BakerZ, Inc. Expansion Strategy for Toronto and New York

Shail Rajput

November 22, 2020

1. Introduction/Business Problem

BakerZ, Inc. is a multi-national corporation that specializes in wholesale manufacturing and distribution of fine baked goods to coffee shops including sandwiches, cakes, pastries, and cookies. Quality, freshness, and punctual delivery of the product have been their key success factors is sustaining their customers. These are the factors that BakerZ does not compromise on.

BakerZ is now looking at expanding their business to New York, NY in the USA and Toronto, ON in Canada. They have tasked us to determine the best neighborhood locations in Toronto and New York where they can set up distribution centers.

Their requirements for each distribution center are as follows:
- It should serve multiple neighborhoods with the maximum number of coffee shops that are closest to each other;
- It should only attempt to target neighborhoods with at least 10 or more coffee shops
- It should itself be located in the neighborhood with the maximum number of coffee shops

Based on these requirements, BakerZ, Inc. has tasked us in determining how many distribution centers they should establish in Toronto and New York, and their neighborhood locations. For each city, they also want us to recommend a priority order in establishing the distribution centers so they have a better return on investment.

2. Data

We primarily need the geolocations (latitude and longitude) for all neighborhoods in Toronto, ON and New York, NY to get started.

Once we have the geolocations of each neighborhood, we shall use the Foursquare API to retrieve the data on coffee shops in each neighborhood of each city.

We shall then use this data to solve our business problem as stated above in Section 1.

2.1 Data sources for Toronto, ON

We have two sources of data for Toronto.
Neighborhood data shall be scraped from Wikipedia located at the following URL:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

This data does not have geolocation information for each neighborhood so we shall use a second source that maps postal codes to geolocations to get the complete data we need to

solve our problem. The second source is located in a CSV file called toronto_postal_code_geoloc.csv at the following URL: https://cocl.us/Geospatial_data

We shall then combine these two data sources using the postal code as key to create a single data frame that holds all the data we need for Toronto, mapping neighborhoods to geolocations.

2.2 Data sources for New York, NY

We shall use a single source for New York for the data mapping neighborhoods to geolocations. This data is available in a file called located at the following URL:

https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json

Similar to Toronto, this data shall be compiled into a data frame that maps New York neighborhoods to their geolocations.

2.3 Sourcing Coffee Shop data using Foursquare API

Getting the total number of coffee shops for each neighborhood in Toronto and New York is critical to solving our business problem.

We shall use the Foursquare Venues Search feature (https://api.foursquare.com/v2/venues/search/<parameters>) and use the geolocation of each neighborhood to determine all the coffee shops in that neighborhood, count them and add that number to our city data frame in a new column to capture the number of coffee shops.

Using these data sources coupled with data extracted using the Foursquare API satisfies our basic data needs to solve our business problem.

3. Methodology

3.1 Basic Data Preparation and Cleaning

To begin our analysis for Toronto and New York, we primarily need to prepare our data so we have consistent data frames that map Neighborhoods to their geolocations (Latitude and Longitude.) Since, we also need data on the number of Coffee Shops in each neighborhood, we shall retrieve that data using the Foursquare API. We shall also keep additional neighborhood data on neighborhood Borough and Postal Codes that might provide additional value for our customer (BakerZ.)

- **Toronto**

We do not have a single source of data so we shall use our two sources - the Wikipedia site that provides a table mapping Toronto postal codes to Boroughs and Neighborhoods, and the Toronto geospatial data available as a CSV file (https://cocl.us/Geospatial_data), mapping postal codes to their respective geolocations. We shall scrape the data table from the Wikipedia site into a data frame, download the geolocation data into a separate data frame and

then using the Postal Code as key, we create a single data frame that has the following columns: **Postal Code**, **Borough**, **Neighborhood**, **Latitude**, and **Longitude**. We then iterate over each Neighborhood in the data frame and use its geolocation to get the number coffee shops in that neighborhood (done using Foursquare API Venues Search feature) which is added to a new column called NumCoffeeShops. We use a search radius of 500 meters for each neighborhood and limit the results to 100 for each query. Our final data frame for analysis then consists of the following columns:
Postal Code, **Borough**, **Neighborhood**, **Latitude**, **Longitude,** and **NumCoffeeShops**

- **New York**

We have a single source of data so we shall use our source - a JSON file containing the data and available at - https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json. We can easily convert the dictionary obtained from the JSON file into a data frame that has the following columns: **Postal Code**, **Borough**, **Neighborhood**, **Latitude**, and **Longitude**. Similar to Toronto, we then iterate over each Neighborhood in the data frame and use its geolocation to get the number coffee shops in that neighborhood (done using Foursquare API Venues Search feature) which is added to a new column called NumCoffeeShops. We use a search radius of 500 meters for each neighborhood and limit the results to 100 for each query. Our final data frame for analysis then consists of the following columns:
Postal Code, **Borough**, **Neighborhood**, **Latitude**, **Longitude,** and **NumCoffeeShops**

Our data for Toronto and New York is now present in consistent data frames so we can now proceed with some basic visualization and inferential statistics to get a feel of our data. We shall use Folium for map visualization and render neighborhood markers on top.

