Comparing Neighborhoods in New York City and Toronto

IBM Applied Data Science Capstone

March 2nd, 2021

Introduction/Problem Statement

This report analyzes and compares two important cities of the world: New York City and Toronto. Both cities are financial centers of the United States and Canada, respectively. This report will look into the neighborhoods of each city, investigating most popular businesses in different neighborhoods and drawing conclusions on which businesses are most likely to succeed in each city/neighborhood.

The target audience of this project are individuals and/or companies that are interested in investing in the United States or Canada. Apart from finding out the popularity of different types of businesses in each city, the report will also discuss which businesses are more likely to succeed in which neighborhoods. For example, if an individual is considering opening up a coffee shop in Toronto, he/she could use this information to figure out in which neighborhoods of Toronto are coffee shops the most popular business.

New York City (NYC) is the most densely populated city in the United States, with the population of over 8 million. New York City is the center of the New York metropolitan area, which is the largest metropolitan area in the world. NYC has significant influence on the commerce, technology, politics, art and fashion industries in the country. The city is composed of five boroughs - Brooklyn, Queens, Manhattan, the Bronx, and Staten Island.

Toronto is the capital city of the Canadian province of Ontario and the most populous city in Canada. Toronto is an international center of business and finance. Toronto is also home to the Toronto Stock Exchange and headquarters of large Canadian and multinational corporations. Its economy is highly diversified and performs especially well in technology, financial services, tourism, and aerospace.

Data Description/How it Will Be Used to Solve the Problem

This section focuses on the data that is being collected for analysis. There are two main types of data used: (1) neighborhood data with the zip codes and latitude/longitude values, and (2) venues data.

New York: The neighborhood data for New York (linked here as a json file) was provided by the IBM data science team and the latitude/longitude values were extracted from the Geopy library. The venue data for NYC was extracted from Foursquare API. Foursquare lists top 100 venues of each city by

neighborhood and even allows us to categorize venues (different categories of venues include coffee shops, restaurants, parks, etc).

Toronto: Neighborhood data for Toronto was scraped from a Wikipedia page (linked here), which includes postal codes, boroughs and neighborhoods columns. Next, latitude and longitude values for the neighborhood data were loaded from a csv file which was also provided by the IBM data science team. Regarding the venue data for Toronto, the acquisition process was similar as that of New York: Foursquare API was used to get top 100 venues of Toronto which were then categorized.

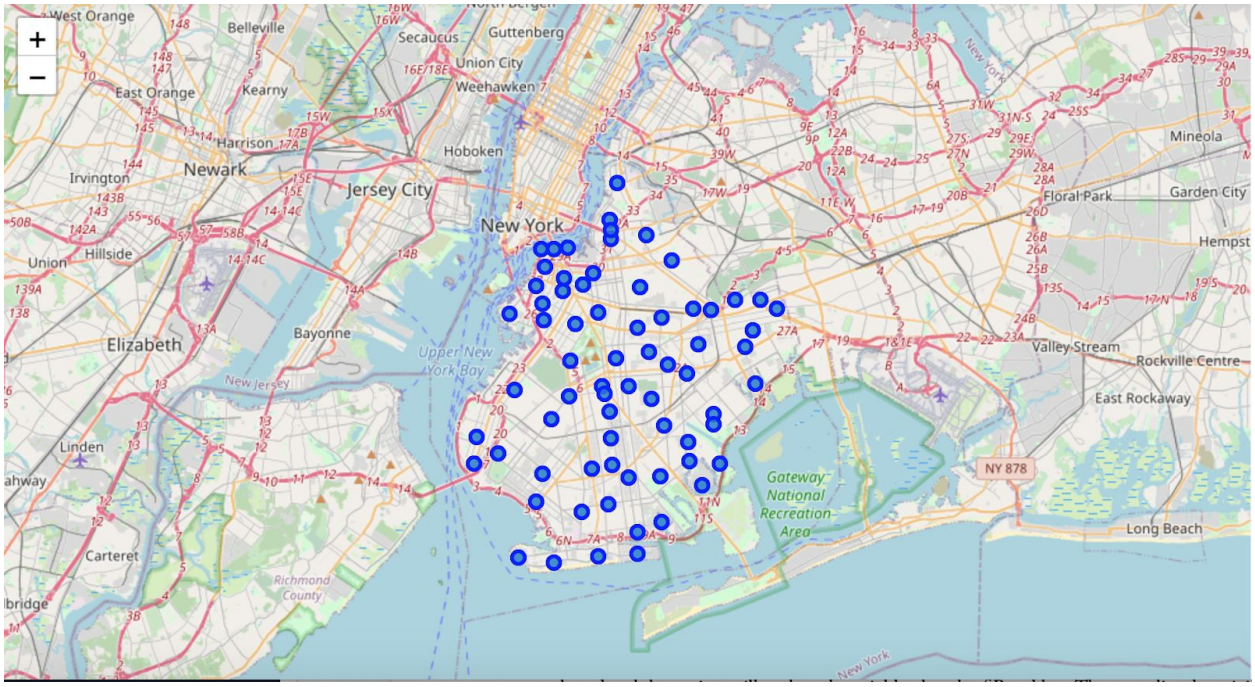
Finally, in addition to the Foursquare API and geopy library, the Folium library was used to create maps for both cities with neighborhoods superimposed on top. This data and libraries/tools will help us find out which categories of venues/businesses tend to be most frequent in the neighborhoods of New York City and Toronto. These results can be useful for (1) comparison purposes - it will let us find out which venues are most popular in each city; (2) investing purposes - individuals or businesses thinking about investing in either New York or Toronto, which are most populous cities of their respective countries, will find it useful to know which categories of venues tend to more successful in each city and even neighborhood.

