Coursera Capstone Project
The Battle of the Neighborhoods (Week 2)
FIFTH ASSIGNMENT

Anna Fabiola Cedeno Maselli |Data Science Professional Certificate | 03/02/2021

1. Introduction: Business Problem.

How is it known Toronto, the capital of the province of Ontario, is a major Canadian city along Lake Ontario's northwestern shore. It's a dynamic metropolis with a core of skyscrapers, all dwarfed by the iconic, free-standing CN Tower.

On the other hand New York City comprises 5 boroughs sitting where the Hudson River meets the Atlantic Ocean. At its core is Manhattan, a densely populated borough that's among the world's major commercial, financial and cultural centers. Its iconic sites include skyscrapers such as the Empire State Building and sprawling Central Park.

Let's say that an Italian firm located in Texas City, United States, decides to move its headquarters to New York City or Toronto, Canada. They don't know which city is the best for them. The firm wants to know local businesses and neighborhoods to locate the company. The project will analyze the neighborhoods between New York City and Toronto, understand the differences and similarities, group the neighborhoods, visualize these groups on a map, and provide the best decision.

Also for this project the target audience are the investors interested in moving their headquarters to the best city and may need an objective advice to choose the location for the company and its employees.

2. Data and Methodology

The dataset used for this project and analyse the information are:

a. The websites that collect the information about Toronto and New York borough and their locations. The pages are: https://geo.nyu.edu/catalog/nyu_2451_34572 and https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.

b. The Foursquare API, that collect venues and their categories for each location within a radius of 700 meters.

c. The Geopy and Folium libraries to get the coordinates of every location of Toronto and New York City. The page is https://cocl.us/Geospatial_data

d. Cluster venues of each neighborhood using k-means algorithm and analyze the top 10 most common venue in each cluste.

e. Visualize clusters on the map, thus showing the best locations.

f. The pandas library used for data manipulation and analysis.

g. The Numpy library used to work with arrays.

h. The Requests used to send HTTP/1.1 requests.

i. Matplotlib library used to create static, animated, and interactive visualizations in Python.

j. The json used to transfer data as text that can be sent over a network.

k. The Urllib used to fetch URLs (Uniform Resource Locators).

l. Bs4 library used to pulling data out of HTML and XML files.

The work flow for this project begins:

- Exploring Toronto and New York City through the websites.
- Getting the latitude and longitude coordinates from both cities.
- Creating maps with neighborhoods superimposed on top.
- Exploring the neighborhoods with Foursquare API.
- Getting the first neighborhood's name in the data frame.
- Getting the neighborhood's latitude and longitude values.
- Top 100 venues in a radius of 700 meters.
- Creating the GET request URL.
- How many venues were returned for each neighborhood.
- How many unique categories can be curated from all the returned venues.
- Analyzing Each Neighborhood.
- Grouping rows by neighborhood and by taking the mean of the frequency of occurrence of each category.
- Printing each neighborhood along with the top 5 most common venues.
- Displaying the top 10 venues for each neighborhood.
- Clustering Neighborhoods using K-means.
- New data frame that includes the cluster as well as the top 10 venues for each neighborhood.
- Examining Clusters.

3. Results

Downtown Toronto

It is used k-means to group the neighborhoods into 5 clusters and show the 10th most common venues. The results are:

a. Cluster 0 (Red dots): Coffee Shop, Hotel, Restaurant, Japanese Restaurant, Asian Restaurant, Seafood Restaurant, Gym, Theater and American Restaurant.

b. Cluster 1 (Purple dots): Park, Playground, Trail, Dog Run, Fast Food Restaurant, Farmers Market, Farm, Falafel Restaurant, Event Space and Ethiopian Restaurant.

c. Cluster 2 (Blue dots): Rental Car Location, Airport Service, Airport Terminal, Boat or Ferry, Sculpture Garden, Coffee Shop, Harbor / Marina, Airport Lounge, Airport Gate and Airport Food Court.

d. Cluster 3 (Green dots): Grocery Store, Café, Park, Restaurant, Bakery, Nightclub, Baby Store, Candy Store and Athletics & Sports.

