Applied Data Science Capstone Project:

Cluster and Segment Neighborhoods in Major Cities for Expats

1. Introduction

As many people leave their home countries to move and work abroad they become expats to that country. To define an expat or expatriate, it is any person who lives temporarily or permanently in a country other than their country of citizenship. Leaving your neighborhood behind and moving to a new neighborhood in a new country can be quiet challenging and someone can still get confused while transitioning into a new culture/social etiquette. Moving into a new country but a similar neighborhood can help expats to quickly adapt to their new environment. Therefore, the approach proposed is to segment and cluster neighborhoods of two major cities. Providing this type of guidance can help expats find a suitable neighborhood in a new country and adjust much faster.

The business problem is clearly addressed to expats and that is the target audience. As an expat, before moving into a new country you are expected to do research. This research would consist of recognizing which neighborhood is most suitable to your needs and life style. The easier you make it for you to settle in, meet people with similar hobbies and start to feel at home, the better.

2. Data

To solve this problem, multiple datasets will be used in combination with the Foursquare location data. Data will be used to cluster and segment neighborhoods in two major cities. The two major cities to be taken as an example are Toronto, Canada and New York City, U.S.

The first step in data collection is to extract the list of Toronto and New York City neighborhoods. Luckily, the datasets exist for free on the web. The New York City neighborhoods dataset is published by the New York (City). Department of City Planning and can be found on geo.nyu.edu website which is a spatial data repository maintained by New York University (NYU). The Toronto neighborhoods dataset can be scraped online from Wikipedia which includes the Postcode, Borough, and Neighborhood name.

Next, the Geocoder library can be used to fetch latitude and longitude coordinates for each of the neighborhoods. Adding the geographical coordinates (latitude and longitude) allows to map these neighborhoods using the folium API. For example, mapping these coordinates provides a better visual to understanding the distribution in each city.

Finally, the Foursquare location API will be used to extract the list of venues surrounding each of the neighborhoods and this list, which contains venues like restaurants/gym/coffee shops/parks, will be used to cluster and segment neighborhoods in Toronto and New York City. The data will be merged, and further analysis will be performed to clean and prepare it for modeling.

Toronto neighborhood data with latitude and longitude:

	Postcode	Borough	Neighbourhood	Latitude	Longitude
0	M4E	East Toronto	The Beaches	43.676357	-79.293031
1	M4K	East Toronto	The Danforth West, Riverdale	43.679557	-79.352188
2	M4L	East Toronto	The Beaches West, India Bazaar	43.668999	-79.315572
3	M4M	East Toronto	Studio District	43.659526	-79.340923
4	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790

Neighborhood data merged with venues data:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	The Beaches	43.676357	-79.293031	Glen Manor Ravine	43.676821	-79.293942	Trail
1	The Beaches	43.676357	-79.293031	The Big Carrot Natural Food Market	43.678879	-79.297734	Health Food Store
2	The Beaches	43.676357	-79.293031	Grover Pub and Grub	43.679181	-79.297215	Pub
3	The Beaches	43.676357	-79.293031	Glen Stewart Ravine	43.676300	-79.294784	Other Great Outdoors
4	The Beaches	43.676357	-79.293031	Upper Beaches	43.680563	-79.292869	Neighborhood

Sample Foursquare response to extract venues list:

```
results = requests.get(url).json()
results
```


```
{'meta': {'code': 200, 'requestId': '5d23a06ba6ec98002c2ccada'},
 'response': {'warning': {'text': "There aren't a lot of results near you. Try som
area."},
  'headerLocation': 'Malvern',
  'headerFullLocation': 'Malvern, Toronto',
  'headerLocationGranularity': 'neighborhood',
  'totalResults': 2,
  'suggestedBounds': {'ne': {'lat': 43.8111863045, 'lng': -79.18812958073042},
   'sw': {'lat': 43.80218629549999, 'lng': -79.2005772192696}},
  'groups': [{'type': 'Recommended Places',
    'name': 'recommended',
    'items': [{'reasons': {'count': 0,
       'items': [{'summary': 'This spot is popular',
         'type': 'general',
         'reasonName': 'globalInteractionReason'}]},
      'venue': {'id': '4bb6b9446edc76b0d771311c',
       'name': "Wendy's",
       'location': {'crossStreet': 'Morningside & Sheppard',
        'lat': 43.80744841934756,
        'lng': -79.19905558052072,
        'labeledLatLngs': [{'label': 'display',
          'lat': 43.80744841934756,
          'lng': -79.19905558052072}],
        'distance': 387,
        'cc': 'CA',
        'city': 'Toronto',
        'state': 'ON',
        'country': 'Canada',
        'formattedAddress': ['Toronto ON', 'Canada']},
```