Upon inspection, we see that New York has a total of 306 neighborhoods and Toronto has a total of 180 neighborhoods.

For Toronto, we see that some Boroughs and Neighborhoods are unassigned (having a “Not Assigned” value.) Also the spelling of the Neighborhood column is British-style (spelled as Neighbourhood) which is not consistent with out New York data frame.

For cleanup, we rename the column to Neighborhood, assign Borough name to all unassigned Neighborhood names, and finally remove all rows for which Borough is unassigned.

After the cleanup, we see that we are left with only 103 Neighborhoods to consider in Toronto.

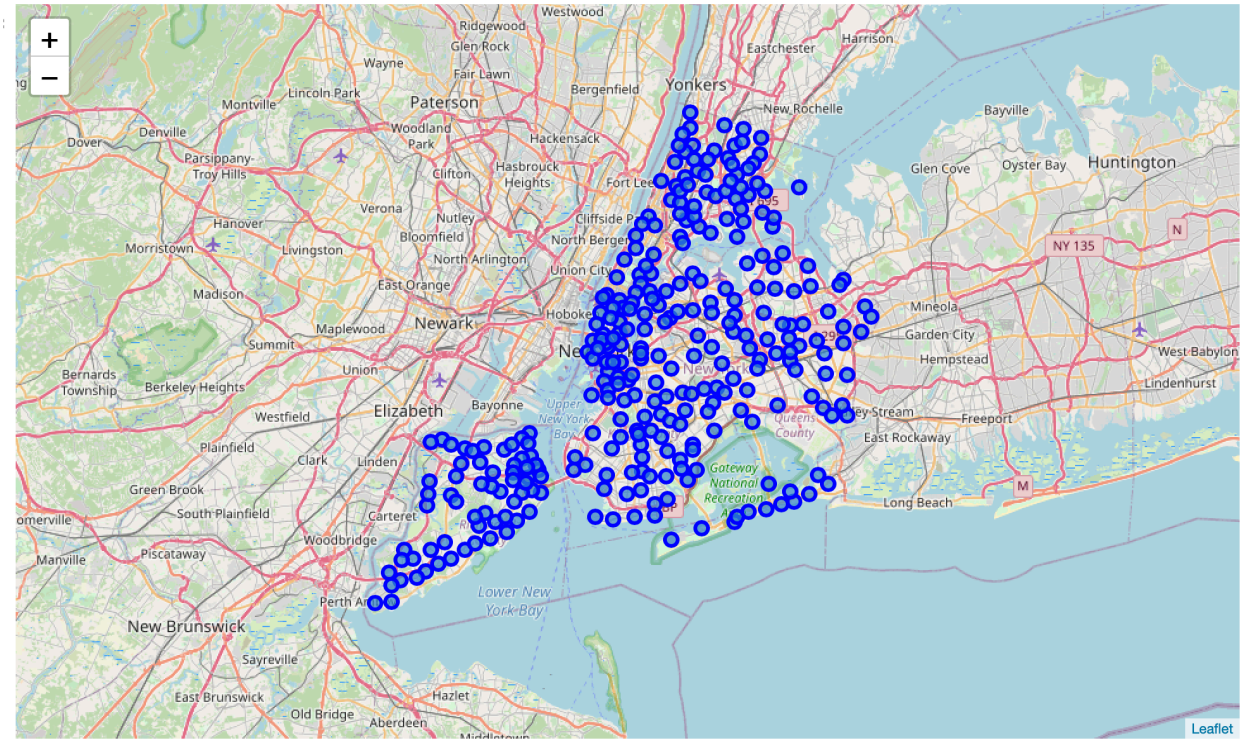
For our analysis, we will consider 180 neighborhoods for New York and 103 neighborhoods for Toronto.

