Compare the similarities of financial world cities (Manhattan and Toronto)

Jiaqing Chen 2020/08/23

1. Introduction

1.1. Background

A global city, also called a power city, world city, alpha city or world center, is a city which is a primary node in the global economic network. The concept comes from geography and urban studies, and the idea that globalization is created and furthered in strategic geographic locales according to a hierarchy of importance to the operation of the global system of finance and trade. A large city carries the development opportunities of a region. A world-class city is a place for investors and job seekers alike. There are many common characteristics between these cities. Looking for these characteristics not only helps us to analyze the future development of big cities, but also helps us to measure which city has these common parts and will develop into a big city in the future. This is a very meaningful thing for many people, especially investors. This will help trend the direction of capital development.

1.2. Problem

In part of the previous work, I explored New York City and the city of Toronto and segmented and clustered their neighborhoods. Both cities are very diverse and are the financial capitals of their respective countries. One interesting idea is to compare the neighborhoods of the two cities and determine how similar they are. The problem the project is trying to solve is in which aspects and to what extent the two cities are similar. Through comparison and analysis, the project attempts to draw a conclusion in which aspects can we see whether a city has the potential to develop into a large financial city.

2. Data

2.1. Description

The datasets needed to complete this project are what I used in previous work. They include the data from New York and Toronto, respectively. For the New York dataset, the Neighborhood has a total of 5 boroughs and 306 neighborhoods. In order to

segment the neighborhoods and explore them, I essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood. The dataset exists for free the web and the link to the dataset is on shown https://geo.nyu.edu/catalog/nyu 2451 34572. For the Toronto neighborhood data, a Wikipedia page exists that has all the information we need to explore and cluster the neighborhoods in Toronto. In general, it is important to first scrape the Wikipedia page and wrangle the data, clean it, and then read it into a pandas dataframe so that it is in a structured format like the New York dataset. The link to the Wikipedia page of the dataset is shown as:

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.

2.2. Usage Method

These two datasets are first faced with some cleaning work. After downloading the data, the first work is to transform the data into a panda dataframe. Next, we need to use some geopy libraries to get the latitude and longitude values of the two cities and create two maps of them with neighborhoods superimposed on top. Here, we need to use a library called Folium, which is a great visualization library. We could zoom into the above map, and click on each circle mark to reveal the name of the neighborhood and its respective borough. Next, we are going to start utilizing the Foursquare API to explore the neighborhoods and segment them. After completing the above steps, we will face the most important part, analysis work. We need group rows by neighbor and by taking the mean of the frequency of occurrence of each category. Finally, we use the k-means method for clustering neighborhoods. This includes the work of comparison and analysis. This project would exchange test datasets to determine how similar the two models are. The exchanged test datasets are compared with the non-exchanged test datasets. At the same time, the two models are compared intuitively.

3. Methodology

3.1. Overview

The whole data analysis and comparison section covers data download, cleaning, merging and analysis. As the project is mainly for data analysis of the two regions, the analysis steps and processes are basically similar. The difference between them is mainly the data acquisition part. One data is downloaded from Wikipedia. The

processing of this data is complicated. The other data is loaded directly in the form of CSV. This project is also designed to show two different ways to load data. After the data is loaded, the project cleans up and merges the data appropriately. This process clearly sorts out the data. Next, the data is analyzed step by step, and the results of a single regional data set are obtained. Finally, the results of the two regions are compared and the analysis report is output. In the process of data analysis, this project uses the Foursquare library. This will be described later in the report.

