The Battle of Neighborhoods

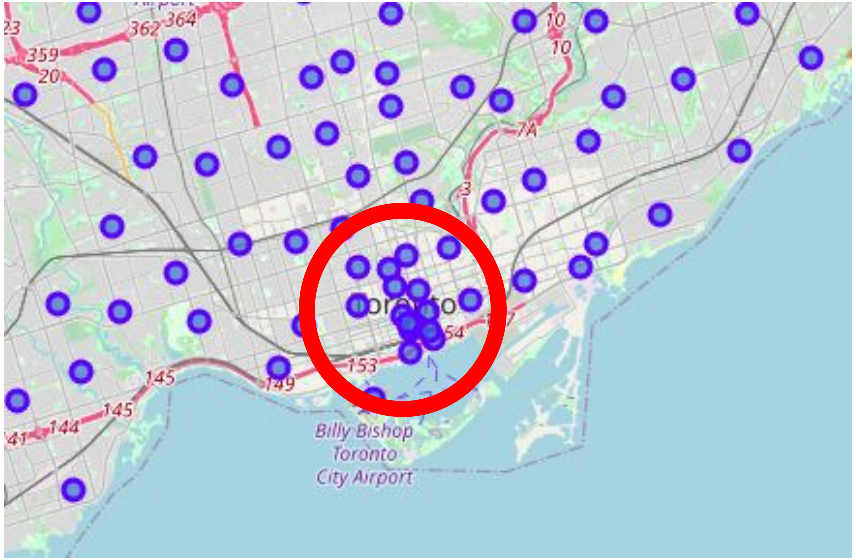
New York - Toronto Comparison

Introduction

The main interest is to analyze the similarity of different groups of neighborhoods in two cities in different countries. Many businesses are interested in being able to expand internationally. However, when it comes to finding a place to establish a new business, it may not be enough that it corresponds to a popular place.

Introduction

We could see that the most "central" places can respond to different interests depending on the country in which they are located, so choosing a more appropriate neighborhood can improve future projections of the business, a neighborhood that before this analysis I would never have thought of .



Data - New York

First of all, the information of the neighborhoods of New York is obtained, which is obtained from https://cocl.us/new_york_dataset. With these data a table is constructed only with the data of interest; Borough Neighborhood Latitude Longitude.

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Data - Toronto

In the same way, the data for Toronto neighborhoods are then obtained from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. So we are only interested in the neighborhood name and the coordinates. The rest of the information will be obtained from Foursquare API.

Borough	Neighborhood	Latitude	Longitude
North York	Parkwoods	43.753259	-79.329656
North York	Victoria Village	43.725882	-79.315572
Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

Data - Foursquare

With the Foursquare API, giving the coordinates we can get the all the info of the venues around the neighborhood.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant
Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop

Methodology

To analyze the data, a single final dataset was built that included information from Las Venues and all neighborhoods in both cities. For that dataset, we sort by most common venues categories nearby of the neighborhoods.

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
Allerton	Pizza Place	Deli / Bodega	Department Store	Cosmetics Shop	Supermarket	Spa
Annadale	Food	Diner	Pizza Place	Train Station	Pub	Sports Bar
Arden Heights	Hotel	Pharmacy	Smoke Shop	Coffee Shop	Pizza Place	Field
Arlington	Bus Stop	Deli / Bodega	Intersection	Liquor Store	Boat or Ferry	American Restaurant
Arrochar	Deli / Bodega	Italian Restaurant	Bus Stop	Pizza Place	Middle Eastern Restaurant	Food Truck

Methodology

Then, we group the data to use K-Means Clustering with ***sklearn*** python library.

