Comparative analytics project between New York and Toronto – What makes a city unique

Forrest Yao

April 11th, 2020

1. Introduction

1.1 Background

I am a big fan of travel. I have lived in Toronto for 5 years and have mixed feellings about it. I also travelled to New York once and fell in love with it within 5 days. I am personally really interested into urbanology, therefore, I would like to utilize what I have learned from IBM Data Science course and conduct a comparative analytics project between these two cities, the ultimate goal of this project is to find out what factors could make a city unique by comparing their neighborhoods, their communities, venues, and infrastructures. For example, New York may have more universities near its financial area in order to take advantage of the Wall street; however, universities in Toronto may be uniformly distributed given that there is large population based on North. Another aspect this project aims to explore is to provide more information for entrepreneurs who want to start a business in these cities by using some data science techniques such as clustering.

1.2 Interest

This project is for all travel-lovers, urbanologists, entrepreneurs and anyone who lives in or is interested in these two cities since it could provide a comprehensive viewpoint about multiple aspects of these two cities. Hopefully I will find something interesting.

2. Data acquisition and cleaning

2.1 Data sources

In this project, I will mainly use Toronto geographical data and New York geographical data which are actually given by this course. As all data are available throughout this project, it would save a lot of time searching for data and import them. I will also integrate Foursquare location data into this project and I will specify all necessary data sources if I use them later in this project.

3. Methodology

3.1 Clustering

First of all, I am going to briefly talk about what is a cluster. A cluster is a group of data points or objects in a dataset that are similar to other objects in the group, and dissimilar to datapoints in other clusters. The difference between clustering and classification is that: classification algorithms predict categorical classed labels. This means assigning instances to predefined classes such as defaulted or not defaulted. For example, if an analyst wants to analyze customer data in order to know which customers might default on their payments, she uses a labeled dataset as training data and uses classification approaches such as a decision tree, Support Vector Machines or SVM, or logistic regression, to predict the default value for a new or unknown customer. Generally speaking, classification is a supervised learning where each training data instance belongs to a particular class.
In clustering however, the data is unlabeled and the process is unsupervised. For example, we can use a clustering algorithm such as k-means to group similar customers as mentioned, and assign them to a cluster, based on whether they share similar attributes, such as; age, education, and so on.

3.2 K means method

K-means method aims to form To form clusters in such a way that similar samples go into a cluster, and dissimilar samples fall into different clusters. The principle it follows is that it should minimize the intra cluster distances and maximize the inter cluster distance and it should divide the data into non overlapping clusters without any cluster internal structure
Now, lets compare hierarchical clustering with K-means. K-means is more efficient for large data sets. In contrast to K-means, hierarchical clustering does not require the number of cluster to be specified. Hierarchical clustering gives more than one partitioning depending on the resolution or as K-means gives only one partitioning of the data. Hierarchical clustering always generates the same clusters, in contrast with K-means, that returns different clusters each time it is run, due to random initialization of centroids.

