DataScience Capstone Project

Finding Similar Neighborhoods between New York and Toronto

Table of contents

1. Introduction : Business Problem

Living in a big city provides a number of options in terms of neighborhood choices. When selecting a neighborhood to live in a person might have a number of considerations, including rental costs, housing prices, transportation , walkability, restaurants in neighborhood, Gyms, running routes etc.

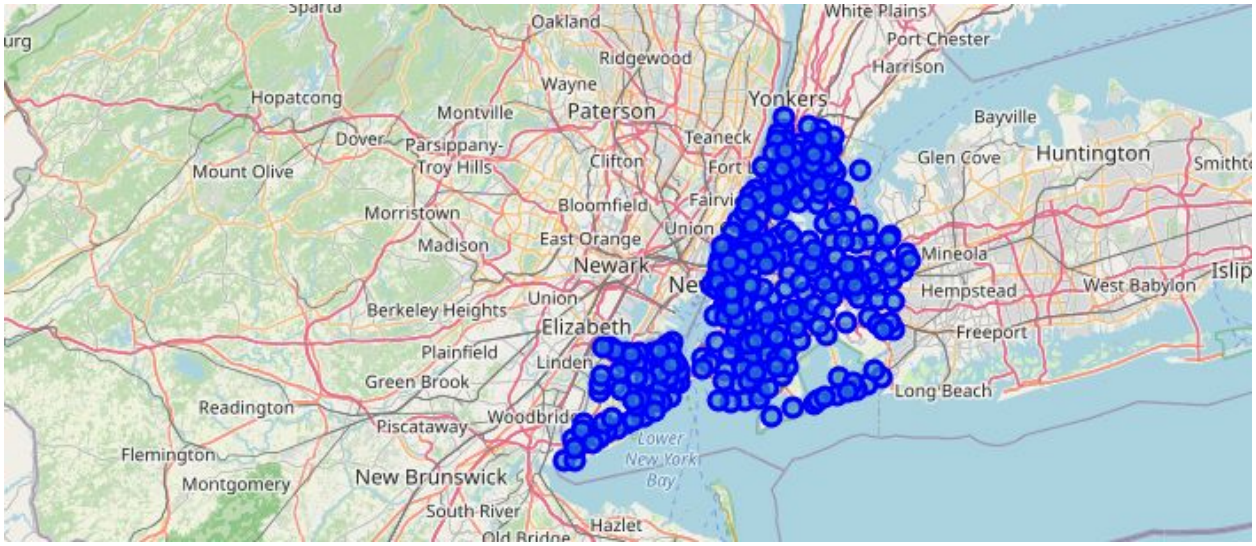
If someone is relocating from one city to another while they might have a good idea of the neighborhoods in one city they might be unfamiliar with the neighborhoods in another city. In this case study we are going to look at neighborhoods in New York and try to find similar neighborhoods in Toronto as an example.

While the intention for this Capstone project is only to do this for two sample cities, this can easily be extended to add additional cities for comparison