3.2. Data Collation and Analysis

For the data download and collation part, this project will describe the two methods adopted by itself. The full version of this code can be found in the link to the project. In the method of downloading by link, the data must be imported into the dataframe through the link, which is as shown below.

```
Scrape the List of postal codes of Canada from URL

In [2]:

ca_url = "https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M"

ca_source = requests.get(ca_url).text

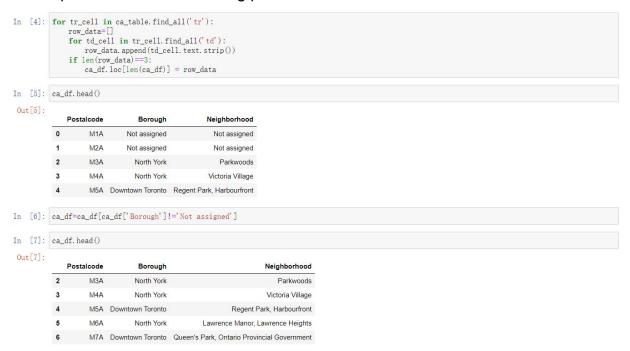
ca_soup = BeautifulSoup(ca_source, 'xml')

ca_table=ca_soup.find('table')

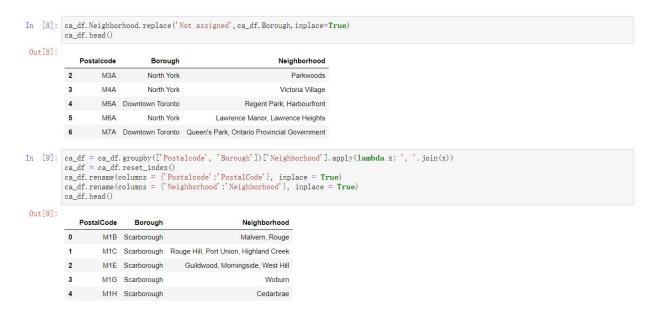
ca_column_names = ['Postalcode', 'Borough', 'Neighborhood']

ca_df = pd.DataFrame(columns = ca_column_names)
```

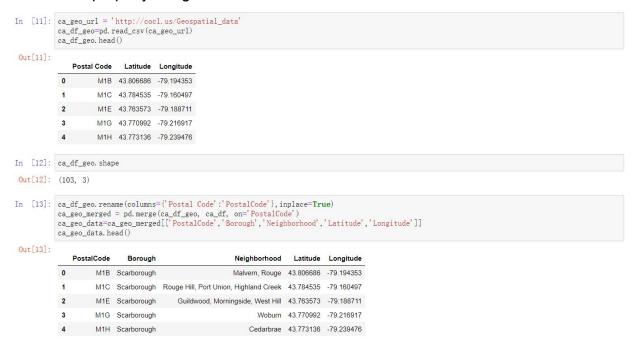
Next, the data needs to be further cleaned by searching and sorting. The operation of each step is shown in the following process.



Finally, we will get the following results through the analysis of this part.



At this point, we need to download the data from the new link and merge the two parts of the data. Before merging two parts of data, we should make sure that the two parts of data can be properly merged.



Next, we will define a new function to get the top 100 venues that are in Toronto within a radius of 500 meters.

Get the top 100 venues that are in Toronto within a radius of 500 meters

```
In [24]: def getNearbyVenues(names, latitudes, longitudes):
                 radius=500
LIMIT=100
                 venues_list=[]
                 for name, lat, lng in zip(names, latitudes, longitudes):
                     print(name)
                      url = 'https://api.foursquare.com/v2/venues/explore?&client_id={} &client_secret={} &v={} &11={}, {} &radius={} &limit={}'.format(CLIENT_ID,
                           CLIENT_SECRET,
                           VERSION,
                           lat,
                           lng,
                           radius,
                           LIMIT)
                      results = requests.get(url).json()["response"]['groups'][0]['items']
                       # return only relevant information for each nearby venue
                       venues_list.append([(
                           name,
                           lat,
                          lng,
v['venue']['name'],
v['venue']['location']['lat'],
v['venue']['location']['lng'],
v['venue']['categories'][0]['name']) for v in results])
                 nearby_venues = pd.DataFrame([item for venue_list in venues_list for item in venue_list])
                 nearby_venues.columns = ['Neighborhood', 'Neighborhood Latitude', 'Neighborhood Longitude',
                                  'Venue Latitude',
'Venue Longitude',
                                  'Venue Category']
                 return(nearby_venues)
```

