Neighborhoods Similarity between NYC and Chicago

Wansu Zhan
11/25/2018

Contents

Introduction

Chicago based firms, such as the one I am working for, are constantly hiring talents from all over the country. One common question they are asked by candidates graduated from universities that are based in New York, or are near New York (such as Columbia University, New York University, Cornell University, etc.), is that how Chicago is compared to New York. Those students, who get themselves quite familiar to the New York City and its lifestyle, want to know if they can fit well in Chicago. This project tries to answer such question from the neighborhood similarity perspective. We would like to utilize cluster analysis to help us determine (1) whether there are similarity at all between New York and Chicago neighborhoods (2) If there is, which neighborhoods between New York City and Chicago are similar to each other.

The report may be interesting to people who are considering moving from New York to Chicago, or the other way around. It would also help Chicago and New York based firms answer candidates' question regarding the similarity between the two cities.

Data Description

There are mainly two types of data that are used in this project.

(1) The neighborhoods and their coordinates of New York City and Chicago.

- For New York, the data is got from the Week3 Lab of the Coursera Course "Applied data Science Capstone".
- For Chicago, we first use python to scrape the neighborhoods of Chicago from Wikipedia "***https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Chicago***". Then we got the coordinates of each neighborhood from the combination of below two approaches
    i. Use python to scrape the coordinates of each Chicago neighborhood from the Google Search link. For example, to look up the coordinates of the Neighborhood "Printer's Row", whose Community Area is "The Loop", we would use python to scrape the link https://www.google.com/search?ei=GTT3W6_kKY_vjwTlg66oDQ&q=Printer%27s+Row+The+Loop+chicago+coordinates. Generally, we would use the link https://www.google.com/search?ei=GTT3W6_kKY_vjwTlg66oDQ&q={%s}+{%s}+chicago+coordinates, where the first "%s" would be replaced by the name of the neighborhood, and the second "%s" would be replaced by the name of the Community Area/Borough.
        ✧ Note that we tried to use the package Nominatim of geopy as well, however it fails to get results for a lot of cases. Therefore, we switched to use the google search method.
    ii. Manual correction. Most of the results got from above approach are correct. However, there are a few wrong coordinates that need to be manually corrected

- The total number of distinct neighborhoods we have for New York City is 302, while the total number of distinct neighborhoods we have for Chicago is 119.
- Below shows examples of neighborhoods data of New York and Chicago with their coordinates.

```
[173]: df_ny.head(2)
```

[173]:

	Borough	Neighborhood	Latitude	Longitude	City
0	Bronx	Wakefield	40.894705	-73.847201	New York
1	Bronx	Co-op City	40.874294	-73.829939	New York

```
[174]: df_chi.head(2)
```

[174]:

	Neighborhood	Borough	Latitude	Longitude	City
0	The Loop	The Loop	41.878635	-87.625055	Chicago
1	Near East Side	The Loop	41.886100	-87.617800	Chicago

(2) The venues for each neighborhood. This data is acquired from Foursquare API by using the "explore" functionality.

- Foursquare is a technology company that built a massive dataset of location data. They crowd-sourced their data and had people use their app to build their dataset and add venues and complete any missing information they had in their dataset. Currently its location data is the most comprehensive out there, and quite accurate that it powers location data for many popular services like Apple Maps, Uber, Snapchat, Twitter and many others, and is currently being used by over 100,000 developers, and this number is only growing.
- For our purpose, we use Foursquare API to grab nearby venues that are within 1 km of each neighborhood. The returned venue information include venue's coordinate, name, venue category, etc.
- We can further determine the popular venue categories for each neighborhood, and utilize this data for our cluster analysis. The methodology session below will discuss detailed description of the approach.
- Below shows examples of venue data returned by Foursquare API for each neighborhood, combined with Neighborhood coordinates, Borough and City information.

```
[175]: df_venues_ny.head(2)
```

[175]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Borough	Venue	Venue Latitude	Venue Longitude	Venue Category	City
0	Wakefield	40.894705	-73.847201	Bronx	Lollipops Gelato	40.894123	-73.845892	Dessert Shop	New York
1	Wakefield	40.894705	-73.847201	Bronx	Ripe Kitchen & Bar	40.898152	-73.838875	Caribbean Restaurant	New York

```
[176]: df_venues_chi.head(2)
```

[176]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Borough	Venue	Venue Latitude	Venue Longitude	Venue Category	City
0	The Loop	41.878635	-87.625055	The Loop	The Art Institute of Chicago	41.879665	-87.62363	Art Museum	Chicago
1	The Loop	41.878635	-87.625055	The Loop	Symphony Center (Chicago Symphony Orchestra)	41.879275	-87.62468	Concert Hall	Chicago

