A Comparative Analysis of Neighborhood Characteristics between Toronto, CA and New York City, USA

Table of Contents

1.Introduction
2.Data and Methodology
3.Results
4.Discussion and Conclusion
5.Limitations and Perspective
6.Reference

1 Introduction

Both Toronto and New York are two big cities in North America. With the close connection between Canada and the U.S. governments, there are a strong need of population immigration for both cities' dwellers. It is important for those people interested in two places to know what the similarities are as well as the differences between two cities. In this lab, I will use all the knowledge and coding skills I learned from previous courses to conduct a comparative analysis for the targeted cities. The major problems I am interested here includes:

1. **Is the city urban structure of Toronto alike or different from the NYC?**
2. **Which dimensions are those similarities or differences?**

2 Data and Methodology

2.1 Data

The following table shows the details of the datasets used in this analysis. The Toronto data is in spatial format (GeoJSON) and I will use the GeoPandas library to preprocess it while the NYC data is in JSON format which is cleaned by json package. Both datasets will have the same columns at the end, namely Neighborhood Name, Latitude, Longitude, which will be sent to Foursquare API to obtain venues information.

City Name	Year	Total # of Neighborhoods	Data Format	Data Source
Toronto Neighborhood Boundary	2021	140	GeoJSON	https://ckan0.cf.opendata.inter.prod-toronto.ca/dataset/neighbourhood-crime-rates
New York Neighborhood Boundary	NA	303	JSON	https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json

Besides above-mentioned datasets, Foursquare venues data for each neighborhood location within 500m is obtained through the API calls. The data has 9 different layers (show in following table), which will be used to compare two cities and to conduct the K-means clustering analysis.

Urban Element	Foursquare Layer ID
Outdoors & Recreation	4d4b7105d754a06377d81259
Shop & Service	4d4b7105d754a06378d81259
Arts & Entertainment	4d4b7104d754a06370d81259
College & University	4d4b7105d754a06372d81259
Event	4d4b7105d754a06373d81259
Food	4d4b7105d754a06374d81259
Travel & Transport	4d4b7105d754a06379d81259
Nightlife Spot	4d4b7105d754a06376d81259
Professional & Other Places	4d4b7105d754a06375d81259

2.2 Python Library List

The following table shows all the libraries which are used to complete the project.

Library	Function
GeoPandas	manipulation of spatial data
Pandas	For creating and manipulating dataframes
Folium	to visualize the neighborhoods cluster distribution of using interactive leaflet map spatially
Scikit Learn	implementing k-means clustering
JSON	handle JSON files
Requests	handle http requests
Matplotlib	Python Plotting Module