2. Data

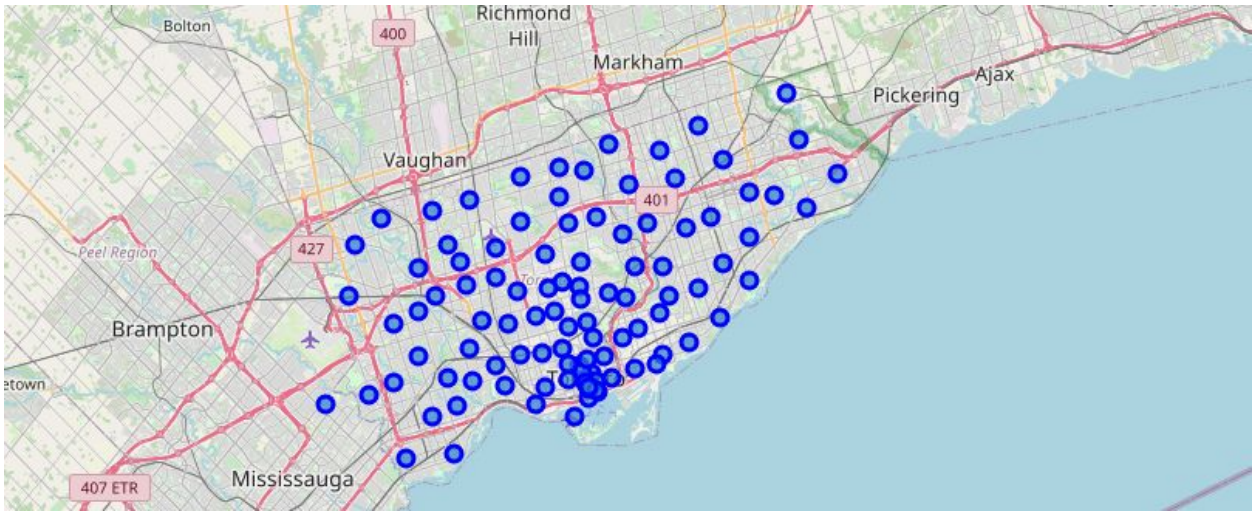
To group neighborhoods into different categories and find comparable neighborhoods from the other city, we can use the data from foursquare API to find venues type and number of venues in each neighborhood.
We first gather neighbourhood data for both Toronto and and New York for along with zip code and latitude and longitude data. We then plot this on a map to get an idea of the spread.
Below is the map of all the neighborhoods we'll be looking at for New york. We will include all five boroughs for our project.

New York Neighborhoods



Similarly we gather neighborhood data for all toronto neighborhoods including latitude and longitude and have plotted it below on a map.

Toronto Neighborhoods



We then gather venue data from foursquare API for each of these neighbourhoods, we are limiting it to a maximum of 100 venues for each neighborhood in a 500 m radius for simplicity.

3. Exploratory Data Analysis

Looking at the total number of neighborhoods we find we have 306 different neighborhoods in new york and 103 unique neighborhoods in Toronto we will be looking at.

After getting the venue data from foursquare we further inspect it and found that the number of unique venues are very high. FourSquare uses a Category hierarchy with all venues falling into one of the following top level categories. Some of the hierarchies are upto 4 levels deep but they all roll up into one of the following:

1. Arts & Entertainment
2. College & University
3. Event
4. Food
5. Nightlife Spot
6. Outdoors & Recreation
7. Professional & Other Places
8. Residences
9. Shop & Service
10. Travel & Transport

Using these top level Categories will better help us in segregating neighborhoods which might otherwise be dominated by one category.

Foursquare provides their Category Hierarchy at the following URL: https://developer.foursquare.com/docs/build-with-foursquare/categories/ We used Beautiful soup to do some web scraping and create a Panda Data frame with the category structure.
After converting all venues into one of the categories we found the following distribution across all neighborhoods:

Venue Type	Number of Venues
Arts & Entertainment	446
College & University	16
Food	6389
Nightlife Spot	768
Outdoors & Recreation	1172
Professional & Other Places	100
Residence	10
Shop & Service	2717
Travel & Transport	471
Events	0

4. Methodology

We run K-means clustering on the neighbourhoods across the two cities and classify them into six different clusters. Providing us with the most similar neighbourhoods based on the make-up of the type of venues in each neighborhood.

The six neighborhood clusters end up with the following distribution based on the venue type make up in these neighborhoods

Cluster Label	Number of Neighborhoods
Cluster 1	156
Cluster 2	21
Cluster 3	14
Cluster 4	14
Cluster 5	51
Cluster 6	148