Methodology

Exploratory Data Analysis for New York

This project explores the neighborhood and venues data for New York City and Toronto. First, we start by creating a dataframe for New York that includes Borough, Neighborhood, Latitude, and Longitude columns. As discussed in the previous section, data on New York City's neighborhoods and boroughs were already provided by IBM and the latitude and longitude values were generated by Geopy library.

After we ensure that the data frame includes New York's 5 boroughs and 306 neighborhoods, we narrow down search to one of the boroughs of New York City. Given its size and diversity, Brooklyn was selected and the project will explore the neighborhoods of Brooklyn. Thus, we slice the original data frame and create a new dataframe specifically for the Brooklyn data. Then, we use geolocator to get latitude and longitude values of Brooklyn and use Folium to create a map of Brooklyn with the neighborhoods superimposed on top.



Next, we start using the Foursquare API to explore the venues of the Brooklyn neighborhoods. First neighborhood of Brooklyn is "Bay Ridge", for which we get latitude and longitude values. Then, we get the top 100 venues of the Bay Ridge neighborhood using Foursquare data. Next, since we want to know the categories of venues rather than venues themselves, we borrow the get_category_type function from the Foursquare lab. As a result, we get venues and categories of venues for the Bay Ridge neighborhood.

	name	categories	lat	lng
0	Pilo Arts Day Spa and Salon	Spa	40.624748	-74.030591
1	Bagel Boy	Bagel Shop	40.627896	-74.029335
2	Cocoa Grinder	Juice Bar	40.623967	-74.030863
3	Leo's Casa Calamari	Pizza Place	40.624200	-74.030931
4	Pegasus Cafe	Breakfast Spot	40.623168	-74.031186

Since we are interested in Brooklyn in general and not just the Bay Ridge neighborhood, we create a function that repeats the same process for all the neighborhoods in Brooklyn. After running the Foursquare function for Brooklyn neighborhoods, we create a new dataframe for Brooklyn venues which has 5434 rows and 7 columns.

