Dessert Island

Recommendations for New Restaurant Locations in New York City

Sean Wagner

March 10, 2020

Table of Contents

Executive Summary

Our client has a small chain of dessert restaurants in Toronto, and is looking to expand to New York City. To help identify possible locations in New York City where the business would be successful, data on both cities was examined. The information included data about the locations of various neighbourhoods in Toronto and New York City, and data from Foursquare on the existing venues located within reach of those neighbourhoods. The similarity of neighbourhoods between Toronto and New York City was determined by applying the k-means clustering algorithm to place each neighbourhood into one of seven clusters based on features derived from the mix of venue types found in those neighbourhoods. It was found that the client's successful locations in Toronto were in Cluster 0 and 5, while their unsuccessful location was found in Cluster 6. Cluster 0 neighbourhoods were characterized as having an abundance of Pizza Places. Cluster 5 neighbourhoods were characterized as having a lot of Bars and an overall large number of nearby venues. Pizza Places and Bars may be complimentary to the success of client's dessert cafes in Toronto. Cluster 6 neighbourhoods were found to have an abundance of incumbent Dessert Shops and Cafes, which likely place competitive pressure on the client's business. Hence, it is recommended that the client open new locations in New York City where the neighbourhoods belong to Cluster 0 and/or 5, and avoid Cluster 6 neighbourhoods.

Introduction

The project will examine the problem of deciding where to locate a business. In the case to be examined, a client that already has an established business in one city is working to expand to another city. However, with little knowledge about the other city, it can be difficult to determine the best possible locations such that they would be successful. Here, data on the composition of different areas in two different cities is analyzed to help with this problem.

Business Problem

Our client has a small chain of boutique dessert restaurants in different neighbourhoods in the City of Toronto, Canada. The client is looking to expand to other cities, the first being New York City in the United States. Of the three existing restaurant locations in Toronto, two are very successful having a large number of customers, while one location has been less successful. These cafes and their neighbourhood locations are listed in Table 1. It is not clear why there is a difference in success, but the client would like to open new locations in New York that would have the best chance of attracting a large number of customers. One possible strategy is to locate new restaurants in neighbourhoods in New York that are similar to the neighbourhoods in Toronto in which the already successful restaurants are located. Our goal is to recommend appropriate neighbourhoods in New York City by examining data about New York City and Toronto.

Location ID	**Neighbourhood**	**Success level**
1	Annex	High
2	Roncesvales	High
3	Woodbine Corridor	Low

Table 1: Client's current locations in Toronto

Data

Data Sources

Solving the business problem at hand will require information about the characteristics of neighbourhoods in both Toronto and New York City. There are two main sets of data for Toronto and New York City that were examined to help solve the business problem: 1) listings of neighbourhoods, and 2) listing of existing venues. For both data sets, spatial coordinates were provided as latitude and longitude.

According to the City of Toronto Web site, Toronto has 140 city-designated neighbourhoods [1]. Neighbourhood listings with coordinates are provided for Toronto in .csv format from the following link:

https://open.toronto.ca/dataset/neighbourhoods/

For New York City, location data on neighbourhoods is provided by New York University (NYU) at the at the following link in JSON format:

https://geo.nyu.edu/download/file/nyu-2451-34572-geojson.json

In each neighbourhood location data set, spatial coordinates are provided as latitude and longitude.

Data on venues such as restaurants, shops, and sight-seeing attractions were obtained from Foursquare (https://foursquare.com/). In particular, the `explore` function in Foursquare's RestAPI provides listings of venues within a specified radius of given latitude and longitude coordinates. Each venue is listed with a category that describes the type of venue (e.g. "Japanese Restaurant", "Movie Theatre", "Park", etc.). Those categories and their hierarchical organization are described by Foursquare's documentation here:

https://developer.foursquare.com/docs/resources/categories

Data Preparation

Neighbourhood Location Data

Data on the names and locations of neighbourhoods from the Toronto Open Data portal was well structured and clean. Each row represented one neighbourhood. While there were several fields contained in the data set, the only fields kept were the those listed in Table 2.

Field	**Description**
AREA_NAME	Neighbourhood name
LONGITUDE	Longitude coordinate
LATITUDE	Latitude coordinate

Table 2: Toronto neighbourhood location data set fields used

The locations of Toronto neighbourhoods are plotted in Figure 1. There were 140 different neighbourhoods for Toronto

