The battle of cities

Qihang Cai

April 1st,2020

1. Introduction

This project will mainly focus on the comparison of three districts: New York City, Toronto and Los Angeles. They are all economic and cultural centers in their area. They are all districts of my interests so in this project I will compare the segmentations of these three cities and find out the similarity of these three cities.

New York City (NYC), often called the City of New York or simply New York (NY), is the most populous city in the United States. With an estimated 2018 population of 8,398,748 distributed over about 302.6 square miles (784 km2), New York is also the most densely populated major city in the United States. (https://en.wikipedia.org/wiki/New_York_City)

Los Angeles(Spanish for '"The Angels"') ,officially the City of Los Angeles and often known by its initials L.A., is the largest city in the U.S. state of California. With an estimated population of nearly four million people, it is the country's second most populous city (after New York City) and the third most populous city in North America (after Mexico City and New York City).(https://en.wikipedia.org/wiki/Los_Angeles)

Toronto is the provincial capital of Ontario and the most populous city in Canada, with a population of 2,731,571 as of 2016.Current to 2016, the Toronto census metropolitan area (CMA), of which the majority is within the Greater Toronto Area (GTA), held a population of 5,928,040, making it Canada's most populous CMA. The city is the anchor of the Golden Horseshoe, an urban agglomeration of 9,245,438 people (as of 2016) surrounding the western end of Lake Ontario. Toronto is an international center of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world.(https://en.wikipedia.org/wiki/Toronto)

Based on the analysis done before, we will add Los Angeles to the project and compare the clusters of these three districts and find out the new information.

2. Data acquisition and cleaning

2.1Distict data

We get the data of three districts from different sources. Since the data of New York city can be obtain from the Internet, we can use it directly, but we should turn other sources to get the data of other districts. We combine the postal code in Wikipedia and coordinates from official website to form the data. After that we get that the Los Angeles county has 88 cities so we combine the data to form a csv.

Besides, we get the venue's data from the Foursquare API and Google map API geo-data we get have the information contains many sites in the neighborhoods of a certain location like

a park and restaurant.

I once wanted to explore the neighborhood of the city I live in, but unfortunately the Foursquare does not have the geo-data of China, so I take a step forward from the project I made about the Toronto. The main goal is to find the neighborhood that suits me in Toronto. If I have a job offer from a Toronto company, there are many factors that I will take into consideration except from the geo-information in the previous project like the people, the population and so on.

There are additional information about the condition of neighborhood in Toronto. Chinese means the Chinese population in the area, since I am from China and I will surely take this into consideration. Tenant Average Rent means the rent price level in the area, since I may not afford buying the house when I arrive in Toronto so I will rent an apartment in the area. Population density is another factor I will think about. Because I am a person interested in healthy life style, so Healthy Food Index may influence my judgment of an area. Social Housing Units,PFR,Community Space Use Early Development Instrument (EDI) Medical resources are important under current circumstances so Hospital Readmissions is an index I will look into.

2.2Data Cleaning

The data we get is raw and may cause some problems in the processing of our analysis,so the cleaning step is of necessity. We may meet following problems so we will give certain solutions.Since the same postal code may have different boroughs,we will combine the data with the same postal code and focus on one.There also some boroughs that are too far away from the center of the district that may be the outliner of the district so such boroughs are excluded from the dataframes.

Besides,the data we get from three sources are in different forms,so we will reform the dataframes and make them in the same format,which is the form of Neighborhood-Latitude-Longitude.

Here is a sample of the New York's data.

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Here is a sample of the Toronto's data.

	Postalcode	Borough	Neighbourhood	Latitude	Longitude
58	M5H	Downtown Toronto	Adelaide,King,Richmond	43.650571	-79.384568
12	M1S	Scarborough	Agincourt	43.794200	-79.262029
14	M1V	Scarborough	Agincourt North,L'Amoreaux East,Milliken,Steel...	43.815252	-79.284577
101	M9V	Etobicoke	Albion Gardens,Beaumond Heights,Humbergate,Jam...	43.739416	-79.588437
89	M8W	Etobicoke	Alderwood,Long Branch	43.602414	-79.543484

Here is a sample of the Los Angeles's data.

	Neighborhood	Latitude	Longitude
0	Agoura Hills	34.13600	-118.77500
1	Alhambra	34.09500	-118.12700
2	Arcadia	33.98095	-118.14697
3	Artesia	33.86607	-118.08294
4	Avalon	33.34300	-118.32800

If we want to get a map using the folium package, what we need is coordinates and label. The data above give us the information we need.

