

2. Acquiring and Pre-processing data

2.1 Data Acquisition

Data regarding the List of postal codes, Boroughs and Neighborhoods of Toronto, Canada was obtained by scraping data from the following Wikipedia page:

[List of postal codes of Canada: M](#)

The obtained response was transformed into a pandas dataframe:

	PostalCode	Borough	Neighborhood
0	M1A\n	Not assigned\n	\n
1	M2A\n	Not assigned\n	\n
2	M3A\n	North York\n	Parkwoods\n
3	M4A\n	North York\n	Victoria Village\n
4	M5A\n	Downtown Toronto\n	Regent Park / Harbourfront\n
...
175	M5Z\n	Not assigned\n	\n
176	M6Z\n	Not assigned\n	\n
177	M7Z\n	Not assigned\n	\n
178	M8Z\n	Etobicoke\n	Mimico NW / The Queensway West / South of Bloo...
179	M9Z\n	Not assigned\n	\n

Geospatial Data corresponding to each postal code was obtained from the link provided in the course: [Geospatial data](#)

	PostalCode	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

All nearby venues within 1 KM radius of each neighborhood was collected using the explore call to the Foursquare API:

	Neighborhood	Venue	Latitude	Longitude	Category
0	Parkwoods	Allwyn's Bakery	43.759840	-79.324719	Caribbean Restaurant
1	Parkwoods	Brookbanks Park	43.751976	-79.332140	Park
2	Parkwoods	Tim Hortons	43.760668	-79.326368	Café
3	Parkwoods	A&W	43.760643	-79.326865	Fast Food Restaurant
4	Parkwoods	Bruno's valu-mart	43.746143	-79.324630	Grocery Store
5	Parkwoods	High Street Fish & Chips	43.745260	-79.324949	Fish & Chips Shop
6	Parkwoods	Food Basics	43.760549	-79.326045	Supermarket
7	Parkwoods	Shoppers Drug Mart	43.745315	-79.325800	Pharmacy
8	Parkwoods	Shoppers Drug Mart	43.760857	-79.324961	Pharmacy
9	Parkwoods	Variety Store	43.751974	-79.333114	Food & Drink Shop
10	Parkwoods	Pizza Pizza	43.760231	-79.325666	Pizza Place
11	Parkwoods	DVP at York Mills	43.758899	-79.334099	Road
12	Parkwoods	TTC Stop #09083	43.759655	-79.332223	Bus Stop
13	Parkwoods	TTC Stop 9083	43.759251	-79.334000	Bus Stop
14	Parkwoods	Sandover Park	43.760277	-79.333305	Park
15	Parkwoods	TTC Stop #9075	43.757596	-79.338155	Train Station
16	Parkwoods	Dollarama	43.760341	-79.325519	Discount Store
17	Parkwoods	Parkwoods Coin Laundry	43.760386	-79.324894	Laundry Service
18	Parkwoods	Spicy Chicken House	43.760639	-79.325671	Chinese Restaurant
19	Parkwoods	La Notre	43.760704	-79.325396	Coffee Shop
20	Parkwoods	Underhill Mini Mart Convenience	43.745836	-79.324835	Convenience Store
21	Parkwoods	Parkwoods Village Centre	43.760735	-79.324873	Shopping Mall
22	Parkwoods	Family Food Fair Convenience	43.760620	-79.324459	Convenience Store
23	Parkwoods	Broadlands Skating Rink	43.746689	-79.322678	Skating Rink
24	Parkwoods	Parkway Valley Tennis Club	43.754481	-79.318285	Tennis Court

2.2 Data Cleaning

The data obtained by scraping the Wikipedia page was not ready to be used for analysis straightaway. The following pre-processing operations were performed on data before it could be used for further analysis:

- Removing newline characters
- Replace '/' with a ',' as a separator for multiple neighborhoods
- Removing Rows where the Borough is 'Not assigned'
- Removing extra spaces from the 'PostalCode' column
- Merging the two dataframes based on the PostalCode Column

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park , Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor , Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park , Ontario Provincial Government	43.662301	-79.389494
...
98	M8X	Etobicoke	The Kingsway , Montgomery Road , Old Mill North	43.653654	-79.506944
99	M4Y	Downtown Toronto	Church and Wellesley	43.665860	-79.383160
100	M7Y	East Toronto	Business reply mail Processing Centre	43.662744	-79.321558
101	M8Y	Etobicoke	Old Mill South , King's Mill Park , Sunnylea , ...	43.636258	-79.498509
102	M8Z	Etobicoke	Mimico NW , The Queensway West , South of Bloo...	43.628841	-79.520999

Similarly, the venues data returned from the Foursquare API was also cleaned removing any unnecessary spaces and characters that would hinder further comparison and analysis.