Coursera Capstone Project

IBM Applied Data-Science Capstone Project

The Battle of Neighborhoods : Toronto vs New York

By : **Maruf Ansari**
Date : 31st July 2020

The link to my Jupyter Notebook with the coding for this data analysis is available below:

Introduction

Business Problem

Every city is built on different circumstances, environment and different constraints. The aim of this project is to explore the neighborhoods of two major economic capitals, Toronto (CA) and New York City (USA), and group them by common nearby venues. In this project i have compared the neighborhoods of both the cities with respect to places to eat, better connectivity to several useful regions and how they are distributed around both cities. The places I considered are airports, metros, coffee-shops, restaurants, schools, colleges, general stores, hospitals etc.

This information will be very useful for anyone moving to an unknown city, it will help narrow down the list of areas to search for a new home, profitable businesses or trending venues to explore. Thus it will speed up the relocation process and avoid long and pricey stays in hotels or other temporary living arrangements. Along with it, this information could be useful for the tourists visiting between the two cities in deciding the best location for a vacation rental or hotel booking, based on the interests and priorities of the traveler(s).

Target audience of this project

The audience will be anyone visiting, relocating or expanding their businesses between the financial capitals of the two countries to search for better neighborhoods suited for their needs and therefore might offer their preferred range of amenities.

Data

To solve the problem, i will need the following data :

- In order to segment the neighborhoods and explore them, I will essentially need a dataset that contains the boroughs and the neighborhoods exist in each borough.
- It will also need the latitude and longitude coordinates of each neighborhood. This is required to plot the map and to get data about the venues.
- Along with it it will require complete venue data, i.e. venue name, location and category in order to perform clustering of the neighborhoods.

Sources of data and method to extract them :

The data set for Toronto is available at wikipedia. It includes postal-code, boroughs and neighborhood names. Here is the link :
Toronto Data set with postal-codes, boroughs, neighborhoods and their locations
I have scrapped this data from wikipedia into a dataframe. But it needs to clean and prepare this data as it contains some 'Not assigned' boroughs and neighborhoods. Along with it i have to add location coordinates of each neighborhood, i found it on a coursera given link :
Geospatial data of Toronto

The data set for New York was given by coursera. It includes boroughs, neighborhood names and location of each neighborhood. Here is the link :
New York data set
This data set was already cleaned and formatted, so I just loaded it into pandas dataframe.

Resultant Datasets :

	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

```
[15] #Shape of our table
     print('Shape of New York dataframe is {} with {} unique boroughs and {} neighborhoods.'.format(
             newyork_df_new.shape, len(newyork_df_new['Borough'].unique()), newyork_df_new.shape[0] ))
```

```
Shape of New York dataframe is (306, 4) with 5 unique boroughs and 306 neighborhoods.
```

	Borough	Neighborhood	Latitude	Longitude
0	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497
2	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	Scarborough	Woburn	43.770992	-79.216917
4	Scarborough	Cedarbrae	43.773136	-79.239476

```
[8] # Shape of our table
    print('Shape of Toronto dataframe is {} with {} unique boroughs and {} neighborhoods.'.format(
            toronto_df_new.shape, len(toronto_df_new['Borough'].unique()), toronto_df_new.shape[0] ))
```

```
Shape of Toronto dataframe is (103, 4) with 10 unique boroughs and 103 neighborhoods.
```

In between the process I reduced the list of boroughs by removing duplicate values so that there would only be one occurrence of each neighborhood and their corresponding data.
Here I used the Foursquare API to analyze the nearby places to these neighborhoods and see the proximity of important places from the corresponding neighborhoods.

Methodology

Analyze each neighborhood

Firstly I obtained our data and cleaned it and prepared it for further processes and took initial visualization of the neighborhood dataset for both the cities with a folium map in order to understand the layout of the neighborhoods and their location in the respective cities using generated coordinates and i was able to code markers onto each city map of the corresponding neighborhood coordinates using the data from the previously created data frames.
For analyzing the neighborhoods of Toronto and New York, I was able to generate a list of popular/trending places near each neighborhood based on the corresponding map coordinates in the data sets using the explore section of the Foursquare API. I set the radius to 500 and limited the venue results to 100 per neighborhood or set of coordinates. I then transformed this venue data into Pandas data frames (see below example of Toronto neighborhood venue data generated using Foursquare API.) and then merged this info to our dataset. The process was repeated for New York neighborhoods.