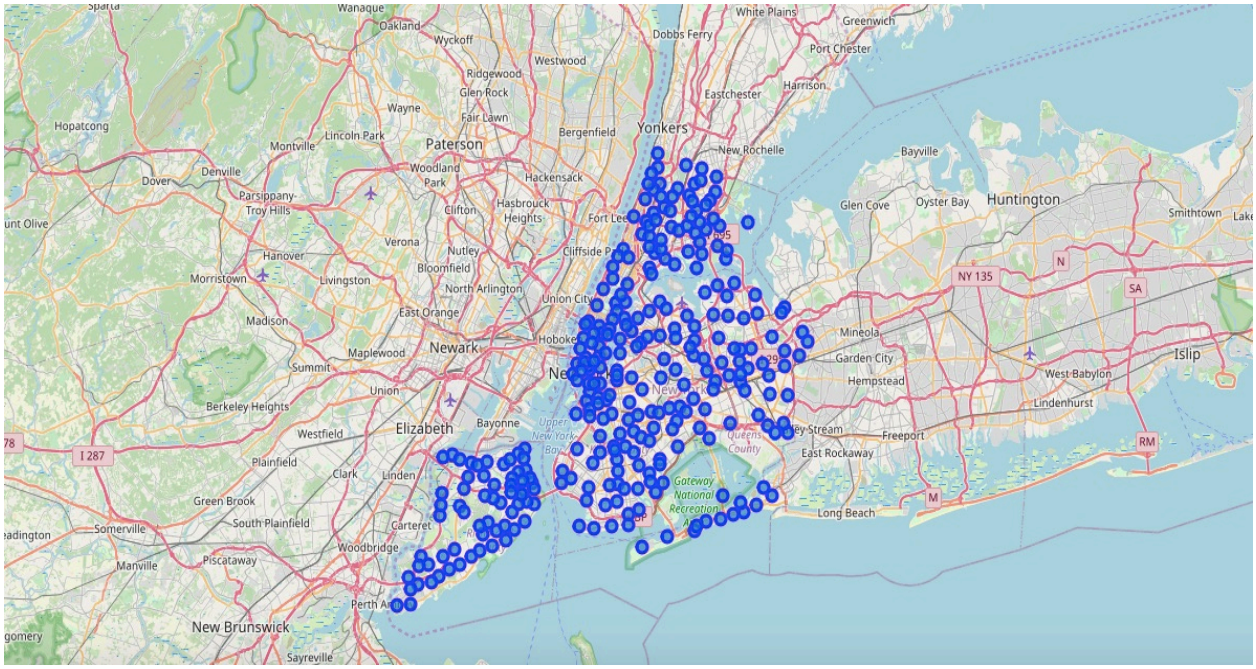
4. Exploratory Data Analysis

This part is the major part of the project which divided into two sub-sections. The first section tries explore the layout of two most important city infrastructures: university and

hospital. By comparing the location and density of university and hospital within these two cities, Audience could have a better understanding about which neighborhood could have education and medical potentials so that they could decide where to live in these two cities. The second sub-section of this part is to utilize clustering methodology I learned from IBM data science certification, and my target areas include Manhattan and Downtown Toronto as they are the most famous and populous neighborhoods within these two cities. I will use k-mean-Clustering technique introduced by last section to determine which neighborhoods could be grouped together based on certain characteristics within these two districts.

Let's first explore these two cities. Figure 1 lists all neighborhoods recorded by New York government, and Figure 2 lists all neighborhoods recorded by Toronto government. New York has a population of 8398748 distributed over about 302 square miles, and it has 250 to 300 neighborhoods grouped by 5 boroughs. Toronto, on the other hand, has a population of 5928040 distributed over about 243 square miles, and it has 140 neighborhoods. As we can see, the density of neighborhoods and population in New York is relatively high than it in Toronto, we can visually find more information from Figure 1 and 2.

Figure1 Neighborhoods distribution in New York City



Before we took a closer look at two major boroughs within each city, let's focus on the distribution of universities in these two cities. Although New York is a relatively crowded city, it has many great universities almost uniformly distributed in New York city.