Yoga Studio	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Methodology

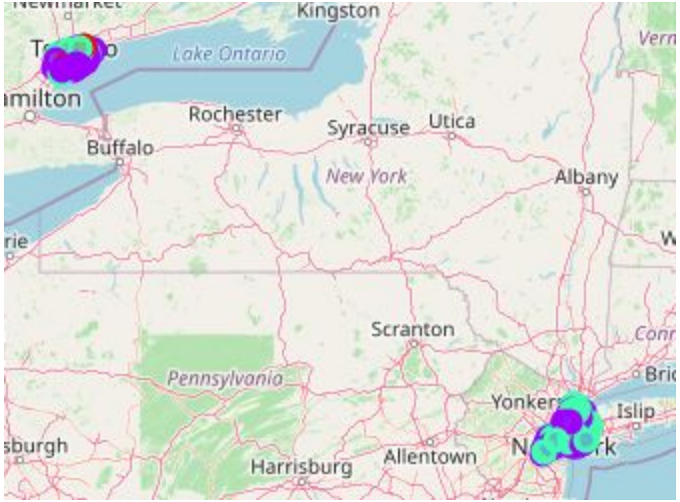
Iterating by **k** values, we check that k-means method return best clustering with **3** total clusters. We can check the result of try with 3 clusters and with 7 clusters. The right column for each table, show the number of neighborhoods in each cluster. With **k=7** we have a lot of "too small clusters".

```
In [124]: df_merged['Cluster Labels'].value_counts()

Out[124]: 2.0    198
          1.0    195
          0.0     12
          Name: Cluster Labels, dtype: int64
```

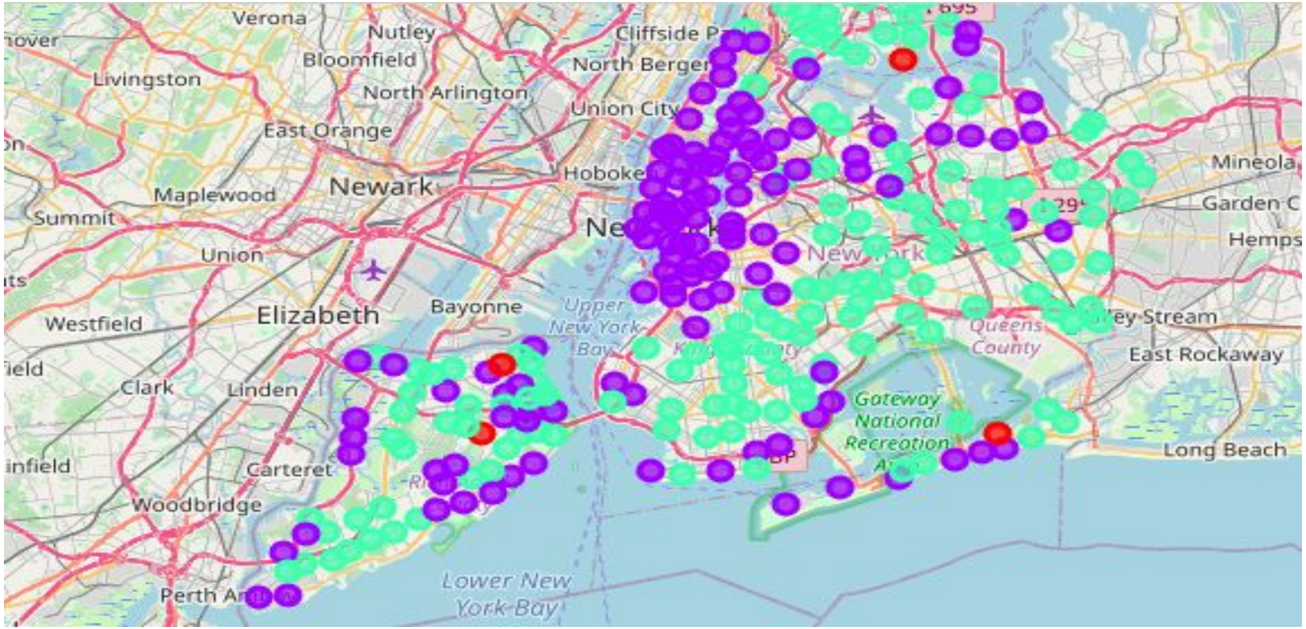
```
In [119]: df_merged['Cluster Labels'].value_counts()

Out[119]: 2.0    286
          4.0     91
          0.0     12
          1.0     11
          3.0      2
          5.0      2
          6.0      1
          Name: Cluster Labels, dtype: int64
```

