The Battle of Neighborhoods

Guanming Shen

August 6th, 2019

1. Introduction

1.1 Background

Since I'm from Shanghai, I just want to use this capstone project to show you guys some aspects about this amazing city in some possible ways. I hope that finding out some similarities among NYC, Toronto and Shanghai may attract you guys to know more about Shanghai so that some of you who have already shown interest in moving to this very metropolis could make decision after reading my project and the report.

1.2 Interest

Obviously, my biggest interest would be finding out all the similarities within these three big cities.

2. Data acquisition and cleaning

2.1 Data sources

The neighborhood data of New York:
https://geo.nyu.edu/catalog/nyu_2451_34572
The neighborhoods of Toronto:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
The neighborhoods of Shanghai:
https://en.wikipedia.org/wiki/List_of_township- level_divisions_of_Shanghai

2.2 Data cleaning

I carried out the process of data cleaning including removing some wrong location data, dropping some NaN values and removing some outliers.

3. Methodology

3.1 Data acquisition

The Foursquare API was used to search for nearby venues of each neighborhood in radius of 1000 meters. Only the venues name and venues category (i.e. café, restaurant, school, etc.) are extracted. After obtaining all the venues, the total number of venues in each category is counted for each neighborhood.

3.3 Data Analysis

I applied hierarchical agglomerative clustering method to compare neighborhoods among cities since unlike k-means clustering or other machine learning clustering methods, hierarchical agglomerative clustering doesn't require the number of clusters at the beginning. What's more, it also tells us whether the dataset would support clustering at first glance. I carried out a hierarchical clustering dendrogram with SciPy library. Let's take a look at Fig.1, it's easy to just to separate the neighborhoods into 2 or 3 clusters. However, 2 or 3 clusters will just separate the neighborhoods into city center and suburb area. To have a fuller understanding from the dataset I chose to separate the neighborhoods into 9 clusters by cutting at the distance of 31 (horizontal black line is the cutoff at distance 31 to separate the neighborhoods into 9 clusters.). After that, I applied hierarchical agglomerative clustering to cluster neighborhoods.

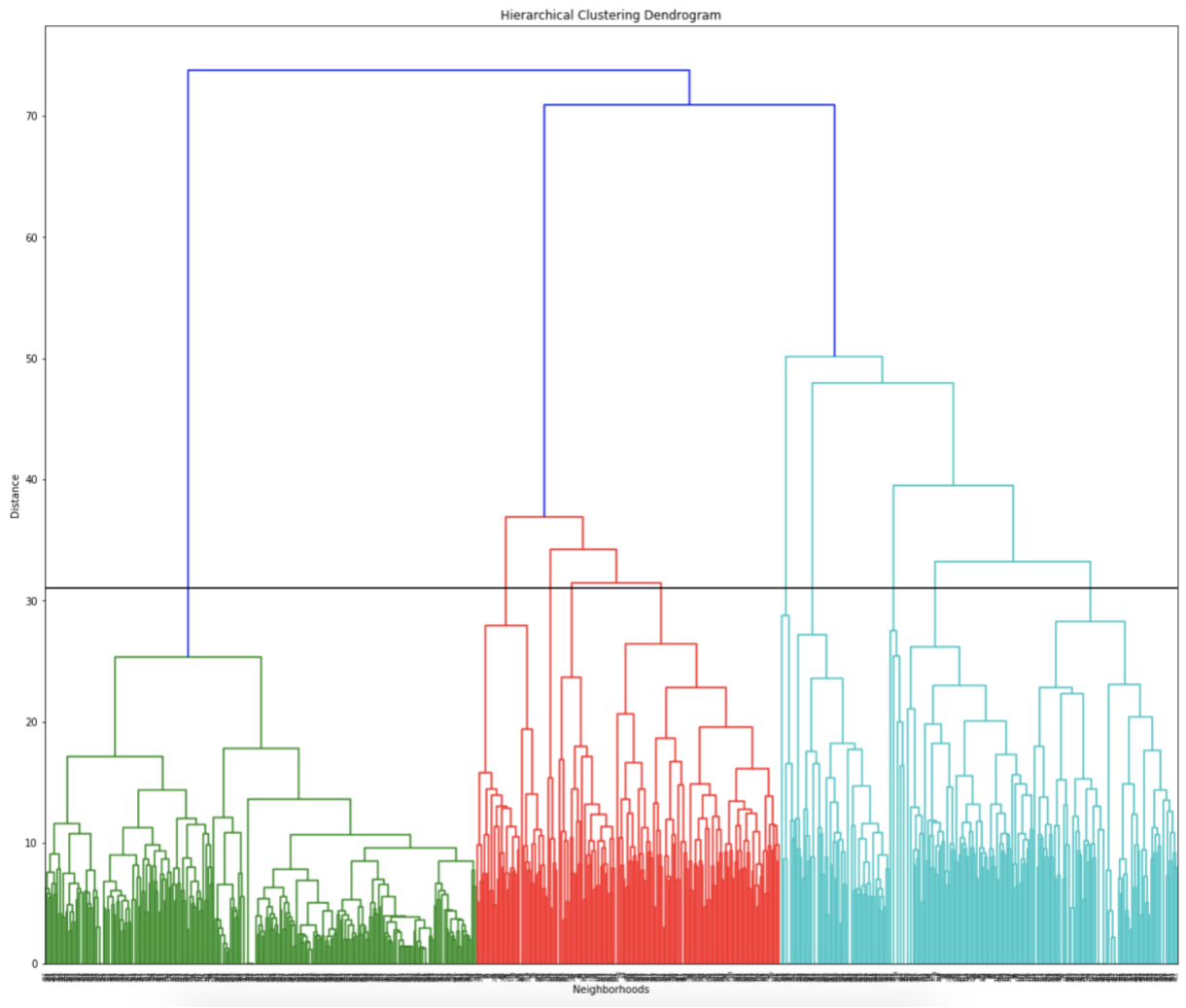


Fig.1 Hierarchical clustering dendrogram

4. Results

4.1 Total number of neighborhoods in each cluster

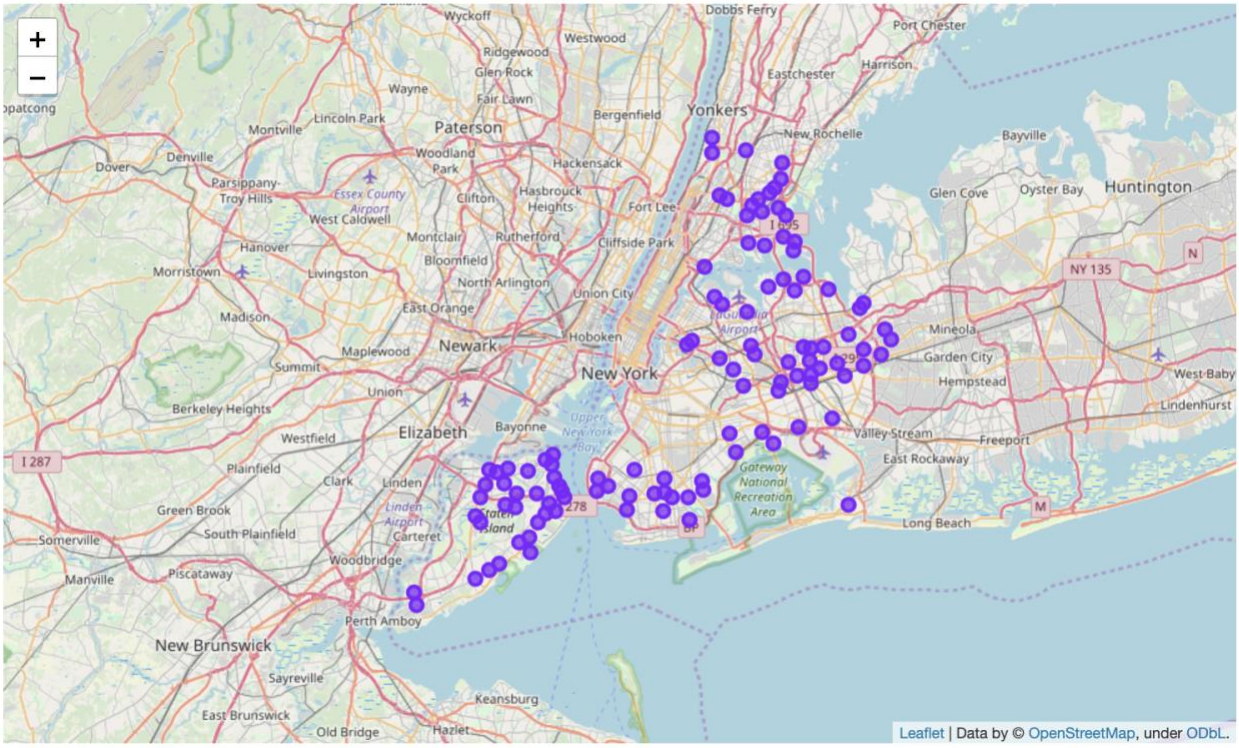
Let’s look at the total number of neighborhoods in each cluster to make sure they are properly clustered.