Figure2 Neighborhoods distribution in Toronto

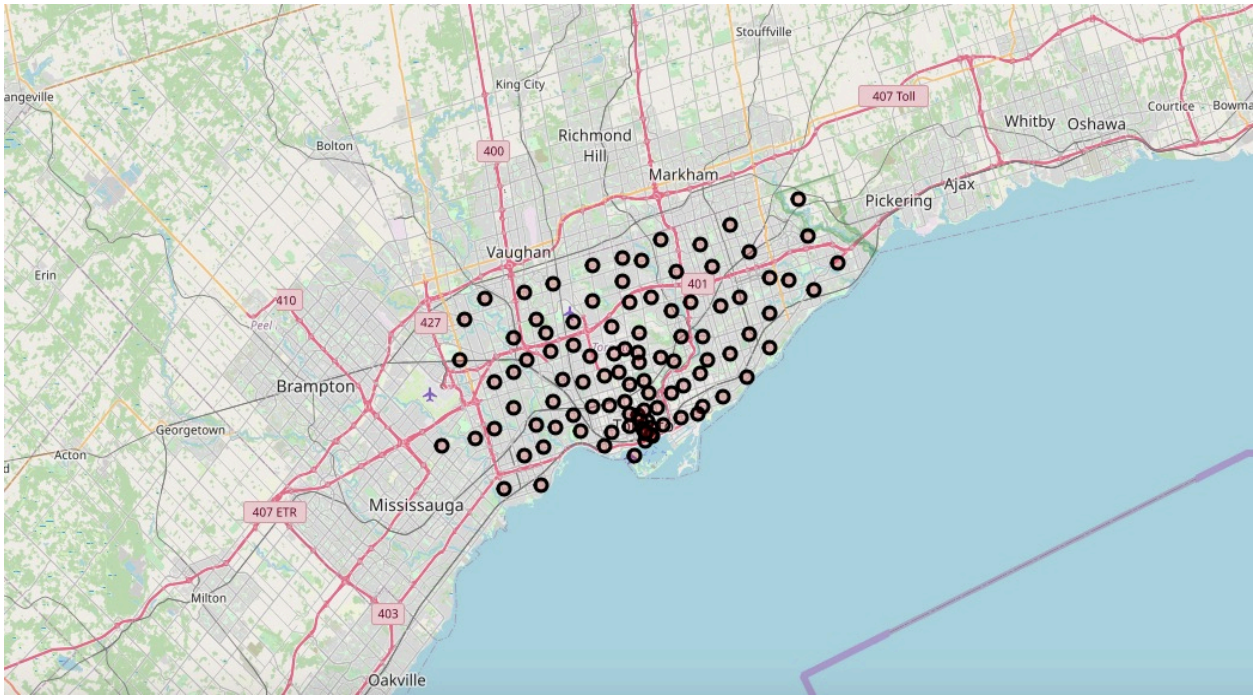
