Study of the Neighborhoods of Toronto and New York

Introduction

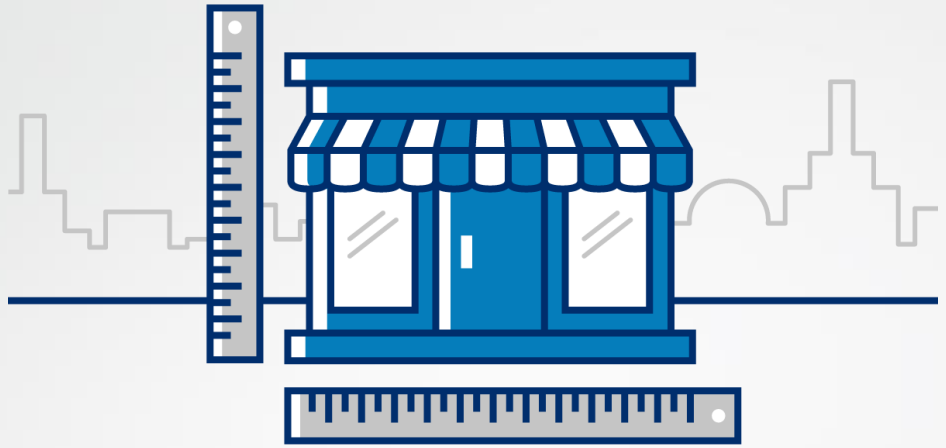


Image Source: https://www.sba.gov/build/71bd57d80e04f91d53641835ce6d7acc.png

- Objective: Extract useful information of differences between cities and opportunities, for people looking to open a business or even for instance, wants to move to a similar place where he can make the most of it.
- When someone is looking to invest in a profitable business, first must look for opportunities, needs that have not been covered yet and discover the most successful cases. Using analytics is possible to extract this information.
- The neighborhoods for the cities of Toronto and New York are compared and is decided if is better to study Toronto and New York, or Manhattan instead.
- For this purpose, I'm using the machine learning clustering algorithm of k means. This allow to find groups of neighborhoods with similar features and with some other methods then find the optimal number of clusters.

Data Acquisition

Image Source: https://www.mindseyetech.net/database-management

- Data Sources:
  - Foursquare API to get the most common venues categories for every neighborhoods
  - Postal code data from Wikipedia for Toronto
  - Geospatial coordinates of latitude and longitude
  - New York University spatial data repository published by the New York (City). Department of City Planning.

Feature selection

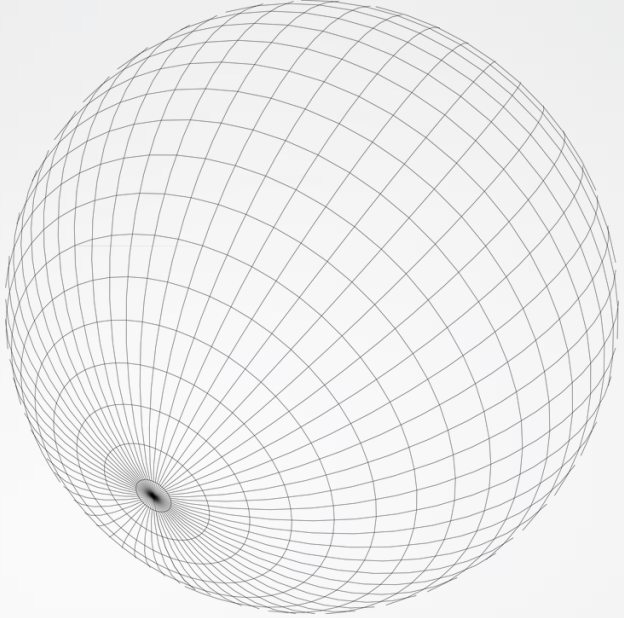
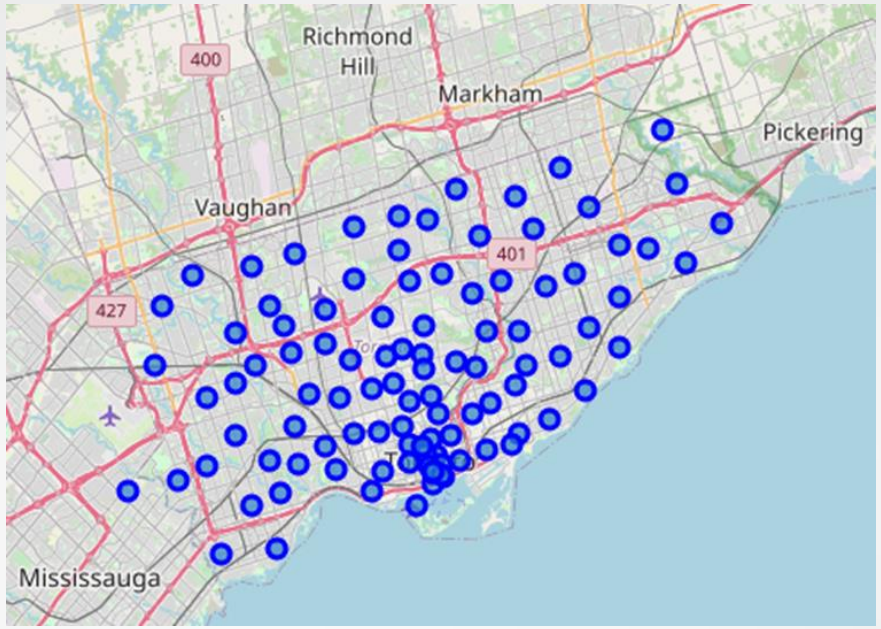


Image Source: https://cdn.pixabay.com/photo/2012/04/16/11/33/globe-35588_960_720.png

- New York dataset: Stored in a .json file with all the relevant data inside the feature key, containing the list of neighborhoods.
- Toronto: Borough, Neighborhood, Latitude and Longitude are used.
- Foursquare API: The top 100 venues for every neighborhood within a radius of 500 meters are added. Finally the 10 most common venues are selected.