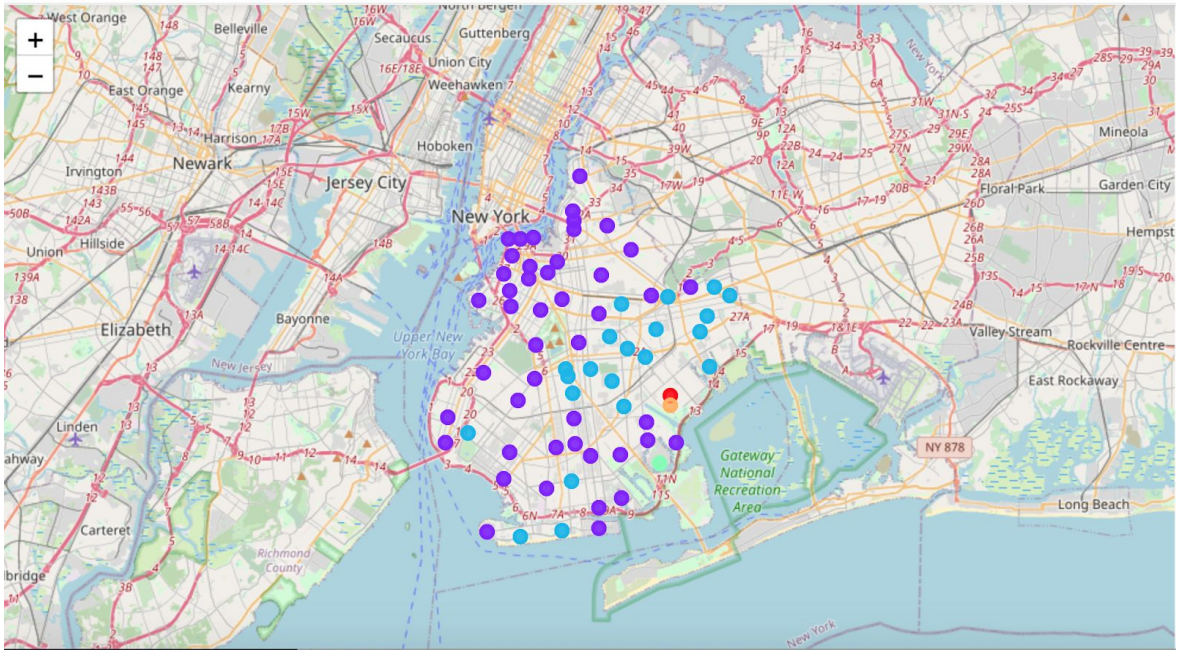
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bay Ridge	40.625801	-74.030621	Pilo Arts Day Spa and Salon	40.624748	-74.030591	Spa
1	Bay Ridge	40.625801	-74.030621	Bagel Boy	40.627896	-74.029335	Bagel Shop
2	Bay Ridge	40.625801	-74.030621	Cocoa Grinder	40.623967	-74.030863	Juice Bar
3	Bay Ridge	40.625801	-74.030621	Leo's Casa Calamari	40.624200	-74.030931	Pizza Place
4	Bay Ridge	40.625801	-74.030621	Pegasus Cafe	40.623168	-74.031186	Breakfast Spot

Then, we move on to analyzing each neighborhood. As a first step in this process, we take the mean of the frequency of occurrence of each category and group rows by neighborhood. Then, we display top 10 venues for each neighborhood and create a function that sorts the venues in descending order. Finally, we use the same function to create a new dataframe and display the top 10 venues for each neighborhood.

	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	Bath Beach	Pizza Place	Pharmacy	Chinese Restaurant	Fast Food Restaurant	Cantonese Restaurant	Italian Restaurant	Bubble Tea Shop	Peruvian Restaurant	Gas Station	Sushi Restaurant
1	Bay Ridge	Spa	Pizza Place	Italian Restaurant	Bagel Shop	Greek Restaurant	American Restaurant	Bar	Sandwich Place	Café	Chinese Restaurant
2	Bedford Stuyvesant	Coffee Shop	Pizza Place	Deli / Bodega	Bar	Café	Park	Discount Store	Cocktail Bar	New American Restaurant	Thrift / Vintage Store
3	Bensonhurst	Chinese Restaurant	Pizza Place	Ice Cream Shop	Dessert Shop	Sushi Restaurant	Donut Shop	Italian Restaurant	Bagel Shop	Liquor Store	Factory
4	Bergen Beach	Park	Harbor / Marina	Playground	Athletics & Sports	Baseball Field	Food	Food & Drink Shop	Flower Shop	Flea Market	Fish Market

Machine Learning Used for New York: K-means Clustering

For New York, we use a k-clustering algorithm to cluster Brooklyn neighborhoods into five clusters and use folium to visualize them. The goal of using k-means clustering algorithm is to provide cluster-level output as opposed to neighborhood-level one. For example, if an individual is thinking about opening up a coffee shop in Brooklyn, he/she will have more generalized data on the clusters of neighborhoods and will proceed to choose one of the neighborhoods from the given cluster.



