprints.

Toronto:

First, we used Beautifulsoup on the wikipedia page to obtain the Boroughs and Neighborhoods in Toronto and create a dataframe. Then, using nominatim geolocator (geopy) addon, we proceeded to get each neighborhoods Latitude and Longitude coordinates

Link: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

The boroughs in the data set was manually cleaned, for example EtobicokeNorthwest was converted to Etobicoke Northwest with a manual replace line in the code.

Data 2 (Venue):

Using the above location data, we proceeded to use the explore function of Foursquare API. We defined the search radius for a list of 100 venues within 500m of each neighborhood. Other important details of the venue were not explored in this sprint

example of FourSquare API output in dataframe

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Roselawn	43.7113	-79.4195	Ceiling Champions	43.713891	-79.420702	Home Service
1	Roselawn	43.7113	-79.4195	BAGEL TIME MONTREAL STYLE	43.709067	-79.415858	Fast Food Restaurant
2	Roselawn	43.7113	-79.4195	The Bewding	43.707047	-79.417756	Spa
3	Summerhill West, Rathnelly, South Hill, Forest...	43.6861	-79.4025	The Market By Longo's	43.686711	-79.399536	Supermarket
4	Summerhill West, Rathnelly, South Hill, Forest...	43.6861	-79.4025	LCBO	43.686991	-79.399238	Liquor Store

At times, there were issues with a limitation of only 23 venues per neighborhood regardless of the limit we place in the get subcommand. However, a rerun of the code managed to get us to that 100 limit per neighborhood.

Data 3: Population/ethnicity and Neighborhood boundaries

New York :

a) Ethnicity data : census 2010 data from the official website of the City of New York. PDF format, needed to convert to table and did some data shaping using excel
The PDF was converted to csv format externally and table formatting was done using excel. This was then imported into a dataframe in python.

https://www1.nyc.gov/assets/planning/download/pdf/data-maps/nyc-population/census2010/t_pl_p3a_nta.pdf

	Borough	2010\nCensus FIPS County Code	Code	Name	Total\nPopulation	White	Black/ African American	American Indian and Alaska Native	Asian	Native\nHawaiian and Other Pacific Islander	Some Other Race	Total	Hispanic Origin (of any race)
0	Bronx	5.0	BX01	Claremont-Bathgate	31078	370	13036	82	108	7	60	221	17194
1	Bronx	5.0	BX03	Eastchester-Edenwald-Baychester	34517	1111	24381	147	232	10	148	509	7979
2	Bronx	5.0	BX05	Bedford Park-Fordham North	54415	3637	9805	145	2727	29	490	623	36959
3	Bronx	5.0	BX06	Belmont	27378	5381	5059	38	620	18	84	249	15929
4	Bronx	5.0	BX07	Bronxdale	35538	5559	10594	102	1404	9	169	430	17271
...	...	...	...	...	...	...	...	...	...	...	...	...	...
190	Staten Island	NaN	SI37	Stapleton-Rosebank	26453	9910	5097	72	2565	12	102	427	8268
191	Staten Island	NaN	SI45	New Dorp-Midland Beach	21896	17136	261	28	1148	2	31	215	3075
192	Staten Island	NaN	SI48	Arden Heights	25238	20328	373	22	1675	4	27	249	2560
193	Staten Island	NaN	SI54	Great Kills	40720	35649	169	26	1233	8	56	331	3248
194	Staten Island	NaN	SI99	park-cemetery-etc-Staten Island	0	0	0	0	0	0	0	0	0

195 rows × 13 columns

Figure 1 2010 NYC census dataframe

b) Population data : also 2010. Taken from NYC open data website
https://data.cityofnewyork.us/City-Government/New-York-City-Population-By-Neighborhood-Tabulatio/swpk-hqdp

c) GeoJSON of neighborhoods : the neighborhood boundaries taken from GeoJSON from the city of New York website
https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-nynta.page#:~:text=Archived%20Data%20Sets-,Neighborhood%20Tabulation%20Areas%20(Formerly%20%22Neighborhood%20Projection%20Areas%22),plan%20for%20New%20York%20City

Toronto :

a)2016 census data : taken from toronto open data portal home ,
https://open.toronto.ca/dataset/neighbourhood-profiles/

The neighborhood profiles dataset contains a lot of information from census taken in 2016 in Toronto and tabulated as per the neighborhood demarcations. Information such as age profile,ethnicity, income etc. are contained in this dataset. However, we are only using the population by ethnicity data. This also required some reshaping of the data arrangement to be able to get the desired dataframe

_id	Category	Topic	Data Source	Characteristic	Agincourt North
1	Neighbourhood Inform	Neighbourhood Info	City of Toronto	Neighbourhood Number	129
2	Neighbourhood Inform	Neighbourhood Info	City of Toronto	TSNS2020 Designation	No Designation
3	Population	Population and dwel	Census Profile 98-316-X2016001	Population, 2016	29,113
4	Population	Population and dwel	Census Profile 98-316-X2016001	Population, 2011	30,279
5	Population	Population and dwel	Census Profile 98-316-X2016001	Population Change 2011-2016	-3.90%
6	Population	Population and dwel	Census Profile 98-316-X2016001	Total private dwellings	9,371
7	Population	Population and dwel	Census Profile 98-316-X2016001	Private dwellings occupied by usual	9,120
8	Population	Population and dwel	Census Profile 98-316-X2016001	Population density per square kilom	3,929

