## Neighborhood Segementation and Clustering

## 1. Introduction

### 1.1 Background

New York and Toronto are very diverse and financial capitals of their respective countries. Lying in the continent of North America, both cities have a complex urban structure made up of various neighborhoods that consists of all sorts of public utilities and tourist attractions. With restaurants, theatres, cinemas, office towers, parks, and all other places for work, food, and entertainment, these components of the neighborhoods play a vital role in New York and Toronto residents, retaining a direct relationship with the quality of their lives. For example, for a neighborhood with plenty of well-developed cinemas but no karaoke, it is more likely that residents in that neighborhood choose watching movies over singing for leisure. Having established that people’s lives are strongly related to their neighborhoods, it may be worth exploring the neighborhoods in New York and Toronto, and drawing some conclusions about how similar or dissimilar they are.

### 1.2 Problem

This project aims to find out the similarities and differences between the neighborhoods in New York and those in Toronto. To be more precise, we will first analyze the basic demographics of the two cities, including population, land area, city planning, GDP, etc. The above information will present us with a good preliminary understanding of the two cities. Then, we carry onto locating and plotting specific neighborhoods in the two cities to visualize how these neighborhoods, along with the popular venues in the cities, are composed. Finally, we will perform neighborhood segmentation and clustering on each of the city’s neighborhoods, and draw both qualitative and quantitative conclusions on the characteristics of each cluster of neighborhoods.

### 1.3 Interest

This project will definitely be of interest to all New York and Toronto citizens who wish to better understand the city and neighborhood they reside in. In addition, the clustering results will help citizens who would like to change a neighborhood within the city to choose the one that is most suitable for them. Besides, for all potential visitors to these two cities, it will give them sufficient information and advices of which neighborhood to stay in and which venues to visit, as they are able to see the geographical relationship between all these places clearly in the project. We also hope that our analysis of New York and Toronto will help the global society to eliminate stereotypes about the two cities, as any label that seems suitable for describing one neighborhood is likely not to be matching to another.

## 2. Data acquisition and cleaning

### 2.1 Data sources

The basic demographics data of New York and Toronto, including population, land area, GDP, are obtained directly from Wikipedia pages of the two cities, respectively. We understand that information such as population and GDP changes from time to time, so we try to compare figures that are from the same period of time while ensuring that they are both up to date. For the list of boroughs and neighborhoods of New York and Toronto, we fetch data from the NYU Spatial Data Repository (https://geo.nyu.edu), the Google Maps Geocoding API, and the

List of Postal Codes of Canada: M (Toronto Postal Code) Wikipedia page (https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M). All the above data sets are publicly available and thus can be accessed directly. Finally, we use Foursquare API to get the information of the venues within each neighborhood. Due to our restriction of a free account, we are entitled to only very limited number of premium calls of Foursquare API which are necessary for detailed information of the venues, such as posts, menus, etc. Hence, our analysis will stay at a broader level which primarily includes the location and the type of the venue, all of which can be accessed through regular calls.

### 2.2 Data cleaning

All basic demographics data that involve only one or several numbers will be used directly for bar charts and pie charts. There is more subtlety involved when dealing with data frames. First, we need to extract the data frames from the source website with beautiful soup, while cleaning out all other unnecessary information on the webpage. Afterwards, the New York and Toronto borough and neighborhood data is combined with their geospatial data where the names, postal codes, longitude and latitude of a neighborhood are included in one single data frame. Specifically, there are postal codes without any borough assigned, which we ignore for our analysis. There are also boroughs without any named neighborhoods inside it, in which case we name the neighborhood that is inside that borough the same name as the borough. We also merge all neighborhoods with the same postal code into one instance for the purpose of a more direct and insightful visualization. Finally, a sanity check is performed that the type of each variable of interest is correct, for example, the longitude and latitude values are floats that fall into a specific range, and there is no absent values (NA, NaN, null) inside the data frame. Besides, data processing is needed before passing data into k-means clustering, which we will discuss in more detail in the next section.