Exploratory Data Analysis for Toronto

For Toronto, after we scrape data from a Wikipedia page and get latitude and longitude values for each neighborhood, we proceed to create a map using Folium with neighborhoods superimposed on top.

Then, we move on to using Foursquare API and exploring our first neighborhood of Toronto, which is Regent Park. We get latitude and longitude values from the Geopy library and get top 100 venues of this neighborhood. Next, like we did for New York, we borrow the get_category_type from the Foursquare lab to extract the category of each venue. After extracting categories for the Regent Park venues, we create a function called getNearbyVenues which will repeat the same process for all the neighborhoods in Toronto. Finally, we print each neighborhood along with the top 10 most common venues.

	Neighbourhood	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	Berczy Park	Coffee Shop	Cocktail Bar	Bakery	Seafood Restaurant	Beer Bar	Farmers Market	Pharmacy	Cheese Shop	Restaurant	Juice Bar
1	Brockton, Parkdale Village, Exhibition Place	Café	Breakfast Spot	Coffee Shop	Pet Store	Bar	Climbing Gym	Stadium	Furniture / Home Store	Italian Restaurant	Intersection
2	Business reply mail Processing Centre, South C...	Yoga Studio	Auto Workshop	Smoke Shop	Burrito Place	Light Rail Station	Farmers Market	Fast Food Restaurant	Butcher	Restaurant	Skate Park
3	CN Tower, King and Spadina, Railway Lands, Har...	Airport Service	Airport Lounge	Airport	Harbor / Marina	Coffee Shop	Sculpture Garden	Boat or Ferry	Airport Terminal	Airport Gate	Airport Food Court
4	Central Bay Street	Coffee Shop	Sandwich Place	Café	Italian Restaurant	Thai Restaurant	Japanese Restaurant	Burger Joint	Bubble Tea Shop	Salad Place	Ramen Restaurant

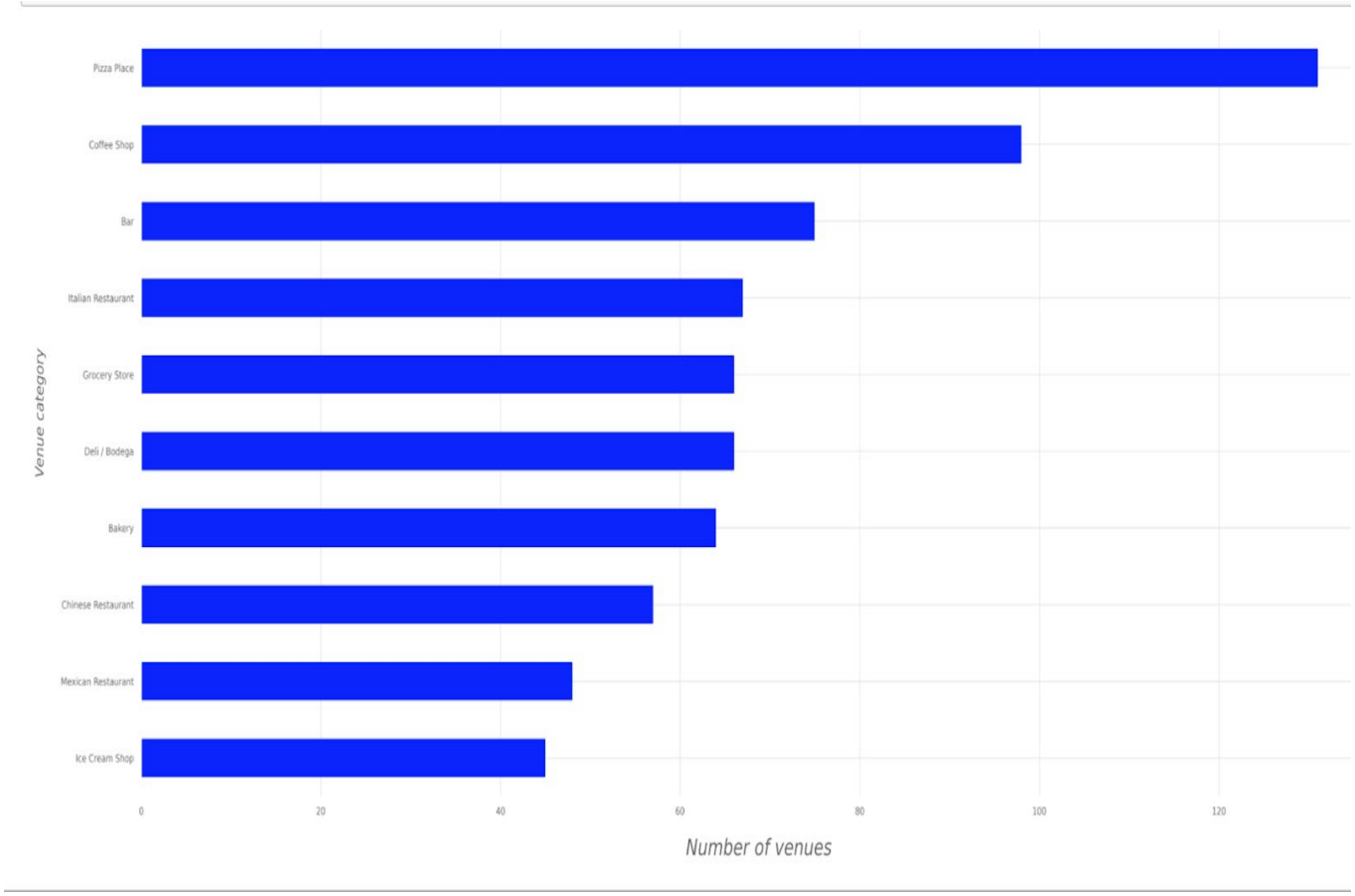
Machine Learning Used for Toronto: K-means Clustering

