

A Tale of Two Cities through ML Techniques

Baran Sen

18 Oct 2020

1 Introduction: Business Problem

1.1 Background

In today's age of globalization and rapid adoption of high-end technology in communication and transportation systems, the physical boundaries of the countries and geographical distances have virtually disappeared. The entire world has become one single large arena of human civilization.

One major impact of this phenomenon is that it has encouraged a continuous movement of people from one place to the other with the intention of relocating. Such movements are occurring due to change in job location, higher education, settling into a retired life, migrating to a better quality of life etc.

There is, therefore, a constant search of appropriate destination for relocation. Our project deals with relocation issues faced while **moving from one city to another** of your choice.

1.2 Problem

The problem in relocation lies in finding out the correct destination which suits one's purpose of relocation. Choosing the right city is not enough, finding the right neighborhood is equally important, if not more. This is because the environment would vary from place to place, even within the city. The neighborhoods will have different sets of amenities, cost of living, safety, quality of life etc. Each of us would like to relocate to a place which is better than the current location in terms of these factors. The problem statement, therefore, is as follows: -

Given the purpose of relocation, which neighborhood in the destination city is the most suitable?

In this project, two cities chosen for the study are **New York** and **Toronto**. **We will attempt to evolve an approach to identify suitable Neighborhoods in each city which the individual relocating from the other city would prefer.**

1.3 Interest

Quite obviously, the target audience will be the people living in New York and wanting to relocate to Toronto for reasons such as the following: -

- a. People migrating for a similar or better quality of life
- b. Professionals shifting to the other city looking to stay close to the office
- c. International students looking for a place to live with certain amenities within a budget
- d. Businessman looking to open shop in the other city in a location which would offer good customer base and minimum competition

Likewise, the people living in Toronto and wanting to relocate to New York are also part of the target audience for this project. **In general, individuals looking to relocate to another city would be interested in this project.**

Our approach for suggesting suitability of a neighborhood for relocation would be by mapping the objective of relocation to the categories of venues available in the neighborhood. This, in my opinion, would be a novel approach since there are other parameters (such as crime rate, cost of living, property cost etc) which are not being considered; rather, these aspects would be inferred from the venue details.

2 Data

2.1 Data sources

The idea of the project is to first understand user's (i.e. Target audience) requirements for relocation, thereafter, know his /her present neighborhood and finally recommend a suitable neighborhood in the destination city (New York or Toronto). To achieve the above objectives, following data sources were used: -

a. **Understand user's requirements:** The following framework was adopted to "formalize" the user requirements: -

i. **Purpose of relocation**

- new business
- better quality of life
- Higher education
- Profession

Each option will imply a different set of neighborhood selection criteria

ii. For the selected purpose of relocation, **specify few characteristics** of the neighborhood he / she would like to move in. This would help in narrowing down the recommended options: -

- **New business:** Clientele would be high-income group, medium income group or low-income group
- **Better quality of life:** Countryside, uptown or downtown
- **Education:** University location
- **Profession:** Location of the Office where he / she would work

b. **Know his /her Present Neighborhood.** The user's current residential and office address. The current residential location forms the benchmark for the "quality of life" (as a function of amenities available in the neighborhood). For this purpose, details of present neighborhood of the individual will be extracted out of his / her residential address using Foursquare app.

Neighborhood details obtained using **Foursquare** app have been discussed in succeeding paragraphs.

c. **Recommend a Suitable Neighborhood.** To recommend a suitable neighborhood, we would need details of all the neighborhoods in New York and Toronto. Neighborhood details would comprise of the following: -

i. **Neighborhood details of Toronto and New York** - Borough, Neighborhood name, Neighborhood Latitude, Neighborhood Longitude

For this purpose, following data sources were used:-

List of all neighborhoods of Toronto will be scraped from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

However, the neighborhood dataset of Toronto available at this website does not contain Latitude, Longitude data of the neighborhoods. The Latitude, Longitude data were retrieved from the site http://cocl.us/Geospatial_data.

List of all neighborhoods of New York were scraped from the database maintained at https://cocl.us/new_york_dataset (courtesy: "Segmenting and Clustering Neighborhoods in New York City" Lab Exercise of week 3 of this Module)

2.2 Data cleaning

Data downloaded or scraped from the abovementioned sources were combined into one table for each city. Following data cleaning activities were carried out: -

- a. **Borough not assigned.** In this case, the neighborhood records were deleted.
- b. **Neighborhood not assigned.** In such cases, the neighborhood name was made the same as the borough name.
- c. **Duplicate entries of neighborhood** Duplicate entries of neighborhoods with different Postal Codes were found. On further investigation, it was seen that the neighborhoods with the same name were geographically distinct, i.e., North, South, East, West, Central etc. Corrections were made by adding the location information (North, South etc) as prefix.

One example is that of “Downsview” which had four records: -

PostalCode	Borough	Neighborhood
M3K	North York	Downsview
M3L	North York	Downsview
M3M	North York	Downsview
M3N	North York	Downsview

The details were corrected as follows: -

PostalCode	Borough	Neighborhood
M3K	North York	Downsview East
M3L	North York	Downsview West
M3M	North York	Downsview Central
M3N	North York	Downsview North

2.3 Create combined database of two cities

The database of Toronto and New York neighborhoods were combined into one database "combined_data" so that the two cities can be compared and contrasted.

2.4 List of prominent venues in neighborhoods

This data was used to characterise various neighborhoods and compare them with the user's requirements for relocation. Following types of venues have been obtained for each neighborhood using **Foursquare** application: -

- a. Public amenities (Departmental stores, Hospital, School, Bank, Post Office etc)
- b. Recreational & sports facilities (Sports centre, Cinema, Museum, Gym, Yoga studio etc.)
- c. Shopping malls, airport terminal, railway station
- d. Food joints (restaurant of various cuisines)

2.4.1 Venue Database using Foursquare App

For each neighborhood in the combined database, details of **all venues within 500 m** radius were obtained using Foursquare app. The venue details include **name, latitude/longitude** and **category** of the venue.

2.4.2 Venue Database Cleaning

Wrong venue details. In some cases, the venue category was mentioned as "neighborhood" in the Foursquare data. Such records have been deleted.

2.5 Pre-processing of Venues data

Following pre-processing was performed with the venues data obtained by using Foursquare: -

a. **Generalisation of venue categories.** Venue details received through Four Square application was too specific. As a result, there were about **462 'Venue Category'**. Such very detailed information on the venues were not necessary for meeting the project objectives, rather, it made the analysis very complex. Hence, the venues were further categorised into **16 generic categories**. E.g., there were more than **95 types of Restaurant**, besides another **33+ venues of various types of eatery**. These were classified into only **03 categories**, viz., '**Bakery and Cafe**', '**Budget Food Joint**' and '**Restaurant**'.

b. **Additional attributes of the venues.** Additional attributes of the venue were recorded, e.g., the cuisine (such as Indian, Thai etc) in case of Food Joints. These will be used only if the individual has such specific requirements for stay / relocation.

Output of the above processing was the following **Generic Categories**: -

- | | |
|--------------------------------------|----------------------------|
| • Academic Institutes and Facilities | • Lung Space |
| • Bakery and Café | • Market Place |
| • Bar Pub etc | • Public Tpt System |
| • Basic Needs | • Restaurant |
| • Budget Food Joint | • Lifestyle |
| • Business Facility | • mall_plaza_supermarket |
| • Essential Services | • sports and entertainment |
| • Health and Wellness | • Hotel Lodge etc |
| | • Ignored |

As can be seen, there is one category named "**Ignored**". These venues were not much useful in characterizing the neighborhoods. Although few of them sounded relevant, but online scrutiny revealed these were mis-categorised. The venues belonging to "**Ignored**" category were excluded from further processing.

- Aquarium
- Art Gallery
- Art Museum
- Auditorium
- Beach
- Bridge
- Building
- Child Care Service
- Church
- Escape Room
- Farm
- Field
- Financial or Legal Service
- Fountain
- Historic Site
- History Museum
- Intersection
- Memorial Site
- Monument / Landmark
- Moving Target
- Museum
- Newsstand
- Outdoor Sculpture
- Pier
- Plane
- Professional & Other Places
- Recording Studio
- Rental Service
- Residential Building (Apartment / Condo)
- Rest Area
- River
- Ski Area
- Soup Place
- Speakeasy
- Storage Facility
- Street Art
- Tech Startup
- Tourist Information Center
- Track
- Trail
- Train
- Waste Facility

c. One-hot encoding to **convert Generic Categories into numeric data** was carried out, since clustering and other mathematical processing would require numeric values.

