Battle of Neighborhoods

BY XUE YANG

Content

- Python Libraries List
- Introduction/Business Problem
- Data acquisition and cleaning
- Methodologies
- Result
- Conclusion

Python Libraries List

pandas	Data analysis
numpy	Processing data in a vectorized manner
matplotlib	Plotting module
JSON	Processing json file
geopy	To get the location data for cities
requests	Handle the http request
sklearn	Machine learning lib
folium	Produce map

Introduction / Business Problem

Where to start a business is a key question to all the business owners. 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 NYC, 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 NYV in restaurant options, since the higher level of diverse, the higher level the acceptance of new brand the city has.

Data Acquisition and Cleaning

Data Acquisition

- **Neighborhoods and coordinates**

NYC neighborhoods: https://cocl.us/new_york_dataset

Chicago: List of neighborhoods in Chicago - Wikipedia

Coordinates: by applying library Geopy's functions to get the coordinates, including the location, latitudes and longitudes of each neighborhood

- **Venue**

We then explore each neighborhood with the coordinates acquired using the Foursquare API and the explore endpoint, a free call request. We will get venue id, coordinates, and categories of food. We define radius of 1 km for each neighborhood as the range, and we get a max of 100 of the most popular venues at the time the notebook is run for each neighborhood.

Data Acquisition and Cleaning

- ✓ Neighborhoods and Venues of New York City

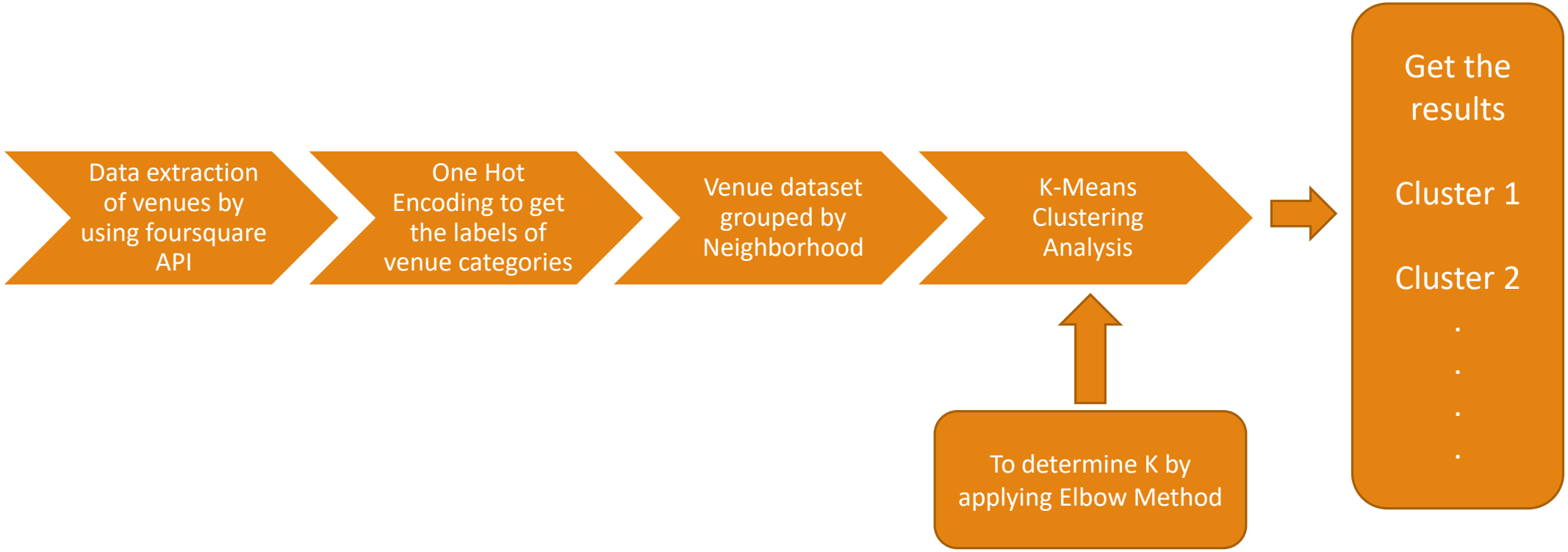
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Ripe Kitchen & Bar	40.898152	-73.838875	Caribbean Restaurant
1	Wakefield	40.894705	-73.847201	Ali's Roti Shop	40.894036	-73.856935	Caribbean Restaurant
2	Wakefield	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