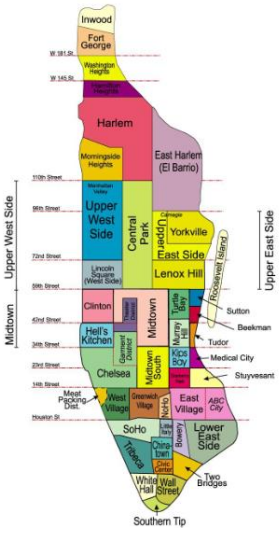
Methodology

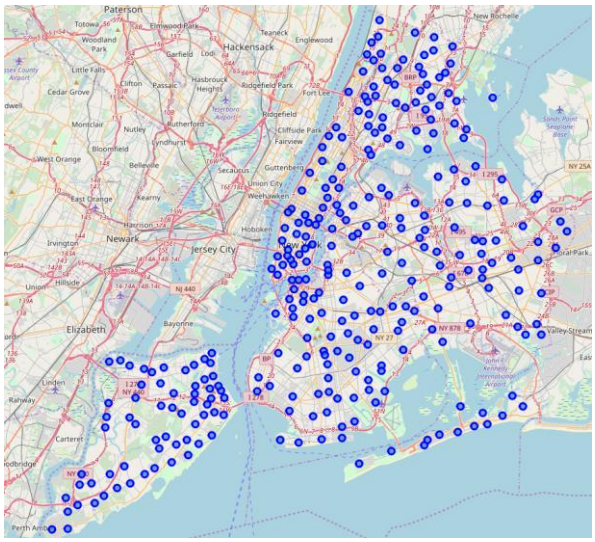
Neighborhood Exploration

We first explore the neighborhood data we have by visualizing graphically for each neighborhood of New York and Chicago as below.

New York City

The graph on the left shows the neighborhood/Borough names of New York Manhattan[1], while the graph on the right shows the coordinates of each New York Neighborhood we choose to perform analysis for in this project.



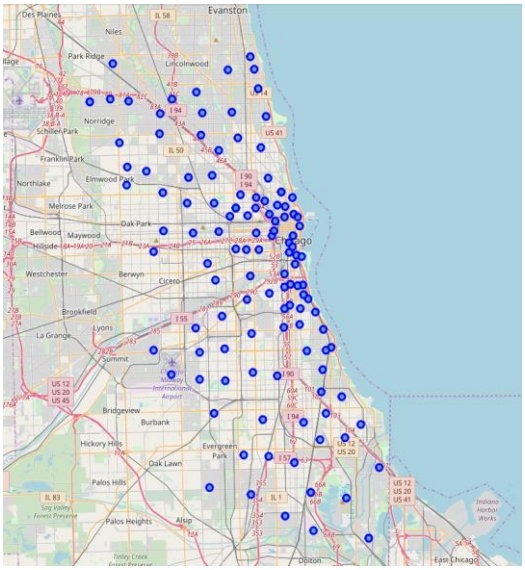


Chicago

The graph on the left shows the neighborhood/Borough names of Chicago[2], while the graph on the right shows the coordinates of each Chicago Neighborhood we choose to perform analysis for in this project.

[1] See https://www.pinterest.com/pin/302233824963969541/

[2] See https://sk.wikipedia.org/wiki/S%C3%BAbor:Chicago_community_areas_map.svg



Venues Exploration

For each neighborhood, we use Foursquare API to get the names and categories of the venues that are within 1km of the neighborhood coordinates. There are 462 and 355 unique venue categories for New York and Chicago respectively and 501 unique venue categories in total for both of them.

At this point, we combined New York and Chicago venues data, and used One-Hot-Encoding for all the venue categories. After this, we got a dataframe with each venue category being a column, and each venue of a neighborhood being a row entry. The value of the dataframe would be either 0 or 1, indicating whether the venue belongs to the venue category within 1km. We then aggregate the dataframe by Neighborhood, and calculate the mean of the Venue Categories, so that we get the portion of each venue category for each neighborhood. This will prepare us for performing clustering later.

[28]:
```
# one hot encoding
onehot_all = pd.get_dummies(df_venues_all[['Venue Category']], prefix="", prefix_sep="")

# add neighborhood column back to dataframe
onehot_all['Neighborhood'] = df_venues_all['Neighborhood']
onehot_all['Borough'] = df_venues_all['Borough']
onehot_all['City'] = df_venues_all['City']

# move Borough and City columns to the first two columns
fixed_columns = list(onehot_all.columns[-2:]) + list(onehot_all.columns[:-2])
onehot_all = onehot_all[fixed_columns]

df_grouped = onehot_all.groupby(['Neighborhood', 'Borough', 'City']).mean().reset_index()
df_grouped.head(2)
```

[28]:

	Neighborhood	Borough	City	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Lounge	...	Weight Loss Center	Whisky Bar	Wine Bar	Wine Shop	Winery	Wings Joint	Women's Store	Yoga Studio	Zoo	Zoo Exhibit
0	Albany Park	Albany Park	Chicago	0.0	0.011628	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.011628	0.0	0.0	0.0	0.0
1	Allerton	Bronx	New York	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0