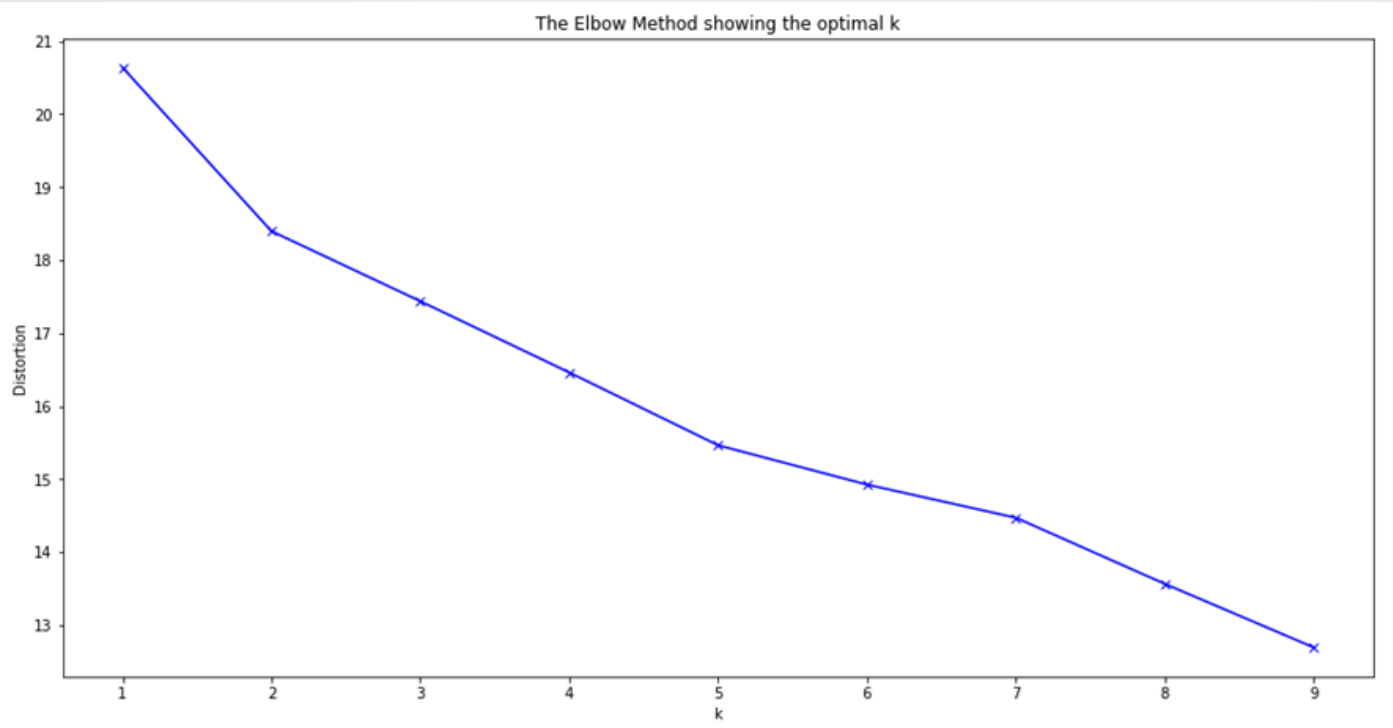
FOURSQUARE

Data Analysis



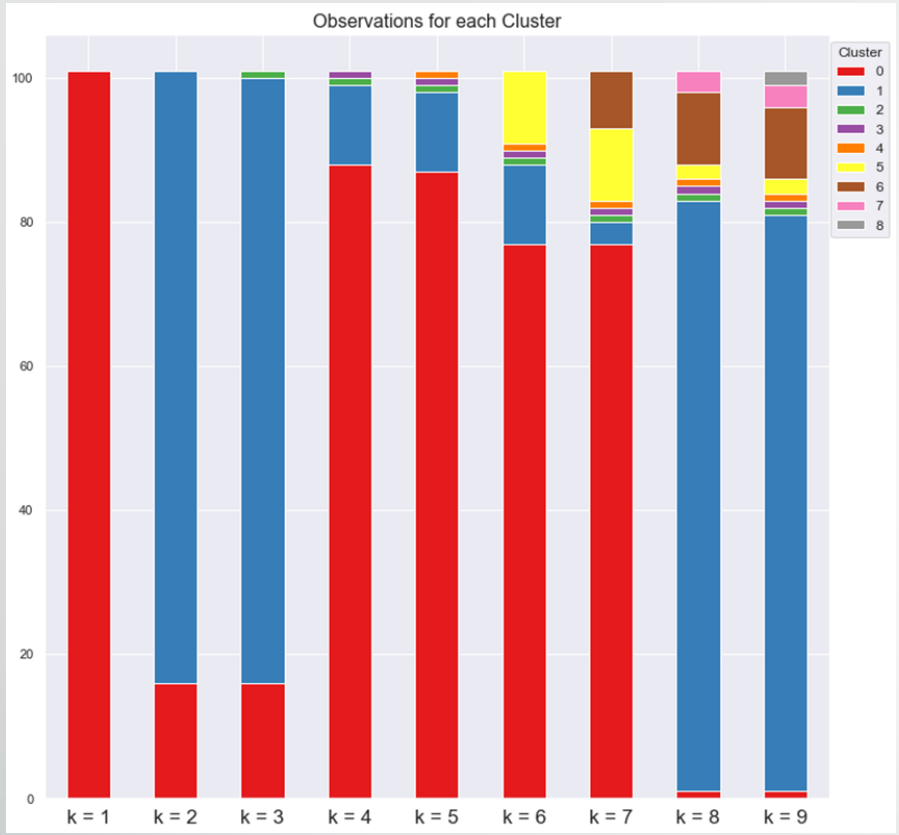
- To perform the data analysis the k-means algorithm is used to create groups of similar neighborhoods and then categorize them.
- The most common venues in each neighborhood are grouped in a cluster.
- Starting with the neighborhoods from Toronto, the coordinates are plotted into a map pointing at their locations. This map would be useful later to displaying the clusters and distribution of datapoints.

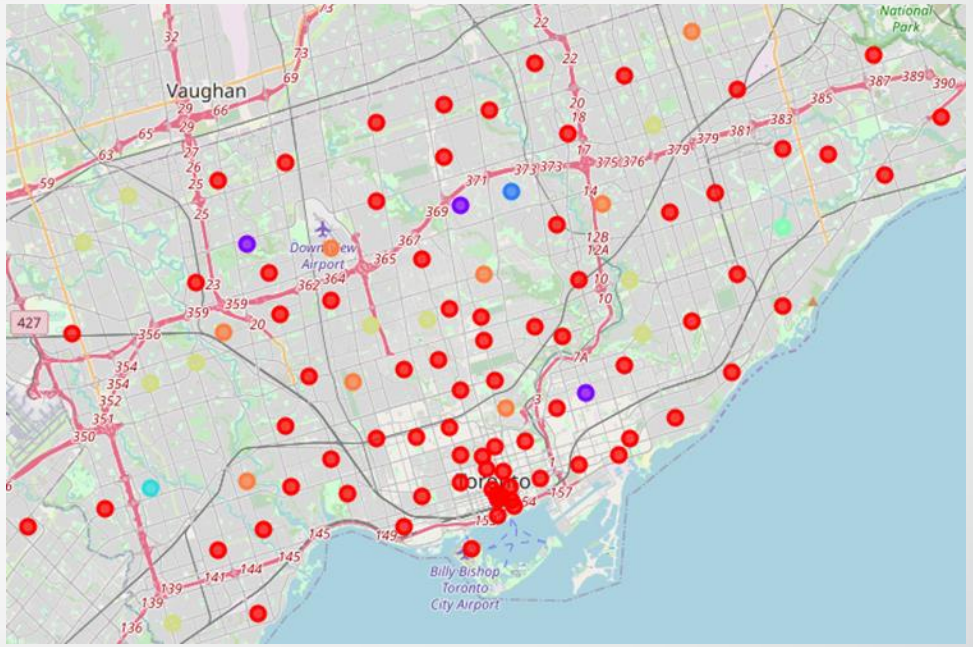
K means



- To find the optimal amount of K the model is tested with 1 to 10 clusters and the results are plotted using the elbow method to find the best K. As clearly seen, there is not definitive answer for this problem.
- The same happens for New York. An alternative method is used in his place.

