IBM Coursera Data Science Specialization Capstone Project - Clustering and exploring neighbourhoods of New York and Toronto

by Aisha Kala

Introduction

New York City as well as Toronto City are recognised cosmopolitan cities of the world, and as such, they both attract many people who come to visit or to live and work in these cities. For this project I have decided to cluster and explore the neighbourhoods of both New York, USA and Toronto, Canada to compare the two cities in terms of venues such as restaurants, art galleries, parks, coffee shops, café's, and so forth as this information will be useful to people who are either looking to visit any of these cities or people that may intend to relocate to one of these cities.

Background

New York City is one of the most densely populated cities in the United States, with an estimated population of 8 336 817 distributed over 784 square kilometres. New York City has been described as the cultural, financial and media capital of the world and many popular television shows have made New York City a "must-see" destination for people from all over the world. New York is also considered the financial hub of the United States of America. The city comprises of 5 boroughs: Brooklyn, Queens, Manhattan, The Bronx and Staten Island and is situated on one of the world's largest natural harbours. Many districts and landmarks in New York City are well known tourist destinations, attracting 65 million tourists annually which is why I have chosen to use New York City in my project.

Toronto is the capital city of Ontario in Canada, situated in Northern America. The city is also the most populous city in Canada, with an estimated population of 6 139 000 distributed over 630.2 square kilometres. Toronto has been recognised as one of the most multicultural and cosmopolitan cities of the world. Toronto's diverse population is a reflection of its current and historical role as an important destination for immigrants to Canada. Like New York, Toronto is considered the financial and business capital of Canada. The city comprises of 10 boroughs: Etobicoke, North York, Scarborough, York, East York, Central Toronto, East Toronto, West Toronto, Downtown Toronto and old Toronto. Toronto city attracts over 43 million tourists annually and is home to many museums, art galleries, historic sites, festivals and sporting activities.

From the above it can be seen that there are many similarities between New York City and Toronto City. Both of these cities are large, multicultural, diverse and attractive and have many points of interest for residents as well as tourists, and both cities offer potential for entrepreneurs wanting to open a business in either city. The aim of this project is to compare the neighbourhoods of the 2 cities and find neighbourhoods that are similar based on the venues in these neighbourhood's that make them similar.

Target Audience:

- Potential job seekers wanting to relocate to a large, multicultural city
- Tourists intending to visit either one or both of these cities and would like to know which areas would be of interest to them

- Entrepreneurs intending to start a business in either one or both of these cities

Data Acquisition and processing

Neighbourhood data

New York City

Data for the boroughs and neighbourhoods of New York was obtained from the Coursera skills lab as New York was one of the cities explored during the lab (**NYU Spatial Data Repository**). The dataset is in a json file that contains the name of each borough, the neighbourhoods, and the geographical coordinates and so on. In order to utilise the data, it had to be stored in a pandas dataframe.

Load the New York data:

[4]:
```
with open('newyork_data.json') as json_data:
    newyork_data = json.load(json_data)
```

All the relevant neighbourhood data for New York is in the features key, so we have to define a new variable containing this data:

[5]:
```
ny_neighborhoods_data = newyork_data['features']
ny_neighborhoods_data[0]
```

[5]:
```
{'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]}}
```

Figure 1. New York data extraction from json file

Toronto City

Unlike New York, there was no json file for Toronto, therefore the data for the boroughs of Toronto was obtained by scraping a Wikipedia page (**List of Postal Codes of Canada_M.**) This was merged with a csv file provided in the Coursera skills lab containing geographical coordinates and postal codes of the neighbourhoods. The resulting dataframe contained the postal codes, boroughs, neighbourhoods and geographical coordinates for the city.

Create a dataframe containing the Postal Code, Borough, Neighbourhood, Latititude & Longitude in 1 dataframe

```
t_df_merged = pd.merge(toronto_df, toronto_df.cood, on='Postal Code', how='left')
t_df_merged
```

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
5	M9A	Etobicoke	Islington Avenue, Humber Valley Village	43.667856	-79.532242

Figure 2. Toronto dataframe

Venues data: Foursquare location data

Foursquare is a social location service that allows users to explore places around them. Foursquare API provides location based experiences with diverse information about venues, users, photos, and check-ins. Foursquare was utilised to extract venue based information for all of the neighbourhoods in this project, by utilising the geographical coordinates from the dataframes. For this project, the top 5 most common venue data was extracted for both New York and Toronto.

Use Foursquare to get the nearby venues of the neighborhoods in New York with the given location data.

```
def getNearbyVenues(names, latitudes, longitudes, radius=1000):

    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'],
```

Figure 3. Foursquare nearby venues data for New York using given geographical coordinates

Methodology

Exploratory Data Analysis

After obtaining the data for both cities, and combining the data for Toronto from the data scraped off the Wikipedia page as well as the geographical coordinates obtained from the csv file, the shape of the data had to be checked to ensure that all the boroughs and neighbourhoods were present.

Check the number of boroughs and neighborhoods in the dataframe:

```
[7]: # ensure that the dataset contains all 5 boroughs and 306 neighbourhoods:

print('The dataframe has {} boroughs and {} neighborhoods.'.format(
        len(ny_neighborhoods['Borough'].unique()),
        ny_neighborhoods.shape[0]))
```

```
The dataframe has 5 boroughs and 306 neighborhoods.
```

Figure 4. Checking the shape of the New York dataframe

Data pre-processing

For the New York data, the data from the json file was placed into a pandas dataframe. While doing the exploratory data analysis, it was discovered that for New York, certain neighbourhoods had the same name but were in different boroughs. To correct this, the borough name had to be added to these neighbourhood names in order to differentiate them. For Toronto, certain boroughs were not assigned therefore these boroughs were removed. For the neighbourhoods in Toronto that we not assigned, the neighbourhood name was changed from “not assigned” to the respective borough name.

```
[8]: # there are neighborhoods that share the same name but are located in different boroughs:
nyc = ny_neighborhoods['Neighborhood'].value_counts()
nyc[nyc > 1]
```

```
[8]: Bay Terrace    2
Chelsea        2
Murray Hill    2
Sunnyside      2
Name: Neighborhood, dtype: int64
```

```
[11]: # To resolve the above, for neighborhoods that have the same name but are located in different boroughs,
for i in range(ny_neighborhoods.shape[0]):
    nyn_ = ny_neighborhoods.loc[i, 'Neighborhood']
    if ny_neighborhoods[ny_neighborhoods['Neighborhood'] == nyn_].shape[0] > 1:
        ind_ = ny_neighborhoods[ny_neighborhoods['Neighborhood'] == nyn_].index.tolist()
        for j in ind_:
            nyb__ = ny_neighborhoods.loc[j, 'Borough']
            ny_neighborhoods.loc[j, 'Neighborhood'] = nyn_ + ', ' + nyb__

ny_neighborhoods[ny_neighborhoods['Neighborhood'].str.startswith('Bay Terrace')]
```