Figure 2 Toronto Neighborhood Profile example

Category	index	Population	Latin American origins	Asian origins
0	Agincourt North	29113.0	470.0	24305.0
1	Agincourt South-Malvern West	23757.0	480.0	17955.0
2	Alderwood	12054.0	315.0	2055.0
3	Annex	30526.0	765.0	6485.0
4	Banbury-Don Mills	27695.0	585.0	12025.0
...	...	...	...	...
135	Wychwood	14349.0	645.0	2500.0
136	Yonge-Eglinton	11817.0	370.0	2895.0
137	Yonge-St.Clair	12528.0	300.0	2330.0
138	York University Heights	27593.0	2055.0	12550.0
139	Yorkdale-Glen Park	14804.0	1025.0	4090.0

140 rows × 4 columns

Figure 3 Toronto neighborhood profile after data shaping

b) GeoJSON of neighborhoods : GeoJSON of Toronto neighborhoods taken from link below.
https://nad.carto.com/tables/neighbourhoods_toronto/public/map

Methodology

Business Understanding:

Exploratory analysis of composition of neighborhoods and availability/popularity of certain cuisines in them.

Analytic Approach:

New York city has a total of 5 boroughs and 306 neighborhoods. We are focusing only on Manhattan borough for this sprint. Toronto is unofficially known as “ the city of neighbourhoods”. There are over 140 neighbourhoods officially recognized by the City of Toronto and upwards of 240 official and unofficial neighbourhoods within city limits, divided into 5 districts. For Toronto, we are only focusing on neighborhoods containing ‘Toronto’.

Exploratory Data Analysis

Neighborhoods

1. We load data fron newyork_data.json file
2. Transform data of nested python dictionaries into a pandas dataframe
    a. This dataframe contains the geographical coordinates of New York city neighborhoods
3. Use the above coordinates to feed into the ‘get’ command using the Foursquare API
4. Use Geopy and folium libraries to create map of New York city with neighborhoods superimposed

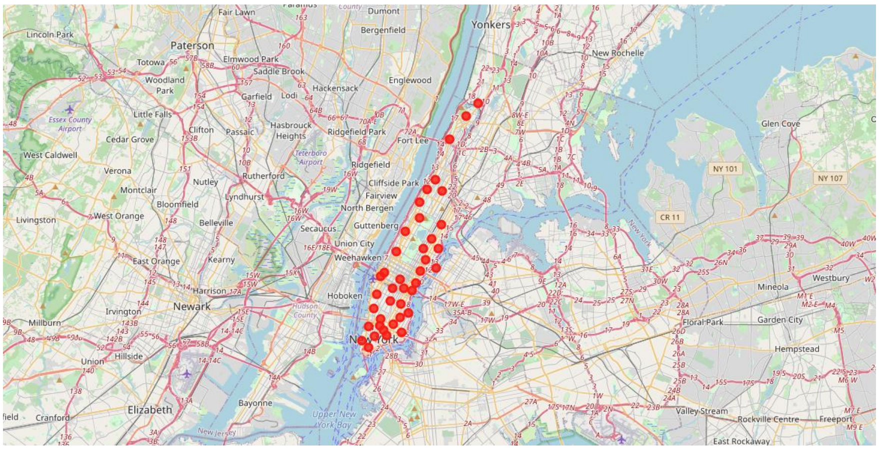


Figure 4 Neighborhood samples in Manhattan, New York

The same is done for Toronto, with minor differences, as can be seen below

1. The list of neighborhoods is scraped using beautifulsoup from wikipedia
2. Transform data from the beautiful soup table into a pandas dataframe
    a. This dataframe contains only the list of neighborhoods with the postal codes
3. Using Geopy, querying the postalcodes, we add the Latitude and Longitude coordinates of the neighborhoods
4. Use the above coordinates to feed into the ‘get’ command using the Foursquare API
5. Use Geopy and folium libraries to create map of New York city with neighborhoods superimposed

Figure 5 Neighborhood samples in Toronto

A brief check on the number of neighborhoods being compared show that 40 neighborhoods are listed under the Manhattan borough while we have 39 neighborhoods in Toronto being listed spread over multiple boroughs

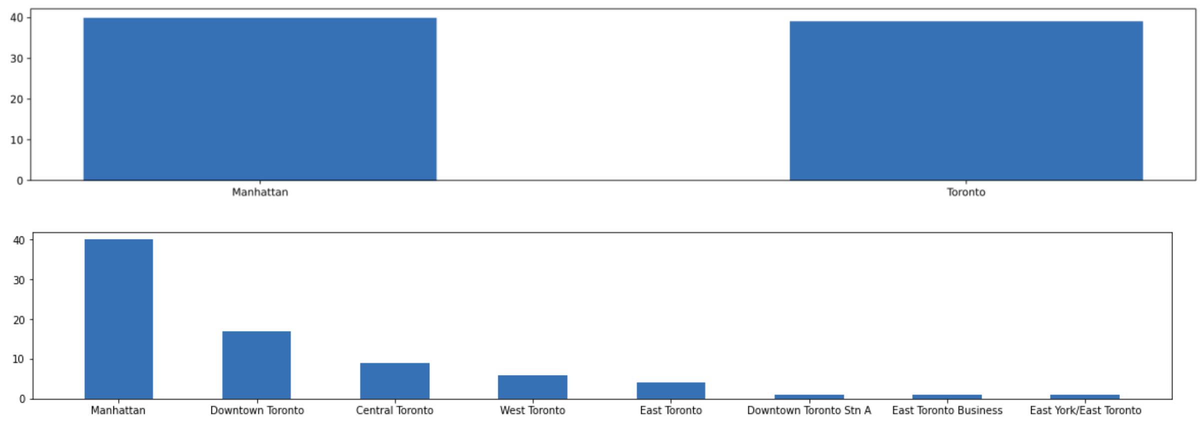


Figure 6 Distribution of neighborhoods by boroughs in sample

Foursquare Venues

The number of venue results from the Foursquare API are as follows

	Neighborhood	Venue	Venue Category
city			
Manhattan	40	2787	332
Toronto	39	969	217

There are far more venues within roughly the same amount of neighbourhoods with the same search parameters (500m radius, 100 max venues per neighborhood). This is probably due to the sheer size and density of New York city (pop 8.8 million, 11,232/km^2) compared to Toronto (pop 2.7 million, 4334/km^2). The distribution of the venues can also be seen in the graph below.