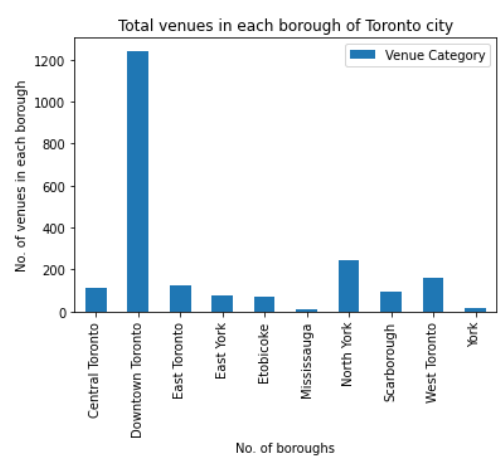
Here there are a total 2152 venues in all the 103 neighborhoods of Toronto city :

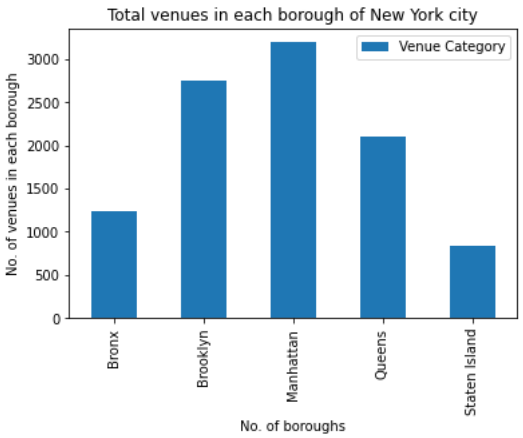
	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue Name	Venue Latitude	Venue Longitude	Venue Category
0	Scarborough	Malvern, Rouge	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	RIGHT WAY TO GOLF	43.785177	-79.161108	Golf Course
2	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
3	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711	RBC Royal Bank	43.766790	-79.191151	Bank
4	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711	G & G Electronics	43.765309	-79.191537	Electronics Store

And total 10125 venues in all 306 neighborhoods of New York city :

	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue Name	Venue Latitude	Venue Longitude	Venue Category
0	Bronx	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Bronx	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
2	Bronx	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
3	Bronx	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
4	Bronx	Wakefield	40.894705	-73.847201	Shell	40.894187	-73.845862	Gas Station

After it, I have generated tables showing the number of venues in each of the boroughs and then visualize them using a bar chart of matplotlib library to get a better idea about it.





Finally i have generated tables for the top 10 most common venues around each of the neighborhoods of both the cities by using one hot encoding and then display them after grouping them on neighborhoods.

Toronto top 10 venues by neighborhood :

(96, 12)

	Borough	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	Central Toronto	Davisville	Pizza Place	Sandwich Place	Dessert Shop	Coffee Shop	Café	Toy / Game Store	Sushi Restaurant	Italian Restaurant	Gym	Japanese Restaurant
1	Central Toronto	Davisville North	Hotel	Food & Drink Shop	Park	Pizza Place	Breakfast Spot	Sandwich Place	Department Store	Yoga Studio	Dim Sum Restaurant	Diner
2	Central Toronto	Forest Hill North & West, Forest Hill Road Park	Trail	Park	Sushi Restaurant	Jewelry Store	Yoga Studio	Diner	Discount Store	Distribution Center	Dog Run	Doner Restaurant
3	Central Toronto	Lawrence Park	Park	Swim School	Bus Line	Yoga Studio	Doner Restaurant	Diner	Discount Store	Distribution Center	Dog Run	Donut Shop
4	Central Toronto	Moore Park, Summerhill East	Gym	Summer Camp	Park	Restaurant	Dim Sum Restaurant	Diner	Discount Store	Distribution Center	Dog Run	Doner Restaurant

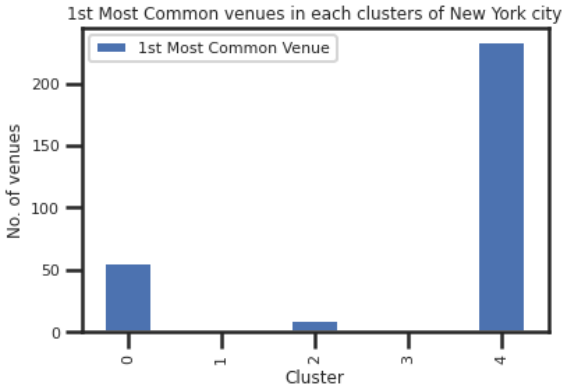
New York top 10 venues by neighborhood :