[11]:

	Borough	Neighborhood	Latitude	Longitude
175	Queens	Bay Terrace, Queens	40.782843	-73.776802
235	Staten Island	Bay Terrace, Staten Island	40.553988	-74.139166

Figure 5. Renaming neighbourhoods in New York that appear in more than 1 borough

Drop the rows where borough = "not assigned" and replace the "not assigned" neighborhood values with the corresponding borough names:

```
toronto_df.drop(toronto_df.loc[toronto_df['Borough']=="Not assigned"].index, inplace=True)


toronto_df=toronto_df.replace('Not assigned', toronto_df['Borough'])
toronto_df
```

	Postal Code	Borough	Neighbourhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront

Figure 6. Removing boroughs in Toronto dataframe that are not assigned

The Foursquare API was utilised to explore both cities and find venues within a given radius, because it is free and easily accessible. Foursquare is one of the world largest sources of location and venue data. In order to use the Foursquare API, a free developer's account needs to be created with Foursquare**.** Then a sample app needs to be created in order to collect the location data. The Client Secret, Client ID and Version is used to obtain the venues that are present in different neighbourhoods. To retrieve the venues and their categories in a given neighbourhood, the geographical coordinates of the neighbourhood are sent in the API request. The resulting venues are placed into a new dataframe which is then combined with the dataframe containing the borough, neighbourhood and geographical coordinates.

One-hot encoding

Machine learning algorithms cannot work with categorical data directly therefore categorical data must be converted to numbers. This is required for both input and output variables that are categorical. This is done using one-hot encoding. One hot encoding is a representation of categorical variables as binary vectors. This first requires that the categorical values be mapped to integer values. Then, each integer value is represented as a binary vector that is all zero values except the index of the integer, which is marked with a 1. One hot encoding allows the representation of categorical data to be more expressive.

In order to explore the venue data and use it for analysis, the foursquare venue data was arranged into a pandas dataframe as follows:

- Creation of a dataframe with pandas one hot encoding for each of the venue categories
- Obtained the mean of each of the one-hot encoded venue category using pandas groupby method on the neighbourhood column
- Used the venue category mean to obtain a venue-based dataframe for each city giving the five most common venues for each neighbourhood

Perform one-hot encoding on the dataframe:

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

# add neighborhood column back to dataframe
toronto_onehot['Neighborhood'] = toronto_venues['Neighborhood']

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

toronto_onehot.head()
```

	Zoo	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Lounge	American Restaurant	Amphitheater	Animal Shelter	Antique Shop	Aquarium	Art Gallery	Art Museum
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 7. Performing one-hot encoding on the Toronto dataframe

Data Visualisation

The neighbourhoods of New York and Toronto were plotted on a map using Folium. Folium is an interactive maps generator package in pandas. Maps were created for New York and Toronto neighbourhoods and also for the clustering of the neighbourhoods. For the mapping on New York and Toronto, geopy was used to obtain the latitude and longitude of both cities and this was subsequently utilised to generate the maps of New York and Toronto City.

Use Geopy to get the latitude & longitude coordinates for Toronto:

```
# Get the latitude & longitude for Toronto using the geolocator:
address = 'Toronto, CA'
geolocator = Nominatim(user_agent="ny_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The geograpical coordinate of Toronto are {}, {}.'.format(latitude, longitude))
```

```
The geograpical coordinate of Toronto are 43.6534817, -79.3839347.
```

Figure 8. Obtaining the geographical coordinates of Toronto using geopy geolocator

Machine Learning – K-means clustering

The neighbourhoods were compared using clustering and segmentation by an unsupervised machine learning technique called k-means clustering. The *k*-means clustering method is an unsupervised machine learning technique used to identify clusters of data objects in a dataset. I chose the k-means clustering technique for this project because it is relatively simple to implement, it generalizes to clusters of different shapes and sizes and it can identify unknown groups in complex datasets. The disadvantages of k-means clustering is that it has a strong sensitivity to outliers and noise and the number of clusters need to be specified beforehand. Sci-Kit Learn is the package used for this algorithm and for this project I chose the number of cluster for both cities to be five. I did not utilise the elbow method or silhouette score to determine the optimum amount of clusters. The output of the clustering is a label for each neighbourhood indicating to which cluster this neighbourhood belongs. These clusters were then visualized using folium.

After clustering, the clusters were explored to find similarities and/or to find the top venues per cluster as well as to determine what sort of category the neighbourhoods in the clusters belong to. Clustering allowed an easier visualisation of the occurrence of various venues in the clusters.

The dataframes for New York and Toronto were then combined and clustering was performed using k-means clustering in order to determine which neighbourhood's in both New York and Toronto fell within the same clusters and what made them similar.

	Postal Code	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	M3A	North York	Parkwoods	43.753259	-79.329656	2	Park	Pharmacy	Shopping Mall	Convenience Store	Café
1	M4A	North York	Victoria Village	43.725882	-79.315572	0	Coffee Shop	Portuguese Restaurant	Intersection	Men's Store	Lounge
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	0	Coffee Shop	Theater	Park	Café	Restaurant
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	0	Clothing Store	Coffee Shop	Restaurant	Fast Food Restaurant	Furniture / Home Store
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	0	Coffee Shop	Park	Café	Sushi Restaurant	Pizza Place

Figure 9. Clustering and cluster labels of Toronto neighbourhoods and the top 5 most common venues per neighbourhood

Results

New York

This city was clustered into 5 clusters.

Cluster 1

This is the largest cluster and from the top 5 venues, it can be seen that this cluster is quite diverse as the venues are many and varied, with neighbourhoods such as Chelsea, Upper East Side, Upper West Side, Tribeca, Soho, Greenwich Village and Brooklyn Heights, to name a few. The most common venues are wide and varied, ranging from coffee shops, Italian and Greek restaurants, and bakeries to art galleries, bookstores, parks, scenic trails, beaches, fitness centres and hotels. There is a wide variety of food places such as Italian, Greek, Chinese, Southern, African, Mexican, French, Turkish and Caribbean. This reflects the melting pot of cultures and backgrounds of the residents in these neighbourhoods. Coffee shops, cocktails bars and cafe's seem to be the most common venue in this cluster, followed by food venues. This cluster would appeal to tourists, visitors and residents alike as there are many things to do and see in these neighbourhoods.

