Starting a New Business: Comparing New York and Toronto Neighborhoods

Waad Kahouli

May 2020

I. Introduction

Background:

When it comes to the best cities to live in North America, NYC and Toronto top the list. As the vibrant multicultural melting pots they are, both cities are considered some of the most attractive cities to live, especially for young millennials. In this project, we will explore the differences and similarities in both cities for someone who is interested in starting a new business. At the end of the report, we will recommend what would be the best business of each city.

The first city of interest in New York City. NYC, located in New York state, is the business capital of the United States. Located on the north east coast of the US, the city boasts a vibrant and diverse economy. From Wallstreet to a thriving tourism sector, NYC is the home to many attractions and continuous influx of visitors. However, many people call the city home and find it a very desirable place to live because of all the amenities it offers. NYC, has five boroughs: Manhattan, Queens, Bronx, Brooklyn and Staten Island. Each borough has its own identity and its unique combination of amenities and attractions.

An aerial view of NYC, source: https://www.pods.com/blog/2019/03/nyc-moving-guide-the-new-york-city-boroughs-explained/

The second city of interest is Toronto. Located on the west shore of lake Ontario, Toronto is Canada's most populous city. Much like NYC, Toronto is renowned for its highly diversified economy. From technology, to financial services, to performing arts and tourism, the city has it all. Its business friendliness attracted many entrepreneurs to move to the city and contribute to the local economy. Many people describe Toronto as a great place to live because it has a balance between having all the amenities of a big city yet having a small town feel to it. Similar

to NYC, Toronto has 6 boroughs or districts: East York, Etobicoke, North York, Old Toronto, Scarborough and York.

A Helicopter view of the city of Toronto:https://unsplash.com/photos/qlKaN7eqay8

Problem:

In this project, we will answer the question of what would be the best type business to open in each city. In order to do this, we will first look at the most common businesses in each city. Second, we will also look at the unique businesses combination in each city and finally we will compare both cities for a given type of business of interest.

II. Data

For this project we will use two types of accessible data. First we will use geographical data for both cities. This data encompases data about the boroughs, the neighborhoods and location data (longitude and latitude). Second, we will rely on venue data which is data that will inform us about the type of businesses or attractions available for a given neighborhood.
The combination of these two sets of data will allow us to perform the analysis necessary to determine the recommendation at the end of the report.

1. Neighborhood Data:

The neighborhood data for both cities refer to data that includes the neighborhoods and the boroughs data in addition to latitude and longitude coordinates.

1.1 New York

The data for the city of New York is a JSON file provided in the IBM Applied Data Science capstone Lab. The raw data format is shown in Figure 1. For each neighborhood, we have information about its respective borough and followed by the coordinates. The file also specifies that there are 306 neighborhoods in NY.


```
Out[5]: {'type': 'FeatureCollection',
         'totalFeatures': 306,
         'features': [{'type': 'Feature',
           'id': 'nyu_2451_34572.1',
           'geometry': {'type': 'Point',
            'coordinates': [-73.84720052054902, 40.89470517661]},
           'geometry_name': 'geom',
           'properties': {'name': 'Wakefield',
            'stacked': 1,
            'annoline1': 'Wakefield',
            'annoline2': None,
            'annoline3': None,
            'annoangle': 0.0,
            'borough': 'Bronx',
            'bbox': [-73.84720052054902,
             40.89470517661,
             -73.84720052054902,
             40.89470517661]}},
          {'type': 'Feature',
```


Figure 1. A sniped from the JSON file provided in Week 3 in the Applied Data Science capstone course designed by IBM.

In order to parse the data, the course mentioned provides some useful code in python. Figure 2. Demonstrates a sample of the code that allows to organize the JSON file into a pandas data frame called *neighborhoods* with four columns: Borough, Neighborhood, Latitude, Longitude. The first 5 rows of the dataframe are presented in Figure 3.

```
for data in neighborhoods_data:
    borough = neighborhood_name = data['properties']['borough']
    neighborhood_name = data['properties']['name']

    neighborhood_latlon = data['geometry']['coordinates']
    neighborhood_lat = neighborhood_latlon[1]
    neighborhood_lon = neighborhood_latlon[0]

    neighborhoods = neighborhoods.append({'Borough': borough,
                                          'Neighborhood': neighborhood_name,
                                          'Latitude': neighborhood_lat,
                                          'Longitude': neighborhood_lon}, ignore_index=True)
```

Figure 2. A sample of the code provided the Applied Data Science capstone course to parse the JSON file in order to extract NY neighborhood data.

	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

Figure 3. The first five rows form the 'neighborhoods' dataframe containing NY neighborhood data

In addition, it is always helpful to visualize the data and to verify that the coordinates do reflect the coordinates of NY neighborhoods. Guided by the lab in this course, I was able to generate the map shown in figure 4 using Folium python library. The map shows the neighborhoods in Manhattan, Bronx, Brooklyn, Staten Island and Queen boroughs.

