Capstone Project - The Battle of the Neighborhoods

Introduction & Business Problem

Analysis of a locality based on the various venues available in that locality in give useful insights into the kind of Business thriving in that area. This profiling can be used to come up with the type of business which is likely to succeed in that locality.

This project aims at profiling the neighborhoods to come up with the best location for starting a new restaurant. The project aims to analyze the localities of New York and Toronto. This will involve profiling these neighborhoods. The profiling will be based on the number and category of venues of various types present in an area.

Data & Tools

New York and Toronto are the two cities which are planned to be analyzed as part of this assignment. New York data was provided as part of the previous assignment in the course. Information of the neighborhoods names of Toronto will be extracted from Wikipedia article as was done in assignment in week 3. Coordinates will be extracted using the Geocoder API, which will then be used as input for Foursquare to obtain venue information and map generation.

Methodology

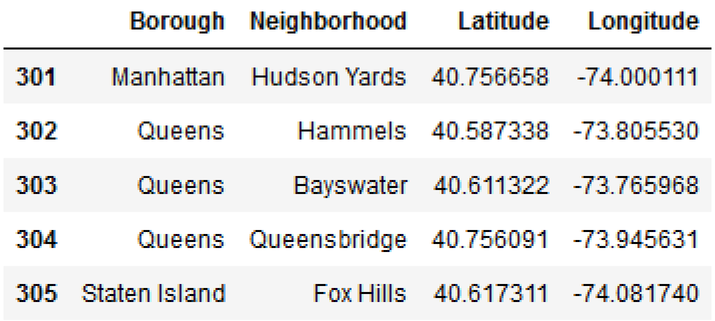
Stage 1 - Business Understanding: As stated in the previous section of this report, our main goal is to create a reliable profile of the neighborhoods in New York City and Toronto. Our fictional business sponsors are two entrepreneurs, one looking to open a new restaurant in New York City and another one looking to open a new bar in Toronto.

Stage 2 - Analytic Approach: To decide the ideal neighborhood for the new business, we must classify the neighborhoods into three main different kinds of regions based on the proportion of venue categories present in each one:
a) Residential
b) Services
c) "Going Out"

After the necessary data preparation (collection, encoding and normalization) the neighborhoods will be clustered into three groups using the k-means clustering algorithm. To solve our business problem, the third cluster "Going Out" will be further studied, and the venue categories in these neighborhoods in this group will be expanded, to give insight in the kinds of places that do not already exist in these neighborhoods. The information can help our business sponsors decide what kind of restaurant or bar are lacking and are probable business opportunities.

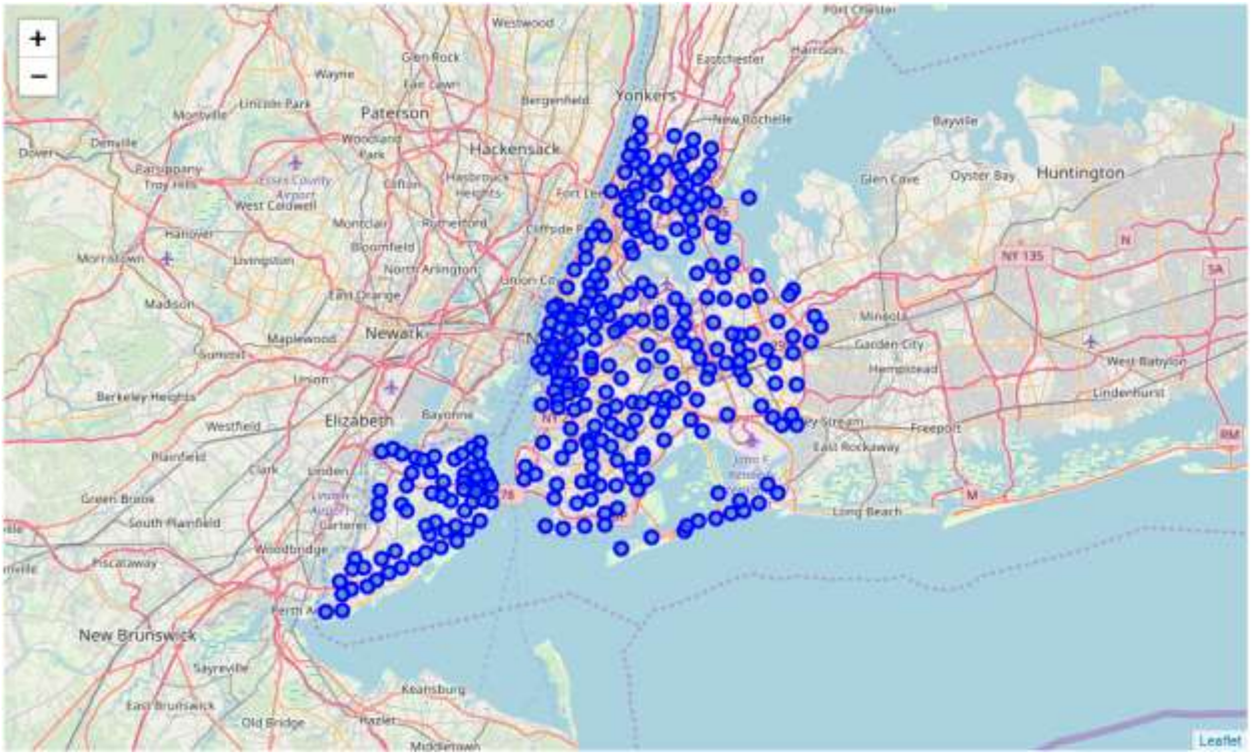
<u>Stage 3 - Data Requirements:</u> As stated in the Data & Tools section, the data requirements for this research are the venue information for each neighborhood in Toronto and New York City. Consequently, information about the neighborhoods (names and geographical coordinates) are also necessary.

<u>Stage 4 & 5 - Data Collection & Understanding:</u> The required data is collected in the first parts of the Jupyter Notebook. Toronto boroughs and neighborhoods are scrapped from the wikipedia link, using the BeautifulSoup package, and the New York City boroughs and neighborhoods information is scrapped from the JSON file. At this point the data is organized in a Pandas DataFrame like the following:

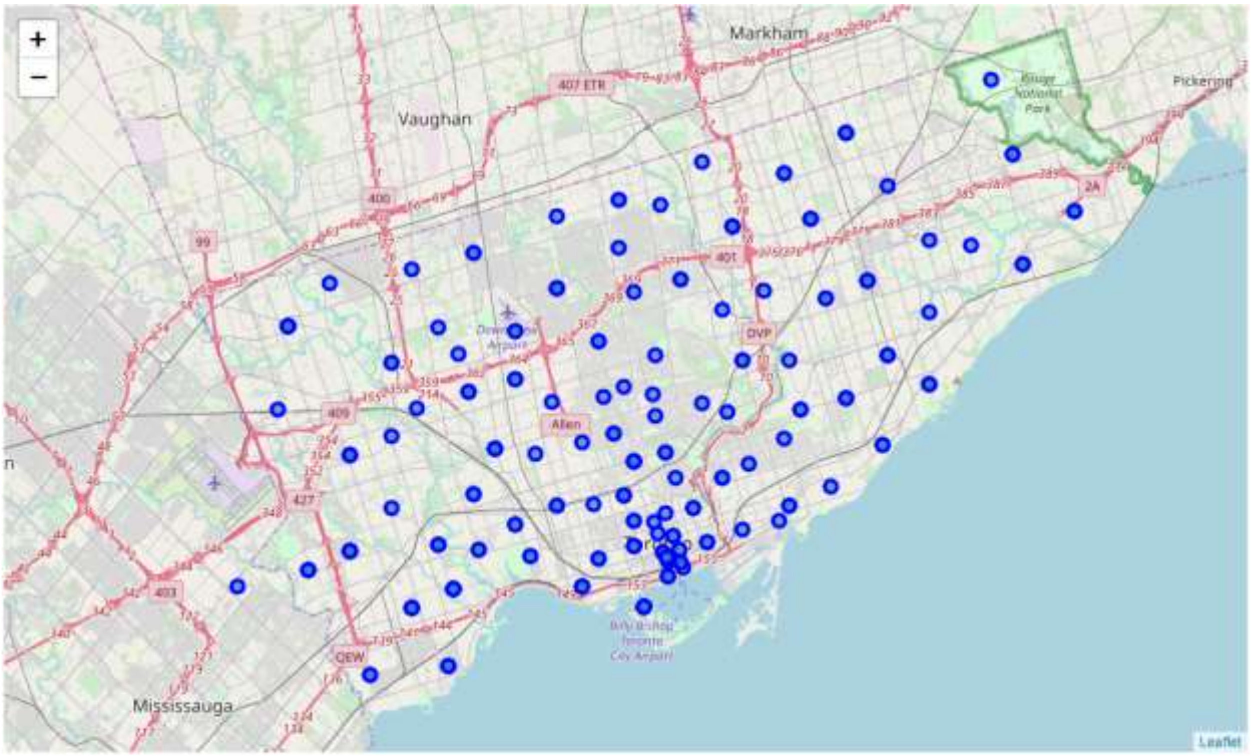


	Borough	Neighborhood	Latitude	Longitude
301	Manhattan	Hudson Yards	40.756658	-74.000111
302	Queens	Hammels	40.587338	-73.805530
303	Queens	Bayswater	40.611322	-73.765968
304	Queens	Queensbridge	40.756091	-73.945631
305	Staten Island	Fox Hills	40.617311	-74.081740

The "New York City, NY" dataframe has 5 boroughs and 302 neighborhoods, and the "Toronto, ON" dataframe has 11 boroughs and 210 neighborhoods. With the data collected at this point we can already visualize geographically each neighborhood using the Folium package to generate interactive Leaflet maps.



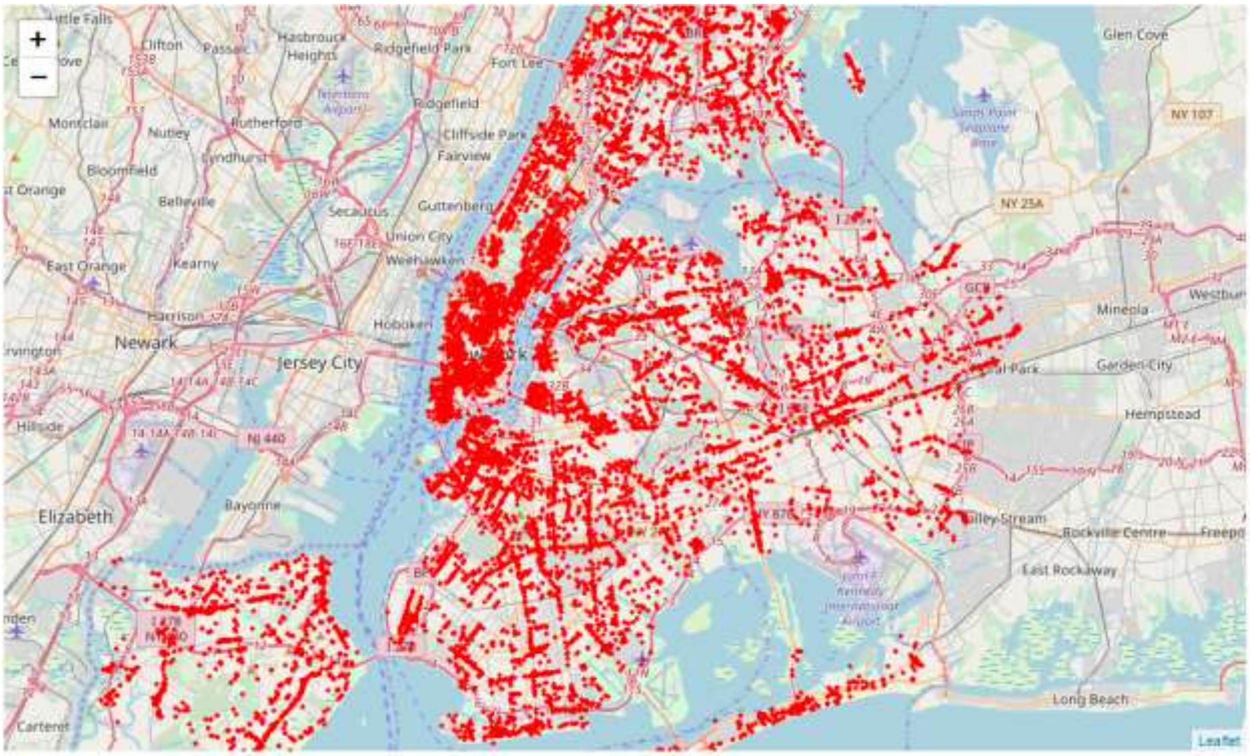
New York Neighborhoods Visualization



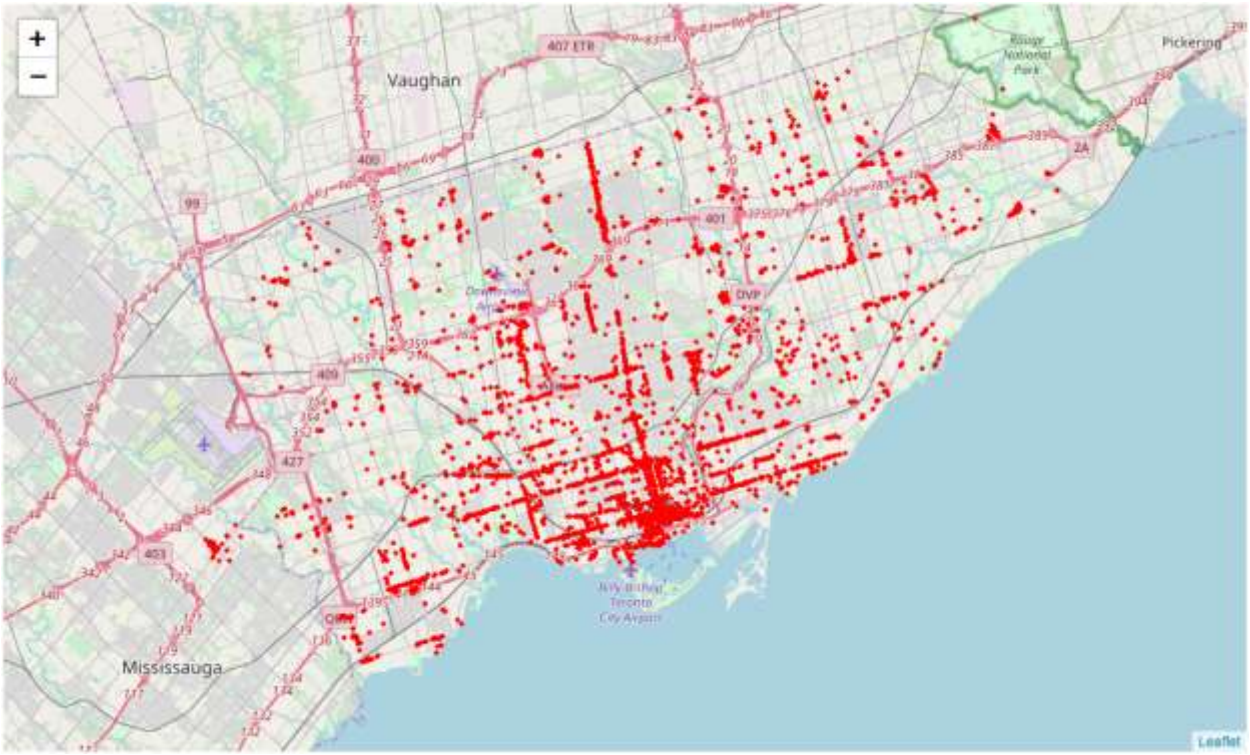
Toronto Neighborhoods Visualization

Departing from the same DataFrame, we now use the Foursquare API to collect venue data. Using the geographical coordinates of each neighborhood, API calls are made requesting the top 200 venues in a radius of 1000 meters. The results are inserted in a new pandas dataframe.