3.2 Visualization and Inferential Statistics

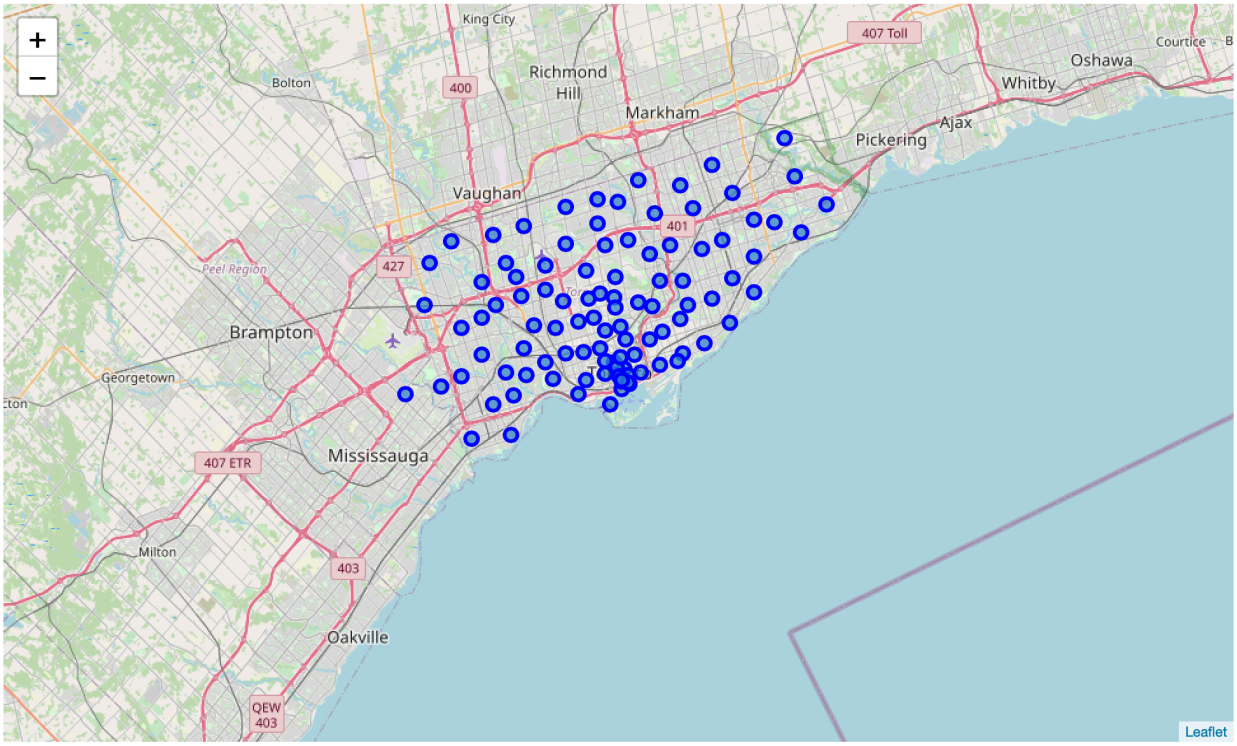
We now use Folium to visualize the New York and Toronto neighborhoods that form the basis for our analysis.

We use Nominatim from geolocator module to determine Latitude and Longitude values for New York and Toronto.