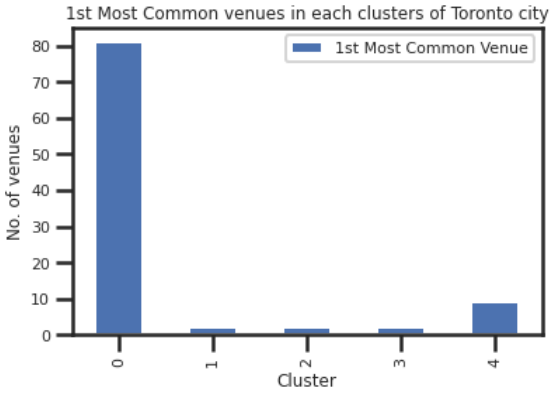
(302, 12)

	Borough	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	Bronx	Allerton	Pizza Place	Deli / Bodega	Supermarket	Check Cashing Service	Mexican Restaurant	Martial Arts Dojo	Fast Food Restaurant	Pharmacy	Used Auto Dealership	Gas Station
1	Bronx	Baychester	Donut Shop	Music Venue	Supermarket	Electronics Store	Pet Store	Fast Food Restaurant	Other Great Outdoors	Bank	Pizza Place	Mexican Restaurant
2	Bronx	Bedford Park	Chinese Restaurant	Diner	Deli / Bodega	Pizza Place	Mexican Restaurant	Sandwich Place	Fried Chicken Joint	Spanish Restaurant	Supermarket	Train Station
3	Bronx	Belmont	Italian Restaurant	Pizza Place	Deli / Bodega	Bakery	Bank	Liquor Store	Grocery Store	Donut Shop	Dessert Shop	Mexican Restaurant
4	Bronx	Bronxdale	Chinese Restaurant	Italian Restaurant	Bank	Pizza Place	Eastern European Restaurant	Performing Arts Venue	Paper / Office Supplies Store	Spanish Restaurant	Breakfast Spot	Supermarket

Cluster Neighborhood

I imported the scikit-learn library in order to use the k-means algorithm to cluster venues and see similarities for different neighborhoods of each of the cities. K-Means is one of the most common methods of unsupervised machine learning for clustering.
Using this algorithm i have successfully labeled each of the neighborhood venues in suitable clusters and generated a well formatted data frame containing all the required information.
Along with it, I have visualized the total no. of 1st most common venues in each of the clusters of both the cities using a bar graph of matplotlib library.





After this I have also plotted the count of all unique venues in each cluster of 1st most common venues and visualized it using seabron library.

	Cluster Labels	1st Most Common Venue	Venue Count
0	0	Pizza Place	32
1	0	Deli / Bodega	22
2	0	Bar	17
3	0	Deli / Bodega	22
4	0	Pizza Place	32
5	0	Dog Run	1
6	0	Deli / Bodega	22
7	0	Pizza Place	32
8	0	Bank	11
9	0	Pizza Place	32

	Cluster Labels	1st Most Common Venue	Venue Count
0	0	Pizza Place	4
1	0	Pharmacy	5
2	0	Ramen Restaurant	1
3	0	Hockey Arena	1
4	0	Furniture / Home Store	1
5	0	Clothing Store	2
6	0	Gym	4
7	0	Athletics & Sports	1
8	0	Bakery	2
9	0	Latin American Restaurant	1

