

# New York vs Toronto

Sherif Essam

## 1. Introduction

Each city has its venues that determine the nature and the characteristics of it. But what can you conclude when I tell you that Paris has for example 40 gardens and parks? could you decide whether Paris is filled with green areas or it lacks it?

These numbers of different kind of venues could be useful to obtain a good overview about the city and could be used to understand the city's characteristics if we have a reference or if we compare between two cities according to these numbers. Comparing cities together would give us a good intuition about understanding the pros and cons of each city.

## 2. Data acquisition and cleaning

### 2.1 Data sources

Toronto and New York city are the chosen cities to be compared together. Datasets for their neighborhood could be the starting point for our project and both datasets could be found on Wikipedia ([Toronto dataset](#) - [New York city dataset](#)).

The two web pages contain two large tables about the neighborhood of each city and these tables would be our raw dat. Such tables could be scraped using BeautifulSoup library after inspecting the website to know its class.

### 2.2 Data Cleaning

Removing rows containing None values :

After scraping the tables from the web pages and converting them into dataframes, the dataframes contained None rows that should be removed.

Adjusting the values of all remaining rows :

All remaining rows contain (\n) at the end that should be removed to deal just with the names of the neighborhoods and avoid these symbols from affecting any upcoming process.

## 2.3 Adding additional necessary columns :

Till now, we have two dataframes that contain only neighborhood column for each city (also Postal code and Borough for Toronto dataframe). Now we want to get the latitude and longitude for all neighborhoods for both cities. We can do that using Geocoder library.

After getting latitude and longitude columns for both cities neighborhoods, the two columns for each city should be concatenated with the city dataframe that contains the neighborhoods names. So, the final dataframe for each city would contain at least Neighborhoods, latitude and longitude columns.

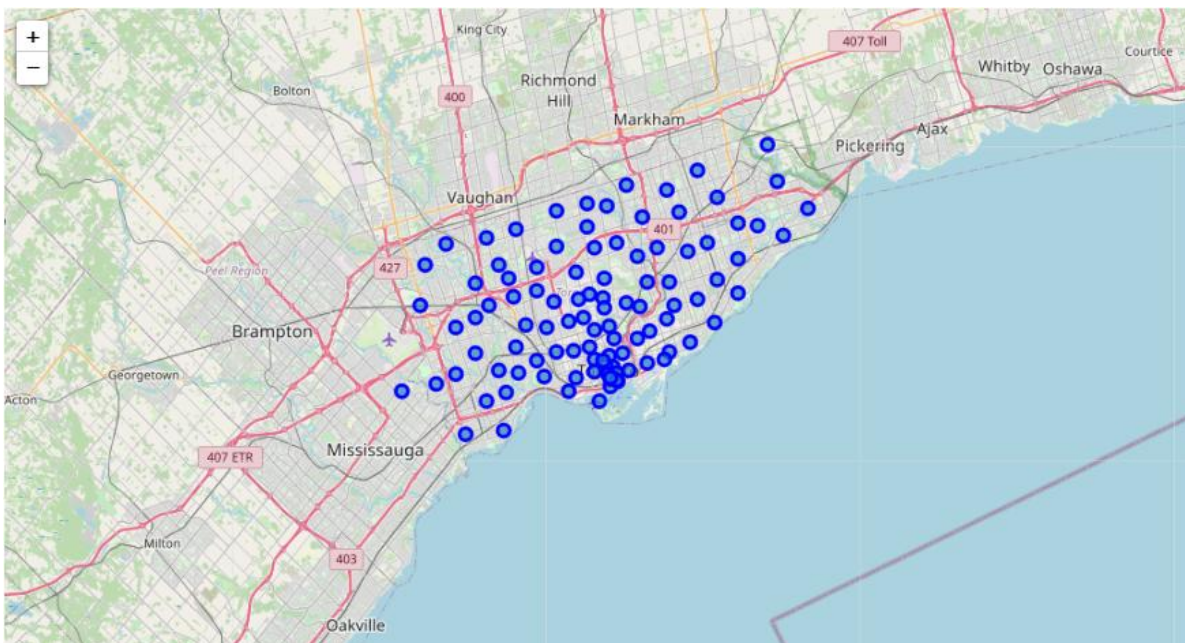
## 2.4 Data visualization (map drawing)

Using Geocoder library, we can obtain the geographical coordinate of both cities which are as following :

- The geographical coordinate of Toronto are 43.6534817, -79.3839347.
- The geographical coordinate of New York are 40.7127281, -74.0060152.