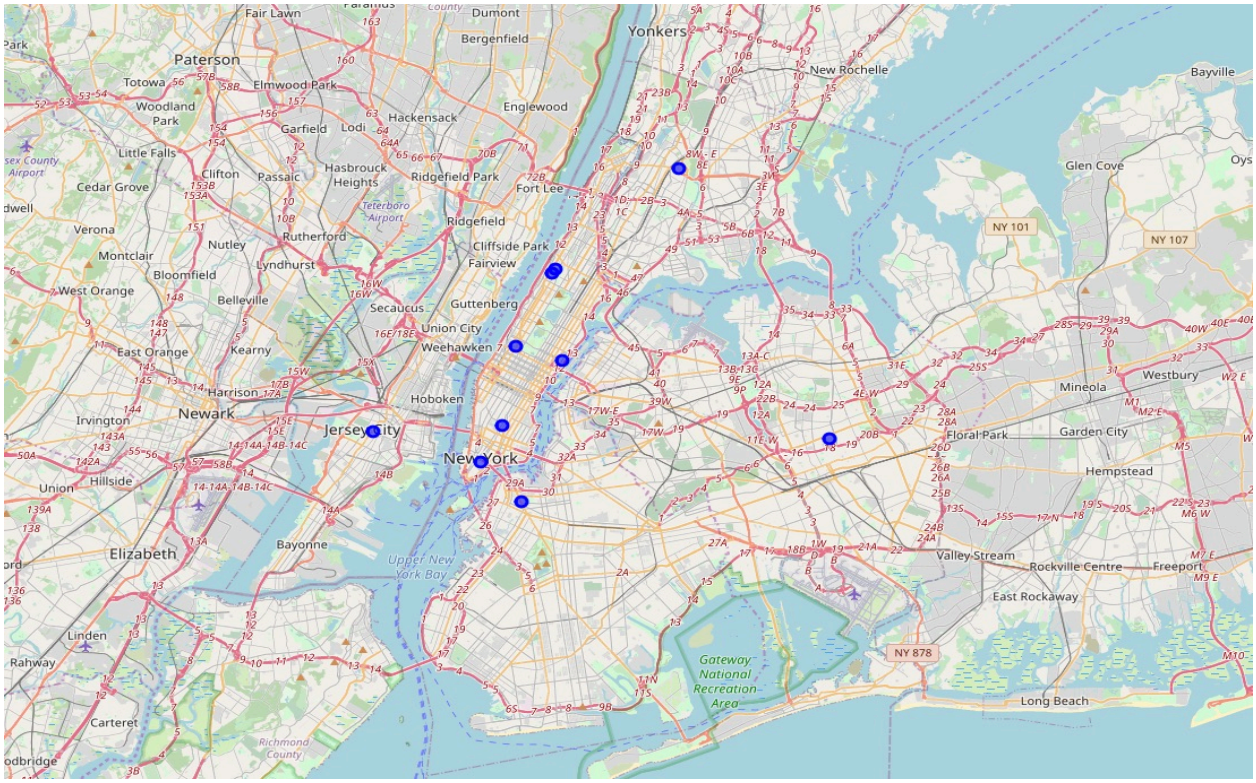
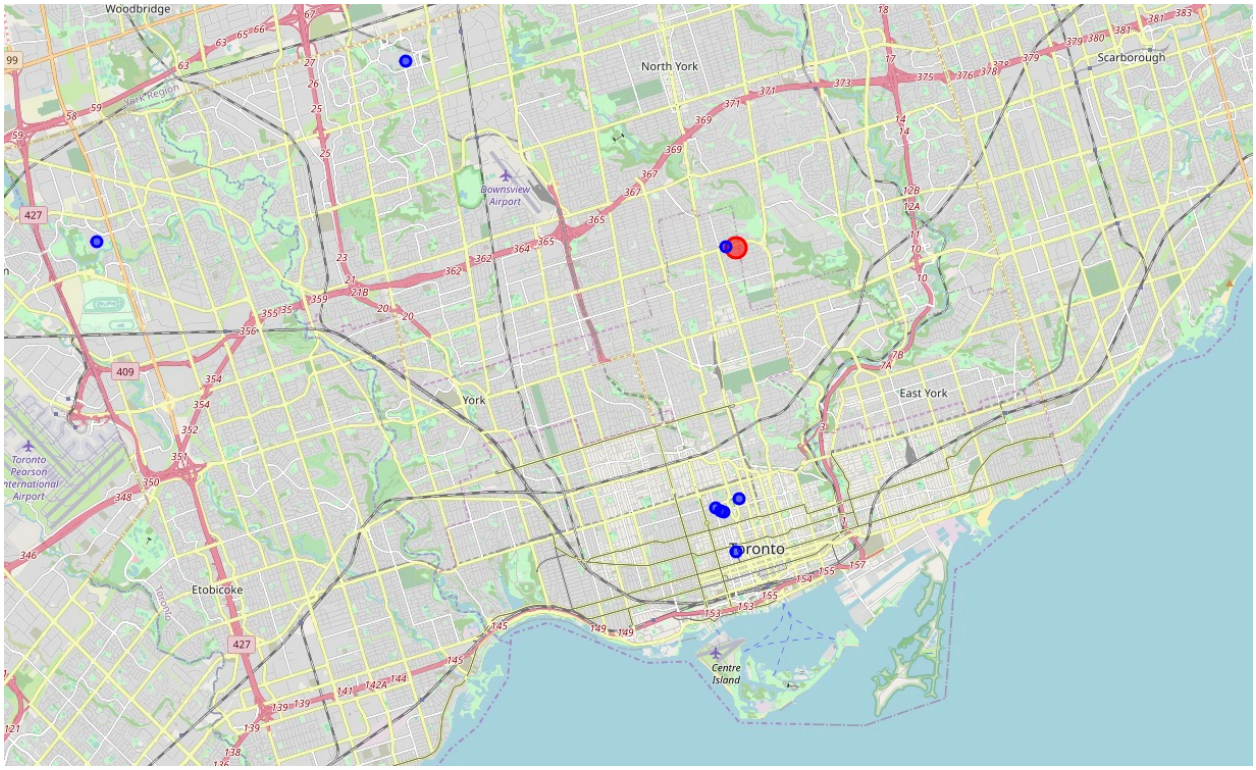