2.6 Processed databases:

So, now we have the following databases, cleaned up and pre-processed, for further analysis and conclusion: -

- a. **combined_data** : Master database which contains details of all neighborhoods (limitation: as per data available through the Internet sources)
- b. **combined_venues** : Master database of venues within 500m radius of each neighborhood of the two cities.
- c. **combined_venues_dummies** : Master database of venues combined into generic categories. Only the relevant generic categories were retained in the database.

3 Methodology

3.1 In the project, we will attempt to characterise each neighborhood of the two cities based on the venues such as restaurant, shops for daily need, gym and other health facilities, essential services etc. Availability and density of these venues will provide an indication of nature of the neighborhood. For example, availability of large number of restaurants and shops indicate that the neighborhood is a densely populated residential area. The analysis will be performed using the following two methods: -

- a. **The conventional approach**: Using Exploratory Data Analysis techniques
- b. **Machine Learning approach**: Using unsupervised learning technique - Clustering

3.2 Aggregation of Venue Characteristics into Generic Categories and Features.

The venue database would be systematically reduced into a generalized set of **Features** in two stages. Firstly, the venues will be grouped into **generic categories** based on similarity of the types. Thereafter, subsets of the Generic Categories will be taken to define a **Feature**. For example, restaurants of various cousins will be grouped into **Restaurant** generic category. Restaurant will be part of the **eatery** feature which would also include generic categories "**Bakery and Cafe**", "**Budget Food Joint**" and "**Bar Pub etc**".

3.3 Feature selection

Appropriate features will be selected which would map the neighborhood characteristics to the purpose of move of the migrating people. The features would be a sub-set of the neighborhood characteristics that meets the needs. An example mapping is tabulated below. Actual mapping will evolve at the time of data analysis.

Purpose of Move	Neighborhood Features
Quality of life	Lung Space (Park, Jogging track etc), lifestyle, shopping, Recreational facilities
Professionals	Hotel Lodge etc, Public Tpt System, Food Joints, public amenities
International students	Basic Needs, Budget Food Joint, sports and entertainment, transport facility
Businessman	availability of potential customer

Obviously, the neighborhoods whose features satisfy the purpose of move will be preferred by the individual for relocation. The entire **conventional approach is depicted in Figure 1** below.

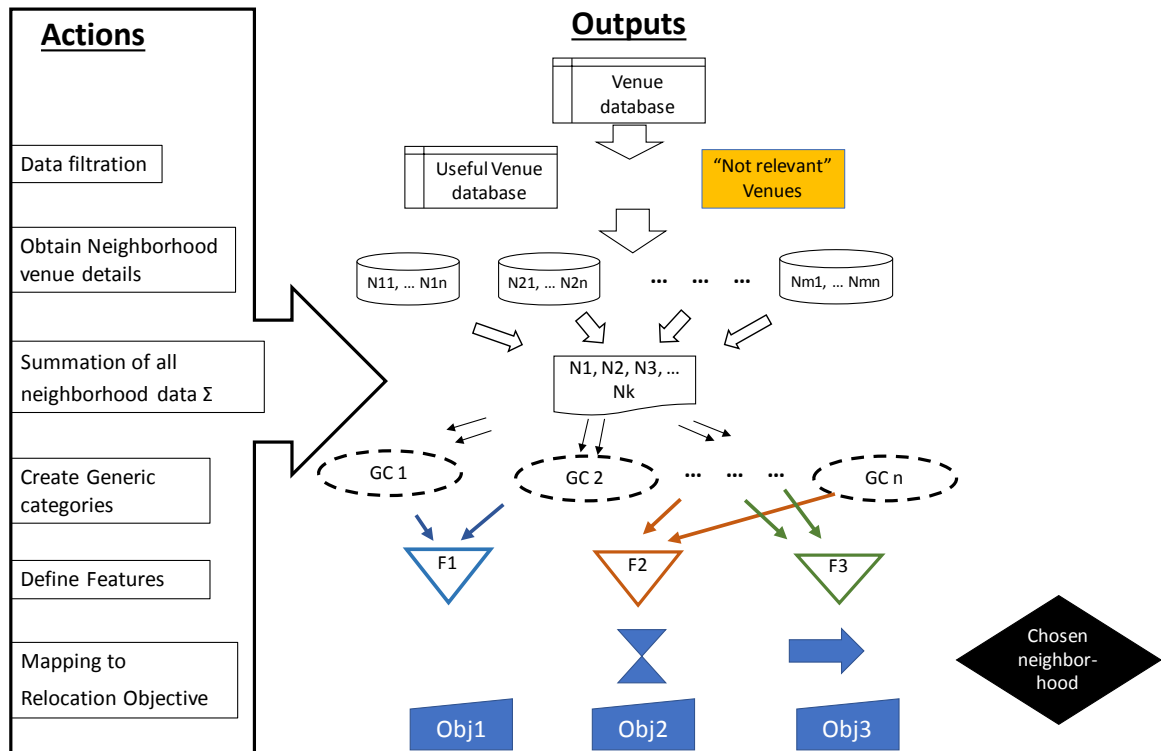


Fig 1. Concept of Neighborhood selection based on Venue Characteristics by Conventional approach

3.4 Clustering

Attempt will be made to cluster the neighborhoods into various groups based on their features. This would provide an opportunity to analyse the neighborhoods for their similarity and differences. **The objective is to try to replace the entire process depicted in Fig 1 through Clustering.**

4 Analysis

4.1 Exploratory Data Analysis

We performed the following exploratory data analysis: -

- The exploratory data analysis was carried out by grouping the dataset into Neighborhoods and summing up the venues under various Generic Categories.
- Database of **top venues** in all neighborhoods was created which consisted of the **top 10 venue categories by number** for each neighborhood.
- top_venues** was analysed to characterise each neighborhood. The database was also used to arrive at the most common venue categories available in all the neighborhoods.
- Principal Component Analysis** was carried out to establish correlations between the various venue categories. The analysis provided a clue to what venues can be expected to co-exist in a neighborhood.

e. **Comparison between New York and Toronto.** Profiling of the two cities was carried out, i.e., which city has higher density of different venue categories, city-wise box-plot of venue distribution in various neighborhoods etc.

4.1.1 Neighborhood analysis based on Features

After performing the abovementioned preliminary analysis, features were defined as described in **Para 3.3** and all neighborhoods were examined to figure out which features are present in each neighborhood. List of **Top 15** neighborhoods in each **Feature** was prepared.

These lists are the most important output of the Exploratory Data Analysis, which will be used by the individuals to make decision regarding which neighborhood to move in.

4.2 **Most common venue categories** were identified by aggregating the venue counts across all the neighborhoods and plotted in decreasing order. Graphs of Topmost, Second Topmost and 3rd Topmost venues are depicted in **Figure 2** below.

