Capstone Project - The Battle of Neighborhoods

1. Introduction

1.1 Background

We will cluster New York City and the city of Toronto neighbourghoods. In the clusters we will see how similar or dissimilar they are. For this task we select New York City borough Queens. Based on wikipedia information „Queens is the most ethnically diverse urban area in the world It is the most ethnically diverse county in the United States.“. In Toronto we will choose borough North York. Based on wikipedia information „North York is highly multicultural and diverse. In 2016, 56% of North York's residents were not born in Canada, and 60% were classified as belonging to a visible minority.The neighbourhoods of North York are highly diverse, inhabited by people of many different cultures.“

1.2 Problem

We have to gather publically available data and merge it with the Foursquare API to explore neighborhoods in New York City and Toronto. We will use the explore function to get the most common venue categories in each neighborhood, and then use this feature to group the neighborhoods into clusters. We will use the k-means clustering algorithm to complete this task. From Foursquare API explore endpoint we choose section „food“, because in Canada Foursquare API return more less data in compare with USA.

1.3 Interest

Similarities or dissimilarities between cities and they neighborhoods could be very useful for companies whose trying to expand they markets. Lets imagine that company has food business in New York borough Queens and wish to explore possibilities in Toronto borough North York. Based on our report will be possibble to choose neigbourghoods where the same food type industries exist or not.

2. Data section

2.1 New York data

We receive New York neighbourghoods data from https://cocl.us/new_york_dataset. Data exist on JSON format so we have to parse data and put into our dataframe. Finally data format will be

Borough Neighborhood Latitude Longitude

2.2 Toronto data

We receive Toronto neighbourghoods data from web page https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. To gather data we will us Beautiful Soup a Python library for pulling data out of HTML . Because in this page we don't have information about Lattitude and Longitude we download csv file http://cocl.us/Geospatial_data/Geospatial_Coordinates.csv and merge this data to format the same we used in New York :

Borough Neighborhood Latitude Longitude

2.3 Toronto North York and New York Queens data

Finally we merge Toronto and NewYork dataframes leaving neigborhood whoose depends to Toronto North York and New York Queens boroughs. Data format the same:

Borough Neighborhood Latitude Longitude

	Postal Code	Borough	Neighbourghood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M6A	North York	Lawrence Heights, Lawrence Manor	43.718518	-79.464763
2	M3B	North York	Don Mills North	43.745906	-79.352188
3	M6B	North York	Glencairn	43.709577	-79.445073
4	M3C	North York	Flemingdon Park, Don Mills South	43.725900	-79.340923
5	M2H	North York	Hillcrest Village	43.803762	-79.363452
6	M3H	North York	Bathurst Manor, Downsview North, Wilson Heights	43.754328	-79.442259
7	M2J	North York	Fairview, Henry Farm, Oriole	43.778517	-79.346556
8	M3J	North York	Northwood Park, York University	43.767980	-79.487262
9	M3L	North York	Downsview West	43.739015	-79.506944
10	M6L	North York	Downsview, North Park, Upwood Park	43.713756	-79.490074
11	M9L	North York	Humber Summit	43.756303	-79.565963
12	M5M	North York	Bedford Park, Lawrence Manor East	43.733283	-79.419750
13	M2N	North York	Willowdale South	43.770120	-79.408493
14	M2R	North York	Willowdale West	43.782736	-79.442259
15	NYC	Queens	Astoria	40.768509	-73.915654
16	NYC	Queens	Woodside	40.746349	-73.901842
17	NYC	Queens	Jackson Heights	40.751981	-73.882821
18	NYC	Queens	Elmhurst	40.744049	-73.881656
19	NYC	Queens	Howard Beach	40.654225	-73.838138
20	NYC	Queens	Corona	40.742382	-73.856825
21	NYC	Queens	Forest Hills	40.725264	-73.844475
22	NYC	Queens	Kew Gardens	40.705179	-73.829819
23	NYC	Queens	Richmond Hill	40.697947	-73.831833

2.4 Foursquare data

Use the Foursquare API to explore neighborhoods in New York City and Toronto. We will use the explore function to get the most common venue categories in each neighborhood . We add additional parameter in Foursquare Explore endpoint. Section parameter "food". So we on request receive information based on food industry. Finnally managed data will be formated

Neighborhood 1st Most Common Venue 2nd Most Common Venue 3rd Most Common Venue ....