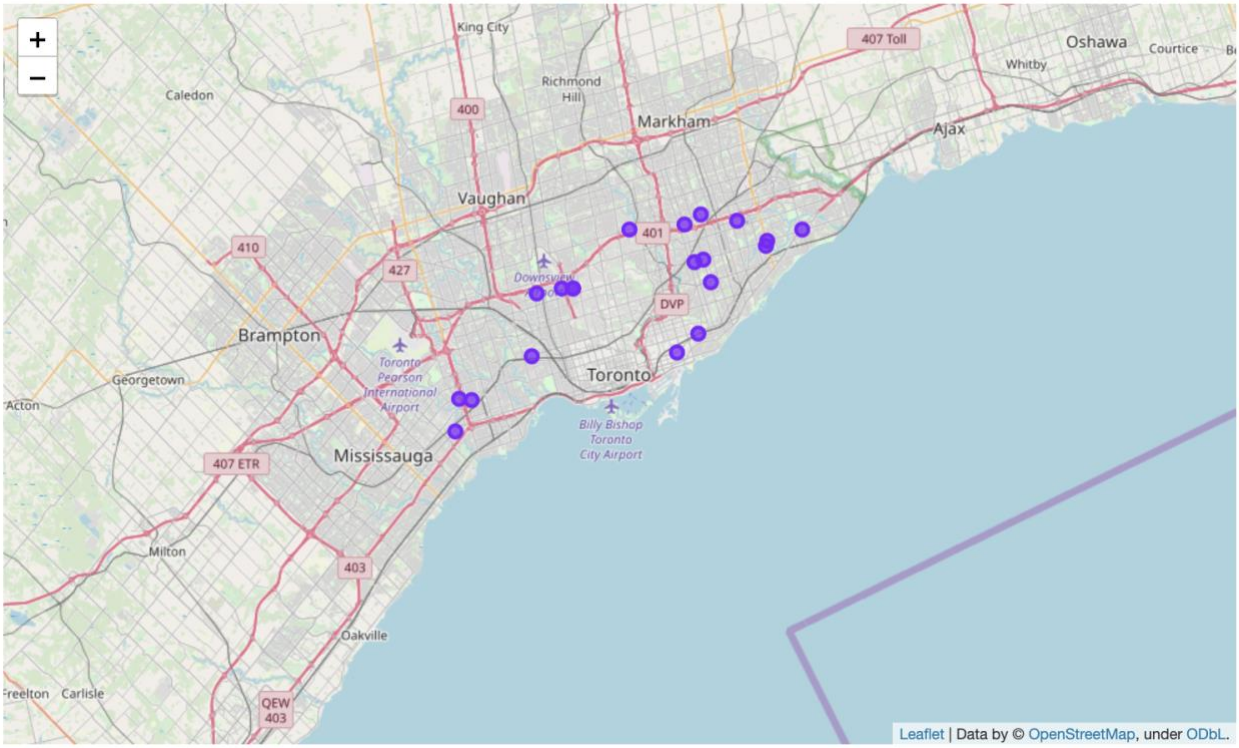
```
Total number of neighborhoods in cluster 0 is 131
Total number of neighborhoods in cluster 1 is 9
Total number of neighborhoods in cluster 2 is 85
Total number of neighborhoods in cluster 3 is 75
Total number of neighborhoods in cluster 4 is 55
Total number of neighborhoods in cluster 5 is 10
Total number of neighborhoods in cluster 6 is 42
Total number of neighborhoods in cluster 7 is 6
Total number of neighborhoods in cluster 8 is 254
```

4.2 Details of cluster 0 to 8

Cluster 0

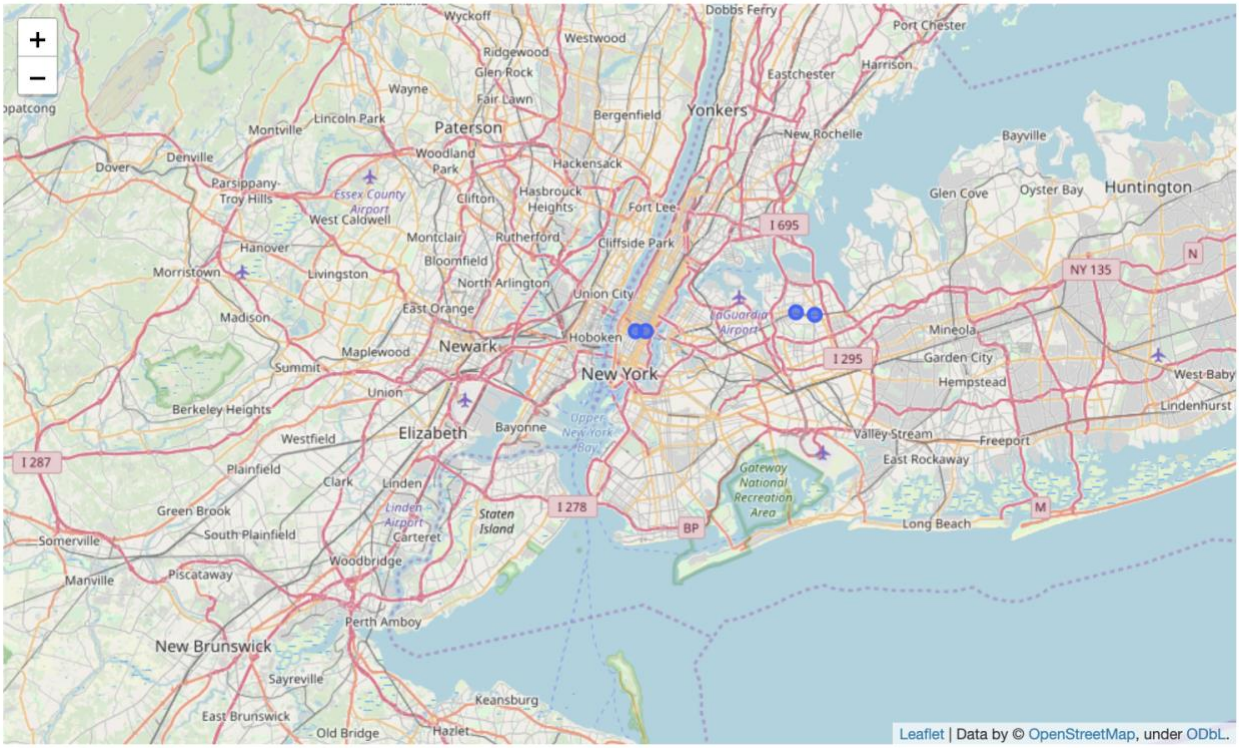
Cluster 0 contains a lot of restaurants and bars. Mostly Italian restaurants, coffee shop and bars. Basically the downtown areas where people grab quality food are covered.