New York Neighborhoods for BakerZ Analysis



Toronto Neighborhoods for BakerZ Analysis



3.3 Retrieving Coffee Shop Data using Foursquare API

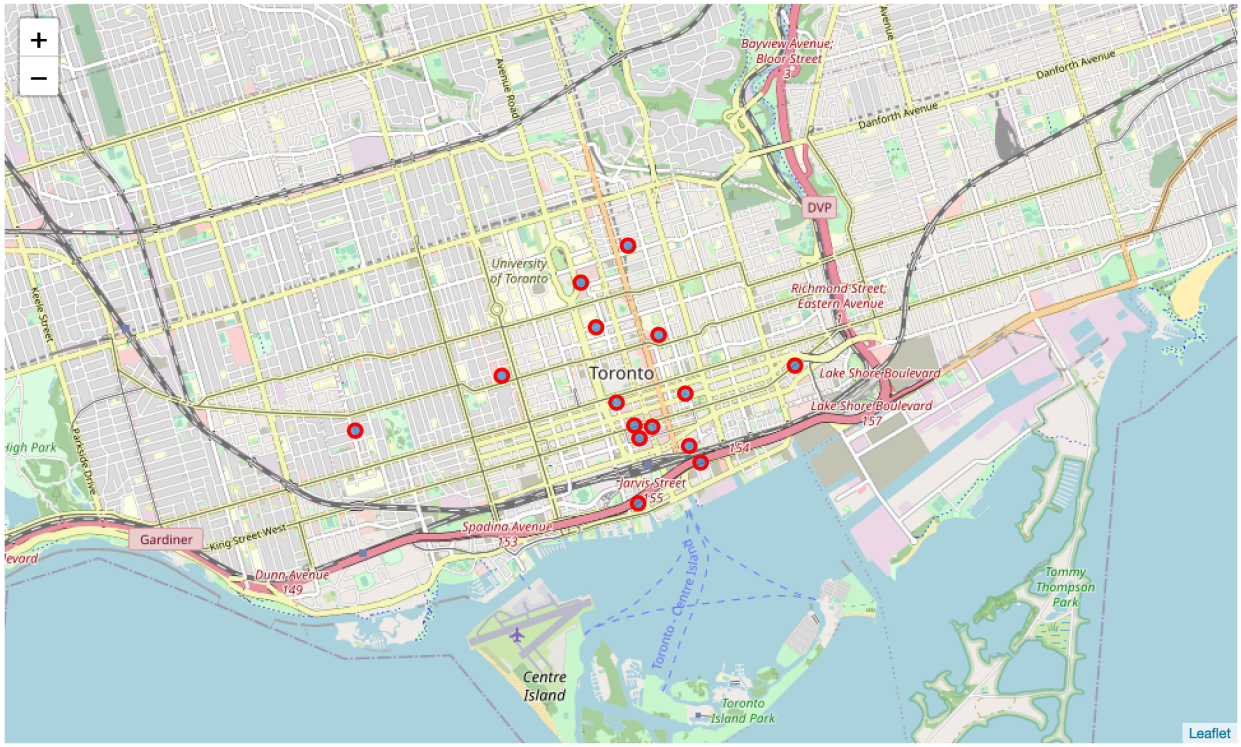
Our data at this time is essentially missing the information on the number of Coffee Shops in each neighborhood. We shall accomplish this by using the Foursquare API, specifically its Venues Search feature. Using the geolocation (Latitude, Longitude) of each neighborhood, we will iterate over the data frames of both New York and Toronto. We shall count the number of coffee shops returned by the API and populate the values in a new column called “NumCoffeeShops”.

BakerZ has requested that they are interested only in neighborhoods that have 10 or more Coffee Shops, so we filter off the neighborhoods that have less than ten Coffee Shops for both New York and Toronto and look at the visualizations again.

From our data, we see that New York has only 38 neighborhoods with 10 or more coffee shops each and Toronto has 15 neighborhoods with the same criterion.

New York Neighborhoods with more than 10 Coffee Shops

Toronto Neighborhoods with more than 10 Coffee Shops



3.4 Application of Machine Learning to Cluster Neighborhoods

We want to establish Neighborhood Clusters that have the highest density of Coffee Shops and are closet to each other. To accomplish this, we use the NumCoffeeShops, Latitude, and Longitude data for each of the New York and Toronto data sets.

We shall use DBSCAN from the sklearn toolkit which provides a form of Unsupervised Machine Learning to create clusters. NumCoffeeShops, Latitude, and Longitude values will form the basis on which the clustering happens to create clusters of Neighborhoods that have the highest number of coffee shops and are closet to each other.

A critical parameter for DBSCAN is the Epsilon parameter whose value is data dependent. We use the NearestNeighbor model from sklearn to create distance data and then plot it. We find the best Epsilon value from the Elbow point of the plot.

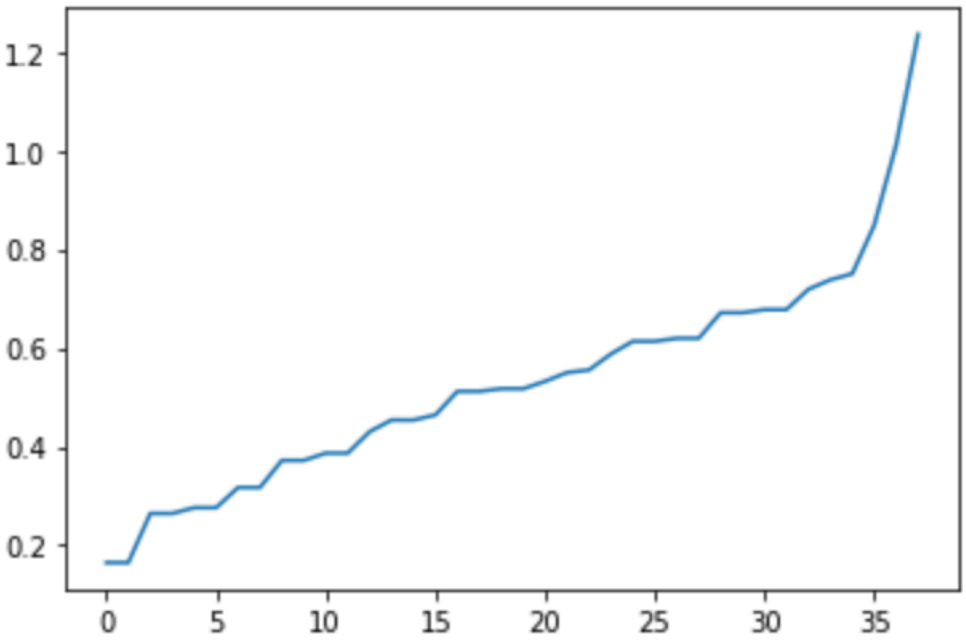
Using NearestNeighbor, we determined the best Epsilon parameter values for New York and Toronto.