In addition, we also get the top 10 common venue categories for each neighborhood for further cluster results examination. The data looks like below

```
[151]: neighborhoods_venues_sorted.head(2)
```

	Neighborhood	Borough	City	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
0	Albany Park	Albany Park	Chicago	Mexican Restaurant	Grocery Store	Discount Store	Pizza Place	Hookah Bar	Donut Shop	Sandwich Place	Chinese Restaurant	Bakery	Middle Eastern Restaurant
1	Allerton	Bronx	New York	Pizza Place	Fast Food Restaurant	Donut Shop	Caribbean Restaurant	Bus Station	Sandwich Place	Pharmacy	Mexican Restaurant	Dessert Shop	Supermarket

Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics[3]. Clustering is suitable for our problem here, as it can help us to understand the similarity between neighborhoods by determining which neighborhoods belong to the same clusters.

One of the simplest models among clustering approaches is K means. In this project, we will use K means to perform cluster analysis. The feature we choose to use is the portion of each venue category for each neighborhood, as calculated in previous section.

We will first perform cluster analysis for all neighborhoods of New York and Chicago. Then we will perform cluster analysis for only the downtown area of New York and Chicago. Here for New York, a neighborhood belongs to “downtown” if its borough is Manhattan; while for Chicago, a neighborhood belongs to “downtown” if its borough is The Loop, Near North Side, Near South Side, or Near West Side[4].

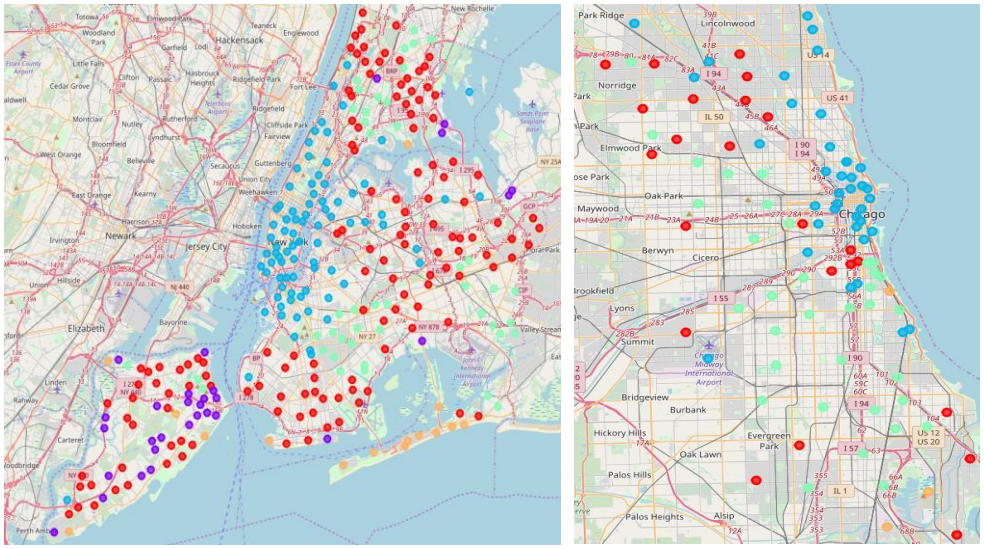
Results

Using All Neighborhoods

Using 5 clusters, below graphs show the clustering results for all neighborhoods of New York and Chicago neighborhoods, with each color representing a different cluster. New York is on the left while Chicago is on the right.

