1. Data Acquisition and Cleaning

1.1 Data Sources

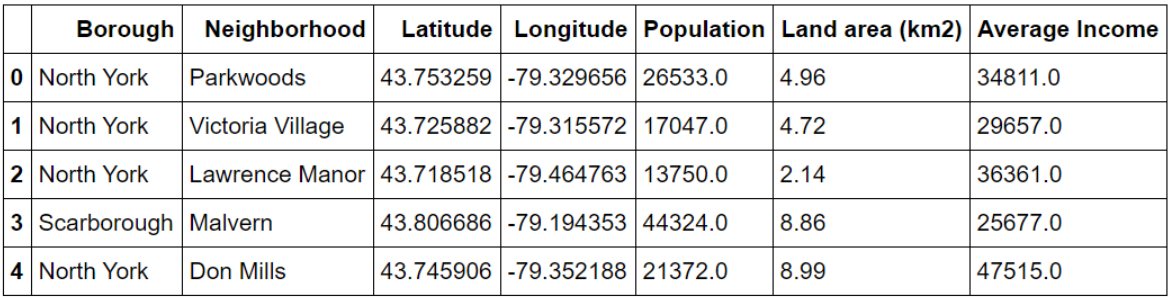
Based on definition of our problem, factors that will influence our decision include, number of existing gymnasiums in the neighbourhoods, population in the neighbourhood, population density, average income etc. Toronto neighbourhood data and population information could be obtained by scraping Wikipedia. That is from this link and this link.

Nearby venues of each neighbourhood can be obtained from www.foursquare.com. Thus, we can filter out the number of gymnasiums in every neighbourhoods from the received data.

1.2 Data Cleaning

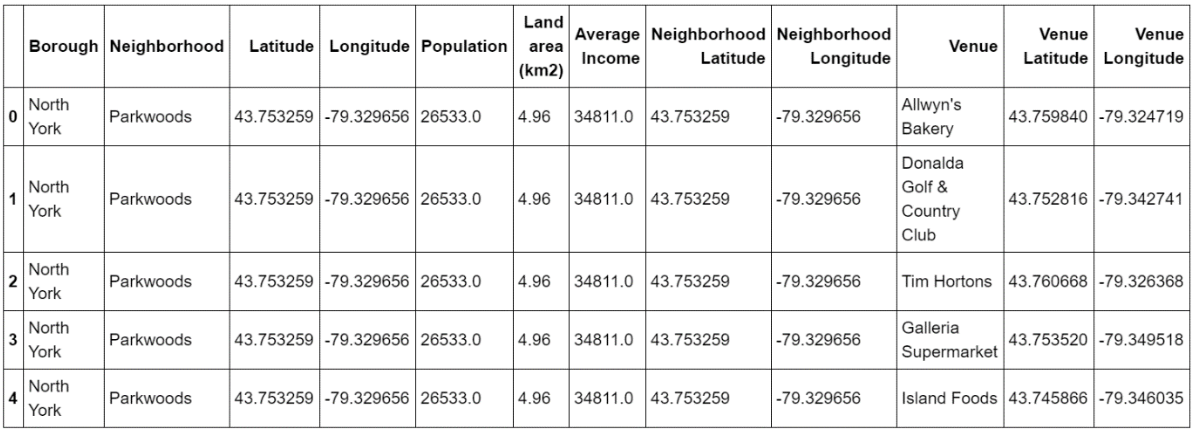
Toronto neighbourhood data scraped contains some neighbourhood's data missing and I decided to drop it. Another problem was multiple entries for single neighbourhood field, so I separated it into multiple fields. Collected latitude and longitude information of each neighbourhoods and merged both together.

Population related data was also scraped and merged with the above dataset. After this step dataset looked like this:



	Borough	Neighborhood	Latitude	Longitude	Population	Land area (km2)	Average Income
0	North York	Parkwoods	43.753259	-79.329656	26533.0	4.96	34811.0
1	North York	Victoria Village	43.725882	-79.315572	17047.0	4.72	29657.0
2	North York	Lawrence Manor	43.718518	-79.464763	13750.0	2.14	36361.0
3	Scarborough	Malvern	43.806686	-79.194353	44324.0	8.86	25677.0
4	North York	Don Mills	43.745906	-79.352188	21372.0	8.99	47515.0

Next step is to find nearby venues of each neighbourhood from foursquare API. By sending location details (latitude, longitude) we will get venues closer to each neighbourhood and add that data with the above table. So, our final table looks as:



	Borough	Neighborhood	Latitude	Longitude	Population	Land area (km2)	Average Income	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude
0	North York	Parkwoods	43.753259	-79.329656	26533.0	4.96	34811.0	43.753259	-79.329656	Allwyn's Bakery	43.759840	-79.324719
1	North York	Parkwoods	43.753259	-79.329656	26533.0	4.96	34811.0	43.753259	-79.329656	Donalda Golf & Country Club	43.752816	-79.342741
2	North York	Parkwoods	43.753259	-79.329656	26533.0	4.96	34811.0	43.753259	-79.329656	Tim Hortons	43.760668	-79.326368
3	North York	Parkwoods	43.753259	-79.329656	26533.0	4.96	34811.0	43.753259	-79.329656	Galleria Supermarket	43.753520	-79.349518
4	North York	Parkwoods	43.753259	-79.329656	26533.0	4.96	34811.0	43.753259	-79.329656	Island Foods	43.745866	-79.346035