## 3. Methodology

In this section, I will discuss the methods that are involved in visualizing the demographics data as well as those in breaking the neighborhoods into clusters. Specifically, I will analyze why a bar chart is used for presenting the demographics data instead of a pie chart, and why I choose k-means clustering over hierarchical clustering when categorizing the neighborhoods of the two cities.

### 3.1 Demographics data: bar charts over pie charts

For demographics data such as population and GDP, we usually have a single number for each of the two cities that we want to compare. Our purpose is to show these numbers in a chart so that readers are able to easily view them and sense the difference between them. Bar charts and pie charts have thus risen two possible options for us to visualize these data. However, for pie charts, it is better to use them when “there is a specific and clear point related to the share of the total that we are trying to get across and the individual values of each slice is not important”. In our case, the total population of New York and Toronto does not seem to have a significant meaning when viewed as a whole. Moreover, as it is very difficult for the human eye to estimate the magnitude of angles, visualization of pie charts not very clear, which is one of the main criticisms for pie charts. For these reasons, bar charts are chosen as the way of visualizing demographics data in this project.

### 3.2 Clustering: k-means clustering over hierarchical clustering

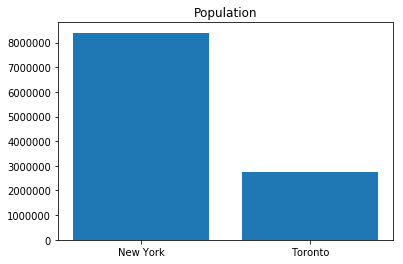
The main part of this project is to break the neighborhoods of New York and Toronto into clusters, so choosing a wise algorithm for clustering is essential. The two commonly used clustering algorithms are k-means clustering and hierarchical clustering. Since we have a desired number of clusters for the area to be broken into, k-means clustering seems to be a more suitable method than hierarchical clustering. Specifically, dendrograms produced by hierarchical needs to be further processed in order to retrieve the number of clusters, the process of which involves difficult decisions for how to wisely choose the points to break. Another compelling reason to choose k-means clustering over hierarchical clustering is that the computational time of k-means clustering is much faster than hierarchical clustering over the same set of data. Henceforth, we will perform k-means clustering over hierarchical clustering for neighborhood segmentation of New York and Toronto.

## 4. Results

In this section we present a summarization of the demographics data and the clusters of neighborhoods in New York and Toronto.