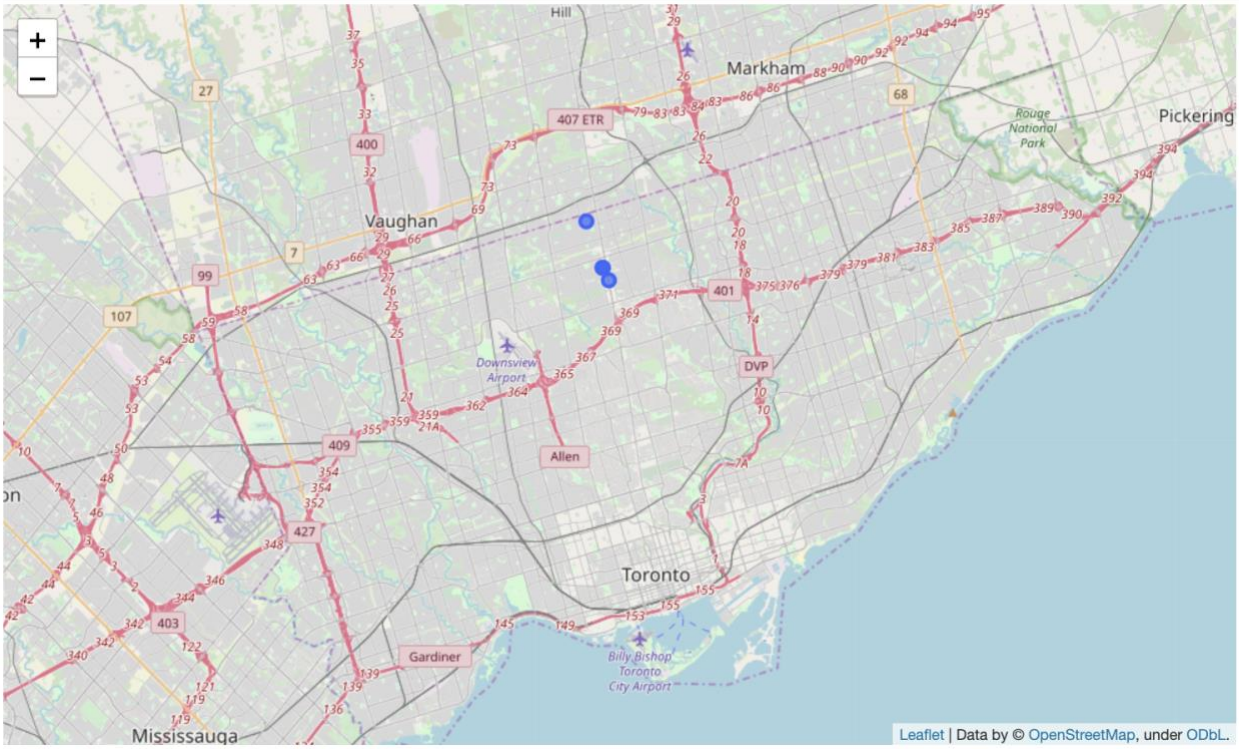



King City
Caledon
Richmond Hill
400
Markham
407 Toll
Oshawa
Whitby
Courtice
Ajax
Vaughan
401
410
427
Downsview
DVP
Brampton
Toronto Pearson International Airport
Georgetown
Toronto
Billy Bishop Toronto City Airport
Mississauga
407 ETR
Milton
403
Oakville
Carlisle
QEW
403
Leaflet | Data by © OpenStreetMap, under ODbL.

Cluster 1

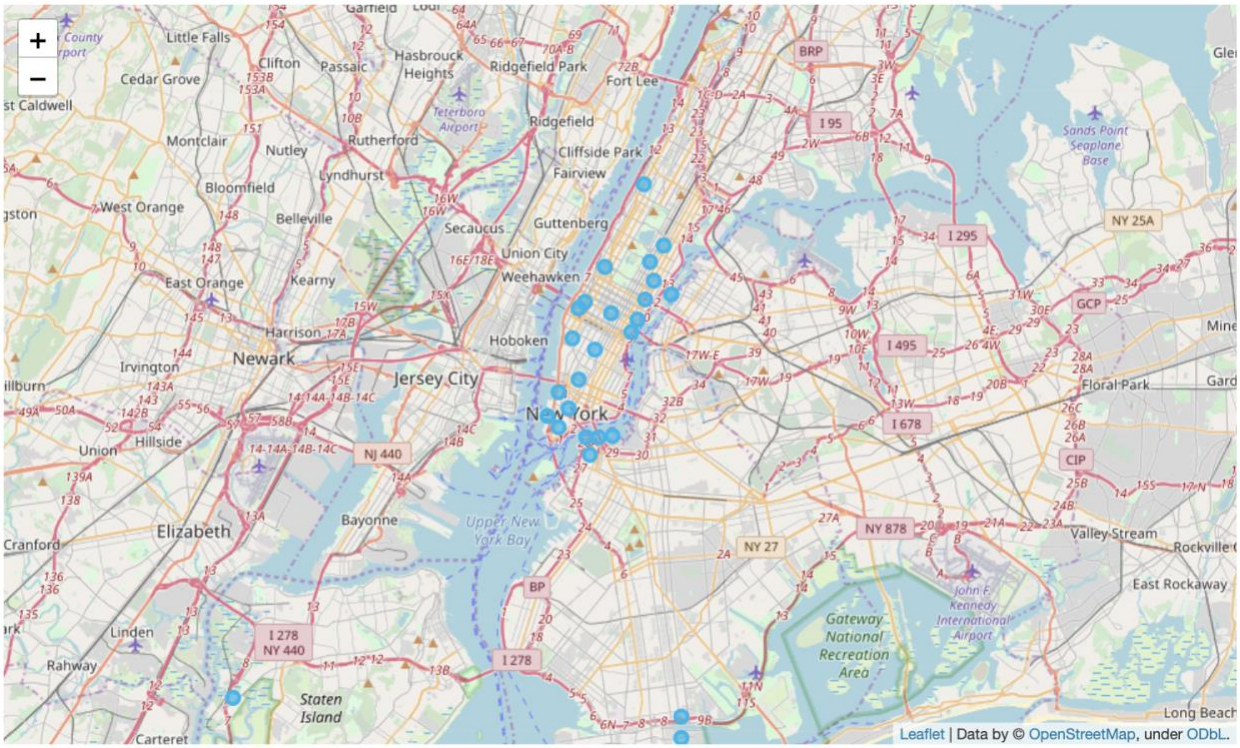
Cluster 1 contains almost all the Korean restaurants both in Toronto and NYC!

