Applied Data Science Capstone

Inhalt

Introduction

My capstone projects looks at the Neighborhood of different cities in different countries and compares them by building clusters. The research question is rather a sociological one than a business problem. The idea is to check, whether the neighborhoods of one city will end up in one cluster or whether there are similar neighborhood structures that exists in different countries. This can be a signal of whether the national culture is dominant or local cultures develop on their own. The two cities analyzed are New York and Toronto. For further research, the analysis could be run multiple times to see whether characteristics of the country have an impact on which culture (national or local) is the more important one.

Data

The data used in this capstone are the foursquare location data for developers. The data are accessible via API. They contain location-based data on the venues in a requested area. I will restrict my analysis to use

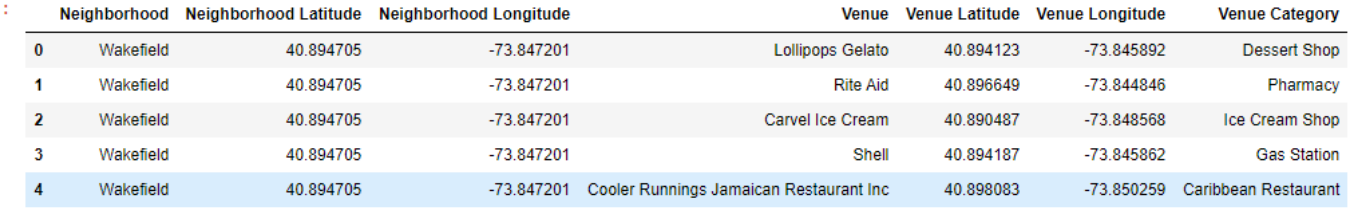
- the name of a venue (not used in clustering, but for the purpose of documentation)
- the location (longitude and latitude)
- the category (e.g. Italian)

Analyzed venues are restricted to those in a radius of 500.

I analyze the neighborhoods of Toronto and New York. Both places are rather similar, as they belong to the western world and are in North America. Still they belong to different countries and may have different local traditions, which are reflected in the local venue structure.

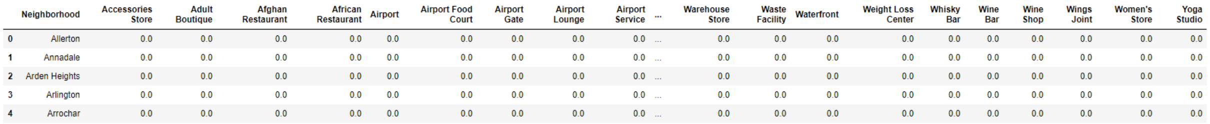
Methodology

First all relevant Neighborhoods of Toronto and New York are identified. For those neighborhoods, all available venues are listed including category.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
2	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
3	Wakefield	40.894705	-73.847201	Shell	40.894187	-73.845862	Gas Station
4	Wakefield	40.894705	-73.847201	Cooler Runnings Jamaican Restaurant Inc	40.898083	-73.850259	Caribbean Restaurant

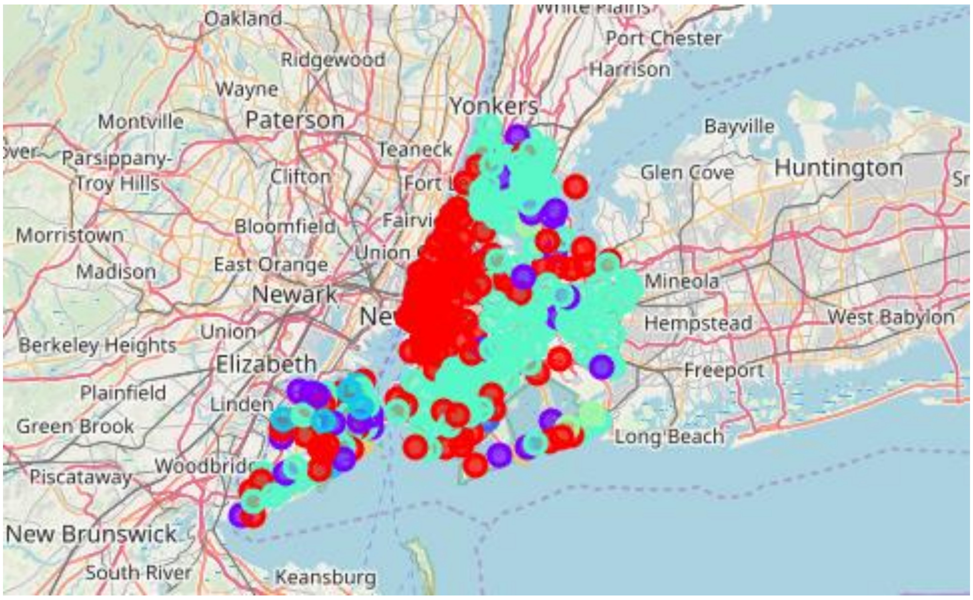
As Input for Clustering, the data are normalized by calculating the share of the venue categories by Neighborhood.



