SIMILAR NEIGHBORHOODS IN NEW YORK AND TORONTO

Coursera IBM Applied Data Science Capstone Project

Majube Chavoshi

PROBLEM:

In recent years there has been many immigrants from New York to Toronto and vice versa. On the other hand, moving to a new city is not an easy task to do, specially if one is already satisfied with her neighborhood and is reluctant to leave it. But due to emerging Machine Learning techniques and rich location APIs (such as the one provided by Foursquare), possible solutions are found. In this mini-project, different neighborhoods of New York and Toronto are compared and similar neighborhoods between the two states are found. In this way if one is moving to another city, she can find the most similar neighborhood to her current one and simply relocate there.

DATA:

The data used in this mini-project consists of two dataframes for each city. Each one has the name, latitude and longitude of a single neighborhood. We then use Foursquare API to retrieve top popular venues in each neighborhood.

METHODOLOGY:

The category of each venue is assigned and all venues of a neighborhood are one-hot encoded based on their category. Then the two dataframes are joined together. In order to find the similar neighborhoods in both cities based on popular venues the best machine learning algorithm is clustering, therefore a clustering model is applied on the data and neighborhoods in the same cluster will be the most similar ones to each other.

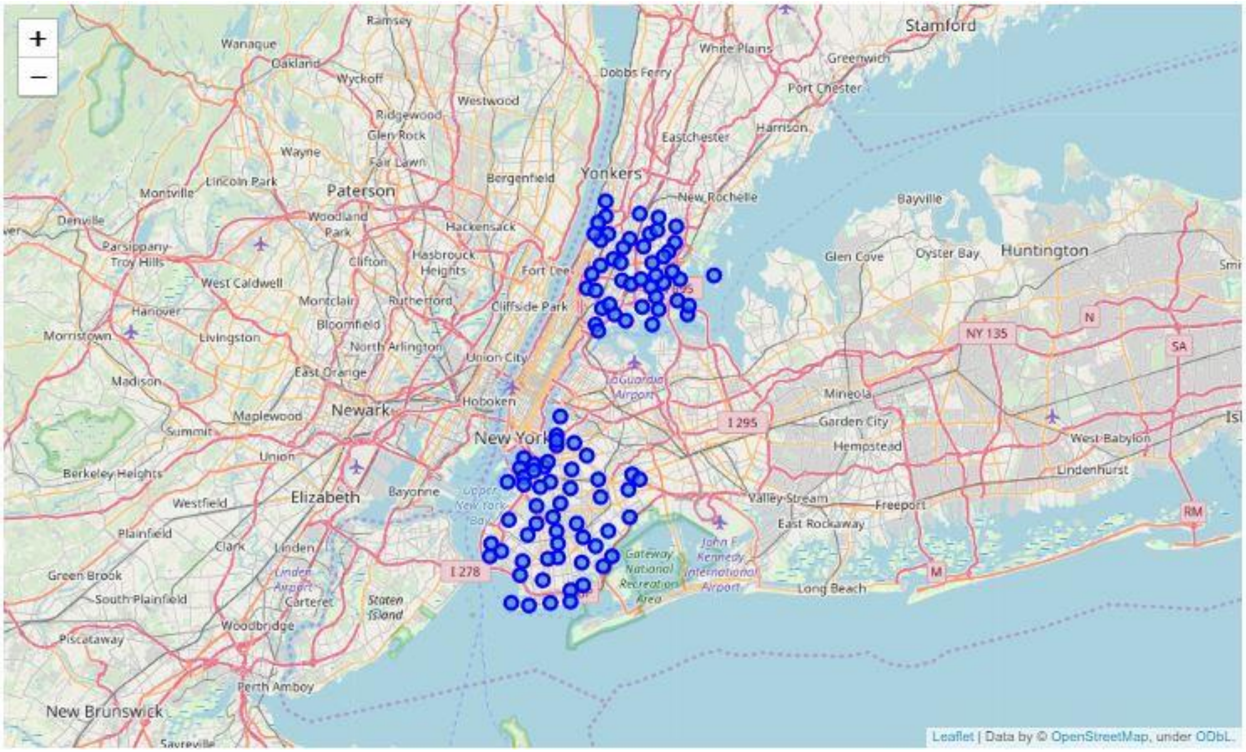
The data for Toronto districts contained different neighborhoods with the same location, these duplicates have been removed.