Figure3 Famous Universities in New York City



For example, Cornell University, Columbia university, New York University and so on. New York city boasts a world-class selection of universities, and eight of them are featured within the QS World University Rankings 2020. From Figure3, it is desired that all these universities are assessible and distributed perfectly so that if you live in New York City and want to walk around the campus, there is always one beautiful campus around you. On the other hand, Figure 4 represents universities in Toronto.

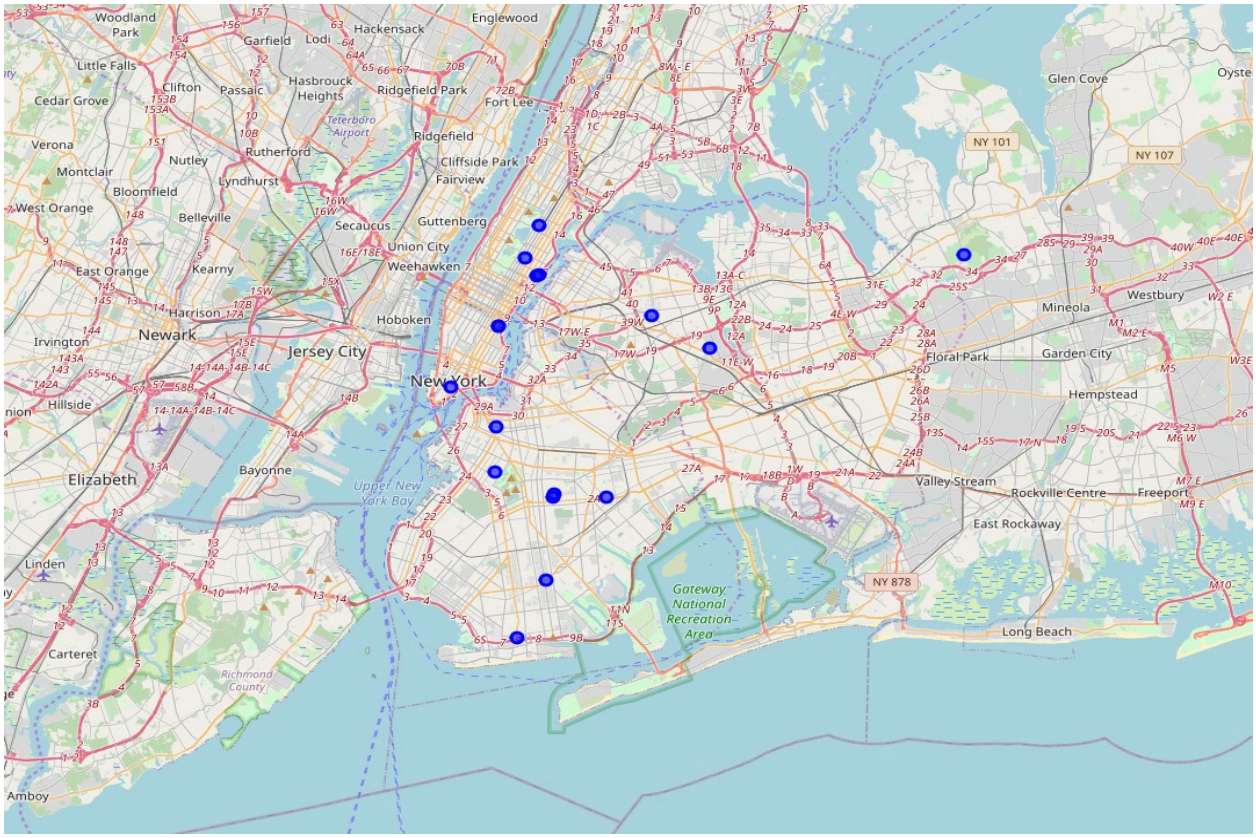
Figure4 Famous Universities in Toronto



From the map we can see, there are less than 5 universities distributed un-uniformly in the Big Toronto Area. Specifically, University of Toronto, OCAD University and Ryerson University are very close to each other and they are all based on Downtown Toronto, which could arise some potential problems. First of all, too many students live in Downtown which could significantly lead to a high rental expense. Moreover, every year, when students are graduating from their universities, thousands of students are looking for jobs in downtown area which enhance the intense of employment.

