The Battle of Neighborhoods

Sebastian Toledo

March 08, 2020

1. Introduction

1.1 Background

Today, technology represents an incredible and fundamental tool for society that goes beyond what was previously considered possible and one of the most exciting and requested areas is data science, since thanks to various techniques such as Data Analysis, Machine Learning and Deep Learning have allowed to solve innumerable problems in a very efficient and fast way. That is why my interest and motivation in contributing to society in a responsible way, starting with this first contribution that focuses on analyzing and comparing neighborhoods in two of the most exponential and important cities in the world, New York and Toronto.

1.2 Problem

Through Data Analysis and some Machine Learning algorithms we will look for patterns that help us understand the similarities and differences that the neighborhoods of both cities have. Information that is very valuable for tourists, business people or interested in living there.

2. Data

For this project we will use information provided by two sources that I will cite below.

Dataset containing relevant information from the city of New York: https://geo.nyu.edu/catalog/nyu_2451_34572

Website containing relevant information about the city of Toronto: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.

This information is essential to respond to our analysis as it contains longitude, latitude and corresponding neighborhood of each city, we also have the Foursquare API that provides us with additional information as more relevant places that make up each neighborhood and their respective information provided by visitors.

To respond to our problem, we will use the information provided and we will categorize and segment it to later graph it and be able to explain through a statistical summary the main factors that occur in each neighborhood and how they are related, we can also give a clear perspective of the similarities and differences between New York and Toronto.

3. Methodology

In this section we will explore and analyze both cities. First, we will individually analyze each city with the help of Foursquare, we will look for the most visited places by neighborhood, together with important statistics that will help us understand its main characteristics, second we will compare both cities in search of similarities and differences, we will use the k-means clustering algorithm To segment the neighborhoods and discover which neighborhoods in both cities share fundamental characteristics in relation to the most emblematic places of each cluster, we will also use bar graphs that will help us visualize each cluster.

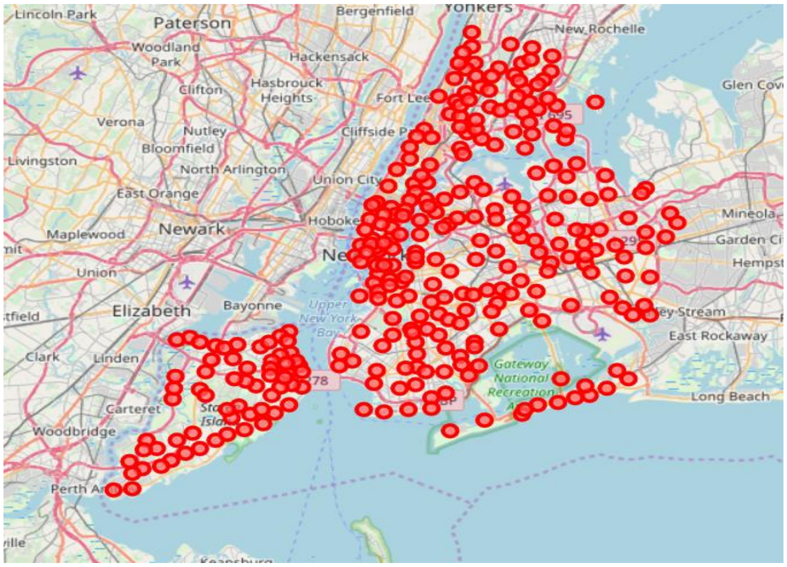


Figure 1. Map of New York with neighborhoods superimposed on top.

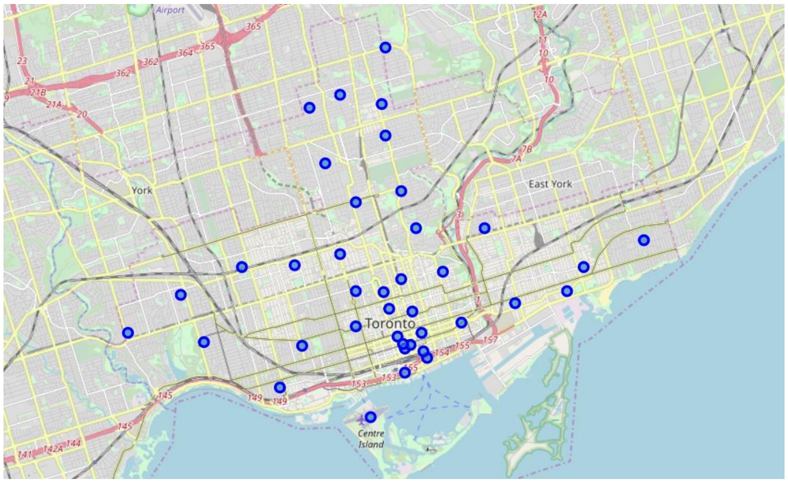


Figure 2. Map of Toronto with neighborhoods superimposed on top.

