**Applied Data Science Capstone Assignment – Week 4**

**1. A description of the problem and a discussion of the background**

From the history, we observed that civilization was based on various processes, such as industrial revolution in England, agricultural innovation in ancient Egypt, and transport hub development in Hong Kong. Even though the remarkable difference between current metropolitan areas is insignificant anymore as all multinational companies are operating everywhere, the foundational difference in civil planning, road construction, or constitutional categorization, still exists. For example, you can find easily many small plazas all around Barcelona city while Lisbon has one big plaza in the middle of central area, towards to the harbor are directly. Together, geographical change might influence the formation of cities. Thus, by categorizing what kinds of stores and venues in each neighborhood, we can discuss how two different cities formed differently and how that difference connected to fundamental issues of two cities through the history.

To narrow down the range, I will choose two cities, New York and Toronto. Both cities are the biggest city in the country though not the capital. They are historically important and active harbor cities from 20th centuries. Furthermore, two cities have a lot of similarity, such as having world top class level educational institute(NYU, Columbia, UToronto), positioning as top populous city in their states, and cultural and entertainment hub of each side of the country.

Besides many similarities between two historically and geographically, they are completely different cities in their brand image. While New York is positioning as a vivid, liberal, even chaotic and financial business hub of the east coast, Toronto stands as a safe and peaceful, as well as culturally harmonized community among races.

By looking at composition of venues and popularities, we can discover how two cities are different and what factor may cause that deeply.

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

For the approach, I will collect all local code, neighborhood, latitude, longitude information from two different resources. Wiki and NYU studies will offer free resources on local code and neighborhood. From Foursquare API, we will access to all geographical and venue information of venues in two cities. To conglomerate raw dataset, we will match postal code or latitude and longitude to clarify the list between neighborhood and venue positioning.

In the end, by categorizing venues in each neighborhood and by clustering them, we will segment each city area under a few criteria including coffee shops & bakery, museum, sports and entertainment, finance to observe which part of the city area is specialized to which activity. Also, city clusters comparison will be a key to understand the formation of each city and to find the reason of fundamental difference between them.

For Analysis, I simplified segmentation by neighborhoods. There are 306 neighborhoods In New York city area and 103 neighborhoods in Toronto. From Foursqure API, I imported 6149 venues in New York city with 390 different venue categories and 1319 venues in Toronto with 234 venue categories. For comparison and simplification, I selected top 10 venue categories by occurrences in total. Frequencies and clustering will be used from this simplified total dataset. Here is the example of dataset for further analysis.

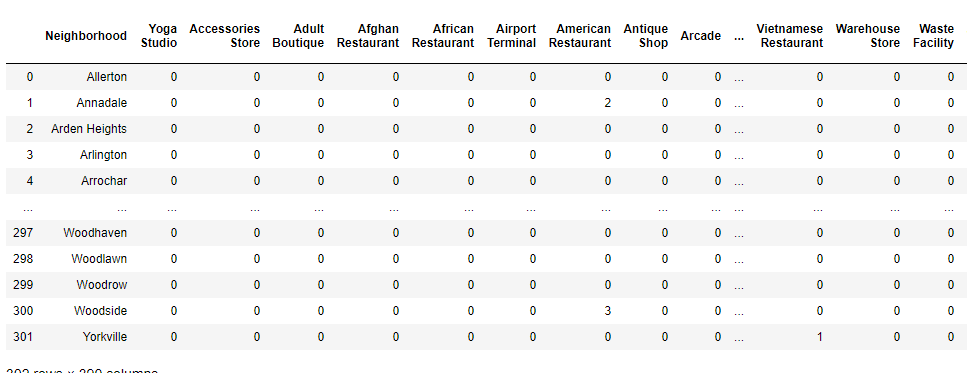


Figure 1. Frequencies of venues categorized by venue categories and neighborhoods in New York city