# The battle of Neighborhoods: Investigate the similarity or dissimilarity of New York and Toronto City

Dejene Techane

December 28, 2020

## 1. Introduction

## 1.1 Identifying the Business Problem(Introduction)

New York and Toronto are the most densely populated areas in the USA and Canada. Both cities are very diverse and are financial capitals of their respective countries. The problem description is how could we determine the similarity or dissimilarity of New York and Toronto by comparing their neighborhoods. Is New York City more like Toronto city?

To answer these questions, we were clustered the cities into groups of similar items. We can then use the results of this analysis to understand the similarities and differences of neighborhoods of these cities by identifying the most common venue category.

The objective of this project is to leverage Foursquare API and machine learning to determine the similarities and differences of these cities by comparing their neighborhoods. We have to use Foursquare location data and a clustering algorithm to explore and identify the most common venue by grouping into similar items by their category venues and neighborhood information.

## 2. Description of the Data

Getting data is one of the parts of this project. We have used different sources of data to address this problem. We can get the list of Postal Codes of Canada[2] and New York City[3]. The data is unstructured and needs to be scrapped and cleaned to make it structured and ready for data analysis.

- We can also get geographical coordinates of neighborhoods[4] for getting latitude and longitude data for Toronto City neighborhoods.
- Another source of data is the Foursquare Database. We will use the Foursquare API[4] to get all venues. We will filter out all venue category data from JSON files into structured and tabular forms.

Now that we have equipped with the data we want to use and the tools to use location data to explore a geographical location to get venue information of both cities.

# 3. Methodology

## 3. 1 Data Preparation and Cleaning

Data preparation is one of the first methodologies that is the process of collecting data from different sources. In this case, I have downloaded and scrapped the Toronto and New York data that contains boroughs and neighborhoods along with latitude and longitude data from Wiki, cleaned, combined, and transformed it into a dataframe using Pandas. I have built a dataframe with neighborhood, borough, latitude and longitude information of both Toronto and New York. The following shows the results of the cleaned data of both Toronto and New York.

| | Postal Code | Borough | Neighbourhood | Latitude | Longitude |
|---|---|---|---|---|---|
| 0 | M1B | Scarborough | Malvern, Rouge | 43.806686 | -79.194353 |
| 1 | M1C | Scarborough | Rouge Hill, Port Union, Highland Creek | 43.784535 | -79.160497 |
| 2 | M1E | Scarborough | Guildwood, Morningside, West Hill | 43.763573 | -79.188711 |
| 3 | M1G | Scarborough | Woburn | 43.770992 | -79.216917 |
| 4 | M1H | Scarborough | Cedarbrae | 43.773136 | -79.239476 |

Table 1.1 Neighborhoods of Toronto City

| | Borough | Neighborhood | Latitude | Longitude |
|---|---|---|---|---|
| 0 | Bronx | Wakefield | 40.894705 | -73.847201 |
| 1 | Bronx | Co-op City | 40.874294 | -73.829939 |
| 2 | Bronx | Eastchester | 40.887556 | -73.827806 |
| 3 | Bronx | Fieldston | 40.895437 | -73.905643 |
| 4 | Bronx | Riverdale | 40.890834 | -73.912585 |

Table 1.2 Neighborhoods of New York City

**Folium** is a great Python based visualization tool that is used to visualize the geographic details of Toronto and New York neighborhoods and their boroughs on a map. As shown below, I created the map of Toronto and New York using latitude and longitude with neighborhoods superimposed on top.