```
[24]: ny_merged.loc[ny_merged['Cluster Labels'] == 0, ny_merged.columns[[1] + list(range(5, ny_merged.shape[1]))]]
```

[24]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
6	Marble Hill	Park	Mexican Restaurant	Pizza Place	Café	Sandwich Place
19	High Bridge	Baseball Stadium	Lounge	Park	Burger Joint	Plaza
24	Hunts Point	Coffee Shop	Food	Park	Nightclub	Grocery Store
46	Bay Ridge	Spa	Pizza Place	Italian Restaurant	Bar	Coffee Shop
49	Greenpoint	Coffee Shop	Cocktail Bar	Bar	Pizza Place	Café
51	Brighton Beach	Beach	Restaurant	Grocery Store	Russian Restaurant	Eastern European Restaurant
54	Flatbush	Pizza Place	Caribbean Restaurant	Mexican Restaurant	Coffee Shop	Bar
55	Crown Heights	Café	Pizza Place	Southern / Soul Food Restaurant	Caribbean Restaurant	Grocery Store
58	Windsor Terrace	Deli / Bodega	Italian Restaurant	Café	Playground	Wine Shop
59	Prospect Heights	Bar	Mexican Restaurant	Wine Shop	Cocktail Bar	Plaza
61	Williamsburg	Pizza Place	Bar	Coffee Shop	American Restaurant	Wine Bar
62	Bushwick	Bar	Coffee Shop	Pizza Place	Cocktail Bar	Mexican Restaurant
63	Bedford Stuyvesant	Coffee Shop	Pizza Place	Bar	Café	Wine Shop

Figure 10. Cluster 1 of New York

Cluster 2

This is quite a small cluster comprising mainly of neighbourhoods at/near the harbour/marina. The top venues are boat/ferry rides, the harbour/marina and food venues such as fast food restaurants. This venue would be an ideal location for residents working in these neighbourhoods or for those wanting to be near the marina.

Cluster 2: Purple Cluster

```
This is quite a small cluster comprising mainly of neighborhoods at/near the harbour/marina. The top venues are boat/ferry rides, the
harbour/marina and food venues such as fast food restaurants. This venue would be an ideal location for residents working in these
neighborhoods or for those wanting to be near the marina.
```

```
[5]: ny_merged.loc[ny_merged['Cluster Labels'] == 1, ny_merged.columns[[1] + list(range(5, ny_merged.shape[1]))]]
```

[5]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
12	City Island	Harbor / Marina	Seafood Restaurant	Boat or Ferry	Italian Restaurant	Bar
27	Clason Point	Park	Gym / Fitness Center	Discount Store	Pool	Zoo
207	Port Ivory	Boat or Ferry	Snack Place	Intersection	Harbor / Marina	Fish Market
227	Arlington	Boat or Ferry	Hardware Store	Department Store	Donut Shop	Fast Food Restaurant
257	Howland Hook	Harbor / Marina	Lounge	Resort	Border Crossing	Peruvian Restaurant

Figure 11. Cluster 2 of New York

Cluster 3

This is the second largest cluster of neighbourhoods in New York. The most common venues appear to be pizza places and Italian restaurants, which may reflect the inhabitants of these neighbourhoods. There are also many bars, deli/bodega's, donut shops, restaurants and banks in this cluster.

```
[26]: ny_merged.loc[ny_merged['Cluster Labels'] == 2, ny_merged.columns[[1] + list(range(5, ny_merged.shape[1]))]]
```

[26]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
3	Fieldston	Bar	Deli / Bodega	Pizza Place	Plaza	Mexican Restaurant
4	Riverdale	Bar	Park	Pizza Place	Bank	Mexican Restaurant
5	Kingsbridge	Bar	Pizza Place	Mexican Restaurant	Sandwich Place	Donut Shop
7	Woodlawn	Pizza Place	Pub	Bar	Deli / Bodega	Bank
10	Baychester	Clothing Store	Donut Shop	Pizza Place	Pharmacy	Department Store
11	Pelham Parkway	Pizza Place	Deli / Bodega	Bakery	Donut Shop	Sandwich Place
13	Bedford Park	Pizza Place	Park	Diner	Donut Shop	Deli / Bodega
16	Fordham	Italian Restaurant	Pizza Place	Spanish Restaurant	Mobile Phone Shop	Diner
21	Mott Haven	Pizza Place	Mexican Restaurant	Donut Shop	Baseball Field	Furniture / Home Store
22	Port Morris	Baseball Field	Donut Shop	Pizza Place	Department Store	Grocery Store
28	Throgs Neck	Italian Restaurant	Deli / Bodega	Donut Shop	Pizza Place	Restaurant
29	Country Club	Sandwich Place	Italian Restaurant	Bank	Bakery	Pizza Place
31	Westchester Square	Pizza Place	Fast Food Restaurant	Donut Shop	Sandwich Place	Bar
32	Van Nest	Pizza Place	Deli / Bodega	Park	Playground	Donut Shop

Figure 12. Cluster 3 of New York

Cluster 4

This is a large cluster comprising many neighbourhoods, with Caribbean restaurants being the most common venue in this cluster. This is not surprising as some of the neighbourhoods in this cluster are Jamaica Center, South Jamaica and Jamaica Estates. The other popular venues in this cluster are pizza places, grocery stores and fast food restaurants.