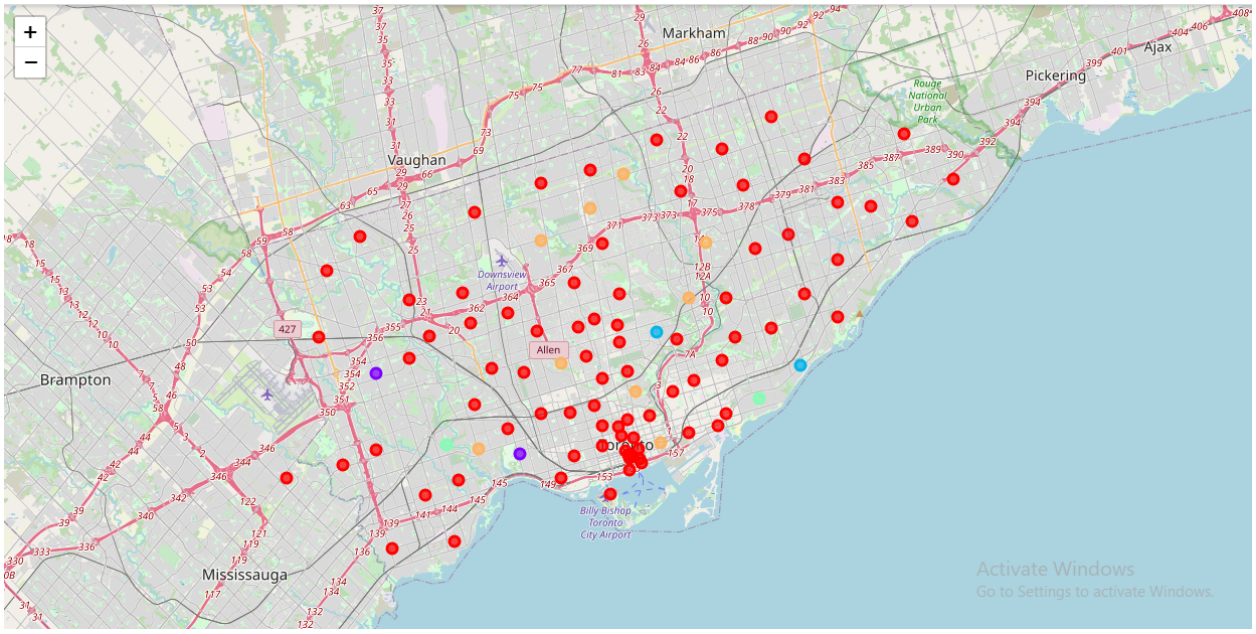
Finally the visualization for both cities' neighborhood clusters is done through the folium library. To find similarities an unsupervised learning is used which helps in clustering places and is quite effective.

Result

The results obtained after applying the k-means clustering on both the dataset show that we can categorise the neighborhoods into 5 clusters based on the frequency of occurrence of venues.
Here is the map of both cities containing clusters of top 10 most common venues obtained after visualising the clustered dataset :-

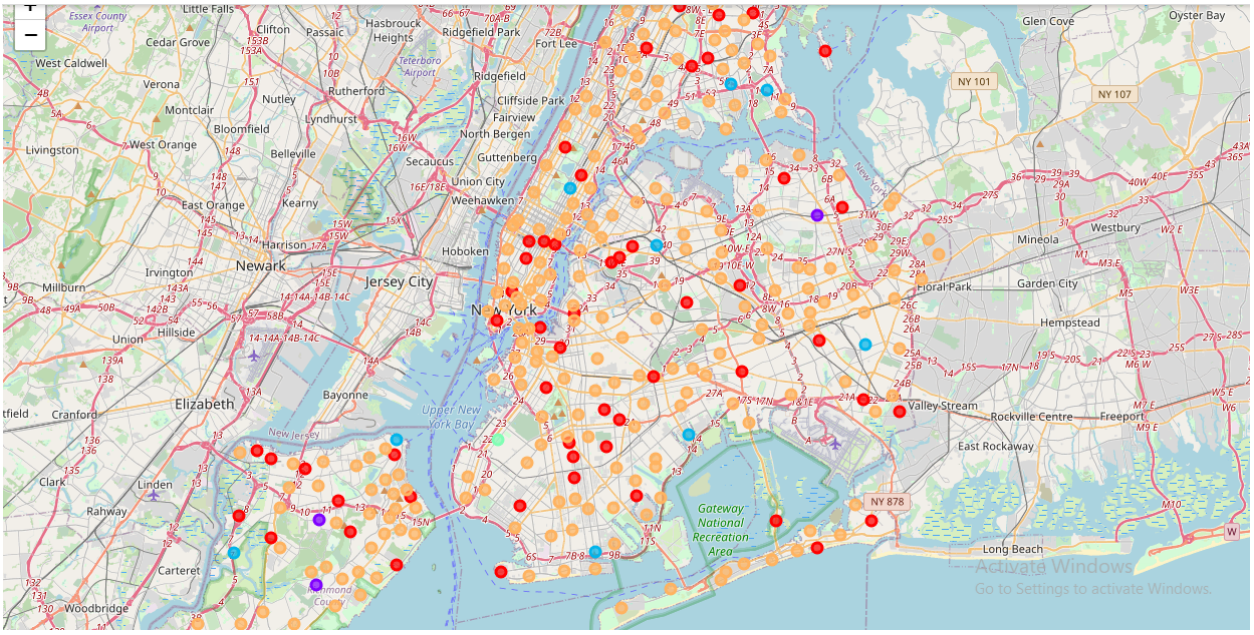
Toronto city cluster map :