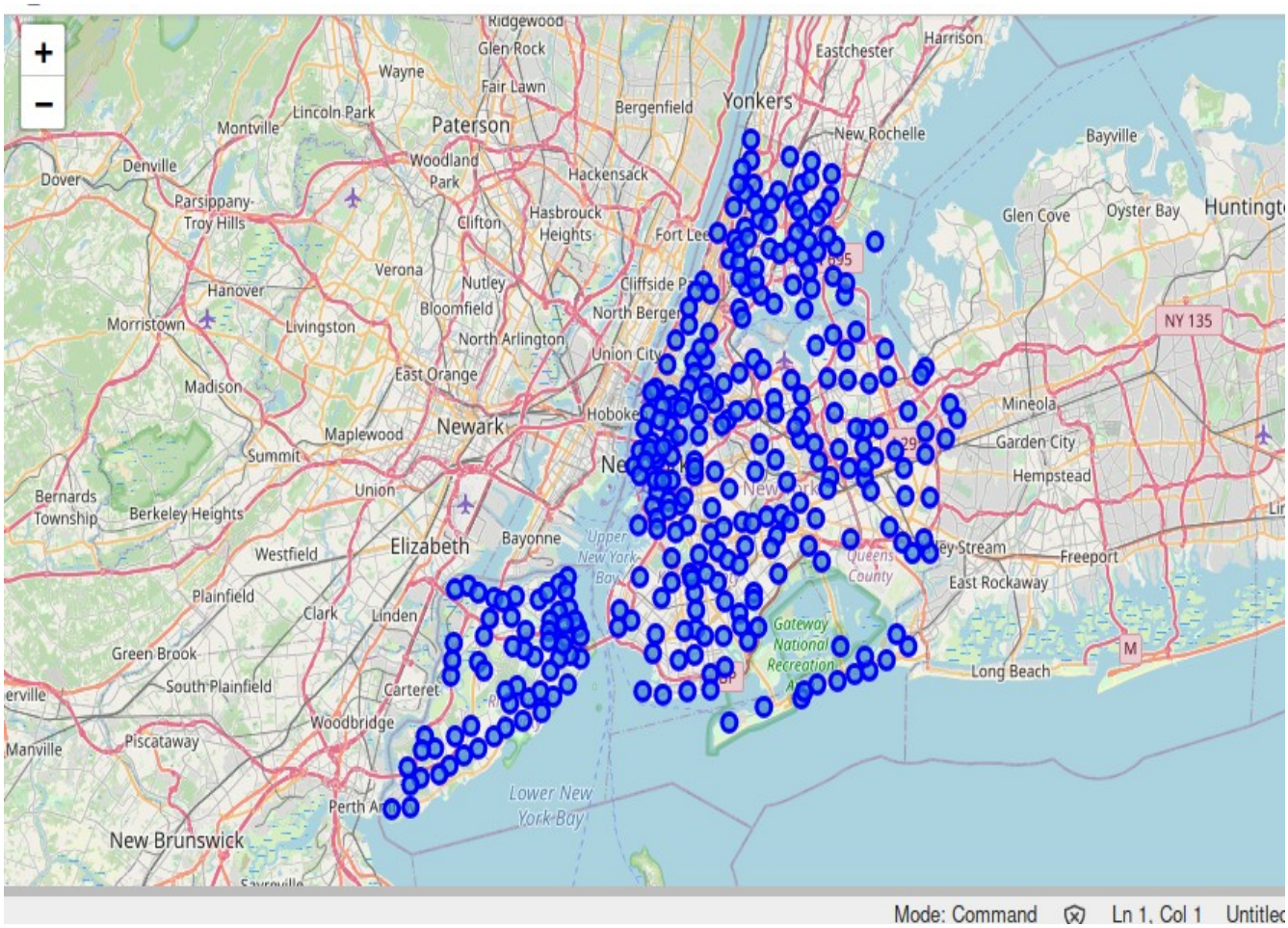


Fig 1.1 Visualization of New York Neighborhoods without Clustering

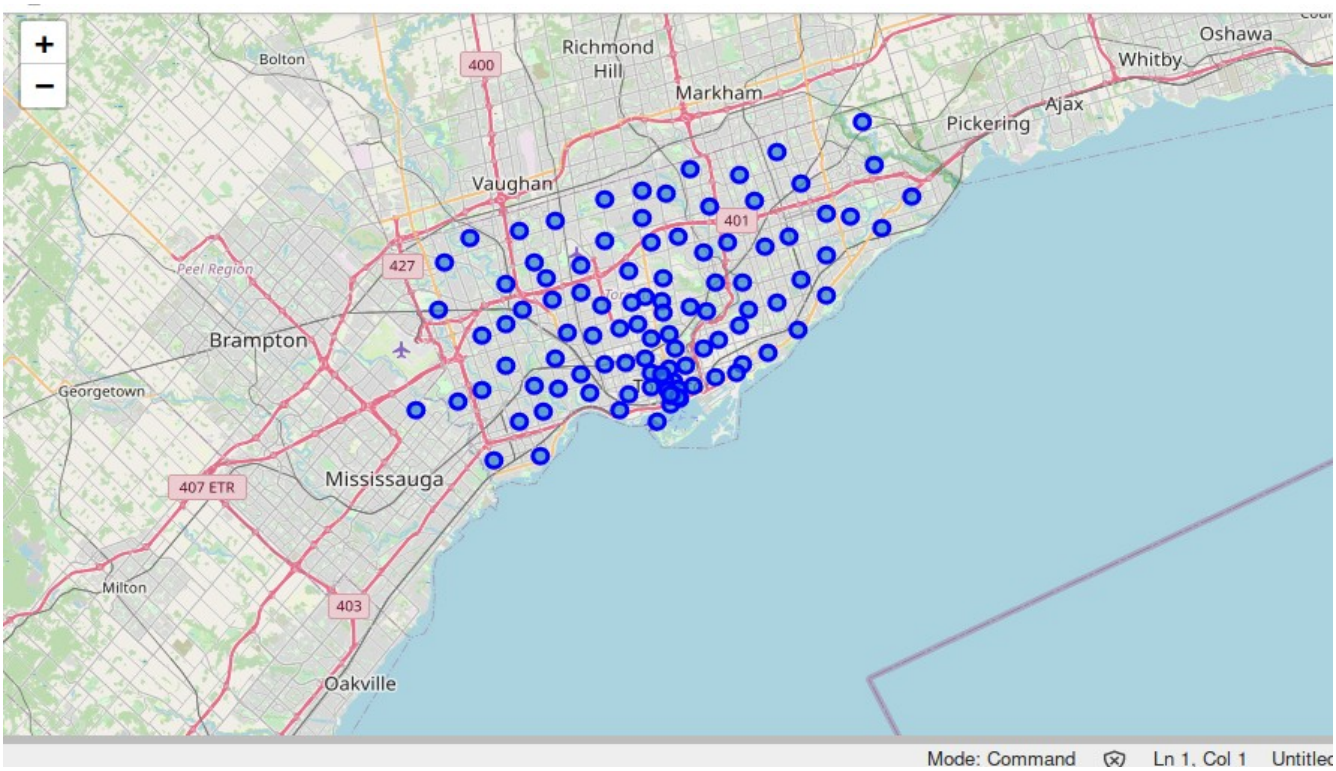


Fig 1.2 Visualization of Toronto Neighborhoods without Clustering

I utilized the Foursquare API to explore the boroughs of each neighborhood and segment them. I designed the limit as **100 venues** and **a 500-meter radius** for each borough from their given latitude and longitude information. The data we obtained from Foursquare API is in the form of JSON file format, and then extracted all the venue information from the JSON file and transformed it into pandas dataframe which contains the list of venues name, category name, latitude, and longitude information as shown below.

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

Table 1. 3 Venues information of Toronto City

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

Table 1.4: Venues information of New York City

Now, we have prepared the data and made ready for data analysis. The next step is data analysis.

## 3.2 Exploratory Data Analysis(EDA)

EDA is one of the data science methodologies that can be performed to get a better understanding of the venue data. In this stage, we found the most common and widespread venue categories information in New York and Toronto. To have a better understanding of the difference between most common and widespread: for example, if we have found 186 venues with the category of "Coffee Shop" and these venues exist in 48 neighborhoods out of 50 neighborhoods;  also, there are 49 venues with the "Pizza Palace" category and these venues exist in 34 neighborhoods; each one of them in a different neighborhood. That means the "Coffee Shop" category is the most common than the "Pizza Palace" category because; there are more venues under this category. Also "Coffee shop" venue category is most widespread than the "Pizza Palace" category because; there are more neighborhoods under this category.

Before we get started EDA, a data preparation was performed, and venues such as Building, Office, Bus Line, Bus Station, Bus Stop, or Road were excluded in this analytical process because these are not added analytical value to this project. The following result shows that the most common and widespread venue categories in New York and Toronto City.

