

# Battle of Neighborhoods – Capstone Project – Full Report

## Problem Background

An established well-known business man and investor from India, who wants to stay anonymous until an official public announcement, has invited multiple sales, marketing and engineering teams from all over the world to help him in providing with a decision to choose a city in United States or Canada to expand his widely famous and high quality Indian sweets and delicacies manufacturing center.

Indian sweets either be it in Canada or United States are being accepted and tasted around the world with all types of cultures and societies, and by opening up this famous manufacturing center it's expected to attract more customers from a global diaspora. Overall, this will be a win-win scenario for both the business and the customers too. Due to their high quality and extremely tasty food items, the customers will get a chance to taste a new variety without travelling the world and the business will benefit from an increased audience outside Indian communities.

In order to do an experiment, the investor wishes to open up a small food-stall with all the signature delicacies and later expand to build a full size manufacturing unit. All the teams need to determine the best location to open the food-stall so that it can be reached to a wide audience.

## Problem Description

One of the marketing and engineering teams together form a Data Science team to help this investor in getting the data required for the investor to make an appropriate decision based on the data they will collect by comparing the facts from locations with a high cultural diversity in both the countries - New York City (United States) and Toronto (Canada).

The team has to gather the available data, do the required data processing, analysis, segmentation and exploration of different parts of the cities and provide a report which will encompass information such as:

Number / Density of current restaurants or food stalls

Possible locations for growth etc.

There must be enough information out of this report, including the analysis and summarizations done on each location, which should help the investor to make an appropriate decision.

## Data Collection

I will use the publicly available data on New York city and Toronto, which was used in previous modules, along with the Foursquare API to gather all the required information. If required, I will also search for other available data sets which are publicly available and can provide an additional insight into creating a report for a wise decision making.

The reference URLs which I will begin to use are mentioned below. I will also add more data sets as and when I progress with creating the report.

- [https://en.wikipedia.org/wiki/List\\_of\\_postal\\_codes\\_of\\_Canada:\\_M](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)
- [https://geo.nyu.edu/catalog/nyu\\_2451\\_34572](https://geo.nyu.edu/catalog/nyu_2451_34572)

