## 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.