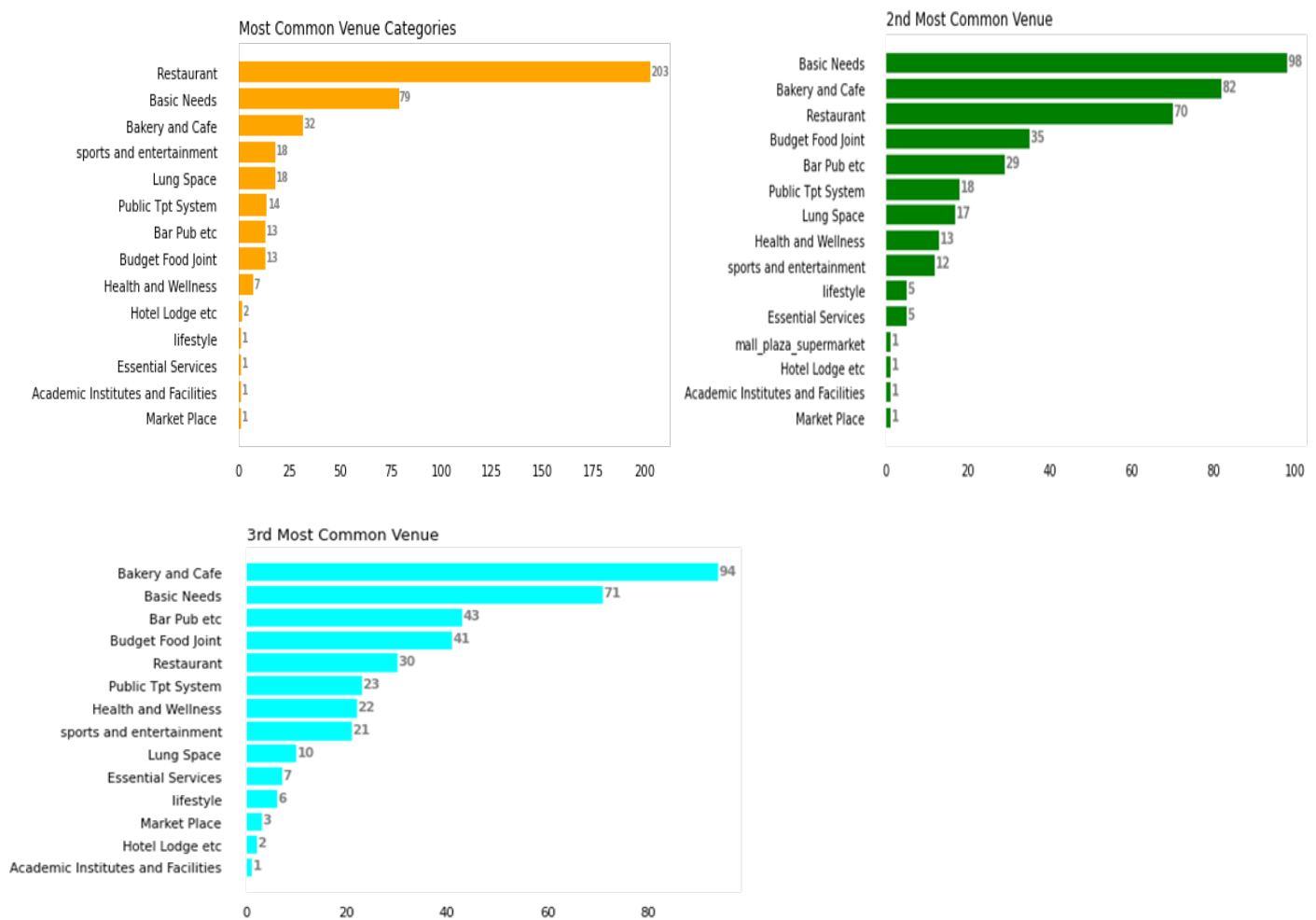


Fig 2. Most Common Venue Categories

4.2.1 As it emerged, **“Restaurant”** was the topmost Common Venue category, i.e., number of neighborhoods where Restaurant was the most common venue was highest. Further analysis was carried out with these neighborhoods and it was found that **“Basic Needs”** was the Second most Common Venue. Likewise, out of the neighborhoods with **“Basic Needs”** as the Second most common venue, **“Bakery and Café”** emerged as the Third Most Common Venue. The results are depicted in **Figure 3** below.

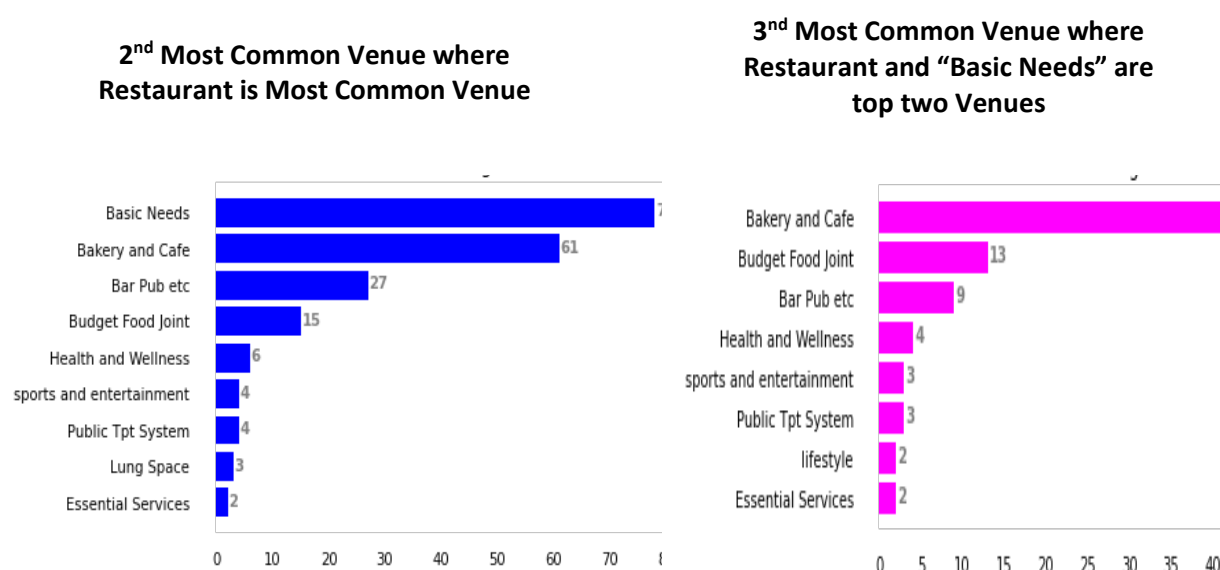


Fig 3. 2nd and 3rd Most Common Venues in Neighborhoods with Restaurant as Topmost Venue

4.2.2 Findings:

a. **“Restaurant”, “Basic Needs” and “Bakery and Cafe”** are the **three most common venues** in majority of the neighborhoods. Availability of these venues as the most common ones indicate that such neighborhoods are well-populated residential areas. Preferred by those who would relocate for jobs and decent living.

b. Neighborhoods with most common venues **“Lifestyle”** and **“Lung Space”** are the high-cost-of-living areas. Suitable for those looking to relocate for enjoying a relaxed and premium lifestyle.

c. The neighborhoods with **“Budget Food Joint”** as the most common venue offers affordable living. Such neighborhoods would be preferred by students for relocation.

4.3 Principal Component Analysis

Principal Component Analysis was carried out by computing the correlation among the Venue Category counts.

4.3.1 Findings:

a. Following venue categories have **high correlation (> 0.6)**: -

i. **“Bakery and Cafe”, “Bar Pub etc”, “Basic Needs”, “Health and Wellness”, “Restaurant”**. This implies that neighborhoods which have any of these categories as the most common venues, are likely to have venues of the remaining categories as well.

b. Following venue categories are moderately correlated (0.45 to 0.55): -

i. "lifestyle", "Health and Wellness"

ii. "Health and Wellness", "sports and entertainment"

iii. "Academic Institutes and Facilities" do not have good correlation with any other venue category. This could imply that the university campus is self-sufficient and/or other venues (such as restaurant, basic needs etc) are located beyond radius of 500 metres. My database is built based on venues within a radius of 500 meters.

4.4 New York and Toronto data compare

4.4.1 **Cumulative numbers of venues in different categories** Stacked Bar graphs were plotted for cumulative numbers of venues in different categories for New York and Toronto. The graph provided a comparative study of the venues in the two cities. Refer **Figure 4**.

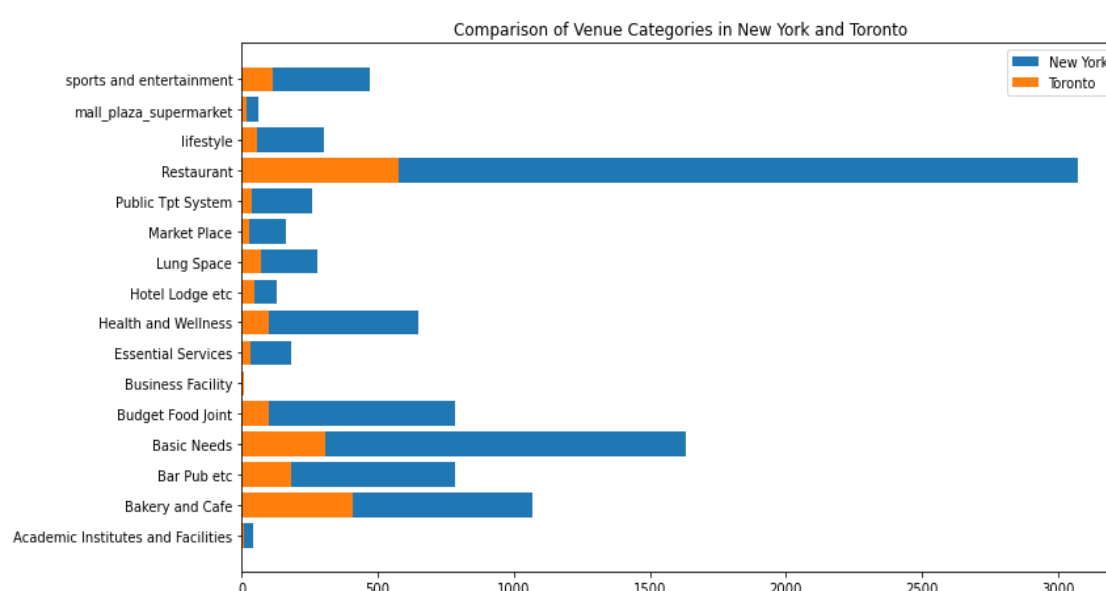


Fig 4. Stacked Bar Graphs of Total Venues in Different Categories

Findings: -

- In all generic categories, number of venues in New York are much more than Toronto.
- For both New York and Toronto, the relative proportions of various generic category venues are very similar to each other
- In both the cities, Restaurant is the most common venue among all generic categories.