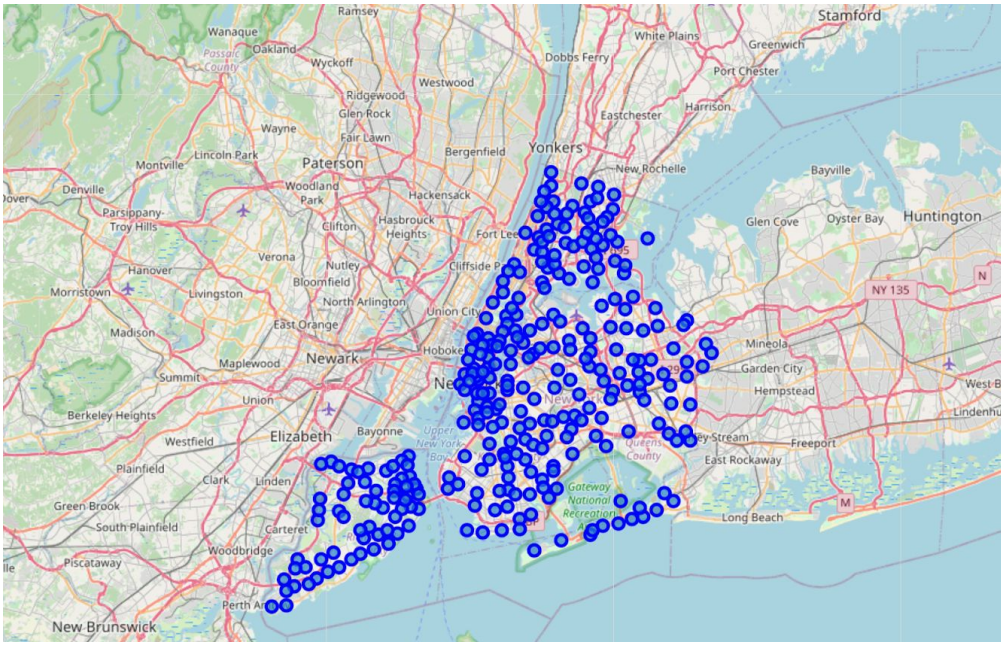

Figure 4: a map of NY neighborhoods (the blue dots) generated using Folium

1.2 Toronto

The first section of the neighborhood data for the city of Toronto was obtained from Wikipedia. The link content focuses mainly on the postal codes and their respective neighborhoods in Toronto and can be found here:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
Figure 5. Shows the format of the data downloaded from the website. The wikitable object was the item of interest. In order to parse the data the bs4 library was used to create a soup.

Afterwards the data was organized into a pandas dataframe called df as shown in figure 6. Note that the postal codes with no boroughs assigned were ignored.

Figure 5. The data format from the Wikipedia website:https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

	Postal Code	Borough	Neighborhood
1	M1A\n	Not assigned\n	\n
2	M2A\n	Not assigned\n	\n
3	M3A\n	North York\n	Parkwoods\n
4	M4A\n	North York\n	Victoria Village\n
5	M5A\n	Downtown Toronto\n	Regent Park, Harbourfront\n
6	M6A\n	North York\n	Lawrence Manor, Lawrence Heights\n
7	M7A\n	Downtown Toronto\n	Queen's Park, Ontario Provincial Government\n
8	M8A\n	Not assigned\n	\n
9	M9A\n	Etobicoke\n	Islington Avenue\n

Figure 6: The first ten rows from the df dataframe used to organize the data from the wikipedia website with the postal codes of the city of Toronto.

The second section of the neighborhood data was given as part of the course week3 Lab material. An alternative method of making API calls using the FourSquare database. However, this method required multiple calls that exceeds the free plan limits. The file had a csv format and was uploaded and stored into a dataframe as shown in Figure 7.

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

Figure 7. The first five rows of the file provided with the geographical coordinates to each postal area code in the city of Toronto.

The final step of this process is to combine the first and second section into a single dataframe. The two dataframes were cleaned and sorted into descending order. The df dataframe was used as the final dataframe to which two columns were added for the geographical coordinates. The final result is shown in Figure 8. Note that this process resulted in 103 neighborhoods and 10 boroughs.

	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	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

Figure 8. The first five rows of the final df dataframe that contains all of Toronto neighborhoods and boroughs.

Similar to NY data visualization, figure 9. Shows a map of the city of Toronto neighborhood using the Folium library.