3. Methodology

3.1clustering

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.

3.2K-means

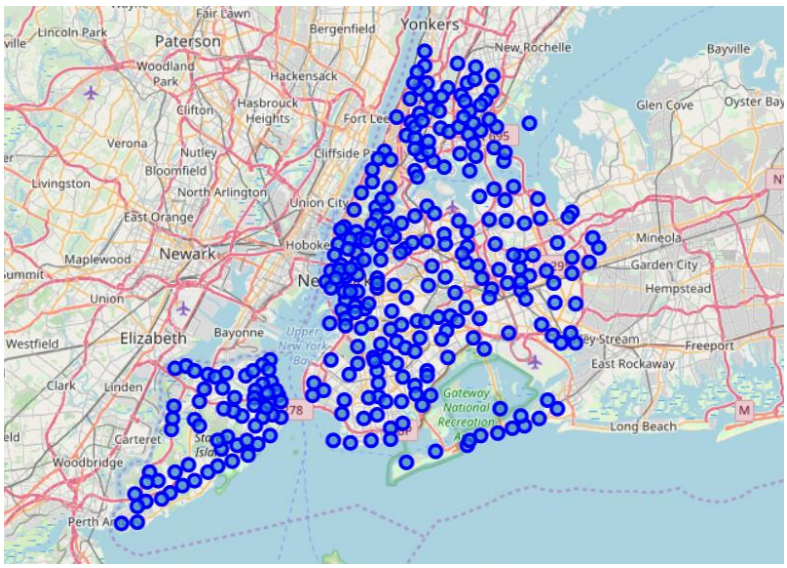
k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. It is popular for cluster analysis in data mining. k-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. For instance, Better Euclidean solutions can be found using k-medians and k-medoids.

4. Analysis

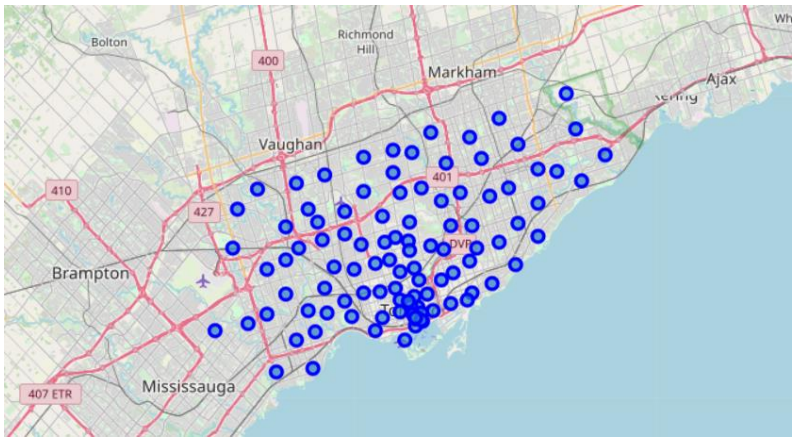
4.1Data visualization

First we give a glance to the map of the boroughs of the three districts.

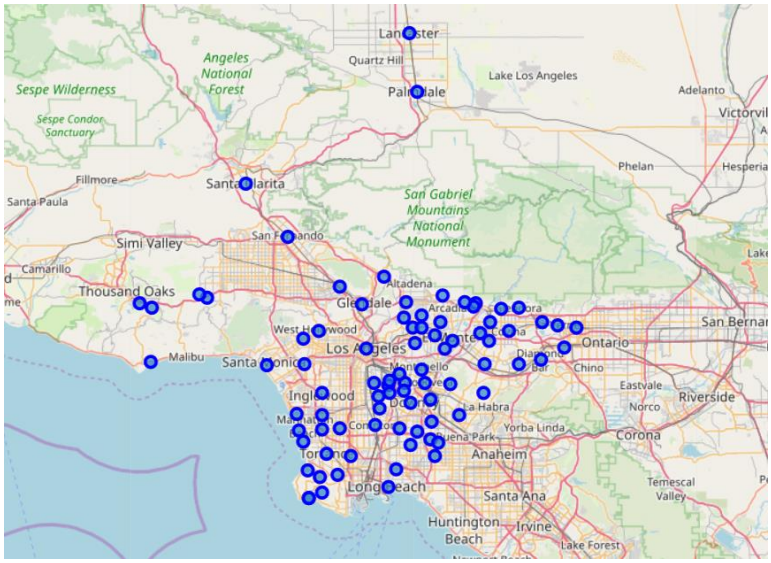
New York city



Toronto



Los Angeles



Because of the kind of the data we get,the boroughs in the New York city is much denser

than those in the other two districts.