2.3 Methodology

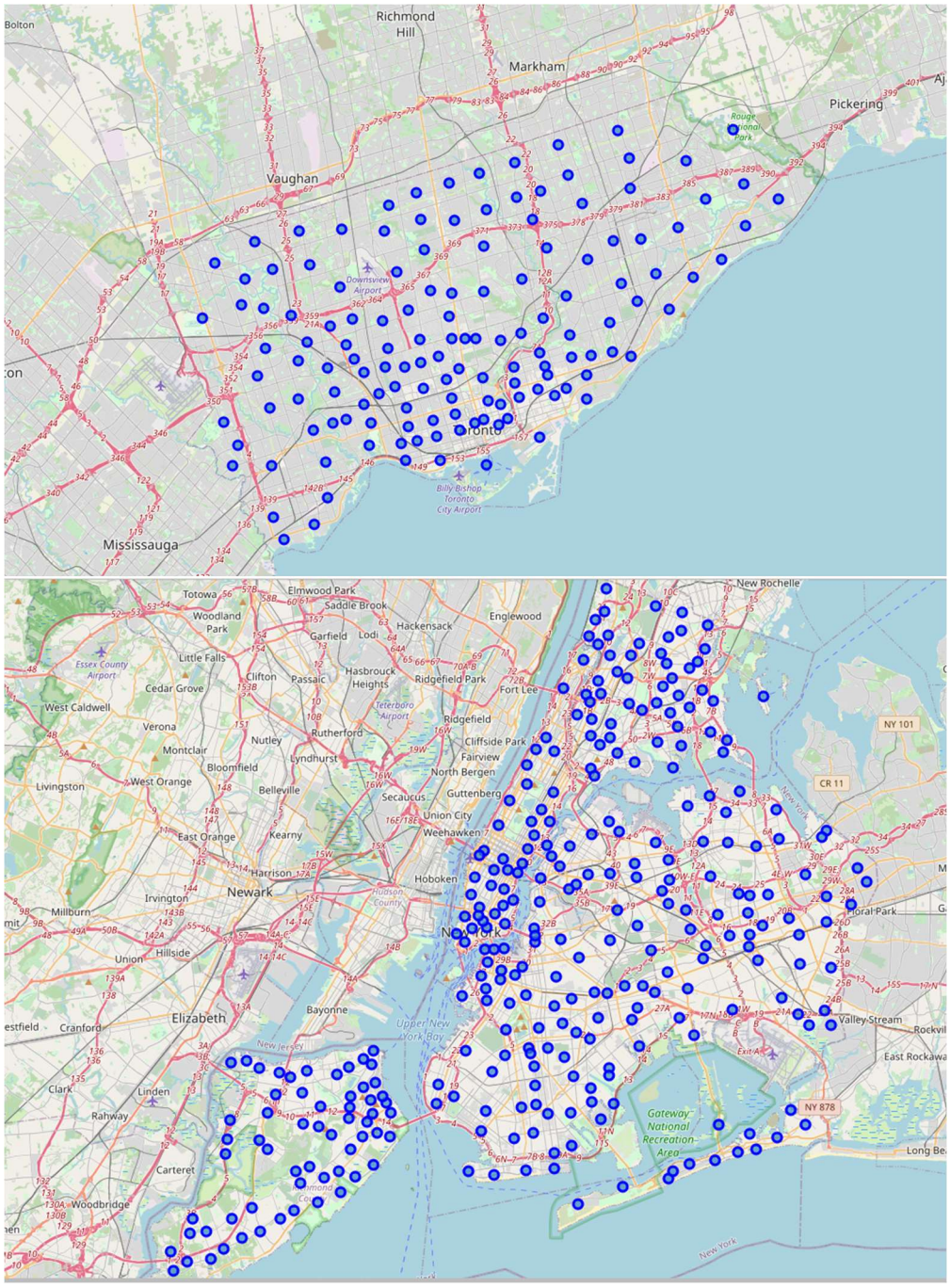
This lab will use two neighborhood boundary datasets from the municipalities and the neighborhood venues data obtained with the Foursquare API. First, I will convert spatial and non-spatial data (neighborhood names/addresses) into their equivalent latitude and longitude values. I will use the explore function to get the most common venue categories in each neighborhood and use search function to get the specific category of venue and then use this feature to group the neighborhoods into clusters. The k-means clustering algorithm is adopted in the clustering analysis to capture the similarities and differences between two cities. Finally, I will use the Folium library to visualize the neighborhoods in Toronto and NYC and their emerging clusters.

2.3.1 Download and Explore Neighborhood Dataset

Neighborhood spatial data of NYC is downloaded from the course server and loaded locally. The data has a total of 5 boroughs and 306 neighborhoods. While cleaning the data, two more columns (Latitude and Longitude) are generated in order to feed the location information into Foursquare API.

There are 140 neighborhoods in 47 wards of Toronto, according to the newest city boundary administration system. I used the spatial information of crime data for the city as I can't find the existing boundary data.

Folium library is adopted to visualize the spatial locations of all the neighborhoods. The following figures show the spatial distributions of neighborhoods in Toronto (Top) and New York City (Bottom). It is obvious that New York City has a bigger spatial coverage with more than double of total neighborhoods. One similarity can be observed that there are concentrations of neighborhoods in some small regions where are densely populated, such as downtowns.