The City of Toronto

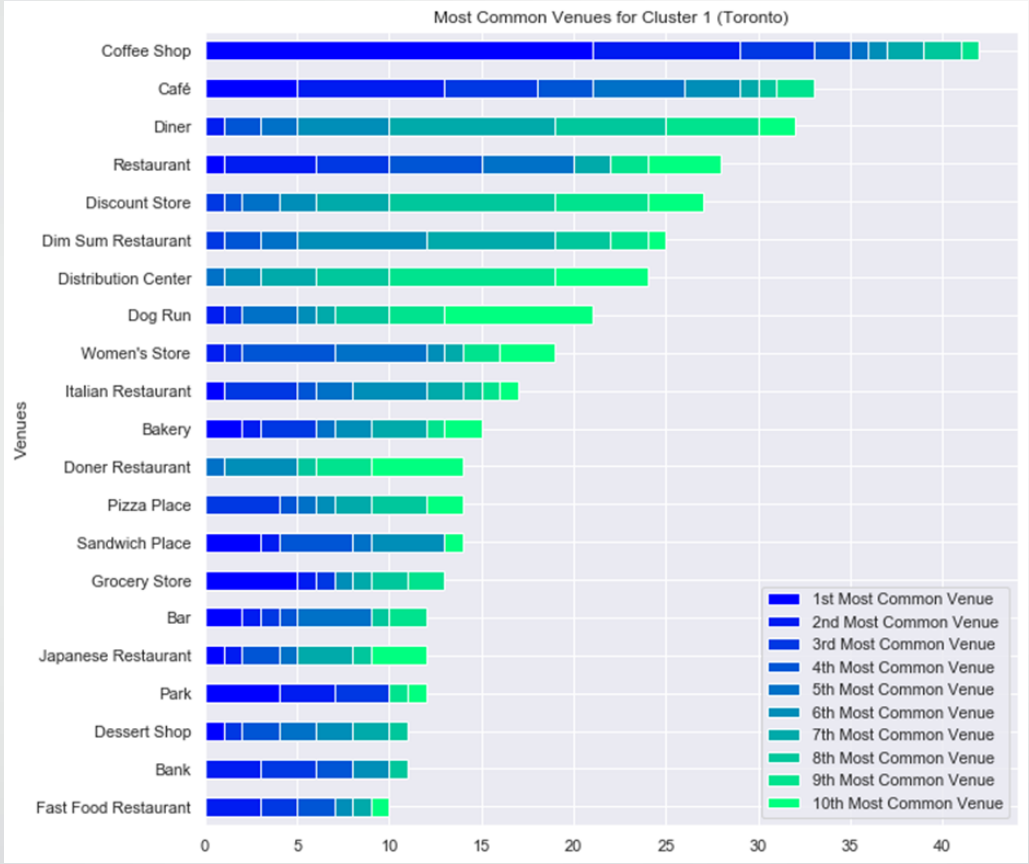




- With 2 – 5 centroids, only 2 clusters are created, the smaller ones having only one data point. With 6 and 7 centroids, I have 3 and 4 clusters.
- Using 7 centroids is possible to find a good balance between the amount of clusters and data points in them.

Data points for each Cluster

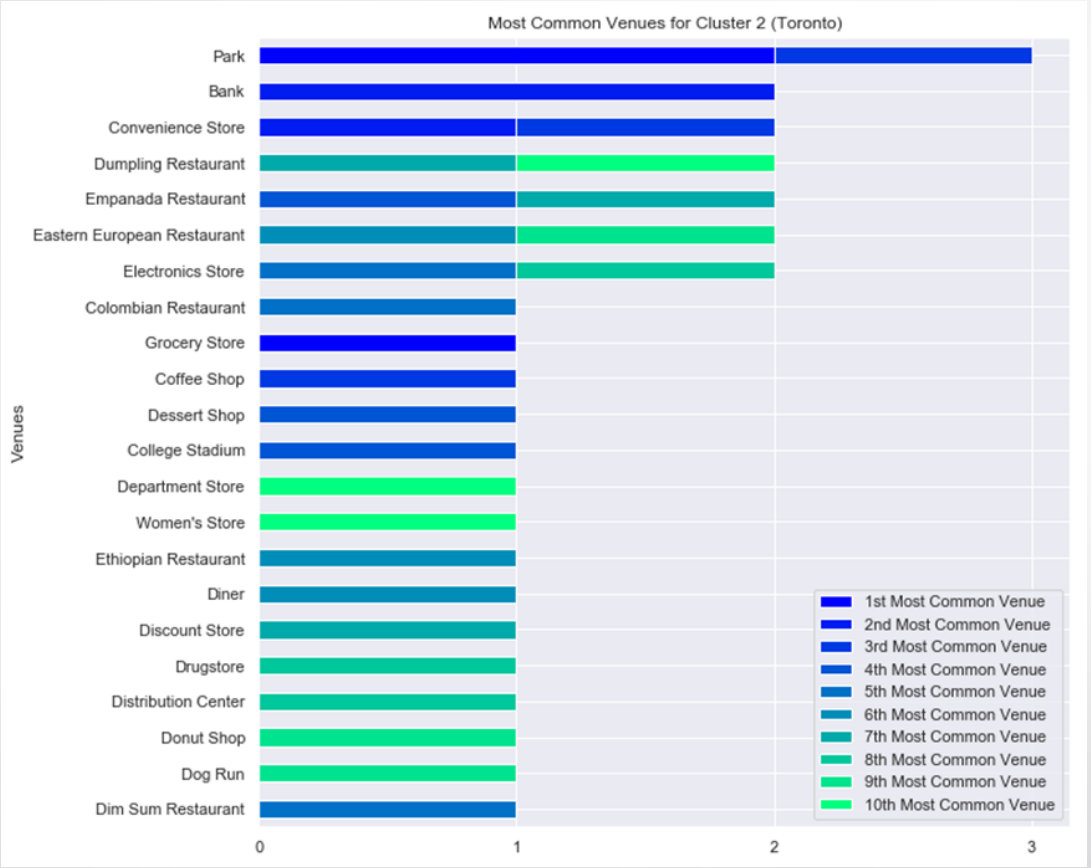
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Data Points	76.24%	2.97%	0.99%	0.99%	0.99%	9.90%	7.92%



- Cluster 1 is dominated by Cafes (Coffee Shops and Cafés), Restaurants (and Diners), including from Chinese (Dim Sum), Italian, Turkish (Doner) and of Japanese origin.
- Strong presence of Discount Stores, Distribution Centers and Women's Stores.
- Dog Runs are of important to note, being very common for all the clusters studied, the same happens with Dim Sum restaurants.