4.4.2 **Comparison of Neighborhood-wise distribution of venues in New York and Toronto.** Four generic categories viz, "Restaurant", "sports and entertainment", "Lung Space" and "Basic Needs" were chosen for the analysis. Box Plot of the data are depicted in **Figure 5**.

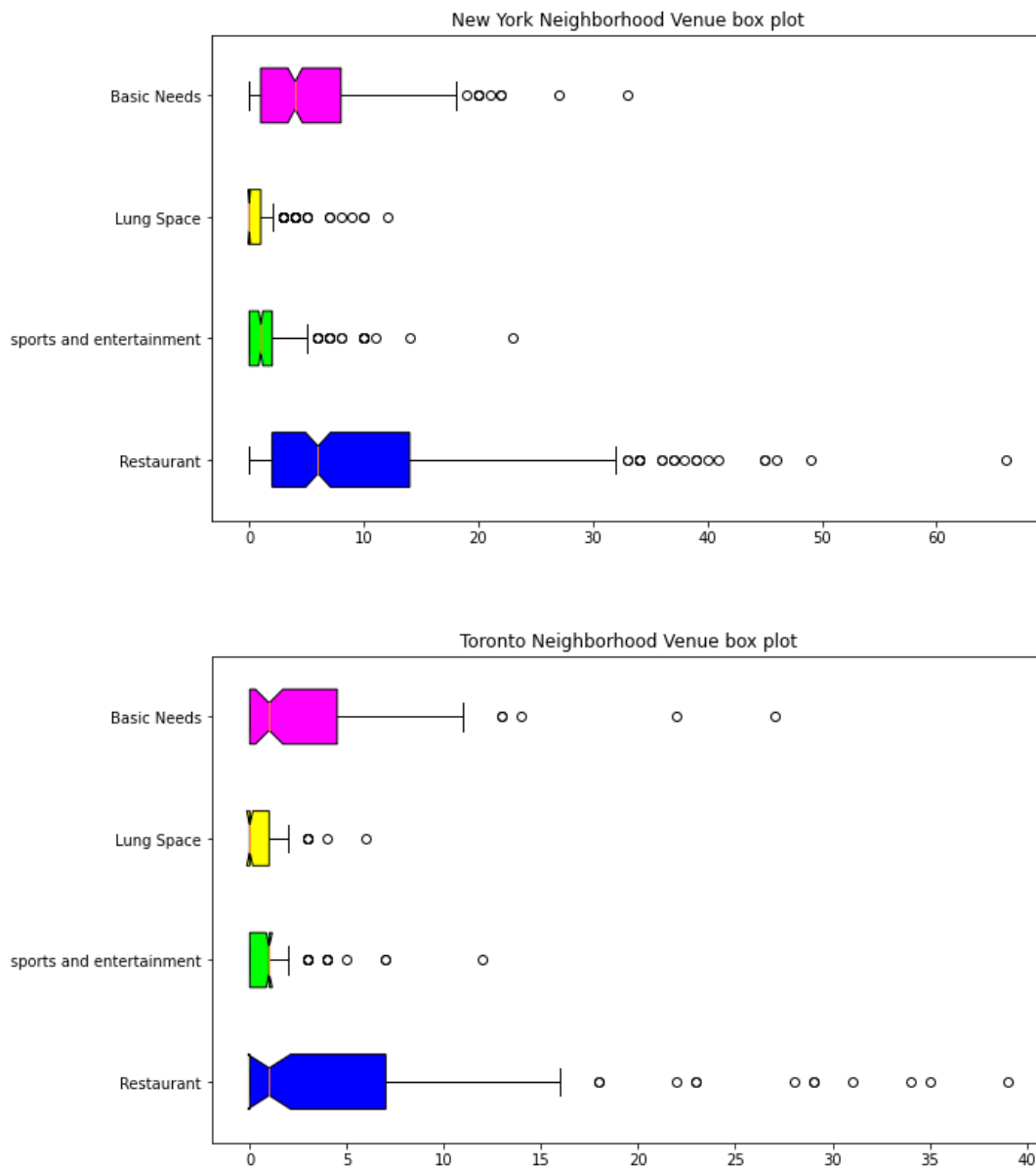


Fig 5. Box Plot of Distribution of Venues in New York and Toronto

Findings: -

- In both the cities, the distribution is similar - with the inter-quartile range being quite close to the first quartile.
- In both the cities, the inter-quartile ranges of "**Restaurant**" and "**Basic Needs**" are much larger than "**Lung Space**" and "**sports and entertainment**". Moreover, the inter-quartile ranges of "**Lung Space**" and "**sports and entertainment**" are very close to zero. This indicates that majority of the neighborhoods do not have these two venue categories (within 500 m radius).
- In case of New York, the median value for "**Basic Need**" and "**Restaurant**" are approximately **5** and **6** respectively. However, in case of Toronto, these values are just about **one**. This indicates that the venue concentration in Toronto neighborhoods are much lower, as compared to New York.
- In both the cities, Restaurant is the most common (generic category) venue

4.5 Neighborhood Comparison based on Features

Effort was made to compare the neighborhoods based on the following features, which are representative of needs of the individual who is relocating: -

- Eatery (database name: eatery): This is aggregation of "Restaurant", "Bakery and Cafe", "Budget Food Joint" and "Bar Pub etc"
- Lifestyle (database name: lifestyle): Set of "Lung Space" and "lifestyle" venues
- Daily Life (database name: daily_life): Aggregation of "Basic Needs", "Essential Services", "Market Place" and "mall_plaza_supermarket"
- Fitness (database name: fitness): Set of "Health and Wellness" and "sports and entertainment" venues
- Student (database name: student_needs): Set of "Academic Institutes and Facilities", "Basic Needs", "Budget Food Joint" and "sports and entertainment"
- Professional (database name: professionals_need): Set of "Business Facility", "Hotel Lodge etc" and "Public Tpt System" venues.

It is to be noted that, **eatery** and **daily_life** features are aggregations of various venue categories, which are similar in nature. The remaining features are collections i.e., set of few venue categories which would collectively define that feature.

The databases of features so created have been sorted in descending order to obtain the list of **top 15 neighborhoods for each feature**. In case of the features which are sets of venue categories, one category (which is considered as the Most Important Attribute (MIA) of the feature) has been chosen as the key for sorting and generating list of top 15 neighborhoods. The outputs are tabulated in Table 1, 2 and 3 below.

Table 1. Top 15 in Eatery and Lifestyle Features

S. No.	eatery_15			lifestyle_15		
	Borough	Neighborhood	Venue Count	Borough	Neighborhood	Venue Count (MIA)
1	Queens	Astoria	83	Brooklyn	Fulton Ferry	12
2	Manhattan	East Village	76	Manhattan	Tribeca	10
3	Brooklyn	South Side	74	Brooklyn	Dumbo	10
4	Manhattan	Murray Hill	73	Manhattan	Tudor City	9
5	Downtown Toronto	Commerce Court, Victoria Hotel	73	Manhattan	Battery Park City	8
6	Downtown Toronto	First Canadian Place, Underground city	73	Brooklyn	Brooklyn Heights	7
7	Downtown Toronto	Toronto Dominion Centre, Design Exchange	70	Manhattan	Turtle Bay	7
8	Brooklyn	North Side	69	Downtown Toronto	Harbourfront East, Union Station,	6

S. No.	eatery_15			lifestyle_15		
	Borough	Neighborhood	Venue Count	Borough	Neighborhood	Venue Count (MIA)
					Toronto Islands	
9	Brooklyn	Greenpoint	68	Bronx	Clason Point	5
10	Brooklyn	Downtown	68	Manhattan	West Village	5
11	Manhattan	Financial District	67	Manhattan	Roosevelt Island	5
12	Brooklyn	Prospect Heights	67	East Toronto	Business reply mail Processing Centre,	4
13	Manhattan	Upper West Side	66	Queens	Hunters Point	4
14	Manhattan	Turtle Bay	65	Manhattan	Sutton Place	4
15	Brooklyn	Carroll Gardens	65	Bronx	Spuyten Duyvil	4

