Applied Data Science Capstone

Final Assignment

Where to move in?
(Toronto vs NYC neighborhoods)

Julieta Cordara

May, 2020

Table of Contents

1. Discussion and Background of the Business Problem:

1.1. Problem Statement: Make long distance moving an easier decision

In these times of globalization is very usual that people move from one city to another, even from one country to another for professional or academical reasons. For some, this decision is pleasant and a choice, but for some others is hard and stressful. Leaving all the family, friends, traditions behind makes the task much more difficult. However, we could try to make this decision a bit easier by comparing multiple cities and find similarities of their neighborhoods in order to choose the one that will make you feel more comfortable.

Between 1993 and 2013 USA was in the TOP 5 countries of immigration to Canada and Canada in the TOP 15 countries to USA. For this reason, I find it representative to center the analysis in two cities: NYC and Toronto. So, if you currently live in East Toronto, which NYC neighborhood would be the most suitable for you to move in? This is the question we will answer at the end of the analysis, but first let me walk you through all the steps of the project in this long journey from NYC to Toronto or vice versa.

2. Description of the data used for the analysis:

2.1. Data Sources

- Data of migration between Canada and USA has been obtained from the **UN** *'International migration flows to and from selected countries: The 2015 revision.'* (here).
- To get NYC neighborhoods and coordinates I have used from **NYU** (IBM downloaded version) *'2014 New York City Neighborhood Names'* (here).
- From **Wikipedia** *'List of postal codes of Canada: M.'* It contains full list of the neighborhoods around Toronto by Postal Code (here).
- **Geocoder API** was used to get the coordinates of Toronto neighborhoods and NYC
- Lastly **Foursquare** venues and categories APIs will give us the list of venues for each neighborhood in order to make the comparison between them and a list of categories to group the venues.

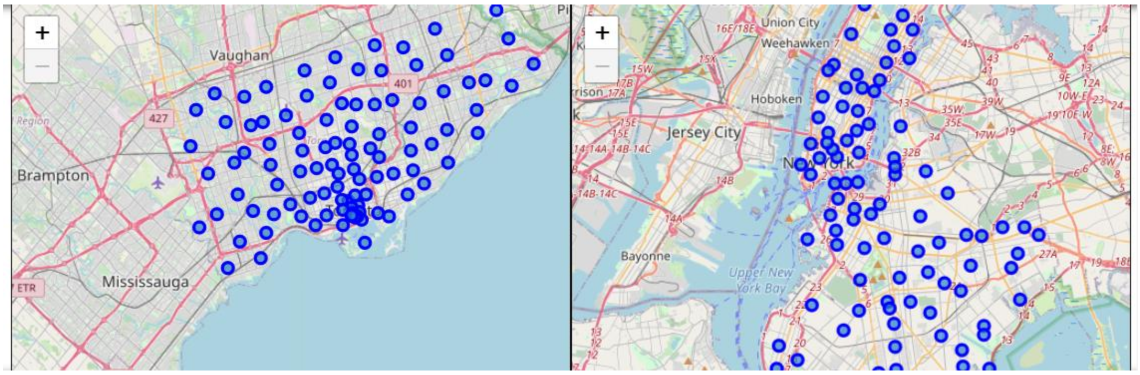
2.2. Data cleaning

First, I had to obtain the coordinates of NYC and Toronto neighborhoods. For that purpose I used *2014 New York City Neighborhood Names*. As there were many neighborhoods in the NYC area, I decided to focus the analysis in Manhattan and Brooklyn boroughs. These boroughs represent almost the same area than Toronto. The total number of neighborhoods obtained were 110.