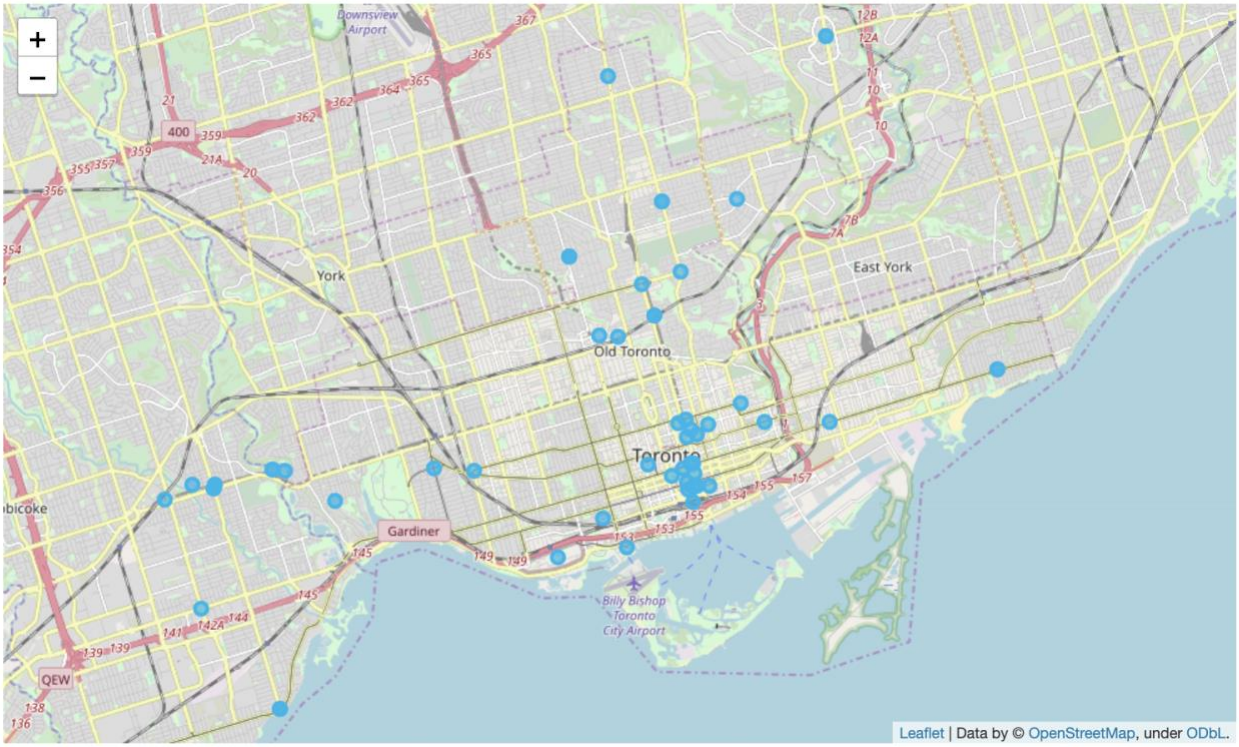




Cluster 2

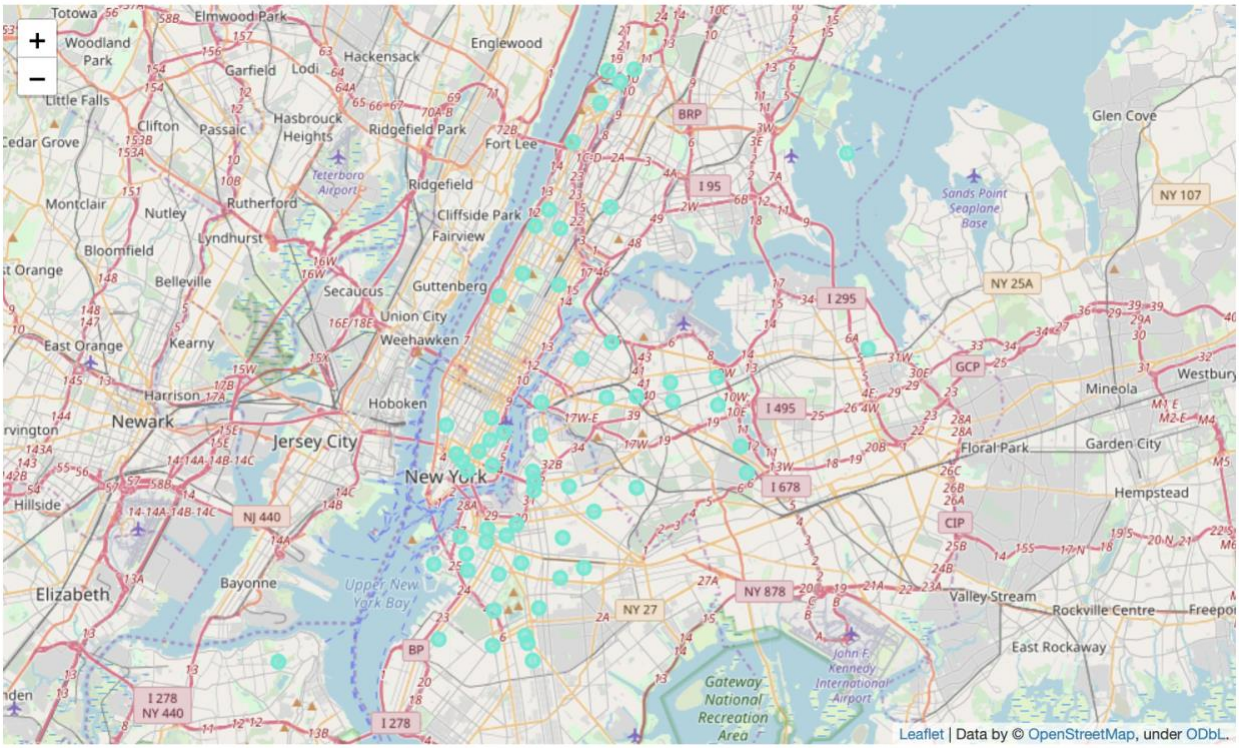
Cluster 2 contains mostly bars, coffee shops, pizza shops and almost all kinds of popular restaurants from different countries such as Japanese cuisines.