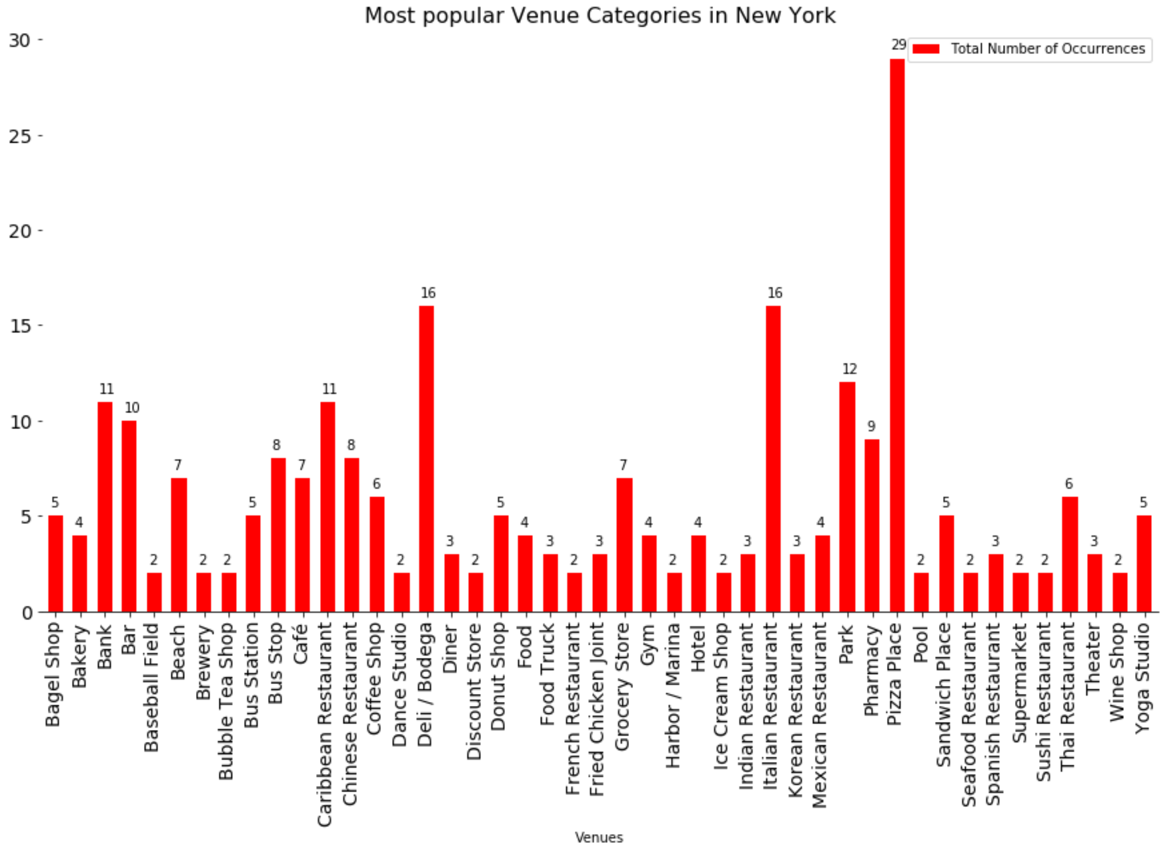


Figure 3. Most popular Venue categories in New York.

As we can see I have compiled the most influential categories related to the most qualified and visited places in New York City, interestingly the most popular category is Pizza Place followed by Italian Restaurant.

