"Tale of Two Cities: Toronto vs. New York Comparison Analysis"

Applied Data Science Capstone Project -Week 5

Prepared by Alice Yang

June 2021

Table of Contents

1 Introduction ---------- 2
1.1 Background ---------- 2
1.2 Objective ---------- 2
1.3 Libraries and Packages ---------- 3
2 Data ---------- 4
2.1 Data Sources ---------- 4
2.2 Data Description ---------- 4
3 Methodology ---------- 5
3.1 Data preprocessing ---------- 5
3.2 Foursquare API Data ---------- 5
4 Results ---------- 6
4.1 Maps ---------- 6
4.2 Analyze Each Neighbourhood ---------- 7
5 Discussion ---------- 8
6 Conclusion ---------- 9

1. Introduction

1.1 Background

A Tale of Two cities, a novel written by Charles Dickens was set in two European cities which takes pla ce during the French Revolution. Similar story between two cities were both happening then and now. In the year of COVID19 pandemic, a lot has changed, and we now take a look at how Toronto and New Y ork, two North America cities, have looked like now.

Toronto and New York are quite the popular tourist and vacation destinations for people all around the world. Both cities are international centre of business, finance, arts, and culture and share some common characteristics. They are diverse and multicultural and offer a wide variety of experiences that is widely sought after. The purpose of this project is to group the neighbourhoods of Toronto and New York respe ctively and draw insights to how they recovered approximately one year after the pandemic outbreak.

1.2 Objective

The project aims to create an analysis of features for new graduate student live in the North America eas t coast to search a best neighborhood as comparative analysis between two main cities in Canada and Un ited States. The features include housing, crime rates, road connectivity, management for emergency, an d excrement conveyed in sewers and recreational facilities.

This will help a new graduate student to have a better understanding of workplace after the pandemic an d realize how the neighbourhoods in the east coasts look like before starting their careers, taking two po pular cities as an example.

1.3 Libraries and Packages

The libraries used in this project are Pandas, Numpy, requests, matplotlib, folium, geocoder. The packages used are BeautifulSoup, Nominatim, json_normalize, matplotlib.cm, matplotlib.colors, sklearn.cluster.

2. Data

2.1 Data Source

Most data used in this project come from web source. Neighbourhood data for Toronto come from web s ource here and the location data from here Neighbourhood data for New York come from here. We will preprocess these two neighbourhoods data so they contain same format and therefore for easy analysis. The feature information about neighbourhoods in both cities come from Foursquare.

2.2 Data Description

2.1.1 Toronto Neighbourhood data

Data Source 1: We will use web scraping technique to get the data. There are several changes made to th e raw web scaping data:

- We will only extract three pieces of information: PostalCode, Borough, and Neighborhood
- Only process the cells that have an assigned borough.
- Ignore cells with a borough that is Not assigned.
- More than one neighborhood can exist in one postal code area. For example, in the table on the Wikipedia page, you will notice that M5A is listed twice and has two neighborhoods: Harbourfro nt and Regent Park. These two rows will be combined into one row with the neighborhoods sepa rated with a comma as shown in row 11 in the above table.
- If a cell has a borough but a Not assigned neighborhood, then the neighborhood will be the same as the borough.

Data Source 2: We will combine the above data set with Toronto location csv dataset. The csv dataset ha s three columns: PostalCode, Latitude, and Longitude. PostalCode is consistent with the above data sour ce and therefore can be used for matching purpose.

3. Methodology

3.1 Data Preprocessing

Using the data cleaning techniques and data description above, we are able to generate geography data frames for Toronto and New York. Each data frame contains four columns: Borough, Neighborhood, Latitude, Longitude. The desired data frame should look like this:

	Borough	Neighborhood	Latitude	Longitude
0	North York	Parkwoods	43.753259	-79.329656
1	North York	Victoria Village	43.725882	-79.315572
2	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	Queen's Park	Ontario Provincial Government	43.662301	-79.389494

3.2 Foursquare API Data

We use foursquare API to get venues near the neighbourhoods with radius of 500 meters. The function used to generate suck venues data frame is getNearbyVenues taking 3 parameters: neighbourhood name, latitudes, longitude.