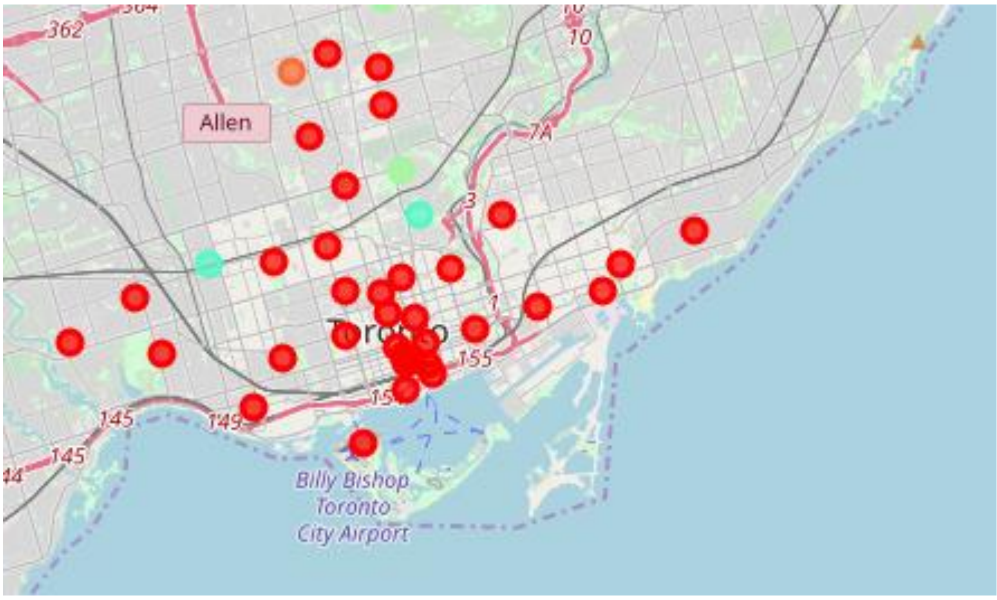
	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	...	Warehouse Store	Waste Facility	Waterfront	Weight Loss Center	Whisky Bar	Wine Bar	Wine Shop	Wings Joint	Women's Store	Yoga Studio
0	Allerton	0.0	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
1	Annadale	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
2	Arden Heights	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
3	Arlington	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
4	Arrochar	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

The preprocessed dataset results in **11,957 different venues** from **337 different neighborhoods** that will be analyzed.

The Clustering of neighborhoods is done using the kmeans algorithm. This method builds clusters that are most similar within and diverse between clusters. The analysis is carried out with only 8 clusters and a random starting point for each cluster center. As clustering is a technique of unsupervised learning the results are interpreted and discussed but no statistical testing is applied.



Resulting clusters New York



Resulting clusters Toronto

If the national culture dominate the local one, then I would expect to see clusters either only in New York Neighborhoods or in Toronto neighborhoods. If the local culture is the stronger one, then we may find neighborhoods of the same cluster in both New York and Toronto. The results will be discussed in the next section.

Results

The resulting clusters cover the Boroughs as follows

- Cluster 0 covers 119 neighborhoods from New York and 34 neighborhoods from Toronto
- Cluster 1 covers 25 neighborhoods from New York
- Cluster 2 covers 1 neighborhood from New York
- Cluster 3 covers 7 neighborhoods from New York
- Cluster 4 covers 145 neighborhoods from New York and 2 neighborhoods from Toronto
- Cluster 5 covers 4 neighborhoods from New York and 2 neighborhoods from Toronto
- Cluster 6 covers 2 neighborhoods from New York
- Cluster 7 covers 1 neighborhood from Toronto

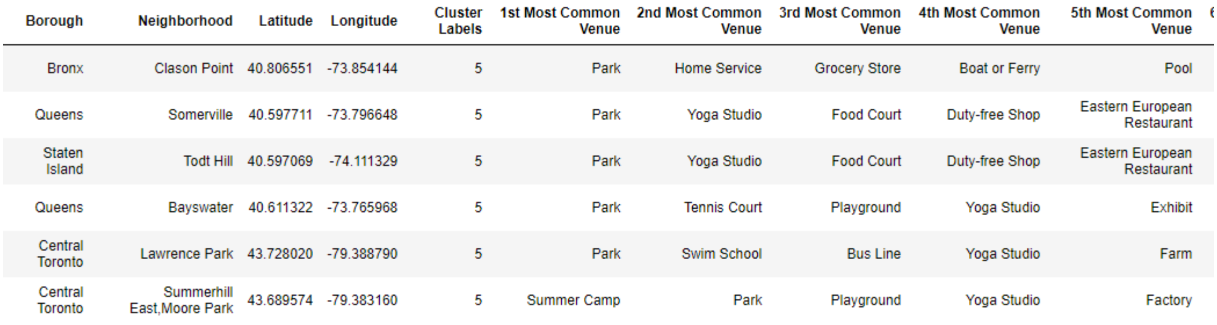
We can see that there are two clusters which cover most of the neighborhoods. Cluster 0 (red dots) appears in both Toronto and New York. This might be a "North American urban culture" Cluster. The second big cluster (number 4, turquoise dots) includes also a wide range of neighborhoods but nearly only appears in New York. This might represent an urban US Cluster. It can further be seen, that New York is by far more diverse than Toronto. In New York it is interesting to note, that the two dominant clusters (red and blue) are split locally. You can see that on the map.