4.2venues

Here is a screenshot of the ny's venues example.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
2	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
3	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

We use groupby command to count the amount of the venue category.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Allerton	28	28	28	28	28	28
Annadale	10	10	10	10	10	10
Arden Heights	4	4	4	4	4	4
Arlington	7	7	7	7	7	7
Arrochar	20	20	20	20	20	20
Arverne	18	18	18	18	18	18
Astoria	50	50	50	50	50	50
Astoria Heights	12	12	12	12	12	12
Auburndale	18	18	18	18	18	18
Bath Beach	49	49	49	49	49	49
Battery Park City	50	50	50	50	50	50

Like the picture shows, Yoga Studio, Accessories Store, Adult Boutique, Afghan Restaurant, African Restaurant, Airport Terminal ,American Restaurant, Antique Shop, Arcade Arepa Restaurant, Argentinian Restaurant, Art Gallery, Art Museum Arts Crafts Store, Arts& Entertainment, Asian Restaurant,Athletics&Sports,Auditorium,Australian Restaurant are some examples of the venues.

	Yoga Studio	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Terminal	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Mu
0	0	0	0	0	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	0	0	

Here is the labeled venues.

	Neighborhood	Yoga Studio	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Terminal	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant
0	Allerton	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
1	Annadale	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
2	Arden Heights	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
3	Arlington	0.0	0.0	0.0	0.0	0.0	0.0	0.142857	0.0	0.0	0.0	0.0
4	Arrochar	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0

Here is the most common venues in each neighborhood, this table can provide us with the most useful information about what kind of the neighborhood it is.They can also provide us with the sequences of the venues that will interests us.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Allerton	Pizza Place	Supermarket	Chinese Restaurant	Deli / Bodega	Check Cashing Service
1	Annadale	Pizza Place	Park	Train Station	Cosmetics Shop	Sports Bar
2	Arden Heights	Pizza Place	Pool	Pharmacy	Coffee Shop	Field
3	Arlington	Bus Stop	Food	Intersection	Deli / Bodega	American Restaurant
4	Arrochar	Deli / Bodega	Italian Restaurant	Bus Stop	Pizza Place	Food Truck

4.3Clustering

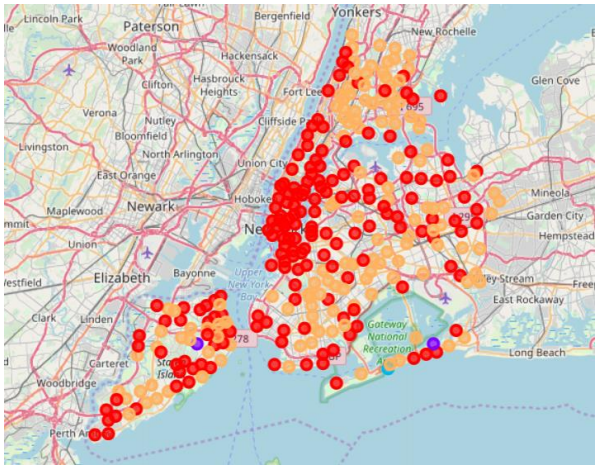
Finally we use the K-means algorithm to compute the cluster and find out how the borough will cluster together.As mentioned before,the k-means algorithm is a kind of algorithm in machine learning to do unsupervised learning and find how boroughs are similar to each other.

We add the clustering label to the dataframe and use that to the application in the folium map.