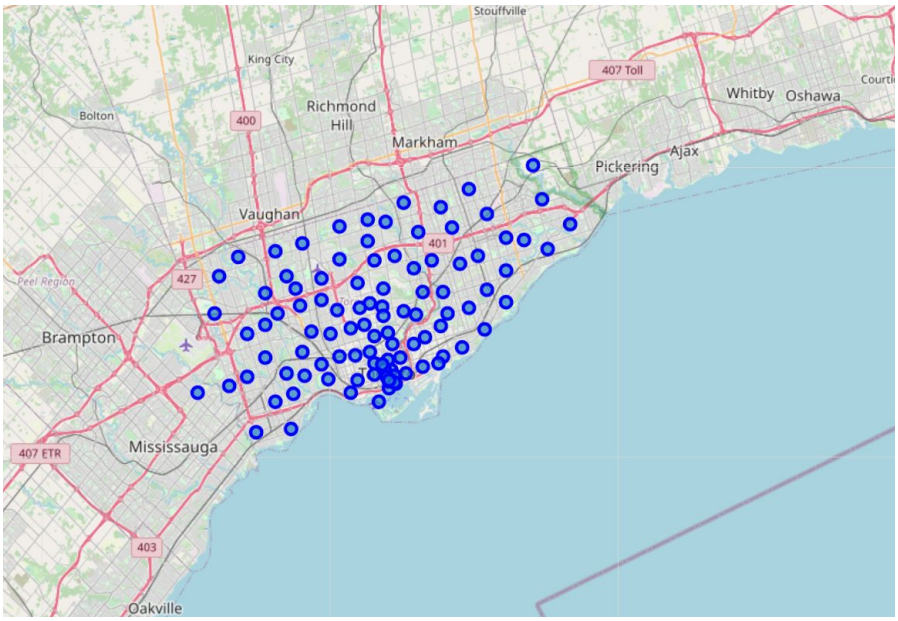

Figure 9. A map of the city of Toronto neighborhoods generated using the Folium library

2. Venue Data

The venues' data was retrieved from the website FourSquare database. Foursquare is a platform that has location and venues data. It is similar to google maps or Yelp. In order to access the data on Foursquare, a developer account must be created to which a client ID and secret is assigned as shown in Figure 10.

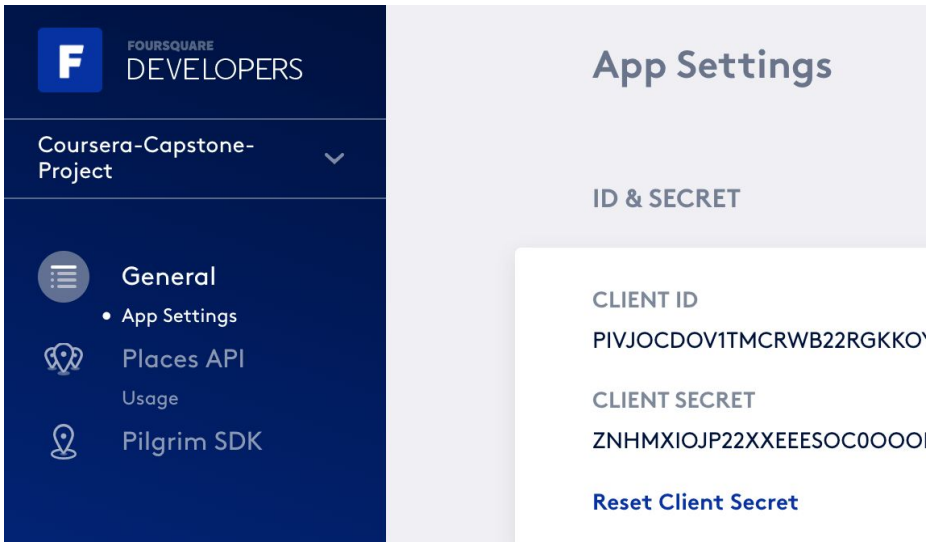

Figure 10: the credential data necessary to make API calls for the Foursquare database.

As an important resource, week3 lab provided a function called getNearbyVenues. Once called, this function will provide all venues for a given latitude and longitude. The function is shown in figure 11.

```
def getNearbyVenues(names, latitudes, longitudes, radius=500, LIMIT=100):

    venues_list=[]
    for name, lat, lng in zip(names, latitudes, longitudes):
        print(name)

        # create the API request URL
        url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll
={},{}&radius={}&limit={}'.format(
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            radius,
            LIMIT)

        # make the GET request
        results = requests.get(url).json()["response"]['groups'][0]['items']

        # return only relevant information for each nearby venue
        venues_list.append([(
            name,
            lat,
            lng,
            v['venue']['name'],
            v['venue']['location']['lat'],
            v['venue']['location']['lng'],
            v['venue']['categories'][0]['name']) for v in results])

    nearby_venues = pd.DataFrame([item for venue_list in venues_list for item in venue_list])
    nearby_venues.columns = ['Neighborhood',
                  'Neighborhood Latitude',
                  'Neighborhood Longitude',
                  'Venue',
                  'Venue Latitude',
                  'Venue Longitude',
                  'Venue Category']

    return(nearby_venues)
```

Figure 11. The getNearbyVenues function provided in the Applied data science Capstone course.