Data points for each Cluster

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Data Points	76.24%	2.97%	0.99%	0.99%	0.99%	9.90%	7.92%



- Cluster 2 is dominated by Parks, Banks, Convenience and Electronic stores.
- Empanadas (Hispanic), Dumpling (Chinese), Eastern European restaurants, as well Colombian, Ethiopian and Dim Sum (Chinese) restaurants.

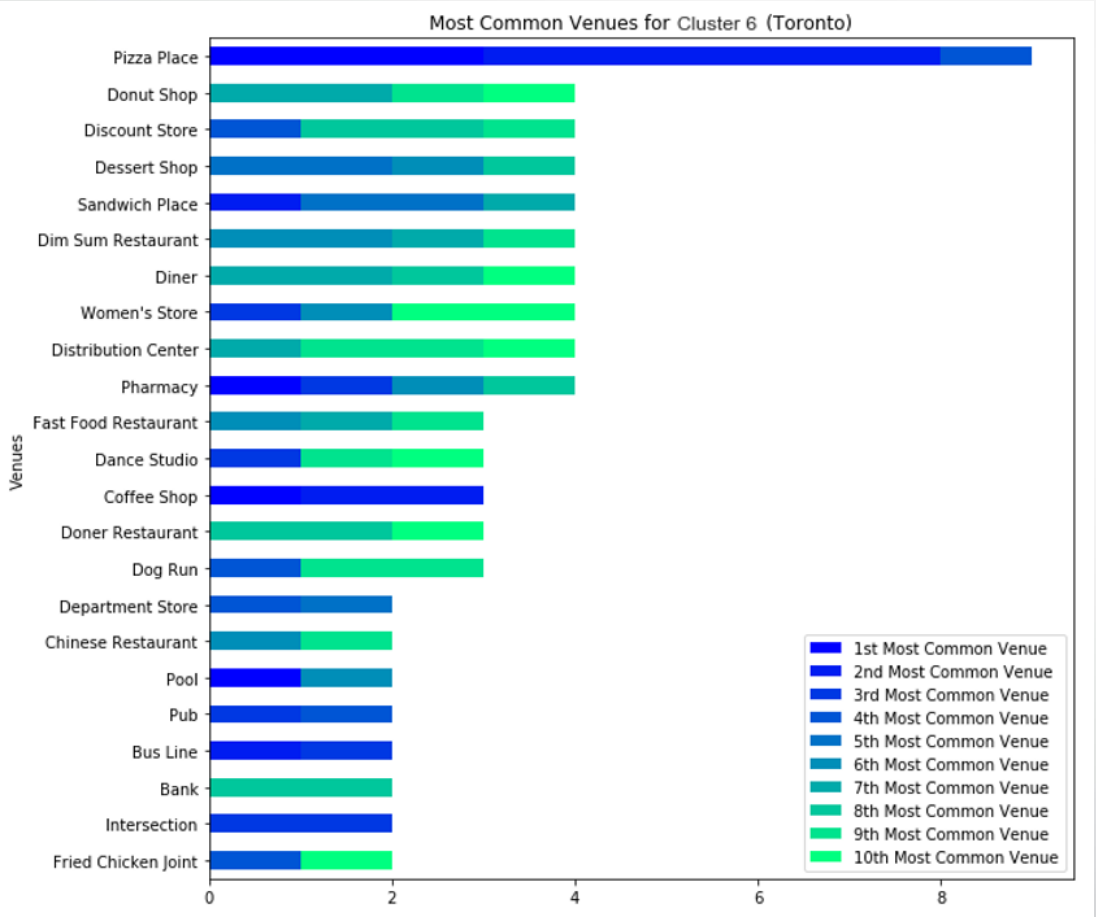
Data points for each Cluster

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Data Points	76.24%	2.97%	0.99%	0.99%	0.99%	9.90%	7.92%

- Clusters 3, 4 and 5 are very similar, each one with 1% or one data point.
- Cluster 3 is dominated by Cafeterias, Cluster 4 Golf Courses and Cluster 5 Playgrounds.
- Then are followed by Women's Stores, Doner and Dim Sum restaurants, Diner and Discount Stores.

Data points for each Cluster

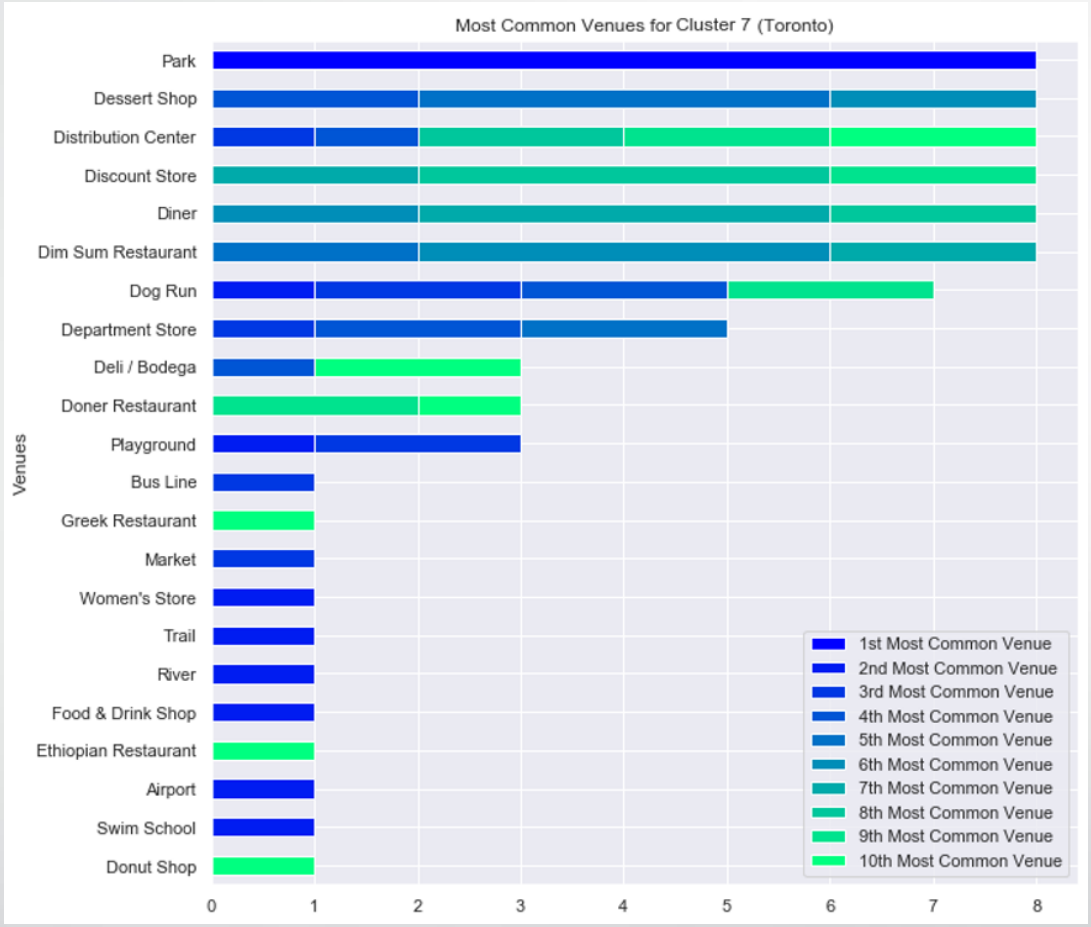
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Data Points	76.24%	2.97%	0.99%	0.99%	0.99%	9.90%	7.92%