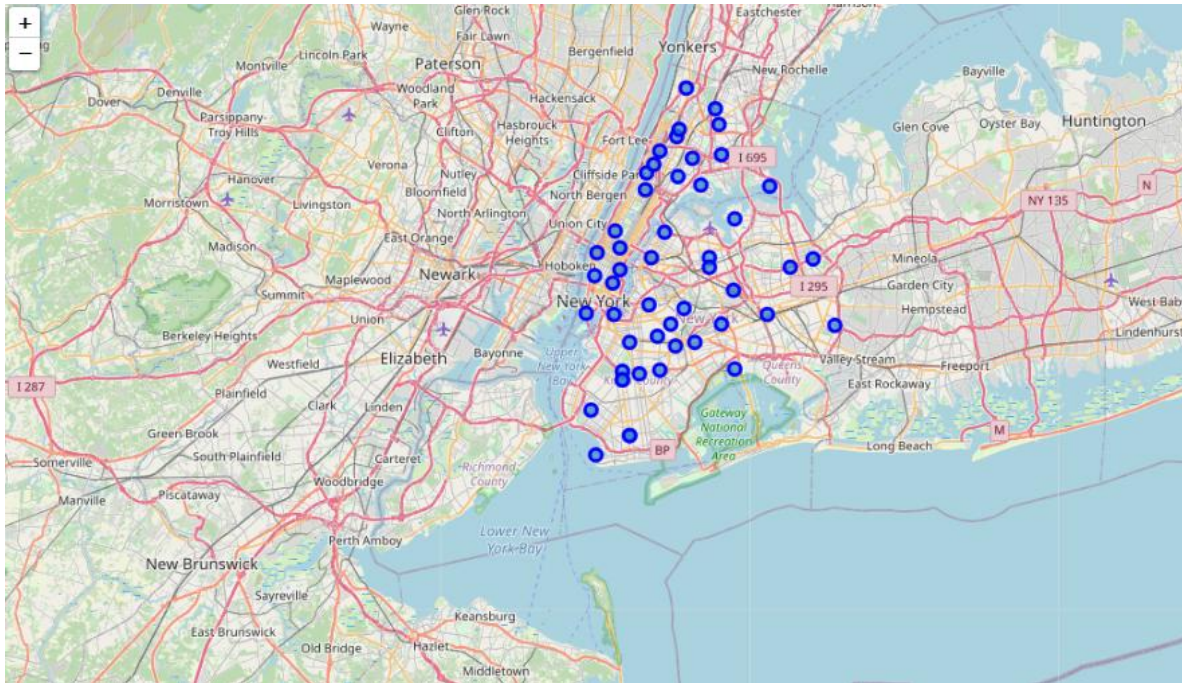
These geographical coordinates for both cities and also latitude and longitude columns of their neighborhoods could be used to draw maps for both cities using Folium library. This would help us make sure that the coordinates of all neighborhoods were obtained correctly.

Map of Toronto :



Map of Toronto with its neighborhoods

Map of New York city :



Map of New York city with its neighborhoods

### 3. Exploratory Data Analysis

First, Foursquare API was used to obtain the data for all venues located in each city (the venue name and its type).

VenueCategory column obtained after using Foursquare API would be the base of the dataframe that would be used for comparing cities together.

Using Value\_counts function on the VenueCategory column would obtain the number of each venue type exists in the city. Storing the result in a dataframe would result in a dataframe like the following :

	count	VenueCategory
<b>Pizza Place</b>	80	<b>Pizza Place</b>
<b>Bar</b>	63	<b>Bar</b>
<b>Coffee Shop</b>	58	<b>Coffee Shop</b>
<b>Deli / Bodega</b>	53	<b>Deli / Bodega</b>
<b>Bakery</b>	52	<b>Bakery</b>
<b>Mexican Restaurant</b>	47	<b>Mexican Restaurant</b>
<b>Chinese Restaurant</b>	40	<b>Chinese Restaurant</b>
<b>Café</b>	40	<b>Café</b>

Venues value counts for New York

The dataframe above contains the counts of all the venues categories. These categories may include Mexican restaurants, Chinese restaurants, Italian restaurants and more. But when comparing between cities, it is only important the total number of restaurants that would give us some indications about the difference between the two cities ignoring the type of each restaurant. So, the dataframe above requires some preprocessing.

