A Tale of Two Cities through ML Techniques

Using Exploratory Data Analysis and Clustering

People Relocate

- Globalization resulted in movement of people for **relocating**
- Appropriate destination is a key factor which depends on **purpose** of relocation
- Evolving a tool for ascertaining appropriate **neighborhood** will help people in
  - Comparing the destination neighborhood with the current one
  - Choose neighborhood meeting relocation purpose
- People interested in the solution include
  - People migrating for a similar or better quality of life
  - Professionals relocating to new place of work
  - International students
  - Businesspersons looking for new market
- Cities selected for project: **New York** and **Toronto**

Data Acquisition and Cleaning

1. **Neighborhood database**

- Toronto neighborhoods’ data scraped from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- Latitude, Longitude data for Toronto neighborhoods retrieved from http://cocl.us/Geospatial_data.
- New York neighborhoods’ data scraped from https://cocl.us/new_york_dataset (courtesy: "Segmenting and Clustering Neighborhoods in New York City" Wk 3 Lab Exercise)
- Toronto and New York databases were merged to create a master database of neighborhoods
- Records with **Borough not assigned** were dropped, **Neighborhood not assigned** were assigned borough name, duplicate entries due to name truncation were corrected
- Cleaned master database contained 409 records of neighborhoods

Data Acquisition and Cleaning

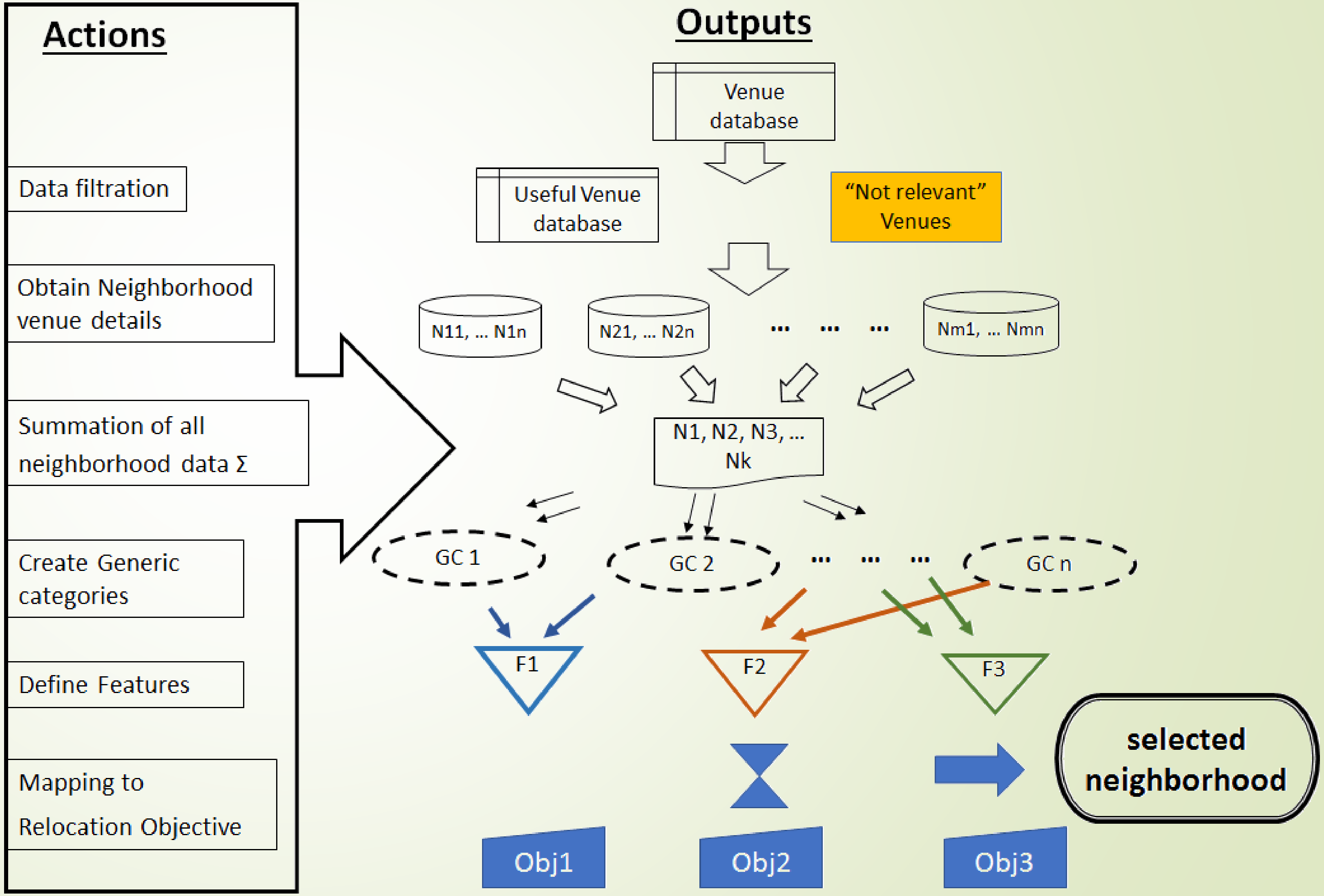
2. **database of venues in neighborhoods**