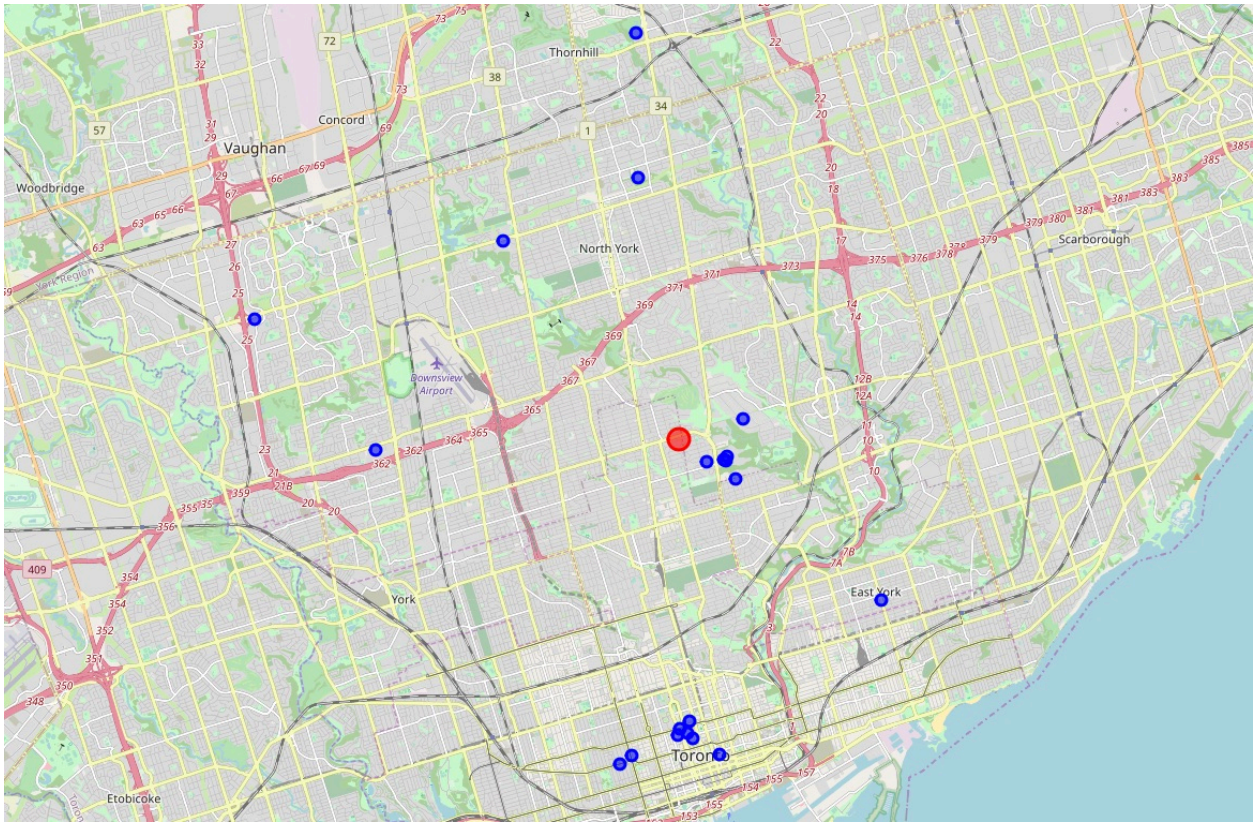
Besides for education level in these two cities, medical level is also another important factor when considering moving to a new city. Therefore, I also generated distribution graph for hospitals in these two cities. Figure 5 is distribution map for large hospital in New York City.

Figure5 Large hospitals in New York City



As we can see, similar to universities, the distribution of hospital in New York City is satisfying as well. However, in Jersey City, there is no large hospital shown in graph, so this is an area that NYC could improve in the future. Moving to Toronto, this time, the situation is much better, we can see that there are enough number of large hospitals based on Downtown Toronto and North York, and these two areas are the most specious and populous areas in GTA.

Figure6 Large hospitals in Toronto



Now it's time to conduct the second part analysis of this project which is cluster analysis about a major borough of each city. The detailed process will be discussed in Jupyter notebook associated with this report. In summary, we divide all neighborhoods in Manhattan and Toronto into five clusters. As shown in Figure 7, we could see the first cluster is neighborhood Queen's Park, Ontario Provincial Government, which is marked in red color, the most three appeared venue in this are coffee shop, diner and sandwich place. The second cluster also contain only one neighborhood which is Christie which marked in purple. The most appeared venues for neighborhood are grocery store, care and park. The third cluster only contains one neighborhood which is Rosedale marked in blue. The most appeared venues for neighborhood are park, playground and trail, so this area must be a great place for tourists to explore. The fourth cluster contains the largest number of neighborhoods in Downtown Toronto. The most appeared venues for these cluster are highly likely coffee shop, café, bar and restaurants, so this is definitely a place for food lovers. The last cluster also only contain one neighborhood which could be referred to as "financial district". It is marked by yellow color.

Now let's take a look at New York City. The first cluster which marked by red color, contains the following five neighborhoods: Central Harlem, East Village, Manhattan Valley, Morningside Heights and Gramercy. They are recognized by having bars, all kinds of restaurants and coffee shops. The second cluster is marked by purple color contains more than 20 neighborhoods and they are mainly based on south of Manhattan. The third cluster only contains one neighborhood and it is also the most north neighborhood. There is one special cluster which is near the sea, it is Stuyvesant Town, and the 1st most common venue in this neighborhood is Boat or Ferry, the 2nd most common venue is park, so this place could be a place where tourists could enjoy. And the final cluster contains also 5 neighborhood and it extends from north to south. It has a lot of different café, restaurants and park for people to relax.

