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

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 1. 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 2. Dataset of Chicago

3 Methodology

3.1 Overall working process

The project requires to analyze the ranking of different venue categories of each cluster which is produced by the unsupervised machine learning method, K-Means clustering.

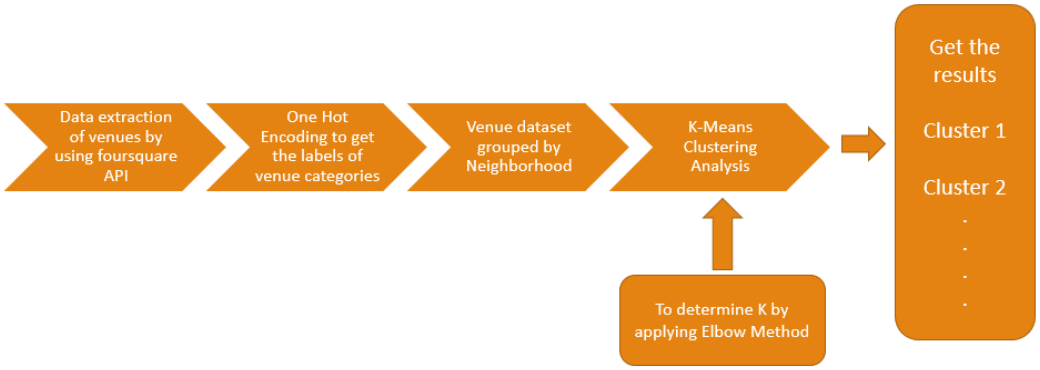


Figure 3. Overall working process

The elbow method will be applied during the analysis to support the system to determine the optimal value of K, which means the number of clusters. Method of K-Means will cluster the venues of these two cities by venue categories, and we will see how diversified the restaurants are in these two cities, which represents the culture diverse. That will help our client to decide if Chicago is a good city to welcome the new sparkling water brand.

3.2 K-Means clustering

Clustering is the process of dividing the entire data into groups based on the patterns in the data. K-Means is one of the most popular “clustering” algorithms. K-Means stores K centroids that it uses to define clusters. A point is considered to be in a particular cluster if it is closer to that cluster’s centroid than any other centroid. The K-Means clustering can be applied in 2-dimensions space, and it is also applied in multi-dimensions’ space as well.

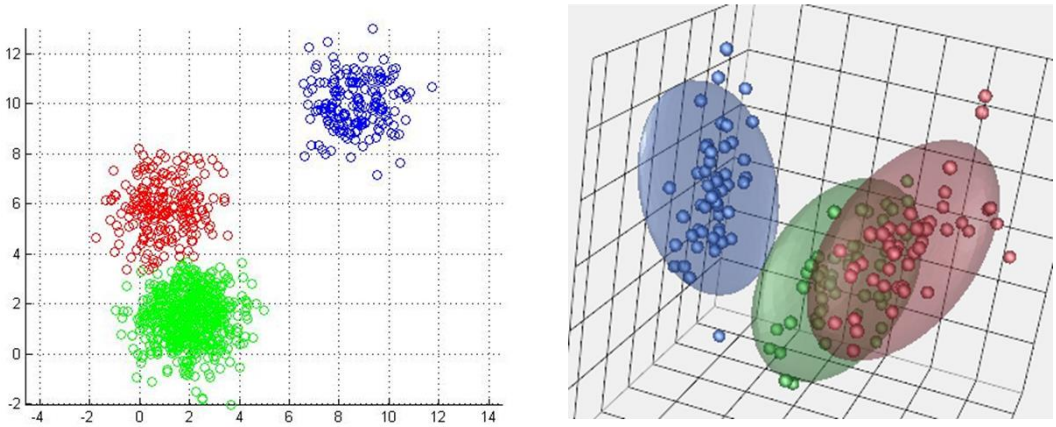


Figure 4 K-Means clustering (source: Stanford University)

In this case, the “venue category” will be used as the feature to be analyzed with K-Means. And we will see how the venues to be clustered based on the result.

3.3 One hot encoding

One hot encoding is a process by which categorical variables are converted into a form that could be provided to machine learning algorithms. The categorical value represents the numerical value of the entry in the dataset. In our project, the venue category is the categorical variables, and the K-Means will be processed with venue category to make the clusters. As the categorical, the venue category cannot be processed by machine learning algorithm directly, in our case is K-Means. Therefore, the categorical values need to be converted. One hot encoding creates new (binary) columns, indicating the presence of each possible value from the original data.