- For each **neighborhood** in the combined database, details of all **venues** within **500 m radius** were obtained using **Foursquare** app.
- Venue details include **name, latitude/longitude** and **category** of the venue.
- Raw database contained 12288 records
- Venues with categories mentioned as neighborhood were dropped
- Cleaned database has 12279 records
- More than 462 Venues categories generalized to 16 categories
- 42 categories of Venues found not relevant for the project were ignored

Neighborhood selection using Venue-based Features

- Venue **categories** mapped to **Features**
- Features used to characterize neighborhoods
- **Relocation objectives** were **mapped** to **neighborhood features**

Objective 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



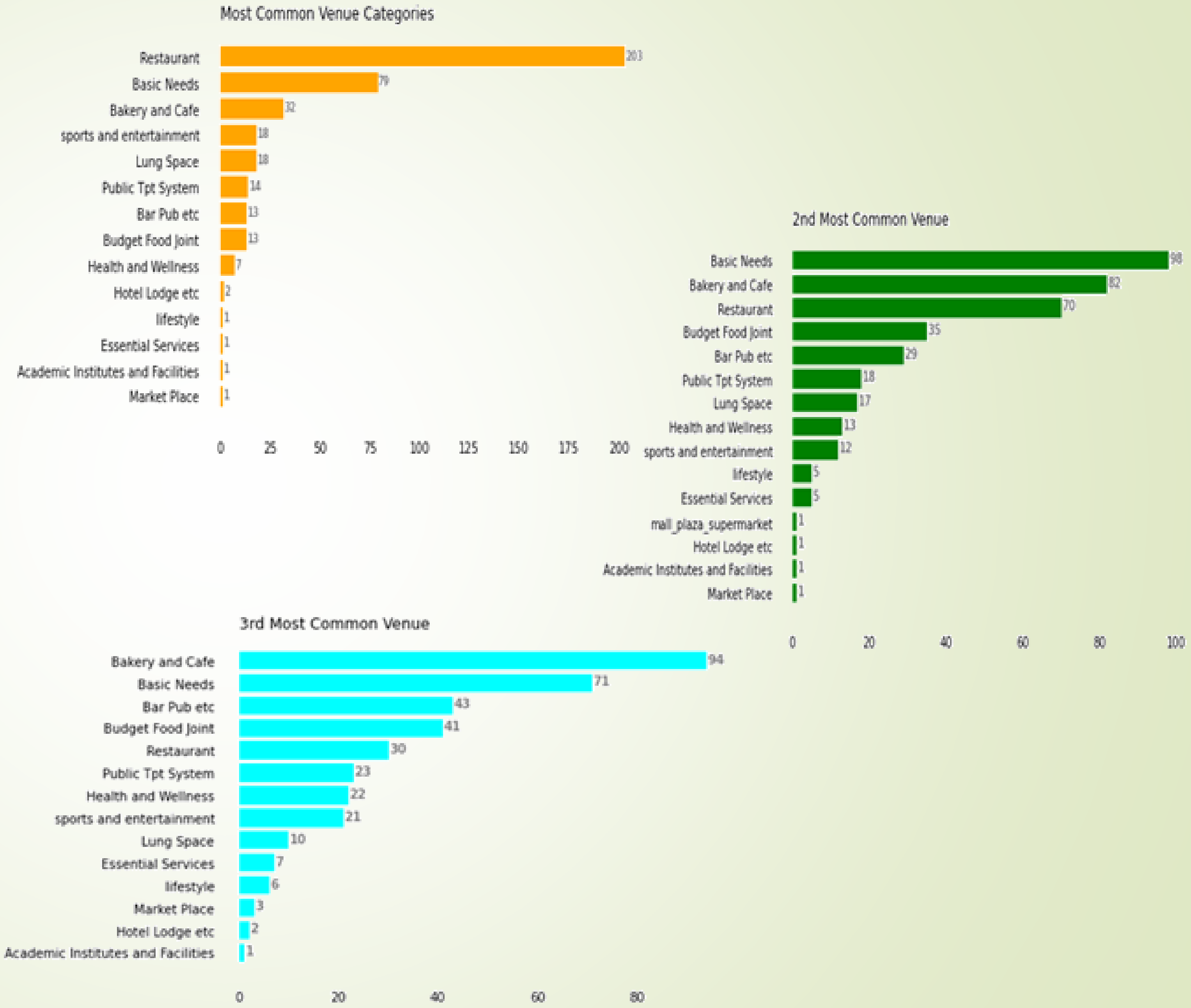