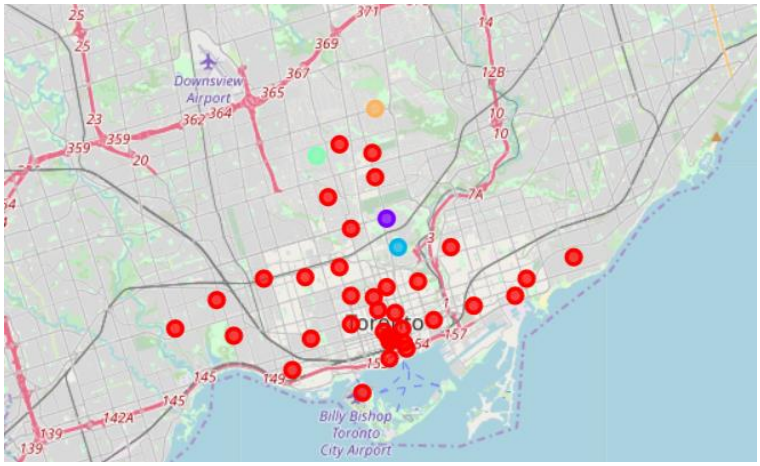
	Borough	Neighborhood	Latitude	Longitude	Clustering Label	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Bronx	Wakefield	40.894705	-73.847201	4	Pharmacy	Laundromat	Donut Shop	Pizza Place	Sandwich Place
1	Bronx	Co-op City	40.874294	-73.829939	4	Baseball Field	Bus Station	Restaurant	Park	Chinese Restaurant
2	Bronx	Eastchester	40.887556	-73.827806	4	Caribbean Restaurant	Diner	Intersection	Fast Food Restaurant	Bakery
3	Bronx	Fieldston	40.895437	-73.905643	4	Bus Station	River	Plaza	Women's Store	Field
4	Bronx	Riverdale	40.890834	-73.912585	0	Park	Baseball Field	Plaza	Medical Supply Store	Bank

We use the label we get from the above action to make a map.

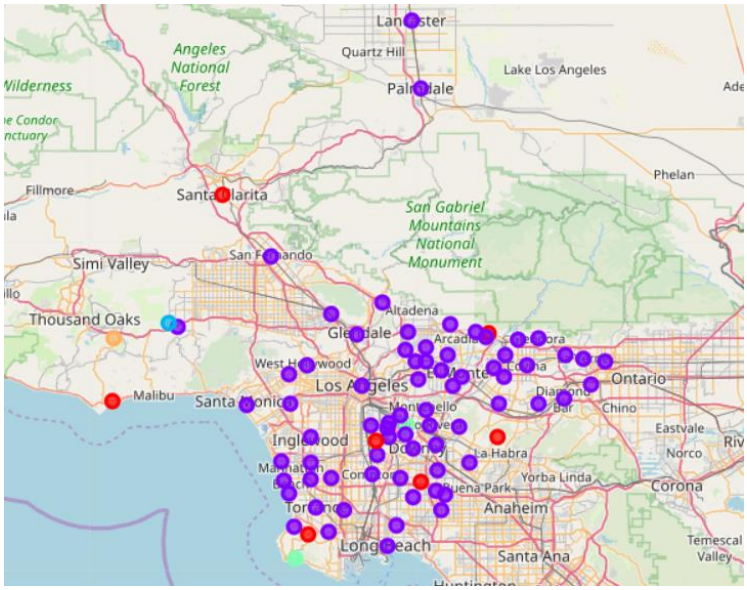
New York city



Toronto



Los Angeles



4.4Examine the clustering

New York city

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
4	Riverdale	Park	Baseball Field	Plaza	Medical Supply Store	Bank
6	Marble Hill	Sandwich Place	Gym	Coffee Shop	Yoga Studio	Diner
10	Baychester	Donut Shop	Supermarket	Gym / Fitness Center	Music Venue	Men's Store
12	City Island	Harbor / Marina	Thrift / Vintage Store	Boat or Ferry	Seafood Restaurant	Pharmacy
16	Fordham	Mobile Phone Shop	Bank	Shoe Store	Supplement Shop	Donut Shop
22	Port Morris	Latin American Restaurant	Furniture / Home Store	Storage Facility	Restaurant	Distillery
24	Hunts Point	Juice Bar	Restaurant	Home Service	Gourmet Shop	Bank

The picture above shows the major clustering in the New York city,as we examine the clustering,we can find that it is the most prosperous cluster in the New York city.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
192	Somerville	Park	Women's Store	Field	Entertainment Service	Ethiopian Restaurant
203	Todt Hill	Park	Women's Store	Field	Entertainment Service	Ethiopian Restaurant

The second clustering is a relaxing clustering and it will provide people with the relaxing and recreational activities after their work.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
179	Neponsit	Beach	Women's Store	Filipino Restaurant	Ethiopian Restaurant	Event Space

The third clustering is a tourism location and it is for the tourist.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Wakefield	Pharmacy	Laundromat	Donut Shop	Pizza Place	Sandwich Place
1	Co-op City	Baseball Field	Bus Station	Restaurant	Park	Chinese Restaurant
2	Eastchester	Caribbean Restaurant	Diner	Intersection	Fast Food Restaurant	Bakery
3	Fieldston	Bus Station	River	Plaza	Women's Store	Field
5	Kingsbridge	Pizza Place	Sandwich Place	Mexican Restaurant	Supermarket	Spanish Restaurant
7	Woodlawn	Pub	Pizza Place	Deli / Bodega	Playground	Convenience Store
8	Norwood	Pizza Place	Park	Bank	Pharmacy	Deli / Bodega
9	Williamsbridge	Caribbean Restaurant	Bar	Soup Place	Nightclub	Entertainment Service

The fifth clustering is full of the human being atmosphere, restaurants and transportation sites give the facility to people's live.