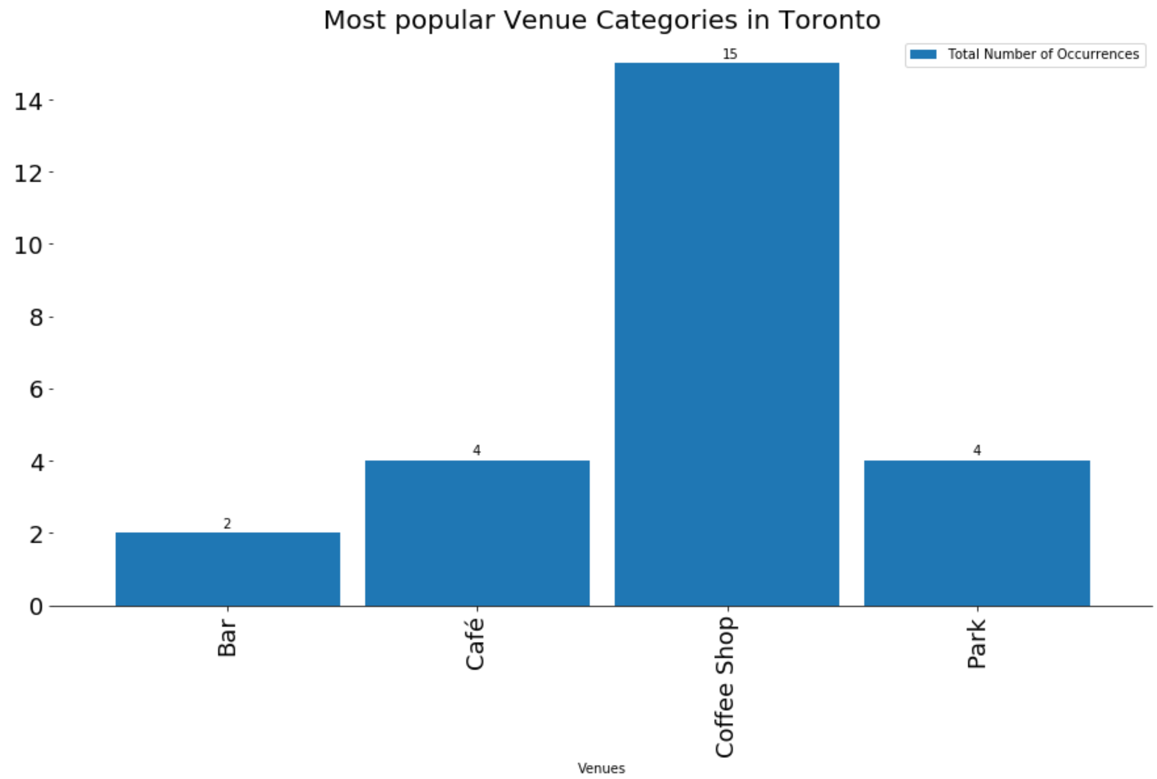


Figure 3. Most popular Venue categories in Toronto.

Otherwise, in Toronto the most popular category is Coffee Shop and Park.

3.1 Clustering Model

As a predictive model we will use a Machine Learning algorithm belonging to the unsupervised learning called K-Means, which mainly consists of the following steps.

1) Randomly placing *k* centroids, one for each cluster.
2) Calculate the distance of each point from each centroid.
3) Assign each data point (object) to its closest centroid, creating a cluster.
4) Recalculate the position of the *k* centroids.
5) Repeat the steps 2-4, until the centroids no longer move.

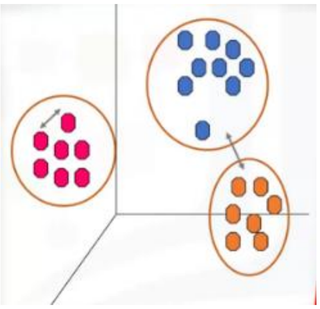


Figure 4. K-Means algorithm.

To determine its accuracy, we use the internal approach called "elbow method", which consists of average the distance between data points within a cluster or the Mean Square Error (MSE).