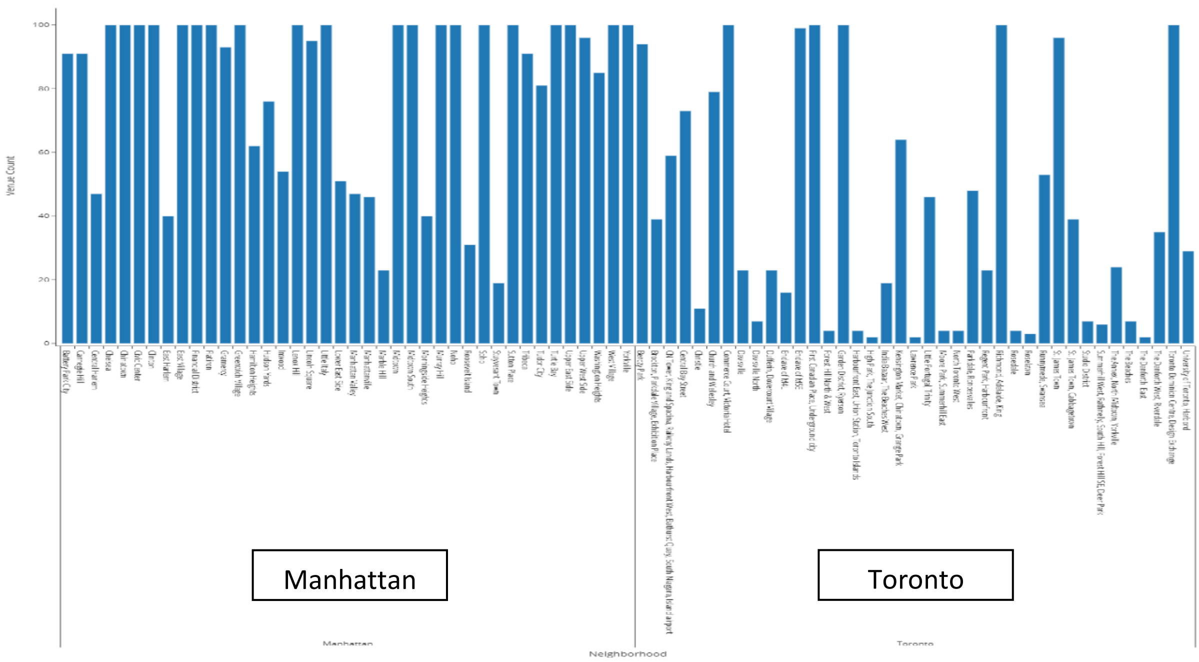


Figure 7 Distribution of venue count by neighborhood and city

In terms of the most numerous venue types, it can be seen in the graph below.

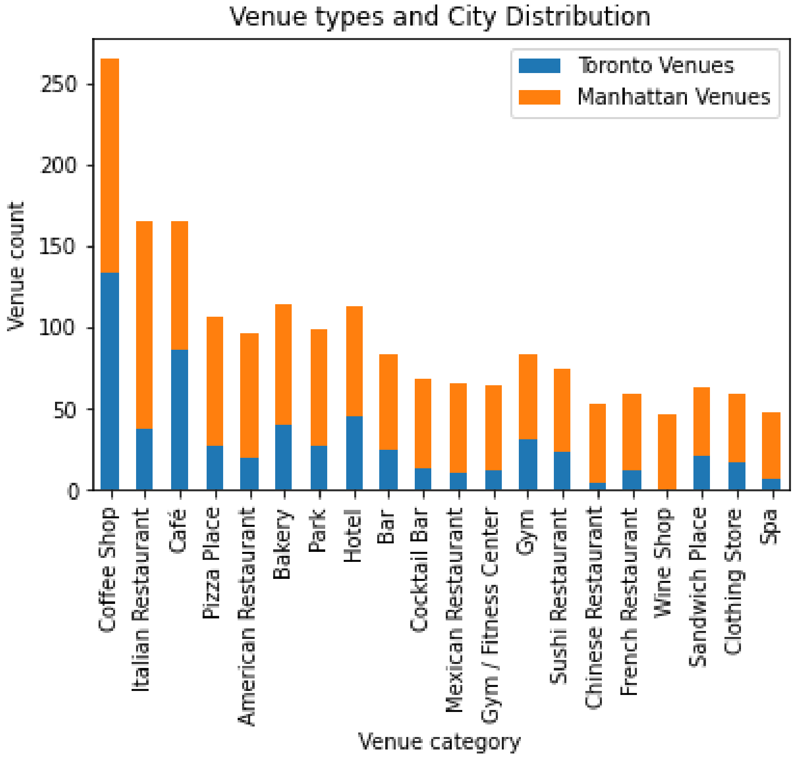


Figure 8 Venue count by type and city

The same graphs created for individual cities give a better picture of the preference of the city. We can see similarities in that:

1. Italian restaurants are highly popular in both areas
2. Asian restaurants (Sushi, Japanese, Chinese, Asian, Ramen, Vietnamese etc.) are also highly popular in both
3. Considering the high number of Latin American population in the New York city area (similar to Asian population), the proportion of Latin American restaurants are not so high with only Mexican restaurants being popular.