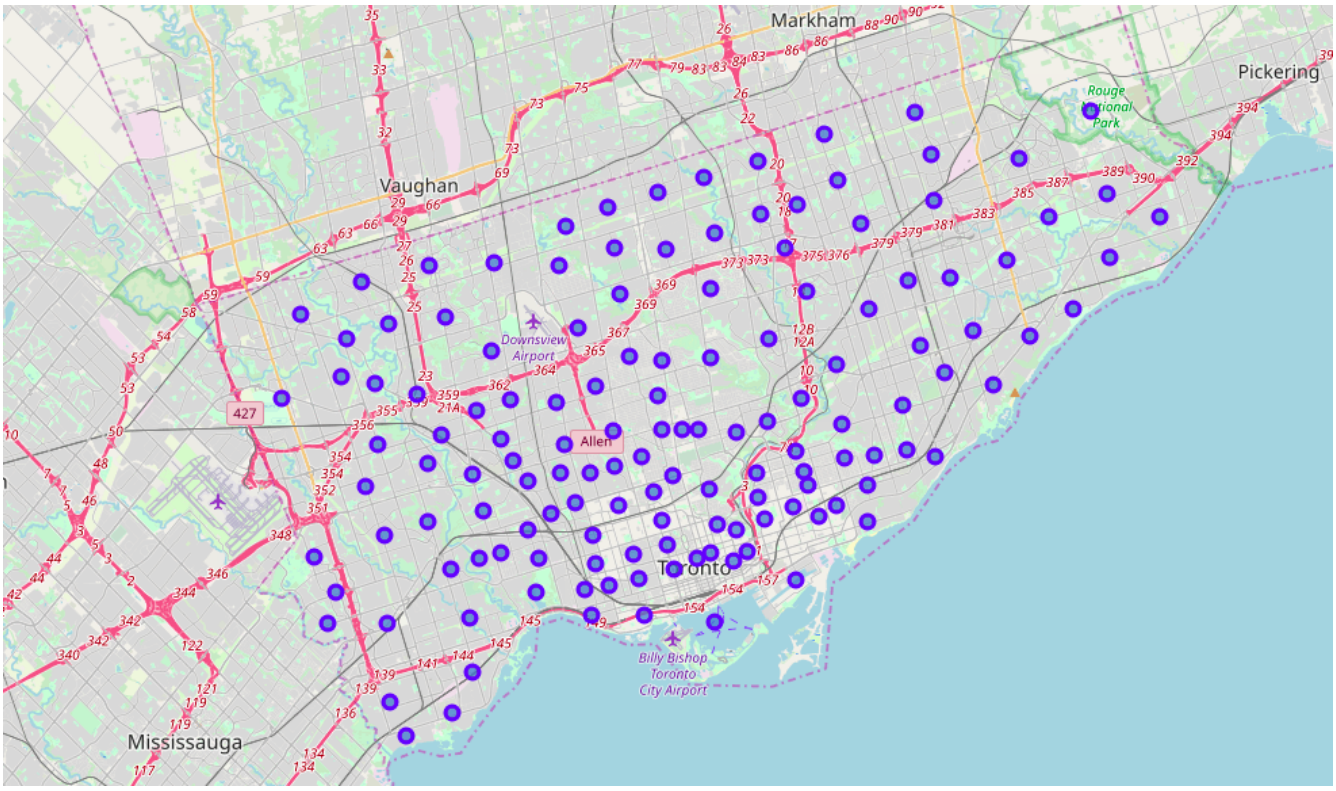

Figure 1: Map of Toronto neighbourhood locations.

For New York City, the data from NYU was also well structured in JSON format and required no cleaning. However, the JSON format was adapted to a Pandas DataFrame to allow for analysis. While there were several data fields for each neighbourhood listed, only the fields listed in Table 3 were retained.

Field	**Description**
borough	Borough name
name	Neighbourhood name
coordinates[0]	Longitude coordinate
coordinates[1]	Latitude coordinate

Table 3: New York City neighbourhood location data fields kept

There were 306 neighbourhoods in the New York City data set. A map of New York City neighbourhood locations is shown in Figure 2.