Actions

- Data filtration
- Obtain Neighborhood venue details
- Summation of all neighborhood data Σ
- Create Generic categories
- Define Features
- Mapping to Relocation Objective

Outputs

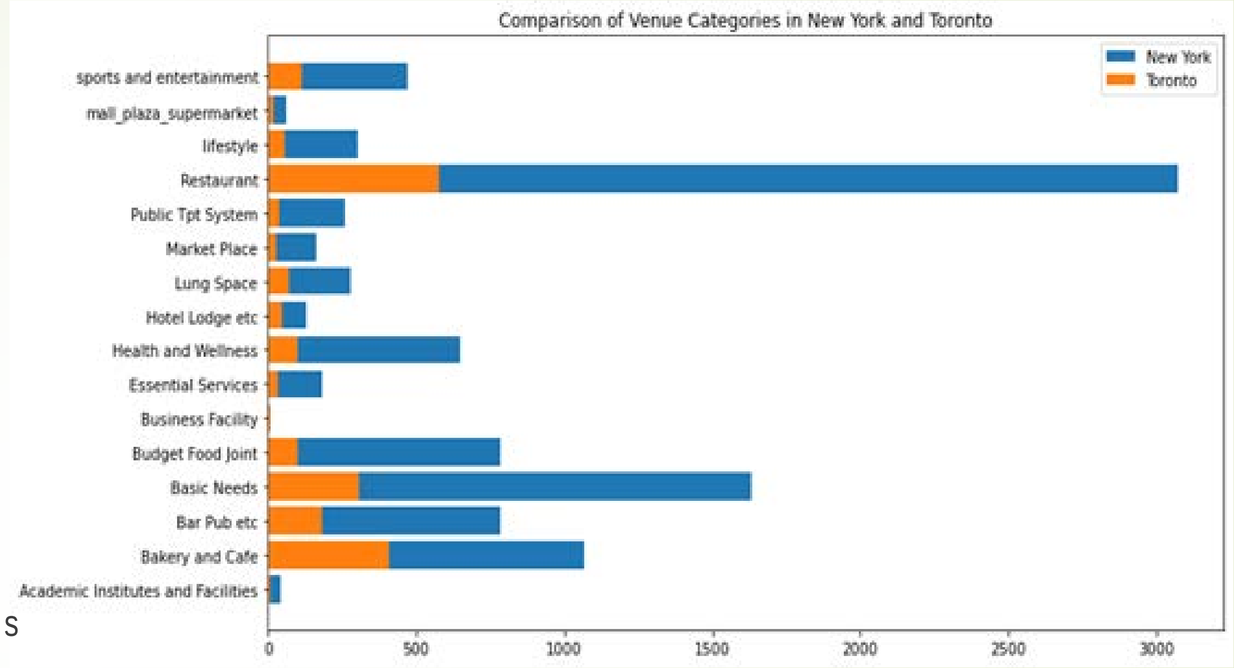
Most common venue categories

- **"Restaurant"**, **"Basic Needs"** and **"Bakery and Cafe"** are **Top three venues** in neighborhoods
- Neighborhoods with **"Lifestyle"** and **"Lung Space"** venues are high-cost-of-living areas
- Neighborhoods with **"Budget Food Joint"** venues offer affordable living
- **Principal Component Analysis** shows "Bakery and Cafe", "Bar Pub etc", "Basic Needs", "Health and Wellness“ and "Restaurant“ have high correlation
- Neighborhoods which have any of these venue categories likely to have venues of the remaining categories