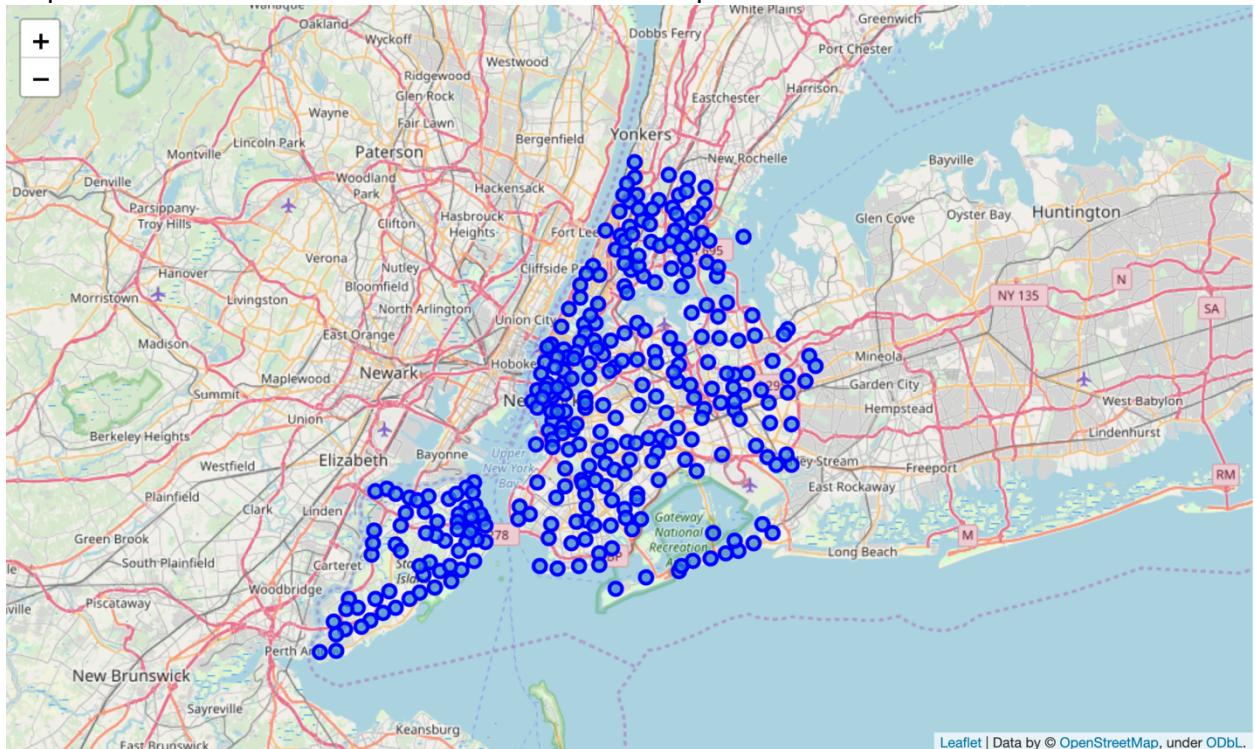
## Methodology

### New York Data Set

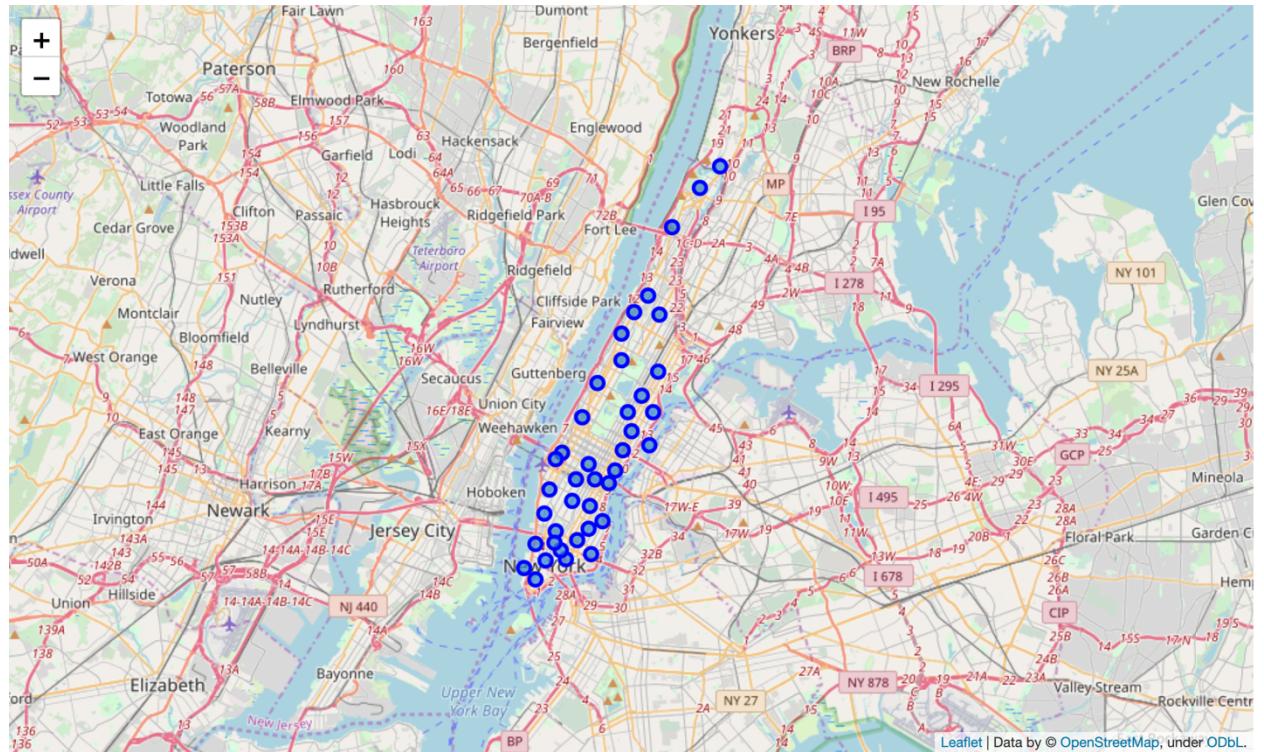
- **Data Cleansing:** I have started with the data collection for the city of New York by using the URL given above ([https://geo.nyu.edu/catalog/nyu\\_2451\\_34572](https://geo.nyu.edu/catalog/nyu_2451_34572)).
- For this data, I have pre-processed it using the techniques learnt in the modules of the Data science course, where I had cleared off the “Not assigned” data from the columns Borough and Neighborhood columns.
- For the Boroughs, the entire rows were removed which had “Not assigned” values.
- For the Neighborhoods, the values were copied from the borough values when it was having a “not assigned” value.
- Also, to simplify the grouping, I had grouped the similar postal codes together and separated the multiple neighborhoods with a comma separated values list.
- Finally, the data was perfect to be used for further analysis. Example:

	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

- **Data Analysis:** For the data fetched and pre-processed, I used the “geopy” library to get the coordinate values for latitude and longitude for all the neighborhoods in New York city.
- With this data, I will be gathering all the required information about the number of restaurants, hotels, coffee shops, places to visit, etc. in each area/neighborhood.
- The more the density of these top places in the vicinity of an area, the more will be the active lifestyle of the people and communities.
- This data was to be used for processing by the Foursquare API to get the nearby venues and attractions.
- The initial data which was projected for New City analysis was mapped with Folium maps and was marked with all the observed location points as shown below:



- Since Manhattan is the most busy area, let's dive to focus over the marks in Manhattan area:

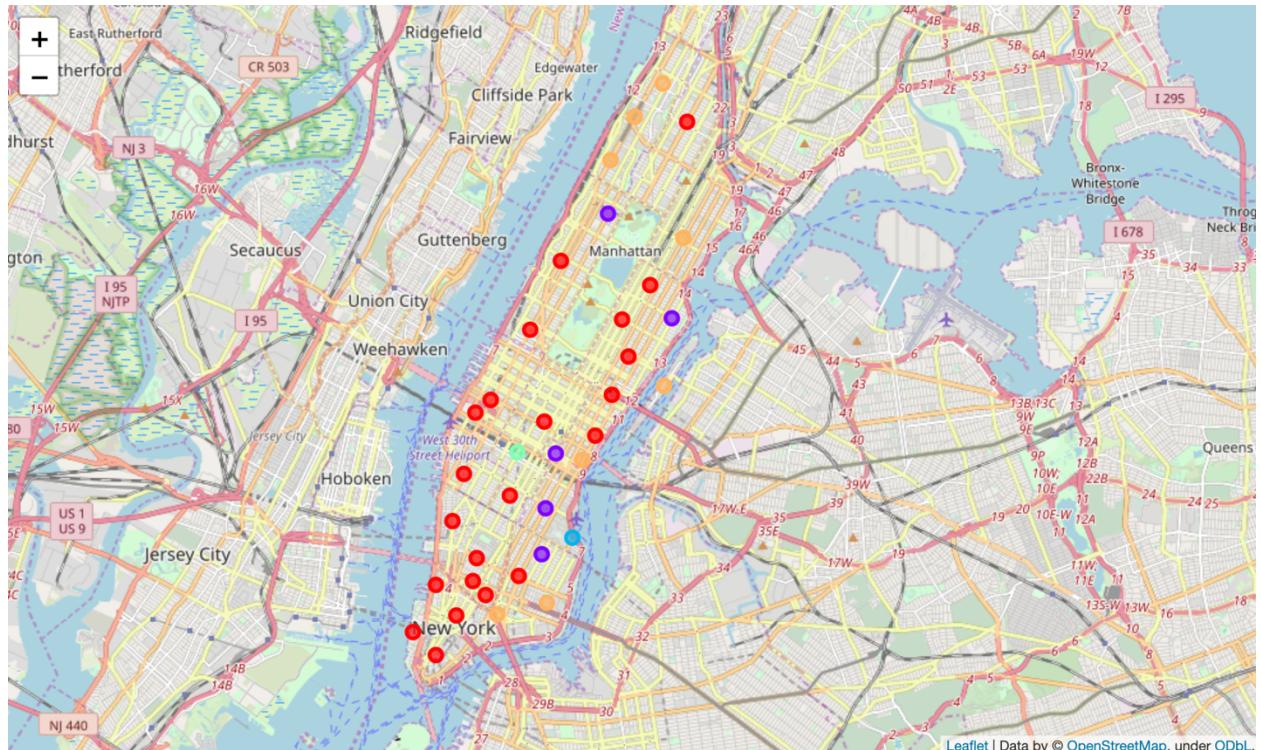


- **Analyzing each individual Neighborhood:** In order to get information about each location, I used the Foursquare API to trigger calls to the server and fetch information about each marked location at Manhattan:

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Animal Shelter	Antique Shop	Arcade	Arepas Restaurant	Argentinian Restaurant
0	Battery Park City	0.000000	0.00	0.00	0.000000	0.010000	0.00	0.00	0.00	0.00	0.000000
1	Carnegie Hill	0.000000	0.00	0.00	0.000000	0.010000	0.00	0.00	0.00	0.00	0.000000
2	Central Harlem	0.000000	0.00	0.00	0.065217	0.043478	0.00	0.00	0.00	0.00	0.000000
3	Chelsea	0.000000	0.00	0.00	0.000000	0.030000	0.00	0.00	0.00	0.00	0.000000
4	Chinatown	0.000000	0.00	0.00	0.000000	0.040000	0.00	0.00	0.00	0.00	0.000000
5	Civic Center	0.000000	0.00	0.00	0.000000	0.030000	0.00	0.01	0.00	0.00	0.000000
6	Clinton	0.000000	0.00	0.00	0.000000	0.040000	0.00	0.00	0.00	0.00	0.000000
7	East Harlem	0.000000	0.00	0.00	0.000000	0.000000	0.00	0.00	0.00	0.00	0.000000
8	East Village	0.000000	0.00	0.00	0.000000	0.020000	0.00	0.01	0.00	0.02	0.010000

Financial

- This process was repeated for all the neighborhoods and obtained the top 5 locations as per the Foursquare data trends.



	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
6	Central Harlem	African Restaurant	American Restaurant	Seafood Restaurant	Chinese Restaurant	Bar	French Restaurant	Fried Chicken Joint	Southern / Soul Food Restaurant	Market	Tapas Restaurant
8	Upper East Side	Italian Restaurant	Exhibit	Bakery	Art Gallery	Coffee Shop	Gym / Fitness Center	Hotel	French Restaurant	Juice Bar	Yoga Studio
10	Lenox Hill	Coffee Shop	Italian Restaurant	Pizza Place	Sushi Restaurant	Café	Burger Joint	Sporting Goods Shop	Cocktail Bar	Gym / Fitness Center	Gym
12	Upper West Side	Italian Restaurant	Bar	Coffee Shop	Wine Bar	Mediterranean Restaurant	Bakery	Café	Gym / Fitness Center	Ice Cream Shop	Indian Restaurant
13	Lincoln Square	Theater	Café	Plaza	Italian Restaurant	Gym / Fitness Center	Performing Arts Venue	Concert Hall	Park	French Restaurant	Indie Movie Theater
14	Clinton	Theater	Gym / Fitness Center	Italian Restaurant	American Restaurant	Hotel	Coffee Shop	Wine Shop	Sandwich Place	Gym	Spa
15	Midtown	Hotel	Theater	Clothing Store	Coffee Shop	Sporting Goods Shop	Japanese Restaurant	Bakery	Steakhouse	Spa	Bookstore