2.1 New York venue data

In order to retrieve the New York city venue data, it is sufficient to make a call using the function above passing as parameters the neighborhoods dataframe and the latitude and longitude columns as shown below:

```
nyc_venues = getNearbyVenues(names=neighborhoods['Neighborhood'],
                             latitudes=neighborhoods['Latitude'],
                             longitudes=neighborhoods['Longitude']
                            )
```

The result of the function call is saved in a dataframe called nyc_venues. There are 9569 venues in New York city. There are many categories of venues and venues such as “bus stations”, “Roads” and “Offices” were eliminated because they are not relevant to the analysis.
Figure 12. Shows the final nyc_venues dataframe. For each venue, we are given the name, the latitude and longitude and category in addition to the neighborhood and its coordinates. For instance, Lollipops Gelato is the name of the venue and its category is Dessert shop. There are 421 different venue categories in NYC.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
2	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
3	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

Fig 12. The New York city venues dataframe.

2.2 Toronto venue data

Similar to the process of retrieving the venues for NYC, the same function was used to extract the venues information for Toronto neighborhoods.

```
to_venues = getNearbyVenues(names=df['Neighborhood'],
                            latitudes=df['Latitude'],
                            longitudes=df['Longitude'])
```

All the data is saved in a dataframe called to_venues. There are 266 types of venue categories in Toronto. Figure 13. Shows the 12 neighborhoods with the most number of venues. Finally, figure 14. Shows the final data frame complete with the all the columns mentioned in the previous section.

Neighborhood	Venue
First Canadian Place, Underground city	100
Commerce Court, Victoria Hotel	100
Harbourfront East, Union Station, Toronto Islands	100
Garden District, Ryerson	100
Toronto Dominion Centre, Design Exchange	100
Stn A PO Boxes	94
Richmond, Adelaide, King	93
St. James Town	77
Church and Wellesley	75
Fairview, Henry Farm, Oriole	65
Central Bay Street	63
Berczy Park	56

Figure 13. The most popular neighborhoods with the highest number of neighborhoods

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Malvern, Rouge	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
2	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	Affordable Toronto Movers	43.787919	-79.162977	Moving Target
3	Guildwood, Morningside, West Hill	43.763573	-79.188711	RBC Royal Bank	43.766790	-79.191151	Bank
4	Guildwood, Morningside, West Hill	43.763573	-79.188711	G & G Electronics	43.765309	-79.191537	Electronics Store

Figure 13. The Toronto venues dataframe.

III. Methodology:

Now that the data for both cities is prepared, an exploratory data analysis was performed. The first step is to use some visualization tools to help us understand the data.

1. Most common venues:

For each city, it is helpful to get a sense for what the most common categories in each city are.

1.1 New York:

Figure 14. Shows a horizontal bar plot of the top 15 most common venues. For instance, a pizza place is the most common venue in New York. In fact, there are 430 pizza restaurants. The second category is coffee shops with a little less than 300 shops, followed closely by the number of italian restaurants. In the fourth rank, we find deli/bodegas which are small grocery stores. Although the number of pizza places is quite high compared to other categories, it makes sense. NYC is known for its famous New York pizza. The city is also known for its diverse restaurants, this explains there are so many italian, chinese and mexican restaurants. Lasly, for NY residents, because of the expensive real estate, bodegas are very important for daily grocery rans.

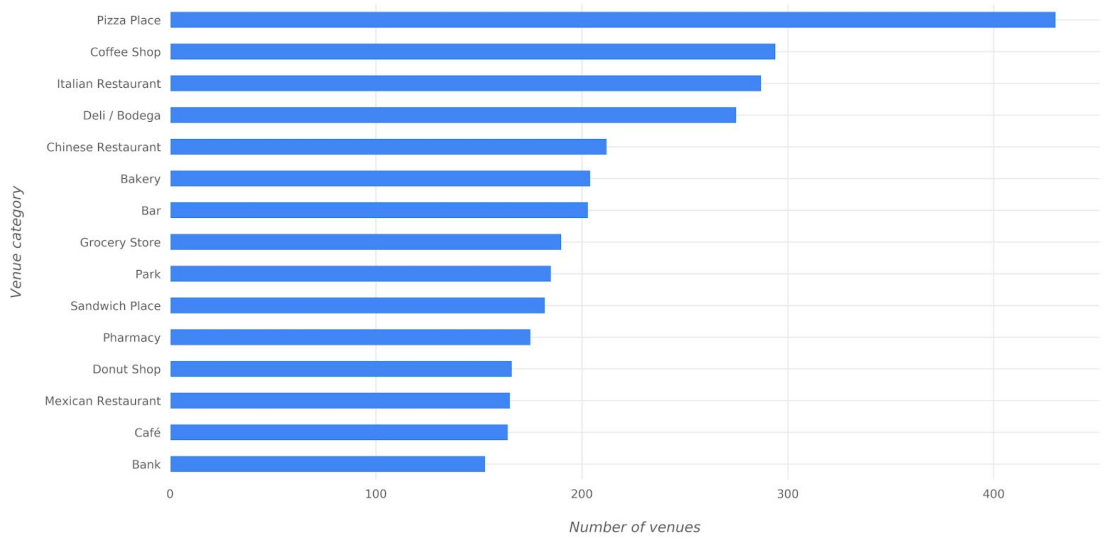

Figure 14. NYC most common venues by category

1.2 Toronto

Performing the same process, we produced figure 15 for the city of Toronto venues. Unlike NYC, the most common venue is coffee shops with over 190 shops. The second category is cafes with over 100 shops. The difference between the two categories is a little confusing but unlike coffee shops, cafes serve full meals. In third place we find parks but in terms of businesses, italian restaurants is the thirst most popular business with 48 restaurants.

