Compare the similarities of financial world cities

Jiaqing Chen
2020/08/23

Introduction

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.

A large city carries the development opportunities of a region. A world-class city is a place for investors and job seekers alike. 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.

Problems

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.

Data Description

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. The dataset exists for free on the web and the link to the dataset is shown as: 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. The link to the Wikipedia page of the dataset is shown as: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.

Data Usage Method

The two datasets will be processed in an incompatible way.

Here, we need to use a library called Folium, which is a great visualization library.

We'll segment it using the Foursquare API.

We need to group by adjacent rows and average the frequency of each category.

Use k-means method to cluster the neighborhood.

Methodology Overview

The whole data analysis and comparison section covers data download, cleaning, merging and analysis.

This project is 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.

Data Collation

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)
```

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	2
1	Manhattan	Chinatown	40.715618	-73.994279	2
2	Manhattan	Washington Heights	40.851903	-73.936900	4
3	Manhattan	Inwood	40.867684	-73.921210	3
4	Manhattan	Hamilton Heights	40.823604	-73.949688	0

In [16]:
```
manhattan_data.tail()
```

Out[16]:

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels
35	Manhattan	Turtle Bay	40.752042	-73.967708	3
36	Manhattan	Tudor City	40.746917	-73.971219	3
37	Manhattan	Stuyvesant Town	40.731000	-73.974052	4
38	Manhattan	Flatiron	40.739673	-73.990947	3
39	Manhattan	Hudson Yards	40.756658	-74.000111	2

Data Analysis

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)

        # create the API request URL
        url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            radius,
            LIMIT)

        # make the GET request
        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',
                  'Venue Latitude',
                  'Venue Longitude',
                  'Venue Category']

    return(nearby_venues)
```

Data Analysis

In [33]:
```
def return_most_common_venues(row, num_top_venues):
    row_categories = row.iloc[1:]
    row_categories_sorted = row_categories.sort_values(ascending=False)
    return row_categories_sorted.index.values[0:num_top_venues]
```

In [37]:
```
num_top_venues = 10

indicators = ['st', 'nd', 'rd']

# create columns according to number of top venues
columns = ['Neighborhood']
for ind in np.arange(num_top_venues):
    try:
        columns.append('{}{} Most Common Venue'.format(ind+1, indicators[ind]))
    except:
        columns.append('{}th Most Common Venue'.format(ind+1))

# create a new dataframe
neighborhoods_venues_sorted = pd.DataFrame(columns=columns)
neighborhoods_venues_sorted['Neighborhood'] = ca_toronto_grouped['Neighborhood']

for ind in np.arange(ca_toronto_grouped.shape[0]):
    neighborhoods_venues_sorted.iloc[ind, 1:] = return_most_common_venues(ca_toronto_grouped.iloc[ind, :], num_top_venues)

neighborhoods_venues_sorted.head()
```

Foursquare

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:

    package                    |            build
    ---------------------------|-----------------
    conda-4.8.4                |   py38h32f6830_2         3.1 MB  conda-forge
    ------------------------------------------------------------
                                           Total:         3.1 MB

The following packages will be UPDATED:

  conda                              4.8.4-py38h32f6830_1 --> 4.8.4-py38h32f6830_2



Downloading and Extracting Packages
```

Results

Out[49]:

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	Ice 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

In [50]:

```
manhattan_merged.head()
```

Out[50]:

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
Marble Hill	40.876551	-73.910660	2	Coffee Shop	Discount Store	Yoga Studio	Steakhouse	Supplement Shop	Tennis Stadium	Shoe Store	Gym	Bank	Seafood Restaurant
Chinatown	40.715618	-73.994279	2	Chinese Restaurant	Cocktail Bar	Dim Sum Restaurant	American Restaurant	Vietnamese Restaurant	Salon / Barbershop	Noodle House	Bakery	Bubble Tea Shop	Ice Cream Shop
Washington Heights	40.851903	-73.936900	4	Café	Bakery	Mobile Phone Shop	Pizza Place	Sandwich Place	Park	Gym	Latin American Restaurant	Tapas Restaurant	Mexican Restaurant
Inwood	40.867684	-73.921210	3	Mexican Restaurant	Lounge	Pizza Place	Café	Wine Bar	Bakery	American Restaurant	Park	Frozen Yogurt Shop	Spanish Restaurant
Hamilton Heights	40.823604	-73.949688	0	Mexican Restaurant	Coffee Shop	Café	Deli / Bodega	Pizza Place	Liquor Store	Indian Restaurant	Sushi Restaurant	Sandwich Place	Yoga Studio

Results

Compare the information from the two regions

In [56]:
```
! pip install datacompy

import datacompy, pandas as pd, sys

compare = datacompy.Compare(toronto_merged, manhattan_merged, join_columns=['1st Most Common Venue','2nd Most Common Venue','3rd Most Comm
print(compare.matches())
print(compare.report())
```

```
28  Downtown Toronto                                              Stn A PO Boxes  43.646435 -79.374846          0.0
Coffee Shop                 Pub                Café              Beer Bar                 Restaurant
9    Central Toronto  Summerhill West, Rathnelly, South Hill, Forest Hill SE, Deer Park  43.686412 -79.400049          0.0
Coffee Shop                 Pub             Bagel Shop            Supermarket        Vietnamese Restaurant
10  Downtown Toronto                                                    Rosedale  43.679563 -79.377529          2.0
Park                 Trail            Playground          Dance Studio  Eastern European Restaurant
31       West Toronto                                  Dufferin, Dovercourt Village  43.669005 -79.442259          0.0
Pharmacy                 Bakery          Grocery Store    Athletics & Sports          Gym / Fitness Center
3        East Toronto                                               Studio District  43.659526 -79.340923          0.0
Café            Coffee Shop            Gastropub               Bakery                    Brewery
23    Central Toronto                        Forest Hill North & West, Forest Hill Road Park  43.696948 -79.411307          3.0
Jewelry Store               Trail     Mexican Restaurant      Sushi Restaurant                Yoga Studio
33       West Toronto                          Brockton, Parkdale Village, Exhibition Place  43.636847 -79.428191          0.0
Café         Breakfast Spot           Coffee Shop           Yoga Studio                      Gym
13  Downtown Toronto                                         Regent Park, Harbourfront  43.654260 -79.360636          0.0
Coffee Shop                Café                     Park                 Pub                       Bakery
4     Central Toronto                                                 Lawrence Park  43.728020 -79.388790          2.0
Park                Bus Line           Swim School      Department Store          Electronics Store
5     Central Toronto                                              Davisville North  43.712751 -79.390197          0.0
Park       Department Store                Hotel          Dance Studio                 Food & Drink Shop
```

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

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.

Future Work

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?

Thank You For Your Time