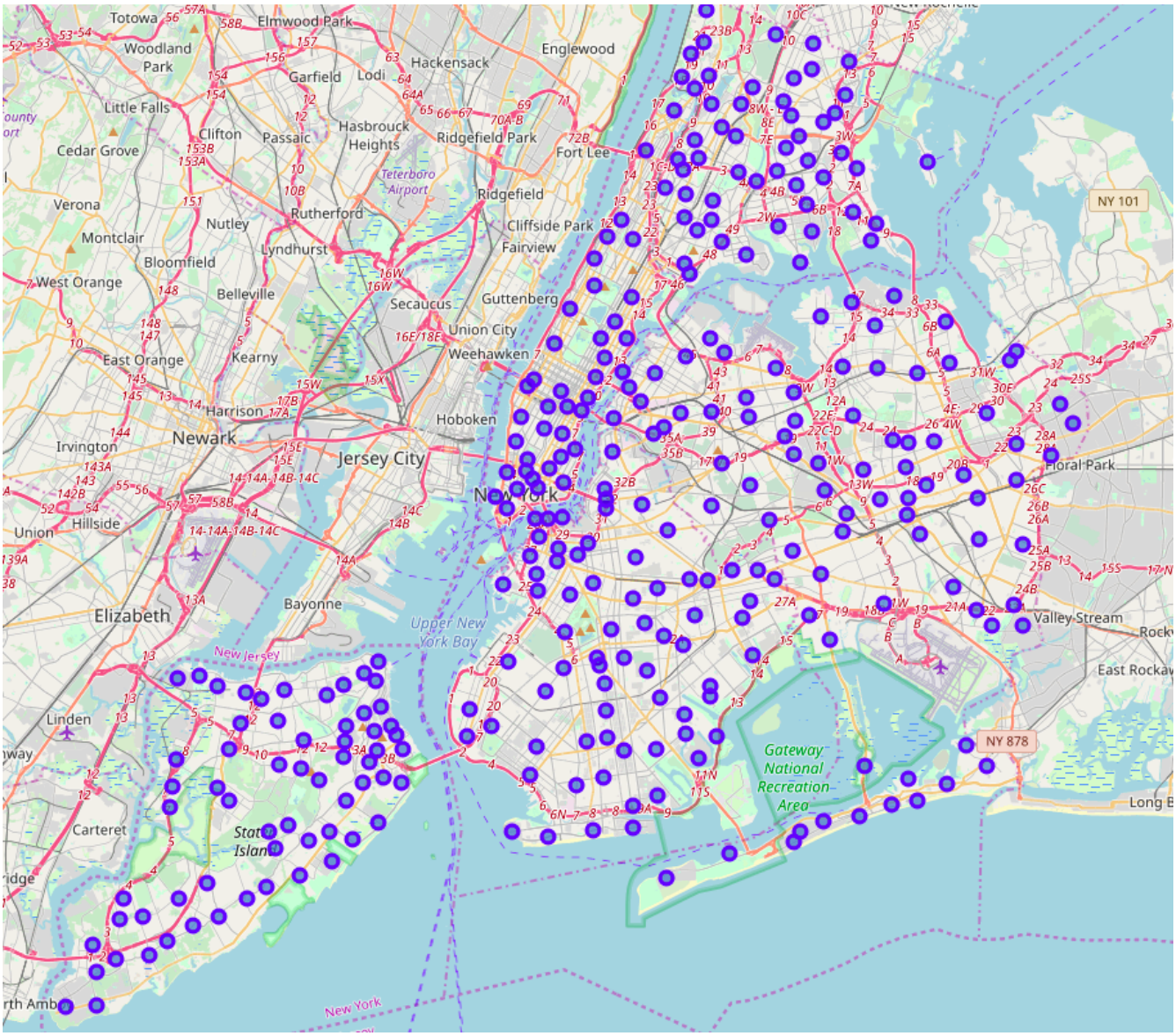

Figure 2: Map of New York City neighbourhood locations

Neighbourhood Venues Data

Data on venues within each neighbour were obtained via the Foursquare by calling the `explore` RestAPI function for each neighbourhood in both Toronto and New York City. The latitude/longitude of each neighbourhood was used as the center point for each query, and venues within 500 m of the center point were requested. Note that some neighbourhoods are separated by less than 500 m, hence there was some overlap in the returned venues. While it could be said that a particular venue belongs to one neighbourhood and not another, the aim of the analysis is to characterize neighbourhoods based on the venues that are within quick reach of a neighbourhood, not based on what is contained only within their boundaries.

Data returned from the `explore` RestAPI function is in JSON format. The following fields for each venue were extracted:

Field	Description
name	Venue name
location.lat	Latitude coordinate of venue
location.lng	Longitude coordinate of venue
categories[0].name	Venue category (first entry)

Table 4: Foursquare venue data fields used

Overall, there were 2102 venues found for Toronto, and 10301 venues for New York City.

The key features used in further analyses are the venue categories. For each venue, only one category was returned for each venue and was listed as the primary category. Between both Toronto and New York City, the venues returned belonged to one of 462 unique categories. In early analysis, this large number of categories proved to be problematic when attempting to cluster the neighbourhoods using the k-means algorithm. It was determined that using less categories as the feature set could improve the clustering algorithm and the overall data analysis. Upon further inspection of the categories, it was found that many were too specific. For instance, a venue categorized as a "Cantonese Restaurant" is just a more specific type of Chinese restaurant. This high level of specificity resulted in many venue categories having few members. Hence, the feature set of the data was very sparse. Instead of using the listed primary venue categories as the features, it was determined that generalizing the categories would allow for better grouping of the data into fewer features that were more usable.

Foursquare lists a hierarchy of categories that allows for generalizing the categories for each venue. It can be obtained through the Foursqaure `category` RestAPI function. With this hierarchy, the parent categories for each venue's listed category were found. For this analysis, the parent category used was the second category level down from the top level. For instance, the category "Cantonese Restaurant" was mapped to a parent category of "Asian Restaurant", which is one level down from the top level category "Food". After adding the parent categories to each venue entry in the data set, the number of categories to deal with was reduced to 299.

Methodology

The goal of the analysis is to find neighbourhood similarities between Toronto and New York City in order to identify where to open new locations of the client's dessert cafes. One such way to accomplish this is to use clustering algorithms on the neighbourhoods based on the venues that within their reach. The k-means clustering algorithm was chosen as the clustering algorithm. Venue parent categories were chosen as the feature set.

First, the data was prepared for the clustering algorithm. The neighbourhood venue data of both Toronto and New York City was combined into a single data frame. One-hot encoding was used to mark each venue's parent category. For each neighbourhood, the number of venues belonging to each parent category was calculated as a percentage of the total number of venues in that neighbourhood. The resulting data frame listed the proportion of each venue parent category in each neighbourhood.

Next, the category proportions were extracted as the feature set for the k-means algorithm. Features were scaled prior to k-means clustering by using the `MinMaxScaler` in SciKit-learn. The number of clusters was chosen to be 7. The `KMeans` module from SciKit-learn library in Python was used.

Results

Results from the k-means clustering process were analyzed to gather insights about the various neighbourhood clusters. Table 5 list the number of neighbourhoods in each cluster for each city, and their percentages of the total number of neighbourhoods in each city. The data is summarized in Figure 3. Note that New York City has more than twice the amount of neighbourhoods than Toronto, and hence the number of neighbourhoods in each cluster is larger for New York City in 5 out of the 7 clusters. Cluster 1 contains the largest number of neighbourhoods, followed by Cluster 2. Cluster 3 has only one neighbourhood, which is in New York City.

Cluster	**Toronto**			**New York City**		
	# Neighbourhoods		**Avg. # venues**	**# Neighbourhoods**		**Avg. # venues**
0	25	17.9%	10.1	49	16.05	25.3
1	72	51.4%	18.8	175	57.2%	31.4
2	5	3.57%	12.4	31	10.1%	17.8
3	0	0%	0	1	0.3%	5.0
4	23	16.4%	3.4	5	1.6%	3.6
5	6	4.23%	48.7	35	11.4%	83.2
6	6	4.35	10.3	4	1.3%	20.3

Table 5: Number of neighbourhoods in each cluster

The number of venues in each neighbourhood can also help to characterize the clusters. Figure 4 shows the average number of venues per neighbourhood in each cluster. The following observations were made:

1. Cluster 5 neighbourhoods by far have the most number of venues per neighbourhood.
2. Cluster 4 has the least number of venues on average.

3. In all cases, New York City neighbourhoods have a larger number of venues on average in all clusters than in Toronto neighbourhoods.