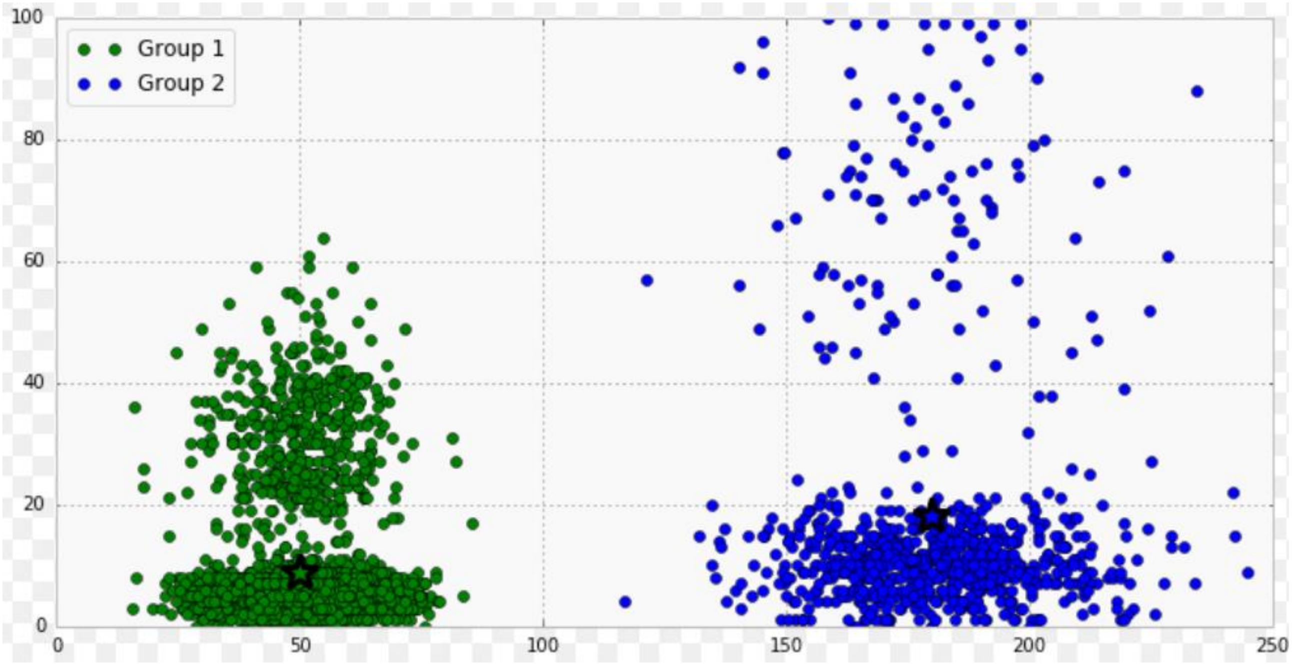
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Astoria	Hookah Bar	Bar	Pub
1	Astoria Heights	Hotel Bar	Bowling Alley	Cocktail Bar
2	Auburndale	Bar	Hookah Bar	Wine Bar
3	Bathurst Manor, Downsview North, Wilson Heights	Bar	Wine Bar	Caribbean Restaurant
4	Bay Terrace	American Restaurant	Whisky Bar	Hotel
5	Bayside	Bar	Pub	Wine Bar
6	Bedford Park, Lawrence Manor East	American Restaurant	Comfort Food Restaurant	Pub
7	Bellaire	Nightlife Spot	Wine Bar	Hotel Bar
8	Belle Harbor	Bar	Beach Bar	Pub
9	Bellerose	Bar	Sports Bar	Pub
10	Blissville	Bar	Wine Bar	Caribbean Restaurant
11	Briarwood	Nightlife Spot	Wine Bar	Hotel Bar
12	Broad Channel	Bar	Dive Bar	Pub
13	Cambria Heights	Nightlife Spot	Bar	Lounge
14	College Point	Bar	Karaoke Bar	Hotel Bar
15	Corona	Wine Bar	Nightclub	Hotel Bar
16	Douglaston	Bar	Pub	Lounge
17	East Elmhurst	Hotel Bar	Lounge	Wine Bar