	Venue CategoryAfghan Restaurant	Venue CategoryAfrican Restaurant	Venue CategoryAmerican Restaurant	Venue CategoryArepa Restaurant	Venue CategoryArgentinian Restaurant	Venue CategoryAsian Restaurant	CategoryAus Rest
0	0	0	0	0	0	0	
1	0	0	0	0	0	0	
2	0	0	0	0	0	0	
3	0	0	0	0	0	0	
4	0	0	0	0	0	0	
...	...	...	...	...	...	...	
17741	0	0	0	0	0	0	
17742	0	0	0	0	0	0	
17743	0	0	0	0	0	0	
17744	0	0	0	0	0	0	
17745	0	0	0	0	0	0	

Figure 5. The dataset of New York Venue Category after one hot encoding

3.4 Elbow method

As we know, K-Means is a type of unsupervised learning and one of the popular methods of clustering unlabeled data into K clusters. One of the trickier tasks in K-Means is identifying the appropriate number of clusters K. The idea of the elbow method is to run k-means clustering on the dataset for a range of values of K (in our case, K is from 1 to 10), and for each value of K calculate the sum of the squared errors (SSE). Then, we need to plot a line chart of the SSE for each value of K. If the line chart looks like an arm, then the "elbow" on the arm is the value of K that is the best.

We want a small SSE, but that the SSE tends to decrease toward 0 as we increase K. so our goal is to choose the small value of K that still has a low SSE, and the elbow usually represents where are start to have diminishing returns by increasing K.

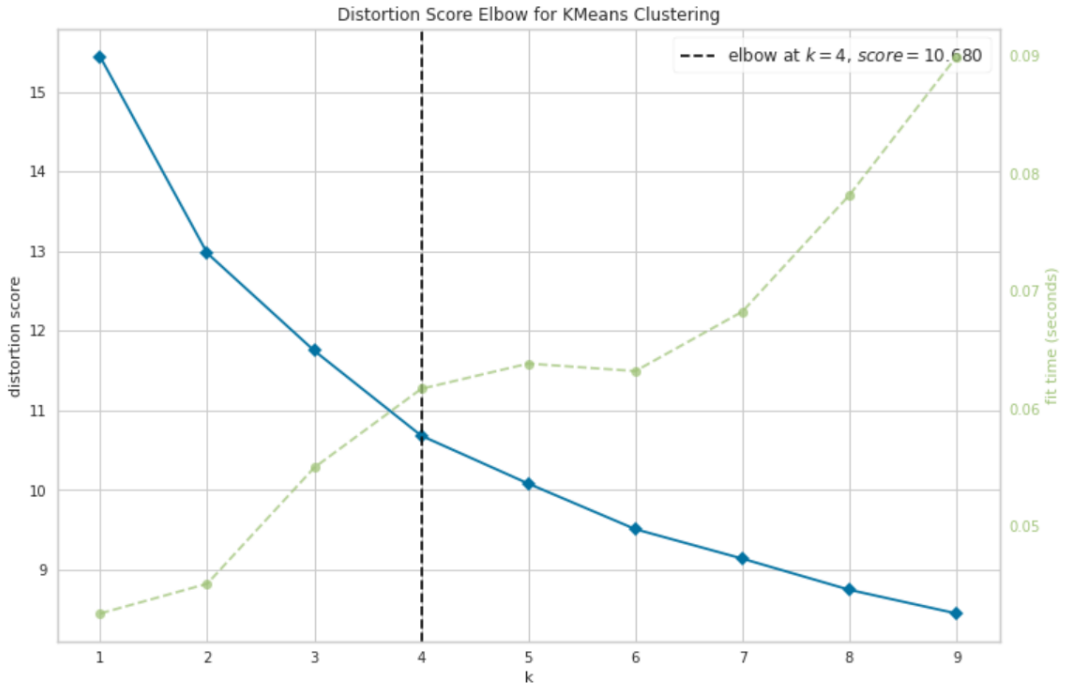


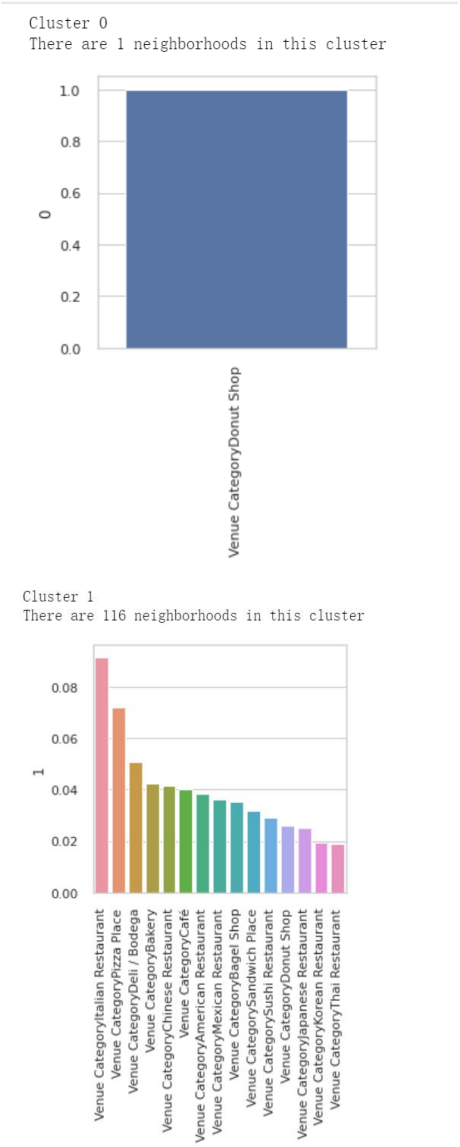
Figure 6. Line chart after elbow applied with K-Means for Chicago