```
Cluster 4: Green Cluster

This is a large cluster comprising may neighborhoods, with Caribbean restaurants being the most common venue in this cluster. This is not
surpirising as some of the neighborhoods in this cluster are Jamaica Center, South Jamaica and Jamaica Estates. The other popular venues
in this cluster are pizza places, grocery stores and fast food restaurants.

7]: ny_merged.loc[ny_merged['Cluster Labels'] == 3, ny_merged.columns[[1] + list(range(5, ny_merged.shape[1]))]]
```

7]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Wakefield	Pharmacy	Supermarket	Caribbean Restaurant	Fast Food Restaurant	Pizza Place
1	Co-op City	Department Store	Mobile Phone Shop	Pizza Place	Shoe Store	Shopping Mall
2	Eastchester	Caribbean Restaurant	Diner	Pizza Place	Fast Food Restaurant	Shopping Mall
8	Norwood	Pizza Place	Bank	Donut Shop	Mexican Restaurant	Gym
9	Williamsbridge	Pizza Place	Caribbean Restaurant	Supermarket	Bakery	Deli / Bodega
14	University Heights	Grocery Store	Pizza Place	Donut Shop	Fried Chicken Joint	Spanish Restaurant
15	Morris Heights	Deli / Bodega	Grocery Store	Food Truck	Pizza Place	Pharmacy
17	East Tremont	Donut Shop	Pizza Place	Park	Fast Food Restaurant	Zoo Exhibit
18	West Farms	Park	Donut Shop	Pizza Place	Zoo	Supermarket
20	Melrose	Mexican Restaurant	Donut Shop	Pizza Place	Sandwich Place	Grocery Store

Figure 13. Cluster 4 of New York

Cluster 5

This is a small cluster and almost all of these neighbourhoods have a beach as the most common venue. This cluster also includes 3 neighbourhoods with a baseball field as the most common venue, which sets this cluster apart from the other clusters. Neighbourhoods in this cluster would appeal to those who enjoy outdoor activities and the beach. There are also many food venues such as pizza places, fast food joints, Chinese, American, Italian and Polish restaurants, the theatre and bars/pubs.

Cluster 5: Orange Cluster

```
This is a small cluster and almost all of these neighborhoods have a beach as the most common venue. This cluster also includes 3
neighborhoods with a baseball field as the most common venue, which sets this cluster apart from the other clusters. Neighborhoods in
this cluster would appeal to those who enjoy outdoor activites and the beach. There are also many food venues such as pizza places, fast
food joints, Chinese, American, Italian and Polish restaurants, the theater and bars/pubs.
```

```
ny_merged.loc[ny_merged['Cluster Labels'] == 4, ny_merged.columns[[1] + list(range(5, ny_merged.shape[1]))]]
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
85	Sea Gate	Beach	Pharmacy	Donut Shop	Spa	Paper / Office Supplies Store
171	Broad Channel	Metro Station	Chinese Restaurant	Food	Sporting Goods Shop	Pizza Place
172	Breezy Point	Pizza Place	American Restaurant	Beach	Trail	Surf Spot
177	Arverne	Surf Spot	Beach	Deli / Bodega	Donut Shop	Caribbean Restaurant
178	Rockaway Beach	Beach	Bar	Ice Cream Shop	Pharmacy	Boat or Ferry
179	Neponsit	Beach	Pub	Deli / Bodega	Spa	Restaurant
190	Belle Harbor	Beach	Deli / Bodega	Pub	Smoke Shop	Spa
191	Rockaway Park	Beach	Pizza Place	Deli / Bodega	Donut Shop	Bagel Shop
204	South Beach	Beach	Pier	Cafeteria	Food	Park
228	Arrochar	Baseball Field	Beach	Italian Restaurant	Polish Restaurant	Cosmetics Shop

Figure 14. Cluster 5 of New York

Toronto

This city was clustered into 5 clusters.

Cluster 1

This is a small cluster comprising of various venues with the most common venues being the park and food venues such as pizza places, Italian, Chinese, Easter European and Caribbean restaurants. There are also a number of shopping malls, ice skating rinks and pharmacies in this cluster. This venues in this cluster are what you would expect of a metropolitan city, however the number of parks were pleasantly surprising as many cities have very limited numbers of parks.

Cluster 1: Red Cluster

```
This is a small cluster comprising of various venues with the most common venues being the park and food venues such as pizza places,
Italian, Chinese, Easter European and Caribbean restaurants. There are also a number of shopping malls, ice skating rinks and pharmacies
in this cluster. This venues in this cluster are what you would expect of a metropolitan city, however the number of parks were
pleasantly surprising as many cities have very limited numbers of parks.
```

[55]:
```
toronto_merged.loc[toronto_merged['Cluster Labels'] == 0, toronto_merged.columns[[2] + list(range(5, toronto_merged.shape[1]))]]
```

[55]:

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Parkwoods	0	Park	Pharmacy	Shopping Mall	Pizza Place	Grocery Store
5	Islington Avenue, Humber Valley Village	0	Pharmacy	Café	Bank	Golf Course	Shopping Mall
11	West Deane Park, Princess Gardens, Martin Grov...	0	Pizza Place	Park	Hotel	Bank	Fish & Chips Shop
12	Rouge Hill, Port Union, Highland Creek	0	Italian Restaurant	Breakfast Spot	Park	Playground	Burger Joint
16	Humewood-Cedarvale	0	Park	Pizza Place	Convenience Store	Coffee Shop	Field
21	Caledonia-Fairbanks	0	Pharmacy	Park	Pizza Place	Cosmetics Shop	Bank
22	Woburn	0	Park	Coffee Shop	Chinese Restaurant	Pharmacy	Mobile Phone Shop
27	Hillcrest Village	0	Pharmacy	Coffee Shop	Park	Convenience Store	Intersection

Figure 15. Cluster 1 of Toronto

Cluster 2

We will ignore this cluster as it contains only 1 neighbourhood.

