Battle of Neighborhoods

Xue YANG

January 2, 2021

1 Introduction (Business Problem)

1,1 Background

Where to start a business is a key important question to all the business owners. Different cities and areas are suitable for different business. The local economy, the cultures, the races of residents, all these can influence the decision. And there are no unique criteria for all industries or business to find the clue of decision-making. The factors which influence the decision-making vary from business to business.

1.2 Business problem

Our client is a bottled sparkling water supplier, and they offer such water to the local restaurants. They have already a warehouse and delivery center in New York because of the diversified food restaurants and availability. And now they need to decide if Chicago is another good place to have a new center. The key importance for making decision in such business is availability of restaurants venues, and if Chicago is as diversified as New York in restaurant options, since the higher level of diverse, the higher level the acceptance of new brand the city has.

1.3 Target audience of this project

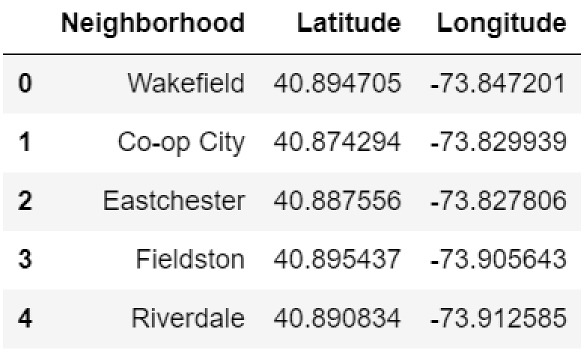
This project may be interested to the people who have the similar business, and they have not had any solution to make the final decision of which city should be the best to them to open a new beverage supply center. This project will also give some ideas to people who need to make the choice among the options.

2 Data acquisition and cleaning

2.1 Data sources and acquisition

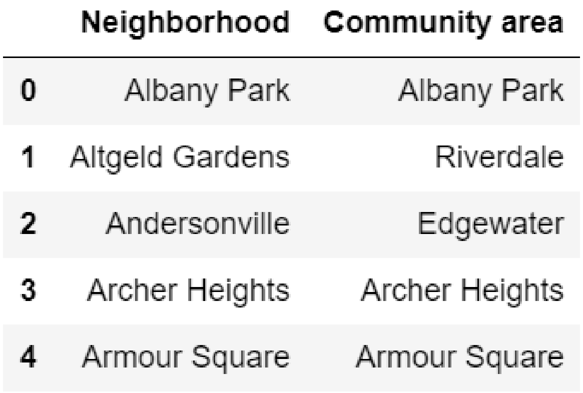
The data for this project includes neighborhoods names and their coordinates, the venues names and their categories of each city.

New York neighborhoods data source
https://cocl.us/new_york_dataset
Chicago neighborhoods data source
List of neighborhoods in Chicago - Wikipedia



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

Figure 1 Neighborhoods and coordinates of New York



	Neighborhood	Community area
0	Albany Park	Albany Park
1	Altgeld Gardens	Riverdale
2	Andersonville	Edgewater
3	Archer Heights	Archer Heights
4	Armour Square	Armour Square

Figure 2 Neighborhoods and coordinates of Chicago

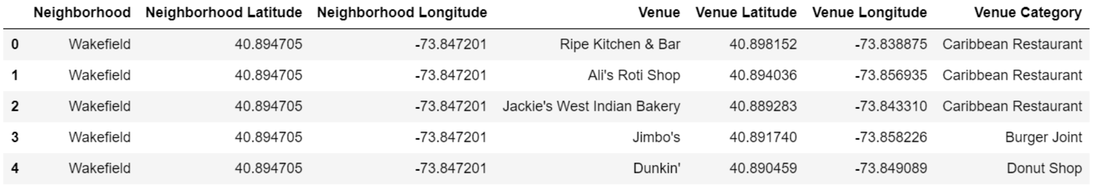
To get the coordinates of the neighborhoods, geopy library of python language will be applied in this project. The coordinates which are the latitude and longitude of each neighborhood will be obtained with the help of geopy based on the neighborhoods names.
Foursquare API can give us the access to the venue data which contains latitude, longitude and category.
This project requires several skills for data acquisition, they are beautifulsoup for web scraping, requests and get for foursquare API, json for processing the downloaded json file.

2.2 Data cleaning

The data of this project are from difference data sources with different format, including json, online table. All the data with different format need to be merged into one DataFrame format of Python.

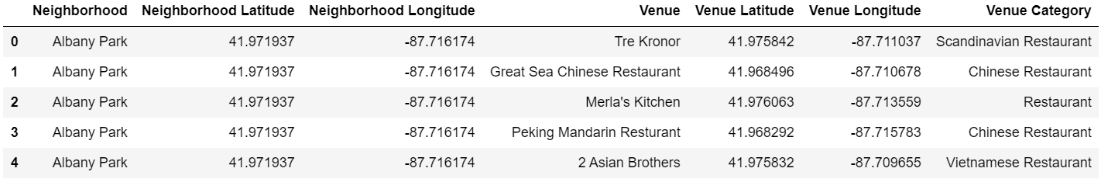
Since the coordinates of neighborhoods are obtained based on the neighborhoods names, so the coordinates are easily linked with neighborhoods. Venue data is extracted from Fousquare, and this dataset also contain the coordinates and the own neighborhood for each venue, so the venues dataset contains venues names, latitude, longitude, neighborhood and category.

These two datasets can be merged by the overlap column, neighborhood name.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Ripe Kitchen & Bar	40.898152	-73.838875	Caribbean Restaurant
1	Wakefield	40.894705	-73.847201	Ali's Roti Shop	40.894036	-73.856935	Caribbean Restaurant
2	Wakefield	40.894705	-73.847201	Jackie's West Indian Bakery	40.889283	-73.843310	Caribbean Restaurant
3	Wakefield	40.894705	-73.847201	Jimbo's	40.891740	-73.858226	Burger Joint
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

Figure 3 Venue dataset of New York



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Albany Park	41.971937	-87.716174	Tre Kronor	41.975842	-87.711037	Scandinavian Restaurant
1	Albany Park	41.971937	-87.716174	Great Sea Chinese Restaurant	41.968496	-87.710678	Chinese Restaurant
2	Albany Park	41.971937	-87.716174	Merla's Kitchen	41.976063	-87.713559	Restaurant
3	Albany Park	41.971937	-87.716174	Peking Mandarin Resturant	41.968292	-87.715783	Chinese Restaurant
4	Albany Park	41.971937	-87.716174	2 Asian Brothers	41.975832	-87.709655	Vietnamese Restaurant

Figure 4 Venue dataset of Chicago

The dataset will be used with K-Means to analyze the venue categories. The breakdown of the venues categories will indicate the presence of each possible value from the original data. The method to process it is one hot encoding. After the venue categories being processed by one hot encoding, the categorical variables (venue categories) will be converted to numerical variables, and that can be handled by K-Means clustering analysis. And then we can get the result from k-means.

Python packages and Dependencies:

- Pandas - Library for Data Analysis
- NumPy – Library to handle data in a vectorized manner
- JSON – Library to handle JSON files
- Geopy – To retrieve Location Data
- Requests – Library to handle http requests
- Matplotlib – Python Plotting Module
- Sklearn – Python machine learning Library
- Folium – Map rendering Library