### 4.1 Demographics data

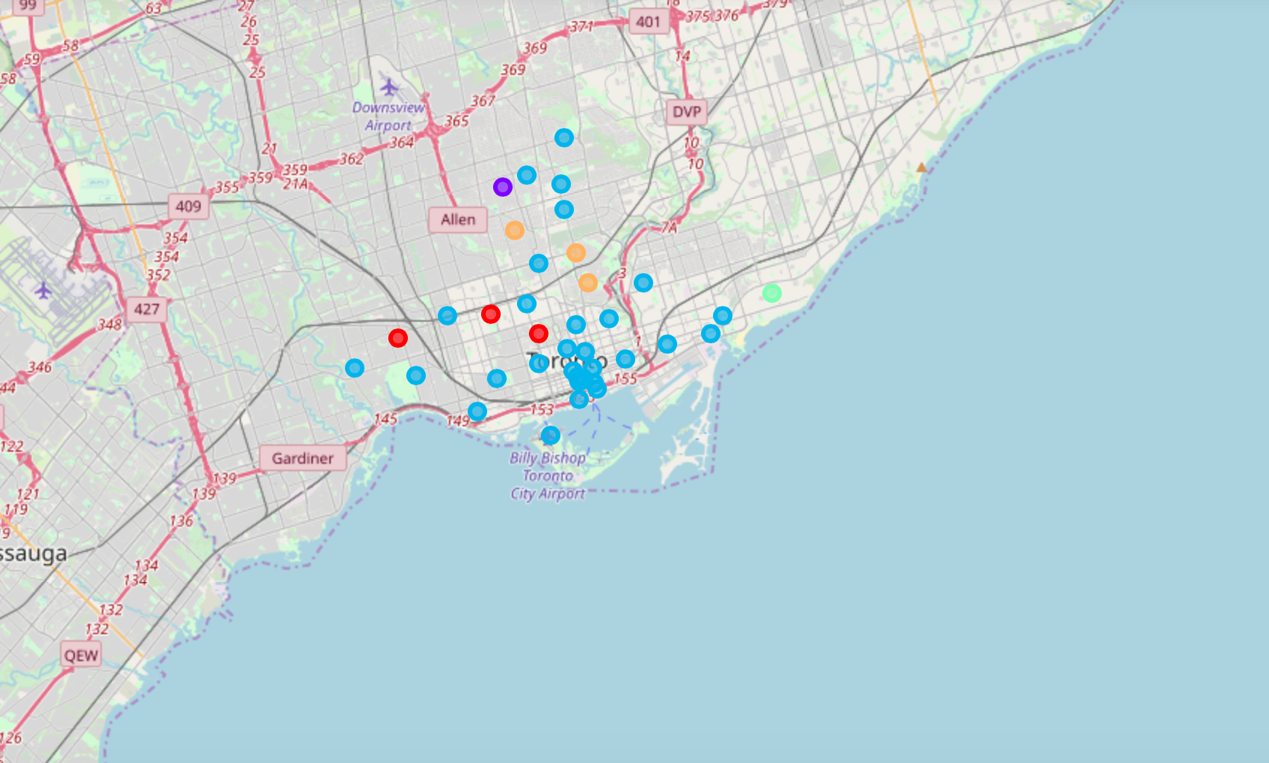
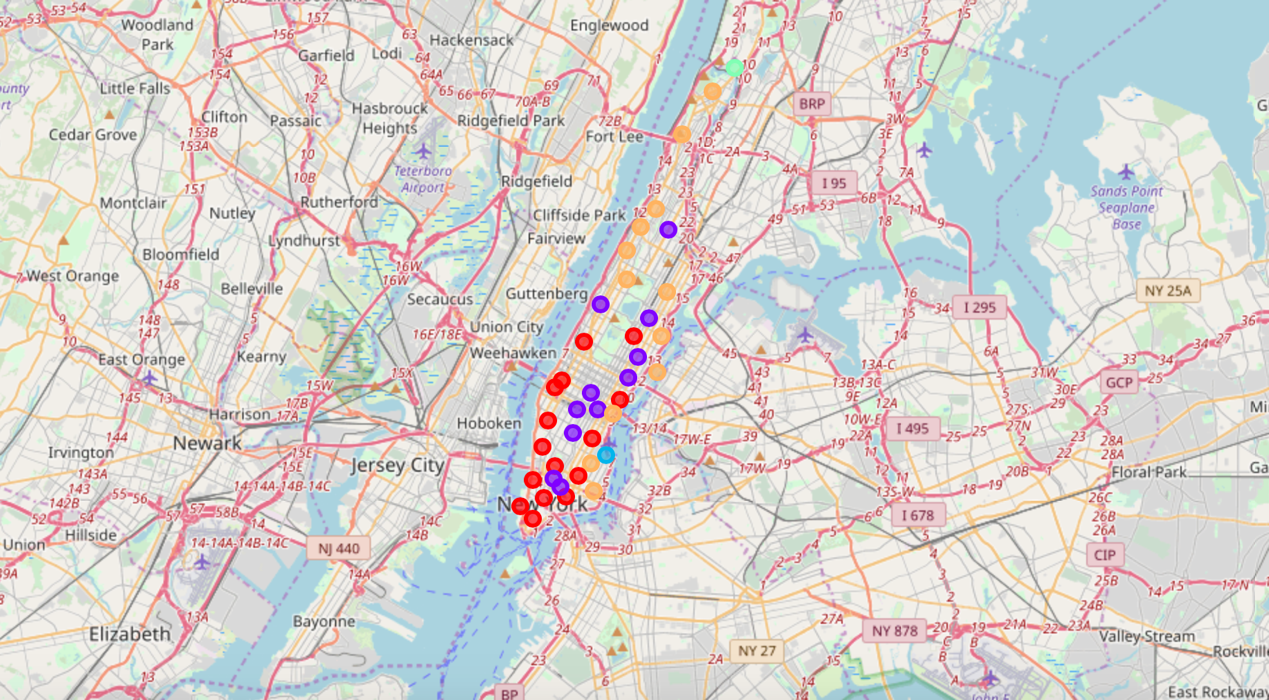
The results of the demographics data consist of three main parts: population, land area, and GDP.



