Data:

Mainly two types of data will be used for this project.

1. The first one is the file containing the neighbourhoods of New York. Neighborhood has a total of 5 boroughs and 306 neighborhoods. In order to segement the neighborhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the the latitude and logitude coordinates of each neighborhood.

Luckily, this dataset exists for free on the web. Here is the link to the dataset: https://geo.nyu.edu/catalog/nyu_2451_34572

a) And this will be converted to a DataFrame.

Assumptions made to a DataFrame of neighbourhoods in New York:

- Dataframe will consist of three columns: PostalCode, Borough, and Neighborhood
- Only the cells that have an assigned borough will be processed. Borough that is not assigned are ignored.
- More than one neighborhood can exist in one postal code area. These two rows will be combined into one row with the neighborhoods separated with a comma as shown in row 11 in the above table.
- If a cell has a borough but a Not assigned neighborhood, then the neighborhood will be the same as the borough.

b) Adding geographical coordinates to the neighborhoods

Next important step is adding the geographical coordinates to these neighborhoods. To do so we will be extracting the data present in the Geospatial Data csv file and combining it with the existing neighborhood dataframe by merging them both based on the postal code.

Neighbourhood level data from a variety of other sources are also available through the City's mapping application and here on the Open Data portal.

Each data point in this file is presented for the City's neighbourhoods, as well as for the City of Toronto as a whole. The data is sourced from several Census tables released by Statistics. The general Census Profile is the main source table for this data, but other Census tables have also been used to provide additional information.

c) Scrap the distribution of population from Wikipedia

Another factor that can help us in deciding which neighborhood would be best option to open a restaurant is, the distribution of population based on the ethnic diversity for each neighborhood. As this helps us in identifying the neighborhoods which are densely populated with Indian crowd since that neighborhood would be an ideal place to open an Indian restaurant.

2) The other type of data is from FourSquare to get the location data of venues in the neighbourhoods. Foursquare API is very useful online application used my many developers & other applications like Uber etc. In this project we will use it to retrieve information about the places present in the neighborhoods of New York. The API returns a JSON file and we need to turn that into a data-frame. Here we will choose 100 popular spots for each neighbourhood within a radius of 1km.