We generate geographical visualizations of the venue data as well, presented below in red dots. The "ny_venues" dataframe has 20537 venues and 466 unique venue types, and the "to_venues" dataframe has 9306 venues and 334 unique venue types. The proportion in number of venues are expected, considering the population and population density of these two cities.



New York Venue Visualization



Toronto Venue Visualization

Manually group Foursquare's venues categories found in New York City and Toronto. The Venue Category data extracted with the Foursquare API is very granular, to facilitate the visualization of data the 334 unique types of venues in Toronto and the 466 unique types of venues in New York City will be grouped into eight larger categories:

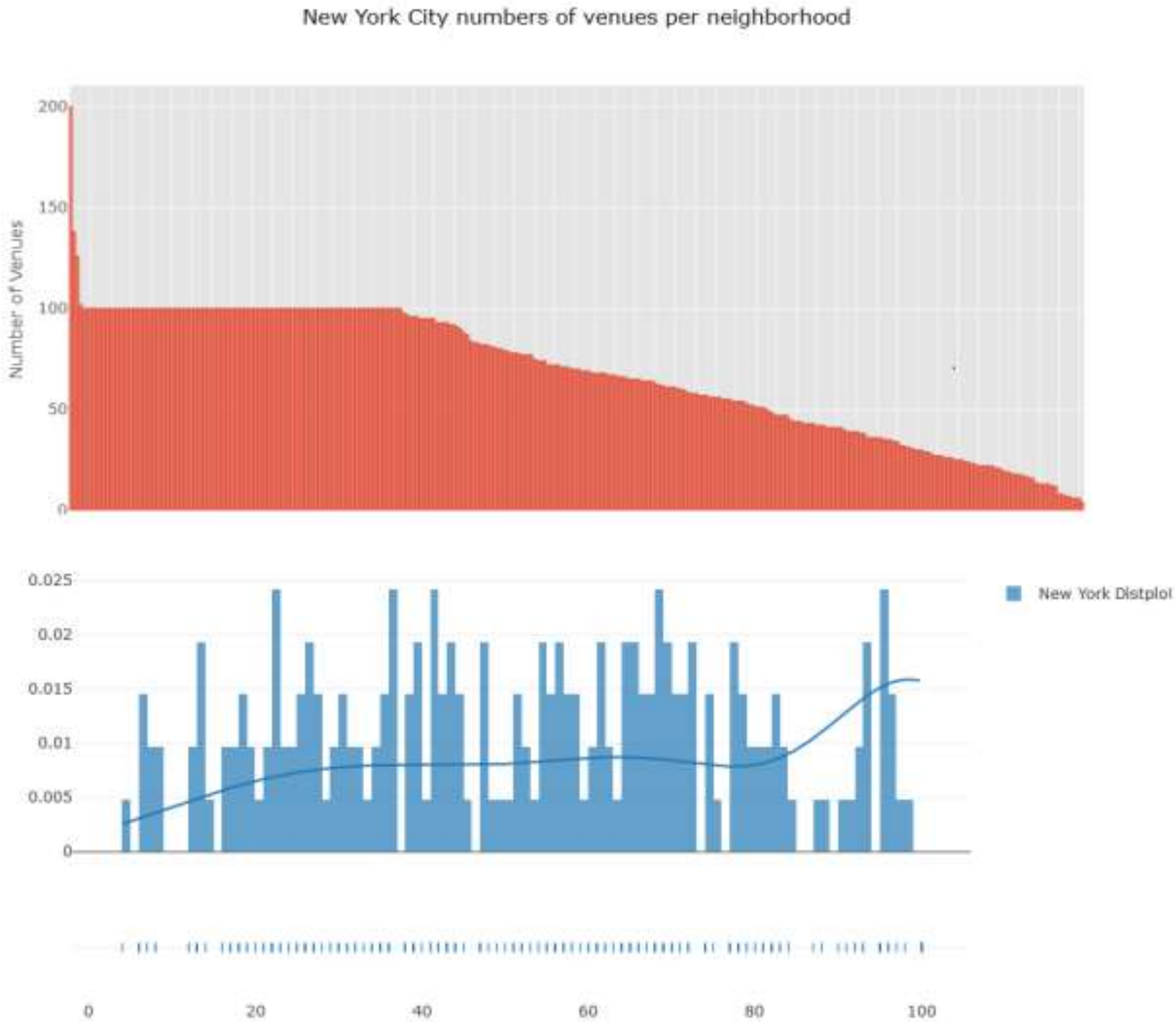
a) Bars and Clubs
b) Restaurants
c) General Services
d) Leisure & Sports
e) Culture & Education
f) Parks & Nature
g) Transportation Infrastructure
h) Residential

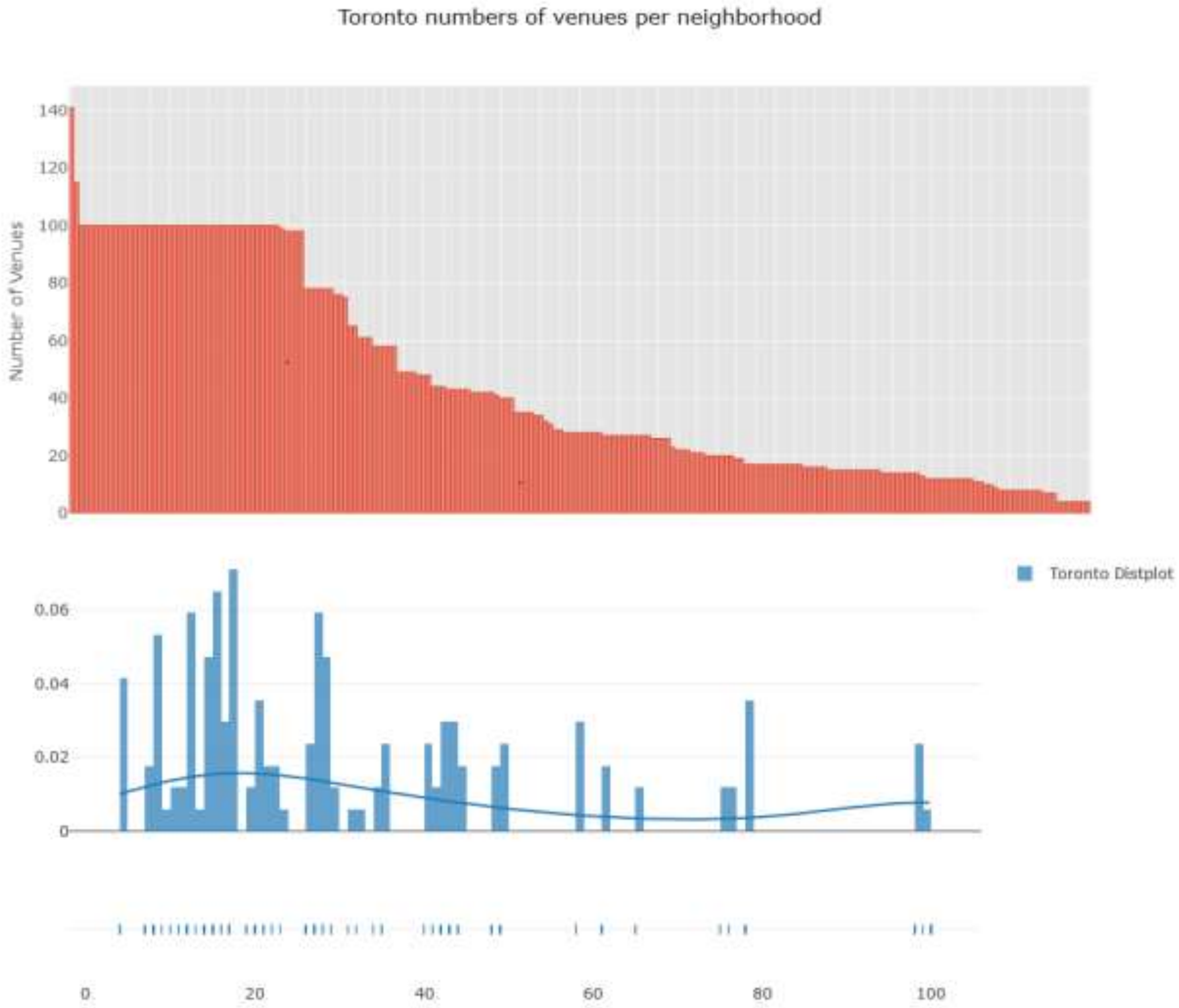
As it can be noted, New York City have slightly more variety of venues than Toronto. This is expected as the population in New York City is much larger than in Toronto, and also the number of neighborhoods.

