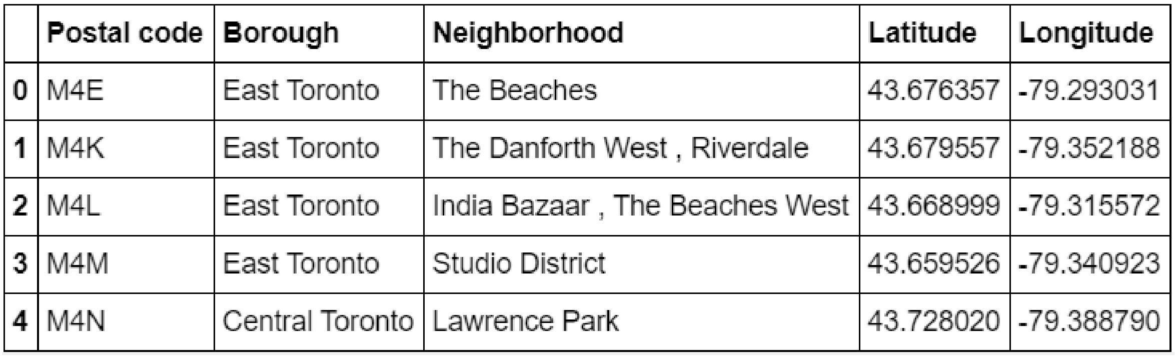
The Battle of Neighborhoods

(Marco Gallegos)

Data Description

The goal of this project is the segmentation, exploration and clustering of neighborhoods in Toronto. The key data for this task to be achieved is the Toronto neighborhood data. The structured data for this purpose is not readily available on the inter-net in structured format. Therefore we will scrap the data from the inter-net using scraping libraries in python to achieve a structured dataset from Wikipedia Toronto neighborhood data page which contains all the necessary information needed for this project. The structured dataset should contain the latitudes and longitudes of the each and every neighborhood for exploring and analyzing further venues. The data needs to be cleaned and prepared and the key information factors will be needed in the final recommendation report to the management. The following data is critical for the project:

1. Neighborhood Name
2. Neighborhood Latitude
3. Neighborhood Longitude
4. Number of residences in each Neighborhood



	Postal code	Borough	Neighborhood	Latitude	Longitude
0	M4E	East Toronto	The Beaches	43.676357	-79.293031
1	M4K	East Toronto	The Danforth West , Riverdale	43.679557	-79.352188
2	M4L	East Toronto	India Bazaar , The Beaches West	43.668999	-79.315572
3	M4M	East Toronto	Studio District	43.659526	-79.340923
4	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790

Data Features

We use a reliable location information provider such as the Foursquare.com to explore the various types of venues and its categories available in each neighborhood. We will also understand the trending of these venues in their respective neighborhoods. The informative dataset gathered should be containing the following information columns in a structured form:

1. Neighborhood
2. Neighborhood Latitude
3. Neighborhood Longitude
4. Venue Name

5. Venue Category
6. Venue Latitude
7. Venue Longitude

Conclusion

Clustering methods like K-means Clustering can be used to segment and cluster these neighborhoods for grouping them and understanding their similarities and creating the model accordingly.

With the above defined features and techniques with all the collected and prepared data, we can come up with the clusters of neighborhoods which are best and worse suited to be suggested to the management of the (XYZ) on-line grocery store for starting the first on-line delivery service for their store. Neighborhoods with high competition and less demands will be avoided by the model as they are not suited best for the company for profit maximization and cost cutting.