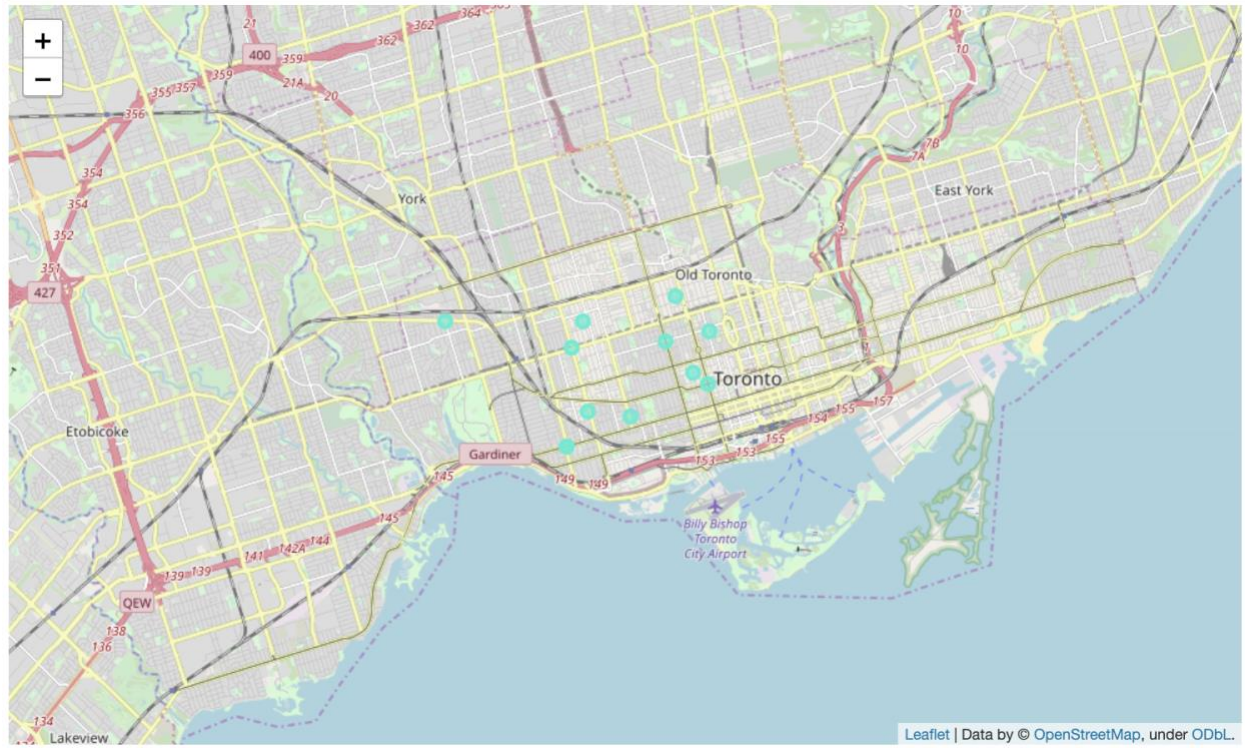




Cluster 3

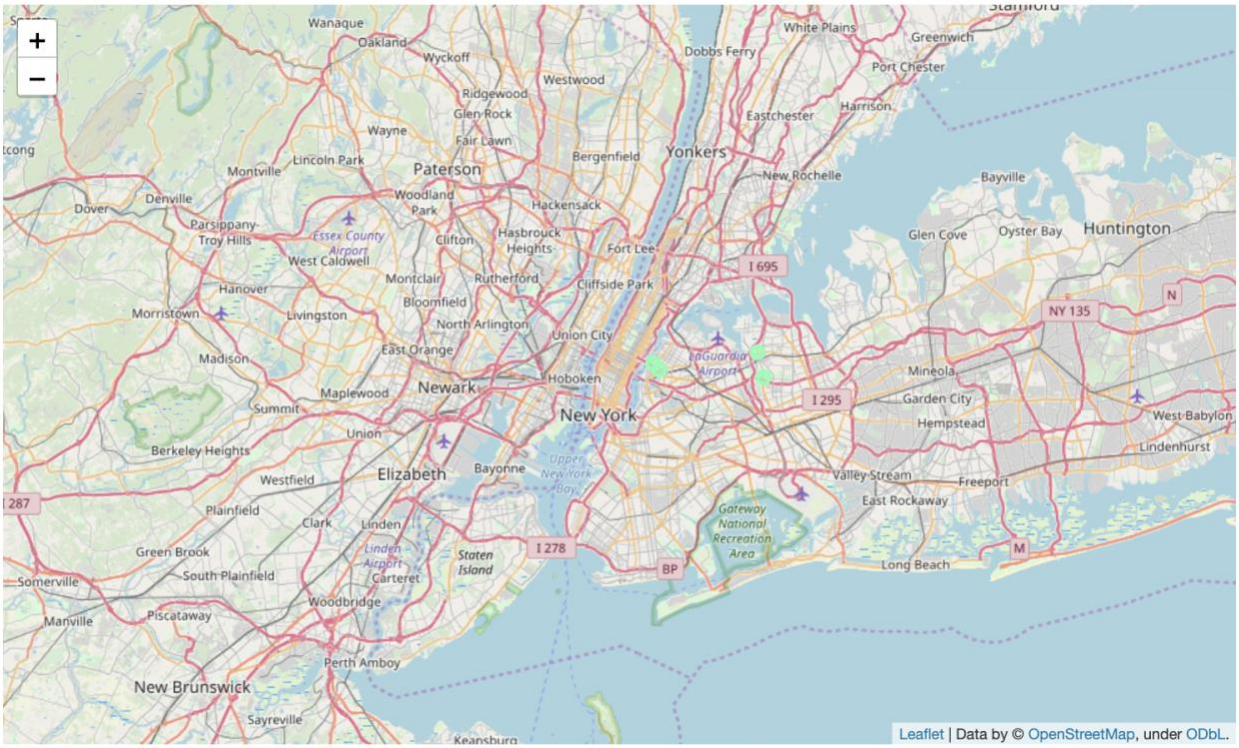
Cluster 3 contains mostly fast food restaurants in NYC and Toronto. Other than that, some other rare exotic cuisines like Caribbean and Indian cuisines are also among the list as well.