3. Methodology

In this project we can direct our efforts on detecting areas of Toronto and NYC that have similar common venues, particularly clustering them based on venues categories. We can limit our analysis to the area of Manhattan for NYC and specific areas in Toronto (such as East Toronto, Central Toronto, Downtown Toronto …)

In first step the required data was collected: Extracting the list of Toronto and New York City neighborhoods by scraping the web. Once the datasets are extracted, dataframes were populated accordingly. Further cleaning and data preparation were performed on each of the dataframes to analyze the data generated and yield better results from the clustering model. In this part of data cleaning, the extra steps performed are to ignore cells with a borough that is Not assigned, groupby postcode and combine neighbourhood comma separated, and if a cell has a borough but a Not assigned neighborhood, then the neighborhood will be the same as the borough. Dealing with missing as well as null values was also done in this step. Once the data is clean, we can create a map to better visualize and understand the distribution of these areas. To map the areas, we need to get the coordinates of the neighborhoods. This was done using the Geocoder library to fetch latitude and longitude coordinates for each area.

Second step in our analysis was calculation and exploration of most common venues across different areas of Toronto and NYC - maps from folium API were used to easily identify a few promising areas similar to the expats current home and focus our attention on those areas. The Foursquare location API was used to extract the list of venues surrounding each of the neighborhoods and this list, which contains venues like restaurants/gym/coffee shops/ parks, will be used to cluster and segment neighborhoods in Toronto and New York City.

In third and final step the focus was on the most promising areas and within those create clusters of locations that share similar common venues: Taking into consideration locations with same types of dining options, coffee shops, and gym/parks. Then, a map of all such locations is presented that also creates clusters (using K-means clustering) of those locations to identify general zones / neighborhoods which should be a starting point for final 'street level' exploration and search for optimal venue location.

Group rows by neighborhood and by taking the mean of the frequency of occurrence of each category

```
manhattan_grouped = manhattan_onehot.groupby('Neighborhood').mean().reset_index()
manhattan_grouped
```

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Animal Shelter	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum
0	Battery Park City	0.000000	0.00	0.00	0.00000	0.010000	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.00
1	Carnegie Hill	0.000000	0.00	0.00	0.00000	0.010000	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.01
2	Central Harlem	0.000000	0.00	0.00	0.06383	0.042553	0.00	0.00	0.00	0.000000	0.000000	0.042553	0.00
3	Chelsea	0.000000	0.00	0.00	0.00000	0.030000	0.00	0.01	0.00	0.000000	0.000000	0.030000	0.00
4	Chinatown	0.000000	0.00	0.00	0.00000	0.040000	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.00
5	Civic Center	0.000000	0.00	0.00	0.00000	0.030000	0.00	0.01	0.00	0.000000	0.000000	0.020000	0.00
6	Clinton	0.000000	0.00	0.00	0.00000	0.040000	0.00	0.00	0.00	0.000000	0.000000	0.010000	0.00
7	East Harlem	0.000000	0.00	0.00	0.00000	0.000000	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.00
8	East Village	0.000000	0.00	0.00	0.00000	0.020000	0.00	0.01	0.00	0.020000	0.010000	0.010000	0.00
9	Financial District	0.010000	0.00	0.00	0.00000	0.040000	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.00
10	Flatiron	0.000000	0.00	0.00	0.00000	0.040000	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.00
11	Gramercy	0.000000	0.00	0.00	0.00000	0.040000	0.00	0.00	0.01	0.000000	0.000000	0.010000	0.00

Figure 1: Dataframe contains each neighborhood and the frequency of the venues in that area

From the above dataframe, we can notice the frequency of each venue in a certain neighborhood. We can also extract the top-N most common venues for each neighborhood as follows:

```
----Battery Park City----
           venue  freq
0           Park  0.08
1    Coffee Shop  0.07
2          Hotel  0.05
3            Gym  0.04
4  Memorial Site  0.04


----Carnegie Hill----
                venue  freq
0         Pizza Place  0.06
1         Coffee Shop  0.06
2                Café  0.04
3  Japanese Restaurant  0.03
4    French Restaurant  0.03


----Central Harlem----
                 venue  freq
0    African Restaurant  0.06
1  Gym / Fitness Center  0.04
2           Art Gallery  0.04
3     French Restaurant  0.04
4        Cosmetics Shop  0.04
```