2.3.2 Foursquare Venus Information Extraction for Neighborhoods

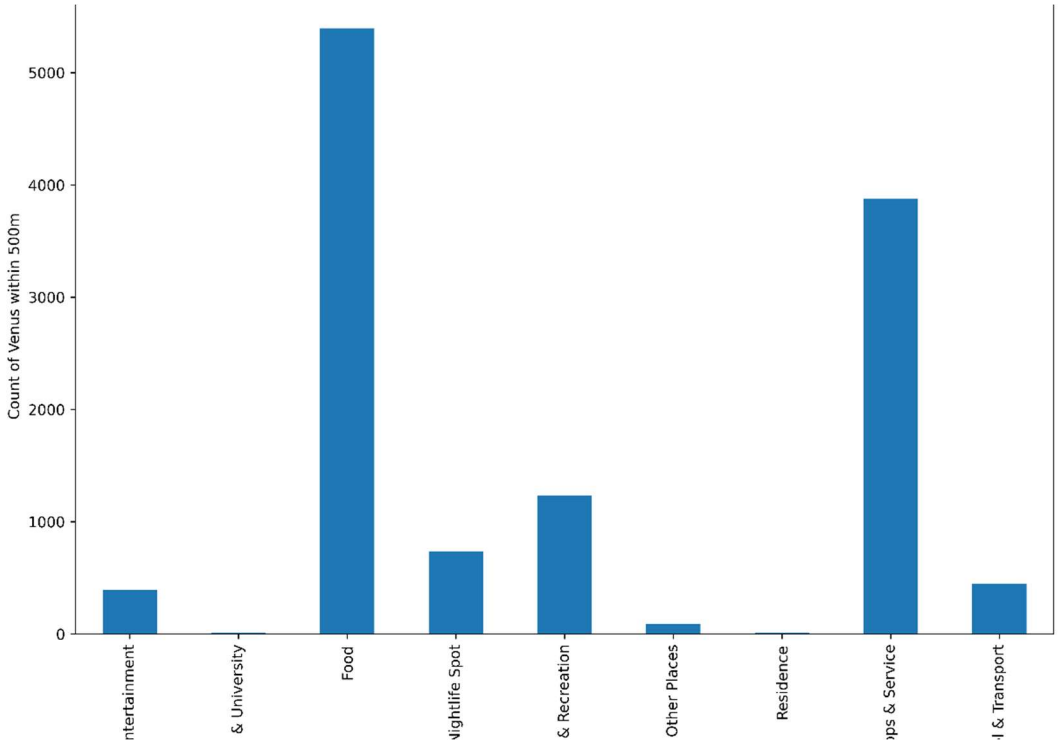
As Foursquare provides venue data for regular user without a cost, in this analysis I use the API to collect the venue locations within 500m of the neighborhood centroids. In order to process the data efficiently, I combine two dataframes into one and add two more columns, namely nb_id and City. nb_id gives unique neighborhood id for each neighborhood in two cities while the city column indicates which city the neighborhood is located at. The data of paired latitude and longitude for each neighborhood will be sent to Foursquare server to obtain the needed venues information. The API call endpoint is shown as below.

```
# create the API request URL
url =https://api.foursquare.com/v2/venues/explore?&client_id={}&
client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
    CLIENT_ID,
    CLIENT_SECRET,
    VERSION,
    lat,
    lng,
    radius,
    LIMIT)
# make the GET request
results = requests.get(url).json()["response"]['groups'][0]['items']
```

A total of 12,194 venues are obtained within 500 m of the neighborhoods' centroids. The venue data has the detailed type and the spatial location. The detailed categories are recoded based on the Foursquare documentation into 9 larger categories (https://docs.foursquare.com/docs/legacy-venue-categories). During the processing, I have to hard code few types as the official document does not include those detailed categories. The following table is the first 5 rows of the dataset.

	neighborhood_name	City	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Category
0	Wakefield	New York City	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop	Shops & Service
1	Wakefield	New York City	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy	Shops & Service
2	Wakefield	New York City	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop	Shops & Service
3	Wakefield	New York City	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy	Shops & Service
4	Wakefield	New York City	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop	Shops & Service

Nearly half of the venues fall in food category, followed by shop & service category and the outdoor & recreational category. The distribution of each category of venues is shown in the following bar plot.



