Applied Data Science Capstone Project:

Cluster and Segment Neighborhoods in Major Cities for Expats

1. Introduction/Business Problem

As many people leave their home countries to move and work abroad they become expats to that country. To define an expat or expatriate, it is any person who lives temporarily or permanently in a country other than their country of citizenship. Leaving your neighborhood behind and moving to a new neighborhood in a new country can be quiet challenging and someone can still get confused while transitioning into a new culture/social etiquette. Moving into a new country but a similar neighborhood can help expats to quickly adapt to their new environment. Therefore, the approach proposed is to segment and cluster neighborhoods of two major cities. Providing this type of guidance can help expats find a suitable neighborhood in a new country and adjust much faster.

As an expat, before moving into a new country you are expected to do research. This research would consist of recognizing which neighborhood is most suitable to your needs and life style. The easier you make it for you to settle in, meet people with similar hobbies and start to feel at home, the better.

2. Data

To solve this problem, multiple datasets will be used in combination with the Foursquare location data. Data will be used to cluster and segment neighborhoods in two major cities. The two major cities to be taken as an example are Toronto, Canada and New York City, U.S.

The first step in data collection is to extract the list of Toronto and New York City neighborhoods. Luckily, the datasets exist for free on the web. The New York City neighborhoods dataset is published by the New York (City). Department of City Planning and can be found on geo.nyu.edu website which is a spatial data repository maintained by New York University (NYU). The Toronto neighborhoods dataset can be scraped online from Wikipedia which includes the Postcode, Borough, and Neighborhood name.

Next, the Geocoder library can be used to fetch latitude and longitude coordinates for each of the neighborhoods. Adding the geographical coordinates (latitude and longitude) allows to map these neighborhoods using the folium API. Mapping these coordinates provides a better visual to understanding the distribution in each city.

Finally, the Foursquare location API will be used to extract the list of venues surrounding each of the neighborhoods and this list, which contains venues like restaurants/gym/coffee shops/ parks, will be used to cluster and segment neighborhoods in Toronto and New York City. The data will be merged, and further analysis will be performed to clean and prepare it for modeling.