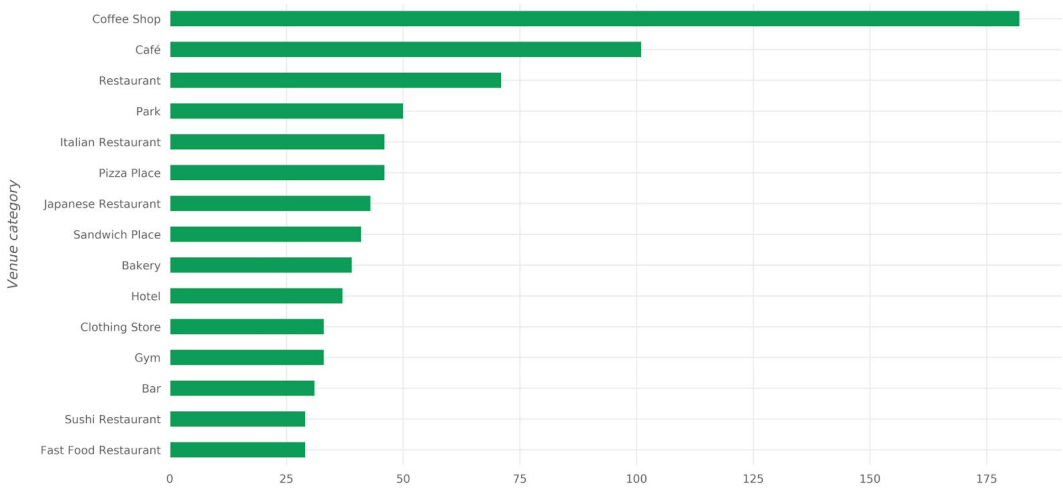

Figure 15. Toronto most common venues by category

2. Clustering the venues:

The goal of this second is to use clustering and some machine learning techniques to find neighborhoods between the two cities. Using in particular K-means clusterings will be used in this analysis. K-means is defined as a method to partition data into clusters with similar properties into clusters . The python library used is sklearn.cluster.

2.1 New York

First step for the clustering process is to perform one-hot encoding on first the NYC data and then the Toronto data and then combine the two. Since clustering can not use values other than numerical then we must perform one-hot encoding to map all the venue data into numerical values. The code to generate the one-hot encoding data frame shown in Figure 17 is presented in figure 16. For example, if we examine the dataframe in Figure 17, we can notice that for the Annandale neighborhood in NYC, there is one American restaurant because the value corresponding to the American restaurant category is non-zero. Similarly, we can also see that Arlinghton neighborhood also has an American restaurant. Note that Figure 17 array is a result of grouping the one-hot encoding NYC venue dataframe by neighborhood and applying a mean function so that we don't have multiple rows for a single neighborhood.

```
# one hot encoding
ny_onehot = pd.get_dummies(nyc_venues[['Venue Category']], prefix="", prefix_sep="")

# add neighborhood column back to dataframe
# we used Neighborhood_" instead of just "Neighborhood" because
# there is a venue category called "Neighborhood"
ny_onehot['Neighborhood'] = nyc_venues['Neighborhood']

# move neighborhood column to the first column
fixed_columns = [ny_onehot.columns[-1]] + list(ny_onehot.columns[:-1])
ny_onehot = ny_onehot[fixed_columns]

ny_onehot.head()
```

Figure 16. Code used to generate a dataframe that organizes the result of one hot encoding of the NYC venue hot-encoding

	Neighborhood	Yoga Studio	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Terminal	American Restaurant	Animal Shelter	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store
0	Allerton	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	Annadale	0.0	0.0	0.0	0.0	0.0	0.0	0.125000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	Arden Heights	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	Arlington	0.0	0.0	0.0	0.0	0.0	0.0	0.333333	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	Arrochar	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 17. Dataframe for the result of applying one-hot encoding on the NYC venues data

2.2 Toronto

Using the same approach for the NYC venue data, the one-hot encoding was applied on the Toronto venue data. In addition, the same method of grouping was employed to obtain the data frame shown in figure 18. Below is the code used:

to_grouped = to_onehot.groupby("Neighborhood").mean().reset_index()

	Neighborhood	Korean Restaurant	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant
0	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
1	Alderwood, Long Branch	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
2	Bathurst Manor, Wilson Heights, Downsview North	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
3	Bayview Village	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
4	Bedford Park, Lawrence Manor East	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.043478	0.0	0.0	0.0	0.0	0.0	0.0

Figure 18. Dataframe for the result of applying one-hot encoding on the Toronto venues data

2.3 Aggregating the NYC and Toronto data

In order to apply clustering we need to merge both cities datasets. During this merger, we must still be able to distinguish the NYC neighborhoods from the Toronto neighborhoods. Hence markers such as "_NYC" and "_Toronto" to make their respective neighborhoods. An additional step to matching the venues categories for both cities was needed. This step required adding missing categories for each city and assigning a value of zero. The result is a data frame that has 408 neighborhoods. Figure 19 shows the top five rows of the merger dataframe. Note that columns now show a mixtures of columns from figure 17 and figure 18.