Data Encoding: The next important step is the preparation of the data for the clustering/classification algorithms we are going to use later. Usually, only numeric inputs are valid in these algorithms, so in this section of our Juypter Notebook the dataframes with venue data collected and classified so far is encoded, creating a bigger dataframe following the model.

This data is then grouped for each Neighborhood, resulting in a dataframe with the number of venues in each category for each neighborhood. With this data prepared, we can generate several rich visualizations about the statistical venue makeup of New York and Toronto.

<u>Understanding the Data Collected so Far</u>. In this section we list some visualizations and distributions relevant to the topic of this work. First, a bar chart about the number of venues collected for each neighborhood and the correspondent distribution of neighborhoods based on the number of venues collected.

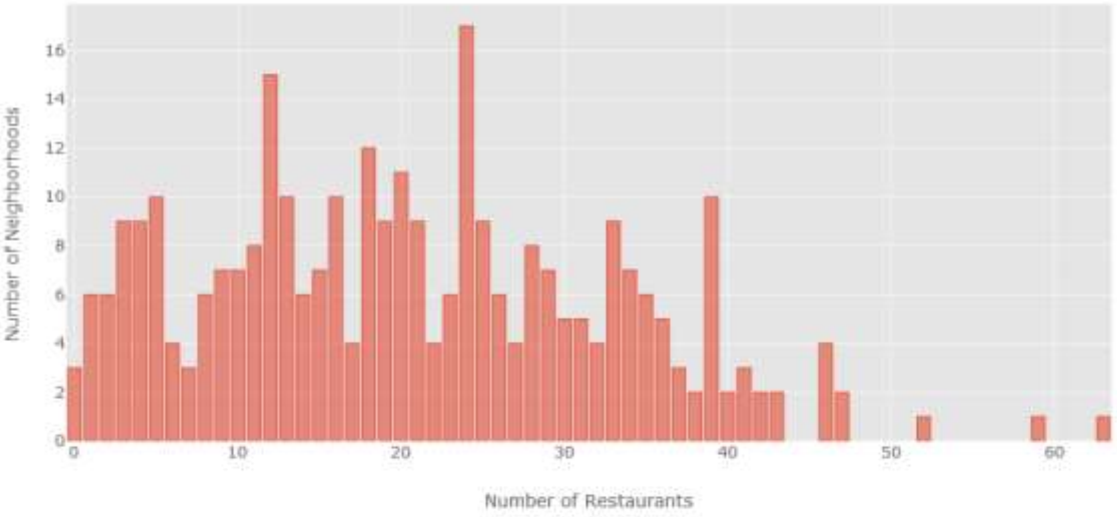


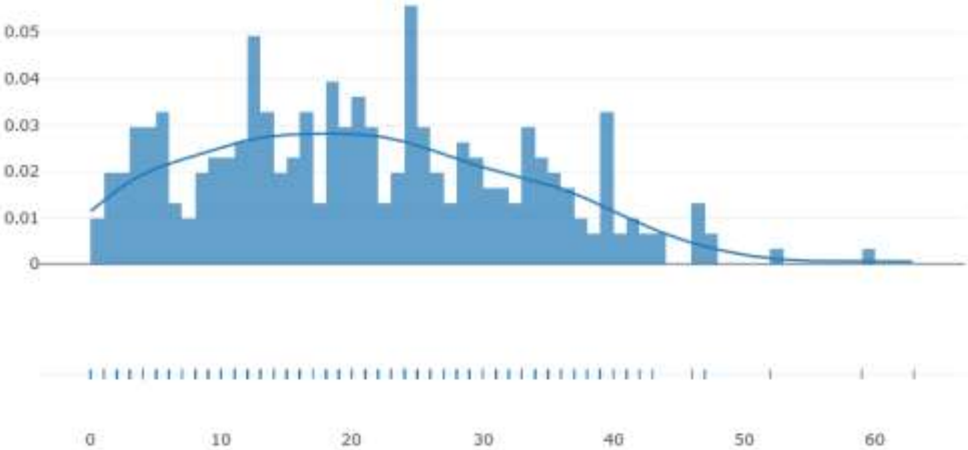


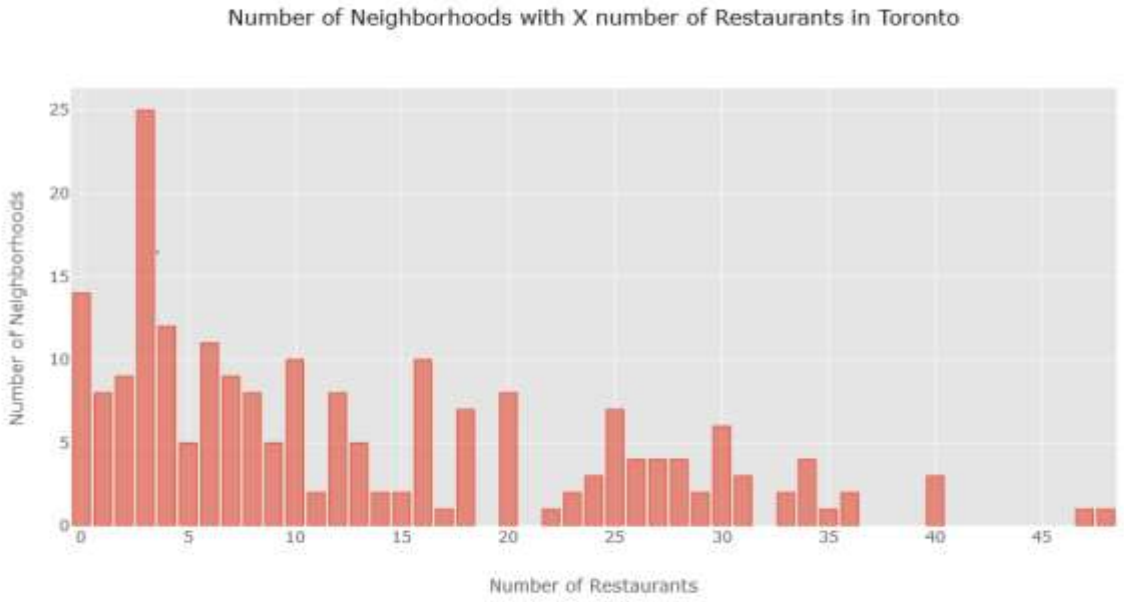
As it can be noted, New York City have several neighborhoods with "100" venues. A curious detail is the fact that for some reason the Foursquare API returned only 100 venues for each neighborhood, even after declaring explicitly that we wanted 200 in the API request. Nevertheless, several neighborhoods do not have more than one hundred venues returned and this won't be a big problem for our objectives.