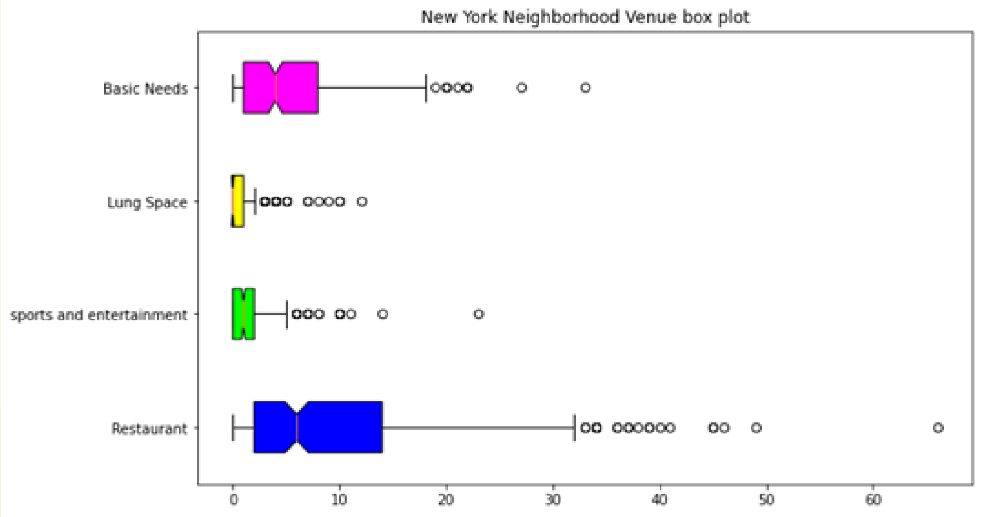


New York and Toronto data compare

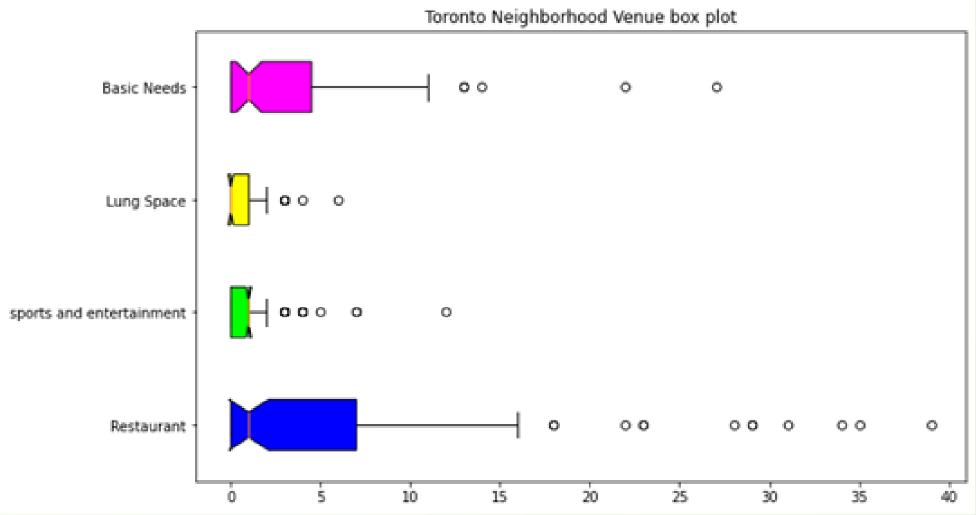
- Venues in New York are much more than Toronto.
- Relative proportions of various generic category venues similar in both cities
- Restaurant is the most common venue
- Majority of the neighborhoods do not have **"Lung Space"** and ***sports and entertainment**” (within 500 m radius)
- Venue concentration in Toronto neighborhoods are much lower than New York