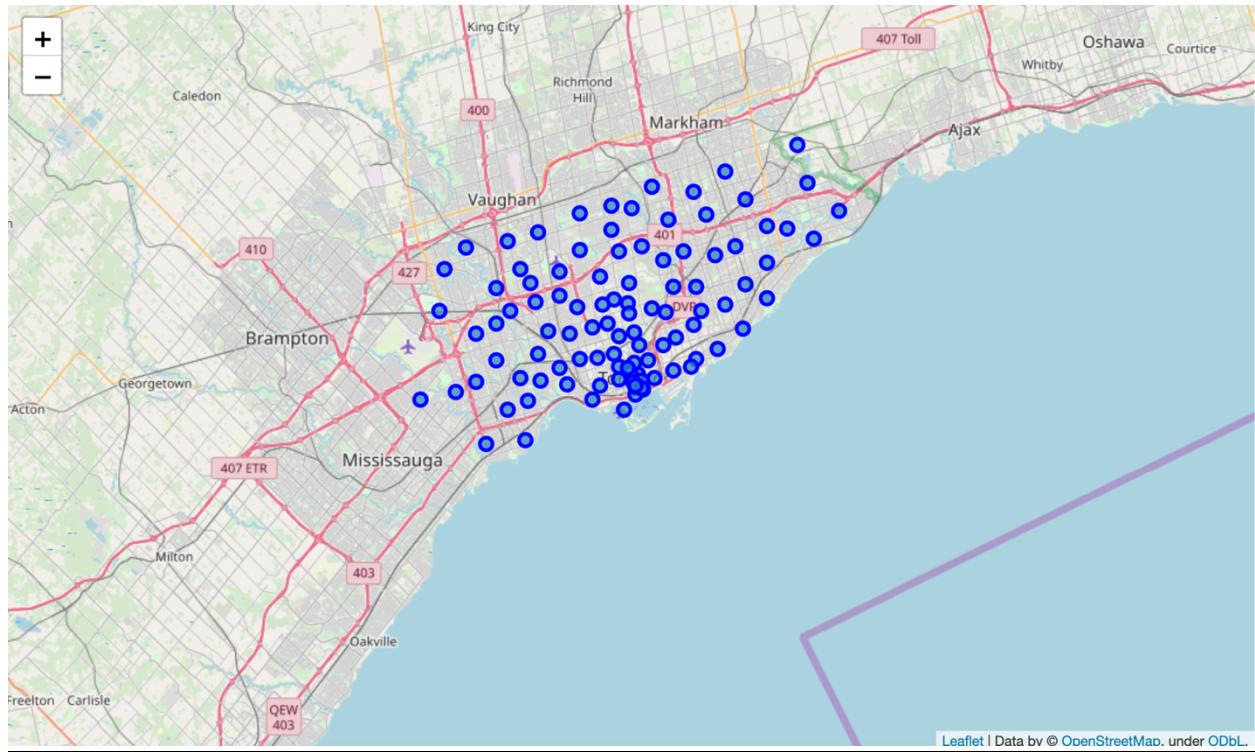
## Toronto Data Set

- **Data Cleansing:** I have started with the data collection for the city of Toronto for one particular zip code location by using the URL given above ([https://en.wikipedia.org/wiki/List\\_of\\_postal\\_codes\\_of\\_Canada:\\_M](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M))
- The data for Toronto city was not directly and immediately usable as that of the New York data set. This had to be pre-processed with more steps.