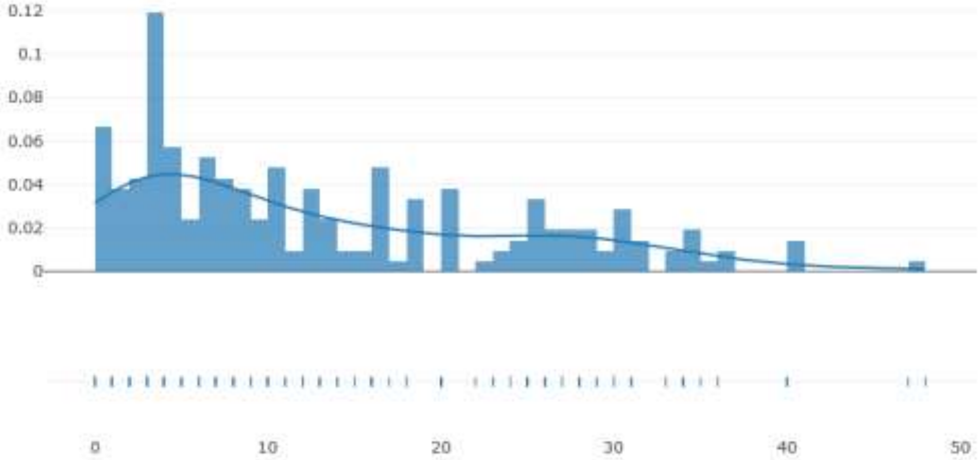
<u>Restaurants Distribution in Each City</u>: Repeating the plots but only considering venues of the "Restaurant" category, we can have an idea about the number of these types of business in each city and also their distribution between different neighborhoods. Both cities have some neighborhoods with several restaurants, indicating that there are agglomerations of business of the same kind at certain locations that will be further studied.

Number of Neighborhoods with X number of Restaurants in New York City

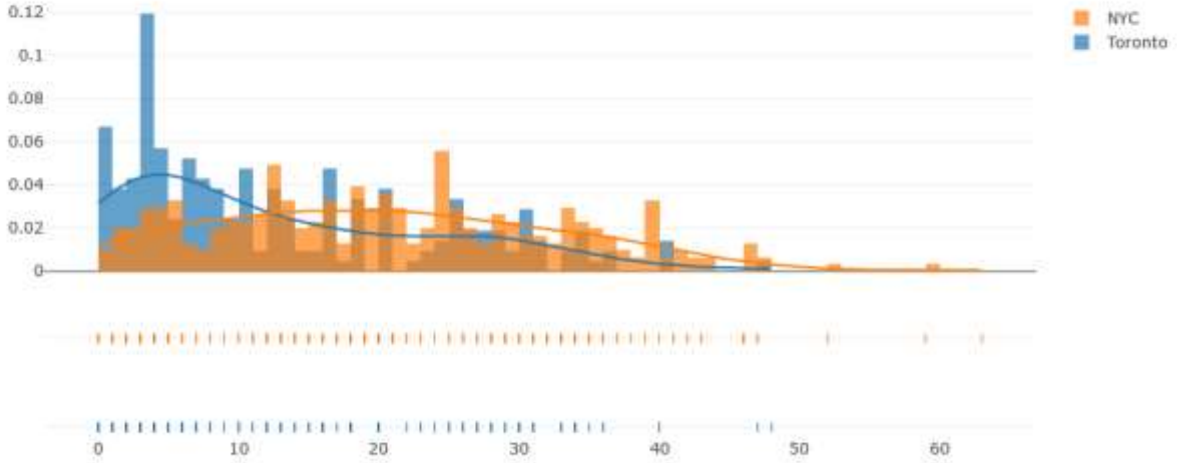




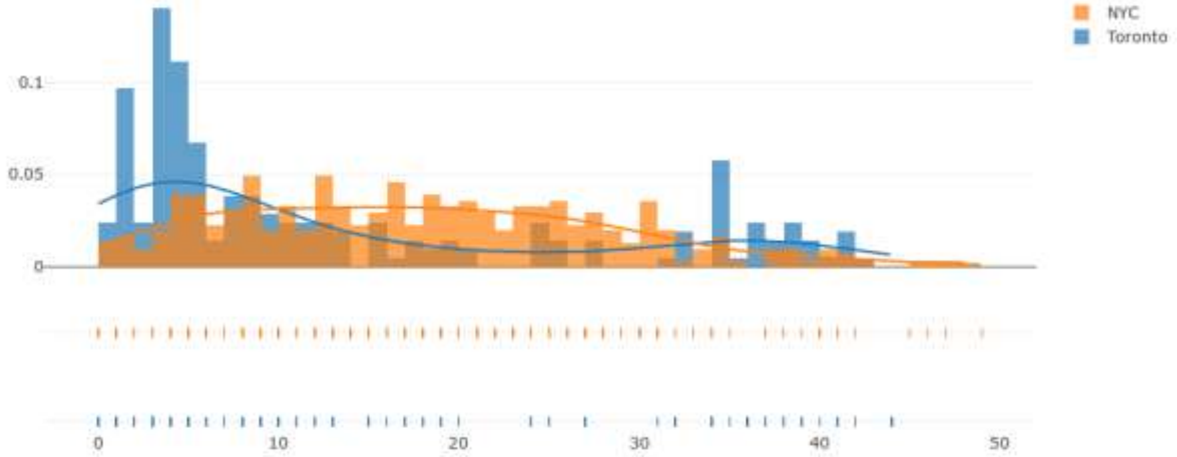




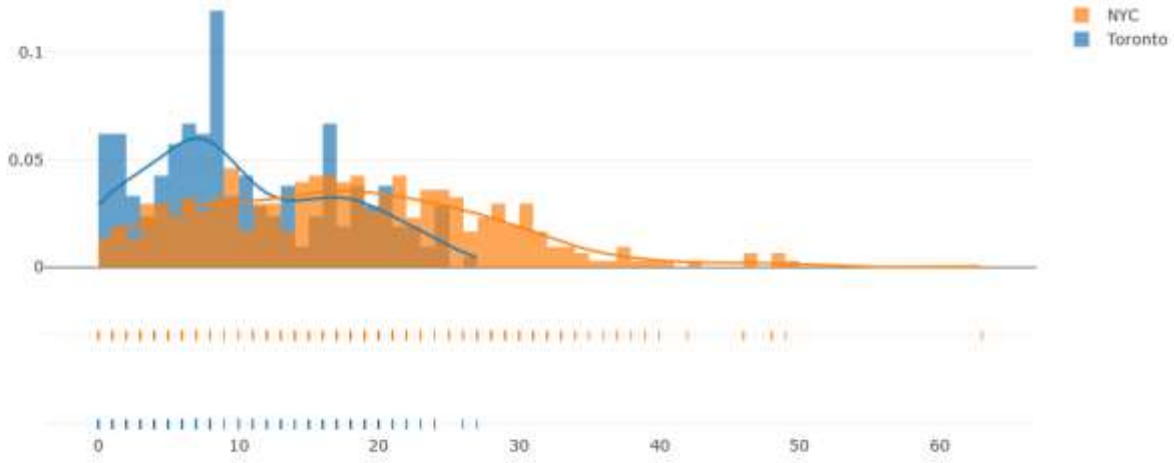
Comparing both distributions we can conclude that the restaurant business in New York City is much more saturated than in Toronto, as only few neighborhoods have been "taken" by these establishments.