With this data we have to do clustering tasks

3. Methodology section

3.1 K means clustering

To segment Toronto and New York neighborhoods we will use K-means clustering. K-means can group data only unsupervised based on the similarity of neigborhoods to each other. K-means is a type of partitioning clustering. That is, it divides the data into k non-overlapping subsets or clusters without any cluster internal structure or labels. This means, it's an unsupervised algorithm. Objects within a cluster are very similar and objects across different clusters are very different or dissimilar. Though the objective of K-means is to form clusters in such a way that similar samples go into a cluster and dissimilar samples fall into different clusters, it can be shown that instead of a similarity metric, we can use dissimilarity metrics. In other words, conventionally, the distance of samples from each other is used to shape the clusters. So, we can say, K-means tries to minimize the intra-cluster distances and maximize the inter-cluster distances. Clustering example :

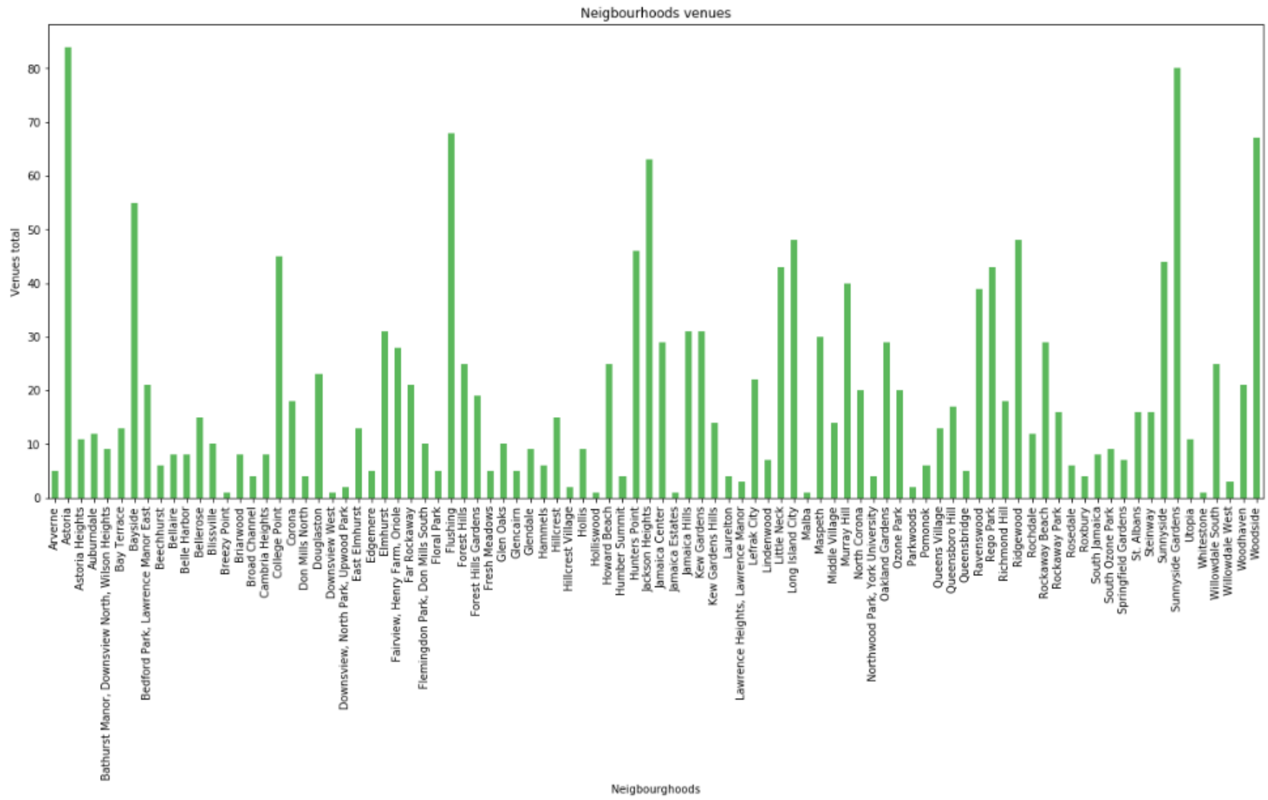


4. Results section

We have to check how many venues were returned for each neighborhood. Returned results based on Foursquare API explore enpoint with selection criteria food.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Arverne	5	5	5	5	5	5
Astoria	84	84	84	84	84	84
Astoria Heights	11	11	11	11	11	11
Auburndale	12	12	12	12	12	12
Bathurst Manor, Downsview North, Wilson Heights	9	9	9	9	9	9
Bay Terrace	13	13	13	13	13	13
Bayside	55	55	55	55	55	55
Bedford Park, Lawrence Manor East	21	21	21	21	21	21
Beechhurst	6	6	6	6	6	6
Bellaire	8	8	8	8	8	8
Belle Harbor	8	8	8	8	8	8
Bellerose	15	15	15	15	15	15

Managed graph for more detail visualisation:



Based on k-means clustering method we received four different clusters.

Cluster 1 results.

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
66	Queens	0	Indian Restaurant	Chinese Restaurant	Pizza Place
68	Queens	0	Indian Restaurant	Wings Joint	Filipino Restaurant

Cluster 2 results

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
10	North York	1	Deli / Bodega	Bakery	Filipino Restaurant
14	North York	1	Pizza Place	Bakery	Wings Joint
28	Queens	1	Deli / Bodega	Chinese Restaurant	Pizza Place
29	Queens	1	Deli / Bodega	Bakery	Italian Restaurant
30	Queens	1	Deli / Bodega	Pizza Place	Bakery
32	Queens	1	Deli / Bodega	Mexican Restaurant	Pizza Place
34	Queens	1	Deli / Bodega	Fast Food Restaurant	Donut Shop
36	Queens	1	Deli / Bodega	Filipino Restaurant	Dim Sum Restaurant
38	Queens	1	Korean Restaurant	Deli / Bodega	Italian Restaurant
42	Queens	1	Deli / Bodega	Chinese Restaurant	Pizza Place
45	Queens	1	Deli / Bodega	Fast Food Restaurant	Indian Restaurant
56	Queens	1	Pizza Place	Deli / Bodega	Chinese Restaurant
57	Queens	1	Pizza Place	Deli / Bodega	Halal Restaurant
59	Queens	1	Deli / Bodega	Sushi Restaurant	Café
60	Queens	1	Chinese Restaurant	Deli / Bodega	Donut Shop
62	Queens	1	Pizza Place	Deli / Bodega	Asian Restaurant
72	Queens	1	Bakery	Deli / Bodega	Donut Shop
74	Queens	1	Deli / Bodega	Bakery	Donut Shop
75	Queens	1	Deli / Bodega	Donut Shop	Chinese Restaurant
77	Queens	1	Chinese Restaurant	Deli / Bodega	Greek Restaurant
78	Queens	1	Deli / Bodega	Pizza Place	Chinese Restaurant
81	Queens	1	Deli / Bodega	Afghan Restaurant	Bakery
83	Queens	1	Deli / Bodega	Chinese Restaurant	Italian Restaurant
86	Queens	1	Deli / Bodega	Donut Shop	Restaurant
87	Queens	1	Pizza Place	Fast Food Restaurant	Deli / Bodega
91	Queens	1	Sandwich Place	Spanish Restaurant	Bakery

Cluster 3 results.