Figure 2: Top 5 most common venues for analysis

The resulting dataframe for top 10 most common venues in the Manhattan area is as shown below:

	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	Battery Park City	Park	Coffee Shop	Hotel	Gym	Memorial Site	Wine Shop	Clothing Store	Italian Restaurant	Department Store	Women's Store
1	Carnegie Hill	Coffee Shop	Pizza Place	Café	Yoga Studio	Bookstore	Cosmetics Shop	French Restaurant	Bar	Japanese Restaurant	Spa
2	Central Harlem	African Restaurant	Art Gallery	Seafood Restaurant	American Restaurant	Gym / Fitness Center	French Restaurant	Cosmetics Shop	Chinese Restaurant	Public Art	Grocery Store
3	Chelsea	Coffee Shop	Italian Restaurant	Ice Cream Shop	Nightclub	Bakery	Seafood Restaurant	American Restaurant	Theater	Art Gallery	Hotel
4	Chinatown	Chinese Restaurant	American Restaurant	Cocktail Bar	Salon / Barbershop	Dim Sum Restaurant	Spa	Vietnamese Restaurant	Dumpling Restaurant	Ice Cream Shop	Bubble Tea Shop

Figure 3: Top 10 most common venues in Manhattan area

The clustering model that was used is K-means clustering. Each of the neighborhoods were grouped into clusters as seen below:

	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	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
0	Manhattan	Marble Hill	40.876551	-73.910660	1	Coffee Shop	Discount Store	Sandwich Place	Yoga Studio	Tennis Stadium	Supplement Shop	Steakhouse	Spa	Seafood Restaurant
1	Manhattan	Chinatown	40.715618	-73.994279	1	Chinese Restaurant	American Restaurant	Cocktail Bar	Salon / Barbershop	Dim Sum Restaurant	Spa	Vietnamese Restaurant	Dumpling Restaurant	Ice Cream Shop
2	Manhattan	Washington Heights	40.851903	-73.936900	0	Café	Mobile Phone Shop	Bakery	Spanish Restaurant	Deli / Bodega	Mexican Restaurant	Sandwich Place	New American Restaurant	Park
3	Manhattan	Inwood	40.867684	-73.921210	0	Mexican Restaurant	Café	Lounge	Bakery	Pizza Place	Park	Frozen Yogurt Shop	Chinese Restaurant	American Restaurant
4	Manhattan	Hamilton Heights	40.823604	-73.949688	0	Deli / Bodega	Café	Mexican Restaurant	Pizza Place	Chinese Restaurant	Coffee Shop	Sushi Restaurant	Caribbean Restaurant	Bank

Figure 4: Dataframe resulting from K-means clustering

4. Results

We can compare the clusters generated from the Toronto and NYC datasets to check for similarities in terms of most common venues. Some of the neighborhoods share the same type of venues. For instance, we can notice that some of the 10 most common venues in one cluster of NYC can also be found in the 10 most common venues in another cluster of Toronto. This is an indicator that the two neighborhoods are similar since they share the same categories of venues. As a new expat in NYC, you can find the same venues (stores, Italian restaurant, gym, park, ...) you would also find in your neighborhood in Toronto. By comparing the list of most common venues for each of the neighborhoods, you can tell which neighborhood is closest to the current one you live in.

The maps below are generated to visually translate the results that were found after running the K-means clustering model:

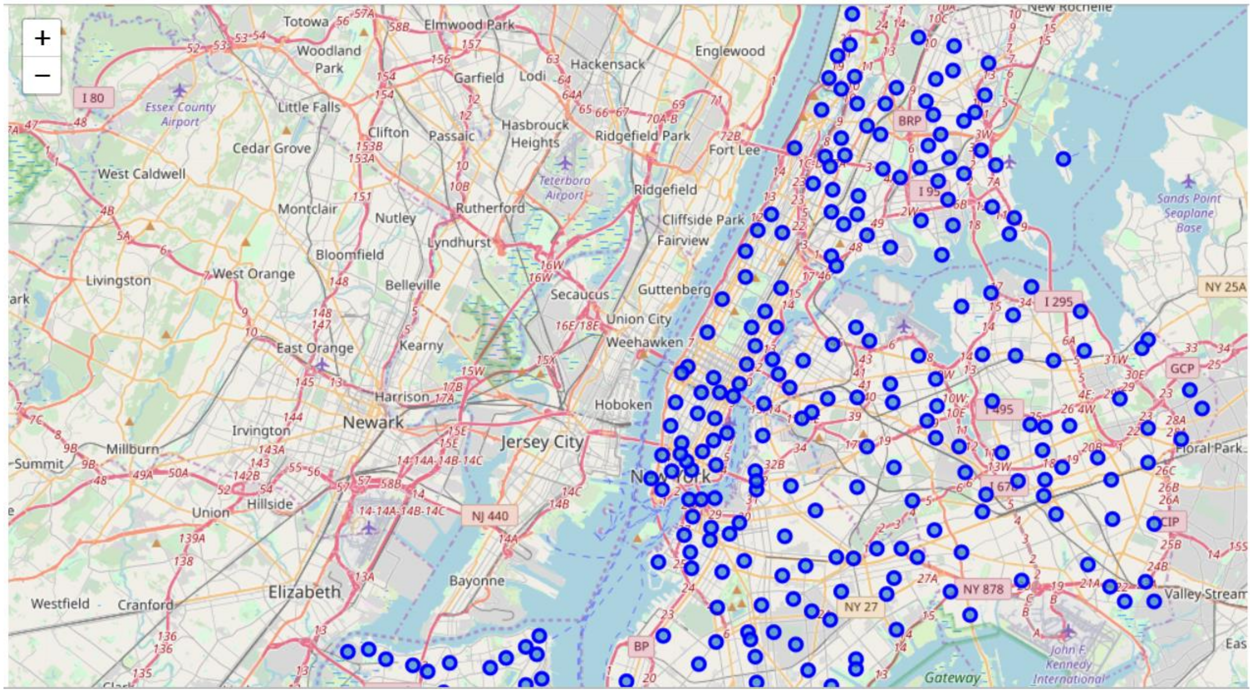

Figure 5: New York City Neighborhoods

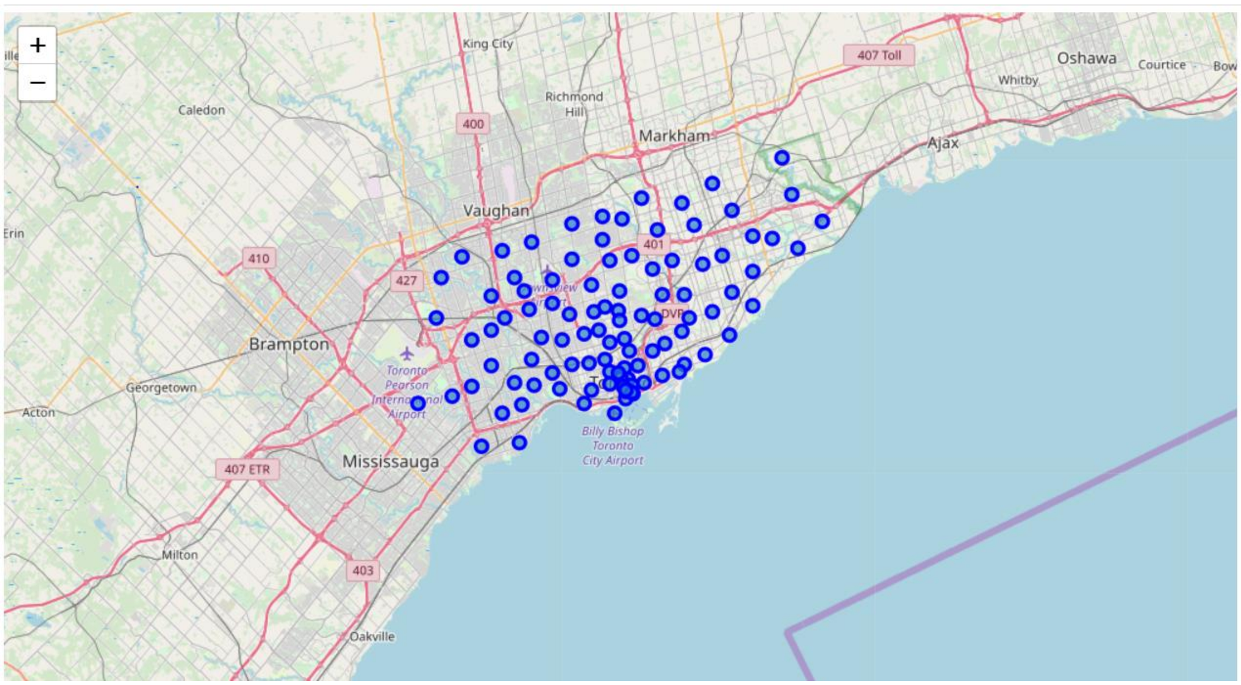