<u>Bars and Clubs Distribution in Each City</u>: The Bars and Clubs distribution are much more different between Toronto and NYC. In NYC this kind of business seems also saturated, with lots of places distributed between several neighborhoods, while in Toronto there are much more inequality in bars and club's distribution between the neighborhoods - In Toronto this kind of business is concentrated in few neighborhoods.



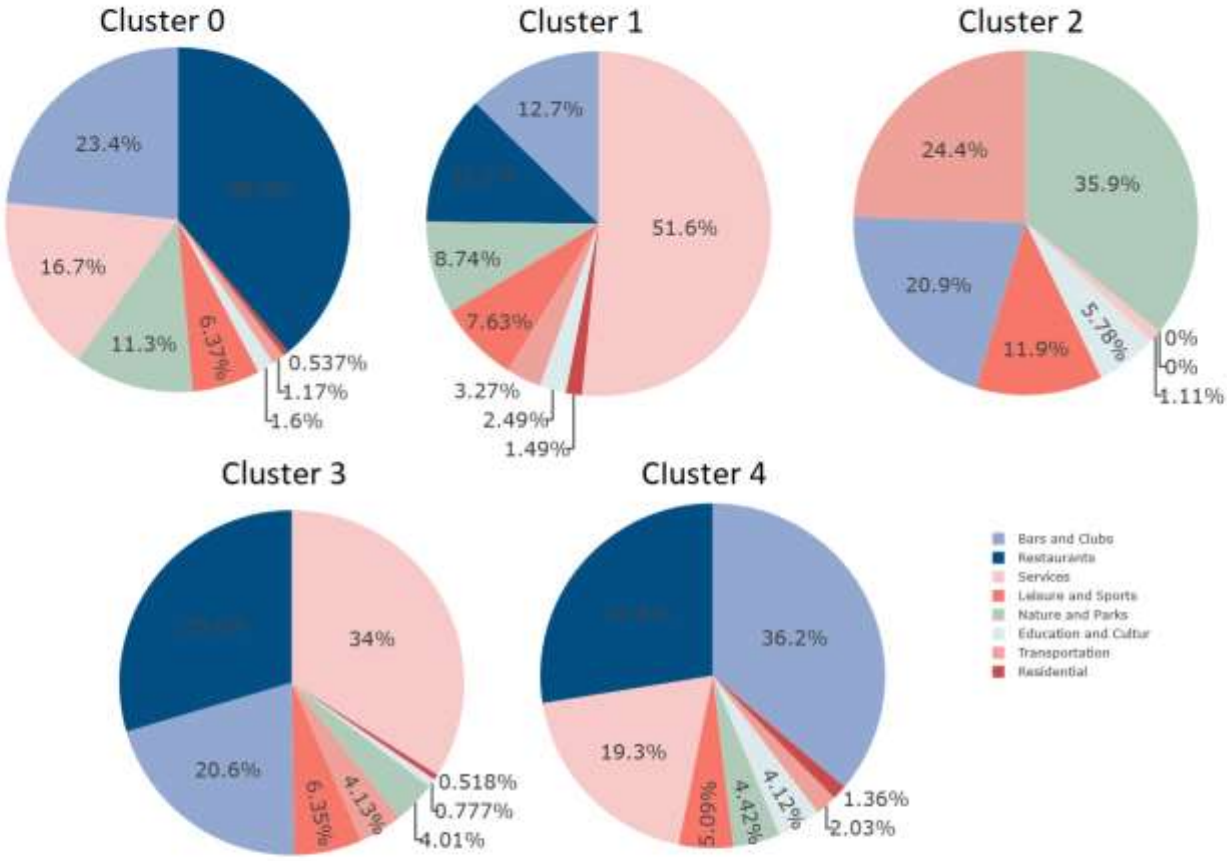
<u>General Services Distribution in Each City</u>: The general services distribution is somewhat similar between the two cities, but still following the same pattern than the previous ones - New York seems more saturated while in Toronto some neighborhoods have a deficiency of services.



Results

<u>Neighborhood K-Means Clustering based on Mean Occurrence of Each Venue Category</u>. With the previously encoded data, we will now aim to cluster the neighborhoods into five clusters, each one with different major characteristics. K-means clustering is the algorithm that will be used - this algorithm aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. K-Means uses an iterative refinement technique, and it is also referred to as Lloyd's algorithm. In the next pictures we show how the five clusters are characterized, in terms of median percent share of each kind of neighborhood.

<u>Clusters in Toronto</u>. The Toronto city clusters are presented below.



Cluster 0 aggregates neighborhoods with a huge proportion of restaurants, followed by bars and then services. The parks and nature proportional share in this cluster is higher than similar clusters in New York, indicating few residential venues like closed residential buildings, and bigger parks and forested areas (as is expected in some parts of Toronto than in an urban sprawl like NYC).

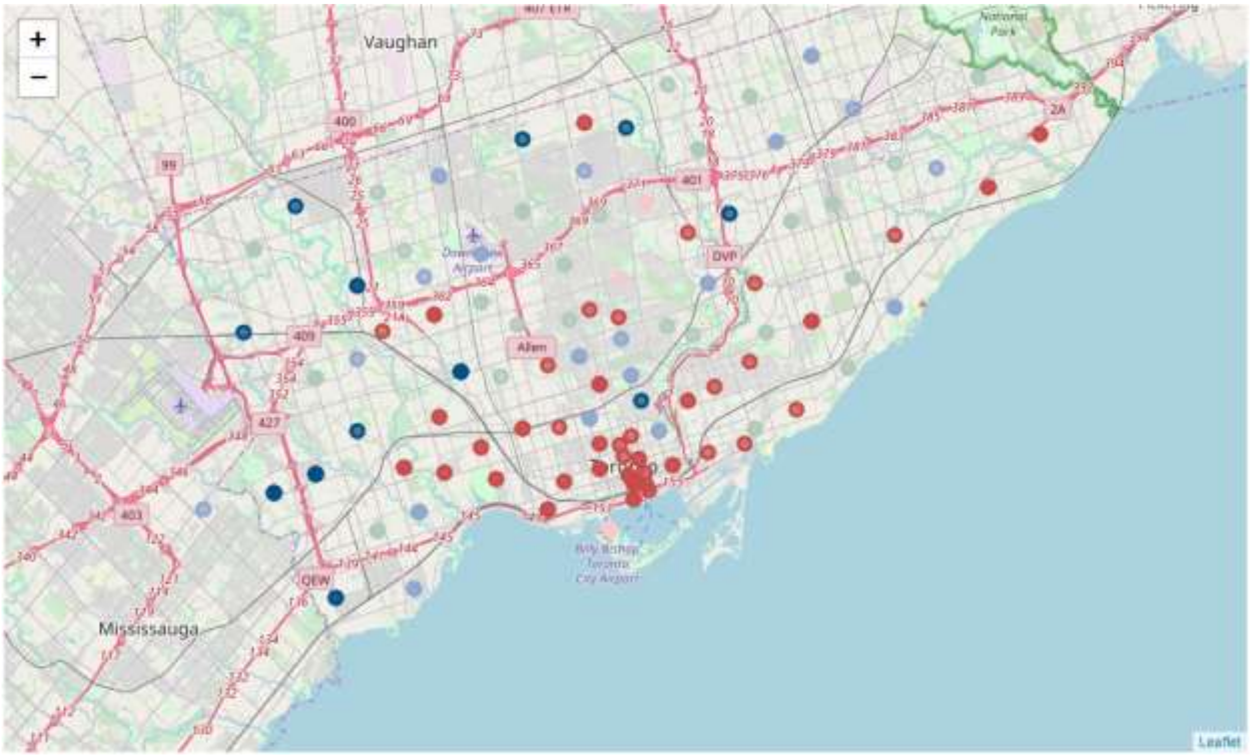
Cluster 1 aggregates neighborhoods with a huge proportional share of services (more than 50%!) indicating that neighborhoods classified in this cluster probably are commercial districts or city centers.

