Capstone Project - The Battle of Neighborhoods

2. DATA SECTION

For this project, the following data will be taken in account.

1) List of Boroughs and their corresponding Neighborhoods in Toronto, Canada which has been taken from the following Wikipedia page.
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

2) Geographical coordinates of every postal code in Toronto, Canada which have been taken from the following csv file.
https://cocl.us/Geospatial_data

3) Dataset containing the Boroughs and the Neighborhoods that exist in each borough along with the geographical coordinates of each neighborhood which has been taken from the following json file.
https://cocl.us/new_york_dataset

It contains 227,428 check-ins in New York city. The data contains two files in tsv format. Each file contains 8 columns, which are:

1. User ID (anonymized)
2. Venue ID (Foursquare)
3. Venue category ID (Foursquare)
4. Venue category name (Foursquare)
5. Latitude
6. Longitude
7. Time zone offset in minutes (The offset in minutes between when this check-in occurred and the same time in UTC)
8. UTC time

After extracting and reading the data, we will translate the above data into a Pandas data frame for processing which would look like this. These are the data elements that are needed when we call Foursquare web service call in order to get the venues available in that neighborhood (Neighborhoods are not included here)

Then, the machine learning technique, “Clustering” is used to segment the neighborhoods with similar objects on the basis of each neighborhood data. These objects are given priority on the basis of foot traffic (activity) in their respective neighborhoods. This will help to locate the tourist’s areas and hubs, and then we can judge the similarity or dissimilarity between two cities on that basis.