[3] See https://en.wikipedia.org/wiki/Cluster_analysis

[4] The east side of Chicago Loop is the Michigan Lake ☺

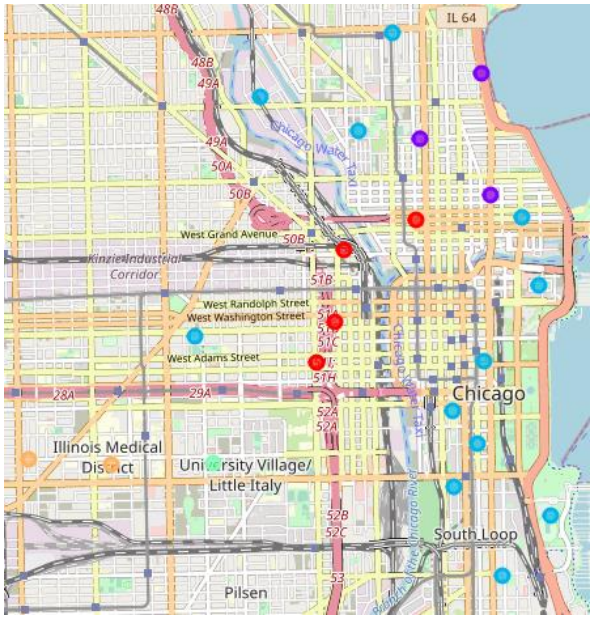


As shown from the graph, we see that there are indeed similarities between New York and Chicago neighborhoods, from the perspective of nearby venues. For example, as one might have guessed, Manhattan area is very similar to Chicago downtown area, as indicated by blue dots in the graphs. In addition, some New York neighborhoods in Bronx, Staten Island, Brooklyn and Queens are similar to neighborhoods in Chicago suburbs in the south and west, indicating by red dots and green dots. We also found that some neighborhoods in Staten Island at New York is not similar to any Chicago neighborhoods, as indicated by purple dots. Please see Appendix for more information of the cluster results. We also vary the number of clusters used, and the main results will not change much.

Using Downtown Neighborhoods Only

We then performed a similar but more detailed analysis to only downtown area for both cities, with the attempt to further classify these areas. Note that there are 40 New York neighborhoods and 22 Chicago neighborhoods belong to the downtown area we defined.

The clustered results using 5 clusters are as below, with New York on the left and Chicago on the right.



Below table summarizes the proportion of clusters that each Borough belong to for New York/Chicago downtown area, with color of each cluster corresponding to the color shown in the map.

Borough	City	#Neighborhoods	1	2	3	4	5
Manhattan	NYC	40	0.10	0.35	0.23	0.30	0.03
Near North Side	CHI	8	0.13	0.38	0.50	-	-
Near South Side	CHI	3	-	-	1.00	-	-
Near West Side	CHI	7	0.43	-	0.14	0.14	0.29
The Loop	CHI	4	-	-	1.00	-	-

Now let's examine each cluster further.

Cluster1 (red dots)

There are 8 neighborhoods in total belong to this cluster, as listed below.

- New York
  - Flatiron, Gramercy, Midtown South, Murray Hill
- Chicago
  - Fulton River District, Greektown, River North, West Loop

Below table shows the top 10 common venues for neighborhoods that belong to this cluster:

```
[160]: df_merged_downtown.loc[df_merged_downtown['Cluster Labels'] == 0, :]
```

[160]:

	Neighborhood	Borough	City	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	Cluster Labels	Latitude	Longitude
12	Flatiron	Manhattan	New York	Gym	Gym / Fitness Center	Cycle Studio	New American Restaurant	American Restaurant	Italian Restaurant	Spa	Park	Wine Shop	Vegetarian / Vegan Restaurant	0	40.739673	-73.990947
13	Fulton River District	Near West Side	Chicago	New American Restaurant	Italian Restaurant	Coffee Shop	Bar	Restaurant	Mexican Restaurant	Gym	Burger Joint	Latin American Restaurant	Café	0	41.889495	-87.643364
16	Gramercy	Manhattan	New York	American Restaurant	New American Restaurant	Restaurant	Mediterranean Restaurant	Cheese Shop	Park	Mexican Restaurant	Cosmetics Shop	Cocktail Bar	Juice Bar	0	40.737210	-73.981376
17	Greektown	Near West Side	Chicago	Greek Restaurant	New American Restaurant	Coffee Shop	Italian Restaurant	Pizza Place	Sandwich Place	Café	Bar	Grocery Store	Gym	0	41.878564	-87.647050
33	Midtown South	Manhattan	New York	Korean Restaurant	Gym / Fitness Center	Coffee Shop	Sandwich Place	Japanese Restaurant	Yoga Studio	Pizza Place	Hotel	Italian Restaurant	Spa	0	40.748510	-73.988713
35	Murray Hill	Manhattan	New York	Gym / Fitness Center	Japanese Restaurant	Korean Restaurant	Coffee Shop	Gym	Pizza Place	Chinese Restaurant	Sandwich Place	American Restaurant	Gourmet Shop	0	40.748303	-73.978332
44	River North	Near North Side	Chicago	Steakhouse	Italian Restaurant	Restaurant	Mexican Restaurant	Gym	Coffee Shop	Bar	Hotel	Sushi Restaurant	Wine Bar	0	41.892385	-87.634075
59	West Loop	Near West Side	Chicago	New American Restaurant	Coffee Shop	Grocery Store	Breakfast Spot	Greek Restaurant	Pizza Place	Bar	Sandwich Place	Italian Restaurant	Mediterranean Restaurant	0	41.882457	-87.644678

The venues that appear in the top 10 categories by frequency for this cluster are displayed as below, where venues like Italian Restaurant, Coffee Shop, New American Resaurant, Gym etc. show up.