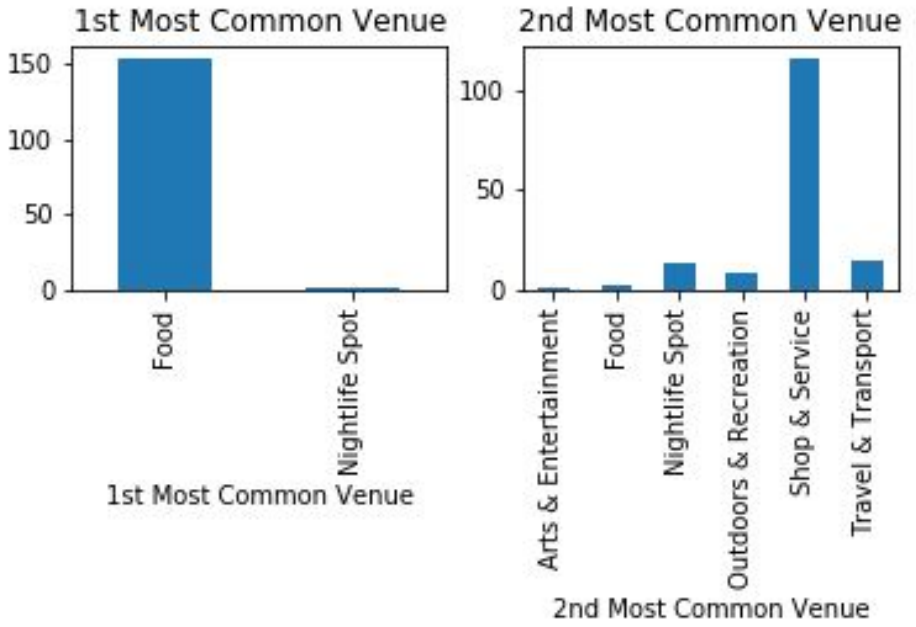
We selected six clusters based on trying to find unique Neighborhood properties. We experimented with different cluster sizes from three thru seven and found six clusters to be the ideal number.

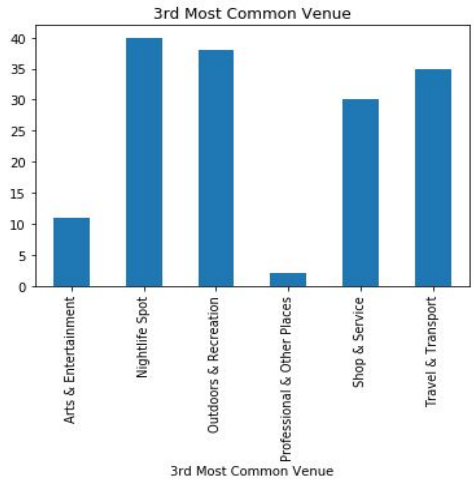
5. Analysis

Let's look at the individual clusters in a bit more detail and see which type of venues are the top 3 most common for each Cluster

Cluster 1

Total Number of Neighborhoods: 156. Based on the first and second most common venue types in these neighborhoods we can classify Cluster 1 as **Shopping districts with good food options and some nightlife.**

