Using Data Science to Compare Toronto and Manhattan

Adrian Van Meerbeeck

June 19, 2019

1. Introduction

Toronto is the financial capital of Canada and one of the most vibrant and active cities in the world. People always compare it with Manhattan, NY and considers it the equivalent for Canada. I will use the data collected for both cities as well as the techniques acquired in this course to make a comparison of both cities based in data rather than perceptions.

2. Data acquisition and cleaning

2.1 Data sources

The data for needed to explore, segment, and cluster the neighborhoods in the city of Toronto is not available in the internet in a form ready to be used in data science. I used the data scraped from the Wikipedia page https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M for this analysis. In later sections I will elaborate on the steps I took to acquire and clean the data.

I used BeautifulSoup to scrape the Wikipedia page mentioned above and get the information related to Toronto's boroughs and neighborhoods.

The information for the city of New York and its 5 boroughs and 306 neighborhoods is readily available via the link https://geo.nyu.edu/catalog/nyu_2451_34572. This link contains a dataset with all the boroughs and neighborhoods of New York as well as their latitude and longitude.

For your convenience, I downloaded the files and placed it on the server, so you can simply run a wget command and access the data. So let's go ahead and do that.

2.2 Data Cleaning and Review

Toronto

For Toronto I combined the data scraped from BeautifulSoup into a new dataframe. After deleting the rows with values "Not assigned" for Boroughs I generated a new dataframe called *df_final.* At this stange the data frame contains three columns “Postal Code”, “Borough” and “Neighborhood”

Next step was to add columns Latitude and Longitude to this dataframe. In order to populate the latitude and longitude columns I pulled the information from a csv file named Geospatial_Coordinates.csv available in the site below

http://cocl.us/Geospatial_data/

After reading the geolocation data I created a new dataframe called *df_geocord.* Next, I combined the dataframe *df_geocord* and added the geolocation data to *df_final*.

The resulting *df_final* dataframe is shown below.

	postalcode	borough	neighborhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

Figure 1. *Dataframe containing Toronto’s boroughs and neighborhoods and their geolocations*

Finally I checked the number of *boroughs* and *neighborhoods* in the dataframe and verified that Toronto has 11 boroughs and 103 neighborhoods.

New York

New York data was available from the site mentioned in section 1 and came in a form of a JSON file named “newyork_data.json”.

After reviewing the downloaded data, the next step was to move them into a *pandas* dataframe.

Started by creating an empty dataframe with column names “Borough”, “Neighborhood”, “Latitude” and “Longitude”. Then looped through the data and filled the dataframe one row at a time.

The resulting dataframe called *neighborhoods* is shown below

```
In [45]: neighborhoods.head()
```

Out[45]:

	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

Figure 2. *Dataframe containing New York City's boroughs and neighborhoods and their geolocations*

Finally I checked the number of *boroughs* and *neighborhoods* in the dataframe and verified that New York has 5 boroughs and 306 neighborhoods.

At this stage both sets of data are ready for next step.

2.3 Using Foursquare to explore the cities

At this stage I used Foursquare to collect additional data on both cities. Using Foursquares' API gives access to locations and venues near a particular address or city, recommendations on these venues and much more.

After registering for a developer account and getting my Foursquare API credentials (Client IC and Client Secret), I proceeded to explore both cities.

Starting with Toronto; I collected the nearby venues for each neighborhood and created a new dataframe called *toronto_venues.*

The results showed a total of 2258 venues and of are represented in Figure 3.

```
In [84]:
print(toronto_venues.shape)
toronto_venues_total=toronto_venues.shape[0]
toronto_venues.head()
```

```
(2258, 7)
```

Out[84]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rouge, Malvern	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Rouge, Malvern	43.806686	-79.194353	Interprovincial Group	43.805630	-79.200378	Print Shop
2	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	Chris Effects Painting	43.784343	-79.163742	Construction & Landscaping
3	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
4	Guildwood, Morningside, West Hill	43.763573	-79.188711	Swiss Chalet Rotisserie & Grill	43.767697	-79.189914	Pizza Place

Figure 3. *Toronto venues*

Next step was to group the venues returned from each neighbourhood into categories and create a new dataframe summarizing the neighborhoods, as well as the number of venue categories in each. Counting the number of unique categories in all the neighborhoods of Toronto yields a number of 280 unique categories.

2.4 Analyzing neighborhoods

Next step was to analyse each Toronto neighborhood. I created a one-hot encoding matrix with 2258 rows (number of venues) and 280 columns (number of unique categories). Next, I grouped rows by neighborhood using the mean of the frequency of occurrence for each category. This resulted in a dataframe with dimensions 100 x 280.