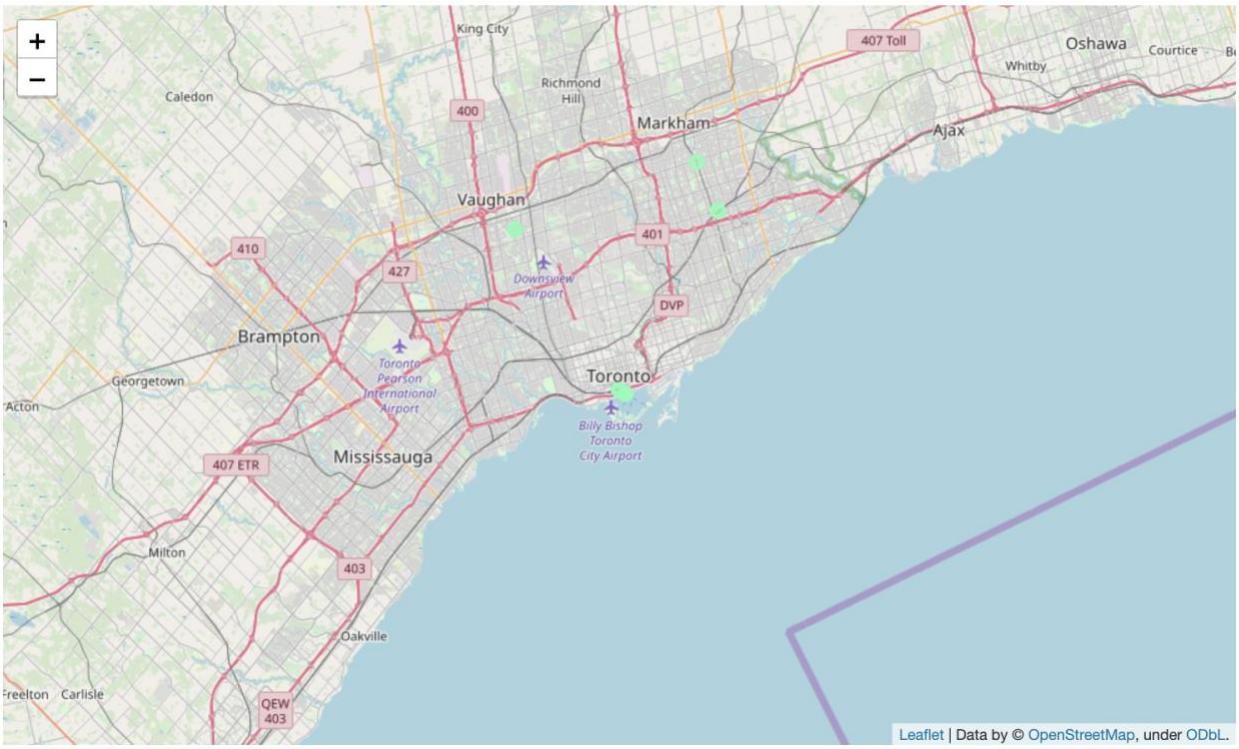




Cluster 4

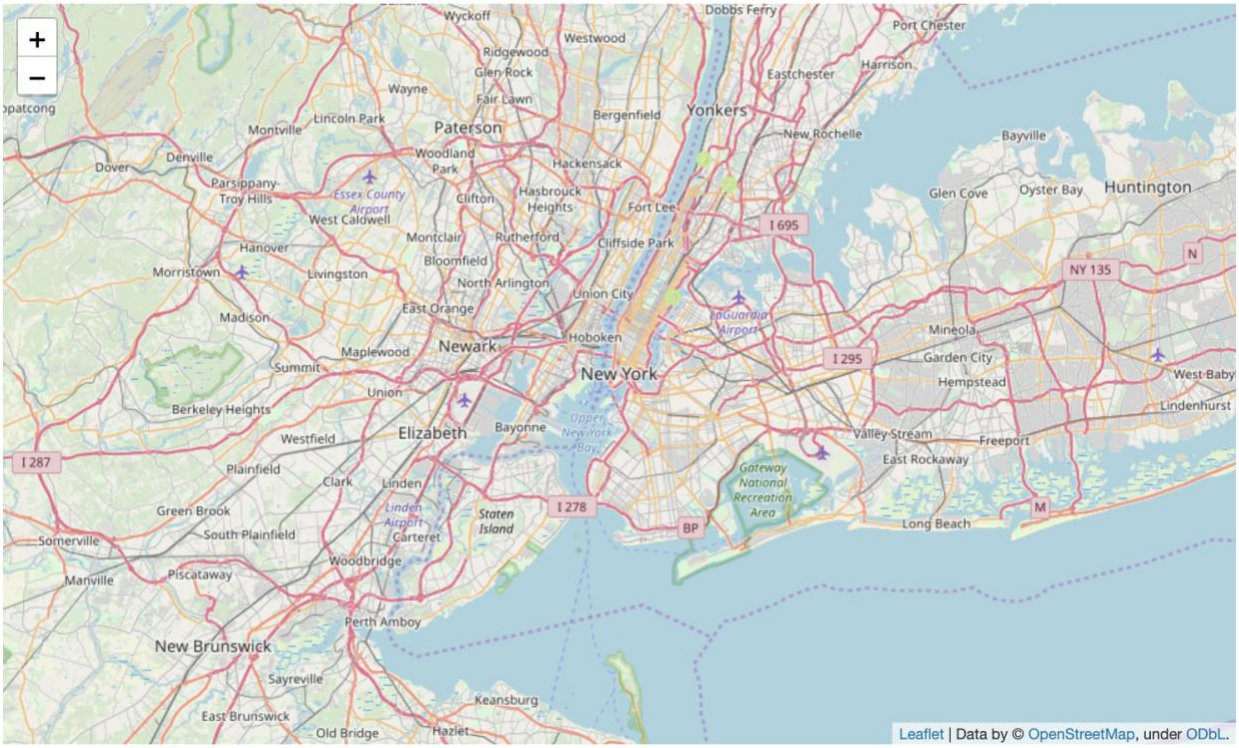
Cluster 4 contains almost all of the quality bars in NYC and Toronto, in which areas like Flushing are included as well.

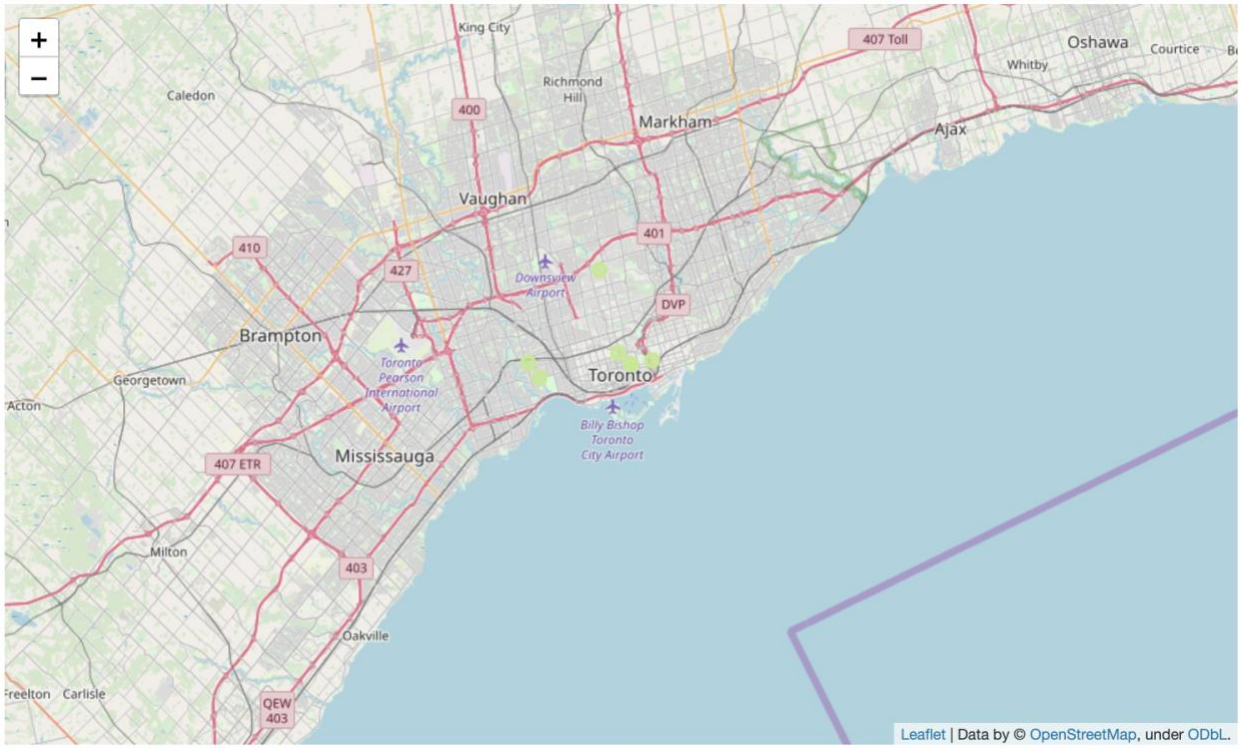




Cluster 5

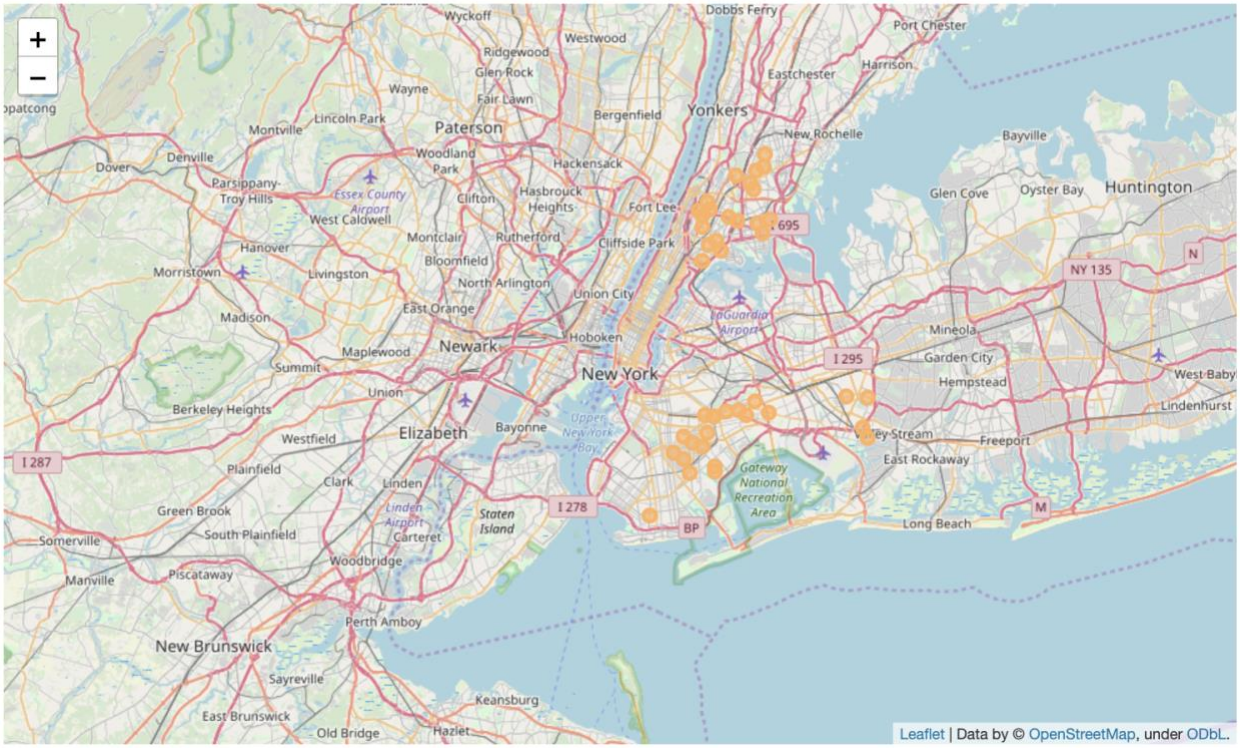
Cluster 5 contains mostly Chinese restaurants in NYC, Toronto and of course, Shanghai.