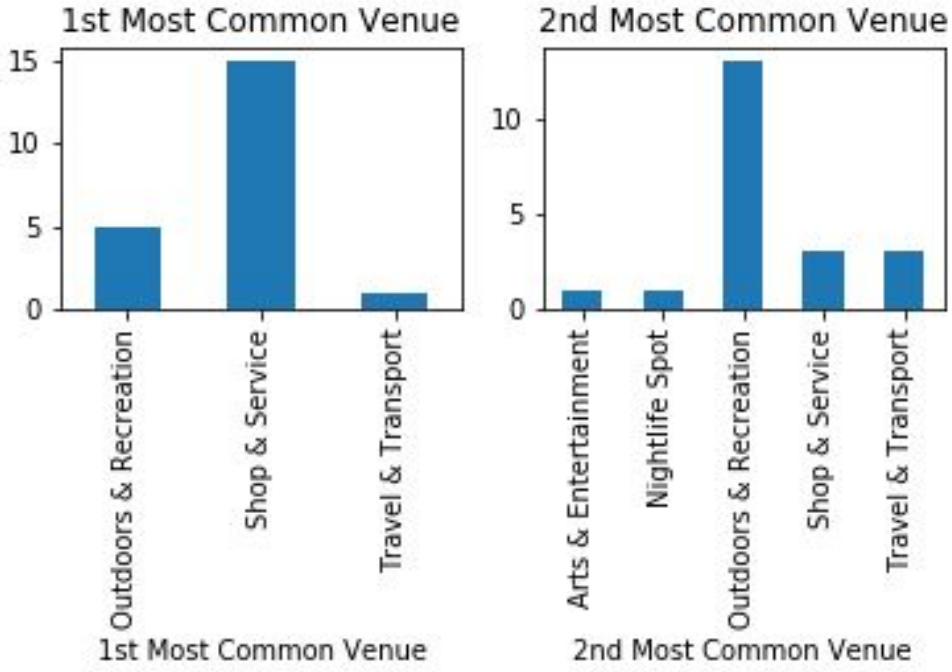


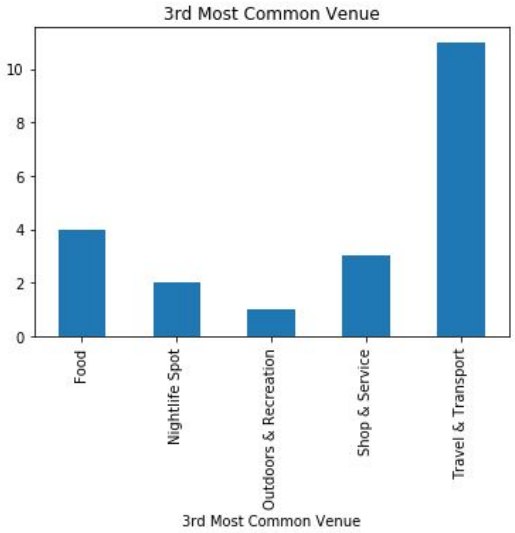


Cluster 2

Total Number of Neighborhoods: 21

We can classify Cluster 2 Neighborhoods ar primarily have lot of outdoor and recreation with some shopping & service.