Table 2. Top 15 in Daily life and Fitness Features

S. No.	daily_life_15			fitness_15		
	Borough	Neighborhood	Venue Count	Borough	Neighborhood	Venue Count (MIA)
1	Bronx	Fordham	42	Manhattan	Civic Center	16
2	Downtown Toronto	Garden District, Ryerson	30	Manhattan	Yorkville	14
3	Manhattan	Soho	29	Brooklyn	Brooklyn Heights	14
4	Bronx	Belmont	29	Queens	Forest Hills	12
5	Manhattan	Washington Heights	27	Manhattan	Carnegie Hill	11
6	Manhattan	Chelsea	26	Manhattan	Clinton	11
7	North York	Fairview, Henry Farm, Oriole	25	Manhattan	Sutton Place	11
8	Queens	Sunnyside Gardens	23	Manhattan	Murray Hill	10
9	Queens	Bay Terrace	22	Brooklyn	Boerum Hill	10
10	Brooklyn	Boerum Hill	21	Manhattan	Upper East Side	10
11	Queens	Woodside	21	Manhattan	Midtown South	9
12	Manhattan	Flatiron	21	Manhattan	Midtown	9
13	Manhattan	Chinatown	20	Manhattan	Flatiron	9
14	Manhattan	Little Italy	20	Manhattan	Financial District	8
15	Brooklyn	Carroll Gardens	19	Manhattan	Lincoln Square	8

Table 3. Top 15 in Student Needs and Professional Needs Features

S. No.	student_needs_15			professionals_need_15		
	Borough	Neighborhood	Venue Count	Borough	Neighborhood	Venue Count (MIA)
1	Manhattan	Lincoln Square	3	Downtown Toronto	Harbourfront East, Union Station, Toronto Islands	2
2	Queens	Hillcrest	2	Manhattan	Chelsea	1
3	Downtown Toronto	University of Toronto, Harbord	2	Brooklyn	Vinegar Hill	1
4	Manhattan	Hamilton Heights	2	Downtown Toronto	Commerce Court, Victoria Hotel	1
5	Brooklyn	Gowanus	1	Downtown Toronto	Richmond, Adelaide, King	1
6	Brooklyn	North Side	1	Queens	Long Island City	1
7	Central Toronto	Lawrence Park	1	Downtown Toronto	St. James Town	1
8	Downtown Toronto	Church and Wellesley	1	Downtown Toronto	Garden District, Ryerson	1
9	Downtown Toronto	Garden District, Ryerson	1	Queens	Steinway	1
10	Downtown Toronto	Queen's Park, Ontario Provincial Government	1	Downtown Toronto	Toronto Dominion Centre, Design Exchange	1
11	Manhattan	Flatiron	1	Downtown Toronto	Central Bay Street	1
12	Brooklyn	Gerritsen Beach	1	Manhattan	Tudor City	1
13	Manhattan	Battery Park City	1	Manhattan	Flatiron	1
14	Manhattan	Carnegie Hill	1			
15	Manhattan	Central Harlem	1			

Findings: -

Scrutiny of the top 15 lists for various features reveals that these lists are largely **dis-joint (i.e., orthogonal)**, except for a few neighborhoods which are as follows: -

- a. Neighborhoods **Chelsea** and **Flatiron** are among the top 15 in **Professional** and **student** features

- b. Neighborhoods **Fordham, Boerum Hill** and **Flatiron** are among the **top 15** in **daily_life** and **fitness** features
- c. **Financial District** neighborhood is among the top 15 in **lifestyle** and **eatery**
- d. There is no neighborhood which is in **top 15** in all features.

Remarks: The above findings imply that each neighborhood (except 05 neighborhoods mentioned above) has characteristics which quite uniquely belong to only one feature. So, an individual, who is relocating from another neighborhood will choose distinctly different neighborhood depending on his / her purpose of relocation.

4.5.1 Outliers

- a. The following neighborhoods have very low density of **eatery venues** (zero within 500 m radius). One needs to be very selective / cautious in choosing these neighborhoods for relocation, since conveniences may not be located near-by. Details are tabulated below (**Table 4**).

Table 4. Outliers in Eatery Feature

Neighborhood	Restaurant	Bakery and Cafe	Budget Food Joint	Bar Pub etc	Total Food venues
Bayswater	0	0	0	0	0
Bergen Beach	0	0	0	0	0
Breezy Point	0	0	0	0	0
Butler Manor	0	0	0	0	0
Downsview South	0	0	0	0	0
Downsview West	0	0	0	0	0
East Toronto, Broadview North (Old East York)	0	0	0	0	0
Fieldston	0	0	0	0	0
Howland Hook	0	0	0	0	0
Humberlea, Emery	0	0	0	0	0
Humewood-Cedarvale	0	0	0	0	0
Jamaica Estates	0	0	0	0	0
Lawrence Park	0	0	0	0	0
Malba	0	0	0	0	0
Mill Island	0	0	0	0	0

Neighborhood	Restaurant	Bakery and Cafe	Budget Food Joint	Bar Pub etc	Total Food venues
Milliken, Agincourt North, Steeles East, L'Amoreaux East	0	0	0	0	0
Port Ivory	0	0	0	0	0
Roselawn	0	0	0	0	0
Scarborough Village	0	0	0	0	0
Somerville	0	0	0	0	0
The Kingsway, Montgomery Road...	0	0	0	0	0
Todt Hill	0	0	0	0	0
Westerleigh	0	0	0	0	0
York Mills West	0	0	0	0	0

b. The following neighborhoods, which are among the last in the **lifestyle** feature list have **zero "Lung Space" and zero / one "lifestyle" venues**. The **daily_life** feature of these neighborhoods were explored. It was found that even the **daily_life** feature is not prominent, indicating that these neighborhoods may not be suitable for relocation from another city. List of these neighborhoods are placed in Table 5 below.

Table 5. Outliers in Lifestyle Feature

Neighborhood	Lung Space	lifestyle	Basic Needs	Essential Services	Market Place	mall_plaza_sup ermarket
Agincourt	0	1	0	0	0	0
Mount Hope	0	0	6	0	1	0
Mott Haven	0	1	6	0	0	0
Morrisania	0	0	7	0	1	0
Morris Park	0	0	2	1	0	0
Morris Heights	0	0	1	1	0	0
Mimico NW, The Queensway West...	0	1	5	0	0	0
Mill Island	0	0	0	0	0	0

Neighborhood	Lung Space	lifestyle	Basic Needs	Essential Services	Market Place	mall_plaza_supermarket
Midwood	0	0	4	0	0	0

Exploratory Data Analysis has helped us to identify top 15 neighborhoods for each feature that an individual would look for to relocate in the neighborhood. The analysis has also provided an insight into the most prominent venues. Besides, the two cities could be compared in terms of the venues to figure out which one is busier, more populated and with more amenities.

4.6 Neighborhood Clustering

Next, we applied **k-means Clustering Machine Learning algorithm** to create clusters of the neighborhoods based on the **combined_venues_grp** database. The database was clustered into **seven clusters** as follows: -

- Scaling of database was performed since the counts of few venues (such as restaurant) were too high. Without scaling, the clustering would have been dominated by such venues.
- Number of clusters (**knn**) was selected by applying **Elbow Method** using **inertia** and **distortion** parameters
- For optimum result, **init='k-means++'** was chosen

Output of Elbow Method is depicted in **Figure 6** below.

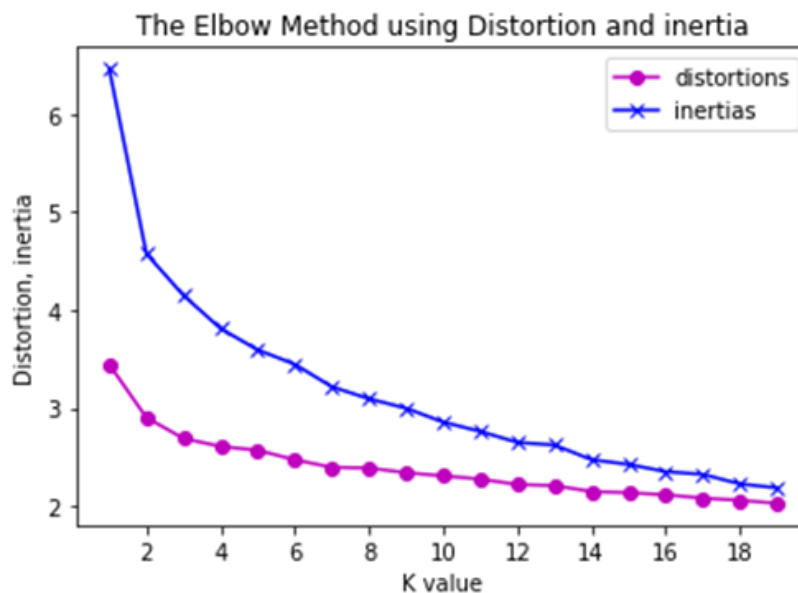


Fig. 6 Obtaining Optimum knn using Elbow Method

The cluster labels (obtained through **k-means Clustering**) were inserted into **top_venues** and **combined_venues_grp** databases. The two databases were combined to obtain a neighborhood database **combined_data_cluster** which contains the **cluster label** of each neighborhood and venue counts in each category.