[162]:

	Counts	VenueCategory
35	6	Italian Restaurant
34	6	Coffee Shop
33	5	New American Restaurant
32	5	Gym
31	4	Sandwich Place
30	4	Pizza Place
29	4	Bar
26	3	Gym / Fitness Center
25	3	American Restaurant
28	3	Restaurant
27	3	Mexican Restaurant
19	2	Hotel
24	2	Spa
23	2	Park
22	2	Mediterranean Restaurant
21	2	Korean Restaurant
20	2	Japanese Restaurant
18	2	Grocery Store

Cluster2 (purple dots)

One can see from the graph that New York neighborhoods in lower Manhattan, such as Soho and Little Italy, as well as some neighborhoods in the middle of Manhattan, such as Lincoln Square and Upper East Side, are similar to Chicago neighborhoods in the near north, such as Magnificent Mile and Gold Coast, and they all belong to this purple cluster.

Below table shows the top 10 common venues for neighborhoods that belong to this cluster:

```
[130]: df_merged_downtown.loc[df_merged_downtown['Cluster Labels'] == 1, :]
```

[130]:

	Neighborhood	Borough	City	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	Cluster Labels	Latitude	Longitude
4	Chelsea	Manhattan	New York	Art Gallery	American Restaurant	Coffee Shop	Hotel	Seafood Restaurant	Nightclub	Italian Restaurant	Bakery	Bagel Shop	Tapas Restaurant	1	40.744035	-74.003116
5	Chinatown	Manhattan	New York	Chinese Restaurant	Cocktail Bar	Ice Cream Shop	Wine Bar	Sandwich Place	Café	American Restaurant	French Restaurant	Shoe Store	Thai Restaurant	1	40.715618	-73.994279
6	Civic Center	Manhattan	New York	French Restaurant	Bakery	Coffee Shop	Hotel	Spa	Chinese Restaurant	Cocktail Bar	Men's Store	American Restaurant	Ice Cream Shop	1	40.715229	-74.005415
14	Gold Coast	Near North Side	Chicago	Hotel	Italian Restaurant	Gym	American Restaurant	Salon / Barbershop	Café	Coffee Shop	Steakhouse	Yoga Studio	Restaurant	1	41.906699	-87.625331
18	Greenwich Village	Manhattan	New York	Italian Restaurant	Coffee Shop	Seafood Restaurant	American Restaurant	Pizza Place	Clothing Store	Spa	Café	Indie Movie Theater	French Restaurant	1	40.726933	-73.999914
23	Lenox Hill	Manhattan	New York	Italian Restaurant	Sushi Restaurant	French Restaurant	Gym / Fitness Center	Dessert Shop	Bakery	Coffee Shop	Café	Burger Joint	Spanish Restaurant	1	40.768113	-73.958860
24	Lincoln Square	Manhattan	New York	Italian Restaurant	Gym / Fitness Center	French Restaurant	Jazz Club	Gym	Bakery	Theater	Coffee Shop	Performing Arts Venue	Concert Hall	1	40.773529	-73.985338
25	Little Italy	Manhattan	New York	Men's Store	Clothing Store	Café	Italian Restaurant	Cocktail Bar	Chinese Restaurant	Coffee Shop	Shoe Store	Pizza Place	Women's Store	1	40.719324	-73.997305

The venues that appear in the top 10 categories by frequency for this cluster are displayed as below, where venues like Italian Restaurant, Coffee Shop, American Resaurant, as well as Hotel, Gym, Men's Store etc. show up.

```
[151]: df_venues1.sort_values('Counts', ascending=False)
```

[151]:

	Counts	VenueCategory
52	15	Italian Restaurant
51	15	Coffee Shop
50	11	American Restaurant
49	8	French Restaurant
48	8	Bakery
47	7	Pizza Place
46	7	Hotel
45	6	Café
44	5	Gym / Fitness Center
43	5	Cocktail Bar
38	4	Men's Store
35	4	Chinese Restaurant
36	4	Gym
37	4	Ice Cream Shop
40	4	Shoe Store
39	4	Seafood Restaurant
41	4	Steakhouse

Full list of the 17 neighborhoods belong to this cluster are as below:

- New York
  - Chelsea, Chinatown, Civic Center, Greenwich Village, Lenox Hill, Lincoln Square, Little Italy, Lower East Side, Noho, Soho, Sutton Place, Tribeca, Upper East Side, West Village
- Chicago
  - Gold Coast, Magnificent Mile, Near North Side

Cluster3 (blue dots)

There are 21 neighborhoods in total belong to this cluster, as listed below.

- New York
  - Battery Park City, Carnegie Hill, Central Harlem, Clinton, Financial District, Hudson Yards, Midtown, Tudor City, Turtle Bay
- Chicago
  - Cabrini-Green, Dearborn Park, Goose Island, Museum Campus, Near East Side, Near West Side, Old Town, Prairie Avenue Historic District, Printer's Row, South Loop, Streeterville, The Loop

Note that all neighborhoods of the two Chicago Boroughs, "Near South Side" and "The Loop",

belong to this cluster.