- Cluster 1 (Red) : neighborhoods with highest concentration of most common venues
- Cluster 5 (Orange) : neighborhoods with moderate concentration of most common venues
- Clusters 2,3,4 (Purple, Light-blue, Teal) : neighborhoods with lowest concentration of most common venues



New York city cluster map :

- Cluster 5 (Orange) : neighborhoods with highest concentration of most common venues
- Cluster 1 (Red) : neighborhoods with moderate concentration of most common venues
- Clusters 2,3,4 (Purple, Light-blue, Teal) : neighborhoods with lowest concentration of most common venues



One of my aims was also to show the top 5 venues information for each cluster of both the cities in a single table. Thus, I combined both the tables of each city by the top 5 venues. For which I have created two dataframes to show the top 5 most common venues of each cities' clusters and then merged them, which will help us to easily compare and analyze the clusters of both the cities.

Cluster Labels	Toronto Cluster 0	Toronto Cluster 1	Toronto Cluster 2	Toronto Cluster 3	Toronto Cluster 4	|	New York Cluster 0	New York Cluster 1	New York Cluster 2	New York Cluster 3	New York Cluster 4
1st Most Common Venue	Pizza Place	Baseball Field	Garden	Playground	Park	---- ----	Pizza Place	Park	Bus Station	Italian Restaurant	Tennis Court
2nd Most Common Venue	Sandwich Place	Yoga Studio	Home Service	Park	Convenience Store	---- ----	Deli / Bodega	South American Restaurant	Bus Stop	Yoga Studio	Playground
3rd Most Common Venue	Dessert Shop	Donut Shop	Dim Sum Restaurant	Yoga Studio	Yoga Studio	---- ----	Supermarket	Grocery Store	Grocery Store	Fish & Chips Shop	Yoga Studio
4th Most Common Venue	Coffee Shop	Diner	Diner	Doner Restaurant	Donut Shop	---- ----	Check Cashing Service	Bus Stop	Yoga Studio	Event Space	Ethiopian Restaurant
5th Most Common Venue	Café	Discount Store	Discount Store	Dim Sum Restaurant	Diner	---- ----	Mexican Restaurant	Boat or Ferry	Fish & Chips Shop	Exhibit	Event Space

The resultant clustered map and the above table will allow us to identify which neighborhoods have higher concentration of venues and places to visit while which neighborhoods have lower concentration. Based on this, it will help us to answer the question as to which neighborhoods are most suitable based on our requirements.

Discussion

The results show that there is a similarity for both cities' first few common places. The 1st cluster and 5th cluster of Toronto city and New York city respectively, shows that they have a very high concentration of venues around their neighborhoods and that's why they can be compared directly. But the results may also show that there are many clusters that do not have direct comparison. This could be due to a number of factors, such as the geographical size and layout of the neighborhoods and differences in culture and lifestyle between the US and the CA. But the common places for both the cities had coffee shops, cafes and restaurants for certain clusters. Both cities' common places are compared among different clusters and some similarities in distributions are apparent.

Limitations and future research

Here it may be necessary to better clean the venue data returned by Foursquare API. As we can see, some of the top venues listed include 'Bus Stop' or 'Miscellaneous Shop' or 'Discount Store'. This may or may not be a significant venue and could be excluded for more statistically significant venues. However based on the results it is clear we need more holistic data to improve the accuracy and usefulness of our neighborhood clustering. Along with this, it can be considered to include data on population, cost of living, demographics, schools and transportation. In addition, these data can also be accessed dynamically from specific platforms or packages.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing K-means machine learning algorithm for clustering and lastly providing recommendations for stakeholders i.e. one who wants to relocate, expands business or to change the work location or to just visit to explore the other city.

This project has given us some insight into the amenities in the selected neighborhoods of both the cities most common places. As a result, people are turning to big cities as they can achieve better outcomes through their access to the platforms where such information is provided. Not only for investors but also city managers can manage the city more regularly by using similar data analysis types or platforms.

To the future,

Maruf Ansari

Reference

1. Toronto Data set with postal-codes, boroughs, neighborhoods and their locations
2. Geospatial data of Toronto
3. New York data set
4. Foursquare API

Thank you for reading!

This project was created for my Coursera capstone project to complete a IBM Professional Certificate course in Data Science.