Note that because of Foursquare API limitations only the first 100 neighborhoods in each city is retrieved and worked on.

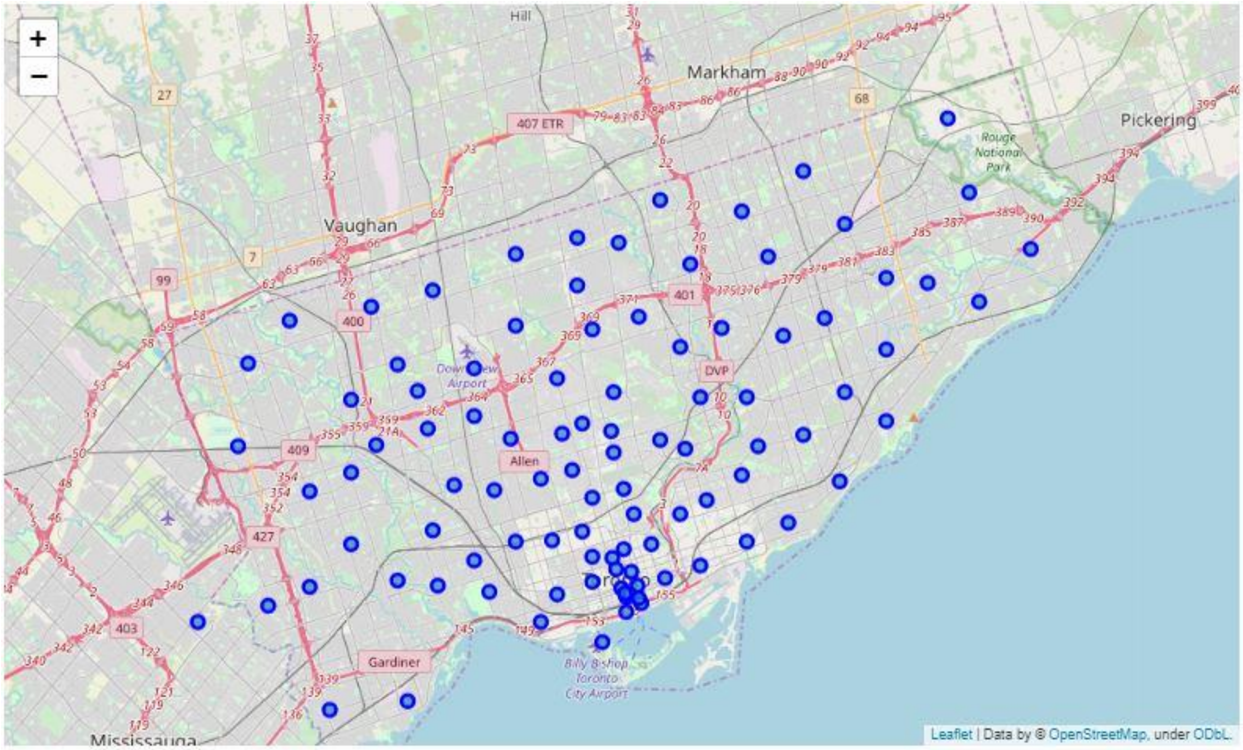
Number of the clusters is fixed on 10 as it is quite optimal in the sense that it gives us more diverse clusters.

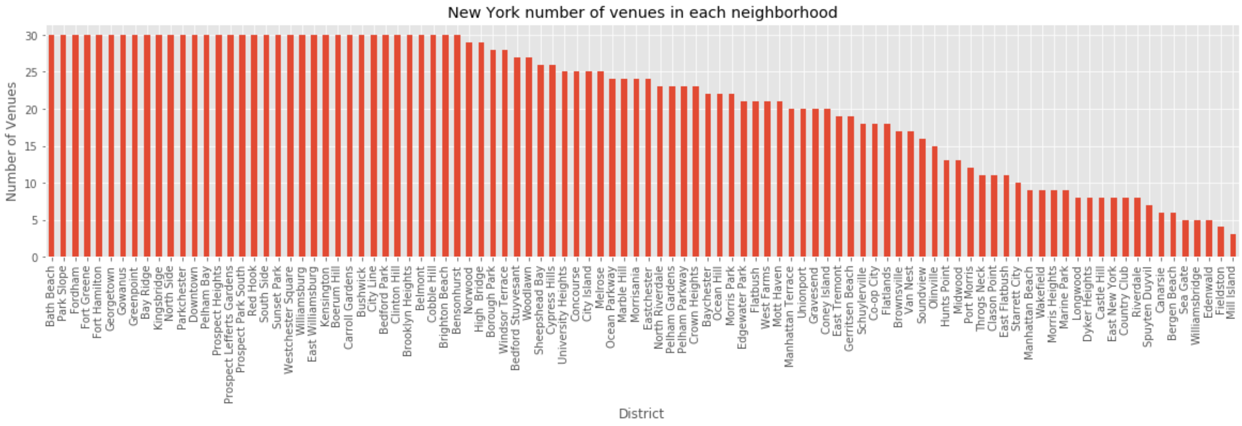
Here’s a map of considered neighborhoods in both cities:

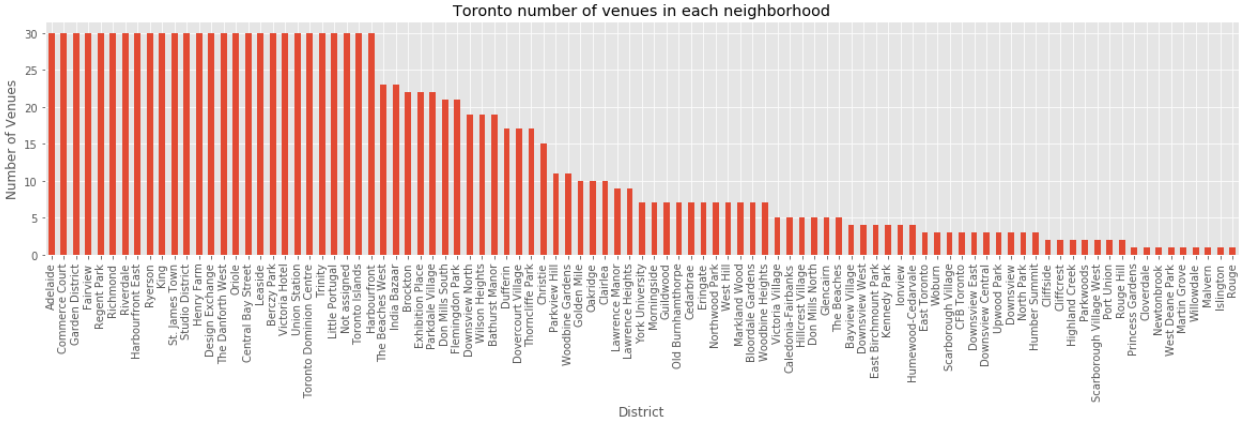
NEW YORK:

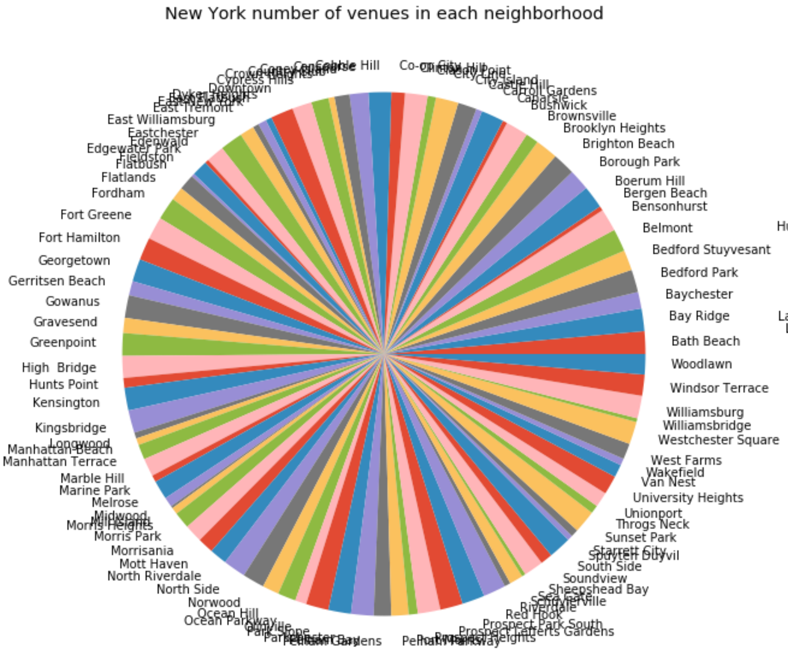


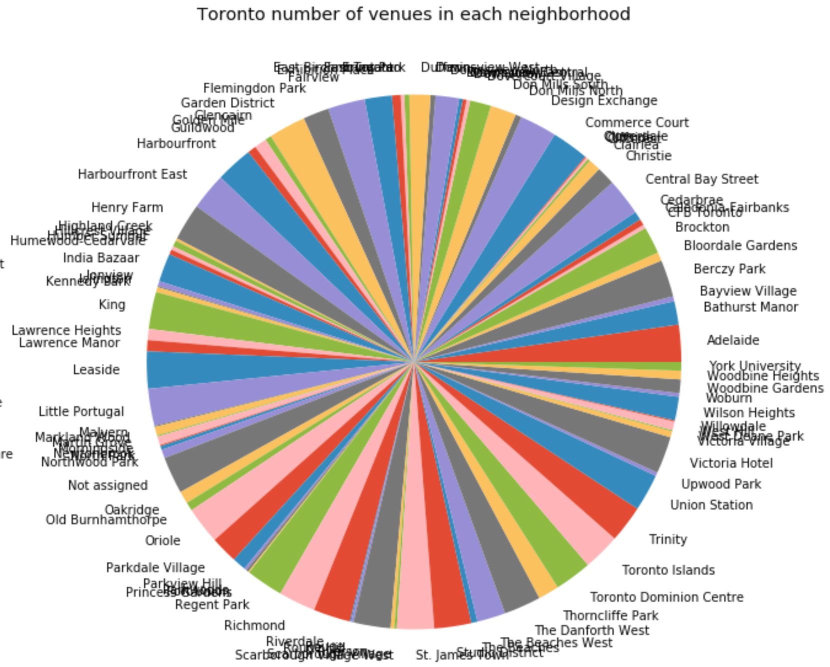
TORONTO:



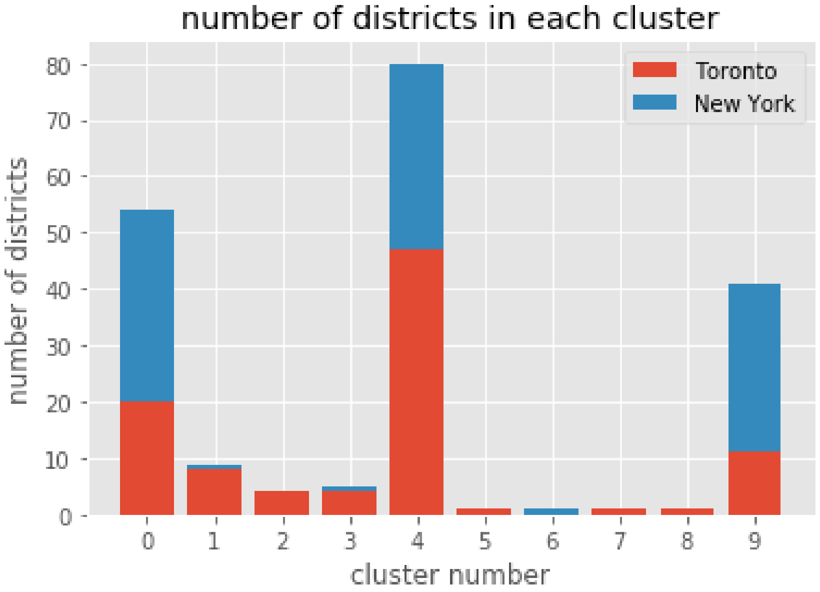

New York number of venues in each neighborhood


Toronto number of venues in each neighborhood


New York number of venues in each neighborhood

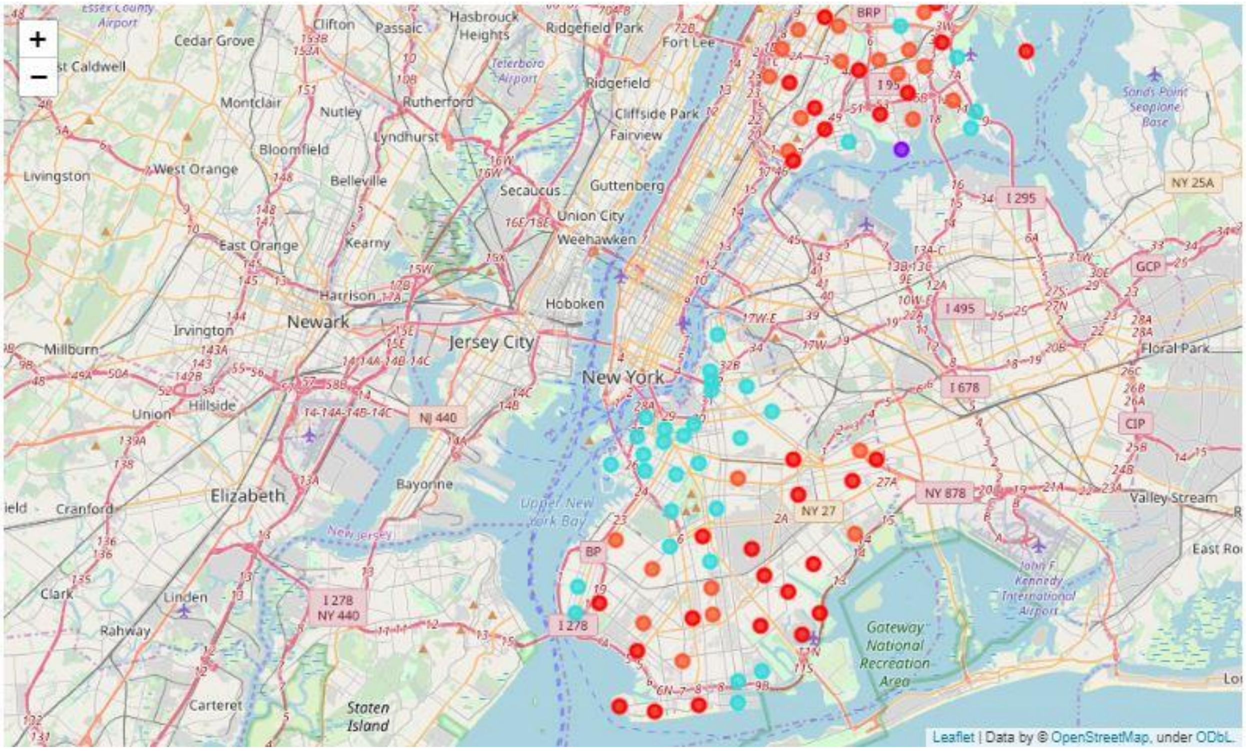

Toronto number of venues in each neighborhood

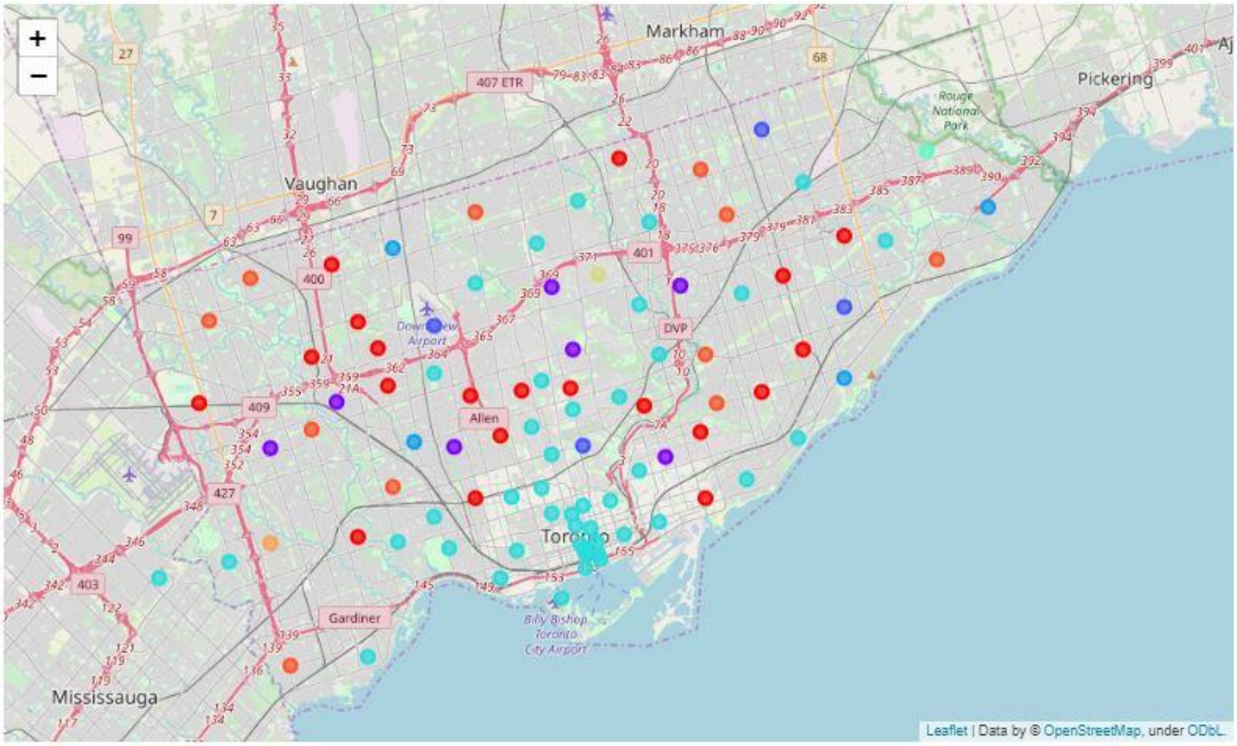
RESULTS:



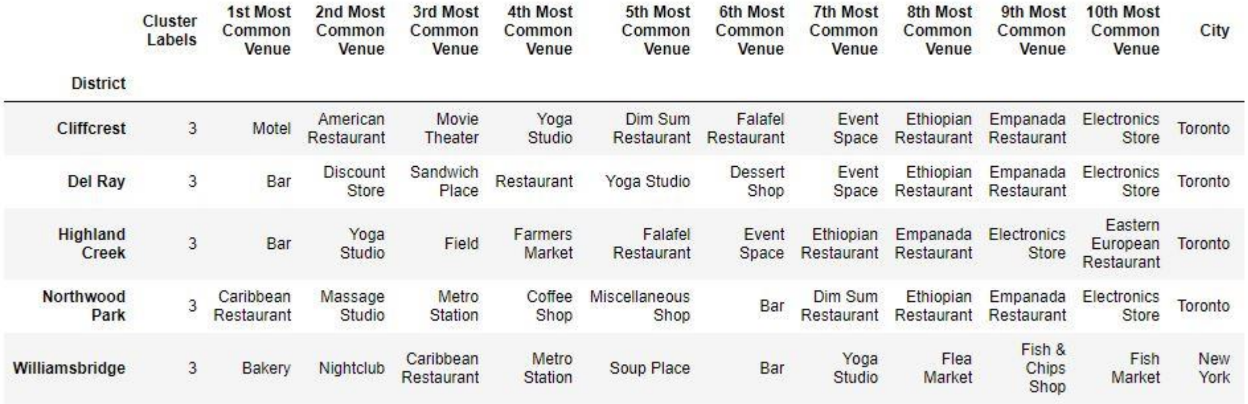
Here’s the number of districts in each cluster is shown. Note that the bar chart is stacked.

And here’s the clusters shown on the map. Districts in the same cluster are shown with same color markers.





Here's an example of a cluster:



District	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	City
Cliffcrest	3	Motel	American Restaurant	Movie Theater	Yoga Studio	Dim Sum Restaurant	Falafel Restaurant	Event Space	Ethiopian Restaurant	Empanada Restaurant	Electronics Store	Toronto
Del Ray	3	Bar	Discount Store	Sandwich Place	Restaurant	Yoga Studio	Dessert Shop	Event Space	Ethiopian Restaurant	Empanada Restaurant	Electronics Store	Toronto
Highland Creek	3	Bar	Yoga Studio	Field	Farmers Market	Falafel Restaurant	Event Space	Ethiopian Restaurant	Empanada Restaurant	Electronics Store	Eastern European Restaurant	Toronto
Northwood Park	3	Caribbean Restaurant	Massage Studio	Metro Station	Coffee Shop	Miscellaneous Shop	Bar	Dim Sum Restaurant	Ethiopian Restaurant	Empanada Restaurant	Electronics Store	Toronto
Williamsbridge	3	Bakery	Nightclub	Caribbean Restaurant	Metro Station	Soup Place	Bar	Yoga Studio	Flea Market	Fish & Chips Shop	Fish Market	New York

DISCUSSION AND CONCLUSION:

I think because of the limitation in number of districts (due to Foursquare API) the clusters are not diverse enough and some clusters are pure. I have also tried Postcode instead of District for Toronto data but the results are not much different.