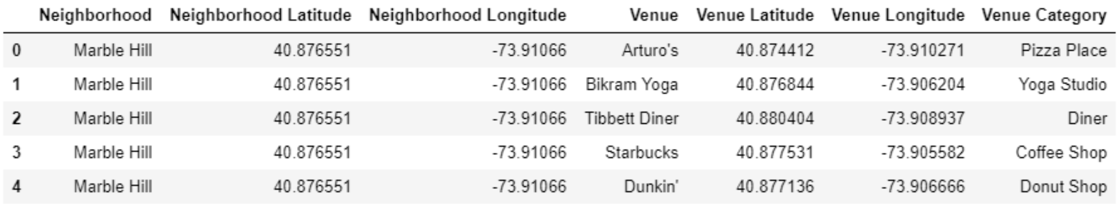
To obtain the Toronto neighborhoods the **Wikipedia** table for Postal Codes was used. Not Assigned Boroughs were excluded, there were some neighborhoods duplicated with different Postal Codes, for those cases the neighborhood and postal code were concatenated in the neighborhood column. The number of neighborhoods obtained for Toronto were 103.

For these I had to do an extra step, because coordinates were not available for Toronto's neighborhoods. That's how I used Geocoder to get the coordinates from the Postal Codes given in the previous list. After that, I merged Toronto's and NYC coordinates list into the same dataset in order to manage altogether.

Let's take a look at the neighborhoods chosen for the analysis

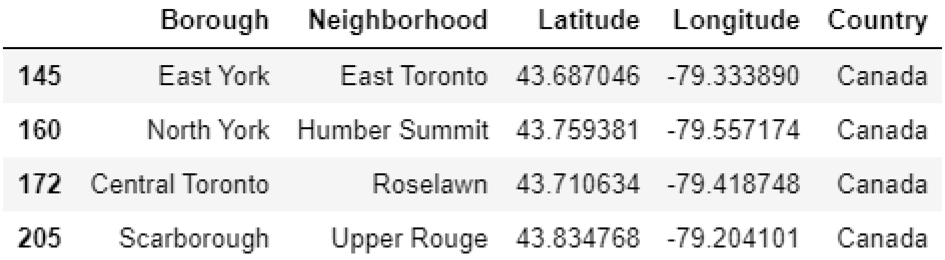


With the list of places obtained in the previous step, I ran the Foursquare API in order to find all the venues located around those places. I got the recommended nearby venues for each location in a radio of 500m and with a limit of 100 venues per spot. 7965 venues were found.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
1	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	Coffee Shop
4	Marble Hill	40.876551	-73.91066	Dunkin'	40.877136	-73.906666	Donut Shop

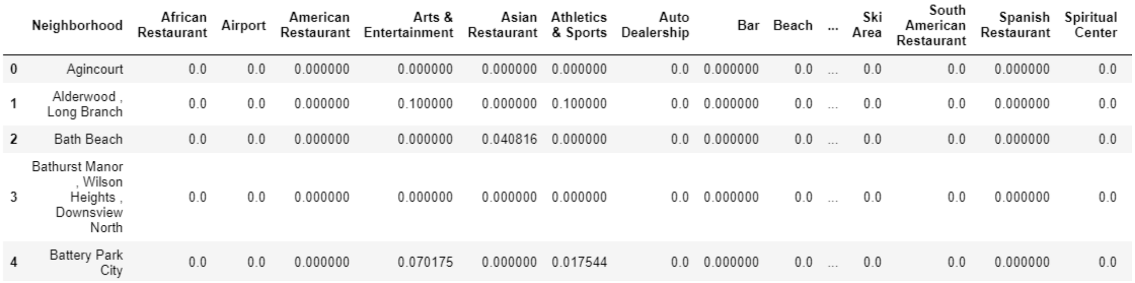
If no venues were found for a neighborhood, we kept it out of scope. For this reason, there were 4 neighborhoods out of scope of analysis:



	Borough	Neighborhood	Latitude	Longitude	Country
145	East York	East Toronto	43.687046	-79.333890	Canada
160	North York	Humber Summit	43.759381	-79.557174	Canada
172	Central Toronto	Roselawn	43.710634	-79.418748	Canada
205	Scarborough	Upper Rouge	43.834768	-79.204101	Canada

After counting the list of distinct venue categories obtained, we got a total of 416. Hence, I tried to group them using the Forquare API Categories. This API returns a hierarchy with the categories available in the current version of the API. For those that matched, I used the 1st category in the hierarchy (for the venue category), for those that were not found, I kept the venue category. After that I had a total of 65 categories.