Best Epsilon value for New York: 0.75

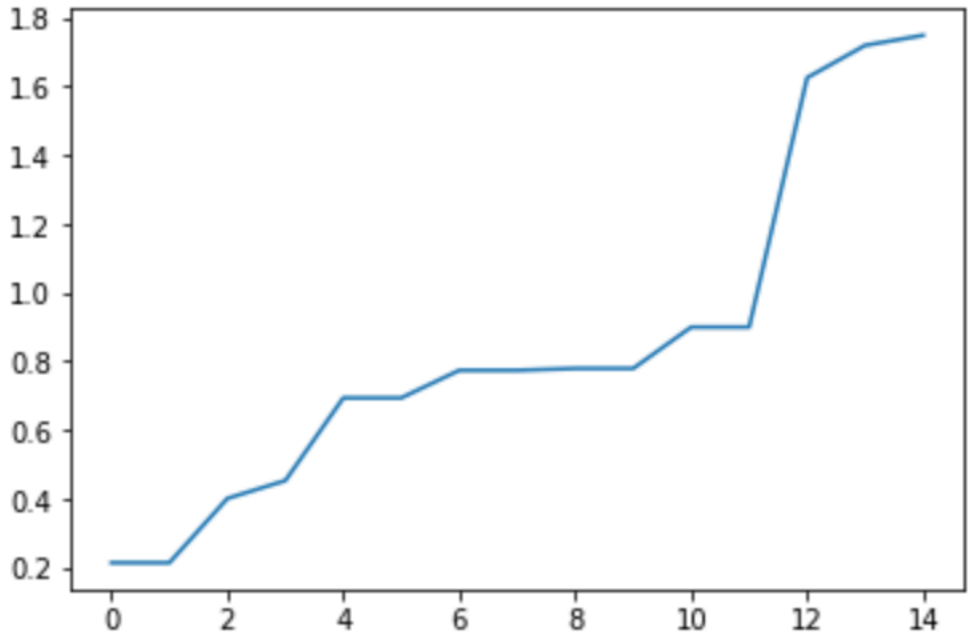
Best Epsilon value for Toronto: 0.90

These were obtained based on the plots shown below after our modeling.

Plot to Determine Best Epsilon value for New York



Plot to Determine Best Epsilon value for Toronto



We also used minimum sample values of 3 for New York (38 neighborhoods) and 2 for Toronto (15 neighborhoods).

Then we run actual DBSCAN for New York and Toronto using their corresponding Epsilon and min_sample parameter values.

Based on this, we obtained the Neighborhood Clusters from DBSCAN and removed the outlier neighborhoods whose labels are -1. These are populated into new data frames for New York and Toronto with an added column to indicate the Cluster label values.

This results in the following clustering:

New York: 22 Neighborhoods grouped into 3 Clusters (Labeled as 0, 1, and 2)

Toronto: 12 Neighborhoods grouped into 5 Clusters (Labeled as 0, 1, 2, 3, and 4)

The resulting data frames are shown below:

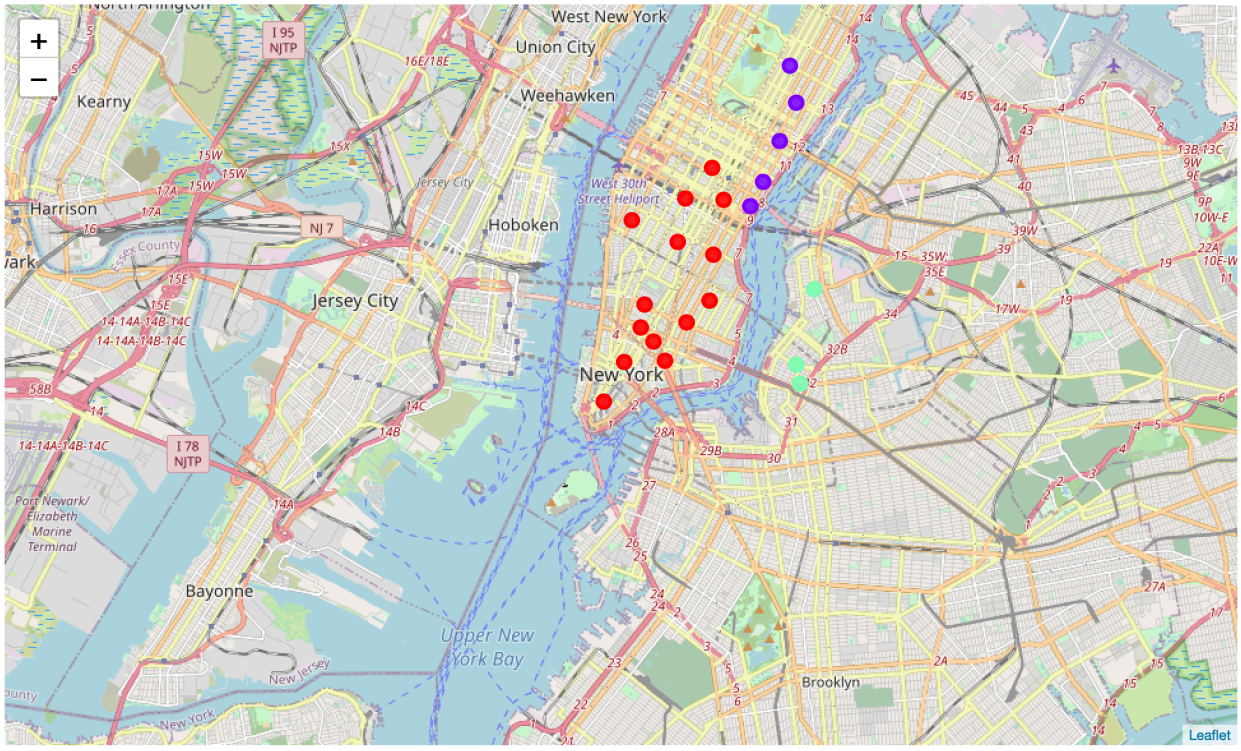
```
Number of Neighborhood Clusters in New York is 3
```

	ClusterID	Borough	Neighborhood	Latitude	Longitude	NumCoffeeShops
0	0	Manhattan	Murray Hill	40.748303	-73.978332	50.0
1	0	Manhattan	Greenwich Village	40.726933	-73.999914	50.0
2	0	Manhattan	Flatiron	40.739673	-73.990947	50.0
3	0	Manhattan	Midtown	40.754691	-73.981669	50.0
4	0	Manhattan	Soho	40.722184	-74.000657	50.0
5	0	Manhattan	Midtown South	40.748510	-73.988713	50.0
6	0	Manhattan	Financial District	40.707107	-74.010665	50.0
7	0	Manhattan	Chelsea	40.744035	-74.003116	48.0
8	0	Manhattan	Civic Center	40.715229	-74.005415	46.0
9	0	Manhattan	Little Italy	40.719324	-73.997305	41.0
10	0	Manhattan	Noho	40.723259	-73.988434	36.0
11	1	Manhattan	Turtle Bay	40.752042	-73.967708	35.0
12	1	Manhattan	Upper East Side	40.775639	-73.960508	32.0
13	1	Manhattan	Lenox Hill	40.768113	-73.958860	31.0
14	1	Manhattan	Tudor City	40.746917	-73.971219	31.0
15	0	Manhattan	Chinatown	40.715618	-73.994279	27.0
16	1	Manhattan	Sutton Place	40.760280	-73.963556	25.0
17	0	Manhattan	East Village	40.727847	-73.982226	24.0
18	0	Manhattan	Gramercy	40.737210	-73.981376	21.0
19	2	Brooklyn	North Side	40.714823	-73.958809	20.0
20	2	Brooklyn	South Side	40.710861	-73.958001	14.0
21	2	Brooklyn	Greenpoint	40.730201	-73.954241	10.0