Cluster 2 aggregates neighborhoods with a high prevalence of parks and nature, and also transportation infrastructure. There are services, restaurants and bars but they aren't a common occurrence in these areas. Here we also have the neighborhoods with the highest proportional share of leisure, education, cultural, and sports venues (probably indicating some kind of university presence).

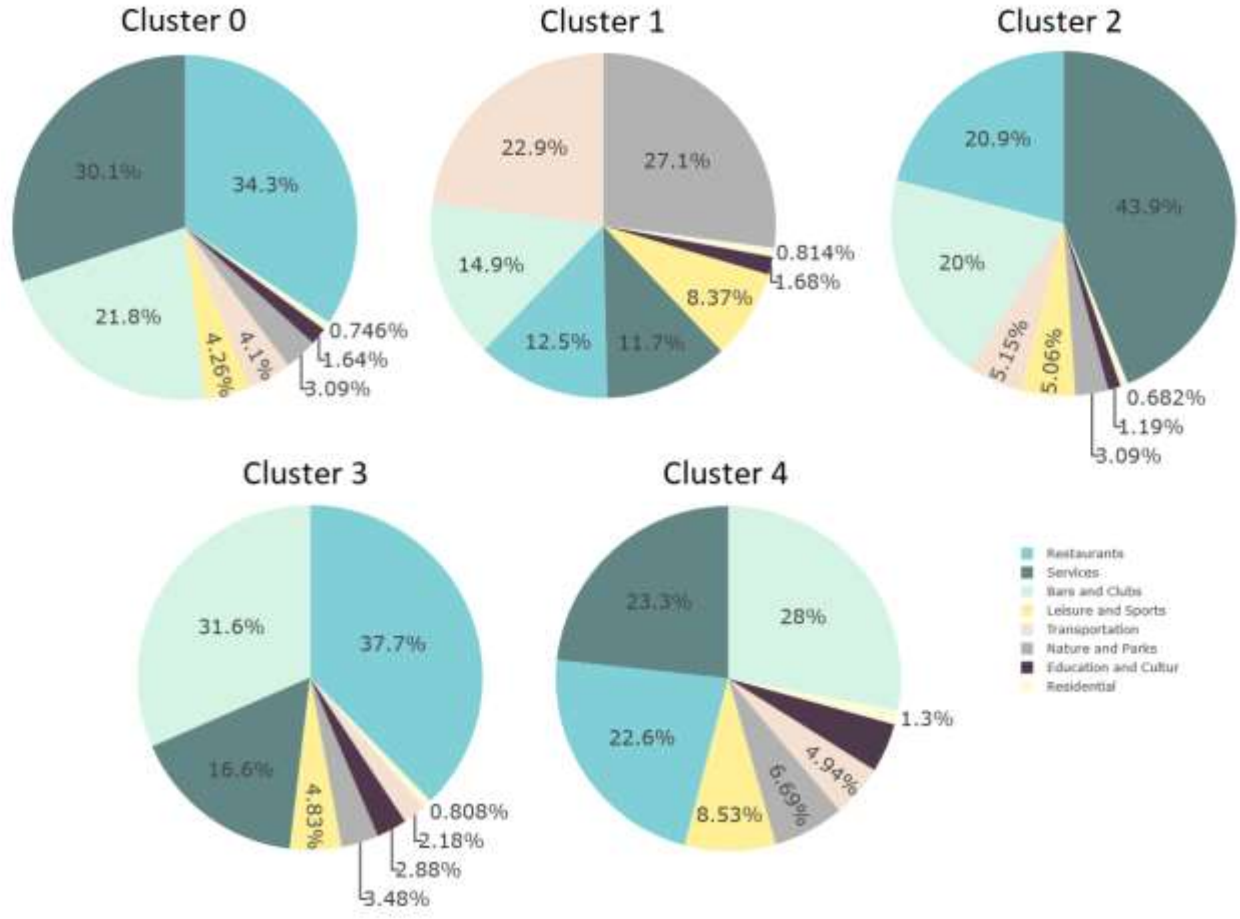
Cluster 3 aggregates neighborhoods with balanced shares of restaurants, bars, clubs and leisure. This cluster also has the lowest parks and nature proportional share, indicating that neighborhoods in these areas are highly urbanized parts of the city, indicating high development.

Cluster 4 aggregates the highest proportional share of bars and clubs in Toronto, indicating that these regions for some reason are attracting business related to nightlife and food. This cluster will be further analyzed to answer our specified business problem.

In the map below we can see the geographical visualization of the different types of clusters created using K-Means for Toronto. The colors indicate the biggest proportional share of venue category (using the same legend from the pie charts, except for the red one that indicate the balanced Cluster 3).



Clusters in New York City. The New York City clusters are presented below.



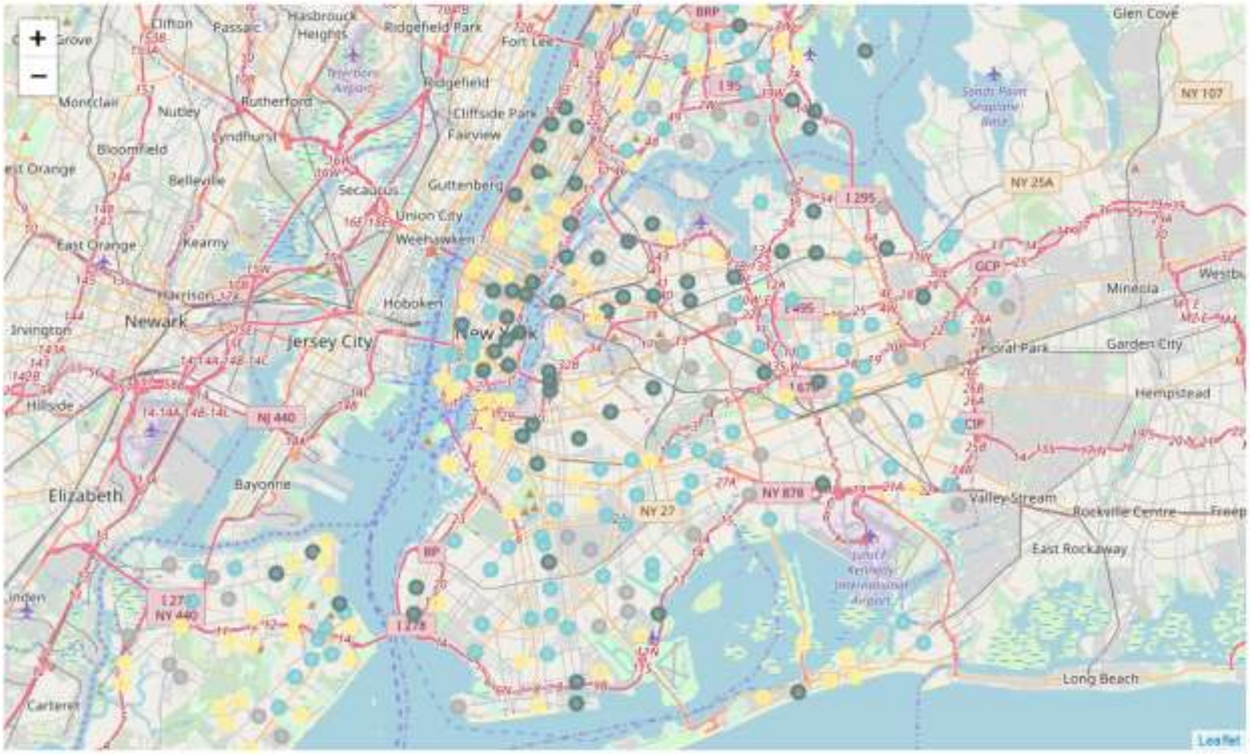
Cluster 0 aggregates neighborhoods with a near equal rate of restaurants, bars, and services venues, It is similar to Cluster 4 in the sense that the proportions are similar between these same three venue categories, except Cluster 4 have a bigger proportion of leisure, sports and cultural venues (probably indicating universities and schools present).

