# Methodology

## Description of Data and Solution

**We would be needing the Neighborhood information of 2 Cities here**

1) **New York**
2) **Toronto**

**New York neighborhood data with latitude and longitudes are already available in the link - https://cocl.us/new_york_dataset**

**This data is json format, this must be converted into Tabular format.**

**Toronto dataset is not readily available, but we can find the neighborhood information with zip codes Wikipedia page**

**https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M**

**This must be web scraped and cleansed before the same can be used.**

- **Drop the postcodes with Borough's 'Not Assigned'**
- **Group the Neighborhood with more than one post code**
- **Replace Neighborhood names which are 'Not Assigned' with respective Borough names.**

**Latitude and longitude data for Toronto neighborhoods can be found in this location**

**http://cocl.us/Geospatial_data**

**Once the above 2 data sets are gathered, we would require Venue details which are selling coffee around 500 meters of the above-mentioned neighborhoods. This data can be acquired from foursquare using explore api with relevant filters.**

**Once the data is cleansed and Venue details are captured, the same can be split into 2 data frames**

1) **coffee companies selling coffee**
2) **Other Outlets selling coffee**

**The same needs to be visualized using Folium on a New York map to see the distribution of Non-Coffee shops, and the same can be used to estimate the possibility of opening a coffee shop in the vicinity.**

**We can also do the same analysis for Toronto, which later can be used to cluster with New York to see which Coffee Drinking Neighborhood in Toronto is similar to a neighborhood in New York.**