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Animal Shelter	Antique Shop	Aquarium	Arcade
0	Allerton_NYC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
1	Annadale_NYC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.125000	0.0	0.0	0.0	0.0
2	Arden Heights_NYC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
3	Arlington_NYC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.333333	0.0	0.0	0.0	0.0
4	Arrochar_NYC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0

Figure 19. Resulting dataframe from merging the venue one-hot-encoding data frames of NYC and Toronto.

Now that we combined data from both cities, it is interesting to look at the top most popular venues in both cities. Figure 20 shows that pizza places, coffee shops and italian restaurants are the top three most common venues. However, the data can be skewed due to the fact that we have a much higher number of venues in NYC.

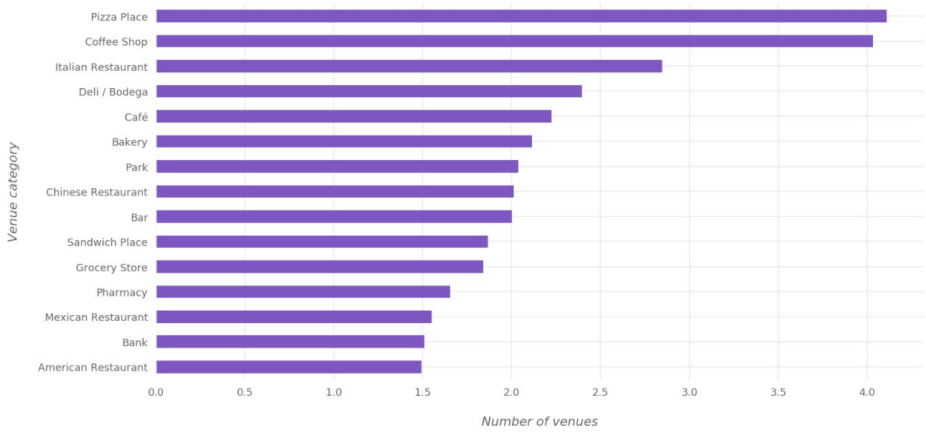

Figure 20. Most common venues in NYC and Toronto

2.4 Clustering results:

Using similar code to that we used for clustering NY and Toronto data in previous labs in the course, k-means was performed on the combined data. The number of clusters was selected to be 5. The code in figure below shows the sequence that results in the clustering process dataframe (shown in Figure 22).

```
# the number of clusters
kclusters = 5

nyc_tor_grouped_clustering = nyc_tor_grouped.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(nyc_tor_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

Figure 21. Code sequence to perform k-means clustering on the NYC and Toronto combined data

The resultant of the clustering operation is five clusters with cluster labels 0, 1, 2, 3, and 4. Each cluster or group contains similar neighborhoods based on the categories of the venues in each neighborhood. The clustering algorithm was performed on 408 neighborhoods from both cities.

Neighborhood	Cluster Labels	1st Most Common Category	2nd Most Common Category	3rd Most Common Category	4th Most Common Category	5th Most Common Category	6th Most Common Category	7th Most Common Category	8th Most Common Category	9th Most Common Category	10th Most Common Category
Agincourt_Toronto	4	Breakfast Spot	Lounge	Skating Rink	Latin American Restaurant	Clothing Store	Dry Cleaner	Duty-free Shop	Eastern European Restaurant	Egyptian Restaurant	Electronics Store
Alderwood, Long Branch_Toronto	0	Pizza Place	Gym	Skating Rink	Athletics & Sports	Sandwich Place	Pharmacy	Coffee Shop	Pub	Pool	Cupcake Shop
Bathurst Manor, Wilson Heights, Downsview North_Toronto	0	Bank	Coffee Shop	Bridal Shop	Deli / Bodega	Gas Station	Chinese Restaurant	Sandwich Place	Diner	Fried Chicken Joint	Pharmacy
Bayview Village_Toronto	0	Chinese Restaurant	Bank	Japanese Restaurant	Café	Yoga Studio	Eye Doctor	Eastern European Restaurant	Egyptian Restaurant	Electronics Store	Empanada Restaurant
Bedford Park, Lawrence Manor East_Toronto	4	Juice Bar	Italian Restaurant	Sandwich Place	Coffee Shop	Restaurant	Liquor Store	Thai Restaurant	Grocery Store	Café	Pharmacy
Berczy Park_Toronto	4	Coffee Shop	Cocktail Bar	Café	Restaurant	Beer Bar	Bakery	Cheese Shop	Seafood Restaurant	Shopping Mall	Pub
Birch Cliff, Cliffside West_Toronto	4	Skating Rink	College Stadium	Café	General Entertainment	Yoga Studio	Exhibit	Duty-free Shop	Eastern European Restaurant	Egyptian Restaurant	Electronics Store
Brockton, Parkdale Village, Exhibition Place_Toronto	4	Café	Coffee Shop	Breakfast Spot	Bakery	Pet Store	Furniture / Home Store	Performing Arts Venue	Restaurant	Climbing Gym	Gym
Business reply mail Processing Centre_Toronto	4	Light Rail Station	Auto Workshop	Smoke Shop	Restaurant	Skate Park	Pizza Place	Garden Center	Burrito Place	Garden	Farmers Market
CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara, Island airport_Toronto	4	Airport Service	Airport Lounge	Airport Terminal	Rental Car Location	Bar	Airport	Airport Food Court	Airport Gate	Coffee Shop	Sculpture Garden

Figure 22: NYC and Toronto neighborhoods, their clusters, and their most common categories

The clustering groups are as follows: 169 under cluster 1, 19 under cluster 2, 3 under cluster 3 and 1 under cluster 4 and finally 202 under cluster 5. Here are examples of clusters.

Neighborhood	Cluster Labels	1st Most Common Category	2nd Most Common Category	3rd Most Common Category	4th Most Common Category	5th Most Common Category	6th Most Common Category	7th Most Common Category	8th Most Common Category	9th Most Common Category	10th Most Common Category
Bayswater_NYC	1	Construction & Landscaping	Park	Playground	Exhibit	Dumpling Restaurant	Duty-free Shop	Eastern European Restaurant	Egyptian Restaurant	Electronics Store	Empanada Restaurant
Bloomfield_NYC	1	Theme Park	Recreation Center	Discount Store	Park	Yoga Studio	Event Space	Dumpling Restaurant	Duty-free Shop	Eastern European Restaurant	Egyptian Restaurant
Clason Point_NYC	1	Park	Boat or Ferry	South American Restaurant	Pool	Grocery Store	Fish & Chips Shop	Financial or Legal Service	Dumpling Restaurant	Duty-free Shop	Eastern European Restaurant
New Brighton_NYC	1	Deli / Bodega	Park	Bowling Alley	Discount Store	Playground	Yoga Studio	Exhibit	Duty-free Shop	Eastern European Restaurant	Egyptian Restaurant
Randall Manor_NYC	1	Pizza Place	Playground	Park	Bagel Shop	Fish & Chips Shop	Event Space	Dry Cleaner	Dumpling Restaurant	Duty-free Shop	Eastern European Restaurant
Riverdale_NYC	1	Park	Bank	Food Truck	Baseball Field	Plaza	Gym	Event Space	Dumpling Restaurant	Duty-free Shop	Eastern European Restaurant
Somerville_NYC	1	Park	Yoga Studio	Food	Dumpling Restaurant	Duty-free Shop	Eastern European Restaurant	Egyptian Restaurant	Electronics Store	Empanada Restaurant	English Restaurant
Todt Hill_NYC	1	Park	Yoga Studio	Food	Dumpling Restaurant	Duty-free Shop	Eastern European Restaurant	Egyptian Restaurant	Electronics Store	Empanada Restaurant	English Restaurant
Caledonia-Fairbanks_Toronto	1	Park	Women's Store	Pool	Eye Doctor	Dumpling Restaurant	Duty-free Shop	Eastern European Restaurant	Egyptian Restaurant	Electronics Store	Empanada Restaurant

Figure 23. Sample of cluster 2 neighborhoods from both NYC and Toronto

Neighborhood	Cluster Labels	1st Most Common Category	2nd Most Common Category	3rd Most Common Category	4th Most Common Category	5th Most Common Category	6th Most Common Category	7th Most Common Category	8th Most Common Category	9th Most Common Category	10th Most Common Category
Butler Manor_NYC	2	Baseball Field	Pool	Convenience Store	Factory	Duty-free Shop	Eastern European Restaurant	Egyptian Restaurant	Electronics Store	Empanada Restaurant	English Restaurant
Mill Island_NYC	2	Locksmith	Pool	Eye Doctor	Dry Cleaner	Dumpling Restaurant	Duty-free Shop	Eastern European Restaurant	Egyptian Restaurant	Electronics Store	Empanada Restaurant
Humberlea, Emery_Toronto	2	Baseball Field	Yoga Studio	Dry Cleaner	Duty-free Shop	Eastern European Restaurant	Egyptian Restaurant	Electronics Store	Empanada Restaurant	English Restaurant	Entertainment Service

Figure 24. Sample of cluster 3 neighborhoods from both NYC and Toronto

IV. Results

In this section, we will discuss some results of the clustering performed on venue data from both the city of New York and Toronto. The following tables show the most common venues in each cluster. All the neighborhoods in the cluster have similarities that resulted in these combinations of venues.

Table 1. Cluster 1

Category	% of venues
Pizza places	7.2
Pharmacy	4.1
Chinese restaurant	3.6
Bank	3.6
Donut shop	3.4
Sandwich place	3.4