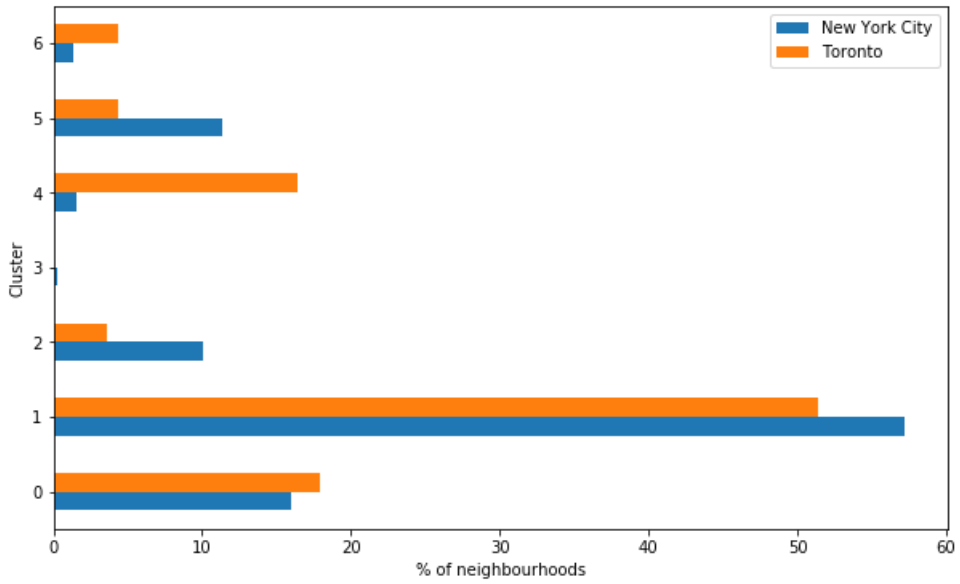

Figure 3: Percentages of neighbourhoods in each cluster for each city

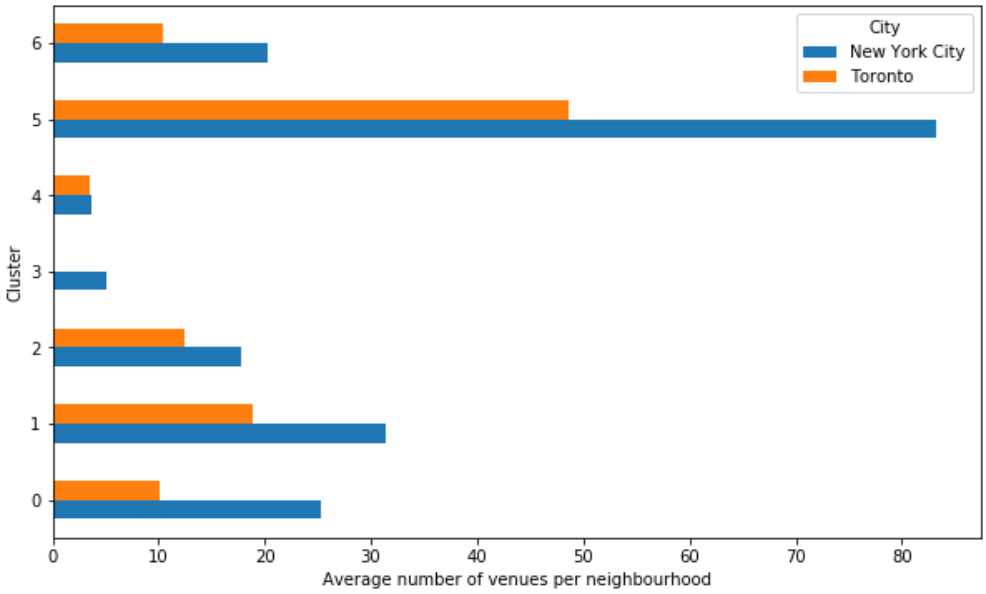

Figure 4: Average number of venues per neighbourhood in each cluster

Mapping the clusters geographically provides additional insights. Figure 5 shows the neighbourhood locations in Toronto coloured by their corresponding cluster, while Figure 6 shows the same information for New York City. The following general observations on the geographic distribution of the cluster were made:

1. In both Toronto and New York City, Cluster 5 neighbourhoods tend to be close to the city centre.
2. Cluster 0, 2 and 6 neighbourhoods tend to be outside of the city centre for both cities.
3. Neighbourhoods in Cluster 1 are distributed all around both cities.