- ✓ Neighborhoods and Venues of Chicago

	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

Methodologies

From data extraction to clustering analysis



Methodologies

- One hot encoding

This method is applied to integer representation of features containing the categorical data, since the machine learning cannot understand the logical relationship between such data. It may response a result with underperformance. One hot encoding can produce the ‘dummy variables’ which takes only the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcomes.

	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	
...	...	...	...	...	...	...	
17748	0	0	0	0	0	0	
17749	0	0	0	0	0	0	
17750	0	0	0	0	0	0	

In our project, there is breakdown of ‘venue category’, for example African Restaurant, is a kind of food in venue category, 0 represents the absence.

Methodologies

- K-Means Clustering

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled outcomes.

In our project as below, K represents the number of clusters to create.

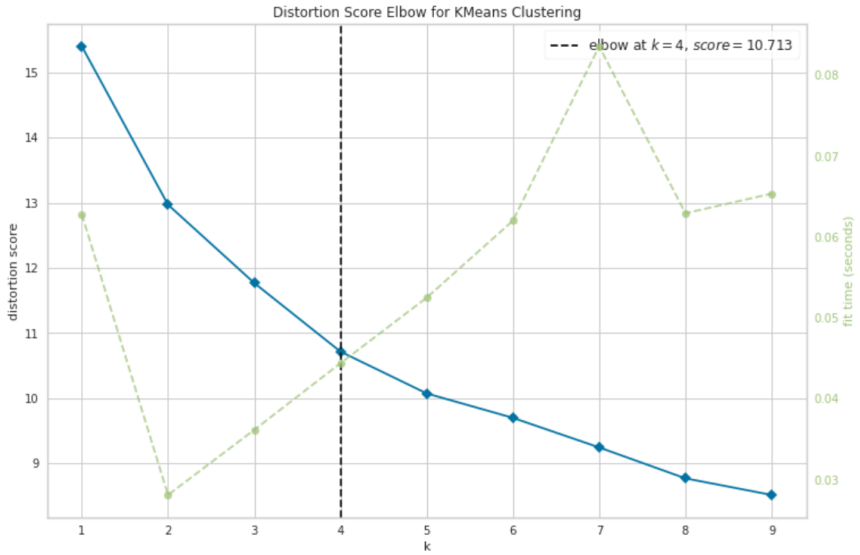
In [62]:
```
# clustering analysis
k_ny = 4
ny_kmeans = KMeans(n_clusters=k_ny, random_state=1).fit(ny_grouped_cluster)
ny_kmeans.labels_[0:10]
```

Out[62]: array([0, 1, 1, 0, 1, 0, 2, 1, 2, 1], dtype=int32)

Methodologies

- Elbow method

Determining K is a problem while using K-Means clustering. How to find the optimal value of K is the key point the K-Means application. Elbow method is one way as the solution.