Cluster 1 aggregates neighborhoods with a high prevalence of parks and nature, and also transportation infrastructure. There are services, restaurants and bars but they aren't a common occurrence in these areas.

Cluster 2 aggregates neighborhoods with a huge proportion of services, followed by restaurants and bars, indicating that these neighborhoods are probably commercial districts or city centers.

Cluster 3 aggregates the largest proportion of restaurants and bars of all clusters, meaning that neighborhoods grouped in this cluster are important entities of study for our business problem related to restaurants and bars distribution.

In the map below we can see the geographical visualization of the different types of clusters created using K-Means for New York City. The colors indicate the biggest proportional share of venue category (using the same legend from the pie charts, except for the yellow one that indicate Cluster 4).



After the study of the data presented, we selected some clusters of interest in each city.

Toronto Analysis. For Toronto, we selected the clusters 3 and 4. The cluster 4 aggregates neighborhoods with high numbers of bars and good demand and accessibility for the public (suburban areas), indicating places with lower rents and property prices - relative to city center neighborhoods, of the also selected cluster 3. Cluster 3 groups the more urbanized and developed parts of Toronto, with several services category venues - making these areas great neighborhoods with high demand for restaurants, bars, nightclubs, etc. The list of neighborhoods in this cluster is presented below, and they basically form a list of places with well-established business in the restaurant/bar/club's segment. The optimal location for a new business in the restaurant or bar category can be further studied with the granular data about the specific themes of restaurants and bars. High demand signals high offerings and also higher competitivity, meaning that it's probably better to start a "new" kind of venue, in an untapped market in an underdeveloped or suburban area.

Good neighborhoods for establishing new restaurant venues in Toronto (possible untapped markets): Richview Gardens, Roselawn, Rouge, Royal York South West, Scarborough Town Centre, Silverstone, Parkview Hill, Mimico NW, Mount Olive, Oriole, Maryvale, South Steeles, Wexford Heights, Wilson Heights, Woodbine Gardens, York Mills West, Wexford, Thorncliffe Park, Thistletown, South of Bloor, St. Phillips, Steeles West, The Beaches West, The Queensway West, Martin Grove Gardens, Cliffside West, Downsview North, Downsview Northwest, Downsview West, East Birchmount Park, Dorset Park, Albion Gardens, Bathurst Manor, Beaumond Heights, Bedford Park, Birch Cliff, Caledonia-Fairbanks, Cedarbrae, Jamestown, Kennedy Park, Kingsview Village, Kingsway Park South West, L'Amoreaux West, Lawrence Heights, Lawrence Manor, Lawrence Manor East, Leaside, Malvern, Ionview, India Bazaar, Glencairn, Henry Farm, Hillcrest Village, Humbergate, Fairview

Good neighborhoods for restaurant venues in Toronto (but probably saturated markets):

Railway Lands, Silver Hills, York Mills, South Niagara, Bathurst Quay, CN Tower, King and Spadina, Lawrence Park, Island airport, Harbourfront West

New York City Analysis. For NYC, we selected the clusters 0 and 4. The cluster 4 aggregates neighborhoods with high but balanced proportions of services, restaurants, and bars/clubs. Comparing cluster 4 with cluster 0, that aggregates city center neighborhoods with very high proportional share of restaurants and also relatively high proportion of services, we can notice that cluster 4 is behind cluster 0 in the gentrification, or urban development process. This information can be used to plan the best locations for a restaurant business based on the intentions of our business sponsor: does he want to open a restaurant in some place with high demand, but also high price of entry or he wants to bet in a place with less entrenched competitors and also good demand? We list the possible neighborhoods for each group in the next subsection.

Good neighborhoods for establishing new restaurant venues in NYC (not so much competition and good demand - "safe bets"):

Oakland Gardens, North Side, North Riverdale, North Corona, Noho, Prospect Heights, Pelham Parkway, New Brighton, Murray Hill, Manhattanville, Manhattan Valley, Manhattan Beach, Lower East Side, Murray Hill, Morningside Heights, Midtown South, Ravenswood, Upper West Side, Turtle Bay, Tudor City, Tottenville, Throgs Neck, Sunnyside Gardens, Williamsburg, Whitestone, West Village, West Brighton, Sunnyside, Stuyvesant Town, Roosevelt Island, Rockaway Beach, Riverdale, Ridgewood, Schuylerville, Sheepshead Bay, Steinway, South Side, South Ozone Park, City Island, Chinatown, Central Harlem, Bushwick, Clifton, Clinton Hill, Bayside, Bay Ridge, Astoria, Annadale, Bedford Stuyvesant, Briarwood, Bergen Beach, Woodside, Hamilton Heights, Great Kills, Gramercy, Kingsbridge, Jackson Heights, Inwood, Hunters Point, Elmhurst, Edgewater Park, East Village, East Harlem, Fieldston, Fort Hamilton, Fort Greene, Forest Hills Gardens, Flushing, Flatbush, Yorkville

Good neighborhoods for establishing new clubs and bars in NYC (not so much competition and good demand - "safe bets"):

Ozone Park, New Springville, New Dorp Beach, New Dorp, Queens Village, Pleasant Plains, Malba, Lindenwood, Marine Park, Mount Hope, Morrisania, Mill Basin, Middle Village, Maspeth, Westerleigh, Rosedale, Richmond Valley, Sea Gate, Starrett City, South Jamaica, Soundview, Lefrak City, Charleston, Castle Hill, Co-op City, Baychester, Bay Terrace, Arlington, Bellerose, Bloomfield, Heartland Village, Holliswood, Hunts Point, Glendale, Elm Park, Eastchester, East Flatbush, Erasmus, Georgetown, Floral Park

Good neighborhoods for restaurant venues in NYC (but probably saturated markets - entrenched business):

Allerton, Olinville, Old Town, Ocean Parkway, Ocean Hill, Norwood, New Lots, Paerdegat Basin, Park Hill, Parkchester, Queensboro Hill, Prospect Park South, Prospect Lefferts Gardens, Prince's Bay, Port Richmond, Park Slope, Manhattan Terrace, Madison, Longwood, Little Neck, Mariner's Harbor, Melrose, Mount Eden, Mott Haven, Morris Park, Midwood, Rego Park, Utopia, University Heights, Unionport, Tribeca, Sutton Place, Sunset Park, Washington Heights, Woodhaven, Williamsbridge, Westchester Square, Weeksville, Wakefield, Rugby, Rosebank, Rochdale, Richmond Hill, Remsen Village, St. Albans, Spuyten Duyvil, Soho, Laurelton, City Line, Canarsie, Cambria Heights, Civic Center, Dongan Hills, Ditmas Park, Cypress Hills, Corona, Concord, College Point, Brownsville, Bay Terrace, Bath Beach, Auburndale, Bayswater, Bronxdale, Borough Park, Douglaston, Bensonhurst, Belmont, Bellaire, Hillcrest, Highland Park, Greenwich Village, Greenridge, Grasmere, Grant City, Graniteville, Hollis, Howard Beach, Kew Gardens Hills, Kensington, Jamaica Hills, Jamaica Estates, Jamaica Center, Homecrest, Glen Oaks, Eltingville, Edenwald, East Tremont, East New York, Far Rockaway, Fresh Meadows, Fordham, Flatlands, Flatiron, Forest Hills