- Cluster 6 unlike previous clusters, the most common venue is much more prominent: Pizza Places.
- Followed by Pharmacies, Women's Stores, Discount stores and Food: Dim Sum (Chinese) Restaurants, Diner, Sandwich, Donuts, Dessert.

Data points for each Cluster

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Data Points	76.24%	2.97%	0.99%	0.99%	0.99%	9.90%	7.92%



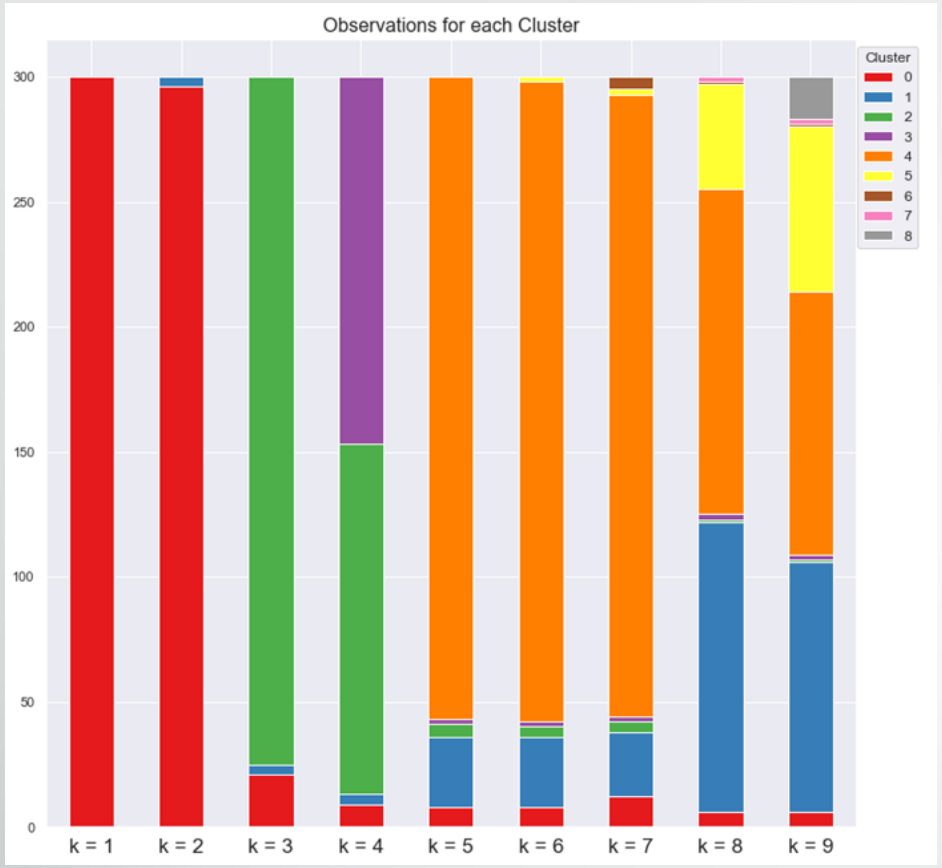
- Cluster 7 is dominated by Parks and Dessert Shops.
- Then Distribution Centers, Discount Stores, Diner, Dog Runs, Dim Sum (Chinese) Restaurants and Department Stores.

All of New York or Manhattan

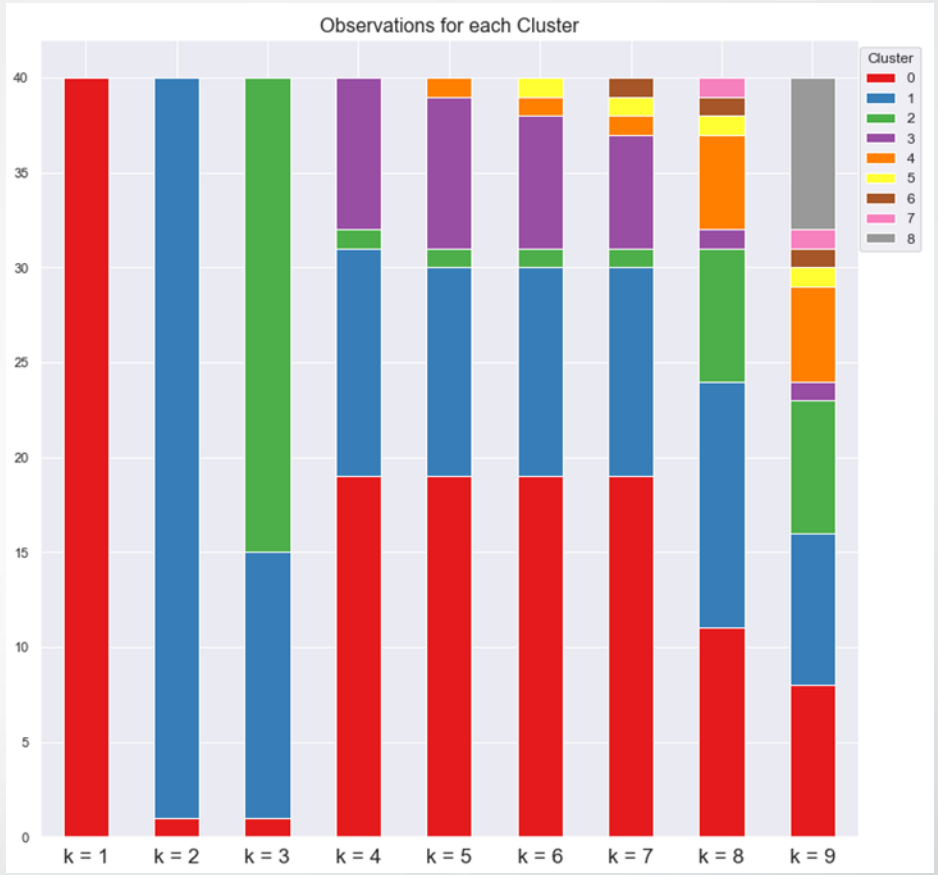
- When creating clusters for New York and Manhattan, I concluded that Manhattan and Toronto are more suited to perform comparatives.
- Manhattan serves as New York economic and administrative center, cultural identifier.
- Described as the cultural, financial, media, and entertainment capital of the world.
- Similarly Toronto is an international center of business, finance, arts, and culture.
- Its home to the Toronto Stock Exchange, the headquarters of Canada's five largest banks.