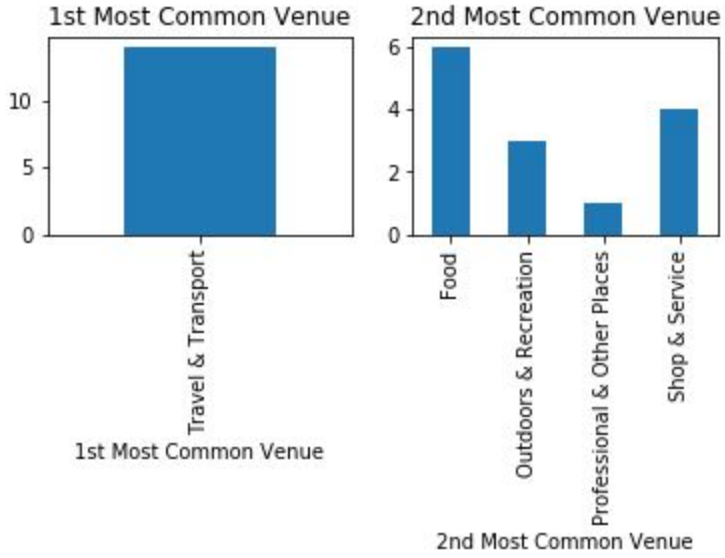


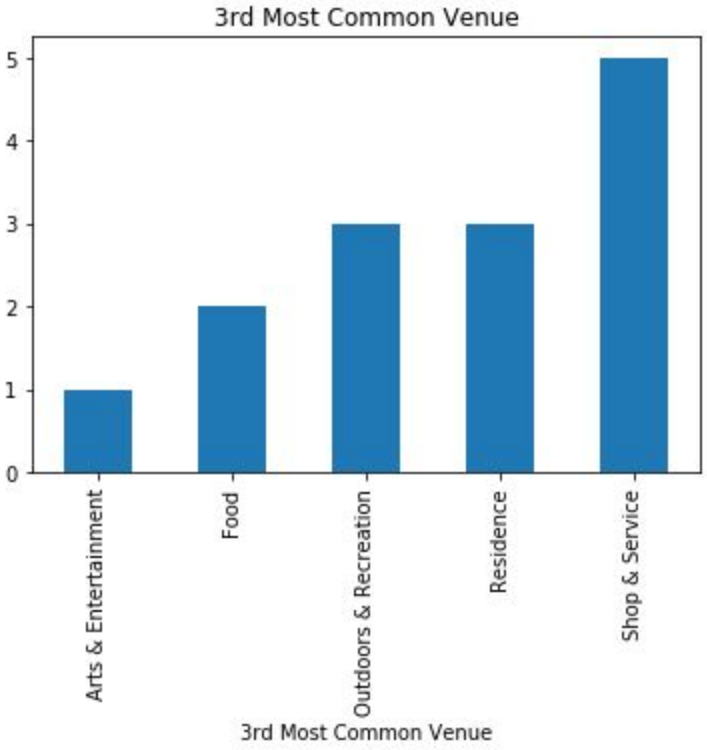


Cluster 3

Total Number of Neighborhoods: 14

Cluster 3 Neighborhoods can be classified as Travel and Transportation Neighborhoods.