In order to get more high-level data, we use k-means clustering algorithm and cluster Toronto's neighborhoods into five clusters and create a dataframe that shows most common venues for each neighborhood along with their cluster labels.

	Postcode	Borough	Neighbourhood	Latitude	Longitude	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
0	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	0	Coffee Shop	Bakery	Park	Breakfast Spot	Café	Pub	Theater
1	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	0	Coffee Shop	Sushi Restaurant	Diner	College Cafeteria	Bar	Beer Bar	Smoothie Shop
2	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937	0	Coffee Shop	Clothing Store	Hotel	Bubble Tea Shop	Café	Middle Eastern Restaurant	Italian Restaurant
3	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418	0	Café	Coffee Shop	Gastropub	American Restaurant	Cocktail Bar	Clothing Store	Park
4	M4E	East Toronto	The Beaches	43.676357	-79.293031	0	Neighborhood	Health Food Store	Trail	Pub	Yoga Studio	Dog Run	Diner

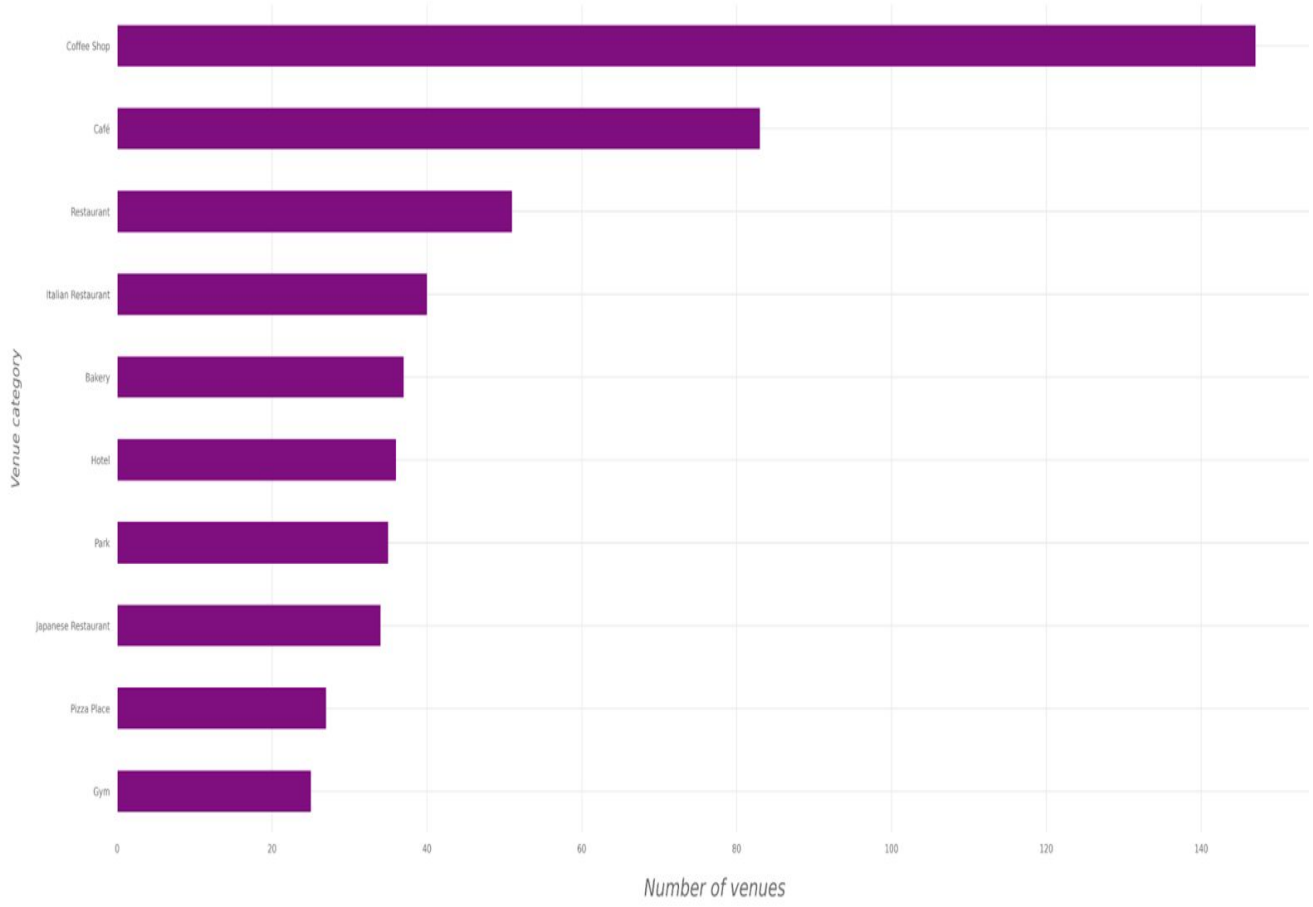
Results

New York: The graph below shows that the most common venues in Brooklyn (including all neighborhoods for Brooklyn) are Pizza Place, Coffee Shop, Bar, Italian Restaurant, Grocery Store, Deli/Bodega, Bakery, Chinese Restaurant, Mexican Restaurant, and Ice Cream Shop. These results imply that if an individual or company thinks about opening up a restaurant in Brooklyn, they should focus on the types of restaurants/shops listed above. This would guarantee that the business has consumers and succeeds.