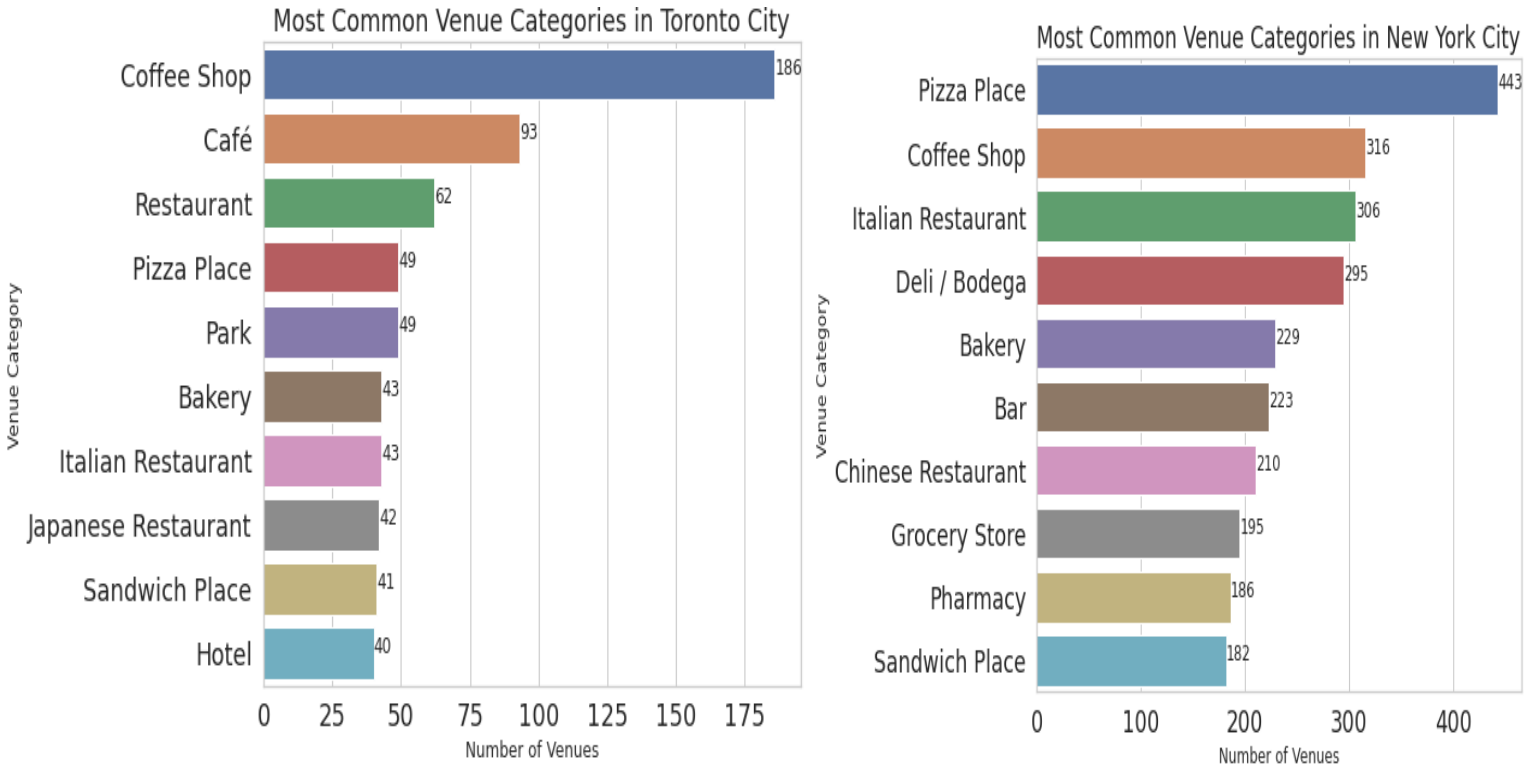


Fig 1.3 Most Common Venue Categories in Toronto and New York City

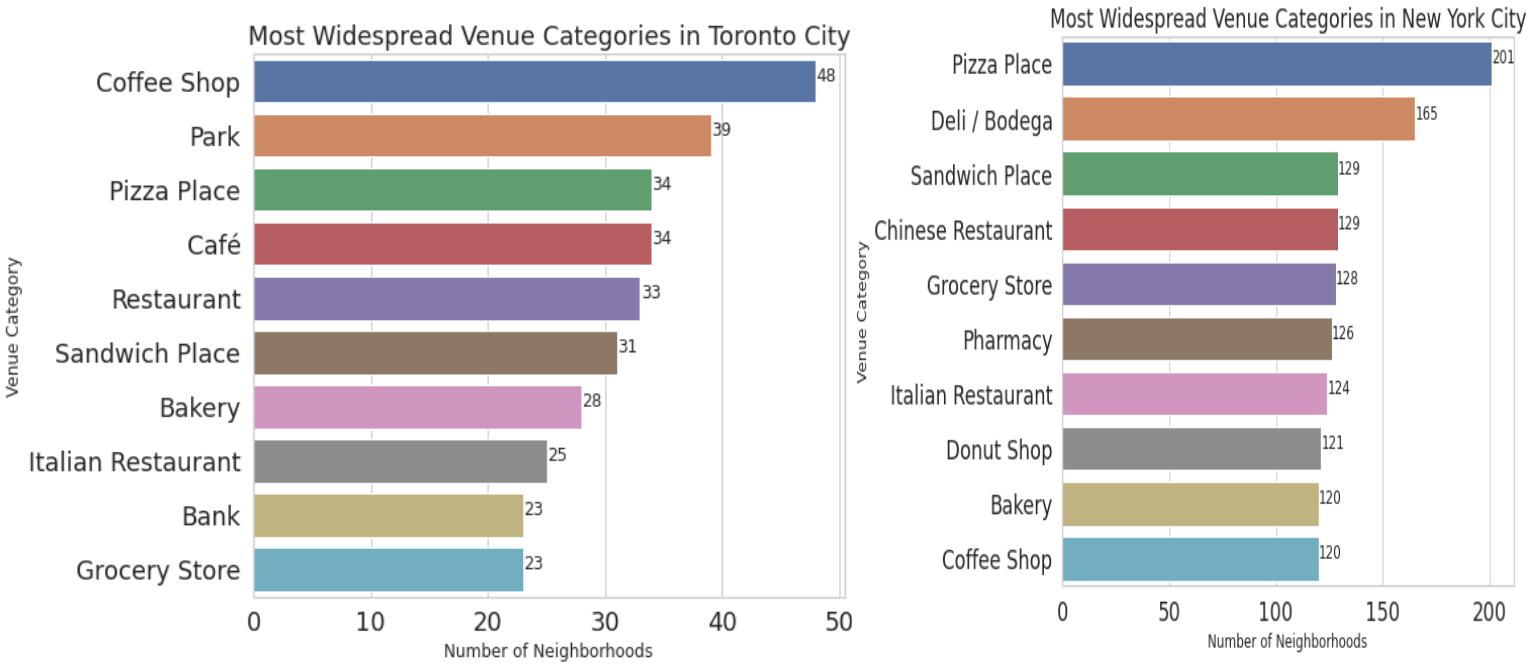


Fig 1.4 Most Widespread Venue Categories in Toronto and New York City

# 3. 3 Cluster Analysis

## 3.3.1 Clustering Neighborhoods

We use a machine learning algorithm called k-means clustering is used to cluster venue categories into similar groups. In this analysis, k-means is a popular algorithm for clustering neighborhoods. Before clustering, the categorical values were transformed to numerical values using the One-hot encoding method because we can't be able to analyze categorical variables. Then, the rows are grouped by neighborhood and by taking the mean of the frequency of occurrence of each category and returned the category of the most common and widespread venues and created the new dataframe. We can see the results in the above visualization.

To make the analysis more interesting, we wanted to cluster the neighborhoods based on the neighborhoods that had similar averages of venues in that Neighborhood. To do this we used **K-Means** clustering. Now, the neighborhoods are clustered into 5 similar groups, and combined the results of both cities shown below dataframe.

| Neighborhood | Cluster Labels | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Allerton_NYC** | 0 | Breakfast Spot | Supermarket | Martial Arts School | Check Cashing Service | Chinese Restaurant | Donut Shop | Gas Station | Pharmacy | Bus Station | Discount Store |
| **Annadale_NYC** | 2 | Food | Dance Studio | Pharmacy | Train Station | Diner | Pizza Place | Restaurant | American Restaurant | Park | Fast Food Restaurant |
| **Arden Heights_NYC** | 2 | Bus Stop | Pizza Place | Pharmacy | Deli / Bodega | Coffee Shop | Egyptian Restaurant | Electronics Store | Empanada Restaurant | English Restaurant | Entertainment Service |
| **Arlington_NYC** | 2 | Bus Stop | Arcade | Deli / Bodega | Scenic Lookout | American Restaurant | Event Service | Factory | Exhibit | Event Space | Yoga Studio |
| **Arrochar_NYC** | 0 | Food Truck | Mediterranean Restaurant | Middle Eastern Restaurant | Sandwich Place | Supermarket | Pizza Place | Bus Stop | Deli / Bodega | Pharmacy | Bagel Shop |