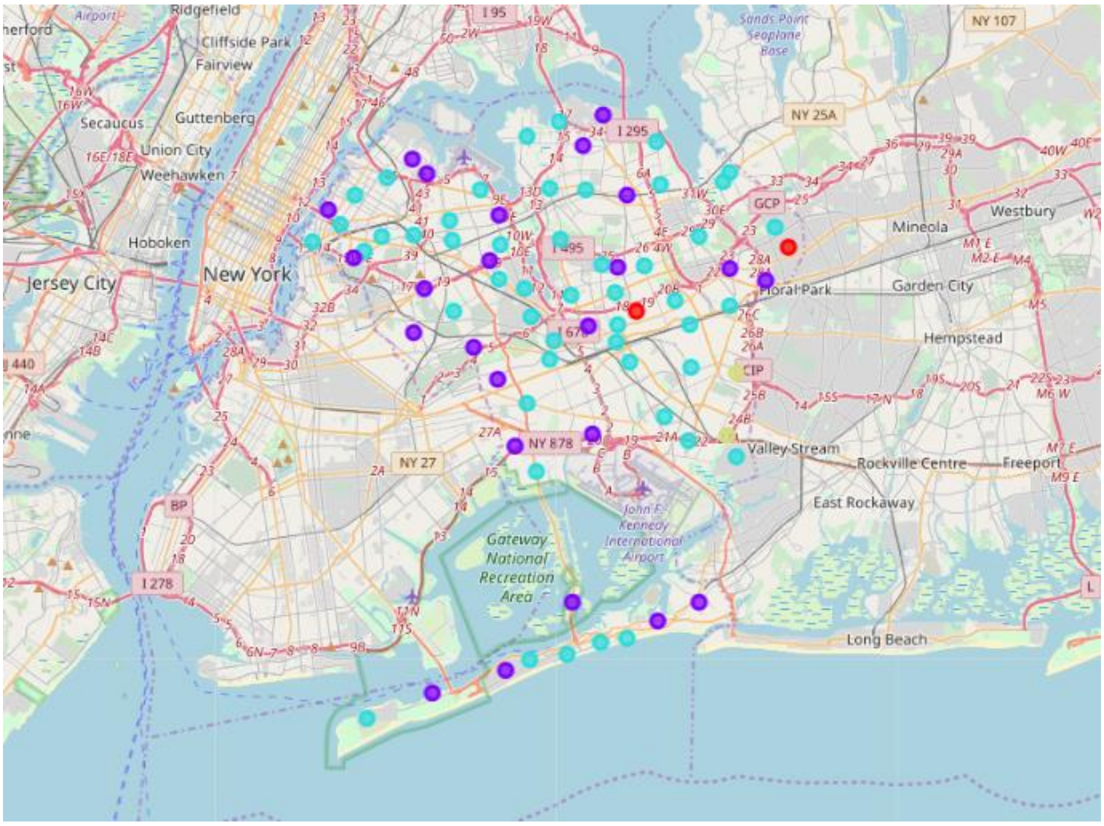
	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	North York	2	BBQ Joint	Fast Food Restaurant	Wings Joint
1	North York	2	Hot Dog Joint	Vietnamese Restaurant	Filipino Restaurant
2	North York	2	Restaurant	Caribbean Restaurant	Japanese Restaurant
3	North York	2	Pizza Place	Asian Restaurant	Sushi Restaurant
4	North York	2	Asian Restaurant	Dim Sum Restaurant	Café
5	North York	2	Mediterranean Restaurant	Fast Food Restaurant	Wings Joint
6	North York	2	Pizza Place	Deli / Bodega	Fried Chicken Joint
7	North York	2	Fast Food Restaurant	Asian Restaurant	Restaurant
8	North York	2	Pizza Place	Caribbean Restaurant	Eastern European Restaurant
9	North York	2	Wings Joint	Filipino Restaurant	Dim Sum Restaurant
11	North York	2	Pizza Place	Food Truck	Italian Restaurant
12	North York	2	Italian Restaurant	Pizza Place	Fast Food Restaurant
13	North York	2	Ramen Restaurant	Restaurant	Sushi Restaurant
15	Queens	2	Middle Eastern Restaurant	Bakery	Deli / Bodega
16	Queens	2	Deli / Bodega	Bakery	Thai Restaurant
17	Queens	2	Latin American Restaurant	South American Restaurant	Peruvian Restaurant
18	Queens	2	Thai Restaurant	Mexican Restaurant	Chinese Restaurant
19	Queens	2	Italian Restaurant	Bagel Shop	Chinese Restaurant
20	Queens	2	Mexican Restaurant	Chinese Restaurant	Bakery
21	Queens	2	Food Truck	Deli / Bodega	Thai Restaurant
22	Queens	2	Chinese Restaurant	Bakery	Deli / Bodega
23	Queens	2	Pizza Place	Chinese Restaurant	Latin American Restaurant
24	Queens	2	Chinese Restaurant	Korean Restaurant	Bakery
25	Queens	2	Pizza Place	Deli / Bodega	Food Truck
26	Queens	2	Deli / Bodega	Italian Restaurant	Pizza Place
27	Queens	2	Donut Shop	Deli / Bodega	Snack Place