Based on the figure above, the optimal value for K is 4, and we choose 4 clusters to run K-Means for Chicago venue dataset.

4. Result

After running the K-Means with the help of elbow, the K for New York is 3, and for Chicago is 4.

Firstly, here comes the result of New York



Cluster 2
There are 81 neighborhoods in this cluster

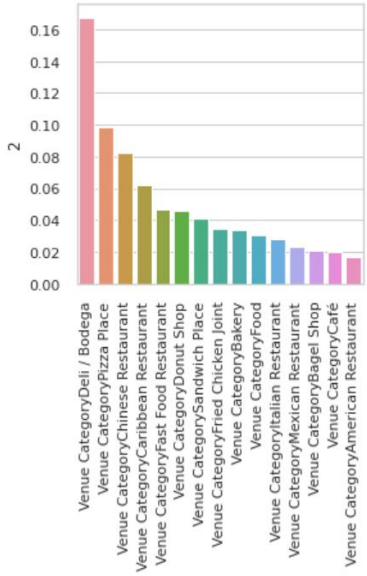
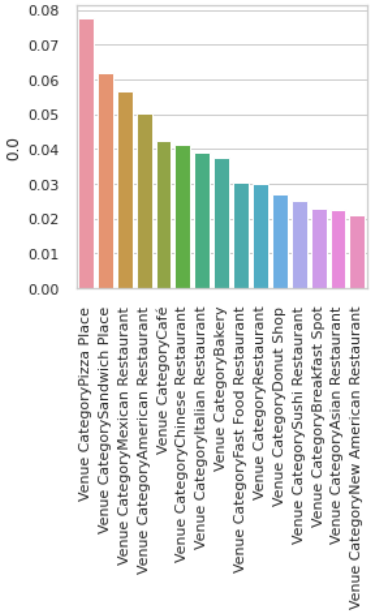


Figure 7. Clustering result of New York

Here comes the result of Chicago



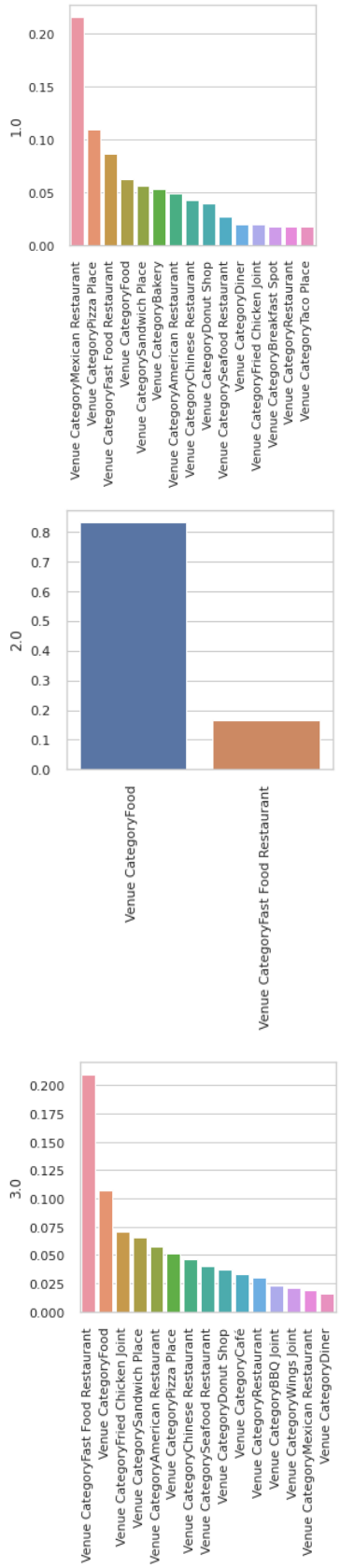


Figure 8. Clustering result of Chicago

The result of K-Means clustering shows that the venues of Chicago is as diversified as

New York. Compared with New York, Chicago has more weight on Mexico food, and less on Asian food. Meanwhile, Chicago has more clusters than New York, which means the city is no less diversified than New York, though both have a cluster without much information.