4. Results

4.1 Maps

4.1.1 Toronto Venues Map

4.1.2 New York Venues Map

4.2 Analyze Each Neighbourhood

4.2.1 One Hot Encoding

```
11  toronto_onehot.head()
```

	Yoga Studio	Accessories Store	Adult Boutique	Afghan Restaurant	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	...	Truck Stop
0	0	0	0	0	0	0	0	0	0	0	...	0
1	0	0	0	0	0	0	0	0	0	0	...	0
2	0	0	0	0	0	0	0	0	0	0	...	0
3	0	0	0	0	0	0	0	0	0	0	...	0
4	0	0	0	0	0	0	0	0	0	0	...	0

5 rows × 252 columns

```
11  newyork_onehot.head()
```

	Yoga Studio	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Animal Shelter	Antique Shop	Arcade	Arepa Restaurant	...	Warehouse Store
0	0	0	0	0	0	0	0	0	0	0	...	0
1	0	0	0	0	0	0	0	0	0	0	...	0
2	0	0	0	0	0	0	0	0	0	0	...	0
3	0	0	0	0	0	0	0	0	0	0	...	0
4	0	0	0	0	0	0	0	0	0	0	...	0

5 rows × 426 columns

4.2.2 Top 5 Most Common Venues in Toronto Neighbourhoods

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Agincourt	Latin American Restaurant	Skating Rink	Clothing Store	Lounge	Breakfast Spot
1	Alderwood, Long Branch	Pizza Place	Coffee Shop	Sandwich Place	Dance Studio	Pub
2	Bathurst Manor, Wilson Heights, Downsview North	Coffee Shop	Bank	Frozen Yogurt Shop	Pharmacy	Pizza Place
3	Bayview Village	Chinese Restaurant	Japanese Restaurant	Café	Bank	Distribution Center
4	Bedford Park, Lawrence Manor East	Sandwich Place	Coffee Shop	Pizza Place	Restaurant	Pub

4.2.3 Top 5 Most Common Venues in New York Neighbourhoods

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Allerton	Deli / Bodega	Pizza Place	Discount Store	Pharmacy	Spa
1	Annadale	Pizza Place	Bakery	Liquor Store	Food	Train Station
2	Arden Heights	Pharmacy	Pizza Place	Bus Stop	Coffee Shop	Home Service
3	Arlington	Deli / Bodega	Bus Stop	Boat or Ferry	Grocery Store	Fruit & Vegetable Store
4	Arrochar	Pizza Place	Italian Restaurant	Bus Stop	Athletics & Sports	Food Truck

5. Discussion

The major purpose of this porject is to create an analysis of features for new graduate students live in the North America east coast to search a best neighborhood as comparative analysis between two main citie s in Canada and United States. Even though the whole world has been hit but the COVID19 global pand emic since 2019/2020, it seems like the two cities in North America are running like normal until now. According to the analysis above, we can observe that Toronto city is far less crowded compares to New York city. Stores, banks, restaurants are running as usual in both countries. New graduate students need t o take their own consideration and carefully plan their future career.

6. Conclusion

In this project, I use the Foursquare API to find popular venues around the neighbourhoods

I feel rewarded with the efforts and believe this course with all the topics covered is well worthy of appr eciation. This project has shown me a practical application to resolve a real situation that has impacting personal and financial impact using Data Science tools. The mapping with Folium is a very powerful tec hnique to consolidate information and make the analysis and decision better with confidence.

Future Works:

This project can be continued for making it more precise in terms to find popular venues as the radius of searches is only set to 500. It will take a very long time to search for venues in larger radius.