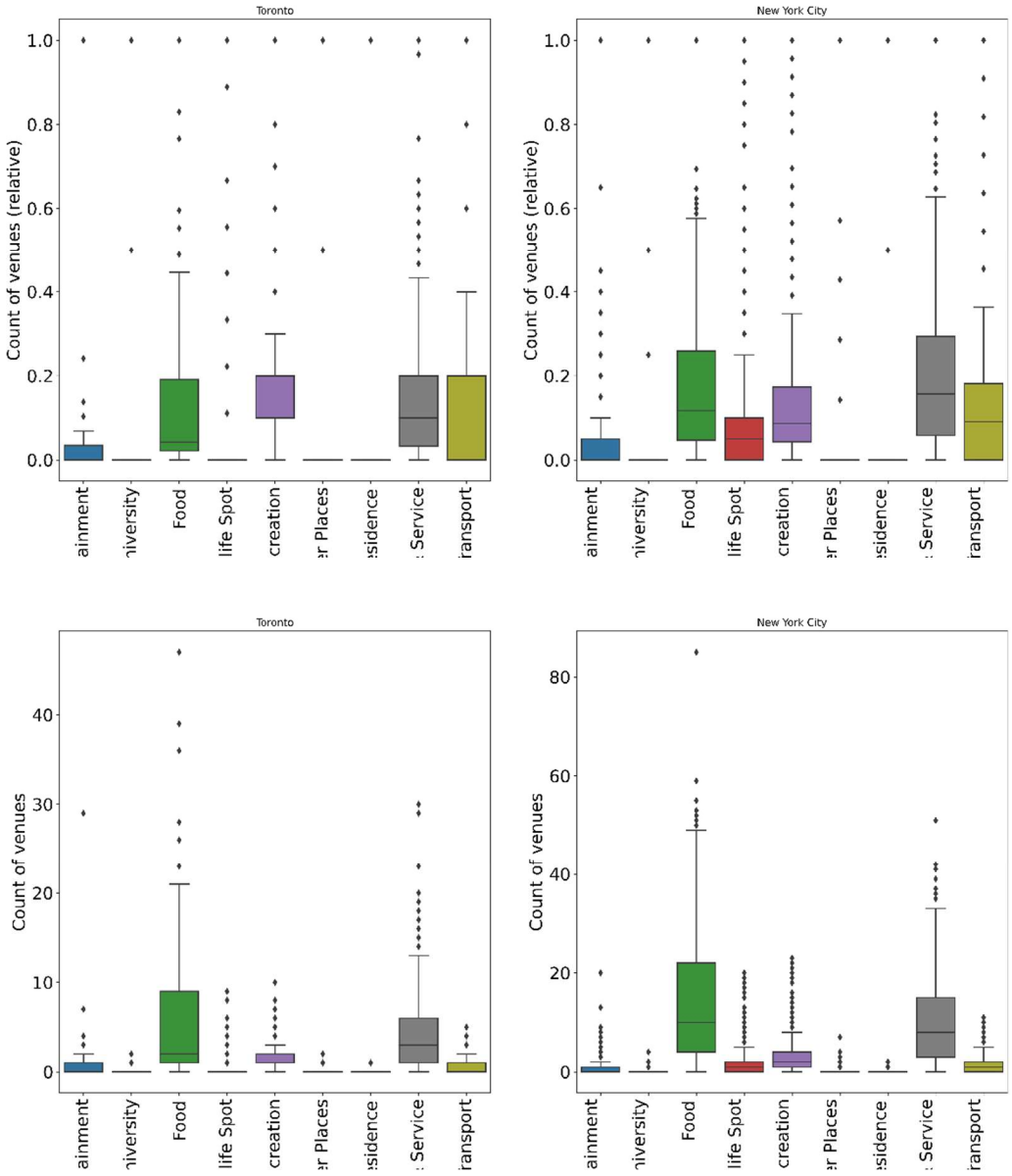
In the next step, I calculate the count of venues in each category for each neighborhood in two cities. As discussed in the previous sections, NYC is much larger with significant bigger population than is Toronto. It will yield some inaccuracy if we directly use the number of venues in each category for comparison purpose. A standardization process is necessary here for each city to make the data comparable. Here I use the MinMaxScaler to stretch the data. The formula is shown below.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