4.6.1 The individual clusters were analysed to obtain the following: -

- Which **features** (out of 'eatery', 'lifestyle', 'professionals', 'student', 'fitness' and 'daily_life') are dominant in each cluster.
- Which are the **top three most common venue categories** in each cluster and which venue categories are **least common**
- How many venue categories each cluster has? For example, a cluster might have 10 different venue categories, whereas another cluster may have only two types of venues.

Objective of the above analysis is to characterise each cluster in terms of the various features and venue types. This will aid in choosing appropriate neighborhood based on the individual's needs (i.e., features) for relocation.

4.6.2 Neighborhood venue characteristics

Efforts were made to characterise each cluster by mapping onto the various features identified during the **Exploratory Data Analysis**. Further, the two cities were compared through map depiction using **Folium**. The neighborhoods were represented using coloured dots, each colour representing one cluster Id. The outputs for New York and Toronto cities are displayed in **Fig 7** and **8** respectively.

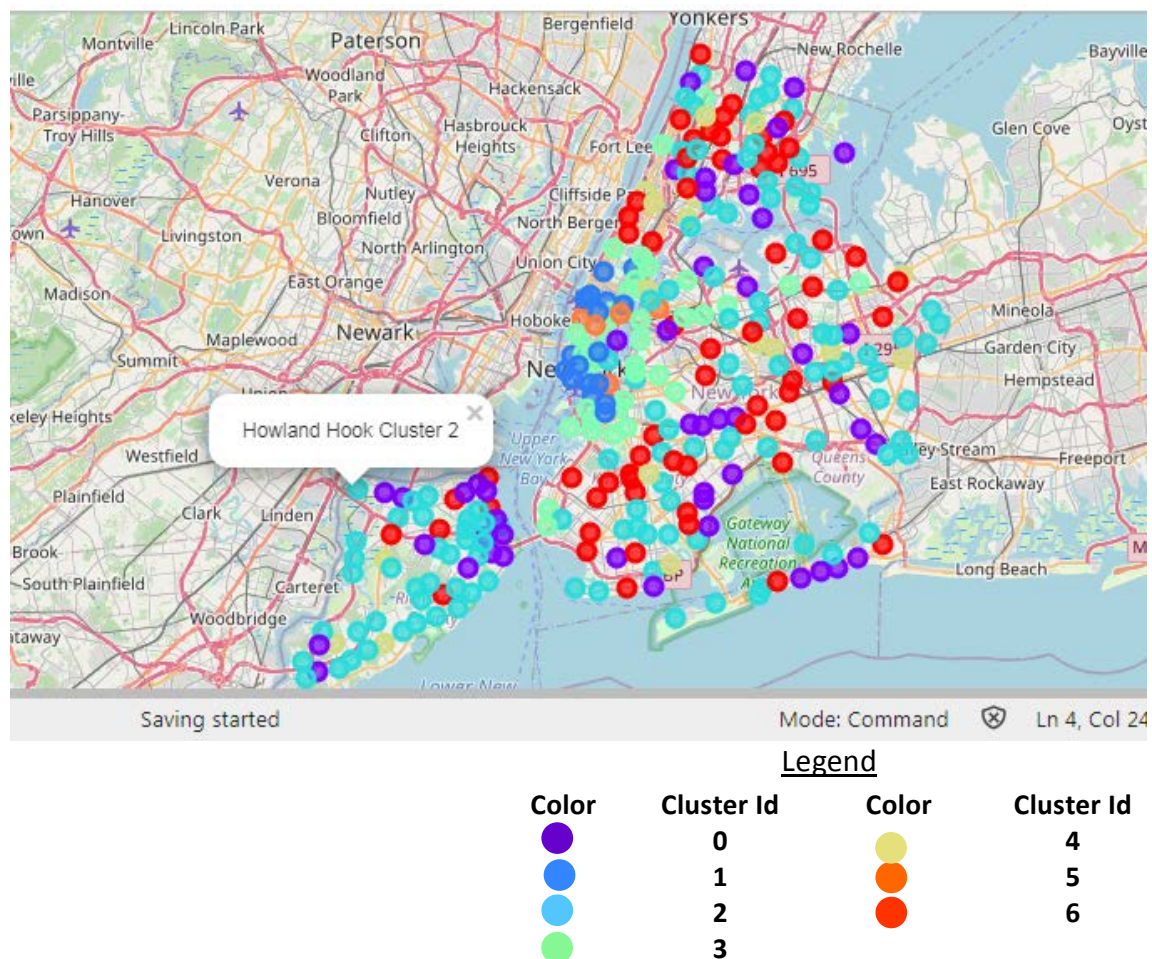
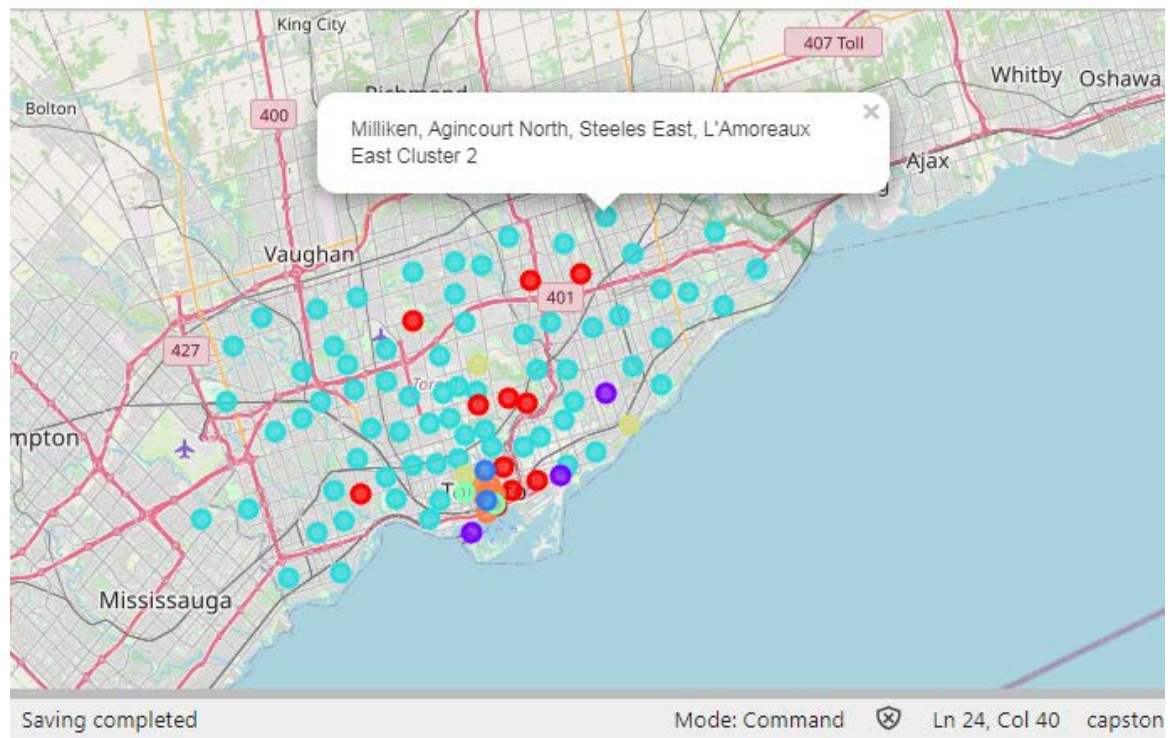


Fig. 7 Folium Map of New York City depicting the Clusters









Legend			
Color	Cluster Id	Color	Cluster Id
	0		4
	1		5
	2		6
	3		

Fig. 8 Folium Map of Toronto City depicting the Clusters

Findings: - Comparison of the two maps reveals the following: -

- New York has larger number of neighborhoods as compared to Toronto. This indicates that New York is bigger and more crowded.
- New York has a good mix of all the clusters. Whereas, Toronto has predominantly **Cluster 2** (marked by **Cyan** markers). This indicates that New York has more options in choosing neighborhoods suiting to different needs.

4.6.3 Quantitative analysis

- Topmost and least common venues in each cluster.** The database `combined_data_cluster` was used to compute frequencies of various venue types in each cluster and arrange them in decreasing order. The output was used to generate a Heatmap of Cluster Vs Venue Categories. The output is depicted in **Figure 9** below.