So, the next step would be grouping all similar venues types together :

1- Grouping all food places together and the result would be :

Count of all food places in New York = 785

Count of all food places in Toronto = 572

venues that contain (restaurants - Pizza - Sandwich - Chicken - Food - Burger - Diner - Deli - Steakhouse - Joint - Taco - Buffet - Breakfast - Snack - Soup - Gastropub - poke - Burrito - Creperie - Noodle).

2- Grouping all bars together and the result would be :

Count of all Bars in New York = 162

Count of all Bars in Toronto = 117

venues that contain (Bar - pub - Nightclub - Pub).

3- Grouping all shops together and the result would be :

Count of all shops in New York = 570

Count of all shops in Toronto = 466

venues that contain (Shop - Shopping - Bakery - Store - Boutique - Pharmacy - Supermarket - Market - Salad - Butcher - Cleaner - Bookstore - Laundromat - Grocery - Auto - Car).

4- Grouping all sports places together and the result would be :

Count of all sports places in New York = 115

Count of all sports places in Toronto = 70

venues that contain (Gym - Spa - Basketball - Yoga - Playground - Martial - Sports - Bike - Tennis - Stadium).

5- Grouping all Entertaining places together and the result would be :

Count of all Entertaining places in New York = 104

Count of all Entertaining places in Toronto = 81

venues that contain (Circus - Club - Opera - Fountain - Arcade - Pier - Pool - Boat - Gallery - Boutique - Art - Arts - Theater - Hall - Plaza - Dance - Entertainment - Music - Bowling - Golf - Beach - Skating).

6- Grouping all Hotels together and the result would be :

Count of all Hotels places in New York = 35

Count of all Hotels places in Toronto = 47

venues that contain (Hotel - Motel - Hostel - Lounge).

7- Grouping all coffee shops together and the result would be :

Count of all coffee shops in New York = 44

Count of all coffee shops in Toronto = 108

venues that contain (Coffee - Café - Tea - Bistro - Cafe - Cafeteria).

8- Grouping all green places together and the result would be :

Count of all green places in New York = 49

Count of all green places in Toronto = 41

venues that contain (Park - Garden - Field - Farm).

9- Grouping all transportation spaces together and the result would be :

Count of all transportation spaces in New York = 31

Count of all transportation spaces in Toronto = 20

venues that contain (Bus - Station - Bridge - Heliport - Intersection - Airport - Trail).

10- Grouping all tourism places together and the result would be :

Count of all tourism places in New York = 18

Count of all tourism places in Toronto = 14

venues that contain (Museum - Temple - Landmark - Scenic - Historic).

11- Grouping all banks together and the result would be :

Count of all banks places in New York = 30

Count of all banks places in Toronto = 11

12- Grouping all remaining venues together and the result would be :

Count of all banks places in New York = 63

Count of all banks places in Toronto = 68

And the final dataframes would be as follow :

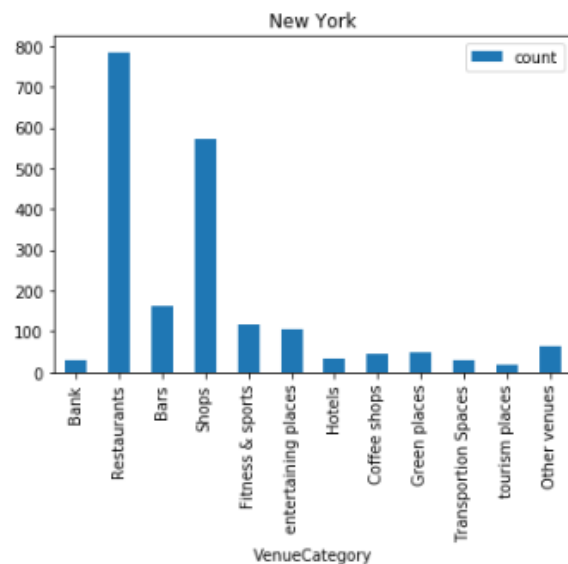
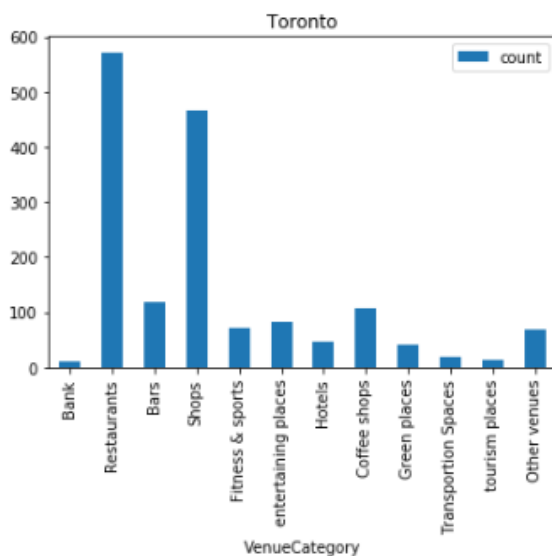
	count	VenueCategory
Bank	30	Bank
Restaurants	785	Restaurants
Bars	162	Bars
Shops	570	Shops
Fitness & sports	115	Fitness & sports
entertaining places	104	entertaining places
Hotels	35	Hotels
Coffee shops	44	Coffee shops
Green places	49	Green places
Transportion Spaces	31	Transportion Spaces
tourism places	18	tourism places
Other venues	63	Other venues