3 Results

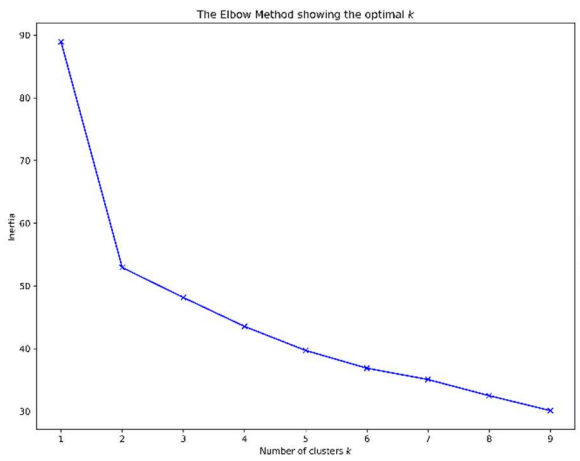
3.1 Variation of Count of Venues

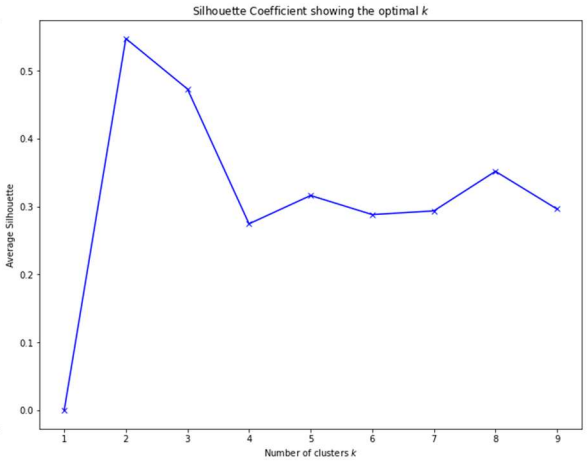
The boxplots for original and scaled count of venues show the variations of different venues categories in two cities. On average, Toronto has more food, shops, and transport related venues, which is quite similar in New York City. However, the magnitude in NYC is much bigger than Toronto. Noticeably, NYC is much richer on nightlife spot category, indicating a total different night culture in the city.



3.2 KMeans Clustering Analysis

In the analysis, I use two methods to find the optimal number for the following cluster analysis, namely the elbow plot method and the Silhouette coefficient Method. Both methods suggest that 2 clusters is the best number for the analysis. However, I continue my analysis with a cluster number of 5. The reason for me not using 2 or 3 is I want to show readers what are the answers for my initial questions. It is evident that NYC and Toronto are significantly different from each other which will automatically make 2 clusters. There is no educational purpose if I choose 2.



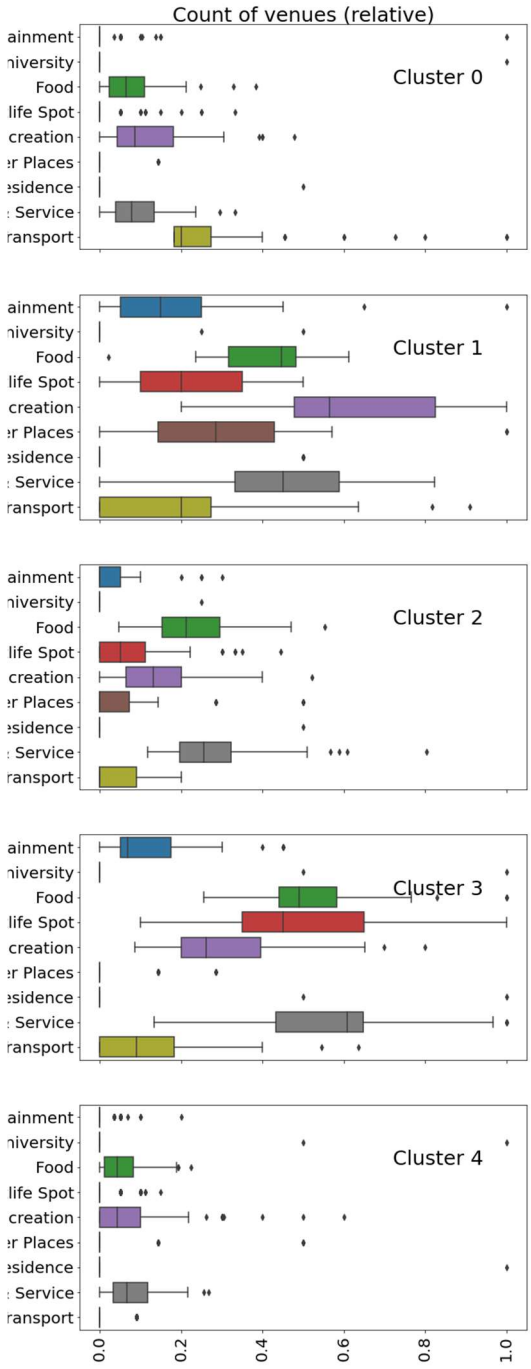


The following table shows the number of neighborhoods from two cities in each cluster. For example, cluster 4 includes 119 NYC neighborhoods and 69 neighborhoods in Toronto while cluster 1 only includes 19 from NYC and 6 from Toronto.