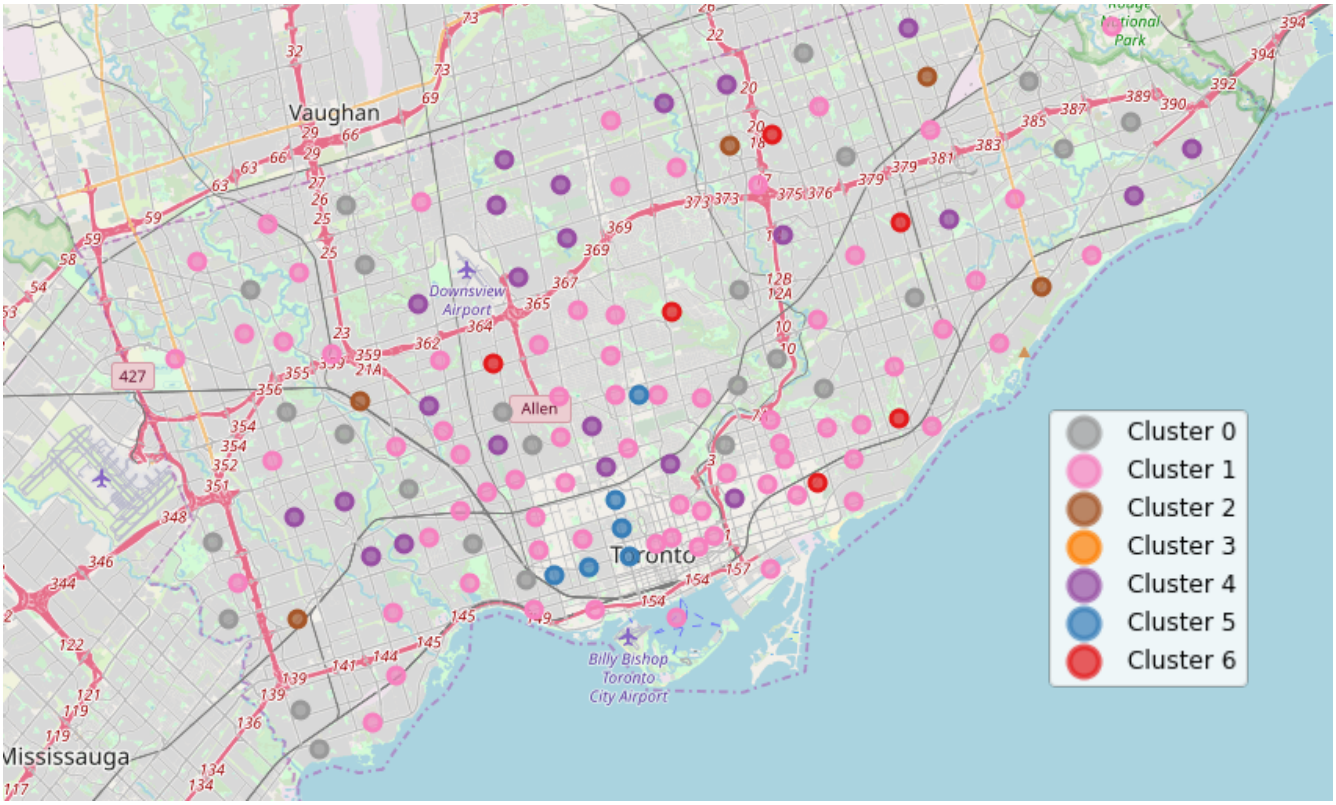

Figure 5: Map of neighbourhood clusters in Toronto

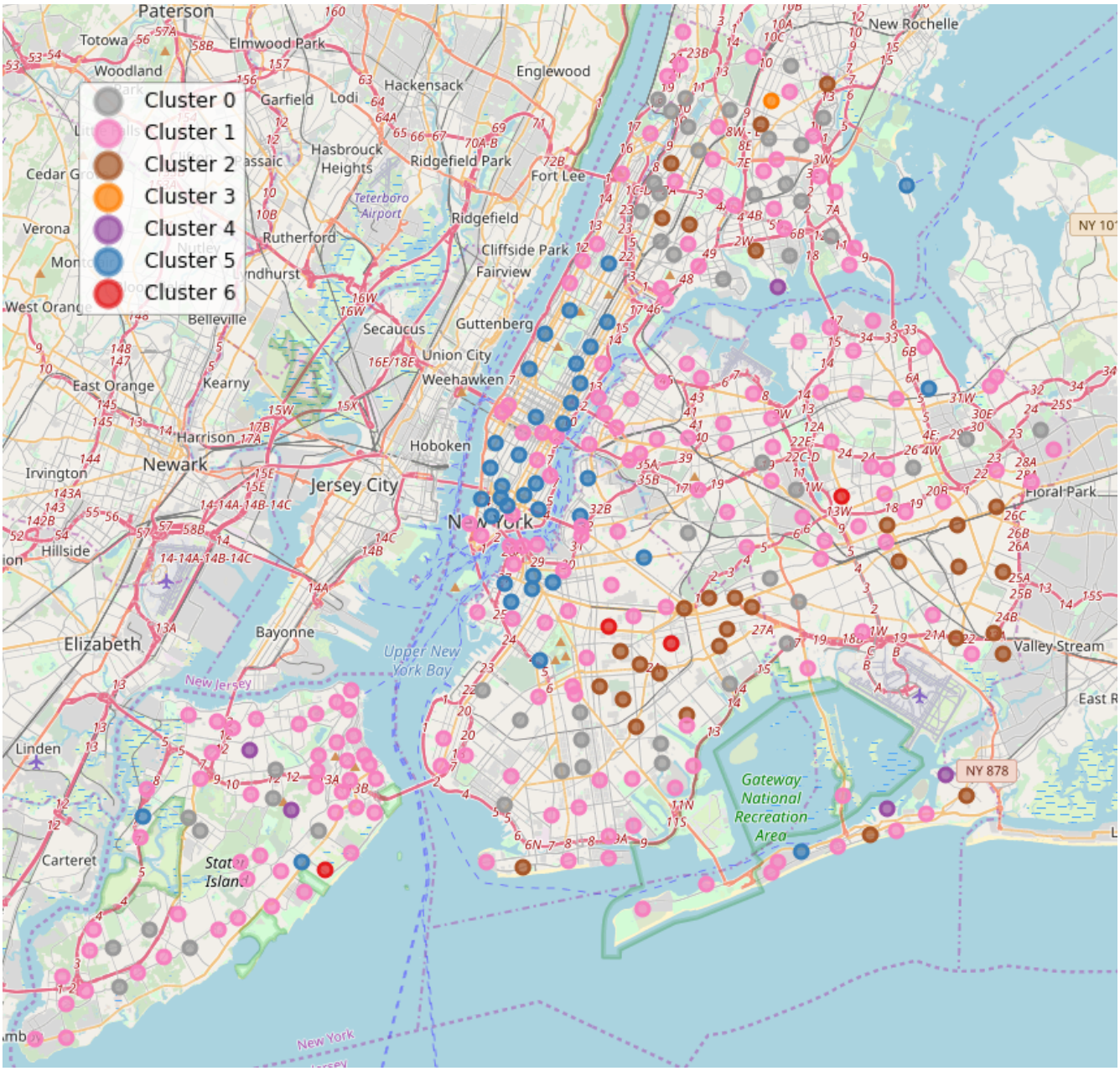

Figure 6: Map of neighbourhood clusters in New York City

The types of venues and their abundance in each cluster was also examined as a means to characterize them. Venue categories in each cluster were ranked by the average of their percentage make-up of the total venue count in each neighbourhood. Figure 7 list the top 5 ranked venue types in each cluster. The following general observations were made:

1. Five of the seven clusters have Asian Restaurants in their top five venue types, with cluster 1 having the highest proportion compared to other venues types in that cluster.
2. Athletics & Sports venues show up in the top 5 in five out of the seven clusters.