(Source: Wikipedia)

New York

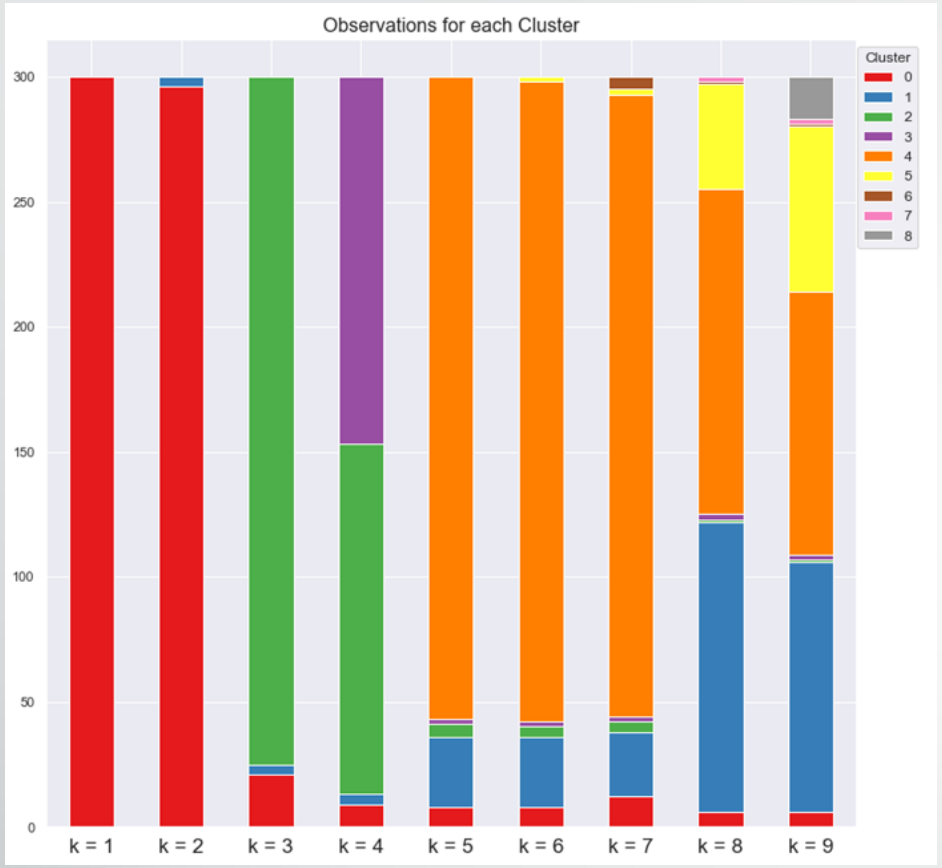


Manhattan

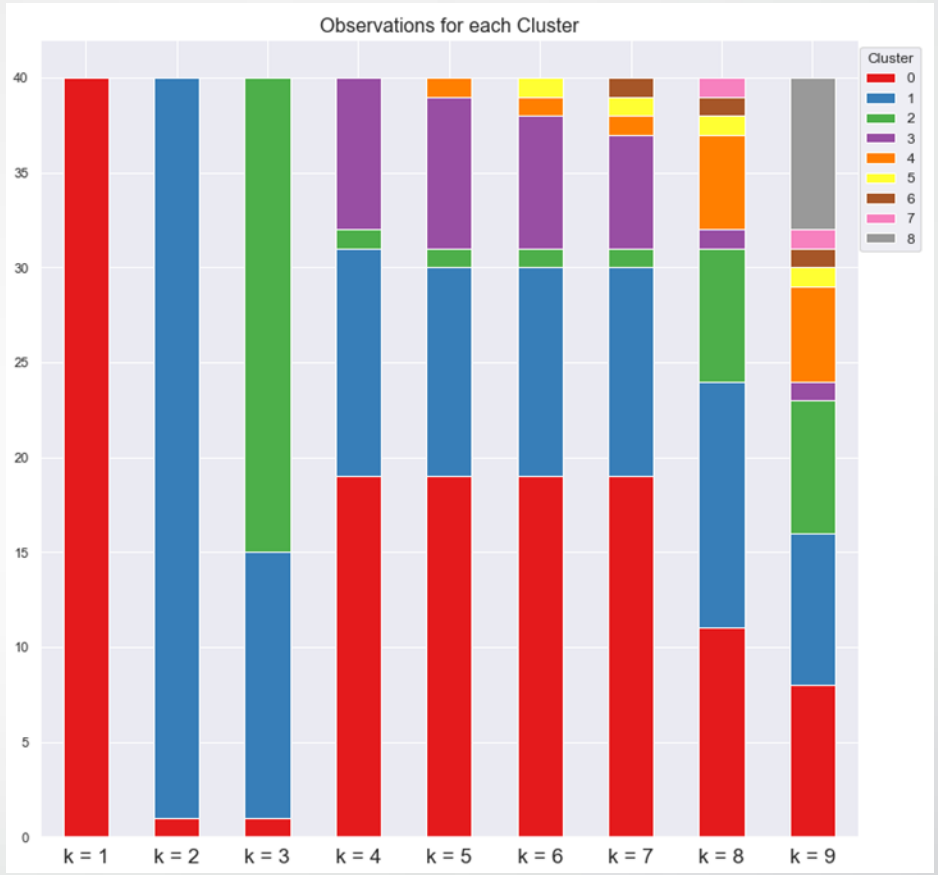


- With Toronto 7 centroids where used, but only was possible to find 4 clusters, the 3 other clusters having 1 data point or neighborhood. New York have 300 neighborhoods vs 100 of Toronto, having enough data points to create clusters with no problems.
- Instead Manhattan have 40 Neighborhoods and like with Toronto, is harder to find the right number of centroids. Was possible to distinguish 2 clusters with 2 – 3 centroids, and 3 clusters with 4 – 7 centroids. With 8 and 9 centroids more clusters where created. I chose 4 centroids, finding a good balance between the amount of clusters and data points in them.