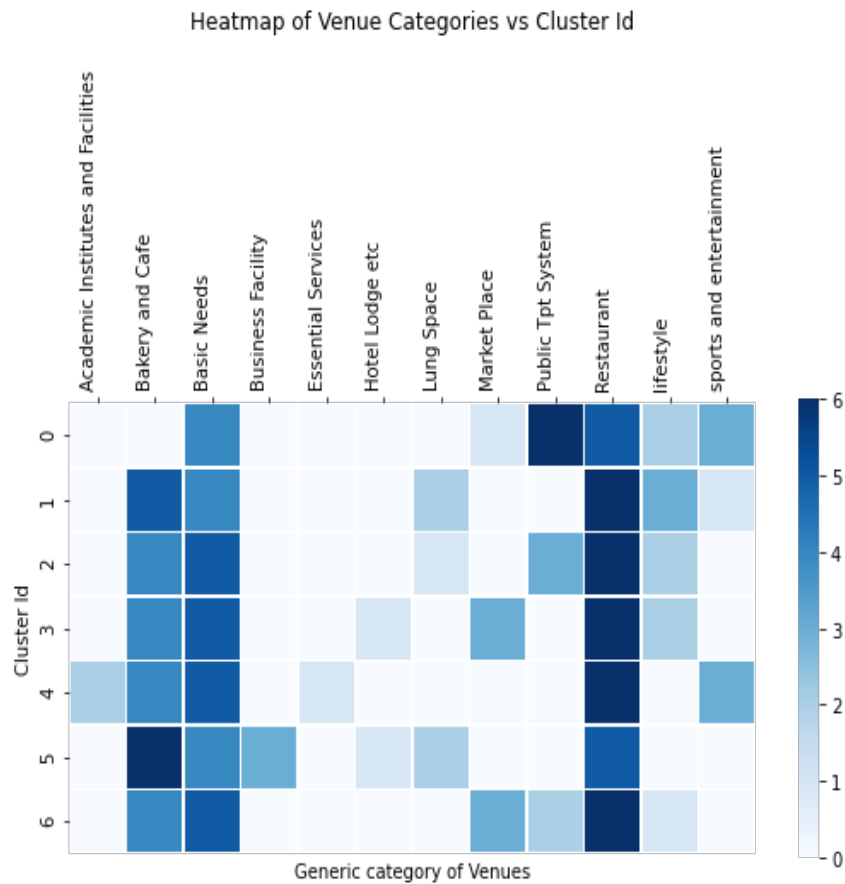


Fig 9. Heatmap of Topmost Venues in Different Clusters

Findings: -

The above Heatmap depicts the Venue Categories in shades of Blue - darkest shade (value 6) indicates **Topmost Common Venue** and lightest shade (score 0) indicates **Least Common Venue**.

- It is seen that **Restaurant, Basic Needs** and **Bakery and Cafe** are the most prominent venue categories.
- A rather curious and unexpected observation is that **Essential Services** and **Hotel Lodge etc** are the **Least Common Venues**. This is not the case in reality. The reason most probably is that the database in this project contains venues within 500 metre radius of the neighborhood centers. In most localities, these facilities would be located beyond 500 metres.

4.6.3.2 Clusterwise statistics of top venue distribution in the two cities. The database **combined_data_cluster** was split into New York and Toronto databases. For each city, **line plot** was generated - one for each cluster by joining the **frequency plot** of each venue type. The venue types were arranged in decreasing order as **1st Most Common Venue, 2nd Most Common Venue, ... 10th Most Common Venue**. The output graphs are depicted in **Figure 10** below.

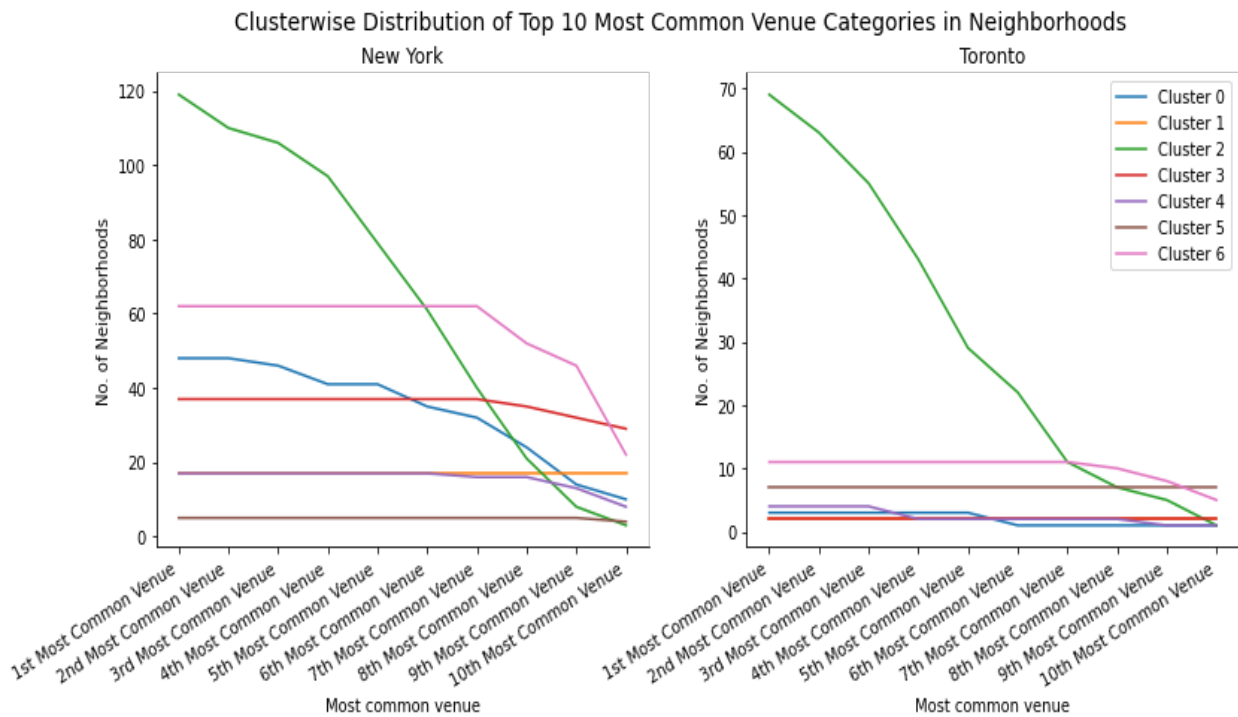


Fig 10. Cluster-wise Top Venue Distribution in the two Cities

Findings:

The graphs depict some interesting results, as brought out below: -

- Graph of **Cluster 2** drops sharply beyond 4th Most Common Venue. This indicates that most neighborhoods in this cluster have only four types of venues which are prominent.
- In both the cities, Cluster 2 contains the largest number of neighborhoods. This implies that the remaining clusters either have exclusive localities (better lifestyle) or not-so-developed ones.
- Toronto** has very few neighborhoods in **Cluster 0, 1, 3, 4 and 5** (in single digit). **Cluster 6** has 11 neighborhoods. Interestingly, most neighborhoods in all these clusters have all the ten types of venues included in the analysis.
- Likewise, most of the **New York** neighborhoods in **Cluster 1, 3, 4 and 5** have all the ten types of venues that were considered for the analysis.

4.6.4 Cluster to Feature mapping

This analysis is a **crucial** one - it is the culmination of all the analysis done so far. Objective of this analysis is to identify the **Features** present in each cluster. This would **enable one to recommend cluster (and neighborhoods in the cluster) for relocation, depending on the individual's needs.**

Computation was performed to find out **how many of the top 15 neighborhoods in each Feature belong to a certain cluster.** Assuming that each cluster is internally homogenous, these statistics would provide an indication of the cluster characteristics.

Results are depicted in **Figure 11** below.

