

Introduction/Business Problems

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

People moving to different locations are constantly searching for information about their surrounding area. When buying a new home, people always look for what is in proximity to their neighborhood, how much homes in that area cost, nearby schools, crime rates, restaurants, and fun things to do in the area. This project is designed to inform people of the 'important information' when determining whether they will be happy with a move to a new location. In this project, I will continue to use the Scarborough data from the previous assignments to analyze and explore the data to make a well-rounded opinion on a particular neighborhood in the Scarborough area, based on key location information from FourSquare.

We can consider a 'good neighborhood' as one that has a considerable amount of commercial presence, including dining, entertainment, as well as house pricing, schools, and general safety.

Data Description and how it will be used

Scarborough is a popular destination in Canada. It is said to be one of the most diverse and multicultural areas Toronto. This diverse culture is comparable to New York, due to the amounts of immigration. This trend seems to be continuing, especially since immigration in both the United States and Canada have been increasing.

Where am I getting the data?

This project will use FourSquare API as its prime data gathering source. FourSquare has a database of millions of places acquired through crowdsourcing. As a free user of FourSquare's API, certain features are limited and therefore, certain calls to the API are limited throughout the project.

Mainly, to compare the similarities of two cities, I am going to explore the neighborhoods, segment them using KMeans, and then group them into clusters to find similar neighborhoods in New York and Toronto.

This will be done in a Jupyter Notebook, using Python, which requires several libraries such as Pandas, Folium, SciKit Learn, Geopy, and Requests. Pandas will be used to create and manipulate dataframes. Folium will be used to create maps to visualize the neighborhoods. Scikit Learn is the machine learning tool for importing k-means clustering. Geopy will be used to retrieve the location data. Finally, Requests will be used to handle http requests.