To normalize the data, the table was converted into one hot encoding and the venues were grouped by neighborhood given the mean of the frequency of occurrence of each category.



	Neighborhood	African Restaurant	Airport	American Restaurant	Arts & Entertainment	Asian Restaurant	Athletics & Sports	Auto Dealership	Bar	Beach	...	Ski Area	South American Restaurant	Spanish Restaurant	Spiritual Center
0	Agincourt	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.0	...	0.0	0.0	0.000000	0.0
1	Alderwood , Long Branch	0.0	0.0	0.000000	0.100000	0.000000	0.100000	0.0	0.000000	0.0	...	0.0	0.0	0.000000	0.0
2	Bath Beach	0.0	0.0	0.000000	0.000000	0.040816	0.000000	0.0	0.000000	0.0	...	0.0	0.0	0.000000	0.0
3	Bathurst Manor , Wilson Heights , Downsview North	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.0	...	0.0	0.0	0.000000	0.0
4	Battery Park City	0.0	0.0	0.000000	0.070175	0.000000	0.017544	0.0	0.000000	0.0	...	0.0	0.0	0.000000	0.0

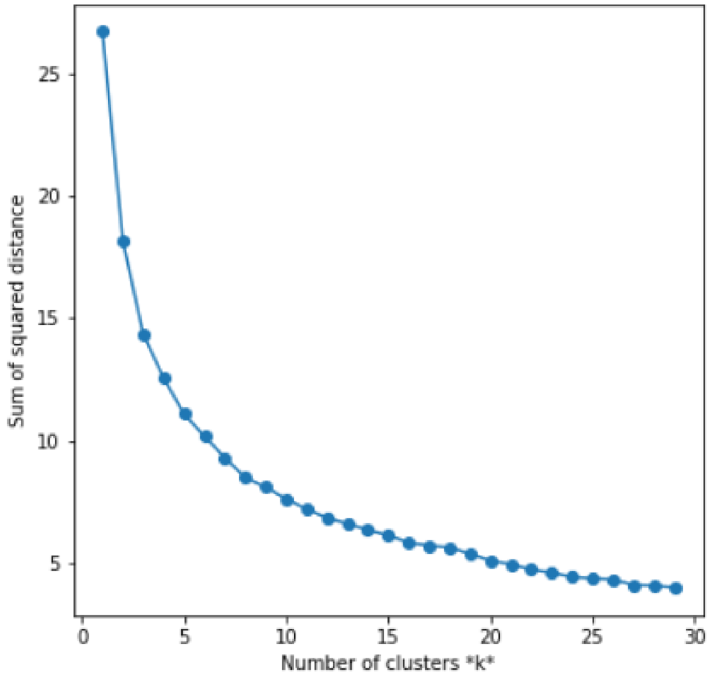
3. Modelling k-means Clusters

For this analysis we only wanted to investigate the structure of the data by grouping the data points into distinct subgroups trying to find similarities in the neighborhoods. Therefore I used an unsupervised learning method.