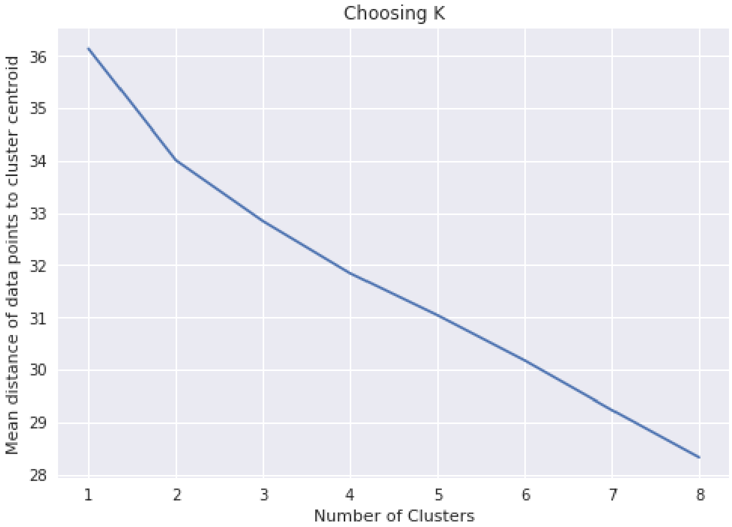


Figure 5. K-Means Accuracy.

Based on the results obtained by the accuracy analysis, we decided to test the model with k = 8, that is, eight clusters or centroids, and we obtained the following as a result.

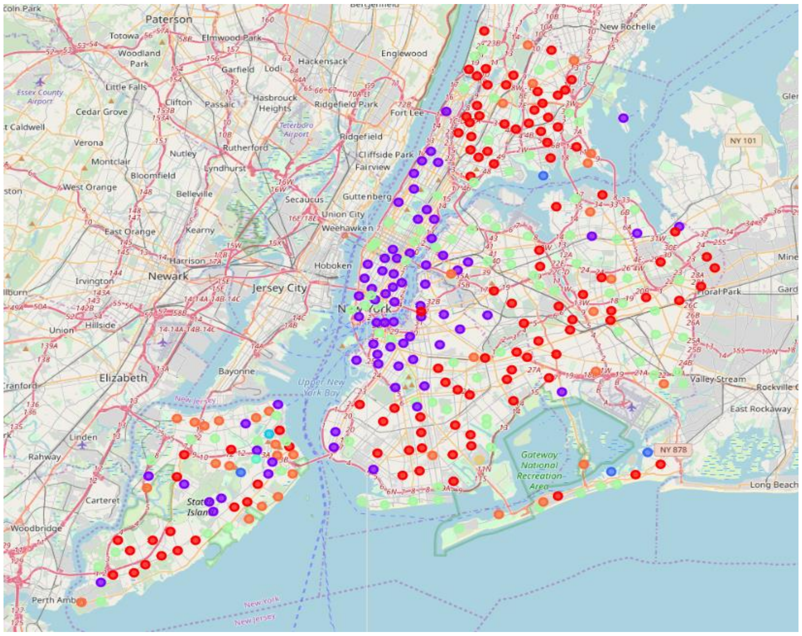


Figure 6. New York clustering results.

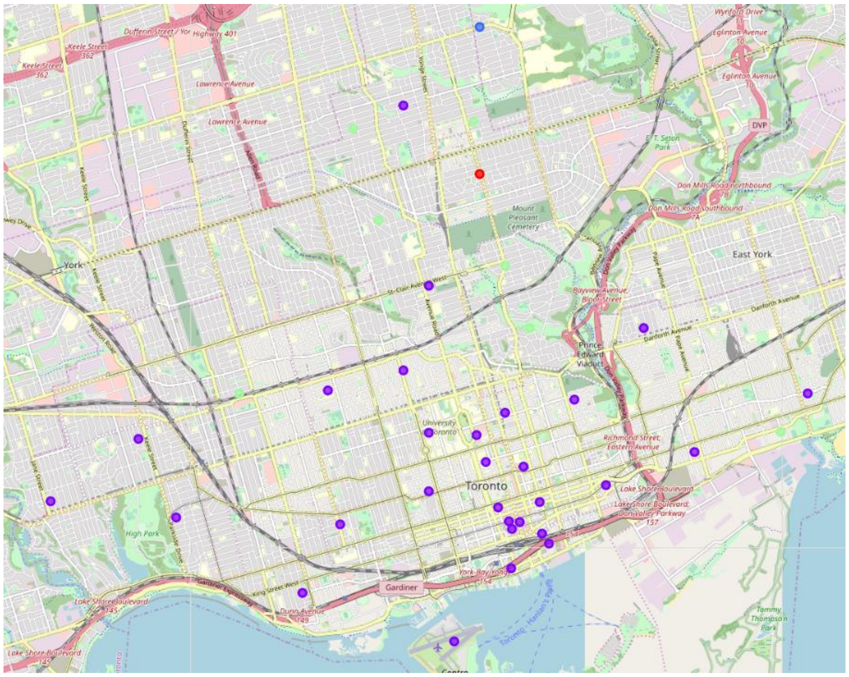


Figure 6. Toronto clustering results.

