**Finding the similar neighborhoods in New York and Toronto cities**

**Nikhil Thomas**

**Aug 6, 2021**

Table of Contents

[Introduction 1](#_Toc79421361)

[Data 1](#_Toc79421362)

[Methodology 2](#_Toc79421363)

[Results 3](#_Toc79421364)

[Discussion 16](#_Toc79421365)

[Conclusion 16](#_Toc79421366)

## Introduction

As people move from New York City to Toronto and vice versa, many are worried about the new changes they would bring in their lives. They are concerned about their neighborhoods and wish to move to a similar area they live in. It would be beneficial if they could know the neighborhood that matches their current one. This would ease their nervousness and help them properly focus on the actual neighborhood of their liking rather than regretting the choice after the move. Therefore, we would try to identify and compare each area in New York and Toronto and create a similar grouping to help people identify their similar neighborhoods in the other city.

We would plan to capture all the neighborhoods in New York and Toronto cities. Using the Foursquare integration, we could then find all types of places surrounding these neighborhoods. Break them into religious, restaurants, schools/colleges, parks, movie theaters, etc. Based on this information, we could then utilize K-means clustering to similar group neighborhoods together.

## Data

For this project we have used data from public sources and Foursquare data through its API interface.

**Public Data:**

The New York data about their neighborhoods would be downloaded from NYC Planning website. The data would list all the neighborhoods and the geographic data.

[Link](https://www1.nyc.gov/site/planning/data-maps/open-data.page) : <https://www1.nyc.gov/site/planning/data-maps/open-data.page>

The Toronto zip codes along with their neighborhood names can be downloaded from the wiki page.

[Link](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) : <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>

**Foursquare API:**

Foursquare provides a valuable and publicly accessible location information like the amenities in nearby locations. We use their developer tools to access the required information about the neighborhoods in a city. Using this accessed information, we then rank the neighborhoods based on the amenities they have. This service is free of charge.

## Methodology

**First Step** of the data analysis is to retrieve information about all the borough and its neighborhoods for both cities (New York and Toronto). This information is captured from the public data sources. Then we would pass the coordinates of neighborhoods in New York and Toronto to the Foursquare and retrieve the information all the venues in about the 500-mile radius. The venue information would include its name, location coordinates and the category of the venue.

**Once we have all the information,** we would be to merge the data into a single data frame. We still would retain all the information about the locations parent. We would be to identify all the top 10 venues in each neighborhood based on each venue’s frequency. i.e., by identifying the frequency of venue category in each location. This would help us to identify the common venues like parks, type of restaurants, types of business etc. This would help us to understand what the people in each location likes and provides a character to a location. So, if we see any locations having a lot of parks, we could assume that the people in this location like to go to parks and hence would tend to be more connected with nature.

Once we have identified the top venues in each location, We will use K-Means Clustering algorithm to group venues in a neighborhoods, then we can reduce the number of individual venue comparisons to be done against each neighborhood. We will do these comparisons against the types of venue, individually, collectively, or altogether. The reason why er choose the K-means clustering is because its easier to define and use.

We canexamine each cluster and determine the distinguishing characteristics on venue categories that distinguish each cluster. This will help us to provide an alternate if anyone wants to find the similar neighborhood in another city. This should also help if anyone is planning to move within the city.

## Results

###### Understanding the New York Layout

A picture containing text

Description automatically generated

A picture containing text

Description automatically generated

Map

Description automatically generated

###### Getting the venue locations from Foursquare for New York

**Table

Description automatically generated**

###### Understanding the Toronto Layout

**Table

Description automatically generated with medium confidence**

**Map

Description automatically generated**

###### Getting the venue locations from Foursquare for Toronto

Table

Description automatically generated

###### Merging the New York and Toronto Data

**Table

Description automatically generated**

###### Analyzing each neighborhood

**Table

Description automatically generated**

###### Identifying the top 10 venues in each neighborhood

**Table

Description automatically generated**

###### Clustering neighborhoods to find the similarities

**Graphical user interface, application, table

Description automatically generated**

###### Creating an aggregate view of cluster per city and borough

Table

Description automatically generated

###### Analyzing each cluster

A computer screen capture

Description automatically generated with low confidence

## Discussion

Based on our analysis, we found that New York has close to 10199 venues compared to Toronto, which had 2149 venues. Therefore, we can assume that New York is going to be bigger than Toronto. Consequently, we should find alternate neighborhoods that are very active and crowded; however, we might find it difficult to find remote and less populated spots.

New York has most of its neighborhoods spread around Cluster 0 and 4. Unfortunately, we could not define a similar location in cluster 1 for New York, and we did find that there are only two neighborhoods that qualified for Toronto.

Bronx and Brooklyn's neighborhoods dominate in Cluster 0, whereas Brooklyn and Manhattan neighborhoods dominate in Cluster 4. Most of the downtown area of Toronto belongs to Cluster 4. This illustrates that anyone moving from the downtown area in either place should find it easier and comfortable in Manhattan or vice-versa.

Cluster 2 represents the prime locations of restaurants, so it should be suitable for people who would want to be somewhere in the middle but not very close to the action of the downtown.

Cluster 0 represents neighborhoods that are a bit far away from the business neighborhoods and hence rely on public transport for commuting.

Cluster 3 represents people who love parks or like physical activities; this is perhaps meant for young families with kids. Therefore, they should find quite a selection in neighborhoods away from the busy and buzzing life of the city.

Cluster 4 represents restaurants and shopping neighborhoods. People who love to dine or shop would love these neighborhoods.

## Conclusion

Therefore, based on our analysis, we are hopeful that this algorithm model would benefit people who are moving between two cities of New York and Toronto. We are optimistic that this would comfort people who are new to the city and want to find a similar environment as they were in their previous city. This should also be beneficial to those looking to move within the neighborhoods in the city as the model can also list similar neighborhoods in the same city.

As we control the number of clusters, we did not increase the cluster count as it increases the risk of outliers for additional clusters, like what we see in Cluster 1. However, we could expand the number to fine-tune the similarities closer to its accuracy grouping.

We were limited with the venue information provided by Foursquare and adding additional information from other sources would help the K-means clustering model to be more accurate with new additional information. However, keeping in mind, the K-means is not great when dealing with the outliers as it would assign it to cluster based on the weakest similarities. Hence this model should be seen as helpful in shortlisting the neighborhoods. There should be additional research performed by the user to finalize and ensure the actual match.