The following are specific observations about each cluster:

Cluster 0: Pizza Places are the number 1 type of venue in these neighbourhoods with an average of over 10% of the venues in each neighbourhood.

Cluster 1: This cluster has the largest number of neighbourhoods. Asian restaurants led in this cluster by a slight margin. The overall proportion of the top 5 categories is lower than in other clusters, suggesting that the top 5 venue categories in Cluster 1 do not dominate and that there is a more balanced mix of other venue types in these neighbourhoods.

Cluster 2: Dominated by Food and Drink Shops (e.g. grocery stores, wine stores, etc.). Carribean Restaurants are the dominant restaurant type.

Cluster 3: As discovered previously, Cluster 3 has only one neighbourhood in it, Williamsbridge. There were only 5 venues in this neighbourhood, all of different types.

Cluster 4: Parks are very dominant in this cluster. There are no food or drink venues in this cluster's top 5. Contrary to the trend in other clusters, Toronto has more neighbourhoods in Cluster 4 than does New York City.

Cluster 5: This cluster shares a similar venue mix in it's top 5 venue types as in Cluster 1. However, the proportions of each venue type is in a different order with bars being the lead venue type. Also, the overall proportions of the top 5 are larger than in Cluster 1, suggesting that the proportions of other venues type below the top 5 are lower.

Cluster 6: Four of the five venue categories in Cluster 6 are food and drink venues. The generic "Restaurant" venue category dominates, which is somewhat ambiguous and not specific enough to draw good conclusions about the mix of restaurant types. This cluster also contains the Dessert Shop category and Cafe category in it’s top 5.

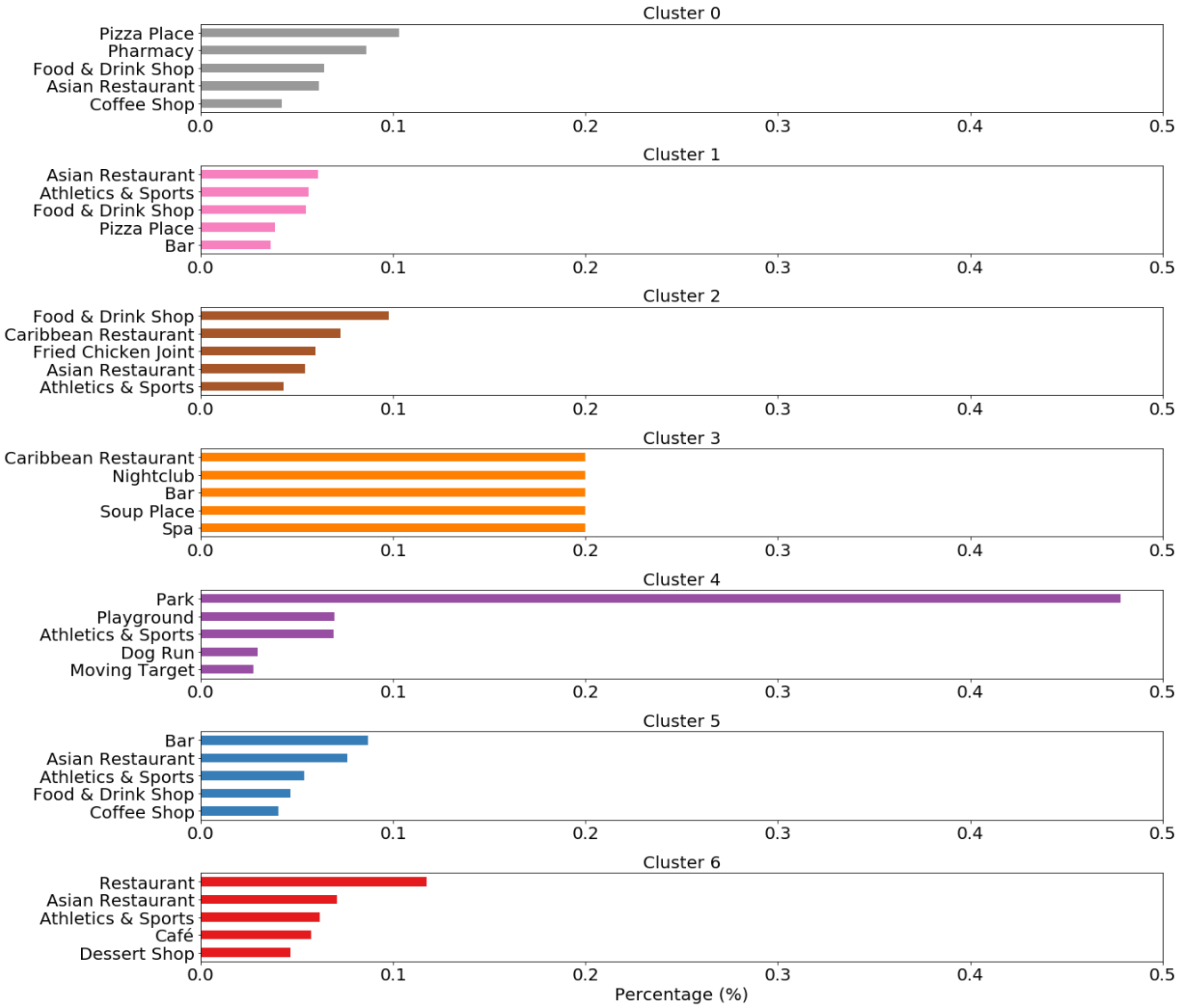

Figure 7: Top 5 venue categories in each cluster by average percentage

In the analysis, it was found that the Coffee Shops show up in the top 5 in Cluster 0 and 5, while Cafes show up in the top 5 venue type for Cluster 6. These may be synonymous terms, hence ambiguous. To determine if there is a distinction, statistics on the abundance of Coffee Shops and Cafes were examined. Table 6 shows the counts for the top five Coffee Shops by their name, while Table 7 shows the same information for Cafes. Coffee Shops are dominated in count by two companies, Starbucks and Tim Hortons, which are large national chains focused on quick service food and drinks. By contrast, venues categorized as Cafes are not dominated by any national chain as they all tend to have far fewer locations each. It's likely that Cafes are more akin to boutique venues.