On the condition of the Toronto, we analyze the 5 clustering of the Toronto.

	Borough	Clustering Label	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
37	East Toronto	0	Trail	Pub	Health Food Store	Wings Joint	Cupcake Shop
41	East Toronto	0	Greek Restaurant	Coffee Shop	Italian Restaurant	Ice Cream Shop	Bookstore
42	East Toronto	0	Sandwich Place	Pet Store	Intersection	Steakhouse	Restaurant
43	East Toronto	0	Café	Coffee Shop	Bakery	Gastropub	Brewery
45	Central Toronto	0	Food & Drink Shop	Hotel	Park	Breakfast Spot	Gym
46	Central Toronto	0	Clothing Store	Sporting Goods Shop	Coffee Shop	Restaurant	Mexican Restaurant
47	Central Toronto	0	Sandwich Place	Dessert Shop	Pizza Place	Gym	Café
49	Central Toronto	0	Pub	Coffee Shop	Sushi Restaurant	Bank	Sports Bar
51	Downtown Toronto	0	Coffee Shop	Park	Bakery	Restaurant	Italian Restaurant
52	Downtown Toronto	0	Gastropub	Gay Bar	Coffee Shop	Men's Store	Restaurant
53	Downtown Toronto	0	Coffee Shop	Pub	Park	Mexican Restaurant	Breakfast Spot

The first clustering is more like cluster5 in the New York example.It is of restaurant and café shop but not transportation.Other clustering only have one borough and they are highly

diverse.

	Latitude	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
9	33.88200	Concert Hall	Farmers Market	Dumpling Restaurant	Eastern European Restaurant
11	34.14695	Home Service	Fast Food Restaurant	Flower Shop	Flea Market
38	33.96100	Home Service	Yoshoku Restaurant	Dry Cleaner	Flea Market
49	34.02600	Beach	Harbor / Marina	Yoshoku Restaurant	Dry Cleaner
65	33.78792	Park	Farm	Donut Shop	Flea Market
71	34.39200	Home Service	Construction & Landscaping	Carpet Store	Farm
77	33.95500	Park	Liquor Store	Farm	Eastern European Restaurant

In the example of Los Angeles,we can see from the first clustering that it mainly consist of the relaxing sites

	Latitude	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	34.09500	Cocktail Bar	Chinese Restaurant	Seafood Restaurant	Café
3	33.86607	Chinese Restaurant	Grocery Store	Bubble Tea Shop	Vietnamese Restaurant
4	33.34300	Seafood Restaurant	Mexican Restaurant	Pizza Place	Ice Cream Shop
5	34.13400	Pizza Place	Pharmacy	Big Box Store	Bank
6	34.08501	Ice Cream Shop	Discount Store	Pharmacy	Fast Food Restaurant
7	33.97800	Discount Store	South American Restaurant	Latin American Restaurant	Shoe Store
8	33.96507	Seafood Restaurant	Mexican Restaurant	Fried Chicken Joint	Latin American Restaurant
10	34.07400	Boutique	Park	Salon / Barbershop	American Restaurant
12	34.18100	Japanese Restaurant	Pizza Place	Sandwich Place	Ice Cream Shop

After that in the cluster2 we can find there exist the same clustering that have the restaurant and café shop. It have the similarity of that in Toronto.

	Latitude	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
2	33.98095	Wine Bar	Dry Cleaner	Flower Shop	Flea Market
62	33.74400	Dry Cleaner	Flower Shop	Flea Market	Fish & Chips Shop
63	33.74400	Dry Cleaner	Flower Shop	Flea Market	Fish & Chips Shop

In clustering 4 we have some chores sites.

5. Results and conclusion

We can see from our analysis that the three districts have the similarities and the differences. If we divide the area in the districts of the following parts-food, recreation&relaxing,chores and so on. We find that both New York city and Los Angeles have the clustering that contains the living area. The sites are highly concentrated and New York city and Los Angeles both

have an area of the living cluster.