Figure 6: Toronto Neighborhoods

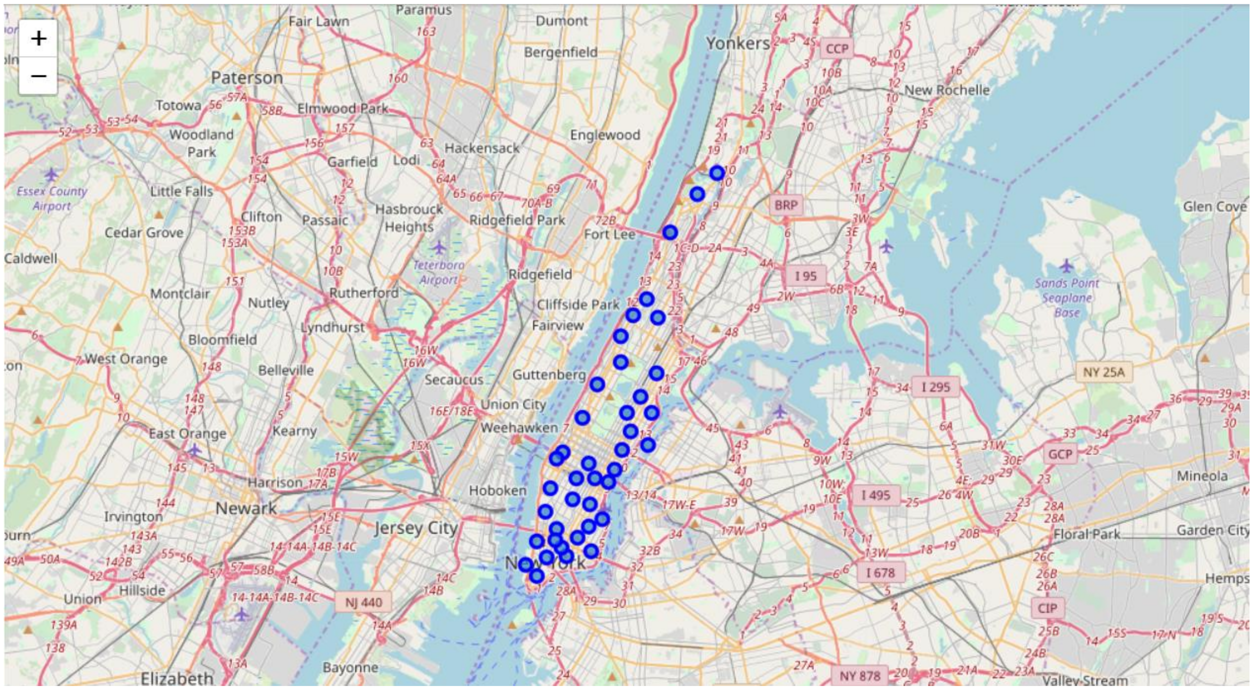

Figure 7: Manhattan Neighborhoods (Only Neighborhoods in Manhattan)