Cluster 2: Purple Cluster

```
We will ignore this cluster as it contains only 1 neighborhood.
```

```
56]: toronto_merged.loc[toronto_merged['Cluster Labels'] == 1, toronto_merged.columns[[2] + list(range(5, toronto_merged.shape[1]))]]
```

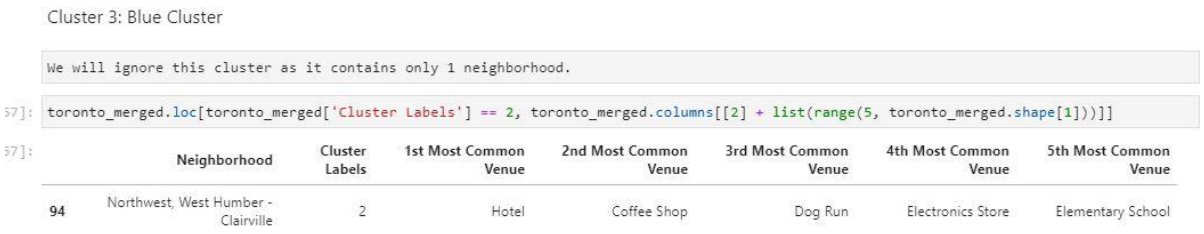
56]:

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
45	York Mills, Silver Hills	1	Park	Pool	Yoga Studio	Farmers Market	Eastern European Restaurant

Figure 16. Cluster 2 of Toronto

Cluster 3

We will ignore this cluster as it contains only 1 neighbourhood.



Cluster 3: Blue Cluster

```
We will ignore this cluster as it contains only 1 neighborhood.
```

```
57]: toronto_merged.loc[toronto_merged['Cluster Labels'] == 2, toronto_merged.columns[[2] + list(range(5, toronto_merged.shape[1]))]]
```

57]:

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
94	Northwest, West Humber - Clairville	2	Hotel	Coffee Shop	Dog Run	Electronics Store	Elementary School

Figure 17. Cluster 3 of Toronto

Cluster 4

This is the second largest cluster for Toronto. There are many different venues in the 5 most common venues. Coffee shops are the most common venue, followed by food venues such as pizza places, followed by fast food restaurants, Chinese, Vietnamese and Korean restaurants. This cluster also contains a brewery, soccer stadium, bowling alley, gastropub and a few parks. It is a diverse cluster and would appeal to a large number of people, both residents and tourists alike.

```
[58]: toronto_merged.loc[toronto_merged['Cluster Labels'] == 3, toronto_merged.columns[[2] + list(range(5, toronto_merged.shape[1]))]]
```

[58]:

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
3	Lawrence Manor, Lawrence Heights	3	Clothing Store	Furniture / Home Store	Coffee Shop	Restaurant	Fast Food Restaurant
6	Malvern, Rouge	3	Coffee Shop	Restaurant	Fast Food Restaurant	Trail	Park
8	Parkview Hill, Woodbine Gardens	3	Pizza Place	Brewery	Soccer Stadium	Gastropub	Bank
10	Glencairn	3	Grocery Store	Pizza Place	Italian Restaurant	Fast Food Restaurant	Gas Station
18	Guildwood, Morningside, West Hill	3	Pizza Place	Fast Food Restaurant	Bank	Restaurant	Food & Drink Shop
23	Leaside	3	Coffee Shop	Sporting Goods Shop	Furniture / Home Store	Burger Joint	Electronics Store
26	Cedarbrae	3	Bakery	Bank	Coffee Shop	Gas Station	Grocery Store
28	Bathurst Manor, Wilson Heights, Downsview North	3	Pizza Place	Park	Coffee Shop	Bank	Pharmacy
29	Thorncliffe Park	3	Coffee Shop	Indian Restaurant	Grocery Store	Sandwich Place	Brewery
32	Scarborough Village	3	Ice Cream Shop	Bowling Alley	Intersection	Convenience Store	Fast Food Restaurant
33	Fairview, Henry Farm, Oriole	3	Clothing Store	Coffee Shop	Sandwich Place	Juice Bar	Restaurant
34	Northwood Park, York University	3	Coffee Shop	Pizza Place	Furniture / Home Store	Fast Food Restaurant	Chinese Restaurant

Figure 18. Cluster 4 of Toronto

Cluster 5

This is the largest cluster of neighbourhoods in Toronto and this cluster includes some of the trendiest neighbourhoods such as Yorkville, Chinatown, Kensington Market and The Annex. Much like New York's trendiest neighbourhoods, this clusters venues are wide and varied, ranging from coffee shops, cafe's, restaurants, pubs and bars to a hockey arena, hotels, the theatre, spas, parks and gyms. This cluster would appeal to tourists and also to those wanting to live in a trendy, hip neighbourhood.

```
59]: toronto_merged.loc[toronto_merged['Cluster Labels'] == 4, toronto_merged.columns[[2] + list(range(5, toronto_merged.shape[1]))]]
```

59]:

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Victoria Village	4	Coffee Shop	Park	Lounge	Hockey Arena	Sporting Goods Shop
2	Regent Park, Harbourfront	4	Coffee Shop	Restaurant	Café	Theater	Park
4	Queen's Park, Ontario Provincial Government	4	Coffee Shop	Sushi Restaurant	Park	Café	Ramen Restaurant
7	Don Mills	4	Restaurant	Japanese Restaurant	Coffee Shop	Gym	Supermarket
13	Don Mills	4	Restaurant	Japanese Restaurant	Coffee Shop	Gym	Supermarket
9	Garden District, Ryerson	4	Coffee Shop	Café	Gastropub	Japanese Restaurant	Diner
14	Woodbine Heights	4	Coffee Shop	Pizza Place	Café	Sandwich Place	Park
15	St. James Town	4	Coffee Shop	Café	Gastropub	Restaurant	Italian Restaurant
17	Eringate, Bloordale Gardens, Old Burnhamthorpe...	4	Coffee Shop	Breakfast Spot	Pet Store	Gas Station	Shopping Mall
19	The Beaches	4	Coffee Shop	Pub	Pizza Place	Breakfast Spot	Beach
20	Berczy Park	4	Coffee Shop	Café	Restaurant	Hotel	Japanese Restaurant
24	Central Bay Street	4	Coffee Shop	Café	Ramen Restaurant	Clothing Store	Sushi Restaurant
25	Christie	4	Korean Restaurant	Café	Coffee Shop	Grocery Store	Pizza Place

Figure 19. Cluster 5 of Toronto

New York and Toronto combined

The combined dataframe was segmented into 5 clusters.