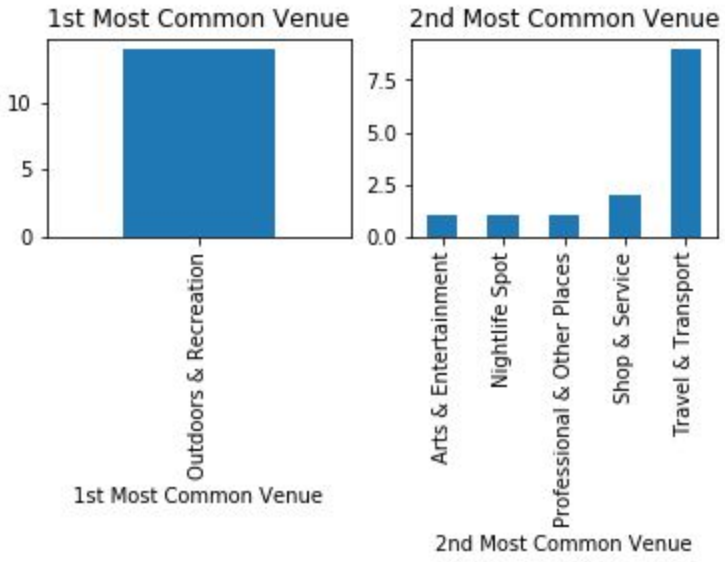


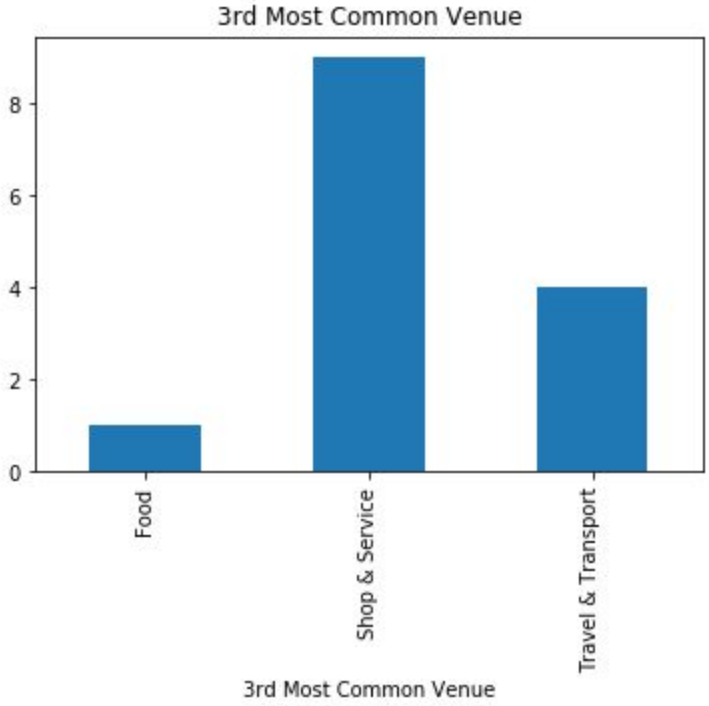


Cluster 4

Total Number of Neighborhoods: 14

Cluster 4 Neighborhoods can be classified as Outdoors & Recreation with travel and transport. This Mix would indicate more Isolated neighborhoods with parks.

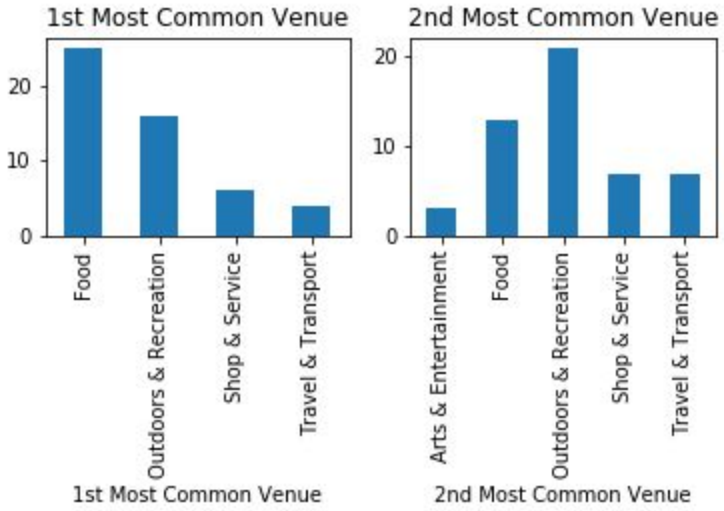


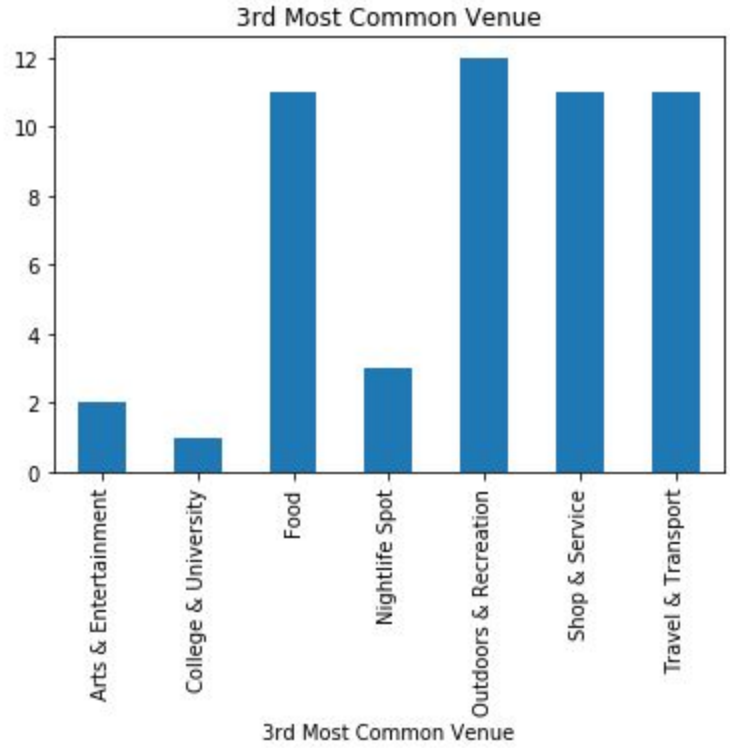


Cluster 5

Total Number of Neighborhoods: 51

Cluster5 seem to be neighborhoods with a good mix of all venues with a number of outdoor and recreation areas.

