APPLIED DATA SCIENCE CAPSTONE PROJECT:

CLUSTER AND SEGMENT NEIGHBORHOODS IN MAJOR CITIES FOR EXPATS

INTRODUCTION/BUSINESS PROBLEM:

As many people leave their home countries to move and work abroad they become expats to that country.

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.

Providing some type of guidance can help expats find a suitable neighborhood in a new country and adjust much faster.

The solution is 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.

DATA COLLECTION AND CLEANING

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 New York City neighborhoods dataset is published by the New York (City). Department of City Planning. The Toronto neighborhoods dataset can be scraped online from Wikipedia.

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.

The Foursquare location API will be used to extract the list of venues surrounding each of the neighborhoods.

METHODOLOGY

01

In this project we can direct our efforts on detecting areas of Toronto and NYC that have similar common venues, particularly clustering them.

02

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 while still dealing with missing as well as null values was also done in this step.

03

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.

04

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.

RESULTS

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

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

```
----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
```

We can also extract the top-N most common venues for each neighborhood

RESULTS

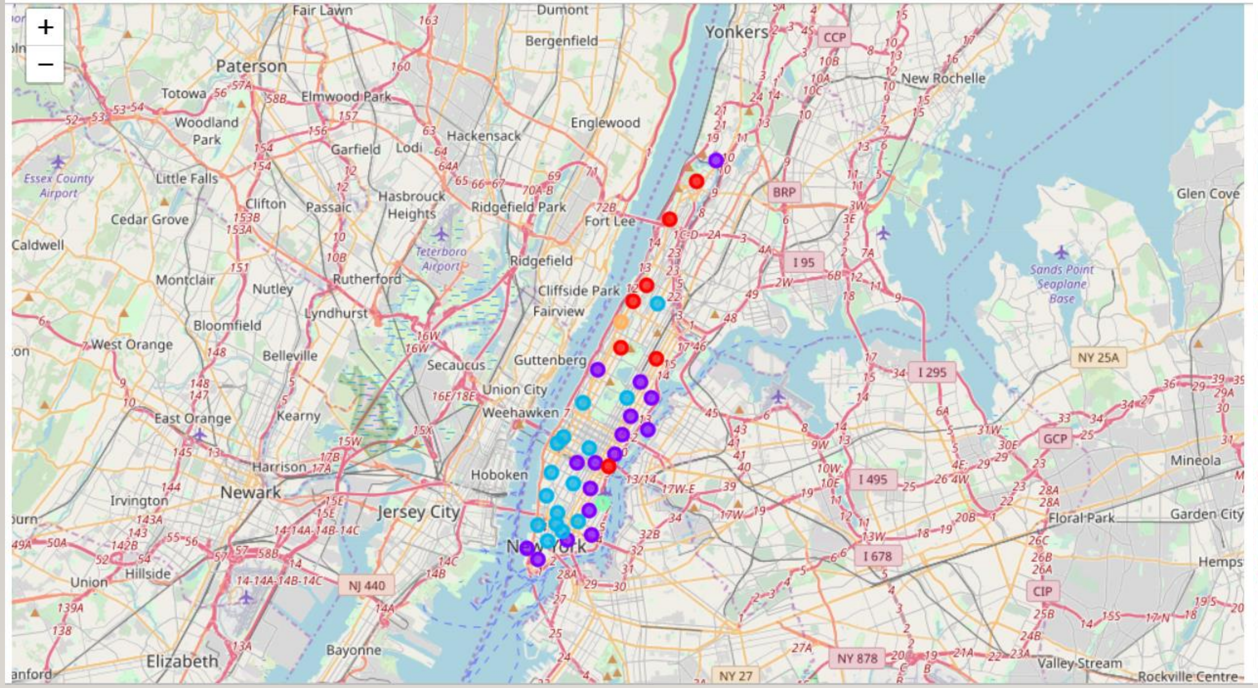
Before clustering

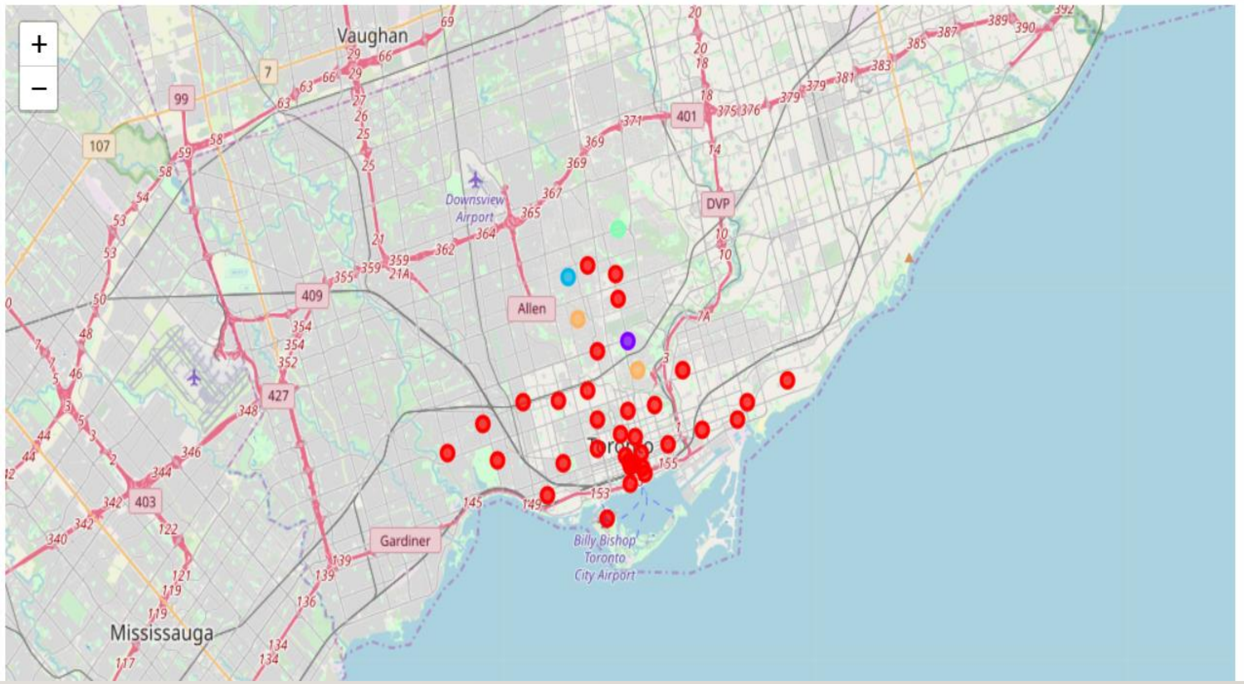
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Battery Park City	Park	Coffee Shop	Hotel	Gym	Memorial Site	Wine Shop	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

After clustering

	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

RESULTS





Manhattan Neighborhoods Clustered by Venues Category

Toronto Neighborhood Clustered

RESULTS

- 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 results.

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

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

CONCLUSION

The analysis performed on the Toronto and NYC datasets was used to address the problem expats face when moving to a new country.

The solution was 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.

Providing this type of guidance can help expats find a suitable neighborhood in a new country and adjust much faster.