I extracted the top 5 most common categories for each neighborhood in Toronto. A few examples are shown in Figure 4 below.

```
----Adelaide, King, Richmond----
                 venue  freq
0                Café  0.05
1          Coffee Shop  0.05
2                  Bar  0.04
3           Steakhouse  0.04
4  American Restaurant  0.04


----Agincourt----
                 venue  freq
0      Sandwich Place  0.25
1              Lounge  0.25
2      Breakfast Spot  0.25
3  Chinese Restaurant  0.25
4         Yoga Studio  0.00
```

Figure 4. *Example of top 5 most common categories for each neighborhood in Toronto*

After putting this data in a pandas dataframe and sorting them in descending order, I created a new dataframe with the top 10 common venues in each neighborhood. See Figure 5.

```
neighborhoods_venues_sorted.head()
```

Out[28]:

	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	Adelaide, King, Richmond	Coffee Shop	Café	Steakhouse	Bar	American Restaurant	Gym	Cosmetics Shop	Hotel	Burger Joint	Restaurant
1	Agincourt	Chinese Restaurant	Lounge	Sandwich Place	Breakfast Spot	Women's Store	Discount Store	Dog Run	Doner Restaurant	Donut Shop	Drugstore
2	Agincourt North, L'Amoreaux East, Milliken, St...	Park	Asian Restaurant	Playground	Women's Store	Donut Shop	Dim Sum Restaurant	Diner	Discount Store	Dog Run	Doner Restaurant
3	Albion Gardens, Beaumond Heights, Humbergate, ...	Grocery Store	Liquor Store	Sandwich Place	Fried Chicken Joint	Video Store	Coffee Shop	Pharmacy	Pizza Place	Beer Store	Fast Food Restaurant
4	Alderwood, Long Branch	Pizza Place	Coffee Shop	Gym	Skating Rink	Pharmacy	Pub	Dance Studio	Pool	Sandwich Place	Women's Store

Figure 5. *Dataframe with top 10 common venues for each neighborhood in Toronto*

The process was repeated for New York city, where I focused mainly in the borough of Manhattan.

The figure below shows a few examples from the extracted top 5 most common categories for each neighborhood in Manhattan.

```
----Battery Park City----
            venue  freq
0            Park  0.08
1     Coffee Shop  0.07
2           Hotel  0.05
3  Memorial Site  0.04
4             Gym  0.03


----Carnegie Hill----
             venue  freq
0      Coffee Shop  0.06
1      Pizza Place  0.06
2             Café  0.04
3  Cosmetics Shop  0.03
4        Bookstore  0.03
```

Figure 6. *Example of top 5 most common categories for each neighborhood in Manhattan*

For Manhattan the dataframe with the top 10 venues in each neighborhood is shown in Figure 7.

Out[61]:

	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	Memorial Site	Wine Shop	Italian Restaurant	Clothing Store	Gym	Plaza	Men's Store
1	Carnegie Hill	Coffee Shop	Pizza Place	Café	Yoga Studio	Bookstore	Wine Shop	Cosmetics Shop	French Restaurant	Bar	Japanese Restaurant
2	Central Harlem	African Restaurant	Public Art	Art Gallery	Seafood Restaurant	Chinese Restaurant	Gym / Fitness Center	French Restaurant	American Restaurant	Cosmetics Shop	Liquor Store
3	Chelsea	Coffee Shop	Ice Cream Shop	Italian Restaurant	Bakery	Nightclub	Theater	Seafood Restaurant	American Restaurant	Hotel	Art Gallery
4	Chinatown	Chinese Restaurant	American Restaurant	Cocktail Bar	Salon / Barbershop	Spa	Bubble Tea Shop	Dumpling Restaurant	Vietnamese Restaurant	Ice Cream Shop	Hotpot Restaurant
5	Civic Center	Italian Restaurant	Gym / Fitness Center	Coffee Shop	French Restaurant	Sandwich Place	Bakery	Yoga Studio	Sporting Goods Shop	American Restaurant	Spa
6	Clinton	Theater	Italian Restaurant	Gym / Fitness Center	American Restaurant	Hotel	Wine Shop	Spa	Coffee Shop	Sandwich Place	Bar
7	East Harlem	Mexican Restaurant	Deli / Bodega	Bakery	Latin American Restaurant	Thai Restaurant	Convenience Store	Café	Gas Station	Taco Place	Steakhouse
8	East Village	Bar	Wine Bar	Chinese Restaurant	Mexican Restaurant	Ice Cream Shop	Cocktail Bar	Vegetarian / Vegan Restaurant	Pizza Place	Ramen Restaurant	Coffee Shop
9	Financial District	Coffee Shop	Steakhouse	Wine Shop	Gym	Bar	Gym / Fitness Center	Hotel	American Restaurant	Cocktail Bar	Café