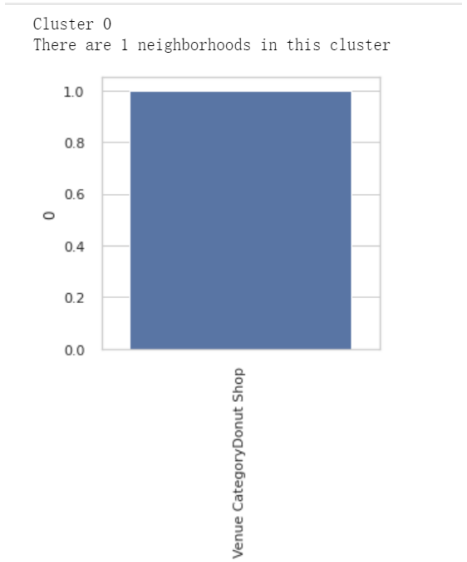


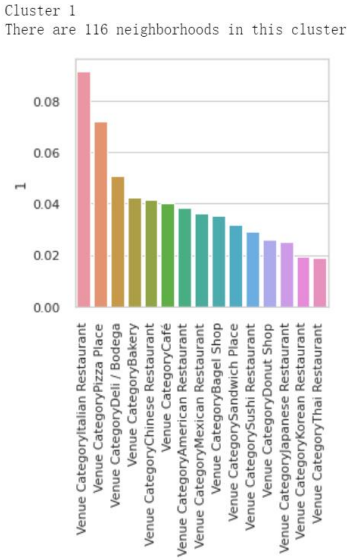
In our project, the elbow method is used to find the optimal value of K, we define the K is from 0 to 10, and algorithm of elbow responses the result with 4. It means that the optimal value of K is 4, and we have 4 clusters for Chicago when we analysis its character of venue categories.

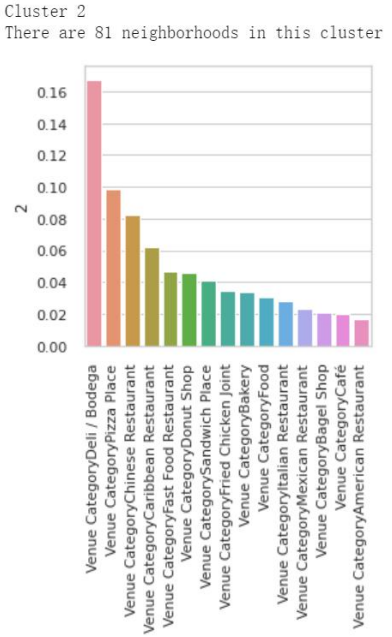
Result

Let’s examine the result of each cluster, to check the frequency of each venue category for each neighborhood.