5 Discussion and Conclusion

From what I have conducted by utilizing clustering techniques and comparative analytics methodology, we could obtain many insights between New York City and Toronto. To be short, first of all, the universities are almost uniformly distributed throughout the whole New York city whereas universities in Toronto are mainly based on Downtown area which could arise several social problems such as high rental expense and intense job searching pressure. In the second place, both cities have well-developed medical system where there are many large and assessible hospitals distributed in these two cities. Moreover, both cities have specific neighborhoods for specific functions, for example, by conducting cluster analytics, both cities have neighborhoods which have a large number of restaurants and bars. More importantly, both cities are near the sea; therefore, there are two special neighborhood built for Ferry or Airport for these two cities, and my cluster model correctly recognize these two neighborhoods. Furthermore, there is one thing I do want to point out is that deeper insights could be figured out by further analysis based on my project, I welcome everyone who would like to use my analysis as a base to obtain more interesting findings and insights.

Figure7 Cluster Analytics in Manhattan

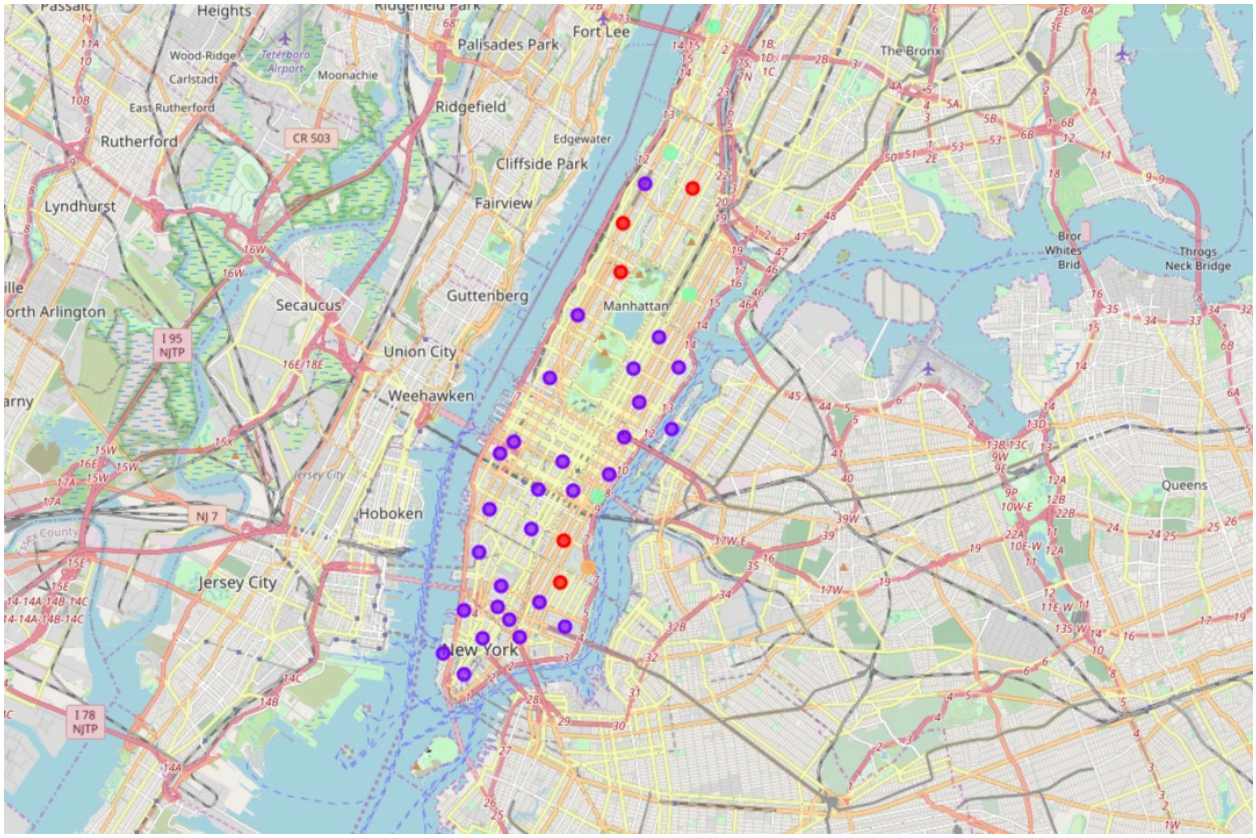


Figure8 Cluster Analytics in Toronto