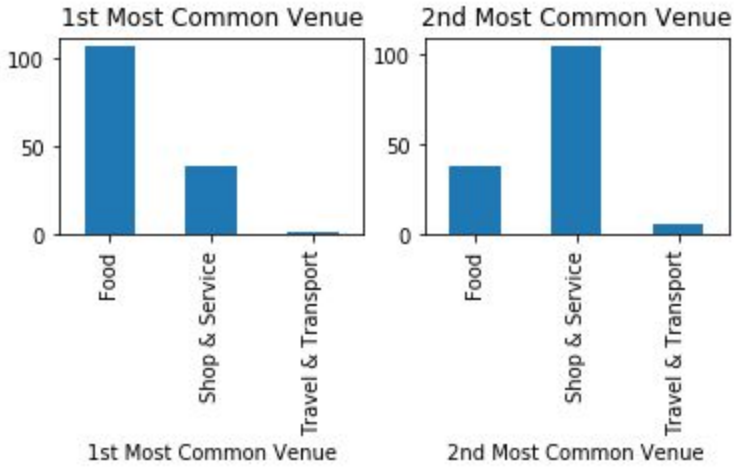


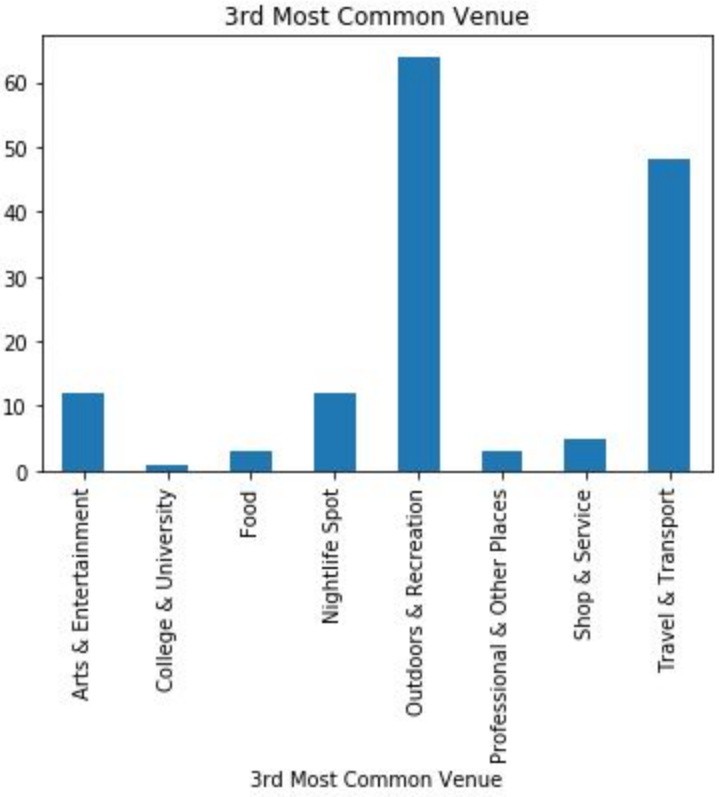


Cluster 6

Total Number of Neighborhoods: 148

These Neighborhoods would seem to have great dining and shopping venues.





6. Results and Discussion

Based on our Clustering we have come up with this set of six neighborhoods.

Cluster Number	Neighborhood Types
Cluster 1	Primarily Shopping districts with good food options and some nightlife.
Cluster 2	Good Outdoor and recreation with some shopping & service.
Cluster 3	Travel and Transportation
Cluster 4	Outdoors & Recreation with travel and transport. This Mix would indicate more Isolated neighborhoods with parks.
Cluster 5	Good mix of all venues with a number of outdoor and recreation areas.
Cluster 6	Great for dining and shopping venues.

7.Conclusion

We used the number and type of venues to come up with similar neighborhoods between the cities of Toronto and New York. This allows us to get a good idea of which neighborhoods would have a similar feel in a new city. This can be further expanded in the future by adding rental rates, property prices walkscore etc.