New York



Manhattan

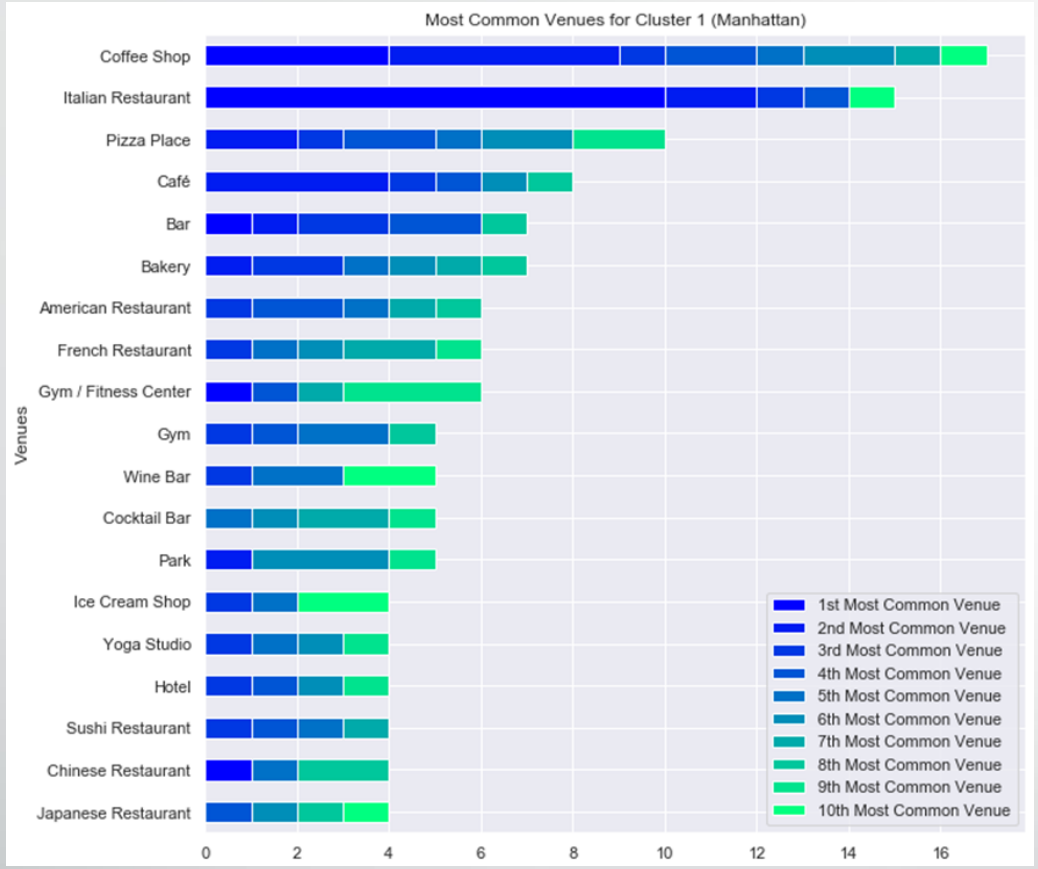


- With Toronto 7 centroids where used, but only was possible to find 4 clusters, the 3 other clusters having 1 data point or neighborhood. New York have 300 neighborhoods vs 100 of Toronto, having enough data points to create clusters with no problems.
- Instead Manhattan have 40 Neighborhoods and like with Toronto, is harder to find the right number of centroids. Was possible to distinguish 2 clusters with 2 – 3 centroids, and 3 clusters with 4 – 7 centroids. With 8 and 9 centroids more clusters where created. I chose 4 centroids, finding a good balance between the amount of clusters and data points in them.

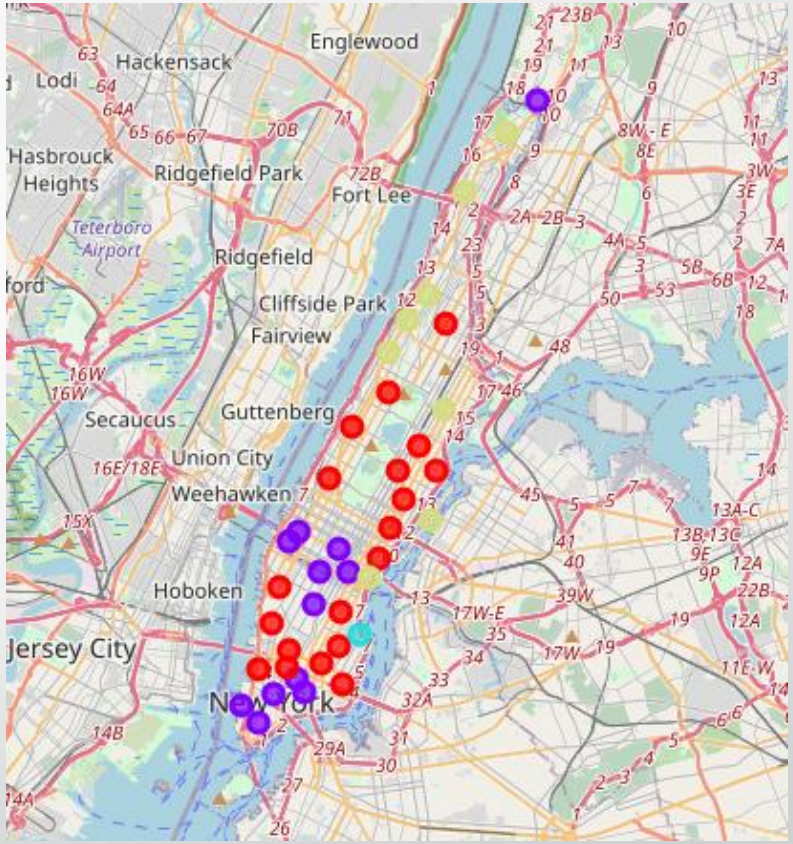
The analysis of New York was performed in the notebook. Instead I'm focusing in Toronto and Manhattan here.

Data points for each Cluster

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Data Points	47.5%	30.0%	2.5%	20.0%