Cluster 4 results.

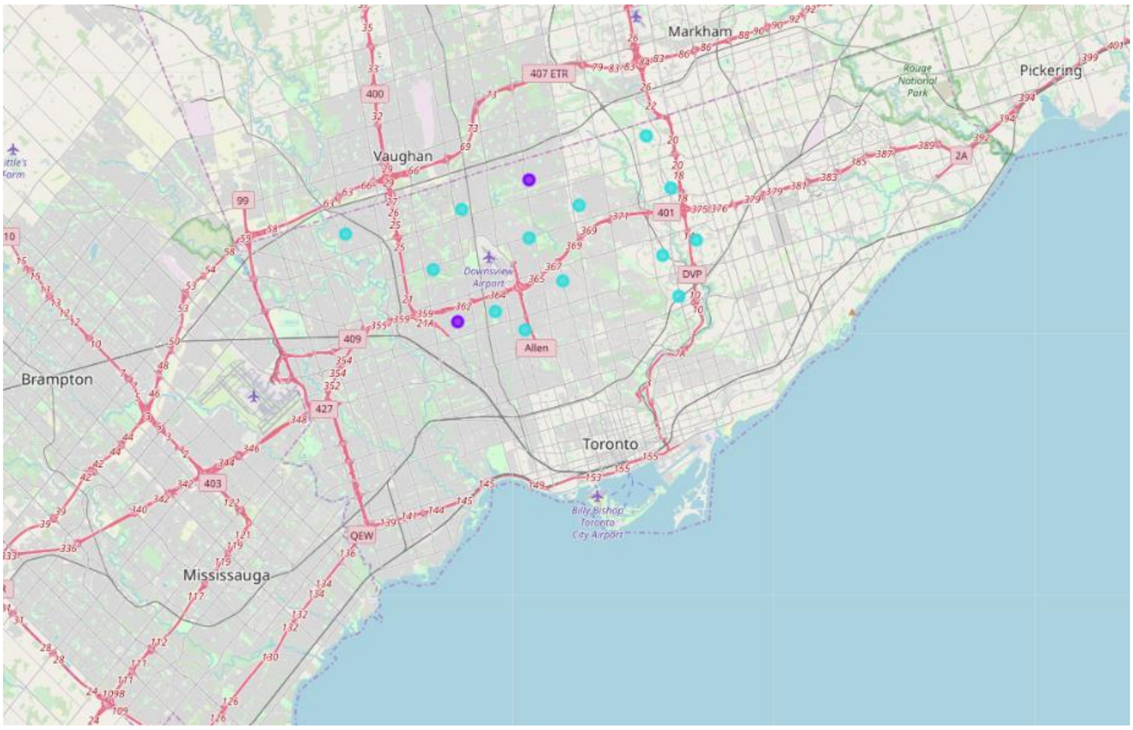
	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
54	Queens	3	Caribbean Restaurant	Restaurant	Chinese Restaurant
73	Queens	3	Caribbean Restaurant	Wings Joint	Cuban Restaurant

You can see a clustered map boroughs of Toronto borough North York and New York City borough Queens in the below.