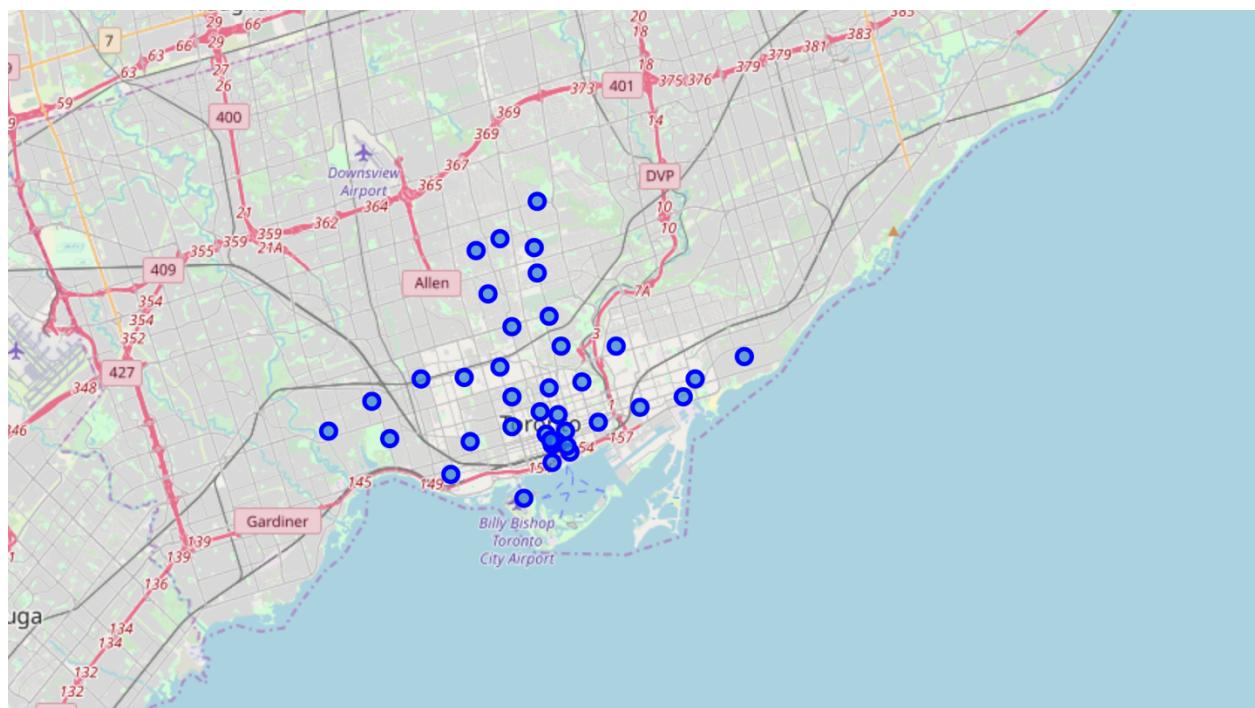
- For this data, I have pre-processed it using the techniques learnt in the modules of the Data science course. It required HTML to JSON processing where I used the libraries available in python to achieve that.
- I had cleared off the “Not assigned” data from the columns Borough and Neighborhood columns.
- For the Boroughs, the entire rows were removed which had “Not assigned” values.
- For the Neighborhoods, the values were copied from the borough values when it was having a “not assigned” value.
- Also, to simplify the grouping, I had grouped the similar postal codes together and separated the multiple neighborhoods with a comma separated values list.

	<b>PostalCode</b>	<b>Borough</b>	<b>Neighborhood</b>	<b>Latitude</b>	<b>Longitude</b>
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

- **Data Analysis:** For the data fetched and pre-processed, I used the “geopy” library to get the coordinate values for latitude and longitude for all the neighborhoods in New York city.
- With this data, I will be gathering all the required information about the number of restaurants, hotels, coffee shops, places to visit, etc. in each area/neighborhood.
- The more the density of these top places in the vicinity of an area, the more will be the active lifestyle of the people and communities.
- This data was to be used for processing by the Foursquare API to get the nearby venues and attractions.