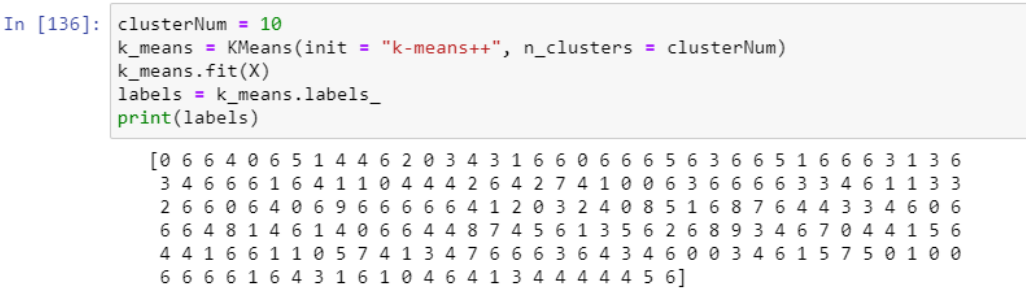
"Clustering is one of the most common exploratory data analysis technique used to get an intuition about the structure of the data. It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different. In other words, we try to find homogeneous subgroups within the data such that data points in each cluster are as similar as possible according to a similarity measure such as euclidean-based distance or correlation-based distance. The decision of which similarity measure to use is application-specific." source In this case, k-means method was chosen.

3.1 Model Evaluation (using Elbow method)

To evaluate the model I used Elbow method, which gives an idea on what a good k number of clusters would be, based on the sum of squared distance (SSE) between data points and their assigned clusters' centroids.



As we can see, it was a bit hard to figure out a good number of clusters to use, because the curve was monotonically decreasing and there was no obvious point where the curve started flattening out. However, it seemed that the number was around 8 and 12. For this reason I applied the k-means cluster method with a k = 10.



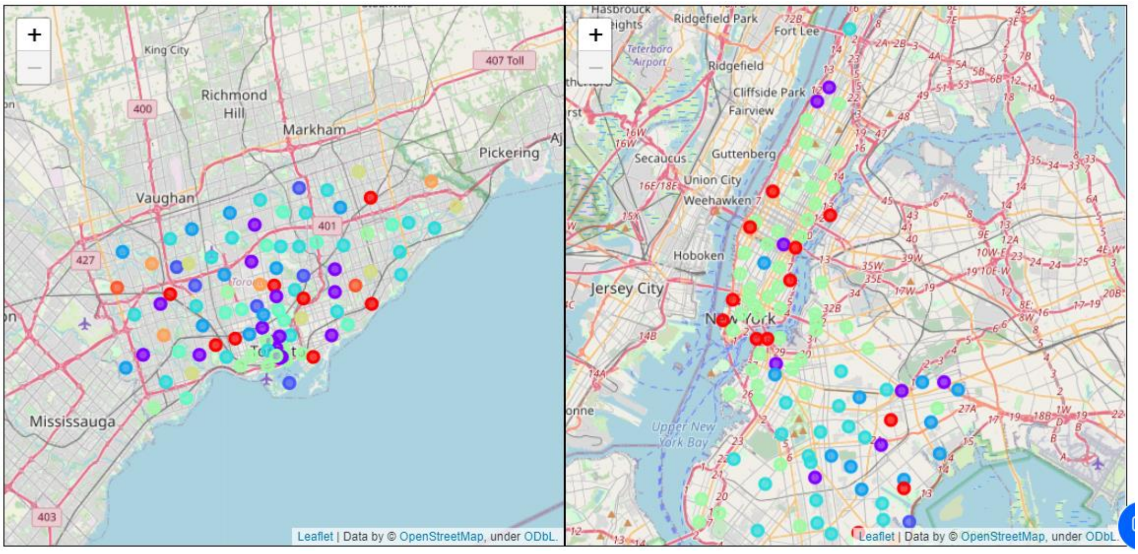
```
In [136]: clusterNum = 10
          k_means = KMeans(init = "k-means++", n_clusters = clusterNum)
          k_means.fit(X)
          labels = k_means.labels_
          print(labels)

[0 6 6 4 0 6 5 1 4 4 6 2 0 3 4 3 1 6 6 0 6 6 6 5 6 3 6 6 5 1 6 6 6 3 1 3 6
 3 4 6 6 6 1 6 4 1 1 0 4 4 4 2 6 4 2 7 4 1 0 0 6 3 6 6 6 6 3 3 4 6 1 1 3 3
 2 6 6 0 6 4 0 6 9 6 6 6 6 6 4 1 2 0 3 2 4 0 8 5 1 6 8 7 6 4 4 3 3 4 6 0 6
 6 6 4 8 1 4 6 1 4 0 6 6 4 4 8 7 4 5 6 1 3 5 6 2 6 8 9 3 4 6 7 0 4 4 1 5 6
 4 4 1 6 6 1 1 0 5 7 4 1 3 4 7 6 6 6 3 6 4 3 4 6 0 0 3 4 6 1 5 7 5 0 1 0 0
 6 6 6 6 1 6 4 3 1 6 1 0 4 6 4 1 3 4 4 4 4 4 5 6]
```