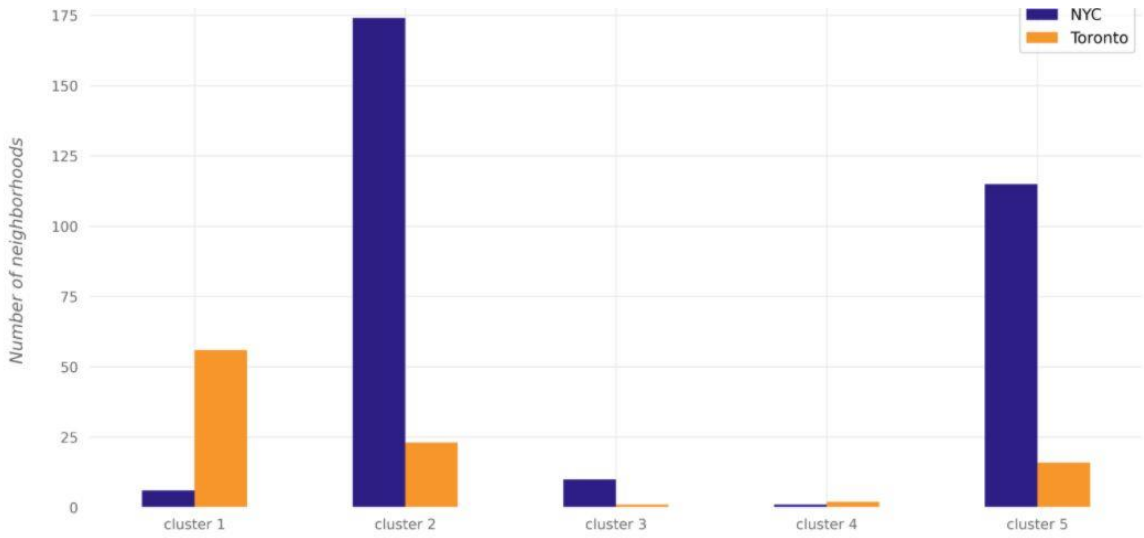


Figure 20. Bar plot of the New York and Toronto combined clusters

Cluster 1

This cluster comprises of a large number of neighbourhoods. The bulk of these appear to be from Toronto. The most common venues in this cluster are coffee shops, followed by cafes and parks. There are also bars, hotels, gyms, a few memorial sites, theatres, a concert hall and shopping malls. The food venues cover a wide range, such as Italian, Chinese, Japanese, Thai, Vietnamese, Middle-Eastern, Mediterranean and Indian restaurants. The neighbourhoods in these clusters would be suitable for those wanting to live in a city wanting to be close to a wide range of venues.

```
c1 = ny_toronto_merged.loc[ny_toronto_merged['Cluster Labels'] == 0, :]
c1
```

Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Arden Heights_NYC	0	Park	Mexican Restaurant	Italian Restaurant	Sandwich Place	Sushi Restaurant
Battery Park City_NYC	0	Coffee Shop	Park	Hotel	Plaza	Memorial Site
Financial District_NYC	0	Coffee Shop	Park	Hotel	Memorial Site	Gym
Hunts Point_NYC	0	Food	Coffee Shop	Park	Juice Bar	Travel & Transport
Long Island City_NYC	0	Coffee Shop	Hotel	Café	Bar	Restaurant
Queensbridge_NYC	0	Hotel	Coffee Shop	Café	Deli / Bodega	Bar
Bayview Village_Toronto	0	Gas Station	Japanese Restaurant	Bank	Grocery Store	Skating Rink
Bedford Park, Lawrence Manor East_Toronto	0	Italian Restaurant	Coffee Shop	Fast Food Restaurant	Bank	Sandwich Place
Berczy Park_Toronto	0	Coffee Shop	Café	Hotel	Restaurant	Park
Birch Cliff, Cliffside West_Toronto	0	Park	Convenience Store	Diner	Thai Restaurant	Café
CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara, Island airport_Toronto	0	Coffee Shop	Café	Harbor / Marina	Garden	Track
Canada Post Gateway Processing Centre_Toronto	0	Coffee Shop	Middle Eastern Restaurant	Chinese Restaurant	Hotel	Burrito Place

Figure 21. Cluster 1 of the combined New York-Toronto data

Cluster 2

This is the largest cluster of neighbourhoods in both New York and Toronto. The bulk of the neighbourhoods in this cluster are from New York. The most common venues are pizza places, followed by donut shops and pharmacies. There are also many stores such as hardware, mobile phones, cosmetics, women's' clothing, discount stores, shopping malls and department stores. This

cluster contains trails, parks, baseball fields, skate parks, golf courses and playgrounds as activity-based venues.

```
c2 = ny_toronto_merged.loc[ny_toronto_merged['Cluster Labels'] == 1, :]
c2
```

Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Allerton_NYC	1	Pizza Place	Donut Shop	Fast Food Restaurant	Pharmacy	Fried Chicken Joint
Annadale_NYC	1	Pizza Place	Trail	Restaurant	Bar	Diner
Arlington_NYC	1	Boat or Ferry	Hardware Store	Snack Place	Intersection	Donut Shop
Auburndale_NYC	1	Korean Restaurant	Greek Restaurant	Pizza Place	Ice Cream Shop	Bank
Bath Beach_NYC	1	Pizza Place	Chinese Restaurant	Japanese Restaurant	Bank	Cantonese Restaurant
Bay Terrace, Queens_NYC	1	Clothing Store	Bank	Women's Store	Cosmetics Shop	Mobile Phone Shop
Bay Terrace, Staten Island_NYC	1	Italian Restaurant	Donut Shop	Supermarket	Food Truck	Spanish Restaurant
Baychester_NYC	1	Clothing Store	Pizza Place	Pharmacy	Donut Shop	Department Store
Bayswater_NYC	1	Park	Construction & Landscaping	Hardware Store	Other Great Outdoors	Grocery Store
Bedford Park_NYC	1	Pizza Place	Park	Diner	Bank	Baseball Field

Figure 22. Cluster 2 of the combined New York-Toronto data

Cluster 3

This cluster contains neighbourhoods in New York and Toronto that have a beach and the most common venue is the beach, followed by food venues. There are many outdoor venues such as surf spots, parks, trails, skate parks, soccer fields and baseball fields and would be ideal for those wanting to visit or reside close to venues such as these and still be in the city.