Deli/Bodega	3.1

Table 2. Cluster 2

Category	% of venues
Park	38.88
Playground	6.9
Construction/landscaping	4.2
Deli/Bodega	4.2
Bank	2.7
Pool	2.7
Coffee shop	2.7

Table 3. Cluster 3

Category	% of venues
Pool	37.5
Basketball field	37.5

Locksmith	12.5
Convenience store	12.5

Table 4. Cluster 4

Category	% of venues
Business service	100

Table 5. Cluster 5

Category	% of venues
Coffee shop	5.1
Italian restaurant	3.4
Pizza place	2.8
Cafe	2.7
Bar	2.4
Bakery	2.2
Deli/Bodega	2.1

Each cluster in the tables presented has a different combination of venues. For example cluster 1 has a diverse set of businesses that range from Pizza places, Coffee shops and bodegas and sandwich places. It seems that the neighborhoods in this cluster have a commercial dynamism and plenty of places where people go to find a place to eat. Cluster 1 is very similar to cluster 5 where there are also many eateries and places to socialize. This contrasts with cluster 3 with most of the neighborhoods having parks (38%) and playgrounds (7%). These neighborhoods might be more residential and less commercial. Cluster 4 also seems to be less commercial and can possibly be a cluster of residential neighborhoods. Figure 25 represents the number of neighborhoods in each cluster for both cities. Note that most neighborhoods in both cities fall in cluster 1 and 5. This makes sense, due to the population densities of the two cities, there are

many businesses in many neighborhoods versus a central small town layout where most businesses cluster in specific areas (usually near the center).

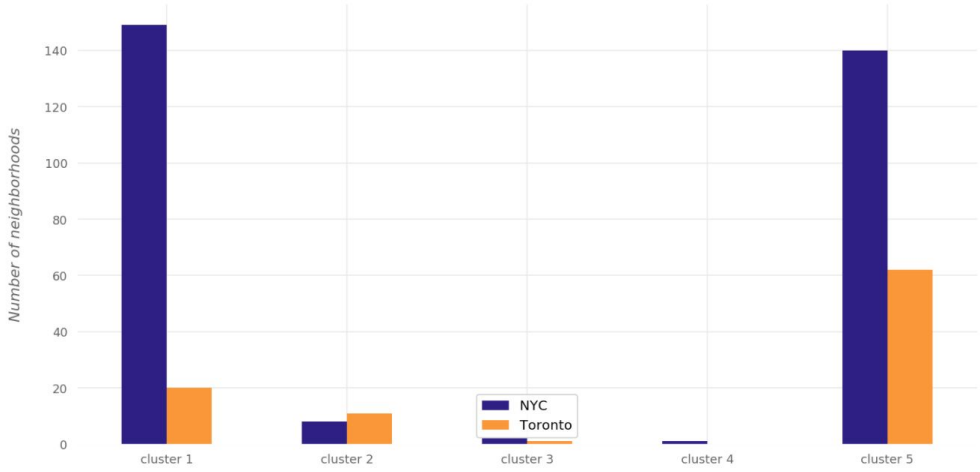

Figure 25. The number of neighborhoods from NYC and Toronto populating each cluster

V. Discussion

Based on the analysis presented in this report, we can issue the following recommendations:

1. If we are interested in opening a business, neighborhoods in cluster 1 and cluster 5 can be good candidates for a location of opening a business such as restaurant, cafe, coffee shop or bar. There is a lot of food traffic.
2. The neighborhoods that have amenities like in cluster 1 and cluster 5 are quite spread out and geographically diverse. Hence, it is possible to find more affordable neighborhoods within these clusters to rent a venue for a business
3. If the business of interest is a restaurant, it is recommended to avoid a pizza place, italian restaurant, american restaurant, mexican restaurant or chinese restaurant. There are plenty of these restaurants in both cities and it will be very difficult to compete with very established restaurants.
4. Thorough analysis of real estate value is needed in order to recommend one neighborhood versus the other.
5. Depending on the client's specifications, NYC or Toronto can be recommended for a place to open a business. Specifications such as location preference, budget and ease of starting a business will help guide the final decision. This is beyond the scope of this project.

VI. Conclusion

In this report, we present our work regarding clustering the neighborhoods from the city of New York, NY and the city of Toronto, ON. The clusters constitue groups that have a similar combination of businesses or venues. The project resulted in a list of neighborhoods where it might be appealing to open a certain business like a restaurant of a less common cuisine or even a new cuisine. Additional analysis of real estate data for both cities and ease of business indices, in addition to client specifications, can help issue a strong recommendation for one city versus the other. [1]

[1] The layout of this work was inspired by Ammar AlYousfi report that can be found in this link https://github.com/ammar1y/Clustering-and-Comparing-the-Neighborhoods-of-New-York-City-and-Toronto/blob/master/Report.pdf