Comparison of Venue Categories in New York and Toronto



New York Neighborhood Venue box plot



Toronto Neighborhood Venue box plot

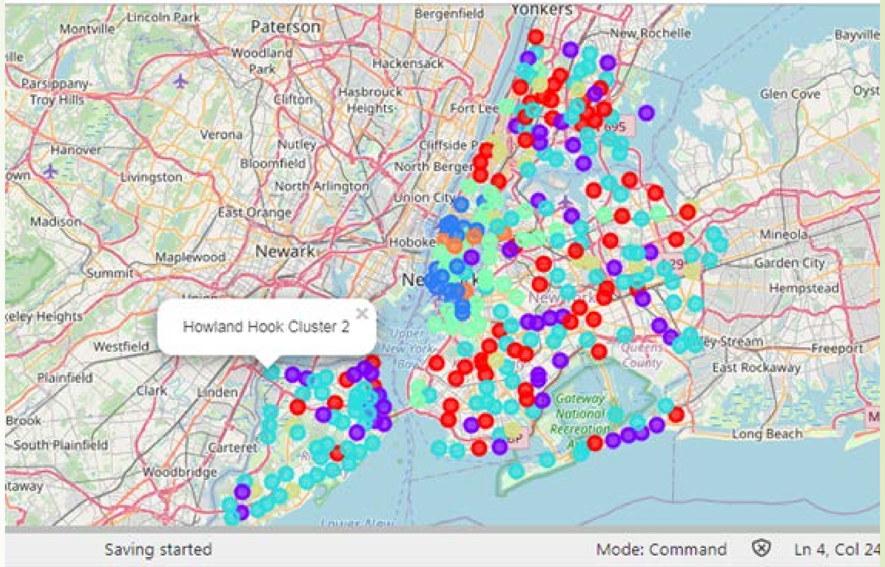
Neighborhood analysis based on Features

- Neighborhoods **Chelsea** and **Flatiro**n are among the top 15 in **Professional** and **student** features
- Neighborhoods **Fordham, Boerum Hill** and **Flatiron** are among the **top 15** in **daily_life** and **fitness** features
- **Financial District** neighborhood is among the top 15 in **lifestyle** and **eatery**
- There is no neighborhood which is in **top 15** in all features
- There are **Outliers** – not recommended as a relocation destination
  - Neighborhoods with very low density of **eatery venues** (zero within 500 m radius)
  - Neighborhoods,with very less venues of **"Lung Space“, "lifestyle" and daily_life** features

Feature	Definition
Eatery	Aggregation of "Restaurant", "Bakery and Cafe", "Budget Food Joint" and "Bar Pub etc“ venues
Lifestyle	Set of "Lung Space" and "lifestyle" venues
Daily Life	Aggregation of "Basic Needs", "Essential Services", "Market Place" and "mall_plaza_supermarket“ venues
Fitness	Set of "Health and Wellness" and "sports and entertainment" venues
Student	Set of "Academic Institutes and Facilities", "Basic Needs", "Budget Food Joint" and "sports and entertainment“ venues
Professional	Set of "Business Facility", "Hotel Lodge etc" and "Public Tpt System" venues.

Neighborhood Clustering

- Database clustered into seven clusters using **k-means Clustering**
- Scaling of database performed
- Elbow Method used for selecting **knn**
- Used **Folium** to create maps with clusters
- New York has larger number of neighborhoods as compared to Toronto. This indicates that New York is bigger and more crowded.


Howland Hook Cluster 2

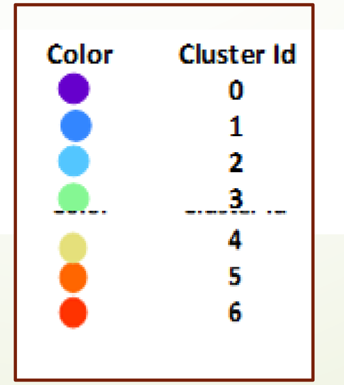
Saving started   Mode: Command   Ln 4, Col 24

Folium Map of New York City depicting the Clusters

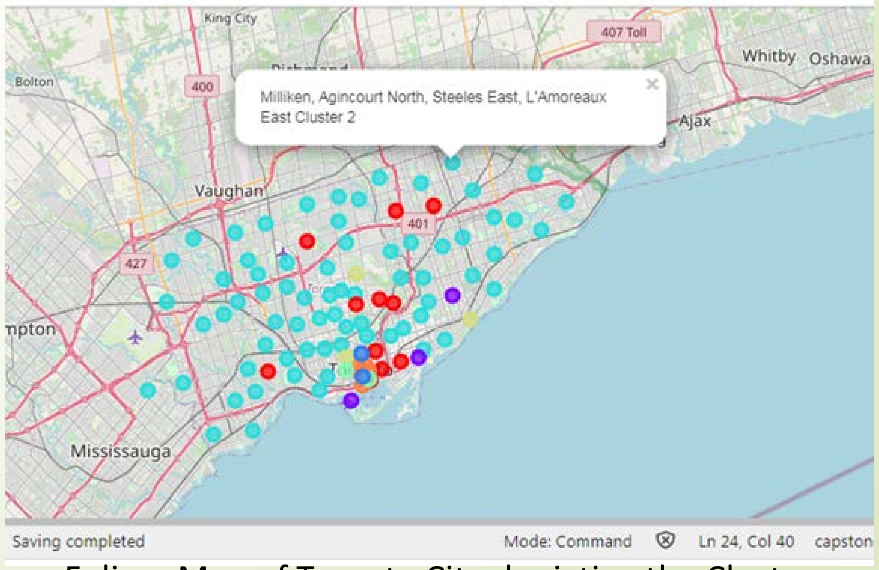
Cluster results

- New York has a good mix of all the clusters. Whereas, Toronto has predominantly **Cluster 2**
- New York has more options in choosing neighborhoods suiting to different needs