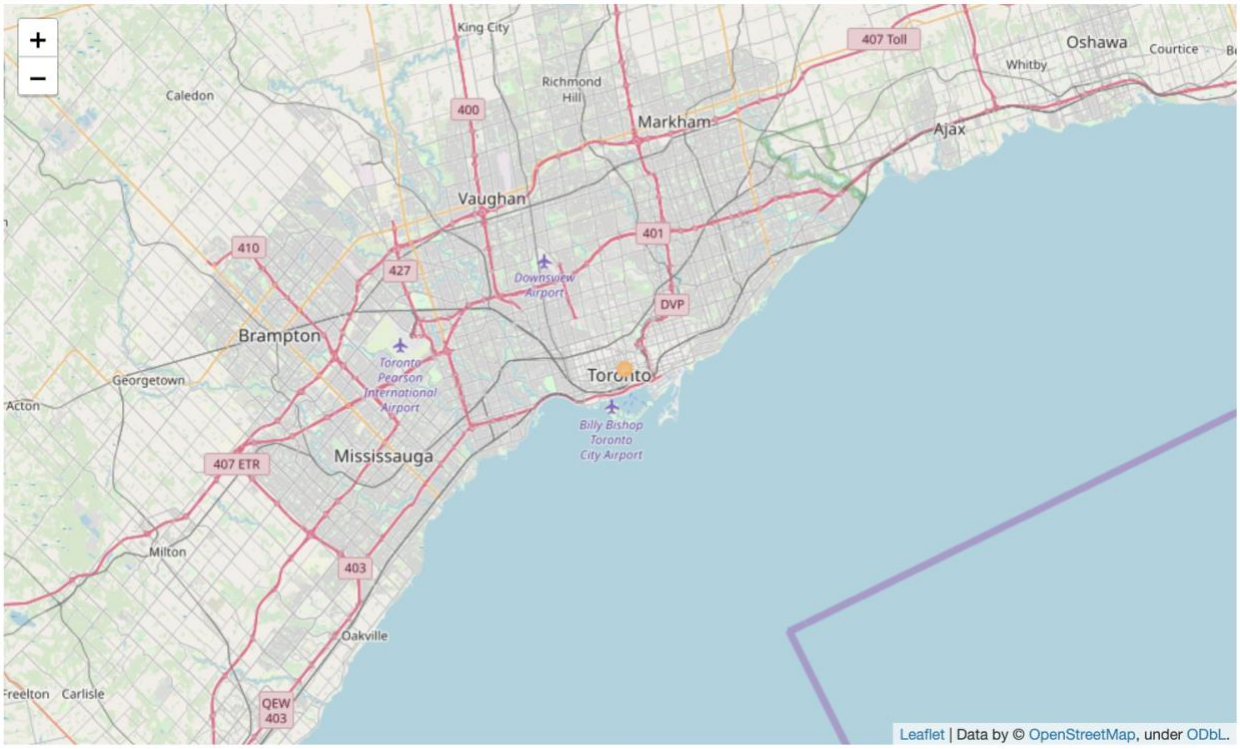




Cluster 6

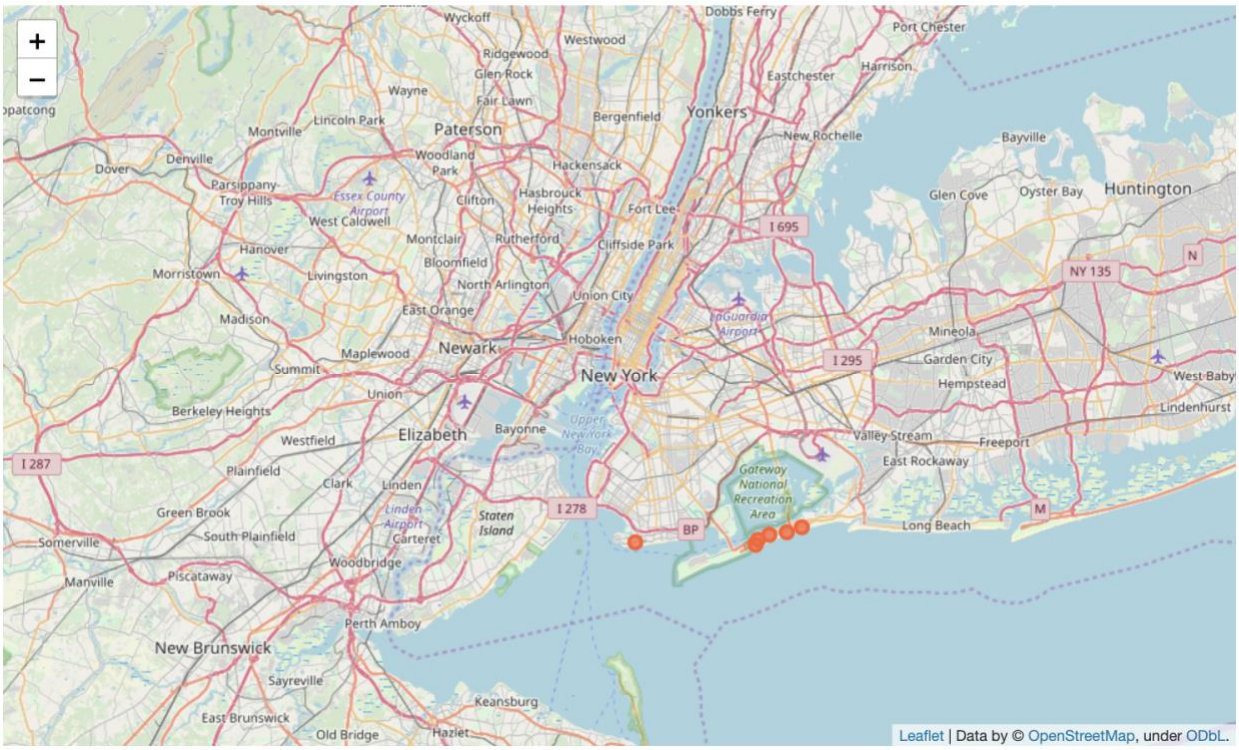
Cluster 6 includes mostly the coffee shops within NYC and Toronto.