Table 1.5 Clustering the most common Venues of both Cities

## 3.2.2 Clustering Analysis

We have a total of 5 clusters and added columns that show the most common venue categories in each neighborhood with its cluster label(0, 1, 2, 3, 4).

| Cluster Labels | Number of Neighborhoods |
|---|---|
| 0 | 39 |
| 1 | 118 |
| 2 | 183 |
| 3 | 22 |
| 4 | 35 |

Table 1.6 Number of Neighborhoods in each cluster

The k-means clustering algorithm grouped neighborhoods of New York and Toronto in 5 clusters based on the similarity between their venues. Now, these clusters will be investigated to see the most common categories in each clusters. The following fig show the most common 7 venue categories in each cluster; for each common category, the percentage of venues of that category in the neighborhoods of the cluster is shown also.

Cluster 1

| | % of venues |
|---|---|
| Coffee Shop | 4.589995 |
| Italian Restaurant | 3.145952 |
| Bar | 2.965446 |
| Pizza Place | 2.810727 |
| Café | 2.733368 |
| Bakery | 2.320784 |
| Park | 1.985560 |

Cluster 2

| | % of venu |
|---|---|
| Pizza Place | 5.7600 |
| Pharmacy | 3.5122 |
| Sandwich Place | 3.2031 |
| Deli / Bodega | 3.0345 |
| Italian Restaurant | 2.9502 |
| Bank | 2.9502 |
| Donut Shop | 2.8659 |

Cluster 3

| | % of venues |
|---|---|
| Deli / Bodega | 4.759592 |
| Pizza Place | 4.711025 |
| Park | 3.545410 |
| Chinese Restaurant | 2.428363 |
| Grocery Store | 2.379796 |
| Coffee Shop | 2.282661 |
| Bus Stop | 2.136960 |

Cluster 4

| | % of venues |
|---|---|
| Coffee Shop | 5.457826 |
| Italian Restaurant | 3.608480 |
| Bakery | 2.796572 |
| Pizza Place | 2.480830 |
| Korean Restaurant | 2.255300 |
| American Restaurant | 2.210194 |
| Café | 2.210194 |

Cluster 5

| | % of venues |
|---|---|
| Coffee Shop | 10.865191 |
| Café | 5.633803 |
| Hotel | 5.030181 |
| Restaurant | 4.426559 |
| Japanese Restaurant | 2.414487 |
| Deli / Bodega | 2.414487 |
| American Restaurant | 2.414487 |

Table 1.7 Most common venue category in each cluster

As we can see the differences between the clusters in the above fig, each cluster has different distributions of common venue categories than the other clusters. Some of the observations are explained as follows:-

1. 1st cluster commonly consists of Coffee Shop and Italian Restaurant and these make about 9% of Venues
2. 2nd cluster consists of Pizza Palace and Pharmacy as most common categories
3. Deli/Bodega and Pizza Palace are appear as the most common categories of the 3rd cluster.
4. Coffee Shop is the most common in the 4th and 5th clusters.

Moreover, the following bar plot shows the number of similar neighborhoods in both cities.


Fig 1.6 Similar Neighborhoods in both cities

# 4. Conclusions

From the above analysis, I concluded that the areas of New York City and Toronto were clustered into various groups dependent on the classification of the venues in these areas. The outcomes of the

clusters are demonstrated that there are venue categories that are more common in certain groups than the others. Also, these top most common venue categories contrast from one cluster to the others. So using this information, we can make decisions and can be able to find similar neighborhoods in the City. It will also help business people to get answers to the questions like what type of businesses are more likely to thrive, what are the neighborhoods that are suitable for each type of business, and what types of businesses are not desirable in each city. This allows business people to take better and more effective decisions regarding where to open new businesses. In the future, if we investigate further analysis by taking more aspects into account, it may bring about discovering various styles in each cluster based on the most common venue categories.

# 5. References


1. https://ich.unesco.org/en/RL/ethiopian-epiphany-01491
2. https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
3. https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm
4. http://cocl.us/Geospatial_data
5. https://Foursquare.com
6. https://foursquare.com/developers/signup