A description of the data and how it will be used to solve the problem.

D ATA S E T S: To solve the problem at hand, the information was extracted from data available online. Specifically, this study takes the advantage of geo-localized activities exposed on Foursquare. An extensive exploration of the data collected leads me to choose venue as the elementary unit of information. Therefore, I represented them by the most common venues in the rest of this work. Besides this, this study also gathered neighborhoods latitudinal and longitudinal data.

To limit the scope of our study, this study focusses on cities located in the United States (New York) and Canada (Toronto). They were chosen for their high activity according to sqstat.com. The other parts of the world were deliberately excluded as I am not familiar enough with them to correctly evaluate our results. At the time of writing, this dataset is available online.

The New York neighborhood has a total of 5 boroughs and 306 neighborhoods. To segment the neighborhoods and explore them, a dataset containing the 5 boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood. On the other hand, the Canada neighborhood has a total of 5 boroughs and 210 neighborhoods.

Data Exploration: The datasets were explored for micro-analysis, in order to discover features that could characterize places and help cluster them into similar groups. By examining these similar groups with respect to two specific boroughs/areas in the two cities, using k-means clustering analysis this project aims to reveal interesting trends in urban development where the role of geography and culture remains important, even in an era where centralized city planning is dominant.

Foursquare is the most popular LBSN. Its two main purposes are: (i) enable users discover new places, and (ii) let their friends or the world know where they are. The first goal is achieved through venues. Venues have their own web page that displays basic facts (like name, address and type), but also user contributed information (such as photos, reviews, likes and tips) which can be used for rating or recommendation.

K-Means clustering: Since urban growth was supposed to be defined by clustering with population density and local spatial entropy, a handy tool for local area classification would be preferred. Although there exist many statistical packages, the commonly used K-Means clustering was chosen according to Vicker’s comparison on cluster analysis methods. K-Means clustering is capable to partition units by spatial correlation, aiming to make differences among the units in a group, overall groups, is minimized. Furthermore, the K-means algorithm is suitable for the situation that the number of clusters has already been designated. Besides, the K-means algorithm is distance-based, taking distance as the evaluation index of similarity, that is, the closer the distance between two objects is, the larger the similarity is. The algorithm considers the cluster to be composed of objects that are close together, so the compact and independent cluster is the ultimate target.