Cluster 4

This is a very small cluster and will be ignored.

Cluster 5

This cluster contains the trendiest neighbourhoods for both New York and Toronto. It is a large cluster comprising of many neighbourhoods and a wide variety of common venues such as baseball fields, beaches, parks, zoo, arcades, bookstores, gyms, coffee shops, art galleries, a theme park, jazz clubs, golf courses, mountains, marina's and many other venues. This cluster would appeal to tourists and residents who enjoy a wide variety of interests and being in a trendy district.

```
c5 = ny_toronto_merged.loc[ny_toronto_merged['Cluster Labels'] == 4, :]
c5
```

Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Arrochar_NYC	4	Baseball Field	Italian Restaurant	Beach	Deli / Bodega	Taco Place
Astoria_NYC	4	Bar	Coffee Shop	Seafood Restaurant	Middle Eastern Restaurant	Greek Restaurant
Astoria Heights_NYC	4	Airport Service	Rental Car Location	Italian Restaurant	Donut Shop	Park
Bay Ridge_NYC	4	Pizza Place	Spa	Italian Restaurant	Bar	Bagel Shop
Bayside_NYC	4	Pizza Place	Bar	Sandwich Place	Sushi Restaurant	Bakery
Bedford Stuyvesant_NYC	4	Coffee Shop	Bar	Pizza Place	Café	Wine Shop
Belmont_NYC	4	Italian Restaurant	Pizza Place	Deli / Bodega	Bakery	Zoo
Bloomfield_NYC	4	Baseball Field	Arcade	Pizza Place	Burger Joint	Gymnastics Gym
Boerum Hill_NYC	4	Coffee Shop	Bar	Bakery	Grocery Store	Bookstore
Brighton Beach_NYC	4	Beach	Restaurant	Russian Restaurant	Pharmacy	Eastern European Restaurant

Figure 23. Cluster 5 of the combined New York-Toronto data

Discussion

After clustering the neighbourhood's one can easily explore the clusters and from the most common venues determine what sort of cluster it is. Both New York and Toronto were divided into 5 clusters. For this project, I did not utilise the elbow method or silhouette score to determine the prime number of clusters. This may have resulted in better clusters, as some of these clusters were quite small and thus had to be ignored.

As can be seen from the clusters, both New York and Toronto clusters have a wide range of venues as would be expected of a large cosmopolitan city. New York City clusters had more food venues than Toronto city clusters. In comparison, Toronto City clusters had more parks and outdoor activities. This could be a reflection of the priorities of the residents in these cities..

The trendy neighbourhoods in New York had similar venues to the trendy neighbourhoods in Toronto, and these neighbourhoods fell into one cluster when the cities were combined for clustering.

For New York, the overall top 3 venues were pizza places, coffee shops and Italian restaurants.

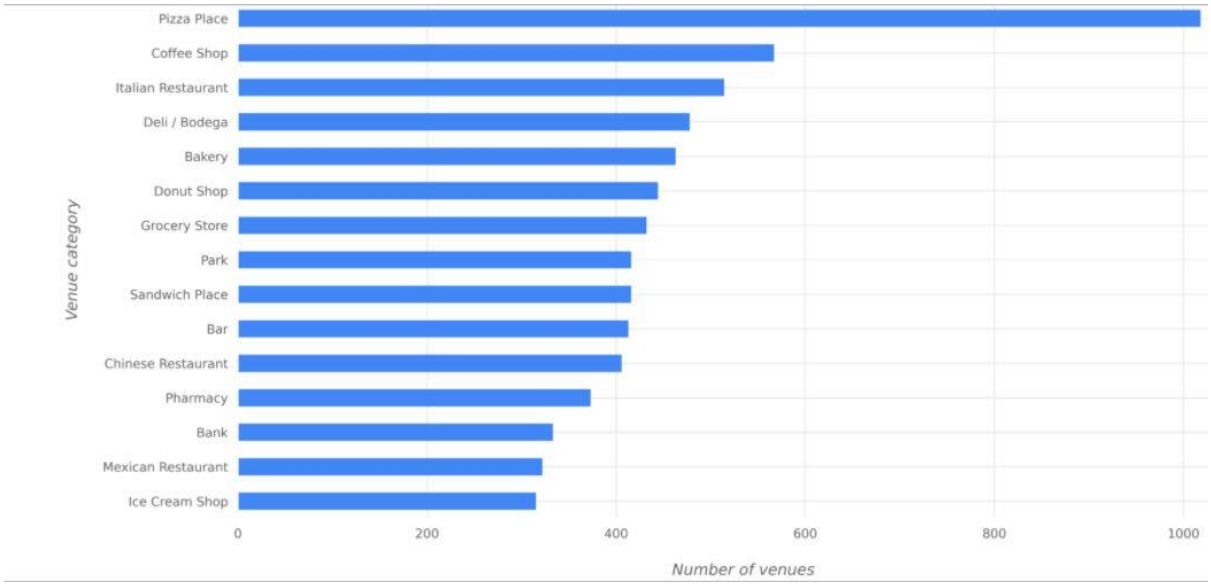


Figure 24. Bar plot of the overall top venues for New York in terms of number of venues

For Toronto, the overall top 3 venues were coffee shops, cafes and parks.