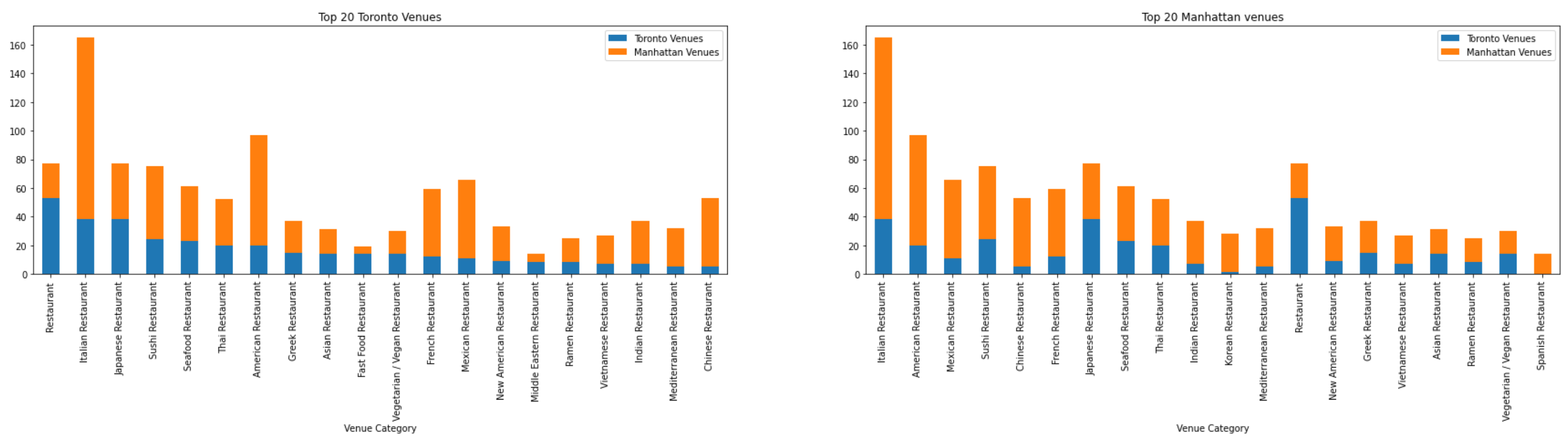
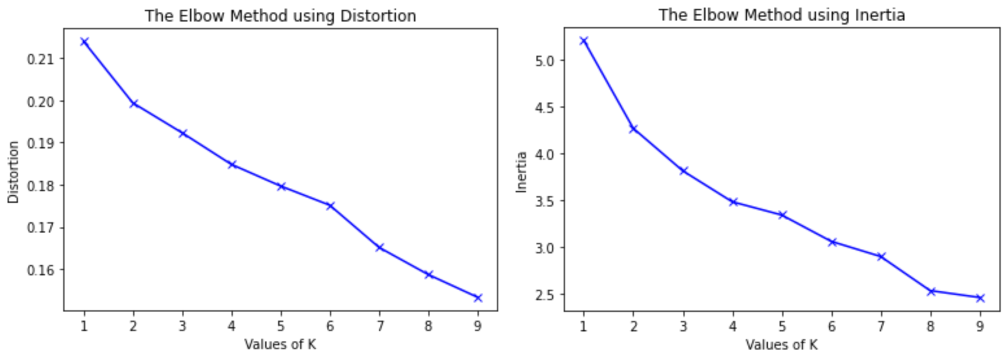


Figure 9 Venue count by type and city sorted by popularity in each city

Results

From this venues data, we clustered the entire venues data to group together similar neighborhood types together across both cities. A K-Means clustering method was used on the mean occurrence of venue categories. In this case, the venues categories still encompassed all the venue categories.

The optimum number of K folds is first checked using the elbow method using both distortion and inertia parameters . Following the inertia method, an estimate of around K=3 is found. However, we chose to create the model with k=5.



Below are the outcomes of the clusters.

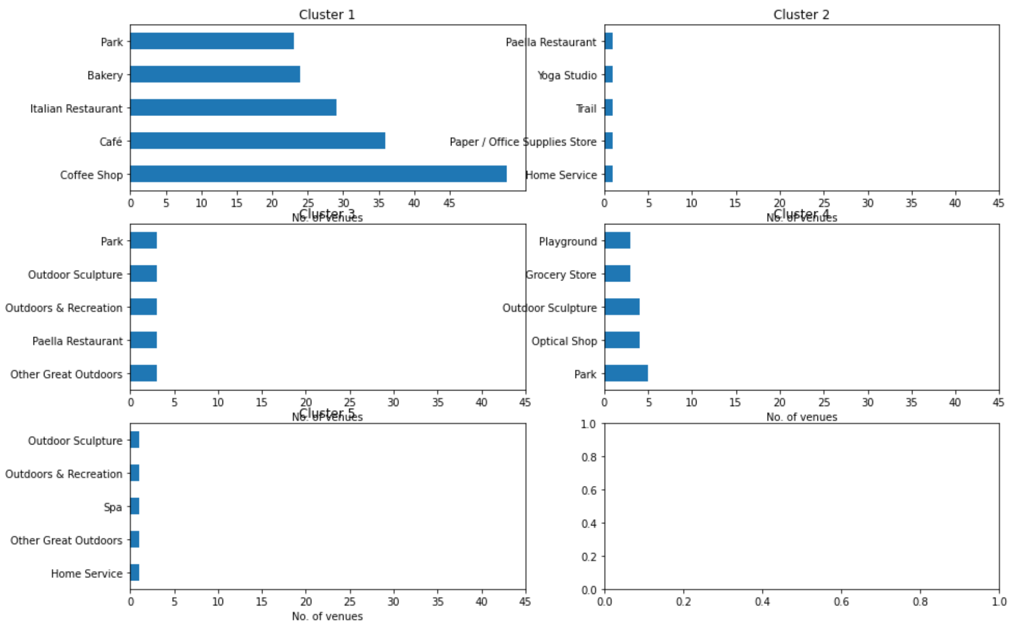


Figure 10 Neighborhood cluster profile.
Note that cluster 1 is actually cluster 0