Next, the results obtained represented in bar graphs corresponding to the most common Venues of each cluster will be shown.

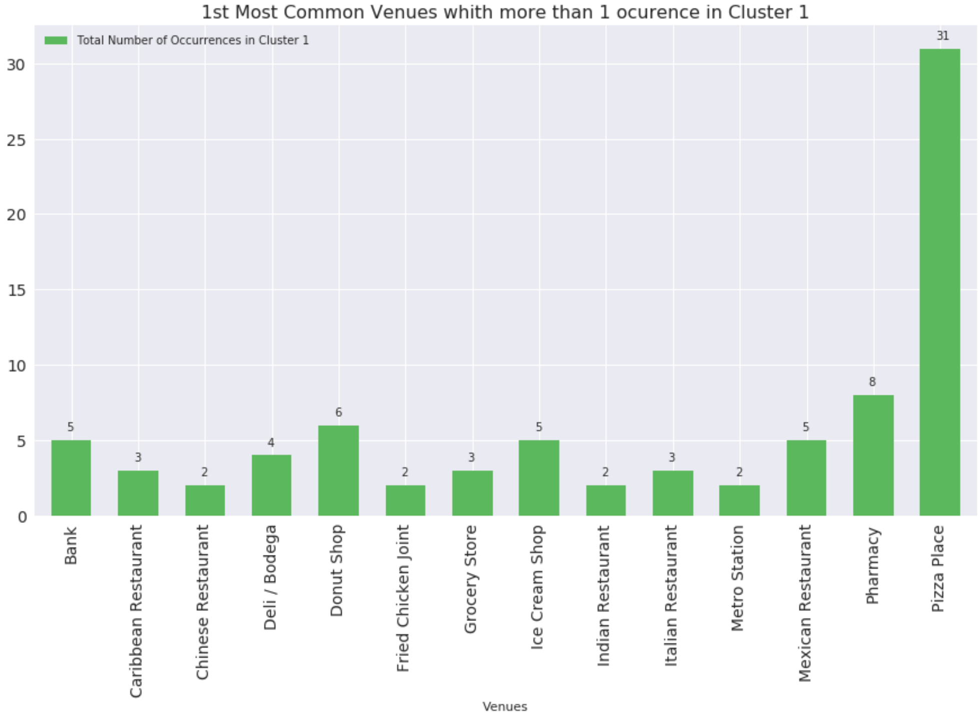


Figure 7. Cluster No 1.