Then, we will call this function to analyze and process the dataset and get new results.

```
In [25]: ca_toronto_venues = getNearbyVenues(names=ca_toronto_data['Neighborhood'],
                                                 latitudes=ca_toronto_data['Latitude'],
                                                 longitudes=ca_toronto_data['Longitude']
           The Beaches
           The Danforth West, Riverdale
           India Bazaar, The Beaches West
           Studio District
           Lawrence Park
           Davisville North
           North Toronto West, Lawrence Park
           Davisville
           Moore Park, Summerhill East
           Summerhill West, Rathnelly, South Hill, Forest Hill SE, Deer Park
           Rosedale
           St. James Town, Cabbagetown
           Church and Wellesley
           Regent Park, Harbourfront
           Garden District, Ryerson
           St. Tames Town
           Berczy Park
           Central Bay Street
           Richmond, Adelaide, King
           Harbourfront East, Union Station, Toronto Islands
Toronto Dominion Centre, Design Exchange
           Commerce Court, Victoria Hotel
           Roselawn
           Forest Hill North & West, Forest Hill Road Park
           The Annex, North Midtown, Yorkville
           University of Toronto, Harbord
           Kensington Market, Chinatown, Grange Park
CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara, Island airport
           Stn A PO Boxes
           First Canadian Place, Underground city
           Christie
           Dufferin, Dovercourt Village
           Little Portugal, Trinity
           Brockton, Parkdale Village, Exhibition Place
           High Park, The Junction South
Parkdale, Roncesvalles
           Runnymede, Swansea
           Queen's Park, Ontario Provincial Government
           Business reply mail Processing Centre, South Central Letter Processing Plant Toronto
```

At this time, we need to sort out the data and get the counting results that are conducive to our further classification and sorting. Here, this report will not show this part of the steps one by one. The specific process is in the project link. Only the final statistics are shown here.

	Neighborhood	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	 Theme Restaurant	Toy / Game Store	Trail	Train Station	Vegetariar / Vegar Restauran
0	Berczy Park	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.0	0.0	0.0	 0.0	0.0	0.0	0.0	0.017241
1	Brockton, Parkdale Village, Exhibition Place	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.0	0.0	0.0	 0.0	0.0	0.0	0.0	0.000000
	Business reply mail Processing Centre, South C	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.0	0.0	0.0	 0.0	0.0	0.0	0.0	0.000000
3	CN Tower, King and Spadina, Railway Lands, Har	0.066667	0.066667	0.066667	0.133333	0.2	0.133333	0.0	0.0	0.0	 0.0	0.0	0.0	0.0	0.000000
4	Central Bay Street	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.0	0.0	0.0	 0.0	0.0	0.0	0.0	0.01538

Through the above data, we further develop a new statistical function.

By executing this part of the function, we get the final result of one of the datasets.

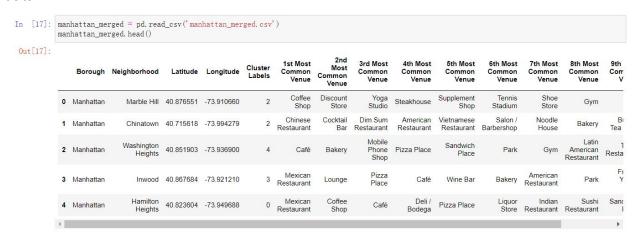
] :													
Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
The Beaches	43.676357	-79.293031	0	Health Food Store	Pub	Trail	Dog Run	Dessert Shop	Dim Sum Restaurant	Diner	Discount Store	Distribution Center	Yoga Studio
The Danforth West, Riverdale	43.679557	-79.352188	0	Greek Restaurant	Coffee Shop	Italian Restaurant	Restaurant	Ice Cream Shop	Furniture / Home Store	Fruit & Vegetable Store	Pub	Pizza Place	Lounge
India Bazaar, The Beaches West	43.668999	-79.315572	0	Sushi Restaurant	Pub	Sandwich Place	Light Rail Station	Board Shop	Liquor Store	Burrito Place	Italian Restaurant	Restaurant	Cream Shop
Studio District	43.659526	-79.340923	0	Café	Coffee Shop	Gastropub	Bakery	Brewery	American Restaurant	Yoga Studio	Convenience Store	Sandwich Place	Cheese Shop
Lawrence Park	43.728020	-79.388790	2	Park	Bus Line	Swim School	Department Store	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Donut Shop	Doner Restaurant	Dog Run
4)

For another dataset, the steps we take will be simpler. The two files here are already saved and exist in the project directory. We first import the first batch of data.