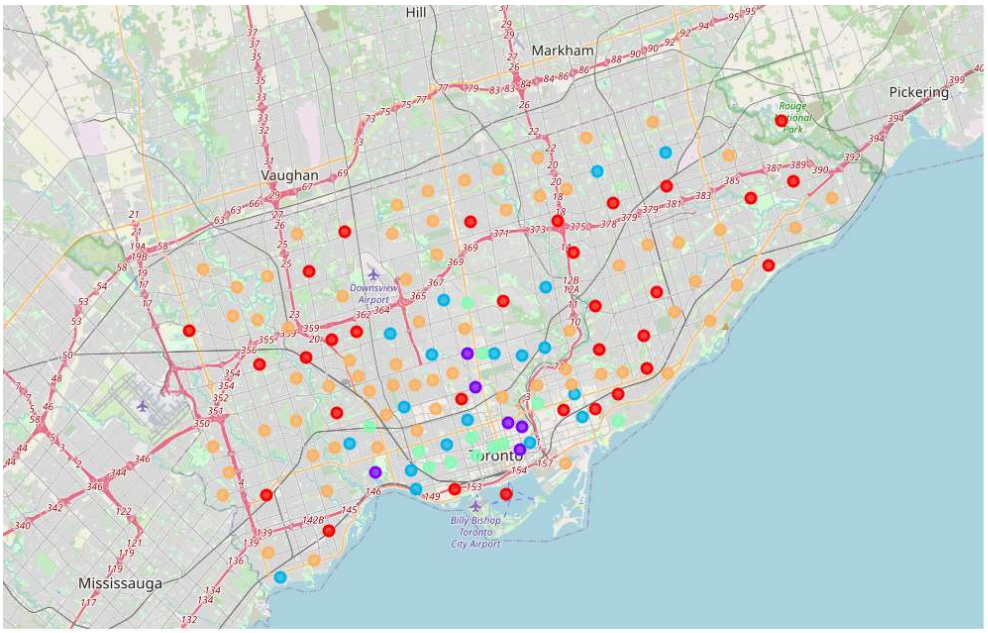
cluster	0	1	2	3	4
City					
New York City	61	19	68	35	119
Toronto	31	6	19	12	69

3.3 Cluster Visualize using Seaborn Boxplot and Folium Leaflet Map

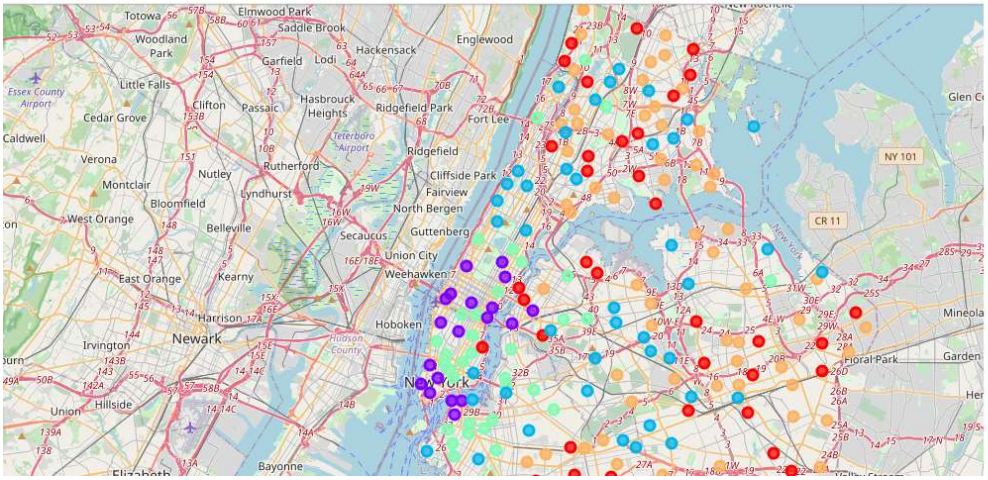
From Seaborn vertical boxplots for clusters, we can clearly see the difference among clusters. Some clusters have significantly more certain category of venues than others, such as cluster 3 has a lot of nightlife spots and cluster 1 the most Outdoors & Recreation venues.