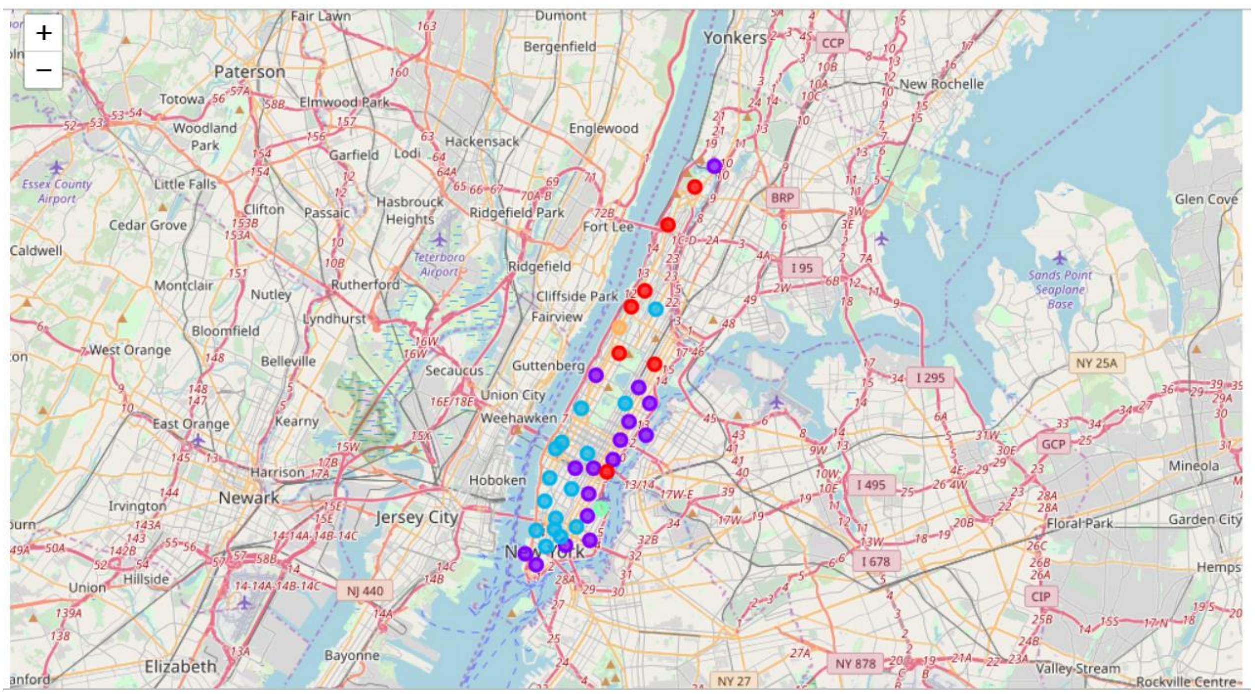

Figure 8: Manhattan Neighborhoods Clustered by Venues Category (Only Neighborhoods in Manhattan are Clustered)

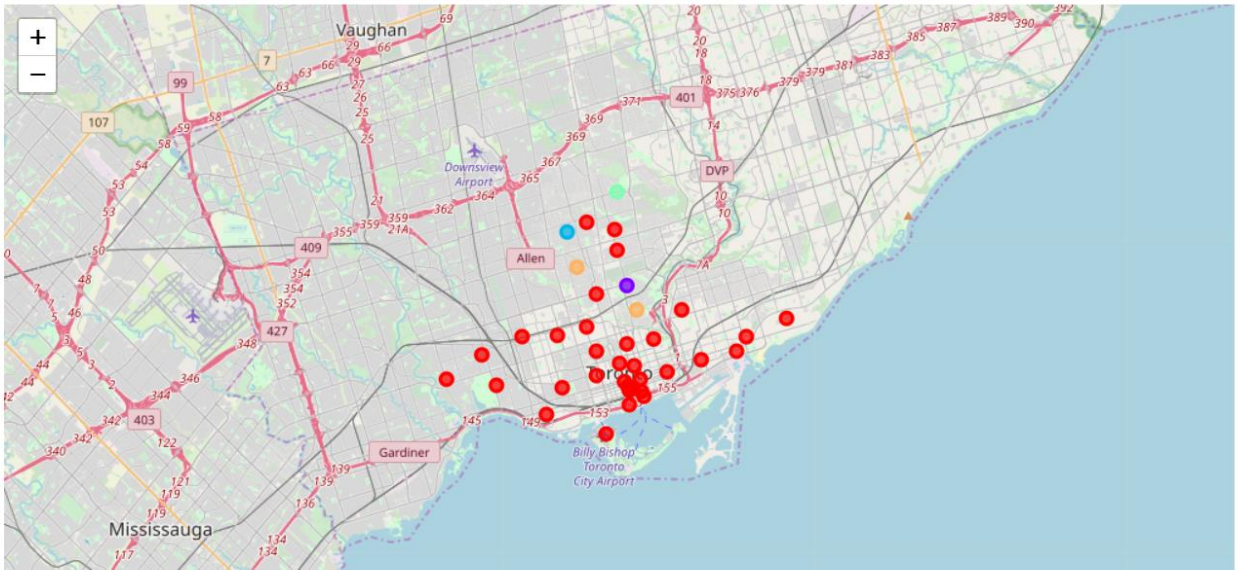

Figure 9: Toronto Neighborhood Clustered (Only Neighborhoods that contain the word Toronto are Clustered)