Name	Count	Percentage (%)
Starbucks	78	32.5
Tim Hortons	62	25.8
Blue Bottle Coffee	10	4.2
Second Cup	9	3.8
Hungry Ghost	6	2.5

Table 6: Statistics on Coffee Shops in Toronto and New York

Name	Count	Percentage (%)
Bluestone Lane	7	3.1
Aroma Espresso Bar	4	1.8
Gotan	4	1.8
Maman	3	1.3
Thobors Boulangerie Patisserie Café	2	0.9

Table 7: Statistics on Cafes in Toronto and New York

Discussion

The analysis detailed in the previous section characterized the neighbourhoods according to their makeup by the venues that are within reach of each neighbourhood. The observations made suggest that there are commonalities between neighbourhoods in Toronto and New York City such that recommendations on potential new locations for our client's business can be made.

The client's current successful locations' neighbourhoods in Toronto were in the Annex and Roncesvalles neighbourhoods, which were were placed in Cluster 5 and in Cluster 0, respectively. The unsuccessful location's neighbourhood was in Woodbine Corridor, a Cluster 6 neighbourhood. These neighbourhoods are somewhat close to the city centre of Toronto, so geographic location alone is not sufficient to draw any conclusions.

The characteristics of the venue mix in each neighbourhood cluster provides better insights. In all three clusters where the client has current locations in Toronto, all have Asian Restaurants in their top 5 in similar proportions, so this is not likely a determining factor for success. Cluster 0 neighbourhoods have a large number of Pizza Places. It's reasonable to assume that they may draw clientele at their Roncesvalles location from Pizza Places, and hence these may be complimentary businesses. Cluster 5 neighbourhoods have an abundance of Bars, which may too be complimentary to dessert cafes and hence a good location. This cluster is also characterized by having a much larger average total count of venues in both cities, which is likely due to these neighbourhoods being located in or near the city centres. Cluster 6 neighbourhoods have Dessert Shops in their top 5 venue types. This suggests that

these shops in these neighbourhoods present competition with each other and divide the market. These may not be good neighbourhoods in which to set up a new location for the client given the presence of similar incumbents.

Clusters 0 and 5 were identified has having an abundance of existing Coffee Shops. While these may present some competition to prospective locations of the client's dessert cafes, venues in the Coffee Shop category were identified as being quick serve restaurants instead of boutique cafes like the client's business. Nevertheless, existing Toronto locations of the client's business in those clusters with an abundance of Coffee Shops are successful. However, Cluster 6 has both Cafes and Dessert Shops in it's top 5. Cafes were identified as potential boutique-type businesses, which could place additional competitive pressure in that neighbourhood for the client's new dessert cafe locations.

Cluster 1 neighbourhoods were identified as being similar to Cluster 5 neighbourhoods in venue mix. However, given the relative abundance of neighbourhoods of this type, more information may be needed in order to determine which specific neighbourhoods in Cluster 1 would be most appropriate for the client's new locations in New York City.

Recommendations

Based on the findings of the analysis, it is recommended that the client start build new locations in New York City that fall into either Cluster 0 or Cluster 5. This is based on the success of current locations in Toronto that are present in neighbourhoods belonging to those two clusters. Cluster 6 should be avoided in New York City because of the lack of success of the client's location in a Toronto neighbourhood belonging to Cluster 6 and because of the identified competitive characteristics identified in neighbourhoods belonging to that cluster. New York City neighbourhoods are listed with their assigned cluster in the Appendix for the client's reference.

Conclusions

The client is interested in opening locations of their dessert cafes in New York City, but required more information to select neighbourhoods where it's business is likely to be successful. Based on the success levels of the client's locations in Toronto, the analysis presented in this report looked at relevant data on both cities to uncover similarities in the make up of neighbourhoods in each city. This was done by clustering the neighbourhoods using the k-means algorithm on information about existing venues in each city derived from Foursquare's data. Based on this analysis, it is recommended that the client open new locations in New York City in neighbourhoods that belong to Cluster 0 and Cluster 5, while avoiding neighbourhoods in Cluster 6.

Appendix: List of New York Neighbourhoods

Neighbourhood	Cluster
Allerton	0
Arden Heights	0
Bath Beach	0
Borough Park	0
Castle Hill	0
Castleton Corners	0
Co-op City	0
Concourse Village	0
Dongan Hills	0
Eltingville	0
Flatbush	0
Fordham	0
Forest Hills	0
Fresh Meadows	0
Georgetown	0
Glen Oaks	0
Heartland Village	0
High Bridge	0
Hunts Point	0
Kingsbridge	0
Kingsbridge Heights	0
Lefrak City	0
Lindenwood	0
Manhattan Terrace	0
Manor Heights	0
Marble Hill	0
Melrose	0
Midwood	0
Mill Basin	0

Neighbourhood	Cluster
Morris Heights	0
Morris Park	0
Mount Eden	0
New Springville	0
Norwood	0
Oakland Gardens	0
Ocean Parkway	0
Ozone Park	0
Pelham Gardens	0
Prince's Bay	0
Ridgewood	0
Schuylerville	0
Spuyten Duyvil	0
Starrett City	0
Sunset Park	0
Van Nest	0
Wakefield	0
Westchester Square	0
Woodhaven	0
Woodrow	0
Annadale	1
Arlington	1
Arrochar	1
Arverne	1
Astoria	1
Astoria Heights	1
Auburndale	1
Battery Park City	1
Bay Ridge	1
Bay Terrace	1
Baychester	1
Bedford Park	1