Read csv file with clustered neighborhoods with geodata of Manhattan

In [15]: manhattan_data = pd. read_csv('mh_neigh_data.csv') manhattan_data.head() Out[15]: Borough Neighborhood Latitude Longitude Cluster Labels **0** Manhattan Marble Hill 40.876551 -73.910660 Chinatown 40.715618 -73.994279 2 Manhattan Washington Heights 40.851903 -73.936900 3 Manhattan Inwood 40.867684 -73.921210 3 4 Manhattan Hamilton Heights 40.823604 -73.949688 In [16]: manhattan_data.tail() Out[16]: Borough Neighborhood Latitude Longitude Cluster Labels **35** Manhattan Turtle Bay 40.752042 -73.967708 36 Manhattan Tudor City 40,746917 -73,971219 37 Manhattan Stuyvesant Town 40.731000 -73.974052 38 Manhattan Flatiron 40.739673 -73.990947 3 39 Manhattan Hudson Yards 40.756658 -74.000111

Next, we can directly get the data results of this region by importing the second batch of data.



After the two parts of data are loaded and sorted out, we will compare and analyze the two groups of data. We wanted to apply the data to maps and analyze them. However, through practice, I found that this can not get good results. On the contrary, it is more appropriate to analyze the data table directly.

3.3. Foursquare

As mentioned earlier, the library Foursquare is used in this project. This part is mainly used to segment and cluster the neighborhoods. Foursquare, which is short for Foursquare City Guide, is a local search-and-discovery mobile app developed by Foursquare Labs Inc. The app provides personalized recommendations of places to go near a user's current location based on users' previous browsing history and check-in history.

```
In [19]: !conda install -c conda-forge geocoder --yes
           !conda install -c conda-forge geopy --yes
          !pip install lxml
          import geocoder
          from geopy.geocoders import Nominatim
          address = 'Toronto, Ontario'
           geolocator = Nominatim(user_agent="toronto_explorer")
          location = geolocator.geocode(address)
latitude = location.latitude
          longitude = location.longitude
          Collecting package metadata (current_repodata.json): ...working... done
          Solving environment: ...working... done
          ## Package Plan ##
            environment location: D:\Anaconda
           added / updated specs:
               - geocoder
          The following packages will be downloaded:
                                                  build
             package
                                         py38h32f6830_2 3.1 MB conda-forge
             conda-4. 8. 4
                                                   Total:
                                                                  3.1 MB
          The following packages will be UPDATED:
                                                4.8.4-py38h32f6830_1 --> 4.8.4-py38h32f6830_2
```

The usage of this library will not be discussed here. This part of the content can be found in the course materials.

4. Results

Downloading and Extracting Packages

Since the effect of map display is not as good as expected. And the map display can not clearly get the contrast results. Finally, this project adopts the method of direct comparison of tables, which are dataframes to get the final result. Here we will show the data analysis results of the two regions and the comparison results of the data analysis results of the two regions. The comparison results can not be fully displayed. Specific data run results can be found in the project link. The statistical results of the two regions are shown below.