New York City

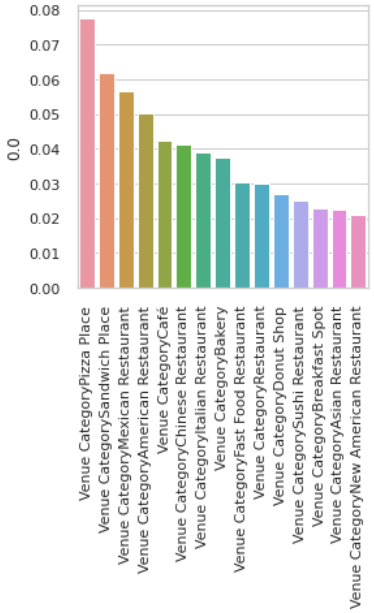


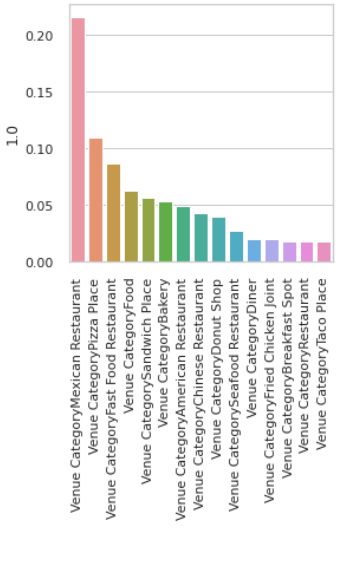


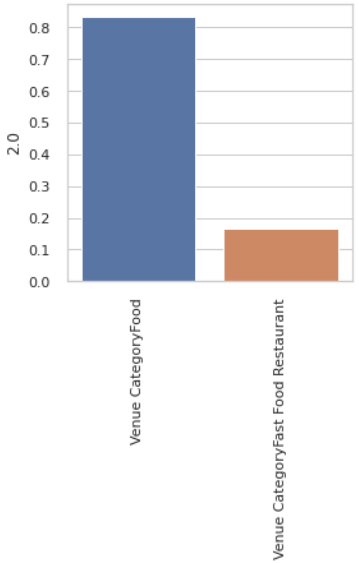


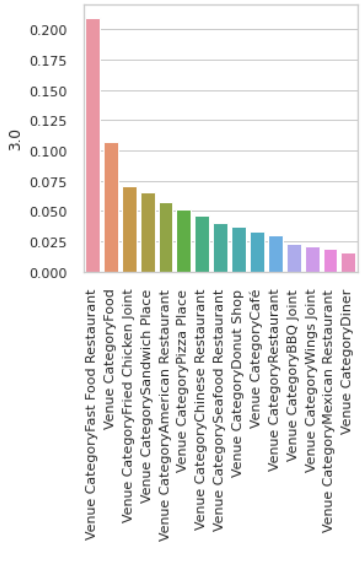
Result

Chicago









Result

In our project, New York has three clusters and Chicago has 4.

New York

Cluster 0 is not informative, there is only one neighborhood.

Cluster 1 has 116 neighborhoods, and top venues categories contain Italian, Chinese, Japanese, Korean, Thai and Mexico restaurants. Italian restaurants dominate.

Cluster 2 has 81 neighborhoods, and top venues categories contain Chinese, Italian, fast food and Caribbean food. Chinese and Italian restaurants have leading position.

Result

Chicago

Cluster 0 has 119 neighborhoods , and top venues categories contain Italian, Mexico, Chinese and other Asian food. Italian and Mexico restaurant are leading position.

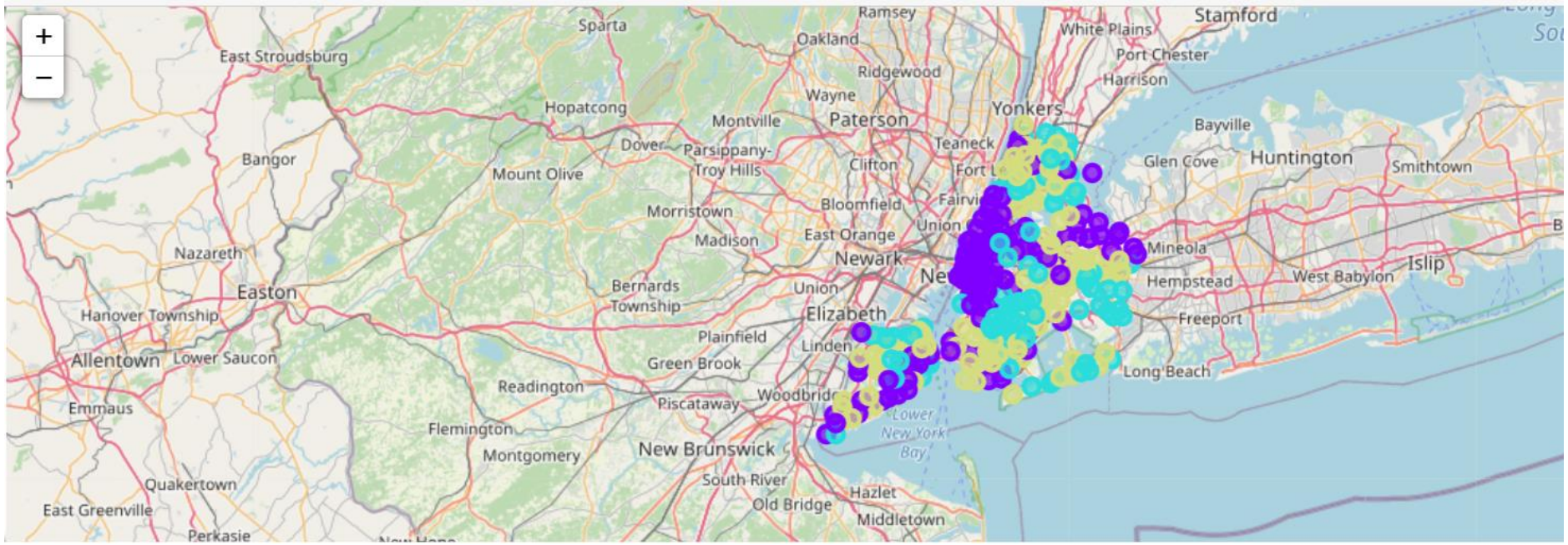
Cluster 1 has 46 neighborhoods, and top venues categories contain Mexico, Italian, Chinese, fast food and seafood restaurants. Mexico restaurants dominate.