Results

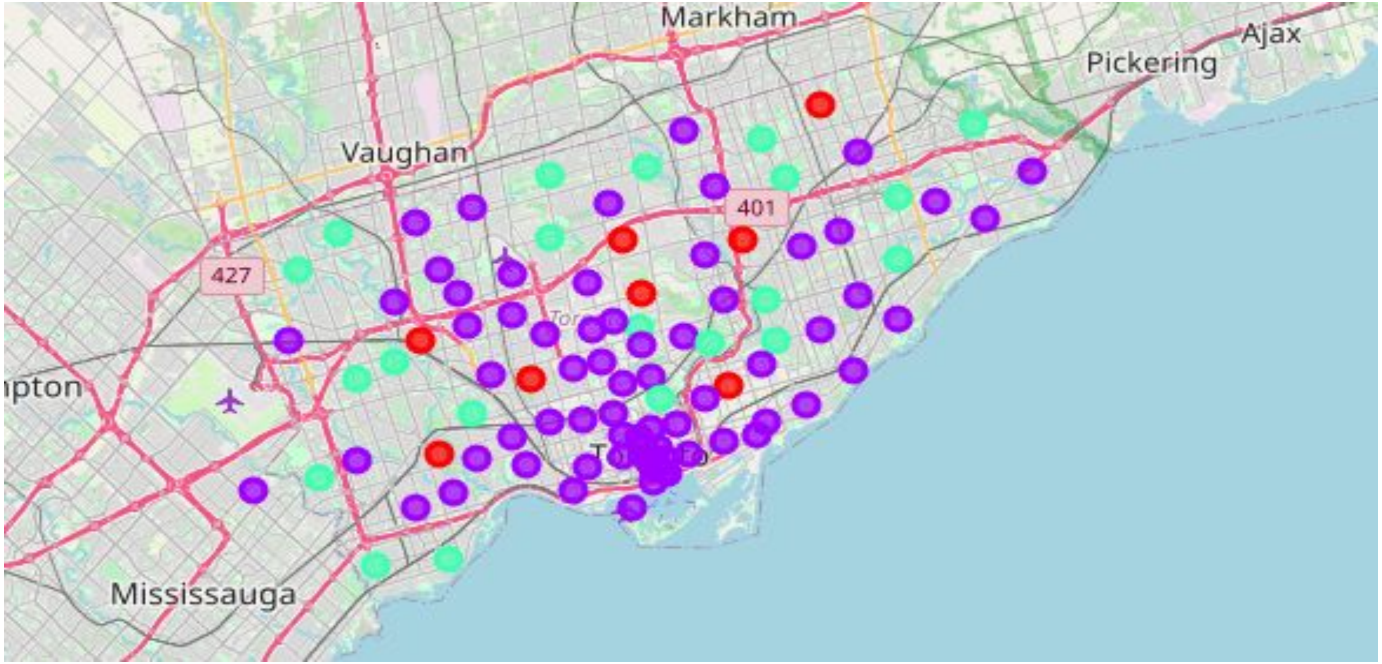
After run k-means method, we check the maps with the generated data.



Results - New York



Results - Toronto



Results

- We can check by the colored marks, the different of the distribution of the neighborhoods around the each city.
- Now we want to know about the resulting clusters.

Results - Cluster 0

1st Common Venues:

Park	9
Food & Drink Shop	1
Convenience Store	1

2nd Common Venues:

Women's Store	4
Park	3
Convenience Store	2

3rd Common Venues:

Dumpling Restaurant	3
Pizza Place	2
Pool	2

Results - Cluster 1

1st Common Venues:

Coffee Shop	27
Italian Restaurant	20
Bar	16

2nd Common Venues:

Coffee Shop	24
Park	11
Café	9

3rd Common Venues:

Bar	12
Coffee Shop	11
Café	9

Results - Cluster 2

1st Common Venues:

Pizza Place	37
Chinese Restaurant	15
Bank	12

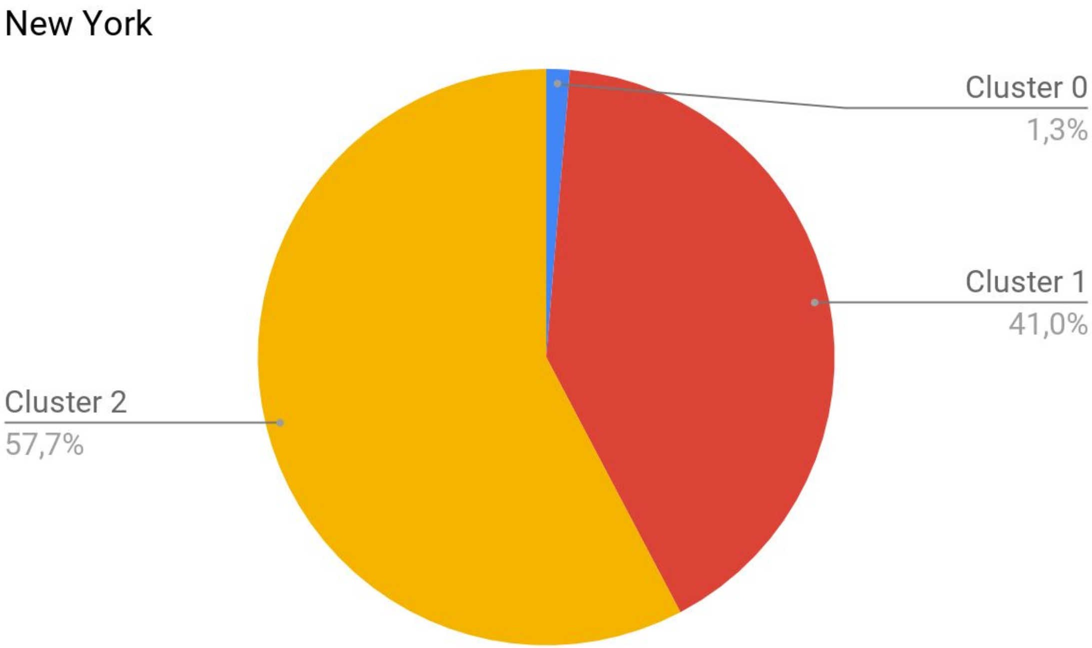
2nd Common Venues:

Pizza Place	22
Deli / Bodega	18
Grocery Store	10

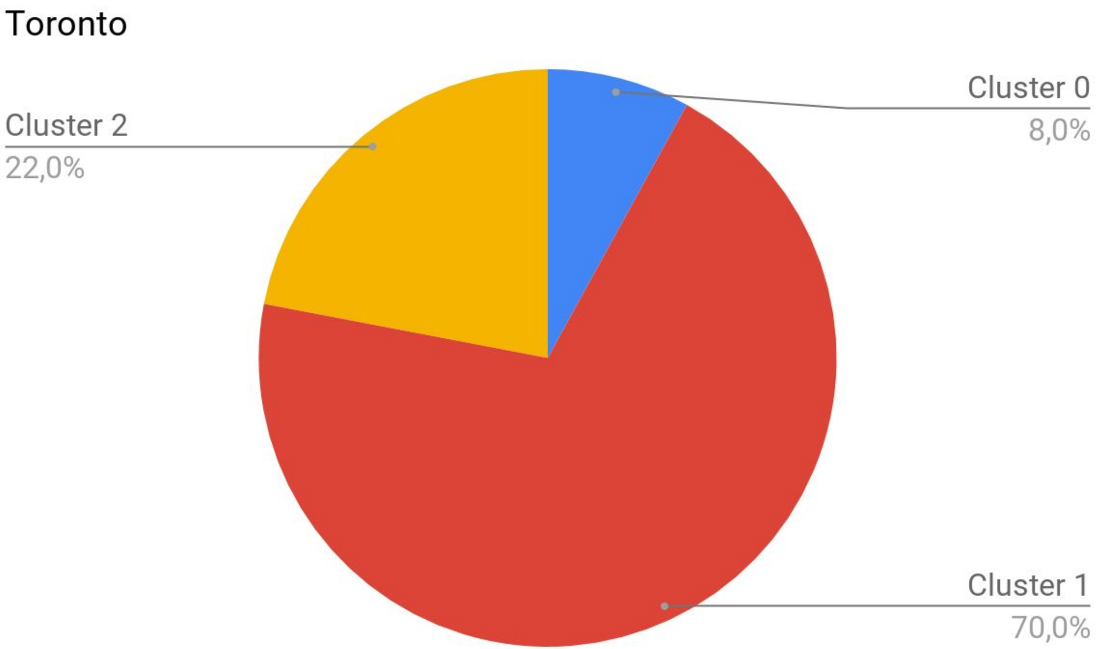
3rd Common Venues:

Pizza Place	13
Bakery	12
Bank	10

Results - New York Distribution



Results - Toronto Distribution

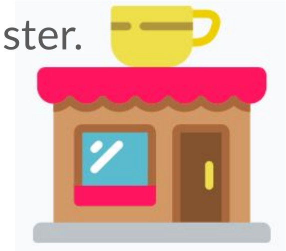


Discussion

Given the results, we analyze the main clusters. While cluster number 1 corresponds to neighborhoods that have preferences for coffee shops, parks and iItalian food restaurants; Cluster number 2 contains the neighborhoods that show a preference for pizza places, Chinese food, banks and bakeries.

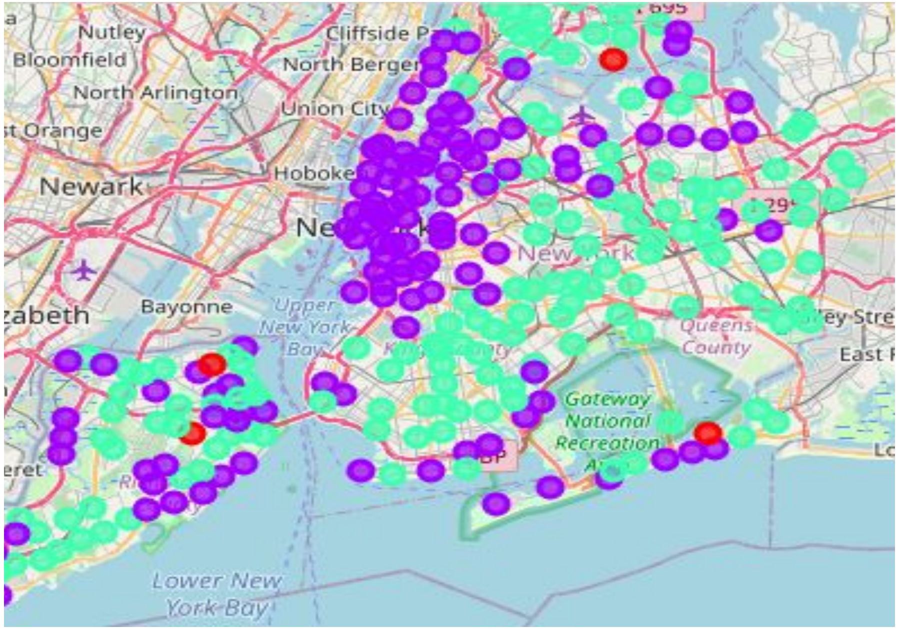
Discussion

We can see from the results graph that both cities have notable differences between the proportions of their neighborhoods for each type of cluster.



Discussion

We can see to that the most neighborhoods with type cluster 1 are located in Manhattan. While in the rest of New York, most are type cluster 2. So, i.e. if you want open a new coffee shop, probably your best option in New York are the Manhattan's neighborhoods.



Conclusion

As we proposed, different cultures of countries suppose different distributions and locations of neighborhoods. As we proposed, different cultures of countries suppose different distributions and locations of neighborhoods. On the other hand, many times the information overcomes the intuition and when making decisions it is necessary to review all the available information, visualize it and see behaviors according to empirical data. In our case, we achieved differences in clear preferences for the most populated areas of each city.

Thanks!

By Enrique Urrutia