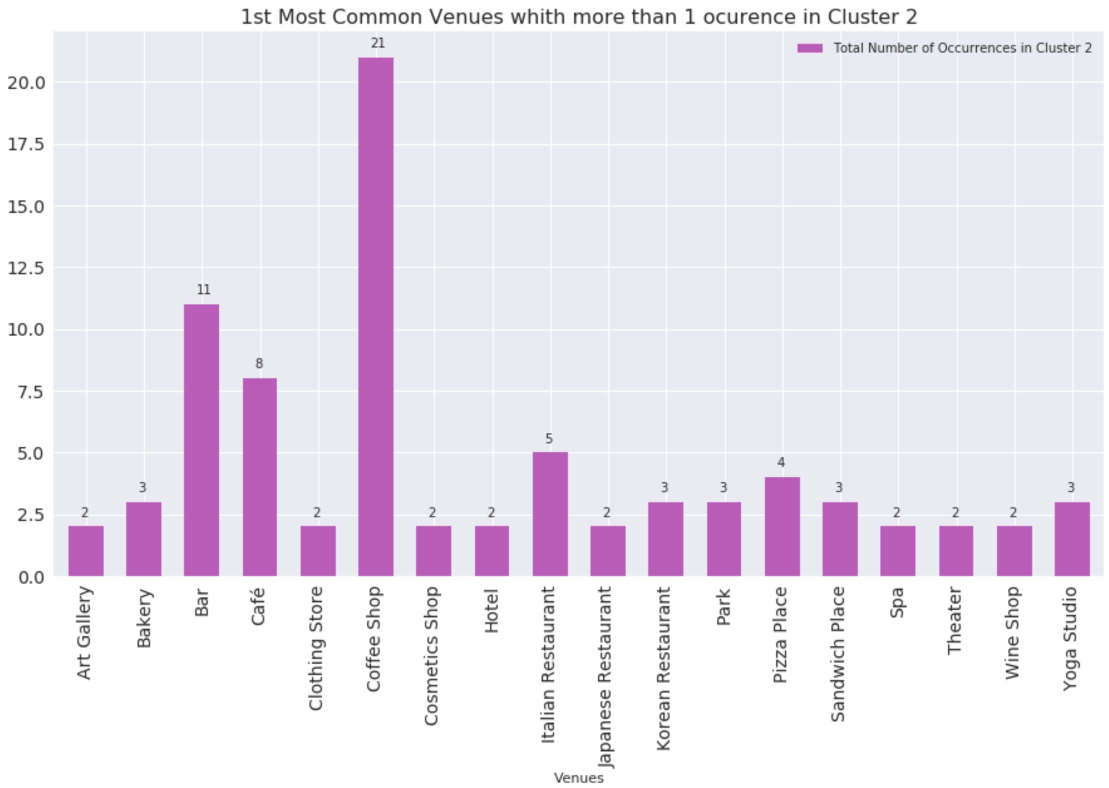
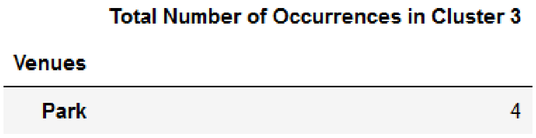
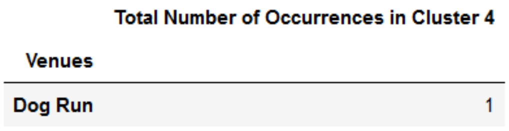


Figure 8. Cluster No 2.



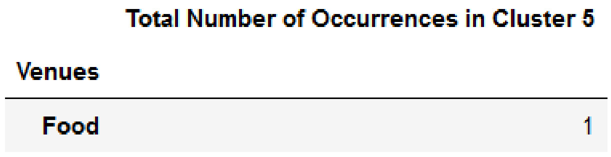
Venues	Total Number of Occurrences in Cluster 3
Park	4

Figure 8. Cluster No 3.



Venues	Total Number of Occurrences in Cluster 4
Dog Run	1

Figure 9. Cluster No 4.



Venues	Total Number of Occurrences in Cluster 5
Food	1

Figure 10. Cluster No 5.

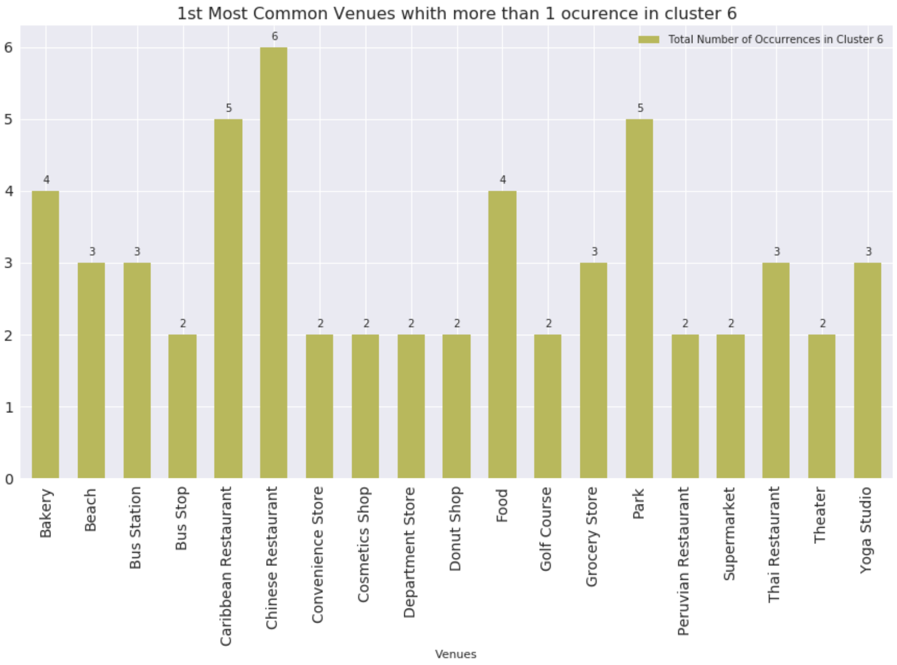
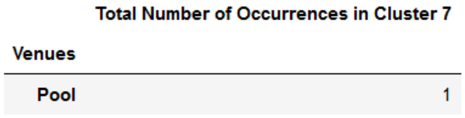