Figure 7. *Dataframe with top 10 common venues for each neighborhood in Manhattan*

2.4 Clustering the neighborhoods in both cities

Next step in the analysis is to group the neighborhoods into clusters. I used k-means clustering algorithm to cluster the neighborhoods into 5 clusters (k = 5).

Finally, I used the Folium library to visualize the neighborhoods in Toronto and Manhattan and their emerging clusters.

The results of the clustering is shown in the two figures below.

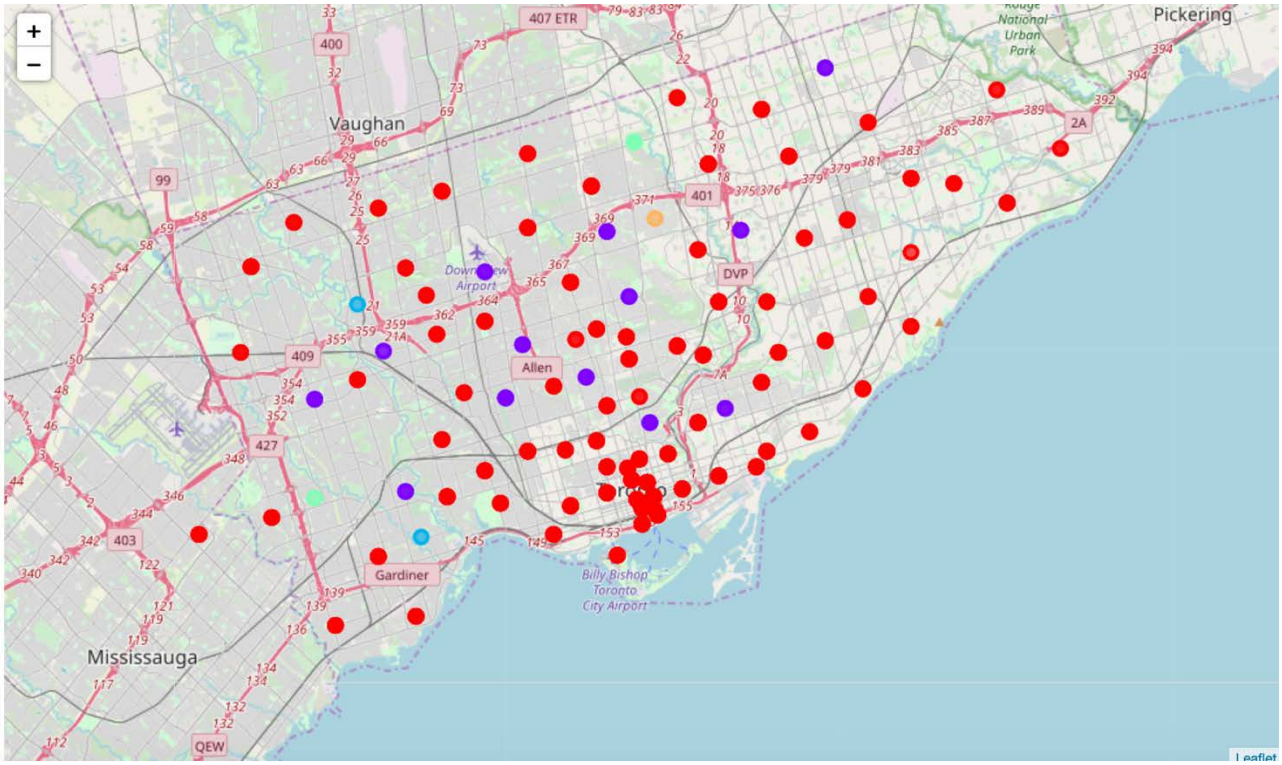

Figure 8. *Clusters in Toronto*

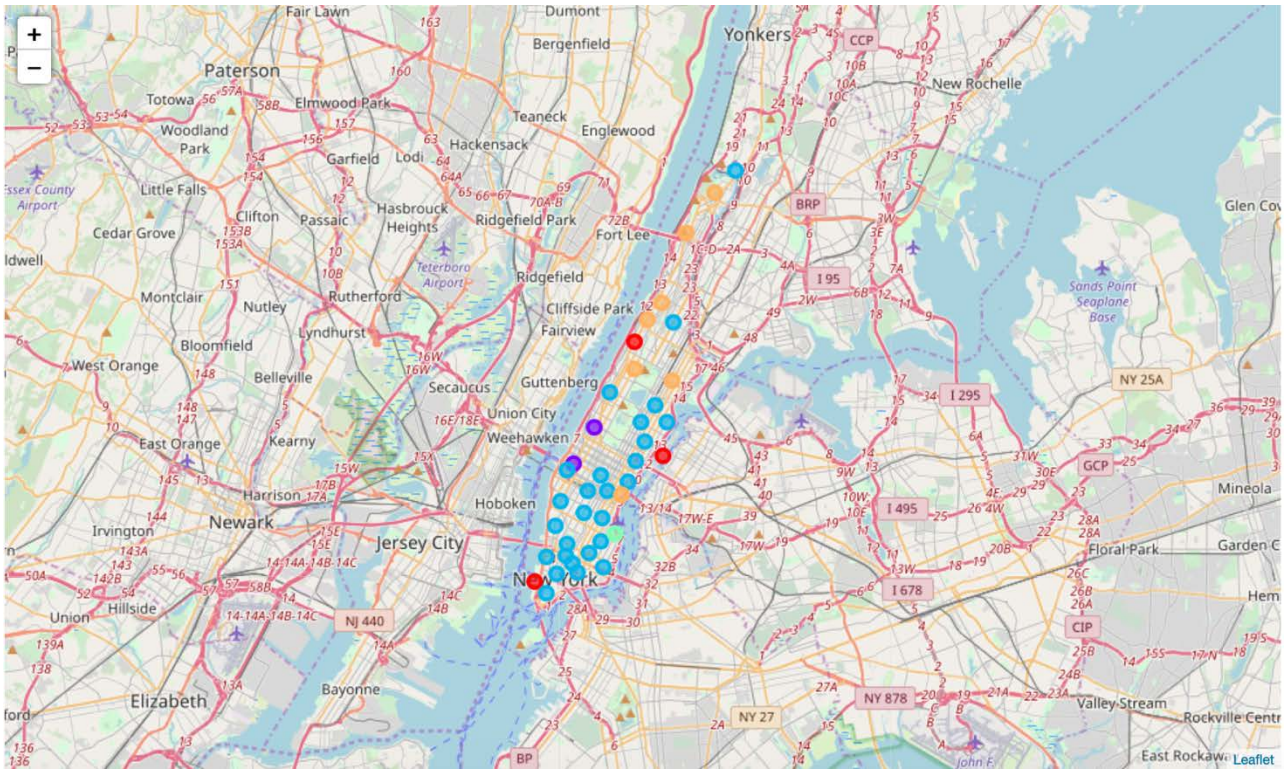

Figure 9. *Clusters in Manhattan*

3 Comparing the two cities - Conclusions

The results of the data from Toronto and Manhattan show several similarities between the two cities:

An indication that both cities have a rich and vibrant environment is indicated by the number of venues returned for each city: Toronto with 2258 and Manhattan with 3331 venues.