The results also suggest that in terms of clusters, the most common venues are pizza places for the first cluster and spas and bars for the second cluster. This is more detailed information that can be used by potential investors. While we knew that pizza place was the most popular one in general in Brooklyn, now we know that the first cluster would be better for opening up a pizza place than the second cluster.

Toronto: The graph below shows that the most common venue categories in Toronto are Coffee Shop, Cafe, Restaurant, Italian Restaurant, Bakery, Hotel, Park Japanese Restaurant, Pizza Place, and Gym. As such, an individual/company would be better off opening up a coffee shop of a cafe in Toronto rather than a pizza place or a bar.



The results also suggest that in the very first cluster, coffee shop is the most common venue category, indicating that if an individual is opening up a coffee shop in Toronto, he/she would be better off by doing in the first cluster of neighbourhoods.

Discussion

Results for New York City and Toronto showed us that the two cities are similar in the sense that the most common venues are food places in both, however, residents have different preferences for the

types of foods. For example, New York residents prefer Italian, Chinese, and Mexican cuisines while Toronto residents like Italian and Japanese cuisine more. Thus, if an Italian chef is thinking about opening a restaurant, he/she would succeed in both cities while a Mexican chef would be better off in New York City than in Toronto.

Conclusion

Finally, this project compared and contrasted New York City and Toronto, focusing on Brooklyn borough from New York and on the boroughs of Toronto that have Toronto in them, for example "East Toronto". The results showed which venue categories are most popular in each city. This information can be used by potential investors to figure out what type of restaurant/business should be opened up in New York City and Toronto.