Clusters



Color	Cluster Id
🟣	0
🔵	1
🔵	2
🟢	3
🟡	4
🟠	5
🔴	6

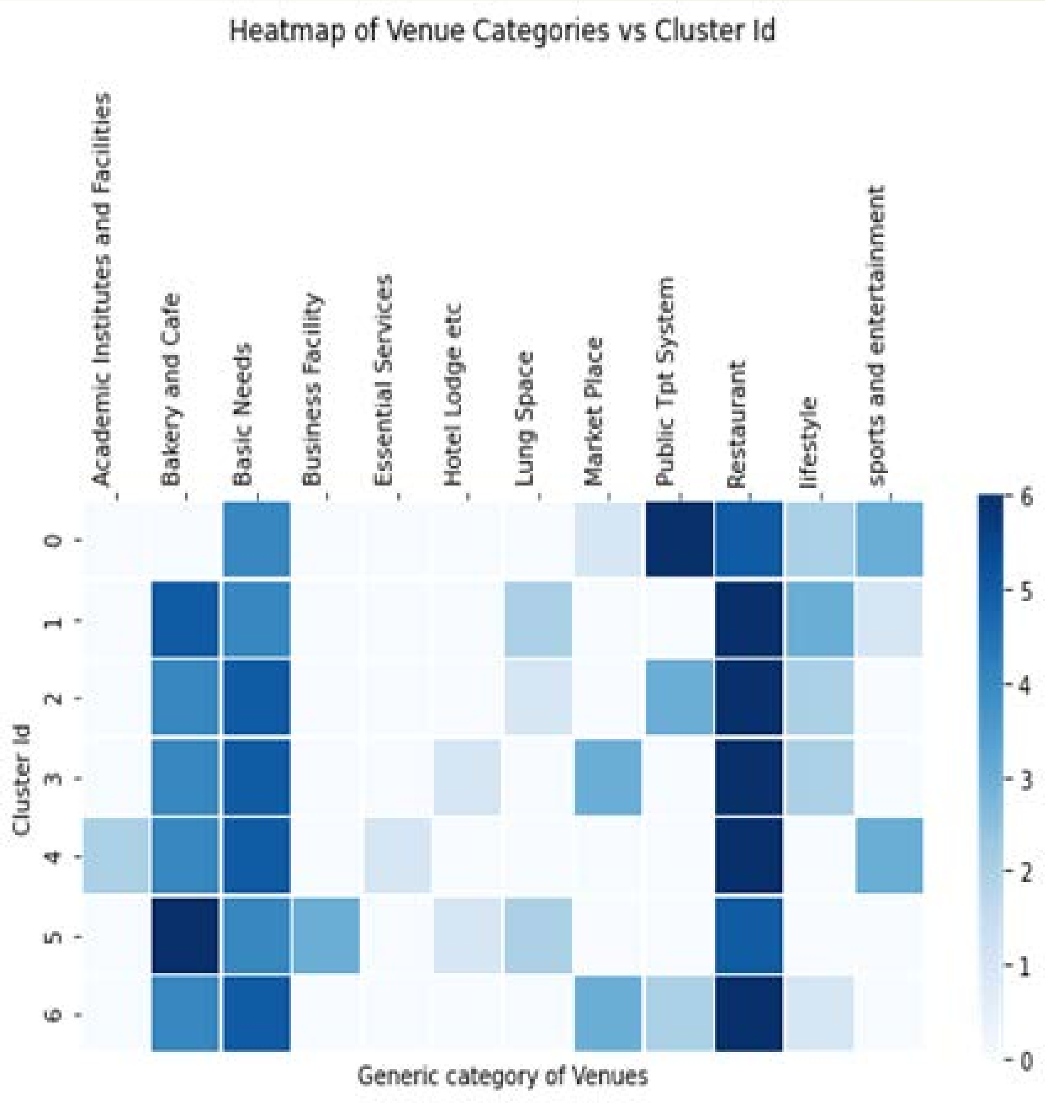

Milliken, Agincourt North, Steeles East, L'Amoreaux East Cluster 2

Saving completed   Mode: Command   Ln 24, Col 40   capston

Folium Map of Toronto City depicting the Clusters

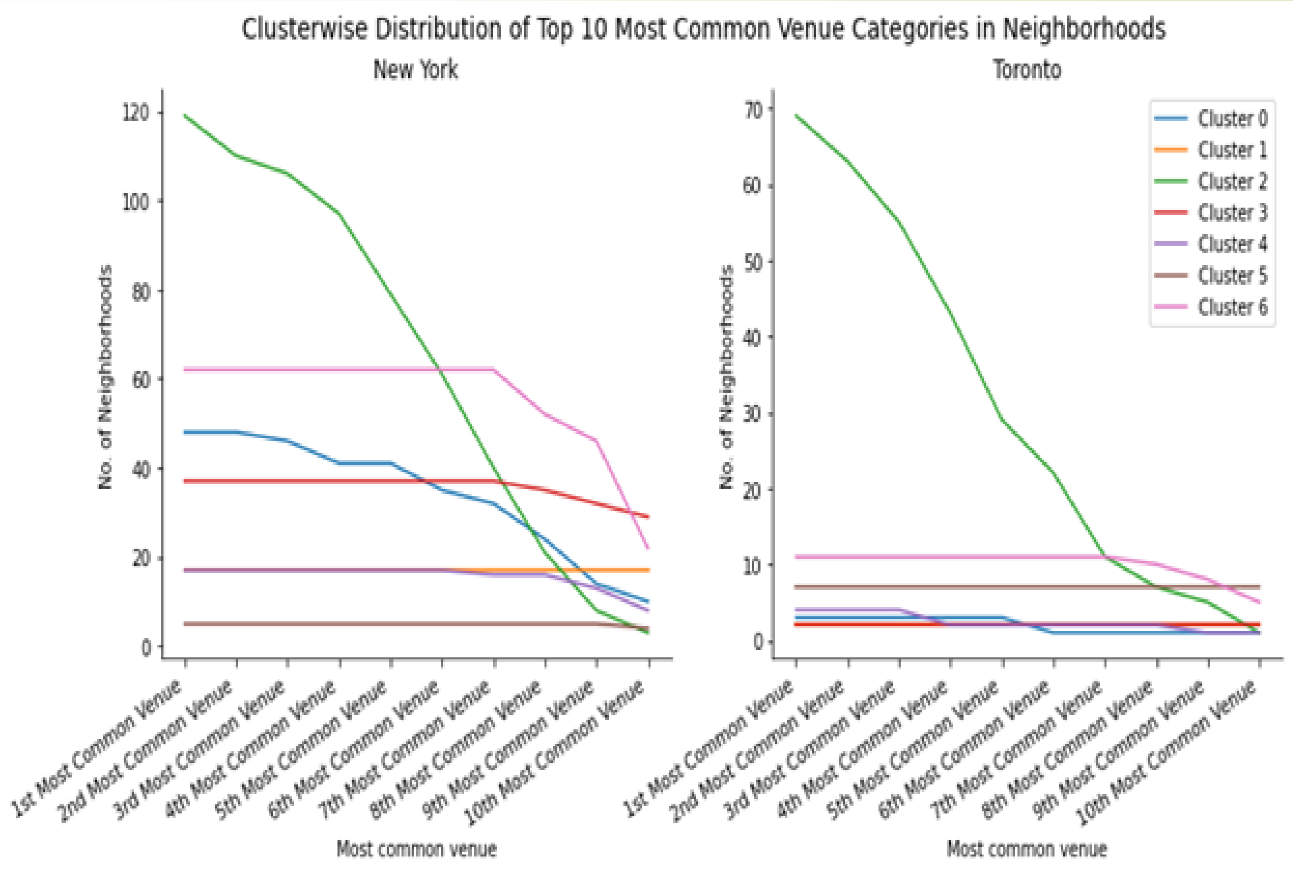
Topmost and least common venues in each cluster

- **Restaurant**, **Basic Needs** and **Bakery and Cafe** are most prominent venue categories
- **Essential Services** and **Hotel Lodge etc** are the **Least Common Venues**.
  - Contradicts reality
  - Could be due to restriction of 500 metre for obtaining venue details



