Name: Parth Parikh

Course: Applied Data Science Capstone

2. Data Acquisition and Processing:

2.1 Data Sources:

The neighborhoods and coordinate (latitude and longitude) data for the New York city can be found in this .json file that was provided by IBM previously in this course. The neighborhood data for Toronto was found in parts at two different places. The first part containing the postal codes, borough names, and respective neighborhood names was scraped from a Wikipedia page found under this link. The second part of the data containing the latitude and longitude coordinates of each postal code was used from a previous project where IBM had provided the data here in this .csv format.

Next, the geopy library was used in order to find the coordinate data of a selected address/location when needed.

The Folium library was used to plot the neighborhoods onto a map. The restaurant related data such as their location, coordinates, trends, categories, etc. for each neighborhood in both New York City and Toronto were obtained via extensive use of Foursquare API.

2.2 Data Cleaning:

The raw json file for New York neighborhood data was converted into a pandas dataframe after which all the columns were dropped except 'Neighborhood Name', 'Latitude', and 'Longitude'. For the Toronto data, the Wikipedia scraped data was converted from html to a pandas dataframe. The second part from the csv file was also loaded as a pandas dataframe. Both sources were then merged into a single dataframe by postal codes, after which all columns were dropped except the ones same as New York data.

The Foursquare API provided the list of top-10 most visited restaurants for each neighborhood in both cities. This data was merged with the previous data of each city, thus giving a dataframe containing neighborhood name, coordinates of neighborhood, name of 10-most visited restaurants, coordinates of each restaurant in list, and the category of restaurant.

The data was then prepared for further analysis as needed.