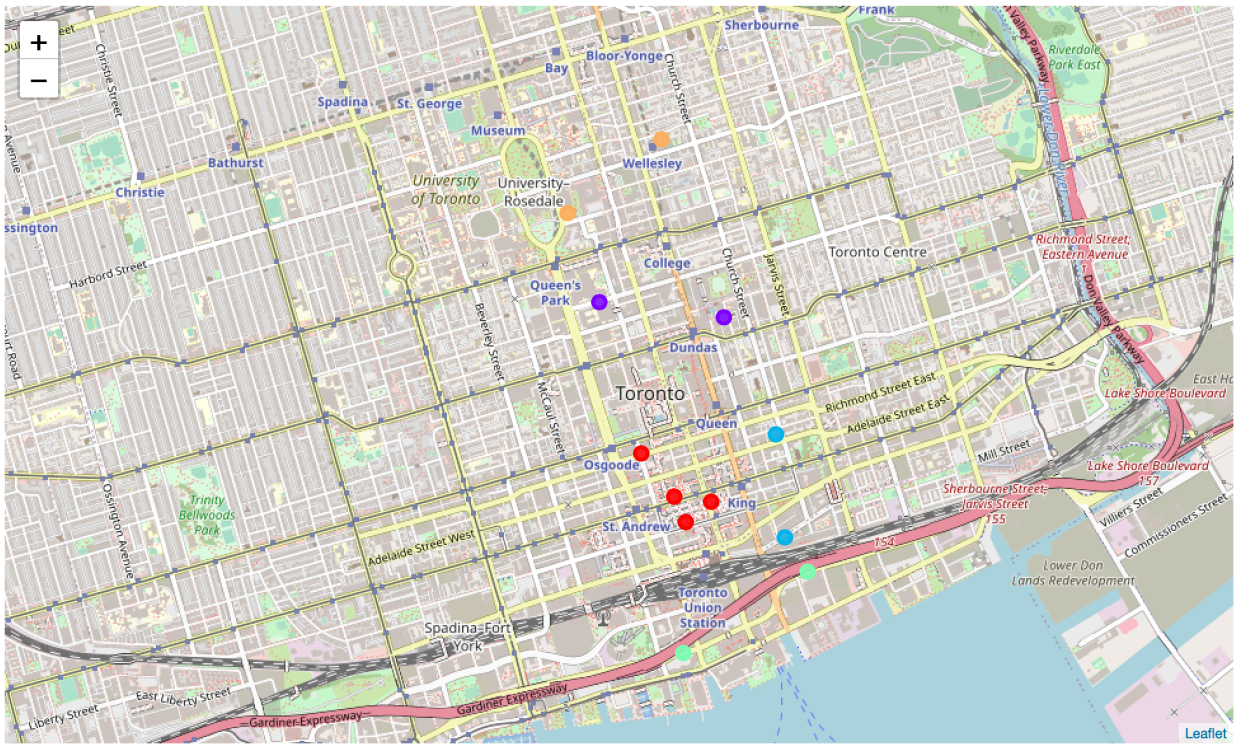
Number of Neighborhood Clusters in Toronto is 5

	ClusterID	Postal Code	Borough	Neighborhood	Latitude	Longitude	NumCoffeeShops
0	0	M5H	Downtown Toronto	Richmond, Adelaide, King	43.650571	-79.384568	43.0
1	0	M5X	Downtown Toronto	First Canadian Place, Underground city	43.648429	-79.382280	40.0
2	0	M5L	Downtown Toronto	Commerce Court, Victoria Hotel	43.648198	-79.379817	39.0
3	0	M5K	Downtown Toronto	Toronto Dominion Centre, Design Exchange	43.647177	-79.381576	35.0
4	1	M5G	Downtown Toronto	Central Bay Street	43.657952	-79.387383	32.0
5	1	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937	30.0
6	2	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418	26.0
7	2	M5W	Downtown Toronto	Stn A PO Boxes	43.646435	-79.374846	26.0
8	3	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306	14.0
9	4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662302	-79.389494	14.0
10	3	M5J	Downtown Toronto	Harbourfront East, Union Station, Toronto Islands	43.640816	-79.381752	13.0
11	4	M4Y	Downtown Toronto	Church and Wellesley	43.665860	-79.383160	11.0

Visualization of Neighborhood Clusters for New York using Folium



Visualization of Neighborhood Clusters for Toronto using Folium



3.5 Compiling New York and Toronto Data into a Single Data Frame to determine Distribution Center Rollout Strategy for BakerZ expansion.

We compile the data we have so far into a new data frame where each row represents:

- A Neighborhood Cluster using the City column (either New York or Toronto)
- Identifies the Neighborhood Cluster with ClusterID of respective city
- The Neighborhood in that Cluster with the highest number of Coffee Shops
- The total number of Coffee Shops in that Cluster

We then sort this data frame on basis of the TotalCoffeeShopsInCluster column values.