Cluster 4 **NYC**

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

	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
37	Stuyvesant Town	Bar	Park	Playground	Pet Service	Farmers Market	Baseball Field	Fountain	Harbor / Marina	Cocktail Bar	Coffee Shop

Cluster 5 **NYC**

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

	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
26	Morningside Heights	Park	Bookstore	American Restaurant	Coffee Shop	Food Truck	Burger Joint	New American Restaurant	Tennis Court	Deli / Bodega	College Cafeteria

Figure 10: Example of clusters generated from NYC dataset

5. Discussion

As an expat, deciding on a new location where to live really matters. The approach implemented was to analyze neighborhoods in the home city and cluster them by venues. Then, analyze neighborhoods in the destination city and cluster them by venues. Once we generated the clusters using K-means clustering, we can compare the most common venues in each of the destination city clusters and identify one or more similar clusters to the one I identify with from the home city list of clusters.

The two figures below illustrate the approach employed in solving this problem:

Cluster 5 **Toronto**

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

	Borough	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
50	Downtown Toronto	4	Park	Playground	Trail	Building	Diner	Farmers Market	Falafel Restaurant	Event Space	Ethiopian Restaurant	Electronics Store
64	Central Toronto	4	Trail	Jewelry Store	Park	Sushi Restaurant	Electronics Store	Doner Restaurant	Donut Shop	Dumpling Restaurant	Eastern European Restaurant	Women's Store

Figure 11: Cluster 5 for Toronto

Cluster 1 **NYC**

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

	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
2	Washington Heights	Café	Mobile Phone Shop	Bakery	Spanish Restaurant	Deli / Bodega	Mexican Restaurant	Sandwich Place	New American Restaurant	Park	Supplement Shop
3	Inwood	Mexican Restaurant	Café	Lounge	Bakery	Pizza Place	Park	Frozen Yogurt Shop	Chinese Restaurant	American Restaurant	Wine Bar
4	Hamilton Heights	Deli / Bodega	Café	Mexican Restaurant	Pizza Place	Chinese Restaurant	Coffee Shop	Sushi Restaurant	Caribbean Restaurant	Bank	Bakery
5	Manhattanville	Deli / Bodega	Park	Mexican Restaurant	Coffee Shop	Seafood Restaurant	Italian Restaurant	Ramen Restaurant	Café	Bike Trail	Lounge
7	East Harlem	Mexican Restaurant	Bakery	Deli / Bodega	Thai Restaurant	Latin American Restaurant	Café	French Restaurant	Steakhouse	Spanish Restaurant	Taco Place
25	Manhattan Valley	Indian Restaurant	Coffee Shop	Pizza Place	Yoga Studio	Mexican Restaurant	Café	Bar	Thai Restaurant	Deli / Bodega	Szechuan Restaurant
36	Tudor City	Park	Mexican Restaurant	Café	Greek Restaurant	Asian Restaurant	Deli / Bodega	Pizza Place	Hotel	Dog Run	Spa

Figure 12: Cluster 1 for NYC

After directing our attention to more specifically the two closest clusters that share the most similarities in terms of venues, we were able to narrow down the list of similar neighborhoods between the home city and the destination city.

If we assume that I live in Downtown Toronto, which is the home city, and that belongs to cluster 5 as noted from the figure 7 above. The corresponding cluster 1 of NYC, which is the destination city, is shown in figure 8. By comparing the two clusters, I can identify which neighborhood in NYC is most suited to my life style by looking at the most common venues between my current neighborhood and another neighborhood in NYC from cluster 5.

6. Conclusion

To conclude with, the analysis performed on the Toronto and NYC datasets was used to address the problem expats face when moving to a new country. As mentioned before, when moving to a new country, expats are expected to do research. This research would consist of recognizing which neighborhood is most suitable to their needs and life style. The solution was to extract the neighborhoods for the home city and the destination city to perform clustering taking mainly into account the most common venues in each neighborhood. Neighborhoods were classified into clusters depending on their similarities (in terms of most common venues), then the clusters of each country were compared, and two closest clusters were identified for further analysis. Finally, once we have the similar clusters we can go one level deeper and compare each neighborhood from the two clusters to determine which neighborhood is closest to my current neighborhood.