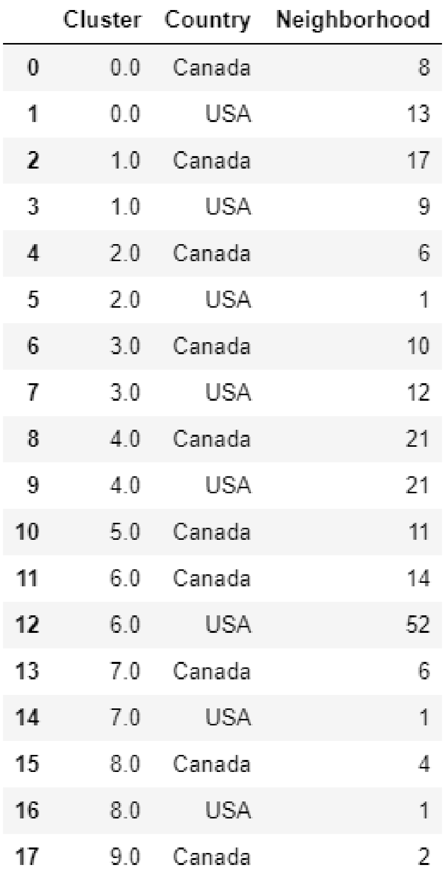
4. Results and Discussion

4.1 Visualize the results

To visualize the resulting clusters in a map, I used folium library. This allowed to see if there were similarities between NYC and Toronto's neighborhoods



Below you can find the list of clusters that were built. Meaning how many neighborhoods of each country where clustered together:



	Cluster	Country	Neighborhood
0	0.0	Canada	8
1	0.0	USA	13
2	1.0	Canada	17
3	1.0	USA	9
4	2.0	Canada	6
5	2.0	USA	1
6	3.0	Canada	10
7	3.0	USA	12
8	4.0	Canada	21
9	4.0	USA	21
10	5.0	Canada	11
11	6.0	Canada	14
12	6.0	USA	52
13	7.0	Canada	6
14	7.0	USA	1
15	8.0	Canada	4
16	8.0	USA	1
17	9.0	Canada	2

4.2 Discussion

According to this analysis we can conclude that if you are planning to move from NYC to Toronto or viceversa:

- Almost always there's a chance to find a neighborhood that resembles to your current
- If you live in Toronto's city centre you might probably want to move to Manhattan and viceversa
- If you live outside Toronto, then you should look for something in Brooklyn
- If you live in Brooklyn, depending on the neighborhood, you could find places inside Toronto or in the outskirts of the city

5. Conclusion

This has been a small glimpse of how real life data-science projects look like. Using some python libraries to scrap web-data, APIs to explore these cities and saw their similar neighborhoods, using k-means algorithm to cluster and Folium leaflet maps to visualize them. The results obtained resembles a lot to what I have expected. As Toronto centre it not as large as Manhattan in terms of density, Manhattan neighborhoods were grouped in fewer clusters. For that reason it might be easier to find a Neighborhood in NYC that resembles to one in Toronto, than the other way around. However, There's a lot of potential and chance for improvements to represent even more realistic comparisons using other kind of data.