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.

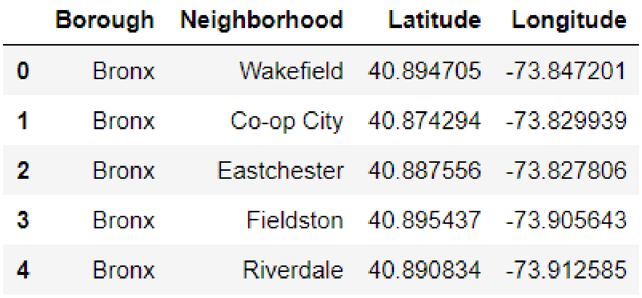
There are additional information about the condition of neighborhood in Toronto. Chinese means the Chinese population in the area, since I am from China and I will surely take this into consideration. Tenant Average Rent means the rent price level in the area, since I may not afford buying the house 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.

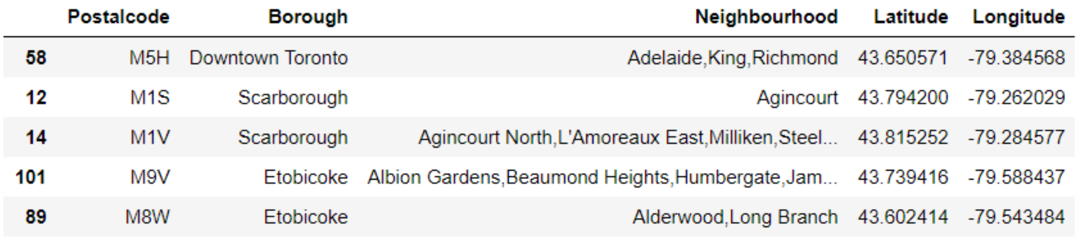
Besides,the data we get from three sources are in different forms,so we will reform the dataframes and make them in the same format,which is the form of Neighborhood-Latitude-Longitude.

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



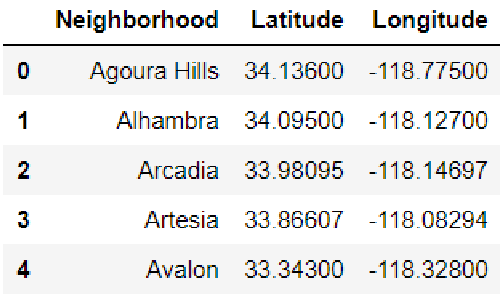
	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

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.