Clusterwise Proportion of Neighborhoods in Top 15 of Each Feature

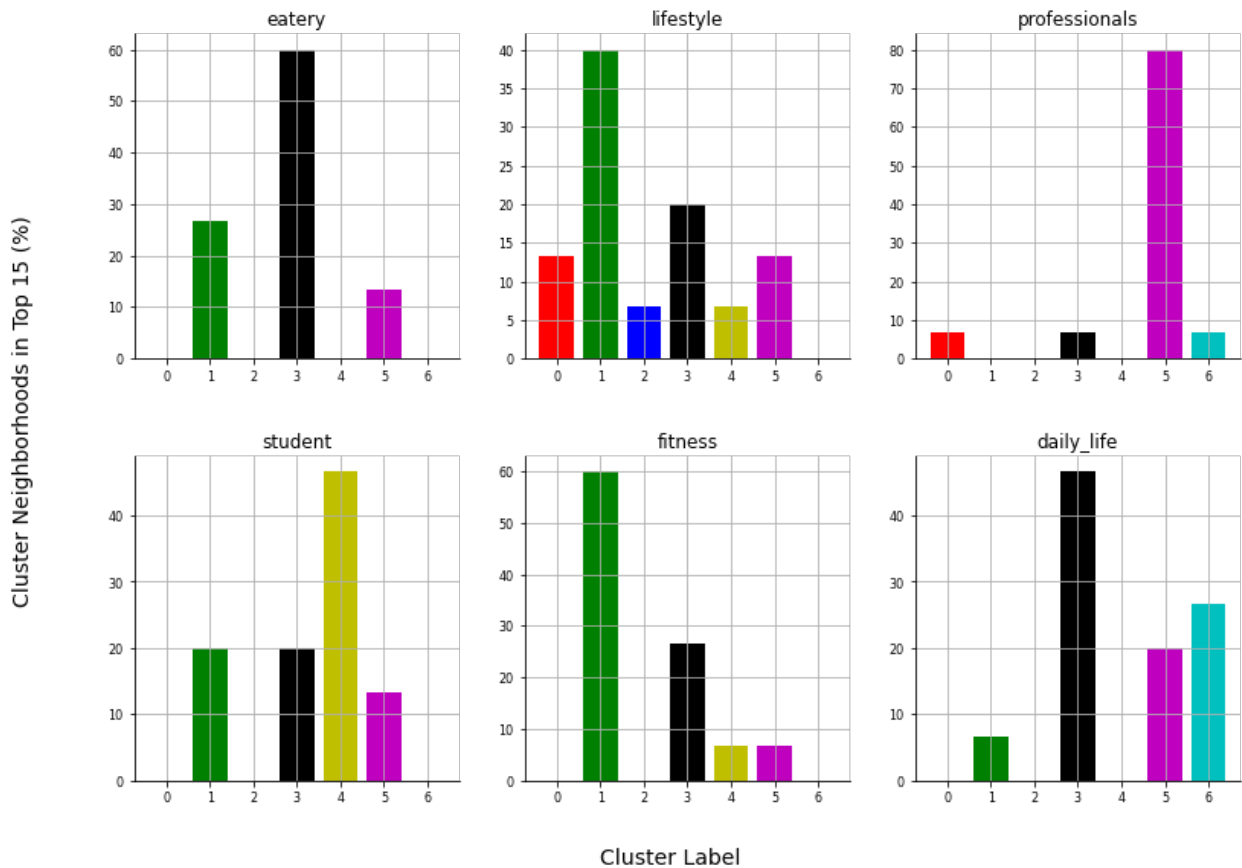


Fig 11. Cluster-to-Feature Mapping for Top 15 Neighborhoods in Each Feature

Findings:

The graphs depict proportion of neighborhoods out of the Top 15 list present in each cluster. Separate graphs have been plotted for each **Feature**. The following emerge from study of the graphs: -

- Cluster 1** has the highest proportion of **Top 15** neighborhoods in "**lifestyle**" and "**fitness**" Features
- Likewise, highest proportion of neighborhoods out of **Top 15** in "**eatery**" **Feature** and "**daily_life**" **Feature** are present in **Cluster 3**
- Cluster 4** has the highest proportion of **Top 15** neighborhoods in "**Student**" Feature
- Cluster 5** has the highest proportion of **Top 15** neighborhoods in "**Professionals**" **Feature**
- An interesting observation is that **Cluster 2** doesn't have any presence in the **Top 15** in any **Feature**, except "**lifestyle**" (only 6.67%). it was observed his implies that **Cluster 2** has moderate density of venues in all the categories, offering a **decent, affordable lifestyle and moderately crowded neighborhoods**.
- As was observed in the map, **Cluster 2 is the most prominent cluster in Toronto**.

5 Results and Discussion

5.1 The analysis was done in three levels: -

- a. **City-level:** Comparison of the two cities in terms of the venue types and venue density (within a radius of 500 metre from neighborhood centre)
- b. **Cluster-level:** Clustering, characterisation of each cluster and mapping them to the Features
- c. **Neighborhood-level:**
 - i. Comparison of neighborhoods in terms of **venue categories** and **venue density**
 - ii. Comparison of neighborhoods in terms of **Features**

5.2 Salient results obtained are summarised below. Based on the available venue data, it was possible to characterise neighborhoods, cities and clusters. The characterisation was helpful in classifying/ recommending the neighborhoods based on the relocation objectives (defined as **Features**).

5.2.1 City-level: -

- a. New York has large number of neighborhoods as compared to Toronto. Number of venues in New York are also much more than Toronto. These findings indicate that New York is bigger and more crowded
- b. New York has a good mix of all the clusters. Whereas, Toronto has predominantly Cluster 2 (marked by Blue markers). This indicates that New York has more options in choosing neighborhoods suiting to different needs.
- c. In both the cities, "Restaurant" is the most common venue among all generic categories.

5.2.2 **Cluster-level.** Clusters have been identified for relocation, based on the needs of the individuals. The suitability matrix is given in **Table 6** below.

Table 6. Cluster-to-Feature Mapping for Relocation

Cluster Id	Suitability for Relocation (Feature)
Cluster 0	lifestyle
Cluster 1	eatery, lifestyle, student, fitness
Cluster 2	average lifestyle
Cluster 3	eatery, lifestyle, student, fitness, daily_life
Cluster 4	student
Cluster 5	lifestyle, professionals, daily_life
Cluster 6	daily_life

5.2.3 Neighborhood-level: -

- e. **"Restaurant", "Basic Needs" and "Bakery and Cafe"** are the **three most common venues** in majority of the neighborhoods. Availability of these venues as the most common ones indicate that such neighborhoods are well-populated residential areas. Preferred by those who would relocate for jobs and decent living.
- f. Neighborhoods with most common venues **"Lifestyle"** and **"Lung Space"** are the high-cost-of-living areas. Suitable for those looking to relocate for enjoying a relaxed and premium lifestyle.
- g. The neighborhoods with **"Budget Food Joint"** as the most common venue offers affordable living. Such neighborhoods would be preferred by students for relocation.
- h. Venue categories **"Bakery and Cafe", "Bar Pub etc", "Basic Needs", "Health and Wellness"** and **"Restaurant"** have **high correlation** (> 0.6). This implies that neighborhoods which have any of these categories as the most common venues, are likely to have the other categories as well.
- i. Database of Top 15 Neighborhoods suited for relocation in each need category (i.e., Feature) are : -

**[eatery_15, lifestyle_15, professionals_need_15, student_needs_15, fitness_15,
daily_life_15]**

- j. There are outliers which do not fit into any relocation criteria due to lack of the required features.

5.3 Discussions

There was good coherence between **Exploratory Data Analysis** and **Clustering** results. Both analysis established that **New York has more diversity of neighborhood types as compared to Toronto**. The Clustering algorithm created clusters based on the needs of the individuals (defined as Features in the project).

However, **clusters were not entirely distinct**; there are overlaps of multiple Features. This is expected, because, the various needs of relocation cannot be totally isolated from each other. For example, an individual looking to relocate for better **"quality of life"** would also look for **"daily needs", "restaurant"** and **"health facilities"**. Of course, he / she would not like if these are present in excessive numbers, thereby making the neighborhood crowded and noisy,

For the data gathered through **Foursquare** in this project, **the elbow-method graph did not have well-defined elbow**. As a result, the number of clusters may not be the most optimum. This would affect formation of well-defined clusters which are distinct from each other. Improving the clustering result may be possible if the radius around the neighborhood centre is increased for venue analysis or alternate algorithms are chosen for clustering. However, this was not attempted as it was out of the scope of the project.

5.4 Use of Foursquare Database for Clustering.

One needs to be careful about using Foursquare app to generate real-time data for the analysis. This is because, the data is dynamic. Each day, new venues get added up, some existing venues may get deleted or modified. It was observed that a fresh computation with live Foursquare database leads to minor modifications in the results. When the project was applied on a fresh database taken after three months, the clustering algorithm allotted different Cluster Ids to few neighborhoods due to change in venues. However, the major findings and analysis remained more or less the same.

6 Conclusion

- 6.1 The objective of this project was to evolve a method to identify distinct characteristics of the neighborhoods in the two cities, based solely on details of venues in the neighborhoods. Once such characteristics were established, the neighborhoods were to be categorised for relocation such that each category fulfills a definite set of needs (such as better quality of life, living within a budget etc).
- 6.2 Towards this end, we have been successful to evolve metrics which can be useful to categorise the neighborhoods. We were also able to draw indirect inferences as to the cost of living, whether the residents are high income group etc., without even referring to the Census data.
- 6.3 The results of the project can be useful in many ways, beyond just making a decision on relocating to a neighborhood (which was the main objective of the project). One such example is that the outliers which do not have adequate "eatery", "lung space" or "health facility" etc can be taken up for development and provisioning of public amenities. Another utility of the project results is that, a businessperson can target such outlier neighborhoods for setting up his/her business.
- 6.4 Success of the approach adopted in the project would depend to a large extent on the venue database and how the appropriate attributes for clustering are chosen. Another aspect is the choice of the Clustering algorithm (e.g. DBSCAN). The implementation can be made more robust through experimentation in these regards.
- 6.5 The database generated by Foursquare is dynamic. Hence, the venue database instance created using Foursquare app needs to be "frozen" so as to obtain consistency in the results.