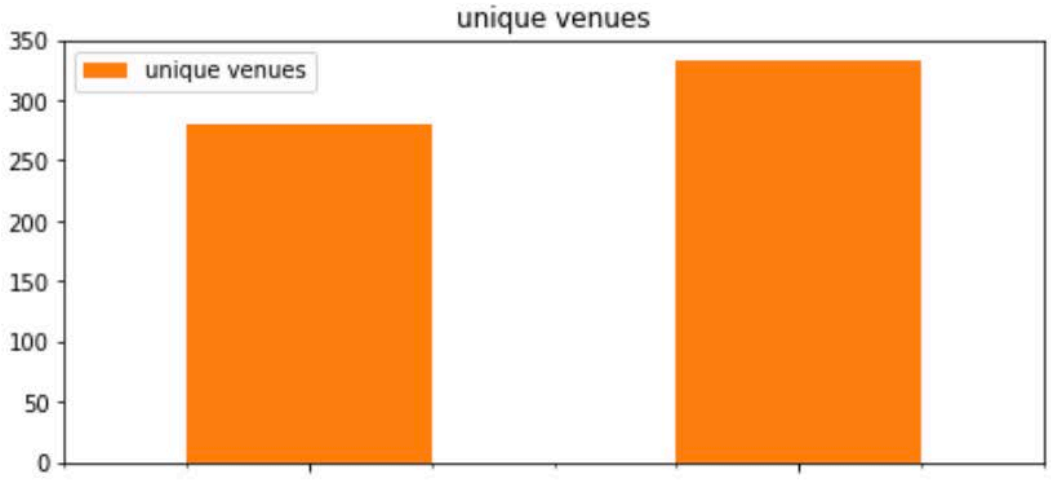

Figure 10. *Unique venues in Toronto and Manhattan*

Both cities are also very diverse and this is demonstrated by the number of unique type venues returned for both; Toronto has 280 and Manhattan has 333 unique type venues.

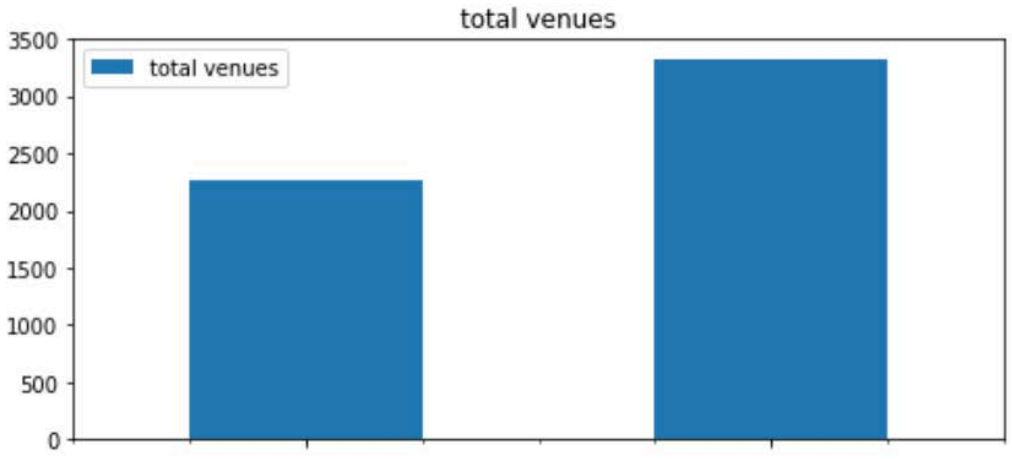

Figure 11. *Unique venues in Toronto and Manhattan*

On the other hand, considering the difference in size between the cities (Manhattan has an area of 33.58 square miles while Toronto is much larger with 243.33 square miles) it seems like Manhattan has a clear edge over Toronto on the number of venues per square miles.

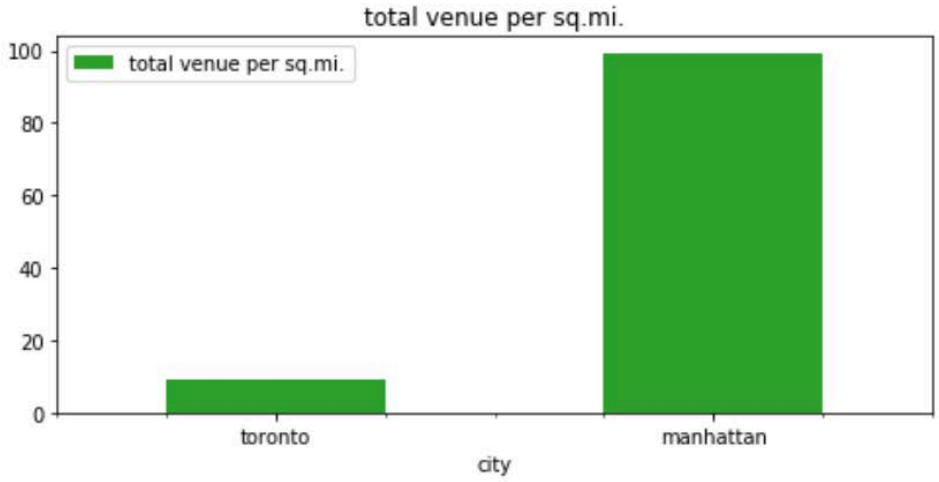

Figure 12. *Venues per square mile in Toronto and Manhattan*

To summarize; both cities are similar and offer a wide variety of venue choices. One can find almost everything they need, or can imagine, in either cities. The difference between the two cities is that while in Manhattan one simply need to walk around in a neighborhood to find almost all type of venues, in Toronto one has to drive or take public transportation since they are spread over a much larger surface.