e. Cluster 4 (Orange dots): Coffee Shop, Japanese Restaurant, Sandwich Place, French Restaurant, Pizza Place, Gastropub, Gym and Italian Restaurant
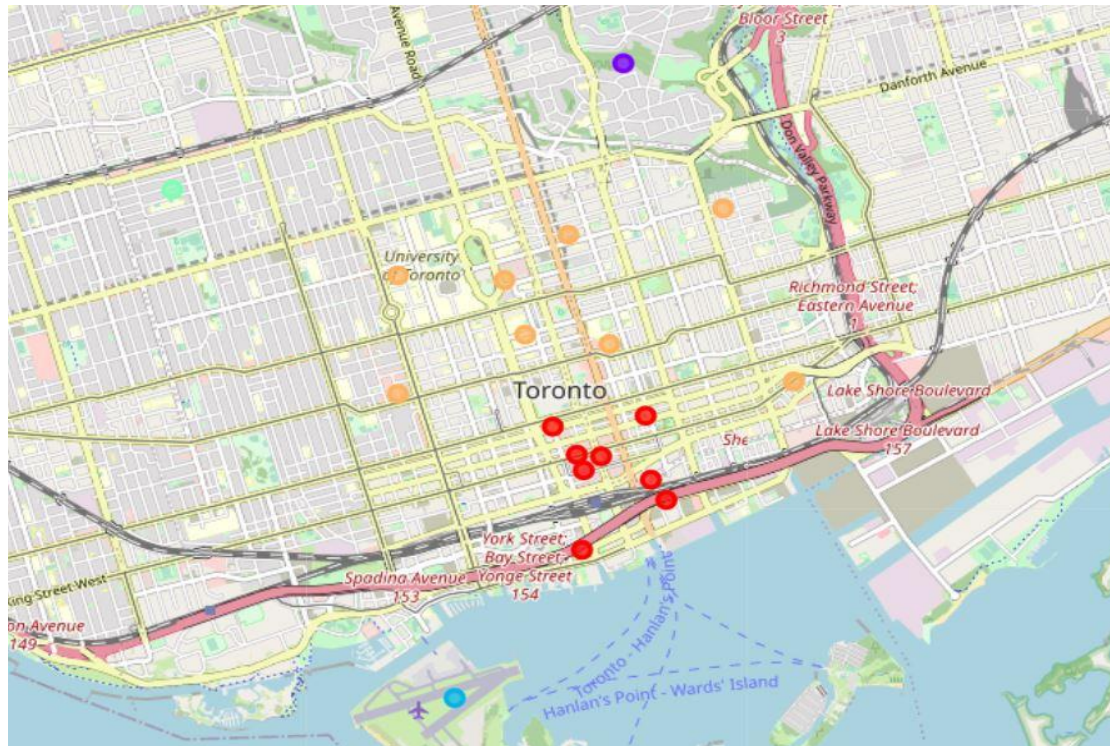
Queens, New York City

It is used k-means to group the neighborhoods into 5 clusters and show the 10th most common venues. The results are:

a. Cluster 0 (Red dots): Pizza Place, Deli / Bodega, Park, Sandwich Place, Donut Shop, Supermarket, Metro Station and Supermarket.

b. Cluster 1 (Purple dots): Beach, Park, Pizza Place, Zoo, Event Space, Falafel Restaurant, Farm, Farmers Market, Fast Food Restaurant and Filipino Restaurant.

c. Cluster 2 (Blue dots): Playground, Park, Men's Store, Athletics & Sports, Event Space, Falafel Restaurant, Farm, Farmers Market, Fast Food Restaurant and Filipino Restaurant.

d. Cluster 3 (Green dots): Chinese Restaurant, Pizza Place, Bakery, Donut Shop, Sandwich Place, Bar and Fast Food Restaurant.