Cluster 2 is not informative, as there are only two neighborhoods.

Cluster 3 has 62 neighborhoods, and top venues categories contain fast food, Italian, Chinese and seafood restaurants. Fast food restaurants dominate.

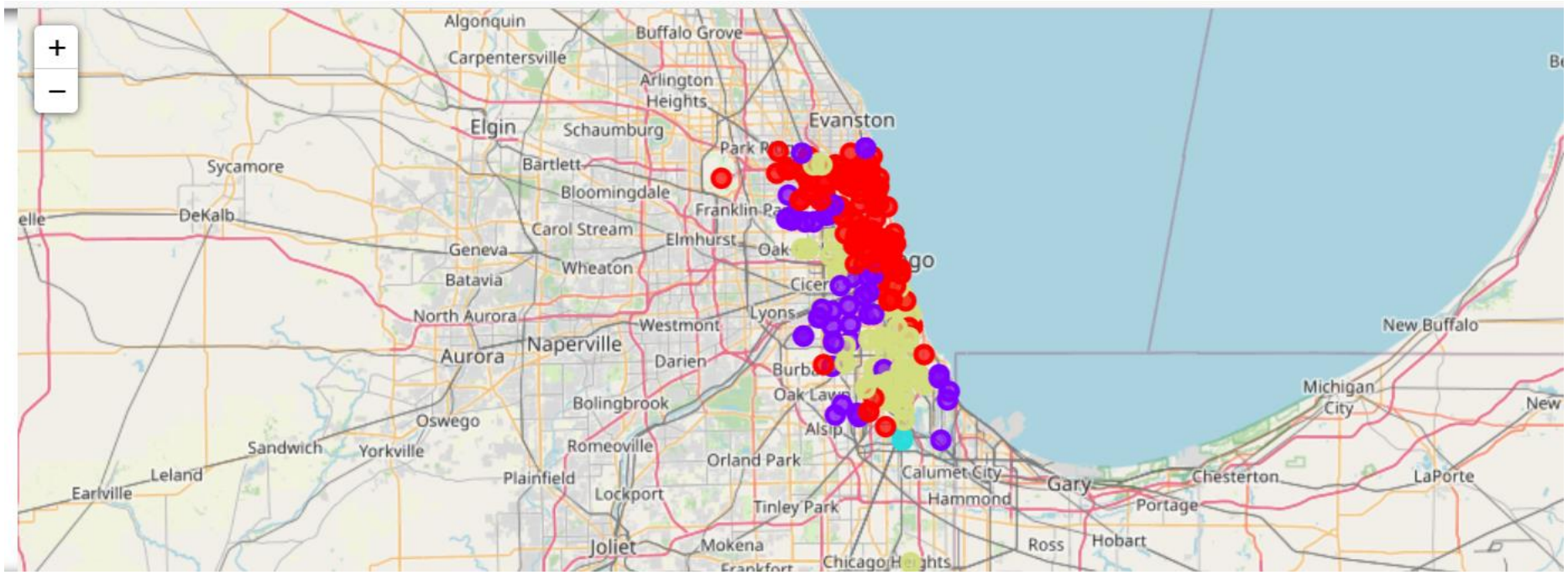
Result

New York clustering map



Result

Chicago clustering map



Conclusion

The business problem is to check if Chicago has the same diversified food venues as New York. And it even has more clusters than New York. This factor will determine if our client, a sparkling water supplier, is going to set up a new warehouse and delivery center in Chicago, since they have already one in New York.

Our project applies K-Means clustering to analyze the neighborhoods and venues of New York and Chicago, and then the result tells that Chicago is also a very active city with diversified food restaurants from different cultures. Setting up a new center in Chicago is good idea.

Thank You !