We observe that New York surpasses Toronto in all these three aspects. However, this does not bring a conclusion that New York is a better place to live in. Since New York has more land, it makes sense it to hold more population, thus leading to a higher population. In fact, when we look land area and GDP per person, we will observe that the difference between the two numbers will be subtle – Toronto is slightly higher than New York in land area per person, and New York is slightly higher than Toronto in GDP. Henceforth, we will expect New York to be a little bit faster-more paced and developed.

### 4.2 Clustering

We divide both New York and Toronto neighborhoods into 5 clusters, each of which we represent with a different color. Each point on the plot indicates one neighborhood.



We observe a few differences here. First, all neighborhoods along the waters for Toronto falls into the same cluster, i.e., the one indicated by the blue color, while neighborhoods along the waters for Manhattan consist of all of the 5 categories. Although some may argue that the geographic structure of the two cities are different – Manhattan is almost entirely surrounded by water, I suppose the reasons behind the clustering results are not that simple. A deeper analysis into the characteristics of each neighborhood provides us with more insights: the blue cluster of neighborhoods in Toronto are not that uncommon in terms of public facilities such as restaurants, bars, and cinemas, while the New York clusters are. The cluster seems to perfectly divide New York into continuous segments, each of which with its own very characteristics: most developed area, entertainment area, quiet area, etc. The types of public facilities also differ vastly in these areas: cinemas and bars are much denser in the entertainment area than other places, while the quiet areas consist mainly of households, supermarkets, with very few tourist attractions.

Another observation that aligns with this line of reasoning is the resulting size of the clusters. We clearly see that the sizes of New York clusters are relatively comparable, while more than 80% of the Toronto neighborhoods fall into one cluster. This may also be viewed as an indication that the New York neighborhoods are more categorizable given their own features. Nevertheless, for both New York and Toronto, there exists clusters with only one neighborhood in it. This type of neighborhood is usually far from the rest of the neighborhoods geographically or simply devoid of sufficient data for it to be categorized together with the others.

## 5. Discussion

While we may conclude that the demographic data do not point to any significant difference, the New York and Toronto neighborhood compositions are indeed disparate. The sizes of neighborhoods in New York are more equal within themselves while the whole Toronto is mainly characterized by one giant cluster. Based on the previous results, there are a few recommendations to give for potential visitors to the two cities. First, to fully explore New York, one should go through all of the upper, middle, and lower Manhattan areas so as to experience the different features of each neighborhood cluster; on the other hand, there may not be huge regret for those who fail to cover all boroughs of Toronto, since the composition of neighborhoods there are relatively similar. Second, for visitors to New York, the clustering graph gives a good indication of where to go based on your interest: the entertainment area is very suitable for those who enjoy watching films or theaters, e.g., the Broadway area, while other residence areas are less interesting and quieter.

For residents in the two cities, the message is there exist many differences both within and between the two cities, so it is best to experience the lifestyles and public facilities in the different cluster at least once for each before making a decision of where is the best place to live. Since different people have different predilection for the place to live, more detailed recommendations can be provided given a list of the preferences of the user, for example, close to cinema, full of restaurants, little noise, etc.

## 6. Conclusion

In this project, we performed neighborhood segmentation and clustering on New York and Tokyo, with a detailed analysis of the clustering results that leads to some of our recommendations to the potential visitors and the residents. In fact, the same method we use above can be applied to analyze any city, as long as borough data is available and can be processed and cleaned according to our needs. Therefore, one future work to perform is to extend the number of cities to 3 or more, and analyze which pairs of cities are the most similar or the most different. Another possibility is to delve deeper into our current clustering results. With the help of more premium calls, which involves posts and menus, we can delve deeper into the differences between the public facilities of the same type: for example, one region may be full of restaurant of Indian cuisine and another of Italian cuisine. These regions, categorized as similar currently, will turn out to be very different. Overall, neighborhood clustering and segmentation provides us with a perfect way to analyze the similarities and differences between two or more cities.