Below table shows the top 10 common venues for neighborhoods that belong to this cluster:

```
[164]: df_merged_downtown.loc[df_merged_downtown['Cluster Labels'] == 2, :]
```

[164]:

	Neighborhood	Borough	City	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	Cluster Labels	Latitude	Longitude
0	Battery Park City	Manhattan	New York	Park	Coffee Shop	Wine Shop	Hotel	American Restaurant	Gym	Plaza	Dog Run	Gym / Fitness Center	BBQ Joint	2	40.711932	-74.016869
1	Cabrini-Green	Near North Side	Chicago	Coffee Shop	Gym / Fitness Center	Gym	American Restaurant	Breakfast Spot	Bar	Italian Restaurant	Café	Bakery	Park	2	41.901091	-87.641464
2	Carnegie Hill	Manhattan	New York	Pizza Place	Coffee Shop	Gym	Yoga Studio	Italian Restaurant	Bakery	Art Museum	Cocktail Bar	Café	Spa	2	40.782683	-73.953256
3	Central Harlem	Manhattan	New York	Southern / Soul Food Restaurant	Café	French Restaurant	African Restaurant	Seafood Restaurant	Theater	Sushi Restaurant	American Restaurant	Coffee Shop	Gym / Fitness Center	2	40.815976	-73.943211
7	Clinton	Manhattan	New York	Theater	Italian Restaurant	American Restaurant	Burger Joint	Hotel	Coffee Shop	Indie Theater	Wine Shop	Gym / Fitness Center	Bakery	2	40.759101	-73.996119
8	Dearborn Park	Near South Side	Chicago	Pizza Place	Grocery Store	Coffee Shop	Gym / Fitness Center	Park	Breakfast Spot	Burger Joint	American Restaurant	Sushi Restaurant	Yoga Studio	2	41.866429	-87.629000

The venues that appear in the top 10 categories by frequency for this cluster are displayed as below, where venues like Coffee Shop, American Resaurant, Hotel, Pizza Place, Park etc. show up.

	Counts	VenueCategory
68	18	Coffee Shop
67	12	American Restaurant
66	11	Hotel
65	10	Pizza Place
64	10	Park
63	9	Gym / Fitness Center
62	8	Italian Restaurant
60	7	Gym
59	7	Burger Joint
61	7	Theater
58	6	Café
57	6	Bar
56	5	Grocery Store
55	5	Bakery
51	4	Seafood Restaurant
49	4	Breakfast Spot

Cluster4 (green dots)

There are 13 neighborhoods in total belong to this cluster, as listed below.

- New York
  - East Harlem, East Village, Hamilton Heights, Inwood, Manhattan Valley, Manhattanville, Morningside Heights, Roosevelt Island, Stuyvesant Town, Upper West Side, Washington Heights, Yorkville
- Chicago
  - Little Italy/University Village

Below table shows the top 10 common venues for neighborhoods that belong to this cluster:

```
[167]: df_merged_downtown.loc[df_merged_downtown['Cluster Labels'] == 3, :]
```

[167]:

	Neighborhood	Borough	City	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	Cluster Labels	Latitude	Longitude
9	East Harlem	Manhattan	New York	Bakery	Mexican Restaurant	Pizza Place	Café	Plaza	Thai Restaurant	Latin American Restaurant	Deli / Bodega	Park	Cocktail Bar	3	40.792249	-73.944182
10	East Village	Manhattan	New York	Cocktail Bar	Ice Cream Shop	Coffee Shop	Wine Bar	Italian Restaurant	Japanese Restaurant	Vegetarian / Vegan Restaurant	Bar	Chinese Restaurant	Pizza Place	3	40.727847	-73.982226
19	Hamilton Heights	Manhattan	New York	Coffee Shop	Mexican Restaurant	Café	Bar	Yoga Studio	Park	Scenic Lookout	Chinese Restaurant	Caribbean Restaurant	American Restaurant	3	40.823604	-73.949688
22	Inwood	Manhattan	New York	Latin American Restaurant	Pizza Place	Mexican Restaurant	Café	Wine Bar	Lounge	Deli / Bodega	Spanish Restaurant	Chinese Restaurant	Bakery	3	40.867684	-73.921210
26	Little Italy/University Village	Near West Side	Chicago	Sandwich Place	Italian Restaurant	Bar	Thai Restaurant	Pizza Place	Park	Breakfast Spot	Theater	Pharmacy	Donut Shop	3	41.868607	-87.660579
29	Manhattan Valley	Manhattan	New York	Park	Coffee Shop	Indian Restaurant	Pizza Place	Bar	Grocery Store	Mexican Restaurant	Chinese Restaurant	Ice Cream Shop	Dog Run	3	40.797307	-73.964286
30	Manhattanville	Manhattan	New York	Mexican Restaurant	American Restaurant	Seafood Restaurant	Italian Restaurant	Park	Café	Tennis Court	Coffee Shop	Southern / Soul Food Restaurant	Art Gallery	3	40.816934	-73.957385
34	Morningside Heights	Manhattan	New York	Coffee Shop	Italian Restaurant	Park	American Restaurant	Seafood Restaurant	Bakery	Mexican Restaurant	Bookstore	Restaurant	Café	3	40.808000	-73.963896

The venues that appear in the top 10 categories by frequency for this cluster are displayed as below, where venues like Park, Pizza Place, Mexican Restaurant, Coffee Shop, Bar etc. show up.