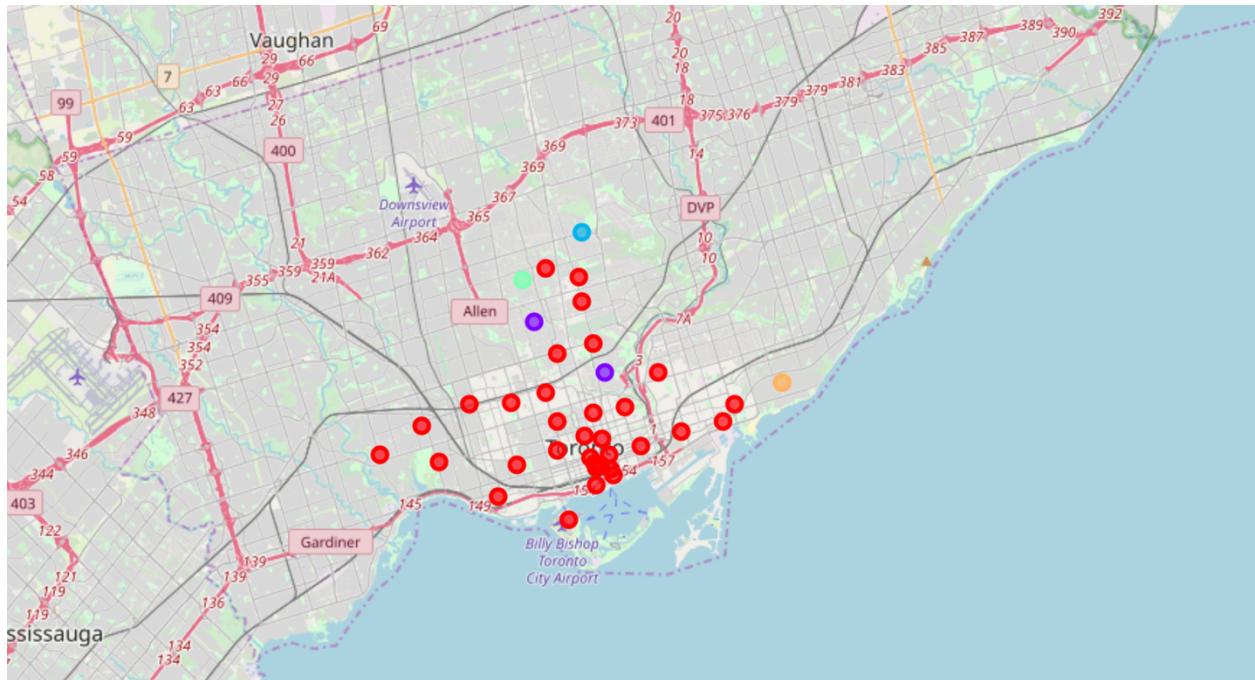
- The next was to get the top destinations for Toronto in the marked areas. Using the Foursquare API, I was able to mark the locations into the Folium map as depicted below.



- Similar to the New York dataset, I had also segmented and clustered the data for the neighborhoods in Toronto for getting the top results in each location.

	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	Adelaide, King, Richmond	Coffee Shop	Café	Bar	Thai Restaurant	Bakery	Steakhouse	Sushi Restaurant	Asian Restaurant	Burger Joint	American Restaurant
1	Berczy Park	Coffee Shop	Café	Cheese Shop	Farmers Market	Bakery	Steakhouse	Seafood Restaurant	Italian Restaurant	Beer Bar	Cocktail Bar
2	Brockton, Exhibition Place, Parkdale Village	Café	Coffee Shop	Breakfast Spot	Grocery Store	Furniture / Home Store	Burrito Place	Sandwich Place	Restaurant	Stadium	Italian Restaurant
3	Business Reply Mail Processing Centre 969 Eastern	Light Rail Station	Yoga Studio	Spa	Gym / Fitness Center	Garden Center	Garden	Fast Food Restaurant	Farmers Market	Comic Shop	Park
4	CN Tower, Bathurst Quay, Island airport, Harbo...	Airport Lounge	Airport Service	Airport Terminal	Boat or Ferry	Airport	Airport Food Court	Airport Gate	Bar	Sculpture Garden	Harbor / Marina