Venues categories counts in New York

	count	VenueCategory
Bank	11	Bank
Restaurants	572	Restaurants
Bars	117	Bars
Shops	466	Shops
Fitness & sports	70	Fitness & sports
entertaining places	81	entertaining places
Hotels	47	Hotels
Coffee shops	108	Coffee shops
Green places	41	Green places
Transportion Spaces	20	Transportion Spaces
tourism places	14	tourism places
Other venues	68	Other venues

Venues categories counts in Toronto

Bar Plots for the two dataframes :



#### **4. Observations :**

At first, let's consider the area of each city to help us compare between the characteristic of each city in a better way :

- New York total area = 783.8 km<sup>2</sup>

- Toronto total area = 630.2 km<sup>2</sup>

New York is bigger than Toronto with about 150 Km<sup>2</sup>. So, it is ordinary that New York would have more specific venues and the two cities would share the same characteristic according to that venues.

#### Coffee shops and entertaining places :

First clear observation is the number of coffee and tea shops in the two cities, It is mentioned that New York is bigger than Toronto and that would affect the number of all types of venues. But despite of that, Toronto has much more coffee shops than New York.

It can be observed that people in Toronto spend much time just sitting in coffee shops drinking some hot drinks and this leads to large number of that kind of shops.

And although the number of entertaining places would be near for both cities, it seems that people in New York use these places more often than people in Toronto who prefer coffee shops in their ordinary days.

#### Green places and transportation stations :

Number of green areas are so close in both cities, which means that Toronto is more greener city than New York. This also reflects on the number of transportation places (Bus stations, Metro Stations and Gas stations) which is greater in New York.

#### Sports and Fitness :

New York has more sports and fitness places and also more different kinds of these places (if we display all kinds of venues categorized in sports and fitness places). These places including Gyms, Yoga studios, Spas, Boxing, Basketball, Tennis and Martial arts.

This variety make it easier for people live in New York to practice whatever the sport each one loves.

Where to spend vacation? (entertaining places - tourist attractions - Hotels) :

New York has much more entertaining and also Tourist attractions, which makes it a better decision to spend your vacation in. But it would also be easier to stay in Toronto than New York because of its large number of hotels that would make it easier to find an available one.

Other observations :

New York is more financial city than Toronto according to the number of Banks. Also we can see that Toronto is famous for Brewery (11 Brewery in it and none in New York).

## **5. Discussion**

After obtaining venues data using Foursquare API, I have tried clustering the neighborhoods according to the top ten venues categories they contain using K-Means clustering technique but the results were not helpful that more than 90% of neighborhood were assigned to one cluster and other clusters may include just 1 neighborhood. I could not conclude or extract any information from the way the neighborhoods were assigned to the clusters so I decided to make my observations based only on the final count of all venues categories.

## **6. Conclusion**

Starting with a dataframe containing the names of neighborhoods of a city, the characteristics of the city could be obtained by comparing the analytical data with the data of a reference city that we already know about it.

By comparing New York city with Toronto, we can say that they are both big cities that share a lot of similar characteristics but some observations could be made about the difference between the two cities.



## **7. Future directions**

The final categories of venues could be subcategorized into more specific subcategories. Also some venues could have specific categories instead of combining them together in (other venues) row. This all will enhance the comparison between cities and make it more accurate.