	Counts	VenueCategory
44	10	Park
43	9	Pizza Place
42	9	Mexican Restaurant
41	9	Coffee Shop
40	9	Bar
39	8	Italian Restaurant
38	7	Café
37	5	Ice Cream Shop
36	5	Bakery
35	5	American Restaurant
33	4	Chinese Restaurant
34	4	Deli / Bodega
32	3	Wine Bar
31	3	Thai Restaurant
30	3	Seafood Restaurant
29	3	Latin American Restaurant
28	3	Cocktail Bar
27	2	Grocery Store

Cluster5 (orange dots)

There are only 3 neighborhoods in total belong to this cluster, as listed below.

- New York
  - Marble Hill
- Chicago
  - Illinois Medical District, Tri-Taylor

Below table shows the top 10 common venues for neighborhoods that belong to this cluster:

```
[171]: df_merged_downtown.loc[df_merged_downtown['Cluster Labels'] == 4, :]
```

[171]:

	Neighborhood	Borough	City	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	Cluster Labels	Latitude	Longitude
21	Illinois Medical District	Near West Side	Chicago	Sandwich Place	Pizza Place	Thai Restaurant	Mexican Restaurant	Train Station	Sports Bar	Middle Eastern Restaurant	Italian Restaurant	Bar	Hot Dog Joint	4	41.868494	-87.673975
31	Marble Hill	Manhattan	New York	Park	Mexican Restaurant	Pharmacy	Spanish Restaurant	Supermarket	Donut Shop	Café	Pizza Place	Coffee Shop	Shoe Store	4	40.876551	-73.910660
52	Tri-Taylor	Near West Side	Chicago	Fast Food Restaurant	Café	Bus Station	Sandwich Place	Park	Pizza Place	Hot Dog Joint	Bus Line	Lounge	Bike Rental / Bike Share	4	41.869102	-87.684744

The venues that appear in the top 10 categories by frequency for this cluster are displayed as

below, where venues like Pizza Place, Sandwich Place, Park etc. show up.

	Counts	VenueCategory
22	3	Pizza Place
21	2	Sandwich Place
20	2	Park
19	2	Mexican Restaurant
18	2	Hot Dog Joint
17	2	Café
12	1	Spanish Restaurant
16	1	Train Station
15	1	Thai Restaurant
14	1	Supermarket
13	1	Sports Bar
0	1	Bar
1	1	Bike Rental / Bike Share

Discussion

From the clustering results, we can see that when using all neighborhoods data, we see that Manhattan area is very similar to Chicago downtown area. In addition, some New York neighborhoods in Bronx, Staten Island, Brooklyn and Queens are similar to neighborhoods in Chicago suburbs in the south and west. That being said, some New York neighborhoods, such as those in Staten Island, are not similar to any Chicago neighborhoods.

When using only downtown area data, we further understand the similarity between New York and Chicago neighborhoods. For example, New York neighborhoods in lower Manhattan, such as Soho and Little Italy, as well as some neighborhoods in the middle of Manhattan, such as Lincoln Square and Upper East Side, are similar to Chicago neighborhoods in the near north, such as Magnificent Mile and Gold Coast.

Note that the similarity in this analysis is all based on the proportion of nearby venues categories. If one has more time, we can further investigate the similarity between neighborhoods of the two cities from different angles. Using similar venue dataset, we can explore similarity using convenience score, or focusing on specific venues, like restaurants. Using different dataset, we can explore similarity of neighborhoods from other perspectives, such as weather, relative house prices, lifestyle of people, etc.

Conclusion

In this paper, we use K means to perform cluster analysis to help us understand the similarity of neighborhoods of New York and Chicago. The data shows that there are indeed similarities among neighborhoods of these two cities. When using all neighborhoods data, one can see that Manhattan is very similar to Chicago downtown area. In addition, some New York neighborhoods in Bronx, Staten Island, Brooklyn and Queens are similar to neighborhoods in Chicago suburbs in

the south and west. That being said, some New York neighborhoods, such as those in Staten Island, are not similar to any Chicago neighborhoods.

When using only downtown area data, we further identified the similarity between New York and Chicago neighborhoods. For example, New York neighborhoods in lower Manhattan, such as Soho and Little Italy, as well as some neighborhoods in the middle of Manhattan, such as Lincoln Square and Upper East Side, are similar to Chicago neighborhoods in the near north, such as Magnificent Mile and Gold Coast.

