2. Data

For the purpose of this research, we obtain geographical data from the cities of Toronto, New York, and London. This includes information regarding the neighborhoods of each city and their coordinates (longitude and latitude) which were scraped and transformed into *pandas* data frames from the following websites.

1. **Toronto**: The Wikipedia page for the list of postal codes in Toronto and this csv file from Coursera for the geospatial coordinates.
2. **New York**: This Geonames website for the list of New York postal codes and coordinates, which were linked to their respective neighborhoods through this NYV government pdf file.
3. **London**: This Wikipedia page for the list of areas of London wherein the coordinates were easily obtained from Britain's OS Grid.

This resulted in a dataset of 103 neighborhoods in Toronto, 42 neighborhoods in New York, and 531 neighborhoods in London. Toronto had 77 non-assigned postal codes which we opted to exclude from the dataset.

The coordinates are used to determine the venues within a 500-meters radius of the neighborhoods. Herein, the data on the venue's names and venue categories around each coordinate is gathered from the Foursquare API[1], and are limited to 100 venues for each neighborhood. After filtering out the neighborhoods with no venues within a 500-meters radius which resulted in *NaN* values, there were 100 neighborhoods in Toronto, 42 neighborhoods in New York, and 527 neighborhoods in London.

Table 1. Number of Observations Before and After Data Cleanup

Toronto Not Assigned: 77

	City	NaN Rows	Cleaned up Rows	Total Rows
0	London	4	527	531
1	New York	0	42	42
2	Toronto	3	100	103

[1] Using the following URL and explore query:
https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}