- I used the Folium map again to mark the clustered top locations for Toronto area as shown below:



- For all the top locations, I have also generated the top venues chosen by the people by looking at the trending results from Foursquare API and got all of them as shown below

Borough	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
1	East Toronto	0	Greek Restaurant	Coffee Shop	Ice Cream Shop	Italian Restaurant	Furniture / Home Store	Cosmetics Shop	Brewery	Bubble Tea Shop	Café
2	East Toronto	0	Park	Sandwich Place	Brewery	Steakhouse	Sushi Restaurant	Food & Drink Shop	Fish & Chips Shop	Ice Cream Shop	Movie Theater
3	East Toronto	0	Café	Coffee Shop	Bakery	Italian Restaurant	American Restaurant	Convenience Store	Brewery	Seafood Restaurant	Sandwich Place
5	Central Toronto	0	Gym	Convenience Store	Food & Drink Shop	Hotel	Sandwich Place	Breakfast Spot	Clothing Store	Park	Eastern European Restaurant
6	Central Toronto	0	Clothing Store	Sporting Goods Shop	Coffee Shop	Health & Beauty Service	Burger Joint	Café	Chinese Restaurant	Dessert Shop	Diner
7	Central Toronto	0	Sandwich Place	Dessert Shop	Pizza Place	Sushi Restaurant	Gym	Coffee Shop	Café	Italian Restaurant	Chinese Restaurant
8	Central Toronto	0	Playground	Tennis Court	Restaurant	Gym	Airport Food Court	Farmers Market	Event Space	Ethiopian Restaurant	Electronics Store
											Eastern European Restaurant

## Results

Given the information about the 2 cities of New York and Toronto in the above section, I would like to summarize as below:

1. Both the cities show a large population interested in multiple themes of restaurants.
  - a. This indicates there are a lot of different cultures and communities located in the vicinity of the top sought locations.
  - b. This also indicates people having lot of interest in trying new cuisines and food styles.
  - c. This also indicates quality is very important
2. New York city has a wide variety of venues (other than restaurants)
  - a. This indicates that people are a lot outgoing and willing to spend an extra dollar for experiencing a good quality event.
3. Toronto city has a lot of Indian restaurants
  - a. This indicates that there is a huge population density of Indians as compared to that of New York city.
  - b. This also indicates that the cultural diversity, though being skewed with Indians and Chinese, is a little less diverse than that of New York city.
4. Other influencing factors:
  - a. Taxes for sales in Toronto is higher than that of New York city.
  - b. Number of people visiting New York city is very much higher than that of Toronto.
  - c. New York and Toronto are referred as the financial capitals of their countries, but New York has an advantage of being more attractive to a varied food cuisines.

Overall, I would recommend New York city to be the recommended choice for our investor to start up a new Indian dessert themed restaurant

## Discussion

Observations noted:

- The data to be compared was limited to the top areas with trending population. Would it have been easier to consider less prominent areas so that the operating costs would be lower.
- Is our investor ready to accommodate losses for the first few months since as per the recent observations on market trends, the New York city businesses attract customers by offering low priced / free / discounted items and later increase the sales by making them getting used to their high quality products.

## Conclusion

The choice of a location totally depends on the business needs and the customer requirements. The business in this case is an Indian cuisine with Indian delicacies which are quite famous for the taste as well as the quality of the food ingredients. This analysis helps in deciding what location to choose for based on the data we have from the Foursquare API and which was the New York city. Furthermore improvements can be made to our analysis by including other financial aspects such as the real-estate costs, taxes, cost of owning/renting a property etc.