The paper should help interested readers to understand the neighborhood similarity between New York and Chicago from nearby venue categories perspective. Given more time, one can further explore the similarities from different perspectives, as mentioned in the Discussion section above.

Appendix

Below table shows the proportion of cluster each Borough of New York and Chicago belongs to. There are 70 boroughs in total for both cities. The columns 1 to 5 below represent the 5 clusters, and the colors correspond to the colors used in the map.

Borough	City	#Neighborhoods	1	2	3	4	5
Albany Park	CHI	1	1.00	–	–	–	–
Archer Heights	CHI	1	–	–	–	1.00	–
Armour Square	CHI	3	0.67	–	0.33	–	–
Ashburn	CHI	1	–	–	–	1.00	–
Auburn Gresham	CHI	1	–	–	–	1.00	–
Austin	CHI	4	0.50	–	–	0.50	–
Austin, Humboldt Park	CHI	1	–	–	–	1.00	–
Avalon Park	CHI	1	–	–	–	1.00	–
Avondale	CHI	1	1.00	–	–	–	–
Belmont Cragin	CHI	1	1.00	–	–	–	–
Beverly	CHI	1	1.00	–	–	–	–
Bridgeport	CHI	1	1.00	–	–	–	–
Brighton Park	CHI	1	–	–	–	1.00	–
Bronx	NYC	52	0.60	0.06	0.06	0.27	0.02
Brooklyn	NYC	69	0.39	0.01	0.36	0.22	0.01
Burnside	CHI	1	–	–	–	1.00	–
Calumet Heights	CHI	1	–	–	–	1.00	–
Chatham	CHI	1	–	–	–	1.00	–
Chicago Lawn	CHI	1	–	–	–	1.00	–
Clearing	CHI	1	–	–	1.00	–	–
Douglas	CHI	7	0.14	–	0.14	0.71	–
Dunning	CHI	1	1.00	–	–	–	–
East Garfield Park	CHI	1	–	–	–	1.00	–

East Side	CHI	1	1.00	–	–	–	–
Edgewater	CHI	1	–	–	1.00	–	–
Edison Park	CHI	1	–	–	1.00	–	–
Englewood	CHI	1	–	–	–	1.00	–
Forest Glen	CHI	1	–	–	1.00	–	–
Fuller Park	CHI	1	–	–	–	1.00	–
Gage Park	CHI	1	–	–	–	1.00	–
Garfield Ridge	CHI	1	1.00	–	–	–	–
Grand Boulevard	CHI	1	–	–	–	1.00	–
Greater Grand Crossing	CHI	1	–	–	–	1.00	–
Hegewisch	CHI	1	1.00	–	–	–	–
Hermosa	CHI	1	1.00	–	–	–	–
Humboldt Park	CHI	1	–	–	–	1.00	–
Hyde Park	CHI	2	–	–	1.00	–	–
Irving Park	CHI	1	1.00	–	–	–	–
Jefferson Park	CHI	1	–	–	1.00	–	–
Kenwood	CHI	1	–	–	–	1.00	–
Lake View	CHI	1	–	–	1.00	–	–
Lincoln Park	CHI	2	–	–	1.00	–	–
Lincoln Square	CHI	1	–	–	1.00	–	–
Logan Square	CHI	1	–	–	1.00	–	–
Lower West Side	CHI	1	–	–	1.00	–	–
Manhattan	NYC	40	0.03	–	0.98	–	–
McKinley Park	CHI	1	–	–	–	1.00	–
Montclare	CHI	1	–	–	–	1.00	–
Morgan Park	CHI	1	–	–	–	1.00	–
Mount Greenwood	CHI	1	1.00	–	–	–	–
Near North Side	CHI	8	–	–	1.00	–	–
Near South Side	CHI	3	–	–	1.00	–	–
Near West Side	CHI	7	0.14	–	0.71	0.14	–
New City	CHI	1	–	–	–	1.00	–
North Center	CHI	1	–	–	1.00	–	–
North Lawndale	CHI	1	–	–	–	1.00	–
North Park	CHI	1	1.00	–	–	–	–
Norwood Park	CHI	2	0.50	–	–	0.50	–
O'Hare	CHI	1	1.00	–	–	–	–
Oakland	CHI	1	–	–	–	–	1.00
Portage Park	CHI	1	1.00	–	–	–	–
Pullman	CHI	1	–	–	–	1.00	–
Queens	NYC	81	0.53	0.04	0.17	0.15	0.11
Riverdale	CHI	1	–	–	–	–	1.00
Rogers Park	CHI	2	–	–	1.00	–	–
Roseland	CHI	1	–	–	–	1.00	–
South Chicago	CHI	1	1.00	–	–	–	–

South Deering	CHI	1	-	-	-	-	1.00
Staten Island	NYC	63	0.51	0.37	0.02	0.05	0.06
The Loop	CHI	4	-	-	1.00	-	-

1.