

Lucy Herr
IBM Applied Data Science Capstone
November 22, 2020

Week 4: Capstone Project - The Battle of Neighborhoods (Part 1 of 2)

1. Clearly define a problem or an idea of your choice, where you would need to leverage the Foursquare location data to solve or execute. Remember that data science problems always target an audience and are meant to help a group of stakeholders solve a problem, so make sure that you explicitly describe your audience and why they would care about your problem.

This submission will eventually become your Introduction/Business Problem section in your final report. So I recommend that you push the report (having your Introduction/Business Problem section only for now) to your Github repository and submit a link to it.

Problem/idea leveraging Foursquare location data:

In 2019, the Greater Toronto Area (GTA) experienced the largest population growth of any city in either the U.S. or Canada ([Petramala & Chan Smyth, 2020](#)). Of the 127,575 individuals who moved to Toronto last year, the majority were immigrants to Canada. According to [CIC News](#), Toronto's growth from immigration is expected to continue in the long-term despite the impacts of the Covid-19 pandemic.

People who move to Toronto from other countries are faced with the task of making important decisions about where they will live, even when they might lack the detailed knowledge of Toronto's neighborhoods that its local residents might have. To address this problem, the plan for my project is to use the city of Toronto's Neighbourhood Profiles dataset and Foursquare location data to generate rankings of the neighborhoods in Toronto based on attributes that might be relevant to people moving into the city, such as average housing costs or top-rated restaurant cuisine.

Target audience:

The audience that this solution targets is the diverse range of individuals who move to Toronto and are looking for information to assist in their decision-making about which neighborhood to move to. Toronto's recent immigrants come from countries all over the world, including India, China, the Philippines, Nigeria, the United States, Pakistan, Syria, Eritrea, South Korea, and Iran ([CIC News](#)). While many are within the 25- to 64-year age range, others are younger or older ([Statistics Canada](#)).

Why the audience would care about your problem:

Regardless of their specific demographic characteristics, people who immigrate to Toronto can benefit from a solution that offers them the information to evaluate neighbourhoods based on the criteria that are most important to them. Toronto features a wide variety of neighborhoods for newcomers to relocate to, and the diversity of both the city and its immigrants means that many different factors might be taken into account when deciding where to move. Decisions might focus

on factors like housing costs, neighborhood safety ratings, or school rankings; they might also involve restaurants and businesses in the area, religious and cultural centers, or parks and recreation areas.

2. Describe the data that you will be using to solve the problem or execute your idea. Remember that you will need to use the Foursquare location data to solve the problem or execute your idea. You can absolutely use other datasets in combination with the Foursquare location data. So make sure that you provide adequate explanation and discussion, with examples, of the data that you will be using, even if it is only Foursquare location data.

This submission will eventually become your Data section in your final report. So I recommend that you push the report (having your Data section) to your Github repository and submit a link to it.

As we did in last week's assignment, I will first create a *pandas* dataframe of Toronto's neighborhoods, boroughs, and postal codes using the **Wikipedia data table** (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) and add longitude and latitude coordinate values using the **Toronto geospatial data CSV file** (http://cocl.us/Geospatial_data).

Next, I plan to leverage **Foursquare location data** for each neighborhood, including neighbourhood, borough, city/borough limit, latitude/longitude, and venue category/rating data. To give a specific example, I will create dataframe columns that highlight top-rated restaurants based on cuisine, top-rated parks and recreation areas, and cultural or community centers, as these attributes might be more relevant to people deciding between the neighborhoods they might move to.

I will also integrate relevant data from the **City of Toronto Neighbourhood Profile Dataset** (<https://www.toronto.ca/city-government/data-research-maps/neighbourhoods-communities/neighbourhood-profiles/>), which is based mainly on Census data and provides detailed info about various neighborhood characteristics. I plan to include median shelter (housing) cost, average commute time, and residents' ethnic origins and languages (among others) in order to round out the Foursquare data with additional information that might be useful to Toronto newcomers.

Ultimately, my goal is to use the integrated dataframe I create to generate Folium maps ranking the top neighborhoods for each criterion (such as restaurant cuisine or housing cost) and, if time permits, specific combinations of criteria (such as best neighborhoods for parents of young children).