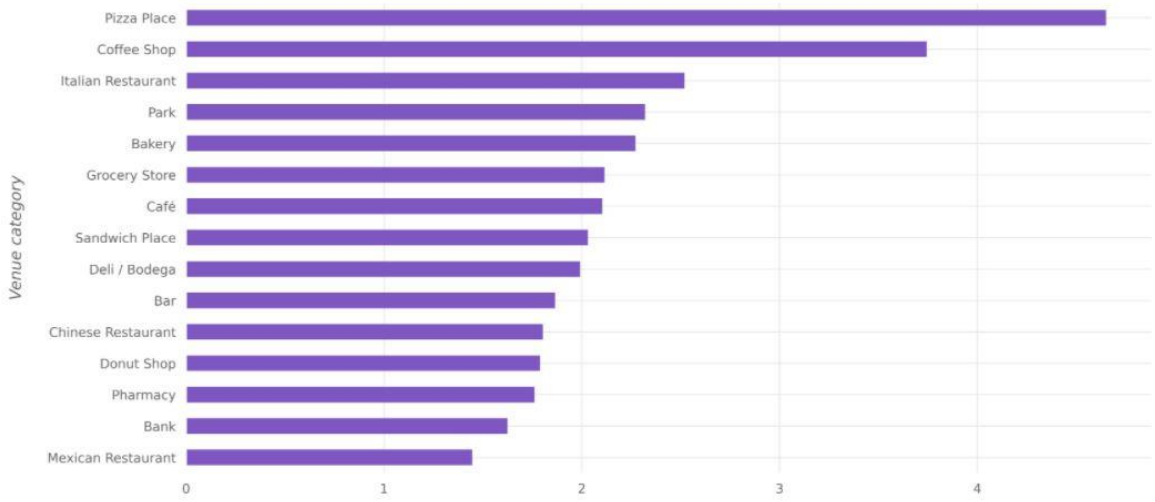


Figure 25. Bar plot of the overall top venues for Toronto in terms of number of venues

For the combined New York and Toronto data, the top 3 venues were pizza places, coffee shops and Italian restaurants.

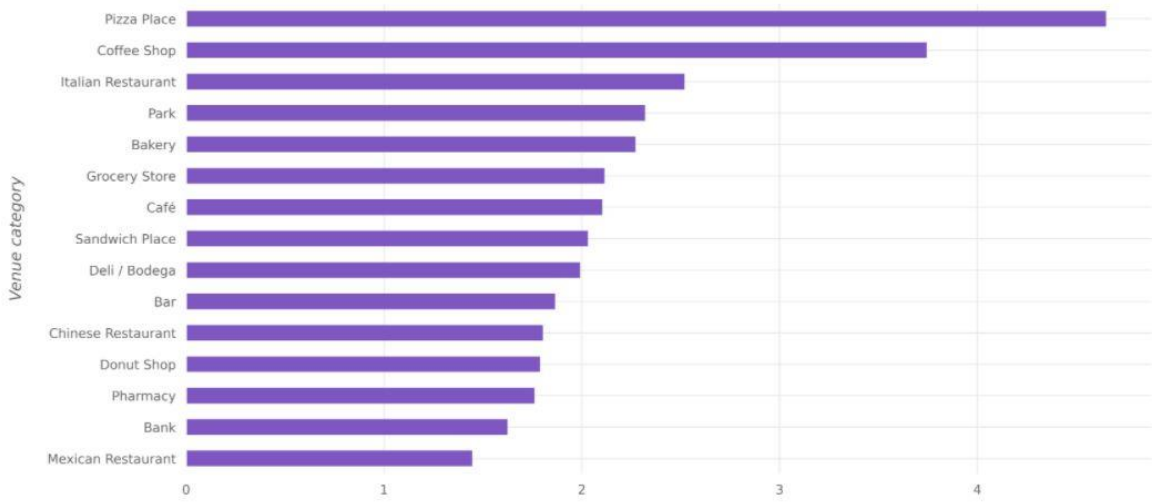


Figure 26. Bar plot of the overall top venues for the combined New York-Toronto data in terms of number of venues

For the combined New York and Toronto dataframe the most common venues for the 5 clusters were as follows:

Cluster 1:

Category	% of venues
Coffee Shop	9.236488
Café	4.232199
Park	3.460109
Restaurant	2.945382
Pizza Place	2.545039
Italian Restaurant	2.516443
Hotel	2.487847

Cluster 2:

Category	% of venues
Pizza Place	6.644486
Donut Shop	3.601135
Pharmacy	3.287993
Deli / Bodega	3.053136
Chinese Restaurant	2.994422
Sandwich Place	2.965065
Bank	2.925922

Cluster 3:

Category	% of venues
Beach	19.512195
Pizza Place	3.963415
Deli / Bodega	3.963415
Surf Spot	3.353659
Pharmacy	3.048780
Donut Shop	3.048780
Bar	2.439024

Cluster 4:

Category	% of venues
Park	47.058824
Pool	11.764706
Gym / Fitness Center	11.764706
Ice Cream Shop	5.882353
Discount Store	5.882353
Shopping Mall	5.882353
Eastern European Restaurant	5.882353

Cluster 5:

Category	% of venues
Coffee Shop	3.967243
Pizza Place	3.494086
Bar	3.039126
Italian Restaurant	2.975432
Bakery	2.793449
Café	2.711556
Park	2.374886

This project did not look at user recommendations for common venues which can be sourced from Foursquare. Another limitation of this project is that the venue data was extracted from Foursquare during COVID-19, which may have affected the data as many places were restricting public outings as well as the number of people allowed per venue, and tourist numbers were quite low due to travel restrictions.

The data in this study only focused on the most common venues per neighbourhood however this project could be broadened to include data such as top 10 places to visit as per data from a platform such a TripAdvisor in order to be more relevant for tourists.

The data could also include average real estate prices per neighbourhood if one would like to compare rent prices of New York with that of Toronto, and this could also extend to include cost of basics such as utilities, groceries, fuel and so forth which would be useful for anyone exploring relocation to any of these cities.

Conclusion:

Analysing, clustering and exploring cities and neighbourhoods of two large cities has revealed a basic idea of the types if venues and activities available in these cities, however this was a very generalised analysis and does not provide comprehensive information to objectively compare both cities. It was interesting to note the similarities and differences in the venues for both cities.

This project provided an understanding of the application of data science principles to a real-life scenario. There are many ways to improve this analysis by exploring further areas such as average income, real estate pricing and cost of living.