The Folium Leaflet map shows the spatial locations of the clustered neighborhoods in two cities. Some clusters (0 and 1) are more confined in the urban central regions while cluster 3 and 4 are more spread out to the sub-urban areas.



Toronto



New York City

3.4 Examine the most common n venues for each cluster

It is important for neighborhoods to be known for. Hence in this section, I investigate the top 3 most common venues in those clusters. The following output is the result for the top 1 most common venue. Those common venues make the neighborhoods special and unique. While transportation hubs might be the impression for cluster 0 neighborhoods, cluster 4 is known for its recreational resources and the convenience of shopping. Cluster 2 and cluster 3 are both mixed-used multi-cultural neighborhoods which might be very attractive for regular people.

==========
== Cluster 0 (NYC:61 TOR: 31) ==
==========

	Venue Category	1st Most Common Venue
0	Travel & Transport	76
1	Outdoors & Recreation	9
2	Shops & Service	3
3	Residence	1
4	Food	1
5	College & University	1
6	Arts & Entertainment	1

==========
== Cluster 1 (NYC:19 TOR: 6) ==
==========

	Venue Category	1st Most Common Venue
0	Outdoors & Recreation	12
1	Shops & Service	4
2	Professional & Other Places	3
3	Travel & Transport	2
4	Food	2
5	Residence	1
6	Arts & Entertainment	1

==========
== Cluster 2 (NYC:68 TOR: 19) ==
==========

	Venue Category	1st Most Common Venue
0	Shops & Service	44
1	Food	28
2	Outdoors & Recreation	5
3	Nightlife Spot	5
4	Professional & Other Places	4
5	Residence	1

==========
== Cluster 3 (NYC:35 TOR: 12) ==
==========

	Venue Category	1st Most Common Venue
0	Shops & Service	20
1	Nightlife Spot	15
2	Food	6
3	Outdoors & Recreation	3
4	Residence	2
5	College & University	1

==========
== Cluster 4 (NYC:119 TOR: 69) ==
==========

	Venue Category	1st Most Common Venue
0	Outdoors & Recreation	67
1	Shops & Service	66
2	Food	18
3	Travel & Transport	14
4	Nightlife Spot	8
5	Professional & Other Places	6
6	Arts & Entertainment	6
7	College & University	2
8	Residence	1

1st Most Common Venue

4 Discussion and Conclusion

As big cities in North America, NYC and Toronto share many things, such as the multi-culture and the convenient urban venues. Their neighborhoods are clearly defining themselves based on the accessible venues around them. The various special attractiveness of different neighborhoods can provide the opportunities of immigrants or tourists to experience the culture they want to experience. However, some neighborhoods in NYC are known for nightlife spots, which might be favored by certain younger generations. From that perspective, Toronto might not be a good destination for them. Most of neighborhoods in Toronto has a balanced distribution of food, recreation and shopping venues, which might indicate the city is very convenient and livable.

Those similarities and differences between NYC and Toronto can be useful for decision-makers too. They can understand their city better and know where they might invest more if they want the city to be more embraced by young people.

5 Limitations and Perspective

This study has many limitations. First, the data of venues collected is not comprehensive. And when we are considering the neighborhood, we simply abstracted it as a centroid point. In the real life, the most active area of a neighborhood is not necessary in the center of neighborhood. Therefore, the analysis is not 100% impartial and rational spatially. Second, the KMeans cluster technique or the optimal number of clusters might be not suitable for the analysis. Third, due to the time, I only investigated the bigger category of the venues. It will absolutely yield more

accurate and convincible results to reveal the insights of the urban structure in two cities. This analysis is very preliminary and rough in some ways but it opens a door for us to understand the similarities and differences between cities.

6 Reference

- Peer-graded Assignment: Capstone Project - The Battle of Neighborhoods. Available at https://www.coursera.org/learn/applied-data-science-capstone/peer/3a01f/capstone-project-the-battle-of-neighborhoods-week-2. Visited July 28th, 2021.
- Scikit-learn Machine Learning in Python. Available at https://scikit-learn.org/stable/modules/clustering.html#clustering. Visited July 28th, 2021.
- J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.
- Waskom, M. L., (2021). seaborn: statistical data visualization. Journal of Open-Source Software, 6(60), 3021, https://doi.org/10.21105/joss.03021