Below are the map of clusters for New York and Chicago.

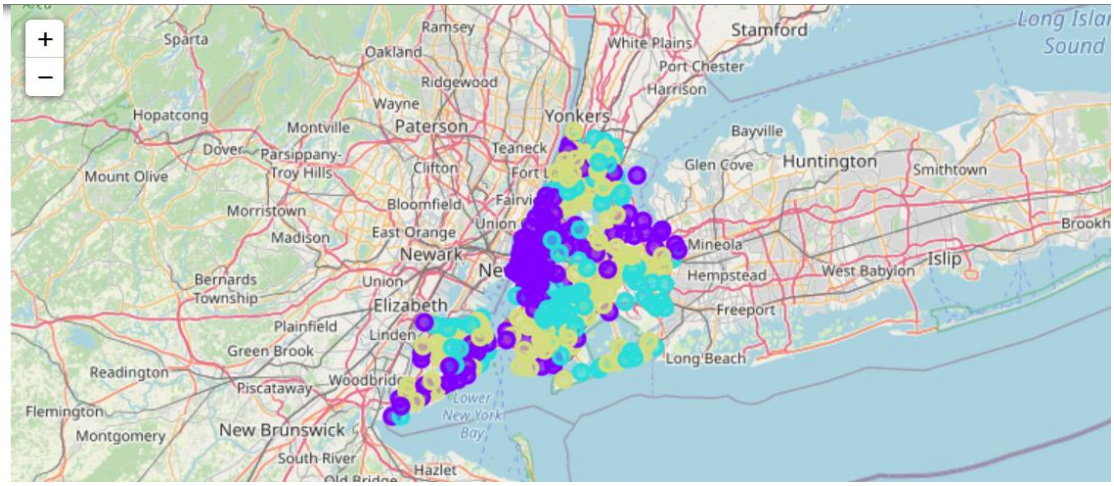


Figure 9. Map of New York clustering

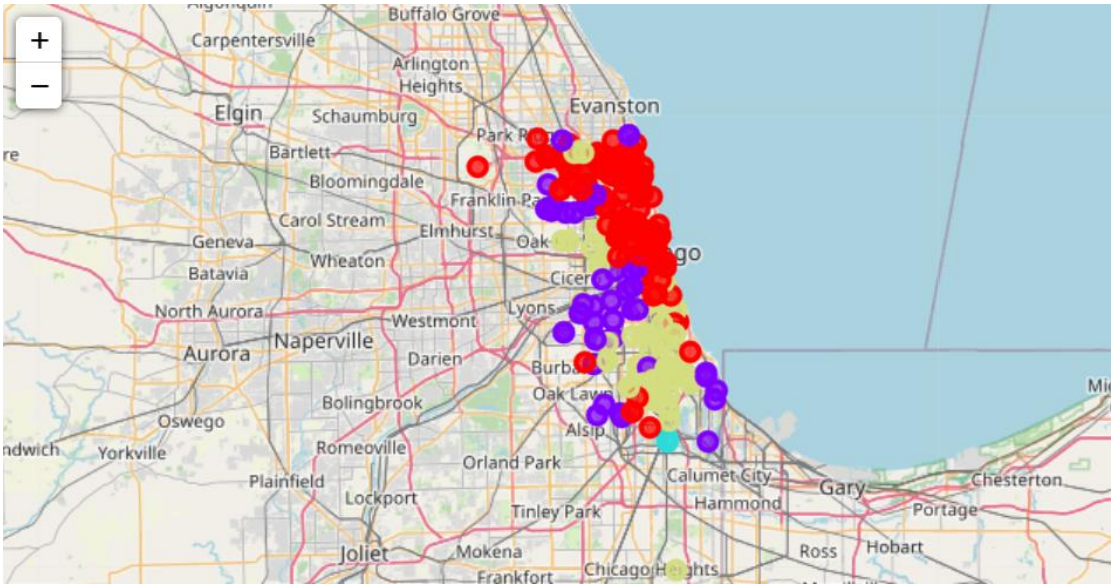


Figure 10. Map of Chicago clustering

5 Conclusion

The goal of this project is to help our client to decide if Chicago is a good place to have another warehouse and delivery center for the products. We collected the data and merged them into one dataset. And then we applied K-Means clustering to analyze the venue category and to see the frequency of the category in each cluster. From the result, we can tell that Chicago is another good city for our client to open a new bottled sparkling water warehouse and delivery center.

6 Discussion

The K-Means analysis for these two cities' venue categories tells the initial result of how diversified the restaurants are in both cities. And it gives some advice and ideas to business owners where to start the business. However, we can add more features into analysis process, for example, the popular brands of sparkling water in Chicago, and the behavior of water consumption among the top venue categories. Then we can get a more detailed analysis and will get a better prediction.