Discussion

In the results section I called cluster 4 an urban US Cluster. In order to test this assumption it would be a good idea to further analyze other American cities, e.g. LA or San Francisco. If the cluster is a typical urban American Cluster we would find it in these cities as well. Otherwise it might be a special New York culture.

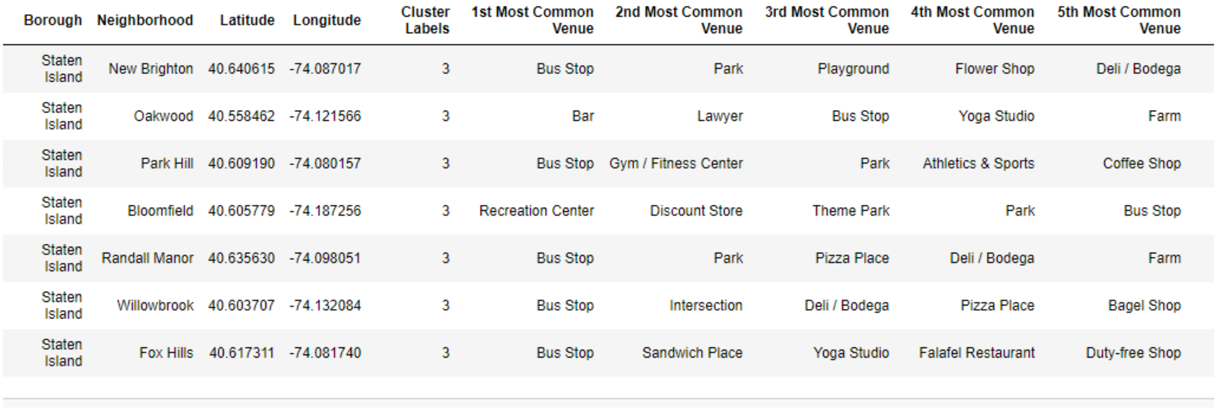
Apart from cluster 0 and 4, the others are much smaller. They seem to be somehow special. Let's look at the very special clusters and check whether we can find some interesting facts

Cluster 5 is small but appears both in Toronto an New York. It contains neighborhoods which most dominant venues are parks



Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Bronx	Clason Point	40.806551	-73.854144	5	Park	Home Service	Grocery Store	Boat or Ferry	Pool
Queens	Somerville	40.597711	-73.796648	5	Park	Yoga Studio	Food Court	Duty-free Shop	Eastern European Restaurant
Staten Island	Todt Hill	40.597069	-74.111329	5	Park	Yoga Studio	Food Court	Duty-free Shop	Eastern European Restaurant
Queens	Bayswater	40.611322	-73.765968	5	Park	Tennis Court	Playground	Yoga Studio	Exhibit
Central Toronto	Lawrence Park	43.728020	-79.388790	5	Park	Swim School	Bus Line	Yoga Studio	Farm
Central Toronto	Summerhill East,Moore Park	43.689574	-79.383160	5	Summer Camp	Park	Playground	Yoga Studio	Factory

Cluster 3 covers 7 neighborhoods but only exists in New York. They are all in Staten Island.



Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Staten Island	New Brighton	40.640615	-74.087017	3	Bus Stop	Park	Playground	Flower Shop	Deli / Bodega
Staten Island	Oakwood	40.558462	-74.121566	3	Bar	Lawyer	Bus Stop	Yoga Studio	Farm
Staten Island	Park Hill	40.609190	-74.080157	3	Bus Stop	Gym / Fitness Center	Park	Athletics & Sports	Coffee Shop
Staten Island	Bloomfield	40.605779	-74.187256	3	Recreation Center	Discount Store	Theme Park	Park	Bus Stop
Staten Island	Randall Manor	40.635630	-74.098051	3	Bus Stop	Park	Pizza Place	Deli / Bodega	Farm
Staten Island	Willowbrook	40.603707	-74.132084	3	Bus Stop	Intersection	Deli / Bodega	Pizza Place	Bagel Shop
Staten Island	Fox Hills	40.617311	-74.081740	3	Bus Stop	Sandwich Place	Yoga Studio	Falafel Restaurant	Duty-free Shop

Conclusion

We can conclude that a type of culture exists, that is not bond to national borders. We find this neighborhoods (red clustered dot) both in Toronto and New York. Apart from that both cities have neighborhoods which are very special and have no common part in the other city. This is for example Staten Island which seems to have strong local subculture.

Extending the analysis may help to find more interesting results. I recommend analyzing at least two cities for each country and further analyzing a country that differs from Canada and US. E.g. adding a central European country and an Asian one.