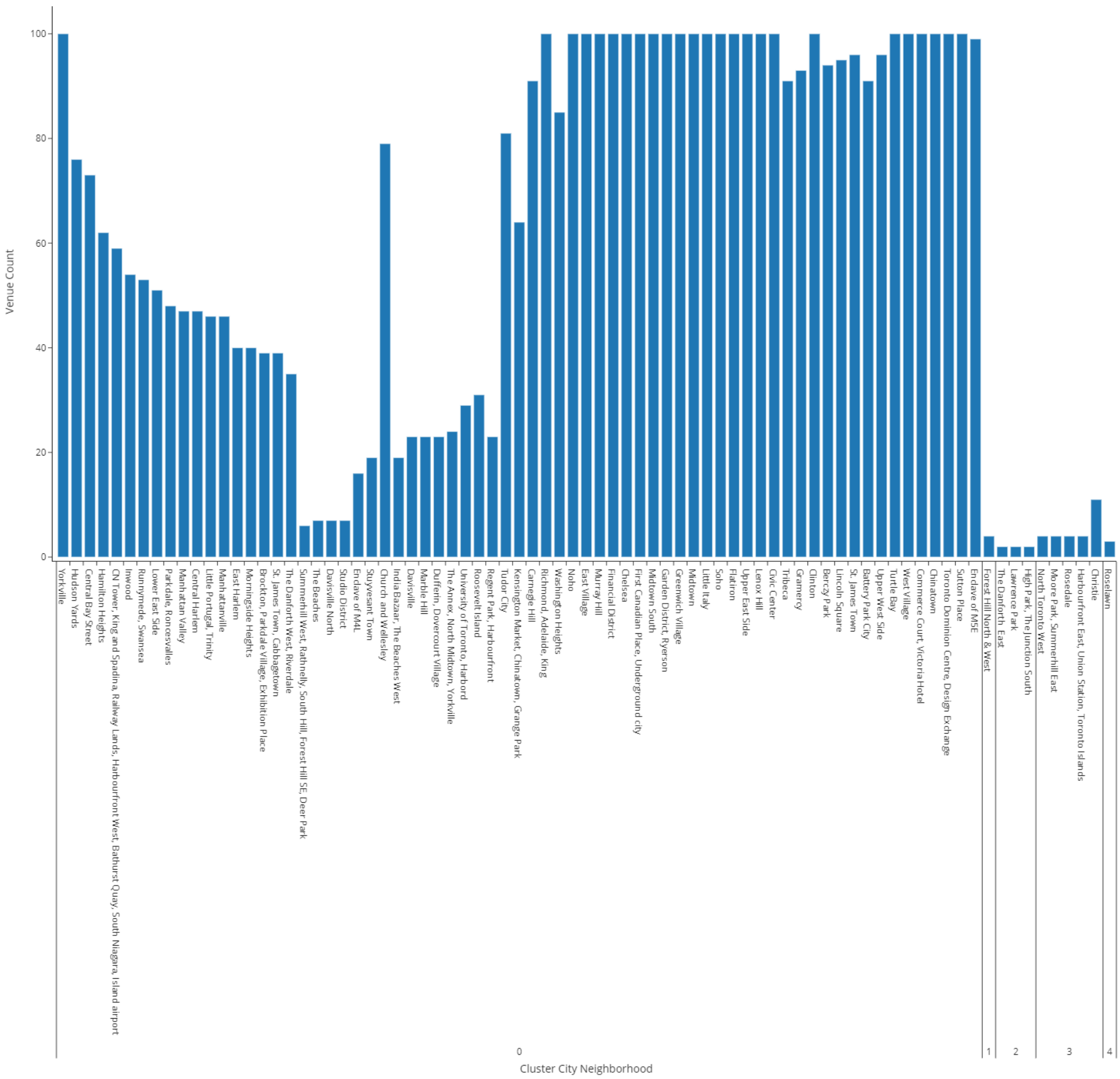


Figure 11 Clustering of Neighborhoods vs quantity of venues

From the clustering of the neighborhoods, we can see

1. Cluster 0 is the one with the highest number of neighborhoods. It is also the one where the composition does not differ much where coffee shops,cafes and eateries are highly popular. This signifies a much more cosmopolitan/central sort of neighborhood where plenty of business activity is performed.
2. All the other clusters have much smaller densities in terms of venues denoting much less population, economic activity or simply because a large part of the neighborhood consists of open spaces
3. Cluster 3 seems to consist mostly of residential neighborhoods, with grocery parks, stores, and playgrounds being the popular venues.
4. From the maps below, we can see the red dots (cluster 0) being primarily concentrated in high population areas especially in Toronto. This differentiation is not seen in the Manhattan data set as all the associated neighborhoods are in the same data cluster.