Combined New York and Toronto data frame with compiled information (Unsorted)

	City	ClusterID	Neighborhood	NumCoffeeShops	TotalCoffeeShopsInCluster
0	Toronto	0	Richmond, Adelaide, King	43.0	157.0
1	Toronto	1	Central Bay Street	32.0	62.0
2	Toronto	2	St. James Town	26.0	52.0
3	Toronto	2	Stn A PO Boxes	26.0	52.0
4	Toronto	3	Berczy Park	14.0	27.0
5	Toronto	4	Queen's Park, Ontario Provincial Government	14.0	25.0
6	New York	0	Murray Hill	50.0	593.0
7	New York	0	Greenwich Village	50.0	593.0
8	New York	0	Flatiron	50.0	593.0
9	New York	0	Midtown	50.0	593.0
10	New York	0	Soho	50.0	593.0
11	New York	0	Midtown South	50.0	593.0
12	New York	0	Financial District	50.0	593.0
13	New York	1	Turtle Bay	35.0	154.0
14	New York	2	North Side	20.0	44.0

Combined New York and Toronto data frame with compiled information (Sorted)

	City	ClusterID	Neighborhood	NumCoffeeShops	TotalCoffeeShopsInCluster
0	New York	0	Murray Hill	50.0	593.0
1	New York	0	Greenwich Village	50.0	593.0
2	New York	0	Flatiron	50.0	593.0
3	New York	0	Midtown	50.0	593.0
4	New York	0	Soho	50.0	593.0
5	New York	0	Midtown South	50.0	593.0
6	New York	0	Financial District	50.0	593.0
7	Toronto	0	Richmond, Adelaide, King	43.0	157.0
8	New York	1	Turtle Bay	35.0	154.0
9	Toronto	1	Central Bay Street	32.0	62.0
10	Toronto	2	St. James Town	26.0	52.0
11	Toronto	2	Stn A PO Boxes	26.0	52.0
12	New York	2	North Side	20.0	44.0
13	Toronto	3	Berczy Park	14.0	27.0
14	Toronto	4	Queen's Park, Ontario Provincial Government	14.0	25.0

4. Results

Based on our final compiled and sorted data frame representing the combined data from New York and Toronto, we can confidently recommend the following Expansion Strategy to BakerZ to open new Distribution Centers in New York and Toronto.

Plan to open a total of 15 Distribution Centers across New York and Toronto.

To maximize return on investment, the Distribution Centers should be opened in the following order:

	City	ClusterID	Neighborhood	NumCoffeeShops	TotalCoffeeShopsInCluster
0	New York	0	Murray Hill	50.0	593.0
1	New York	0	Greenwich Village	50.0	593.0
2	New York	0	Flatiron	50.0	593.0
3	New York	0	Midtown	50.0	593.0
4	New York	0	Soho	50.0	593.0
5	New York	0	Midtown South	50.0	593.0
6	New York	0	Financial District	50.0	593.0
7	Toronto	0	Richmond, Adelaide, King	43.0	157.0
8	New York	1	Turtle Bay	35.0	154.0
9	Toronto	1	Central Bay Street	32.0	62.0
10	Toronto	2	St. James Town	26.0	52.0
11	Toronto	2	Stn A PO Boxes	26.0	52.0
12	New York	2	North Side	20.0	44.0
13	Toronto	3	Berczy Park	14.0	27.0
14	Toronto	4	Queen's Park, Ontario Provincial Government	14.0	25.0

5. Discussion/Observations

We observe that while both New York and Toronto are large cities, the number of coffee shops in New York are much higher than Toronto. So it definitely makes sense for BakerZ to first invest in and prioritize New York. Based on the data above, we clearly see that BakerZ should open their first 7 Distribution Centers in New York in the neighborhoods mentioned above. This will give them the maximum return on investment as they will be able to cater to the bulk of the coffee shops. After that they can interleave opening of distribution centers in Toronto and New York as per the priority list shown above.

Note: The Foursquare API seems to be capping the number of search results for coffee shops to 50 so that is the maximum number of shops we see in a neighborhood and you may notice that it results in the same number of coffee shops in multiple neighborhoods.

6. Conclusion

We now have a clean expansion strategy to present to BakerZ. Starting with our separate raw data for New York and Toronto, we cleaned and prepared it properly into consistent data frames. We then retrieved addition coffee shop data using Foursquare and joined it to our city data frames that allowed us to analyze and visualize are our data. We then proceeded to utilize Unsupervised Machine Learning (DBSCAN) to create special density based clustering of neighborhoods on the basis of number of coffee shops in each neighborhood and its geolocation. We could then visualize the neighborhood clusters on New York and Toronto maps. Finally, we were able to combine the final data from New York and Toronto into a single data frame by computing all the information we needed to extract business insights and create an expansion strategy for BakerZ. We are confident of our data-driven results and looking forward to presenting them to BakerZ.