Figure 11. Cluster No 6.



	Total Number of Occurrences in Cluster 7
Venues	
Pool	1

Figure 12. Cluster No 7.

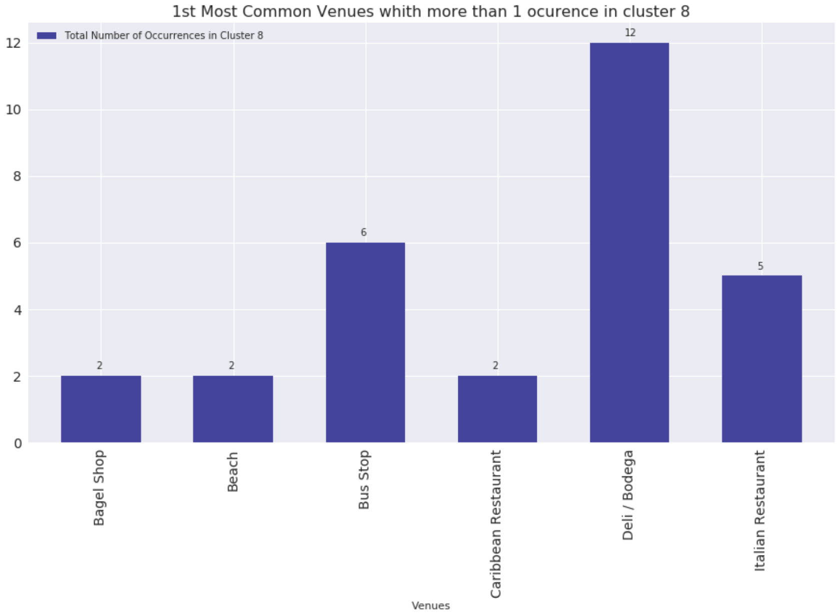


Figure 13. Cluster No 8.

Results and Discussion

Our analysis shows a large number of similarities in both cities and without a doubt Foursquare has provided us with relevant information and notorious similarity between neighborhoods that were segmented using unsupervised learning with the Kmeans algorithm in which we divided the information into eight clusters related to the categories of most important and visited places. On the one hand, New York with a population of around eight million, has five boroughs and three hundred six neighborhoods. As we could see in the bar chart the three most important categories within the most visited and qualified places in the city are:

1. Pizza Place
2. Italian Restaurant
3. Deli / Winery

On the other hand, Toronto has a population of around three million, four boroughs and thirties and nine neighborhoods. If we look at the information provided by Foursquare and later in the bar chart we see that the three most important categories within the most visited and qualified places in the city are:

1. Cofee Shop
2. Cafe
3. Park

However, at the moment that we join both datasets and through the Kmeans algorithm we form eight clusters which indicate important characteristics that we can appreciate in similarities of both cities, without a doubt this information is clearly explained with the bar graphs, which I invite to you to observe.

Conclusion

Purpose of this project was to identify the similarities and differences between the city of New York and the city of Toronto, a task that was not simple due to the large number of factors that can influence the outcome in addition to the vision and purpose have when analyzing the data. As a result of this project we can conclude that despite the geographic differences between the two cities together with the population difference, we were able to successfully segment both cities into eight clusters which show us remarkable characteristics that the neighborhoods between both cities share, mainly for the most common places that thanks at foursquare we have been able to recover.

I hope this notebook is useful to someone who aspires to visit or live in either of the two New York or Toronto cities, or even to someone who has a business interest.