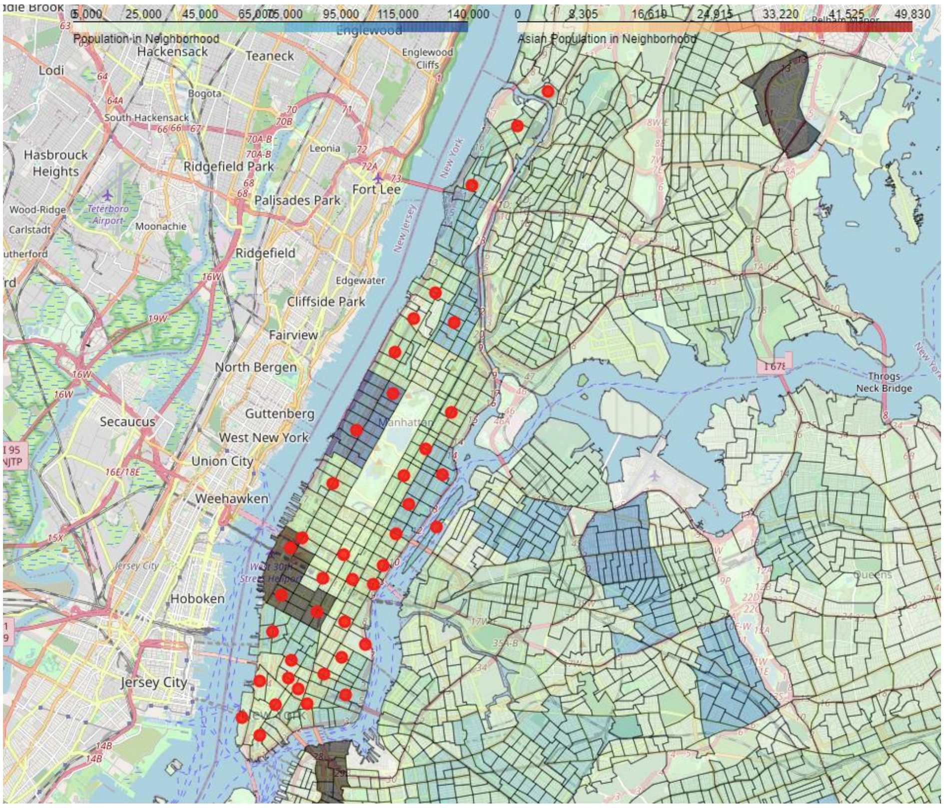


Figure 12 Map of New York with choropleth of population

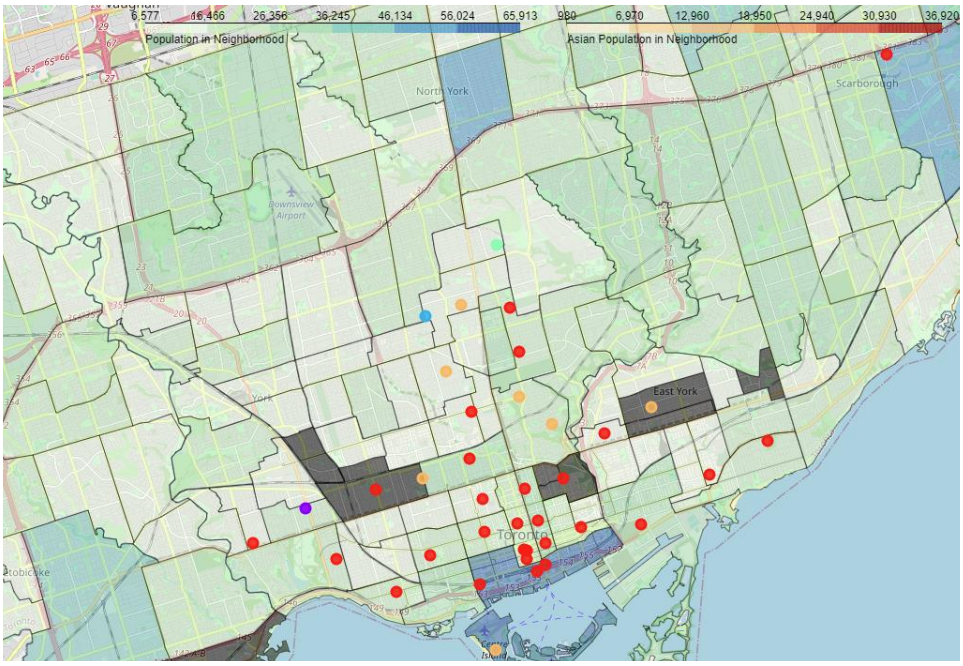


Figure 13 Map of Toronto with choropleth of population

A further check into Asian and Latin American population distribution into the two cities might assist ABC Trading LTD further into where best to invest a restaurant in.