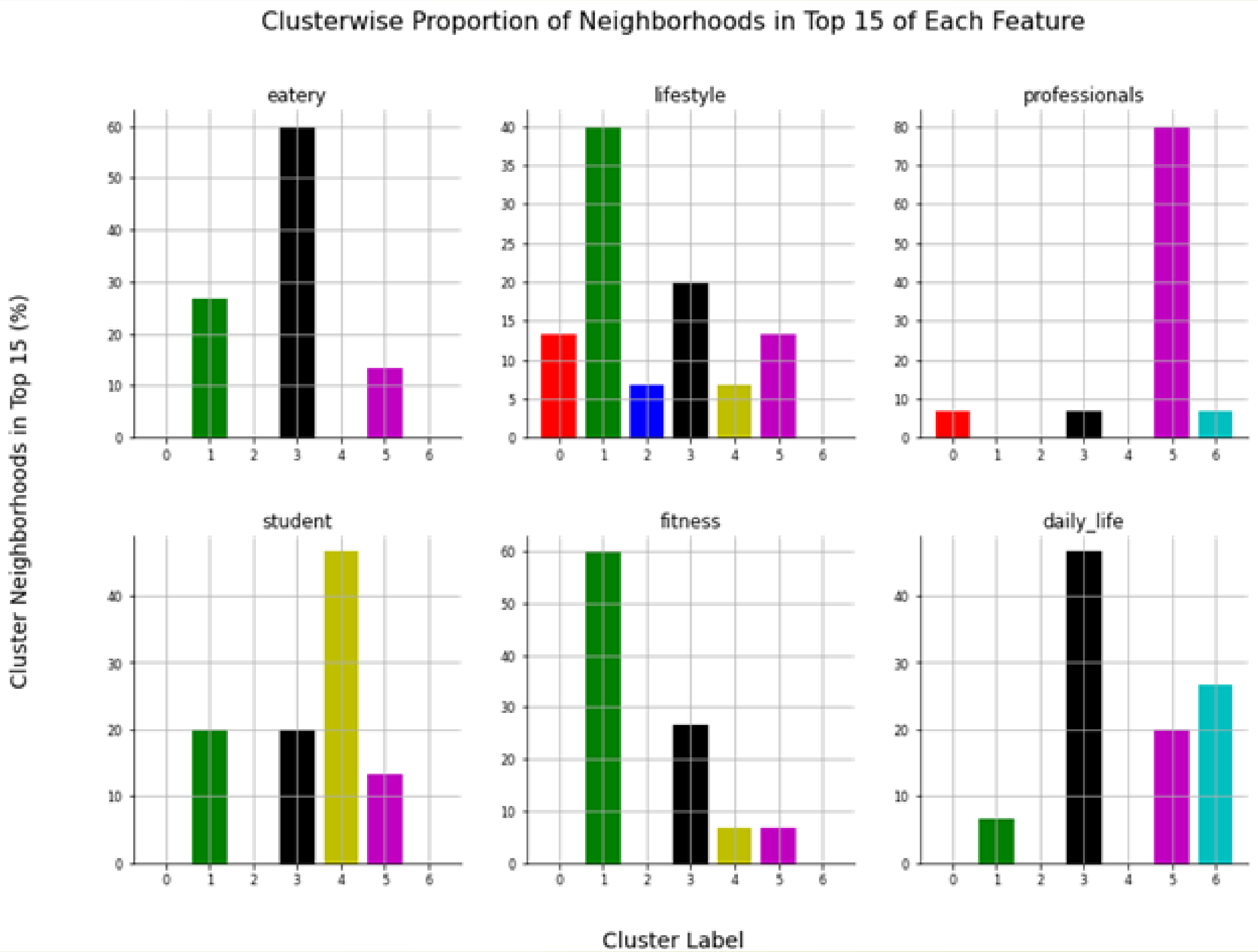
Cluster-wise statistics of top venue distribution in the two cities

- Most neighborhoods in **Cluster 2** have only four prominent venue categories
- In both cities, Cluster 2 has largest number of neighborhoods. Remaining clusters either have exclusive localities or not so developed
- **Toronto** has very few neighborhoods in **Cluster 0, 1, 3, 4 and 5** (in single digit). **Cluster 6** has 11 neighborhoods
- Most neighborhoods in all these clusters have all the ten types of venues included in the analysis
- 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



Cluster to Feature mapping

- **Cluster 1** has highest proportion of **Top 15** neighborhoods in **"lifestyle"** and **"fitness"** Features
- **Cluster 3** has highest proportion of neighborhoods out of **Top 15** in **"eatery"** Feature and **"daily_life" Feature**
- **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**
- **Cluster 2** has moderate density of venues and offers decent, affordable lifestyle and moderately crowded neighborhoods



Clustering summary

- good coherence between **Exploratory Data Analysis** and **Clustering** results
- Clusters have overlaps of multiple Features. This is expected since basic human needs will be common across various relocation objectives

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

Conclusion and future directions

- Project objectives successful
  - To identify neighborhood characteristics based on venue details
  - Categorise neighborhoods such that each fulfills a definite set of needs (such as better quality of life, living within a budget etc)
  - Drew indirect inferences to cost of living, high income group residents etc
- Database generated by Foursquare is dynamic
  - Venue database instance to be “frozen” for consistency in results
- Project results useful in many ways
  - Decision-making on relocating to a neighborhood
  - Outliers can be taken up for development and provisioning of public amenities
  - Businessperson can target outlier neighborhoods for business
- Project scope can be enhanced
  - Larger venue database
  - Choice of attributes for clustering
  - Choice of the Clustering algorithm (e.g. DBSCAN)