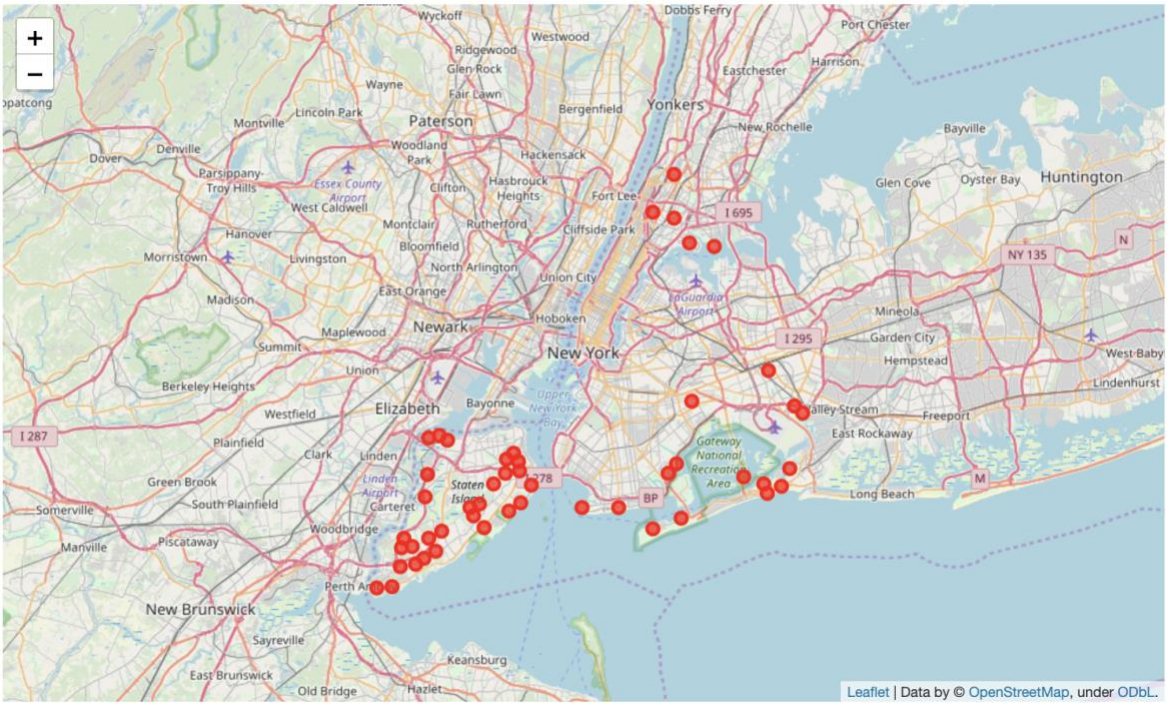
Cluster 7

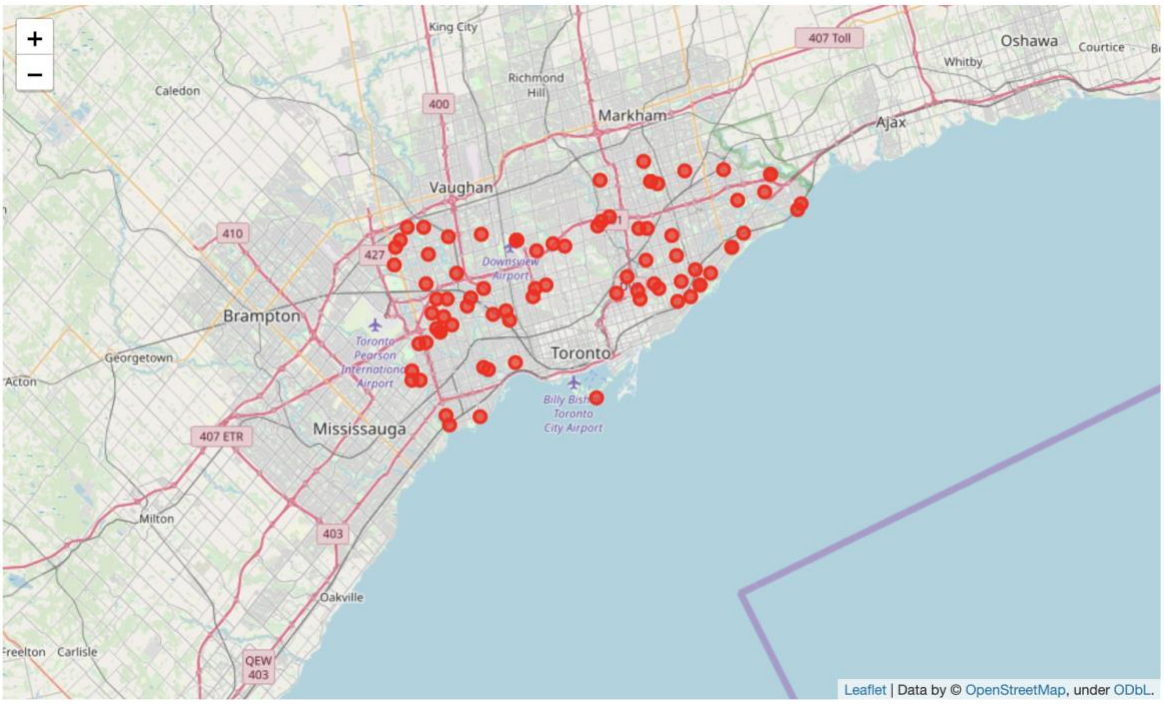
Cluster 7 only contains restaurants in NYC, mostly south and east regions.

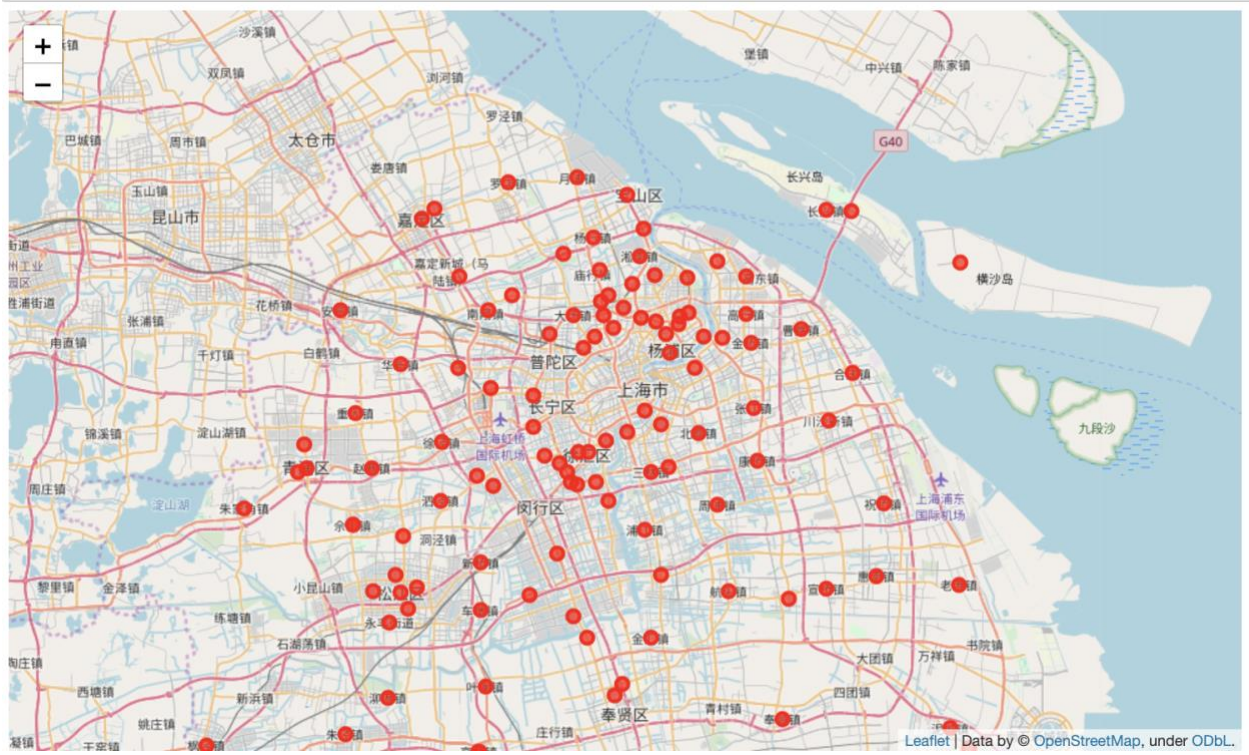


Cluster 8

Cluster 8 contains some recreational public places like parks, hotels, shopping malls and coffee shops, bakery shops as well in these three cities and they basically are all over the geographical region of all three cities.







4.3 Overall conclusion

To be honest, I think that the dataset of Shanghai and the information hidden is not even on the same scale compared to those of NYC and Toronto, which restricts the result of finding more valuable insights. I guess that's probably because Foursquare is more U.S. based and does not have much more relevant data of cities out of North America.

However, I can still tell from the existing results and findings that city of Shanghai is on the same level and is as developed as NYC and Toronto from all aspects, especially from cluster 8. No matter it's about cuisines and restaurants of all countries or other recreational spaces, Shanghai has them all, just like the other two do.