9]:						2nd								10th
	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	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	Mos Common Venue
	The Beaches	43.676357	-79.293031	0	Health Food Store	Pub	Trail	Dog Run	Dessert Shop	Dim Sum Restaurant	Diner	Discount Store	Distribution Center	Yog Studi
	The Danforth West, Riverdale	43.679557	-79.352188	0	Greek Restaurant	Coffee Shop	Italian Restaurant	Restaurant	Ice Cream Shop	Furniture / Home Store	Fruit & Vegetable Store	Pub	Pizza Place	Loung
	India Bazaar, The Beaches West	43.668999	-79.315572	0	Sushi Restaurant	Pub	Sandwich Place	Light Rail Station	Board Shop	Liquor Store	Burrito Place	Italian Restaurant	Restaurant	Crear Sho
	Studio District	43.659526	-79.340923	0	Café	Coffee Shop	Gastropub	Bakery	Brewery	American Restaurant	Yoga Studio	Convenience Store	Sandwich Place	Cheese
	Lawrence Park	43.728020	-79.388790	2	Park	Bus Line	Swim School	Department Store	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Donut Shop	Doner Restaurant	Dog Ru
	4													
0]:	↑ manhattan_m	erged. head	1()											
0]:		erged. head Latitude	Longitude	Cluster Labels	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 Mos
0]:	manhattan_m		Longitude		Common	Most Common	Common	Common	Common	Common	Common	Common	Common Venue	10th Mos Commo Venu Seafoo
0]:	manhattan_m Neighborhood Marble Hill	Latitude 40.876551	Longitude	Labels	Common Venue Coffee	Most Common Venue	Common Venue Yoga	Common Venue	Common Venue Supplement	Tennis Stadium	Common Venue Shoe	Common Venue	Common Venue	10th Mos Common Venue Seafoor Restauran Ice Crean
0]:	manhattan_m Neighborhood Marble Hill	Latitude 40.876551	-73.910660 -73.994279	Labels 2	Common Venue Coffee Shop Chinese	Most Common Venue Discount Store Cocktail	Yoga Studio Dim Sum	Common Venue Steakhouse American	Common Venue Supplement Shop Vietnamese	Tennis Stadium	Shoe Store Noodle House	Common Venue Gym	Common Venue Bank Bubble	10th Mos Common Venuu Seafoot Restauran Ice Crean Shop Mexicar Restauran
0]:	manhattan_m leighborhood Marble Hill Chinatown Washington	Latitude 40.876551 40.715618	-73.910660 -73.994279 -73.936900	Labels 2 2	Common Venue Coffee Shop Chinese Restaurant	Most Common Venue Discount Store Cocktail Bar	Yoga Studio Dim Sum Restaurant Mobile Phone	Common Venue Steakhouse American Restaurant	Supplement Shop Vietnamese Restaurant	Common Venue Tennis Stadium Salon / Barbershop	Shoe Store Noodle House	Gym Bakery Latin American	Bank Bubble Tea Shop Tapas Restaurant Frozen	10th Mos Common Venue Seafoon Restauran Ice Crean Shop Mexican

The comparison results of the statistical results of the two regions are as follows.

Compare the information from the two regions

[56]	: ! pip install datacompy											
	<pre>import datacompy, pandas as pd, sys compare = datacompy. Compare(toronto_merged, manhattan_merged, join_columns=['1st Most Common Venue', '2nd Most Common Venue' print(compare.matches())</pre>											
	print(compare.report())											
	4											
	28 Downtown Toronto	D.I.	-			43. 646435 -79. 374846	0.0					
	Coffee Shop 9 Central Toronto Su		, Rathnelly, So	outh Hill, Forest H			0.0					
	Coffee Shop	Pub	Bagel S	Shop Supe		etnamese Restaurant	(2002)					
	10 Downtown Toronto					43. 679563 -79. 377529	2. 0					
		ail	Playground	Dance Studio								
	31 West Toronto	D 1				43. 669005 -79. 442259	0. 0					
		Bakery	Grocery Store	Athletics & Sp			0.0					
	3 East Toronto Café Coffee Sh	1	0 1	Bakerv		43. 659526 -79. 340923	0. 0					
		5.5. * (2)					3. 0					
	Jewelry Store				estaurant	43.696948 -79.411307 Yoga Studio	3. 0					
	33 West Toronto	IIaII				43. 636847 -79. 428191	0. 0					
	Café Breakfast Sy	oot		Yoga Studio		Gvm	0.0					
	13 Downtown Toronto	JOC .	corree bhop			43. 654260 -79. 360636	0. 0					
	Coffee Shop	Café	p	ark	Pub	Bakery	0.0					
	4 Central Toronto					43, 728020 -79, 388790	2. 0					
	Park Bus L:	ine	Swim School	Department Store		ronics Store	2. 0					
	5 Central Toronto	555.5×				43. 712751 -79. 390197	0.0					

Thank you for your valuable time and careful browsing. Thanks to Coursera Team and Peers.

5. Discussion and Conclusion

This project is a good example of how to do data analysis and modeling. This project shows the whole process of data analysis from the initial data download or loading to data cleaning, to data analysis and final comparison. The only regret is that in the process of the project, I gradually realized that the initial assumption was unreasonable to a certain extent. In the process of the project, I constantly think about and gradually improve the direction and method of data analysis and modeling. In the end, the project gave me my own comparison goals and results.

There are also some possible future work on this project. For example, whether it is possible to show the final comparison results in the map. Or, if we take the lead in visual analysis on the map, is it possible to compare the visual analysis to get a higher level of conclusion?

I am very grateful to myself for this learning opportunity and to peers who have graded the quizzes and projects for me. Thanks to all the people who helped me in the learning process.