New York City borough Queens



Toronto borough North York

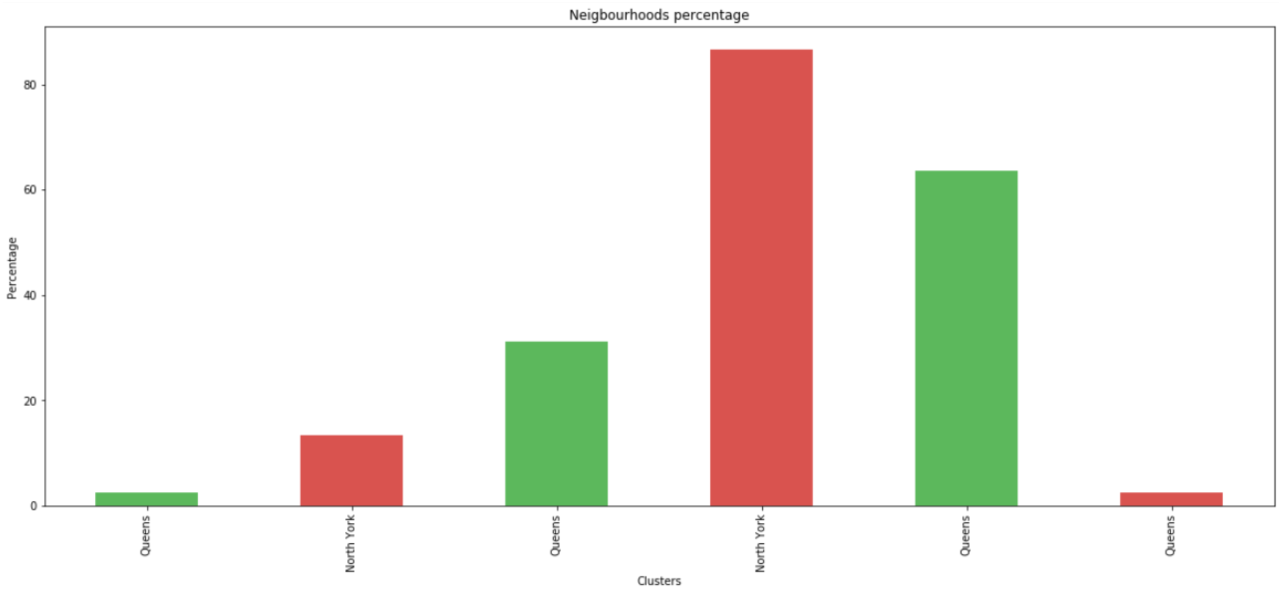


5. Discussion section

Neigbourghoods in Toronto *North York* and New York Queens has different numbers of Neigborhoods so we have to normalyse this doing this in Percentage representation. So after normalisation we have possibilyte to see results in mor common way. Normalisation table I did look bellow.

		Cluster Labels	Borough	Count	Percentage
Cluster Labels	**Borough**				
0	**Queens**	0	Queens	2	2.60
1	**North York**	1	North York	2	13.33
	Queens	1	Queens	24	31.17
2	**North York**	2	North York	13	86.67
	Queens	2	Queens	49	63.64
3	**Queens**	3	Queens	2	2.60

Results graphical representation



6. Conclusion section

In this study, I analyzed the relationship between cities Toronto and New York. From both cities I am selected New York City borough Queens and Toronto North York . From introduction we know that this boroughs are most ethnically diverse urban areas. Looking in the results we can conclude that basically to largest clusters shows more similarity than asimiliraty.This models allow us to shoose which food industry venue will be good in one or other neighbourghood. These models can be very useful in helping food markets management and expansion between different countries and cities.