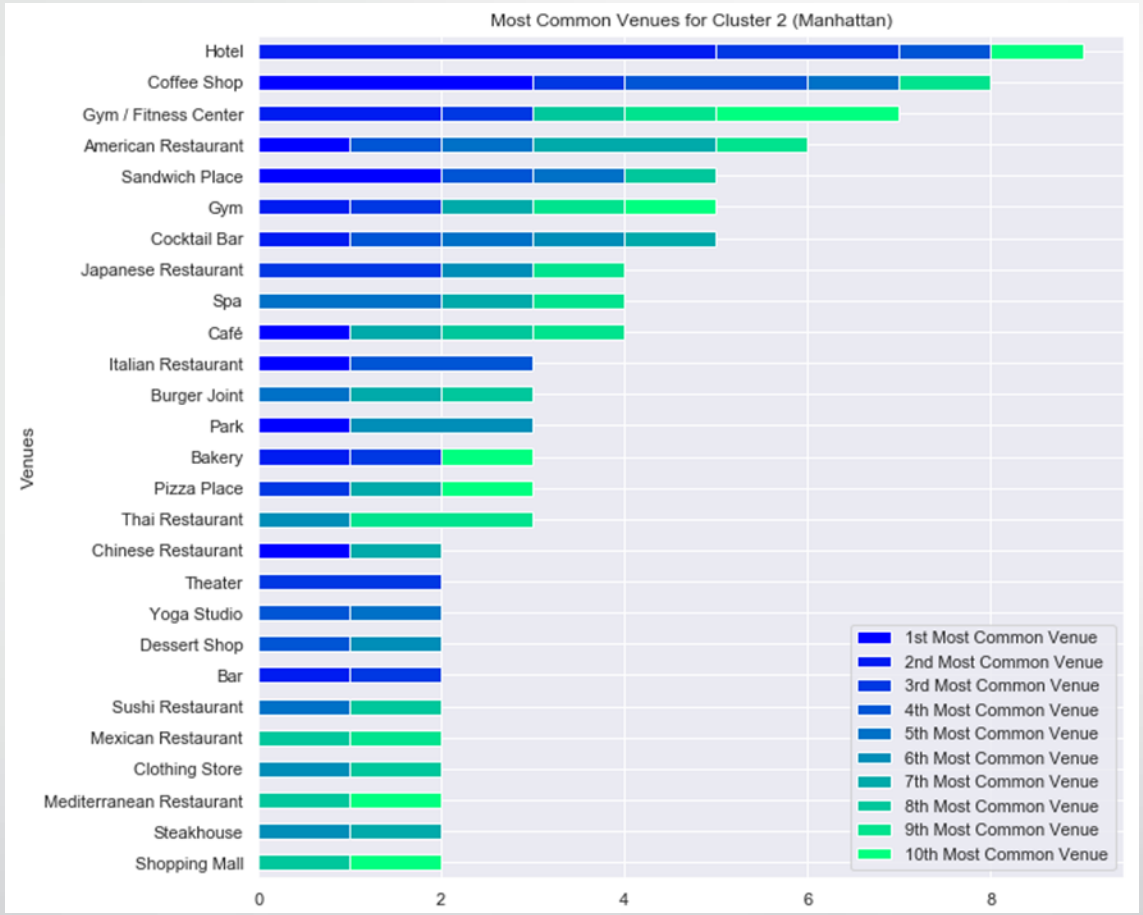
Manhattan Clusters



- Cluster 1 is dominated by Coffee Shops and Italian Restaurants. Followed by Pizza places, Cafes, Bars, Bakery, Gyms and Yoga.
- Restaurants: American, French, Japanese and Chinese.

Data points for each Cluster

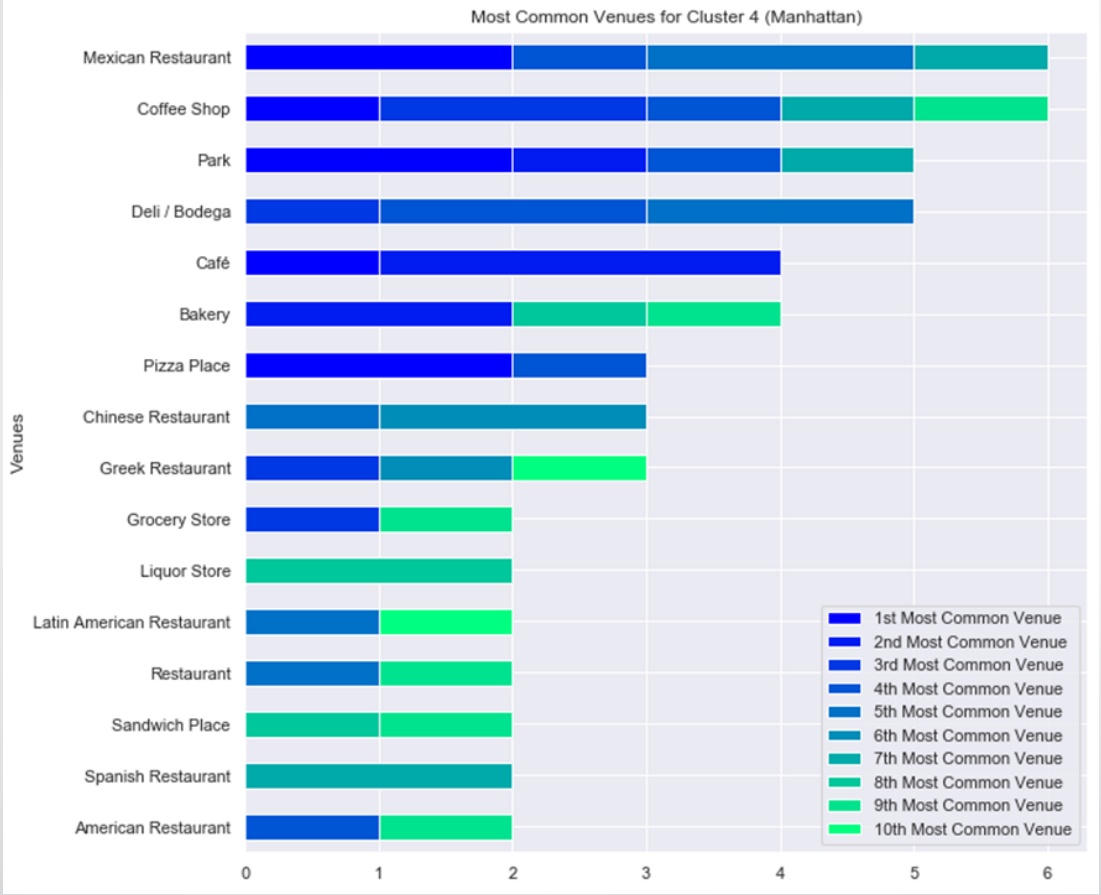
	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Data Points	47.5%	30.0%	2.5%	20.0%



- Cluster 2 is dominated by Hotels, Coffees, Gyms, Cocktail Bars, and Spa.
- Restaurants: American, Japanese, Italian and Thai. Then Chinese, Mexican and Mediterranean.

Data points for each Cluster

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Data Points	47.5%	30.0%	2.5%	20.0%



- Cluster 4 is dominated by Mexican Restaurants, Cafes, Parks, Bodegas, Bakery, Pizza places.
- Restaurants: Chinese, Greek, Latin American, Spanish and American.
- Cluster 3 consists of only one neighborhood, with Boats or Ferry the most common venue, followed by Parks, Gas Stations, Baseball Fields and Harbor/Marina.

Conclusions

- Chinese Restaurants, Coffee shops, Dog Runs and Women Stores are very characteristic of Toronto.
- For Manhattan Coffee Shops, Gyms, Pizzas and Bakery.
- This information denotes preferences, the cultural differences and similarities between these two great financial centers of their respective nations.
- Distribution center and discount stores have great success in Toronto given the amount of them.
- In Toronto opening stores directed to dog owners can be a big opportunity.
- Coffee is very good option in both Manhattan and Toronto.
- The number of venues for this business makes them great opportunities to invest, especially in cluster with lower presence, with the potential to be easier to compete and dominate the market.