Neighbourhood	Cluster
Bedford Stuyvesant	1
Beechhurst	1
Bellaire	1
Belle Harbor	1
Bellerose	1
Belmont	1
Bensonhurst	1
Bergen Beach	1
Blissville	1
Bloomfield	1
Breezy Point	1
Briarwood	1
Brighton Beach	1
Broad Channel	1
Bronxdale	1
Brooklyn Heights	1
Brookville	1
Bulls Head	1
Butler Manor	1
Charleston	1
Chinatown	1
Clifton	1
Clinton	1
Clinton Hill	1
College Point	1
Concord	1
Corona	1
Country Club	1
Ditmas Park	1
Douglaston	1
Dumbo	1
Dyker Heights	1

Neighbourhood	Cluster
East Elmhurst	1
East Tremont	1
East Williamsburg	1
Edenwald	1
Edgemere	1
Edgewater Park	1
Egbertville	1
Elm Park	1
Elmhurst	1
Emerson Hill	1
Fieldston	1
Financial District	1
Floral Park	1
Flushing	1
Forest Hills Gardens	1
Fort Hamilton	1
Fox Hills	1
Fulton Ferry	1
Gerritsen Beach	1
Glendale	1
Gowanus	1
Gramercy	1
Graniteville	1
Grasmere	1
Gravesend	1
Great Kills	1
Greenridge	1
Grymes Hill	1
Hamilton Heights	1
Hillcrest	1
Holliswood	1
Homecrest	1

Neighbourhood	Cluster
Howard Beach	1
Hudson Yards	1
Huguenot	1
Hunters Point	1
Inwood	1
Jackson Heights	1
Jamaica Center	1
Jamaica Estates	1
Kensington	1
Kew Gardens	1
Lighthouse Hill	1
Little Neck	1
Long Island City	1
Longwood	1
Madison	1
Malba	1
Manhattan Beach	1
Manhattanville	1
Marine Park	1
Mariner's Harbor	1
Maspeth	1
Middle Village	1
Midtown South	1
Mill Island	1
Morningside Heights	1
Morrisania	1
Mott Haven	1
Mount Hope	1
Murray Hill	1
Neponsit	1
New Brighton	1
New Dorp	1

Neighbourhood	Cluster
New Dorp Beach	1
North Corona	1
North Riverdale	1
Oakwood	1
Ocean Hill	1
Old Town	1
Paerdegat Basin	1
Park Hill	1
Park Slope	1
Parkchester	1
Pelham Bay	1
Pelham Parkway	1
Pleasant Plains	1
Pomonok	1
Port Morris	1
Port Richmond	1
Prospect Heights	1
Prospect Lefferts Gardens	1
Prospect Park South	1
Queensboro Hill	1
Queensbridge	1
Randall Manor	1
Ravenswood	1
Red Hook	1
Rego Park	1
Richmond Hill	1
Richmond Town	1
Richmond Valley	1
Riverdale	1
Rochdale	1
Rockaway Beach	1
Roosevelt Island	1

Neighbourhood	Cluster
Rosebank	1
Rossville	1
Roxbury	1
Sandy Ground	1
Sea Gate	1
Sheepshead Bay	1
Shore Acres	1
Silver Lake	1
South Beach	1
South Ozone Park	1
South Side	1
St. George	1
Stapleton	1
Steinway	1
Stuyvesant Town	1
Sunnyside	1
Sunnyside Gardens	1
Throgs Neck	1
Tompkinsville	1
Tottenville	1
Travis	1
Tudor City	1
Unionport	1
Utopia	1
Vinegar Hill	1
Washington Heights	1
Weeksville	1
West Brighton	1
West Farms	1
Whitestone	1
Williamsburg	1
Willowbrook	1

Neighbourhood	Cluster
Woodlawn	1
Woodside	1
Yorkville	1
Broadway Junction	2
Cambria Heights	2
Canarsie	2
City Line	2
Claremont Village	2
Concourse	2
Coney Island	2
Cypress Hills	2
East Flatbush	2
East New York	2
Eastchester	2
Erasmus	2
Far Rockaway	2
Flatlands	2
Hammels	2
Highland Park	2
Hollis	2
Jamaica Hills	2
Laurelton	2
New Lots	2
Olinville	2
Queens Village	2
Remsen Village	2
Rosedale	2
Rugby	2
Soundview	2
South Jamaica	2
Springfield Gardens	2
St. Albans	2

Neighbourhood	Cluster
University Heights	2
Wingate	2
Williamsbridge	3
Bayswater	4
Clason Point	4
Somerville	4
Todt Hill	4
Westerleigh	4
Bayside	5
Boerum Hill	5
Bushwick	5
Carnegie Hill	5
Carroll Gardens	5
Central Harlem	5
Chelsea	5
City Island	5
Civic Center	5
Cobble Hill	5
Downtown	5
East Harlem	5
East Village	5
Flatiron	5
Fort Greene	5
Grant City	5
Greenpoint	5
Greenwich Village	5
Lenox Hill	5
Lincoln Square	5
Little Italy	5
Lower East Side	5
Manhattan Valley	5
Midtown	5

Neighbourhood	Cluster
Noho	5
North Side	5
Rockaway Park	5
Soho	5
Sutton Place	5
Tribeca	5
Turtle Bay	5
Upper East Side	5
Upper West Side	5
West Village	5
Windsor Terrace	5
Brownsville	6
Crown Heights	6
Kew Gardens Hills	6
Midland Beach	6

List of References

[1] “Neighbourhood Profiles,” [URL] https://www.toronto.ca/city-government/data-research-maps/neighbourhoods-communities/neighbourhood-profiles/ .