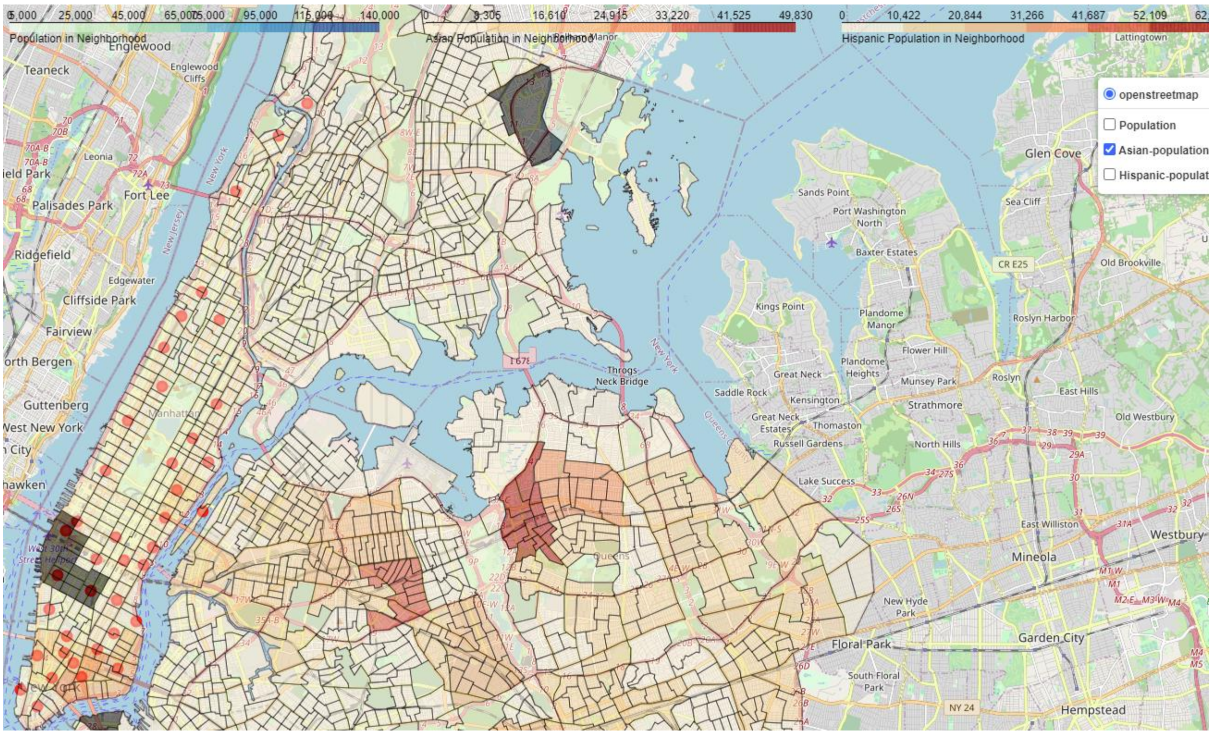


Figure 14 Map of New York with choropleth of Asian population

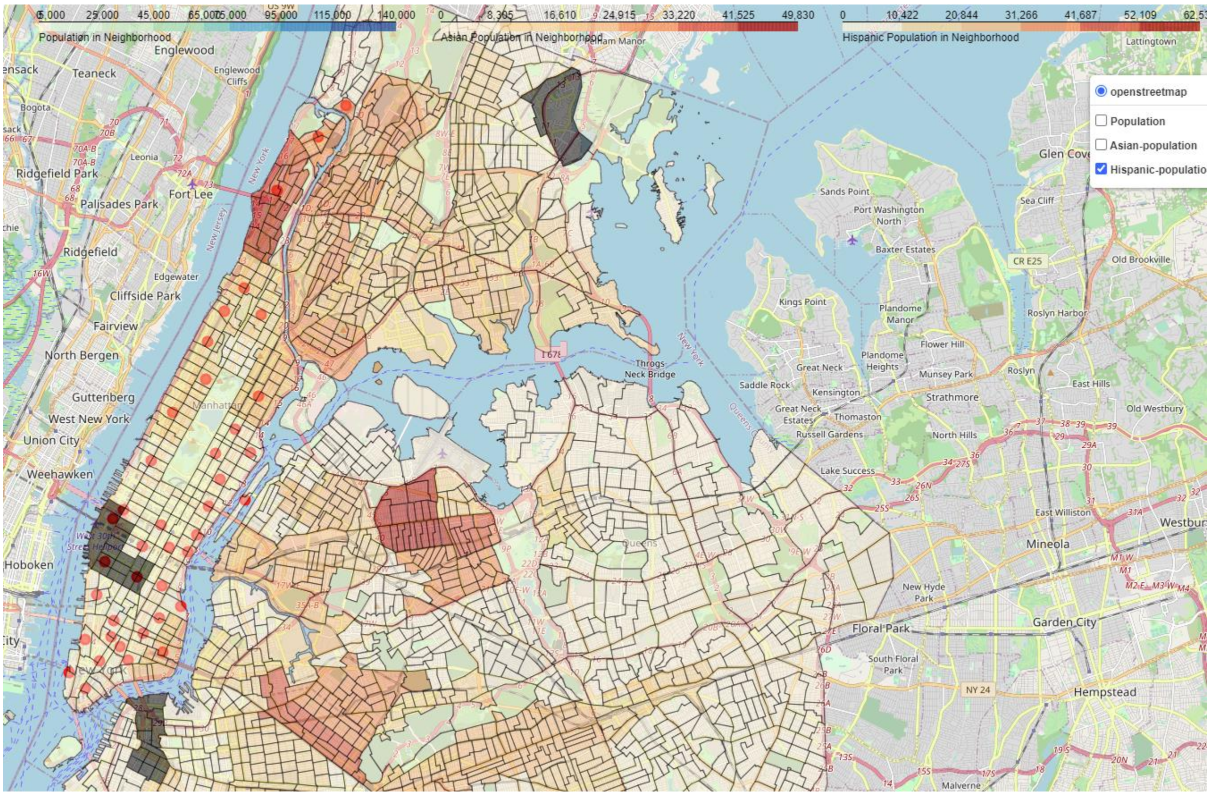


Figure 15 Map of New York with choropleth of Hispanic population

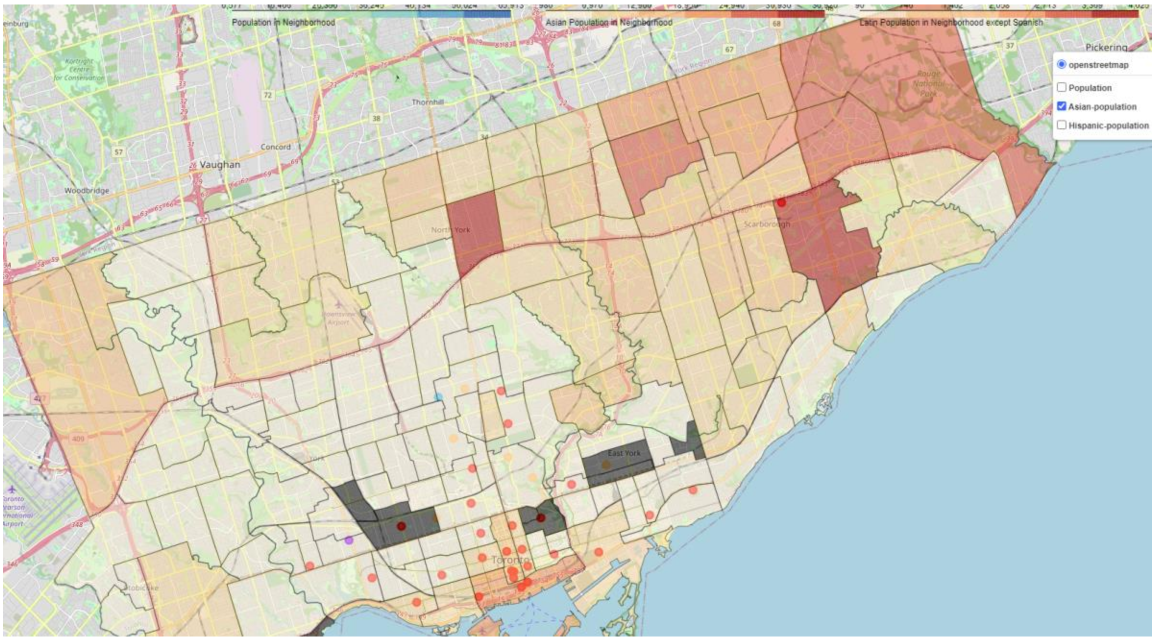


Figure 16 Map of Toronto with choropleth of asian population

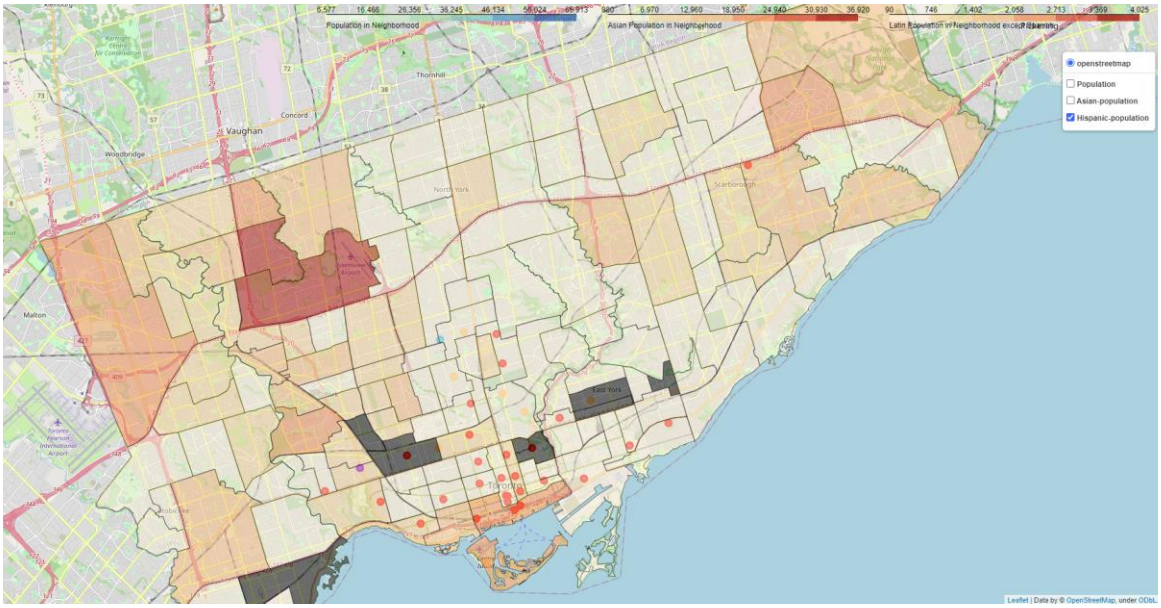


Figure 17 Map of Toronto with choropleth of Hispanic population

Discussion

1. For a prospective investor into the restaurant in either city, another useful information would be real estate prices either for rental or purchase
2. There is much more competition in Manhattan but this is proportionate to the density of the population compared to Toronto
3. With the pandemic, a different approach might be considered. A restaurant does not need to be located exactly in the middle of a population dense area if the location can be easily serviced by food delivery riders.
4. The data is inconclusive as the current data set is not yet fully explored. There is scope for
    a. investigating on current popularity, pricing, menu of the venues using foursquare

b. investigating the status of more neighborhoods especially ones that have high contrast of ethnicity, tourist footfall etc.

Conclusion

This analysis was performed on limited time and data. There is plenty more open access data that can be utilized to give a better picture of the neighborhoods, that would take more time for us to utilize. Expanding the areas investigated and finding more differences and similarities between the two cities for other purposes might be interesting as well. For example, if we could find a specific type of venue that was popular in one city but not easily found in another, could we possibly find similar neighborhood profiles to give it a better chance of success?