e. Cluster 4 (Orange dots): Beach, Supermarket, Donut Shop, Dog Run, Bar, Beach Bar, Mexican Restaurant, Bagel Shop, Fast Food Restaurant and Seafood Restaurant.
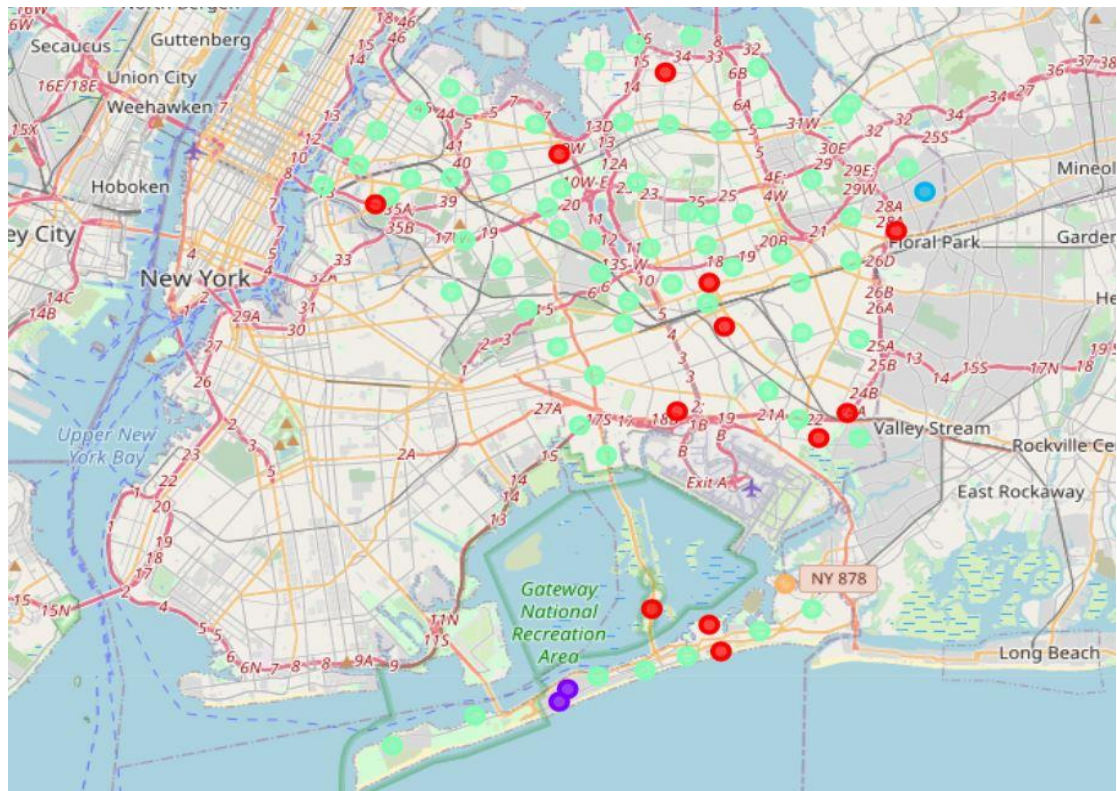
4. Discussion.

Toronto has 10 boroughs and 103 neighborhoods and its geographical coordinate are for the latitude: 43.6534817 and longitude: -79.3839347. Downtown Toronto has 19 neighborhoods and 224 venues, further there are 816 distinct venues in 224 categories.

On the other hand New York City has 5 boroughs and 306 neighborhoods and its geographical coordinate are for the latitude: 40.7127281 and longitude -74.0060152. Queens borough has 81 neighborhoods and 302 venues, further there are 2559 distinct venues in 302 categories.

5. Conclusion.

In this project it was collect the information about Toronto and New York boroughs from the websites, using geospatial libraries to mapped them, using Foursquare API to collect the venues and their types for each location within a radius of 700 meters. It was collected neighborhoods and venues by location and preparing them for clustering and finally by the k-means algorithm, analyze the top 10 most common venues in each cluster and visualized them on the map. In conclusion both cities are good but based on the quantity of venues and neighborhoods is the best for the Italian firm to choose Queens over Downtown Toronto to move its headquarters because offer more options for the company and its employees.